Analiza prodaje motora višefaktorskom analizom varijance i analizom korespondencije

Iljadica-Rapo, Anđela

Master's thesis / Diplomski rad

2017

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Science / Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:217:066015

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2025-03-22**

Repository / Repozitorij:

Repository of the Faculty of Science - University of Zagreb

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Anđela Iljadica-Rapo

ANALIZA PRODAJE MOTORA VIŠEFAKTORSKOM ANALIZOM VARIJANCE I ANALIZOM KORESPONDENCIJE

Diplomski rad

Voditelj rada:
prof. dr. sc. Anamarija Jazbec

Zagreb, rujan, 2017

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Za mog tatu

Sadržaj

Popis engleskog nazivlja iz SAS tablica

Adj R-Sq - prilagođeni koeficijent determinacije (*engl. Adjusted R-square*)
Analysis of Variance (ANOVA) - analiza varijance
Coeff Var - koeficijent varijacije
Col Pct - relativna frekvencija stupca
Column Profiles - profili stupaca
Contingency Coefficient - kontingencijski koeficijent, mjera statističke jačine izvedena iz Pearsonove χ^2-statistike
Contingency Table - kontingencijska tablica
Contributions to the Total χ^2-Statistics - doprinosi ukupnoj χ^2-statistici
Corrected Total - ukupno ispravljeno/suma kvadrata ukupnih odstupanja
Cramer's V - Cramerov V, mjera statističke jačine izvedena iz Pearsonove χ^2-statistike
Cumulative Frequency - kumulativna frekvencija
Degrees of Freedom (DF) - stupnjevi slobode
Distribution - razdioba
Error - greška
Expected Values - očekivane vrijednosti
Frequency (Freq) - frekvencija
Indices of the Coordinates - oznaka koordinata
Inertia - inercija
Intercept -slobodni član
Least Square Means - očekivanje najmanjih kvadrata
Likelihood Ratio Chi-Square - χ^2 omjer vjerodostojnosti
Mantel-Haenszel Chi-Square - Mantel-Haenszelova χ^2 statistika
Mass - masa
Mean - očekivanje
Mean Square - procjenitelj varijance (sredina kvadrata)
Model - model
Parameter Estimate - procjena parametara
Percent - relativna frekvencija
Phi Coefficient - koeficijent koji daje korijen χ^2-statistike podijeljen s veličinom uzorka
Prob - vjerojatnost (*engl. Probability*)
R-Square ($\mathbb{R}^2$) - koeficijent determinacije
Root MSE - drugi korijen prosječne kvadratne greške modela
Row Pct - relativna frekvencija retka
Singular Value - singularna vrijednost
Source - izvor

Standard Error - standardna pogreška
Statistic - statistika
Sum of Squares - suma kvadrata
Summary Statistics for the Row/Column Points - rezime statistike za retke/stupce
Type I SS - tip sume kvadrata I
Type III SS - tip sume kvadrata III
Value - vrijednost , *F-Value* - vrijednost F-statistike, *t-Value* - vrijednost t-statistike
Variable - varijabla, *Dependent Variable* - zavisna varijabla
Quality - kvaliteta

Uvod

U ovom radu se bavimo analizom prodaje motora firme Iljadica-Rapo d.o.o iz Šibenika. Firma je osnovana 1989. godine i prvi je prodavatelj Piaggio, Vespa i Gilera motora na hrvatskom tržištu. Krajem 2008. godine u Republici Hrvatskoj je nastupila globalna ekonomska kriza koja je poprilično pogodila sektor trgovine. Prije krize broj prodanih motora u Šibensko-kninskoj i Splitsko-dalmatinskoj županiji bio je znatno veći nego u svim ostalim županijama RH, a upravo je Šibensko-kninska županija imala najveću prodaju motora po broju stanovnika. Za usporedbu, jedna trgovina u Šibeniku prodavala je koliko četiri trgovine u Zagrebu zajedno.
Prvih godina rada firme podaci o prodaji su bili na papirima, no zahvaljujući tehnološkom napretku, počeli su se spremati na računala u baze podataka. Danas se gotovo sve može prikazati pomoću brojeva, odnosno statističkih podataka i analiza koje mogu pomoći firmi u donošenju učinkovitijih poslovnih strategija. Razmotrit ćemo kako se kretala prodaja, koje su marke bile popularnije prije desetak godina a koje su danas. Također, analizirati ćemo mijenja li se jačina motora koji se prodaju i koje boje su najpopularnije. Testirat ćemo što utječe na prodaju, te grafičkim prikazima prikazati odnose nekih varijabli baze podataka.
Otprilike znamo da je prodaja prvih godina rada firme naglo rasla. Prodavali su se motori svih boja i svih jačina, no početkom pada prodaje jači motori su se skoro prestali prodavati, nastavili su se prodavat samo najslabiji, ujedno i najjeftiniji. Od početka je najpopularnija marka bila Piaggio, a Gilera i Aprilia su je pratile. Moto Guzzi marka ima samo najjače motore, dakle ona se u jednom trenutku prestala prodavati. Također, u prodaju je uvedena marka Derbi koju simboliziraju najslabiji i najjeftini motori. Bez statistike otprilike znamo situaciju na tržištu, no statistika nam može pomoći u vjerodostojnijem prikazu prodaje.

Poglavlje 1

Opisna statistika

1.1 O bazi podataka

Baza koju koristimo izvučena je iz baze podataka Piaggio Centra Iljadica-Rapo koja se zasniva na ORACLE bazi podataka koju održava poduzeće POS d.o.o. iz Splita. Uzeti su podaci o motorima prodanim od 2006. do 2016. godine. Baza sadrži 580 observacija. Svaka observacija sadrži godinu, marku motora, kubikažu (tj. jačinu motora), boju, cijenu, tip motora te broj prodanih motora s tim svojstvima. Na Slici 1.1 prikazane su prve tri observacije baze.

BAZA

Obs	godina	boja	kubikaza	tip	br_prodanih	cijena	marka
1	2016	crna	50	Fly	9	10700	Piaggio
2	2016	bijela	50	Fly	11	10970	Piaggio
3	2016	crna	50	Motard	5	11650	Aprilia

Slika 1.1: Isječak iz baze (ispis iz SAS-a)

Marke su: Piaggio, Aprilia, Derbi, Gilera i Moto Guzzi.
Boje su: bijela, crna, crvena, ostalo, special, gdje pod ostalo spadaju ne toliko popularne boje motora (smeđa, narančasta, siva, zelena), a pod special spadaju boje s posebnim efektima.

Cijene su u rasponu od 6890,00 kn do 78500,00 kn. Pomoću PROC FORMATA u programskom sustavu SAS cijene su podijeljene u 4 kategorije: $\leqslant$9999,00 kn, 10000,00-19999,00 kn, 20000,00-29999,00 kn, $\geqslant$30000,00 kn.

Raspon kubikaže je od 50 do 1100 kubika, i one su podijeljene u 3 kategorijske skupine: moped (50 kubika), skuter (100 do 599 kubika) te motor ($\geqslant$ 600 kubika). Ima 40 različitih tipova motora prodanih u razdoblju od 2006. do 2016. godine, no kako svih godina nisu bili u prodaji izdvojeno je 8 najpopularnijih Beverly, Boulevard, NRG, Pegaso, Runner, Sportcity, Vespa, ZIP.

1.2 Programski sustav SAS

Programski sustav SAS (Statistical Analysis System) se počeo razvijati između 1966. i 1976. godine na sveučilištu North Carolina u SAD-u za potrebe poljoprivrednih istraživanja. Ubrzo se pokazao jako korisnim u svim granama industrije. Tvorci sustava su A. Barr, J. Goodnight i J. Sall.
SAS je integrirani aplikacijski sustav koji može mijenjati podatke i upravljati njima sa različitih izvora i u različitim formatima te vršiti statističku analizu nad njima. Na vrlo jednostavan način omogućuje elementarnu, ali i sofisticiranu analizu podataka uporabom "point and click" tehnike rada preko grafičkih korisničkih sučelja (*engl. Grafical User Interface - GUI*), gotovih programa - SAS procedura, ili programiranjem. Kako se statističke metode usavršavanju, tako programski sustav SAS ide u korak s njima.[8], [2]
U izradi ovog rada korišteni su SAS University Edition i Sas On Demand for Academcis. SAS University je dostupan za skidanje preko web pretraživača te se može koristiti preko virtualne mašine bez upotrebe interneta, dok se u SAS On Demandu radi preko pretraživača weba.[9]

1.3 Tablice frekvencija

Za sve kategorijske varijable napravljene su tablice frekvencija. Frekvencija je zastupljenost jedne kategorije u uzorku. Relativna frekvencija je omjer frekvencije i ukupnog broja podataka. Tablice frekvencija nam služe kao alati za prikaz podataka.
Prva znamenka u tablici (*engl. Frequency*) pokazuje frekvenciju, druga relativnu frekvenciju (*engl. Percent*) s obzirom na ukupni broj podataka, treća relativnu frekvenciju (*engl. Row Percent*) s obzirom na ukupni broj podataka retka i četvrta relativnu frekvenciju (*engl. Col Percent*) s obzirom na ukupni broj podataka stupca.

Kod u SAS-u

Koristimo PROC FREQ proceduru, s *data = baza* označavamo skup podataka na kojem radimo. Naredbom TABLES označavamo koje tablice frekvencija želimo, naredbom WEIGHT brojimo prodane motore po observaciji, inače bi se svaka observacija gledala kao jedan subjekt. SAS kod:

PROC FREQ *data = baza;*
*TABLES marka*godina kubikaza*godina boja*godina cijena*godina marka*cijena boja*cijena;*
WEIGHT brprodanih;
run;

Provjerimo kako je tekla prodaja različitih marki motora u promatranom razdoblju.

Tablica 1.1: Tablica frekvencija marki motora po godinama promatranog razdoblja (ispis iz SAS-a)

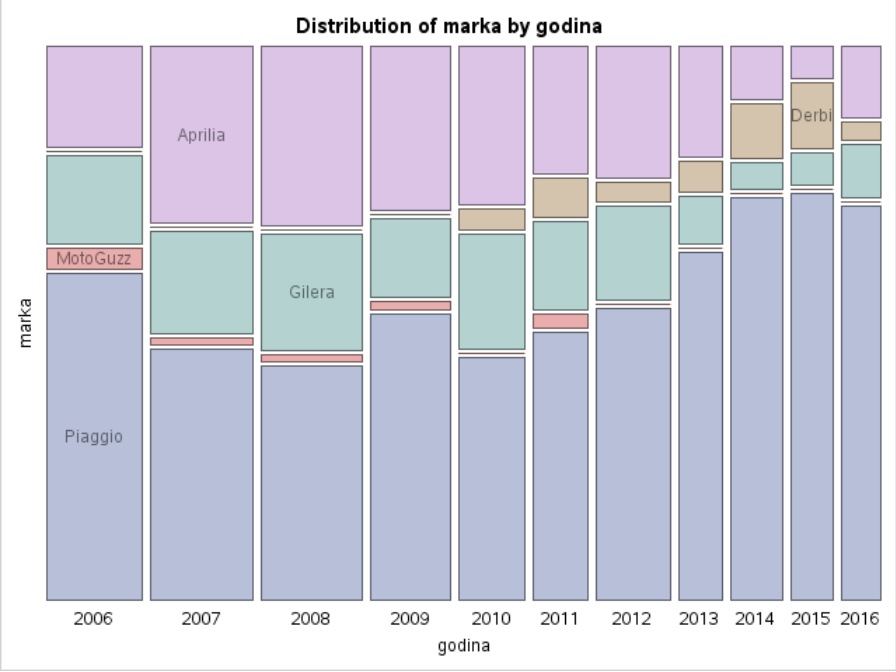
Frequency
Percent
Row Pct
Col Pct

Table of marka by godina

marka	godina 2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	Total
Aprilia	59 2.01 12.80 10.97	109 3.72 23.64 18.32	73 2.49 15.84 15.70	45 1.53 9.76 18.91	50 1.70 10.85 22.83	47 1.60 10.20 22.71	44 1.50 9.54 21.05	18 0.61 3.90 12.24	4 0.14 0.87 3.31	3 0.10 0.65 3.37	9 0.31 1.95 8.57	461 15.72
Derbi	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	4 0.14 6.78 1.83	22 0.75 37.29 10.63	9 0.31 15.25 4.31	2 0.07 3.39 1.36	15 0.51 25.42 12.40	6 0.20 10.17 6.74	1 0.03 1.69 0.95	59 2.01
Gilera	146 4.98 29.55 27.14	107 3.65 21.66 17.98	88 3.00 17.81 18.92	29 0.99 5.87 12.18	37 1.26 7.49 16.89	30 1.02 6.07 14.49	25 0.85 5.06 11.96	7 0.24 1.42 4.76	4 0.14 0.81 3.31	7 0.24 1.42 7.87	14 0.48 2.83 13.33	494 16.84
MotoGuzz	7 0.24 63.64 1.30	1 0.03 9.09 0.17	1 0.03 9.09 0.22	1 0.03 9.09 0.42	0 0.00 0.00 0.00	1 0.03 9.09 0.48	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	11 0.38
Piaggio	326 11.11 17.09 60.59	378 12.89 19.81 63.53	303 10.33 15.88 65.16	163 5.56 8.54 68.49	128 4.36 6.71 58.45	107 3.65 5.61 51.69	131 4.47 6.87 62.68	120 4.09 6.29 81.63	98 3.34 5.14 80.99	73 2.49 3.83 82.02	81 2.76 4.25 77.14	1908 65.05
Total	538 18.34	595 20.29	465 15.85	238 8.11	219 7.47	207 7.06	209 7.13	147 5.01	121 4.13	89 3.03	105 3.58	2933 100.00

U zadnjem stupcu Tablice 1.1 vidimo koliko ukupne prodaje u promatranom razdoblju otpada na pojedinu marku. Marka motora Piaggio je očigledno dominantna, 65,05% ukupnog tržišta otpada na nju, a slijede ju marke Gilera (16,84%) i Aprilia (15,72%). Do 2009. godine, Piaggio motora se prodavalo više od 300 motora po godini, no tada je nastupila kriza i 2009. ih se prodalo 163, a nakon 2009. godine

taj broj pada na ispod 100 prodanih motora godišnje.
Marka motora Gilera je početkom promatranog razdoblja činila između 20 i 30% ukupnog tržišta, ali nakon toga je doživjela pad koji je trajao do 2016. godine. Marka motora Aprilia se dobro održala u krizi obzirom na ostale marke, no ipak od 2014. godine nastupa pad u prodaji marke motora Aprilia. Derbi marka motora je uvedena na tržište 2010. godine, imala je dvije dobre godine prodaje, no zadnje godine njen udio na tržištu je manji od 1%. Marka motora Moto Guzzi se prestala prodavati nakon 2011. godine. Možemo vidjeti da se 2016. godine povećala prodaja Piaggio, Aprilia i Gilera marki motora.
Podaci iz Tablice 1.1 prikazani su mozaik prikazom na Slici 1.2. Mozaik prikaz (*engl. mosaic plot*) je grafički prikaz koji se sastoji od pravokutnika. Vrlo je sličan strukturiranom stupičastom grafu, samo što je širina stupaca proporcionalna relativnoj frekvenciji varijable na x-osi. Uočimo kako se pravokutnici smanjuju kroz godine (pad prodaje), te kako se Piaggio marka najbolje održala. Iako se prodaja marke Piaggio drastično smanjila, zadnje četiri godine se povećao njen udio na tržištu. Iz Tablice 1.1 vidimo da je prije činila 60-70% tržišta, a zadnjih godina 80% tržišta čini Piaggio motori.



Slika 1.2: Mozaik prikaz marki prodanih motora po godinama promatranog razdoblja (ispis iz SAS-a)

U Tablici 1.2 prikazani su podaci prodaje motora po jačinskim razredima.

Tablica 1.2: Tablica frekvencija kubikaže (jačine) motora po godinama promatranog razdoblja (ispis iz SAS-a)

The FREQ Procedure

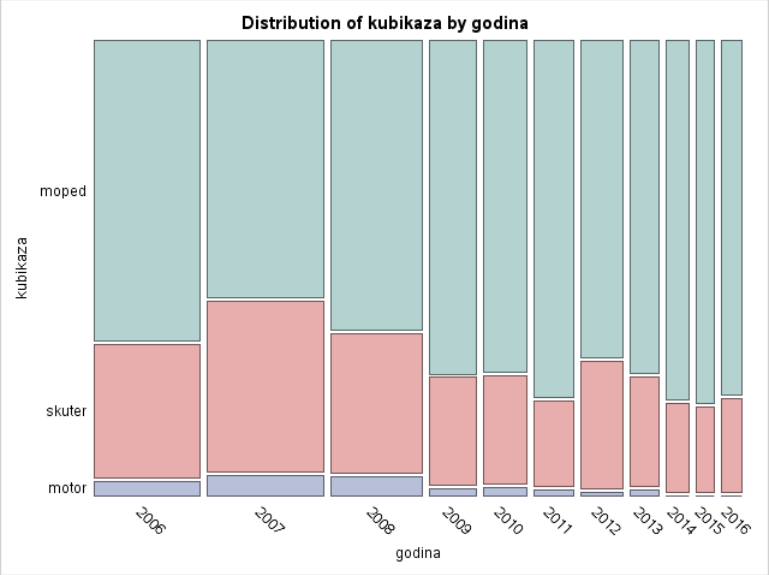
Table of kubikaza by godina

Frequency
Percent
Row Pct
Col Pct

kubikaza	godina											
	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	Total
moped	361 12.31 17.91 67.10	342 11.66 16.96 57.48	300 10.23 14.88 64.52	177 6.03 8.78 74.37	162 5.52 8.04 73.97	165 5.63 8.18 79.71	148 5.05 7.34 70.81	109 3.72 5.41 74.15	97 3.31 4.81 80.17	72 2.45 3.57 80.90	83 2.83 4.12 79.05	2016 68.74
skuter	159 5.42 19.00 29.55	226 7.71 27.00 37.98	145 4.94 17.32 31.18	57 1.94 6.81 23.95	53 1.81 6.33 24.20	39 1.33 4.66 18.84	59 2.01 7.05 28.23	36 1.23 4.30 24.49	24 0.82 2.87 19.83	17 0.58 2.03 19.10	22 0.75 2.63 20.95	837 28.54
motor	18 0.61 22.50 3.35	27 0.92 33.75 4.54	20 0.68 25.00 4.30	4 0.14 5.00 1.68	4 0.14 5.00 1.83	3 0.10 3.75 1.45	2 0.07 2.50 0.96	2 0.07 2.50 1.36	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	80 2.73
Total	538 18.34	595 20.29	465 15.85	238 8.11	219 7.47	207 7.06	209 7.13	147 5.01	121 4.13	89 3.03	105 3.58	2933 100.00

Iz zadnjeg stupca (Total) Tablice 1.2 vidimo da mopedi čine čak 68,74% tržišta, skuteri slijede nakon njih s 28,54% , a motori se najmanje prodaju, 2,73% tržišta otpada na njih. Ipak je to najskuplja skupina motora, najvećih kubikaža.
Gledajući frekvencije možemo uočiti kako se prodaja mopeda, skutera i motora dosta smanjuje od 2006. godine prema 2016. Uočimo drastičan pad u prodaji nakon 2008. godine. Do tada se prodavalo više od 300 mopeda godišnje, a nakon tog broj prodanih mopeda nije prešao 180 po godini. Prodaja skutera također pada nakon 2008. godine, a prodaja motora je stala nakon 2013. godine.
Porast u prodaji mopeda i skutera možemo uočiti tek 2016. godine. Promotrimo li relativne frekvencije s obzirom na ukupan broj prodanih motora po svakoj od godina, možemo uočiti da se padom prodaje povećava postotak mopeda na tržištu. Do 2008. godine činili su do 70% tržišta, dok u novije vrijeme čine otprilike 80% ukupnog tržišta. U zadnjem retku (Total) Tablice 1.2 čitamo koliko je od ukupnog broja prodanih motora prodano u određenoj godini.

Sve što smo upravo analizirali iz tablice frekvencija možemo vidjeti na mozaik prikazu, Slika 1.3. Uočimo kako se pravokutnici smanjuju po godinama (pad prodaje), koliki udio na tržištu pripada motorima najmanjih jačina (mopedima) i kako se taj udio povećao kad je nastupila kriza.
Također, dolaskom krize vidi se pad prodaje motora najvećih kubikaža, koji potom i nestaju iz prodaje.



Slika 1.3: Mozaik prikaz kubikaža prodanih motora po godinama promatranog razdoblja (ispis iz SAS-a)

Promotrimo kako se kreće prodaja motora po bojama.

Tablica 1.3: Tablica frekvencija boja motora po godinama promatranog razdoblja (ispis iz SAS-a)

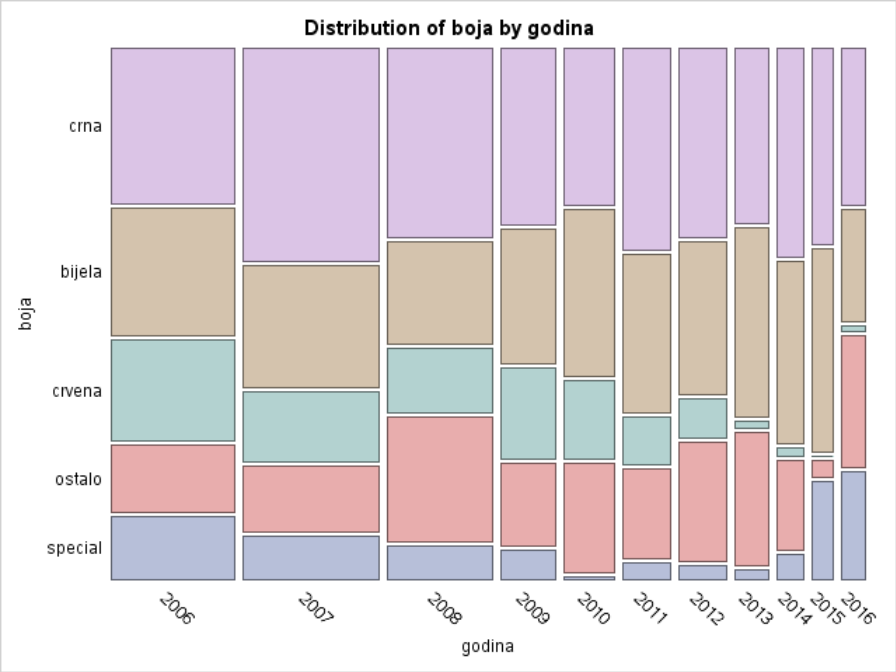
Frequency
Percent
Row Pct
Col Pct

Table of boja by godina

boja	godina 2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	Total
crna	163 5.56 15.49 30.30	246 8.39 23.38 41.34	171 5.83 16.25 36.77	82 2.80 7.79 34.45	67 2.28 6.37 30.59	81 2.76 7.70 39.13	77 2.63 7.32 36.84	50 1.70 4.75 34.01	49 1.67 4.66 40.50	34 1.16 3.23 38.20	32 1.09 3.04 30.48	1052 35.87
bijela	133 4.53 17.03 24.72	142 4.84 18.18 23.87	92 3.14 11.78 19.78	62 2.11 7.94 26.05	71 2.42 9.09 32.42	64 2.18 8.19 30.92	62 2.11 7.94 29.67	54 1.84 6.91 36.73	43 1.47 5.51 35.54	35 1.19 4.48 39.33	23 0.78 2.94 21.90	781 26.63
crvena	106 3.61 29.44 19.70	80 2.73 22.22 13.45	59 2.01 16.39 12.69	42 1.43 11.67 17.65	33 1.13 9.17 15.07	19 0.65 5.28 9.18	16 0.55 4.44 7.66	2 0.07 0.56 1.36	2 0.07 0.56 1.65	0 0.00 0.00 0.00	1 0.03 0.28 0.95	360 12.27
ostalo	70 2.39 13.54 13.01	76 2.59 14.70 12.77	113 3.85 21.86 24.30	38 1.30 7.35 15.97	47 1.60 9.09 21.46	36 1.23 6.96 17.39	48 1.64 9.28 22.97	38 1.30 7.35 25.85	21 0.72 4.06 17.36	3 0.10 0.58 3.37	27 0.92 5.22 25.71	517 17.63
special	66 2.25 29.60 12.27	51 1.74 22.87 8.57	30 1.02 13.45 6.45	14 0.48 6.28 5.88	1 0.03 0.45 0.46	7 0.24 3.14 3.38	6 0.20 2.69 2.87	3 0.10 1.35 2.04	6 0.20 2.69 4.96	17 0.58 7.62 19.10	22 0.75 9.87 20.95	223 7.60
Total	538 18.34	595 20.29	465 15.85	238 8.11	219 7.47	207 7.06	209 7.13	147 5.01	121 4.13	89 3.03	105 3.58	2933 100.00

Iz Tablice 1.3 čitamo da je najzastupljenija boja crna, nakon nje slijedi bijela boja, crvenih motora je znatno manje, dok je motora special boje najmanje. Ostatak prodaje čine motori koje smo svrstali u skupinu ostalo, gdje spadaju boje motora koje nisu zastupljene u svim markama.
Gotovo svih godina, crna boja je vodeća u prodaji, no nekih godina su ipak bijeli motori bili popularniji, primjerice 2010. godine. Iste godine možemo uočiti pad u prodaji boje special. Također možemo uočiti da je boja special bila dosta popularna 2016. godine. Što se tiče crvene boje, vidimo da je u prvoj polovici promatranog razdoblja činila 15-20% ukupne prodaje po godini, no u drugoj polovici čini znatno manje, a 2015. godine nije prodan nijedan crveni motor.

Na slici 1.4 vidimo mozaik prikaz boja prodanih motora u promatranom razdoblju.



Slika 1.4: Mozaik prikaz boja prodanih motora po godinama promatranog razdoblja (ispis iz SAS-a)

U Tablici 1.4 prikazane su frekvencije prodanih motora po cijenovnim razredima u promatranom razdoblju.

Tablica 1.4: Tablica frekvencija cijenovnih razreda po godinama promatranog razdoblja (ispis iz SAS-a)

Frequency
Percent
Row Pct
Col Pct

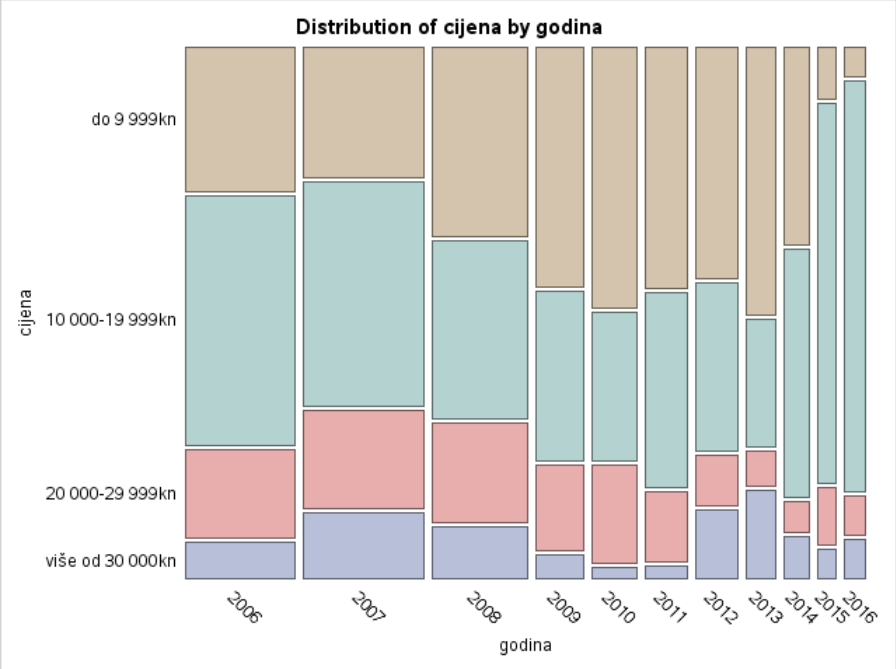
Table of cijena by godina

cijena	godina											
	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	Total
do 10 000kn	149 5.08 14.71 27.70	149 5.08 14.71 25.04	169 5.76 16.68 36.34	110 3.75 10.86 46.22	110 3.75 10.86 50.23	96 3.27 9.48 46.38	93 3.17 9.18 44.50	76 2.59 7.50 51.70	46 1.57 4.54 38.02	9 0.31 0.89 10.11	6 0.20 0.59 5.71	1013 34.54
10 000-19 999kn	259 8.83 21.48 48.14	258 8.80 21.39 43.36	160 5.46 13.27 34.41	78 2.66 6.47 32.77	63 2.15 5.22 28.77	78 2.66 6.47 37.68	68 2.32 5.64 32.54	36 1.23 2.99 24.49	58 1.98 4.81 47.93	65 2.22 5.39 73.03	83 2.83 6.88 79.05	1206 41.12
20 000-29 999kn	92 3.14 20.13 17.10	113 3.85 24.73 18.99	89 3.03 19.47 19.14	39 1.33 8.53 16.39	41 1.40 8.97 18.72	28 0.95 6.13 13.53	20 0.68 4.38 9.57	10 0.34 2.19 6.80	7 0.24 1.53 5.79	10 0.34 2.19 11.24	8 0.27 1.75 7.62	457 15.58
više od 30 000kn	38 1.30 14.79 7.06	75 2.56 29.18 12.61	47 1.60 18.29 10.11	11 0.38 4.28 4.62	5 0.17 1.95 2.28	5 0.17 1.95 2.42	28 0.95 10.89 13.40	25 0.85 9.73 17.01	10 0.34 3.89 8.26	5 0.17 1.95 5.62	8 0.27 3.11 7.62	257 8.76
Total	538 18.34	595 20.29	465 15.85	238 8.11	219 7.47	207 7.06	209 7.13	147 5.01	121 4.13	89 3.03	105 3.58	2933 100.00

Najveći udio prodaje motora otpada na motore cijena 10000,00-19999,00 kn (čine 41,12% ukupne prodaje motora), dok na razred motora cijena manjih od 10000,00 kn otpada nešto više od trećine tržišta. Prodanih motora cijene 20000,00-29999,00 kn je 15,58% na tržištu, dok je najskuplja kategorija najmanje zastupljena.

Promotrimo cijenovni razred 10000,00-19999,00 kn. Gledajući relativne frekvencije po broju prodanih motora u određenoj godini (četvrte znamenke u tablici), možemo vidjeti da su prvih godina promatranog razdoblja prodani motori tih cijena činili 30-50% tržišta, no tada je taj postotak počeo opadati (do posljednje dvije godine) i u prodaji su prevladali motori najjeftinijeg cijenovog razreda. Posljednje dvije godine prodani motori cijenovnog razreda 10000,00-19999,00 kn čine 73,03% odnosno 79,05% tržišta dok je najjeftinijih motora znatno manje. Možemo zaključiti da se tržište prodaje 2015. godine počelo oporavljati od strašne krize koja ga je pogodila nakon 2008. godine.

Na slici 1.5 prikazan je mozaik prikaz cijenovnih razreda po godinama. Može se uočiti kako se od 2009. do 2014. godine povećao udio najjeftinijih motora na tržištu, a od 2014. godine vidimo porast prodaje motora cijenovnog razreda 10000,00-19999,00 kn.



Slika 1.5: Mozaik prikaz cijena prodanih motora po godinama promatranog razdoblja (ispis iz SAS-a)

U Tablici 1.5 možemo vidjeti kojih su približno cijena marke motora na tržištu.

Motori marke Derbi koji se prodaju nisu skuplji od 20000,00 kn, dok marka motora Moto Guzzi pripada najskupljem cijenovnom razredu.
Piaggio motora ima svih cijena, a čine čak 93,58% prodanih motora najnižeg cijenovnog razreda te između 48% i 50% svih ostalih cijenovnih razreda.
Marka Gilera motora je uglavnom srednjih cijenovnih razreda, tj. cijena između 10000,00 kn i 30000,00 kn.
Marke Aprilia motora se prodaju u svim cijenovim razredima te čini 36,58% od ukupnog tržišta prodanih motora najskupljeg cijenovnog razreda.

Na slici 1.6 prikazan je mozaik prikaz marka motora po cijenovnim razredima.

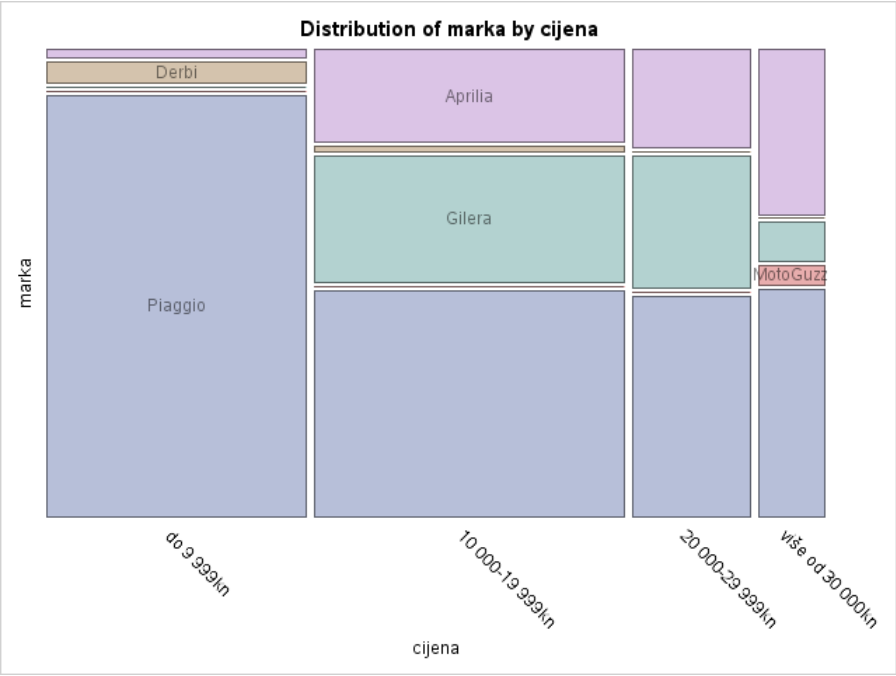
Jasno se vidi da je Piaggio marka motora najzastupljenija u svim cijenovnim razredima. Primjetimo širine pravokutnika cijenovnih razreda. Najširi pravokutnik čini cijenovni razred 10000,00-19999,00 kn, dok najuži pravokutnik pripada najskupljem cijenovnom razredu.

Tablica 1.5: Tablica frekvencija marki motora po cijenovnim razredima (ispis iz SAS-a)

Frequency
Percent
Row Pct
Col Pct

Table of marka by cijena

marka	cijena: do 10 000kn	10 000-19 999kn	20 000-29 999kn	više od 30 000kn	Total
Aprilia	19 0.65 4.12 1.88	248 8.46 53.80 20.56	100 3.41 21.69 21.88	94 3.20 20.39 36.58	461 15.72
Derbi	46 1.57 77.97 4.54	13 0.44 22.03 1.08	0 0.00 0.00 0.00	0 0.00 0.00 0.00	59 2.01
Gilera	0 0.00 0.00 0.00	338 11.52 68.42 28.03	134 4.57 27.13 29.32	22 0.75 4.45 8.56	494 16.84
MotoGuzz	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	11 0.38 100.00 4.28	11 0.38
Piaggio	948 32.32 49.69 93.58	607 20.70 31.81 50.33	223 7.60 11.69 48.80	130 4.43 6.81 50.58	1908 65.05
Total	1013 34.54	1206 41.12	457 15.58	257 8.76	2933 100.00



Slika 1.6: Mozaik prikaz marki prodanih motora po cijenovnim razredima (ispis iz SAS-a)

Provjerimo još kojih su cijenovnih razreda boje prodanih motora na tržištu.

Uočimo da je crna boja motora podjednako zastupljena u svim cijenovnim razredima i na nju otpada između 34% i 39% tržišta. Bijelih prodanih motora ima gotovo dvostruko više u najjeftinijim cijenovnim razredima, dok su crveni motori jako zastupljeni u cijenovnom razredu 20000,00 do 29999,00 kn. Netipičnih boja, odnosno skupina boja ostalo je znatno zastupljenija u cijenovnim razredima višim od 20000,00 kn, dok special boja varira, a najmanje je ima u najjeftinijem cijenovnom razredu.

Tablica 1.6: Tablica frekvencija boja prodanih motora po cijenovnim razredima (ispis iz SAS-a)

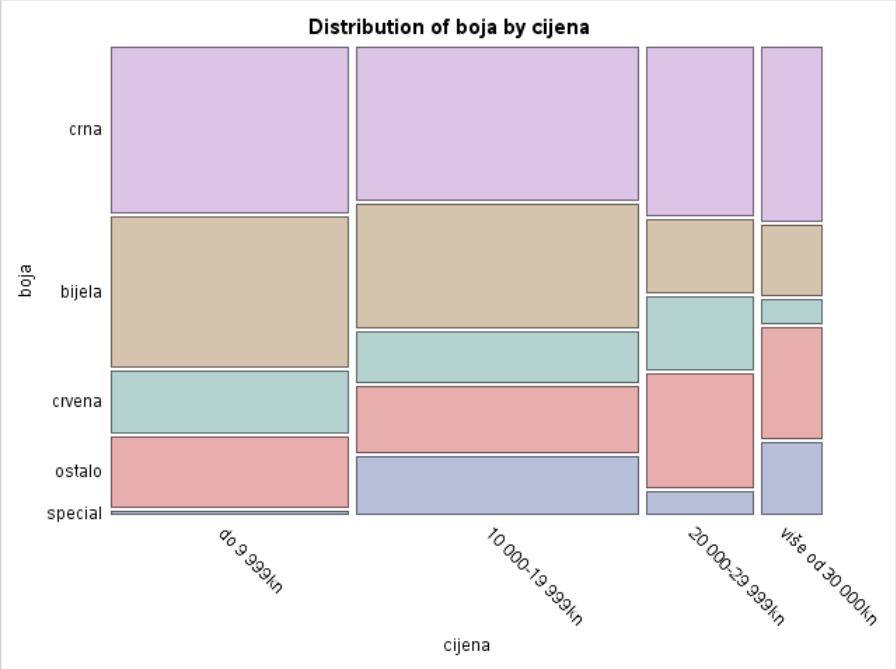
The FREQ Procedure

Table of boja by cijena

Frequency Percent Row Pct Col Pct	cijena				
boja	do 9 999kn	10 000-19 999kn	20 000-29 999kn	više od 30 000kn	Total
crna	371 12.65 35.27 36.62	411 14.01 39.07 34.08	171 5.83 16.25 37.42	99 3.38 9.41 38.52	1052 35.87
bijela	339 11.56 43.41 33.46	328 11.18 42.00 27.20	74 2.52 9.48 16.19	40 1.36 5.12 15.56	781 26.63
crvena	137 4.67 38.06 13.52	135 4.60 37.50 11.19	74 2.52 20.56 16.19	14 0.48 3.89 5.45	360 12.27
ostalo	160 5.46 30.95 15.79	179 6.10 34.62 14.84	115 3.92 22.24 25.16	63 2.15 12.19 24.51	517 17.63
special	6 0.20 2.69 0.59	153 5.22 68.61 12.69	23 0.78 10.31 5.03	41 1.40 18.39 15.95	223 7.60
Total	1013 34.54	1206 41.12	457 15.58	257 8.76	2933 100.00

Na slici 1.7 prikazan je mozaik prikaz cijenovnih razreda po bojama motora.

Jasno se vidi da je crna boja motora dominantna u svim cijenovnim razredima, da je bijela boja zastupljenija kod jeftinijih cijenovnih razreda, te da special boja motora zauzima najmanji dio najjeftinijeg cijenovnog razreda.



Slika 1.7: Mozaik prikaz boja prodanih motora po cijenovnim razredima (ispis iz SAS-a)

Poglavlje 2

χ^2-testovi

2.1 Uvod

χ^2-test jedan je od prvih statističkih testova, razvio ga je Pearson 1900. godine pa se često naziva Pearsonovim χ^2-testom. On je neparametrijski test koji testira razliku između opaženih i očekivanih frekvencija. Pomoću χ^2-testa testiramo nultu hipotezu da neko obilježje ima navedenu razdiobu naspram alternative da nema tu razdiobu te ispitujemo nezavisnost ili homogenost dvije varijable ili faktora. Za sve navedeno, test statistika je

$$H = \sum_{i=1}^{n} \frac{(f_i - f_{t_i})^2}{f_{t_i}} \tag{2.1}$$

gdje su f_i opažene (promatrane) frekvencije, f_{t_i} teorijske (očekivane) frekvencije, a n veličina uzorka. [6]
Uz pretpostavku da je nulta hipoteza točna, za velike n vrijedi

$$H \approx \chi^2(n-1) \tag{2.2}$$

Testiramo nultu hipotezu (H_0) naspram alternativne hipoteze (H_a):

H_0 *: nema statistički značajne razlike između dva uzorka kategorijskih varijabli*

H_a *: razlika postoji*

2.2 Rezultati

Kod u SAS-u

Koristimo PROC FREQ proceduru s opcijom CHISQ za testiranje χ^2-testom. PROC FREQ računa sljedeće χ^2-testove: Pearson χ^2, likelihood ratio χ^2, Mantel-Haenszel χ^2-test.

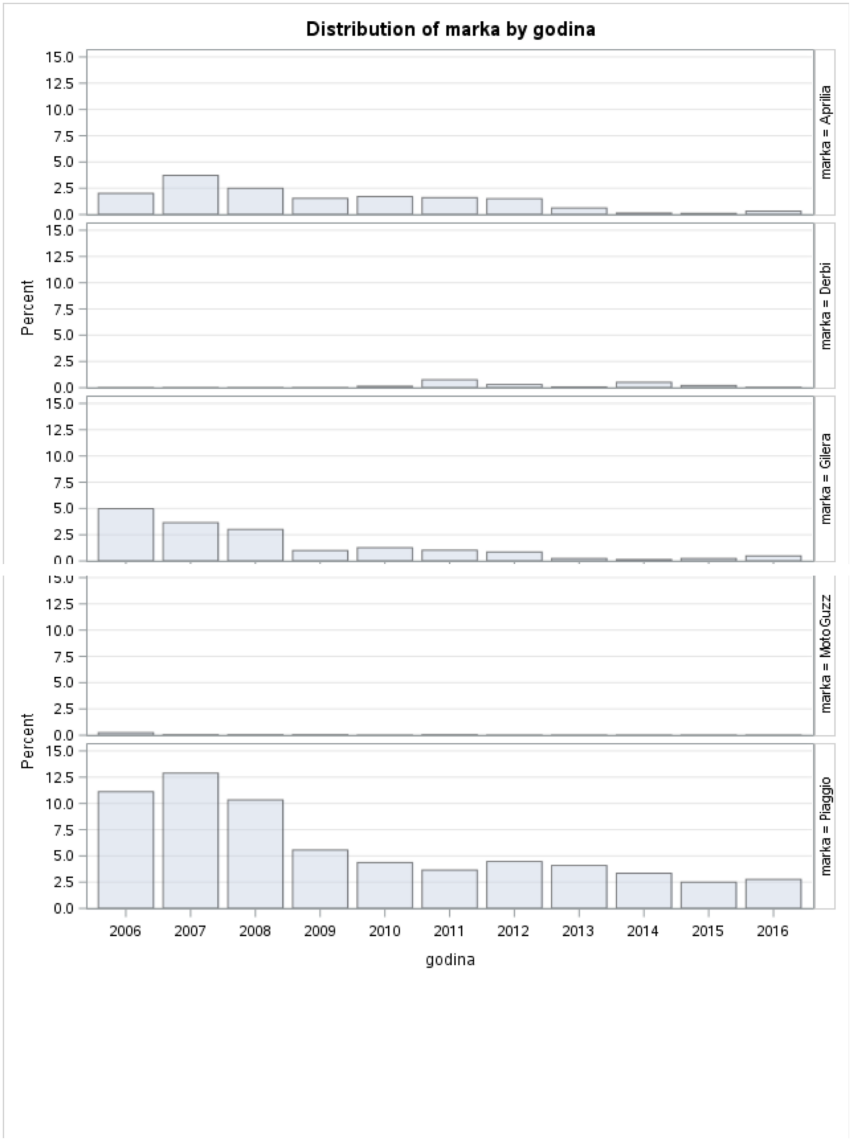
PROC FREQ *data=baza;*
WEIGHT brprodanih;
*TABLES marka*godina / CHISQ measures*
PLOTS=(freqplot(twoway=groupvertical scale=percent));
run;

Naredbom *TABLES marka*godina* u SAS-u ukazujemo da želimo da varijabla marka određuje retke, a varijabla godina stupce kontigencijske tablice. *CHISQ* naredbom se računa χ^2-test. Opcijom *PLOTS=(freqplot(twoway=groupvertical scale=percent))* crtamo stupičaste grafičke prikaze (Slika 2.1) u kojima je visina svakog pravokutnika relativna frekvencija određene marke u promatranoj godini. Relativne frekvencije marki motora po godinama promatranog razdoblja čitamo iz Tablice 1.1.

Sada provjerimo postoji li statistički značajna razlika u prodaji motora po markama obzirom na godine promatranog razdoblja.

H_0 *: ne postoji statistički značajna razlika u prodaji motora između marki motora i godina promatranog razdoblja*
H_a *: razlika postoji*

Prema rezultatima χ^2-testa (Tablica 2.1) vidimo da postoji statistički značajna razlika po markama prodanih motora s obzirom na godine promatranog razdoblja. Naime, χ^2=365,34, broj stupnjeva slobode=40, p<0,0001, stoga odbacujemo nultu hipotezu da ne postoji razlika u prodaji između marki motora po godinama promatranog razdoblja.



Slika 2.1: Stupičasti grafički prikaz marki prodanih motora po godinama promatranog razdoblja (ispis iz SAS-a)

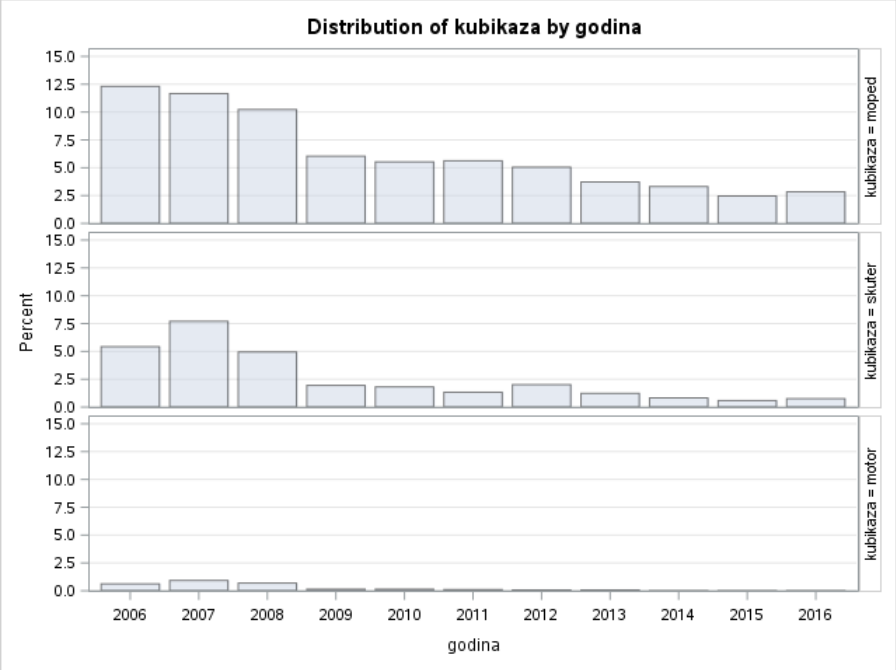
Kada bi vrijedila nulta hipoteza to bi značilo da su relativne frekvencije za svaku marku motora u godini promatranog razdoblja približno jednake, no sa Slike 2.1 vidimo da nisu. Također možemo primjetiti da kod svih marki motora, osim Derbi marke postoji velika razlika u prodaji između prve tri godine promatranog razdoblja i ostalih godina, kada je nastupila kriza.

Tablica 2.1: Rezultati χ^2-testa za nezavisnost marki prodanih motora po godinama promatranog razdoblja (ispis iz SAS-a)

Statistics for Table of marka by godina

Statistic	DF	Value	Prob
Chi-Square	40	365.3431	<.0001
Likelihood Ratio Chi-Square	40	338.3432	<.0001
Mantel-Haenszel Chi-Square	1	9.2674	0.0023
Phi Coefficient		0.3529	
Contingency Coefficient		0.3328	
Cramer's V		0.1765	

H_0 : ne postoji statistički značajna razlika u prodaji motora po jačinama u promatranom razdoblju
H_a : razlika postoji



Slika 2.2: Stupičasti grafički prikaz jačina (kubikaža) prodanih motora po godinama promatranog razdoblja (ispis iz SAS-a)

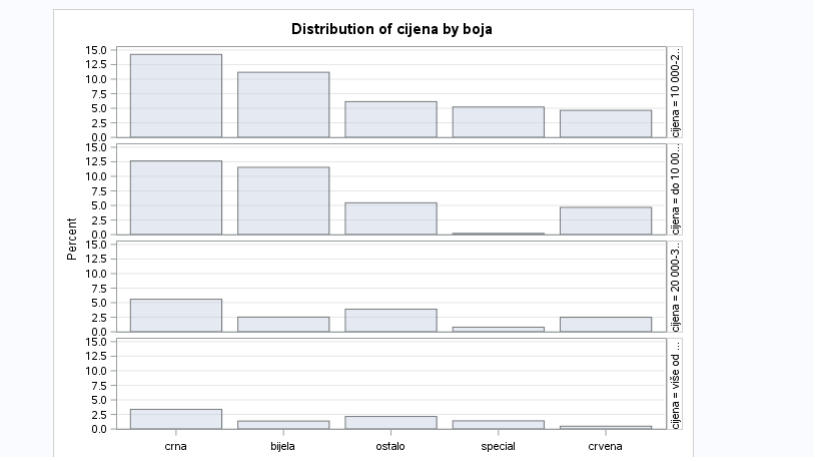
Tablica 2.2: Rezultati χ^2-testa za nezavisnost jačina prodanih motora po godinama promatranog razdoblja (ispis iz SAS-a)

Statistics for Table of kubikaza by godina

Statistic	DF	Value	Prob
Chi-Square	20	90.4662	<.0001
Likelihood Ratio Chi-Square	20	99.1748	<.0001
Mantel-Haenszel Chi-Square	1	33.6818	<.0001
Phi Coefficient		0.1756	
Contingency Coefficient		0.1730	
Cramer's V		0.1242	

Prema rezultatima χ^2-testa (Tablica 2.2) vidimo da postoji statistički značajna razlika u prodaji motora po jačinama s obzirom na godine promatranog razdoblja. χ^2=90,47, broj stupnjeva slobode=20, p<0,0001, pa odbacujemo nultu hipotezu. Iz stupičastog grafičkog prikaza na Slici 2.2 možemo uočiti da postoji razlika koju smo komentirali u Tablici 1.2 (tablici frekvencija broja prodanih motora po jačinama u promatranom razdoblju).

H_0 : ne postoji statistički značajna razlika u prodaji motora po bojama motora između cijenovnih razreda
H_a : razlika postoji



Slika 2.3: Stupičasti grafički prikaz boja prodanih motora po cijenovnim razredima (ispis iz SAS-a)

Tablica 2.3: Rezultati χ^2-testa za nezavisnost boja i cijenovnih kategorija prodanih motora (ispis iz SAS-a)

Statistics for Table of cijena by boja

Statistic	DF	Value	Prob
Chi-Square	12	228.6857	<.0001
Likelihood Ratio Chi-Square	12	265.9403	<.0001
Mantel-Haenszel Chi-Square	1	4.3317	0.0374
Phi Coefficient		0.2792	
Contingency Coefficient		0.2689	
Cramer's V		0.1612	

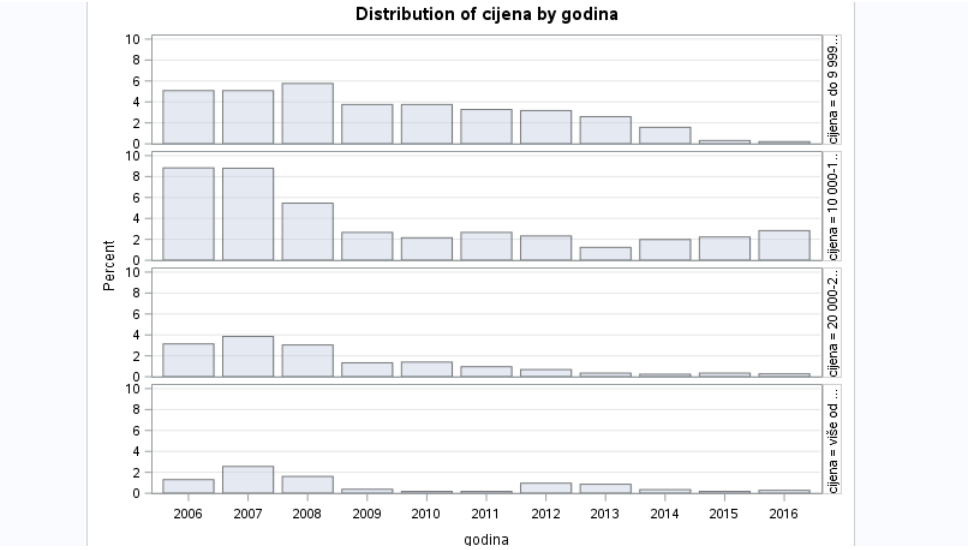
Prema rezultatima χ^2-testa (Tablica 2.3) vidimo da postoji statistički značajna razlika u bojama prodanih motora s obzirom na cijenovne razrede (χ^2=228,69, broj stupnjeva slobode=12, p<0,0001).

Stupčastim grafičkim prikazom na Slici 2.3 imamo jako dobar vizualni prikaz koliko je prodanih motora određene boje u kojem cijenovnom razredu. Kada bi vrijedila nulta hipoteza da ne postoji razlika u prodaji motora između kategorijskih varijabli boja i cijena, to bi značilo da su relativne frekvencije svih boja u svakom od cijenovnih razreda približno jednake, no sa Slike 2.3 je jasno da nisu.

Na sličan način provjerimo postoji li razlika u prodaji motora između cijenovnih razreda obzirom na godine prodaje. Testiramo:

H_0 : ne postoji statistički značajna razlika u prodaji motora po godinama između cijenovnih razreda
H_a : razlika postoji

Prema rezultatima χ^2-testa (Tablica 2.4) vidimo da postoji statistički značajna razlika u cijenovnim razredima prodanih motora s obzirom na godine promatranog razdoblja. Naime, χ^2=306,21, broj stupnjeva slobode=30, p<0,0001, stoga odbacujemo nultu hipotezu.



Slika 2.4: Stupičasti grafički prikaz cijena prodanih motora po godinama promatranog razdoblja (ispis iz SAS-a)

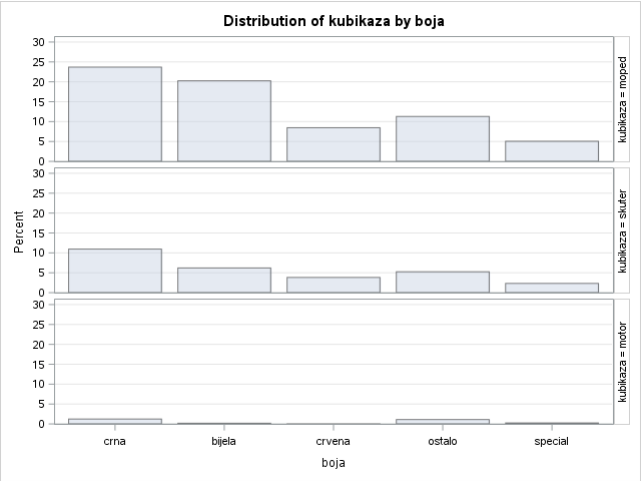
Tablica 2.4: Rezultati χ^2-testa za nezavisnost cijenovnih razreda prodanih motora po godinama promatranog razdoblja (ispis iz SAS-a)

Statistics for Table of cijena by godina

Statistic	DF	Value	Prob
Chi-Square	30	306.2099	<.0001
Likelihood Ratio Chi-Square	30	324.0124	<.0001
Mantel-Haenszel Chi-Square	1	12.6361	0.0004
Phi Coefficient		0.3231	
Contingency Coefficient		0.3075	
Cramer's V		0.1865	

Iz stupičastog grafičkog prikaza na Slici 2.4 možemo odmah uočiti da postoji razlika koju smo komentirali u Tablici 1.4 (tablici frekvencija cijenovnih razreda po godinama promatranog razdoblja). Očit je pad u prodaji motora po svim cijenovnim razredima nakon 2008. godine. Nakon 2013. godine postoji još jedan pad po svim cijenovnim razredima osim razreda cijena 10000,00-19999,00 kn kod kojeg možemo uočiti rast u posljednje tri godine.

H_0 : ne postoji statistički značajna razlika u prodaji motora između kategorijskih varijabli boja motora i njihovih jačina
H_a : razlika postoji



Slika 2.5: Stupičasti grafički prikaz boja prodanih motora po jačinama motora (ispis iz SAS-a)

Tablica 2.5: Rezultati χ^2-testa za nezavisnost boja i jačinskih kategorija prodanih motora (ispis iz SAS-a)

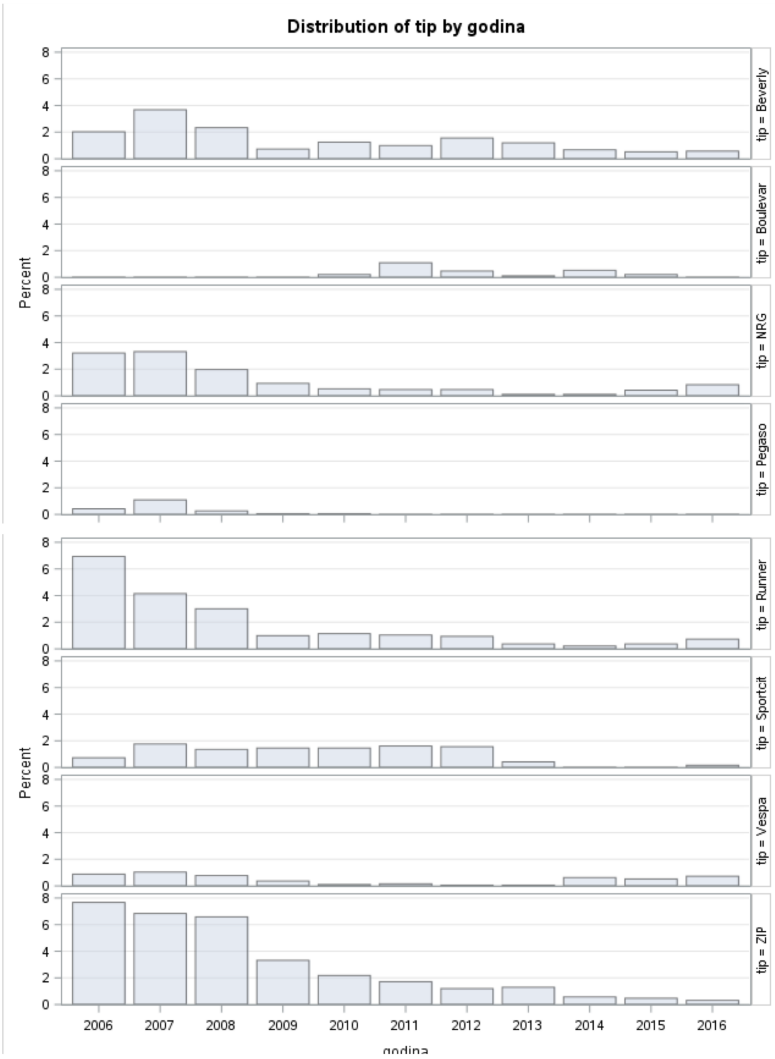
Statistics for Table of kubikaza by boja

Statistic	DF	Value	Prob
Chi-Square	8	66.4000	<.0001
Likelihood Ratio Chi-Square	8	75.9273	<.0001
Mantel-Haenszel Chi-Square	1	3.7388	0.0532
Phi Coefficient		0.1505	
Contingency Coefficient		0.1488	
Cramer's V		0.1064	

Prema rezultatima χ^2-testa (Tablica 2.5) vidimo da postoji statistički značajna razlika po bojama prodanih motora s obzirom na jačinske razrede motora. Naime, χ^2=66,40, broj stupnjeva slobode=8, p<0,0001, stoga odbacujemo nultu hipotezu. Na Slici 2.5 je jasno da postoji razlika.

Za kraj, testirajmo postoji li razlika u prodaji motora po 8 tipova koje smo odabrali obzirom na promatrano razdoblje.

H_0 : ne postoji statistički značajna razlika u prodaji motora po tipu motora u promatranom razdoblju
H_a : razlika postoji



Slika 2.6: Stupičasti grafički prikaz tipova prodanih motora po godinama promatranog razdoblja (ispis iz SAS-a)

Tablica 2.6: Rezultati χ^2-testa za nezavisnost tipova motora po godinama promatranog razdoblja (ispis iz SAS-a)

Statistics for Table of tip by godina

Statistic	DF	Value	Prob
Chi-Square	70	589.0201	<.0001
Likelihood Ratio Chi-Square	70	513.5732	<.0001
Mantel-Haenszel Chi-Square	1	24.3009	<.0001
Phi Coefficient		0.5524	
Contingency Coefficient		0.4836	
Cramer's V		0.2088	

Prema rezultatima χ^2-testa (Tablica 2.6) vidimo da postoji statistički značajna razlika u prodaji navedenih 8 tipova motora s obzirom na godine promatranog razdoblja. Naime, χ^2=589,02, broj stupnjeva slobode=70, p<0,0001, stoga odbacujemo nultu hipotezu da ne postoji razlika u prodaji određenih 8 tipova motora po godinama promatranog razdoblja.

Na Slici 2.6 vidimo distribucije prodaje tipova motora za analizirani period.

Beverly i NRG su se 2006. i 2007. relativno dobro prodavali, dok od 2008. dolazi do pada prodaje, pogotovo za tip motora NRG. Boulevard tip motora, tip marke Derbi, se počeo prodavati za vrijeme krize i tada je bio popularan u prodaji zbog niske cijene. 2016. godine se prestao prodavati. Pegaso je tip motora koji spada u skupinu najjačih motora i 2007. godine je prodaja Pegasa bila najveća.
Tip motora Runner je jako popularan među mladima, jačinskih razreda do 500 kubika. Prijašnjih godina se jako dobro prodavao, no možemo vidjeti kako mu prodaja dosta opada kroz godine. Sportcity se "dobro" prodavao tijekom najgorih godina prodaje, no krajem promatranog razdoblja njegova prodaja opada.
Proučimo prodaju motora poznatog u cijelom svijetu, bezvremenske Vespe. Godinama predstavlja talijansku eleganciju, osvaja populaciju svakim svojim novim dizajnom i godinama ostaje omiljeno prijevozno sredstvo svake nove generacije. Nastala je u tvornici Rinalda Piaggia u Genovi nakon prvog svjetskog rata. Vespa je doživjela pad u prodaji nakon 2008. godine, ali vidimo da se zadnje tri godine povećala prodaja Vespa motora. Tip motora ZIP je također jako popularan, jačina od 50 do 150 kubika i ne prevelikih cijena. Vidimo da je njegov pad skoro linearan.

Poglavlje 3

Analiza varijance

3.1 Uvod

Analiza varijance (ANOVA) jedna je od najkorištenijih statističkih metoda koja ima dugu povijest. Razvio ju je engleski statističar R.A. Fisher (1890.-1962.) i koristio kao praktičnu tehniku za istraživanje nekih bioloških fenomena. ANOVA je specijalan slučaj linearne regresije, koja je pak specijalan slučaj generaliziranih linearnih modela kojima je zajedničko da minimiziraju grešku modela.
ANOVA je statistička metoda kojom se testiraju jednakosti više očekivanja i donosi se zaključak o postojanju (ili ne) razlika između očekivanja više populacija. Analizira se utjecaj jedne ili više kategorijskih (nezavisnih) varijabli na jednu numeričku kontinuiranu (zavisnu) varijablu. Kategorijske varijable se nazivaju faktorima pa govorimo o jednofaktorskoj, dvofaktorskoj ili višefaktorskoj analizi varijance. ANOVA se bavi analizom varijabilnosti po čemu je i dobila ime. Problem razlike više populacija svodimo na analizu varijabilnosti unutar svakog uzorka (ona varijabilnost koju ne možemo objasniti) i varijabilnosti između uzoraka (ona varijabilnost koju možemo objasniti i točno znamo od kuda dolazi). Dakle, preko varijabilnosti, tj. prosječnog odstupanja od aritmetičke sredine uspoređujemo same aritmetičke sredine. Ako nam je varijabilnost između uzoraka veća od varijabilnosti unutar uzoraka možemo pretpostaviti da se radi o različitim populacijama.

Pretpostavke koje moraju biti zadovoljene da bi mogli koristiti ANOVA-u su:

1) uzorci moraju biti nezavisni
2) homogenost varijance, tj. populacije moraju imati približno jednake varijance
3) uzorci su iz normalno distribuirane populacije

Za tumačenje značajnosti razlika važno je napomenuti *post hoc testiranje*. Naime, ako dobijemo F-omjer koji je statistički značajan, to znači da se promatrane skupine statistički značajno razlikuju u istraživanoj varijabli. No, još ne možemo tvrditi između kojih parova je razlika statistički značajna. Da bi provjerili između kojih parova postoji razlika, moramo primijeniti jedan od naknadih testova (post hoc) koji slijede nakon analize varijance. Postoje različiti post hoc testovi (Scheffe, Tukey, Duncan, Newman-Keuls...), ali mi ćemo koristiti Tukey-ev test. [5]

Jednofaktorska analiza varijance

Promatra se utjecaj jednog faktora koji ima m razina.
Neka su:

$X_{11}, X_{12}, ..., X_{1n_1}$ slučajni uzorak za $X_1 \sim N(\mu_1, \sigma^2)$
$X_{21}, X_{22}, ..., X_{2n_2}$ slučajni uzorak za $X_2 \sim N(\mu_2, \sigma^2)$
..
$X_{m1}, X_{m2}, ..., X_{mn_m}$ slučajni uzorak za $X_m \sim N(\mu_m, \sigma^2)$,

m nezavisnih slučajnih uzoraka, svaki za obilježje X reprezentirano sa X_i u populaciji i, $i = 1, 2, .., m$.

Testiramo:

$H_0 : \mu_1 = \mu_2 = ... = \mu_m$

H_a : barem se dva uzorka od njih m statistički značajno razlikuju

Stavimo:

$$\begin{aligned} n &= \sum_{i=1}^{m} n_i \\ \mu &= \frac{1}{n} \sum_{i=1}^{m} n_i \mu_i \\ \delta_i &= \mu_i - \mu \qquad i = 1, .., m \end{aligned}$$

δ_i je efekt i-te razine djelujućeg faktora, μ je opća srednja vrijednost.

Sada se nulta hipoteza svodi na:

$H_0 : \delta_1 = \delta_2 = ... = \delta_m = 0$, tj. efekti djelujućeg faktora su beznačajni. Dalje, za svaki uzorak posebno izračunamo aritmetičke sredine $\overline{X_i}$ i uzoračke varijance S_i^2. Računamo aritmetičku sredinu svih podataka i sumu kvadrata zbog tretmana:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n_i} X_{ij} = \frac{1}{n} \sum_{i=1}^{m} n_i \overline{X_i}$$

$$SST = \sum_{i=1}^{m} n_i (\overline{X_i} - \overline{X})^2$$

Dalje računamo sumu kvadrata pogrešaka:

$$SSE = \sum_{i=1}^{m} (n_i - 1) S_i^2$$

Ukupna pogreška:

$$SS = SST + SSE$$

Srednjekvadratno odstupanje tretmana:

$$MST = \frac{SST}{m-1}$$

Srednje kvadratna pogreška:

$$MSE = \frac{SSE}{n-m}$$

Testna statistika glasi:

$$F = \frac{MST}{MSE} \quad \backsim \quad F(m-1, n-m)$$

(uz H_0)
Kritično područje je oblika : $[f_\alpha(m-1, n-m), \infty\rangle$
P-vrijednost: $p = P(F > f || H_0)$
[7], [5]

Tablica 3.1: Tablica rezultata za jednofaktorsku analizu varijance

Izvor varijabilnosti	Broj stupnjeva slobode	Sume kvadrata	Varijanca	Test statistika F	p-vrijednost
Između grupa	m-1	SST	MST	F	p
Unutar grupa	n-m	SSE	MSE		
Ukupno	n-1	SS			

Procedura PROC GLM

Modeliranje veze između zavisne varijable i jedne ili više nezavisnih varijabli u SAS-u možemo provesti koristeći tri procedure: PROC REG, PROC ANOVA i PROC GLM. Želi se provjeriti koje od nezavisnih varijabli (jačina motora, marka motora, boja motora, godina u kojoj je prodan, tip motora) utječu na zavisnu varijablu *prodaju* (naziv u SAS-kodu *brprodanih*).
PROC REG je procedura opće linearne regresije, odgovara metodi najmanjih kvadrata i računa p-vrijednost pomoću t-testa. PROC ANOVA odgovara analizi varijance, višestrukoj analizi varijance i analizi varijance ponovljenih mjerenja. Daje istu p-vrijednost kao PROC REG, samo u formi F-testa. Međutim, rad u PROC ANOVA-i može biti problematičan ako nemamo balansiran dizajn.
Procedura PROC GLM, generaliziranih linearnih modela, je mješavina regresije i analize varijance, može rukovati sa svim tipovima varijabli i može dati četiri tipa sume kvadrata, no bazično, PROC GLM daje dva tipa sume kvadrata (*engl. Type I, Type III*) za razliku od druge dvije procedure. Tip sume kvadrata I je hijerarhijski (sekvencijalni) tip. Važan je redoslijed varijabli (efekata) u modelu. Svaki efekt uzima u obzir efekt koji se pojavio prije u modelu, ali ne one koji se pojavljuju kasnije. Tip I SS se računa tako da se suma kvadrata pogrešaka (SSE) stalno smanjuje kada se novi efekt uvrsti u model. Tip sume kvadrata III se ”ponaša” kao da svaka varijabla ulazi zadnja u model. U jednofaktorskim analizama varijance tip sume kvadrata I jednak je tipu sume kvadrata III jer samo jedna varijabla ulazi u model. Kako su naše zavisne varijable koje želimo ispitivati različitih tipova i nemamo balansiran dizajn, radit ćemo u PROC GLM proceduri. [8].

3.2 Rezultati jednofaktorske ANOVA-e

Ispitajmo kako svaka od varijabli baze (marka, jačina, boja, cijena, godina prodaje, tip motora) utječe na prodaju motora.

SAS kod za jednofaktorsku ANOVA-u:

PROC GLM *data=baza;*
CLASS varijabla;
MODEL brprodanih=varijabla;
LSMEANS varijabla/pdiff adjust=tukey;
run;

Prvom naredbom određujemo koju proceduru koristimo (PROC GLM), i na kojem skupu podataka radimo (*data=baza*). Naredbom CLASS definiramo kategorijsku varijablu, a zatim naredbom *MODEL brprodanih = varijabla* želimo provjeriti utječe li ta varijabla na zavisnu varijablu prodaje.
Ukoliko varijabla statistički značajno utječe na prodaju motora, posljednjom naredbom *LSMEANS varijabla/pdiff adjust=tukey;* ispitujemo između kojih parova postoji razlika. LSMEANS računa *least-square means* fiksnih efekata, to je procjena marginalne aritmetičke sredine populacije (očekivanja) za balansirane populacije. Za sve analize nivo značajnosti od 5% (greška tipa I) smatran je statistički značajnim. [5].

Analiza prodaje motora prema marki motora

$H_0 : \mu_{Piaggio} = \mu_{Aprilia} = \mu_{Gilera} = \mu_{Derbi} = \mu_{MotoGuzzi}$
H_a : barem dva očekivanja se statistički značajno razlikuju

Prva tablica u ispisu rezultata ANOVA-e (Tablica 3.2) testira sveukupnu značajnost modela. Iz nje vidimo da je varijabilnost između prodaje marki motora 5 puta veća nego varijabilnost prodaje unutar marki, te zaključujemo da postoji statistički značajna razlika između prosječnog broja prodanih motora prema analiziranim markama (F=5, p=0,0006).
Druga tablica u ispisu rezultata ANOVA-e (Tablica 3.2) nam daje deskriptivne kvantifikatore: koeficijent determinacije (*engl. R-Square*), koeficijent varijacije (*engl.*

Coefficient of Variation), i aritmetičku sredinu svih izmjerenih vrijednosti.

Posljednje dvije tablice u ispisu ANOVA-e (Tablica 3.2) su ANOVA tablice koje se razlikuju u tipovima suma kvadrata, no kako je model jednofaktorski, one su jednake ne samo međusobno nego i sa prvom tablicom u ispisu rezultata.

Tablica 3.2: Rezultati ANOVA-e broja prodanih motora prema markama motora (ispis iz SAS-a)

The GLM Procedure
Dependent Variable: br_prodanih

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	838.53882	209.63470	5.00	0.0006
Error	575	24094.58360	41.90362		
Corrected Total	579	24933.12241			

R-Square	Coeff Var	Root MSE	br_prodanih Mean
0.033632	128.0094	6.473301	5.056897

Source	DF	Type I SS	Mean Square	F Value	Pr > F
marka	4	838.5388151	209.6347038	5.00	0.0006

Source	DF	Type III SS	Mean Square	F Value	Pr > F
marka	4	838.5388151	209.6347038	5.00	0.0006

Kako još ne znamo koje marke motora čine razliku, provedimo Tukey-ev post hoc test. Iz rezultata Tukey-evog post hoc testiranja (Tablica 3.3) vidimo da se marke motora Piaggio i Aprilia statistički značajno razlikuju (p=0,0006), a između ostalih marki nema statistički značajne razlike (sve ostale p-vrijednosti su veće od razine značajnosti).

Tablica 3.3: Rezultati Tukey-evog post hoc testa za prodaju motora prema markama motora (ispis iz SAS-a)

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

marka	br_prodanih LSMEAN	LSMEAN Number
Aprilia	3.26950355	1
Derbi	3.27777778	2
Gilera	5.42857143	3
MotoGuzz	1.57142857	4
Piaggio	5.90712074	5

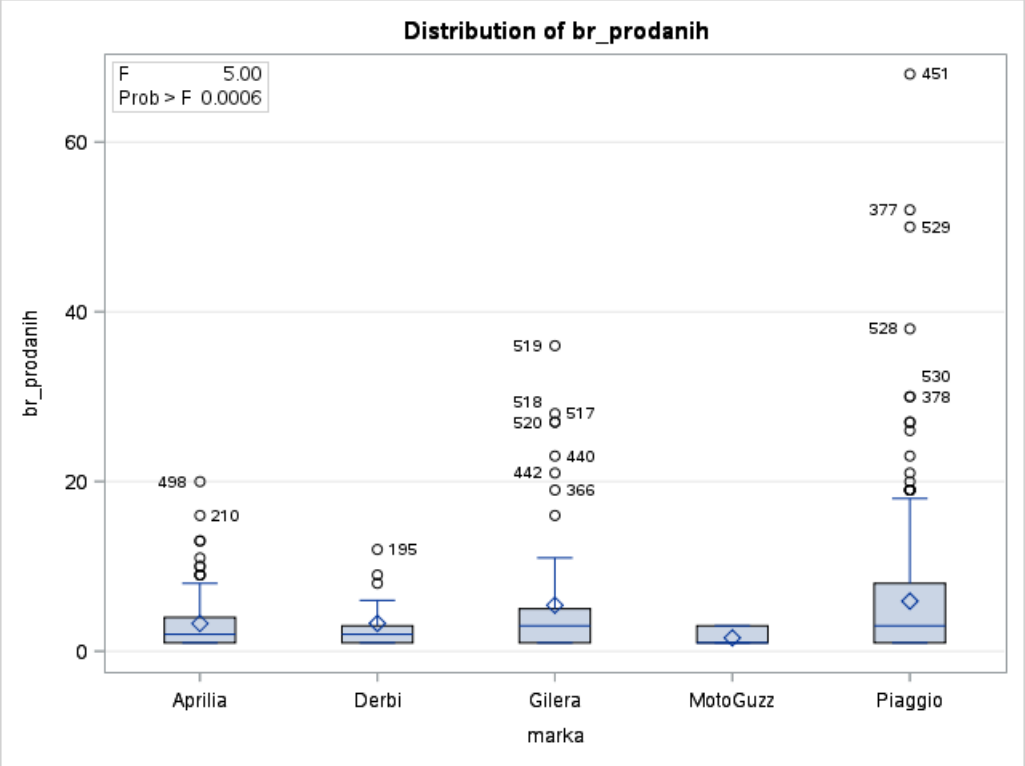
Least Squares Means for effect marka
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: br_prodanih

i/j	1	2	3	4	5
1		1.0000	0.0965	0.9613	0.0006
2	1.0000		0.6988	0.9763	0.4489
3	0.0965	0.6988		0.5504	0.9714
4	0.9613	0.9763	0.5504		0.4023
5	0.0006	0.4489	0.9714	0.4023	

Jednofaktorska analiza varijance standardno crta Box-plot pravokutnike.

Slika 3.1 prikazuje Box-plot grafički prikaz koji se sastoji od pravokutnika koji prikazuje podatke od donjeg do gornjeg kvartila za svaku od marki motora. Horizontalna crta po pravokutniku označava median (vrijednost središnjeg podatka koja podatke poredane po veličini dijeli u dva jednakobrojna dijela), a simbol unutar pravokutnika označava očekivanu vrijednost te marke (*engl. mean*).
Sve točke izvan pravokutnika se crtaju posebno i smatraju se netipičnim vrijednostim (*engl. outlierima*). Netipične vrijednosti su označene brojem observacije, pa primjerice broj 210 iznad pravokutnika kod marke motora Aprilia označava observaciju baze koja se sastoji od 16 prodanih motora tipa Sportcity (marke Aprilia) 2011. godine.
Izgled Box-plota ukazuje na stupanj raspršenosti i asimetričnosti (*engl. skewness*), te nam ukazuje na netipične vrijednosti među podacima. Uočimo kako marka motora Moto Guzzi nema netipičnih vrijednosti, dok ih marka motora Piaggio ima mnogo i poprimaju najviše vrijednosti od svih marki motora.



Slika 3.1: Box-plot grafički prikaz broja prodanih motora po markama motora (ispis iz SAS-a)

Analiza prodaje motora prema jačini motora

$H_0 : \mu_{mopedi} = \mu_{skuteri} = \mu_{motori}$
H_a : barem dva očekivanja se statistički značajno razlikuju

Prva tablica u ispisu rezultata ANOVA-e (Tablica 3.4) testira sveukupnu značajnost modela. Iz nje vidimo da je varijabilnost između prodaje jačinskih razreda motora 31,47 puta veća nego varijabilnost prodaje unutar jačinskih razreda, te zaključujemo da postoji statistički značajna razlika između prosječnog broja prodanih motora prema analiziranim markama (F=31,47, p<0,0001).
Druga tablica u ispisu rezultata ANOVA-e (Tablica 3.4) daje deskriptivne kvantifikatore.
Treća i četvrta tablica (Tablice 3.4) su ANOVA tablice koje se razlikuju u tipovima suma kvadrata koje su i ovdje identične jer je model jednofaktorski.

Tablica 3.4: Rezultati ANOVA-e broja prodanih motora prema jačinskim razredima motora (ispis iz SAS-a)

The GLM Procedure
Dependent Variable: br_prodanih

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2452.18194	1226.09097	31.47	<.0001
Error	577	22480.94047	38.96177		
Corrected Total	579	24933.12241			

R-Square	Coeff Var	Root MSE	br_prodanih Mean
0.098350	123.4341	6.241936	5.056897

Source	DF	Type I SS	Mean Square	F Value	Pr > F
kubikaza	2	2452.181944	1226.090972	31.47	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
kubikaza	2	2452.181944	1226.090972	31.47	<.0001

Tablica 3.5: Rezultati Tukey-evog post hoc testa za prodaju motora prema jačinskim razredima motora (ispis iz SAS-a)

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

kubikaza	br_prodanih LSMEAN	LSMEAN Number
moped	7.17437722	1
motor	2.58064516	2
skuter	3.12313433	3

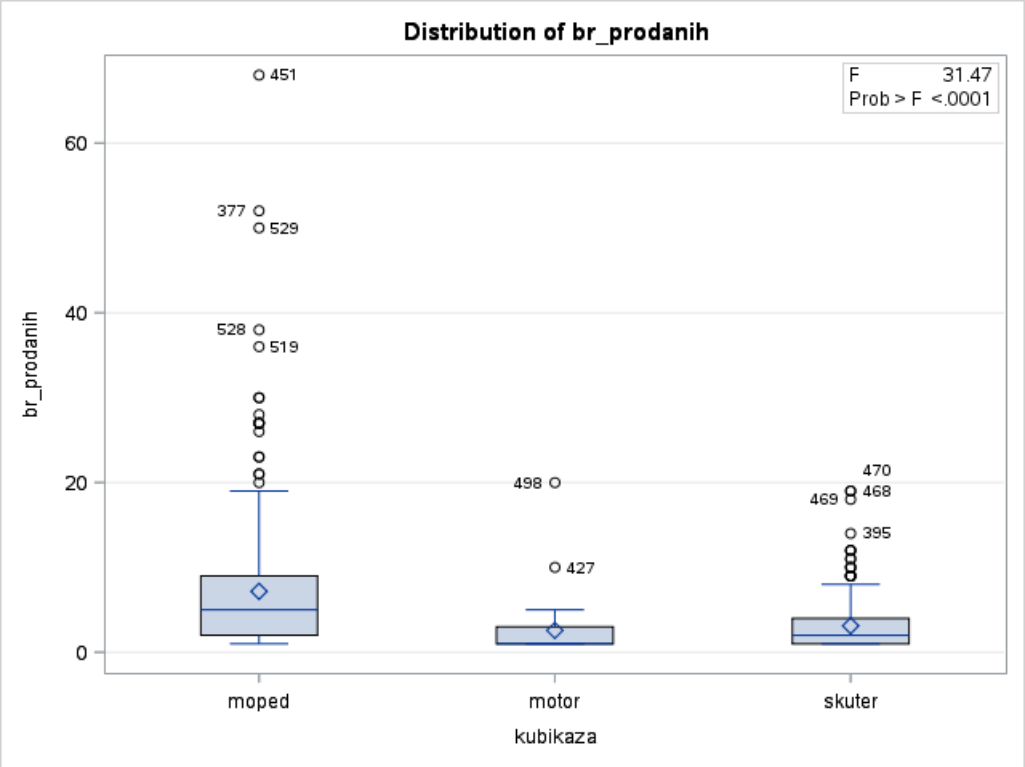
Least Squares Means for effect kubikaza
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: br_prodanih

i/j	1	2	3
1		0.0003	<.0001
2	0.0003		0.8908
3	<.0001	0.8908	

Iz Tablice 3.5, rezultata Tukey-evog post hoc testiranja vidimo da postoji statistički značajna razlika između sredina jačinskih kategorija mopeda i motora (p=0,0003), te između mopeda i skutera (p<0,0001). Između jačinskih kategorija skutera i motora ne postoji statistički značajna razlika.

Slika 3.2 prikazuje Box-plot koji se sastoji od pravokutnika koji prikazuje podatke od donjeg do gornjeg kvartila za svaku od jačinskih kategorija motora.



Slika 3.2: Box-plot grafički prikaz broja prodanih motora po jačinama motora (ispis iz SAS-a)

Analiza prodaje motora prema boji motora

$H_0 : \mu_{crna} = \mu_{bijela} = \mu_{crvena} = \mu_{ostalo} = \mu_{special}$
H_a : barem dva očekivanja se statistički značajno razlikuju

Tablica 3.6: Rezultati ANOVA-e broja prodanih motora prema boji motora (ispis iz SAS-a)

The GLM Procedure
Dependent Variable: br_prodanih

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	108.00850	27.00212	0.63	0.6445
Error	575	24825.11392	43.17411		
Corrected Total	579	24933.12241			

R-Square	Coeff Var	Root MSE	br_prodanih Mean
0.004332	129.9354	6.570701	5.056897

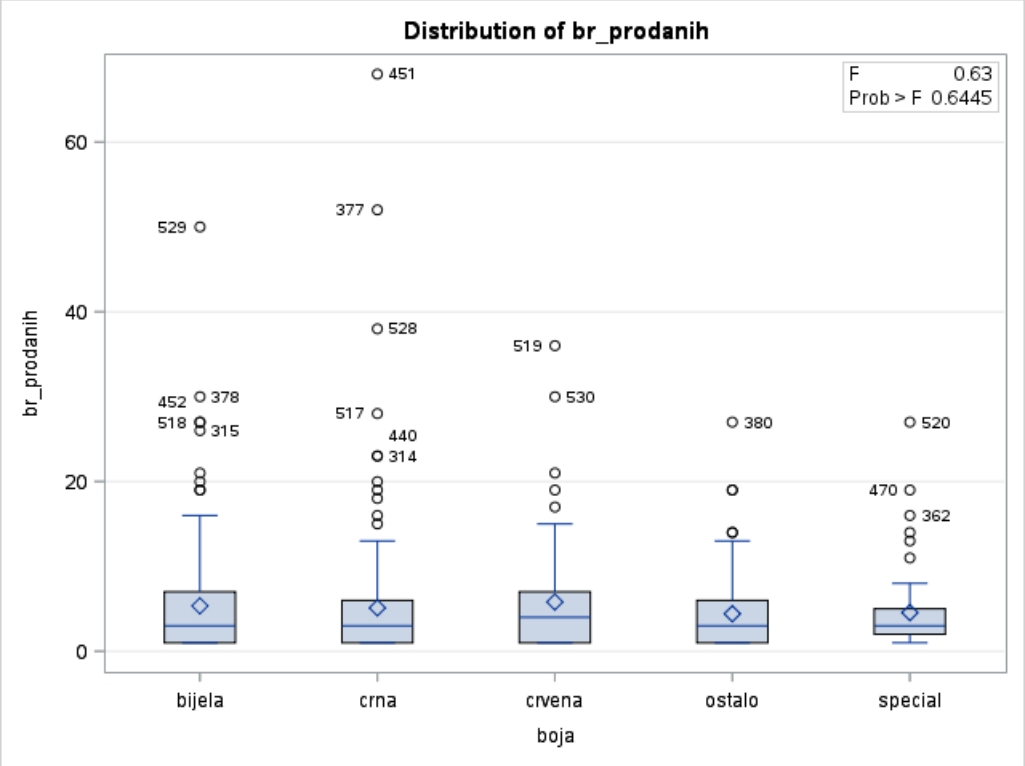
Source	DF	Type I SS	Mean Square	F Value	Pr > F
boja	4	108.0084960	27.0021240	0.63	0.6445

Source	DF	Type III SS	Mean Square	F Value	Pr > F
boja	4	108.0084960	27.0021240	0.63	0.6445

Prema rezultatima ANOVA-e (Tablica 3.6) zaključujemo da ne postoji statistički značajna razlika u broju prodanih motora s obzirom na boju motora. (F=0,63, p=0,64)

Na Slici 3.3 vidimo Box-plot grafički prikaz broja prodanih motora po bojama motora.

Vidjeli smo kako su na Box-plot grafičkim prikazima prijašnja dva modela kod kojih smo odbacili nultu hipotezu o jednakosti sredina pravokutnici bili različitih veličina, na ovom Box-plot grafičkom prikazu možemo uočiti da su pravokutnici sličnih veličina što se poklapa s zaključkom o prihvaćanju nulte hipoteze da boja ne utječe na prodaju motora. Uočimo da najviše netipičnih podataka (*engl. outliera*) ima kod crne i bijele boje motora, dok im skupina boja koju smo nazvali ostalo nije podložna.



Slika 3.3: Box-plot grafički prikaz broja prodanih motora po bojama motora (ispis iz SAS-a)

Analiza prodaje motora prema cijeni motora

Označimo s:

μ_1 očekivani broj prodanih motora cijenovnog razreda $\leqslant$9999,00 kn,
μ_2 očekivani broj prodanih motora cijenovnog razreda 10000,00-19999,00 kn,
μ_3 očekivani broj prodanih motora cijenovnog razreda 20000,00-29999,00 kn te
μ_4 očekivani broj prodanih motora cijena $\geqslant$30000,00 kn.

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$
H_a : barem dva očekivanja se statistički značajno razlikuju

Iz rezultata ANOVA-e (Tablica 3.7) vidimo da postoji statistički značajna razlika u broju prodanih motora s obzirom na cijenovne razrede (F=30,29, p<0,0001).

Tablica 3.7: Rezultati ANOVA-e broja prodanih motora prema cijenovnim razredima (ispis iz SAS-a)

The GLM Procedure
Dependent Variable: br_prodanih

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	3397.09904	1132.36635	30.29	<.0001
Error	576	21536.02337	37.38893		
Corrected Total	579	24933.12241			

R-Square	Coeff Var	Root MSE	br_prodanih Mean
0.136248	120.9170	6.114649	5.056897

Source	DF	Type I SS	Mean Square	F Value	Pr > F
cijena	3	3397.099041	1132.366347	30.29	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
cijena	3	3397.099041	1132.366347	30.29	<.0001

Tablica 3.8: Rezultati Tukey-evog post hoc testa za broj prodanih motora prema cijenovnim razredima (ispis iz SAS-a)

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

cijena	br_prodanih LSMEAN	LSMEAN Number
10 000-19 999kn	4.6926070	1
20 000-29 999kn	3.4621212	2
do 9 999kn	10.1300000	3
više od 30 000kn	2.8241758	4

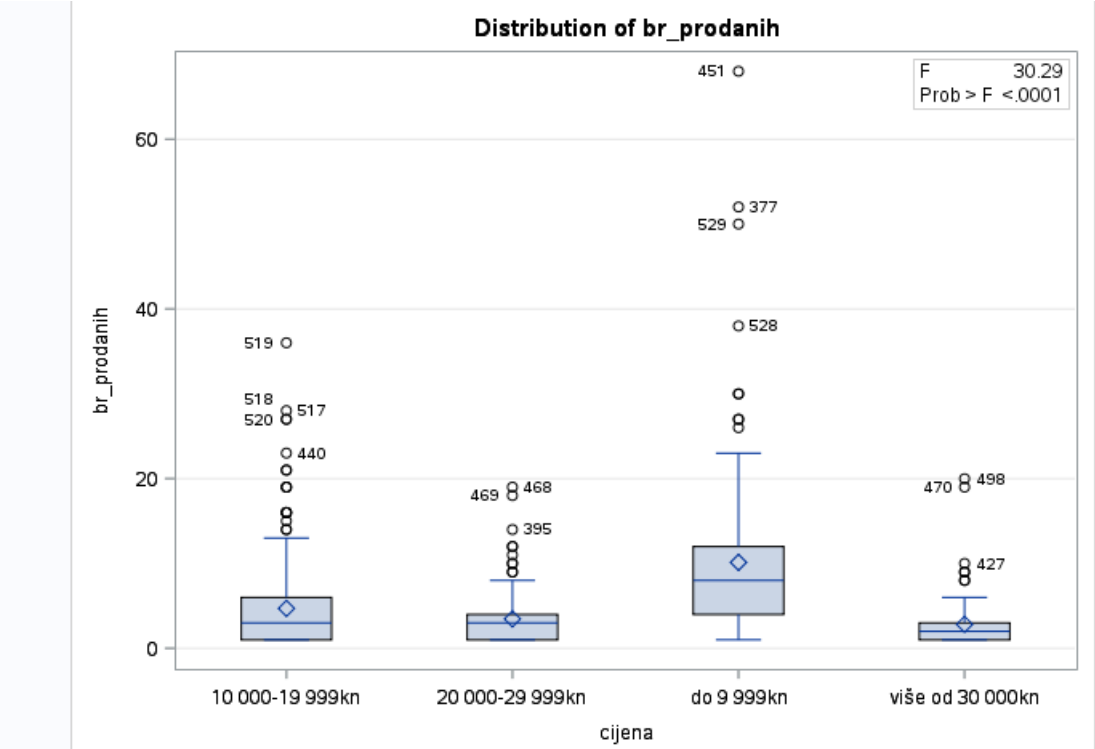
Least Squares Means for effect cijena
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: br_prodanih

i/j	1	2	3	4
1		0.2381	<.0001	0.0602
2	0.2381		<.0001	0.8699
3	<.0001	<.0001		<.0001
4	0.0602	0.8699	<.0001	

Iz Tablice 3.8, rezultata Tukey-evog post hoc testiranja vidimo da statistički značajnu razliku čini cijenovni razred ⩽9999,00 kn s razredima: 10000,00-19999,00 kn, 20000,00-29999,00 kn i razredom ⩾30000,00 kn. Za sve navedene razlike p<0,0001.

Na Slici 3.4 dan je Box-plot grafički prikaz broja prodanih motora po cijenovnim razredima.



Slika 3.4: Box-plot grafički prikaz broja prodanih motora po cijenovnim razredima motora (ispis iz SAS-a)

Analiza prodaje motora prema godini promatranog razdoblja

$H_0 : \mu_{2006} = \mu_{2007} = \mu_{2008} = ... = \mu_{2016}$
H_a : barem dva očekivanja se statistički značajno razlikuju

Vidimo (Tablica 3.9) da postoji statistički značajna razlika u prodaji motora po godinama.
Varijabilnost između prodaje motora po godinama promatranog razdoblja je 3,86 puta veća nego varijabilnost prodaje unutar godina promatranog razdoblja, te zaključujemo da postoji statistički značajna razlika između prosječnog broja prodanih motora prema godinama promatranog razdoblja (F=3,86, p<0,0001).

Tablica 3.9: Rezultati ANOVA-e broja prodanih motora prema godinama promatranog razdoblja (ispis iz SAS-a)

The GLM Procedure
Dependent Variable: br_prodanih

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	1582.65873	158.26587	3.86	<.0001
Error	569	23350.46368	41.03772		
Corrected Total	579	24933.12241			

R-Square	Coeff Var	Root MSE	br_prodanih Mean
0.063476	126.6799	6.406069	5.056897

Source	DF	Type I SS	Mean Square	F Value	Pr > F
godina	10	1582.658731	158.265873	3.86	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
godina	10	1582.658731	158.265873	3.86	<.0001

Iz Tablice 3.10, rezultata Tukey-evog post hoc testiranja vidimo da 2006. godina čini statistički značajnu razliku s godinama 2014.,2015., te 2007. godina čini statistički značajnu razliku s godinama 2009.,2012.,2014.,2015.

Tablica 3.10: Rezultati Tukey-evog post hoc testa broja prodanih motora prema godinama promatranog razdoblja (ispis iz SAS-a)

godina	br_prodanih LSMEAN	LSMEAN Number
2006	7.27027027	1
2007	7.53164557	2
2008	5.96153846	3
2009	3.83870968	4
2010	4.29411765	5
2011	4.92857143	6
2012	3.66666667	7
2013	4.32352941	8
2014	3.02500000	9
2015	2.69696970	10
2016	3.50000000	11

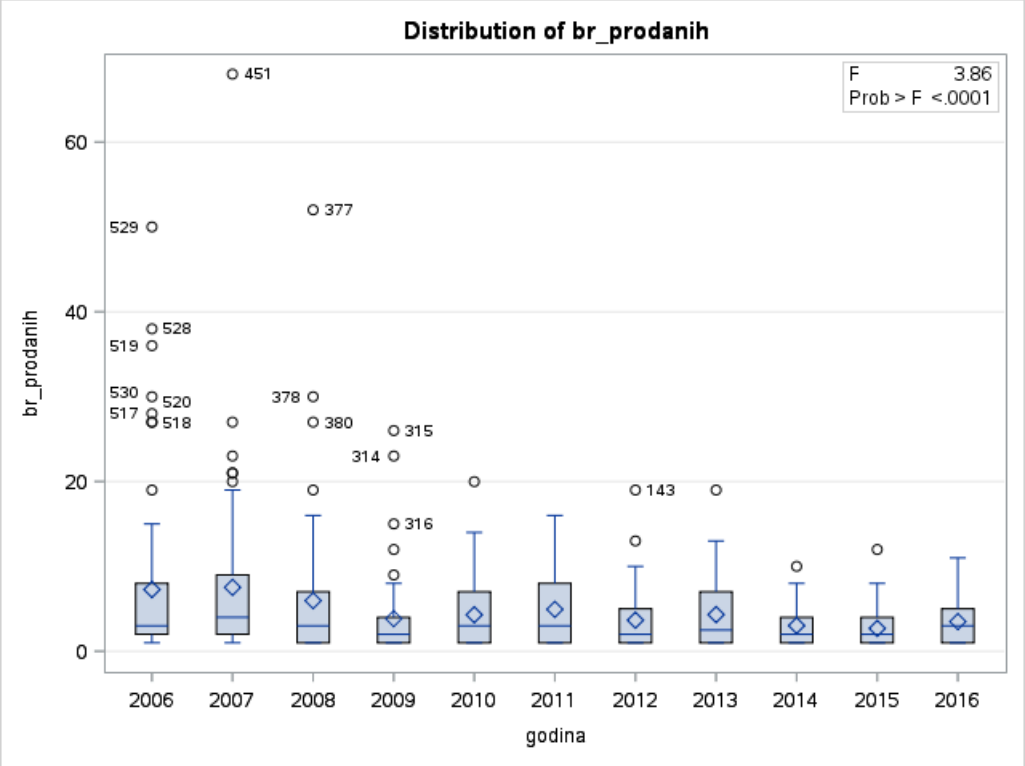
Least Squares Means for effect godina
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: br_prodanih

i/j	1	2	3	4	5	6	7	8	9	10	11
1		1.0000	0.9750	0.0711	0.2765	0.7224	0.0563	0.4913	0.0319	0.0286	0.1938
2	1.0000		0.9073	0.0299	0.1555	0.5577	0.0234	0.3420	0.0139	0.0131	0.1149
3	0.9750	0.9073		0.6853	0.9363	0.9990	0.6095	0.9771	0.3965	0.3343	0.7867
4	0.0711	0.0299	0.6853		1.0000	0.9989	1.0000	1.0000	0.9999	0.9991	1.0000
5	0.2765	0.1555	0.9363	1.0000		1.0000	1.0000	1.0000	0.9975	0.9898	1.0000
6	0.7224	0.5577	0.9990	0.9989	1.0000		0.9967	1.0000	0.9604	0.9204	0.9976
7	0.0563	0.0234	0.6095	1.0000	1.0000	0.9967		1.0000	1.0000	0.9998	1.0000
8	0.4913	0.3420	0.9771	1.0000	1.0000	1.0000	1.0000		0.9987	0.9942	1.0000
9	0.0319	0.0139	0.3965	0.9999	0.9975	0.9604	1.0000	0.9987		1.0000	1.0000
10	0.0286	0.0131	0.3343	0.9991	0.9898	0.9204	0.9998	0.9942	1.0000		1.0000
11	0.1938	0.1149	0.7867	1.0000	1.0000	0.9976	1.0000	1.0000	1.0000	1.0000	

Slika 3.5 prikazuje Box-plot grafički prikaz broja prodanih motora po godinama promatranog razdoblja.

Može se primjetiti kako prvih godina promatranog razdoblja, kada je prodaja motora bila znatno veća nego u ostalim godinama, ima mnogo netipičnih vrijednosti.



Slika 3.5: Box-plot grafički prikaz broja prodanih motora prema godinama promatranog razdoblja (ispis iz SAS-a)

Analiza prodaje motora prema tipu motora

$H_0 : \mu_{Runner} = \mu_{Beverly} = \mu_{Vespa} = ... = \mu_{ZIP}$
H_a : barem dvije sredine se statistički značajno razlikuju

Jednofaktorsku ANOVA-u prema tipu motora radimo za 8 odabranih tipova motora. Naime, naša baza podataka sadrži više od 40 različitih tipova motora, no nisu se svih godina prodavali svi tipovi motora, stoga smo odabrali 8 značajnijih.

Analiziramo li jednakost prodaje prema tipu motora (Tablica 3.11), odbacujemo nultu hipotezu i prihvaćamo alternativnu, tj. postoji statistički značajna razlika u broju prodanih motora za analizirane tipove motora (F=10,09, p<0.0001).

Tablica 3.11: Rezultati ANOVA-e broja prodanih motora prema tipu motora (ispis iz SAS-a)

The GLM Procedure
Dependent Variable: br_prodanih

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	4009.90845	572.84406	10.09	<.0001
Error	293	16637.00849	56.78160		
Corrected Total	300	20646.91694			

R-Square	Coeff Var	Root MSE	br_prodanih Mean
0.194213	117.5203	7.535357	6.411960

Source	DF	Type I SS	Mean Square	F Value	Pr > F
tip	7	4009.908450	572.844064	10.09	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
tip	7	4009.908450	572.844064	10.09	<.0001

Tablica 3.12: Rezultati Tukey-evog post hoc testa broja prodanih motora prema tipovima motora (ispis iz SAS-a)

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

tip	br_prodanih LSMEAN	LSMEAN Number
Beverly	4.6718750	1
Boulevar	4.1666667	2
NRG	6.2631579	3
Pegaso	4.5000000	4
Runner	7.3653846	5
Sportcit	4.3913043	6
Vespa	2.6153846	7
ZIP	14.7619048	8

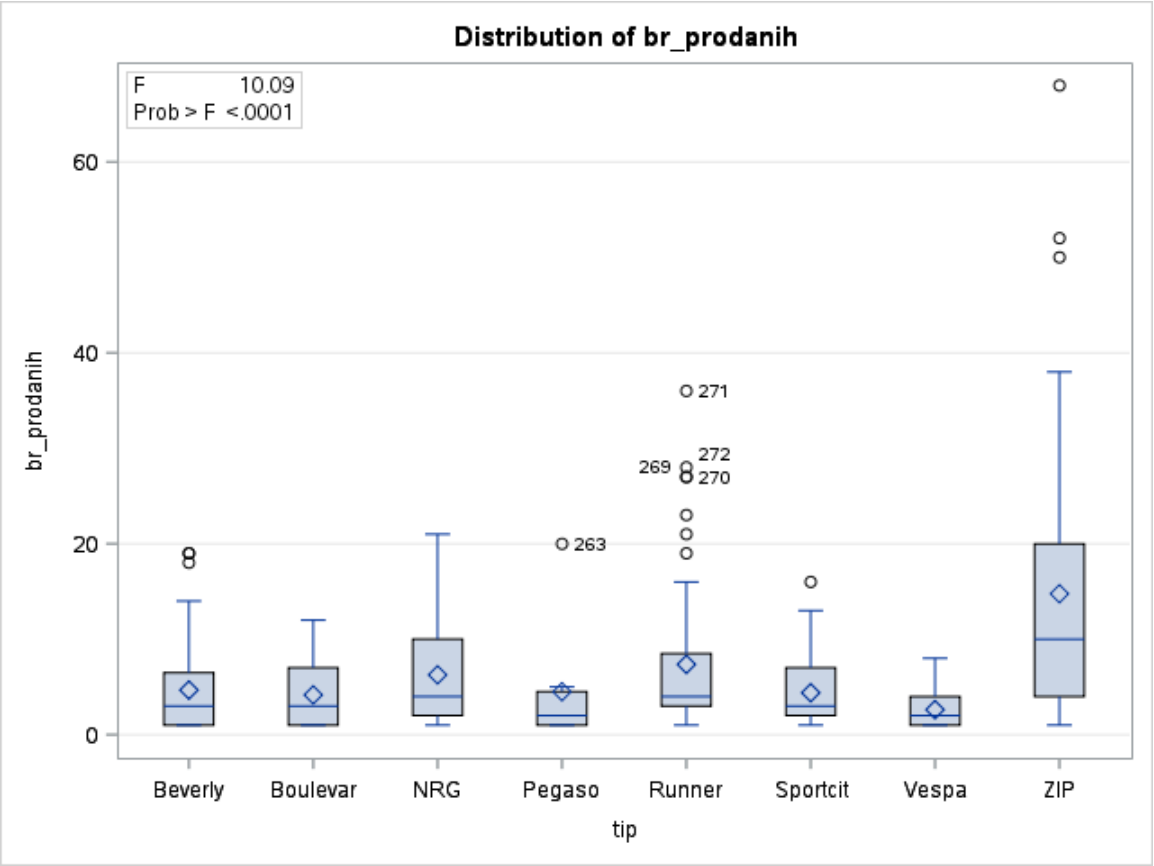
Least Squares Means for effect tip
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: br_prodanih

i/j	1	2	3	4	5	6	7	8
1		1.0000	0.9694	1.0000	0.5420	1.0000	0.8814	<.0001
2	1.0000		0.9906	1.0000	0.8886	1.0000	0.9985	0.0006
3	0.9694	0.9906		0.9988	0.9973	0.9490	0.4021	<.0001
4	1.0000	1.0000	0.9988		0.9740	1.0000	0.9982	0.0112
5	0.5420	0.8886	0.9973	0.9740		0.5177	0.0621	<.0001
6	1.0000	1.0000	0.9490	1.0000	0.5177		0.9600	<.0001
7	0.8814	0.9985	0.4021	0.9982	0.0621	0.9600		<.0001
8	<.0001	0.0006	<.0001	0.0112	<.0001	<.0001	<.0001	

Iz Tablice 3.12 vidimo da se tip motora ZIP statistički značajno razlikuje od svih ostalih tipova. Naime, ZIP je motor marke Piaggio, male jačine i niske cijene koji je jako popularan u prodaji radi navedenih svojstava.

Slika 3.6 prikazuje Box-plot grafički prikaz broja prodanih motora prema tipovima motora.



Slika 3.6: Box-plot grafički prikaz broja prodanih motora prema tipovima motora (ispis iz SAS-a)

3.3 Višefaktorska analiza varijance

Znamo da na prodaju motora ne utječe samo jedan faktor, utječe ih mnogo više. Nakon što smo obradili utjecaj svih faktora baze podataka zasebno na prodaju, prirodno je pitati se možemo li u isto vrijeme ispitati utječu li dva ili više faktora na zavisnu varijablu (prodaju motora) te vidjeti ne samo kako utječe svaki od faktora na zavisnu varijablu, već i njihove zajedničke interakcije (međudjelovanje). Ako nam je ta interakcija značajna, to znači da se zavisna varijabla ne ponaša jednako u oba faktora. Za to ćemo koristiti višefaktorsku ANOVA-u. [5].

Naš model višefaktorske ANOVA-e

Želimo provjeriti kako na prodaju motora utječu godina, marka, cijena, boja motora, jačina motora te interakcija godine i cijene te boje i cijene.

$$\begin{aligned} Y = \mu + godina_i + marka_j + cijena_k + boja_l + kubikaza_m \\ +(godina * cijena)_{ik} + (boja * cijena)_{lk} \\ +\epsilon_{ijklm} \end{aligned} \quad (3.1)$$

gdje su $i = 1, 2, .., 11$, $j = 1, 2, .., 5$, $k = 1, 2, .., 4$, $l = 1, 2, .., 5$, $m = 1, 2, 3$

$godina_i$ predstavlja utjecaj i-tog nivoa efekta godina (ima ih ukupno 11)
$marka_j$ predstavlja utjecaj j-tog nivoa efekta marka (ima ih ukupno 5)
....
$(godina * cijena)_{ik}$ predstavlja utjecaj ik-tog nivoa efekta interakcije godine i cijene
ϵ_{ijklm} je greška od $ijklm$ observacije

SAS kod za model (3.1):

```
PROC GLM data=baza;
CLASS godina marka cijena boja kubikaza;
MODEL brprodanih=godina marka cijena kubikaza boja godina*cijena
boja*cijena;
LSMEANS godina/pdiff adjust=tukey;
LSMEANS marka/pdiff adjust=tukey;
LSMEANS cijena/pdiff adjust=tukey;
LSMEANS kubikaza/pdiff adjust=tukey;
run;
```

Dakle, zanima nas postoji li statistički značajna razlika u prodaji motora obzirom na pojedinačne varijable baze i obzirom na pojedine interakcije varijabli. Prva tablica u ispisu rezultata višefaktorske ANOVA-e (Tablica 3.13) testira sveukupnu značajnost modela. F-vrijednost iznosi 5,85, broj stupnjeva slobode je 78, p<0,0001, stoga zaključujemo da je model statistički značajan.

Tablica 3.13: Rezultati višefaktorske ANOVA-e za model (3.1) (ispis iz SAS-a)

The GLM Procedure
Dependent Variable: br_prodanih

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	78	11888.48740	152.41651	5.85	<.0001
Error	501	13044.63502	26.03720		
Corrected Total	579	24933.12241			

R-Square	Coeff Var	Root MSE	br_prodanih Mean
0.476815	100.9051	5.102666	5.056897

Tablica 3.14: Rezultati višefaktorske ANOVA-e prema modelu (3.1) (ispis iz SAS-a)

Source	DF	Type I SS	Mean Square	F Value	Pr > F
godina	10	1582.658731	158.265873	6.07	<.0001
marka	4	1141.871858	285.467964	10.96	<.0001
cijena	3	3548.989154	1182.996385	45.41	<.0001
kubikaza	15	2323.030669	154.868711	5.94	<.0001
boja	4	106.498331	26.624583	1.02	0.3954
godina*cijena	30	2264.283697	75.476123	2.90	<.0001
cijena*boja	12	913.631972	76.135998	2.92	0.0006

Source	DF	Type III SS	Mean Square	F Value	Pr > F
godina	10	3011.863155	301.186315	11.56	<.0001
marka	4	425.722369	106.430592	4.09	0.0029
cijena	3	686.712408	228.904136	8.79	<.0001
kubikaza	15	4049.602150	269.973477	10.36	<.0001
boja	4	166.605541	41.651385	1.60	0.1733
godina*cijena	30	2459.469245	81.982308	3.15	<.0001
cijena*boja	12	913.631972	76.135998	2.92	0.0006

U Tablici 3.14 se nalaze dvije ANOVA tablice za naš model, u prvoj se računa s tipom sume kvadrata I, a u drugoj se računa s tipom sume kvadrata III. Podsjetimo se, tip sume kvadrata I je hijerarhijski tip, važan je redoslijed varijabli u modelu, svaka varijabla uzima u obzir varijablu koja se pojavila prije nje u modelu. Tip sume kvadrata III se "ponaša" kao da svaka varijabla ulazi zadnja u model.
Proučimo prvo prvu tablicu Tablice 3.14. Iz nje zaključujemo da postoji statistički značajna razlika u prodaji motora po *godini, marki, cijeni, kubikaži* ($p<0,0001$) te u odnosu na interakcije *godina*cijena* ($p<0,0001$) i *cijena*boja* ($p=0,0006$). Interakcije su statistički značajne, to znači da efekt *godine* zavisi od djelovanja efekta *cijene* na zavisnu varijablu *prodaje*. Također, interakcija *cijena*boja* je statistički značajna, tj. djelovanje efekta *cijene* zavisi od djelovanja efekta *boje*. No, ne postoji statistički značajna razlika u odnosu na *boju* ($p=0,39$ što je veće od razine značajnosti).
Iz druge tablice Tablice 3.14 dolazimo do istih zaključaka, no kako se tu računa s tipom sume kvadrata III dobivamo malo drugačije p-vrijednosti za varijable marku i boju.

Tablica 3.15: Rezultati Tukey-evog post hoc testa broja prodanih motora prema markama motora (ispis iz SAS-a)

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

marka	br_prodanih LSMEAN	LSMEAN Number
Aprilia	1.66069040	1
Derbi	-1.22400774	2
Gilera	4.15739743	3
MotoGuzz	-1.37472403	4
Piaggio	2.11123492	5

Least Squares Means for effect marka
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: br_prodanih

i/j	1	2	3	4	5
1		0.2780	0.0192	0.8774	0.9609
2	0.2780		0.0044	1.0000	0.1019
3	0.0192	0.0044		0.4583	0.0564
4	0.8774	1.0000	0.4583		0.8239
5	0.9609	0.1019	0.0564	0.8239	

ANOVA nam kaže da postoji razlika, no još ne znamo koje grupe čine razliku. Stoga, uradimo Tukey-eve post hoc testove.

Jednofaktorskom analizom varijance, kada se promatralo utječe li marka motora na prodaju motora, zaključili smo da utječe te da statistički značajnu razliku čine marke motora Aprilia i Piaggio. Iz Tukey-evog post hoc testa za naš model (3.1), iz Tablice 3.15, zaključujemo da statistički značajnu razliku čini marka motora Gilera obzirom na marke motora Apriliu, Derbi te Piaggio.

Tablica 3.16: Rezultati Tukey-evog post hoc testa broja prodanih motora prema jačinama motora (ispis iz SAS-a)

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

kubikaza	br_prodanih LSMEAN	LSMEAN Number
moped	7.73545789	1
motor	1.08930574	2
skuter	1.55802850	3

Least Squares Means for effect kubikaza
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: br_prodanih

i/j	1	2	3
1		0.0004	<.0001
2	0.0004		0.9470
3	<.0001	0.9470	

Što se tiče jačine motora, u našem modelu (3.1) jačina motora je statistički značajna za prodaju, razliku čini jačinska skupina motora moped s preostale dvije skupine (Tablica 3.16), što je isti rezultat kao kod jednofaktorske ANOVA-e s nezavisnom varijablom jačine motora.

U jednofaktorskom tipu, cijenovni razredi bili su statistički značajni za prodaju motora. Tukey-evim post hoc testom vidjeli smo da statistički značajnu razliku čini cijenovni razred ⩽9999,00 kn sa svim ostalima. Kod našeg modela višefaktorske ANOVA-e, cijenovni razredi također utječu na prodaju. Iz rezultata Tukey-evog post-hoc testa (Tablica 3.17) vidimo da statistički značajnu razliku čini najniži cijenovni razred (⩽9999,00 kn) sa svim ostalima, jednako kao i u jednofaktorskoj ANOVA-i.

Tablica 3.17: Rezultati Tukey-evog post hoc testa broja prodanih motora prema cijenovnim razredima (ispis iz SAS-a)

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

cijena	br_prodanih LSMEAN	LSMEAN Number
10 000-19 999kn	0.20262814	1
20 000-29 999kn	-1.20060909	2
do 9 999kn	5.24375627	3
više od 30 000kn	0.01869745	4

Least Squares Means for effect cijena
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: br_prodanih

i/j	1	2	3	4
1		0.6973	<.0001	0.9996
2	0.6973		0.0004	0.8752
3	<.0001	0.0004		0.0510
4	0.9996	0.8752	0.0510	

Kada smo promatrali kako samo nezavisna varijabla godina utječe na zavisnu varijablu prodaju, zaključili smo da 2006. godina čini statistički značajnu razliku s godinama 2014., 2015, te 2007. godina čini statistički značajnu razliku s godinama 2009.,2012.,2014.,2015.
Kada smo uključili druge varijable te neke interakcije (višefaktorski model (3.1)), iz Tablice 3.18 vidimo da 2006. i 2007. godina čine statistički značajnu razliku sa svim godinama osim s jedna s drugom. Osim toga, 2008. godina čini statistički značajnu razliku sa svim godinama iznad nje osim 2011. i 2013.

Da zaključimo, kod jednofaktorskih ANOVA-i, za sve varijable baze osim *boje motora* postojala je statistički značajna razlika između prosječnog broja prodanih motora i svake od varijabli zasebno. U višefaktorskom modelu (3.1) kojeg smo opisali, postoji statistički značajna razlika između prosječnog broja prodanih motora i svih varijabli (uključujući i obje interakcije) osim varijable *boje motora*.

Tablica 3.18: Rezultati Tukey-evog post hoc testa broja prodanih motora prema godinama promatranog razdoblja (ispis iz SAS-a)

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

godina	br_prodanih LSMEAN	LSMEAN Number
2006	7.42525632	1
2007	7.41613843	2
2008	4.03047928	3
2009	0.79085381	4
2010	0.02652994	5
2011	0.73362276	6
2012	-0.65555256	7
2013	-0.00089592	8
2014	-1.62477054	9
2015	-3.02294345	10
2016	-3.39141794	11

Least Squares Means for effect godina
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: br_prodanih

i/j	1	2	3	4	5	6	7	8	9	10	11
1		1.0000	0.0177	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
2	1.0000		0.0129	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
3	0.0177	0.0129		0.0315	0.0135	0.1454	0.0011	0.0955	0.0022	0.0006	0.0008
4	<.0001	<.0001	0.0315		0.9999	1.0000	0.9689	1.0000	0.8084	0.3686	0.3304
5	<.0001	<.0001	0.0135	0.9999		1.0000	1.0000	1.0000	0.9906	0.7768	0.7034
6	<.0001	<.0001	0.1454	1.0000	1.0000		0.9905	1.0000	0.8893	0.4991	0.4503
7	<.0001	<.0001	0.0011	0.9689	1.0000	0.9905		1.0000	0.9997	0.8916	0.8269
8	<.0001	<.0001	0.0955	1.0000	1.0000	1.0000	1.0000		0.9916	0.7680	0.6982
9	<.0001	<.0001	0.0022	0.8084	0.9906	0.8893	0.9997	0.9916		0.9991	0.9964
10	<.0001	<.0001	0.0006	0.3686	0.7768	0.4991	0.8916	0.7680	0.9991		1.0000
11	<.0001	<.0001	0.0008	0.3304	0.7034	0.4503	0.8269	0.6982	0.9964	1.0000	

Poglavlje 4

Analiza korespondencije

4.1 Uvod

Analiza korespondencije, ili skraćeno CA (*engl. Correspondence Analysis*), je multivarijatna statistička metoda kojom ćemo grafički predočiti varijable iz baze podataka. CA je vrlo slična faktorskoj analizi ali na kategorijskim varijablama i umjesto euklidske udaljenosti uvodi se χ^2-udaljenost. Nekad je lakše razumjeti odnose varijabli ako ih lijepo predočimo grafom, a grafički prikaz može poslužiti kao dobra podloga za dobivanje novih informacija. CA je statistička metoda čiji je razvoj počeo gotovo stoljeće nakon razvoja ANOVA-e. Osnivačem se smatra H. O. Hirschfeld koji joj je postavio temelje 1935. godine, a veliki razvoj i primjena metode kakvu danas koristimo potječe od francuskog statističara i lingvističara Jean-Paul Benzecra (1960). Osim njih, veliki doprinos korespondencijskoj analizi donio je M. J. Greenacre, Benzecreov učenik. Ova analiza je dobila ime korespondencijska jer je *korespondencija* bila oznaka za sistem povezanosti između dva skupa (između redaka i stupaca).
Analiza korespondencije je statistička metoda vizualizacije podataka pomoću koje analiziramo povezanost dviju (jednostavna CA) ili više kategorijskih varijabli (višestruka CA). Kategorije prikazujemo kao točke u nisko-dimenzionalnom prostoru koje će biti blizu jedna drugoj ukoliko su to točke kategorija sa sličnom distribucijom, dok će kategorije koje imaju potpuno različite distribucije biti udaljenije. Prednost metode je mogućnost uspoređivanja kvalitativnih i kvantitativnih varijabli. Udaljenosti u korespondencijskoj analizi mjerimo pomoću χ^2-udaljenosti. [1], [3]

4.2 Jednostavna analiza korespondencije

Analiza korespondencije u SAS-u

Procedurom CORRESP u programskom sustavu SAS radimo jednostavnu i višestruku analizu korespondencije. Grafički prikaz rezultata prikazuje se pomoću ODS Graphics, a ispis rezultata uključuje i dekompoziciju inercije i koordinate. Postoje dvije forme unosa u navedenu proceduru, prva je naredbom VAR ako već imamo podatke u formi tablice, a druga forma je naredbom TABLES ako unosimo sirove kategorijske podatke. Mi ćemo koristiti drugu naredbu. Koristimo i naredbu WEIGHT jer moramo brojiti prodane motore po svakoj observaciji, u protivnom bi svaka observacija predstavljala jedan subjek što nije slučaj naše baze podataka.[8]

SAS kod za jednostavnu CA:

PROC CORRESP *all data=baza outc=Coor;*
TABLES marka,cijena;
WEIGHT brprodanih;
run;

Navedeni SAS kod nam daje Sliku 4.2.

Opis jednostavne analize korespondencije

χ^2-udaljenost je temelj analize korespondencije. Sjetimo se formule za χ^2-statistiku

$$H = \sum \frac{(f_o - f_t)^2}{f_t} \tag{4.1}$$

gdje su f_o opažene (promatrane) frekvencije, f_t teorijske (očekivane) frekvencije. Interpretacija konfiguracije točaka bazira se na χ^2-udaljenostima između točaka. Ako npr. dvije redčane točke leže jedna blizu druge tada su i *profili* (sinonim za relativne frekvencije u CA) tih točaka slični. Sa

$$\frac{\chi^2}{n}$$

označavamo *inerciju*, to je mjera varijacije u kontigencijskoj tablici. Geometrijski gledano, inercija mjeri koliko je svaki od redaka udaljen od prosječnog retka. Tablica 4.1 prikazuje broj prodanih motora po markama motora u razdoblju od 2006.

do 2016. godine obzirom na cijenovne razrede. Primjerice, u promatranom razdoblju je prodano 248 motora marke Aprilia koji su cijene između 10000,00 kn i 19999,00 kn. Naime, primjenom korespondencijske analize nad ovom tablicom dobiti ćemo geometrijsku interpretaciju podataka prikazanu na Slici 4.2 koju nazivamo *simetričnom mapom*. Opišimo postupak analize korespondencije.
Analizu provodimo posebno nad retcima, odnosno stupcima. Usredotočiti ćemo se na analizu nad retcima jer je postupak analogan za stupce. Prvo trebamo iz kontigencijske Tablice 4.1 izračunati tablicu relativnih frekvencija (profila). Profil retka za marku motora Aprilia dobije se tako da svaki element u retku marke Aprilia podijelimo sa sumom toga retka (248/461 = 0,53796 , 100/461 = 0,21692, 19/461 = 0,04121, 94/461 = 0,2039). Nakon što to napravimo za sve retke, dobijemo Tablicu 4.2. Svaki profil retka možemo prikazati kao vektor kojeg možemo promatrati kao točku u prostoru gdje svaki profil predstavlja pojedinu koordinatu. Najsličniji profili retka bit će prikazani kao dvije bliske točke dok će različitiji predstavljati udaljenije točke. Prosječni profil retka predstavlja težiniski prosjek te tu točku nazivamo *centroidom* i prikazujemo je kao ishodište koordinatnih osi.

Tablica 4.1: Kontigencijska tablica marki motora prema cijenovnim razredima (ispis iz SAS-a)

The CORRESP Procedure

Contingency Table

	10 000-19 999kn	20 000-29 999kn	do 9 999kn	više od 30 000kn	Sum
Aprilia	248	100	19	94	461
Derbi	13	0	46	0	59
Gilera	338	134	0	22	494
MotoGuzz	0	0	0	11	11
Piaggio	607	223	948	130	1908
Sum	1206	457	1013	257	2933

Tablica 4.2: Profili redaka kontigencijske Tablice 4.1 (ispis iz SAS-a)

Row Profiles

	10 000-19 999kn	20 000-29 999kn	do 9 999kn	više od 30 000kn
Aprilia	0.53796	0.21692	0.04121	0.20390
Derbi	0.22034	0.00000	0.77966	0.00000
Gilera	0.68421	0.27126	0.00000	0.04453
MotoGuzz	0.00000	0.00000	0.00000	1.00000
Piaggio	0.31813	0.11688	0.49686	0.06813

Tablica 4.3: Profili stupaca kontigencijske Tablice 4.1 (ispis iz SAS-a)

Column Profiles				
	10 000-19 999kn	20 000-29 999kn	do 9 999kn	više od 30 000kn
Aprilia	0.205638	0.218818	0.018756	0.365759
Derbi	0.010779	0.000000	0.045410	0.000000
Gilera	0.280265	0.293217	0.000000	0.085603
MotoGuzz	0.000000	0.000000	0.000000	0.042802
Piaggio	0.503317	0.487965	0.935834	0.505837

Slično se izračunaju stupčani profili (Tablica 4.3). Proučimo li prva dva stupca Tablice 4.3, možemo zaključiti da će cijenovni razredi 10000,00-19999,00 kn i 20000,00-29999,00 kn biti jako blizu na grafu za razliku od preostala dva cijenovna razreda.
Drugi stupac u Tablici 4.4 predstavlja *mase* retka koje dobijemo kao profile sume redaka (za marku Aprilia 461/2933 = 0,15717). Masa svakog retka predstavlja mjeru značajnosti te marke u analizi. Profil retka koji malo odstupa od prosjeka bit će bliže ishodištu dok profil retka koji dosta odstupa od prosjeka, bit će udaljeniji od ishodišta. Vidimo da će marka motora Piaggio biti blizu ishodišta, dok marka Moto Guzzi neće. Slično, drugi stupac u Tablici 4.5 predstavlja mase stupaca tj. cijenovnih razreda.

Tablica 4.4: Rezime statistike za retčane točke (ispis iz SAS-a)

Summary Statistics for the Row Points			
	Quality	Mass	Inertia
Aprilia	1.0000	0.1572	0.2495
Derbi	0.9785	0.0201	0.0578
Gilera	0.9999	0.1684	0.3489
MotoGuzz	0.9998	0.0038	0.1277
Piaggio	0.9996	0.6505	0.2161

Tablica 4.5: Rezime statistike za stupčane točke (ispis iz SAS-a)

Summary Statistics for the Column Points			
	Quality	Mass	Inertia
10 000-19 999kn	0.9977	0.4112	0.1756
20 000-29 999kn	0.9894	0.1558	0.0923
do 9 999kn	1.0000	0.3454	0.5095
više od 30 000kn	1.0000	0.0876	0.2225

Dalje, računamo očekivane frekvencije za kontigencijsku tablicu marki motora prema cijenovnim razredima, od koje smo krenuli, Tablicu 4.1. Očekivane frekvencije su prikazane u Tablici 4.6.

Tablica 4.6: Očekivane frekvencije kontigencijske Tablice 4.1 (ispis iz SAS-a)

Chi-Square Statistic Expected Values				
	10 000-19 999kn	20 000-29 999kn	do 9 999kn	više od 30 000kn
Aprilia	189.555	71.830	159.220	40.394
Derbi	24.260	9.193	20.377	5.170
Gilera	203.124	76.972	170.618	43.286
MotoGuzz	4.523	1.714	3.799	0.964
Piaggio	784.537	297.292	658.985	167.186

Potom računamo razlike između očekivanih i opaženih frekvencija da bi mogli računati χ^2-statistiku.

Tablica 4.7: Doprinosi pojedinih točaka χ^2-statistici (ispis iz SAS-a)

Contributions to the Total Chi-Square Statistic					
	10 000-19 999kn	20 000-29 999kn	do 9 999kn	više od 30 000kn	Sum
Aprilia	18.020	11.048	123.488	71.137	223.692
Derbi	5.226	9.193	32.218	5.170	51.807
Gilera	89.558	42.252	170.618	10.467	312.896
MotoGuzz	4.523	1.714	3.799	104.501	114.537
Piaggio	40.176	18.565	126.755	8.271	193.767
Sum	157.503	82.772	456.877	199.546	896.698

Pomoću formule (4.1) za χ^2-udaljenost imamo χ^2-statistiku. Iz Tablice 4.7 vidimo da ona iznosi 896,70 te usporedimo taj rezultat s graničnom vrijednosti χ^2-razdiobe na razini značajnosti od 5% i s 12 stupnjeva slobode, koje računamo na način:

$$(broj\ redaka - 1) * (broj\ stupaca - 1) = 4 * 3 = 12$$

$\chi^2_{0,05}(12) = 21{,}03$. Kao što smo i očekivali, χ^2 statistika dosta odstupa od granične, podaci nisu homogeni, odnosno prodaja marki motora po cijenovnim razredima se razlikuje.

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent
0.48835	0.23849	699.486	78.01	78.01
0.25848	0.06681	195.966	21.85	99.86
0.02062	0.00043	1.247	0.14	100.00
	0.30573	896.698	100.00	

Degrees of Freedom = 12

Slika 4.1: Singularne vrijednosti, inercija, postotak objašnjene inercije po dimenzijama (ispis iz SAS-a)

Slika 4.1 je dio SAS ispisa i prikazuje dekompoziciju svojstvenih vrijednosti za konačno rješenje u dvije dimenzije. Iz nje čitamo χ^2-statistiku, koliko je inercije objašnjeno prvom dimenzijom (78,01%), koliko drugom dimenzijom (21,85%) te koliko je ostalo neobjašnjeno.
SAS nam daje i najbolju dimenziju za sve marke motora (Tablica 4.8) i za sve cijenovne razrede (Tablica 4.9), te ispada da su sve marke motora osim Moto Guzzi dobro objašnjene prvom dimenzijom kao i svi cijenovni razredi osim najvišeg.

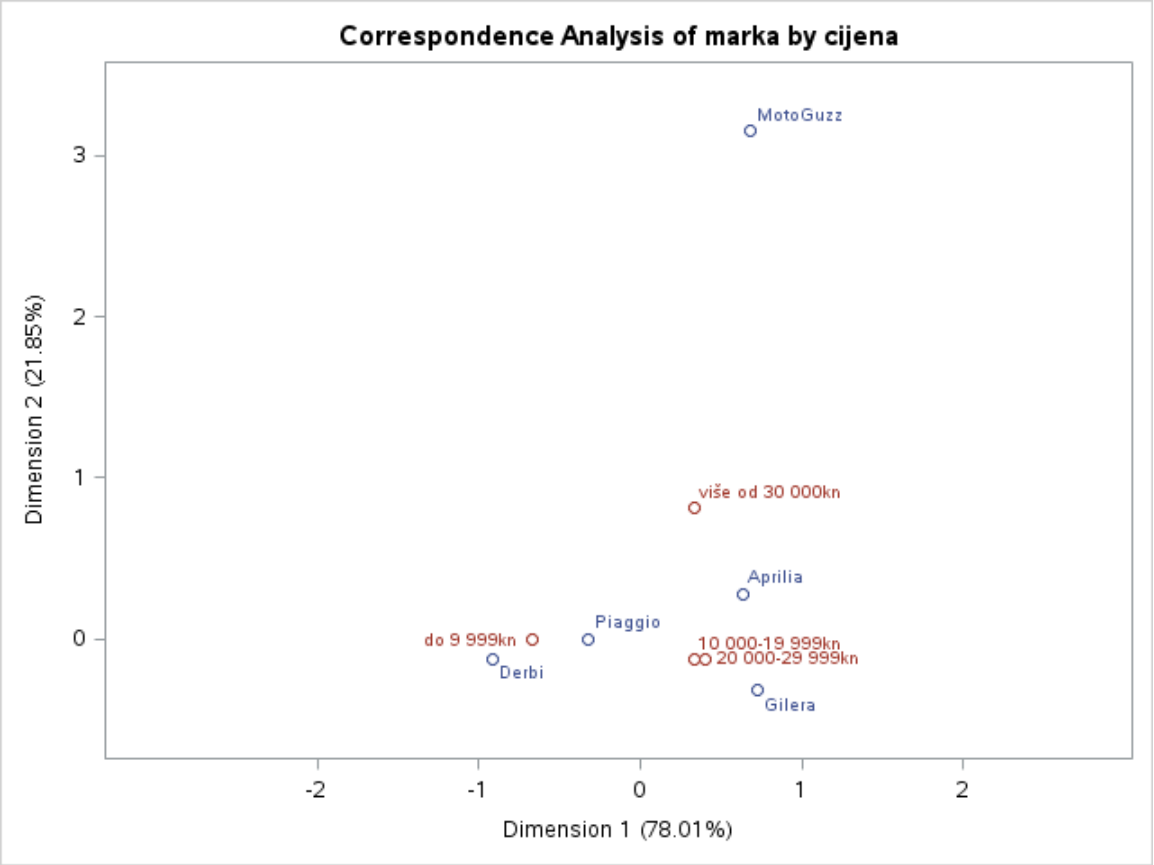
Tablica 4.8: Najbolje dimenzije za marke motora (ispis iz SAS-a)

Indices of the Coordinates That Contribute Most to Inertia for the Row Points			
	Dim1	Dim2	Best
Aprilia	1	0	1
Derbi	0	0	1
Gilera	1	1	1
MotoGuzz	0	2	2
Piaggio	1	0	1

Tablica 4.9: Najbolje dimenzije za cijenovne razrede (ispis iz SAS-a)

Indices of the Coordinates That Contribute Most to Inertia for the Column Points			
	Dim1	Dim2	Best
10 000-19 999kn	1	0	1
20 000-29 999kn	0	0	1
do 9 999kn	1	0	1
više od 30 000kn	0	2	2

Prva dimenzija simetrične mape na Slici 4.2 opisuje 78,01% inercije, a druga 21,85% inercije što znači da prve dvije dimenzije objašnjavaju 99.86% ukupne inercije što je jako dobro jer će se zaključci izvedeni iz grafa temeljiti na skoro savršenim informacijama dobivenim iz Tablice 4.1.



Slika 4.2: Rezultati korespondencijske analize primjenjene nad Tablicom 4.1 (ispis iz SAS-a)

Uočimo da su, obzirom na prvu dimenziju, cijene više od 10000,00 kn pozitivne dok su najniže cijene negativne. Također možemo vidjeti da je oko najnižeg cijenovnog razreda marka Derbi, dok je Moto Guzzi marka najudaljenija od svih, i grupira se sa najvišim cijenovnim razredom. Položaj točaka ukazuje da su marke Gilera i Aprilia povezane sa srednjim cijenovnim razredima (10000,00 do 30000,00 kn) koji su jako blizu kao što smo i pretpostavili promatrajući njihove profile u Tablici 4.3. Ipak, marka Aprilia više teži najskupljem cijenovnom razredu. Marka motora

Piaggio najbliža je najnižem cijenovnom razredu, zatim srednjim cijenovnim razredima. Kada bi povukli dva okomita pravca kroz ishodišta dimenzija, vidjela bi se upravo opisana podjela marki motora i cijenovnih razreda.

Cilj korespondencijske analize je transformirati Tablicu 4.1 u simetričnu mapu (Slika 4.2) u kojoj svaki redak i stupac predstavljaju točku. U većim kontigencijskim tablicama pokušavamo opisati podatke sa što manje parametara, tj. u što manje dimenzija, ali tako da je opis što potpuniji odnosno da je inercija objašnjena u što većem postotku. Inercija je mjera za preciznost prikaza. Koordinate su baza grafičkog prikaza i preko njih su prikazane osnovne informacije o položaju točaka.

4.3 Višestruka analiza korespondencije

Jednostavnu analizu korespondencije koristimo kad imamo dvije kategorijske varijable, ali ova metoda je primjenjiva i na više od dvije kategorijske varijable.
Za višestruku analizu korespondencije, ili skraćeno MCA (*engl. Multiple Correspondence Analysis*), postoje dva pristupa. Prvi je s *indikatorskom matricom* (zamislimo da analiziramo skup podataka u kojem su pomoćne varijable (*engl. dummy*)), drugi je pomoću *Burtove matrice* (kontigencijske tablice svih mogućih kombinacija varijabli), odnosno analizom međusobnog odnosa (*engl. crosstabulation*). Ideja MCA je prikazati više kategorijskih varijabli u dvodimenzionalnom prostoru uz što manji gubitak informacija.

Višestruka analiza korespondencije za tip motora, cijenovni razred i boju:

SAS kod za višestruku CA:

PROC CORRESP *MCA observed data=baza outc=Coor;*
TABLES tip cijena boja;
WEIGHT brprodanih;
run;

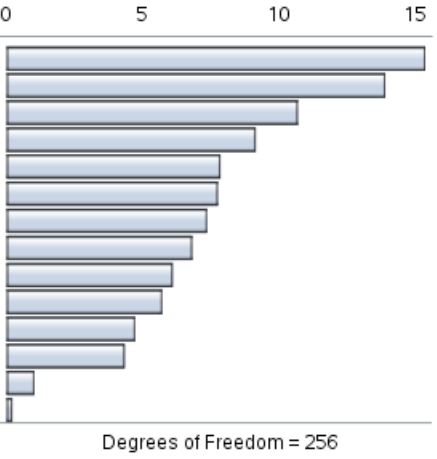
Navedeni SAS kod izvršava MCA pomoću Burtove tablice. Kao i kod jednostavne analize korespondencije, radimo u PROC CORRESP proceduri u SAS-u i koristimo slične naredbe. Za razliku od jednostavne analize korespondencije, nakon imena procedure u kojoj radimo koristimo *MCA* da bi SAS-u dali naredbu da želimo

višestruku analizu korespondencije. U višestrukoj CA ne stavljamo zareze između kategorijskih varijabli u naredbi TABLES.
Kao i prije, naredbom WEIGHT brojimo prodane motore po observacijama.

Slika 4.3 dio je SAS ispisa, daje nam singularne vrijednosti, inerciju, χ^2-statistiku, postotak objašnjene inercije po svakoj od dimenzija.

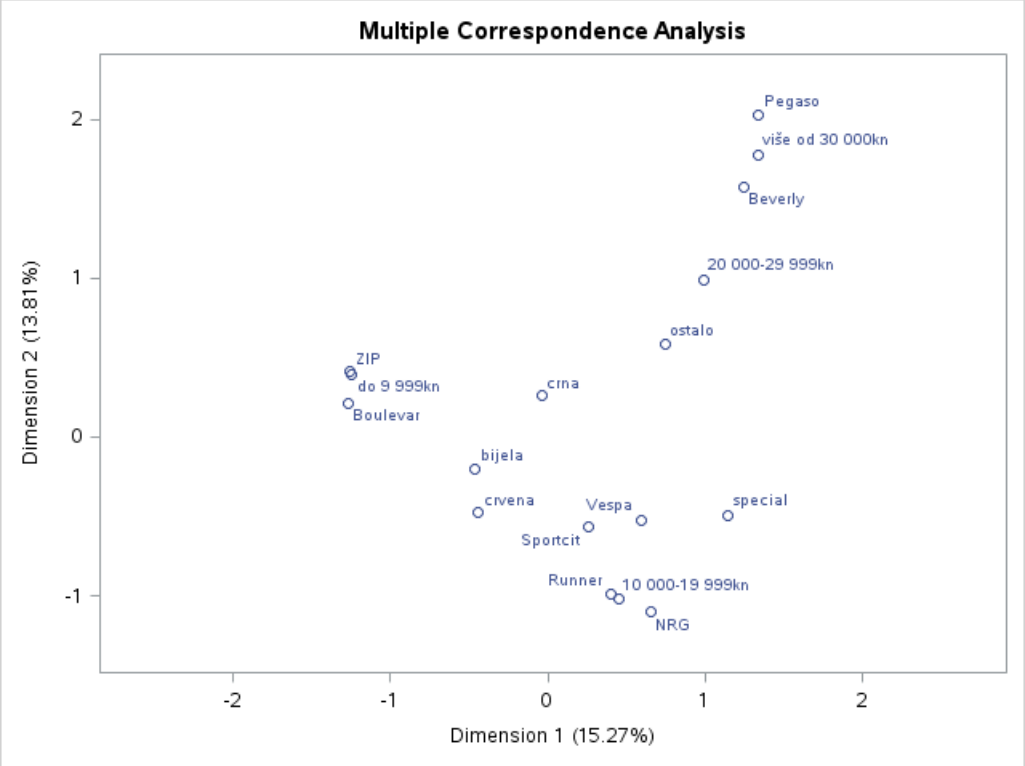
Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent
0.84410	0.71251	5463.9	15.27	15.27
0.80267	0.64428	4940.6	13.81	29.07
0.70374	0.49525	3797.9	10.61	39.69
0.65054	0.42321	3245.4	9.07	48.76
0.60240	0.36288	2782.8	7.78	56.53
0.59887	0.35865	2750.3	7.69	64.22
0.58315	0.34006	2607.8	7.29	71.50
0.56116	0.31490	2414.8	6.75	78.25
0.53073	0.28167	2160.0	6.04	84.29
0.51338	0.26355	2021.1	5.65	89.93
0.46611	0.21725	1666.0	4.66	94.59
0.44693	0.19975	1531.8	4.28	98.87
0.21254	0.04517	346.4	0.97	99.84
0.08680	0.00753	57.8	0.16	100.00
	4.66667	35786.5	100.00	



Slika 4.3: Singularne vrijednosti, inercija, postotak objašnjene inercije po dimenzijama za MCA (ispis iz SAS-a)

Na Slici 4.4 prikazana je MCA za kategorijske varijable tip motora (8 izdvojenih tipova), boja motora i cijenovni razred. Prvom dimenzijom opisano je 15,27% ukupne inercije, a drugom 13,81% ukupne inercije.
Na jednostavan način, MCA daje uvid u kompleksne veze između navedene tri kategorijske varijable.



Slika 4.4: MCA za kategorijske varijable: tip, boja, cijenovni razred (ispis iz SAS-a)

Primjetimo kako su točke koje predstavljaju dva niža cijenovna razreda ($\leqslant$9999,00 kn i 10000,00-19999,00 kn) i one koje predstavljaju dva viša cijenovna razreda grupirane u klastere. Oko najnižeg cijenovnog razreda grupirali su se tipovi motora ZIP i Boulevard za koje smo rekli da su najnižih cijena i od svih tipova motora, oni se jedini nalaze na negativnoj strani horizontalne dimenzije. Kod drugog klastera (oko cijenovnog razreda 10000,00-19999,00 kn) grupirali su se tipovi motora Vespa, Sportcity, Runner i NRG, a od boja, special je najbliža tom razredu odnosno tim tipovima motora. Kod najviših cijenovnih razreda nalazimo tipove motora Pegaso i Beverly, s tim da je Pegaso najbliži najvišem cijenovnom razredu, dok Beverly ”naginje” tom cijenovnom razredu, ali nije daleko od cijenovnog razreda 20000,00-29999,00 kn. Uočimo da je crna boja između navedena tri klastera, ima je u svim cijenovnim razredima i postoje gotovo svi tipovi crnih motora, a bijela boja se nalazi između dva jeftinija cijenovna razreda. Skupina boja koju smo nazvali ostalo nalazi se između klastera određenog najskupljim motorima i klastera određenog cijenovnim razredom 10000,00-19999,00 kn.

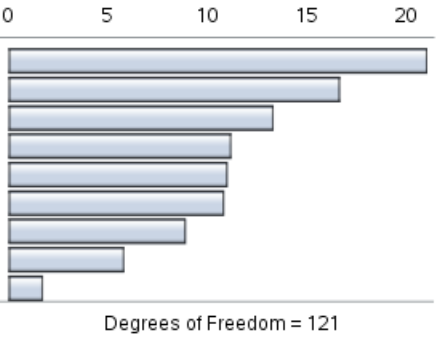
Višestruka analiza korespondencije za jačinu, boju i cijenovni razred

Kako smo se u radu dosta usredotočili na kretanje prodaje motora po njihovim markama, uradili smo višestruku analizu korespondencije motora bez obzira na marke i tipove motora. Želimo vidjeti kako se grupiraju jačine motora, boje i cijenovni razredi.

Slika 4.5 dio je SAS ispisa, daje nam singularne vrijednosti, inerciju, χ^2-statistiku, postotak objašnjene inercije po svakoj od dimenzija.

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent
0.79355	0.62972	6816.0	20.99	20.99
0.70587	0.49825	5393.0	16.61	37.60
0.63072	0.39781	4305.8	13.26	50.86
0.57802	0.33410	3616.3	11.14	62.00
0.57330	0.32867	3557.5	10.96	72.95
0.56870	0.32341	3500.6	10.78	83.73
0.51543	0.26567	2875.6	8.86	92.59
0.41531	0.17248	1866.9	5.75	98.34
0.22334	0.04988	539.9	1.66	100.00
	3.00000	32471.5	100.00	

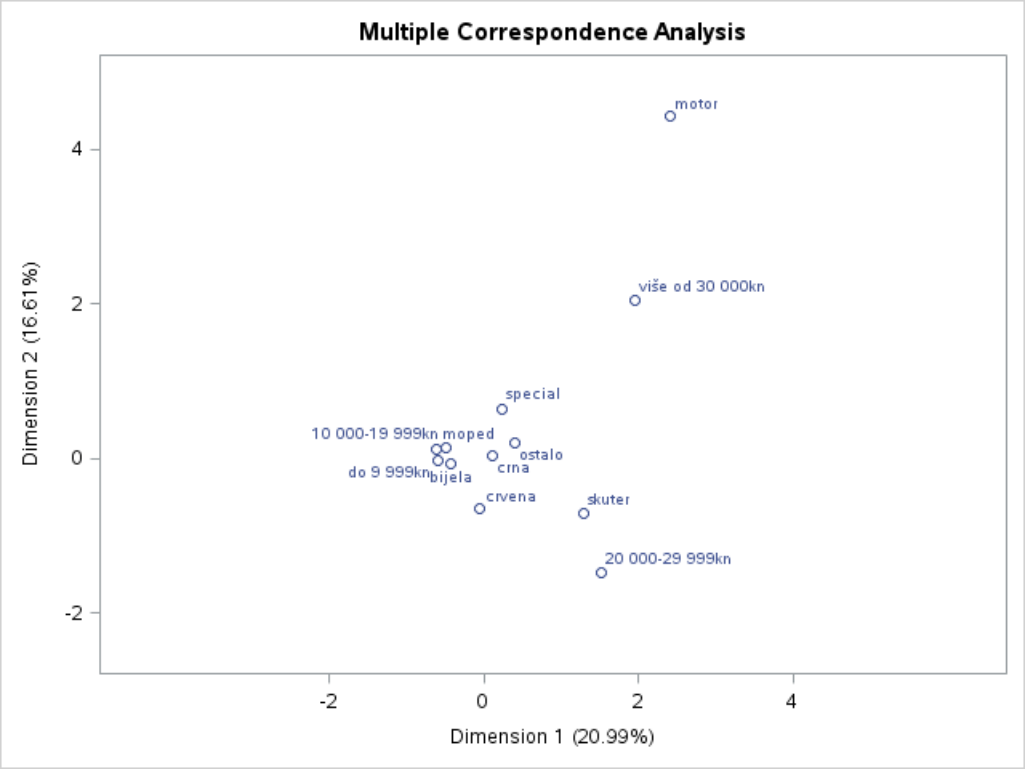


Slika 4.5: Singularne vrijednosti, inercija, postotak objašnjene inercije po dimenzijama za MCA (ispis iz SAS-a)

Višestrukom analizom korespondencije dobijemo simetričnu mapu prikazanu na Slici 4.6.

Prvom dimenzijom opisano je 20,99% ukupne inercije, a drugom 16,61%.
Prvo što možemo uočiti je da postoji netipična vrijednost (*engl. outlier*) motor, koji je određen motorima najvećih jačina. Oko njega se grupira cijenovni razred cijena viših od 30000,00 kn.
Dalje primjećujemo središnji klaster i četiri točke unutar njega koje se skoro pa preklapaju. Te četiri točke predstavljaju motori jačinskog razreda moped (dakle motori najslabijih jačina), bijela boja motora te dva niža cijenovna razreda (cijene motora do 20000,00 kn). Preostale boje su jako blizu te četiri grupirane točke, s tim da crvena boja motora i boja koju smo nazvali ostalo teži jačinskom razredu skutera, a special boja teži najvišem cijenovnom razredu. Primjetimo da je crna

boja motora opet najbliža ”središtu”, što se moglo i očekivati jer smo vidjeli da je najdominantnija boja.



Slika 4.6: MCA za kategorijske varijable: jačina, boja, cijena (ispis iz SAS-a)

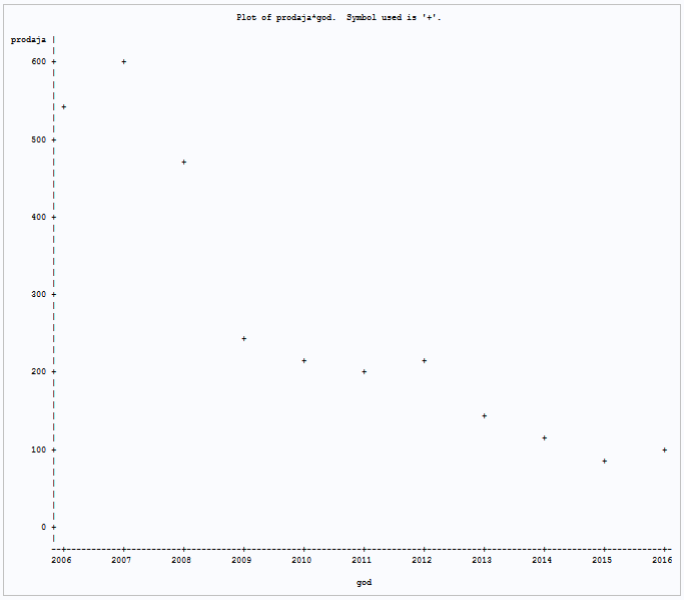
Obzirom na prvu dimenziju, dva jeftinija cijenovna razreda nalaze se na negativnoj strani obzirom na ishodište, a dva skuplja su s pozitivne strane. Najslabija jačinska skupina motora, mopedi, se nalazi na negativnoj strani obzirom na prvu dimenziju, a ostale jačinske skupine (skuteri i motori) se nalaze na pozitivnoj strani prve dimenzije.

Poglavlje 5

Linearna regresija i predikcija

5.1 Linearna regresija

Pojave promatrane u vremenu najčešće pokazuju određeno kretanje i tendenciju. Znamo kako se kretala prodaja unazad 11 godina (pad prodaje nakon 2008. godine), stoga možemo pretpostaviti da imamo padajući linearni trend, da se prodaja nije odvijala na slučajan način već postoji neka povezanost među godinama.



Slika 5.1: Graf broja prodanih motora od 2006. do 2016. godine (ispis iz SAS-a)

Nacrtamo li u SAS-u pomoću PROC PLOT naredbe broj prodanih motora po

godinama promatranog razdoblja, možemo uočiti da postoji linearna veza prodaje motora po godinama promatranog razdoblja (Slika 5.1). Statistički postupak za procjenu i kvantifikaciju odnosa među varijablama zovemo regresijska analiza. Cilj istraživanja odnosa je utvrditi statističku ovisnost i pokazatelje jakosti takve ovisnosti.

Jednadžba jednostavne linearne regresije glasi : $Y = a + bX + \epsilon$.

Y je zavisna varijabla, u našem slučaju će Y biti broj prodanih motora. X je nezavisna varijabla tj. vrijeme (u godinama), ϵ je nepoznata komponenta greške koja je dodana na linearnu vezu, a a i b su nepoznati parametri pretpostavljene veze koje treba procijeniti.
Do procjene parametara se najčešće dolazi metodom najmanjih kvadrata. Parametar a je *regresijska vrijednost* zavisne varijable ako je nezavisna jednaka nuli (konstantni član), a b je *regresijski koeficijent* koji nam pokazuje koliko se u prosjeku mijenja vrijednost zavisne varijable Y za jediničnu promjenu vrijednosti nezavisne varijable X.

Pretpostavke koje moraju biti zadovoljene za linearni regresijski model su:
1) veza između zavisne varijable i nezavisne je (barem približno) linearna
2) greške ϵ imaju očekivanje 0 i konstantnu varijancu σ^2
3) greške su nekorelirane
4) greške su normalno distribuirane

Pretpostavke 3) i 4) zajedno povlače da su greške nezavisne slučajne varijable. Za linearnu regresijsku analizu trebaju se provjeriti gornje pretpostavke i ispitati adekvatnost dobivenog modela. Metode ispitivanja adekvatnosti modela temelje se na proučavanju reziduala modela.

SAS kod:

Radimo u PROC REG proceduri u SAS-u. Varijabla *prodaja* označava broj prodanih motora po godini, a varijabla *t* broji godine počevši od 2006. Naredbom MODEL u PROC REG proceduri zadajemo model kojeg želimo, tako da s lijeve strane jednakosti stavimo nezavisnu varijablu, a s desne zavisnu (jednu zavisnu ili više njih).

PROC REG *data=baza;*
MODEL prodaja=t;
run;

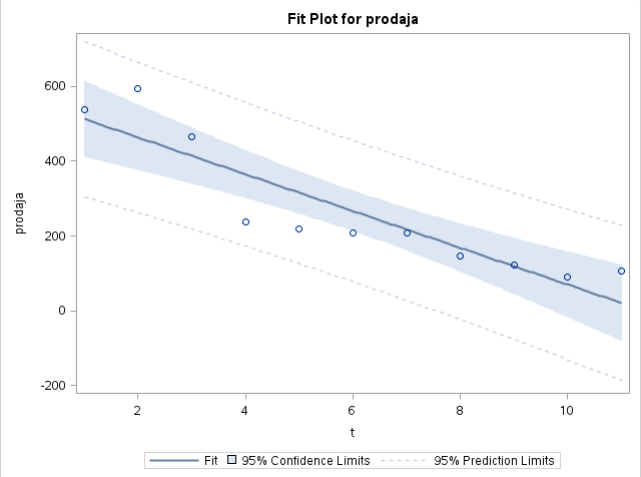
Rezultati

Uobičajeno se univarijatni regresijski model gdje je nezavisna varijabla vrijeme zove *trend*. Analizirajmo linearnu regresiju za ukupni broj prodanih motora od 2006. do 2016. godine da vidimo koliki je pad prodaje motora. Testiramo ima li godina utjecaj na prodaju.

Trend prodaje motora

Procijenjena jednadžba glasi : $prodaja = 561,89 - 49,21 * t$

Prva tablica Tablice 5.1 je ANOVA tablica, iz rezultata ANOVA-e je jasno da je model statistički značajan (postoji značajna promjena u prodaji motora po godinama), odbacujemo nulhipotezu da je regresijski koeficijent jednak nuli (p<0,0001). Iz druge tablice Tablice 5.1 čitamo da R^2 iznosi 82,22%, što znači da je odabranim modelom protumačeno 82,22% svih odstupanja. Također vidimo da je procijenjena standardna devijacija (*engl. Root MSE*) jednaka 80,01 što znači da je prosječno odstupanje empirijskih od regresijskih vrijednosti prodaje 80 motora po godini.
Iz posljednje tablice procjene parametara (*engl. Parameter Estimates*) Tablice 5.1 dobivamo procijenjene parametre. Konstantni član je dan pod *Intercept*. Negativan i dosta velik regresijski koeficijent, *b*=-49,21 ukazuje na drastičan pad prodaje motora u promatranom razdoblju. Procijenjeni parametri su statistički značajni. Iz procijenjene jednadžbe čitamo da će se povećanjem vremena za jednu jedinicu (godinu), prodaja smanjiti za gotovo 50 motora.[5]



Slika 5.2: Grafički prikaz linearnog trenda prodaje motora s 95% pouzdanim intervalom (ispis iz SAS-a)

Na Slici 5.2 vidimo jednadžbu negativnog linearnog trenda s pripadnim 95% pouzdanim intervalom.

Tablica 5.1: Rezultati linearne regresije za broj prodanih motora po godinama promatranog razdoblja (ispis iz SAS-a)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	266369	266369	41.61	0.0001
Error	9	57612	6401.30404		
Corrected Total	10	323981			

Root MSE	80.00815	R-Square	0.8222
Dependent Mean	266.63636	Adj R-Sq	0.8024
Coeff Var	30.00647		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	561.89091	51.73885	10.86	<.0001
t	1	-49.20909	7.62848	-6.45	0.0001

Trend prodaje marke motora Aprilia

$Aprilia = 93,24 - 8,55 * t$

Ovaj model je statistički značajan, p=0,0004 (Tablica 5.2). Uočimo da prodaja motora po marki motora Aprilia pada po godinama, regresijski koeficijent je negativan i iznosi -8,55, što nam govori da se godišnje broj prodanih motora Aprilia marke u prosjeku smanjuje za 8,55 komada. Iz druge tablice Tablice 5.2 čitamo da R^2 iznosi 76,88%, što znači da je odabranim modelom protumačeno 76,88% svih odstupanja. Iz treće tablice Tablice 5.2 vidimo da su procijenjeni parametri statistički značajni.

Tablica 5.2: Rezultati linearne regresije za prodaju marke motora Aprilia u ovisnosti u godinama (ispis iz SAS-a)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	8049.82727	8049.82727	29.92	0.0004
Error	9	2421.08182	269.00909		
Corrected Total	10	10471			

Root MSE	16.40150	R-Square	0.7688
Dependent Mean	41.90909	Adj R-Sq	0.7431
Coeff Var	39.13589		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	93.23636	10.60635	8.79	<.0001
t	1	-8.55455	1.56382	-5.47	0.0004

Trend prodaje marke motora Piaggio

Tablica 5.3: Rezultati linearne regresije za prodaju marke motora Piaggio u ovisnosti o godinama (ispis iz SAS-a)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	89804	89804	29.21	0.0004
Error	9	27671	3074.51616		
Corrected Total	10	117475			

Root MSE	55.44832	R-Square	0.7645
Dependent Mean	173.45455	Adj R-Sq	0.7383
Coeff Var	31.96706		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	344.89091	35.85675	9.62	<.0001
t	1	-28.57273	5.28679	-5.40	0.0004

$Piaggio = 344,89 - 28,57 * t$

Iz Tablice 5.3 se vidi da je model statistički značajan. Prodaja motora po marki motora Piaggio pada kroz godine, regresijski koeficijent iznosi -28,57. Tumačimo da se prodaja Piaggio motora po godini smanjuje za otprilike 28,57 motora. Sjetimo se da Piaggio marka čini velik dio tržišta stoga se moglo i očekivati da će regresijski koeficijent za marku motora Piaggio biti najveći.

Trend prodaje marke motora Gilera
$Gilera = 119,53 - 12,44 * t$

Tablica 5.4: Rezultati linearne regresije za prodaju marke motora Gilera u ovisnosti o godinama (ispis iz SAS-a)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	17013	17013	28.59	0.0005
Error	9	5355.96364	595.10707		
Corrected Total	10	22369			

Root MSE	24.39482	R-Square	0.7606
Dependent Mean	44.90909	Adj R-Sq	0.7340
Coeff Var	54.32044		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	119.52727	15.77539	7.58	<.0001
t	1	-12.43636	2.32595	-5.35	0.0005

Iz Tablice 5.4 vidimo da je model statistički značajan, p-vrijednost iznosi 0,0005.

Trend prodaje marke motora Moto Guzzi

$MotoGuzzi = 3,40 - 0,40 * t$

Za marku motora Moto Guzzi model je također statistički značajan. Iz Tablice 5.5 vidimo da je p=0,0313 što je manje od razine značajnosti. Odbacujemo nul-hipotezu da je regresijski koeficijent jednak nuli, tj. godina ima utjecaj na prodaju

marke motora Moto Guzzi.
Regresijski koeficijent marke motora Moto Guzzi je najmanji od svih regresijskih koeficijenata marki kojima prodaja pada. Moto Guzzi marka je marka najjačih i najskupljih motora koja se tijekom tržišne krize prestala prodavati (oko 2011. godine), a prije krize nije brojala velik broj prodanih motora po godini. Stoga, jasno je da ima najmanji regresijski koeficijent.

Tablica 5.5: Rezultati linearne regresije za prodaju marke motora Moto Guzzi u ovisnosti o godinama (ispis iz SAS-a)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	17.60000	17.60000	6.49	0.0313
Error	9	24.40000	2.71111		
Corrected Total	10	42.00000			

Root MSE	1.64655	R-Square	0.4190
Dependent Mean	1.00000	Adj R-Sq	0.3545
Coeff Var	164.65452		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.40000	1.06477	3.19	0.0110
t	1	-0.40000	0.15699	-2.55	0.0313

Trend prodaje marke motora Derbi

Iz Tablice 5.6 vidimo da model nije statistički značajan za marku motora Derbi, p-vrijednost iznosi 0,3009 što je veće od razine značajnosti, stoga ne odbacujemo nulhipotezu da je regresijski koeficijen b jednak nuli. Jednadžba za marku Derbi jedina ima pozitivan regresijski koeficijent (0,75) tj. uočavamo rast prodaje samo kod marke motora Derbi. No, na žalost taj rezultat nije statistički značajan.

$$Derbi = 0,84 + 0,75 * t$$

Tablica 5.6: Rezultati linearne regresije za prodaju marke motora Derbi u ovisnosti o godinama (ispis iz SAS-a)

Analysis of Variance

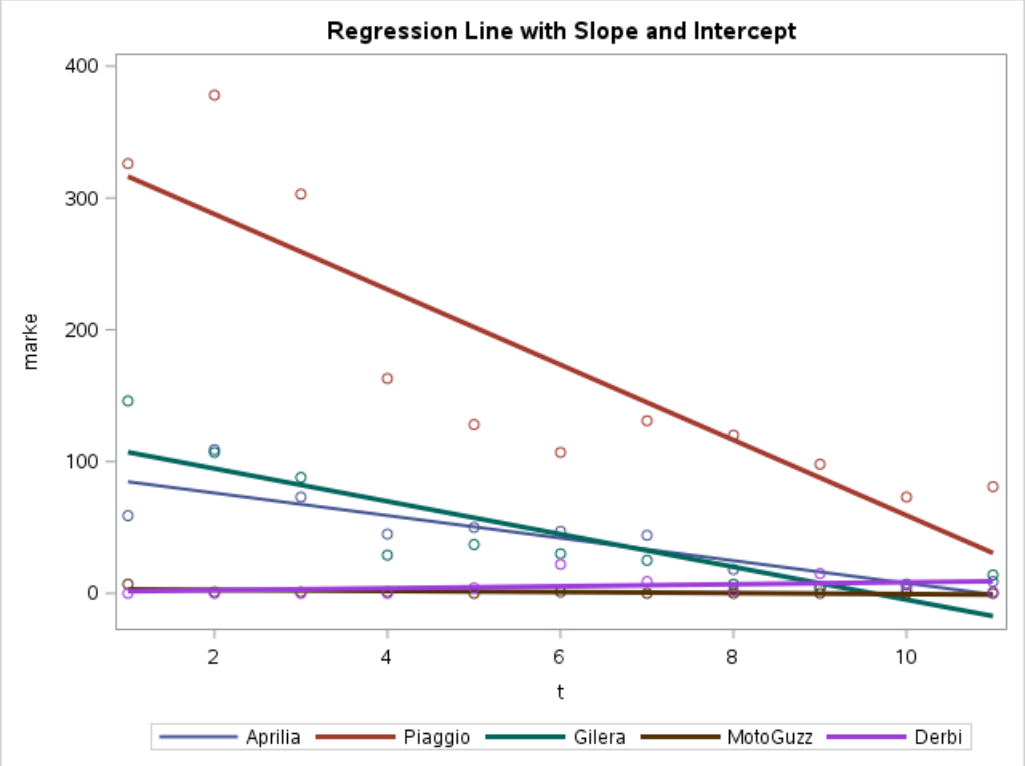
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	62.62727	62.62727	1.20	0.3009
Error	9	467.91818	51.99091		
Corrected Total	10	530.54545			

Root MSE	7.21047	R-Square	0.1180
Dependent Mean	5.36364	Adj R-Sq	0.0200
Coeff Var	134.43253		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.83636	4.66279	0.18	0.8616
t	1	0.75455	0.68749	1.10	0.3009

Očigledno je marka motora Piaggio marka koja ima najveći pad u prodaji, no to je marka koja se najviše prodaje od početka promatranog razdoblja pa smo mogli i očekivati da će ona imati najveći pad. Na Slici 5.3 se jasno vidi da je nagib pravca Piaggio marke motora najveći.
Vidjeli smo da se u promatranom razdoblju prodavao sličan broj marki motora Gilera i Aprilia, no iz rezultata linearne regresije marka motora Gilera ipak ima veći regresijski koeficijent, tj. veći je pad u prodaji marke motora Gilera nego Aprilia. Na Slici 5.4 možemo uočiti presijecanje pravaca navedene dvije marke. Do 2012. godine prodavalo se više Gilera marki motora, ali 2013. godine marka motora Aprilia ju je prestigla. Također možemo uočiti lagani rast prodaje motora marke Derbi.



Slika 5.3: Jednadžbe linearne regresije za sve marke motora (ispis iz SAS-a)

5.2 Predikcija buduće prodaje

Predviđanje pomoću modela jednostavne linearne regresije

Jedan od osnovnih ciljeva regresijske analize je predviđanje. Pod ”prognostičkom” vrijednosti varijable Y na osnovi regresijskog modela podrazumijeva se njezina procijenjena vrijednost za novu (stvarnu ili pretpostavljenu) vrijednost regresorske varijable na osnovi prošlih i sadašnjih informacija ugrađenih u model. Razlikujemo vremenski i prostorni regresijski model, no kako je naš model vremenski proučavat ćemo samo njega. Zadatak predviđanja je doći do što efikasnije prognoze kako bi se mogle računati ”buduće” vrijednosti, te provesti odgovarajući statistički testovi. Naši podaci ”traju” samo 11 godina, što predstavlja problem jer je to razdoblje prekratko za predikciju. No, ipak ćemo pokušati predvidjeti prodaju marke motora Piaggio koja svih godina promatranog razdoblja ima najveći udio na tržištu prodaje motora.

Durbin-Watson test

U modelu linearne regresije pretpostavlja se da su slučajne varijable (greške relacije) ϵ_t međusobno nezavisne i identično distribuirane normalne slučajne varijable s varijancom σ^2 . Ako pretpostavka o nezavisnosti nije ispunjena, javlja se problem autokorelacije (korelacija slučajnih varijabli unutar jednog stohastičkog procesa). Koeficijent autokorelacije prvog reda pokazuje smjer i jakost linearne veze među članovima procesa razmaknutih za jedno vremensko razdoblje, koeficijent autokorelacije drugog reda pokazuje smjer i jakost linearne veze među članovima procesa razmaknutih za dva vremenska razdoblja, itd. Da bi se uočile posljedice autokorelacije, najčešće se pretpostavlja da su greške (ϵ_t) generirane autoregresijskim modelom prvog reda, AR(1). Problem autokorelacije može se uočiti na temelju dijagrama rasipanja ili korelograma rezidualnih odstupanja (grafičkog prikaza autokorelacijeske funkcije).

Durbin-Watsonovim testom ispitujemo postoji li problem autokorelacije. [4]

Testiramo:

H_0 : ne postoji autokorelacija
H_a : postoji autokorelacija

Ako su greške relacije autokorelirane vrijedi: $\epsilon_t = \rho\epsilon_{t-1} + u_t$, gdje su u_t centrirane, identično normalno distribuirane slučajne varijable, koje su međusobno nekorelirane i nekorelirane s varijablama ϵ_t. Stohastički proces $\{u_t : t = ... - 2, -1, 0, 1, 2....\}$ zovemo *bijelim šumom*. Empirijska test veličina za DW-test je:

$$DW = \frac{\sum_{t=2}^{n}(\hat{\epsilon}_t - \hat{\epsilon_{t-1}})^2}{\sum_{t=1}^{n} \hat{\epsilon}_t^{\,2}} \tag{5.1}$$

AUTOREG procedura

AUTOREG procedura u SAS-u služi za procjenu i predikciju linearnih regresijskih modela čije su greške autokorelirane ili heteroshedastične koristeći vremenski trend.
Naredba *DWPROB* bazično ispituje problem autokorelacije prvog reda, a naredbom *DW = option* ispitujemo probleme autokorelacije viših redova.

SAS kod za autokorelaciju s legom 1:

PROC AUTOREG *data=baza;*
MODEL Piaggio=god/dw=1 archtest ***DWPROB****;*
run;

Iz druge tablice, Durbin-Watsonove statistike (Tablica 5.7), vidimo da postoji pozitivna autokorelacija (DW=0,93, p=0,0058). U trećoj tablici procijenjenih parametara (Tablica 5.7) dani su procijenjeni parametri koje smo već dobili kad smo tražili trend prodaje marke motora Piaggio.

Tablica 5.7: Rezultati analize DW testa za marku motora Piaggio (ispis iz SAS-a)

Ordinary Least Squares Estimates			
SSE	27670.6455	DFE	9
MSE	3075	Root MSE	55.44832
SBC	122.144992	AIC	121.349201
MAE	40.3570248	AICC	122.849201
MAPE	28.0052251	HQC	120.847567
Durbin-Watson	0.9262	Total R-Square	0.7645

Durbin-Watson Statistics			
Order	DW	Pr < DW	Pr > DW
1	0.9262	0.0058	0.9942

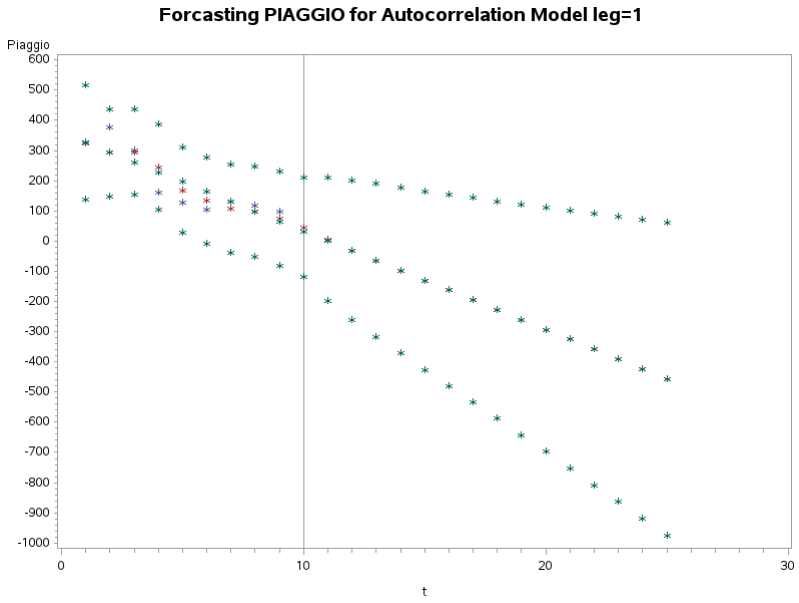
NOTE: Pr<DW is the p-value for testing positive autocorrelation, and Pr>DW is the p-value for testing negative autocorrelation.

Parameter Estimates					
Variable	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	1	344.8909	35.8568	9.62	<.0001
t	1	-28.5727	5.2868	-5.40	0.0004

SAS kod za crtanje predviđenih vrijednosti:

```
title"Forcasting PIAGGIO for Autocorrelation Model leg=1";
PROC GPLOT data=p;
PLOT piaggio*t=1
yhat*t=2
piaggiotrend*t=3
lcl*t=3 ucl*t=3
/overlay href=10;
where t >= 1;
symbol1 v=star i=none;
symbol2 v=circle i=join;
symbol3 v=none i=join;
run;
```

Na Slici 5.4 se vidi grafički prikaz predikcije broja prodanih motora marke Piaggio sa legom 1.



Slika 5.4: Grafički prikaz predikcije broja prodanih Piaggio motora sa legom 1 (ispis iz SAS-a)

Nakon što smo provjerili autokorelaciju s legom 1 za marku motora Piaggio, Durbin-Watsonovim testom provjerimo postoji li autokorelacija s legom 2.

SAS kod za autokorelaciju s legom 2:

PROC AUTOREG *data=forcast;*
MODEL Piaggio=t/nlag=2 ***DW=2 DWPROB*** *method=ml;*
output out=p p=yhat pm=piaggiotrend lcl=lcl ucl=ucl;
run;

Iz druge tablice Tablice 5.8 se vidi da p-vrijednost za leg 2 iznosi 0,64, što je veće od razine značajnosti, tj. leg 2 nije statistički značajan. Dakle, ostajemo kod modela s legom 1 i prema njemu je predikcija Slika 5.5.

Tablica 5.8: Rezultati analize DW testa za marku motora Piaggio (ispis iz SAS-a)

The AUTOREG Procedure

Ordinary Least Squares Estimates			
SSE	22177.6222	DFE	7
MSE	3168	Root MSE	56.28705
SBC	100.221873	AIC	99.8274239
MAE	43.2493827	AICC	101.827424
MAPE	27.8527999	HQC	98.9762039
		Total R-Square	0.7658

Durbin-Watson Statistics			
Order	DW	Pr < DW	Pr > DW
1	1.1089	0.0192	0.9808
2	2.1661	0.6359	0.3641

NOTE: Pr<DW is the p-value for testing positive autocorrelation, and Pr>DW is the p-value for testing negative autocorrelatio

Parameter Estimates					
Variable	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	1	368.7222	40.8916	9.02	<.0001
t	1	-34.7667	7.2666	-4.78	0.0020

Bibliografija

[1] S.E. Clausen, *Applied Correspondence analysis*, Sage University Papers Series on Quantitative Applications in the Social Sciences, Thousand Oaks 1988.

[2] J. Lawson, *Design and Analysis of Experiments with SAS*, Chapman and Hall/CRC Press, 2010.

[3] M. Greenacre, *Correspondence analysis in practice*, Chapman and Hall/CRC

[4] V. Bahovec, N. Erjavec, *Uvod u ekonometrijsku analizu*, 1. izdanje, Zagreb 2009.

[5] A. Jazbec, Materijali s predavanja iz Odabranih stat. metoda u biomedicini, Zagreb, 2015./2016.

[6] Materijali s predavanja iz Statistike, Zagreb, 2014./2015.

[7] Materijali s predavanja iz Statistickog Praktikuma, Zagreb, 2015./2016.

[8] `https://www.sas.com/en_us/home.html`.

[9] `https://support.sas.com/software/products/ondemand-academics/#s1=2`.

Sažetak

U ovom radu opisana je analiza prodaje motora firme iz Šibenika koja se uspjela održati za vrijeme ”prve globalne financijske krize 21. stoljeća”, kad su mnogi prodavatelji motora u RH ”propali”. Podaci s kojima se radi sežu od 2006. do 2016. godine te zapravo opisuju veliki pad u prodaji. Poznate su jačine, boje, marke i tipovi prodanih motora u navedenom razdoblju.
Rad se sastoji od 5 poglavlja u kojima su obrađene različite statističke metode. Prvo se opisnom statistikom opisala baza podataka, navedene su tablice frekvencija i mozaik prikazi varijabli baze. Potom se uradio χ^2-test, jedan od prvih statističkih testova. Glavna tema trećeg poglavlja je analiza varijance. Nadalje, opisana je novija, jako fleksibilna statistička metoda, analiza korespondencije. U posljednjem poglavlju je obrađena linearna regresija kojom se izračunalo koliki pad u prodaji se dogodio svakoj od marki motora. Također, pokušala se predvidjeti daljna prodaja.
Prilikom obrade uzetih podataka korišten je programski sustav SAS, te su navedeni potrebni kodovi. Iako se radi o velikom padu u prodaji motora, iz tablica frekvencija se može vidjeti lagani porast u prodaji prošle godine (2016.), a djelatnici firme kažu da je porast još i veći ove godine.

Summary

This paper describes motor sales analysis of a company from Šibenik which survived during ”the first global financial crisis of the 21st century”, when most of Croatian motor dealers collapsed. The data extend from 2006 to 2016 and actually, sales come down is showed. Strengths, colors, brands and types of sold motors in the specified period are known.
The paper consists of five chapters which show different statistic methods. First, the data base is described with descriptive statistics, and also frequency tables and mosaic plots of the base variables are showed. Secondly, χ^2-test, one of the first statistic tests is done. The main theme of the third chapter was variance analysis. Then is described the newer, very flexible statistic method, correspondence analysis. The last chapter describes linear regression, which is used to calculate the exact sales come down for all of the motor brands. Also, we tried to make a prediction of the future sales.
For all data analysis program system SAS is used and there are all required codes. Although the paper is about the big sales collapse, from the frequency table we can see light increase in sales from the last year (2016), and the company employees state that sale increase is even greater this year.

Životopis

Rođena sam 8. veljače 1993. godine u Šibeniku. Nakon završene osnovne škole Fausta Vrančića u Šibeniku, u istom gradu upisujem Gimnaziju Antuna Vrančića, opći smjer. Maturirala sam 2011. godine te iste godine upisujem sveučilišni preddiplomski studij Matematika na Prirodoslovno-matematičkom fakultetu u Zagrebu. Nakon završenog preddiplomskog studija, 2015. godine upisujem diplomski sveučilišni studij Matematička statistika na istom odsjeku.