Илустрација истраживања о аутомобилима

Павле Вилотијевић, 78/2021

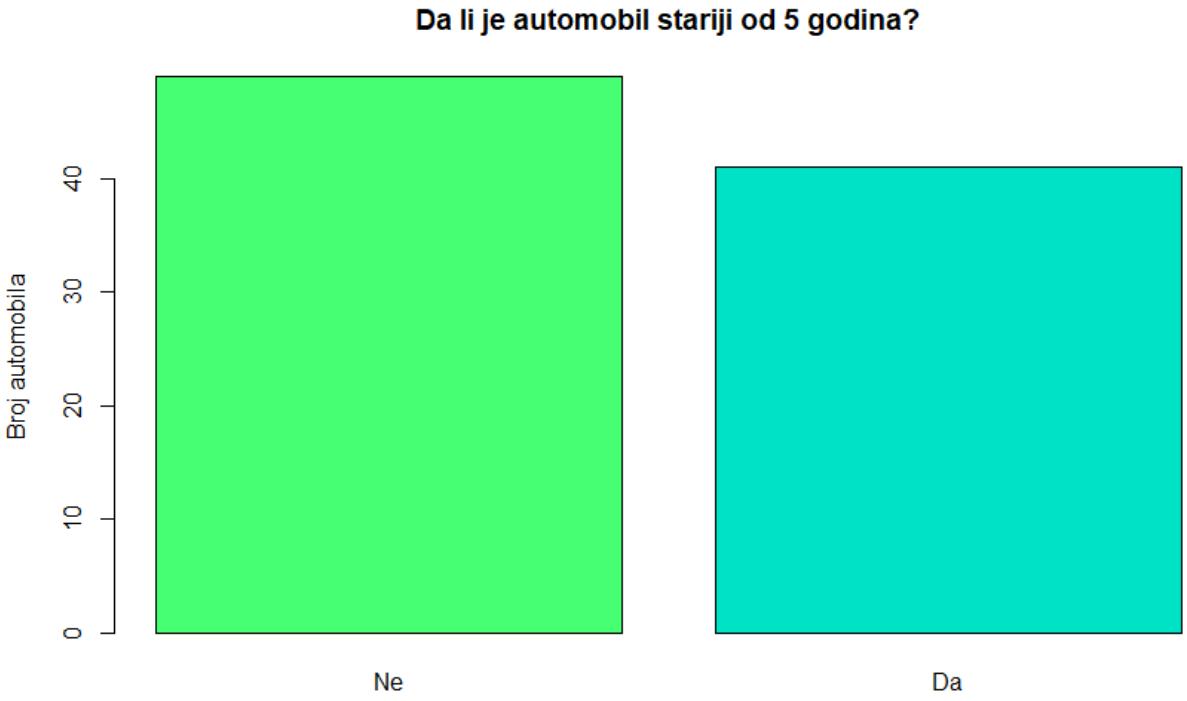
1. Опис базе података

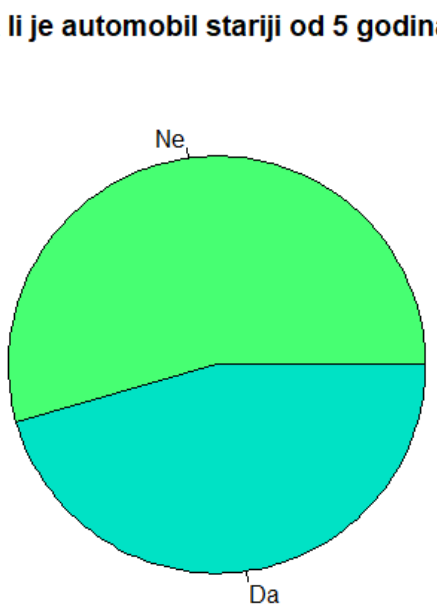
НАПОМЕНА: уколико није другачије наглашено, приликом доношења закључака за статистичке тестове за ниво значајности узимана је вредност 0.05,

У питању су подаци о ценама, старостима, моделима и пређеном путу аутомобила, које су скупљала 2 студента 2007 године, са сајта „autotrader.com“. База садржи 90 опсервација са 4 променљиве: price – нумеричка променљива која означава цену аутомобила, изражена у хиљадама долара; age – категоричка променљива која означава да ли је аутомобил старији од 5 година (0 – мање или једнако 5 година, 1 – више од 5 година); mileage – нумеричка променљива која означава пређени пут аутомобила, изражен у хиљадама миља; car – категоричка променљива којим је изражен произвођач аутомобила (0 – Porsche, 1 – Jaguar, 2 – BMW).

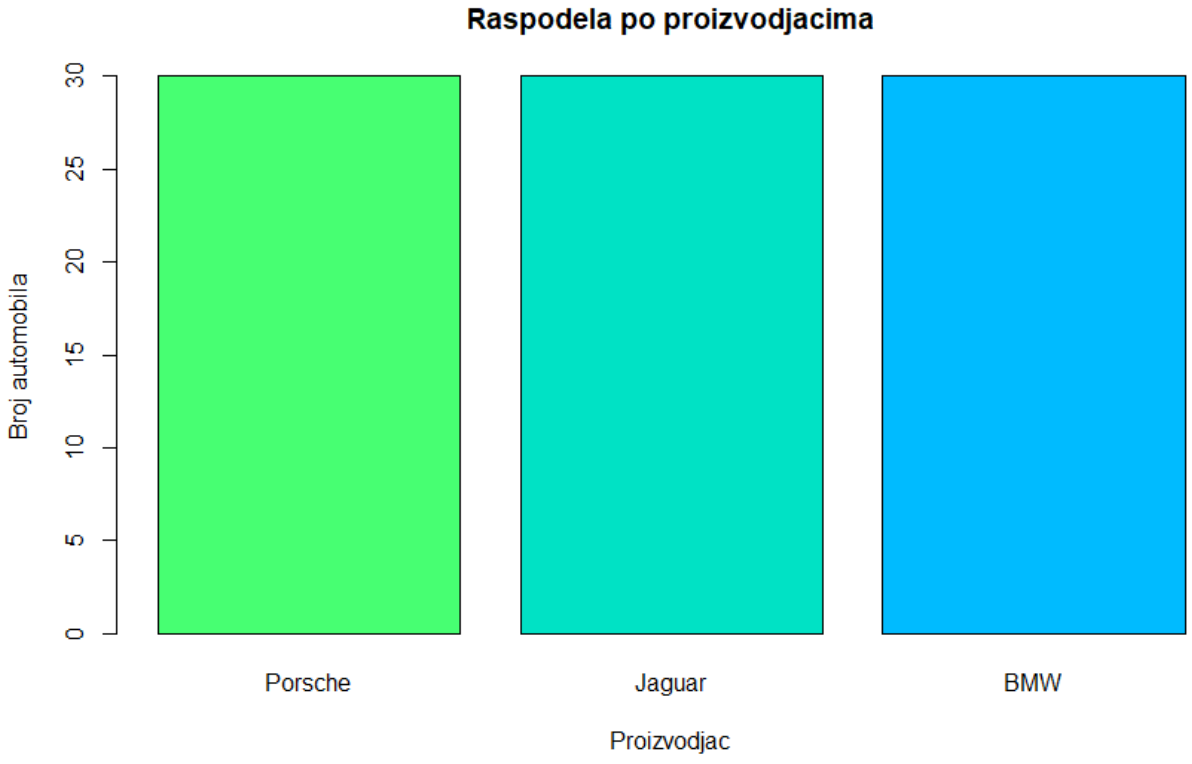
2. Категоричке променљиве

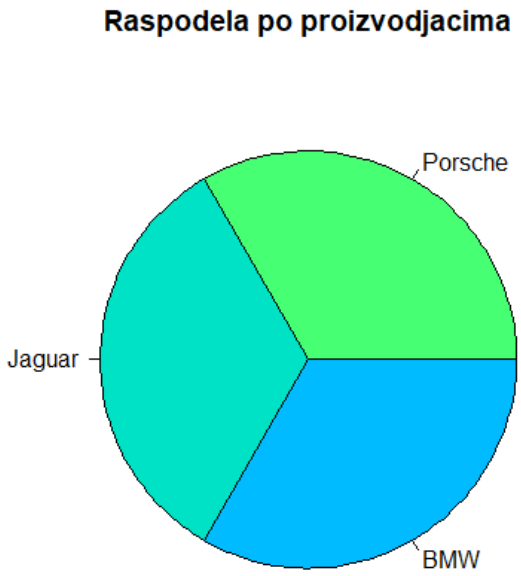
База података садржи 2 категоричке променљиве – age и car, које су у делу изнад описане. Битно је напоменути да је age оригинално била дискретна нумеричка величина, али сам поделио податке према томе да ли је аутомобил старији од 5 година или није, како бих добио категоричку величину. Разлог зашто сам одабрао баш 5 година је зато што је то била медијана ове величине. Погледајмо њихове графичке приказе:



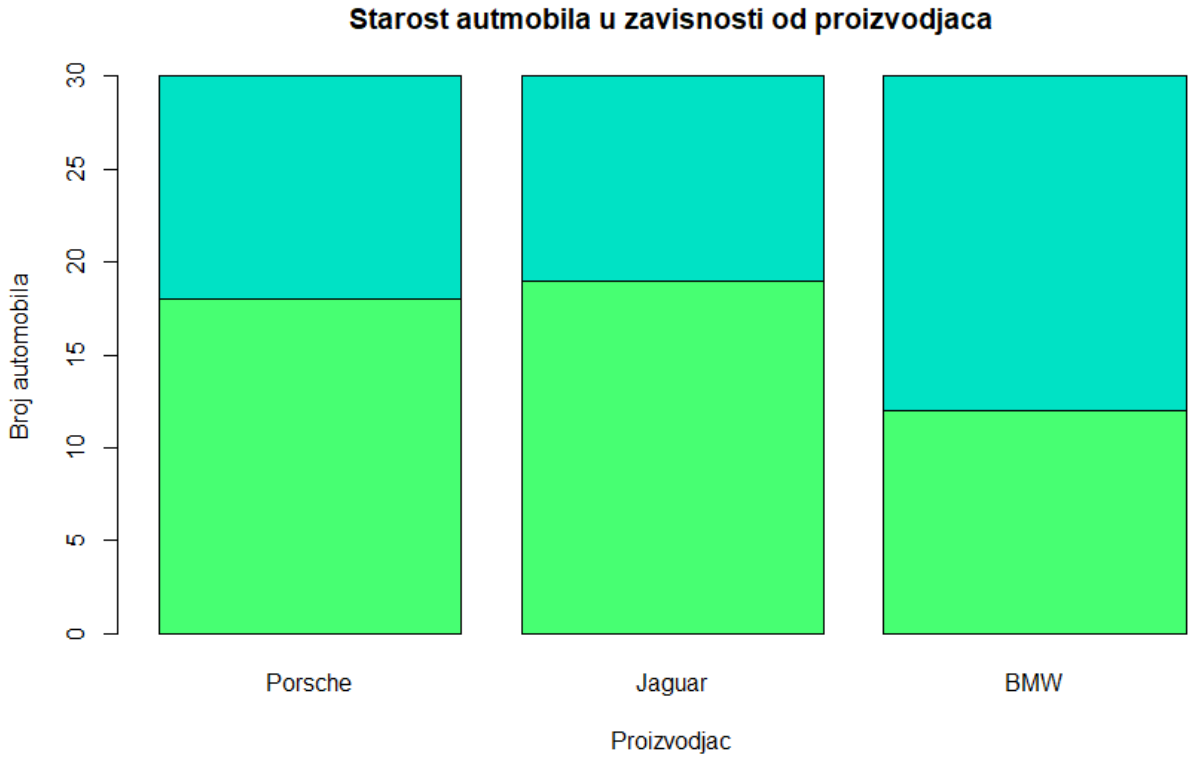


Можемо приметити да је нешто мање од пола аутомобила старије од 5 година, али да су бројеви прилично блиски, што је било очекивано узевши у обзир начин на који је ова категоричка променљива „направљена“. Посматрајмо сада графичке приказе променљиве car:





На основу ових слика јасно се види да су аутомобили једнако распоређени по произвођачима – по 30 у свакој категорији. Посматрајмо сада график зависности ове две променљиве:



На основу графика изгледа као да су BMW аутомобили нешто старији, док код Porsche и Jaguar већи удео чине млађи. Поставља се питање - да ли између ових података заиста постоји зависност, или су они независни? Спровођењем хи квадрат теста (нулта хипотеза- подаци су независни, алтернативна хипотеза- подаци нису независни), добија се p-вредност

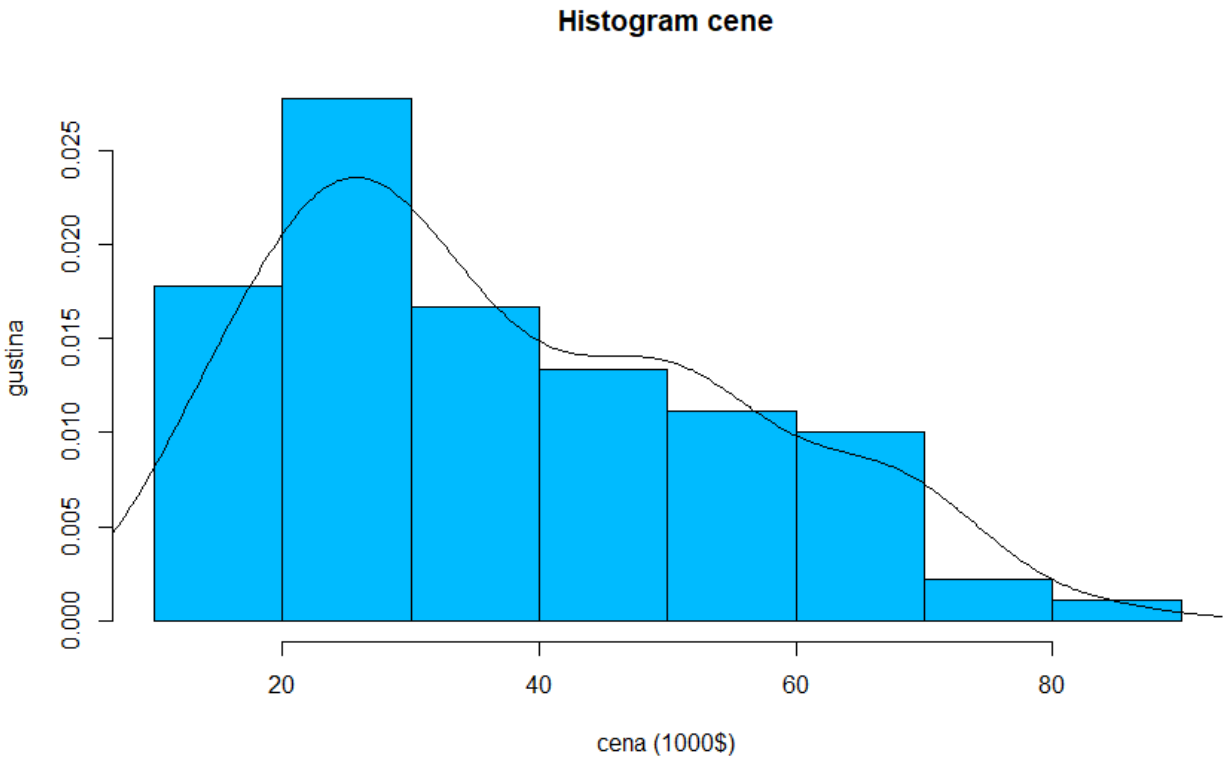
p=0.1457, што је релативно велика вредност, па како не одбацујемо нулту хипотезу, (на основу ових података) можемо закључити да старост аутомобила не зависи од произвођача.

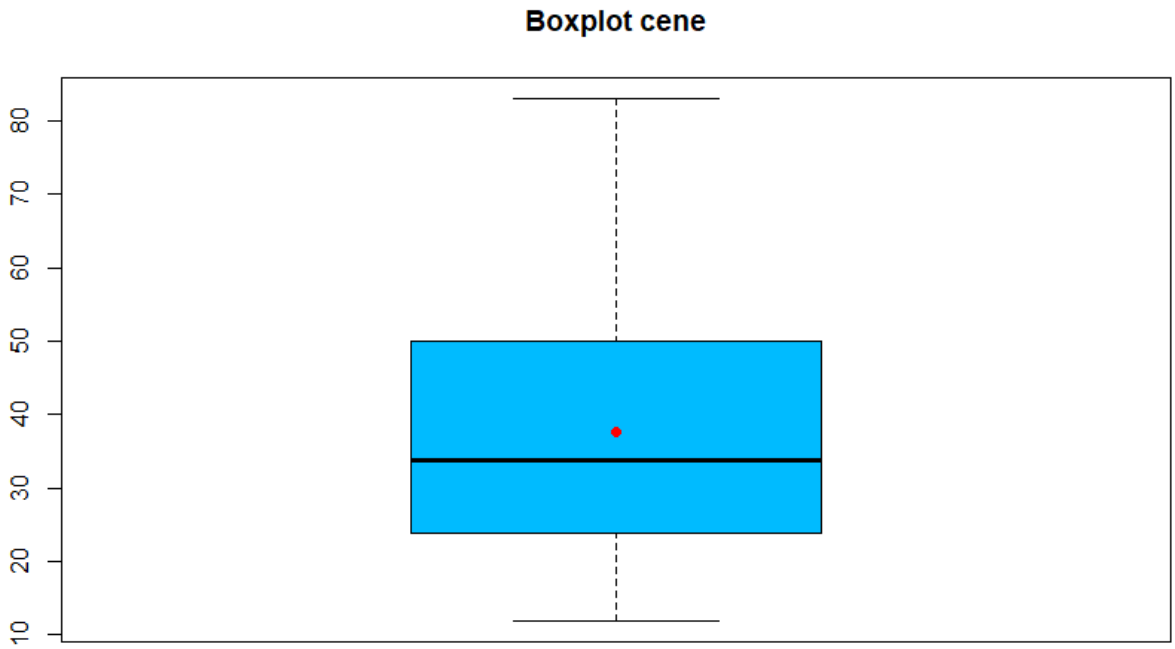
3. Нумеричке променљиве

Нумеричке променљиве у овој бази података су price и mileage. Њихове основне статистике су следеће:

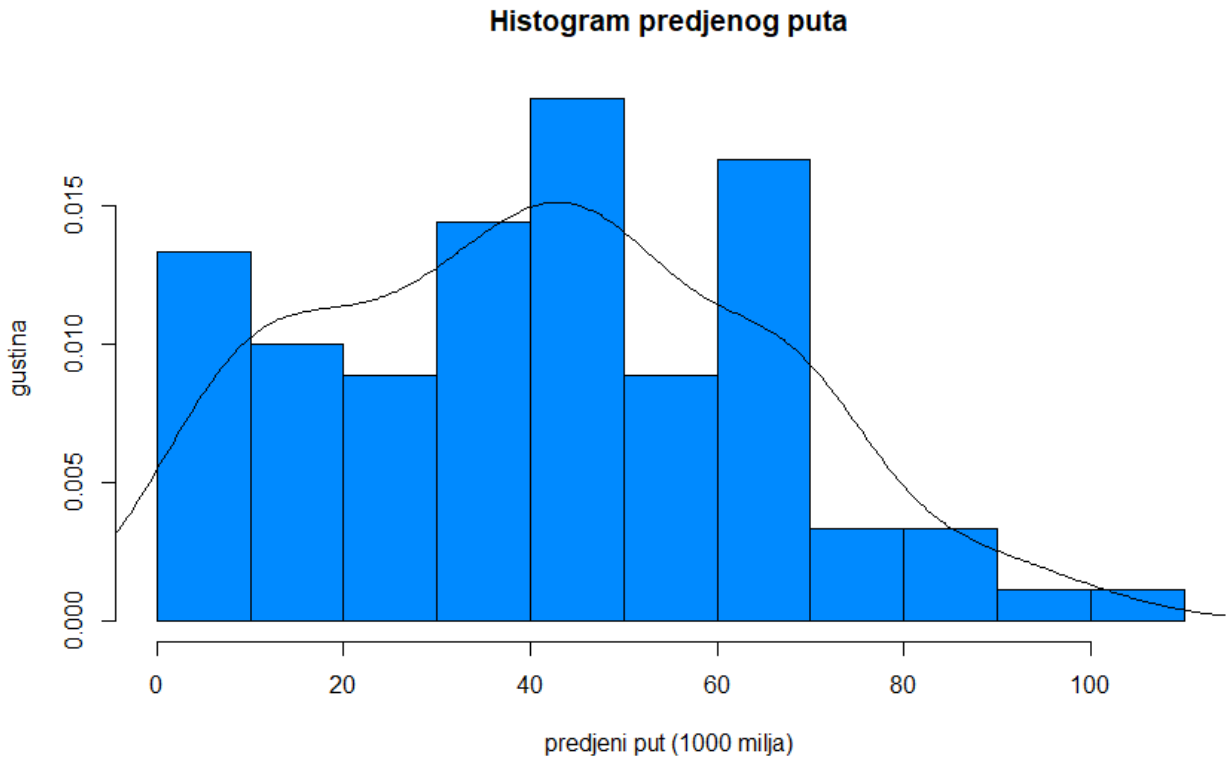
	min	max	mean	median	sd
price	12	83	37.58	33.7	17.64
mileage	0.67	100.7	41.32	42.85	23.52

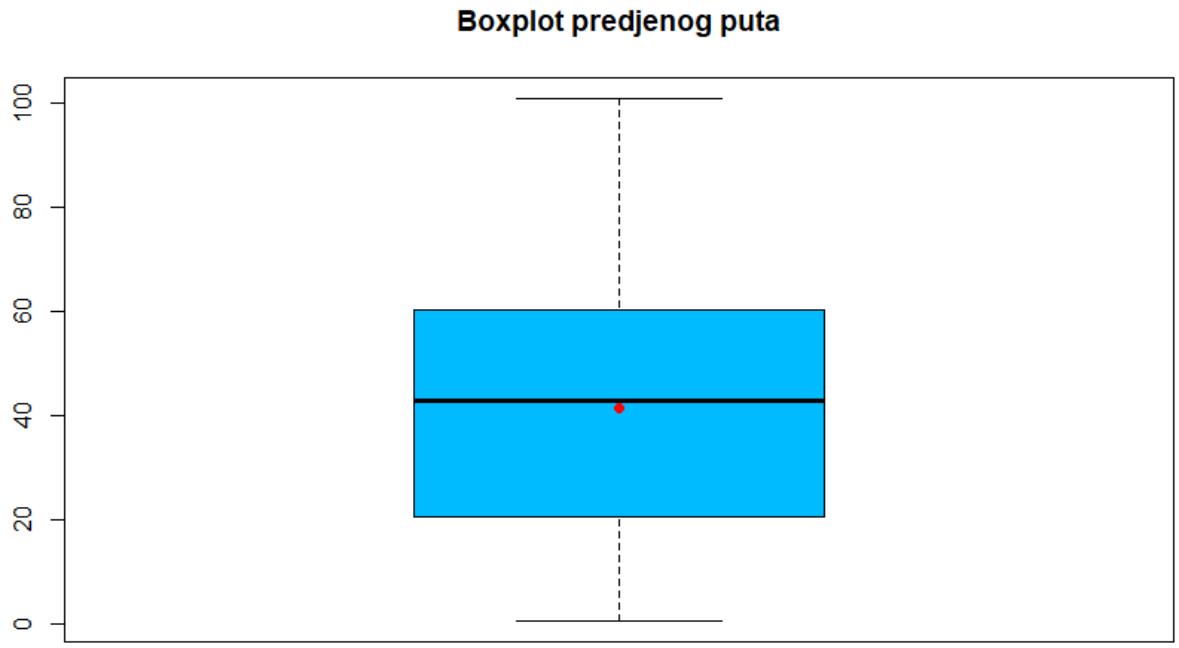
Графички приказ променљиве price:



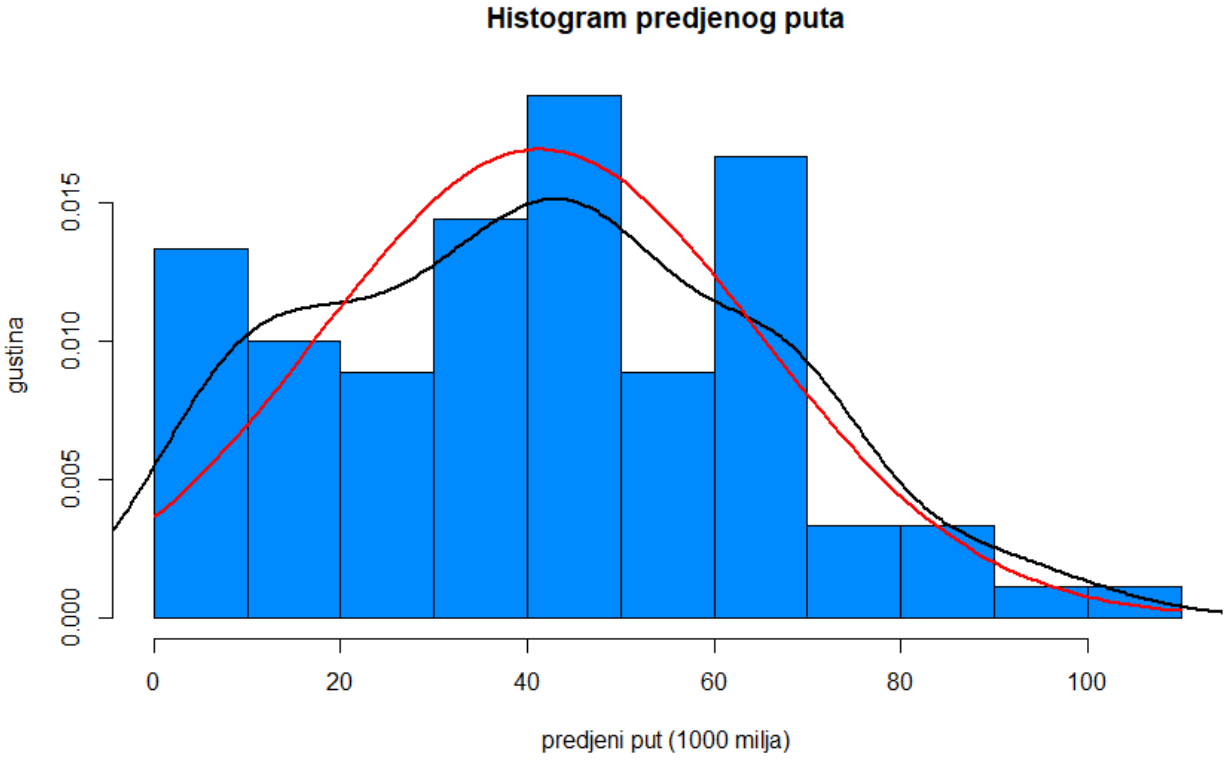


Графички приказ промењиве mileage:





На основу графичких приказа можемо видети да су обе расподеле асиметричне, додуше нешто мање код променљиве mileage. На boxplot-у те променљиве можемо видети да су медијана и узорачка средина веома близу, па нас то може упутити да проверимо припадност нормалној расподели N(41.32, 23.52):



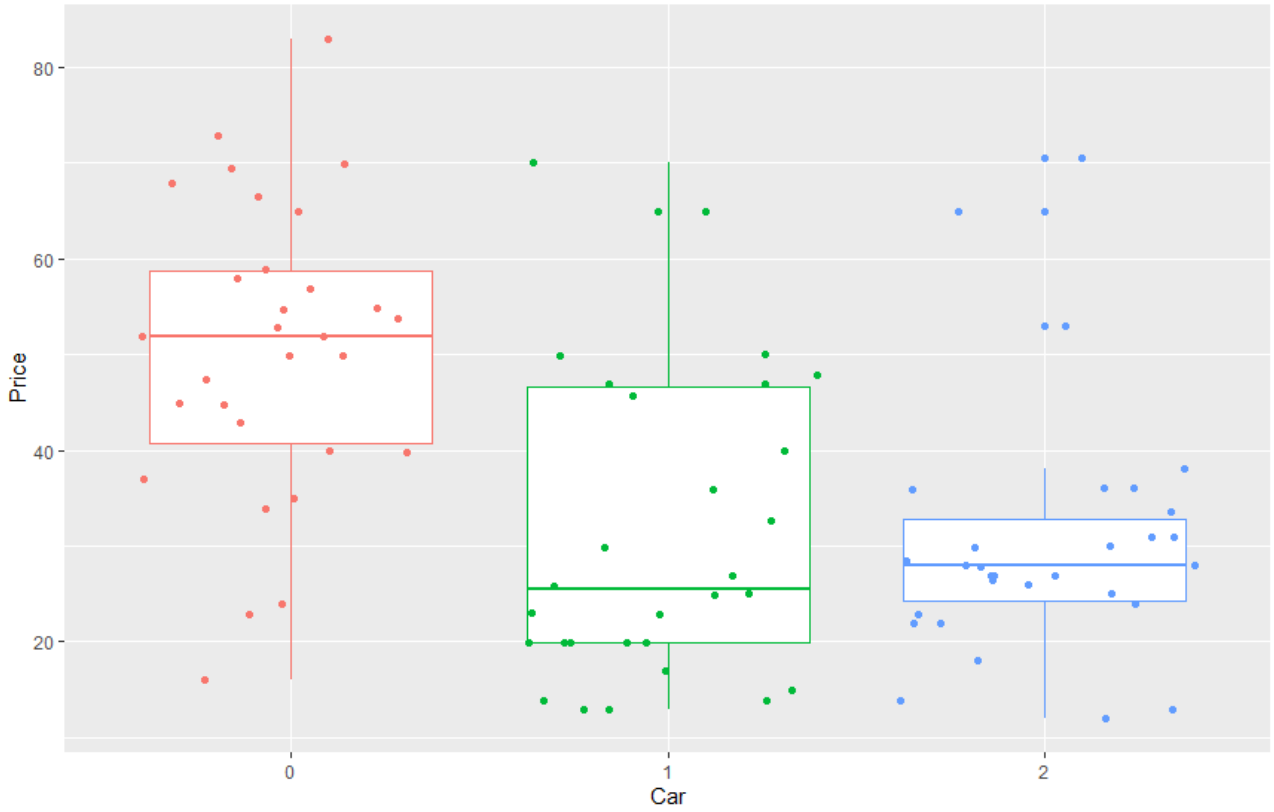
Додавањем графика функције густине нормалне расподеле на хистограм, можемо видети да је он (релативно) близу емпиријске фунцкије густине ове случајне величине. Ово можемо тестирати KC тестом (нулта хипотеза – случајна величина има дату нормалну расподелу, алтернативна – случајна велилина нема дату нормалну расподелу). Када га применимо, добијамо p-вредност 0.8055, па како је она велика нећемо одбацити нулту хипотезу. Закључак је да ова случајна величина има нормалну расподелу N(41.32, 23.52). Уколико исти тест спроведемо за променљиву price добијамо p-вредност 0.1189, што је и даље велико али ни близу колико за mileage, па (са нешто већим устручавањем) можемо закључити да и она има приближно нормалну расподелу (у овом случају N(37.58,17.64)).

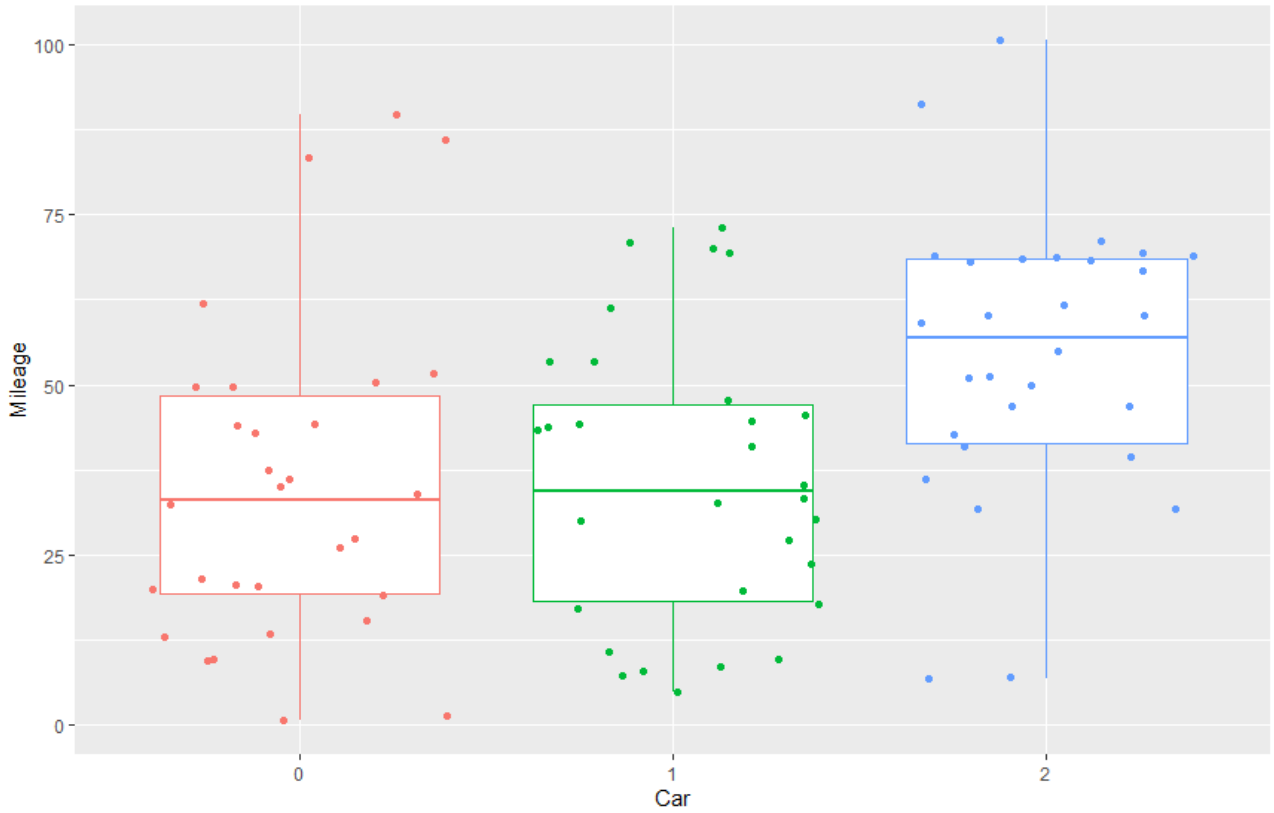
Податке можемо поделити према томе да ли су аутомобили старији од 5 година или нису. Урадимо у том случају тест једнакости средњих вредности за обележја price и mileage (нулта хипотеза – средње вредности су једнаке, алтернативна хипотеза – средње вредности нису једнаке). У оба случаја добијамо p-вредност која је практично 0 (4.331e-10 и 3.991e-13), па закључујемо да се средње вредности разликују. Логично је да је цена нових аутомобила већа, а пређени пут мањи, па можемо урадити исте тестове са измењеним алтернативним хипотезама како бисмо видели да ли је ово случај. За price (нулта хипотеза – средње вредности су једнаке, алтернативна хипотеза – средња вредност млађих аутомобила већа је од средње вредности старијих) добијамо p-вредност 2.166e-10, а за mileage (нулта хипотеза – средње вредности су једнаке, алтернативна хипотеза – средња вредност млађих аутомобила мања је од средње вредности старијих) p-вредност 1.995e-13, па је закључак да су новији аутомобили заиста скупљи и да им је укупан пређени пут мањи.

4. Имплементација ANOVA теста

Напомена: када би ово било право истраживање, неопходно би било да будемо доста стриктнији при оправдавању коришћења овог теста, али како је у питању илустрација на неким местима ћу поједностављивати испитивања претпоставки неопходних да би се тест применио.

ANOVA (ANalisys Of VAriance) тест је статистички тест који се користи када наше податке можемо поделити на k група (углавном $k \geq 3$, за 2 групе се користи t-тест) и желимо да испитамо нулту хипотезу да су средње вредности између свих група неког нумеричког обележја(у нашем случају mileage, из разлога који ће бити објашњени касније) једнаке ($m_1 = m_2 = \ldots = m_k$), против алтернативне да нису. Како се ови подаци могу поделити на 3 групе (према произвођачу аутомобила), ово је тест који сам се одлучио да имплементирам у R-у. Да бисмо применили овај тест, неопходно нам је пар претпоставки. Прво, подаци су приближно нормално расподељени – већ смо тестирали ову хипотезу у претходном делу, додуше тада смо је тестирали када нису били подељени у групе. Али, како је узорак релативно велик (свака група има 30 опсервација), то је довољно да можемо сматрати да је овај услов задовољен. Следеће је да су подаци независни, што смо претпостављали и за претходне тестове. Последњи услов је једнакост дисперзија – све групе би требало да имају једнаке дисперзије (напомена: постоји и верзија овог теста која не претпоставља овај услов, тзв. Welch ANOVA). Овај услов је најједноставније проверити графички: направимо boxplot дијаграме за све 3 групе и видимо да ли је варијација између њих слична.





На првој слици налазе се boxplot-ови цена, а на другој пређеног пута. Изгледа да су дисперзије пређеног пута макар приближно једнаке, па ћемо стога применити тест на нумеричко обележје mileage. У R-у овај тест се врши позивом функције aov(), на чији излаз позивамо summary() како бисмо видели резултате. Позив функције је облика aov(formula, data), где је formula формула која означава модел који испитујемо (у нашем случају Mileage ~

Car), a data je data frame у којем се налазе подаци (R може и аутоматски да их нађе уколико је data изостављен из улаза, у нашем случају ово је сама база података – ThreeCars). Као излаз добијамо објекат класе aov, на који можемо применити метод summary како бисмо добили податке који су нам потребни при доношењу закључака. Саму тест статистику најлакше је описати „успут“ то јест описаћу је док описујем имплементацију теста „пешке“, а њена расподела под нултом хипотезом биће Фишерова F(k-1, n-k) расподела, где је са k означен број група (код нас k=3), а са n број опсервација (код нас n=90). Пређимо сада на имплементацију у R-u:

Пре него што имплементирамо овај тест у R-у, згодно је представити све податке који ће нам бити потребни. Често се користи оваква (или њој слична) табела:

	Degrees of freedom	Sum of squares	Mean square	F (тест статистика)
Treatments	k-1	SST	SST/(k-1)	MST/MSE
Error	n-k	SSE	SSE/(n-k)	/
Total	n-1	SS(Total)	/	/

Степене слободе лако израчунавамо. SS(Total) је збир квадратних одступања целог узорка, а $SST = \sum_{i=1}^{k} n_i(\overline{X}_i - \overline{X})^2$ се односи на варијабилитет између група, док се $SSE = \sum_{i=1}^{k} (n_i - 1)s_i^2$ односи на варијабилитет унутар група. Такође, знамо да важи SS(Total) = SST + SSE. Средњеквадратно одступање добијамо када збир квадрата поделимо степенима слободе, а наша тест статистика је управо количник два добијена средњеквадратна одступања, и под нултом хипотезом имаће Фишерову F(k-1, n-k) расподелу. Када спроведемо сав овај рачун у R-у, добијамо F=6.4485, док је вредност коју функција aov враћа 6.448, што значи да је тест добро имплементиран. p-вредност је 0.0024656, док је при позиву функције aov она 0.00245 (претпостављам да до ових разлика долази услед заокруживања). У сваком случају, како је p-вредност мала, долазимо до закључка да постоји разлика између средњих вредности међу групама, а на основу boxplot-а од малочас можемо претпоставити да BMW аутомобили у просеку пређу већи пут (ако бисмо хтели то да докажемо један од начина био би да урадимо t-тест на Porsche и Jaguar аутомобилима, видимо да је ту средња вредност иста, а потом опет применимо t-тест на BMW и једног од осталих произвођача и видимо да је средња вредност код BMW-а већа).

5. Закључак

На примеру базе података о ценама, пређеном путу, врстама и старости аутомобила видели смо како се нумерички и категорички подаци могу графички представити, као и како се различити статистички тестови на њима могу применити и имплементирати у циљу доношења закључака.