به نام خدا

فاز اول پروژه

امیرمحمد رنجبر پازکی ۸۱۰۱۹۵۴۰۲

دکتر بهرک
استباط آماری

دانشکده مهندسی برق و کامپیوتر دانشگاه تهران

بهار ۹۹

سوال ۰

a. مجموعه داده مورد بررسی مجموعه used car است که داده‌های مربوط به ۳۸۵۰۶ ماشین کارکرده است که برای فروش گذاشته‌شده‌اند. این اطلاعات شامل قیمت و اطلاعات فنی ماشین‌هاست. بررسی این مجموعه داده می‌تواند نتایج جالبی داشته باشد. از جمله آن‌ها می‌توان به ساخت تخمینگر قیمت ماشین‌های دست دوم با استفاده از یادگیری ماشین اشاره کرد. این بررسی می‌تواند ویژگی‌های موثر در قیمت ماشین را به ما نشان‌دهد و علاوه بر دادن دید در این زمینه، ویژگی‌های مناسب مدل‌کردن این مسئله را آشکار کند.

b. این مجموعه داده شامل ۳۸۵۰۶ داده و شامل ۱۸ متغیر(ستون/ ویژگی) است که اطلاعات هرکدام برای هر داده جمع‌آوری شده‌است.

```
print(paste("Number of records: ", nrow(usedCar)))
print(paste("Number of features: ", ncol(usedCar)))
```

c. با استفاده از دستور زیر تعداد مقادیر گمشده هر ویژگی را می‌توان به‌دست آورد. تکه کد و خروجی در پایین آمده‌است.

```
print(sapply(usedCar, function(x) sum(is.na(x))))
```

```
     rownum       price acquisition_date       badge   body_type
          0           3                0           0           0
   category      colour        cylinders     economy        fuel
          0           0             2488        3920           0
last_updated      litres         location        make       model
          0        2488                0           0           0
   odometer transmission             year
       1550           0                0
```

همان‌طور که مشاهده می‌شود، ویژگی‌های price، cylinders، economy، liters و odometer دارای داده‌های گمشده هستند.

Price یکی از مهمترین ویژگی‌های این مجموعه داده‌ست و تعداد داده‌های گمشده آن بسیار کم است. اگر آن مقادیر را پر کنیم، ممکن است تاثیرگذار در بررسی‌ها باشد. به این دو دلیل، این داده‌ها را حذف می‌کنیم.

```
usedCars <- usedCar[!is.na(usedCar$price),]
```

در بقیه ویژگی‌ها، تعداد داده‌های دارای مقادیر گمشده قابل توجه هستند. به همین دلیل، این داده‌ها را حذف نمی‌کنیم و با مقادیر میانی جایگزین می‌کنیم. برای مقادیر عددی با میانگین و برای مقادیر کیفی با میانه مقادیر گمشده را پر می‌کنیم. (البته می‌توانیم از regression برای پر کردن این مقادیر استفاده کنیم.)

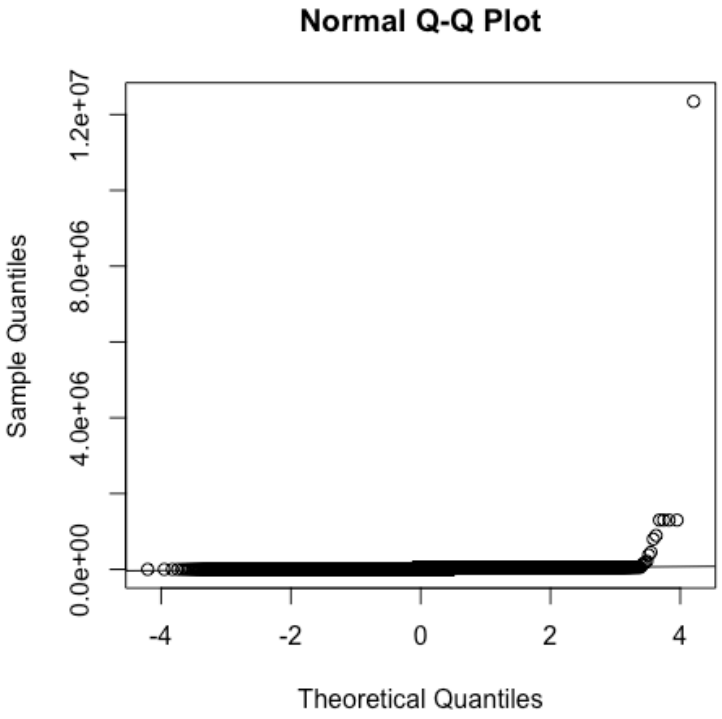
مقادیر عددی ویژگی‌های economy و odometer هستند. ویژگی‌های دسته‌بندی‌شده ویژگی‌های cylinders و liters هستند.

```
usedCars$economy <- ifelse(is.na(usedCars$economy), mean(usedCars$economy, na.rm=TRUE), usedCars$economy)
usedCars$odometer <- ifelse(is.na(usedCars$odometer), mean(usedCars$odometer, na.rm=TRUE), usedCars$odometer)
usedCars$litres <- ifelse(is.na(usedCars$litres), median(usedCars$litres, na.rm=TRUE), usedCars$litres)
usedCars$cylinders <- ifelse(is.na(usedCars$cylinders), median(usedCars$cylinders, na.rm=TRUE), usedCars$cylinders)
```

d. از میان ویژگی‌هـــا، ویژگی‌هـــای price ، colour، make، model، year ، odometer و transmission دارای اطلاعات مهمی هستند. Price ویژگی اصلی ماشین‌هاست که مورد توجه قرار می‌گیرد. make، model و transmission و year بیانگر ویژگی‌های ابتدایی ماشین‌ها هستند که وجه تمایز اولیه آن‌ها با هم هستند. Odometer برای ماشین‌های کارکرده تعریف می‌شود و بین ماشین‌های یک مدل تمایز ایجاد می‌کند و بـه همین دلیل، اطـلاعـات ارزشـمندی در آن وجـود دارد. همچنین، ویژگی rownum ویژگی کارآمدی نیست. به همین دلیل، از مجموعه داده آن را حذف می‌کنیم.

سوال ۱. ویژگی عددی انتخاب شده قیمت(price) است.

a. در این نمودار میبینیم، که داده‌های پایین روی یک خط قرار دارند و توزیع نرمال پیروی می‌کنند اما چند داده خیلی زیاد این نمودار را تا این حد فشرده کرده‌اند.

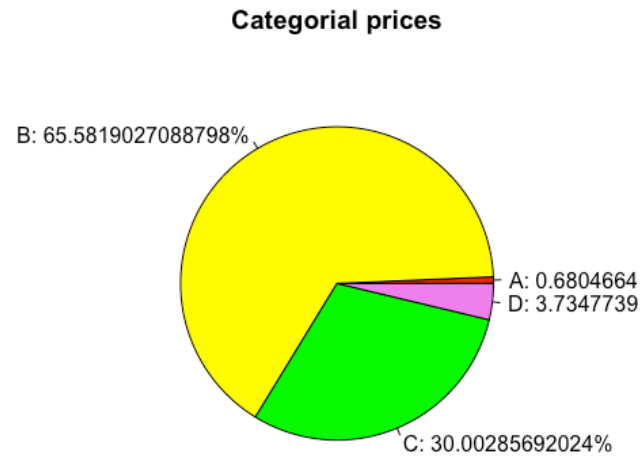



```
qqnorm(usedCars$price)
qqline(usedCars$price)
```

b. برای این منظور یک سطر categorial_price به مجموعه داده خود اضافه می‌کنیم. دسته‌بندی این قیمت‌ها به صورت زیر است:

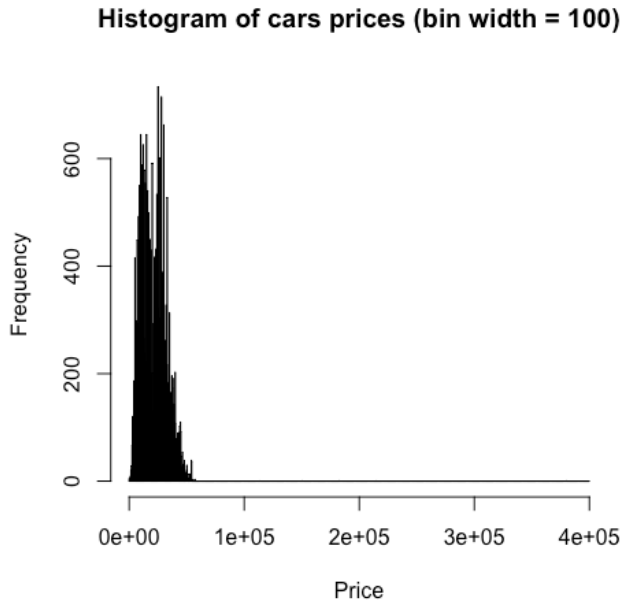
Range	< 5000	5000 - 15000	15000 - 50000	50000 <=
Class	D	C	B	A

نمودار این دسته‌بندی قیمت‌ها به شکل زیر است. همچنین، کد نیز در ادامه آمده است.

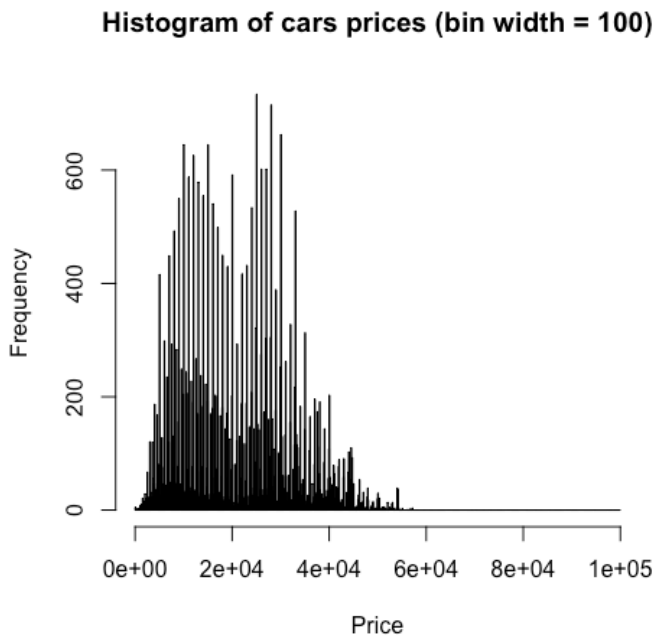



```r
usedCars$categorial_price <- NA
usedCars$categorial_price[usedCars$price < 50000] <- "B"
usedCars$categorial_price[usedCars$price < 15000] <- "C"
usedCars$categorial_price[usedCars$price < 5000] <- "D"
usedCars$categorial_price[usedCars$price >= 50000] = "A"
colors = c("red", "yellow", "green", "violet")
classes = c("A: ", "B: ", "C: ", "D: ")
percentages = paste(prop.table(table(usedCars$categorial_price))*100, "%", sep="")
labels = paste0(classes, percentages)
pie(table(usedCars$categorial_price), col = colors, labels = labels, main = "Categorial prices")
```

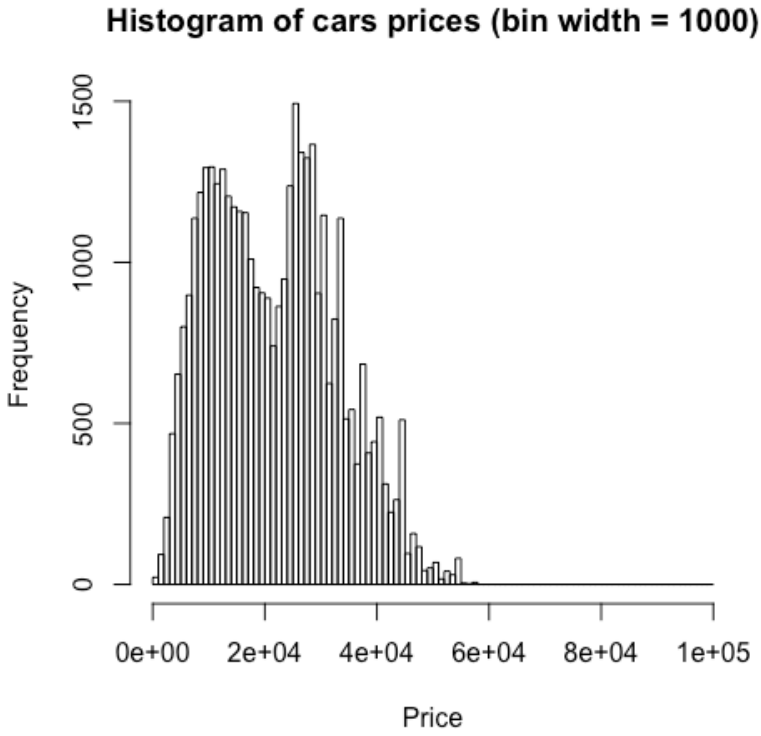
c. ابتدا سعی شد بر روی تمامی داده‌ها histogram رسم شود ولی به دلیل، داده‌های خیلی بزرگ این کار موفق نبود. به همین دلیل، داده‌های زیر ۴۰۰۰۰۰ رسم شده‌اند چراکه تعداد انگشت‌شماری داده بالای این مقدار وجود داشت. نمودار به شکل زیر درآمد.



همانطور که در این نمودار دیده‌می‌شود، سه چهارم نمودار خالی است. به همین دلیل، داده‌های زیر ۱۰۰ هزار تنها نمایش می‌دهیم چراکه تعداد داده‌های بالای چهارصدهزار بسیار کم است.



اندازه binها کوچک هستند و به همین دلیل، نمودار به این شکل درآمده‌است. به همین‌دلیل، سایز bin هزار قرار داده‌شد.

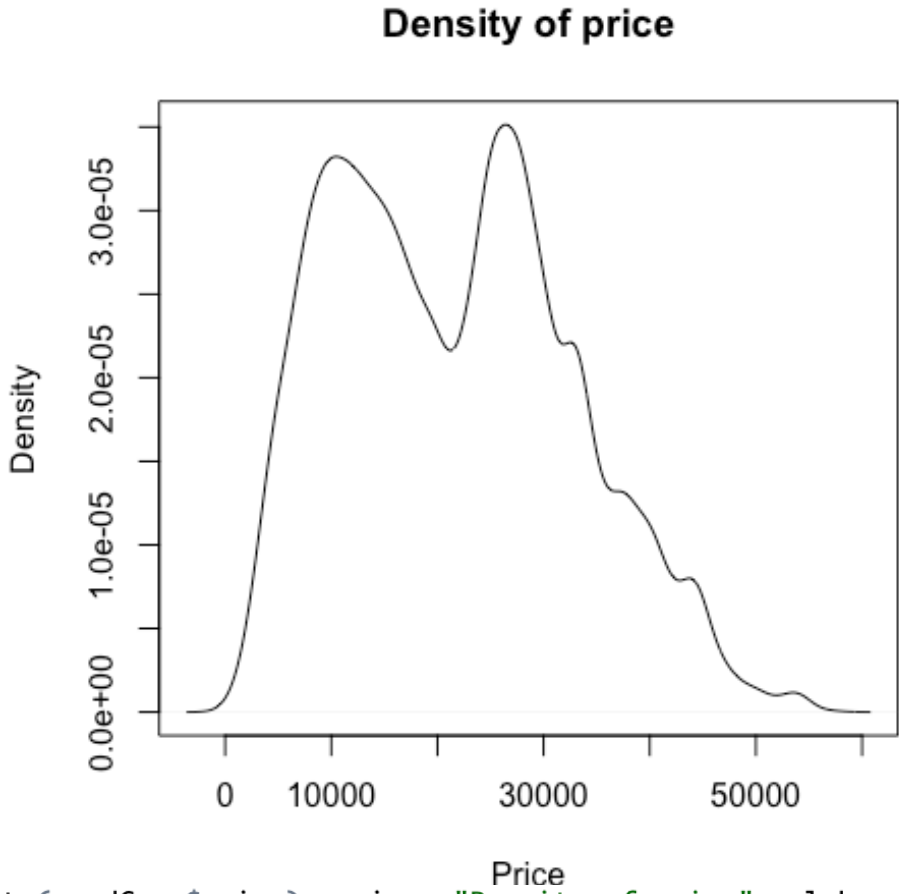



```
bins <- seq(0, 100000, 1000)
hist(usedCars$price[usedCars$price < 100000], breaks=bins, main = "Histogram of cars prices (bin width = 1000)"
```

d. در این سوال سعی شد نمودار توزیع این متغیر رسم شود و باز هم به دلیل داده پرت این امکان وجود نداشت. برای همین این داده‌های خیلی پرت به کل از مجموعه داده حذف می‌کنیم. (بالای ۱۰۰۰۰۰ در مجموع ۱۳ داده وجود دارد.)

```
usedCars <- usedCars[usedCars$price < 100000,]
```

نمودار توزیع این‌داده‌ها به صورت زیر است.




```
plot(density(usedCars$price), main = "Density of price", xlab = "Price")
```

e. این متغیر bimodal است و همچنین، right skewed (راست چوله) است. مقدار skewness با استفاده از تابع skewness کتابخانه PerformanceAnalytics ۰.۳۴۴۰۰۲ به‌دست‌آمد. فرمول استفاده‌شده برای محاسبه به صورت زیر است.

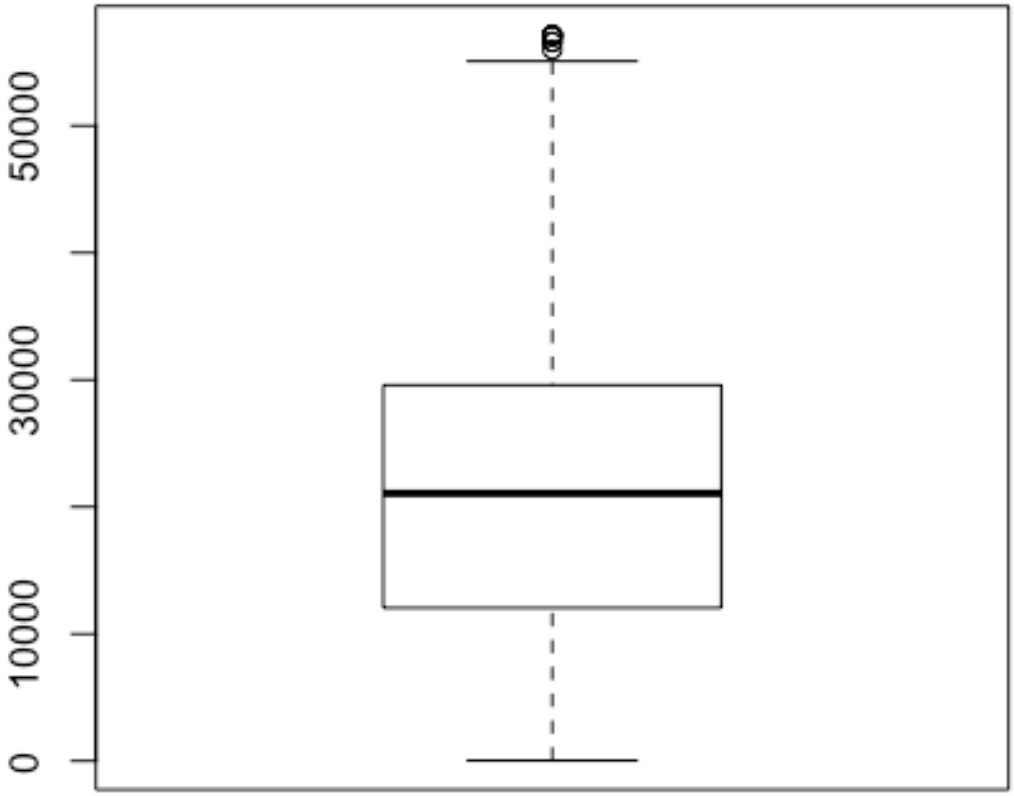
$$g_1 = m_3 / m_2^{3/2}$$

```
library(PerformanceAnalytics)
print(skewness(usedCars$price))
```

f. Skewness در سوال قبل محاسبه شد. مقادیر میانگین، واریانس و انحراف معیار در زیر آمده‌است.

```
"Mean:  21639.3219537542"
"Variance:  125917052.260915"
"Standard deviation:  11221.2767660777"
```

g. Box plot این متغیر در زیر آمده‌است.



چارک اول مقدار ۱۲۰۴۴ و چارک سوم مقدار۲۹۵۸۴ را دارد. میانه بر روی ۲۱۰۶۲ قرار دارد. Whisker پایین و بالا به ترتیب مقادیر ۳۳ و ۵۵۰۷۷ را دارند. مقدار IQR برابر است با:

$$IQR = Q_3 - Q_1 = 29584 - 12044 = 17540$$

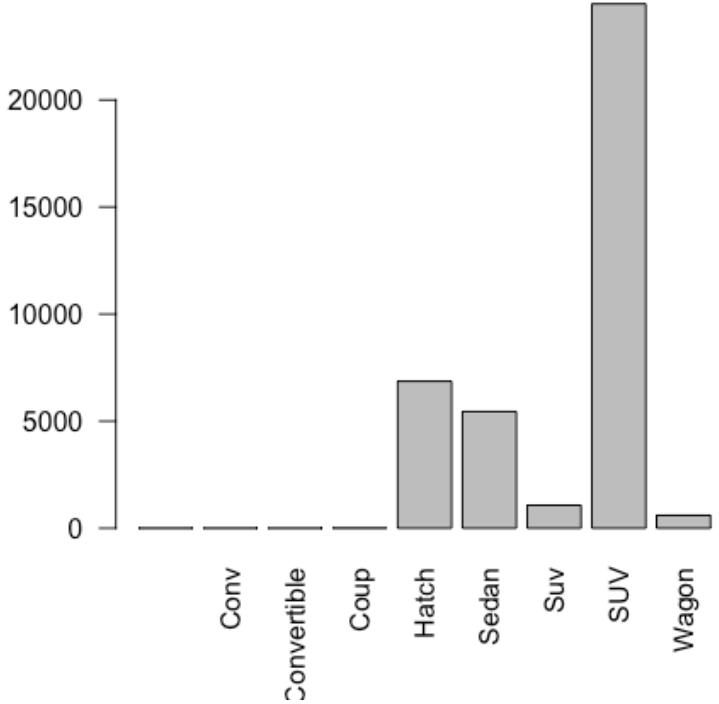
```
plt <- boxplot(usedCars$price)
print(plt)
```

h. داده‌های خارج از بازه whiskerها، داده‌های پرت محسوب می‌شوند. نه داده پرت با مقادیر ۵۶۵۸۹-۵۶۵۹۸-۵۷۱۳۲-۵۷۰۷۸-۵۷۱۰۴-۵۷۱۱۰-۵۷۱۱۰-۵۷۰۶۵-۵۵۹۸۷ وجـــود دارد. این مقادیر از خروجی box plot به‌دست‌آمد.

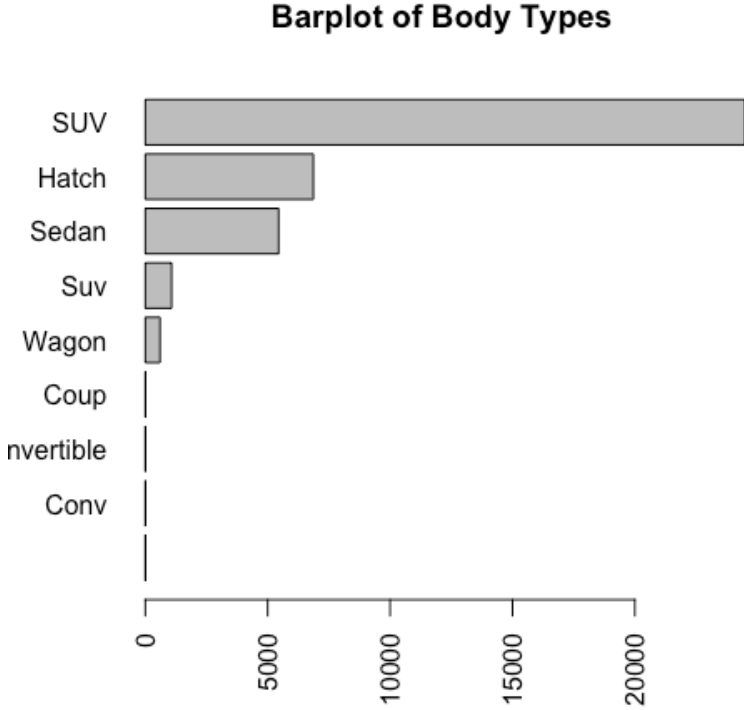
سوال ۲. ویژگی categorial انتخاب شده نوع بدنه(body_type) است.

a. Bar plot آن به صورت زیر است.

```
body_types <- table(usedCars$body_type)
barplot(body_types, las=2)
```



b. Bar plot افقی مرتب‌شده به صورت زیر است.



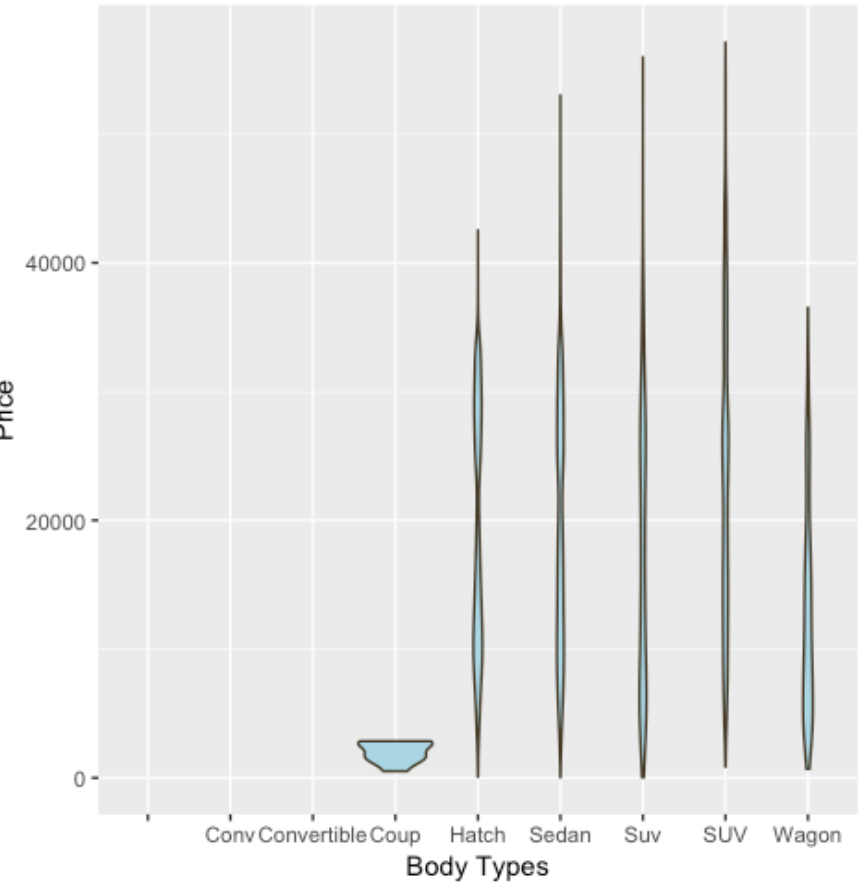

```
body_types <- sort(body_types)
barplot(body_types, las=2, horiz = TRUE, main = "Barplot of Body Types")
```

c. جدول تکرار برای این متغیر با استفاده از تابع count کتابخانه plyr به‌دست‌آمد که کد متناظر و خروجی‌آن به صورت یک dataframe در زیر آمده‌است.

```
library(plyr)
print(count(usedCars, "body_type"))
```

	body_type	freq
1		2
2	Conv	2
3	Convertible	2
4	Coup	8
5	Hatch	6865
6	Sedan	5447
7	Suv	1074
8	SUV	24488
9	Wagon	602

d. Violin plot مربوط به این متغیر به همراه کد متناظر رسم آن در زیر آمده‌است. البته لازم به ذکر است که برای محور عمودی باید یک متغیر عددی انتخاب شود که متغیر قیمت برای این مورد انتخاب شده‌است.

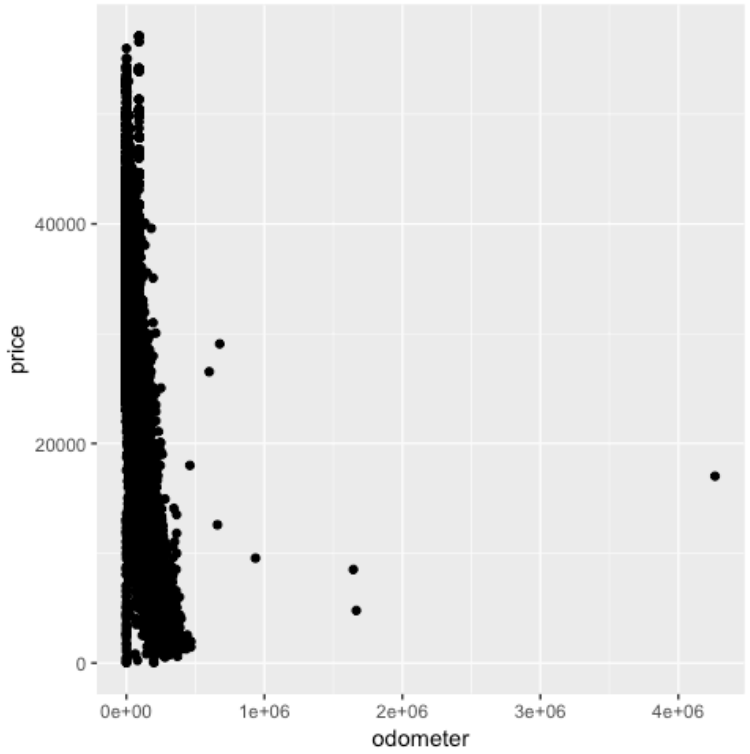



```
violin <- ggplot(usedCars, aes(x = body_type,
                               y = price)) +
  geom_violin(fill = "lightBlue", color = "#473e2c") +
  labs(x = "Body Types", y = "Price")
plot(violin)
```

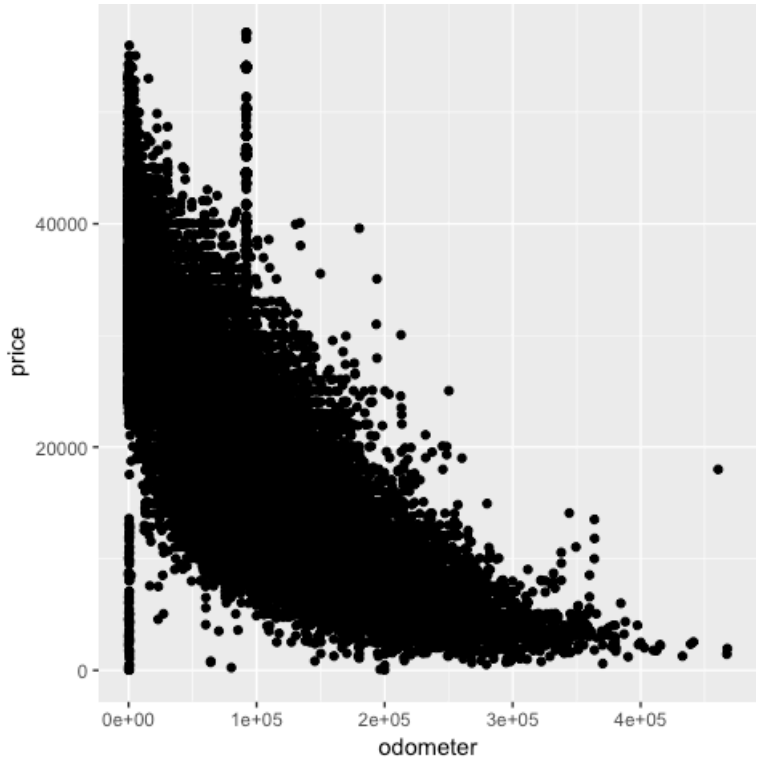
سوال ۳.

a. دو متغیر عددی انتخاب شده price و odometer هستند. Scatter plot آن‌ها در زیر آمده‌است. همانطور که مشاهده می‌کنید، داده‌ها پرت در متغیر odometer اجازه تحلیل درست نمودار را به ما نمی‌دهند. به همین دلیل، داده‌های پرت این متغیر را حذف می‌کنیم. کد این قسمت به صورت مقابل است. این داده‌های پرت تنها ۷ مورد بودند.

```
usedCars <- usedCars[usedCars$odometer < 500000,]
```



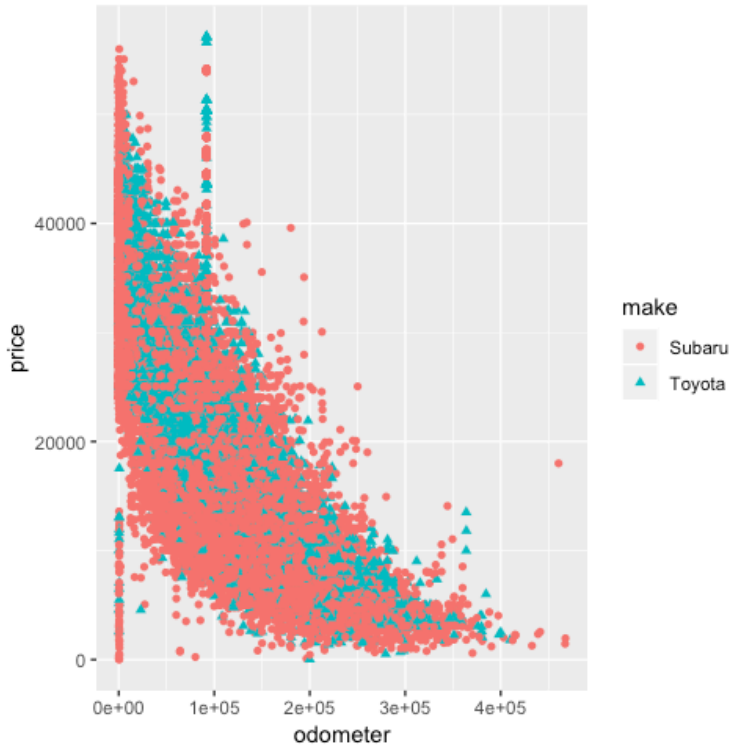
نمودار جدید به صورت زیر است. همانطور که مشاهده می‌شود این دو متغیر رابطه معکوس با یکدیگر دارند. دلیل آن‌هم این است که هرچه کارکرد ماشین بالاتر برود، قیمت ان پایین‌تر می‌آید.




```
usedCars <- usedCars[usedCars$odometer < 500000,]
plot(ggplot(usedCars, aes(x=odometer, y=price)) + geom_point())
```

```
plot(ggplot(usedCars, aes(x=odometer, y=price, shape=make, color=make)) + geom_point())
```

b. متغیر categorial انتخاب‌شده متغیر make است. هم شکل و هم رنگ نقاط نمودار قبل در نمودار زیر بر اساس سازنده متمایز شده‌اند. نمودار در شکل زیر، نمودار قابل مشاهده‌است.



c. ضریب همبستگی با استفاده از کد زیر -0.7881534 محاسبه‌شده‌است.

```
print(cor(usedCars$odometer, usedCars$price))
```

با استفاده از کد زیر و تابع cor.test همبستگی این دو متغیر آزموده‌شد و نتایج آن در ادامه آمده‌است.

```
print(cor.test(usedCars$odometer, usedCars$price))
```

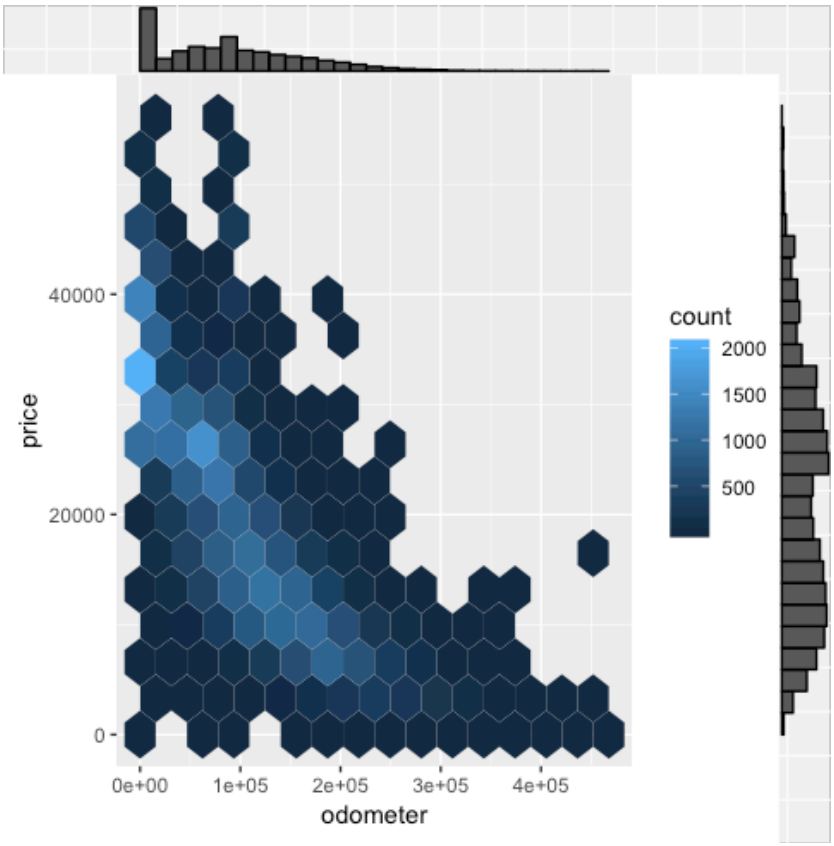
```
        Pearson's product-moment correlation

data:  usedCars$odometer and usedCars$price
t = -251.2, df = 38481, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.7919087 -0.7843386
sample estimates:
       cor
-0.7881534
```

مقدار p-value بسیار کم به‌دست‌آمد. پس، فرض صفر ما رد می‌شود. این نشان‌می‌دهد که دو متغیر ما به هم همبستگی دارند.

d. برای رسم نمودار hex از کتابخانه ggplot و تابع geom_hex با تعداد bin پانزده استفاده شد. همچنین، برای اضافه کردن توزیع‌های حاشیه‌ای از تابع ggMarginal کتابخانه ggExtra استفاده شد. کد و نمودار در پایین آمده است.

همان طور که در تصویر مشاهده‌می‌شود، مقادیر میانی نمودار بیشتر هستند. این‌ها مقادیری هستند که در خط رابطه هستند. چراکه گفته‌شد، این عکس همانطور که یکدیگر رابطه دو متغیر با همچنین، قابل عکس دارند. است که مقادیر پیش‌بینی پایین کمتر خیلی بالا و مقادیر میانی باشند و دلیل میل به نمودار به باشند. مرکز بیشتر

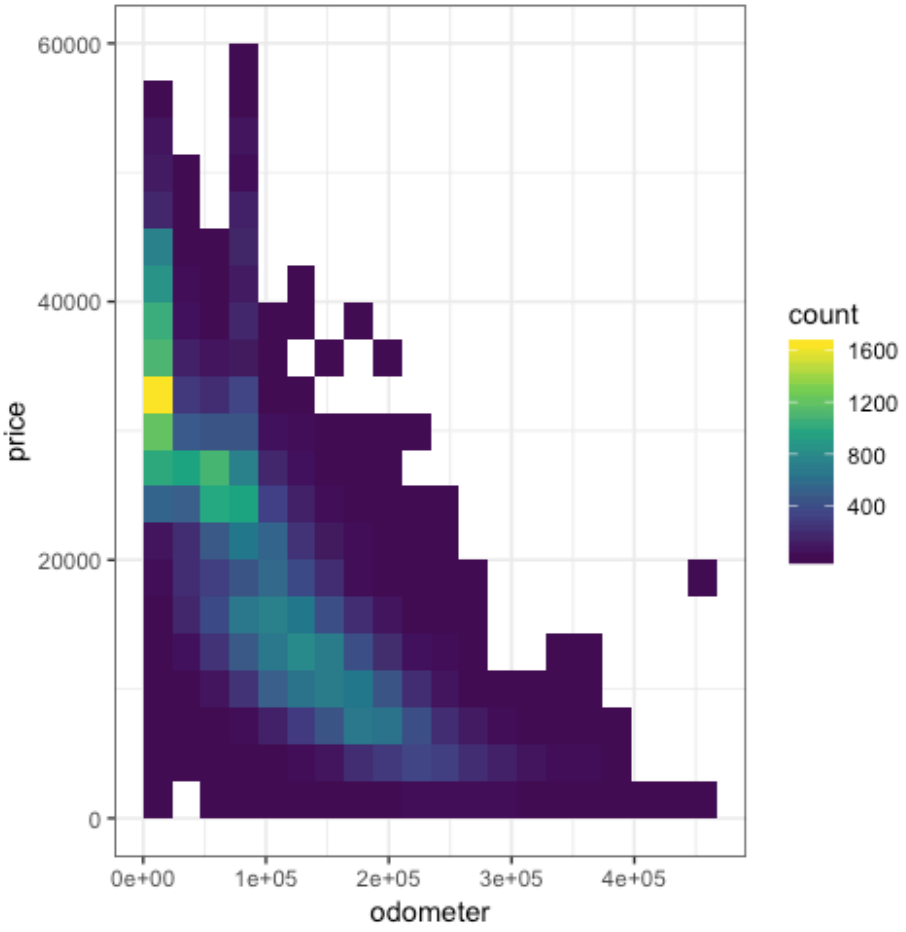



```
library(ggExtra)
hex <- ggplot(usedCars, aes(odometer, price))
plot(ggMarginal(hex + geom_point(col="transparent") + geom_hex(bins=15), type="histogram", size=10))
```

اگر تعداد bin ها بیشتر شود، ریزدانگی خیلی بالا می‌رود و تعداد داخل هر bin بسیار کم می‌وشد و این‌گونه نمی‌توان تفاوت‌های binها را دید و این طرح را تحلیل کرد. اگر تعداد binها کمتر شود، تعداد داخل هر bin خیلی بالا می‌رود و همچنین، ناحیه گسترده‌ای را پوشش می‌دهد و این‌گونه نیز طرح خاصی برای تحلیل وجود ندارد.

e. نمودار و کد در زیر آمده‌است.

```
plot(ggplot(usedCars, aes(x=odometer, y=price) ) +
  geom_bin2d(bins = 20) +
  scale_fill_continuous(type = "viridis") +
  theme_bw())
```

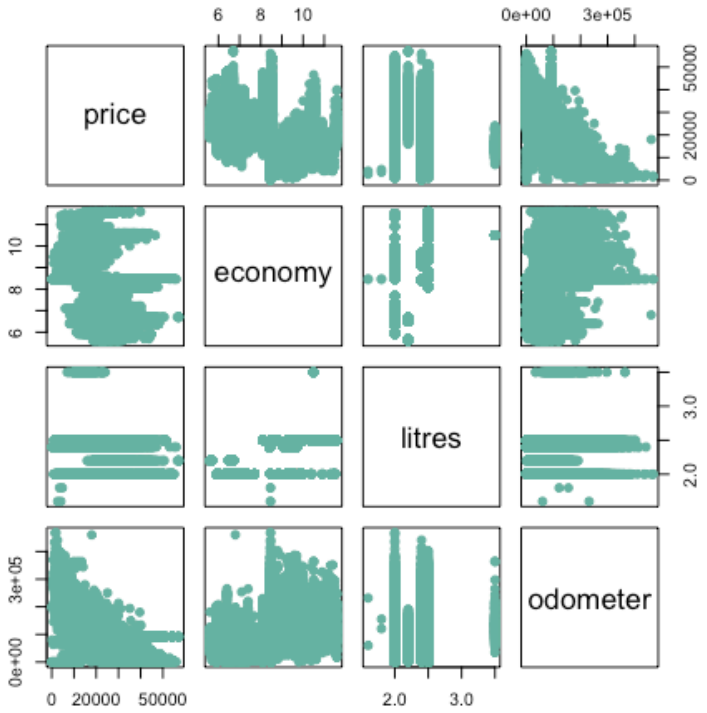


همانطور که مشاهده می‌شود این نمودار نیز تعبیر مشابه قسمت قبل دارد. یعنی قسمت میانی و روی خط رابطه عکس تعداد داده بیشتری را به خود اختصاص داده‌اند.

این نوع density plotها ساده‌ترین نوع آن‌هاست ولی hex bin راه بسیار سودمند تری است. زیرا این شش ضلعی به دایره شبیه تر است تا مربع. دایره بهترین شکل برای مدل کردن نقاط دور یک مرکز است و به همین دلیل، شش ضلعی بهینه ترین مدل برای این کار است.

سوال ۴.

a. متغیرهای عددی انتخاب شده price و odometer و economy و litres هستند. نمودار رابطه این متغیر و تکه کد مربوط به این قسمت در زیر آمده‌است. برای رسم نمودار از plot استفاده شد.

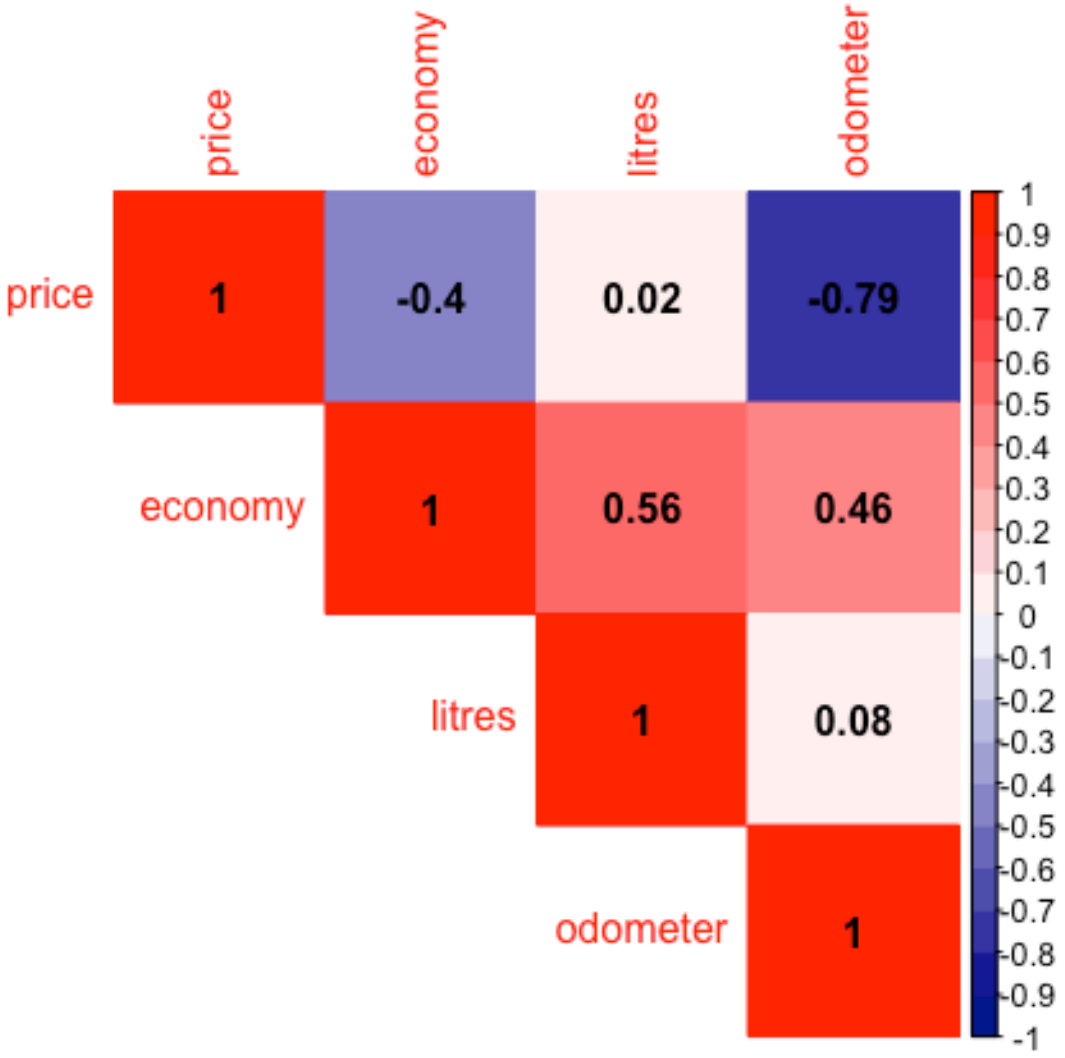


```r
cor_mat <- usedCars[ , c(1, 8, 11, 15)]
plot(cor_mat , pch=20 , cex=1.5 , col="#69b3a2")
```

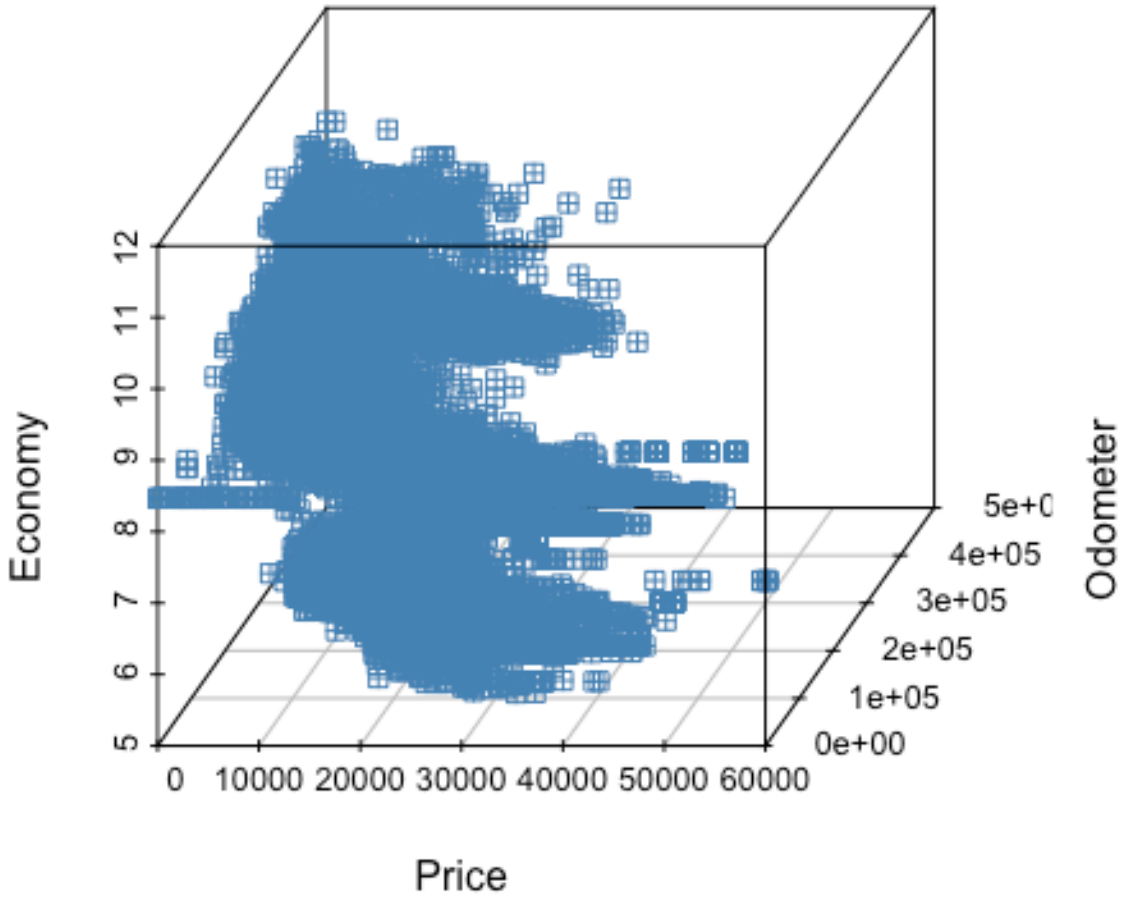
b. لازم به ذکر است متغیر liters نیز به نوعی categorical است ولی متاسفانه مجموعه داده متغیر عددی دیگری نداشت. به همین دلیل، سطر و ستون سوم گسستگی آشکار دارند. به نظر economy و price رابطه عکس پنهانی با یکدیگر دارند. Litres به نظر رابطه‌ای با هیچکدام به جز economy ندارد. با economy رابطه اندک مثبتی دارد. Odometer و economy نیز به نظر رابطه اندک مثبت و مستقیمی با یکدیگر دارند. جالب‌ترین طرح بین odometer و price دیده‌می‌شود که در سوال قبل نیز به آن پرداخته‌شد.

c. نمودار این heat map همبستگی به همراه کد آن در پایین آمده‌است.

```r
library(corrplot)
col<- colorRampPalette(c("darkblue", "white", "red"))(20)
corrplot(cor(cor_mat), type="upper", method="color", addCoef.col = "black", col=col)
```



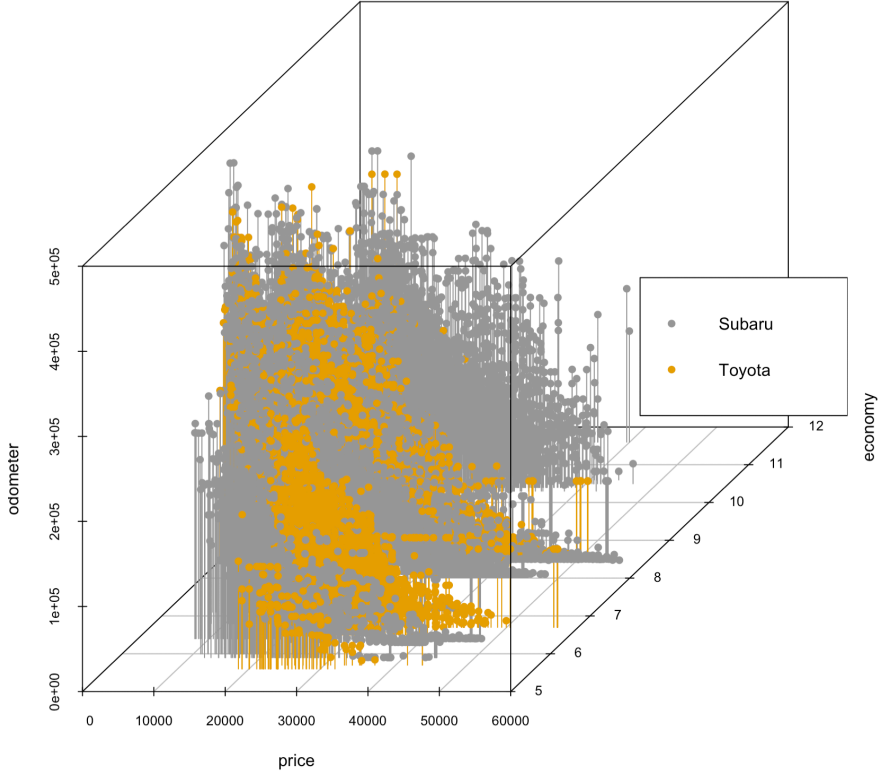
d. متغیر litres را از متغیرهای قبلی حذف می‌کنیم و برای سه‌تای دیگر نمودار ۳d می‌کشیم.
نمودار و کد مربوطه در پایین آمده‌است.




```
library("scatterplot3d")
scatterplot3d(cor_mat$price, cor_mat$odometer, cor_mat$economy, angle = 60, pch = 12, color="steelblue",
              xlab = "Price", ylab = "Odometer", zlab = "Economy")
```

همان روابطی که در بخش‌های قبل بیان‌شده‌بود، در این نمودار نیز تاحدی قابل مشاهده است.
شاخص ترین چیزی که مشخص است، رابطه عکس odometer و price است.
e. نمودار رنگی با استفاده از متغیر categorial سازنده به شکل زیر درآمد.

```
colors <- c("#999999", "#E69F00")
colors <- colors[as.numeric(usedCars$make)]
scatterplot3d(usedCars[,c(1, 8, 15)], pch = 16, color=colors, type="h")
legend("right", legend = levels(usedCars$make), col =  c("#999999", "#E69F00"), pch = 16)
```



سوال ۵.

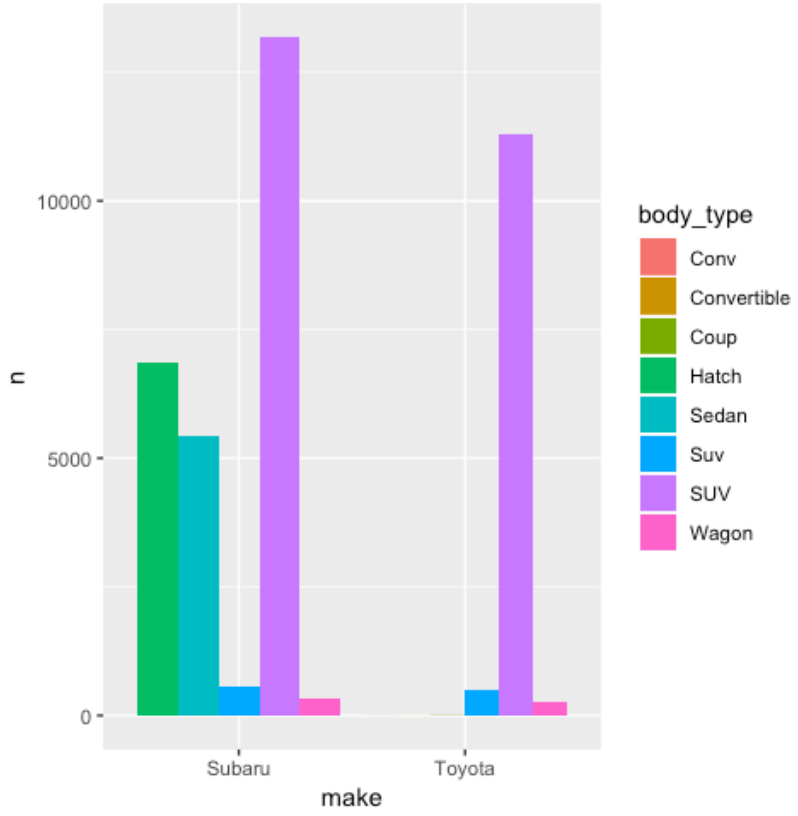
a. برای این بخش دو متغیر body_type و make انتخاب شده‌اند. سازنده دو ماشین رشته خالی ثبت شده‌است که به دلیل کم بودن تعداد(۲ عدد)، حذفشان می‌کنیم.

	Subaru	Toyota
Conv	0	2
Convertible	0	2
Coup	0	8
Hatch	6864	0
Sedan	5444	0
Suv	577	497
SUV	13177	11308
Wagon	335	267

```
usedCars <- usedCars[!usedCars$body_type=="",]
usedCars$body_type <- droplevels(usedCars$body_type)
print(table(usedCars$body_type, usedCars$make))
```

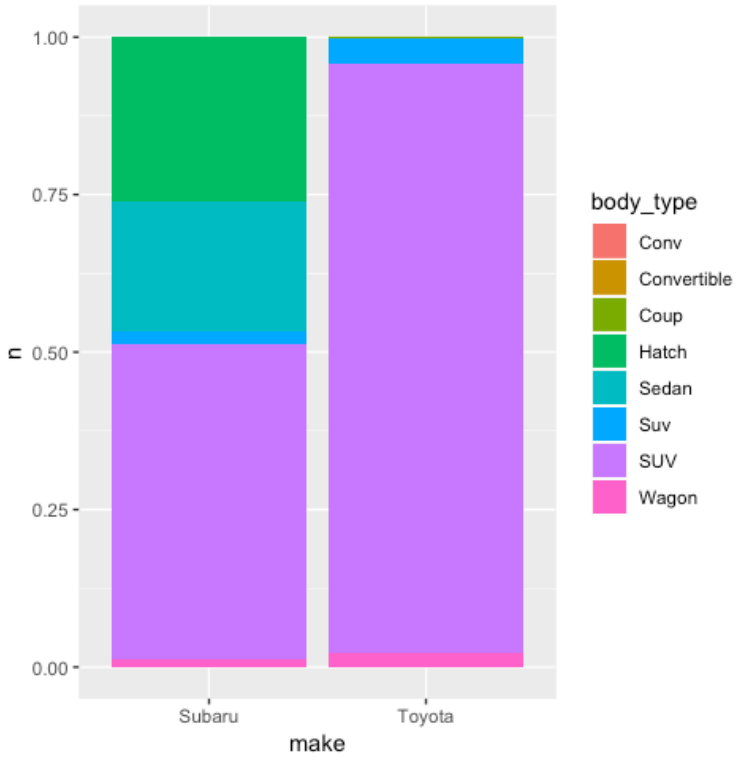
b. برای این بخش دو متغیر body_type و make انتخاب شده‌اند. برای این منظور ابتدا تعداد ترکیب دو دسته را می‌شماریم.

```
group_bar <- usedCars %>% group_by(body_type, m
plot(ggplot(group_bar, aes(fill=body_type, y=n,
  geom_bar(position="dodge", stat="identity"))
```



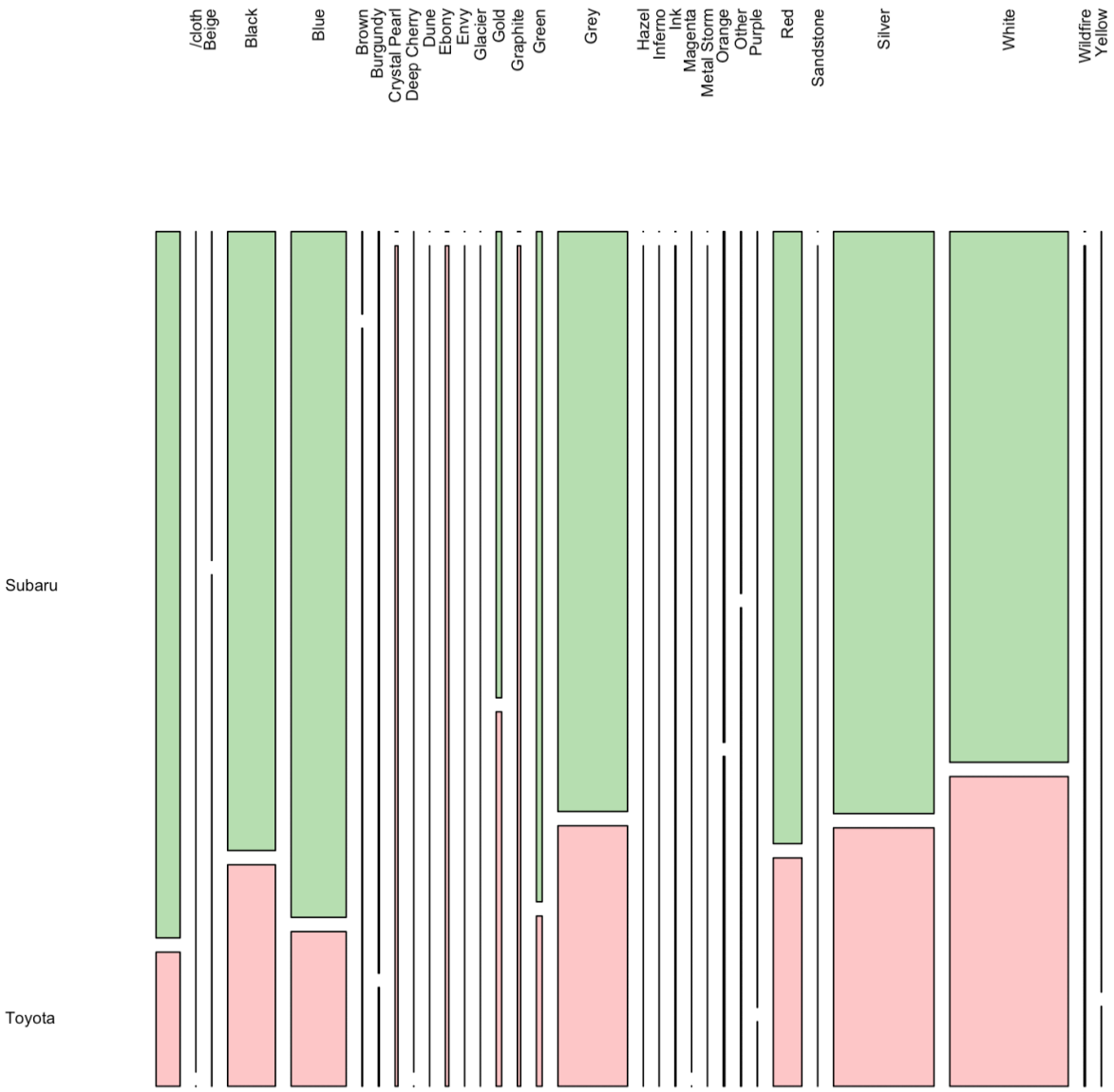
c. برای این بخش دو متغیر body_type و make انتخاب شده‌اند.

```
plot(ggplot(group_bar, aes(fill=body_type, y=n, x=make)) +
  geom_bar(position="fill", stat="identity"))
```



d. برای این بخش دو متغیر colour و make انتخاب شده‌اند.

```
mosaicplot(table(usedCars$colour, usedCars$make), col = hcl(c(120, 10)), las=2)
```



سوال ۶. برای این بخش متغیر price انتخاب شده‌است.

a. بازه اطمینان ۹۸ درصد به صورت زیر به‌دست می‌آید.

$$\bar{x} \pm z^* \times \frac{s}{\sqrt{n}}$$

کد و نتیجه در زیر آمده‌است.

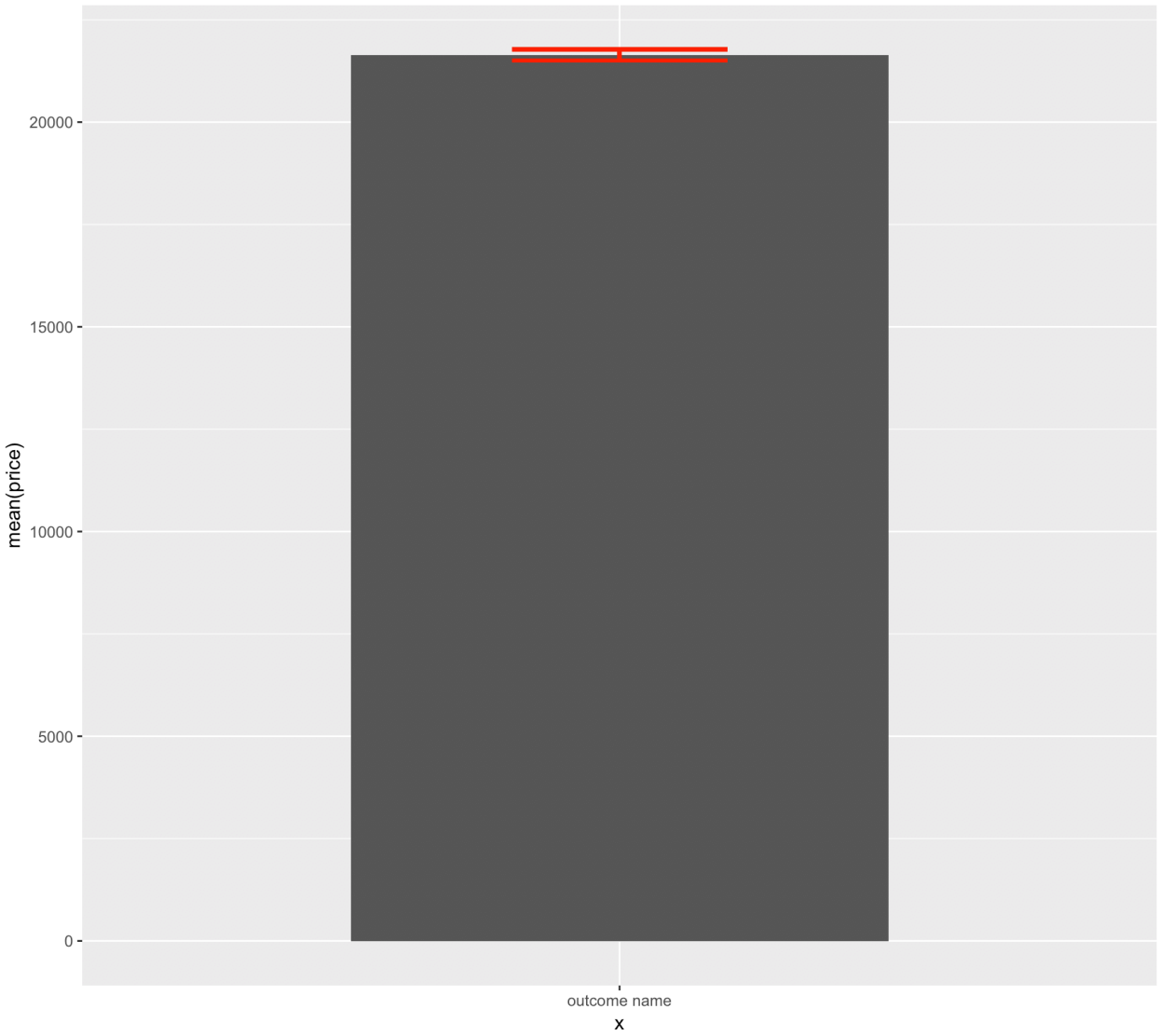
```
x_bar <- mean(usedCars$price)
z_star <- -qnorm(0.01)
s <- sd(usedCars$price)
SE <- s / sqrt(length(usedCars$price))
print(paste("98% CI: [", x_bar - (z_star * SE), ", " , x_bar + (z_star * SE), "]"))
```

```
"98% CI: [ 21507.129359354 ,  21773.2860664405 ]"
```

b. ۹۸ درصد اطمینان داریم که میانگین قیمت ماشین‌های دست دوم در این بازه باشد. بازه اطمینان می‌گوید اگر تعداد زیادی نمونه برداریم ۹۸ درصد میانگین آن‌ها در این بازه می‌افتد.

c. نمودار و کد مربوطه در زیر آمده‌است.

```
plot(ggplot(usedCars, aes(x="outcome name", y=mean(price))) +
  geom_bar(position = 'dodge', stat='identity', width=.5) +
  geom_errorbar(aes(ymin = lower_bound, ymax = upper_bound),
                width = 0.2,
                linetype = "solid",
                position = position_dodge(width = 0.5),
                color="red", size=1))
```



d. آزمون فرض را به صورت زیر تعریف می‌کنیم.

فرض صفر: میانگین برابر ۲۱۰۰۰ است. $\mu_0 = 21000$
فرض جایگزین: میانگین بیشتر از ۲۱۰۰۰ است. $\mu_a > 21000$
کد و مقدار p value در زیر آمده‌است.

```
sample_indices <- sample(nrow(usedCars), 80)
sampled_prices <- usedCars$price[sample_indices]
x_bar <- mean(sampled_prices)
se <- sd(usedCars$price) / sqrt(80)
z_stat <- (x_bar - 21000) / se
p_value <- pnorm(z_stat, lower.tail = FALSE)
print(paste("p-value: ", p_value))
```

```
"p-value:  0.0133554386772776"
```

همانطور که مشاهده می‌شود، مقدار p_value کمتر از ۰.۰۵ است و بنابراین، فرض صفر را رد می‌شود و فرض جایگزین پذیرفته می‌شود. این p_value می‌گوید که به احتمال ۱.۳ درصد با فرض درستی فرض صفر، این میانگین نمونه می‌توانست دیده‌شود.

e. بازه اطمینان و کد مربوطه در زیر آمده‌است.

```
x_bar <- mean(usedCars$price)
z_star <- -qnorm(0.025)
s <- sd(usedCars$price)
SE <- s / sqrt(length(usedCars$price))
upper_bound <- x_bar + (z_star * SE)
lower_bound <- x_bar - (z_star * SE)
print(paste("95% CI: [", lower_bound, ", " , upper_bound, "]"))
```

```
"95% CI: [ 21528.0882904424 ,  21752.3271353522 ]"
```

همانطور که مشاهده می‌شود، نقطه فرض صفر در این بازه اطمینان قرار نگرفت. بنابراین، این نمونه فرض صفر را رد می‌کند و فرض جایگزین را می‌پذیرد.

f. g. این دو قسمت با هم انجام شد. ابتدا power محاسبه شد و سپس، با متمم گیری مقدار β به‌دست‌آمد. کد و نتایج در زیر قابل مشاهده است. فرض شد که میانگین واقعی میانگین تمام مجموعه داده‌است.

```
actual_average <- mean(usedCars$price)
s <- sd(usedCars$price)
SE <- s / sqrt(80)
boundary <- -qnorm(0.025)
z_stat_bar <- SE * boundary + 21000
z_stat <- (z_stat_bar - actual_average) / SE
power <- pnorm(z_stat, lower.tail = FALSE)
print(paste("Power: ", power))
beta <- 1 - power
print(paste("Type II error: ", beta))
```

```
"Power:  0.07357356350576"
"Type II error:  0.92642643649424"
```

Effect size بیانگر اختلاف بین تخمین نقطه‌ای و فرض صفر است. هنگامی که n زیاد باشد، یک اختلاف کم میان این دو می‌تواند به عنوان اختلاف آماری شناسایی شود. هنگامی که effect size کم باشد، power کم می‌شود چراکه تشخیص دادن این اختلاف سخت تر می‌شود و احتمال پذیرفتن فرض صفر بالاتر می‌رود. در نتیجه، بتا زیاد و توان کم می‌شود. برای بالاتر بردن توان تست باید تعداد نمونه‌برداری را بالا ببریم. همانطور که در زیر قابل مشاهده‌است، افزایش تعداد از ۸۰ به ۸۰۰ و ۲۰۰۰ توان را به ترتیب ۲۹ و ۶۵ درصد افزایش داده‌است تا به حد قابل قبولی رسیده‌است.

```
"Power:  0.364554242037165"
"Type II error:  0.635445757962835"
```

```
"Power:  0.722888622752378"
"Type II error:  0.277111377247622"
```

سوال ۷. دو متغیر عددی انتخاب شده price و odometer هستند.

a. از تست t استفاده می‌کنیم چراکه تعداد نمونه‌های ما کم است و نمی‌توانیم فرض کنیم نمونه ما به اندازه کافی بزرگ است و با استفاده از قضیه حد مرکزی آن را با نمودار نرمال مدل کنیم. به همین دلیل، از توزیع t با درجه آزادی استفاده می‌کنیم تا تقریب درست تری داشته باشیم. این توزیع دیرتر به صفر میل می‌کند.

b. آزمون فرض به صورت زیر است.

فرض صفر: میانگین این دو متغیر با یکدیگر تفاوت ندارد. $\mu_p = \mu_o$

فرض جایگزین: میانگین این دو متغیر با یکدیگر تفاوت دارد. $\mu_p \neq \mu_o$

مقدار p-value و بازه اطمینان ۹۵ درصد به همراه کد مربوطه در زیر آمده‌است.

```r
point_estimate <- mean(sampled_data$odometer) - mean(sampled_data$price)
df <- min(length(sampled_data$odometer), length(sampled_data$price)) - 1
t_star <- -qt(0.25, df = df)
se <- sqrt((sd(sampled_data$price)**2 / length(sampled_data$price)) + (sd(sampled_data$odometer)**2 / length(sampled_data
lower_bound <- point_estimate - t_star * se
upper_bound <- point_estimate + t_star * se
print(paste("CI: [", lower_bound, ",", upper_bound, "]"))
t_stat <- point_estimate / se
p_value <- pt(t_stat, df=df, lower.tail = FALSE)
print(paste("P-value: ", p_value))
```

```
"CI: [ 52072.9710770883 , 75583.160512769 ]"
"P-value:  0.000534678121182339"
```

همانطور که مشاهده می‌شود، مقدار p-value بسیار کوچک است و بنابراین، فرض صفر رد می‌شود. یعنی این دو میانگین با یکدیگر تفاوت دارند. بازه اطمینان نیز موید این موضوع است چراکه فرض صفر یعنی نقطه صفر داخل این بازه نیست.

سوال ۸. متغیر عددی انتخاب شده price است. این متغیر دارای outlier هایی است که تاثیر به سزایی روی میانگین دارند. پس، شاخص مرکزی مناسب میانه است. برای میانه توزیع

مشخصی نداریم. بنابراین، باید از bootstrapping استفاده کنیم. برای این کار از کتابخانه boot استفاده می‌کنیم.

c. در این روش کافی است میانه‌های نمونه‌ها را مرتب کنیم و داده‌ی ششم و نود و پنجم را به عنوان مرز انتخاب کنیم. کد و بازه در زیر قابل مشاهده‌است.

```
resamples <- lapply(1:100, function(i) sample(usedCars$price, replace = TRUE))
samples.median <- sapply(resamples, median)
## a ##
samples.median <- sort(samples.median)
print(paste("CI: [", samples.median[6], ",", samples.median[95], "]"))
```

```
"CI: [ 21014 , 21505 ]"
```

b. کد و بازه اطمینان مربوطه در زیر آمده‌است.

```
median_mean <- mean(samples.median)
median_se <- sqrt(var(samples.median)) / 10
t_star <- -qt(0.025, df=99)
upper_bound <- median_mean + t_star * median_se
lower_bound <- median_mean - t_star * median_se
print(paste("CI: [", lower_bound, ",", upper_bound, "]"))
```

```
"CI: [ 20812.8366832056 , 21417.9033167944 ]"
```

c. همانطور که مشاهده می‌شود، تفاوت بسیار کمی میان این دو بازه اطمینان وجود دارد. دلیل این امر خوب بودن نمونه‌های گرفته‌شده‌است چراکه انحراف معیار آن‌ها معین نحوه پخش‌شدن این‌داده‌هاست و داده‌های مرزی به نقاط به‌دست‌آمده از روی توزیع (انحراف معیار) نزدیک هستند.

سوال ۹. این متغیر categorial با دو سطح automatic و manual است. پس باید میانگین این دو گروه را با یکدیگر مقایسه کنیم و ببینیم آیا تفاوت معناداری با یکدیگر دارند یا خیر. با توجه به اینکه دو گروه داریم نیازی به استفاده از تست ANOVA نیست و کافی است از همان t-test استفاده کنیم. برای اجرای t-test هر کدام از شرط‌ها را بررسی می‌کنیم.

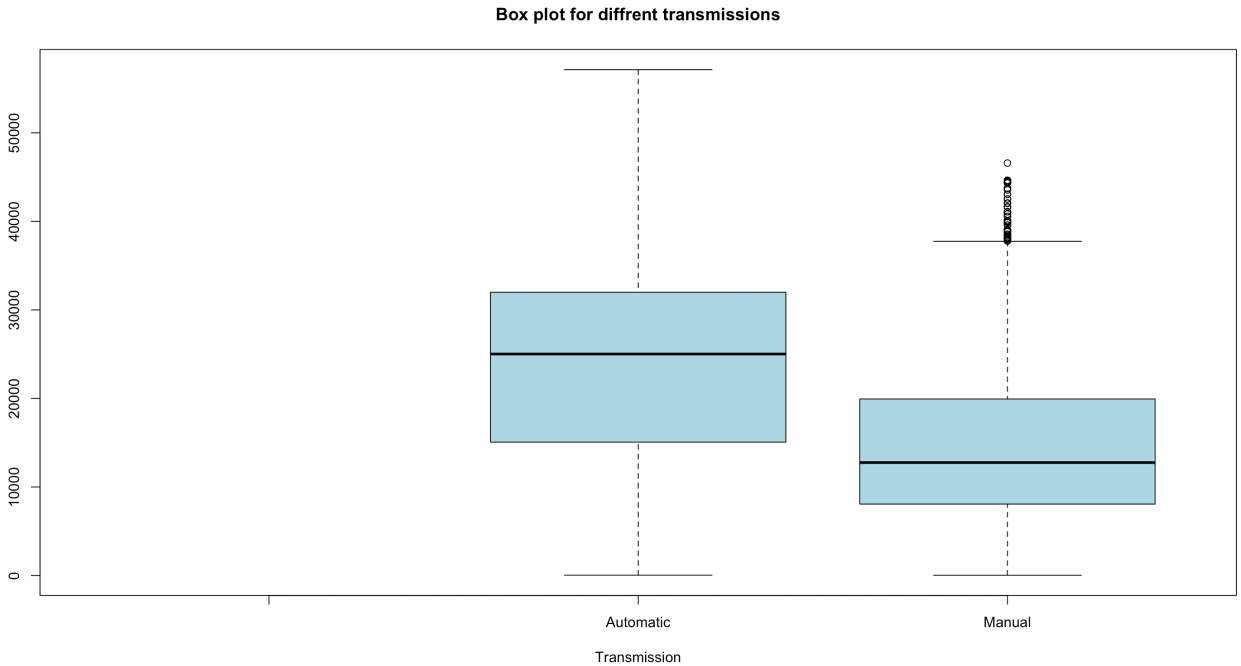
استقلال درون گروهی: مشاهده‌ها از هم مستقلند. سایز هر نمونه از ۱۰ درصد جامعه کمتر است.
استقلال بین گروهی: این نمونه‌ها از یکدیگر مستقلند چراکه یک ماشین همزمان نمی‌تواند هم دنده‌ای و هم اتوماتیک باشد.
برای اندازه نمونه نیز تمام نمونه‌های گروه‌ها را در نظر می‌گیریم چراکه خود این مجموعه داده خود نمونه‌برداری شده از کل جامعه است و تعداد آن نیز به اندازه کافی بزرگ است. به دلیل تعداد بالا، عملا t-test ما تبدیل به z-test می‌شود.
آزمون فرض به صورت زیر است.
فرض صفر: میانگین دو گروه با یکدیگر برابر است. $\mu_a - \mu_m = 0$
فرض مقابل: میانگین قیمت ماشین‌های اتوماتیک بیشتر از ماشین‌های دنده‌ای است. $\mu_a - \mu_m > 0$
کد آزمون فرض و مقدار p-value در زیر آمده‌است.

```r
manual_cars <- usedCars$price[usedCars$transmission == "Manual"]
auto_cars <- usedCars$price[usedCars$transmission == "Automatic"]
man_mean <- mean(manual_cars)
man_s <- sqrt(var(manual_cars))
man_len <- length(manual_cars)
auto_mean <- mean(auto_cars)
auto_s <- sqrt(var(auto_cars))
auto_len <- length(auto_cars)
point_estimate <- auto_mean - man_mean
null_h <- 0
se <- sqrt((man_s^2/man_len)+(auto_s^2/auto_len))
df <- min(man_len, auto_len) - 1
t_stat <- (point_estimate - null_h) / se
p_value <- pt(t_stat, df=df, lower.tail = FALSE)
print(paste("P-value: ", p_value))
```

```
"P-value:  0"
```

مقدار P-value صفر به‌دست‌آمد که کمتر از هر significance level ای می‌باشد. بنابراین، فرض صفر ما رد می‌شود. نتیجه می‌گیریم میانگین ماشین‌های اتوماتیک از ماشین‌های دستی بالاتر است. در حقیقت، میانگین این دو گروه با یکدیگر برابر نیست و فرض مقابل در برابر فرض صفر پذیرفته می‌شود.

برای واضح ترشدن این رابطه box plot مربوط به هر دسته رسم شده‌است که در زیر قابل مشاهده است. همانطور که می‌بینید این اختلاف در این نمودار نیز قابل مشاهده‌است.




```r
boxplot(price~transmission,
        data=usedCars[!usedCars$transmission == "", ],
        main="Box plot for diffrent transmissions",
        xlab="Transmission",
        ylab="price",
        col="lightblue",
        border="black"
)
```