Report

Project Phase I

Car Insurance Claim Prediction

Statistical Inference

Fatemeh Nadi

810101285

December, 2022

Contents

1. Question 0

1.1 A

Briefly describe your dataset and why studying your dataset can be interesting?

Dataset: Car Insurance Claim Prediction

There are attributes in the dataset about policyholders, including tenure of the policy, age of the car, age of the owner, population density, car make and model, power, engine type, etc., and the target variable indicates whether a claim will be filed in the next six months.

	X	policy_id	policy_tenure	age_of_car	age_of_policyholder	area_cluster	population_density	segment	model	fuel_type	engine_type
1	5756	ID05757	0.082997707	0.07	0.3269231	C11	6108	B2	M6	Petrol	K Series Dual jet
2	21942	ID21943	0.097829849	0.03	0.5769231	C9	17804	B2	M7	Petrol	1.2 L K Series Engine
3	53860	ID53861	0.059199281	0.01	0.7019231	C3	4076	A	M1	CNG	F8D Petrol Engine
4	38217	ID38218	0.541986698	0.17	0.6153846	C2	27003	B2	M6	Petrol	K Series Dual jet
5	7724	ID07725	0.093530064	0.09	0.5096154	C5	34738	C2	M4	Diesel	1.5 L U2 CRDi
6	31567	ID31568	1.109789324	0.23	0.4038462	C2	27003	B2	M6	Petrol	K Series Dual jet
7	42661	ID42662	0.080624964	0.00	0.5000000	C14	7788	B2	M7	Petrol	1.2 L K Series Engine
8	2989	ID02990	0.337275748	0.18	0.6730769	C8	8794	B2	M6	Petrol	K Series Dual jet
9	39633	ID39634	0.598877634	0.12	0.4807692	C8	8794	B2	M6	Petrol	K Series Dual jet
10	42488	ID42489	1.231427992	0.12	0.4134615	C8	8794	B2	M7	Petrol	1.2 L K Series Engine

Figure 1: First 10 rows of the dataset

Analyzing this data help the insurance companies detects some cases which haven't enough revenue.
If a case has a high risk of accident or destruction so reject insurance it.
And finally, make the best decision to prevent failure in it's market.

Studying and analyzing the information in this database lets us determine what influences consumers' claims most. It also allows us to study any two columns and see if there is an interesting statistical relationship between them.
Also exploring the statistical distribution of the factors can also help us discover the most meaningful factors.

1.2 B

How many variables (features) and cases does your dataset have?

We have a dataset of 30000 policyholders, each policyholder has 28 features which we explain in detail below:

- policy_id: Unique identifier of the policyholder.
- policy_tenure: Time period of the policy.
- age_of_car: Normalized age of the car in years.
- age_of_policyholder: Normalized age of policyholder in years.
- area_cluster: Area cluster of the policyholder.
- population_density: Population density of the city (Policyholder City).
- segment: Segment of the car (A/ B1/ B2/ C1/ C2).
- model: Encoded name of the car.
- fuel_type: Type of fuel used by the car.
- engine_type: Type of engine used in the car.
- airbags: Number of airbags installed in the car.
- is_parking_sensors: Boolean flag indicating whether parking sensors are present in the car or not.
- is_parking_camera: Boolean flag indicating whether the parking camera is present in the car or not.
- rear_brakes_type: Type of brakes used in the rear of the car
- displacement: Engine displacement of the car (cc).
- cylinder: Number of cylinders present in the engine of the car.
- transmission_type: Transmission type of the car.
- gear_box: Number of gears in the car.
- steering_type: Type of the power steering present in the car.
- turning_radius: The space a vehicle needs to make a certain turn (Meters).
- length: Length of the car (Millimetre).
- width: Width of the car (Millimetre).
- height: Height of the car (Millimetre).
- gross_weight: The maximum allowable weight of the fully-loaded car, including passengers, cargo and equipment (Kg).

- is_speed_alert: Boolean flag indicating whether the speed alert svstem is available in the car or not.
- ncap_rating: Safety rating given by NCAP (out of 5).
- is_claim: Outcome: Boolean flag indicating whether the policyholder file a claim in the next 6 months or not.

```
> summary(car)
       X          policy_id         policy_tenure       age_of_car      age_of_policyholder area_cluster
 Min.   :    0   Length:30000       Min.   :0.00275   Min.   :0.00000   Min.   :0.2885      Length:30000
 1st Qu.:14647   Class :character   1st Qu.:0.20760   1st Qu.:0.02000   1st Qu.:0.3654      Class :character
 Median :29397   Mode  :character   Median :0.57231   Median :0.06000   Median :0.4519      Mode  :character
 Mean   :29314                      Mean   :0.60925   Mean   :0.06948   Mean   :0.4697
 3rd Qu.:43948                      3rd Qu.:1.03817   3rd Qu.:0.11000   3rd Qu.:0.5481
 Max.   :58590                      Max.   :1.38652   Max.   :0.82000   Max.   :0.9808
 population_density   segment             model            fuel_type         engine_type           airbags
 Min.   :  290      Length:30000       Length:30000       Length:30000       Length:30000       Min.   :1.00
 1st Qu.: 6112      Class :character   Class :character   Class :character   Class :character   1st Qu.:2.00
 Median : 8794      Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :2.00
 Mean   :18816                                                                                 Mean   :3.14
 3rd Qu.:27003                                                                                 3rd Qu.:6.00
 Max.   :73430                                                                                 Max.   :6.00
 is_parking_sensors is_parking_camera  rear_brakes_type    displacement     cylinder     transmission_type
 Length:30000       Length:30000       Length:30000       Min.   : 796   Min.   :3.000   Length:30000
 Class :character   Class :character   Class :character   1st Qu.: 796   1st Qu.:3.000   Class :character
 Mode  :character   Mode  :character   Mode  :character   Median :1197   Median :4.000   Mode  :character
                                                          Mean   :1163   Mean   :3.628
                                                          3rd Qu.:1493   3rd Qu.:4.000
                                                          Max.   :1498   Max.   :4.000
    gear_box     steering_type      turning_radius      length         width          height      gross_weight
 Min.   :5.000   Length:30000       Min.   :4.500   Min.   :3445   Min.   :1475   Min.   :1475   Min.   :1051
 1st Qu.:5.000   Class :character   1st Qu.:4.600   1st Qu.:3445   1st Qu.:1515   1st Qu.:1475   1st Qu.:1185
 Median :5.000   Mode  :character   Median :4.800   Median :3845   Median :1735   Median :1530   Median :1335
 Mean   :5.245                      Mean   :4.853   Mean   :3851   Mean   :1672   Mean   :1553   Mean   :1385
 3rd Qu.:5.000                      3rd Qu.:5.000   3rd Qu.:3995   3rd Qu.:1755   3rd Qu.:1635   3rd Qu.:1510
 Max.   :6.000                      Max.   :5.200   Max.   :4300   Max.   :1811   Max.   :1825   Max.   :1720
 is_speed_alert      ncap_rating      is_claim
 Length:30000       Min.   :0.000   Min.   :0.00000
 Class :character   1st Qu.:0.000   1st Qu.:0.00000
 Mode  :character   Median :2.000   Median :0.00000
                    Mean   :1.762   Mean   :0.06377
                    3rd Qu.:3.000   3rd Qu.:0.00000
                    Max.   :5.000   Max.   :1.00000
```

Figure 2: Summary of the dataset

Except for X as index and *policy_id*, we have 15 numerical and 11 categorical variables.

1.3 C

Is there any missing value in your data? Provide a summary of a portion of missing values for each variable (feature) and describe how you handle these missing values for each variable (on what basis).

There are no missing values in this dataset.

```
> any(is.na(car))
[1] FALSE
```

Figure 3: Detecting missing values

Portion of missing values for each variable (feature):

X	0
policy_id	0
policy_tenure	0
age_of_car	0
age_of_policyholder	0
area_cluster	0
population_density	0
segment	0
model	0
fuel_type	0
engine_type	0
airbags	0
is_parking_sensors	0
is_parking_camera	0
rear_brakes_type	0
displacement	0
cylinder	0
transmission_type	0
gear_box	0
steering_type	0
turning_radius	0
length	0
width	0
height	0
gross_weight	0
is_speed_alert	0
ncap_rating	0
is_claim	0

Figure 4: Missing values for each feature

The dataset shows all the values haven't any missing values. I also looked for special characters indicating missing values, such as ' ', '-', 'NA', and 'NAN', by unique on each column.
If the dataset has missing values, these could be fixed by deleting, replacing with mean, median, mod, etc., or using some algorithm like fill forward or estimating or backward eliminating.

1.4 D

Using this elementary view of your dataset, which variables do you think may be the most relevant (contain some important information)? Why?

Age of the car: if a car is too old, it's maybe destroyed as soon as possible, and generally, people care about new cars more than old cars.

Displacement: High displacement (compared to the car's age) shows this policyholder driving more, so the probability of accidents increases.

Policy tenure: one or two months is an alarm for insurance companies that may be this case, show fraud.

Population density: usually, an accident occurs in a big city with a vast population.

Airbags: number of airbags can reduce the risk of death.

Model: some cars have more security, and some have less, so it's essential to which model of the car you insure.

2. Question 1

Chosen Numerical Variable : age_of_car: Normalized age of the car in years.

2.1 A

Plot a histogram with an appropriate bin size, then overlay that with the curve of a density, is the distribution approximately normal?

I use two different methods to estimate the number of bins.

1. Freedman Diaconis rule:
The general equation for the rule is: $2\frac{IQR(x)}{\sqrt[3]{n}}$

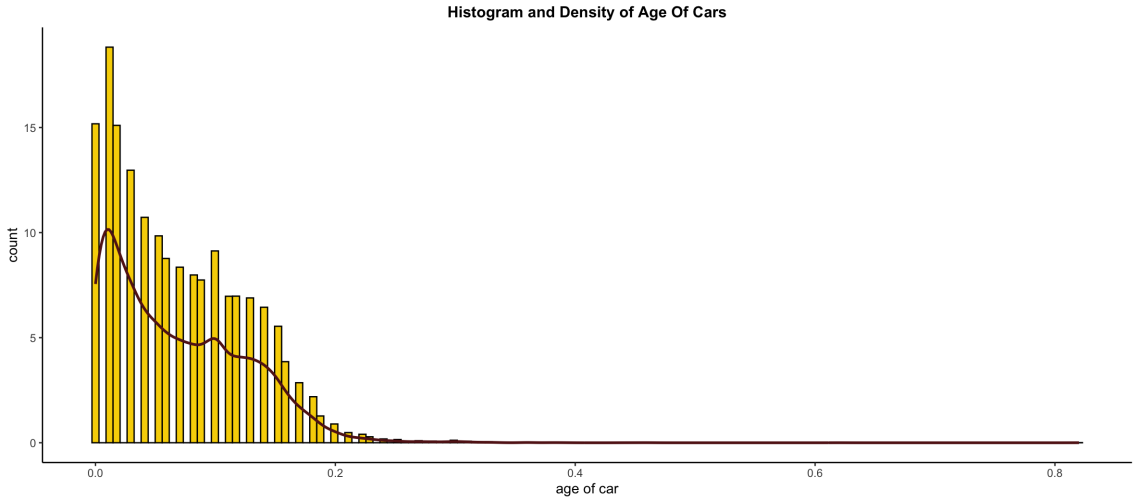

Figure 5: Histogram of Age Of Car with Freedman Diaconis rule bandwidth

2. $\sqrt[2]{n}$ rule
The number of bins corresponds to the square root of the number of observations

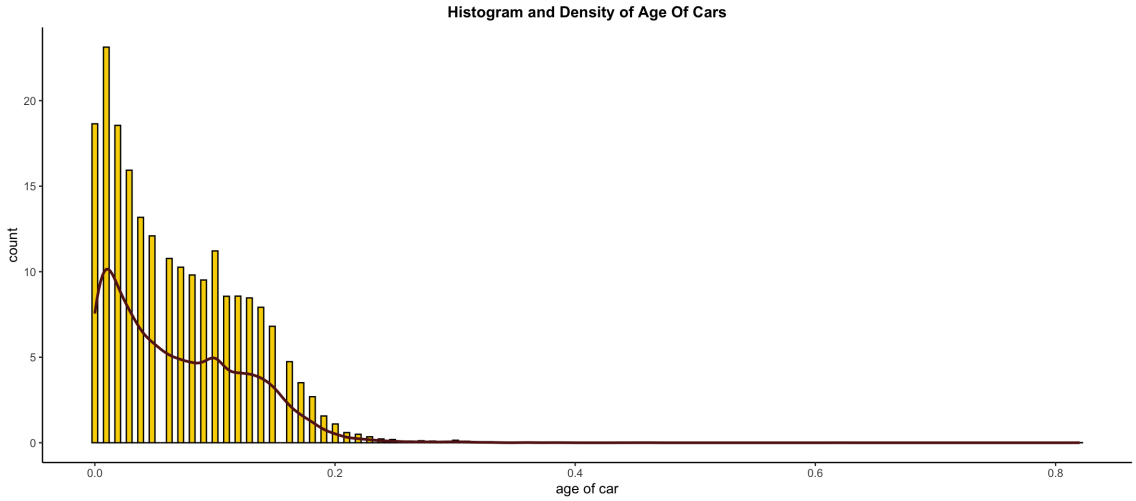

Figure 6: Histogram of Age Of Car with $\sqrt[2]{n}$ rule bandwidth

3. default in R

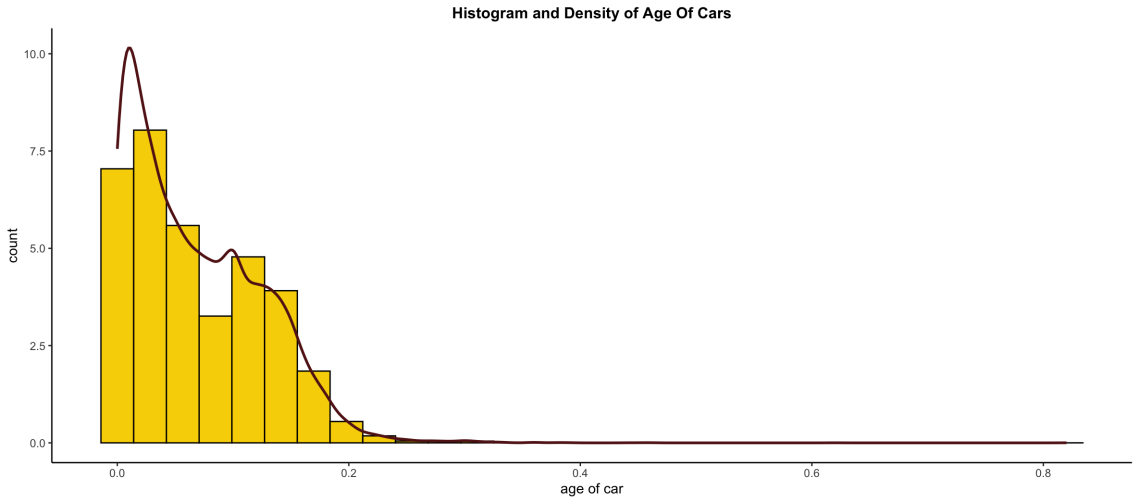

Figure 7: Histogram of Age Of Car with default in R bandwidth

It seems to be the default setting in r is a better choice for select bandwidth, and I think this because the age of the car variable is normalized.
We can see in the histogram that the distribution is highly right-skewed normal

To more accurately compare it to a normal distribution, we use QQ Plot.

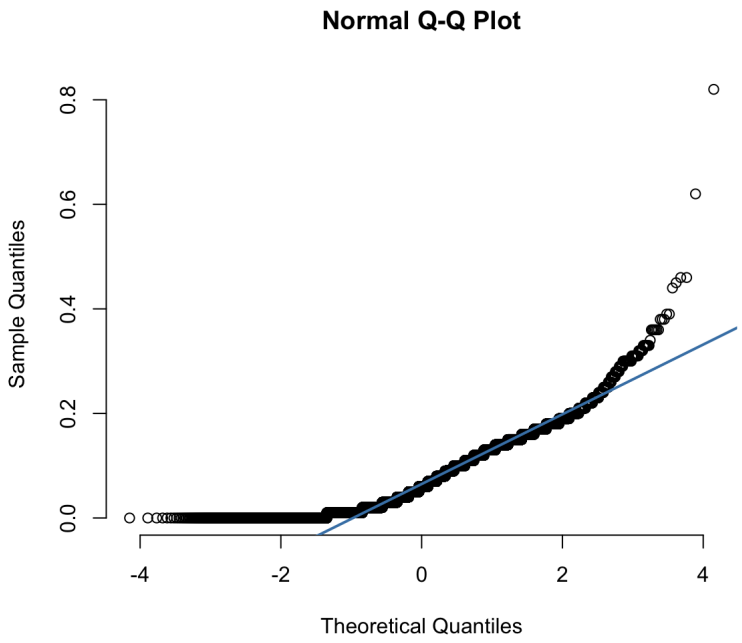

Figure 8: Normal Q-Q plot of Age Of Car

This plot confirms that the distribution is definitely right-skewed.

2.2 B

Based on the previous plot in section "A" talk about the modularity and skewness of this variable and describe it.

The plot above shows a right-skewed unimodal distribution which means that there are more recent cars in this dataset.
We rarely have worn-out cars either because, naturally, after a while, cars are replaced with other cars, which must have a different reason, like an accident or people like append changes in their life and so on or insurance companies just insure just new cars.

2.3 C

Determine the upper and lower quartiles, whiskers, and IQR by drawing a boxplot.

```
> quantile(car$age_of_car)
  0%  25%  50%  75% 100%
0.00 0.02 0.06 0.11 0.82
```

Figure 9: Important percentiles

$min = 0$, $lower\ extreme = 0$
$Q1 = 0.02$ so the lower whisker is of length 0.02, and we don't have outlier here.
$Q3 = 0.11$, $max = 0.84$
$upper\ extreme = 0.24$, is not equal max which means 1.5 * IQR + Q3 is 0.24 and the rest of upper part are outliers. the upper whisker is of length 0.13.
$IQR = Q3 - Q1 = 0.9$

```
> boxplot
$stats
     [,1]
[1,] 0.00
[2,] 0.02
[3,] 0.06
[4,] 0.11
[5,] 0.24

$n
[1] 30000

$conf
           [,1]
[1,] 0.05917901
[2,] 0.06082099
```

Figure 10: five-number information about age of car

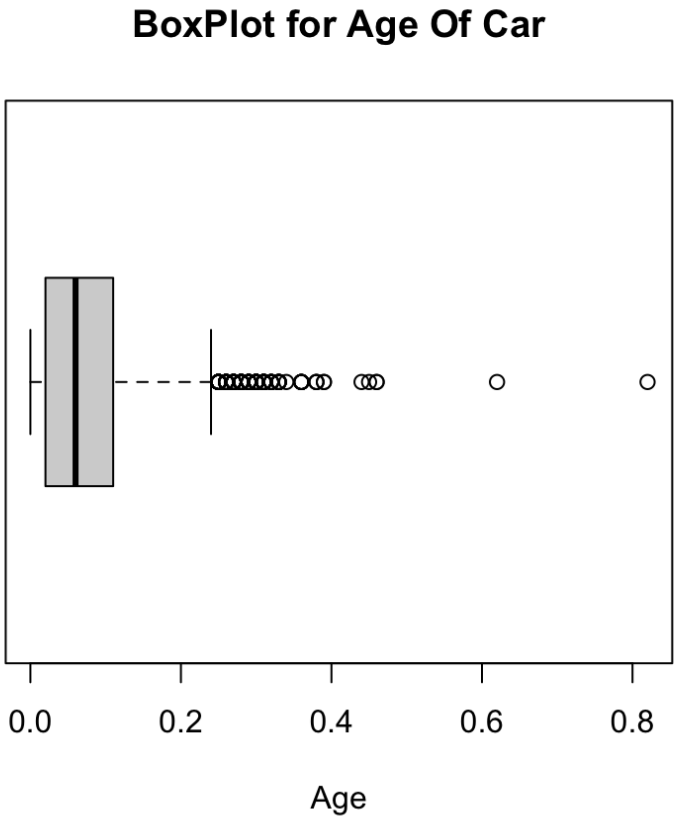

Figure 11: Important percentiles

2.4 D

What are the outliers in this variable? Determine the outliers and their quantity, then try to find what is the meaning of them.

Mention these in the last part. On the left, we haven't outliers, but on the right, we have 147 outliers which means 147 out of 30,000 cars are relatively worn out.

```
> aoc <- car$age_of_car
> length(aoc[which(aoc < boxplot$stats[1] | aoc > boxplot$stats[5])])
[1] 147
```

Figure 12: Important percentiles

Because the numbers of the age of cars are normalized, it is impossible to express the actual number, but if the age of your car is more than 0.24, it means that your car has a longer lifespan compared to all existing cars or If it doesn't work correctly, it is worn out.

2.5 E

Calculate the mean, median, variance, and standard deviation and describe each one.

Mean : mean is also known as average of all the numbers in the data set which is calculated by below equation.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

So the sum of cars' age is divided by the population's size.

Median : median is mid value in this ordered data set.
So, we can see half of the cars are 0.06, which means that almost cars are new.

Mean is larger than median so the distribution is right-skewed.
Variance : variance is the numerical values that describe the variability of the observations from its arithmetic mean.
Variance measure how far individuals in the group are spread out, in the set of data from the mean.

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

Standard Deviation : It is a measure of dispersion of observation within dataset relative to their mean.It is square root of the variance.
Standard deviation is expressed in the same unit as the values in the dataset so it measure how much observations of the data set differs from its mean, on average, data have 0.056, far from the mean.

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

```
> mean(car$age_of_car)
[1] 0.06948467
> median(car$age_of_car)
[1] 0.06
> var(car$age_of_car)
[1] 0.003199781
> sd(car$age_of_car)
[1] 0.05656661
```

Figure 13: Some Statistics for Age Of Car

2.6 F

Categorize this variable into four intervals based on its mean and plot a pie chart that visualizes the frequency of these four categories. Your chart should be colorized, and the labels should contain each category with its percentage.

Our data are categorized according to a range of values.

$$\begin{cases} min \leq \ \leq \frac{\mu}{2} & First \\ \frac{\mu}{2} \leq \ < \mu & Second \\ \frac{\mu}{2} \leq \ < \frac{3\mu}{4} & Third \\ \frac{3\mu}{4} \leq \ \leq max & Fourth \end{cases}$$

```r
aoc <- car$age_of_car
mean = mean(aoc)

Fo = aoc[aoc > .75*mean]
Th  = aoc[aoc >= .5*mean & aoc<= .75*mean]
Se = aoc[aoc < .5*mean & aoc >= .25*mean]
Fi = aoc[aoc < .25*mean]

frequencies<-c(length(Fo),length(Th),length(Se),length(Fi))
percentage  <- round(100*frequencies/sum(frequencies), 2)

aoc.categorized <- data.frame(names = c("Fourth", "Third", "Second","
    First"),value = percentage)

ggplot(aoc.categorized, aes(x="", y = value, fill = names)) +
  geom_bar(stat = "identity") + coord_polar("y") +
  geom_text(aes(label = paste0(value, "%")),
            position = position_stack(vjust = 0.5)) +
  labs(title="Age Of Car Pie Chart", x = 'Frequency', y = 'Age Of Car'
    )
```

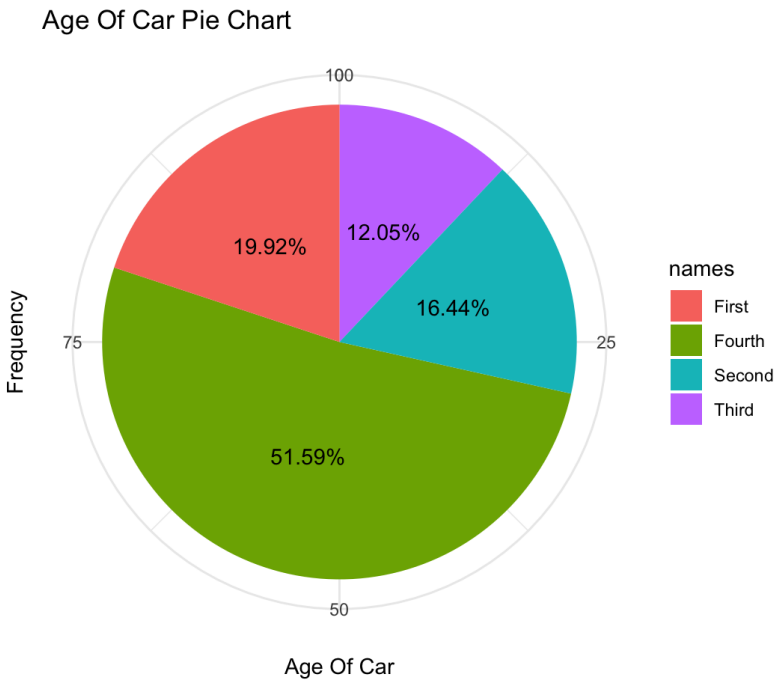

Figure 14: Age of Car Pie Chart

2.7 G

Draw a density plot of this variable and add lines for the mean and median to it. What is the relationship between the mean, median, and density of this variable?

```
ggplot(car, aes(x = age_of_car)) + geom_density( size = 1) +
  geom_vline(aes(xintercept = median(age_of_car)), linetype = "dashed"
   , size = .7,col= 'red') +
  geom_vline(aes(xintercept = mean(age_of_car)), linetype = "dashed",
   size = .7,col= 'green') +
  theme_classic() +
  theme(
    plot.title = element_text(size = 12, face = "bold", hjust = 0.5),
    plot.caption = element_text(face = "italic")
  )
+
  annotate("text", x = mean(car$age_of_car) + .02 , label = "mean", y
   = 6, size = 3, angle = 90 , color = 'green') +
  annotate("text", x = median(car$age_of_car) - 0.02 , label = "median
   ", y = 0.6, size = 3, angle = 90, color = 'red')
+
  labs(
    title = "Age of Car Density",
    caption = "green line is about mean and red is about median
   variable on the diagram.",
    x = "Age Of The Car",
    y = "Count"
  )
```

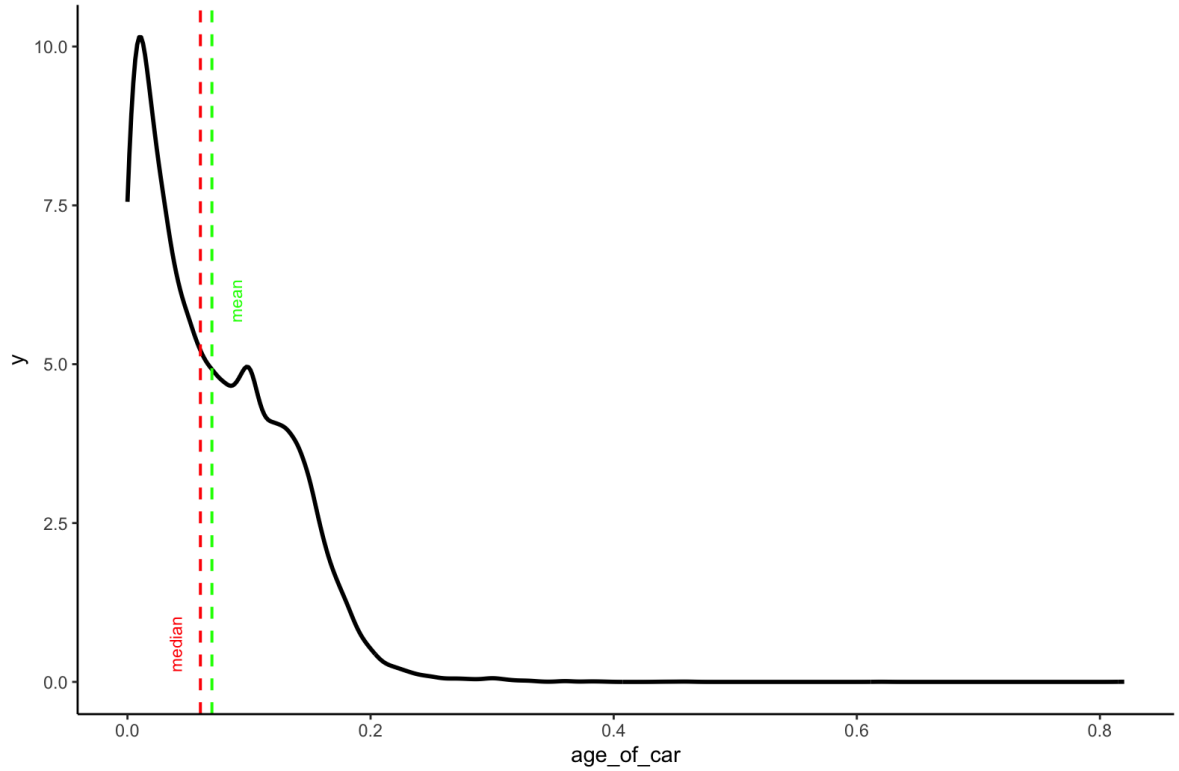

Figure 15: Age of Car Density diagram

Mean is larger than median so the distribution is right-skewed.
We can see this fact by calculating skewness:

```
> skewness(car$age_of_car)
[1] 0.9014617
```

Figure 16: skewness

It's closed to 1 so we have strong right-skewed.

3. Question 2

Chosen Categorical Variable : transmission_type: Transmission type of the car

3.1 A

Create a frequency table for this variable.

```
transmission <- car$transmission_type
gp <- unique(transmission)

Manual = transmission[transmission =='Manual']
Automatic = transmission[transmission =='Automatic']

frequencies<-c(length(Manual),length(Automatic))
percentage  <- round(100*frequencies/sum(frequencies), 2)

transmission.categorized <- data.frame(types = c("Manual", "Automatic"
   ),percentage = percentage, value = frequencies)
```

	types	percentage	value
1	Manual	64.99	19498
2	Automatic	35.01	10502

Figure 17: frequency table for transmission_type.

As you can see, most of the cars are still manual. It's about 64% of the population.

3.2 B

Plot a horizontal bar plot, sort by frequency, and use different colors for each category.

```
transmission.categorized <- transmission.categorized[order(percentage)
    ,]
ggplot(data=transmission.categorized, aes(x=types, y=frequencies, fill
    =types)) +
  geom_bar(stat="identity", alpha = 0.7, width = 0.7)+
  labs(title="Barplot of Transmission Type")+
  coord_flip()
```

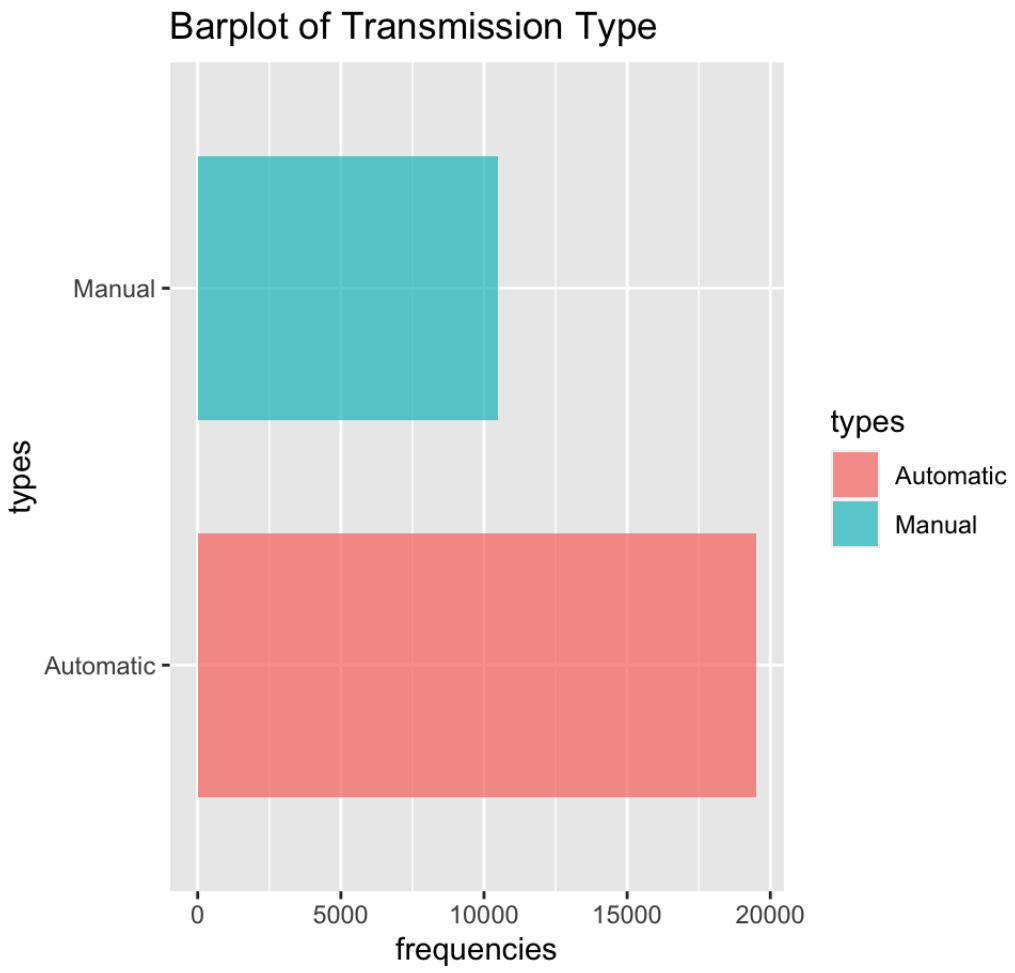

Figure 18: Horizontal Bar Plot of Transmission Type

3.3 C

Plot a bar plot for this variable and add percentage marks to it.

```
ggplot(data=transmission.categorized, aes(x=types, y=percentage, fill=
    types)) +
  geom_bar(stat="identity", alpha = 0.7, width = 0.6)+
  labs(title="Barplot of Transmission Type")+
  geom_text(aes(label=paste(percentage, "\%")),vjust=-0.4,size=4)
```

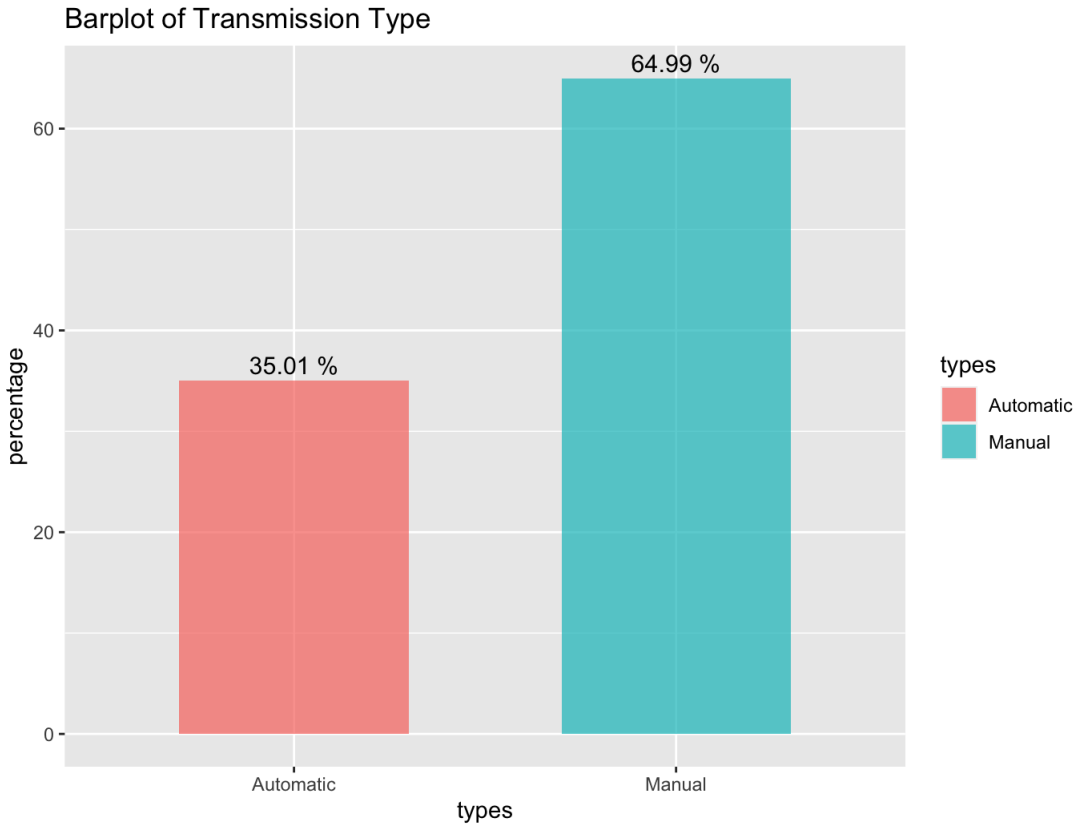

Figure 19: Bar Plot of Transmission Type

3.4 D

Plot a violin plot for this variable.

```
ggplot(data = car, aes(x=transmission_type, y=age_of_car, fill=
    transmission_type))+
  geom_violin( trim=FALSE, alpha = 0.7)+
  labs(title="Violin Plot of age of car by transmission type")
```

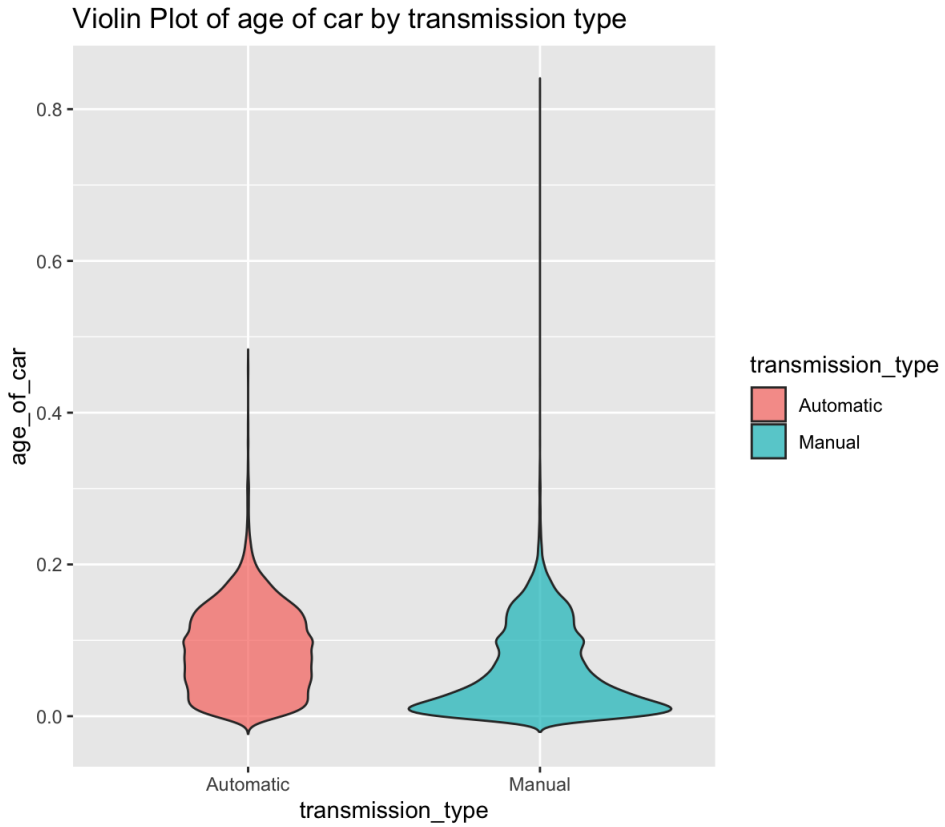

Figure 20: Violin Plot

Automatic cars have mostly normal distribution and show that these models produce with a fixed range during the time.

Manual cars recently produced or bought more.

4. Question 3

Chosen Numerical Variable : age_of_car and age_of_policyholder

4.1 A

Draw a scatter plot for two variables and describe the relationship between them based on the plot.

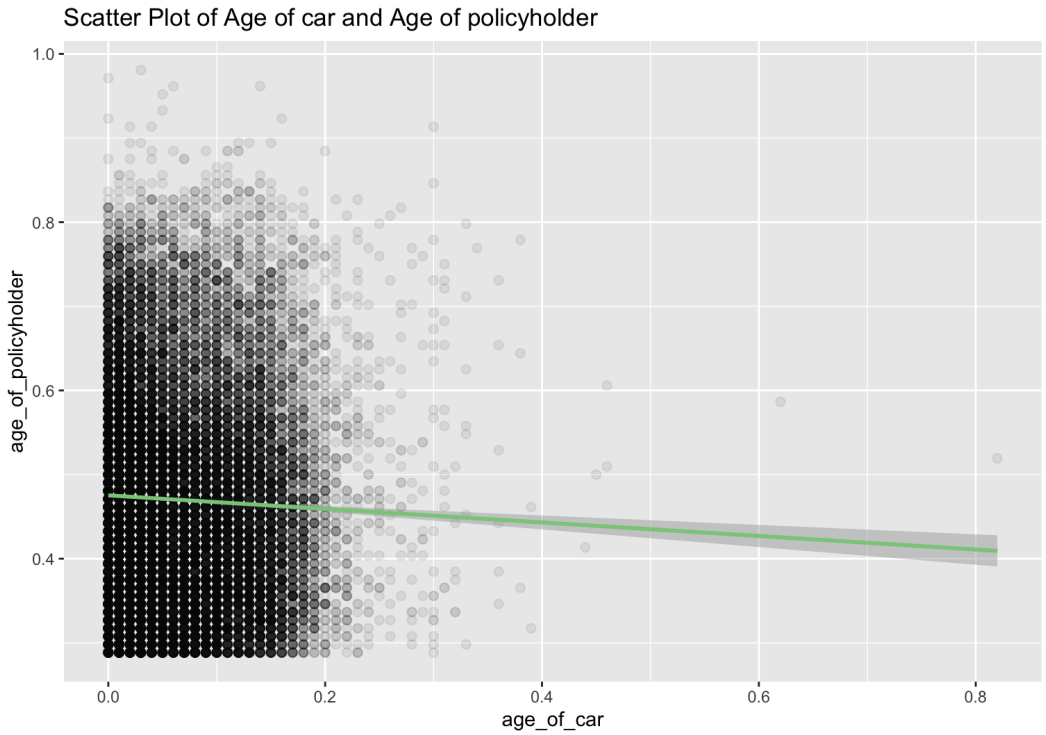

Figure 21: Scatter plot

As you can see, the relationship between these two variables is not easy. We can't describe a clear function between them. Some outliers are detected.
Because both of my variables represent age, the majority of points can be found in a specific area between 0 to 0.2 for cars' age and 0 to 0.8 for policyholders' age.
About outliers: people who are old can't drive, and old cars should be scrapped, so they occur in my dataset rarely.

4.2 B

Select a categorical variable, and determine the samples either by the symbol or by the color (or by both) in a scatter plot that has been drawn in section "A". Does the relation still hold for different categories?

Select fuel type as categorical variable.

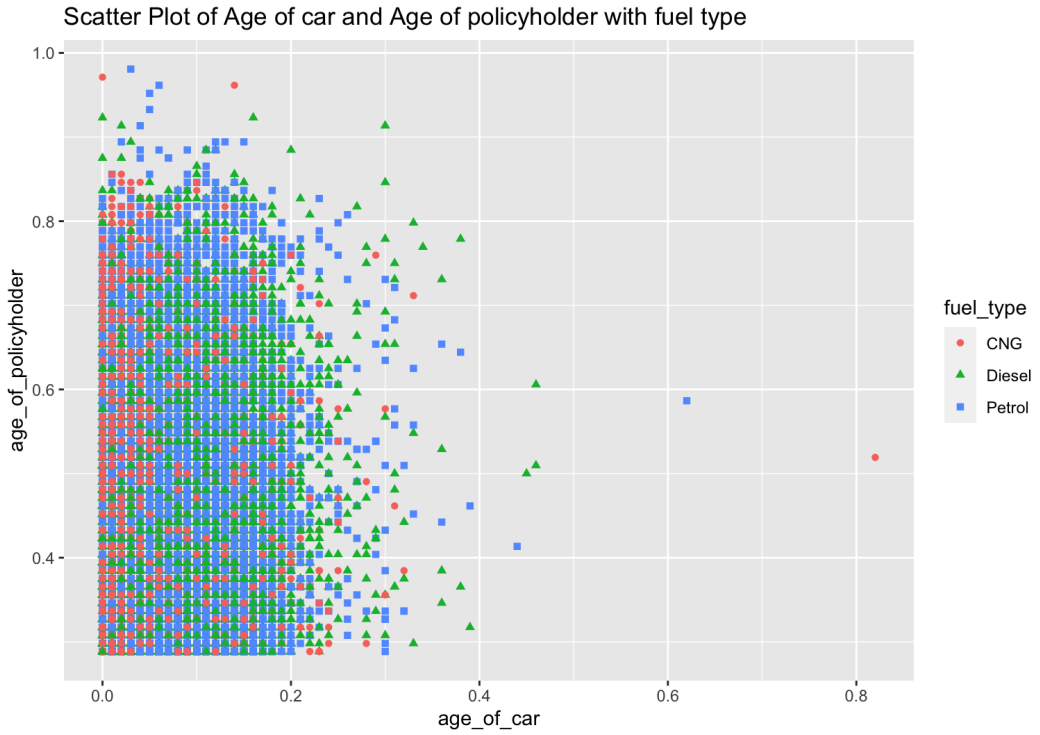

Figure 22: Scatter plot

Because there was no relationship in the previous part, this is not observed either. It can only be said that newer cars use CNG fuel.

4.3 C

Calculate the correlation coefficient for these two variables. Using the "cor.test" function, we can also test the significance of a correlation. Are the variables correlated? According to the test, what is shown by the p value, and what is the intuition of the p value?

```
> cor.test(car$age_of_car, car$age_of_policyholder)

        Pearson's product-moment correlation

data:  car$age_of_car and car$age_of_policyholder
t = -6.3982, df = 29998, p-value = 1.595e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.04821195 -0.02561091
sample estimates:
        cor
-0.03691615
```

Figure 23: Scatter plot

As expected, there is no correlation between these two variables, and it is close to zero.
P value define on hypothesis test.
On average, is there a relationship between two means of these variables?
H0 : means are close two each other.
Ha : means are not close two each other.

And we can see p value is close to 0, so we reject H0.
These two expressions have the same Concept in this context, the population has no powerful linear correlation between age of car and age of policyholder.

4.4 D

A hexbin (Figure 1) plot with marginal distribution is like a two-dimensional histogram. The data is divided into bins, and the color strength of each bin represents the number of data points in that bin. Also, each dimension has its own distribution in front of its axis. Draw the hexbin plot with marginal distribution for chosen variables. What is your interpretation? Discuss the bin size and how it changes the result.

```
p12 <- ggplot(car, aes(age_of_car, age_of_policyholder)) +
  geom_point() +
  geom_hex(bins = 12) +
  geom_smooth(color='#FFFF00', se=FALSE) +
  ggtitle("Hexbin Plot - Age of car and policyholder with binsize 12")

ggMarginal(p12, type="histogram", size=3, fill='lightblue')

```

Figure 24: hexbin plot with bin size 7

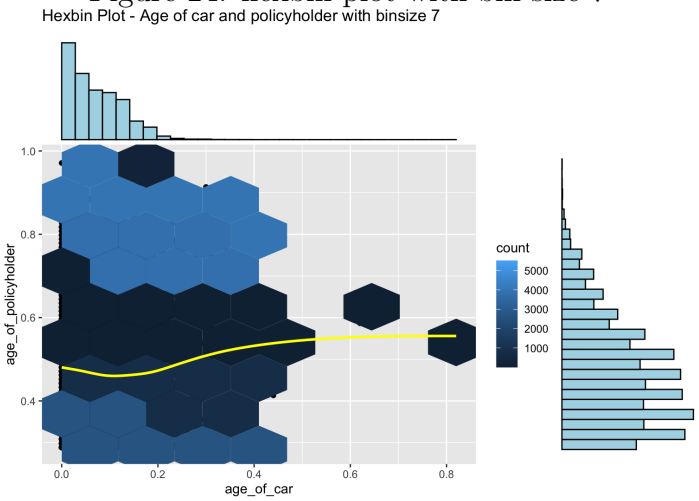

Figure 25: hexbin plot with bin size 12

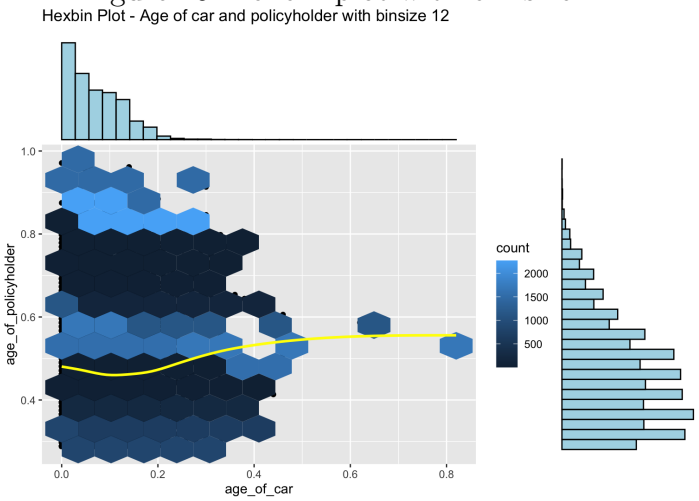

Figure 26: hexbin plot with bin size 30

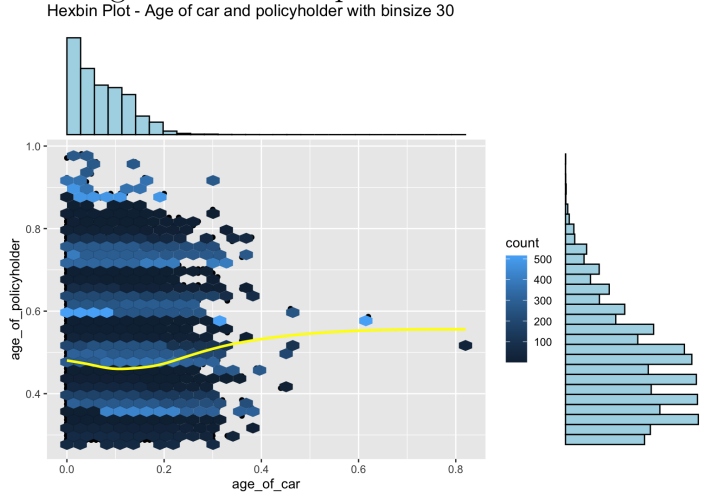

Figure 27: hexbin plot of age_of_car and age_of_policyholder with different bin size.

This graph is used when the number of samples is large, and the scatter plot cannot determine the distribution of the data locally.
In this diagram, the coordinate space is divided into bins, where the number of samples in each bin is indicated by color.
So we can see some interesting information, for example, older people who are policy-holders support more cars and, in other words, have more consumers, and this may be because of their experience in this market or people trust them more because of their age.

About bin size, when the number of bins increases, the hexbin size decreases, and the hexbin becomes sparser.
According to the figure, it seems bin size 12 is more informative, and when increasing bin size, this graph is close to a scatter plot.

4.5 E

Draw the 2D density plot for chosen variables. How do you interpret the resulting graph? Describe the advantages and disadvantages of the 2D density and hexbin graph.

Density can be represented in the form of 2D density graphs or density plots.
A 2d density chart displays the relationship between 2 numeric variables, where one variable is represented on the X-axis, the other on the Y axis, like for a scatterplot.

```
ggplot(car, aes(x=age_of_car, y=age_of_policyholder) ) +
  stat_density_2d(aes(fill = ..level..), geom = "polygon") +
  scale_fill_continuous(type = "viridis") +
  labs(title = "2D density plot Age of car and Age of policyholder")

```

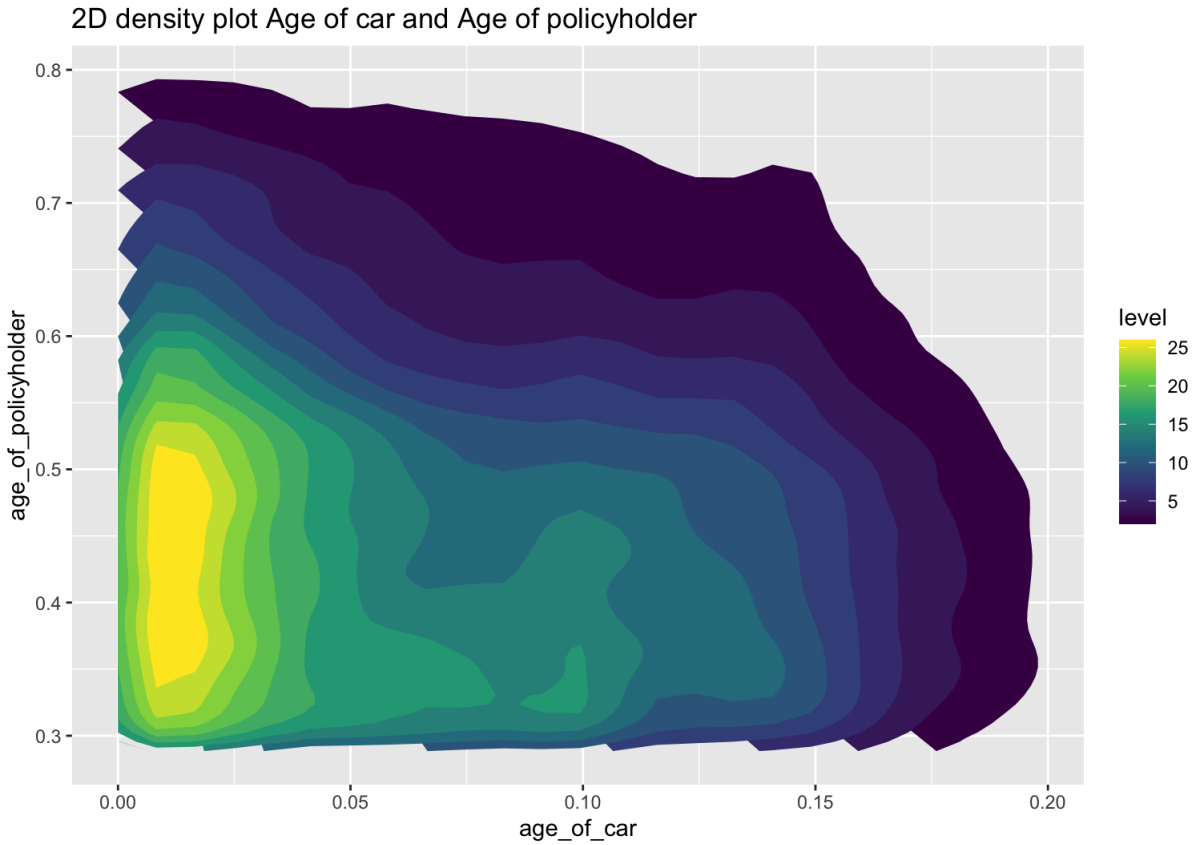

Figure 28: 2D density plot

As can be concluded from the graph, this data and relation is too depth, and learning this is so hard.

In hexbin, the number of bins can be controlled, which changes the shape of the distribution but not in the density plot.
In density and hexbin plots can see the pick of data but in hexbin show the local distribution in a better way by use color.

5. Question 4

5.1 A

Create a heatmap correlogram from your variables.
Annotate each cell with their corresponding Pearson's correlation coefficients and p-value as well.
Use red for the positive correlation and blue for the negative correlation.
Highlight significant correlations

```
#some preprocessing
car$is_claim <- as.logical(car$is_claim)
car <- subset(car, select = -c(X))

#separate numerical features
car.numeric <- Filter(is.numeric, car)

#correlation matrix
cor_mat <- cor(as.matrix(car.numeric), method="pearson")
cor_mat <- round(cor_mat, 2)

#calc p_values
p.mat <- cor_pmat(car.numeric)

ggcorrplot(
  cor_mat, hc.order = TRUE, type = "upper",
  lab = TRUE , p.mat = p.mat, sig.level = 0.05,
  title = "heatmap correlogram",
  colors = c("blue", "white", "red"),
  hc.method = "complete", lab_size = 2,
  as.is = FALSE)

```

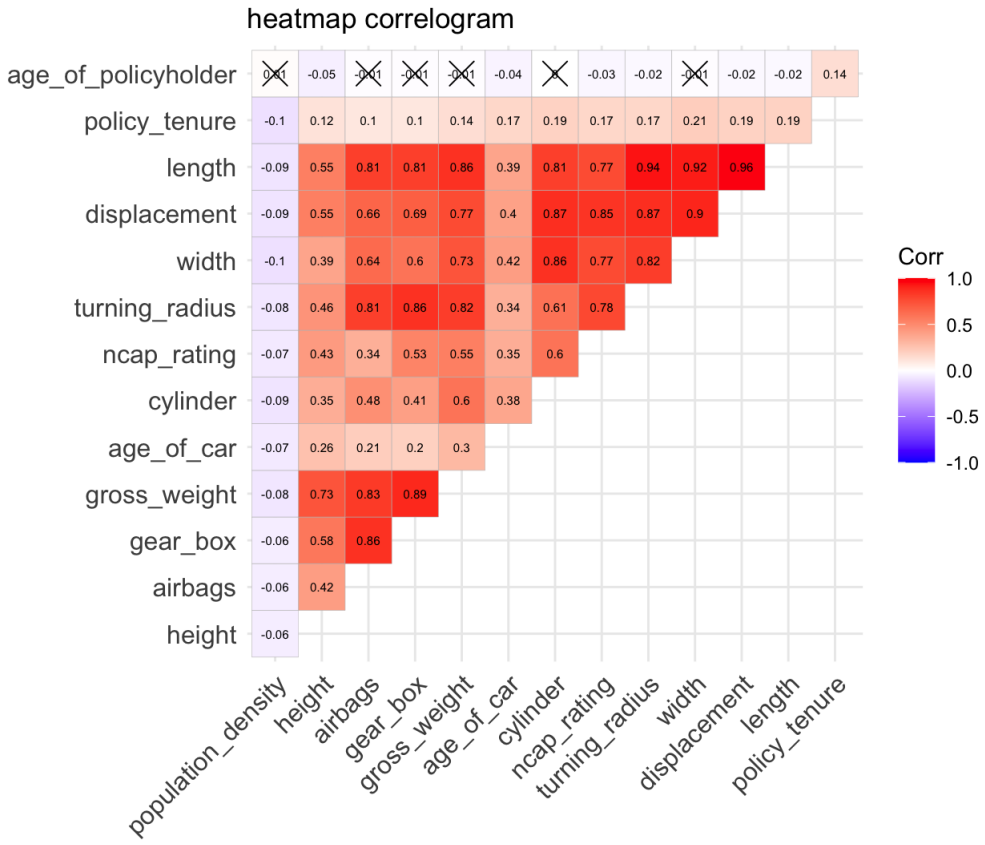

Figure 29: heatmap correlogram for numeric features

$p - value > 0.05$ show with red box.

	policy_tenure	age_of_car	age_of_policyholder	population_density	airbags	displacement	cylinder	gear_box	turning_radius	length	width	height	gross_weight	ncap_rating
policy_tenure	0.000000E+00	8.161691E-187	4.944304E-138	7.665155E-65	1.356004E-70	1.098616E-251	3.290997E-238	2.883208E-61	4.609704E-187	2.386836E-242	2.315608E-299	6.097440E-94	6.375608E-127	6.443607E-203
age_of_car	8.161691E-187	0.000000E+00	1.595233E-10	8.414501E-34	7.760921E-291	0.000000E+00	0.000000E+00	2.962429E-270	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00
age_of_policyholder	4.944304E-138	1.595233E-10	0.000000E+00	2.023697E-01	5.564222E-02	3.051618E-05	4.615045E-01	3.789340E-01	2.807660E-03	2.293577E-04	2.574386E-01	4.039960E-20	1.888151E-01	1.402864E-07
population_density	7.665155E-65	8.414501E-34	2.023697E-01	0.000000E+00	2.965172E-29	3.844415E-56	3.280059E-57	9.925498E-23	2.903633E-41	1.885420E-57	9.383007E-64	7.551083E-29	5.423320E-43	6.334509E-32
airbags	1.356004E-70	7.760921E-291	5.564222E-02	2.965172E-29	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00
displacement	1.098616E-251	0.000000E+00	3.051618E-05	3.844415E-56	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00
cylinder	3.290997E-238	0.000000E+00	4.615045E-01	3.280059E-57	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00
gear_box	2.883208E-61	2.962429E-270	3.789340E-01	9.925498E-23	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00
turning_radius	4.609704E-187	0.000000E+00	2.807660E-03	2.903633E-41	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00
length	2.386836E-242	0.000000E+00	2.293577E-04	1.885420E-57	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00
width	2.315608E-299	0.000000E+00	2.574386E-01	9.383007E-64	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00
height	6.097440E-94	0.000000E+00	4.039960E-20	7.551083E-29	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00
gross_weight	6.375608E-127	0.000000E+00	1.888151E-01	5.423320E-43	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00
ncap_rating	6.443607E-203	0.000000E+00	1.402864E-07	6.334509E-32	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00

Figure 30: report p values

Most of the features have little p value.

5.2 B

Display all the bivariate relations between the variables using a correlogram where each element is a scatter plot between two variables. Can you find any meaningful pattern between them?

Select some important features because the computaion is high.

```
selected_var = data.frame("policy tenure" = car$policy_tenure,
    "car's age" = car$age_of_car,
    "policyholder's age"=car$age_of_policyholder,
    "population density"= car$population_density,
    "displacement"=car$displacement,
    "turning radius"=car$turning_radius,
    "length"=car$length,
    "width"=car$width,
    "height"=car$height,
    "gross_weight" = car$gross_weight)

ggpairs(selected_var, title = "Correlogram",
        upper = list(continuous = wrap("density", alpha = 0.3)),
        lower = list(continuous = wrap("points",colour="grey", alpha =
    0.3), combo = "dot_no_facet"))
```

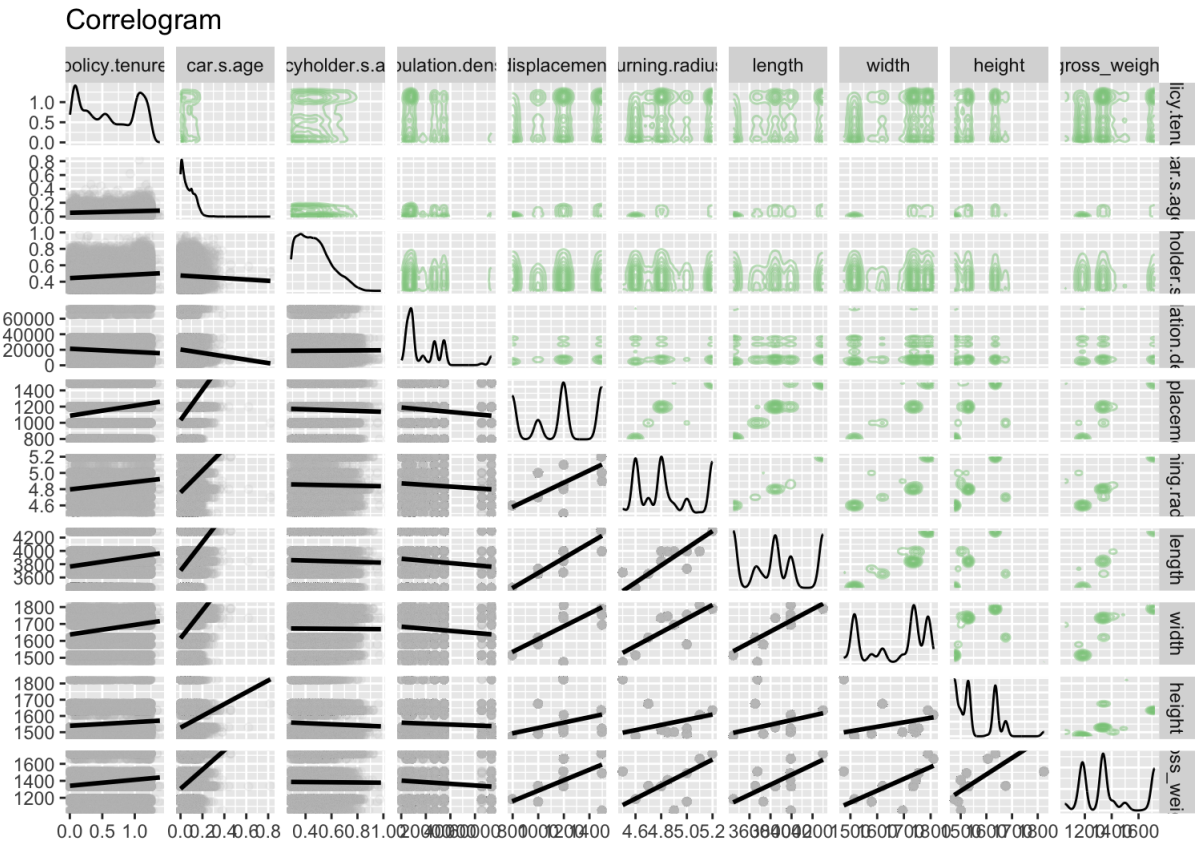

Figure 31: correlogram

Some conclusions:
Age has right-skew distribution.
As expected, width, length, and height have a positive correlation. the higher the width, the higher the length and height, and some categories, like driving cars and cranes.
Policy tenure has two picks because of short-term and long-term options for consumers.
Population show we have one crowded and two medium population and others are Low population.
Gross weight has a positive correlation with the physical features of the cars.
The length has a high tail, and the width has a low tail.

5.3 C

Choose 3 numerical and 1 categorical variable from your dataset. Draw a 3D scatter plot for the numerical variables and use the categorical variable as the points' color. Describe the relation between them

Choose population-density, age-of-policyholder and displacement as numerical variable and fuel-type as categorical.

```
colors <- c("#999999", "#E69F00", "#56B4E9")
fuel_type <- colors[as.numeric(factor(car$fuel_type))]

scatterplot3d(
  main="3D Scatter Plot",
  car$population_density,
  car$age_of_policyholder,
  car$displacement,
  color=fuel_type,
  xlab = "population_density",
  ylab = "age_of_policyholder ",
  zlab = "displacement",
  pch = 16)

legend("right", legend = unique(car$fuel_type), pch = 16, col =
   colors)
```

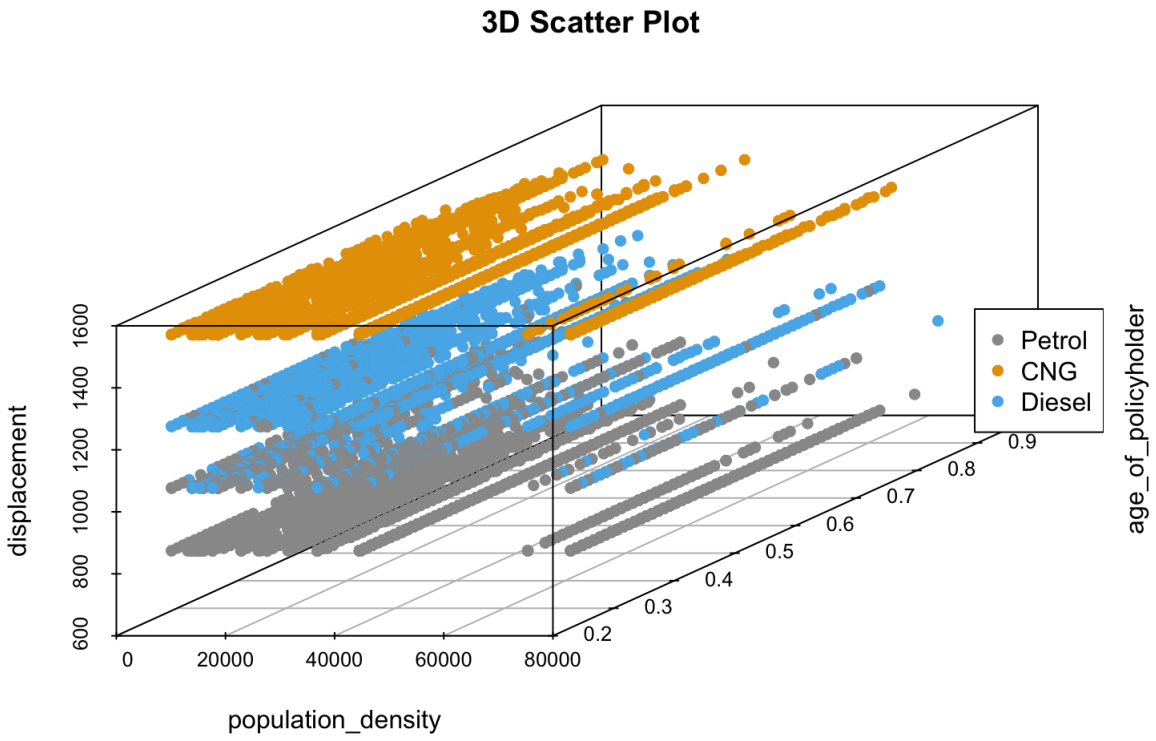

Figure 32: 3D scatter plot

As you can see, the type of fuel and displacement have a good relationship, cars with low displacement use diesel and cars with medium displacement use petrol, and cars with high displacement use CNG.

Cars like trucks usually use CNG, which drives between the city and travels a long distance, and they don't pollute the air.
Cars in the town typically use diesel fuel because their price is low, like vans or mini trucks.
And people use petrol fuel for daily transportation, which usually has low displacement.

6. Question 5

Chosen Categorical Variables : fuel_type and transmission_type

6.1 A

Contingency table

```
Total <- sum
t<-as.data.frame.matrix(addmargins(table(car[,c('fuel_type', '
    transmission_type')]), FUN = Total))

```

	Automatic	Manual	Total
CNG	0	10370	10370
Diesel	7171	1931	9102
Petrol	3331	7197	10528
Total	10502	19498	30000

Figure 33: Contingency table of fuel_type and transmission_type

Automatic cars don't use CNG.
Diesel is used in the automatic car more than manual ones.
This relation is revers for petrol fuel.

6.2 B

Grouped Bar Charts

```
ggplot(car, aes(x =fuel_type , fill = transmission_type)) +
  geom_bar(position = "dodge", alpha = 0.6)  +
  geom_text(aes(label = ..count..), stat = "count", vjust = -.3,  size
    = 3, color = 'black', position = position_dodge(width = 1)) +
  labs(title="Grouped Bar Chart", x="fuel_type")

```

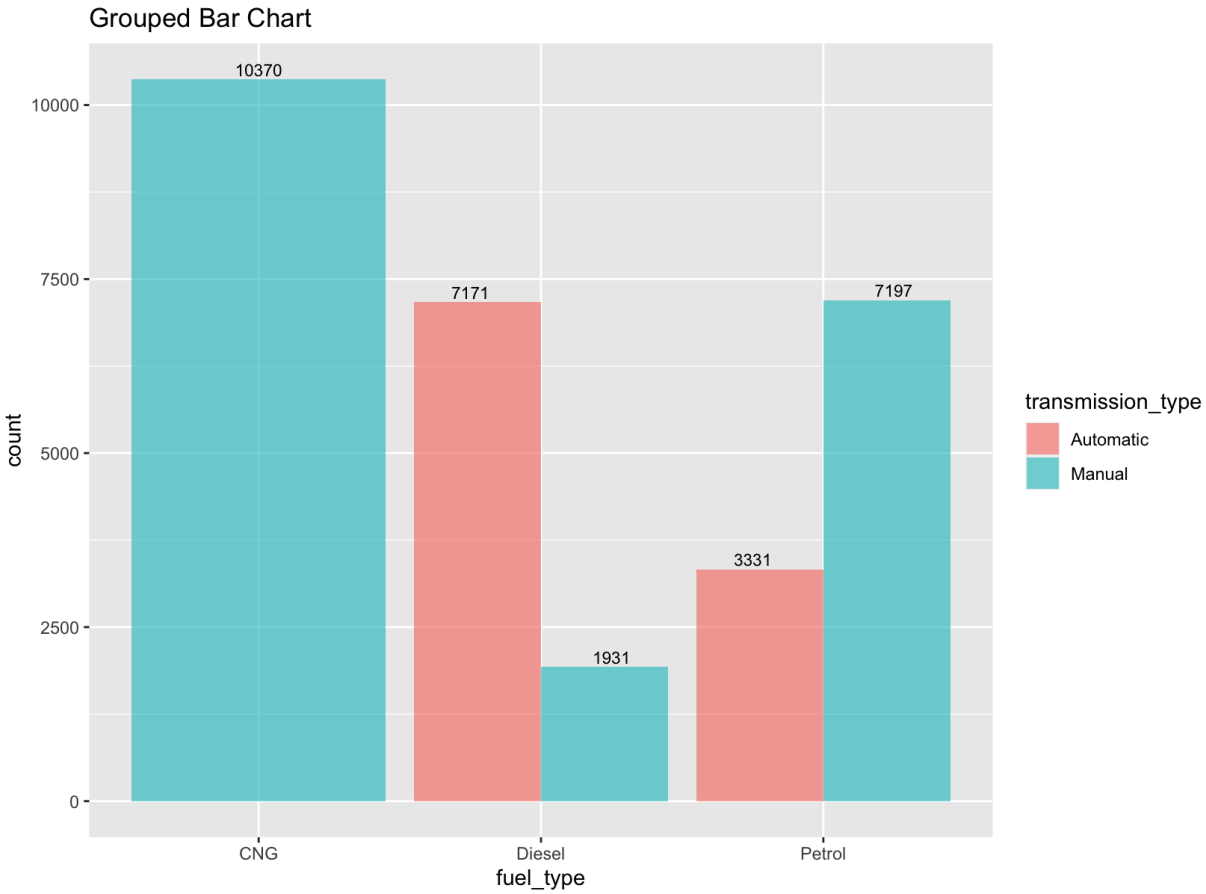

Figure 34: Grouped Barplot of fuel_type and transmission_type

6.3 C

Segmented Bar Chart

```
ggplot(car, aes(x = fuel_type, fill = transmission_type)) +
  geom_bar(position = "stack", alpha = 0.6) +
  geom_text(aes(label=..count..), stat='count', vjust = -.3,  size =
   3, color = 'black', position = position_stack(0.5)) +
  labs(title=paste("Segmented barplot"))
```

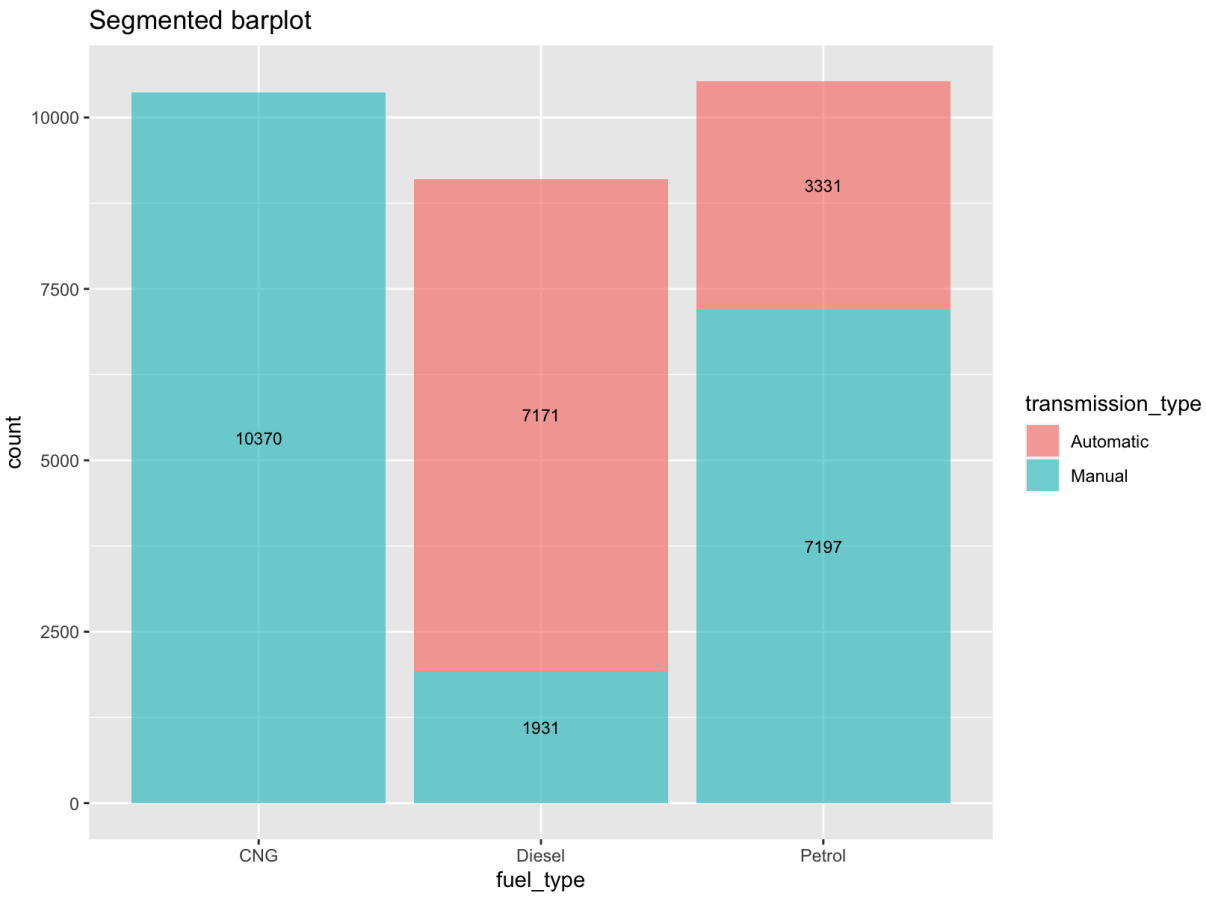

Figure 35: Segmented Barplot of fuel_type and transmission_type

6.4 D

Mosaic Plott

```
ggplot(data = car) +
  geom_mosaic(aes(x = product(transmission_type), fill = fuel_type ),
   alpha = 0.6) +
  annotate("text", x = 0.195, y = 0.35,label=paste(round(t[2,1]/t
   [4,1],2)*100,"\%"),fontface = "bold" ,size=4,color="black") +
  annotate("text", x = 0.195, y = 0.855,label=paste(round(t[3,1]/t
   [4,1],2)*100,"\%"), fontface = "bold",size = 4,color="black") +
  annotate("text", x = 0.7, y = 0.25,label=paste(round(t[1,2]/t
   [4,2],2)*100,"\%"),fontface = "bold" ,size = 4,color="black") +
  annotate("text", x = 0.7, y = 0.57,label=paste(round(t[2,2]/t
   [4,2],2)*100,"\%"), fontface = "bold",size = 4,color="black") +
  annotate("text", x = 0.7, y = 0.8,label=paste(round(t[3,2]/t[4,2],2)
   *100,"\%"), fontface = "bold",size = 4,color="black") +
  labs(title=paste("Mosaic plot"))

```

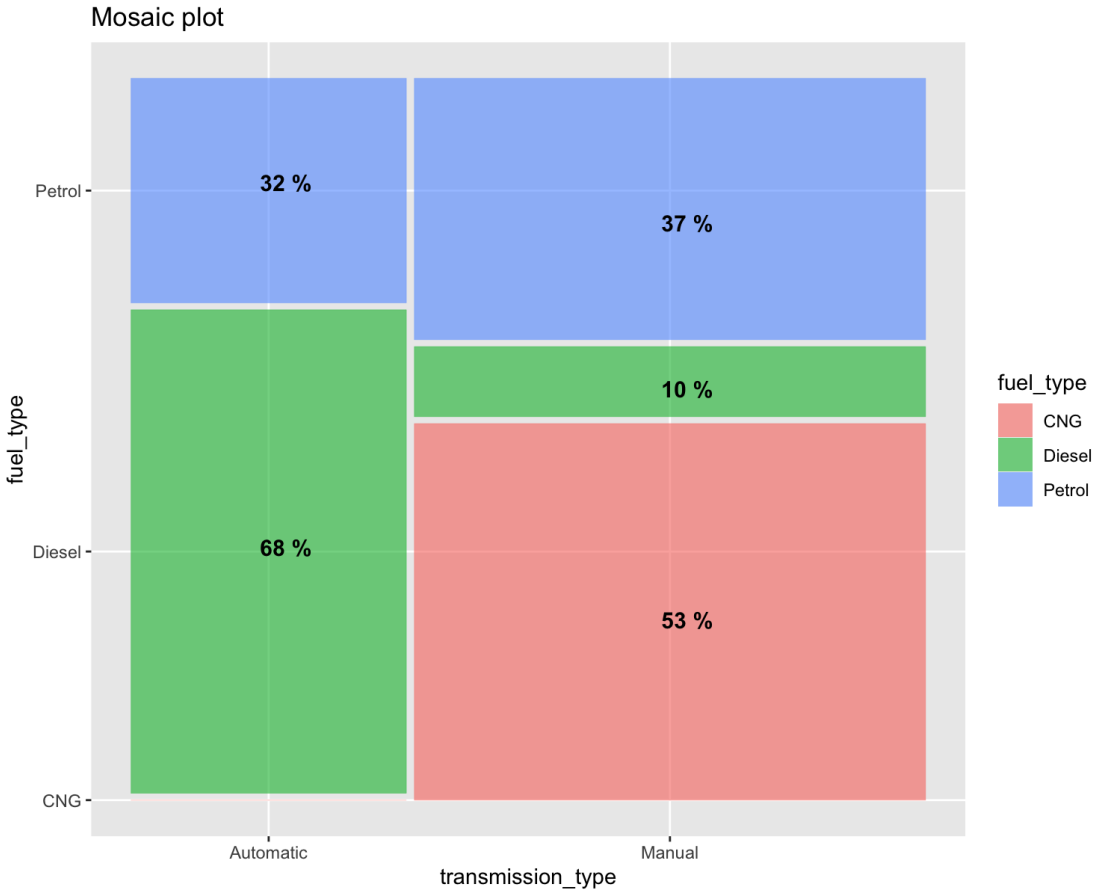

Figure 36: Mosaic plot of fuel_type and transmission_type

7. Question 6

In this question, you will conduct a hypothesis test for two numerical variables.
Choose a random sample of 25 data points from the dataset and choose two numerical variables that are not of a corresponding quantity.
This data can be used to compare the average quantity between the two variables.

Chosen Numerical Variables : displacement and policy_tenure

7.1 A

What is the best method for testing our hypothesis? Is it a t-test or a z-test? Explain it.

$n = 25$

Checking conditions:

1. Independence: random sample and $n < 10\%$ of all population (30000)
* Suppose the tenures of policies are independent because they could be dependent on one another. For example, a company has two types of insurance policies for insuring cars in a small town, but we assume, they are independence. 2. Sample size/skew: $n < 30$.

So, we can't assume that the sampling distribution from this dataset will be nearly normal.
But we can use Student's t-distribution for this problem.

7.2 B

Design a hypothesis test to see if these data provide convincing evidence of a difference between mean values. Does the result agree with the 95% confidence interval?

Conditions for inference for comparing two independent means

1. Independence:
1.1 within groups: sampled observations must be independent
1.1.1 random sample/assignment
1.1.2 if sampling without replacement, $n < 10\%$ of population

1.2 between groups: the two groups must be independent of each other (non-paired)

2. Sample size/skew: The more skew in the population distributions, the higher the sample size needed

All conditions are met.

Let's estimating the Mean

Reject $H_0 \rightarrow$, $P_{value} < 0.05$ so H_0 reject,
it means between mean of $\mu_{displacement}$ and $\mu policy\ tenure$ are not equal.

Calculate Confidence interval 95%:

$df = n - 1 = 24$

$$SE_{displacement-policytenure} = \sqrt{\frac{s^2_{displacement}}{n_{displacement}} + \frac{s^2_{policy\ tenure}}{n_{policy\ tenure}}} = 29249.43$$

```
> margin <- qt(0.975,df=n+n-1)*sqrt(sp/n + sp/n)
> margin
[1] 97.20933
>
> lowerinterval <- (xbar1-xbar2) - margin
> lowerinterval
[1] -1197.122
>
> upperinterval <- (xbar1-xbar2) + margin
> upperinterval
[1] -1002.703
>
> cat("95% confidence interval is= (", lowerinterval,",",upperinterval,")")
95% confidence interval is= ( -1197.122 , -1002.703 )
```

Figure 37: Caculate CI 95%

$0 \notin (-1197.122, -1002.703)$

T-test result and 95% confidence interval agree.

8. Question 7

Chosen Numerical Variable : age_of_policyholder

$n = 100,\ \bar{x} = 0.4760577,\ s = 0.01150644$

Checking conditions:

1. Independence: random sample and $n < 10\%$ of all population of people who has car in this dataset($100 < 3000$).
2. Sample size/skew: $n \geq 30$.

Can see histogram of age_of_policyholder below:

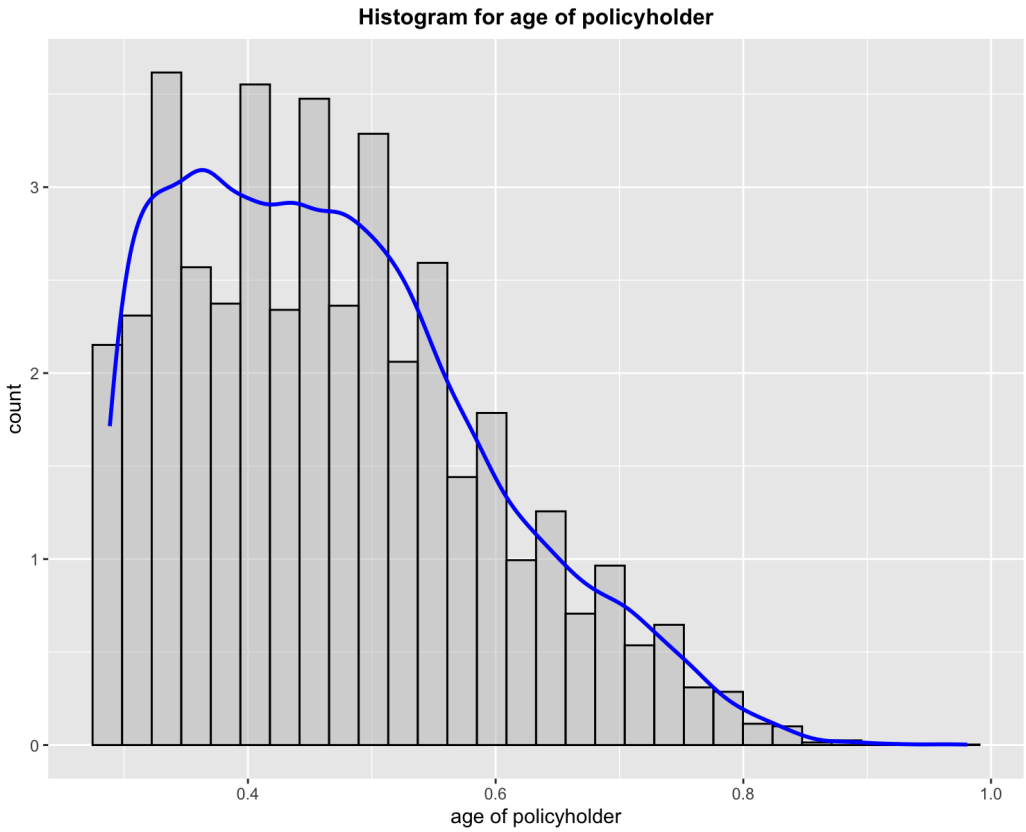

Figure 38: Histogram of age_of_policyholder

We have a right-skewed distribution, and it's far from normal, so for using CLT, we must increase the number of samples by choosing n = 100, and we can assume that the sampling distribution of size 100 will be nearly normal.

Central Limit Theorem (CLT)

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right) = N(0.476057, 0.01150644)$$

8.1 A

Calculate a 98% confidence interval for the mean of this variable

```
n = 100

#reads the dataset 'car' and take the 100 rows as sample
sdf<- sample(1:nrow(car), n)

#sample 10 rows
sub_car <- car[sdf,]

xbar = mean(sub_car$age_of_policyholder)

ci = 0.98
z = qnorm((1-ci)/2)

se <- sd(sub_car$age_of_policyholder) / sqrt(n) #SE = s/sqrt(n)

lower <- xbar + z * se
upper <- xbar - z * se

cat("98\% confidence interval is = (", lower,",",upper,")")

```

approximate 98% CI for μ: 0.4760577 $\pm$2.32 SE

$$\boxed{(0.4492897, 0.5028257)}$$

We are 98% confidence that the mean point of age of policyholder is between \$0.4492897 and \$0.5028257.

8.2 B

Interpret this confidence interval. In this context, what does a 98% confidence level mean?

We are 98% confidence that the mean point of age of policyholder is between $0.4492897 and $0.5028257.

```
> mean(car$age_of_policyholder)
[1] 0.4696782
```

Figure 39: Actual mean of age of policyholder.

As we can see, this claim is valid, and the actual mean is in this interval.

8.3 C

Plot the histogram of the variable and mark the mean of all the samples as a vertical line on top of the histogram plot. You must also mark the confidence intervals on the plot as two vertical lines

```
ggplot(sub_car, aes(x = age_of_policyholder)) +
  geom_histogram(aes(y = ..density..),alpha=0.5, colour = "black")+
  labs(
    title = "Histogram for sample of age of policyholder",
    x = "age of policyholder",
    y = "count"
  ) +
  geom_vline(aes(xintercept = mean(age_of_policyholder)), linetype = "
   dashed", size = .7,col= 'red') +
  geom_vline(aes(xintercept = lower), linetype = "dashed", size = .7,
   col= 'green') +
  geom_vline(aes(xintercept = upper), linetype = "dashed", size = .7,
   col= 'blue') +
  geom_vline(aes(xintercept = mean(car$age_of_policyholder)), linetype
    = "dashed", size = .7,col= 'orange') +
  theme(
    plot.title = element_text(size = 12, face = "bold", hjust = 0.5),
    plot.caption = element_text(face = "italic") ) +
  annotate("text", x = mean(sub_car$age_of_policyholder) + 0.007 ,
   label = "sample mean", y = 6, size = 4, angle = 90 , color = 'red'
   ) +
  annotate("text", x = lower - 0.01 , label = "lower", y = 6, size =
   4, angle = 90, color = 'green') +
  annotate("text", x = upper - 0.01 , label = "upper", y = 6, size =
   4, angle = 90, color = 'blue') +
  annotate("text", x = mean(car$age_of_policyholder) - 0.01 , label =
   "actual mean", y = 6, size = 4, angle = 90, color = 'orange')
```

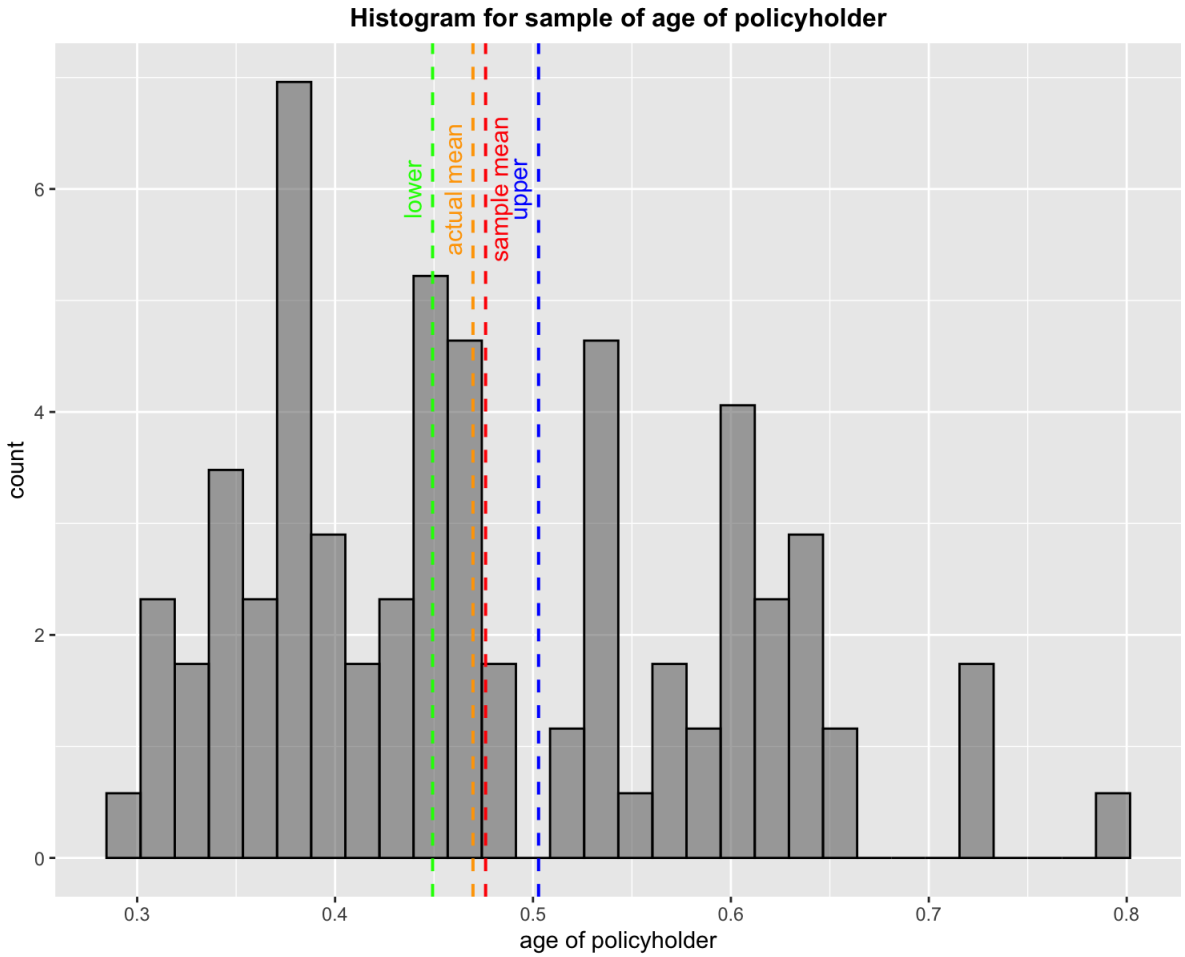

Figure 40: Histogram of age of policyholder marked with CI and sample mean

8.4 D

For the mean value of this numerical variable, design a hypothesis test and by finding the p-value, confirm or reject your assumption. What does this p-value signify?

The actual mean of the population is equal to 0.469. The mean of the sample is equal to 0.476. It is therefore necessary to design a test to measure it. Is it bigger than the actual one?

In the first part, we checked the central limit condition.

We have a random variable with a Normal Distribution. We should standardize it with Z score:

$$Z = \frac{observation - \mu}{SE} = \frac{0.469 - 0.476}{0.0115} = -0.608$$

Setting the hypothesis:

$$H_0 : \mu = 0.469 \quad on\ average,\ the\ age\ of\ policyholder\ is\ 0.469.$$
$$H_a : \mu > 0.469 \quad on\ average,\ the\ age\ of\ policyholder\ is\ bigger\ than\ 0.469.$$

and p-value like:

$$\begin{aligned} p - value &= P\left(observed\ or\ more\ extreme\ outcome \mid H_0\ true\right) \\ &= P\left(\bar{x} \geq 0.476 \mid H_0 :\ \mu = 0.469\right) \\ &= P\left(Z \geq -0.608\right) \\ &\simeq 0.27 \quad \alpha = 0.02 \end{aligned}$$

Null Hypothesis is not rejected in favor of H_a. we can believes that mean age of policyholder is not overestimate and the mean age of policyholder is less than 0.476.

```
null.value = 0.469
alpha = 0.02

x_bar <-mean(sub_car$age_of_policyholder)
s <- sd(sub_car$age_of_policyholder)
se <- s/sqrt(length(sub_car$age_of_policyholder))

z_score <- abs((x_bar - null.value)) / se

pvalue <- pnorm(z_score, lower.tail = FALSE)

if (pvalue < alpha) {
  print("Reject null hypothesis.")
} else {
  print("Fail to reject null hypothesis.")
}
```

Result: Fail to reject null hypothesis. There needs to be more evidence to say the mean of the samples is higher.

8.5 E

Based on the confidence interval you calculated in part "A", does the data support the hypothesis that you have designed? Explain it.

Statistical results can be analyzed using P-values and confidence intervals, and their effects usually agree.
We cannot reject H_0 because the sample is in CI with a confidence level of 98%.
In addition, we cannot reject H_0 at a significance level of 0.02.
Since H_0 represents the population mean, it was expected to be the same.

8.6 F

Calculate type II error. What does this value mean?

We know that the actual mean number is 0.4696782 so:

$$\mu_a = 0.4696782 \rightarrow \bar{x} \sim N\left(\mu = 0.4696782, SE = \frac{s}{\sqrt{n}} = 0.79\right)$$

$\alpha = 0.02,\ one\ sided\ test \rightarrow z_a = ?$

$P\left(Z < z_a\right) = 0.02 \rightarrow z_a = qnorm(0.02) = -2.053$

$$\begin{aligned} Type\ II\ Error = \beta &= P\left(fail\ to\ reject\ H_0 | \mu = \mu_a\right) \\ &= P\left(\frac{\bar{x} - 0.476}{0.79} < -2.053 |\ \bar{x} \sim N\left(\mu = 0.4696782,\ SE = 0.79\right)\right) \\ &= P\left(\bar{x} < -1.45 |\ \bar{x} \sim N\left(\mu = 0.4696782,\ SE = 0.79\right)\right) \\ &= P\left(Z < \frac{-1.45 - 0.4696782}{0.79}\right) \\ &= pnorm(-1.79) = 0.036 \end{aligned}$$

$\boxed{Power = 1 - \beta = 0.96}$

```r
ci = 0.98 # alpha = 0.02
z = qnorm((1-ci)/2)

s = sd(sub_car$age_of_policyholder)
se <- s / sqrt(n) #SE = s/sqrt(n)

lower <- xbar + z * se
upper <- xbar - z * se

muactual = mean(car$age_of_policyholder)

Zleft <- (lower-muactual)/(se)
Zright <-(upper-muactual)/(se)
b <- pnorm(Zright)-pnorm(Zleft,lower.tail = FALSE)

power = 1-b
power
```

Result: Error Type II = 0.034 and Power is 96.3%.
Because the value we considered in the null hypothesis was very close to the actual mean value. As a result, the type 2 error is expected to be very small.

8.7 G

Calculate the power and Explain the relationship between the power and the effect size

It was calculated that 96% of the power was obtained from the previous section.
Power and effect size are closely related in statistical tests.
It is easy to reject the null hypothesis since there are significant differences between the two groups.
The null hypothesis becomes more likely to fail with increasing effect size.
Test power is increased as a result.

You can see relationship between effect size and power below, when we reach 0.3 effect size power increases to 100%.

```
# calc effectsize ~ power
d_seq <- seq(0, 2, by = 0.1)
pwr_list <- lapply(d_seq, function(d){
  power.t.test(n = 100, delta = d, sd = sd(sub_car$age_of_policyholder
   ), type="one.sample")
})
pwr <- sapply(pwr_list, '[[', 'power')

dfpwr <- data.frame(power = pwr, effect.size = d_seq)

ggplot(dfpwr, aes(effect.size, power)) +
  geom_point(size = 2, colour = "#8ACA88") +
  geom_line(size = 0.5, colour = "black") +
  scale_y_continuous(labels = scales::percent) +
  xlab("effect size") +
  ylab(expression("test power =" ~ 1 - beta))
```

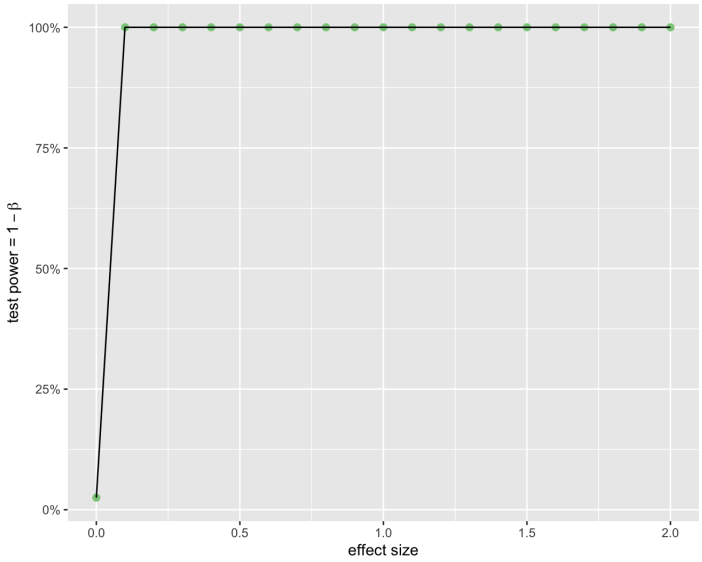

Figure 41: Effect size vs. Power

9. Question 8

Choose a numerical variable that has outliers, and we cannot apply CLT-based methods we have learned so far.

Chosen Numerical Variables : age_of_car

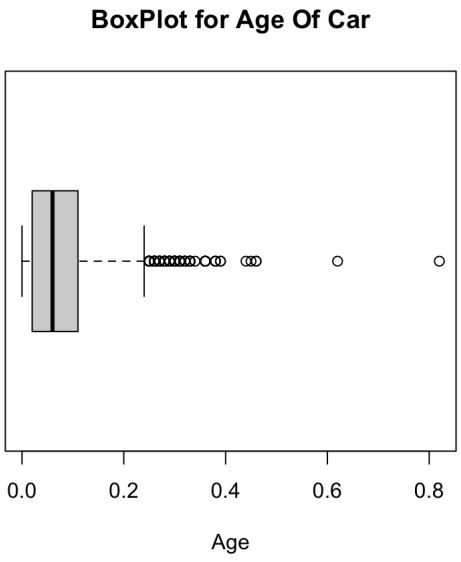

Figure 42: Boxplot of age of car

This variable has some outliers so we can't use CLT for this.

9.1 A

Calculate a 95% confidence interval for the mean of this variable using the percentile method and show the interval on the histogram

```
> quantile(car$age_of_car,c(.025,.975))
 2.5% 97.5%
 0.00  0.19
```

Figure 43: Percentile Method to find CI

Result: Confidence Interval for Original population!! : (0, 0.19)

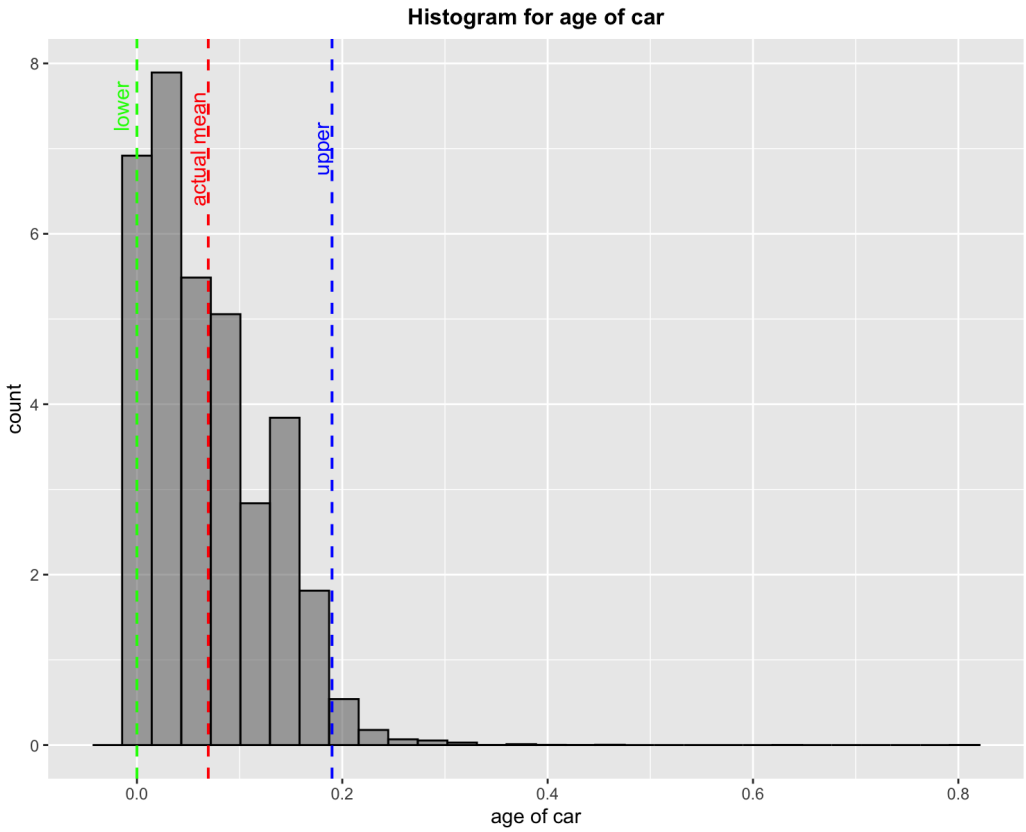

Figure 44: Histogram of age of car with CI 95% and actual mean

Using the percentile method, we can see that the confidence interval is between 0 and 0.19, so this interval covers the minimum and nearly maximum (because it has a right-skewed distribution), so it is ineffective and useless since it encompasses the entire range of possible values.

9.2 B

Pick a random sample of size 20. Then, using the bootstrapping method, calculate a 95% confidence interval for the mean of this variable using the standard error method.Also, plot the bootstrap distribution on the dot-plot.

```
car.sample = sample_n(car, 20)
boots.dist <- replicate(1000, mean(sample(car.sample$age_of_car, size
    = 20, replace=TRUE) ))

boots.mean <- mean(boots.dist)
boots.sd = sd(boots.dist)
boots.se = boots.sd / sqrt(1000)

upper = boots.mean + qt((1-0.95)/2, df=1000-1, lower.tail = FALSE) *
    boots.se
lower = boots.mean - qt((1-0.95)/2, df=1000-1, lower.tail = FALSE) *
    boots.se

cat('Confidence Interval for bootstrap dusribution : (', lower,upper,
    ')')
```

Result: Confidence Interval for bootstrap dusribution : (0.06881241, 0.07061859).

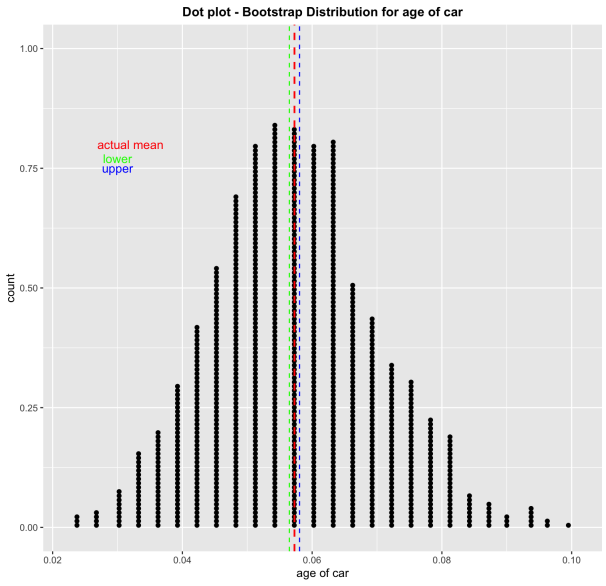

Figure 45: Bootstrap distribution for age of car - find CI with standard error method

Based on 20 samples of the age of a car, calculate the margin and bootstrap distribution using the standard error method.
This interval is much better and more informative than the previous part, as seen by the confidence interval between 0.06881241 and 0.07061859. As a result, we have a better and more accurate idea of what to expect.

9.3 C

Is there any noticeable difference between these two calculated confidence intervals? Explain your reasoning

Yes, there is a noticeable difference between these two intervals. Can see qqplot from two distributions.

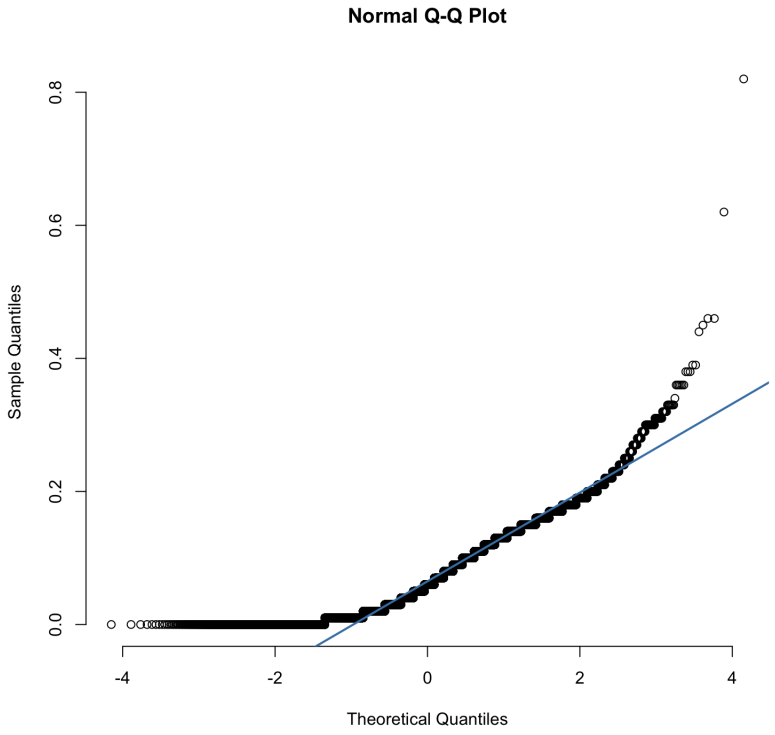

Figure 46: QQPlot - Original distribution - age of car

As you can see, this distribution has outliers. It is skewed, so when we use the percentile method, the result is not accurate and informative because we, only in normal distributions, can use this method.
Therefore use bootstrapping method to reach a normal distribution. Furthermore, use the standard error to approximate the margin of error, which is better than the percentage method.

Here is the Q-Q plot of bootstrap, which is a nearly normal distribution and has no skewness:

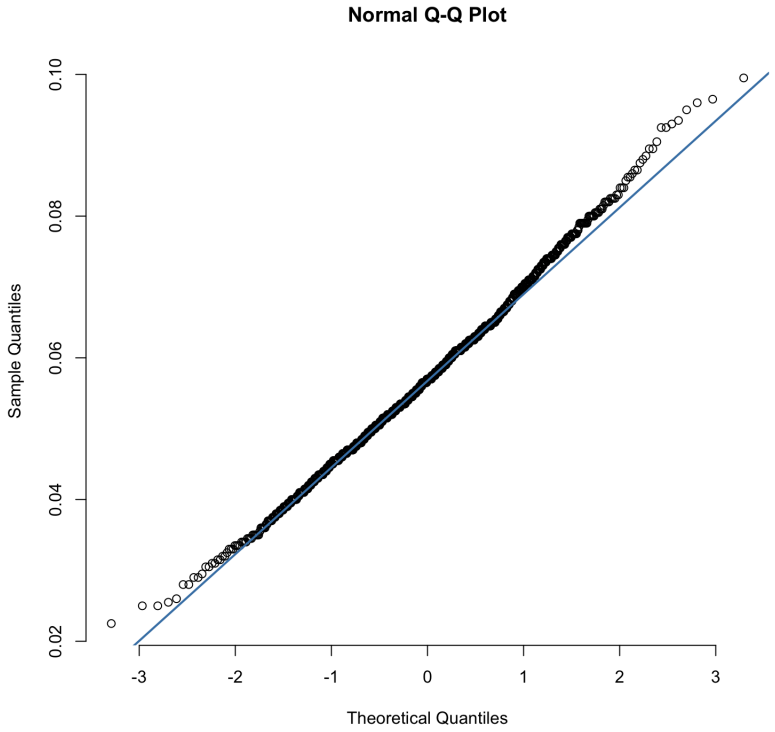

Figure 47: QQPlot - bootstrap distribution - age of car

10. Question 9

Choose a numerical and a categorical variable with more than two levels.
Divide observations of this dataset into different groups such that each group represents a level of the chosen categorical variable.

Chosen Numerical Variables : age_of_car Chosen Categorical Variables : fuel_type
A random sample without replacement was taken from the original sample, n = 200.
You can see the Box plot of three variables by fuel type below:

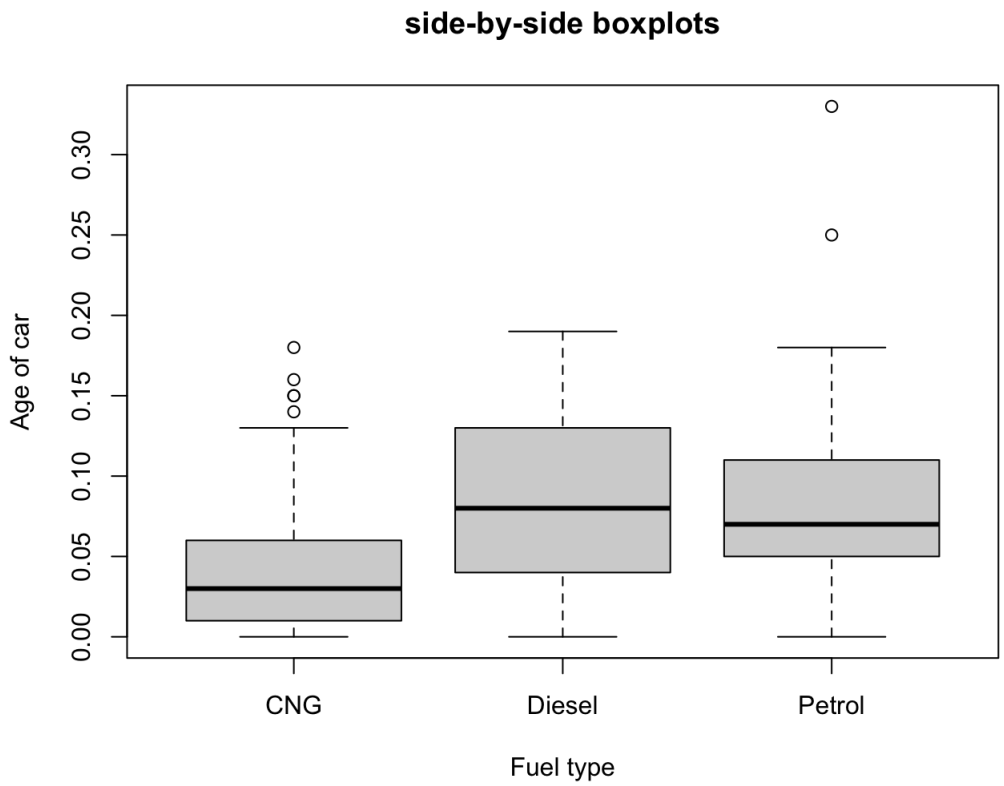

Figure 48: Three Types of fuel

Compared to petrol and diesel, CNG has a significant difference. In the next section, we will examine these cases in more detail.

10.1 A

Use the ANOVA test and compare the mean value of the numerical variable in the groups

In ANOVA, the null hypothesis is that there is no difference among group means. If any group differs significantly from the overall group mean, then the ANOVA will report a statistically significant result.

Significant differences among group means are calculated using the F statistic, which is the ratio of the mean sum of squares (the variance explained by the independent variable) to the mean square error (the variance left over).

If the F statistic is higher than the critical value (the value of F that corresponds with your alpha value, usually 0.05), then the difference among groups is deemed statistically significant.

Conditions for ANOVA:

1. Independence:

1.1 within groups: sampled observations is independent.
1.1.1 random sample / assignment.
1.1.2 each n_j less than 10% of population.

1.2 between groups: the groups is independent of each other (non-paired).

2. Approximate normality: distributions should be nearly normal within each group.

3. Equal variance: groups should have roughly equal variability.

Based on these conditions, we can assume that they are met.

Based on the probability distribution, establish the null hypothesis and alternative hypothesis.

$$H_0 : \ On\ average\ the\ mean\ of\ all\ group\ means\ are\ equal.$$
$$H_a : \ At\ least\ one\ group\ mean\ is\ different\ from\ the\ rest.$$

```
> car.sample = sample_n(car, 200)
> fisher <- aov(age_of_car  ~  as.factor(fuel_type), data = car.sample)
> summary(fisher)
                     Df Sum Sq Mean Sq F value   Pr(>F)
as.factor(fuel_type)   2 0.1586 0.07931   36.48 3.32e-14 ***
Residuals            197 0.4282 0.00217
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 49: ANOVA Test

Because p-value is smaller than significance level, we reject the null hypothesis.

Conclusion:
This means that there is a difference between at least two of the groups.

10.2 B

Choose two of the groups, perform a hypothesis test for the mean difference of the selected numerical variable in these groups and calculate the p-value.
Make a decision and explain the result using a significance level of 5%.

Based on the probability distribution, establish the null hypothesis and alternative hypothesis.
We have independence and approximate normality conditions.

$$H_0 : \mu_A - \mu_B = 0 \quad On\ average\ the\ mean\ of\ thesetwo\ groups\ are\ equal.$$
$$H_a : \mu_A - \mu_B \neq 0 \quad On\ average\ the\ mean\ of\ these\ two\ groups\ are\ not\ equal.$$

First check Petrol and Diesel

```
gp <- unique(car.sample$fuel_type)

Petrol <- car.sample$age_of_car[which(car.sample$fuel_type==gp[1])]
Diesel <- car.sample$age_of_car[which(car.sample$fuel_type==gp[2])]

car.t.test <- t.test(Petrol , Diesel)
car.t.test
```

```
        Welch Two Sample t-test

data:  Petrol and Diesel
t = -0.2742, df = 101.22, p-value = 0.7845
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.02250769  0.01704102
sample estimates:
 mean of x  mean of y
0.08226667 0.08500000
```

Figure 50: T-test for Petrol and Diesel

$$\alpha^* = \frac{\alpha}{K} = \frac{0.05}{3} = 0.016, \alpha = 0.05, K = \binom{3}{2} = 3$$

$\alpha^* = 0.78 > 0.016 \rightarrow RejectH_0$ mean of type Petrol is different from the mean of type Diesel.

Based on t-test there is no differences between the means of them.

Second check Petrol and CNG

```
Petrol <- car.sample$age_of_car[which(car.sample$fuel_type==gp[1])]
CNG <- car.sample$age_of_car[which(car.sample$fuel_type==gp[3])]

car.t.test <- t.test(Petrol , CNG)
car.t.test
```

```
        Welch Two Sample t-test

data:  Petrol and CNG
t = 8.3535, df = 118.76, p-value = 1.425e-13
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.04374431 0.07092601
sample estimates:
 mean of x  mean of y
0.08226667 0.02493151
```

Figure 51: T-test for Petrol and CNG

$\alpha^* = 0 < 0.016 \rightarrow RejectH_0$ mean of type Petrol is different from the mean of type CNG.

Based on t-test there is a differences between the means of them.

* that the t-test was used because when the df is increased, it goes toward normal, and there is no matter.

11. R codes

R Codes

```r
# Import Libraries

require(qqplotr)
library(gridExtra)
library(moments)
library(magrittr)
library(ggfortify)
library(ggplot2)
library(Hmisc)
library(plyr)
library(hexbin)
library("plot3D")
library(plotly)
library(scatterplot3d)
library(RNHANES)
library(GGally)
library(dplyr)
require(ggpubr)
library(ggmosaic)
require(corrplot)
library(patchwork)
library(ggExtra)
library(ggcorrplot)
library(reshape)
theme_set(theme_minimal())


# read dataset
path = './Car_Insurance_Claim_Prediction.csv'
car<-read.csv(path)


#-----> Question 0 <-----
#Part A ---------
#Part B ---------
summary(car)

#Part C ---------
any(is.na(car))
missingValue <- data.frame(colSums(is.na.data.frame(car)))

#Part D ---------

#-----> Question 1 <-----
#Part A ---------
# Freedman-Diaconis rule
bins_fd <- function(vec) {
  diff(range(vec)) / (2 * IQR(vec) / length(vec)^(1/3))}
bin_fd = bins_fd(car$age_of_car)

# plot
```

```r
ggplot(car, aes(x = age_of_car)) +
  geom_histogram(aes(y = ..density..), colour = "black", fill = "#F7D302",
                 bins = bin_fd) +
  labs(
    title = "Histogram and Density of Age Of Cars",
    x = "age of car",
    y = "count"
  ) +
  geom_density(color = "#631919", size = 1) +
  theme_classic() +
  theme(
    plot.title = element_text(size = 12, face = "bold", hjust = 0.5),
    plot.caption = element_text(face = "italic")
  )

# sqrt(n) rule
bins_sqrt <- function(vec) {
  sqrt(length(vec))}
bin_sq <- bins_sqrt(car$age_of_car)

# plot
ggplot(car, aes(x = age_of_car)) +
  geom_histogram(aes(y = ..density..), colour = "black", fill = "#F7D302",
                 bins = bin_sq) +
  labs(
    title = "Histogram and Density of Age Of Cars",
    x = "age of car",
    y = "count"
  ) +
  geom_density(color = "#631919", size = 1)  +
  theme_classic() +
  theme(
    plot.title = element_text(size = 12, face = "bold", hjust = 0.5),
    plot.caption = element_text(face = "italic")
    )

# default
ggplot(car, aes(x = age_of_car)) +
  geom_histogram(aes(y = ..density..), colour = "black", fill = "#F7D302") +
  labs(
    title = "Histogram and Density of Age Of Cars",
    x = "age of car",
    y = "count"
  ) +
  geom_density(color = "#631919", size = 1) +
  theme_classic() +
  theme(
    plot.title = element_text(size = 12, face = "bold", hjust = 0.5),
    plot.caption = element_text(face = "italic")
  )

# qqplot
qqnorm(car$age_of_car, pch = 1, frame = FALSE)
qqline(car$age_of_car, col = "steelblue", lwd = 2)
```

```r

#Part A ---------

#Part C ---------
quantile(car$age_of_car)

boxplot(car$age_of_car,
      main = "BoxPlot for Age Of Car",
      names = c("age_of_car"),
      xlab = "Age",
      horizontal = TRUE)


#Part D ---------
aoc <- car$age_of_car
length(aoc[which(aoc < boxplot$stats[1] | aoc > boxplot$stats[5])])

#Part E ---------
mean(car$age_of_car)
median(car$age_of_car)
var(car$age_of_car)
sd(car$age_of_car)

#Part F ---------
aoc <- car$age_of_car
mean = mean(aoc)


Fo = aoc[aoc > .75*mean]
Th  = aoc[aoc >= .5*mean & aoc<= .75*mean]
Se = aoc[aoc < .5*mean & aoc >= .25*mean]
Fi = aoc[aoc < .25*mean]

frequencies<-c(length(Fo),length(Th),length(Se),length(Fi))
percentage  <- round(100*frequencies/sum(frequencies), 2)

aoc.categorized <- data.frame(names = c("Fourth", "Third", "Second","
    First"),value = percentage)

ggplot(aoc.categorized, aes(x="", y = value, fill = names)) +
  geom_bar(stat = "identity") + coord_polar("y") +
  geom_text(aes(label = paste0(value, "%")),
            position = position_stack(vjust = 0.5)) +
  labs(title="Age Of Car Pie Chart", x = 'Frequency', y = 'Age Of Car'
    )

#Part G ---------
ggplot(car, aes(x = age_of_car)) + geom_density( size = 1) +
  geom_vline(aes(xintercept = median(age_of_car)), linetype = "dashed"
    , size = .7,col= 'red') +
  geom_vline(aes(xintercept = mean(age_of_car)), linetype = "dashed",
    size = .7,col= 'green') +
  theme_classic() +
  theme(
    plot.title = element_text(size = 12, face = "bold", hjust = 0.5),
    plot.caption = element_text(face = "italic")
  ) +
```

```
  annotate("text", x = mean(car$age_of_car) + .02 , label = "mean", y = 6, size = 3, angle = 90 , color = 'green') +
  annotate("text", x = median(car$age_of_car) - 0.02 , label = "median", y = 0.6, size = 3, angle = 90, color = 'red') +
  labs(
    title = "Age of Car Density",
    caption = "green line is about mean and red is about median variable on the diagram.",
    x = "Age Of The Car",
    y = "Count"
  )

#-----> Question 2 <-----
#Part A ---------
transmission <- car$transmission_type
gp <- unique(transmission)

Manual = transmission[transmission =='Manual']
Automatic = transmission[transmission =='Automatic']

frequencies<-c(length(Manual),length(Automatic))
percentage  <- round(100*frequencies/sum(frequencies), 2)

transmission.categorized <- data.frame(types = c("Manual", "Automatic"),percentage = percentage, value = frequencies)

#Part B ---------
transmission.categorized <- transmission.categorized[order(percentage),]
ggplot(data=transmission.categorized, aes(x=types, y=frequencies, fill=types)) +
  geom_bar(stat="identity", alpha = 0.7, width = 0.7) +
  labs(title="Barplot of Transmission Type") +
  coord_flip()

#Part C ---------
ggplot(data=transmission.categorized, aes(x=types, y=percentage, fill=types)) +
  geom_bar(stat="identity", alpha = 0.7, width = 0.6)+
  labs(title="Barplot of Transmission Type")+
  geom_text(aes(label=paste(percentage, "%")),vjust=-0.4,size=4)

#Part D ---------
ggplot(data = car, aes(x=transmission_type, y=age_of_car, fill=transmission_type))+
  geom_violin( trim=FALSE, alpha = 0.7)+
  labs(title="Violin Plot of age of car by transmission type")


#-----> Question 3 <-----
#Part A ---------
ggplot(car, aes(y=age_of_policyholder, x=age_of_car)) +
  geom_point(size=2, alpha = .07) +
  geom_smooth(method=lm, col = '#8ACA88') +
  labs(title=paste("Scatter Plot of Age of car and Age of policyholder"))

```

```
#Part B ---------
ggplot(car, aes(y=age_of_policyholder, x=age_of_car, shape = fuel_type
    , color = fuel_type)) +
  geom_point(size=1.5, alpha = 1) +
  labs(title=paste("Scatter Plot of Age of car and Age of policyholder
    with fuel type"))

#Part C ---------
cor.test(car$age_of_car, car$age_of_policyholder)

#Part D ---------
# bin 7
p7 <- ggplot(car, aes(age_of_car, age_of_policyholder)) +
  geom_point() +
  geom_hex(bins = 7) +
  geom_smooth(color='#FFFF00', se=FALSE) +
  ggtitle("Hexbin Plot - Age of car and policyholder with binsize 7")

ggMarginal(p7, type="histogram", size=3, fill='lightblue')
#-------

# bin 12
p12 <- ggplot(car, aes(age_of_car, age_of_policyholder)) +
  geom_point() +
  geom_hex(bins = 12) +
  geom_smooth(color='#FFFF00', se=FALSE) +
  ggtitle("Hexbin Plot - Age of car and policyholder with binsize 12")

ggMarginal(p12, type="histogram", size=3, fill='lightblue')
#-------

# bin 30
p30 <- ggplot(car, aes(age_of_car, age_of_policyholder)) +
  geom_point() +
  geom_hex(bins = 30) +
  geom_smooth(color='#FFFF00', se=FALSE) +
  ggtitle("Hexbin Plot - Age of car and policyholder with binsize 30")

ggMarginal(p30, type="histogram", size=3, fill='lightblue')
#----

# bin 100
p100 <- ggplot(car, aes(age_of_car, age_of_policyholder)) +
  geom_point() +
  geom_hex(bins = 100) +
  geom_smooth(color='#FFFF00', se=FALSE) +
  ggtitle("Hexbin Plot - Age of car and policyholder with binsize 100"
    )

ggMarginal(p100, type="histogram", size=3, fill='lightblue')

#Part E ---------
ggplot(car, aes(x=age_of_car, y=age_of_policyholder) ) +
  stat_density_2d(aes(fill = ..level..), geom = "polygon") +
  scale_fill_continuous(type = "viridis") +
  labs(title = "2D density plot Age of car and Age of policyholder")

```

```r
#-----> Question 4 <-----
#Part A ---------

# Some preprocessing
car$is_claim <- as.logical(car$is_claim)
car <- subset(car, select = -c(X))

# Separate Numerical features
car.numeric <- Filter(is.numeric, car)

# correlation matrix
cor_mat <- cor(as.matrix(car.numeric), method="pearson")
cor_mat <- round(cor_mat, 2)

# calc p_values
p.mat <- cor_pmat(car.numeric)

ggcorrplot(
  cor_mat, hc.order = TRUE, type = "upper",
  lab = TRUE , p.mat = p.mat, sig.level = 0.05,
  title = "heatmap correlogram",
  colors = c("blue", "white", "red"),
  hc.method = "complete", lab_size = 2,
  as.is = FALSE)

#Part B ---------
selected_var = data.frame("policy tenure" = car$policy_tenure,
                          "car's age" = car$age_of_car,
                          "policyholder's age_"=car$age_of_
   policyholder,
                          "population density"= car$population_density
   ,
                          "displacement"=car$displacement,
                          "turning radius"=car$turning_radius,
                          "length"=car$length,
                          "width"=car$width,
                          "height"=car$height,
                          "gross_weight" = car$gross_weight
                          )


#ggpairs(dplyr::select_if(car, is.numeric), title = "Correlogram",
   options(expressions=10000))

ggpairs(selected_var, title = "Correlogram",
        lower = list(continuous = wrap("smooth", alpha = 0.1, colour="
   gray")),
        upper = list(continuous = wrap("density",colour="#8ACA88",
   alpha = 0.5), combo = "dot_no_facet"))


#Part C ---------
colors <- c("#999999", "#E69F00", "#56B4E9")

fuel_type <- colors[as.numeric(factor(car$fuel_type))]

scatterplot3d(
```

```
  main="3D Scatter Plot",
  car$population_density,
  car$age_of_policyholder,
  car$displacement,
  color=fuel_type,
  xlab = "population_density",
  ylab = "age_of_policyholder ",
  zlab = "displacement",
  pch = 16)

legend("right", legend = unique(car$fuel_type), pch = 16, col =
    colors)

#-----> Question 5 <-----
#Part A ---------
Total <- sum
t<-as.data.frame.matrix(addmargins(table(car[,c('fuel_type', '
    transmission_type')]), FUN = Total))

#Part B ---------
ggplot(car, aes(x =fuel_type , fill = transmission_type)) +
  geom_bar(position = "dodge", alpha = 0.6)  +
  geom_text(aes(label = ..count..), stat = "count", vjust = -.3,  size
    = 3, color = 'black', position = position_dodge(width = 1)) +
  labs(title="Grouped Bar Chart", x="fuel_type")

#Part C ---------
ggplot(car, aes(x = fuel_type, fill = transmission_type)) +
  geom_bar(position = "stack", alpha = 0.6) +
  geom_text(aes(label=..count..), stat='count', vjust = -.3,  size =
    3, color = 'black', position = position_stack(0.5)) +
  labs(title=paste("Segmented barplot"))

#Part D ---------
ggplot(data = car) +
  geom_mosaic(aes(x = product(transmission_type), fill = fuel_type ),
    alpha = 0.6) +
  annotate("text", x = 0.195, y = 0.35,label=paste(round(t[2,1]/t
    [4,1],2)*100,"%"),fontface = "bold" ,size=4,color="black") +
  annotate("text", x = 0.195, y = 0.855,label=paste(round(t[3,1]/t
    [4,1],2)*100,"%"), fontface = "bold",size = 4,color="black") +
  annotate("text", x = 0.7, y = 0.25,label=paste(round(t[1,2]/t
    [4,2],2)*100,"%"),fontface = "bold" ,size = 4,color="black") +
  annotate("text", x = 0.7, y = 0.57,label=paste(round(t[2,2]/t
    [4,2],2)*100,"%"), fontface = "bold",size = 4,color="black") +
  annotate("text", x = 0.7, y = 0.8,label=paste(round(t[3,2]/t[4,2],2)
    *100,"%"), fontface = "bold",size = 4,color="black") +
  labs(title=paste("Mosaic plot"))

#-----> Question 6 <-----
#Part A ---------
car_sample <- sample_n(car, 25)
car_sample$diff <- car_sample$policy_tenure - car_sample$displacement

#Part B ---------
t.test(car_sample$policy_tenure, car_sample$displacement, paired =
    TRUE, var.equal = TRUE)
```

```r

# calculate CI
xbar1 <- mean(car_sample$policy_tenure)
s1 <- sd(car_sample$policy_tenure)
xbar2 <-  mean(car_sample$displacement)
s2 <- sd(car_sample$displacement)

sp = ((n-1)*s1^2+(n-1)*s2^2)/(n+n-2)
sp

margin <- qt(0.975,df=n+n-1)*sqrt(sp/n + sp/n)
margin

lowerinterval <- (xbar1-xbar2) - margin
lowerinterval

upperinterval <- (xbar1-xbar2) + margin
upperinterval

cat("95\% confidence interval is= (", lowerinterval,",",upperinterval,
    ")")

#-----> Question 7 <-----

ggplot(car, aes(x = age_of_policyholder)) +
  geom_histogram(aes(y = ..density..),alpha=0.5, colour = "black",
    fill = "gray")+
  labs(
    title = "Histogram for age of policyholder",
    x = "age of policyholder",
    y = "count"
  ) +
  geom_density(color = "blue", size = 1)  +
  theme(
    plot.title = element_text(size = 12, face = "bold", hjust = 0.5)
  )

#Part A ---------
n = 100

# Reads the dataset 'car' and take the 100 rows as sample
sdf<- sample(1:nrow(car), n)

#sample 10 rows
sub_car <- car[sdf,]

xbar = mean(sub_car$age_of_policyholder)

ci = 0.98
z = qnorm((1-ci)/2)

se <- sd(sub_car$age_of_policyholder) / sqrt(n) #SE = s/sqrt(n)

lower <- xbar + z * se
upper <- xbar - z * se

cat("98\% confidence interval is= (", lower,",",upper,")")
```

```r

#Part B ---------
ggplot(sub_car, aes(x = age_of_policyholder)) +
  geom_histogram(aes(y = ..density..),alpha=0.5, colour = "black")+
  labs(
    title = "Histogram for sample of age of policyholder",
    x = "age of policyholder",
    y = "count"
    ) +
  geom_vline(aes(xintercept = mean(age_of_policyholder)), linetype = "dashed", size = .7,col= 'red') +
  geom_vline(aes(xintercept = lower), linetype = "dashed", size = .7, col= 'green') +
  geom_vline(aes(xintercept = upper), linetype = "dashed", size = .7, col= 'blue') +
  geom_vline(aes(xintercept = mean(car$age_of_policyholder)), linetype = "dashed", size = .7,col= 'orange') +
  theme(
    plot.title = element_text(size = 12, face = "bold", hjust = 0.5),
    plot.caption = element_text(face = "italic")
  ) +
  annotate("text", x = mean(sub_car$age_of_policyholder) + 0.007 , label = "sample mean", y = 6, size = 4, angle = 90 , color = 'red' ) +
  annotate("text", x = lower - 0.01 , label = "lower", y = 6, size = 4, angle = 90, color = 'green') +
  annotate("text", x = upper - 0.01 , label = "upper", y = 6, size = 4, angle = 90, color = 'blue') +
  annotate("text", x = mean(car$age_of_policyholder) - 0.01 , label = "actual mean", y = 6, size = 4, angle = 90, color = 'orange')

#Part D ---------
null.value = 0.469
alpha = 0.02

x_bar <-mean(sub_car$age_of_policyholder)
s <- sd(sub_car$age_of_policyholder)
se <- s/sqrt(length(sub_car$age_of_policyholder))

z_score <- abs((x_bar - null.value)) / se

pvalue <- pnorm(z_score, lower.tail = FALSE)

if (pvalue < alpha) {
  print("Reject null hypothesis.")
} else {
  print("Fail to reject null hypothesis.")
}

#Part F ---------
ci = 0.98 # alpha = 0.02
z = qnorm((1-ci)/2)

s = sd(sub_car$age_of_policyholder)
se <- s / sqrt(n) #SE = s/sqrt(n)

lower <- xbar + z * se
```

```r
upper <- xbar - z * se

muactual = mean(car$age_of_policyholder)

Zleft <- (lower-muactual)/(se)
Zright <-(upper-muactual)/(se)
b <- pnorm(Zright)-pnorm(Zleft,lower.tail = FALSE)

power = 1-b
power
#-------

# calc effectsize ~ power
d_seq <- seq(0, 2, by = 0.1)
pwr_list <- lapply(d_seq, function(d){
  power.t.test(n = 100, delta = d, sd = sd(sub_car$age_of_policyholder
   ), type="one.sample")
})
pwr <- sapply(pwr_list, '[[', 'power')

dfpwr <- data.frame(power = pwr, effect.size = d_seq)

ggplot(dfpwr, aes(effect.size, power)) +
  geom_point(size = 2, colour = "#8ACA88") +
  geom_line(size = 0.5, colour = "black") +
  scale_y_continuous(labels = scales::percent) +
  xlab("effect size") +
  ylab(expression("test power =" ~ 1 - beta))

#-----> Question 8 <-----
boxplot(car$age_of_car,
        main = "BoxPlot for Age Of Car",
        names = c("age_of_car"),
        xlab = "Age",
        horizontal = TRUE)

#Part A ---------
q <- quantile(car$age_of_car,c(.025,.975))
cat('Confidence Interval for Original population!! : (', q[1],q[2], ')
   ')

ggplot(car, aes(x = age_of_car)) +
  geom_histogram(aes(y = ..density..),alpha=0.5, colour = "black")+
  labs(
    title = "Histogram for age of car",
    x = "age of car",
    y = "count"
  ) +
  geom_vline(aes(xintercept = mean(age_of_car)), linetype = "dashed",
   size = .7,col= 'red') +
  geom_vline(aes(xintercept = q[1]), linetype = "dashed", size = .7,
   col= 'green') +
  geom_vline(aes(xintercept = q[2]), linetype = "dashed", size = .7,
   col= 'blue') +
  theme(
    plot.title = element_text(size = 12, face = "bold", hjust = 0.5),
    plot.caption = element_text(face = "italic")
```

```
  ) +
  annotate("text", x = mean(car$age_of_car) - 0.01 , label = "actual
    mean", y = 7, size = 4, angle = 90 , color = 'red') +
  annotate("text", x =  q[1] - 0.015 , label = "lower", y = 7.5, size
    = 4, angle = 90, color = 'green') +
  annotate("text", x = q[2] - 0.01 , label = "upper", y = 7, size = 4,
     angle = 90, color = 'blue')

#Part B ---------
car.sample = sample_n(car, 20)

boots.dist <- replicate(1000, mean(sample(car.sample$age_of_car, size
    = 20, replace=TRUE) ))

boots.mean <- mean(boots.dist)
boots.sd = sd(boots.dist)
boots.se = boots.sd / sqrt(1000)

upper = boots.mean + qt((1-0.95)/2, df=1000-1, lower.tail = FALSE) *
    boots.se
lower = boots.mean - qt((1-0.95)/2, df=1000-1, lower.tail = FALSE) *
    boots.se

cat('Confidence Interval for bootstrap dusribution : (', lower,upper,
    ')')
df.boots.dist = data.frame(age_of_car = matrix(data = boots.dist, ncol
    = 1, byrow = TRUE))

ggplot(df.boots.dist, aes(x = age_of_car)) +
  geom_dotplot( stackratio = 1.01,dotsize = .25)+
  labs(
    title = "Dot plot - Bootstrap Distribution for age of car",
    x = "age of car",
    y = "count"
  ) +
  geom_vline(aes(xintercept = mean(age_of_car)), linetype = "dashed",
    size = .8,col= 'red') +
  geom_vline(aes(xintercept = lower), linetype = "dashed", size = .5,
    col= 'green') +
  geom_vline(aes(xintercept = upper), linetype = "dashed", size = .5,
    col= 'blue') +
  theme(
    plot.title = element_text(size = 12, face = "bold", hjust = 0.5),
    plot.caption = element_text(face = "italic")
  ) +
  annotate("text", x = .032 , label = "actual mean", y = .8, size = 4,
     angle = 0 , color = 'red') +
  annotate("text", x = .03 , label = "lower", y = .77, size = 4, angle
    = 0, color = 'green') +
  annotate("text", x = .03 , label = "upper", y = .75, size = 4, angle
    = 0, color = 'blue')

#Part C ---------
qqnorm(car$age_of_car, pch = 1, frame = FALSE)
qqline(car$age_of_car, col = "steelblue", lwd = 2)
```

```r
qqnorm(df.boots.dist$age_of_car, pch = 1, frame = FALSE)
qqline(df.boots.dist$age_of_car, col = "steelblue", lwd = 2)


#-----> Question 9 <-----
car.sample = sample_n(car, 200)

#draw box plot
boxplot(car.sample$age_of_car  ~ car.sample$fuel_type, ylab = "Age of
    car", xlab = "Fuel type",
         main = "side-by-side boxplots")

#Part A ---------
fisher <- aov(age_of_car  ~  as.factor(fuel_type), data = car.sample)
summary(fisher)

#Part B ---------
gp <- unique(car.sample$fuel_type)
gp
# petrol vs, diesel
Petrol <- car.sample$age_of_car[which(car.sample$fuel_type==gp[1])]
Diesel <- car.sample$age_of_car[which(car.sample$fuel_type==gp[2])]

car.t.test <- t.test(Petrol , Diesel)
car.t.test

# petrol vs, CNG
Petrol <- car.sample$age_of_car[which(car.sample$fuel_type==gp[1])]
CNG <- car.sample$age_of_car[which(car.sample$fuel_type==gp[3])]

car.t.test <- t.test(Petrol , CNG)
car.t.test
```