UTM

UNIVERSITI TEKNOLOGI MALAYSIA

School of Computing

Session 2020/21

Probability and Statistical Data Analysis

GROUP PROJECT-2 INFERENTIAL STATISTIC ANALYSIS

LECTURER: Dr. Izyan Izzati Kamsani

Section:06

Group Members	
MD Masqurul Hasan	A20EC4031
Fuad Jama Abdillahi	A20EC4021
Mohamed Ahmed Labib Eltelb	A18CS4028

Contents

1.0 Introduction

Cars play a vital role in our everyday lives since we travel to our destinations every day. However, having a great car in terms of characteristics such as fuel type, horsepower, and pricing is also essential, especially for travelers. Therefore, the primary goal of this study is to provide key data about a car, such as its manufacture, engine size, horsepower peak rpm, and so on. as well as to utilize statistical analysis skills in the dataset to demonstrate if there is a relationship between the data. In order to achieve this goal, a few potential variables are chosen, and a series of test analysis are performed.

2.0 Background of Study

The car specifications dataset is a secondary data source collected from the Kaggle website. this data was collected by Eleanor Xu who works as an analytic from Coursera, San Francisco, California, United States. This dataset was collected to show the data of 205 sample cars, in terms of their car companies, car aspiration, Types of fuel, doors number, location of engine, width, length height , type of engine, horsepower, price and other specifications.

3.0 Objective of Study

The study was conducted to meet the following objectives:
- Applying and carrying out statistical test analysis on secondary data sources
- To demonstrate if the dataset's selected variables are dependent on one another

4.0 Description of Data

Population : Cars from different companies
Sample : 205 Cars

Variables (Description)	Type of Variable	Measurement Level
Car number from sample 1 to 205)	Quantitative	Nominal
Make (name of car company)	Qualitative	Nominal
Fuel type(gas or diesel fuel used)	Qualitative	Nominal
Aspiration (standard or turbo-aspirated car)	Qualitative	Nominal
Door numbers (number of car doors)	Quantitative	Ratio
Body style (sedan, wagon, hatchback, etc.)	Qualitative	Nominal
Wheels drive (front-wheel drive, 4-wheel drive)	Qualitative	Nominal
Engine location (front or rear)	Qualitative	Nominal
Wheel base (distance from front to rear wheels)	Quantitative	Ratio
length (length of car)	Quantitative	Ratio
Width (width of car)	Quantitative	Ratio
Height (height of car)	Quantitative	Ratio
Curb weight (total mass of car)	Quantitative	Ratio
Engine type (DOHC, OHC, OHCV, etc.)	Qualitative	Nominal
Num cylinders (number of cylinders)	Quantitative	Ratio
Engine size (size of car engine)	Quantitative	Interval
Fuel system (MPFI, MFI, 2BBL, etc.)	Qualitative	Nominal
Compression ratio (cylinder volume when piston at top to cylinder volume when piston at bottom)	Quantitative	Ratio
Horsepower (car horsepower)	Quantitative	Ratio
Peak rpm (car engine rev/min at max power)	Quantitative	Ratio
City mpg (car fuel consumption on city streets)	Quantitative	Ratio
Highway mpg (car fuel consumption on highways)	Quantitative	Ratio
price (car sale price)	Quantitative	Ratio

5.0 Test Statistical Analysis

Selected Variables	Objectives	Test Analysis , Expected Outcome
Engine size, price	At a 95% confidence level, use Pearson's Product-Moment Correlation Coefficient to determine whether a linear relationship exists between engine size and car price.	**Analysis:** Correlation Analysis **Expected Outcome:** There is strong linear relationship between the engine size and the car price, at confidence level 95%. The larger the engine size, the higher the car price.
Aspiration and horsepower	To determine if the mean horsepower of turbo-aspirated cars is greater than the mean horsepower of standard-aspirated cars at a 95% level of confidence, considering different variances.	**Analysis:** 2 Sample Hypothesis Testing (Test on Mean, Variance Unknown) **Expected Outcome:** The mean of horsepower of turbo-aspirated cars is larger than the mean of horsepower of standard-aspirated cars, at confidence level 95% and assuming variances unequal.
Fuel type	To test whether there is difference between the observed frequency and expected frequency of fuel type used by cars, that is gas or diesel fuel, at 95% confidence level.	**Analysis:** Goodness of Fit Test (One Way Contingency Table) **Expected Outcome:** There is difference between the observed frequency and expected frequency of fuel type used by cars, at 95% confidence level. The observed frequency is not a good fit to the assumed distribution.

Engine size, horsepower	To test whether the value of horsepower depend on the value of car engine size, using engine size as the independent variable(x) and horsepower as the dependent variable(y).	**Analysis:** Regression Analysis **Expected Outcome:** The value of horsepower depends on the value of engine size. The larger the car engine size, the larger the car horsepower.
Numbers of doors, aspiration	To test whether the number of doors and car aspiration are related using Two Way Contingency Table, at 95% confidence level.	**Analysis:** Chi-Square Test of Independence **Expected Outcome:** The number of doors and the car aspiration are irrelated and independent at 95% confidence level.

6.0 Analysis andDiscussion

✓ 2 Sample Hypothesis Testing

In this analysis, we will use the variables aspiration and horsepower to see if the mean horsepower of turbo-aspirated cars is greater than the mean horsepower of standard-aspirated cars at a 95 % confidence level, assuming differences with variances. From the data, frequency(n), mean($\bar{x}$), standard deviation(s) are calculated.

```
> mean(horsepower[aspiration=="turbo"])
[1] 124.4324
> mean(horsepower[aspiration=="std"])
[1] 100
> sd(horsepower[aspiration=="turbo"])
[1] 31.24059
> sd(horsepower[aspiration=="std"])
[1] 39.89927
```

	aspiration	count	mean	sd
	<chr>	<int>	<dbl>	<dbl>
1	std	168	100	39.9
2	turbo	37	124.	31.2

Calculating mean and standard deviation

We have calculated the required data, we can group them now:

$\overline{X1}$=124.4324	$\overline{X2}$=100.00
S_1= 31.24059	S_2= 39.89927
N_1=37	N_2=168

where group$_1$ is for turbo-aspirated cars, while group$_2$ is for std-aspirated cars.

1. **Hypothesis statement:**

   H_0: $\mu_1 = \mu_2$

   H_1: $\mu_1 > \mu_2$

   μ_1 equals the mean of horsepower of turbo-aspirated cars, and μ_2 equals the mean of horsepower of std-aspirated cars.

2. **To calculate the test statistics, t_0 with given 95% confidence level, α = 0.05 formula is given below**

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

   By using RStudio, test statistics, t_0 = 4.0804.

3. **Formula for calculating degree of freedom is :**

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}$$

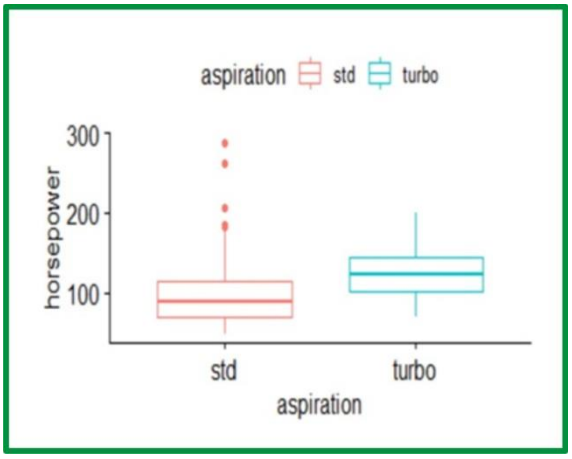
```
> alpha=0.05
> t.alpha=qt(alpha,floor(v))

t.alpha    -1.66901302502409
```

Critical Value of T using R-Studio

Degree of freedom by using RStudio, v = 64.711.
Therefore, using α = 0.05, we reject if H_0 if $t_0 > t_{0.025,64.711}$ = 1.669.

∴Critical value, $t_{0.025,64.711}$ = 1.669, *p-value* = 0.0001258.

4. **Decision:**
   Since test statistics t_0 = 4.0804 > critical value $t_{0.025,64.711}$ = 1.669, we **reject** the null hypothesis. So there is sufficient evidence to support that mean of horsepower of turbo aspirated cars is greater than the mean of horsepower of standard aspirated cars, at significance level α = 0.05.



Visualizing data using box plots

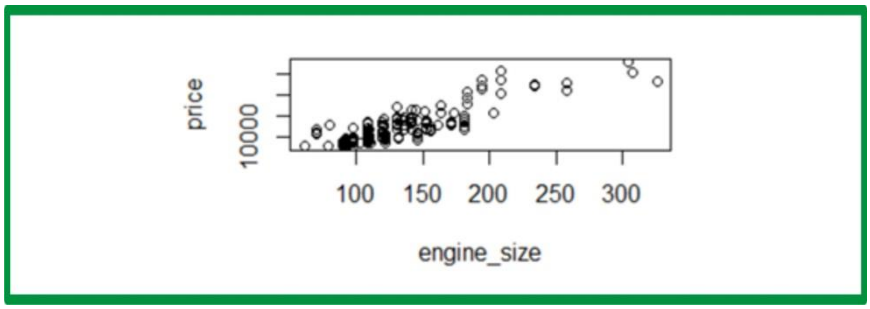
```
> t.test(horsepower[aspiration=="turbo"],
 horsepower[aspiration=="std"])

        Welch Two Sample t-test

data:  horsepower[aspiration == "turbo"]
 and horsepower[aspiration == "std"]
t = 4.0804, df = 64.711,
p-value = 0.0001258
alternative hypothesis: true difference i
n means is not equal to 0
95 percent confidence interval:
 12.47298 36.39188
sample estimates:
mean of x mean of y
 124.4324  100.0000
```

P-*value* is shown, which equals to 0.0001258.

✓ Correlation Analysis

We are using Pearson's Product-Moment Correlation Coefficient to evaluate whether there is a linear relationship between **engine size** and **car pricing** in this analysis at confidence level 95%.

The strength of association between two variables is measured using correlation analysis. We used Pearson's Product-Moment Correlation Coefficient for the correlation coefficient since the variables engine size and price are ratio-type data.



Visualize data in scatter plot using RStudio

In the above figure Left side scatter plot indicate that there exist positive correlation relation between car price and engine size and that is the higher the car price the larger the engine size while there are also some outliers on the top right side of scatter plot.

1. **Using Pearson's method calculate the sample correlation coefficient by R studio :**

```
> cor(x,y)
[1] 0.8731717
```
Calculate *r* using RStudio

we get sample correlation coefficient, $r = 0.8731717$. Which express that there is relatively strong positive linear correlation between x and y.

2. **Significance Test for Correlation**

- Hypothesis Statement:
  H_0: $\rho = 0$ no linear correlation
  H_1: $\rho \neq 0$ linear correlation exists

- Calculate test statistic using Rstudio:

```
> n <- 205
> r <- cor(x,y)
> t <- r/(sqrt((1-(r^2))/(n-2)))
```

t	25.5241267815105

Calculate test statistic using RStudio

We have test statistic t = 25.5241.
From the table find critical value, using α = 0.05, df = n-2 = 203,this is a two-tailed test so, there are two critical values:

Critical value $-t_{\alpha/2=0.025,\ df=203}$ = -2.25815 Lower tail and Critical value $t_{\alpha/2=0.025,\ df=203}$ = 2.25815 Upper tail

We also got $p\text{-}value = 2.2e-16$ in Rstudio

If test statistics > 2.25815 / test statistics < -2.25815, reject H_0. Otherwise fail to reject H_0.

Decision: Since test statistics t = 25.5241 > upper tail critical value $t_{\alpha/2=0.025,\ df=203}$ = 2.25815, we **reject** the null hypothesis. There is sufficient evidence to conclude that there is a linear relationship between car engine size and car price, at α =0.05.

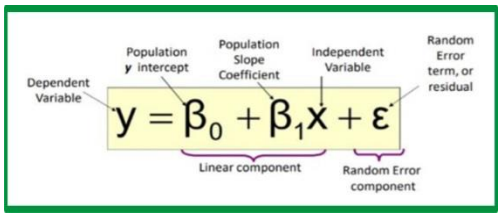
```
> cor.test(car_data$engine_size, car_data$price, method="pearson")

        Pearson's product-moment
        correlation

data:  car_data$engine_size and car_data$price
t = 25.524, df = 203, p-value <
2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8361913 0.9022482
sample estimates:
      cor
0.8731717
```

Significance test for correlation using RStudio

✓ Regression Analysis

In this analysis, we are using variables **engine size** and **horsepower**, where we will test whether the value of horsepower depend on the value of engine size, using engine size as the independent variable(x) and horsepower as the dependent variable(y).
Since our regression model is linear, we use simple linear regression. The change in horsepower values are assumed to be caused by changes in engine size values.

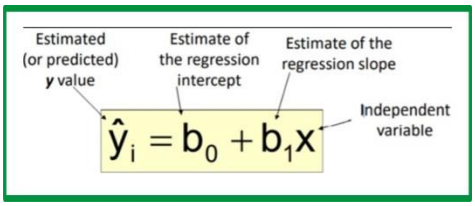
Equation for population linear Regression is:



We assume that:

- Error values(Ɛ) are normally distributed and statistically independent for any *x*
- Probability distribution of errors has constant variance
- Underlying relationship between variable *x* and variable *y* is linear

1. Estimated Regression Model:



From above equation b_0 is the estimated average value of *y* when the value of *x is zero*. Whereas b_1 is the estimated change in the average value of *y* due to a one-unit change in *x*.

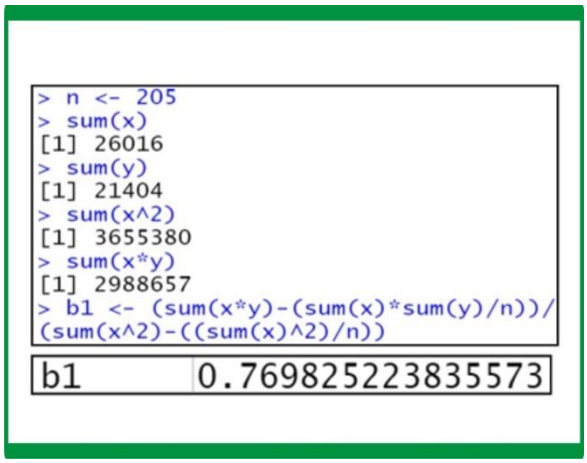
- Find least squares criterion:

  we can find the values of b_0 and b_1 by:

$$b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} \qquad b_0 = \bar{y} - b_1 \bar{x}$$

```
> mean(x)
[1] 126.9073
> mean(y)
[1] 104.4098
> b0 <- mean(y)-(b1*mean(x))
```

b0	6.71330232533525



By using RStudio, we get b_1 = 0.7698, b_0 = 6.7133.

Substituting the values of b_0 and b_1 into the regression model equation:

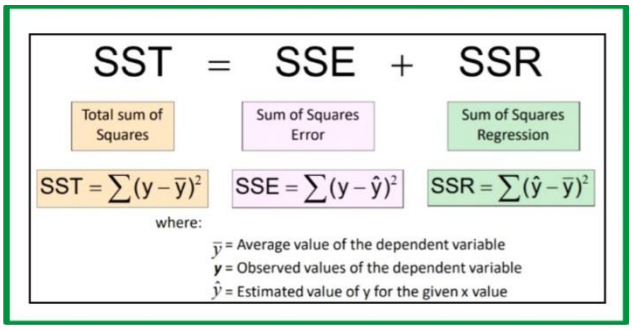
$$\hat{y}i = 6.7133 + 0.7698x$$

From above equation, we can do interpretation of the intersection coefficient b_0, and slope coefficient b_1.

In the data, no cars had 0 engine size, so b_0 = 6.7133 indicates that, for cars within the range of engine sizes observed, 6.7133 is the portion of horsepower not explained by engine size.

Whereas b_1 = 0.7698 indicate that the average value of car horsepower increases by 0.7698 on average, for each additional one-unit engine size.

Explained and Unexplained Variation:



By using RStudio, we get:

SSR = 209648.6475, SST = 319091.5805, SSE = 109442.9330

Coefficient of Determination, R^2 using R studio:

```
> R2 <- SSR/SST
R2        0.657017170843358
```

Coefficient of Determination, R^2 = 0.6570
Hence, we can interpret it as 65.7% of the variation in horsepower is explained by variation in engine size.

Find Standard Error of Estimate by:

$$s_\varepsilon = \sqrt{\frac{SSE}{n-k-1}}$$

```
> k <- 1
> Se <- sqrt(SSE/(n-k-1))
Se        23.2191246377784
```

U sing RStudio, Error of Estimate, s_ε = 23.2191.

Find Standard Deviation of Regression Slope by:

$$s_{b_1} = \frac{s_\varepsilon}{\sqrt{\sum(x-\bar{x})^2}} = \frac{s_\varepsilon}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}}$$

```
> Sb1 <- Se/(sqrt(sum((x-mean(x))^2)))
Sb1       0.0390383943909781
```

Using RStudio,Standard Deviation of Regression Slope, s_{b1}= 0.03904.

2. Slope Inference: t-Test

- Hypothesis Statement:
  H_0: $\beta_1 = 0$ no linear relationship
  H_1: $\beta_1 \neq 0$ linear relationship exist

- Using $\alpha = 0.05$ finding critical value df = n-2 = 203
  since this is a two-tailed test, there are two critical values:
  critical value $-t_{\alpha/2=0.025,\ df=203} = -2.25815$ Lower tail
  critical value $t_{\alpha/2=0.025,\ df=203} = 2.25815$ Upper tail

  we also get *p-value* = 2.2*e* - 16.

  Hence, we reject H_0 if test statistics > 2.25815 / test statistics < -2.25815.

- Calculate test statistic by:

$$t = \frac{b_1 - \beta_1}{s_{b_1}}$$

```
> t <- (b1-0)/Sb1
t       19.7196948246796
```

  we get test statistic t = 19.7197.

- **Decision**:
  Since test statistics t = 19.7197 > upper tail critical value $t_{\alpha/2=0.025,\ df=203} = 2.25815$, we **reject** the null hypothesis. There is sufficient evidence that engine size affects car horsepower, at $\alpha = 0.05$.

∴∴ Linear Regression Model: $\hat{y}i = 6.7133 + 0.7698x$

We use the **lm()** function to perform linear regression in RStudio:

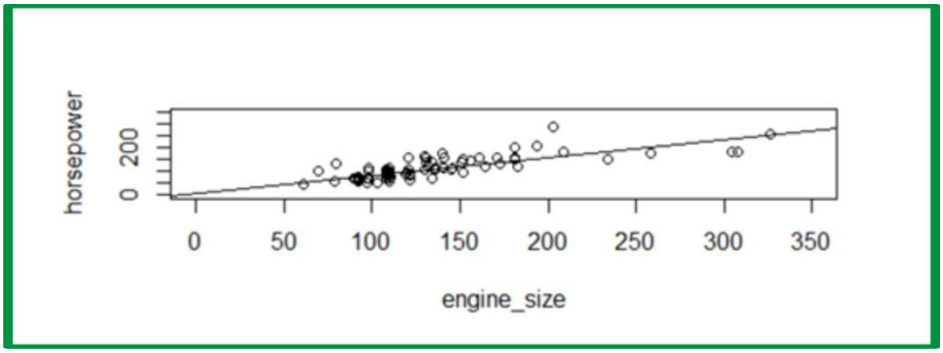
```
> model <- lm(y~x)
> model

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     6.7133       0.7698
```

we can see values of intersection coefficient from the image above, (Intercept) and slope coefficient (x).

```
> summary(model)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-59.819 -12.386  -5.624  10.138 125.012

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.71330    5.21292   1.288    0.199
x            0.76983    0.03904  19.720   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.22 on 203 degrees of freedom
Multiple R-squared:  0.657,     Adjusted R-squared:  0.6553
F-statistic: 388.9 on 1 and 203 DF,  p-value: < 2.2e-16
```

we can get the values of intersection coefficient b_0 = 6.7133, slope coefficient b_1 = 0.7698, Standard Deviation of Regression Slope, $sb1$= 0.03904, Standard Error of Estimate, sЄ = 23.2191, df = 203, Coefficient of Determination, R2 = 0.6570 and p-value = 2.2e – 16. When we view the summary of our linear regression model,

Finally using the **plot()** function we can plot a scatter plot, and add the linear regression model into the plot using **abline()** function:



✓ Chi-Square Test of Independence

For this Analysis we are using variables **aspiration** and **num_of_doors**
where we will test whether number of doors and car aspiration are related using Two Way Contingency Table.At 95% confidence level.
Therefore, we use Chi-Square Test of Independence, with two-way contingency table.

```
> table(car_data$num_of_doors, car_data$aspiration)

      std turbo
four   93    23
two    75    14
```

1. **State the test hypothesis:**
   H_0: No relationship between variables.
   H_1: Variables are related and dependent.

2. **Find the critical value:**
3.

```
> alpha <- 0.05
> x2.alpha <- qchisq(alpha, df=1, lower.tail=FALSE)
```

x2.alpha	3.84145882069413

Finding critical value x^2 using RStudio

Critical value x^2 = 3.841 , (with df=(2-1)(2-1)=1, α = 0.05)

4. **Calculate the expected counts:**

num_of_doors	aspiration				Total
	std		turbo		
	Obs.	Exp.	Obs.	Exp.	
four	93	$\frac{116 \times 168}{205} = 95.1$	23	$\frac{116 \times 37}{205} = 20.9$	116
two	75	$\frac{89 \times 168}{205} = 72.9$	14	$\frac{89 \times 37}{205} = 16.1$	89
Total	168	168	37	37	205

***Remarks**: $e_{ij} \geq 5$ in all cells

5. **Calculate the test statistic value:**

- Calculate manually:

Cell, ij	Observed Count, o_{ij}	Expected Count, e_{ij}	$\frac{(o_{ij} - e_{ij})^2}{e_{ij}}$
1, 1	93	$\frac{116 \times 168}{205} = 95.1$	$\frac{(93 - 95.1)^2}{95.1} = 0.0464$
1, 2	23	$\frac{116 \times 37}{205} = 20.9$	$\frac{(23 - 20.9)^2}{20.9} = 0.2110$
2, 1	75	$\frac{89 \times 168}{205} = 72.9$	$\frac{(75 - 72.9)^2}{72.9} = 0.0605$
2, 2	14	$\frac{89 \times 37}{205} = 16.1$	$\frac{(14 - 16.1)^2}{16.1} = 0.2739$
		x^2	0.5918

When we calculate test statistic manually, we get test statistic $x^2 = 0.5918$.

- Using RStudio:

```
> chisq.test(tbl, correct=FALSE)

        Pearson's Chi-squared test

data:  tbl
X-squared = 0.57158, df = 1, p-value = 0.4496
```

When we calculate test statistic using RStudio, we get test statistic $x^2 = 0.57158$, with *p-value* = 0.4496.

6. **Decision**:
Since test statistic value ($x^2 = 0.57158$) < critical value($x^2_{k=1, \alpha = 0.05} = 3.841$), it does not fall within the critical region. Thus, we **fail to reject** H_0. There is sufficient evidence to conclude that there is no relationship between the variables num_of_doors and aspiration, at $\alpha = 0.05$.

7.0 Conclusion

We observed that the mean of horsepower of turbo-aspirated cars is greater than the mean of horsepower of standard-aspirated cars using two-sample hypothesis testing, where we test on the mean assuming unequal variances. , hence we **reject** null hypothesis. In the real world, this valid argument might be true because turbo-aspirated cars require more horsepower to improve car performance in terms of engine performance and acceleration, especially for sport cars.

Next, for the correlation analysis, we found out that there is a linear relationship between car engine size and car price, hence we also **reject** null hypothesis. The relationship indicates a relatively strong positive linear correlation, where sample correlation coefficient $r = 0.8731717$.

In the real world, although car engine size is not the main element influencing car pricing, larger engine size affects, with larger engine leading to higher car price, because larger engine has more equipment standard and tends to be more expensive.

For the regression analysis, we found out that engine size affects car horsepower, with our Linear Regression Model equation: $\hat{y} = 6.7133 + 0.7698x$, hence we **reject** null hypothesis. We may conclude that the engine size and car horsepower have a positive linear connection. In the real world, a larger engine is typically more powerful; this is an important feature to consider for racing drivers because a larger engine size boasts more power to make the car more nimble and performance-enhanced.

Lastly, for chi-square test of independence, we found out that there is no relationship between the number of car doors and car aspiration, hence we **fail to reject** null hypothesis. In real world, number of car doors also do not affect whether the car is std-aspirated or turbo-aspirated. The difference in the number of car doors is just probably for aesthetic purpose