UTM UNIVERSITI TEKNOLOGI MALAYSIA

SCHOOL OF COMPUTING
Faculty of Engineering

Project 2

SECI2143 PROBABILITY & STATISTICAL DATA ANALYSIS

SEMESTER II, SESSION 2021/2022

Lecturer: Dr Rozilawati Dollah @ Md Zain

Group Name: 4 for a kind

Name	Matric No.
LIM SHI KAI	A21EC0196
PHANG SENG SOON	A21EC0220
TAN CHUN MING	A21EC0229
NG KENG KEAT	A21EC0211

Section: 09

Table of Contents

1.0 Introduction and Background

Cars, an automobiles on wheels that are primarily used for transportation. During the 20th century, cars were invented and became widely used since they were essential since developed countries' economies depend on them. In 2022, around 1.446 billion cars will be in the world. So, it is clearly shown that cars play a crucial role in our daily lives. Nowadays, consumers' choice expands as automakers release many car models. Every car has its specifications in terms of horsepower, car body configuration, fuel consumption, price and many more. So, the consumers would consider those specifications, whether they suit their driving style or other aspects, before buying the car. As a result, the primary goal of this study is to display crucial facts about a vehicle, such as its manufacturer, fuel type, aspiration, number of doors and so on. Then, we will apply statistical analysis skills to the dataset to determine whether the data is linked. A few candidate variables are chosen to achieve this goal, and a series of test analyses are performed.

We retrieved this dataset from Kaggle, and the data was gathered by Eleanor Xu, an analyst from Coursera in San Francisco, California, United States. This dataset is secondary data since it is obtained online. This dataset contains 205 samples from several manufacturers, and each car has its specifications, including fuel type, aspiration, number of doors, etc. Next, we will choose a few specifications considered variables in this project for testing purposes, such as 2 sample hypothesis testing, correlation analysis, regression analysis, the goodness of fit test, and chi-square test of independence will be used to test the chosen variables in our project.

2.0 Description of Data

Dataset URL: https://www.kaggle.com/datasets/jingbinxu/sample-of-car-data

Population: Cars from various manufacturers

Sample: 205 cars

Data Description:

Variables	Description	Type of Variable	Measurement Level
#	Number of cars	Quantitative	Ratio
make	Car manufacturer name	Qualitative	Nominal
fuel_type	Gas or diesel	Qualitative	Nominal
aspiration	Standard or turbocharged	Qualitative	Nominal
num_of_doors	Number of car doors	Quantitative	Ratio
body_style	Car body configuration	Qualitative	Nominal
drive_wheels	Drivetrain	Qualitative	Nominal
engine_location	Location of engine	Qualitative	Nominal
wheel_base	Distance between the centres of the front and rear wheels	Quantitative	Ratio
length	Length of car	Quantitative	Ratio
width	Width of car	Quantitative	Ratio
height	Height of car	Quantitative	Ratio
curb_weight	Weight of car	Quantitative	Ratio
engine_type	Types of camshaft drives	Qualitative	Nominal
num_of_cylinders	Number of engine cylinders	Quantitative	Ratio
engine_size	Engine capacity	Quantitative	Interval
fuel_system	Type of fuel injection	Qualitative	Nominal
compression_ratio	Ratio between the volume of the cylinder and the combustion chamber	Quantitative	Ratio
horsepower	Car horsepower	Quantitative	Ratio
peak_rpm	RPM range that produces max horsepower	Quantitative	Ratio
city_mpg	Fuel consumption in the city	Quantitative	Ratio
highway_mpg	Fuel consumption on the highway	Quantitative	Ratio
price	Car price	Quantitative	Ratio

Statistical Test Analysis

Selected Variables	Objectives	Test Analysis and Expected Outcome
aspiration, horsepower	To determine if the mean horsepower of the turbocharged car is greater than the mean horsepower of natural aspirated (standard) cars at a 95% confidence level, assuming unequal variances.	**Analysis:** 2 sample hypothesis testing **Expected Outcome:** The mean horsepower of turbocharged cars is greater than standard cars at a 95% confidence level.
engine_size, price	To measure the strength of the linear relationship between engine size and car price at a 95% confidence level.	**Analysis:** Correlation Analysis **Expected Outcome:** There is a linear relationship between engine size and car price: the larger the engine size, the higher the car price at a 95% confidence level.
engine_size, horsepower	To identify the relationship between a dependent variable (engine size) and an independent variable (horsepower).	**Analysis:** Regression Analysis **Expected Outcome:** There is a relationship between engine size and horsepower; the larger the engine size, the higher the horsepower.
fuel_type	To test the difference between the observed frequency and expected frequency of fuel type used by cars at 95% confidence level.	**Analysis:** The goodness of Fit Test **Expected Outcome:** At a 95% confidence level, there is a difference between observed frequency and expected frequency of fuel type.
num_of_doors, aspiration	To confirm that the relationship between the two qualitative variables, which are the number of doors and aspiration at 95% confidence level.	**Analysis:** Chi-Square Test of Independence **Expected Outcome:** At a 95% confidence level, the number of car doors and aspirations are unrelated.

3.0 Data Analysis

3.1 Hypothesis Testing by using 2-sample

In the 2-sample hypothesis testing, we will use the data of horsepower and the data of aspiration, where we will calculate the mean, standard deviation and frequency for both variables. In the meantime, we will test whether the mean of the horsepower of turbo-aspirated is the same as the mean of the horsepower of standard-aspirated by using a 95% confidence level and assuming the variances are unequal.

Sample 1: Horsepower with turbo-aspirated

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

Sample 2: Horsepower with standard-aspirated

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

Tabulation data

Values	
n1	37
n2	168
s1	31.24059
s2	39.89927
xbar1	124.4324
xbar2	100

1. **Hypothesis Testing**

   H_0: $\mu_1 = \mu_2$

   H_1: $\mu_1 > \mu_2$

   where μ_1 is the mean of the horsepower of turbo-aspirated and μ_2 is the mean of the horsepower of standard-aspirated.

2. **Given, $\alpha = 0.05$. The test statistic is**

$$t_0 = \frac{\overline{x_1} - \overline{x_2} - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$t_0 = \frac{124.4324 - 100 - 0}{\sqrt{\frac{31.24059^2}{37} + \frac{39.89927^2}{168}}} = 4.08037$$

from the value calculated by using the RStudio.

3. **We also need to calculate the degrees of freedom**

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

$$v = \frac{\left(\frac{31.24059^2}{37}+\frac{39.89927^2}{168}\right)^2}{\frac{\left(\frac{31.24059^2}{37}\right)^2}{37-1}+\frac{\left(\frac{39.89927^2}{168}\right)^2}{168-1}} = 64.7109 \approx 64$$

from the value calculated by using RStudio.

Given α = 0.05, critical value t was calculated by using RStudio

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

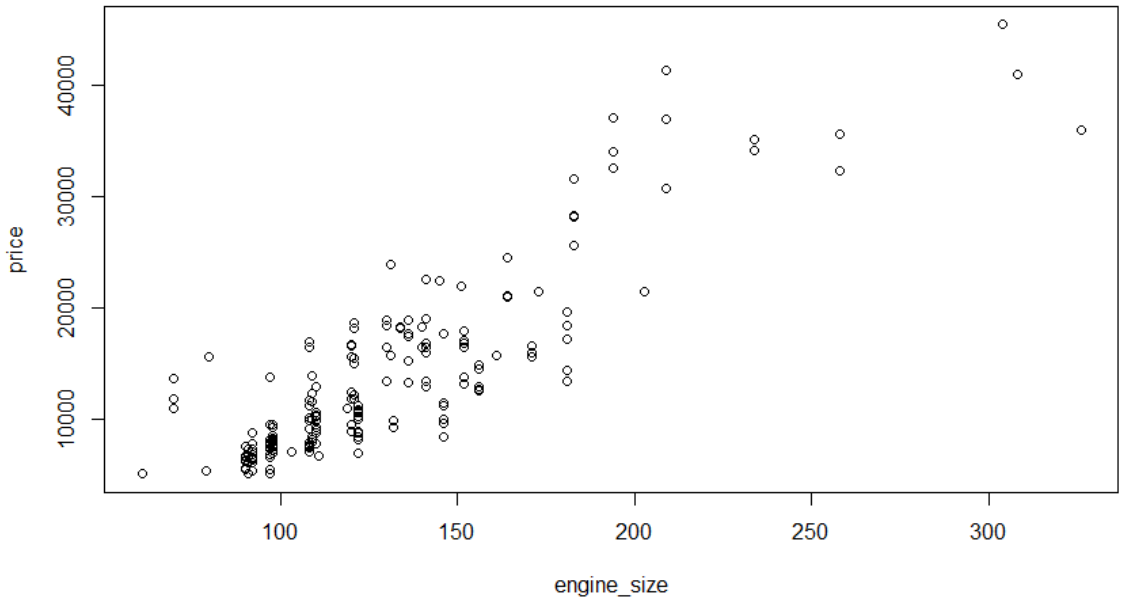
t.alpha	-1.66901302502409

4. **Conclusion**

Since $t_0 > t_\alpha$, we reject the H_0. Therefore, critical value, $t_0 = 1.669$. There is no sufficient evidence to conclude that the mean of horsepower of turbo-aspirated is equal to the mean of horsepower of standard-aspirated.

3.2 Correlation

Correlation analysis is used to test the relationship between quantitative variables or categorical variables. In other words, it's a measurement of the degree of a relationship. In the correlation analysis, we will use the variable engine_size and the variable price. At $\alpha = 0.05$, we will test whether there is a linear relationship between engine size and the car price. In fact, since both variables engine_size and price are ratio data types, we will use Pearson's Product-Moment Correlation Coefficient, which measures the strength and direction of association between two variables measured on at least an interval scale.



We used a scatter plot with the variable engine_size and price. The plot indicates a positive relationship between engine size and car price; the larger the engine size, the higher the car price. However, outliers still appear when the engine size hits above 300.

1. **Sample correlation coefficient using Pearson's Product-Moment Correlation Coefficient**

$$r = \frac{\frac{\sum xy - (\sum x \sum y)}{n}}{\sqrt{\left[(\sum x^2) - \left(\frac{\sum x^2}{n}\right)\right]\left[(\sum y^2) - \left(\frac{\sum y^2}{n}\right)\right]}}$$

```
> cor(x,y)
[1] 0.8731717
> r <- cor(x,y)
```

r	0.873171748808439

From the RStudio, we will get the sample correlation coefficient, r = 0.8732. This indicates a strong positive linear correlation between x and y.

2. Significance Test for Correlation

I. Hypothesis Testing

H_0: $\rho = 0$ (no linear correlation)

H_1: $\rho \neq 0$ (linear correlation exists)

II. The test statistic is

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

$$t = \frac{0.873171748808439}{\sqrt{\frac{1-0.873171748808439^2}{205-2}}} = 25.52413$$

from the value calculated by using RStudio.

III. Given α = 0.05, critical value t was calculated by using RStudio

```
> 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
```

From the RStudio, we can get the p-value = 2.2e – 16.

From the t table, as we are using the two-tailed test with the degree of freedom = n – 2 = 205 – 2 = 203 and significance level = $\frac{\alpha}{2} = \frac{0.05}{2} = 0.025$, the critical value, $t_{\alpha/2}$ calculated is ±2.2582.

3. Conclusion

Since $t > t_{\alpha/2}$, we reject the H_0. There is insufficient evidence to conclude that there is no linear relationship between car engine size and car price at a 95% confidence level.

3.3 Regression

Regression analysis is a set of statistical methods used to estimate the relationship between a dependent variable and one or more independent variables. In this case, we will use the variable engine_size to be the independent variable, while the dependent variable will be horsepower. Since the regression model is linear, we will use simple linear regression. In the simple linear regression, we will investigate whether engine size impacts how much horsepower an engine produces. Assuming that variations in engine size are caused for changes in horsepower values.

Mathematical Equation of Population Linear Regression:

$y = \beta_0 + \beta_1 x + \varepsilon,$

Where y is the dependent variable, β_0 (Population y-intercept), β_1(Population slope coefficient) and x (Independent variable) is the linear component ε is the random error component

1. Estimated Regression Model

$\hat{y}_i = b_0 + b_1 x$

I. First, we need to find the least-squares criterion. From the formula of the estimated regression model, we can find the values of b_0 and b_1 by using the formula

II.

a. $b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$

```
> n <- 205
> sum(x)
[1] 26016
> sum(y)
[1] 21404
> sum(x^2)
[1] 3655380
> sum(x*y)
[1] 2988657
> b1 <- (sum(x*y)-(sum(x)*sum(y)/n))/(sum(x^2)-((sum(x)^2)/n))
```

b1	0.769825223835573
n	205
x	int [1:205] 130 130 152 109 136 136 136 136 131 131 ...
y	int [1:205] 111 111 154 102 115 110 110 110 140 160 ...

From RStudio, we get b_1 = 0.7698.

b. $b_0 = \overline{y} - b_1 \overline{x}$

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

b0	6.71330232533525

From RStudio, we get $b_0 = 6.7133$.

III. Substitute the values b_0 and b_1 into the formula estimated regression model.

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

We may interpret the equation's intersection coefficient b0 and slope coefficient b1. Since there were no vehicles with a 0 engine size in the data, we can say that the amount of horsepower that cannot be explained by engine size for vehicles within the range of observed engine size is 6.7133. On the other hand, b_1 informs us that for every extra unit of the engine size, the average value of the horsepower will rise by 0.7698.

IV. Next up, we need to do the explained and unexplained variation. The total variation is made up of two parts:

$$SST\left(\sum(y-\overline{y})^2\right) = SSE\left(\sum(y-\hat{y})^2\right) + SSR\left(\sum(\hat{y}-\overline{y})^2\right)$$

where $\overline{y}$ is the average value of the dependent variable

y is the observed values of the dependent variable

$\hat{y}$ Is the estimated of y for the given x value

```
> yhat <- b0 + (b1*x)
> SSR <- sum((yhat-mean(y))^2)

  SSR        209648.647452033
> SST <- sum((y-mean(y))^2)
> SSE <- SST-SSR

SST          319091.580487805
 SSE         109442.933035772
```

From the RStudio, we will get the data SSR = 209648.6475, SST = 319091.5805 and SSE = 109442.9330

V. By using the data SSR and SST, we need to find the Coefficient of Determination, R^2, by using the formula

$$R^2 = \frac{SSR}{SST}$$

```
> R2 <- SSR/SST

 R2          0.657017170843358
```

From the RStudio, we will get the data $R^2 = 0.6570$.

Therefore, we can say that 65.7% of the variation in horsepower is explained by variation in engine size.

VI. We also need to find the Standard Error of Estimate by using the data SSE and the formula

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

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

From the RStudio, we will get the data s_ε =23.2191.

VII. Other than the Standard Error of Estimate, we also need to find the Standard Deviation of the Regression Slope by using the data s_ε and the formula

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

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

From the RStudio, we will get the data s_{b_1} = 0.0390.

2. t-test

I. Hypothesis Testing

H_0: $\beta_1 = 0$ (no linear relationship)

H_1: $\beta_1 \neq 0$ (linear relationship exists)

II. The test statistic is

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

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

From the RStudio, we will get the data t = 19.7197.

III. From the t table, as we are using the two-tailed test with the degree of freedom = n – 2 = 205 – 2 = 203 and significance level = $\frac{\alpha}{2} = \frac{0.05}{2}$ = 0.025, the critical value, $t_{\alpha/2}$ calculated is ±2.2582.

IV. Conclusion

Since $t > t_{\alpha/2}$, we reject the H_0. There is not sufficient evidence to conclude that engine size will not affect the car's horsepower at a 95% confidence level.

3. To perform linear regression on RStudio, we will use the lm() function.

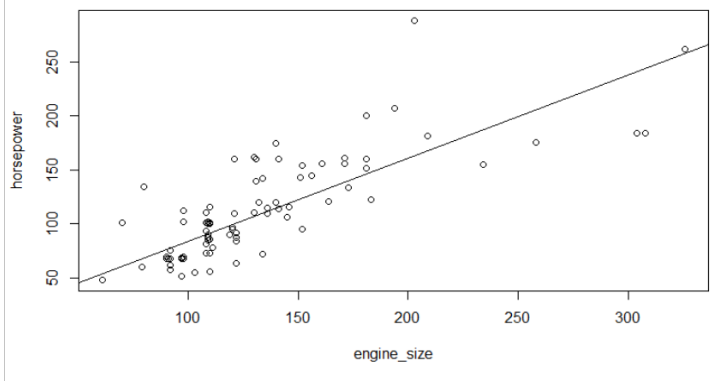
```
> model <- lm(y~x)
> model

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     6.7133       0.7698
```

From the image on the left, we can obviously read the values of the intersection coefficient,(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
```

From the summary of the linear regression model, we can get the values of $b_0 = 6.71330$, $b_1 = 0.76983$, $s_{b_1} = 0.03904$, $s_\varepsilon = 23.2191$, degree of freedom = 203, $R^2 = 0.6570$ and p-value = 2.2e - 16.

```
> plot(x, y, xlab="engine_size", ylab="horsepower")
> abline(model)
```



The figure above is the scatter plot. By using the function plot() and the function abline(), we can add the linear regression model to the scatter plot.

3.4 Goodness of the test

We are using the variable, fuel_type, to test the difference between the observed frequency and expected frequency of fuel type used by cars at a 95% confidence level. So, the goodness of fit test or the chi-square test with a one-way contingency table and the unequal probabilities will be the choice we use in the test.

Based on our dataset, we know that there are two types of fuel: gas fuel and diesel fuel. The percentages of fuel type are claimed with 85% gas fuel and 15% diesel fuel.

```
> table(car_data$fuel_type)

diesel    gas
    20    185
```

Observed frequencies for both fuel types

```
> fuel_type <- c(185, 20)
> prob <- c(0.85, 0.15)
```

Our claim:

ρ_{gas}=0.85 ,ρ_{diesel}=0.15

1. Statement of test hypothesis
   H_0: $\rho_{gas} = 0.85$, $\rho_{diesel} = 0.15$
   H_1: At least one of the two proportions is different from the claimed value.

2. Calculated Value

When E are **not equal**, E = np;

	Gas	**Diesel**	**Total**
Observed Frequency, O	185	20	205
Expected Frequency, E	np=(205)(0.85) = 174.25	np=(205)(0.15) = 30.75	205

3. Calculate the test statistics@chi-square value by:

$$x^2 = \sum \frac{(O-E)^2}{E}$$

   By using Rstudio, we get the test statistics value, x^2 = 4.4213.

4. Find the critical value:

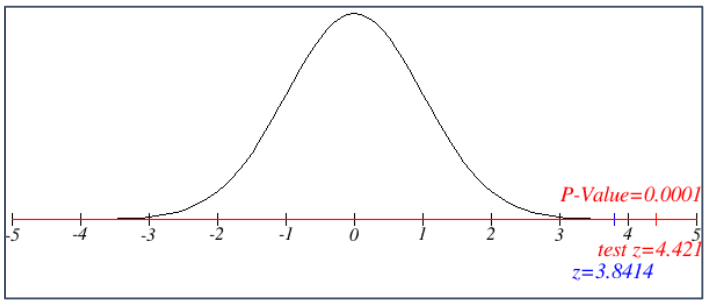
```
> alpha <- 0.05
> x2.alpha <- qchisq(alpha, df=1, lower.tail=FALSE)
> x2.alpha
[1] 3.841459
```

Critical value, x^2 get from RStudio

Critical value, $x^2 = 3.8414$ (with df = k-1 = 1 and $\alpha = 0.05$)

5. Conclusion

Since test statistics value, $x^2 = 4.4213$ is larger than critical value, $x^2{}_{1,0.05} = 3.8414$.



The test statistics value falls within the critical region, so we reject H_0.

There is sufficient evidence to conclude that the fuel types are distributed with the given percentages is different between the observed frequency and expected frequency of fuel type used by cars.

```
> chisq.test(fuel_type, p=prob, correct=FALSE)

        Chi-squared test for given probabilities

data:  fuel_type
X-squared = 4.4213, df = 1, p-value = 0.03549
```

Perfoming chi-square test using RStudio

Value of test statistics, $x^2 = 4.4213$, p-value = 0.03549

3.5 Chi-square test of independence

We are using the variable **num_of_doors** and **aspiration** to confirm that the relationship between the two qualitative variables, which are the number of doors and aspiration, exists at a 95% confidence level when using Two Way Contingency Table.

Based on our dataset, we know that there are two types of aspiration, one is standard, and the other one is turbo. In addition, the door of the car can be two or four.

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

      std turbo
four   93    23
two    75    14
```

Observed frequencies for variables num_of_doors and aspiration

1. Statement of test hypothesis
   H_0: No relationship between variables
   H_1: Variables have a relationship

2. Find the critical value:

```
> alpha <- 0.05
> x2.alpha <- qchisq(alpha, df=1, lower.tail=FALSE)
> x2.alpha
[1] 3.841459
```

Critical value, x^2 get from RStudio

Critical value, $x^2 = 3.8414$ (with df = (2-1)(2-1) = 1 and $\alpha = 0.05$)

3. Calculated Value:

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

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

4. Calculate the test statistic value:

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	$\frac{23 \times 37}{205} = 20.9$	$\frac{(23 - 20.9)^2}{20.9} = 0.2110$
2,1	75	$\frac{89 \times 168}{205} = 72.9$	$\frac{(75 - 72.9)^2}{72.9} = 0.0605$
2,2	14	$\frac{14 \times 37}{205} = 16.1$	$\frac{(14 - 16.1)^2}{16.1} = 0.2739$
		$x^2 =$	0.5918

When we calculate the test statistic value by formula, we get the value for $x^2 = 0.5918$

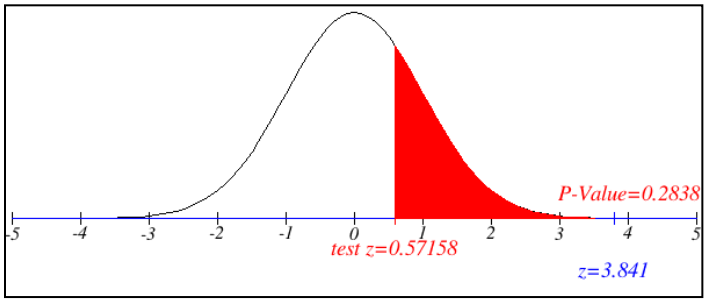
RStudio

```
> tbl = table(car_data$num_of_doors, car_data$aspiration)
> chisq.test(tbl, correct=FALSE)

        Pearson's Chi-squared test

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

We get the test statistics value, $x^2 = 0.57158$ with p-value = 0.4496 when we going through RStudio calculation.

5. Conclusion

Since test statistics value, $x^2 = 0.57158$ is smaller than the critical value, $x^2_{1,0.05} = 3.841$.



The test statistics value does not falls within the critical region, so we fail to reject H_0.

There is sufficient evidence to conclude that there is no relationship between both variables at a 95% confidence value.

4.0 Conclusion

After performing 2 Sample Hypothesis Testing, we found that test statistics is greater than the critical value. Therefore, we reject the null hypothesis since sufficient evidence shows that the mean horsepower of turbocharged cars is larger than naturally aspirated (standard) cars. In the real-world, turbocharged engines often have larger horsepower than naturally aspirated engines as the turbocharger will compress more air into the cylinder, considerably increasing engine horsepower significantly.

Next, for Correlation Analysis, we have concluded that test statistics is greater than the upper tail critical value; hence we reject the null hypothesis. There is sufficient evidence to show a linear relationship between car engine capacity and car price. For example, in real life, the 2022 Ford Mustang GT with a 5.0-litre engine pricing at $37,545 is higher than the 2022 Ford Mustang EcoBoost with a 2.3-litre engine at $32,495.

Besides that, for Regression Analysis, we discovered that test statistics is greater than the upper tail critical value, so we reject the null hypothesis. We claimed that the engine capacity and car horsepower have a positive linear relationship. In reality, too, usually bigger engine capacity will produce larger horsepower. It is because a greater engine displacement allows for more air and fuel to be burnt simultaneously, resulting in much more power than a smaller engine.

Furthermore, for the Goodness of Fit Test, we found out that the test statistics value is larger than the critical value and falls within the critical region, so we reject the null hypothesis. There is adequate evidence to disprove the claim that the fuel types are distributed with the provided percentages, which means there is a difference between observed frequency and expected frequency of fuel type. As in real life, we can see that the number of cars that use gas fuel is always greater than those that use diesel fuel.

Last but not least, using the Chi-Square Test of Independence, we discovered that the test statistics value is smaller than the critical value. It means there is no association between the number of car doors and aspiration. Therefore, we fail to reject the null hypothesis. In the actual world, the number of car doors has no bearing on whether the car is naturally aspirated or turbocharged. We see that in real life, mostly sports cars have only 2 doors as they appear nicer to the eye rather than 4 doors. Next, sports cars with 2 doors will provide more handling as the turning radius will be more precise. Family cars usually have 4 doors as it is easier for the family to get on and off the car.

In a nutshell, after completing this project, we have performed several inference statistical analyses successfully by using RStudio. Throughout this project, we have gained a lot of knowledge on finding suitable secondary data and calculating test statistics values. It is a very fruitful experience for us as we have learned a lot of things about statistics. Lastly, we would like to express our deep appreciation to our lecturer, Dr Rolizawati Dollah @ Md Zain, for giving us instructions and guidance on this project.

5.0 Reference

1. (n.d.). 2022 Ford Mustang Sports Car | Compare Models. Retrieved June 26, 2022, from https://www.ford.com/cars/mustang/compare-models/#gtfastback|ecoboostpremiumfastback
2. *8.1: The T-sTaTisTic*. (2022, May 13). Statistics LibreTexts. Retrieved June 25, 2022, from https://bit.ly/3ymSVkp
3. (n.d.). ACKO Drive | Buy car at the lowest rates. Retrieved June 26, 2022, from https://ackodrive.com/car-guide/what-is-turbo-engine-in-cars/
4. Brad, R. (2018, July 20). Why is it Most Sports Cars Have Only Two Doors?. Retrieved June 26, 2022, from https://www.kendalldodgechryslerjeepram.com/why-is-it-most-sports-cars-have-only-two-doors/
5. *Correlation in statistics: Correlation analysis explained*. (2021, June 20). Statistics How To. Retrieved June 25, 2022, from https://www.statisticshowto.com/probability-and-statistics/correlation-analysis/
6. *How many cars are there in the world in 2021? Stats by country*. (2021, October 4). Hedges & Company. Retrieved June 25, 2022, from https://hedgescompany.com/blog/2021/06/how-many-cars-are-there-in-the-world/
7. *Hypothesis test graph generator*. (n.d.). IMathAS. Retrieved June 26, 2022, from https://www.imathas.com/stattools/norm.html
8. *Pearson's product-moment correlation in SPSS statistics - Procedure, assumptions, and output using a relevant example*. (n.d.). URL Shortener - Short URLs & Custom Free Link Shortener | Bitly. Retrieved June 25, 2022, from https://bit.ly/39Vloo9

9. School of Computing, Universiti Teknologi Malaysia. (n.d.). *Chapter 6: Chi-Square Test & Contingency Analysis* [PDF Document].
10. School of Computing, Universiti Teknologi Malaysia. (n.d.). *Chapter 7 : Correlation Analysis* [PDF Document].
11. School of Computing, Universiti Teknologi Malaysia. (n.d.). Chapter 5 Part 3 : Hypothesis Testing (Two Samples Test) [PDF Document].

6.0 Video Link

Please access our presentation video via this link : https://youtu.be/ndLghfK_3-Q

7.0 e-Portfolio Link

Lim Shi Kai
https://eportfolio.utm.my/view/view.php?t=BzHK5t1nAv6TJWEobwyp

Phang Seng Soon
https://eportfolio.utm.my/view/view.php?t=9J6EZjCHlkhQ8Yxb1g2B

Ng Keng Keat
https://eportfolio.utm.my/view/view.php?t=Ib3gwSyiW8P4ZverqKkd

Tan Chun Ming
https://eportfolio.utm.my/user/tan-chun-ming/probability-statistical-data-analysis