UNIVERSITI TEKNOLOGI MALAYSIA

SCHOOL OF COMPUTING

SESSION 2019/2020 SEMESTER 2

COURSE CODE

SECI 2143 – Probability and Statistical Data Analysis

LECTURER'S NAME

Dr. Chan Weng Howe

INDIVIDUAL PROJECT (PROJECT 2)

Report on Study of Secondary Data of Car Specifications

STUDENT'S NAME

Lee Tong Ming

MATRIC NO

A19EC0069

SECTION

02

Contents

1.0 Introduction

Cars are very important in our daily life because we drive cars to our destinations every day. However, having a great car in terms of its specifications such as fuel type, horsepower and price is also very important, especially for travelers. Therefore, the main purpose of this study is to show the key data of a car such as its company, engine size, horsepower peak rpm, etc., as well as to apply the uses of statistical analysis skill in the dataset, to prove whether there is relationship between the data. In the process of achieving this objective, a few potential variables are selected, and a series of test analysis are carried out.

2.0 Background of Study

The dataset regarding the car specifications is a secondary data source retrieved 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, fuel types, car aspiration, number of doors, engine location, length, width, height, engine type, horsepower, price, and other specifications.

3.0 Objective of Study

The study was conducted to meet the following objectives:

- To apply and perform statistical test analysis on the secondary data source
- To prove whether the selected variables from the dataset are dependent on each other.

4.0 Description of Data

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

Population : Cars from various companies

Sample : 205 Cars

Data Description:

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
num_of_doors (number of car doors)	Quantitative	Ratio
body_style (sedan, wagon, hatchback, etc.)	Qualitative	Nominal
drive_wheels (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_of_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 Statistical Test Analysis

Selected Variables	Objectives	Test Analysis and Expected Outcome
aspiration, horsepower	To test whether the mean of horsepower of turbo-aspirated cars is larger than the mean of horsepower of standard-aspirated cars at 95% confidence level, assuming unequal 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.
engine_size, price	To test whether linear relationship exists between the engine size and the car price using Pearson's Product-Moment Correlation Coefficient, at 95% confidence level.	**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.
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.
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.

num_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 and Discussion

- 2 Sample Hypothesis Testing

In this analysis, we are using variables **aspiration** and **horsepower**, where we will test whether the mean of horsepower of turbo-aspirated cars is larger than the mean of horsepower of standard-aspirated cars at 95% confidence level, assuming unequal variances. From the data, frequency(n), mean($\bar{x}$), standard deviation(s) are calculated.

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

```
> 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
```

Calculating the mean and standard deviation

Now that we have calculated the data required, we can group them:

$\bar{x}_1$ = 124.4324	$\bar{x}_2$ = 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:

   $\mathbf{H_0}$: $\mu_1 = \mu_2$

   $\mathbf{H_1}$: $\mu_1 > \mu_2$

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

2. Given 95% confidence level, α = 0.05. The test statistics, t_o can be calculated by:

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

By using RStudio, test statistics, t_o = 4.0804.

3. Calculate the degree of freedom by:

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

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

t.alpha	-1.66901302502409

Finding critical value t using RStudio

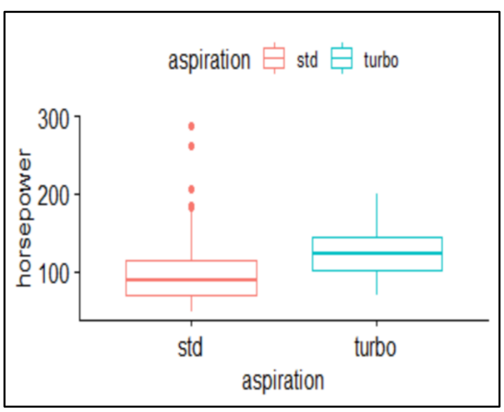
By using RStudio, degree of freedom, v = 64.711.

Therefore, using α = 0.05, we reject if H_0 if $t_o > t_{o.025,\ 64.711}$ = 1.669.

∴ Critical value, $t_{o.025,\ 64.711}$ = 1.669, *p-value* = 0.0001258.

4. Conclusion:

Since test statistics t_o = 4.0804 > critical value $t_{o.025,\ 64.711}$ = 1.669, we **reject** the null hypothesis. There is sufficient evidence to conclude that the mean of horsepower of turbo-aspirated cars is larger than the mean of horsepower of standard-aspirated cars, at α = 0.05.



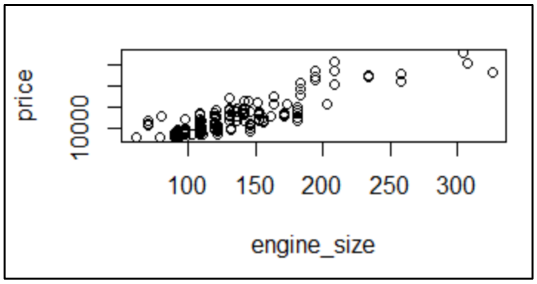
Visualizing data using box plots

```
> 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
```

Performing t-test in RStudio, *p-value* is shown, which equals to 0.0001258.

- Correlation Analysis

In this analysis, we are using variables **engine_size** and **price**, where we will test whether there is linear relationship between the engine size and the car price using Pearson's Product-Moment Correlation Coefficient, at 95% confidence level. Correlation analysis is used to measure the strength of association(linear relationship) between two variables. For the correlation coefficient, we use Pearson's Product-Moment Correlation Coefficient since variables engine_size and price are ratio-type data.



Visualize data in scatter plot using RStudio

From the scatter plot on the left, it indicates that there is positive correlation relationship between engine size and the car price, that is the larger the engine size, the higher the car price. However, there are a few outliers also on the top right side of the plot.

1. Calculate the sample correlation coefficient using Pearson's method by:

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

where:
- r = Sample correlation coefficient
- n = Sample size
- x = Value of the independent variable
- y = Value of the dependent variable

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

Calculate r using RStudio

By using RStudio, we get sample correlation coefficient, r = 0.8731717, which indicates that there is a 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 by:

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

By using RStudio, test statistic t = 25.5241.

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

t	25.5241267815105

Calculate test statistic using RStudio

- Find critical value, using $\alpha = 0.05$, $df = n-2 = 203$
  From t-table, since this is a two-tailed test, there are two critical values:
  Lower tail critical value $-t_{\alpha/2=0.025,\ df=203} = -2.25815$
  Upper tail critical value $t_{\alpha/2=0.025,\ df=203} = 2.25815$
  From RStudio, we also get *p-value* = 2.2*e* - 16.
  Hence, if test statistics > 2.25815 / test statistics < -2.25815, reject H_0. Otherwise fail to reject H_0.

- State the 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 $\alpha = 0.05$.

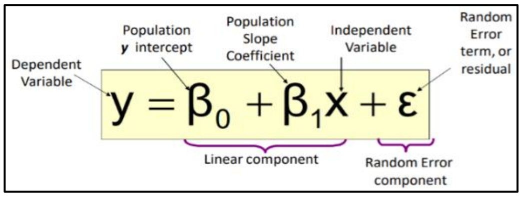
```
> cor.test(car_data$engine_size, car_data$price, method="pearson")

        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
```

Performing 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). Our regression model is a linear model, hence simple linear regression is used. The changes in the values of horsepower are assumed to be caused by the changes in the values of engine size.

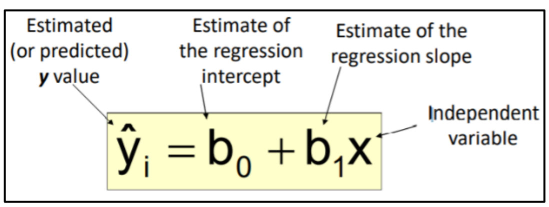
The mathematical equation for Population Linear Regression:



We assume that:

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

1. Estimated Regression Model:



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

- Find least squares criterion:
  From the above formula, we can find the values of b_0 and b_1 by:

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

$$b_0 = \bar{y} - b_1\bar{x}$$

```
> mean(x)
[1] 126.9073
> mean(y)
[1] 104.4098
> b0 <- mean(y)-(b1*mean(x))
b0    6.71330232533525
```

```
> 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
```

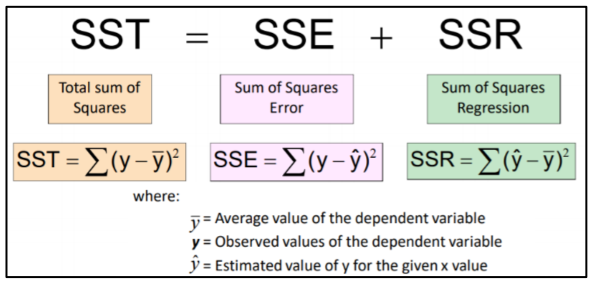
By using RStudio, we get b_1 = 0.7698, b_0 = 6.7133.

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

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

From the 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 $b_1 = 0.7698$ tells us 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

Calculate **SSR** using RStudio:

```
> yhat <- b0 + (b1*x)
> SSR <- sum((yhat-mean(y))^2)
SSR     209648.647452033
```

Calculate **SST** & **SSE** using RStudio:

```
> SST <- sum((y-mean(y))^2)
SST     319091.580487805
SSE <- SST-SSR
SSE     109442.933035772
```

- Find Coefficient of Determination, $\boldsymbol{R^2}$, by:

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

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

Finding R^2 using RStudio

By using RStudio, we get:

Coefficient of Determination, $\boldsymbol{R^2} = 0.6570$

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

By using RStudio, we get Standard Error of Estimate, $\boldsymbol{s_\varepsilon}$ = 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

By using RStudio, we get Standard Deviation of Regression Slope, $\boldsymbol{s_{b_1}}$ = 0.03904.

2. Inference about the Slope: **t-Test**
   - Hypothesis Statement:

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

     H_1: $\beta_1 \neq 0$ (linear relationship does exist)

   - Find critical value, using α = 0.05, df = n-2 = 203

     From *t*-table, since this is a two-tailed test, there are two critical values:

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

     Upper tail critical value $t_{\alpha/2=0.025,\ df=203}$ = 2.25815

     From RStudio, 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

By using RStudio, we get test statistic t = 19.7197.

- State the 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 α = 0.05.

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

- To perform linear regression in RStudio, we use the **lm()** function:

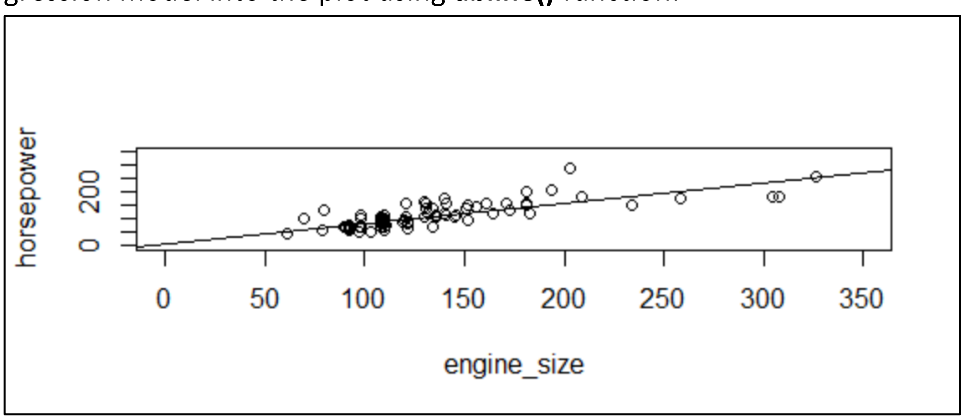
```
> model <- lm(y~x)
> model

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     6.7133       0.7698
```

From the image above, we can see the values of 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
```

When we view the summary of our linear regression model, we can get the values of intersection coefficient $\boldsymbol{b_0}$ = 6.7133, slope coefficient $\boldsymbol{b_1}$ = 0.7698, Standard Deviation of Regression Slope, $\boldsymbol{s_{b_1}}$ = 0.03904, Standard Error of Estimate, $\boldsymbol{s_\varepsilon}$ = 23.2191, *df* = 203, Coefficient of Determination, $\boldsymbol{R^2}$ = 0.6570 and ***p-value*** = 2.2*e* – 16.

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



- Goodness of Fit Test

In this analysis, we use variable **fuel_type**, where we will test whether there is difference between the observed frequency and expected frequency of fuel type used by cars, at 95% confidence level. Hence, we use goodness of fit test, or also known as the chi-square test with one-way contingency table, and we will be using unequal probabilities. There are two fuel types, that is gas fuel and diesel fuel. The fuel type percentages are claimed to be distributed with 85% gas fuel and 15% diesel fuel.

```
> table(car_data$fuel_type)

diesel    gas
    20    185
```

Observed frequencies for both fuel types

Our claim:
$p_{gas} = 0.85$, $p_{diesel} = 0.15$

1. Statement of test hypothesis:

   H_0: $p_{gas} = 0.85$, $p_{diesel} = 0.15$
   H_1: At least one of the proportions is different from the claimed value.

2. Calculate the expected frequency:

	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 statistic @chi-square value by:

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

   By using RStudio, test statistics value $x^2 = 4.4213$.

4. Find the critical value:

   Critical value $x^2 = 3.841$
   (with k-1 = 1 and $\alpha = 0.05$)

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

x2.alpha	3.84145882069413

Finding critical value x^2 using RStudio

5. State the decision:
   Since test statistic value (x^2 = 4.4213) > critical value($x^2_{k=1,\ \alpha = 0.05}$ = 3.841), it falls within the critical region. Thus, we **reject** H_0. There is sufficient evidence to warrant rejection of the claim that the fuel types are distributed with the given percentages, this shows that there is difference 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
```

Performing chi-square test using RStudio, value of test statistics x^2 = 4.4213, *p-value* = 0.03549

- Chi-Square Test of Independence

In this analysis, we are using variables **num_of_doors** and **aspiration**, where we will test whether number of doors and car aspiration are related using Two Way Contingency Table, at 95% confidence level. Hence, 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
```

Observed Frequencies for variables num_of_doors and aspiration

1. State the test hypothesis:
   H_0: There is no relationship between variables num_of_doors and aspiration.
   H_1: Variables num_of_doors and aspiration are related and dependent.

2. Find the critical value:
   Critical value x^2 = 3.841
   (with *df*=(2-1)(2-1)=1, α = 0.05)

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

x2.alpha	3.84145882069413

Finding critical value x^2 using RStudio

3. 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

4. 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.

5. State the 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 α = 0.05.

7.0 Conclusion

For the 2 sample hypothesis testing, where we test on the mean assuming unequal variances, we found out that the mean of horsepower of turbo-aspirated cars is larger than the mean of horsepower of standard-aspirated cars, hence we **reject** null hypothesis. In real world, this conclusive statement could be true considering that turbo-aspirated cars need to have larger horsepower to enhance car performance, in terms of the engine efficiency, car 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 real world, although car engine size is not the only factor affecting car price, but larger engine size truly affects car price, where larger engine will result in higher car price, because larger engine would have 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: $\hat{y}_i = 6.7133 + 0.7698x$, hence we **reject** null hypothesis. We can say that there is a positive linear relationship between the engine size and the car horsepower. In real world, this statement is also true, where larger engine is usually more powerful, this is a very important specification to be considered for race drivers as larger engine size boasts more power to make the car more agile and performance-enhanced.

For the goodness of fit test, we found out that v, that is gas fuel and diesel fuel, hence we **reject** null hypothesis. We can conclude that the observed frequency is not a good fit to the assumed distribution. In real world, if we compare gas-fueled cars and diesel-fueled cars, gas-fueled cars are more than diesel-fueled cars. Although diesel engines are more efficient, they are more expensive, less readily available, and not environmental-friendly, this could be the reason for larger number of gas-fueled cars in real world.

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 purposes, where we normally see sportier cars with only two car doors and normal sedan-cars with four car doors.

In conclusion, I can perform test analysis such as 2 sample hypothesis testing, correlation analysis, regression analysis, goodness of fit test and chi-square test of independence using RStudio. I believe this project is very useful for my future as this project has developed my data analysis skills. Also, special thanks to our lecturer, Dr. Chan Weng Howe for his help and guidance throughout this project.

8.0 References

Eleanor Xu (2016, Dec 25). Sample of Car Data. Retrieved from: https://www.kaggle.com/jingbinxu/sample-of-car-data