Title Page

Clover Pike High School

PSDA Project 2 Presentation

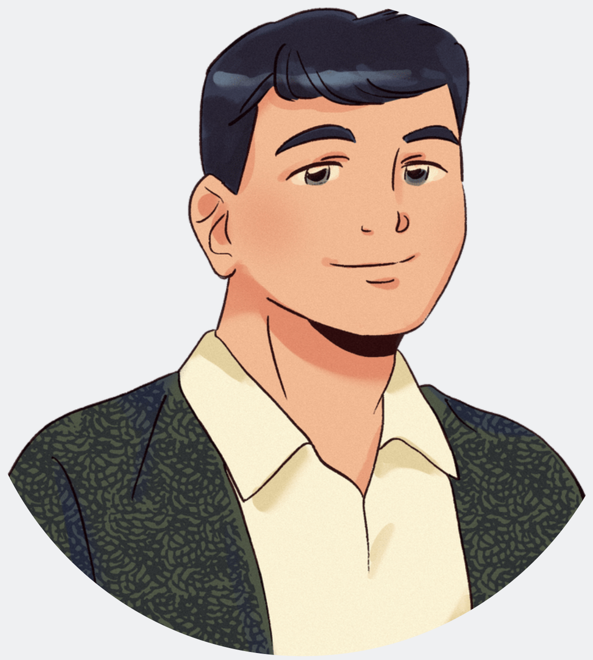
4 of a kind

Title Page

Proprietors

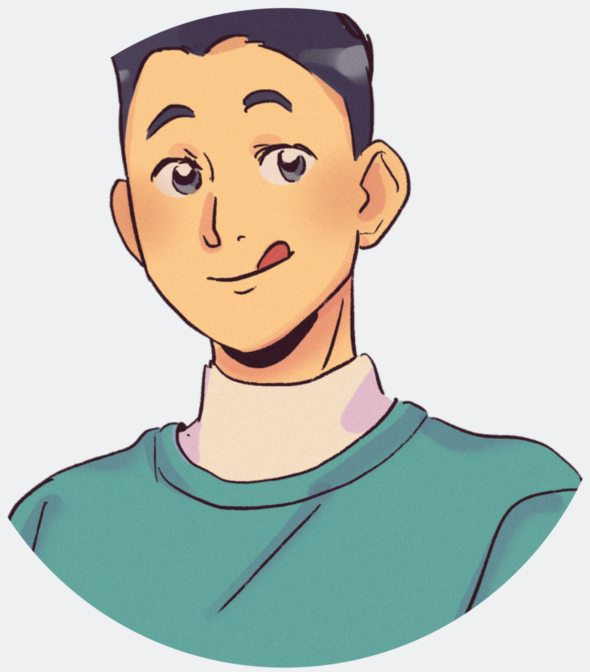
Team Members

4 of a kind

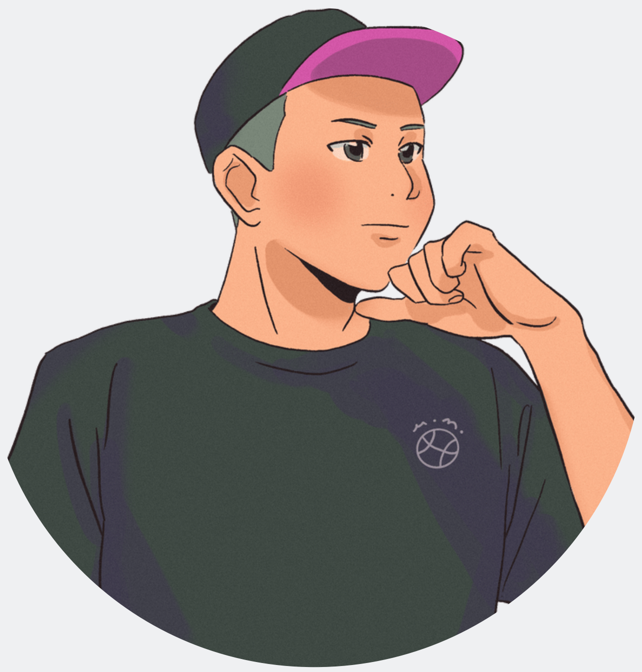


LIM SHI KAI

Group Leader



PHANG SENG SOON



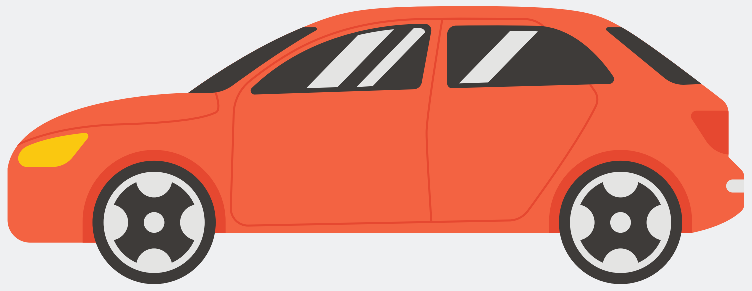
TAN CHUN MING

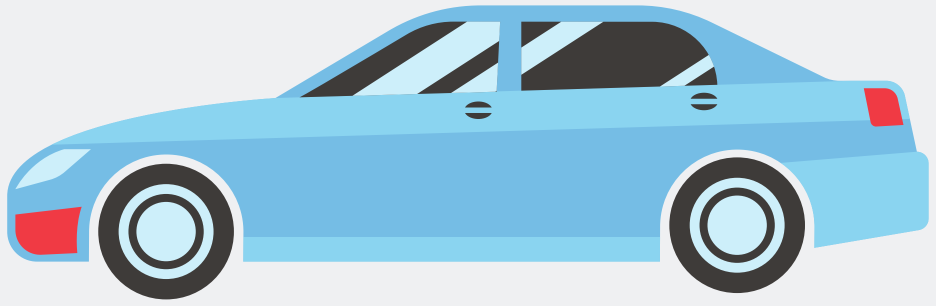
NG KENG KEAT

Introduction

Cars, an automobile 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.

Introduction

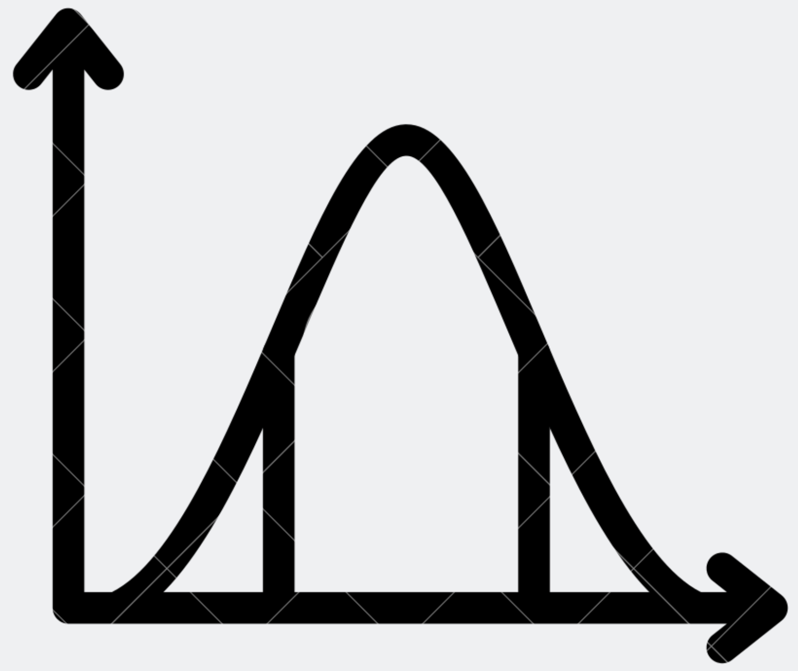




Nowadays, consumers' choice expands as automakers release an increasing number of car models. Every car has its specifications in terms of horsepower, car body configuration, fuel consumption, price and many more.

The primary goal of this study is to display crucial facts about a vehicle which is its specification. 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.

Background



We retrieved this dataset from Kaggle. 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. 2 sample hypothesis testing, correlation analysis, regression analysis, goodness of fit test and chi-square test of independence will be used to test the chosen variables in our project.

Data Analysis

2-sample hypothesis testing

To determine if the mean horsepower of the turbocharged car is greater than the mean horsepower of naturally aspirated cars at a 95% confidence level, assuming unequal variances.

Correlation

To measure the strength of the linear relationship between engine size and car price at 95% confidence level.

Regression

To identify is there a relationship between a dependent variable (engine size) and independent variable (horsepower).

Title Page Proprietors Introduction Data Analysis

Data Analysis

Data Analysis

Goodness of Fit Test

To test the difference between the observed frequency and expected frequency of fuel type used by cars at 95% confidence level.

Chi-Square Test of Independence

To confirm that the relationship between
the two qualitative variables, which are the number of doors and aspiration at 95%
confidence level.

Insert your topic here

Hypothesis Testing by using 2-sample

A method used to test whether the unknown population means of two groups are equal or not

Objective: To test whether the mean of the horsepower of turbo-aspirated is the same with the mean of the horsepower of standard-aspirated by using 95% confidence level and assuming the variances are unequal

Sample 1

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

Sample 2

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

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

Hypothesis Testing

H0: μ1 = μ2

H1: μ1 > μ2

Test Statistic

$$t_0 = \frac{\overline{x}_1 - \overline{x}_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = 4.08037$$

Insert your topic here

Degree of freedom

$$v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}} = 64.7109 \approx 64$$

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

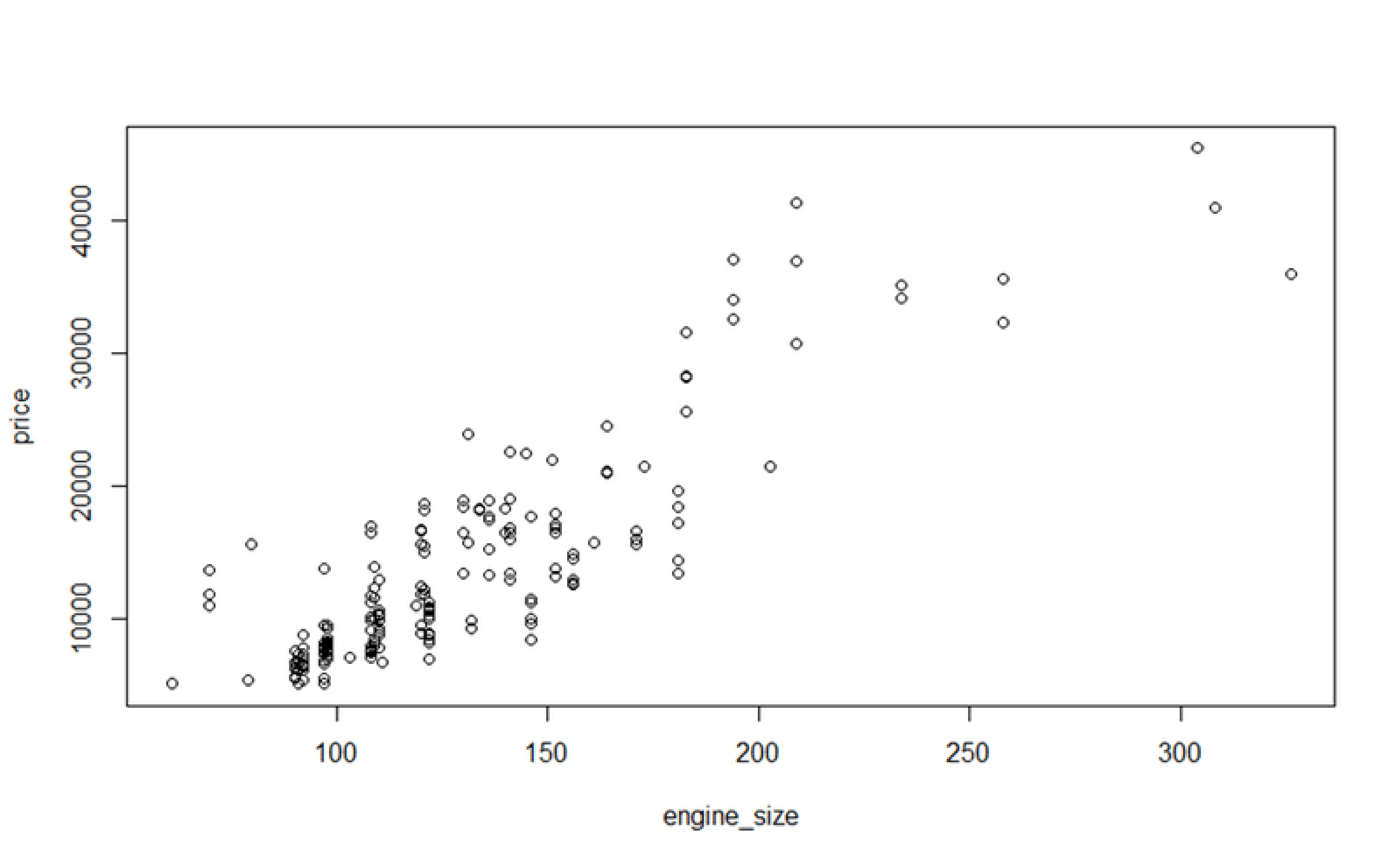
t.alpha	-1.66901302502409

$$\therefore \ Reject \ H_0$$

Correlation

To test relationship between quantitative variables or categorical variables

Objective: To test is there a linear relationship between engine size and car price



Sample correlation coefficient

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

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

r	0.873171748808439

Significance Test for Correlation

Hypothesis Testing

H0: ρ = 0 (no linear correlation)

H1: ρ ≠ 0 (linear correlation exists)

Test Statistic

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

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

Critical value, t

```
> 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
```

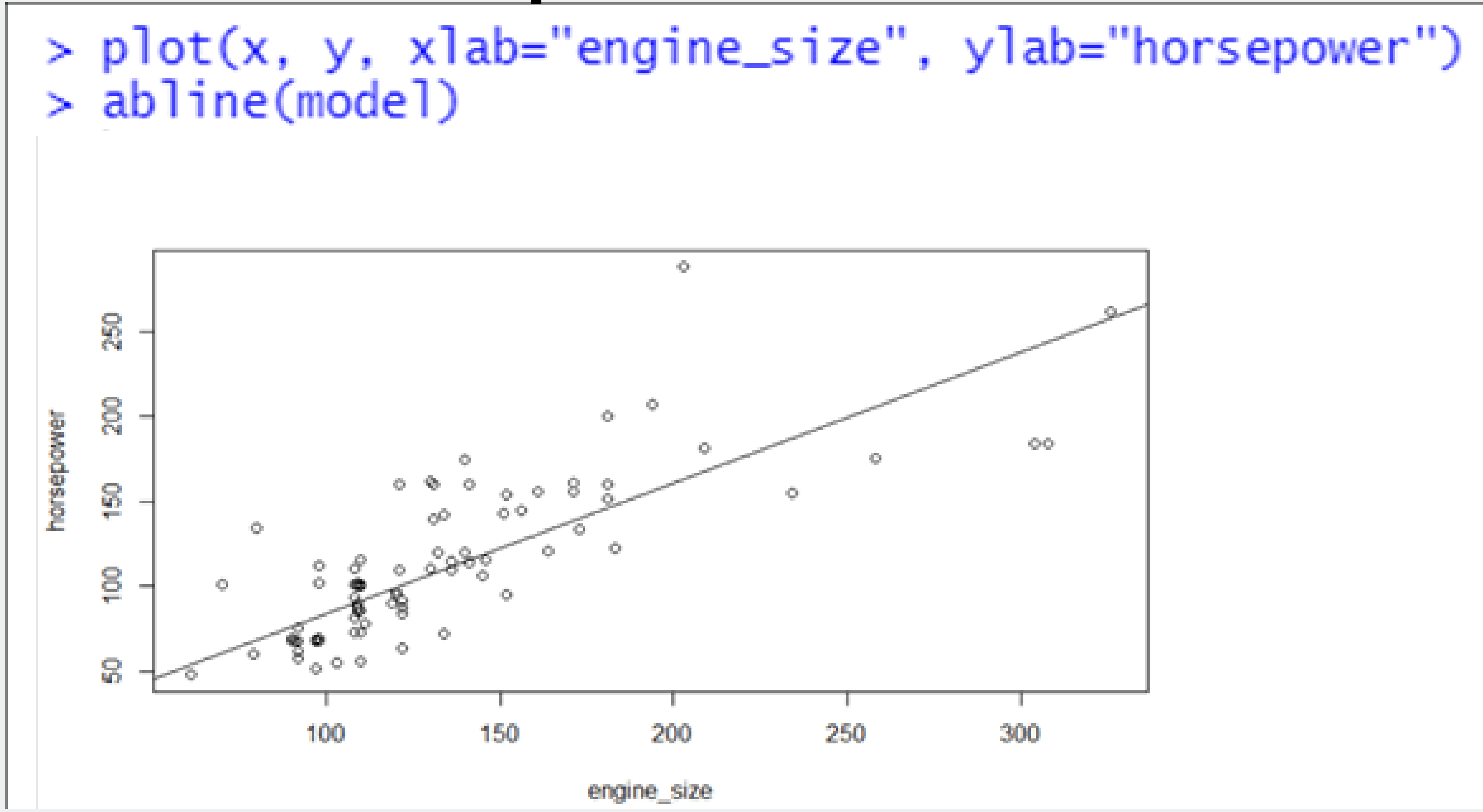
Title Page | Proprietors | Introduction | Data Analysis | Hypothesis.... | Correlation | Regression

Insert your topic here

Regression

A set of statistical methods ued for the estimation of relationship between a dependent variable and one or more independent variable

Objective: To investigate whether engine size has an impact on how much horsepower an engine produces



Insert your topic here

Estimated Regression model

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

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

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

```
> 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 ...

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

b0	6.7133023253525

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

Explained and unexplained variation

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

```
> 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

Coefficient of Determination

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

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

R2	0.657017170843358

Standard Error of Estimate

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

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

Se	23.219124637784

Standard Deviation of Regression Slope

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

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

Sb1	0.0390383943909781

t-test

Hypothesis Testing

H0: β1 = 0 (no linear relationship)

H1: β1≠ 0 (linear relationship exists)

Test Statistic

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

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

t	19.7196948246796

$\therefore Reject\ H_0$

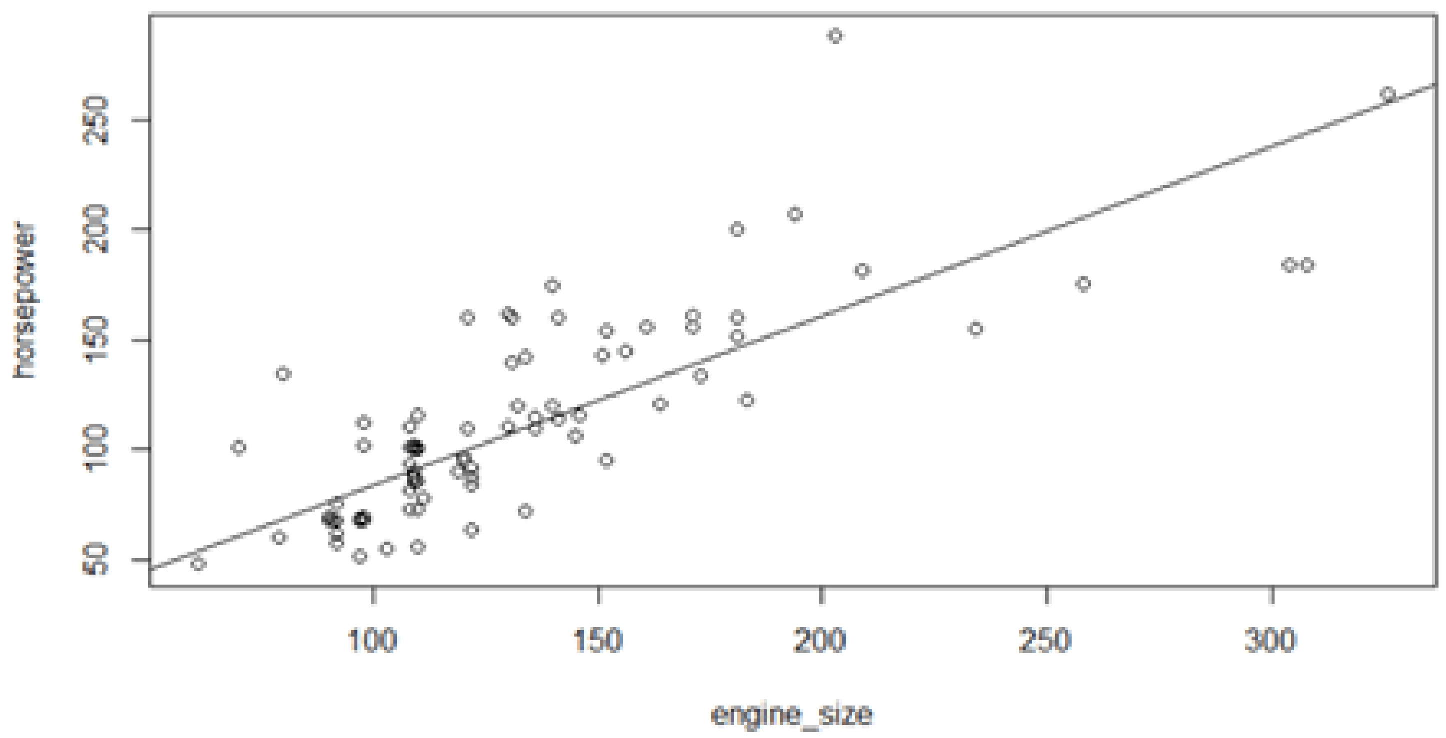
Performing linear regression on RStudio using lm() function

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

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     6.7133       0.7698
```

Insert your topic here

```
> 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
```

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



Goodness of Fit Test

Variable used: fuel_type

Objectives: To test the difference between the observed frequency and expected frequency of fuel type used by cars at a 95% confidence level.

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

Our claim:

$\rho_{gas}=0.85$, $\rho_{diesel}=0.15$

85% gas fuel

15% diesel fuel

Insert your topic here

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

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
Observed Frequency, O	
Expected Frequency, E	np=(

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

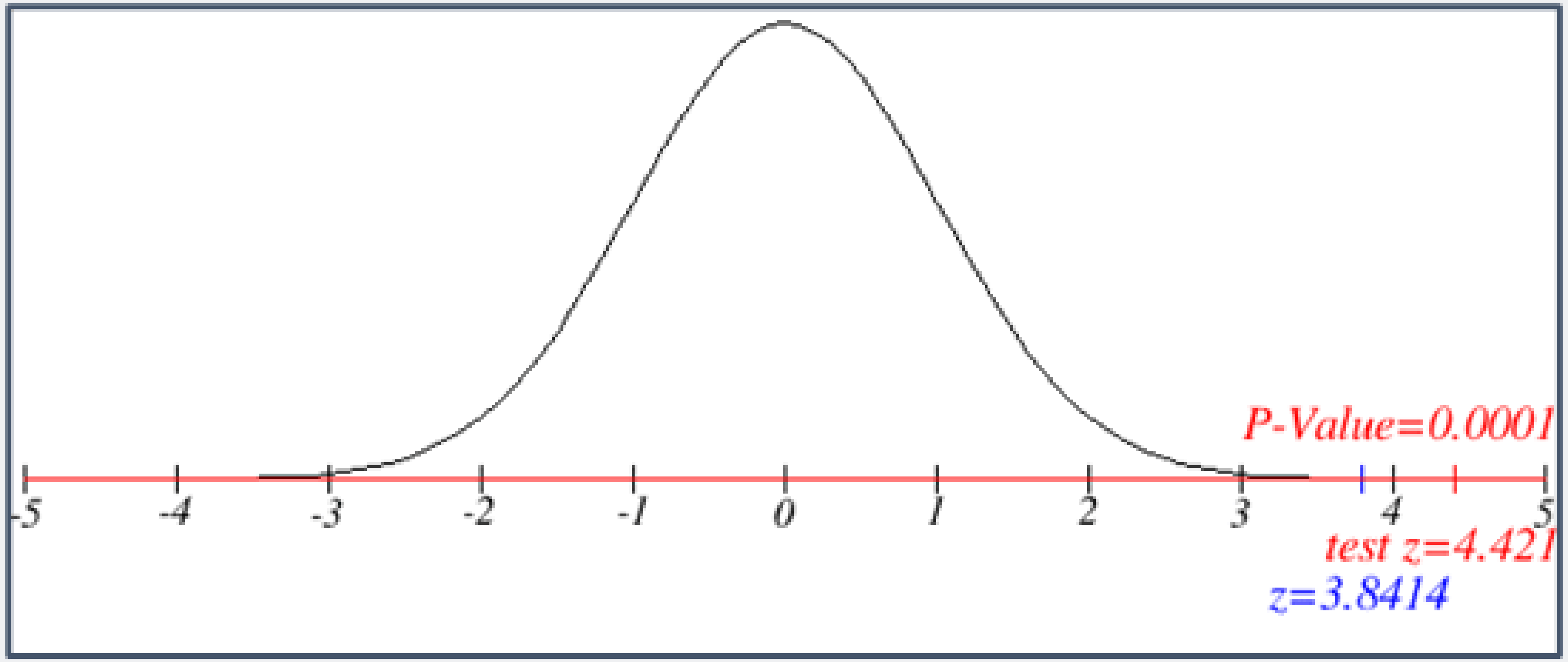
Critical value, $x^2 = 3.8414$ (with df = k-1 = 1 and $\alpha = 0.05$)

Title Pag | Proprietors | Introduction | Data Analysis | Hypothesi | Corre | Regressio | Goodne | Chi-square test of ...

Insert your topic here

5. Conclusion

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



Chi-square test of independence

Variable used: num_of_doors and aspiration

Objectives: To confirm that the relationship between the two qualitative variables exists at a 95% confidence level when using 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

Insert your topic here

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 $\alpha = 0.05$)

Insert your topic here

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 <- q
> x2.alpha
[1] 3.841459
```

Critical v

Critical value, x^2 =

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

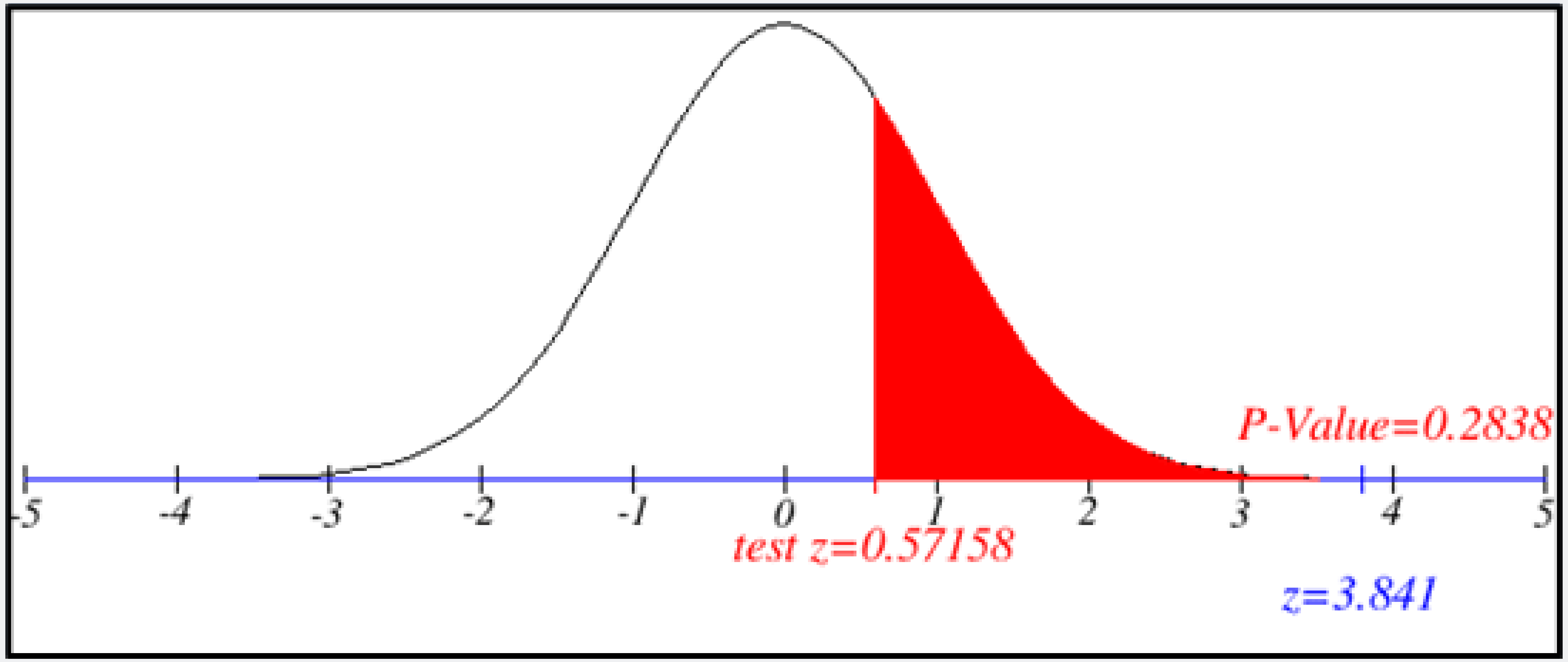
```
> tbl = table(car_data$num_of_doors, car_data$aspiration)
> chisq.test(tbl, correct=FALSE)

        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.

Insert your topic here

5. Conclusion

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



Insert your topic here

Conclusion

1.Two Sample Hypothesis Testing

- Test Statistics value > Critical Value
- So, reject the null hypothesis

2. Correlation Analysis

- Test Statistics value > Upper tail critical value
- So, reject the null hypothesis

Insert your topic here

3. Regression

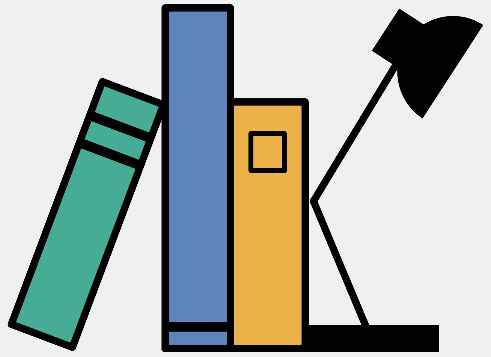
- Test Statistics value > upper tail critical value
- So, reject the null hypothesis

4. Goodness of Fit Test

- Test Statistics value > critical value and falls within the critical region
- So, reject the null hypothesis

5. Chi-square Test of Independence

- Test Statistics value < critical value
- So, fail to reject the null hypothesis



References

01 **Lecture Note - Chapter 5 Part 3**

02 **Lecture Note - Chapter 6**

03 **Lecture Note - Chapter 7**

04 **Online Resources - Statistics LibreTexts**

THANK YOU