FINAL

library(tidyverse)  
data(mpg)

## Soalan 2 / *Question2*

Jawab soalan di bawah dengan menulis kod-kod R yang bersesuaian. Ubahkan menjadi dokumen word markdown dengan penerangan yang munasabah untuk setiap langkah. Hantarkan fail R markdown tersebut satu hari selepas soalan dibuka di ruangan penghantaran di bawah.

*Answer the question below by writing the appropriate R codes. Transform them into an R markdown word document with proper explanation for each step. Submit the R markdown file one day after the question is opened to the submission area below.*

### Question a

1. Mulakan Markdown anda dengan pembuka kata menerangkan apa itu Pemvisualan Data dan kegunaannya.

* *Begin your markdown with an introduction explaining what is data visualization and its use.*

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. As the “age of Big Data” kicks into high-gear, visualization is an increasingly key tool to make sense of the trillions of rows of data generated every day. A good visualization tells a story, removing the noise from data and highlighting the useful information.

### Question b

1. Gunakan set data yang sama untuk semua geraf. Berikan penerangan mengenai set data tersebut.

* *Use the same dataset for all graphs. Provide some explanation regarding the dataset.*

We can see that the dataset has 234 rows and 11 columns, and it contains 0 null field.

dim(mpg)

## [1] 234 11

sum(is.na(mpg))

## [1] 0

As the dataset is from **ggplot2**, we can see the explanation of each variable using the **?*dataset*** command

?mpg

|  |  |
| --- | --- |
| Features | Explanation |
| manufacturer | manufacturer name |
| model | model name |
| displ | engine displacemnt, in litres |
| year | year of manufacture |
| cyl | number of cylinders |
| trans | type of transmission |
| drv | the type of drive train, where f=font-wheel drive, r=rear wheel drive, 4=4wd |
| cty | city miles per gallon |
| hwy | highway miles per gallon |
| fl | fuel type |
| class | “type” of car |

Out of 11 variables:

6 are categorical variables

* **manufacturer**, **model**, **trans**, **drv**, **f1**, and **class**

5 are numerical variables

* **displ**, **year**, **cyl**, **cty**, and **hwy**

summary(mpg)

## manufacturer model displ year   
## Length:234 Length:234 Min. :1.600 Min. :1999   
## Class :character Class :character 1st Qu.:2.400 1st Qu.:1999   
## Mode :character Mode :character Median :3.300 Median :2004   
## Mean :3.472 Mean :2004   
## 3rd Qu.:4.600 3rd Qu.:2008   
## Max. :7.000 Max. :2008   
## cyl trans drv cty   
## Min. :4.000 Length:234 Length:234 Min. : 9.00   
## 1st Qu.:4.000 Class :character Class :character 1st Qu.:14.00   
## Median :6.000 Mode :character Mode :character Median :17.00   
## Mean :5.889 Mean :16.86   
## 3rd Qu.:8.000 3rd Qu.:19.00   
## Max. :8.000 Max. :35.00   
## hwy fl class   
## Min. :12.00 Length:234 Length:234   
## 1st Qu.:18.00 Class :character Class :character   
## Median :24.00 Mode :character Mode :character   
## Mean :23.44   
## 3rd Qu.:27.00   
## Max. :44.00

Let’s have a peek at the first five rows of **mpg** dataset

head(mpg)

## # A tibble: 6 x 11  
## manufacturer model displ year cyl trans drv cty hwy fl class   
## <chr> <chr> <dbl> <int> <int> <chr> <chr> <int> <int> <chr> <chr>   
## 1 audi a4 1.8 1999 4 auto(l5) f 18 29 p compa~  
## 2 audi a4 1.8 1999 4 manual(m5) f 21 29 p compa~  
## 3 audi a4 2 2008 4 manual(m6) f 20 31 p compa~  
## 4 audi a4 2 2008 4 auto(av) f 21 30 p compa~  
## 5 audi a4 2.8 1999 6 auto(l5) f 16 26 p compa~  
## 6 audi a4 2.8 1999 6 manual(m5) f 18 26 p compa~

### Question c

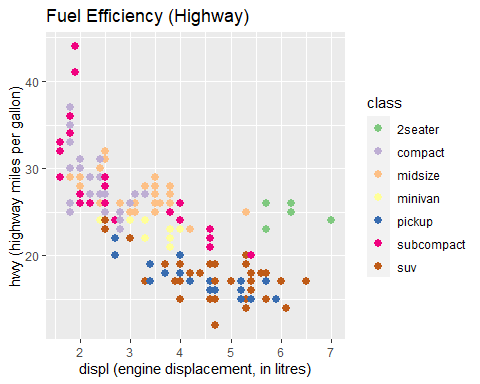
1. Tuliskan kod-kod R yang menghasilkan geraf-geraf scatterplot, histogram, dan boxplot. Jelaskan kegunaan setiap geraf dan berikan penerangan ringkas bagi setiap geraf yang terhasil.

* *Write R codes to produce a scatterplot, a histogram, and a boxplot. Explain the use of each graph and provide a short explanation for each produced graph.*

**Scatter Plot**

Scatter plots are used to plot data points on a horizontal and a vertical axis in the attempt to show how much one variable is affected by another.

ggplot(mpg, aes(x=displ, y=hwy)) +   
 geom\_point(mapping=aes(color=class), size=2.5) +   
 labs(title='Fuel Efficiency (Highway)',   
 x='displ (engine displacement, in litres)',   
 y='hwy (highway miles per gallon)') +   
 scale\_color\_brewer(palette='Accent')

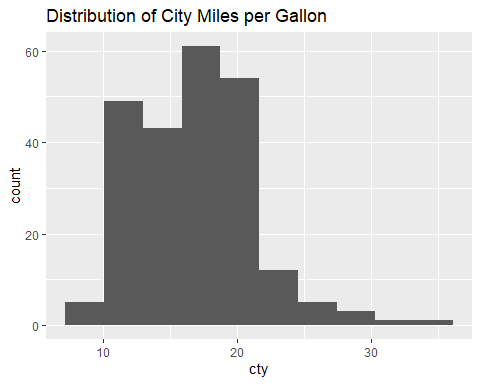


Here, we investigate the fuel efficiency by plotting **hwy** against **displ** and we see that fuel efficiency generally decreases with engine size. Nevertheless, two seaters (sport cars) are an exception maybe because of their light weight.

**Histogram**

Histograms are used to graphically summarize the distribution of a variable.

ggplot(mpg) +   
 geom\_histogram(mapping=aes(cty), bins=10) +   
 labs(title='Distribution of City Miles per Gallon')

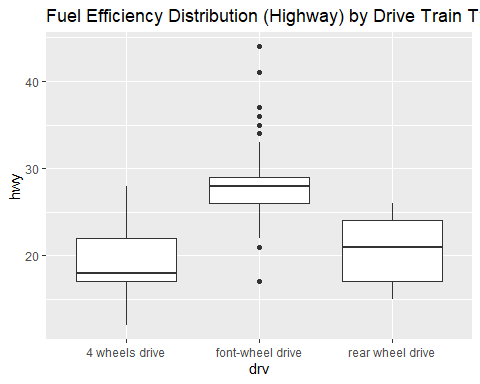


From the plot, we can see that the distribution of cty (city miles per gallon) is not symmetric but skewed to the right, this means that the distribution of the variable is most probably not normal. However, normalizing transformation such as the box-cox transformation can be considered.

**Boxplot**

Boxplots are used in explanatory data analysis to visually show the distribution of numerical data and skewness through displaying the data quartiles (or percentiles) and averages. It can also be useful when comparing distributions between many groups.

ggplot(mpg) +   
 geom\_boxplot(aes(drv, hwy)) +   
 labs(title='Fuel Efficiency Distribution (Highway) by Drive Train Types') +   
 scale\_x\_discrete(labels=c('4'='4 wheels drive', 'f'='font-wheel drive', 'r'='rear wheel drive'))



By using boxplot here, we compare the distribution of highway miles per gallon for three different types of drive train. We can see that the distributions are all not quite similar. Front-wheel drive (f) with the highest median outperforms two other types, and there are some outliers in it which could be caused by classes of the vehicles.