Introduction to Data Science (11372 & G 11516)

Semester 1 2021

UNIVERSITY OF CANBERRA

INTRODUCTION TO DATA SCIENCE

Lecture 10

Dr. Ibrahim Radwan

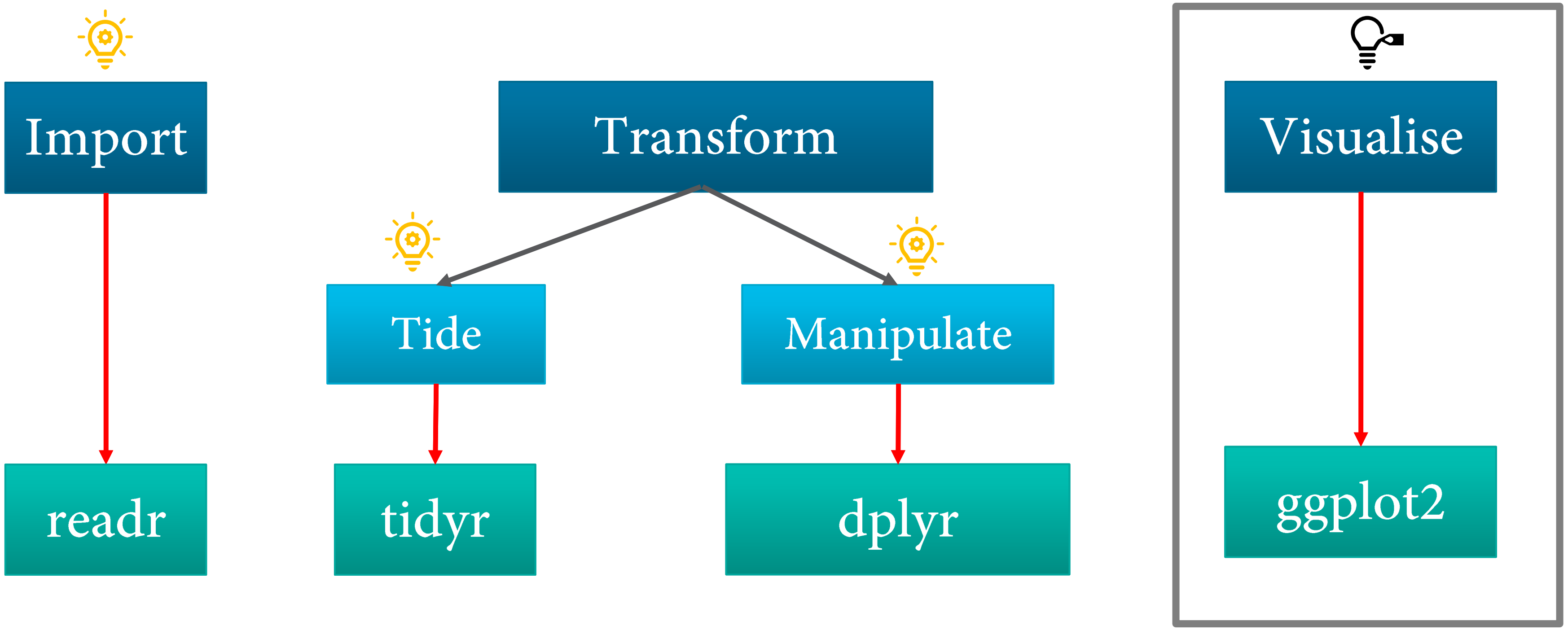
DISTINCTIVE BY DESIGN

OUTLINE

- Data Wrangling, a recap
- Exploratory Data Analysis, a recap
- EDA; Univariate Analysis
- EDA; Bivariate Analysis
  - Continuous
  - Continuous + discrete
  - Discrete

DATA WRANGLING

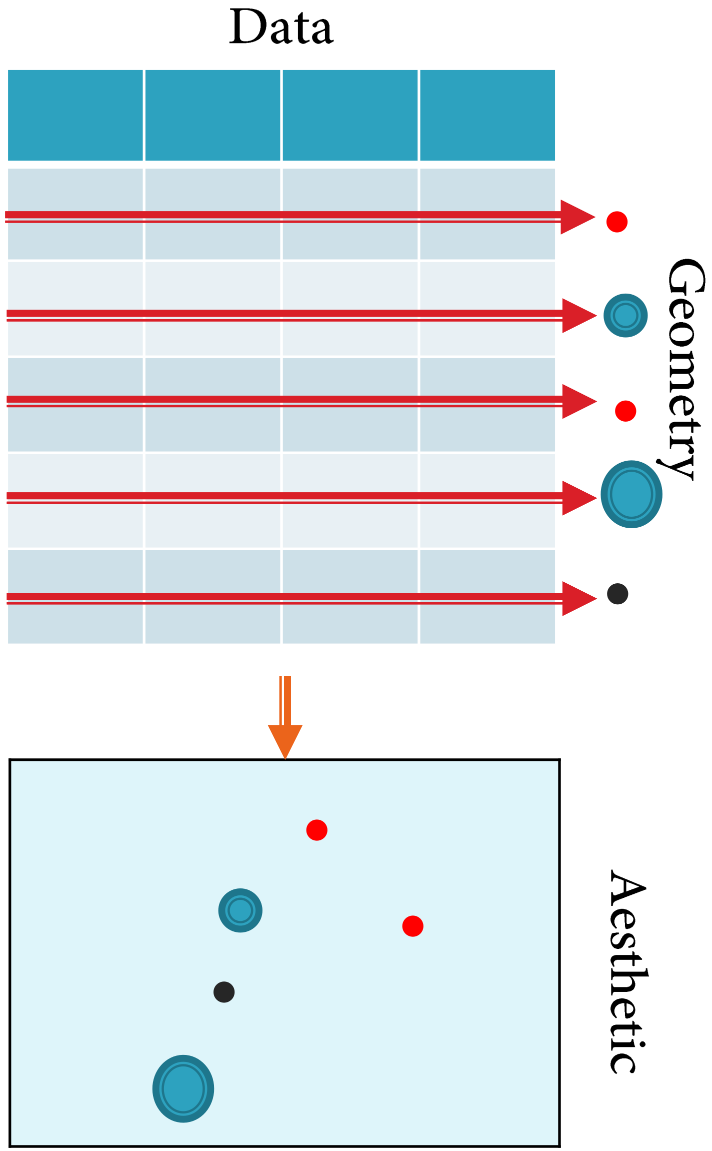
- Practically, we have three main processes to wrangle the data



GRAPH COMPONENTS

- To build a graph in R , you will need to specify three components:

  1. **Data:** the set of records/variables that we need to represent with a graph
  2. **Geometry:** the type of the plot, which will be generated, usually it is a function such as (scatterplot, boxplot, barplot, histogram, smooth density, etc.)
  3. **Aesthetic mapping**: the coordinate map and the other visual cues, such as size, scale and color.



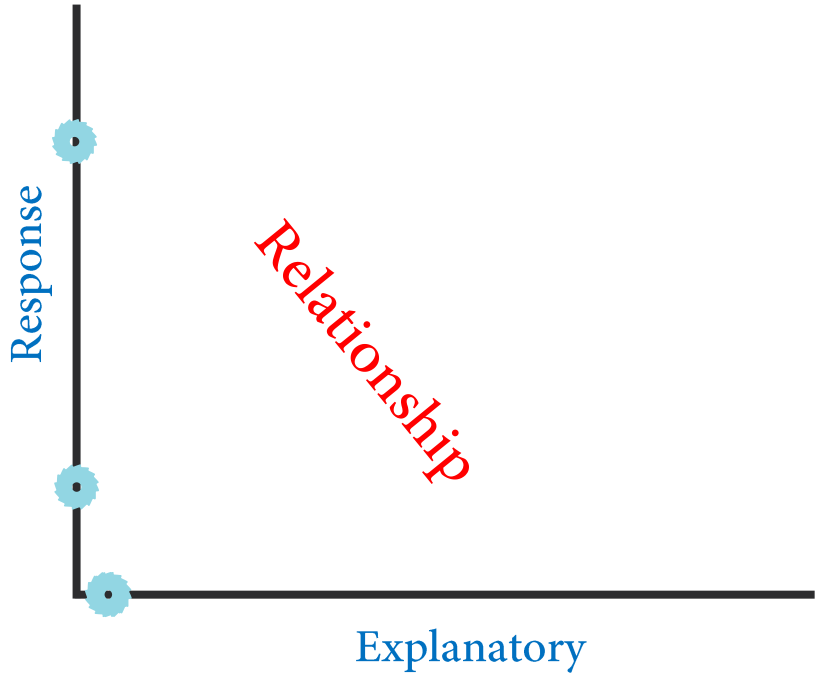
EXPLORATORY DATA ANALYSIS (EDA)

- To extract the relationships between variables or to discover the patterns/distributions of the variables, we need to check on the types of these variables.
- To conduct the data analysis on variables for sake of understanding their relationships, this analysis can be either:
  - Uni-variate analysis
    - Discover the variations of the data into **one** variable
  - Multi-variate analysis
    - Discover the co-variation of **multiple** variables
    - Bi-variate analysis is a special type of this analysis with only **two** variables.

EDA – UNIVARIATE (RECAP)

- There are two types of visualization-based univariate analysis:
  - Variation of a continuous variable
  - Variation of a discrete variable
- Examples of univariate continuous :
  - Histograms, etc.
- Examples of univariate discrete:
  - bar plots, etc.

EDA – BI-VARIATE ANALYSIS



©Dr. Ibrahim Radwan – University of Canberra

EDA – BI-VARIATE ANALYSIS (2)

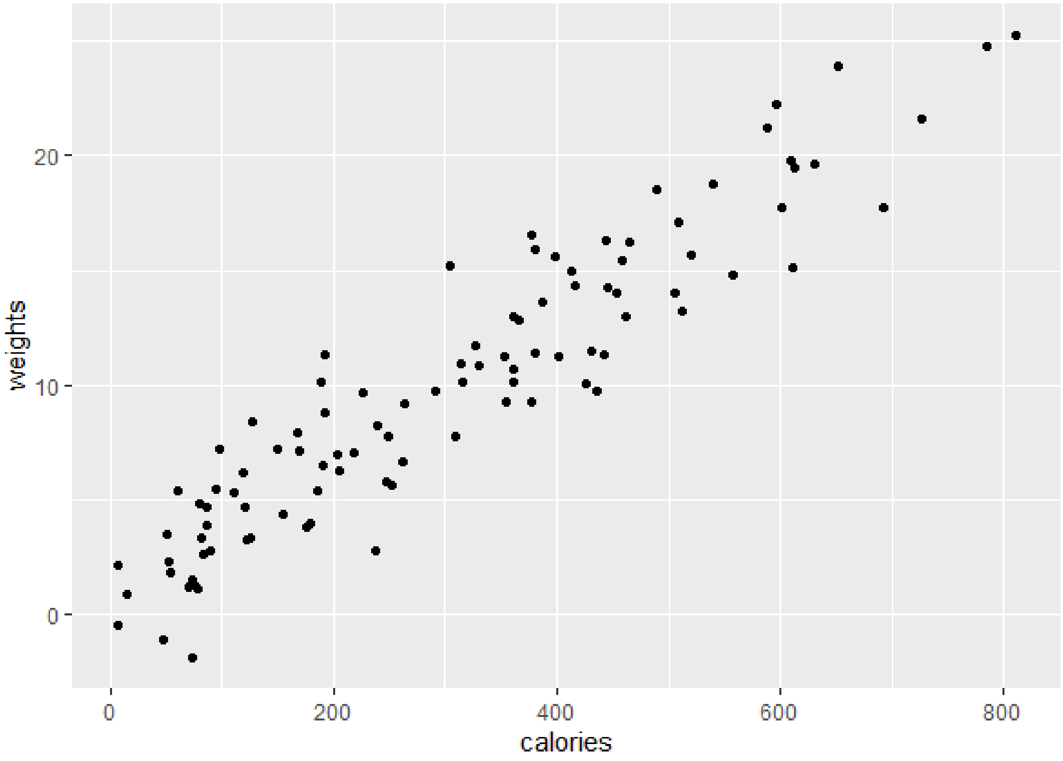
- The bivariate analysis will be done for each of the following pairs:
  - Bivariate analysis of a continuous variable with respect to another continuous variable
  - Bivariate analysis of a continuous variable with respect to a discrete variable
  - Bivariate analysis of a discrete variable with respect to another discrete variable

EDA – BIVARIATE, CONTINUOUS

- The analysis is done to inspect the relationship between two variables, where both of them are 'numerical', continuous variables.
- ggplot2 provides many functions to plot the variation between these two variables such as:
    - *geom_point()*, for a scatter plot
    - *geom_jitter(),* for scatter plot with displacing the overlapped points a bit away from each other
    - Etc.

SCATTER PLOTS

- Summarise the relationship between two variables as scatter dots in the cartesian plane.
- A scatter plot can be used to check properties, such as:
  - Correlation between variables;
  - Presence of outliers



BUILDING A SCATTERPLOT

1- a) First layer contains the data specification
   b) Also specifies the two variables

2- a) Second layer specifies the type of plot; the geometric function

3- a) Third and following layers specify axis labels and further visual cues

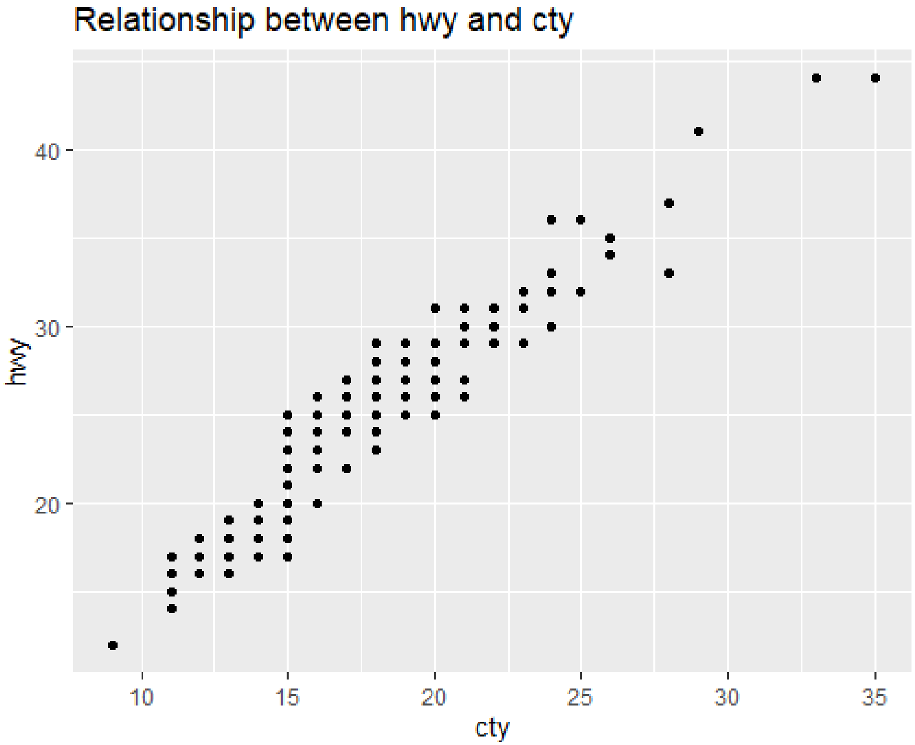
```
ggplot(data,aes(x= explanatory, y= response))+
  geom_point() +
  xlab(“x axis label”) +
  ylab(“y axis label”)+
  ggtitle(“plot title”)
```

©Dr. Ibrahim Radwan – University of Canberra

EXAMPLE

Basic scatterplot visualization

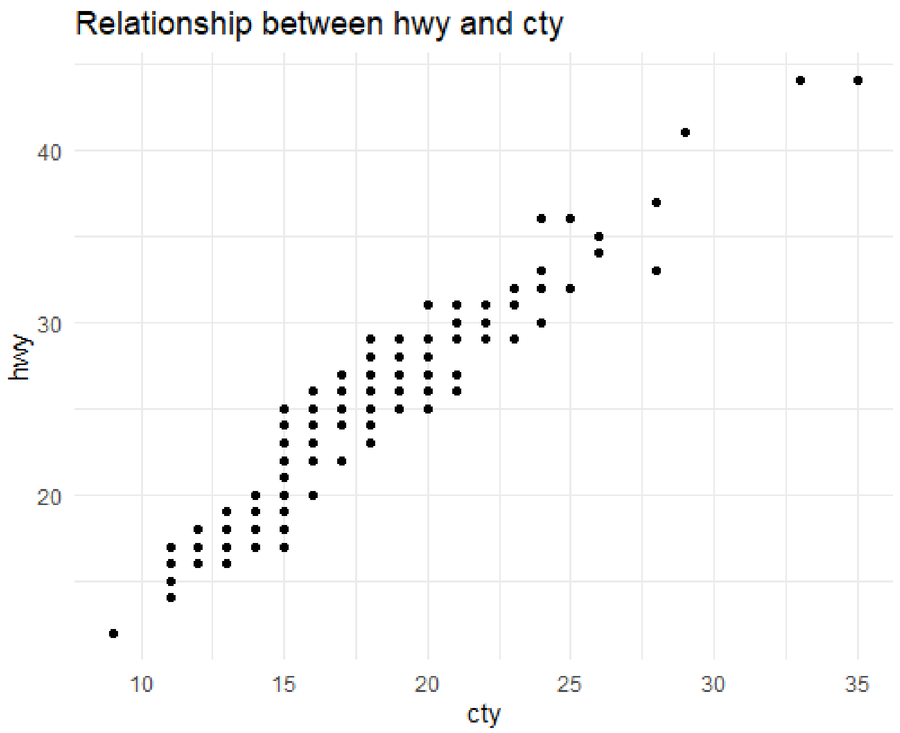
```
ggplot(data= mpg, aes(x= cty, y= hwy)) +
  geom_point() +
  theme_minimal() +
  ggtitle("Relationship between hwy and cty")
```

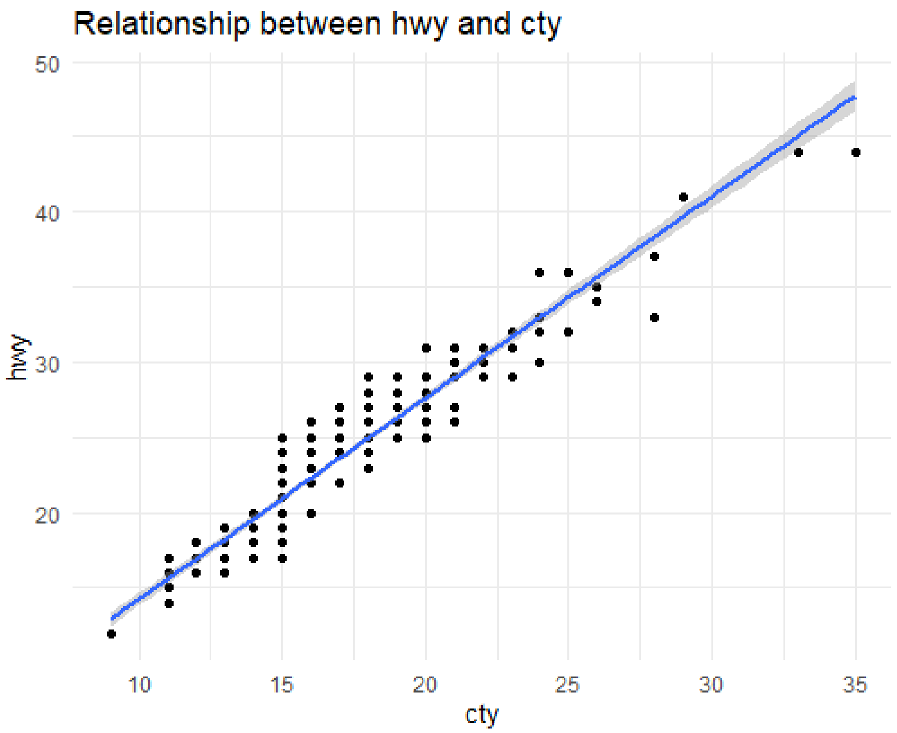


©Dr. Ibrahim Radwan – University of Canberra

EXAMPLE (2)

Adding extra visual cues and more options

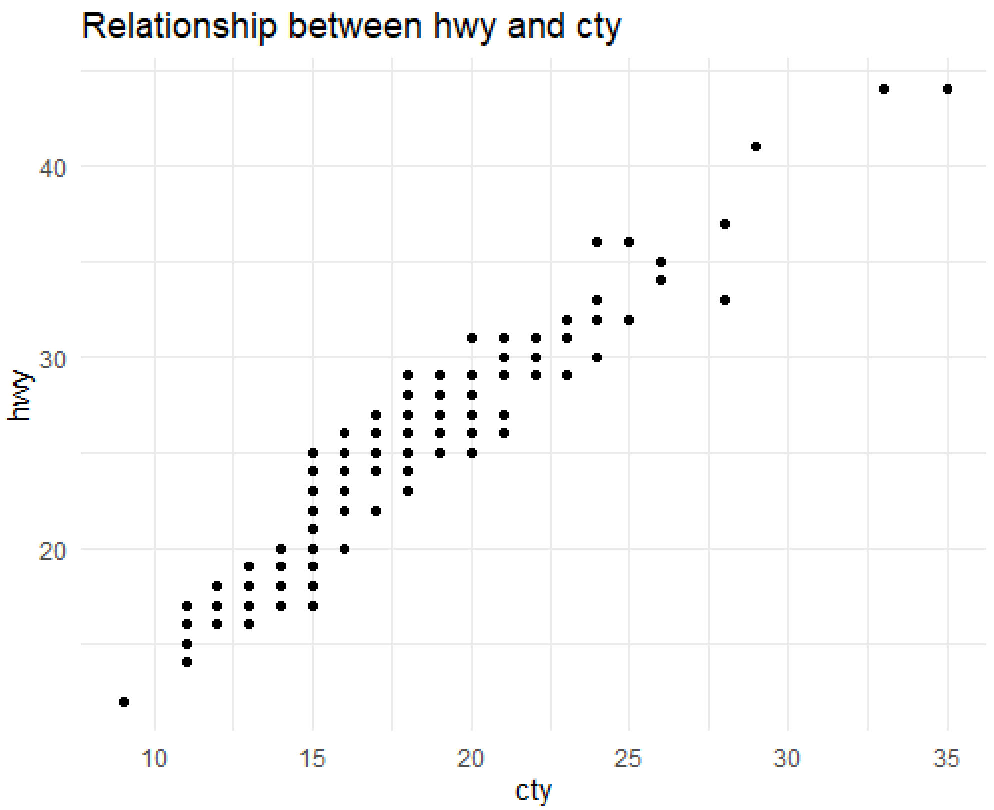

Relationship between hwy and cty


Relationship between hwy and cty

The full code is shared under week 10 on Canvas

©Dr. Ibrahim Radwan – University of Canberra

SCATTER PLOT INTERPRETATION



Scatter plot can be used to answer these questions:

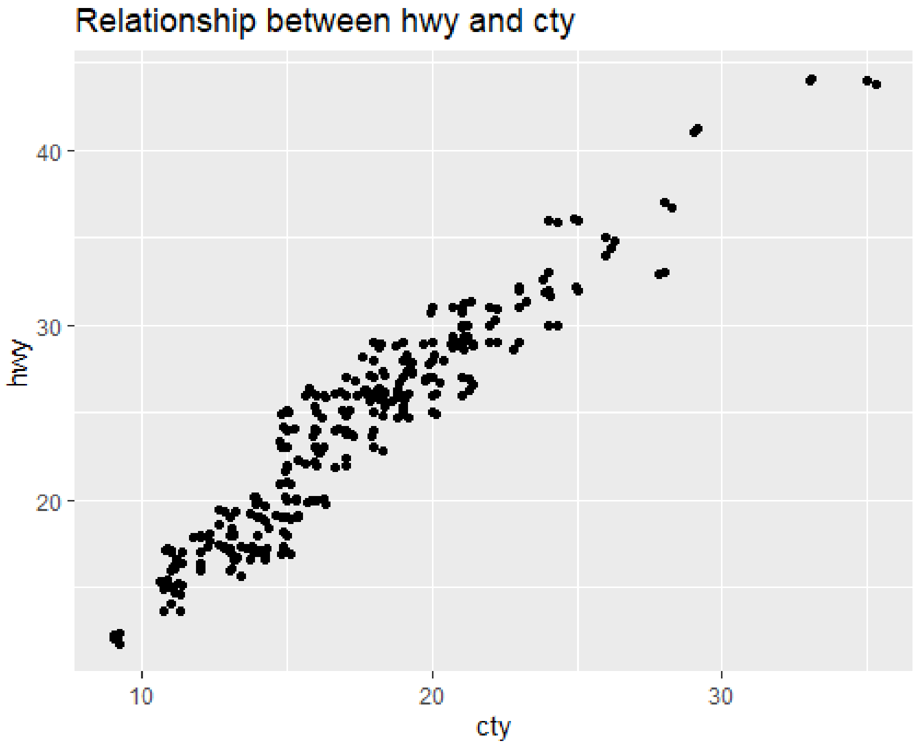
- Are the variables related? [**Form**]
- Are they positively or negatively related? [**Direction**]
- What is the strength of the correlation?[**Strength**]
- Are there outliers?

©Dr. Ibrahim Radwan – University of Canberra

EXAMPLE OF JITTER PLOT

Basic jitterplot visualization

```
ggplot(data= mpg, aes(x= cty, y= hwy)) +
  geom_jitter() +
  theme_minimal() +
  ggtitle("Relationship between hwy and cty")
```

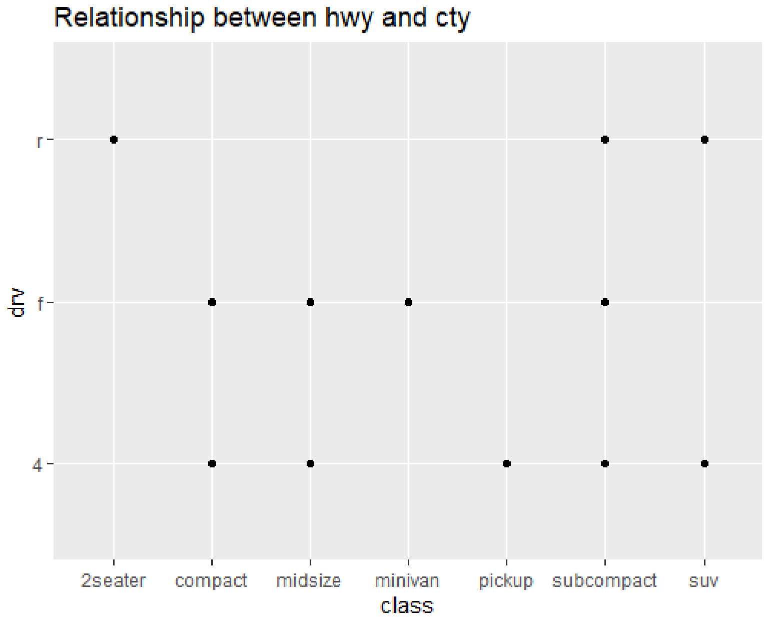


©Dr. Ibrahim Radwan – University of Canberra

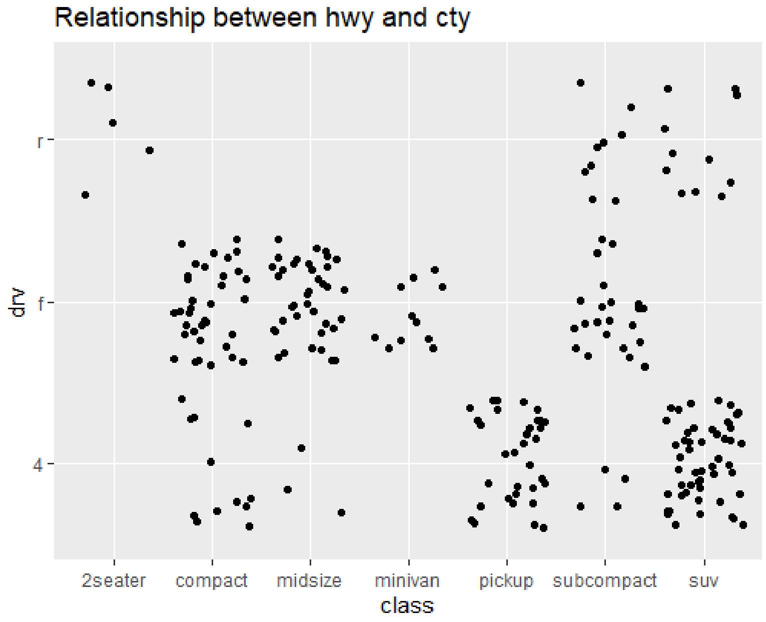
EXAMPLE OF JITTER PLOT (2)

UNIVERSITY OF CANBERRA

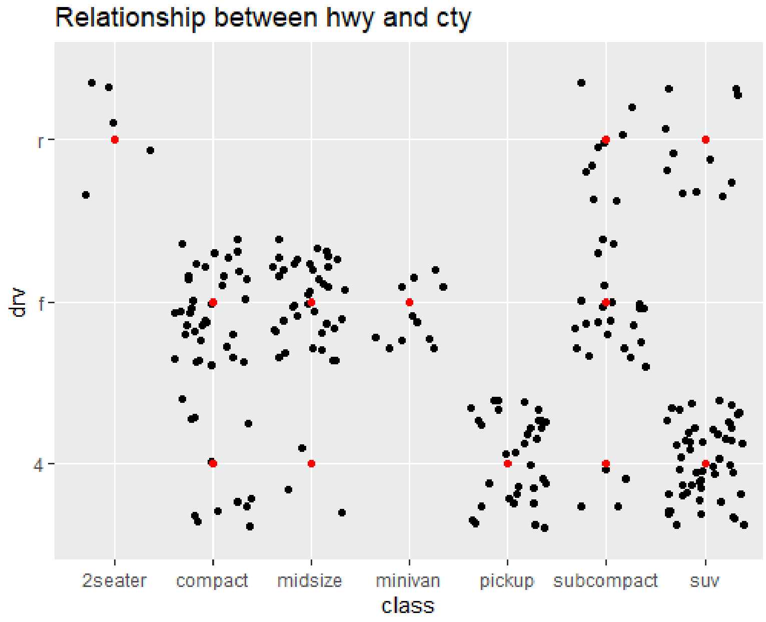
Adding extra visual cues and more options



```
mpg %>%
  ggplot(aes(class, drv)) +
  geom_point() +
  labs(title = "Relationship between class and drv")
```



```
mpg %>%
  ggplot(aes(class, drv)) +
  geom_jitter() +
  labs(title = "Relationship between class and drv")
```



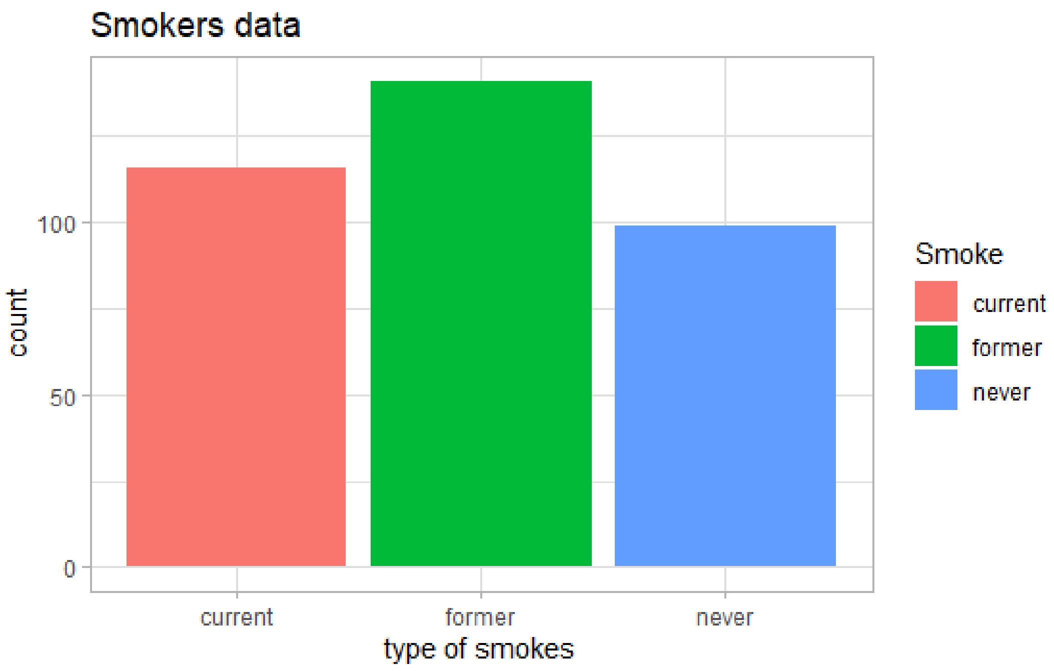
```
mpg %>%
  ggplot(aes(class, drv)) +
 geom_point(colour=“red”)+
  geom_jitter() +
  labs(title = "Relationship between class and drv")
```

BIVARIATE, CONTINUOUS & DISCRETE

- The analysis is done to inspect the relationship between two variables, where the explanatory variable is discrete and the response variable is continuous.
- ggplot2 provides some functions to plot this variation such as:
  - *geom_bar()*, for a bar plot
  - *geom_boxplot(),* for visualising the quantiles of response variable
  - Etc.

BAR PLOTS

- Bar plot is usually used to visualize the counts or the frequencies of the categories of the explanatory variable
- In this case, the visualisation is represented by bars
- The response variable is the counts or the frequencies



©Dr. Ibrahim Radwan – University of Canberra

BOXPLOTS

- Boxplot usually allows us to compare the distribution of a continuous variable across categories of a discrete variable
- It shows the distributions by marking the following values:
  - Minimum
  - Maximum
  - Median
  - IQR: inter Quantile Range
  - 25% percentile
  - 27% percentile

BUILDING A BOXPLOT

1- a) First layer contains the data specification
b) Also specifies the two variables

2- a) Second layer specifies the type of plot; the geometric function

3- a) Third and following layers specify axis labels and further visual cues

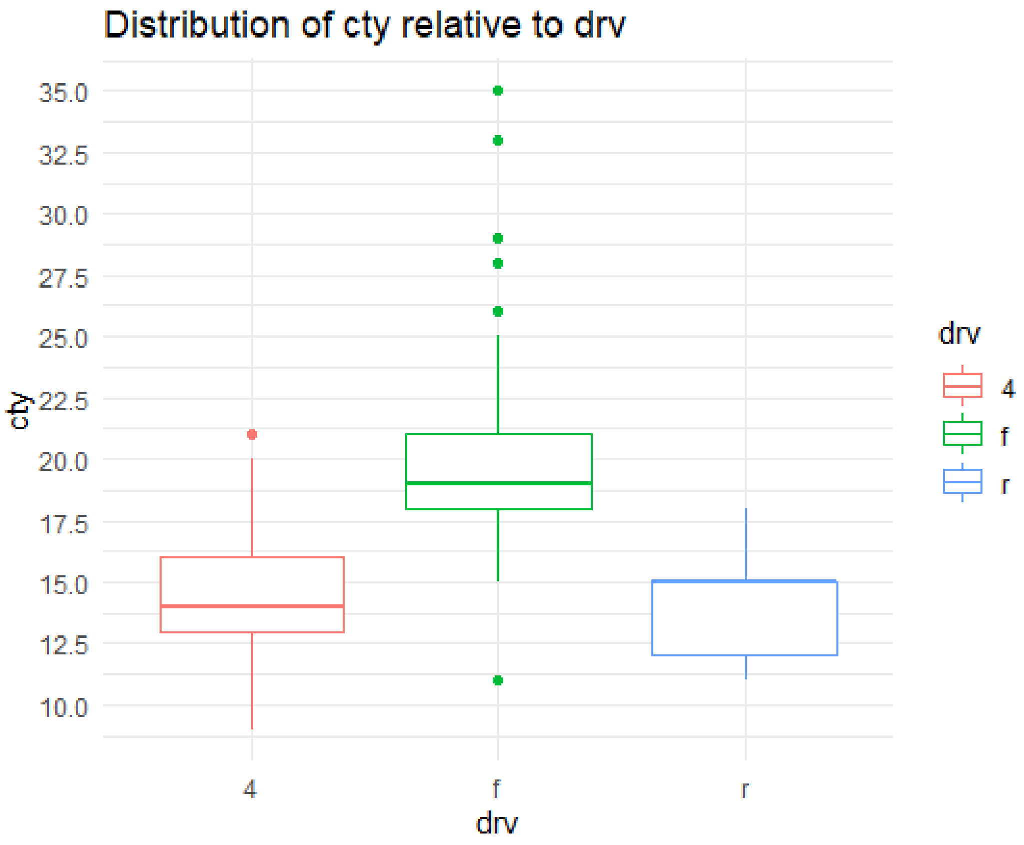
```
ggplot(data,aes(x= explanatory, y= response))+
  geom_boxplot(colour=classification) +
  xlab(“x axis label”) +
  ylab(“y axis label”)+
  ggtitle(“plot title”)
```

©Dr. Ibrahim Radwan – University of Canberra

EXAMPLE

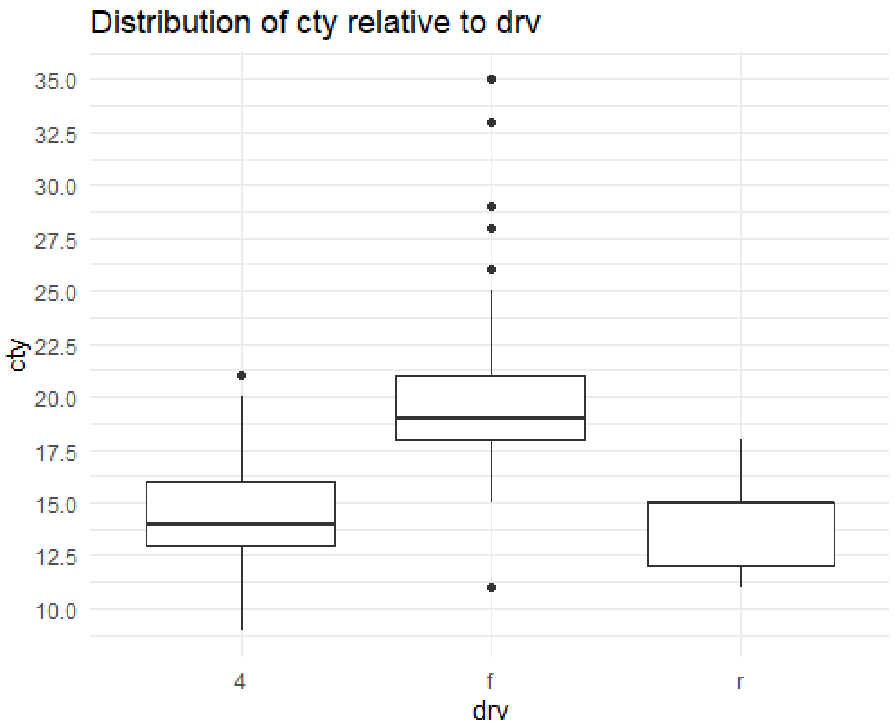
Basic boxplot visualization

```
ggplot(data= mpg, aes(x= drv, y= cty)) +
  geom_boxplot(colour=) +
  theme_minimal() +
  ggtitle("Relationship between drv and cty")
```

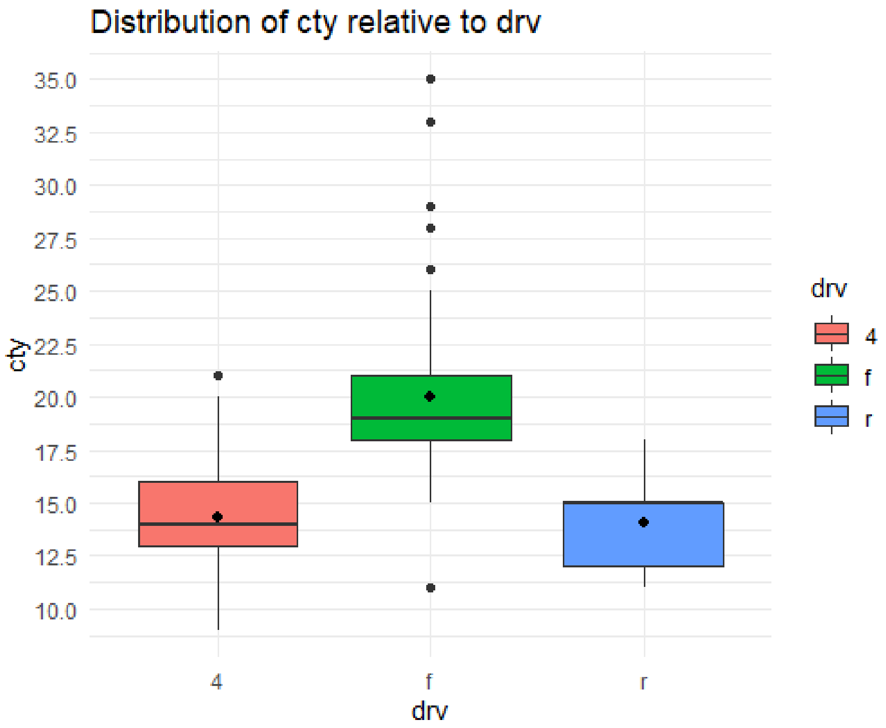


EXAMPLE (2)

Adding extra visual cues and more options



Basic plot



Combined with stat_summary

©Dr. Ibrahim Radwan – University of Canberra

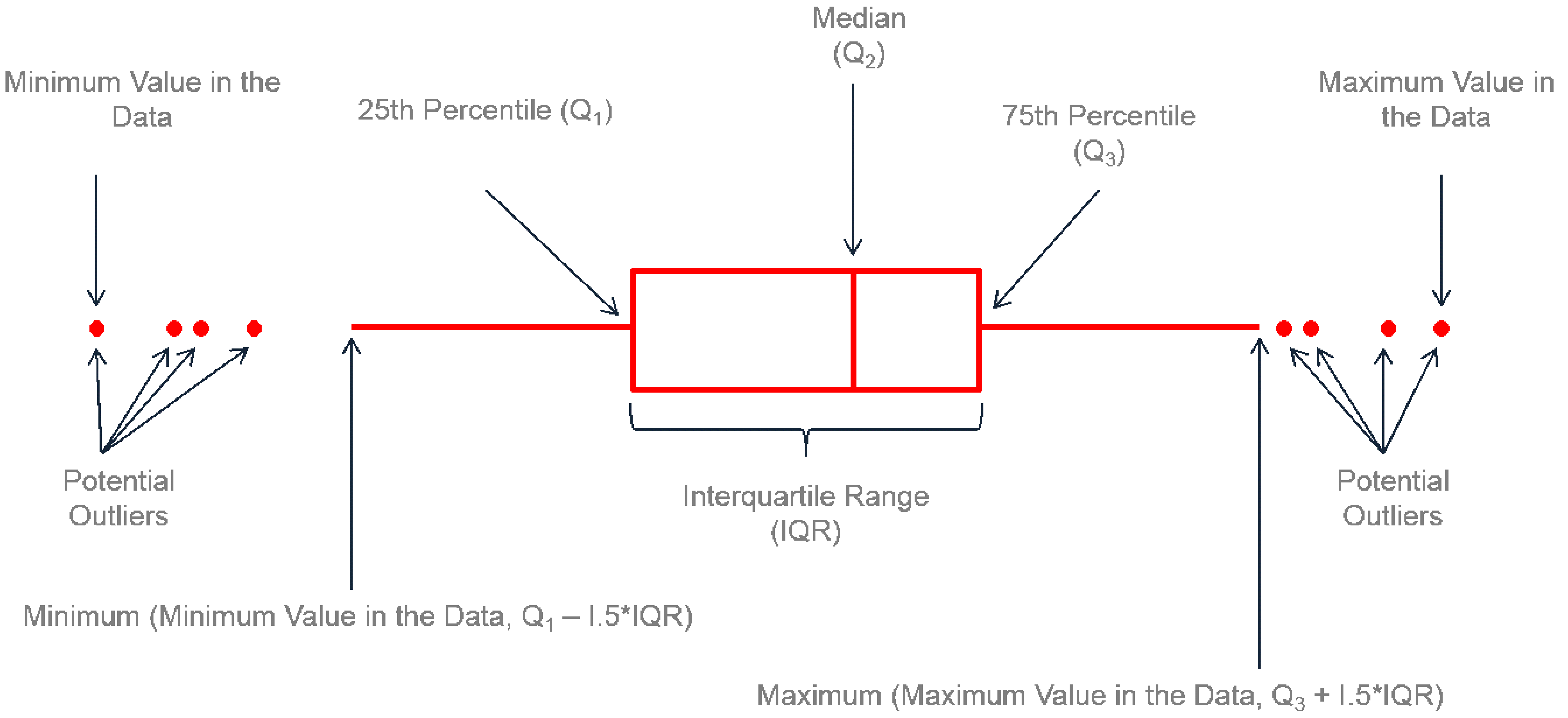
BOXPLOTS INTERPRETATION

UNIVERSITY OF CANBERRA

- Bivariate box plots combined with stat_summary() can help to understand the distribution of the variables.
- They can also show the distribution of several levels, allowing you to efficiently identify outliers, find patterns and spread of the observations
- Excellent for conveying the variation and particularly detecting and illustrating the variation changes between different levels of data
- Boxplots can help answer the following questions:
  - Is a category significant over the others?
  - Does the center differ between categories?
  - Does the variation differ between categories?
  - Are there any outliers?

©Dr. Ibrahim Radwan – University of Canberra

BOXPLOTS INTERPRETATION (2)


Median
(Q_2)
Minimum Value in the Data
25th Percentile (Q_1)
75th Percentile (Q_3)
Maximum Value in the Data
Potential Outliers
Interquartile Range (IQR)
Potential Outliers
Minimum (Minimum Value in the Data, Q_1 – 1.5*IQR)
Maximum (Maximum Value in the Data, Q_3 + 1.5*IQR)

Credit: https://www.leansigmacorporation.com/box-plot-with-minitab

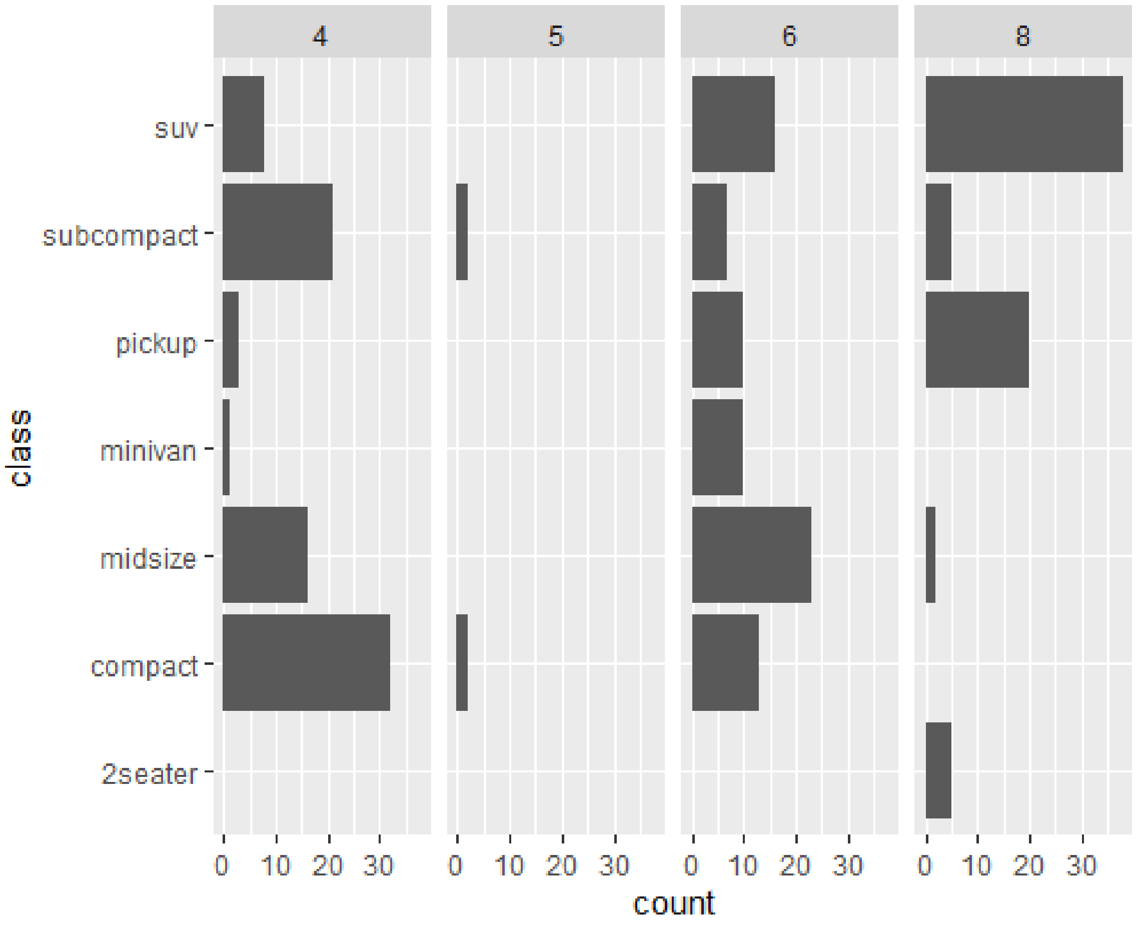
©Dr. Ibrahim Radwan – University of Canberra

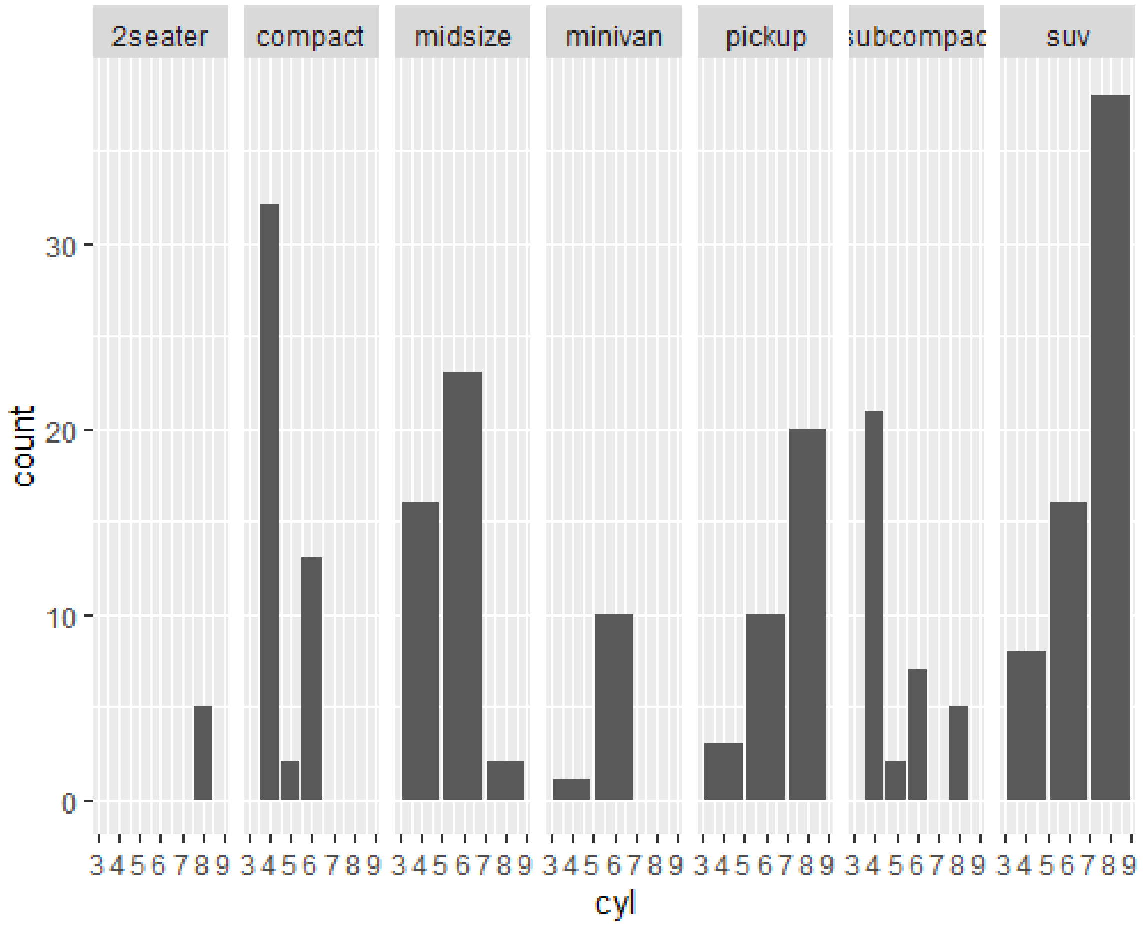
BIVARIATE, DISCRETE

- The analysis is done to inspect the relationship between two variables, where both of them are discrete.
- ggplot2 provides some functions to plot this variation such as:
  - *facet + geom_bar()*, for a bar chart for different categories as subplots
  - *geom_bin2d()*

EXAMPLE OF DISCRETE BIVARIATE

Facet + barplot





©Dr. Ibrahim Radwan – University of Canberra

RECOMMENDED READING

- You are recommended to read chapters 7 from the *“R for Data Science”* book:
  - https://r4ds.had.co.nz/exploratory-data-analysis.html
  - https://beta.rstudioconnect.com/content/3350/dplyr_tutorial.html