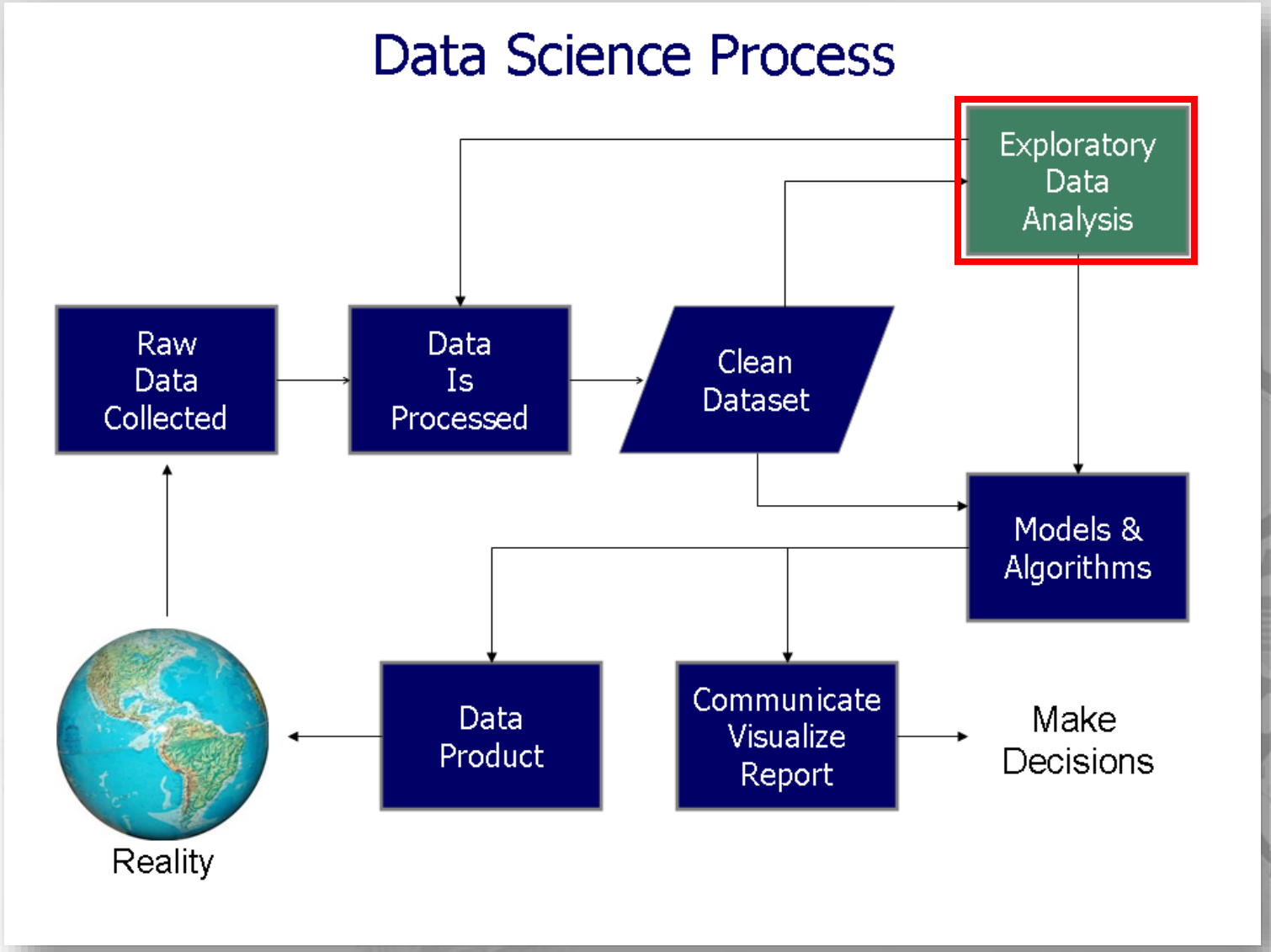
*If you can't measure it, you can't manage it.* –彼得・杜拉克

# 資料科學產業應用
# Data Science & Business Application
# Exploratory Data Analysis

郭俊良 博士

逢甲大學–勞動部雲端運算與數位轉型培訓班

# Recap


Data Science Process
Exploratory Data Analysis
Raw Data Collected
Data Is Processed
Clean Dataset
Models & Algorithms
Reality
Data Product
Communicate Visualize Report
Make Decisions

# Outlines

- **Introduction**
  - Anscombe's quartet
  - Understanding Data Structure
- **Steps of Exploratory Data Analysis**
  - Variable Identification
  - Univariate Analysis
  - Bivariate Analysis
  - Missing Values Treatment
  - Outlier Treatment
  - Feature Engineering (if necessary)
    - **Data Normalization**
    - **Feature / Variable Creation**
- Case Studies

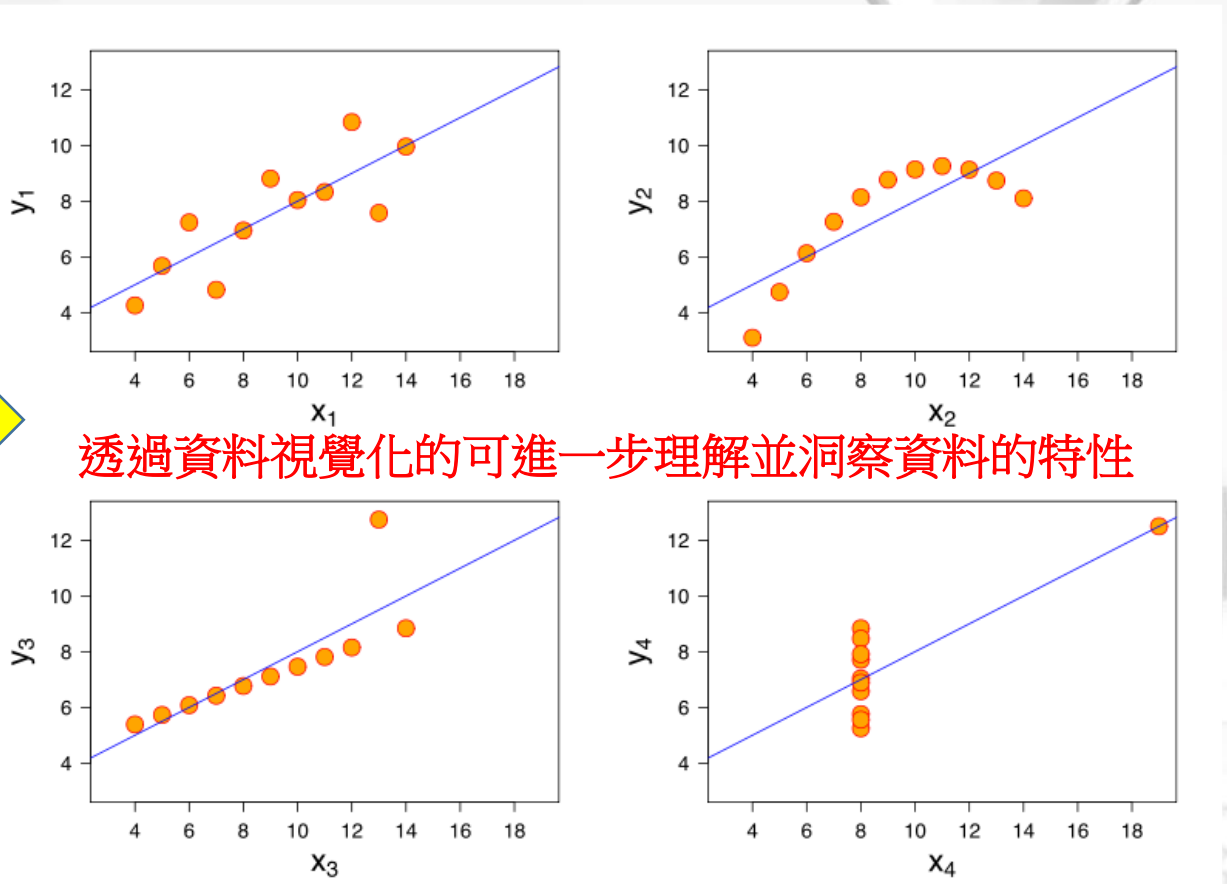
# Introduction

- Definition: Describing the data by means of statistical visualization techniques in order to identify important patterns
  - Preliminary step of data analysis ➔ Understanding data
  - Statistically analyzing data sets to summarize main characteristics
  - **Uncover relationships between variables** and **extract important features**
- Methods
  - Explore main characteristics ➔ Descriptive Statistics
  - Complete View of data ➔ Data Visualization
- Objectives
  - Examine the characteristics of univariate and correlation of bivariate
  - Discover Patterns ➔ **In order to apply proper Modeling method later**
  - Spot Anomalies (missing data & outliers)
  - Frame Hypothesis (Identify the research issues)
  - Check Assumptions (initial insights)

# Anscombe's quartet

- Same mean, variance, correlation, and linear regression line

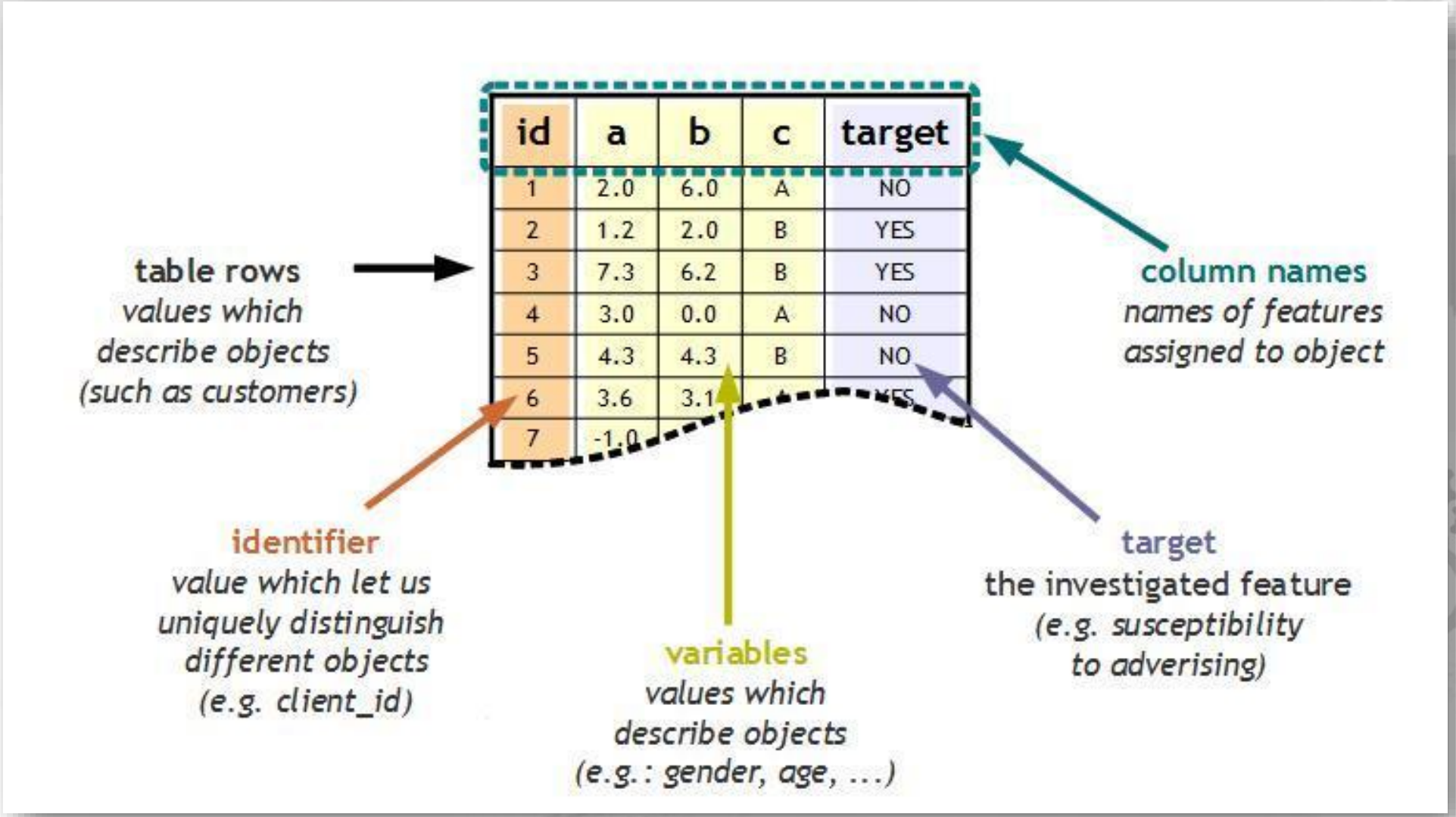
| | I | | II | | III | | IV | |
|---|---|---|---|---|---|---|---|---|
| | *x* | *y* | *x* | *y* | *x* | *y* | *x* | *y* |
| 1 | 10 | 8.04 | 10 | 9.14 | 10 | 7.46 | 8 | 6.58 |
| 2 | 8 | 6.95 | 8 | 8.14 | 8 | 6.77 | 8 | 5.76 |
| 3 | 13 | 7.58 | 13 | 8.74 | 13 | 12.74 | 8 | 7.71 |
| 4 | 9 | 8.81 | 9 | 8.77 | 9 | 7.11 | 8 | 8.84 |
| 5 | 11 | 8.33 | 11 | 9.26 | 11 | 7.81 | 8 | 8.47 |
| 6 | 14 | 9.96 | 14 | 8.1 | 14 | 8.84 | 8 | 7.04 |
| 7 | 6 | 7.24 | 6 | 6.13 | 6 | 6.08 | 8 | 5.25 |
| 8 | 4 | 4.26 | 4 | 3.1 | 4 | 5.39 | 19 | 12.5 |
| 9 | 12 | 10.84 | 12 | 9.13 | 12 | 8.15 | 8 | 5.56 |
| 10 | 7 | 4.82 | 7 | 7.26 | 7 | 6.42 | 8 | 7.91 |
| 11 | 5 | 5.68 | 5 | 4.74 | 5 | 5.73 | 8 | 6.89 |




透過資料視覺化的可進一步理解並洞察資料的特性

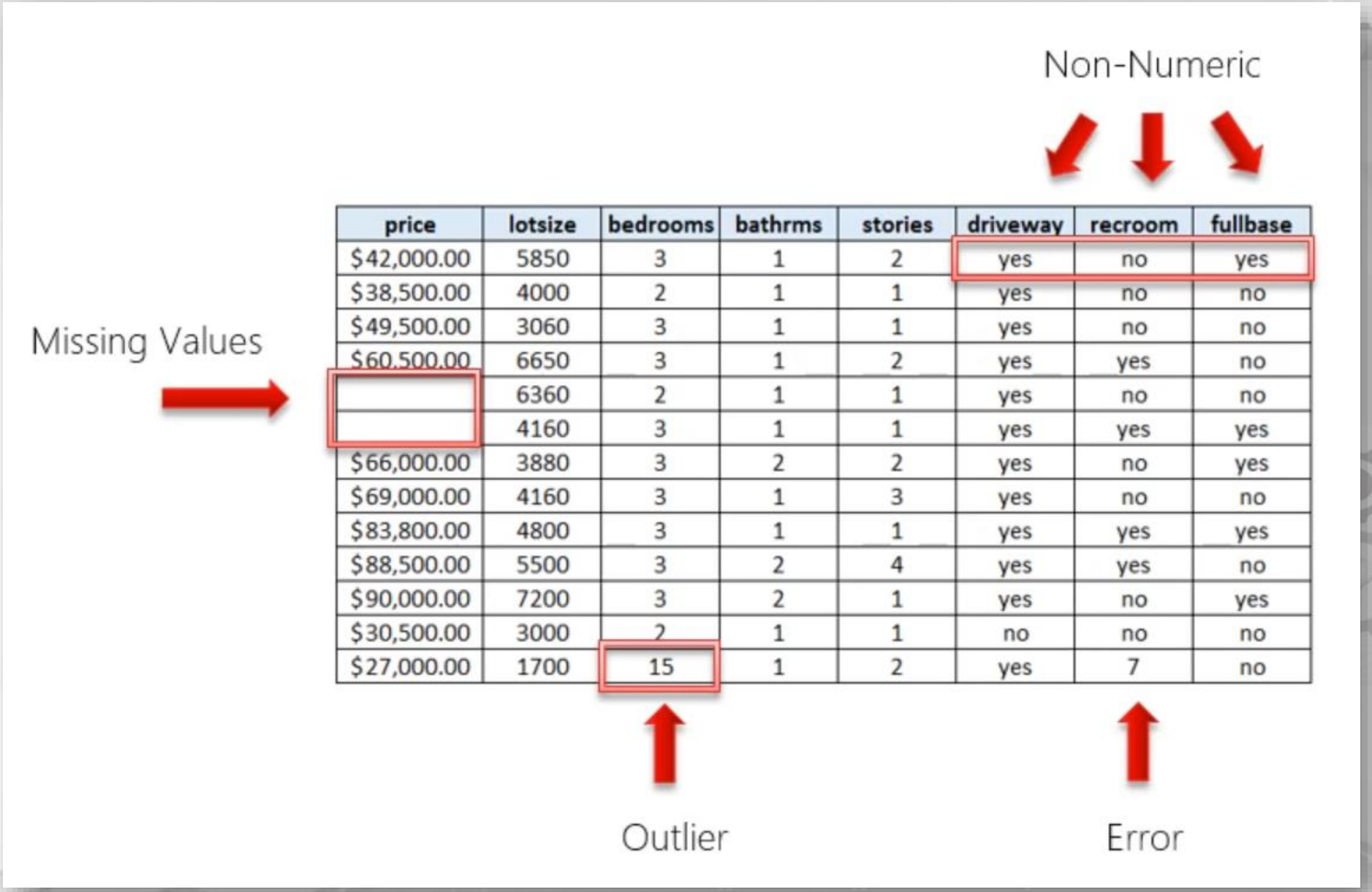
Source: Wikipedia, https://en.wikipedia.org/wiki/Anscombe's_quartet

# Understanding Data Structure

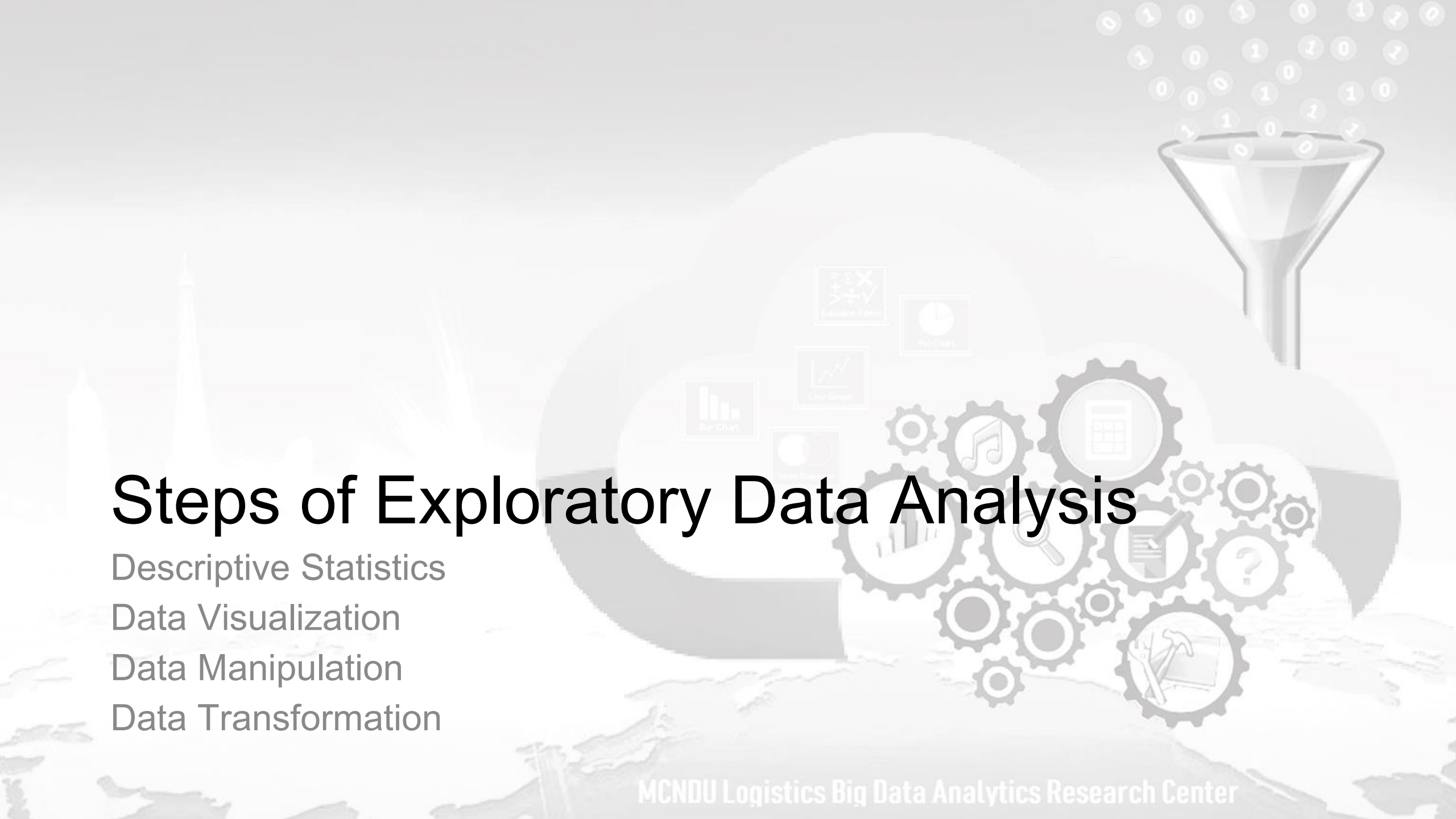



Source : https://www.datasciencecentral.com/profiles/blogs/predictive-analytics-for-beginners-part-1

# Understanding Data Structure




| price | lotsize | bedrooms | bathrms | stories | driveway | recroom | fullbase |
|---|---|---|---|---|---|---|---|
| $42,000.00 | 5850 | 3 | 1 | 2 | yes | no | yes |
| $38,500.00 | 4000 | 2 | 1 | 1 | yes | no | no |
| $49,500.00 | 3060 | 3 | 1 | 1 | yes | no | no |
| $60,500.00 | 6650 | 3 | 1 | 2 | yes | yes | no |
| | 6360 | 2 | 1 | 1 | yes | no | no |
| | 4160 | 3 | 1 | 1 | yes | yes | yes |
| $66,000.00 | 3880 | 3 | 2 | 2 | yes | no | yes |
| $69,000.00 | 4160 | 3 | 1 | 3 | yes | no | no |
| $83,800.00 | 4800 | 3 | 1 | 1 | yes | yes | yes |
| $88,500.00 | 5500 | 3 | 2 | 4 | yes | yes | no |
| $90,000.00 | 7200 | 3 | 2 | 1 | yes | no | yes |
| $30,500.00 | 3000 | 2 | 1 | 1 | no | no | no |
| $27,000.00 | 1700 | 15 | 1 | 2 | yes | 7 | no |



# Steps of Exploratory Data Analysis

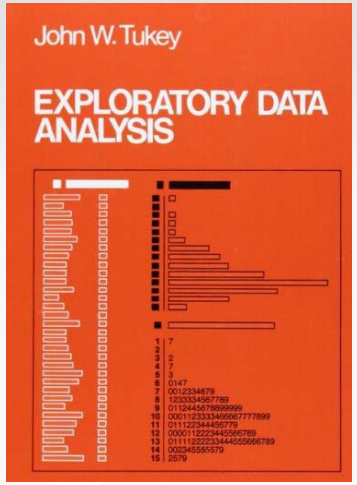
Descriptive Statistics

Data Visualization

Data Manipulation

Data Transformation

# Exploratory Data Analysis (EDA)




John Tukey (1915~2000) (統計學界的畢卡索)

「對正確的問題有個近似的答案，
勝過對錯的問題有精確的答案。」

- **Summarize** for large, complicated data sets
- **Reveal** structure, patterns, features, trends, outliers, anomalies, and relationships in data
- **Extract** important variables
- **Examine** assumptions in statistical models
- **Interaction** between the researcher and the data
- **Identify** the areas of interest
- **Visualization = Graphing for Data + Fitting + Graphing for Model**

# Steps of EDA

- Variable Identification
- Univariate Analysis
- Bivariate Analysis
- Missing Values Treatment
- Outlier Treatment
- Feature Engineering
  - **Data Normalization**
  - **Feature / Variable Creation**

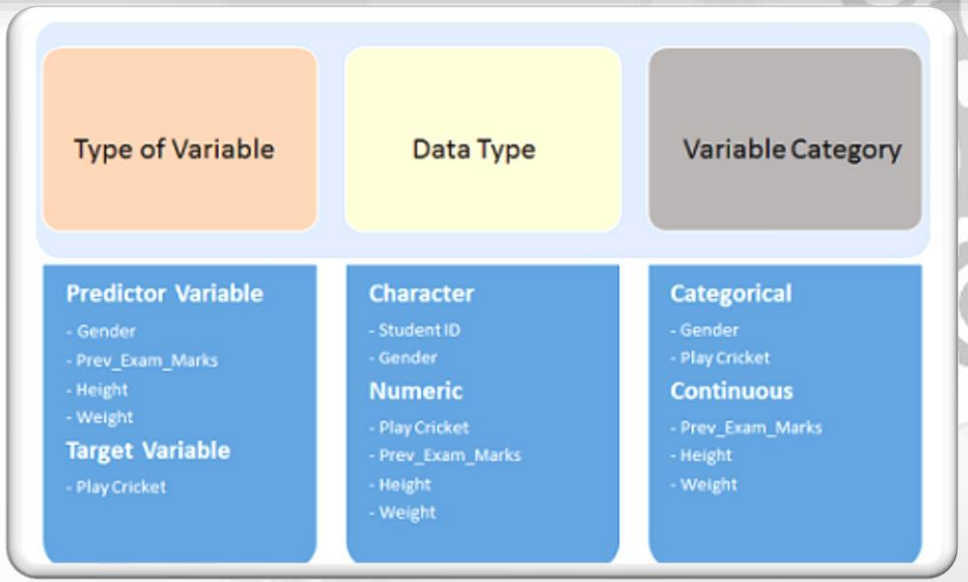
# Statistics for EDA

- **Descriptive Statistics**
  - 單變量統計分析 (Univariate Analysis of EDA)
    - 資料集中趨勢量數
      - 平均數、中位數、峰態係數 (Kurtosis)、偏態係數 (Skewness)
    - 資料分佈及變異
      - 機率密度函數、常態性檢定 (Normality Test)、全距(上/下界)、四分位數、標準差、變異數、ANOVA
  - 二元變量統計分析 (Bivariate Analysis of EDA)
    - 資料相關性分析 (Correlation Analysis)
      - Pearson 相關係數
  - 多元資料分佈及變異
    - MANOVA
- **PS. 資料需先進行遺失值 (Missing Value)、離群值 (Outlier) 處理**

# Variable Identification

- First, identify **Predictor** (Input) and **Target** (output) variables
- Next, identify the data type and category of the variables
- Example:

| Student_ID | Gender | Prev_Exam_Marks | Height (cm) | Weight Caregory (kgs) | Play Cricket |
|---|---|---|---|---|---|
| S001 | M | 65 | 178 | 61 | 1 |
| S002 | F | 75 | 174 | 56 | 0 |
| S003 | M | 45 | 163 | 62 | 1 |
| S004 | M | 57 | 175 | 70 | 0 |
| S005 | F | 59 | 162 | 67 | 0 |

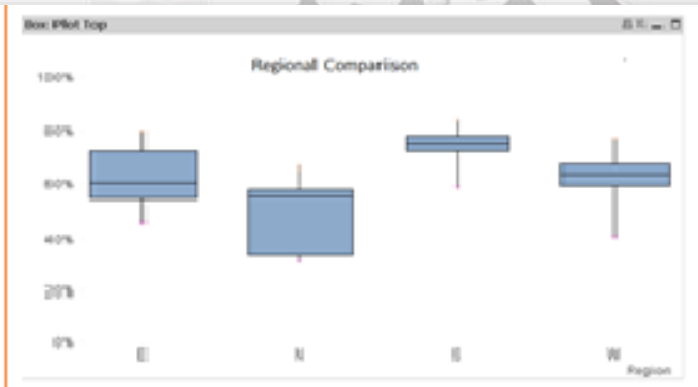


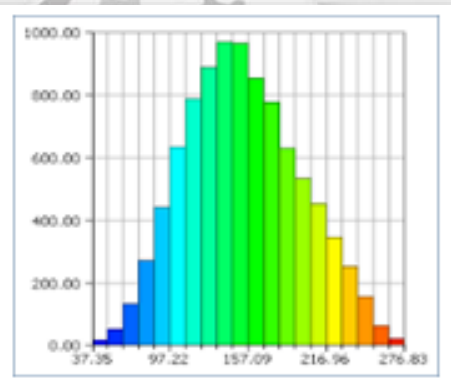

Source: [Analytics Vidhya - A Comprehensive Guide to Data Exploration]

# Univariate Analysis

- Explore variables one by one and to highlight the missing value and outliers
- Variables could be either *categorical* or *numerical*
  - Categorical Variables
    - Data is discrete and can be counted with predefined groups or levels
  - Numerical Variables
    - Numerical variables is continuous and can be transformed into categorical counterparts by a process called binning or discretization
    - It is also possible to transform a categorical variable into its numerical counterpart by a process called encoding

| Central Tendency | Measure of Dispersion | Visualization Methods |
|---|---|---|
| Mean | Range | Histogram |
| Median | Quartile | Box Plot |
| Mode | IQR | |
| Min | Variance | |
| Max | Standard Deviation | |
| | Skewness and Kurtosis | |



**Source:** [Analytics Vidhya - A Comprehensive Guide to Data Exploration]

- Again, handling of missing values & outliers is an important issue before mining data

# Bivariate Analysis

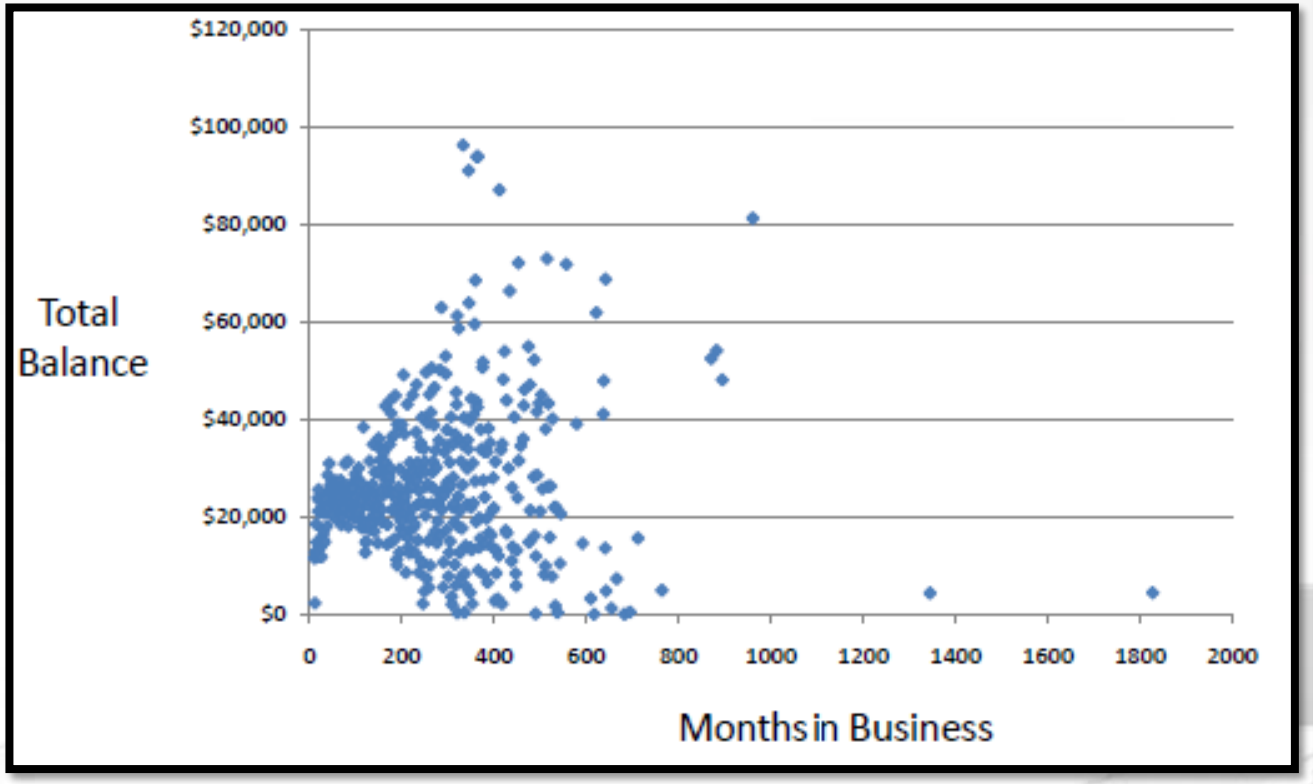
- Find out the relationship between two variables (attributes/features)
- Explores the relationship between two variables, whether there exists an association and the strength of this association, or whether there are differences between two variables and the significance of these differences
- Three Scenarios for Bivariate Analysis
  - Numerical & Numerical
  - Categorical & Categorical
  - Numerical & Categorical

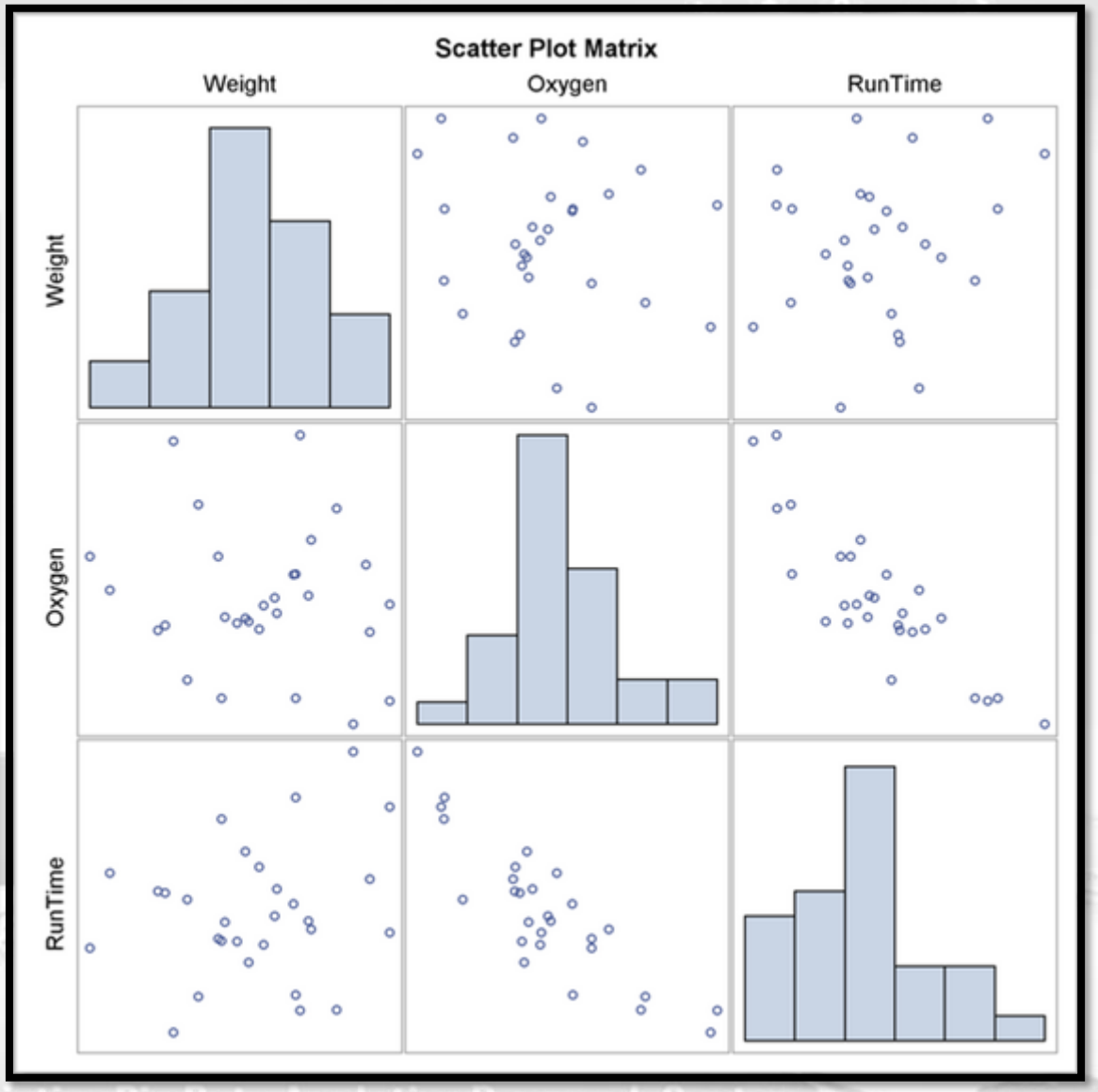
# Numerical & Numerical (1/7)

- **Scatter Plot**
  - Visual representation of the relationship between two numerical variables (attributes)
  - Drawn before working out a linear correlation or fitting a regression line
  - The resulting pattern indicates the type (linear or non-linear) and strength of the relationship between two variables
- **Linear Correlation:** ***Pearson correlation***
  - Quantifies the strength of a linear relationship between two numerical variables
  - When there is no correlation between two variables, there is no tendency for the values of one quantity to increase or decrease with the values of the second quantity

# Numerical & Numerical (2/7)

- **Scatter Plot**



# Numerical & Numerical (3/7)

- Linear Correlation Analysis
  - Measures to what extent different variables are interdependent
  - Correlation Matrix
  - Data Visualization - Correlation Heat-maps
  - **Pearson Correlation**: $r_{xy}$

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

- For example:
  - Lung cancer ➔ Smoking
  - Rain ➔ Umbrella
- Correlation doesn't imply causation
- Other Methods
  - Kendall rank correlation
  - Spearman Correlation

# Numerical & Numerical (4/7)

**Pearson Correlation**: Measure the strength of the correlation between two variables (features)

## Correlation coefficient ($r_{xy}$)

- Close to +1: Large Positive relationship
- Close to -1: Large Negative relationship
- Close to 0: No relationship

## P-value

- P-value < 0.001 Strong certainty in the result
- P-value < 0.05 Moderate certainty in the result
- P-value < 0.1 Weak certainty in the result
- P-value > 0.1 No certainty in the result

## Strong Correlation:

- Correlation coefficient close to 1 or -1
- P-value is less than 0.001

$$r = \frac{Covar(x,y)}{\sqrt{Var(x)Var(y)}}$$

$$Covar(x,y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{n}$$

$$Var(x) = \frac{\sum(x-\bar{x})^2}{n}$$

$$Var(y) = \frac{\sum(y-\bar{y})^2}{n}$$

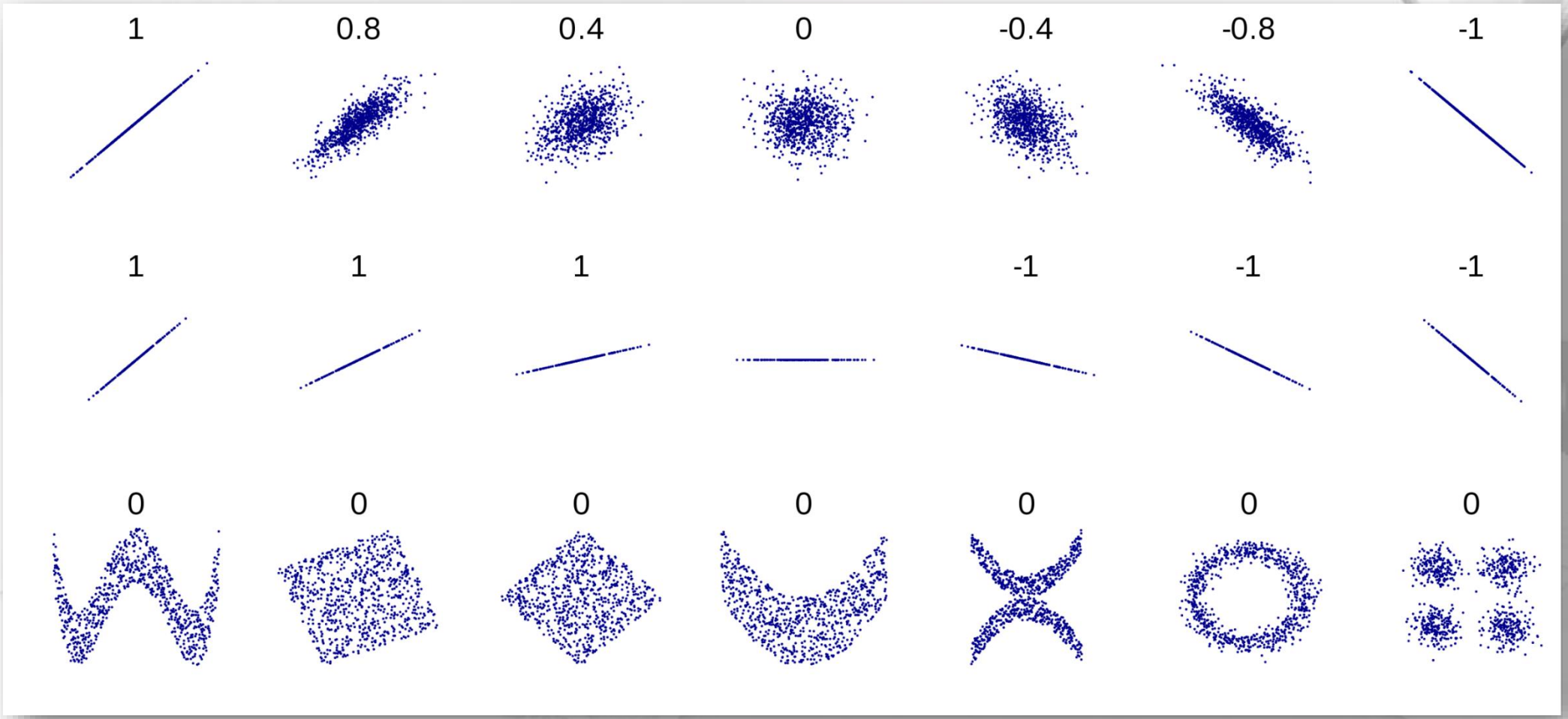
$r$ : Linear Correlation

$Covar$ : Covariance

$Var$ : Variance

# Numerical & Numerical (5/7)

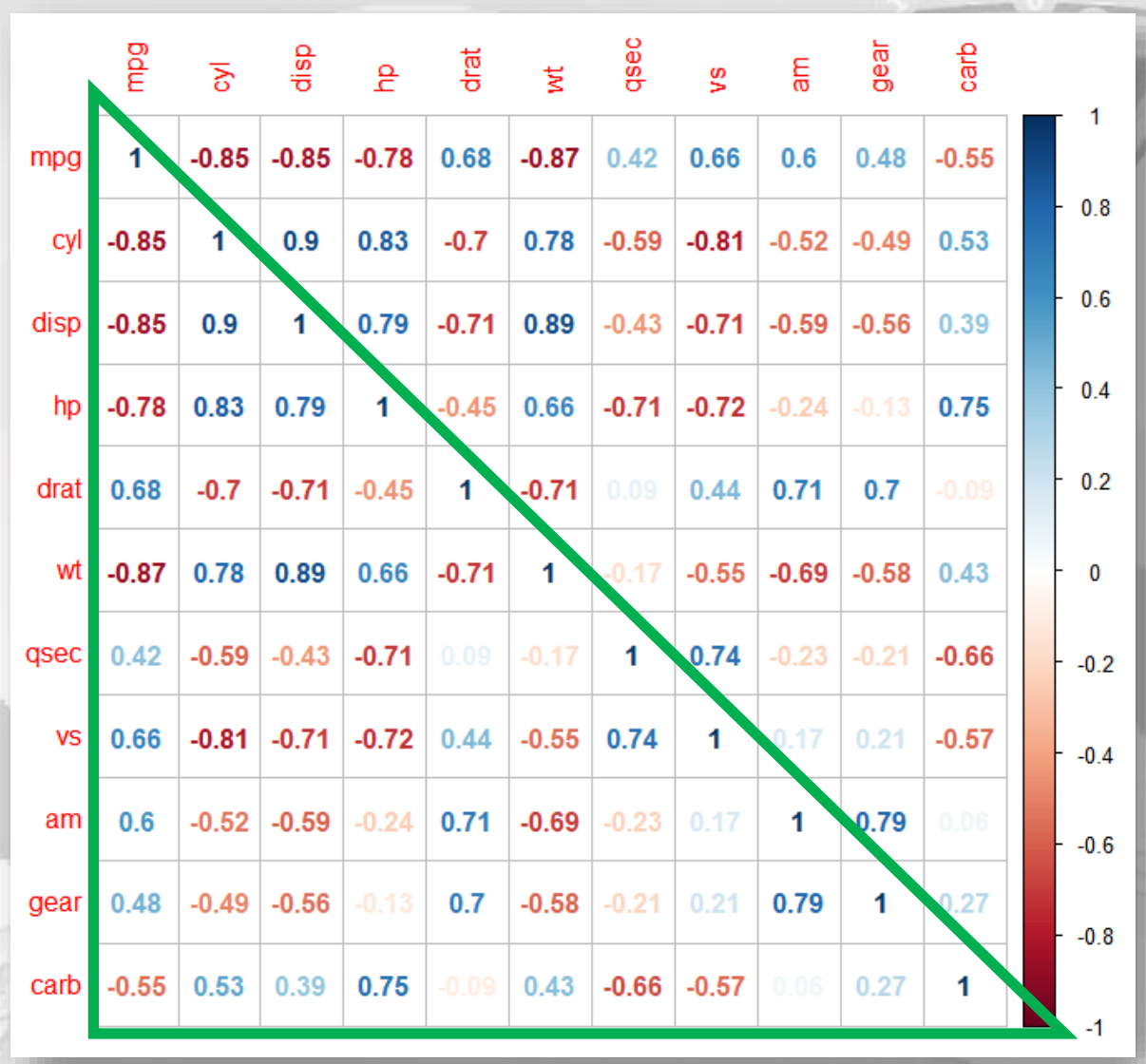
Visualization of Correlation




Source: https://en.wikipedia.org/wiki/Correlation_and_dependence

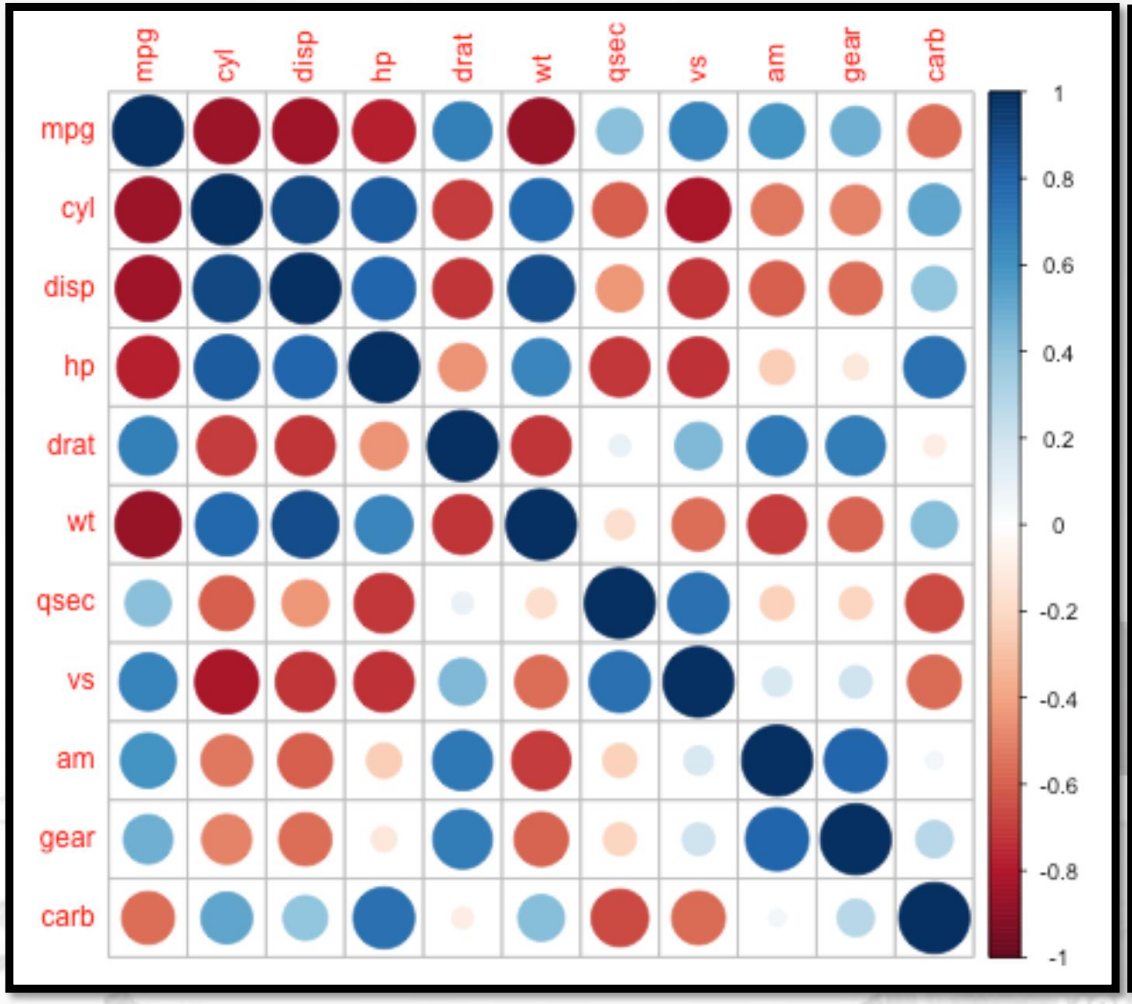
# Numerical & Numerical (6/7)

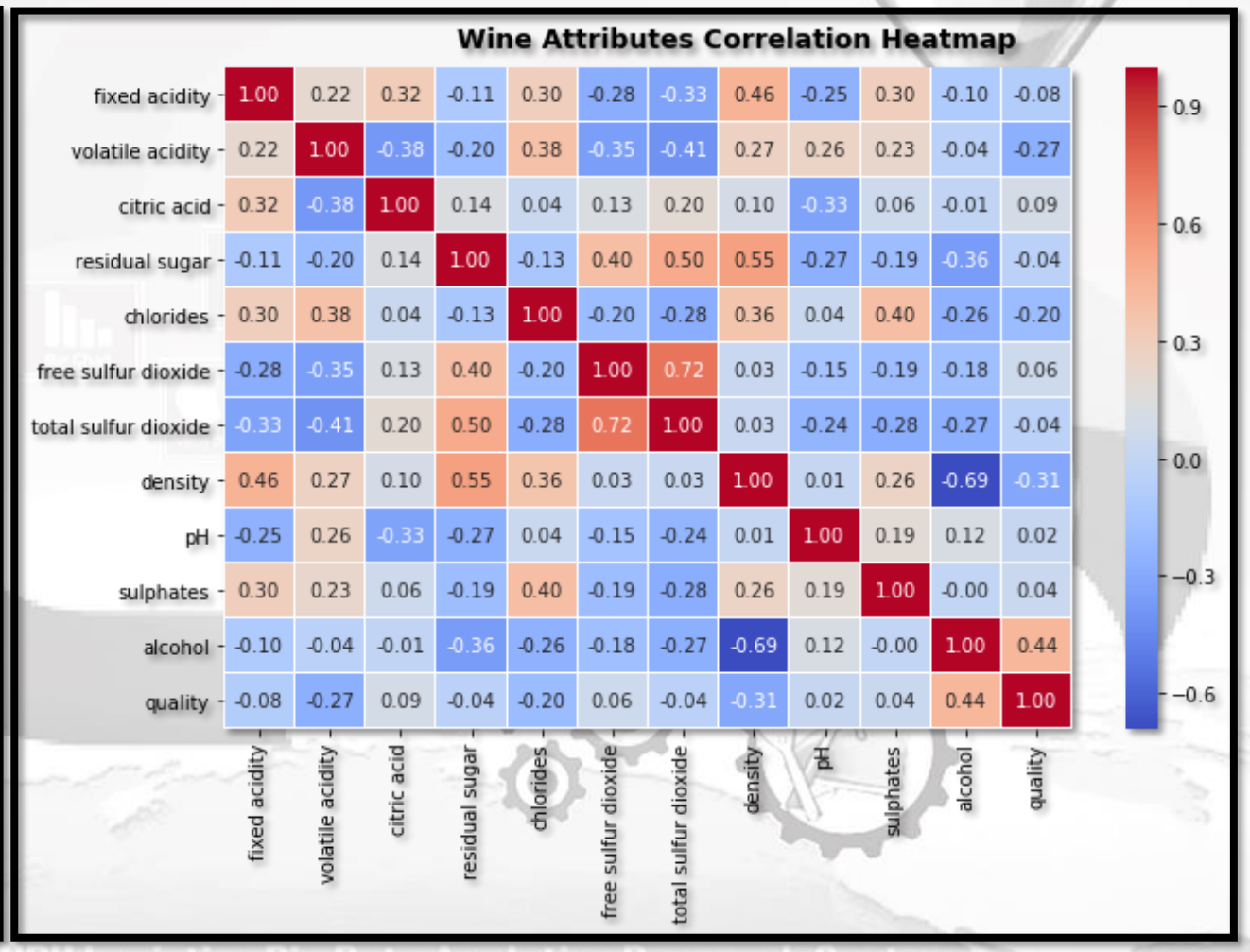
- Correlation Matrix
- Correlation matrix is a table showing **correlation coefficients** between variables (numerical)
- Each cell in the table shows the correlation between two variables
- A correlation matrix is used as a way to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses



# Numerical & Numerical (7/7)

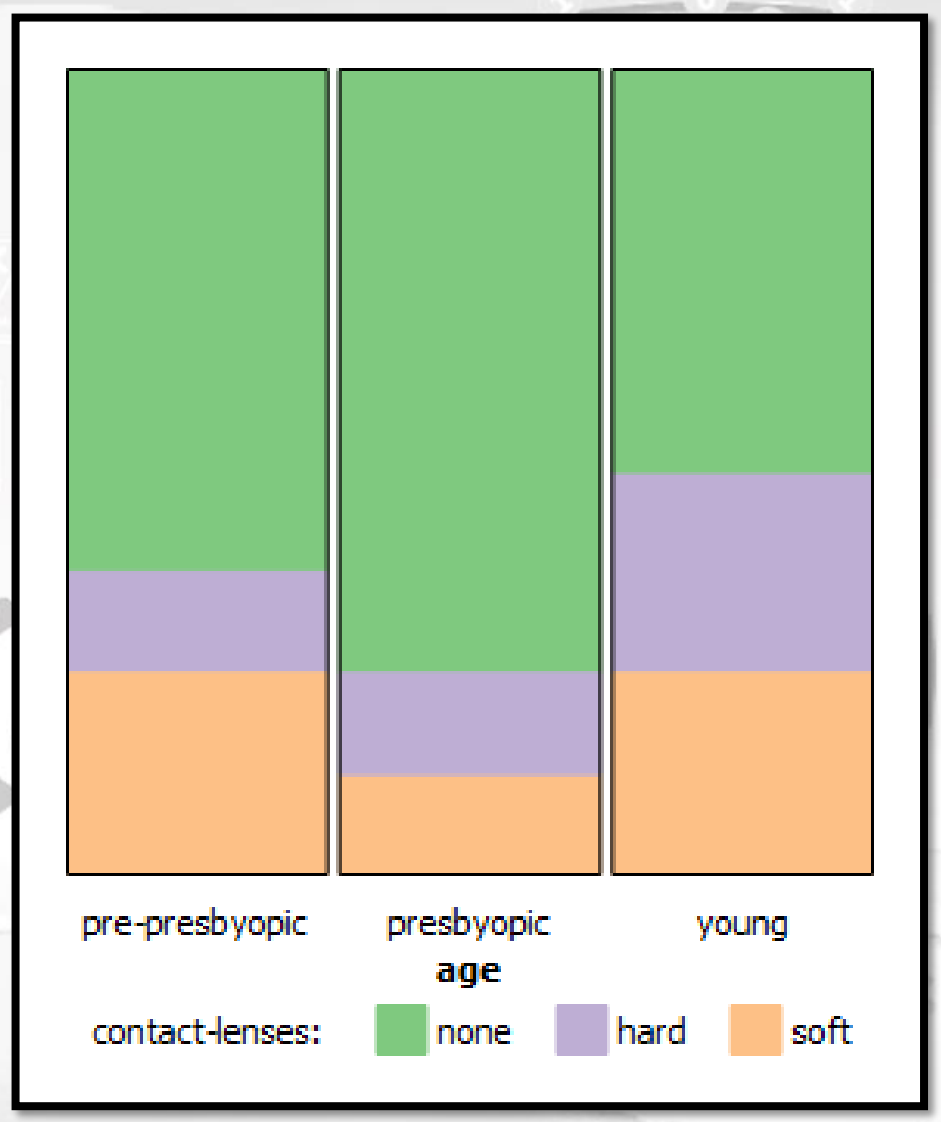
## Visualization of Correlation Matrix



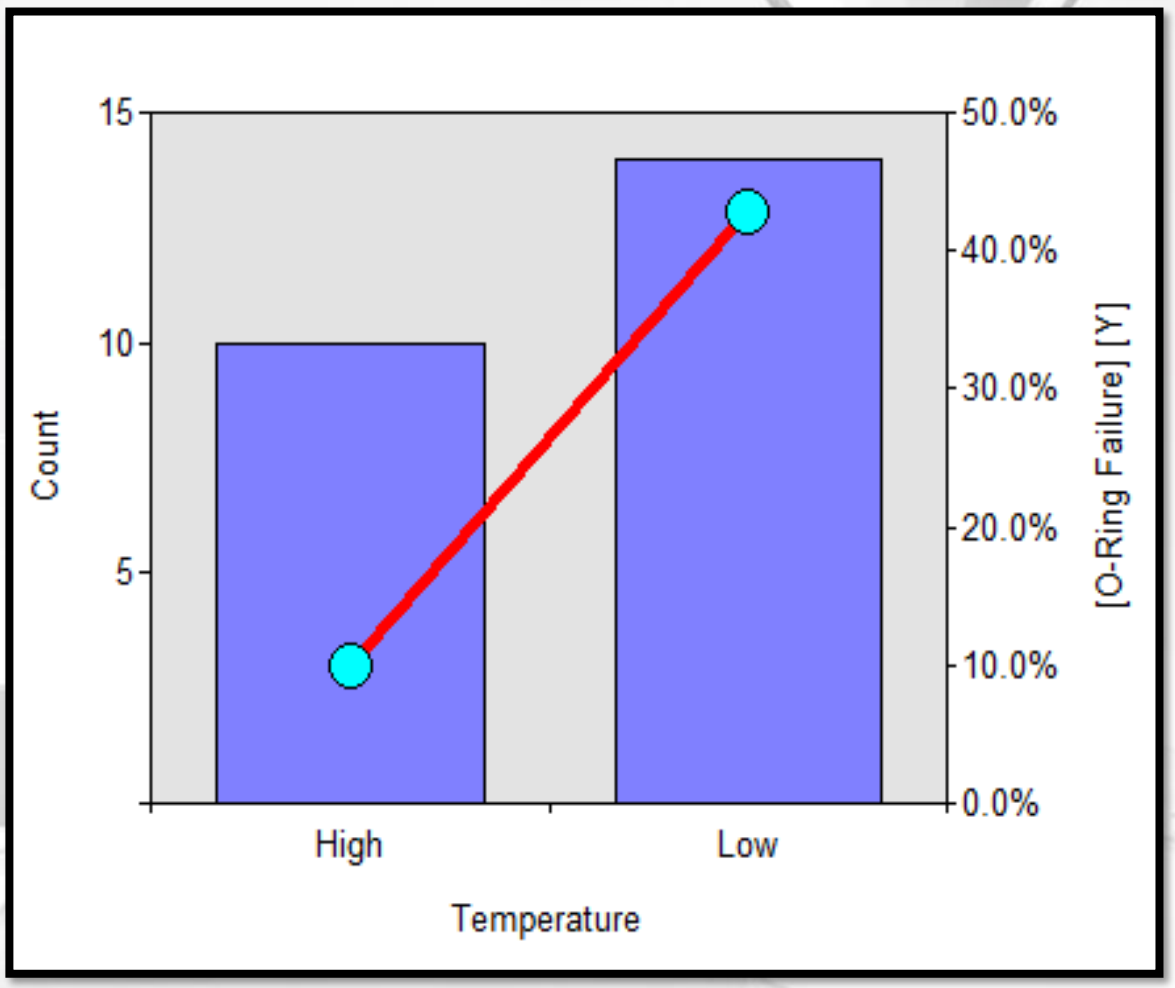
# Categorical & Categorical (1/4)

- **Two-way Table**
- **Stacked Column Chart**
  - Visualize the relationship between two categorical variables
  - Compares the percentage that each category from one variable contributes to a total across categories of the second variable
- **Chi-Square Test**



# Categorical & Categorical (2/4)

- **Two-way Table**
- **Stacked Column Chart**
- **Combination Chart**
  - Uses two or more chart types to emphasize that the chart contains different kinds of information
  - Example: use a bar chart to show the distribution of one categorical variable and a line chart to show the percentage of the selected category from the second categorical variable
  - The best visualization method to demonstrate the predictability power of a predictor (X-axis) against a target (Y-axis)



# Categorical & Categorical (3/4)

- **Chi-square Test**
  - Used to determine the association between categorical variables
  - Based on the difference between the expected frequencies (*e*) and the observed frequencies (*n*) in one or more categories in the frequency table

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-e_{ij})^2}{e_{ij}}$$

$$e_{ij} = \frac{n_{i.}n_{.j}}{n}$$

$$df = (r-1)(c-1)$$

Tchouproff Contingency Coefficient

$$\rho_\tau = \sqrt{\frac{\chi^2}{n\sqrt{(c-1)(r-1)}}}$$

# Categorical & Categorical (4/4)

- Chi-square Test: Example
  - The following frequency table (contingency table) with a chi-square of 10.67, degree of freedom (*df*) of 2 and probability of 0.005 shows a significant dependency between two categorical variables (hair and eye colors).

**e**=(44 * 52)/95=24.1

| | | *Hair* | | |
|---|---|---|---|---|
| | | Light | Dark | |
| *Eye* | Black | **32** **(24.1)** | **12** **(19.9)** | 44 |
| | Green/Blue | **14** **(19.7)** | **22** **(16.3)** | 36 |
| | Others | **6** **(8.2)** | **9** **(6.8)** | 15 |
| | | 52 | 43 | 95 |

$$\chi^2 = 10.67$$

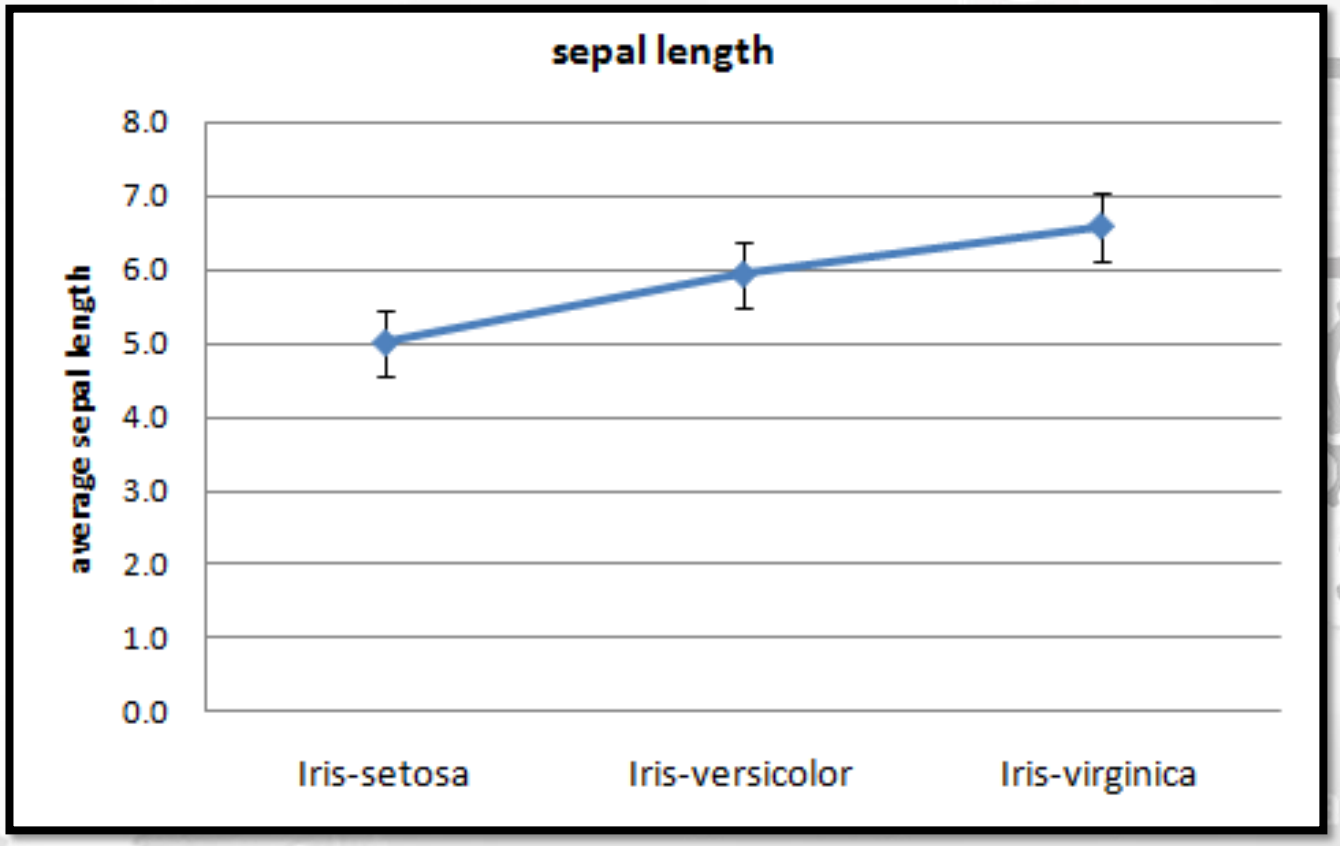
$$df = (r-1)(c-1) = (3-1)(2-1) = 2$$

$$p = 0.005$$

$$\rho_c = \sqrt{\frac{10.67}{95\sqrt{(3-1)(2-1)}}} = 0.28$$

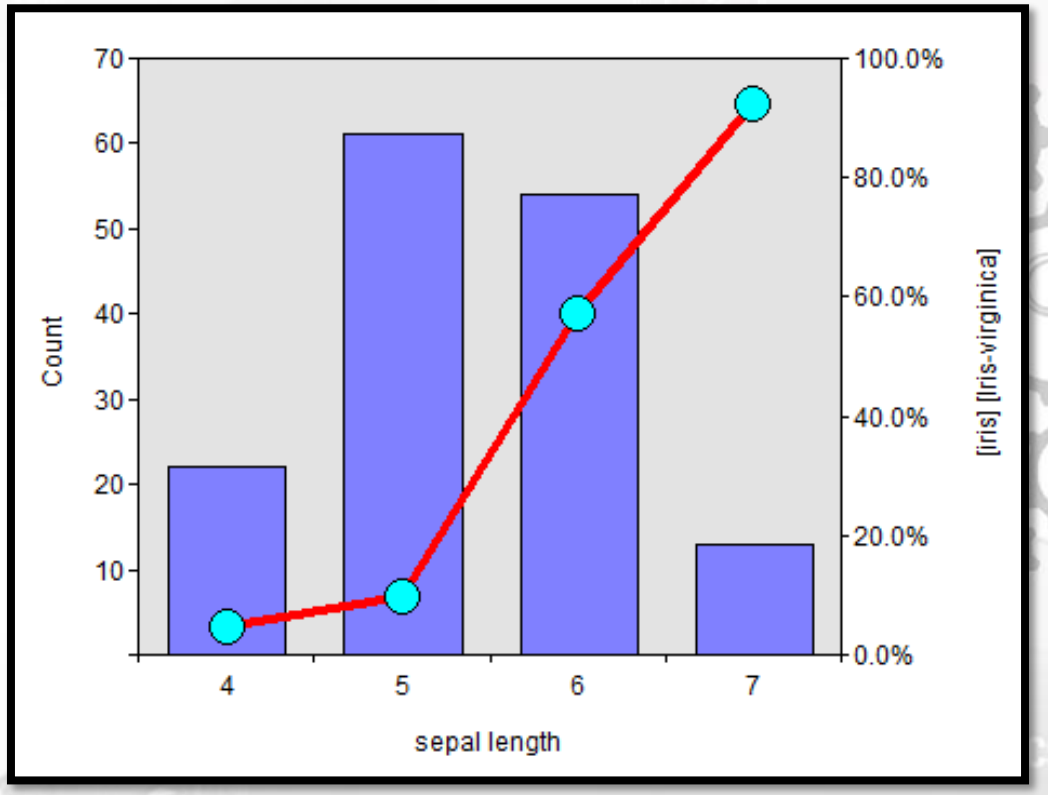
# Numerical & Categorical (1/6)

- **Line Chart with Error Bars (折線圖誤差線)**
  - Displays information as a series of data points connected by straight line segments
  - Each data point is average of the numerical data for the corresponding category of the categorical variable with error bar showing standard error
  - Summarize how pieces of information are related and how they vary depending on one another



# Numerical & Categorical (2/6)

- **Combination Chart**
  - Uses two or more chart types to emphasize that the chart contains different kinds of information
  - Here, we use a bar chart to show the distribution of a binned numerical variable and a line chart to show the percentage of the selected category from the categorical variable
  - The combination chart is the best visualization method to demonstrate the predictability power of a predictor (X-axis) against a target (Y-axis)



# Numerical & Categorical (3/6)

- **Z-test and t-test**
  - They assess whether the averages of two groups are statistically different from each other
  - For comparing the averages of a numerical variable for two categories of a categorical variable

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}}$$

where:

- $\overline{X}_1, \overline{X}_2$: *Averages*
- $S_1^2, S_2^2$: *Variances*
- $N_1, N_2$: *Counts*
- $Z$: *Standard Normal Distribution*

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{S^2\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

$$S^2 = \frac{(N_1 - 1)S_1^2 + (N_2 - 1)S_2^2}{N_1 + N_2 - 2}$$

where:

- $\overline{X}_1, \overline{X}_2$: *Averages*
- $S_1^2, S_2^2$: *Variances*
- $N_1, N_2$: *Counts*
- $t$: has t distribution with $N_1 + N_2 - 2$ degree of freedom

# Numerical & Categorical (4/6)

- **t-test: example**
  - Is there a significant difference between the means (averages) of the numerical variable (Temperature) in two different categories of the categorical variable (O-Ring Failure)?

| O-Ring Failure | Temperature |
|---|---|
| Y | 53 56 57 70 70 70 75 |
| N | 63 66 67 67 67 68 69 70 72 73 75 76 76 78 79 80 81 |

| *t-test* | O-Ring Failure | |
|---|---|---|
| **Temperature** | Y | N |
| Count | 7 | 17 |
| Mean | 64.43 | 72.18 |
| Variance | 76.95 | 30.78 |
| t | -2.62 | |
| *df* | 22 | |
| Probability | 0.0156 | |

  - Result:
  - The low probability (0.0156) means that the difference between the average temperature for failed O-Ring and the average temperature for intact O-Ring is **significant**

# Numerical & Categorical (5/6)

- **Analysis of Variance (ANOVA)**
  - Assesses whether the averages of more than two groups are statistically different from each other
  - For comparing the averages of a numerical variable for more than two categories of a categorical variable

| Source of Variation | Sum of Squares | Degree of Freedom | Mean Square | F | P |
|---|---|---|---|---|---|
| Between Groups | $SS_B$ | $df_B$ | $MS_B = SS_B/df_B$ | $F=MS_B/MS_W$ | P(F) |
| Within Groups | $SS_W$ | $df_w$ | $MS_W = SS_W/df_w$ | | |
| Total | $SS_T$ | $df_T$ | | | |

$$SS_B = \sum_{i=1}^{k} \frac{(\sum X)_i^2}{N_i} - \frac{\left(\sum_{i=1}^{k} (\sum X)_i\right)^2}{\sum_{i=1}^{k} N_i}$$

$$SS_W = \sum_{i=1}^{k} (\sum X^2)_i - \sum_{i=1}^{k} \frac{(\sum X)_i^2}{N_i}$$

$$SS_T = \sum_{i=1}^{k} (\sum X^2)_i - \frac{\left(\sum_{i=1}^{k} (\sum X)_i\right)^2}{\sum_{i=1}^{k} N_i}$$

$$df_W = \sum N_i - k$$

$$df_B = k - 1$$

$$df_T = \sum N_i - 1$$

$F$ : has F distribution with $df_B$ and $df_w$ degree of freedom

# Numerical & Categorical (6/6)

- **Analysis of Variance (ANOVA) – example**
  - Is there a significant difference between the averages of the numerical variable (Humidity) in the three categories of the categorical variable (Outlook)?

| Outlook | Humidity |
|---|---|
| overcast | 86 65 90 75 |
| rainy | 96 80 70 80 91 |
| sunny | 85 90 95 70 70 |

| Outlook | Count | Mean | Variance |
|---|---|---|---|
| overcast | 4 | 79.0 | 127.3 |
| rainy | 5 | 83.4 | 104.8 |
| sunny | 5 | 82.0 | 132.5 |

| Source of Variation | Sum of Squares | Degree of freedom | Mean Square | F Value | Probability |
|---|---|---|---|---|---|
| Between Groups | 44.0 | 2 | 22.0 | 0.182 | 0.836 |
| Within Groups | 1331.2 | 11 | 121.0 | | |
| Total | 1375.2 | 13 | | | |

- There is no significant difference between the averages of Humidity in the three categories of Outlook

# Missing Values Treatment

- Missing data in the training data set can reduce the power / fit of a model or can lead to a biased model
- It can lead to wrong prediction or classification

| Name | Weight | Gender | Play Cricket/ Not |
|---|---|---|---|
| Mr. Amit | 58 | M | Y |
| Mr. Anil | 61 | M | Y |
| Miss Swati | 58 | F | N |
| Miss Richa | 55 | | Y |
| Mr. Steve | 55 | M | N |
| Miss Reena | 64 | F | Y |
| Miss Rashmi | 57 | | Y |
| Mr. Kunal | 57 | M | N |

| Gender | #Students | #Play Cricket | %Play Cricket |
|---|---|---|---|
| F | 2 | 1 | 50% |
| M | 4 | 2 | 50% |
| Missing | 2 | 2 | 100% |

| Name | Weight | Gender | Play Cricket/ Not |
|---|---|---|---|
| Mr. Amit | 58 | M | Y |
| Mr. Anil | 61 | M | Y |
| Miss Swati | 58 | F | N |
| Miss Richa | 55 | F | Y |
| Mr. Steve | 55 | M | N |
| Miss Reena | 64 | F | Y |
| Miss Rashmi | 57 | F | Y |
| Mr. Kunal | 57 | M | N |

| Gender | #Students | #Play Cricket | %Play Cricket |
|---|---|---|---|
| F | 4 | 3 | 75% |
| M | 4 | 2 | 50% |

**Source:** [Analytics Vidhya – A Comprehensive Guide to Data Exploration]

# Methods to Treat Missing Values

- Deletion
- Mean/ Mode/ Median Imputation
- Prediction Model
- KNN Imputation

**List wise deletion**

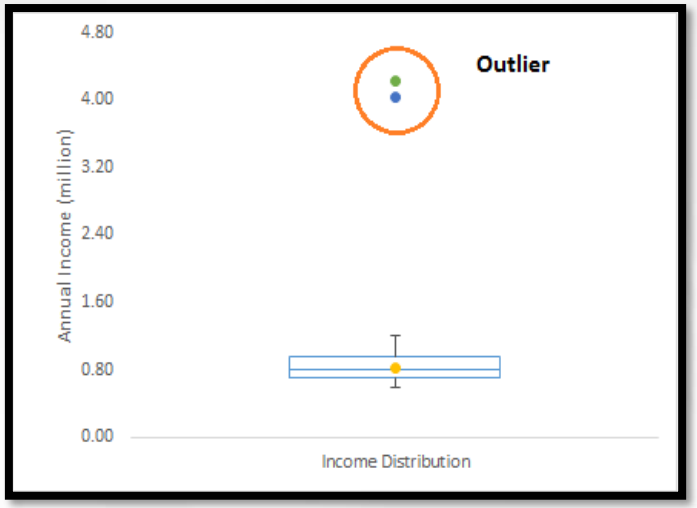
| Gender | Manpower | Sales |
|---|---|---|
| M | 25 | 343 |
| F | . | 280 |
| M | 33 | 332 |
| M | . | 272 |
| F | 25 | . |
| M | 29 | 326 |
| | 26 | 259 |
| M | 32 | 297 |

**Pair wise deletion**

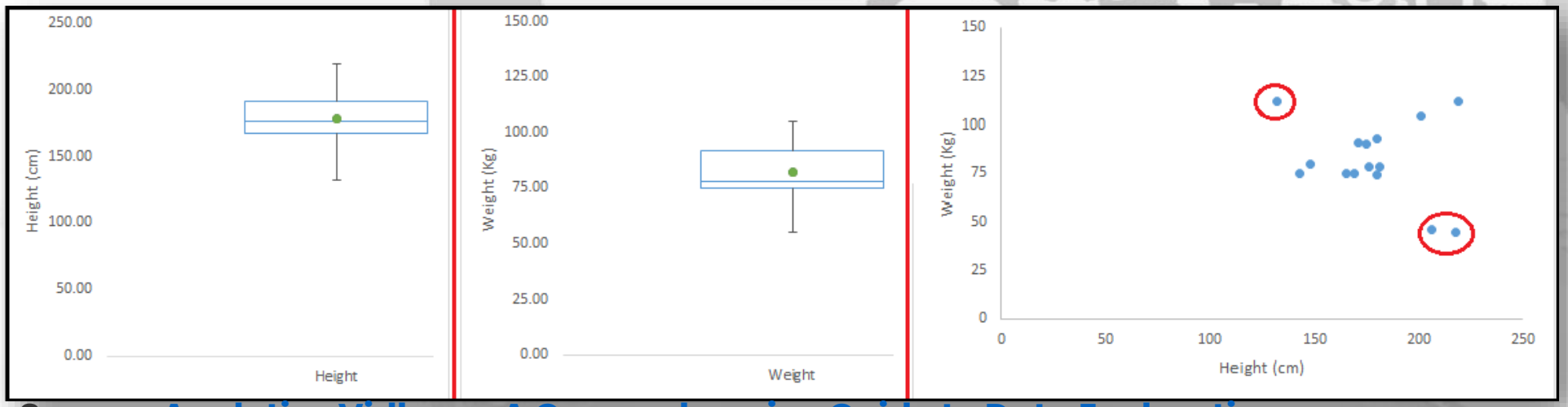
| Gender | Manpower | Sales |
|---|---|---|
| M | 25 | 343 |
| F | . | 280 |
| M | 33 | 332 |
| M | . | 272 |
| F | 25 | . |
| M | 29 | 326 |
| | 26 | 259 |
| M | 32 | 297 |

Source: [Analytics Vidhya – A Comprehensive Guide to Data Exploration]

# Outlier Detection and Treatment




- Outlier is an observation that appears far away and diverges from an overall pattern in a sample
- Outliers can drastically change the results of the data analysis and statistical modeling
- Outlier can be of two types: **Univariate** and **Multivariate**




**Source:** [Analytics Vidhya – A Comprehensive Guide to Data Exploration]

# Impact of Outliers

- It increases the error variance and reduces the power of statistical tests
- If the outliers are non-randomly distributed, they can decrease normality
- They can bias or influence estimates that may be of substantive interest
- They can also impact the basic assumption of Regression, ANOVA and other statistical model assumptions
- Example:

| Without Outlier | With Outlier |
|---|---|
| 4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7 | 4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7,300 |
| Mean = 5.45 | Mean = 30.00 |
| Median = 5.00 | Median = 5.50 |
| Mode = 5.00 | Mode = 5.00 |
| Standard Deviation = 1.04 | Standard Deviation = 85.03 |

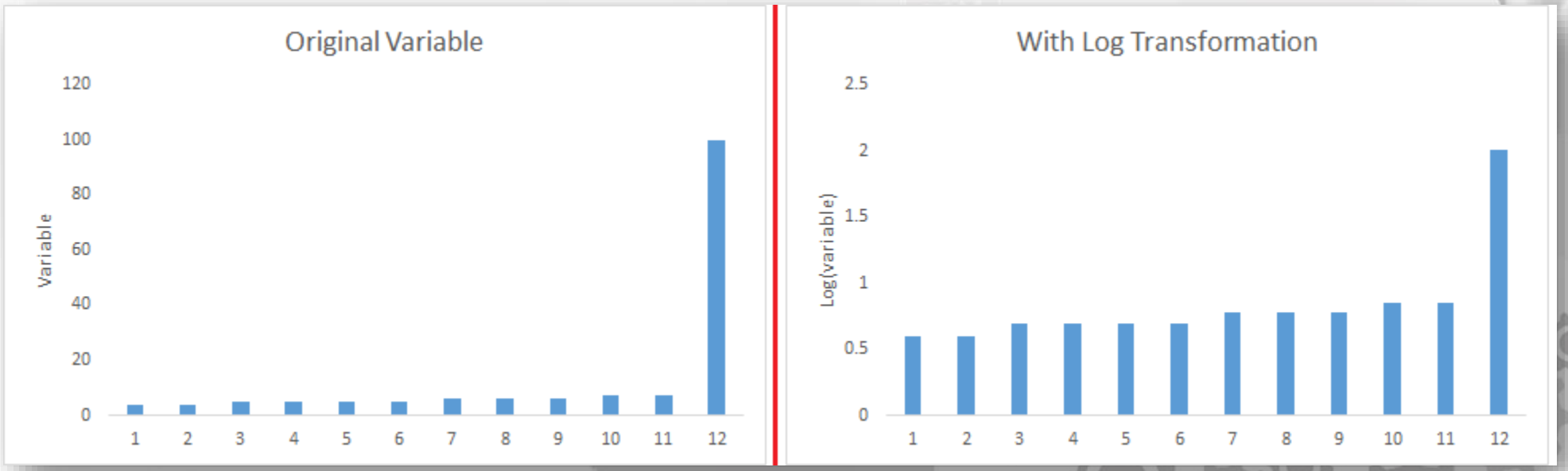
**Source:** [Analytics Vidhya - A Comprehensive Guide to Data Exploration]

# How to Detect Outliers

- Most commonly used method to detect outliers is visualization
- Use various visualization methods, like **Box-plot**, **Histogram**, **Scatter Plot**
- **Other Rules:**
  - Any value, which is beyond the range of -1.5 x IQR to 1.5 x IQR
  - Use capping methods. Any value which out of range of 5th and 95th percentile can be considered as outlier
  - Data points, three or more standard deviation away from mean are considered outlier
  - Outlier detection is merely a special case of the examination of data for influential data points and it also depends on the business understanding
  - Bivariate and multivariate outliers are typically measured using either an index of influence or leverage, or distance. Popular indices such as Mahalanobis' distance and Cook's *D* are frequently used to detect outliers

# Handling of Outliers

- Deleting observations
- Transforming variables & Binning values




- Treat them as a separate group
- Imputing values and other statistical methods

**Source:** [Analytics Vidhya – A Comprehensive Guide to Data Exploration]

# Feature Engineering

- The science (and art) of extracting more information from existing data by using transforming techniques
- Conduct after proceeding the basics of EDA
- **Process of Feature Engineering**
- Variable Transformation
  - **Change the scale** of a variable ➔ **Data Normalization**
  - **Standardize the values** of a variable
  - **Transform complex non-linear relationships into linear relationships**
  - **Symmetric distribution is preferred over skewed distribution**
  - **Implementation point of view** (Human involvement)
- Variable / Feature creation
  - process to generate a new variables / features based on existing variable(s)

**Source:** Analytics Vidhya – A Comprehensive Guide to Data Exploration

# Methods of Variable Transformation

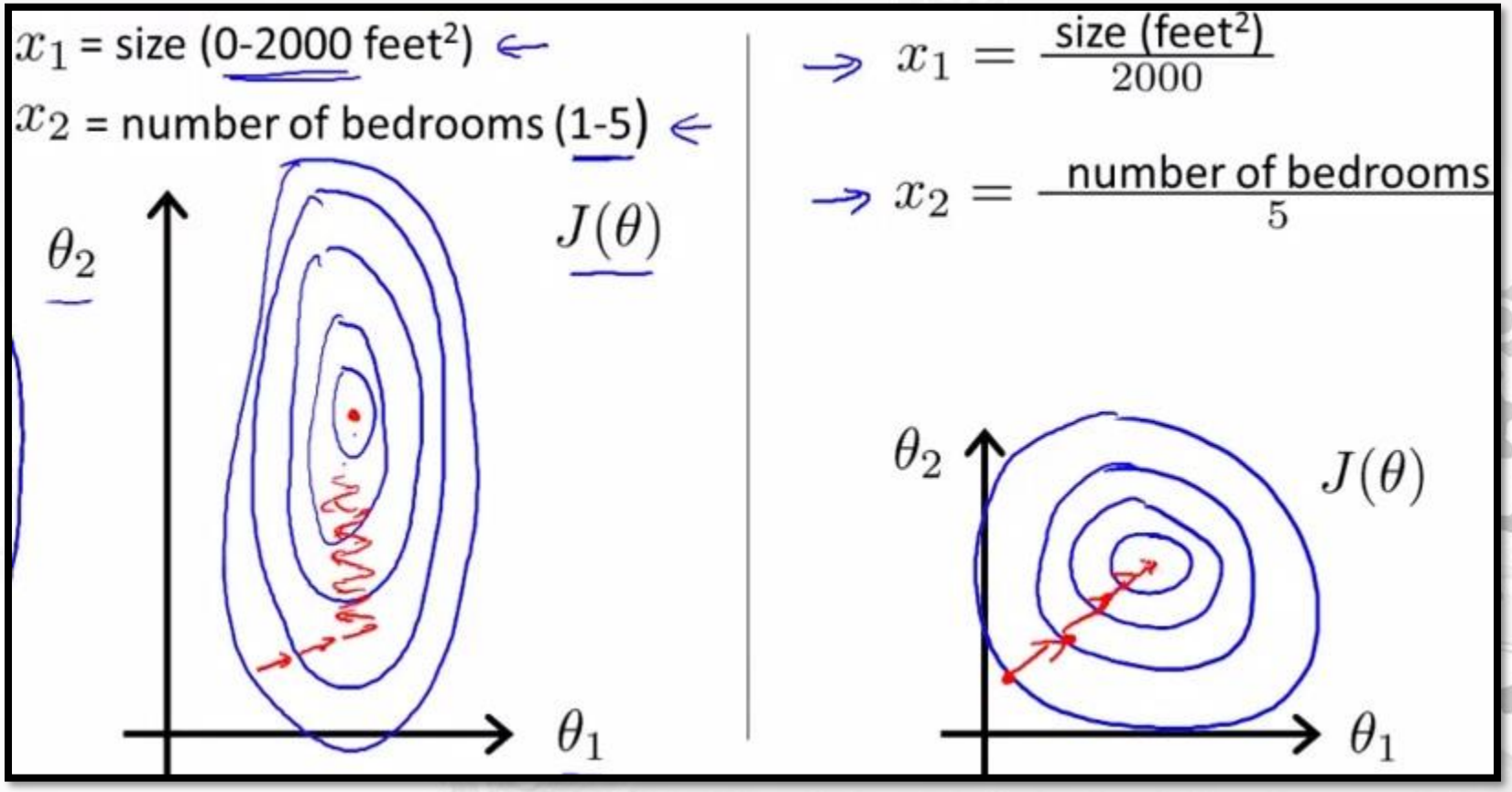
- **Logarithm:** a common transformation method used to change the shape of distribution of the variable on a distribution plot. It is generally used for reducing right skewness of variables. Though, It can't be applied to zero or negative values as well.
- **Square / Cube root:** The square and cube root of a variable has a sound effect on variable distribution. However, it is not as significant as logarithmic transformation. Cube root has its own advantage. It can be applied to negative values including zero. Square root can be applied to positive values including zero.
- **Binning:** It is used to categorize variables. It is performed on original values, percentile or frequency. Decision of categorization technique is based on business understanding.

# Data Normalization

- 資料特徵標準化(normalization)是將特徵資料按比例縮放，讓資料落在某一特定的區間
- 目的：提高後續資料分析的精準度
- Example
  - Given a dataset with a minimum and maximum values, say -23.89 and 7.54990767
  - Normalize a data of 5.6878 to meet the scale between 0 to 1
- Method
  - Min-max normalization：(value-min)/(max-min)
  - Standard deviation normalization:將所有特徵數據轉換成平均為 0、變異數為 1

# Example – using in Gradient Decent

藍色圈圈代表的是特徵的等高線，左圖的特徵x1,x2區間相差非常大，所以對應的等高線非常尖，會導致在使用梯度下降法尋求最佳解時，需要很迭代多次才可以收斂



# Feature / Variable Creation

- **Creating derived variables**

| Emp_Code | Gender | Date | New_Day | New_Month | New_Year |
|---|---|---|---|---|---|
| A001 | Male | 21-Sep-11 | 21 | 9 | 2011 |
| A002 | Female | 27-Feb-13 | 27 | 2 | 2013 |
| A003 | Female | 14-Nov-12 | 14 | 11 | 2012 |
| A004 | Male | 07-Apr-13 | 7 | 4 | 2013 |
| A005 | Female | 21-Jan-11 | 21 | 1 | 2011 |
| A006 | Male | 26-Apr-13 | 26 | 4 | 2013 |
| A007 | Male | 15-Mar-12 | 15 | 3 | 2012 |

- **Creating dummy variables**
  - Dummy variables are also called Indicator Variables. It is useful to take categorical variable as a predictor in statistical models

| Emp_Code | Gender | Var_Male | Var_Female |
|---|---|---|---|
| A001 | Male | 1 | 0 |
| A002 | Female | 0 | 1 |
| A003 | Female | 0 | 1 |
| A004 | Male | 1 | 0 |
| A005 | Female | 0 | 1 |
| A006 | Male | 1 | 0 |
| A007 | Male | 1 | 0 |

Source: Analytics Vidhya – A Comprehensive Guide to Data Exploration

# End Notes

- As mentioned in the beginning, quality and efforts invested in data exploration differentiates a good model from a bad model
- The aim of EDA was to provide an in depth and step by step guide to an extremely important process in data science
- Treat seriously to the quality of data

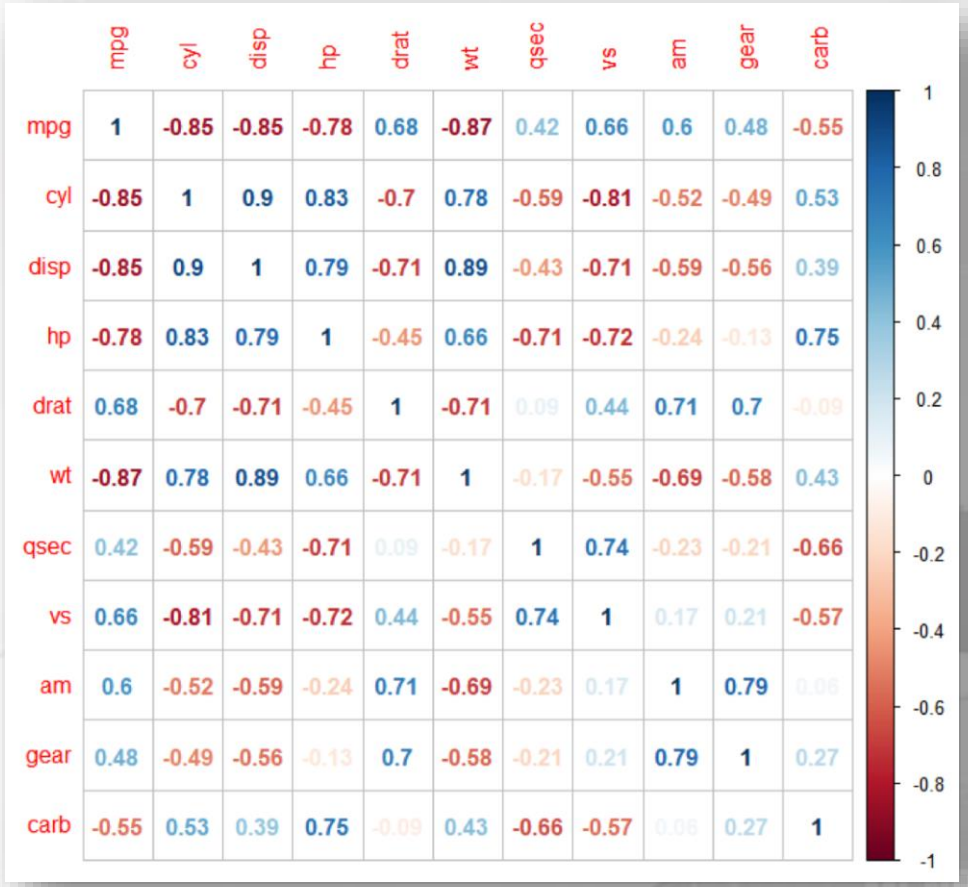
# Workshop – Hands-on Practice

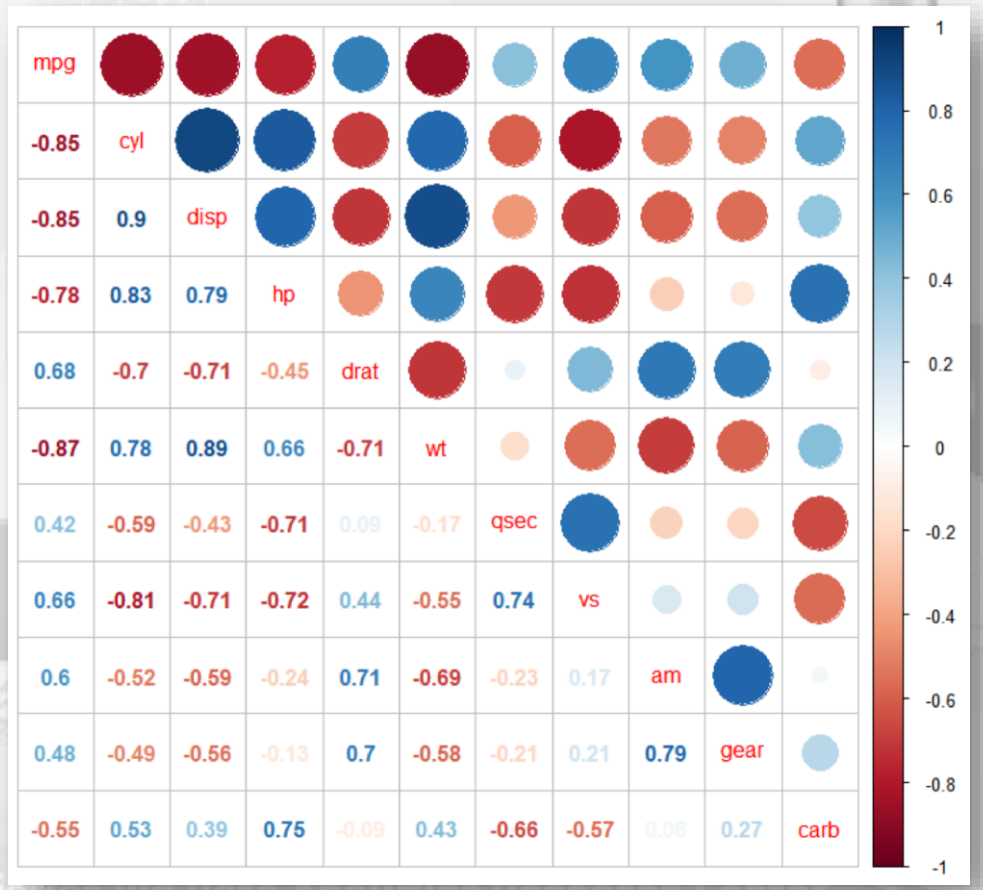
1. 相關係數矩陣圖Correlation Analysis :
   - 2.EDA_Case 1-Basic-Cor_plot.R
2. 圖資視覺化 (GIS地理資訊系統)
   - leaflet套件於GIS的基本應用(逢甲大學為例)
3. 犯罪巨量資料分析(Crime Analysis using Data Visualization)
   - Case Study: Crime Analysis of San Francisco

# 1. Correlation Analysis

- Data source: default dataset-mtcars
- Goal: plot correlation matrix
- Using R function- cor() & Package-corrplot

| | mpg | cyl | disp | hp | drat | wt | qsec | vs | am | gear | carb |
|---|---|---|---|---|---|---|---|---|---|---|---|
| mpg | 1 | -0.85 | -0.85 | -0.78 | 0.68 | -0.87 | 0.42 | 0.66 | 0.6 | 0.48 | -0.55 |
| cyl | -0.85 | 1 | 0.9 | 0.83 | -0.7 | 0.78 | -0.59 | -0.81 | -0.52 | -0.49 | 0.53 |
| disp | -0.85 | 0.9 | 1 | 0.79 | -0.71 | 0.89 | -0.43 | -0.71 | -0.59 | -0.56 | 0.39 |
| hp | -0.78 | 0.83 | 0.79 | 1 | -0.45 | 0.66 | -0.71 | -0.72 | -0.24 | -0.13 | 0.75 |
| drat | 0.68 | -0.7 | -0.71 | -0.45 | 1 | -0.71 | 0.09 | 0.44 | 0.71 | 0.7 | -0.09 |
| wt | -0.87 | 0.78 | 0.89 | 0.66 | -0.71 | 1 | -0.17 | -0.55 | -0.69 | -0.58 | 0.43 |
| qsec | 0.42 | -0.59 | -0.43 | -0.71 | 0.09 | -0.17 | 1 | 0.74 | -0.23 | -0.21 | -0.66 |
| vs | 0.66 | -0.81 | -0.71 | -0.72 | 0.44 | -0.55 | 0.74 | 1 | 0.17 | 0.21 | -0.57 |
| am | 0.6 | -0.52 | -0.59 | -0.24 | 0.71 | -0.69 | -0.23 | 0.17 | 1 | 0.79 | 0.06 |
| gear | 0.48 | -0.49 | -0.56 | -0.13 | 0.7 | -0.58 | -0.21 | 0.21 | 0.79 | 1 | 0.27 |
| carb | -0.55 | 0.53 | 0.39 | 0.75 | -0.09 | 0.43 | -0.66 | -0.57 | 0.06 | 0.27 | 1 |



# 2. 圖資視覺化

leaflet套件GIS的基本應用

# 以逢甲大學為例

- 逢甲大學的經緯度
- 載入leaflet套件
- 建立以逢甲大學為中心的圖資視覺化

# 3. Crime Analysis

Case Study: San Francisco

# Case Study - San Francisco

- Overview: San Francisco (SF, CA, USA)
  - 「Fog City」is famous with its fog. One of the most populous city of USA with the population 884,363
- Goal
  - Aims to explore the crime rate using EDA
  - Perform EDA of gaining insight
  - Mapping the Hot Spot of LA Crime Event
- Data Source
  - Incidents of crime in SF dating back to 2003
- Data Scale
  - 2007-2016

# Q & A