



COS10022 – DATA SCIENCE PRINCIPLES

Dr Pei-Wei Tsai (Lecturer, Unit Convenor)
ptsai@swin.edu.au, EN508d

SWIN
BUR
NE

SWINBURNE
UNIVERSITY OF
TECHNOLOGY

The background is a blurred collage of data-related graphics. It includes a world map composed of blue dots, orange bar charts, and orange line graphs with markers. The overall color palette is dominated by blue, orange, and white.

Week 09

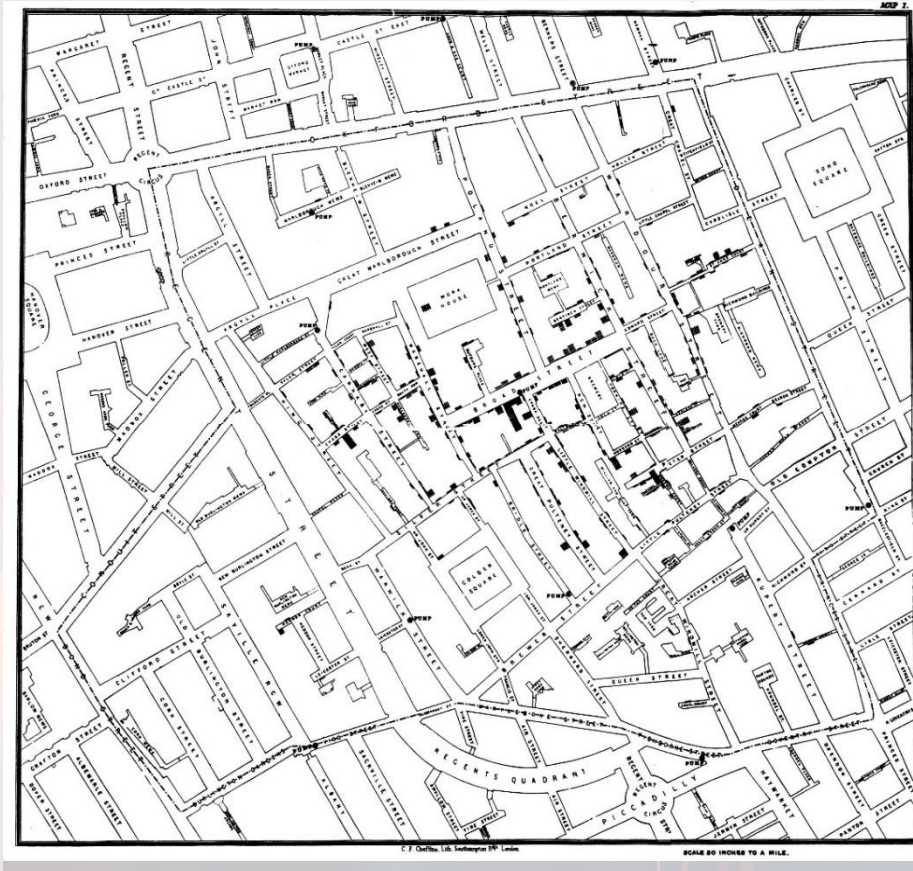
Basic Data Analytics Methods

COS10022 - Data science Principles

Outline

- **Part A: Exploratory Data Analysis**
 - Descriptive statistics
 - Visualization before Analysis
 - Dirty Data
 - Visualizing a Single Variable
 - Examining Multiple Variables
 - Data Exploration versus Presentation
- **Part B: Statistical Methods for Evaluation**
 - Hypothesis Testing
 - Difference of Means
 - Wilcoxon Rank-Sum Test
 - Type I and Type II Errors
 - Power and Sample Size
 - ANOVA (Analysis of Variance)

Survey and Visualize



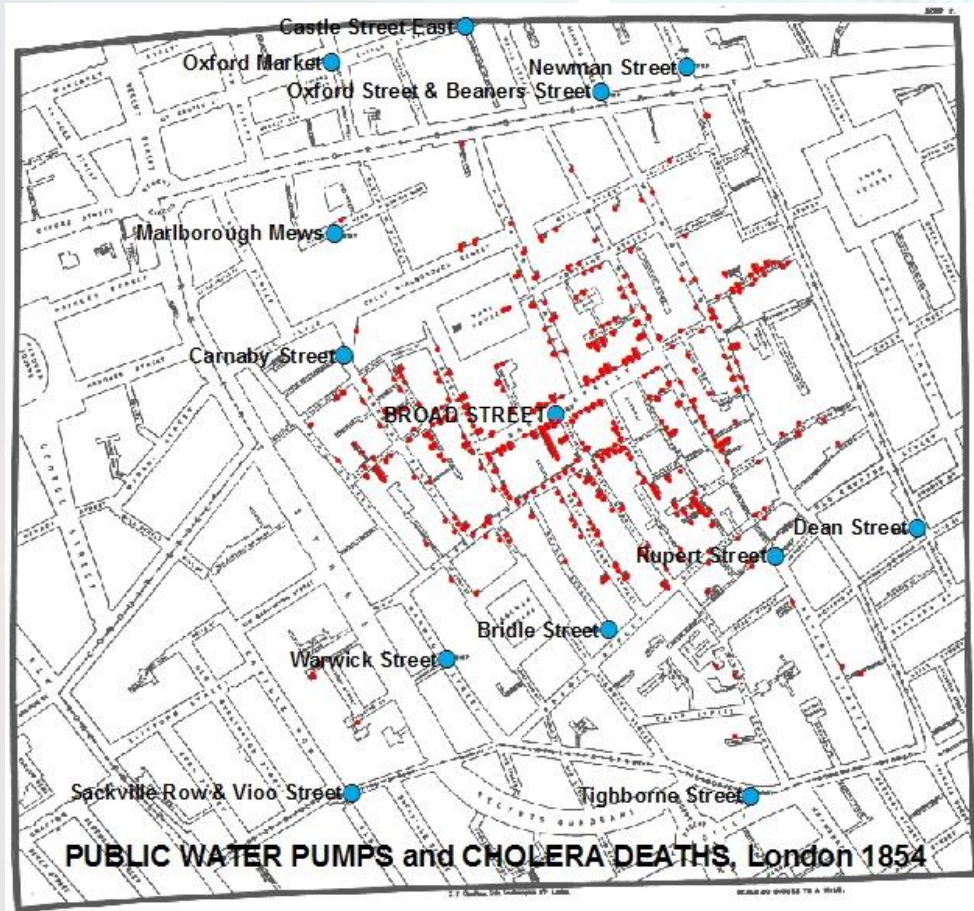
1. Dr. John Snow is one of the founding fathers of modern epidemiology.
2. On the left is John Snow's map showing the clusters of cholera cases in the London epidemic of 1854.

By John Snow - Published by C.F. Cheffins, Lith, Southampton Buildings, London, England, 1854 in Snow, John.

On the Mode of Communication of Cholera, 2nd Ed, John Churchill, New Burlington Street, London, England, 1855.

(This image was originally from en.wikipedia; description page is/was here. Image copied from http://matrix.msu.edu/~johnsnow/images/online_companion/chapter_images/fig1-2-5.jpg), Public Domain, <https://commons.wikimedia.org/w/index.php?curid=2278605>

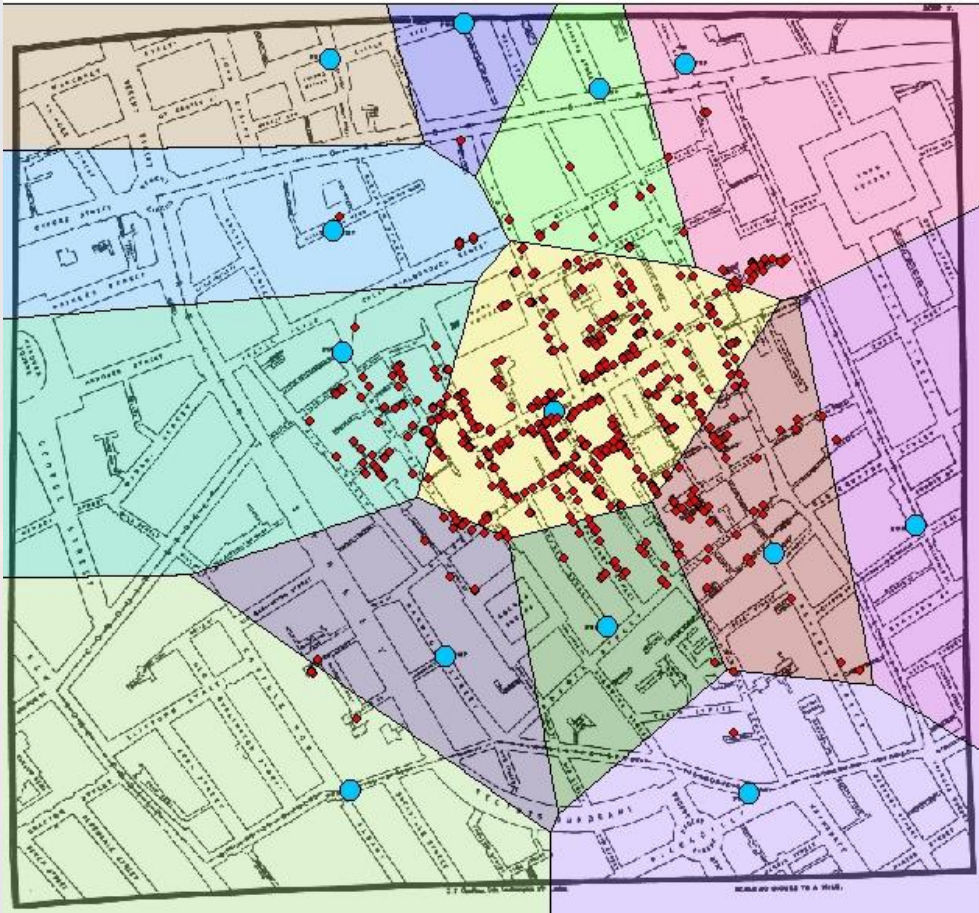
Survey and Visualize



1. Dr. John Snow is one of the founding fathers of modern epidemiology.
2. On the left is John Snow's map showing the clusters of cholera cases in the London epidemic of 1854.
3. The map shows the locations of 13 public wells (blue dots) surrounding the areas where 578 cholera deaths (red bars) were recorded.

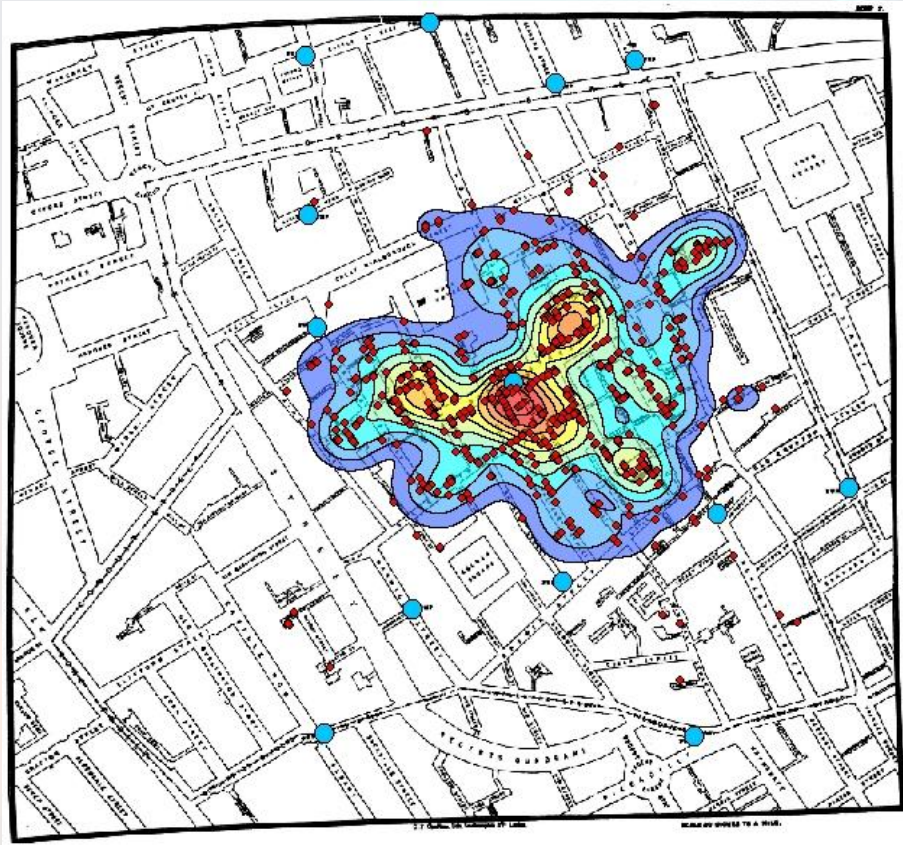
Source: <https://www1.udel.edu/johnmack/frec682/cholera/cholera2.html>

Survey and Visualize



1. Dr. John Snow is one of the founding fathers of modern epidemiology.
2. On the left is John Snow's map showing the clusters of cholera cases in the London epidemic of 1854.
3. The map shows the locations of 13 public wells (blue dots) surrounding the areas where 578 cholera deaths (red bars) were recorded.
4. Snow noted a particular clustering of deaths (yellow Thiessen polygon) around a single public well on the southwest corner of the intersection of Broad Street and Cambridge Street.

Survey and Visualize



Source: <https://www1.udel.edu/johnmack/frec682/cholera/cholera2.html>

1. Dr. John Snow is one of the founding fathers of modern epidemiology.
2. On the left is John Snow's map showing the clusters of cholera cases in the London epidemic of 1854.
3. The map shows the locations of 13 public wells (blue dots) surrounding the areas where 578 cholera deaths (red bars) were recorded.
4. Snow noted a particular clustering of deaths (yellow Thiessen polygon) around a single public well on the southwest corner of the intersection of Broad Street and Cambridge Street.
5. Kernel density visualization shows the high-density of deaths in areas adjacent to the Broad Street well.

Survey and Visualize

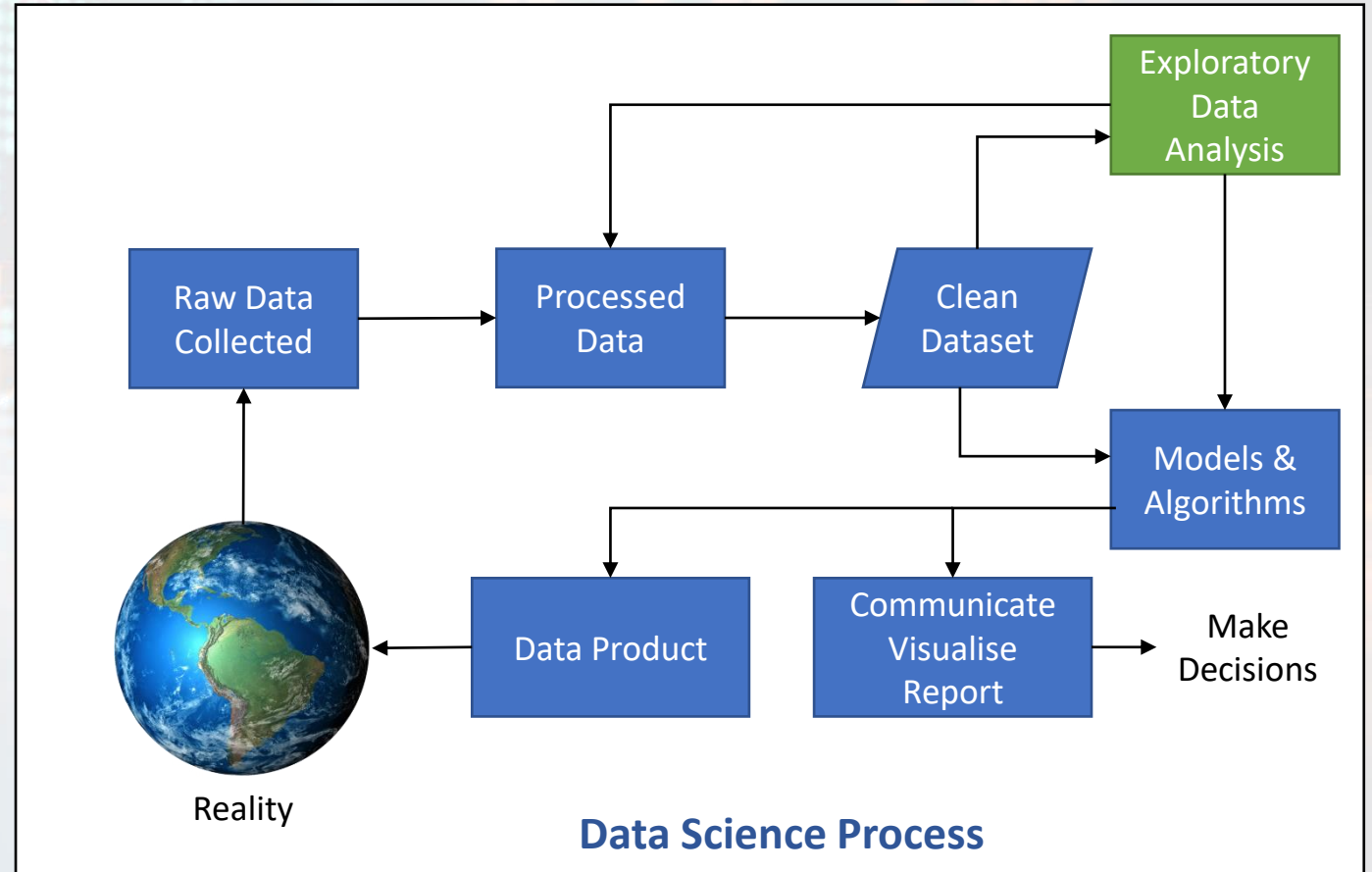
Aisch, G 2012, 'Using Data Visualization to Find Insights in Data', in *Data Journalism Handbook* (ed.), O'Reilly Media
Link: http://datajournalismhandbook.org/1.0/en/understanding_data_7.html

Four (4) important types of data visualisation:

- **Tables** are very powerful in dealing with a relatively small number of data points.
- **Charts** allow mapping multiple dimensions of the data to visual properties of geometric shapes.
- **Maps** can powerfully connect data to the physical world.
- **Graphs (networks)** show the interconnections between various types of real-world objects.

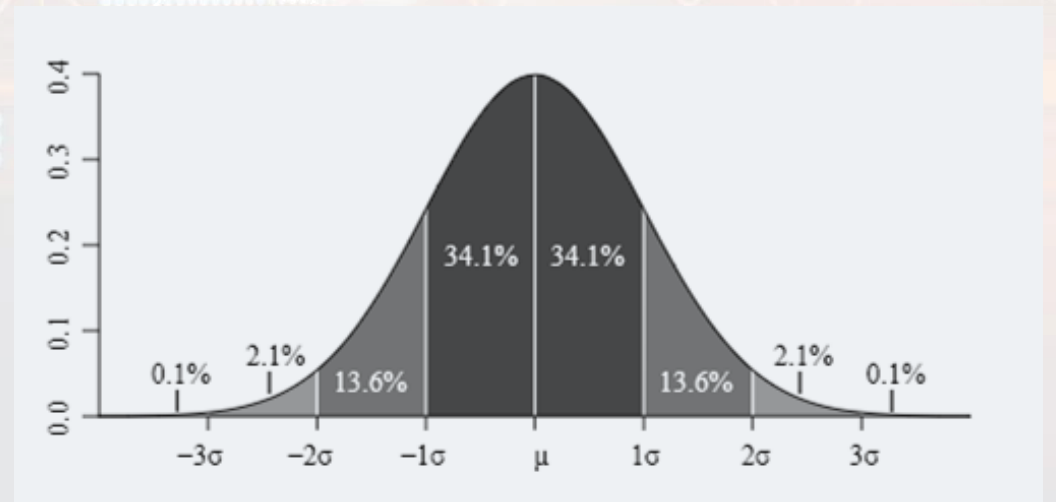
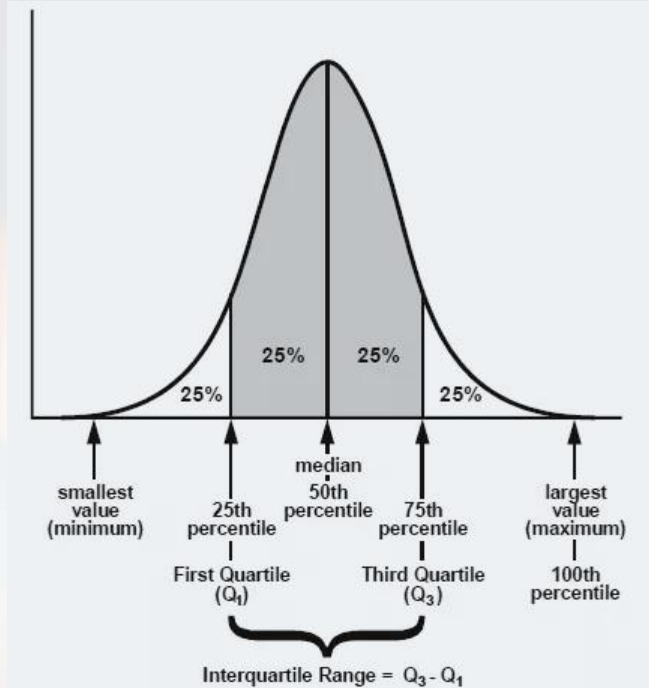
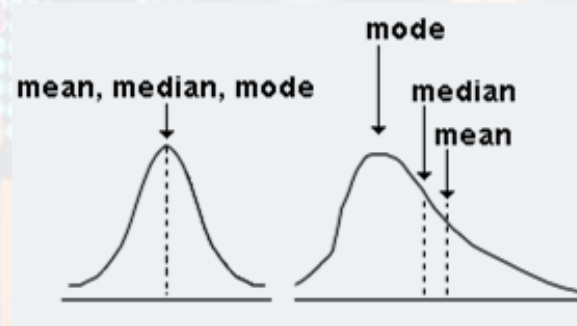
Part A: Exploratory Data Analysis (EDA)

- **EDA** is an approach to analysing datasets to summarise their main characteristics, often with visual methods.
- **Why is EDA important?**
 - Gain new insight
 - Explore data structures
 - Detect missing data
 - Check significant variables
 - Examine relationship between variables
 - Select an appropriate model
 - Check model assumptions



Descriptive Statistics

- **Descriptive statistics** quantitatively describe the main features of data.
- **Main data features:**
 - **Measures of central tendency** – represent a ‘centre’ around which measurements are distributed.
 - E.g. mean and median
 - **Measures of variability** – represent the ‘spread’ of the data from the ‘centre’.
 - E.g. standard deviation
 - **Measures of relative standing** – represent the ‘relative position’ of specific measurements in the data.
 - E.g. quantiles



Descriptive Statistics

Mean

- Sum all the numbers and divide by their count

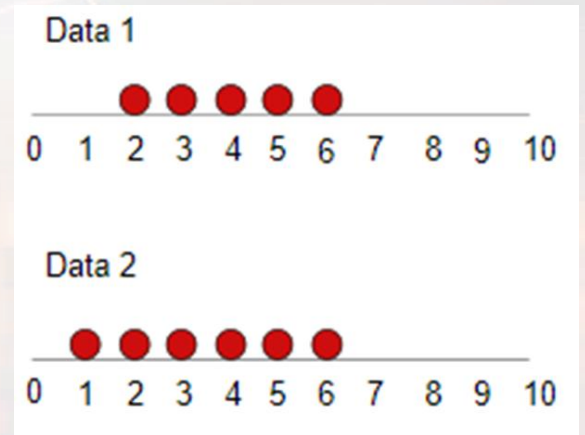
- $$X = (x_1 + x_2 + \dots + x_n) / n$$

- E.g. Mean = $(2 + 3 + 4 + 5 + 6) / 5 = 4$*



Median

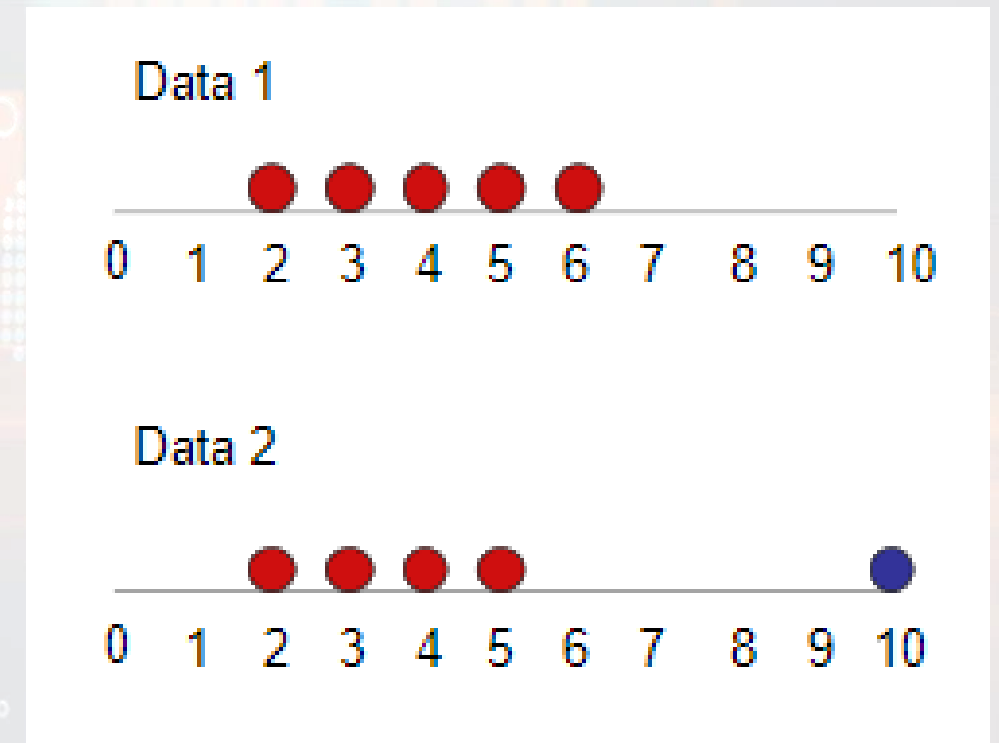
- The exact middle value.
- When the count is odd, find the middle value of sorted data.
 - E.g. For Data 1, the median is 4.
- When the count is even, find the means of the middle two values.
 - E.g. For Data 2, the median is $(3+4)/2 = 3.5$.



Descriptive Statistics

Mean VS. Median

- When data distribution is **skewed**, **median** is **more meaningful** than **mean**.
- When data has **outliers**, **median** is more robust.
 - The blue data point is the outlier in data 2.
- For Data 1,
 - Mean = 4, Median = 4
- For Data 2,
 - Mean = 4.8, Median = 4



Descriptive Statistics

Standard Deviation (SD)

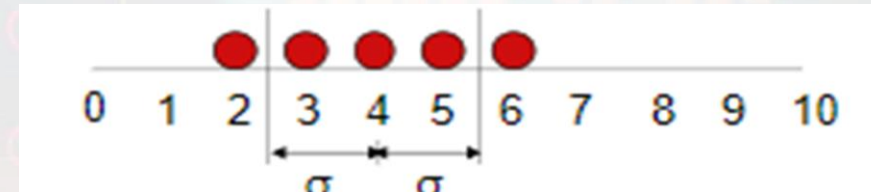
- Computation steps:
 - Compute mean
 - Compute the deviation of each measurement from the mean
 - Square the deviations
 - Sum the squared deviations
 - Divide by (count-1) or by count.
 - Compute the square root.

For Population:

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

For Sample:

$$SD = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$



Mean = 4

Deviations: -2, -1, 0, 1, 2

Squared deviations: 4, 1, 0, 1, 4

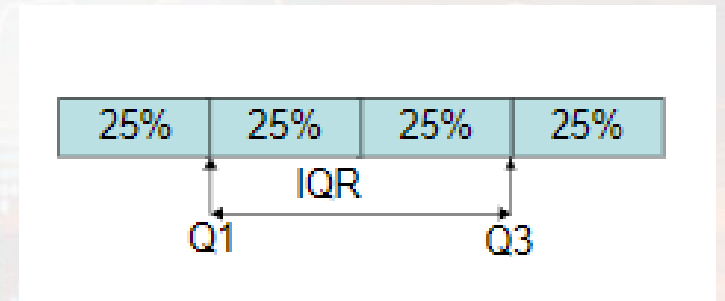
Sum = 10

Standard deviation = $\sqrt{10/4} = 1.58$

Descriptive Statistics

Quartiles

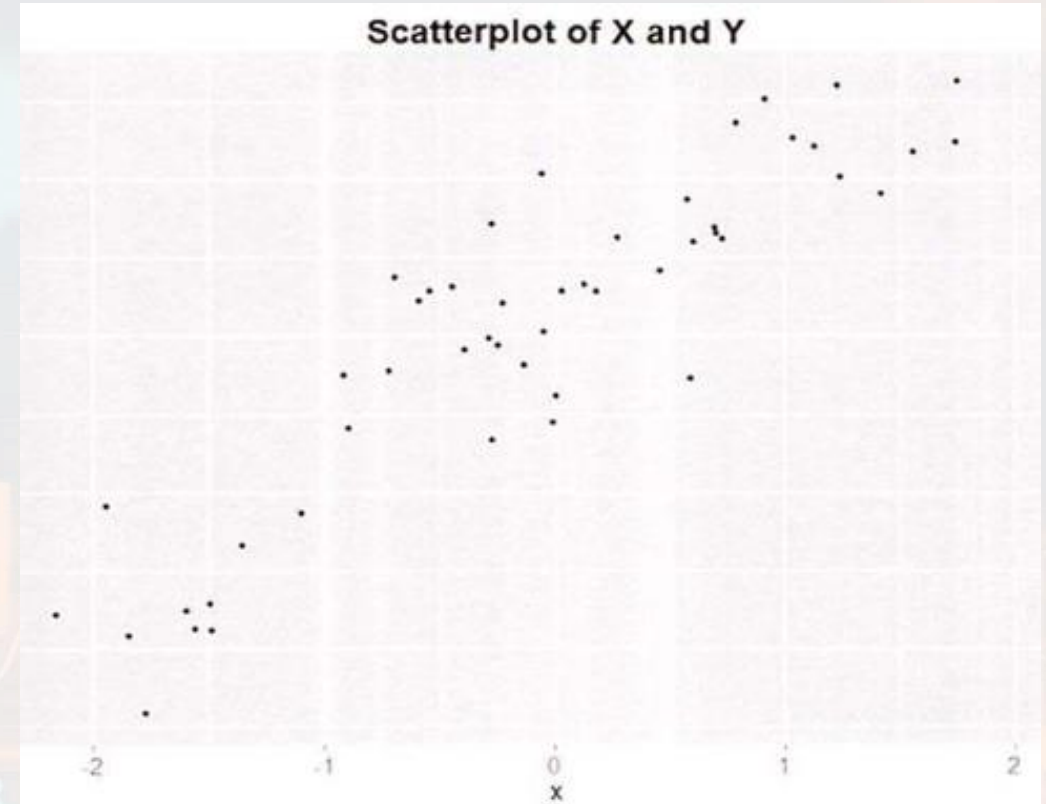
- 1st quartile is the measurement with 25% measurements smaller and 75% larger – lower quartile (Q1).
- 2nd quartile is the median.
- 3rd quartile is the measurements with 75% measurements smaller and 25% larger – upper quartile (Q3).
- Inter quartile range (IQR) is a measure of variability between quartiles. For example, the difference between Q3 and Q1, i.e. $Q3 - Q1$.



Why Visualisation?

The descriptive statistics below shows the range of x and y, but it is not clear what the relationship may be between these two variables.

x		y	
Min.	: -1.90483	Min.	: -2.16545
1 st Qu.	: -0.66321	1 st Qu.	: -0.71451
Median	: 0.09367	Median	: -0.03797
Mean	: 0.02522	Mean	: -0.02153
3 rd Qu.	: 0.65414	3 rd Qu.	: 0.55738
Max.	: 2.18471	Max.	: 1.70199



A useful way to detect **patterns** and **anomalies** in the data is through the exploratory data analysis with visualization. The **scatterplot** above depicts the relationship between x and y.

Visualization before Analysis

Importance of graphs in statistical analyses:

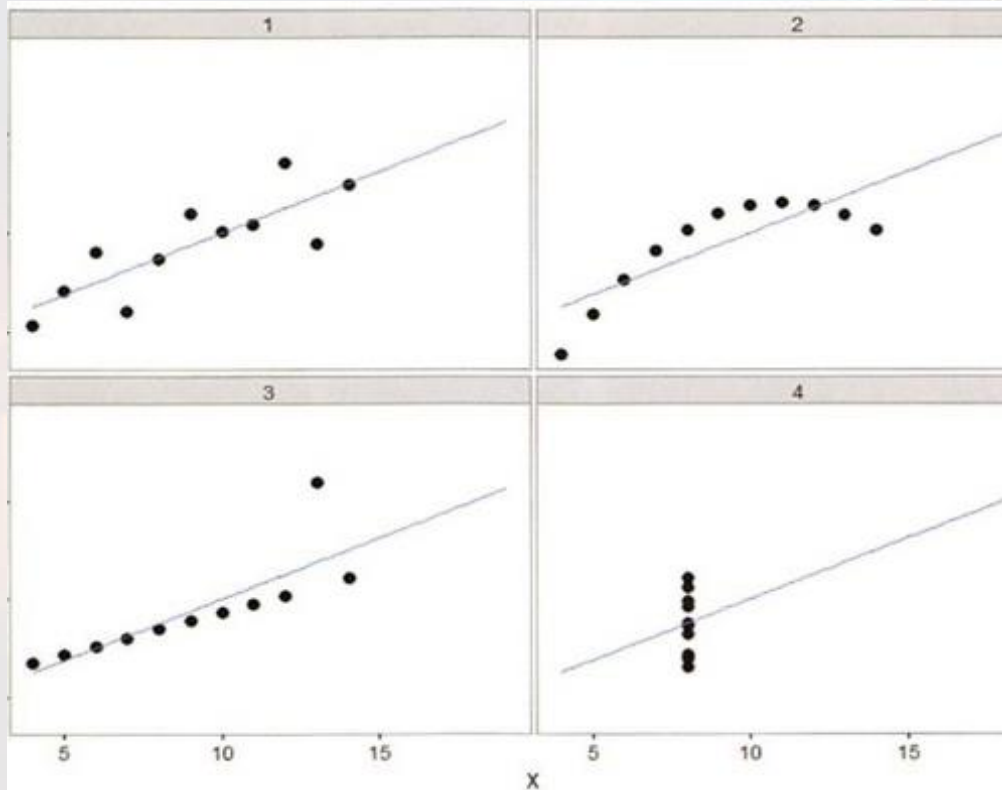
Anscombe's Quartet consists of four datasets with nearly identical statistical properties.

Statistical Property		Value
Mean of x	:	9
Variance of x	:	11
Mean of y	:	7.50 (to 2 decimal points)
Variance of y	:	4.12 or 4.13 (to 2 decimal points)
Correlations between x and y	:	0.816
Linear regression line	:	$y = 3.00 + 0.50 x$ (to 2 decimal points)

One might concluded that these four datasets are quite similar. However...

Visualisation before Analysis

Each dataset is plotted as a scatterplot, and the fitted lines are the result of applying linear regression models.

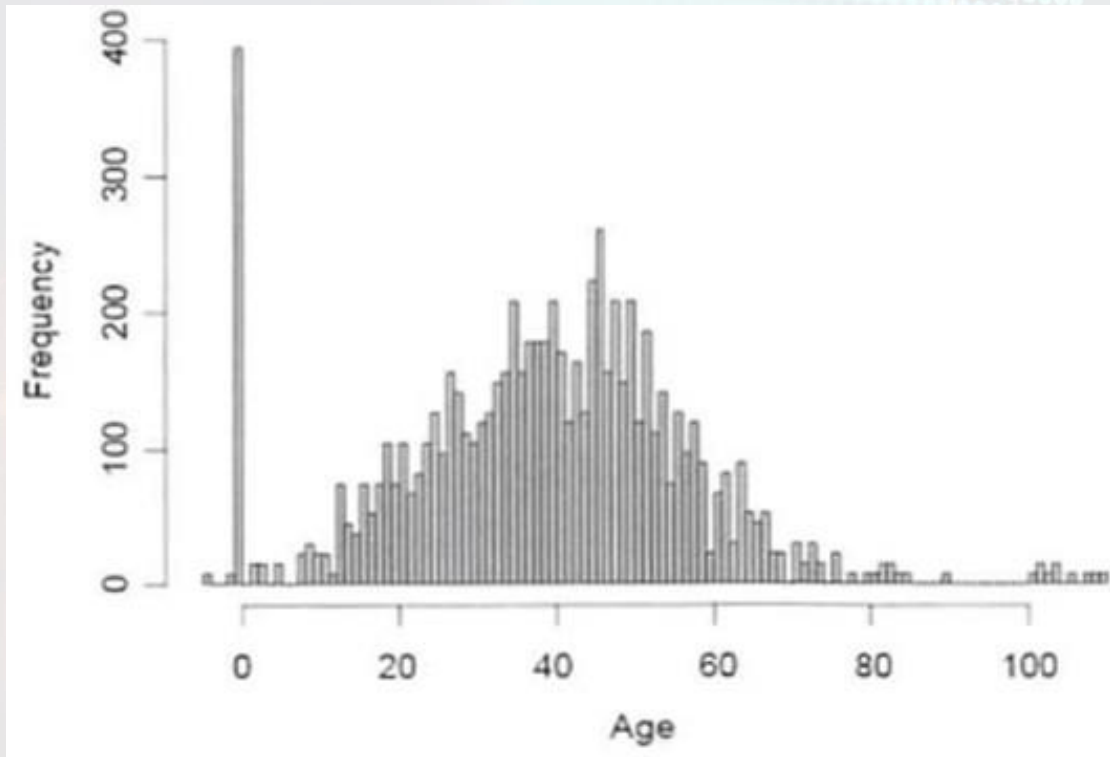


The figures show that:

- The regression line fits Dataset 1 reasonably well.
- Dataset 2 is definitely nonlinear.
- Dataset 3 exhibits a linear trend, with one apparent outlier at $x = 13$.
- With only points at two x values, it is not possible to determine that the linearity assumption is proper.

Dirty Data

Age distribution of bank account holders.

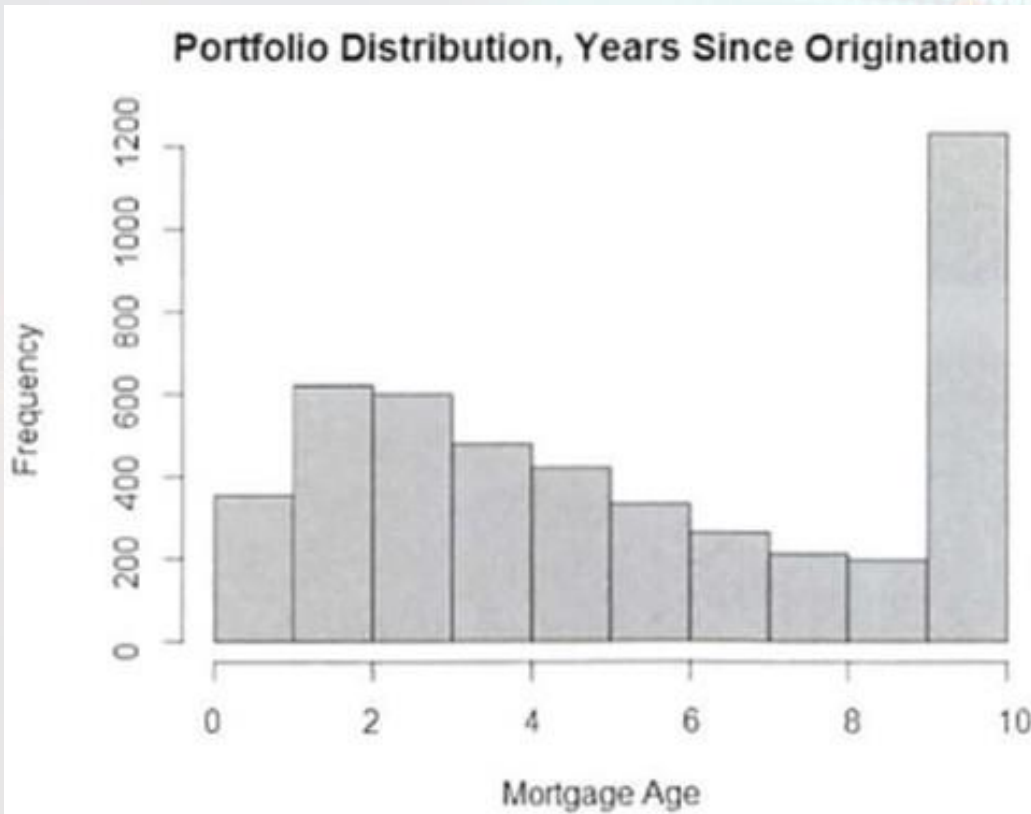


An example of how dirty data can be detected in the data exploration phase with visualization. The figure shows that:

- The **median** age of the account holders is **around 40**.
- A huge spike of customers who are **zero y.o. or have negative ages**. This is likely to be evidence of **missing data**.
- Account holders older than 100 **may be** due to bad data caused by **typos**.
- The dirty data should be **retained** for further analyses. Besides, data **cleansing** need to be performed.

Dirty Data

Age of Mortgage in a bank's home loan portfolio.



Another example of **dirty data**. The figure shows that:

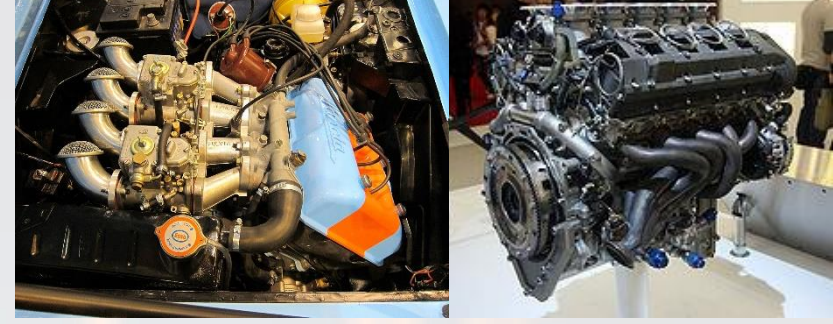
- The mortgages are **no more than 10 years old**.
- 10-year-old mortgages have a **disproportionate frequency** compared to the rest of the population.
- **Possible explanation**: Mortgages aged more than 10 years were **included** into the 10th bin.
- Analysts the data further and decide the most appropriate way to perform data cleansing.

Visualizing a Single Variable

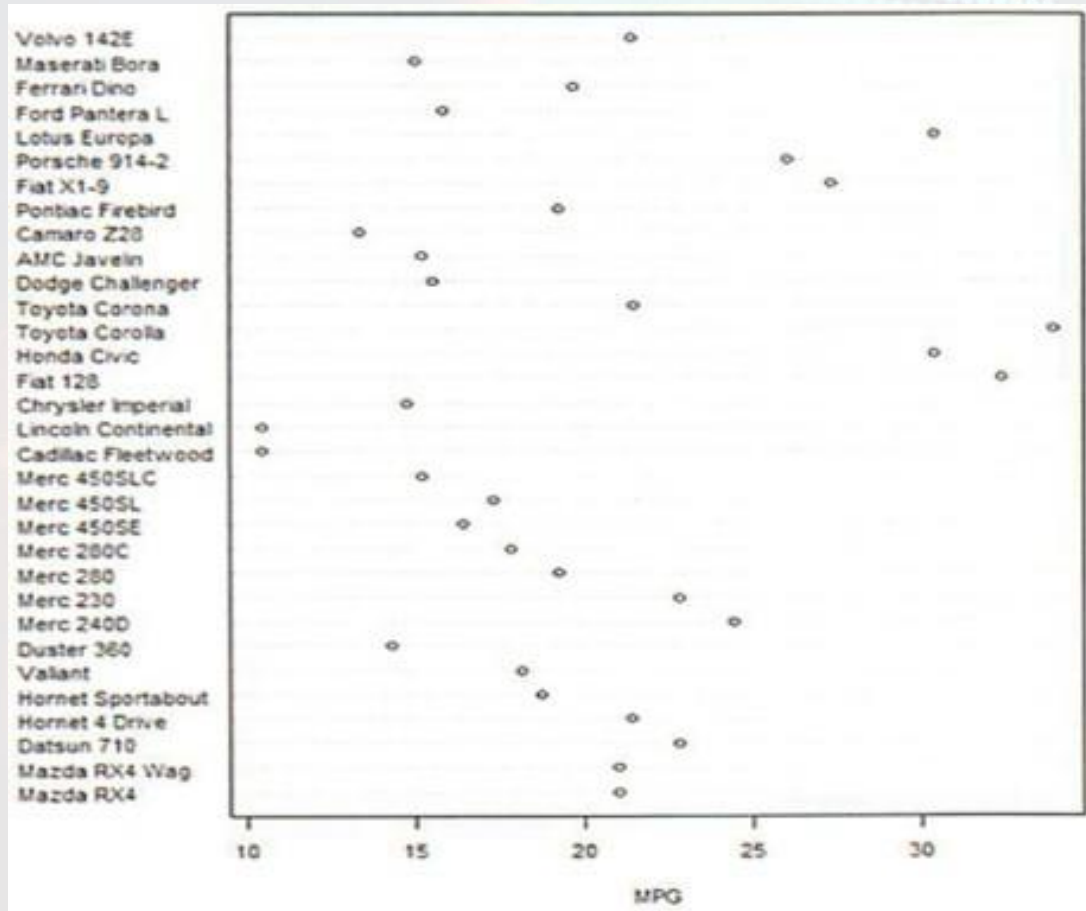
Example functions for visualizing a single variable using R.

Function	Purpose
<code>plot(data)</code>	Scatterplot where x is the index and y is the value; suitable for low-volume data
<code>barplot(data)</code>	Barplot with vertical or horizontal bars
<code>dotchart(data)</code>	Cleveland dot plot [12]
<code>hist(data)</code>	Histogram
<code>plot(density(data))</code>	Density plot (a continuous histogram)
<code>stem(data)</code>	Stem-and-leaf plot
<code>rug(data)</code>	Add a rug representation (1-d plot) of the data to an existing plot

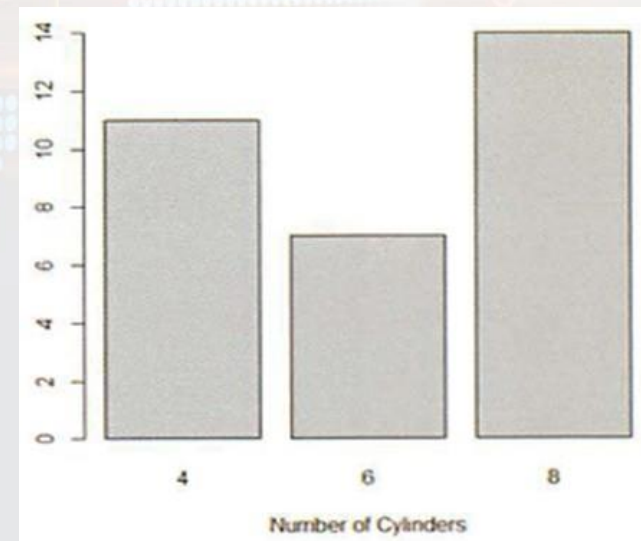
Visualizing a Single Variable



Dotchart on the Miles Per Gallon of Cars



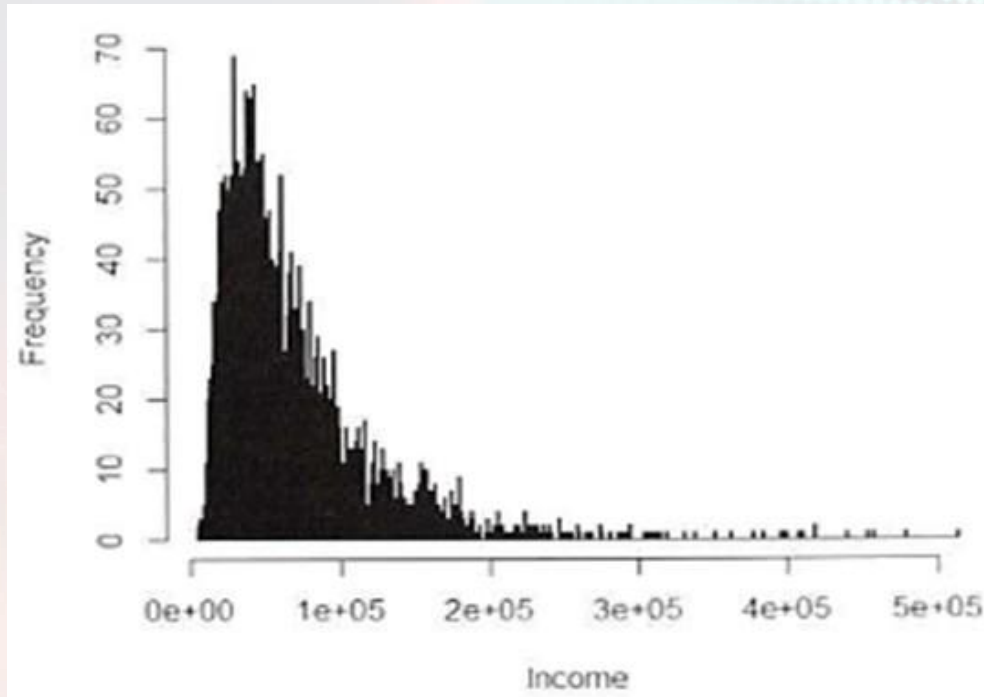
- A dotchart shows each item of numerical data above a number line, or horizontal axis. Dotcharts make it easy to see **gaps** and **clusters** in a dataset, as well as **how the data spreads along the axis**.
- Barplots gather data into categories, allowing us to quickly compare values for each category.



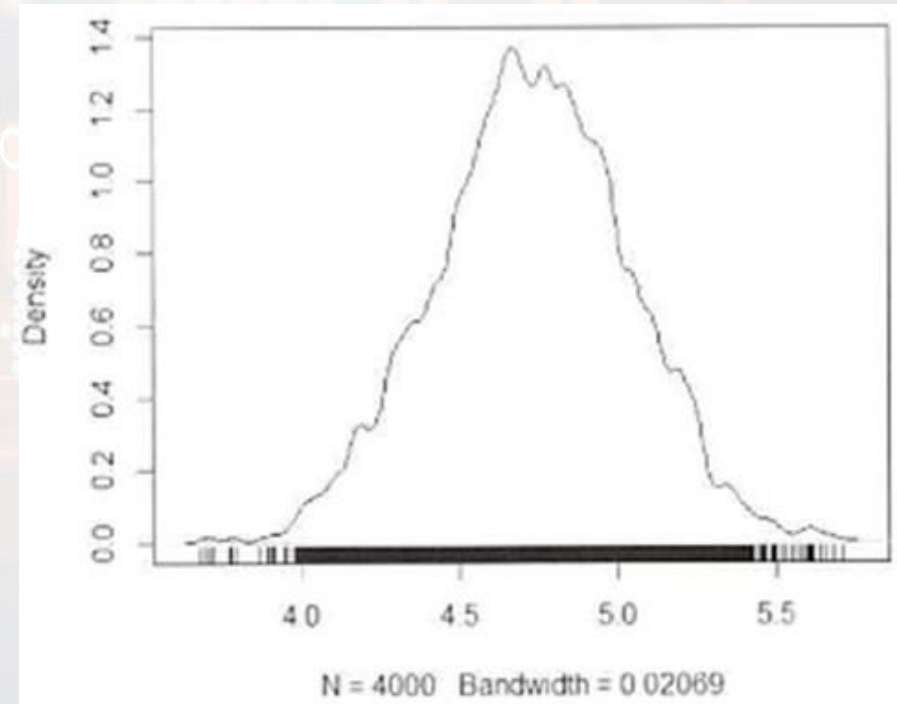
Barplot on the Distribution of Car Cylinder Counts

Visualizing a Single Variable

Histogram of Income



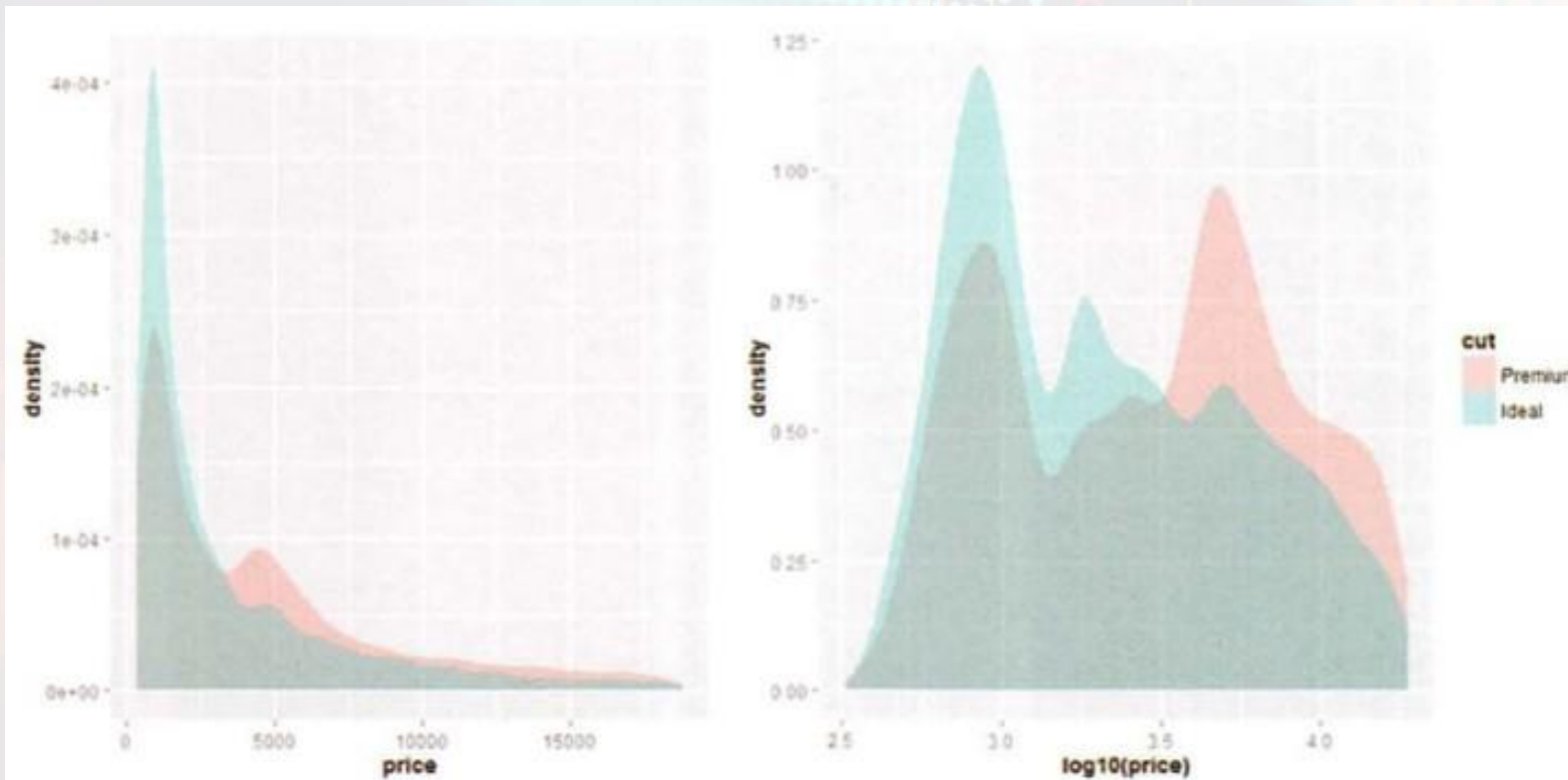
Density Plot of Income (log10 scale)



- If the data is **skewed**, viewing the **logarithm** of the data can help detect structures that might otherwise be overlooked in a graph with a regular, nonlogarithm scale.

Visualizing a Single Variable

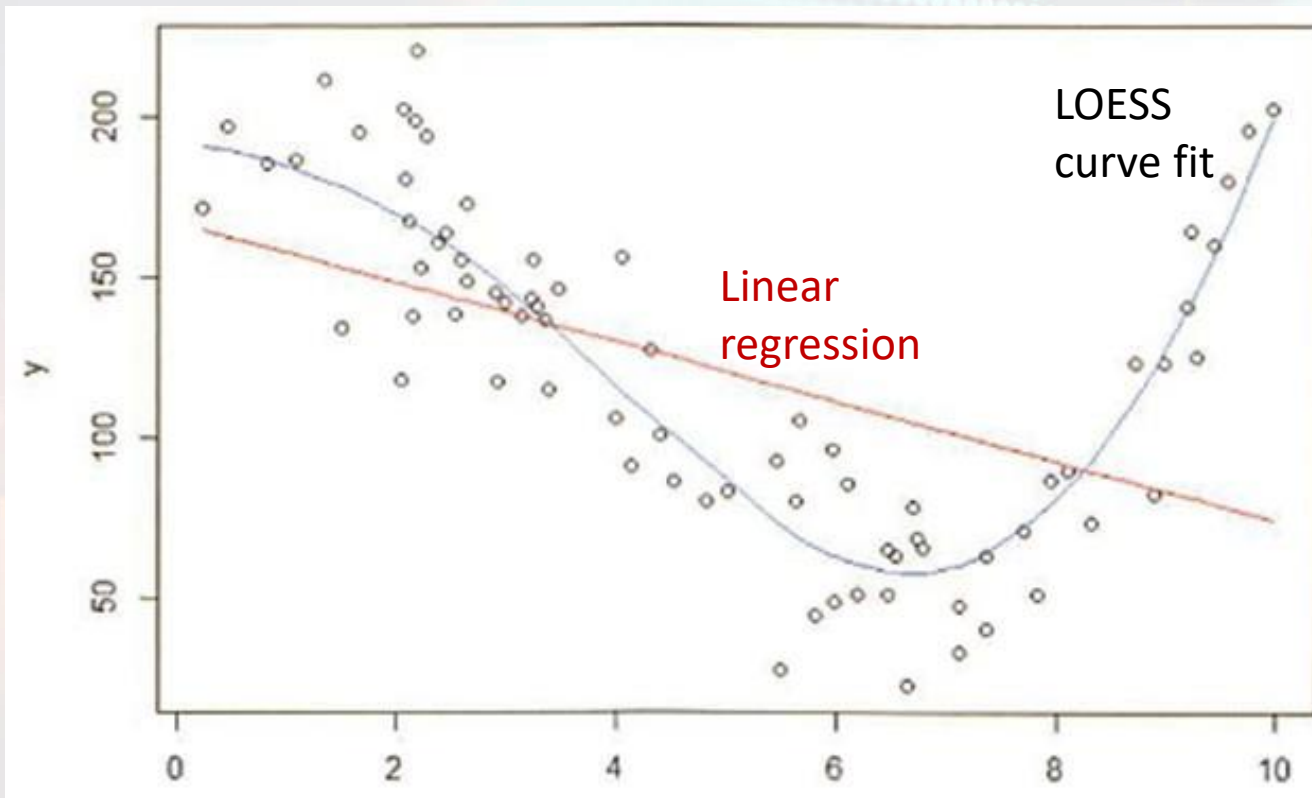
Density Plots for Premium and Ideal Cuts of Diamonds



- The right density plot shows more detail of the diamond prices than the left density plot.
- The two humps in the premium cut represent two distinct groups of diamond prices:
 - One group centres around $\log_{10} \text{ price} = 2.9$ (~\$794).
 - The other group centres around $\log_{10} \text{ price} = 3.7$ (~\$5,012).

Examining Multiple Variables

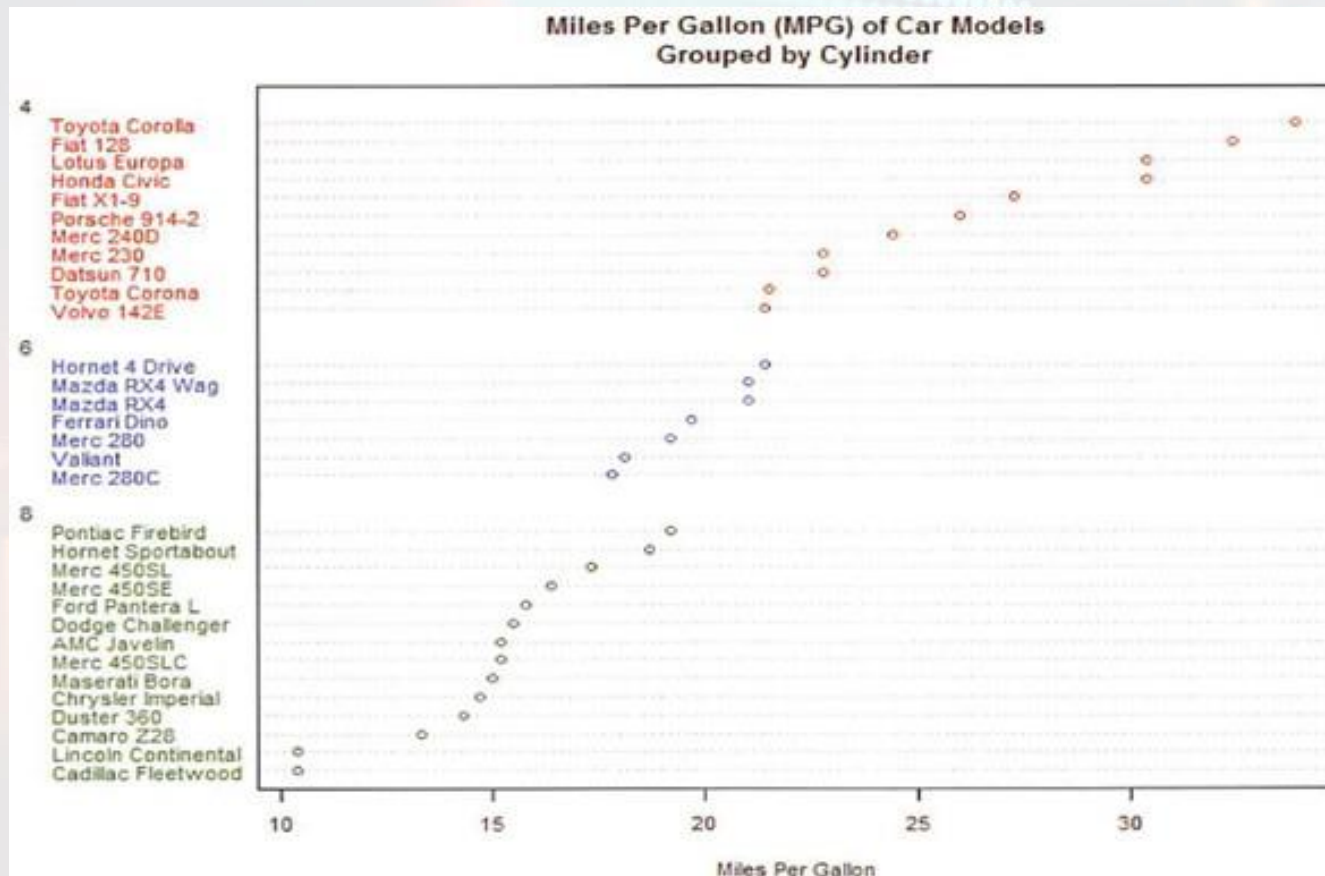
Examining two variables with regression



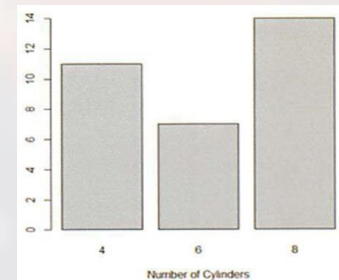
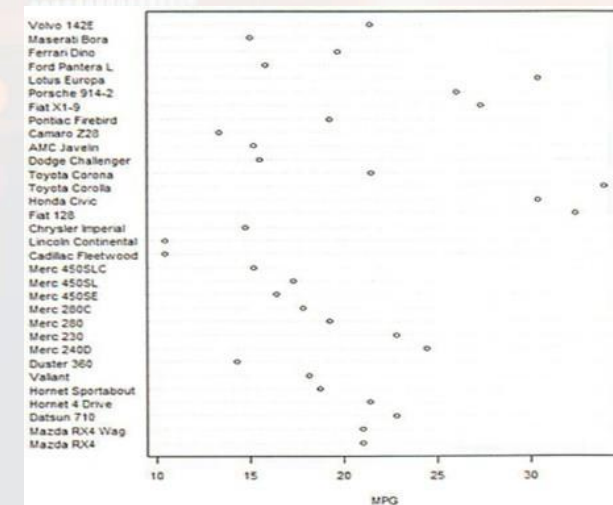
- The scatterplot portrays the relationship of x and y.
- The red line is the fitted line from the linear regression. The regression line does not fit the data well.
- The LOESS() function is used to fit a nonlinear line to the data. The LOESS curve fits the data better than linear regression.
- This is a case in which linear regression cannot model the relationship between the two variables.

Examining Multiple Variables

Dotplot to visualize multiple variables

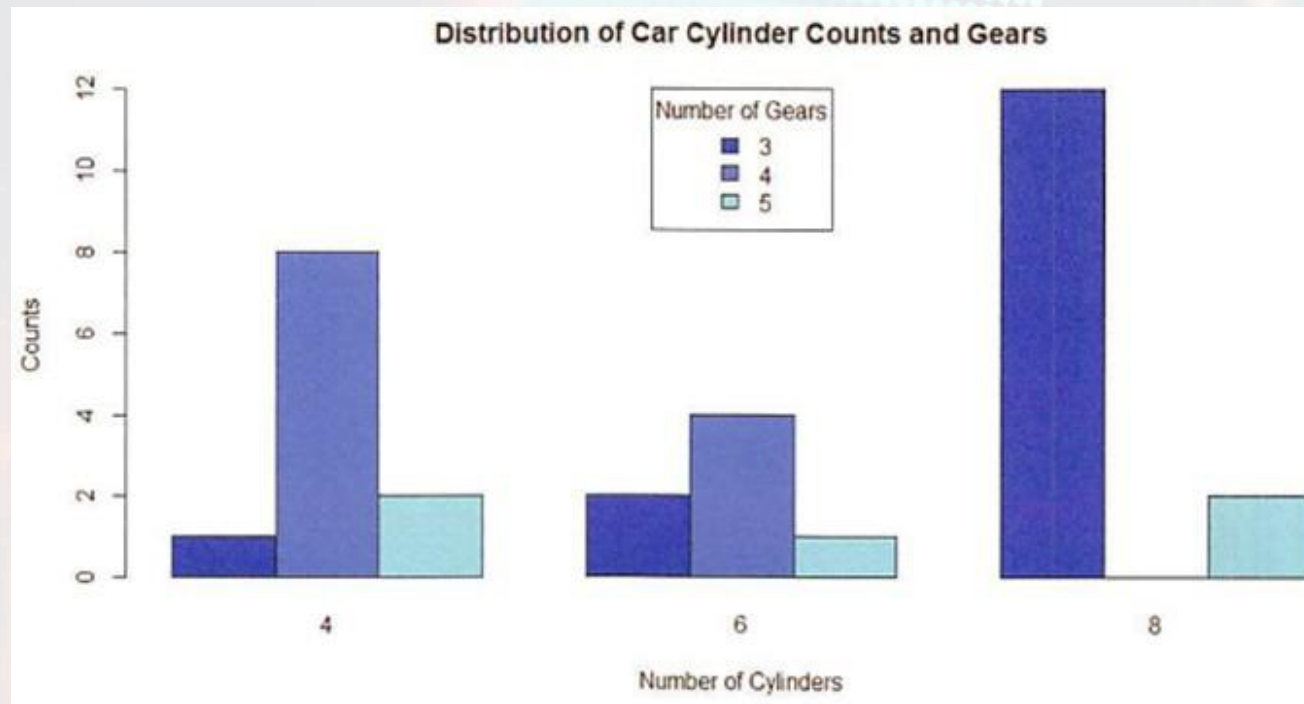


- This figure shows a dotchart that groups vehicle cylinders at the y-axis and uses colours to distinguish different cylinders.
- The vehicles are **sorted** according to their Miles Per Gallon (MPG) values.



Examining Multiple Variables

Barplot to visualize multiple variables



- This figure shows a barplot that visualizes the distribution of car cylinder counts and number of gears.
- The x-axis represents the number of cylinders.
- The colour represents the number of gears.
- Y-axis is the number of cars belongs to the group.

Examining Multiple Variables

Box-and-whisker plot of mean household income and geographical region

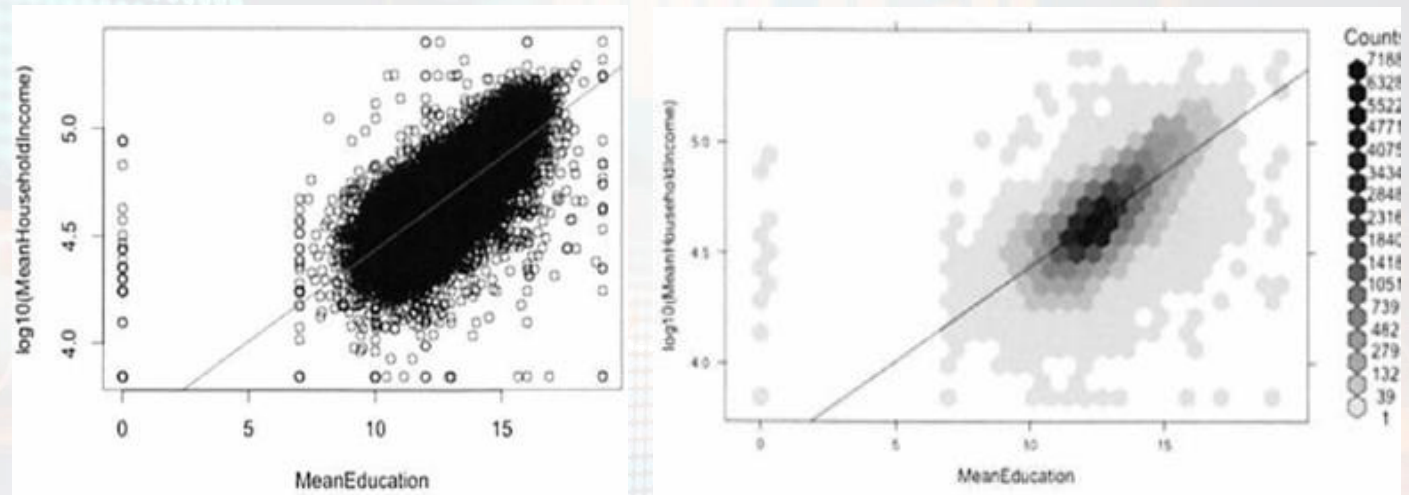


- The box-and-whisker plots shows the distribution of a continuous variable (i.e. logarithm of mean household income) for each of a discrete variable (i.e. the U.S. postal 'ZIP' code).
- The 'box' shows the range that contains the central 50% of the data, and the line inside the box is the location of the median value.
- The upper and lower hinges of the boxes correspond to the first and third quartiles of the data.
- The points outside the whiskers can be considered possible outliers.

Examining Multiple Variables

Scatterplot (left) and Hexbinplot (right) of household income against years of education

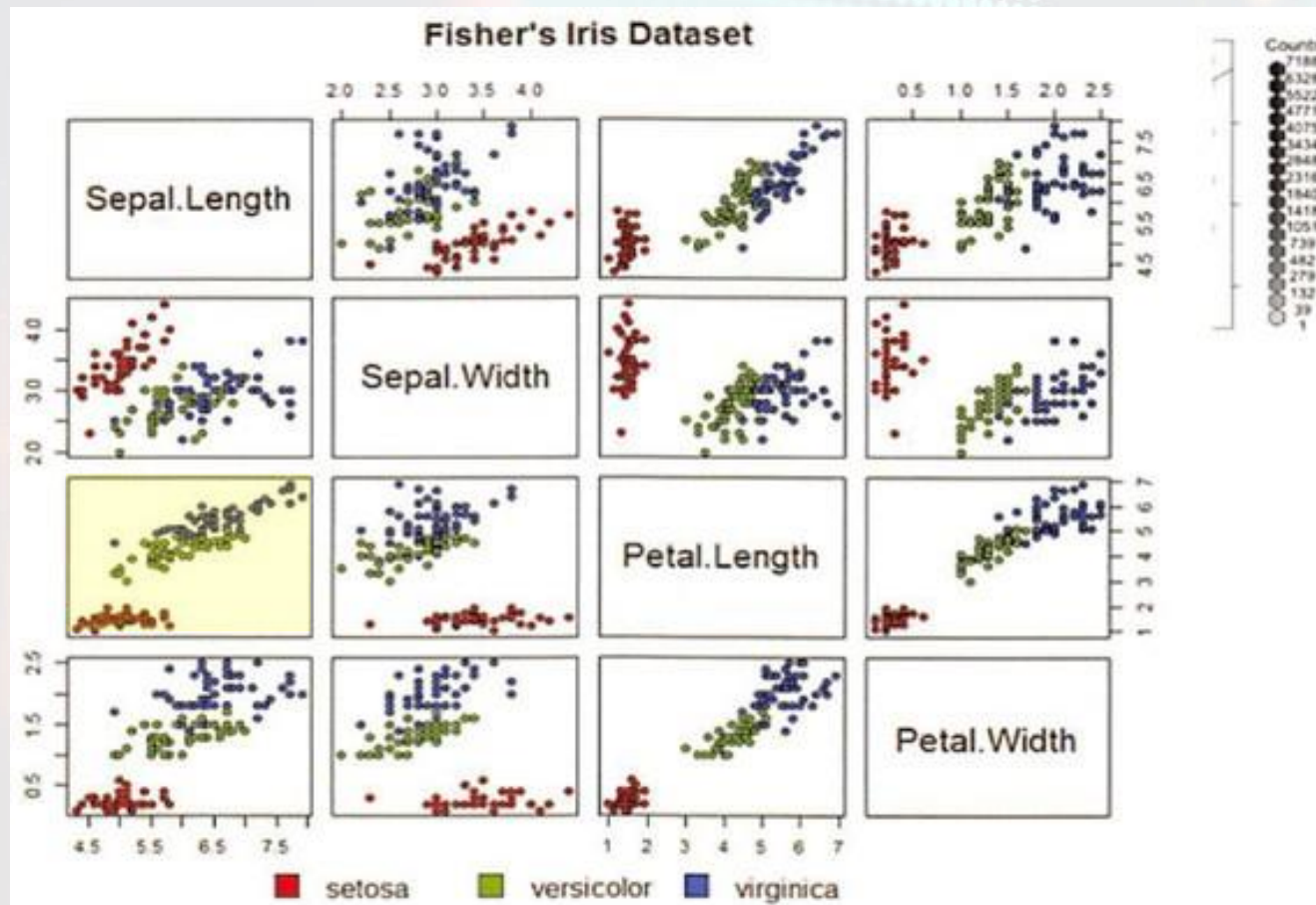
- The cluster in the scatter plot on the left suggests a **somewhat linear relationship** of the two variables.
- However, it is hard to see the structure of **how the data is distributed inside the cluster**. This is a Big Data type of problem.
- For high volume data, a **hexbinplot** may be better than scatterplot.



- A hexbinplot combines the ideas of scatterplot and histogram. Data is placed into hexbins, and the third dimension uses shading to represent the concentration of data in each hexbin.
- The hexbinplot shows that the biggest concentration is around 12 years of education, extending to about 15 years.

Examining Multiple Variables

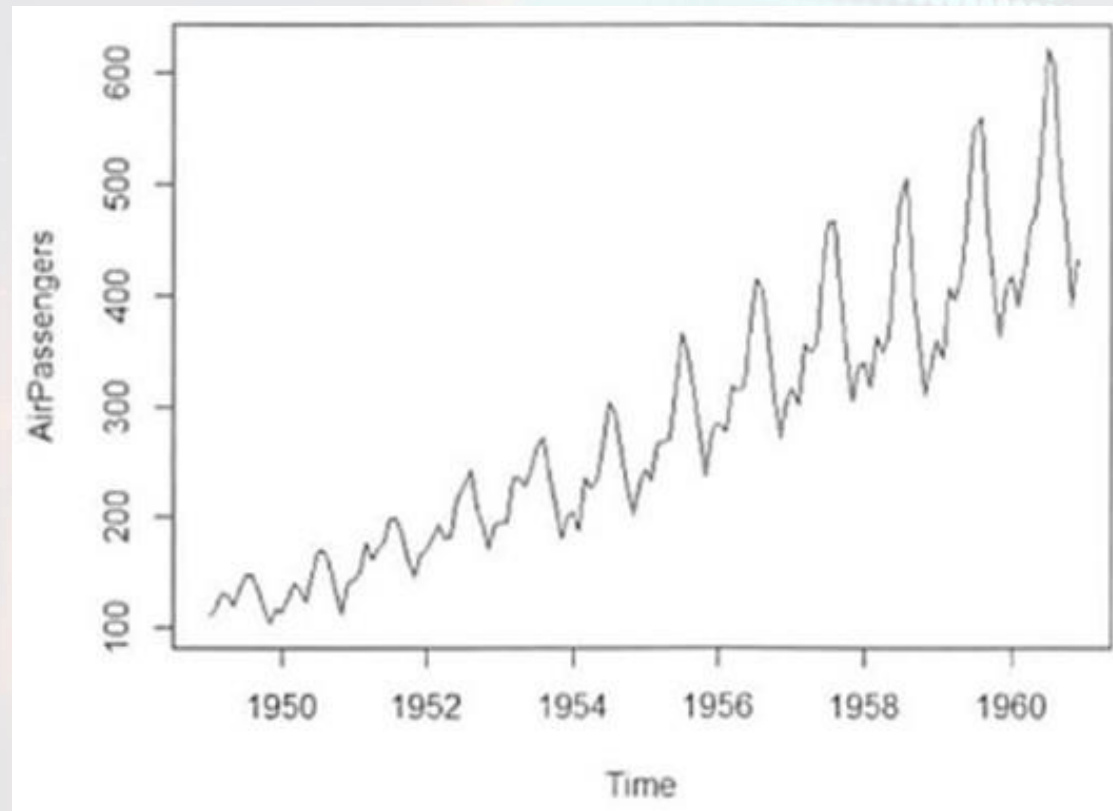
Scatterplot Matrix of Fisher's Iris dataset.



- All variables of Fisher's iris are compared in a scatterplot matrix.
- The three different colours represent three species of iris flowers.
- The first row (y-axis) and the third column (x-axis) compares sepal length against petal length.
- The scatterplot shows that *versicolor* and *virginica* share similar sepal and petal lengths (although the later has longer petals).
- The petal length of all *Setosa* are about the same, and the petal lengths are remarkably shorter than the other two species.

Examining Multiple Variables

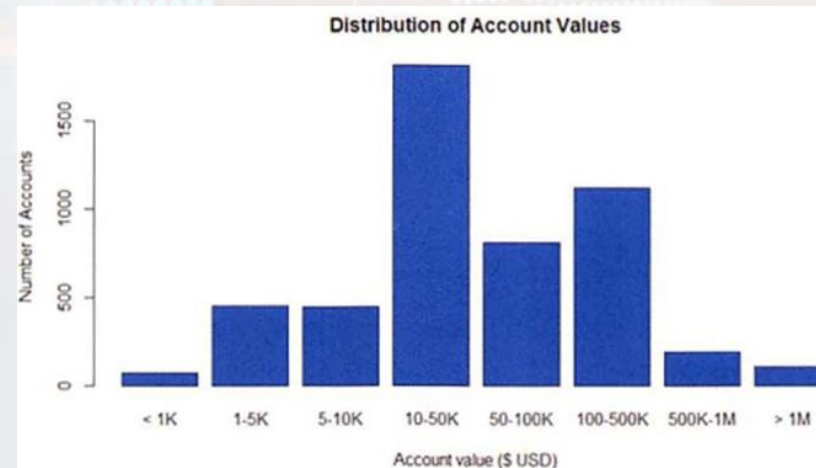
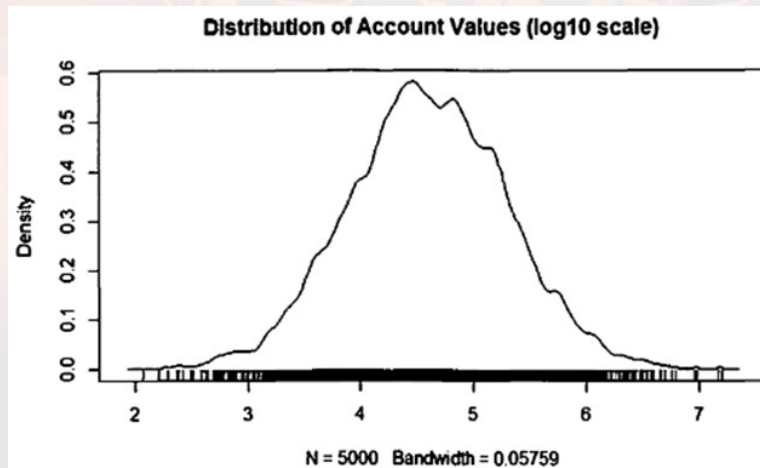
Airline passenger counts from 1949 to 1960



- The goal of visualizing a variable (in this case: airline passengers in thousands) **over time** is to identify **time-specific patterns**.
- The plot shows that, for each year, a **large** peak occurs **mid-year around July and August**, and a **small** peak happens around the **end of the year**, possibly due to the **holidays** – a phenomenon known as a **seasonality effect**.

Exploration vs. Presentation

- Data visualization for data **exploration** is **different** from **presenting** results to stakeholders.
 - Data scientists prefer graphs that are technical in nature.
 - Nontechnical stakeholders prefer simple, clear graphics that focus on the **message** rather than the data.



Density plot better for data scientists and histograms better to show to stakeholders.

Outline

- **Part A: Exploratory Data Analysis**
 - Descriptive statistics
 - Visualization before Analysis
 - Dirty Data
 - Visualizing a Single Variable
 - Examining Multiple Variables
 - Data Exploration versus Presentation
- **Part B: Statistical Methods for Evaluation**
 - Hypothesis Testing
 - Difference of Means
 - Wilcoxon Rank-Sum Test
 - Type I and Type II Errors
 - Power and Sample Size
 - ANOVA (Analysis of Variance)

Statistics can help answer the following data analytics questions:

Model Building

- What are the best input variables for the model?
- Can the model predict the outcome given the input?

Model Evaluation

- Is the model accurate?
- Does the model perform better than an obvious guess?
- Does the model perform better than other models?

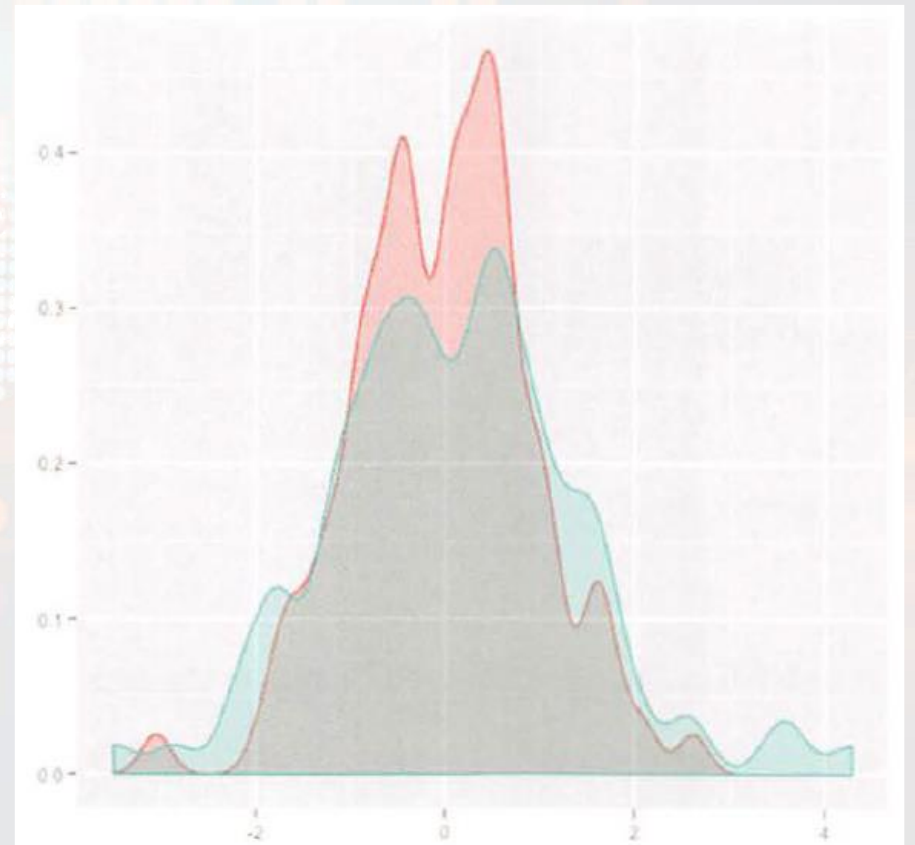
Model Deployment

- Is the prediction sound?
- Does the model have the desired effect (e.g. reducing cost)?

Hypothesis Testing

- A common techniques used to assess the difference of the **means** from two samples of data or the significance of the difference.
- The basic concept is to form an assertion (hypothesis) and test it with data.
- The common assumption is that **there is no difference between two samples**. Statisticians refer this as the **null hypothesis (H_0)**.
- The **alternative hypothesis (H_A or H_1)** is that there is a difference between two samples.

Distributions of two samples of data.



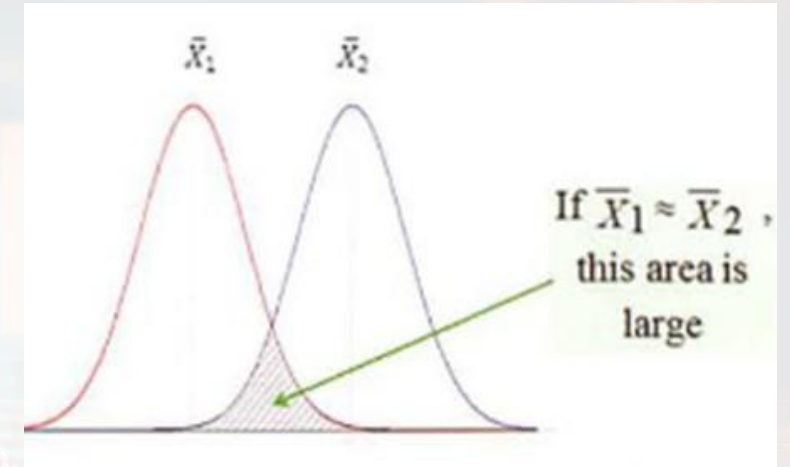
Hypothesis Testing

A hypothesis test leads to either rejecting the H_0 in favour of the H_A or not rejecting the H_0 .

Application	Null Hypothesis	Alternative Hypothesis
Accuracy Forecast	Model X <i>does not predict</i> better than the existing model.	Model X <i>predicts</i> better than the existing model.
Recommendation Engine	Algorithm Y <i>does not produce</i> better recommendations than the current algorithm being used.	Algorithm Y <i>produces</i> better recommendations than the current algorithm being used.
Regression Modeling	This variable <i>does not affect</i> the outcome because its coefficient is zero.	This variable <i>affects</i> outcome because its coefficient is not zero.

Difference of Means

- Hypothesis tests :
 - Is a common approach to draw inferences on whether the two populations (*pop1* and *pop2*) are different from each other.
 - compare the means of the respective populations based on samples randomly drawn from each population.
 - Consider the following H_0 and H_A :
 - $H_0: \mu_1 = \mu_2$
 - $H_A: \mu_1 \neq \mu_2$where \bar{X}_1 and \bar{X}_2 denotes the population means of *pop1* and *pop2*.
- The basic testing approach is to compare the observed sample means, $\bar{X}_1 = \bar{X}_2$, corresponding to each population.



- If $\bar{X}_1 = \bar{X}_2$, the distributions overlap substantially.
 - H_0 is accepted and H_A is rejected.
- If $\bar{X}_1 \neq \bar{X}_2$, a large difference between the sample means indicates that the H_0 should be rejected.
 - H_0 is rejected and H_A is accepted.

Difference of Means

Two Parametric Methods to test the difference in means:

- **Student's T-Test**

- Assumes two normally distributed populations have equal but unknown variance.

- **Welch's T-test**

- Assumes two normally distributed populations have unequal variance.

Difference of Means

Student's T-Test

- Student's t-test assumes that **distributions** of the two populations have **equal but unknown variances**.
- Suppose n_1 and n_2 samples are randomly and independently selected from two populations, pop1 and pop2 , respectively.
- If each population is normally distributed with **the same mean** ($\mu_1 = \mu_2$) and with **the same variance**, then T (the t-statistic), follows a t-distribution with $n_1 + n_2 - 2$ **degrees of freedom (df)**.

$$T = \frac{\bar{x}_1 - \bar{x}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\text{where } S_p = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

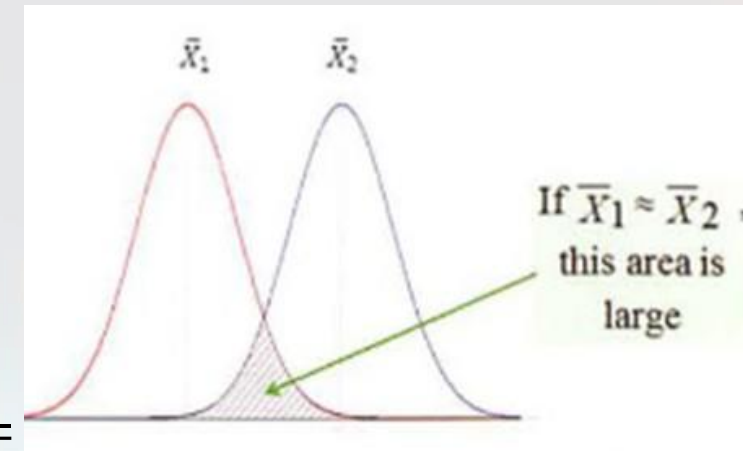
S_p is pooled variance

Significance level, $\alpha = 0.05$

Degree of freedom, $df = n_1 + n_2 - 2$

T^* is critical value found using df (from table)

**If $T \geq T^*$
the null hypothesis is rejected**



Difference of Means

Welch's T-test

- Also known as **unequal variances t-test**
- When the equal population variance assumption is **not justified** in performing Student's t-test for the difference of means, Welch's t-test can be used based on T expressed in:

$$T_{welch} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

where \bar{X}_i , S_i^2 , and n_i correspond to the i-th sample mean, sample variance, and sample size.

- Notice that **Welch's t-test uses the sample variance (s^2) for each population instead of the pooled sample variance.**

- The degree of freedom for Welch's t-test is calculated using:

$$df = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Difference of Means

Example (Student's T-test)

Some brown hairs were found on the clothing of a victim at a crime scene. The five of the hairs were measured: 46, 57, 54, 51, 38 μm .

A suspect is the owner of a shop with similar brown hairs. A sample of the hairs has been taken and their widths measured: 31, 35, 50, 35, 36 μm .

Is it possible that the hairs found on the victim were left by the suspects? Test at the 5% level.

Calculation Steps:

1. Calculate the mean and standard deviation for the data sets.

	From Crime Scene	From a suspect
	46	31
	57	35
	54	50
	51	35
	38	36
Total	246	187
Mean	49.2	37.4
Standard deviation	7.463	7.301

2. Calculate the magnitude of the difference between the two means.

$$49.2 - 37.4 = 11.8$$

Difference of Means

Calculation Steps:

3. Calculate the standard error in the difference.

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{7.463^2}{5} + \frac{7.301^2}{5}} = 4.669 \approx 4.67 \text{ (3 s.f.)}$$

4. Calculate the value of T.

$$\begin{aligned} T &= \text{difference between the means} \div \text{standard error in the difference} \\ &= 11.8 \div 4.669 = 2.527 \approx 2.53 \text{ (3 s.f.)} \end{aligned}$$

5. Calculate the degrees of freedom.

$$df = n_1 + n_2 - 2 = 5 + 5 - 2 = 8$$

Difference of Means

Calculation Steps:

6. Find the critical value T^* for the particular significance you are working to from the table.

df	PROPORTION IN ONE TAIL			
	0.25	0.10	0.05	0.025
df	PROPORTION IN TWO TAILS COMBINED			
	0.50	0.20	0.10	0.05
1	1.000	3.078	6.314	12.706
2	0.816	1.886	2.920	4.303
3	0.765	1.638	2.353	3.182
4	0.741	1.533	2.132	2.776
5	0.727	1.476	2.015	2.571
6	0.718	1.440	1.943	2.447
7	0.711	1.415	1.895	2.365
8	0.706	1.397	1.860	2.306
9	0.703	1.383	1.833	2.262

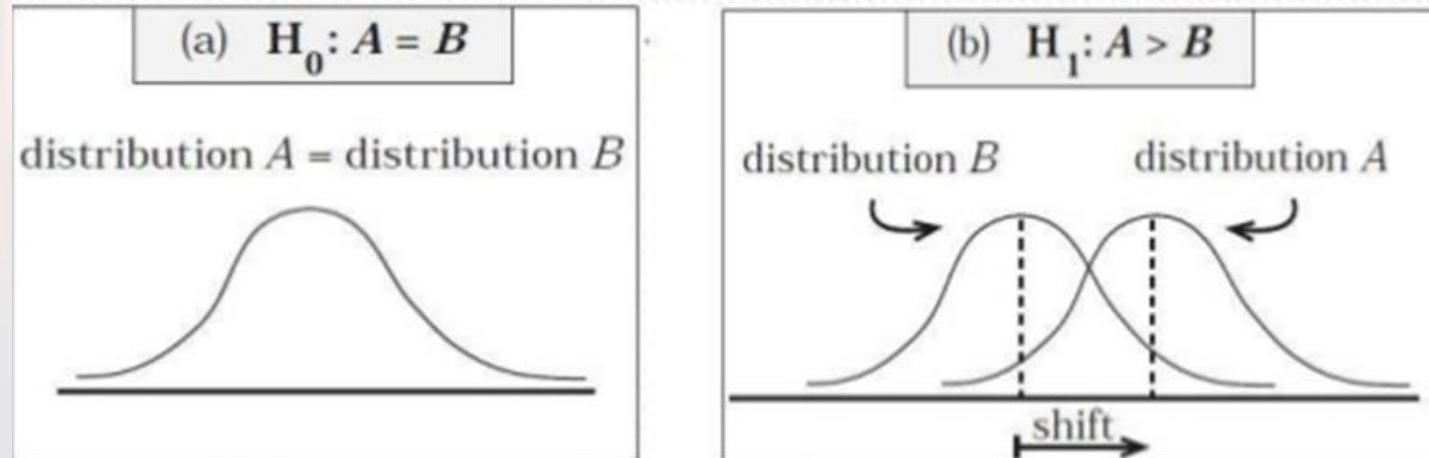
If $T < T^*$ (critical value) then there is no significant difference between the two sets of data ,i.e. null hypothesis is Accepted.

If $T \geq T^*$ (critical value) then there is a significant difference between the two sets of data i.e. null hypothesis is Rejected.

2.53 > 2.306, H_A is accepted, H_0 is reject.

Wilcoxon Rank-Sum Test

- Generally used when normality assumption for the sample does not hold and **sample size is small**.
- Non-parametric statistical hypothesis test for assessing whether one of two samples of independent observations tends to have larger values than the other.



Wilcoxon Rank-Sum Test

Steps:

1. Select a **random sample** from each of the populations.
2. Let n_1 and n_2 be the number of observations in the smaller and larger sample, respectively.
3. Arrange the combined $n_1 + n_2$ observations in **ascending order** and substitute a rank of 1, 2... to the $n_1 + n_2$ observations.
4. In the case of ties, we assign the conflicting observations with their mean ranks.
5. Our decision is based on the value of the test statistics, U (the random variable for u).

- For one-tail test: u_1 or u_2
- For two-tail test: $u = \min(u_1, u_2)$

H_0	H_1	Compute
$\mu_1 = \mu_2$	$\mu_1 < \mu_2$	u_1
	$\mu_1 > \mu_2$	u_2
	$\mu_1 \neq \mu_2$	u

6. Null hypothesis will be rejected whenever the appropriate statistic U_1 , U_2 or U assumes a value less than or equal to the desired critical value.
 - Test Statistic \leq Critical Value

Wilcoxon Rank-Sum Test

Example (Two-tail):

The nicotine content of two brands of cigarette, measured in mg, was found to be as given the table. Test the hypothesis, at 0.05 level of significance, that the median nicotine content of two brands are equal against the alternative that they are unequal.

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

$$\alpha = 0.05$$

$$n_1 = 8$$

$$n_2 = 10$$

Critical Region: $\mu \leq 17$ (From Table)

Brand A	2.1	4.0	6.3	5.4	4.8	3.7	6.1	3.3		
Brand B	4.1	0.6	3.1	2.5	4.0	6.2	1.6	2.2	1.9	5.4

Table A.18 (continued) Critical Values for the Wilcoxon Rank-Sum Test

One-Tailed Test at $\alpha = 0.025$ or Two-Tailed Test at $\alpha = 0.05$

n_1	n_2																
	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																	
2					0	0	0	0	1	1	1	1	1	2	2	2	2
3																	
4	I)	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
5		1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	13
6		2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
7			5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
8				8	10	12	14	16	18	20	22	24	26	28	30	32	34
9					13	15	17	19	22	24	26	29	31	34	36	38	41
10						17	20	23	26	28	31	34	37	39	42	45	48
11							23	26	29	33	36	39	42	45	48	52	55
12								30	33	37	40	44	47	51	55	58	62

Wilcoxon Rank-Sum Test

Example (Two-tail):

- Computation Steps:
 - Arranging observations in ascending order and assigning ranks from 1 to 18.
 - $w_1 = 4+8+9+10.5+13+14.5+16+18 = 93$
 - $w_2 = 1+2+3+5+6+7+10.5+12+14.5+17 = 78$
 - Therefore,
 - $u_1 = 93 - ((8*9)/2) = 57$
 - $u_2 = 78 - ((10*11)/2) = 23$
 - $\text{Min}(u_1, u_2) = 23$ (not ≤ 17)
- **Decision:** Don't reject H_0 and conclude that there is no significant difference in median nicotine contents of two brands of cigarettes at 0.05 significance level.

DATA	RANKS	BRAND
0.6	1	B
1.6	2	B
1.9	3	B
2.1	4	A
2.2	5	B
2.5	6	B
3.1	7	B
3.3	8	A
3.7	9	A
4.0	10.5	A
4.0	10.5	B
4.1	12	B
4.8	13	A
5.4	14.5	A
5.4	14.5	B
6.1	16	A
6.2	17	B
6.3	18	A

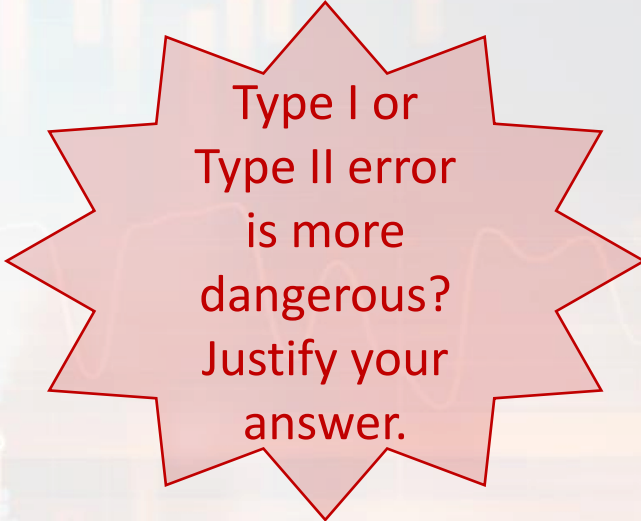
Type I and Type II Errors

- A hypothesis test may result in two types of errors
 - **Type I Error:** Rejection of the null hypothesis when the null hypothesis is TRUE
 - **Type II Error:** Acceptance of the null hypothesis when the null hypothesis is FALSE

	H_0 is true	H_0 is false
H_0 is accepted	Correct outcome	Type II Error
H_0 is rejected	Type I error	Correct outcome

Type I and Type II Errors

- The significant level is equivalent to the Type I Error.
- For a significance level such as $\alpha = 0.05$, if the H_0 is TRUE, there is a 5% chance that the observed T value based on the sample data will be large enough to reject the H_0 .
- By **selecting an appropriate significance level**, the probability of committing a Type I error can be defined before any data is collected and analyzed.
- To **reduce** the probability of a **Type II** error to a reasonable level, it is often **necessary to increase the sample size**.



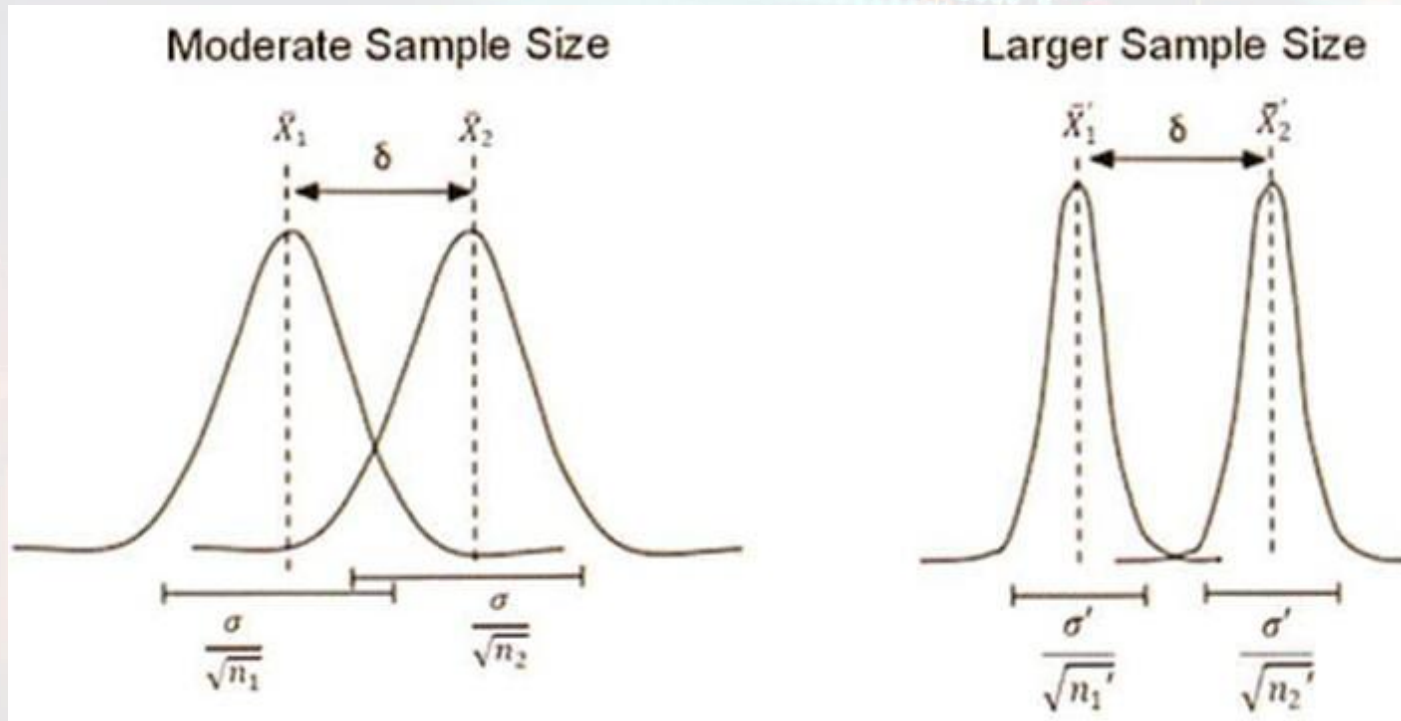
Type I or
Type II error
is more
dangerous?
Justify your
answer.

Power and Sample Size

- The *power of a test* is the ***probability of correctly rejecting the null hypothesis***.
- It is denoted by $1 - \beta$, where β is the probability of a ***Type II error***.
- The power of a test **improves** as the sample size **increases**. It is **used to determine the necessary sample size**.
- In the difference of means, the power of a hypothesis test depends on the true difference of the population means.
 - In other words, for a fixed significance level, a larger sample size is required to detect a smaller difference in the means.
- In general, ***Effect size δ*** = the magnitude of the difference between the means.
 - As the sample size becomes larger, it is easier to detect a given effect size.

Power and Sample Size

A larger sample size better identifies a fixed effect size.



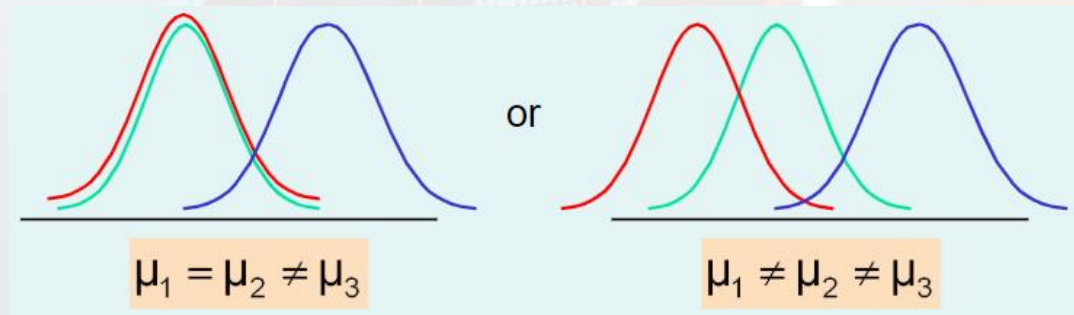
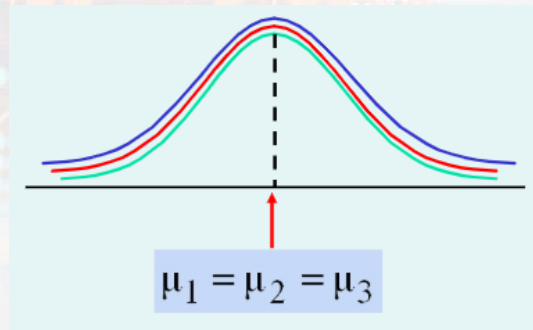
- Large sample size: any effect size can appear statistically significant.
- Very small sample size: maybe useless in practical sense.
- It is important to consider an appropriate effect size for the problem at hand.

Analysis of Variance (ANOVA)

- Consider an example of testing the impact of nutrition and exercise on 60 candidates between age 18 and 50.
- The candidates are randomly split into **six groups** to determine which of the following strategy is the most effective:
 - Group 1 only eats junk foods.
 - Group 2 only eats healthy foods.
 - Group 3 eats junk food and does cardio exercise every other day.
 - Group 4 eats healthy food and does cardio exercise every other day.
 - Group 5 eats junk food and does both cardio and strength training every other day.
 - Group 6 eats healthy food and does both cardio and strength training every other day.
- Multiple t-tests could be applied to each pair of weight loss strategies (e.g. Group 1 vs. Group 2., 3, 4, 5, or 6). **A total of 15 t-tests would need to be performed!**

Analysis of Variance (ANOVA)

- ANOVA is designed to test whether the means of **more than 2** quantitative populations are equal.
- The H_0 of ANOVA is that all the **population means are equal** and the H_A is that **at least one pair** of the population mean is not equal.
 - $H_0 : \mu_1 = \mu_2 = \dots = \mu_n$
 - $H_1 : \mu_i \neq \mu_j$ for at least one pair of i, j .



Analysis of Variance (ANOVA)

$$F_{obt} = \frac{s^2_B}{s^2_W}$$

where $s^2_B = \frac{\sum n(\bar{x} - \bar{x}_G)^2}{k-1}$ and $s^2_W = \frac{\sum SS}{N-k}$

Key Steps:

Find the mean for each of the groups.

Find the overall mean (the mean of the groups combined).

Find the **Within Group Variation**: The total deviation of each member's score from the Group Mean.

Find the **Between Group Variation**: The deviation of each Group Mean from the Overall Mean.

Find the **F critical** and **F statistic**: the ratio of Between Group Variation to Within Group Variation.

F statistic < F critical accept H_0 else reject H_0 and accept H_A .

Examples: <https://medium.com/@hussein.sajid7/understanding-analysis-of-variance-anova-6aebd01d44c8>

End of Lecture

- D. Lucy. (2005). *Introduction to Statistics for Forensic Scientists* Chichester: Wiley.
- EMC Education Services. (2015). *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. Wiley.
- H. Sajid. (2019). Understanding Analysis of Variance: ANOVA. Available at: <https://medium.com/@hussein.sajid7/understanding-analysis-of-variance-anova-6aebd01d44c8>