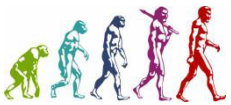
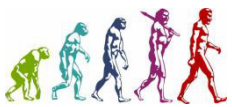
Exploratory Data Analysis and Visualization



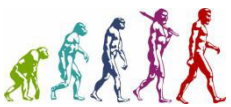
Exploratory Data Analysis

- The purpose of exploratory data analysis (EDA) is to convert the available data from their raw form to an informative one, in which the main features of the data are illuminated
- We display and summarize data that are obtained from a sample.
- By means of this visualization and summaries, we try to explore the information hidden in the data
- These summaries are applied only to the data at hand; we do not attempt to make claims about the larger population from which the data is obtained



Exploratory Data Analysis

- When performing EDA, we should always:
  - use visual displays (graphs or tables) plus numerical summaries
  - describe the overall pattern and mention any striking deviations from that pattern
  - interpret the results we got in context
- While exploring the Data, we should explore:
  - Each variable individually (all Xs and Y)
  - Each X with Y in pair
  - Xs in pair
  - All Xs and Y together

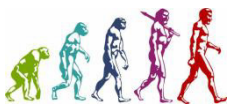


Data Visualization

Data visualization is a general term that describes any effort to help people understand the significance of data by placing it in a visual context. Patterns, trends and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization

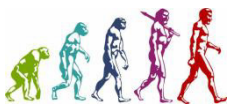
Why Data Visualization?

- To describe, analyse and summarize data in visual form
- To understand Data Properties
- To find patterns in data
- To suggest modelling strategies
- To “debug” analyses
- To communicate results



Univariate Analysis

- When examining the distribution of a single variable, it is important to distinguish on type of variable i.e. between a categorical variable and a quantitative variable
  - Variables are visualized and summarized differently based on type of variable

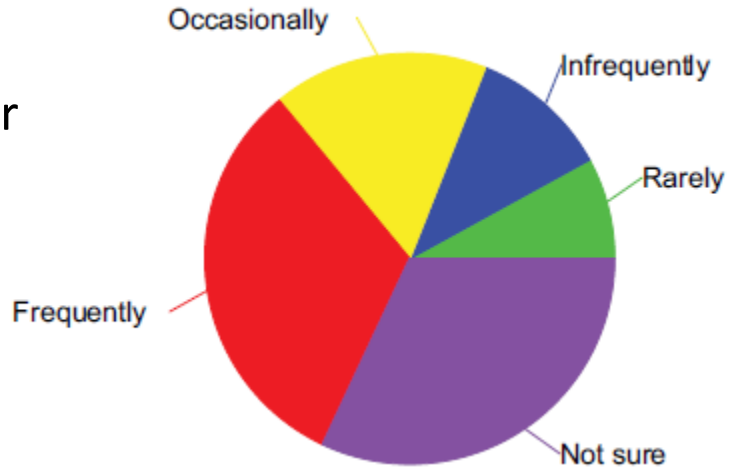


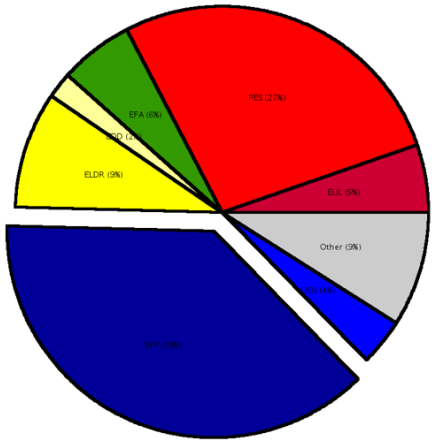
Univariate Analysis – Categorical Variable

- The distribution of a categorical variable is summarized using:
  - Display: pie-chart or bar-chart
  - Numerical summaries: category (group) count, percentages and proportions

Pie Chart

- A pie chart is circular and uses pie wedges to represent the proportion of the whole for some category
- It gives a good idea of proportion of the constituents
- Pie is a divided bar in polar coordinate
- A pie chart is usually better than bar chart for judging the proportion of the whole
- Exploded pie chart separate one or more sections to highlight a sector or smaller segments
- It is difficult to numerically compare different sections of pie chart or to compare data across different pie charts
- Not suitable if number of sections are large

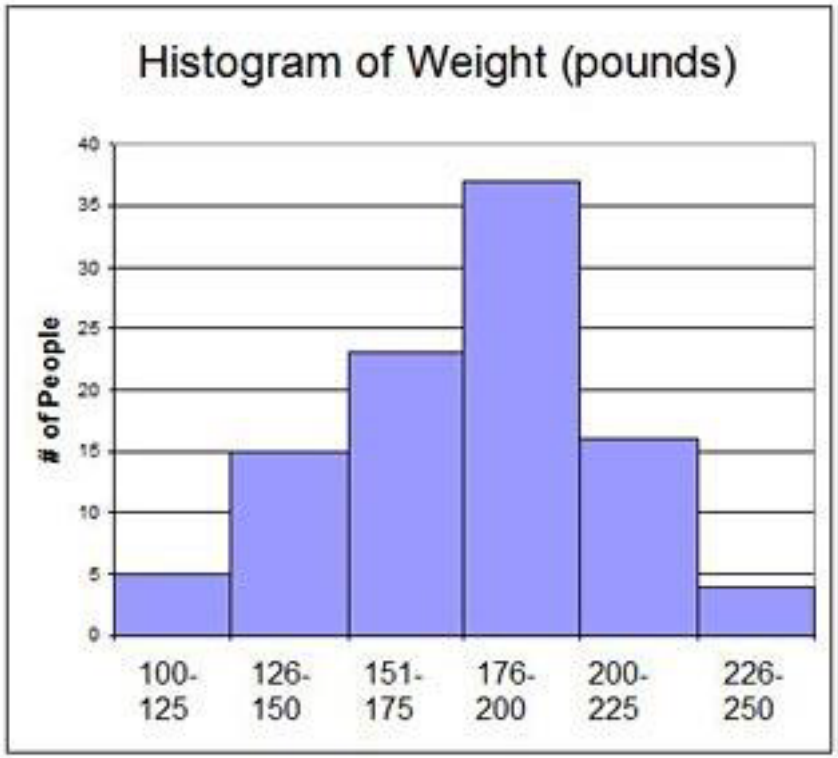


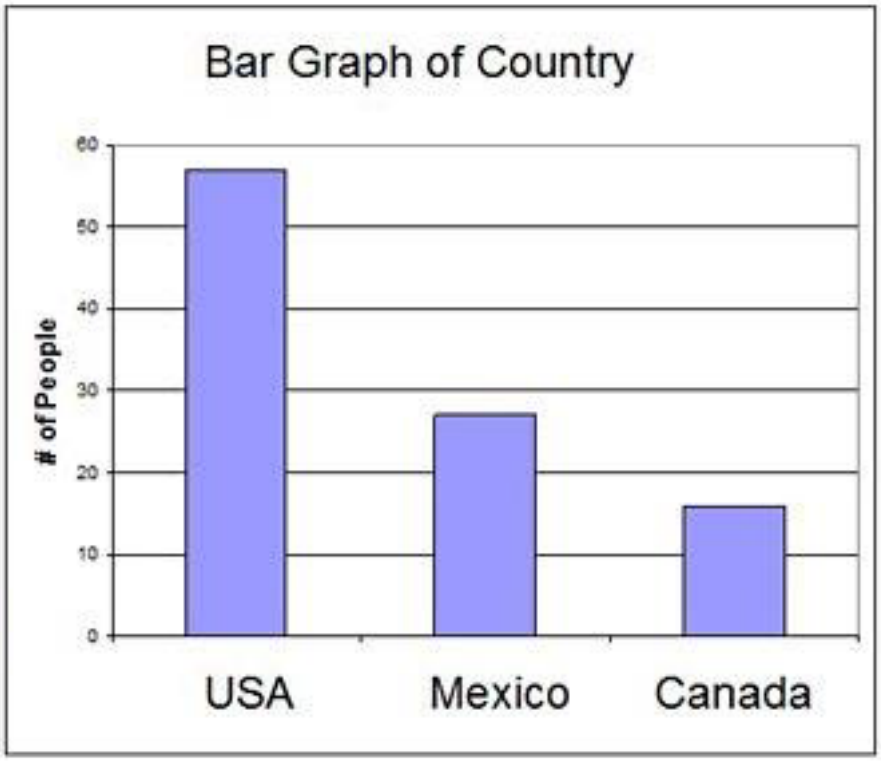


Univariate Analysis – Quantitative Variable

- The distribution of a quantitative variable is summarized using:
  - Display: histogram (or stemplot, mainly for small data sets). When describing the distribution as displayed by the histogram, we should describe the:
    - Overall pattern → shape, centre, spread
    - Deviations from the pattern → outliers

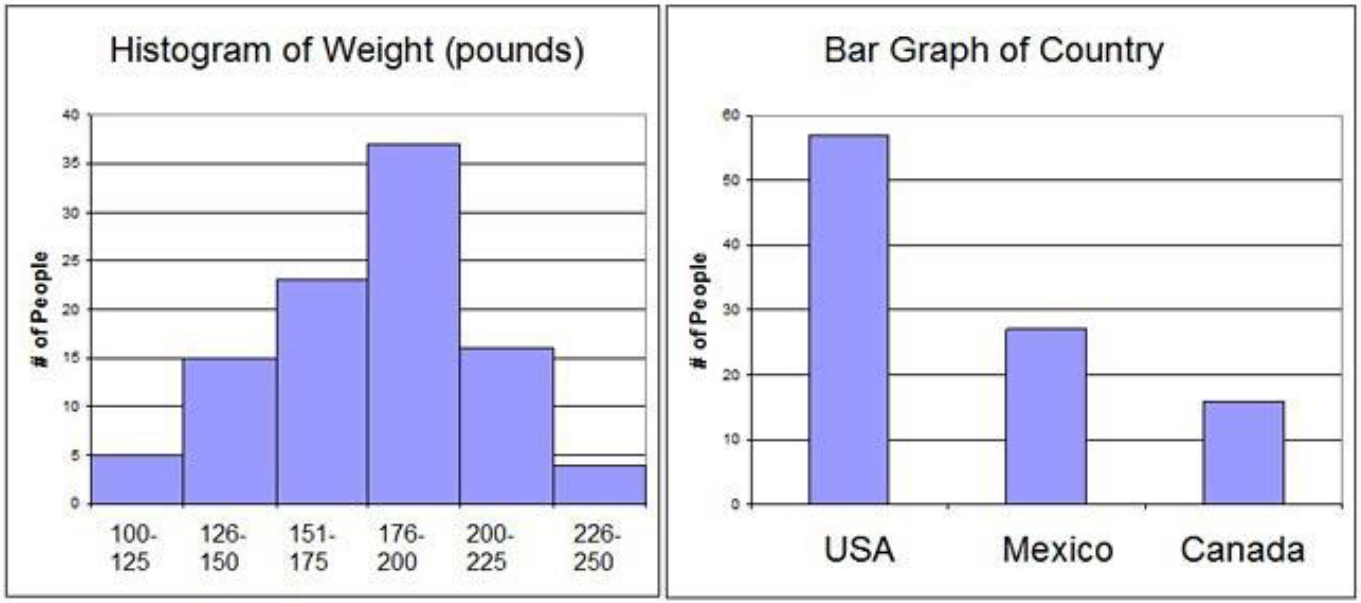
Histogram & Bar Graph/Chart





- Histogram gives the distribution of weight of adults taken randomly from three different countries. Bar chart compares no of people for the three countries
- Histograms show distributions of variables while bar charts compare variables.

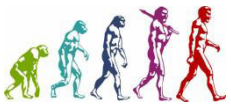
Histogram & Bar Graph/Chart



- Histograms plot quantitative data with ranges of the data grouped into bins or class intervals while bar charts plot categorical data
- Bars can be reordered in bar charts but not in histogram
- No spaces between the bars of histogram but there are gaps in Bar Chart bars
- The bars of bar charts typically have the same width. The widths of the bars in a histogram need not be the same. Values in bar charts are given by the length of the bar while values in histograms are given by areas

Histogram

- Histogram is a graph that displays the distribution of data
- Histogram is characterized by three constituents
  - A center (mean)
  - A width (spread)
  - An over all shape
- Application of Histogram
  - Shape and Smoothness
  - Comparison to Specification limits
  - Comparison to Sources of Variability
  - Outlier Detection: points lying outside 3 may be considered outlier
  - Before and After Comparison
- The original data is not preserved



Stem Leaf Plot (or stem plot)

- Graphical method of displaying data which clarify shape of the distribution while preserving the original data
- Particularly useful when data is not too numerous
- Two portions: Left is stem and right is leaves
- Stem consists of one or more leading digits and leaves contain remaining digits. Many times last digit is treated as leaves and rest as stem
- Leaves are generally arranged by magnitude (sorted)
- When rotated by 90^{o} counterclockwise, it resembles a histogram
- If the data includes numbers with three or more digits or decimals, they may also be rounded to two digit accuracy

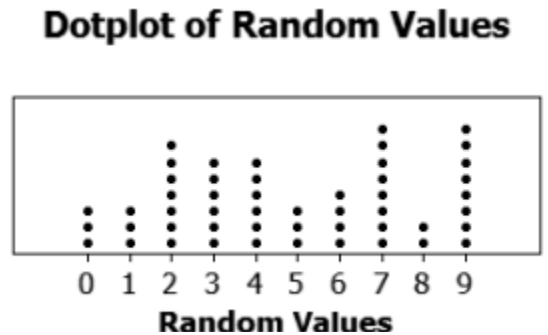
```
37, 33, 33, 32, 29, 28, 28, 23, 22,
22, 22, 21, 21, 21, 20, 20, 19, 19,
18, 18, 18, 18, 16, 15, 14, 14, 14,
12, 12, 9, 6
```

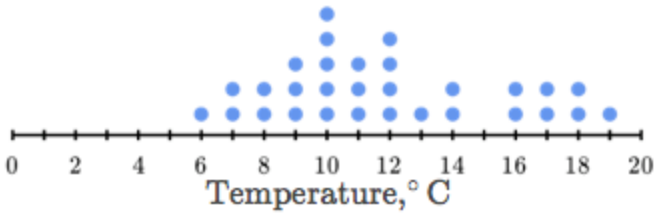
```
3|2337
2|001112223889
1|2244456888899
0|69
```

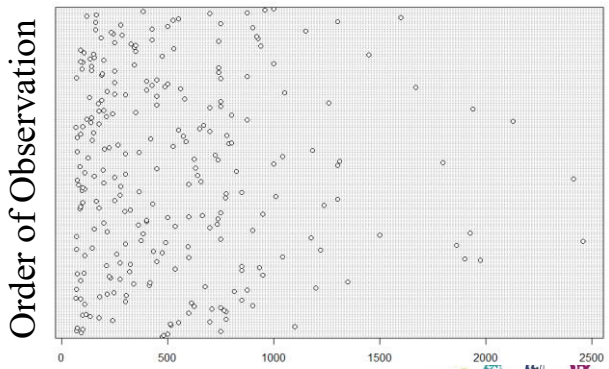
```
3|7
3|233
2|889
2|001112223
1|56888899
1|22444
0|69
```

Dot Plot

- A dot plot (also dot chart), is a type of simple histogram-like chart for relatively small univariate data sets where values fall into a number of discrete bins.
- Data points are plotted on fairly simple scale
- To draw a dot plot, count the number of data points falling in each bin
- The dots are placed one above the other so that the height of the column of dots represents the frequency of that value
- Like stem plot but displays dots instead of individual values
- Useful for highlighting clusters and gaps, skews as well as outliers
- Sometimes values are plotted in the order they occur
- For data sets larger than 50 points, it is better to use other chart types like histogram as Dot plot becomes too cluttered

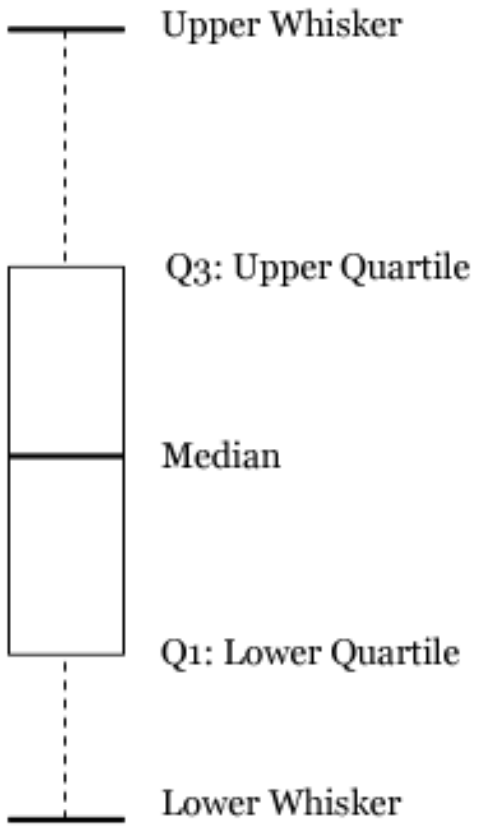




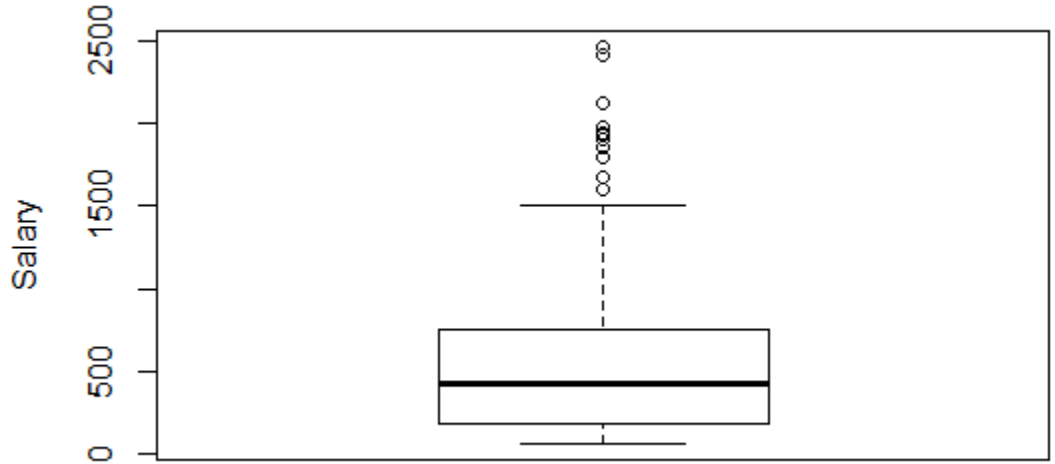


Box Plot

- Box Plot provides another visual summary of a distribution
- The plot is also called a Box and Whisker plot.
- The plot consists of a box with two ends as the 3rd and 1st quartile respectively. The median is shown as a line inside the box. Two whiskers are drawn from the 3rd and 1st quartiles to the maximum and minimum values respectively.
- Thus in one figure the box plot displays the central tendency, the dispersion as well as the skew
- Box plot is not only a great way to present summary description of a random variable. It is an excellent tool to compare a number of distributions as well



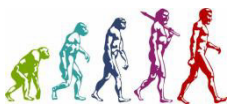
Box Plot – IQR and Finding Outliers



- While the range quantifies the variability by looking at the range covered by ALL the data, the IQR measures the variability of a distribution by giving us the range covered by the MIDDLE 50% of the data
- Inter Quartile Range (IQR) is thus given by (Q3 – Q1)
- An observation is considered a suspected outlier if it is
  - Below Q1 – 1.5 (IQR) or
  - Above Q3 + 1.5 (IQR)
- If max or min point lie beyond these points, then the whiskers of box plot is extended only upto these points and all points lying beyond them are shown as outliers

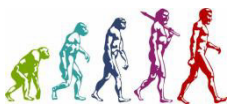
Univariate Analysis – Quantitative Variable

- Numerical summaries: descriptive statistics (measure of centre plus measure of spread):
  - If distribution is symmetric with no outliers, use mean and standard deviation
  - Otherwise, use the five-number summary, in particular, median and IQR (inter-quartile range)
- Five number summaries are:
  - Minimum
  - Q1
  - Median
  - Q3
  - Maximum
- Inter Quartile Range is Q3 – Q1



Quantitative Variables - Numerical Summaries

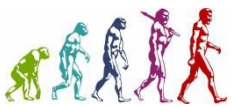
- The five-number summary and the 1.5(IQR) Criterion for detecting outliers are the ingredients we need to build the boxplot.
- Boxplots are most effective when used side-by-side for comparing distributions



Bivariate Analysis

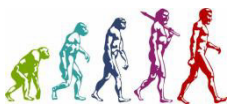
- When examining the relationship between two variables, the first step is to classify the two relevant variables according to their role and type:

- and only then to determine the appropriate tools for summarizing the data



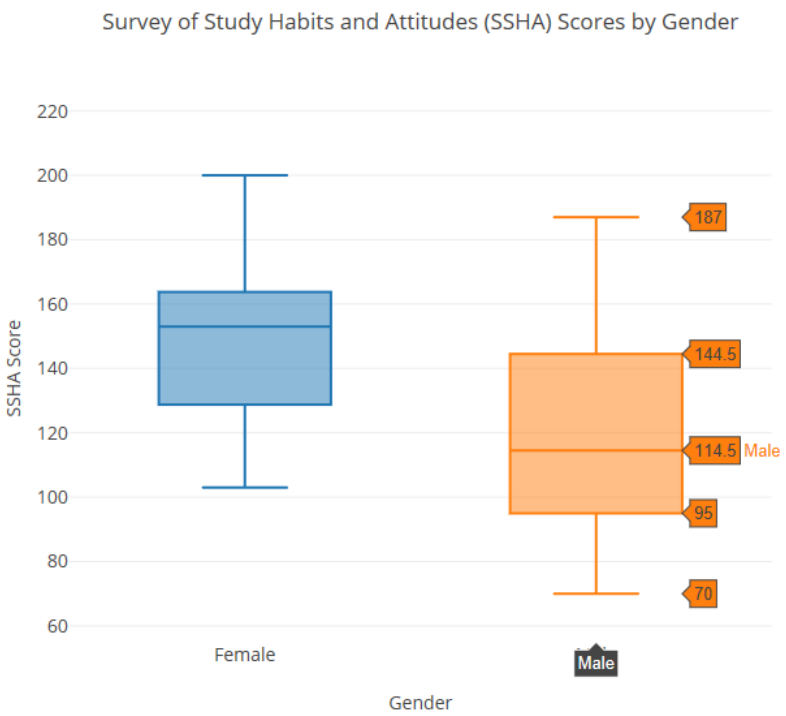
Bivariate Analysis: C -> Q

- Exploring the relationship amounts to comparing the distributions of the quantitative response variable for each category of the explanatory variable. To do this, we use:
  - Display: side-by-side boxplots
  - Numerical summaries: descriptive statistics of the response variable, for each value (category) of the explanatory variable separately



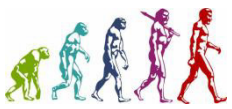
Side-by-Side Boxplot

- Graphically show relationship between two variables Categorical vs Quantitative
- We essentially compare the distributions of the quantitative response for each category of the explanatory variable



Bivariate Analysis: C -> C

- Exploring the relationship amounts to comparing the distributions of the categorical response variable, for each category of the explanatory variable. To do this, we use:
  - Display: two-way table
  - Numerical summaries: conditional percentages (of the response variable for each value (category) of the explanatory variable separately)



Bivariate Analysis: C-->C

- Exploring the relationship between two categorical variables amounts to comparing the distributions of the response variable across the different values of the explanatory variable
- We need to supplement our display, the contingency table (or two-way table), with some numerical summaries that will allow us to compare the distributions. These numerical summaries are found by simply converting the counts to percentages within (or restricted to) each value of the explanatory variable separately.

Explanatory Variable: Gender
Response Variable: Maths Difficulty Level

Compare these Distributions

Gender	Maths Difficulty Level			
	Average	Hard	Easy	Total
Female	560	163	37	760
Male	295	72	73	440
Total	855	235	110	1200

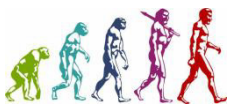
Two Way Table

Gender	Maths Difficulty Level			
	Average	Hard	Easy	Total
Female	560/760 = 73.7 %	163/760 = 21.5 %	37/760 = 4.9 %	760/760 = 100 %
Male	295/440 = 67 %	72/440 = 16.4 %	73/440 = 16.6 %	440/440 = 100 %

Conditional Percentages

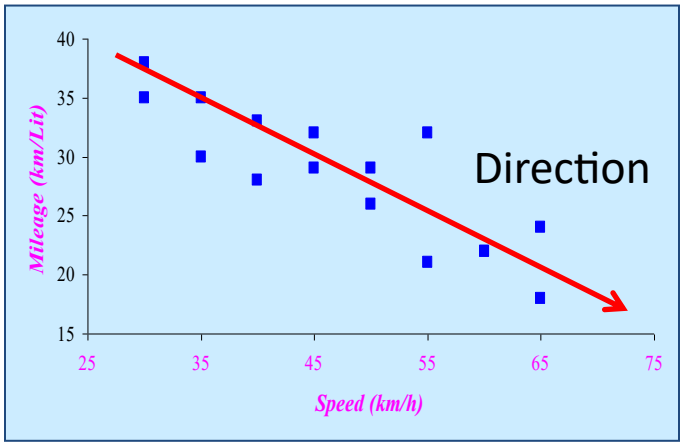
Bivariate Analysis: Q -> Q

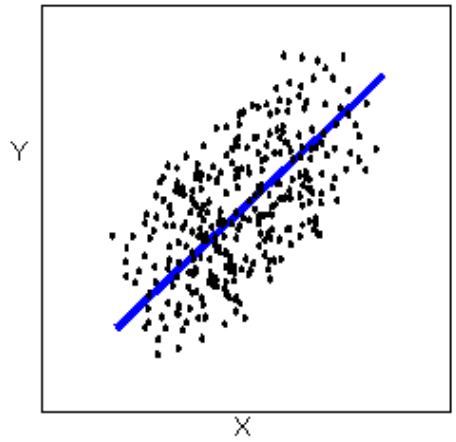
- We examine the relationship using:
  - Display: scatterplot. When describing the relationship as displayed by the scatterplot, be sure to consider:
    - Overall pattern → direction, form, strength.
    - Deviations from the pattern → outliers.
- Labelling the scatterplot (including a relevant third categorical variable in our analysis) called Stratified scatter plot, might add some insight into the nature of the relationship.
- Matrix plot is a matrix of scatterplots and is used to check interesting scatterplots



Scatter Plot

- A scatter plot is a chart that graphically depicts the relationship between two quantitative data types X and Y. When describing the overall pattern of the relationship we look at its direction, form and strength

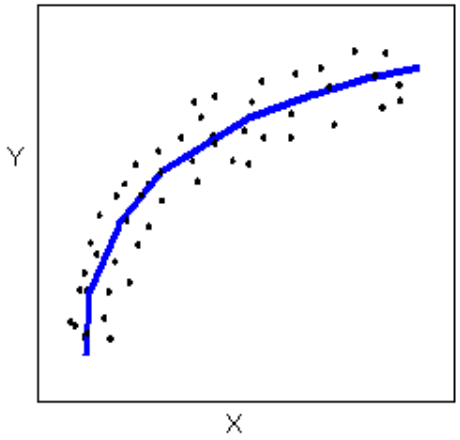




← Weak

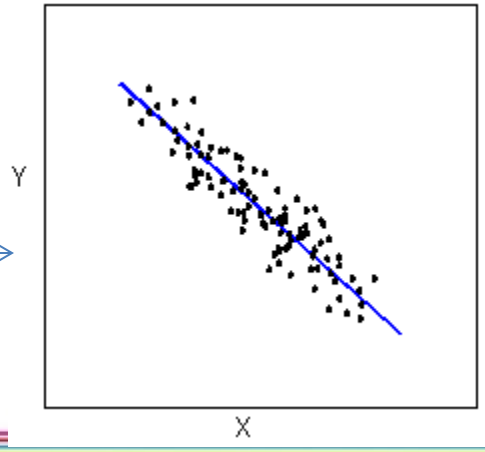
Strength

Form



Linear →

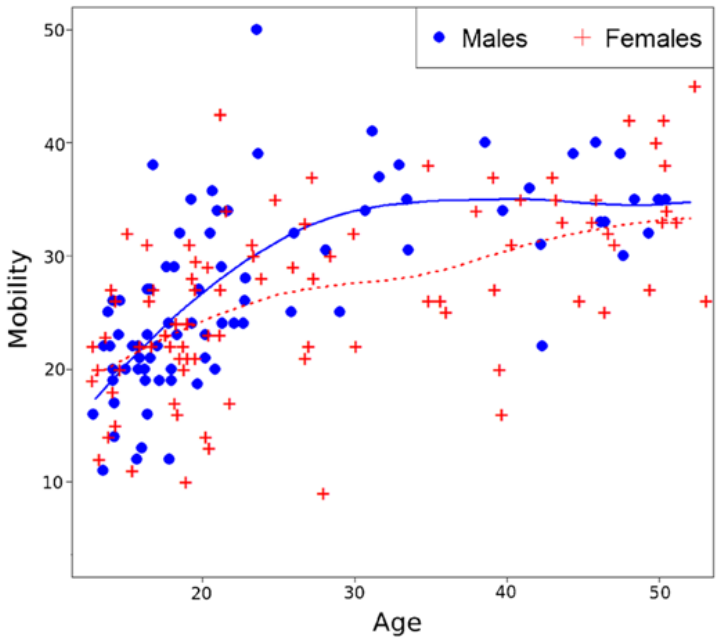
← Curvilinear

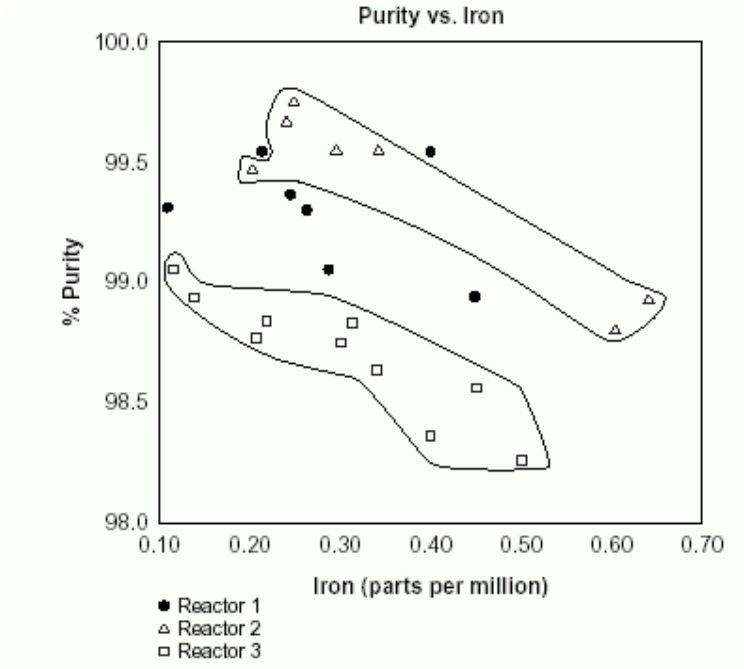


← Strong

Stratified Scatter Plot

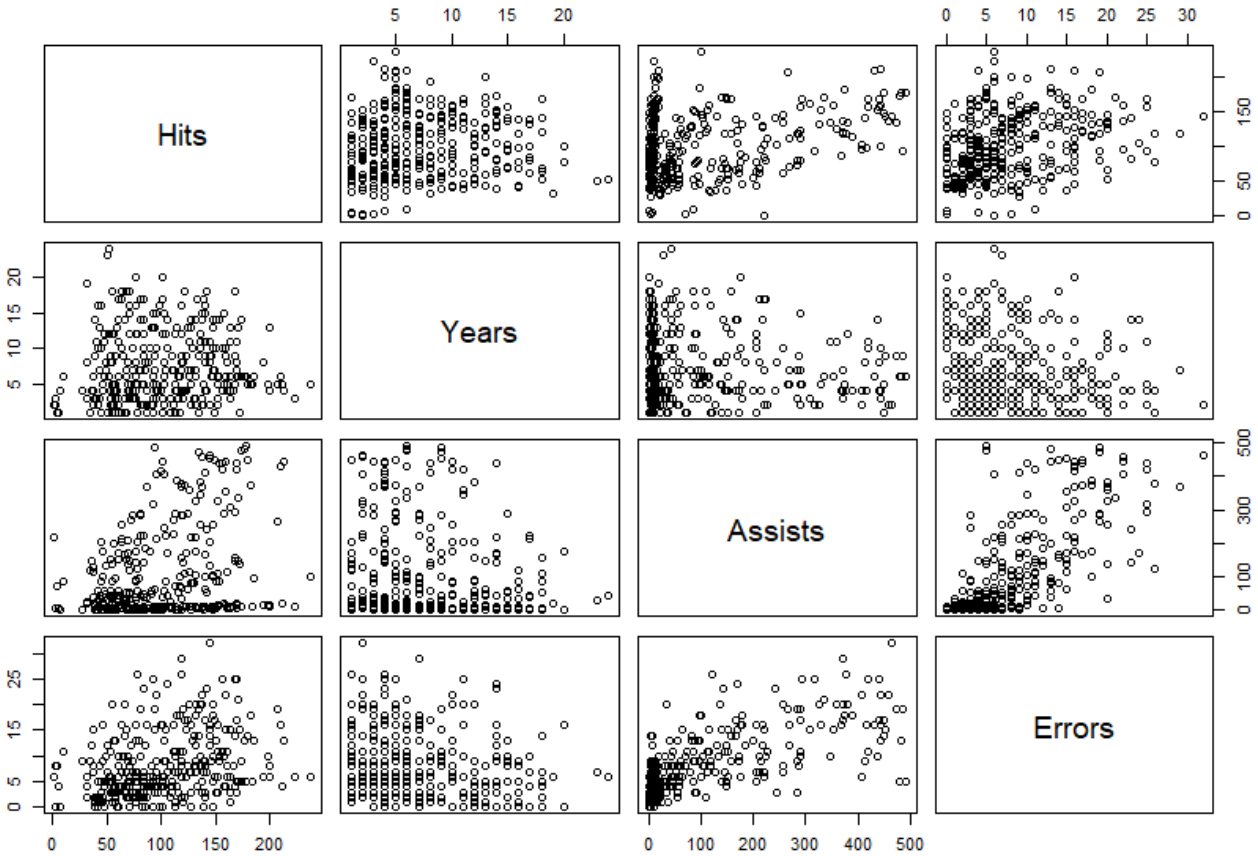
- Stratified Scatter plot is used to make patterns visible when the data is coming from variety of sources or multiple categories have been lumped together
- The technique separates the data so that patterns can be seen





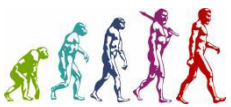
Matrix Plot

- It is a Matrix of scatterplots. The ijth scatter plot containing ith variable plotted against jth variable. It can be used to visually check possible correlated values

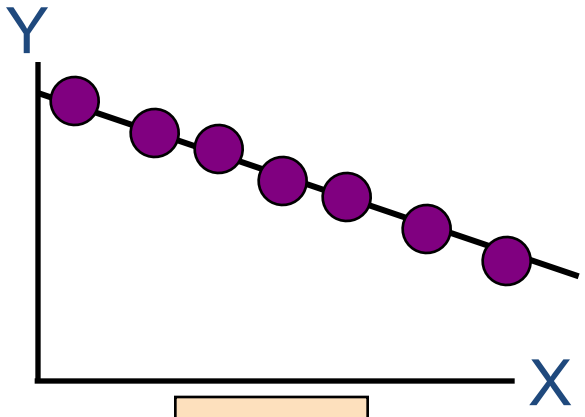


Correlation

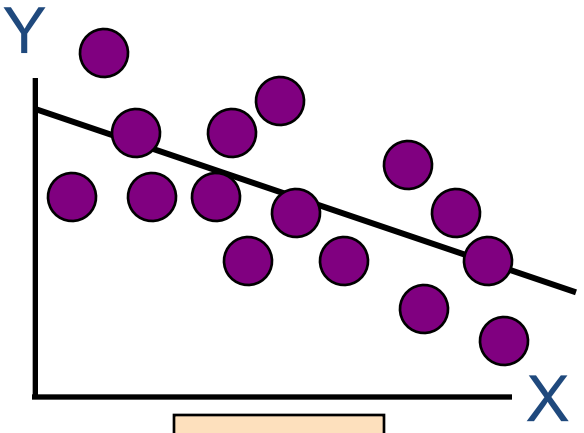
- In the special case that the scatterplot displays a linear relationship (and only then), we supplement the scatterplot with:
  - Numerical summaries: the correlation coefficient (r) measures the direction and, more importantly, the strength of the linear relationship
  - The closer r is to 1 (or -1), the stronger the positive (or negative) linear relationship
  - r is unit-less, influenced by outliers, and should be used only as a supplement to the scatterplot
- When the relationship is linear (as displayed by the scatterplot, and supported by the correlation r), we can summarize the linear pattern using the least squares regression line.



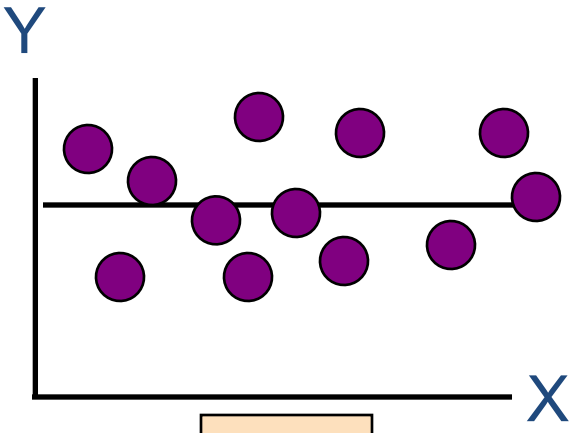
Scatter Plots of Data with Various Correlation Coefficients



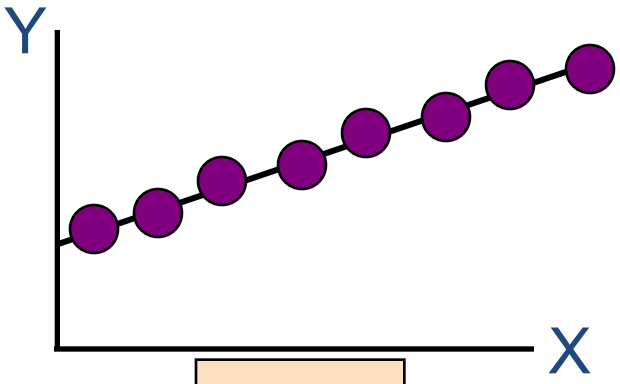
r = -1



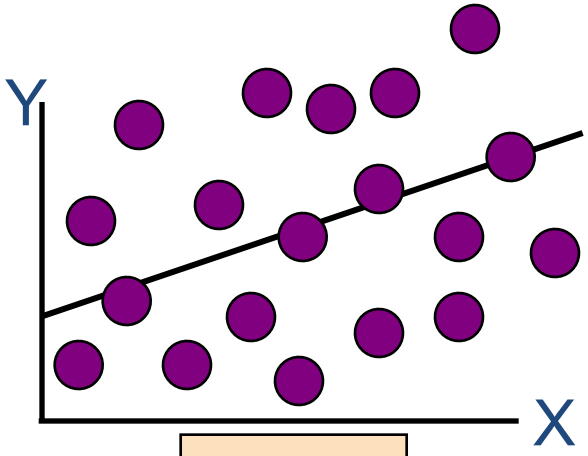
r = -.6



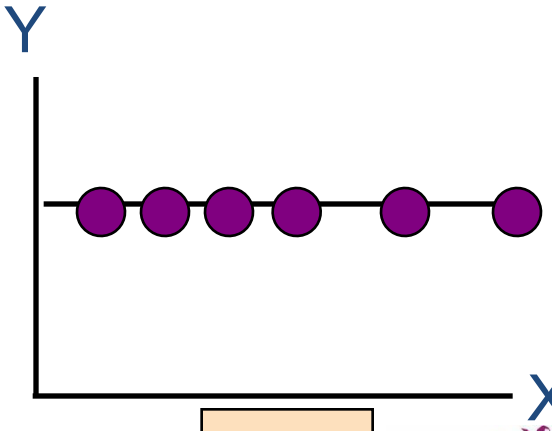
r = 0



r = +1



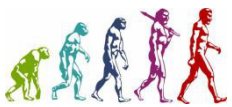
r = +.3



r = 0

Association vs Causation

- When examining the relationship between two variables (regardless of the case), any observed relationship (association) does not imply causation, due to the possible presence of lurking variables (Third variable)



Thank You