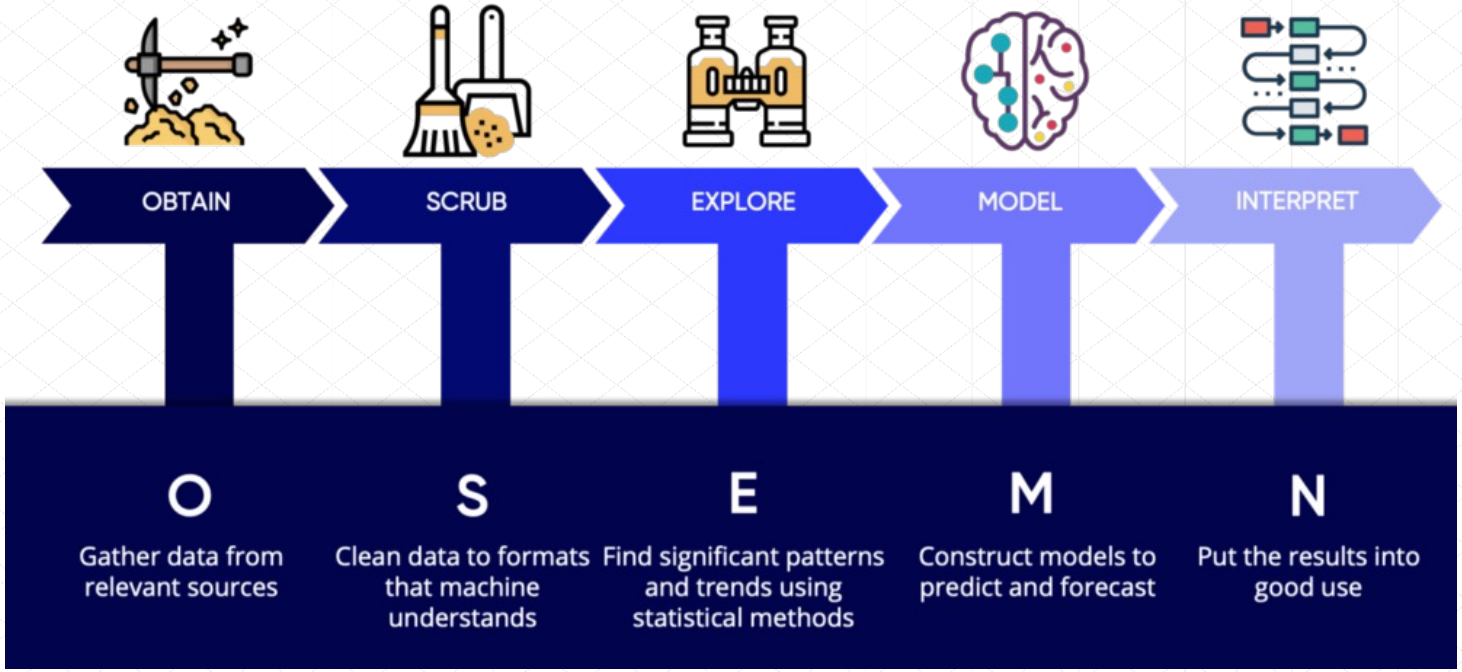
8 February 2023

UNIVERSITY OF GLOUCESTERSHIRE

CT4031
Maths for Data Science

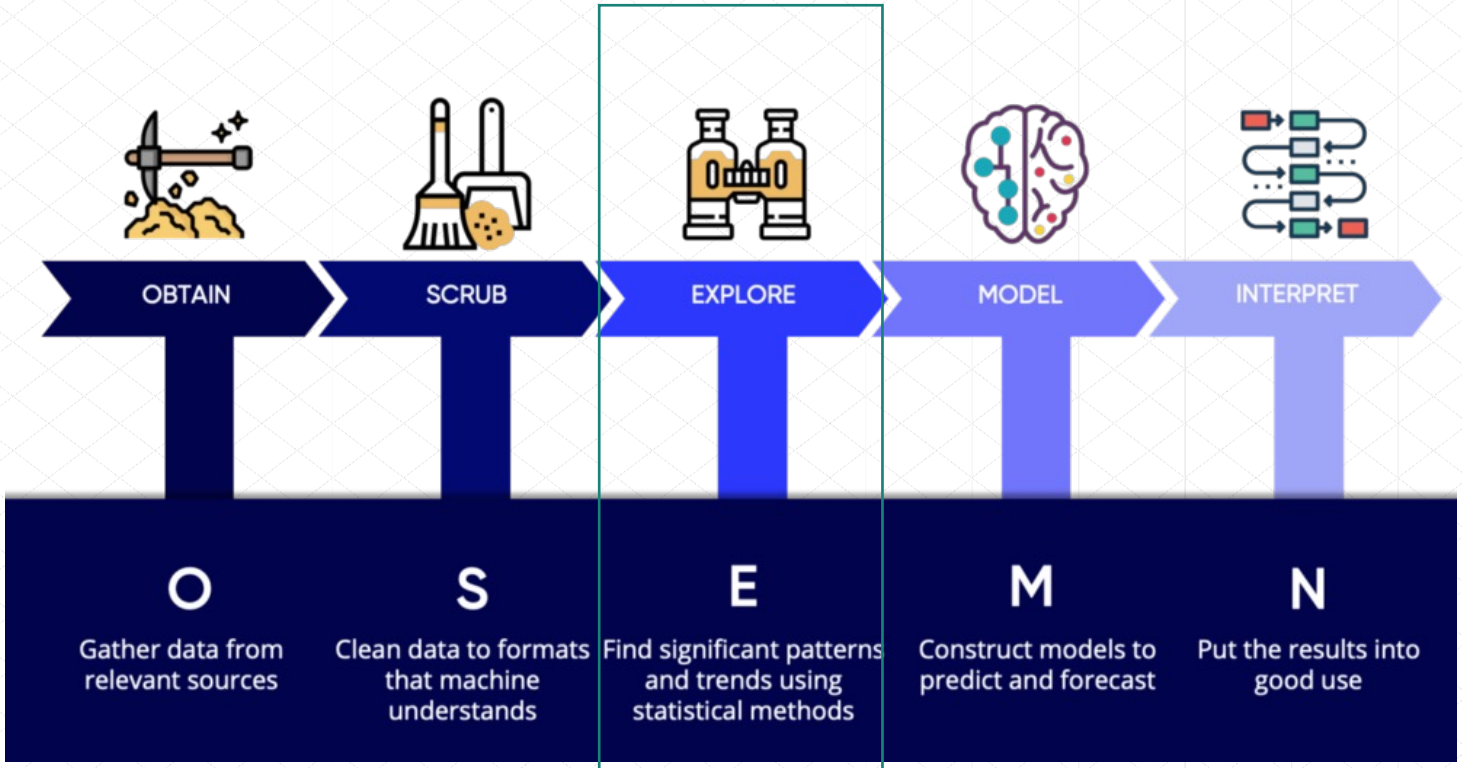
Week 4 – EDA

Data science process



Source: https://bit.ly/3rlxxqV

Data science process



Source: https://bit.ly/3rlxxqV

Exploratory Data Analysis: Introduction

- Foundation stone
- a critical first step in analyzing the data from an experiment.
- a systematic way to investigate relevant information from multiple perspectives.
- Like an investigation carried out by a detective.
- Digging deep into piles of data to find clues that would aid in actual data analysis.
- The better you know your data (have more clues) the better is your analysis (outcome).
- A philosophy or an attitude about how data analysis should be carried out, rather than being a fixed set of techniques.

Introduction

If you are going to find out anything about a data set you must first understand the data

Basically getting a feel for you numbers

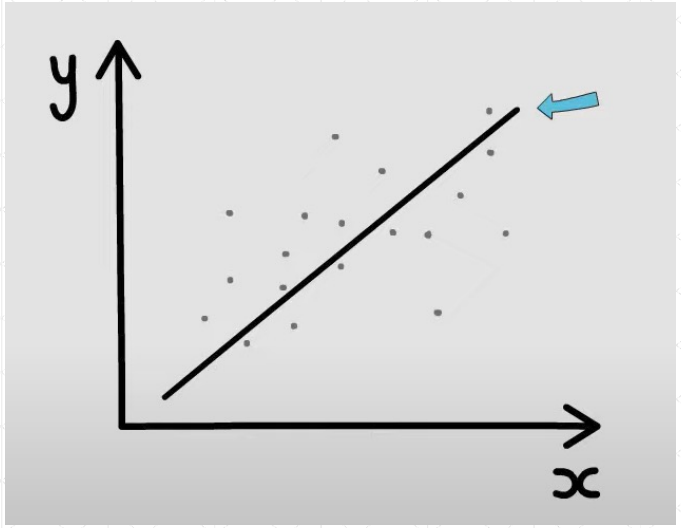
- Easier to find mistakes
- Easier to guess what actually happened
- Easier to find odd values

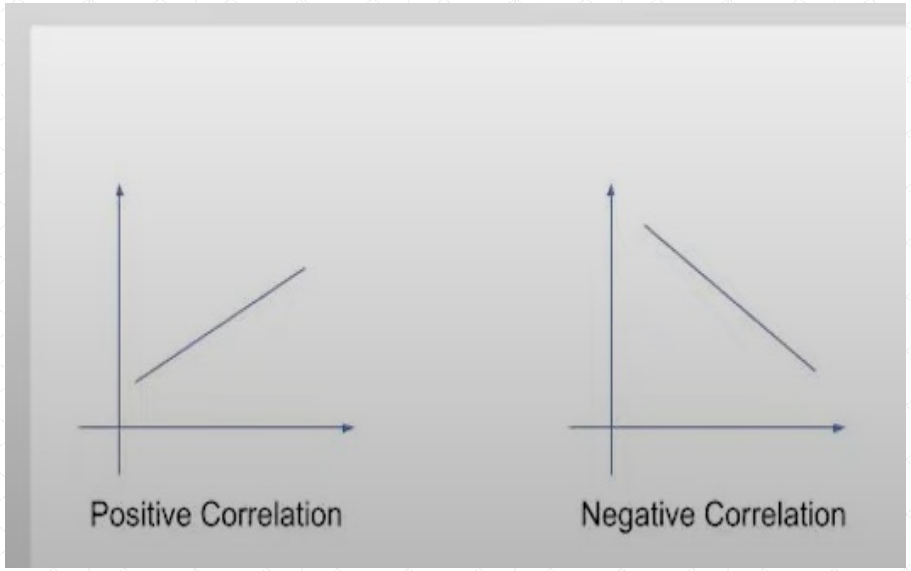
Objectives of EDA

- **Confirm if the data is making sense in the context of a problem domain.**
  - If not collect more data and change the strategy.
- **Uncover and resolve data quality issues.**
  - It is highly unlikely that you will receive a clean data. So EDA is helpful in fixing the data cleaning issues such as missing data, duplicate values, incorrect values, Anomalies, incorrect data types.
- **Get information about the data summary.**
  - Mean, Median, Mode, Variance, Skewness, Range, Minimum, Maximum, sum, count, standard deviation
  - Helps to understand the relationship variables which gives us a wider perspective on data

Objectives of EDA

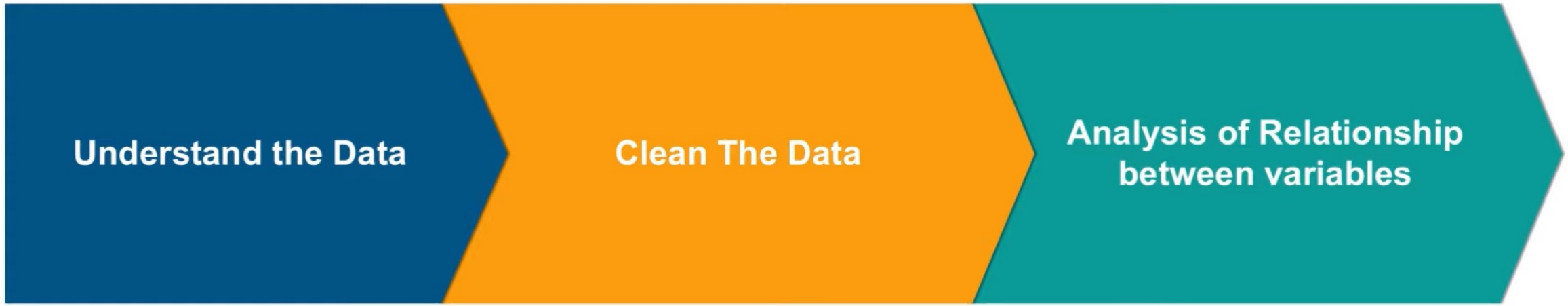
- **Detect outliers and anomalies**
  - Anomalies and outliers may lead to miscalculations.
  - Should be detected and removed before the actual data analysis.
- **Uncover and resolve data quality issues.**
  - It is highly unlikely that you will receive a clean data. So EDA is helpful in fixing the data cleaning issues such as missing data, duplicate values, incorrect values, Anomalies, incorrect data types.
- **Get information about the data summary.**
  - Mean, Median, Mode, Variance, Skewness, Range, Minimum, Maximum, sum, count, standard deviation
- **Drop unwanted columns and derive new variables**
  - Age=current year-DoB





Steps involved in EDA

- Follows a systematic set of steps to explore the data in a most efficient way possible.



Hey look, numbers!

x (the value)	f (frequency)
10	1
23	2
25	5
30	2
33	1
35	1

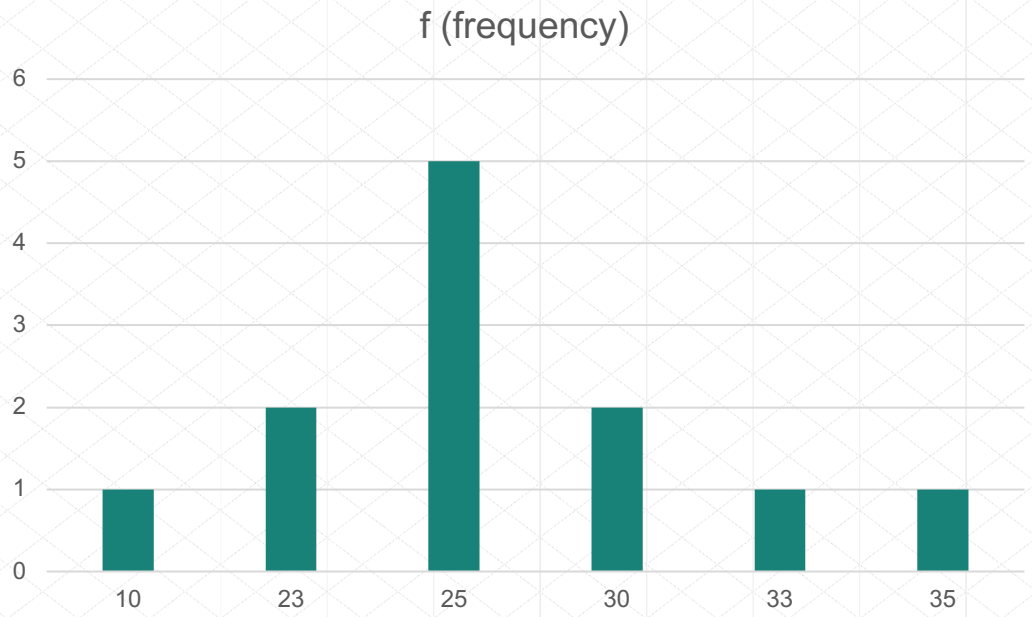
Frequency tables make stuff easy

10(1)+23(2)+25(5)+30(2)+33(1)+35(1)= 309

$$\sum xf$$

Relative Frequency Histogram

- You can use this to make a relative frequency histogram
- Lose no richness in the data
- Easy to reconstruct data set
- Allows you to spot oddities



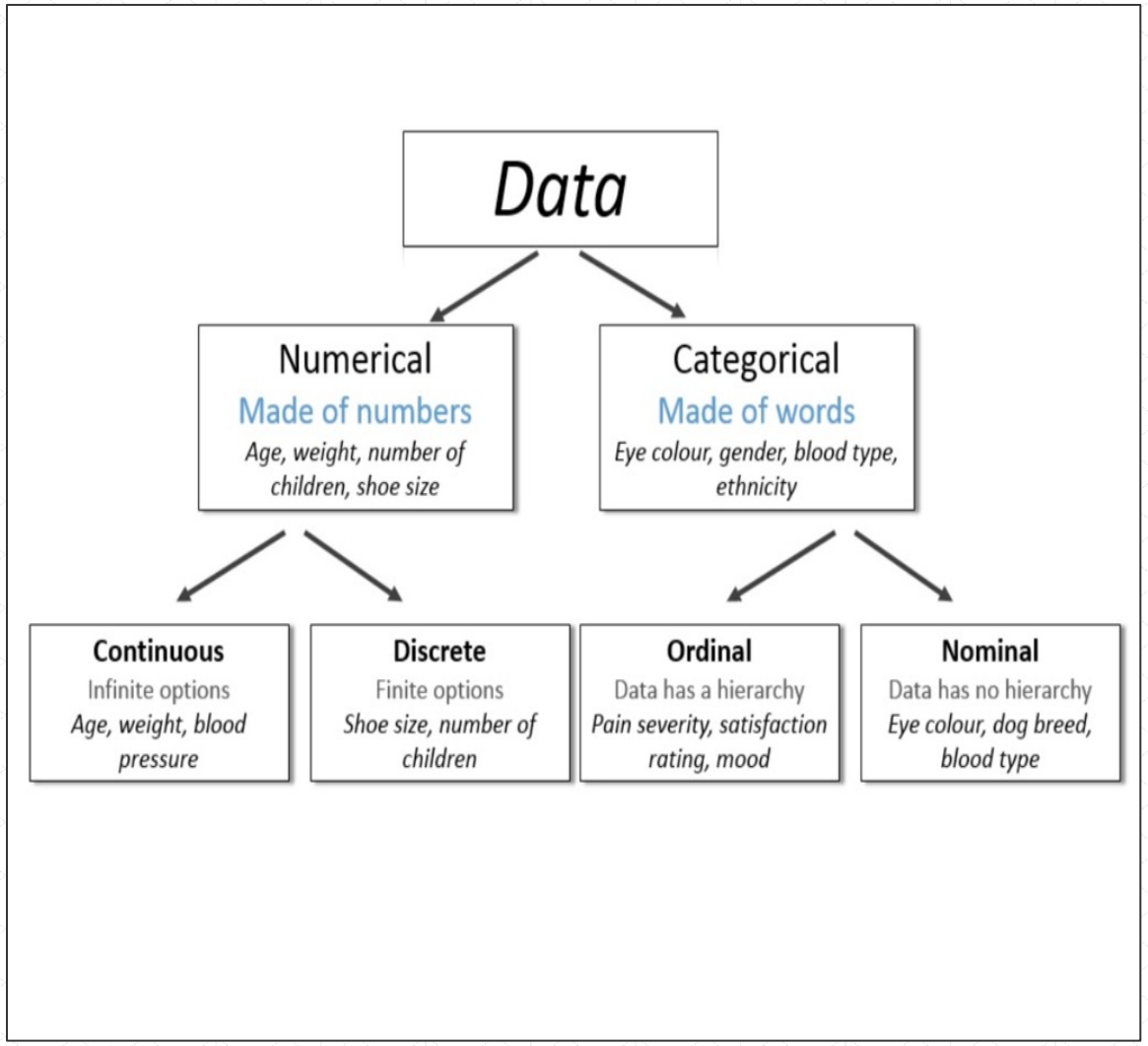
Types of Data in a Dataset

Datasets are composed by different types of data
Task

- Does this matter?
- Does this affect the cleaning process?
- What are the main different types?
- How EDA and Visualization differs when it comes to different types of Data in a Dataset?

Types of Data in a Dataset

- We have to see first what kind of data we need to deal with while doing EDA.
- you will get numerical or Categorical data, if not (such as an image data we have to convert the image into the numerical form).

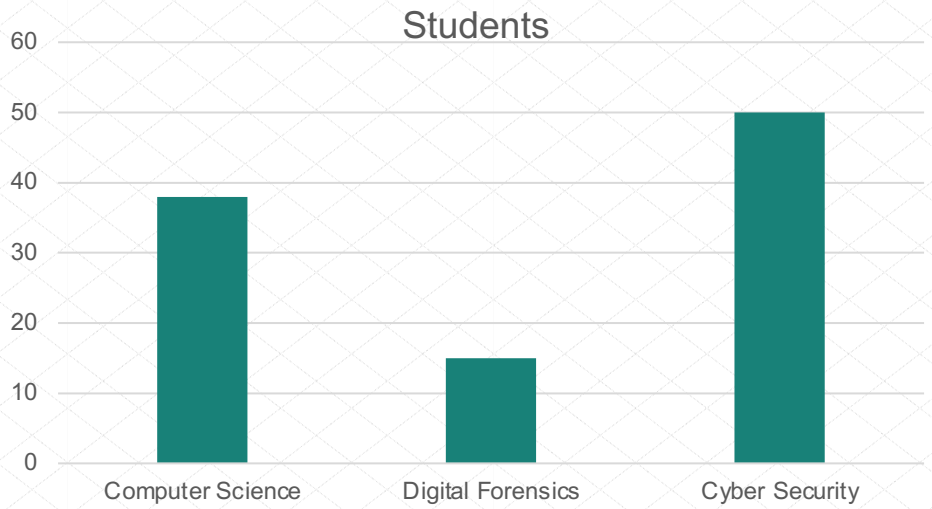


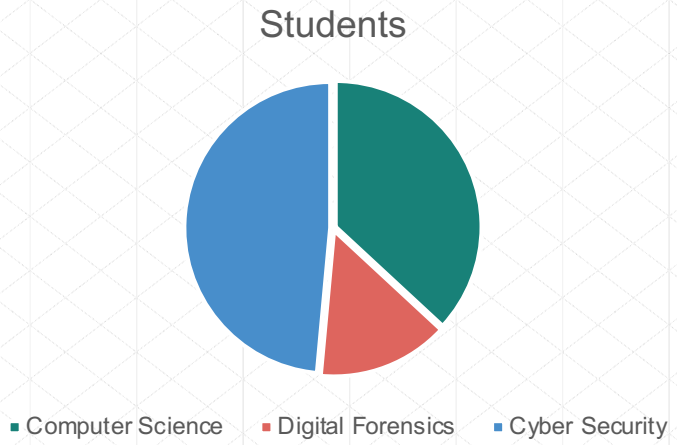
Categorical Data

- This is any data that isn't a number.
  - **Ordinal:** have a set of order e.g rating happens on scale 1-10
  - **Binary:** Have only two values 0 or 1
  - **Nominal:** No set of orders. For example countries
- With categorical data you do not get a histogram, you get a bar graph.
- You could do a pie chart too (btw, I love pie)
- Pretty much the same thing, but the X axis does not have a scale so to speak

Like this

Course	Students
Computer Science	38
Digital Forensics	15
Cyber Security	50





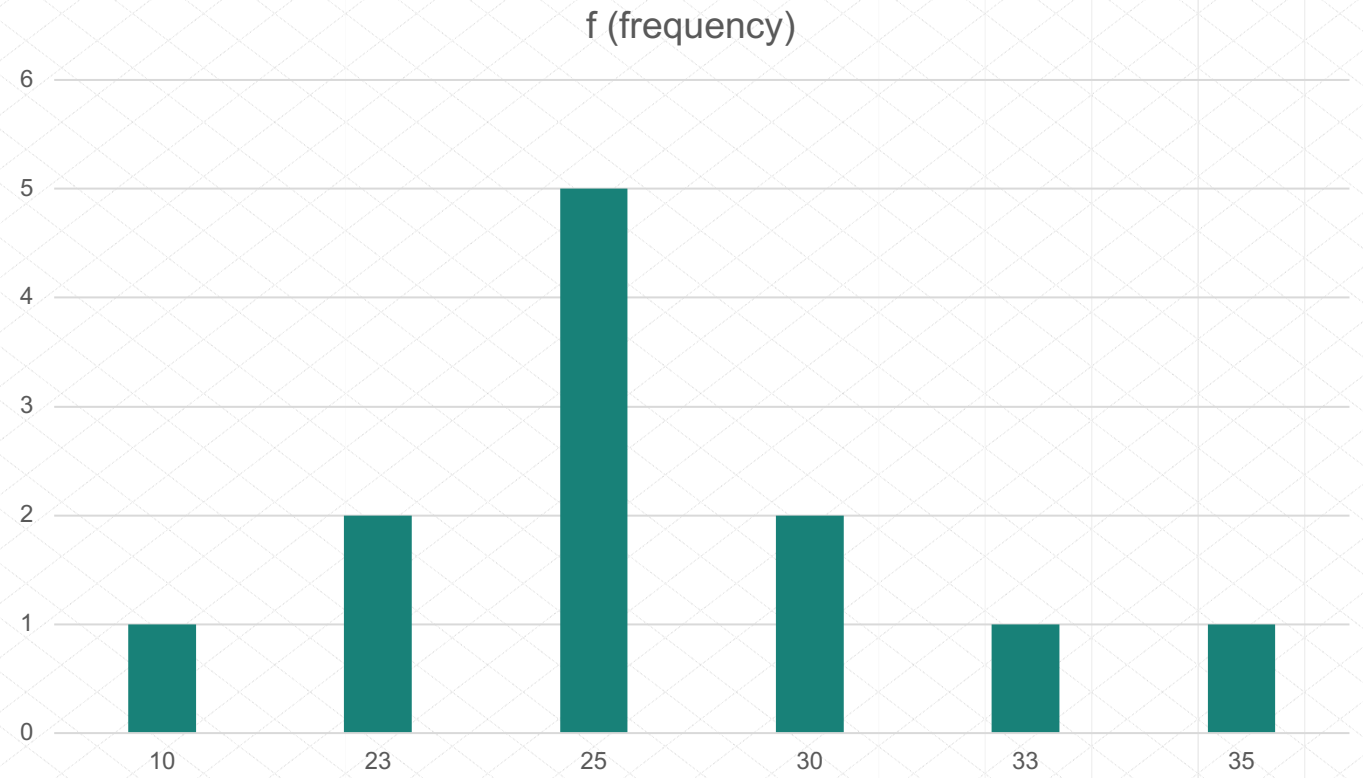
Numerical Data

- Data consists of numbers
  - Continuous: temperature, height,
  - Discrete: Non continuous.
- So with these of course we use a histogram
- We can see central tendency (mean, mode, median)
- Spread and Shape (standard deviation, variance, range, max, min, etc)
- Percentiles and Quartiles

Hey look, numbers!

x (the value)	f (frequency)
10	1
23	2
25	5
30	2
33	1
35	1

Relative Frequency Histogram

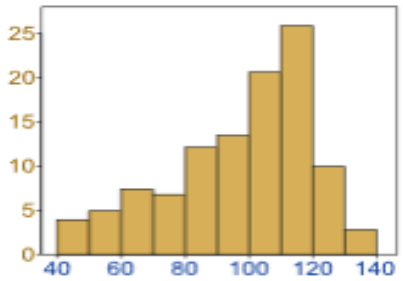


Skewness and Kurtosis

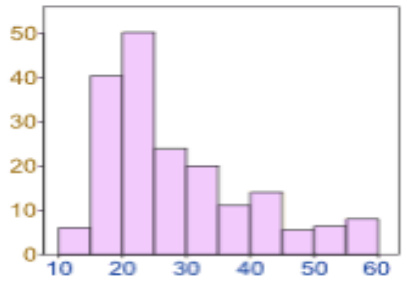
- A fundamental task in many statistical analyses is to characterize the ***location*** and ***variability*** of a dataset.
- Skewness is a measure of symmetry
  - A dataset, is symmetric if it looks the same to the left and right.
- Kurtosis is a measure of distribution
  - Datasets with high kurtosis tend to have outliers.
  - Datasets with low kurtosis tend to have lack of outliers.

Skewness

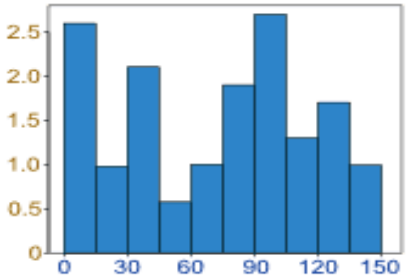
It can be spread out more on the left



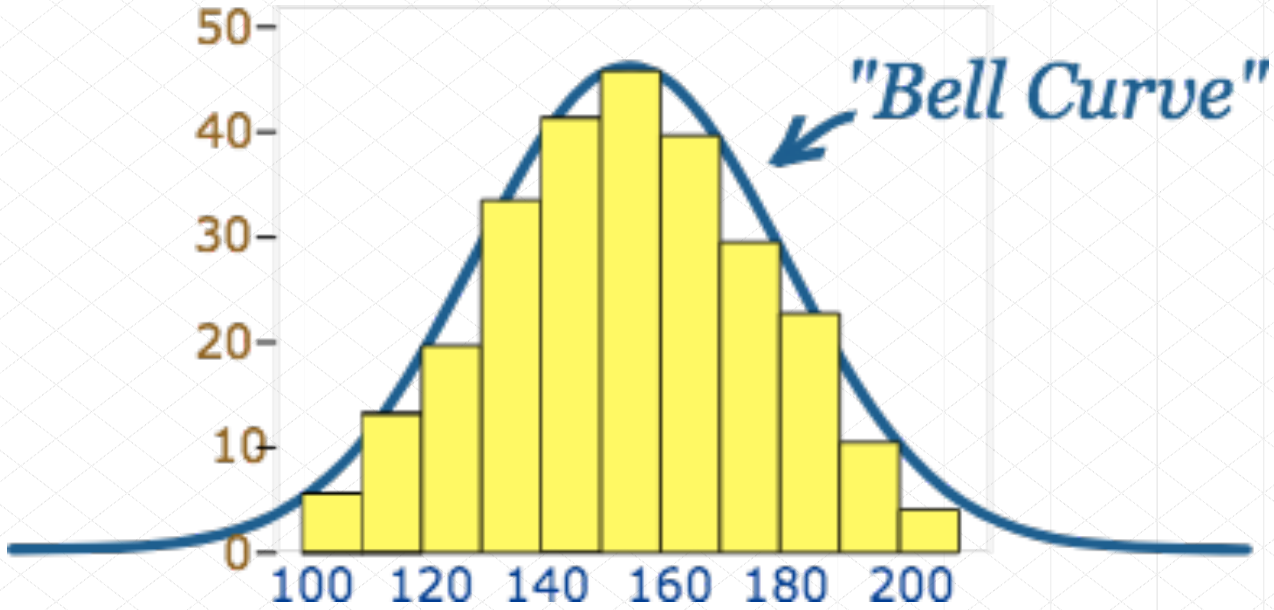
Or more on the right



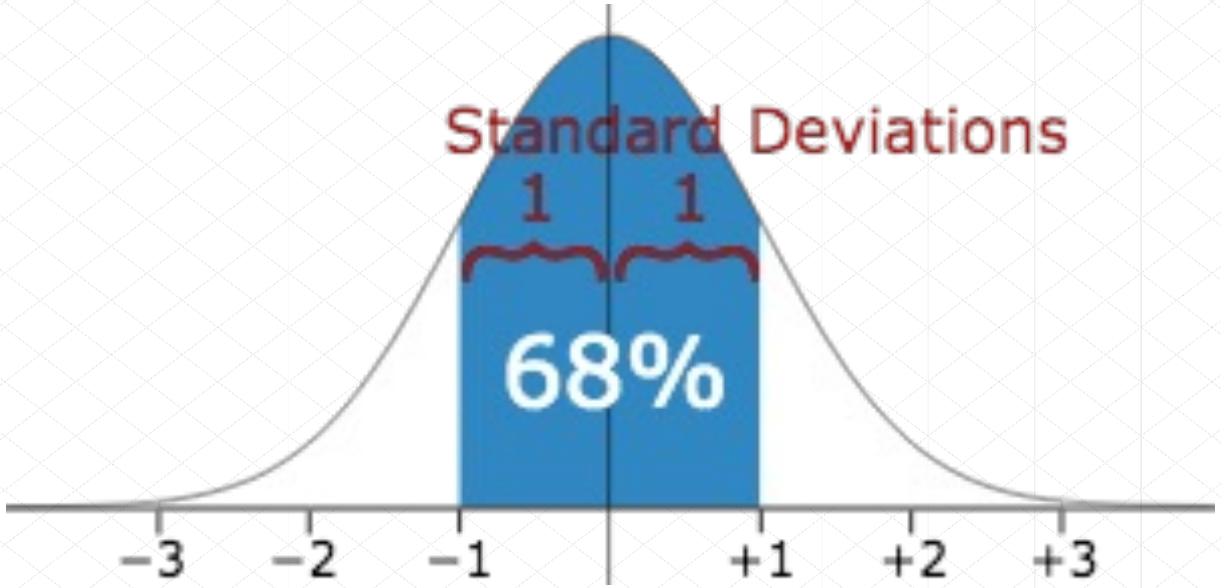
Or it can be all jumbled up



Skewness, Kurtosis and normal distribution



Skewness, Kurtosis and normal distribution



Visualization

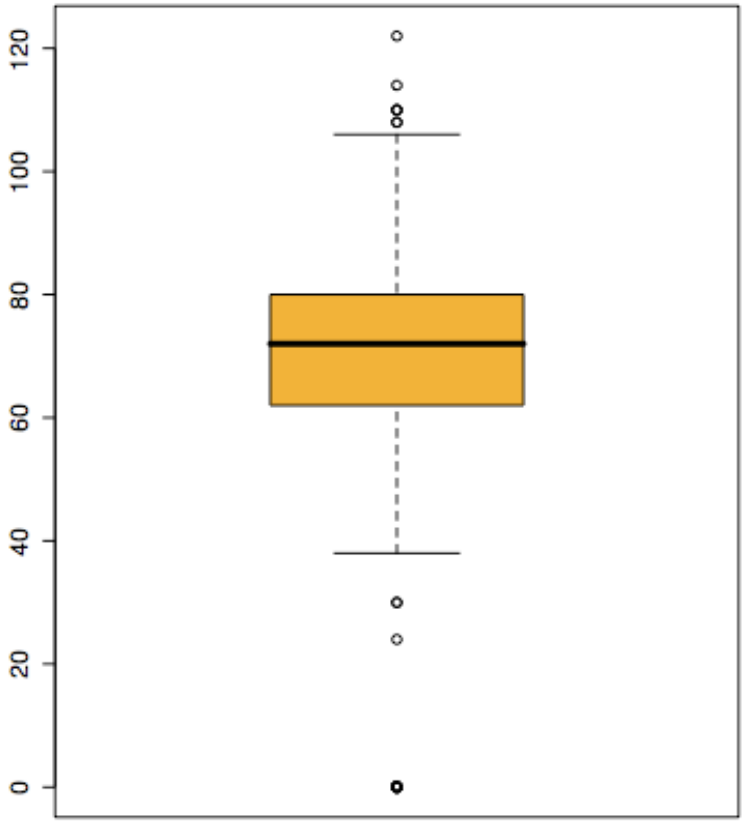
Boxplots

Shows a lot of information about a variable in one plc

- Median
- Quartiles
- Outliers
- Range

Negatives

- Overplotting
- Hard to tell distributional shape

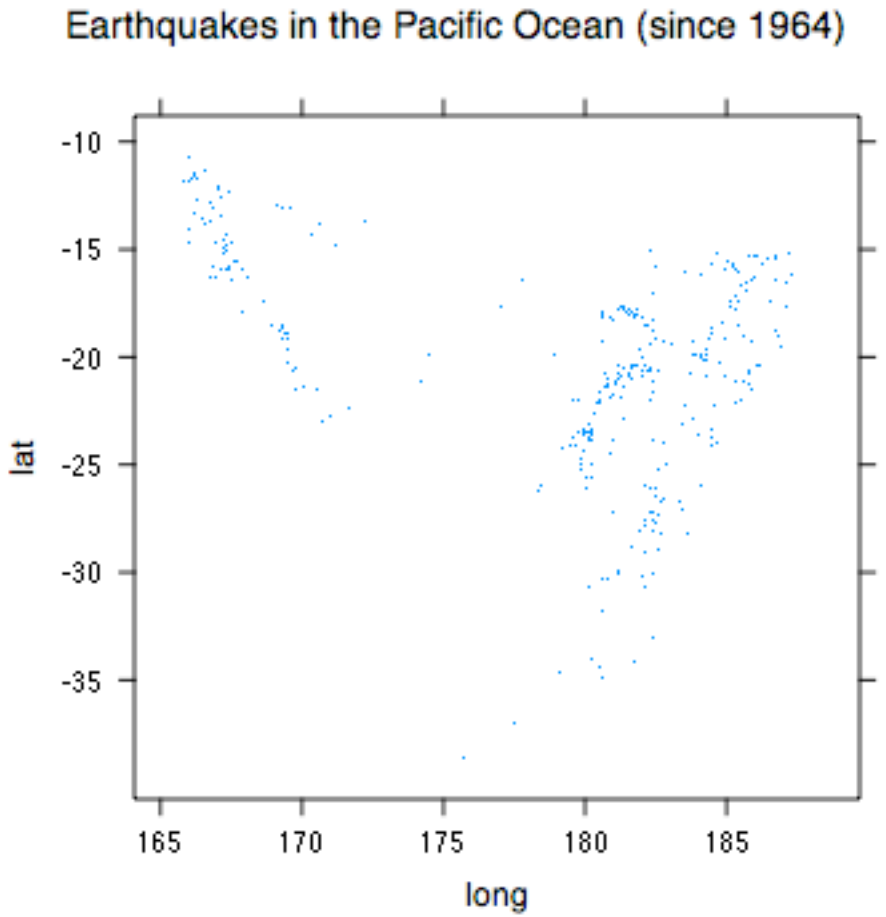


Time Series

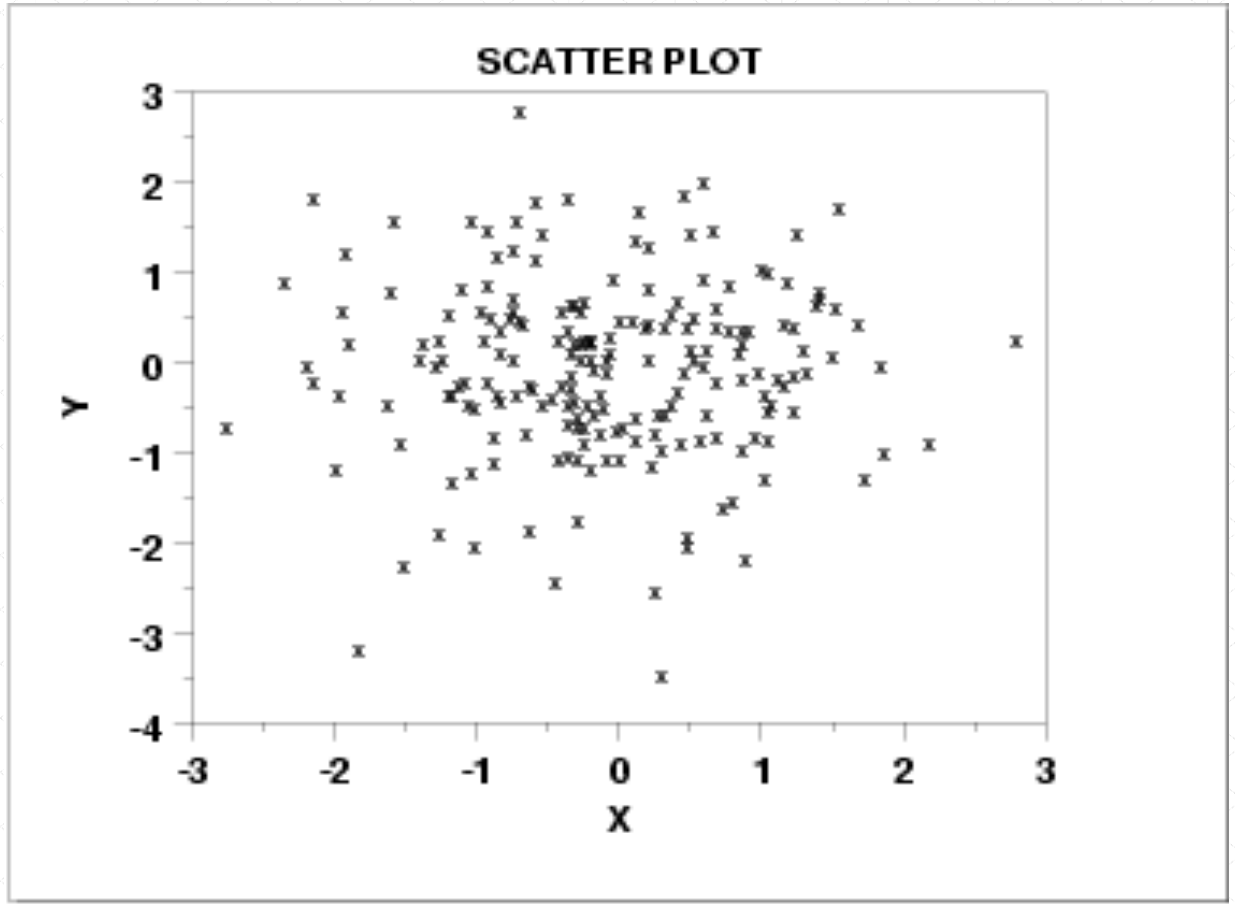
If your data has a temporal component, be sure to exploit it

Spatial Data

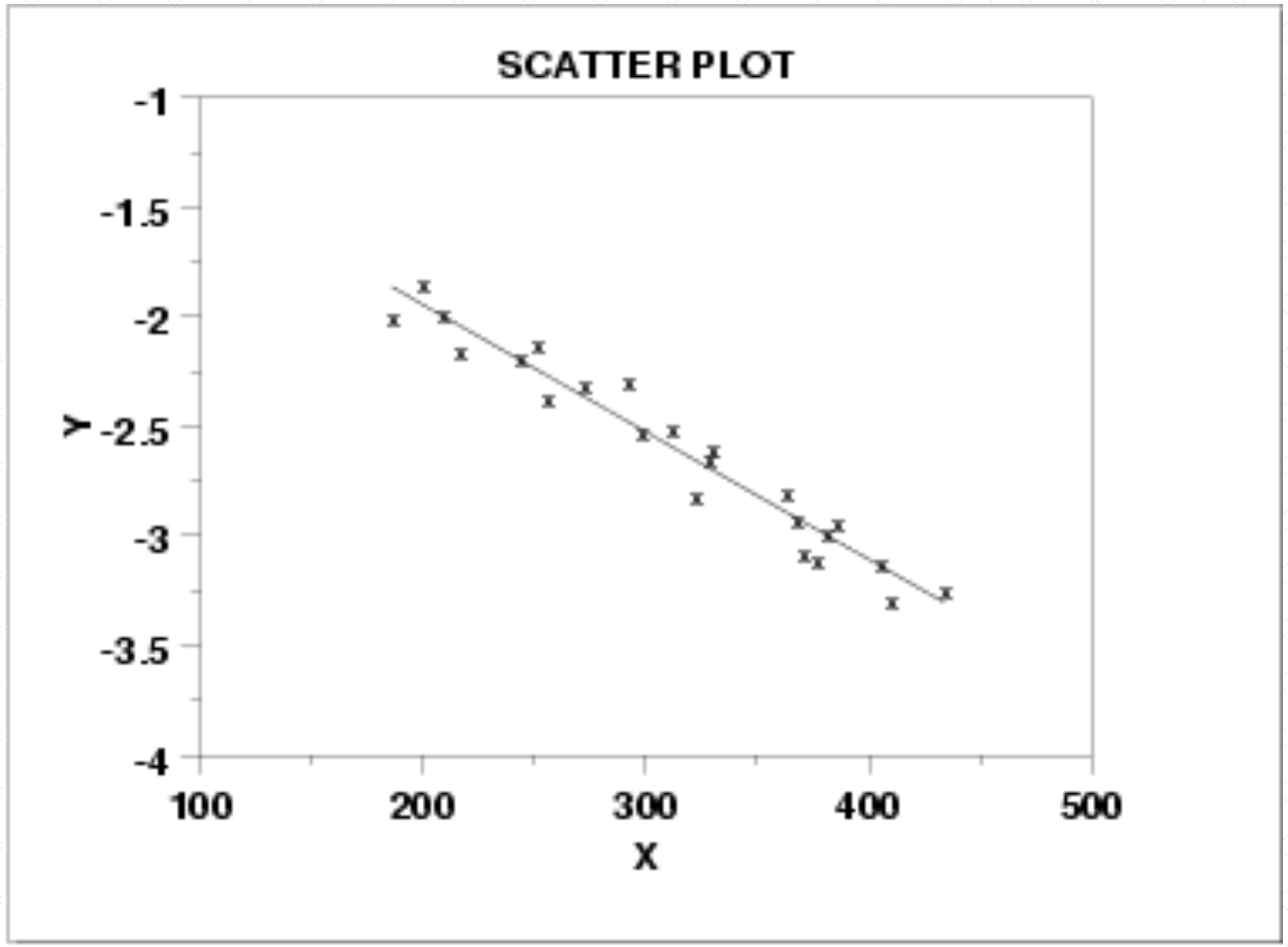
- If your data has a geographic component, be sure to exploit it
- Scatterplot



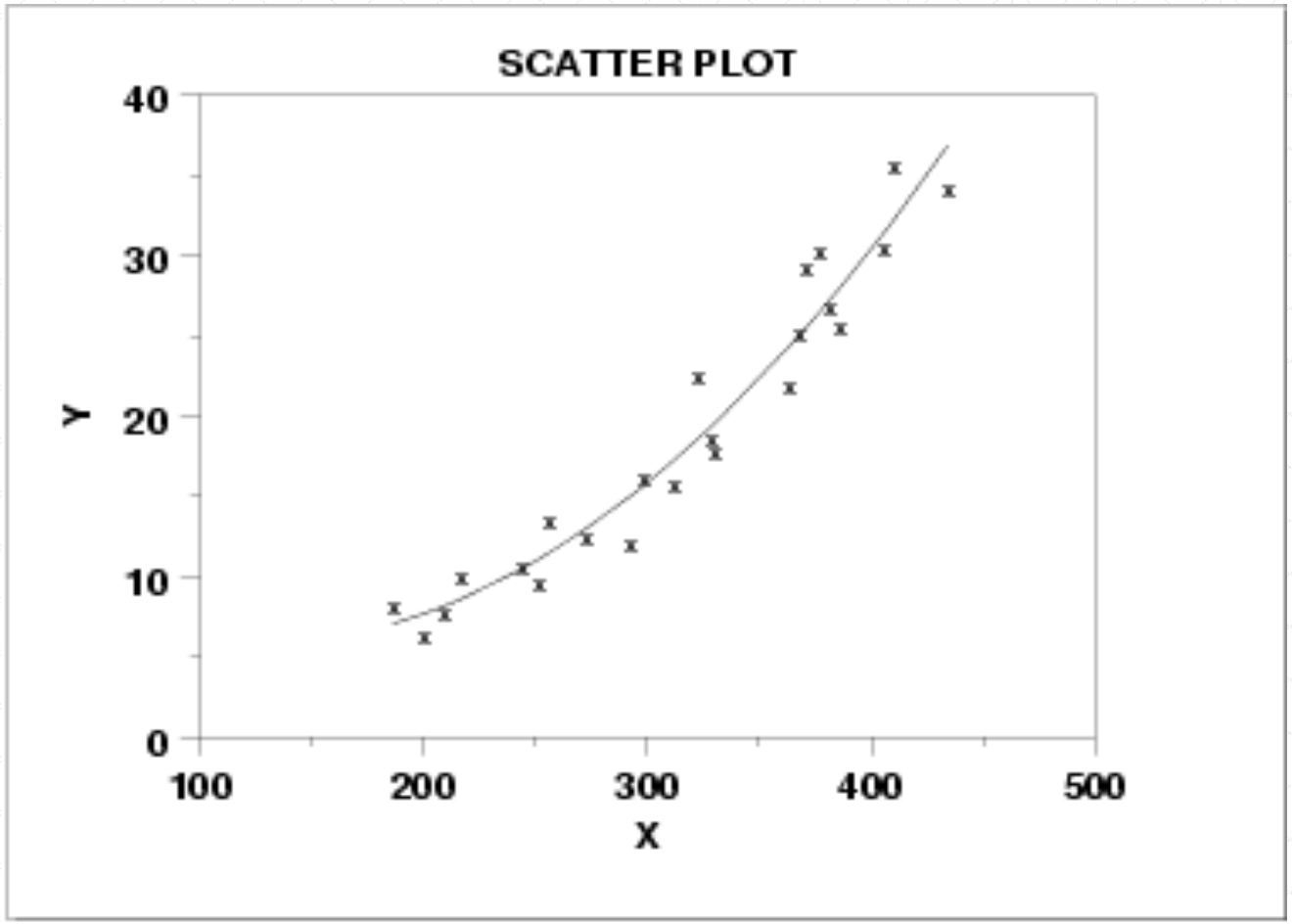
Scatter Plot: No apparent relationship



Scatter Plot: Linear relationship



Scatter Plot: Quadratic relationship



Python Packages for EDA

Seaborn

Post-sessional work

Research and provide examples of probabilistic independence.

References

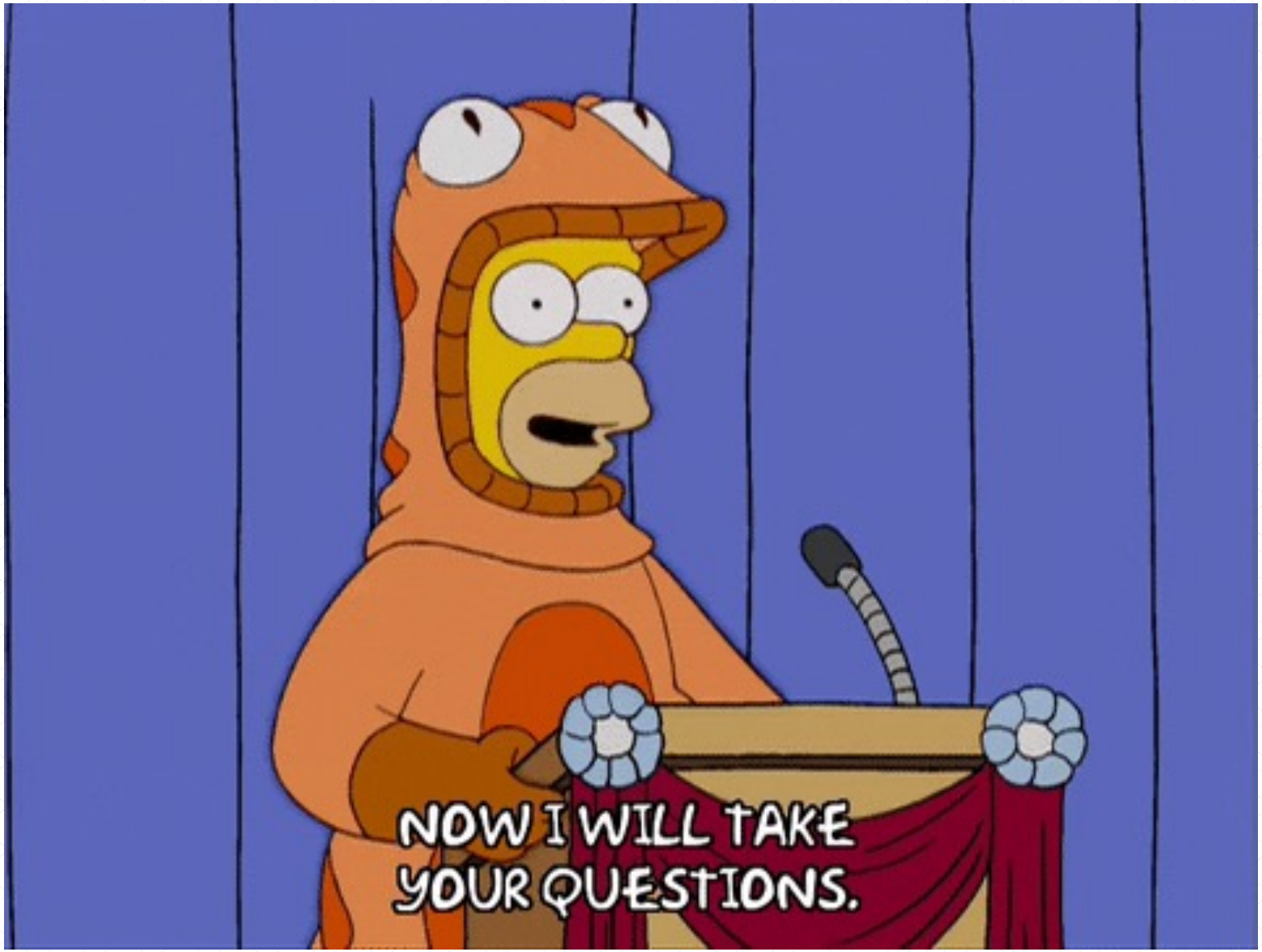
‘Exploratory Data Analysis. Psychology’. https://docplayer.net/15063292-Exploratory-data-analysis-psychology-3256.html (accessed Jul. 14, 2021).

‘Data Mining and Predictive Modelling . https://bit.ly/3ibTO62. (accessed Jul. 14, 2021).

https://www.oxfordbibliographies.com/view/document/obo-9780199828340/obo-9780199828340-0200.xml

Next Session!

- Probability Part 1


NOW I WILL TAKE
YOUR QUESTIONS.