ENCS5341
Machine Learning and Data Science

Introduction to Data Science

Data Science

- **Data science** is a field of study and practice that involves the collection, storage, and processing of data in order to derive important insights into a problem or a phenomenon.

- Data may be generated by humans (surveys, logs, etc.) or machines (weather data, road vision, etc.), and could be in different formats (text, audio, video, augmented or virtual reality, etc.)

Where Do We See Data Science?

- Finance
- Public Policy
- Politics
- Healthcare
- Urban Planning
- Education
- Libraries
- ….
- …

Where Do We See Data Science?

Finance

- Data scientists capture and analyze new sources of data
- Building predictive models and running real-time simulations of market events.
- Help the finance industry obtain the information necessary to make accurate predictions.
- Fraud detection and risk reduction.
- Minimize the chance of loan defaults via information such as customer profiling, past expenditures, and other essential variables that can be used to analyze the probabilities of risk and default.
- Analyze a customer's purchasing power to more effectively try to sell additional banking products.
- Identify the creditworthiness of potential customers

Where Do We See Data Science?

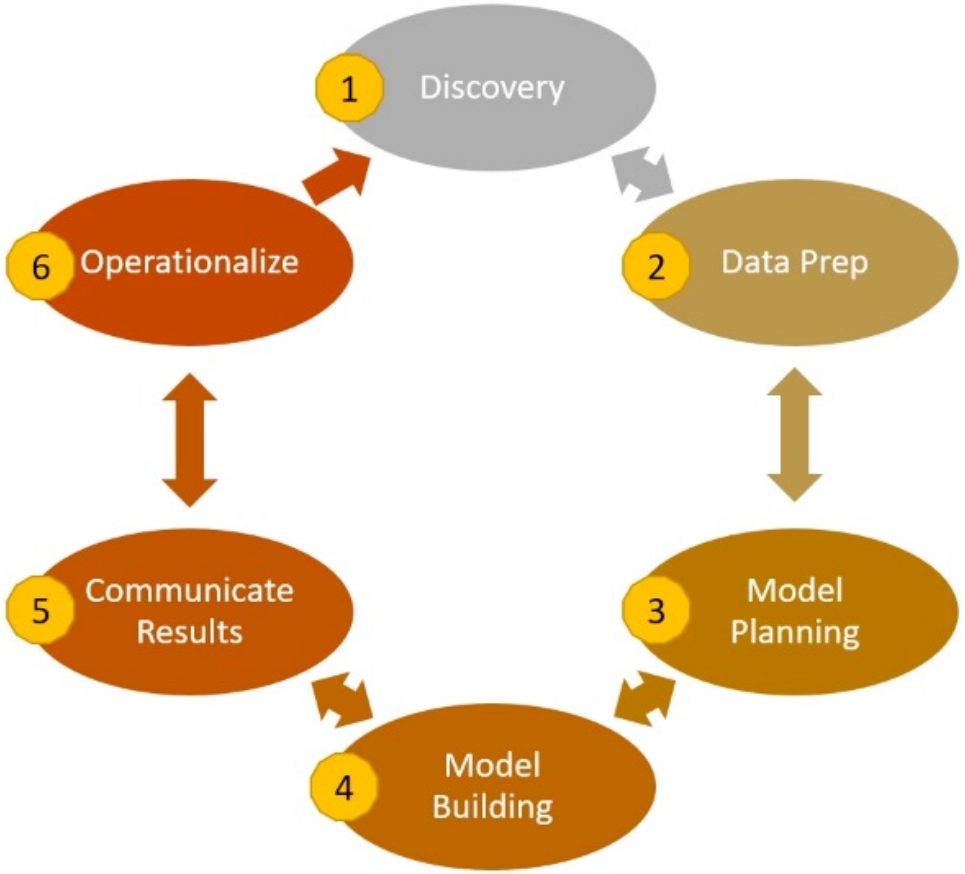
Public Policy

- Data science helps governments and agencies gain insights into citizen behaviors that affect the quality of public life, including traffic, public transportation, social welfare, community wellbeing, etc. This information, or data, can be used to develop plans that address the betterment of these areas.

Politics

- Data scientists have been quite successful in constructing accurate voter targeting models and increasing voter participation.

Data Analytics Life Cycle



Data Analytics Life Cycle

- **Phase 1—Discovery**: In Phase 1, the team learns the business domain, including attempted similar projects in the past from which they can learn. The team assesses the resources available to support the project in terms of people, technology, time, and data. Important activities in this phase include framing the business problem as an analytics challenge that can be addressed in subsequent phases and formulating initial hypotheses to test and begin learning the data.

- **Phase 2—Data preparation**: Phase 2 the team works with data and perform analytics for the duration of the project. The team needs to execute extract, load, and transform (ELT) or extract, transform and load (ETL) to get data into the sandbox. The ELT and ETL are sometimes abbreviated as ETLT. Data should be transformed in the ETLT process so the team can work with it and analyze it. In this phase, the team also needs to familiarize itself with the data thoroughly and take steps to condition the data.

Data Analytics Life Cycle

- **Phase 3—Model planning**: Phase 3 is model planning, where the team determines the methods, techniques, and workflow it intends to follow for the subsequent model building phase. The team explores the data to learn about the relationships between variables and subsequently selects key variables and the most suitable models.

- **Phase 4—Model building**: In Phase 4, the team develops datasets for testing, training, and production purposes. In addition, in this phase the team builds and executes models based on the work done in the model planning phase. The team also considers whether its existing tools will suffice for running the models, or if it will need a more robust environment for executing models and workflows (for example, fast hardware and parallel processing, if applicable).

Data Analytics Life Cycle

- **Phase 5—Communicate results**: In Phase 5, the team, in collaboration with major stakeholders, determines if the results of the project are a success or a failure based on the criteria developed in Phase 1. The team should identify key findings, quantify the business value, and develop a narrative to summarize and convey findings to stakeholders.

- **Phase 6—Operationalize**: In Phase 6, the team delivers final reports, briefings, code, and technical documents. In addition, the team may run a pilot project to implement the models in a production environment.

What Do Data Scientists Do?

- Data collection
- Descriptive statistics
- Correlation
- Data visualization
- Model building
- Extrapolation and regression analysis

Data

- Data be viewed as the raw material from which information is obtained.

- Two main types of data:
  - Structured data.
  - unstructured data.

- **Structured data** refers to highly organized information that can be seamlessly included in a database and readily searched via simple search operations; whereas **unstructured** data is essentially the opposite, devoid of any underlying structure
- Structured data is data that has been predefined and formatted to a set structure before being placed in data storage.
- Unstructured data is data stored in its native format and not processed until it is used. Example: social media posts, presentations, chats, and satellite imagery

How to Collect Data?

Method	Effort
Download	Low
API (Application program interface)	Medium
Scrape/Crawl	High

Data Storage and Presentation

- Depending on its nature, data is stored in various formats.

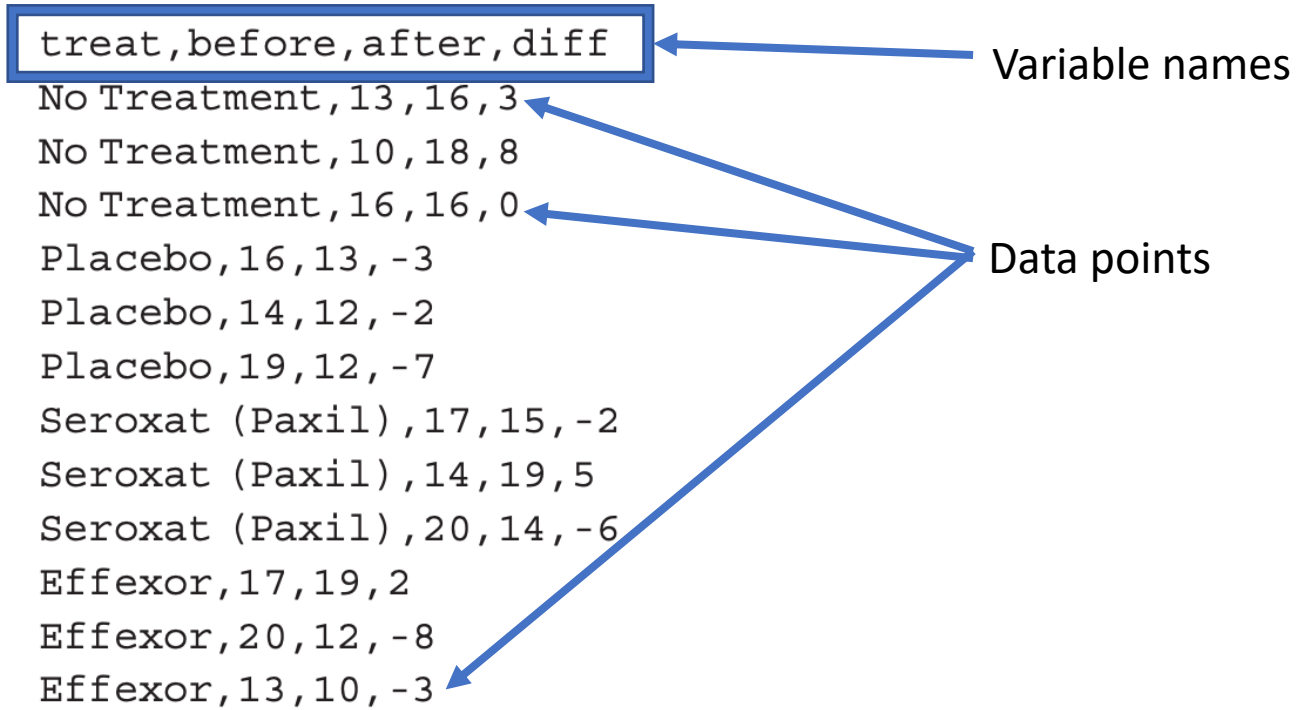
- We will start with simple kinds – data in text form. If such data is structured, it is common to store and present it in some kind of delimited way. That means various fields and values of the data are separated using delimiters, such as commas or tabs

- Structured text data can be stored in the following formats:
  - Comma Separated Values **CSV**
  - Tab Separated Value **TSV**
  - eXtensible Markup Language **XML**
  - Really Simple Syndication **RSS**
  - JavaScript Object Notation **JSON**

- The two most commonly used formats that store data as simple text – comma-separated values (CSV) and tab-separated values (TSV).

CSV (Comma-Separated Values)

- the most common import and export format for spreadsheets and databases.
- the first row mentions the variable names. The remaining rows each individually represent one data point.
- Example: A snippet from a dataset that represents the effectiveness of different treatment procedures on separate individuals with clinical depression



TSV (Tab-Separated Values)

- TSV files are used for raw data and can be imported into and exported from spreadsheet software. Tab-separated values files are essentially text files, and the raw data can be viewed by text editors, though such files are often used when moving raw data between spreadsheets.
- Example

```
Name<TAB>Age<TAB>Address
Ryan<TAB>33<TAB>1115 W Franklin
Paul<TAB>25<TAB>Big Farm Way
Jim<TAB>45<TAB>W Main St
Samantha<TAB>32<TAB>28 George St
```

- An advantage of TSV format is that the delimiter (tab) will not need to be avoided because it is unusual to have the tab character within a field. In fact, if the tab character is present, it may have to be removed. On the other hand, TSV is less common than other delimited formats such as CSV.

XML(eXtensibleMarkupLanguage)

- XML was designed to be both human and machine readable, and can thus be used to store and transport data. In the real world, computer systems and databases contain data in incompatible formats. As the XML data is stored in plain text format, it provides a software and hardware independent way of storing data. This makes it much easier to create data that can be shared by different applications.

- Example

```
<?xml version="1.0" encoding="UTF-8"?>
<bookstore>
     <book category="information science" cover="hardcover">
          <title lang="en">Social Information Seeking</title>
          <author>Chirag Shah</author>
          <year>2017</year>
          <price>62.58</price>
     </book>
     <book category="data science" cover="paperback">
          <title lang="en">Hands-On Introduction to Data
            Science</title>
          <author>Chirag Shah</author>
          <year>2019</year>
          <price>50.00</price>
     </book>
</bookstore>
```

JSON (JavaScript Object Notation)

- JSON is a lightweight data-interchange format. It is not only easy for humans to read and write, but also easy for machines to parse and generate.
- JSON is built on two structures:
  - A collection of name–value pairs. In various languages, this is realized as an object, record, structure, dictionary, hash table, keyed list, or associative array.
  - An ordered list of values. In most languages, this is realized as an array, vector, list, or sequence.
- Example


```
[
    {
        "trackid": "AA-1234",
        "reported_dt": "12/31/2019 23:59:59",
        "longitude": -111.12500000,
        "latitude": 33.37500000
    },
    {
        "trackid": "BB-7890",
        "reported_dt": "12/31/2019 23:59:59",
        "longitude": -113.67500000,
        "latitude": 35.87500000
    },
    {
        "trackid": "CC-4545",
        "reported_dt": "12/31/2019 23:59:59",
        "longitude": -115.57500000,
        "latitude": 37.67500000
    }
]
```

Data Pre-processing

- Data in the real world is often *dirty* – not ready to be used for the desired purpose.

- How dirty is real data?

  Examples

  - Jan 19, 2016
  - January 19, 16
  - 1/19/16
  - 2006-01-19
  - 19/1/16

Data Pre-processing

- factors that indicate that data is not clean or ready to process:

  - **Incomplete**. When some of the attribute values are lacking, certain attributes of interest are lacking, or attributes contain only aggregate data.

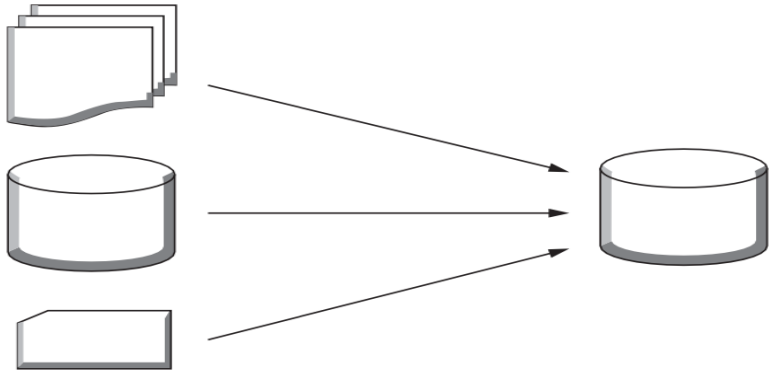
  - **Noisy**. When data contains errors or outliers. For example, some of the data points in a dataset may contain extreme values that can severely affect the dataset's range. (e.g., Salary="-10")

  - **Inconsistent**. Data contains discrepancies in codes or names. For example, if the "Name" column for registration records of employees contains values other than alphabetical letters, or if records do not start with a capital letter, discrepancies are present. (e.g., Age="42" Birthday="03/07/1997"; Was rating "1,2,3", now rating "A, B, C")

Forms of data pre-processing

- Data Cleaning

- Data Integration



- Data Transformation

–17, 25, 39, 128, –39 → 0.17, 0.25, 0.39, 1.28, –0.39

- Data Reduction

	A1	A2	A3	....	A200
T1					
T2					
T3					
....					
T200					

→

	A1	A2	A3	...	A120
T1					
T2					
T3					
....					
T150					

Data Cleaning

Following are different ways to clean data:

- **Munging or Wrangling**: Involves reformatting data into usable format.

  Often, the data is not in a format that is easy to work with. For example, it may be stored or presented in a way that is hard to process. Thus, we need to convert it to something more suitable for a computer to understand. To accomplish this, there is no specific scientific method. The approaches to take are all about manipulating or wrangling (or munging) the data to turn it into something that is more convenient or desirable. This can be done manually, automatically, or, in many cases, semi-automatically.

- **Handling missing data**: Sometimes data may be in the right format, but some of the values are missing may be due to **problems with the process of collecting data, or an equipment malfunction**. Or, **comprehensiveness** may not have been considered important at the time of collection. Furthermore, some data may get lost due to **system or human error** while storing or transferring the data.

  Strategies to combat missing data: Global constant, Ignoring, Imputing or Inference approach

- **Filtering noisy data**: Identifying outliers and removing from database.

Data Cleaning - Missing Values

Strategies to handle missing data:

- **Ignore the tuple**: This is usually done when the class label is missing. This method is not very effective, unless the tuple contains several attributes with missing values. By ignoring the tuple, we do not make use of the remaining attributes in the tuple. Such data could have been useful to the task at hand.

- **Fill in the missing value manually**: In general, this approach is time consuming and may not be feasible given a large data set with many missing values.

- **Use a global constant to fill in the missing value**: Replace all missing attribute values by the same constant such as a label like “Unknown” or $-\infty$. If missing values are replaced by, say, “Unknown,” then the mining program may mistakenly think that they form an interesting concept, since they all have a value in common. Hence, although this method is simple, it is not foolproof.

Data Cleaning - Missing Values (Cont.)

Strategies to handle missing data:

- **Use a measure of central tendency for the attribute (e.g., the mean or median) to fill in the missing value**: For normal (symmetric) data distributions, the mean can be used, while skewed data distribution should employ the median. We will talk about skewed distribution later in the lecture.

- **Use the most probable value to fill in the missing value**: This may be determined with regression, inference-based tools using a Bayesian formalism, or decision tree induction.

Data Cleaning – Noisy Data

Noise is a random error or variance in a measured variable. some basic statistical description techniques, and methods of data visualization (e.g., boxplots and scatter plots) can be used to identify **outliers**, which may represent noise.

Here are some of the common data smoothing techniques:

- **Binning**: Binning methods smooth a sorted data value by consulting its "neighborhood," that is, the values around it.
  - first sort data and partition into (equal-frequency, or equal-width) bins
  - then one can smooth by bin mean, smooth by bin median, smooth by bin boundaries, etc.

- **Regression**: Data smoothing can also be done by regression, a technique that conforms data values to a function.

- **Outlier analysis**: Outliers may be detected by clustering, for example, where similar values are organized into groups, or "clusters." Intuitively, values that fall outside of the set of clusters may be considered outliers.

Binning methods for data smoothing

- **Smoothing by bin means:** each value in a bin is replaced by the mean value of the bin. Similarly, smoothing by bin **medians** can be employed, in which each bin value is replaced by the bin median.

- **Smoothing by bin boundaries**: the minimum and maximum values in a given bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value.

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:

Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

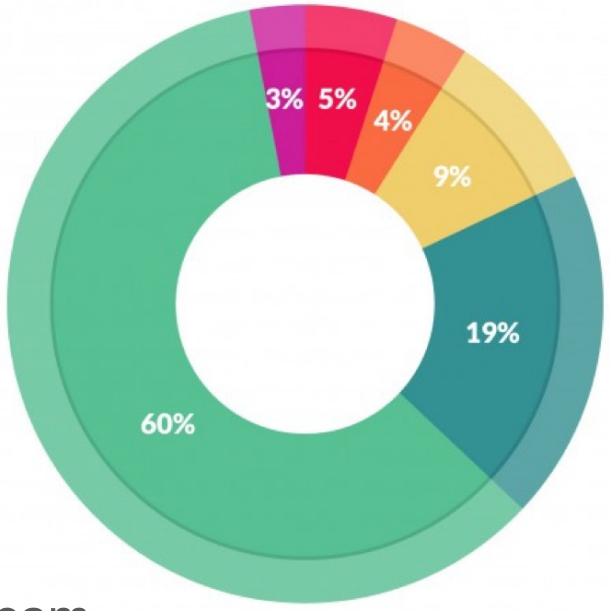
Smoothing by bin boundaries:

Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

Importance of Data Cleaning

“80%” Time Spent on Data Preparation

Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says [Forbes]
http://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#73bf5b137f75



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

Data Integration

To be as efficient and effective for various data analyses as possible, data from various sources commonly needs to be integrated. The following steps describe how to integrate multiple databases or files.

- Combine data from multiple sources into a coherent storage place (e.g., a single file or a database).

- Engage in schema integration, or the combining of metadata from different sources.

- Detect and resolve data value conflicts. For example:
  - A conflict may arise; for instance, such as the presence of different attributes and values from various sources for the same real-world entity.
  - Reasons for this conflict could be different representations or different scales; for example, metric vs. British units.

- Address redundant data in data integration. Redundant data is commonly generated in the process of integrating multiple databases. For example:
  - The same attribute may have different names in different databases.
  - One attribute maybe a “derived” attribute in another table; for example, annual revenue.
  - Correlation analysis may detect instances of redundant data.

Data Transformation

Some of the typical data manipulation and transformation techniques are:

- Reduction
  - Data Cube Aggregation: use the smallest representation that is sufficient to address the given task.

    For example, suppose you have the data of All Electronics sales per quarter for the year 2018 to the year 2022. If you want to get the annual sale per year, you just have to aggregate the sales per quarter for each year.

  - Dimensionality Reduction: dimensionality reduction method works with respect to the nature of the data. Here, a dimension or a column in your data spreadsheet is referred to as a "feature," and the goal of the process is to identify which features to remove or collapse to a combined feature. (later in this course)

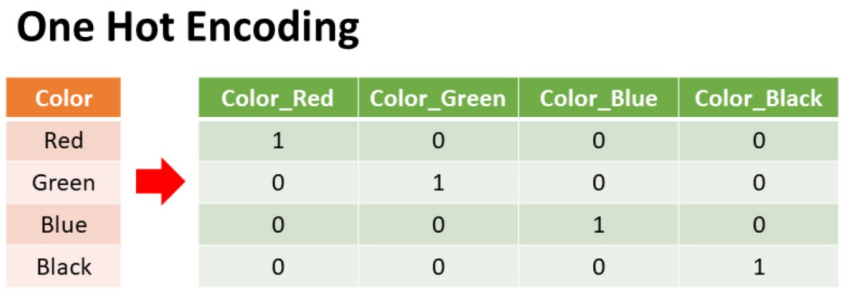
- Discretization: There are three types of attributes involved in discretization:
  - Nominal: Values from an unordered set. E.g., colors
  - Ordinal: Values from an ordered set. E.g., grades in a mark sheet.
  - Continuous: Real numbers

  To achieve discretization, divide the range of continuous attributes into intervals. For instance, we could decide to split the range of temperature values into cold, moderate, and hot, or the price of company stock into above or below its market valuation.

Data Transformation

Some of the typical data manipulation and transformation techniques are:

- Conversion
  - One Hot Encoding



  - Feature Construction

    New attributes constructed from the given ones

Data Transformation

Some of the typical data manipulation and transformation techniques are:

- Normalization: Scaled to fall within a small, specified range and aggregation. Some of the techniques that are used for accomplishing normalization
  - min-max

$$v' = \frac{v - min_A}{max_A - min_A}(newMax_A - newMin_A) + newMin_A$$

  - Z-score

$$v' = \frac{v - \mu_A}{\sigma_A}$$

  - Scaling

$$v' = \frac{v}{10^j}$$

Data Transformation: Normalization

- Since the range of values of raw data varies widely, in some *machine learning* algorithms, objective functions will not work properly without normalization.

- For example, the majority of *classifiers* calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature.

- Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.

- Another reason why feature scaling is applied is that *gradient descent* converges much faster with feature scaling than without it.

Data Transformation: Min-Max Normalization

Min-Max Normalization:

$$v' = \frac{v - min_A}{max_A - min_A}(newMax_A - newMin_A) + newMin_A$$

- works as a linear transformation to a given range.
- In this approach, the data is scaled to a fixed range - usually 0 to 1.
- The technique keeps relationship among original data.
- The cost of having this bounded range - in contrast to standardization - is that we will end up with smaller standard deviations, which can suppress the effect of outliers.
- Two of the problems associated with this approach
  - The value range should be known and predefined both in training and in real world applications
  - The closest the distribution is to a linear form the better it will work

Data Transformation: z-Score Normalization

- The result of standardization (or Z-score normalization) is that the features will be rescaled so that they'll have the properties of a standard normal distribution with μ = 0 and σ = 1

$$v' = \frac{v - \mu_A}{\sigma_A}$$

- Standardizing the features so that they are centered around 0 with a standard deviation of 1 is not only important if we are comparing measurements that have different units, but it is also a general requirement for many machine learning algorithms.
  - Gradient descent as a prominent example (an optimization algorithm often used in logistic regression, SVMs, perceptrons, neural networks etc

- This method will work good assuming the following:
  - The data are normally distributed
  - The mean and sigma can be considered the same for training and real world data

Z-score standardization or Min-Max scaling?

- There is no obvious answer to this question: it really depends on the application.

- In clustering analyses, standardization may be especially crucial in order to compare similarities between features based on certain distance measures.

- Another prominent example is the Principal Component Analysis, where we usually prefer standardization over Min-Max scaling, since we are interested in the components that maximize the variance

- However, this doesn't mean that Min-Max scaling is not useful at all! A popular application is image processing, where pixel intensities have to be normalized to fit within a certain range (i.e., 0 to 255 for the RGB color range). Also, typical neural network algorithm require data that on a 0-1 scale.

Data Transformation: Normalization by decimal scaling

- Normalizes by moving the decimal point of values of feature X. The number of decimal points moved depends on the maximum absolute value of X. A modified value corresponding to v is obtained using

$$v' = \frac{v}{10^j}$$

  - Usually the normalization factor is chosen such that max(|new_v|) < 1.

- This approach needs (similarly to min – max method) the maximum value to be predefined

- Example: suppose the range of attribute X is -500 to 45. The maximum absolute value of X is 500. To normalize by decimal scaling we will divide each value by 1,000 (j = 3). In this case, -500 becomes -0.5 while 45 will become 0.045.

Data Analysis and Data Analytics

- These two terms – **data analysis** and **data analytics** – are often used interchangeably and could be confusing.
- However, there are some subtle but important differences between analysis and analytics.

Data Analysis:

> It is hands-on data exploration and evaluation
> It looks backwards, providing historical view of what has happened
> It is not involved in prediction of what will happen

Data Analytics:

> It is a broader term and includes data analysis
> It is the science behind the analysis
> It is concerned with the entire methodology

Categorization of Analysis and Analytics

Data Analysis	Data Analytics
Descriptive Analysis	Diagnostic Analytics
Exploratory Analysis	Predictive Analytics
Mechanistic Analysis	Prescriptive Analytics

Descriptive Analysis

- Descriptive analysis is about: “What is happening now based on incoming data.” It is a method for quantitatively describing the main features of a collection of data.
    - Example: categorize customers by their likely product preferences and purchasing patterns.
- Typically, it is the first kind of data analysis performed on a dataset.
- Usually it is applied to large volumes of data.
- Description and interpretation processes are different steps.
- Descriptive statistics come into play to facilitate analyzing and summarizing the data (e.g. mean, median, or mode, …. etc.)

Summary Statistics/Quantitative Methods

- **Central Tendency measures**. They are computed to give a “center” around which the measurements in the data are distributed.

- **Variation or Variability measures**. They describe “data spread” or how far away the measurements are from the center.

- **Relative Standing measures**. They describe the relative position of specific measurements in the data.

Central Tendency Measures - Mean

- Mean is commonly known as average, though they are not exactly synonyms. Mean is most often used to measure the central tendency of continuous data as well as a discrete dataset. If there are n number of values in a dataset and the values are x1, x2, . . ., xn, then the mean is calculated as

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}.$$

- There is a significant drawback to using the mean as a central statistic: it is susceptible to the influence of outliers. Also, mean is only meaningful if the data is normally distributed, or at least close to looking like a normal distribution.

- The mean is useful for predicting future results when there are no extreme values in the data set

Central Tendency Measures - Median

- Median is the middle value of the data that has been sorted according to the values of the data.
- When the data has even number of values, median is calculated as the average of the two middle values.
- Typically, median is less susceptible to the presence of outliers (compared to mean).
- Example

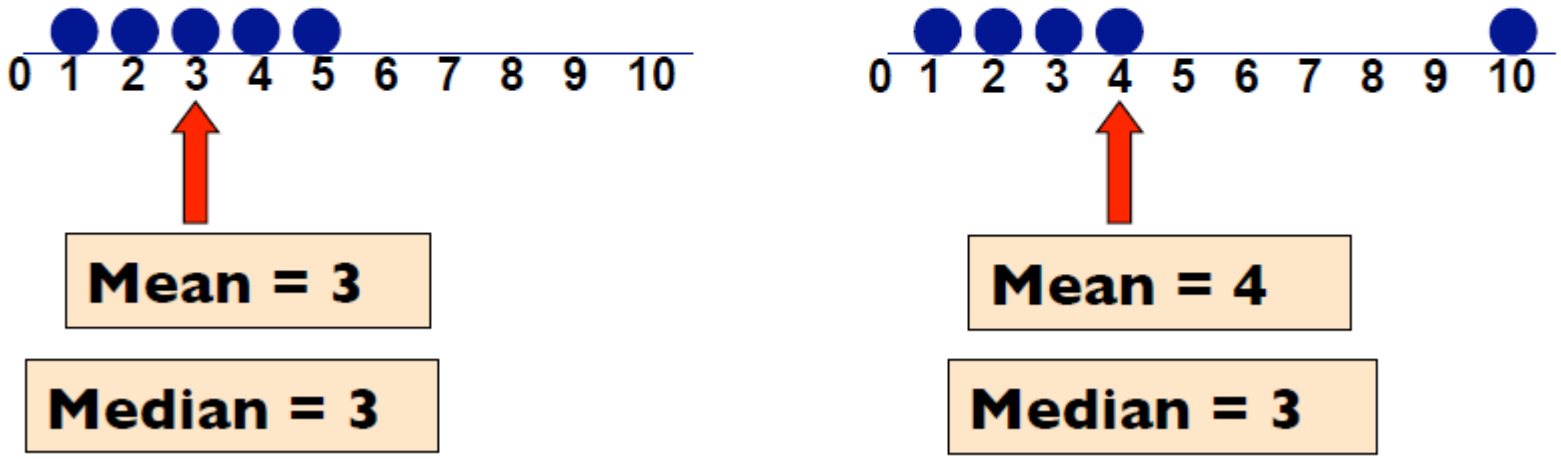
Some data:
Age of participants: 17 19 21 22 23 23 23 38

Median = (22+23)/2 = 22.5

Which Location Measure Is Best?

- Mean is best for symmetric distributions without outliers.

- Median is useful for skewed distributions or with outliers.



Central Tendency Measures - Mode

- Mode is the most frequently occurring value in a dataset.

- Example: Find the mode of the following data set:

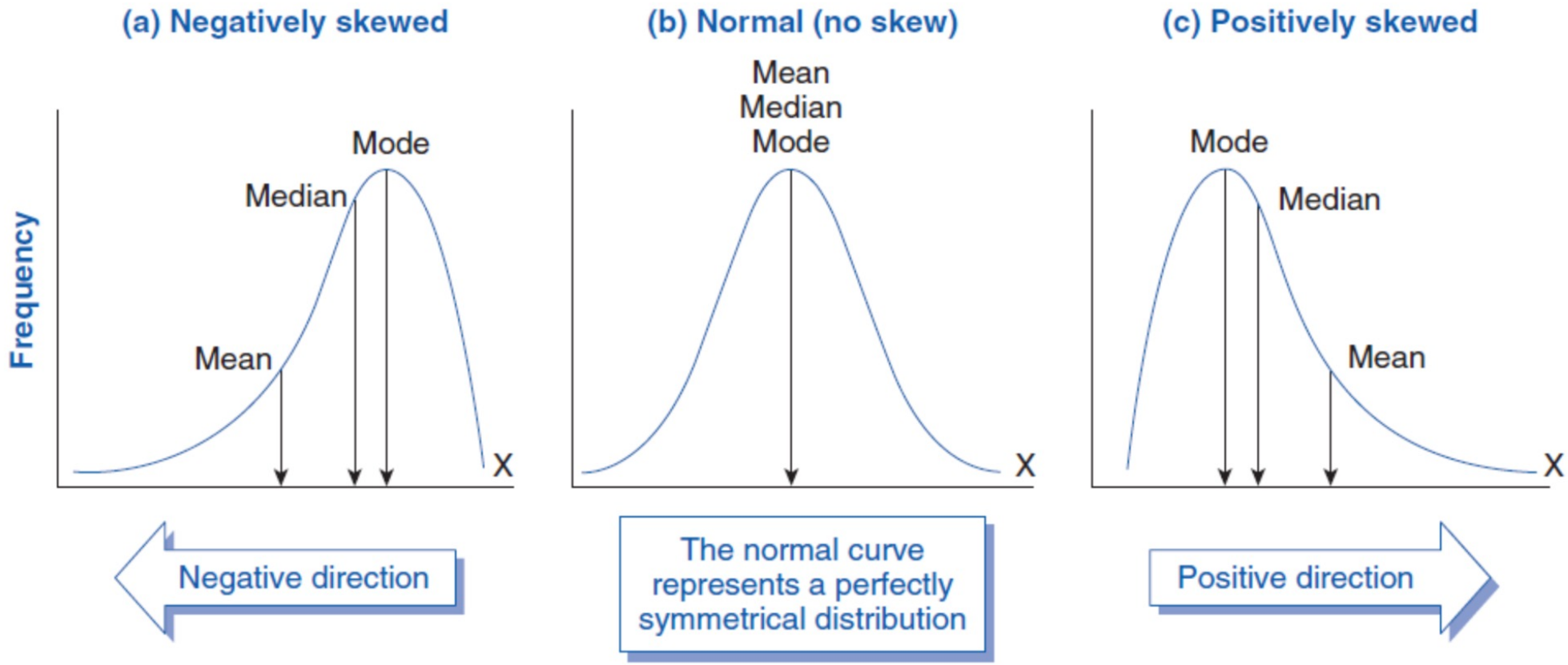
  48 44 48 45 42 49 48

  Solution: The mode is 48 since it occurs most often.

- Typically, mode is used for non-numerical data.

- The mode is useful when the most common item, characteristic or value of a data set is required.

Mean vs. Median vs. Mode



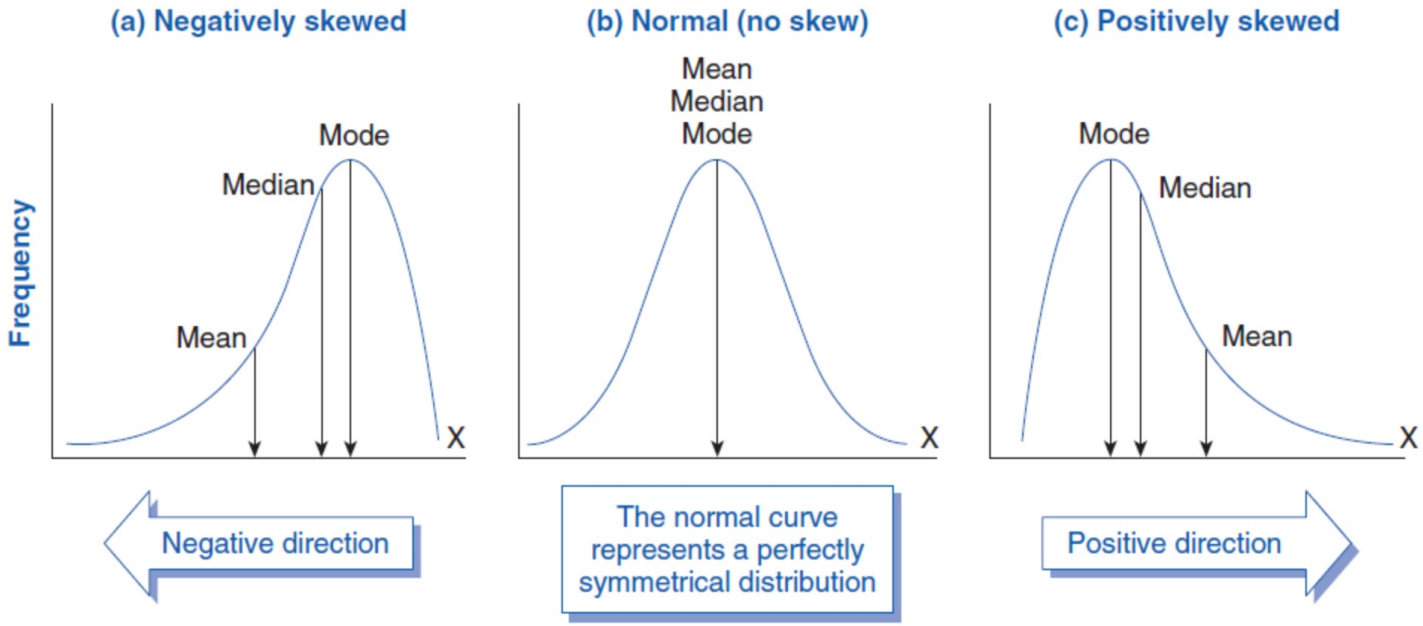
Measures of Shape: Skewness

Skewness

- Absence of symmetry
- Extreme values in one side of a distribution

- There are many measures for skewness, one used measure is the Pearson's second skewness coefficient (median skewness)

$$S = \frac{3\,(mean - median)}{standard\ deviation}$$



- If S < 0, the distribution is negatively skewed (skewed to the left).
- If S = 0, the distribution is symmetric (not skewed).
- If S > 0, the distribution is positively skewed (skewed to the right).

Variation Measures - Variance

- It is a measure used to indicate how spread out the data points are.
- If the individual observations vary greatly from the group mean, then the variance is big; and vice versa.
- The variance of the population is defined by the following formula:

$$\sigma^2 = \frac{\sum(X_i - \overline{X})^2}{N}$$

Where N is the number of elements in the population

- The variance of the sample is defined by the following formula:

$$s^2 = \frac{\sum(x_i - \overline{x})^2}{n-1}$$

Where n is the number of elements in the sample

Variation Measures - Standard Deviation

- It is the square root of the variance.

- It is computed as:

$$s = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n-1}}$$

Where n is the number of elements in the sample

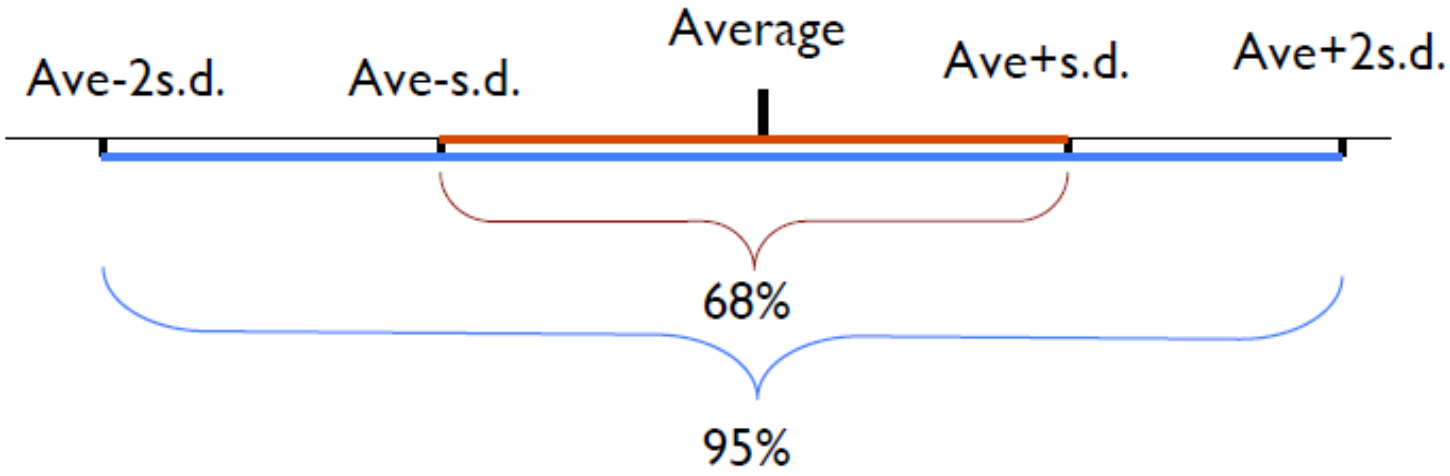
- The advantage: the units of standard deviation is same as the units of the data, which is not the case for the variance.

Standard Deviation: Interesting Theoretical Result

For many lists of observations, especially if their histogram is bell-shaped

- Roughly 68% of the observations lie within 1 standard deviation of the mean.

- 95% of the observations lie within 2 standard deviations of the mean.



Diagnostic Analytics

- Diagnostic analytics are used for discovery, or to determine why something happened.
- Sometimes this type of analytics is also known as causal analysis.
- It involves at least one cause (usually more than one) and one effect.
- There are many techniques available in diagnostic analytics, which should be capable of recognizing patterns, detecting anomalies, surfacing ‘unusual’ events.
- correlation is one of the most frequently used technique in diagnostic analysis.

Diagnostic Analytics - Correlations

- Correlation is a statistical analysis that is used to measure and describe the strength and direction of the relationship between two variables.
- Strength indicates how closely two variables are related to each other, and direction indicates how one variable would change its value as the value of the other variable changes.
- An important statistic, the **Pearson's r correlation**, is widely used to measure the degree of the relationship between linear related variables. The following formula is used to calculate the Pearson's r correlation:

$$r = \frac{N\sum xy - \sum x \sum y}{\sqrt{\left[N\sum x^2 - \left(\sum x\right)^2\right]\left[N\sum y^2 - \left(\sum y\right)^2\right]}},$$

where

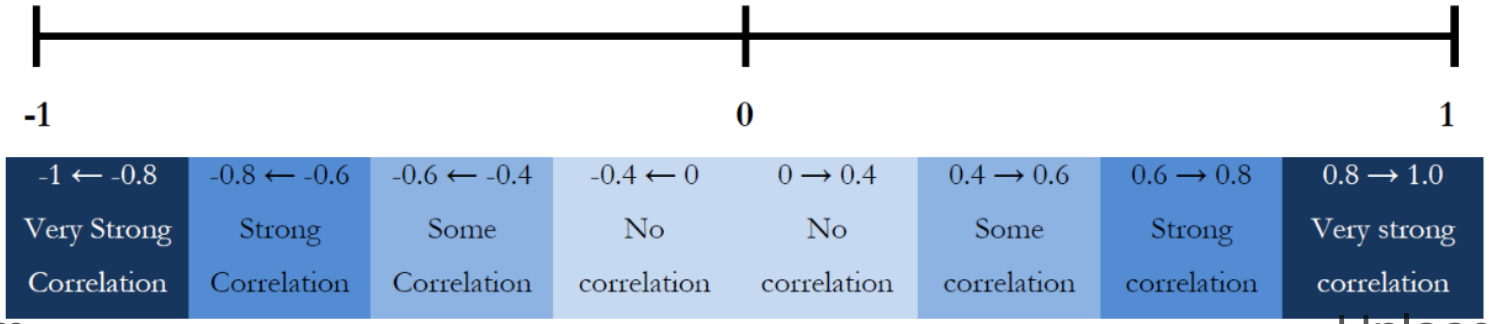
r = Pearson's r correlation coefficient,
N = number of values in each dataset,
$\sum xy$ = sum of the products of paired scores,
$\sum x$ = sum of x scores,
$\sum y$ = sum of y scores,

Diagnostic Analytics - Correlations

$$r = \frac{N\sum xy - \sum x \sum y}{\sqrt{\left[N\sum x^2 - \left(\sum x\right)^2\right]\left[N\sum y^2 - \left(\sum y\right)^2\right]}},$$

- **Directions**: All correlation coefficients between 0 and 1 represent positive correlations, while all coefficients between 0 and -1 are negative correlations. Positive relation means if one increase(decrease), then other will increase(decrease). Negative relation means if one increase(decrease), then other will decrease(increase).

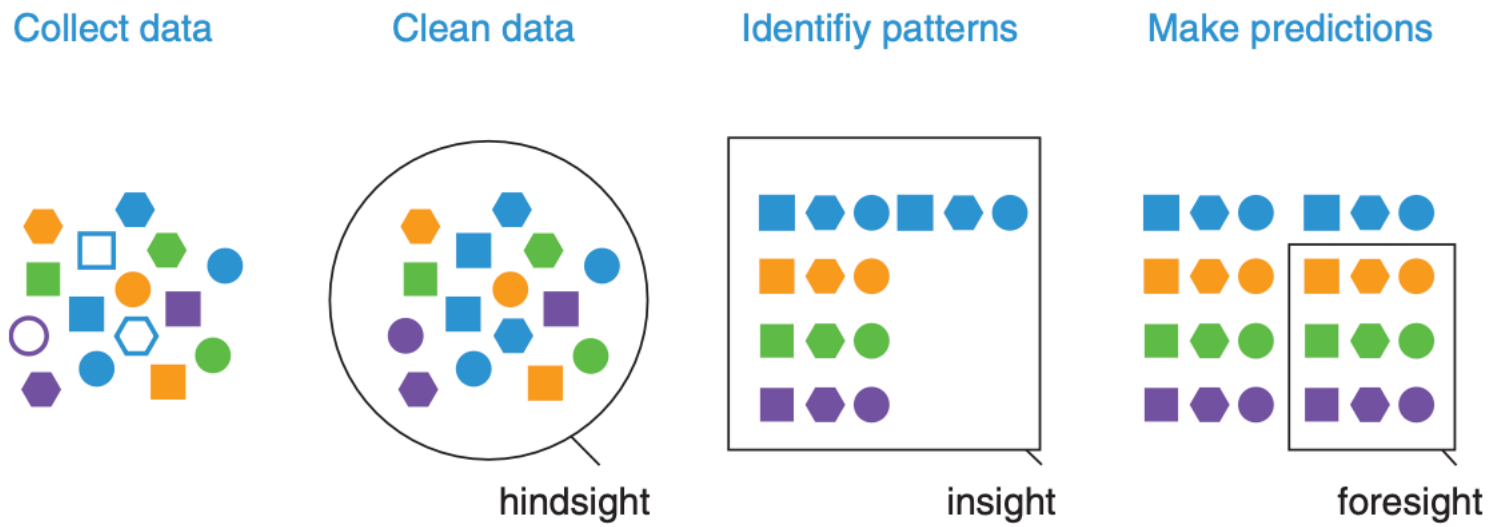
- **Strength**: The closer a correlation coefficient is to 1 or to -1, the stronger it is. Following picture suggests a guideline to interpret the strength.



Predictive Analytics

- **Predictive analytics** has its roots in our ability to predict what might happen. These analytics are about understanding the future using the data and the trends we have seen in the past, as well as emerging new contexts and processes.

- Predictive analytics is done in stages.



Prescriptive Analytics

- **Prescriptive analytics** is the area of business analytics dedicated to finding the best course of action for a given situation.

- It involves the following steps:
  - Analyze potential decisions and their interactions
  - Analyze the influences that bear upon these decisions
  - Prescribe an optimal course of action in real time

- Specific techniques used in prescriptive analytics include optimization, simulation, game theory, and decision-analysis methods.

Exploratory Analysis

- Often when working with data, we may not have a clear understanding of the problem or the situation. And yet, we may be called on to provide some insights. In other words, we are asked to provide an answer without knowing the question! This is where we go for an exploration.

- Exploratory analysis is an approach to analyzing datasets to find previously unknown relationships.

- It typically involves data visualization techniques. The idea is that, plotting the data in different forms could provide us with some clues regarding what we may (or want to) find in data. It consists of range of techniques. The most common goal is looking for patterns in the data.

- Exploratory data analysis is an approach that postpones the usual assumptions about what kind of model the data follows with the more direct approach of allowing the data itself to reveal its underlying structure in the form of a model.

Mechanistic Analysis

- Mechanistic analysis involves understanding the exact changes in variables that lead to changes in other variables for individual objects.

- For instance, studying how the increased amount of CO2 in the atmosphere is causing the overall temperature to change.

- Such relationships are often explored using regression.

Introduction to Data Visualization

What is data visualization

- Data visualization means encoding information about the data in a graphical way.
- It helps to highlight the most useful insights from a dataset, making it easier to spot trends, patterns, outliers, and correlations.
- The ultimate goal of data visualization is to tell a story. We are trying to convey information about the data as efficiently as possible.
- Using graphics allows the reader to quickly and accurately understand the message we are trying to transmit.

Benefits of effective data visualization

Data visualization allows you to:

- Get an initial understanding of your data by making trends, patterns, and outliers easily visible to the naked eye.
- Comprehend large volumes of data quickly and efficiently.
- Communicate insights and findings to non-data experts, making your data accessible and actionable.
- Tell a meaningful and impactful story, highlighting only the most relevant information for a given context.

[Source: Emily Stevens]

Benefits of effective data visualization (Cont.)

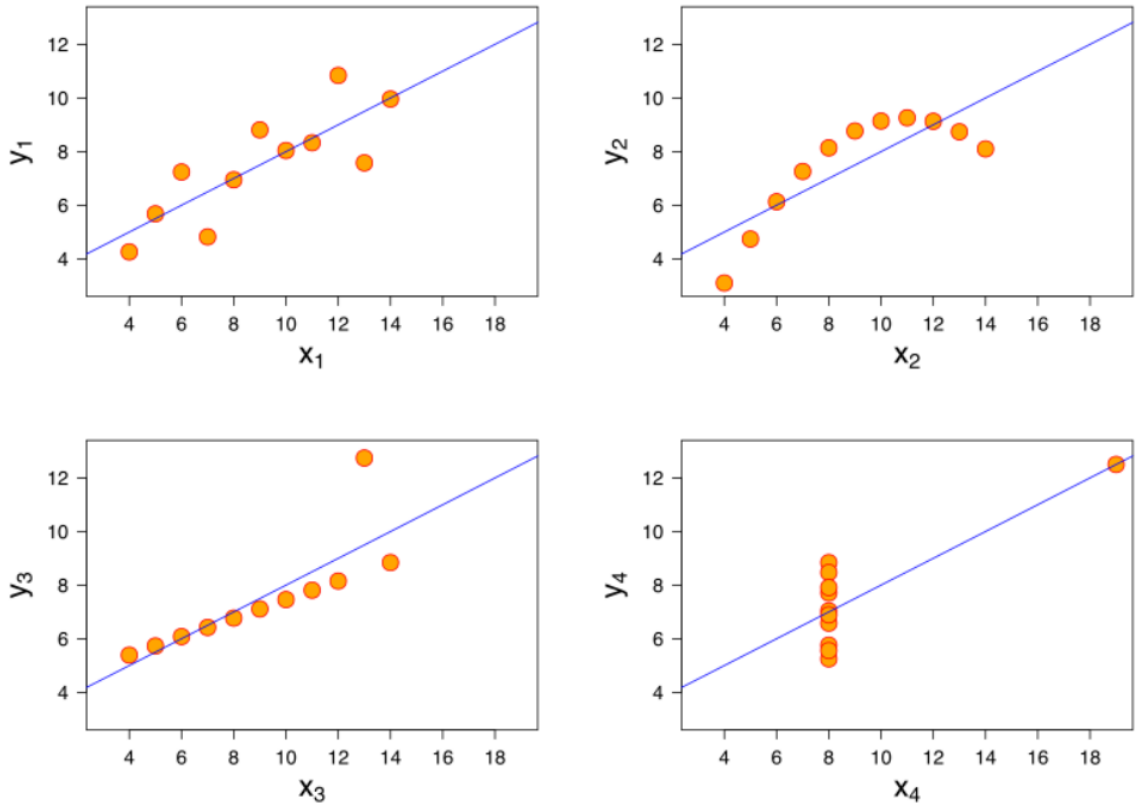
A good example of this is "Anscombe's quartet", four datasets that share the same descriptive statistics, including mean, variance, and correlation.

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

[Source: https://guides.library.jhu.edu/datavisualization]

Benefits of effective data visualization (Cont.)

Upon visual inspection, it becomes immediately clear that these datasets, while seemingly identical according to common summary statistics, are each unique. This is the power of effective data visualization: it allows us to bypass cognition by communicating directly with our perceptual system.



[Source: https://guides.library.jhu.edu/datavisualization]

What is data visualization used for?

Within the broader goal of conveying key insights, different visualizations can be used to tell different stories. Data visualizations can be used to:

- **Convey changes over time**: For example, a line graph could be used to present how the value of Bitcoin changed over a certain time period.
- **Determine the frequency of events**: You could use a histogram to visualize the frequency distribution of a single event over a certain time period (e.g. number of internet users per year from 2007 to 2021).
- **Highlight interesting relationships or correlations between variables**: If you wanted to highlight the relationship between two variables (e.g. marketing spend and revenue, or hours of weekly exercise vs. cardiovascular fitness), you could use a scatter plot to see, at a glance, if one increases as the other decreases (or vice versa).
- **Examine a network**: If you want to understand what's going on within a certain network (for example, your entire customer base), network visualizations can help you to identify (and depict) meaningful connections and clusters within your network of interest.
- **Analyze value and risk**

[Source: Emily Stevens]

Types of data visualization

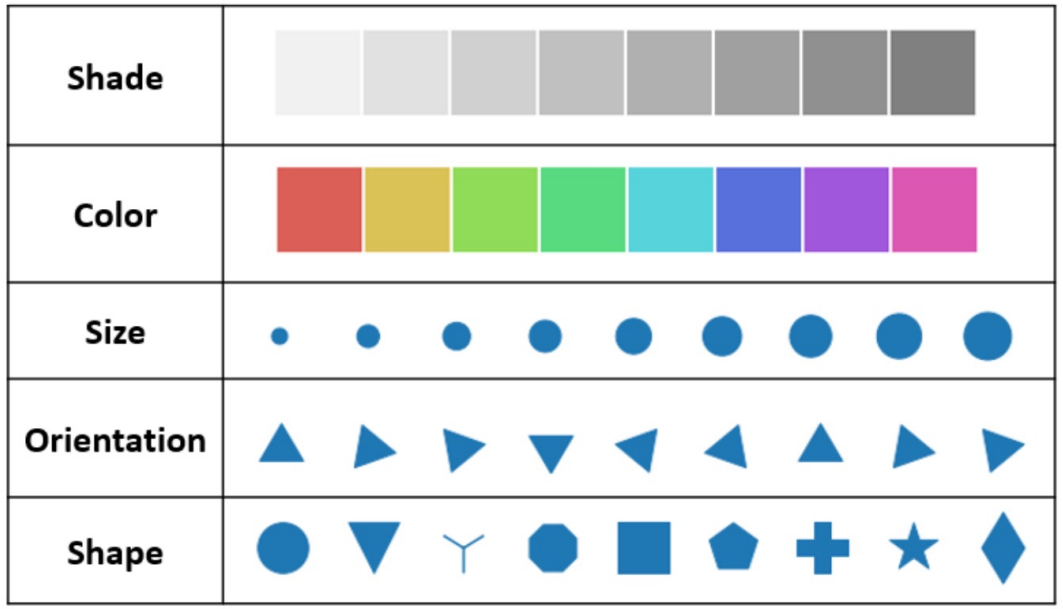
- It is important to distinguish between two main types: exploratory and explanatory data visualization.
- In a nutshell, exploratory data visualization helps you figure out what's in your data, while explanatory visualization helps you to communicate what you've found.
- Exploration takes place while you're still analyzing the data, while explanation comes towards the end of the process when you're ready to share your findings.

[Source: Emily Stevens]

Visual encodings

There are many ways to encode information. In data visualization we use visual encodings.

- Retinal Visual Encodings: These encodings are quickly picked up by our retina. Ex.: shade, color, size, …etc.
  - You could use all of them in one chart but it would be hard for the reader to grasp all the information quickly.
  - 1 or 2 retinal encodings per chart is optimal.
  - Think about what encodings are best for what kind of
    For example, color would work pretty badly for a continuous variable, but would work very well for a discrete variable

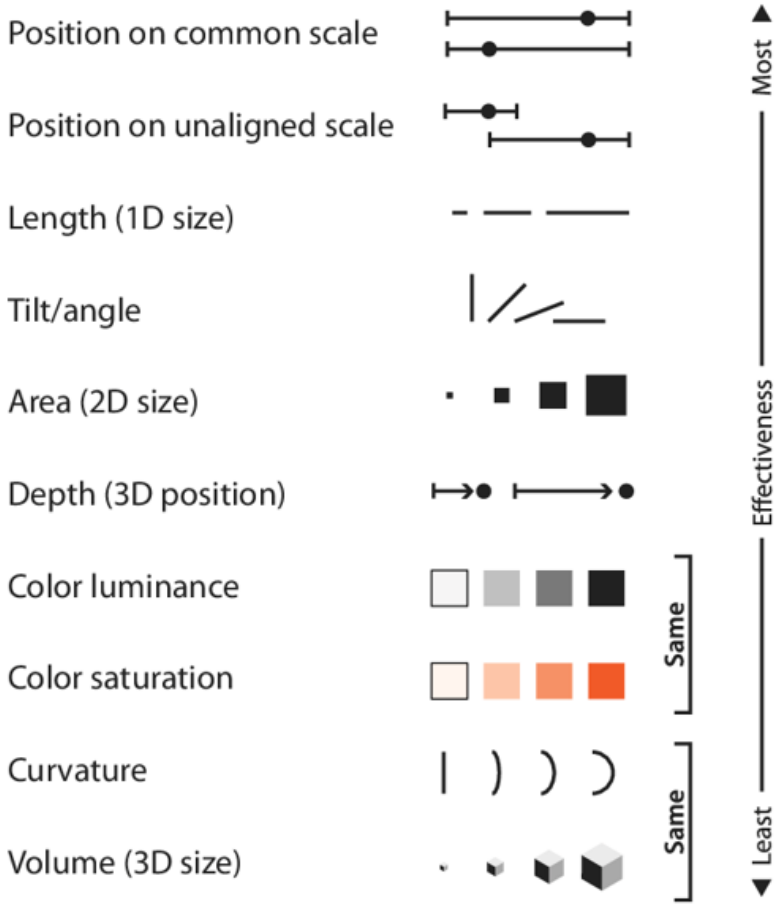


- Other Visual Encodings: For example “spatial” encodings exploit the cortex’s spatial awareness to encode information. This can be achieved through position in a scale, length, area, volume.

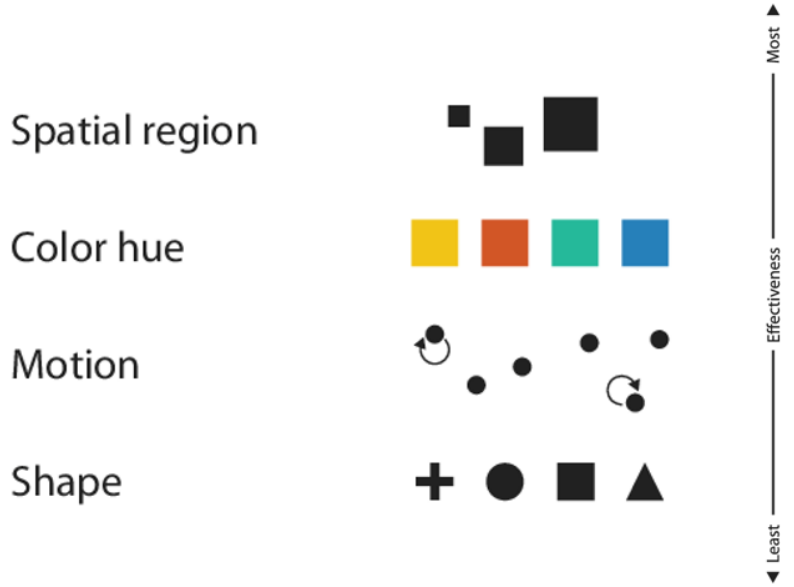
[Source: Diego Unzueta]

What's the best encoding?

- It depends on what you're trying to achieve. The following charts show the effectiveness of encodings for continuous and discrete variables:



Continuous Variables



Discrete Variables

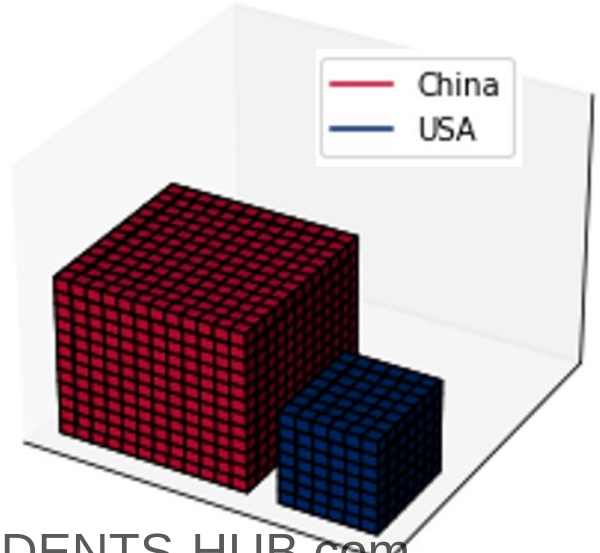
[Source: Diego Unzueta]

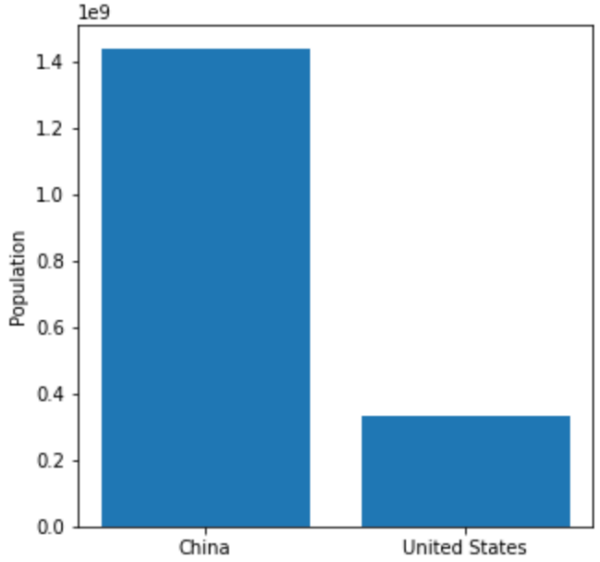
What's the best encoding?

- Example: population of China and USA

In the 1st figure, The population is encoded in volume, whilst category is encoded in color. Color works very well here for the discrete variable, however, volume is a terrible choice because it is really difficult for the reader to know exactly what population each country has.

Now let's look at a bar chart in the 2nd figure. Here we are using 2 encodings for the continuous variable, and o 1 encoding for the discrete. For the continuous variable (population), the encodings used are the position in a common scale and the length. For the discrete variable (class membership), we are using spatial region instead of color, which works best.





[Source: Diego Unzueta]

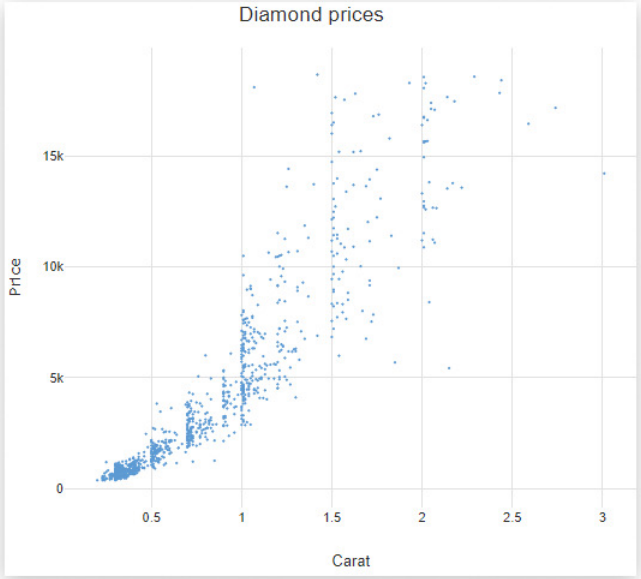
Five data visualization categories

- **Temporal data visualizations** are linear and one-dimensional. Examples include scatterplots, timelines, and line graphs.
- **Hierarchical visualizations** organize groups within larger groups, and are often used to display clusters of information. Examples include tree diagrams, ring charts, and sunburst diagrams.
- **Network visualizations** show the relationships and connections between multiple datasets. Examples include matrix charts, word clouds, and node-link diagrams.
- **Multidimensional or 3D visualizations** are used to depict two or more variables. Examples include pie charts, Venn diagrams, stacked bar graphs, and histograms.
- **Geospatial visualizations** convey various data points in relation to physical, real-world locations (for example, voting patterns across a certain country). Examples include heat maps, cartograms, and density maps.

[Source: Emily Stevens]

Common types of data visualization - Scatterplots

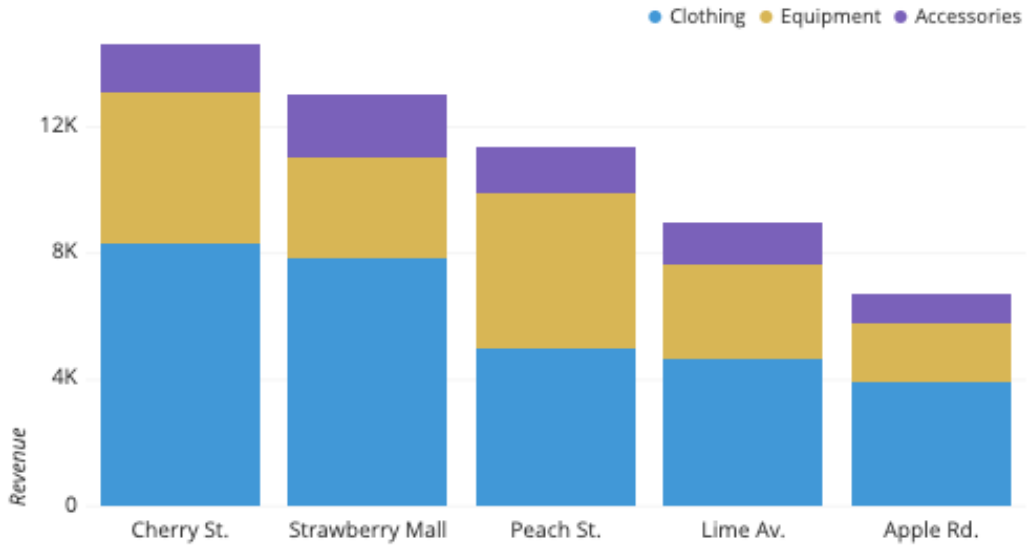
- Scatterplots (or scatter graphs) visualize the relationship between two variables. One variable is shown on the x-axis, and the other on the y-axis, with each data point depicted as a single “dot” or item on the graph. This creates a “scatter” effect, hence the name.
- Scatterplots are best used for large datasets when there’s no temporal element. For example, if you wanted to visualize the relationship between a person’s height and weight, or between how many carats a diamond measures and its monetary value, you could easily visualize this using a scatterplot. It’s important to bear in mind that scatterplots simply describe the correlation between two variables; they don’t infer any kind of cause-and-effect relationship.



[Source: Emily Stevens]

Common types of data visualization - Bar charts

- Bar charts are used to plot categorical data against discrete values. Categorical data refers to data that is not numeric, and it's often used to describe certain traits or characteristics. Some examples of categorical data include things like education level (e.g. high school, undergrad, or post-grad) and age group (e.g. under 30, under 40, under 50, or 50 and over).



- So, with a bar chart, you have your categorical data on the x-axis plotted against your discrete values on the y-axis. The height of the bars is directly proportional to the values they represent, making it easy to compare your data at a glance.

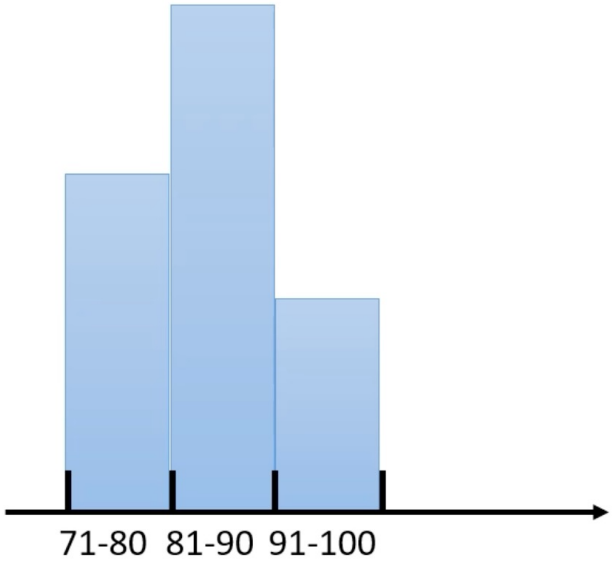
[Source: Emily Stevens]

Common types of data visualization - Histograms

- Plot values of observation on one of the axes (typically x-axis). The values can be ordered (if applicable).
- Plot perpendicular bars to the above axis.
- The height of the bar shows how many times each value occurred in the dataset.
- When you have numeric values, it does not make sense to count the occurrences of each value. Thus, we introduce the concept of buckets or bins.
- The idea is to plot the bars for each bin/bucket, where the height of the bar indicates the number of values that belongs to the bin/bucket.

Stu ID	Score
1	75
2	81
3	95
4	85
5	81
6	82
7	97
8	89
9	76
10	100
11	77
12	79
13	87
14	87
15	79
16	75
17	100
18	82
19	83
20	98
21	86
22	87
23	71
24	78
25	81

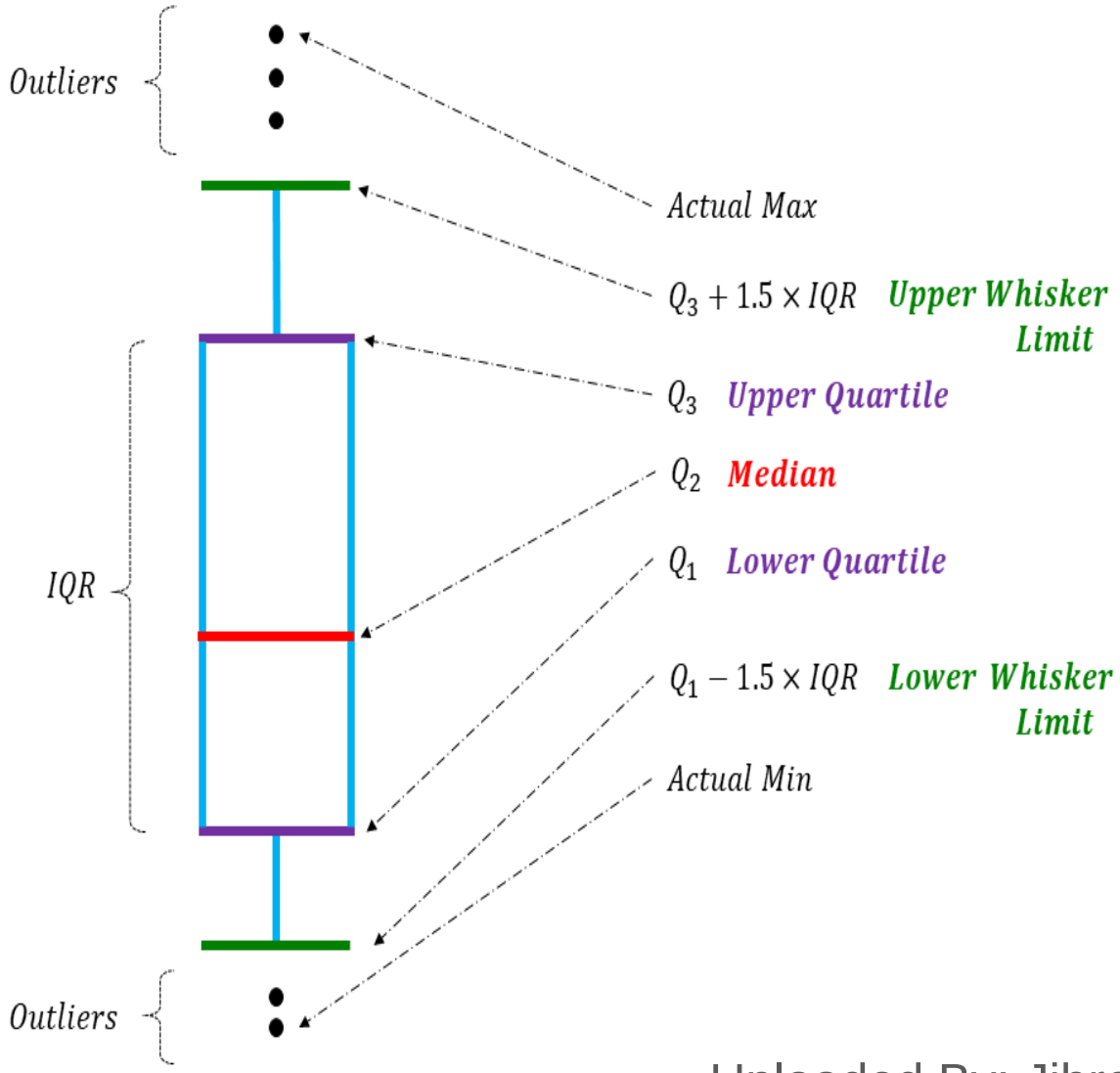
Bin	Occurrence
71-80	8
81-90	12
91-100	5



Common types of data visualization - Box Plots

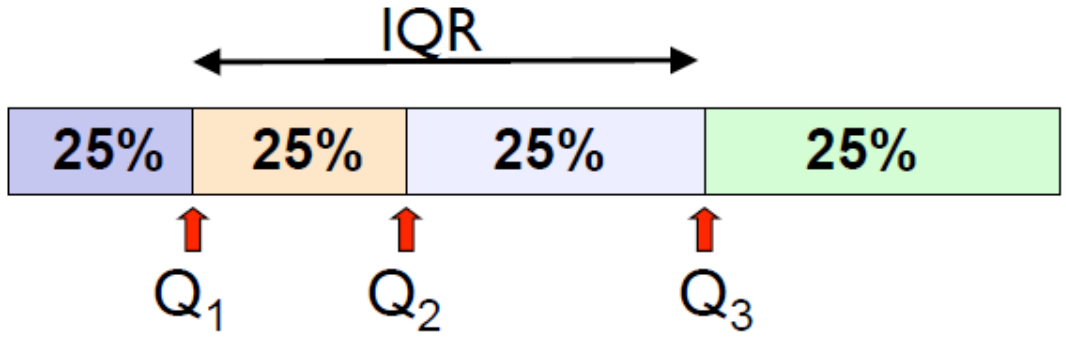
To build a boxplot, following things are required from the data.

- The min and max values in the data.
- The first and third quartile of the data.
- The median of the data.



Quartiles and IQR

- The quartiles of a population or a sample are the three values which divide the distribution or observed data into even fourths.
  - The first quartile, Q1, is the value for which 25% of the observations are smaller and 75% are larger
  - Q2 is the same as the median (50% are smaller, 50% are larger)
  - Only 25% of the observations are greater than the third quartile



- The IQR is used when the researcher wishes to eliminate the influence of extreme values and consider the variation for the more typical cases in a distribution.

Quartiles and IQR - Example

Find the quartiles of this data set: 6, 47, 49, 15, 43, 41, 7, 39, 43, 41, 36

- You first need to arrange the data points in increasing order.

idx	1	2	3	4	5	6	7	8	9	10	11
value	6	7	15	36	39	41	41	43	43	47	49

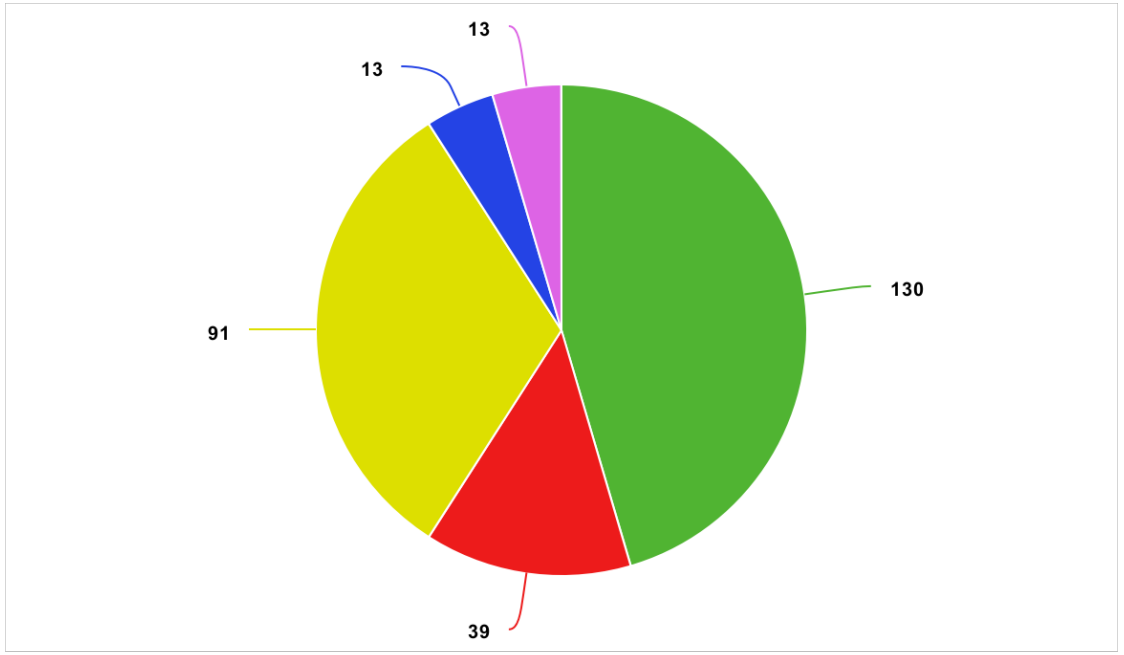
- Then you need to find the rank of the median (Q2) to split the data set in two.

  rank = (n+1)/2 = (11+1)/2 = 6 → Q2 = 41

- Then you need to split the lower half of the data in two again to find the lower quartile. The lower quartile will be the point of rank (5 + 1) ÷ 2 = 3. The result is Q1 = 15. The second half must also be split in two to find the value of the upper quartile. The rank of the upper quartile will be 6 + 3 = 9. So Q3 = 43.
- Once you have the quartiles, you can easily measure the spread. The interquartile range will be Q3 - Q1, which gives 28 (43-15).

Common types of data visualization - Pie charts

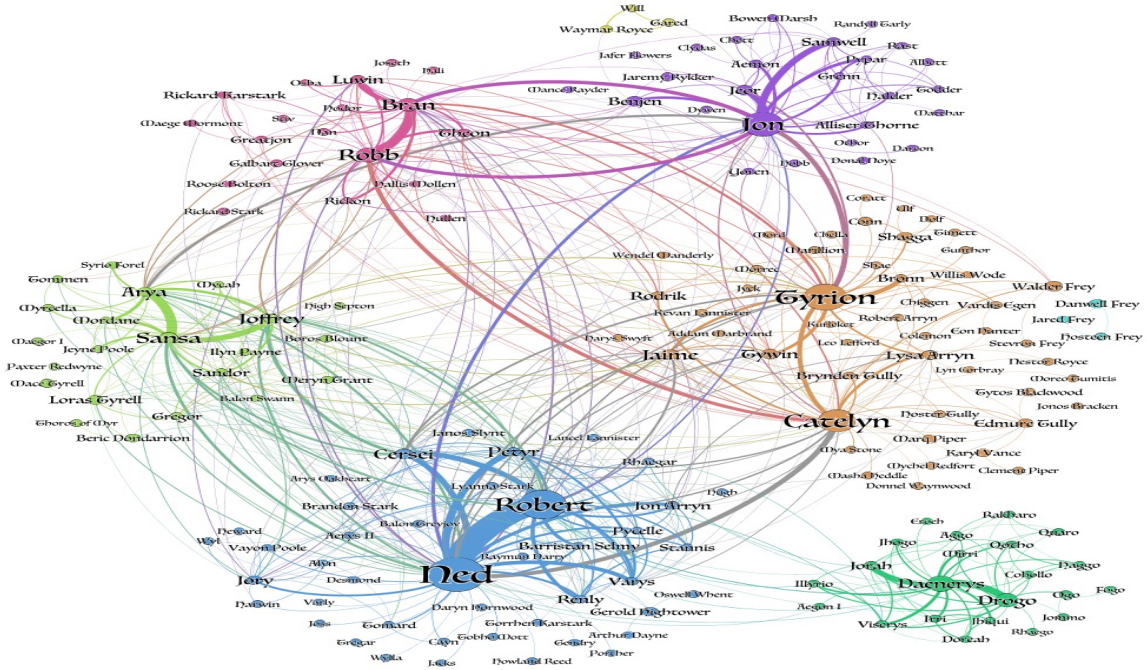
- Just like bar charts, pie charts are used to visualize categorical data. However, while bar charts represent multiple categories of data, pie charts are used to visualize just one single variable broken down into percentages or proportions.
- A pie chart is essentially a circle divided into different “slices,” with each slice representing the percentage it contributes to the whole. Thus, the size of each pie slice is proportional to how much it contributes to the whole “pie.”



[Source: Emily Stevens]

Common types of data visualization - Network graphs

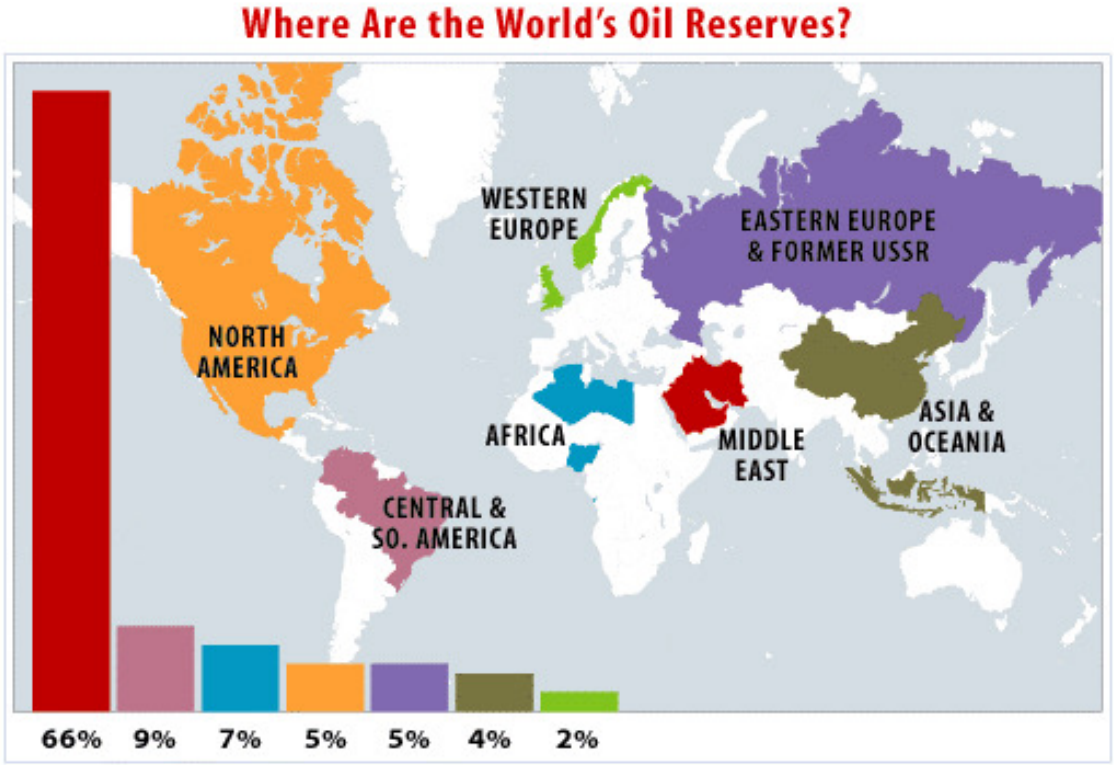
- Network graphs show how different elements or entities within a network relate to one another, with each element represented by an individual node. Nodes are connected to other, related nodes via lines.
- Network graphs are great for spotting and representing clusters within a large network of data. Let's imagine you have a huge database filled with customers, and you want to segment them into meaningful clusters for marketing purposes. You could use a network graph to draw connections and parallels between all your customers or customer groups. With any luck, certain clusters and patterns would emerge, giving you a logical means by which to group your audience.



[Source: Emily Stevens]

Common types of data visualization - Geographical maps

- Geo maps are used to visualize the distribution of data in relation to a physical, geographical area. For example, you could use a color-coded map to see how natural oil reserves are distributed across the world, or to visualize how different states voted in a political election.



[Source: Emily Stevens]

Top data visualization tools and libraries

Tools

- Power BI
- Tableau
- …

Python Packages

- Seaborn
- Matplotlib
- Pandas
- Plotly
- …