Statistical Learning

Introduction to Statistics

Gurumoorthy Pattabiraman

Outline

1. Understanding AI & ML
2. What is Statistics?
3. Why Statistics?
4. Business Statistics-Tools
5. Types of Statistics - Descriptive and Inferential Statistics
6. Data Sources and Types of Datasets
7. Attributes of Datasets
8. Key Takeaways
9. Key Steps in Business Analytics

Understanding AI & ML

Artificial Intelligence (AI) and **Machine Learning (ML)** are two very hot buzzwords right now, and often seem to be used interchangeably.

They are not quite the same thing, but the perception that they are can sometimes lead to some confusion.

AI - Definition in short
Artificial Intelligence is the broader concept of machines being able to carry out tasks in a way that we would consider “smart”.

ML - Definition in short
Machine Learning is a current application of AI based around the idea that we should really just be able to give machines access to data and let them learn for themselves.

Artificial Intelligence

Artificial Intelligence has been around for a long time – the Greek myths contain stories of mechanical men designed to mimic our own behaviour!

Work in the field of AI concentrates on mimicking human decision making processes and carrying out tasks in ever more human ways.

Often classified into one of two fundamental groups – applied or general.

Applied AI

Applied AI is far more common – systems designed to intelligently trade stocks and shares, or manoeuvre an autonomous vehicle would fall into this category.

Generalized AI

Generalized AIs – systems or devices which can in theory handle any task – are less common, but this is where some of the most exciting advancements are happening today. It is also the area that has led to the development of Machine Learning. Often referred to as a subset of AI, it’s really more accurate to think of it as the current state-of-the-art.

Rise of Machine Learning

Two important breakthroughs led to the emergence of Machine Learning as the vehicle which is driving AI development forward with the speed it currently has.

One of these was the realization – credited to Arthur Samuel in 1959 – that rather than teaching computers everything they need to know about the world and how to carry out tasks, it might be possible to teach them to learn for themselves.

The second, more recently, was the emergence of the internet, and the huge increase in the amount of digital information being generated, stored, and made available for analysis.

Once these innovations were in place, engineers realized that rather than teaching computers and machines how to do everything, it would be far more efficient to code them to think like human beings, and then plug them into the internet to give them access to all of the information in the world.

Neural Networks

The development of neural networks has been key to teaching computers to think and understand the world in the way we do, while retaining the innate advantages they hold over us such as speed, accuracy and lack of bias.

A Neural Network is a computer system designed to work by classifying information in the same way a human brain does. It can be taught to recognize, for example, images, and classify them according to elements they contain.

Essentially it works on a system of probability – based on data fed to it, it is able to make statements, decisions or predictions with a degree of certainty. The addition of a feedback loop enables "learning" – by sensing or being told whether its decisions are right or wrong, it modifies the approach it takes in the future.

Machine Learning Applications

Machine Learning applications can read text and work out whether the person who wrote it is making a complaint or offering congratulations. They can also listen to a piece of music, decide whether it is likely to make someone happy or sad, and find other pieces of music to match the mood. In some cases, they can even compose their own music expressing the same themes, or which they know is likely to be appreciated by the admirers of the original piece.

Another field of AI – Natural Language Processing (NLP) – has become a source of hugely exciting innovation in recent years, and one which is heavily reliant on ML.

NLP

NLP applications attempt to understand natural human communication, either written or spoken, and communicate in return with us using similar, natural language. ML is used here to help machines understand the vast nuances in human language, and to learn to respond in a way that a particular audience is likely to comprehend.

What is Statistics?

Your company has created a new drug that may cure arthritis. How would you conduct a test to confirm the drug's effectiveness?

The latest sales data have just come in, and your boss wants you to prepare a report for management on places where the company could improve its business. What should you look for? What should you not look for?

You and a friend are at a baseball game, and out of the blue he offers you a bet that neither team will hit a home run in that game. Should you take the bet?

You want to conduct a poll on whether your school should use its funding to build a new athletic complex or a new library. How many people do you have to poll? How do you ensure that your poll is free of bias? How do you interpret your results?

A widget maker in your factory that normally breaks 4 widgets for every 100 it produces has recently started breaking 5 widgets for every 100. When is it time to buy a new widget maker? (And just what is a widget, anyway?)

Statistics - Defined

Statistics, in short, is the study of data[1]. It includes **descriptive statistics** (the study of methods and tools for collecting data, and mathematical models to describe and interpret data) and **inferential statistics** (the systems and techniques for making probability-based decisions and accurate predictions based on incomplete (sample) data).

Classic Definition of Statistics

“ By Statistics, we mean methods specially adopted to the elucidation of quantitative data affected to a marked extent by multiplicity of causes”.

Yule and Kendal

It is interesting to see what *Thomas Davenport* means by Business Analytics and note the similarities and dissimilarities between the two.

“Business Analytics (BA) can be defined as the broad use of data and quantitative analysis for decision making within organizations”.

Why Statistics is So Important?

Three significant events triggered the current meteoric growth in the use of analytics in decision making and *Statistics is the Heart of Analytics*

Event1

- Technological developments, Revolution of Internet and social networks, data generated from mobile phones and other electronic devices, produce large amount of data from which insights will have to be sifted.
- The discovery of pattern and trends from these data for organizations will pave the way for improving profitability, understanding customer expectations, and appropriately pricing their products so that they can gain competitive advantage in the marketplace.

Why Statistics is So Important?

Event 2

- Advances in enormous computing power to effectively process and analyze massive amounts of data

- Sophisticated and faster algorithms for solving problems

- Data Visualization for Business Intelligence

Why Statistics is So Important?

Event 3

- Large data storage capability
- Parallel computing, and cloud computing coupled with better computer hardware have enabled businesses to solve large scale problems faster than ever before without sacrificing

Big Data

Big data

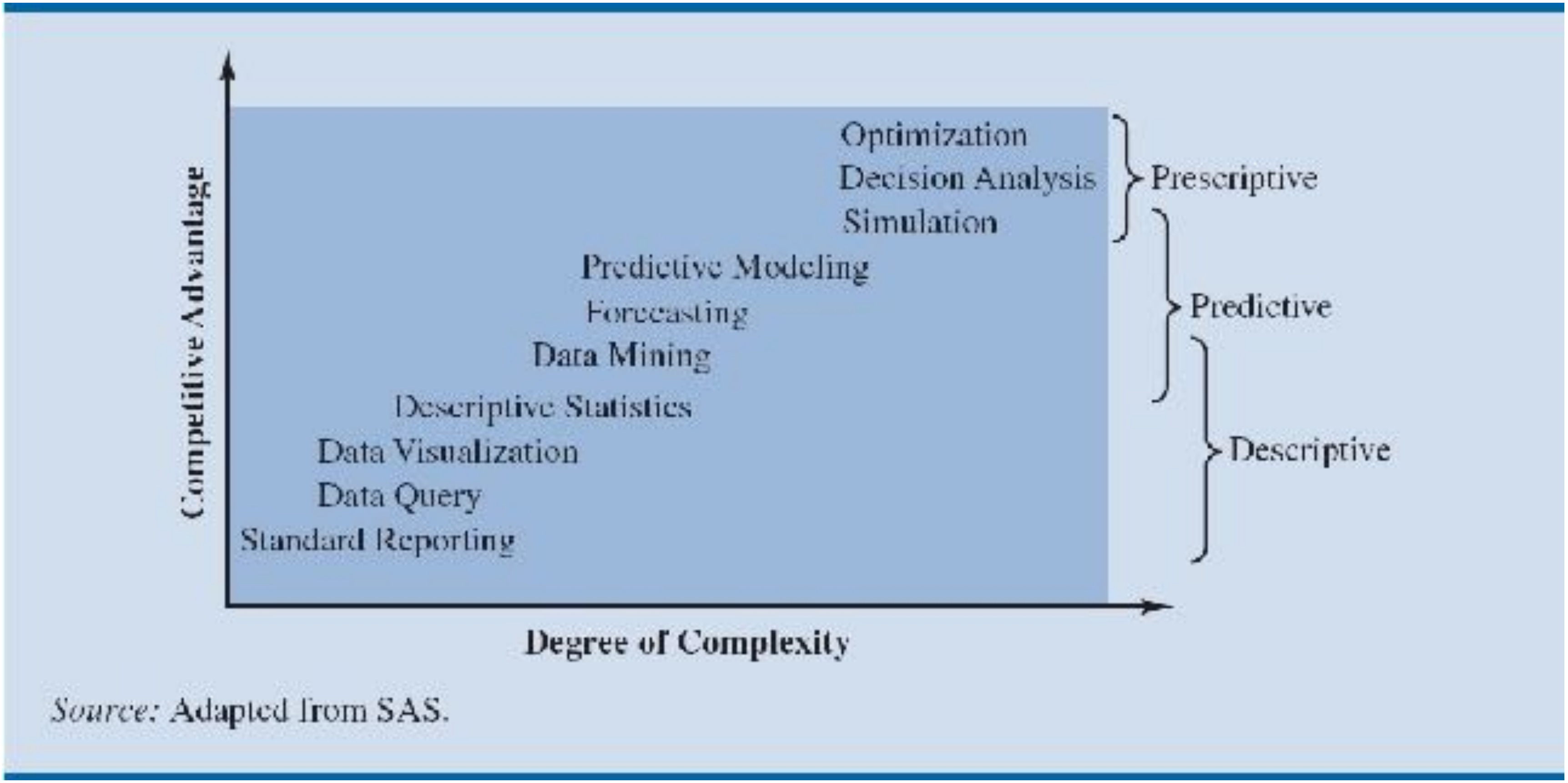
- A set of data that cannot be managed, processed, or analyzed with traditional software/algorithms within a reasonable amount of time.
- Big data revolves around

Volume Velocity VarietyValue Veracity

Walmart handles over one million purchase transactions per hour.

Facebook processes more than 250 million picture uploads per day.

Statistics in The Spectrum of Business Analytics



Source: Adapted from SAS.

Design of Experiments

One of the most neglected aspects of statistics—and maybe the single greatest reason that Statisticians drink—is Experimental Design.

So often a scientist will bring the results of an important experiment to a statistician and ask for help analyzing results only to find that a flaw in the experimental design rendered the results useless. So often we statisticians have researchers come to us hoping that we will somehow magically "rescue" their experiments.

Design of Experiments

A friend provided me with a classic example of this. In his psychology class he was required to conduct an experiment and summarize its results. He decided to study whether music had an impact on problem solving. He had a large number of subjects (myself included) solve a puzzle first in silence, then while listening to classical music and finally listening to rock and roll. He measured how long it would take to complete each of the tasks and then summarized the results.

What my friend failed to consider was that the results were highly impacted by a *learning effect* he hadn't considered. The first puzzle always took longer because the subjects were first learning how to work the puzzle. By the third try (when subjected to rock and roll) the subjects were much more adept at solving the puzzle, thus the results of the experiment would seem to suggest that people were much better at solving problems while listening to rock and roll!

Design of Experiments

The simple act of randomizing the order of the tests would have isolated the "learning effect" and in fact, a well-designed experiment would have allowed him to measure both the effects of each type of music *and* the effect of learning. Instead, his results were meaningless. A careful experimental design can help preserve the results of an experiment, and in fact some designs can save huge amounts of time and money, maximize the results of an experiment, and sometimes yield additional information the researcher had never even considered!

Sampling

Similar to the Design of Experiments, the study of sampling allows us to find a most effective statistical design that will optimize the amount of information we can collect while minimizing the level of effort. Sampling is very different from experimental design however. In a laboratory we can design an experiment and control it from start to finish. But often we want to study something outside of the laboratory, over which we have much less control.

Bias

- While sampling is a more cost effective method of determining a result, small samples or samples that depend on a certain selection method will result in a bias within the results.
  The following are common sources of bias:

  - Sampling bias or statistical bias, where some individuals are more likely to be selected than others (such as if you give equal chance of cities being selected rather than weighting them by size)

  - Systemic bias, where external influences try to affect the outcome (e.g. funding organi- zations wanting to have a specific result)

Modern Regression

Regression models relate variables to each other in a linear fashion. For example, if you recorded the heights and weights of several people and plotted them against each other, you would find that as height increases, weight tends to increase too. You would probably also see that a straight line through the data is about as good a way of approximating the relationship as you will be able to find, though there will be some variability about the line. Such linear models are possibly the most important tool available to statisticians. They have a long history and many of the more detailed theoretical aspects were discovered in the 1970s. The usual method for fitting such models is by "least squares" estimation, though other methods are available and are often more appropriate, especially when the data are not normally distributed.

What happens, though, if the relationship is not a straight line? How can a curve be fit to the data? There are many answers to this question. One simple solution is to fit a quadratic relationship, but in practice such a curve is often not flexible enough. Also, what if you have many variables and relationships between them are dissimilar and complicated?

Modern regression methods aim at addressing these problems. Methods such as generalized additive models, projection pursuit regression, neural networks and boosting allow for very general relationships between explanatory variables and response variables, and modern computing power makes these methods a practical option for many applications

Classification

Some things are different from others. How? That is, how are objects classified into their respective groups? Consider a bank that is hoping to lend money to customers. Some customers who borrow money will be unable or unwilling to pay it back, though most will pay it back as regular repayments. How is the bank to classify customers into these two groups when deciding which ones to lend money to?

The answer to this question no doubt is influenced by many things, including a customer’s income, credit history, assets, already existing debt, age and profession. There may be other influential, measurable characteristics that can be used to predict what kind of customer a particular individual is. How should the bank decide which characteristics are important, and how should it combine this information into a rule that tells it whether or not to lend the money?

This is an example of a classification problem, and statistical classification is a large field containing methods such as linear discriminant analysis, classification trees, neural networks and other methods.

Time Series

Many types of research look at data that are gathered over time, where an observation taken today may have some correlation with the observation taken tomorrow. Two prominent examples of this are the fields of finance (the stock market) and atmospheric science.

We’ve all seen those line graphs of stock prices as they meander up and down over time. Investors are interested in predicting which stocks are likely to keep climbing (i.e. when to buy) and when a stock in their portfolio is falling. It is easy to be misled by a sudden jolt of good news or a simple "market correction" into inferring—incorrectly—that one or the other is taking place!

In meteorology scientists are concerned with the venerable science of predicting the weather. Whether trying to predict if tomorrow will be sunny or determining whether we are experiencing true climate changes (i.e. global warming) it is important to analyze weather data over time.

Survival Analysis

Suppose that a pharmaceutical company is studying a new drug which it is hoped will cause people to live longer (whether by curing them of cancer, reducing their blood pressure or cholesterol and thereby their risk of heart disease, or by some other mechanism). The company will recruit patients into a clinical trial, give some patients the drug and others a placebo, and follow them until they have amassed enough data to answer the question of whether, and by how long, the new drug extends life expectancy.

Such data present problems for analysis. Some patients will have died earlier than others, and often some patients will not have died before the clinical trial completes. Clearly, patients who live longer contribute informative data about the ability (or not) of the drug to extend life expectancy. So how should such data be analyzed?

Survival analysis provides answers to this question and gives statisticians the tools necessary to make full use of the available data to correctly interpret the treatment effect.

Categorical Analysis

In laboratories we can measure the weight of fruit that a plant bears, or the temperature of a chemical reaction. These data points are easily measured with a yardstick or a ther- mometer, but what about the color of a person’s eyes or her attitudes regarding the taste of broccoli? Psychologists can’t measure someone’s anger with a measuring stick, but they can ask their patients if they feel "very angry" or "a little angry" or "indifferent". Entirely different methodologies must be used in statistical analysis from these sorts of experiments. Categorical Analysis is used in a myriad of places, from political polls to analysis of census data to genetics and medicine.

Clinical Trials

In the United States, the FDA[3] requires that pharmaceutical companies undergo rigorous procedures called Clinical Trials[4] and statistical analyses to assure public safety before allowing the sale of use of new drugs.

In fact, the pharmaceutical industry employs more statisticians than any other business!

Types of Statistics

Descriptive Statistics is concerned with Data Summarization, Graphs/Charts, and Tables

Inferential Statistics is a method used to talk about a Population Parameter from a Sample.

Some Key Terms Used in Statistics

Population is the collection of all possible observations of a specified characteristic of interest. An example is all students in the Quantitative Methods course in an MBA program

Parameter is the population characteristic of interest. For example, you are interested in the average income of a particular class of people. The average income of this entire class of people is called a parameter.

Sample is a subset of the population. Suppose you want to select a team of 20 students from 200 students in an MBA program for participating in a management quiz. The 200 students is the total population. 20 students selected for the quiz is the sample.

Statistic is based on a sample to make inferences about the population parameter. If you look at the previous example, the average income in the population can be estimated by the average income based on the sample. This sample average is called a statistic.

Data Sources

Primary Data are collected by the organization itself for a particular purpose. The benefits of primary data are that they fit the needs exactly, are up to date, and reliable.

Secondary Data are collected by other organizations or for other purposes. Any data, which are not collected by the organization for the specified purpose, are secondary data. These may be published by other organizations, available from research studies, published by the government, web, social media and so on.

Types of Data

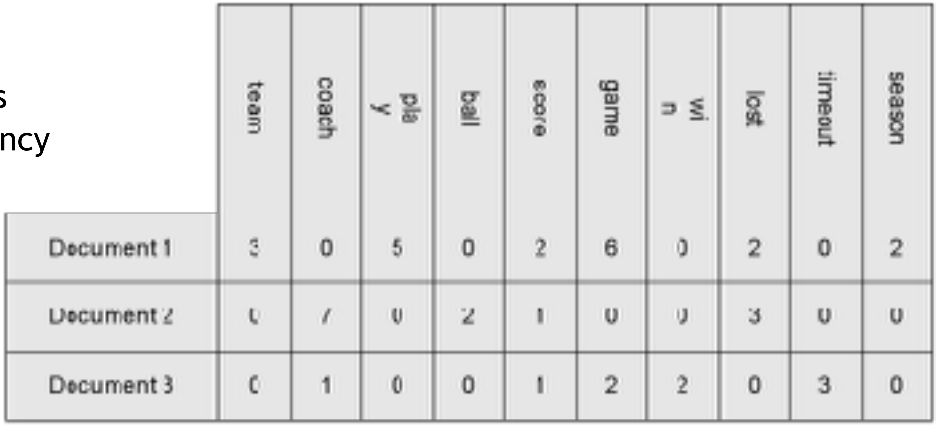
Qualitative Data are nonnumeric in nature and can't be measured. Examples are gender, religion, and place of birth.

Quantitative Data are numerical in nature and can be measured. Examples are balance in your savings bank account, and number of members in your family.

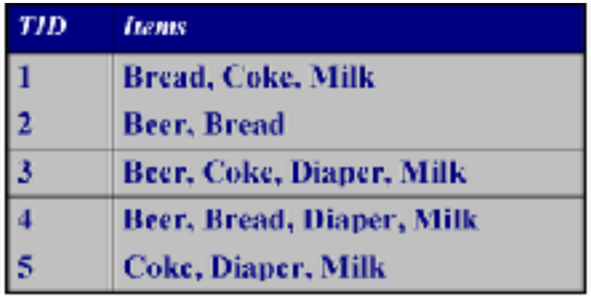
Quantitative data can be classified into discrete type or continuous type. **Discrete type** can take only certain values, and there are discontinuities between values, such as the number of rooms in a hotel, which cannot be in fraction. **Continuous type** can take any value within a specific interval, such as the production quantity of a particular type of paper (measured in kilograms).

Types of Data Sets

- **Record**
  - Relational records
  - Data matrix, e.g., numerical matrix, crosstabs
  - Document data: text documents: term-frequency vector
  - Transaction data
- **Graph and network**
  - World Wide Web
  - Social or information networks
  - Molecular Structures
- **Ordered**
  - Video data: sequence of images
  - Temporal data: time-series
  - Sequential Data: transaction sequences
  - Genetic sequence data
- **Spatial, image and multimedia:**
  - Spatial data: maps
  - Image data
  - Video data



	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0



TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Data Objects

- Data sets are made up of data objects.
- A data object represents an entity.
- Examples:
  - sales database: customers, store items, sales
  - medical database: patients, treatments
  - university database: students, professors, courses
- Also called *samples* , *examples, instances, data points, objects, tuples*.
- Data objects are described by attributes.
- Database rows -> data objects; columns ->attributes.

Attributes

- Attribute (or dimensions, features, variables): a data field, representing a characteristic or feature of a data object.
    - *E.g., customer _ID, name, address*
- Types:
    - Nominal
    - Binary
    - Ordinal
    - Numeric: quantitative
        - Interval-scaled
        - Ratio-scaled

Attribute Types

- Nominal: categories, states, or “names of things”
  - *Hair_color* = {*auburn*, *black*, *blond*, *brown*, *grey*, *red*, *white*}
  - marital status, occupation, ID numbers, zip codes
- Binary
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - e.g., gender
  - Asymmetric binary: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)
- Ordinal
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - *Size* = {*small*, *medium*, *large*}, grades, army rankings

Numeric Attribute Types

- Quantity (integer or real-valued)
- Interval
  - Measured on a scale of equal-sized units
  - Values have order
    - E.g., *temperature in C˚or F˚, calendar dates*
  - No true zero-point
- Ratio
  - Inherent zero-point
  - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K˚ is twice as high as 5 K˚).
    - e.g., *temperature in Kelvin, length, counts, monetary quantities*

Take-aways

- Statistics has become a very important field now.
- We briefly discussed about various business stats tools like classification, pattern recognition, association, and predictive modeling.
- Statistics now has much more focus on predictive and prescriptive aspects.
- Types of datasets, Data types.

The Key Steps in Business Analytics

Golden Rules to follow

1. Frame the Question

- Descriptive - "seeks to summarize a characteristic of a set of data"
- Exploratory - "analyze the data to see if there are patterns, trends, or relationships between variables" (hypothesis generating)
- Inferential - "a restatement of this proposed hypothesis as a question and would be answered by analyzing a different set of data" (hypothesis testing)
- Predictive - "determine the impact on one factor based on other factor in a population - to make a prediction"
- Causal - "asks whether changing one factor will change another factor in a population - to establish a causal link"
- Mechanistic - "establish how the change in one factor results in change in another factor in a population - to determine the exact mechanism"

2. Acquire

Ways to acquire data (typical data source)

- Download from an internal system
- Obtained from client, or other 3rd party
- Extracted from a web-based API
- Scraped from a website
- Extracted from a PDF file
- Gathered manually and recorded

Data Formats

- Flat files (e.g. csv)
- Excel files
- Database (e.g. MySQL)
- JSON
- HDFS (Hadoop)

3. Refine the Data

- Remove e.g. remove redundant data from the data frame
- Derive e.g. State and City from the market field
- Parse e.g. extract date from year and month column
- Other stuff you may need to do to refine are…
- Missing e.g. Check for missing or incomplete data
- Quality e.g. Check for duplicates, accuracy, unusual data

4. Transform Data

- Convert e.g. free text to coded value
- Calculate e.g. percentages, proportion
- Merge e.g. first and surname for full name
- Aggregate e.g. rollup by year, cluster by area
- Filter e.g. exclude based on location
- Sample e.g. extract a representative data
- Summary e.g. show summary stats like mean

5. Explore the Data

- Why do visual exploration?
- Understand Data Structure & Types
- Explore single variable graphs - Quantitative, Categorical
- Explore dual variable graphs - (Q & Q, Q & C, C & C)
- Explore multi variable graphs

We want to first visually explore the data to see if we can confirm some of our initial hypotheses as well as make new hypothesis about the problem we are trying to solve.

For this we will start by loading the data and understanding the data structure of the data frame we have.

6. Model

- The power and limits of models
- Tradeoff between Prediction Accuracy and Model Interpretability
- Assessing Model Accuracy
- Regression models (Simple, Multiple)
- Classification model

7. Insights

- Why do we need to communicate insight?
- Types of communication - Exploration vs. Explanation
- Explanation: Telling a story with data - https://public.tableau.com/views/EarthquakeTrendStoryExample/Earthquakestory?%3AshowVizHome=no
- Exploration: Building an interface for people to find stories - https://shiny.rstudio.com/gallery/movie-explorer.html