Course 1 Module 4

Types of Data Science Questions

The types of data analysis are:

- Descriptive
- Exploratory
- Inferential
- Predictive
- Causal
- Mechanistic

Descriptive Analysis

The goal here is to summarize a set of data. This is useful for early analysis when we receive new data. It can help generate summaries about the samples and their measures. Measures could mean those of central tendency (mean, median, mode) or variability (standard deviation, range, variance). Descriptive analytics is not useful for generalizing results of the analysis to a larger population or drawing conclusions. Descriptions and interpretations of data are different. Generalizations require additional statistical steps.

e.g. Censuses

Exploratory Analysis

The goal of exploratory analysis is to examine or explore the data and find relationships that weren't previously known. Exploratory analyses explore how different measures might be related to each other but do not confirm that relationship is causative. Just because you observed a relationship between two variables during exploratory analysis, it does not mean that one necessarily causes the other. Because of this, exploratory analysis, while useful for discovering new connections, should not be the final say in answering a question. It can allow you to formulate hypotheses and drive the design of future studies and data collection. But exploratory analysis alone should never be used as the final say on why or how data might be related to each other.All exploratory analysis can tell us is that a relationship exist, not the cause.

Inferential Analysis

The goal of inferential analyses is to use a relatively small sample of data to infer or say something about the population at large. Inferential analysis is commonly the goal of statistical modelling. Where you have a small amount of information to extrapolate and generalize that information to a larger group. An inferential analysis typically involves using the data you have to estimate that value in the population, and then give a measure of uncertainty about your estimate. The ability to accurately infer information about the larger population depends heavily on the sampling scheme. If the data you collect is not from a representative sample of the population, the generalizations you infer won't be accurate for the population.

e.g.A study in which a subset of the US population wasn't safe, for their life expectancy given the level of air pollution they experienced.

Predictive analysis

The goal is to use current and historical data to make predictions about future data. Like in inferential analysis, your accuracy and predictions is dependent on measuring the right variables. If you aren't measuring

the right variables to predict an outcome, your predictions aren't going to be accurate. Additionally, there are many ways to build up prediction models with some being better or worse for specific cases. But in general, having more data and a simple model, generally performs well at predicting future outcomes.

The caveat to a lot of the analyses we've looked at so far is we can only seecorrelations and can't get at the cause of the relationships we observe.

Causal analysis The goal of causal analysis is to see what happens to one variable when we manipulate another variable, looking at the cause and effect of the relationship.

Generally, causal analysis are fairly complicated to do with observed data alone. There will always be questions as to whether are these correlation driving your conclusions, or that the assumptions underlying your analysis are valid. More often, causal analysis are applied to the results of randomized studies that were designed to identify causation. Causal analysis is often considered the gold standard in data analysis, and is seen frequently in scientific studies where scientists are trying to identify the cause of a phenomenon. But often getting appropriate data for doing a causal analysis is a challenge. One thing to note about causal analysis is that the data is usually analyzed in aggregate and observed relationships are usually average effects.

e.g. Randomized controlled trials: A trial to examine the effect of a new drug on a treating infants with spinal muscular atrophy. Comparing a sample of infants receiving the drug versus a sample receiving a mock control. They measure various clinical outcomes in the babies and look at how the drug impacts the outcome.

Mechanistic analysis The goal of mechanistic analysis is to understand the exact changes in variables that lead to exact changes in other variables. These analyses are exceedingly hard to use to infer much, except in simple situations or in those that are nicely modeled by deterministic equations. Mechanistic analyses are most commonly applied to physical or engineering sciences. Biological sciences. For example, are far too noisy datasets to use mechanistic analysis. Often, when these analyses are applied,the only noise in the data is measurement error, which can be accounted for. You can generally find examples of mechanistic analysis in material science experiments. They are able to do mechanistic analysis through a careful balance of controlling and manipulating variables with very accurate measures of both those variables and the desired outcome.

Experimental Design

Formulate your question → Design your experiment → Identify problems and sources of error → Collect the data

Some important terminology related to experiments

- Independent variable (factor) is the variable that the experimenter manipulates.It does not depend on other variables being measured, often displayed on the x-axis.
- Dependent variables are those that are expected to change as a result of changes in the independent variable, often displayed on the y-axis.

When designing an experiment, one has to decide the variables to be manipulated to effect changes in other measured variables. Additionally, a hypothesis must be developed.

- Hypothesis is essentially the theory / educated guess on the relationship between the variables and the outcome of the experiment.
- Sample size is the number of experimental subjects you will include in your experiment.

Big Data