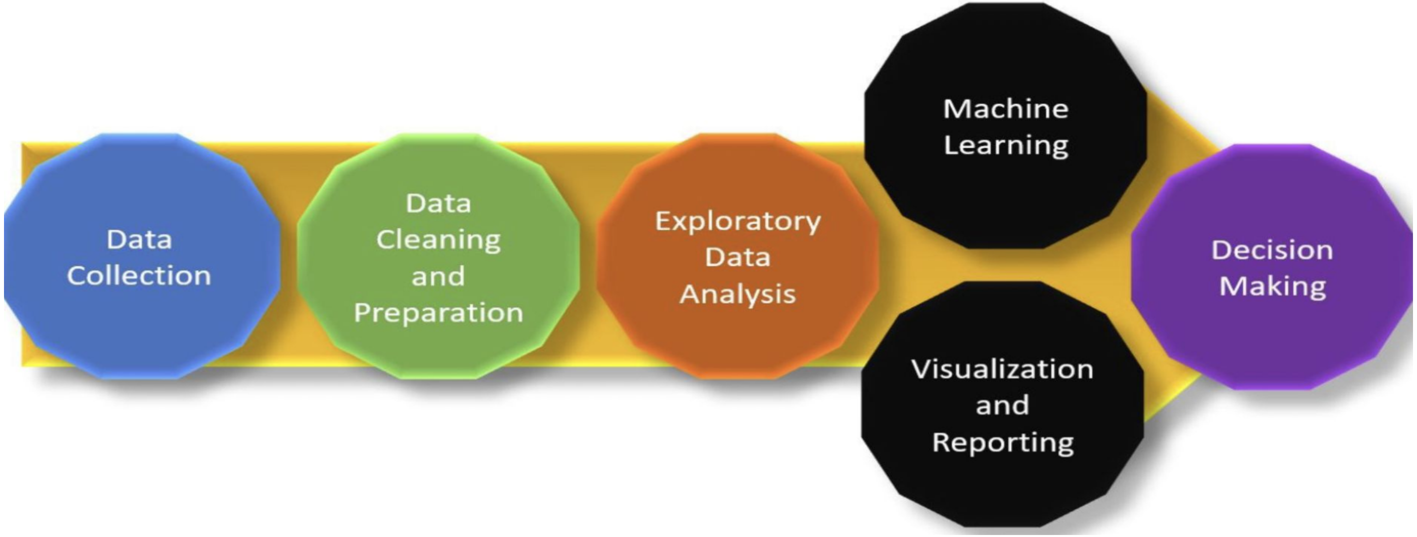
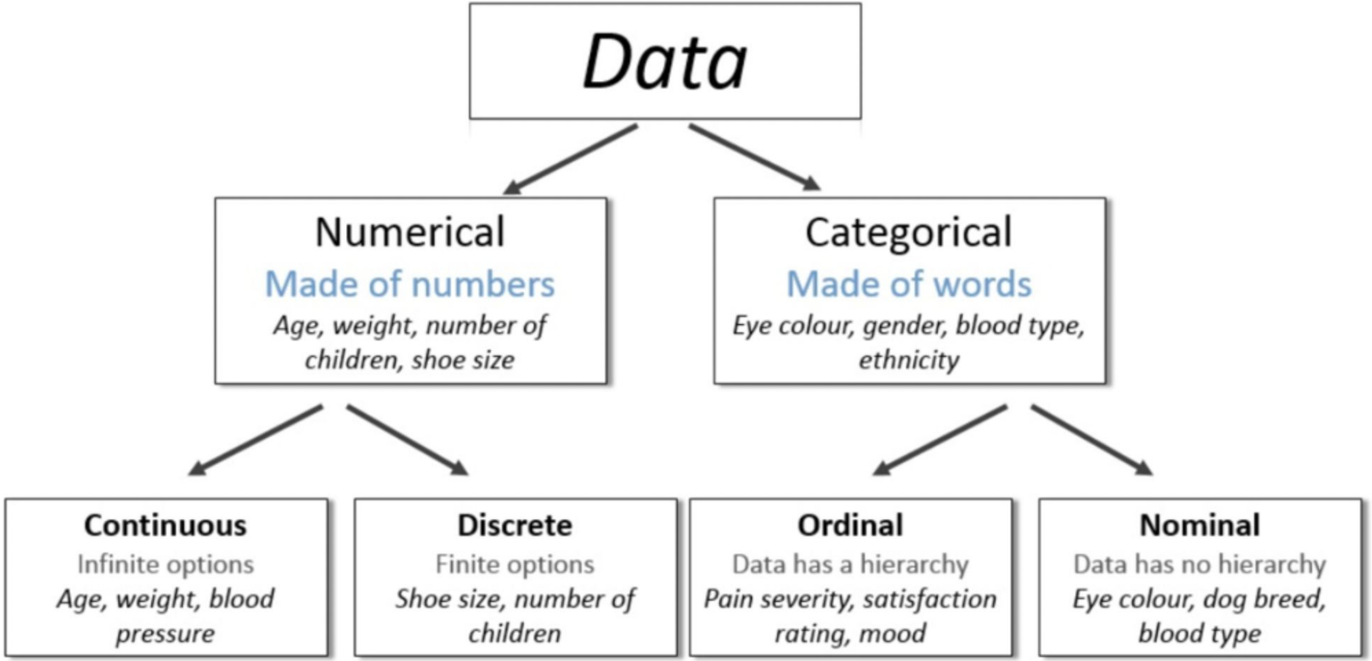
MSDS610 Final Presentation - Fall 2022
Exploratory Data Analysis

Abhradeep Mukherjee
Stephen Louis
Ensun Pak

Problem Solving Journey


Data Collection
Data Cleaning and Preparation
Exploratory Data Analysis
Machine Learning
Visualization and Reporting
Decision Making

Data Types



Data

Numerical
Made of numbers
Age, weight, number of children, shoe size

Categorical
Made of words
Eye colour, gender, blood type, ethnicity

Continuous
Infinite options
Age, weight, blood pressure

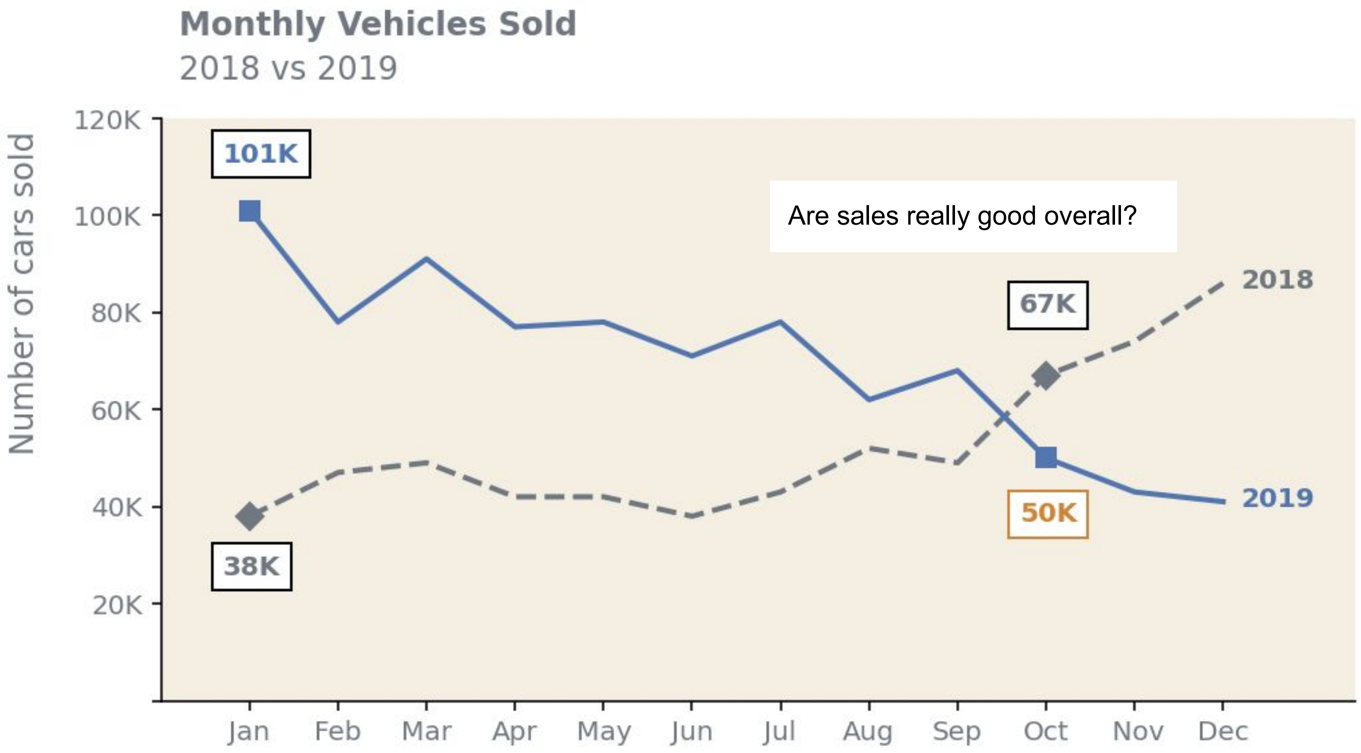
Discrete
Finite options
Shoe size, number of children

Ordinal
Data has a hierarchy
Pain severity, satisfaction rating, mood

Nominal
Data has no hierarchy
Eye colour, dog breed, blood type

Data Visualization

Myths about Data Analysis

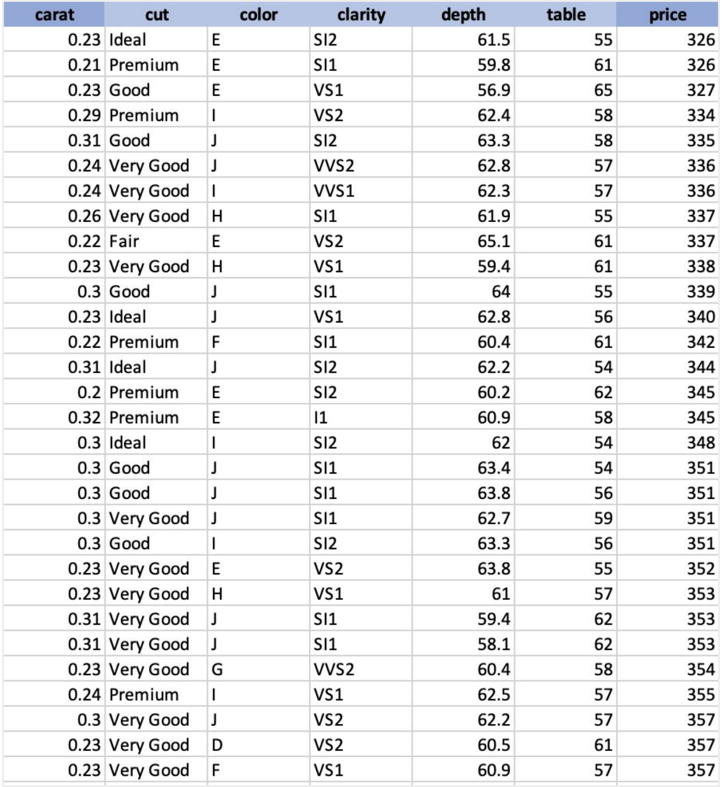


Are sales really good overall?

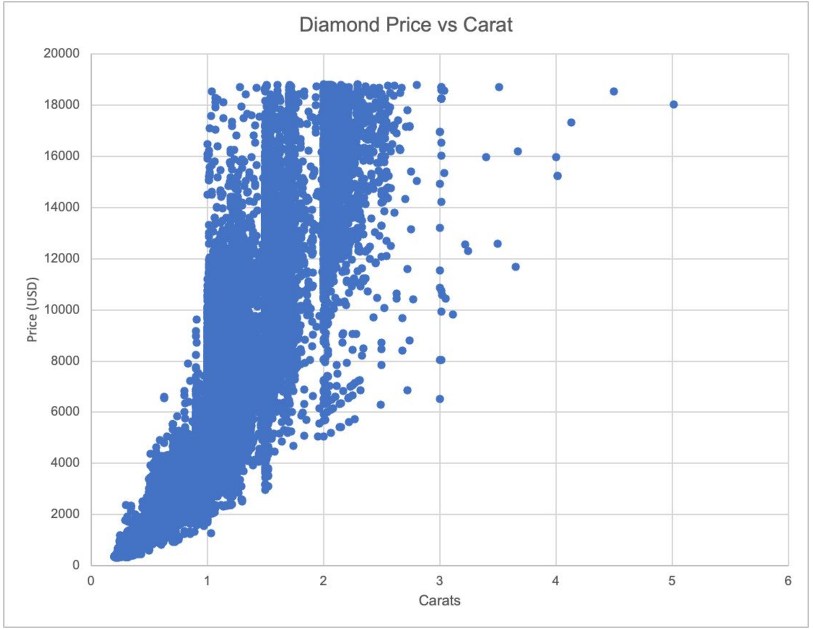
Myths about Data Analysis

- Numbers are always right
- Tables are always better
- Colors are not important
- We do EDA only during the beginning

Myths about Data Analysis



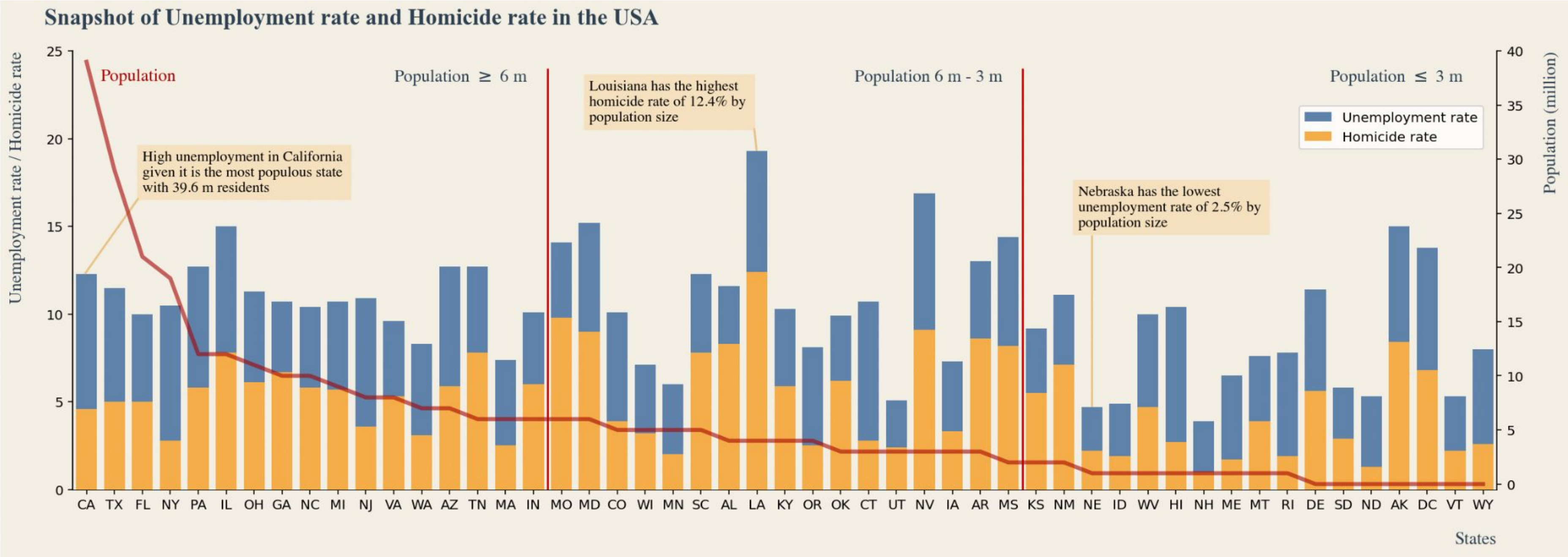
carat	cut	color	clarity	depth	table	price
0.23	Ideal	E	SI2	61.5	55	326
0.21	Premium	E	SI1	59.8	61	326
0.23	Good	E	VS1	56.9	65	327
0.29	Premium	I	VS2	62.4	58	334
0.31	Good	J	SI2	63.3	58	335
0.24	Very Good	J	VVS2	62.8	57	336
0.24	Very Good	I	VVS1	62.3	57	336
0.26	Very Good	H	SI1	61.9	55	337
0.22	Fair	E	VS2	65.1	61	337
0.23	Very Good	H	VS1	59.4	61	338
0.3	Good	J	SI1	64	55	339
0.23	Ideal	J	VS1	62.8	56	340
0.22	Premium	F	SI1	60.4	61	342
0.31	Ideal	J	SI2	62.2	54	344
0.2	Premium	E	SI2	60.2	62	345
0.32	Premium	E	I1	60.9	58	345
0.3	Ideal	I	SI2	62	54	348
0.3	Good	J	SI1	63.4	54	351
0.3	Good	J	SI1	63.8	56	351
0.3	Very Good	J	SI1	62.7	59	351
0.3	Good	I	SI2	63.3	56	351
0.23	Very Good	E	VS2	63.8	55	352
0.23	Very Good	H	VS1	61	57	353
0.31	Very Good	J	SI1	59.4	62	353
0.31	Very Good	J	SI1	58.1	62	353
0.23	Very Good	G	VVS2	60.4	58	354
0.24	Premium	I	VS1	62.5	57	355
0.3	Very Good	J	VS2	62.2	57	357
0.23	Very Good	D	VS2	60.5	61	357
0.23	Very Good	F	VS1	60.9	57	357



Which would you analyze for trend between carat and price?

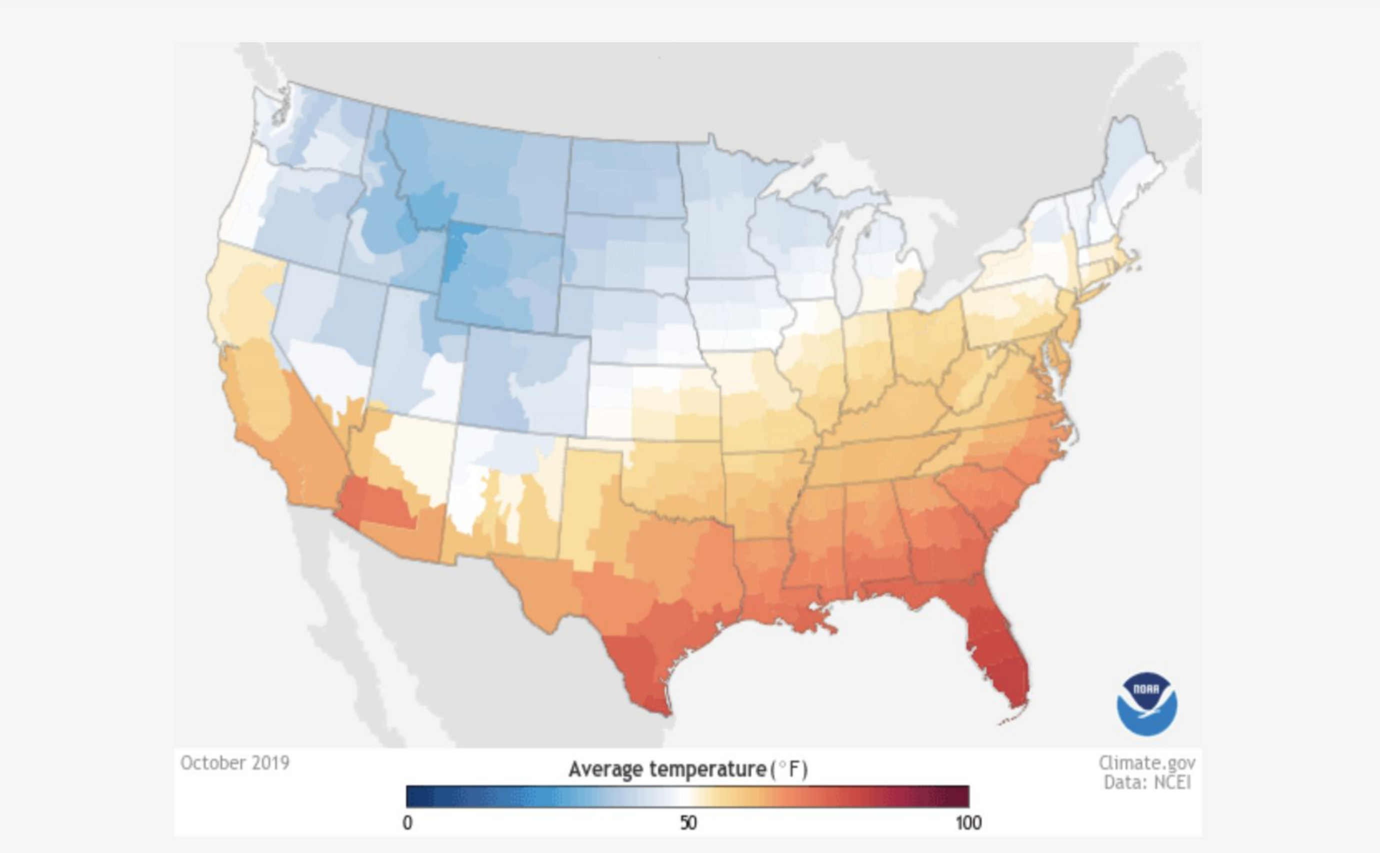
By the way, there are 50k+ rows in the dataset.

Myths about Data Analysis



We can incorporate many different types of visualization in one place to tell a story. Imagine trying to read that information in table format.

What does the variation in color tell us ?



Take this visualization, for example, looking at weather temperatures. Blue and red are readily understood without any explanation, and are easily distinguishable.

Simple Real World Analogy

What happens without EDA?

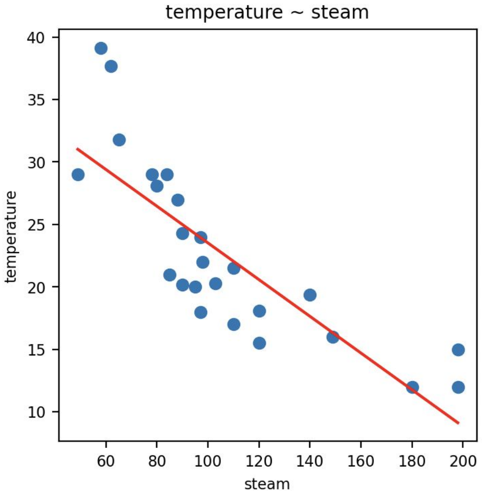
- An energy company was using field data in oil operations.
- In order to pump oil, steam is used to warm up the oil in order to ensure that the oil flows more easily.
  - In order to determine the amount of steam needed, infrared readings take the temperature of the lines.
- However, the lines can become dirty and insulated causing the temperature readings to be way off.
  - Because this problem went unnoticed, more steam was constantly used.
  - This resulted in excessive operational expenses that exceeded tens of millions of dollars.

Think of a Storybook

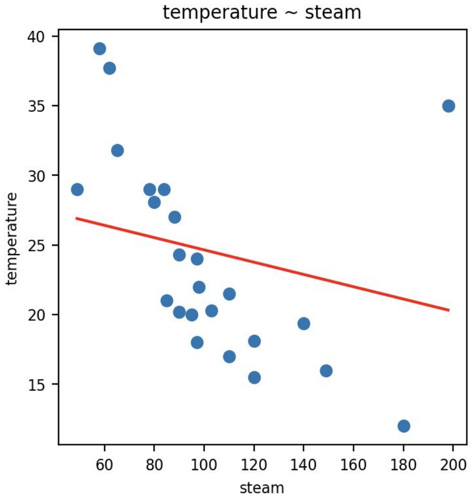
- Why visual is more appealing than text
- Importance of Colors
- Combination of Charts but keeping it simple
- Presence of EDA throughout the problem solving journey

An Example

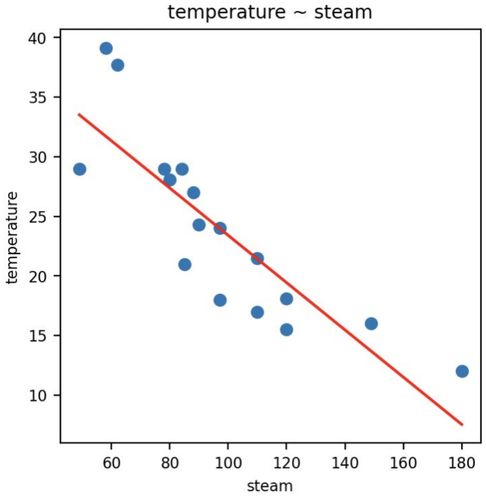
Imagine the steam input is driven by signals from a SLR model dataset through these three scenarios.



With proper EDA



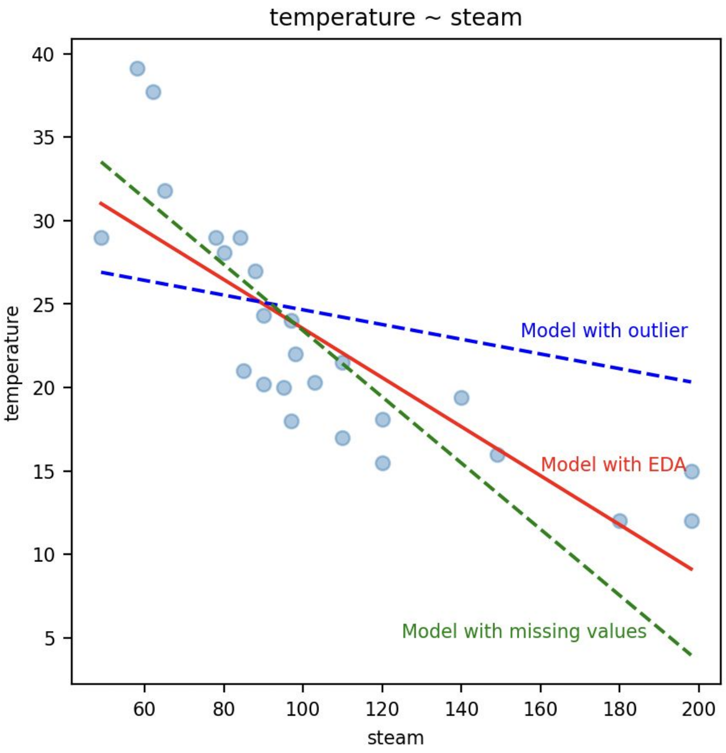
With outlier data



With missing data

Dataset - https://www.kaggle.com/datasets/uciml/autompg-dataset

Scenarios Compared



Jnderstanding he data does natter!

Messy Real World Data
Kedeisha Bryan

Thank you!

Important Steps in EDA

- Missing values
- Identifying and converting to the right data types
- Calculating summary statistics
- Creating plots for visualizations

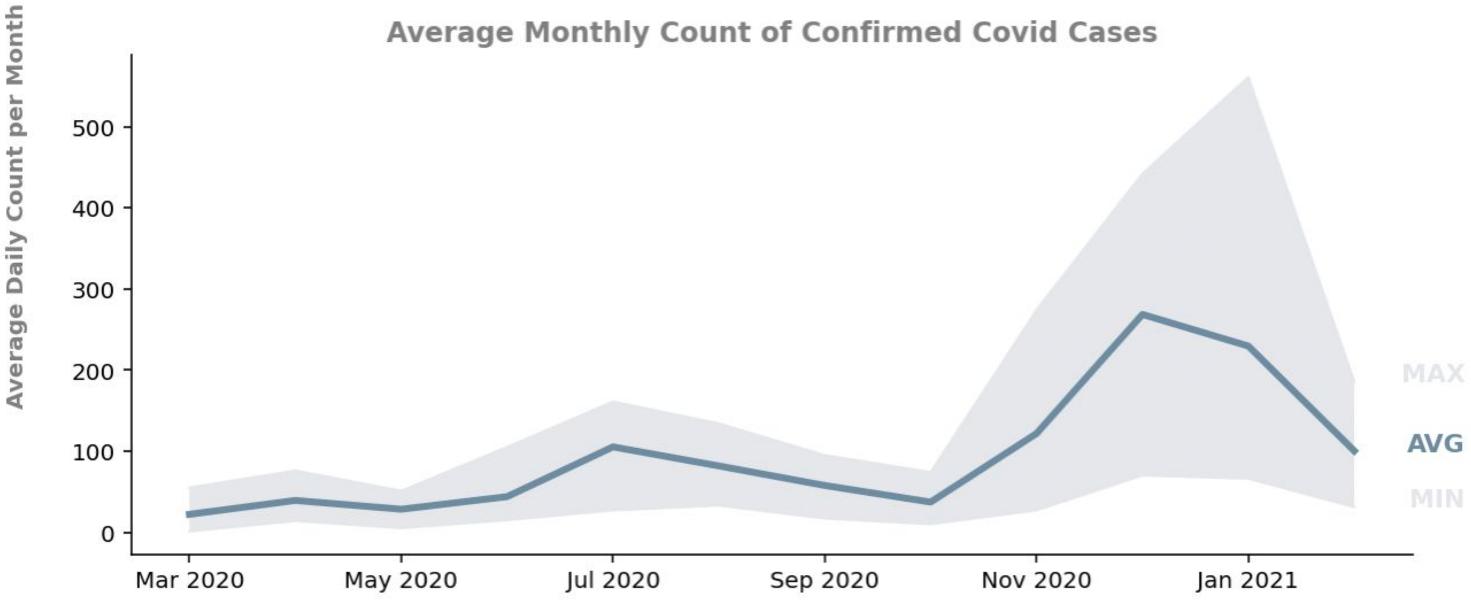
Statistical Measures

In descriptive statistics, summary statistics are used to summarize a set of observations, in order to communicate the largest amount of information as simply as possible.

Statisticians commonly try to describe the observations in:

- A measure of location, or central tendency, such as the arithmetic mean.
- A measure of statistical dispersion like the standard mean absolute deviation.
- A measure of the shape of the distribution like skewness or kurtosis.
- If more than one variable is measured, a measure of statistical dependence such as a correlation coefficient.

Myths about Data Analysis



We can incorporate many different types of visualization in one place to tell a story. Imagine trying to read that in table format.