Fundamentals of Data Analysis

Data Science Tools 1

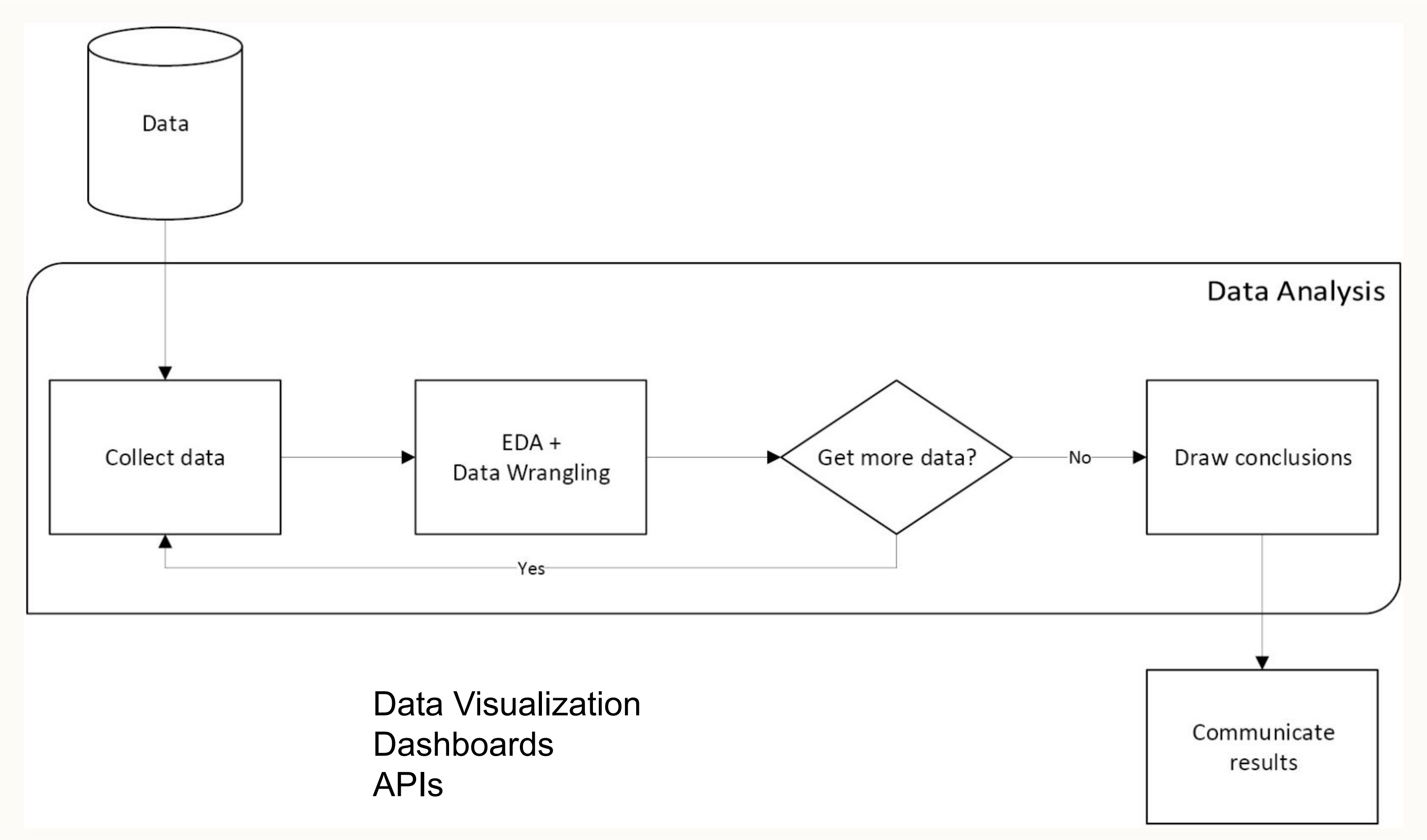
Fall 2021

UNIVERSITY *of* DENVER

Agenda

- Fundamentals of Data Analysis
- Review Wk5 Homework

Data Analysis Workflow



Data Visualization
Dashboards
APIs

Data Collection

Web Scraping (scrapy, BeautifulSoup)
APIs (requests)
Downloads (wget)
RDBMS (mysql, postgresql)
NoSQL (mongodb, HBASE)
FTP (File Transfer Protocol)
Hard Drives
Cloud Drives
Excel Spreadsheets

Data Collection - Sampling

Our sample must be a **random sample** that is representative of the population => our data should not be biased

Resampling - taking a random sample from a random sample

Stratified Random Sample - preserves the proportion of the groups in the data

Bootstrap Sample - beyond the scope of this course

Data Wrangling

Data wrangling is the process of preparing the data and getting it into a format that can be used for analysis.

What are some issues we may encounter with our data?

Data Wrangling

Data wrangling is the process of preparing the data and getting it into a format that can be used for analysis.

What are some issues we may encounter with our data?

- Human Errors (e.g., typos)
- Computer Errors (e.g., server down)
- Unexpected Values (e.g., USD 100 => $100)
- Incomplete information(e.g., survey results)
- Resolution(The data may have been collected per second, while we need hourly data for our analysis)
- Relevance of the fields (e.g, split or combine fields)
- Format of the data(e.g., may require reshaping of the data)
- Bad ordering (e.g., sensor data not arriving in order)

Exploratory Data Analysis

After Data Wrangling, you want to learn about your data and the process of process of doing it is called Exploratory Data Analysis (EDA).

This is where we bring our **Statistics and Visualizations skills**

- Statistics:
  - Descriptive - describes about the data
  - Inferential - infers about the data

Descriptive Statistics

- Perform Univariate Descriptive Statistics:
  - Measure of central tendency
    - Mean
    - Median
    - Mode
  - Measure of spread
    - Range
    - Variance
    - Standard Deviation
    - Coefficient of variation
    - Interquartile range

Descriptive Statistics - Measure of central tendency

Mean

$$\bar{x} = \frac{\sum_{1}^{n} x_i}{n}$$

Caution: very sensitive to outliers

What is the problem with looking at average US household income?

Median

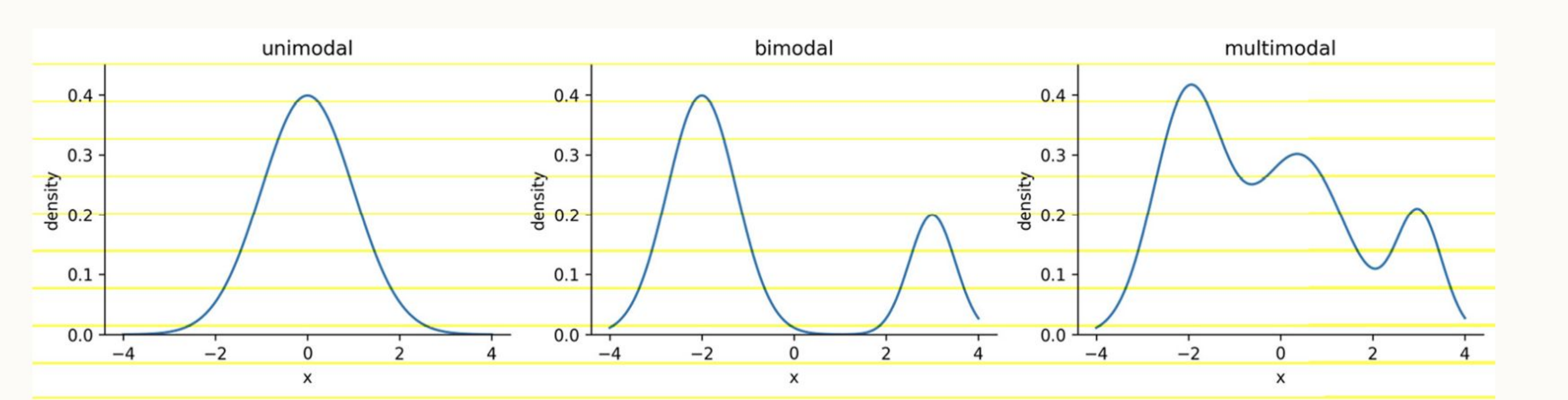
This is 50th percentile of the data. 50% of the values are greater than the median and 50% are less than the median.

Robust to outliers

Good metric to look at to understand the central tendency of US Household income

Mode

The **mode** is the most common value in the data.
If we have the numbers 0, 1, 1, 2, and 9, then 1 is the mode



How can it be done in Pandas?

```
df.group_by(col).agg(count)
df.value_counts
df.mode
```

Descriptive Statistics - Measure of Spread

Range

range = max(Xi) - min(Xi)

Caution: sensitive to outlier

Standard Deviation

Answers how spread out the data is from the mean

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

Range

range = max(Xi) - min(Xi)

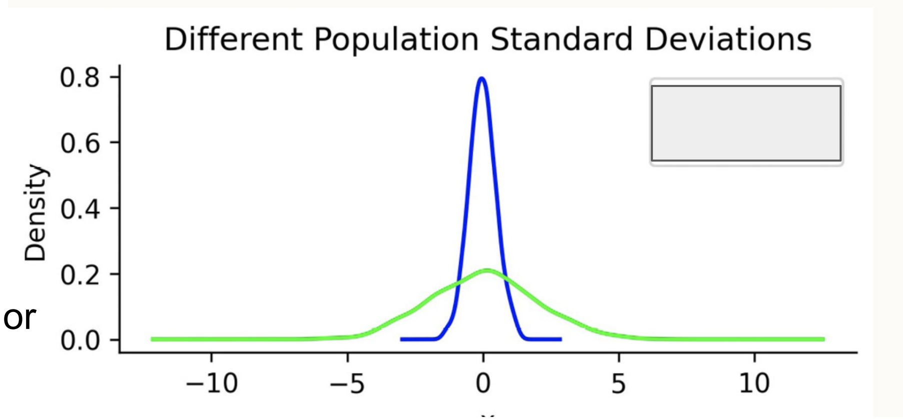
Caution: sensitive to outlier

Standard Deviation

Answers how spread out the data is from the mean

$$s = \sqrt{\frac{\sum_1^n (x_i - \bar{x})^2}{n-1}} = \sqrt{s^2}$$

Which curve has higher std deviation? blue or green? Ans = green



Coefficient of Variation

used in comparing the dispersion (or spread) of two different datasets
note that this is a unitless quantity

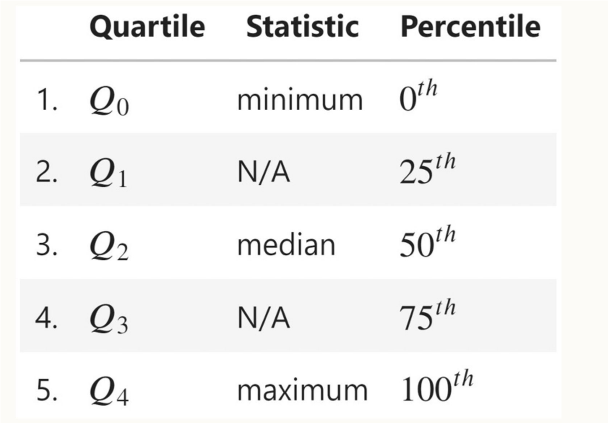
$$CV = \frac{s}{\bar{x}}$$

Interquartile range

this provides median based dispersion in the dataset

$$IQR = Q_3 - Q_1$$

5-number summary of the data -



	Quartile	Statistic	Percentile
1.	Q_0	minimum	0^{th}
2.	Q_1	N/A	25^{th}
3.	Q_2	median	50^{th}
4.	Q_3	N/A	75^{th}
5.	Q_4	maximum	100^{th}

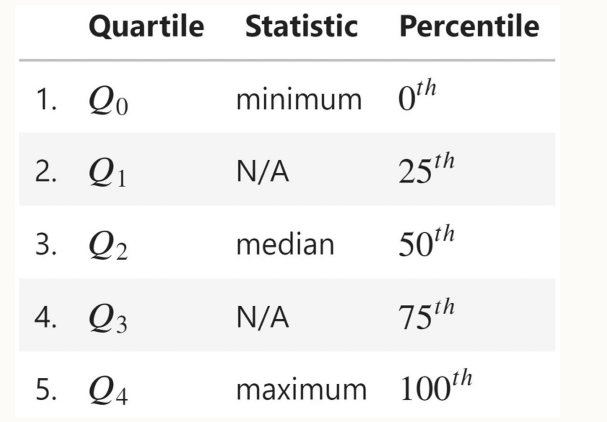
Descriptive Statistics - Measure of Spread

Interquartile range

this provides median based dispersion in the dataset

$$IQR = Q_3 - Q_1$$

5-number summary of the data -



	Quartile	Statistic	Percentile
1.	Q_0	minimum	0^{th}
2.	Q_1	N/A	25^{th}
3.	Q_2	median	50^{th}
4.	Q_3	N/A	75^{th}
5.	Q_4	maximum	100^{th}

Box Plot

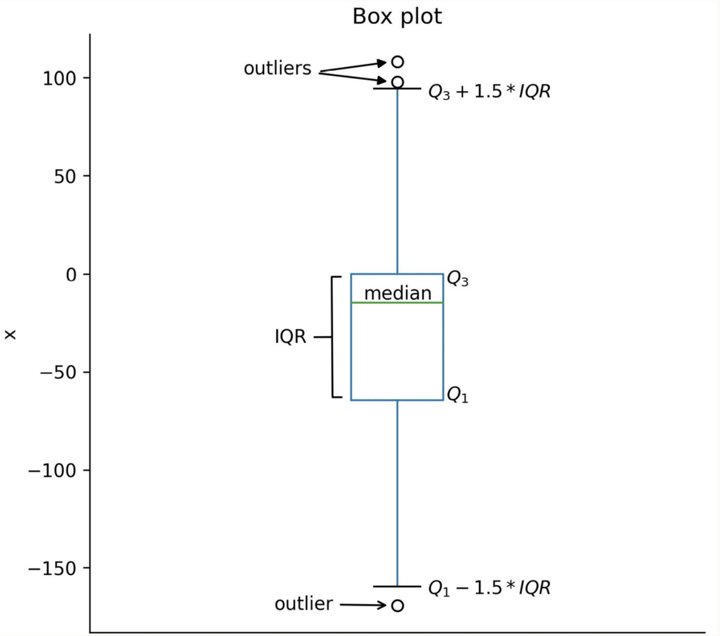
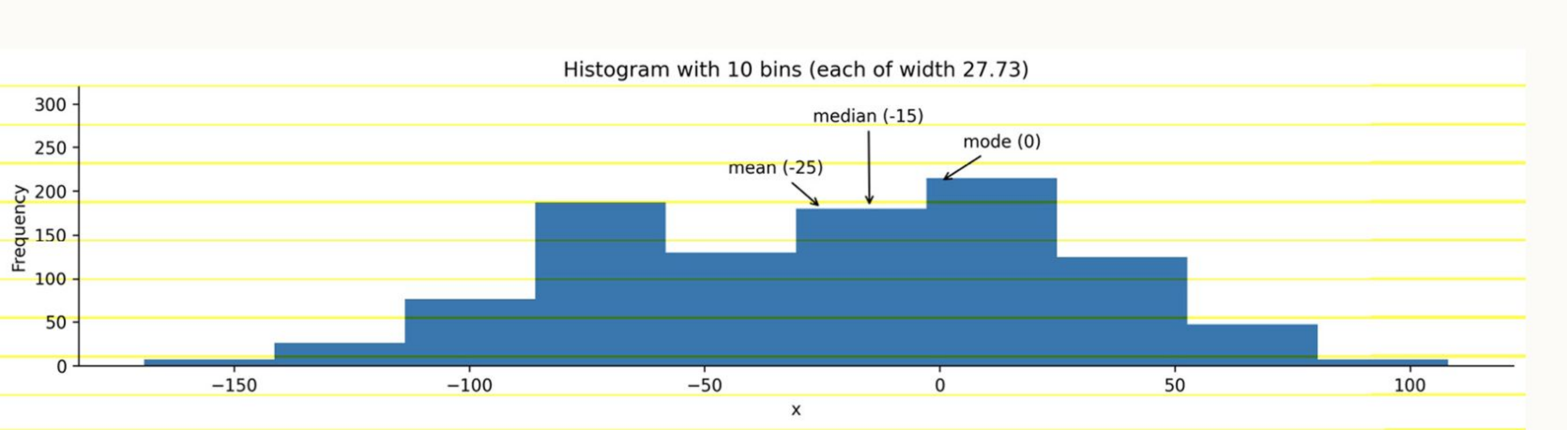
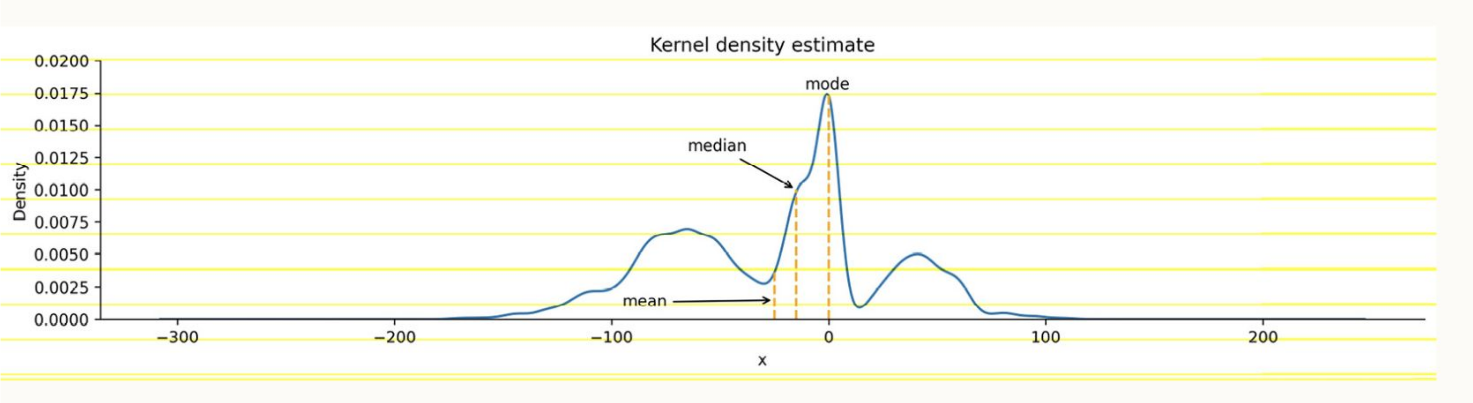


Figure 1.6 – The Tukey box plot

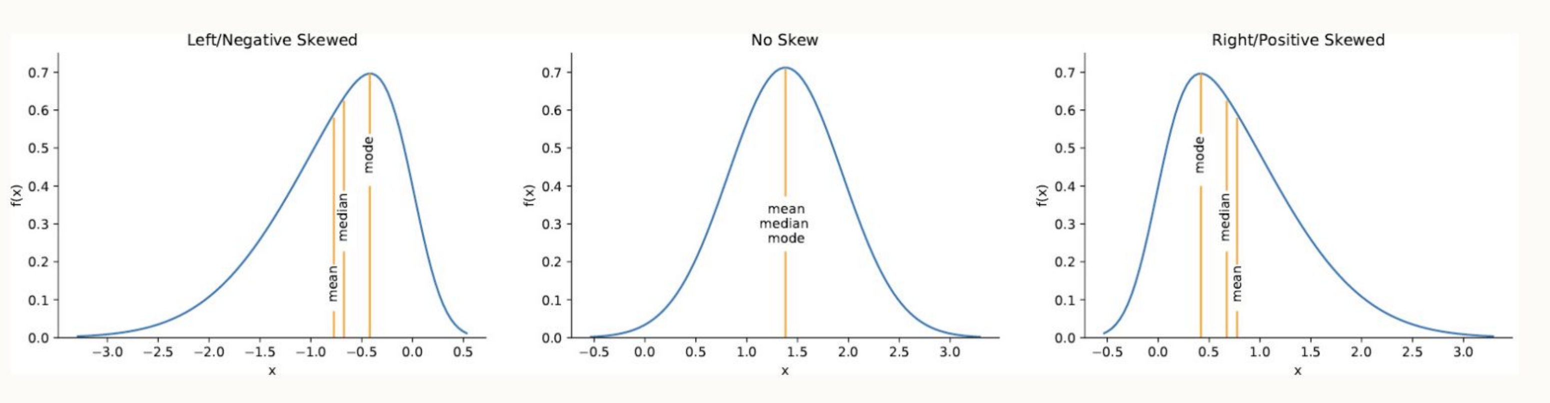
Histogram Plot



KDE (kernel density estimate) Plot



Also can be used to visualize skewness

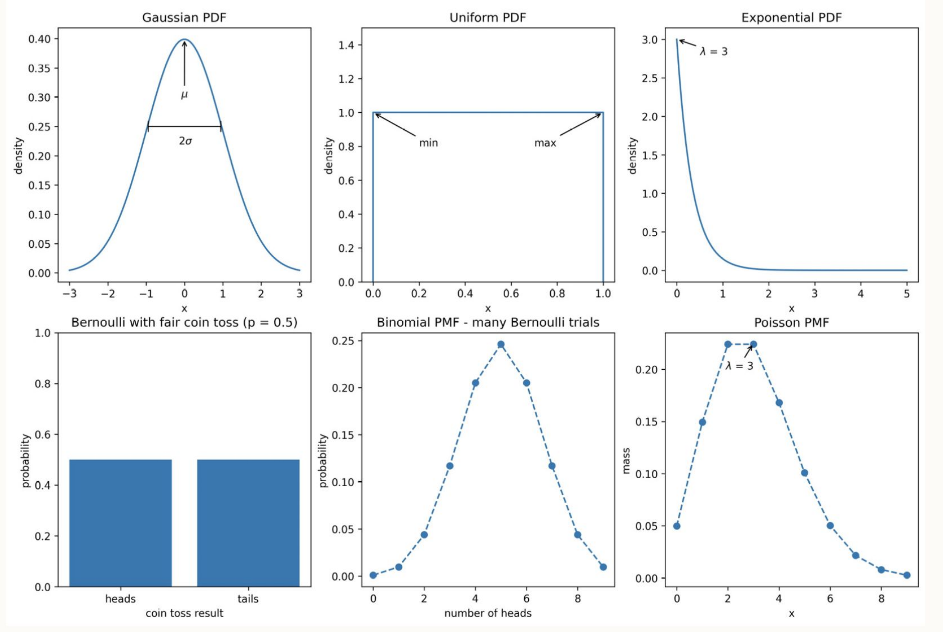


Descriptive Statistics - Common Distributions

Gaussian - normally found in nature (e.g., heights)
Uniform distribution places equal likelihood on each value within its bounds
When we generate a random number to simulate a single success/failure outcome, it is called a **Bernoulli** trial. This is parameterized by the probability of success (p). When we run the same experiment multiple times (n), the total number of successes is then a **binomial** random variable. Both the Bernoulli and binomial distributions are discrete.

Poisson distribution is a discrete distribution that is often used to model arrivals. The time between arrivals can be modeled with the **exponential distribution**. Both are defined by their mean, lambda (λ)

Note: discrete distributions give us a probability mass function (PMF) instead of a PDF



Gaussian PDF

Uniform PDF

Exponential PDF

Bernoulli with fair coin toss (p = 0.5)

Binomial PMF - many Bernoulli trials

Poisson PMF

Descriptive Statistics - Scaling

min-max scaling

normalizes data from 0 to 1

$$x_{scaled} = \frac{x - \min(X)}{range(X)}$$

z-score scaling

uses mean and standard deviation

$$z_i = \frac{x_i - \bar{x}}{s}$$

Descriptive Statistics - measure variables relationship

Covariance

- provides relationship between two variables
- The magnitude of the covariance isn't easy to interpret, but its sign tells us whether the variables are positively or negatively correlated.

$$cov(X,Y) = E[(X - E[X])(Y - E[Y])]$$

What if we want to know how strongly the variables are related?

Covariance

- provides relationship between two variables
- The magnitude of the covariance isn't easy to interpret, but its sign tells us whether the variables are positively or negatively correlated.

$$cov(X,Y) = E[(X - E[X])(Y - E[Y])]$$

What if we want to know strongly the variables are related?

- **Correlation** tells us how variables change together both in direction (same or opposite) and magnitude (strength of the relationship). To find the correlation, we calculate the **Pearson correlation coefficient**
- Bounds values from -1 to 1

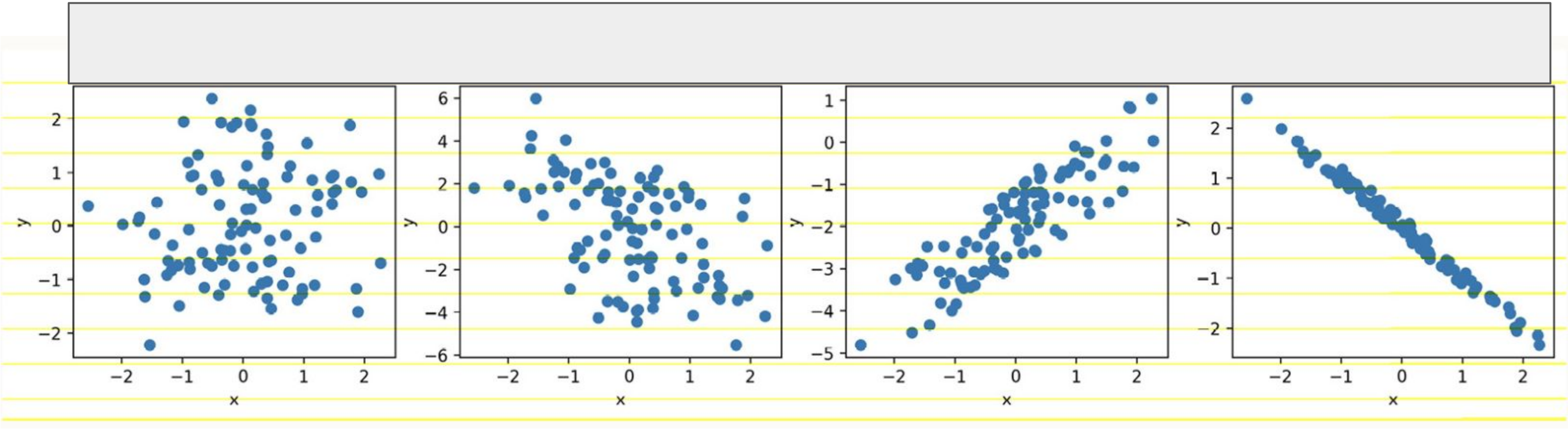
$$\rho_{X,Y} = \frac{cov(X,Y)}{s_X s_Y}$$

Descriptive Statistics - measure variables relationship

Which figures have high and low correlations between x and y?

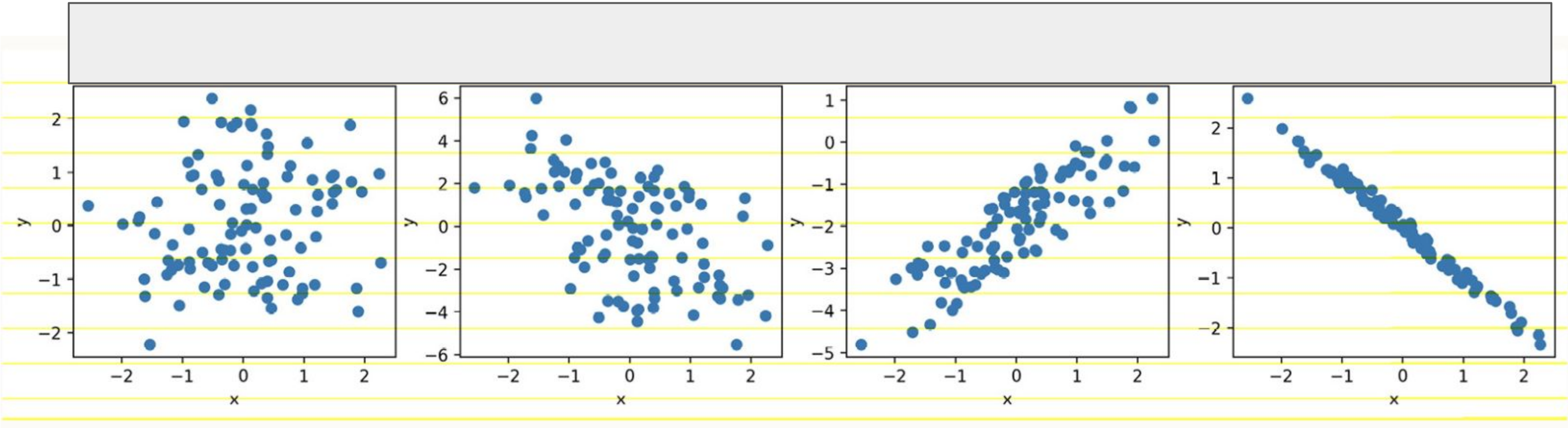
low => first, second

high => third (+), fourth(-)



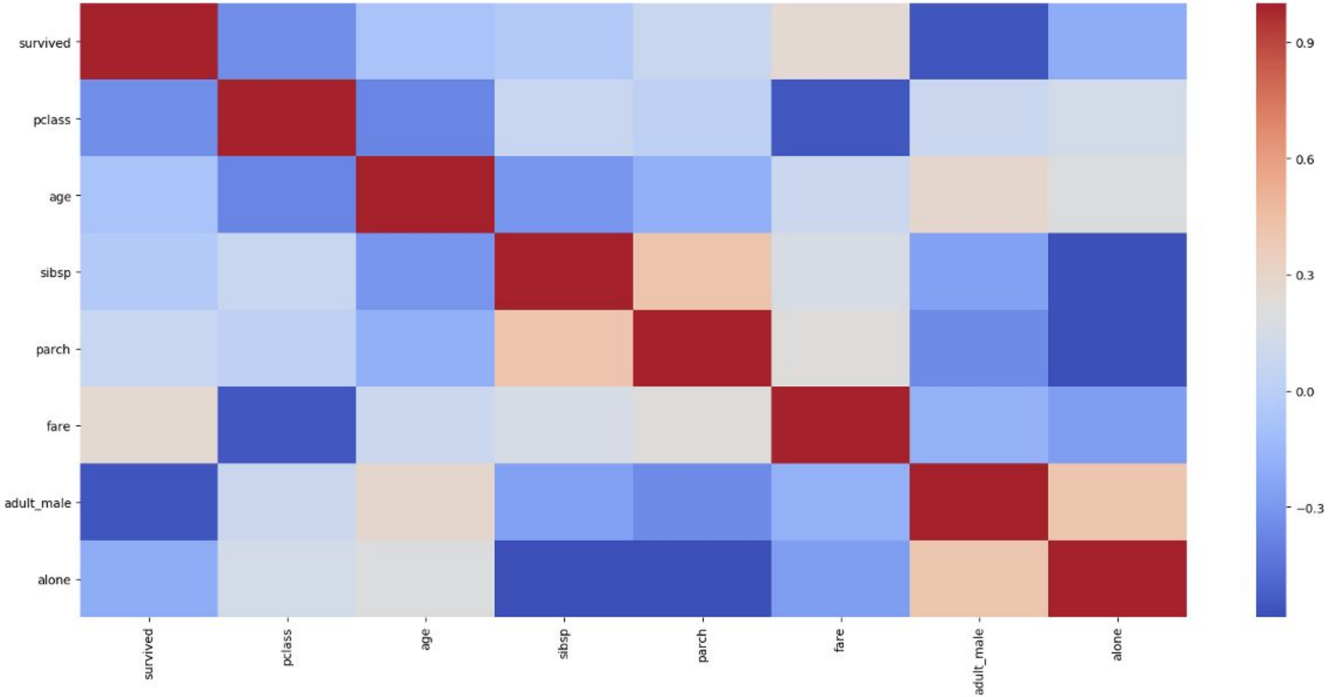
Descriptive Statistics - measure variables relationship

Correlations with Pandas?
df.corr



```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
df=sns.load_dataset('titanic')
print(df.head())
sns.heatmap(df.corr(), cmap='coolwarm')
plt.tight_layout()
plt.show()
```

Descriptive Statistics - measure variables relationship



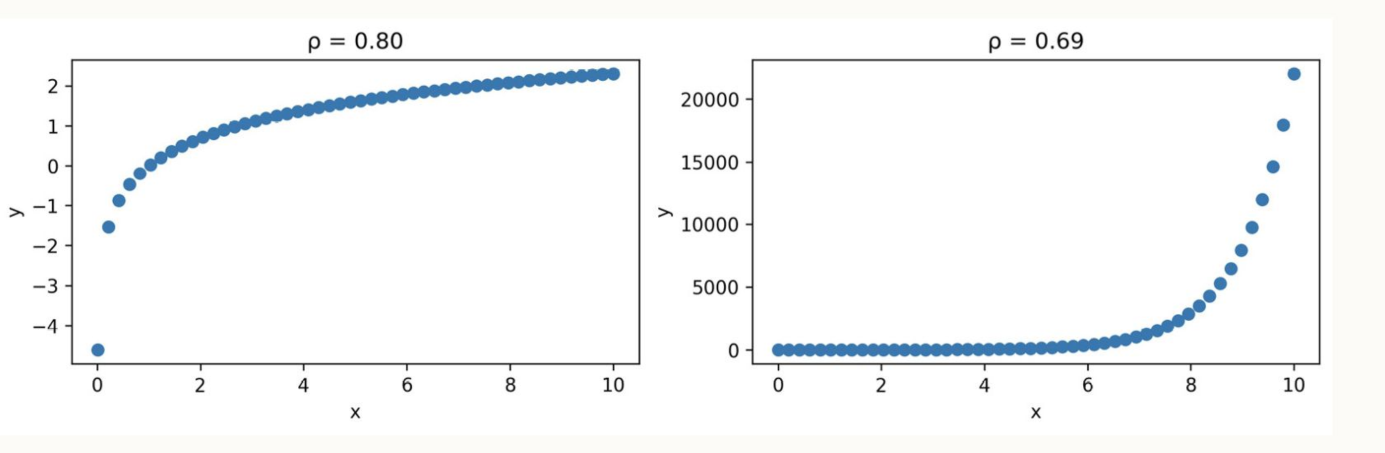
Limitation of correlations coefficient:

Doesn’t tell you the nature of correlation

Limitation of correlations coefficient:

Doesn't tell you the nature of correlation

What is the other limitation?

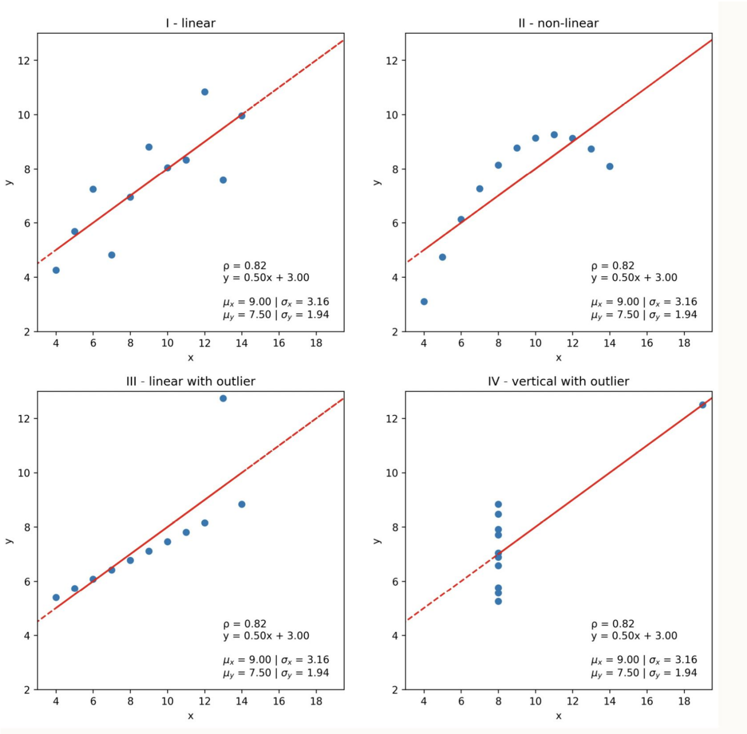


Descriptive Statistics -Visualization

Visualization is not optional

Summary Statistics can be limiting

Demonstrated by **Anscombe's quartet** is a collection of four different datasets that have identical summary statistics and correlation coefficients

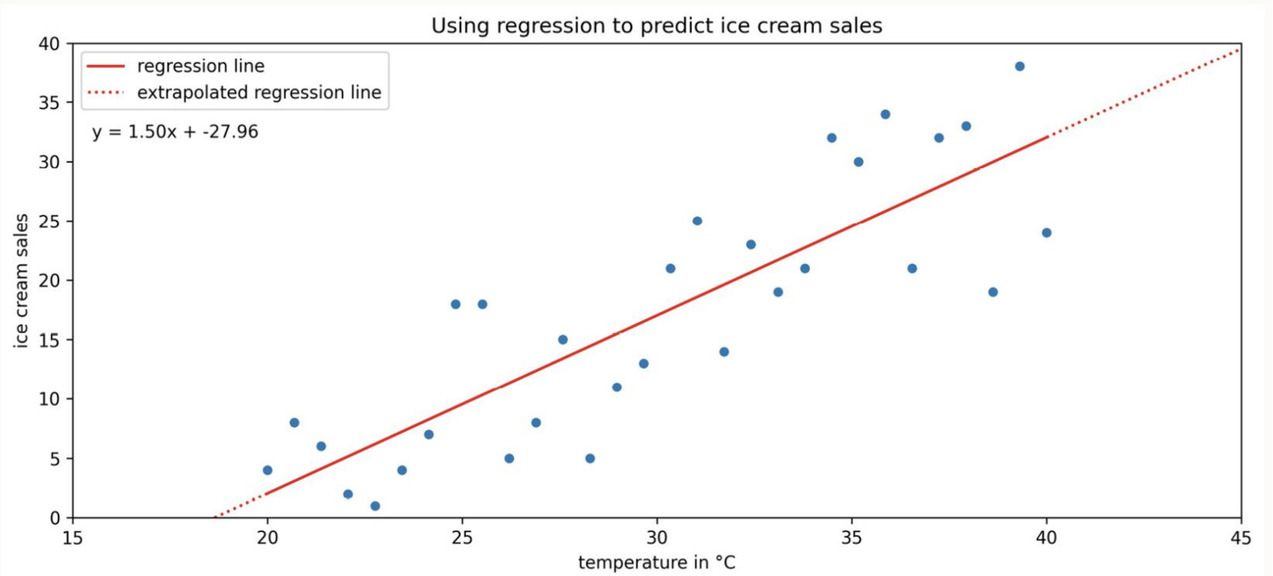


Inferential Statistics

We infer something about the data:

- Hypothesis Tests
    - Null hypothesis

- Estimating parameters
    - Regression, Machine Learning (Deep or Shallow)
    - Neural Network



Drawing Conclusions

- Did we notice any patterns or relationships when visualizing the data?
- Does it look like we can make accurate predictions from our data? Does it make sense to move to modeling the data?
- Should we handle missing data points? How?
- How is the data distributed?
- Does the data help us answer the questions we have or give insight into the problem we are investigating?
- Do we need to collect new or additional data?
- How often do we need to refresh our data?

Week 5 Homework

5 mins Break