Preprocessing and Data Cleaning

Prep work for machine learning

Machine Learning Plan

- Machine Learning Intro
- Preprocessing
- Making Pipelines
- Comparing Evaluation Metrics
- Feature Selection

Preprocessing Plan

- Exploring the Dataset
- Find and handle missing data
- Encoding Categorical and String Data into Numbers
- Outlier Removal
- Standardization and Normalization

In Class Example

- Dataset that should predict whether someone purchased a product or not
- Load in dataset as a class and look at it.

1. Identify the Features and the Target

- What are features and Target here?

2. Explore Dataset

- Ex:
  - What are counts of different values in each column?
  - What are averages of each column?
  - What columns follow a normal distribution?
  - How many NA values are there?

2. Explore Dataset

- You can use pandas's **describe** method to get total count and means for numerical columns
- You can use pandas's **value_counts** method to get count of a column's individual value count.
- Make bar chart function with docstring to explore this as well.

2. Explore Dataset

- Which columns follows a normal distribution?
  - This information is important to know for later when we have to normalize or standardize the data
  - It also may tell us that the columns do not have any “blind spots”

3. Handling NA values

- Different methods
  - Remove any row with missing data
  - Remove any column with missing data
  - Impute using the average value for numerical data
  - Impute using the most frequent value for categorical data
  - Impute using random selection for categorical data
  - Impute with a new category for categorical data- eg category called “missing”
  - **Impute**- Replace the value based on inference

3. Handling NA values

- Pros and Cons of removing rows with missing values.
- What do you think?

3. Handling NA values

- Removing Rows with NA values
- Pros
  - Easy and simple
- Cons
  - Could lose valuable information
- Rule of thumb
  - Don’t do it if you are losing more then 10% of data

3. Handling NA values

- Removing Cols with NA values
- When would you do it vs not?

3. Handling NA values

- Removing Cols with NA values
- When do it?
  - When you may want to get rid of that column anyway
    - Adds noise
    - Not relevant or helpful
- Disadvantage
  - You lose an entire feature! This could have a lot of information!!

3. Handling NA values

- Imputing average value for numerical results
- Pros and Cons?

3. Handling NA values

- Imputing average value for numerical results
- Pros
  - Allows you to use more data for your training
  - With normal distribution, the underlying pattern the model is predicting will not change
- Cons
  - You are inserting “fake” data- could skew results if this is done for a lot of values

3. Handling NA values

- Imputing most frequent value for categorical data
- Talk about what this means with Ex
- What do you think pros and cons are?

3. Handling NA values

- Imputing most frequent value for categorical data
- Pros
  - You get to use more data!
- Cons
  - Could just be wrong
  - Recommended to use this if you have one category that is the much bigger than the others
  - May make more sense if you look at other factors in the data to confirm that you are filling in where there is

3. Handling NA values

- Replacing certain problematic categories with a new category- Ex: “missing”, “other”
- Pros and Cons?

3. Handling NA values

- Replacing certain problematic categories with a new category- Ex: “missing”, “other”
- Pros
  - Your not putting in anything false that will skew the data
- Cons
  - Assuming there are not a lot of these, they will not affect the ml training much.
  - If there are a lot, the “missing” category may actually play an important role in how the model performs. You probably don’t want that

4. Encoding Categories into Numbers

- Machine Learning models need all features to be numbers
  - All ml models are just math functions
- What do we do with data that are strings?
  - We make them into numbers!
  - This process is called “encoding”

4. Encoding Categories into Numbers

- Common methods
  - Label Encoding
  - One Hot Encoding

4. Encoding Categories into Numbers

- Label Encoding
  - Do example on whiteboard
- What are pros and cons?

4. Encoding Categories into Numbers

- Label Encoding
- Pros
  - Simple
  - Does not change amount of columns
- Assumptions it has
  - That there is a sequence in the categories

4. Encoding Categories into Numbers

- Ex:
  - France, Germany, Spain
  - Encoded:
    - F = 0
    - G = 1
    - S = 2
  - What is average of the 3 countries? A: Germany
    - It makes no sense in this case

4. Encoding Categories into Numbers

- One Hot Encoding:
  - Makes a Column for each possible value
  - Show Example on whiteboard
- Pros and cons?

4. Encoding Categories into Numbers

- One Hot Encoding:
  - Makes a Column for each possible value
  - Show Example on whiteboard
- Pros and cons?

4. Encoding Categories into Numbers

- One Hot Encoding
- Pros
  - Does not prioritize some variables over others
- Cons
  - Makes a lot of columns if you have a lot of categories
  - Unintuitive, but not necessarily a bad thing for machine learning

Outlier Removal- ie Anomaly Detection

- Ex:
  - HW where someone ran over 100,000 km in a session
- Why would you want to remove outliers?

Outlier Removal

- Why would you NOT want to remove outliers?

Outlier Removal

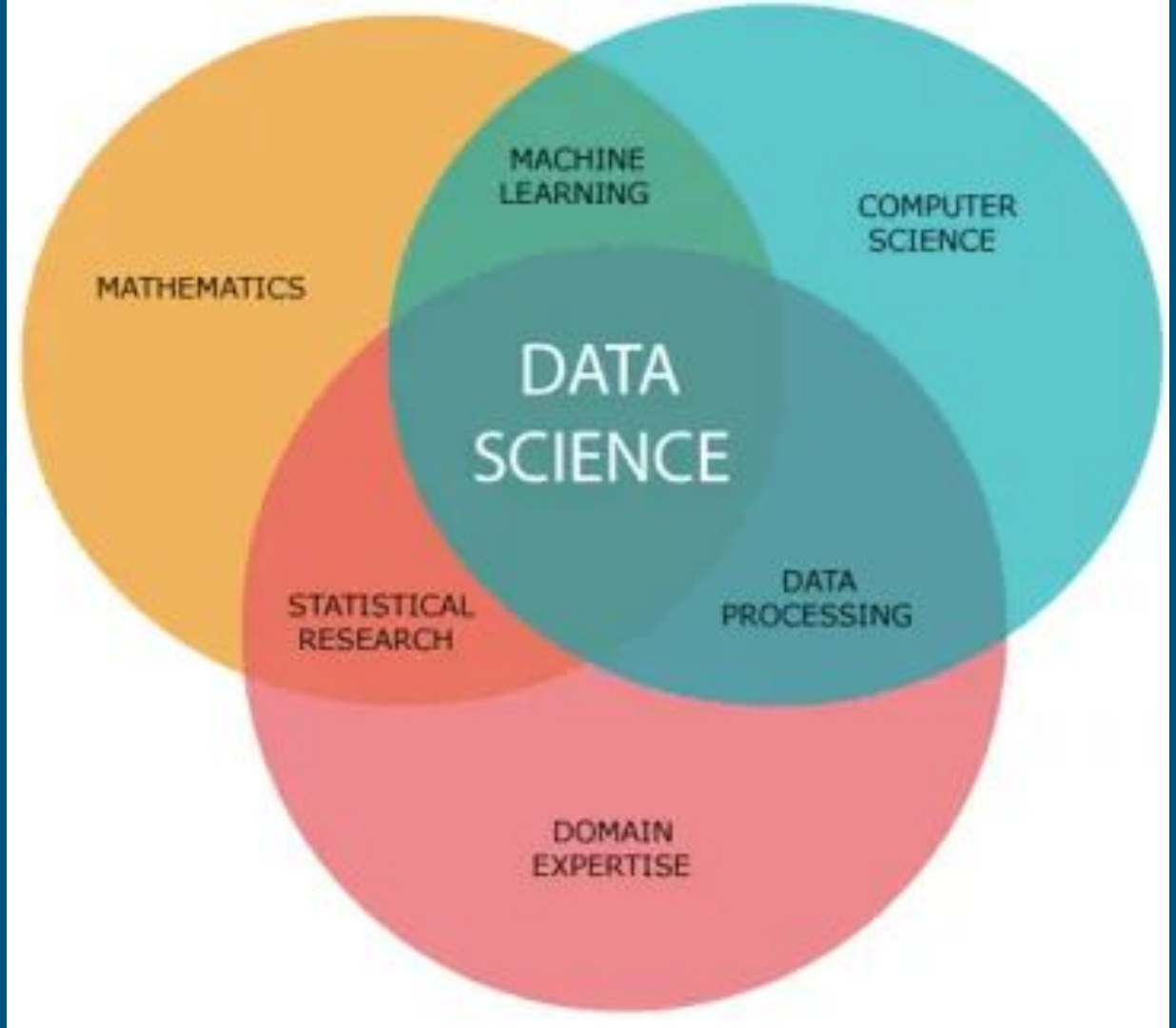
- Why would you NOT want to remove outliers?
  - Ex: You want to predict the population of different groups of endangered species
    - If you see an extremely small number, you don’t want to ignore it, these may be the groups you care most about!

Outlier Removal

- TLDR
  - You want to remove outliers that don’t match the underlying function/pattern you want to model
  - You want to keep outliers

How do you know when to remove outliers?

- Domain Knowledge

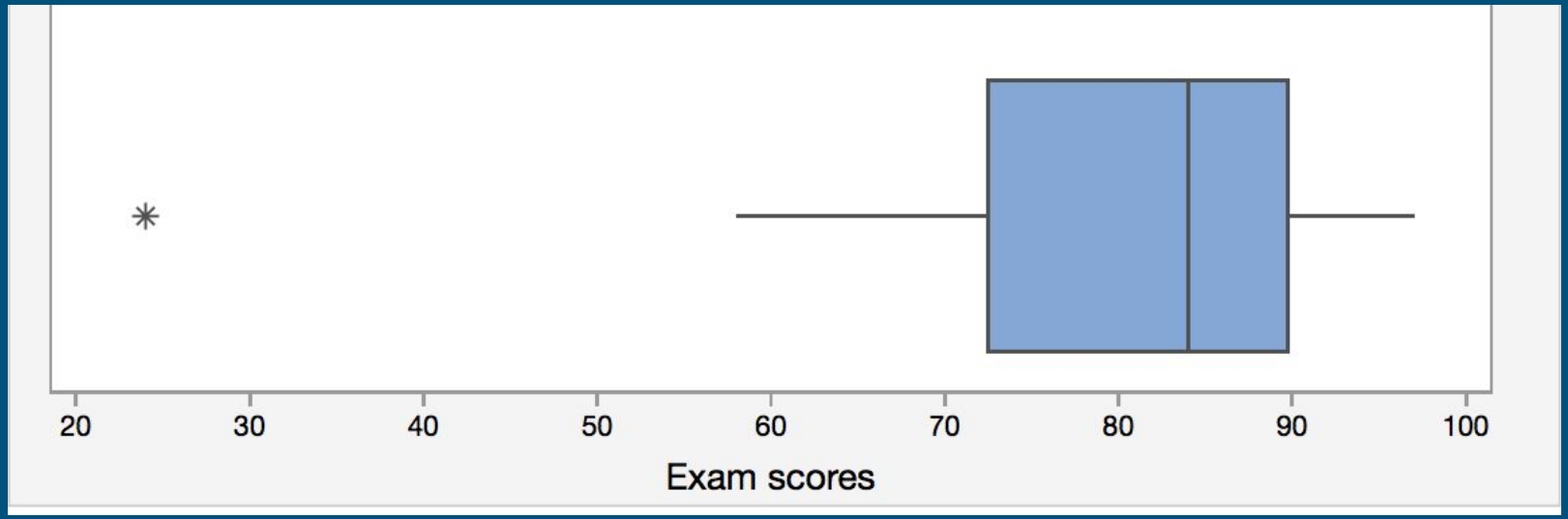

MACHINE
LEARNING
COMPUTER
SCIENCE
MATHEMATICS
DATA
SCIENCE
STATISTICAL
RESEARCH
DATA
PROCESSING
DOMAIN
EXPERTISE

3 Methods we will explore

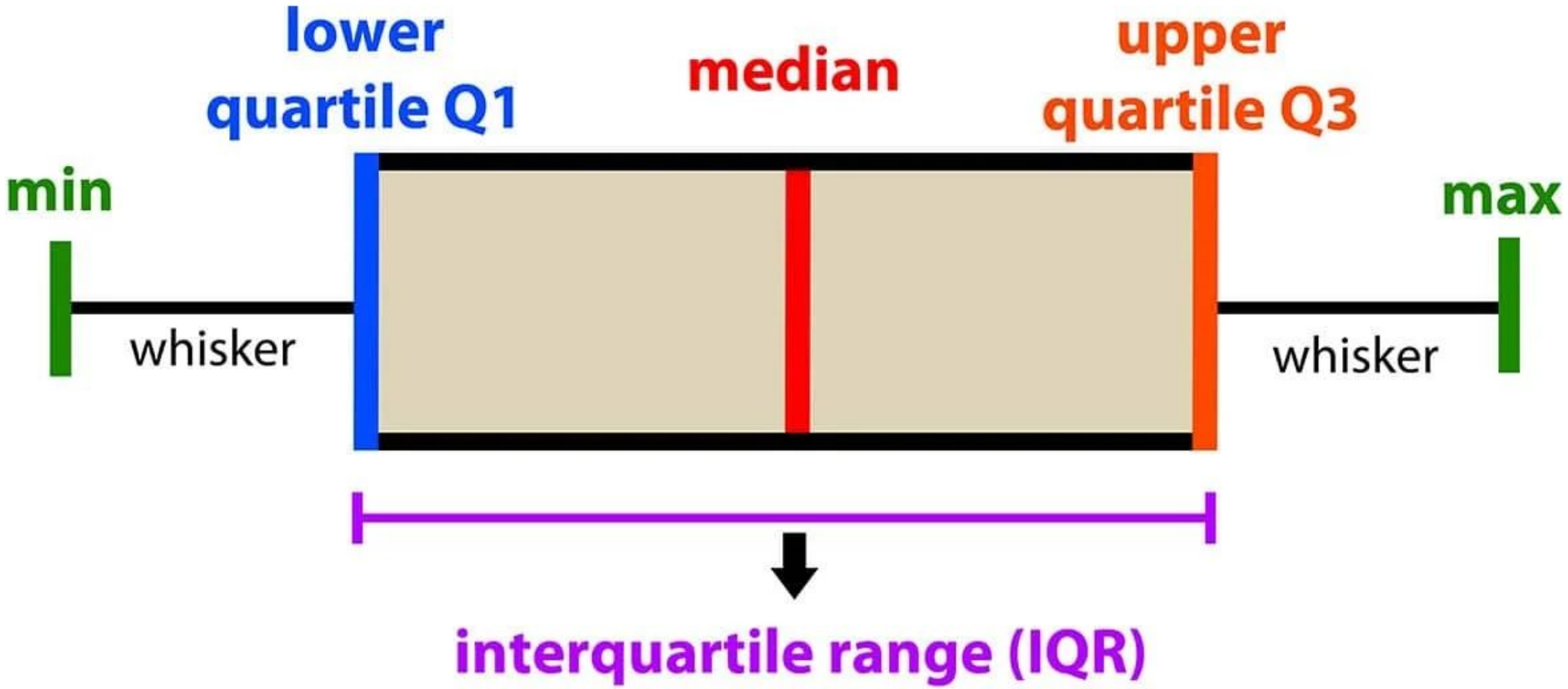
1. Box plot with hard coded thresholding
2. Interquartile Range
3. Z-score

Box plot

- What is a box plot? How do you read one?



introduction to data analysis: Box Plot



$$\text{Max} = Q3 + (1.5)(IQR)$$

$$\text{Min} = Q1 - (1.5)(IQR)$$

Do Interquartile Range Ex

- Use ziteboard with 7 numbers

Making a Boxplot with Seaborn

- Seaborn is a wrapper of matplotlib
  - Much easier to customize
  - Default settings are better
  - Generally easier to use then matplotlib
  - BUT it is built on top of matplotlib

Removing Outliers with hard coded thresholding

- Show example from notebook

Removing Outliers with hard coded Thresholds

- Pros and Cons?

Removing Outliers with hard coded Thresholds

- Pros
  - Expert/boss can tell you what they think you should ignore
- Cons
  - You gotta just pick the thresholds
  - Not mathematical
  - May be hard to defend if you are not domain expert

Removing Outliers based on Interquartile Range

- Show example from notebook

Removing Outliers based on Interquartile Range

- Pros and Cons?

Removing Outliers based on Interquartile Range

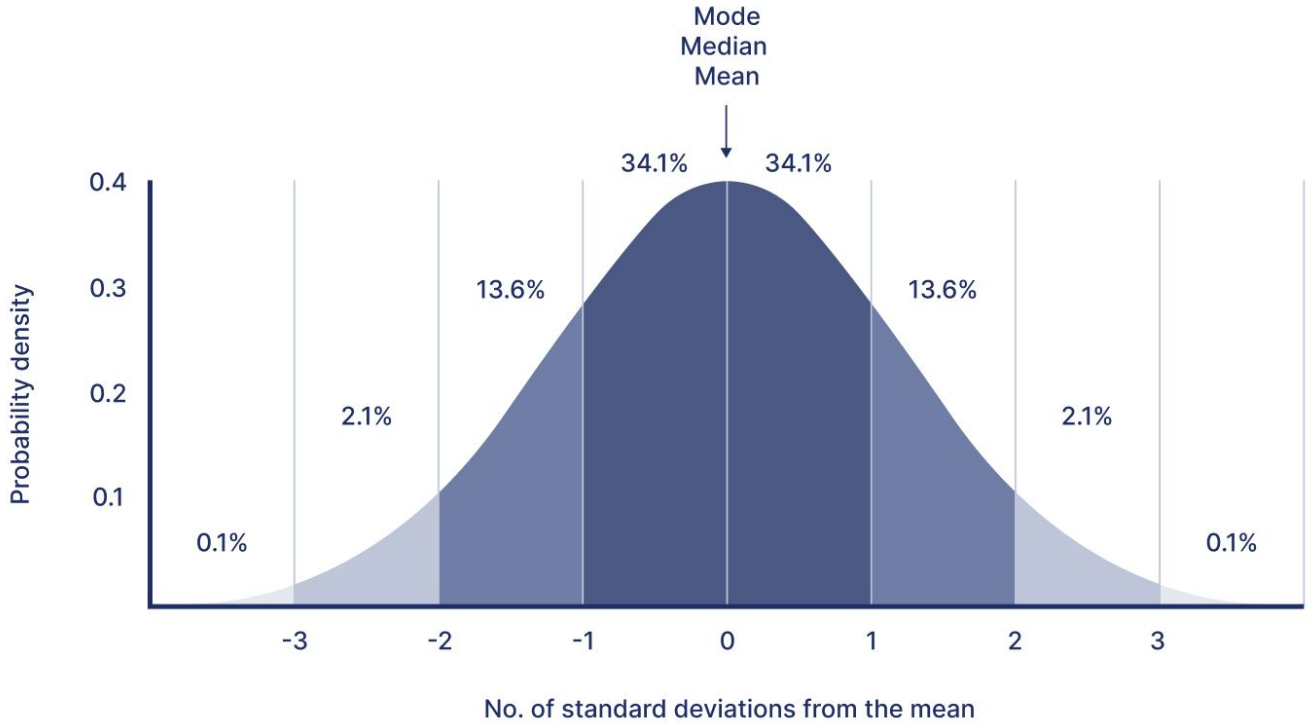
- Pros
  - Standard statistical technique
  - Works for features that are not normally distributed
    - Boxplot takes distribution into account
- Cons
  - If data does follow a normal distribution, this method can remove more outliers then another method (z-score). This may/may not be a con

What if the data is normally distributed?

- Is there another way?
- Yep! Using something called the Z-score!

Z-score- good for normally distributed data

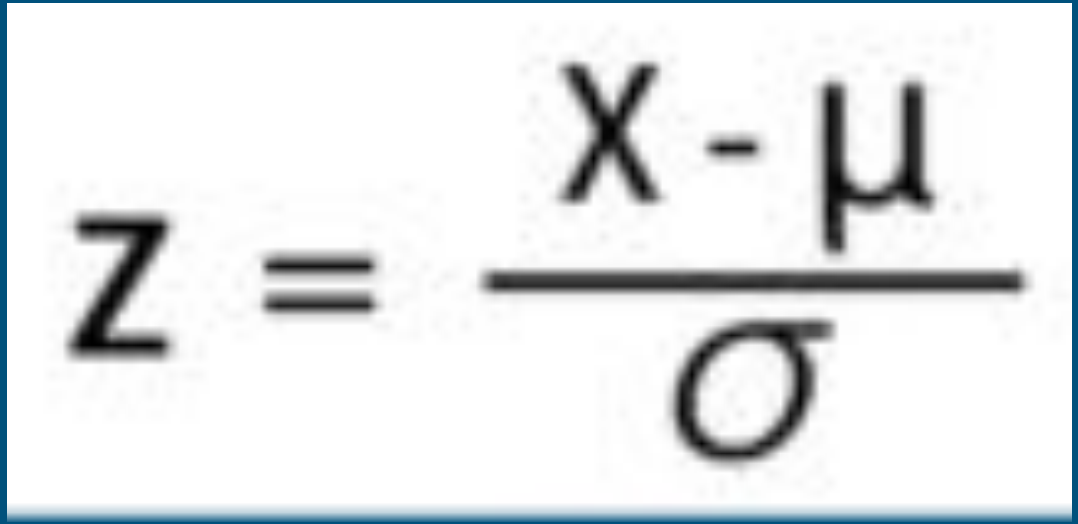
Standard normal distribution



Scribbr

What is Z-score?

- X- value in your data
- u- mean of feature
- σ- standard deviation



What is Z-score?

- What does a z = 0 mean?
- z = 1?
- z = 2?
- z = -1?

$$z = \frac{x - \mu}{\sigma}$$

What does Z-score mean?

- It tells you how far you are from the mean.
- Lets look at example in code

Z-score

- Pros and Cons? When to use it vs not?

Z-score

- Pros
  - Established recognized mathematical procedure
  - Still gives you control about how much to prune off
- Should only be used on normally distributed data

Feature Scaling

- What is it and why do we care?
- Ex: predict if someone wants to buys a house given age and salary
  - What are ranges?

Why do Feature Scaling after Outlier Removal?

- What do you think?

Outliers can impact how Feature Scaling Works

- Ex:
  - If you leave extreme outliers, then those are treated as the min and max and the distribution of your data may become very squished

Feature Scaling

- Two main methods:
  a. Normalization
  b. Standardization

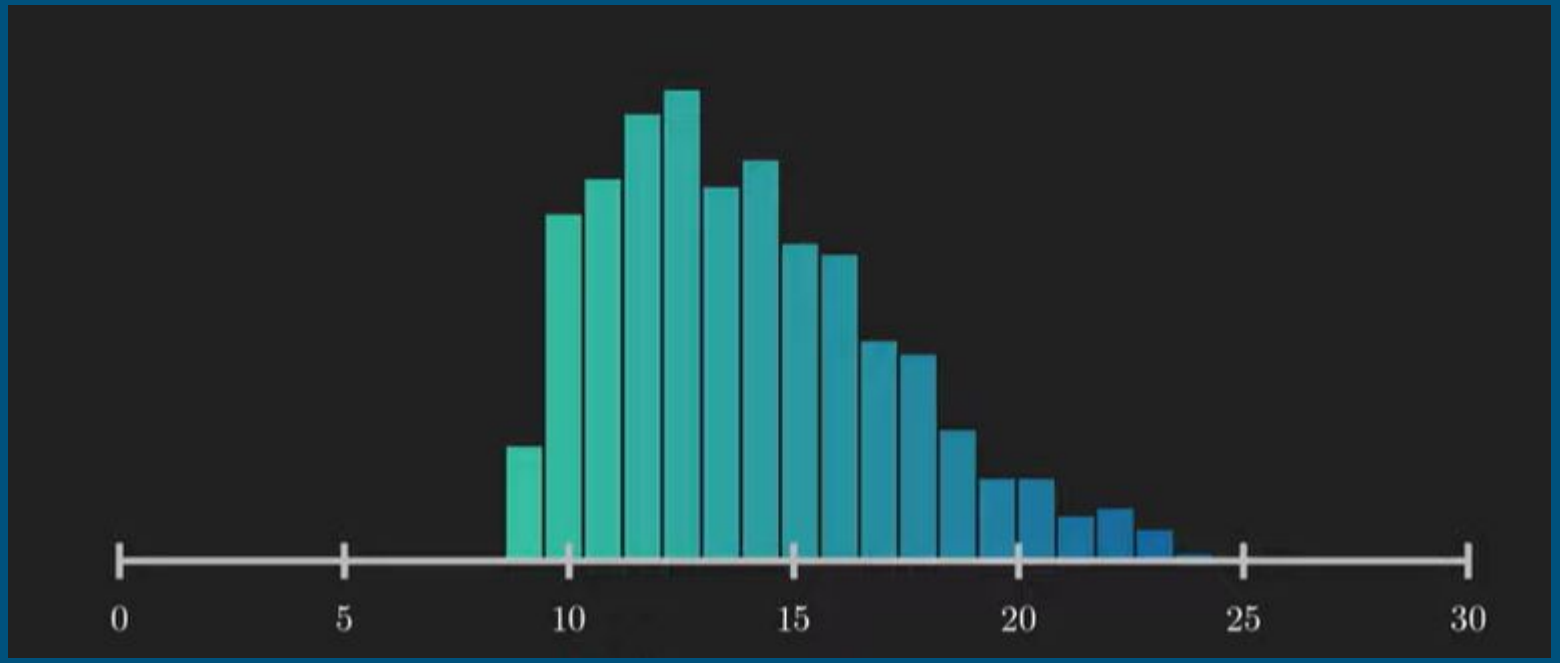
Normalization

- Maps the max number in dataset to 1
- Maps the smallest number in dataset to 0
- Lays everything else out proportionately between 0 and 1.

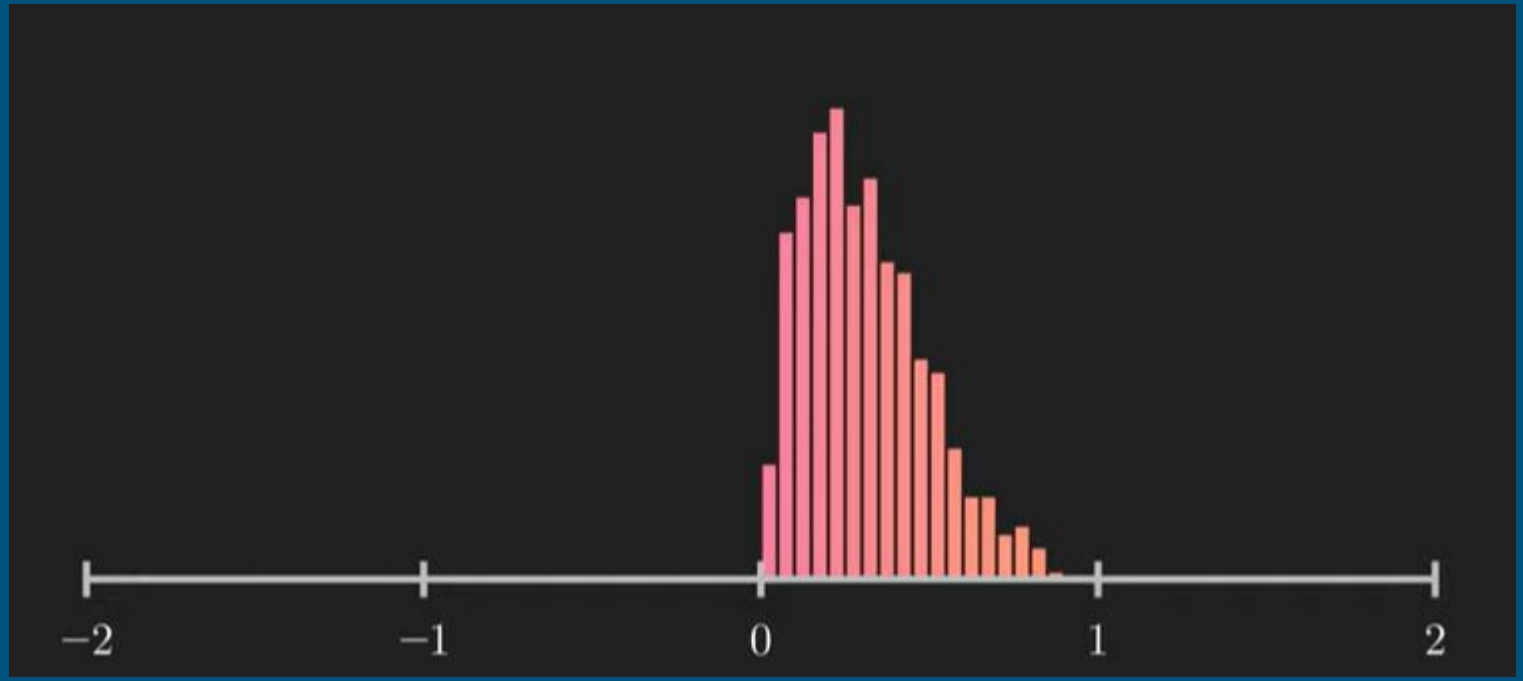
Normalization

$$x' = \frac{x - \min(x)}{\max(x) - min(x)}$$

Data before Normalization

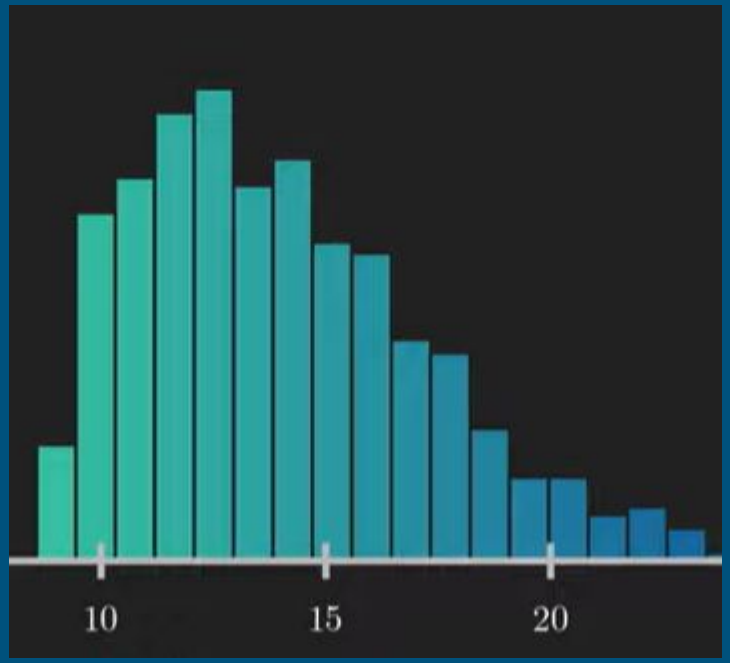


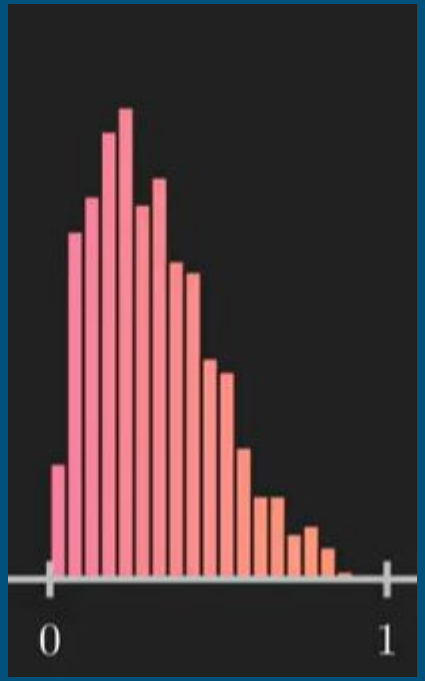
Data after Normalization



Comparison

- What do you notice?





Example

- Do it in notebook

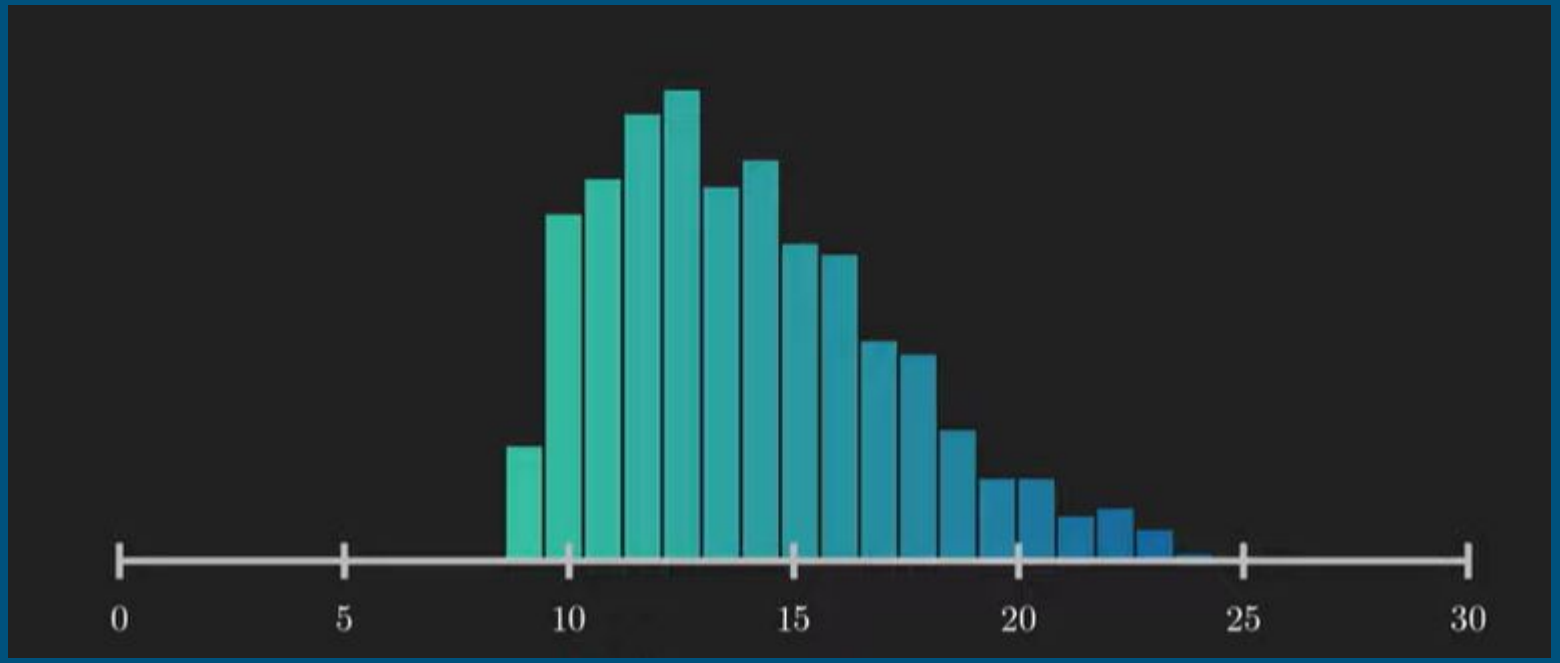
Standardization

- Rescales data so that
  - The mean of the data is 0
  - The standard deviation 1
- The min and max could be whatever

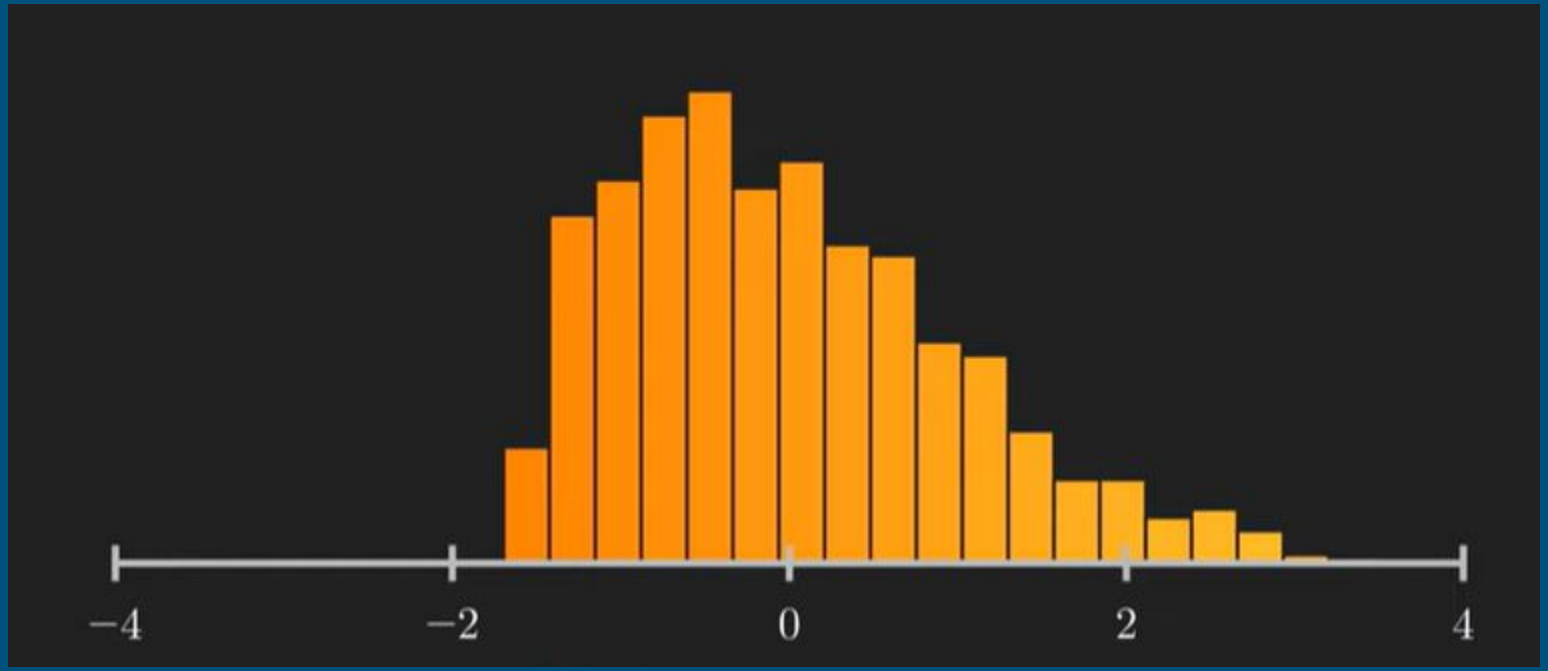
Standardization

$$z = \frac{x - \bar{x}}{\sigma}$$

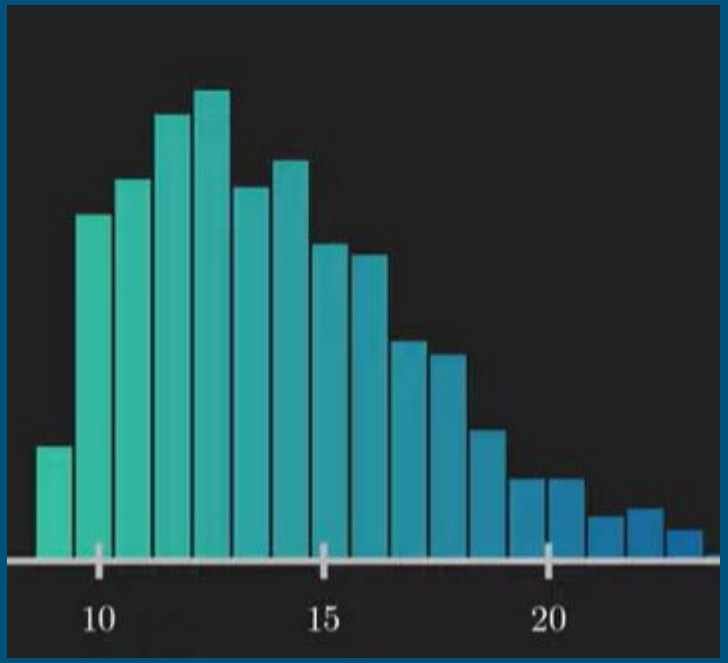
Data before Normalization

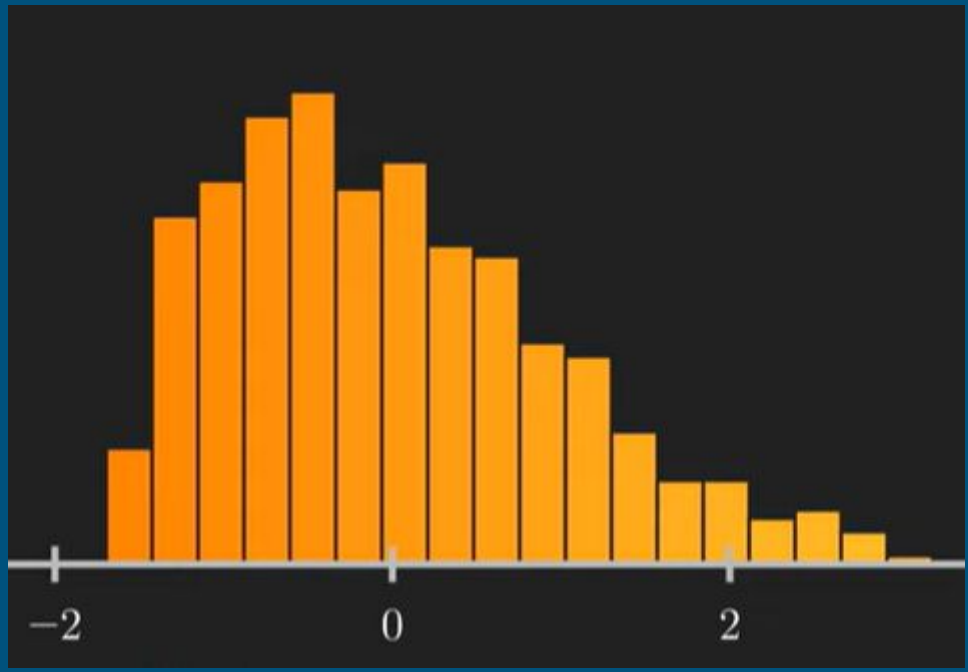


After Standardization



Comparison- What do you notice?





Example

- Do it in notebook