Introduction To Machine Learning

Lecture 1

Instructors

Priyanjana Bengani - pb2616@columbia.edu

Elizabet Doliar - ed2758@columbia.edu

Course Structure

- We will be using Python * Pandas * Scikit-learn * Tensorflow(maybe)
- We will mostly work with notebooks
- We will have HWs after every class / HWs are due at the beginning of next class

What is Machine Learning?

What is Machine Learning?

Arthur Samuel (1959):
Machine Learning is the field of study that gives the computer the ability to learn without being explicitly programmed.

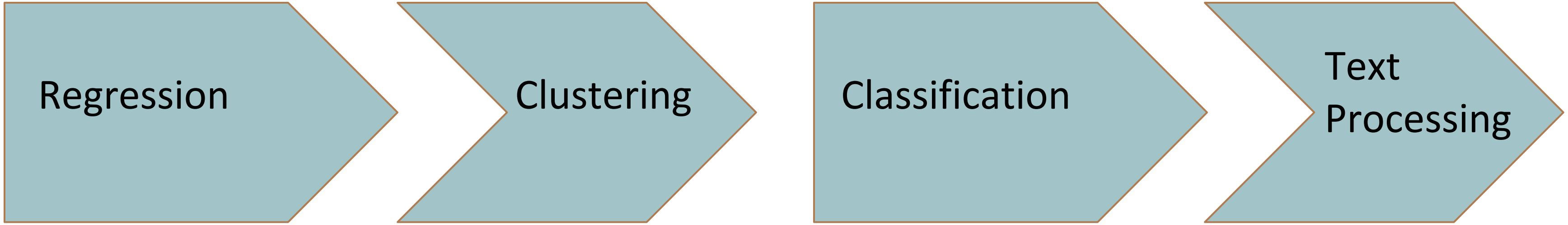
Kasparov vs Computer

Public awareness of AI increased greatly when an IBM computer named Deep Blue beat world chess champion Garry Kasparov in the first game of a match.

Most Common ML Myths

1. *Machine learning is magic, it is incomprehensible to humans*
2. *Machine learning is just about summarizing data*
3. *Machine learning can't predict previously unseen events*
4. *The more data you have, the more likely you are to predict nonexisting patterns*
5. *The patterns computers discover can be taken at face value*

Examples of Machine Learning Problems



Regression

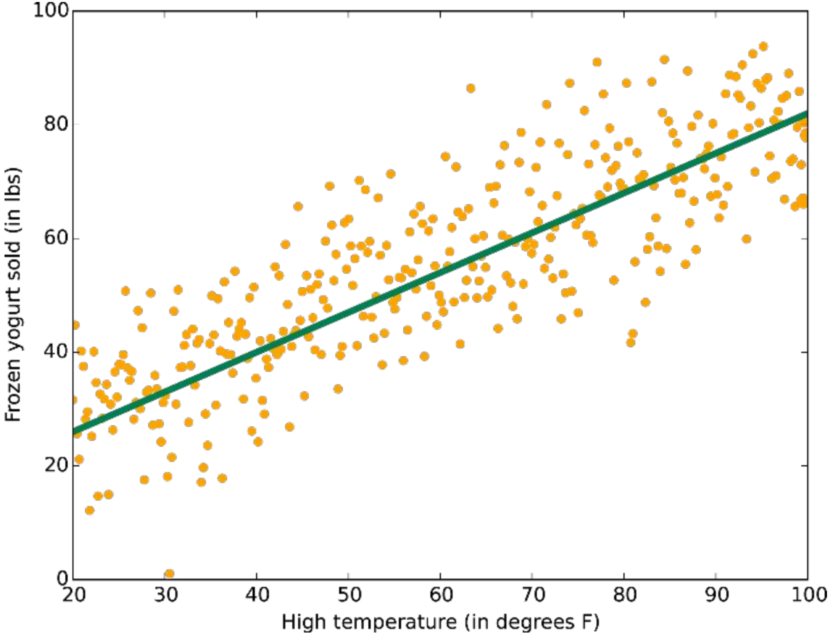
Clustering

Classification

Text Processing

Regression

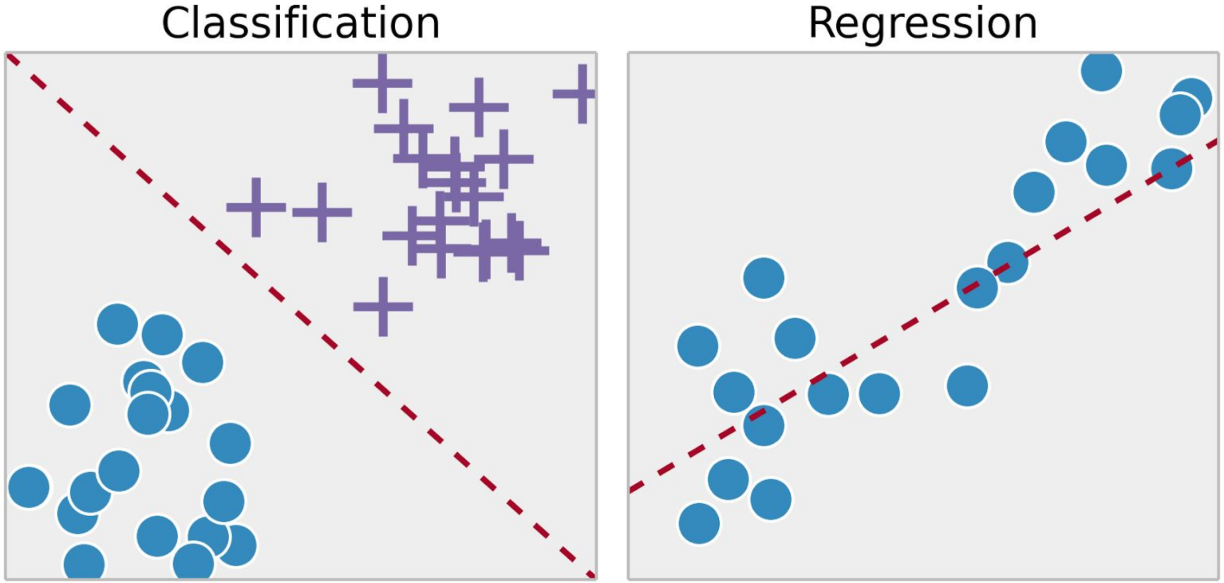
A **regression problem** is one where the output variable is a real or continuous value, such as “salary” or “weight”



Classification

A **classification problem** is when the output variable is a category, such as “red” or “blue” / “disease” or “no disease”.

Classification vs Regression



Exercise: Determine Regression vs Classification tasks

- Predicting age of a person
- Predicting nationality of a person
- Predicting whether stock price of a company will increase tomorrow
- Predict the number of copies a music album will be sold next month
- Predicting whether a document is related to sighting of UFOs?
- Predicting house price based on area

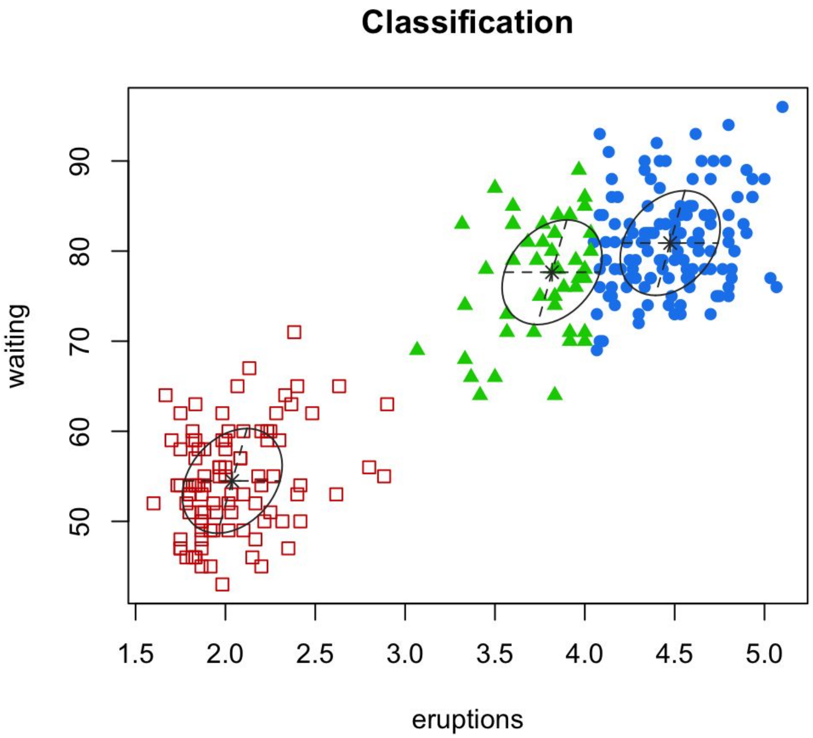
Clustering

Clustering is the assignment of a set of observations into subsets (called clusters) so that observations in the same cluster are similar in some sense.

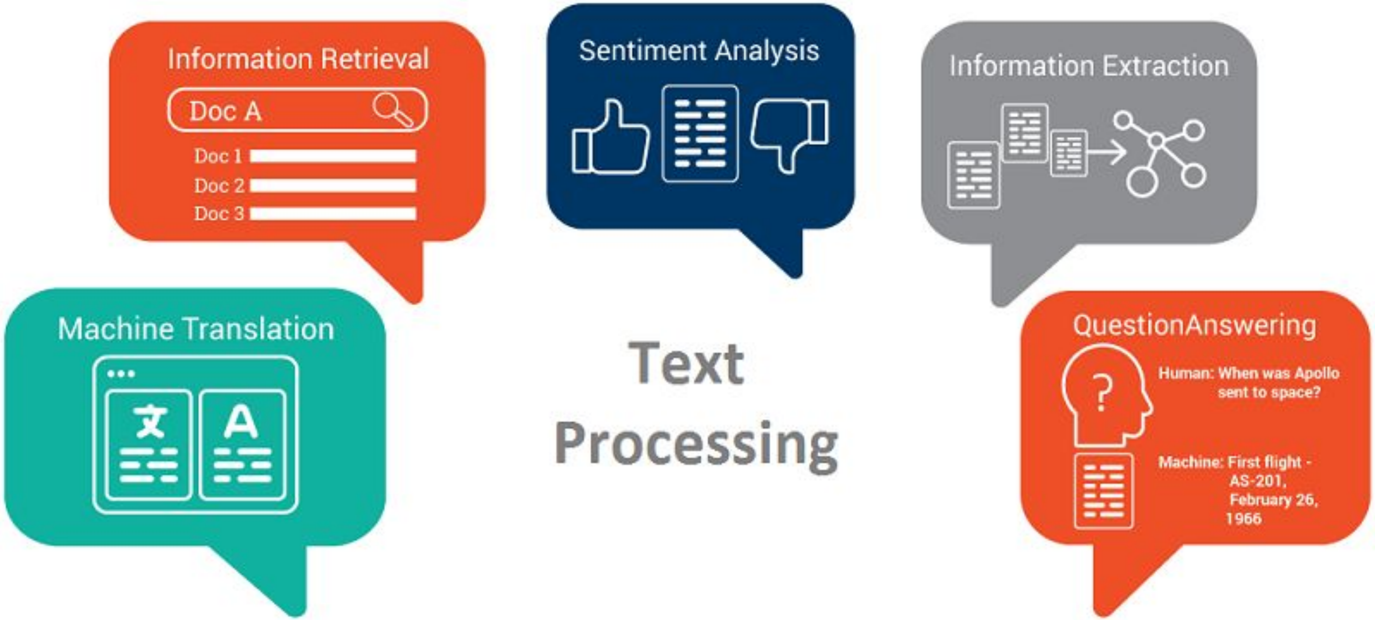
Clustering - Examples

- Understanding consumer preferences to scale up the business
- A Hospital Care chain wants to open a series of Emergency-Care wards, keeping in mind areas with highest number of accidents
- Identifying groups of houses according to their house type, value, and geographical location - city planning
- By learning the earthquake affected areas we can determine the dangerous zones.

Clustering


Classification
waiting
90
80
70
60
50
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
eruptions

Text Preprocessing



Discussion

How Machine Learning can be used in Journalism?

Influence of ML on Computational Journalism

- Change the set of tools journalists use to discover, tell or distribute stories - “reporting by, through and about algorithms.”

- Create visualizations that provide critical context for data

- Find ways to hold algorithms accountable - understanding algorithms that are delivering information to us

Discussion

Name 2 major expectations you have from this class

Topics Covered in This Course

I. Linear Regression
II. Logistic Regression
III. Decision Trees & Feature Engineering
IV. Vectors, Clusters, Visualisations
V. Overview, Sentiment Analysis, Text Processing & TF-IDF
VI. Entity Recognition & Topic Modeling
VII. Feature Engineering
VIII. Machine Bias

Questions?

Introduction to Statistics

Mean, Mode & Median

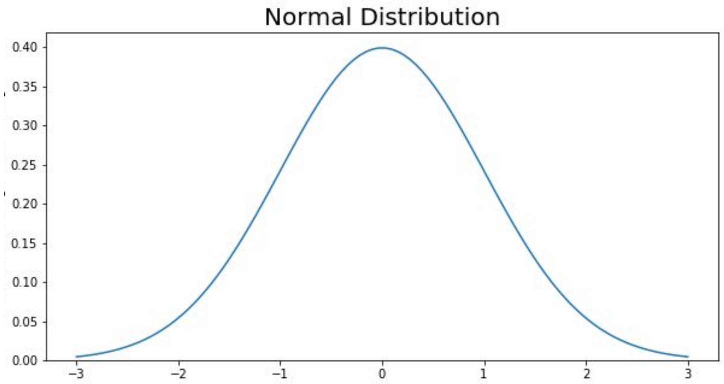
Mean: The "average" number; found by adding all data points and dividing by the number of data points.

Median: The middle number; found by ordering all data points and picking out the one in the middle (or if there are two middle numbers, taking the mean of those two numbers).

Mode: The most frequent number—that is, the number that occurs the highest number of times.

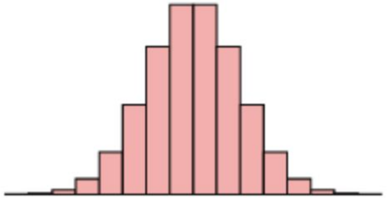
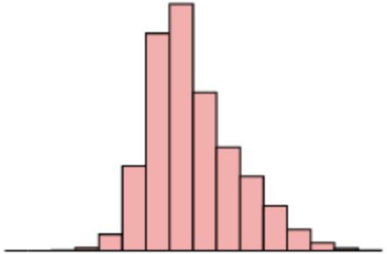
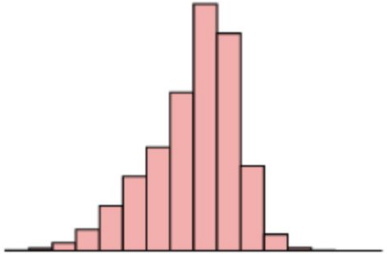
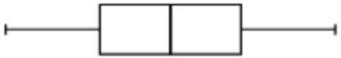
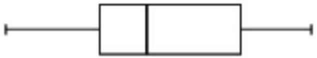
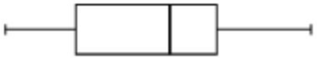
Normal Distribution

A normal distribution of data means that most of the examples in a set of data are close to the "average," while relatively few examples tend to one extreme or the other.



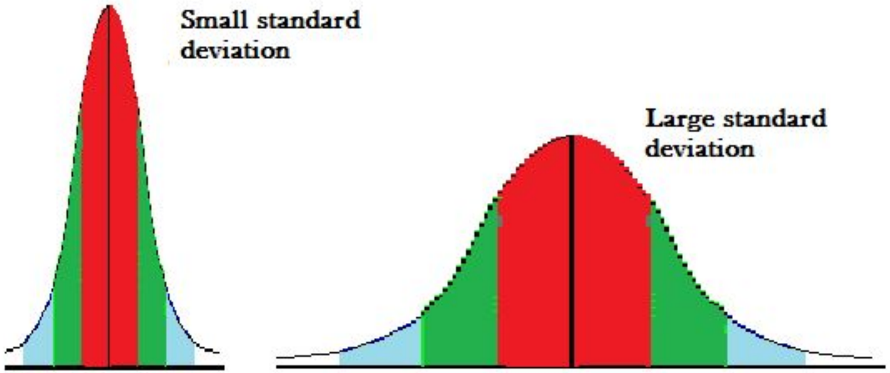
Skewness

Not all sets of data will have graphs that look this perfect. Some will have relatively flat curves, others will be pretty steep. Sometimes the mean will lean a little bit to one side or the other - Skewness

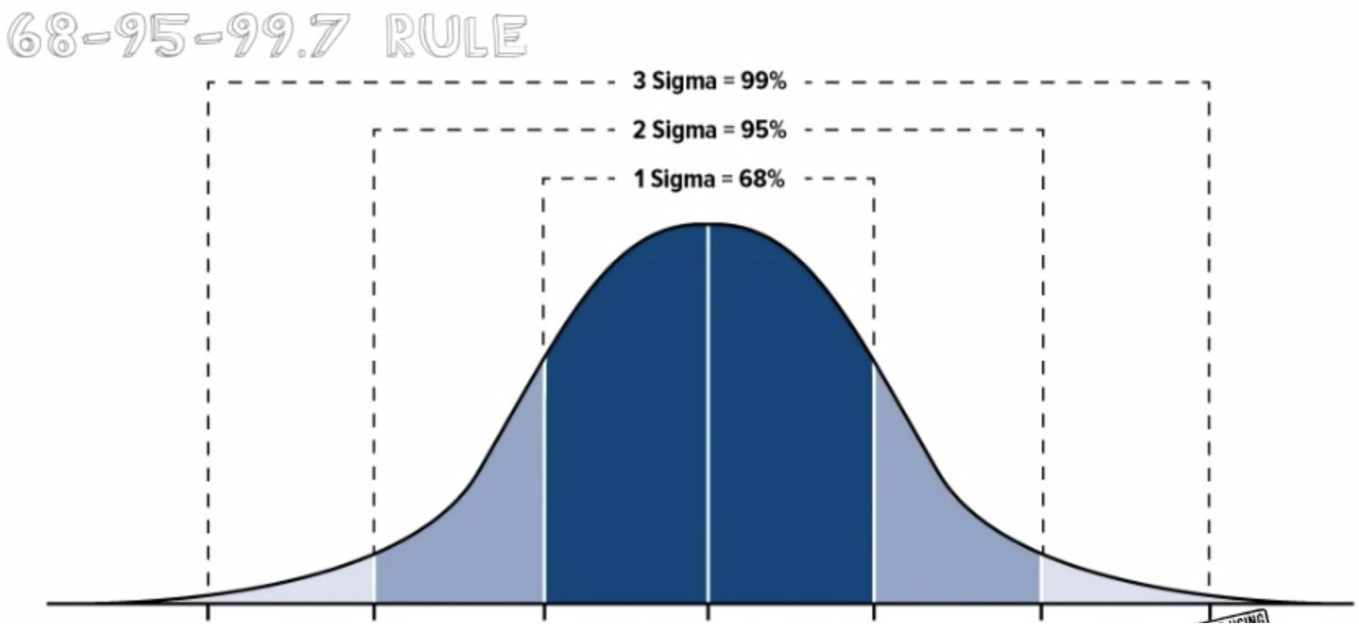
Symmetric	Skewed right (positive)	Skewed left (negative)
		
		

Standard Deviation

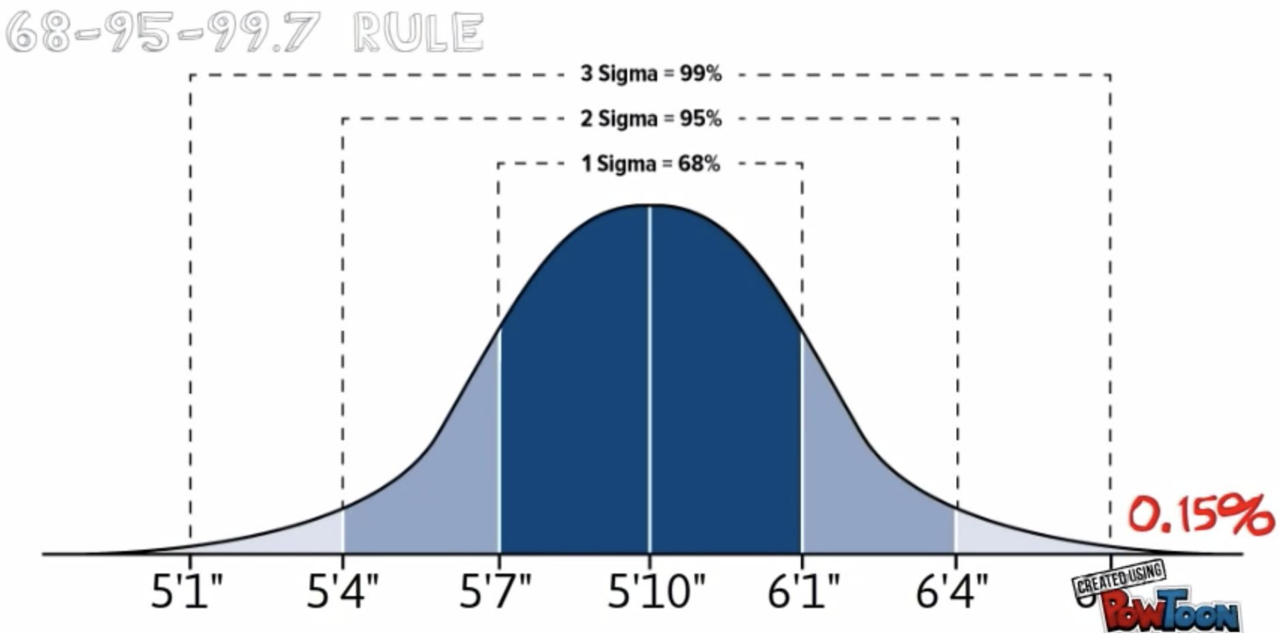
The **standard deviation** is a statistic that tells you how tightly all the various examples are clustered around the mean in a set of data.



Standard Deviation



Standard Deviation



Normalizing Data

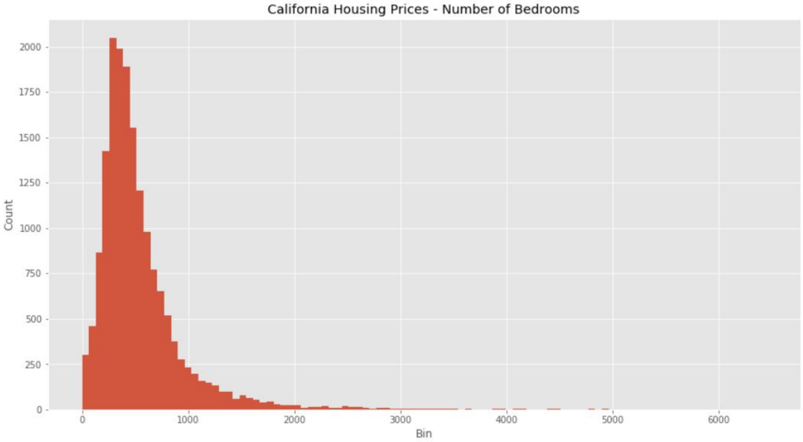
Normalization usually means to scale a variable to have values between 0 and 1

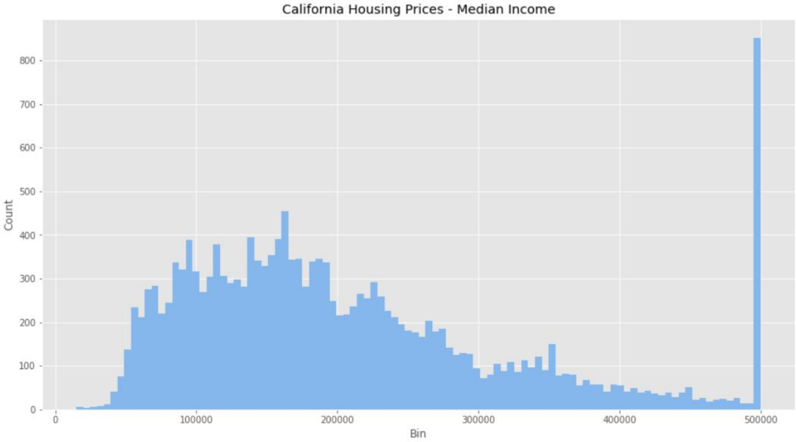
$$z = \frac{x - \min(x)}{\max(x) - \min(x)}$$

There are other formulas for normalization

Example

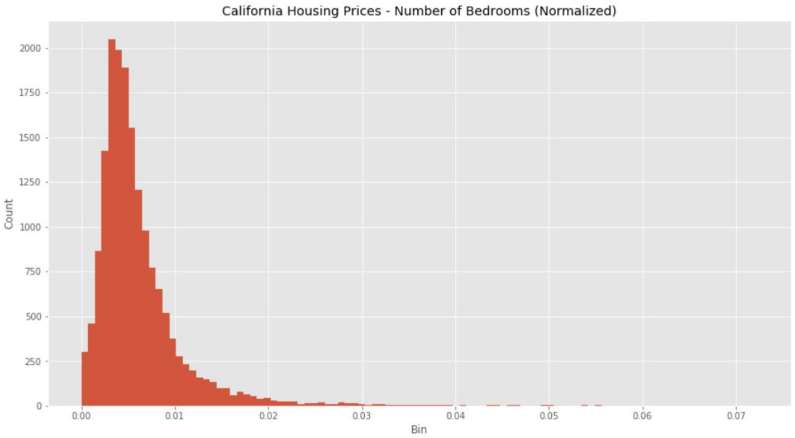
Let’s start by looking at both features without normalization.

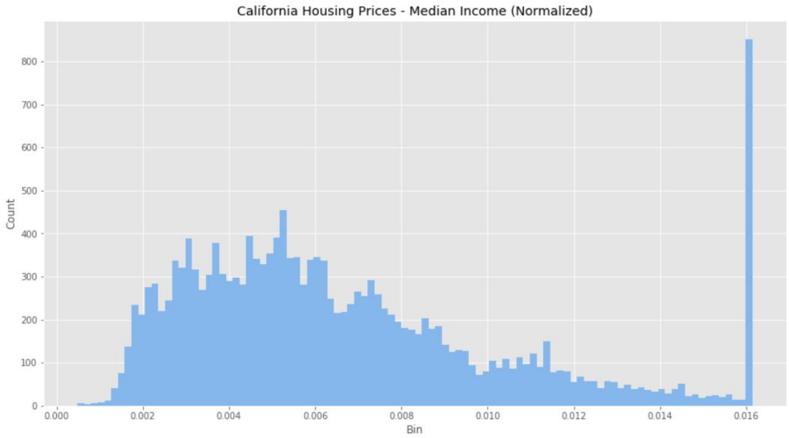




Raw Values

With Normalization


California Housing Prices - Number of Bedrooms (Normalized)


California Housing Prices - Median Income (Normalized)

Normalized Values

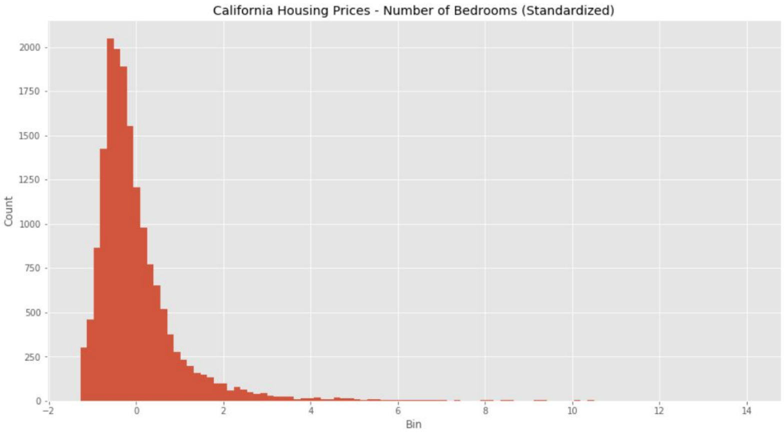
Standardizing Data

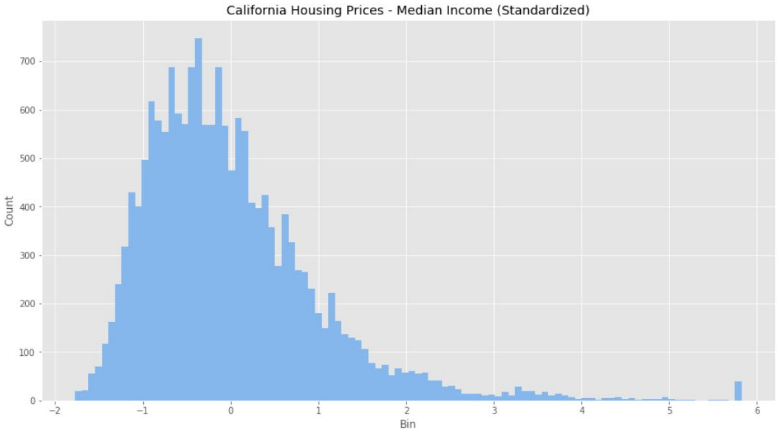
Standardization transforms data to have a mean of zero and a standard deviation of 1

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

Standardization Formula

Example: Standardization





Standardized Values

Questions?