8/21/23

Data Science Lifecycle

1. Data scientists use techniques to transform data into a visual representation that can be easier to understand by humans
2. Data science platform market is expected to grow by upwards of 20% annually.
3. Data science generally falls under math, statistics, and computer science.
4. The Life Cycle
    a. Question
    b. Collect Data
    c. Wrangle Data
    d. Analyze Data
    e. Visualize Information
    f. Communicate Information
5. All steps in the life cycle are all fluid

9/5/23

Python Fundamentals

1. Datasets
    a. The collection of data
    b. Types of datasets
        i. Lists
            1. Ordered, changeable, duplicates allowed
        ii. Dictionaries
            1. Ordered, changeable, duplicates not allowed
        iii. Sets
            1. Unordered, unchangeable*, duplicates not allowed
        iv. Tuples
            1. Unordered, unchangeable, duplicates allowed
2. Representing datasets with code
    a. Column-oriented
        i. Grouping by features
    b. Row-oriented
        i. Grouping by a single observation
3. Indexing

    a. Used to access values of a collection type
    b. Python syntax to access values
        i. List
            1. name[index]
        ii. Dictionary
            1. name[key]
        iii. Set
            1. for loop
        iv. Tuple
            1. Name[index]
4. Iteration
    a. Can repeat processes with loops or recursion in Python
    b. Python loop types
        i. While loop
            1. while condition: statements
        ii. For loop
            1. for thing in collection: statements
5. Useful methods
    a. Dictionaries
        i. values()
        ii. items()
        iii. keys()
    b. Lists
        i. len()
        ii. append()
        iii. sort()
    c. Other
        i. range()
        ii. print()
        iii. split()
        iv. type()
        v. int()
        vi. str()

9/5/23

Central Tendency

1. Measures of Central Tendency
    a. Statistical measures that help describe the behavior of a collection of data points
    b. Mean
        i. The average of all the values in a dataset
        ii. Summation of all the values, divided by the count of values
        iii. Can be misleading, because outliers skew the result
    c. Median
        i. The value in the direct center of a sorted dataset
        ii. Gives a more proportional representation of data that excludes outliers
    d. Mode
        i. The most frequently occurring value in a dataset
        ii. Most useful in a relatively large sample size
2. The center of a dataset is a good measure of determining the behavior or distribution of a dataset
    a. Gives examples of a whole dataset, not data points individually.
3. Distribution
    a. Shows how often data occurs in a dataset
4. Outliers
    a. Unusually large or small values
    b. Skew the result of a mean in a dataset
5. Bimodal
    a. When two values are most common
6. Unimodal
    a. When one value is the most common value
7. Symmetric distribution
    a. When the mean and median are the same
8. Skewed distribution
    a. When the dataset is offset by outliers, causing the mean to be an inaccurate representation of the population

9/15/23

Pandas Fundamentals

1. Pandas
    a. Pandas is a python library that can make analyzing data easier.
2. Dataframes
    a. A pandas object that is used to store a dataset
    b. Information is organized in rows and columns
    c. Dataframs simplify common operations such as sorting data
3. Series
    a. A pandas object used to create dataframes
    b. Seen as a one-dimensional list of data
        i. Think of it as a single column of a dataframe
4. Indexing into Dataframes
    a. df.loc[]
        i. name.loc[row_label, col_label]
    b. df.iloc[]
        i. name.iloc[row_index, col_index]
5. Selection
    a. The process of accessing a subset of a dataframe
    b. Can select subsets using loc and iloc
6. Filtering
    a. Selecting values of a dataset where certain conditions are true
    b. df[condition]
7. Combining Dataframes
    a. Concatenating
        i. Naively combines along an axis
    b. Merge
        i. Combine through a shared column
    c. Join
        i. Combine using shared indices
        ii. Inner Join
            1. Keep similar pieces
        iii. Left Outer Join
            1. Keep the left
        iv. Right Outer Join
            1. Keep the right

v. Full Outer Join
    1. Keep everything

9/19/23

Distributions

1. Distributions
    a. Graphs that tell us about some characteristic of a population
    b. Mean and median are important parts of the graphs
    c. Tells us about the shape and spread of data
2. Normal distribution
    a. The mean, median, and mode are all the same
    b. Empirical Rule
        i. 68% of data is within 1 standard deviation from the mean
        ii. 95% within 2 standard deviations
        iii. 99.7% within 3 standard deviations
    c. Unimodal
        i. Only one peak
3. Standard deviation
    a. The average distance between any point and the mean
4. Skewed distribution
    a. Skew is towards outliers
        i. Can be seen on graphs by a “tail”
5. Bimodal
    a. Has two peaks on a graph
6. Uniform distribution
    a. Each value has the same frequency