8/21/23

Data science Lifecycle

6 steps:

1. Question
2. Collect data
3. Wrangle data
4. Analyze data
5. Visualize information
6. Communicate information

9/5/23

Data sciences fundamentals

Collection types are very useful in data science because data sets are the collection of data.

Pythons collection types are:

- List
  Ordered, changeable,duplicates allowed
- Set
  Ordered, changeable, duplicates not allowed
- Dictionary
- Tuple

You can represent data sets with code using one of the following:

Column oriented- grouping by features

Row oriented- grouping by a single observation

Eg.

Name	City	Age
Matt	Los Angeles	27
Dave	San Francisco	30
Tim	Oakland	33

Row oriented would be

Matt, Los Angeles, 27 Dave ,San Francisco Tim, Oakland, 33

Column oriented would be

Matt, Dave, Tim Los Angeles, San Francisco, Oakland 27,30,33

Indexing

Inorder to access values in a collection type, we need to index into it

List – name[index(whole number only]
Dictionary – name[key]

Iterations

You can repeat processes with loops or recursion in python.

There are: for loops and while loops.

Useful methods

Dictionary; values(), items(), keys()

Lists : len(), append(), sort()

Other: range(), split(), type(), int(), str()

Central Tendency

Measures of central tendency

Mean, Median and Mode are statistical measures that help us describe the behavior of a collection of data points.

Pandas fundamentals.

Pandas:

A python library that can make analyzing data easier.

Dataframes:

- A pandas object that is used to store a dataset.
- Information is organized in rows and columns
- Dataframes simplify common operations, like sorting data.

Series:

- A pandas object used to create dataframes
- Seen as a one dimensional list of data.
  Its as if

Indexing into Dataframes

Main techniques

1. df.loc [ ]
2. df.iloc[ ]

19/9/23

Selection : the process of accessing a subset of a dataframe. You can select subsets using loc and iloc.

Filtering: Selecting values of a dataset where certain conditions are true.

Popular pattern;
Df[condition]

Combining data frames

Three techniques:

Concatenate: Naively combines along an axis

Merge: combine through shared column

Join: Combine using shared indices.

Inner join, outer, left, right,

Distributions:

Distributions are graphs that tell us about some characteristic of a population. Mean and median are important parts of these graphs!

Data visualizations.

A graph or picture that helps humans understand important patterns in a dataset.

Seaborn Fundamentals

A python library that can make visualizing data easier.

Bar Charts

A bar graph is a graph type that uses bars to depict a value associated with a category.

Histograms

A graph that shows the frequency distribution of a variable in a dataset
Sns.distplot(), sns.distplot(), sns.ecdfplot(), sns.kdeplot(), sns.rugplot(), sns.jointplot().

Scatter plots

A graph that uses points to show the relationship between 2 quantitative variables in a dataset.

Scatterplot(), lineplot(), stripplot(), swarmplot()

Techniquesn to collect data
Observe a sample-
Survey a sample- ask people directly. Also with questionnaires.

Experiment on a sample
Usedata someone else has responsibly collected

Sourcing Digital Datasets: API Requests:
The act of using HTTP requests in order to access datasets collected and maintained by other people.
Common API requests:
Get - see
Post - modify data and create resources
,Put,- modify data
Delete- remove existing resources

Sourcing Digital Datasets: Web Scraping
The act of extracting data from websites using the structure of its HTML.