PRESENTATION ABOUT PANDAS

Made by : Rawan Hatem

INTRODUCTION

- pandas is a data manipulation library in Python.
- loads the survey results CSV file into a pandas data frame using the command DF = PD.read_csv('path/to/file.csv').
- the ‘shape’ attribute uses to see the number of rows and columns in the data frame.
- the ‘info’ method uses to see the number of rows, columns, and data types of each column.
- sets the maximum number of columns displayed in the Jupyter notebook to 85 using the command PD.set_option('display.max_columns', 85).
- sets the maximum number of rows displayed in the Jupyter notebook to 85 using the command PD.set_option('display.max_rows', 85).

DATA FRAMES & SERIES

- Data Frames
  - A data frame is a 2-dimensional data structure with rows and columns
  - It's similar to a dictionary of lists, but with more functionality
  - Data frames can be created from dictionaries
  - Each column in a data frame is a series object
- Series Objects
  - A series is a 1-dimensional array of data
  - It's similar to a list, but with more functionality
  - Series objects have an index, which can be used to access specific values
- Accessing Data in Data Frames
  - Data can be accessed using bracket notation (e.g. DF['column_name'])
  - Multiple columns can be accessed by passing a list of column names
  - Rows can be accessed using the loc or iloc indexers
  - loc uses labels, while iloc uses integer locations
  - Slicing can be used to access multiple rows and columns

SET CUSTOM INDEXES

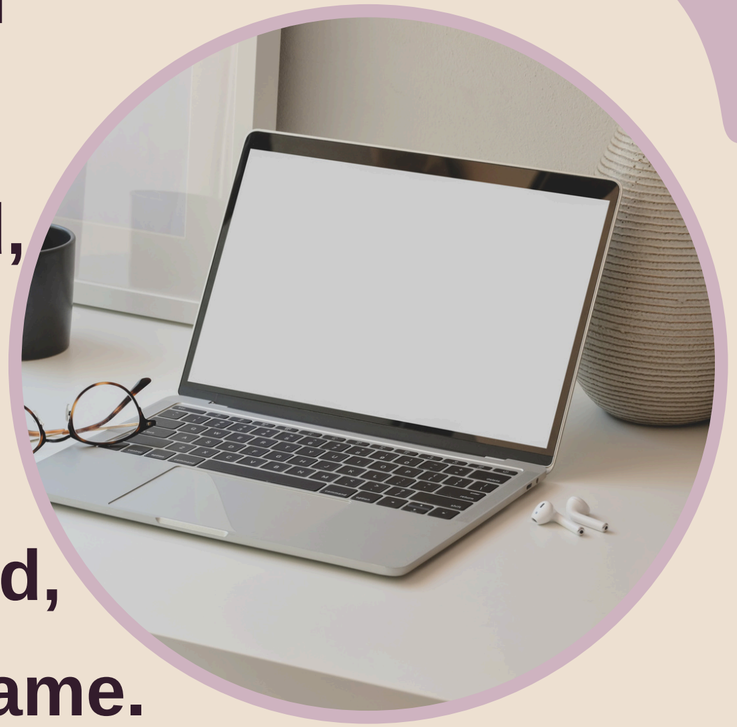
- To set column as the index, using the set_index() method and passes in the name of the column.
- pandas does not enforce indexes being unique, but most of the time they will be.
- to reset the index to its default value we using the reset_index() method.
- to sort the index alphabetically we using the sort_index() method.
- setting the index can make it easier to search for specific rows or columns by using the loc() method.

FILTERING DATA

- Filtering is done to view specific data from the dataframe by excluding unwanted data.
- A basic comparison in pandas returns a series object with true/false values, where true values correspond to the rows that met the filter criteria and false values correspond to the rows that didn't meet the filter criteria.
- The filter can be applied to the dataframe to return all the rows that meet the filter criteria.
- The ‘and’ and ‘or’ operators can be used in filters to get specific results.
- The ‘tilde’ symbol can be used to get the opposite of a filter.

ADDING & REMOVING VALUES

- Adding columns is done by creating a new column and passing in a series of values .
- Removing columns is done using the drop method, which can be set to in-place to make the changes permanent.
- Adding rows can be done using the append method, which can append a single row or an entire data frame.
- Removing rows can be done using the drop method, which can drop rows based on index or a conditional.



UPDATING AND MODIFING

- Updating a single row of data :
  - update all columns values in the row e.g.(df.loc[2] = [’john’, ’Simth’,’john@gmail.com’]
  - updating a specific columns values in specific row e.g.(df.loc[2,[’first,’last’]]= [’John’,’Smith’]
  - updating one value e.g.(df.loc[2,’last’] = ’John’
- Updating a multiple rows of data :
  - change all column values to lower case
    - e.g.(df[’email’] = df[’email’].str.lower()

SORTING DATA

- To sort a data frame by a single column, you can use the sort_values() method and set the by argument to the column name.
- To sort in descending order, pass ascending=False to the sort_values() method.
- To sort by multiple columns, pass a list of column names to the by argument.
- If you want to sort on multiple columns with different orders, pass a list of boolean values to the ascending argument, where True means ascending order and False means descending order.
- To save the sorted data frame permanently, set the inplace=True argument in the sort_values() method.
- If you only want to sort a single column, you can access that column directly and use the sort_values() method on the series.
- To view the largest or smallest values from a data frame, use the nlargest() or nsmallest() methods on the specific column.
- To get the entire row of the largest or smallest values, use the nlargest() or nsmallest() methods on the data frame and pass the column name to the n argument.

METHODS

- the value_counts()
  - method in pandas to count the occurrences of each item
- the idxmax() :
  - finding the index with the maximum count using the idxmax() method.
- the nunique()
  - method to count the number of unique values in column