DSO 545: Statistical Computing and Data Visualization

Abbass Al Sharif

Fall 2019

Lab 02: Data Manipulation using Pandas

Contents

- Introduction
- Pandas Python Package
- Pandas Series Data Structure
- Pandas DataFrame Data Structure

---

1. Import the `pandas` Python package. If you don't have the package installed, then you might use the following line to install it using the terminal (Mac) or command prompt (windows):

Pandas Series Data Structure

2. Create two series `name` ("Dona", "James", "Alice", "USC Marshall") and `age` (25, 35, 19, 99).
3. Update the last element of the `age` series to 100.
4. Find the average, min, and max age.
5. Get the summary statistics for age.
6. Find all ages that are below the average age.

Pandas DataFrame Data Structure

7. Create a dataframe called `data` that has the two series `name` and `age` as its columns.
8. Add a third column ("Person") to the dataframe defined before. The cells in this column has a value of "Yes" if the corresponding entity is a person, and "No" otherwise.
9. Rearrange the columns in the dataframe as follows: Name, Person, and Age.
10. Load the dataset `orders.csv` into a Pandas dataframe called `orders`.
11. Print the first five observations in the orders dataset.
12. Print the last 10 observations in the orders dataset.
13. For the `orders` dataframe, run the two methods `.describe()` and `.info()`. What does each method do?
14. Extract the customer column from the orders dataframe.
15. What are the different products in the orders dataset?
16. Create a dataset (called data) that has only the customers, products, and sales information from the orders dataset.

Indexing Techniques

There are two indexers that are very useful in slicing and dicing your dataset.

- loc[ ] : subset using index value/label
- iloc[ ] : subset using index position (integer)

17. Print the sales value for the first column in the orders dataset.
18. Print the sales, costs, and sales date for the second order in the orders dataset.
19. Print the sales, costs, and sales date for the **second and fifth** orders:
20. Print all information for the first order.
21. Print the sales for the first 15 observations using both `iloc[]` and `.loc[]`.

Vectorized Computations in Python with Pandas

22. Create a list called ages (20, 45, 23, 40, 26).
23. Add one year to each age in the `ages` list created in the previous question.
24. Create a pandas Series called ages (20, 45, 23, 40, 26).
25. Add one year to each age in the `ages` pandas series created in the previous question.

Conditional Subsetting the Data

26. Create a variable called `myage` which stores `78`. What is the output of the following Python code (`age < 90`)?
27. What is the output of comparting the `ages` pandas series to 30 (`ages <30`)?
28. Find all ages in the `ages` pandas series defined earlier that are less than 30.
29. Which orders in the `orders` dataset has sales greater than or equal to $97,000?
30. Create a subset dataset called (orders_GIN) where the customer is `GIN ON THE RUN CO` and include only the columns: products, sales, and costs.

Numerical Summary Statistics

31. What are the average, median, min, and max sales for “SOFT DRINKS” for all years.
32. What is the 25th percentile for all the sales for “SOFT DRINKS” sold by “Michael Jackson”?