assignment_2

December 11, 2022

1 Assignment #2

1.1 Pandas and Visualization

1.1.1 Getting Data

Select a dataset from Toronto Open Data or another data portal of your choice, and download it. Some suggested datasets are linked below and additionally available for download in the course repo /data folder. A good dataset for this exercise will have a mix of data types.

Some sugested datasets: * TTC bus delays: Fewer columns, not well documented, some NaNs. Similar to data we've worked with in class. Recommend choosing a full year of data. * Apartment building evaluations: Lots of columns, well-documented, some NaNs. * Daily shelter overnight service occupancy and capcity: The largest of the datasets suggested. Lots of columns, well-documented, more NaNs.

1.1.2 Metadata Review (3 marks)

1. What organization publishes this dataset?
2. How frequently is the dataset updated?
3. What metadata is available (e.g., column names, data types, descriptions)?
4. Is there documentation about who or what produces the data? About who collects it? Through what processes?
5. Is there documentation about limitations of the data, such as possible sources of error or omission?
6. Are there any restrictions concerning data access or use? (e.g.,registraton required or non-commercial use only)

1.1.3 Getting started (3 marks)

1. Load the data to a single DataFrame.
2. Profile the DataFrame.

- What are the column names?
- What are the dtypes when loaded? Do any not make sense?
- How many NaNs are in each column?
- What is the shape of the DataFrame?

3. Generate some summary statistics for the data.

- For numeric columns: What are the max, min, mean, and median?
- For text columns: What is the most common value? How many unique values are there?
- Are there any statistics that seem unexpected?

4. Rename one or more columns in the DataFrame.
5. Select a single column and find its unique values.
6. Select a single text/categorical column and find the counts of its values.
7. Convert the data type of at least one of the columns. If all columns are typed correctly, convert one to `str` and back.
8. Write the DataFrame to a different file format than the original.

1.1.4 More data wrangling, filtering (3 marks)

1. Create a column derived from an existing one. Some possibilities:

- Bin a continuous variable
- Extract a date or time part (e.g. hour, month, day of week)
- Assign a value based on the value in another column (e.g. TTC line number based on line values in the subway delay data)
- Replace text in a column (e.g. replacing occurrences of "Street" with "St.")

2. Remove one or more columns from the dataset.
3. Extract a subset of columns and rows to a new DataFrame

- with the `.query()` method and column selecting `[[colnames]]`
- with `.loc[]`

4. Investigate null values

- Create and describe a DataFrame containing records with NaNs in any column
- Create and describe a DataFrame containing records with NaNs in a subset of columns
- If it makes sense to drop records with NaNs in certain columns from the original DataFrame, do so.

1.1.5 Grouping and aggregating (3 Marks)

1. Use `groupby()` to split your data into groups based on one of the columns.
2. Use `agg()` to apply multiple functions on different columns and create a summary table. Calculating group sums or standardizing data are two examples of possible functions that you can use.

1.1.6 Plot (3 Marks)

1. Plot two or more columns in your data using `matplotlib`, `seaborn`, or `plotly`. Make sure that your plot has labels, a title, a grid, and a legend.

1.2 References

1.2.1 Data Sources

- Open Data Toronto. *TTC Bus Delay Data*. https://open.toronto.ca/dataset/ttc-bus-delay-data/
- Open Data Toronto. *Apartment Building Evaluation*. https://open.toronto.ca/dataset/apartment-building-evaluation/
- Open Data Toronto. *Daily Shelter & Overnight Service Occupancy & Capacity*. https://open.toronto.ca/dataset/daily-shelter-overnight-service-occupancy-capacity/