WeJapa-DS-Capstone_Project:

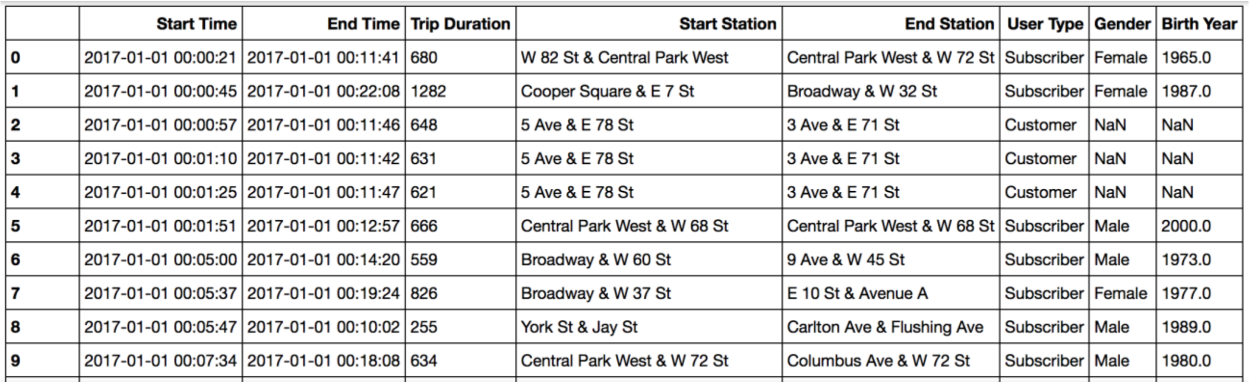
Basic Data Exploration with pandas on Bikeshare Data

Project Overview:

Project Overview   In this project, I made use of Python to explore data related to bike share systems for three major cities in the United States—Chicago, New York City, and Washington. I wrote codes to import the data and answer interesting questions about it by computing descriptive statistics. I wrote a script that takes in raw input to create an interactive experience in the terminal to present these statistics.

Bike Share Data Over the past decade, bicycle-sharing systems have been growing in number and popularity in cities across the world. Bicycle-sharing systems allow users to rent bicycles on a very short-term basis for a price. This allows people to borrow a bike from point A and return it at point B, though they can also return it to the same location if they'd like to just go for a ride. Regardless, each bike can serve several users per day.

Thanks to the rise in information technologies, it is easy for a user of the system to access a dock within the system to unlock or return bicycles. These technologies also provide a wealth of data that can be used to explore how these bike-sharing systems are used. In this project, you will use data provided by Motivate - h ttps://www.motivateco.com/ ...a bike share system provider for many major cities in the United States, to uncover bike share usage patterns. You will compare the system usage between three large cities: Chicago, New York City, and Washington, DC.



	Start Time	End Time	Trip Duration	Start Station	End Station	User Type	Gender	Birth Year
0	2017-01-01 00:00:21	2017-01-01 00:11:41	680	W 82 St & Central Park West	Central Park West & W 72 St	Subscriber	Female	1965.0
1	2017-01-01 00:00:45	2017-01-01 00:22:08	1282	Cooper Square & E 7 St	Broadway & W 32 St	Subscriber	Female	1987.0
2	2017-01-01 00:00:57	2017-01-01 00:11:46	648	5 Ave & E 78 St	3 Ave & E 71 St	Customer	NaN	NaN
3	2017-01-01 00:01:10	2017-01-01 00:11:42	631	5 Ave & E 78 St	3 Ave & E 71 St	Customer	NaN	NaN
4	2017-01-01 00:01:25	2017-01-01 00:11:47	621	5 Ave & E 78 St	3 Ave & E 71 St	Customer	NaN	NaN
5	2017-01-01 00:01:51	2017-01-01 00:12:57	666	Central Park West & W 68 St	Central Park West & W 68 St	Subscriber	Male	2000.0
6	2017-01-01 00:05:00	2017-01-01 00:14:20	559	Broadway & W 60 St	9 Ave & W 45 St	Subscriber	Male	1973.0
7	2017-01-01 00:05:37	2017-01-01 00:19:24	826	Broadway & W 37 St	E 10 St & Avenue A	Subscriber	Female	1977.0
8	2017-01-01 00:05:47	2017-01-01 00:10:02	255	York St & Jay St	Carlton Ave & Flushing Ave	Subscriber	Male	1989.0
9	2017-01-01 00:07:34	2017-01-01 00:18:08	634	Central Park West & W 72 St	Columbus Ave & W 72 St	Subscriber	Male	1980.0

REQUIREMENTS:

- Language: Python 3.8 or above
- Libraries: Pandas, Numpy, time
- Editors/ Development Environment: Anaconda, a text editor like Sublime - https://www.sublimetext.com/ or Atom - https://atom.io/ A terminal application (Terminal on Mac and Linux or Cygwin on Windows).

THE DATASETS

Randomly selected data for the first six months of 2017 are provided for all three cities. All three of the data files contain the same core **six (6)** columns:

- Start Time (e.g., 2017-01-01 00:07:57)
- End Time (e.g., 2017-01-01 00:20:53)

- Trip Duration (in seconds - e.g., 776)
- Start Station (e.g., Broadway & Barry Ave)
- End Station (e.g., Sedgwick St & North Ave)
- User Type (Subscriber or Customer)

The Chicago and New York City files also have the following two columns:

- Gender
- Birth Year

STATISTICS COMPUTED

You will learn about bike share use in Chicago, New York City, and Washington by computing a variety of descriptive statistics. In this project, you'll write code to provide the following information:

Popular times of travel (i.e., occurs most often in the start time)

- most common month
- most common day of week
- most common hour of day

Popular stations and trip

- most common start station
- most common end station
- most common trip from start to end (i.e., most frequent combination of start station and end station)

Trip duration

- total travel time
- average travel time

User info

- counts of each user type
- counts of each gender (only available for NYC and Chicago)
- earliest, most recent, most common year of birth (only available for NYC and Chicago)

PROJECT DATA:

- chicago.csv - Stored in the data folder, the chicago.csv file is the dataset containing all bikeshare information for the city of Chicago provided by Udacity.
- new_york_city.csv - Dataset containing all bikeshare information for the city of New York provided by Udacity.
- washington.csv - Dataset containing all bikeshare information for the city of Washington provided by Udacity. Note: This does not include the 'Gender' or 'Birth Year' data.

NOTE:

In this program we were was required to create and app from which the user can access the program and at the end be able to see the raw data of the selected city.

SOURCES:

Pandas Documentation: https://pandas.pydata.org/panda-docs/stable/user_guide/10min.html

https://pandas.pydata.org/panda-docs/stable/user_guide/timeseries.html

https://pandas.pydata.org/panda-docs/stable/user_guide/io.html

Python Docs: https://docs.python.org/3/library/

Others:

Stackoverflow

www.geeksofgeeks.org/convert-the-column-type-from-strings-to-datetime-format-in-pandas-dataframe/

www.interviewqs.com/ddi_code_snippets/extract_month_year_pandas

Udemy:

Python for Data Science and Data Analysis Masterclass (2020) by Vijay Gadhave (Mostly Section 7)

Basic Data Science with Numpy, Pandas and Matplotlib by Akbar Khan (Mostly Section 2)

FINAL REPORT

To complete this project, I used my knowledge of Python programming and Pandas gained throughout the study sessions of the WeJapa Data Science. I also referred to very many various sources for help along the way.

When I saw the project overview, it looked relatively easy on paper until I opened the bikeshare-2.py file. I realized that I had some knowledge gaps to fill.

I encountered a lot of bugs along the way especially when dealing with dataframe related components, the pandas documentation and some tutorial videos helped me with dealing with **Datatime**.

One of the biggest issues I faced was when trying to extract the day of the week from Start time, I was using the dt.weekday_name. took a fews days and help from a teammate to know that the dt.weekday_name doesn’t work rather dt.day_name.

Most of the other issues I encountered where related to time conversions, apparently that is an area I need to learn more about. This project was **HARD** but a real eye opener on what one can encounter on the field.

I learnt a lot, even new ways for asking questions on stackoverflow, a new YouTube person for my data science related tutorials and patience. It was quiet the learning curve.