Citibike Technical Write-Up

1. Introduction: Project completed during a 3 month immersive Data Analytics course from General Assembly. Students were tasked to analyze a large dataset (1.3M+ records) containing route information collected from users by several RideShare companies. The project was done over the course of two weeks while simultaneously learning many of skills that were required to complete the project (that were new to me at the time).

2. Background: This is the prompt students were given:

“Study the popular routes, population centers and destination attractions, for both amusement and employment. Create recommendations based on the data for adding new stations. Include in your evaluation operational research estimated cost of opening a new station and the amount of time and activity required to return to breakeven.

Based on rider patterns, discover any stations that become excessively popular destination stations. Contrast those statistics with the most popular trip start locations. Recommend a program for redistributing the bike inventory to balance the supply with demand. Consider the factors surrounding the identified stations. Does that research provide insight on the variables involved?”

3. Technical details: A total of 1.3M records were cleaned and analyzed. Four databases, six tables each for the years 2016 through 2019. Python was used for cleaning, SQL (PostgreSQL) was used for querying, and Tableau was used to visualize the data.

4. Implementation: I remember feeling overwhelmed yet excited since this was my first real project analyzing data. The fact that it was over a million rows left me with a constant fear that I would skew the data by forgetting to include important parameters in my queries. Luckily my instructors were very helpful with extensive knowledge about the dataset and would alert me whenever something looked “off”. This allowed me to focus on understanding the data and asking better questions.

5. Results: Looking back at my first project I can see I did not make the best use of space, ideas were redundant, and the code was not optimized. Ultimately this was done in two weeks while learning SQL and Python, and as far as objective it answered the prompt, giving actionable insights.

6. Conclusion: I was able to learn a lot with this project, and have greatly improved technically and analytically since.

7. References: 2016-2019 Citibike Trip data

Power BI Desktop, PostgreSQL, Jupyter Notebook, Python, SQL