Recommendations to NYC citi bike sharing program to mitigate shortfall in sponsorship revenue

Anuj Agrawal

Recommendations to NYC citi bike sharing program

- From EDA it is evident that there are many stations which are low traffic and rake up a lot of operational costs while contributing less to the revenue
- Ranked the stations in order to total start + total end trips = defined as total trips for that station
- Mean Ridership = 28000, Median Ridership = 15000 , 25th Percentile = 4000 trips for whole of 2016 to 2017
- There has to be a fixed cost as well as variable costs of operating a station
- Assumptions: Fixed Cost- Land + Upkeep = 50000 $ per year per station, Variable Costs: Number of Bike * 100 as maintenance for bikes, Number of Docks * 100 as maintenance for Docks
- Assumptions: Revenue per Ride is: 5 $ per Ride
- Assumptions: Upon Selling: 1 Bike is worth 400$, Docs is worth 200$,

Recommendations to NYC citi bike sharing program

- My Idea is to recommend close down stations which are low in ridership and thereby are low margin for the company. It depends on short team priority of the company as to how many stations they would like to close. There is a threshold of ridership above which the station would start generating positive cash flow, ideally stations below this threshold should be closed
- However, a further improvement to this analysis is to take into account the distance of the nearest stations to every station which is being considered for closure. If the distance to nearest station is too high, it might be a good idea to keep the station open to allow for expansion into underpenetrated geographies in the future - i have created a dataframe which has the distance and name/location of the closest station for all the stations

Recommendations to NYC citi bike sharing program

- Based on the analysis the average trip distance is : 1.87 kms , median trip distance is:
- For every station we can calculate this KPI for every station:
- Annual Costs: = 50000 + 100* Number of Bikes + 100* No of docks
- Revenue: Annual Ridership * 5
- Closing Renevue: 400 * Number of Bikes + 200 * Number of Docks
- There can be other ways a station generates income like kiosks, partnerships etc which are beyond the scope of this analysis
- The numbers are just my assumptions and the results with really vary a lot based on the correct accurate assumptions
- My final table also gives us a rough estimate of the annual ridership % loss based on the number of stations that we choose to close down

Final DataFrame :

- I have put together a final dataframe which has these columns for every station in order to aid the decision making
- Station_id, Station_name, Station_location (lat & lon), Annual Ridership - Sum of rides originating from the station + sum of rides returning to the staton,, Overall Ridership % of the entire network, no of bikes, no of docks,
- Based on my assumptions on Revenue per ride, fixed and variable costs - it can be easily calculated whether a station is profitable or not and by how much, what is its margin per ride, and the money saved in the near term on closing it
- As mentioned before another factor in analysis is to keep in mind the average trip distance and whether there is a station closeby or not as it might not be worth closing a station which is very distant from others

Some Notes on my sql and python code:

- All of my Code, including Bigquery SQL scripts and python code for deep dive analysis is in the jupyter notebook
- GCP has a feature to use bigquery sql code using %%bigquery inside the notebook and save it as a dataframe
- I would request if i can be given an opportunity to present my work and walk through the team on it, as i will be able to present my ideas and methodology
- I have included some sql code which use self joins, complex joining conditions, subqueries, CTEs, window functions in a separate file.
- Despite the 300$ GP credit and the free cluster i ran into some issues as there were some out of memory issues while doing some calculations in pandas
- Data Quality Issue: station_id or station_name do not match between trips data and stations data - found several cases