Big Data Case Study

In Round 1, you proved your proficiency and aptitude in big data analytics theory. In this round, you will apply your skills to a real-world data set and generate business insights in a case study.

Case Background

You are a newly hired data scientist for an Analytics-as-a-Service startup company looking to monetise open-source data available on Taxi Cabs within New York City, as well as other open data sources on NYC itself.

The business target customer segments include (but is not restricted to):

- **Individual Taxi Drivers** - They pay a rental fee per shift for the use of a medallion and take home the remaining profits. They are interested in most profitable times and routes, locations to avoid, etc
- **Taxi fleet companies** - They want to optimise their fleet of taxi's to keep their drivers and customers happy by balancing coverage for customers with the profitability of the routes for the drivers.
- **Taxi and Limousine Commission** - They want to ensure that enough taxi's are available on the roads to meet demand, primarily using the levers of medallion price and taxi ride fares.
- **Local councils** – They have jurisdiction over car parking zones, public transport routes (bus and subway) and special event demands, such as events at Madison Square Garden. They also want to ensure there is enough supply of taxi’s to cover the demand.

You are the first hire and your boss has asked you to do some initial exploration of the data and to find some "interesting insights" that will eventually be used as direct insights to sell to one of the above customers. These may also inform the startup in it's goal to find monetisation opportunities, such as what kinds of insights are available. On this particular day, your boss would like to see some outputs before an important lunch meeting with some investors.

As part of the briefing, your boss has mentioned that:

- At minimum, she needs some data profiling done and some descriptive data analysis written up as a report on the dataset. A good starting point is descriptive statistics on the data set such as average trip times per borough or plots on peak activity times.
- Preferably, she wants evidence that there are interesting insights in the data so the investors are convinced that there is a viable monetisation opportunity here. For example, any patterns in tipping behaviour or times and places locations should be avoided would be interesting to drivers.
- She has expressed interest in more advanced profiling of the data, such as creating customer segments or finding the profile of customers who use taxis to commute so customers can be targeted by marketing.
- Your boss has encouraged you to wear multiple hats and potentially suggest something novel in terms of creating or packaging an insights product. The only restrictions are that the insights utilise open-source data sets and the monetisation is related to the Taxi industry

As further background on NYC Taxi’s, there is a notorious issue with shortages during the evening peak hour. It has been identified that the traditional shift change during 4-5pm is a major cause. This time was chosen as it was seen as the fairest point to evenly split profitability over 12-hour shifts for drivers. For the purposes of this case, you may use the $120-130 range as the cost of rent for a medallion for one shift.

TEXATA WORLD CHAMPIONSHIP

WORLD'S BIG DATA SHOWDOWN

Round 2 – Big Data Case Study Information Pack

The Question

You will be required to submit a report on your analysis in the format below. These sections will appear as individual questions in the HackerRank platform:

11. **Analysis Methodology:** A description of the approach that you took and the methodology you used in your analysis with a rationale on why you thought your approach would be the best use of time.
12. **Analysis Outputs:** A scratchpad of your analysis, covering your observations on the data and the potential significance of your results. All findings and interesting insights should go here.
13. **Business Case:** A summary of relevant insights into a cohesive story on monetisation opportunities using the data along with potentially other open-source data.
14. **Code:** Upload the code you wrote to produce the outcome.

The Data Set

The data set is the 2013 NYC Yellow Taxi data. It consists of two separate tables, namely:

- Trip Data – where and when customers were picked up and dropped off and by which taxi
- Fare Data – fares paid including tolls, surcharges, taxes and tips.

Preliminary Data

A data dictionary and a sample of the first 100 rows for each of the tables can be accessed in the information pack **here**. Download this first to get started.

Amazon Web Services

The dataset has been uploaded to multiple AWS regions namely Oregon, Ireland, Singapore and Sydney.

- S3 – The gzipped csv files are separated into months and placed in public access buckets. The list of links for the files can be found in the **texata_2015_round2_aws_information.txt** file.
- Redshift – The data has been preloaded into Redshift clusters. The connection details, regions and schemas can be found in the **texata_2015_round2_redshift_schemas.txt** file. Please try connecting to you closest region cluster first as they have been sized appropriately.

Google BigQuery

- The data is directly queryable from Google BigQuery **here**.

Final Remarks

As an open data set, you will notice that there is already some analysis that has been done by various sources. Although you are free to research publicly available work, we encourage you to begin your own analysis and getting things written down as soon as possible.