Data Proposal

Avishek Saha and Noman Mohammad

The data for this project was uploaded to kaggle by the user RaviMunde and licensed under CC0: Public Domain, which allows for unrestricted use and distribution of the data. The data set we are analyzing includes simulated ride prices for Uber and Lyft, which were obtained through custom application queries collected using their respective API's. In parallel, additional data was also collected for weather conditions corresponding to the source destination for the cab rides. The data was collected for the interval from 2018-11-25 to 2018-12-18. A custom application was built in Scala to query data at regular intervals of 5 minutes for cab ride estimates and 1 hour for weather data. The data was then saved to DynamoDB. The chosen interval was to query as much data as possible without unnecessary redundancy. Finally, the data collected is stored into two separate CSV files: cab_rides.csv and weather.csv

Statistical Description

The cab_rides.csv consists of 10 variables and 693,071 rows:
- `cab_type`(String): Represents the type of ride sharing service used, such as Uber or Lyft.
- `destination`(String): Drop off location of the ride.
- `distance`(Float): Distance traveled from the pickup location to the drop off location in miles.
- `id`(String): An unique identifier for each ride.
- `name`(String): Type of cab used, such as Lyft XL or UberX.
- `price`(Float): Cost of the ride in US dollars.
- `product_id`(String): An unique identifier for cab-type.
- `source`(String): Pickup location of the ride.
- `surge_multiplier`(String): How much the ride price was multiplied based on current ride demand. Default `surge_multiplier` is 1.
- `time_stamp`(Long): Time of the ride as an epoch time stamp in seconds.

Summary of the cab_rides dataset

```
##     distance       time_stamp            price       surge_multiplier
##  Min.   :0.020   Min.   :1.543e+12   Min.   : 2.50   Min.   :1.000
##  1st Qu.:1.280   1st Qu.:1.543e+12   1st Qu.: 9.00   1st Qu.:1.000
##  Median :2.160   Median :1.544e+12   Median :13.50   Median :1.000
##  Mean   :2.189   Mean   :1.544e+12   Mean   :16.55   Mean   :1.014
##  3rd Qu.:2.920   3rd Qu.:1.545e+12   3rd Qu.:22.50   3rd Qu.:1.000
##  Max.   :7.860   Max.   :1.545e+12   Max.   :97.50   Max.   :3.000
##                                      NA's   :55095
```

The weather.csv consists of 8 variables and 6276 rows:
- `clouds`(Float): Percentage of cloud cover in the sky at the time the weather was recorded.
- `humidity`(Float): Percentage of humidity in the air at the time the weather was recorded.
- `time_stamp`(Float): Time when the weather was recorded, as an epoch time stamp in seconds.
- `location`(String): Location where the weather was recorded.
- `temp`(String): Temperature in Fahrenheit at the time the weather was recorded.
- `pressure`(Float): Air pressure in millibars at the time the weather was recorded.

- `wind`(Float): Wind speed in miles per hour at the time the weather was recorded.
- `rain`(Float): Amount of rain in inches for the last hour at the time the weather was recorded.

Summary of the weather dataset

```
##       temp           clouds          pressure           rain
##  Min.   :19.62   Min.   :0.0000   Min.   : 988.2   Min.   :0.000
##  1st Qu.:36.08   1st Qu.:0.4400   1st Qu.: 997.7   1st Qu.:0.005
##  Median :40.13   Median :0.7800   Median :1007.7   Median :0.015
##  Mean   :39.09   Mean   :0.6778   Mean   :1008.4   Mean   :0.058
##  3rd Qu.:42.83   3rd Qu.:0.9700   3rd Qu.:1018.5   3rd Qu.:0.061
##  Max.   :55.41   Max.   :1.0000   Max.   :1035.1   Max.   :0.781
##                                                    NA's   :5382
##    time_stamp           humidity          wind
##  Min.   :1.543e+09   Min.   :0.450   Min.   : 0.290
##  1st Qu.:1.543e+09   1st Qu.:0.670   1st Qu.: 3.518
##  Median :1.544e+09   Median :0.760   Median : 6.570
##  Mean   :1.544e+09   Mean   :0.764   Mean   : 6.803
##  3rd Qu.:1.545e+09   3rd Qu.:0.890   3rd Qu.: 9.920
##  Max.   :1.545e+09   Max.   :0.990   Max.   :18.180
##
```

It is worth noting that the data obtained through simulation may have certain limitations and scientific methods underlying it. Since the data was simulated, there is a possibility that it only provides a partial representation of the actual demand and supply for rides during the time period under consideration. This is because the simulated data may not accurately capture the usage and demand patterns for ride-sharing services across all areas in Boston. Furthermore, human behavior such as commuting patterns, differences in travel patterns between weekdays and weekends, and responses to weather conditions, which heavily influence the demand for ride-sharing services, may affect the accuracy of hypotheses based on locations. Another aspect to consider is that since the actual source and destination of rides are not known due to the simulated nature of the data, it may limit insight regarding location-based patterns.

Hypothesis

Using a combination of exploratory data analysis and creating predictive models, we want to address a number of scientific questions regarding ride-sharing services like Uber and Lyft in Boston.

1. What is the extent of price variation between Uber and Lyft for rides that originate and terminate at the same location and are initiated at the same time? Are there any patterns or factors that contribute to the observed price variations between these ride-sharing services in such situations?
2. Can we accurately predict the fare of a ride based on various factors such as the ride-hailing service provider, the time of day, cab type, and weather conditions? How do these factors interact with each other to impact the ride fare?
3. How can we predict the surge pricing multiplier for Uber and Lyft rides in Boston, taking into account various variables such as time of day, day of the week, and weather conditions? What are the key factors driving the variation in surge pricing among these ride-sharing services in Boston?
4. To what extent do the source and destination locations impact the cab prices in Boston, after controlling for factors such as distance and weather conditions? Can we identify any specific locations or geographic patterns that have a significant impact on cab prices in the city? How do these factors vary between Uber and Lyft?