**ALY6110 DATA MANAGEMENT & BIG DATA**

**CRN -82268**

**“TAXI FARE ANALYSIS”**

**SANJAY LOKULA (lokula.s@northeastern.edu)  
NIKHIL REDDY DEPA (depa.n@husky.neu.edu)  
PUI LING ADA SUNG(sung.p@husky.neu.edu)  
CHARAN KUMAR PATAM(patam.c@husky.neu.edu)  
RISHABH SHARMA (sharma.ris@northeastern.edu)**

**CPS-2020 SPRING QUARTER**

A picture containing drawing

Description automatically generated

**ASSIGNMENT DATE- 22 JUNE 2020**

**INSTRUCTOR- MR.VALERIY SHEVCHENKO**

**SUMMARY**

Bigdata is a collection of structured, unstructured data collected by organizations to mine information based on requirements using machine learning and big data analysis tools, this accumulated information is used to improve customer service, increase market share and hold a potential competitive advantage. In the present digital world, every second generates a ton of data with different velocity and variety i.e. The entire world is connected to big data systems through GPS, support services and social networking. Around 1.7 zettabytes of data has been generated in 2012 and there’s no end for it, as organizations like Facebook and google work on collecting data from customers and improve services based on the insights attained from the data, about 90% of the worlds data is being generated in the past 3 years with about 3.2 quintillion of new data is added each day but the problem with this collected data is it being unstructured. Handling this data requires a different computation which laid the foundation for bigdata analysis that can handle the complexity and challenges of this data which shifted the business focus to a data driven model which rely on big data tools to make any decisions. Tools like Hive on spark provides a means to deploy this bigdata and process in batches for Hive on Apache Spark increasing the performance compared to traditional MapReduce methodology.

Transportation is a major obstacle in present world of increasing population making it near to impossible to provide transportation for everyone, this lead to raise of new market by providing cab/Taxi services for ones who need with a fare-price based on the service provided as this new market was segregated many new organizations have emerged to bring fair pricing and easy access to every individual. Organizations like Uber turned out to be the prominent in the new market by analyzing the data collected from various sources and performing bigdata analytics and building a models based on region and time to charge for the trips which helps the organization to grow better and improve the service based on constant bigdata analysis of the data collected from rides.

For this analysis we considered the New York city to estimate the fare pricing of yellow taxis over years and the parameters that influences the fares as the number of riders have increased from 3 million to over 8 million in a year which made organizations move to datacentric in understanding the pulse and requirements of customers.

**ABOUT THE DATASET**

New York city data is a big query dataset based on 2016 NYC yellow cab trip fare published by NYC Taxi and Limousine Commission(TLC).The dataset comprises of 55 million data entries with 9 features which brings the requirement of Bigdata technologies to analyze the data. The features of the dataset comprises of key a number to identify the row, pick up datetime, pick up latitude, pick up longitude, drop off longitude, drop off latitude along with fare price and passenger count. The link to the dataset is <https://www.kaggle.com/c/new-york-city-taxi-fare-prediction/data> and A sample dataset is provided below for understanding.

A close up of a piece of paper

Description automatically generated

To analyze this dataset and answer few business questions we considered Databricks a unified data analytics cloud platform to handle bigdata.

**OBJECTIVES**

The main objectives of the taxi fare analysis is analyzing the data to solve the business problems and provide recommendations based on insights.

* **Following are the Business questions of New York Taxi fare Analysis:**

1. What is the trend in no of rides and total fare over years?
2. What is the average trip fare for every month over years?
3. What is the busiest day of the week in terms of no of rides?
4. What is the average fare per mile for day of the week?
5. Is there any difference in fare prices during weekdays and weekends?
6. What is the average fare amount over years during peak times?

**ANALYSIS**

* + **Initializing Cluster and Loading Data File:**

The first stage of the analysis is creating a community data bricks account by following the link provided <https://databricks.com/try-databricks> and then navigating to the cluster tab and creating a cluster as shown below

A screenshot of a cell phone

Description automatically generated

Once the cluster is fired up navigate to the data tab and upload the dataset our considered data file is greater than 2GB we have zipped the file and uploaded it as displayed below.

A screenshot of a cell phone

Description automatically generated

After creating a cluster and uploading the dataset, we need to unzip the datafile to do that we used the shell scripting with %sh

A screenshot of a social media post

Description automatically generated

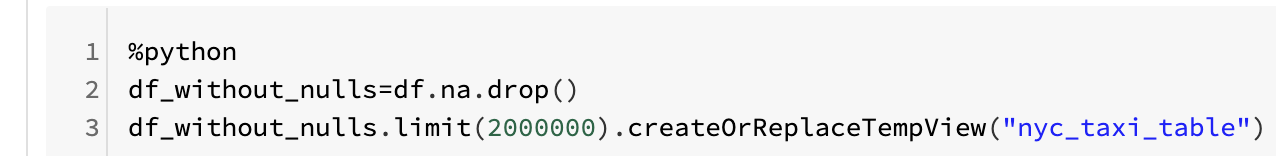
As the datafile is unzipped now to read the file we selected python as the language with %python as the key and performed the pre cleaning summary analysis.

A screenshot of a social media post

Description automatically generated

* + **Cleaning the Dataset:**

From the pre-dataset cleaning summary analysis, we observed that there are many null values which needs to be dealt with, we used the drop.na() method of python to drop all the unnecessary null values and observed the count.



There are many parameters to be worked on while cleaning the dataset i.e. Passengers count needs to be greater than 0 and less than 3 standard deviation, with pick up and drop of latitudes, longitudes under 3 standard deviations to do that we considered the python with spark.sql functions and observed the count post data cleaning as the dataset has more than 50 Million entries we have sampled the dataset on random and considered 2 Million entries for quick processing.

A screenshot of a social media post

Description automatically generated

* + **Processing the Dataset:**

Now that the dataset is cleaned, we can move ahead with performing EDA/preliminary processing of the dataset as the business question shown above revolves around day, time, week and year we need to segregate these data from the data time column available or the key column available in the dataset. To do that we considered python as the language and used the pyspark sql functions and imported day of month, col, hour, month, year, minute, second, day of week and expanded the dataset apart from that we also installed necessary library geopy.geocoders and imported Nominatim as shown below.

A screenshot of a social media post

Description automatically generated

* + **Dataset Analysis:**

Now that the dataset is cleaned and pre-processed, we can move ahead and perform the analysis required to answer the business question mentioned.

* **What is the trend in no of rides and total fare over years?**

To answer this, we consider the sql workbook and considered scripted a query to calculate the total fare amount generated over years considering every month.

A screenshot of a cell phone

Description automatically generated

A close up of a map

Description automatically generated

From the analysis it is observed that the total fare collected is highest in the year 2013

Now to find the no of rides over year we considered the count of rides vs year and plotted it using sql queries.

A screenshot of a cell phone

Description automatically generated

A close up of a map

Description automatically generated

From the above attachment we observe that the no of rides are maximum in the year 2012

* **What is the average trip fare for every month over years?**

Now that we know the trend of fare over years we can move further and find the month in the year that contributes the most to revenue through ride fares. To do that we used the sql query to segregate the data into average fare amount and grouped by name of month, month, year and ordered by month and year as shown below.

A screenshot of a cell phone

Description automatically generated

A close up of a map

Description automatically generated

From the above line plot, we observe that the average fare remained stationary over the months for a year, but the total fare amount increased from year 2009 to 2014.

* **What is the busiest day of the week in terms of no of rides?**

To evaluate the busiest day of the week based on the no of total successful rides we considered the same sql workbook and scripted a query by selecting day name of the week ,month, year and counting the no of trips and day of the day of the week and later grouped them by day of the week, day name of the week, month and year. The sql query is provided below as an attachment for further understanding.

A screenshot of a cell phone

Description automatically generated

A close up of a map

Description automatically generated

From the analysis we observed that the busiest day is Friday ,while the least busy day being Monday.

* **What is the average fare per mile for day of the week?**

To estimate the average fare per mile we considered the pickup latitude, longitude along with drop off latitude and longitude and using geopy library estimated the trip distance. Now that we have trip distance, we divided the fare amount with distance travelled to get the average fare per mile then we plotted it considering the day of the week as a parameter.

A screenshot of a cell phone

Description automatically generated

A close up of a sign

Description automatically generated

From the plot provided we can observe that the average fare per mile for day of the week is high on Sunday with fare about 35.

* **Is there any difference in fare prices during weekdays and weekends?**

To evaluate any differences in fare prices during weekdays and weekends we split the days of the week from Monday to Friday as weekdays with Saturday and Sunday as weekends. Then we move ahead and plot a line plot over hours for both categories using python as the language.

A screenshot of a cell phone

Description automatically generated

A close up of a map

Description automatically generated

From the analysis we observe that the weekday fare is higher from 5:00 am to 7:00 am with a max fare of “$ 13” while the weekend fare is higher than weekday from 8:00 am to 6:00 pm with the highest being “$12” . It could be interpolated from the graph that fares on weekend’s are mostly less than Weekday’s for each hour, but during the early hours and the late hours, the prices are skyrocketing, one reason could be lack of supply / availability of cabs

* **What is the average fare amount over years during peak times?**

To evaluate the fare amount during peak times we considered the peak hours as 4:00 pm to 8:00 pm during weekdays which is because New York government provides a privilege to impose surge charges on trips during peak hours, we aimed to analyze how this average fare amount changed over years during these peak times. To do we considered plot for average fare amount during peak times over years as provided below.

A screenshot of a cell phone

Description automatically generated

A picture containing drawing

Description automatically generated

From the plot displayed above we can observe that the average fare during peak time increased over years from 2009-2015 with highest during 2014 and 2015 at $12. Which depicts that the increase demand of the taxi and changing economic conditions it is expected that the average fare amount could be show a positive slope over years. The same could be proved using the plot.

**CHALLENGES AND RESOLUTION**

* The major challenge faced while working on this dataset would be importing the data set as the original file size is greater than 2GB and data bricks cap the file size for uploading in the community edition to be at 2GB we had to upload the zip file and later then write a shell script to unzip the file.
* The date and time were merged in the same variable, to extract each element, we had to clean the variable accordingly in Python to get the desired values.
* Many corrections were necessary to work on the dataset i.e. filtering out the pickup latitudes and longitudes to New York region, and making sure that passenger count is no greater than the permittable number, to do this we considered the limits to be with in 3 standard deviations and cleaned the data using the spark sql queries.
* The major challenge faced with the dataset is importing the necessary libraries for analysis as only few python libraries are supported by data bricks, to fix this we had to find a library which performs the same analysis and supported by data bricks and had to write various functions to get the similar job done.
* The visualizations attained from the analysis offered no customization options which made us to go ahead with different kind visualization rather than customizing the existing one.
* The final challenge faced while performing this analysis is with the run time as the original dataset had around 55Million entries and took more than 8 hrs. to run and there were scenarios where few steps were skipped automatically, we had to consider a sample data with 1 million entries and perform the analysis.

**RECOMMENDATIONS**

* From the analysis performed to estimate the fare price of taxis in New York city we can definitely recommend the organization Yellow cabs to increase the base fares.
* It is recommended for yellow cabs to run many cabs during rush hours as they contribute high to the revenue generated in a week
* The weekday and weekend fare during the hours 2:00 am to 4:00 am is very high this could be due to unavailability of cabs to increase the range profits we can setup more cab zone especially around airport zones as there’s no public transportation available in New York after 12:00 am.
* It is observed that the average fare per mile is low on the second day of the week which could be due to few/no riders to improve the situation, organization could come up with an offer of travel-Tuesday plan where every rider gets points to avail a free ride.
* We observe that average fare pricing remained constant for the second consecutive year which could be difficult as economic conditions changes every year the organization should reiterate the surge charges imposed to avoid loss or at least attain breakeven.

**CONCLUSION**

* To conclude this analysis provided a clear understanding on the parameters considered by Yellow cabs organization to determine the ride fare.
* Although data bricks provides a convenient way to analyze the data set it has its own disadvantages as the use of JVM memory there is a cost in efficiency when working with python.
* Data bricks provide an efficient method to switch between languages as tasks like segregating data and cleaning can be perform with ease writing queries scripts in SQL while the analysis can be done on python all we need to do is use %sql or %python to switch between languages.
* The sole goal of the analysis is to find out if the organization is running in loss and what measure to be considered to improve the performance from the insights attained and observed that the organization is having a break-even for the year 2014 and 2015.
* There is a significant raise in the average fare pricing during early hours which could be due to no availability of taxis which can be fixed by dividing the city into zones and making taxis available in these zones (Near Airports and clubs).
* This analysis can be used to identify and come up with strategies for organizations to expand outside New York by setting a mean fare price and surge price based of the city.
* Finally, to conclude if the dataset is more than 10 million entries and if there’s lot to analyze its highly recommended to consider AWS/Azure with data bricks rather than go with community edition as in community edition only one cluster is made available at a single time which takes hours to compute.

**REFERENCES**

* New York City Taxi Fare Prediction. (n.d.). Retrieved from <https://www.kaggle.com/c/new-york-city-taxi-fare-prediction/data> - **Link to Dataset**
* About the Author Regulatory Reform Team. (n.d.). Case Study: New York City Taxis. Retrieved from <https://datasmart.ash.harvard.edu/news/article/case-study-new-york-city-taxis-596>
* Rouse, M. (2019, October 25). What is Big Data and Why is it Important? Retrieved from <https://searchdatamanagement.techtarget.com/definition/big-data>
* Schneider, T. (2015, November 17). Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance. Retrieved from <https://toddwschneider.com/posts/analyzing-1-1-billion-nyc-taxi-and-uber-trips-with-a-vengeance/>
* Combessie, G. N. (n.d.). Taxi Fare Prediction Machine Learning in Real Time. Retrieved from <https://blog.dataiku.com/predicting-taxi-fares-in-new-york-using-machine-learning-in-real-time>
* New York City Taxi Trip Duration. (n.d.). Retrieved from <https://www.kaggle.com/c/nyc-taxi-trip-duration/data>
* Jing Guo Analysis and comparison of Uber, Taxi and Uber request via Transit Retrieved from <http://www.jics.utk.edu/files/images/recsem-reu/2018/transit/Report-B.pdf>
* (n.d.)(2020). Retrieved 22 June 2020, from <https://data.cityofnewyork.us/Transportation/NYC-Taxi-Zones/d3c5-ddgc>