* 94693 Big Data Engineering
* [- Spring 2023](https://canvas.uts.edu.au/courses/27806)
* AT2

Done by:

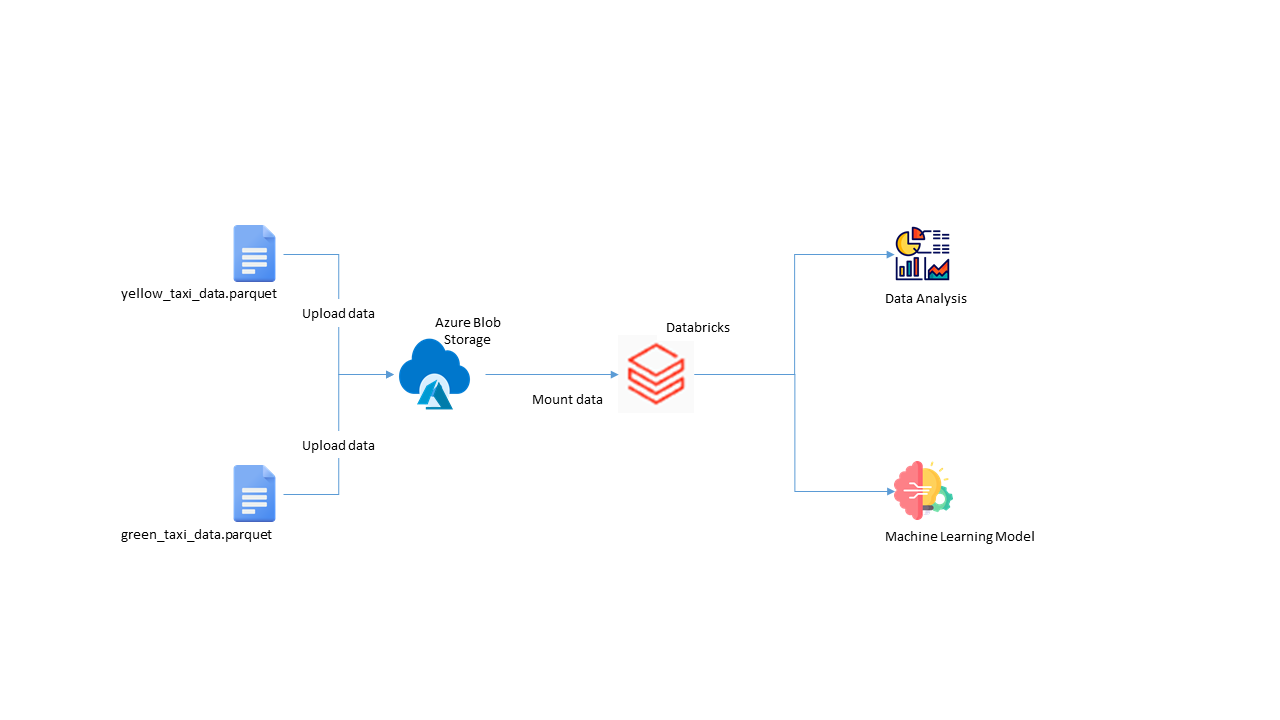
Name: Tahmidul Islam

ID: 24587139

1. **Introduction**

The New York Taxi and Limousine Commission (TLC) was established in 1971 with the primary role of licensing and overseeing various transportation services in New York City, including taxis, for-hire vehicles, commuter vans, and paratransit vehicles. Starting from 2009, the TLC has been making trip records available to the public, and these records have grown to include millions of rows of trip data. The objective of this project is to work with the dataset encompassing both yellow and green taxicabs, covering the period from January 2019 to April 2022. The project aims to conduct a thorough analysis of this dataset and ultimately develop a machine-learning model that can predict the total fare amount for each trip.

1. **Overall Architecture**

****

*Fig: Overall Architecture of the Project*

We first download Data from the Official Website of New York City TLC which is then uploaded to Azure Blob Storage. Following this we mount them to Databricks.

The storage blob container is then mounted to the DBFS (Database File Systems) which allow us to access all the files stored in the Azure Blob Storage directly from Databricks.

The data can then be loaded, cleaned and manipulated to prepare for analysis to find answers to Business Questions. Following this, machine-learning models (Linear Regression & DT) are implemented on the dataset to predict the Total Amount paid in a trip.

1. **Data Ingestion and Preparation**

Ideally, the project aimed to ingest data by making use of Azure Data Cloud Storage before mounting to Databricks File Storage. However, the dataset was extremely large and upload errors were coming up repeatedly after waiting for a long time. Thus, the parquet files were directly mounted to the DBFS from google drive, provided by the Course Lecturer.

**Data Ingestion:**

Data was mounted using the ‘*Download Taxi Dataset from Google Drive.ipynb’* notebook provided already. It was run to upload all the .parquet files for green taxi cabs and yellow taxi cabs into respective folders of ‘green’ and ‘yellow’ inside the FileStore of DBFS.

(NOTE: Running any form of operations inside Databricks requires the creation of a computing cluster. For the community edition used in the project, the cluster would terminate within 1 hour of inactivity.)

The row count for green taxi was **66200401** while for yellow it was **663055251.**

**Data Preparation:**

The datasets for yellow and green taxis underwent a cleaning process to eliminate unrealistic trips. Here's how the cleaning was performed:

Trips that concluded before they started were removed. This involved excluding rows where the drop-off time was earlier than the pick-up time.

Trips with negative speeds were filtered out. Rows indicating a trip with a speed less than zero were eliminated.

Trips exceeding the New York state speed limit of 55 mph were excluded. Rows showing speeds greater than 55 mph were removed.

Trips of very short or excessively long durations were removed. Rows with trip durations less than a minute or longer than 2.5 hours were eliminated. These time limits were selected because only 1% of trips were shorter than a minute, and another 1% of trips exceeded 2.5 hours.

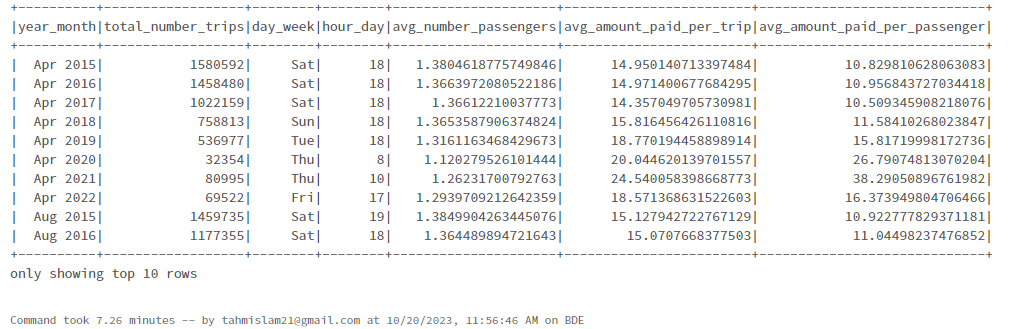
Trips of very short or very long distances were also removed. Rows with trip distances less than 0.45 miles or greater than 50 miles were excluded.

Trips that fell outside the timeframe of January 2015 to December 2022 were eliminated. Rows with pick-up dates earlier than January 1st, 2015, or later than December 31st 2022, were removed to ensure that the data only included trips within the specified period.

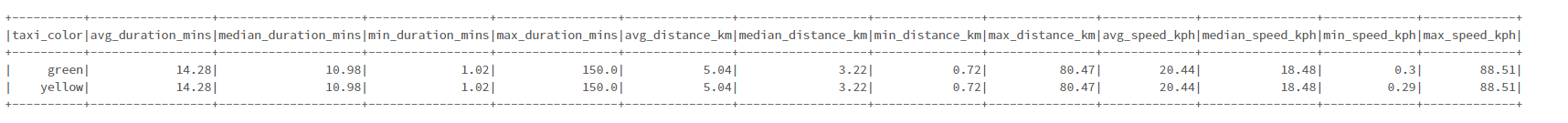
**Merging Dataset:**

Following the clean-up of yellow and green datasets, they were merged on the columns that were found common. This way, the discrepancies that were present in their data structure were settled. The combined data was saved as a parquet file and loaded again for performing some queries for business questions.

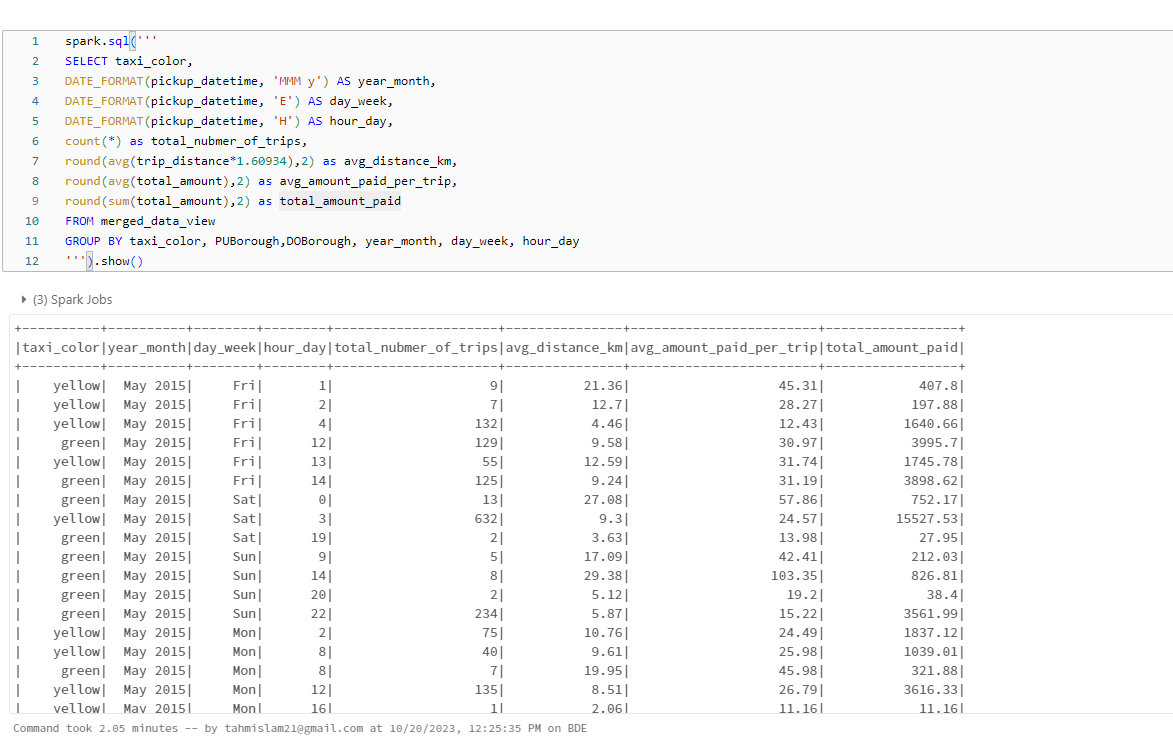
1. **Business Questions**
   1. All these aggregate columns in this business question needed to be found out separately and then were finally combined together in a single data frame.

****

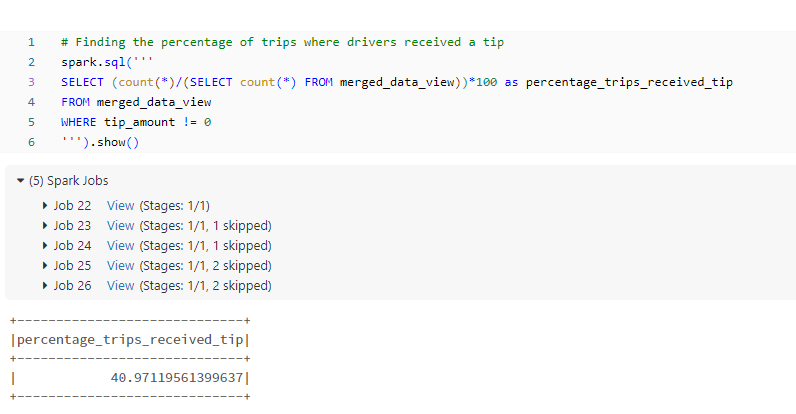
* 1. Each of these columns were found using aggregate functions like avg(), max() and min()



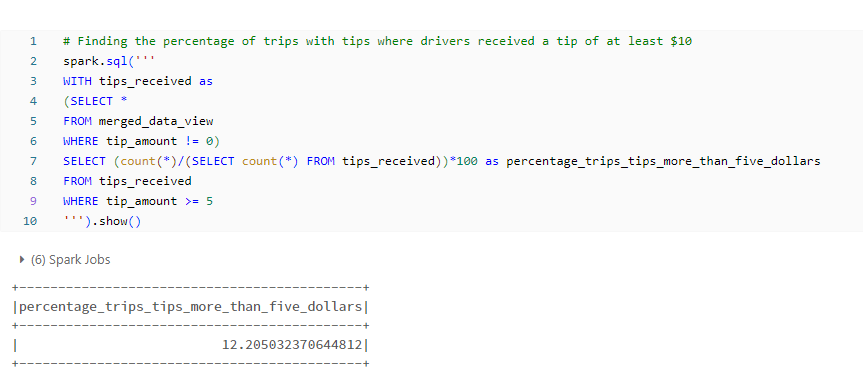
* 1. The columns were found by aggregate functions as well and they are grouped by taxi color, PU and DO location Boroughs, month day of week and hours



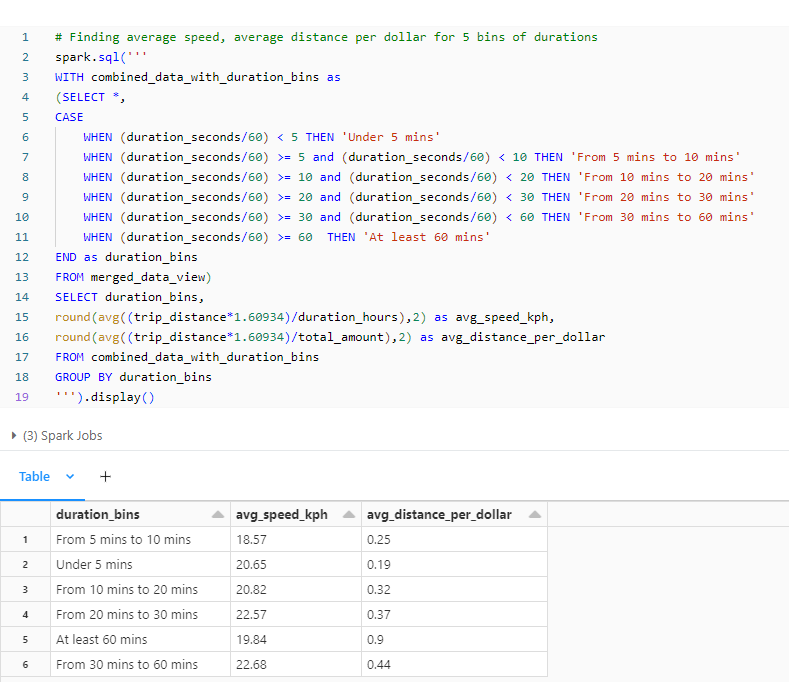
* 1. The table below shows the percentage of trips that received a tip.



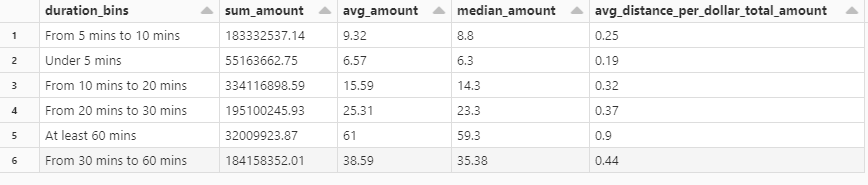
* 1. The table below shows the percentage of trips that received more than 5 dollars



* 1. The table below shows bins and their corresponding aggregate values like average speed and average distance



* 1. As seen from the table below drivers can earn the highest average amount if they drive atleast 60 minutes. However, they have to travel 0.9 kms for every dollar, which is also the highest. This can increase operation cost and is thus not optimal.  
       
     Interestingly, if we look at the 6th row, we can find the second highest average amount in payment, and they only need to travel 0.44 km for every dollar. Thus driving between 30 to 60 minutes would be most recommended for maximum profit.



1. **Machine Learning**

In this part of the project, we applied two Machine Learning models – Linear Regression and Decision Tree to the combined dataset. The aim was to predict the total amount paid for each trip.

**Data Preparation steps:**

Data set ranging from October 2023 to December 2023 was separated as the test data set while the remaining were reserved for training.

The column values "trip\_distance", "duration\_hours", and "taxi\_color" were used for the modelling part as taking in more features increased the computation of the dataBricks clusters, thereby terminating it after 1 hour. The rows containing missing values were removed.

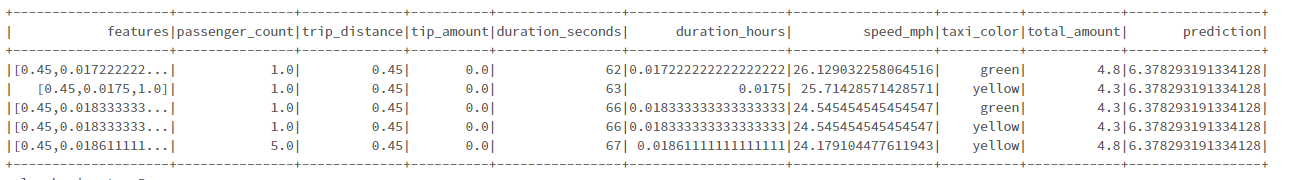
The taxi color column was one-hot encoded because of having nominal values (green and yellow). Following this, the other features were vectorized for the efficiency of Machine Learning Algorithms.

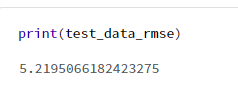
**Results:**

|  |  |  |
| --- | --- | --- |
| **Model** | **RMSE (train)** | **RMSE(test)** |
| Linear Regression | 4.13 | 4.75 |
| Decision Tree | 4.66 | 5.22 |

Both Linear Regression and Decision Tree were overfitting because of having low errors on the Training set while having high errors on the Test Set.

The final predictions on the test set was made by Decision Tree Algorithm which gave an RMSE of 5.22 (again).





1. **Conclusion**

Performance of the models were mediocre, but they can be increased by several more experiments after tuning hyperparameters for each models. This can reduce overfitting of the models and give us enhanced accuracy of the total amount per trip.