|  |
| --- |
| Assignment 1  Big Data Engineering |
| StudentID: 12964298  Group: Group 14 |



Table of Contents

[Overview 3](#_Toc69142151)

[Data Processing 4](#_Toc69142152)

[Importing, Merging & Cleaning Data 4](#_Toc69142153)

[Issues/Problems 5](#_Toc69142154)

[Business Questions 6](#_Toc69142155)

[Machine Learning Model 7](#_Toc69142156)

# Overview

The goal of this project was to analyse New York’s taxi service between 2017 and 2018. The dataset is very large, consisting of over 200 million records. To complete this project, pyspark was used for the data processing and machine learning components of the project.

The project is broken down into 4 major steps:

1. Importing, merging and cleaning the data
2. Performing feature engineering and saving the data
3. Answering business questions about the dataset
4. Training a machine learning model to predict the fare amount.

# Data Processing

## Importing, Merging & Cleaning Data

* Downloaded and saved csvs
* Aligned column names of pick up and drop off
* Adjusted datatypes

## Feature Engineering

## Business Questions – SQL Queries

## Machine Learning

# Issues/Problems

## Saving Parquet Files

### Problem

Various problems arose when attempting to save a parquet file. Two major problems, which also occurred elsewhere in the codebase, are described below. A similar error to the one described in this stake overflow thread[[1]](#footnote-1)[[2]](#footnote-2) was also experienced.

### Solution

Adjusting the memory for spark seemed to help. This error was more common when the memory allocated was larger. Note that this did not resolve the issue completely, however restarting the kernel appeared to resolve the issue sporadically.

## ‘Hanging’ Functions/Excessive Runtimes

### Problem

This problem occurs when the

### Solution

This was encountered quite commonly throughout the SQL queries. To avoid this, I reset the view by running df.createOrReplaceTempView("combined\_data"). This line of code reset the temporary view of the dataframe[[3]](#footnote-3).

## JVM ‘heap memory’ error

### Problem

### Solution

# Business Questions

### For each Year & Month

1. What was the total number of trips?
2. Which day of the week had the most trips?
3. Which hour of the day had the most trips?
4. What was the average number of passengers?
5. What was the average amount paid per trip (total\_amount)?
6. What was the average amount paid per passenger (total\_amount)?

### For each taxi colour

1. What was the average, median, minimum and maximum trip duration in seconds?
2. What was the average, median, minimum and maximum trip distance in km?
3. What was the average, median, minimum and maximum speed in km per hour?

### What was the percentage of trips where the driver received tips?

### For trips where the driver received tips, what was the percentage where the driver received tips of at least $10.

### Classify each trip into bins of durations & calculate

1. Average Speed (km/h)
2. Average distance per dollar (km per $USD)

### Which duration bin will you advise a taxi driver to target to maximise his income?

# Machine Learning Model

# Appendix

## Saving Parquet Error

## Text Description automatically generated

1. (Various, 2018), https://stackoverflow.com/questions/51187904/py4jjavaerror-an-error-occurred-while-calling-o26-parquet-reading-parquet-fil [↑](#footnote-ref-1)
2. Note that the solution provided in the stake overflow thread did not appear to be causing the error on my machine [↑](#footnote-ref-2)
3. Developer note: I don’t know why this worked. In tutorials it appears this step was not necessary, however it seemed to resolve the issue. [↑](#footnote-ref-3)