MIT 805 – ASSIGNMENT 1

U18239201

INTRODUCTION

The New York City yellow taxi business has been under permanent stress since the arrival of Uber taxis. Ameliorate its business design has been one of its challenges lately. Embracing new technology, running promotions, are ones of the many examples an industry can use to grow its revenue. However, it might be interesting to look at the data already available within the company to figure out if one can detect pattern that may help realize profits.

In this paper, we will be discussing the initial exploratory analysis of datasets relating to New York City yellow taxis.

DATA DIMENSIONS

The definition of big data has been refined over the time. Recent literature describe big data using the 3V's which are volume, velocity and variety (Kumar, Gupta, Charu and Jangir, 2014) . Due to the fact that more people are getting into the area of big data, six other V's have been added to the three above mentioned. Sami and Sael, (2016) have broken down the 9V's into five different categories. See image below.

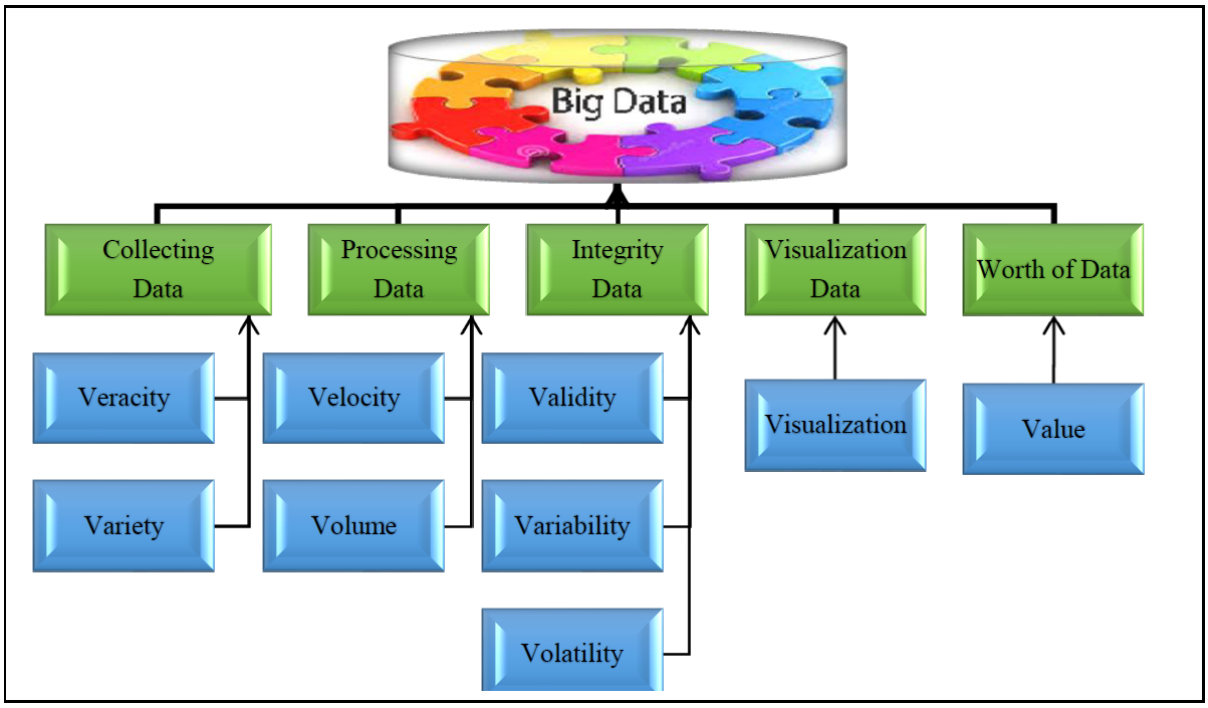

Fig. 1. (Sami and Sael, 2016)

- Variety and Veracity fall under collection. Veracity pertains to the trust we have in the data. Because the data come from different sources we must ensure the quality of these sources. Variety relates to the structure of the data. Data is usually structured, semi structured or unstructured.
- Processing consists of velocity which is the speed at which the data is processed and volume which represents how big our data is.

The three other categories will be discussed during the next assignment.

DISCUSSION OF OUR DATASET

Kuiler (2014) defines big data as a tool to depict large amounts of data generated by machines. Big data tend to have three main properties: velocity, volume and variety. Our dataset (https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page) is considered as big data because it fulfills the three V's.

- Volume: Our data set is 2GB which is not vast but enough to be considered big data
- Velocity: Our data lacks of velocity because the data is static. It is stored on the New York City Taxi website
- Variety: Our data comes from different sources. The website mentions that the data is collected from authorized technology providers
- Since our data set has not been collected directly by the NYC Taxi website, the quality/veracity of the data should be tested to avoid possible issues during our analysis

DATA UNDERSTANDING

Three sets of data (Jan2019 - March 2019) were downloaded in csv format. The files were initially opened in Excel and reviewed to understand the format of the data.

The three files have been imported and combined into one in Python through Jupyter Notebook. The following are the first information we got about our dataset.

- Shape: our data set consists of 22519712 rows and 18 columns
- Data types: We have 15 columns imported as numeric and 3 categorical columns
- Missing Values: There is only one column with missing values. 22% of data are missing in that column (congestion surcharge). The decision has been to keep it as it for now. We will decide later in the project what transformation to do on that column

DATA CLEANING

The years greater than 2020 (2038, 2041, 2088) have been dropped from our dataset. Same for the months after March since the data downloaded on the website ranged from January to March.

VISUALISATION OF SOME VARIABLES

A view of the distribution of some variables has been performed:

- Passenger Count: Figure 2 illustrates the distribution of passengers amongst the trips. Trips with one passenger are the most frequent ones.

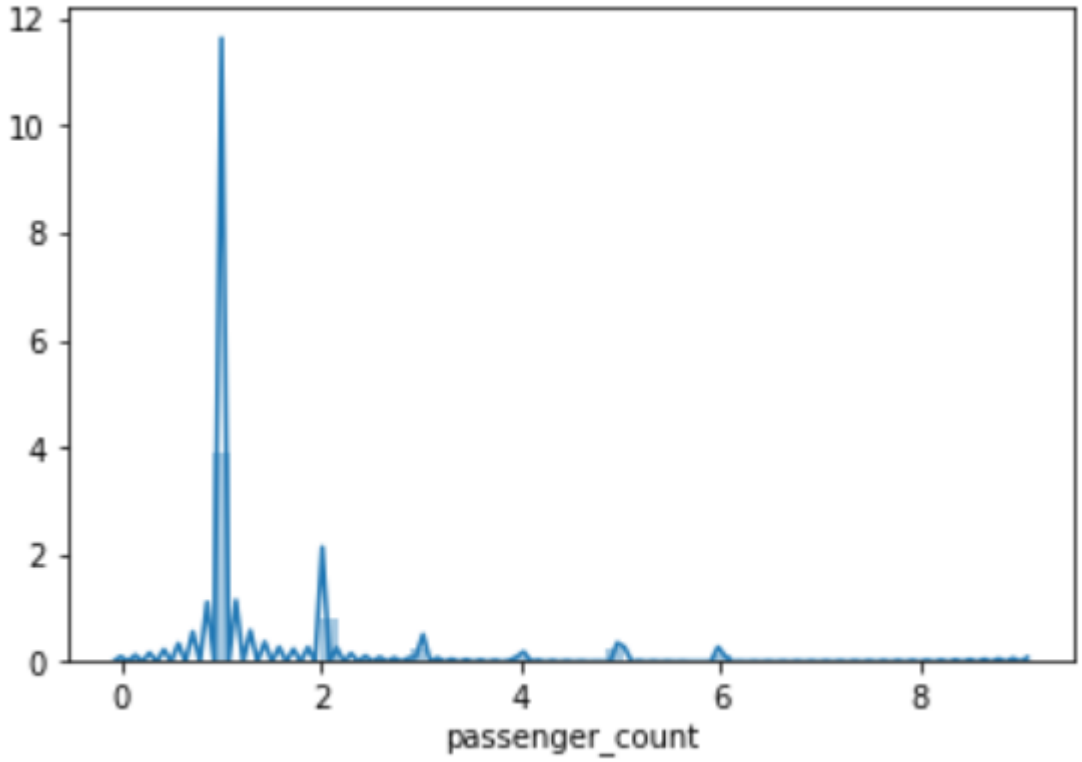

Fig. 2. Distribution of Passengers

- Payment Type: Credit card (code 1) is the preferred method of payment. A significant amount of none charges were given out to passengers. This is part of the New York City royalty program.

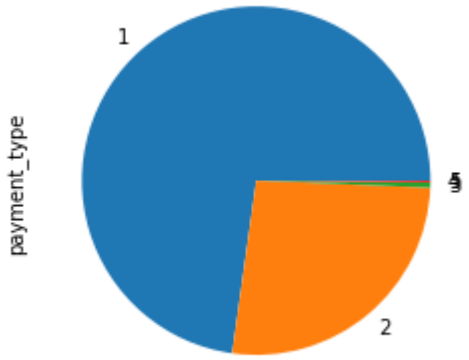

Fig. 3. Payment Type

- Trips per Month: Figures 4 illustrates a comparison between the three months for the numbers of trips recorded during these months. No big difference between the three. A deeper analysis to see what time of the month/day we have more passengers will be made during the visualisation part of the assignment.

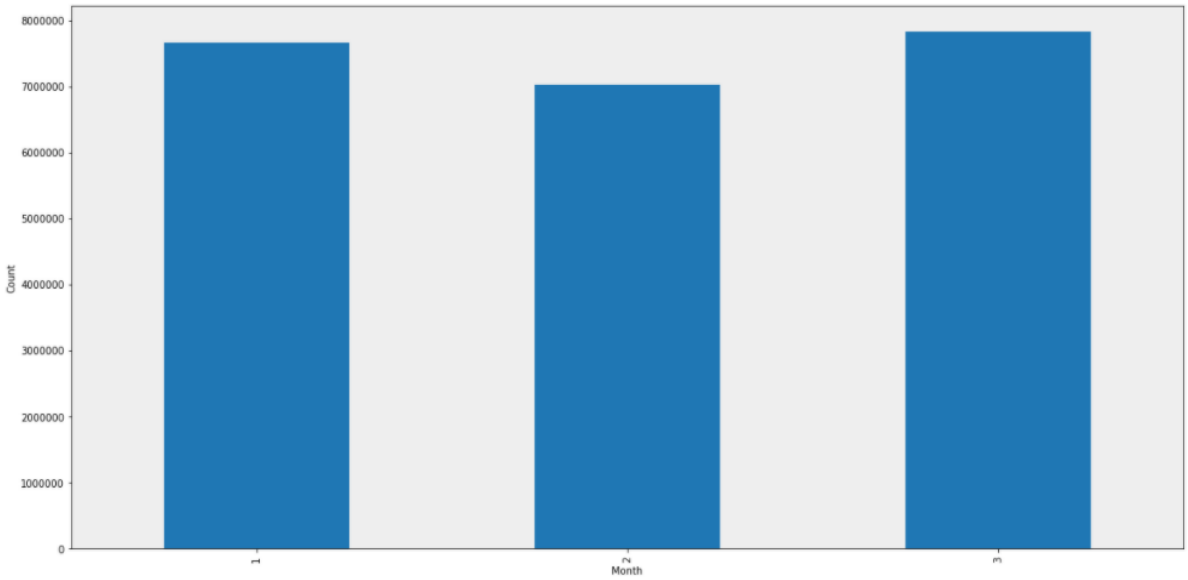

Fig. 4. Count of Trips Per Month

TECHNOLOGY AND FRAMEWORK

Since we are dealing with static data, we are going to use batch processing to process our dataset. Our dataset has been stored over many years, hence batch processing is more suitable. We'll try to get insights from historical data to improve the New York City Taxi business. We don't need real time results. Batch processing requires high CPU and RAM, a new laptop may be required for this project. Hadoop Map reduce is the chosen framework because there are lot of materials available out there and the literature says it is the best technology for batch processing.

REFERENCES

Kuiler, E., 2014. From Big Data to Knowledge: An Ontological Approach to Big Data Analytics. Review of Policy Research, 31(4), pp.311-318.

Kumar, R., Gupta, N., Charu, S. and Jangir, S., 2014. Manage Big Data through NewSQL.

Sami, S. and Sael, N., 2016. Extract Five Categories CPIVW from the 9V's Characteristics of the Big Data. International Journal of Advanced Computer Science and Applications, 7(3).