Logo

Description automatically generated

Week 3 Assignment 3 – New York City Taxi Fare Prediction

Craig Perkins – perkins.cr@northeastern.edu

Northeastern University

December 2021 – ALY6610.70531

**Week 3 Assignment 3 – New York City Taxi Fare Prediction**

For this assignment, I chose to work with a New York City Taxi Fare dataset available on Kaggle (<https://www.kaggle.com/c/new-york-city-taxi-fare-prediction/overview>). There are many reasons why I am choosing this dataset, but one of the primary reasons is that I had a tough time finding datasets that met the main criteria of over 1M records and 5 dimensions (See the Dataset Selection section for more details on the discovery process). This is a very large dataset with over 55 million records and 7 dimensions including:

1. Fare
2. Pickup Datetime
3. Pickup Longitude
4. Pickup Latitude
5. Dropoff Longitude
6. Dropoff Latitude
7. Passenger Count

Other than strictly meeting the criteria for the final project, I was also drawn to this dataset because I am a New York City resident and mobility is an important issue to me. A big interest of mine is in city planning and urban development, so I am often thinking about how to make cities more livable and how to sustain a larger population as cities tend to be concentrated areas of innovation. It’s important to me that people of all abilities and income levels can move about in the city they live in in order to access resources and move up the socioeconomic ladder. To that end, I’ve noticed a startling increase in the costs of rideshare and taxi services in the last few months and as a Data Scientist, I’ve wanted to look into the raw data to try to illustrate the story of why prices were increasing drastically (<https://www.nytimes.com/article/uber-lyft-surge.html>). In this project, I would like to take a deep dive into the raw data of New York City Cab Fares and see what insights can be derived using Big Data techniques and systems we learn about in class like the applications in the Hadoop Ecosystem. For the initial exploratory data analysis, a random subset will be taken to be analyzed up front in a Jupyter notebook which will help in gathering some insights initially before performing the same analysis on the full dataset in Hadoop.

**Introduction**

New York City is the largest city in America and amongst the largest megacities of the world. New York City is world renowned for its iconic skyline, famous landmarks and its subway system which also happens to be the oldest in the world. With 428 stations, and more under development, New York City has the largest metro system in the world in terms of number of stations and near the top in ridership. In addition to the subway system, New York City also has one of the largest taxi and ridesharing fleets in the world topping out near 98,000 in 2019 before the pandemic (<https://toddwschneider.com/dashboards/nyc-taxi-ridehailing-uber-lyft-data/>). Orchestrating this fleet poses a big challenge and rideshare apps like Uber and Lyft face challenging optimization problems like ensuring their drivers have an adequate amount of fares to remain happy driving for the service and also ensuring that passengers are picked up in a timely manner and delivered to their destination safely and in a reasonable amount of time. Drivers can also now take advantage of traffic updates to avoid routes that may have rode closures or have unexpected congestion like at the scene of an accident.

In an effort to reduce the amount of traffic below 60th street, New York City has implemented congestion pricing for rideshare vehicles during peak hours. Between the 2010 and 2020 census, New York City has gained 629,000 residents further showing the urgency of mobility solutions in the nation’s largest city. With the population of NYC projected to increase over the next decade, it’s imperative that city leaders have the most information at hand in order to optimize how city resources are used to further mobility.

# Dataset Selection Process

I had a hard time finding open datasets with over 1 million rows and 5 dimensions. Since I work at a cancer hospital, the first thing that came to mind was to see if NAACCR (North American Association of Central Cancer Registries) had an open dataset to use. They do have an open dataset, but it required contacting NAACCR which I did and received access, but then I got blocked because it also require software that is only available for Windows (<https://www.naaccr.org/cina-public-use-data-set/>). After failing to find the dataset I wanted to work with there, I then went back to a resource I’ve used in prior classes in the EAI program here: <https://github.com/awesomedata/awesome-public-datasets>. I tried a few of these datasets like the one on Crop Yields (<https://doi.pangaea.de/10.1594/PANGAEA.909132>) and the US Department of Educations “College Scorecard” dataset (<https://collegescorecard.ed.gov/data/>), but either found that the dataset was too small or not in a familiar format. The Crop Yields dataset was in NetCDF (.nc4) and I was not able to figure out a way to read it on my mac. After trying those datasets, I then moved to AWS’ public registry of open datasets. I again found a dataset I was interested in called “First Street Foundation (FSF) Flood Risk Summary Statistics” (<https://registry.opendata.aws/fsf-flood-risk/>), but again found that it was too small and also contained “shapefiles” which I am unfamiliar with. After that I turned to Kaggle and went through a number of submissions before finally stumbling upon the New York City Taxi fare dataset. After taking a look at the NYC dataset, I was not only happy to see that it made the requirements but also something that is directly applicable to my own life being a resident of New York City and also an interesting technical challenge. I think we will be able to really show how the tools in the Hadoop ecosystem can shine given that this is a dataset with over 55 million records and is multiple Gigabytes (5.7GB) in size.

**Exploratory Data Analysis on a Subset**

For this first analysis I wanted to stick to a tool I was first comfortable with just to get acclimated with the data, but then quickly moved on to analyzing the entire dataset with Spark. I also tried uploading the data in the Cloudera VM, but ran into many issues described further on in the process. I will be putting more time in to try to find a solution, but I have already spent many hours on the Azure VM trying to get the data onto the Cloudera VM. I attached the python notebook and data in a zipfile in the submission if you wish to run the same analysis. For this analysis I took a random subset of 100k rows and used python + pandas to quickly see the range of dates in the dataset as well as the distribution of fares. I also used a geospatial library in python to plot the longitude and latitude on a map to visualize the trips. Without knowing the specific routes the taxis took, it was interesting to see if I could try to predict the route that was taken just by looking at the pickup and drop off locations. I used the command from the following StackOverflow post (<https://unix.stackexchange.com/a/328241>) to get a 0.2% sample of the 5.7GB file to work with.

The first analyses I ran on the sample was to find the min and max pickup date and plot the distribution of fares using a boxplot.

A picture containing graphical user interface

Description automatically generated

After running those analyses, I wanted to see if I could find a way in python to overlap the pickup and dropoff longitude and latitude over a map of New York City. In the sections above I described not knowing what a shapefile was, but in order to overlap the longitudes and latitudes over a map of NYC I needed to download a shapefile of all 5 boroughs of New York City. Using a package called Geopandas I was able to plot the points on top of a map of New York City. It may not be a statistical analysis like getting the min/max of the distribution, seeing the shape or getting other distribution statistics like the standard deviation, but it does help us visualize the dataset we are analyzing and I thought it would contribute when we tell the story of the data. See the map below of the first 100 pickup locations:

A picture containing graphical user interface

Description automatically generated

From there it is easy to see that Manhattan is the dominant borough for pickup locations. I got the shapefile from NYC Open Data, which coincidentally we were using in Week 1 of the class. The shapefile can be downloaded here (<https://data.cityofnewyork.us/City-Government/Borough-Boundaries/tqmj-j8zm>) and is already included in the attached zipfile.

In the next section I discuss some first preliminary attempts at analyzing this 5.7GB NYC Taxi Fare Dataset using tools taught in class. I really enjoy learning new technology, so it was interesting trying the different techniques. I’m eager to work with my group for the final project to see what we can come up with and figuring out what tools are most effective for our analysis.

**Initial Attempts Using Big Data Techniques from Class**

For the Exploratory Data Analysis, I also wanted to initially try a few of the Big Data Management techniques to get a headstart on the final project. The first one I tried was to install **Spark** on my Mac and to use **PySpark** in the Jupyter notebook.

Below is the Spark Web UI on my computer running on localhost:4040. Use the README for installation instructions on a mac, installation instructions on Windows may be slightly different.

Graphical user interface, table

Description automatically generated

Spark provides an interface for programming against entire clusters and comes with parallelism and fault tolerance built in. I’m still heavily learning the PySpark interface and Spark in greater detail, but the first thing I’ve noticed with PySpark is that you must create a Session with the SparkConfig. This SparkConfig lets you specify the URL of the master node of your cluster running Spark, a name of an app and any parameters you want to give Spark like max parallelization, resources like memory consumption, etc. When running commands in PySpark, PySpark will use this config to parallelize the commands for faster processing. See the screenshot below for an example local config for Spark that takes advantage of multiple cores. I read a few different blogs in an effort to further understand PySpark and this post (<https://www.section.io/engineering-education/getting-started-with-pyspark-spark-part2/>) particularly helped in understanding the different config parameters.

Graphical user interface, text, application

Description automatically generated

If my group chooses to use this Taxi Dataset, then I plan to further my understanding of Spark and perhaps even try to spin up a cluster on AWS to I can see Spark working across different worker nodes.

The next Big Data Management tool I tried during the Exploratory Data Analysis was to upload my 5.7GB CSV file into **Hive** using the commands from the following post (<https://stackoverflow.com/a/32795093>). After a number of attempts to try to download the Dataset onto the Cloudera VM, I was not able to successfully upload data to Hadoop. The issues I ran into were that I could not get to Kaggle inside the Cloudera VM to download the dataset so I had to find another way of getting data onto the VM. The first thing I tried was to download another browser like Chrome or update the Firefox browser inside the Cloudera VM, but I was not able to find out how to do so. The next thing I tried was to attach the VHD (Virtual Hard Drive) as a letter drive on windows and copy the files from the Windows VM onto the Cloudera VM, but I was not able to do that successfully either (I believe its because Windows does not understand the Cloudera VM filesystem) After a couple of hours trying different methods to get the data onto the Cloudera VM I had to discontinue and hope to resume trying more attempts later in the week so I can use Hadoop on the Final Project.

**Building a real-time system**

Getting insight from analyzing the historical data provides tremendous value, but what can unlock even more value is to be able to analyze the data in real-time to support just-in-time decision making. If cab drivers in New York could be alerted in real-time to traffic congestion or predict traffic congestion along their route before the route is congested, then they will be able to optimize the route driven to avoid clogged roads. Building these real-time systems that collect data from millions of nodes is complex and requires scalable architecture to accomplish the task well. Companies like Uber contribute to Open Source (<https://uber.github.io/#/>) so it gives a small glimpse into some of the systems they have built around the challenge of rideshare mobility. I would be really interested in talking to engineers at the companies running at massive scale to understand how they build real-time systems that deliver value to passengers and drivers.

**Conclusion**

As cities around the world continue to grow, it’s important that residents and leaders are thinking about new ways to improve mobility to avoid overcrowding and congestion that can often bring cities to a halt. With our national discourse centered around Equity, it is now more imperative than ever to ensure that we create Equitable cities where all citizens can move around freely to get to new opportunities and climb the economic ladder. In New York City, Taxis are one of the most convenient modes of transportation but historically have been associated with denying rides to some individuals in times due to disability, destination or race. Now that we can collect data with more ease, we can start to shed a light on taxicab practices and learn from patterns we can only start to get insight on by analyzing large sums of data. We can start to find patterns to optimize where taxis can position themselves for optimal pickup based on time of day, seasonal trends and on holidays. If my group chooses this project than I will be eager to see what we are able to accomplish with the different big data tools taught in class. I hope to be able to start to paint a picture of taxi practices in NYC that can ultimately help to make the system more efficient which can help to alleviate traffic and reduce unnecessary emissions. In addition to creating Equitable cities, city leaders are also talking about creating Sustainable cities and reducing congestion is one of the best ways we can remove unnecessary emissions.

References

1. New York City Taxi Fare Prediction - https://www.kaggle.com/c/new-york-city-taxi-fare-prediction/overview
2. Prepare to Pay More for Uber and Lyft Rides - https://www.nytimes.com/article/uber-lyft-surge.html
3. Taxi and Ridehailing Usage in New York City - https://toddwschneider.com/dashboards/nyc-taxi-ridehailing-uber-lyft-data/