Logo

Description automatically generated

Week 6 Final Writeup – Impact of Covid-19 on Yellow Cab Ridership in New York City

Craig Perkins – perkins.cr@northeastern.edu

Northeastern University

December 2021 – ALY6610.70531

**Week 6 Final Writeup – Impact of Covid-19 on Yellow Cab Ridership in New York City**

For our final project we chose to study the impact of the COVID-19 Pandemic on Yellow Cab Ridership in New York City using open data provided by the Taxi and Limousine Commission of New York City`. Using data on Yellow Cab rides from January 2019 – July 2021, we sought to quantify the impact of the global pandemic on the ridership of Yellow Cabs in New York City and assess the financial impact on the total revenue of the industry. In addition to analyzing the impact of the pandemic on ridership and revenue, we also tried to get insight into changing consumer behavior. At the onset, we hypothesized that the pandemic would result in more single-passenger trips with more cashless transactions. Our dataset is composed of 124.3M (11.46GB in total) records corresponding to a single trip taken in a yellow cab.

**Introduction**

The New York City Yellow Cab industry has been reeling long before the pandemic2 and the pandemic has exacerbated the existing issues of the industry2. Despite the challenges in the industry, taxis and rideshare provide an essential service to New Yorkers by providing an unmet mobility need for many in the city. During the pandemic, rideshare was seen as a safer alternative to crowded public transportation by many and provided a way to get around the city while minimizing one’s proximity to others. Rideshare services also provide a necessary service for New Yorkers with disability who often may live away from accessible options for public transportation. With that in mind, it’s imperative to understand how the pandemic has affected the lives of the New Yorkers that keep the city moving. For this analysis we use Big Data techniques, like Apache Spark and Azure Databricks, to analyze data from over 100 million yellow cab trips in New York City from January 2019 to July 2021.

# Dataset Sourcing and Details

The full 11.46GB dataset can be downloaded at the following link: <https://northeastern-my.sharepoint.com/:u:/g/personal/perkins_cr_northeastern_edu/EXnkAbD4KDBMn1meaOQ6WCwBKaBDDkcDu5RtpWoQAKhpxQ?e=1dUmFk>

Our dataset comes from the Taxi and Limousine Commission (TLC) of New York City. The TLC publishes monthly CSV files for each service in the TLC’s purview which encompasses Yellow Cabs, Green Cabs, For-Hire Vehicle (lower volume rideshare service) and High Volume For-Hire Vehicles (Uber, Lyft, Via and Juno). For our analysis we focused on Yellow Cabs and wrote a python file to download all of the months from January 2019 to July 2021. The code for downloading the files from the TLC website and the accompanying code for analysis can be found on Github at: <https://github.com/cwperks/ALY6110>. After downloading all of the CSV files from TLC we then combined them with the following command to get one big CSV file with 124.3M records.

*head -n 1 file1.csv > combined.out && tail -n+2 -q \*.csv >> combined.out*

Once our dataset was assembled we then sought to understand the contents of the dataset and read the Data Dictionary provided by the TLC (<https://www1.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf>)

These are the dimensions and a description of each in the dataset:

|  |  |
| --- | --- |
| **Field Name** | **Description** |
| **VendorID** | A code indicating the TPEP provider that provided the record.  **1= Creative Mobile Technologies, LLC; 2= VeriFone Inc.** |
| **tpep\_pickup\_datetime** | The date and time when the meter was engaged. |
| **tpep\_dropoff\_datetime** | The date and time when the meter was disengaged. |
| **Passenger\_count** | The number of passengers in the vehicle.  This is a driver-entered value. |
| **Trip\_distance** | The elapsed trip distance in miles reported by the taximeter. |
| **PULocationID** | TLC Taxi Zone in which the taximeter was engaged |
| **DOLocationID** | TLC Taxi Zone in which the taximeter was disengaged |
| **RateCodeID** | The final rate code in effect at the end of the trip.  **1= Standard rate**  **2=JFK**  **3=Newark**  **4=Nassau or Westchester**  **5=Negotiated fare**  **6=Group ride** |
| **Store\_and\_fwd\_flag** | This flag indicates whether the trip record was held in vehicle memory before sending to the vendor, aka “store and forward,” because the vehicle did not have a connection to the server.  **Y= store and forward trip N= not a store and forward trip** |
| **Payment\_type** | A numeric code signifying how the passenger paid for the trip.  **1= Credit card**  **2= Cash**  **3= No charge**  **4= Dispute**  **5= Unknown**  **6= Voided trip** |
| **Fare\_amount** | The time-and-distance fare calculated by the meter. |
| **Extra** | Miscellaneous extras and surcharges. Currently, this only includes the $0.50 and $1 rush hour and overnight charges. |
| **MTA\_tax** | $0.50 MTA tax that is automatically triggered based on the metered rate in use. |
| **Improvement\_surcharge** | $0.30 improvement surcharge assessed trips at the flag drop. The improvement surcharge began being levied in 2015. |
| **Tip\_amount** | Tip amount – This field is automatically populated for credit card tips. Cash tips are not included. |
| **Tolls\_amount** | Total amount of all tolls paid in trip. |
| **Total\_amount** | The total amount charged to passengers. Does not include cash tips. |

Below is a random sample of 10 records from the dataset:

Table

Description automatically generated

**.Data Analysis using Spark**

Given the size of our dataset, our group needed to brainstorm Big Data solutions to effectively analyze our large dataset. We set out to design a solution that would scale to handle the full volume of TLC data including information from ride-shares and scale with increasing volumes in the future. In an effort learn more about the Big Data solutions available, our group chose to try and use multiple methods to analyze the dataset. 2 of those methods include:

1. Running Apache Spark in a cluster using Docker with 1 Master and 2 worker nodes
2. Upload data to Azure Databricks Delta Lake and use Spark with a cluster in Azure
3. **Connecting to a Spark Cluster running in Docker**

The code for this analysis and instructions to run it are located at: <https://github.com/cwperks/ALY6110>

For the proposal from week 3, many members of our group had expressed being comfortable with Jupyter Notebooks and performing Exploratory Data Analysis in a Jupyter Notebook. From the start, it was obvious that pandas would not be able to load 11.46GB worth of data into memory so pandas was a non-starter for the analysis. That led us to PySpark which allowed us to try spark in Local Mode. Local Mode allowed us to load the CSV file using PySpark and run analyses while getting comfortable with the PySpark API (Application Programming Interface). Using Spark in Local Mode showed us the power of spark and gave us some immediate benefits by using up the full number of cores on our local machines and parallelizing computation. Below is the PySpark code we used to initialize spark and read in the Taxi Cab dataset:Graphical user interface, text

Description automatically generated

1. **Cleaning**

Prior to creating this visualization, we also found that we needed to “clean” the dataset of invalid entries and null values. The first analysis we ran to show PySpark working in Local Mode was to get the value\_counts of the passenger\_counts in our dataset. Below is the distribution of passenger\_counts in the full dataset:

Graphical user interface, table

Description automatically generated

With each step of the process we were able to use the Spark UI at <http://localhost:4040> to see the spark jobs progress.

After running the analysis above to get the distribution of passenger counts, we realized that there were null values in the dataset. The first analysis we wanted to visualize was to get the count of rides by month based on the pickup date of the ride. In order to accomplish that visualization we needed to remove the dataset of records where the ` tpep\_pickup\_datetime` was null. To do that we ran the following code:

|  |
| --- |
| *import pyspark.sql.functions as F*  *taxi\_fare\_filtered = taxi\_fare\_df.where(F.col('tpep\_pickup\_datetime').isNotNull())* |

After removing records with *null* pickup time, we then needed to engineer a new column in the dataset corresponding to the pickup month. To do that we added a new column called `month` and used the following datetime format (YYYY-mm) to represent the month of pickup. To engineer the `month` column we ran the following code:

|  |
| --- |
| *from dateutil import parser*  *import pyspark.sql.functions as F*  *taxi\_fare\_filtered = taxi\_fare\_filtered.withColumn("month", F.date\_format(F.to\_date(F.col("tpep\_pickup\_datetime")), "yyyy-MM"))* |

Once we computed the new column we wanted to see how many rides occurred in each month. To accomplish that we took advantage of the RDD `.groupBy` function to group the DataFrame by month and perform a `.count()` aggregation. Below is a screenshot of the output we got when running the `.groupBy()`.

Graphical user interface, text, application, email

Description automatically generated

We immediately noticed that there were some errant months in our dataset, so we performed some additional cleaning on the data to ensure that only months from the range January 2019 – July 2021 were included in the dataset.

After we were done cleaning, we then wanted to visualize the data. Below is the output of our first analysis of our first analysis of loading the full dataset (all 124.3M records) with PySpark and visualizing the total number of rides by month for each month in the dataset.Chart, bar chart, histogram

Description automatically generated

**From Local Mode to Cluster Mode**

After a correspondence between Fernando and the professor, we realized that running Spark in Local Mode was not sufficient for the Final Analysis. From this point we then tried 2 new methods of analysis for the dataset. Some members of the Group continued using Databricks from week 4 to analyze the data in a cluster on the cloud and other members of the group simulated a spark cluster locally on their computers using Docker containers. This section goes into details about creating a Spark cluster environment with Docker containers and contains screenshots showing the UI of the Master node and the 2 worker nodes in the cluster.

The screenshot below shows the 3 nodes running in Docker containers.

Graphical user interface, text, application, email

Description automatically generated

Below is the UI from the Master Node:

Graphical user interface

Description automatically generated with medium confidence

Below is the Web UI from Worker 1:

Table

Description automatically generated with medium confidence

Below is the Web UI from Worker 2:

Table

Description automatically generated

It was quite challenging to setup Spark to run in Docker containers, but once it was achieved it was informative to see Spark running in a Cluster environment. The biggest challenge of running Spark locally with Docker containers is that you don’t get to take advantage of the Hadoop Distributed File System which is often the backend used with Spark. The Hadoop Distributed File System is useful because it means that all of the Spark workers and Master nodes have access to the same resources. In order to simulate that in the Docker cluster, we needed to do a Volume mount to all of the containers to ensure that the same data was available across all nodes. See the README.md file in our repository (<https://github.com/cwperks/ALY6110>) for instructions on setting up Apache Spark to run in a Cluster mode locally.

The below code snippet is the same Spark Config from earlier but altered to assign the Master node of the cluster.

Graphical user interface, text, application, email

Description automatically generated

Despite the stress and challenge of setup, we found this approach to be rewarding and a great learning experience. It was a closer simulation to cloud cluster environment and we got to see real-time improvements in processing speed as we allocated more resources to the Docker containers.

**Analysis Continued**

After switching from Local Mode to Cluster Mode, we then continued our analysis on the dataset seeking to answer the questions we posed at the beginning: What has the pandemic’s impact been on ridership, revenue and have consumer behaviors changed?

In an earlier section we showed a chart showing the changing ridership month-to-month over the course of the pandemic. The chart below shows a similar chart, but shows revenue month-to-month:

Chart, bar chart, histogram

Description automatically generated

You can clearly see revenue go from >$140M from March of 2019 to <$10M in April of 2020. Its clear that the pandemic had a devastating toll on the Taxi industry in New York City. Furthermore, you can clearly see in this chart that revenue has not yet bounced back to pre-pandemic levels.

The next analysis we performed was trying to understand changes in consumer behavior. To do this we analyzed the count of the method of payment used by month, in particular, Credit Cards vs. Cash. Below is a chart of the count of transactions for both Credit Card and Cash from January 2019 – July 2021 month-to-month.

Chart, bar chart

Description automatically generated

While the chart may not show an observable drastic change in consumer behavior, there has been a change in the percent of rides paid for by credit card. Prior to the pandemic 70-72% of rides were paid for by credit card and post-pandemic that has risen to 75-79%. One interesting observation is that there were a higher proportion of cash transactions in the early days of the pandemic. Our hunch is that if we combine this dataset with the High-Volume For-Hire dataset from the TLC that we may be able to paint a clearer narrative of all of the competing for-hire services in NYC, but it is interesting to note the higher percentage of cash transactions in the early days of the pandemic.

Given more time to study this dataset we would attempt to find the most lucrative routes, analyze trends by time of day and also combine the Yellow-Cab dataset with the other datasets provided by TLC to get a clearer impact of the Pandemic on the entire industry.

**Conclusion**

COVID-19 has changed the world in many ways and for Yellow Cab Drivers in New York City, the impact has been particularly harmful. While this study showed us the power of Big Data analytics tools in obtaining insight from large datasets, it also illustrates how the data is used to illuminate the reality of a situation, good or bad. That is the power of Big Data, to be able to process and analyze ever increasing amounts of data to ultimately obtain insights and drive decision making. This was an interesting dataset to study in the context of our Final Project to try out different Big Data analytics tools, but in the eyes of New York City leadership this can help them deploy resources to where it is needed most. The use of Big Data analytics tools helped us take a large dataset and perform computations efficiently by breaking down the dataset for parallel computation to run across a cluster. It was not even possible for us to load the dataset we were using into memory, so this was a great example of how Big Data analytics tools differentiate from using vanilla Python with pandas. Big Data is changing the world and we are excited to see the new capabilities and insights companies and organizations will continue to achieve by deploying Big Data analytics tools like Hadoop and Apache Spark.

References

1. TLC Trip Record Data - <https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page>
2. How much is a NYC taxi medallion worth these days? - https://www.cbsnews.com/news/how-much-is-a-nyc-taxi-medallion-worth-these-days/
3. Pandemic Pushes N.Y.C. Cabbies to the Brink: ‘I Can’t Hold On’ - https://www.nytimes.com/2020/11/12/nyregion/nyc-taxi-drivers-coronavirus.html
4. Prepare to Pay More for Uber and Lyft Rides - https://www.nytimes.com/article/uber-lyft-surge.html
5. Taxi and Ridehailing Usage in New York City - https://toddwschneider.com/dashboards/nyc-taxi-ridehailing-uber-lyft-data/