

**DAT560M - Big Data and Cloud  
Computing**

Final Project Report

Data Set: Newyork Taxi Demand

## **Group :**

Ylfei Hu 491843

Xiaoxiang Liang 489743

Qinyi Liu 489537

Jingdi Han 489554

## **Table of Contents**

1. Data description
2. Problem Statement
3. Methods & Results
4. Conclusion
5. Appendix

### **● Data Description**

This structured dataset is from Kaggle available at

[https://www.kaggle.com/vishnurapps/newyork-taxi-](https://www.kaggle.com/vishnurapps/newyork-taxi-demand/download/yellow_tripdata_2016-03.csv)

[demand/download/yellow\\_tripdata\\_2016-03.csv](https://www.kaggle.com/vishnurapps/newyork-taxi-demand/download/yellow_tripdata_2016-03.csv) (Vishnu R, 2020). This dataset

collect trips log information of yellow taxi in March 2016. It is a 1.78 Gb enormous CSV file that has 12210953 rows and 19 columns including vendorID, pickup, and dropoff date, passengers amount, distance, longitude, and latitude of pick up and drop-off, RatecodeID INT, payment\_type, fare\_amount, tip\_amount, tolls\_amount, improvement surcharge, total payment. To analyze this dataset, big data-related tools are called because traditional data analysis tools such as MS Excel, Python, MySQL cannot process and execute queries to handle a mass of statistics in a tolerable time period. In order to analyze this dataset, Hadoop is used to download and reserve the CSV file from the website, and hive and impala are mainly utilized to solve the problem.

- **Problem Statement**

Once upon a time, yellow taxi, which shuttles through the streets of New York City, was a symbol of this metropolis. It witnessed the changes of New York and is also one of the important transportation modes for every new Yorker.

In recent years, with the rise of the sharing economy, the taxi industry in New York has begun to lose its glory. As in China, the rise of online car Hailing has triggered a round of discussions about the threat of shared travel to the traditional taxi industry.

Chart 2: BAC card spending on taxi/limo s, controlling for disruptors (% yoy)



Source: BAC internal data

According to the above chart, by observing the monthly consumption of debit and credit cards of all bank of America customers, it is found that although the overall consumption of taxi / limousine services has increased sharply, the expenditure on traditional taxis has continued to decline.

Uber not only challenges the traditional taxi business model, but also excels in the passenger experience. The user calls the car by mobile phone, Uber schedules the car according to the distance between the vehicle and the user, the driver picks up the order and carries passengers by Uber software, and the passengers complete the payment and score the driver on Uber software.

In this report, We are aimed to use big data tools to analyze trip data to help taxi drivers find some tricks to increase their income such as providing passenger attendance rates and increase received tips. To explore the features in dataset, research questions are listed below.

1. Which payment method is the most popular for taxi passengers in New York?
2. What is the relationship between the amount of tip and the trip distance?
3. Whether tips are related to the payment methods?
4. Is there a relationship between the number of passengers and the trip distance?

## ● **Method & Results**

In this research, we started off cleaning the Newyork taxi demand dataset using bash code. Specifically, we deleted the index column and header row. Then, we create a table in Hive and loaded the dataset, and synchronized it into Impala. We ran most of the queries in Impala and the rest in Hive. After answering some of the questions with Cloudera, we managed to use Tableau to plot charts to help us make a better analysis of the taxi demand data.

1. Which payment method is the most popular for taxi passengers in New York?

In this question, we used Impala to group number of passengers by payment type and order the data by the number of passengers reversely.

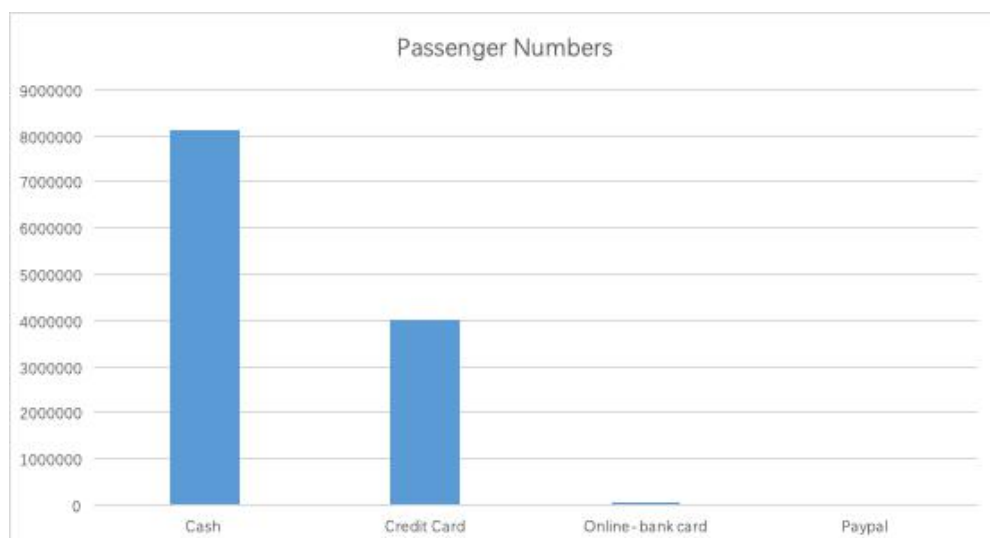
The result is listed as follows:

```
1 select payment_type, count(vendorid) as passenger_num
2 from group16_taxi
3 group by payment_type
4 order by passenger_num desc;
```

Query History Saved Queries Results (4)

|   | payment_type | passenger_num |
|---|--------------|---------------|
| 1 | 1            | 8127391       |
| 2 | 2            | 4020408       |
| 3 | 3            | 46913         |
| 4 | 4            | 16240         |

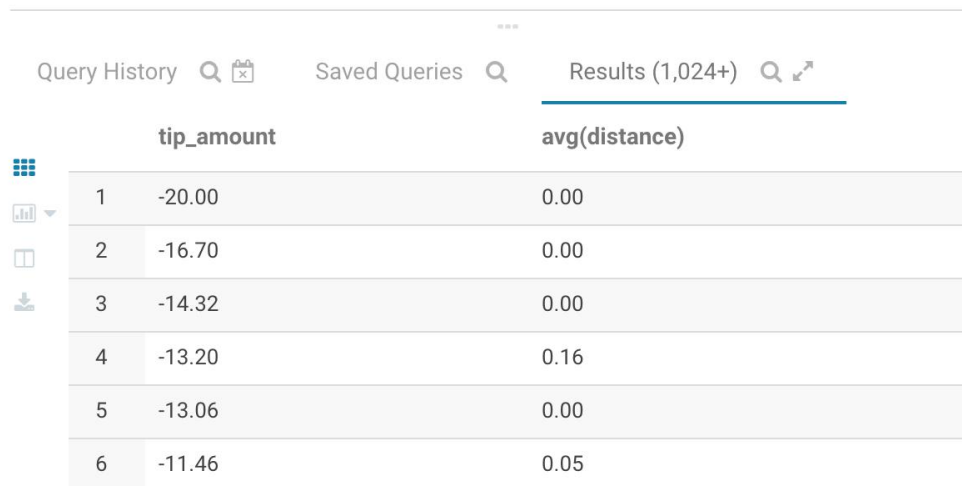
Where 1 stands for cash, 2 stands for credit card, 3 stands for online-bank card, 4 stands for Paypal.



As we can see from the chart, methods used to make by the taxi passengers are cash, credit card, online-bank card, and PayPal. The two main payment methods are cash and credit card. Cash payment took the dominant position, reached 8 million times in a single month. Payment by credit card didn't even reach half of the time of cash payment, by nearly 4 million times. Online payments like online-bank card and PayPal were only a few compared to cash and credit card, credit card payment was 47000 times and PayPal payment was only about 16000 times.

## 2. What is the relationship between the amount of tip and the trip distance?

```
11 select tip_amount, avg(distance) from group16_taxi
12 group by tip_amount
13 order by tip_amount asc;
```

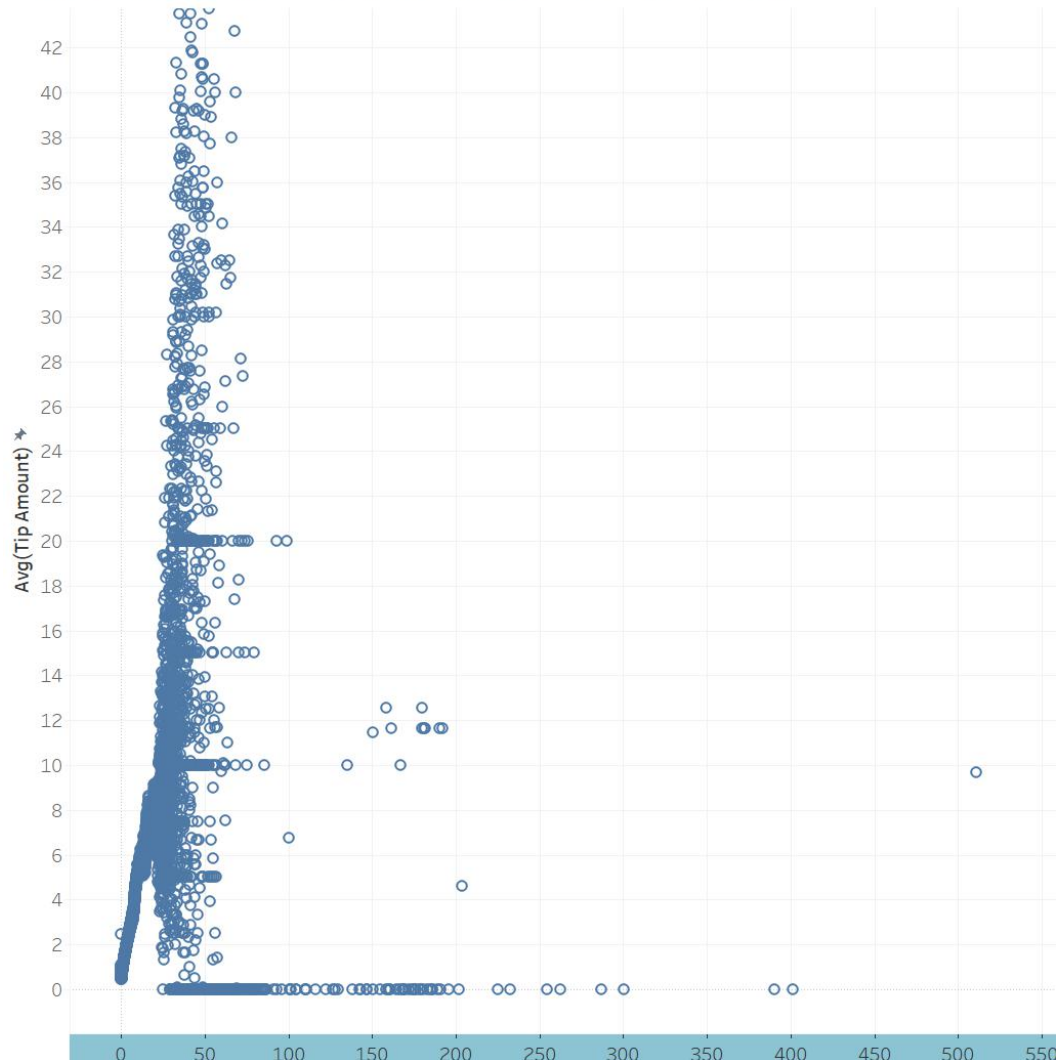


|   | tip_amount | avg(distance) |
|---|------------|---------------|
| 1 | -20.00     | 0.00          |
| 2 | -16.70     | 0.00          |
| 3 | -14.32     | 0.00          |
| 4 | -13.20     | 0.16          |
| 5 | -13.06     | 0.00          |
| 6 | -11.46     | 0.05          |

Impala is used to execute query about the relationship between distance and average tips and Tableau is utilized to visualize the dataset. As the scatter plot illustrated, there was a pattern that long-distance trips tend to give none tip that may result from large fare amount make passengers mean with tips.

Furthermore, in short trips, tips are positively related to distance while there is a large violation in tips given in middle-distance trips.

Scatter plot between distance and average tip\_amount



### 3. Whether tip is related to the payment method?

In this question, we used Impala to group the average tip amount by payment type and order the data by the average tip\_amount reversely. The result is listed as follows:



24.008 DBTCloud TEXT

```
1 SELECT avg(tip_amount),payment_type
2 from group16_taxi
3 group by payment_type
4 order by avg(tip_amount) desc;
```

Query c64fc750e330a241:b893b97700000000 100% Complete (4 out of 4) [c64fc750e330a241:b893b977000000](#)

Query History Saved Queries Results (4)

|   | avg(tip_amount) | payment_type |
|---|-----------------|--------------|
| 1 | 2.69            | 1            |
| 2 | 0.00            | 3            |
| 3 | 0.00            | 2            |
| 4 | 0.00            | 4            |

From that conclusion, we can see that the first payment\_type ----cash will make taxi drivers get more tips.

4. Is there a relationship between the number of passengers and the trip distance?

```

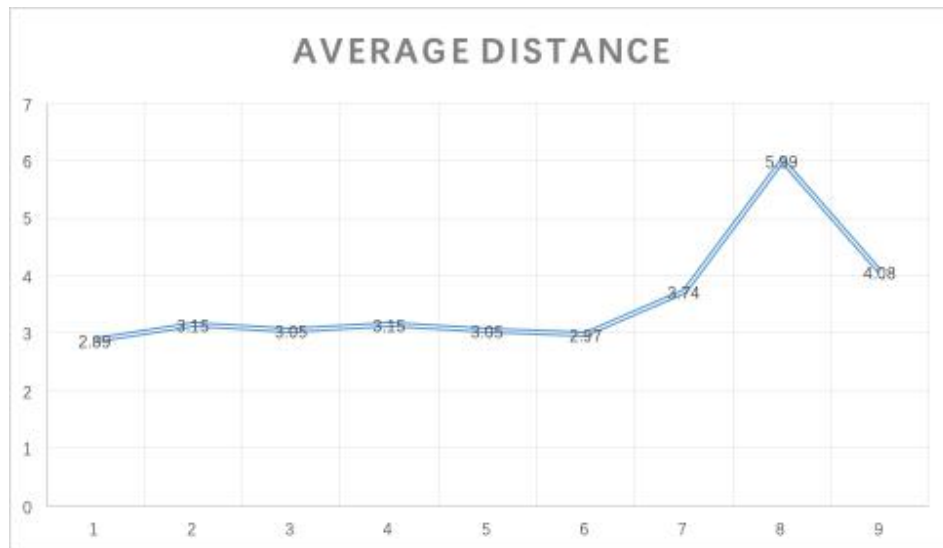
6 select passengers, avg(distance)
7 from group16_taxi
8 group by passengers
9 order by passengers desc limit 9;

```

Query History Saved Queries Results (9)

|   | passengers | avg(distance) |
|---|------------|---------------|
| 1 | 9          | 4.08          |
| 2 | 8          | 5.99          |
| 3 | 7          | 3.74          |
| 4 | 6          | 2.97          |
| 5 | 5          | 3.05          |
| 6 | 4          | 3.15          |
| 7 | 3          | 3.05          |
| 8 | 2          | 3.15          |
| 9 | 1          | 2.89          |

Impala is used to execute query about the relationship between trip distance and passenger numbers, Excel is utilized to visualize the dataset. As the line chart illustrated, passenger numbers between 1-6, these groups of passengers all traveled near 3 miles per car. When the number comes to 7 or more people in the same car, people in New York are more likely to travel to a further destination. Among these groups of people, when 8 people take a taxi together, they are likely to take trips around 6 miles, which are nearly doubled compared to trips under 6 people.



- **Conclusion**

After carefully analyzing the data, we came up with the conclusion that:

. As cash and credit cards are the most popular payment methods. In the most cases, passengers tend to give more tips if their trips are longer when less than 50-foot distance, but drivers tend to receive tips from passengers who pay by cash. Also, if the passenger number is under 6, people tend to take a shorter trip; if the number rises beyond 7 people, they tend to take a longer trip. This shows that people in New York consider about the efficiency of traffic and get more people involved into a longer trip.

## Appendix

Problem 1:

```
select payment_type, count(vendorid) as passenger_num  
from group16_taxi  
group by payment_type  
order by passenger_num desc;
```

Problem 2:

```
select tip_amount, avg(distance) from group16_taxi  
group by tip_amount  
order by tip_amount asc;
```

Problem 3:

```
SELECT avg(tip_amount), payment_type  
from group16_taxi  
group by payment_type  
order by avg(tip_amount) desc;
```

Problem 4:

```
select passengers, avg(distance)  
from group16_taxi  
group by passengers
```

