

# NYC Taxi and Limousine Commission Analysis & Passenger Count Prediction

Data Science TalentHub Kemnaker  
Final Project

Muhammad Randa Yandika






# Project Backgorund

In this project, we study a comprehensive analysis of a taxi trip dataset. Dataset comprises a wide range of variables related to taxi trips, including details such as pick-up and drop-off times, distance, fare, type of payment and number of passengers.

The main objective of the project is to uncover insights and patterns within the data that could provide valuable information for increasing the efficiency of taxi operations,

By building a predictive model, we intend to offer insights to taxi service providers that can help them allocate resources more effectively, optimize fleet management, and cater to varying demand levels at different times of the day.





# TABLE OF CONTENTS

**01** Data  
Preparation

**02** EDA

**03** Modelling

**04** Conclusion





01

# DATA PREPARATION

# DATASET

## Train

2 Attributes  
6305 Instances

## Test


2 Attributes  
1576 Instances

## Trip


20 Attributes  
54224 Instances




# Dataset Trip Description




<b>vendor_id</b>	Code indicating the LPEP provider. 1 = Creative Mobile Technologies, LLC; 2 = VeriFone Inc.
<b>pickup_datetime</b>	Date and time when the meter was activated.
<b>dropoff_datetime</b>	Date and time when the meter was deactivated.
<b>passenger_count</b>	Number of passengers in the taxi including the driver.
<b>trip_distance</b>	Reported distance of the trip on the meter.
<b>rate_code</b>	Final rate code in effect at the end of the trip. 1=standard rate; 2=JFK; 3=newark; 4=nassau/westchester;5=negotiatedfare;6=group ride
<b>store_and_fwd_flag</b>	Flag indicating whether the trip record was stored in vehicle memory before being sent to the vendor.
<b>payment_type</b>	Numeric code indicating how the passenger paid the trip fare.
<b>fare_amount</b>	Time and distance fare calculated with the meter.



# Dataset Trip Description (2)



<b>extra</b>	Various extras and surcharges.
<b>mta_tax</b>	Automatically triggered \$0.50 MTA tax based on metered rate used.
<b>tip_amount</b>	Amount of tip. This variable is automatically filled for credit card tips.
<b>tolls_amount</b>	Total amount of all tolls paid in the trip.
<b>imp_surcharge</b>	\$0.30 improvement surcharge assessed on trips booked at the beginning of the trip.
<b>airport_fee</b>	Airport fee
<b>total_amount</b>	Total fare charged to the passenger excluding cash tips.
<b>pickup_location_id</b>	TLC taxi zone where the meter was activated.
<b>dropoff_location_id</b>	TLC taxi zone where the meter was deactivated.



# Dataset Passenger (Train)



<b>pickup_datetime</b>	Date and time when the meter was activated.
<b>passenger_count</b>	Number of passengers in the taxi including the driver.





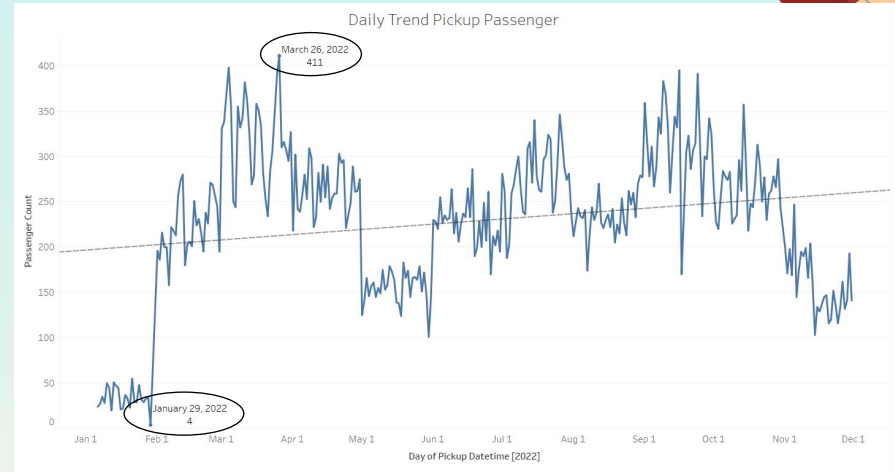
02

# EXPLORATORY DATA ANALYSIS



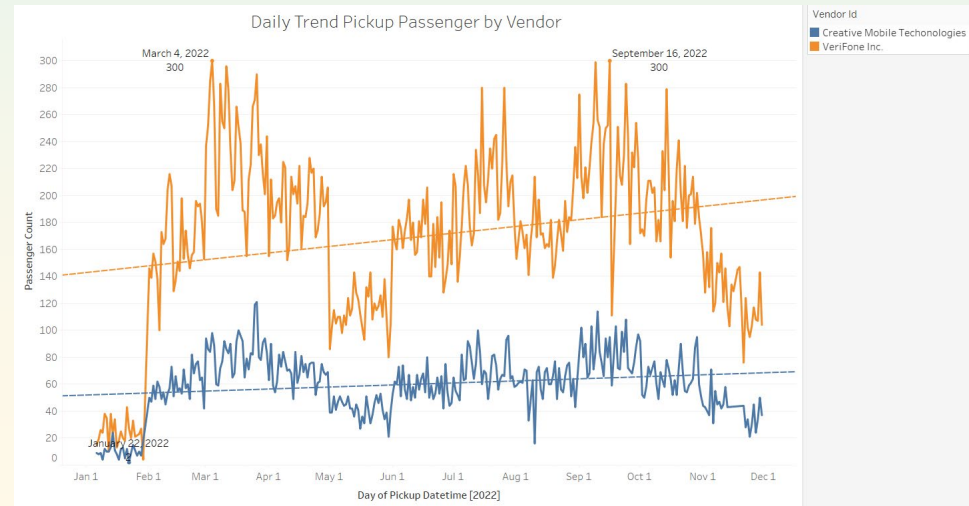
## What is the all-time trend of the daily passenger pickup?

There was an increasing trend this year; a drastic increase occurred in early February and decreased again in May. March 26 got the most passengers, namely 411 people, and there were only 4 passengers on January 29.



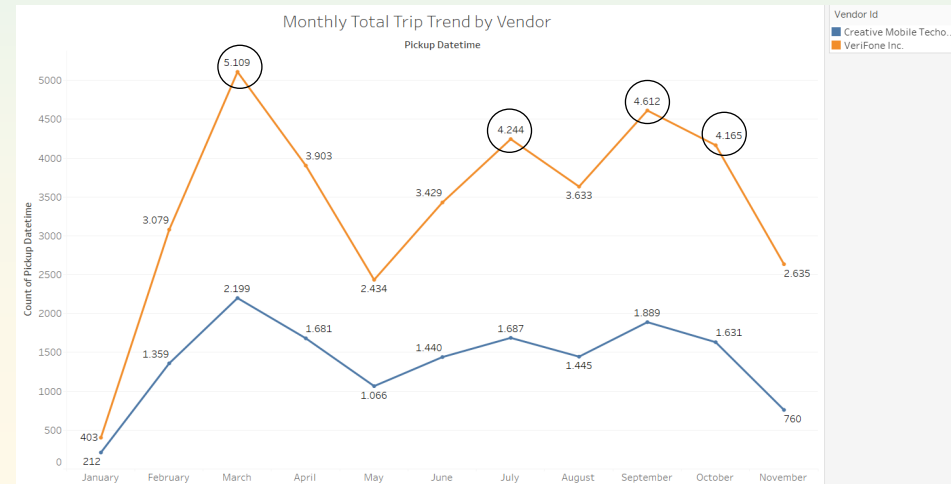
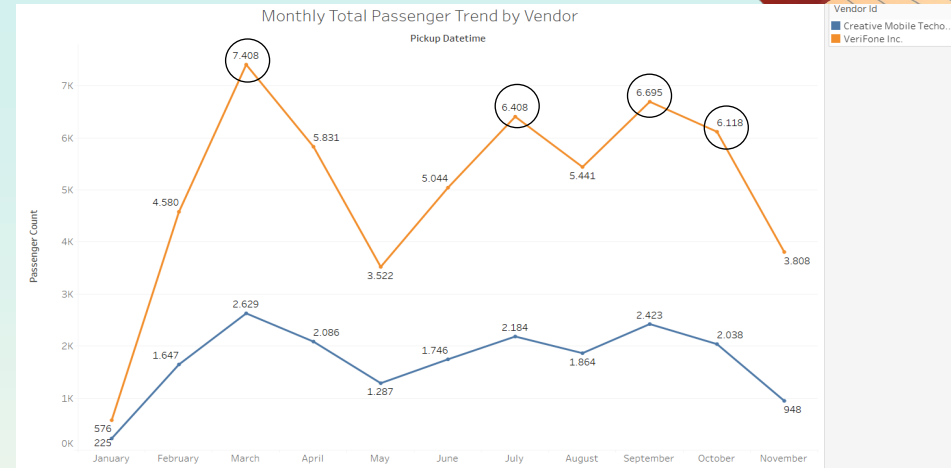
## What is the all-time trend of the daily passenger pickup by each vendor?

The trends for the two vendors are slightly different; Verifone has an increasing trend, and CMT has been more stable this year. With the highest number of passengers on March 4 and September 16, 300 passengers, and the lowest on January 22, 2 people.



What is the monthly trend for the number of trips and the number of passengers recorded?

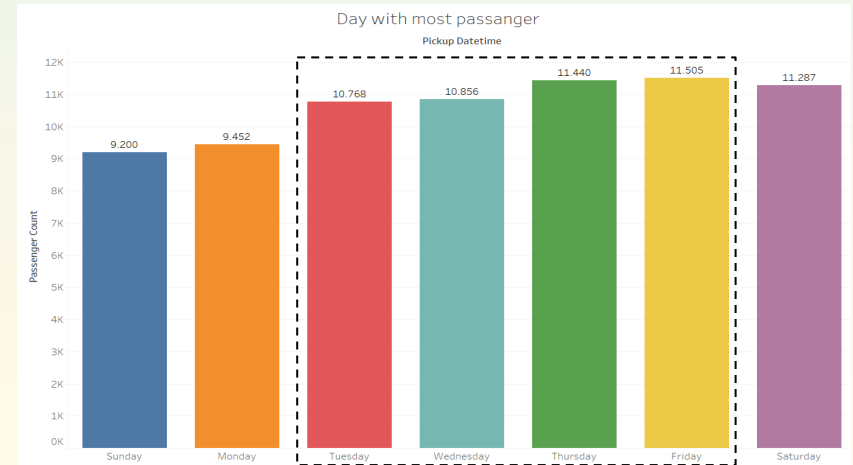
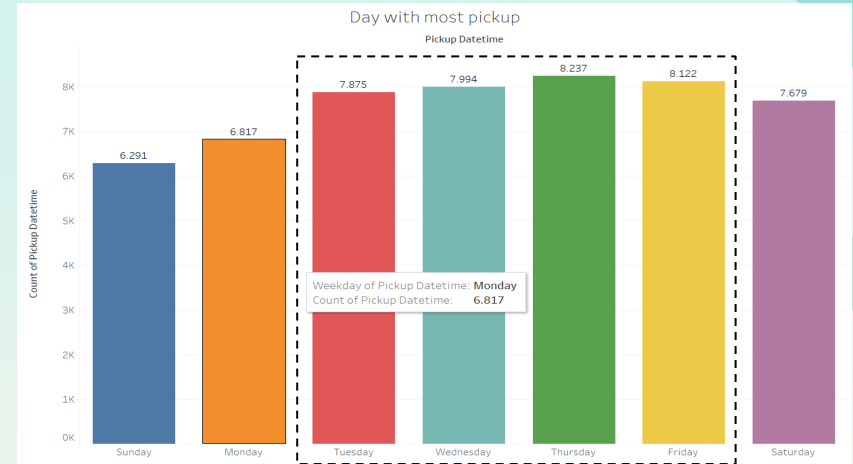
March, July, September, and October make the most trips and carry the most passengers. Both of these have a positive correlation and have the same trend for both.



## When are the days when there are many trips and carry many passengers

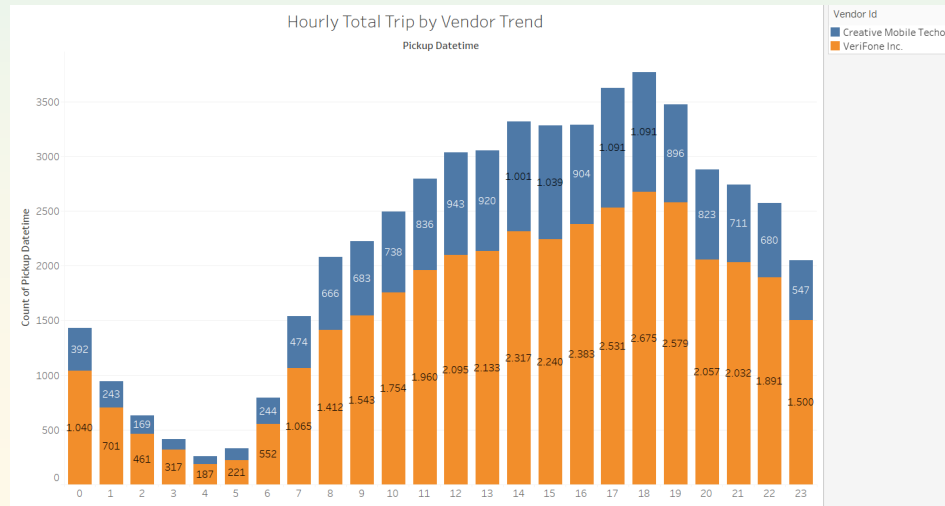
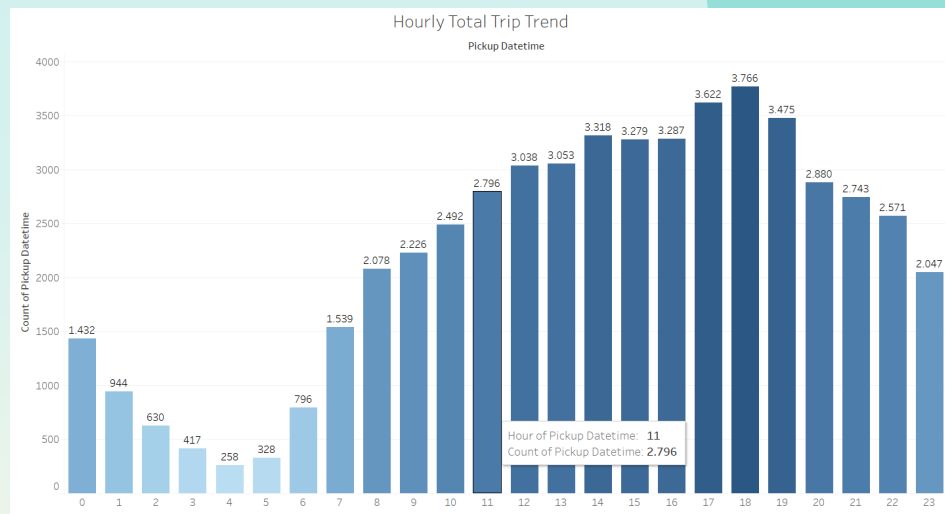
The most frequent trips and the highest number of passengers are from Tuesday to Friday;

From this insight, vendors can adjust the number of taxis available to be more than on weekends and Mondays.



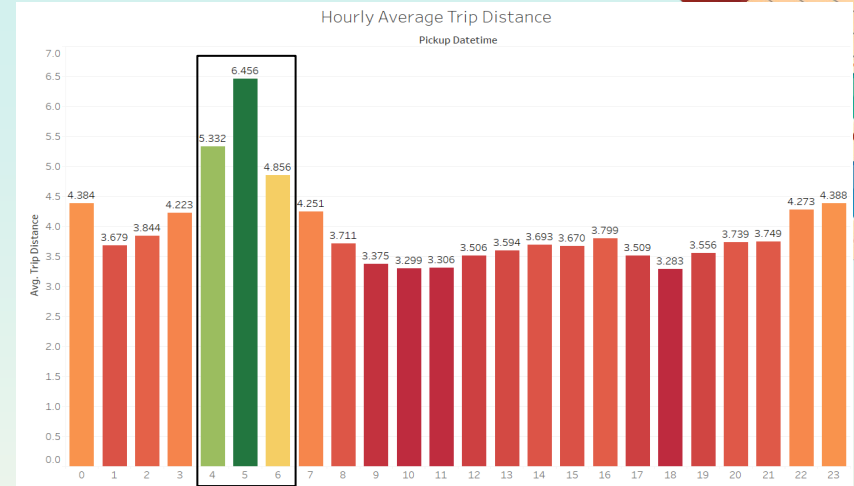
## What is the total trip per hour and what is it for each vendor?

There is an increase in taxi trips from 4 a.m. to 6 p.m., then a decrease again until 4 a.m. the next day. Most trips are during peak hours, around 4 to 7 p.m. The number of trips is dominated by Verifone, and CMT only makes 50% of the trips.



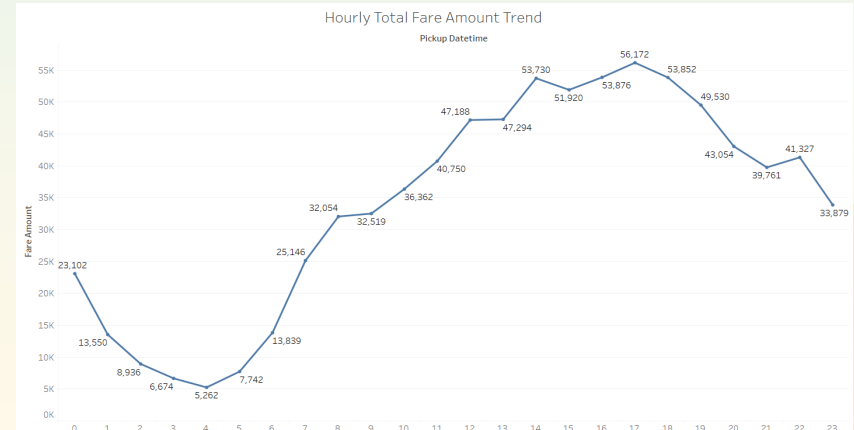
## When is the time with the longest average trip distance?

In the morning, around 4 to 6 o'clock is the time that has the farthest trip distance in one day. With the farthest average of 6,456 miles in the morning and the lowest at 6 p.m. with an average of 3,283 miles.



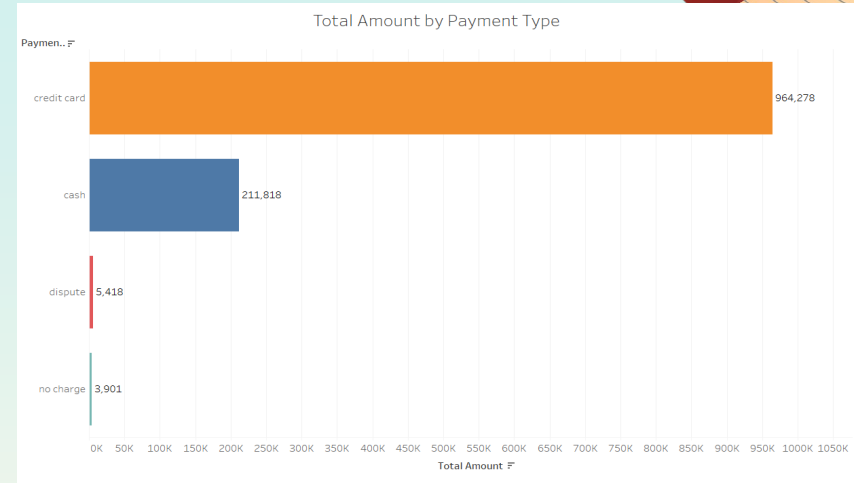
## How is the trend fare amount in every hour?

Fare amount, trip, and passenger count have similar trends, proving that these three features are correlated. Most fare amounts are given during peak hours, around 2 p.m. to 6 p.m.



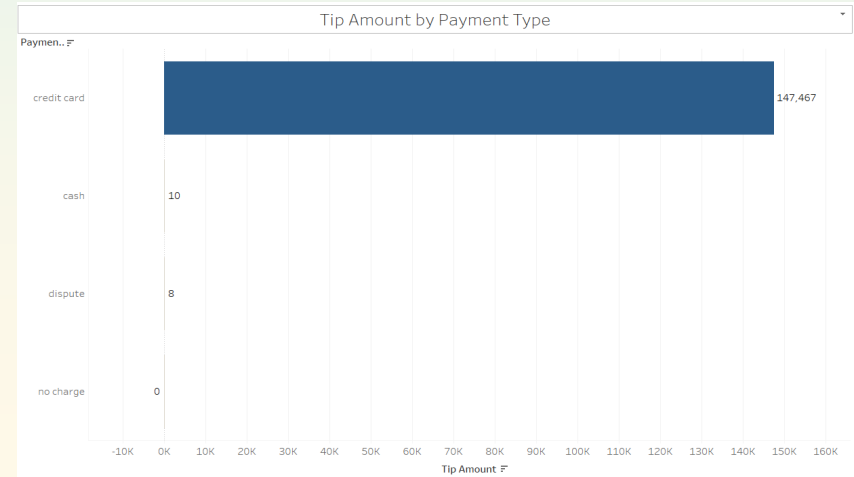
## What is the total amount in each type of payment?

Payment using a credit card is most in demand in this case. Very far compared to other payments, maybe the convenience factor in paying is the reason for passengers to pay by credit card.



## What is the tip amount in each type of payment?

Even when giving tips to drivers, passengers prefer to use a credit card. Of the total amount with CC, around 10% is a tip for drivers.

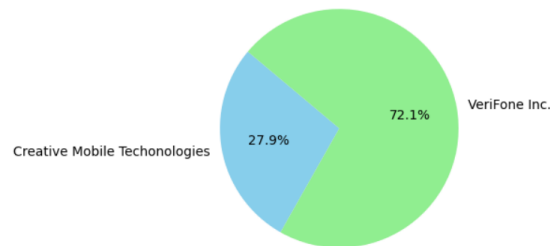


## What is the percentage of the total amount that each vendor gets?

Verifone Inc. earned 854,465.88 dollars, or 72.1% of the total amount earned, and CMT earned 339,949.55 dollars, or 27.9% of the total amount. Verifone dominates, with the possibility of having more taxi fleets.

```
vendor_id  
Creative Mobile Technologies    330949.55  
VeriFone Inc.                  854465.88  
Name: total_amount, dtype: float64
```

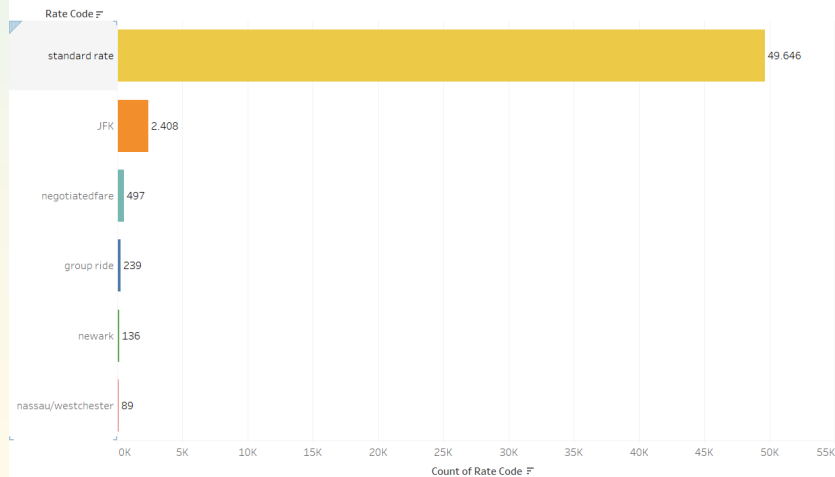
Total Amount per Vendor



## What is the number of rate\_code for trips in this year

The standard rate dominates the type of rate code on trips this year, with as many as 49,646 trips with a standard rate. The standard rate is still in great demand by many passengers.

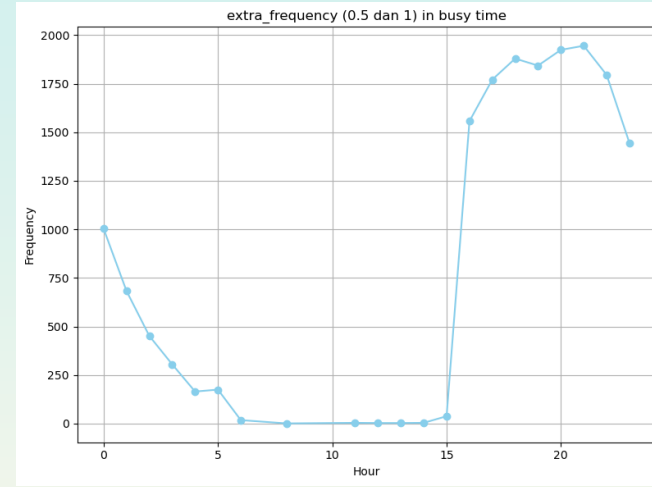
Rate Code Trip Count





## At what time will driver get extra for rush hour ? (0.5 & 1)

Rush hour will occur from 4 pm to 10 pm, if the driver picks up passengers at that hour, it will be possible to get extra on the trip. And the one who gave the most extra was the trip at 9 pm 1945 times giving extra.




## What percentage of taxi will get extra, tip, and both?

Every trip, a taxi is likely to get, Percentage getting extra: 60.15%.  
Tipping percentage: 75.10%  
Percentage to get both extra and a tip: 46.10%




# Conclusion

**March, July, September, and October** are the 4 months with the **highest passenger count**, the **most trips**, and the **largest total amount**.

- In that month, it is better for vendors to provide more taxis to prepare in case of high demand. Apart from these 4 months, **vendors can adjust the number of taxis based on existing insights for each month**.
  - From **4 a.m. to 6 p.m.**, **there is an increase in demand for taxis**. Therefore, vendors must provide sufficient taxis during these hours. And during these hours, there will also be an increase in the number of passengers; until 6 p.m., the number of taxis on standby can be adjusted.
  - Passenger count, total pickup, and fare amount **have the same pattern or trend every day**. The increase occurred from 4 a.m. to 6 p.m., then decreased in the evening.
  - The most **preferred type of payment** by passengers is using a **credit card**, with a resulting **amount of \$964,278.23**, followed by cash, with a resulting amount of 211,818.16. Likewise, **giving tips to drivers** using a **credit card** with a total of **147,466 dollars**
- 



# Conclusion

- Verifone Inc. **generated 72.1%, or 854,465.88, of the total amount**, while Creative Mobile Technologies only **generated 330,949.55, or 27.9%, of the total amount**. Verifone Inc. seems to have a larger number of taxis because of the number of trips, the fare amount, and the total amount generated, which is far more than Creative Mobile Technologies.
  - The **most frequent trips** and the **highest number of passengers** are from **Tuesday to Friday**; from this insight, **vendors can adjust the number of taxis** available to be more than on **weekends and Mondays**.
  - **Giving extra peak hours (0.5 and 1)** mostly on pickups **from 4 p.m. to midnight**, then the amount of extra giving decreases thereafter.
  - **Standard Rate is the highest rate code** on the trip with 49,646, followed by JFK with 2408.
  - **JFK** has a **flat fare** amount of 52 dollars.
  - **75.1% chance** that the driver **will get tips** from passengers and **60.15% chance** that they **will get extras** for both rush hours, overnight fees, and other extras. and a **46.16% chance of getting both**.
- 

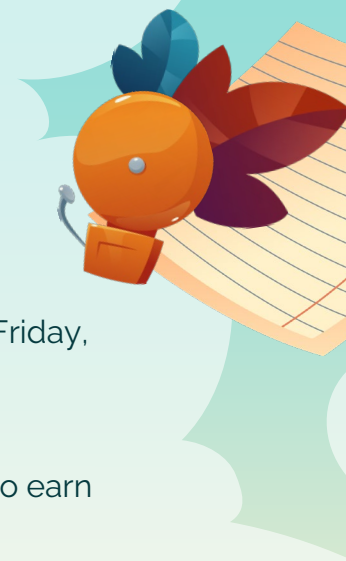
# Business Recommendation



- 1. Adjustment of Number of Taxis by Month:** By knowing that certain months have higher demand and income, vendors can make adjustments to the number of taxis they provide in those months. This will help them optimize their service and ensure that they have enough taxis to meet demand.
- 2. Optimization of Taxi during Rush Hour:** During peak hours between 4 a.m. and 6 p.m., there is an increase in demand and the number of passengers. Vendors can ensure that during these hours, they have sufficient taxis available on the street. This will help them earn more income.
- 3. Promotion for Payment by Credit Card:** Considering that the majority of passengers prefer payment by credit card and also provide more tips by credit card, vendors can carry out promotions or incentives to encourage more passengers to use this payment method. This can increase tipping revenue and make transactions easier.
- 4. Seasonal Promotions:** Based on the patterns found in the analysis, vendors can launch seasonal promotions during high-demand months, such as providing discounts in certain months such as March, July, September, and October.



# Business Recommendation



5. **Daily Taxi Adjustment:** Considering certain days have higher demand, such as Tuesday to Friday, vendors may increase the number of taxis available during this period.
6. **Extras and Tips:** Given the large number of passengers giving extras and tips, vendors can provide training and incentives to drivers to provide better service and get the opportunity to earn more extras and tips.
7. **Extra Rush Hour Promo:** During peak hours, vendors can provide discounts or additional services for passengers who still choose taxis when extra peak hour rates apply.
8. **Weekend Promotions:** Vendors may run special promotions for weekends, such as discounts on Saturday and Sunday trips. This will encourage more passengers to use taxi services during weekends and holidays.
9. **Loyalty Programs:** Vendors may consider loyalty programs where customers who frequently use their taxi services earn points or special discounts. This will encourage customers to remain loyal to the taxi service.



# Anomaly in data





## Total amount data does not match the time difference between pickup and drop off

From top 10 the biggest time differences in this table, spending 24 hours only for 1 trip, and also the trip distance and amount obtained is very little not in accordance with the time spent.

This proves that the driver did not set the taxi meter when dropping off passengers.

	pickup_datetime	dropoff_datetime	trip_distance	fare_amount	total_amount	selisih_waktu_menit
21350	2022-01-25 07:50:00	2022-01-26 07:49:55	0.00	0.0	0.00	1439.916667
34634	2022-03-21 17:34:53	2022-03-22 17:33:52	0.85	5.5	12.25	1438.983333
23670	2022-02-11 10:45:11	2022-02-12 10:43:40	2.43	11.0	17.16	1438.483333
32377	2022-04-28 17:33:38	2022-04-29 17:31:58	2.46	14.5	18.80	1438.333333
45867	2022-06-23 19:20:54	2022-06-24 19:19:10	2.25	10.5	17.76	1438.266667
49446	2022-11-12 15:52:54	2022-11-13 15:50:03	1.70	10.5	15.18	1437.150000
31124	2022-10-29 18:25:17	2022-10-30 18:21:55	3.43	14.5	17.80	1436.633333
30550	2022-09-10 03:28:23	2022-09-11 03:24:47	2.87	12.0	15.80	1436.400000
20782	2022-10-18 13:28:26	2022-10-19 13:24:30	16.32	52.0	55.30	1436.066667
33774	2022-09-13 17:19:25	2022-09-14 17:14:52	5.02	26.0	30.30	1435.450000





## Records data with trip distance = 0 and total amount = 0

In this condition, some drivers immediately set a drop-off time when a passenger cancels a taxi service so that drivers can immediately look for other passengers.

However, there are drivers who don't set the drop-off time for up to 24 hours, which makes two taxis actually available and can carry other passengers, but because they don't set the drop-off time, the taxis are still in a state of running the previous trip.

	trip_distance	total_amount	selisih_waktu_minit
15215	0.0	0.0	0.300000
16180	0.0	0.0	2.683333
16672	0.0	0.0	0.000000
19034	0.0	0.0	0.166667
19707	0.0	0.0	0.066667
21350	0.0	0.0	1439.916667
22902	0.0	0.0	0.300000
23379	0.0	0.0	0.000000
34193	0.0	0.0	1432.283333
38423	0.0	0.0	0.066667
47399	0.0	0.0	0.250000
54029	0.0	0.0	0.600000







What is the fare amount if the taxi does not set the pickup time and dropoff time correctly?

57 times the driver didn't set the meter when dropping off passengers.

with a trip time of more than 200 minutes, the fare amount that should be obtained will be more than 1105 dollars

```
Total data with minutes difference > 200 and fare amount < 100: 57
```

```
Total fare amount for data with a difference of minutes > 200 and fare amount < 100: 1105.40
```





## See time difference based on trip distance and total amount

From the table with the 10 highest fare amounts. With a large amount it will be directly proportional to the distance on a trip. In the table there are also 3 data with trip distance = 0.

This proves that the sensor that records the taxi trip needs to be repaired to find out how far the trip is and what fare amount should be generated from the trip distance.

	pickup_datetime	dropoff_datetime	trip_distance	fare_amount	total_amount	selisih_waktu_menit
30114	2022-07-22 07:01:20	2022-07-22 09:04:15	113.56	385.0	420.60	122.916667
41736	2022-04-08 00:47:44	2022-04-08 02:02:13	48.41	300.0	322.14	74.483333
53860	2022-03-25 14:22:36	2022-03-25 18:23:15	28.43	288.5	289.30	240.650000
42881	2022-06-21 22:58:10	2022-06-21 22:58:24	0.00	250.0	300.96	0.233333
34342	2022-08-25 20:23:34	2022-08-25 20:23:45	0.00	240.0	270.30	0.183333
14258	2022-01-23 08:14:26	2022-01-23 08:14:39	0.00	225.0	225.30	0.216667
6320	2022-07-20 20:14:52	2022-07-20 21:05:32	37.40	222.0	250.10	50.666667
11385	2022-09-10 14:14:08	2022-09-10 14:14:12	0.00	217.0	240.30	0.066667
29060	2022-07-28 07:27:37	2022-07-28 10:06:26	79.20	215.5	224.10	158.816667
30318	2022-02-25 12:07:38	2022-02-25 13:14:05	47.56	214.5	216.55	66.450000





## Which vendor makes a lot of mistakes not setting drop off time?

A trip with a time of more than 200 minutes only resulted in a total amount of 1841.74 dollars for Verifone Inc. Proving that this vendor needs to provide directions and return notifications to their drivers regarding pickup and drop times.

-off

vendor_id	selisih_hari	selisih_waktu_menit	total_amount
Creative Mobile Techonologies	1	556.933333	23.00
VeriFone Inc.	58	75587.033333	1841.74

The difference of 28 days shows that there were 28 trips that took more than 200 minutes but only got a total of less than 20 dollars on each trip.



With this time span, drivers should pick up more passengers if they set the drop off time correctly and can get a higher amount.

-

vendor_id	selisih_hari	selisih_waktu_menit
VeriFone Inc.	28	37445.766667

## How much data with trip distance = 0 and what is the total amount obtained in the trip?

There were 597 trips whose mileage was not recorded, but they received a total of 26605.26 dollars for zero trip distance.

In this case, the vendor cannot know how far the 597 trips have gone.

And make vendors unable to monitor the condition of their taxis so that they remain in good condition.

```
Total Trip Distance is 0: 597  
Total Amount when Trip Distance is 0: 26605.260000000002
```

## How many trips with distance = 0 are not connected to the server (store and forward trip)?

Most of these zero -trip distances do not have a connection to the server to keep records of their trips. There are only 16 trips connected to the server, even though the trip distance is 0.

```
store_and_fwd_flag  
N      581  
Y       16  
Name: trip_distance, dtype: int64
```



## How many trips with a travel time of less than 1 minute?

There are 566 trips that only take less than 1 minute; this needs to be further investigated, whether the distance is close or the passengers cancel their trips.

```
Total data with time difference <= 1 minute: 566
```

## What is the total amount generated? and how much for each store\_and\_fwd\_flag ?

With a trip that is less than 1 minute, the amount still comes in at as much as 26477.96 dollars, but most of the trips are still not connected to the server to record their travel records.

```
Total Amount: 26477.96
The total_amount for each store_and_fwd_flag:
store_and_fwd_flag
N      26334.90
Y       143.06
```

## How many trips with a travel time of less than 1 minute?

The average distance in less than 1 minute is 0.3 miles, with an average fare amount of 39.03 dollars. A very large amount of fare for a trip that is less than 1 minute.

```
average trip_distance: 0.36528268551236737
average fare_amount: 39.03441696113074
```





## Records data with trip distance = 0 and total amount = 0

In this condition, some drivers immediately set a drop-off time when a passenger cancels a taxi service so that drivers can immediately look for other passengers.

However, there are drivers who don't set the drop-off time for up to 24 hours, which makes two taxis actually available and can carry other passengers, but because they don't set the drop-off time, the taxis are still in a state of running the previous trip.

	trip_distance	total_amount	selisih_waktu_minit
15215	0.0	0.0	0.300000
16180	0.0	0.0	2.683333
16672	0.0	0.0	0.000000
19034	0.0	0.0	0.166667
19707	0.0	0.0	0.066667
21350	0.0	0.0	1439.916667
22902	0.0	0.0	0.300000
23379	0.0	0.0	0.000000
34193	0.0	0.0	1432.283333
38423	0.0	0.0	0.066667
47399	0.0	0.0	0.250000
54029	0.0	0.0	0.600000



# Recommendation to avoid this mistake and error data



- 1) **Fix the sensor** to detect the distance traveled on a trip **to prevent zero trip distance**.
- 2) **Connecting taxis** that are **not yet connected to the server** so that they can **keep driving records** so that the data obtained will be more **accurate and appropriate**.
- 3) **Giving announcements and notices** to drivers who often **forget to set the drop-off time** so that the time difference is not far **away** and drivers can **look for other passengers**.
- 4) **Provide notifications to drivers** who travel for **more than 5 hours**, with the choice of whether they are **still on a trip** or **have forgotten to set the drop-off time**.
- 5) **Refund passengers** who do not end up using the taxi service **if they cannot cancel the service** and **their money remains deducted** from their credit card.





03

# MODELLING PROCESS



# Modelling Process Outline

01

Data  
Preprocessing

Data Cleaning,  
Dtype Adjustment,  
Feature engineering

02

Stationary  
Test

Stationary test  
using  
Augmented  
Dickey-Fuller

03

Modelling  
Process

Create Model  
using xgboost  
Regressor

04

Predict with  
New Datatest

Predict  
passenger  
count with new  
data test

# DATA PREPROCESSING

Some process we do:

## Data Cleaning:

- Handling Missing Value, Duplicated Value,

## Change Data Type

- Change pickup\_datetime dtype from object to datetime64[ns]

## Feature Engineering

- Split pickup\_datetime to several feature such as, hour, month, year, quarter, day of week

# STATIONARY TEST

Based on the results of the Dickey-Fuller test given, the conclusion that can be drawn is that the p-value is 0.000007.

A small p-value (smaller than the specified significance level, for example 0.05) indicates that the null hypothesis can be rejected, so the data can be considered stationary.

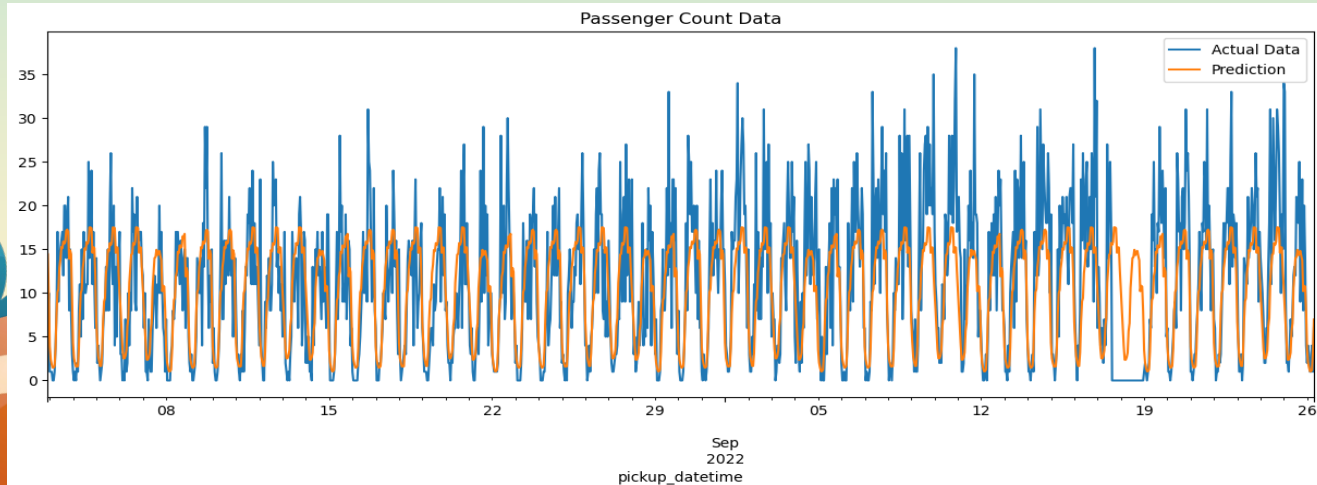
```
Observations of Dickey-fuller test
Test Statistic          -5.234426
p-value                  0.000007
#lags used               33.000000
number of observations used 6272.000000
critical value (1%)      -3.431393
critical value (5%)      -2.862001
critical value (10%)     -2.567015
dtype: float64
```

# Modelling Result

The algorithm to be used is the XGBoost Regressor, using parameters such as `n_estimators=1000`, `booster='gbtree'`, `max_depth=4`, etc.

From the test data tested, this model produces an RMSE score of 5.88.

The following is a graphic display of the model results on the test data.





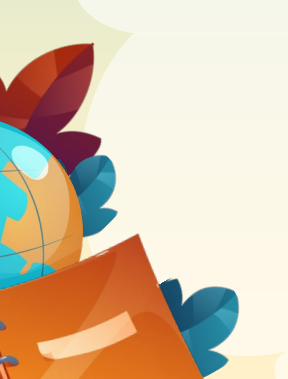
# Predict with New Data Test

Before predicting the number of passengers on the test data every hour. Data Test must go through the same preprocessing as was done on the data train

After the data is clean, then the number of passengers will be predicted.

This is first 5 records prediction result stored in the csv file

	index	pickup_datetime	passenger_count
0	0	2022-09-26 08:00:00+00:00	9.177842
1	1	2022-09-26 09:00:00+00:00	9.463750
2	2	2022-09-26 10:00:00+00:00	11.904019
3	3	2022-09-26 11:00:00+00:00	12.468629
4	4	2022-09-26 12:00:00+00:00	14.721057



04

# CONCLUSION





# Conclusion & Recommendation



- Consistent increase in the number of passengers over the 65-day period, it is advisable for taxi service providers to ensure sufficient vehicle availability to accommodate the increased demand. This will help avoid situations where passengers have to wait a long time or have trouble getting a taxi.
- It is important for taxi companies to have an efficient and coordinated operational system. Good coordination will help ensure smooth operations during periods of spikes in demand.
- After doing passenger forecasting and calculating the current capacity based on the number of cars owned and their ability to serve demand every hour, Then compare the results of the predicted demand for taxis with the capacity of the taxis owned. If demand consistently exceeds current capacity at certain hours, this can be taken as an indicator that capacity caps will be reached at those times.

