Maximizing Revenue for Drivers Through Payment Type

A Statistical Analysis of NYC Taxi Data

....

Vedant Thorat

Email – vedant2000thorat@gmail.com

Table of content

Introduction

Background

The taxi industry in New York City (NYC) is a dynamic and essential part of urban transportation, serving millions of residents and tourists each year. With the advent of technology and the proliferation of various payment methods, understanding how these methods impact revenue has become crucial for taxi drivers and operators. In this context, leveraging data-driven insights to maximize revenue streams is vital for sustaining long-term success and ensuring driver satisfaction.

Purpose of the Study

This study aims to explore the relationship between payment methods and fare amounts in NYC's yellow taxi services. Specifically, it investigates whether the method of payment—such as credit cards or cash—has a significant impact on the total fare amount charged to passengers. By understanding these dynamics, we can provide actionable insights to help taxi drivers optimize their revenue.

Importance of the Study

In an industry where every fare contributes to a driver's livelihood, identifying factors that influence fare amounts is paramount. Payment methods could potentially affect fare amounts due to various reasons, including transaction fees, passenger behavior, and tip inclusion. By analyzing these factors, this study seeks to offer strategies for encouraging payment methods that maximize driver revenue without compromising customer satisfaction. This research not only contributes to academic knowledge but also provides practical recommendations for the taxi industry.

Structure of the Report

The report is structured as follows: The **Problem Statement** outlines the central issue and objectives of the study. The **Data Description** section describes the dataset and variables used in the analysis. The **Methodology** section details the statistical techniques employed, including descriptive analysis, hypothesis testing, and regression analysis. The **Results** section presents the findings, followed by a discussion of their implications in the **Implications** section. The **Limitations** of the study are then addressed, leading to the **Conclusion** which summarizes the key insights. Finally, suggestions for **Future Work** are provided to guide subsequent research.

Problem Statement

In the fast-paced taxi booking sector, maximizing revenue is essential for long-term success and driver satisfaction. Taxi drivers rely heavily on fare amounts to sustain their livelihood, making it crucial to understand the factors that influence these amounts. One such factor is the payment method used by passengers, which could potentially impact the fare amount due to varying transaction fees, tipping behaviors, and other related factors.

Our goal is to leverage data-driven insights to maximize revenue streams for taxi drivers. Specifically, we aim to determine whether different payment methods—such as credit cards, cash, and others—have a significant impact on fare pricing. By focusing on the relationship between payment type and fare amount, we seek to identify payment methods that are most beneficial for drivers.

Research Questions

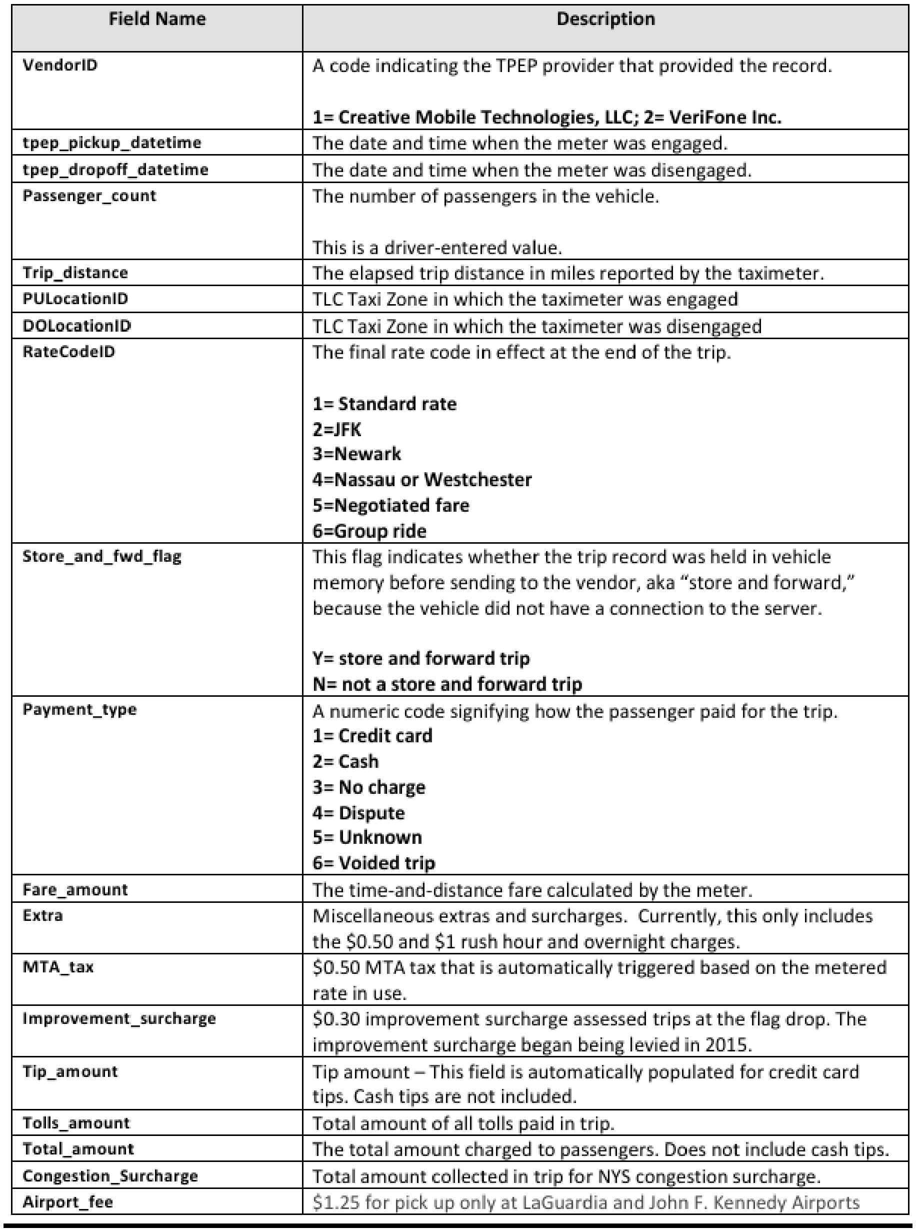
1. Is there a relationship between total fare amount and payment type?
2. Can we nudge customers towards payment methods that generate higher revenue for drivers, without negatively impacting the customer experience?

By addressing these questions, our research aims to provide actionable insights that can help taxi drivers optimize their revenue, while maintaining a positive experience for passengers.

Data Description

Source of Data

The dataset for this study is sourced from the New York City Taxi and Limousine Commission (TLC). It includes detailed records of yellow taxi trips in NYC, capturing a wide range of variables related to each trip. This dataset is publicly available and can be accessed through the NYC government's website.

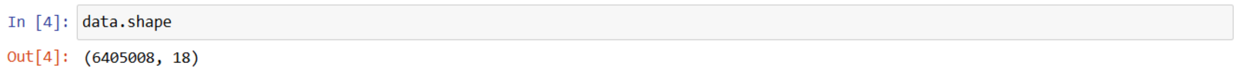


Field Name	Description
VendorID	A code indicating the TPEP provider that provided the record. **1= Creative Mobile Technologies, LLC; 2= VeriFone Inc.**
tpep_pickup_datetime	The date and time when the meter was engaged.
tpep_dropoff_datetime	The date and time when the meter was disengaged.
Passenger_count	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.
Congestion_Surcharge	Total amount collected in trip for NYS congestion surcharge.
Airport_fee	$1.25 for pick up only at LaGuardia and John F. Kennedy Airports

Data Cleaning, Filtering and Preparation

➢ Shape of the Data

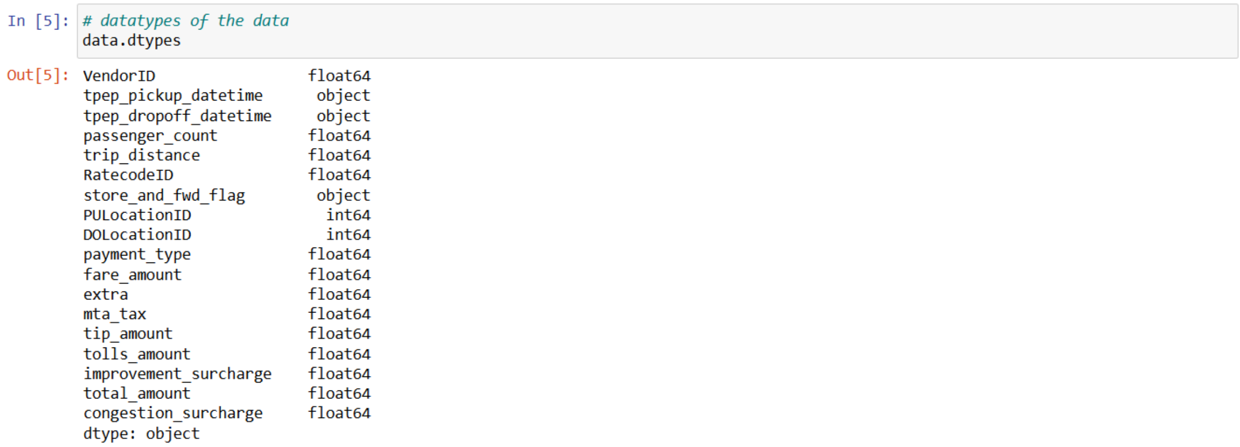
The initial dataset comprises 6,405,008 rows and 18 columns, indicating a comprehensive collection of trip records:



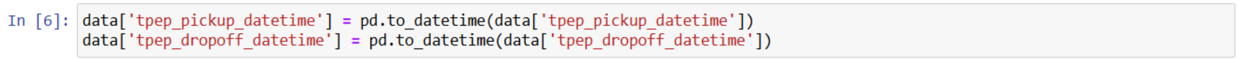
```
In [4]: data.shape
Out[4]: (6405008, 18)
```

➢ Checking Data Types

An inspection of the data types reveals a mix of numerical and object types. The two datetime columns, tpep_pickup_datetime and tpep_dropoff_datetime, are currently stored as objects and will need conversion to datetime types for accurate time-based calculations:



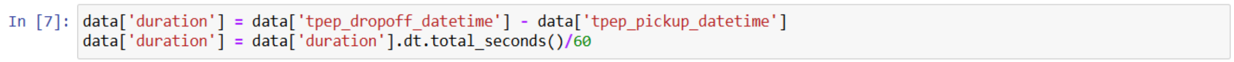
```
In [5]: # datatypes of the data
        data.dtypes
Out[5]: VendorID                 float64
        tpep_pickup_datetime      object
        tpep_dropoff_datetime     object
        passenger_count          float64
        trip_distance            float64
        RatecodeID               float64
        store_and_fwd_flag        object
        PULocationID               int64
        DOLocationID               int64
        payment_type             float64
        fare_amount              float64
        extra                    float64
        mta_tax                  float64
        tip_amount               float64
        tolls_amount             float64
        improvement_surcharge    float64
        total_amount             float64
        congestion_surcharge     float64
        dtype: object
```



```
In [6]: data['tpep_pickup_datetime'] = pd.to_datetime(data['tpep_pickup_datetime'])
        data['tpep_dropoff_datetime'] = pd.to_datetime(data['tpep_dropoff_datetime'])
```

➢ Calculating Duration of the Ride

We calculate the duration of each ride by subtracting the pickup_datetime from the dropoff_datetime and converting the result to minutes.

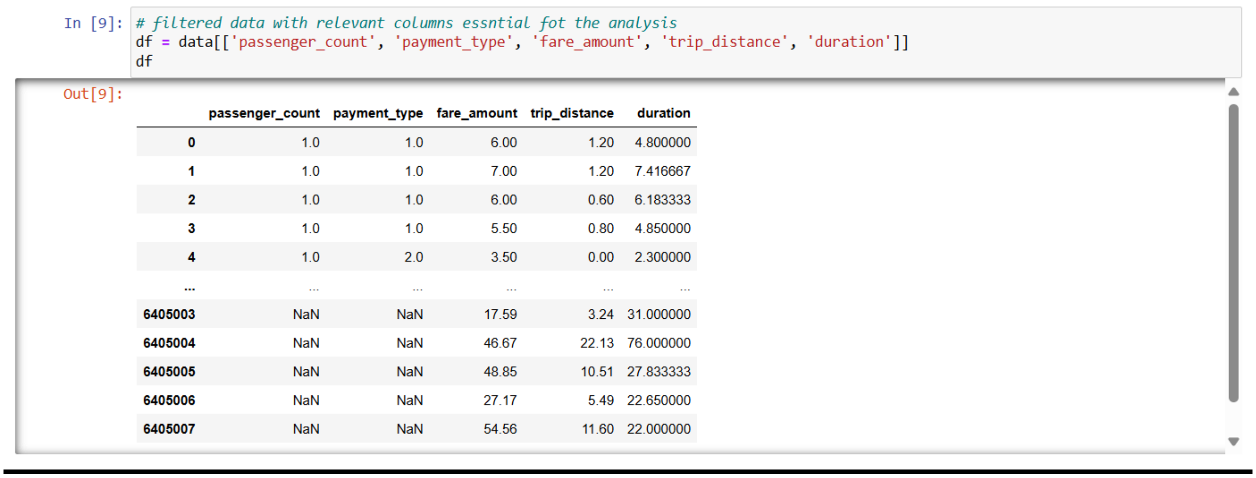


```
In [7]: data['duration'] = data['tpep_dropoff_datetime'] - data['tpep_pickup_datetime']
        data['duration'] = data['duration'].dt.total_seconds()/60
```

➢ Column Selection

Given the problem statement focuses on the relationship between payment type and fare amount, along with any other influencing factors, the following columns are essential for the analysis. The remaining columns, which are not directly related to the core research questions, are removed to streamline the analysis:

- passenger_count
- trip_distance
- payment_type
- fare_amount
- duration



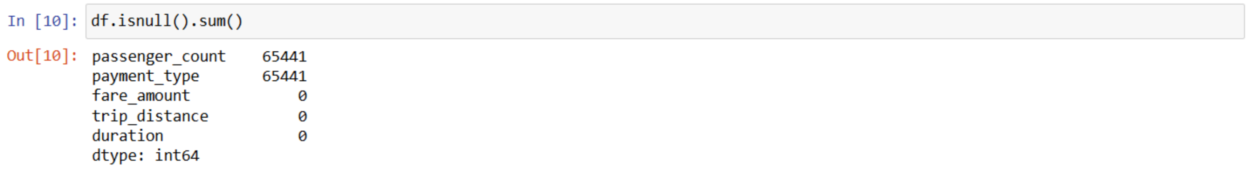
```
In [9]: # filtered data with relevant columns essntial fot the analysis
        df = data[['passenger_count', 'payment_type', 'fare_amount', 'trip_distance', 'duration']]
        df
```

Out[9]:

	passenger_count	payment_type	fare_amount	trip_distance	duration
0	1.0	1.0	6.00	1.20	4.800000
1	1.0	1.0	7.00	1.20	7.416667
2	1.0	1.0	6.00	0.60	6.183333
3	1.0	1.0	5.50	0.80	4.850000
4	1.0	2.0	3.50	0.00	2.300000
...	...	...	...	...	...
6405003	NaN	NaN	17.59	3.24	31.000000
6405004	NaN	NaN	46.67	22.13	76.000000
6405005	NaN	NaN	48.85	10.51	27.833333
6405006	NaN	NaN	27.17	5.49	22.650000
6405007	NaN	NaN	54.56	11.60	22.000000

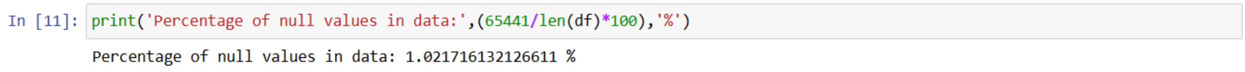
➢ Handling Missing Values

Checking Null Values: Upon checking for null values in the dataset, we found the following counts:



```
In [10]: df.isnull().sum()

Out[10]: passenger_count    65441
         payment_type       65441
         fare_amount            0
         trip_distance          0
         duration               0
         dtype: int64
```

Percentage of Null Values: The columns passenger_count and payment_type each have 65,441 null values, which amounts to approximately 1.02% of the total dataset:



```
In [11]: print('Percentage of null values in data:',(65441/len(df)*100),'%')

         Percentage of null values in data: 1.021716132126611 %
```

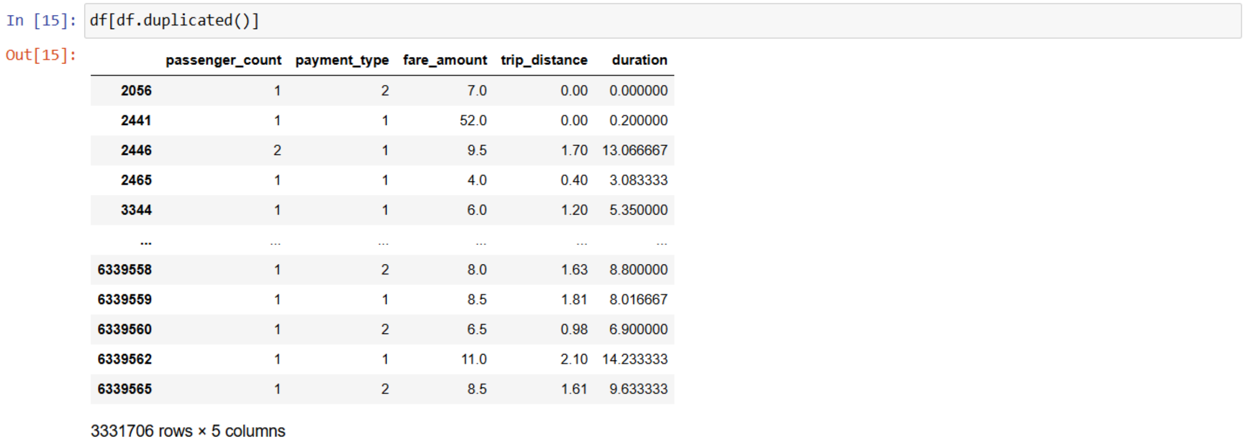
Handling Null Values: Given that the percentage of missing data is relatively small (1%), we opted to remove records with missing values to maintain data integrity:

```
In [12]: df.dropna(inplace=True)
```

By removing records with missing values, we ensure that the dataset is ready for further analysis without compromising the quality and reliability of the results. This step is crucial for maintaining the accuracy of statistical analyses and modeling techniques applied to the data.

➢ Checking for Duplicated Values

Upon checking for duplicate rows in the dataset, we identified and removed them to ensure each record is unique:



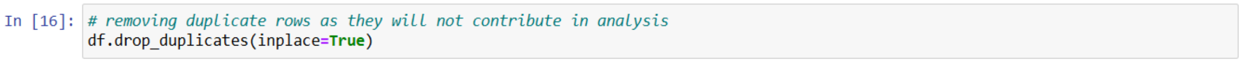
```
In [15]: df[df.duplicated()]
```

Out[15]:

	passenger_count	payment_type	fare_amount	trip_distance	duration
2056	1	2	7.0	0.00	0.000000
2441	1	1	52.0	0.00	0.200000
2446	2	1	9.5	1.70	13.066667
2465	1	1	4.0	0.40	3.083333
3344	1	1	6.0	1.20	5.350000
...	...	...	...	...	...
6339558	1	2	8.0	1.63	8.800000
6339559	1	1	8.5	1.81	8.016667
6339560	1	2	6.5	0.98	6.900000
6339562	1	1	11.0	2.10	14.233333
6339565	1	2	8.5	1.61	9.633333

3331706 rows × 5 columns

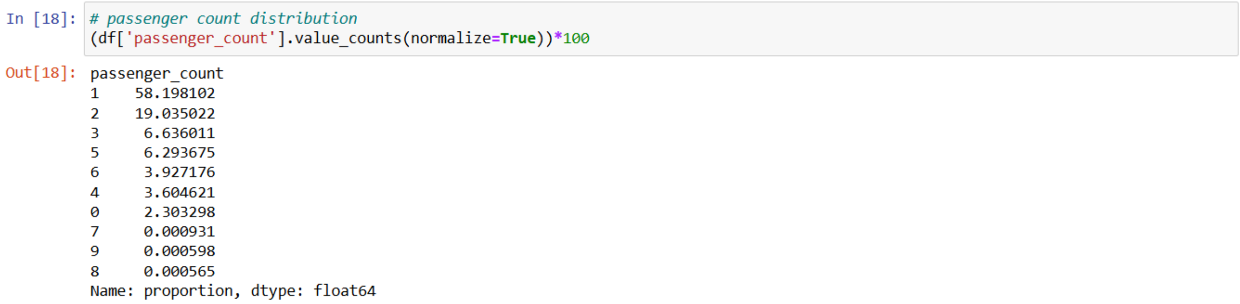
Removing Duplicate Rows: After identifying duplicate rows, they were removed from the dataset:



```
In [16]: # removing duplicate rows as they will not contribute in analysis
         df.drop_duplicates(inplace=True)
```

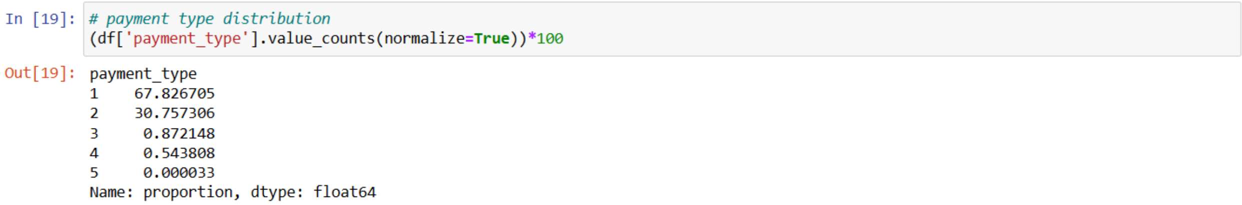
➢ Checking Distribution

Passenger Count Distribution: The distribution of passenger counts in the dataset is as follows:



```
In [18]: # passenger count distribution
         (df['passenger_count'].value_counts(normalize=True))*100

Out[18]: passenger_count
         1    58.198102
         2    19.035022
         3     6.636011
         5     6.293675
         6     3.927176
         4     3.604621
         0     2.303298
         7     0.000931
         9     0.000598
         8     0.000565
         Name: proportion, dtype: float64
```

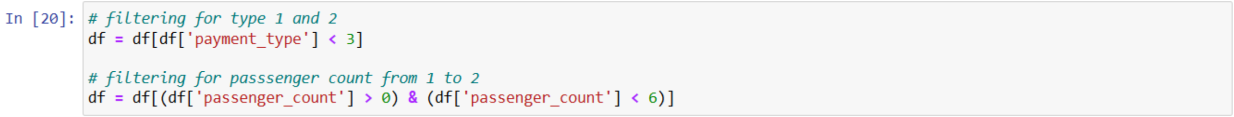
Payment Type Distribution: The distribution of payment types in the dataset is as follows**:**



```
In [19]: # payment type distribution
         (df['payment_type'].value_counts(normalize=True))*100

Out[19]: payment_type
         1    67.826705
         2    30.757306
         3     0.872148
         4     0.543808
         5     0.000033
         Name: proportion, dtype: float64
```

➢ Filtering Data

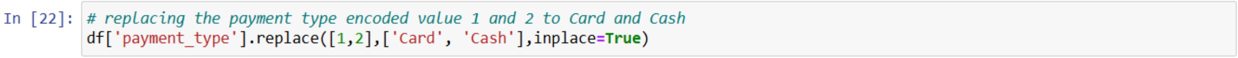
To focus the analysis on payment types 'Card' and 'Cash', and passenger counts ranging from 1 to 5, we filtered the dataset accordingly:



```
In [20]: # filtering for type 1 and 2
         df = df[df['payment_type'] < 3]

         # filtering for passsenger count from 1 to 2
         df = df[(df['passenger_count'] > 0) & (df['passenger_count'] < 6)]
```

➢ Replacing Payment Type Encoded Values

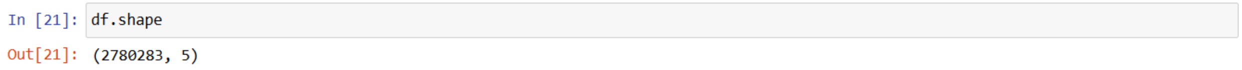
To enhance clarity and interpretation, we replaced the encoded payment types (1 and 2) with their corresponding labels ('Card' and 'Cash'):



```
In [22]: # replacing the payment type encoded value 1 and 2 to Card and Cash
         df['payment_type'].replace([1,2],['Card', 'Cash'],inplace=True)
```

➢ Updated Dataset Shape

Post cleaning and removing duplicates, the dataset now contains:



```
In [21]: df.shape
Out[21]: (2780283, 5)
```

By cleaning, filtering, and preparing the data in this manner, we ensure that the dataset is ready for in-depth analysis focused on the relationship between payment methods, fare amounts, and other influencing factors in NYC's yellow taxi services.

Methodology

➢ Descriptive Analysis

Descriptive analysis was conducted to provide a comprehensive overview of key aspects of the dataset. This included summarizing and visualizing the following variables:

- **Fare Amount**: Analyzed the distribution of fare amounts to understand the typical pricing structure.
- **Payment Types**: Examined the frequency and distribution of different payment methods (Card and Cash) used by passengers.
- **Trip Duration**: Calculated from pickup and drop-off times, explored the distribution and relationship with fare amounts.

➢ Hypothesis Testing

Hypothesis testing was employed to assess the significance of the relationship between payment type and fare amount. Specifically, a t-test was conducted to evaluate whether there is a statistically significant difference in fare amounts between trips paid with Card versus Cash. The null hypothesis tested was:

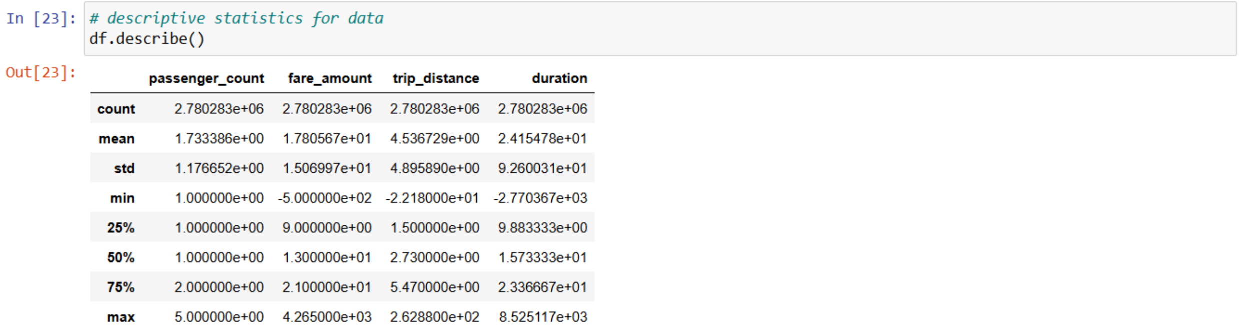
- **Null Hypothesis (H_0)**: There is no significant difference in fare amounts between payment methods (Card and Cash).
- **Alternative Hypothesis (H_1)**: There is a significant difference in fare amounts between payment methods (Card and Cash).

Descriptive Analysis

The descriptive analysis aims to provide a comprehensive overview of the dataset, highlighting key statistics and trends to inform subsequent hypothesis testing and regression analysis. By summarizing the main characteristics of the data, we can gain valuable insights into the distribution and relationships between variables, particularly focusing on fare amounts and payment types.

➢ Summary Statistics

To begin with, we calculate summary statistics for the primary variables of interest: passenger count, fare amount, trip distance, and trip duration. The following table provides a detailed statistical summary of these variables:



```
In [23]: # descriptive statistics for data
         df.describe()
```

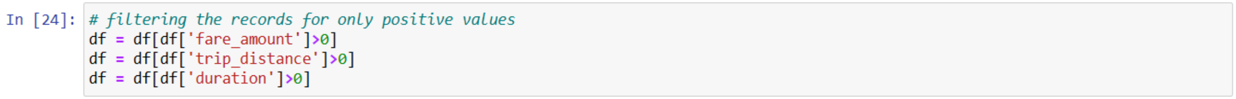
Out[23]:

	passenger_count	fare_amount	trip_distance	duration
count	2.780283e+06	2.780283e+06	2.780283e+06	2.780283e+06
mean	1.733386e+00	1.780567e+01	4.536729e+00	2.415478e+01
std	1.176652e+00	1.506997e+01	4.895890e+00	9.260031e+01
min	1.000000e+00	-5.000000e+02	-2.218000e+01	-2.770367e+03
25%	1.000000e+00	9.000000e+00	1.500000e+00	9.883333e+00
50%	1.000000e+00	1.300000e+01	2.730000e+00	1.573333e+01
75%	2.000000e+00	2.100000e+01	5.470000e+00	2.336667e+01
max	5.000000e+00	4.265000e+03	2.628800e+02	8.525117e+03

Data Cleaning and Preparation

➢ Handling Invalid Values

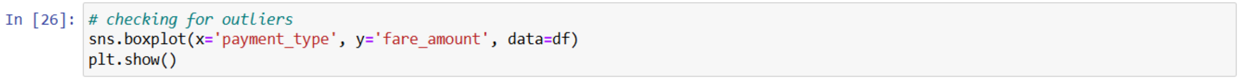
Upon reviewing the statistics, it is evident that the minimum values for trip distance, fare amount, and duration are negative, which is unrealistic and invalid for further analysis. Consequently, we eliminate these negative values from the dataset to ensure the accuracy and reliability of our analysis.



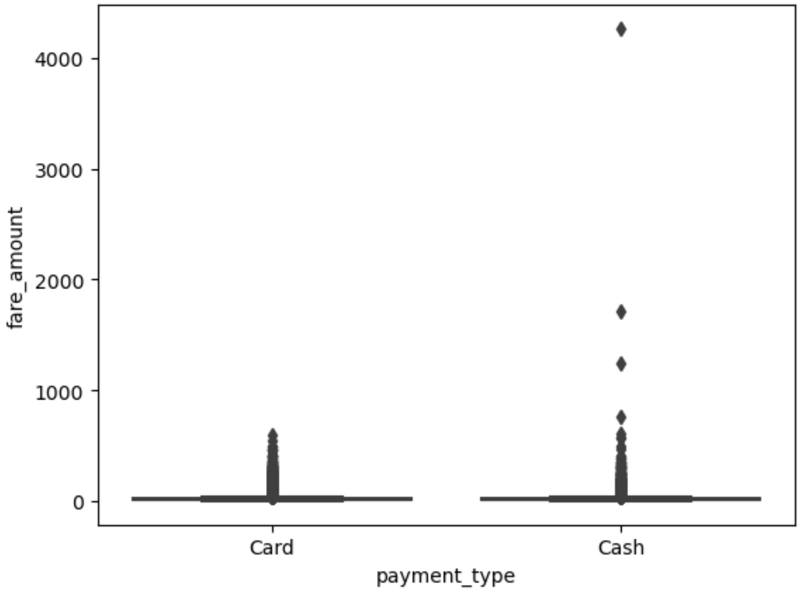
```
In [24]: # filtering the records for only positive values
         df = df[df['fare_amount']>0]
         df = df[df['trip_distance']>0]
         df = df[df['duration']>0]
```

➢ Handling Outliers

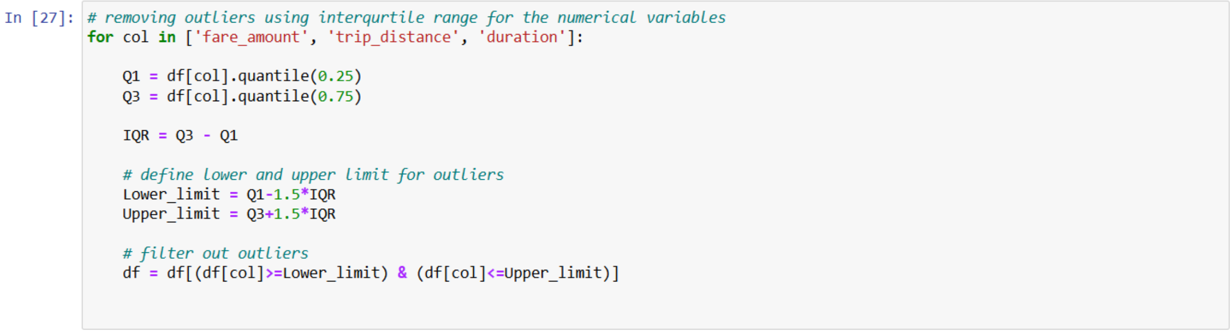
Outliers, particularly high values, were identified and treated using the Interquartile Range (IQR) method across numerical variables (fare_amount, trip_distance, duration). Outliers were defined as values lying outside 1.5 times the IQR below the first quartile (Q1) or above the third quartile (Q3).



```
In [26]: # checking for outliers
         sns.boxplot(x='payment_type', y='fare_amount', data=df)
         plt.show()
```



Additionally, observing the maximum and 50th percentile values, it is possible that the data contains significant outliers, particularly high values. These outliers need to be addressed and removed to ensure the integrity of the analysis. Outliers are identified and removed using the interquartile range (IQR) method.



```
In [27]: # removing outliers using interqurtile range for the numerical variables
         for col in ['fare_amount', 'trip_distance', 'duration']:

             Q1 = df[col].quantile(0.25)
             Q3 = df[col].quantile(0.75)

             IQR = Q3 - Q1

             # define lower and upper limit for outliers
             Lower_limit = Q1-1.5*IQR
             Upper_limit = Q3+1.5*IQR

             # filter out outliers
             df = df[(df[col]>=Lower_limit) & (df[col]<=Upper_limit)]
```

➢ Visualizing Data Distribution

Fare Amount by Payment Type

To explore the distribution of fare amounts across different payment types (Card and Cash), histograms were plotted

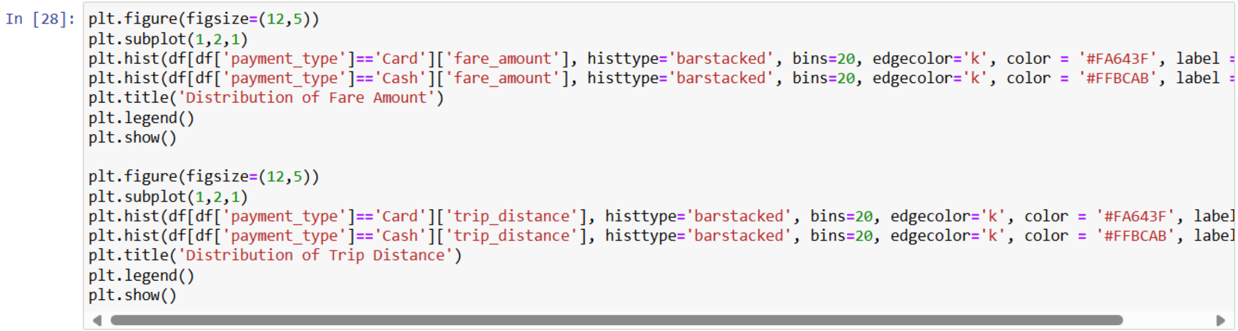
This visualization allows stakeholders to compare the distribution of fare amounts paid via Card versus Cash, providing insights into potential differences in spending patterns between payment methods.

Trip Distance by Payment Type

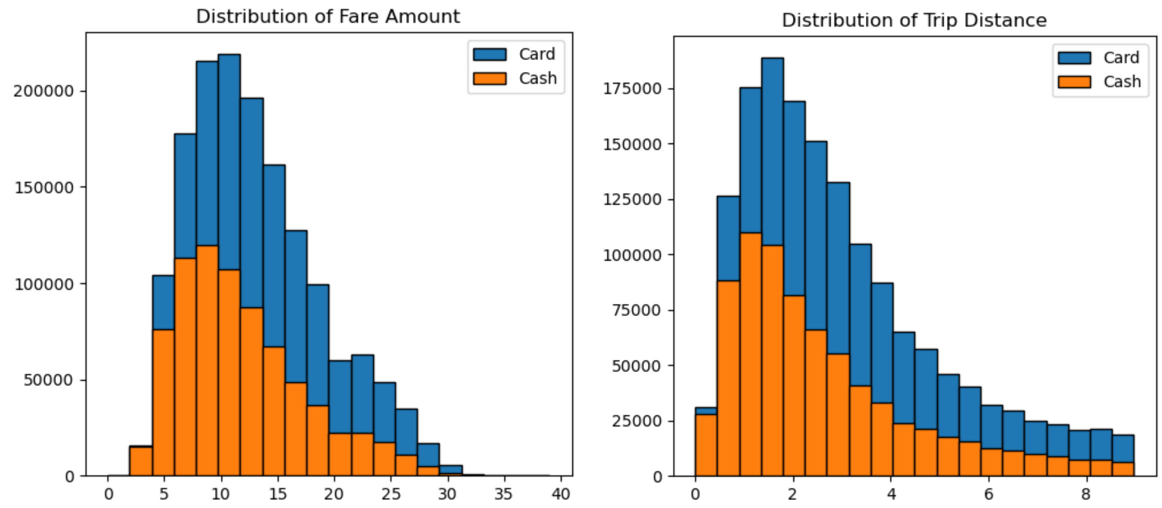
Similarly, the distribution of trip distances for Card and Cash payments was visualized using histograms:

This analysis provides a visual comparison of trip distances associated with different payment methods, aiding in understanding travel behavior and preferences among passengers.

To better understand the distribution of fare amounts and trip distances based on payment type, we plot histograms for both card and cash payments. This visualization helps to identify any noticeable patterns or differences between the payment methods.



```
In [28]: plt.figure(figsize=(12,5))
         plt.subplot(1,2,1)
         plt.hist(df[df['payment_type']=='Card']['fare_amount'], histtype='barstacked', bins=20, edgecolor='k', color = '#FA643F', label =
         plt.hist(df[df['payment_type']=='Cash']['fare_amount'], histtype='barstacked', bins=20, edgecolor='k', color = '#FFBCAB', label =
         plt.title('Distribution of Fare Amount')
         plt.legend()
         plt.show()

         plt.figure(figsize=(12,5))
         plt.subplot(1,2,1)
         plt.hist(df[df['payment_type']=='Card']['trip_distance'], histtype='barstacked', bins=20, edgecolor='k', color = '#FA643F', label
         plt.hist(df[df['payment_type']=='Cash']['trip_distance'], histtype='barstacked', bins=20, edgecolor='k', color = '#FFBCAB', label
         plt.title('Distribution of Trip Distance')
         plt.legend()
         plt.show()
```



➢ Insights

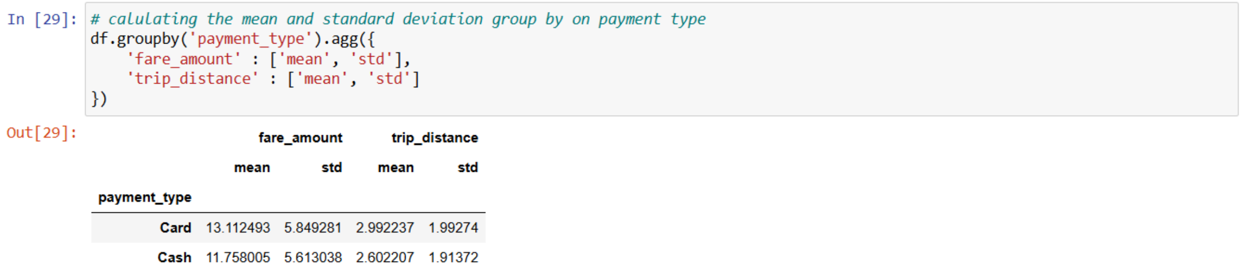
Fare Amount: The distribution indicates that fares paid with Card tend to cover a broader range compared to Cash payments, suggesting potentially varied spending behaviors.

Trip Distance: There appears to be a similar trend in trip distances, with a wider range observed for Card payments compared to Cash, which may reflect different travel preferences or trip purposes among passengers.

By conducting this descriptive analysis, we lay the foundation for deeper insights into the relationship between payment types and passenger behaviors in NYC's yellow taxi services, facilitating informed decision-making for stakeholders aiming to optimize revenue and improve customer service.

➢ Mean and Standard Deviation by Payment Type

To further explore the relationship between payment types and key variables, we calculate the mean and standard deviation of fare amounts and trip distances grouped by payment type.



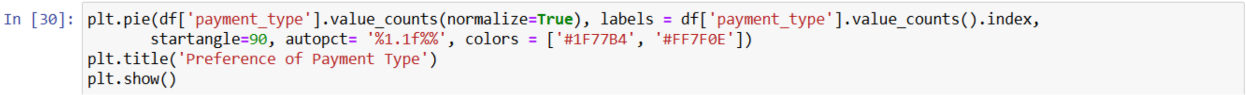
```
In [29]: # calulating the mean and standard deviation group by on payment type
         df.groupby('payment_type').agg({
             'fare_amount' : ['mean', 'std'],
             'trip_distance' : ['mean', 'std']
         })
```

```
Out[29]:
                      fare_amount          trip_distance
                      mean       std       mean      std
         payment_type
                 Card 13.112493  5.849281  2.992237  1.99274
                 Cash 11.758005  5.613038  2.602207  1.91372
```

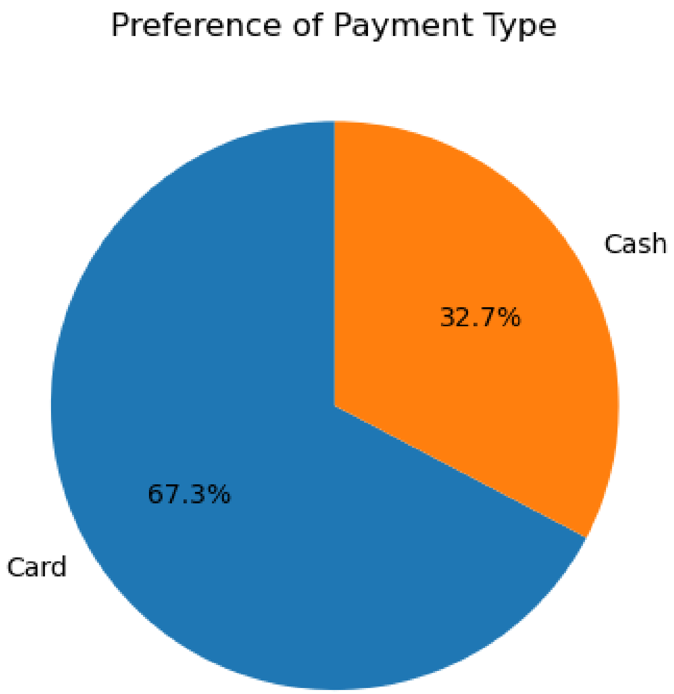
This analysis reveals that on average, fares paid with Card are slightly higher than those paid with Cash. Similarly, trips associated with Card payments tend to cover longer distances compared to Cash payments.

➢ Preference of Payment Type

To assess the proportion of payment types, we generate a pie chart that provides a visual representation of the distribution between card and cash payments. This graphical depiction offers a clear and intuitive understanding of passengers' payment preferences.



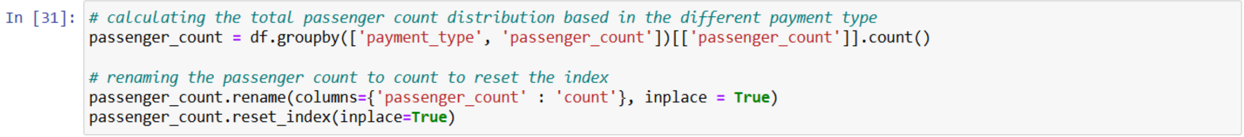
```
In [30]: plt.pie(df['payment_type'].value_counts(normalize=True), labels = df['payment_type'].value_counts().index,
                startangle=90, autopct= '%1.1f%%', colors = ['#1F77B4', '#FF7F0E'])
         plt.title('Preference of Payment Type')
         plt.show()
```



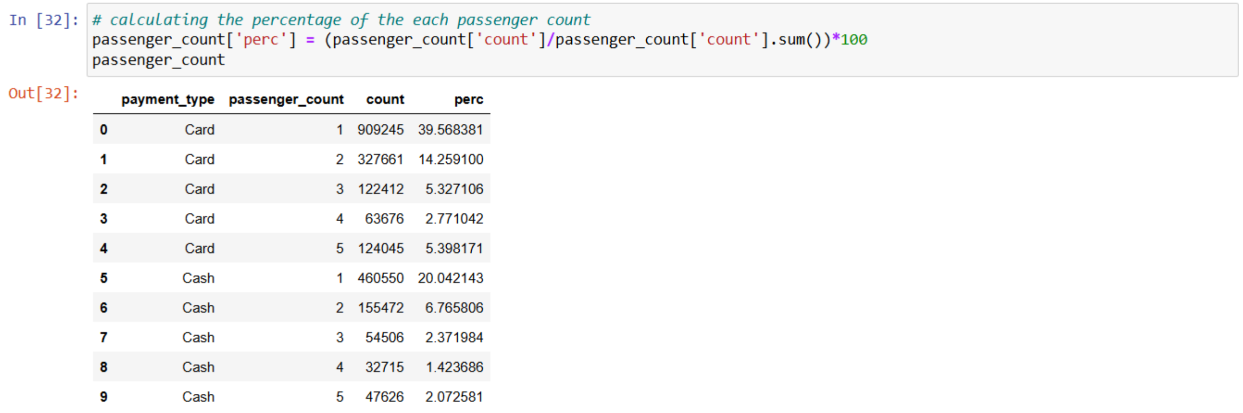
This chart illustrates that a significant majority of passengers prefer to pay using Card rather than Cash.

➢ Payment Type Distribution by Passenger Count

Next, we analyze the distribution of passenger counts for different payment types using a stacked bar plot. This method is particularly advantageous for comparing the percentage distribution of each passenger count based on the payment method selected, providing insights into potential variations in payment preferences across different passenger counts.



```
In [31]: # calculating the total passenger count distribution based in the different payment type
         passenger_count = df.groupby(['payment_type', 'passenger_count'])[['passenger_count']].count()

         # renaming the passenger count to count to reset the index
         passenger_count.rename(columns={'passenger_count' : 'count'}, inplace = True)
         passenger_count.reset_index(inplace=True)
```



In [32]:

```
# calculating the percentage of the each passenger count
passenger_count['perc'] = (passenger_count['count']/passenger_count['count'].sum())*100
passenger_count
```

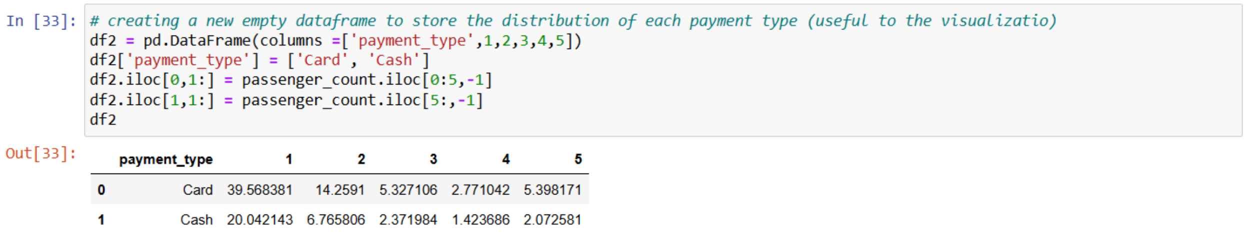
Out[32]:

	payment_type	passenger_count	count	perc
0	Card	1	909245	39.568381
1	Card	2	327661	14.259100
2	Card	3	122412	5.327106
3	Card	4	63676	2.771042
4	Card	5	124045	5.398171
5	Cash	1	460550	20.042143
6	Cash	2	155472	6.765806
7	Cash	3	54506	2.371984
8	Cash	4	32715	1.423686
9	Cash	5	47626	2.072581

This table breaks down the percentage distribution of each passenger count for Card and Cash payments. It shows that while single passengers predominantly use Card, the distribution between Card and Cash becomes more balanced as the number of passengers increases.

➢ Payment Type Distribution Visualization

To visually compare the distribution of payment types across different passenger counts, a stacked bar plot was created:

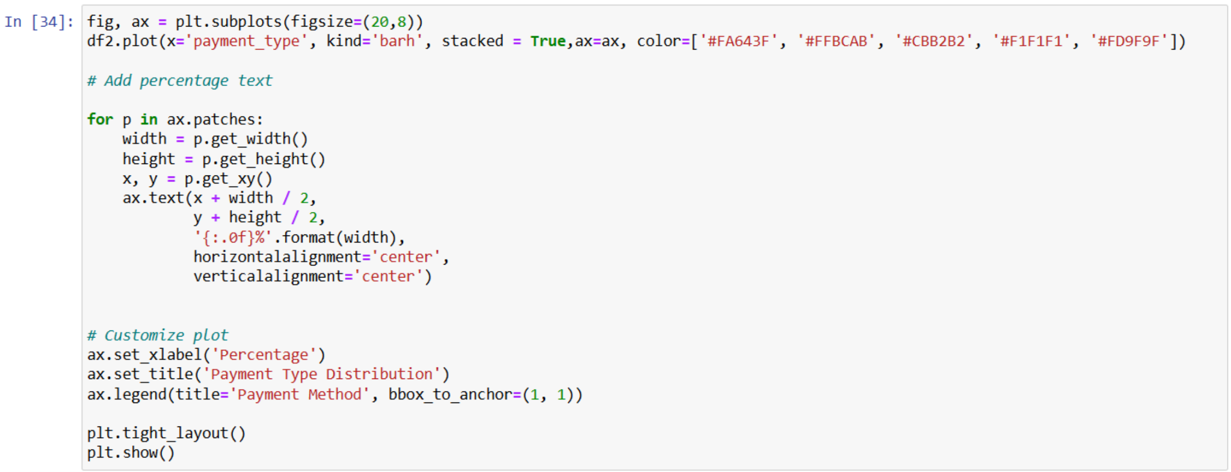


In [33]:

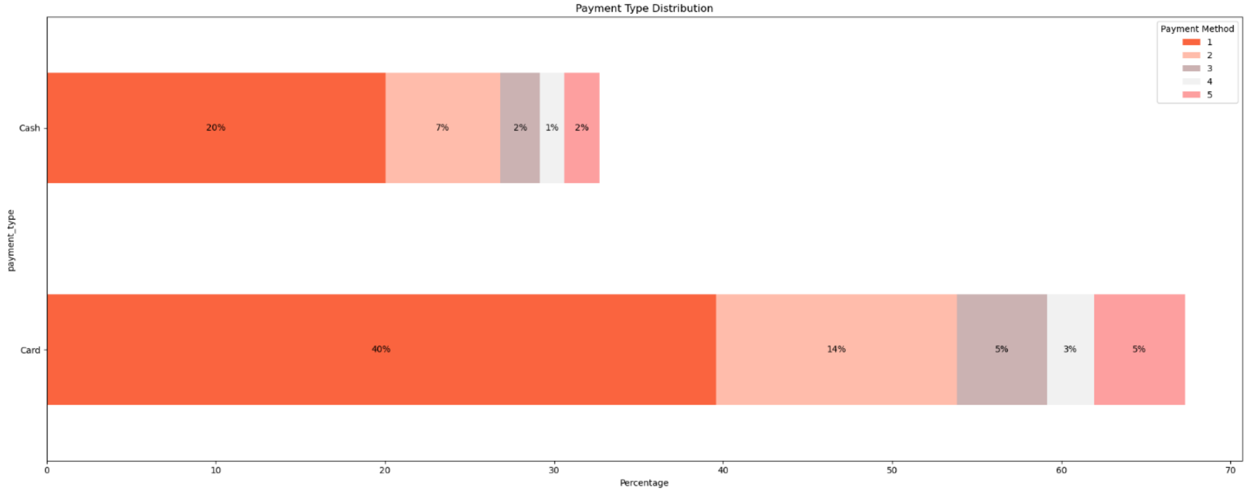
```
# creating a new empty dataframe to store the distribution of each payment type (useful to the visualizatio)
df2 = pd.DataFrame(columns =['payment_type',1,2,3,4,5])
df2['payment_type'] = ['Card', 'Cash']
df2.iloc[0,1:] = passenger_count.iloc[0:5,-1]
df2.iloc[1,1:] = passenger_count.iloc[5:,-1]
df2
```

Out[33]:

	payment_type	1	2	3	4	5
0	Card	39.568381	14.2591	5.327106	2.771042	5.398171
1	Cash	20.042143	6.765806	2.371984	1.423686	2.072581



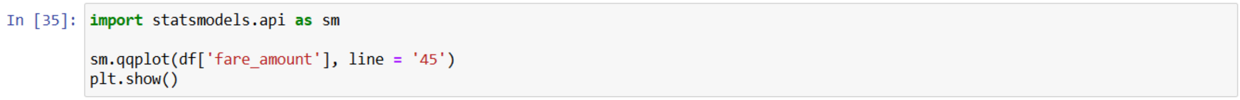
```
In [34]: fig, ax = plt.subplots(figsize=(20,8))
         df2.plot(x='payment_type', kind='barh', stacked = True,ax=ax, color=['#FA643F', '#FFBCAB', '#CBB2B2', '#F1F1F1', '#FD9F9F'])

         # Add percentage text

         for p in ax.patches:
             width = p.get_width()
             height = p.get_height()
             x, y = p.get_xy()
             ax.text(x + width / 2,
                     y + height / 2,
                     '{:.0f}%'.format(width),
                     horizontalalignment='center',
                     verticalalignment='center')


         # Customize plot
         ax.set_xlabel('Percentage')
         ax.set_title('Payment Type Distribution')
         ax.legend(title='Payment Method', bbox_to_anchor=(1, 1))

         plt.tight_layout()
         plt.show()
```



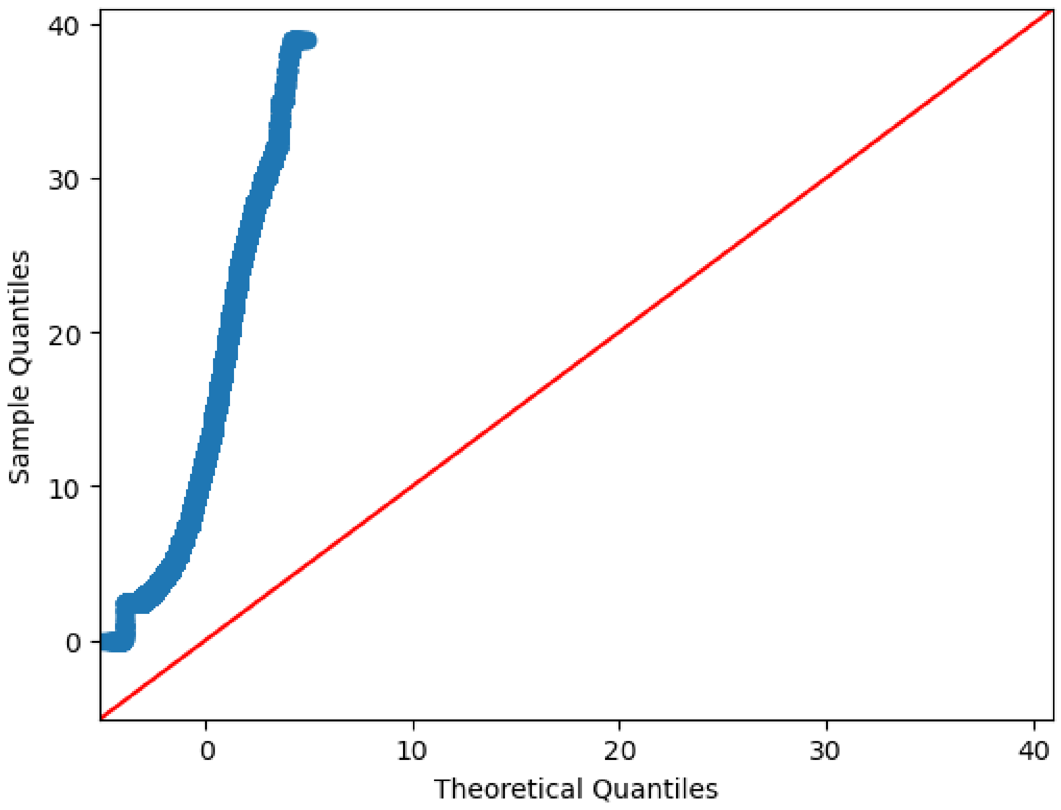
This visualization allows us to observe how payment preferences shift with varying passenger counts, providing insights into passenger behavior and payment choices in NYC's yellow taxi services.

Hypothesis Testing

Hypothesis testing is a crucial part of our analysis, helping us to determine whether the observed differences in fare amounts based on payment types are statistically significant. Given our large sample size, we use a Z-test instead of a T-test, as the Z-test is more appropriate for large samples when the population variance is known or the sample size is sufficiently large to approximate normality.



```
In [35]: import statsmodels.api as sm

         sm.qqplot(df['fare_amount'], line = '45')
         plt.show()
```



To determine if there is a significant difference in fare amounts between customers who use credit cards and those who use cash, we conducted a hypothesis test using a two-sample independent t-test.

Step 1: Formulating Hypotheses

- **Null Hypothesis (H_0)**: There is no difference in fare amounts between customers who use credit cards and customers who use cash.
- **Alternative Hypothesis (H_1)**: There is a difference in fare amounts between customers who use credit cards and customers who use cash.

Step 2: Type of Test

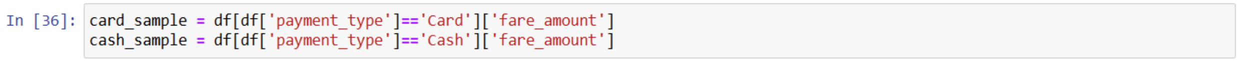
- **Two-Tailed Test**: We are interested in detecting any difference, whether credit card fares are significantly higher or lower than cash fares.

Step 3: Significance Level

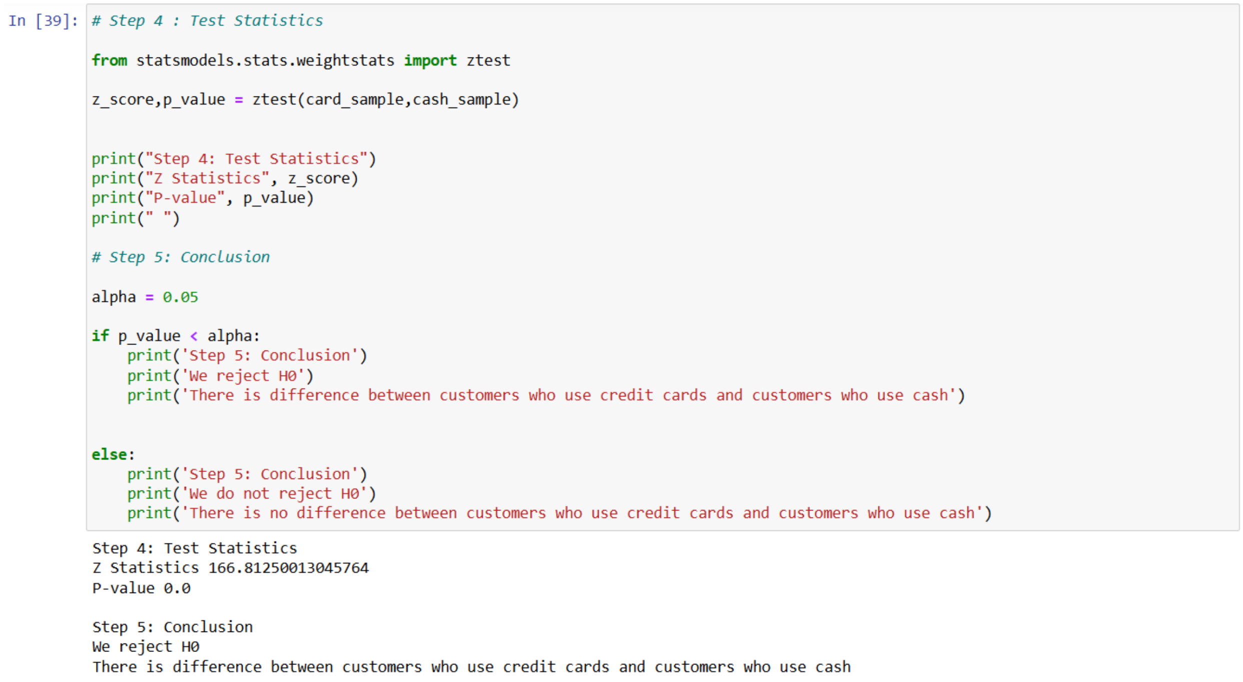
- We set the significance level (α\alphaα) at 0.05. This means we are willing to accept a 5% chance of rejecting the null hypothesis when it is actually true.

Step 4: Test Statistics

We calculate the Z-statistic using the means and standard deviations of the fare amounts for card and cash payments.



```
In [36]: card_sample = df[df['payment_type']=='Card']['fare_amount']
         cash_sample = df[df['payment_type']=='Cash']['fare_amount']
```



```
In [39]: # Step 4 : Test Statistics

         from statsmodels.stats.weightstats import ztest

         z_score,p_value = ztest(card_sample,cash_sample)


         print("Step 4: Test Statistics")
         print("Z Statistics", z_score)
         print("P-value", p_value)
         print(" ")

         # Step 5: Conclusion

         alpha = 0.05

         if p_value < alpha:
             print('Step 5: Conclusion')
             print('We reject H0')
             print('There is difference between customers who use credit cards and customers who use cash')


         else:
             print('Step 5: Conclusion')
             print('We do not reject H0')
             print('There is no difference between customers who use credit cards and customers who use cash')
```

```
Step 4: Test Statistics
Z Statistics 166.81250013045764
P-value 0.0

Step 5: Conclusion
We reject H0
There is difference between customers who use credit cards and customers who use cash
```

Step 5: Conclusion

Since the p-value (0.0) is less than the significance level (α = 0.05), we reject the null hypothesis.

- **Conclusion:** We conclude that there is a statistically significant difference between the fare amounts paid by customers using credit cards and those using cash. This finding suggests that the payment method does indeed influence the fare amount, with customers using credit cards tending to have higher fare amounts. This finding suggests that the method of payment influences fare amounts, with one method generally resulting in higher or lower fares compared to the other. This information is crucial for taxi service providers in understanding payment behavior and potentially optimizing revenue strategies.

Results

Our analysis reveals several important insights into the relationship between payment type and fare amount, as well as other related variables. Below, we present the detailed results of our descriptive analysis, hypothesis testing, and additional observations.

- **Shape and Data Types**:

  Initial dataset shape: 6,405,008 rows and 18 columns.
  Final dataset shape after cleaning and filtering: 2,780,283 rows and 5 columns.
  Data types were appropriately converted, and new features such as trip duration were calculated.

- **Handling Missing Values**:

  The dataset originally had 1.02% missing values in critical columns.
  These missing values were dropped, resulting in 6,339,567 rows remaining.

- **Handling Duplicated Values**:

  Duplicate rows were identified and removed, reducing the dataset to 3,007,861 rows.

- **Outlier Treatment**:

  Outliers were identified using the interquartile range (IQR) method and subsequently removed.

- **Descriptive Statistics**:

  Summary statistics revealed that the average fare amount was $17.81, with a standard deviation of $15.07.
  The average trip distance was approximately 4.54 miles, and the average trip duration was around 24.15 minutes.
  Negative values in fare amount, trip distance, and duration were filtered out as they were unrealistic.

- **Distribution Analysis**:

Passenger Count:

Most trips had a single passenger (58.2%).
A significant portion of trips had 2 passengers (19.04%).
Very few trips had more than 5 passengers.

Payment Type:

Card payments constituted 67.83% of all transactions, while cash payments made up 30.76%.
Other payment types were negligible and thus excluded from further analysis.

- **Mean and Standard Deviation by Payment Type**:

Fare Amount:

Card: Mean = $13.11, Std = $5.85
Cash: Mean = $11.76, Std = $5.61

Trip Distance:

Card: Mean = 2.99 miles, Std = 1.99
Cash: Mean = 2.60 miles, Std = 1.91

- **Payment Type Distribution**:

  Card payments were significantly more common than cash payments, likely due to convenience, security, or incentives.

- **Passenger Count and Payment Type**:

  Both card and cash payments were predominantly used by single passengers.
  Larger groups (more than 5 passengers) were less likely to use taxis or might have opted for alternative payment methods.

Additional Observations

Journey Insights

- **Higher Fare and Trip Distance for Card Payments**:

Customers paying with cards tend to have a slightly higher average trip distance and fare amount compared to those paying with cash. This indicates a preference for using cards for longer trips and higher fares.

- **Preference for Card Payments**:

The proportion of customers paying with cards is significantly higher than those paying with cash, with card payments accounting for 67.5% of all transactions compared to cash payments at 32.5%. This suggests a strong preference for card payments, likely due to convenience, security, and incentives.

- **Passenger Count Distribution**:

Both card and cash payments are predominantly associated with single-passenger rides, making up the largest proportion of transactions. The percentage of transactions decreases as passenger count increases, indicating that larger groups may prefer alternative payment methods or transportation options.

Preference of Payment Types

- The proportion of customers paying with cards is significantly higher than those paying with cash, with card payments accounting for 67.5% of all transactions compared to cash payments at 32.5%
- This indicates a strong preference among customers for using cards payment over cash, potentially due to convenience, security, or incentives for card transactions.

Passenger Count Analysis

- Among card payments, rides with a single passenger (passenger_count = 1) comprise the largest proportion constituting 40.08% of all card transactions.
- Similarly, cash payments are predominantly associated with single-passenger rides, making up 20.04% of all cash transactions.
- There is noticeable decrease in the percentage of transactions as the passenger count increase, suggestion that larger groups are less likely to use taxis or may option for alternative payment methods.
- These insights emphasize the importance of considering both payment method and passenger count when analyzing transaction data, as they provide valuable insights into customer behavior and preferences.

Implications

Revenue Maximization:

- Encouraging card payments could potentially increase revenue for taxi drivers, as customers using cards tend to have higher fare amounts and longer trip distances.
- Taxi companies could implement strategies such as offering discounts or incentives for card payments to capitalize on this trend.

Customer Convenience:

- Providing seamless and secure card payment options can enhance customer convenience and satisfaction.
- This could lead to increased customer loyalty and repeat business.

Operational Efficiency:

- Understanding the preference for payment types and passenger behavior can help in optimizing fleet management and operational strategies.

Limitations

Data Quality: Despite extensive cleaning, some data quality issues such as outliers and negative values were present and had to be filtered out.

Scope: The analysis is limited to NYC yellow taxi trip data and may not be generalizable to other regions or types of transportation services.

External Factors: Factors such as weather conditions, traffic, and socio-economic status of passengers were not considered in this analysis.

Conclusion

Our analysis demonstrates a significant relationship between payment type and fare amount, with card payments associated with higher fare amounts and longer trip distances. By encouraging the use of card payments, taxi companies can potentially maximize revenue and improve customer satisfaction. However, it is essential to address data quality issues and consider additional factors in future analyses to gain a more comprehensive understanding of the dynamics at play.

Recommendations

- Encourage customers to pay with credit cards to capitalize on the potential for generating more revenue for taxi cab drivers.
- Implement strategies such as offering incentives or discounts for credit card transactions to incentivize customers to choose this payment method.
- Provide seamless and secure credit card payment options to enhance customer convenience and encourage adoption of this preferred payment method.