# TAXI FARE HYPOTHESIS TESTING REPORT

A comprehensive statistical analysis comparing fare amounts between card and cash taxi payments

## TABLE OF CONTENTS



## 1. PROBLEM STATEMENT

The goal is to determine whether there is a statistically significant difference in taxi fare amounts between passengers who pay using card versus cash. This analysis helps understand pricing behavior and customer payment preferences.

## 2. RESEARCH QUESTION

Is there any significant difference in the mean fare amount between card and cash taxi trips?

## 3. METHODOLOGY

This analysis includes: - Feature selection - Outlier removal using the IQR method – Exploratory data analysis - Two-sample independent t-test for fare_amount and duration. Below is an image of the initial dataset preview.

## *4. DATA OVERVIEW- FEATURE SELECTION*

The dataset includes these columns used for analysis: passenger_count, trip_distance, payment_type, fare_amount,duration.

```
df = df[['passenger_count','trip_distance','payment_type','fare_amount','duration']]
df.head()
```

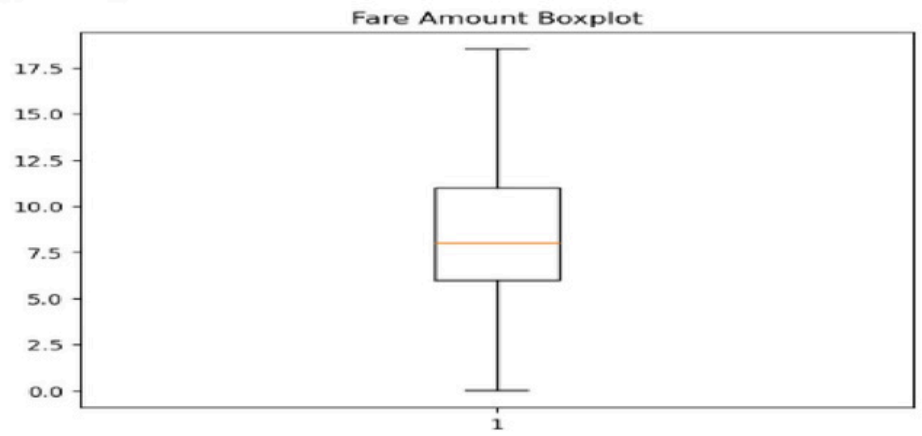
| | passenger_count | trip_distance | payment_type | fare_amount | duration |
|---|---|---|---|---|---|
| **0** | 1.0 | 1.2 | 1.0 | 6.0 | 0 days 00:04:48 |
| **1** | 1.0 | 1.2 | 1.0 | 7.0 | 0 days 00:07:25 |
| **2** | 1.0 | 0.6 | 1.0 | 6.0 | 0 days 00:06:11 |
| **3** | 1.0 | 0.8 | 1.0 | 5.5 | 0 days 00:04:51 |
| **4** | 1.0 | 0.0 | 2.0 | 3.5 | 0 days 00:02:18 |

## *5. OUTLIER REMOVAL IQR METHOD*

### Removing Outliers by IQR Method:

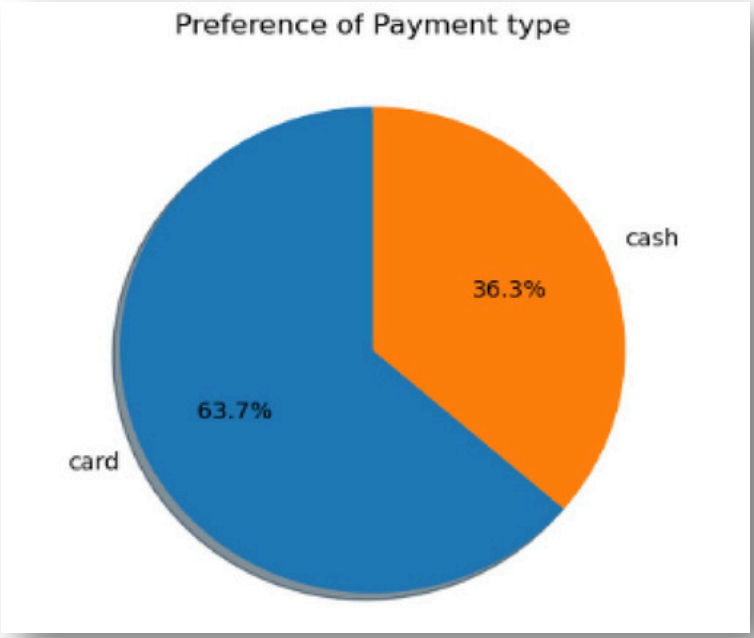
```
for col in ['fare_amount','trip_distance','duration']:
    q1 = df[col].quantile(0.25)
    q3 = df[col].quantile(0.75)
    IQR = q3-q1
    lower_bound = q1-1.5*IQR
    upper_bound = q3+1.5*IQR
    df= df[(df[col]>=lower_bound) & (df[col]<=upper_bound)]
```

```
plt.boxplot(df['fare_amount'])
plt.title('Fare Amount Boxplot')
plt.show()
```

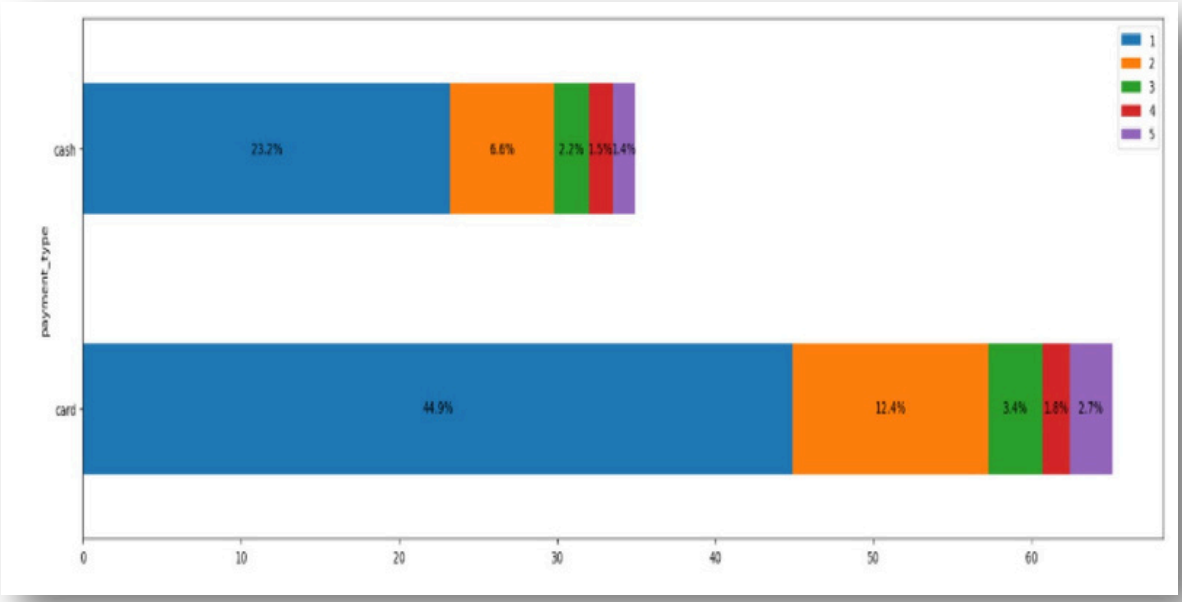


## *6. EXPLORATORY ANALYSIS*

### A) PAYMENT TYPE DISTRIBUTION: CARD VS CASH




### B) PASSENGER COUNT VS PAYMENT TYPE



## *7. HYPOTHESIS TESTING RESULTS*

A two-sample t-test was performed to compare mean fare amounts between card and cash payments.

Null Hypothesis (H0): There is no difference in mean fare amount.

Alternative Hypothesis (Ha):There is a difference in mean fare amount.

t-statistic = 6.9854p-value = 2.856e-12. Since p < 0.05, we reject the null hypothesis. Card payments tend to have higher fare amounts.

```
card_sample = df[df['payment_type'] == 'card']['fare_amount']
cash_sample = df[df['payment_type'] == 'cash']['fare_amount']

t_stats, p_value = st.ttest_ind(card_sample, cash_sample, equal_var=False)

print("T-statistic:", t_stats)
print("p-value:", p_value)
```

```
T-statistic: 6.985426003862178
p-value: 2.8561867684480994e-12
```

## *8. INSIGHTS & RECOMMENDATIONS*

- Card users form the majority (≈64%). - Outliers were successfully removed using IQR. - Statistical testing confirms a significant difference in mean fare. - Card payments are associated with slightly higher fares.

## *CONCLUSION*

The hypothesis testing results show a meaningful difference in fare amounts between payment methods. Further analysis can explore time-of-day, distance segmentation, and multivariate regression.