TASK 2

Preprocessing and EDA

- we had 5 null values in 9 features each so we deleted the null rows
- no duplicates found
- we generated feature (Minutes) to handle dates
- Drop for names/keys features
- used LabelEncoder for object values

Outliers handling

- Fare amount: remove negatives, clip extreme high (100)
- Valid passenger count: 1 to 6
- Distance-related features (clip top 99th percentile)

```python
#Are most cars in good condition?
df['Car Condition'].value_counts()
```

```
Car Condition
Very Good    125310
Bad          124977
Good         124967
Excellent    124741
Name: count, dtype: int64
```

```python
#What is the most common weather condition during trips?
df['Weather'].value_counts()
```

```
Weather
sunny     100433
cloudy    100060
rainy      99971
stormy     99955
windy      99576
Name: count, dtype: int64
```

```python
#Which traffic condition occurs most frequently?
df['Traffic Condition'].value_counts()
```

```
Traffic Condition
Congested Traffic    166846
Dense Traffic        166581
Flow Traffic         166568
Name: count, dtype: int64
```

```python
#Which year has the highest number of trips?
df['year'].value_counts()
```

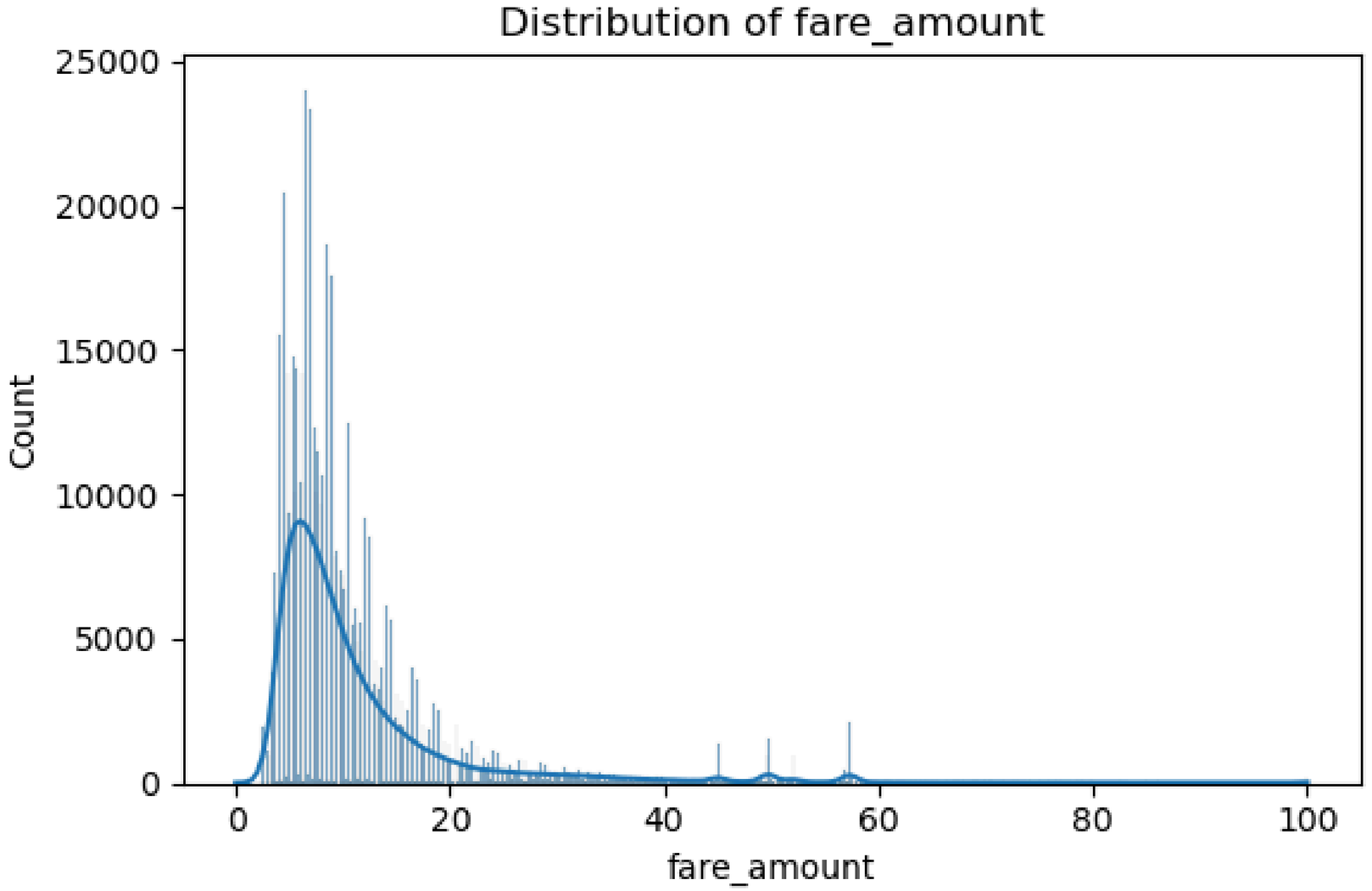
```
year
2012    80222
2011    79528
2013    78033
2009    77002
2010    75791
2014    74608
2015    34811
Name: count, dtype: int64
```

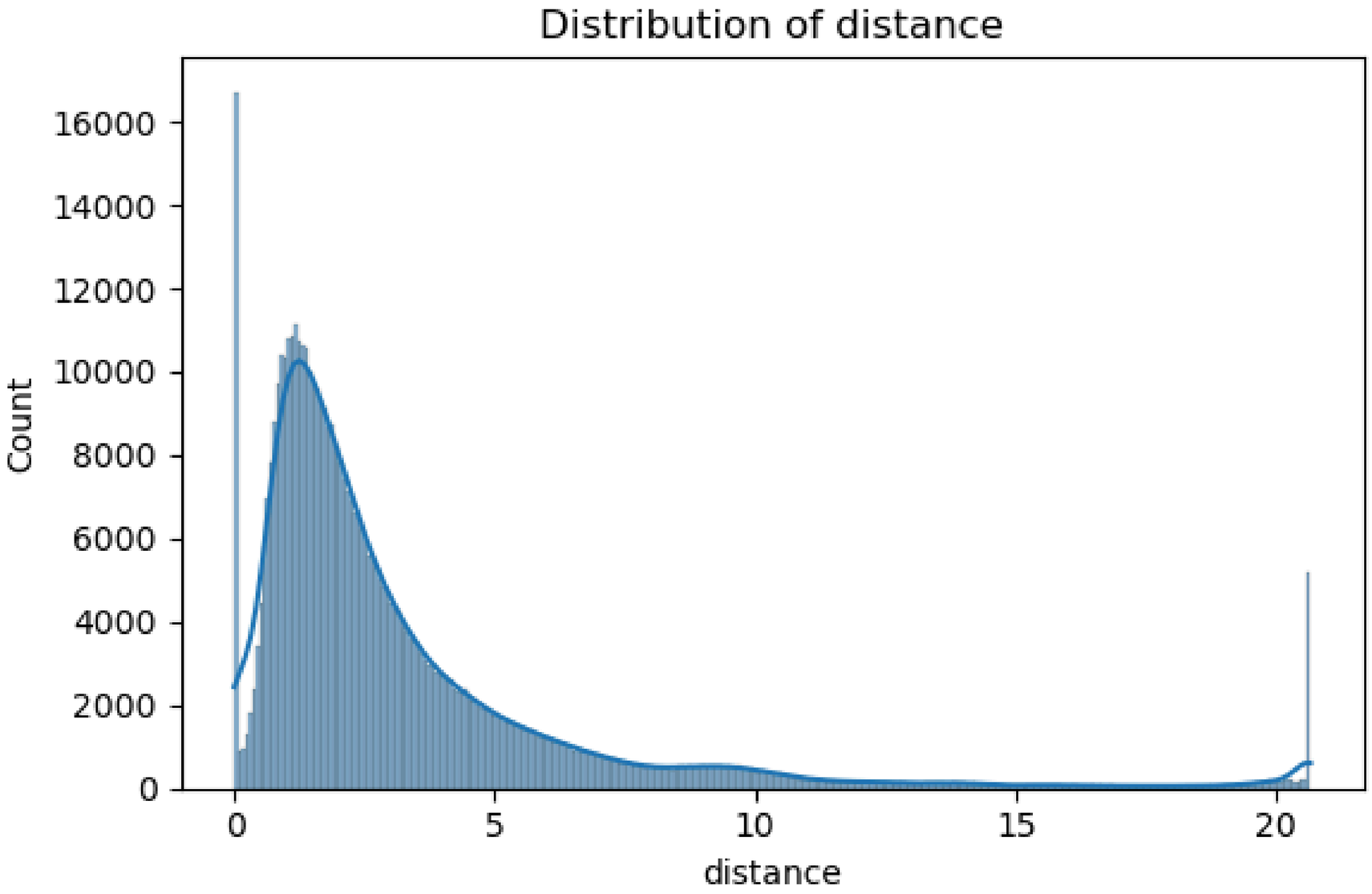
```python
#What is the most popular day for trips?
df['weekday'].value_counts()
```

```
weekday
Friday       77224
Saturday     76212
Thursday     74776
Wednesday    72230
Tuesday      69929
Sunday       65387
Monday       64237
Name: count, dtype: int64
```

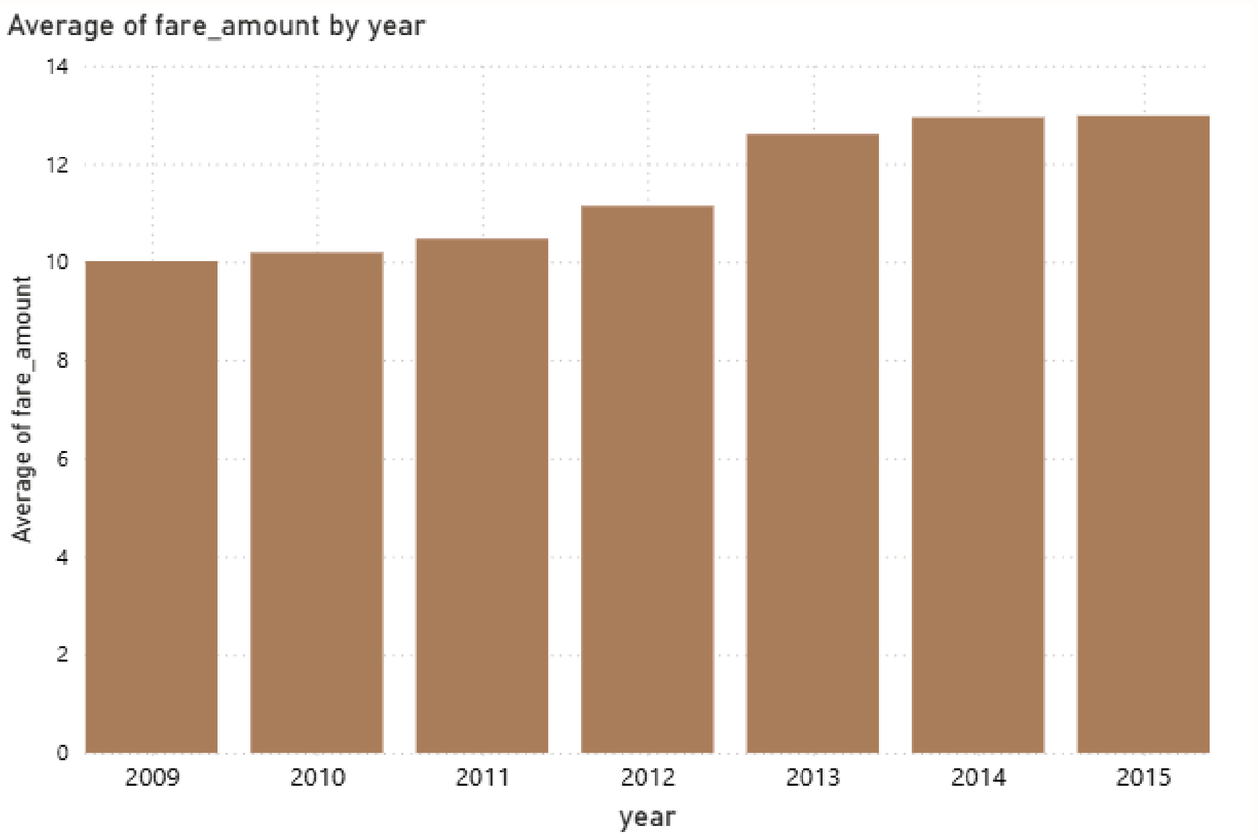
```python
#Which month has the highest number of trips?
df['month'].value_counts()
```

```
month
5     46733
3     46714
4     45959
6     44827
1     44547
2     42454
10    40551
12    38476
7     38127
9     37979
11    37757
8     35871
Name: count, dtype: int64
```

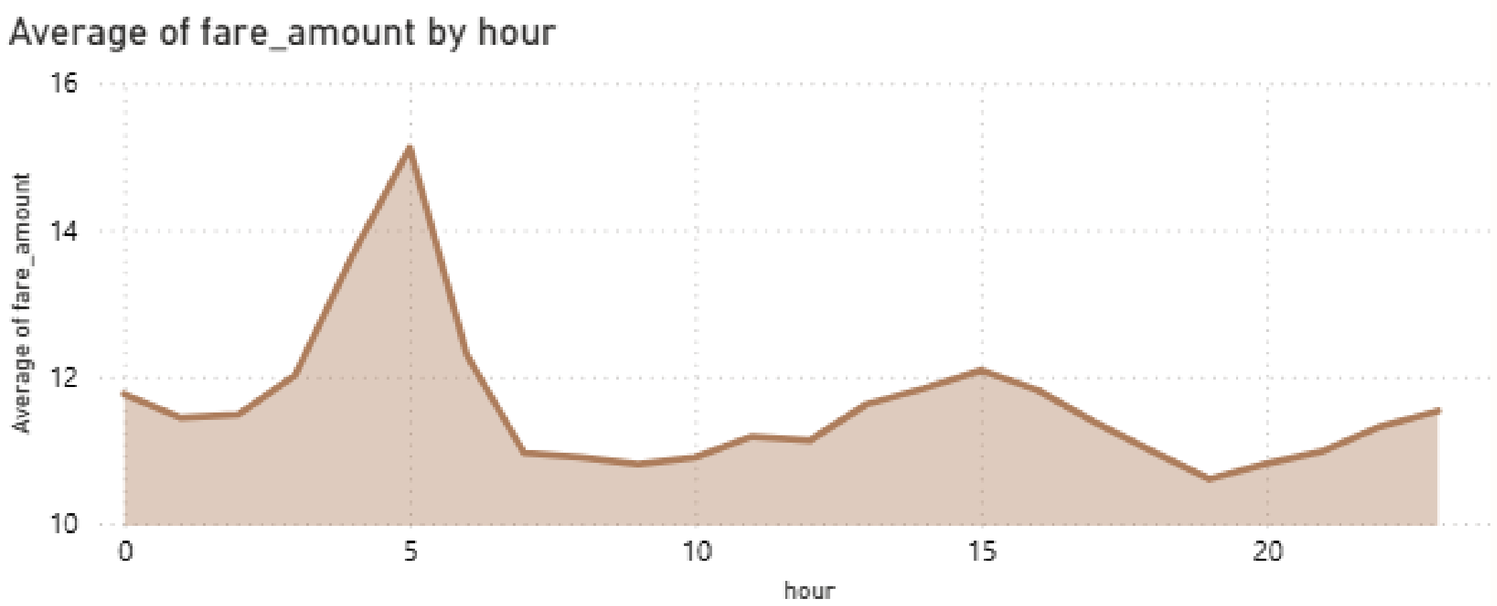

Distribution of fare_amount
25000
20000
15000
10000
5000
0
Count
0
20
40
60
80
100
fare_amount


Distribution of distance
16000
14000
12000
10000
8000
6000
4000
2000
0
Count
0
5
10
15
20
distance

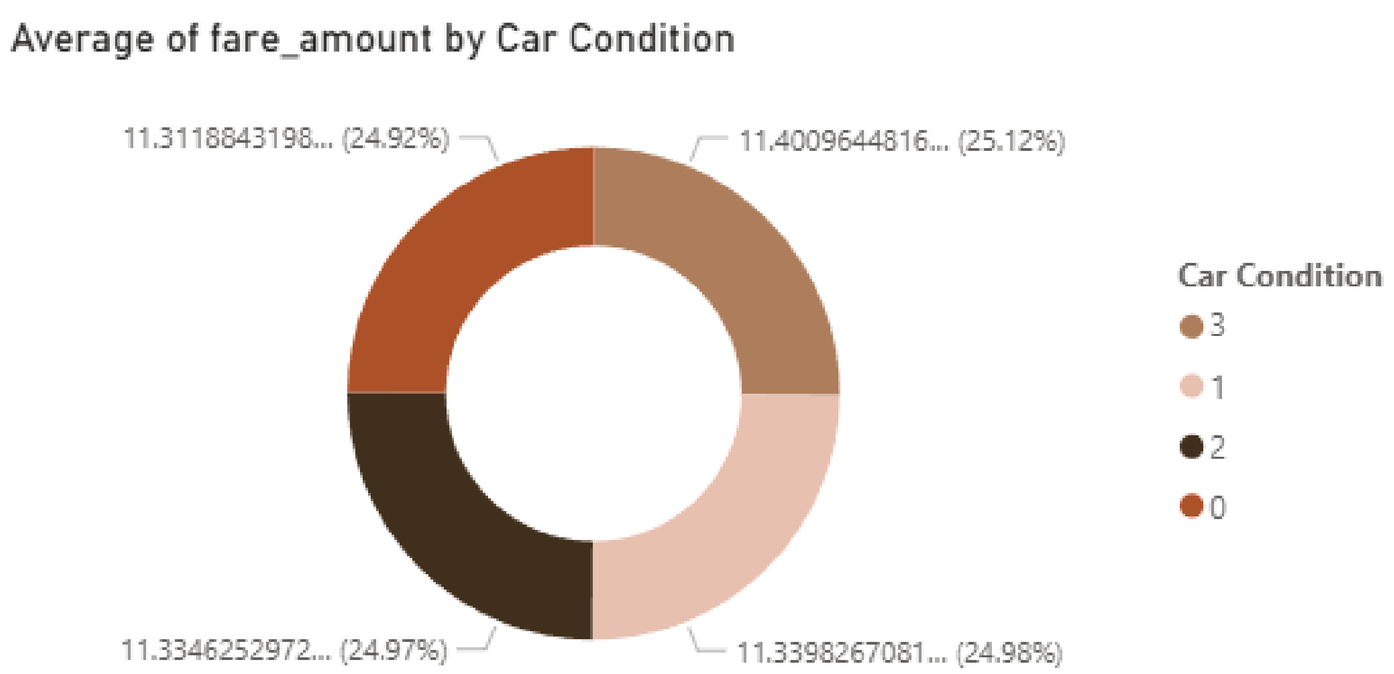
HOW DOES THE AVERAGE FARE AMOUNT VARY ACROSS DIFFERENT YEARS?



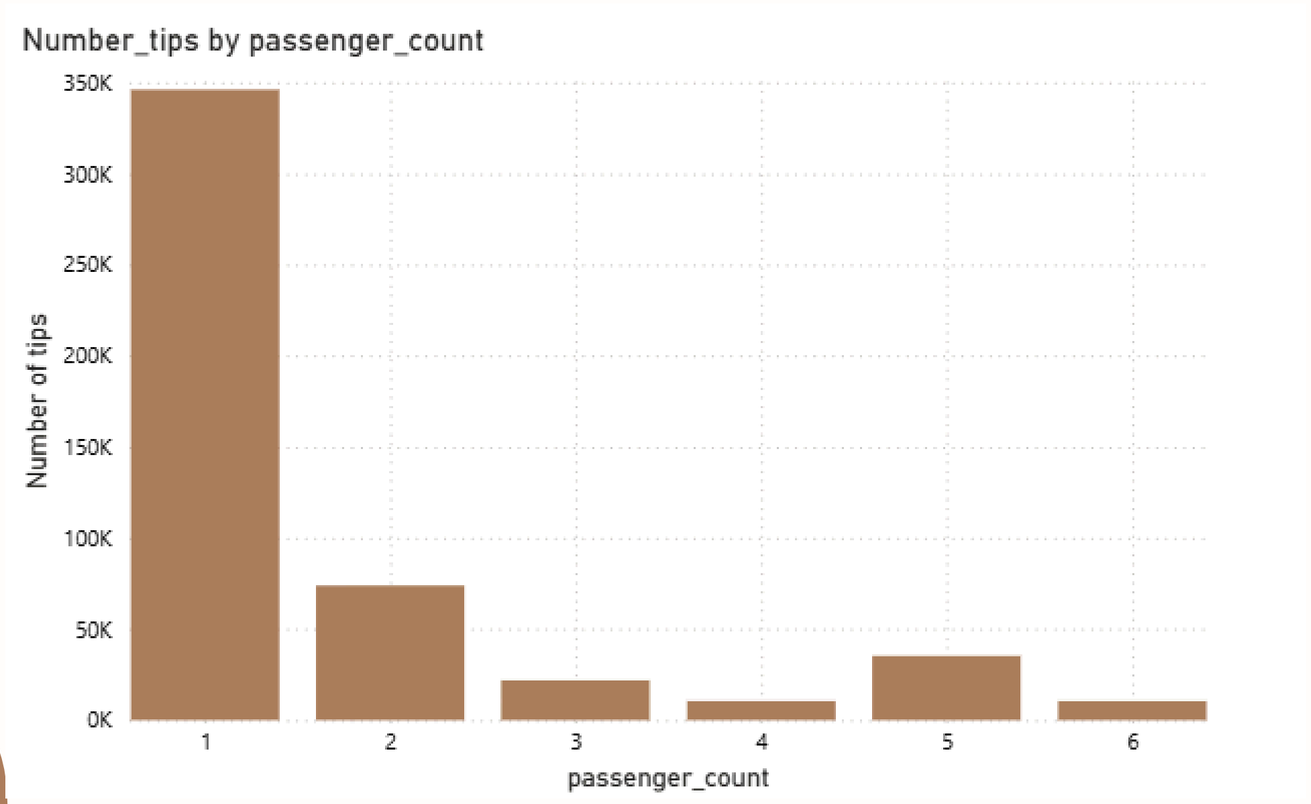
DOES THE AVERAGE FARE AMOUNT VARY BY WEEKDAY?



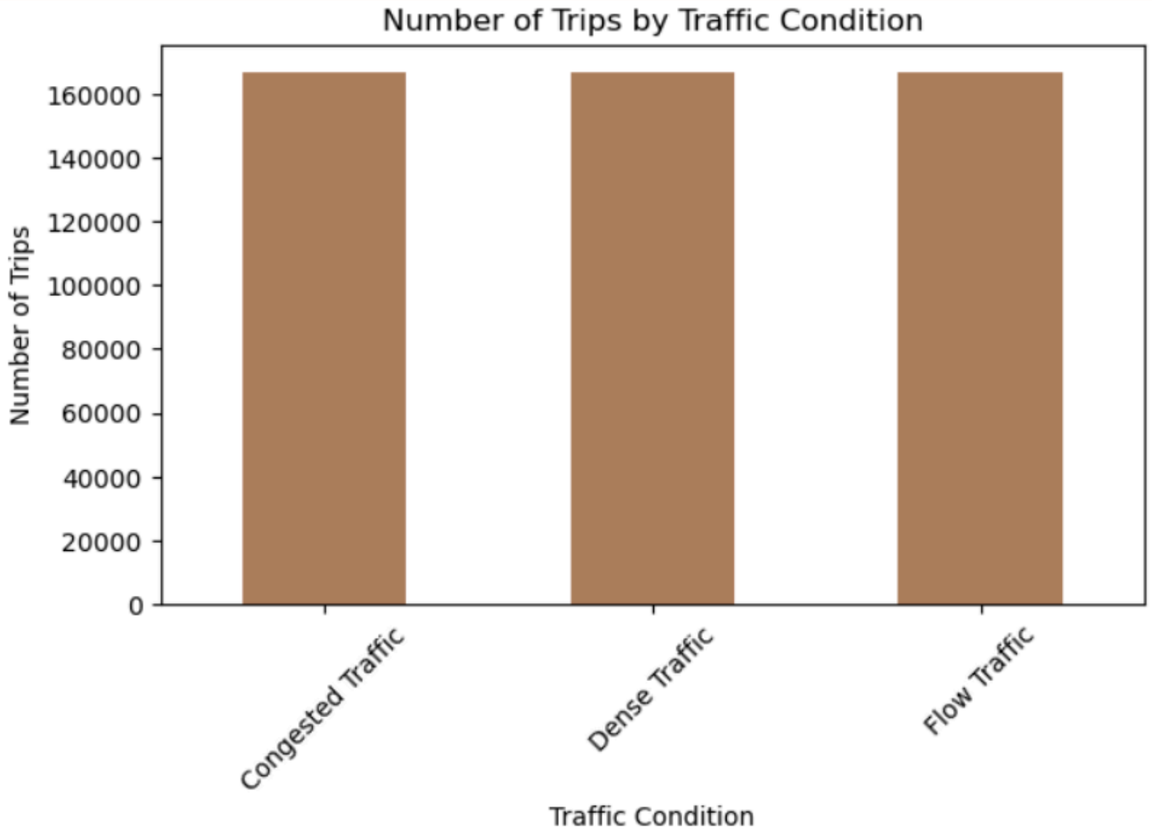
DOES THE AVERAGE FARE AMOUNT VARY DEPENDING ON THE CAR CONDITION?



WHICH PASSENGER COUNT HAS THE HIGHEST NUMBER OF TRIPS?



WHAT IS NUMBER OF TRIPS FOR EACH TRAFFIC CONDITION?

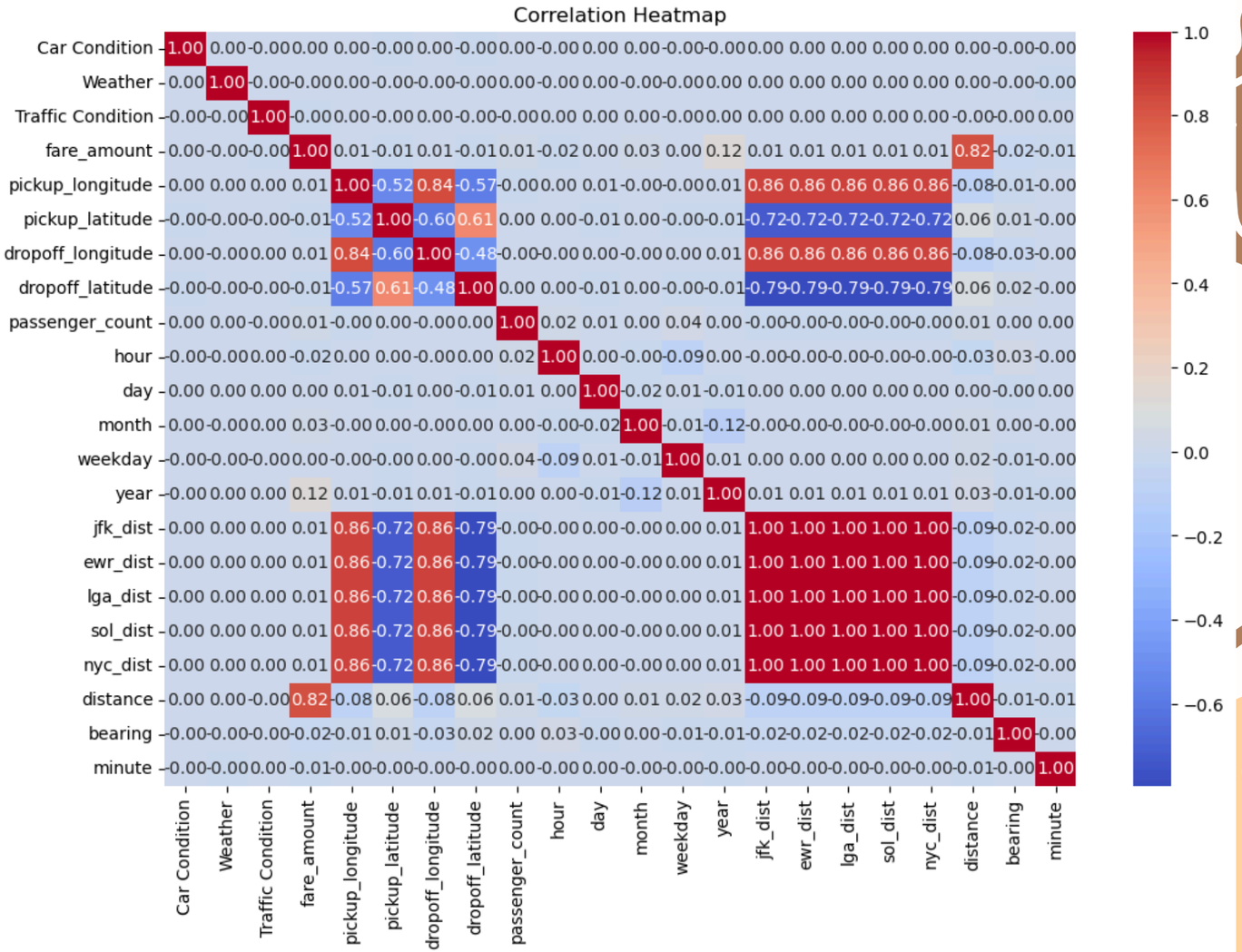


FARE_AMOUNT IS STRONGLY CORRELATED WITH DISTANCE.

PICKUP_LONGITUDE, PICKUP_LATITUDE, DROPOFF_LONGITUDE, AND DROPOFF_LATITUDE ARE HIGHLY CORRELATED WITH DISTANCE-RELATED COLUMNS (JFK_DIST, EWR_DIST, ...)INDICATING GPS-BASED DISTANCES ARE EFFECTIVE PREDICTORS.

FEATURES LIKE WEATHER, CAR CONDITION, AND TRAFFIC CONDITION HAVE VERY WEAK CORRELATIONS WITH FARE, SUGGESTING THEY MAY NOT SIGNIFICANTLY INFLUENCE FARE PREDICTION.

THE YEAR HAS A MODERATE CORRELATION WITH FARE_AMOUNT (0.12), PERHAPS DUE TO INFLATION OR CHANGING FARE POLICIES.



THANK YOU