Taxi Demand Prediction

Yolcular

Ceren TİMURKAN
Çiğdem KILIÇ KOÇER

AMAÇ

Lokasyon bazlı taksi talep yoğunluğunun tahmini

01 EDA

02 Model

03 Prediction

Data İçeriği

- 201901 ve 202002 dönemleri arasındaki NYC sarı taksi datası
- Taksi biniş-iniş lokasyon ve zamanı
- Yolculuk süresi ve mesafesi
- Taksi tarife, bahşiş vb ücretlendirmeleri

NYC
Taxi & Limousine Commission

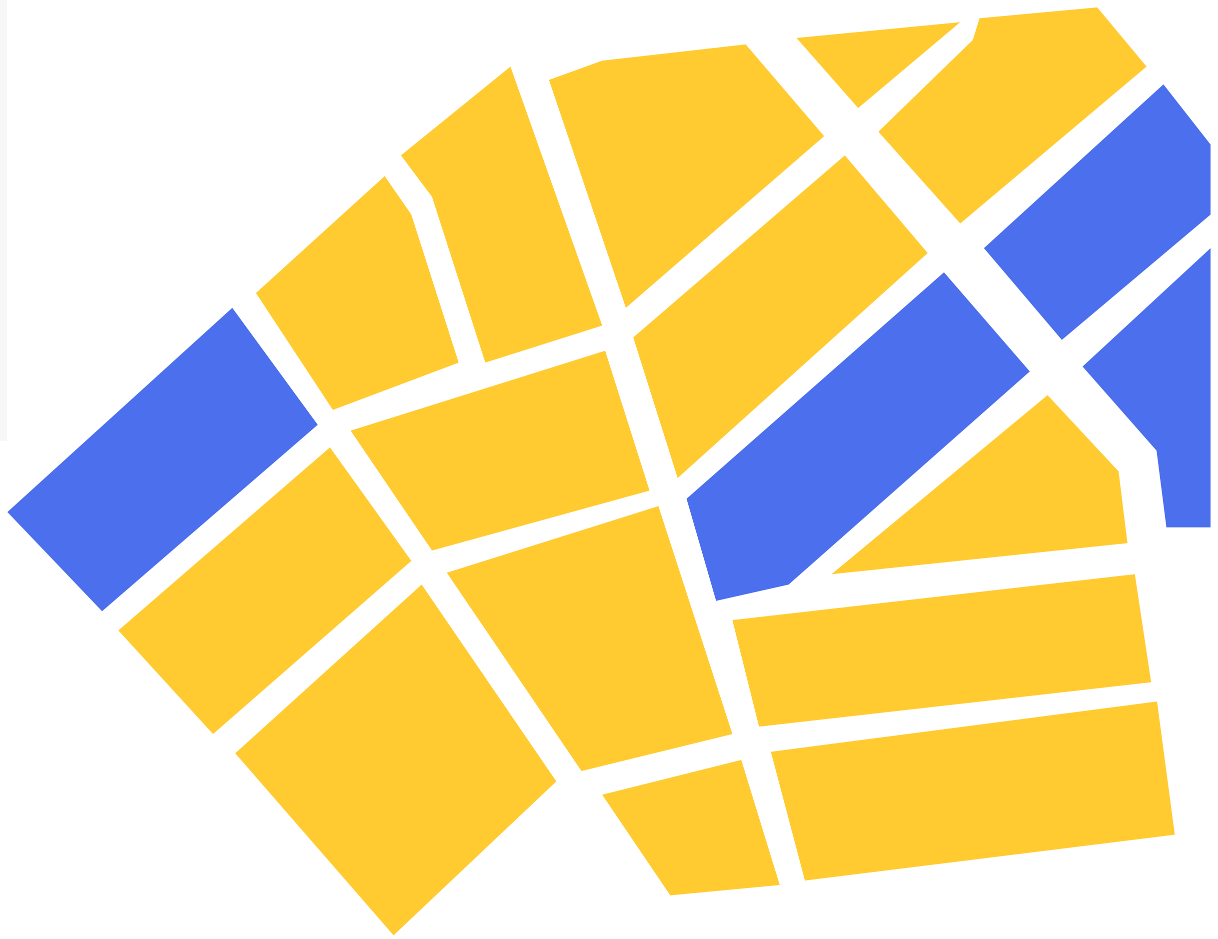
Data Describe

```
973131   -73.9777   40.7580
###################### NA ######################
Unnamed: 0                    0
Unnamed: 0.1                  0
Unnamed: 0_x                  0
VendorID                      0
tpep_pickup_datetime          0
tpep_dropoff_datetime         0
passenger_count            5626
trip_distance                 0
RatecodeID                 5626
store_and_fwd_flag         5626
PULocationID                  0
DOLocationID                  0
payment_type                  0
fare_amount                   0
extra                         0
mta_tax                       0
tip_amount                    0
tolls_amount                  0
improvement_surcharge         0
total_amount                  0
congestion_surcharge      54093
airport_fee              973132
Unnamed: 0_y               9164
zone                       9164
borough                    9164
Longitude                  9164
Latitude                   9164
dtype: int64
###################### Quantiles ######################
```

```
###################### Shape ######################
(973132, 27)
###################### Types ######################
Unnamed: 0                 int64
Unnamed: 0.1               int64
Unnamed: 0_x               int64
VendorID                   int64
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               int64
fare_amount              float64
extra                    float64
mta_tax                  float64
tip_amount               float64
tolls_amount             float64
improvement_surcharge    float64
total_amount             float64
congestion_surcharge     float64
airport_fee              float64
Unnamed: 0_y             float64
zone                      object
borough                   object
Longitude                float64
Latitude                 float64
dtype: object
```

Özellik Mühendisliği

```
df_final["day"] = df_final["tpep_pickup_datetime"].dt.day

df_final["pickup_day"] = df_final["tpep_pickup_datetime"].dt.day_name()

df_final["week"] = df_final["tpep_pickup_datetime"].dt.week

df_final["pickup_hour"] = df_final["tpep_pickup_datetime"].dt.hour

df_final["pickup_month"] = df_final["tpep_pickup_datetime"].dt.month

df_final["year"] = df_final["tpep_pickup_datetime"].dt.year

df_final["period"] = df_final["year"] * 100 + df_final["pickup_month"]

df_final["trip_n"]=1
```

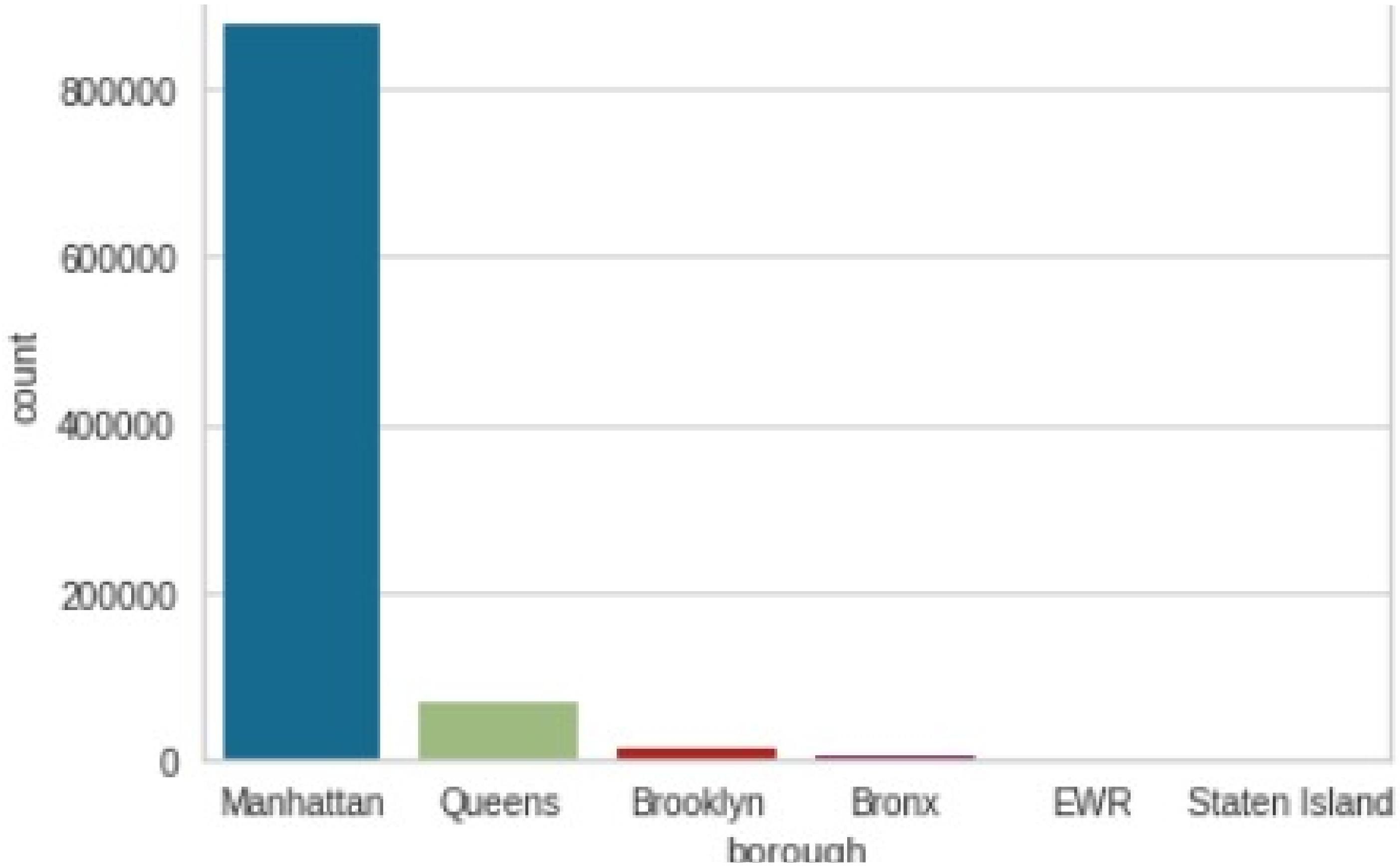


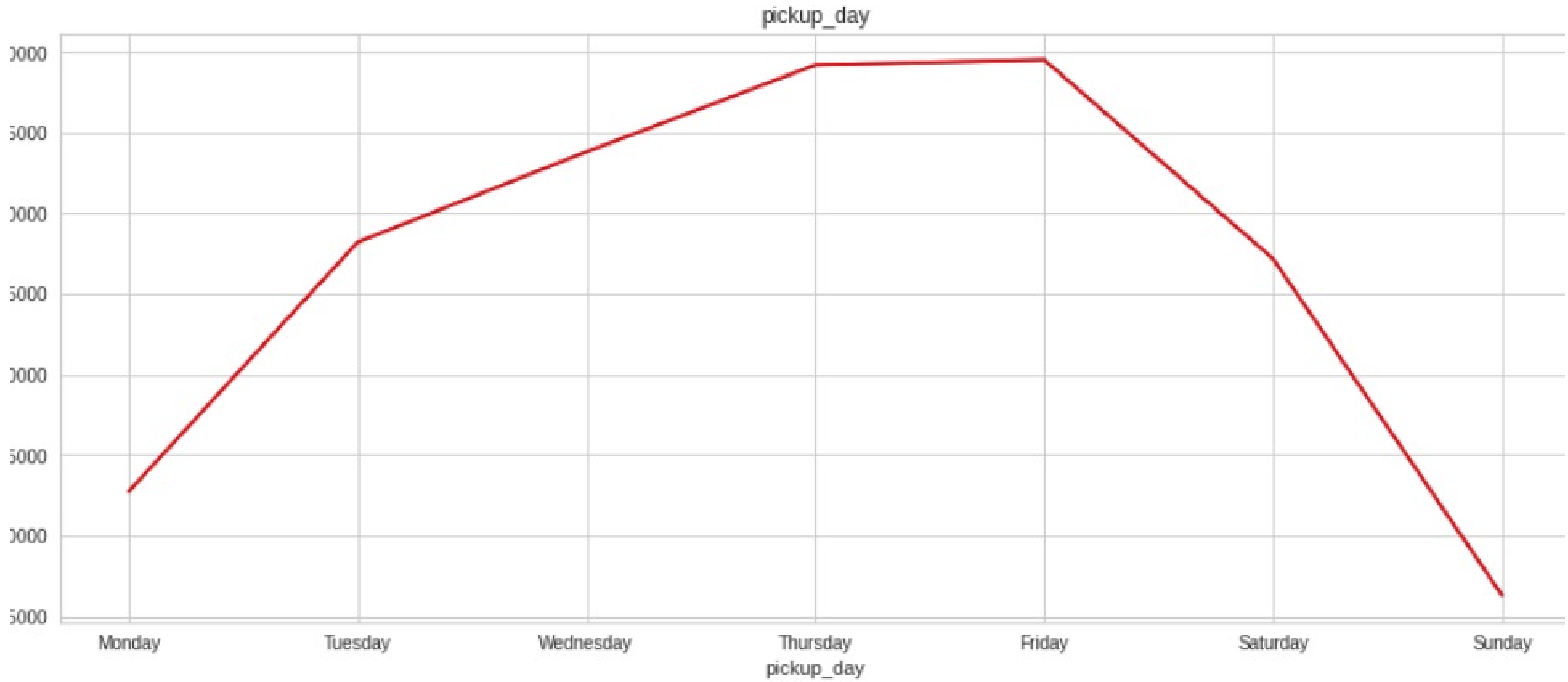
Data Describe

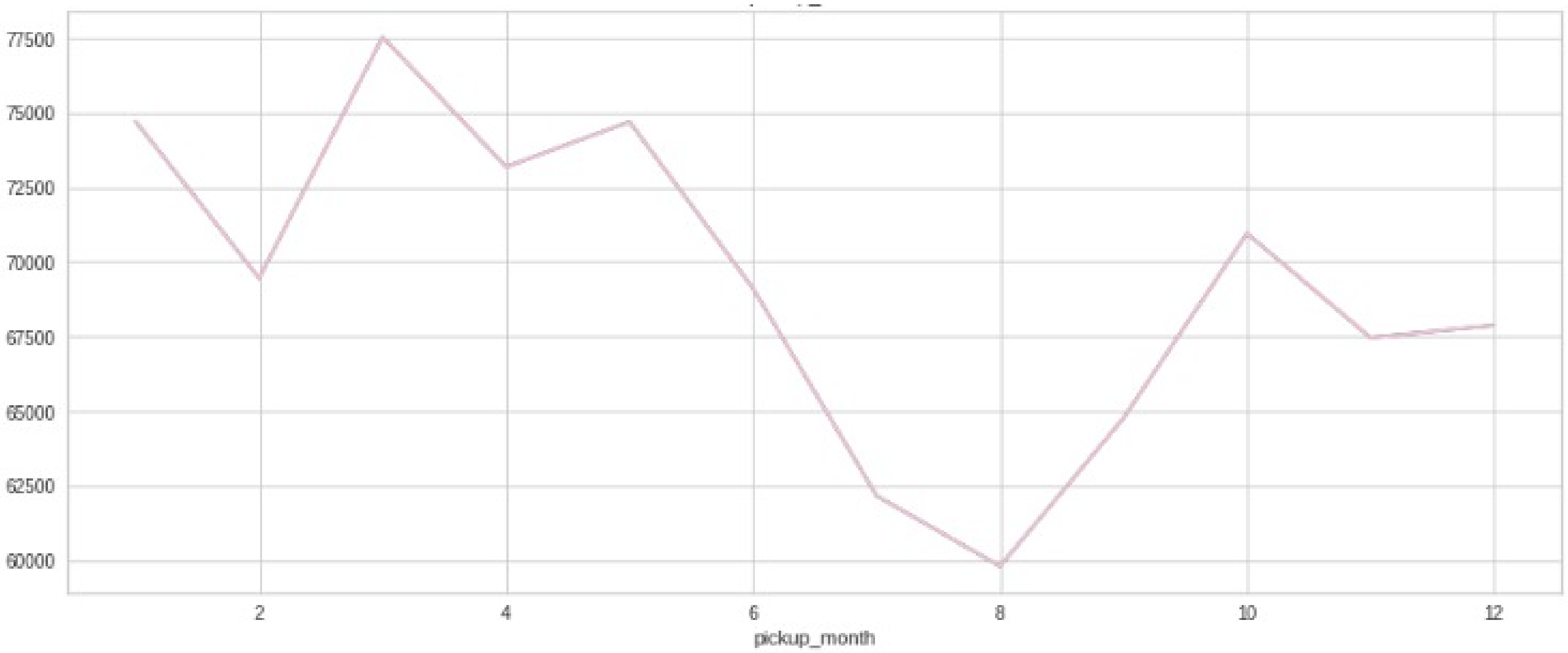
	count	mean	std	min	25%	50%	75%	max
trip_distance	956,824.0000	2.9620	3.8622	0.0000	0.9800	1.6200	3.0000	133.5200
RatecodeID	956,824.0000	1.0524	0.3425	1.0000	1.0000	1.0000	1.0000	6.0000
store_and_fwd_flag	956,824.0000	0.0084	0.0913	0.0000	0.0000	0.0000	0.0000	1.0000
PULocationID	956,824.0000	162.3488	65.4783	1.0000	114.0000	162.0000	232.0000	263.0000
DOLocationID	956,824.0000	160.9035	69.9032	1.0000	107.0000	162.0000	233.0000	265.0000
payment_type	956,824.0000	1.2834	0.4712	1.0000	1.0000	1.0000	2.0000	4.0000
tip_amount	956,824.0000	2.1983	2.8415	-10.5600	0.0000	1.8600	2.9500	600.0000
total_amount	956,824.0000	18.9192	14.5089	-109.9200	11.1600	14.6300	20.3000	700.3000
congestion_surcharge	956,824.0000	2.1172	0.9003	0.0000	2.5000	2.5000	2.5000	2.5000
Longitude	956,824.0000	-73.9705	0.0424	-74.2335	-73.9905	-73.9786	-73.9656	-73.7110
Latitude	956,824.0000	40.7520	0.0300	40.5255	40.7403	40.7567	40.7686	40.8995
day	956,824.0000	15.5659	8.6857	1.0000	8.0000	15.0000	23.0000	31.0000
week	956,824.0000	23.0694	15.7722	1.0000	8.0000	21.0000	37.0000	52.0000
weekday	956,824.0000	2.9836	1.9250	0.0000	1.0000	3.0000	5.0000	6.0000
pickup_hour	956,824.0000	13.9055	6.0096	0.0000	10.0000	15.0000	19.0000	23.0000
pickup_month	956,824.0000	5.7240	3.6417	1.0000	2.0000	5.0000	9.0000	12.0000
year	956,824.0000	2,019.1308	0.3372	2,019.0000	2,019.0000	2,019.0000	2,019.0000	2,020.0000
period	956,824.0000	201,918.8011	32.2397	201,901.0000	201,904.0000	201,907.0000	201,911.0000	202,002.0000
trip_n	956,824.0000	1.0000	0.0000	1.0000	1.0000	1.0000	1.0000	1.0000
duration	956,824.0000	17.3330	70.5275	-57.0000	6.0000	11.0000	18.0000	1,439.0000

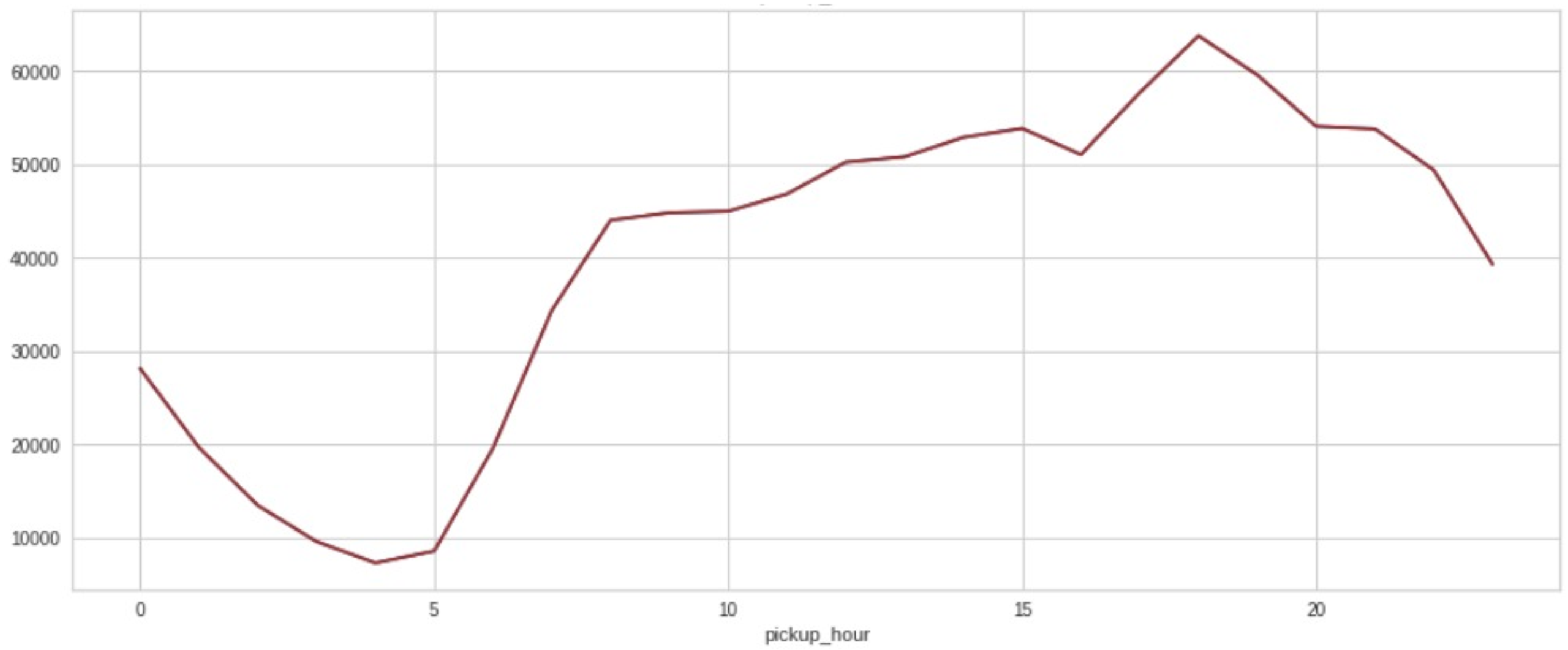
```
#################### Quantiles ####################
                          0.0000      0.0100      0.0500      0.5000  \
trip_distance             0.0000      0.1000      0.4800      1.6200
RatecodeID                1.0000      1.0000      1.0000      1.0000
store_and_fwd_flag        0.0000      0.0000      0.0000      0.0000
PULocationID              1.0000     13.0000     48.0000    162.0000
DOLocationID              1.0000      7.0000     43.0000    162.0000
payment_type              1.0000      1.0000      1.0000      1.0000
tip_amount              -10.5600      0.0000      0.0000      1.8600
total_amount           -109.9200      5.8000      7.8000     14.6300
congestion_surcharge      0.0000      0.0000      0.0000      2.5000
Longitude               -74.2335    -74.0130    -74.0075    -73.9786
Latitude                 40.5255     40.6470     40.7068     40.7567
day                       1.0000      1.0000      2.0000     15.0000
week                      1.0000      1.0000      2.0000     21.0000
weekday                   0.0000      0.0000      0.0000      3.0000
pickup_hour               0.0000      0.0000      2.0000     15.0000
pickup_month              1.0000      1.0000      1.0000      5.0000
year                  2,019.0000  2,019.0000  2,019.0000  2,019.0000
period              201,901.0000 201,901.0000 201,901.0000 201,907.0000
trip_n                    1.0000      1.0000      1.0000      1.0000
duration                -57.0000      1.0000      3.0000     11.0000

                          0.7500      0.9500      0.9900      1.0000
trip_distance             3.0000     11.0800     19.1300    133.5200
RatecodeID                1.0000      1.0000      2.0000      6.0000
store_and_fwd_flag        0.0000      0.0000      0.0000      1.0000
PULocationID            232.0000    249.0000    263.0000    263.0000
DOLocationID            233.0000    249.0000    263.0000    265.0000
payment_type              2.0000      2.0000      2.0000      4.0000
tip_amount                2.9500      7.0000     12.2800    600.0000
total_amount             20.3000     50.4700     73.9200    700.3000
congestion_surcharge      2.5000      2.5000      2.5000      2.5000
Longitude               -73.9656    -73.8736    -73.7865    -73.7110
Latitude                 40.7686     40.7917     40.8095     40.8995
day                      23.0000     29.0000     31.0000     31.0000
week                     37.0000     49.0000     52.0000     52.0000
weekday                   5.0000      6.0000      6.0000      6.0000
pickup_hour              19.0000     22.0000     23.0000     23.0000
pickup_month              9.0000     12.0000     12.0000     12.0000
year                  2,019.0000  2,020.0000  2,020.0000  2,020.0000
period              201,911.0000 202,002.0000 202,002.0000 202,002.0000
trip_n                    1.0000      1.0000      1.0000      1.0000
duration                 18.0000     37.0000     62.0000  1,439.0000
```

Data Analizi - Kategorik Değişkenler



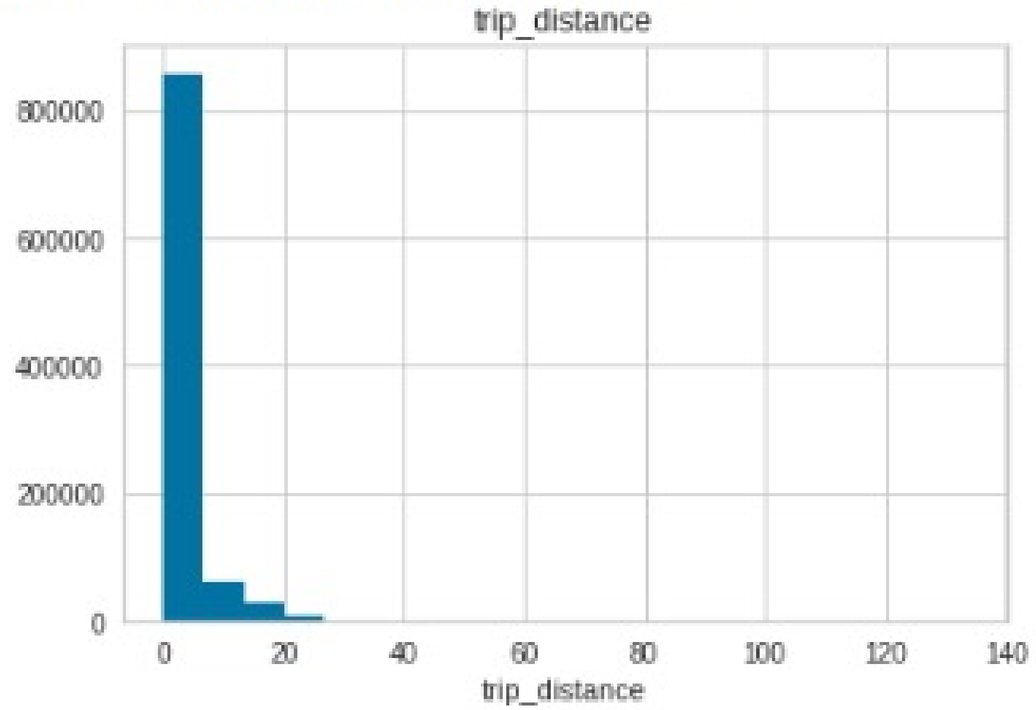




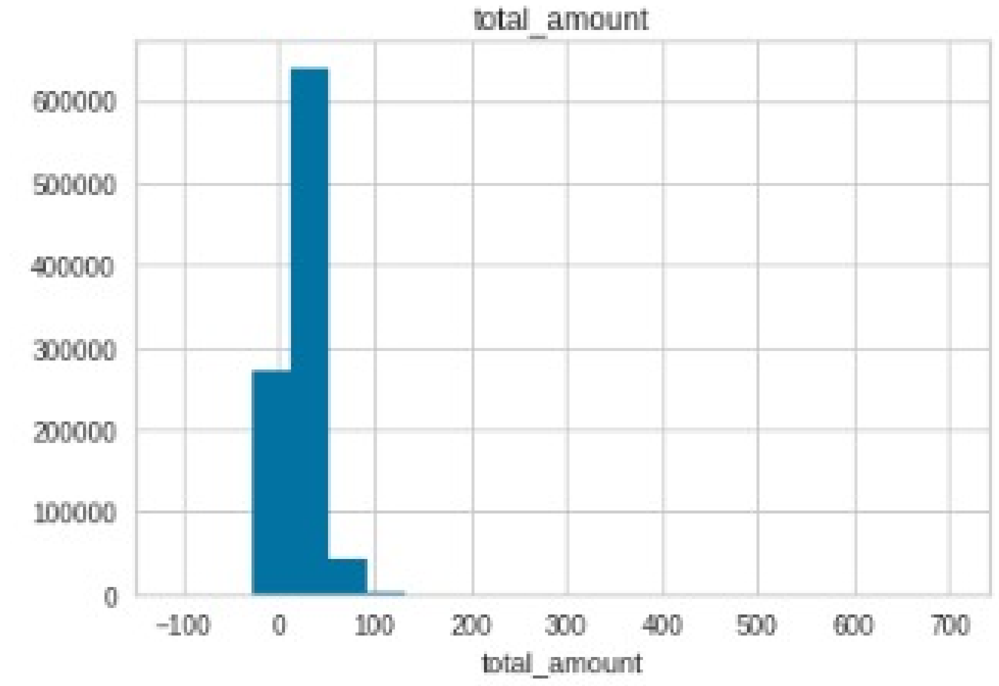


Data Analizi - Numerik Değişkenler

```
count   956,824.0000
mean          2.9620
std           3.8622
min           0.0000
1%            0.1000
5%            0.4800
10%           0.6100
20%           0.8600
30%           1.1000
40%           1.3200
50%           1.6200
60%           2.0000
70%           2.6000
80%           3.6400
90%           6.9000
95%          11.0800
99%          19.1300
100%        133.5200
max         133.5200
Name: trip_distance, dtype: float64
```

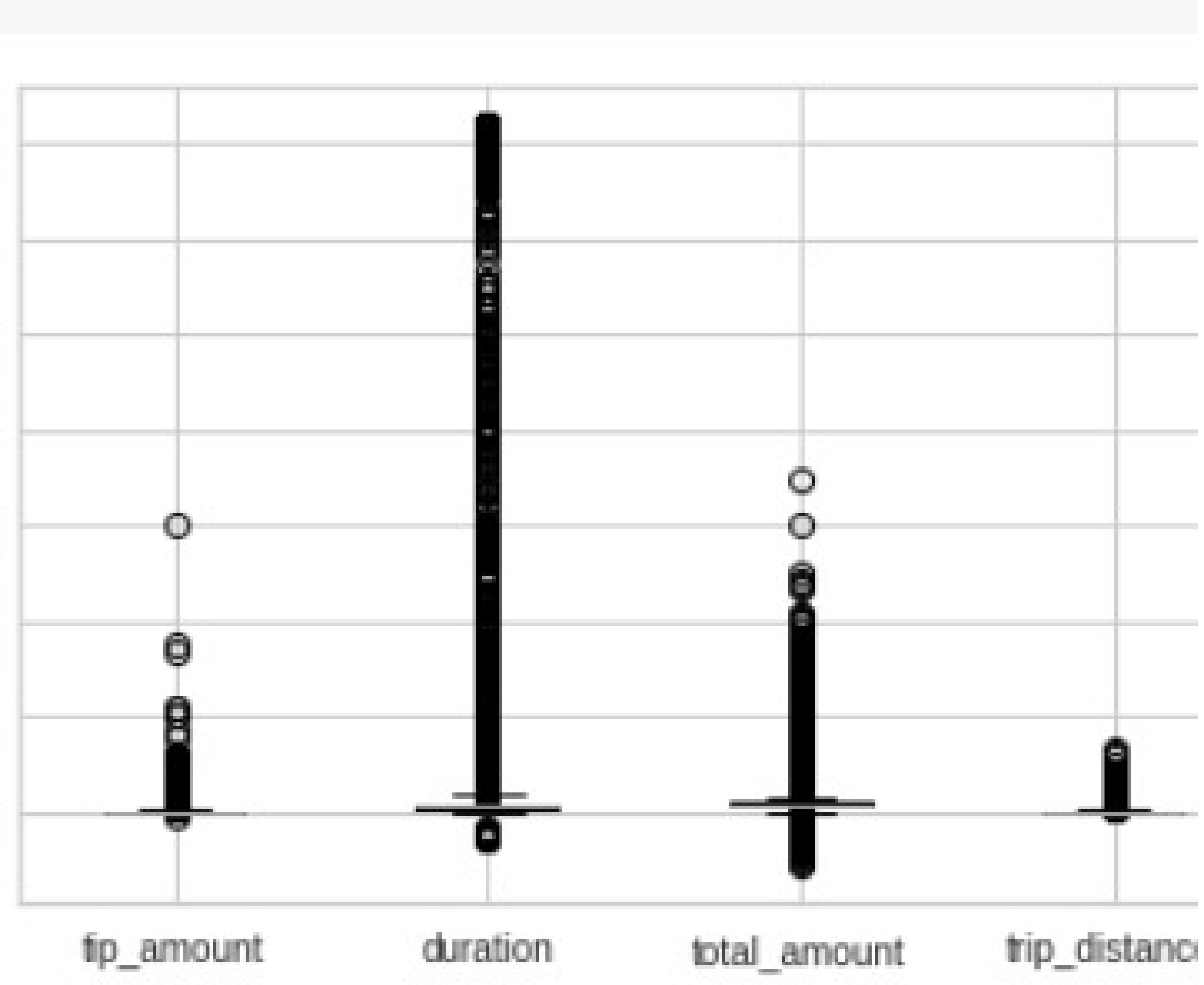



```
count   956,824.0000
mean         18.9192
std          14.5089
min        -109.9200
1%            5.8000
5%            7.8000
10%           8.8000
20%          10.3800
30%          11.7600
40%          12.9600
50%          14.6300
60%          16.3000
70%          18.8000
80%          22.8000
90%          33.3500
95%          50.4700
99%          73.9200
100%        700.3000
max         700.3000
Name: total_amount, dtype: float64
```



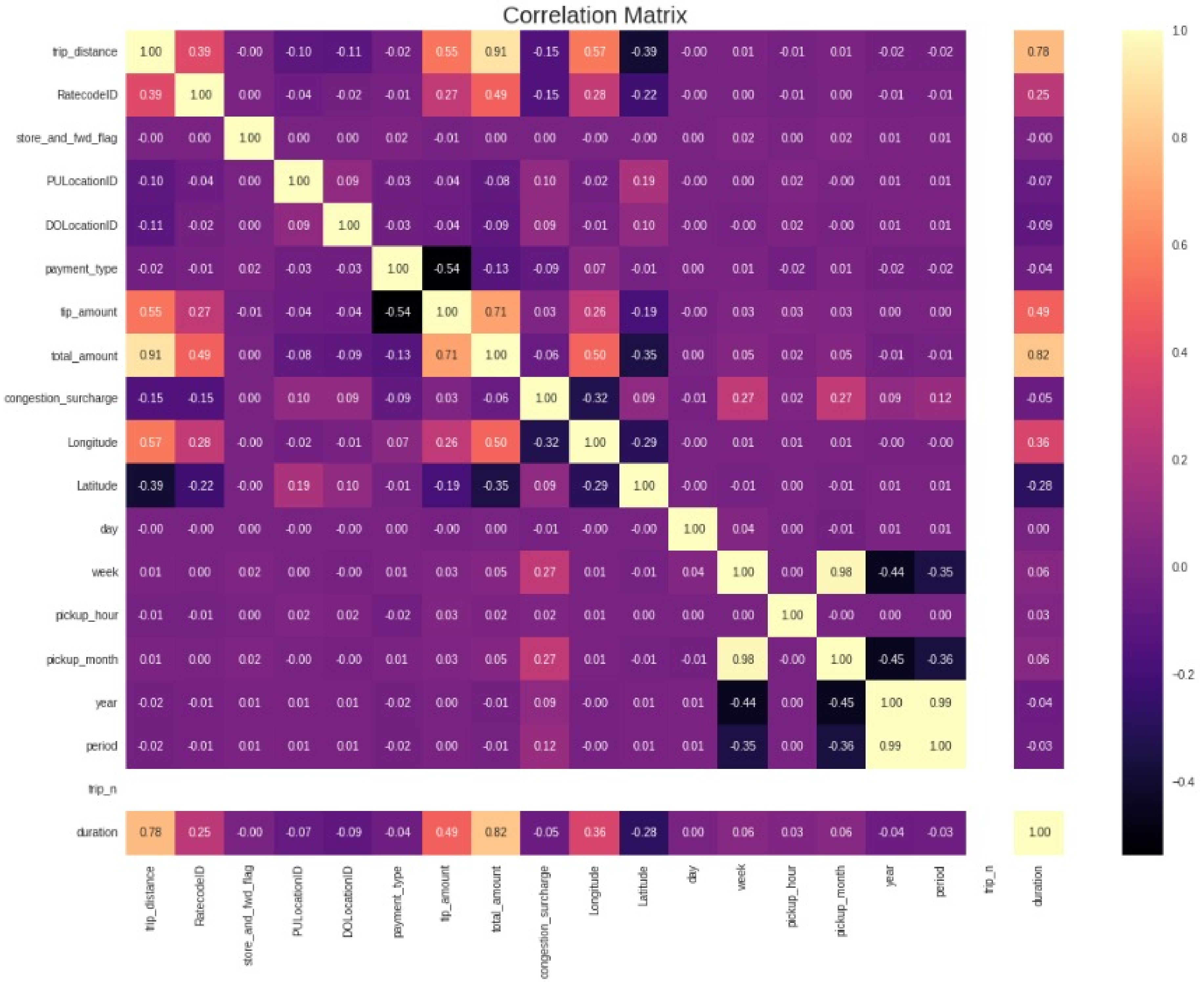
Outlier Analizi

Outlier grafikleri çizdirilerek 0.01 ve 0.99 quantile limitleri dışındaki değerler datanın dışında bırakıldı



Korelasyon Analizi

Korelasyon analizi sonrasında yüksek korelasyonu olan week, total_amount,tip_amount ve period değişkenleri datadan çıkarıldı

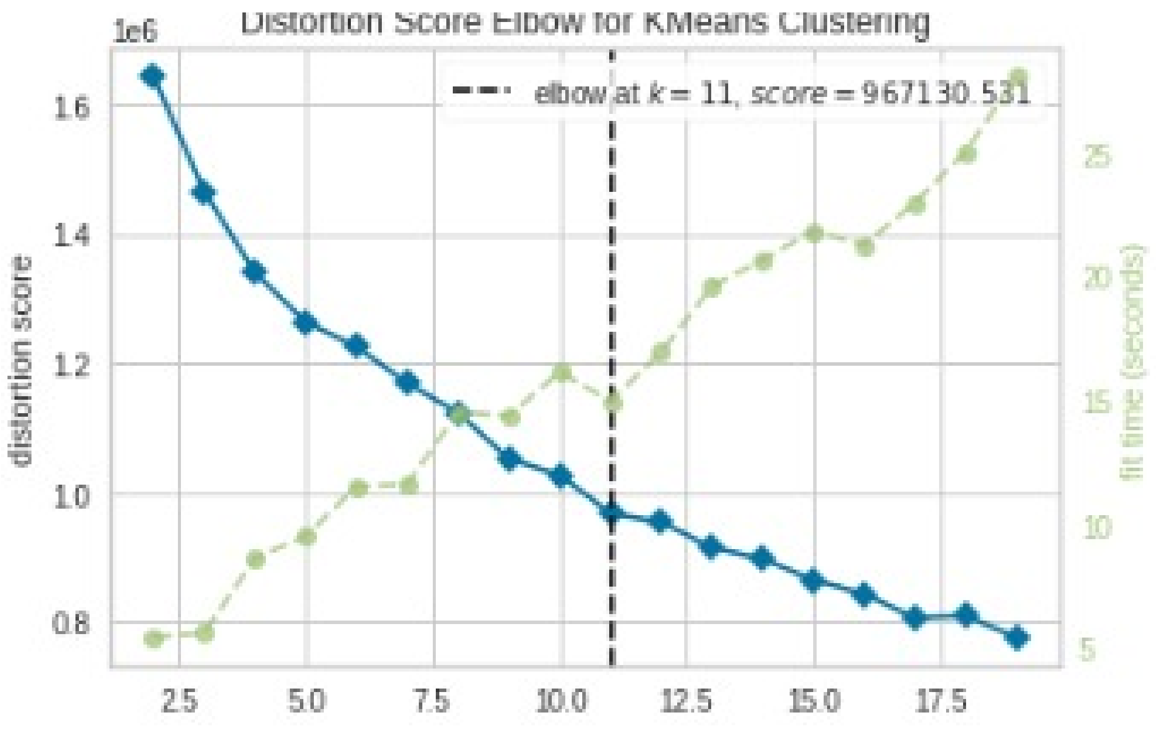


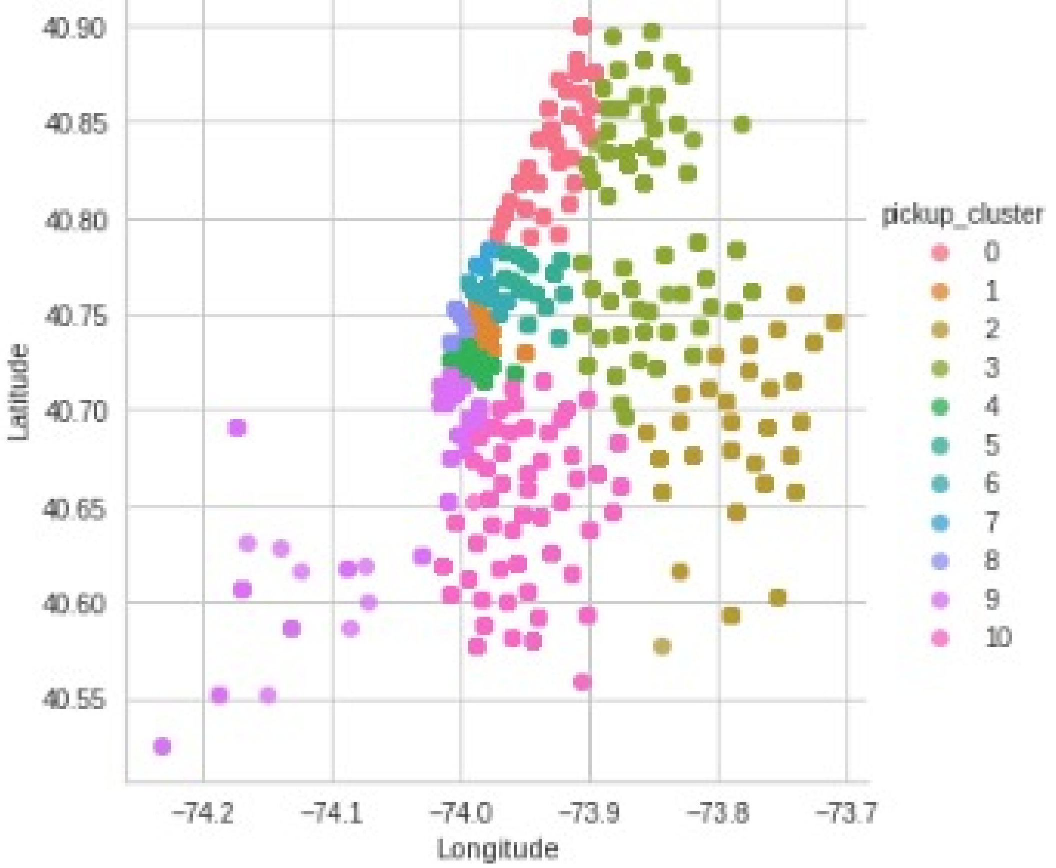
KMeans

Bölge Gruplaması

Bölgeleri cluster edebilmek amacı ile Kmeans modeli uygulandı

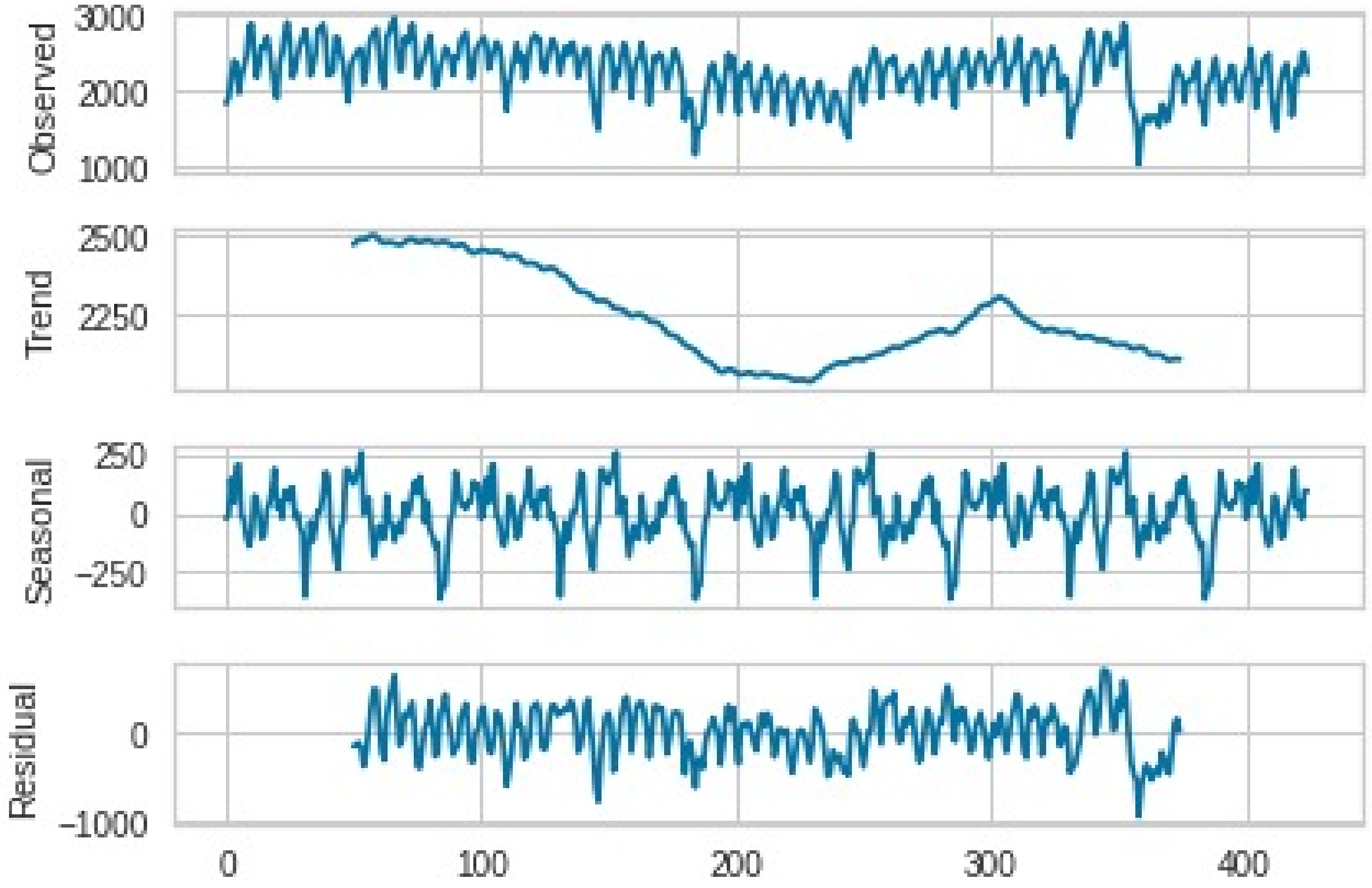
- One-hot encoding
- Min-Max Scaler
- Elbow analizi yapılarak optimum bölünme sayısı 11 olarak bulundu.





Time Series

Zaman Serileri ile
prediction




```
cols = [col for col in df3.columns if col not in df3[['trip_n',"tpep_pickup_date","geo"]]]
```

```
train = df3.loc[(df3["tpep_pickup_date"] < "2020-01-01"), :]

val = df3.loc[(df3["tpep_pickup_date"] >= "2020-01-01"), :]
```

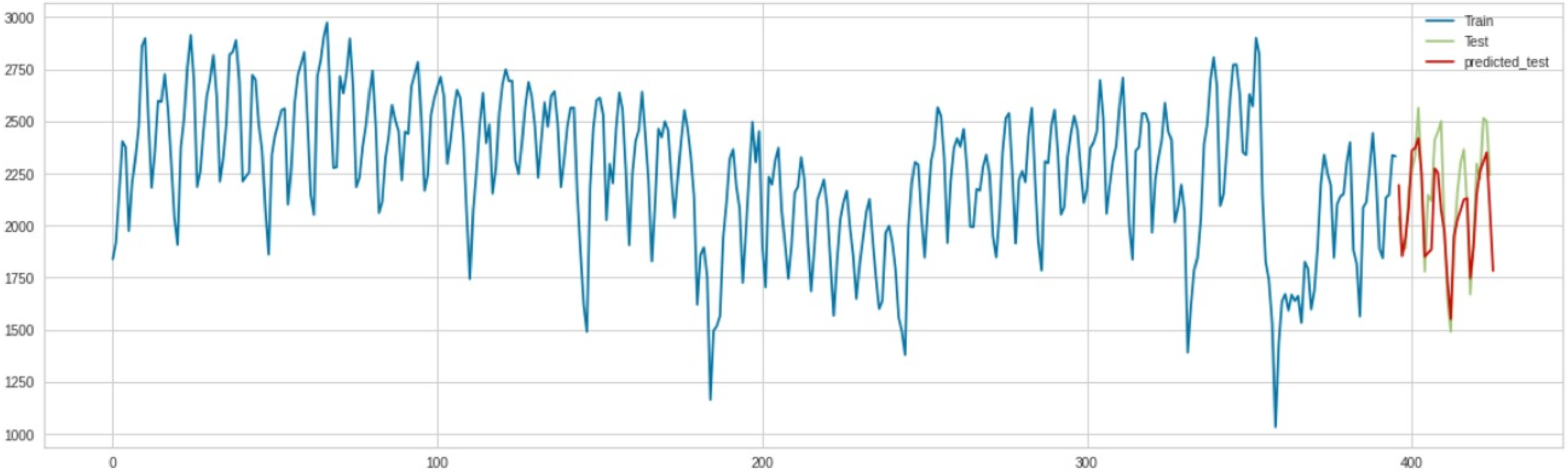
```
Y_train = df3['trip_n']
X_train = df3[cols]
```

```
Y_val = val['trip_n']
X_val = val[cols]
```

Light GBM Model Sonucu

```
Training until validation scores don't improve for 50 rounds.
[50]    training's mape: 0.511865       valid_1's mape: 0.666186
[100]   training's mape: 0.493037       valid_1's mape: 0.656036
[150]   training's mape: 0.486948       valid_1's mape: 0.650123
Early stopping, best iteration is:
[140]   training's mape: 0.487598       valid_1's mape: 0.648818
CPU times: user 6.36 s, sys: 2.31 s, total: 8.67 s
Wall time: 4.28 s
```

TES YÖNTEMİ



Datanın trend, seasonality ve artıkları incelenerek sonrasında TES yönetmi ve LightGBM modeli kuruldu.

Bu yöntemler üzerinden tahminlere bakıldığında Light GBM "mape" oranlarını yüksek olduğu görülürken TES yönteminde tahminin gözleme daha yakın olduğu tespit edildi.

TEŞEKKÜRLER...