Delhivery_BCS

September 19, 2024

1 About Delhivery

Delhivery is the largest and fastest-growing fully integrated player in India by revenue in Fiscal 2021. They aim to build the operating system for commerce, through a combination of world-class infrastructure, logistics operations of the highest quality, and cutting-edge engineering and technology capabilities.

The Data team builds intelligence and capabilities using this data that helps them to widen the gap between the quality, efficiency, and profitability of their business versus their competitors.

2 Problem Statement

We have been given data on trips performed by parcels for Delhivery, which has attributes like trip_creation_time, routes, source and destination places, and open-source routing engine time. We need to clean, sanitize and manipulate data and get useful features and provide data to help them build forecasting models.

Importing required Python Libraries

[3]:
```
import numpy as np
import pandas as pd
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels
from scipy.special import comb
from scipy.stats import binom
from scipy.stats import norm,t
from scipy.stats import poisson, expon,geom, ttest_1samp,␣
 ↪ttest_ind,ttest_ind_from_stats,boxcox
from scipy.stats import shapiro, levene, kruskal, chi2,␣
 ↪chi2_contingency,pearsonr, spearmanr
from statsmodels.graphics.gofplots import qqplot
from sklearn.preprocessing import LabelEncoder, StandardScaler, MinMaxScaler,␣
 ↪OneHotEncoder
from warnings import filterwarnings
filterwarnings('ignore')
```

```
[4]: df = pd.read_csv('delhivery_data.csv')
```

Observations

on shape of data, data types of all the attributes, conversion of categorical attributes to ‘category’, missing value detection, statistical summary

```
[8]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 144867 entries, 0 to 144866
Data columns (total 24 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   data                            144867 non-null  object
 1   trip_creation_time              144867 non-null  object
 2   route_schedule_uuid             144867 non-null  object
 3   route_type                      144867 non-null  object
 4   trip_uuid                       144867 non-null  object
 5   source_center                   144867 non-null  object
 6   source_name                     144574 non-null  object
 7   destination_center              144867 non-null  object
 8   destination_name                144606 non-null  object
 9   od_start_time                   144867 non-null  object
 10  od_end_time                     144867 non-null  object
 11  start_scan_to_end_scan          144867 non-null  float64
 12  is_cutoff                       144867 non-null  bool
 13  cutoff_factor                   144867 non-null  int64
 14  cutoff_timestamp                144867 non-null  object
 15  actual_distance_to_destination  144867 non-null  float64
 16  actual_time                     144867 non-null  float64
 17  osrm_time                       144867 non-null  float64
 18  osrm_distance                   144867 non-null  float64
 19  factor                          144867 non-null  float64
 20  segment_actual_time             144867 non-null  float64
 21  segment_osrm_time               144867 non-null  float64
 22  segment_osrm_distance           144867 non-null  float64
 23  segment_factor                  144867 non-null  float64
dtypes: bool(1), float64(10), int64(1), object(12)
memory usage: 25.6+ MB
```

```
[9]: df.shape
```

```
[9]: (144867, 24)
```

```
[10]: df.head()
```

```
[10]:        data          trip_creation_time  \
      0  training  2018-09-20 02:35:36.476840
      1  training  2018-09-20 02:35:36.476840
      2  training  2018-09-20 02:35:36.476840
      3  training  2018-09-20 02:35:36.476840
      4  training  2018-09-20 02:35:36.476840

                                   route_schedule_uuid route_type  \
      0  thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3…    Carting
      1  thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3…    Carting
      2  thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3…    Carting
      3  thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3…    Carting
      4  thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3…    Carting

                       trip_uuid source_center                 source_name  \
      0  trip-153741093647649320  IND388121AAA  Anand_VUNagar_DC (Gujarat)
      1  trip-153741093647649320  IND388121AAA  Anand_VUNagar_DC (Gujarat)
      2  trip-153741093647649320  IND388121AAA  Anand_VUNagar_DC (Gujarat)
      3  trip-153741093647649320  IND388121AAA  Anand_VUNagar_DC (Gujarat)
      4  trip-153741093647649320  IND388121AAA  Anand_VUNagar_DC (Gujarat)

        destination_center               destination_name  \
      0       IND388620AAB  Khambhat_MotvdDPP_D (Gujarat)
      1       IND388620AAB  Khambhat_MotvdDPP_D (Gujarat)
      2       IND388620AAB  Khambhat_MotvdDPP_D (Gujarat)
      3       IND388620AAB  Khambhat_MotvdDPP_D (Gujarat)
      4       IND388620AAB  Khambhat_MotvdDPP_D (Gujarat)

                      od_start_time  …            cutoff_timestamp  \
      0  2018-09-20 03:21:32.418600  …         2018-09-20 04:27:55
      1  2018-09-20 03:21:32.418600  …         2018-09-20 04:17:55
      2  2018-09-20 03:21:32.418600  …  2018-09-20 04:01:19.505586
      3  2018-09-20 03:21:32.418600  …         2018-09-20 03:39:57
      4  2018-09-20 03:21:32.418600  …         2018-09-20 03:33:55

         actual_distance_to_destination  actual_time  osrm_time osrm_distance  \
      0                       10.435660         14.0       11.0       11.9653
      1                       18.936842         24.0       20.0       21.7243
      2                       27.637279         40.0       28.0       32.5395
      3                       36.118028         62.0       40.0       45.5620
      4                       39.386040         68.0       44.0       54.2181

           factor  segment_actual_time  segment_osrm_time  segment_osrm_distance  \
      0  1.272727                 14.0               11.0                11.9653
      1  1.200000                 10.0                9.0                 9.7590
      2  1.428571                 16.0                7.0                10.8152
      3  1.550000                 21.0               12.0                13.0224
```

```
4  1.545455                  6.0                5.0                 3.9153

   segment_factor
0        1.272727
1        1.111111
2        2.285714
3        1.750000
4        1.200000

[5 rows x 24 columns]
```

```
[11]: df.nunique() # number of unique values in columns
```

```
[11]: data                                   2
      trip_creation_time                 14817
      route_schedule_uuid                 1504
      route_type                             2
      trip_uuid                          14817
      source_center                       1508
      source_name                         1498
      destination_center                  1481
      destination_name                    1468
      od_start_time                      26369
      od_end_time                        26369
      start_scan_to_end_scan              1915
      is_cutoff                              2
      cutoff_factor                        501
      cutoff_timestamp                   93180
      actual_distance_to_destination    144515
      actual_time                         3182
      osrm_time                           1531
      osrm_distance                     138046
      factor                             45641
      segment_actual_time                  747
      segment_osrm_time                    214
      segment_osrm_distance             113799
      segment_factor                      5675
      dtype: int64
```

```
[12]: df.isna().sum() #missing values in columns
```

```
[12]: data                                0
      trip_creation_time                  0
      route_schedule_uuid                 0
      route_type                          0
      trip_uuid                           0
      source_center                       0
```

```
source_name                       293
destination_center                  0
destination_name                  261
od_start_time                       0
od_end_time                         0
start_scan_to_end_scan              0
is_cutoff                           0
cutoff_factor                       0
cutoff_timestamp                    0
actual_distance_to_destination      0
actual_time                         0
osrm_time                           0
osrm_distance                       0
factor                              0
segment_actual_time                 0
segment_osrm_time                   0
segment_osrm_distance               0
segment_factor                      0
dtype: int64
```

```
[13]: df.describe() #Statistical summary of the dataset
```

```
[13]:        start_scan_to_end_scan  cutoff_factor  actual_distance_to_destination  \
      count           144867.000000  144867.000000                   144867.000000
      mean               961.262986     232.926567                      234.073372
      std               1037.012769     344.755577                      344.990009
      min                 20.000000       9.000000                        9.000045
      25%                161.000000      22.000000                       23.355874
      50%                449.000000      66.000000                       66.126571
      75%               1634.000000     286.000000                      286.708875
      max               7898.000000    1927.000000                     1927.447705

               actual_time      osrm_time  osrm_distance         factor  \
      count  144867.000000  144867.000000  144867.000000  144867.000000
      mean      416.927527     213.868272     284.771297       2.120107
      std       598.103621     308.011085     421.119294       1.715421
      min         9.000000       6.000000       9.008200       0.144000
      25%        51.000000      27.000000      29.914700       1.604264
      50%       132.000000      64.000000      78.525800       1.857143
      75%       513.000000     257.000000     343.193250       2.213483
      max      4532.000000    1686.000000    2326.199100      77.387097

             segment_actual_time  segment_osrm_time  segment_osrm_distance  \
      count        144867.000000      144867.000000           144867.00000
      mean             36.196111          18.507548               22.82902
      std              53.571158          14.775960               17.86066
      min            -244.000000           0.000000                0.00000
```

```
25%              20.000000          11.000000               12.07010
50%              29.000000          17.000000               23.51300
75%              40.000000          22.000000               27.81325
max            3051.000000        1611.000000             2191.40370

       segment_factor
count   144867.000000
mean         2.218368
std          4.847530
min        -23.444444
25%          1.347826
50%          1.684211
75%          2.250000
max        574.250000
```

[14]:

```
df.describe(include=object)
```

[14]:

```
            data          trip_creation_time  \
count     144867                      144867
unique         2                       14817
top     training  2018-09-28 05:23:15.359220
freq      104858                         101

                                      route_schedule_uuid route_type  \
count                                              144867     144867
unique                                               1504          2
top     thanos::sroute:4029a8a2-6c74-4b7e-a6d8-f9e069f…        FTL
freq                                                 1812      99660

                      trip_uuid source_center                    source_name  \
count                    144867        144867                         144574
unique                    14817          1508                           1498
top     trip-153811219535896559  IND000000ACB  Gurgaon_Bilaspur_HB (Haryana)
freq                        101         23347                          23347

       destination_center               destination_name  \
count              144867                         144606
unique               1481                           1468
top          IND000000ACB  Gurgaon_Bilaspur_HB (Haryana)
freq                15192                          15192

                     od_start_time                 od_end_time  \
count                       144867                      144867
unique                       26369                       26369
top     2018-09-21 18:37:09.322207  2018-09-24 09:59:15.691618
freq                            81                          81
```

```
           cutoff_timestamp
count                144867
unique                93180
top     2018-09-24 05:19:20
freq                     40
```

[15]:
```
#Checking for source center value for which source name is null
df[(df["source_name"].notnull()) & (df["source_center"].
 ↪isin(df[df["source_name"].isnull()]["source_center"]))]
```

[15]:
```
Empty DataFrame
Columns: [data, trip_creation_time, route_schedule_uuid, route_type, trip_uuid,
source_center, source_name, destination_center, destination_name, od_start_time,
od_end_time, start_scan_to_end_scan, is_cutoff, cutoff_factor, cutoff_timestamp,
actual_distance_to_destination, actual_time, osrm_time, osrm_distance, factor,
segment_actual_time, segment_osrm_time, segment_osrm_distance, segment_factor]
Index: []

[0 rows x 24 columns]
```

[16]:
```
#Checking for destination center value for which destination name is null
df[(df["destination_name"].notnull()) & (df["destination_center"].
 ↪isin(df[df["destination_name"].isnull()]["destination_center"]))]
```

[16]:
```
Empty DataFrame
Columns: [data, trip_creation_time, route_schedule_uuid, route_type, trip_uuid,
source_center, source_name, destination_center, destination_name, od_start_time,
od_end_time, start_scan_to_end_scan, is_cutoff, cutoff_factor, cutoff_timestamp,
actual_distance_to_destination, actual_time, osrm_time, osrm_distance, factor,
segment_actual_time, segment_osrm_time, segment_osrm_distance, segment_factor]
Index: []

[0 rows x 24 columns]
```

[17]:
```
#Here we can observe that minimum value of segment_actual_time and␣
 ↪segment_factor is negative,
#which seems false values as time can not be negative, so we will drop that data
df.drop(df[df["segment_actual_time"]<0].index, inplace=True)
```

[18]:
```
df.describe()
```

[18]:
```
       start_scan_to_end_scan  cutoff_factor  actual_distance_to_destination  \
count           144846.000000  144846.000000                   144846.000000
mean               961.226537     232.911057                      234.057171
std               1036.993595     344.740981                      344.974984
min                 20.000000       9.000000                        9.000045
25%                161.000000      22.000000                       23.354927
```

```
50%                449.000000      66.000000                        66.126234
75%               1634.000000     286.000000                       286.706673
max               7898.000000    1927.000000                      1927.447705

         actual_time      osrm_time  osrm_distance         factor  \
count  144846.000000  144846.000000  144846.000000  144846.000000
mean      416.908724     213.853002     284.750969       2.120190
std       598.085058     307.997702     421.101831       1.715508
min         9.000000       6.000000       9.008200       0.144000
25%        51.000000      27.000000      29.909925       1.604288
50%       132.000000      64.000000      78.524600       1.857143
75%       513.000000     257.000000     343.062075       2.213589
max      4532.000000    1686.000000    2326.199100      77.387097

       segment_actual_time  segment_osrm_time  segment_osrm_distance  \
count        144846.000000      144846.000000          144846.000000
mean             36.207427          18.507304              22.828528
std              53.561259          14.775870              17.860268
min               0.000000           0.000000               0.000000
25%              20.000000          11.000000              12.070100
50%              29.000000          17.000000              23.513000
75%              40.000000          22.000000              27.812975
max            3051.000000        1611.000000            2191.403700

       segment_factor
count   144846.000000
mean         2.219084
std          4.847144
min         -1.000000
25%          1.347826
50%          1.684211
75%          2.250000
max        574.250000
```

Dataset Information data: It contains whether the data is testing or training type

trip_creation_time: It is the timestamp of trip_creation. It ranges from ‘2018-09-12 00:25:19.499696’ to ‘2018-10-03 23:59:42.701692’

oute_schedule_uuid: it is unique_id for particular route schedule

route type: It contains whether the route is Full Truck Load or Carting type

trip_uuid: It is a unique id associated with a particular trip

source_center: It is the ID of the origin of the trip

source_name: Its the name of the origin of the trip

destination_center: It is the ID of the destination of the trip destination_name: It is the name of the destination of the trip

od_start_time: It is the trip start time

od_end_time: It is the trip end time

Start_scan_to_end_scan: It gives the time taken to deliver from source to destination. It ranges from 20 to 7898.

is_cutoff: It is an unknown field, which is boolean

cutoff_factor: It is the rounded value of the actual_distance_to_destination, it ranges from 9 to 1927

cutoff_timestamp: It is an unknown field

actual_distance_to_destination: It is the distance between the source and destination warehouses, it ranges from 9.00 to 1927.44

actual_time: It contains the actual time taken to complete the delivery (cumulative), it ranges from 9 to 4532.

osrm_time: It is an open-source routing engine time calculator which computes the shortest path between points in a given map and gives the time (cumulative), it ranges from 6 to 1686

osrm_distance: It contains the distance to the destination based on osrm, it ranges from 9.00 to 2326.199

factor: It is a ratio of actual_time to osrm_time, it ranges from 0.144 to 77.38.

segment_actual_time: It is a segment time, a time taken by a subset of package delivery, It ranges from -244 to 3051

segment_osrm_time: It contains the orsm time taken by a subset of the package delivery. It ranges from 0 to 1611

segment_osrm_distance: It contains OSRM distance, the distance covered by a subset of package delivery, it ranges from 0 to 2191.40

segment_factor: It is a ratio between segment_actual_time to segment_osrm_time, it ranges from -23.544 to 574.25

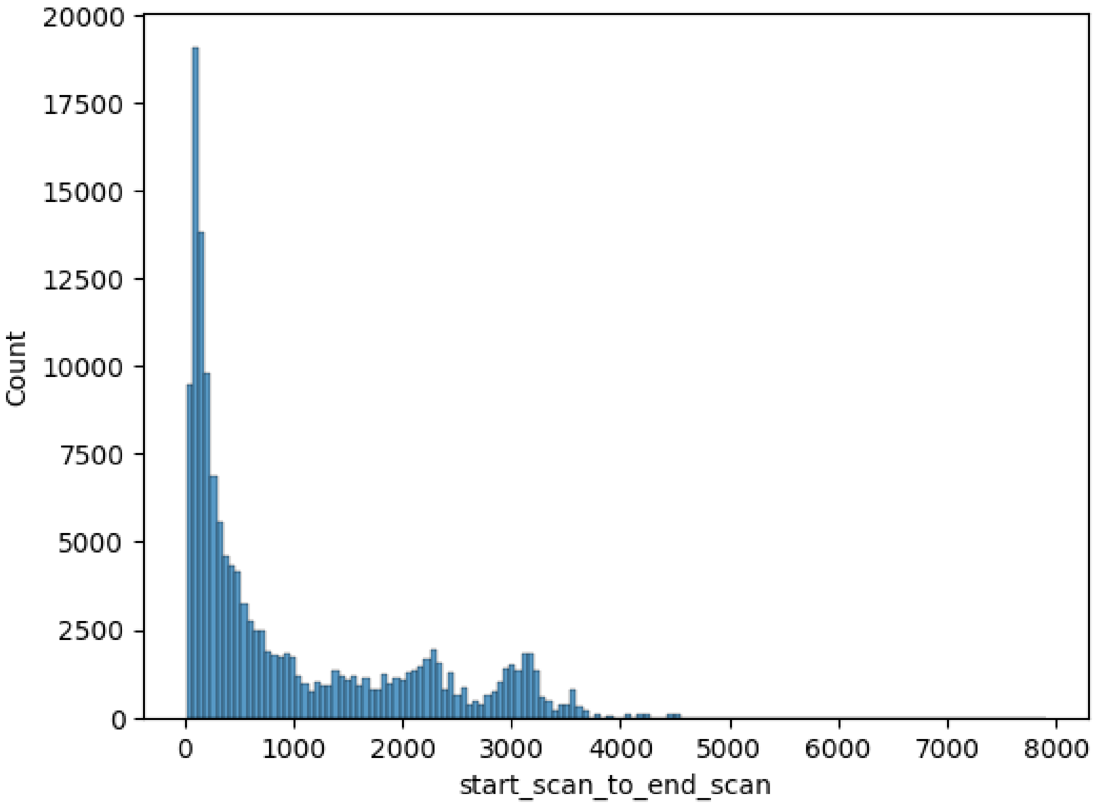
3 Univariate Analysis

[19]:

```
#Histplot for start_to_scan_to_end_scan attribute
sns.histplot(df["start_scan_to_end_scan"])
```

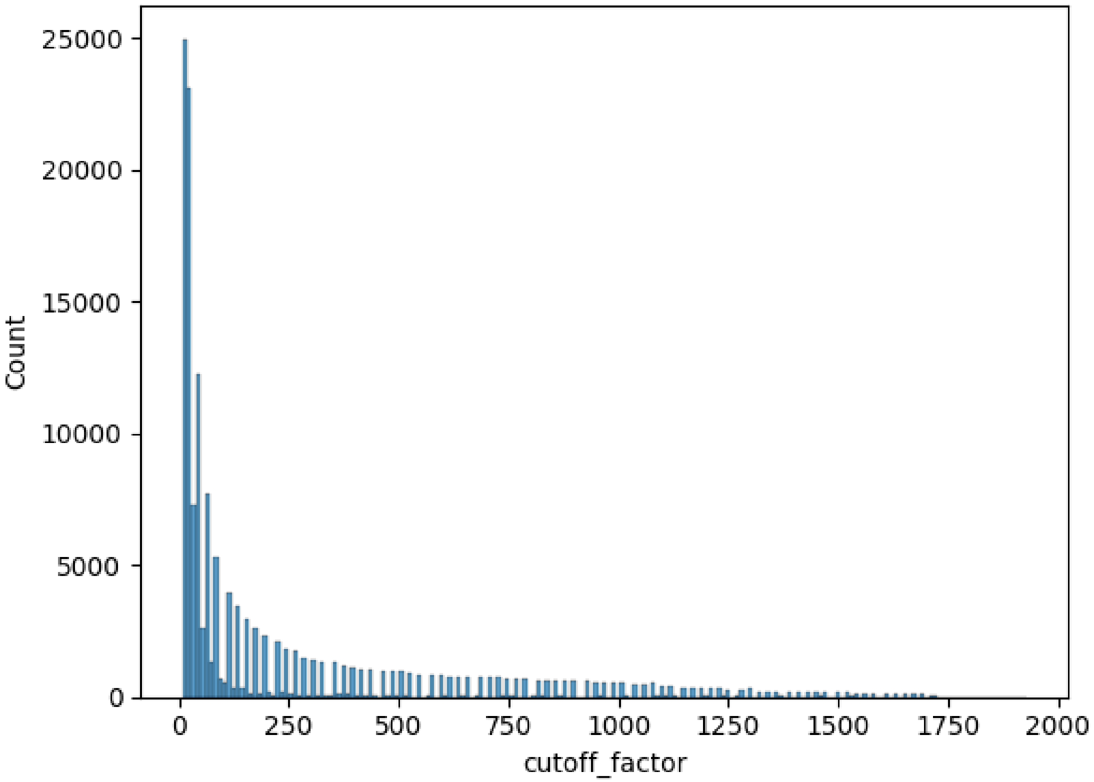
[19]:

```
<Axes: xlabel='start_scan_to_end_scan', ylabel='Count'>
```



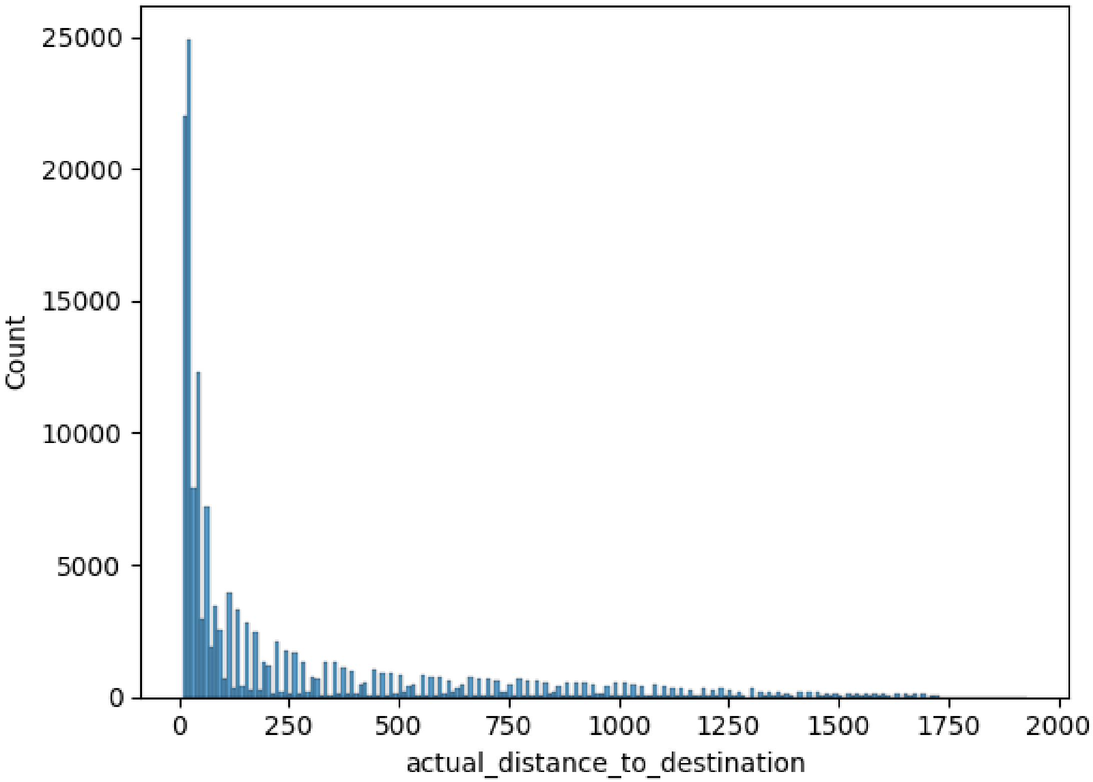
```
[20]: #Histplot for cutoff_factor attribute
      sns.histplot(df["cutoff_factor"])
```

```
[20]: <Axes: xlabel='cutoff_factor', ylabel='Count'>
```



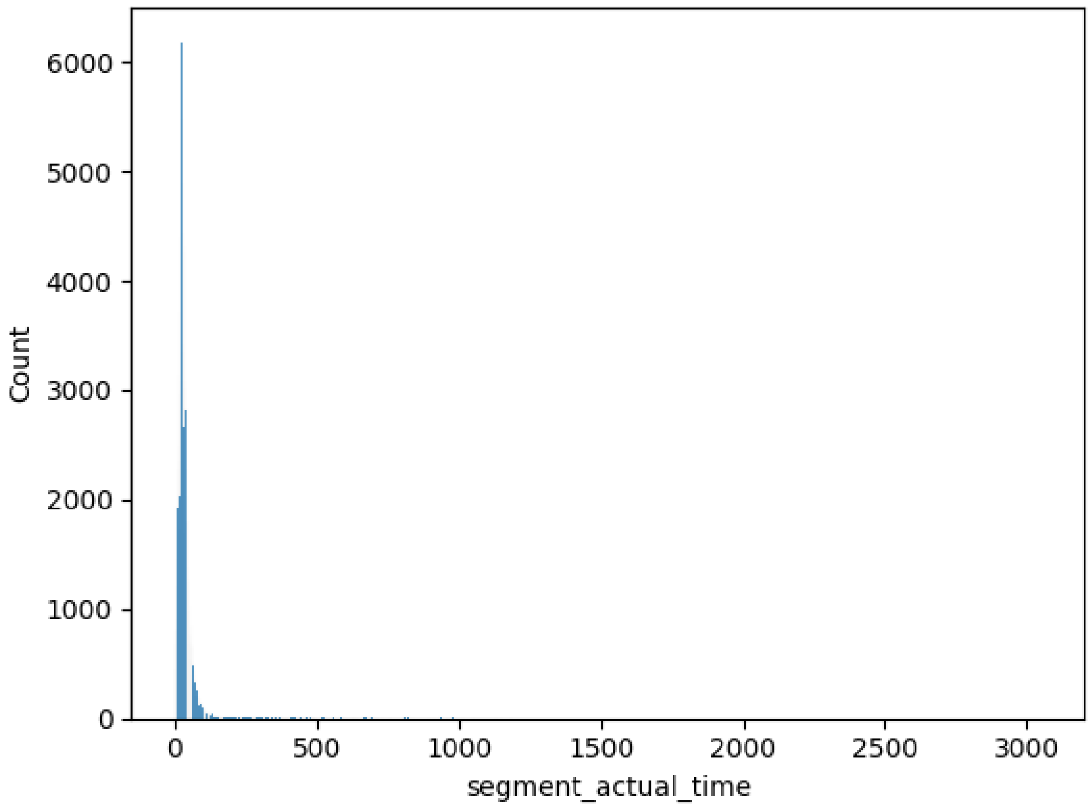
```
[21]: #Histplot for actual_distance_to_destination attribute
      sns.histplot(df["actual_distance_to_destination"])
```

```
[21]: <Axes: xlabel='actual_distance_to_destination', ylabel='Count'>
```



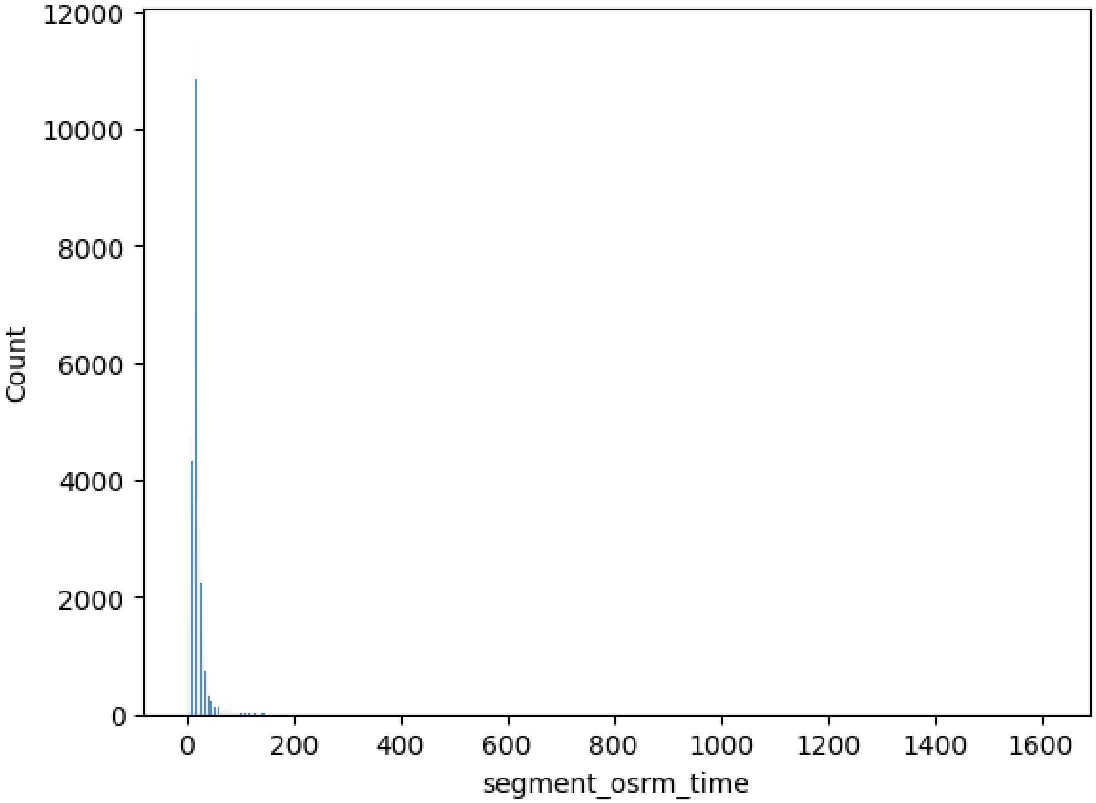
[22]:
```
#Histplot for segment_actual_time attribute
sns.histplot(df["segment_actual_time"])
```

[22]:
```
<Axes: xlabel='segment_actual_time', ylabel='Count'>
```



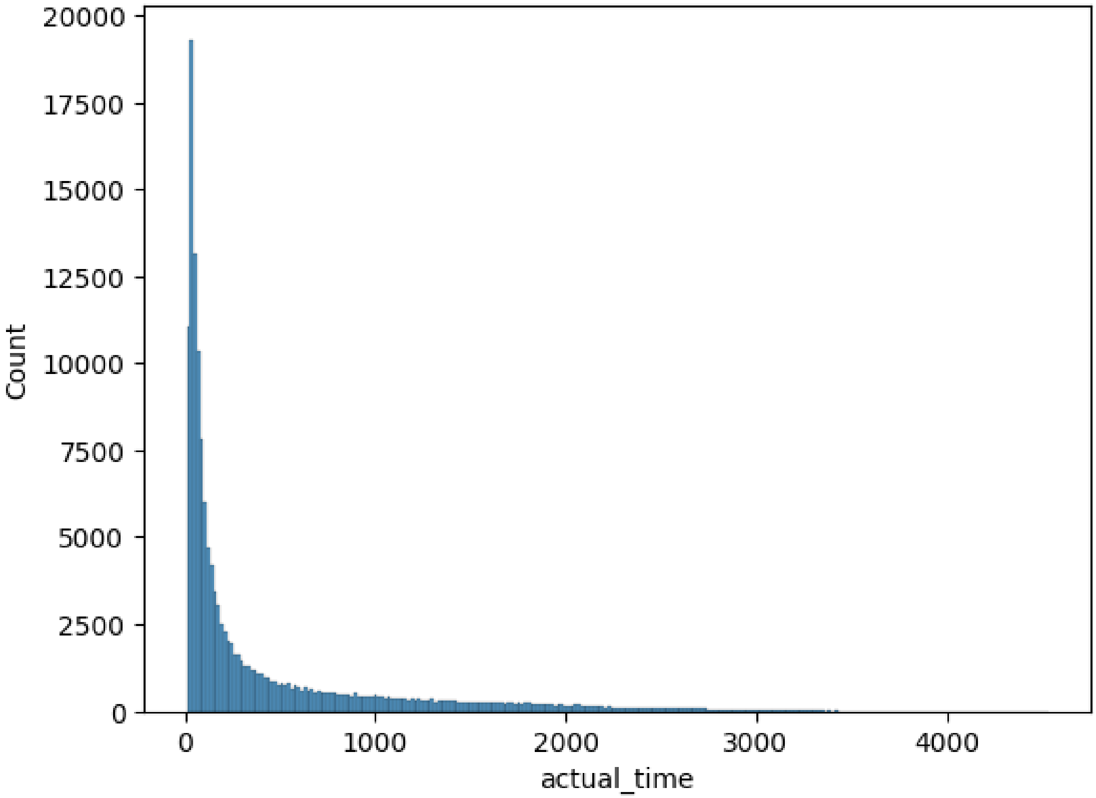
```
[23]: #Histplot for segment_osrm_time attribute
      sns.histplot(df["segment_osrm_time"])
```

```
[23]: <Axes: xlabel='segment_osrm_time', ylabel='Count'>
```



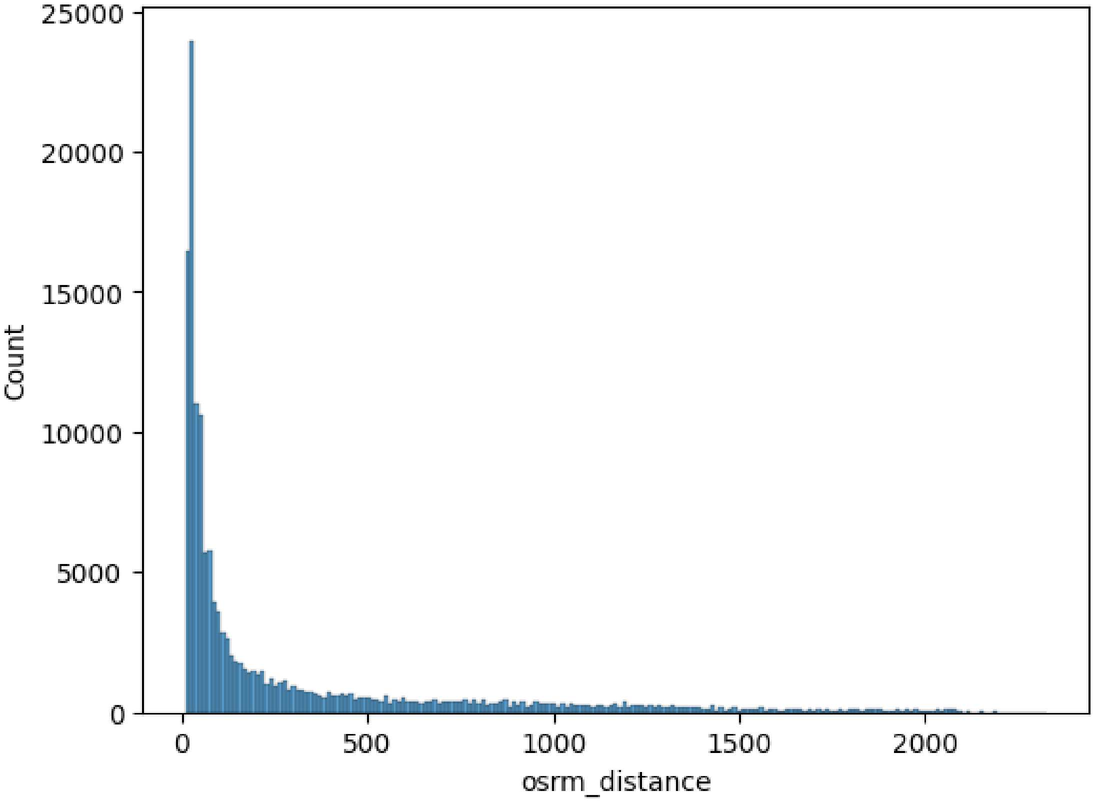
```
[24]: #Histplot for actual_time attribute
      sns.histplot(df["actual_time"])
```

```
[24]: <Axes: xlabel='actual_time', ylabel='Count'>
```



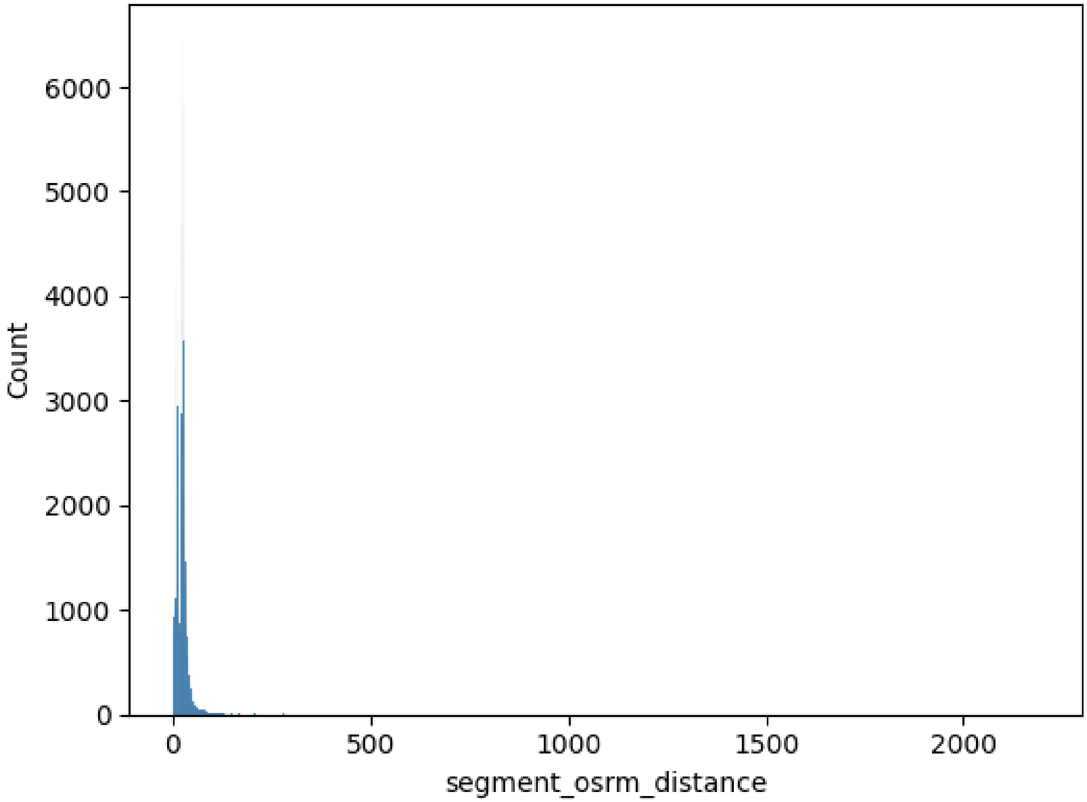
```
[25]: #Histplot for osrm_distance attribute
      sns.histplot(df["osrm_distance"])
```

```
[25]: <Axes: xlabel='osrm_distance', ylabel='Count'>
```



```
[26]: #Histplot for segment_osrm_distance attribute
      sns.histplot(df["segment_osrm_distance"])
```

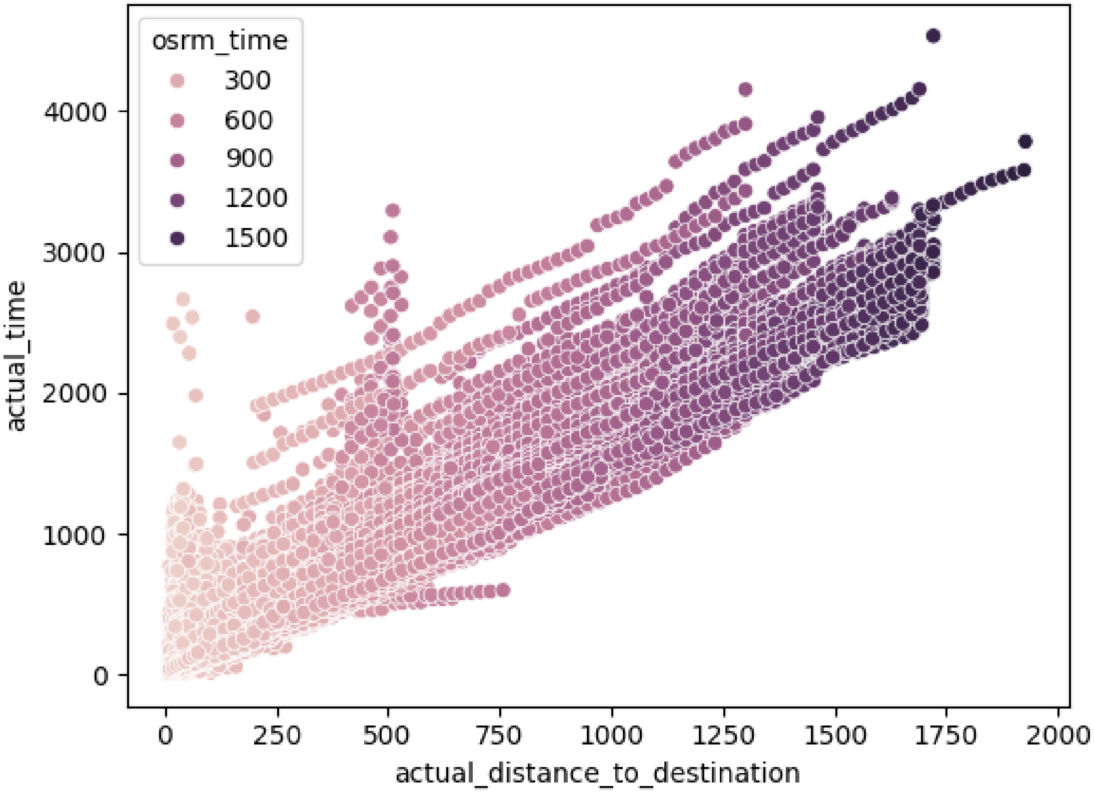
```
[26]: <Axes: xlabel='segment_osrm_distance', ylabel='Count'>
```



4 Bivariate Analysis

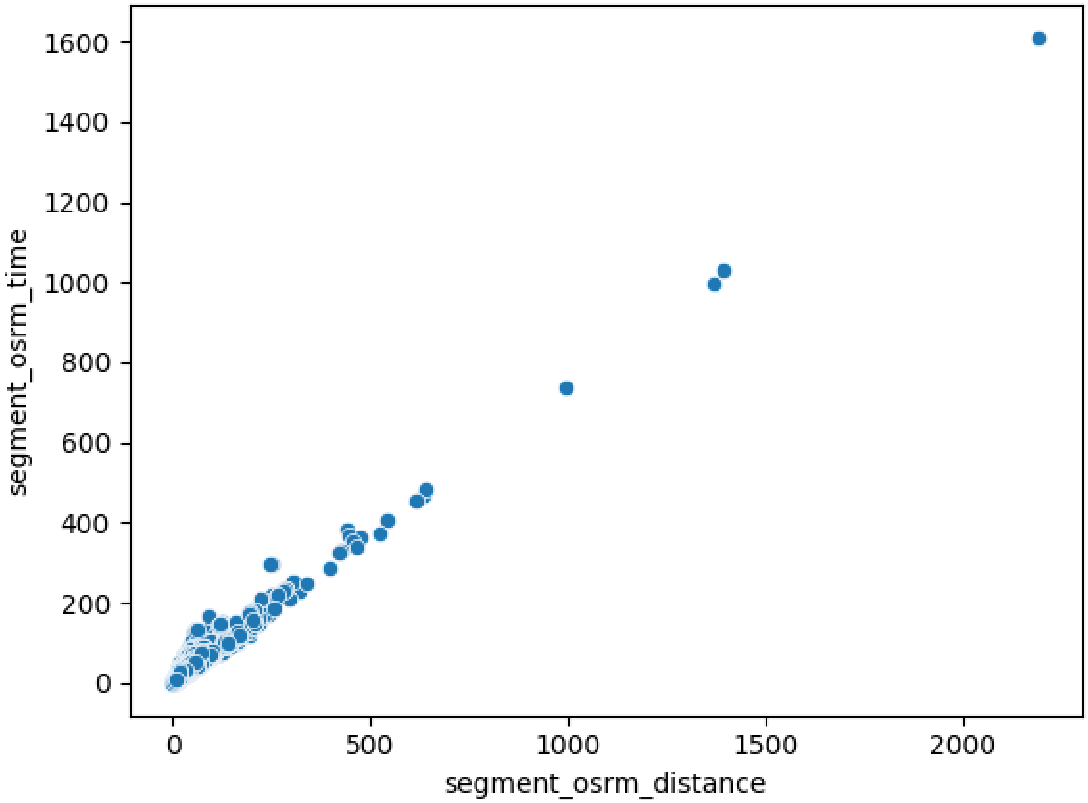
```
[27]: #Scatterplot between actual-distance_to_destination, actual_time and osrm_time
      sns.scatterplot(data=df, x="actual_distance_to_destination",y="actual_time",␣
       ↪hue="osrm_time")
```

```
[27]: <Axes: xlabel='actual_distance_to_destination', ylabel='actual_time'>
```



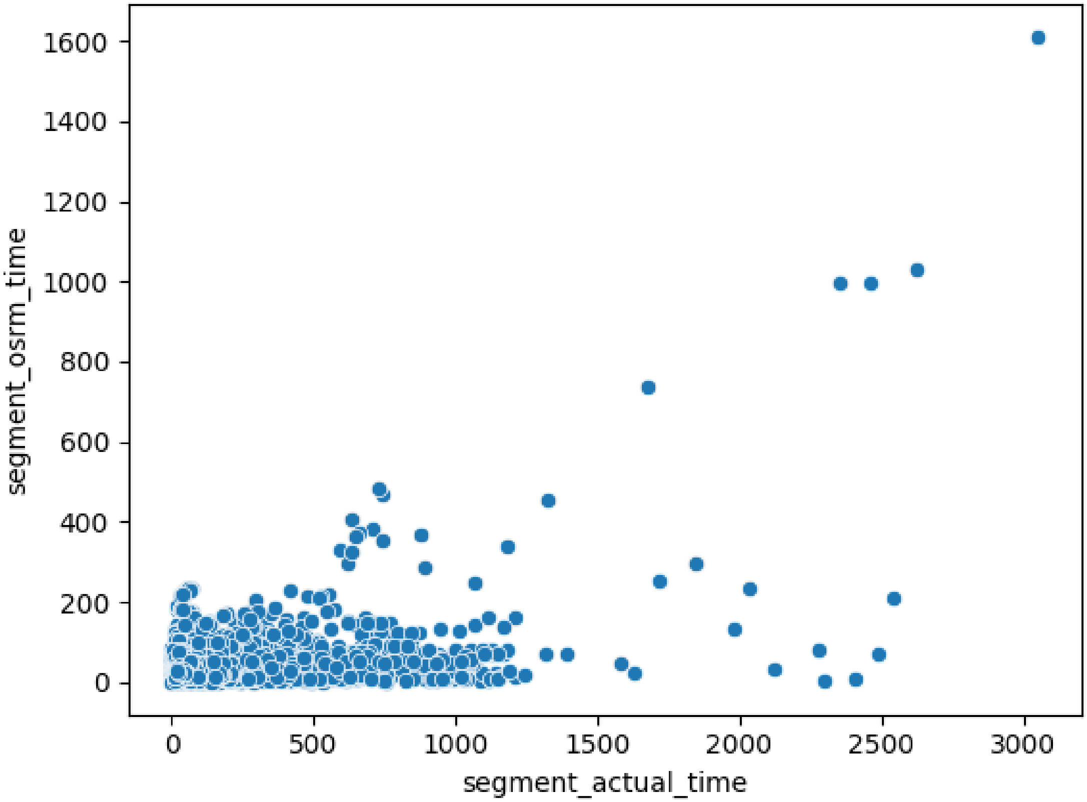
[28]:
```
#scatterplot between segment_osrm_distance and segment_osrm_time
sns.scatterplot(data=df, x="segment_osrm_distance",y="segment_osrm_time")
```

[28]:
```
<Axes: xlabel='segment_osrm_distance', ylabel='segment_osrm_time'>
```



```
[29]: #scatterplot between segment_actual_time and segment_osrm_time
      sns.scatterplot(data=df, x="segment_actual_time",y="segment_osrm_time")
```

```
[29]: <Axes: xlabel='segment_actual_time', ylabel='segment_osrm_time'>
```



[35]:
```
#Drop non-numeric columns
df_numeric = df.select_dtypes(include=['float64', 'int64'])
```

[34]:
```
df_numeric.corr()
```

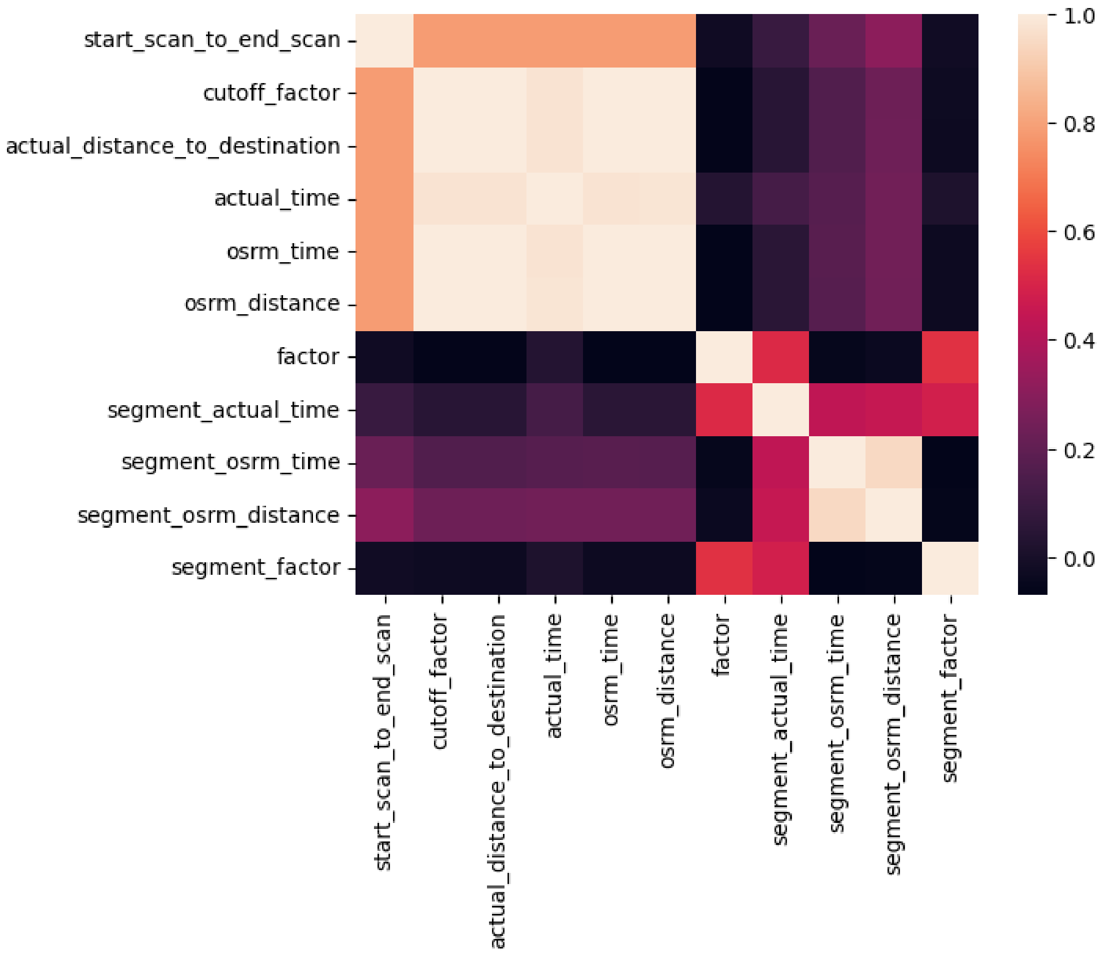
[34]:
```
                                start_scan_to_end_scan  cutoff_factor  \
start_scan_to_end_scan                        1.000000       0.784656
cutoff_factor                                 0.784656       1.000000
actual_distance_to_destination                0.784988       0.999986
actual_time                                   0.785924       0.978719
osrm_time                                     0.785283       0.995833
osrm_distance                                 0.784120       0.997116
factor                                       -0.023192      -0.064559
segment_actual_time                           0.093372       0.045063
segment_osrm_time                             0.219844       0.157942
segment_osrm_distance                         0.306972       0.231109
segment_factor                               -0.020225      -0.031439

                                actual_distance_to_destination  actual_time  \
start_scan_to_end_scan                                0.784988     0.785924
cutoff_factor                                         0.999986     0.978719
```

```
actual_distance_to_destination                        1.000000     0.978658
actual_time                                           0.978658     1.000000
osrm_time                                             0.995872     0.977996
osrm_distance                                         0.997148     0.979398
factor                                               -0.064743     0.033498
segment_actual_time                                   0.045320     0.124483
segment_osrm_time                                     0.158836     0.171480
segment_osrm_distance                                 0.232119     0.242296
segment_factor                                       -0.031588     0.017570

                                osrm_time  osrm_distance    factor  \
start_scan_to_end_scan           0.785283       0.784120 -0.023192
cutoff_factor                    0.995833       0.997116 -0.064559
actual_distance_to_destination   0.995872       0.997148 -0.064743
actual_time                      0.977996       0.979398  0.033498
osrm_time                        1.000000       0.999119 -0.069081
osrm_distance                    0.999119       1.000000 -0.065391
factor                          -0.069081      -0.065391  1.000000
segment_actual_time              0.049977       0.048787  0.518451
segment_osrm_time                0.177074       0.169157 -0.053154
segment_osrm_distance            0.242288       0.239672 -0.036724
segment_factor                  -0.033038      -0.031786  0.540448

                                segment_actual_time  segment_osrm_time  \
start_scan_to_end_scan                     0.093372           0.219844
cutoff_factor                              0.045063           0.157942
actual_distance_to_destination             0.045320           0.158836
actual_time                                0.124483           0.171480
osrm_time                                  0.049977           0.177074
osrm_distance                              0.048787           0.169157
factor                                     0.518451          -0.053154
segment_actual_time                        1.000000           0.433604
segment_osrm_time                          0.433604           1.000000
segment_osrm_distance                      0.449167           0.948520
segment_factor                             0.483699          -0.068472

                                segment_osrm_distance  segment_factor
start_scan_to_end_scan                       0.306972       -0.020225
cutoff_factor                                0.231109       -0.031439
actual_distance_to_destination               0.232119       -0.031588
actual_time                                  0.242296        0.017570
osrm_time                                    0.242288       -0.033038
osrm_distance                                0.239672       -0.031786
factor                                      -0.036724        0.540448
segment_actual_time                          0.449167        0.483699
segment_osrm_time                            0.948520       -0.068472
segment_osrm_distance                        1.000000       -0.059317
```

```
segment_factor                                    -0.059317         1.000000
```

[36]:
```
# Now ploting the heatmap
sns.heatmap(df_numeric.corr())
```

[36]:
```
<Axes: >
```



5 Data Wrangling

[37]:
```
#merging of rows based on trip_id and source and destination details
data=df.
 ↪groupby(["route_type","trip_uuid","trip_creation_time","source_center","source_name","desti
 ↪aggregate({"cutoff_factor":"max","actual_distance_to_destination":
 ↪"max","segment_actual_time":"sum", "segment_osrm_time":"sum", "actual_time":
 ↪"max", "osrm_time":"max","osrm_distance":"max","segment_osrm_distance":
 ↪"sum"}).reset_index()
```

```
data
```

[37]:

```
      route_type                trip_uuid          trip_creation_time  \
0        Carting  trip-153671042288605164  2018-09-12 00:00:22.886430
1        Carting  trip-153671042288605164  2018-09-12 00:00:22.886430
2        Carting  trip-153671046011330457  2018-09-12 00:01:00.113710
3        Carting  trip-153671055416136166  2018-09-12 00:02:34.161600
4        Carting  trip-153671055416136166  2018-09-12 00:02:34.161600
…           …                       …                          …
26218        FTL  trip-153861014185597051  2018-10-03 23:42:21.856227
26219        FTL  trip-153861023893369544  2018-10-03 23:43:58.933947
26220        FTL  trip-153861023893369544  2018-10-03 23:43:58.933947
26221        FTL  trip-153861118270144424  2018-10-03 23:59:42.701692
26222        FTL  trip-153861118270144424  2018-10-03 23:59:42.701692

      source_center                         source_name destination_center  \
0      IND561203AAB   Doddablpur_ChikaDPP_D (Karnataka)       IND562101AAA
1      IND572101AAA       Tumkur_Veersagr_I (Karnataka)       IND561203AAB
2      IND400072AAB            Mumbai Hub (Maharashtra)       IND401104AAA
3      IND600056AAA    Chennai_Poonamallee (Tamil Nadu)       IND602105AAB
4      IND600116AAB      Chennai_Porur_DPC (Tamil Nadu)       IND600056AAA
…             …                                 …                  …
26218  IND462022AAA  Bhopal_Trnsport_H (Madhya Pradesh)       IND209304AAA
26219  IND382715AAA           Kadi_KaranNGR_D (Gujarat)       IND382430AAB
26220  IND384205AAA        Mehsana_Panchot_IP (Gujarat)       IND382715AAA
26221  IND583119AAA       Sandur_WrdN1DPP_D (Karnataka)       IND583101AAA
26222  IND583201AAA                  Hospet (Karnataka)       IND583119AAA

                            destination_name               od_start_time  \
0          Chikblapur_ShntiSgr_D (Karnataka)  2018-09-12 02:03:09.655591
1          Doddablpur_ChikaDPP_D (Karnataka)  2018-09-12 00:00:22.886430
2             Mumbai_MiraRd_IP (Maharashtra)  2018-09-12 00:01:00.113710
3      Chennai_Sriperumbudur_Dc (Tamil Nadu)  2018-09-12 02:12:10.755603
4           Chennai_Poonamallee (Tamil Nadu)  2018-09-12 00:02:34.161600
…                                        …                         …
26218     Kanpur_Central_H_6 (Uttar Pradesh)  2018-10-03 23:42:21.856227
26219           Ahmedabad_East_H_1 (Gujarat)  2018-10-04 01:48:54.382343
26220              Kadi_KaranNGR_D (Gujarat)  2018-10-03 23:43:58.933947
26221                 Bellary_Dc (Karnataka)  2018-10-04 03:58:40.726547
26222          Sandur_WrdN1DPP_D (Karnataka)  2018-10-04 02:51:44.712656

                      od_end_time  start_scan_to_end_scan  cutoff_factor  \
0      2018-09-12 03:01:59.598855                    58.0             24
1      2018-09-12 02:03:09.655591                   122.0             48
2      2018-09-12 01:41:29.809822                   100.0             17
3      2018-09-12 03:13:03.432532                    60.0             15
4      2018-09-12 02:12:10.755603                   129.0              9
```

```
…                           …                      …            …
26218  2018-10-04 19:57:34.928573                  1215.0          442
26219  2018-10-04 04:01:41.425627                   132.0           50
26220  2018-10-04 01:48:54.382343                   124.0           34
26221  2018-10-04 08:46:09.166940                   287.0           40
26222  2018-10-04 03:58:40.726547                    66.0           25

       actual_distance_to_destination  segment_actual_time  segment_osrm_time  \
0                           24.644021                 46.0               26.0
1                           48.542890                 95.0               39.0
2                           17.175274                 59.0               16.0
3                           15.325529                 39.0               12.0
4                            9.271519                 21.0               11.0
…                               …                  …                …
26218                      442.024575                991.0              425.0
26219                       50.473578                129.0               55.0
26220                       34.270235                 57.0               37.0
26221                       40.546740                233.0               42.0
26222                       25.534793                 41.0               25.0

       actual_time  osrm_time  osrm_distance  segment_osrm_distance
0             47.0       26.0        28.1994                28.1995
1             96.0       42.0        56.9116                55.9899
2             59.0       15.0        19.6800                19.8766
3             40.0       12.0        16.2225                16.2225
4             21.0       11.0        11.8422                11.8422
…            …        …            …                    …
26218        997.0      395.0       545.1256               573.6479
26219        130.0       54.0        61.9571                67.2659
26220         57.0       38.0        40.4257                40.4256
26221        233.0       42.0        52.5303                52.5303
26222         42.0       26.0        28.0484                28.0484

[26223 rows x 18 columns]
```

[38]:

```
#Merging rows based on trip_id
data=data.groupby(["route_type","trip_uuid","trip_creation_time"]).
  ↪aggregate({"source_center":"first","source_name":
  ↪"first","destination_center":"last",
                                     "destination_name":"last", "od_start_time":
  ↪"first",
                                     "od_end_time":"last","cutoff_factor":
  ↪"sum","actual_distance_to_destination":"sum","osrm_distance":"sum",
                                     "start_scan_to_end_scan":"sum",␣
  ↪"segment_actual_time":"sum",
                                    "segment_osrm_time":"sum","actual_time":
  ↪"sum",
                                     "osrm_time":"sum","segment_osrm_distance":
  ↪"sum"}).reset_index()
data
```

[38]:

```
      route_type                trip_uuid          trip_creation_time  \
0        Carting  trip-153671042288605164  2018-09-12 00:00:22.886430
1        Carting  trip-153671046011330457  2018-09-12 00:01:00.113710
2        Carting  trip-153671055416136166  2018-09-12 00:02:34.161600
3        Carting  trip-153671066201138152  2018-09-12 00:04:22.011653
4        Carting  trip-153671066826362165  2018-09-12 00:04:28.263977
…           …                      …                         …
14782        FTL  trip-153861004148234782  2018-10-03 23:40:41.482736
14783        FTL  trip-153861007249500192  2018-10-03 23:41:12.495257
14784        FTL  trip-153861014185597051  2018-10-03 23:42:21.856227
14785        FTL  trip-153861023893369544  2018-10-03 23:43:58.933947
14786        FTL  trip-153861118270144424  2018-10-03 23:59:42.701692

      source_center                        source_name destination_center  \
0      IND561203AAB  Doddablpur_ChikaDPP_D (Karnataka)       IND561203AAB
1      IND400072AAB           Mumbai Hub (Maharashtra)       IND401104AAA
2      IND600056AAA   Chennai_Poonamallee (Tamil Nadu)       IND600056AAA
3      IND600044AAD  Chennai_Chrompet_DPC (Tamil Nadu)       IND600048AAA
4      IND560043AAC          HBR Layout PC (Karnataka)       IND560043AAC
…             …                                 …                …
14782  IND814101AAB        Dumka_Dudhani_D (Jharkhand)       IND815351AAA
14783  IND842001AAA        Muzaffrpur_Bbganj_I (Bihar)       IND842001AAA
14784  IND206001AAA  Etawah_MhraChng_D (Uttar Pradesh)       IND209304AAA
14785  IND382715AAA          Kadi_KaranNGR_D (Gujarat)       IND382715AAA
14786  IND583119AAA      Sandur_WrdN1DPP_D (Karnataka)       IND583119AAA

                         destination_name               od_start_time  \
0       Doddablpur_ChikaDPP_D (Karnataka)  2018-09-12 02:03:09.655591
1          Mumbai_MiraRd_IP (Maharashtra)  2018-09-12 00:01:00.113710
2        Chennai_Poonamallee (Tamil Nadu)  2018-09-12 02:12:10.755603
3        Chennai_Vandalur_Dc (Tamil Nadu)  2018-09-12 00:04:22.011653
4               HBR Layout PC (Karnataka)  2018-09-12 00:04:28.263977
…                                    …                         …
14782               Jamtara_D (Jharkhand)  2018-10-04 04:22:21.025250
14783         Muzaffrpur_Bbganj_I (Bihar)  2018-10-03 23:41:12.495257
14784  Kanpur_Central_H_6 (Uttar Pradesh)  2018-10-05 02:44:50.858859
14785           Kadi_KaranNGR_D (Gujarat)  2018-10-04 01:48:54.382343
14786       Sandur_WrdN1DPP_D (Karnataka)  2018-10-04 03:58:40.726547

                      od_end_time  cutoff_factor  \
0      2018-09-12 02:03:09.655591             72
1      2018-09-12 01:41:29.809822             17
2      2018-09-12 02:12:10.755603             24
```

```
3      2018-09-12 01:42:22.349694              9
4      2018-09-12 03:00:55.163423             22
…                           …            …
14782  2018-10-04 02:24:41.382263            167
14783  2018-10-04 16:40:41.713085            192
14784  2018-10-04 19:57:34.928573            835
14785  2018-10-04 01:48:54.382343             84
14786  2018-10-04 03:58:40.726547             65

       actual_distance_to_destination  osrm_distance  start_scan_to_end_scan  \
0                           73.186911        85.1110                   180.0
1                           17.175274        19.6800                   100.0
2                           24.597048        28.0647                   189.0
3                            9.100510        12.0184                    98.0
4                           22.424210        28.9203                   146.0
…                               …            …                     …
14782                      168.396341       207.4975                   428.0
14783                      194.552260       229.2052                  1017.0
14784                      836.072017       997.7577                  2180.0
14785                       84.743813       102.3828                   256.0
14786                       66.081533        80.5787                   353.0

       segment_actual_time  segment_osrm_time  actual_time  osrm_time  \
0                    141.0               65.0        143.0       68.0
1                     59.0               16.0         59.0       15.0
2                     60.0               23.0         61.0       23.0
3                     24.0               13.0         24.0       13.0
4                     64.0               34.0         64.0       34.0
…                    …                …          …        …
14782                347.0              220.0        349.0      220.0
14783                845.0              178.0        847.0      178.0
14784               1660.0              891.0       1674.0      724.0
14785                186.0               92.0        187.0       92.0
14786                274.0               67.0        275.0       68.0

       segment_osrm_distance
0                    84.1894
1                    19.8766
2                    28.0647
3                    12.0184
4                    28.9203
…                      …
14782               209.4499
14783               232.5811
14784              1166.3614
14785               107.6915
14786                80.5787
```

```
[14787 rows x 18 columns]
```

[39]:
```
data.nunique() # Unique values in the dataset
```

[39]:
```
route_type                            2
trip_uuid                         14787
trip_creation_time                14787
source_center                       930
source_name                         930
destination_center                 1035
destination_name                   1035
od_start_time                     14787
od_end_time                       14787
cutoff_factor                       684
actual_distance_to_destination    14771
osrm_distance                     14706
start_scan_to_end_scan             2203
segment_actual_time                1887
segment_osrm_time                  1242
actual_time                        1850
osrm_time                           827
segment_osrm_distance             14724
dtype: int64
```

[40]:
```
data.isna().sum() #nullvalues in the data frame
```

[40]:
```
route_type                        0
trip_uuid                         0
trip_creation_time                0
source_center                     0
source_name                       0
destination_center                0
destination_name                  0
od_start_time                     0
od_end_time                       0
cutoff_factor                     0
actual_distance_to_destination    0
osrm_distance                     0
start_scan_to_end_scan            0
segment_actual_time               0
segment_osrm_time                 0
actual_time                       0
osrm_time                         0
segment_osrm_distance             0
dtype: int64
```

```
[41]: data.describe() #statistical summary of dataset
```

```
[41]:        cutoff_factor  actual_distance_to_destination  osrm_distance  \
      count   14787.000000                    14787.000000   14787.000000
      mean      163.379523                      164.290730     204.631953
      std       305.558531                      305.678137     370.953239
      min         9.000000                        9.002461       9.072900
      25%        22.000000                       22.840056      30.875600
      50%        48.000000                       48.376934      65.575600
      75%       162.000000                      163.685113     207.087600
      max      2185.000000                     2187.483994    2840.081000

             start_scan_to_end_scan  segment_actual_time  segment_osrm_time  \
      count            14787.000000         14787.000000       14787.000000
      mean               529.442754           353.118618         180.482924
      std                658.286556           556.439155         314.622727
      min                 23.000000             9.000000           6.000000
      25%                149.000000            66.000000          30.000000
      50%                279.000000           147.000000          65.000000
      75%                632.000000           364.000000         184.000000
      max               7898.000000          6230.000000        2564.000000

              actual_time     osrm_time  segment_osrm_distance
      count  14787.000000  14787.000000            14787.00000
      mean     356.316224    161.667072              222.66823
      std      561.528033    272.406218              416.76499
      min        9.000000      6.000000                9.07290
      25%       67.000000     29.000000               32.57885
      50%      148.000000     60.000000               69.78420
      75%      367.000000    168.000000              216.46395
      max     6265.000000   2032.000000             3523.63240
```

```
[42]: data.describe(include=object)
```

```
[42]:        route_type                trip_uuid          trip_creation_time  \
      count       14787                    14787                       14787
      unique          2                    14787                       14787
      top       Carting  trip-153671042288605164  2018-09-12 00:00:22.886430
      freq         8906                        1                           1

             source_center                    source_name destination_center  \
      count          14787                          14787              14787
      unique           930                            930               1035
      top      IND000000ACB  Gurgaon_Bilaspur_HB (Haryana)       IND000000ACB
      freq            1052                           1052                821

                                destination_name               od_start_time  \
      count                          14787                       14787
      unique                          1035                       14787
      top     Gurgaon_Bilaspur_HB (Haryana)  2018-09-12 02:03:09.655591
      freq                             821                           1

                             od_end_time
      count                       14787
      unique                      14787
      top     2018-09-12 02:03:09.655591
      freq                            1
```

6 Feature Generation

[43]:

```
#Feature generation like source_state and destination_state
data["source_state"]=data["source_name"].apply(lambda x: str(x).split("(")[1][:
 ↪-1])
data["destination_state"]=data["destination_name"].apply(lambda x: str(x).
 ↪split("(")[1][:-1])
data
```

[43]:

```
      route_type                trip_uuid          trip_creation_time  \
0        Carting  trip-153671042288605164  2018-09-12 00:00:22.886430
1        Carting  trip-153671046011330457  2018-09-12 00:01:00.113710
2        Carting  trip-153671055416136166  2018-09-12 00:02:34.161600
3        Carting  trip-153671066201138152  2018-09-12 00:04:22.011653
4        Carting  trip-153671066826362165  2018-09-12 00:04:28.263977
…          …                        …                          …
14782        FTL  trip-153861004148234782  2018-10-03 23:40:41.482736
14783        FTL  trip-153861007249500192  2018-10-03 23:41:12.495257
14784        FTL  trip-153861014185597051  2018-10-03 23:42:21.856227
14785        FTL  trip-153861023893369544  2018-10-03 23:43:58.933947
14786        FTL  trip-153861118270144424  2018-10-03 23:59:42.701692

      source_center                         source_name destination_center  \
0      IND561203AAB  Doddablpur_ChikaDPP_D (Karnataka)       IND561203AAB
1      IND400072AAB           Mumbai Hub (Maharashtra)       IND401104AAA
2      IND600056AAA   Chennai_Poonamallee (Tamil Nadu)       IND600056AAA
3      IND600044AAD  Chennai_Chrompet_DPC (Tamil Nadu)       IND600048AAA
4      IND560043AAC          HBR Layout PC (Karnataka)       IND560043AAC
…            …                                 …                 …
14782  IND814101AAB        Dumka_Dudhani_D (Jharkhand)       IND815351AAA
14783  IND842001AAA        Muzaffrpur_Bbganj_I (Bihar)       IND842001AAA
14784  IND206001AAA  Etawah_MhraChng_D (Uttar Pradesh)       IND209304AAA
14785  IND382715AAA          Kadi_KaranNGR_D (Gujarat)       IND382715AAA
14786  IND583119AAA      Sandur_WrdN1DPP_D (Karnataka)       IND583119AAA
```

```
                         destination_name               od_start_time  \
0       Doddablpur_ChikaDPP_D (Karnataka)  2018-09-12 02:03:09.655591
1          Mumbai_MiraRd_IP (Maharashtra)  2018-09-12 00:01:00.113710
2        Chennai_Poonamallee (Tamil Nadu)  2018-09-12 02:12:10.755603
3        Chennai_Vandalur_Dc (Tamil Nadu)  2018-09-12 00:04:22.011653
4               HBR Layout PC (Karnataka)  2018-09-12 00:04:28.263977
…                                   …                         …
14782               Jamtara_D (Jharkhand)  2018-10-04 04:22:21.025250
14783         Muzaffrpur_Bbganj_I (Bihar)  2018-10-03 23:41:12.495257
14784  Kanpur_Central_H_6 (Uttar Pradesh)  2018-10-05 02:44:50.858859
14785           Kadi_KaranNGR_D (Gujarat)  2018-10-04 01:48:54.382343
14786       Sandur_WrdN1DPP_D (Karnataka)  2018-10-04 03:58:40.726547

                      od_end_time  cutoff_factor  \
0      2018-09-12 02:03:09.655591             72
1      2018-09-12 01:41:29.809822             17
2      2018-09-12 02:12:10.755603             24
3      2018-09-12 01:42:22.349694              9
4      2018-09-12 03:00:55.163423             22
…                           …            …
14782  2018-10-04 02:24:41.382263            167
14783  2018-10-04 16:40:41.713085            192
14784  2018-10-04 19:57:34.928573            835
14785  2018-10-04 01:48:54.382343             84
14786  2018-10-04 03:58:40.726547             65

       actual_distance_to_destination  osrm_distance  start_scan_to_end_scan  \
0                           73.186911        85.1110                   180.0
1                           17.175274        19.6800                   100.0
2                           24.597048        28.0647                   189.0
3                            9.100510        12.0184                    98.0
4                           22.424210        28.9203                   146.0
…                               …            …                     …
14782                      168.396341       207.4975                   428.0
14783                      194.552260       229.2052                  1017.0
14784                      836.072017       997.7577                  2180.0
14785                       84.743813       102.3828                   256.0
14786                       66.081533        80.5787                   353.0

       segment_actual_time  segment_osrm_time  actual_time  osrm_time  \
0                    141.0               65.0        143.0       68.0
1                     59.0               16.0         59.0       15.0
2                     60.0               23.0         61.0       23.0
3                     24.0               13.0         24.0       13.0
4                     64.0               34.0         64.0       34.0
…                    …                …          …        …
14782                347.0              220.0        349.0      220.0
```

```
14783                845.0              178.0        847.0      178.0
14784               1660.0              891.0       1674.0      724.0
14785                186.0               92.0        187.0       92.0
14786                274.0               67.0        275.0       68.0

       segment_osrm_distance   source_state destination_state
0                    84.1894      Karnataka         Karnataka
1                    19.8766    Maharashtra       Maharashtra
2                    28.0647     Tamil Nadu        Tamil Nadu
3                    12.0184     Tamil Nadu        Tamil Nadu
4                    28.9203      Karnataka         Karnataka
…                      …             …               …
14782               209.4499      Jharkhand         Jharkhand
14783               232.5811          Bihar             Bihar
14784              1166.3614  Uttar Pradesh     Uttar Pradesh
14785               107.6915        Gujarat           Gujarat
14786                80.5787      Karnataka         Karnataka

[14787 rows x 20 columns]
```

[44]:
```
data.describe(include=object)
```

[44]:
```
       route_type                trip_uuid          trip_creation_time  \
count       14787                    14787                       14787
unique          2                    14787                       14787
top       Carting  trip-153671042288605164   2018-09-12 00:00:22.886430
freq         8906                        1                           1

       source_center                    source_name destination_center  \
count          14787                          14787              14787
unique           930                            930               1035
top     IND000000ACB  Gurgaon_Bilaspur_HB (Haryana)       IND000000ACB
freq            1052                           1052                821

                     destination_name               od_start_time  \
count                           14787                       14787
unique                           1035                       14787
top     Gurgaon_Bilaspur_HB (Haryana)  2018-09-12 02:03:09.655591
freq                              821                           1

                       od_end_time source_state destination_state
count                        14787        14787             14787
unique                       14787           29                31
top     2018-09-12 02:03:09.655591  Maharashtra       Maharashtra
freq                             1         2714              2561
```

[45]:
```
data.nunique() #unique value in dataframe
```

[45]:

```
route_type                            2
trip_uuid                         14787
trip_creation_time                14787
source_center                       930
source_name                         930
destination_center               1035
destination_name                 1035
od_start_time                     14787
od_end_time                       14787
cutoff_factor                       684
actual_distance_to_destination    14771
osrm_distance                     14706
start_scan_to_end_scan             2203
segment_actual_time                1887
segment_osrm_time                  1242
actual_time                        1850
osrm_time                           827
segment_osrm_distance             14724
source_state                         29
destination_state                    31
dtype: int64
```

[46]:

```
data["source_state"].value_counts() #source-statewise trip count
```

[46]:

```
source_state
Maharashtra               2714
Karnataka                 2143
Haryana                   1823
Tamil Nadu                1039
Telangana                  784
Uttar Pradesh              760
Gujarat                    750
Delhi                      725
West Bengal                665
Punjab                     536
Rajasthan                  514
Andhra Pradesh             435
Bihar                      351
Madhya Pradesh             318
Kerala                     289
Assam                      268
Jharkhand                  160
Uttarakhand                114
Orissa                     107
Chandigarh                  93
Goa                         65
Chhattisgarh                43
```

```
Himachal Pradesh            34
Jammu & Kashmir             17
Dadra and Nagar Haveli      15
Pondicherry                 12
Nagaland                     5
Arunachal Pradesh            4
Mizoram                      4
Name: count, dtype: int64
```

```
[47]: data["destination_state"].value_counts() #destination-statewise trip count
```

```
[47]: destination_state
Maharashtra               2561
Karnataka                 2294
Haryana                   1640
Tamil Nadu                1084
Uttar Pradesh              805
Telangana                  784
Gujarat                    734
West Bengal                697
Delhi                      657
Punjab                     617
Rajasthan                  550
Andhra Pradesh             442
Bihar                      367
Madhya Pradesh             350
Kerala                     270
Assam                      232
Jharkhand                  181
Uttarakhand                122
Orissa                     119
Chandigarh                  65
Goa                         52
Chhattisgarh                43
Himachal Pradesh            42
Arunachal Pradesh           25
Jammu & Kashmir             20
Dadra and Nagar Haveli      17
Meghalaya                    8
Mizoram                      6
Nagaland                     1
Daman & Diu                  1
Tripura                      1
Name: count, dtype: int64
```

```
[48]: data["source_name"].value_counts().head()
```

```
[48]: source_name
      Gurgaon_Bilaspur_HB (Haryana)        1052
      Bhiwandi_Mankoli_HB (Maharashtra)     697
      Bangalore_Nelmngla_H (Karnataka)      624
      Bengaluru_Bomsndra_HB (Karnataka)     455
      Pune_Tathawde_H (Maharashtra)         396
      Name: count, dtype: int64
```

```
[49]: data["source_name"].value_counts().tail()
```

```
[49]: source_name
      Chikodi_IndraNgr_D (Karnataka)          1
      Atmakur_IndraNgr_D (Andhra Pradesh)     1
      Jetpur_DC (Gujarat)                     1
      Bantwal_Trmltmpl_D (Karnataka)          1
      Sandur_WrdN1DPP_D (Karnataka)           1
      Name: count, dtype: int64
```

```
[50]: data["destination_name"].value_counts().head()
```

```
[50]: destination_name
      Gurgaon_Bilaspur_HB (Haryana)        821
      Bangalore_Nelmngla_H (Karnataka)     548
      Bhiwandi_Mankoli_HB (Maharashtra)    403
      Bengaluru_Bomsndra_HB (Karnataka)    342
      Hyderabad_Shamshbd_H (Telangana)     280
      Name: count, dtype: int64
```

```
[51]: data["source-destination"]=data["source_name"] + data["destination_name"]
```

```
[52]: data["source-destination"].value_counts() #Busiest Corridors
```

```
[52]: source-destination
      Bangalore_Nelmngla_H (Karnataka)Bengaluru_KGAirprt_HB (Karnataka)      151
      Gurgaon_Bilaspur_HB (Haryana)Gurgaon_Bilaspur_HB (Haryana)             123
      Bengaluru_Bomsndra_HB (Karnataka)Bengaluru_KGAirprt_HB (Karnataka)     121
      Bengaluru_KGAirprt_HB (Karnataka)Bangalore_Nelmngla_H (Karnataka)      108
      Bhiwandi_Mankoli_HB (Maharashtra)Mumbai Hub (Maharashtra)              105
                                                                            …
      Khammam_NSTRoad_I (Telangana)Nalgonda_HydRoad_DC (Telangana)             1
      Kolkata_Dankuni_HB (West Bengal)Tarkeshwar_Naraynpr_D (West Bengal)      1
      Bamangola_Central_D_1 (West Bengal)Malda_krshnPly_DC (West Bengal)       1
      Nalbari_Bhgtpura_D (Assam)Dhubri_Tetultol_D (Assam)                      1
      Sandur_WrdN1DPP_D (Karnataka)Sandur_WrdN1DPP_D (Karnataka)               1
      Name: count, Length: 2165, dtype: int64
```

[53]:

```
#Average distance
data[data["source-destination"]=="Bangalore_Nelmngla_H␣
↪(Karnataka)Bengaluru_KGAirprt_HB␣
↪(Karnataka)"]["actual_distance_to_destination"].mean()
```

[53]:

```
28.03163476896394
```

[54]:

```
#Average time
data[data["source-destination"]=="Bangalore_Nelmngla_H␣
↪(Karnataka)Bengaluru_KGAirprt_HB (Karnataka)"]["actual_time"].mean()
```

[54]:

```
87.87417218543047
```

[55]:

```
data.drop("source-destination",axis=1,inplace=True)
```

[56]:

```
data["trip_creation_time"]=pd.to_datetime(df["trip_creation_time"]) #␣
↪conversion to datetime datatype
```

[57]:

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14787 entries, 0 to 14786
Data columns (total 20 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   route_type                      14787 non-null  object
 1   trip_uuid                       14787 non-null  object
 2   trip_creation_time              14783 non-null  datetime64[ns]
 3   source_center                   14787 non-null  object
 4   source_name                     14787 non-null  object
 5   destination_center              14787 non-null  object
 6   destination_name                14787 non-null  object
 7   od_start_time                   14787 non-null  object
 8   od_end_time                     14787 non-null  object
 9   cutoff_factor                   14787 non-null  int64
 10  actual_distance_to_destination  14787 non-null  float64
 11  osrm_distance                   14787 non-null  float64
 12  start_scan_to_end_scan          14787 non-null  float64
 13  segment_actual_time             14787 non-null  float64
 14  segment_osrm_time               14787 non-null  float64
 15  actual_time                     14787 non-null  float64
 16  osrm_time                       14787 non-null  float64
 17  segment_osrm_distance           14787 non-null  float64
 18  source_state                    14787 non-null  object
 19  destination_state               14787 non-null  object
dtypes: datetime64[ns](1), float64(8), int64(1), object(10)
memory usage: 2.3+ MB
```

[58]:

```
#Feature generation year
data["trip_creation_year"]=data["trip_creation_time"].dt.year
data["trip_creation_year"].value_counts()
```

[58]:

```
trip_creation_year
2018.0    14783
Name: count, dtype: int64
```

[59]:

```
#Feature generation month
data["trip_creation_month"]=data["trip_creation_time"].dt.month
data["trip_creation_month"].value_counts()
```

[59]:

```
trip_creation_month
9.0     13092
10.0     1691
Name: count, dtype: int64
```

[60]:

```
#Feature generation day
data["trip_creation_day"]=data["trip_creation_time"].dt.day
data["trip_creation_day"].value_counts()
```

[60]:

```
trip_creation_day
25.0    1024
17.0    1000
20.0     854
23.0     820
15.0     809
12.0     779
14.0     762
28.0     731
3.0      695
24.0     674
16.0     657
21.0     657
26.0     642
18.0     580
19.0     571
30.0     552
22.0     544
1.0      539
13.0     516
29.0     463
27.0     457
2.0      457
Name: count, dtype: int64
```

[61]:

```
#Feature generation triptime
data["od_start_time"]=pd.to_datetime(data["od_start_time"])
data["od_end_time"]=pd.to_datetime(data["od_end_time"])
data["trip_time"]=data["od_end_time"]-data["od_start_time"]
data
```

[61]:

```
      route_type                trip_uuid         trip_creation_time  \
0        Carting  trip-153671042288605164 2018-09-20 02:35:36.476840
1        Carting  trip-153671046011330457 2018-09-20 02:35:36.476840
2        Carting  trip-153671055416136166 2018-09-20 02:35:36.476840
3        Carting  trip-153671066201138152 2018-09-20 02:35:36.476840
4        Carting  trip-153671066826362165 2018-09-20 02:35:36.476840
…          …                      …                         …
14782        FTL  trip-153861004148234782 2018-09-24 05:06:56.558662
14783        FTL  trip-153861007249500192 2018-09-24 05:06:56.558662
14784        FTL  trip-153861014185597051 2018-09-24 05:06:56.558662
14785        FTL  trip-153861023893369544 2018-09-24 05:06:56.558662
14786        FTL  trip-153861118270144424 2018-09-24 05:06:56.558662

      source_center                         source_name destination_center  \
0      IND561203AAB   Doddablpur_ChikaDPP_D (Karnataka)       IND561203AAB
1      IND400072AAB            Mumbai Hub (Maharashtra)       IND401104AAA
2      IND600056AAA    Chennai_Poonamallee (Tamil Nadu)       IND600056AAA
3      IND600044AAD   Chennai_Chrompet_DPC (Tamil Nadu)       IND600048AAA
4      IND560043AAC           HBR Layout PC (Karnataka)       IND560043AAC
…             …                                …                …
14782  IND814101AAB         Dumka_Dudhani_D (Jharkhand)       IND815351AAA
14783  IND842001AAA         Muzaffrpur_Bbganj_I (Bihar)       IND842001AAA
14784  IND206001AAA   Etawah_MhraChng_D (Uttar Pradesh)       IND209304AAA
14785  IND382715AAA           Kadi_KaranNGR_D (Gujarat)       IND382715AAA
14786  IND583119AAA       Sandur_WrdN1DPP_D (Karnataka)       IND583119AAA

                         destination_name              od_start_time  \
0       Doddablpur_ChikaDPP_D (Karnataka) 2018-09-12 02:03:09.655591
1          Mumbai_MiraRd_IP (Maharashtra) 2018-09-12 00:01:00.113710
2        Chennai_Poonamallee (Tamil Nadu) 2018-09-12 02:12:10.755603
3        Chennai_Vandalur_Dc (Tamil Nadu) 2018-09-12 00:04:22.011653
4               HBR Layout PC (Karnataka) 2018-09-12 00:04:28.263977
…                                   …                         …
14782               Jamtara_D (Jharkhand) 2018-10-04 04:22:21.025250
14783         Muzaffrpur_Bbganj_I (Bihar) 2018-10-03 23:41:12.495257
14784  Kanpur_Central_H_6 (Uttar Pradesh) 2018-10-05 02:44:50.858859
14785           Kadi_KaranNGR_D (Gujarat) 2018-10-04 01:48:54.382343
14786       Sandur_WrdN1DPP_D (Karnataka) 2018-10-04 03:58:40.726547

                     od_end_time  cutoff_factor  …  segment_osrm_time  \
0     2018-09-12 02:03:09.655591             72  …               65.0
```

```
1     2018-09-12 01:41:29.809822             17  …               16.0
2     2018-09-12 02:12:10.755603             24  …               23.0
3     2018-09-12 01:42:22.349694              9  …               13.0
4     2018-09-12 03:00:55.163423             22  …               34.0
…                          …            …  …                …
14782 2018-10-04 02:24:41.382263            167  …              220.0
14783 2018-10-04 16:40:41.713085            192  …              178.0
14784 2018-10-04 19:57:34.928573            835  …              891.0
14785 2018-10-04 01:48:54.382343             84  …               92.0
14786 2018-10-04 03:58:40.726547             65  …               67.0

       actual_time  osrm_time  segment_osrm_distance   source_state  \
0            143.0       68.0                84.1894      Karnataka
1             59.0       15.0                19.8766    Maharashtra
2             61.0       23.0                28.0647     Tamil Nadu
3             24.0       13.0                12.0184     Tamil Nadu
4             64.0       34.0                28.9203      Karnataka
…            …        …                    …            …
14782        349.0      220.0               209.4499      Jharkhand
14783        847.0      178.0               232.5811          Bihar
14784       1674.0      724.0              1166.3614  Uttar Pradesh
14785        187.0       92.0               107.6915        Gujarat
14786        275.0       68.0                80.5787      Karnataka

       destination_state  trip_creation_year  trip_creation_month  \
0              Karnataka              2018.0                  9.0
1            Maharashtra              2018.0                  9.0
2             Tamil Nadu              2018.0                  9.0
3             Tamil Nadu              2018.0                  9.0
4              Karnataka              2018.0                  9.0
…                  …                 …                  …
14782          Jharkhand              2018.0                  9.0
14783              Bihar              2018.0                  9.0
14784      Uttar Pradesh              2018.0                  9.0
14785            Gujarat              2018.0                  9.0
14786          Karnataka              2018.0                  9.0

      trip_creation_day               trip_time
0                  20.0          0 days 00:00:00
1                  20.0   0 days 01:40:29.696112
2                  20.0          0 days 00:00:00
3                  20.0   0 days 01:38:00.338041
4                  20.0   0 days 02:56:26.899446
…                 …                      …
14782              24.0 -1 days +22:02:20.357013
14783              24.0   0 days 16:59:29.217828
14784              24.0 -1 days +17:12:44.069714
```

```
14785              24.0          0 days 00:00:00
14786              24.0          0 days 00:00:00

[14787 rows x 24 columns]
```

[62]:

```
data.isnull().sum()
```

[62]:

```
route_type                        0
trip_uuid                         0
trip_creation_time                4
source_center                     0
source_name                       0
destination_center                0
destination_name                  0
od_start_time                     0
od_end_time                       0
cutoff_factor                     0
actual_distance_to_destination    0
osrm_distance                     0
start_scan_to_end_scan            0
segment_actual_time               0
segment_osrm_time                 0
actual_time                       0
osrm_time                         0
segment_osrm_distance             0
source_state                      0
destination_state                 0
trip_creation_year                4
trip_creation_month               4
trip_creation_day                 4
trip_time                         0
dtype: int64
```

[63]:

```
data.dropna(inplace=True)
```

[64]:

```
data.isnull().sum()
```

[64]:

```
route_type                        0
trip_uuid                         0
trip_creation_time                0
source_center                     0
source_name                       0
destination_center                0
destination_name                  0
od_start_time                     0
od_end_time                       0
cutoff_factor                     0
```

```
actual_distance_to_destination    0
osrm_distance                     0
start_scan_to_end_scan            0
segment_actual_time               0
segment_osrm_time                 0
actual_time                       0
osrm_time                         0
segment_osrm_distance             0
source_state                      0
destination_state                 0
trip_creation_year                0
trip_creation_month               0
trip_creation_day                 0
trip_time                         0
dtype: int64
```

[65]:
```
data.shape
```

[65]:
```
(14783, 24)
```

[66]:
```
#conversion of triptime to float type data
data["triptime_sec"]=data["trip_time"].dt.total_seconds()
```

[67]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 14783 entries, 0 to 14786
Data columns (total 25 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   route_type                      14783 non-null  object
 1   trip_uuid                       14783 non-null  object
 2   trip_creation_time              14783 non-null  datetime64[ns]
 3   source_center                   14783 non-null  object
 4   source_name                     14783 non-null  object
 5   destination_center              14783 non-null  object
 6   destination_name                14783 non-null  object
 7   od_start_time                   14783 non-null  datetime64[ns]
 8   od_end_time                     14783 non-null  datetime64[ns]
 9   cutoff_factor                   14783 non-null  int64
 10  actual_distance_to_destination  14783 non-null  float64
 11  osrm_distance                   14783 non-null  float64
 12  start_scan_to_end_scan          14783 non-null  float64
 13  segment_actual_time             14783 non-null  float64
 14  segment_osrm_time               14783 non-null  float64
 15  actual_time                     14783 non-null  float64
 16  osrm_time                       14783 non-null  float64
 17  segment_osrm_distance           14783 non-null  float64
 18  source_state                    14783 non-null  object
 19  destination_state               14783 non-null  object
 20  trip_creation_year              14783 non-null  float64
 21  trip_creation_month             14783 non-null  float64
 22  trip_creation_day               14783 non-null  float64
 23  trip_time                       14783 non-null  timedelta64[ns]
 24  triptime_sec                    14783 non-null  float64
dtypes: datetime64[ns](3), float64(12), int64(1), object(8), timedelta64[ns](1)
memory usage: 2.9+ MB
```

[68]:
```python
data[data["triptime_sec"]<0]
```

[68]:
```
      route_type                trip_uuid        trip_creation_time  \
5        Carting  trip-153671074033284934 2018-09-20 02:35:36.476840
14       Carting  trip-153671202698783427 2018-09-23 06:42:06.021680
16       Carting  trip-153671225291120891 2018-09-14 15:42:46.437249
31       Carting  trip-153671440490445199 2018-09-13 20:44:19.424489
35       Carting  trip-153671508851597828 2018-09-29 22:21:45.149226
…          …                      …                          …
14768        FTL  trip-153860767482259863 2018-09-24 05:06:56.558662
14779        FTL  trip-153860945742225615 2018-09-24 05:06:56.558662
14781        FTL  trip-153860985527721606 2018-09-24 05:06:56.558662
14782        FTL  trip-153861004148234782 2018-09-24 05:06:56.558662
14784        FTL  trip-153861014185597051 2018-09-24 05:06:56.558662

      source_center                        source_name destination_center  \
5      IND395009AAA        Surat_Central_D_12 (Gujarat)       IND395004AAB
14     IND395001AAA         Surat_Central_D_9 (Gujarat)       IND395006AAA
16     IND712103AAA        Hoogly_Bandel_D (West Bengal)      IND712124AAA
31     IND140501AAA           Lalru_OnkarDPP_D (Punjab)       IND134203AAA
35     IND360530AAB        Jamjodhpur_Court_D (Gujarat)       IND360575AAA
…          …                                  …                   …
14768  IND505122AAA  Jammikunta_ConduDPP_D (Telangana)       IND505467AAA
14779  IND140001AAA      RoopNagar_ChotiHvl_DC (Punjab)       IND140301AAA
14781  IND814133AAB       Godda_Central_D_2 (Jharkhand)       IND815301AAA
14782  IND814101AAB         Dumka_Dudhani_D (Jharkhand)       IND815351AAA
14784  IND206001AAA  Etawah_MhraChng_D (Uttar Pradesh)       IND209304AAA

                         destination_name              od_start_time  \
5              Surat_Central_D_3 (Gujarat) 2018-09-12 02:31:39.246238
14             Surat_Varachha_DC (Gujarat) 2018-09-12 02:37:19.832796
16                Hooghly_DC (West Bengal) 2018-09-12 03:09:08.473151
31          Naraingarh_Ward2DPP_D (Haryana) 2018-09-12 07:36:00.152620
35                   Porbandar_DC (Gujarat) 2018-09-12 06:04:58.698852
…                                        …                          …
14768     Husnabad_Greenmkt_D (Telangana) 2018-10-04 03:51:10.928009
```

```
14779   Chandigarh_Kharar_DC (Chandigarh) 2018-10-04 03:46:12.300247
14781      Giridih_Shivalya_D (Jharkhand) 2018-10-04 08:29:20.440999
14782               Jamtara_D (Jharkhand) 2018-10-04 04:22:21.025250
14784  Kanpur_Central_H_6 (Uttar Pradesh) 2018-10-05 02:44:50.858859

                     od_end_time  cutoff_factor  …  actual_time  osrm_time  \
5     2018-09-12 02:01:41.638015             25  …        161.0       29.0
14    2018-09-12 02:04:22.360575             19  …        170.0       29.0
16    2018-09-12 02:16:17.710493             51  …        222.0       58.0
31    2018-09-12 03:55:15.023521             47  …        147.0       64.0
35    2018-09-12 03:43:56.169739            178  …        553.0      192.0
…                          …            …  …          …        …
14768 2018-10-04 02:25:04.243970            104  …        380.0      119.0
14779 2018-10-04 02:52:02.434753            183  …        281.0      207.0
14781 2018-10-04 03:01:57.954149            226  …        511.0      248.0
14782 2018-10-04 02:24:41.382263            167  …        349.0      220.0
14784 2018-10-04 19:57:34.928573            835  …       1674.0      724.0

       segment_osrm_distance   source_state  destination_state  \
5                    30.9358        Gujarat            Gujarat
14                   30.5457        Gujarat            Gujarat
16                   71.3328    West Bengal        West Bengal
31                  103.6903         Punjab            Haryana
35                  245.2043        Gujarat            Gujarat
…                      …            …                …
14768               140.2444      Telangana          Telangana
14779               216.3882         Punjab         Chandigarh
14781               378.6774      Jharkhand          Jharkhand
14782               209.4499      Jharkhand          Jharkhand
14784              1166.3614  Uttar Pradesh      Uttar Pradesh

       trip_creation_year  trip_creation_month  trip_creation_day  \
5                  2018.0                  9.0               20.0
14                 2018.0                  9.0               23.0
16                 2018.0                  9.0               14.0
31                 2018.0                  9.0               13.0
35                 2018.0                  9.0               29.0
…                   …                  …                …
14768              2018.0                  9.0               24.0
14779              2018.0                  9.0               24.0
14781              2018.0                  9.0               24.0
14782              2018.0                  9.0               24.0
14784              2018.0                  9.0               24.0

                     trip_time  triptime_sec
5     -1 days +23:30:02.391777  -1797.608223
14    -1 days +23:27:02.527779  -1977.472221
```

```
16    -1 days +23:07:09.237342  -3170.762658
31    -1 days +20:19:14.870901 -13245.129099
35    -1 days +21:38:57.470887  -8462.529113
…                          …           …
14768 -1 days +22:33:53.315961  -5166.684039
14779 -1 days +23:05:50.134506  -3249.865494
14781 -1 days +18:32:37.513150 -19642.486850
14782 -1 days +22:02:20.357013  -7059.642987
14784 -1 days +17:12:44.069714 -24435.930286

[891 rows x 25 columns]
```

[69]:

```
#Here Triptime can not be negative values as travelling time should always be␣
 ↪positive, so we will drop that rows as its false values
data.drop(data[data["triptime_sec"]<0].index,inplace=True)
```

[70]:

```
data.describe()
```

[70]:

```
                    trip_creation_time                    od_start_time  \
count                            13892                            13892
mean   2018-09-22 13:07:02.641116416  2018-09-22 13:44:17.722910976
min       2018-09-12 00:25:19.499696     2018-09-12 00:01:00.113710
25%    2018-09-17 03:21:23.289982976  2018-09-17 04:12:26.345313536
50%    2018-09-22 04:51:35.609723904  2018-09-22 04:52:36.690818048
75%    2018-09-27 17:28:45.461110016  2018-09-27 19:53:56.263667968
max       2018-10-03 23:59:42.701692     2018-10-06 04:27:23.392375
std                                NaN                              NaN

                           od_end_time  cutoff_factor  \
count                            13892   13892.000000
mean   2018-09-22 20:29:21.842143744     159.986251
min       2018-09-12 00:50:10.814399       9.000000
25%    2018-09-17 10:16:06.118688512      21.000000
50%    2018-09-22 12:22:51.642557184      46.000000
75%    2018-09-28 02:16:22.134860288     148.000000
max       2018-10-08 03:00:24.353479    2185.000000
std                                NaN     307.520122

       actual_distance_to_destination  osrm_distance  start_scan_to_end_scan  \
count                    13892.000000   13892.000000            13892.000000
mean                       160.852618     200.437664              515.603657
min                          9.002461       9.072900               23.000000
25%                         22.037144      29.802900              144.000000
50%                         46.163919      61.108100              264.000000
75%                        149.281573     193.689125              595.000000
max                       2187.483994    2840.081000             7898.000000
std                        307.627703     373.619022              660.357703
```

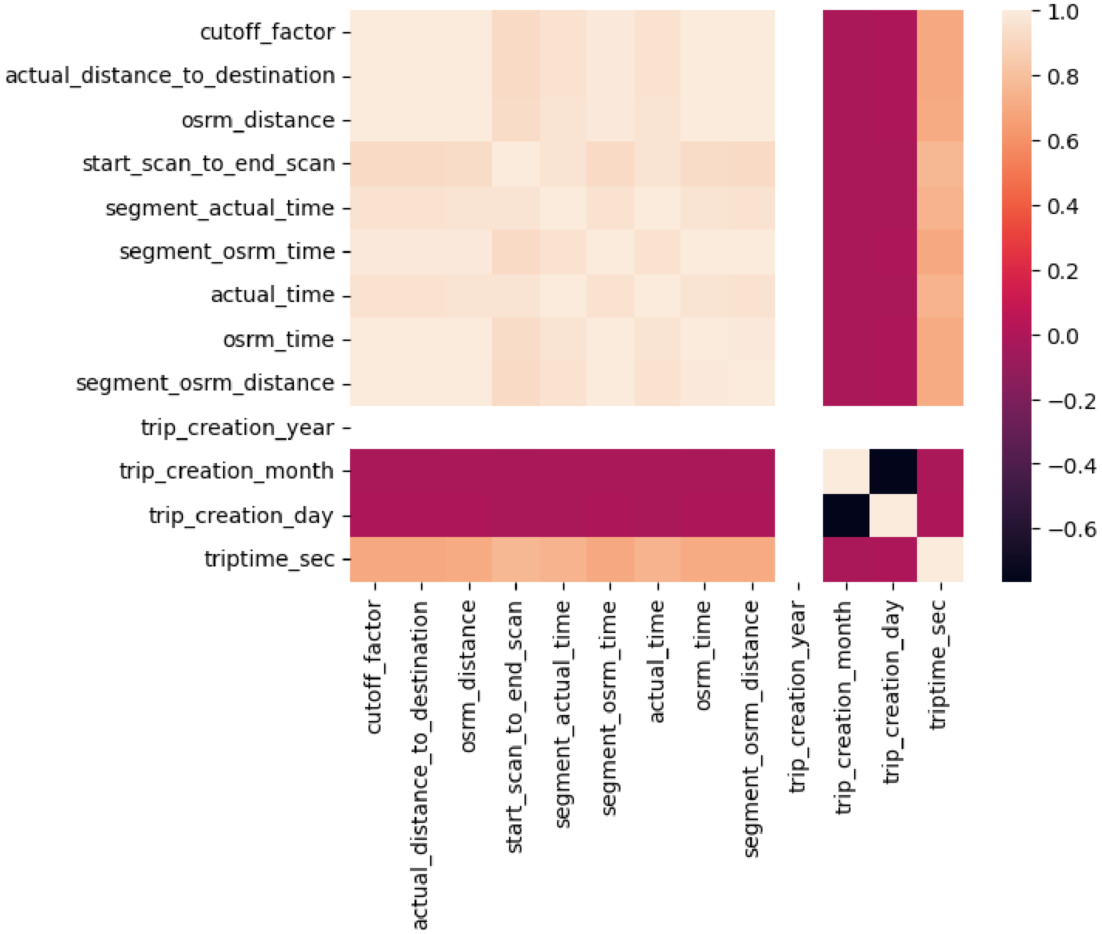
```
       segment_actual_time  segment_osrm_time   actual_time     osrm_time  \
count         13892.000000       13892.000000  13892.000000  13892.000000
mean            346.118557         176.627627    349.267492    157.980564
min               9.000000           6.000000      9.000000      6.000000
25%              64.000000          30.000000     65.000000     29.000000
50%             136.000000          62.000000    138.000000     57.500000
75%             349.000000         176.000000    353.000000    163.250000
max            6230.000000        2564.000000   6265.000000   2032.000000
std             561.918712         316.580388    567.051777    274.050917

       segment_osrm_distance  trip_creation_year  trip_creation_month  \
count           13892.000000             13892.0         13892.000000
mean              218.437771              2018.0             9.114310
min                 9.072900              2018.0             9.000000
25%                31.349950              2018.0             9.000000
50%                65.614850              2018.0             9.000000
75%               204.588700              2018.0             9.000000
max              3523.632400              2018.0            10.000000
std               419.933226                 0.0             0.318199

       trip_creation_day                  trip_time   triptime_sec
count       13892.000000                      13892   13892.000000
mean           18.547653  0 days 06:45:04.119232843   24304.119233
min             1.000000            0 days 00:00:00       0.000000
25%            14.000000  0 days 01:51:56.299656750    6716.299657
50%            20.000000     0 days 03:30:54.417636   12654.417636
75%            25.000000  0 days 07:04:21.465968500   25461.465969
max            30.000000     5 days 11:38:33.117274  473913.117274
std             7.753191  0 days 09:33:47.061593031   34427.061593
```

[72]:

```
#Heatmap of dataframe
data_numeric = data.select_dtypes(include=['float64', 'int64'])

sns.heatmap(data_numeric.corr())
```

[72]:

```
<Axes: >
```



[73]:

```
data_numeric.corr()
```

[73]:

```
                                cutoff_factor  actual_distance_to_destination  \
cutoff_factor                        1.000000                        0.999997
actual_distance_to_destination       0.999997                        1.000000
osrm_distance                        0.997444                        0.997471
start_scan_to_end_scan               0.921081                        0.921345
segment_actual_time                  0.954552                        0.954682
segment_osrm_time                    0.988451                        0.988498
actual_time                          0.955434                        0.955563
osrm_time                            0.994283                        0.994361
segment_osrm_distance                0.993405                        0.993410
trip_creation_year                        NaN                             NaN
trip_creation_month                 -0.019440                       -0.019482
trip_creation_day                   -0.011132                       -0.011104
triptime_sec                         0.703598                        0.703577
```

```
                                osrm_distance  start_scan_to_end_scan  \
cutoff_factor                        0.997444                0.921081
actual_distance_to_destination       0.997471                0.921345
osrm_distance                        1.000000                0.927050
start_scan_to_end_scan               0.927050                1.000000
segment_actual_time                  0.959676                0.963516
segment_osrm_time                    0.992424                0.921375
actual_time                          0.960492                0.963525
osrm_time                            0.997933                0.929408
segment_osrm_distance                0.995036                0.922120
trip_creation_year                        NaN                     NaN
trip_creation_month                 -0.018993               -0.019450
trip_creation_day                   -0.011031               -0.014236
triptime_sec                         0.704835                0.765778

                                segment_actual_time  segment_osrm_time  \
cutoff_factor                              0.954552           0.988451
actual_distance_to_destination             0.954682           0.988498
osrm_distance                              0.959676           0.992424
start_scan_to_end_scan                     0.963516           0.921375
segment_actual_time                        1.000000           0.954571
segment_osrm_time                          0.954571           1.000000
actual_time                                0.999978           0.955367
osrm_time                                  0.959483           0.993647
segment_osrm_distance                      0.957497           0.996487
trip_creation_year                              NaN                NaN
trip_creation_month                       -0.017506          -0.019008
trip_creation_day                         -0.013692          -0.010119
triptime_sec                               0.745170           0.701954

                                actual_time  osrm_time  segment_osrm_distance  \
cutoff_factor                      0.955434   0.994283               0.993405
actual_distance_to_destination     0.955563   0.994361               0.993410
osrm_distance                      0.960492   0.997933               0.995036
start_scan_to_end_scan             0.963525   0.929408               0.922120
segment_actual_time                0.999978   0.959483               0.957497
segment_osrm_time                  0.955367   0.993647               0.996487
actual_time                        1.000000   0.960269               0.958320
osrm_time                          0.960269   1.000000               0.992408
segment_osrm_distance              0.958320   0.992408               1.000000
trip_creation_year                      NaN        NaN                    NaN
trip_creation_month               -0.017533  -0.020042              -0.019023
trip_creation_day                 -0.013676  -0.010269              -0.010496
triptime_sec                       0.745080   0.704238               0.707998

                                trip_creation_year  trip_creation_month  \
cutoff_factor                                  NaN            -0.019440
```

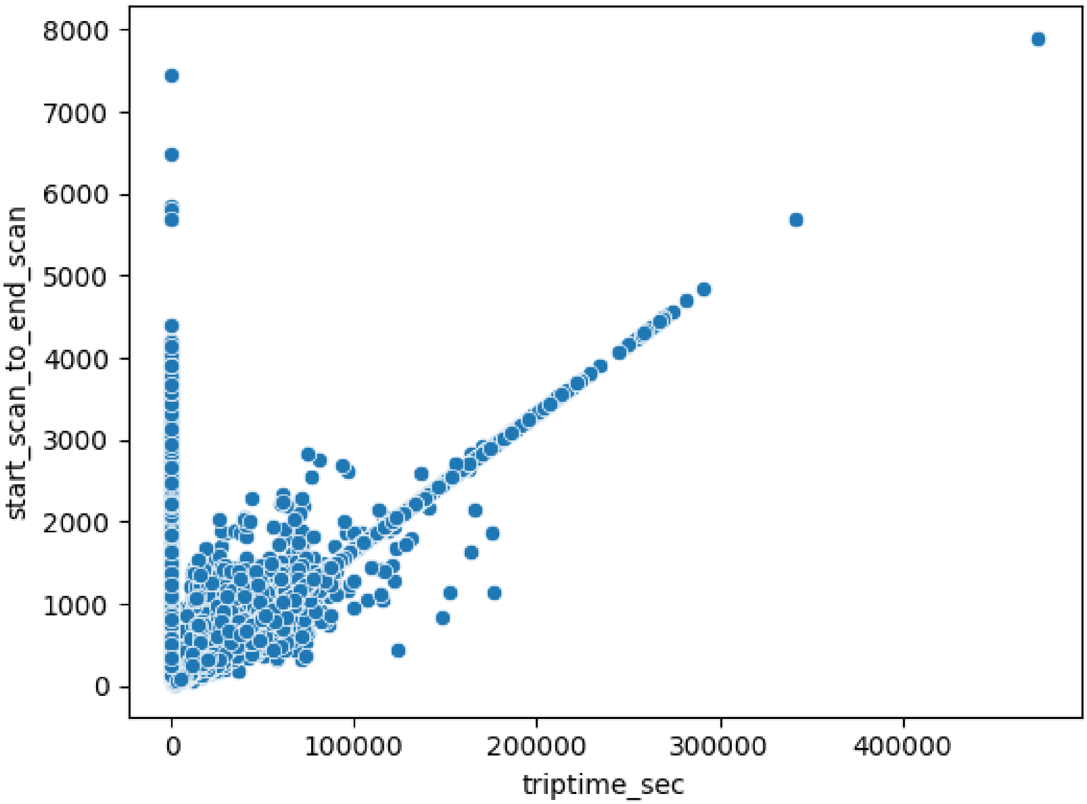
```
actual_distance_to_destination                NaN            -0.019482
osrm_distance                                 NaN            -0.018993
start_scan_to_end_scan                        NaN            -0.019450
segment_actual_time                           NaN            -0.017506
segment_osrm_time                             NaN            -0.019008
actual_time                                   NaN            -0.017533
osrm_time                                     NaN            -0.020042
segment_osrm_distance                         NaN            -0.019023
trip_creation_year                            NaN                  NaN
trip_creation_month                           NaN             1.000000
trip_creation_day                             NaN            -0.762671
triptime_sec                                  NaN            -0.015866

                                trip_creation_day  triptime_sec
cutoff_factor                           -0.011132      0.703598
actual_distance_to_destination          -0.011104      0.703577
osrm_distance                           -0.011031      0.704835
start_scan_to_end_scan                  -0.014236      0.765778
segment_actual_time                     -0.013692      0.745170
segment_osrm_time                       -0.010119      0.701954
actual_time                             -0.013676      0.745080
osrm_time                               -0.010269      0.704238
segment_osrm_distance                   -0.010496      0.707998
trip_creation_year                            NaN           NaN
trip_creation_month                     -0.762671     -0.015866
trip_creation_day                        1.000000     -0.009683
triptime_sec                            -0.009683      1.000000
```

[74]:

```python
#Visualization of triptime and start_scan_to_end_scan
sns.scatterplot(x=data["triptime_sec"], y=data["start_scan_to_end_scan"])
```

[74]:

```
<Axes: xlabel='triptime_sec', ylabel='start_scan_to_end_scan'>
```



7 Hypothesis Testing

- Pearson Test between triptime and start_scan_to_end_scan

H0: Both Variables are not correlated

Ha: Both variables are correlated

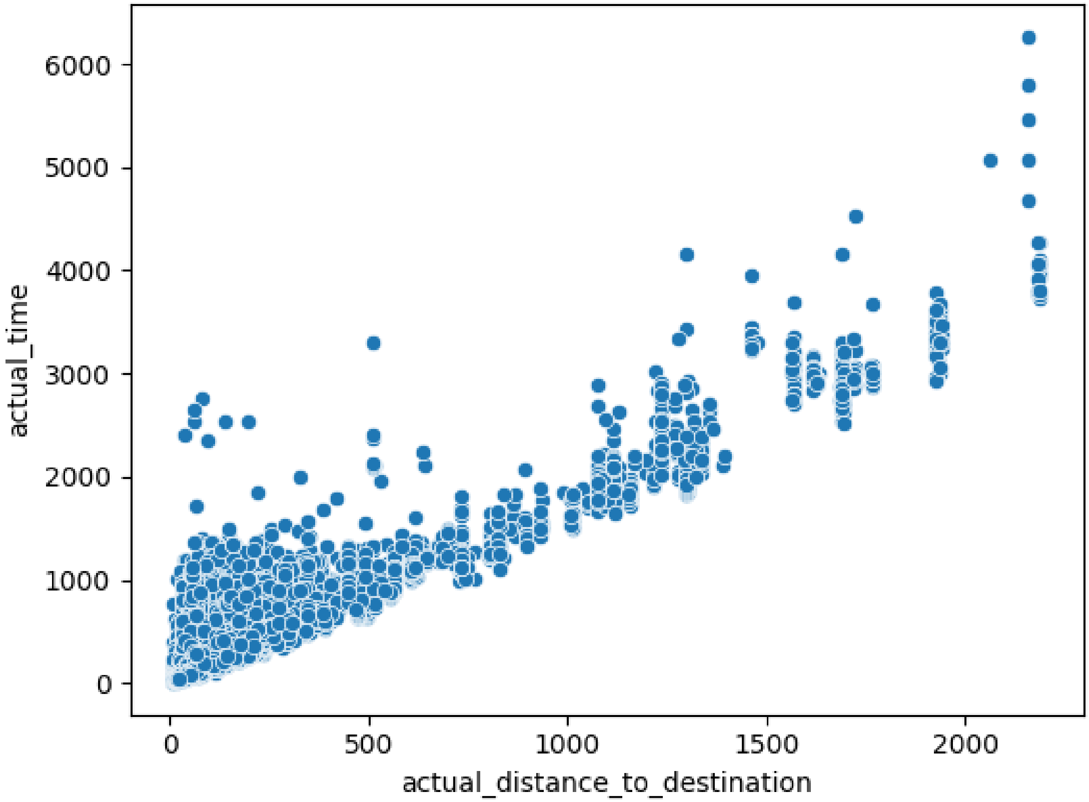
[75]:
```
#Let us set siginificance level 0.05, confidence level 95%
alpha=0.05
```

[76]:
```
test_statistics,p_value=pearsonr(data["triptime_sec"],␣
 ↪data["start_scan_to_end_scan"])
print(p_value)
if p_value < alpha:
    print("Reject Null Hypotheis, Both Variables are correlated")
else:
    print("Fail to Reject Null Hypothesis,Both Variables are not correlated")
```

```
0.0
Reject Null Hypotheis, Both Variables are correlated
```

[77]:
```
#Visualization between distance and time
sns.scatterplot(x=data["actual_distance_to_destination"],y=data["actual_time"])
```

[77]:
```
<Axes: xlabel='actual_distance_to_destination', ylabel='actual_time'>
```



- Pearson Test actual_time and actual_distance_to_destination

H0: Both Variables are not correlated

Ha: Both variables are correlated

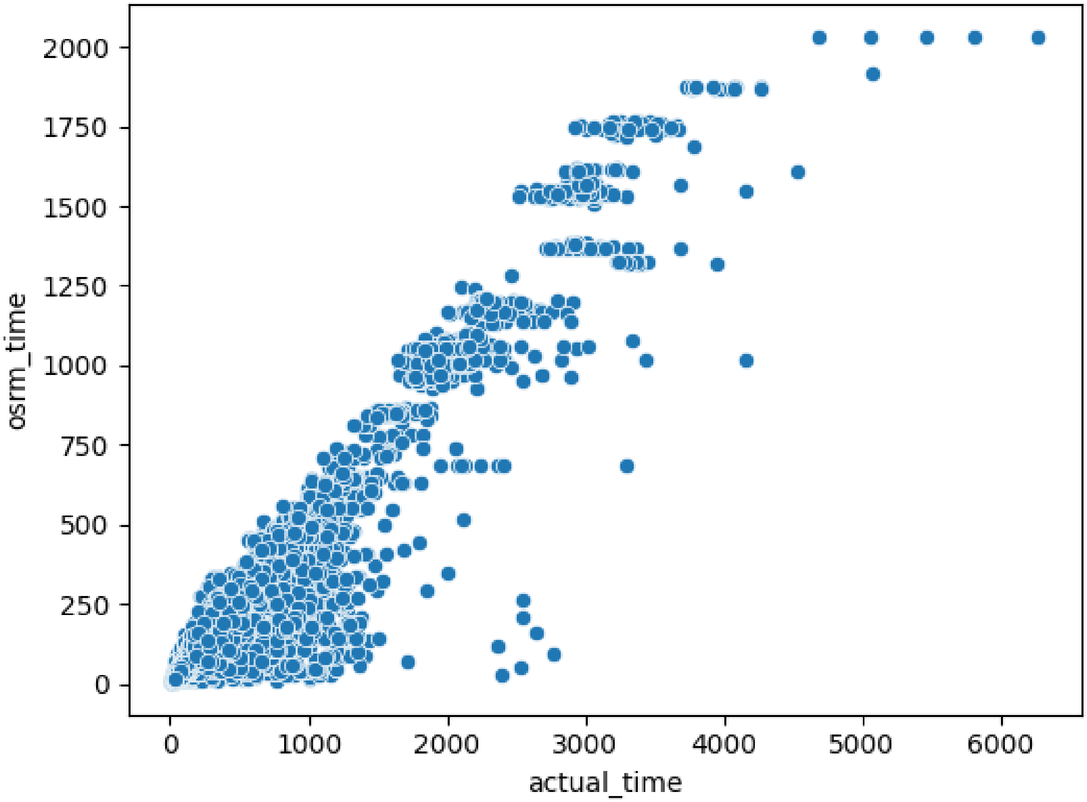
[78]:
```
#Let us set siginificance level 0.05, confidence level 95%
alpha=0.05
```

[79]:
```
test_statistics,p_value=pearsonr(data["actual_time"],data["actual_distance_to_destination"])
print(p_value)
if p_value < alpha:
    print("Reject Null Hypotheis, Both Variables are correlated")
else:
    print("Fail to Reject Null Hypothesis,Both Variables are not correlated")
```

```
0.0
Reject Null Hypotheis, Both Variables are correlated
```

[80]:
```
#Visualization between distance and time
sns.scatterplot(x=data["actual_time"],y=data["osrm_time"])
```

[80]:
```
<Axes: xlabel='actual_time', ylabel='osrm_time'>
```



- T-Test for actual_time and osrm_time

H0: Mean of actual_time and osrm_time are same (mu_1 = mu_2)

Ha: Mean of actual_time is higher than osrm_time (mu_1 > mu_2)

[81]:
```
#Let us set siginificance level 0.05, confidence level 95%
alpha=0.05
```

[82]:
```
test_statistics,p_value=ttest_ind(data["actual_time"],data["osrm_time"],␣
 ↪alternative="greater")
print(p_value)
if p_value < alpha:
    print("Reject Null Hypotheis, Mean of actual_time and osrm_time are same")
else:
    print("Fail to Reject Null Hypothesis,ean of actual_time is higher than␣
 ↪osrm_time")
```

```
1.0113592493195362e-274
Reject Null Hypotheis, Mean of actual_time and osrm_time are same
```

- Pearson Test actual_time and osrm_time

H0: Both Variables are not correlated

Ha: Both variables are correlated

[83]:

```
#Let us set siginificance level 0.05, confidence level 95%
alpha=0.05
```

[84]:

```
test_statistics,p_value=pearsonr(data["actual_time"],data["osrm_time"])
print(p_value)
if p_value < alpha:
    print("Reject Null Hypotheis, Both Variables are correlated")
else:
    print("Fail to Reject Null Hypothesis,Both Variables are not correlated")
```

```
0.0
Reject Null Hypotheis, Both Variables are correlated
```

- T-Test for actual_time and segment_actual_time

H0: Mean of actual_time and segment_actual_time are same (mu_1 = mu_2)

Ha: Mean of actual_time and segment_actual_time are not same (mu_1 != mu_2)

[85]:

```
#Let us set siginificance level 0.05, confidence level 95%
alpha=0.05
```

[86]:

```
test_statistics,p_value=ttest_ind(data["actual_time"],data["segment_actual_time"])
print(p_value)
if p_value < alpha:
    print("Reject Null Hypotheis, Mean of actual_time and segment_actual_time␣
 ↪are not same")
else:
    print("Fail to Reject Null Hypothesis,Mean of actual_time and␣
 ↪segment_actual_time are same")
```

```
0.6419956696137739
Fail to Reject Null Hypothesis,Mean of actual_time and segment_actual_time are
same
```

- T-Test for osrm_time and segment_osrm_time

H0: Mean of osrm_time and segment_osrm_time are same (mu_1 = mu_2)

Ha: Mean of osrm_time and segment_osrm_time are not same (mu_1 != mu_2)

[87]:

```
#Let us set siginificance level 0.05, confidence level 95%
alpha=0.05
```

[88]:

```
test_statistics,p_value=ttest_ind(data["osrm_time"],data["segment_osrm_time"])
print(p_value)
if p_value < alpha:
    print("Reject Null Hypotheis, Mean of osrm_time and segment_osrm_time are␣
 ↪not same")
else:
    print("Fail to Reject Null Hypothesis,Mean of osrm_time and␣
 ↪segment_osrm_time are same")
```

```
1.5413271810594524e-07
Reject Null Hypotheis, Mean of osrm_time and segment_osrm_time are not same
```

- T-Test for osrm_distance and segment_osrm_distance

H0: Mean of osrm_distance and segment_osrm_distance are same (mu_1 = mu_2)

Ha: Mean of osrm_distance and segment_osrm_distance are not same (mu_1 != mu_2)

[89]:

```
#Let us set siginificance level 0.05, confidence level 95%
alpha=0.05
```

[90]:

```
test_statistics,p_value=ttest_ind(data["osrm_distance"],data["segment_osrm_distance"])
print(p_value)
if p_value < alpha:
    print("Reject Null Hypotheis, Mean of osrm_distance and␣
 ↪segment_osrm_distance are not same")
else:
    print("Fail to Reject Null Hypothesis,Mean of osrm_distance and␣
 ↪segment_osrm_distance are same")
```

```
0.00016066670222265932
Reject Null Hypotheis, Mean of osrm_distance and segment_osrm_distance are not
same
```

8 Outliers Detection Using IQR Method

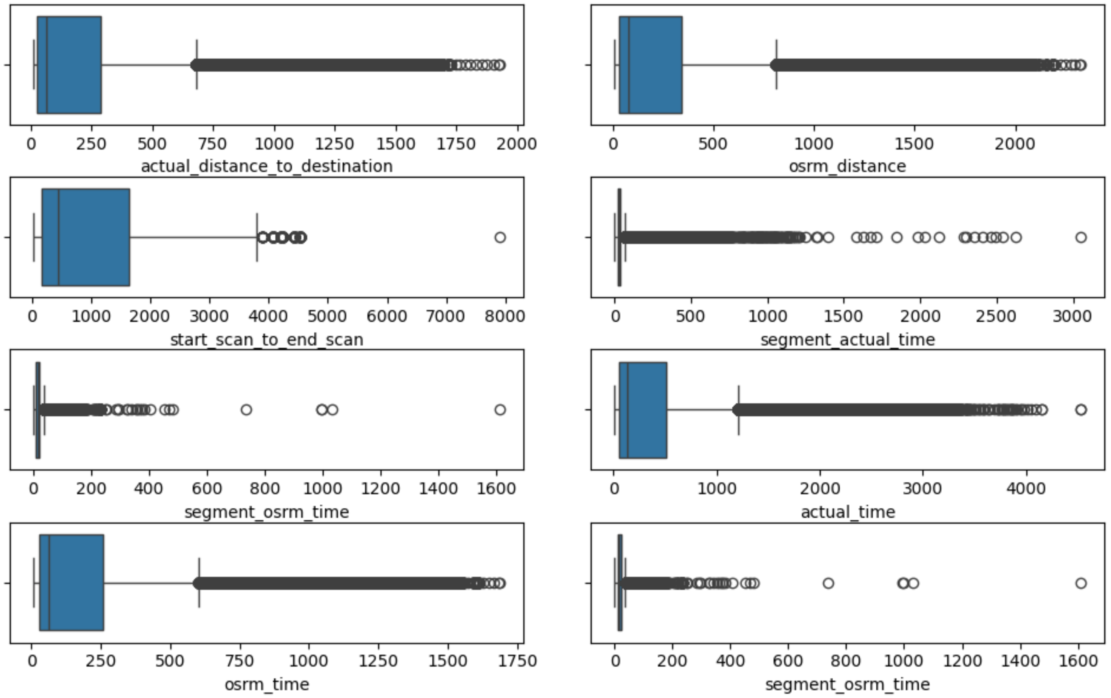
[96]:

```
num_cols = ['actual_distance_to_destination', 'osrm_distance',␣
 ↪'start_scan_to_end_scan', 'segment_actual_time', 'segment_osrm_time',␣
 ↪'actual_time','osrm_time',"segment_osrm_time","triptime_sec"]
```

[110]:

```
fig, axis = plt.subplots(nrows=4, ncols=2, figsize=(10,6))
fig.tight_layout()

index = 0
for row in range(4):
    for col in range(2):
        sns.boxplot(x=df[num_cols[index]], ax=axis[row, col])
        index += 1
```

```
plt.show()
```



[111]:
```
data.describe()
```

[111]:
```
                   trip_creation_time                  od_start_time  \
count                           13892                          13892
mean    2018-09-22 13:07:02.641116416  2018-09-22 13:44:17.722910976
min        2018-09-12 00:25:19.499696     2018-09-12 00:01:00.113710
25%     2018-09-17 03:21:23.289982976  2018-09-17 04:12:26.345313536
50%     2018-09-22 04:51:35.609723904  2018-09-22 04:52:36.690818048
75%     2018-09-27 17:28:45.461110016  2018-09-27 19:53:56.263667968
max        2018-10-03 23:59:42.701692     2018-10-06 04:27:23.392375
std                               NaN                            NaN

                          od_end_time  cutoff_factor  \
count                           13892   13892.000000
mean    2018-09-22 20:29:21.842143744     159.986251
min        2018-09-12 00:50:10.814399       9.000000
25%     2018-09-17 10:16:06.118688512      21.000000
50%     2018-09-22 12:22:51.642557184      46.000000
75%     2018-09-28 02:16:22.134860288     148.000000
max        2018-10-08 03:00:24.353479    2185.000000
std                               NaN     307.520122
```

```
       actual_distance_to_destination  osrm_distance  start_scan_to_end_scan  \
count                    13892.000000   13892.000000            13892.000000   
mean                       160.852618     200.437664              515.603657   
min                          9.002461       9.072900               23.000000   
25%                         22.037144      29.802900              144.000000   
50%                         46.163919      61.108100              264.000000   
75%                        149.281573     193.689125              595.000000   
max                       2187.483994    2840.081000             7898.000000   
std                        307.627703     373.619022              660.357703   

       segment_actual_time  segment_osrm_time   actual_time     osrm_time  \
count         13892.000000       13892.000000  13892.000000  13892.000000   
mean            346.118557         176.627627    349.267492    157.980564   
min               9.000000           6.000000      9.000000      6.000000   
25%              64.000000          30.000000     65.000000     29.000000   
50%             136.000000          62.000000    138.000000     57.500000   
75%             349.000000         176.000000    353.000000    163.250000   
max            6230.000000        2564.000000   6265.000000   2032.000000   
std             561.918712         316.580388    567.051777    274.050917   

       segment_osrm_distance  trip_creation_year  trip_creation_month  \
count           13892.000000             13892.0         13892.000000   
mean              218.437771              2018.0             9.114310   
min                 9.072900              2018.0             9.000000   
25%                31.349950              2018.0             9.000000   
50%                65.614850              2018.0             9.000000   
75%               204.588700              2018.0             9.000000   
max              3523.632400              2018.0            10.000000   
std               419.933226                 0.0             0.318199   

       trip_creation_day                  trip_time   triptime_sec  
count       13892.000000                      13892   13892.000000  
mean           18.547653  0 days 06:45:04.119232843   24304.119233  
min             1.000000            0 days 00:00:00       0.000000  
25%            14.000000  0 days 01:51:56.299656750    6716.299657  
50%            20.000000     0 days 03:30:54.417636   12654.417636  
75%            25.000000  0 days 07:04:21.465968500   25461.465969  
max            30.000000     5 days 11:38:33.117274  473913.117274  
std             7.753191  0 days 09:33:47.061593031   34427.061593  
```

9 Outliers Treatment

[113]:
```
#Let's remove outliers using IQR Method
sses25th = 144
sses75th = 595
iqr = sses75th - sses25th
ssesuw = sses75th +1.5*iqr
ssesuw
```

[113]:
```
1271.5
```

[114]:
```
data[data["start_scan_to_end_scan"]> ssesuw]
```

[114]:
```
      route_type                trip_uuid         trip_creation_time  \
46       Carting  trip-153671813821616145 2018-09-12 01:33:48.711350
433      Carting  trip-153679519504536979 2018-09-21 02:54:55.651098
588      Carting  trip-153681898069638565 2018-10-03 13:07:13.296061
621      Carting  trip-153682425240623736 2018-09-25 06:14:51.782383
836      Carting  trip-153687754564273073 2018-09-27 02:48:03.391966
…          …                      …                        …
14748        FTL  trip-153860451596867762 2018-09-24 05:06:56.558662
14754        FTL  trip-153860570045461434 2018-09-24 05:06:56.558662
14764        FTL  trip-153860698042160875 2018-09-24 05:06:56.558662
14773        FTL  trip-153860879439383883 2018-09-24 05:06:56.558662
14774        FTL  trip-153860880135634048 2018-09-24 05:06:56.558662

      source_center                                 source_name  \
46     IND413517AAA              Udgir_NlgaonRd_D (Maharashtra)
433    IND530012AAA  Visakhapatnam_Gajuwaka_IP (Andhra Pradesh)
588    IND400072AAI             Mumbai_Chndivli_D (Maharashtra)
621    IND421302AAG           Bhiwandi_Mankoli_HB (Maharashtra)
836    IND530012AAA  Visakhapatnam_Gajuwaka_IP (Andhra Pradesh)
…             …                                         …
14748  IND712311AAA            Kolkata_Dankuni_HB (West Bengal)
14754  IND000000ACB               Gurgaon_Bilaspur_HB (Haryana)
14764  IND131028AAB                  Sonipat_Kundli_H (Haryana)
14773  IND000000ACB               Gurgaon_Bilaspur_HB (Haryana)
14774  IND424006AAA              Dhule_MIDCAvdn_I (Maharashtra)

      destination_center                            destination_name  \
46          IND431603AAA             Nanded_Aswningr_I (Maharashtra)
433         IND530012AAA  Visakhapatnam_Gajuwaka_IP (Andhra Pradesh)
588         IND000000AFS             Mumbai_Skynet_INT (Maharashtra)
621         IND401104AAB             Mumbai_MiraRoad_M (Maharashtra)
836         IND530012AAA  Visakhapatnam_Gajuwaka_IP (Andhra Pradesh)
…                  …                                         …
14748       IND712311AAA            Kolkata_Dankuni_HB (West Bengal)
```

```
14754       IND834002AAB                      Ranchi_Hub (Jharkhand)
14764       IND131028AAB                  Sonipat_Kundli_H (Haryana)
14773       IND000000ACB               Gurgaon_Bilaspur_HB (Haryana)
14774       IND425409AAA            Shahada_Nandrbar_D (Maharashtra)

                   od_start_time                od_end_time  cutoff_factor  \
46    2018-09-12 09:44:13.402455 2018-09-13 00:46:22.053872            291
433   2018-09-12 23:33:15.045633 2018-09-13 22:18:58.260294            194
588   2018-09-13 07:10:45.472473 2018-09-14 06:23:48.451756              9
621   2018-09-13 07:37:32.406544 2018-09-14 06:17:06.037037             16
836   2018-09-13 22:25:43.913670 2018-09-14 20:39:35.934714            193
…                          …                        …            …
14748 2018-10-03 22:08:35.968978 2018-10-04 22:26:30.408004            159
14754 2018-10-03 22:28:20.454881 2018-10-05 08:39:47.996375           1010
14764 2018-10-05 08:35:15.664489 2018-10-05 08:35:15.664489           1321
14773 2018-10-06 04:27:23.392375 2018-10-06 04:27:23.392375           1931
14774 2018-10-03 23:20:01.356596 2018-10-04 03:38:11.949795            211

       …  actual_time  osrm_time  segment_osrm_distance    source_state  \
46     …        593.0      235.0               337.1176     Maharashtra
433    …        304.0      168.0               228.2350  Andhra Pradesh
588    …        113.0       12.0                14.7276     Maharashtra
621    …         59.0       27.0                29.6053     Maharashtra
836    …        350.0      167.0               227.4162  Andhra Pradesh
…    …          …        …                    …             …
14748  …       1342.0      145.0               197.2656     West Bengal
14754  …       1625.0      851.0              1222.2127         Haryana
14764  …       2003.0     1166.0              1747.4544         Haryana
14773  …       3307.0     1739.0              2600.9869         Haryana
14774  …       1293.0      185.0               253.9858     Maharashtra

       destination_state  trip_creation_year  trip_creation_month  \
46           Maharashtra              2018.0                  9.0
433       Andhra Pradesh              2018.0                  9.0
588          Maharashtra              2018.0                 10.0
621          Maharashtra              2018.0                  9.0
836       Andhra Pradesh              2018.0                  9.0
…                  …                 …                  …
14748        West Bengal              2018.0                  9.0
14754          Jharkhand              2018.0                  9.0
14764            Haryana              2018.0                  9.0
14773            Haryana              2018.0                  9.0
14774        Maharashtra              2018.0                  9.0

       trip_creation_day              trip_time   triptime_sec
46                  12.0 0 days 15:02:08.651417   54128.651417
433                 21.0 0 days 22:45:43.214661   81943.214661
```

```
588                  3.0 0 days 23:13:02.979283   83582.979283
621                 25.0 0 days 22:39:33.630493   81573.630493
836                 27.0 0 days 22:13:52.021044   80032.021044
…                  …                    …             …
14748               24.0 1 days 00:17:54.439026   87474.439026
14754               24.0 1 days 10:11:27.541494  123087.541494
14764               24.0        0 days 00:00:00       0.000000
14773               24.0        0 days 00:00:00       0.000000
14774               24.0 0 days 04:18:10.593199   15490.593199

[1332 rows x 25 columns]
```

[115]:

```
#We will drop outliers which we have found using IQR Method
data.drop(data[data["start_scan_to_end_scan"]> ssesuw].index, inplace=True)
```

[116]:

```
data.describe()
```

[116]:

```
                  trip_creation_time                 od_start_time  \
count                          12560                         12560
mean   2018-09-22 15:01:55.974861824  2018-09-22 13:14:45.159265536
min       2018-09-12 00:25:19.499696     2018-09-12 00:01:00.113710
25%    2018-09-17 05:23:36.585742080  2018-09-17 04:06:47.181209856
50%    2018-09-22 09:11:30.250937088  2018-09-22 04:19:12.697853440
75%    2018-09-27 19:50:24.957410560  2018-09-27 19:36:40.066671616
max       2018-10-03 23:59:42.701692     2018-10-04 20:15:07.233819
std                              NaN                           NaN

                         od_end_time  cutoff_factor  \
count                          12560   12560.000000
mean   2018-09-22 17:59:39.655986432      82.207006
min       2018-09-12 00:50:10.814399       9.000000
25%    2018-09-17 08:22:35.810087936      21.000000
50%    2018-09-22 08:43:56.948343552      39.000000
75%    2018-09-28 00:21:23.487930368     110.000000
max       2018-10-05 02:38:49.857748     768.000000
std                              NaN      97.457672

       actual_distance_to_destination  osrm_distance  start_scan_to_end_scan  \
count                    12560.000000   12560.000000            12560.000000
mean                        83.047597     105.864573              340.209634
min                          9.002461       9.072900               23.000000
25%                         21.371169      28.290200              135.000000
50%                         39.126693      49.867150              234.000000
75%                        110.677576     140.667500              445.000000
max                        769.326535     879.382200             1271.000000
std                         97.762589     122.336394              286.650425
```

```
       segment_actual_time  segment_osrm_time   actual_time     osrm_time  \
count         12560.000000       12560.000000  12560.000000  12560.000000
mean            199.925080          97.330016    201.728344     88.656290
min               9.000000           6.000000      9.000000      6.000000
25%              60.000000          28.000000     61.000000     27.000000
50%             114.000000          54.000000    116.000000     51.000000
75%             264.000000         136.000000    268.000000    118.000000
max            1212.000000         867.000000   1213.000000    641.000000
std             208.697465         106.831672    210.071577     95.378159

       segment_osrm_distance  trip_creation_year  trip_creation_month  \
count           12560.000000             12560.0         12560.000000
mean              112.928154              2018.0             9.115446
min                 9.072900              2018.0             9.000000
25%                29.320300              2018.0             9.000000
50%                54.772450              2018.0             9.000000
75%               150.009625              2018.0             9.000000
max              1033.678700              2018.0            10.000000
std               131.105959                 0.0             0.319572

       trip_creation_day                  trip_time    triptime_sec
count       12560.000000                      12560    12560.000000
mean           18.591879  0 days 04:44:54.496720952    17094.496721
min             1.000000            0 days 00:00:00        0.000000
25%            14.000000  0 days 01:50:43.119529750     6643.119530
50%            20.000000     0 days 03:18:03.833821    11883.833821
75%            25.000000  0 days 06:02:42.184283250    21762.184283
max            30.000000     2 days 01:09:57.136511   176997.136511
std             7.801503  0 days 04:32:06.822854151    16326.822854
```

Here We can observe that by removing outliers in start_scan_to_end_scan column, other columns max values has been decreased significantly and further dropping columns will lead to loss of valuable data

10 Data Encoding

[117]:
```
data_encoding=data.copy()
```

[118]:
```
data_encoding.shape
```

[118]:
```
(12560, 25)
```

11 Label Encoding

```
[119]: #Here We will use label encoder for encoding route_type column
       le = LabelEncoder()
```

```
[120]: col="route_type"
       data_encoding[col].value_counts()
```

```
[120]: route_type
       Carting    8465
       FTL        4095
       Name: count, dtype: int64
```

```
[121]: data_encoding[col]=le.fit_transform(data_encoding[col])
       data_encoding[col].value_counts()
```

```
[121]: route_type
       0    8465
       1    4095
       Name: count, dtype: int64
```

12 Target Encoding

```
[137]: !pip install category_encoders

       from category_encoders import TargetEncoder
```

```
Requirement already satisfied: category_encoders in
/usr/local/lib/python3.10/dist-packages (2.6.3)
Requirement already satisfied: numpy>=1.14.0 in /usr/local/lib/python3.10/dist-
packages (from category_encoders) (1.26.4)
Requirement already satisfied: scikit-learn>=0.20.0 in
/usr/local/lib/python3.10/dist-packages (from category_encoders) (1.3.2)
Requirement already satisfied: scipy>=1.0.0 in /usr/local/lib/python3.10/dist-
packages (from category_encoders) (1.13.1)
Requirement already satisfied: statsmodels>=0.9.0 in
/usr/local/lib/python3.10/dist-packages (from category_encoders) (0.14.3)
Requirement already satisfied: pandas>=1.0.5 in /usr/local/lib/python3.10/dist-
packages (from category_encoders) (2.1.4)
Requirement already satisfied: patsy>=0.5.1 in /usr/local/lib/python3.10/dist-
packages (from category_encoders) (0.5.6)
Requirement already satisfied: python-dateutil>=2.8.2 in
/usr/local/lib/python3.10/dist-packages (from pandas>=1.0.5->category_encoders)
(2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-
packages (from pandas>=1.0.5->category_encoders) (2024.2)
```

```
Requirement already satisfied: tzdata>=2022.1 in /usr/local/lib/python3.10/dist-
packages (from pandas>=1.0.5->category_encoders) (2024.1)
Requirement already satisfied: six in /usr/local/lib/python3.10/dist-packages
(from patsy>=0.5.1->category_encoders) (1.16.0)
Requirement already satisfied: joblib>=1.1.1 in /usr/local/lib/python3.10/dist-
packages (from scikit-learn>=0.20.0->category_encoders) (1.4.2)
Requirement already satisfied: threadpoolctl>=2.0.0 in
/usr/local/lib/python3.10/dist-packages (from scikit-
learn>=0.20.0->category_encoders) (3.5.0)
Requirement already satisfied: packaging>=21.3 in
/usr/local/lib/python3.10/dist-packages (from
statsmodels>=0.9.0->category_encoders) (24.1)
```

[127]:

```
te=TargetEncoder()
```

[128]:

```
#Here we will do target encoding for␣
 ↪"source_center","source_name","destination_center","destination_name","source_state","desti
 ↪columns
columns=["source_center","source_name","destination_center","destination_name","source_state",
for col in columns:
    data_encoding[col]=te.fit_transform(data_encoding[col],␣
 ↪data_encoding["route_type"])
```

[129]:

```
data_encoding.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 12560 entries, 0 to 14786
Data columns (total 25 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   route_type                      12560 non-null  int64
 1   trip_uuid                       12560 non-null  object
 2   trip_creation_time              12560 non-null  datetime64[ns]
 3   source_center                   12560 non-null  float64
 4   source_name                     12560 non-null  float64
 5   destination_center              12560 non-null  float64
 6   destination_name                12560 non-null  float64
 7   od_start_time                   12560 non-null  datetime64[ns]
 8   od_end_time                     12560 non-null  datetime64[ns]
 9   cutoff_factor                   12560 non-null  int64
 10  actual_distance_to_destination  12560 non-null  float64
 11  osrm_distance                   12560 non-null  float64
 12  start_scan_to_end_scan          12560 non-null  float64
 13  segment_actual_time             12560 non-null  float64
 14  segment_osrm_time               12560 non-null  float64
 15  actual_time                     12560 non-null  float64
 16  osrm_time                       12560 non-null  float64
 17  segment_osrm_distance           12560 non-null  float64
 18  source_state                    12560 non-null  float64
 19  destination_state               12560 non-null  float64
 20  trip_creation_year              12560 non-null  float64
 21  trip_creation_month             12560 non-null  float64
 22  trip_creation_day               12560 non-null  float64
 23  trip_time                       12560 non-null  timedelta64[ns]
 24  triptime_sec                    12560 non-null  float64
dtypes: datetime64[ns](3), float64(18), int64(2), object(1), timedelta64[ns](1)
memory usage: 2.5+ MB
```

```
[130]: data_encoding.
        ↪drop(["trip_uuid","trip_creation_time","od_start_time","od_end_time","trip_time"],axis=1,␣
        ↪inplace=True)
```

```
[131]: data_encoding.head()
```

```
[131]:    route_type  source_center  source_name  destination_center  \
       0           0       0.224956     0.224956        2.317944e-01
       1           0       0.026667     0.026667        4.952686e-08
       2           0       0.256210     0.256210        2.836151e-01
       3           0       0.018690     0.018690        2.178526e-01
       4           0       0.000891     0.000891        4.033957e-06

          destination_name  cutoff_factor  actual_distance_to_destination  \
       0      2.317944e-01             72                       73.186911
       1      4.952686e-08             17                       17.175274
       2      2.836151e-01             24                       24.597048
       3      2.178526e-01              9                        9.100510
       4      4.033957e-06             22                       22.424210

          osrm_distance  start_scan_to_end_scan  segment_actual_time  \
       0        85.1110                   180.0                141.0
       1        19.6800                   100.0                 59.0
       2        28.0647                   189.0                 60.0
       3        12.0184                    98.0                 24.0
       4        28.9203                   146.0                 64.0

          segment_osrm_time  actual_time  osrm_time  segment_osrm_distance  \
       0               65.0        143.0       68.0                84.1894
       1               16.0         59.0       15.0                19.8766
       2               23.0         61.0       23.0                28.0647
       3               13.0         24.0       13.0                12.0184
       4               34.0         64.0       34.0                28.9203

          source_state  destination_state  trip_creation_year  trip_creation_month  \
       0      0.126357           0.156993              2018.0                  9.0
       1      0.198639           0.177807              2018.0                  9.0
```

```
2      0.259448           0.273547           2018.0                   9.0
3      0.259448           0.273547           2018.0                   9.0
4      0.126357           0.156993           2018.0                   9.0

   trip_creation_day  triptime_sec
0               20.0      0.000000
1               20.0   6029.696112
2               20.0      0.000000
3               20.0   5880.338041
4               20.0  10586.899446
```

13 Standardization

[135]:
```
#Here We will use MinMaxScaler method for standardizing dataframe
scaler=MinMaxScaler()
std_data=scaler.fit_transform(data_encoding)
std_data=pd.DataFrame(std_data, columns=data_encoding.columns)
```

[136]:
```
std_data
```

[136]:
```
       route_type  source_center  source_name  destination_center  \
0             0.0       0.227004     0.227004        2.455594e-01
1             0.0       0.026909     0.026909        5.246406e-08
2             0.0       0.258542     0.258542        3.004574e-01
3             0.0       0.018860     0.018860        2.307896e-01
4             0.0       0.000899     0.000899        4.273507e-06
…             …              …            …                   …
12555         1.0       0.425476     0.425476        8.941138e-01
12556         1.0       0.635162     0.635162        1.000000e+00
12557         1.0       0.893970     0.893970        9.463165e-01
12558         1.0       0.635162     0.635162        6.668095e-01
12559         1.0       0.417490     0.417490        5.237077e-01

       destination_name  cutoff_factor  actual_distance_to_destination  \
0          2.455594e-01       0.083004                        0.084417
1          5.246406e-08       0.010540                        0.010749
2          3.004574e-01       0.019763                        0.020510
3          2.307896e-01       0.000000                        0.000129
4          4.273507e-06       0.017128                        0.017653
…                   …            …                              …
12555      8.941138e-01       0.176548                        0.178597
12556      1.000000e+00       0.217391                        0.220359
12557      9.463165e-01       0.241107                        0.244040
12558      6.668095e-01       0.098814                        0.099617
12559      5.237077e-01       0.073781                        0.075072
```

```
       osrm_distance  start_scan_to_end_scan  segment_actual_time  \
0           0.087369                0.125801             0.109726
1           0.012188                0.061699             0.041563
2           0.021822                0.133013             0.042394
3           0.003384                0.060096             0.012469
4           0.022805                0.098558             0.045719
…              …                      …                  …
12555       0.209090                0.664263             0.216958
12556       0.242704                0.524840             0.306733
12557       0.252936                0.796474             0.694929
12558       0.107215                0.186699             0.147132
12559       0.082161                0.264423             0.220283

       segment_osrm_time  actual_time  osrm_time  segment_osrm_distance  \
0               0.068525     0.111296   0.097638               0.073313
1               0.011614     0.041528   0.014173               0.010544
2               0.019744     0.043189   0.026772               0.018536
3               0.008130     0.012458   0.011024               0.002875
4               0.032520     0.045681   0.044094               0.019371
…                   …          …        …                    …
12555           0.150987     0.218439   0.209449               0.159903
12556           0.218351     0.306478   0.292913               0.198886
12557           0.199768     0.696013   0.270866               0.218141
12558           0.099884     0.147841   0.135433               0.096250
12559           0.070848     0.220930   0.097638               0.069789

       source_state  destination_state  trip_creation_year  \
0          0.103203           0.089502                 0.0
1          0.182698           0.112002                 0.0
2          0.249577           0.215500                 0.0
3          0.249577           0.215500                 0.0
4          0.103203           0.089502                 0.0
…              …                …                 …
12555      0.418798           0.342118                 0.0
12556      0.460661           0.442377                 0.0
12557      0.997386           0.935294                 0.0
12558      0.460661           0.442377                 0.0
12559      0.103203           0.089502                 0.0

       trip_creation_month  trip_creation_day  triptime_sec
0                      0.0           0.655172      0.000000
1                      0.0           0.655172      0.034067
2                      0.0           0.655172      0.000000
3                      0.0           0.655172      0.033223
4                      0.0           0.655172      0.059814
…                     …                …           …
12555                  0.0           0.793103      0.289181
```

```
12556                   0.0           0.793103      0.230396
12557                   0.0           0.793103      0.345594
12558                   0.0           0.793103      0.000000
12559                   0.0           0.793103      0.000000

[12560 rows x 20 columns]
```

14 Business Insights

1. On average, osrm_time is lesser than segment_osrm_time
2. On average, osrm_distance is lesser than segment_osrm_distance
3. On average, There is no difference between actual time and segment actual time
4. On average, Actual time is higher than osrm time. While the maximum osrm_time is 400 mins (6.6 hrs), the actual time goes upto 800 mins (13 hrs) which is almost double
5. (9:00am to 12:00pm) and (5:00pm to 10:00pm) have higher delivery time
6. Wednesday is the busiest day of the week with maximum number of trips
7. (10:00pm to 1:00am) is the busiest time of the day having maximum number of trips- probably because the delivery time is least during these hours - less traffic on the roads
8. Carting route type is used for short-distance (0-100km) and short duration (<500 mins) trips while FTLs are used for long-distance (>100km) and long-duration (>300 mins) trips
9. FTL trips are 50% of carting trips in count
10. It can be observed that average speed in inter-state deliveries is much higher than the avg speed in intra-state deliveries
11. Delhi has the lowest intra-state delivery speed while Punjab has the highest

15 Recommendations

1. Since actual time is higher than OSRM time on an average for all trips, the company needs to either improve their forecasting accuracy or identify root cause of delays in deliveries

2. Identify best practices from Maharashtra and Karnataka (states which have the highest volume of deliveries) to increase business in other states

3. To reduce actual_time, dispatch as many deliveries as possible outside of the busy hours

4. Optimise routes along corridors with maximum average speed to shorten delivery time