Business Case Study for Delhivery

Introduction

Delhivery is the largest and fastest-growing fully integrated logistics player in India. With a focus on building the operating system for commerce, they leverage world-class infrastructure, top-notch logistics operations, and advanced engineering and technology capabilities. The data team at Delhivery plays a crucial role in creating intelligence and capabilities using data, which helps them maintain a competitive edge in quality, efficiency, and profitability.

Objective

The primary objective is to help Delhivery understand and process the data emerging from their data engineering pipelines. This involves:

- Cleaning, sanitizing, and manipulating data to derive useful features.
- Analyzing raw data to assist the data science team in building forecasting models.

Approach

Data Preparation

Load the Dataset:

Load the dataset and inspect the first few rows to understand its structure.

Basic Data Cleaning:

Handle missing values.

Analyze data types and convert columns to appropriate types.

Feature Extraction and Transformation:

Extract features from trip_creation_time, source_name, and destination_name.

Calculate additional features like time taken between od_start_time and od_end_time.

Aggregating Data:

Aggregate data based on trip_uuid, source_center, and destination_center using groupby and aggregation functions.

Feature Engineering:

Create new features based on business logic and the dataset's nature.

Handle categorical values using one-hot encoding.

Normalize/standardize numerical features.

Outlier Detection and Treatment:

Identify and treat outliers using visual analysis and the IQR method.

Hypothesis Testing and Visual Analysis:

Conduct hypothesis testing and visual analysis to compare various aggregated values and understand relationships between features.

Import Necessary Libraries, Load, and Inspect Data

Import Libraries: Ensure that all necessary libraries are imported at the beginning of your analysis.

Load Data: Load your dataset into a pandas DataFrame to facilitate data manipulation and analysis.

Inspect Data: Perform an initial inspection of the data to understand its structure and content.

```
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

url =
('https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/001/5
51/original/delhivery_data.csv?1642751181')

df=pd.read_csv(url)

df.shape #Shape of dataset
```

```
(144867, 24)
```

```
df.info() #datatype info of dataset
```

```
<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
```

```
df.nunique() # number of unique values in columns
```

```
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
```

```
df.isna().sum() #missing values in columns
```

```
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
```

```
df.describe() #Statistical summary of the dataset
```

```
       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

df.describe(include=object)

             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
```

```
#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"]))]
```

```
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]
```

```
#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"]))]
```

```
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]
```

```
#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)
```

```
df.describe()
```

```
       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

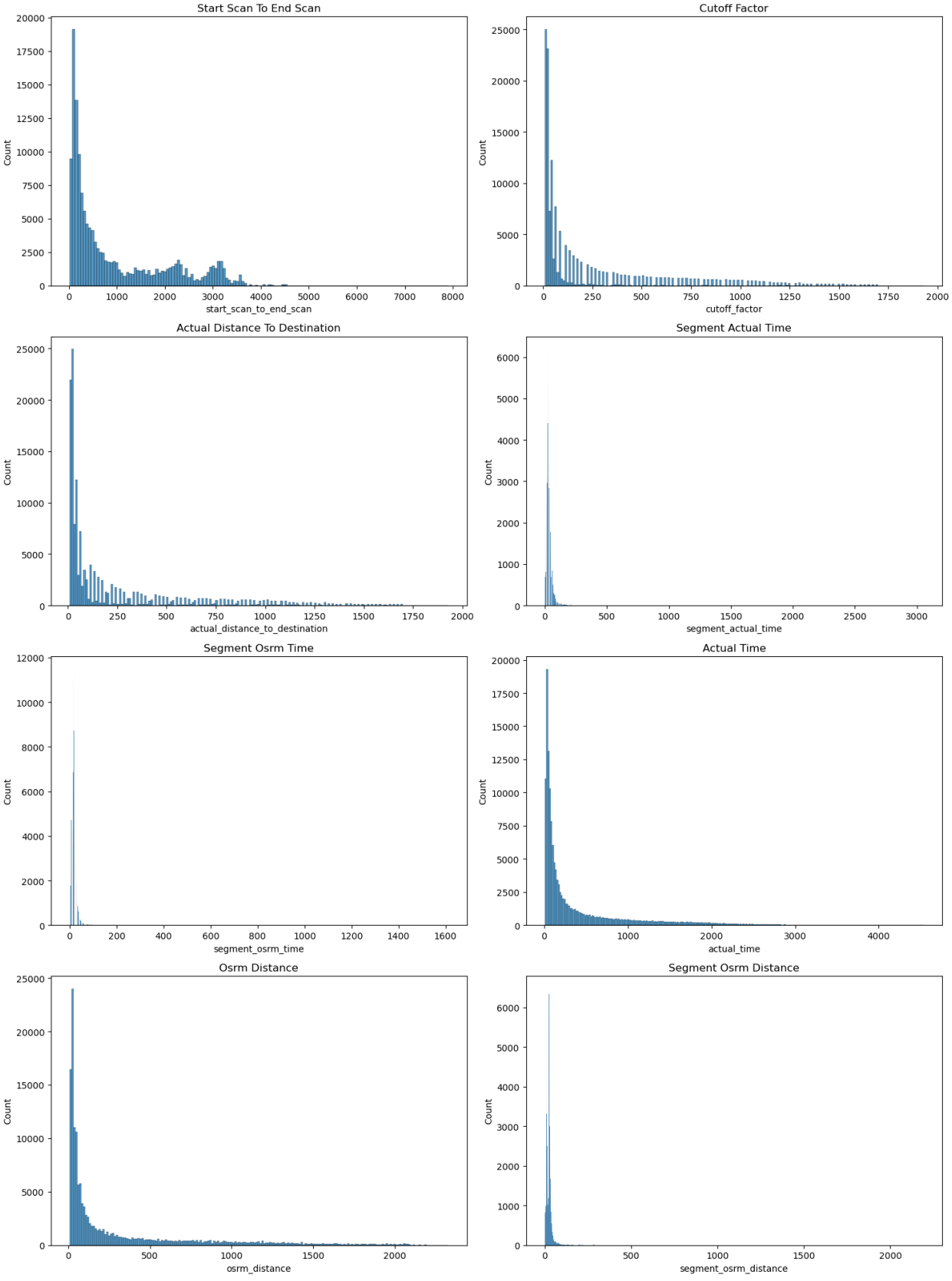
segment_osrm_time: It contains the orsm time taken by a subset of the package delivery. It ranges from 0 to 1611

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

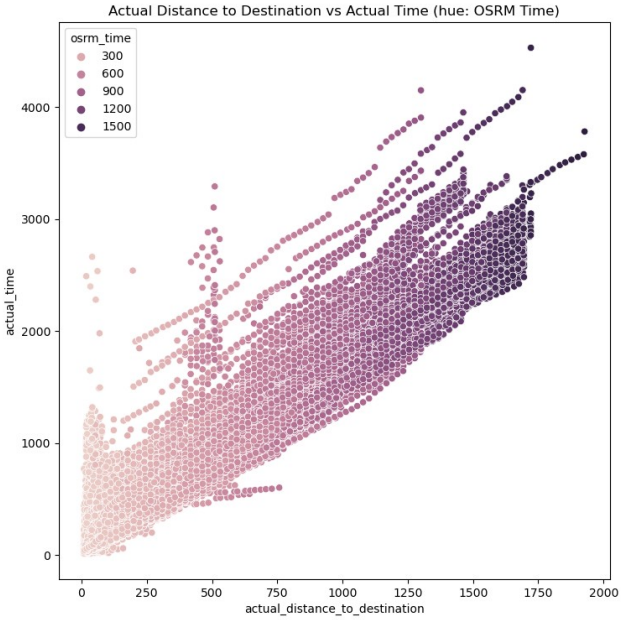
Univariate Analysis

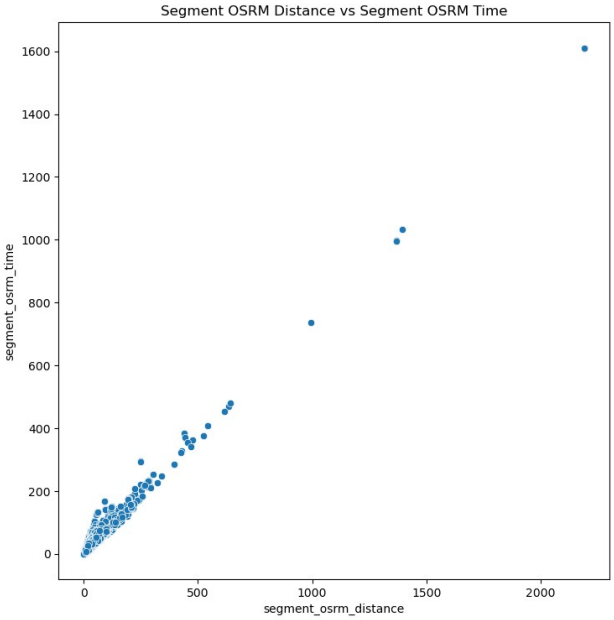
```
# Create a figure and a grid of subplots
fig, axes = plt.subplots(nrows=4, ncols=2, figsize=(15, 20))

# List of attributes to plot
attributes = [
    "start_scan_to_end_scan", "cutoff_factor",
"actual_distance_to_destination",
    "segment_actual_time", "segment_osrm_time", "actual_time",
    "osrm_distance", "segment_osrm_distance"
]

# Plot each attribute in a different subplot
for ax, attr in zip(axes.flatten(), attributes):
    sns.histplot(df[attr], ax=ax)
    ax.set_title(attr.replace("_", " ").title())

# Adjust layout
plt.tight_layout()
plt.show()
```

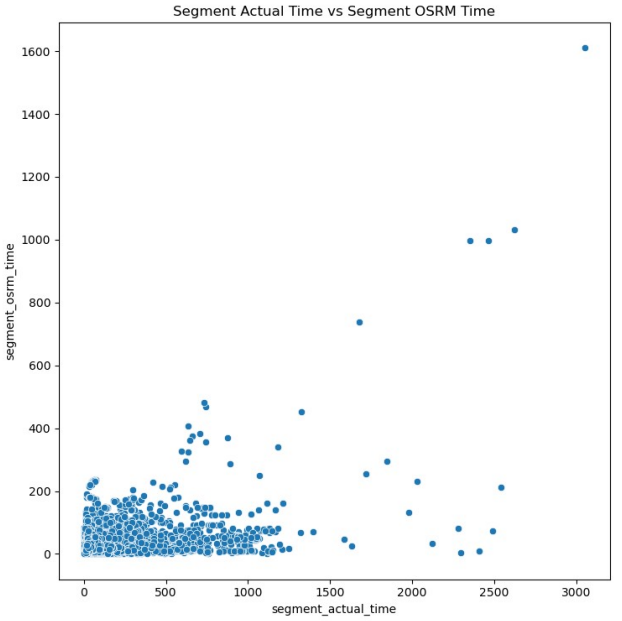

Start Scan To End Scan
Count
start_scan_to_end_scan
Cutoff Factor
Count
cutoff_factor
Actual Distance To Destination
Count
actual_distance_to_destination
Segment Actual Time
Count
segment_actual_time
Segment Osrm Time
Count
segment_osrm_time
Actual Time
Count
actual_time
Osrm Distance
Count
osrm_distance
Segment Osrm Distance
Count
segment_osrm_distance

Bivariate Analysis

```
# Create a figure and a grid of subplots
fig, axes = plt.subplots(nrows=2, ncols=2, figsize=(15, 15))

# 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", ax=axes[0, 0])
axes[0, 0].set_title('Actual Distance to Destination vs Actual Time
(hue: OSRM Time)')

# Scatterplot between segment_osrm_distance and segment_osrm_time
sns.scatterplot(data=df, x="segment_osrm_distance",
y="segment_osrm_time", ax=axes[0, 1])
axes[0, 1].set_title('Segment OSRM Distance vs Segment OSRM Time')

# Scatterplot between segment_actual_time and segment_osrm_time
sns.scatterplot(data=df, x="segment_actual_time",
y="segment_osrm_time", ax=axes[1, 0])
axes[1, 0].set_title('Segment Actual Time vs Segment OSRM Time')

# Hide the empty subplot
axes[1, 1].axis('off')

# Adjust layout
plt.tight_layout()
plt.show()
```








```
# Identify columns with non-numeric values
non_numeric_cols = df.select_dtypes(exclude=['number']).columns
print(non_numeric_cols)

# Option 1: Drop non-numeric columns before calculating correlation
df_numeric = df.drop(non_numeric_cols, axis=1)
correlation_matrix = df_numeric.corr()
print(correlation_matrix)
```

```
Index(['data', 'trip_creation_time', 'route_schedule_uuid',
'route_type',
       'trip_uuid', 'source_center', 'source_name',
```

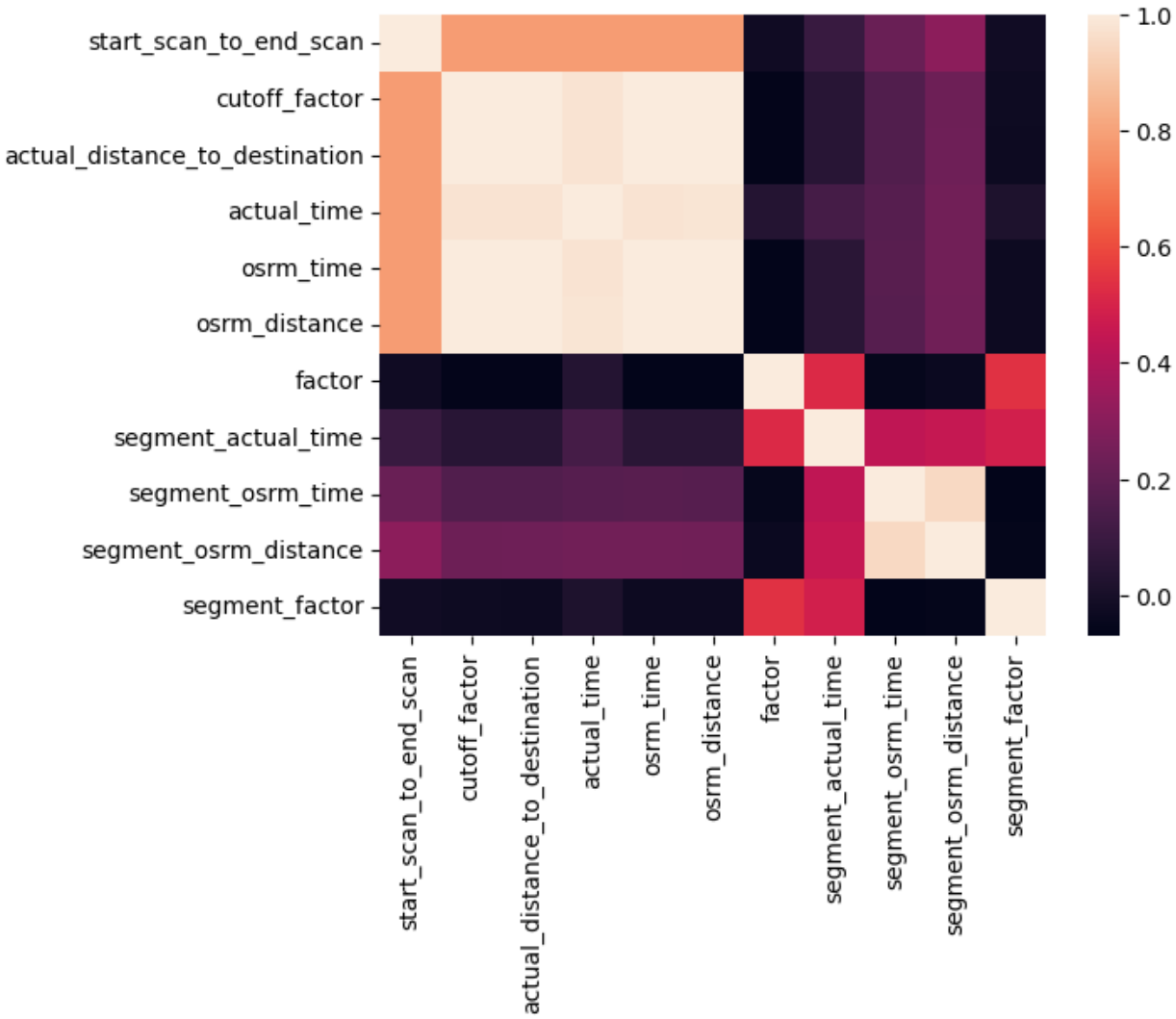
```
'destination_center',
       'destination_name', 'od_start_time', 'od_end_time',
'is_cutoff',
       'cutoff_timestamp'],
      dtype='object')
                                start_scan_to_end_scan  cutoff_factor
\
start_scan_to_end_scan                        1.000000       0.784656

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
```

```
#heatmap of dataframe
sns.heatmap(df.select_dtypes(include=['number']).corr())
```

```
<Axes: >
```



Data Wrangling

```
#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","destination_center","destination_name","od_sta
rt_time","od_end_time","start_scan_to_end_scan"]).aggregate({"cutoff_f
actor":"max","actual_distance_to_destination":"max","segment_actual_ti
me":"sum", "segment_osrm_time":"sum", "actual_time":"max",
"osrm_time":"max","osrm_distance":"max","segment_osrm_distance":"sum"}
).reset_index()
data
```

```
      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]
```

```
#Merging rows based on trip_id
data=data.groupby(["route_type","trip_uuid","trip_creation_time"]).agg
regate({"source_center":"first","source_name":"first","destination_cen
ter":"last",
                                    "destination_name":"last",
"od_start_time":"first",

"od_end_time":"last","cutoff_factor":"sum","actual_distance_to_destina
tion":"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
```

```
      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]
```

```
data.nunique() # Unique values in the dataset
```

```
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
```

```
data.isna().sum() #nullvalues in the data frame
```

```
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
```

```
data.describe() #statistical summary of dataset
```

```
       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
```

#Feature Generation

```
#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
```

```
      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]
```

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

```
       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
```

```
data.nunique() #unique value in dataframe
```

```
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
```

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

```
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: source_state, dtype: int64
```

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

```
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: destination_state, dtype: int64
```

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

```
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: source_name, dtype: int64
```

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

```
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: source_name, dtype: int64
```

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

```
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: destination_name, dtype: int64
```

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

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

```
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: source-destination, Length: 2165, dtype: int64
```

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

```
28.03163476896394
```

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

```
87.87417218543047
```

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

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

```
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

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

2018.0    14783
Name: trip_creation_year, dtype: int64

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

9.0     13092
10.0     1691
Name: trip_creation_month, dtype: int64

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

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: trip_creation_day, dtype: int64
```

```
#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
```

```
      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]
```

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

```
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
```

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

```
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
```

```
data.shape
```

```
(14783, 24)
```

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

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 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

data[data["triptime_sec"]<0]

      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]
```

```
#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)

data.describe()
```

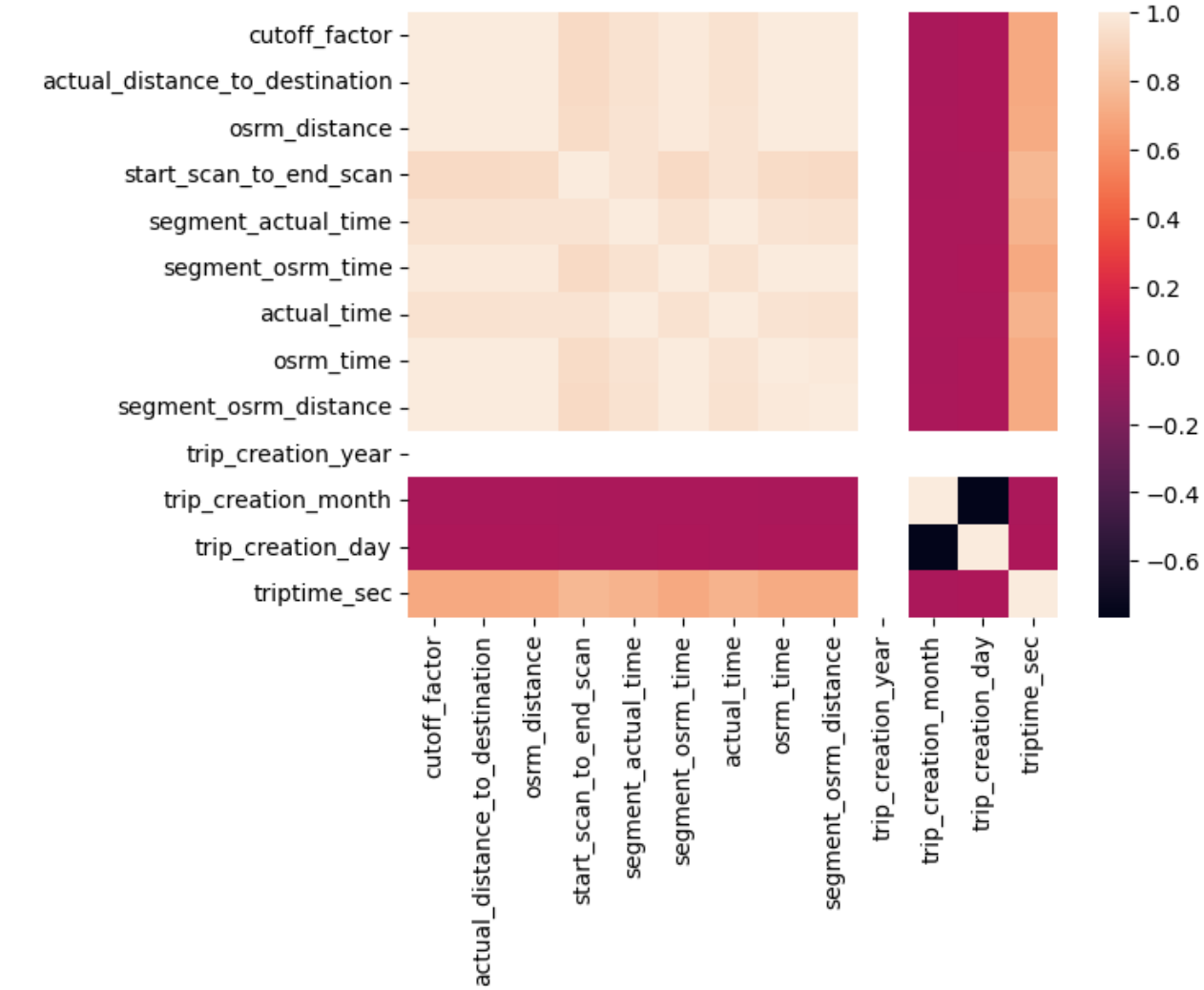
```
       cutoff_factor  actual_distance_to_destination  osrm_distance  \
count   13892.000000                    13892.000000   13892.000000
mean      159.986251                      160.852618     200.437664
std       307.520122                      307.627703     373.619022
min         9.000000                        9.002461       9.072900
25%        21.000000                       22.037144      29.802900
50%        46.000000                       46.163919      61.108100
75%       148.000000                      149.281573     193.689125
max      2185.000000                     2187.483994    2840.081000

       start_scan_to_end_scan  segment_actual_time  segment_osrm_time
\
count            13892.000000         13892.000000       13892.000000

mean               515.603657           346.118557         176.627627

std                660.357703           561.918712         316.580388

min                 23.000000             9.000000           6.000000

25%                144.000000            64.000000          30.000000

50%                264.000000           136.000000          62.000000

75%                595.000000           349.000000         176.000000

max               7898.000000          6230.000000        2564.000000


        actual_time     osrm_time  segment_osrm_distance
trip_creation_year  \
count  13892.000000  13892.000000           13892.000000
13892.0
mean     349.267492    157.980564             218.437771
2018.0
std      567.051777    274.050917             419.933226
```

```
0.0
min        9.000000      6.000000               9.072900
2018.0
25%       65.000000     29.000000              31.349950
2018.0
50%      138.000000     57.500000              65.614850
2018.0
75%      353.000000    163.250000             204.588700
2018.0
max     6265.000000   2032.000000            3523.632400
2018.0

       trip_creation_month  trip_creation_day
trip_time  \
count         13892.000000       13892.000000
13892
mean              9.114310          18.547653  0 days
06:45:04.119232843
std               0.318199           7.753191  0 days
09:33:47.061593031
min               9.000000           1.000000            0 days
00:00:00
25%               9.000000          14.000000  0 days
01:51:56.299656750
50%               9.000000          20.000000     0 days
03:30:54.417636
75%               9.000000          25.000000  0 days
07:04:21.465968500
max              10.000000          30.000000     5 days
11:38:33.117274

        triptime_sec
count   13892.000000
mean    24304.119233
std     34427.061593
min         0.000000
25%      6716.299657
50%     12654.417636
75%     25461.465969
max    473913.117274
```

```
# Select only numerical columns for correlation calculation
numerical_data = data.select_dtypes(include=['number'])

# Generate the heatmap using the numerical data
sns.heatmap(numerical_data.corr())
```

```
C:\Users\aa\AppData\Local\Temp\ipykernel_7456\3759149774.py:5:
FutureWarning: The default value of numeric_only in DataFrame.corr is
deprecated. In a future version, it will default to False. Select only
```

```
valid columns or specify the value of numeric_only to silence this
warning.
  sns.heatmap(numerical_data.corr())

<Axes: >
```




```
# Select only numerical columns for correlation calculation
numerical_data = data.select_dtypes(include=['number'])

# Calculate the correlation matrix on numerical data
correlation_matrix = numerical_data.corr()

# Display the correlation matrix
print(correlation_matrix)

                               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
```

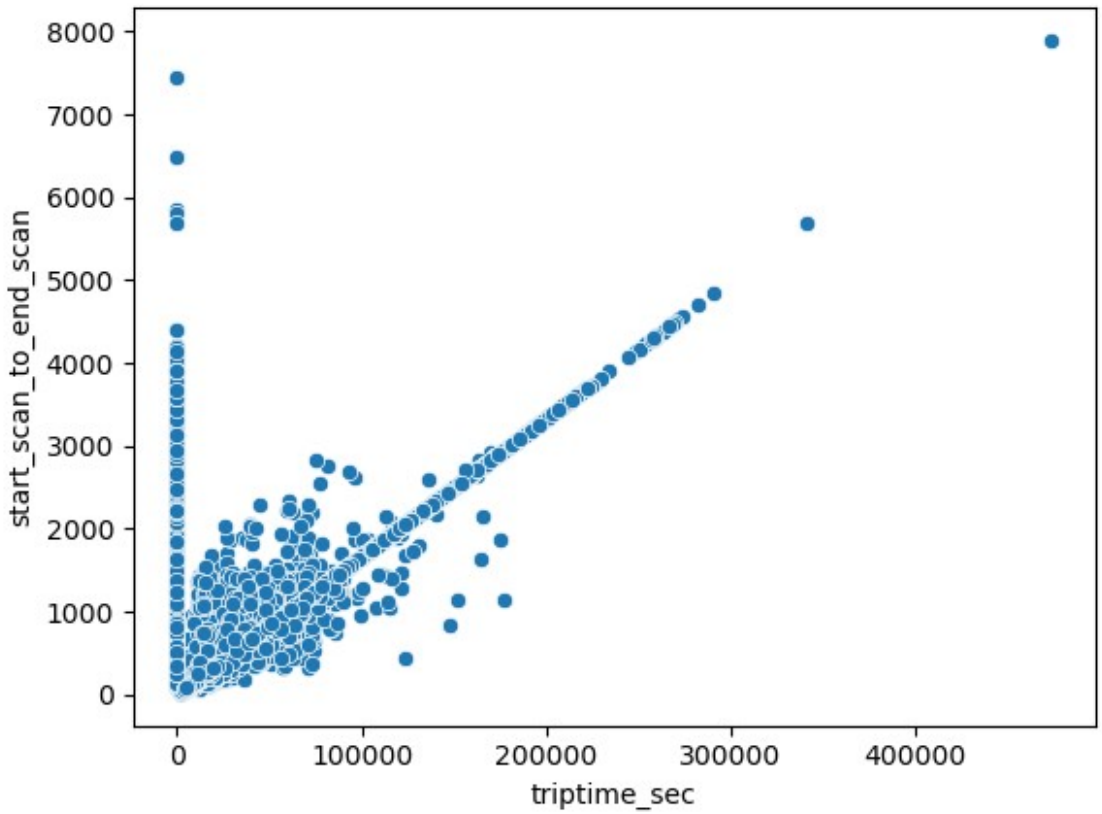
```
0.992408
segment_osrm_distance               0.958320   0.992408
1.000000
trip_creation_year                       NaN        NaN
NaN
trip_creation_month               -0.017533  -0.020042              -
0.019023
trip_creation_day                 -0.013676  -0.010269              -
0.010496
triptime_sec                       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
```

```
C:\Users\aa\AppData\Local\Temp\ipykernel_7456\1601561340.py:5:
FutureWarning: The default value of numeric_only in DataFrame.corr is
deprecated. In a future version, it will default to False. Select only
valid columns or specify the value of numeric_only to silence this
warning.
  correlation_matrix = numerical_data.corr()
```

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

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



Hypothesis Testiing

Pearson Test between triptime and start_scan_to_end_scan

H0: Both Variables are not correlated

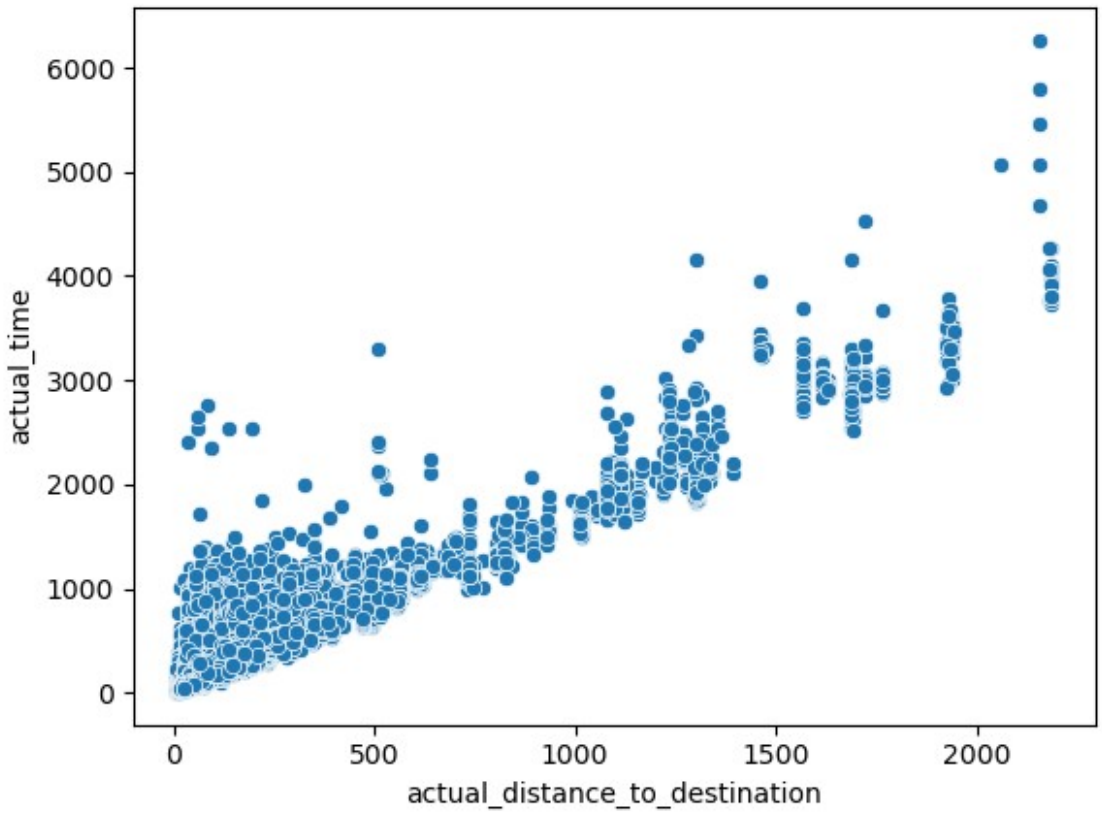
Ha: Both variables are correlated

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

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

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



Pearson Test actual_time and actual_distance_to_destination

H0: Both Variables are not correlated

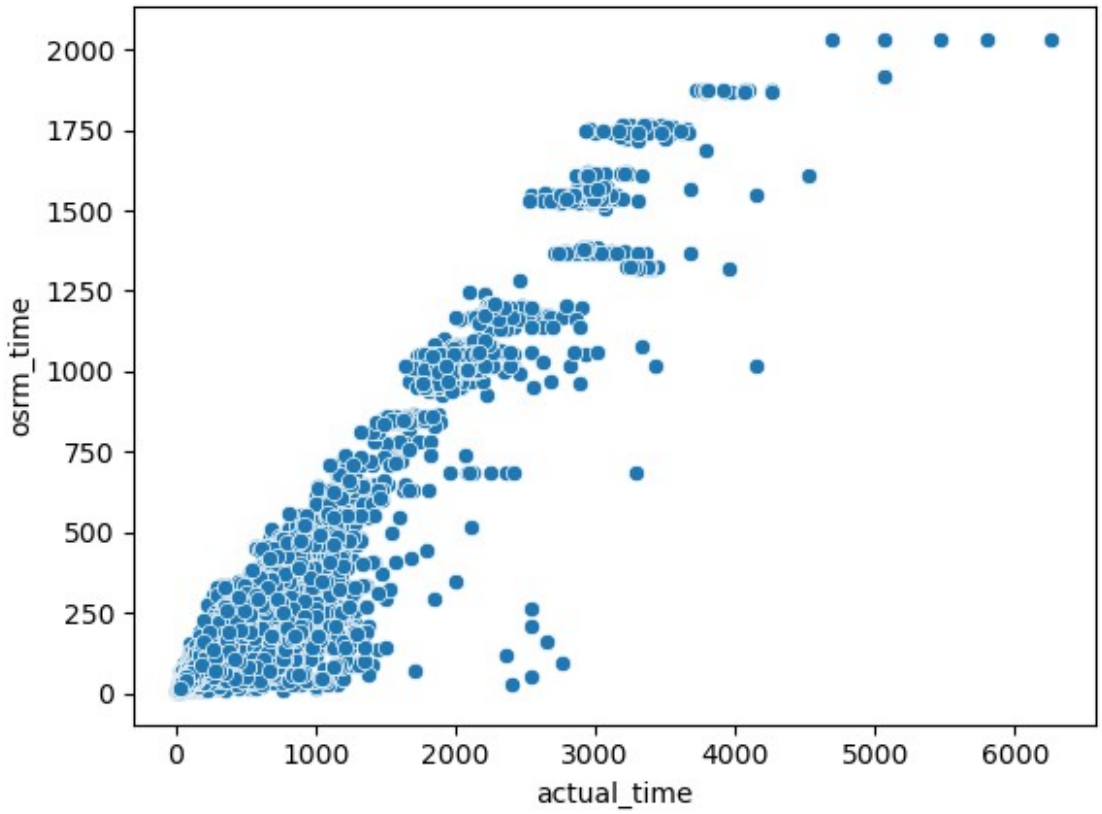
Ha: Both variables are correlated

```
#Let us set siginificance level 0.05, confidence level 95%
alpha=0.05
test_statistics,p_value=pearsonr(data["actual_time"],data["actual_dist
ance_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
```

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

```
<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)

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

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

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

test_statistics,p_value=ttest_ind(data["actual_time"],data["segment_ac
tual_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)

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

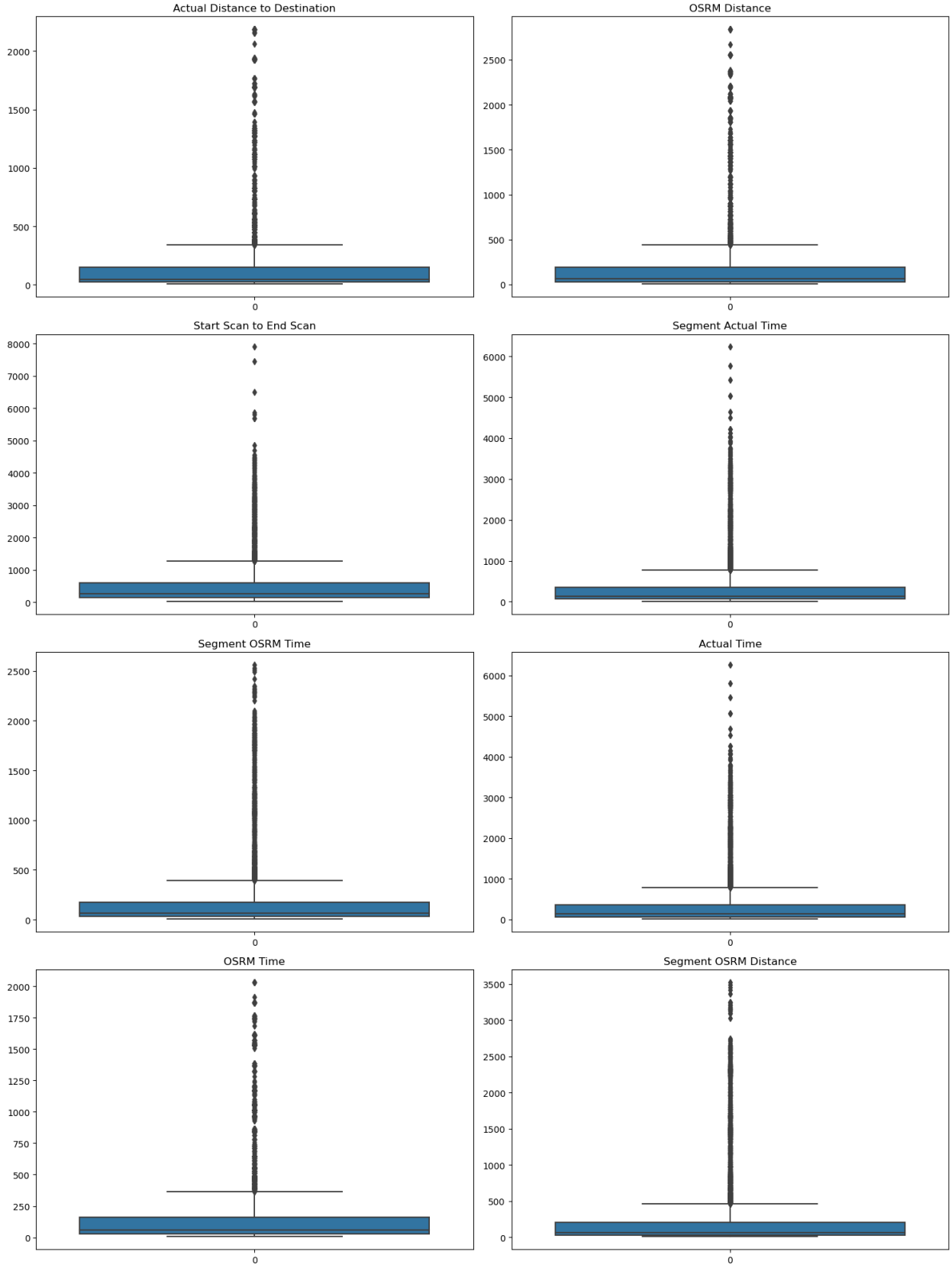
Ha: Mean of osrm_distance and segment_osrm_distance are not same (mu_1 != mu_2)

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

```
0.0001606670222265932
Reject Null Hypotheis, Mean of osrm_distance and segment_osrm_distance
are not same
```

Outliers Detection Using IQR Method

```
fig, axes = plt.subplots(nrows=4, ncols=2, figsize=(15, 20))

sns.boxplot(data["actual_distance_to_destination"], ax=axes[0, 0])
axes[0, 0].set_title('Actual Distance to Destination')

sns.boxplot(data["osrm_distance"], ax=axes[0, 1])
axes[0, 1].set_title('OSRM Distance')

sns.boxplot(data["start_scan_to_end_scan"], ax=axes[1, 0])
axes[1, 0].set_title('Start Scan to End Scan')

sns.boxplot(data["segment_actual_time"], ax=axes[1, 1])
axes[1, 1].set_title('Segment Actual Time')

sns.boxplot(data["segment_osrm_time"], ax=axes[2, 0])
axes[2, 0].set_title('Segment OSRM Time')

sns.boxplot(data["actual_time"], ax=axes[2, 1])
axes[2, 1].set_title('Actual Time')

sns.boxplot(data["osrm_time"], ax=axes[3, 0])
axes[3, 0].set_title('OSRM Time')

sns.boxplot(data["segment_osrm_distance"], ax=axes[3, 1])
axes[3, 1].set_title('Segment OSRM Distance')

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


Actual Distance to Destination
OSRM Distance
Start Scan to End Scan
Segment Actual Time
Segment OSRM Time
Actual Time
OSRM Time
Segment OSRM Distance
0

```
data.describe()
```

```
       cutoff_factor  actual_distance_to_destination  osrm_distance  \
count   13892.000000                    13892.000000   13892.000000
mean      159.986251                      160.852618     200.437664
std       307.520122                      307.627703     373.619022
min         9.000000                        9.002461       9.072900
25%        21.000000                       22.037144      29.802900
50%        46.000000                       46.163919      61.108100
75%       148.000000                      149.281573     193.689125
max      2185.000000                     2187.483994    2840.081000

       start_scan_to_end_scan  segment_actual_time  segment_osrm_time
\
count            13892.000000         13892.000000       13892.000000

mean               515.603657           346.118557         176.627627

std                660.357703           561.918712         316.580388

min                 23.000000             9.000000           6.000000

25%                144.000000            64.000000          30.000000

50%                264.000000           136.000000          62.000000

75%                595.000000           349.000000         176.000000

max               7898.000000          6230.000000        2564.000000


        actual_time     osrm_time  segment_osrm_distance
trip_creation_year  \
count  13892.000000  13892.000000           13892.000000
13892.0
mean     349.267492    157.980564             218.437771
2018.0
std      567.051777    274.050917             419.933226
0.0
min        9.000000      6.000000               9.072900
2018.0
25%       65.000000     29.000000              31.349950
2018.0
50%      138.000000     57.500000              65.614850
2018.0
75%      353.000000    163.250000             204.588700
2018.0
max     6265.000000   2032.000000            3523.632400
2018.0

       trip_creation_month  trip_creation_day
```

```
trip_time  \
count         13892.000000       13892.000000
13892
mean              9.114310          18.547653  0 days
06:45:04.119232843
std               0.318199           7.753191  0 days
09:33:47.061593031
min               9.000000           1.000000            0 days
00:00:00
25%               9.000000          14.000000  0 days
01:51:56.299656750
50%               9.000000          20.000000     0 days
03:30:54.417636
75%               9.000000          25.000000  0 days
07:04:21.465968500
max              10.000000          30.000000     5 days
11:38:33.117274

        triptime_sec
count   13892.000000
mean    24304.119233
std     34427.061593
min         0.000000
25%      6716.299657
50%     12654.417636
75%     25461.465969
max    473913.117274
```

Outliers Treatment

```
#Let's remove outliers using IQR Method
# Given values for 25th and 75th percentiles
sses_25th = 44
sses_75th = 595

# Calculate the Interquartile Range (IQR)
iqr = sses_75th - sses_25th

# Calculate the upper whisker using 1.5*IQR
upper_whisker = sses_75th + 1.5 * iqr

# Display the upper whisker value
upper_whisker
```

```
1421.5
```

```
data[data["start_scan_to_end_scan"]> upper_whisker]
```

```
      route_type                trip_uuid
trip_creation_time  \
875      Carting  trip-153688139008597141 2018-10-03 20:11:38.080296

1187     Carting  trip-153695358056452857 2018-09-28 21:57:19.197648

1221     Carting  trip-153695825653985603 2018-09-30 07:30:32.488364

1316     Carting  trip-153696880159964742 2018-09-21 20:55:53.211057

1402     Carting  trip-153697923702522443 2018-09-23 15:30:38.146740

...          ...                       ...                       ...

14744        FTL  trip-153860352246282031 2018-09-24 05:06:56.558662

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


      source_center                                    source_name  \
875    IND395023AAD                Surat_Central_I_4 (Gujarat)
1187   IND431112AAB              Sillod_ZebaTWR_D (Maharashtra)
1221   IND530012AAA  Visakhapatnam_Gajuwaka_IP (Andhra Pradesh)
1316   IND395023AAD                Surat_Central_I_4 (Gujarat)
1402   IND211002AAB       Allahabad_Central_H_1 (Uttar Pradesh)
...             ...                                       ...
14744  IND000000ACB              Gurgaon_Bilaspur_HB (Haryana)
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)

      destination_center
destination_name  \
875         IND396321AAA                       Bilimora_DC (Gujarat)

1187        IND431203AAA             Jalna_BhgyaNgr_D (Maharashtra)

1221        IND530012AAA  Visakhapatnam_Gajuwaka_IP (Andhra Pradesh)

1316        IND396321AAA                       Bilimora_DC (Gujarat)

1402        IND212402AAA         Phulpur_Shekhpur_D (Uttar Pradesh)

...                  ...                                       ...
```

```
14744       IND712311AAA             Kolkata_Dankuni_HB (West Bengal)

14748       IND712311AAA             Kolkata_Dankuni_HB (West Bengal)

14754       IND834002AAB                        Ranchi_Hub (Jharkhand)

14764       IND131028AAB                   Sonipat_Kundli_H (Haryana)

14773       IND000000ACB                Gurgaon_Bilaspur_HB (Haryana)


                    od_start_time                 od_end_time
cutoff_factor  \
875   2018-09-13 23:29:50.086266 2018-09-14 02:58:24.884963
137
1187  2018-09-15 04:40:33.384907 2018-09-16 01:06:53.762074
207
1221  2018-09-14 20:50:56.540060 2018-09-15 23:00:45.715950
194
1316  2018-09-14 23:46:41.600072 2018-09-15 03:21:56.351119
113
1402  2018-09-15 02:40:37.025475 2018-09-15 07:54:57.848178
93
...                           ...                       ...
...
14744 2018-10-03 21:52:02.463089 2018-10-05 09:06:25.403171
1300
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

       ...  actual_time  osrm_time  segment_osrm_distance
source_state  \
875    ...       1108.0      145.0               271.4230
Gujarat
1187   ...        586.0      169.0               295.9635
Maharashtra
1221   ...        303.0      169.0               228.7358  Andhra
Pradesh
1316   ...        258.0      127.0               206.7191
Gujarat
1402   ...       1204.0       90.0               117.4880   Uttar
Pradesh
...    ...          ...        ...                    ...
```

```
...
14744  ...       1930.0     1016.0              1497.6331
Haryana
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

      destination_state  trip_creation_year  trip_creation_month  \
875             Gujarat              2018.0                 10.0
1187        Maharashtra              2018.0                  9.0
1221      Andhra Pradesh             2018.0                  9.0
1316            Gujarat              2018.0                  9.0
1402       Uttar Pradesh             2018.0                  9.0
...                  ...                ...                  ...
14744        West Bengal              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

      trip_creation_day              trip_time   triptime_sec
875                 3.0 0 days 03:28:34.798697   12514.798697
1187               28.0 0 days 20:26:20.377167   73580.377167
1221               30.0 1 days 02:09:49.175890   94189.175890
1316               21.0 0 days 03:35:14.751047   12914.751047
1402               23.0 0 days 05:14:20.822703   18860.822703
...                 ...                    ...           ...
14744              24.0 1 days 11:14:22.940082  126862.940082
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

[1040 rows x 25 columns]

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

      cutoff_factor  actual_distance_to_destination  osrm_distance  \
count   12852.000000                   12852.000000   12852.000000
mean       88.108077                      88.962952     113.436786
std       108.676646                     108.964263     135.904653
min         9.000000                       9.002461       9.072900
```

```
25%        21.000000                        21.513693      28.613375
50%        39.000000                        39.995569      51.578350
75%       117.000000                       118.535713     151.928175
max       830.000000                       830.517272     970.943400

       start_scan_to_end_scan  segment_actual_time  segment_osrm_time
\
count            12852.000000         12852.000000       12852.000000

mean               363.109088           216.271709         103.763305

std                320.783878           237.050219         117.762542

min                 23.000000             9.000000           6.000000

25%                137.000000            61.000000          28.000000

50%                240.000000           119.000000          56.000000

75%                471.000000           282.000000         144.000000

max               1421.000000          1372.000000         867.000000


        actual_time     osrm_time  segment_osrm_distance
trip_creation_year  \
count  12852.000000  12852.000000           12852.000000
12852.0
mean     218.170557     94.566838             121.064327
2018.0
std      238.554090    105.277546             146.155991
0.0
min        9.000000      6.000000               9.072900
2018.0
25%       62.000000     27.750000              29.653100
2018.0
50%      120.000000     52.000000              56.815650
2018.0
75%      284.000000    125.000000             159.915625
2018.0
max     1372.000000    712.000000            1150.617300
2018.0

       trip_creation_month  trip_creation_day
trip_time  \
count          12852.000000       12852.000000
12852
mean               9.114846          18.584967  0 days
05:00:24.581706652
std                0.318848           7.786337  0 days
```

```
04:58:46.808724515
min               9.000000            1.000000            0 days
00:00:00
25%               9.000000           14.000000  0 days
01:51:31.947565250
50%               9.000000           20.000000     0 days
03:21:36.093887
75%               9.000000           25.000000  0 days
06:17:26.963298250
max              10.000000           30.000000     2 days
01:09:57.136511

        triptime_sec
count   12852.000000
mean    18024.581707
std     17926.808725
min         0.000000
25%      6691.947565
50%     12096.093887
75%     22646.963298
max    176997.136511
```

By removing outliers in the start_scan_to_end_scan column, we observe a significant decrease in the maximum values of other columns. Further dropping columns would lead to a loss of valuable data.

Data Encoding

```
data_encoding=data.copy()

data_encoding.shape

(12852, 25)
```

Label Encoding

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

col="route_type"
data_encoding[col].value_counts()

Carting    8503
FTL        4349
Name: route_type, dtype: int64
```

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

```
0    8503
1    4349
Name: route_type, dtype: int64
```

Target Encoding

```
!pip install category_encoders
from category_encoders import TargetEncoder
```

```
Defaulting to user installation because normal site-packages is not
writeable
Collecting category_encoders
  Downloading category_encoders-2.6.3-py2.py3-none-any.whl (81 kB)
                                              0.0/81.9 kB ? eta
-:--:--
     ---------------------------------      71.7/81.9 kB 2.0 MB/s
eta 0:00:01
     ---------------------------------------- 81.9/81.9 kB 1.1 MB/s
eta 0:00:00
Requirement already satisfied: numpy>=1.14.0 in c:\programdata\
anaconda3\lib\site-packages (from category_encoders) (1.24.3)
Requirement already satisfied: scikit-learn>=0.20.0 in c:\programdata\
anaconda3\lib\site-packages (from category_encoders) (1.2.2)
Requirement already satisfied: scipy>=1.0.0 in c:\programdata\
anaconda3\lib\site-packages (from category_encoders) (1.10.1)
Requirement already satisfied: statsmodels>=0.9.0 in c:\programdata\
anaconda3\lib\site-packages (from category_encoders) (0.13.5)
Requirement already satisfied: pandas>=1.0.5 in c:\programdata\
anaconda3\lib\site-packages (from category_encoders) (1.5.3)
Requirement already satisfied: patsy>=0.5.1 in c:\programdata\
anaconda3\lib\site-packages (from category_encoders) (0.5.3)
Requirement already satisfied: python-dateutil>=2.8.1 in c:\
programdata\anaconda3\lib\site-packages (from pandas>=1.0.5-
>category_encoders) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in c:\programdata\
anaconda3\lib\site-packages (from pandas>=1.0.5->category_encoders)
(2022.7)
Requirement already satisfied: six in c:\programdata\anaconda3\lib\
site-packages (from patsy>=0.5.1->category_encoders) (1.16.0)
Requirement already satisfied: joblib>=1.1.1 in c:\programdata\
anaconda3\lib\site-packages (from scikit-learn>=0.20.0-
>category_encoders) (1.2.0)
Requirement already satisfied: threadpoolctl>=2.0.0 in c:\programdata\
anaconda3\lib\site-packages (from scikit-learn>=0.20.0-
>category_encoders) (2.2.0)
```

```
Requirement already satisfied: packaging>=21.3 in c:\programdata\
anaconda3\lib\site-packages (from statsmodels>=0.9.0-
>category_encoders) (23.0)
Installing collected packages: category_encoders
Successfully installed category_encoders-2.6.3
```

```
te=TargetEncoder()

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

data_encoding.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 12852 entries, 0 to 14786
Data columns (total 25 columns):
 #   Column                           Non-Null Count  Dtype
---  ------                           --------------  -----
 0   route_type                       12852 non-null  int32
 1   trip_uuid                        12852 non-null  object
 2   trip_creation_time               12852 non-null  datetime64[ns]
 3   source_center                    12852 non-null  float64
 4   source_name                      12852 non-null  float64
 5   destination_center               12852 non-null  float64
 6   destination_name                 12852 non-null  float64
 7   od_start_time                    12852 non-null  datetime64[ns]
 8   od_end_time                      12852 non-null  datetime64[ns]
 9   cutoff_factor                    12852 non-null  int64
 10  actual_distance_to_destination   12852 non-null  float64
 11  osrm_distance                    12852 non-null  float64
 12  start_scan_to_end_scan           12852 non-null  float64
 13  segment_actual_time              12852 non-null  float64
 14  segment_osrm_time                12852 non-null  float64
 15  actual_time                      12852 non-null  float64
 16  osrm_time                        12852 non-null  float64
 17  segment_osrm_distance            12852 non-null  float64
 18  source_state                     12852 non-null  float64
 19  destination_state                12852 non-null  float64
 20  trip_creation_year               12852 non-null  float64
 21  trip_creation_month              12852 non-null  float64
 22  trip_creation_day                12852 non-null  float64
 23  trip_time                        12852 non-null  timedelta64[ns]
 24  triptime_sec                     12852 non-null  float64
dtypes: datetime64[ns](3), float64(18), int32(1), int64(1), object(1),
```

```
timedelta64[ns](1)
memory usage: 2.5+ MB
```

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

```
data_encoding
```

```
       route_type  source_center  source_name  destination_center  \
0               0       0.233481     0.233481        2.405789e-01
1               0       0.026667     0.026667        5.140380e-08
2               0       0.265919     0.265919        2.943634e-01
3               0       0.019398     0.019398        2.261087e-01
4               0       0.000924     0.000924        4.186833e-06
...           ...            ...          ...                 ...
14778           1       0.432241     0.432241        8.583065e-01
14780           1       0.636225     0.636225        9.449722e-01
14783           1       0.888157     0.888157        8.955545e-01
14785           1       0.636225     0.636225        6.362248e-01
14786           1       0.424472     0.424472        5.036211e-01

       destination_name  cutoff_factor  actual_distance_to_destination
\
0          2.405789e-01             72                       73.186911

1          5.140380e-08             17                       17.175274

2          2.943634e-01             24                       24.597048

3          2.261087e-01              9                        9.100510

4          4.186833e-06             22                       22.424210

...                 ...            ...                             ...

14778      8.583065e-01            143                      144.794266

14780      9.449722e-01            174                      176.546661

14783      8.955545e-01            192                      194.552260

14785      6.362248e-01             84                       84.743813

14786      5.036211e-01             65                       66.081533


       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
...              ...                     ...                   ...
14778       191.0458                   852.0                 270.0
14780       220.3007                   678.0                 378.0
14783       229.2052                  1017.0                 845.0
14785       102.3828                   256.0                 186.0
14786        80.5787                   353.0                 274.0

       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
...                  ...          ...        ...                      ..
.
14778              136.0        272.0      139.0
172.9107
14780              194.0        378.0      192.0
212.8530
14783              178.0        847.0      178.0
232.5811
14785               92.0        187.0       92.0
107.6915
14786               67.0        275.0       68.0
80.5787

       source_state  destination_state  trip_creation_year  \
0          0.132353           0.167293              2018.0
1          0.213544           0.189456              2018.0
2          0.263959           0.277888              2018.0
3          0.263959           0.277888              2018.0
4          0.132353           0.167293              2018.0
...             ...                ...                 ...
14778      0.417819           0.398413              2018.0
14780      0.456364           0.505843              2018.0
14783      0.940594           0.940594              2018.0
14785      0.456364           0.505843              2018.0
14786      0.132353           0.167293              2018.0

       trip_creation_month  trip_creation_day  triptime_sec
0                      9.0               20.0      0.000000
1                      9.0               20.0   6029.696112
2                      9.0               20.0      0.000000
```

```
3                      9.0                20.0   5880.338041
4                      9.0                20.0  10586.899446
...                    ...                 ...           ...
14778                  9.0                24.0  51184.266566
14780                  9.0                24.0  40779.500123
14783                  9.0                24.0  61169.217828
14785                  9.0                24.0      0.000000
14786                  9.0                24.0      0.000000

[12852 rows x 20 columns]
```

Standardization

```
#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)
std_data
```

```
       route_type  source_center  source_name  destination_center  \
0             0.0       0.234251     0.234251        2.521111e-01
1             0.0       0.026755     0.026755        5.386383e-08
2             0.0       0.266796     0.266796        3.084738e-01
3             0.0       0.019462     0.019462        2.369473e-01
4             0.0       0.000928     0.000928        4.387526e-06
...           ...            ...          ...                 ...
12847         1.0       0.433666     0.433666        8.994497e-01
12848         1.0       0.638322     0.638322        9.902697e-01
12849         1.0       0.891085     0.891085        9.384832e-01
12850         1.0       0.638322     0.638322        6.667225e-01
12851         1.0       0.425871     0.425871        5.277623e-01

       destination_name  cutoff_factor  actual_distance_to_destination
\
0          2.521111e-01       0.076736                        0.078129

1          5.386383e-08       0.009744                        0.009948

2          3.084738e-01       0.018270                        0.018983

3          2.369473e-01       0.000000                        0.000119

4          4.387526e-06       0.015834                        0.016338

...                 ...            ...                             ...

12847      8.994497e-01       0.163216                        0.165294

12848      9.902697e-01       0.200974                        0.203945
```

```
12849      9.384832e-01       0.222899                          0.225863

12850      6.667225e-01       0.091352                          0.092197

12851      5.277623e-01       0.068210                          0.069480


       osrm_distance  start_scan_to_end_scan  segment_actual_time  \
0           0.079052                0.112303             0.096845
1           0.011028                0.055079             0.036684
2           0.019745                0.118741             0.037417
3           0.003062                0.053648             0.011005
4           0.020634                0.087983             0.040352
...              ...                     ...                  ...
12847       0.189186                0.592990             0.191489
12848       0.219601                0.468526             0.270726
12849       0.228859                0.711016             0.613353
12850       0.097009                0.166667             0.129861
12851       0.074340                0.236052             0.194424

       segment_osrm_time  actual_time  osrm_time
segment_osrm_distance  \
0               0.068525     0.098313   0.087819
0.065803
1               0.011614     0.036684   0.012748
0.009464
2               0.019744     0.038151   0.024079
0.016637
3               0.008130     0.011005   0.009915
0.002580
4               0.032520     0.040352   0.039660
0.017386
...                  ...          ...        ...                       ..
.
12847           0.150987     0.192957   0.188385
0.143523
12848           0.218351     0.270726   0.263456
0.178513
12849           0.199768     0.614820   0.243626
0.195795
12850           0.099884     0.130594   0.121813
0.086391
12851           0.070848     0.195158   0.087819
0.062640

       source_state  destination_state  trip_creation_year  \
0          0.108790           0.094519                 0.0
1          0.197365           0.118627                 0.0
2          0.252366           0.214823                 0.0
```

```
3          0.252366           0.214823                 0.0
4          0.108790           0.094519                 0.0
...             ...                ...                 ...
12847      0.420218           0.345928                 0.0
12848      0.462268           0.462789                 0.0
12849      0.990537           0.935705                 0.0
12850      0.462268           0.462789                 0.0
12851      0.108790           0.094519                 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
...                    ...                ...           ...
12847                  0.0           0.793103      0.289181
12848                  0.0           0.793103      0.230396
12849                  0.0           0.793103      0.345594
12850                  0.0           0.793103      0.000000
12851                  0.0           0.793103      0.000000

[12852 rows x 20 columns]
```

Business Insights

1. **Hypothesis Testing**:
   - A hypothesis test between OSRM data and actual data shows that the means of both datasets are significantly different.
2. **Correlation**:
   - Distance and time attributes are highly correlated. Therefore, shorter distances between places contribute to faster delivery times.
3. **Order Distribution**:
   - The majority of orders come from Maharashtra, indicating a high customer base in the state.
   - The fewest trips are from the North-Eastern states, suggesting the need for business improvement in these regions.
4. **Warehouse Activity**:
   - The busiest warehouses are Gurgaon_Bilaspur, Bhiwandi, and Bangalore, which should be prioritized for operational focus.
5. **Busiest Route**:
   - The busiest route is from Bangalore_Nalamangla_H (Karnataka) to Bengaluru_KGAirport_HB (Karnataka), with an average distance of 28.03 km and an average travel time of 87.87 minutes.

Recommendations

1. **Optimize Delivery Times**:

- The actual delivery time is longer than the OSRM time. Services can be optimized using OSRM data to reduce delivery times.

2. **Expand in Bengaluru**:
   - Given the busy route in Bengaluru, consider increasing services by opening more outlets and hiring additional staff.
3. **Increase Outlets in Maharashtra**:
   - With the highest number of trips in Maharashtra, increasing the number of outlets in the state can cater to the growing customer base.
4. **Enhance Presence in North-Eastern States**:
   - The low business volume in the North-Eastern states indicates a need for improved conditions and marketing efforts to boost services in the region.
5. **Improve Warehouse Efficiency**:
   - For the busiest warehouse in Gurgaon_Bilaspur, consider increasing the number of warehouses or boosting manpower to manage the load effectively.
6. **Utilize OSRM for Route Planning**:
   - OSRM provides optimal minimal distance calculations, which can be used to enhance route planning and reduce travel times.