```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv('delhivery_data.csv')

df.head()
```

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

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

                 trip_uuid source_center
source_name  \
0  trip-153741093647649320  IND388121AAA  Anand_VUNagar_DC (Gujarat)

1  trip-153741093647649320  IND388121AAA  Anand_VUNagar_DC (Gujarat)

2  trip-153741093647649320  IND388121AAA  Anand_VUNagar_DC (Gujarat)

3  trip-153741093647649320  IND388121AAA  Anand_VUNagar_DC (Gujarat)

4  trip-153741093647649320  IND388121AAA  Anand_VUNagar_DC (Gujarat)


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

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

   actual_distance_to_destination  actual_time  osrm_time
```

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

     factor  segment_actual_time  segment_osrm_time
segment_osrm_distance  \
0  1.272727                 14.0               11.0
11.9653
1  1.200000                 10.0                9.0
9.7590
2  1.428571                 16.0                7.0
10.8152
3  1.550000                 21.0               12.0
13.0224
4  1.545455                  6.0                5.0
3.9153

   segment_factor
0        1.272727
1        1.111111
2        2.285714
3        1.750000
4        1.200000

[5 rows x 24 columns]
```

Removing null values

```
df = df.dropna(how='any')

df.info()

<class 'pandas.core.frame.DataFrame'>
Index: 144316 entries, 0 to 144866
Data columns (total 24 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   data                            144316 non-null  object
 1   trip_creation_time              144316 non-null  object
 2   route_schedule_uuid             144316 non-null  object
 3   route_type                      144316 non-null  object
 4   trip_uuid                       144316 non-null  object
 5   source_center                   144316 non-null  object 
 6   source_name                     144316 non-null  object 
 7   destination_center              144316 non-null  object 
 8   destination_name                144316 non-null  object 
 9   od_start_time                   144316 non-null  object 
 10  od_end_time                     144316 non-null  object 
 11  start_scan_to_end_scan          144316 non-null  float64
 12  is_cutoff                       144316 non-null  bool   
 13  cutoff_factor                   144316 non-null  int64  
 14  cutoff_timestamp                144316 non-null  object 
 15  actual_distance_to_destination  144316 non-null  float64
 16  actual_time                     144316 non-null  float64
 17  osrm_time                       144316 non-null  float64
 18  osrm_distance                   144316 non-null  float64
 19  factor                          144316 non-null  float64
 20  segment_actual_time             144316 non-null  float64
 21  segment_osrm_time               144316 non-null  float64
 22  segment_osrm_distance           144316 non-null  float64
 23  segment_factor                  144316 non-null  float64
dtypes: bool(1), float64(10), int64(1), object(12)
memory usage: 26.6+ MB
```

```
df.describe()
```

```
       start_scan_to_end_scan  cutoff_factor  
actual_distance_to_destination  \
count           144316.000000  144316.000000   
144316.000000   
mean               963.697698     233.561345   
234.708498   
std               1038.082976     345.245823   
345.480571   
min                 20.000000       9.000000   
9.000045   
25%                161.000000      22.000000   
23.352027   
50%                451.000000      66.000000   
66.135322   
75%               1645.000000     286.000000   
286.919294   
max               7898.000000    1927.000000   
1927.447705   

         actual_time      osrm_time  osrm_distance         factor  \
count  144316.000000  144316.000000  144316.000000  144316.000000   
mean      417.996237     214.437055     285.549785       2.120178   
std       598.940065     308.448543     421.717826       1.717065   
min         9.000000       6.000000       9.008200       0.144000   
25%        51.000000      27.000000      29.896250       1.604545   
50%       132.000000      64.000000      78.624400       1.857143   
```

```
75%       516.000000     259.000000     346.305400       2.212280
max      4532.000000    1686.000000    2326.199100      77.387097

       segment_actual_time  segment_osrm_time
segment_osrm_distance  \
count        144316.000000      144316.000000          144316.000000

mean             36.175379          18.495697              22.818993

std              53.524298          14.774008              17.866367

min            -244.000000           0.000000               0.000000

25%              20.000000          11.000000              12.053975

50%              28.000000          17.000000              23.508300

75%              40.000000          22.000000              27.813325

max            3051.000000        1611.000000            2191.403700


       segment_factor
count   144316.000000
mean         2.218707
std          4.854804
min        -23.444444
25%          1.347826
50%          1.684211
75%          2.250000
max        574.250000
```

```
df['od_start_time'] = pd.to_datetime(df['od_start_time'])
df['od_end_time'] = pd.to_datetime(df['od_end_time'])

df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 144316 entries, 0 to 144866
Data columns (total 24 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   data                            144316 non-null  object
 1   trip_creation_time              144316 non-null  object
 2   route_schedule_uuid             144316 non-null  object
 3   route_type                      144316 non-null  object
 4   trip_uuid                       144316 non-null  object
 5   source_center                   144316 non-null  object
 6   source_name                     144316 non-null  object
 7   destination_center              144316 non-null  object
 8   destination_name                144316 non-null  object
 9   od_start_time                   144316 non-null  datetime64[ns]
 10  od_end_time                     144316 non-null  datetime64[ns]
 11  start_scan_to_end_scan          144316 non-null  float64
 12  is_cutoff                       144316 non-null  bool
 13  cutoff_factor                   144316 non-null  int64
 14  cutoff_timestamp                144316 non-null  object
 15  actual_distance_to_destination  144316 non-null  float64
 16  actual_time                     144316 non-null  float64
 17  osrm_time                       144316 non-null  float64
 18  osrm_distance                   144316 non-null  float64
 19  factor                          144316 non-null  float64
 20  segment_actual_time             144316 non-null  float64
 21  segment_osrm_time               144316 non-null  float64
 22  segment_osrm_distance           144316 non-null  float64
 23  segment_factor                  144316 non-null  float64
dtypes: bool(1), datetime64[ns](2), float64(10), int64(1), object(10)
memory usage: 26.6+ MB
```

```
df['segment_key'] = df['trip_uuid'] + df['source_center'] +
df['destination_center']

segment_cols = ['segment_actual_time', 'segment_osrm_distance',
'segment_osrm_time']

for col in segment_cols:
    df[col + '_sum'] = df.groupby('segment_key')[col].cumsum()
df[[col + '_sum' for col in segment_cols]]
```

```
        segment_actual_time_sum  segment_osrm_distance_sum  \
0                          14.0                    11.9653
1                          24.0                    21.7243
2                          40.0                    32.5395
3                          61.0                    45.5619
4                          67.0                    49.4772
...                         ...                        ...
144862                     92.0                    65.3487
144863                    118.0                    82.7212
144864                    138.0                   103.4265
144865                    155.0                   122.3150
144866                    423.0                   131.1238

        segment_osrm_time_sum
0                        11.0
1                        20.0
2                        27.0
3                        39.0
4                        44.0
...                       ...
144862                   94.0
144863                  115.0
```

```
144864                   149.0
144865                   176.0
144866                   185.0

[144316 rows x 3 columns]

create_segment_dict = {

    'data' : 'first',
    'trip_creation_time' : 'first',
    'route_schedule_uuid' : 'first',
    'route_type' : 'first',
    'trip_uuid' : 'first',
    'source_center' : 'first',
    'source_name' : 'first',

    'destination_center' : 'last',
    'destination_name' : 'last',

    'od_start_time' : 'first',
    'od_end_time' : 'first',
    'start_scan_to_end_scan' : 'first',

    'actual_distance_to_destination' : 'last',
    'actual_time' : 'last',

    'osrm_time' : 'last',
    'osrm_distance' : 'last',

    'segment_actual_time_sum' : 'last',
    'segment_osrm_distance_sum' : 'last',
    'segment_osrm_time_sum' : 'last',

}
```

Grouping mini-trips, sorting by time

```
segment =
df.groupby('segment_key').agg(create_segment_dict).reset_index()
segment = segment.sort_values(by=['segment_key','od_end_time'],
ascending=True).reset_index()

segment

       index                                     segment_key
data  \
0          0  trip-153671041653548748IND209304AAAIND000000ACB
training
1          1  trip-153671041653548748IND462022AAAIND209304AAA
training
```

```
2          2  trip-153671042288605164IND561203AABIND562101AAA
training
3          3  trip-153671042288605164IND572101AAAIND561203AAB
training
4          4  trip-153671043369099517IND000000ACBIND160002AAC
training
...      ...                                              ...        ..
.
26217  26217  trip-153861115439069069IND628204AAAIND627657AAA
test
26218  26218  trip-153861115439069069IND628613AAAIND627005AAA
test
26219  26219  trip-153861115439069069IND628801AAAIND628204AAA
test
26220  26220  trip-153861118270144424IND583119AAAIND583101AAA
test
26221  26221  trip-153861118270144424IND583201AAAIND583119AAA
test

               trip_creation_time  \
0      2018-09-12 00:00:16.535741
1      2018-09-12 00:00:16.535741
2      2018-09-12 00:00:22.886430
3      2018-09-12 00:00:22.886430
4      2018-09-12 00:00:33.691250
...                           ...
26217  2018-10-03 23:59:14.390954
26218  2018-10-03 23:59:14.390954
26219  2018-10-03 23:59:14.390954
26220  2018-10-03 23:59:42.701692
26221  2018-10-03 23:59:42.701692

                                     route_schedule_uuid route_type  \
0      thanos::sroute:d7c989ba-a29b-4a0b-b2f4-288cdc6...        FTL
1      thanos::sroute:d7c989ba-a29b-4a0b-b2f4-288cdc6...        FTL
2      thanos::sroute:3a1b0ab2-bb0b-4c53-8c59-eb2a2c0...    Carting
3      thanos::sroute:3a1b0ab2-bb0b-4c53-8c59-eb2a2c0...    Carting
4      thanos::sroute:de5e208e-7641-45e6-8100-4d9fb1e...        FTL
...                                                  ...        ...
26217  thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...    Carting
26218  thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...    Carting
26219  thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...    Carting
26220  thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...        FTL
26221  thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...        FTL

                     trip_uuid source_center  \
0      trip-153671041653548748  IND209304AAA
1      trip-153671041653548748  IND462022AAA
2      trip-153671042288605164  IND561203AAB
3      trip-153671042288605164  IND572101AAA
```

```
4      trip-153671043369099517  IND000000ACB
...                        ...           ...
26217  trip-153861115439069069  IND628204AAA
26218  trip-153861115439069069  IND628613AAA
26219  trip-153861115439069069  IND628801AAA
26220  trip-153861118270144424  IND583119AAA
26221  trip-153861118270144424  IND583201AAA

                              source_name destination_center  ...  \
0      Kanpur_Central_H_6 (Uttar Pradesh)       IND000000ACB  ...
1      Bhopal_Trnsport_H (Madhya Pradesh)       IND209304AAA  ...
2       Doddablpur_ChikaDPP_D (Karnataka)       IND562101AAA  ...
3           Tumkur_Veersagr_I (Karnataka)       IND561203AAB  ...
4           Gurgaon_Bilaspur_HB (Haryana)       IND160002AAC  ...
...                                   ...                ...  ...
26217  Tirchchndr_Shnmgprm_D (Tamil Nadu)       IND627657AAA  ...
26218   Peikulam_SriVnktpm_D (Tamil Nadu)       IND627005AAA  ...
26219        Eral_Busstand_D (Tamil Nadu)       IND628204AAA  ...
26220       Sandur_WrdN1DPP_D (Karnataka)       IND583101AAA  ...
26221                  Hospet (Karnataka)       IND583119AAA  ...

                   od_start_time                od_end_time  \
0     2018-09-12 16:39:46.858469 2018-09-13 13:40:23.123744
1     2018-09-12 00:00:16.535741 2018-09-12 16:39:46.858469
2     2018-09-12 02:03:09.655591 2018-09-12 03:01:59.598855
3     2018-09-12 00:00:22.886430 2018-09-12 02:03:09.655591
4     2018-09-14 03:40:17.106733 2018-09-14 17:34:55.442454
...                          ...                        ...
26217 2018-10-04 02:29:04.272194 2018-10-04 03:31:11.183797
26218 2018-10-04 04:16:39.894872 2018-10-04 05:47:45.162682
26219 2018-10-04 01:44:53.808000 2018-10-04 02:29:04.272194
26220 2018-10-04 03:58:40.726547 2018-10-04 08:46:09.166940
26221 2018-10-04 02:51:44.712656 2018-10-04 03:58:40.726547

      start_scan_to_end_scan  actual_distance_to_destination
actual_time  \
0                     1260.0                      383.759164
732.0
1                      999.0                      440.973689
830.0
2                       58.0                       24.644021
47.0
3                      122.0                       48.542890
96.0
4                      834.0                      237.439610
611.0
...                      ...                             ...
...
26217                   62.0                       33.627182
51.0
```

```
26218                   91.0                        33.673835
90.0
26219                   44.0                        12.661945
30.0
26220                  287.0                        40.546740
233.0
26221                   66.0                        25.534793
42.0

       osrm_time  osrm_distance  segment_actual_time_sum  \
0          329.0       446.5496                    728.0
1          388.0       544.8027                    820.0
2           26.0        28.1994                     46.0
3           42.0        56.9116                     95.0
4          212.0       281.2109                    608.0
...          ...            ...                      ...
26217       41.0        42.5213                     49.0
26218       48.0        40.6080                     89.0
26219       14.0        16.0185                     29.0
26220       42.0        52.5303                    233.0
26221       26.0        28.0484                     41.0

       segment_osrm_distance_sum  segment_osrm_time_sum
0                       670.6205                  534.0
1                       649.8528                  474.0
2                        28.1995                   26.0
3                        55.9899                   39.0
4                       317.7408                  231.0
...                          ...                    ...
26217                    42.1431                   42.0
26218                    78.5869                   77.0
26219                    16.0184                   14.0
26220                    52.5303                   42.0
26221                    28.0484                   25.0

[26222 rows x 21 columns]
```

Calculate time taken between od_start_time and od_end_time and keep it as a feature

```
segment['od_time_diff_hour'] = (segment['od_end_time'] -
segment['od_start_time']).dt.total_seconds() /(60)
segment['od_time_diff_hour']

0        1260.604421
1         999.505379
2          58.832388
3         122.779486
4         834.638929
             ...
```

```
26217      62.115193
26218      91.087797
26219      44.174403
26220     287.474007
26221      66.933565
Name: od_time_diff_hour, Length: 26222, dtype: float64

segment

       index                                      segment_key
data  \
0          0  trip-153671041653548748IND209304AAAIND000000ACB
training
1          1  trip-153671041653548748IND462022AAAIND209304AAA
training
2          2  trip-153671042288605164IND561203AABIND562101AAA
training
3          3  trip-153671042288605164IND572101AAAIND561203AAB
training
4          4  trip-153671043369099517IND000000ACBIND160002AAC
training
...      ...                                              ...         ..
.
26217  26217  trip-153861115439069069IND628204AAAIND627657AAA
test
26218  26218  trip-153861115439069069IND628613AAAIND627005AAA
test
26219  26219  trip-153861115439069069IND628801AAAIND628204AAA
test
26220  26220  trip-153861118270144424IND583119AAAIND583101AAA
test
26221  26221  trip-153861118270144424IND583201AAAIND583119AAA
test

               trip_creation_time  \
0      2018-09-12 00:00:16.535741
1      2018-09-12 00:00:16.535741
2      2018-09-12 00:00:22.886430
3      2018-09-12 00:00:22.886430
4      2018-09-12 00:00:33.691250
...                           ...
26217  2018-10-03 23:59:14.390954
26218  2018-10-03 23:59:14.390954
26219  2018-10-03 23:59:14.390954
26220  2018-10-03 23:59:42.701692
26221  2018-10-03 23:59:42.701692

                                     route_schedule_uuid route_type  \
0      thanos::sroute:d7c989ba-a29b-4a0b-b2f4-288cdc6...        FTL
1      thanos::sroute:d7c989ba-a29b-4a0b-b2f4-288cdc6...        FTL
```

```
2      thanos::sroute:3a1b0ab2-bb0b-4c53-8c59-eb2a2c0...    Carting
3      thanos::sroute:3a1b0ab2-bb0b-4c53-8c59-eb2a2c0...    Carting
4      thanos::sroute:de5e208e-7641-45e6-8100-4d9fb1e...        FTL
...                                                  ...        ...
26217  thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...    Carting
26218  thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...    Carting
26219  thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...    Carting
26220  thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...        FTL
26221  thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...        FTL

                     trip_uuid source_center  \
0      trip-153671041653548748  IND209304AAA   
1      trip-153671041653548748  IND462022AAA   
2      trip-153671042288605164  IND561203AAB   
3      trip-153671042288605164  IND572101AAA   
4      trip-153671043369099517  IND000000ACB   
...                        ...           ...   
26217  trip-153861115439069069  IND628204AAA   
26218  trip-153861115439069069  IND628613AAA   
26219  trip-153861115439069069  IND628801AAA   
26220  trip-153861118270144424  IND583119AAA   
26221  trip-153861118270144424  IND583201AAA   

                              source_name destination_center  ...  \
0      Kanpur_Central_H_6 (Uttar Pradesh)       IND000000ACB  ...   
1      Bhopal_Trnsport_H (Madhya Pradesh)       IND209304AAA  ...   
2       Doddablpur_ChikaDPP_D (Karnataka)       IND562101AAA  ...   
3           Tumkur_Veersagr_I (Karnataka)       IND561203AAB  ...   
4           Gurgaon_Bilaspur_HB (Haryana)       IND160002AAC  ...   
...                                   ...                ...  ...   
26217  Tirchchndr_Shnmgprm_D (Tamil Nadu)       IND627657AAA  ...   
26218   Peikulam_SriVnktpm_D (Tamil Nadu)       IND627005AAA  ...   
26219        Eral_Busstand_D (Tamil Nadu)       IND628204AAA  ...   
26220       Sandur_WrdN1DPP_D (Karnataka)       IND583101AAA  ...   
26221                  Hospet (Karnataka)       IND583119AAA  ...   

                     od_end_time start_scan_to_end_scan  \
0     2018-09-13 13:40:23.123744                 1260.0   
1     2018-09-12 16:39:46.858469                  999.0   
2     2018-09-12 03:01:59.598855                   58.0   
3     2018-09-12 02:03:09.655591                  122.0   
4     2018-09-14 17:34:55.442454                  834.0   
...                          ...                    ...   
26217 2018-10-04 03:31:11.183797                   62.0   
26218 2018-10-04 05:47:45.162682                   91.0   
26219 2018-10-04 02:29:04.272194                   44.0   
26220 2018-10-04 08:46:09.166940                  287.0   
26221 2018-10-04 03:58:40.726547                   66.0   

      actual_distance_to_destination  actual_time  osrm_time
```

```
osrm_distance  \
0                         383.759164        732.0      329.0
446.5496
1                         440.973689        830.0      388.0
544.8027
2                          24.644021         47.0       26.0
28.1994
3                          48.542890         96.0       42.0
56.9116
4                         237.439610        611.0      212.0
281.2109
...                              ...          ...        ...
...
26217                      33.627182         51.0       41.0
42.5213
26218                      33.673835         90.0       48.0
40.6080
26219                      12.661945         30.0       14.0
16.0185
26220                      40.546740        233.0       42.0
52.5303
26221                      25.534793         42.0       26.0
28.0484

       segment_actual_time_sum  segment_osrm_distance_sum  \
0                        728.0                   670.6205
1                        820.0                   649.8528
2                         46.0                    28.1995
3                         95.0                    55.9899
4                        608.0                   317.7408
...                        ...                        ...
26217                     49.0                    42.1431
26218                     89.0                    78.5869
26219                     29.0                    16.0184
26220                    233.0                    52.5303
26221                     41.0                    28.0484

       segment_osrm_time_sum  od_time_diff_hour
0                      534.0        1260.604421
1                      474.0         999.505379
2                       26.0          58.832388
3                       39.0         122.779486
4                      231.0         834.638929
...                      ...                ...
26217                   42.0          62.115193
26218                   77.0          91.087797
26219                   14.0          44.174403
26220                   42.0         287.474007
26221                   25.0          66.933565
```

```
[26222 rows x 22 columns]

create_trip_dict = {

    'data' : 'first',
    'trip_creation_time' : 'first',
    'route_schedule_uuid' : 'first',
    'route_type' : 'first',
    'trip_uuid' : 'first',

    'source_center' : 'first',
    'source_name' : 'first',

    'destination_center' : 'last',
    'destination_name' : 'last',

    'start_scan_to_end_scan' : 'sum',
    'od_time_diff_hour' : 'sum',

    'actual_distance_to_destination' : 'sum',
    'actual_time' : 'sum',
    'osrm_time' : 'sum',
    'osrm_distance' : 'sum',

    'segment_actual_time_sum' : 'sum',
    'segment_osrm_distance_sum' : 'sum',
    'segment_osrm_time_sum' : 'sum',

}



trip =
segment.groupby('trip_uuid').agg(create_trip_dict).reset_index(drop =
True)


trip

           data          trip_creation_time  \
0      training  2018-09-12 00:00:16.535741
1      training  2018-09-12 00:00:22.886430
2      training  2018-09-12 00:00:33.691250
3      training  2018-09-12 00:01:00.113710
4      training  2018-09-12 00:02:09.740725
...         ...                         ...
14782      test  2018-10-03 23:55:56.258533
14783      test  2018-10-03 23:57:23.863155
14784      test  2018-10-03 23:57:44.429324
14785      test  2018-10-03 23:59:14.390954
```

```
14786      test  2018-10-03 23:59:42.701692

                                     route_schedule_uuid route_type  \
0      thanos::sroute:d7c989ba-a29b-4a0b-b2f4-288cdc6...        FTL
1      thanos::sroute:3a1b0ab2-bb0b-4c53-8c59-eb2a2c0...    Carting
2      thanos::sroute:de5e208e-7641-45e6-8100-4d9fb1e...        FTL
3      thanos::sroute:f0176492-a679-4597-8332-bbd1c7f...    Carting
4      thanos::sroute:d9f07b12-65e0-4f3b-bec8-df06134...        FTL
...                                                  ...        ...
14782  thanos::sroute:8a120994-f577-4491-9e4b-b7e4a14...    Carting
14783  thanos::sroute:b30e1ec3-3bfa-4bd2-a7fb-3b75769...    Carting
14784  thanos::sroute:5609c268-e436-4e0a-8180-3db4a74...    Carting
14785  thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...    Carting
14786  thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...        FTL

                      trip_uuid source_center  \
0      trip-153671041653548748  IND209304AAA
1      trip-153671042288605164  IND561203AAB
2      trip-153671043369099517  IND000000ACB
3      trip-153671046011330457  IND400072AAB
4      trip-153671052974046625  IND583101AAA
...                        ...           ...
14782  trip-153861095625827784  IND160002AAC
14783  trip-153861104386292051  IND121004AAB
14784  trip-153861106442901555  IND208006AAA
14785  trip-153861115439069069  IND627005AAA
14786  trip-153861118270144424  IND583119AAA

                               source_name destination_center  \
0       Kanpur_Central_H_6 (Uttar Pradesh)       IND209304AAA
1        Doddablpur_ChikaDPP_D (Karnataka)       IND561203AAB
2            Gurgaon_Bilaspur_HB (Haryana)       IND000000ACB
3                 Mumbai Hub (Maharashtra)       IND401104AAA
4                   Bellary_Dc (Karnataka)       IND583119AAA
...                                    ...                ...
14782       Chandigarh_Mehmdpur_H (Punjab)       IND160002AAC
14783         FBD_Balabhgarh_DPC (Haryana)       IND121004AAA
14784   Kanpur_GovndNgr_DC (Uttar Pradesh)       IND208006AAA
14785  Tirunelveli_VdkkuSrt_I (Tamil Nadu)       IND628204AAA
14786        Sandur_WrdN1DPP_D (Karnataka)       IND583119AAA

                         destination_name  start_scan_to_end_scan  \
0      Kanpur_Central_H_6 (Uttar Pradesh)                  2259.0
1       Doddablpur_ChikaDPP_D (Karnataka)                   180.0
2           Gurgaon_Bilaspur_HB (Haryana)                  3933.0
3          Mumbai_MiraRd_IP (Maharashtra)                   100.0
4           Sandur_WrdN1DPP_D (Karnataka)                   717.0
...                                   ...                     ...
14782      Chandigarh_Mehmdpur_H (Punjab)                   257.0
14783      Faridabad_Blbgarh_DC (Haryana)                    60.0
```

```
14784  Kanpur_GovndNgr_DC (Uttar Pradesh)                      421.0
14785  Tirchchndr_Shnmgprm_D (Tamil Nadu)                      347.0
14786       Sandur_WrdN1DPP_D (Karnataka)                      353.0

       od_time_diff_hour  actual_distance_to_destination  actual_time
\
0            2260.109800                      824.732854       1562.0

1             181.611874                       73.186911        143.0

2            3934.362520                     1927.404273       3347.0

3             100.494935                       17.175274         59.0

4             718.349042                      127.448500        341.0

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

14782         258.028928                       57.762332         83.0

14783          60.590521                       15.513784         21.0

14784         422.119867                       38.684839        282.0

14785         348.512862                      134.723836        264.0

14786         354.407571                       66.081533        275.0


       osrm_time  osrm_distance  segment_actual_time_sum  \
0          717.0       991.3523                   1548.0
1           68.0        85.1110                    141.0
2         1740.0      2354.0665                   3308.0
3           15.0        19.6800                     59.0
4          117.0       146.7918                    340.0
...          ...            ...                      ...
14782       62.0        73.4630                     82.0
14783       12.0        16.0882                     21.0
14784       48.0        58.9037                    281.0
14785      179.0       171.1103                    258.0
14786       68.0        80.5787                    274.0

       segment_osrm_distance_sum  segment_osrm_time_sum
0                      1320.4733                 1008.0
1                        84.1894                   65.0
2                      2545.2678                 1941.0
3                        19.8766                   16.0
4                       146.7919                  115.0
...                          ...                    ...
14782                    64.8551                   62.0
14783                    16.0883                   11.0
```

```
14784                   104.8866                   88.0
14785                   223.5324                  221.0
14786                    80.5787                   67.0

[14787 rows x 18 columns]
```

```
trip[['actual_time', 'segment_actual_time_sum']]
```

```
       actual_time  segment_actual_time_sum
0           1562.0                   1548.0
1            143.0                    141.0
2           3347.0                   3308.0
3             59.0                     59.0
4            341.0                    340.0
...            ...                      ...
14782         83.0                     82.0
14783         21.0                     21.0
14784        282.0                    281.0
14785        264.0                    258.0
14786        275.0                    274.0

[14787 rows x 2 columns]
```

```
trip[['actual_distance_to_destination', 'osrm_distance']]
```

```
       actual_distance_to_destination  osrm_distance
0                          824.732854       991.3523
1                           73.186911        85.1110
2                         1927.404273      2354.0665
3                           17.175274        19.6800
4                          127.448500       146.7918
...                               ...            ...
14782                       57.762332        73.4630
14783                       15.513784        16.0882
14784                       38.684839        58.9037
14785                      134.723836       171.1103
14786                       66.081533        80.5787

[14787 rows x 2 columns]
```

```
trip['destination_name'] = trip['destination_name'].str.lower() #
lowering all columns
trip['source_name'] = trip['source_name']

def place2state(x):
    # transform "gurgaon_bilaspur_hb (haryana)" into "haryana"
    state = x.split('(')[1]

    return state[:-1] #removing ')' from ending
```

```python
def place2city(x):
    #we will remove state
    city = x.split(' (')[0]

    city = city.split('_')[0]

    # Now daling with edge cases

    if city == 'pnq vadgaon sheri dpc': return 'vadgaonsheri'

    # ['PNQ Pashan DPC', 'Bhopal MP Nagar', 'HBR Layout PC',
    #  'PNQ Rahatani DPC', 'Pune Balaji Nagar', 'Mumbai Antop Hill']

    if city in ['pnq pashan dpc','pnq rahatani dpc', 'pune balaji
nagar']:
        return 'pune'

    if city == 'hbr layout pc' :
        return 'bengaluru'
    if city == 'bhopal mp nagar':
        return 'bhopal'
    if city == 'mumbai antop hill':
        return 'mumbai'

    return city

def place2city_place(x):

    # we will remove state
    x = x.split('(')[0]

    len_ = len(x.split('_'))

    if len_ >= 3:
        return x.split('_')[1]

    # small cities have same city and place name
    if len_ == 2:
        return x.split('_')[0]

    # now we need to deal with edge cases or imporper name convention

    # if len(x.split('_')) == 2:

    return x.split(' ')[0]

def place2code(x):
    # we will remove state
    x = x.split('(')[0]
    if len(x.split('_')) >= 3:
        return x.split('_')[-1]

    return 'none'

trip['destination_state'] = trip['destination_name'].apply(lambda x:
place2state(x))
trip['destination_city']  = trip['destination_name'].apply(lambda x:
place2city(x))
trip['destination_place'] = trip['destination_name'].apply(lambda x:
place2city_place(x))
trip['destination_code']  = trip['destination_name'].apply(lambda x:
place2code(x))

trip[['destination_state','destination_city','destination_place','dest
ination_code']]

      destination_state destination_city destination_place
destination_code
0         uttar pradesh           kanpur           central
6
1             karnataka       doddablpur          chikadpp
d
2               haryana          gurgaon          bilaspur
hb
3           maharashtra           mumbai            mirard
ip
4             karnataka           sandur          wrdn1dpp
d
...                 ...              ...               ...
...
14782            punjab       chandigarh          mehmdpur
h
14783           haryana        faridabad           blbgarh
dc
14784     uttar pradesh           kanpur          govndngr
dc
14785        tamil nadu       tirchchndr          shnmgprm
d
14786         karnataka           sandur          wrdn1dpp
d

[14787 rows x 4 columns]



trip['trip_creation_time'] =
pd.to_datetime(trip['trip_creation_time'])

trip['trip_year'] = trip['trip_creation_time'].dt.year
trip['trip_month'] = trip['trip_creation_time'].dt.month
```

```
trip['trip_hour'] = trip['trip_creation_time'].dt.hour
trip['trip_day'] = trip['trip_creation_time'].dt.day
trip['trip_week'] = trip['trip_creation_time'].dt.isocalendar().week
trip['trip_dayofweek'] = trip['trip_creation_time'].dt.dayofweek
```

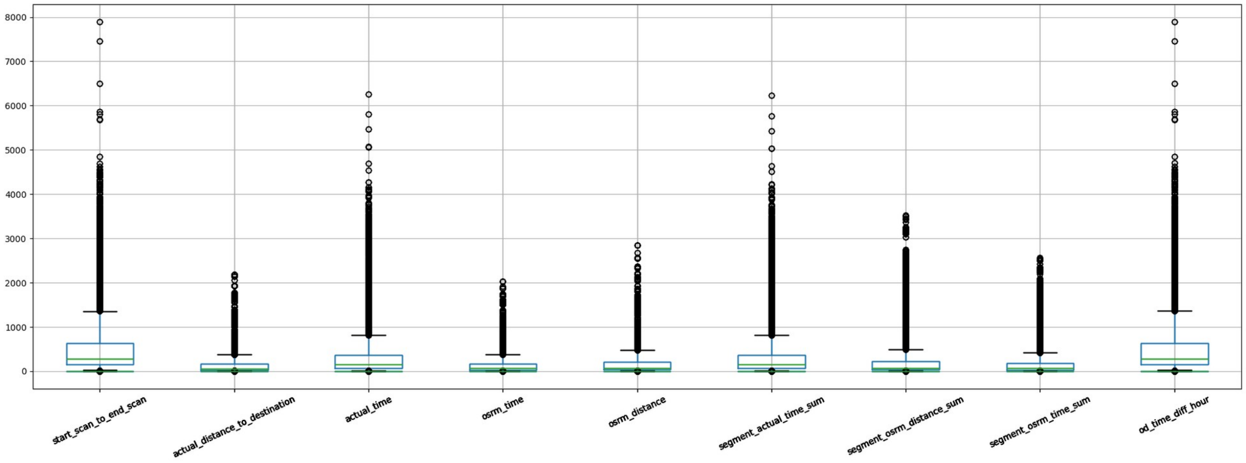
```
trip[['trip_year','trip_month','trip_hour','trip_day','trip_week','tri
p_dayofweek']]
```

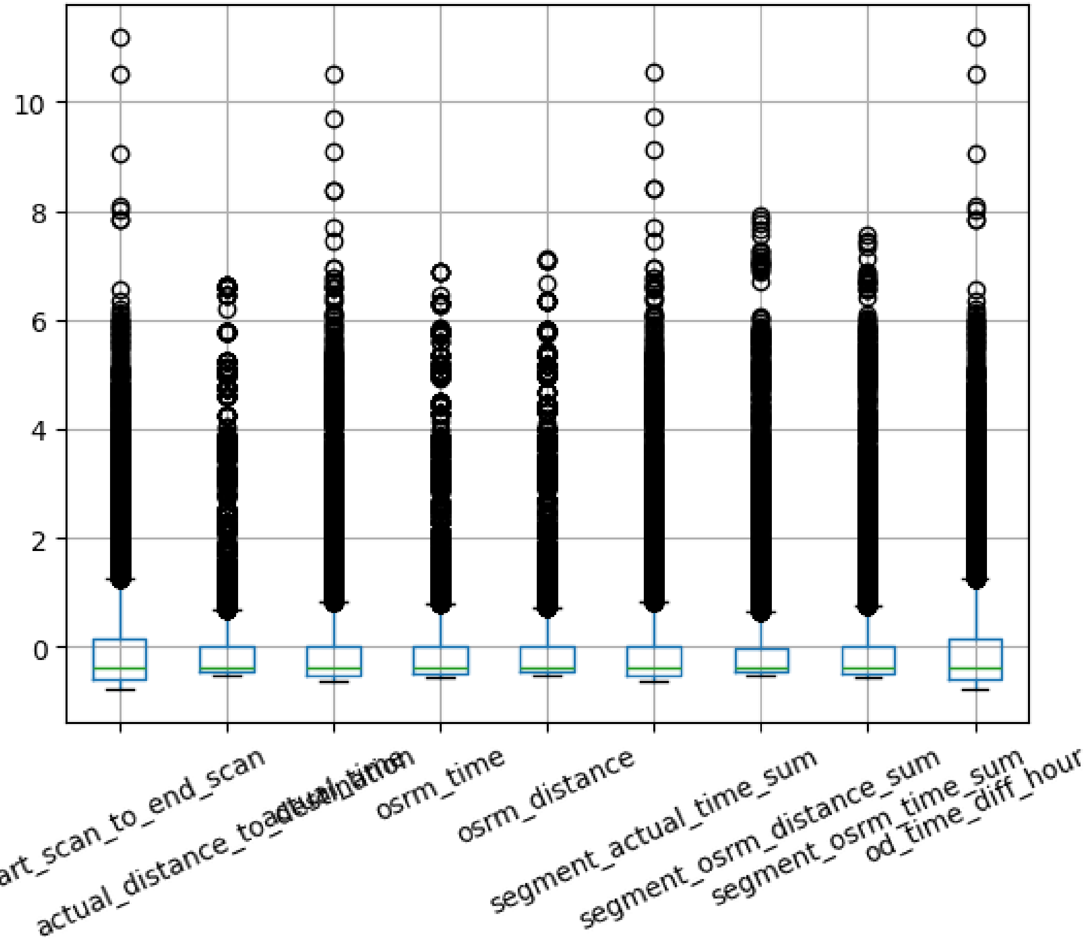
```
       trip_year  trip_month  trip_hour  trip_day  trip_week
trip_dayofweek
0           2018           9          0        12         37
2
1           2018           9          0        12         37
2
2           2018           9          0        12         37
2
3           2018           9          0        12         37
2
4           2018           9          0        12         37
2
...          ...         ...        ...       ...        ...
...
14782       2018          10         23         3         40
2
14783       2018          10         23         3         40
2
14784       2018          10         23         3         40
2
14785       2018          10         23         3         40
2
14786       2018          10         23         3         40
2

[14787 rows x 6 columns]
```

```
num_cols =
['start_scan_to_end_scan','actual_distance_to_destination','actual_tim
e','osrm_time',

'osrm_distance','segment_actual_time_sum','segment_osrm_distance_sum',
           'segment_osrm_time_sum', 'od_time_diff_hour']
```

Find outliers in numericals variable, and visualize it using visual analysis

```
trip[num_cols].boxplot(rot=25, figsize=(25,8))
plt.show()
```





```
trip['route_type'].value_counts()

route_type
Carting    8906
FTL        5881
Name: count, dtype: int64



trip['route_type'] = trip['route_type'].map({'FTL':0, 'Carting':1})
```

Normalize/Standarize the numerical features using MinMaxScaler or StandardScaler

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
scaler.fit(trip[num_cols])

trip[num_cols] = scaler.transform(trip[num_cols])


trip[num_cols]

       start_scan_to_end_scan  actual_distance_to_destination
actual_time  \
0                    2.627598                        2.162548
2.147277
1                   -0.530859                       -0.297563    -
0.379887
2                    5.170772                        5.772034
5.326268
3                   -0.652397                       -0.480911    -
0.529486
4                    0.284962                       -0.119943    -
0.027259
...                       ...                             ...
...
14782               -0.413880                       -0.348054    -
0.486744
14783               -0.713166                       -0.486350    -
0.597162
14784               -0.164728                       -0.410502    -
0.132335
14785               -0.277150                       -0.096128    -
0.164392
14786               -0.268034                       -0.320822    -
0.144802
```

```
       osrm_time  osrm_distance  segment_actual_time_sum  \
0       2.048290       2.125107                 2.147833   
1      -0.342571      -0.320538                -0.381163   
2       5.816936       5.802622                 5.311326   
3      -0.537818      -0.497115                -0.528553   
4      -0.162059      -0.154082                -0.023473   
...          ...            ...                      ...   
14782  -0.364674      -0.351972                -0.487212   
14783  -0.548870      -0.506808                -0.596856   
14784  -0.416249      -0.391263                -0.129522   
14785   0.066344      -0.088455                -0.170863   
14786  -0.342571      -0.332769                -0.142104   

       segment_osrm_distance_sum  segment_osrm_time_sum  
od_time_diff_hour  
0                       2.633597               2.629714           
2.627300  
1                      -0.332307              -0.367090          -
0.529625  
2                       5.571936               5.594737           
5.170237  
3                      -0.486596              -0.522809          -
0.652830  
4                      -0.182120              -0.208192           
0.285598  
...                          ...                    ...                
...  
14782                  -0.378690              -0.376623          -
0.413559  
14783                  -0.495684              -0.538699          -
0.713438  
14784                  -0.282653              -0.293997          -
0.164330  
14785                   0.001984               0.128670          -
0.276128  
14786                  -0.340969              -0.360734          -
0.267175  

[14787 rows x 9 columns]
```

Insights

Recommendations