Delhivery - Feature Engineering

In [ ]:

Importing Libraries

In [67]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
import matplotlib.pyplot as plt
from scipy import stats
from scipy.stats import kruskal,pearsonr,chi2_contingency
```

In [ ]:

Importing Dataset

In [68]:
```
df = pd.read_csv('delhivery_data.csv')
df.head()
```

Out[68]:

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_cer
0	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121A
1	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121A
2	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121A
3	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121A
4	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121A

5 rows × 24 columns

In [ ]:

Basic Exploratory Data Analysis(EDA)

In [69]:
```
df.shape
```

Out[69]: (144867, 24)

In [70]:
```
df.info
```

Out[70]:

```
<bound method DataFrame.info of             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
...          ...                         ...
144862  training  2018-09-20 16:24:28.436231
144863  training  2018-09-20 16:24:28.436231
144864  training  2018-09-20 16:24:28.436231
144865  training  2018-09-20 16:24:28.436231
144866  training  2018-09-20 16:24:28.436231

                                    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
...                                                   ...        ...
144862  thanos::sroute:f0569d2f-4e20-4c31-8542-67b86d5...    Carting
144863  thanos::sroute:f0569d2f-4e20-4c31-8542-67b86d5...    Carting
144864  thanos::sroute:f0569d2f-4e20-4c31-8542-67b86d5...    Carting
144865  thanos::sroute:f0569d2f-4e20-4c31-8542-67b86d5...    Carting
144866  thanos::sroute:f0569d2f-4e20-4c31-8542-67b86d5...    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)
...                         ...           ...                        ...
144862  trip-153746066843555182  IND131028AAB  Sonipat_Kundli_H (Haryana)
144863  trip-153746066843555182  IND131028AAB  Sonipat_Kundli_H (Haryana)
144864  trip-153746066843555182  IND131028AAB  Sonipat_Kundli_H (Haryana)
144865  trip-153746066843555182  IND131028AAB  Sonipat_Kundli_H (Haryana)
144866  trip-153746066843555182  IND131028AAB  Sonipat_Kundli_H (Haryana)

       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)
...                   ...                           ...
144862       IND000000ACB  Gurgaon_Bilaspur_HB (Haryana)
144863       IND000000ACB  Gurgaon_Bilaspur_HB (Haryana)
144864       IND000000ACB  Gurgaon_Bilaspur_HB (Haryana)
144865       IND000000ACB  Gurgaon_Bilaspur_HB (Haryana)
144866       IND000000ACB  Gurgaon_Bilaspur_HB (Haryana)

                    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
...                            ...  ...                         ...
144862  2018-09-20 16:24:28.436231  ...         2018-09-20 21:57:20
144863  2018-09-20 16:24:28.436231  ...         2018-09-20 21:31:18
144864  2018-09-20 16:24:28.436231  ...         2018-09-20 21:11:18
144865  2018-09-20 16:24:28.436231  ...         2018-09-20 20:53:19
144866  2018-09-20 16:24:28.436231  ...  2018-09-20 16:24:28.436231
```

```
        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
...                                ...          ...        ...           ...
144862                       45.258278         94.0       60.0       67.9280
144863                       54.092531        120.0       76.0       85.6829
144864                       66.163591        140.0       88.0       97.0933
144865                       73.680667        158.0       98.0      111.2709
144866                       70.039010        426.0       95.0       88.7319

          factor  segment_actual_time  segment_osrm_time  \
0       1.272727                 14.0               11.0
1       1.200000                 10.0                9.0
2       1.428571                 16.0                7.0
3       1.550000                 21.0               12.0
4       1.545455                  6.0                5.0
...          ...                  ...                ...
144862  1.566667                 12.0               12.0
144863  1.578947                 26.0               21.0
144864  1.590909                 20.0               34.0
144865  1.612245                 17.0               27.0
144866  4.484211                268.0                9.0

        segment_osrm_distance  segment_factor
0                     11.9653        1.272727
1                      9.7590        1.111111
2                     10.8152        2.285714
3                     13.0224        1.750000
4                      3.9153        1.200000
...                       ...             ...
144862                 8.1858        1.000000
144863                17.3725        1.238095
144864                20.7053        0.588235
144865                18.8885        0.629630
144866                 8.8088       29.777778

[144867 rows x 24 columns]>
```

In [ ]:

In [71]:
```
df.isna().sum()
```

Out[71]:
```
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
```

In [ ]:
```
```

In [72]:
```
df.dtypes
```

Out[72]:
```
data                               object
trip_creation_time                 object
route_schedule_uuid                object
route_type                         object
trip_uuid                          object
source_center                      object
source_name                        object
destination_center                 object
destination_name                   object
od_start_time                      object
od_end_time                        object
start_scan_to_end_scan            float64
is_cutoff                            bool
cutoff_factor                       int64
cutoff_timestamp                   object
actual_distance_to_destination    float64
actual_time                       float64
osrm_time                         float64
osrm_distance                     float64
factor                            float64
segment_actual_time               float64
segment_osrm_time                 float64
segment_osrm_distance             float64
segment_factor                    float64
dtype: object
```

Null values are removed.

In [ ]:
```
```

In [73]:
```
df.dropna(how='any')
df.reset_index(drop=True)
```

Out[73]:

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	sou
0	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND3
1	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND3
2	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND3
3	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND3
4	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND3
...	...	...	...	...	...	
144862	training	2018-09-20 16:24:28.436231	thanos::sroute:f0569d2f-4e20-4c31-8542-67b86d5...	Carting	trip-153746066843555182	IND1
144863	training	2018-09-20 16:24:28.436231	thanos::sroute:f0569d2f-4e20-4c31-8542-67b86d5...	Carting	trip-153746066843555182	IND1
144864	training	2018-09-20 16:24:28.436231	thanos::sroute:f0569d2f-4e20-4c31-8542-67b86d5...	Carting	trip-153746066843555182	IND1
144865	training	2018-09-20 16:24:28.436231	thanos::sroute:f0569d2f-4e20-4c31-8542-67b86d5...	Carting	trip-153746066843555182	IND1
144866	training	2018-09-20 16:24:28.436231	thanos::sroute:f0569d2f-4e20-4c31-8542-67b86d5...	Carting	trip-153746066843555182	IND1

144867 rows × 24 columns

In [ ]:

Coverting columns with 'object' data type to 'datetime' data type.

In [74]:
```
df['od_start_time']=pd.to_datetime(df['od_start_time'])
df['od_end_time']=pd.to_datetime(df['od_end_time'])
```

In [75]:
```
df.dtypes
```

Out[75]:
```
data                                    object
trip_creation_time                      object
route_schedule_uuid                     object
route_type                              object
trip_uuid                               object
source_center                           object
source_name                             object
destination_center                      object
destination_name                        object
od_start_time                   datetime64[ns]
od_end_time                     datetime64[ns]
start_scan_to_end_scan                 float64
is_cutoff                                 bool
cutoff_factor                            int64
cutoff_timestamp                        object
actual_distance_to_destination         float64
actual_time                            float64
osrm_time                              float64
osrm_distance                          float64
factor                                 float64
segment_actual_time                    float64
segment_osrm_time                      float64
segment_osrm_distance                  float64
segment_factor                         float64
dtype: object
```

In [76]:
```
df.isna().sum()
```

Out[76]:
```
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
```

In [ ]:
```

```

In [77]:
```
df.describe(include = "all")
```

Out[77]:

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	sour
count	144867	144867	144867	144867	144867	
unique	2	14817	1504	2	14817	
top	training	2018-09-28 05:23:15.359220	thanos::sroute:4029a8a2-6c74-4b7e-a6d8-f9e069f...	FTL	trip-153811219535896559	IND0
freq	104858	101	1812	99660	101	
first	NaN	NaN	NaN	NaN	NaN	
last	NaN	NaN	NaN	NaN	NaN	
mean	NaN	NaN	NaN	NaN	NaN	
std	NaN	NaN	NaN	NaN	NaN	
min	NaN	NaN	NaN	NaN	NaN	
25%	NaN	NaN	NaN	NaN	NaN	
50%	NaN	NaN	NaN	NaN	NaN	
75%	NaN	NaN	NaN	NaN	NaN	
max	NaN	NaN	NaN	NaN	NaN	

13 rows × 24 columns

In [ ]:

```
```

Grouping by sub-journey in the trip.

In [78]:

```
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()
```

In [79]:

```
df[[col + '_sum' for col in segment_cols]]
```

Out[79]:

	segment_actual_time_sum	segment_osrm_distance_sum	segment_osrm_time_sum
0	14.0	11.9653	11.0
1	24.0	21.7243	20.0
2	40.0	32.5395	27.0
3	61.0	45.5619	39.0
4	67.0	49.4772	44.0
...	...	...	...
144862	92.0	65.3487	94.0
144863	118.0	82.7212	115.0
144864	138.0	103.4265	149.0
144865	155.0	122.3150	176.0
144866	423.0	131.1238	185.0

144867 rows × 3 columns

In [ ]:

Aggregating at sub-journey level

In [80]:
```
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',

    }
```

In [ ]:

Groupby mini-trips, sorting by time.

In [81]:
```
segment = df.groupby('segment_key').agg(create_segment_dict).reset_index()
segment = segment.sort_values(by=['segment_key','od_end_time'], ascending=True).res
```

In [82]:
```
segment
```

Out[82]:

	index	segment_key	data	trip_creation_time	rout
0	0	trip-153671041653548748IND209304AAAIND000000ACB	training	2018-09-12 00:00:16.535741	thanos:
1	1	trip-153671041653548748IND462022AAAIND209304AAA	training	2018-09-12 00:00:16.535741	thanos:
2	2	trip-153671042288605164IND561203AABIND562101AAA	training	2018-09-12 00:00:22.886430	thanos::
3	3	trip-153671042288605164IND572101AAAIND561203AAB	training	2018-09-12 00:00:22.886430	thanos::
4	4	trip-153671043369099517IND000000ACBIND160002AAC	training	2018-09-12 00:00:33.691250	thanos::
...	...	...	...	...	
26363	26363	trip-153861115439069069IND628204AAAIND627657AAA	test	2018-10-03 23:59:14.390954	thanos
26364	26364	trip-153861115439069069IND628613AAAIND627005AAA	test	2018-10-03 23:59:14.390954	thanos
26365	26365	trip-153861115439069069IND628801AAAIND628204AAA	test	2018-10-03 23:59:14.390954	thanos
26366	26366	trip-153861118270144424IND583119AAAIND583101AAA	test	2018-10-03 23:59:42.701692	thanos
26367	26367	trip-153861118270144424IND583201AAAIND583119AAA	test	2018-10-03 23:59:42.701692	thanos

26368 rows × 21 columns

In [ ]:
```
```

In [83]:
```
segment[segment['trip_uuid'] == 'trip-153741093647649320']
```

Out[83]:

	index	segment_key	data	trip_creation_time	rout
10374	10374	trip-153741093647649320IND388121AAAIND388620AAB	training	2018-09-20 02:35:36.476840	thanos::
10375	10375	trip-153741093647649320IND388620AABIND388320AAA	training	2018-09-20 02:35:36.476840	thanos::

2 rows × 21 columns

In [ ]:

In [84]:
```
segment.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26368 entries, 0 to 26367
Data columns (total 21 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   index                           26368 non-null  int64
 1   segment_key                     26368 non-null  object
 2   data                            26368 non-null  object
 3   trip_creation_time              26368 non-null  object
 4   route_schedule_uuid             26368 non-null  object
 5   route_type                      26368 non-null  object
 6   trip_uuid                       26368 non-null  object
 7   source_center                   26368 non-null  object
 8   source_name                     26302 non-null  object
 9   destination_center              26368 non-null  object
 10  destination_name                26287 non-null  object
 11  od_start_time                   26368 non-null  datetime64[ns]
 12  od_end_time                     26368 non-null  datetime64[ns]
 13  start_scan_to_end_scan          26368 non-null  float64
 14  actual_distance_to_destination  26368 non-null  float64
 15  actual_time                     26368 non-null  float64
 16  osrm_time                       26368 non-null  float64
 17  osrm_distance                   26368 non-null  float64
 18  segment_actual_time_sum         26368 non-null  float64
 19  segment_osrm_distance_sum       26368 non-null  float64
 20  segment_osrm_time_sum           26368 non-null  float64
dtypes: datetime64[ns](2), float64(8), int64(1), object(10)
memory usage: 4.2+ MB
```

In [ ]:

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

od_time_diff_hour is matching with start_scan_to_end_scan

In [85]:
```
segment['od_time_diff_hour'] = (segment['od_end_time'] - segment['od_start_time']).
segment['od_time_diff_hour']
```

Out[85]:
```
0        1260.604421
1         999.505379
2          58.832388
3         122.779486
4         834.638929
             ...
26363      62.115193
26364      91.087797
26365      44.174403
26366     287.474007
26367      66.933565
Name: od_time_diff_hour, Length: 26368, dtype: float64
```

In [40]:
```
segment
```

Out[40]:

	index	segment_key	data	trip_creation_time	rou
0	0	trip-153671041653548748IND209304AAAIND000000ACB	training	2018-09-12 00:00:16.535741	thanos:
1	1	trip-153671041653548748IND462022AAAIND209304AAA	training	2018-09-12 00:00:16.535741	thanos:
2	2	trip-153671042288605164IND561203AABIND562101AAA	training	2018-09-12 00:00:22.886430	thanos::
3	3	trip-153671042288605164IND572101AAAIND561203AAB	training	2018-09-12 00:00:22.886430	thanos::
4	4	trip-153671043369099517IND000000ACBIND160002AAC	training	2018-09-12 00:00:33.691250	thanos::
...	...	...	...	...	
26363	26363	trip-153861115439069069IND628204AAAIND627657AAA	test	2018-10-03 23:59:14.390954	thanos
26364	26364	trip-153861115439069069IND628613AAAIND627005AAA	test	2018-10-03 23:59:14.390954	thanos
26365	26365	trip-153861115439069069IND628801AAAIND628204AAA	test	2018-10-03 23:59:14.390954	thanos
26366	26366	trip-153861118270144424IND583119AAAIND583101AAA	test	2018-10-03 23:59:42.701692	thanos
26367	26367	trip-153861118270144424IND583201AAAIND583119AAA	test	2018-10-03 23:59:42.701692	thanos

26368 rows × 22 columns

In [ ]:
```
```

In [86]:
```
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',

    }
```

In [87]:
```
trip = segment.groupby('trip_uuid').agg(create_trip_dict).reset_index(drop = True)
```

In [88]:
```
trip
```

Out[88]:

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	sour
0	training	2018-09-12 00:00:16.535741	thanos::sroute:d7c989ba-a29b-4a0b-b2f4-288cdc6...	FTL	trip-153671041653548748	IND2(
1	training	2018-09-12 00:00:22.886430	thanos::sroute:3a1b0ab2-bb0b-4c53-8c59-eb2a2c0...	Carting	trip-153671042288605164	IND56
2	training	2018-09-12 00:00:33.691250	thanos::sroute:de5e208e-7641-45e6-8100-4d9fb1e...	FTL	trip-153671043369099517	IND0(
3	training	2018-09-12 00:01:00.113710	thanos::sroute:f0176492-a679-4597-8332-bbd1c7f...	Carting	trip-153671046011330457	IND4(
4	training	2018-09-12 00:02:09.740725	thanos::sroute:d9f07b12-65e0-4f3b-bec8-df06134...	FTL	trip-153671052974046625	IND58
...	...	...	...	...	...	
14812	test	2018-10-03 23:55:56.258533	thanos::sroute:8a120994-f577-4491-9e4b-b7e4a14...	Carting	trip-153861095625827784	IND16
14813	test	2018-10-03 23:57:23.863155	thanos::sroute:b30e1ec3-3bfa-4bd2-a7fb-3b75769...	Carting	trip-153861104386292051	IND12
14814	test	2018-10-03 23:57:44.429324	thanos::sroute:5609c268-e436-4e0a-8180-3db4a74...	Carting	trip-153861106442901555	IND2(
14815	test	2018-10-03 23:59:14.390954	thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...	Carting	trip-153861115439069069	IND62
14816	test	2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	FTL	trip-153861118270144424	IND58

14817 rows × 18 columns

In [ ]:

In [89]:
```
trip[['actual_time', 'segment_actual_time_sum']]
```

Out[89]:

	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
...	...	...
14812	83.0	82.0
14813	21.0	21.0
14814	282.0	281.0
14815	264.0	258.0
14816	275.0	274.0

14817 rows × 2 columns

In [90]:
```
trip
```

Out[90]:

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	sour
0	training	2018-09-12 00:00:16.535741	thanos::sroute:d7c989ba-a29b-4a0b-b2f4-288cdc6...	FTL	trip-153671041653548748	IND20
1	training	2018-09-12 00:00:22.886430	thanos::sroute:3a1b0ab2-bb0b-4c53-8c59-eb2a2c0...	Carting	trip-153671042288605164	IND56
2	training	2018-09-12 00:00:33.691250	thanos::sroute:de5e208e-7641-45e6-8100-4d9fb1e...	FTL	trip-153671043369099517	IND00
3	training	2018-09-12 00:01:00.113710	thanos::sroute:f0176492-a679-4597-8332-bbd1c7f...	Carting	trip-153671046011330457	IND40
4	training	2018-09-12 00:02:09.740725	thanos::sroute:d9f07b12-65e0-4f3b-bec8-df06134...	FTL	trip-153671052974046625	IND58
...	...	...	...	...	...	
14812	test	2018-10-03 23:55:56.258533	thanos::sroute:8a120994-f577-4491-9e4b-b7e4a14...	Carting	trip-153861095625827784	IND16
14813	test	2018-10-03 23:57:23.863155	thanos::sroute:b30e1ec3-3bfa-4bd2-a7fb-3b75769...	Carting	trip-153861104386292051	IND12
14814	test	2018-10-03 23:57:44.429324	thanos::sroute:5609c268-e436-4e0a-8180-3db4a74...	Carting	trip-153861106442901555	IND20
14815	test	2018-10-03 23:59:14.390954	thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...	Carting	trip-153861115439069069	IND62
14816	test	2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	FTL	trip-153861118270144424	IND58

14817 rows × 18 columns

In [ ]:

In [93]:
```
trip[trip['trip_uuid'] == 'trip-153741093647649320']
```

Out[93]:

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source
5919	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388

In [ ]:

In [94]:
```
trip[['actual_distance_to_destination','osrm_distance']]
```

Out[94]:

	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
...	...	...
14812	57.762332	73.4630
14813	15.513784	16.0882
14814	38.684839	58.9037
14815	134.723836	171.1103
14816	66.081533	80.5787

14817 rows × 2 columns

In [ ]:

Hypothesis testing/ visual analysis between actual_time aggregated value and OSRM

Does segment_actual_time is similer as segment_osrm_time?

In [49]:
```
from scipy.stats import ttest_ind
null_hypothesis = 'mean of actual_time is not higher than mean of osrm_time'
alternative_hypothesis = 'mean of actual_time is higher than mean of osrm_time'
sample1 = trip['actual_time']
sample2 = trip['osrm_time']
t_stat, p_value = ttest_ind(sample1, sample2, equal_var=False, alternative='greater
print(t_stat, p_value)

if(p_value < 0.05):
 print('Since, p-value < 0.05, the null hypothesis is rejected')
 print(alternative_hypothesis)
else:
 print('Since p-value > 0.05, we fail to reject null hypothesis')
 print(null_hypothesis)
```

```
38.21545390583316 1.85234938418568e-309
Since, p-value < 0.05, the null hypothesis is rejected
mean of actual_time is higher than mean of osrm_time
```

In [ ]:

Hypothesis testing/ visual analysis between actual_time aggregated value and segment_osrm_time

Does actual_time is similer as segment_osrm_time?

In [50]:
```
from scipy.stats import ttest_ind
null_hypothesis = 'mean of actual_time is similer as segment_actual_time'
alternative_hypothesis = 'mean of actual_time is different than mean of segment_osr
sample1 = trip['actual_time']
sample2 = trip['segment_actual_time_sum']
t_stat, p_value = ttest_ind(sample1, sample2)
print(t_stat, p_value)
if(p_value < 0.05):
 print('Since, p-value < 0.05, the null hypothesis is rejected')
 print(alternative_hypothesis)
else:
 print('Since p-value > 0.05, we fail to reject null hypothesis')
 print(null_hypothesis)
```

```
0.5008024728897531 0.6165138648224772
Since p-value > 0.05, we fail to reject null hypothesis
mean of actual_time is similer as segment_actual_time
```

In [ ]:

Does osrm_distance is similer as segment_osrm_distance_sum

In [51]:
```
from scipy.stats import ttest_ind
null_hypothesis = 'mean of osrm_distance is similer as mean of segment_osrm_distanc
alternative_hypothesis = 'mean of osrm_distance is higher than mean of segment_osrm
sample1 = trip['osrm_distance']
sample2 = trip['segment_osrm_distance_sum']
t_stat, p_value = ttest_ind(sample1, sample2, equal_var=False, alternative='greater
print(t_stat, p_value)
if(p_value < 0.05):
 print('Since, p-value < 0.05, the null hypothesis is rejected')
 print(alternative_hypothesis)
else:
 print('Since p-value > 0.05, we fail to reject null hypothesis')
 print(null_hypothesis)
```

```
-4.117367046483823 0.9999807861306765
Since p-value > 0.05, we fail to reject null hypothesis
mean of osrm_distance is similer as mean of segment_osrm_distance
```

In [ ]:

In [52]:
```
num_cols = ['actual_time','osrm_time','segment_actual_time_sum','segment_osrm_time_
            'actual_distance_to_destination','osrm_distance','segment_osrm_distance
for i in (num_cols):
 stat, p_value = stats.shapiro(sample1)
 if(p_value < 0.05):
     print(i, ": sample is not normally distributed, do non parametric test")
 else:
     print(i, ": sample is normally distributed, can do parametric test")
```

```
actual_time : sample is not normally distributed, do non parametric test
osrm_time : sample is not normally distributed, do non parametric test
segment_actual_time_sum : sample is not normally distributed, do non parametric te
st
segment_osrm_time_sum : sample is not normally distributed, do non parametric test
actual_distance_to_destination : sample is not normally distributed, do non parame
tric test
osrm_distance : sample is not normally distributed, do non parametric test
segment_osrm_distance_sum : sample is not normally distributed, do non parametric
test
```

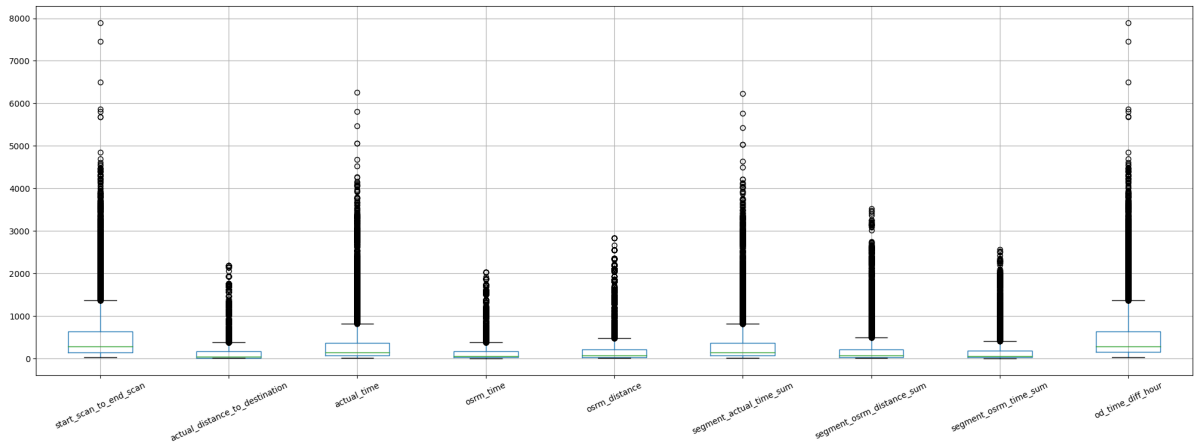
```
C:\Users\Mateen\anaconda3\anoconda\Lib\site-packages\scipy\stats\_morestats.py:1816: UserWarning: p-value may not be accurate for N > 5000.
  warnings.warn("p-value may not be accurate for N > 5000.")
```

In [ ]:

Find outliers in numerical variable (you might find outliers in almost all the variables), and visualize it using visual analysis

In [53]:
```
num_cols = ['start_scan_to_end_scan', 'actual_distance_to_destination', 'actual_tim
            'osrm_distance', 'segment_actual_time_sum', 'segment_osrm_distance_sum'
            'segment_osrm_time_sum', 'od_time_diff_hour']
```

In [55]:
```
trip[num_cols].boxplot(rot=25, figsize=(25,8))
plt.show()
```



In [ ]:

Handle the outliers using IQR method.

In [56]:
```
Q1 = trip[num_cols].quantile(0.25)
Q3 = trip[num_cols].quantile(0.75)

IQR = Q3 - Q1
```

In [57]:
```
trip = trip[~((trip[num_cols] < (Q1 - 1.5 * IQR)) | (trip[num_cols] > (Q3 + 1.5 * I
trip = trip.reset_index(drop=True)
```

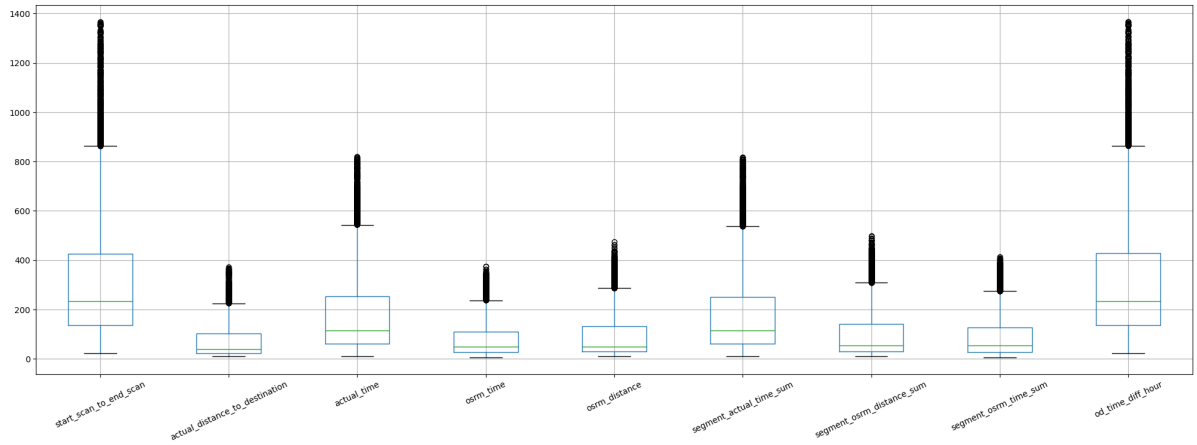
In [58]:
```
trip
```

Out[58]:

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	sour
0	training	2018-09-12 00:00:22.886430	thanos::sroute:3a1b0ab2-bb0b-4c53-8c59-eb2a2c0...	Carting	trip-153671042288605164	IND5
1	training	2018-09-12 00:01:00.113710	thanos::sroute:f0176492-a679-4597-8332-bbd1c7f...	Carting	trip-153671046011330457	IND4
2	training	2018-09-12 00:02:09.740725	thanos::sroute:d9f07b12-65e0-4f3b-bec8-df06134...	FTL	trip-153671052974046625	IND5
3	training	2018-09-12 00:02:34.161600	thanos::sroute:9bf03170-d0a2-4a3f-aa4d-9aaab3d...	Carting	trip-153671055416136166	IND6
4	training	2018-09-12 00:04:22.011653	thanos::sroute:a97698cc-846e-41a7-916b-88b1741...	Carting	trip-153671066201138152	IND6
...	...	...	...	...	...	
12754	test	2018-10-03 23:55:56.258533	thanos::sroute:8a120994-f577-4491-9e4b-b7e4a14...	Carting	trip-153861095625827784	IND1
12755	test	2018-10-03 23:57:23.863155	thanos::sroute:b30e1ec3-3bfa-4bd2-a7fb-3b75769...	Carting	trip-153861104386292051	IND1
12756	test	2018-10-03 23:57:44.429324	thanos::sroute:5609c268-e436-4e0a-8180-3db4a74...	Carting	trip-153861106442901555	IND2
12757	test	2018-10-03 23:59:14.390954	thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...	Carting	trip-153861115439069069	IND6
12758	test	2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	FTL	trip-153861118270144424	IND5

12759 rows × 18 columns

In [59]:
```
trip[num_cols].boxplot(rot=25, figsize=(25,8))
plt.show()
```



Handling categorical variables.

Only two route_type – Do one hot encoding

In [61]:
```
trip['route_type'].value_counts()
```

Out[61]:
```
Carting    8817
FTL        3942
Name: route_type, dtype: int64
```

In [62]:
```
trip['route_type'] = trip['route_type'].map({'FTL':0, 'Carting':1})
```

In [ ]:
```

```

Normalize/ Standardize the numerical features using MinMaxScaler or StandardScaler

In [63]:
```
from sklearn.preprocessing import StandardScaler
```

In [64]:
```
scaler = StandardScaler()
scaler.fit(trip[num_cols])
```

Out[64]:
```
StandardScaler()
```

In [65]:
```
trip[num_cols] = scaler.transform(trip[num_cols])
```

In [66]:
```
trip[num_cols]
```

Out[66]:

	start_scan_to_end_scan	actual_distance_to_destination	actual_time	osrm_time	osrm_distanc
0	-0.551781	0.004976	-0.223508	-0.150681	-0.08060
1	-0.862589	-0.766880	-0.751536	-0.878175	-0.80610
2	1.534514	0.752716	1.021129	0.521909	0.60331
3	-0.516816	-0.664606	-0.738964	-0.768365	-0.71313
4	-0.870359	-0.878152	-0.971547	-0.905628	-0.89105
...	...	...	...	...	
12754	-0.252629	-0.207579	-0.600671	-0.233038	-0.20975
12755	-1.017993	-0.789776	-0.990406	-0.919354	-0.84593
12756	0.384526	-0.470472	0.650252	-0.425207	-0.37119
12757	0.097029	0.852973	0.537103	1.372940	0.87296
12758	0.120340	-0.092938	0.606250	-0.150681	-0.13085

12759 rows × 9 columns

In [ ]:
```

```

Insights :

Mean of actual time is different from mean of segment osrm time. Mean of osrm distance is similar to mean of segment osrm distance. Mean of actual time is higher than mean of segment osrm time. Carting Transportation : 69% FTL Transportation : 31% Most orders are coming and going to same state ie. Maharashtra. Most orders are going to Mumbai. Most orders are coming from Bhindwandi_Mankoli_HB . Trip between Angamaly to Chalakudy saw the least avg time for completion. Trip between Hyderabad to Shamshabad saw the highest avg time for completion. The busiest route is between Bhindwandi_Mankoli_HB to Mumbai.

Recommendations :

Since most orders are coming and going to Maharashtra , company have to expand the strategy used in Maharashtra to other states. Since the busiest route is between Bhindwandi_Mankoli_HB to Mumbai ,company can use more transportation in this route. Company have to analyse the transportation setup between Angamaly and Chalakudy which is the fastest route and setup the same producere to all other routes. Company have to analyse Hyderabad to Shamshabad route and make necessary changes to make the route fast. Proper route type for each route should be implemented.

In [ ]:

In [ ]:

In [ ]:

In [ ]: