Basic data cleaning and exploration:

In [ ]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Set the maximum number of columns and rows to be displayed
pd.options.display.max_columns = 50     # Set the maximum number of columns
pd.options.display.max_rows = 1000      # Set the maximum number of rows
```

In [ ]:
```
df = pd.read_csv("dataset/data.csv")
```

In [ ]:
```
df.head(n=7)
```

Out[ ]:

	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
0	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	IND388620AAB	Khambhat_MotvdDPP_D (Gujarat)	2018-09-20 03:21:32.418600	2018-09-20 04:47:45.236797	86.0	True	9	2018-09-20 04:27:55	10.435660	14.0	11.0	11.9653	1.272727	14.0	11.0	11.9653	1.272727
1	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	IND388620AAB	Khambhat_MotvdDPP_D (Gujarat)	2018-09-20 03:21:32.418600	2018-09-20 04:47:45.236797	86.0	True	18	2018-09-20 04:17:55	18.936842	24.0	20.0	21.7243	1.200000	10.0	9.0	9.7590	1.111111
2	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	IND388620AAB	Khambhat_MotvdDPP_D (Gujarat)	2018-09-20 03:21:32.418600	2018-09-20 04:47:45.236797	86.0	True	27	2018-09-20 04:01:19.505586	27.637279	40.0	28.0	32.5395	1.428571	16.0	7.0	10.8152	2.285714
3	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	IND388620AAB	Khambhat_MotvdDPP_D (Gujarat)	2018-09-20 03:21:32.418600	2018-09-20 04:47:45.236797	86.0	True	36	2018-09-20 03:39:57	36.118028	62.0	40.0	45.5620	1.550000	21.0	12.0	13.0224	1.750000
4	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	IND388620AAB	Khambhat_MotvdDPP_D (Gujarat)	2018-09-20 03:21:32.418600	2018-09-20 04:47:45.236797	86.0	False	39	2018-09-20 03:33:55	39.386040	68.0	44.0	54.2181	1.545455	6.0	5.0	3.9153	1.200000
5	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388620AAB	Khambhat_MotvdDPP_D (Gujarat)	IND388320AAA	Anand_Vaghasi_IP (Gujarat)	2018-09-20 04:47:45.236797	2018-09-20 06:36:55.627764	109.0	True	9	2018-09-20 06:15:58	10.403038	15.0	11.0	12.1171	1.363636	15.0	11.0	12.1171	1.363636
6	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388620AAB	Khambhat_MotvdDPP_D (Gujarat)	IND388320AAA	Anand_Vaghasi_IP (Gujarat)	2018-09-20 04:47:45.236797	2018-09-20 06:36:55.627764	109.0	True	18	2018-09-20 05:47:29	18.045481	44.0	17.0	21.2890	2.588235	28.0	6.0	9.1719	4.666667

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

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

In [ ]:
```
df.drop(["is_cutoff", "cutoff_factor", "cutoff_timestamp", "factor", "segment_factor"], axis=1, inplace=True)
```

In [ ]:
```
df["destination_name"].fillna("missing_"+df["destination_center"], inplace=True)
df["source_name"].fillna("missing_"+df["source_center"], inplace=True)
```

In [ ]:
```
df.head(5)
```

Out[ ]:

	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	actual_distance_to_destination	actual_time	osrm_time	osrm_distance	segment_actual_time	segment_osrm_time	segment_osrm_distance
0	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	IND388620AAB	Khambhat_MotvdDPP_D (Gujarat)	2018-09-20 03:21:32.418600	2018-09-20 04:47:45.236797	86.0	10.435660	14.0	11.0	11.9653	14.0	11.0	11.9653
1	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	IND388620AAB	Khambhat_MotvdDPP_D (Gujarat)	2018-09-20 03:21:32.418600	2018-09-20 04:47:45.236797	86.0	18.936842	24.0	20.0	21.7243	10.0	9.0	9.7590
2	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	IND388620AAB	Khambhat_MotvdDPP_D (Gujarat)	2018-09-20 03:21:32.418600	2018-09-20 04:47:45.236797	86.0	27.637279	40.0	28.0	32.5395	16.0	7.0	10.8152
3	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	IND388620AAB	Khambhat_MotvdDPP_D (Gujarat)	2018-09-20 03:21:32.418600	2018-09-20 04:47:45.236797	86.0	36.118028	62.0	40.0	45.5620	21.0	12.0	13.0224
4	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	IND388620AAB	Khambhat_MotvdDPP_D (Gujarat)	2018-09-20 03:21:32.418600	2018-09-20 04:47:45.236797	86.0	39.386040	68.0	44.0	54.2181	6.0	5.0	3.9153

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 144867 entries, 0 to 144866
Data columns (total 19 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                    144867 non-null  object 
 7   destination_center             144867 non-null  object 
 8   destination_name               144867 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  actual_distance_to_destination 144867 non-null  float64
 13  actual_time                    144867 non-null  float64
 14  osrm_time                      144867 non-null  float64
 15  osrm_distance                  144867 non-null  float64
 16  segment_actual_time            144867 non-null  float64
 17  segment_osrm_time              144867 non-null  float64
 18  segment_osrm_distance          144867 non-null  float64
dtypes: float64(8), object(11)
memory usage: 21.0+ MB
```

In [ ]:
```
## We've to update object type to datetime for datetime columns: trip_creation_time, od_start_time, od_end_time, cutoff_timestamp

# List of columns to convert to datetime
data_columns = ["trip_creation_time", "od_start_time", "od_end_time"]
df[data_columns] = df[data_columns].apply(pd.to_datetime, format='mixed')
```

In [ ]:
```
df.describe(include=["float", "int"])
```

Out[ ]:

	start_scan_to_end_scan	actual_distance_to_destination	actual_time	osrm_time	osrm_distance	segment_actual_time	segment_osrm_time	segment_osrm_distance
count	144867.000000	144867.000000	144867.000000	144867.000000	144867.000000	144867.000000	144867.000000	144867.00000
mean	961.262986	234.073372	416.927527	213.868272	284.771297	36.196111	18.507548	22.82902
std	1037.012769	344.990009	598.103621	308.011085	421.119294	53.571158	14.775960	17.86066
min	20.000000	9.000045	9.000000	6.000000	9.008200	-244.000000	0.000000	0.00000
25%	161.000000	23.355874	51.000000	27.000000	29.914700	20.000000	11.000000	12.07010
50%	449.000000	66.126571	132.000000	64.000000	78.525800	29.000000	17.000000	23.51300
75%	1634.000000	286.708875	513.000000	257.000000	343.193250	40.000000	22.000000	27.81325
max	7898.000000	1927.447705	4532.000000	1686.000000	2326.199100	3051.000000	1611.000000	2191.40370

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

Out[ ]:

	data	route_schedule_uuid	route_type	trip_uuid	source_center	source_name	destination_center	destination_name
count	144867	144867	144867	144867	144867	144867	144867	144867
unique	2	1504	2	14817	1508	1508	1481	1481
top	training	thanos::sroute:4029a8a2-6c74-4b7e-a6d8-f9e069f...	FTL	trip-153784927255069118	IND000000ACB	Gurgaon_Bilaspur_HB (Haryana)	IND000000ACB	Gurgaon_Bilaspur_HB (Haryana)
freq	104858	1812	99660	101	23347	23347	15192	15192

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

Out[ ]:

	trip_creation_time	od_start_time	od_end_time
count	144867	144867	144867
mean	2018-09-22 13:34:23.659819264	2018-09-22 18:02:45.855230720	2018-09-23 10:04:31.395393024
min	2018-09-12 00:00:16.535741	2018-09-12 00:00:16.535741	2018-09-12 00:50:10.814399
25%	2018-09-17 03:20:51.775845888	2018-09-17 08:05:40.886155008	2018-09-18 01:48:06.410121984
50%	2018-09-22 04:24:27.932764928	2018-09-22 08:53:00.116656128	2018-09-23 03:13:03.520212992
75%	2018-09-27 17:57:56.350054912	2018-09-27 22:41:50.285857024	2018-09-28 12:49:06.054018048
max	2018-10-03 23:59:42.701692	2018-10-06 04:27:23.392375	2018-10-08 03:00:24.353479

In [ ]:
```
df.head(n=7)
```

Out[ ]:

	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	actual_distance_to_destination	actual_time	osrm_time	osrm_distance	segment_actual_time	segment_osrm_time	segment_osrm_distance
0	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	IND388620AAB	Khambhat_MotvdDPP_D (Gujarat)	2018-09-20 03:21:32.418600	2018-09-20 04:47:45.236797	86.0	10.435660	14.0	11.0	11.9653	14.0	11.0	11.9653
1	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	IND388620AAB	Khambhat_MotvdDPP_D (Gujarat)	2018-09-20 03:21:32.418600	2018-09-20 04:47:45.236797	86.0	18.936842	24.0	20.0	21.7243	10.0	9.0	9.7590
2	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	IND388620AAB	Khambhat_MotvdDPP_D (Gujarat)	2018-09-20 03:21:32.418600	2018-09-20 04:47:45.236797	86.0	27.637279	40.0	28.0	32.5395	16.0	7.0	10.8152
3	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	IND388620AAB	Khambhat_MotvdDPP_D (Gujarat)	2018-09-20 03:21:32.418600	2018-09-20 04:47:45.236797	86.0	36.118028	62.0	40.0	45.5620	21.0	12.0	13.0224
4	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	IND388620AAB	Khambhat_MotvdDPP_D (Gujarat)	2018-09-20 03:21:32.418600	2018-09-20 04:47:45.236797	86.0	39.386040	68.0	44.0	54.2181	6.0	5.0	3.9153
5	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388620AAB	Khambhat_MotvdDPP_D (Gujarat)	IND388320AAA	Anand_Vaghasi_IP (Gujarat)	2018-09-20 04:47:45.236797	2018-09-20 06:36:55.627764	109.0	10.403038	15.0	11.0	12.1171	15.0	11.0	12.1171
6	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388620AAB	Khambhat_MotvdDPP_D (Gujarat)	IND388320AAA	Anand_Vaghasi_IP (Gujarat)	2018-09-20 04:47:45.236797	2018-09-20 06:36:55.627764	109.0	18.045481	44.0	17.0	21.2890	28.0	6.0	9.1719

In [ ]:
```

```

Try merging the rows

In [ ]:
```
## Grouping by segment

## unique identifier for different segments of a trip
df["segment_key"]=df["trip_uuid"]+df["source_center"]+df["destination_center"]

## Grouping by above created segments
for columns in ["segment_actual_time", "segment_osrm_distance", "segment_osrm_time"]:
    df[columns+"_sum"] = df.groupby("segment_key")[columns].cumsum()

df[[columns+"_sum" for columns in ["segment_actual_time", "segment_osrm_distance", "segment_osrm_time"]]]
```

Out[ ]:

	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 [ ]:
```
df.columns
```

Out[ ]:
```
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',
       'start_scan_to_end_scan', 'actual_distance_to_destination',
       'actual_time', 'osrm_time', 'osrm_distance', 'segment_actual_time',
       'segment_osrm_time', 'segment_osrm_distance', 'segment_key',
       'segment_actual_time_sum', 'segment_osrm_distance_sum',
       'segment_osrm_time_sum'],
      dtype='object')
```

In [ ]:
```
## Aggregating at segment level

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 [ ]:
```
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()
```

In [ ]:
```
segment
```

Out[ ]:

	index	segment_key	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	actual_distance_to_destination	actual_time	osrm_time	osrm_distance	segment_actual_time_sum	segment_osrm_distance_sum	segment_osrm_time_su
0	0	trip-153671041653548748IND209304AAAIND000000ACB	training	2018-09-12 00:00:16.535741	thanos::sroute:d7c989ba-a29b-4a0b-b2f4-288cdc6...	FTL	trip-153671041653548748	IND209304AAA	Kanpur_Central_H_6 (Uttar Pradesh)	IND000000ACB	Gurgaon_Bilaspur_HB (Haryana)	2018-09-12 16:39:46.858469	2018-09-13 13:40:23.123744	1260.0	383.759164	732.0	329.0	446.5496	728.0	670.6205	534
1	1	trip-153671041653548748IND462022AAAIND209304AAA	training	2018-09-12 00:00:16.535741	thanos::sroute:d7c989ba-a29b-4a0b-b2f4-288cdc6...	FTL	trip-153671041653548748	IND462022AAA	Bhopal_Trnsport_H (Madhya Pradesh)	IND209304AAA	Kanpur_Central_H_6 (Uttar Pradesh)	2018-09-12 00:00:16.535741	2018-09-12 16:39:46.858469	999.0	440.973689	830.0	388.0	544.8027	820.0	649.8528	474
2	2	trip-153671042288605164IND561203AABIND562101AAA	training	2018-09-12 00:00:22.886430	thanos::sroute:3a1b0ab2-bb0b-4c53-8c59-eb2a2c0...	Carting	trip-153671042288605164	IND561203AAB	Doddablpur_ChikaDPP_D (Karnataka)	IND562101AAA	Chikblapur_ShntiSgr_D (Karnataka)	2018-09-12 02:03:09.655591	2018-09-12 03:01:59.598855	58.0	24.644021	47.0	26.0	28.1994	46.0	28.1995	26
3	3	trip-153671042288605164IND572101AAAIND561203AAB	training	2018-09-12 00:00:22.886430	thanos::sroute:3a1b0ab2-bb0b-4c53-8c59-eb2a2c0...	Carting	trip-153671042288605164	IND572101AAA	Tumkur_Veersagr_I (Karnataka)	IND561203AAB	Doddablpur_ChikaDPP_D (Karnataka)	2018-09-12 00:00:22.886430	2018-09-12 02:03:09.655591	122.0	48.542890	96.0	42.0	56.9116	95.0	55.9899	39
4	4	trip-153671043369099517IND000000ACBIND160002AAC	training	2018-09-12 00:00:33.691250	thanos::sroute:de5e208e-7641-45e6-8100-4d9fb1e...	FTL	trip-153671043369099517	IND000000ACB	Gurgaon_Bilaspur_HB (Haryana)	IND160002AAC	Chandigarh_Mehmdpur_H (Punjab)	2018-09-14 03:40:17.106733	2018-09-14 17:34:55.442454	834.0	237.439610	611.0	212.0	281.2109	608.0	317.7408	231
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
26363	26363	trip-153861115439069069IND628204AAAIND627657AAA	test	2018-10-03 23:59:14.390954	thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...	Carting	trip-153861115439069069	IND628204AAA	Tirchchndr_Shnmgprm_D (Tamil Nadu)	IND627657AAA	Thisayanvilai_UdnkdiRD_D (Tamil Nadu)	2018-10-04 02:29:04.272194	2018-10-04 03:31:11.183797	62.0	33.627182	51.0	41.0	42.5213	49.0	42.1431	42
26364	26364	trip-153861115439069069IND628613AAAIND627005AAA	test	2018-10-03 23:59:14.390954	thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...	Carting	trip-153861115439069069	IND628613AAA	Peikulam_SriVnktpm_D (Tamil Nadu)	IND627005AAA	Tirunelveli_VdkkuSrt_I (Tamil Nadu)	2018-10-04 04:16:39.894872	2018-10-04 05:47:45.162682	91.0	33.673835	90.0	48.0	40.6080	89.0	78.5869	77
26365	26365	trip-153861115439069069IND628801AAAIND628204AAA	test	2018-10-03 23:59:14.390954	thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...	Carting	trip-153861115439069069	IND628801AAA	Eral_Busstand_D (Tamil Nadu)	IND628204AAA	Tirchchndr_Shnmgprm_D (Tamil Nadu)	2018-10-04 01:44:53.808000	2018-10-04 02:29:04.272194	44.0	12.661945	30.0	14.0	16.0185	29.0	16.0184	14
26366	26366	trip-153861118270144424IND583119AAAIND583101AAA	test	2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	FTL	trip-153861118270144424	IND583119AAA	Sandur_WrdN1DPP_D (Karnataka)	IND583101AAA	Bellary_Dc (Karnataka)	2018-10-04 03:58:40.726547	2018-10-04 08:46:09.166940	287.0	40.546740	233.0	42.0	52.5303	233.0	52.5303	42
26367	26367	trip-153861118270144424IND583201AAAIND583119AAA	test	2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	FTL	trip-153861118270144424	IND583201AAA	Hospet (Karnataka)	IND583119AAA	Sandur_WrdN1DPP_D (Karnataka)	2018-10-04 02:51:44.712656	2018-10-04 03:58:40.726547	66.0	25.534793	42.0	26.0	28.0484	41.0	28.0484	25

26368 rows × 21 columns

In [ ]:
```
segment.iloc[203,:]
```

Out[ ]:
```
index                                                    203
segment_key                trip-153671844300881686IND606105AAAIND606201AAC
data                                                 training
trip_creation_time                    2018-09-12 02:14:03.009204
route_schedule_uuid      thanos::sroute:872d9762-5527-44ee-9fc9-4457999...
route_type                                           Carting
trip_uuid                                 trip-153671844300881686
source_center                                       IND606105AAA
source_name                        Pennadam_EastmnRD_D (Tamil Nadu)
destination_center                                  IND606201AAC
destination_name                 Chinnasalem_VkotRoad_D (Tamil Nadu)
od_start_time                         2018-09-12 06:26:35.800461
od_end_time                           2018-09-12 08:09:20.046912
start_scan_to_end_scan                                    102.0
actual_distance_to_destination                            52.034197
actual_time                                                93.0
osrm_time                                                  54.0
osrm_distance                                             73.1099
segment_actual_time_sum                                    90.0
segment_osrm_distance_sum                                 63.9399
segment_osrm_time_sum                                      67.0
Name: 203, dtype: object
```

In [ ]:
```
segment.describe(include="object")
```

Out[ ]:

	segment_key	data	route_schedule_uuid	route_type	trip_uuid	source_center	source_name	destination_center	destination_name
count	26368	26368	26368	26368	26368	26368	26368	26368	26368
unique	26368	2	1504	2	14817	1508	1508	1481	1481
top	trip-153861007249500192IND847103AAAIND847404AAB	training	thanos::sroute:f8c83fd0-6554-44f3-9408-32465bd...	FTL	trip-153710494321650505	IND000000ACB	Gurgaon_Bilaspur_HB (Haryana)	IND000000ACB	Gurgaon_Bilaspur_HB (Haryana)
freq	1	18947	111	13939	8	1063	1063	928	928

Feature Engineering

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

In [ ]:
```
segment.head()
```

Out[ ]:

	index	segment_key	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	actual_distance_to_destination	actual_time	osrm_time	osrm_distance	segment_actual_time_sum	segment_osrm_distance_sum	segment_osrm_time_sum	od
0	0	trip-153671041653548748IND209304AAAIND000000ACB	training	2018-09-12 00:00:16.535741	thanos::sroute:d7c989ba-a29b-4a0b-b2f4-288cdc6...	FTL	trip-153671041653548748	IND209304AAA	Kanpur_Central_H_6 (Uttar Pradesh)	IND000000ACB	Gurgaon_Bilaspur_HB (Haryana)	2018-09-12 16:39:46.858469	2018-09-13 13:40:23.123744	1260.0	383.759164	732.0	329.0	446.5496	728.0	670.6205	534.0	
1	1	trip-153671041653548748IND462022AAAIND209304AAA	training	2018-09-12 00:00:16.535741	thanos::sroute:d7c989ba-a29b-4a0b-b2f4-288cdc6...	FTL	trip-153671041653548748	IND462022AAA	Bhopal_Trnsport_H (Madhya Pradesh)	IND209304AAA	Kanpur_Central_H_6 (Uttar Pradesh)	2018-09-12 00:00:16.535741	2018-09-12 16:39:46.858469	999.0	440.973689	830.0	388.0	544.8027	820.0	649.8528	474.0	
2	2	trip-153671042288605164IND561203AABIND562101AAA	training	2018-09-12 00:00:22.886430	thanos::sroute:3a1b0ab2-bb0b-4c53-8c59-eb2a2c0...	Carting	trip-153671042288605164	IND561203AAB	Doddablpur_ChikaDPP_D (Karnataka)	IND562101AAA	Chikblapur_ShntiSgr_D (Karnataka)	2018-09-12 02:03:09.655591	2018-09-12 03:01:59.598855	58.0	24.644021	47.0	26.0	28.1994	46.0	28.1995	26.0	
3	3	trip-153671042288605164IND572101AAAIND561203AAB	training	2018-09-12 00:00:22.886430	thanos::sroute:3a1b0ab2-bb0b-4c53-8c59-eb2a2c0...	Carting	trip-153671042288605164	IND572101AAA	Tumkur_Veersagr_I (Karnataka)	IND561203AAB	Doddablpur_ChikaDPP_D (Karnataka)	2018-09-12 00:00:22.886430	2018-09-12 02:03:09.655591	122.0	48.542890	96.0	42.0	56.9116	95.0	55.9899	39.0	
4	4	trip-153671043369099517IND000000ACBIND160002AAC	training	2018-09-12 00:00:33.691250	thanos::sroute:de5e208e-7641-45e6-8100-4d9fb1e...	FTL	trip-153671043369099517	IND000000ACB	Gurgaon_Bilaspur_HB (Haryana)	IND160002AAC	Chandigarh_Mehmdpur_H (Punjab)	2018-09-14 03:40:17.106733	2018-09-14 17:34:55.442454	834.0	237.439610	611.0	212.0	281.2109	608.0	317.7408	231.0	

In [ ]:
```
create_trip_dict = {

    'data' : 'first',
    'trip_creation_time': 'first',
    'route_schedule_uuid' : 'first',
    'route_type' : 'first',
    'trip_uuid' : 'first',
    'od_start_time' : 'first',
    'od_end_time' : '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 [ ]:
```
trip = segment.groupby("trip_uuid").agg(create_trip_dict).reset_index(drop= True)
```

In [ ]:
```
trip.head()
```

Out[ ]:

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	od_start_time	od_end_time	source_center	source_name	destination_center	destination_name	start_scan_to_end_scan	od_time_diff_hour	actual_distance_to_destination	actual_time	osrm_time	osrm_distance	segment_actual_time_sum	segment_osrm_distance_sum	segment_osrm_time_sum
0	training	2018-09-12 00:00:16.535741	thanos::sroute:d7c989ba-a29b-4a0b-b2f4-288cdc6...	FTL	trip-153671041653548748	2018-09-12 16:39:46.858469	2018-09-13 13:40:23.123744	IND209304AAA	Kanpur_Central_H_6 (Uttar Pradesh)	IND209304AAA	Kanpur_Central_H_6 (Uttar Pradesh)	2259.0	2260.109800	824.732854	1562.0	717.0	991.3523	1548.0	1320.4733	1008.0
1	training	2018-09-12 00:00:22.886430	thanos::sroute:3a1b0ab2-bb0b-4c53-8c59-eb2a2c0...	Carting	trip-153671042288605164	2018-09-12 02:03:09.655591	2018-09-12 03:01:59.598855	IND561203AAB	Doddablpur_ChikaDPP_D (Karnataka)	IND561203AAB	Doddablpur_ChikaDPP_D (Karnataka)	180.0	181.611874	73.186911	143.0	68.0	85.1110	141.0	84.1894	65.0
2	training	2018-09-12 00:00:33.691250	thanos::sroute:de5e208e-7641-45e6-8100-4d9fb1e...	FTL	trip-153671043369099517	2018-09-14 03:40:17.106733	2018-09-14 17:34:55.442454	IND000000ACB	Gurgaon_Bilaspur_HB (Haryana)	IND000000ACB	Gurgaon_Bilaspur_HB (Haryana)	3933.0	3934.362520	1927.404273	3347.0	1740.0	2354.0665	3308.0	2545.2678	1941.0
3	training	2018-09-12 00:01:00.113710	thanos::sroute:f0176492-a679-4597-8332-bbd1c7f...	Carting	trip-153671046011330457	2018-09-12 00:01:00.113710	2018-09-12 01:41:29.809822	IND400072AAB	Mumbai Hub (Maharashtra)	IND401104AAA	Mumbai_MiraRd_IP (Maharashtra)	100.0	100.494935	17.175274	59.0	15.0	19.6800	59.0	19.8766	16.0
4	training	2018-09-12 00:02:09.740725	thanos::sroute:d9f07b12-65e0-4f3b-bec8-df06134...	FTL	trip-153671052974046625	2018-09-12 00:02:09.740725	2018-09-12 02:34:10.515593	IND583101AAA	Bellary_Dc (Karnataka)	IND583119AAA	Sandur_WrdN1DPP_D (Karnataka)	717.0	718.349042	127.448500	341.0	117.0	146.7918	340.0	146.7919	115.0

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

In [ ]:
```
trip.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14817 entries, 0 to 14816
Data columns (total 20 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   data                            14817 non-null  object
 1   trip_creation_time              14817 non-null  datetime64[ns]
 2   route_schedule_uuid             14817 non-null  object
 3   route_type                      14817 non-null  object
 4   trip_uuid                       14817 non-null  object
 5   od_start_time                   14817 non-null  datetime64[ns]
 6   od_end_time                     14817 non-null  datetime64[ns]
 7   source_center                   14817 non-null  object
 8   source_name                     14817 non-null  object
 9   destination_center              14817 non-null  object
 10  destination_name                14817 non-null  object
 11  start_scan_to_end_scan          14817 non-null  float64
 12  od_time_diff_hour               14817 non-null  float64
 13  actual_distance_to_destination  14817 non-null  float64
 14  actual_time                     14817 non-null  float64
 15  osrm_time                       14817 non-null  float64
 16  osrm_distance                   14817 non-null  float64
 17  segment_actual_time_sum         14817 non-null  float64
 18  segment_osrm_distance_sum       14817 non-null  float64
 19  segment_osrm_time_sum           14817 non-null  float64
dtypes: datetime64[ns](3), float64(9), object(8)
memory usage: 2.3+ MB
```

In [ ]:
```
def place2state(x):
    # transform  "gurgaon_bilaspur_hb (haryana)" into "haryana""
    state = x.split('(')[1]
    return state[:-1] #removing ')' from ending

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

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

    #Now dealing 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'
```

In [ ]:
```
trip['destination_state'] = trip['destination_name'].apply(lambda x: place2state(x) if x.split("_")[0] != "missing" else "missing")
trip['destination_city']  = trip['destination_name'].apply(lambda x: place2city(x) if x.split("_")[0] != "missing" else "missing")
trip['destination_place'] = trip['destination_name'].apply(lambda x: place2city_place(x) if x.split("_")[0] != "missing" else "missing")
trip['destination_code']  = trip['destination_name'].apply(lambda x: place2code(x) if x.split("_")[0] != "missing" else "missing")
trip['source_state'] = trip['source_name'].apply(lambda x: place2state(x) if x.split("_")[0] != "missing" else "missing")
trip['source_city']  = trip['source_name'].apply(lambda x: place2city(x) if x.split("_")[0] != "missing" else "missing")
trip['source_place'] = trip['source_name'].apply(lambda x: place2city_place(x) if x.split("_")[0] != "missing" else "missing")
trip['source_code']  = trip['source_name'].apply(lambda x: place2code(x) if x.split("_")[0] != "missing" else "missing")
```

In [ ]:
```
trip.describe(include="object")
```

Out[ ]:

	data	route_schedule_uuid	route_type	trip_uuid	source_center	source_name	destination_center	destination_name	destination_state	destination_city	destination_place	destination_code	source_state	source_city	source_place	source_code
count	14817	14817	14817	14817	14817	14817	14817	14817	14817	14817	14817	14817	14817	14817	14817	14817
unique	2	1504	2	14817	938	938	1042	1042	32	840	866	33	30	716	772	32
top	training	thanos::sroute:a16bfa03-3462-4bce-9c82-5784c7d...	Carting	trip-153861118270144424	IND000000ACB	gurgaon_bilaspur_hb (haryana)	IND000000ACB	gurgaon_bilaspur_hb (haryana)	maharashtra	bengaluru	bilaspur	d	maharashtra	gurgaon	bilaspur	hb
freq	10654	53	8908	1	1063	1063	821	821	2561	1221	864	2868	2714	1139	1085	3222

In [ ]:
```
trip[['destination_state', 'destination_city', 'destination_place', 'destination_code']]
```

Out[ ]:

	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
...	...	...	...	...
14812	punjab	chandigarh	mehmdpur	h
14813	haryana	faridabad	blbgarh	dc
14814	uttar pradesh	kanpur	govndngr	dc
14815	tamil nadu	tirchchndr	shnmgprm	d
14816	karnataka	sandur	wrdn1dpp	d

14817 rows × 4 columns

In [ ]:
```
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_time'] = trip['trip_creation_time'].dt.time
trip['trip_start_hour'] = trip['od_start_time'].dt.hour
trip['trip_end_hour'] = trip['od_end_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
```

In [ ]:
```
trip[['trip_year', 'trip_month', 'trip_day', 'trip_week', 'trip_dayofweek']]
```

Out[ ]:

	trip_year	trip_month	trip_day	trip_week	trip_dayofweek
0	2018	9	12	37	2
1	2018	9	12	37	2
2	2018	9	12	37	2
3	2018	9	12	37	2
4	2018	9	12	37	2
...	...	...	...	...	...
14812	2018	10	3	40	2
14813	2018	10	3	40	2
14814	2018	10	3	40	2
14815	2018	10	3	40	2
14816	2018	10	3	40	2

14817 rows × 5 columns

In [ ]:
```
trip.head(5)
```

Out[ ]:

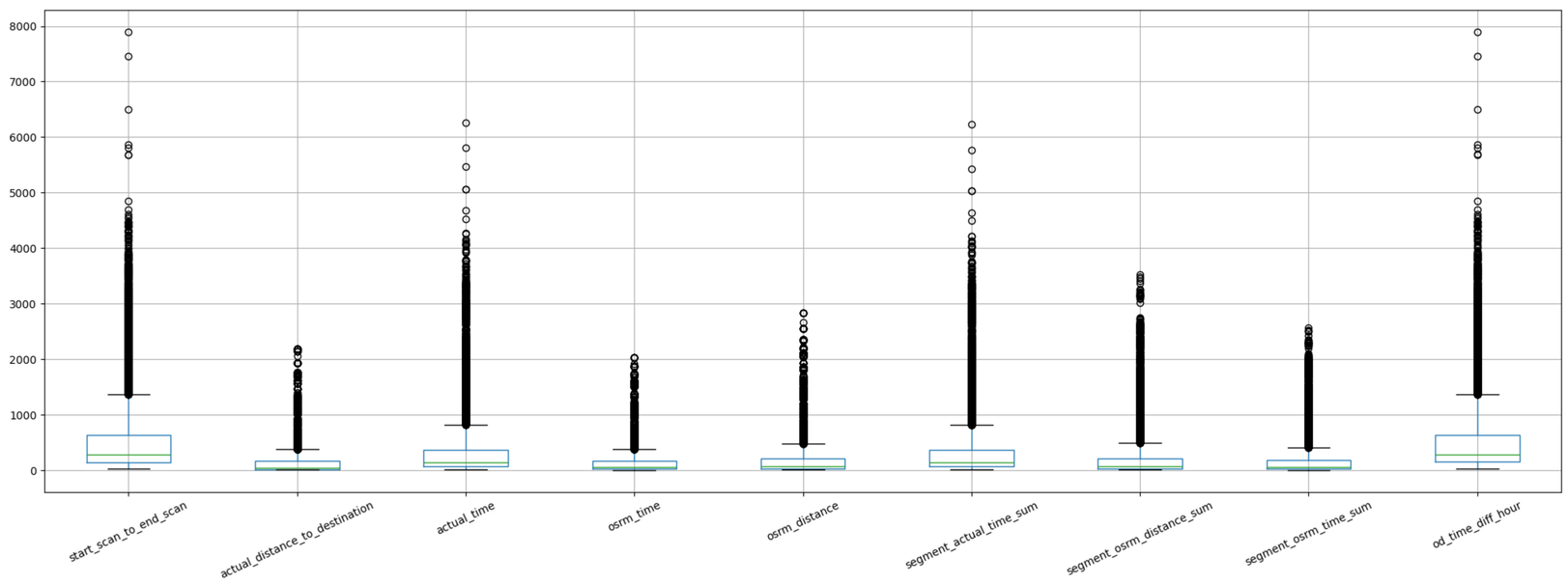
	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	od_start_time	od_end_time	source_center	source_name	destination_center	destination_name	start_scan_to_end_scan	od_time_diff_hour	actual_distance_to_destination	actual_time	osrm_time	osrm_distance	segment_actual_time_sum	segment_osrm_distance_sum	segment_osrm_time_sum	destination_state	destination_city	destina
0	training	2018-09-12 00:00:16.535741	thanos::sroute:d7c989ba-a29b-4a0b-b2f4-288cdc6...	FTL	trip-153671041653548748	2018-09-12 16:39:46.858469	2018-09-13 13:40:23.123744	IND209304AAA	kanpur_central_h_6 (uttar pradesh)	IND209304AAA	kanpur_central_h_6 (uttar pradesh)	2259.0	2260.109800	824.732854	1562.0	717.0	991.3523	1548.0	1320.4733	1008.0	uttar pradesh	kanpur	
1	training	2018-09-12 00:00:22.886430	thanos::sroute:3a1b0ab2-bb0b-4c53-8c59-eb2a2c0...	Carting	trip-153671042288605164	2018-09-12 02:03:09.655591	2018-09-12 03:01:59.598855	IND561203AAB	doddablpur_chikadpp_d (karnataka)	IND561203AAB	doddablpur_chikadpp_d (karnataka)	180.0	181.611874	73.186911	143.0	68.0	85.1110	141.0	84.1894	65.0	karnataka	doddablpur	
2	training	2018-09-12 00:00:33.691250	thanos::sroute:de5e208e-7641-45e6-8100-4d9fb1e...	FTL	trip-153671043369099517	2018-09-14 03:40:17.106733	2018-09-14 17:34:55.442454	IND000000ACB	gurgaon_bilaspur_hb (haryana)	IND000000ACB	gurgaon_bilaspur_hb (haryana)	3933.0	3934.362520	1927.404273	3347.0	1740.0	2354.0665	3308.0	2545.2678	1941.0	haryana	gurgaon	
3	training	2018-09-12 00:01:00.113710	thanos::sroute:f0176492-a679-4597-8332-bbd1c7f...	Carting	trip-153671046011330457	2018-09-12 00:01:00.113710	2018-09-12 01:41:29.809822	IND400072AAB	mumbai hub (maharashtra)	IND401104AAA	mumbai_mirard_ip (maharashtra)	100.0	100.494935	17.175274	59.0	15.0	19.6800	59.0	19.8766	16.0	maharashtra	mumbai	
4	training	2018-09-12 00:02:09.740725	thanos::sroute:d9f07b12-65e0-4f3b-bec8-df06134...	FTL	trip-153671052974046625	2018-09-12 00:02:09.740725	2018-09-12 02:34:10.515593	IND583101AAA	bellary_dc (karnataka)	IND583119AAA	sandur_wrdn1dpp_d (karnataka)	717.0	718.349042	127.448500	341.0	117.0	146.7918	340.0	146.7919	115.0	karnataka	sandur	

In [ ]:

In-depth analysis:

In [ ]:
```
## Numerical Columns

num_cols = ['start_scan_to_end_scan', 'actual_distance_to_destination', 'actual_time', 'osrm_time',
            'osrm_distance', 'segment_actual_time_sum', 'segment_osrm_distance_sum',
            'segment_osrm_time_sum', 'od_time_diff_hour']
```

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



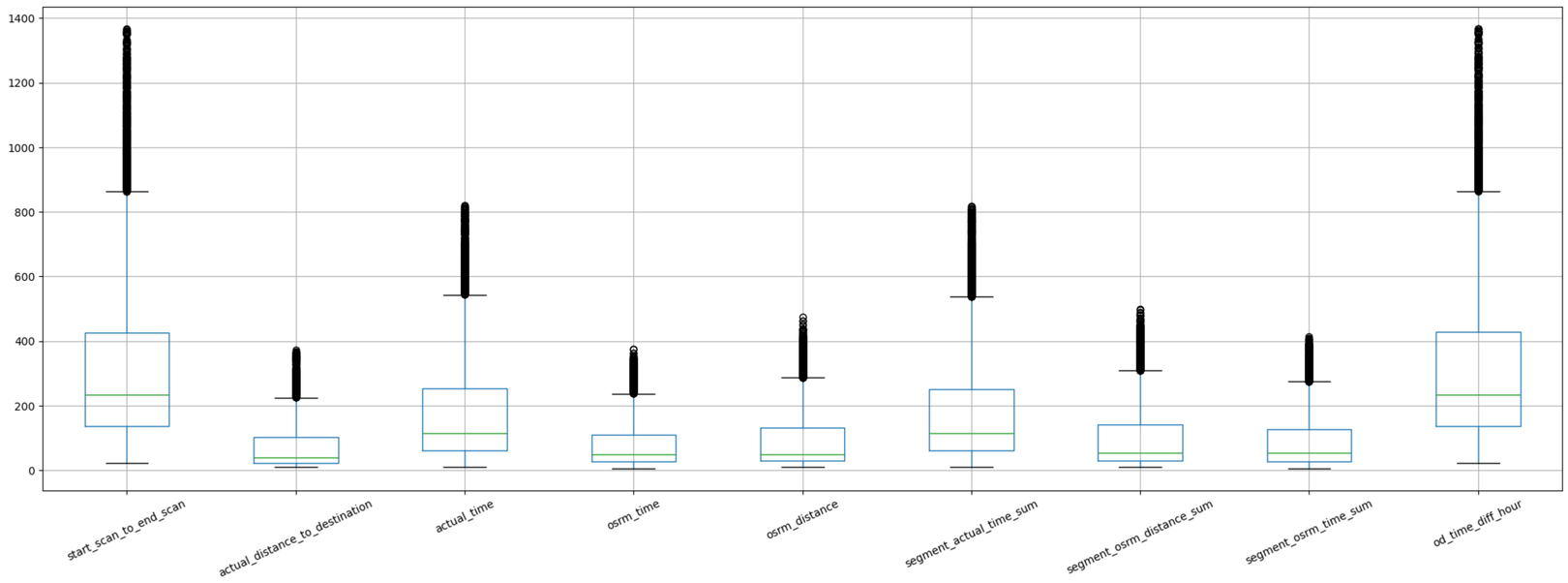
In [ ]:
```
Q1 = trip[num_cols].quantile(0.25)
Q3 = trip[num_cols].quantile(0.75)

IQR = Q3 - Q1
```

In [ ]:
```
trip = trip[~((trip[num_cols] < (Q1 - 1.5 * IQR)) | (trip[num_cols] > (Q3 + 1.5 * IQR))).any(axis=1)]
trip = trip.reset_index(drop=True)
```

In [ ]:
```
trip.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12759 entries, 0 to 12758
Data columns (total 36 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   data                            12759 non-null  object
 1   trip_creation_time              12759 non-null  datetime64[ns]
 2   route_schedule_uuid             12759 non-null  object
 3   route_type                      12759 non-null  object
 4   trip_uuid                       12759 non-null  object
 5   od_start_time                   12759 non-null  datetime64[ns]
 6   od_end_time                     12759 non-null  datetime64[ns]
 7   source_center                   12759 non-null  object
 8   source_name                     12759 non-null  object
 9   destination_center              12759 non-null  object
 10  destination_name                12759 non-null  object
 11  start_scan_to_end_scan          12759 non-null  float64
 12  od_time_diff_hour               12759 non-null  float64
 13  actual_distance_to_destination  12759 non-null  float64
 14  actual_time                     12759 non-null  float64
 15  osrm_time                       12759 non-null  float64
 16  osrm_distance                   12759 non-null  float64
 17  segment_actual_time_sum         12759 non-null  float64
 18  segment_osrm_distance_sum       12759 non-null  float64
 19  segment_osrm_time_sum           12759 non-null  float64
 20  destination_state               12759 non-null  object
 21  destination_city                12759 non-null  object
 22  destination_place               12759 non-null  object
 23  destination_code                12759 non-null  object
 24  source_state                    12759 non-null  object
 25  source_city                     12759 non-null  object
 26  source_place                    12759 non-null  object
 27  source_code                     12759 non-null  object
 28  trip_year                       12759 non-null  int32
 29  trip_month                      12759 non-null  int32
 30  trip_time                       12759 non-null  object
 31  trip_start_hour                 12759 non-null  int32
 32  trip_end_hour                   12759 non-null  int32
 33  trip_day                        12759 non-null  int32
 34  trip_week                       12759 non-null  UInt32
 35  trip_dayofweek                  12759 non-null  int32
dtypes: UInt32(1), datetime64[ns](3), float64(9), int32(6), object(17)
memory usage: 3.2+ MB
```

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



In [ ]:
```
trip["route_type"].value_counts()
```

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

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

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

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

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

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

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

Out[ ]:

	start_scan_to_end_scan	actual_distance_to_destination	actual_time	osrm_time	osrm_distance	segment_actual_time_sum	segment_osrm_distance_sum	segment_osrm_time_sum	od_time_diff_hour
0	-0.551781	0.004976	-0.223508	-0.150681	-0.080602	-0.227130	-0.151645	-0.268226	-0.548105
1	-0.862589	-0.766880	-0.751536	-0.878175	-0.806104	-0.746018	-0.825232	-0.879530	-0.862847
2	1.534514	0.752716	1.021129	0.521909	0.603318	1.032122	0.504029	0.355554	1.534486
3	-0.516816	-0.664606	-0.738964	-0.768365	-0.713134	-0.739690	-0.739473	-0.792201	-0.513666
4	-0.870359	-0.878152	-0.971547	-0.905628	-0.891056	-0.967495	-0.907535	-0.916957	-0.872505
...	...	...	...	...	...	...	...	...	...
12754	-0.252629	-0.207579	-0.600671	-0.233038	-0.209756	-0.600476	-0.354145	-0.305653	-0.251600
12755	-1.017993	-0.789776	-0.990406	-0.919354	-0.845930	-0.986478	-0.864909	-0.941908	-1.017680
12756	0.384526	-0.470472	0.650252	-0.425207	-0.371190	0.658775	0.065130	0.018713	0.385089
12757	0.097029	0.852973	0.537103	1.372940	0.872963	0.513234	1.307779	1.677967	0.099486
12758	0.120340	-0.092938	0.606250	-0.150681	-0.130856	0.614480	-0.189462	-0.243275	0.122358

12759 rows × 9 columns

In [ ]:
```
trip[num_cols].describe()
```

Out[ ]:

	start_scan_to_end_scan	actual_distance_to_destination	actual_time	osrm_time	osrm_distance	segment_actual_time_sum	segment_osrm_distance_sum	segment_osrm_time_sum	od_time_diff_hour
count	1.275900e+04	1.275900e+04	1.275900e+04	1.275900e+04	1.275900e+04	1.275900e+04	1.275900e+04	1.275900e+04	1.275900e+04
mean	7.295329e-17	-3.341372e-18	-2.227581e-17	-7.685155e-17	-2.227581e-18	-3.786888e-17	1.113791e-18	-8.576188e-17	1.102653e-16
std	1.000039e+00	1.000039e+00	1.000039e+00	1.000039e+00	1.000039e+00	1.000039e+00	1.000039e+00	1.000039e+00	1.000039e+00
min	-1.161741e+00	-8.795036e-01	-1.065838e+00	-1.001712e+00	-9.237163e-01	-1.062413e+00	-9.383855e-01	-1.004286e+00	-1.161744e+00
25%	-7.227256e-01	-7.085171e-01	-7.389638e-01	-7.134593e-01	-7.096382e-01	-7.396902e-01	-7.245542e-01	-7.298227e-01	-7.218003e-01
50%	-3.419863e-01	-4.706378e-01	-3.995173e-01	-3.977542e-01	-4.841527e-01	-3.979836e-01	-4.649355e-01	-4.179329e-01	-3.433922e-01
75%	4.078368e-01	4.152532e-01	4.742429e-01	4.395507e-01	4.382701e-01	4.689385e-01	4.480913e-01	5.052611e-01	4.071949e-01
max	4.055940e+00	4.142566e+00	4.032144e+00	4.077023e+00	4.232902e+00	4.056857e+00	4.180276e+00	4.073281e+00	4.052774e+00

In [ ]:
```
trip.describe(include="object")
```

Out[ ]:

	data	route_schedule_uuid	trip_uuid	source_center	source_name	destination_center	destination_name	destination_state	destination_city	destination_place	destination_code	source_state	source_city	source_place	source_code	trip_time
count	12759	12759	12759	12759	12759	12759	12759	12759	12759	12759	12759	12759	12759	12759	12759	12759
unique	2	1366	12759	909	909	1010	1010	32	812	841	32	29	692	749	32	12759
top	training	thanos::sroute:a16bfa03-3462-4bce-9c82-5784c7d...	trip-153861118270144424	IND000000ACB	gurgaon_bilaspur_hb (haryana)	IND000000ACB	gurgaon_bilaspur_hb (haryana)	maharashtra	bengaluru	central	d	maharashtra	bengaluru	central	hb	23:59:42.701692
freq	9112	53	1	679	679	555	555	2286	1206	673	2513	2309	1113	727	2544	1

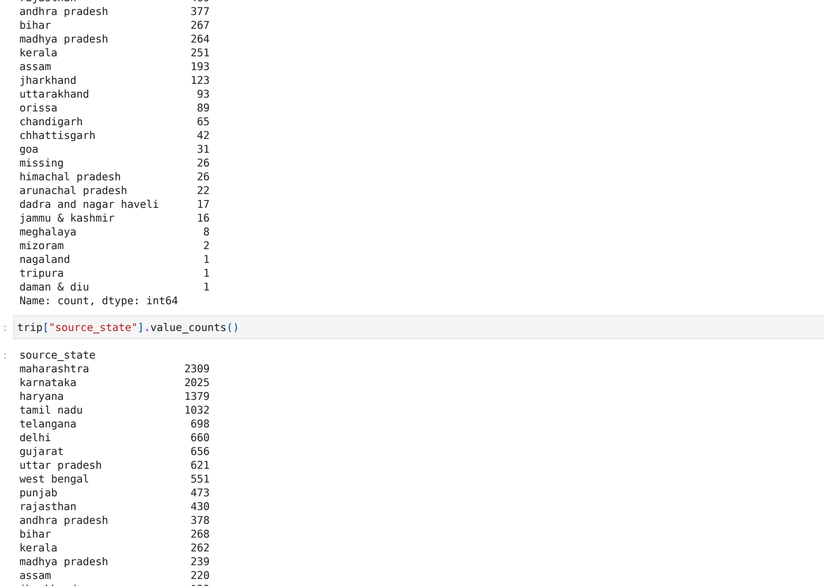
In [ ]:
```
trip["destination_state"].value_counts()
```

Out[ ]:
```
destination_state
maharashtra               2286
karnataka                 2070
haryana                   1337
tamil nadu                1040
telangana                  682
gujarat                    653
uttar pradesh              625
delhi                      574
west bengal                559
punjab                     549
rajasthan                  469
andhra pradesh             377
bihar                      267
madhya pradesh             264
kerala                     251
assam                      193
jharkhand                  123
uttarakhand                 93
orissa                      89
chandigarh                  65
chhattisgarh                42
goa                         31
missing                     26
himachal pradesh            26
arunachal pradesh           22
dadra and nagar haveli      17
jammu & kashmir             16
meghalaya                    8
mizoram                      2
nagaland                     1
tripura                      1
daman & diu                  1
Name: count, dtype: int64
```

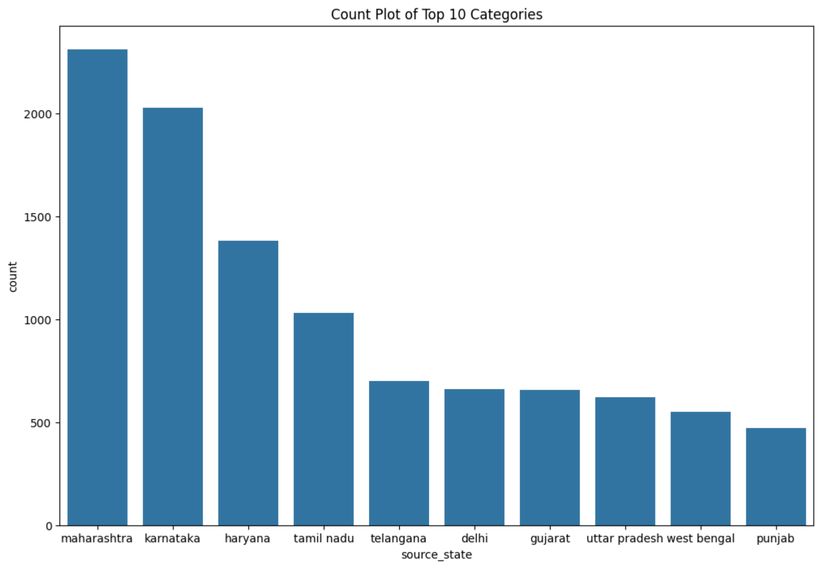
In [ ]:
```
trip["source_state"].value_counts()
```

Out[ ]:
```
source_state
maharashtra               2309
karnataka                 2025
haryana                   1379
tamil nadu                1032
telangana                  698
delhi                      660
gujarat                    656
uttar pradesh              621
west bengal                551
punjab                     473
rajasthan                  430
andhra pradesh             378
bihar                      268
kerala                     262
madhya pradesh             239
assam                      220
jharkhand                  123
uttarakhand                 94
orissa                      94
chandigarh                  93
chhattisgarh                42
goa                         34
jammu & kashmir             16
missing                     16
dadra and nagar haveli      15
pondicherry                 12
himachal pradesh            12
nagaland                     4
arunachal pradesh            3
Name: count, dtype: int64
```

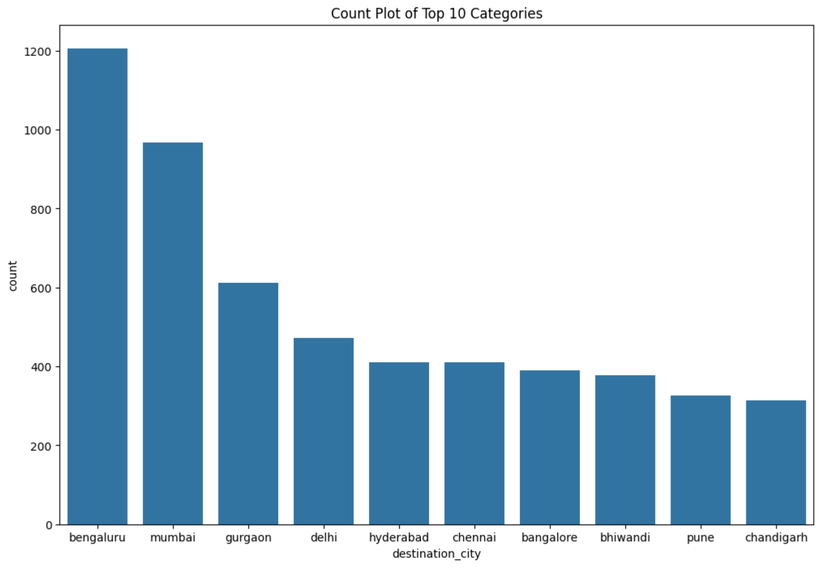
In [ ]:
```
# Filter the DataFrame to include only the top 10 categories
top_categories = trip["destination_state"].value_counts().nlargest(10).index
df_top_categories = trip[trip["destination_state"].isin(top_categories)]

# Create a count plot or bar plot using Seaborn
plt.figure(figsize=(12, 8))
sns.countplot(x="destination_state", data=df_top_categories, order=top_categories)
plt.title('Count Plot of Top 10 Categories')
plt.show()
```



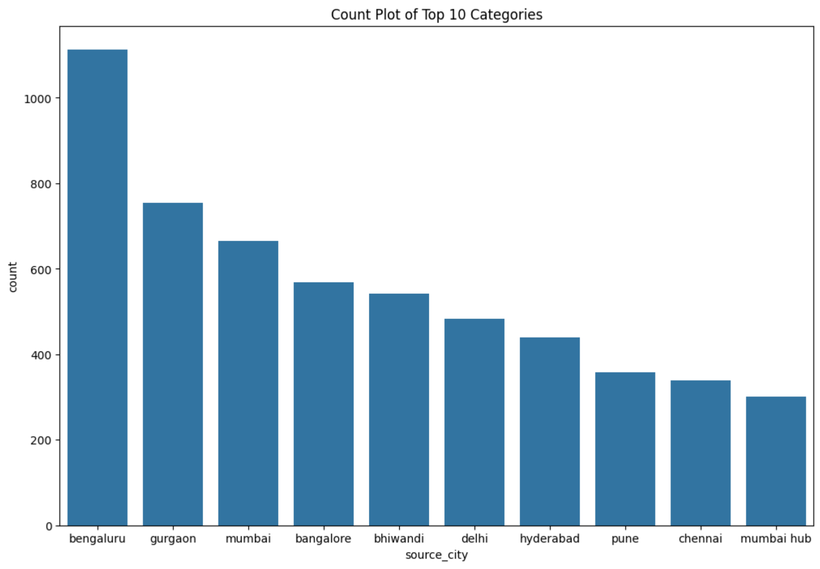
In [ ]:
```
# Filter the DataFrame to include only the top 10 categories
top_categories = trip["source_state"].value_counts().nlargest(10).index
df_top_categories = trip[trip["source_state"].isin(top_categories)]

# Create a count plot or bar plot using Seaborn
plt.figure(figsize=(12, 8))
sns.countplot(x="source_state", data=df_top_categories, order=top_categories)
plt.title('Count Plot of Top 10 Categories')
plt.show()
```



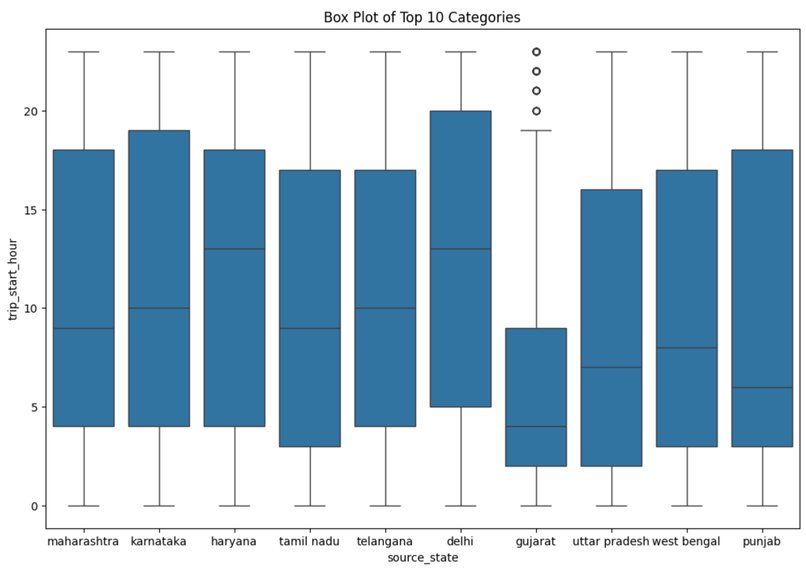
In [ ]:
```
# Filter the DataFrame to include only the top 10 categories
top_categories = trip["destination_city"].value_counts().nlargest(10).index
df_top_categories = trip[trip["destination_city"].isin(top_categories)]

# Create a count plot or bar plot using Seaborn
plt.figure(figsize=(12, 8))
sns.countplot(x="destination_city", data=df_top_categories, order=top_categories)
plt.title('Count Plot of Top 10 Categories')
plt.show()
```



In [ ]:
```
# Filter the DataFrame to include only the top 10 categories
top_categories = trip["source_city"].value_counts().nlargest(10).index
df_top_categories = trip[trip["source_city"].isin(top_categories)]

# Create a count plot or bar plot using Seaborn
plt.figure(figsize=(12, 8))
sns.countplot(x="source_city", data=df_top_categories, order=top_categories)
plt.title('Count Plot of Top 10 Categories')
plt.show()
```



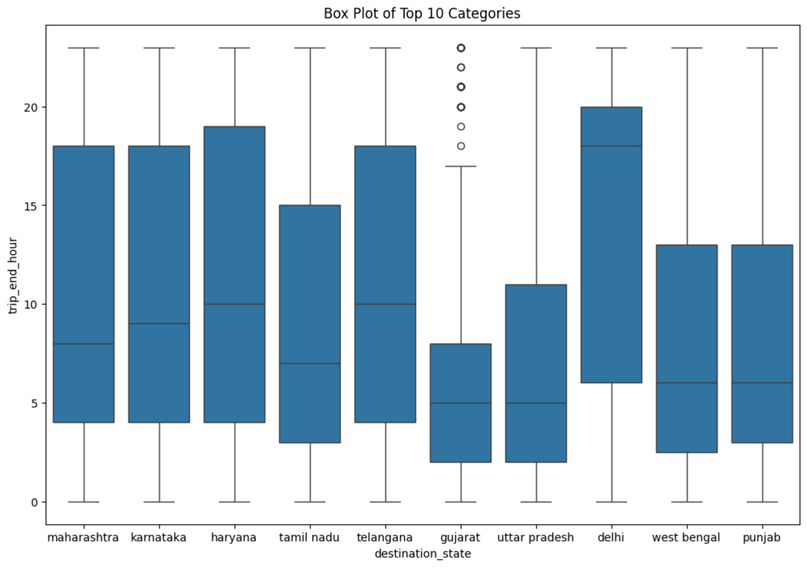
In [ ]:
```
## Time Vs State

# Replace these with your actual column names
category_column = 'source_state'
value_column = 'trip_start_hour'

# Filter the DataFrame to include only the top 10 categories
top_categories = trip[category_column].value_counts().nlargest(10).index
df_top_categories = trip[trip[category_column].isin(top_categories)]

# Create a box plot using Seaborn
plt.figure(figsize=(12, 8))
sns.boxplot(x=category_column, y=value_column, data=df_top_categories, order=top_categories)
plt.title('Box Plot of Top 10 Categories')
plt.show()
```



In [ ]:
```
Q1 = trip[trip["source_state"]=="gujarat"]["trip_start_hour"].quantile(0.25)
Q2 = trip[trip["source_state"]=="gujarat"]["trip_start_hour"].quantile(0.5)
Q3 = trip[trip["source_state"]=="gujarat"]["trip_start_hour"].quantile(0.75)

Q1, Q2, Q3
```

Out[ ]:
```
(2.0, 4.0, 9.0)
```

In [ ]:
```
## Time Vs State

# Replace these with your actual column names
category_column = 'destination_state'
value_column = 'trip_end_hour'

# Filter the DataFrame to include only the top 10 categories
top_categories = trip[category_column].value_counts().nlargest(10).index
df_top_categories = trip[trip[category_column].isin(top_categories)]

# Create a box plot using Seaborn
plt.figure(figsize=(12, 8))
sns.boxplot(x=category_column, y=value_column, data=df_top_categories, order=top_categories)
plt.title('Box Plot of Top 10 Categories')
plt.show()
```



In [ ]:
```
Q1 = trip[trip["source_state"]=="gujarat"]["trip_end_hour"].quantile(0.25)
Q2 = trip[trip["source_state"]=="gujarat"]["trip_end_hour"].quantile(0.5)
Q3 = trip[trip["source_state"]=="gujarat"]["trip_end_hour"].quantile(0.75)

Q1, Q2, Q3
```

Out[ ]:
```
(2.0, 4.0, 8.0)
```

In [ ]:
```
trip.head()
```

Out[ ]:

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	od_start_time	od_end_time	source_center	source_name	destination_center	destination_name	start_scan_to_end_scan	od_time_diff_hour	actual_distance_to_destination	actual_time	osrm_time	osrm_distance	segment_actual_time_sum	segment_osrm_distance_sum	segment_osrm_time_sum	destination_state	destination_city	destina
0	training	2018-09-12 00:00:22.886430	thanos::sroute:3a1b0ab2-bb0b-4c53-8c59-eb2a2c0...	1	trip-153671042288605164	2018-09-12 02:03:09.655591	2018-09-12 03:01:59.598855	IND561203AAB	doddablpur_chikadpp_d (karnataka)	IND561203AAB	doddablpur_chikadpp_d (karnataka)	-0.551781	-0.548105	0.004976	-0.223508	-0.150681	-0.080602	-0.227130	-0.151645	-0.268226	karnataka	doddablpur	
1	training	2018-09-12 00:01:00.113710	thanos::sroute:f0176492-a679-4597-8332-bbd1c7f...	1	trip-153671046011330457	2018-09-12 00:01:00.113710	2018-09-12 01:41:29.809822	IND400072AAB	mumbai hub (maharashtra)	IND401104AAA	mumbai_mirard_ip (maharashtra)	-0.862589	-0.862847	-0.766880	-0.751536	-0.878175	-0.806104	-0.746018	-0.825232	-0.879530	maharashtra	mumbai	
2	training	2018-09-12 00:02:09.740725	thanos::sroute:d9f07b12-65e0-4f3b-bec8-df06134...	0	trip-153671052974046625	2018-09-12 00:02:09.740725	2018-09-12 02:34:10.515593	IND583101AAA	bellary_dc (karnataka)	IND583119AAA	sandur_wrdn1dpp_d (karnataka)	1.534514	1.534486	0.752716	1.021129	0.521909	0.603318	1.032122	0.504029	0.355554	karnataka	sandur	
3	training	2018-09-12 00:02:34.161600	thanos::sroute:9bf03170-d0a2-4a3f-aa4d-9aaab3d...	1	trip-153671055416136166	2018-09-12 02:12:10.755603	2018-09-12 03:13:03.432532	IND600056AAA	chennai_poonamallee (tamil nadu)	IND600056AAA	chennai_poonamallee (tamil nadu)	-0.516816	-0.513666	-0.664606	-0.738964	-0.768365	-0.713134	-0.739690	-0.739473	-0.792201	tamil nadu	chennai	
4	training	2018-09-12 00:04:22.011653	thanos::sroute:a97698cc-846e-41a7-916b-88b1741...	1	trip-153671066201138152	2018-09-12 00:04:22.011653	2018-09-12 01:42:22.349694	IND600044AAD	chennai_chrompet_dpc (tamil nadu)	IND600048AAA	chennai_vandalur_dc (tamil nadu)	-0.870359	-0.872505	-0.878152	-0.971547	-0.905628	-0.891056	-0.967495	-0.907535	-0.916957	tamil nadu	chennai	

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

Out[ ]:
```
route_type
1    8817
0    3942
Name: count, dtype: int64
```

In [ ]:
```
route_id=trip["route_schedule_uuid"].value_counts().reset_index()["route_schedule_uuid"][0]
route_id
```

Out[ ]:
```
'thanos::sroute:a16bfa03-3462-4bce-9c82-5784c7d315e6'
```

In [ ]:
```
trip[trip["route_schedule_uuid"]==route_id]["source_city"].value_counts()
```

Out[ ]:
```
source_city
lowerparel    53
Name: count, dtype: int64
```

In [ ]:
```
trip[trip["route_schedule_uuid"]==route_id]["destination_city"].value_counts()
```

Out[ ]:
```
destination_city
mumbai    53
Name: count, dtype: int64
```

In [ ]:
```
trip[trip["route_schedule_uuid"]==route_id]["source_place"].value_counts()
```

Out[ ]:
```
source_place
lowerparel    53
Name: count, dtype: int64
```

In [ ]:
```
df[df["route_schedule_uuid"]==route_id]['actual_time'].mean()
```

Out[ ]:
```
42.67289719626168
```

In [ ]:
```
df[df["route_schedule_uuid"]==route_id]['actual_distance_to_destination'].mean()
```

Out[ ]:
```
13.179664875398535
```

In [ ]:
```

```

Hypothesis Testing

Why T-Test ?

With a sample size greater than 12k, we can use a t-test even if the sample size is large. The t-test will provide reliable results in this situation. In fact, when sample sizes are very large, the t-test and z-test tend to give similar results.

In [ ]:
```
trip.head()
```

Out[ ]:

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	od_start_time	od_end_time	source_center	source_name	destination_center	destination_name	start_scan_to_end_scan	od_time_diff_hour	actual_distance_to_destination	actual_time	osrm_time	osrm_distance	segment_actual_time_sum	segment_osrm_distance_sum	segment_osrm_time_sum	destination_state	destination_city	destina
0	training	2018-09-12 00:00:22.886430	thanos::sroute:3a1b0ab2-bb0b-4c53-8c59-eb2a2c0...	1	trip-153671042288605164	2018-09-12 02:03:09.655591	2018-09-12 03:01:59.598855	IND561203AAB	doddablpur_chikadpp_d (karnataka)	IND561203AAB	doddablpur_chikadpp_d (karnataka)	-0.551781	-0.548105	0.004976	-0.223508	-0.150681	-0.080602	-0.227130	-0.151645	-0.268226	karnataka	doddablpur	
1	training	2018-09-12 00:01:00.113710	thanos::sroute:f0176492-a679-4597-8332-bbd1c7f...	1	trip-153671046011330457	2018-09-12 00:01:00.113710	2018-09-12 01:41:29.809822	IND400072AAB	mumbai hub (maharashtra)	IND401104AAA	mumbai_mirard_ip (maharashtra)	-0.862589	-0.862847	-0.766880	-0.751536	-0.878175	-0.806104	-0.746018	-0.825232	-0.879530	maharashtra	mumbai	
2	training	2018-09-12 00:02:09.740725	thanos::sroute:d9f07b12-65e0-4f3b-bec8-df06134...	0	trip-153671052974046625	2018-09-12 00:02:09.740725	2018-09-12 02:34:10.515593	IND583101AAA	bellary_dc (karnataka)	IND583119AAA	sandur_wrdn1dpp_d (karnataka)	1.534514	1.534486	0.752716	1.021129	0.521909	0.603318	1.032122	0.504029	0.355554	karnataka	sandur	
3	training	2018-09-12 00:02:34.161600	thanos::sroute:9bf03170-d0a2-4a3f-aa4d-9aaab3d...	1	trip-153671055416136166	2018-09-12 02:12:10.755603	2018-09-12 03:13:03.432532	IND600056AAA	chennai_poonamallee (tamil nadu)	IND600056AAA	chennai_poonamallee (tamil nadu)	-0.516816	-0.513666	-0.664606	-0.738964	-0.768365	-0.713134	-0.739690	-0.739473	-0.792201	tamil nadu	chennai	
4	training	2018-09-12 00:04:22.011653	thanos::sroute:a97698cc-846e-41a7-916b-88b1741...	1	trip-153671066201138152	2018-09-12 00:04:22.011653	2018-09-12 01:42:22.349694	IND600044AAD	chennai_chrompet_dpc (tamil nadu)	IND600048AAA	chennai_vandalur_dc (tamil nadu)	-0.870359	-0.872505	-0.878152	-0.971547	-0.905628	-0.891056	-0.967495	-0.907535	-0.916957	tamil nadu	chennai	

In [ ]:
```
trip["actual_time"]
```

Out[ ]:
```
0       -0.223508
1       -0.751536
2        1.021129
3       -0.738964
4       -0.971547
           ...
12754   -0.600671
12755   -0.990406
12756    0.650252
12757    0.537103
12758    0.606250
Name: actual_time, Length: 12759, dtype: float64
```

In [ ]:
```
trip["segment_osrm_time_sum"]
```

Out[ ]:
```
0       -0.268226
1       -0.879530
2        0.355554
3       -0.792201
4       -0.916957
           ...
12754   -0.305653
12755   -0.941908
12756    0.018713
12757    1.677967
12758   -0.243275
Name: segment_osrm_time_sum, Length: 12759, dtype: float64
```

In [ ]:
```
trip["segment_actual_time_sum"]
```

Out[ ]:
```
0       -0.227130
1       -0.746018
2        1.032122
3       -0.739690
4       -0.967495
           ...
12754   -0.600476
12755   -0.986478
12756    0.658775
12757    0.513234
12758    0.614480
Name: segment_actual_time_sum, Length: 12759, dtype: float64
```

In [ ]:
```
'''
Perform hypothesis testing / visual analysis between :
a. actual_time aggregated value and OSRM time aggregated value.
b. actual_time aggregated value and segment actual time aggregated value.
c. OSRM distance aggregated value and segment OSRM distance aggregated value.
d. OSRM time aggregated value and segment OSRM time aggregated value.
'''
```

Out[ ]:
```
'\nPerform hypothesis testing / visual analysis between :\na. actual_time aggregated value and OSRM time aggregated value.\nb. actual_time aggregated value and segment actual time aggregated value.\nc. OSRM distance aggregated value and segment OSRM distance aggregated value.\nd. OSRM time aggregated value and segment OSRM time aggregated value.\n'
```

In [ ]:
```
# Null Hypothesis: There is no significant difference between the actual time taken and predicted time.
# Alternative Hypothesis: There is a significant difference between the actual time taken and predicted time.

import numpy as np
from scipy.stats import ttest_rel

# Assuming actual_time and predicted_time are NumPy arrays with your data
# Replace these with your actual data
actual_time = trip["actual_time"]
predicted_time = trip["osrm_time"]

# Calculate the differences
differences = actual_time - predicted_time

# Perform the paired t-test
t_statistic, p_value = ttest_rel(actual_time, predicted_time)

# Print the results
print("T-statistic:", t_statistic)
print("P-value:", p_value)

# Decide whether to reject the null hypothesis
alpha = 0.05
if p_value < alpha:
    print("Reject the null hypothesis. There is a significant difference.")
else:
    print("Fail to reject the null hypothesis. There is no significant difference.\nHence we can say that out model is predicting the output correctly")
```

```
T-statistic: 1.2306817515382064e-14
P-value: 0.9999999999999902
Fail to reject the null hypothesis. There is no significant difference.
Hence we can say that out model is predicting the output correctly
```

In [ ]:
```
# Null Hypothesis: There is no significant difference between the actual time taken and segment actual time aggregated.
# Alternative Hypothesis: There is a significant difference between the actual time taken and segment actual time aggregated.

import numpy as np
from scipy.stats import ttest_rel

# Assuming actual_time and predicted_time are NumPy arrays with your data
# Replace these with your actual data
actual_time = trip["actual_time"]
segment_actual_time = trip["segment_actual_time_sum"]

# Calculate the differences
differences = actual_time - segment_actual_time

# Perform the paired t-test
t_statistic, p_value = ttest_rel(actual_time, segment_actual_time)

# Print the results
print("T-statistic:", t_statistic)
print("P-value:", p_value)

# Decide whether to reject the null hypothesis
alpha = 0.05
if p_value < alpha:
    print("Reject the null hypothesis. There is a significant difference.")
else:
    print("Fail to reject the null hypothesis. There is no significant difference.\nHence we can say that out model is predicting the output correctly")
```

```
T-statistic: 2.1913509440793997e-13
P-value: 0.9999999999998251
Fail to reject the null hypothesis. There is no significant difference.
Hence we can say that out model is predicting the output correctly
```

In [ ]:
```python
# Null Hypothesis: There is no significant difference between the OSRM distance aggregated value and segment OSRM distance aggregated value.
# Alternative Hypothesis: There is a significant difference between the OSRM distance aggregated value and segment OSRM distance aggregated value.

import numpy as np
from scipy.stats import ttest_rel

# Assuming actual_time and predicted_time are NumPy arrays with your data
# Replace these with your actual data
predicted_dist = trip["osrm_distance"]
segment_predicted_dist = trip["segment_osrm_distance_sum"]

# Calculate the differences
differences = predicted_dist - segment_predicted_dist

# Perform the paired t-test
t_statistic, p_value = ttest_rel(predicted_dist, segment_predicted_dist)

# Print the results
print("T-statistic:", t_statistic)
print("P-value:", p_value)

# Decide whether to reject the null hypothesis
alpha = 0.05
if p_value < alpha:
    print("Reject the null hypothesis. There is a significant difference.")
else:
    print("Fail to reject the null hypothesis. There is no significant difference.\nHence we can say that out model is predicting the output correctly")
```

```
T-statistic: -5.96838009886474e-15
P-value: 0.9999999999999952
Fail to reject the null hypothesis. There is no significant difference.
Hence we can say that out model is predicting the output correctly
```

In [ ]:
```python
# Null Hypothesis: There is no significant difference between the OSRM time aggregated value and segment OSRM time aggregated value.
# Alternative Hypothesis: There is a significant difference between the OSRM time aggregated value and segment OSRM time aggregated value.

import numpy as np
from scipy.stats import ttest_rel

# Assuming actual_time and predicted_time are NumPy arrays with your data
# Replace these with your actual data
predicted_time = trip["osrm_time"]
segment_predicted_time = trip["segment_osrm_time_sum"]

# Calculate the differences
differences = predicted_time - segment_predicted_time

# Perform the paired t-test
t_statistic, p_value = ttest_rel(predicted_time, segment_predicted_time)

# Print the results
print("T-statistic:", t_statistic)
print("P-value:", p_value)

# Decide whether to reject the null hypothesis
alpha = 0.05
if p_value < alpha:
    print("Reject the null hypothesis. There is a significant difference.")
else:
    print("Fail to reject the null hypothesis. There is no significant difference.\nHence we can say that out model is predicting the output correctly")
```

```
T-statistic: 5.121474559051023e-15
P-value: 0.9999999999999959
Fail to reject the null hypothesis. There is no significant difference.
Hence we can say that out model is predicting the output correctly
```

In [ ]: