Delhivery - Case Study

Problem Statement

The company wants to understand and process the data coming out of data engineering pipelines:

- Clean, sanitize and manipulate data to get useful features out of raw fields
- Make sense out of the raw data and help the data science team to build forecasting models on it

Installing Packages

In [43]:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
from scipy.stats import stats
```

Loading Dataset

In [2]:

```
delhivery = pd.read_csv("https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/001/551/original/delhivery_data.csv?1642751181")

delhivery.head(5)
```

Out[2]:

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name	destination_center	destina
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_M
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_M
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_M
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_M
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_M

5 rows × 24 columns

Understanding Shape and Structure of Data

In [3]:

```
delhivery.shape
```

Out[3]:

```
(144867, 24)
```

There are `144867` rows and `24` columns

In [4]:

```
delhivery.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
```

Missing Values Detection

In [5]:

```
delhivery.isna().sum()
```

Out[5]:

```
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
```

- **`source_name` and `destination_name` contain missing values**

Change data type of feature

In [6]:

```
delhivery["trip_creation_time"] = pd.to_datetime(delhivery["trip_creation_time"])
delhivery["od_start_time"] = pd.to_datetime(delhivery["od_start_time"])
delhivery["od_end_time"] = pd.to_datetime(delhivery["od_end_time"])
```

Range of Datapoint available acc. `trip_creation_time`

In [7]:

```
delhivery["trip_creation_time"].dt.month_name().value_counts()
```

Out[7]:

```
September    127349
October       17518
Name: trip_creation_time, dtype: int64
```

In [8]:

```
delhivery["trip_creation_time"].dt.year.value_counts()
```

Out[8]:

```
2018    144867
Name: trip_creation_time, dtype: int64
```

In [9]:

```
delhivery["trip_creation_time"].dt.day_name().value_counts()
```

Out[9]:

```
Wednesday    26732
Thursday     20481
Friday       20242
Tuesday      19961
Saturday     19936
Monday       19645
Sunday       17870
Name: trip_creation_time, dtype: int64
```

- **Datepoints are from the month of September and October of year 2018**

No. of Unique Categories of Features

In [10]:

```
delhivery.nunique()
```

Out[10]:

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

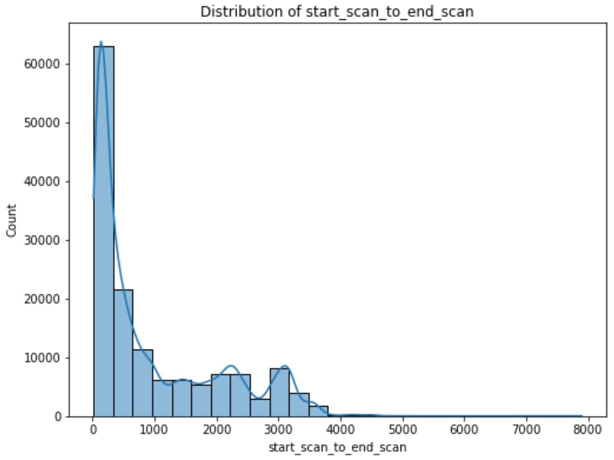
- **There are total 14817 different trips of data available**
- **There are 1508 unique source_center**
- **There are 1481 unique destination_center**
- **There are total 1504 delivery routes**

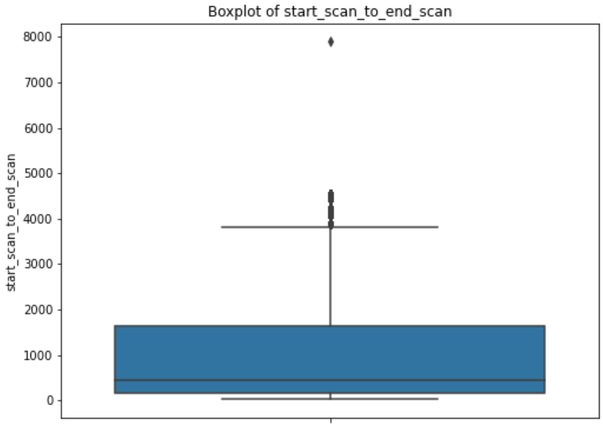
Visual Analysis

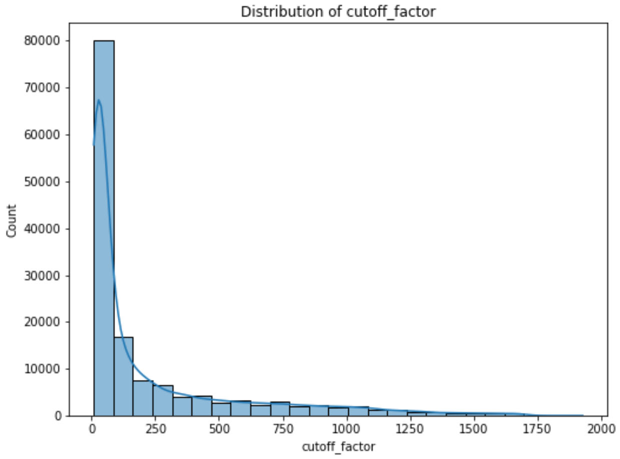
Univariate Continuous

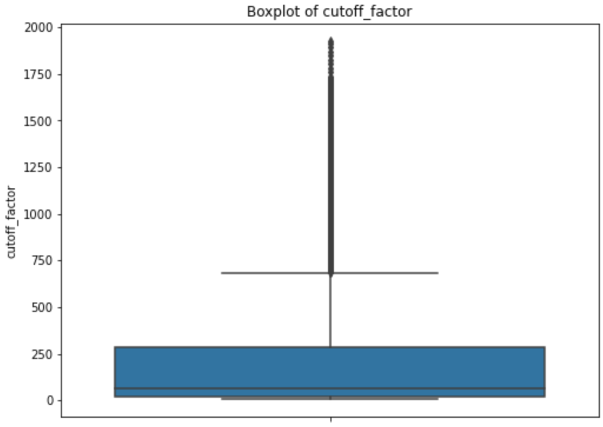
In [11]:

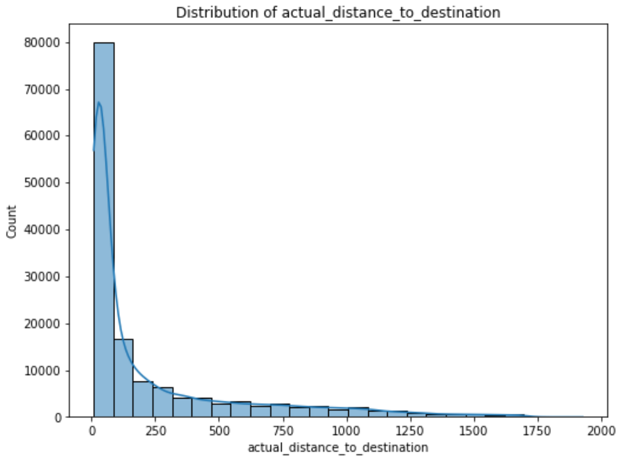
```
num_vars = delhivery.select_dtypes(include=np.number).columns.tolist()

fig, ax = plt.subplots(nrows=11, ncols=2, figsize=(18, 80))

for i in range(len(num_vars)):

    sns.histplot(x=delhivery[num_vars[i]], kde=True, bins = 25, ax=ax[i, 0])
    ax[i, 0].set_title(f"Distribution of {num_vars[i]}")

    sns.boxplot(y = delhivery[num_vars[i]], ax=ax[i, 1], data=delhivery)
    ax[i, 1].set_title(f"Boxplot of {num_vars[i]}")

plt.show()
```

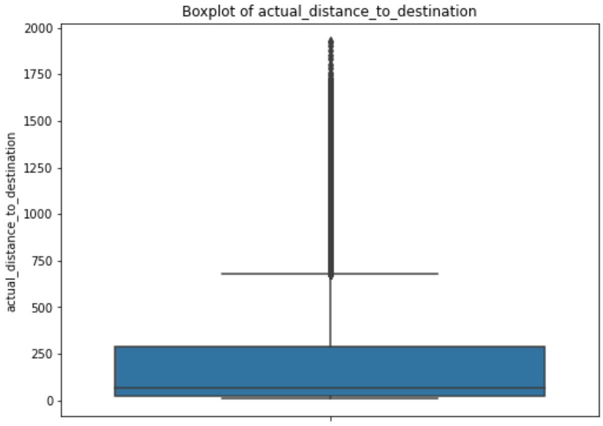

Distribution of start_scan_to_end_scan
Count
start_scan_to_end_scan

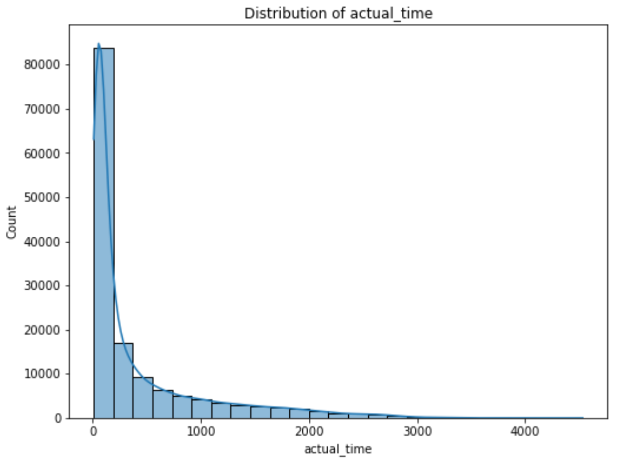

Boxplot of start_scan_to_end_scan
start_scan_to_end_scan

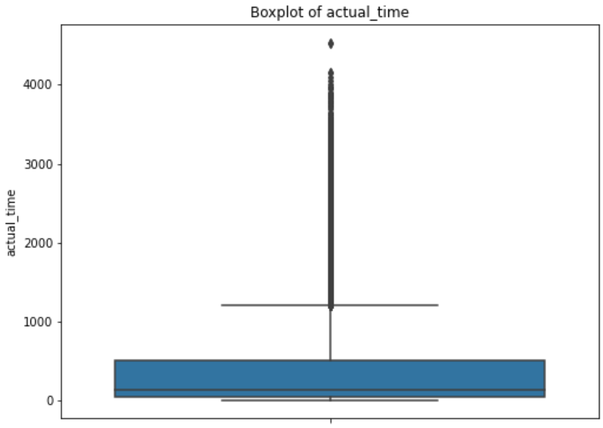

Distribution of cutoff_factor
Count
cutoff_factor

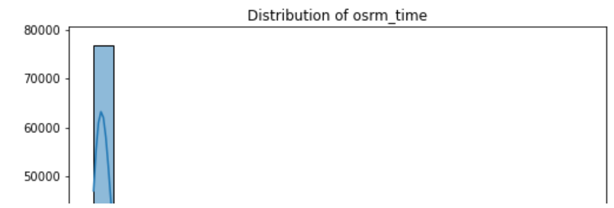

Boxplot of cutoff_factor
cutoff_factor

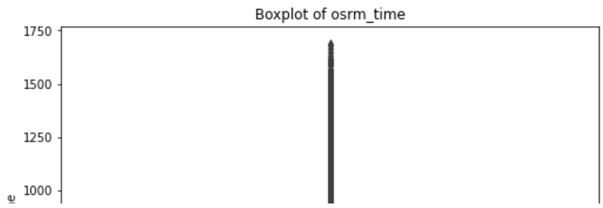

Distribution of actual_distance_to_destination
Count
actual_distance_to_destination


Boxplot of actual_distance_to_destination
actual_distance_to_destination


Distribution of actual_time
Count
actual_time


Boxplot of actual_time
actual_time


Distribution of osrm_time


Boxplot of osrm_time

Featu

Extra ... pin coc ... tion name

In [14]:

```
delhivery["source_city"] = delhivery["source_name"].str.split(" ",n=1,expand=True)[0].str.split("_",n=1,expand=True)[0]
delhivery["source_state"] = delhivery["source_name"].str.split(" ",n=1,expand=True)[1].str.replace("(","").str.replace(")","")

delhivery["destination_city"] = delhivery["destination_name"].str.split(" ",n=1,expand=True)[0].str.split("_",n=1,expand=True)[0]
delhivery["destination_state"] = delhivery["destination_name"].str.split(" ",n=1,expand=True)[1].str.replace("(","").str.replace(")","")

delhivery["source_pincode"] = delhivery["source_center"].apply(lambda x : x[3:9] )
delhivery["destination_pincode"] = delhivery["destination_center"].apply(lambda x : x[3:9] )
```

Time ... nd VS s

In [16]:

```
delhivery["time_taken_btwn_odstart_and_od_end"] = ((delhivery["od_end_time"]-delhivery["od_start_time"])/pd.Timedelta(1,unit="hour"))
```

Converting given time duration features into hours

In [17]:

```
delhivery["start_scan_to_end_scan"] = delhivery["start_scan_to_end_scan"]/60
delhivery["actual_time"] = delhivery["actual_time"]/60
delhivery["osrm_time"] = delhivery["osrm_time"]/60
delhivery["segment_actual_time"] = delhivery["segment_actual_time"]/60
delhivery["segment_osrm_time"] = delhivery["segment_osrm_time"]/60
```

In [18]:

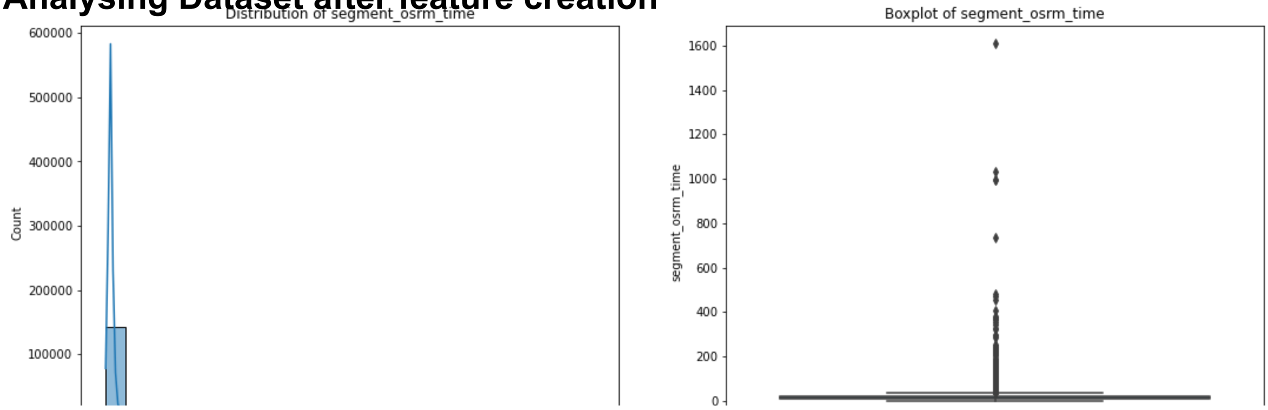
```
delhivery.head()
```

Out[18]:

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name	destination_center	destina
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_M
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_M
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_M
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_M
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_M

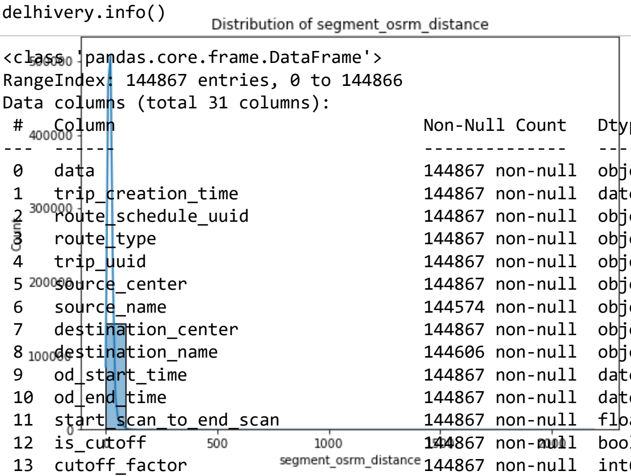
5 rows × 31 columns

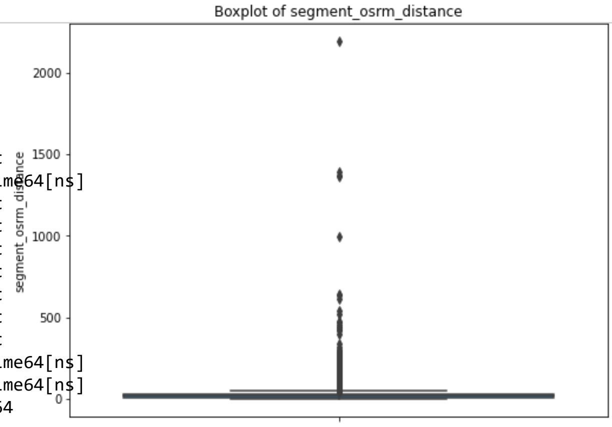
Analysing Dataset after feature creation



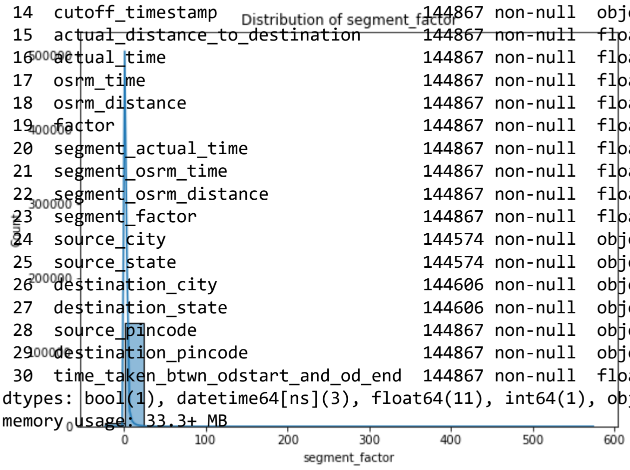
In [19]:

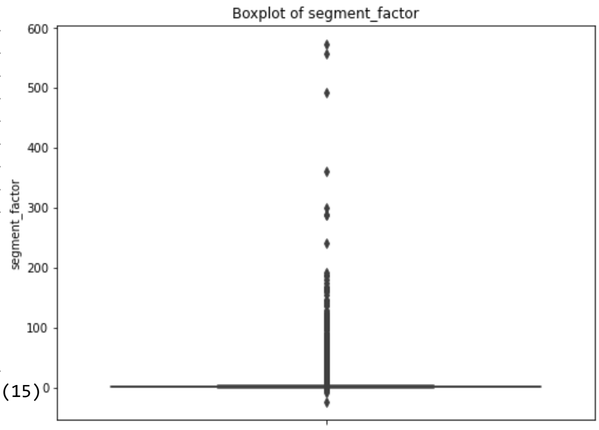
```
delhivery.info()
```






```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 144867 entries, 0 to 144866
Data columns (total 31 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   data                            144867 non-null  object
 1   trip_creation_time              144867 non-null  datetime64[ns]
 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  datetime64[ns]
 10  od_end_time                     144867 non-null  datetime64[ns]
 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
 24  source_city                     144574 non-null  object
 25  source_state                    144574 non-null  object
 26  destination_city                144606 non-null  object
 27  destination_state               144606 non-null  object
 28  source_pincode                  144867 non-null  object
 29  destination_pincode             144867 non-null  object
 30  time_taken_btwn_odstart_and_od_end  144867 non-null  float64
dtypes: bool(1), datetime64[ns](3), float64(11), int64(1), object(15)
memory usage: 33.3+ MB
```





Data Cleaning

In [20]:

```
delhivery["source_state"] = delhivery["source_state"].replace({"Goa Goa":"Goa",
                           "Layout PC Karnataka":"Karnataka",
                           "Vadgaon Sheri DPC Maharashtra":"Maharashtra",
                           "Pashan DPC Maharashtra":"Maharashtra",
                           "City Madhya Pradesh":"Madhya Pradesh",
                           "02_DPC Uttar Pradesh":"Uttar Pradesh",
                           "Nagar_DC Rajasthan":"Rajasthan",
                           "Alipore_DPC West Bengal":"West Bengal",
                            "Mandakni Madhya Pradesh":"Madhya Pradesh",
                            "West _Dc Maharashtra":"Maharashtra",
                            "DC Rajasthan":"Rajasthan",
                            "MP Nagar Madhya Pradesh":"Madhya Pradesh",
                            "Antop Hill Maharashtra":"Maharashtra",
                            "Avenue_DPC West Bengal":"West Bengal",
                            "Nagar Uttar Pradesh":"Uttar Pradesh",
                            "Balaji Nagar Maharashtra":"Maharashtra",
                            "Kothanur_L Karnataka":"Karnataka",
                            "Rahatani DPC Maharashtra":"Maharashtra",
                            "Mahim Maharashtra":"Maharashtra",
                            "DC Maharashtra":"Maharashtra",
                            "_NAD Andhra Pradesh":"Andhra Pradesh",
                                                        })

delhivery["destination_state"] = delhivery["destination_state"].replace({"Goa Goa":"Goa",
                           "Layout PC Karnataka":"Karnataka",
                           "Vadgaon Sheri DPC Maharashtra":"Maharashtra",
                           "Pashan DPC Maharashtra":"Maharashtra",
                           "City Madhya Pradesh":"Madhya Pradesh",
                           "02_DPC Uttar Pradesh":"Uttar Pradesh",
                           "Nagar_DC Rajasthan":"Rajasthan",
                           "Alipore_DPC West Bengal":"West Bengal",
                            "Mandakni Madhya Pradesh":"Madhya Pradesh",
                            "West _Dc Maharashtra":"Maharashtra",
                            "DC Rajasthan":"Rajasthan",
                            "MP Nagar Madhya Pradesh":"Madhya Pradesh",
                            "Antop Hill Maharashtra":"Maharashtra",
                            "Avenue_DPC West Bengal":"West Bengal",
                            "Nagar Uttar Pradesh":"Uttar Pradesh",
                            "Balaji Nagar Maharashtra":"Maharashtra",
                            "Kothanur_L Karnataka":"Karnataka",
                            "Rahatani DPC Maharashtra":"Maharashtra",
                            "Mahim Maharashtra":"Maharashtra",
                            "DC Maharashtra":"Maharashtra",
                            "_NAD Andhra Pradesh":"Andhra Pradesh",
                           "Delhi Delhi":"Delhi",
                           "West_Dc Maharashtra":"Maharashtra",
                           "Hub Maharashtra":"Maharashtra"
                                                        })
```

In [21]:

```
delhivery["destination_city"].replace({
    "del":"Delhi"
},inplace=True)
delhivery["source_city"].replace({
    "del":"Delhi"
},inplace=True)

delhivery["source_city"].replace({
    "Bangalore":"Bengaluru"
        },inplace=True)
delhivery["destination_city"].replace({
    "Bangalore":"Bengaluru"
        },inplace=True)
delhivery["destination_city"].replace({
    "AMD":"Ahmedabad"
        },inplace=True)
delhivery["destination_city"].replace({
    "Amdavad":"Ahmedabad"
        },inplace=True)
delhivery["source_city"].replace({
    "AMD":"Ahmedabad"
        },inplace=True)
delhivery["source_city"].replace({
    "Amdavad":"Ahmedabad"
        },inplace=True)
```

Creating Feature - (Source city + state & Destination city + state)

In [22]:

```
delhivery["source_city_state"] = delhivery["source_city"] + " " + delhivery["source_state"]
delhivery["destination_city_state"] = delhivery["destination_city"] + " " + delhivery["destination_state"]
```

In [23]:

```
delhivery["source_city_state"].nunique()
```

Out[23]:

```
1249
```

In [24]:

```
delhivery["destination_city_state"].nunique()
```

Out[24]:

```
1242
```

In [25]:

```
delhivery["source_state"].nunique()
```

Out[25]:

```
33
```

In [26]:

```
delhivery["destination_state"].nunique()
```

Out[26]:

```
32
```

- **Company delivers in approximately all the states and cities of India**

Dropping Unnecessary columns

In [27]:

```
data = delhivery.copy()
```

In [28]:

```
data.shape
```

Out[28]:

```
(144867, 33)
```

In [29]:

```
data.drop(
    ['source_center',"source_name","destination_center","destination_name","cutoff_timestamp", "od_end_time","od_start_time"],
    axis = 1,
    inplace=True
)
```

In [30]:

```
data.shape
```

Out[30]:

```
(144867, 26)
```

3. Merging of rows and aggregation of fields

In [31]:

```
actual_time = data.groupby(["trip_uuid",
           "start_scan_to_end_scan"])["actual_time"].max().reset_index().groupby("trip_uuid")["actual_time"].sum().reset_index()

actual_time
```

Out[31]:

	trip_uuid	actual_time
0	trip-153671041653548748	26.033333
1	trip-153671042288605164	2.383333
2	trip-153671043369099517	55.783333
3	trip-153671046011330457	0.983333
4	trip-153671052974046625	5.683333
...	...	...
14812	trip-153861095625827784	1.383333
14813	trip-153861104386292051	0.350000
14814	trip-153861106442901555	4.700000
14815	trip-153861115439069069	4.400000
14816	trip-153861118270144424	4.583333

14817 rows × 2 columns

In [32]:

```
segment_osrm_time = data[["trip_uuid","segment_osrm_time"]].groupby("trip_uuid")["segment_osrm_time"].sum().reset_index()
segment_osrm_time
```

Out[32]:

	trip_uuid	segment_osrm_time
0	trip-153671041653548748	16.800000
1	trip-153671042288605164	1.083333
2	trip-153671043369099517	32.350000
3	trip-153671046011330457	0.266667
4	trip-153671052974046625	1.916667
...	...	...
14812	trip-153861095625827784	1.033333
14813	trip-153861104386292051	0.183333
14814	trip-153861106442901555	1.466667
14815	trip-153861115439069069	3.683333
14816	trip-153861118270144424	1.116667

14817 rows × 2 columns

In [33]:

```
segment_actual_time = data.groupby("trip_uuid")["segment_actual_time"].sum().reset_index()
segment_actual_time
```

Out[33]:

	trip_uuid	segment_actual_time
0	trip-153671041653548748	25.800000
1	trip-153671042288605164	2.350000
2	trip-153671043369099517	55.133333
3	trip-153671046011330457	0.983333
4	trip-153671052974046625	5.666667
...	...	...
14812	trip-153861095625827784	1.366667
14813	trip-153861104386292051	0.350000
14814	trip-153861106442901555	4.683333
14815	trip-153861115439069069	4.300000
14816	trip-153861118270144424	4.566667

14817 rows × 2 columns

In [34]:

```
osrm_time = data.groupby(["trip_uuid",
          "start_scan_to_end_scan"])["osrm_time"].max().reset_index().groupby("trip_uuid")["osrm_time"].sum().reset_index()
osrm_time
```

Out[34]:

	trip_uuid	osrm_time
0	trip-153671041653548748	12.383333
1	trip-153671042288605164	1.133333
2	trip-153671043369099517	29.016667
3	trip-153671046011330457	0.250000
4	trip-153671052974046625	1.950000
...	...	...
14812	trip-153861095625827784	1.033333
14813	trip-153861104386292051	0.200000
14814	trip-153861106442901555	0.900000
14815	trip-153861115439069069	3.066667
14816	trip-153861118270144424	1.133333

14817 rows × 2 columns

In [35]:

```
time_taken_btwn_odstart_and_od_end = data.groupby("trip_uuid")["time_taken_btwn_odstart_and_od_end"].unique().reset_index()
time_taken_btwn_odstart_and_od_end
```

Out[35]:

	trip_uuid	time_taken_btwn_odstart_and_od_end
0	trip-153671041653548748	[16.65842298, 21.0100736875]
1	trip-153671042288605164	[2.0463247669444447, 0.9805397955555556]
2	trip-153671043369099517	[51.662059856388886, 13.910648811388889]
3	trip-153671046011330457	[1.6749155866666667]
4	trip-153671052974046625	[2.5335485744444446, 1.3423885633333332, 8.096...
...	...	...
14812	trip-153861095625827784	[2.546464057777778, 1.7540180775]
14813	trip-153861104386292051	[1.0098420219444444]
14814	trip-153861106442901555	[2.895179575833333, 4.1401515375]
14815	trip-153861115439069069	[1.7609491794444445, 0.7362400538888889, 1.035...
14816	trip-153861118270144424	[1.1155594141666667, 4.7912334425]

14817 rows × 2 columns

In [36]:

```
time_taken_btwn_odstart_and_od_end["time_taken_btwn_odstart_and_od_end"] = time_taken_btwn_odstart_and_od_end["time_taken_btwn_odstart_anc
time_taken_btwn_odstart_and_od_end["time_taken_btwn_odstart_and_od_end"]
```

Out[36]:

```
0        37.668497
1         3.026865
2        65.572709
3         1.674916
4        11.972484
           ...
14812     4.300482
14813     1.009842
14814     7.035331
14815     5.808548
14816     5.906793
Name: time_taken_btwn_odstart_and_od_end, Length: 14817, dtype: float64
```

In [37]:

```
start_scan_to_end_scan = ((data.groupby("trip_uuid")["start_scan_to_end_scan"].unique())).reset_index()
start_scan_to_end_scan
```

Out[37]:

	trip_uuid	start_scan_to_end_scan
0	trip-153671041653548748	[16.65, 21.0]
1	trip-153671042288605164	[2.033333333333333, 0.9666666666666667]
2	trip-153671043369099517	[51.65, 13.9]
3	trip-153671046011330457	[1.6666666666666667]
4	trip-153671052974046625	[2.533333333333333, 1.3333333333333333, 8.0833...
...	...	...
14812	trip-153861095625827784	[2.533333333333333, 1.75]
14813	trip-153861104386292051	[1.0]
14814	trip-153861106442901555	[2.8833333333333333, 4.133333333333334]
14815	trip-153861115439069069	[1.75, 0.7333333333333333, 1.0333333333333334,...
14816	trip-153861118270144424	[1.1, 4.783333333333333]

14817 rows × 2 columns

In [38]:

```
start_scan_to_end_scan["start_scan_to_end_scan"] = start_scan_to_end_scan["start_scan_to_end_scan"].apply(sum)
start_scan_to_end_scan["start_scan_to_end_scan"]
```

Out[38]:

```
0        37.650000
1         3.000000
2        65.550000
3         1.666667
4        11.950000
           ...
14812     4.283333
14813     1.000000
14814     7.016667
14815     5.783333
14816     5.883333
Name: start_scan_to_end_scan, Length: 14817, dtype: float64
```

In [39]:

```
osrm_distance = data.groupby(["trip_uuid",
              "start_scan_to_end_scan"])["osrm_distance"].max().reset_index().groupby("trip_uuid")["osrm_distance"].sum().reset_index()

osrm_distance
```

Out[39]:

	trip_uuid	osrm_distance
0	trip-153671041653548748	991.3523
1	trip-153671042288605164	85.1110
2	trip-153671043369099517	2372.0852
3	trip-153671046011330457	19.6800
4	trip-153671052974046625	146.7918
...	...	...
14812	trip-153861095625827784	73.4630
14813	trip-153861104386292051	16.0882
14814	trip-153861106442901555	63.2841
14815	trip-153861115439069069	177.6635
14816	trip-153861118270144424	80.5787

14817 rows × 2 columns

In [40]:

```
actual_distance_to_destination = data.groupby(["trip_uuid",
              "start_scan_to_end_scan"])["actual_distance_to_destination"].max().reset_index().groupby("trip_uuid")["actual_distance_to_de

actual_distance_to_destination
```

Out[40]:

	trip_uuid	actual_distance_to_destination
0	trip-153671041653548748	824.732854
1	trip-153671042288605164	73.186911
2	trip-153671043369099517	1932.273969
3	trip-153671046011330457	17.175274
4	trip-153671052974046625	127.448500
...	...	...
14812	trip-153861095625827784	57.762332
14813	trip-153861104386292051	15.513784
14814	trip-153861106442901555	38.684839
14815	trip-153861115439069069	134.723836
14816	trip-153861118270144424	66.081533

14817 rows × 2 columns

In [41]:

```
segment_osrm_distance = data[["trip_uuid",
                              "segment_osrm_distance"]].groupby("trip_uuid")["segment_osrm_distance"].sum().reset_index()

segment_osrm_distance
```

Out[41]:

	trip_uuid	segment_osrm_distance
0	trip-153671041653548748	1320.4733
1	trip-153671042288605164	84.1894
2	trip-153671043369099517	2545.2678
3	trip-153671046011330457	19.8766
4	trip-153671052974046625	146.7919
...	...	...
14812	trip-153861095625827784	64.8551
14813	trip-153861104386292051	16.0883
14814	trip-153861106442901555	104.8866
14815	trip-153861115439069069	223.5324
14816	trip-153861118270144424	80.5787

14817 rows × 2 columns

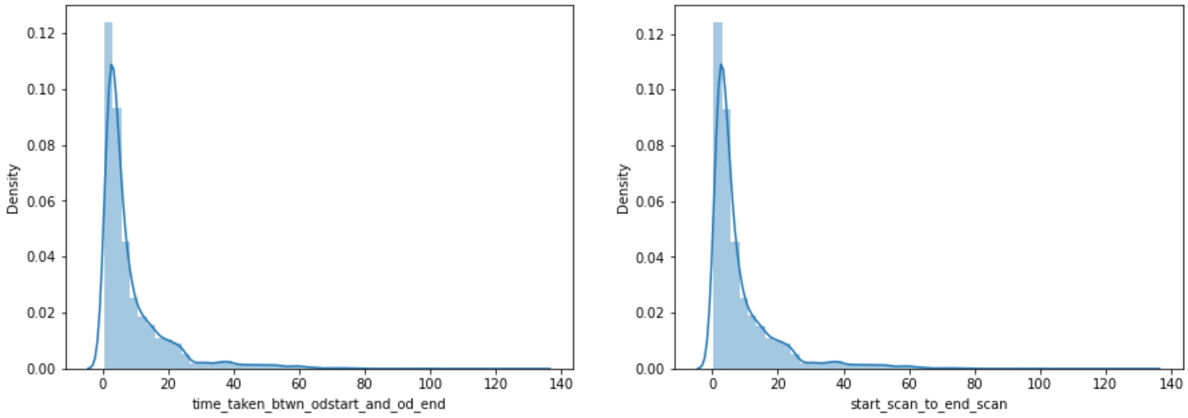
Hypothesis Test

1. Analysing TimeTaken Between OdStart and OdEnd time & StartScanToEndScan :

```
H0: Mean of time taken betweenn trip end ans start time = Mean of start and end scan time
Ha: Mean of time taken betweenn trip end ans start time != Mean of start and end scan time
```

In [42]:

```
plt.figure(figsize=(15,5))
plt.subplot(121)
sns.distplot((time_taken_btwn_odstart_and_od_end["time_taken_btwn_odstart_and_od_end"]))
plt.subplot(122)
sns.distplot((start_scan_to_end_scan["start_scan_to_end_scan"]))

plt.show()
```



In [47]:

```
# KS Test to check the similarity of distribution of these two.
```

In [48]:

```
ks_test, p_value = stats.ks_2samp(time_taken_btwn_odstart_and_od_end["time_taken_btwn_odstart_and_od_end"]
               ,start_scan_to_end_scan["start_scan_to_end_scan"])
```

In [50]:

```
# Ho: The distribution are similar
# Ha: The disbutions are different

if p_value < 0.05:
    print("Reject Ho: The distribution are different.")
else :
    print("Fail to reject Ho: The distribution is same.")
```

```
Fail to reject Ho: The distribution is same.
```

In [44]:

```
for i in range(5):
    print(stats.ttest_ind((time_taken_btwn_odstart_and_od_end["time_taken_btwn_odstart_and_od_end"].sample(3000))
                ,(start_scan_to_end_scan["start_scan_to_end_scan"].sample(3000))))
```

```
Ttest_indResult(statistic=0.6714239945778661, pvalue=0.5019763203283931)
Ttest_indResult(statistic=0.41848264577226674, pvalue=0.6756092552179287)
Ttest_indResult(statistic=0.6447318181807175, pvalue=0.519125650157082)
Ttest_indResult(statistic=0.9823631331057817, pvalue=0.3259606046963722)
Ttest_indResult(statistic=0.464823972038283, pvalue=0.6420743922558231)
```

- **from 2 sample t-test ,we can also conclude that Average time_taken_btwn_odstart_and_od_end for population is also equal to Average start_scan_to_end_scan for population.**

In [45]:

```
rt_and_od_end["time_taken_btwn_odstart_and_od_end"].mean(),time_taken_btwn_odstart_and_od_end["time_taken_btwn_odstart_and_od_end"].std()
```

Out[45]:

```
(8.861857235305067, 10.981665759990623)
```

In [46]:

```
start_scan_to_end_scan["start_scan_to_end_scan"].mean(),start_scan_to_end_scan["start_scan_to_end_scan"].std()
```

Out[46]:

```
(8.835777597804325, 10.97628639143973)
```

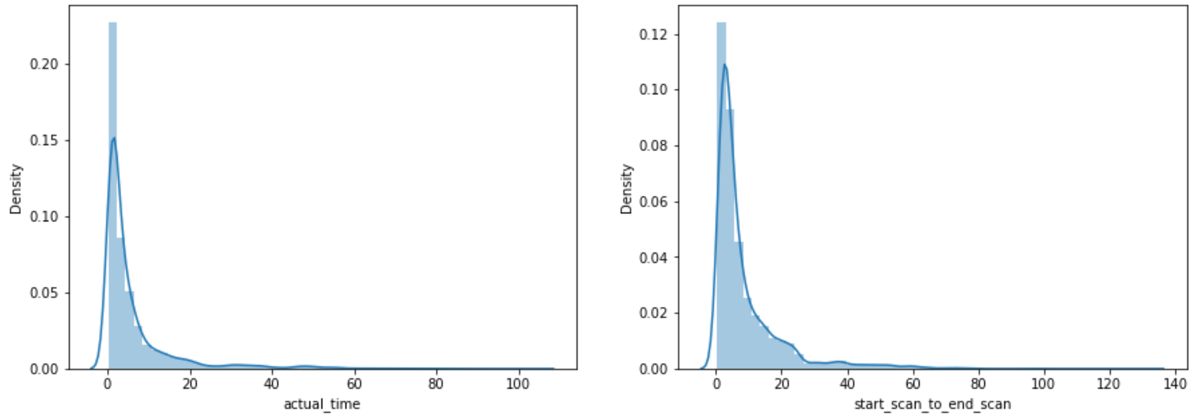
- **variance and means both are closly similar for scan time and trip start and end time taken**

2. Analysing Actual Time taken to complete the delivery & start-scan-end-scan

```
H0: Mean of start and end scan time <= Mean of Actual time taken to complete delivery
Ha: Mean of start and end scan time > Mean of Actual time taken to complete delivery
```

In [51]:

```
plt.figure(figsize=(15,5))
plt.subplot(121)
sns.distplot((actual_time["actual_time"]))
plt.subplot(122)
sns.distplot((start_scan_to_end_scan["start_scan_to_end_scan"]))

plt.show()
```



In [52]:

```
stats.ks_2samp(actual_time["actual_time"],start_scan_to_end_scan["start_scan_to_end_scan"])
```

Out[52]:

```
KstestResult(statistic=0.27387460349598436, pvalue=0.0)
```

In [53]:

```
for i in range(7):
    print(stats.ttest_ind((actual_time["actual_time"].sample(3000))
                ,(start_scan_to_end_scan["start_scan_to_end_scan"].sample(3000)),alternative="less"))
```

```
Ttest_indResult(statistic=-11.48588040898402, pvalue=1.6050039198262936e-30)
Ttest_indResult(statistic=-12.167439991452746, pvalue=5.761438744463318e-34)
Ttest_indResult(statistic=-9.612125347131766, pvalue=5.092064712029623e-22)
Ttest_indResult(statistic=-11.06992305366733, pvalue=1.6441121808032186e-28)
Ttest_indResult(statistic=-10.678067493515393, pvalue=1.1111779058337194e-26)
Ttest_indResult(statistic=-10.683768838802184, pvalue=1.0461814155703182e-26)
Ttest_indResult(statistic=-9.969258593721714, pvalue=1.5738483102437443e-23)
```

In [54]:

```
actual_time["actual_time"].mean(),actual_time["actual_time"].std()
```

Out[54]:

```
(5.945176711435117, 9.35554782297388)
```

In [55]:

```
start_scan_to_end_scan["start_scan_to_end_scan"].mean(),start_scan_to_end_scan["start_scan_to_end_scan"].std()
```

Out[55]:

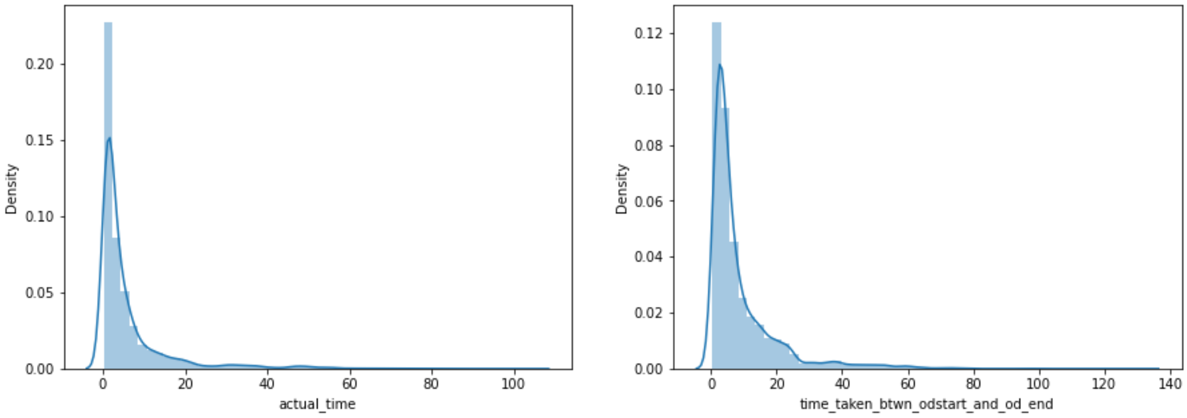
```
(8.835777597804325, 10.97628639143973)
```

Analysing Actual Time & TimeTaken between start and end trip time.

```
H0: Mean of Actual time taken to complete delivery = Mean of time taken betweenn trip end and start time
Ha: Mean of Actual time taken to complete delivery != Mean of time taken betweenn trip end and start time
```

In [57]:

```
plt.figure(figsize=(15,5))
plt.subplot(121)
sns.distplot((actual_time["actual_time"]))
plt.subplot(122)
sns.distplot((time_taken_btwn_odstart_and_od_end["time_taken_btwn_odstart_and_od_end"]))

plt.show()
```



In [58]:

```
stats.ks_2samp(actual_time["actual_time"],time_taken_btwn_odstart_and_od_end["time_taken_btwn_odstart_and_od_end"])
```

Out[58]:

```
KstestResult(statistic=0.2765067152594992, pvalue=0.0)
```

In [59]:

```
for i in range(5):
    print(stats.ttest_ind((actual_time["actual_time"].sample(1000))
                ,(time_taken_btwn_odstart_and_od_end["time_taken_btwn_odstart_and_od_end"].sample(1000))))
```

```
Ttest_indResult(statistic=-6.786852150438458, pvalue=1.5054397246445133e-11)
Ttest_indResult(statistic=-7.715022157505155, pvalue=1.895629127743348e-14)
Ttest_indResult(statistic=-7.094301439066613, pvalue=1.7976235814594463e-12)
Ttest_indResult(statistic=-7.892239905891693, pvalue=4.848911672436884e-15)
Ttest_indResult(statistic=-6.829469810483725, pvalue=1.1270937534469917e-11)
```

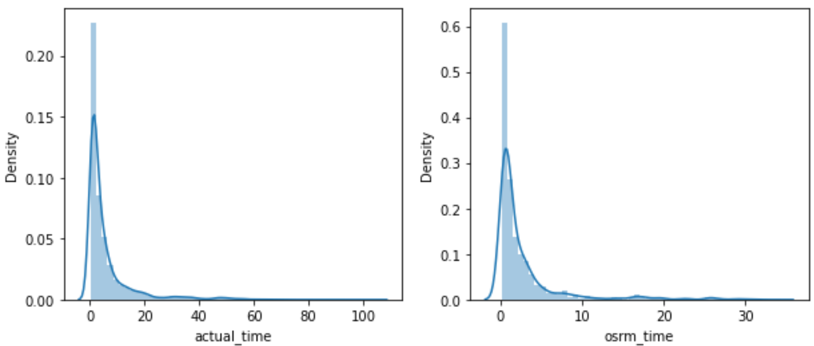
- **from above kstest of distribution and two sample ttest ,**
- *we can conclude that population mean Actual time taken to complete delivery and population mean time_taken_btwn_od_start_and_od_end are also not same.* *

Analysing Actual Time taken to complete delivery from source to destination hub & OSRM measured time :

H0: Mean of OSRM time >= Mean of Actual time taken to complete delivery
Ha: Mean of OSRM time < Mean of Actual time taken to complete delivery

In [61]:

```
plt.figure(figsize=(10,4))
plt.subplot(121)
sns.distplot(((actual_time["actual_time"])))
plt.subplot(122)
sns.distplot(((osrm_time["osrm_time"])))

plt.show()
```



In [62]:

```
stats.ks_2samp(actual_time["actual_time"],
               osrm_time["osrm_time"])
```

Out[62]:

```
KstestResult(statistic=0.2945265573327934, pvalue=0.0)
```

In [63]:

```
for i in range(5):
    print(stats.ttest_ind(actual_time["actual_time"].sample(5000),
               osrm_time["osrm_time"].sample(5000),alternative='greater'))
```

```
Ttest_indResult(statistic=20.874844797587166, pvalue=4.6741525943075676e-95)
Ttest_indResult(statistic=21.06263659333233, pvalue=1.0663252663857793e-96)
Ttest_indResult(statistic=21.90231605667887, pvalue=3.338340966942265e-104)
Ttest_indResult(statistic=22.626137040820716, pvalue=6.95064157206658e-111)
Ttest_indResult(statistic=22.372750465385117, pvalue=1.5968850909928022e-108)
```

- **from two sample ttest can conclude , that population mean actual time taken to complete delivert from source to warehouse and orsm estimate mean time for population are not same.**
- **actual time is higher than the osrm estimated time for delivery.**

In [64]:

```
actual_time["actual_time"].mean(),actual_time["actual_time"].std()
```

Out[64]:

```
(5.945176711435117, 9.35554782297388)
```

In [65]:

```
osrm_time["osrm_time"].mean(),osrm_time["osrm_time"].std()
```

Out[65]:

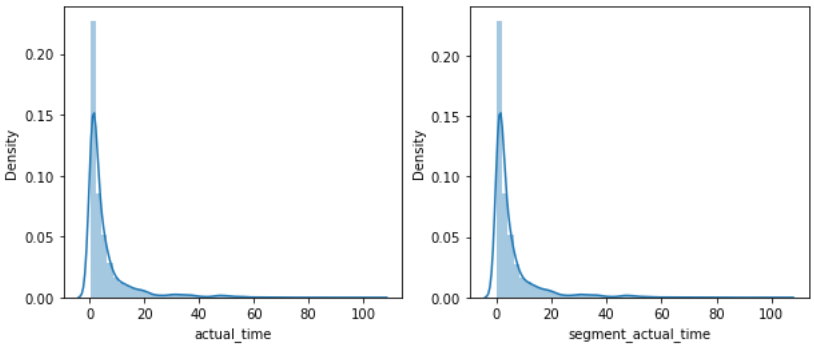
```
(2.697313896200314, 4.537654251845703)
```

Analysing Actual Time taken to complete delivery from source to destination hub & Segment Actual Time :

H0: Actual time = segment actual time
Ha: Actual time != segment actual time

In [66]:

```
plt.figure(figsize=(10,4))
plt.subplot(121)
sns.distplot(((actual_time["actual_time"])))
plt.subplot(122)
sns.distplot(((segment_actual_time["segment_actual_time"])))

plt.show()
```



In [67]:

```
for i in range(7):
    print(stats.ttest_ind((actual_time["actual_time"].sample(3000)),
                (segment_actual_time["segment_actual_time"].sample(3000))))
```

```
Ttest_indResult(statistic=-0.016715959547720472, pvalue=0.9866637710387933)
Ttest_indResult(statistic=0.5871854005027938, pvalue=0.5571012979004615)
Ttest_indResult(statistic=-0.02670849159278006, pvalue=0.9786931287610291)
Ttest_indResult(statistic=1.075612000857682, pvalue=0.2821440815292437)
Ttest_indResult(statistic=0.6217036800005118, pvalue=0.5341603600193086)
Ttest_indResult(statistic=-0.234716811580539, pvalue=0.8144365503619873)
Ttest_indResult(statistic=0.11828520674704442, pvalue=0.9058456392873808)
```

from two sample ttest , we can conclude that

- Population average for
- Actual Time taken to complete delivery trip and segment actual time are same.

In [68]:

```
actual_time["actual_time"].mean(),actual_time["actual_time"].std()
```

Out[68]:

```
(5.945176711435117, 9.35554782297388)
```

In [69]:

```
segment_actual_time["segment_actual_time"].mean(),segment_actual_time["segment_actual_time"].std()
```

Out[69]:

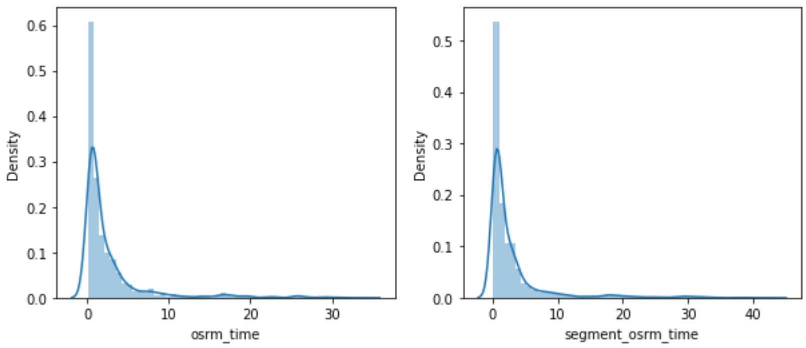
```
(5.898204764797215, 9.270799413152762)
```

Analysing osrm Time & segment-osrm-time :

Ho: segment actual time <= OSRM time
Ha: segment actual time > OSRM time

In [70]:

```
plt.figure(figsize=(10,4))
plt.subplot(121)
sns.distplot(((osrm_time["osrm_time"])))
plt.subplot(122)
sns.distplot(((segment_osrm_time["segment_osrm_time"])))

plt.show()
```



In [71]:

```
for i in range(7):
    print(stats.ttest_ind((osrm_time["osrm_time"].sample(3000)),
                (segment_osrm_time["segment_osrm_time"].sample(3000)),alternative ="less"))
```

```
Ttest_indResult(statistic=-2.5714643684061778, pvalue=0.005075408499762505)
Ttest_indResult(statistic=-1.5691908804660726, pvalue=0.05832810884813956)
Ttest_indResult(statistic=-2.9217546638780645, pvalue=0.0017468230961662475)
Ttest_indResult(statistic=-3.498475612576067, pvalue=0.00023566059986077686)
Ttest_indResult(statistic=-1.0738001172175655, pvalue=0.14147773706130126)
Ttest_indResult(statistic=-2.144958486986873, pvalue=0.015998105644098412)
Ttest_indResult(statistic=-1.4332422913226777, pvalue=0.07592036418712263)
```

from ttest , we can conclude that

- average of osrm Time & segment-osrm-time for population is not same.
- Population Mean osrm time is less than Population Mean segment osrm time.

In [72]:

```
osrm_time["osrm_time"].mean(),osrm_time["osrm_time"].std()
```

Out[72]:

```
(2.697313896200314, 4.537654251845703)
```

In [73]:

```
segment_osrm_time["segment_osrm_time"].mean(),segment_osrm_time["segment_osrm_time"].std()
```

Out[73]:

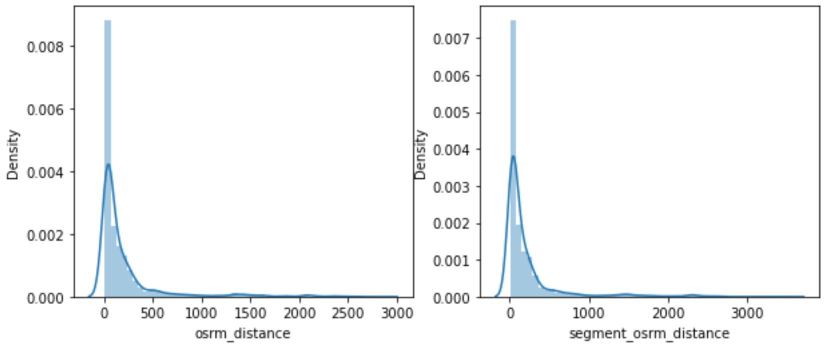
```
(3.0158297901059705, 5.242367441693007)
```

Analysing and Visulizing OSRM Estimated distance and Segment-osrm-distance :

H0 : Segment OSRM distnace <= OSRM distnace
Ha : Segment OSRM distnace > OSRM distnace

In [74]:

```
plt.figure(figsize=(10,4))
plt.subplot(121)
sns.distplot(((osrm_distance["osrm_distance"])))
plt.subplot(122)
sns.distplot(((segment_osrm_distance["segment_osrm_distance"])))

plt.show()
```



In [75]:

```
stats.ks_2samp(osrm_distance["osrm_distance"],segment_osrm_distance["segment_osrm_distance"])
```

Out[75]:

```
KstestResult(statistic=0.03948167645272321, pvalue=1.8042208791084262e-10)
```

In [76]:

```
for i in range(7):
    print(stats.ttest_ind(osrm_distance["osrm_distance"].sample(5000),
                segment_osrm_distance["segment_osrm_distance"].sample(5000),alternative="less"))
```

```
Ttest_indResult(statistic=-3.1205975445015204, pvalue=0.0009049920177902538)
Ttest_indResult(statistic=-2.96098883209724, pvalue=0.0015368676978257379)
Ttest_indResult(statistic=-1.8806924246314503, pvalue=0.030021405926531658)
Ttest_indResult(statistic=-1.8618084413928004, pvalue=0.03132970787636542)
Ttest_indResult(statistic=-2.943881956240802, pvalue=0.0016243461003502275)
Ttest_indResult(statistic=-2.3820303805154133, pvalue=0.008618033390676348)
Ttest_indResult(statistic=-1.4715344413611824, pvalue=0.070589056920015)
```

In [77]:

```
osrm_distance["osrm_distance"].mean(),osrm_distance["osrm_distance"].std()
```

Out[77]:

```
(204.83672531551625, 370.74927471335496)
```

In [78]:

```
segment_osrm_distance["segment_osrm_distance"].mean(),segment_osrm_distance["segment_osrm_distance"].std()
```

Out[78]:

```
(223.20116128771042, 416.6283742907418)
```

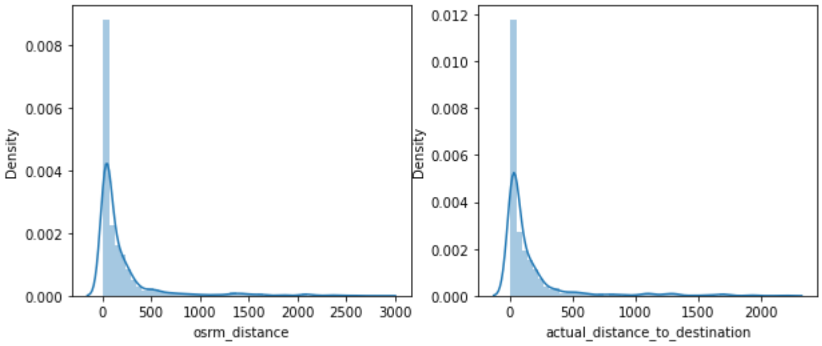
- **from KS test , we can conclude the distributions of segment osrm distance and osrm distnace are not same!**
- **from two sample one sided ttest, we can conclude: Average of osrm distance for population is less than average of segment osrm distance**

Analysing and Visulizing OSRM Estimated distance and Actual Distance between source and destination warehouse :

H0 : Mean OSRM distance <= Mean Actual distnace
Ha : Mean OSRM distance > Mean Actual distnace

In [79]:

```
plt.figure(figsize=(10,4))
plt.subplot(121)
sns.distplot(((osrm_distance["osrm_distance"])))
plt.subplot(122)
sns.distplot(((actual_distance_to_destination["actual_distance_to_destination"])))

plt.show()
```



In [80]:

```
stats.ks_2samp(osrm_distance["osrm_distance"],actual_distance_to_destination["actual_distance_to_destination"])
```

Out[80]:

```
KstestResult(statistic=0.11837753931295136, pvalue=6.578385372142345e-91)
```

In [81]:

```
for i in range(5):
    print(stats.ttest_ind(osrm_distance["osrm_distance"].sample(5000),
               actual_distance_to_destination["actual_distance_to_destination"].sample(5000),alternative="greater"))
```

```
Ttest_indResult(statistic=6.553978031969843, pvalue=2.939322740895829e-11)
Ttest_indResult(statistic=6.750578503769857, pvalue=7.77153615930502e-12)
Ttest_indResult(statistic=4.992393449114795, pvalue=3.0320962145937554e-07)
Ttest_indResult(statistic=6.852074692811769, pvalue=3.8538345486306455e-12)
Ttest_indResult(statistic=5.7440946985473245, pvalue=4.755662267561689e-09)
```

From left sided ttest , we can conclude

- for population OSRM estimated distance is higher than the actual distance from source to destination warehouse.

In [82]:

```
osrm_distance["osrm_distance"].mean(),osrm_distance["osrm_distance"].std()
```

Out[82]:

```
(204.83672531551625, 370.74927471335496)
```

In [83]:

```
actual_distance_to_destination["actual_distance_to_destination"].mean(),actual_distance_to_destination["actual_distance_to_destination"].s
```

Out[83]:

```
(164.4733217454422, 305.5408288910492)
```

Merging

In [84]:

```
distances = segment_osrm_distance.merge(actual_distance_to_destination.merge(osrm_distance,
                                                                              on="trip_uuid"),
                                                                             on="trip_uuid")
```

In [85]:

```
distances
```

Out[85]:

	trip_uuid	segment_osrm_distance	actual_distance_to_destination	osrm_distance
0	trip-153671041653548748	1320.4733	824.732854	991.3523
1	trip-153671042288605164	84.1894	73.186911	85.1110
2	trip-153671043369099517	2545.2678	1932.273969	2372.0852
3	trip-153671046011330457	19.8766	17.175274	19.6800
4	trip-153671052974046625	146.7919	127.448500	146.7918
...	...	...	...	...
14812	trip-153861095625827784	64.8551	57.762332	73.4630
14813	trip-153861104386292051	16.0883	15.513784	16.0882
14814	trip-153861106442901555	104.8866	38.684839	63.2841
14815	trip-153861115439069069	223.5324	134.723836	177.6635
14816	trip-153861118270144424	80.5787	66.081533	80.5787

14817 rows × 4 columns

In [86]:

```
time = segment_osrm_time.merge(osrm_time.merge(segment_actual_time.merge(actual_time.merge(time_taken_btwn_odstart_and_od_end.merge(start_
                                          on="trip_uuid",
                                          ),on="trip_uuid"),on="trip_uuid"),on="trip_uuid"),on="trip_uuid")

time
```

Out[86]:

	trip_uuid	segment_osrm_time	osrm_time	segment_actual_time	actual_time	time_taken_btwn_odstart_and_od_end	start_scan_to_end_sc
0	trip-153671041653548748	16.800000	12.383333	25.800000	26.033333	37.668497	37.650C
1	trip-153671042288605164	1.083333	1.133333	2.350000	2.383333	3.026865	3.000C
2	trip-153671043369099517	32.350000	29.016667	55.133333	55.783333	65.572709	65.550C
3	trip-153671046011330457	0.266667	0.250000	0.983333	0.983333	1.674916	1.666E
4	trip-153671052974046625	1.916667	1.950000	5.666667	5.683333	11.972484	11.950C
...	...	...	...	...	...	...	
14812	trip-153861095625827784	1.033333	1.033333	1.366667	1.383333	4.300482	4.283:
14813	trip-153861104386292051	0.183333	0.200000	0.350000	0.350000	1.009842	1.000C
14814	trip-153861106442901555	1.466667	0.900000	4.683333	4.700000	7.035331	7.016E
14815	trip-153861115439069069	3.683333	3.066667	4.300000	4.400000	5.808548	5.783:
14816	trip-153861118270144424	1.116667	1.133333	4.566667	4.583333	5.906793	5.883:

14817 rows × 7 columns

In [87]:

```
Merge1 = time.merge(distances,on="trip_uuid",
                                        )
Merge1
```

Out[87]:

	trip_uuid	segment_osrm_time	osrm_time	segment_actual_time	actual_time	time_taken_btwn_odstart_and_od_end	start_scan_to_end_sc
0	trip-153671041653548748	16.800000	12.383333	25.800000	26.033333	37.668497	37.650C
1	trip-153671042288605164	1.083333	1.133333	2.350000	2.383333	3.026865	3.000C
2	trip-153671043369099517	32.350000	29.016667	55.133333	55.783333	65.572709	65.550C
3	trip-153671046011330457	0.266667	0.250000	0.983333	0.983333	1.674916	1.6666
4	trip-153671052974046625	1.916667	1.950000	5.666667	5.683333	11.972484	11.950C
...	...	...	...	...	...	...	
14812	trip-153861095625827784	1.033333	1.033333	1.366667	1.383333	4.300482	4.2833
14813	trip-153861104386292051	0.183333	0.200000	0.350000	0.350000	1.009842	1.000C
14814	trip-153861106442901555	1.466667	0.900000	4.683333	4.700000	7.035331	7.0166
14815	trip-153861115439069069	3.683333	3.066667	4.300000	4.400000	5.808548	5.7833
14816	trip-153861118270144424	1.116667	1.133333	4.566667	4.583333	5.906793	5.8833

14817 rows × 10 columns

Merging Location details and route_type and Numerical data on TripID :

In [88]:

```
city = data.groupby("trip_uuid")[["source_city",
                                  "destination_city"]].aggregate({
        "source_city":pd.unique,
    "destination_city":pd.unique,
})

state = data.groupby("trip_uuid")[["source_state",
                                   "destination_state"]].aggregate({
        "source_state":pd.unique,
    "destination_state":pd.unique,
})

city_state = data.groupby("trip_uuid")[["source_city_state",
                                        "destination_city_state"]].aggregate({
        "source_city_state":pd.unique,
    "destination_city_state":pd.unique,
})

locations = city.merge(city_state.merge(state,on="trip_uuid"
                            ,how="outer"),
           on="trip_uuid",
           how="outer")
```

In [91]:

```
route_type = data.groupby("trip_uuid")["route_type"].unique().reset_index()

Merged = route_type.merge(locations.merge(Merge1,on="trip_uuid",
           how="outer"),
                 on="trip_uuid",
           how="outer"
                )
```

In [92]:

```
trip_records = Merged.copy()
```

In [93]:

```
trip_records["route_type"] = trip_records["route_type"].apply(lambda x:x[0])
route_to_merge = data.groupby("trip_uuid")["route_schedule_uuid"].unique().reset_index()
trip_records = trip_records.merge(route_to_merge,on="trip_uuid",how="outer")
trip_records["route_schedule_uuid"] = trip_records["route_schedule_uuid"].apply(lambda x:x[0])
trip_records
```

Out[93]:

	trip_uuid	route_type	source_city	destination_city	source_city_state	destination_city_state	source_state	destination_state	segment_
0	trip-153671041653548748	FTL	[Bhopal, Kanpur]	[Kanpur, Gurgaon]	[Bhopal Madhya Pradesh, Kanpur Uttar Pradesh]	[Kanpur Uttar Pradesh, Gurgaon Haryana]	[Madhya Pradesh, Uttar Pradesh]	[Uttar Pradesh, Haryana]	
1	trip-153671042288605164	Carting	[Tumkur, Doddablpur]	[Doddablpur, Chikblapur]	[Tumkur Karnataka, Doddablpur Karnataka]	[Doddablpur Karnataka, Chikblapur Karnataka]	[Karnataka]	[Karnataka]	
2	trip-153671043369099517	FTL	[Bengaluru, Gurgaon]	[Gurgaon, Chandigarh]	[Bengaluru Karnataka, Gurgaon Haryana]	[Gurgaon Haryana, Chandigarh Punjab]	[Karnataka, Haryana]	[Haryana, Punjab]	
3	trip-153671046011330457	Carting	[Mumbai]	[Mumbai]	[Mumbai Hub Maharashtra]	[Mumbai Maharashtra]	[Hub Maharashtra]	[Maharashtra]	
4	trip-153671052974046625	FTL	[Bellary, Hospet, Sandur]	[Hospet, Sandur, Bellary]	[Bellary Karnataka, Hospet Karnataka, Sandur K...	[Hospet Karnataka, Sandur Karnataka, Bellary K...	[Karnataka]	[Karnataka]	
...	...	...	...	...	...	...	...	...	
14812	trip-153861095625827784	Carting	[Chandigarh]	[Zirakpur, Chandigarh]	[Chandigarh Punjab, Chandigarh Chandigarh]	[Zirakpur Punjab, Chandigarh Punjab]	[Punjab, Chandigarh]	[Punjab]	
14813	trip-153861104386292051	Carting	[FBD]	[Faridabad]	[FBD Haryana]	[Faridabad Haryana]	[Haryana]	[Haryana]	
14814	trip-153861106442901555	Carting	[Kanpur]	[Kanpur]	[Kanpur Uttar Pradesh]	[Kanpur Uttar Pradesh]	[Uttar Pradesh]	[Uttar Pradesh]	
14815	trip-153861115439069069	Carting	[Tirunelveli, Eral, Tirchchndr, Thisayanvilai,...	[Eral, Tirchchndr, Thisayanvilai, Peikulam, Ti...	[Tirunelveli Tamil Nadu, Eral Tamil Nadu, Tirc...	[Eral Tamil Nadu, Tirchchndr Tamil Nadu, Thisa...	[Tamil Nadu]	[Tamil Nadu]	
14816	trip-153861118270144424	FTL	[Hospet, Sandur]	[Sandur, Bellary]	[Hospet Karnataka, Sandur Karnataka]	[Sandur Karnataka, Bellary Karnataka]	[Karnataka]	[Karnataka]	

14817 rows × 18 columns

In [94]:

```
trip_records.isna().sum()
```

Out[94]:

```
trip_uuid                         0
route_type                        0
source_city                       0
destination_city                  0
source_city_state                 0
destination_city_state            0
source_state                      0
destination_state                 0
segment_osrm_time                 0
osrm_time                         0
segment_actual_time               0
actual_time                       0
time_taken_btwn_odstart_and_od_end    0
start_scan_to_end_scan            0
segment_osrm_distance             0
actual_distance_to_destination    0
osrm_distance                     0
route_schedule_uuid               0
dtype: int64
```

Unnesting Data

In [95]:

```
trip_records["source_city"] = trip_records["source_city"].astype("str").str.strip("[]").str.replace("'","")
trip_records["destination_city"] = trip_records["destination_city"].astype("str").str.strip("[]").str.replace("'","")
trip_records["source_city_state"] = trip_records["source_city_state"].astype("str").str.strip("[]").str.replace("'","")
trip_records["destination_city_state"] = trip_records["destination_city_state"].astype("str").str.strip("[]").str.replace("'","")

trip_records["source_state"] = trip_records["source_state"].astype("str").str.strip("[]").str.replace("'","")
trip_records["destination_state"] = trip_records["destination_state"].astype("str").str.strip("[]").str.replace("'","")
```

Statistically Analysis

In [96]:

```
trip_records.corr()
```

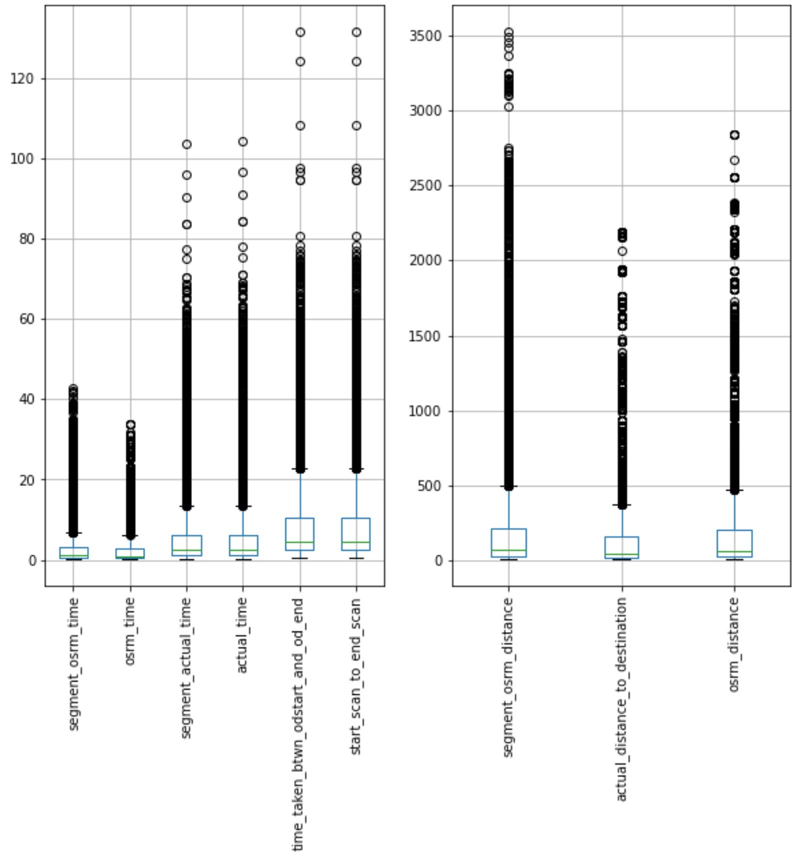
Out[96]:

	segment_osrm_time	osrm_time	segment_actual_time	actual_time	time_taken_btwn_odstart_and_od_end	start_scan_t
segment_osrm_time	1.000000	0.993508	0.953039	0.953800	0.918447	
osrm_time	0.993508	1.000000	0.957747	0.958613	0.926280	
segment_actual_time	0.953039	0.957747	1.000000	0.999920	0.961096	
actual_time	0.953800	0.958613	0.999920	1.000000	0.960958	
time_taken_btwn_odstart_and_od_end	0.918447	0.926280	0.961096	0.960958	1.000000	
start_scan_to_end_scan	0.918493	0.926469	0.961107	0.961163	0.999860	
segment_osrm_distance	0.996092	0.991848	0.956106	0.956949	0.919156	
actual_distance_to_destination	0.987627	0.993556	0.953048	0.954082	0.918373	
osrm_distance	0.992050	0.997610	0.958341	0.959290	0.924093	

Detecting Outliers

In [97]:

```
plt.figure(figsize = (10,8))
plt.subplot(121)
trip_records[['segment_osrm_time', 'osrm_time',
       'segment_actual_time', 'actual_time',
       'time_taken_btwn_odstart_and_od_end', 'start_scan_to_end_scan']].boxplot()
plt.xticks(rotation =90)
plt.subplot(122)
trip_records[['segment_osrm_distance', 'actual_distance_to_destination',
       'osrm_distance']].boxplot()
plt.xticks(rotation =90)
plt.show()
```



In [98]:

```
outlier_treatment  = trip_records.copy()
```

In [99]:

```
outlier_treatment_num = outlier_treatment[['segment_osrm_time', 'osrm_time',
       'segment_actual_time', 'actual_time',
       'time_taken_btwn_odstart_and_od_end', 'start_scan_to_end_scan',
        'segment_osrm_distance', 'actual_distance_to_destination',
        'osrm_distance']]
```

Treating Outliers

In [100]:

```
trip_records_without_outliers = trip_records.loc[outlier_treatment_num[(np.abs(stats.zscore(outlier_treatment_num)) < 3).all(axis=1)].inde
trip_records_without_outliers
```

Out[100]:

	trip_uuid	route_type	source_city	destination_city	source_city_state	destination_city_state	source_state	destination_state	segment_o
0	trip-153671041653548748	FTL	Bhopal Kanpur	Kanpur Gurgaon	Bhopal Madhya Pradesh Kanpur Uttar Pradesh	Kanpur Uttar Pradesh Gurgaon Haryana	Madhya Pradesh Uttar Pradesh	Uttar Pradesh Haryana	1
1	trip-153671042288605164	Carting	Tumkur Doddablpur	Doddablpur Chikblapur	Tumkur Karnataka Doddablpur Karnataka	Doddablpur Karnataka Chikblapur Karnataka	Karnataka	Karnataka	
3	trip-153671046011330457	Carting	Mumbai	Mumbai	Mumbai Hub Maharashtra	Mumbai Maharashtra	Hub Maharashtra	Maharashtra	
4	trip-153671052974046625	FTL	Bellary Hospet Sandur	Hospet Sandur Bellary	Bellary Karnataka Hospet Karnataka Sandur Karn...	Hospet Karnataka Sandur Karnataka Bellary Karn...	Karnataka	Karnataka	
5	trip-153671055416136166	Carting	Chennai	Chennai	Chennai Tamil Nadu	Chennai Tamil Nadu	Tamil Nadu	Tamil Nadu	
...	...	...	...	...	...	...	...	...	
14812	trip-153861095625827784	Carting	Chandigarh	Zirakpur Chandigarh	Chandigarh Punjab Chandigarh Chandigarh	Zirakpur Punjab Chandigarh Punjab	Punjab Chandigarh	Punjab	
14813	trip-153861104386292051	Carting	FBD	Faridabad	FBD Haryana	Faridabad Haryana	Haryana	Haryana	
14814	trip-153861106442901555	Carting	Kanpur	Kanpur	Kanpur Uttar Pradesh	Kanpur Uttar Pradesh	Uttar Pradesh	Uttar Pradesh	
14815	trip-153861115439069069	Carting	Tirunelveli Eral Tirchchndr Thisayanvilai Peik...	Eral Tirchchndr Thisayanvilai Peikulam Tirunel...	Tirunelveli Tamil Nadu Eral Tamil Nadu Tirchch...	Eral Tamil Nadu Tirchchndr Tamil Nadu Thisayan...	Tamil Nadu	Tamil Nadu	
14816	trip-153861118270144424	FTL	Hospet Sandur	Sandur Bellary	Hospet Karnataka Sandur Karnataka	Sandur Karnataka Bellary Karnataka	Karnataka	Karnataka	

14160 rows × 18 columns

Processing Data for One hot encoding :

merging locations details into one columns . and re categorise the data as per highest trips having location as top category

In [101]:

```
trip_records_without_outliers["destination_source_locations"] = trip_records_without_outliers["source_city_state"]+" "+trip_records_withou
trip_records_without_outliers.drop(["source_city_state","destination_city_state"],axis = 1,inplace=True)
```

In [102]:

```
sc_dc = trip_records_without_outliers.groupby(["destination_source_locations"])["trip_uuid"].nunique().sort_values(ascending= False).reset
```

In [103]:

```
def get_cat(H):
    if 0 <= H <= 50:
        return "Category 7"
    elif 51 <= H <= 100:
        return "Category 6"
    elif 101 <= H <= 200:
        return "Category 5"
    elif 201 <= H <= 300:
        return "Category 4"
    elif 301 <= H <= 400:
        return "Category 3"
    elif 401 <= H <= 500:
        return "Category 2"
    else:
        return "Category 1"
```

In [104]:

```
sc_dc["city"]  = pd.Series(map(get_cat,sc_dc["trip_uuid"]))
trip_records_for_encoding = sc_dc.merge(trip_records_without_outliers,
            on="destination_source_locations")
trip_records_for_encoding.drop(["destination_source_locations","trip_uuid_x"],axis = 1,inplace=True)
trip_records_for_encoding.drop(["trip_uuid_y"],axis = 1,inplace=True)
# trip_records_for_encoding.sample(15)
encoded_data = pd.get_dummies(trip_records_for_encoding,
             columns=["route_type","city"] )
encoded_data
```

Out[104]:

	source_city	destination_city	source_state	destination_state	segment_osrm_time	osrm_time	segment_actual_time	actual_time	time_taken_btwn_c
0	Bengaluru	Bengaluru	Karnataka	Karnataka	1.383333	0.950000	3.183333	3.233333	
1	Bengaluru	Bengaluru	Karnataka	Karnataka	1.150000	0.883333	2.666667	2.700000	
2	Bengaluru	Bengaluru	Karnataka	Karnataka	1.183333	0.966667	3.316667	3.333333	
3	Bengaluru	Bengaluru	Karnataka	Karnataka	0.700000	0.733333	1.316667	1.316667	
4	Bengaluru	Bengaluru	Karnataka	Karnataka	0.783333	0.666667	1.750000	1.766667	
...	...	...	...	...	...	...	...	...	
14155	Hyderabad Kadthal Kalwakurthy Devarakonda	Kadthal Kalwakurthy Devarakonda Haliya	Telangana	Telangana	1.966667	1.983333	3.233333	3.250000	
14156	Hyderabad Kadthal	Kadthal Devarakonda	Telangana	Telangana	1.483333	1.433333	2.716667	2.750000	
14157	Hyderabad Kadthal Haliya	Kadthal Kalwakurthy Hyderabad	Telangana	Telangana	2.916667	2.866667	4.950000	4.983333	
14158	Hyderabad Kadthal Haliya	Kadthal Devarakonda Hyderabad	Telangana	Telangana	3.383333	3.333333	10.950000	10.966667	
14159	nan	nan	nan	nan	0.800000	0.816667	2.116667	2.133333	

14160 rows × 23 columns

Column Standardization

In [105]:

```
['segment_osrm_time', 'osrm_time',
       'segment_actual_time', 'actual_time',
       'time_taken_btwn_odstart_and_od_end', 'start_scan_to_end_scan' ,'segment_osrm_distance', 'actual_distance_to_destination','osrm_di
```

Out[105]:

```
['segment_osrm_time',
 'osrm_time',
 'segment_actual_time',
 'actual_time',
 'time_taken_btwn_odstart_and_od_end',
 'start_scan_to_end_scan',
 'segment_osrm_distance',
 'actual_distance_to_destination',
 'osrm_distance']
```

In [106]:

```
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import MinMaxScaler
```

In [107]:

```
scaler = StandardScaler()
std_data = scaler.fit_transform(encoded_data[['segment_osrm_time',
 'osrm_time',
 'segment_actual_time',
 'actual_time',
 'time_taken_btwn_odstart_and_od_end',
 'start_scan_to_end_scan',
 'segment_osrm_distance',
 'actual_distance_to_destination',
 'osrm_distance']])
std_data = pd.DataFrame(std_data, columns=['segment_osrm_time',
 'osrm_time',
 'segment_actual_time',
 'actual_time',
 'time_taken_btwn_odstart_and_od_end',
 'start_scan_to_end_scan',
 'segment_osrm_distance',
 'actual_distance_to_destination',
 'osrm_distance'])
std_data.head()
```

Out[107]:

	segment_osrm_time	osrm_time	segment_actual_time	actual_time	time_taken_btwn_odstart_and_od_end	start_scan_to_end_scan	segment_osrm_distanc
0	-0.269133	-0.409683	-0.220225	-0.214843	-0.394178	-0.391956	-0.36274
1	-0.359785	-0.438916	-0.324535	-0.321822	-0.445632	-0.444397	-0.44886
2	-0.346835	-0.402374	-0.193306	-0.194785	-0.443566	-0.441900	-0.41613
3	-0.534615	-0.504692	-0.597087	-0.599297	-0.318061	-0.317039	-0.53654
4	-0.502239	-0.533926	-0.509601	-0.509034	-0.567441	-0.566761	-0.54929

In [108]:

```
scaler = MinMaxScaler()
MinMax_data = scaler.fit_transform(encoded_data[['segment_osrm_time','osrm_time','segment_actual_time','actual_time',
 'time_taken_btwn_odstart_and_od_end','start_scan_to_end_scan','segment_osrm_distance','actual_distance_to_destination',
 'osrm_distance']])
MinMax_data = pd.DataFrame(MinMax_data,columns=['segment_osrm_time',
 'osrm_time','segment_actual_time','actual_time','time_taken_btwn_odstart_and_od_end','start_scan_to_end_scan',
 'segment_osrm_distance','actual_distance_to_destination','osrm_distance'])
MinMax_data.head()
```

Out[108]:

	segment_osrm_time	osrm_time	segment_actual_time	actual_time	time_taken_btwn_odstart_and_od_end	start_scan_to_end_scan	segment_osrm_distanc
0	0.069369	0.059302	0.098113	0.098719	0.098792	0.098811	0.04642
1	0.056757	0.054651	0.081402	0.081644	0.090329	0.090201	0.03466
2	0.058559	0.060465	0.102426	0.101921	0.090669	0.090611	0.03913
3	0.032432	0.044186	0.037736	0.037353	0.111311	0.111111	0.02269
4	0.036937	0.039535	0.051752	0.051761	0.070296	0.070111	0.02095

In [109]:

```
std_data
```

Out[109]:

	segment_osrm_time	osrm_time	segment_actual_time	actual_time	time_taken_btwn_odstart_and_od_end	start_scan_to_end_scan	segment_osrm_di
0	-0.269133	-0.409683	-0.220225	-0.214843	-0.394178	-0.391956	-0.3
1	-0.359785	-0.438916	-0.324535	-0.321822	-0.445632	-0.444397	-0.4
2	-0.346835	-0.402374	-0.193306	-0.194785	-0.443566	-0.441900	-0.4
3	-0.534615	-0.504692	-0.597087	-0.599297	-0.318061	-0.317039	-0.5
4	-0.502239	-0.533926	-0.509601	-0.509034	-0.567441	-0.566761	-0.5
...	...	...	...	...	...	...	
14155	-0.042502	0.043440	-0.210131	-0.211500	-0.123651	-0.124754	0.1
14156	-0.230282	-0.197738	-0.314441	-0.311792	-0.211977	-0.212156	-0.1
14157	0.326583	0.430787	0.136448	0.136179	0.104495	0.104990	0.3
14158	0.507888	0.635424	1.347789	1.336342	1.031740	1.033953	0.6
14159	-0.495764	-0.468150	-0.435575	-0.435486	-0.732338	-0.731577	-0.4

14160 rows × 9 columns

In [110]:

```
one_hot_encoded_data = encoded_data[["route_type_Carting","route_type_FTL","city_Category 1",
 "city_Category 2","city_Category 3","city_Category 4",
 "city_Category 5","city_Category 6","city_Category 7"]]
```

In [111]:

```
Standardized_Data = pd.concat([std_data,one_hot_encoded_data],axis = 1)
```

In [112]:

```
Min_Max_Scaled_Data = pd.concat([MinMax_data,one_hot_encoded_data],axis = 1)
```

In [113]:

```
Standardized_Data.sample(5)
```

Out[113]:

	segment_osrm_time	osrm_time	segment_actual_time	actual_time	time_taken_btwn_odstart_and_od_end	start_scan_to_end_scan	segment_osrm_di
3158	-0.586416	-0.585085	-0.647560	-0.642757	-0.816804	-0.816482	-0.5
8084	0.423711	0.576956	1.283857	1.279510	2.751270	2.757029	0.4
9733	2.379213	2.535617	2.202458	2.215570	2.255802	2.260084	2.5
6334	-0.690019	-0.694712	-0.772059	-0.773136	-0.921148	-0.918867	-0.6
11059	0.119378	0.065366	0.018678	0.022515	-0.064600	-0.062324	0.1

In [114]:

```
Min_Max_Scaled_Data.sample(5)
```

Out[114]:

	segment_osrm_time	osrm_time	segment_actual_time	actual_time	time_taken_btwn_odstart_and_od_end	start_scan_to_end_scan	segment_osrm_di
13440	0.045946	0.061628	0.063612	0.064568	0.058724	0.058631	0.0
11801	0.210811	0.269767	0.346631	0.345251	0.325108	0.324313	0.2
2648	0.008108	0.010465	0.012399	0.012273	0.030192	0.030340	0.0
14036	0.303604	0.382558	0.370350	0.368730	0.416553	0.416154	0.3
1846	0.004505	0.006977	0.010782	0.010672	0.023550	0.023370	0.0

Route analysis :

In [115]:

```
A = data.groupby("route_schedule_uuid")["route_type"].unique().reset_index()
B = data.groupby("route_schedule_uuid")["destination_city"].unique().reset_index()
B.columns = ["route_schedule_uuid","destination_cities"]
C = data.groupby("route_schedule_uuid")["source_city"].unique().reset_index()
C.columns = ["route_schedule_uuid","source_cities"]
D = data.groupby("route_schedule_uuid")["source_state"].unique().reset_index()
D.columns = ["route_schedule_uuid","source_states"]
E = data.groupby("route_schedule_uuid")["destination_state"].unique().reset_index()
E.columns = ["route_schedule_uuid","destination_states"]
F = data.groupby("route_schedule_uuid")[["source_state",
                                         "destination_state"]].nunique().sort_values(by="source_state",
                                                                                     ascending=False).reset_index()
F.columns = ["route_schedule_uuid","#source_states"
             ,"#destination_states"]
G = trip_records.groupby("route_schedule_uuid")["actual_distance_to_destination"].mean().reset_index()
G.columns = ["route_schedule_uuid","Average_Actual_distance_to_destination"]
H = trip_records["route_schedule_uuid"].value_counts().reset_index()
H.columns = ["route_schedule_uuid","Number_of_Trips"]
I = data.groupby("route_schedule_uuid")[["source_city",
                                         "destination_city"]].nunique().sort_values(by="source_city",
                                                                                    ascending=False).reset_index()
I.columns = ["route_schedule_uuid","#source_cities"
             ,"#destination_cities"]
```

In [116]:

```
route_records = I.merge(H.merge(G.merge(F.merge(E.merge(D.merge(C.merge(A.merge(B,
        on ="route_schedule_uuid",
        how = "outer"),on ="route_schedule_uuid",
        how = "outer"),
       on ="route_schedule_uuid",
        how = "outer"),
       on ="route_schedule_uuid",
        how = "outer"),
       on ="route_schedule_uuid",
        how = "outer"),
       on ="route_schedule_uuid",
        how = "outer"),
       on ="route_schedule_uuid",
        how = "outer"),on ="route_schedule_uuid",
        how = "outer")
```

In [117]:

```
route_records.isna().sum()
```

Out[117]:

```
route_schedule_uuid                    0
#source_cities                         0
#destination_cities                    0
Number_of_Trips                        0
Average_Actual_distance_to_destination    0
#source_states                         0
#destination_states                    0
destination_states                     0
source_states                          0
source_cities                          0
route_type                             0
destination_cities                     0
dtype: int64
```

In [118]:

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

In [119]:

```

route_records["route_type"] = route_records["route_type"].astype("str").str.strip("[]").str.replace("'","")
route_records["source_cities"] = route_records["source_cities"].astype("str").str.strip("[]").str.replace("'","")
route_records["destination_cities"] = route_records["destination_cities"].astype("str").str.strip("[]").str.replace("'","")
route_records["source_states"] = route_records["source_states"].astype("str").str.strip("[]").str.replace("'","")

route_records["destination_states"] = route_records["destination_states"].astype("str").str.strip("[]").str.replace("'","")
```

In [120]:

```
route_records
```

Out[120]:

	route_schedule_uuid	#source_cities	#destination_cities	Number_of_Trips	Average_Actual_distance_to_destination	#source_states	#destination_st
0	thanos::sroute:d010efca-d90d-4977-b987-eae68c5...	13	11	14	281.596486	2	
1	thanos::sroute:4cbecb35-356b-4b68-bf3c-6225b5e...	10	10	12	332.602225	2	
2	thanos::sroute:ae5c430f-6153-48d1-8fe5-d5f0bbc...	10	10	20	351.611796	1	
3	thanos::sroute:f8968c72-5222-4d81-9eed-8a6d88f...	9	9	9	195.257193	1	
4	thanos::sroute:ed5b80be-7abf-424d-b8cd-d81556a...	9	8	20	178.737233	1	
...	...	...	...	...	...	...	
1499	thanos::sroute:9e7bb811-593f-47bc-ac49-ba03ed8...	1	1	19	17.617532	1	
1500	thanos::sroute:46b9641b-55b5-4b15-b039-2612a50...	1	1	15	10.137219	1	
1501	thanos::sroute:b48f633d-15cb-4744-a0b9-21df0a9...	1	1	7	15.467701	1	
1502	thanos::sroute:265efe06-3625-4fba-afee-07b5b64...	0	1	1	236.815038	0	
1503	thanos::sroute:cfb575b8-df26-48f5-8427-6f48f9d...	0	0	1	50.844665	0	

1504 rows × 12 columns

In [121]:

```
route_records["ROUTE"] = route_records["source_cities"] + " -- " + route_records["destination_cities"]
route_records.drop(["route_schedule_uuid"],axis = 1,inplace=True)
first_column = route_records.pop('ROUTE')
route_records.insert(0, 'ROUTE', first_column)
route_records["SouceToDestination_city"] = route_records["source_cities"].str.split(" ").apply(lambda x:x[0]) +" TO " +route_records["des
first_column = route_records.pop('SouceToDestination_city')
route_records.insert(0, 'SouceToDestination_city', first_column)
route_records
```

Out[121]:

	SouceToDestination_city	ROUTE	#source_cities	#destination_cities	Number_of_Trips	Average_Actual_distance_to_destination	#source_states	#
0	Guwahati TO LakhimpurN	Guwahati LakhimpurN Dhemaji Likabali Tezpur Pa...	13	11	14	281.596486	2	
1	Guwahati TO Tura	Guwahati Rangia Kokrajhar Dhubri Bilasipara Tu...	10	10	12	332.602225	2	
2	Jaipur TO Tarnau	Jaipur Chomu Reengus Sikar Bikaner Didwana Suj...	10	10	20	351.611796	1	
3	Mangalore TO Udupi	Mangalore Udupi Kundapura Bhatkal Honnavar Kum...	9	9	9	195.257193	1	
4	Ajmer TO Raipur	Ajmer Beawar Bilara Bijainagar Kekri Nasirabad...	9	8	20	178.737233	1	
...	...	...	...	...	...	...	...	
1499	Mumbai TO Mumbai	Mumbai -- Mumbai	1	1	19	17.617532	1	
1500	Mumbai TO Mumbai	Mumbai -- Mumbai	1	1	15	10.137219	1	
1501	Bengaluru TO Bengaluru	Bengaluru -- Bengaluru	1	1	7	15.467701	1	
1502	nan TO Mainpuri	nan -- Mainpuri	0	1	1	236.815038	0	
1503	nan TO nan	nan -- nan	0	0	1	50.844665	0	

1504 rows × 13 columns

Exploratory Data Analysis : (getting some insights from preprocessed data) :

Busiest Route Analysis :

Number of Trips between cities , sorted highest to lowest

Top 20 source and destination cities wihc have high freqency of trips in between .

In [122]:

```
Number_of_trips_between_cities = data.groupby(["source_city_state",
                                                "destination_city_state"])["trip_uuid"].nunique().sort_values(ascending=False).reset_index(
Number_of_trips_between_cities.head(25)
```

Out[122]:

	source_city_state	destination_city_state	trip_uuid
0	Bengaluru Karnataka	Bengaluru Karnataka	1369
1	Bhiwandi Maharashtra	Mumbai Maharashtra	512
2	Mumbai Maharashtra	Mumbai Maharashtra	361
3	Hyderabad Telangana	Hyderabad Telangana	308
4	Mumbai Maharashtra	Bhiwandi Maharashtra	282
5	Delhi Delhi	Gurgaon Haryana	248
6	Gurgaon Haryana	Delhi Delhi	237
7	Mumbai Hub Maharashtra	Mumbai Maharashtra	227
8	Chennai Tamil Nadu	Chennai Tamil Nadu	205
9	MAA Tamil Nadu	Chennai Tamil Nadu	204
10	Chennai Tamil Nadu	MAA Tamil Nadu	141
11	Bengaluru Karnataka	HBR Karnataka	133
12	Ahmedabad Gujarat	Ahmedabad Gujarat	131
13	Pune Maharashtra	PNQ Maharashtra	122
14	Jaipur Rajasthan	Jaipur Rajasthan	111
15	Delhi Delhi	Delhi Delhi	109
16	Pune Maharashtra	Bhiwandi Maharashtra	107
17	Pune Maharashtra	Pune Maharashtra	101
18	Chandigarh Chandigarh	Chandigarh Punjab	100
19	Kolkata West Bengal	CCU West Bengal	96
20	Gurgaon Haryana	Sonipat Haryana	92
21	Sonipat Haryana	Gurgaon Haryana	86
22	Chandigarh Punjab	Chandigarh Chandigarh	84
23	HBR Karnataka	Bengaluru Karnataka	79
24	Bengaluru Karnataka	BLR Karnataka	78

- **From above table, we can observe that Mumbai Maharashtra ,Delhi ,Gurgaon(Haryana),Bengaluru Karnataka ,Hyderabad Telangana,Chennai Tamil Nadu,Ahmedabad Gujarat,Pune Maharashtra,Chandigarh Chandigarh and Kolkata West Bengal are some cities have higest amount of trips happening states with in the city :**

In [123]:

```
Number_of_trips_between_cities.loc[Number_of_trips_between_cities["source_city_state"] != Number_of_trips_between_cities["destination_city
```

Out[123]:

	source_city_state	destination_city_state	trip_uuid
1	Bhiwandi Maharashtra	Mumbai Maharashtra	512
4	Mumbai Maharashtra	Bhiwandi Maharashtra	282
5	Delhi Delhi	Gurgaon Haryana	248
6	Gurgaon Haryana	Delhi Delhi	237
7	Mumbai Hub Maharashtra	Mumbai Maharashtra	227
9	MAA Tamil Nadu	Chennai Tamil Nadu	204
10	Chennai Tamil Nadu	MAA Tamil Nadu	141
11	Bengaluru Karnataka	HBR Karnataka	133
13	Pune Maharashtra	PNQ Maharashtra	122
16	Pune Maharashtra	Bhiwandi Maharashtra	107
18	Chandigarh Chandigarh	Chandigarh Punjab	100
19	Kolkata West Bengal	CCU West Bengal	96
20	Gurgaon Haryana	Sonipat Haryana	92
21	Sonipat Haryana	Gurgaon Haryana	86
22	Chandigarh Punjab	Chandigarh Chandigarh	84
23	HBR Karnataka	Bengaluru Karnataka	79
24	Bengaluru Karnataka	BLR Karnataka	78
26	Del Delhi	Gurgaon Haryana	76
27	Bhiwandi Maharashtra	Pune Maharashtra	72
28	Ludhiana Punjab	Chandigarh Punjab	71
30	Chandigarh Punjab	Gurgaon Haryana	66
31	Gurgaon Haryana	Bengaluru Karnataka	66
32	LowerParel Maharashtra	Mumbai Maharashtra	65
34	Mumbai Hub Maharashtra	Bhiwandi Maharashtra	63
35	PNQ Maharashtra	Pune Maharashtra	62

If we talk about , not having equal source and destination states , source and destination cities having higest number of trips in between are :

- delhi to gurgao
- Gurgaon,Haryana TO Bengaluru,Karnataka
- Bhiwandi/Mumbai,Maharashtra TO Pune Maharashtra
- Sonipat TO Gurgaon,Haryana
    - it is also been observed that lots of deliveries are happening to airpots
    - like : Chennai to MAA chennai international Airport , Pune to Pune Airport (PNQ),Kolkata to CCU West Bengal Kolkata International Airport , Bengluru to BLR-Bengaluru Internation Airport etc.

In [124]:

```
route_records[["ROUTE","Number_of_Trips",
               "Average_Actual_distance_to_destination",
               "#source_cities",
               "#destination_cities"]].sort_values(by="Number_of_Trips",ascending=False).head(25)
```

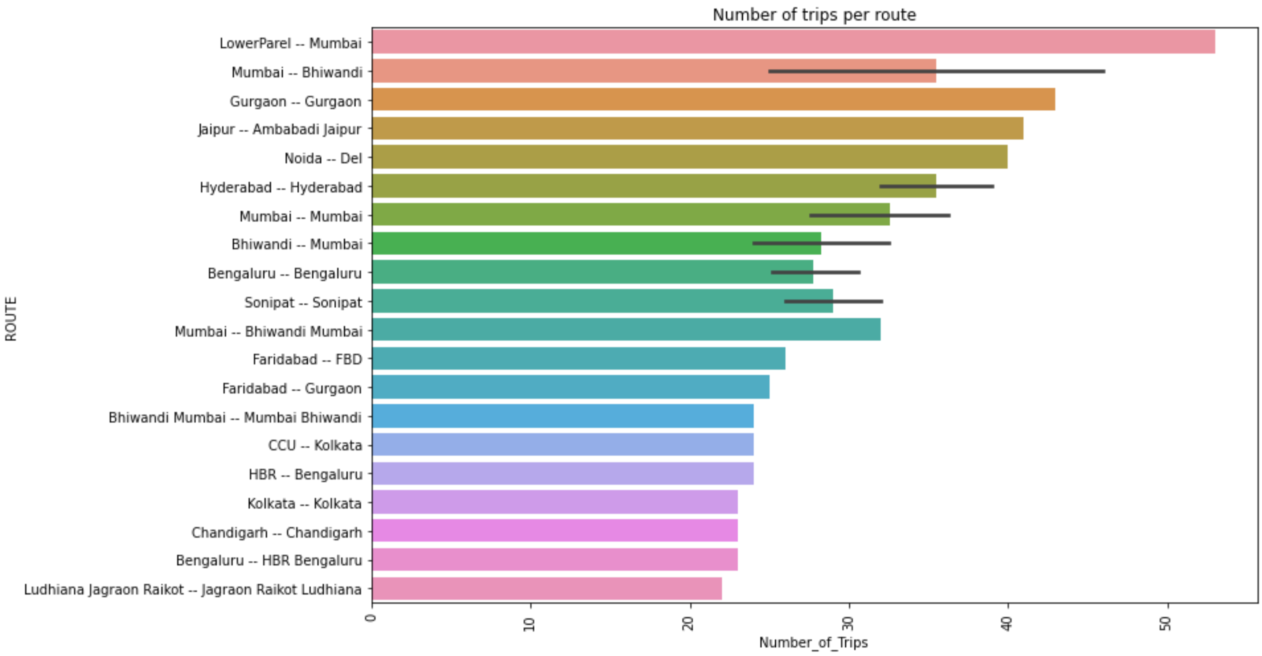
Out[124]:

	ROUTE	Number_of_Trips	Average_Actual_distance_to_destination	#source_cities	#destination_cities
1465	LowerParel -- Mumbai	53	16.428868	1	1
1426	Mumbai -- Bhiwandi	46	20.199445	1	1
808	Gurgaon -- Gurgaon	43	29.740842	1	1
679	Jaipur -- Ambabadi Jaipur	41	15.348495	1	2
1257	Noida -- Del	40	10.882902	1	1
1368	Hyderabad -- Hyderabad	39	35.695641	1	1
1273	Mumbai -- Mumbai	37	13.882863	1	1
1359	Mumbai -- Mumbai	36	17.526251	1	1
1303	Bhiwandi -- Mumbai	35	21.241534	1	1
700	Mumbai -- Mumbai	34	15.906614	1	1
751	Mumbai -- Mumbai	33	15.668726	1	1
1060	Bengaluru -- Bengaluru	33	28.067004	1	1
793	Sonipat -- Sonipat	32	11.691243	1	1
972	Hyderabad -- Hyderabad	32	21.835579	1	1
1184	Mumbai -- Bhiwandi Mumbai	32	21.601109	1	2
874	Bengaluru -- Bengaluru	30	28.055789	1	1
1177	Bhiwandi -- Mumbai	30	21.396002	1	1
1354	Bengaluru -- Bengaluru	27	27.967087	1	1
921	Faridabad -- FBD	26	9.677121	1	1
1480	Sonipat -- Sonipat	26	12.182486	1	1
1041	Mumbai -- Bhiwandi	25	19.942191	1	1
877	Faridabad -- Gurgaon	25	47.091622	1	1
833	Bhiwandi -- Mumbai	25	21.531705	1	1
1249	Bengaluru -- Bengaluru	25	28.019668	1	1
869	Bengaluru -- Bengaluru	24	41.396497	1	1

Top Routes having Maximum Number of Trips between/within the source and destinations .

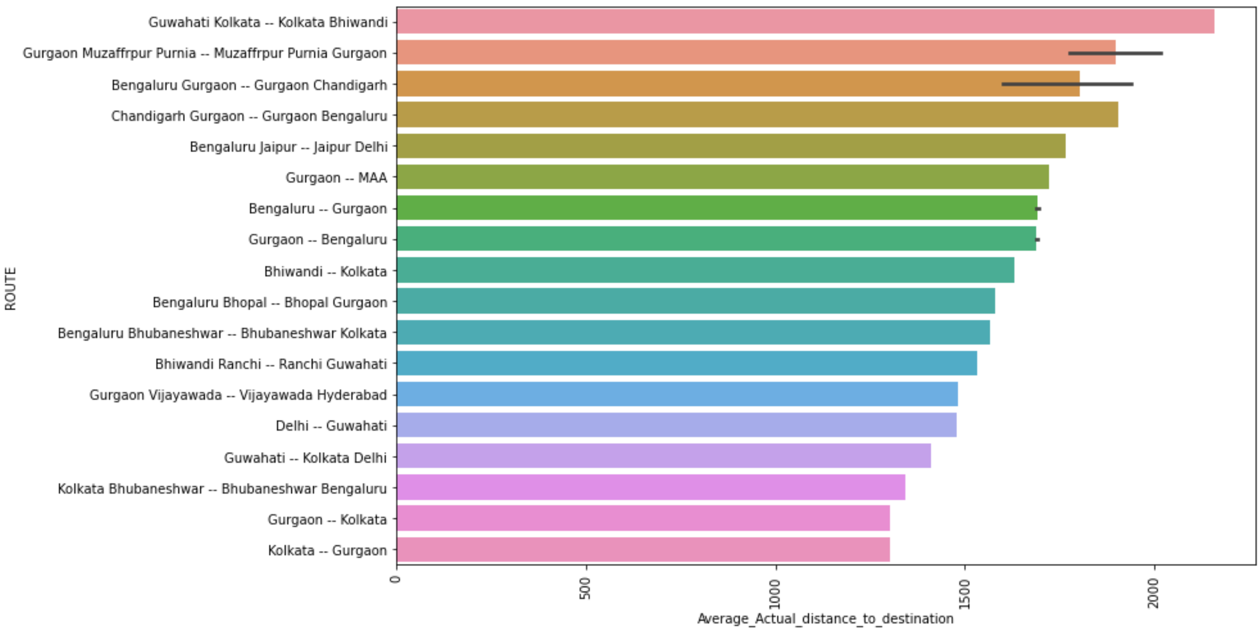
In [125]:

```
plt.figure(figsize=(12,8))

X = route_records[["ROUTE", "Number_of_Trips",
            ]].sort_values(by="Number_of_Trips",ascending=False).head(35)
sns.barplot(y = X["ROUTE"],
           x= X["Number_of_Trips"])
plt.title("Number of trips per route")
plt.xticks(rotation = 90)
plt.show()
```



In [126]:

```
plt.figure(figsize=(12,8))

X = route_records[["ROUTE", "Average_Actual_distance_to_destination",
            ]].sort_values(by="Average_Actual_distance_to_destination",ascending=False).head(25)
sns.barplot(y = X["ROUTE"],
           x = X["Average_Actual_distance_to_destination"])
plt.xticks(rotation = 90)
plt.show()
```



- From above Bar chart , and table , we can observe that higest trips are happening is with in the particular cities.
- in terms of average distnace between destinations , we can observe Guwahati to Mumbai , Benglore to Chandigarh ,Benglore to Delhi , Benglore to Gurgaon are the longest routes .

Busiest and Longest Routes :

In [127]:

```
Busiest_and_Longest_Routes  = route_records[(route_records["Average_Actual_distance_to_destination"] > route_records["Average_Actual_dista
            & (route_records["Number_of_Trips"] > route_records["Number_of_Trips"].quantile(0.75))].sort_values(by="Average_Actual_dista
                                                                                                                   ,ascending=False)
Busiest_and_Longest_Routes_top25 = Busiest_and_Longest_Routes[["source_cities",
                                                               "destination_cities",
                                                               "Number_of_Trips",
                                                               "Average_Actual_distance_to_destination"]].head(25)
Busiest_and_Longest_Routes_top25
```

Out[127]:

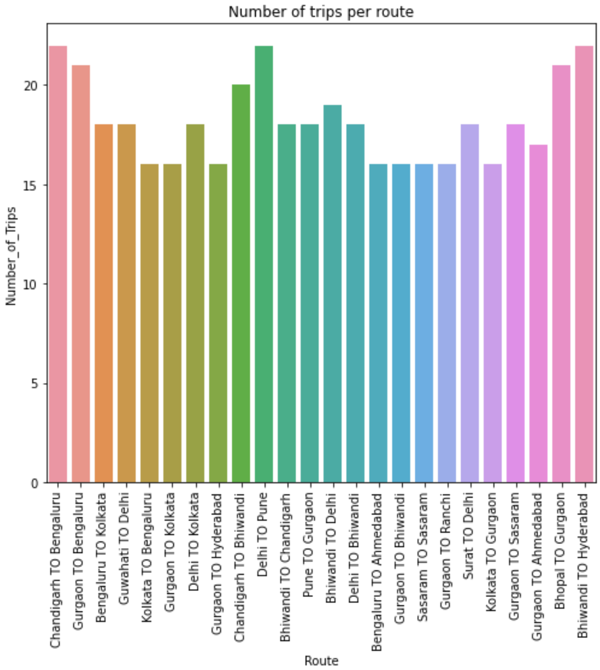
	source_cities	destination_cities	Number_of_Trips	Average_Actual_distance_to_destination
629	Chandigarh Gurgaon	Gurgaon Bengaluru	22	1905.766051
995	Gurgaon	Bengaluru	21	1689.873158
991	Gurgaon	Bengaluru	21	1689.791894
512	Bengaluru Bhubaneshwar	Bhubaneshwar Kolkata	18	1567.577507
745	Guwahati	Kolkata Delhi	18	1411.208424
624	Kolkata Bhubaneshwar	Bhubaneshwar Bengaluru	16	1342.143081
752	Gurgaon	Kolkata	16	1300.572161
588	Delhi Gurgaon	Gurgaon Kolkata	18	1263.113211
826	Gurgaon	Hyderabad	16	1236.572072
541	Chandigarh Gurgaon	Gurgaon Bhiwandi	20	1170.817927
442	Delhi Gurgaon	Gurgaon Pune	22	1151.514940
445	Bhiwandi Sonipat	Sonipat Chandigarh	18	1129.609705
739	Pune	Gurgaon	18	1120.729446
1377	Bhiwandi	Delhi	19	1114.214670
1049	Delhi	Bhiwandi	18	1114.182197
313	Bengaluru Kolhapur Surat	Kolhapur Surat Ahmedabad	16	1110.015339
1219	Gurgaon	Bhiwandi	16	1078.076312
197	Sasaram Kanpur Kolkata Dhanbad	Kanpur Gurgaon Dhanbad Sasaram	16	1028.024726
1136	Gurgaon	Ranchi	16	1010.953223
1286	Surat	Delhi	18	931.980821
439	Kolkata Ranchi	Ranchi Gurgaon	16	881.621264
1108	Gurgaon	Sasaram	18	804.210670
1454	Gurgaon	Ahmedabad	17	735.550450
223	Bhopal Kanpur Auraiya Etawah	Kanpur Auraiya Etawah Gurgaon	21	731.634456
863	Bhiwandi	Hyderabad	22	607.514619

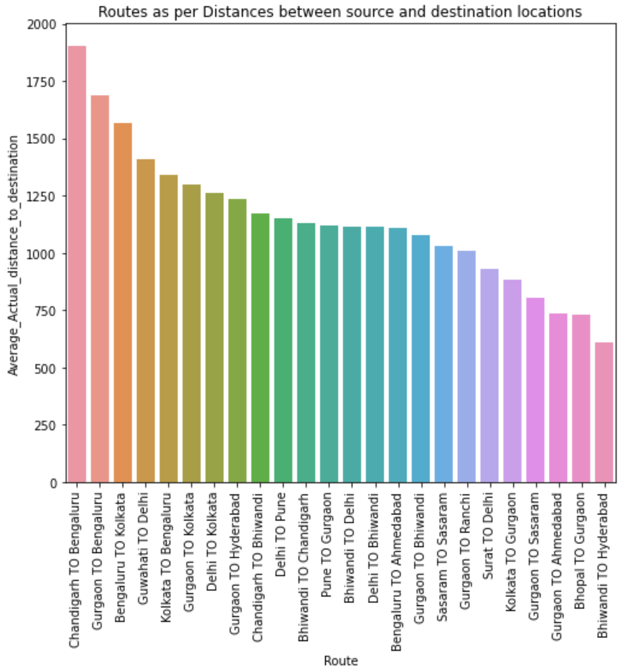
Above Table shows the souce to destination city routes having largest numbers of trip happening having large distnaces : which are :

- Chandigarh TO Bengaluru
- Gurgaon TO Bengaluru
- Bengaluru TO Kolkata
- Guwahati TO Delhi
- Delhi TO Kolkata
- Chandigarh TO Gurgaon
- Gurgaon TO Hydrabad
- Benglore TO Ahmedabad
- Surat TO Delhi
- Gurgaon TO Ahmedabad**

In [128]:

```
Busiest_and_Longest_Routes_top25["Route"] = Busiest_and_Longest_Routes_top25["source_cities"].str.split(" ").apply(lambda x:x[0]) + " TO '
Busiest_and_Longest_Routes_top25.drop(["source_cities","destination_cities"],axis = 1,inplace=True)
plt.figure(figsize=(18,7))

plt.subplot(121)
plt.title("Number of trips per route")
sns.barplot(x=Busiest_and_Longest_Routes_top25["Route"],
           y = Busiest_and_Longest_Routes_top25["Number_of_Trips"])
plt.xticks(rotation = 90)
plt.subplot(122)
plt.title("Routes as per Distances between source and destination locations")
sns.barplot(x=Busiest_and_Longest_Routes_top25["Route"],
           y= Busiest_and_Longest_Routes_top25["Average_Actual_distance_to_destination"])
plt.xticks(rotation = 90)
plt.show()
```





Routes : passing through maxinum number of cities :

In [129]:

```
route_records[["SouceToDestination_city","Number_of_Trips",
               "Average_Actual_distance_to_destination",
               "#source_cities",
               "#destination_cities"]].sort_values(by=["#source_cities",
                                                       "#destination_cities",
                                                       "Number_of_Trips"]
                                                  ,ascending=False).head(25)
```

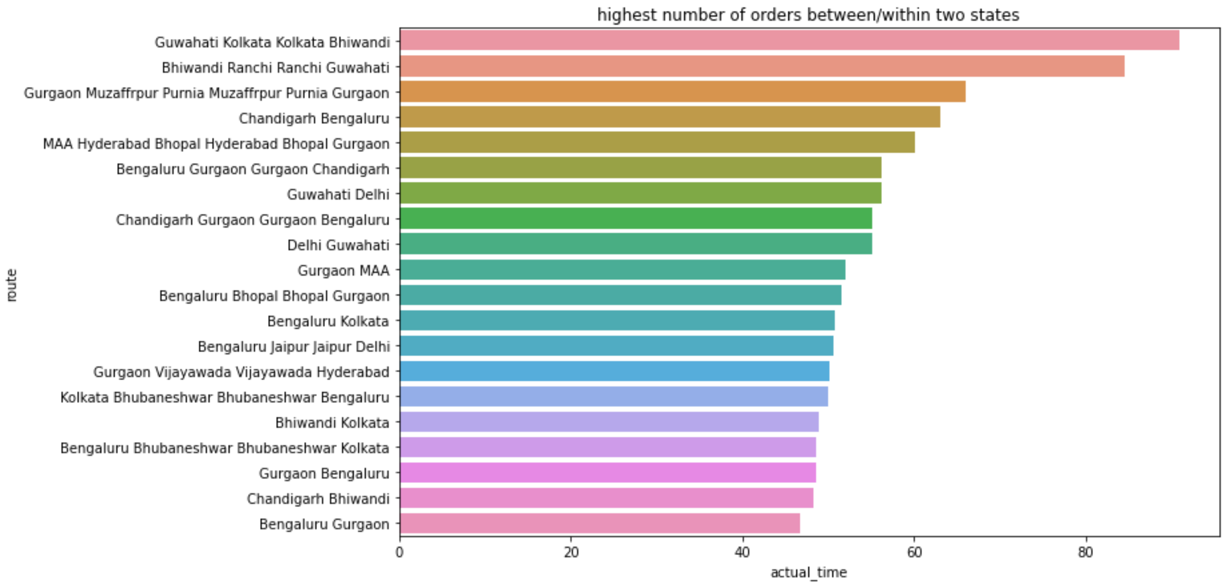
Out[129]:

	SouceToDestination_city	Number_of_Trips	Average_Actual_distance_to_destination	#source_cities	#destination_cities
0	Guwahati TO LakhimpurN	14	281.596486	13	11
2	Jaipur TO Tarnau	20	351.611796	10	10
1	Guwahati TO Tura	12	332.602225	10	10
3	Mangalore TO Udupi	9	195.257193	9	9
4	Ajmer TO Raipur	20	178.737233	9	8
5	Mainpuri TO Tilhar	12	207.247057	8	8
8	Hassan TO Koppa	21	200.497832	7	7
15	Shrirampur TO Sangamner	20	204.509529	7	7
7	Musiri TO Tiruchi	19	219.845121	7	7
9	Bijnor TO Bijnor	17	209.400685	7	7
10	Dausa TO Lalsot	17	232.408310	7	7
17	Tinusukia TO Dibrugarh	16	111.098543	7	7
12	Pondicherry TO Pondicherry	12	230.253602	7	7
14	Mysore TO Mysore	12	154.324190	7	7
6	Golaghat TO Guwahati	11	258.546587	7	7
13	Varanasi TO Varanasi	8	82.545019	7	7
16	Vijayawada TO Suryapet	8	407.029391	7	7
11	Hyderabad TO Miryalguda	7	420.603709	7	7
27	Srikakulam TO Bobbili	22	154.495283	6	6
36	Pukhrayan TO Kanpur	22	139.834945	6	6
48	Dhule TO Shirpur	22	150.016233	6	6
30	Madhupur TO Madhupur	21	252.072259	6	6
38	Kamareddy TO Kamareddy	21	177.923330	6	6
42	Noida TO Khurja	21	208.714043	6	6
20	Junagadh TO Veraval	19	179.538596	6	6

Top 20 Longest Route as per : average actual time taken from one city to another city :

In [130]:

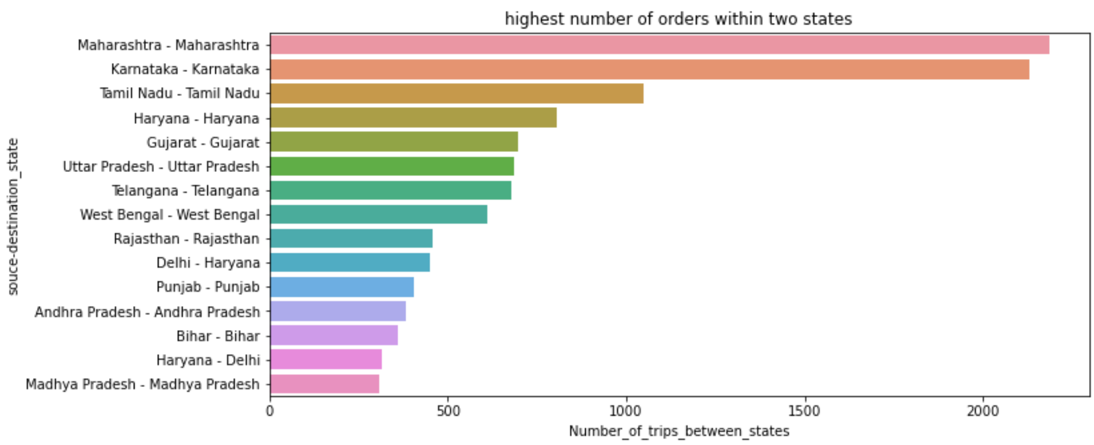
```
Longest_route_as_per_actual_trip_time = trip_records.groupby(["source_city",
                            "destination_city"])["actual_time"].mean().sort_values(ascending=False).head(20).reset_index()
Longest_route_as_per_actual_trip_time["route"] = Longest_route_as_per_actual_trip_time["source_city"] + " " + Longest_route_as_per_actual_
Longest_route_as_per_actual_trip_time.drop(["source_city",
                                           "destination_city"],axis = 1,inplace=True)
Longest_route_as_per_actual_trip_time
plt.figure(figsize=(11,7))
sns.barplot(y = Longest_route_as_per_actual_trip_time["route"],
           x = Longest_route_as_per_actual_trip_time["actual_time"],)
plt.title("highest number of orders between/within two states")
plt.show()
```



highest number of Trips happening between/within two states :

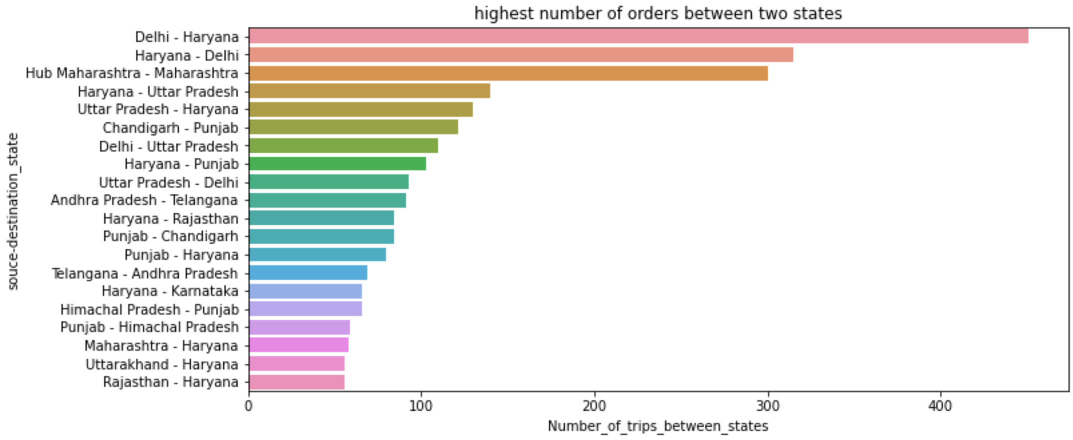
In [131]:

```
highest_order_between_states = data.groupby(["source_state",
                                              "destination_state"])["trip_uuid"].nunique().sort_values(ascending=False).reset_index()
HOBS  = highest_order_between_states.head(15)
HOBS["souce-destination"] = HOBS["source_state"] + " - " + HOBS["destination_state"]
HOBS.drop(["source_state","destination_state"],axis = 1, inplace=True)
HOBS.columns = ["Number_of_trips_between_states","souce-destination_state"]

plt.figure(figsize=(11,5))
sns.barplot(y = HOBS["souce-destination_state"],
           x = HOBS["Number_of_trips_between_states"],)
plt.title("highest number of orders within two states")
plt.show()
```



In [132]:

```
HOBS = data.groupby(["source_state","destination_state"])["trip_uuid"].nunique().sort_values(ascending=False).reset_index()
HOBS = HOBS[HOBS["source_state"]!=HOBS["destination_state"]].head(20)

HOBS["souce-destination"] = HOBS["source_state"] + " - " + HOBS["destination_state"]
HOBS.drop(["source_state","destination_state"],axis = 1, inplace=True)
HOBS.columns = ["Number_of_trips_between_states","souce-destination_state"]

plt.figure(figsize=(11,5))
sns.barplot(y = HOBS["souce-destination_state"],
           x = HOBS["Number_of_trips_between_states"],)
plt.title("highest number of orders between two states")
plt.show()
```



From above charts ,

> Delhi to Haryana is the busiest route, having more than 400 trips in between. Some of such busy routes are Haryana to Uttar Pradesh , Chandigarh to Punjab , Delhi to Uttar Pradesh .

> Within the state , Maharashtra , Karnataka, Tamil Nadu are some states having above 1000 trips.

Top 20 warehouses with heavy traffic :

In [133]:

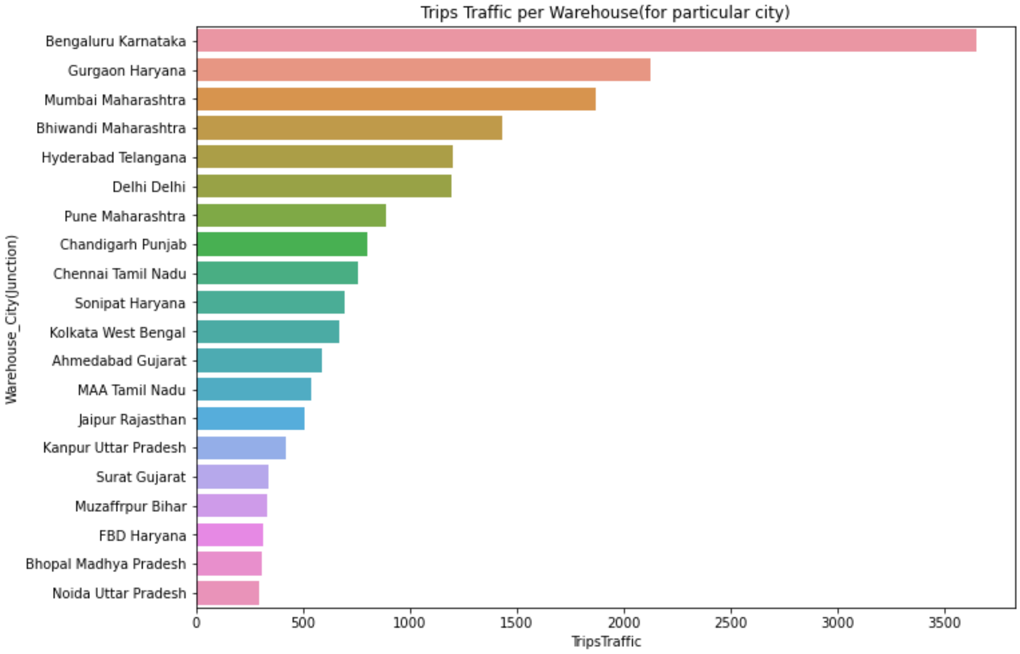
```
destination_traffic = data.groupby(["destination_city_state"])["trip_uuid"].nunique().reset_index()
source_traffic = data.groupby(["source_city_state"])["trip_uuid"].nunique().reset_index()
transactions = source_traffic.merge(destination_traffic,
                               left_on="source_city_state"
                               ,right_on="destination_city_state")
transactions.columns = ["source_city_state","#Trips_s","destination_city_state","#Trips_d"]
transactions["TripsTraffic"] = transactions["#Trips_s"]+transactions["#Trips_d"]
transactions.drop(["#Trips_s","#Trips_d","destination_city_state"],axis = 1,inplace=True)
transactions.columns = ["Warehouse_City(Junction)","TripsTraffic"]
```

In [134]:

```
T = transactions.sort_values(by=["TripsTraffic"],ascending=False).head(20)
```

In [135]:

```
plt.figure(figsize=(11,8))
sns.barplot(y = T["Warehouse_City(Junction)"],
            x = T["TripsTraffic"])
plt.title("Trips Traffic per Warehouse(for particular city)")
plt.show()
```



Top 20 Busiest Warehouse (junctions) as per trips traffic at the juction : are

- 'Bengaluru Karnataka',
- 'Gurgaon Haryana',
- 'Mumbai Maharashtra',
- 'Bhiwandi Maharashtra',
- 'Hyderabad Telangana',
- 'Delhi Delhi',
- 'Pune Maharashtra',
- 'Chandigarh Punjab', -
- 'Chennai Tamil Nadu',
- 'Sonipat Haryana', -
- 'Kolkata West Bengal',
- 'Ahmedabad Gujarat',
- 'MAA Tamil Nadu',
- 'Jaipur Rajasthan',
- 'Kanpur Uttar Pradesh', -
- 'Surat Gujarat',
- 'Muzaffrpur Bihar',
- 'FBD Haryana',
- 'Bhopal Madhya Pradesh',
- 'Noida Uttar Pradesh'

In [136]:

```
trip_records.groupby(["source_state","destination_state"])["trip_uuid"].count().sort_values(ascending=False).head(15).reset_index()
```

Out[136]:

	source_state	destination_state	trip_uuid
0	Maharashtra	Maharashtra	2085
1	Karnataka	Karnataka	2002
2	Tamil Nadu	Tamil Nadu	996
3	Haryana	Haryana	771
4	Telangana	Telangana	627
5	Gujarat	Gujarat	624
6	West Bengal	West Bengal	610
7	Uttar Pradesh	Uttar Pradesh	529
8	Rajasthan	Rajasthan	400
9	Delhi	Haryana	385
10	Andhra Pradesh	Andhra Pradesh	344
11	Punjab	Punjab	342
12	Bihar	Bihar	330
13	Haryana	Delhi	307
14	Hub Maharashtra	Maharashtra	300

Insights

- 14817 different trips happened between source to destinations during 2018 , September and October.
- 1504 delivery routes on which trips are happenig.
- we have 1508 unique source centers and 1481 unique destination centers
- From 14817 total different trips , we have 8908 (60%) of the trip-routes are Carting , which consists of small vehicles and 5909 (40%) of total trip-routes are FTL : which are Full Truck Load get to the destination sooner. as no other pickups or drop offs along the way .

Hypothesis tests Results

```
- from 2 sample t-test ,we can also conclude that

- Average time_taken_btwn_odstart_and_od_end for population is equal to Average start_scan_to_end_scan for population.

- population average actual_time is less than population average start_scan_to_end_scan.

- population mean Actual time taken to complete delivery and population mean time_taken_btwn_od_start_and_od_end are also not sam
e.

- Mean of actual time is higher than Mean of the OSRM estimated time for delivery

- Population average for Actual Time taken to complete delivery trip and segment actual time are same.

- Average of OSRM Time & segment-osrm-time for population is not same.

- Population Mean osrm time is less than Population Mean segment osrm time.

- Average of OSRM distance for population is less than average of segment OSRM distance

- population OSRM estimated distance is higher than the actual distance from source to destination warehouse.
```

EDA Results

Recommendations

- As per analysis, It is recommended to use Carting (small vehicles) for delivery with in the city in order to reduce the delivery time, and Heavy trucks for long distance trips or heavy load. based on this , we can optimize the delivery time as well as increase the revenue as per requirements.
- Incresing the connectivity in tier 2 and tier 3 cities along with profession tie-ups with several e-commerce giants can increase the revenue as well as the reputation on connectivity across borders.
- We can work on optimizing the scanning time on both ends which is start scanning time and end scanning time so that the delivery time can be equated to the OSRM estimated delivery time.
- Revisit information fed to routing engine for trip planning. Check for discrepancies with transporters, if the routing engine is configured for optimum results.
- North, South and West Zones comidors have significant traffic of orders. But, we have a smaller presence in Central, Eastern and North-Eastern zone. However it would be difficult to conclude this, by looking at just 2 months data. It is worth investigating and increasing our presence in these regions.
- From state point of view, we have heavy traffic in Mahrashtra followed by Karnataka. This is a good indicator that we need to plan for resources on ground in these 2 states on priority. Especially, during festive seasons.

In [ ]: