Introduction:

Delhivery, established in 2011, is India's foremost logistics and supply chain service provider, offering a comprehensive range of solutions including express parcel transportation, warehousing, and last-mile delivery.

Leveraging advanced technology and a vast delivery network, Delhivery efficiently manages nationwide movement of goods, earning trust across businesses of all sizes for its dedication to innovation and customer satisfaction.

As the largest fully integrated player in India by revenue in Fiscal 2021, Delhivery aims to lead the industry by pioneering the commerce operating system, driven by top-tier infrastructure, logistics operations, and innovative data intelligence initiatives led by its Data team.

Why this case study?

Delhivery aims to establish itself as the premier player in the logistics industry. This case study is of paramount importance as it aligns with the company's core objectives and operational excellence.

It provides a practical framework for understanding and processing data, which is integral to their operations. By leveraging data engineering pipelines and data analysis techniques, Delhivery can achieve several critical goals.

First, it allows them to ensure data integrity and quality by addressing missing values and structuring the dataset appropriately.

Second, it enables the extraction of valuable features from raw data, which can be utilized for building accurate forecasting models.

Moreover, it facilitates the identification of patterns, insights, and actionable recommendations crucial for optimizing their logistics operations.

Byconducting hypothesis testing and outlier detection, Delhivery can refine their processes and further enhance the quality of service they provide.

Problem Statment

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.

Importing packages

In [63]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import scipy.stats as spy
import re
```

Loading Dataset

In [64]:
```
df =
pd.read_csv("https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/001/551/original/delhivery_data.csv
1642751181")
df
```

Out[64]:

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name	de
0	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	
1	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	
2	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	
3	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	
4	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	
...	...	...	...	...	...	...	...	
144862	training	2018-09-20 16:24:28.436231	thanos::sroute:f0569d2f-4e20-4c31-8542-67b86d5...	Carting	trip-153746066843555182	IND131028AAB	Sonipat_Kundli_H (Haryana)	
144863	training	2018-09-20 16:24:28.436231	thanos::sroute:f0569d2f-4e20-4c31-8542-67b86d5...	Carting	trip-153746066843555182	IND131028AAB	Sonipat_Kundli_H (Haryana)	
144864	training	2018-09-20 16:24:28.436231	thanos::sroute:f0569d2f-4e20-4c31-8542-67b86d5...	Carting	trip-153746066843555182	IND131028AAB	Sonipat_Kundli_H (Haryana)	
144865	training	2018-09-20 16:24:28.436231	thanos::sroute:f0569d2f-4e20-4c31-8542-67b86d5...	Carting	trip-153746066843555182	IND131028AAB	Sonipat_Kundli_H (Haryana)	
144866	training	2018-09-20 16:24:28.436231	thanos::sroute:f0569d2f-4e20-4c31-8542-67b86d5...	Carting	trip-153746066843555182	IND131028AAB	Sonipat_Kundli_H (Haryana)	

144867 rows × 24 columns

Understanding the shape and structure

In [65]:
```
df.shape
```

Out[65]:
```
(144867, 24)
```

There are 144867 Rows and 24 columns

In [66]:
```
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 [67]:
```
df[['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','actual_distance_to_destination','actual_time','segment_osrm_time','segmen
```

Out[67]:

	route_schedule_uuid	route_type	trip_uuid	source_center	source_name	destination_center	des
0	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	IND388620AAB	Khambha
1	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	IND388620AAB	Khambha
2	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	IND388620AAB	Khambha
3	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	IND388620AAB	Khambha
4	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	IND388620AAB	Khambha
5	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388620AAB	Khambhat_MotvdDPP_D (Gujarat)	IND388320AAA	Ana
6	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388620AAB	Khambhat_MotvdDPP_D (Gujarat)	IND388320AAA	Ana
7	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388620AAB	Khambhat_MotvdDPP_D (Gujarat)	IND388320AAA	Ana
8	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388620AAB	Khambhat_MotvdDPP_D (Gujarat)	IND388320AAA	Ana
9	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388620AAB	Khambhat_MotvdDPP_D (Gujarat)	IND388320AAA	Ana
10	thanos::sroute:ff52ef7a-4d0d-4063-9bfe-cc21172...	FTL	trip-153768492602129387	IND421302AAG	Bhiwandi_Mankoli_HB (Maharashtra)	IND411033AAA	Pur
11	thanos::sroute:ff52ef7a-4d0d-4063-9bfe-cc21172...	FTL	trip-153768492602129387	IND421302AAG	Bhiwandi_Mankoli_HB (Maharashtra)	IND411033AAA	Pur
12	thanos::sroute:ff52ef7a-4d0d-4063-9bfe-cc21172...	FTL	trip-153768492602129387	IND421302AAG	Bhiwandi_Mankoli_HB (Maharashtra)	IND411033AAA	Pur
13	thanos::sroute:ff52ef7a-4d0d-4063-9bfe-cc21172...	FTL	trip-153768492602129387	IND421302AAG	Bhiwandi_Mankoli_HB (Maharashtra)	IND411033AAA	Pur
14	thanos::sroute:ff52ef7a-4d0d-4063-9bfe-cc21172...	FTL	trip-153768492602129387	IND421302AAG	Bhiwandi_Mankoli_HB (Maharashtra)	IND411033AAA	Pur
15	thanos::sroute:a16bfa03-3462-4bce-9c82-5784c7d...	Carting	trip-153693976643699843	IND400011AAA	LowerParel_CP (Maharashtra)	IND400072AAD	Mumb
16	thanos::sroute:a16bfa03-3462-4bce-9c82-5784c7d...	Carting	trip-153693976643699843	IND400011AAA	LowerParel_CP (Maharashtra)	IND400072AAD	Mumb
17	thanos::sroute:76951383-1608-44e4-a284-46d92e8...	FTL	trip-153687145942424248	IND562132AAA	Bangalore_Nelmngla_H (Karnataka)	IND560099AAB	Bengaluru
18	thanos::sroute:76951383-1608-44e4-a284-46d92e8...	FTL	trip-153687145942424248	IND562132AAA	Bangalore_Nelmngla_H (Karnataka)	IND560099AAB	Bengaluru
19	thanos::sroute:76951383-1608-44e4-a284-46d92e8...	FTL	trip-153687145942424248	IND560099AAB	Bengaluru_Bomsndra_HB (Karnataka)	IND683511AAA	Al

In [68]:
```
df.dtypes
```

Out[68]:
```
data                                object
trip_creation_time                  object
route_schedule_uuid                 object
route_type                          object
trip_uuid                           object
source_center                       object
source_name                         object
destination_center                  object
destination_name                    object
od_start_time                       object
od_end_time                         object
start_scan_to_end_scan             float64
is_cutoff                             bool
cutoff_factor                        int64
cutoff_timestamp                    object
actual_distance_to_destination     float64
actual_time                        float64
osrm_time                          float64
osrm_distance                      float64
factor                             float64
segment_actual_time                float64
segment_osrm_time                  float64
segment_osrm_distance              float64
segment_factor                     float64
dtype: object
```

In [69]:
```
df.describe(include =object)
```

Out[69]:

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name
count	144867	144867	144867	144867	144867	144867	144574
unique	2	14817	1504	2	14817	1508	1498
top	training	2018-09-28 05:23:15.359220	thanos::sroute:4029a8a2-6c74-4b7e-a6d8-f9e069f...	FTL	trip-153811219535896559	IND000000ACB	Gurgaon_Bilaspur_HB (Haryana)
freq	104858	101	1812	99660	101	23347	23347

Missing value detection

In [70]:
```
df.isna().sum()
```

Out[70]:
```
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
```

Observation

- Source_name and destination_name has null values
- AS we have very few records having missing values , we can drop those rows for our analysis

In [71]:
```
df.drop(df[df['source_name'].isna() | df['destination_name'].isna()].index,inplace=True)
```

In [72]:
```
df.isna().sum()
```

Out[72]:

```
data                                0
trip_creation_time                  0
route_schedule_uuid                 0
route_type                          0
trip_uuid                           0
source_center                       0
source_name                         0
destination_center                  0
destination_name                    0
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
```

Converting the date column into proper datatype

In [73]:

```
df.trip_creation_time = pd.to_datetime(df.trip_creation_time)
df.od_start_time = pd.to_datetime(df.od_start_time)
df.od_end_time = pd.to_datetime(df.od_end_time)
df
```

Out[73]:

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name	de
0	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	
1	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	
2	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	
3	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	
4	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	
...	...	...	...	...	...	...	...	
144862	training	2018-09-20 16:24:28.436231	thanos::sroute:f0569d2f-4e20-4c31-8542-67b86d5...	Carting	trip-153746066843555182	IND131028AAB	Sonipat_Kundli_H (Haryana)	
144863	training	2018-09-20 16:24:28.436231	thanos::sroute:f0569d2f-4e20-4c31-8542-67b86d5...	Carting	trip-153746066843555182	IND131028AAB	Sonipat_Kundli_H (Haryana)	
144864	training	2018-09-20 16:24:28.436231	thanos::sroute:f0569d2f-4e20-4c31-8542-67b86d5...	Carting	trip-153746066843555182	IND131028AAB	Sonipat_Kundli_H (Haryana)	
144865	training	2018-09-20 16:24:28.436231	thanos::sroute:f0569d2f-4e20-4c31-8542-67b86d5...	Carting	trip-153746066843555182	IND131028AAB	Sonipat_Kundli_H (Haryana)	
144866	training	2018-09-20 16:24:28.436231	thanos::sroute:f0569d2f-4e20-4c31-8542-67b86d5...	Carting	trip-153746066843555182	IND131028AAB	Sonipat_Kundli_H (Haryana)	

144316 rows × 24 columns

Checking the range of the dataset available

In [74]:

```
df.trip_creation_time.dt.to_period("M").astype(str).value_counts()
```

Out[74]:

```
trip_creation_time
2018-09    126932
2018-10     17384
Name: count, dtype: int64
```

Data points are from Sep to Oct of 2018

Number of unique categories

In [75]:
```
df.nunique()
```

Out[75]:
```
data                                2
trip_creation_time                14787
route_schedule_uuid                1497
route_type                            2
trip_uuid                         14787
source_center                      1496
source_name                        1496
destination_center                 1466
destination_name                   1466
od_start_time                     26223
od_end_time                       26223
start_scan_to_end_scan             1914
is_cutoff                             2
cutoff_factor                       501
cutoff_timestamp                  92894
actual_distance_to_destination   143965
actual_time                        3182
osrm_time                          1531
osrm_distance                    137544
factor                            45588
segment_actual_time                 746
segment_osrm_time                   214
segment_osrm_distance            113497
segment_factor                     5663
dtype: int64
```

- There are total of 14817 trips have been made
- There are 1496 source center and 1466 destination center

Trip segment analysis

- Creating a segment key to create a unique identifier for different segment of trip trip_uuid, source_center and destination_center
- Based on segment key , we will create new aggregated columns segment_actual_time,segment_osrm_distance,segment_osrm_time

In [76]:
```
df['segment_key'] = df.apply(lambda x
:"#".join([x['trip_uuid'], ['source_center'],x['destination_center']]),axis =1)
df
```

Out[76]:

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name	de
0	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	
1	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	
2	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	
3	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	
4	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	
...	...	...	...	...	...	...	...	
144862	training	2018-09-20 16:24:28.436231	thanos::sroute:f0569d2f-4e20-4c31-8542-67b86d5...	Carting	trip-153746066843555182	IND131028AAB	Sonipat_Kundli_H (Haryana)	
144863	training	2018-09-20 16:24:28.436231	thanos::sroute:f0569d2f-4e20-4c31-8542-67b86d5...	Carting	trip-153746066843555182	IND131028AAB	Sonipat_Kundli_H (Haryana)	
144864	training	2018-09-20 16:24:28.436231	thanos::sroute:f0569d2f-4e20-4c31-8542-67b86d5...	Carting	trip-153746066843555182	IND131028AAB	Sonipat_Kundli_H (Haryana)	
144865	training	2018-09-20 16:24:28.436231	thanos::sroute:f0569d2f-4e20-4c31-8542-67b86d5...	Carting	trip-153746066843555182	IND131028AAB	Sonipat_Kundli_H (Haryana)	
144866	training	2018-09-20 16:24:28.436231	thanos::sroute:f0569d2f-4e20-4c31-8542-67b86d5...	Carting	trip-153746066843555182	IND131028AAB	Sonipat_Kundli_H (Haryana)	

144316 rows × 25 columns

In [77]:
```
df['segment_actual_time_cumsum'] = df.groupby('segment_key')['segment_actual_time'].transform(lambda
x:x.cumsum())
```

In [78]:
```
df['segment_osrm_time_cumsum'] = df.groupby('segment_key')['segment_osrm_time'].transform(lambda x:x.cumsum())
df['segment_osrm_distance_cumsum'] = df.groupby('segment_key')['segment_osrm_distance'].transform(lambda
x:x.cumsum())
```

In [79]:
```
df
```

Out[79]:

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name	de
0	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	
1	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	
2	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	
3	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	
4	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	
...	...	...	...	...	...	...	...	
144862	training	2018-09-20 16:24:28.436231	thanos::sroute:f0569d2f-4e20-4c31-8542-67b86d5...	Carting	trip-153746066843555182	IND131028AAB	Sonipat_Kundli_H (Haryana)	
144863	training	2018-09-20 16:24:28.436231	thanos::sroute:f0569d2f-4e20-4c31-8542-67b86d5...	Carting	trip-153746066843555182	IND131028AAB	Sonipat_Kundli_H (Haryana)	
144864	training	2018-09-20 16:24:28.436231	thanos::sroute:f0569d2f-4e20-4c31-8542-67b86d5...	Carting	trip-153746066843555182	IND131028AAB	Sonipat_Kundli_H (Haryana)	
144865	training	2018-09-20 16:24:28.436231	thanos::sroute:f0569d2f-4e20-4c31-8542-67b86d5...	Carting	trip-153746066843555182	IND131028AAB	Sonipat_Kundli_H (Haryana)	
144866	training	2018-09-20 16:24:28.436231	thanos::sroute:f0569d2f-4e20-4c31-8542-67b86d5...	Carting	trip-153746066843555182	IND131028AAB	Sonipat_Kundli_H (Haryana)	

144316 rows × 28 columns

In [80]:
```
df['segment_actual_time_sum']=df.groupby('segment_key')['segment_actual_time_cumsum'].transform(lambda
x:x.iloc[-1])
df['segment_osrm_time_sum']=df.groupby('segment_key')['segment_osrm_time_cumsum'].transform(lambda
x:x.iloc[-1])
df['segment_osrm_distance_sum']=df.groupby('segment_key')['segment_osrm_distance_cumsum'].transform(lambda
x:x.iloc[-1])
```

In [81]:
```
df=df.sort_values(by=['segment_key','od_end_time'])
df
```

Out[81]:

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name
125002	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)
125003	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)
125004	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)
125005	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)
125006	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)
...	...	...	...	...	...	...	...
86464	test	2018-10-03 23:59:14.390954	thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...	Carting	trip-153861115439069069	IND628801AAA	Eral_Busstand_D (Tamil Nadu)
11572	test	2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	FTL	trip-153861118270144424	IND583119AAA	Sandur_WrdN1DPP_D (Karnataka)
11573	test	2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	FTL	trip-153861118270144424	IND583119AAA	Sandur_WrdN1DPP_D (Karnataka)
11570	test	2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	FTL	trip-153861118270144424	IND583201AAA	Hospet (Karnataka)
11571	test	2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	FTL	trip-153861118270144424	IND583201AAA	Hospet (Karnataka)

144316 rows × 31 columns

Feature Engineering

In [82]:
```
df['od_time_diff_hour']=(df['od_end_time']-df['od_start_time'])/pd.Timedelta(hours=1)
df.od_time_diff_hour
```

Out[82]:
```
125002    21.010074
125003    21.010074
125004    21.010074
125005    21.010074
125006    21.010074
            ...
86464      0.736240
11572      4.791233
11573      4.791233
11570      1.115559
11571      1.115559
Name: od_time_diff_hour, Length: 144316, dtype: float64
```

Extract city, place , code and state information

In [83]:
```
df.columns
```

Out[83]:
```
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', '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',
       'segment_key', 'segment_actual_time_cumsum', 'segment_osrm_time_cumsum',
       'segment_osrm_distance_cumsum', 'segment_actual_time_sum',
       'segment_osrm_time_sum', 'segment_osrm_distance_sum',
       'od_time_diff_hour'],
      dtype='object')
```

In [84]:
```
df['trip_creation_year']=df['trip_creation_time'].dt.year
df['trip_creation_month']=df['trip_creation_time'].dt.month
df['trip_creation_day']=df['trip_creation_time'].dt.day
```

In [85]:
```
def get_state(name):
```

```
    pattern="\([A-Za-z]+\s?\w+\)"
    pattern="\([A-Za-z &]+\s?\w+\)"
    state=re.findall(pattern, name)[0]
    state=state.replace("(","")
    state=state.replace(")","")
    return state
```

In [86]:
```
def get_city(name):
    pattern="\([A-Za-z &]+\s?\w+\)"
    state=re.findall(pattern, name)[0]
    city_place_code=name.replace(state,'')
    city_place_code_parts=city_place_code.split("_")
    if len(city_place_code_parts)==1 :
        city=city_place_code_parts[0].strip()
    elif len(city_place_code_parts)==2:
        city=city_place_code.strip()
    elif len(city_place_code_parts)==3 or len(city_place_code_parts)==4:
        city=city_place_code_parts[0].strip()
    else:
        city=city_place_code
    return city
```

In [87]:
```
def get_place(name):
    pattern="\([A-Za-z &]+\s?\w+\)"
    try:
        state=re.findall(pattern, name)[0]
        city_place_code=name.replace(state,'')
        city_place_code_parts=city_place_code.split("_")
        if len(city_place_code_parts)==3 or len(city_place_code_parts)==4:
            place=city_place_code_parts[1].strip()
        else:
            place=None
        return place
    except Exception as exp:
        return None
```

In [88]:
```
def get_code(name):
    pattern="\([A-Za-z &]+\s?\w+\)"
    try:
        state=re.findall(pattern, name)[0]
        city_place_code=name.replace(state,'')
        city_place_code_parts=city_place_code.split("_")
        if len(city_place_code_parts)==3 :
            code=city_place_code_parts[2].strip()
        elif  len(city_place_code_parts)==4:
            code="_".join(city_place_code_parts[2:]).strip()
        else:
            code=None
        return code
    except Exception as exp:
        return None
```

In [89]:
```
df['destination_state']=df['destination_name'].map(get_state)
df['source_state']=df['source_name'].map(get_state)
df.head()
```

Out[89]:

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name	d
125002	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)	
125003	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)	
125004	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)	
125005	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)	
125006	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)	

5 rows × 37 columns

In [90]:
```
df['destination_city']=df['destination_name'].map(get_city)
df['source_city']=df['source_name'].map(get_state)
```

In [91]:
```
df['destination_place']=df['destination_name'].map(get_place)
df['source_place']= df['source_name'].map(get_place)
```

In [92]:
```
df['destination_code']=df['destination_name'].map(get_code)
df['source_code']= df['source_name'].map(get_code)
```

In [93]:
```
df.destination_name.unique()
```

Out[93]:
```
array(['Gurgaon_Bilaspur_HB (Haryana)',
       'Kanpur_Central_H_6 (Uttar Pradesh)',
       'Chikblapur_ShntiSgr_D (Karnataka)', ...,
       'Kapadvanj_Busstand_D (Gujarat)', 'Lunawada_VrdhriRD_D (Gujarat)',
       'Jaipur_Central_D_1 (Rajasthan)'], dtype=object)
```

In [94]:
```
df
```

Out[94]:

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name
125002	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)
125003	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)
125004	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)
125005	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)
125006	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)
...	...	...	...	...	...	...	...
86464	test	2018-10-03 23:59:14.390954	thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...	Carting	trip-153861115439069069	IND628801AAA	Eral_Busstand_D (Tamil Nadu)
11572	test	2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	FTL	trip-153861118270144424	IND583119AAA	Sandur_WrdN1DPP_D (Karnataka)
11573	test	2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	FTL	trip-153861118270144424	IND583119AAA	Sandur_WrdN1DPP_D (Karnataka)
11570	test	2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	FTL	trip-153861118270144424	IND583201AAA	Hospet (Karnataka)
11571	test	2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	FTL	trip-153861118270144424	IND583201AAA	Hospet (Karnataka)

144316 rows × 43 columns

In-depth analysis:

Grouping and Aggregating at Trip-level

In [95]:
```
trip_actual_time_sum_df=df.groupby(['trip_uuid','segment_key']).nth(-1).groupby(['trip_uuid']).agg({'actual_time':'sum'}).reset_index()
trip_actual_time_sum_df
```

Out[95]:

	trip_uuid	trip_actual_time_sum
0	trip-153671041653548748	1562.0
1	trip-153671042288605164	143.0
2	trip-153671043369099517	3347.0
3	trip-153671046011330457	59.0
4	trip-153671052974046625	341.0
...	...	...
14782	trip-153861095625827784	83.0
14783	trip-153861104386292051	21.0
14784	trip-153861106442901555	282.0
14785	trip-153861115439069069	264.0
14786	trip-153861118270144424	275.0

14787 rows × 2 columns

In [96]:
```
trip_segment_actual_time_sum_df=df.groupby(['trip_uuid','segment_key']).nth(-1).groupby(['trip_uuid']).agg(trip
('segment_actual_time_sum','sum')).reset_index()
trip_segment_actual_time_sum_df
```

Out[96]:

	trip_uuid	trip_segment_actual_time_sum
0	trip-153671041653548748	1548.0
1	trip-153671042288605164	141.0
2	trip-153671043369099517	3308.0
3	trip-153671046011330457	59.0
4	trip-153671052974046625	340.0
...	...	...
14782	trip-153861095625827784	82.0
14783	trip-153861104386292051	21.0
14784	trip-153861106442901555	281.0
14785	trip-153861115439069069	258.0
14786	trip-153861118270144424	274.0

14787 rows × 2 columns

In [97]:
```
trip_segment_osrm_time_sum_df=df.groupby(['trip_uuid','segment_key']).nth(-1).groupby(['trip_uuid']).agg(trip_s
('segment_osrm_time_sum','sum')).reset_index()
trip_segment_osrm_time_sum_df
```

Out[97]:

	trip_uuid	trip_segment_osrm_time_sum
0	trip-153671041653548748	1008.0
1	trip-153671042288605164	65.0
2	trip-153671043369099517	1941.0
3	trip-153671046011330457	16.0
4	trip-153671052974046625	115.0
...	...	...
14782	trip-153861095625827784	62.0
14783	trip-153861104386292051	11.0
14784	trip-153861106442901555	88.0
14785	trip-153861115439069069	221.0
14786	trip-153861118270144424	67.0

14787 rows × 2 columns

In [98]:
```
trip_segment_osrm_distance_sum_df=df.groupby(['trip_uuid','segment_key']).nth(-1).groupby(['trip_uuid']).agg(t
('segment_osrm_distance_sum','sum')).reset_index()
trip_segment_osrm_distance_sum_df
```

Out[98]:

	trip_uuid	trip_segment_osrm_distance_sum
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
...	...	...
14782	trip-153861095625827784	64.8551
14783	trip-153861104386292051	16.0883
14784	trip-153861106442901555	104.8866
14785	trip-153861115439069069	223.5324
14786	trip-153861118270144424	80.5787

14787 rows × 2 columns

In [99]:

```
trip_osrm_time_sum_df = df.groupby(['trip_uuid','segment_key']).nth(-1).groupby(['trip_uuid']).agg(trip_osrm_time
('osrm_time','sum')).reset_index()
trip_osrm_time_sum_df
```

Out[99]:

	trip_uuid	trip_osrm_time_sum
0	trip-153671041653548748	717.0
1	trip-153671042288605164	68.0
2	trip-153671043369099517	1740.0
3	trip-153671046011330457	15.0
4	trip-153671052974046625	117.0
...	...	...
14782	trip-153861095625827784	62.0
14783	trip-153861104386292051	12.0
14784	trip-153861106442901555	48.0
14785	trip-153861115439069069	179.0
14786	trip-153861118270144424	68.0

14787 rows × 2 columns

In [100…

```
trip_osrm_distance_sum_df = df.groupby(['trip_uuid','segment_key']).nth(-1).groupby(['trip_uuid']).agg(trip_osrm
('osrm_distance','sum')).reset_index()
trip_osrm_distance_sum_df
```

Out[100…

	trip_uuid	trip_osrm_distance_sum
0	trip-153671041653548748	991.3523
1	trip-153671042288605164	85.1110
2	trip-153671043369099517	2354.0665
3	trip-153671046011330457	19.6800
4	trip-153671052974046625	146.7918
...	...	...
14782	trip-153861095625827784	73.4630
14783	trip-153861104386292051	16.0882
14784	trip-153861106442901555	58.9037
14785	trip-153861115439069069	171.1103
14786	trip-153861118270144424	80.5787

14787 rows × 2 columns

In [101…

```
trip_total_time_in_hrs_df = df.groupby(['trip_uuid','segment_key']).nth(-1).groupby(['trip_uuid']).agg(trip_total
('od_time_diff_hour','sum')).reset_index()
trip_total_time_in_hrs_df
```

Out[101…

	trip_uuid	trip_total_time_in_hrs
0	trip-153671041653548748	37.668497
1	trip-153671042288605164	3.026865
2	trip-153671043369099517	65.572709
3	trip-153671046011330457	1.674916
4	trip-153671052974046625	11.972484
...	...	...
14782	trip-153861095625827784	4.300482
14783	trip-153861104386292051	1.009842
14784	trip-153861106442901555	7.035331
14785	trip-153861115439069069	5.808548
14786	trip-153861118270144424	5.906793

14787 rows × 2 columns

Combine all trip aggregated dataframes to single trip aggregations dataframe

In [102…

```
trip_aggregated_df=pd.concat(
    objs=(
            iDF.set_index('trip_uuid') for iDF in (trip_actual_time_sum_df,trip_osrm_time_sum_df,
trip_osrm_distance_sum_df,

trip_segment_actual_time_sum_df,trip_segment_osrm_time_sum_df,trip_segment_osrm_distance_sum_df,
                                                   trip_total_time_in_hrs_df)
        ),
    axis=1,
    join='inner'
).reset_index()
trip_aggregated_df
```

Out[102…

	trip_uuid	trip_actual_time_sum	trip_osrm_time_sum	trip_osrm_distance_sum	trip_segment_actual_time_sum	tri
0	trip-153671041653548748	1562.0	717.0	991.3523	1548.0	
1	trip-153671042288605164	143.0	68.0	85.1110	141.0	
2	trip-153671043369099517	3347.0	1740.0	2354.0665	3308.0	
3	trip-153671046011330457	59.0	15.0	19.6800	59.0	
4	trip-153671052974046625	341.0	117.0	146.7918	340.0	
...	...	...	...	...	...	
14782	trip-153861095625827784	83.0	62.0	73.4630	82.0	
14783	trip-153861104386292051	21.0	12.0	16.0882	21.0	
14784	trip-153861106442901555	282.0	48.0	58.9037	281.0	
14785	trip-153861115439069069	264.0	179.0	171.1103	258.0	
14786	trip-153861118270144424	275.0	68.0	80.5787	274.0	

14787 rows × 8 columns

In [103…

```
df=pd.merge(df,trip_aggregated_df, on='trip_uuid' )
df
```

Out[103…

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name
0	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)
1	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)
2	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)
3	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)
4	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)
...	...	...	...	...	...	...	...
144311	test	2018-10-03 23:59:14.390954	thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...	Carting	trip-153861115439069069	IND628801AAA	Eral_Busstand_D (Tamil Nadu)
144312	test	2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	FTL	trip-153861118270144424	IND583119AAA	Sandur_WrdN1DPP_D (Karnataka)
144313	test	2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	FTL	trip-153861118270144424	IND583119AAA	Sandur_WrdN1DPP_D (Karnataka)
144314	test	2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	FTL	trip-153861118270144424	IND583201AAA	Hospet (Karnataka)
144315	test	2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	FTL	trip-153861118270144424	IND583201AAA	Hospet (Karnataka)

144316 rows × 50 columns

Outlier Detection & Treatment

a. Find any existing outliers in numerical features.

b. Visualize the outlier values using Boxplot.

c. Handle the outliers using the IQR method.

Perform one-hot encoding on categorical features.

Normalize/ Standardize the numerical features using MinMaxScaler or StandardScaler.

In [104…

```
trip_aggregated_df
```

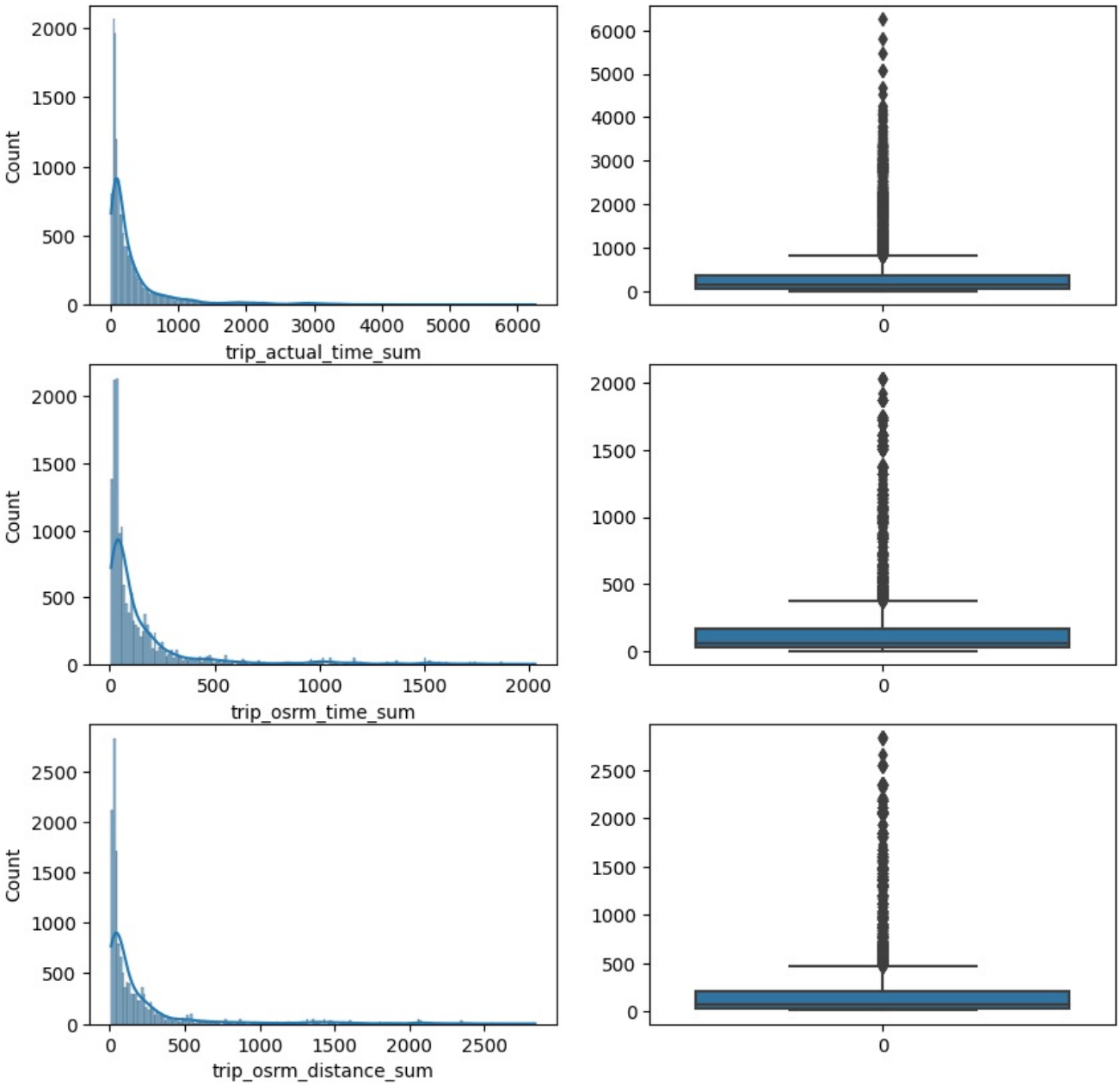
Out[104…

	trip_uuid	trip_actual_time_sum	trip_osrm_time_sum	trip_osrm_distance_sum	trip_segment_actual_time_sum	tri
0	trip-153671041653548748	1562.0	717.0	991.3523	1548.0	
1	trip-153671042288605164	143.0	68.0	85.1110	141.0	
2	trip-153671043369099517	3347.0	1740.0	2354.0665	3308.0	
3	trip-153671046011330457	59.0	15.0	19.6800	59.0	
4	trip-153671052974046625	341.0	117.0	146.7918	340.0	
...	...	...	...	...	...	
14782	trip-153861095625827784	83.0	62.0	73.4630	82.0	
14783	trip-153861104386292051	21.0	12.0	16.0882	21.0	
14784	trip-153861106442901555	282.0	48.0	58.9037	281.0	
14785	trip-153861115439069069	264.0	179.0	171.1103	258.0	
14786	trip-153861118270144424	275.0	68.0	80.5787	274.0	

14787 rows × 8 columns

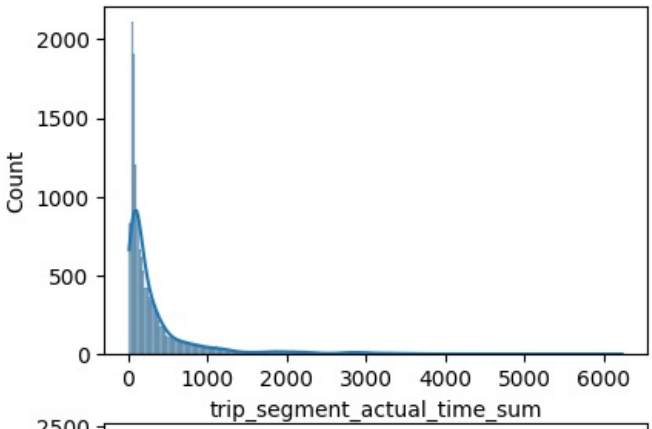
In [105…

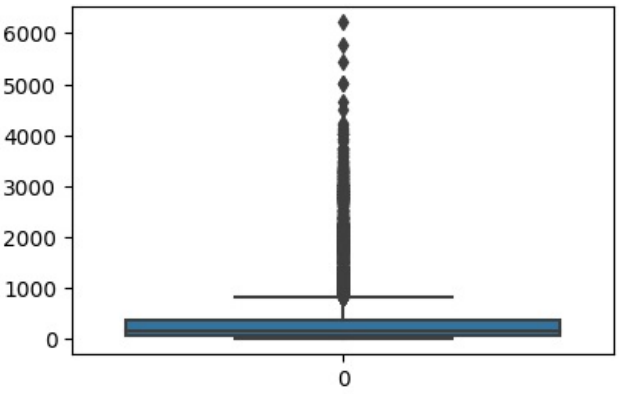
```
import warnings
warnings.filterwarnings("ignore")
fig, axs = plt.subplots(3, 2, figsize=(10,10))
sns.histplot(ax=axs[0,0],data= trip_aggregated_df['trip_actual_time_sum'],kde=True)
sns.boxplot(ax=axs[0,1],data= trip_aggregated_df['trip_actual_time_sum'])

sns.histplot(ax=axs[1,0],data= trip_aggregated_df['trip_osrm_time_sum'],kde=True)
sns.boxplot(ax=axs[1,1],data= trip_aggregated_df['trip_osrm_time_sum'])

sns.histplot(ax=axs[2,0],data= trip_aggregated_df['trip_osrm_distance_sum'],kde=True)
sns.boxplot(ax=axs[2,1],data= trip_aggregated_df['trip_osrm_distance_sum'])


plt.show()
```

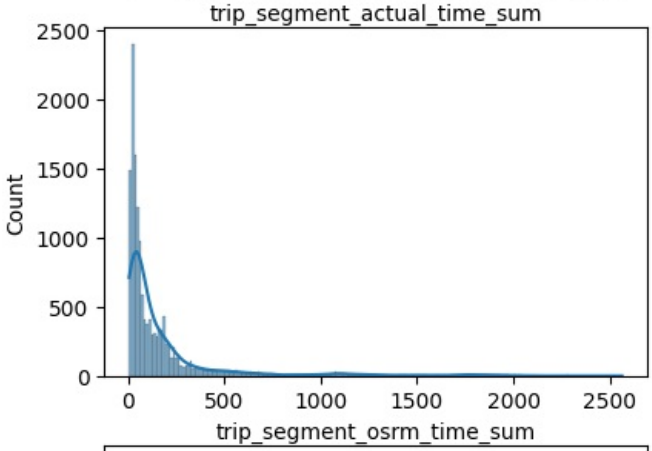


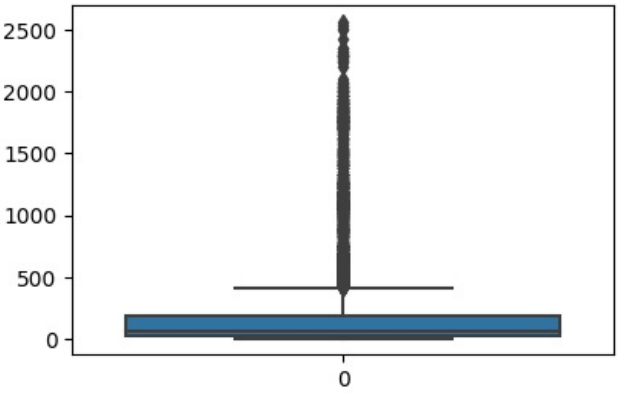
In [106…

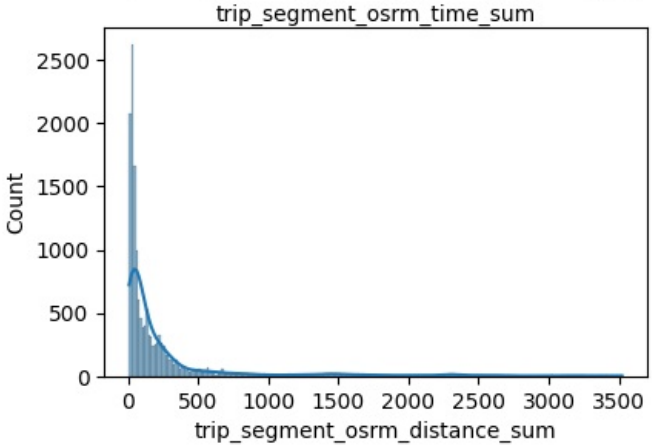
```
import warnings
warnings.filterwarnings("ignore")
fig, axs = plt.subplots(3, 2, figsize=(10,10))

sns.histplot(ax=axs[0,0],data= trip_aggregated_df['trip_segment_actual_time_sum'],kde=True)
sns.boxplot(ax= axs[0,1],data= trip_aggregated_df['trip_segment_actual_time_sum'])


sns.histplot(ax=axs[1,0],data= trip_aggregated_df['trip_segment_osrm_time_sum'],kde=True)
sns.boxplot(ax= axs[1,1],data= trip_aggregated_df['trip_segment_osrm_time_sum'])

sns.histplot(ax=axs[2,0],data= trip_aggregated_df['trip_segment_osrm_distance_sum'],kde=True)
sns.boxplot(ax= axs[2,1],data= trip_aggregated_df['trip_segment_osrm_distance_sum'])


plt.show()
```

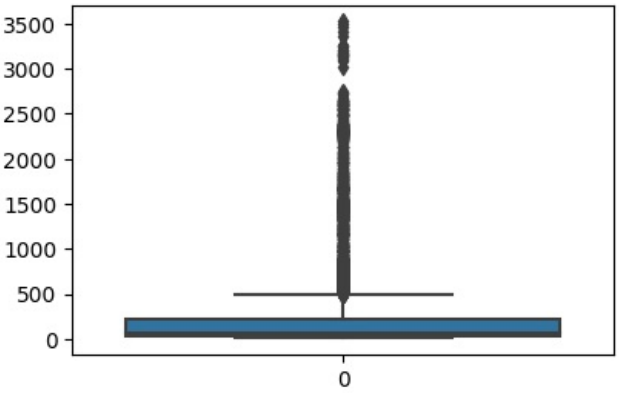

Count
trip_segment_actual_time_sum




Count
trip_segment_osrm_time_sum




Count
trip_segment_osrm_distance_sum



In [107…

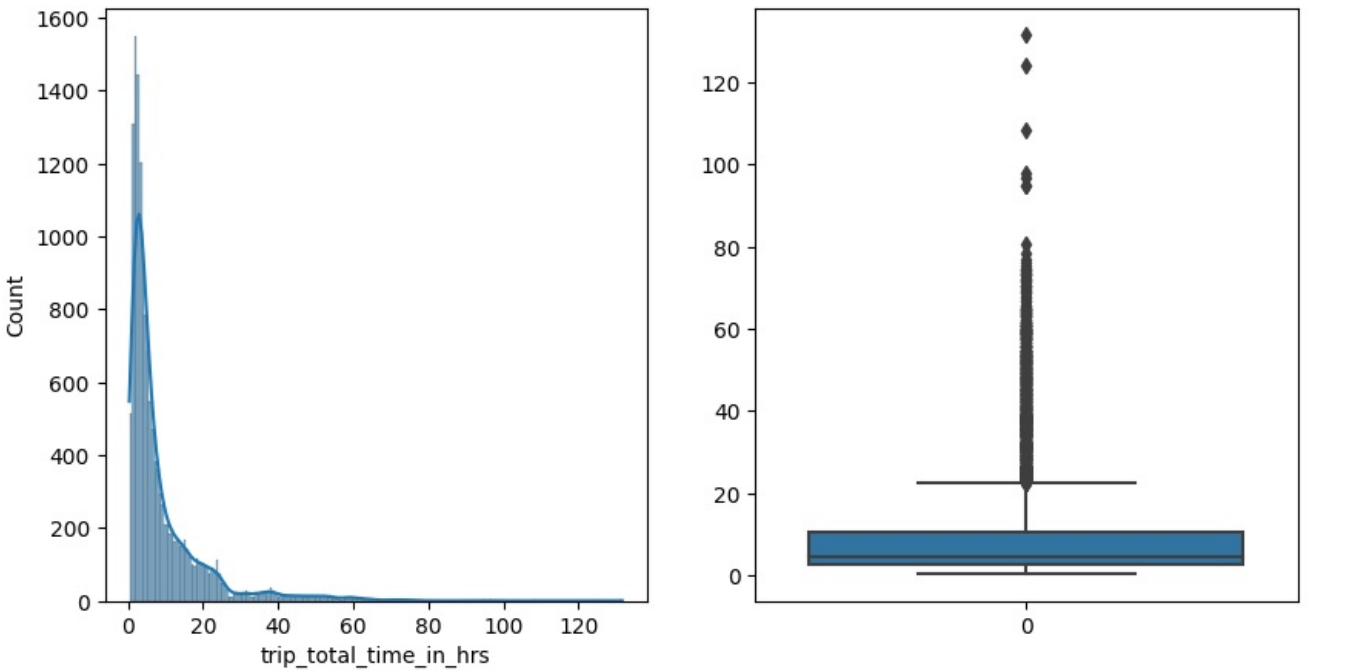
Out[107…

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name
0	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)
1	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)
2	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)
3	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)
4	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)
...	...	...	...	...	...	...	...
144311	test	2018-10-03 23:59:14.390954	thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...	Carting	trip-153861115439069069	IND628801AAA	Eral_Busstand_D (Tamil Nadu)
144312	test	2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	FTL	trip-153861118270144424	IND583119AAA	Sandur_WrdN1DPP_D (Karnataka)
144313	test	2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	FTL	trip-153861118270144424	IND583119AAA	Sandur_WrdN1DPP_D (Karnataka)
144314	test	2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	FTL	trip-153861118270144424	IND583201AAA	Hospet (Karnataka)
144315	test	2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	FTL	trip-153861118270144424	IND583201AAA	Hospet (Karnataka)

144316 rows × 50 columns

In [108…

```
import warnings
warnings.filterwarnings("ignore")
fig, axs = plt.subplots(ncols= 2,  figsize=(10,5))
sns.histplot( ax=axs[0],data= trip_aggregated_df['trip_total_time_in_hrs'],kde=True)
sns.boxplot( ax= axs[1],data= trip_aggregated_df['trip_total_time_in_hrs'])
plt.show()
```

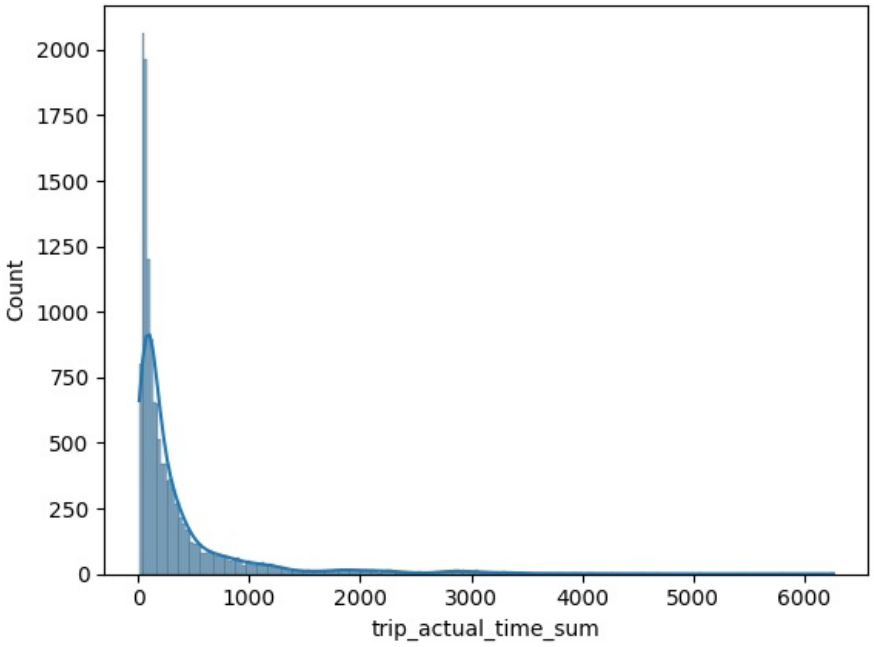


In [109…

```
sns.histplot(data=trip_aggregated_df['trip_actual_time_sum'],kde=True)
```

Out[109…

```
<Axes: xlabel='trip_actual_time_sum', ylabel='Count'>
```



Observations: we have lots of outliers let's remove those outliers Based on IQR range, we will maxout values based on IQR range

In [110…
```
def clip_value_helper(row,cl, Q1,Q3, minval, maxval):
    #     Q1=row[cl].quantile(0.25)
    #     Q3=row[cl].quantile(0.75)
    #     minval=min(row[cl])
    #     maxval=max(row[cl])
    IQR=Q3-Q1
    if row[cl]<Q1-1.5*IQR:
        return min(minval, Q1-1.5*IQR)
    elif row[cl] > Q3+1.5*IQR:
        return min(maxval, Q3+1.5*IQR)
    else:
        return row[cl]
```

In [111…
```
trip_aggregated_df.columns
```

Out[111…
```
Index(['trip_uuid', 'trip_actual_time_sum', 'trip_osrm_time_sum',
       'trip_osrm_distance_sum', 'trip_segment_actual_time_sum',
       'trip_segment_osrm_time_sum', 'trip_segment_osrm_distance_sum',
       'trip_total_time_in_hrs'],
      dtype='object')
```

In [112…
```
for cl in ['trip_actual_time_sum', 'trip_osrm_time_sum', 'trip_osrm_distance_sum',
'trip_segment_actual_time_sum','trip_segment_osrm_time_sum',
'trip_segment_osrm_distance_sum','trip_total_time_in_hrs']:
    Q1=trip_aggregated_df[cl].quantile(0.25)
    Q3=trip_aggregated_df[cl].quantile(0.75)
    minval=min(trip_aggregated_df[cl])
    maxval=max(trip_aggregated_df[cl])
    trip_aggregated_df[cl]=trip_aggregated_df.apply(lambda row:clip_value_helper(row,cl,
                                                    Q1,Q3,minval, maxval) , axis=1)
```

In [113…
```
trip_aggregated_df
```

Out[113…

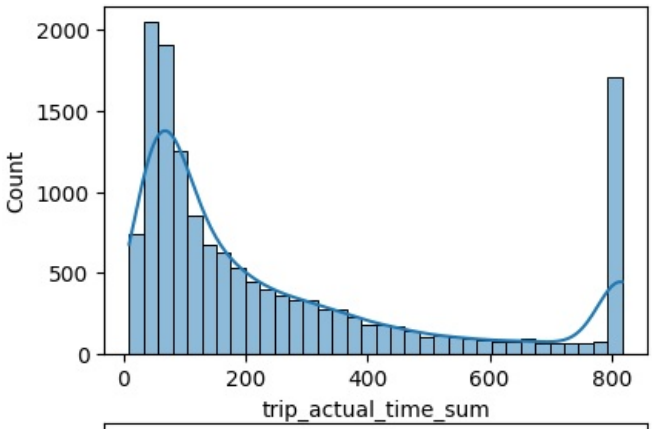
	trip_uuid	trip_actual_time_sum	trip_osrm_time_sum	trip_osrm_distance_sum	trip_segment_actual_time_sum	tri
0	trip-153671041653548748	817.0	376.5	470.47515	811.0	
1	trip-153671042288605164	143.0	68.0	85.11100	141.0	
2	trip-153671043369099517	817.0	376.5	470.47515	811.0	
3	trip-153671046011330457	59.0	15.0	19.68000	59.0	
4	trip-153671052974046625	341.0	117.0	146.79180	340.0	
...	...	...	...	...	...	
14782	trip-153861095625827784	83.0	62.0	73.46300	82.0	
14783	trip-153861104386292051	21.0	12.0	16.08820	21.0	
14784	trip-153861106442901555	282.0	48.0	58.90370	281.0	
14785	trip-153861115439069069	264.0	179.0	171.11030	258.0	
14786	trip-153861118270144424	275.0	68.0	80.57870	274.0	

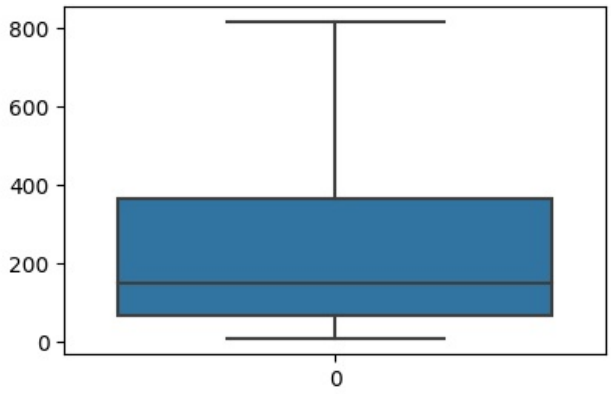
14787 rows × 8 columns

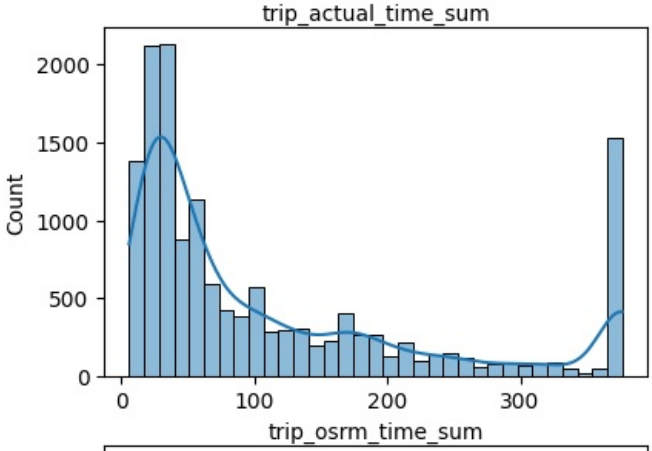
After Clipping Outliers Based On IQR Distribution

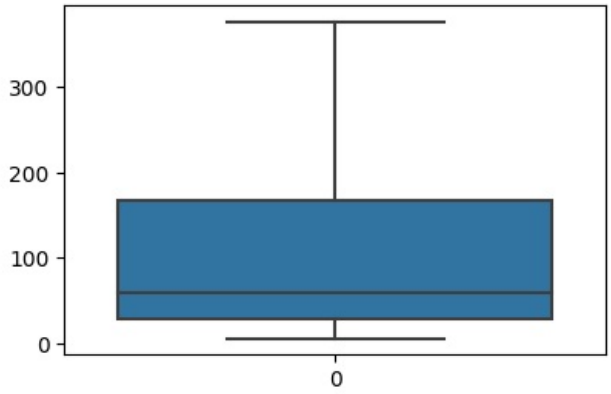
In [114…

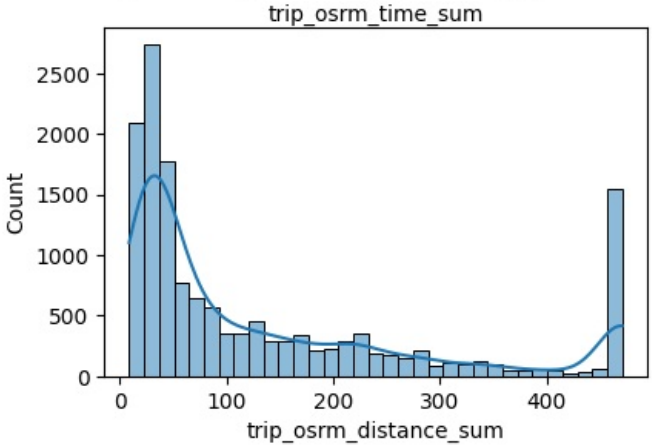
```
import warnings
warnings.filterwarnings("ignore")
fig, axs = plt.subplots(3, 2, figsize=(10,10))
sns.histplot(ax=axs[0,0],data= trip_aggregated_df['trip_actual_time_sum'],kde=True)
sns.boxplot(ax=axs[0,1],data= trip_aggregated_df['trip_actual_time_sum'])

sns.histplot(ax=axs[1,0],data= trip_aggregated_df['trip_osrm_time_sum'],kde=True)
sns.boxplot(ax=axs[1,1],data= trip_aggregated_df['trip_osrm_time_sum'])

sns.histplot(ax=axs[2,0],data= trip_aggregated_df['trip_osrm_distance_sum'],kde=True)
sns.boxplot(ax=axs[2,1],data= trip_aggregated_df['trip_osrm_distance_sum'])


plt.show()
```

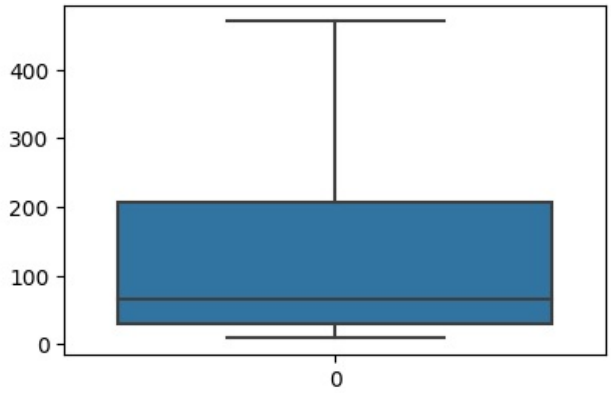






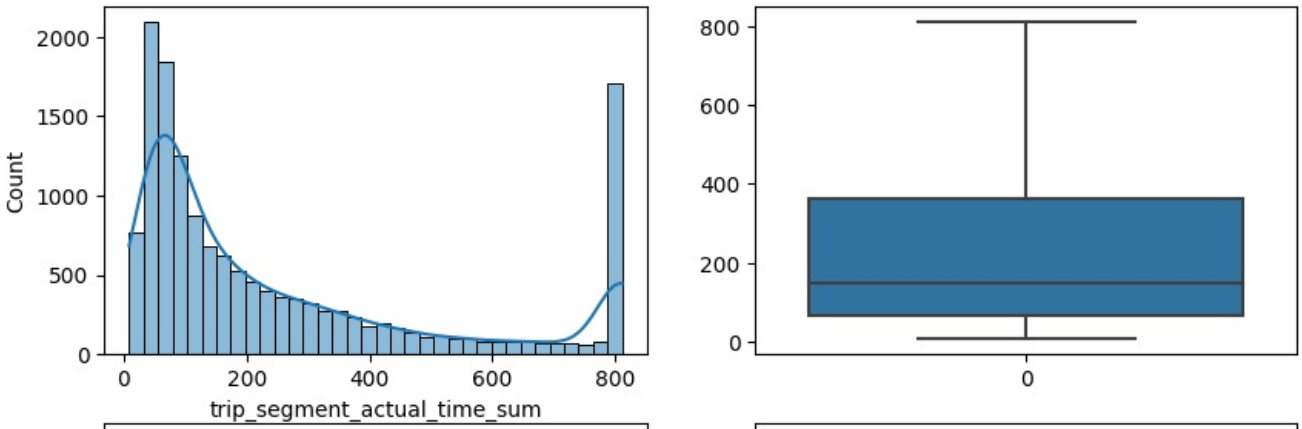


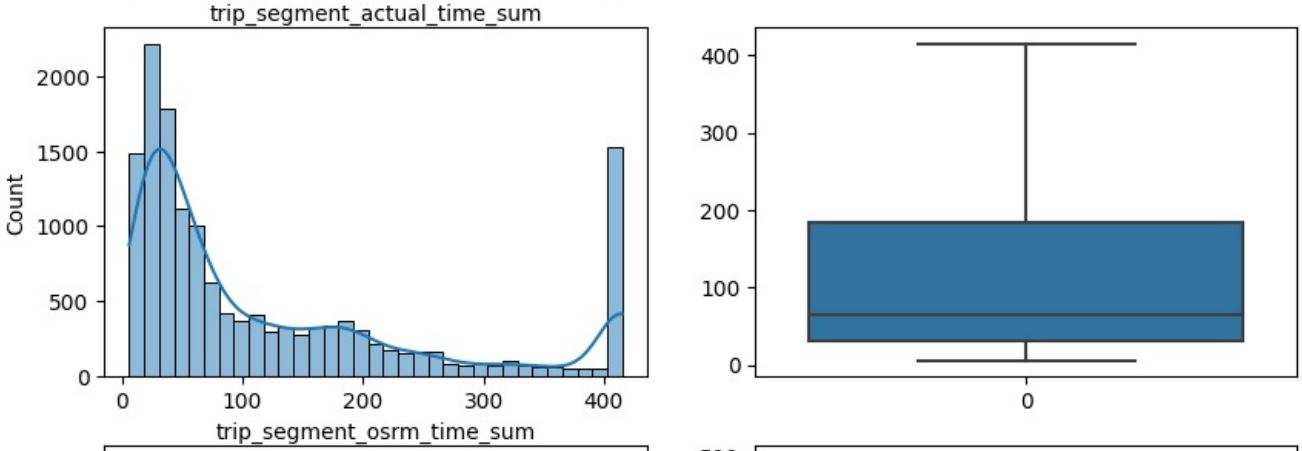


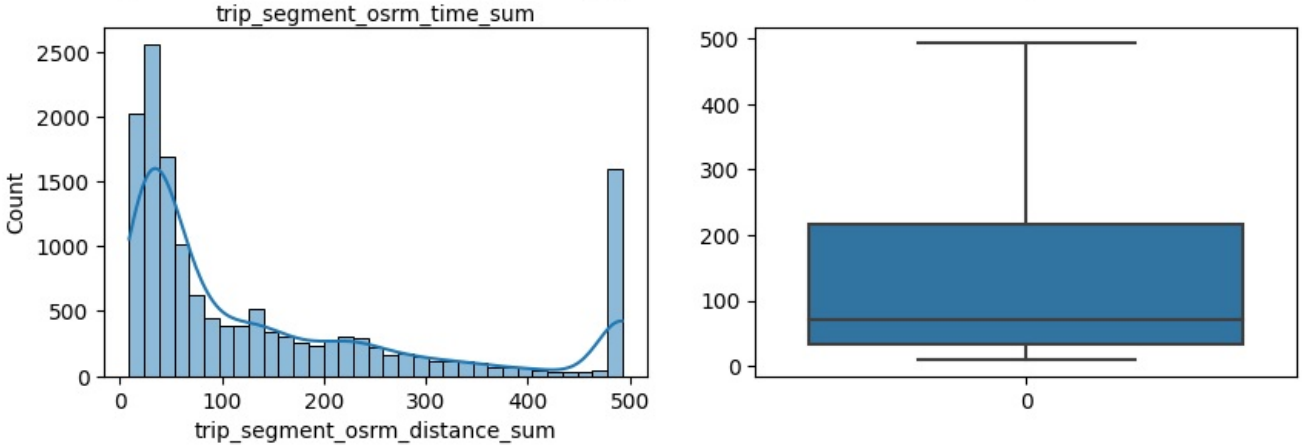


In [115…

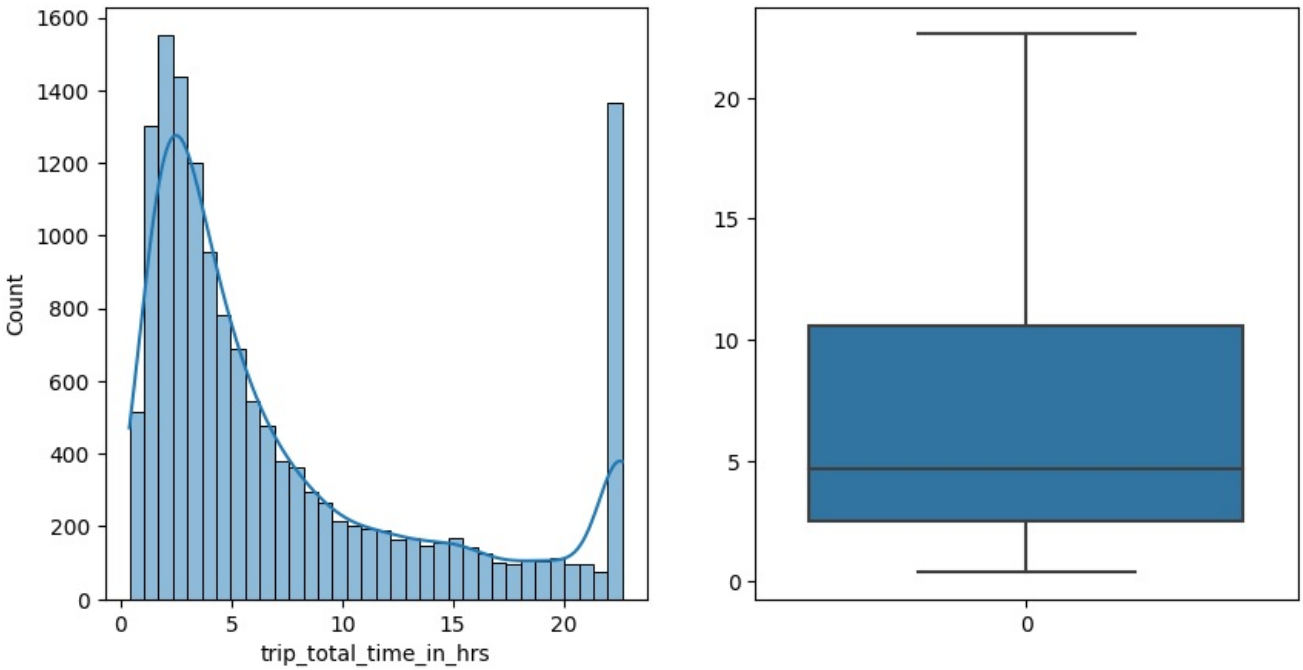
```
import warnings
warnings.filterwarnings("ignore")
fig, axs = plt.subplots(3, 2, figsize=(10,10))

sns.histplot(ax=axs[0,0],data= trip_aggregated_df['trip_segment_actual_time_sum'],kde=True)
sns.boxplot(ax= axs[0,1],data= trip_aggregated_df['trip_segment_actual_time_sum'])


sns.histplot(ax=axs[1,0],data= trip_aggregated_df['trip_segment_osrm_time_sum'],kde=True)
sns.boxplot(ax= axs[1,1],data= trip_aggregated_df['trip_segment_osrm_time_sum'])

sns.histplot(ax=axs[2,0],data= trip_aggregated_df['trip_segment_osrm_distance_sum'],kde=True)
sns.boxplot(ax= axs[2,1],data= trip_aggregated_df['trip_segment_osrm_distance_sum'])


plt.show()
```







In [116…
```
import warnings
warnings.filterwarnings("ignore")
fig, axs = plt.subplots(ncols = 2, figsize=(10,5))
sns.histplot(ax=axs[0],data = trip_aggregated_df['trip_total_time_in_hrs'],kde=True)
sns.boxplot(ax=axs[1],data = trip_aggregated_df['trip_total_time_in_hrs'])
plt.show()
```



Hypothesis Testing

actual_time aggregated value and OSRM time aggregated value.

we will use ttest paired sample test to know if there is significant difference in actual trip aggregated time and OSRM trip aggregated time for each trip

HO : mean Actual time to deliver package from source to destination is lesser than OSRM time for entire trip

HA: mean Actual time to deliver package from source to destination is greater than OSRM time

In [117…
```
stat, pval=spy.ttest_rel(trip_aggregated_df['trip_actual_time_sum'],
                         trip_aggregated_df['trip_osrm_time_sum'],
                         alternative='greater')
print(f"stat {stat} pval {pval}")
```

```
stat 112.89026761644506 pval 0.0
```

In [118…
```
if pval <0.05:
    print("We will reject H0 :")
    print("mean Actual time to deliver package from source to destination is greater than OSRM time")
else:
    print("We fail to  reject H0 :")
    print("mean Actual time to deliver package from source to destination is lesser or equal  than OSRM time")
```

```
We will reject H0 :
mean Actual time to deliver package from source to destination is greater than OSRM time
```

In [119…
```
np.mean(trip_aggregated_df['trip_actual_time_sum']), np.mean(trip_aggregated_df['trip_osrm_time_sum'])
```

Out[119…
```
(262.29289240549133, 114.49563806045852)
```

actual_time aggregated value and segment actual time aggregated value

we will use ttest paired sample test to know if there is significant difference in actual trip aggregated time and segment actual time aggregated value for each trip

HO : mean Actual aggregated trip time to deliver package from source to destination is lesser than segment actual time aggregated value for entire trip

HA: mean Actual aggregated trip time to deliver package from source to destination is greater than segment actual time aggregated value for entire trip

In [120…
```
stat, pval=spy.ttest_rel(trip_aggregated_df['trip_actual_time_sum'],
                         trip_aggregated_df['trip_segment_actual_time_sum'],
                         alternative='greater')
print(f"stat {stat} pval {pval}")
```

```
stat 122.11851987195247 pval 0.0
```

In [121…
```
if pval <0.05:
    print("We will reject H0 :")
    print("mean Actual trip aggregated time   to deliver package from source to destination is greater than
segment actual aggregated time")
else:
    print("We fail to  reject H0 :")
    print("mean Actual trip aggregated time to deliver package from source to destination is lesser or equal
than segment actual aggregated time")
```

```
We will reject H0 :
mean Actual trip aggregated time   to deliver package from source to destination is greater than segment actual
aggregated time
```

OSRM distance aggregated value and segment OSRM distance aggregated value.

we will use ttest paired sample test to know if there is significant difference in OSRM distance aggregated value and segment actual aggregated distance for each trip

HO : mean Actual aggregated OSRM distance for trip to deliver package from source to destination is lesser than segment actual OSRM distance aggregated value for entire trip

HA: mean Actual aggregated OSRM distance for trip to deliver package from source to destination is greater than segment actual OSRM distance aggregated value for entire trip

In [122…
```
stat, pval=spy.ttest_rel(trip_aggregated_df['trip_osrm_distance_sum'],
                         trip_aggregated_df['trip_segment_osrm_distance_sum'],
                         alternative='greater')
print(f"stat {stat} pval {pval}")
```

```
stat -50.07621180430228 pval 1.0
```

In [123…
```
if pval <0.05:
    print("We will reject H0 :")
    print("mean Actual trip OSRM distance for trip  to deliver package from source to destination is greater
than segment actual aggregated OSRM distance for trip")
else:
    print("We fail to  reject H0 :")
    print("mean Actual trip OSRM distance for trip to deliver package from source to destination is lesser or
equal  than segment actual aggregated OSRM distance for trip")
```

```
We fail to  reject H0 :
mean Actual trip OSRM distance for trip to deliver package from source to destination is lesser or equal  than s
egment actual aggregated OSRM distance for trip
```

In [124…
```
trip_aggregated_df
```

Out[124…

	trip_uuid	trip_actual_time_sum	trip_osrm_time_sum	trip_osrm_distance_sum	trip_segment_actual_time_sum	tri
0	trip-153671041653548748	817.0	376.5	470.47515	811.0	
1	trip-153671042288605164	143.0	68.0	85.11100	141.0	
2	trip-153671043369099517	817.0	376.5	470.47515	811.0	
3	trip-153671046011330457	59.0	15.0	19.68000	59.0	
4	trip-153671052974046625	341.0	117.0	146.79180	340.0	
...	...	...	...	...	...	
14782	trip-153861095625827784	83.0	62.0	73.46300	82.0	
14783	trip-153861104386292051	21.0	12.0	16.08820	21.0	
14784	trip-153861106442901555	282.0	48.0	58.90370	281.0	
14785	trip-153861115439069069	264.0	179.0	171.11030	258.0	
14786	trip-153861118270144424	275.0	68.0	80.57870	274.0	

14787 rows × 8 columns

In [125…
```
np.max(trip_aggregated_df['trip_actual_time_sum']), np.max(trip_aggregated_df['trip_segment_actual_time_sum'])
```

Out[125…
```
(817.0, 811.0)
```

OSRM time aggregated value and segment OSRM time aggregated value.

we will use ttest paired sample test to know if there is significant difference in OSRM time aggregated value and segment OSRM aggregated time for each trip

HO : mean Actual aggregated OSRM time aggreated for trip to deliver package from source to destination is lesser than segment OSRM aggregated time value for entire trip

HA: mean Actual aggregated OSRM time aggregated for trip to deliver package from source to destination is greater than segment OSRM aggregated time value for entire trip

In [126…
```
stat, pval=spy.ttest_rel(trip_aggregated_df['trip_osrm_time_sum'],
                         trip_aggregated_df['trip_segment_osrm_time_sum'],
                         alternative='greater')
print(f"stat {stat} pval {pval}")
```

```
stat -63.41875343116358 pval 1.0
```

In [127…
```
if pval <0.05:
    print("We will reject H0 :")
    print("mean Actual trip aggregated  OSRM time  for trip  to deliver package from source to destination is
greater than segment actual aggregated OSRM time for trip")
else:
    print("We fail to  reject H0 :")
    print("mean Actual trip aggregated OSRM time  for trip to deliver package from source to destination is
lesser or equal  than segment actual aggregated OSRM time for trip")
```

```
We fail to  reject H0 :
mean Actual trip aggregated OSRM time  for trip to deliver package from source to destination is lesser or equal
than segment actual aggregated OSRM time for trip
```

Business Insights & Recommendations

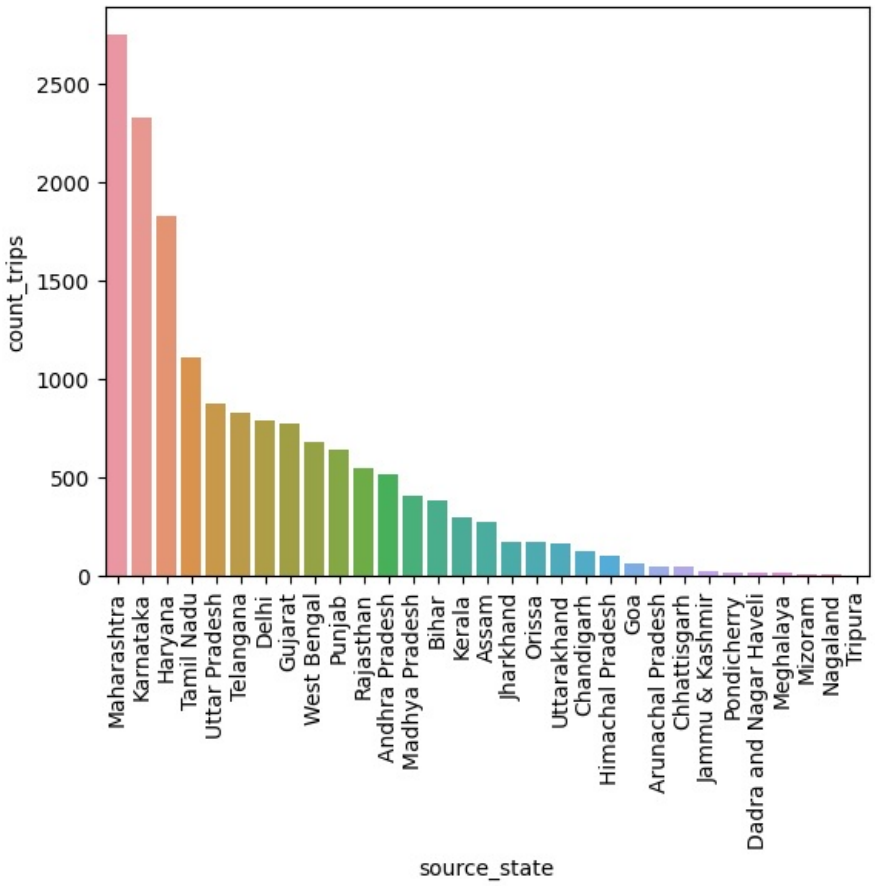
From Where the Most Orders are coming from

In [128…
```
orders_from_df = df.groupby('source_state').agg(count_trips=('trip_uuid','nunique')).reset_index()
orders_from_df = orders_from_df.sort_values(by='count_trips', ascending=False)
orders_from_df
```

Out[128…

	source_state	count_trips
17	Maharashtra	2748
14	Karnataka	2324
10	Haryana	1824
25	Tamil Nadu	1109
28	Uttar Pradesh	873
26	Telangana	825
7	Delhi	790
9	Gujarat	774
30	West Bengal	682
23	Punjab	643
24	Rajasthan	543
0	Andhra Pradesh	516
16	Madhya Pradesh	409
3	Bihar	382
15	Kerala	297
2	Assam	273
13	Jharkhand	175
21	Orissa	170
29	Uttarakhand	164
4	Chandigarh	123
11	Himachal Pradesh	103
8	Goa	65
1	Arunachal Pradesh	44
5	Chhattisgarh	43
12	Jammu & Kashmir	24
22	Pondicherry	19
6	Dadra and Nagar Haveli	15
18	Meghalaya	12
19	Mizoram	5
20	Nagaland	5
27	Tripura	1

In [129…
```
sns.barplot(data=orders_from_df, ='source_state', ='count_trips')
plt.xticks(rotation=90)
plt.show()
```



To Which State Most orders are going

In [130…

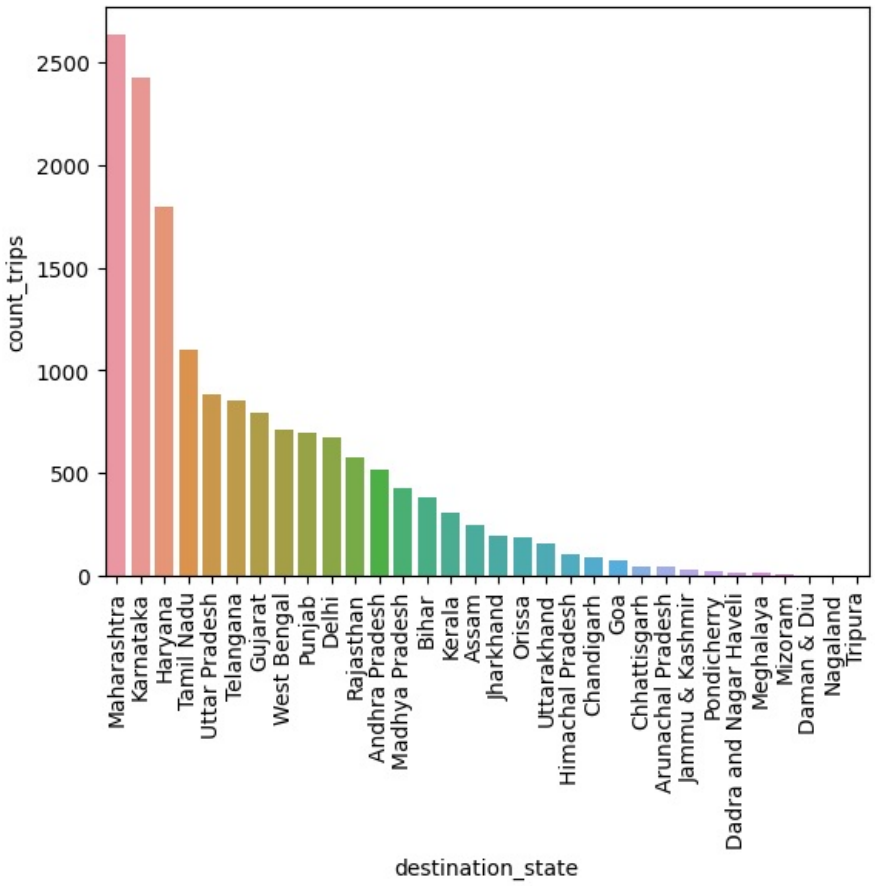
```
orders_to_df = df.groupby('destination_state').agg(count_trips=('trip_uuid','nunique')).reset_index()
orders_to_df = orders_to_df.sort_values(by='count_trips', ascending=False)
orders_to_df
```

Out[130…

	destination_state	count_trips
18	Maharashtra	2637
15	Karnataka	2425
11	Haryana	1800
26	Tamil Nadu	1097
29	Uttar Pradesh	882
27	Telangana	856
10	Gujarat	791
31	West Bengal	713
24	Punjab	693
8	Delhi	674
25	Rajasthan	574
0	Andhra Pradesh	516
17	Madhya Pradesh	423
3	Bihar	384
16	Kerala	303
2	Assam	249
14	Jharkhand	197
22	Orissa	187
30	Uttarakhand	159
12	Himachal Pradesh	101
4	Chandigarh	91
9	Goa	74
5	Chhattisgarh	43
1	Arunachal Pradesh	42
13	Jammu & Kashmir	25
23	Pondicherry	24
6	Dadra and Nagar Haveli	17
19	Meghalaya	11
20	Mizoram	7
7	Daman & Diu	1
21	Nagaland	1
28	Tripura	1

In [131…

```
sns.barplot(data=orders_to_df,x='destination_state',y='count_trips')
plt.xticks(rotation=90)
plt.show()
```



Most Busiest Corridor

In [132…
```
df['corridor']= df.apply(lambda x:"#".join([x['source_center'], ['destination_center']]), axis=1)
df
```

Out[132…

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name
0	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)
1	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)
2	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)
3	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)
4	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)
...	...	...	...	...	...	...	...
144311	test	2018-10-03 23:59:14.390954	thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...	Carting	trip-153861115439069069	IND628801AAA	Eral_Busstand_D (Tamil Nadu)
144312	test	2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	FTL	trip-153861118270144424	IND583119AAA	Sandur_WrdN1DPP_D (Karnataka)
144313	test	2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	FTL	trip-153861118270144424	IND583119AAA	Sandur_WrdN1DPP_D (Karnataka)
144314	test	2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	FTL	trip-153861118270144424	IND583201AAA	Hospet (Karnataka)
144315	test	2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	FTL	trip-153861118270144424	IND583201AAA	Hospet (Karnataka)

144316 rows × 51 columns

In [134…
```
corridor_total_trips = df.groupby(['trip_uuid','corridor']).nth(-1).groupby(['corridor']).agg(total_trips =
```

```
('trip_uuid','nunique')).reset_index()
corridor_total_trips
```

Out[134…

	corridor	total_trips
0	IND000000AAL#IND411033AAA	18
1	IND000000AAQ#IND700028AAB	2
2	IND000000AAS#IND783370AAC	9
3	IND000000AAZ#IND444203AAA	1
4	IND000000AAZ#IND444303AAA	1
...	...	...
2736	IND854326AAB#IND854334AAA	1
2737	IND854334AAA#IND852118AAA	7
2738	IND854334AAA#IND854335AAA	2
2739	IND854335AAA#IND852111AAA	17
2740	IND854335AAA#IND854326AAB	1

2741 rows × 2 columns

In [135…

```
corridor_actual_time_mean_df = df.groupby(['trip_uuid','corridor']).nth(-1).groupby(['corridor']).agg(corridor_ac
('segment_actual_time_cumsum','mean')).reset_index()
corridor_actual_time_mean_df
```

Out[135…

	corridor	corridor_actual_time_mean
0	IND000000AAL#IND411033AAA	87.388889
1	IND000000AAQ#IND700028AAB	84.500000
2	IND000000AAS#IND783370AAC	61.000000
3	IND000000AAZ#IND444203AAA	287.000000
4	IND000000AAZ#IND444303AAA	159.000000
...	...	...
2736	IND854326AAB#IND854334AAA	171.000000
2737	IND854334AAA#IND852118AAA	28.285714
2738	IND854334AAA#IND854335AAA	40.500000
2739	IND854335AAA#IND852111AAA	39.470588
2740	IND854335AAA#IND854326AAB	197.000000

2741 rows × 2 columns

In [136…

```
corridor_osrm_time_mean_df = df.groupby(['trip_uuid','corridor']).nth(-1).groupby(['corridor']).agg(corridor_dsrm
('segment_osrm_time_cumsum','mean')).reset_index()
corridor_osrm_time_mean_df
```

Out[136…

	corridor	corridor_dsrm_time_mean
0	IND000000AAL#IND411033AAA	29.777778
1	IND000000AAQ#IND700028AAB	14.000000
2	IND000000AAS#IND783370AAC	29.000000
3	IND000000AAZ#IND444203AAA	77.000000
4	IND000000AAZ#IND444303AAA	68.000000
...	...	...
2736	IND854326AAB#IND854334AAA	47.000000
2737	IND854334AAA#IND852118AAA	21.428571
2738	IND854334AAA#IND854335AAA	29.500000
2739	IND854335AAA#IND852111AAA	19.294118
2740	IND854335AAA#IND854326AAB	82.000000

2741 rows × 2 columns

In [137…

```
corridor_osrm_distance_mean_df = df.groupby(['trip_uuid','corridor']).nth(-1).groupby(['corridor']).agg(corridor
('segment_osrm_distance_cumsum','mean')).reset_index()
```

```
corridor_osrm_distance_mean_df
```

Out[137…

	corridor	corridor_osrm_distance_mean
0	IND000000AAL#IND411033AAA	28.885561
1	IND000000AAQ#IND700028AAB	13.900700
2	IND000000AAS#IND783370AAC	41.461622
3	IND000000AAZ#IND444203AAA	109.306700
4	IND000000AAZ#IND444303AAA	93.706900
...	...	...
2736	IND854326AAB#IND854334AAA	67.378600
2737	IND854334AAA#IND852118AAA	23.881371
2738	IND854334AAA#IND854335AAA	36.500750
2739	IND854335AAA#IND852111AAA	27.870394
2740	IND854335AAA#IND854326AAB	109.264800

2741 rows × 2 columns

In [138…

```
corridor_aggregated_df=pd.concat(
    objs=(
        iDF.set_index('corridor') for iDF in (corridor_total_trips,
                                              corridor_actual_time_mean_df,
                                              corridor_osrm_distance_mean_df,
                                              corridor_osrm_time_mean_df
                                             )
    ),
    axis=1,
    join='inner'
).reset_index()
corridor_aggregated_df
```

Out[138…

	corridor	total_trips	corridor_actual_time_mean	corridor_osrm_distance_mean	corridor_dsrm_time_mean
0	IND000000AAL#IND411033AAA	18	87.388889	28.885561	29.777778
1	IND000000AAQ#IND700028AAB	2	84.500000	13.900700	14.000000
2	IND000000AAS#IND783370AAC	9	61.000000	41.461622	29.000000
3	IND000000AAZ#IND444203AAA	1	287.000000	109.306700	77.000000
4	IND000000AAZ#IND444303AAA	1	159.000000	93.706900	68.000000
...	...	...	...	...	...
2736	IND854326AAB#IND854334AAA	1	171.000000	67.378600	47.000000
2737	IND854334AAA#IND852118AAA	7	28.285714	23.881371	21.428571
2738	IND854334AAA#IND854335AAA	2	40.500000	36.500750	29.500000
2739	IND854335AAA#IND852111AAA	17	39.470588	27.870394	19.294118
2740	IND854335AAA#IND854326AAB	1	197.000000	109.264800	82.000000

2741 rows × 5 columns

Busiest Corridor By Trips

In [139…

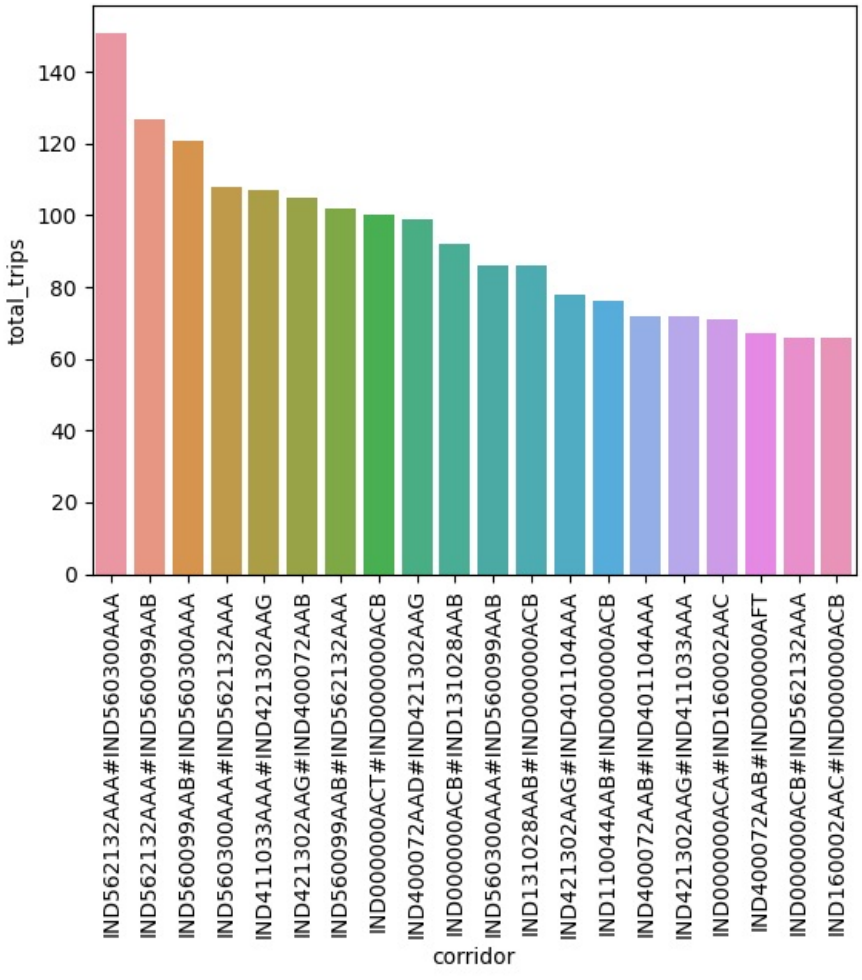
```
top_20Trips_corridor_df=corridor_aggregated_df[['corridor','total_trips']].sort_values(by='total_trips',
ascending=False)[:20]
top_20Trips_corridor_df
```

Out[139…

	corridor	total_trips
1743	IND562132AAA#IND560300AAA	151
1742	IND562132AAA#IND560099AAB	127
1687	IND560099AAB#IND560300AAA	121
1703	IND560300AAA#IND562132AAA	108
1059	IND411033AAA#IND421302AAG	107
1131	IND421302AAG#IND400072AAB	105
1688	IND560099AAB#IND562132AAA	102
66	IND000000ACT#IND000000ACB	100
989	IND400072AAD#IND421302AAG	99
37	IND000000ACB#IND131028AAB	92
1702	IND560300AAA#IND560099AAB	86
205	IND131028AAB#IND000000ACB	86
1137	IND421302AAG#IND401104AAA	78
121	IND110044AAB#IND000000ACB	76
982	IND400072AAB#IND401104AAA	72
1140	IND421302AAG#IND411033AAA	72
16	IND000000ACA#IND160002AAC	71
977	IND400072AAB#IND000000AFT	67
57	IND000000ACB#IND562132AAA	66
322	IND160002AAC#IND000000ACB	66

In [141…

```
sns.barplot(data=top_20Trips_corridor_df, ='corridor', ='total_trips')
plt.xticks(rotation=90)
plt.show()
```



Busiest corridor Actual time

In [142…

```
top_20trip_time_corridor_df = corridor_aggregated_df[['corridor','corridor_actual_time_mean']].sort_values(by='co
  ascending=False)[:20]
```

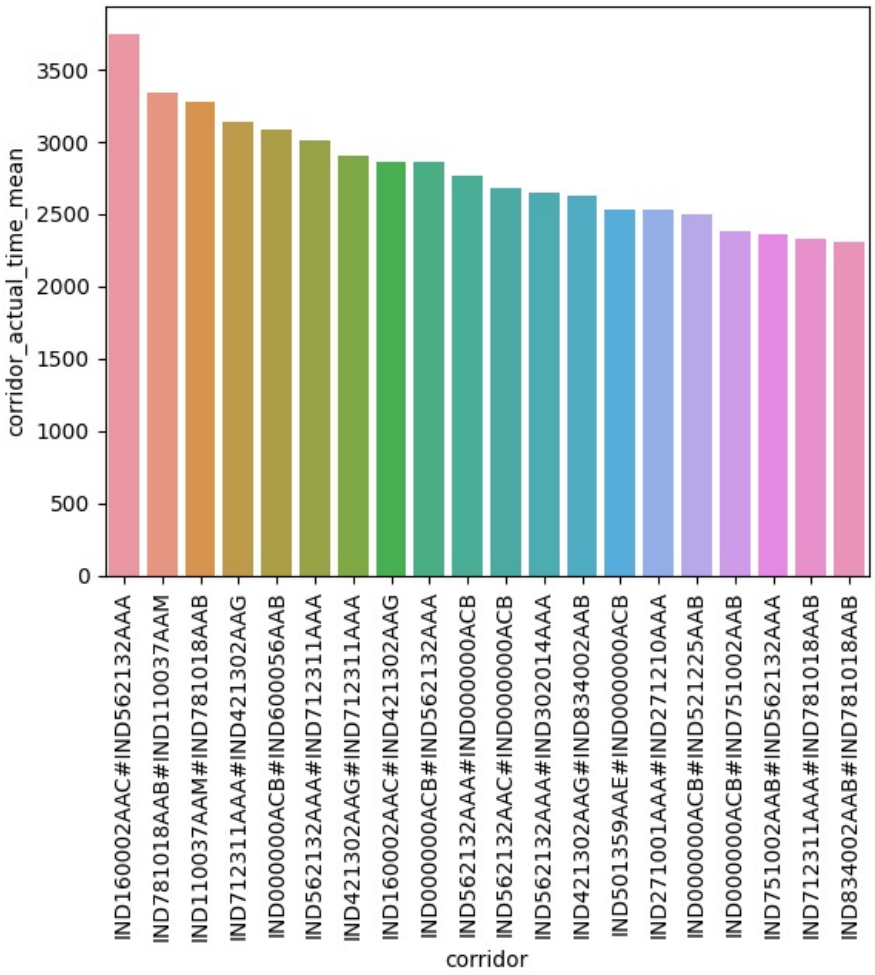
```
top_20trip_time_corridor_df
```

Out[142…

	corridor	corridor_actual_time_mean
349	IND160002AAC#IND562132AAA	3751.000000
2475	IND781018AAB#IND110037AAM	3341.764706
115	IND110037AAM#IND781018AAB	3281.000000
2240	IND712311AAA#IND421302AAG	3141.200000
58	IND000000ACB#IND600056AAB	3090.857143
1754	IND562132AAA#IND712311AAA	3010.333333
1148	IND421302AAG#IND712311AAA	2902.000000
348	IND160002AAC#IND421302AAG	2867.000000
57	IND000000ACB#IND562132AAA	2864.136364
1721	IND562132AAA#IND000000ACB	2766.454545
1756	IND562132AAC#IND000000ACB	2683.000000
1723	IND562132AAA#IND302014AAA	2645.571429
1149	IND421302AAG#IND834002AAB	2627.250000
1381	IND501359AAE#IND000000ACB	2536.000000
608	IND271001AAA#IND271210AAA	2533.000000
56	IND000000ACB#IND521225AAB	2503.272727
60	IND000000ACB#IND751002AAB	2387.466667
2400	IND751002AAB#IND562132AAA	2363.615385
2257	IND712311AAA#IND781018AAB	2331.625000
2645	IND834002AAB#IND781018AAB	2308.000000

In [143…

```
sns.barplot(data=top_20trip_time_corridor_df, x='corridor',y='corridor_actual_time_mean')
plt.xticks(rotation=90)
plt.show()
```



In [144…

```
corridor_aggregated_df[['corridor_actual_time_mean','corridor_osrm_distance_mean','corridor_dsrm_time_mean','to
```

Out[144…

	corridor_actual_time_mean	corridor_osrm_distance_mean	corridor_dsrm_time_mean	total_trips
corridor_actual_time_mean	1.000000	0.926574	0.921375	0.018440
corridor_osrm_distance_mean	0.926574	1.000000	0.995586	0.050239
corridor_dsrm_time_mean	0.921375	0.995586	1.000000	0.050958
total_trips	0.018440	0.050239	0.050958	1.000000

Busiest Corridor by Distance

In [146…

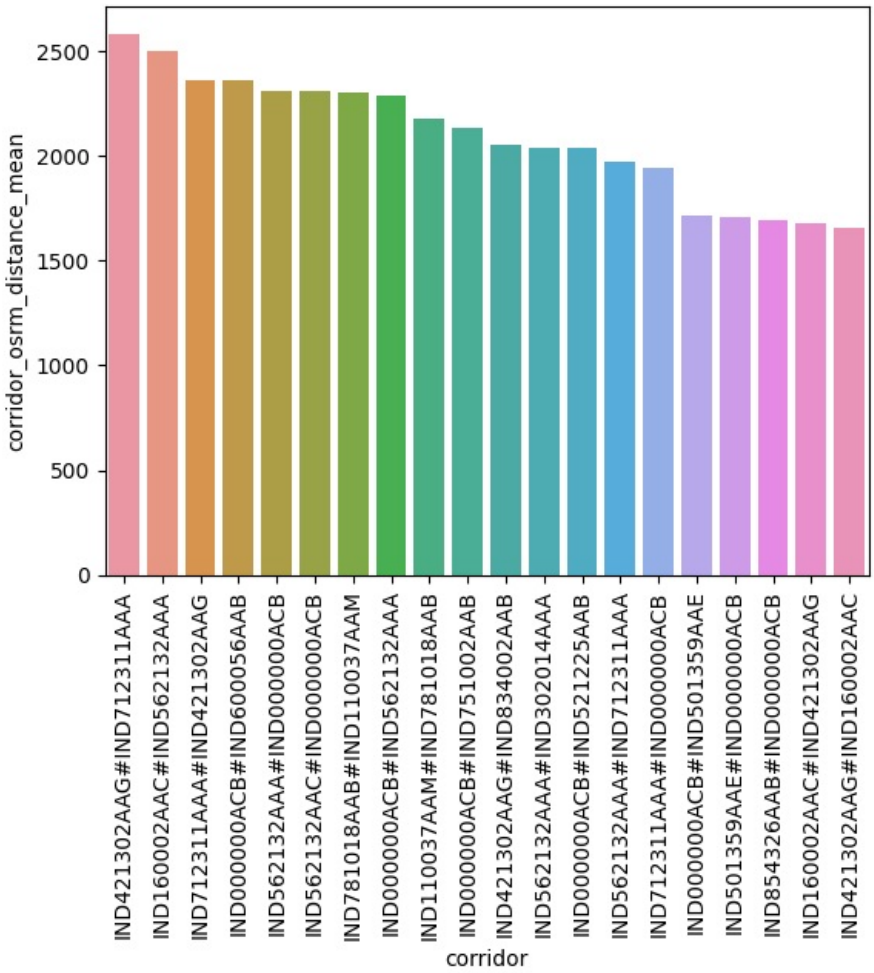
```
top_20trip_dist_corridor_df = corridor_aggregated_df[['corridor','corridor_osrm_distance_mean']].sort_values(by='
 ascending=False)[:20]
top_20trip_dist_corridor_df
```

Out[146…

	corridor	corridor_osrm_distance_mean
1148	IND421302AAG#IND712311AAA	2584.622933
349	IND160002AAC#IND562132AAA	2500.214500
2240	IND712311AAA#IND421302AAG	2363.329580
58	IND000000ACB#IND600056AAB	2361.555264
1721	IND562132AAA#IND000000ACB	2312.589602
1756	IND562132AAC#IND000000ACB	2307.137400
2475	IND781018AAB#IND110037AAM	2300.517159
57	IND000000ACB#IND562132AAA	2288.400620
115	IND110037AAM#IND781018AAB	2181.460700
60	IND000000ACB#IND751002AAB	2134.451707
1149	IND421302AAG#IND834002AAB	2053.603325
1723	IND562132AAA#IND302014AAA	2037.975464
56	IND000000ACB#IND521225AAB	2037.931427
1754	IND562132AAA#IND712311AAA	1972.076000
2239	IND712311AAA#IND000000ACB	1941.393782
55	IND000000ACB#IND501359AAE	1712.413952
1381	IND501359AAE#IND000000ACB	1708.819400
2729	IND854326AAB#IND000000ACB	1691.497550
348	IND160002AAC#IND421302AAG	1679.658000
1125	IND421302AAG#IND160002AAC	1656.298800

In [147…

```
sns.barplot(data=top_20trip_dist_corridor_df, x='corridor',y='corridor_osrm_distance_mean')
plt.xticks(rotation=90)
plt.show()
```



Business Recommendations

- Maharashtra, Karnataka, Tamil Nadu, Utter Pradesh, Telangana and Gujarat States are states where most delivery trips are done. Most Bussiest corridor are in these states.
- Business should focus on identifying best corridors to move packages very quickly, they should focus on potential reasons for difference in actual delivery time and osrm delivery time value.
- If Actual delivery time is higher than osrm time then should focus on hops which are causing delays, if delays are related to processing or logistic that should be quickly fixed.
- If Issue is not related to delivery and logistic process then should focus on identifying best route to move packages quickly.

In [ ]: