Delhivery - Feature Engineering

In [137…
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from scipy.stats import chisquare, chi2, chi2_contingency
from scipy.stats import t, norm, ttest_1samp
from scipy.stats import ttest_ind
from scipy.stats import f_oneway
from scipy.stats import kruskal
from scipy.stats import pearsonr ,spearmanr
from scipy import stats
import statsmodels.api as sm
import statistics
from scipy.stats import poisson, binom
from scipy.stats import levene
from sklearn.preprocessing import StandardScaler, MinMaxScaler
```

Define Problem Statement and perform Exploratory Data Analysis

Analyze the structure of the data

In [73]:
```
df = pd.read_csv("delhivery.csv")
data = pd.read_csv("delhivery.csv")
```

In [6]:
```
df.head(10)
```

Out[6]:

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name	destination_center	d
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	Khamb
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	Khamb
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	Khamb
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	Khamb
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	Khamb
5	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388620AAB	Khambhat_MotvdDPP_D (Gujarat)	IND388320AAA	A
6	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388620AAB	Khambhat_MotvdDPP_D (Gujarat)	IND388320AAA	A
7	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388620AAB	Khambhat_MotvdDPP_D (Gujarat)	IND388320AAA	A
8	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388620AAB	Khambhat_MotvdDPP_D (Gujarat)	IND388320AAA	A
9	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388620AAB	Khambhat_MotvdDPP_D (Gujarat)	IND388320AAA	A

10 rows × 24 columns

Definition of problem

The requirement is 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
- Identify if there is any meaningful relationships among these features and and if so how they are related
- Create new features out of this raw data which is more releveant for forecasting models
- Standardize the data where necessary

Observations on shape of data, data types of all the attributes, conversion of categorical attributes to 'category' (If required), missing value detection, statistical summary.

In [8]:
```
df.shape
```

Out[8]:
```
(144867, 24)
```

In [9]:
```
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 [10]:
```
df.isna().sum()
```

Out[10]:
```
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: As seen above there are some missing source_name and destination_name

In [12]:
```
df.describe() # statistical summary of numeric features
```

Out[12]:

	start_scan_to_end_scan	cutoff_factor	actual_distance_to_destination	actual_time	osrm_time	osrm_distance	factor	segment_actual
count	144867.000000	144867.000000	144867.000000	144867.000000	144867.000000	144867.000000	144867.000000	144867.0(
mean	961.262986	232.926567	234.073372	416.927527	213.868272	284.771297	2.120107	36.1
std	1037.012769	344.755577	344.990009	598.103621	308.011085	421.119294	1.715421	53.5
min	20.000000	9.000000	9.000045	9.000000	6.000000	9.008200	0.144000	-244.0(
25%	161.000000	22.000000	23.355874	51.000000	27.000000	29.914700	1.604264	20.0(
50%	449.000000	66.000000	66.126571	132.000000	64.000000	78.525800	1.857143	29.0(
75%	1634.000000	286.000000	286.708875	513.000000	257.000000	343.193250	2.213483	40.0(
max	7898.000000	1927.000000	1927.447705	4532.000000	1686.000000	2326.199100	77.387097	3051.0(

In [14]:
```
df.describe(include=['object']) # statistical summary of non-numeric values
```

Out[14]:

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name	destination_center	d
count	144867	144867	144867	144867	144867	144867	144574	144867	
unique	2	14817	1504	2	14817	1508	1498	1481	
top	training	2018-09-28 05:23:15.359220	thanos::sroute:4029a8a2-6c74-4b7e-a6d8-f9e069f...	FTL	trip-153811219535896559	IND000000ACB	Gurgaon_Bilaspur_HB (Haryana)	IND000000ACB	Gur
freq	104858	101	1812	99660	101	23347	23347	15192	

Observation: data and route_type are 2 categorical columns

In [18]:
```
df['data'].value_counts()
```

Out[18]:
```
training    104858
test         40009
Name: data, dtype: int64
```

In [19]:
```
df['route_type'].value_counts()
```

Out[19]:
```
FTL        99660
Carting    45207
Name: route_type, dtype: int64
```

Insight:

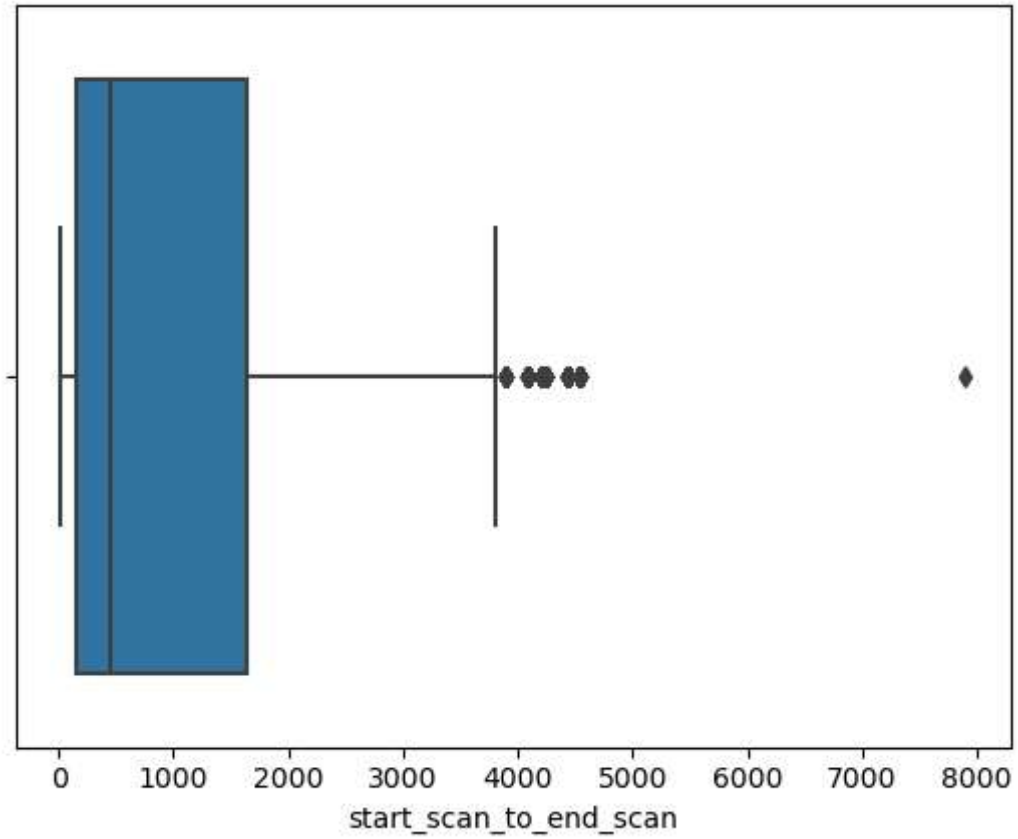
- 'data' column has 2 unique values 'training' and 'test'
- 'route_type' has 2 unique values 'FTL' and 'Carting'

Visual Analysis (distribution plots of all the continuous variable(s), boxplots of all the categorical variables)

range of attributes, outliers of various attributes

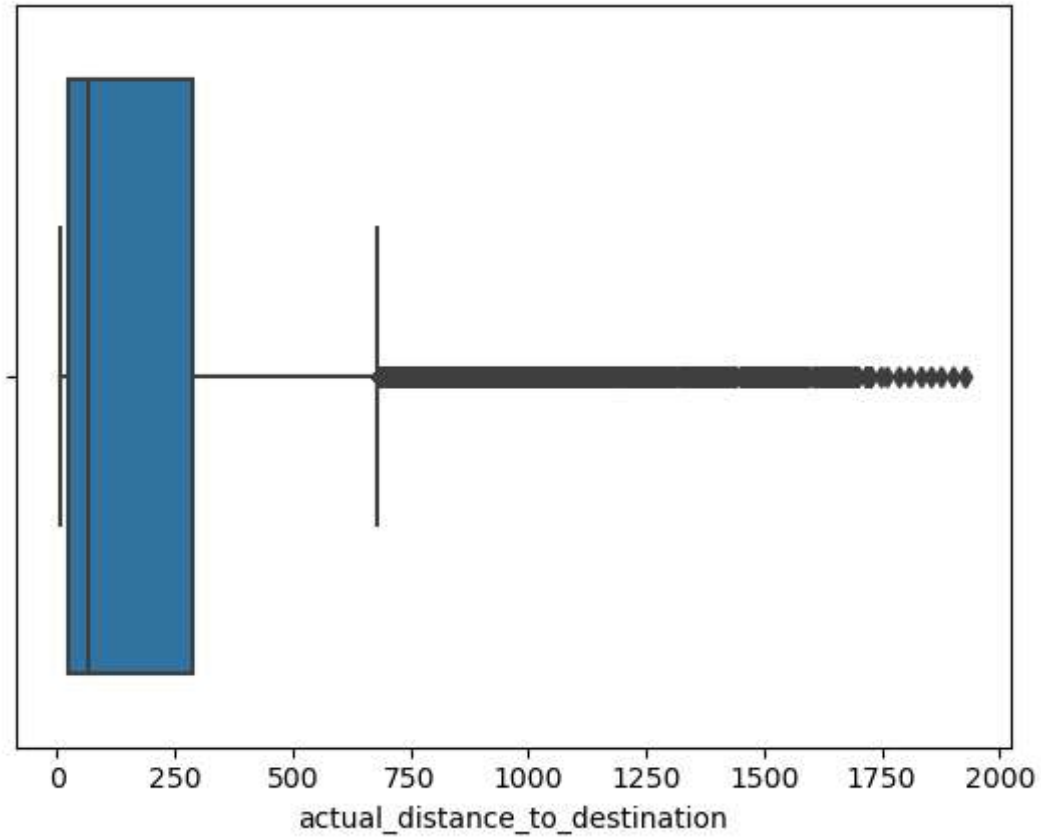
In [28]:
```
sns.boxplot(x=df["start_scan_to_end_scan"])
```

Out[28]:
```
<AxesSubplot:xlabel='start_scan_to_end_scan'>
```



In [29]:
```
sns.boxplot(x=df["actual_distance_to_destination"])
```

Out[29]:
```
<AxesSubplot:xlabel='actual_distance_to_destination'>
```

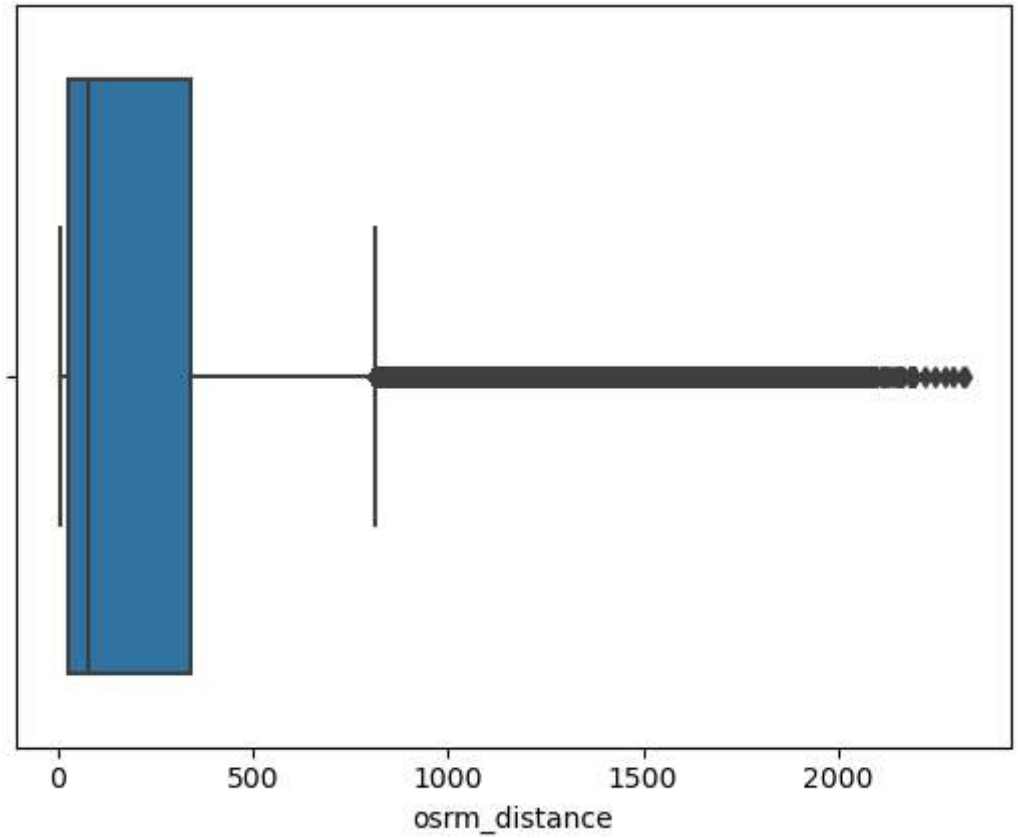


In [138…
```
df[df["actual_distance_to_destination"]>750].count()
```

Out[138]:
```
data                                14969
trip_creation_time                  14969
route_schedule_uuid                 14969
route_type                          14969
trip_uuid                           14969
source_center                       14969
source_name                         14969
destination_center                  14969
destination_name                    14969
od_start_time                       14969
od_end_time                         14969
start_scan_to_end_scan              14969
is_cutoff                           14969
cutoff_factor                       14969
cutoff_timestamp                    14969
actual_distance_to_destination      14969
actual_time                         14969
osrm_time                           14969
osrm_distance                       14969
factor                              14969
segment_actual_time                 14969
segment_osrm_time                   14969
segment_osrm_distance               14969
segment_factor                      14969
destination_city                    14969
destination_place                   14969
destination_code_state              13387
source_city                         14969
source_place                        14969
source_code_state                   13600
trip_month                          14969
trip_year                           14969
trip_day                            14969
total_time                          14969
dtype: int64
```

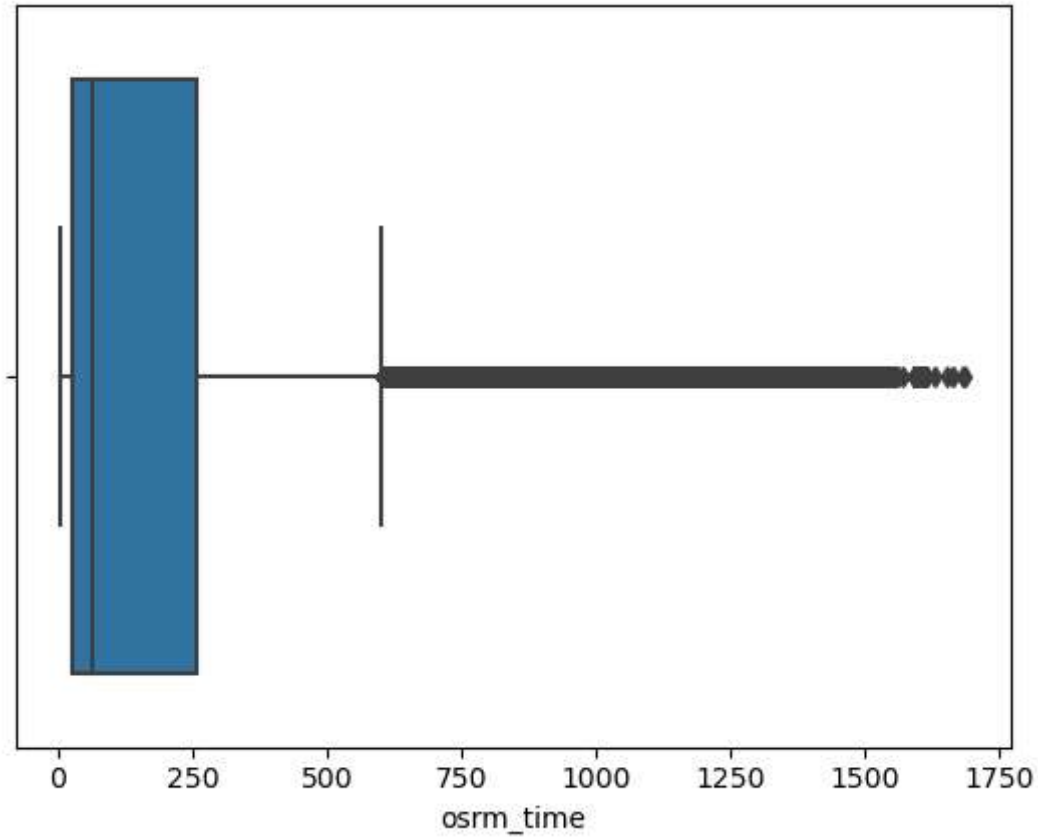
In [30]:
```
sns.boxplot(x=df["osrm_distance"])
```

Out[30]:
```
<AxesSubplot:xlabel='osrm_distance'>
```



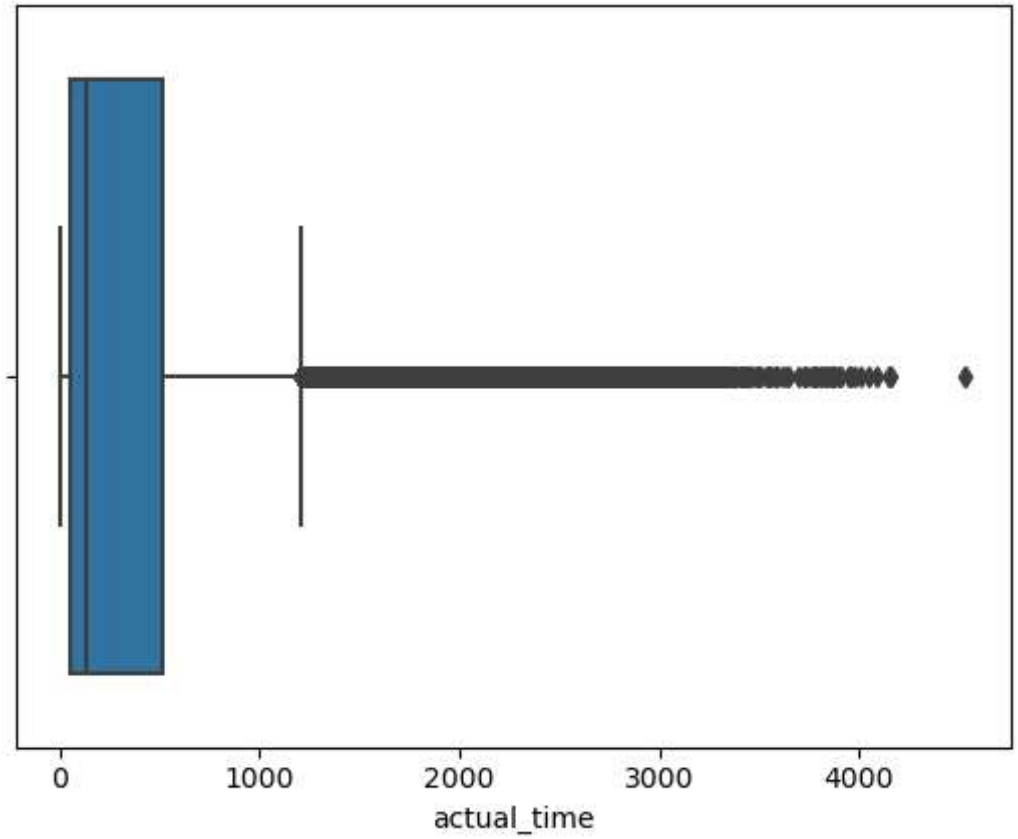
In [31]:
```
sns.boxplot(x=df["osrm_time"])
```

Out[31]:
```
<AxesSubplot:xlabel='osrm_time'>
```



In [32]:
```
sns.boxplot(x=df["actual_time"])
```

Out[32]:
```
<AxesSubplot:xlabel='actual_time'>
```



Insight: start_scan_to_end_scan seems to be the only one having distinct outlier.

distribution of the variables and relationship between them (univariate and bivariate plot)

In [22]:
```
df.describe()
```

Out[22]:

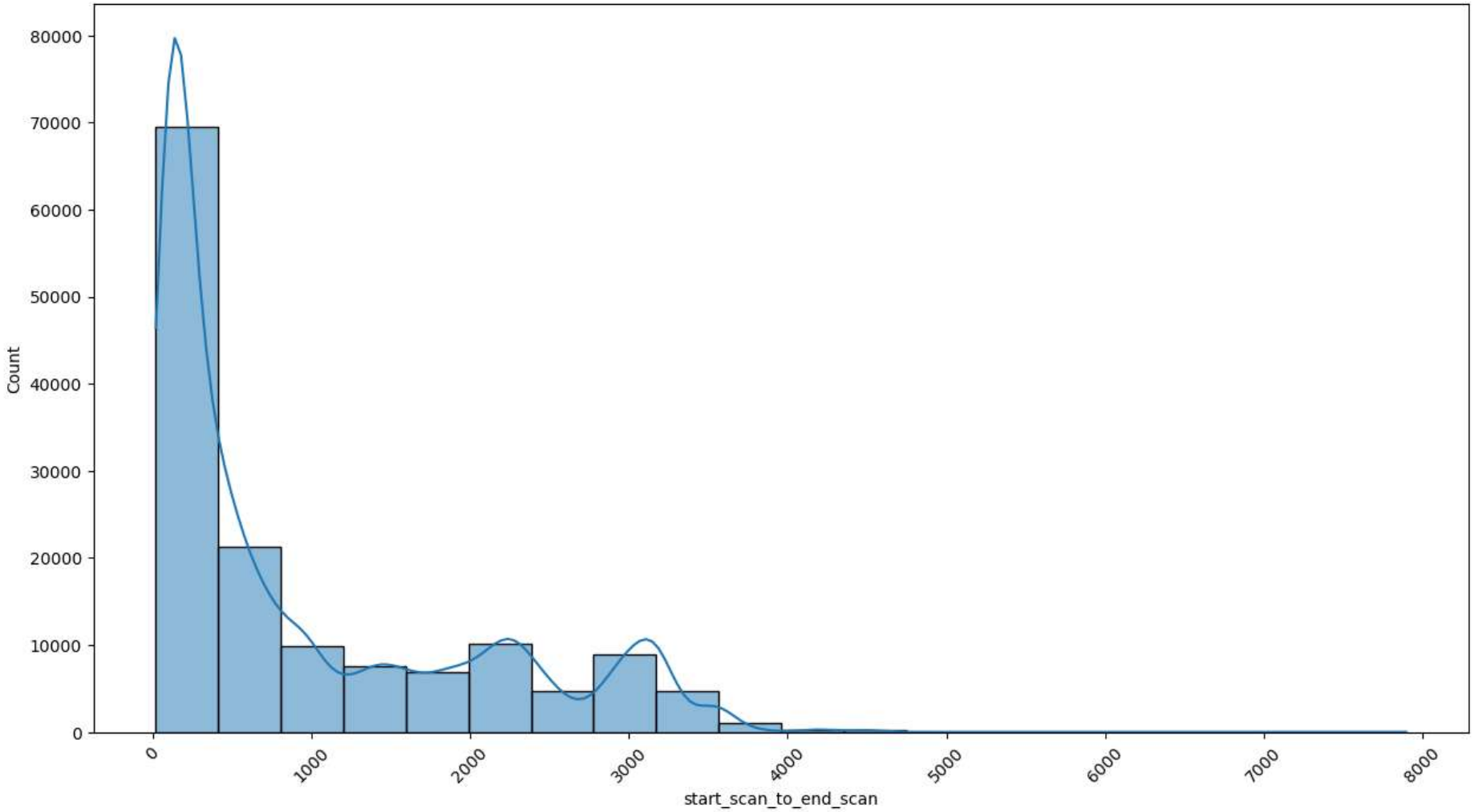
	start_scan_to_end_scan	cutoff_factor	actual_distance_to_destination	actual_time	osrm_time	osrm_distance	factor	segment_actual
count	144867.000000	144867.000000	144867.000000	144867.000000	144867.000000	144867.000000	144867.000000	144867.0(
mean	961.262986	232.926567	234.073372	416.927527	213.868272	284.771297	2.120107	36.1
std	1037.012769	344.755577	344.990009	598.103621	308.011085	421.119294	1.715421	53.5
min	20.000000	9.000000	9.000045	9.000000	6.000000	9.008200	0.144000	-244.0(
25%	161.000000	22.000000	23.355874	51.000000	27.000000	29.914700	1.604264	20.0(
50%	449.000000	66.000000	66.126571	132.000000	64.000000	78.525800	1.857143	29.0(
75%	1634.000000	286.000000	286.708875	513.000000	257.000000	343.193250	2.213483	40.0(
max	7898.000000	1927.000000	1927.447705	4532.000000	1686.000000	2326.199100	77.387097	3051.0(

In [23]:
```
df.describe(include=['object'])
```

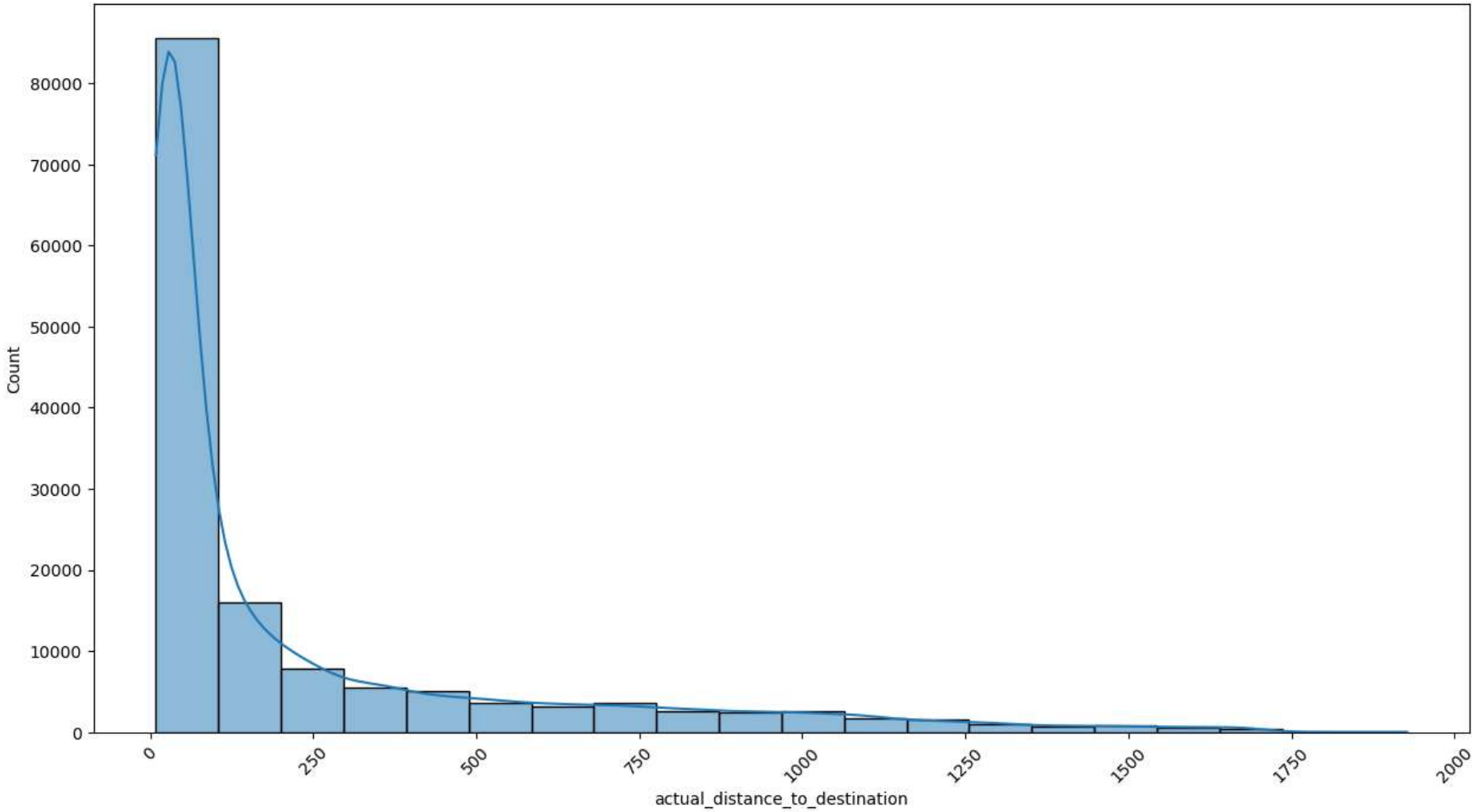
Out[23]:

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name	destination_center	d
count	144867	144867	144867	144867	144867	144867	144574	144867	
unique	2	14817	1504	2	14817	1508	1498	1481	
top	training	2018-09-28 05:23:15.359220	thanos::sroute:4029a8a2-6c74-4b7e-a6d8-f9e069f...	FTL	trip-153811219535896559	IND000000ACB	Gurgaon_Bilaspur_HB (Haryana)	IND000000ACB	Gur
freq	104858	101	1812	99660	101	23347	23347	15192	

In [36]:
```
plt.figure(figsize=(15,8))
plt.xticks(rotation=45)
sns.histplot(x="start_scan_to_end_scan", bins=20, data = df, kde=True)
plt.show()
```

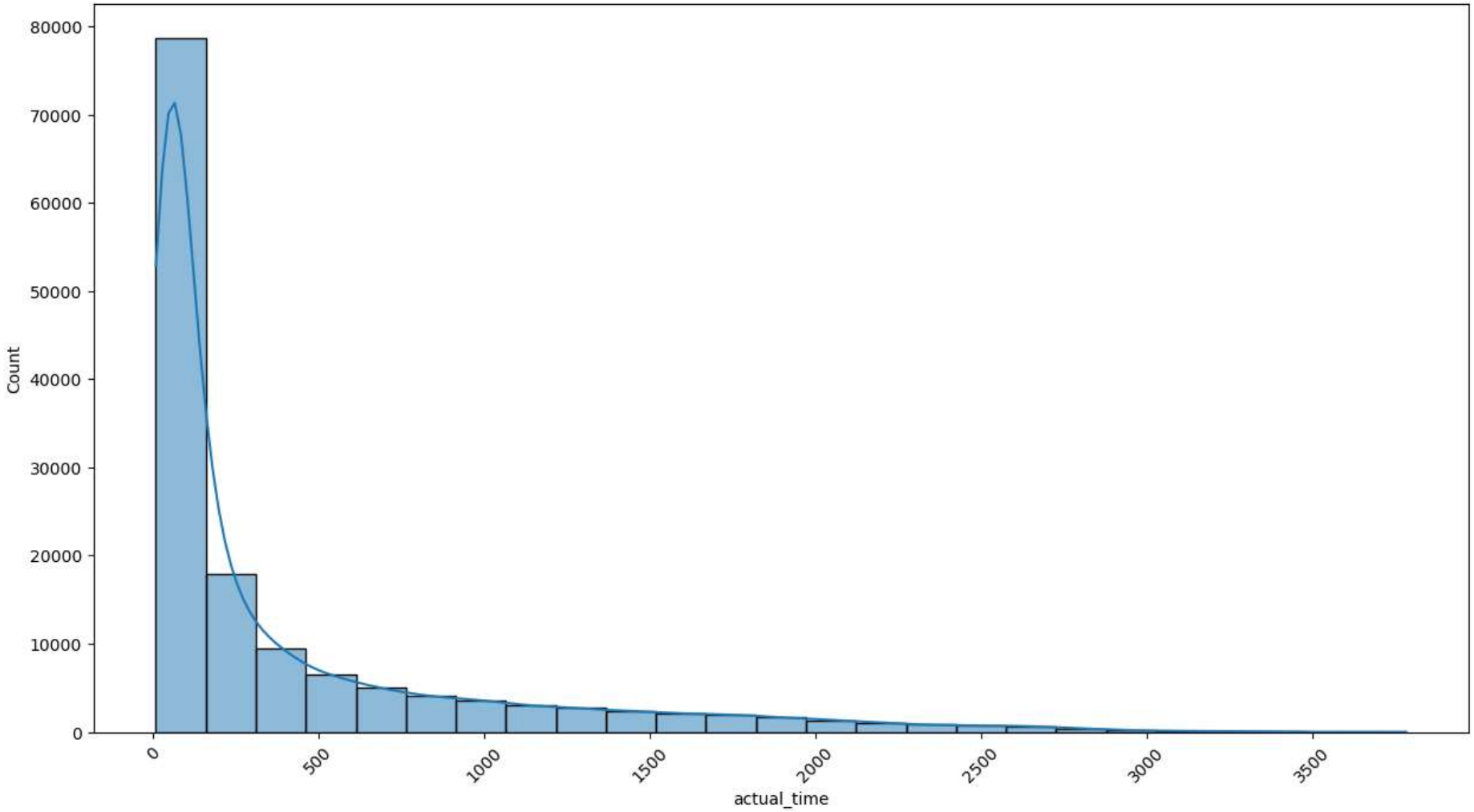


In [37]:
```
plt.figure(figsize=(15,8))
plt.xticks(rotation=45)
sns.histplot(x="actual_distance_to_destination", bins=20, data = df, kde=True)
plt.show()
```

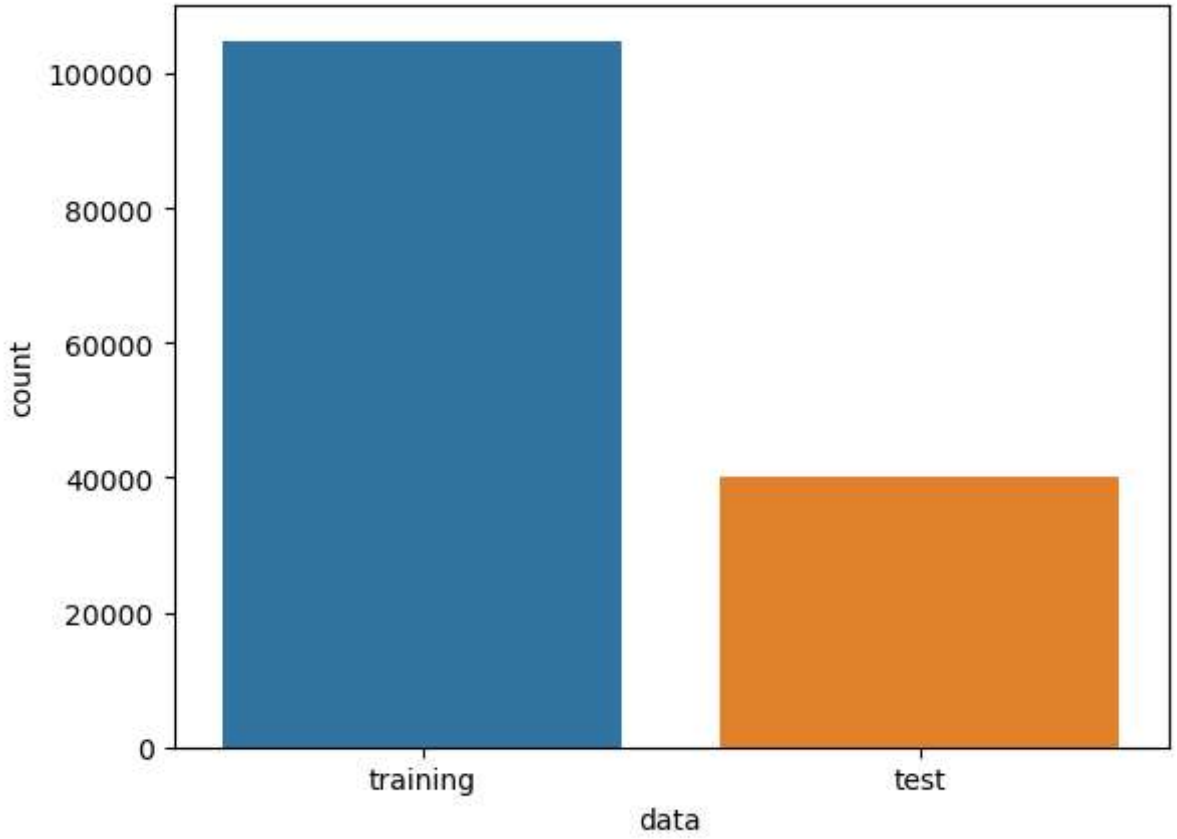


In [214…

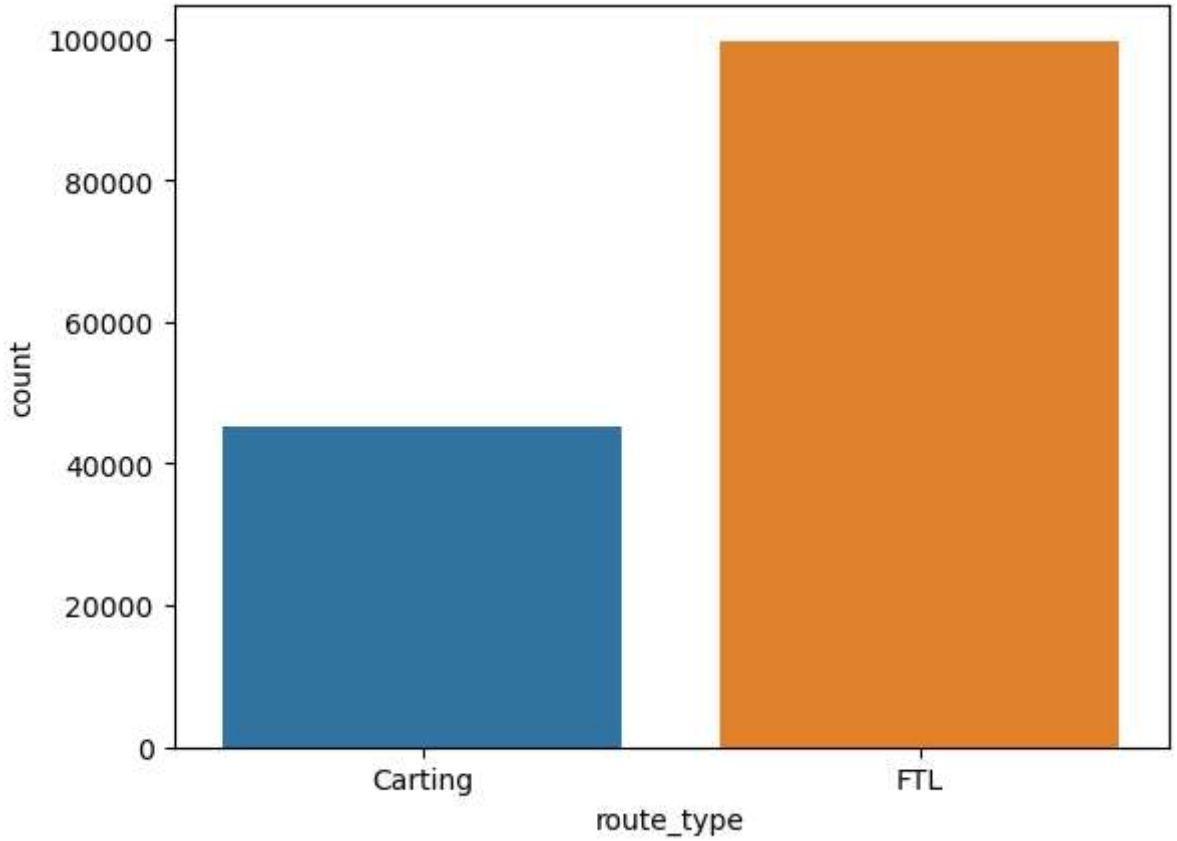
```
plt.figure(figsize=(15,8))
plt.xticks(rotation=45)
sns.histplot(x="actual_time", bins=25, data = df, kde=True)
plt.show()
```



In [40]:
```
sns.countplot(x ='data', data = df)
plt.show()
```

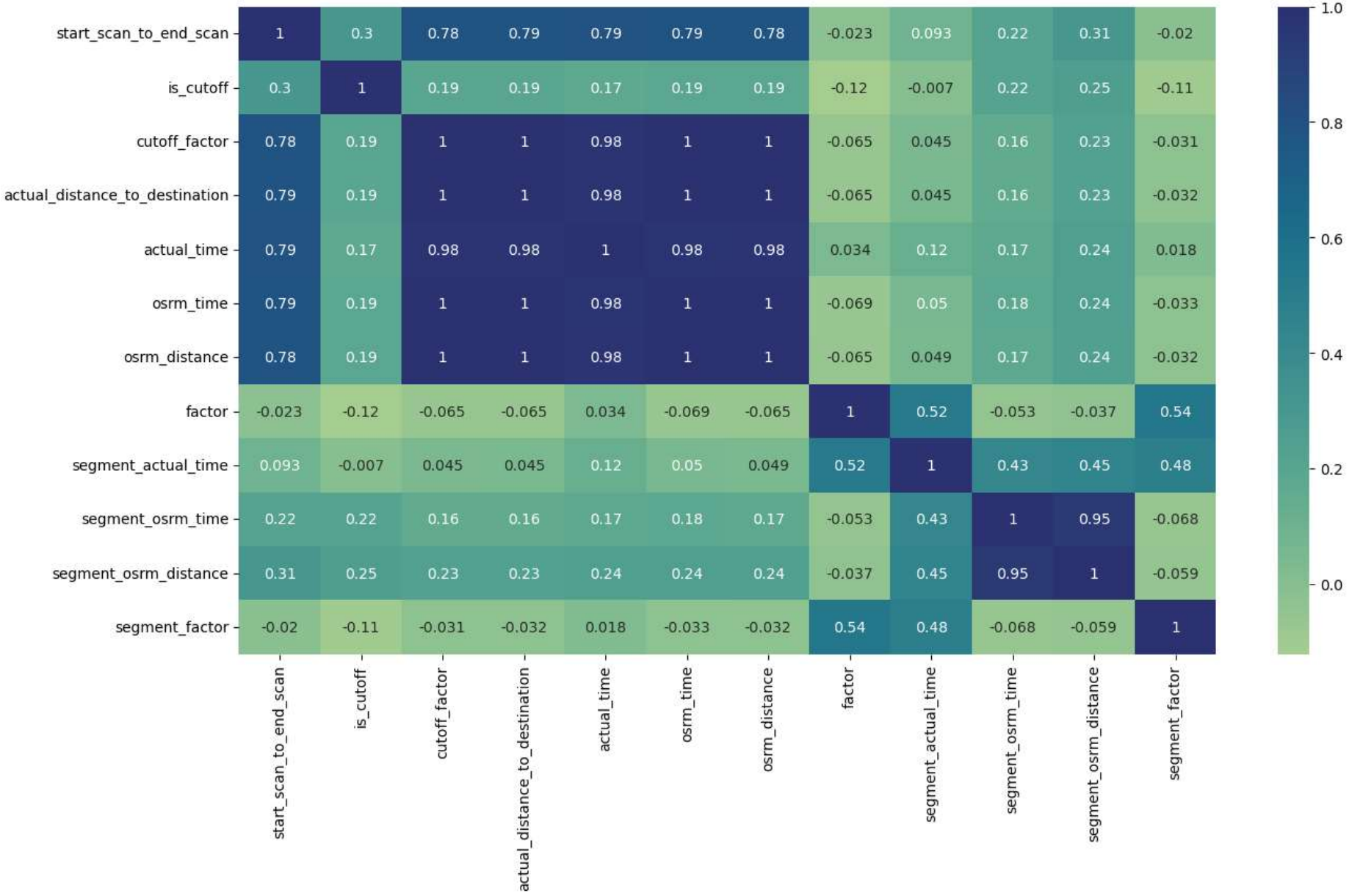


In [41]:
```
sns.countplot(x ='route_type', data = df)
plt.show()
```

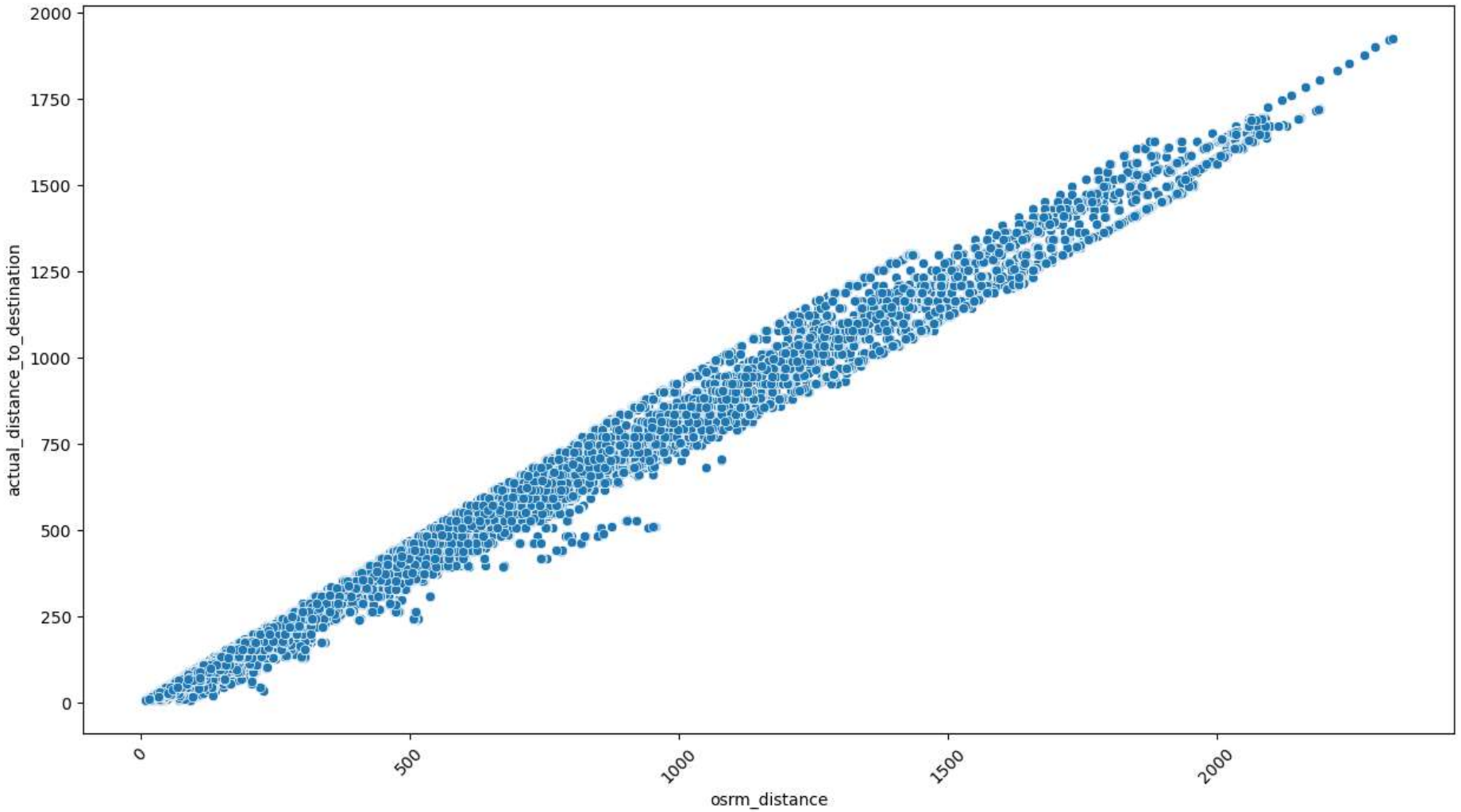


Insight : The amount of FTL is almost twice that of carting visualy. Also most of the data in time or distance are not normally distributed.

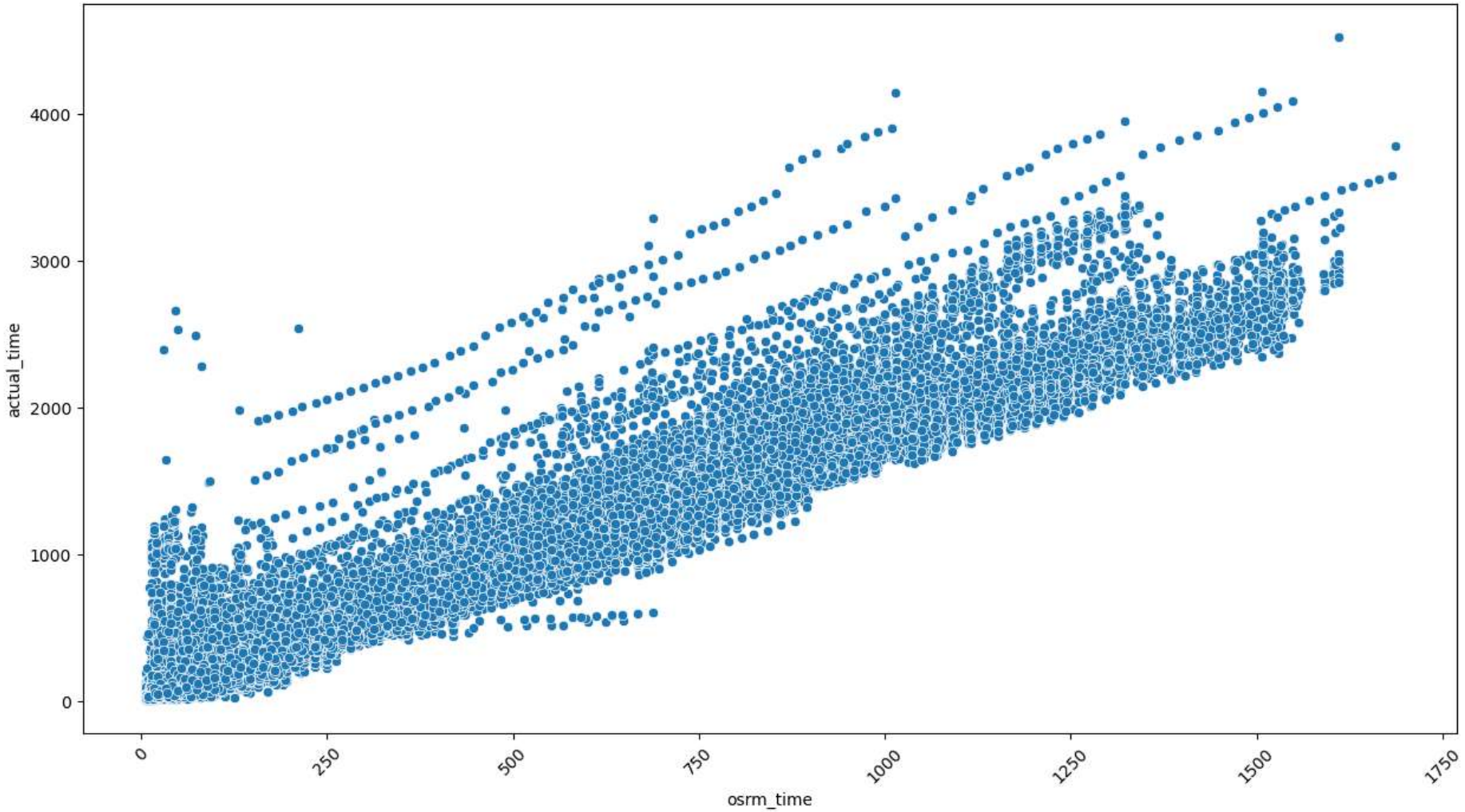
In [45]:
```
plt.figure(figsize=(15,8))
plt.xticks(rotation=45)
sns.heatmap(df.corr(),annot=True, cmap="crest")
plt.show()
```



In [49]:
```
plt.figure(figsize=(15,8))
plt.xticks(rotation=45)
sns.scatterplot(x='osrm_distance',y='actual_distance_to_destination',data=df)
plt.show()
```



In [50]:
```
plt.figure(figsize=(15,8))
plt.xticks(rotation=45)
sns.scatterplot(x='osrm_time',y='actual_time',data=df)
plt.show()
```



Insight: The accuracy of osrm_distance is higher than osrm time since its closer to a straight line

In [ ]:

Missing values Treatment & Outlier treatment (using the IQR method)

Handle missing values in the data

In [51]:
```
df.isna().sum()
```

Out[51]:
```
data                                0
trip_creation_time                  0
route_schedule_uuid                 0
route_type                          0
trip_uuid                           0
source_center                       0
source_name                       293
destination_center                  0
destination_name                  261
od_start_time                       0
od_end_time                         0
start_scan_to_end_scan              0
is_cutoff                           0
cutoff_factor                       0
cutoff_timestamp                    0
actual_distance_to_destination      0
actual_time                         0
osrm_time                           0
osrm_distance                       0
factor                              0
segment_actual_time                 0
segment_osrm_time                   0
segment_osrm_distance               0
segment_factor                      0
dtype: int64
```

In [69]:
```
df[df["source_name"].isna()]["trip_uuid"].unique()
```

Out[69]:
```
array(['trip-153786558437756691', 'trip-153842737815495661',
       'trip-153834519721733970', 'trip-153846056503320607',
       'trip-153852612674280168', 'trip-153785822252799564',
       'trip-153835867702133730', 'trip-153843937115921268',
       'trip-153851526862672465', 'trip-153783153973255752',
       'trip-153786712501643905', 'trip-153818244828109704',
       'trip-153777969957700771', 'trip-153802263936969812',
       'trip-153812396555262982', 'trip-153799300352000726',
       'trip-153800882473542201', 'trip-153781894334349262',
       'trip-153833330949418536', 'trip-153836697913613926',
       'trip-153769166516379642', 'trip-153829753238591840',
       'trip-153860002475779846', 'trip-153854526936264994',
       'trip-153824891534925374', 'trip-153785493712368255',
       'trip-153851268207010003', 'trip-153826788460137094',
       'trip-153860203010589724', 'trip-153809438886343536',
       'trip-153790525985329256', 'trip-153792104124206797',
       'trip-153800051661903546', 'trip-153794907452350443',
       'trip-153826587347960527', 'trip-153799759049764136',
       'trip-153791004076950775', 'trip-153739065339036065',
       'trip-153816053387149067', 'trip-153829881746420606',
       'trip-153818227968802218', 'trip-153752033566751074',
       'trip-153811367563100850', 'trip-153760846587062276',
       'trip-153847102374259925', 'trip-153801047876051786',
       'trip-153809656036701745', 'trip-153775037443441088',
       'trip-153842562860298862', 'trip-153779508492576787',
       'trip-153800731291924640', 'trip-153746918356196062',
       'trip-153804224240745898', 'trip-153808726874610853',
       'trip-153795287270071798', 'trip-153776806236494354',
       'trip-153777973100266904', 'trip-153851243698591845',
       'trip-153755911378016076', 'trip-153843109410376388',
       'trip-153794922985641945', 'trip-153821242026370083',
       'trip-153820032399976293', 'trip-153847015570070605',
       'trip-153812758596408063', 'trip-153855756668984584'], dtype=object)
```

In [62]:
```
len(df["trip_uuid"].unique())
```

Out[62]:
```
14817
```

In [63]:
```
df[df["trip_uuid"]=='trip-153786558437756691']
```

Out[63]:

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name	destination_center	des
91	training	2018-09-25 08:53:04.377810	thanos::sroute:4460a38d-ab9b-484e-bd4e-f4201d0...	FTL	trip-153786558437756691	IND302014AAA	Jaipur_Hub (Rajasthan)	IND305001AAC	Ajm
92	training	2018-09-25 08:53:04.377810	thanos::sroute:4460a38d-ab9b-484e-bd4e-f4201d0...	FTL	trip-153786558437756691	IND302014AAA	Jaipur_Hub (Rajasthan)	IND305001AAC	Ajm
93	training	2018-09-25 08:53:04.377810	thanos::sroute:4460a38d-ab9b-484e-bd4e-f4201d0...	FTL	trip-153786558437756691	IND302014AAA	Jaipur_Hub (Rajasthan)	IND305001AAC	Ajm
94	training	2018-09-25 08:53:04.377810	thanos::sroute:4460a38d-ab9b-484e-bd4e-f4201d0...	FTL	trip-153786558437756691	IND302014AAA	Jaipur_Hub (Rajasthan)	IND305001AAC	Ajm
95	training	2018-09-25 08:53:04.377810	thanos::sroute:4460a38d-ab9b-484e-bd4e-f4201d0...	FTL	trip-153786558437756691	IND302014AAA	Jaipur_Hub (Rajasthan)	IND305001AAC	Ajm
96	training	2018-09-25 08:53:04.377810	thanos::sroute:4460a38d-ab9b-484e-bd4e-f4201d0...	FTL	trip-153786558437756691	IND302014AAA	Jaipur_Hub (Rajasthan)	IND305001AAC	Ajm
97	training	2018-09-25 08:53:04.377810	thanos::sroute:4460a38d-ab9b-484e-bd4e-f4201d0...	FTL	trip-153786558437756691	IND305001AAC	Ajmer_FoySGRRD_I (Rajasthan)	IND306401AAB	P
98	training	2018-09-25 08:53:04.377810	thanos::sroute:4460a38d-ab9b-484e-bd4e-f4201d0...	FTL	trip-153786558437756691	IND305001AAC	Ajmer_FoySGRRD_I (Rajasthan)	IND306401AAB	P
99	training	2018-09-25 08:53:04.377810	thanos::sroute:4460a38d-ab9b-484e-bd4e-f4201d0...	FTL	trip-153786558437756691	IND305001AAC	Ajmer_FoySGRRD_I (Rajasthan)	IND306401AAB	P
100	training	2018-09-25 08:53:04.377810	thanos::sroute:4460a38d-ab9b-484e-bd4e-f4201d0...	FTL	trip-153786558437756691	IND305001AAC	Ajmer_FoySGRRD_I (Rajasthan)	IND306401AAB	P
101	training	2018-09-25 08:53:04.377810	thanos::sroute:4460a38d-ab9b-484e-bd4e-f4201d0...	FTL	trip-153786558437756691	IND305001AAC	Ajmer_FoySGRRD_I (Rajasthan)	IND306401AAB	P

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name	destination_center	des
102	training	2018-09-25 08:53:04.377810	thanos::sroute:4460a38d-ab9b-484e-bd4e-f4201d0...	FTL	trip-153786558437756691	IND305001AAC	Ajmer_FoySGRRD_I (Rajasthan)	IND306401AAB	P
103	training	2018-09-25 08:53:04.377810	thanos::sroute:4460a38d-ab9b-484e-bd4e-f4201d0...	FTL	trip-153786558437756691	IND305001AAC	Ajmer_FoySGRRD_I (Rajasthan)	IND306401AAB	P
104	training	2018-09-25 08:53:04.377810	thanos::sroute:4460a38d-ab9b-484e-bd4e-f4201d0...	FTL	trip-153786558437756691	IND306401AAB	Pali_Nayagaon_I (Rajasthan)	IND342005AAD	J
105	training	2018-09-25 08:53:04.377810	thanos::sroute:4460a38d-ab9b-484e-bd4e-f4201d0...	FTL	trip-153786558437756691	IND306401AAB	Pali_Nayagaon_I (Rajasthan)	IND342005AAD	J
106	training	2018-09-25 08:53:04.377810	thanos::sroute:4460a38d-ab9b-484e-bd4e-f4201d0...	FTL	trip-153786558437756691	IND306401AAB	Pali_Nayagaon_I (Rajasthan)	IND342005AAD	J
107	training	2018-09-25 08:53:04.377810	thanos::sroute:4460a38d-ab9b-484e-bd4e-f4201d0...	FTL	trip-153786558437756691	IND342005AAD	Jodhpur_Basni_I (Rajasthan)	IND342601AAA	Piparci
108	training	2018-09-25 08:53:04.377810	thanos::sroute:4460a38d-ab9b-484e-bd4e-f4201d0...	FTL	trip-153786558437756691	IND342005AAD	Jodhpur_Basni_I (Rajasthan)	IND342601AAA	Piparci
109	training	2018-09-25 08:53:04.377810	thanos::sroute:4460a38d-ab9b-484e-bd4e-f4201d0...	FTL	trip-153786558437756691	IND342005AAD	Jodhpur_Basni_I (Rajasthan)	IND342601AAA	Piparci
110	training	2018-09-25 08:53:04.377810	thanos::sroute:4460a38d-ab9b-484e-bd4e-f4201d0...	FTL	trip-153786558437756691	IND342601AAA	Piparcity_BsstdDPP_D (Rajasthan)	IND342902A1B	
111	training	2018-09-25 08:53:04.377810	thanos::sroute:4460a38d-ab9b-484e-bd4e-f4201d0...	FTL	trip-153786558437756691	IND342601AAA	Piparcity_BsstdDPP_D (Rajasthan)	IND342902A1B	
112	training	2018-09-25 08:53:04.377810	thanos::sroute:4460a38d-ab9b-484e-bd4e-f4201d0...	FTL	trip-153786558437756691	IND342902A1B	NaN	IND302014AAA	

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name	destination_center	des
113	training	2018-09-25 08:53:04.377810	thanos::sroute:4460a38d-ab9b-484e-bd4e-f4201d0...	FTL	trip-153786558437756691	IND342902A1B	NaN	IND302014AAA	
114	training	2018-09-25 08:53:04.377810	thanos::sroute:4460a38d-ab9b-484e-bd4e-f4201d0...	FTL	trip-153786558437756691	IND342902A1B	NaN	IND302014AAA	
115	training	2018-09-25 08:53:04.377810	thanos::sroute:4460a38d-ab9b-484e-bd4e-f4201d0...	FTL	trip-153786558437756691	IND342902A1B	NaN	IND302014AAA	
116	training	2018-09-25 08:53:04.377810	thanos::sroute:4460a38d-ab9b-484e-bd4e-f4201d0...	FTL	trip-153786558437756691	IND342902A1B	NaN	IND302014AAA	
117	training	2018-09-25 08:53:04.377810	thanos::sroute:4460a38d-ab9b-484e-bd4e-f4201d0...	FTL	trip-153786558437756691	IND342902A1B	NaN	IND302014AAA	
118	training	2018-09-25 08:53:04.377810	thanos::sroute:4460a38d-ab9b-484e-bd4e-f4201d0...	FTL	trip-153786558437756691	IND342902A1B	NaN	IND302014AAA	
119	training	2018-09-25 08:53:04.377810	thanos::sroute:4460a38d-ab9b-484e-bd4e-f4201d0...	FTL	trip-153786558437756691	IND342902A1B	NaN	IND302014AAA	
120	training	2018-09-25 08:53:04.377810	thanos::sroute:4460a38d-ab9b-484e-bd4e-f4201d0...	FTL	trip-153786558437756691	IND342902A1B	NaN	IND302014AAA	
121	training	2018-09-25 08:53:04.377810	thanos::sroute:4460a38d-ab9b-484e-bd4e-f4201d0...	FTL	trip-153786558437756691	IND342902A1B	NaN	IND302014AAA	

In [67]:
```
df[(df["source_center"]=="IND342902A1B") | (df["destination_center"]=="IND342902A1B")]
```

Out[67]:

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name	destination_center	de
110	training	2018-09-25 08:53:04.377810	thanos::sroute:4460a38d-ab9b-484e-bd4e-f4201d0...	FTL	trip-153786558437756691	IND342601AAA	Piparcity_BsstdDPP_D (Rajasthan)	IND342902A1B	
111	training	2018-09-25 08:53:04.377810	thanos::sroute:4460a38d-ab9b-484e-bd4e-f4201d0...	FTL	trip-153786558437756691	IND342601AAA	Piparcity_BsstdDPP_D (Rajasthan)	IND342902A1B	
112	training	2018-09-25 08:53:04.377810	thanos::sroute:4460a38d-ab9b-484e-bd4e-f4201d0...	FTL	trip-153786558437756691	IND342902A1B	NaN	IND302014AAA	
113	training	2018-09-25 08:53:04.377810	thanos::sroute:4460a38d-ab9b-484e-bd4e-f4201d0...	FTL	trip-153786558437756691	IND342902A1B	NaN	IND302014AAA	
114	training	2018-09-25 08:53:04.377810	thanos::sroute:4460a38d-ab9b-484e-bd4e-f4201d0...	FTL	trip-153786558437756691	IND342902A1B	NaN	IND302014AAA	
...	...	...	...	...	...	...	...	...	
102433	test	2018-10-02 09:03:43.743009	thanos::sroute:88f3c4f2-ba7c-4817-8dff-e181ba3...	FTL	trip-153847102374259925	IND342902A1B	NaN	IND302014AAA	
102434	test	2018-10-02 09:03:43.743009	thanos::sroute:88f3c4f2-ba7c-4817-8dff-e181ba3...	FTL	trip-153847102374259925	IND342902A1B	NaN	IND302014AAA	
102435	test	2018-10-02 09:03:43.743009	thanos::sroute:88f3c4f2-ba7c-4817-8dff-e181ba3...	FTL	trip-153847102374259925	IND342902A1B	NaN	IND302014AAA	
102436	test	2018-10-02 09:03:43.743009	thanos::sroute:88f3c4f2-ba7c-4817-8dff-e181ba3...	FTL	trip-153847102374259925	IND342902A1B	NaN	IND302014AAA	
102437	test	2018-10-02 09:03:43.743009	thanos::sroute:88f3c4f2-ba7c-4817-8dff-e181ba3...	FTL	trip-153847102374259925	IND342902A1B	NaN	IND302014AAA	

106 rows × 24 columns

In [72]:
```
df[df["source_name"].isna()]["source_center"].unique()
```

Out[72]:
```
array(['IND342902A1B', 'IND577116AAA', 'IND282002AAD', 'IND465333A1B',
       'IND841301AAC', 'IND509103AAC', 'IND126116AAA', 'IND331022A1B',
       'IND505326AAB', 'IND852118A1B'], dtype=object)
```

In [70]:
```
df[df["destination_name"].isna()]["destination_center"].unique()
```

Out[70]:
```
array(['IND342902A1B', 'IND577116AAA', 'IND282002AAD', 'IND465333A1B',
       'IND841301AAC', 'IND505326AAB', 'IND852118A1B', 'IND126116AAA',
       'IND509103AAC', 'IND221005A1A', 'IND250002AAC', 'IND331001A1C',
       'IND122015AAC'], dtype=object)
```

Seems like "IND342902A1B" and some other source and destination center are ones whose names are missing

So instead of dropping missing values we can replace the missing names with the corresponding centre code

In [79]:
```
df.groupby(["source_center","source_name"])
```

Out[79]:
```
<pandas.core.groupby.generic.DataFrameGroupBy object at 0x00000277E81B9D00>
```

In [81]:
```
df.loc[df['source_name'].isna(), 'source_name'] = df["source_center"]
```

In [83]:
```
df.loc[df['destination_name'].isna(), 'destination_name'] = df["destination_center"]
```

In [84]:
```
df.head()
```

Out[84]:

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name	destination_center	desti
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_
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_
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_
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_
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_

5 rows × 24 columns

In [85]:
```
df.isna().sum()
```

Out[85]:
```
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
```

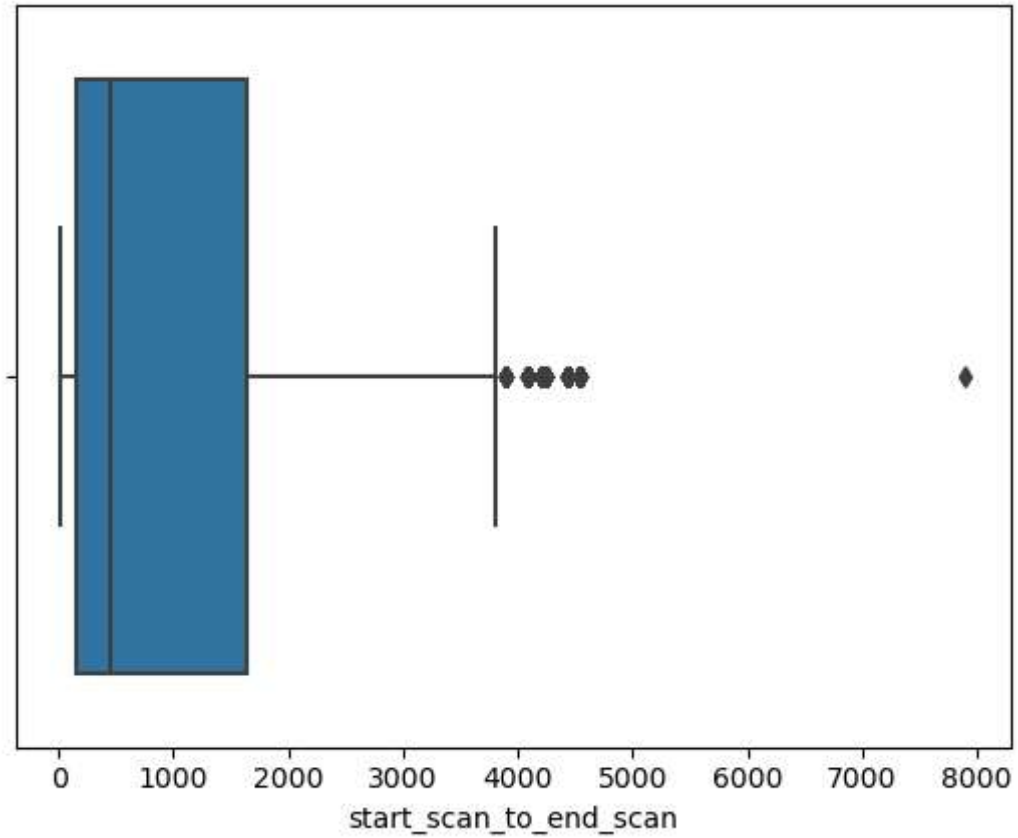
Insight: All missing values has been removed successfully

In [ ]:
```
```

Outlier detection

In [86]:
```
sns.boxplot(x=df["start_scan_to_end_scan"])
```

Out[86]:
```
<AxesSubplot:xlabel='start_scan_to_end_scan'>
```



In [88]:
```
Q1 = np.percentile(df["start_scan_to_end_scan"], 25, interpolation = 'midpoint')
Q2 = np.percentile(df["start_scan_to_end_scan"], 50, interpolation = 'midpoint')
Q3 = np.percentile(df["start_scan_to_end_scan"], 75, interpolation = 'midpoint')

print('Q1 25 percentile of the given data is, ', Q1)
print('Q1 50 percentile of the given data is, ', Q2)
print('Q1 75 percentile of the given data is, ', Q3)

IQR = Q3 - Q1
print('Interquartile range is', IQR)
upperWhisker = Q3 + 1.5*IQR
print('upper limit',upperWhisker)
```

```
Q1 25 percentile of the given data is,  161.0
Q1 50 percentile of the given data is,  449.0
Q1 75 percentile of the given data is,  1634.0
Interquartile range is 1473.0
upper limit 3843.5
```

In [91]:
```
df[df["start_scan_to_end_scan"]>upperWhisker]
```

Out[91]:

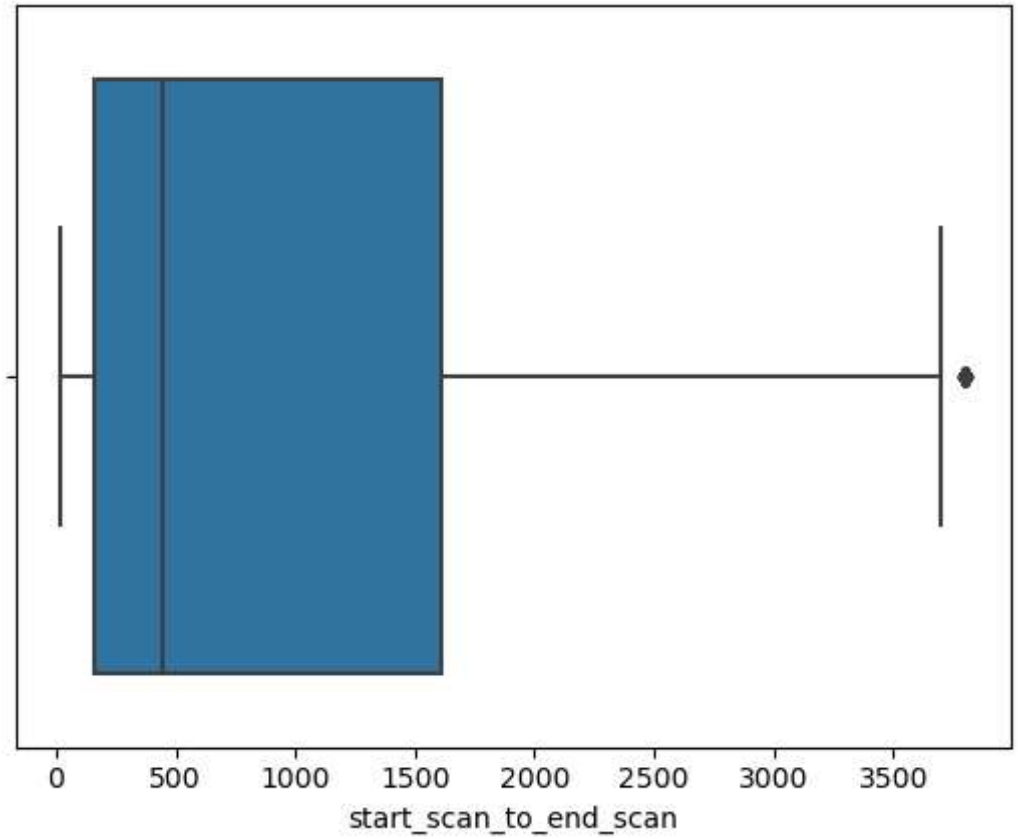
	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name	destination_cen
32950	training	2018-09-13 01:28:45.326644	thanos::sroute:6b87651c-fdf4-432f-bf80-0e394f3...	FTL	trip-153680212532637033	IND712311AAA	Kolkata_Dankuni_HB (West Bengal)	IND781018A
32951	training	2018-09-13 01:28:45.326644	thanos::sroute:6b87651c-fdf4-432f-bf80-0e394f3...	FTL	trip-153680212532637033	IND712311AAA	Kolkata_Dankuni_HB (West Bengal)	IND781018A
32952	training	2018-09-13 01:28:45.326644	thanos::sroute:6b87651c-fdf4-432f-bf80-0e394f3...	FTL	trip-153680212532637033	IND712311AAA	Kolkata_Dankuni_HB (West Bengal)	IND781018A
32953	training	2018-09-13 01:28:45.326644	thanos::sroute:6b87651c-fdf4-432f-bf80-0e394f3...	FTL	trip-153680212532637033	IND712311AAA	Kolkata_Dankuni_HB (West Bengal)	IND781018A
32954	training	2018-09-13 01:28:45.326644	thanos::sroute:6b87651c-fdf4-432f-bf80-0e394f3...	FTL	trip-153680212532637033	IND712311AAA	Kolkata_Dankuni_HB (West Bengal)	IND781018A
...	...	...	...	...	...	...	...	
79524	training	2018-09-19 13:44:58.665210	thanos::sroute:bc7dbb1d-9379-4674-b8d3-f9c3b96...	FTL	trip-153736469866480991	IND000000ACB	Gurgaon_Bilaspur_HB (Haryana)	IND712311A
79525	training	2018-09-19 13:44:58.665210	thanos::sroute:bc7dbb1d-9379-4674-b8d3-f9c3b96...	FTL	trip-153736469866480991	IND000000ACB	Gurgaon_Bilaspur_HB (Haryana)	IND712311A
79526	training	2018-09-19 13:44:58.665210	thanos::sroute:bc7dbb1d-9379-4674-b8d3-f9c3b96...	FTL	trip-153736469866480991	IND000000ACB	Gurgaon_Bilaspur_HB (Haryana)	IND712311A
79527	training	2018-09-19 13:44:58.665210	thanos::sroute:bc7dbb1d-9379-4674-b8d3-f9c3b96...	FTL	trip-153736469866480991	IND000000ACB	Gurgaon_Bilaspur_HB (Haryana)	IND712311A
123196	test	2018-10-01 23:35:54.432745	thanos::sroute:4316e05f-b4cc-4ea7-b801-62a93ae...	Carting	trip-153843695443252828	IND764071AAB	Pappadahandi_Central_DPP_2 (Orissa)	IND530012A

373 rows × 24 columns

In [92]:
```
df.drop(df[df["start_scan_to_end_scan"]>upperWhisker].index , inplace=True)
```

In [95]:
```
sns.boxplot(x=df["start_scan_to_end_scan"])
```

Out[95]:
```
<AxesSubplot:xlabel='start_scan_to_end_scan'>
```



Insight: start_scan_to_end_scan outlier has been removed. If I remove outliers from other columns there will be thousands of data points lost. So we would loose too much data to process further.

In [ ]:
```

```

Feature Creation

Destination Name: Split and extract features out of destination. City-place-code (State)

In [99]:
```
temp=df["destination_name"].str.split("_",  expand = True)
temp
```

Out[99]:

	0	1	2	3
0	Khambhat	MotvdDPP	D (Gujarat)	None
1	Khambhat	MotvdDPP	D (Gujarat)	None
2	Khambhat	MotvdDPP	D (Gujarat)	None
3	Khambhat	MotvdDPP	D (Gujarat)	None
4	Khambhat	MotvdDPP	D (Gujarat)	None
...	...	...	...	...
144862	Gurgaon	Bilaspur	HB (Haryana)	None
144863	Gurgaon	Bilaspur	HB (Haryana)	None
144864	Gurgaon	Bilaspur	HB (Haryana)	None
144865	Gurgaon	Bilaspur	HB (Haryana)	None
144866	Gurgaon	Bilaspur	HB (Haryana)	None

144494 rows × 4 columns

In [100...
```
df['destination_city']=temp[0]
```

In [101...
```
df['destination_place']=temp[1]
```

In [102...
```
df['destination_code_state']=temp[2]
```

In [103...
```
df.head()
```

Out[103]:

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name	destination_center	desti
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_
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_
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_
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_
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_

5 rows × 27 columns

Insight: Destination city , place and state are separated in to separate columns

Source Name: Split and extract features out of destination. City-place-code (State)

In [104…

```
temp=df["source_name"].str.split("_",  expand = True)
temp
```

Out[104]:

	0	1	2	3
0	Anand	VUNagar	DC (Gujarat)	None
1	Anand	VUNagar	DC (Gujarat)	None
2	Anand	VUNagar	DC (Gujarat)	None
3	Anand	VUNagar	DC (Gujarat)	None
4	Anand	VUNagar	DC (Gujarat)	None
...	...	...	...	...
144862	Sonipat	Kundli	H (Haryana)	None
144863	Sonipat	Kundli	H (Haryana)	None
144864	Sonipat	Kundli	H (Haryana)	None
144865	Sonipat	Kundli	H (Haryana)	None
144866	Sonipat	Kundli	H (Haryana)	None

144494 rows × 4 columns

In [105…
```
df['source_city']=temp[0]
df['source_place']=temp[1]
df['source_code_state']=temp[2]
```

In [106…
```
df.head()
```

Out[106]:

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name	destination_center	desti
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_
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_
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_
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_
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_

5 rows × 30 columns

Insight: Source city , place and state are separated in to separate columns

Trip_creation_time: Extract features like month, year and day etc

In [107…

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 144494 entries, 0 to 144866
Data columns (total 30 columns):
 #   Column                          Non-Null Count   Dtype  
---  ------                          --------------   -----  
 0   data                            144494 non-null  object 
 1   trip_creation_time              144494 non-null  object 
 2   route_schedule_uuid             144494 non-null  object 
 3   route_type                      144494 non-null  object 
 4   trip_uuid                       144494 non-null  object 
 5   source_center                   144494 non-null  object 
 6   source_name                     144494 non-null  object 
 7   destination_center              144494 non-null  object 
 8   destination_name                144494 non-null  object 
 9   od_start_time                   144494 non-null  object 
 10  od_end_time                     144494 non-null  object 
 11  start_scan_to_end_scan          144494 non-null  float64
 12  is_cutoff                       144494 non-null  bool   
 13  cutoff_factor                   144494 non-null  int64  
 14  cutoff_timestamp                144494 non-null  object 
 15  actual_distance_to_destination  144494 non-null  float64
 16  actual_time                     144494 non-null  float64
 17  osrm_time                       144494 non-null  float64
 18  osrm_distance                   144494 non-null  float64
 19  factor                          144494 non-null  float64
 20  segment_actual_time             144494 non-null  float64
 21  segment_osrm_time               144494 non-null  float64
 22  segment_osrm_distance           144494 non-null  float64
 23  segment_factor                  144494 non-null  float64
 24  destination_city                144494 non-null  object 
 25  destination_place               141792 non-null  object 
 26  destination_code_state          128688 non-null  object 
 27  source_city                     144494 non-null  object 
 28  source_place                    142094 non-null  object 
 29  source_code_state               129618 non-null  object 
dtypes: bool(1), float64(10), int64(1), object(18)
memory usage: 33.2+ MB
```

In [110…
```
df['trip_creation_time'] = pd.to_datetime(df['trip_creation_time'])
```

In [111…
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 144494 entries, 0 to 144866
Data columns (total 30 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   data                            144494 non-null  object
 1   trip_creation_time              144494 non-null  datetime64[ns]
 2   route_schedule_uuid             144494 non-null  object
 3   route_type                      144494 non-null  object
 4   trip_uuid                       144494 non-null  object
 5   source_center                   144494 non-null  object
 6   source_name                     144494 non-null  object
 7   destination_center              144494 non-null  object
 8   destination_name                144494 non-null  object
 9   od_start_time                   144494 non-null  object
 10  od_end_time                     144494 non-null  object
 11  start_scan_to_end_scan          144494 non-null  float64
 12  is_cutoff                       144494 non-null  bool
 13  cutoff_factor                   144494 non-null  int64
 14  cutoff_timestamp                144494 non-null  object
 15  actual_distance_to_destination  144494 non-null  float64
 16  actual_time                     144494 non-null  float64
 17  osrm_time                       144494 non-null  float64
 18  osrm_distance                   144494 non-null  float64
 19  factor                          144494 non-null  float64
 20  segment_actual_time             144494 non-null  float64
 21  segment_osrm_time               144494 non-null  float64
 22  segment_osrm_distance           144494 non-null  float64
 23  segment_factor                  144494 non-null  float64
 24  destination_city                144494 non-null  object
 25  destination_place               141792 non-null  object
 26  destination_code_state          128688 non-null  object
 27  source_city                     144494 non-null  object
 28  source_place                    142094 non-null  object
 29  source_code_state               129618 non-null  object
dtypes: bool(1), datetime64[ns](1), float64(10), int64(1), object(17)
memory usage: 33.2+ MB
```

In [112…

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

In [114…

```
df['trip_day']=df['trip_creation_time'].dt.day
```

In [115…
```
df.head()
```

Out[115]:

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name	destination_center	desti
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_
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_
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_
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_
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_

5 rows × 33 columns

Insight: extracted month, year and day separately from trip creation time

Calculate the time taken between od_start_time and od_end_time and keep it as a feature. Drop the original columns, if required

In [116…
```
df['od_start_time'] = pd.to_datetime(df['od_start_time'])
df['od_end_time'] = pd.to_datetime(df['od_end_time'])
```

In [117…
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 144494 entries, 0 to 144866
Data columns (total 33 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   data                            144494 non-null  object
 1   trip_creation_time              144494 non-null  datetime64[ns]
 2   route_schedule_uuid             144494 non-null  object
 3   route_type                      144494 non-null  object
 4   trip_uuid                       144494 non-null  object
 5   source_center                   144494 non-null  object
 6   source_name                     144494 non-null  object
 7   destination_center              144494 non-null  object
 8   destination_name                144494 non-null  object
 9   od_start_time                   144494 non-null  datetime64[ns]
 10  od_end_time                     144494 non-null  datetime64[ns]
 11  start_scan_to_end_scan          144494 non-null  float64
 12  is_cutoff                       144494 non-null  bool
 13  cutoff_factor                   144494 non-null  int64
 14  cutoff_timestamp                144494 non-null  object
 15  actual_distance_to_destination  144494 non-null  float64
 16  actual_time                     144494 non-null  float64
 17  osrm_time                       144494 non-null  float64
 18  osrm_distance                   144494 non-null  float64
 19  factor                          144494 non-null  float64
 20  segment_actual_time             144494 non-null  float64
 21  segment_osrm_time               144494 non-null  float64
 22  segment_osrm_distance           144494 non-null  float64
 23  segment_factor                  144494 non-null  float64
 24  destination_city                144494 non-null  object
 25  destination_place               141792 non-null  object
 26  destination_code_state          128688 non-null  object
 27  source_city                     144494 non-null  object
 28  source_place                    142094 non-null  object
 29  source_code_state               129618 non-null  object
 30  trip_month                      144494 non-null  int64
 31  trip_year                       144494 non-null  int64
 32  trip_day                        144494 non-null  int64
dtypes: bool(1), datetime64[ns](3), float64(10), int64(4), object(15)
memory usage: 36.5+ MB
```

In [119…

```
df['total_time'] = (df['od_end_time']-df['od_start_time'])
```

In [132…
```
df['total_time'] = np.round(df['total_time'].dt.seconds / 60 , 2)
```

In [133…
```
df.head()
```

Out[133]:

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name	destination_center	desti
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_
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_
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_
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_
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_

5 rows × 34 columns

In [134…
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 144494 entries, 0 to 144866
Data columns (total 34 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   data                            144494 non-null  object
 1   trip_creation_time              144494 non-null  datetime64[ns]
 2   route_schedule_uuid             144494 non-null  object
 3   route_type                      144494 non-null  object
 4   trip_uuid                       144494 non-null  object
 5   source_center                   144494 non-null  object
 6   source_name                     144494 non-null  object
 7   destination_center              144494 non-null  object
 8   destination_name                144494 non-null  object
 9   od_start_time                   144494 non-null  datetime64[ns]
 10  od_end_time                     144494 non-null  datetime64[ns]
 11  start_scan_to_end_scan          144494 non-null  float64
 12  is_cutoff                       144494 non-null  bool
 13  cutoff_factor                   144494 non-null  int64
 14  cutoff_timestamp                144494 non-null  object
 15  actual_distance_to_destination  144494 non-null  float64
 16  actual_time                     144494 non-null  float64
 17  osrm_time                       144494 non-null  float64
 18  osrm_distance                   144494 non-null  float64
 19  factor                          144494 non-null  float64
 20  segment_actual_time             144494 non-null  float64
 21  segment_osrm_time               144494 non-null  float64
 22  segment_osrm_distance           144494 non-null  float64
 23  segment_factor                  144494 non-null  float64
 24  destination_city                144494 non-null  object
 25  destination_place               141792 non-null  object
 26  destination_code_state          128688 non-null  object
 27  source_city                     144494 non-null  object
 28  source_place                    142094 non-null  object
 29  source_code_state               129618 non-null  object
 30  trip_month                      144494 non-null  int64
 31  trip_year                       144494 non-null  int64
 32  trip_day                        144494 non-null  int64
 33  total_time                      144494 non-null  float64
dtypes: bool(1), datetime64[ns](3), float64(11), int64(4), object(15)
memory usage: 37.6+ MB
```

Insight: catculated total time taken between od start and end time in minutes

Handling categorical values (Do one-hot encoding of categorical variables)

In [136…

```
df_hot_encoded = pd.get_dummies(data = df, columns = ["route_type"], prefix = "is")
df_hot_encoded
```

Out[136]:

	data	trip_creation_time	route_schedule_uuid	trip_uuid	source_center	source_name	destination_center	destination_
0	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	IND388620AAB	Khambhat_Motvd (G
1	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	IND388620AAB	Khambhat_Motvd (G
2	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	IND388620AAB	Khambhat_Motvd (G
3	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	IND388620AAB	Khambhat_Motvd (G
4	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	trip-153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	IND388620AAB	Khambhat_Motvd (G
...	...	...	...	...	...	...	...	
144862	training	2018-09-20 16:24:28.436231	thanos::sroute:f0569d2f-4e20-4c31-8542-67b86d5...	trip-153746066843555182	IND131028AAB	Sonipat_Kundli_H (Haryana)	IND000000ACB	Gurgaon_Bilasp (Ha
144863	training	2018-09-20 16:24:28.436231	thanos::sroute:f0569d2f-4e20-4c31-8542-67b86d5...	trip-153746066843555182	IND131028AAB	Sonipat_Kundli_H (Haryana)	IND000000ACB	Gurgaon_Bilasp (Ha
144864	training	2018-09-20 16:24:28.436231	thanos::sroute:f0569d2f-4e20-4c31-8542-67b86d5...	trip-153746066843555182	IND131028AAB	Sonipat_Kundli_H (Haryana)	IND000000ACB	Gurgaon_Bilasp (Ha
144865	training	2018-09-20 16:24:28.436231	thanos::sroute:f0569d2f-4e20-4c31-8542-67b86d5...	trip-153746066843555182	IND131028AAB	Sonipat_Kundli_H (Haryana)	IND000000ACB	Gurgaon_Bilasp (Ha
144866	training	2018-09-20 16:24:28.436231	thanos::sroute:f0569d2f-4e20-4c31-8542-67b86d5...	trip-153746066843555182	IND131028AAB	Sonipat_Kundli_H (Haryana)	IND000000ACB	Gurgaon_Bilasp (Ha

144494 rows × 35 columns

Insight: hot-encoded if an order is of carting or ftl type in 2 new columns

In [ ]:

Column Normalization /Column Standardization (using MinMaxScaler or StandardScaler)

In [142…

```
scaler = StandardScaler()
df['start_scan_to_end_scan_standard'] = scaler.fit_transform(df[['start_scan_to_end_scan']])
df.head()
```

Out[142]:

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name	destination_center	desti
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_
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_
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_
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_
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_

5 rows × 36 columns

In [143…
```
scaler = MinMaxScaler()
df['actual_distance_to_destination_standard'] = scaler.fit_transform(df[['actual_distance_to_destination']])
df.head()
```

Out[143]:

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name	destination_center	desti
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_
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_
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_
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_
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_

5 rows × 36 columns

Insight: scaler standardized start_scan_to_end_scan and maxmin standardized actual_distance_to_destination

Merging of rows and aggregation of fields

In [152…
```
#cumulative
actual_time_aggregated = df.groupby(['trip_uuid','source_center','destination_center']).aggregate({'actual_time':'sum'}).groupby
actual_time_aggregated
```

Out[152]:

	trip_uuid	source_center	destination_center	actual_time
0	trip-153671041653548748	IND209304AAA	IND000000ACB	6484.0
1	trip-153671041653548748	IND462022AAA	IND209304AAA	15682.0
2	trip-153671042288605164	IND561203AAB	IND562101AAA	96.0
3	trip-153671042288605164	IND572101AAA	IND561203AAB	399.0
4	trip-153671043369099517	IND000000ACB	IND160002AAC	2601.0
...	...	...	...	...
26356	trip-153861115439069069	IND628204AAA	IND627657AAA	376.0
26357	trip-153861115439069069	IND628613AAA	IND627005AAA	549.0
26358	trip-153861115439069069	IND628801AAA	IND628204AAA	600.0
26359	trip-153861118270144424	IND583119AAA	IND583101AAA	278.0
26360	trip-153861118270144424	IND583201AAA	IND583119AAA	350.0

26361 rows × 4 columns

In [153…

```
#cumulative
OSRM_time_aggregated = df.groupby(['trip_uuid','source_center','destination_center']).aggregate({'osrm_time':'sum'}).groupby(leve
OSRM_time_aggregated
```

Out[153]:

	trip_uuid	source_center	destination_center	osrm_time
0	trip-153671041653548748	IND209304AAA	IND000000ACB	3464.0
1	trip-153671041653548748	IND462022AAA	IND209304AAA	7787.0
2	trip-153671042288605164	IND561203AAB	IND562101AAA	55.0
3	trip-153671042288605164	IND572101AAA	IND561203AAB	210.0
4	trip-153671043369099517	IND000000ACB	IND160002AAC	1427.0
...	...	...	...	...
26356	trip-153861115439069069	IND628204AAA	IND627657AAA	316.0
26357	trip-153861115439069069	IND628613AAA	IND627005AAA	424.0
26358	trip-153861115439069069	IND628801AAA	IND628204AAA	446.0
26359	trip-153861118270144424	IND583119AAA	IND583101AAA	59.0
26360	trip-153861118270144424	IND583201AAA	IND583119AAA	106.0

26361 rows × 4 columns

In [154...

```
#cumulative
osrm_distance_aggregated = df.groupby(['trip_uuid','source_center','destination_center']).aggregate({'osrm_distance':'sum'}).grou
osrm_distance_aggregated
```

Out[154]:

	trip_uuid	source_center	destination_center	osrm_distance
0	trip-153671041653548748	IND209304AAA	IND000000ACB	4540.1261
1	trip-153671041653548748	IND462022AAA	IND209304AAA	10577.7647
2	trip-153671042288605164	IND561203AAB	IND562101AAA	60.3157
3	trip-153671042288605164	IND572101AAA	IND561203AAB	269.4308
4	trip-153671043369099517	IND000000ACB	IND160002AAC	1975.7409
...	...	...	...	...
26356	trip-153861115439069069	IND628204AAA	IND627657AAA	312.1457
26357	trip-153861115439069069	IND628613AAA	IND627005AAA	424.0012
26358	trip-153861115439069069	IND628801AAA	IND628204AAA	449.5383
26359	trip-153861118270144424	IND583119AAA	IND583101AAA	76.5169
26360	trip-153861118270144424	IND583201AAA	IND583119AAA	127.8020

26361 rows × 4 columns

In [157…

```
segment_actual_time_aggregated = df.groupby(['trip_uuid','source_center','destination_center']).aggregate({'segment_actual_time'
segment_actual_time_aggregated
```

Out[157]:

	trip_uuid	source_center	destination_center	segment_actual_time
0	trip-153671041653548748	IND209304AAA	IND000000ACB	728.0
1	trip-153671041653548748	IND462022AAA	IND209304AAA	820.0
2	trip-153671042288605164	IND561203AAB	IND562101AAA	46.0
3	trip-153671042288605164	IND572101AAA	IND561203AAB	95.0
4	trip-153671043369099517	IND000000ACB	IND160002AAC	608.0
...	...	...	...	...
26356	trip-153861115439069069	IND628204AAA	IND627657AAA	49.0
26357	trip-153861115439069069	IND628613AAA	IND627005AAA	89.0
26358	trip-153861115439069069	IND628801AAA	IND628204AAA	29.0
26359	trip-153861118270144424	IND583119AAA	IND583101AAA	233.0
26360	trip-153861118270144424	IND583201AAA	IND583119AAA	41.0

26361 rows × 4 columns

In [158…

```
segment_osrm_distance_aggregated = df.groupby(['trip_uuid','source_center','destination_center']).aggregate({'segment_osrm_distar
segment_osrm_distance_aggregated
```

Out[158]:

	trip_uuid	source_center	destination_center	segment_osrm_distance
0	trip-153671041653548748	IND209304AAA	IND000000ACB	670.6205
1	trip-153671041653548748	IND462022AAA	IND209304AAA	649.8528
2	trip-153671042288605164	IND561203AAB	IND562101AAA	28.1995
3	trip-153671042288605164	IND572101AAA	IND561203AAB	55.9899
4	trip-153671043369099517	IND000000ACB	IND160002AAC	317.7408
...	...	...	...	...
26356	trip-153861115439069069	IND628204AAA	IND627657AAA	42.1431
26357	trip-153861115439069069	IND628613AAA	IND627005AAA	78.5869
26358	trip-153861115439069069	IND628801AAA	IND628204AAA	16.0184
26359	trip-153861118270144424	IND583119AAA	IND583101AAA	52.5303
26360	trip-153861118270144424	IND583201AAA	IND583119AAA	28.0484

26361 rows × 4 columns

In [159…

```
segment_osrm_time_aggregated = df.groupby(['trip_uuid','source_center','destination_center']).aggregate({'segment_osrm_time':'su
segment_osrm_time_aggregated
```

Out[159]:

	trip_uuid	source_center	destination_center	segment_osrm_time
0	trip-153671041653548748	IND209304AAA	IND000000ACB	534.0
1	trip-153671041653548748	IND462022AAA	IND209304AAA	474.0
2	trip-153671042288605164	IND561203AAB	IND562101AAA	26.0
3	trip-153671042288605164	IND572101AAA	IND561203AAB	39.0
4	trip-153671043369099517	IND000000ACB	IND160002AAC	231.0
...	...	...	...	...
26356	trip-153861115439069069	IND628204AAA	IND627657AAA	42.0
26357	trip-153861115439069069	IND628613AAA	IND627005AAA	77.0
26358	trip-153861115439069069	IND628801AAA	IND628204AAA	14.0
26359	trip-153861118270144424	IND583119AAA	IND583101AAA	42.0
26360	trip-153861118270144424	IND583201AAA	IND583119AAA	25.0

26361 rows × 4 columns

Insight: aggregated different columns using groupby on trip_uuid, source centre and destination center

Comparison & Visualization of time and distance fields

In [162…

```
# dropping unknown columns
df.drop(['is_cutoff','cutoff_factor','cutoff_timestamp','factor','segment_factor'], axis=1, inplace=True)
df.head()
```

Out[162]:

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name	destination_center	desti
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_
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_
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_
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_
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_

5 rows × 31 columns

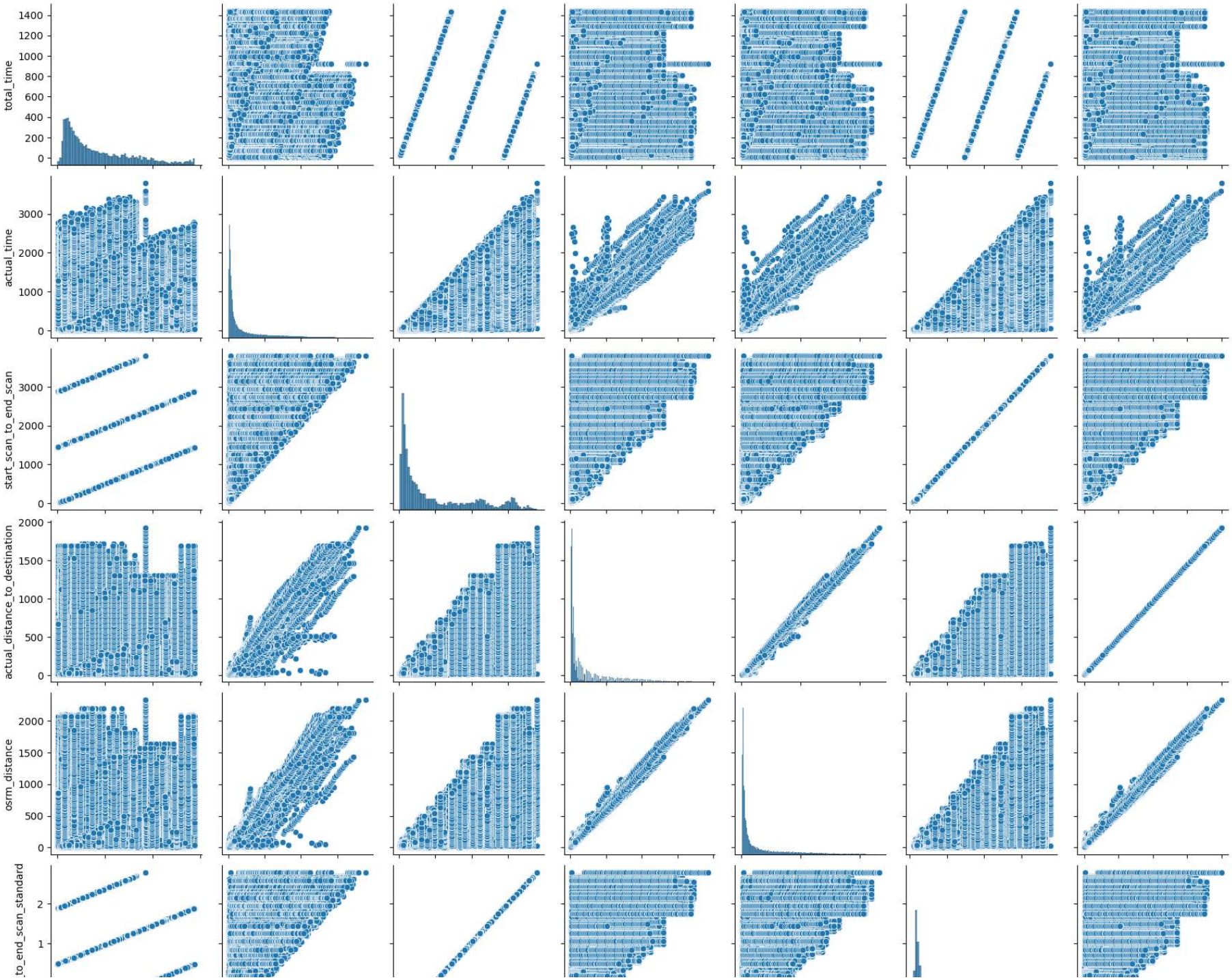
In [166…

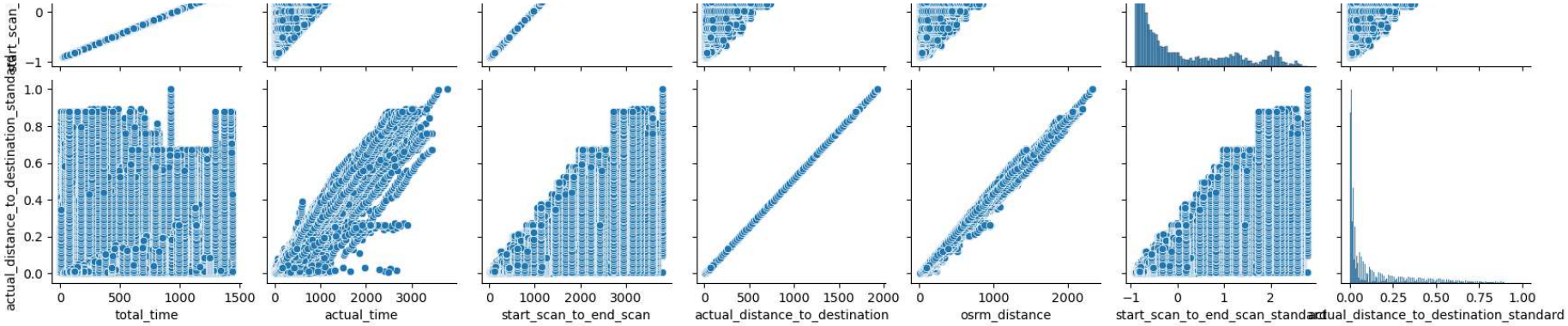
```
plt.figure(figsize=(8,8))
sns.pairplot(data=df[['total_time','actual_time','start_scan_to_end_scan','actual_distance_to_destination','osrm_distance','start
```

Out[166]:

```
<seaborn.axisgrid.PairGrid at 0x277a473dc10>

<Figure size 800x800 with 0 Axes>
```


total_time
actual_time
start_scan_to_end_scan
actual_distance_to_destination
osrm_distance
to_end_scan_standard



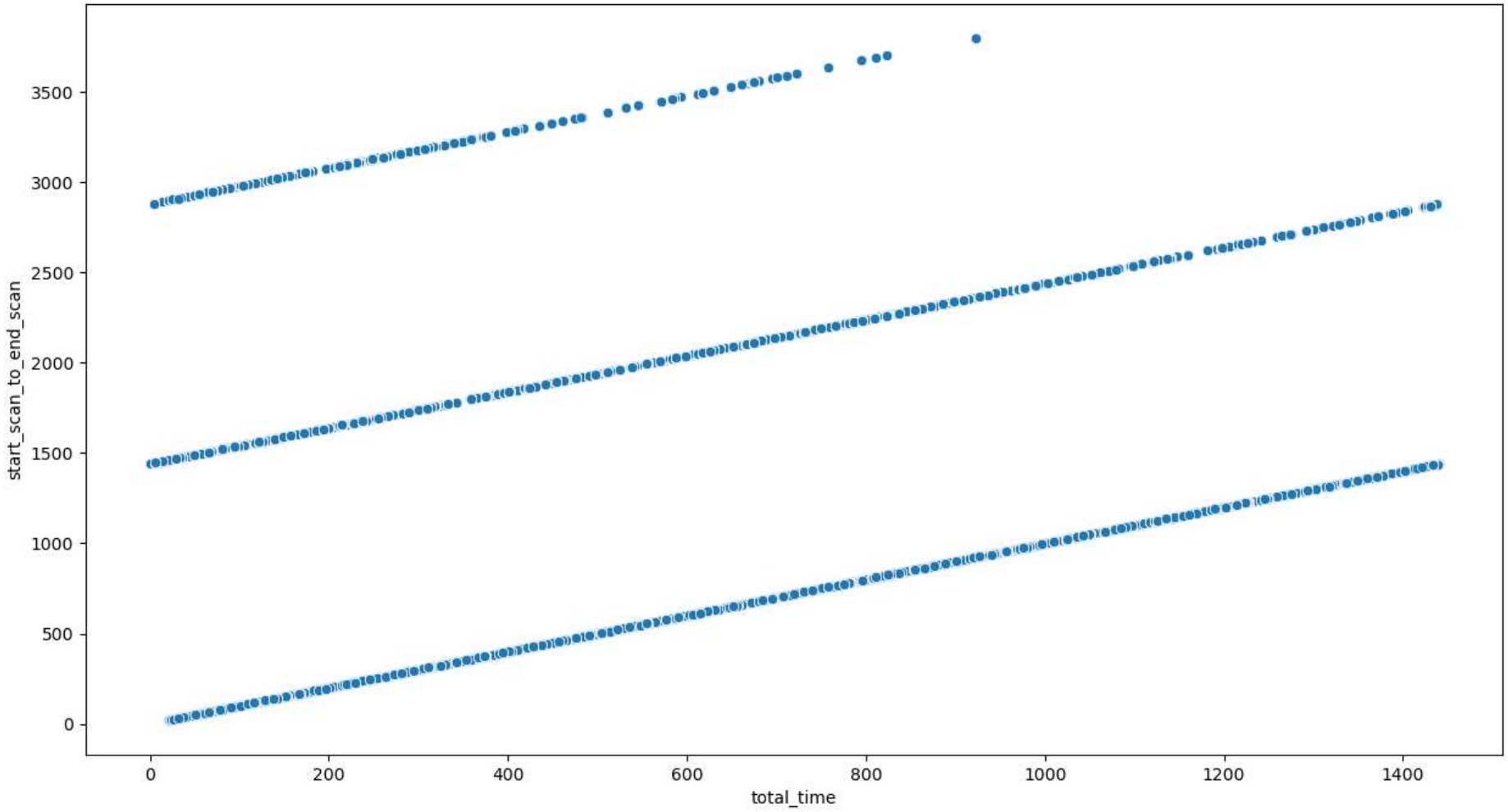
Insight: Comparison of different time and distance columns can be seen in the pairplot visualization above

Checking relationship between aggregated fields

Compare the difference between total_time and start_scan_to_end_scan Do hypothesis testing/ Visual analysis to check.

In [ ]:
```
#H0:The means of the distributions underlying the two samples are the same.
#H1:The means of the distributions underlying the two samples are NOT the same.
```

In [192…
```
plt.figure(figsize=(15,8))
sns.scatterplot(x='total_time',y='start_scan_to_end_scan',data=df)
plt.show()
```



In [193…
```
stat,pvalue = stats.ttest_ind(df['total_time'], df['start_scan_to_end_scan'])
pvalue
```

Out[193]:
```
0.0
```

In [194…
```
alpha = 0.05
if pvalue < alpha:
    print("Reject H0")
```

```
else :
    print("cannot reject H0")
```

```
Reject H0
```

Insight: The means of the distributions underlying the two samples are NOT the same

Do hypothesis testing/ visual analysis between actual_time aggregated value and OSRM time aggregated value

In [ ]:
```
#H0:The means of the distributions underlying the two samples are the same.
#H1:The means of the distributions underlying the two samples are NOT the same.
```

In [197…
```
stat,pvalue = stats.ttest_ind(actual_time_aggregated['actual_time'], OSRM_time_aggregated['osrm_time'])
pvalue
```

Out[197]:
```
1.1384163488225839e-64
```

In [198…
```
alpha = 0.05
if pvalue < alpha:
    print("Reject H0")


else :
    print("cannot reject H0")
```

```
Reject H0
```

Insight: The means of the distributions underlying the two samples are NOT the same

Do hypothesis testing/ visual analysis between actual_time aggregated value and segment actual time aggregated value

In [ ]:
```
#H0:The means of the distributions underlying the two samples are the same.
#H1:The means of the distributions underlying the two samples are NOT the same.
```

In [201…
```
stat,pvalue = stats.ttest_ind(actual_time_aggregated['actual_time'], segment_actual_time_aggregated['segment_actual_time'])
pvalue
```

Out[201]:
```
1.6818148011237537e-298
```

In [202…
```
alpha = 0.05
if pvalue < alpha:
    print("Reject H0")


else :
    print("cannot reject H0")
```

```
Reject H0
```

Insight: The means of the distributions underlying the two samples are NOT the same

Do hypothesis testing/ visual analysis between osrm distance aggregated value and segment osrm distance aggregated value

In [ ]:
```
#H0:The means of the distributions underlying the two samples are the same.
#H1:The means of the distributions underlying the two samples are NOT the same.
```

In [205…
```
stat,pvalue = stats.ttest_ind(osrm_distance_aggregated['osrm_distance'], segment_osrm_distance_aggregated['segment_osrm_distance
pvalue
```

Out[205]:
```
1.4571870772969992e-277
```

In [206…
```
alpha = 0.05
if pvalue < alpha:
    print("Reject H0")


else :
    print("cannot reject H0")
```

```
Reject H0
```

Insight: The means of the distributions underlying the two samples are NOT the same

Do hypothesis testing/ visual analysis between osrm time aggregated value and segment osrm time aggregated value

In [ ]:
```
#H0:The means of the distributions underlying the two samples are the same.
#H1:The means of the distributions underlying the two samples are NOT the same.
```

In [210…
```
stat,pvalue = stats.ttest_ind(OSRM_time_aggregated['osrm_time'], segment_osrm_time_aggregated['segment_osrm_time'])
pvalue
```

Out[210]:
```
9.00889079257682e-286
```

In [211…
```
alpha = 0.05
if pvalue < alpha:
    print("Reject H0")


else :
    print("cannot reject H0")
```

```
Reject H0
```

Insight: The means of the distributions underlying the two samples are NOT the same

Business Insights

Check from where most orders are coming from (State, Corridor etc)

In [173…
```
temp=df["destination_code_state"].str.split(" ",  expand = True)
temp
```

Out[173]:

	0	1	2	3	4
0	D	(Gujarat)	None	None	None
1	D	(Gujarat)	None	None	None
2	D	(Gujarat)	None	None	None
3	D	(Gujarat)	None	None	None
4	D	(Gujarat)	None	None	None
...	...	...	...	...	...
144862	HB	(Haryana)	None	None	None
144863	HB	(Haryana)	None	None	None
144864	HB	(Haryana)	None	None	None
144865	HB	(Haryana)	None	None	None
144866	HB	(Haryana)	None	None	None

144494 rows × 5 columns

In [174...
```
df['destination_state_name']=temp[1]
```

In [175...
```
df.groupby(['destination_state_name']).aggregate({'route_schedule_uuid':'count'}).reset_index().sort_values('route_schedule_uuid
```

Out[175]:

	destination_state_name	route_schedule_uuid
18	(Meghalaya)	8
26	(Tripura)	9
19	(Mizoram)	31
6	(Dadra	34
8	(Goa)	74
21	(Pondicherry)	154
12	(Jammu	167
1	(Arunachal	185
5	(Chhattisgarh)	221
4	(Chandigarh)	282
20	(Orissa)	453
11	(Himachal	543
28	(Uttarakhand)	666
13	(Jharkhand)	1076
2	(Assam)	1250
15	(Kerala)	1835
9	(Gujarat)	2456
3	(Bihar)	2809
23	(Rajasthan)	3335
22	(Punjab)	3751
16	(Madhya	3853
27	(Uttar	4413
7	(Delhi)	5424
0	(Andhra	5983

	destination_state_name	route_schedule_uuid
29	(West	7051
24	(Tamil	7540
25	(Telangana)	8033
17	(Maharashtra)	15608
14	(Karnataka)	19676
10	(Haryana)	19747

In [176…
```
temp=df["source_code_state"].str.split(" ",  expand = True)
temp
```

Out[176]:

	0	1	2	3	4
0	DC	(Gujarat)	None	None	None
1	DC	(Gujarat)	None	None	None
2	DC	(Gujarat)	None	None	None
3	DC	(Gujarat)	None	None	None
4	DC	(Gujarat)	None	None	None
…	…	…	…	…	…
144862	H	(Haryana)	None	None	None
144863	H	(Haryana)	None	None	None
144864	H	(Haryana)	None	None	None
144865	H	(Haryana)	None	None	None
144866	H	(Haryana)	None	None	None

144494 rows × 5 columns

In [177…
```
df['source_state_name']=temp[1]
```

In [178…

```
df.groupby(['source_state_name']).aggregate({'route_schedule_uuid':'count'}).reset_index().sort_values('route_schedule_uuid')
```

Out[178]:

	source_state_name	route_schedule_uuid
26	(Tripura)	5
19	(Mizoram)	26
6	(Dadra	30
21	(Pondicherry)	49
18	(Meghalaya)	77
1	(Arunachal	151
8	(Goa)	165
12	(Jammu	182
5	(Chhattisgarh)	229
4	(Chandigarh)	367
20	(Orissa)	404
11	(Himachal	532
28	(Uttarakhand)	827
2	(Assam)	1126
13	(Jharkhand)	1343
9	(Gujarat)	1867
15	(Kerala)	2038
3	(Bihar)	2206
22	(Punjab)	3381
23	(Rajasthan)	3530
16	(Madhya	3687
7	(Delhi)	4318
29	(West	4645
27	(Uttar	4791

	source_state_name	route_schedule_uuid
0	(Andhra	5087
25	(Telangana)	6212
24	(Tamil	6697
14	(Karnataka)	18534
17	(Maharashtra)	18997
10	(Haryana)	26735

In [179…
```
df.groupby(['source_city']).aggregate({'route_schedule_uuid':'count'}).reset_index().sort_values('route_schedule_uuid')
```

Out[179]:

	source_city	route_schedule_uuid
1200	Tiruchi	1
655	Kothanalloor	1
663	Krishnanagar	1
602	Kayamkulam	1
1155	Sumerpur	1
...	...	...
140	Bengaluru	4237
970	Pune	4269
172	Bhiwandi	9088
102	Bangalore	10104
425	Gurgaon	23458

1272 rows × 2 columns

In [180…
```
df.groupby(['source_place']).aggregate({'route_schedule_uuid':'count'}).reset_index().sort_values('route_schedule_uuid')
```

Out[180]:

	source_place	route_schedule_uuid
139	BnkrGate	1
849	RajpurRD	1
908	Samyaprm	1
32	AnadiDPP	1
914	Sardhnrd	1
...	...	...
1031	Tathawde	4061
172	Central	8988
596	Mankoli	9088
719	Nelmngla	10053
127	Bilaspur	23257

1178 rows × 2 columns

In [181…

```
df.groupby(['destination_city']).aggregate({'route_schedule_uuid':'count'}).reset_index().sort_values('route_schedule_uuid')
```

Out[181]:

	destination_city	route_schedule_uuid
623	Khatauli	1
281	Daman	1
433	Hanskhali	1
353	Falna	1
1052	Salem	1
...	...	...
296	Delhi	5362
170	Bhiwandi	5511
462	Hyderabad	5838
105	Bangalore	11010
418	Gurgaon	15393

1271 rows × 2 columns

In [182…

```
df.groupby(['destination_place']).aggregate({'route_schedule_uuid':'count'}).reset_index().sort_values('route_schedule_uuid')
```

Out[182]:

	destination_place	route_schedule_uuid
426	Kadtmpty	1
921	ShivaDPP	1
436	Kalyan (Maharashtra)	1
37	ArickDPP	1
900	SbhRDDPP	1
...	...	...
912	Shamshbd	5309
581	Mankoli	5511
164	Central	9373
698	Nelmngla	10942
123	Bilaspur	15363

1154 rows × 2 columns

Insight: places like Gurgaon and Bilaspur in Haryana seems to be the place from where most orders are placed as seen above

Busiest corridor, avg distance between them, avg time taken

In [184…

```
df.groupby(['source_city','destination_city']).aggregate({'route_schedule_uuid':'count','actual_distance_to_destination':'mean',
```

Out[184]:

	source_city	destination_city	route_schedule_uuid	actual_distance_to_destination	total_time
2377	Vizag	Vishakhapatnam (Andhra Pradesh)	1	9.228686	231.200000
1335	Kottayam	Kothanalloor	1	15.211748	399.620000
1432	Mahasamund	Durg	1	77.035515	902.830000
596	Delhi	North Delhi (Delhi)	1	9.045083	726.820000
1332	Kothanalloor	Vaikom	1	15.603861	57.380000
...	...	...	...	...	...
193	Bangalore	Bengaluru	1741	21.535381	220.859081
274	Bengaluru	Bengaluru	2062	22.267936	162.984753
844	Gurgaon	Kolkata	2802	672.777483	854.210525
201	Bangalore	Gurgaon	3316	869.072245	375.352325
819	Gurgaon	Bangalore	4899	859.844671	310.120988

2397 rows × 5 columns

Insight: Bangalore - Gurgaon - Bangalore seems to be the bussiest corridor of all

Actionable items for business

- business should focus on routes other than just well developed routes like Gurgaon and bangalore
- They should use a different approach to calculate osrm time since the accuracy is not that high
- data capture for full source and destination name should be improved so that there are no missing values in data

In [ ]: