

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
```

```
!gdown 1ICPz7tBoy_TmnaGWaXz8CdpobqhFZhck
```

Downloading...
From: https://drive.google.com/uc?id=1ICPz7tBoy_TmnaGWaXz8CdpobqhFZhck
To: /content/delhivery_data.csv
100% 55.6M/55.6M [00:01<00:00, 46.7MB/s]

```
df = pd.read_csv('delhivery_data.csv')
pd.set_option('display.max_columns',None)
df.head()
```

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name	dest_center
0	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	
1	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	
2	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	
3	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	
4	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	

```
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
```

▼ 1. Basic Data Cleaning and Exploration

▼ 1. Handling Missing Values

```
df.isna().sum()
```

	0
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
```

```
df = df.dropna(how = 'any')
```

```
df['start_scan_to_end_scan'].sum()
```

```
np.float64(139076997.0)
```

```
df['source_name'].nunique()
```

```
1496
```

```
df['source_center'].nunique()
```

```
1496
```

```
# the source_name and source_center seem to be having 1 to 1 relationship
```

```
source = df.groupby(['source_center'])['source_name'].nunique()
source[source != 1]
```

```
source_name
```

```
source_center
```

```
dtype: int64
```

```
df[df['source_name'].isnull()]['source_center'].unique()
```

```
array([], dtype=object)
```

```
df[df['source_center'] == 'IND342902A1B']['source_name']
```

```
source_name
```

```
dtype: object
```

```
df[df['destination_center'] == 'IND342902A1B'][['destination_center', 'destination_name']]
```

```
destination_center destination_name
```

```
df[df['destination_center'].isin(df[df['source_name'].isnull()]['source_center'].unique())]['destination_name']
```

```
destination_name
```

```
dtype: object
```

```
source[source.index.isin(['IND126116AAA', 'IND282002AAD'])]
```

```
source_name
```

```
source_center
```

```
dtype: int64
```

```
df['source_name'].value_counts().head()
```

```
count
```

```
source_name
```

Gurgaon_Bilaspur_HB (Haryana)	23267
Bangalore_Nelmngla_H (Karnataka)	9975
Bhiwandi_Mankoli_HB (Maharashtra)	9088
Pune_Tathawde_H (Maharashtra)	4061
Hyderabad_Shamsabd_H (Telangana)	3340

```
dtype: int64
```

▼ 2. Converting time columns to pandas datetime

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 144316 entries, 0 to 144866
Data columns (total 24 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   data             144316 non-null   object 
 1   trip_creation_time 144316 non-null   object 
 2   route_schedule_uuid 144316 non-null   object 
 3   route_type         144316 non-null   object 
 4   trip_uuid          144316 non-null   object 
 5   source_center       144316 non-null   object 
 6   source_name         144316 non-null   object 
 7   destination_center 144316 non-null   object 
 8   destination_name    144316 non-null   object 
 9   od_start_time      144316 non-null   object 
 10  od_end_time        144316 non-null   object 
 11  start_scan_to_end_scan 144316 non-null   float64
 12  is_cutoff          144316 non-null   bool   
 13  cutoff_factor      144316 non-null   int64  
 14  cutoff_timestamp   144316 non-null   object 
 15  actual_distance_to_destination 144316 non-null   float64
 16  actual_time         144316 non-null   float64
 17  osrm_time          144316 non-null   float64
 18  osrm_distance       144316 non-null   float64
 19  factor             144316 non-null   float64
 20  segment_actual_time 144316 non-null   float64
 21  segment_osrm_time   144316 non-null   float64
 22  segment_osrm_distance 144316 non-null   float64
 23  segment_factor      144316 non-null   float64
dtypes: bool(1), float64(10), int64(1), object(12)
memory usage: 26.6+ MB
```

```
date_cols = ['trip_creation_time', 'od_start_time', 'od_end_time', 'cutoff_timestamp']
df[date_cols] = df[date_cols].apply(lambda col: pd.to_datetime(col, format='mixed'))
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 144316 entries, 0 to 144866
Data columns (total 24 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   data             144316 non-null   object  
 1   trip_creation_time 144316 non-null   datetime64[ns]
 2   route_schedule_uuid 144316 non-null   object  
 3   route_type         144316 non-null   object  
 4   trip_uuid          144316 non-null   object  
 5   source_center       144316 non-null   object  
 6   source_name         144316 non-null   object  
 7   destination_center 144316 non-null   object  
 8   destination_name    144316 non-null   object  
 9   od_start_time      144316 non-null   datetime64[ns]
 10  od_end_time        144316 non-null   datetime64[ns]
 11  start_scan_to_end_scan 144316 non-null   float64 
 12  is_cutoff          144316 non-null   bool    
 13  cutoff_factor      144316 non-null   int64   
 14  cutoff_timestamp    144316 non-null   datetime64[ns]
 15  actual_distance_to_destination 144316 non-null   float64 
 16  actual_time         144316 non-null   float64 
 17  osrm_time          144316 non-null   float64 
 18  osrm_distance      144316 non-null   float64 
 19  factor             144316 non-null   float64 
 20  segment_actual_time 144316 non-null   float64 
 21  segment_osrm_time  144316 non-null   float64 
 22  segment_osrm_distance 144316 non-null   float64 
 23  segment_factor     144316 non-null   float64 
dtypes: bool(1), datetime64[ns](4), float64(10), int64(1), object(8)
memory usage: 26.6+ MB
```

3. Analyze structure and characteristics of the dataset

```
df.head()
```

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name	dest:
0	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78- b351-4c0e-a951- fa3d5c3...	Carting	153741093647649320	trip- IND388121AAA	Anand_VUNagar_DC (Gujarat)	
1	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78- b351-4c0e-a951- fa3d5c3...	Carting	153741093647649320	trip- IND388121AAA	Anand_VUNagar_DC (Gujarat)	
2	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78- b351-4c0e-a951- fa3d5c3...	Carting	153741093647649320	trip- IND388121AAA	Anand_VUNagar_DC (Gujarat)	
3	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78- b351-4c0e-a951- fa3d5c3...	Carting	153741093647649320	trip- IND388121AAA	Anand_VUNagar_DC (Gujarat)	
4	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78- b351-4c0e-a951- fa3d5c3...	Carting	153741093647649320	trip- IND388121AAA	Anand_VUNagar_DC (Gujarat)	

```
df.groupby('route_schedule_uuid')['trip_uuid'].nunique()
```

	trip_uuid
	route_schedule_uuid
thanos::sroute:0007affd-fd01-4cf0-8a4f-90419df059f7	24
thanos::sroute:00435307-de7f-4439-bd6a-5a2a9a3cd8bf	9
thanos::sroute:00a74fab-a3ac-44df-b83a-cbf181b4cd94	4
thanos::sroute:00b294b8-d2c3-4bca-a3be-684f46278bdd	8
thanos::sroute:01164881-301e-45f8-bacd-ee21c37f1cc4	15
...	...
thanos::sroute:ff52ef7a-4d0d-4063-9bfe-cc211728881b	16
thanos::sroute:ff6d6662-580b-43c3-810c-ba3027000f79	16
thanos::sroute:ff9b1c17-a70d-412a-acd1-5ab51d892d22	1
thanos::sroute:ffaf85f1-2f23-4367-aef7-c58044806911	13
thanos::sroute:ffffa2622-a170-4d08-b60b-38dfbae83869	2

1497 rows × 1 columns

dtype: int64

```
df.groupby('trip_uuid')['route_schedule_uuid'].nunique().sort_values(ascending=False).head()
```

	route_schedule_uuid
	trip_uuid
trip-153861118270144424	1
trip-153671041653548748	1
trip-153671042288605164	1
trip-153861034802474617	1
trip-153861033690433192	1

dtype: int64

Insights:

A route schedule can have many trips, but a trip can belong to only one route schedule

```
df[df['trip_uuid'] == 'trip-153741093647649320']
```

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name	di...
0	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78- b351-4c0e-a951- fa3d5c3...	Carting	trip- 153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	
df.columns[1]								
'trip_creation_time'								
2018-09-20								
group_cols = df.columns[0:12].tolist() group_cols								
df[df['trip_uuid'] == 'trip-153741093647649320'].groupby(group_cols).agg({'actual_distance_to_destination' : 'max','actual_#['actual_distance_to_destination'].max()								
4	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78- b351-4c0e-a951- fa3d5c3297ef	Carting	trip- 153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	
training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78- b351-4c0e-a951- fa3d5c3...	Carting	trip- 153741093647649320	IND388620AAB	Khambhat_MotvdDPP_D (Gujarat)		
0	training	02:35:36.476840	thanos::sroute:eb7bfc78- b351-4c0e-a951- fa3d5c3...	Carting	trip- 153741093647649320	IND388620AAB	Khambhat_MotvdDPP_D (Gujarat)	
df.groupby(group_cols)								
<pandas.core.groupby.generic.DataFrameGroupBy object at 0x7a7f87fb0f90>								
8	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78- b351-4c0e-a951- fa3d5c3...	Carting	trip- 153741093647649320	IND388620AAB	Khambhat_MotvdDPP_D (Gujarat)	
Insights	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78- b351-4c0e-a951- fa3d5c3...	Carting	trip- 153741093647649320	IND388620AAB	Khambhat_MotvdDPP_D (Gujarat)	
The trip start and end time (od_end_time) gives the timings of starting from a source center and destination center which can be multiple for a trip ID								
df								

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

144316 rows × 24 columns

#df['segment_key'] = df['trip_uuid'] + '_' + df['source_center'] + '_' + df['destination_center']

combined_cols = ['trip_uuid', 'source_center', 'destination_center']
df['segment_key2'] = df[combined_cols].agg('_'.join, axis=1)# #df['segment_actual_time_sum'] =
df.groupby('segment_key')[['segment_actual_time', 'segment_osrm_distance', 'segment_osrm_time']].sum()# df['segment_actual_time_sum'] = df.groupby('segment_key')['segment_actual_time'].sum()
df['segment_osrm_distance_sum'] = df.groupby('segment_key')['segment_osrm_distance'].sum()
df['segment_osrm_time_sum'] = df.groupby('segment_key')['segment_osrm_time'].sum()

df

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

144316 rows × 24 columns

df.columns

```
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'],
      dtype='object')
```

```
df2 = df.groupby(group_cols).agg({'actual_distance_to_destination' : 'max', 'actual_time' : 'max', 'osrm_time' : 'max', 'osrm_distance' : 'sum', 'segment_actual_time':'sum', 'segment_osrm_time':'sum', 'segment_osrm_distance':'sum'}).sort_values(['trip_uuid','od_start_time'])
```

		data	trip_creation_time	route_schedule_uuid	route_type		trip_uuid	source_center	source_r
0	training		2018-09-12 00:00:16.535741	thanos::sroute:d7c989ba-a29b-4a0b-b2f4-288cdc6...	FTL		trip-153671041653548748	IND462022AAA	Bhopal_Tnspo (Madhya Prad)
1	training		2018-09-12 00:00:16.535741	thanos::sroute:d7c989ba-a29b-4a0b-b2f4-288cdc6...	FTL		trip-153671041653548748	IND209304AAA	Kanpur_Central_H_6 (L Prad)
2	training		2018-09-12 00:00:22.886430	thanos::sroute:3a1b0ab2-bb0b-4c53-8c59-eb2a2c0...	Carting		trip-153671042288605164	IND572101AAA	Tumkur_Veersa (Karnat)
3	training		2018-09-12 00:00:22.886430	thanos::sroute:3a1b0ab2-bb0b-4c53-8c59-eb2a2c0...	Carting		trip-153671042288605164	IND561203AAB	Doddablpur_ChikaDPI (Karnat)

df2.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26223 entries, 0 to 26222
Data columns (total 19 columns):
 #   Column          Non-Null Count  Dtype    
--- 
 0   data            26223 non-null   object   
 1   trip_creation_time 26223 non-null   datetime64[ns]
 2   route_schedule_uuid 26223 non-null   object   
 3   route_type        26223 non-null   object   
 4   trip_uuid         26223 non-null   object   
 5   source_center     26223 non-null   object   
 6   source_name       26223 non-null   object   
 7   destination_center 26223 non-null   object   
 8   destination_name  26223 non-null   object   
 9   od_start_time    26223 non-null   datetime64[ns]
 10  od_end_time      26223 non-null   datetime64[ns]
 11  start_scan_to_end_scan 26223 non-null   float64 
 12  actual_distance_to_destination 26223 non-null   float64 
 13  actual_time      26223 non-null   float64 
 14  osrm_time        26223 non-null   float64 
 15  osrm_distance   26223 non-null   float64 
 16  segment_actual_time 26223 non-null   float64 
 17  segment_osrm_time 26223 non-null   float64 
 18  segment_osrm_distance 26223 non-null   float64 
dtypes: datetime64[ns](3), float64(8), object(8)
memory usage: 3.8+ MB
```

df2.shape

(26223, 19)

df2.index

RangeIndex(start=0, stop=26223, step=1)

```
df2['trip_time'] = df2['od_end_time'] - df2['od_start_time']
df2['trip_time'] = df2['trip_time'].dt.total_seconds()/60
df2
```

		data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_r
0	training		2018-09-12 00:00:16.535741	thanos::sroute:d7c989ba-a29b-4a0b-b2f4-288cdc6...	FTL	trip-153671041653548748	IND462022AAA	Bhopal_Tnspo (Madhya Prad)
1	training		2018-09-12 00:00:16.535741	thanos::sroute:d7c989ba-a29b-4a0b-b2f4-288cdc6...	FTL	trip-153671041653548748	IND209304AAA	Kanpur_Central_H_6 (L Prad)
2	training		2018-09-12 00:00:22.886430	thanos::sroute:3a1b0ab2-bb0b-4c53-8c59-eb2a2c0...	Carting	trip-153671042288605164	IND572101AAA	Tumkur_Veersa (Karnat)
3	training		2018-09-12 00:00:22.886430	thanos::sroute:3a1b0ab2-bb0b-4c53-8c59-eb2a2c0...	Carting	trip-153671042288605164	IND561203AAB	Doddablpur_ChikaDPI (Karnat)
4	training		2018-09-12 00:00:33.691250	thanos::sroute:de5e208e-7641-45e6-8100-4d9fb1e...	FTL	trip-153671043369099517	IND562132AAA	Bangalore_Nelmngl (Karnat)
...
26218	test		2018-10-03 23:59:14.390954	thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...	Carting	trip-153861115439069069	IND628204AAA	Tirchchndr_Shnmgr (Tamil Na)
26219	test		2018-10-03 23:59:14.390954	thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...	Carting	trip-153861115439069069	IND627657AAA	Thisayanvilai_UdnkdiR (Tamil Na)
26220	test		2018-10-03 23:59:14.390954	thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...	Carting	trip-153861115439069069	IND628613AAA	Peikulam_SriVnkpr (Tamil Na)
26221	test		2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	FTL	trip-153861118270144424	IND583201AAA	Hospet (Karnat)
26222	test		2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	FTL	trip-153861118270144424	IND583119AAA	Sandur_WrdN1DPI (Karnat)

26223 rows × 20 columns

```
# df2['trip_time'] = df2['od_end_time'] - df2['od_start_time']

# df2['trip_time'] = df2['trip_time'].dt.components.apply(lambda row: f"{int(row['hours']):02}:{int(row['minutes']):02}:{int(row['seconds']):02}" if row['seconds'] > 0 else f'{row["hours"]}:{row["minutes"]}:{row["seconds"]}' if row['minutes'] > 0 else f'{row["hours"]}:{row["minutes"]}' if row['seconds'] == 0 and row['minutes'] > 0 else f'{row["hours"]}' if row['minutes'] == 0 and row['seconds'] == 0 else None)
```

```
df2.sort_values('trip_time', ascending = False)
```

		data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name
24024	test		2018-10-01 23:35:54.432745	thanos::sroute:4316e05f- b4cc-4ea7-b801- 62a93ae...	Carting	trip-153843695443252828	IND764071AAB	Pappadahandi_Central
7962	training		2018-09-18 04:59:16.125309	thanos::sroute:6be6529b- f2ad-4714-b7ab- ac58f24...	FTL	trip-153724675612503042	IND000000ACB	Gurgaon_Bilaspur (H)
25658	test		2018-10-03 17:46:03.409692	thanos::sroute:0456b740- 1dad-4929-bbe0- 87d8843...	FTL	trip-153858876340944305	IND000000ACB	Gurgaon_Bilaspur (H)
9420	training		2018-09-19 13:44:58.665210	thanos::sroute:bc7dbb1d- 9379-4674-b8d3- f9c3b96...	FTL	trip-153736469866480991	IND000000ACB	Gurgaon_Bilaspur (H)
23564	test		2018-10-01 15:09:28.129568	thanos::sroute:3592c86e- c3d1-429b-917a- ebe9051...	FTL	trip-153840656812932039	IND712311AAA	Kolkata_Dankuni_H
...
8032	training		2018-09-18 06:31:52.481855	thanos::sroute:9f229200- bc86-4418-90ae- 6534983...	Carting	trip-153725231248161767	IND141010AAA	Ludhiana_DC (P)
15106	training		2018-09-23 23:23:55.668755	thanos::sroute:c446b063- ccbc-453b-9f94- 0697e8c...	FTL	trip-153774503566847992	IND484001AAA	Shahdol_Sohar (Madhya Pradesh)
10090	training		2018-09-19 23:08:07.327096	thanos::sroute:c1cc5ee0- efa2-4d88-b302- 88b6f94...	Carting	trip-153739848732673970	IND752050AAA	Khurdha_JatniDPP_D
2513	training		2018-09-13 23:15:28.857923	thanos::sroute:c57f3600- 80a6-44b6-835e- aad83da...	Carting	trip-153688052885763782	IND382421AAA	Gandhinagar_DC (P)
13265	training		2018-09-22 06:57:59.636076	thanos::sroute:028fad3e- 6945-411a-ab93- b5ac74c...	Carting	trip-153759947963581700	IND603203AAA	Chennai_Potheri (Tamil Nadu)

26223 rows × 20 columns

```
# df3 = df2.groupby(group_cols2).agg({'start_scan_to_end_scan' : 'sum', 'actual_distance_to_destination' : 'sum','actual_time' : 'sum'})
# df3
```

df2[df2['trip_uuid'] == 'trip-153861118270144424']

		data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name	distance
26221	test		2018-10-03 23:59:42.701692	thanos::sroute:412fea14- 6d1f-4222-8a5f- a517042...	FTL	trip-153861118270144424	IND583201AAA	Hospet (Karnataka)	10.0
26222	test		2018-10-03 23:59:42.701692	thanos::sroute:412fea14- 6d1f-4222-8a5f- a517042...	FTL	trip-153861118270144424	IND583119AAA	Sandur_WrdN1DPP_D (Karnataka)	10.0

df44

```
NameError: name 'df44' is not defined
Traceback (most recent call last)
<ipython-input-230-caace1988745> in <cell line: 0>()
----> 1 df44
```

NameError: name 'df44' is not defined

```
#df2['state'] =
df2['source_state'] = df2['source_name'].str.extract(r'\((.*?)\)')
df2['source_city'] = df2['source_name'].str.extract(r'^([_ ]+)')
df2['source_place'] = df2['source_name'].str.extract(r'_(.*)\_'')
```

df2

```
df2['dest_state'] = df2['destination_name'].str.extract(r'\((.*?)\)')
df2['dest_city'] = df2['destination_name'].str.extract(r'^([_ ]+)')
```

```
df2['dest_place'] = df2['destination_name'].str.extract(r'_(.*?)\_')
df2
```

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_r
0	training	2018-09-12 00:00:16.535741	thanos::sroute:d7c989ba-a29b-4a0b-b2f4-288cdc6...	FTL	trip-153671041653548748	IND462022AAA	Bhopal_Trnspo (Madhya Prad)
1	training	2018-09-12 00:00:16.535741	thanos::sroute:d7c989ba-a29b-4a0b-b2f4-288cdc6...	FTL	trip-153671041653548748	IND209304AAA	Kanpur_Central_H_6 (U Prad)
2	training	2018-09-12 00:00:22.886430	thanos::sroute:3a1b0ab2-bb0b-4c53-8c59-eb2a2c0...	Carting	trip-153671042288605164	IND572101AAA	Tumkur_Veersa (Karnat)
3	training	2018-09-12 00:00:22.886430	thanos::sroute:3a1b0ab2-bb0b-4c53-8c59-eb2a2c0...	Carting	trip-153671042288605164	IND561203AAB	Doddablpur_ChikaDPI (Karnat)
4	training	2018-09-12 00:00:33.691250	thanos::sroute:de5e208e-7641-45e6-8100-4d9fb1e...	FTL	trip-153671043369099517	IND562132AAA	Bangalore_Nelmngl (Karnat)
...
26218	test	2018-10-03 23:59:14.390954	thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...	Carting	trip-153861115439069069	IND628204AAA	Tirchchndr_Shnmgr (Tamil Na)
26219	test	2018-10-03 23:59:14.390954	thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...	Carting	trip-153861115439069069	IND627657AAA	Thisayanvilai_UdnkdiRl (Tamil Na)
26220	test	2018-10-03 23:59:14.390954	thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...	Carting	trip-153861115439069069	IND628613AAA	Peikulam_SriVnktp (Tamil Na)
26221	test	2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	FTL	trip-153861118270144424	IND583201AAA	Hospet (Karnat)
26222	test	2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	FTL	trip-153861118270144424	IND583119AAA	Sandur_WrdN1DP (Karnat)

26223 rows × 26 columns

```
df[df['trip_uuid'] == 'trip-153671041653548748']
```

```
df2['trip_creation_hour'] = df2['trip_creation_time'].dt.hour
df2['trip_creation_day'] = df2['trip_creation_time'].dt.day
df2['trip_creation_month'] = df2['trip_creation_time'].dt.month
df2['trip_creation_year'] = df2['trip_creation_time'].dt.year
df2
```

		data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_r
0	training		2018-09-12 00:00:16.535741	thanos::sroute:d7c989ba-a29b-4a0b-b2f4-288cdc6...	FTL	trip-153671041653548748	IND462022AAA	Bhopal_Tnspo (Madhya Prad)
1	training		2018-09-12 00:00:16.535741	thanos::sroute:d7c989ba-a29b-4a0b-b2f4-288cdc6...	FTL	trip-153671041653548748	IND209304AAA	Kanpur_Central_H_6 (L Prad)
2	training		2018-09-12 00:00:22.886430	thanos::sroute:3a1b0ab2-bb0b-4c53-8c59-eb2a2c0...	Carting	trip-153671042288605164	IND572101AAA	Tumkur_Veersa (Karnat)
3	training		2018-09-12 00:00:22.886430	thanos::sroute:3a1b0ab2-bb0b-4c53-8c59-eb2a2c0...	Carting	trip-153671042288605164	IND561203AAB	Doddablpur_ChikaDPI (Karnat)
4	training		2018-09-12 00:00:33.691250	thanos::sroute:de5e208e-7641-45e6-8100-4d9fb1e...	FTL	trip-153671043369099517	IND562132AAA	Bangalore_Nelmngl (Karnat)
...
26218	test		2018-10-03 23:59:14.390954	thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...	Carting	trip-153861115439069069	IND628204AAA	Tirchchndr_Shnmgr (Tamil Na)
26219	test		2018-10-03 23:59:14.390954	thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...	Carting	trip-153861115439069069	IND627657AAA	Thisayanvilai_UdnkdiR (Tamil Na)
26220	test		2018-10-03 23:59:14.390954	thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...	Carting	trip-153861115439069069	IND628613AAA	Peikulam_SriVnkpr (Tamil Na)
26221	test		2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	FTL	trip-153861118270144424	IND583201AAA	Hospet (Karnat)
26222	test		2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	FTL	trip-153861118270144424	IND583119AAA	Sandur_WrdN1DPI (Karnat)

26223 rows × 30 columns

```
create_segment_dict = {
    'data':'first',
    'trip_creation_time':'first',
    'trip_creation_hour':'first',
    'trip_creation_day':'min',
    'trip_creation_month':'min',
    'trip_creation_year':'min',
    'route_schedule_uuid':'first',
    'route_type':'first',
    #'trip_uuid':'first',
    'source_city':'first',
    'source_state':'first',
    'source_place':'first',
    'dest_place':'last',
    'dest_city':'last',
    'dest_state':'last',
    'od_start_time':'first',
    'od_end_time':'last',
    'start_scan_to_end_scan':'sum',
    'actual_distance_to_destination':'sum',
    'actual_time':'sum',
    'osrm_time':'sum',
    'osrm_distance':'sum',
    'segment_actual_time':'sum',
    'segment_osrm_time':'sum',
    'segment_osrm_distance':'sum',
    'trip_time':'sum'
}
```

```
#group_cols2 = df2.columns[0:11].tolist()
df2 = df2.groupby('trip_uuid').agg(create_segment_dict).reset_index()
df3 = df2[::]
df3
```

	trip_uuid	data	trip_creation_time	trip_creation_hour	trip_creation_day	trip_creation_month	trip_crea
0	153671041653548748	trip-training	2018-09-12 00:00:16.535741	0	12	9	
1	153671042288605164	trip-training	2018-09-12 00:00:22.886430	0	12	9	
2	153671043369099517	trip-training	2018-09-12 00:00:33.691250	0	12	9	
3	153671046011330457	trip-training	2018-09-12 00:01:00.113710	0	12	9	
4	153671052974046625	trip-training	2018-09-12 00:02:09.740725	0	12	9	
...
14782	153861095625827784	trip-test	2018-10-03 23:55:56.258533	23	3	10	
14783	153861104386292051	trip-test	2018-10-03 23:57:23.863155	23	3	10	
14784	153861106442901555	trip-test	2018-10-03 23:57:44.429324	23	3	10	
14785	153861115439069069	trip-test	2018-10-03 23:59:14.390954	23	3	10	
14786	153861118270144424	trip-test	2018-10-03 23:59:42.701692	23	3	10	

14787 rows × 26 columns

```
df2[df2['trip_uuid'] == 'trip-153671041653548748']
```

```
df2.columns
```

```
num_cols = df2.columns[14:]
df2[num_cols].boxplot(rot=25, figsize=(25,8))
```

```
q1 = df2[num_cols].quantile(0.25)
q3 = df2[num_cols].quantile(0.75)

iqr = q3-q1

outlier = q3+1.5*iqr

#outlier
iqr
df2 = df2[((df2[num_cols]<=(q1-1.5*iqr)) | (df2[num_cols]>=outlier)).any(axis=1)]
# df2 = df2.reset_index(drop=True)
df2[num_cols].boxplot(rot=25, figsize=(25,8))
```

```
df2
```

```
sns.kdeplot(df2['trip_time'], color = 'blue')
sns.kdeplot(df2['start_scan_to_end_scan'], color = 'pink')
```

Handling Categorical Variables

One Hot Encoding on Route types

```
df2['route_type'].unique()
```

```
df2['route_type'] = df2['route_type'].map({'Carting':0,'FTL':1})
df2['route_type']
```

Normalize/Standardize the numerical features using StandardScaler

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

```
norm_df = df2[::]
```

```
scaler = StandardScaler()
scaler.fit(norm_df[num_cols])
```

```
norm_df[num_cols] = scaler.transform(norm_df[num_cols])
norm_df[num_cols].describe()
```

Hypothesis Testing

Actual_time vs OSRM_time

```
df2[['actual_time','osrm_time']]
```

```
# T-test relative
# Ho: The mean difference between actual time and osrm time is zero
# Ha: There is a significant different between actual time and osrm time
```

```
from statsmodels.graphics.gofplots import qqplot
```

```
qqplot(df2['actual_time'],line='s')
qqplot(df2['osrm_time'],line='s')
plt.show()
```

```
sns.histplot(df2['actual_time'])
sns.histplot(df2['osrm_time'])
```

Insights:

The above two plots (qq & hist) clearly say that both the actual time and osrm time data is not normally distributed

```
stats.anderson(df2['actual_time'], dist = 'norm')
```

```
stats.wilcoxon(df2['actual_time'],df2['osrm_time'])
```

```
differences = df2['actual_time'] - df2['osrm_time']
sns.histplot(differences, edgecolor='black')
```

Insights:

With all the above tests, we can conclude that the difference between actual time and osrm time is significant

Recommendation:

Need to investigate why there is such a large gap between the estimated delivery time and the actual delivery time

Actual time VS Segment Actual time aggregated

```
df2[['actual_time','segment_actual_time']]
```

```
sns.histplot(df2['actual_time'])
sns.histplot(df2['segment_actual_time'])
```

```
qqplot(df2['actual_time'],line='s')
qqplot(df2['segment_actual_time'],line='s')
plt.show()
```

```
stats.anderson(df2['actual_time'],dist='norm')
```

```
stats.anderson(df2['segment_actual_time'],dist = 'norm')
```

Insights:

All the above plots and test say that the data is not normally distributed

```
stats.wilcoxon(df2['actual_time'],df2['segment_actual_time'])
```

```
stats.ttest_rel(df2['actual_time'],df2['segment_actual_time'])
```

```
differences = df2['actual_time'] - df2['segment_actual_time']
```

```
sns.histplot(differences)
```

Insights:

Based on the results of the test we performed, we can conclude that there is a significant difference between the actual time and the segment actual time.

Recommendations:

The data engineering team should probably investigate if the data is properly captured from the source and transferred through the data pipelines

▼ OSRM distance vs segment OSRM distance aggregated value

```
df2[['osrm_distance','segment_osrm_distance']]
```

```
sns.histplot(df2['osrm_distance'])
sns.histplot(df2['segment_osrm_distance'])
```

```
from statsmodels.graphics.gofplots import qqplot
```

```
qqplot(df2['osrm_distance'], line='s')
qqplot(df2['segment_osrm_distance'],line='s')
plt.show()
```

```
stats.anderson(df2['osrm_distance'],dist = 'norm')
stats.anderson(df2['segment_osrm_distance'], dist = 'norm')
```

Insights:

The above plots and test say that the data is not normally distributed

```
stats.wilcoxon(df2['osrm_distance'],df2['segment_osrm_distance'])
```

```
osrm_differences = df2['osrm_distance'] - df2['segment_osrm_distance']
```

```
sns.histplot(osrm_differences)
```

```
sns.boxplot(osrm_differences)
```

Insights:

The boxplot above shows there are outliers. Based on the results of the tests we performed, we can conclude that there is a significant difference between the osrm distance and the segment osrm distance.

Recommendations:

The data engineering team should probably investigate if the data is properly captured from the source and transferred through the data pipelines

▼ OSRM time VS Segment OSRM time

```
df2[['osrm_time', 'segment_osrm_time']]  
  
sns.histplot(df2['osrm_time'], kde = True)  
  
sns.histplot(df2['segment_osrm_time'], kde = True, color = 'red')  
  
from statsmodels.graphics.gofplots import qqplot  
  
qqplot(df2['osrm_distance'], line='s')  
qqplot(df2['segment_osrm_distance'], line='s')  
plt.show()  
  
stats.anderson(df2['osrm_time'])  
  
stats.anderson(df2['segment_osrm_time'])
```

Insights:

The above plots and test say that the data is not normally distributed

```
stats.wilcoxon(df2['osrm_time'],df2['segment_osrm_time'])  
  
time_differences = df2['osrm_time'] - df2['segment_osrm_time']  
  
sns.histplot(time_differences, kde= True)  
  
sns.boxplot(time_differences)
```

Insights:

The boxplot above shows there are outliers. Based on the results of the tests we performed, we can conclude that there is a significant difference between the osrm time and the segment osrm time.

Recommendations:

The data engineering team should probably investigate if the data is properly captured from the source and transferred through the data pipelines

▼ Most ordered States

```
df2['dest_state'].value_counts()  
  
df2['dest_city'].value_counts()
```

Insights:

Most of the orders are coming from Maharashtra and Karnataka states, and Mumbai and Bengaluru cities respectively.

Recommendation:

Should have more distribution centers near these cities to deliver the products quickly and efficiently.

```
df2['trip_creation_hour'].value_counts()
```

Insights:

Most of the trips are created between 8 PM and 1 AM, and the max trips begin around 11 PM

Recommendations:

Need to make sure that the drivers are available during these times

```
df2[['source_city','dest_city']].value_counts()
```

count