

✓ Problem Statement:

Delhivery, India's largest and fastest-growing fully integrated logistics company, seeks to leverage its vast data resources to enhance operational efficiency, improve decision-making, and drive profitability. The company has vast amounts of raw data emerging from its data engineering pipelines, but transforming this raw data into actionable insights remains a challenge. As a Data Scientist the goal is to process raw data to remove inconsistencies, handle missing values, and prepare it for analysis. Extract meaningful features from the raw data that can be used by the Data Science team for forecasting models.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
from IPython.display import Image; Image(url='https://finshiksha.com/wp-content/uploads/2022/05/Delhivery-Banner-Image.jpg')
```



!gdown https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/001/551/original/delhivery_data.csv?1642751181



Downloading...


From: https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/001/551/original/delhivery_data.csv?1642751181

To: /content/delhivery_data.csv?1642751181

100% 55.6M/55.6M [00:02<00:00, 24.7MB/s]

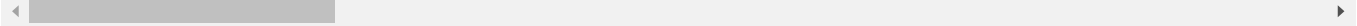
```
data = pd.read_csv('delhivery_data.csv?1642751181')
```

```
data.head()
```



	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name	destination_
0	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78- b351-4c0e-a951- fa3d5c3...	Carting	153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	IND3886
1	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78- b351-4c0e-a951- fa3d5c3...	Carting	153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	IND3886
2	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78- b351-4c0e-a951- fa3d5c3...	Carting	153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	IND3886
3	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78- b351-4c0e-a951- fa3d5c3...	Carting	153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	IND3886
4	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78- b351-4c0e-a951- fa3d5c3...	Carting	153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	IND3886

5 rows × 24 columns




data.shape




(144867, 24)

data.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
```

data.isnull().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

data.describe()



	start_scan_to_end_scan	cutoff_factor	actual_distance_to_destination	actual_time	osrm_time	osrm_distance	fac
count	144867.000000	144867.000000	144867.000000	144867.000000	144867.000000	144867.000000	144867.000
mean	961.262986	232.926567	234.073372	416.927527	213.868272	284.771297	2.120
std	1037.012769	344.755577	344.990009	598.103621	308.011085	421.119294	1.715
min	20.000000	9.000000	9.000045	9.000000	6.000000	9.008200	0.144
25%	161.000000	22.000000	23.355874	51.000000	27.000000	29.914700	1.604
50%	449.000000	66.000000	66.126571	132.000000	64.000000	78.525800	1.857
75%	1634.000000	286.000000	286.708875	513.000000	257.000000	343.193250	2.213
max	7898.000000	1927.000000	1927.447705	4532.000000	1686.000000	2326.199100	77.387

data.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
```

```
data['Segment_Key'] = data['trip_uuid'] + data['source_center'] + data['destination_center']
```

Removing NULL Values

```
data = data.dropna(how='any')
data = data.reset_index(drop=True)

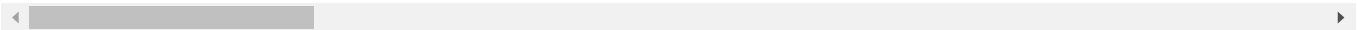
data['od_end_time'] = pd.to_datetime(data['od_end_time'], errors='coerce')
data['od_start_time'] = pd.to_datetime(data['od_start_time'], errors='coerce')
```

```
data.head(5)
```



	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name	destination_
0	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	IND3886
1	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	IND3886
2	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	IND3886
3	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	IND3886
4	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	153741093647649320	IND388121AAA	Anand_VUNagar_DC (Gujarat)	IND3886

5 rows × 25 columns



```
data['od_start_time'].isna().sum()
```



```
0
```

```
data['od_end_time'].isna().sum()
```



```
0
```

```
data['source_name'].value_counts()
```



	count
source_name	
Gurgaon_Bilaspur_HB (Haryana)	23267
Bangalore_Nelmngla_H (Karnataka)	9975
Bhiwandi_Mankoli_HB (Maharashtra)	9088
Pune_Tathawde_H (Maharashtra)	4061
Hyderabad_Shamshbd_H (Telangana)	3340
...	...
Badkulla_Central_DPP_1 (West Bengal)	1
Mahasamund_RajpurRD_D (Chhattisgarh)	1
Tiruchi_Samyaprm_D (Tamil Nadu)	1
Islampure_Central_DPP_2 (West Bengal)	1
Bhadra_GMndiDPP_D (Rajasthan)	1

1496 rows × 1 columns



```
data['destination_name'].value_counts()
```



	count
destination_name	
Gurgaon_Bilaspur_HB (Haryana)	15192
Bangalore_Nelmngla_H (Karnataka)	11019
Bhiwandi_Mankoli_HB (Maharashtra)	5492
Hyderabad_Shamshbd_H (Telangana)	5142
Kolkata_Dankuni_HB (West Bengal)	4892
...	...
Durg_Bhilai_DC (Chhattisgarh)	1
Bhadohi_Rajpura_D (Uttar Pradesh)	1
Vijayawada (Andhra Pradesh)	1
Daman_DC (Daman & Diu)	1
Berhampur_Chatrpr_DC (Orissa)	1

1466 rows × 1 columns



```
data.info()
```



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 144316 entries, 0 to 144315
Data columns (total 25 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 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 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
24 Segment_Key                 144316 non-null object
dtypes: bool(1), datetime64[ns](2), float64(10), int64(1), object(11)
memory usage: 26.6+ MB

```


```
segment_cols = ['segment_actual_time', 'segment_osrm_time', 'segment_osrm_distance']
```

```



for col in segment_cols:
    data[col + '_sum'] = data.groupby('Segment_Key')[col].cumsum()

```

```
data[[col + '_sum' for col in segment_cols]]
```



	segment_actual_time_sum	segment_osrm_time_sum	segment_osrm_distance_sum
0	14.0	11.0	11.9653
1	24.0	20.0	21.7243
2	40.0	27.0	32.5395
3	61.0	39.0	45.5619
4	67.0	44.0	49.4772
...
144311	92.0	94.0	65.3487
144312	118.0	115.0	82.7212
144313	138.0	149.0	103.4265
144314	155.0	176.0	122.3150
144315	423.0	185.0	131.1238

144316 rows x 3 columns

```

create_segment_dict = {
    'data' : 'first',
    'trip_uuid' : 'first',
    'trip_creation_time' : 'first',
    'route_schedule_uuid' : 'first',
    'route_type' : 'first',
    'source_center' : 'first',
    'source_name' : 'first',

    'destination_center' : 'last',
    'destination_name' : 'last',

    'od_start_time' : 'first',
    'od_end_time' : 'first',
    'start_scan_to_end_scan' : 'first',

    'actual_distance_to_destination' : 'last',
    'actual_time' : 'last',

    'osrm_time' : 'last',
    'osrm_distance' : 'last',

    'segment_actual_time_sum' : 'last',
    'segment_osrm_time_sum' : 'last',
    'segment_osrm_distance_sum' : 'last',
}

```

✓ Group the data by Segment_Key and apply the aggregation dictionary

```
segment = data.groupby('Segment_Key').agg(create_segment_dict)
```

```
segment
```

		data	trip_uuid	trip_creation_time	route_schedule_uuid	route_type	so
Segment_Key							
trip-153671041653548748IND209304AAAIND000000ACB	training	153671041653548748	trip-153671041653548748	2018-09-12 00:00:16.535741	thanos::sroute:d7c989ba-a29b-4a0b-b2f4-288cdc6...	FTL	IN
trip-153671041653548748IND462022AAAIND209304AAA	training	153671041653548748	trip-153671041653548748	2018-09-12 00:00:16.535741	thanos::sroute:d7c989ba-a29b-4a0b-b2f4-288cdc6...	FTL	IN
trip-153671042288605164IND561203AABIND562101AAA	training	153671042288605164	trip-153671042288605164	2018-09-12 00:00:22.886430	thanos::sroute:3a1b0ab2-bb0b-4c53-8c59-eb2a2c0...	Carting	IN
trip-153671042288605164IND572101AAAIND561203AAB	training	153671042288605164	trip-153671042288605164	2018-09-12 00:00:22.886430	thanos::sroute:3a1b0ab2-bb0b-4c53-8c59-eb2a2c0...	Carting	IN
trip-153671043369099517IND000000ACBIND160002AAC	training	153671043369099517	trip-153671043369099517	2018-09-12 00:00:33.691250	thanos::sroute:de5e208e-7641-45e6-8100-4d9fb1e...	FTL	IN
...
trip-153861115439069069IND628204AAAIND627657AAA	test	153861115439069069	trip-153861115439069069	2018-10-03 23:59:14.390954	thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...	Carting	IN
trip-153861115439069069IND628613AAAIND627005AAA	test	153861115439069069	trip-153861115439069069	2018-10-03 23:59:14.390954	thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...	Carting	IN
trip-153861115439069069IND628801AAAIND628204AAA	test	153861115439069069	trip-153861115439069069	2018-10-03 23:59:14.390954	thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...	Carting	IN
trip-153861118270144424IND583119AAAIND583101AAA	test	153861118270144424	trip-153861118270144424	2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	FTL	IN
trip-153861118270144424IND583201AAAIND583119AAA	test	153861118270144424	trip-153861118270144424	2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	FTL	IN

26222 rows × 19 columns

Next steps:

Generate code with segment

 View recommended plots

New interactive sheet

```
segment = segment.sort_values(by=['Segment_Key', 'od_end_time'],ascending = True).reset_index()

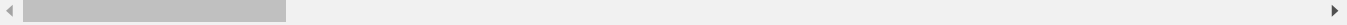
segment.drop(segment['od_start_time'].isna().sum(),inplace=True)

segment
```




	Segment_Key	data	trip_uuid	trip_creation_time	route_schedule_uuid	route_ty
1	trip-153671041653548748IND462022AAAIND209304AAA	training	trip-153671041653548748	2018-09-12 00:00:16.535741	thanos::sroute:d7c989ba-a29b-4a0b-b2f4-288cdc6...	F
2	trip-153671042288605164IND561203AABIND562101AAA	training	trip-153671042288605164	2018-09-12 00:00:22.886430	thanos::sroute:3a1b0ab2-bb0b-4c53-8c59-eb2a2c0...	Cart
3	trip-153671042288605164IND572101AAAINDD561203AAB	training	trip-153671042288605164	2018-09-12 00:00:22.886430	thanos::sroute:3a1b0ab2-bb0b-4c53-8c59-eb2a2c0...	Cart
4	trip-153671043369099517IND000000ACBIND160002AAC	training	trip-153671043369099517	2018-09-12 00:00:33.691250	thanos::sroute:de5e208e-7641-45e6-8100-4d9fb1e...	F
5	trip-153671043369099517IND562132AAAIND000000ACB	training	trip-153671043369099517	2018-09-12 00:00:33.691250	thanos::sroute:de5e208e-7641-45e6-8100-4d9fb1e...	F
...
26217	trip-153861115439069069IND628204AAAINDD627657AAA	test	trip-153861115439069069	2018-10-03 23:59:14.390954	thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...	Cart
26218	trip-153861115439069069IND628613AAAINDD627005AAA	test	trip-153861115439069069	2018-10-03 23:59:14.390954	thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...	Cart
26219	trip-153861115439069069IND628801AAAINDD628204AAA	test	trip-153861115439069069	2018-10-03 23:59:14.390954	thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...	Cart
26220	trip-153861118270144424IND583119AAAINDD583101AAA	test	trip-153861118270144424	2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	F
26221	trip-153861118270144424IND583201AAAINDD583119AAA	test	trip-153861118270144424	2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	F

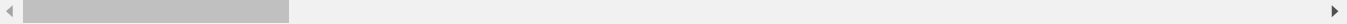
26221 rows × 20 columns



```
segment[segment['trip_uuid'] == 'trip-153671042288605164']
```



	Segment_Key	data	trip_uuid	trip_creation_time	route_schedule_uuid	route_type
2	trip-153671042288605164IND561203AABIND562101AAA	training	trip-153671042288605164	2018-09-12 00:00:22.886430	thanos::sroute:3a1b0ab2-bb0b-4c53-8c59-eb2a2c0...	Carting
3	trip-153671042288605164IND572101AAAINDD561203AAB	training	trip-153671042288605164	2018-09-12 00:00:22.886430	thanos::sroute:3a1b0ab2-bb0b-4c53-8c59-eb2a2c0...	Carting



```
segment.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26221 entries, 1 to 26221
Data columns (total 20 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Segment_Key                          26221 non-null  object
1   data                                 26221 non-null  object
2   trip_uuid                            26221 non-null  object
3   trip_creation_time                   26221 non-null  object
4   route_schedule_uuid                 26221 non-null  object
5   route_type                          26221 non-null  object
6   source_center                       26221 non-null  object
7   source_name                         26221 non-null  object
8   destination_center                  26221 non-null  object
9   destination_name                    26221 non-null  object
10  od_start_time                       26221 non-null  datetime64[ns]
11  od_end_time                         26221 non-null  datetime64[ns]
12  start_scan_to_end_scan              26221 non-null  float64
13  actual_distance_to_destination      26221 non-null  float64
14  actual_time                         26221 non-null  float64
15  osrm_time                           26221 non-null  float64
16  osrm_distance                       26221 non-null  float64
17  segment_actual_time_sum             26221 non-null  float64
18  segment_osrm_time_sum              26221 non-null  float64
19  segment_osrm_distance_sum           26221 non-null  float64
dtypes: datetime64[ns](2), float64(8), object(10)
memory usage: 4.0+ MB
```


✓ Calculate the time difference in hours between od_start_time and od_end_time

```
segment['od_diff_hours'] = (segment['od_end_time'] - segment['od_start_time']).dt.total_seconds() / (60)
segment['od_diff_hours']
```



	od_diff_hours
1	999.505379
2	58.832388
3	122.779486
4	834.638929
5	3099.723591
...	...
26217	62.115193
26218	91.087797
26219	44.174403
26220	287.474007
26221	66.933565

26221 rows × 1 columns

◀		▶
---	--	---

segment



	Segment_Key	data	trip_uuid	trip_creation_time	route_schedule_uuid	route_ty
1	trip-153671041653548748IND462022AAAIND209304AAA	training	trip-153671041653548748	2018-09-12 00:00:16.535741	thanos::sroute:d7c989ba-a29b-4a0b-b2f4-288cdc6...	F
2	trip-153671042288605164IND561203AABIND562101AAA	training	trip-153671042288605164	2018-09-12 00:00:22.886430	thanos::sroute:3a1b0ab2-bb0b-4c53-8c59-eb2a2c0...	Cart
3	trip-153671042288605164IND572101AAAINDD561203AAB	training	trip-153671042288605164	2018-09-12 00:00:22.886430	thanos::sroute:3a1b0ab2-bb0b-4c53-8c59-eb2a2c0...	Cart
4	trip-153671043369099517IND000000ACBIND160002AAC	training	trip-153671043369099517	2018-09-12 00:00:33.691250	thanos::sroute:de5e208e-7641-45e6-8100-4d9fb1e...	F
5	trip-153671043369099517IND562132AAAIND000000ACB	training	trip-153671043369099517	2018-09-12 00:00:33.691250	thanos::sroute:de5e208e-7641-45e6-8100-4d9fb1e...	F
...
26217	trip-153861115439069069IND628204AAAIND627657AAA	test	trip-153861115439069069	2018-10-03 23:59:14.390954	thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...	Cart
26218	trip-153861115439069069IND628613AAAIND627005AAA	test	trip-153861115439069069	2018-10-03 23:59:14.390954	thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...	Cart
26219	trip-153861115439069069IND628801AAAINDD628204AAA	test	trip-153861115439069069	2018-10-03 23:59:14.390954	thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...	Cart
26220	trip-153861118270144424IND583119AAAINDD583101AAA	test	trip-153861118270144424	2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	F
26221	trip-153861118270144424IND583201AAAINDD583119AAA	test	trip-153861118270144424	2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	F

26221 rows × 21 columns

◀		▶
---	--	---

✓ Creating trip dictionary to analyse trip wise data

```
create_trip_dict = {
    'data': 'first',
    'trip_creation_time': 'first',
    'route_schedule_uuid': 'first',
    'route_type': 'first',
    'trip_uuid': 'first',

    'source_center': 'first',
    'source_name': 'first',

    'destination_center': 'last',
    'destination_name': 'last',

    'start_scan_to_end_scan': 'sum',
    'od_diff_hours': 'sum',


    'actual_distance_to_destination': 'sum',
    'actual_time': 'sum',
    'osrm_time': 'sum',
    'osrm_distance': 'sum',

    'segment_actual_time_sum': 'sum',
    'segment_osrm_distance_sum': 'sum',
    'segment_osrm_time_sum': 'sum',

}

trip = segment.groupby('trip_uuid').agg(create_trip_dict).reset_index(drop=True)
```

trip



	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name	dest
0	training	2018-09-12 00:00:16.535741	thanos::sroute:d7c989ba- a29b-4a0b-b2f4- 288cdc6...	FTL	trip- 153671041653548748	IND462022AAA	Bhopal_Trnsport_H (Madhya Pradesh)	
1	training	2018-09-12 00:00:22.886430	thanos::sroute:3a1b0ab2- bb0b-4c53-8c59- eb2a2c0...	Carting	trip- 153671042288605164	IND561203AAB	Doddablpur_ChikaDPP_D (Karnataka)	
2	training	2018-09-12 00:00:33.691250	thanos::sroute:de5e208e- 7641-45e6-8100- 4d9fb1e...	FTL	trip- 153671043369099517	IND000000ACB	Gurgaon_Bilaspur_HB (Haryana)	
3	training	2018-09-12 00:01:00.113710	thanos::sroute:f0176492- a679-4597-8332- bbd1c7f...	Carting	trip- 153671046011330457	IND400072AAB	Mumbai Hub (Maharashtra)	
4	training	2018-09-12 00:02:09.740725	thanos::sroute:d9f07b12- 65e0-4f3b-bec8- df06134...	FTL	trip- 153671052974046625	IND583101AAA	Bellary_Dc (Karnataka)	
...
14782	test	2018-10-03 23:55:56.258533	thanos::sroute:8a120994- f577-4491-9e4b- b7e4a14...	Carting	trip- 153861095625827784	IND160002AAC	Chandigarh_Mehmdpur_H (Punjab)	
14783	test	2018-10-03 23:57:23.863155	thanos::sroute:b30e1ec3- 3bfa-4bd2-a7fb- 3b75769...	Carting	trip- 153861104386292051	IND121004AAB	FBD_Balabhgarh_DPC (Haryana)	
14784	test	2018-10-03 23:57:44.429324	thanos::sroute:5609c268- e436-4e0a-8180- 3db4a74...	Carting	trip- 153861106442901555	IND208006AAA	Kanpur_GovndNgr_DC (Uttar Pradesh)	
14785	test	2018-10-03 23:59:14.390954	thanos::sroute:c5f2ba2c- 8486-4940-8af6- d1d2a6a...	Carting	trip- 153861115439069069	IND627005AAA	Tirunelveli_VdkkuSrt_I (Tamil Nadu)	
14786	test	2018-10-03 23:59:42.701692	thanos::sroute:412fea14- 6d1f-4222-8a5f- a517042...	FTL	trip- 153861118270144424	IND583119AAA	Sandur_WrdN1DPP_D (Karnataka)	

14787 rows × 18 columns


Next steps:

Generate code with trip



 View recommended plots

New interactive sheet

```
trip[['actual_time', 'segment_actual_time_sum']]
```




	actual_time	segment_actual_time_sum
0	830.0	820.0
1	143.0	141.0
2	3347.0	3308.0
3	59.0	59.0
4	341.0	340.0
...
14782	83.0	82.0
14783	21.0	21.0
14784	282.0	281.0
14785	264.0	258.0
14786	275.0	274.0





14787 rows x 2 columns

```
sorted_trip = trip.sort_values(by=['actual_time', 'segment_actual_time_sum'], ascending=False)
sorted_trip[['actual_time', 'segment_actual_time_sum']]
```




	actual_time	segment_actual_time_sum
13310	6265.0	6230.0
10975	5804.0	5768.0
11581	5465.0	5427.0
13719	5067.0	5032.0
9587	5064.0	5028.0
...
8955	10.0	10.0
10909	10.0	10.0
14388	10.0	10.0
11031	9.0	9.0
11074	9.0	9.0




14787 rows x 2 columns

```
trip[trip['trip_uuid']=='trip-153671042288605164']
```





	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name	destination
1	training	2018-09-12 00:00:22.886430	thanos::route:3a1b0ab2- bb0b-4c53-8c59- eb2a2c0...	Carting	trip- 153671042288605164	IND561203AAB	Doddablpur_ChikaDPP_D (Karnataka)	INC


```
trip['route_type'].value_counts().reset_index()
```




	route_type	count
0	Carting	8906
1	FTL	5881



```
trip[['segment_actual_time_sum', 'od_diff_hours']]
```




	segment_actual_time_sum	od_diff_hours
0	820.0	999.505379
1	141.0	181.611874
2	3308.0	3934.362520
3	59.0	100.494935
4	340.0	718.349042
...
14782	82.0	258.028928
14783	21.0	60.590521
14784	281.0	422.119867
14785	258.0	348.512862
14786	274.0	354.407571



14787 rows × 2 columns

```
trip['Actual_diff_hours'] = trip['segment_actual_time_sum'] - trip['actual_time']/ (60)
```

```
trip['Actual_diff_hours']
```



	Actual_diff_hours
0	806.166667
1	138.616667
2	3252.216667
3	58.016667
4	334.316667
...	...
14782	80.616667
14783	20.650000
14784	276.300000
14785	253.600000
14786	269.416667

14787 rows × 1 columns

```
trip
```

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_center	source_name	dest
0	training	2018-09-12 00:00:16.535741	thanos::sroute:d7c989ba-a29b-4a0b-b2f4-288cdc6...	FTL	153671041653548748	trip-IND462022AAA	Bhopal_Trmsport_H (Madhya Pradesh)	
1	training	2018-09-12 00:00:22.886430	thanos::sroute:3a1b0ab2-bb0b-4c53-8c59-eb2a2c0...	Carting	153671042288605164	trip-IND561203AAB	Doddablpur_ChikaDPP_D (Karnataka)	
2	training	2018-09-12 00:00:33.691250	thanos::sroute:de5e208e-7641-45e6-8100-4d9fb1e...	FTL	153671043369099517	trip-IND000000ACB	Gurgaon_Bilaspur_HB (Haryana)	
3	training	2018-09-12 00:01:00.113710	thanos::sroute:f0176492-a679-4597-8332-bbd1c7f...	Carting	153671046011330457	trip-IND400072AAB	Mumbai Hub (Maharashtra)	
4	training	2018-09-12 00:02:09.740725	thanos::sroute:d9f07b12-65e0-4f3b-bec8-df06134...	FTL	153671052974046625	trip-IND583101AAA	Bellary_Dc (Karnataka)	
...
14782	test	2018-10-03 23:55:56.258533	thanos::sroute:8a120994-f577-4491-9e4b-b7e4a14...	Carting	153861095625827784	trip-IND160002AAC	Chandigarh_Mehmdpur_H (Punjab)	
14783	test	2018-10-03 23:57:23.863155	thanos::sroute:b30e1ec3-3bfa-4bd2-a7fb-3b75769...	Carting	153861104386292051	trip-IND121004AAB	FBD_Balabhgarh_DPC (Haryana)	
14784	test	2018-10-03 23:57:44.429324	thanos::sroute:5609c268-e436-4e0a-8180-3db4a74...	Carting	153861106442901555	trip-IND208006AAA	Kanpur_GovndNgr_DC (Uttar Pradesh)	
14785	test	2018-10-03 23:59:14.390954	thanos::sroute:c5f2ba2c-8486-4940-8af6-d1d2a6a...	Carting	153861115439069069	trip-IND627005AAA	Tirunelveli_VdkkuSrt_I (Tamil Nadu)	
14786	test	2018-10-03 23:59:42.701692	thanos::sroute:412fea14-6d1f-4222-8a5f-a517042...	FTL	153861118270144424	trip-IND583119AAA	Sandur_WrdN1DPP_D (Karnataka)	

14787 rows × 19 columns

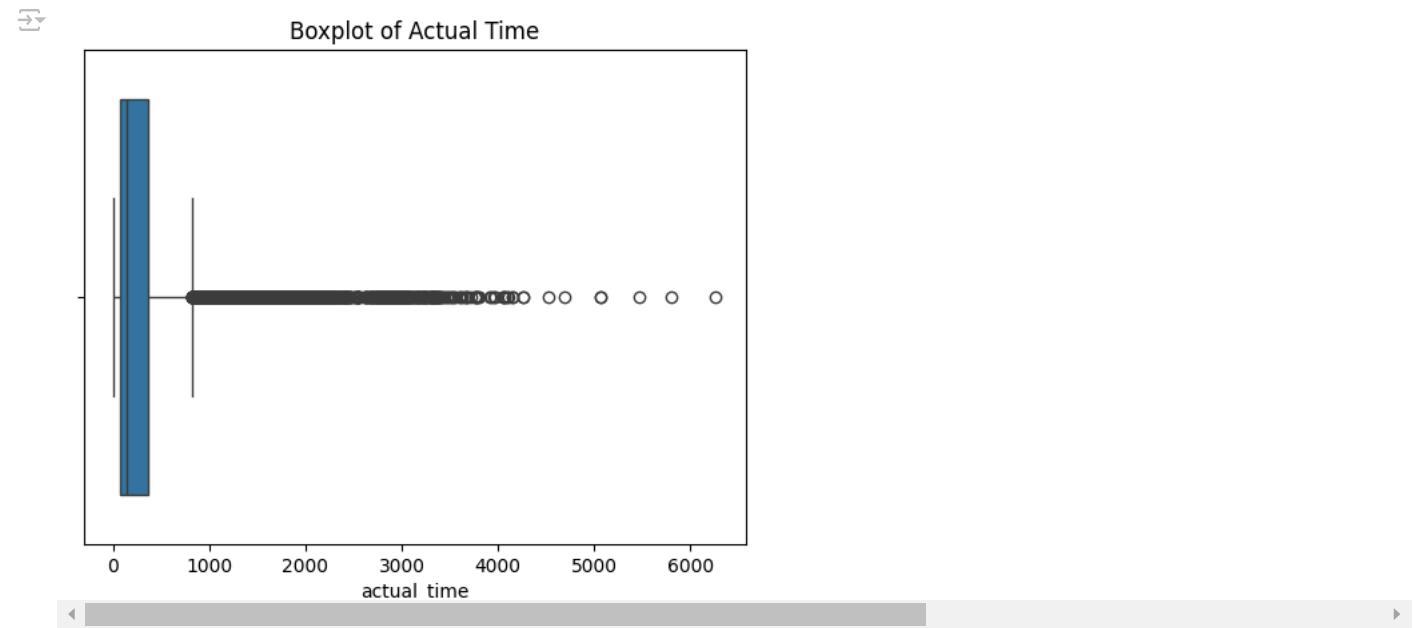
Next steps:

Generate code with trip

View recommended plots

New interactive sheet

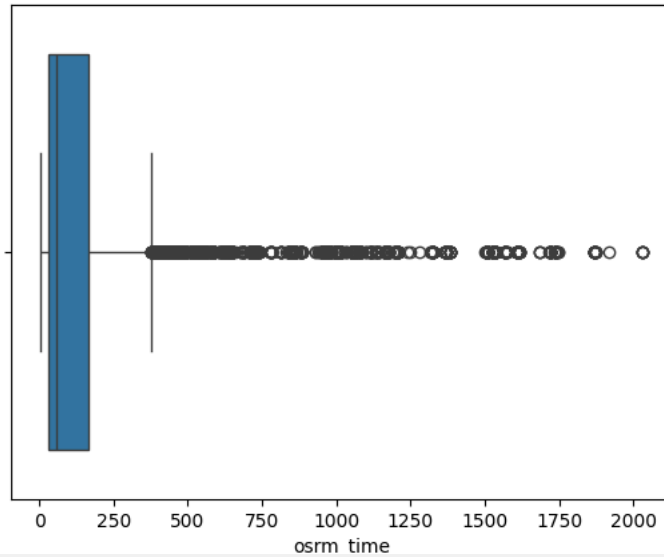
```
# Visualize actual_time outliers
sns.boxplot(x=trip['actual_time'])
plt.title("Boxplot of Actual Time")
plt.show()
```



```
# Visualize osrm_time outliers
sns.boxplot(x=trip['osrm_time'])
plt.title("Boxplot of OSRM Time")
plt.show()
```



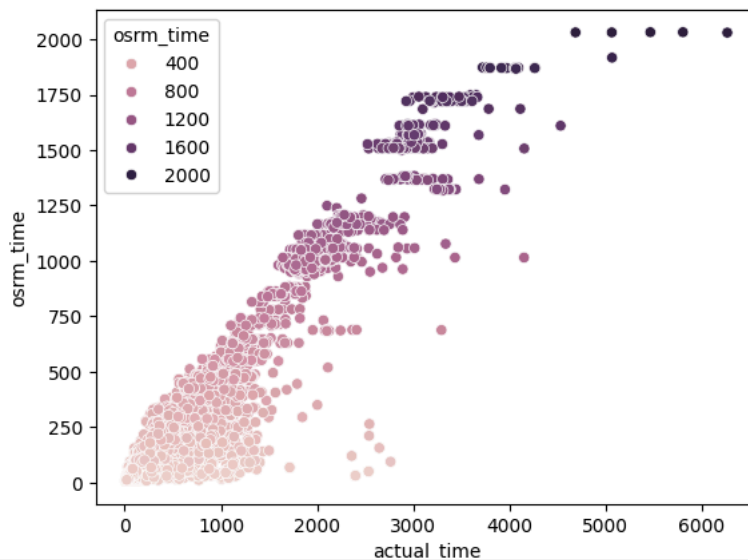
Boxplot of OSRM Time



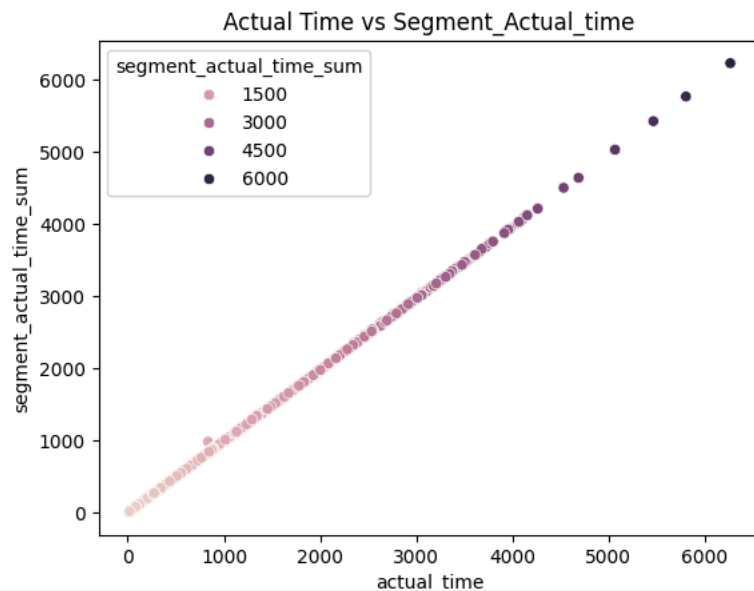
```
# Visualize actual_time vs osrm_time
sns.scatterplot(x='actual_time', y='osrm_time', data=trip, hue='osrm_time')
plt.title("Actual Time vs OSRM Time")
plt.show()
```



Actual Time vs OSRM Time



```
# Visualize actual_time vs Segment_Actual_time
sns.scatterplot(x='actual_time', y='segment_actual_time_sum', data=trip, hue='segment_actual_time_sum')
plt.title("Actual Time vs Segment_Actual_time")
plt.show()
```



✓ Hypothesis Testing

actual_time aggregated value and OSRM time aggregated value.

```
from scipy.stats import ttest_ind, ttest_rel
sampled_data = trip.sample(n=1000, random_state=42)
ttest,pval = ttest_rel(sampled_data['actual_time'],sampled_data['osrm_time'])
print("p-value",pval)
print("ttest values is ",ttest)
if pval<0.05:
    print("Reject null hypothesis")
    print("There is a very significant difference between actual time and osrm time")
else:
    print("Accept null hypothesis")
    print("There is a no difference between actual time and osrm time")
```

p-value 1.3383994234508476e-77
ttest values is 20.39932877587499
Reject null hypothesis
There is a very significant difference between actual time and osrm time

actual_time aggregated value and segment actual time aggregated value.

```
sampled_test_data = trip.sample(n=1000, random_state=42)
ttest,pval = ttest_rel(sampled_test_data['actual_time'],sampled_test_data['segment_actual_time_sum'])
print("p-value",pval)
print("ttest values is ",ttest)
if pval<0.05:
    print("Reject null hypothesis")
    print("There is a very significant difference in between actual time and segement actual time")
else:
    print("Accept null hypothesis")
    print("There is no difference in between actual time and segement actual time")
```

p-value 8.70001488415075e-64
ttest values is 18.139996665565782
Reject null hypothesis
There is a very significant difference in between actual time and segement actual time

```
trip[['actual_time','osrm_time']]
```

	actual_time	osrm_time
0	830.0	388.0
1	143.0	68.0
2	3347.0	1740.0
3	59.0	15.0
4	341.0	117.0
...
14782	83.0	62.0
14783	21.0	12.0
14784	282.0	48.0
14785	264.0	179.0
14786	275.0	68.0

14787 rows x 2 columns

OSRM Distance Vs Segment OSRM Distance Hypothesis Testing

```
sampled_osrm_data = trip.sample(n=1000, random_state=42)
ttest,pval = ttest_rel(sampled_osrm_data['osrm_distance'],sampled_osrm_data['segment_osrm_distance_sum'])
print("p-value",pval)
print("ttest values is ",ttest)
if pval<0.05:
    print("Reject null hypothesis")
    print("There is a very significant difference in between osrm_distance and segment_osrm_distance")
else:
    print("Accept null hypothesis")
    print("There is no difference in between osrm_distance and segment actual segment_osrm_distance")
```

```
p-value 5.847131042782954e-21
ttest values is -9.60644881028483
Reject null hypothesis
There is a very significant difference in between osrm_distance and segment_osrm_distance
```

```
trip[['osrm_distance', 'segment_osrm_distance_sum']]
```

	osrm_distance	segment_osrm_distance_sum
0	544.8027	649.8528
1	85.1110	84.1894
2	2354.0665	2545.2678
3	19.6800	19.8766
4	146.7918	146.7919
...
14782	73.4630	64.8551
14783	16.0882	16.0883
14784	58.9037	104.8866
14785	171.1103	223.5324
14786	80.5787	80.5787

14787 rows x 2 columns

OSRM Time Vs Segment OSRM Time Hypothesis Testing

```
sampled_time_data = trip.sample(n=1000, random_state=42)
ttest,pval = ttest_rel(sampled_time_data['osrm_time'],sampled_time_data['segment_osrm_time_sum'])
print("p-value",pval)
print("ttest values is ",ttest)
if pval<0.05:
    print("Reject null hypothesis")
    print("There is a very significant difference in between osrm_time and segment_osrm_time")
else:
    print("Accept null hypothesis")
    print("There is no difference in between osrm_time and segment actual segment_osrm_time")
```




p-value 4.117974334732554e-28

ttest values is -11.337241440786084

Reject null hypothesis

There is a very significant difference in between norm time and segment norm time

Could not connect to the reCAPTCHA service. Please check your internet connection and reload to get a reCAPTCHA challenge.