Python-Powered Feature Engineering for Delhivery's Operational Optimization

Objective:

- Undertook feature engineering initiative to optimize data utilization for enhanced operational insights and forecasting accuracy.
- Leveraged Delhivery's vast dataset to derive actionable features, aiding in the development of advanced analytics models.

Approach:

- Conducted comprehensive column profiling and data exploration to identify patterns and anomalies.
- Employed advanced techniques including normalization, outlier handling, and categorical encoding for data refinement.
- Generated new features from diverse fields such as destination and source names, trip creation timestamps, and delivery durations.

This summary encapsulates the project's focus on leveraging data to drive business intelligence, showcasing proficiency in feature engineering and data analysis.

```
import pandas as pd, numpy as np
import matplotlib.pyplot as plt, seaborn as sns
from datetime import date
```

```
path = "/content/drive/MyDrive/Dataset/delhivery_data.csv"

data = pd.read_csv(path)
```

```
data.head()
```

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_cente
0	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AA/
1	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AA/
2	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AA/
3	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AA/
4	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AA/

5 rows × 24 columns

```
shape = data.shape
shape
```

```
(144867, 24)
```

```
print(f"No. of Rows: {shape[1]}, \nNo. of Columns: {shape[0]}")
```

```
No. of Rows: 24,
No. of Columns: 144867
```

```
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.describe(include='all')
```

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_
count	144867	144867	144867	144867	144867	
unique	2	14817	1504	2	14817	
top	training	2018-09-28 05:23:15.359220	thanos::sroute:4029a8a2-6c74-4b7e-a6d8-f9e069f...	FTL	trip-153811219535896559	IND000(
freq	104858	101	1812	99660	101	
mean	NaN	NaN	NaN	NaN	NaN	
std	NaN	NaN	NaN	NaN	NaN	
min	NaN	NaN	NaN	NaN	NaN	
25%	NaN	NaN	NaN	NaN	NaN	
50%	NaN	NaN	NaN	NaN	NaN	
75%	NaN	NaN	NaN	NaN	NaN	
max	NaN	NaN	NaN	NaN	NaN	

11 rows × 24 columns

Missing values Treatment

```
data.isnull().any()
```

```
data                            False
trip_creation_time              False
route_schedule_uuid             False
route_type                      False
trip_uuid                       False
source_center                   False
source_name                      True
destination_center              False
destination_name                 True
od_start_time                   False
od_end_time                     False
start_scan_to_end_scan          False
is_cutoff                       False
cutoff_factor                   False
cutoff_timestamp                False
actual_distance_to_destination  False
actual_time                     False
osrm_time                       False
osrm_distance                   False
factor                          False
segment_actual_time             False
segment_osrm_time               False
segment_osrm_distance           False
segment_factor                  False
dtype: bool
```

```
missing_values = data.isnull().sum()
missing_values
```

```
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
```

```
mis_value_percent = missing_values / len(data) * 100
mis_value_percent
```

```
data                            0.000000
trip_creation_time              0.000000
route_schedule_uuid             0.000000
route_type                      0.000000
trip_uuid                       0.000000
source_center                   0.000000
source_name                     0.202254
destination_center              0.000000
destination_name                0.180165
od_start_time                   0.000000
od_end_time                     0.000000
start_scan_to_end_scan          0.000000
is_cutoff                       0.000000
cutoff_factor                   0.000000
cutoff_timestamp                0.000000
actual_distance_to_destination  0.000000
actual_time                     0.000000
osrm_time                       0.000000
osrm_distance                   0.000000
factor                          0.000000
```

```
segment_actual_time              0.000000
segment_osrm_time                0.000000
segment_osrm_distance            0.000000
segment_factor                   0.000000
dtype: float64
```

```
mis_value_table = pd.concat([missing_values, mis_value_percent], axis = 1)
mis_value_table_new = mis_value_table.rename( columns = {0: "Missing Values", 1: "% of Total"})
```

```
mis_value_table_new
```

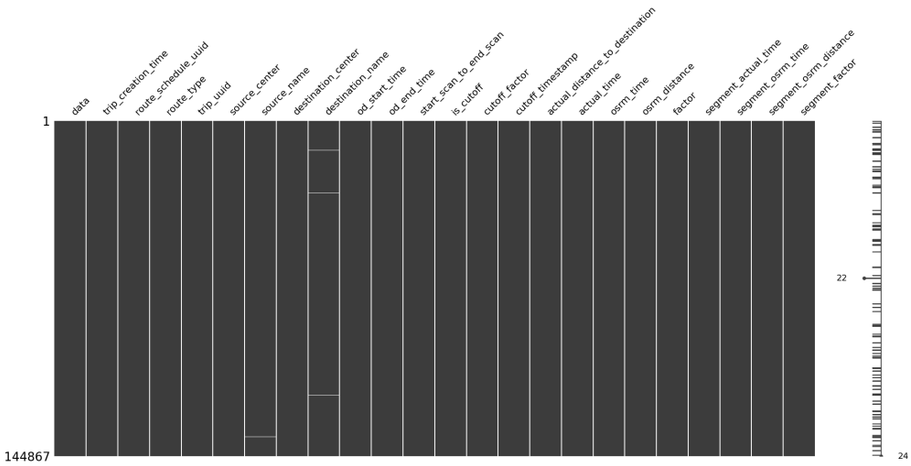
	Missing Values	% of Total
data	0	0.000000
trip_creation_time	0	0.000000
route_schedule_uuid	0	0.000000
route_type	0	0.000000
trip_uuid	0	0.000000
source_center	0	0.000000
source_name	293	0.202254
destination_center	0	0.000000
destination_name	261	0.180165
od_start_time	0	0.000000
od_end_time	0	0.000000
start_scan_to_end_scan	0	0.000000
is_cutoff	0	0.000000
cutoff_factor	0	0.000000
cutoff_timestamp	0	0.000000
actual_distance_to_destination	0	0.000000
actual_time	0	0.000000
osrm_time	0	0.000000
osrm_distance	0	0.000000
factor	0	0.000000
segment_actual_time	0	0.000000
segment_osrm_time	0	0.000000
segment_osrm_distance	0	0.000000
segment_factor	0	0.000000

```
mis_value_table_new = mis_value_table_new[mis_value_table_new.iloc[:,1] != 0].sort_values('% of Total', ascending=False).round(1)
mis_value_table_new
```

	Missing Values	% of Total
source_name	293	0.2
destination_name	261	0.2

```
# Detecting Missing Values Visually
import missingno as msno

msno.matrix(data)

# The source_name and destination_name Columns has very few missing values and donot seem to be correlated with any other column.
# Hence, the missingness in these columns can be attributed as Missing Completely at Random.
```

<Axes: >




```
# Finding reason for missing data using a Heatmap
msno.heatmap(data)

# The heatmap function shows that there are no strong correlations between missing values of different features.
# This is good; low correlations further indicate that the data are Missing At Random.
```

<Axes: >


```
# Finding reason for missing data using Dendrogram
msno.dendrogram(data)

# The correlation is high since Embarked column has a very few missing values.
# This dataset doesn't have much missing values
```

<Axes: >


```
missing_source_name = data.loc[data['source_name'].isnull(), 'source_center'].unique()
missing_source_name
```

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

```
for i in missing_source_name:
 unique_source_name = data.loc[data['source_center'] == i, 'source_name'].unique()
 if pd.isna(unique_source_name):
  print("Source Center :", i, "-" * 10, "Source Name :", 'Not Found')
 else :
  print("Source Center :", i, "-" * 10, "Source Name :", unique_source_name)
```

```
Source Center : IND342902A1B ---------- Source Name : Not Found
Source Center : IND577116AAA ---------- Source Name : Not Found
Source Center : IND282002AAD ---------- Source Name : Not Found
Source Center : IND465333A1B ---------- Source Name : Not Found
Source Center : IND841301AAC ---------- Source Name : Not Found
Source Center : IND509103AAC ---------- Source Name : Not Found
Source Center : IND126116AAA ---------- Source Name : Not Found
Source Center : IND331022A1B ---------- Source Name : Not Found
Source Center : IND505326AAB ---------- Source Name : Not Found
Source Center : IND852118A1B ---------- Source Name : Not Found
```

```
missing_destination_name = data.loc[data['destination_name'].isnull(), 'destination_center'].unique()
missing_destination_name
```

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

```
for i in missing_destination_name:
 unique_dest_name = data.loc[data['destination_center'] == i, 'destination_name'].unique()
 if pd.isna(unique_dest_name):
  print("Destination Center :", i, "-" * 10, "Destination Name :", 'Not Found')
 else :
  print("Destination Center :", i, "-" * 10, "Destination Name :", unique_dest_name)
```

```
Destination Center : IND342902A1B ---------- Destination Name : Not Found
Destination Center : IND577116AAA ---------- Destination Name : Not Found
Destination Center : IND282002AAD ---------- Destination Name : Not Found
Destination Center : IND465333A1B ---------- Destination Name : Not Found
Destination Center : IND841301AAC ---------- Destination Name : Not Found
Destination Center : IND505326AAB ---------- Destination Name : Not Found
Destination Center : IND852118A1B ---------- Destination Name : Not Found
Destination Center : IND126116AAA ---------- Destination Name : Not Found
Destination Center : IND509103AAC ---------- Destination Name : Not Found
Destination Center : IND221005A1A ---------- Destination Name : Not Found
Destination Center : IND250002AAC ---------- Destination Name : Not Found
Destination Center : IND331001A1C ---------- Destination Name : Not Found
Destination Center : IND122015AAC ---------- Destination Name : Not Found
```

```
#  Checking if the Missing Source Name also has all the destination missing or not.

np.all(data.loc[data['source_name'].isnull(), 'source_center'].isin(missing_destination_name))
```

```
False
```

```
# Treating missing values
data1 = data.copy()
```

```
data1.head()
```

data1.head()

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_cente
0	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AA/
1	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AA/
2	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AA/
3	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AA/
4	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AA/

5 rows × 24 columns

```
data1.isnull().sum()
```

```
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
```

```
data1.dropna(subset = ["source_name"], how= 'any', inplace = True)
```

```
data1['source_name'].isnull().sum()
```

```
0
```

```
data1.dropna(subset= ['destination_name'], how= "any", inplace = True)
```

```
data1['destination_name'].isnull().sum()
```

```
0
```

```
data1.isnull().sum()

# Here missing values are treated and now no missing values in the dataset
```

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

Converting time columns into pandas datetime

```
data1['od_start_time'] = pd.to_datetime(data1['od_start_time'])
```

```
data1['od_end_time'] = pd.to_datetime(data1['od_end_time'])
```

```
data1.head()
```

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_cente
0	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AA/
1	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AA/
2	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AA/
3	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AA/
4	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121AA/

5 rows × 24 columns

```
data1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 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  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
dtypes: bool(1), datetime64[ns](2), float64(10), int64(1), object(10)
memory usage: 26.6+ MB
```

Grouping by sub-journey in the trip

```
data1['segment_key'] = data1['trip_uuid'] + data1['source_center'] + data1['destination_center']

segment_columns = ['segment_actual_time' , 'segment_osrm_distance', 'segment_osrm_time']

for column in segment_columns:
  data1[column + 'sum'] = data1.groupby('segment_key')[column].cumsum()

data1[[column + 'sum' for column in segment_columns]]
```

	segment_actual_timesum	segment_osrm_distancesum	segment_osrm_timesum
0	14.0	11.9653	11.0
1	24.0	21.7243	20.0
2	40.0	32.5395	27.0
3	61.0	45.5619	39.0
4	67.0	49.4772	44.0
...	...	...	...
144862	92.0	65.3487	94.0
144863	118.0	82.7212	115.0
144864	138.0	103.4265	149.0
144865	155.0	122.3150	176.0
144866	423.0	131.1238	185.0

144316 rows × 3 columns

Aggregating at sub-jouney lebel

```
create_segment_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',

    '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_timesum' : 'last',
    'segment_osrm_distancesum' : 'last',
    'segment_osrm_timesum' : 'last',

}
```

Grouping mini-trips, sorting by time

```
segment = data1.groupby('segment_key').agg(create_segment_dict).reset_index()
segment = segment.sort_values(by=['segment_key', 'od_end_time'], ascending = True).reset_index()
```

```
segment.head()
```

	index	segment_key	data	trip_creation_time	route_schedu
0	0	trip-153671041653548748IND209304AAAIND000000ACB	training	2018-09-12 00:00:16.535741	thanos::sroute:d a29b-4 2
1	1	trip-153671041653548748IND462022AAAIND209304AAA	training	2018-09-12 00:00:16.535741	thanos::sroute:d a29b-4 2
2	2	trip-153671042288605164IND561203AABIND562101AAA	training	2018-09-12 00:00:22.886430	thanos::sroute:3 bb0b-4 e
3	3	trip-153671042288605164IND572101AAAIND561203AAB	training	2018-09-12 00:00:22.886430	thanos::sroute:3 bb0b-4 e
4	4	trip-153671043369099517IND000000ACBIND160002AAC	training	2018-09-12 00:00:33.691250	thanos::sroute:d 7641-45

5 rows × 21 columns

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

	index	segment_key	data	trip_creation_time	route_sc
10370	10370	trip-153741093647649320IND388121AAAIND388620AAB	training	2018-09-20 02:35:36.476840	thanos::sro b35
10371	10371	trip-153741093647649320IND388620AABIND388320AAA	training	2018-09-20 02:35:36.476840	thanos::sro b35

2 rows × 21 columns

```
segment.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26222 entries, 0 to 26221
Data columns (total 21 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   index                           26222 non-null  int64
 1   segment_key                     26222 non-null  object
 2   data                            26222 non-null  object
 3   trip_creation_time              26222 non-null  object
 4   route_schedule_uuid             26222 non-null  object
 5   route_type                      26222 non-null  object
 6   trip_uuid                       26222 non-null  object
 7   source_center                   26222 non-null  object
 8   source_name                     26222 non-null  object
 9   destination_center              26222 non-null  object
 10  destination_name                26222 non-null  object
 11  od_start_time                   26222 non-null  datetime64[ns]
 12  od_end_time                     26222 non-null  datetime64[ns]
 13  start_scan_to_end_scan          26222 non-null  float64
 14  actual_distance_to_destination  26222 non-null  float64
 15  actual_time                     26222 non-null  float64
 16  osrm_time                       26222 non-null  float64
 17  osrm_distance                   26222 non-null  float64
 18  segment_actual_timesum          26222 non-null  float64
 19  segment_osrm_distancesum        26222 non-null  float64
 20  segment_osrm_timesum            26222 non-null  float64
dtypes: datetime64[ns](2), float64(8), int64(1), object(10)
memory usage: 4.2+ MB
```

Calculate time taken between od_start_time and od_end_time and keep it as a feature

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

```
segment['od_time_diff_hour']
```

```
0        1260.604421
1         999.505379
2          58.832388
```

```
3          122.779486
4          834.638929
             ...
26217       62.115193
26218       91.087797
26219       44.174403
26220      287.474007
26221       66.933565
Name: od_time_diff_hour, Length: 26222, dtype: float64
```

```
segment.head()
```

	index	segment_key	data	trip_creation_time	route_schedu
0	0	trip-153671041653548748IND209304AAAIND000000ACB	training	2018-09-12 00:00:16.535741	thanos::sroute:d a29b-4 2
1	1	trip-153671041653548748IND462022AAAIND209304AAA	training	2018-09-12 00:00:16.535741	thanos::sroute:d a29b-4 2
2	2	trip-153671042288605164IND561203AABIND562101AAA	training	2018-09-12 00:00:22.886430	thanos::sroute:3 bb0b-4 e
3	3	trip-153671042288605164IND572101AAAIND561203AAB	training	2018-09-12 00:00:22.886430	thanos::sroute:3 bb0b-4 e
4	4	trip-153671043369099517IND000000ACBIND160002AAC	training	2018-09-12 00:00:33.691250	thanos::sroute:d 7641-45 4

5 rows × 22 columns

```
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_time_diff_hour' : 'sum',

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

    'segment_actual_timesum' : 'sum',
    'segment_osrm_distancesum' : 'sum',
    'segment_osrm_timesum' : 'sum',

}
```

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

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_cente
0	training	2018-09-12 00:00:16.535741	thanos::sroute:d7c989ba-a29b-4a0b-b2f4-288cdc6...	FTL	trip-153671041653548748	IND209304AA
1	training	2018-09-12 00:00:22.886430	thanos::sroute:3a1b0ab2-bb0b-4c53-8c59-eb2a2c0...	Carting	trip-153671042288605164	IND561203AA
2	training	2018-09-12 00:00:33.691250	thanos::sroute:de5e208e-7641-45e6-8100-4d9fb1e...	FTL	trip-153671043369099517	IND000000AC
3	training	2018-09-12 00:01:00.113710	thanos::sroute:f0176492-a679-4597-8332-bbd1c7f...	Carting	trip-153671046011330457	IND400072AA
4	training	2018-09-12 00:02:09.740725	thanos::sroute:d9f07b12-65e0-4f3b-bec8-df06134...	FTL	trip-153671052974046625	IND583101AA

```
trip[['actual_time', 'segment_actual_timesum']]
```

	actual_time	segment_actual_timesum
0	1562.0	1548.0
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 × 2 columns

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

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_ce
5917	training	2018-09-20 02:35:36.476840	thanos::sroute:eb7bfc78-b351-4c0e-a951-fa3d5c3...	Carting	trip-153741093647649320	IND388121

```
trip[['actual_distance_to_destination', 'osrm_distance']]
```

	actual_distance_to_destination	osrm_distance
0	824.732854	991.3523
1	73.186911	85.1110
2	1927.404273	2354.0665
3	17.175274	19.6800
4	127.448500	146.7918
...	...	...
14782	57.762332	73.4630
14783	15.513784	16.0882
14784	38.684839	58.9037
14785	134.723836	171.1103
14786	66.081533	80.5787

14787 rows × 2 columns

Hypothesis Testing

```
trip['destination_name'] = trip['destination_name'].str.lower()
trip['source_name'] = trip['source_name']


def place2state(x):
  state = x.split('(')[1]
  return state[:-1]



def place2city(x):
  city = x.split('(')[0]

  city = city.split('_')[0]

  if city == 'pnq vadagaon sheri dpc':
    return 'vadagaonsheri'

  if city in ['pnq pashan dpc', 'pnq rahatani dpc', 'pune balaji nagar']:
    return 'pune'

  if city == 'hbr layout pc':
    return 'bengaluru'

  if city == 'bhopal mp nagar':
    return 'bhopal'

  if city == 'mumbai antop hill':
    return 'mumbai'

  return city
```

```
def place2city_place(x):
  x = x.split('(')[0]
  len_ = len(x.split('_'))

  if len_ >= 3:
    return x.split('_')[1]

  if len_ == 2:
    return x.split('_')[0]

  return x.split(' ')[0]




def place2code(x):
  x = x.split('(')[0]

  if len(x.split('_')) >= 3:
    return x.split('_')[-1]

  return 'none'


trip['destination_state'] = trip['destination_name'].apply(lambda x: place2state(x))
trip['destination_city'] = trip['destination_name'].apply(lambda x: place2city(x))
trip['destination_place'] = trip['destination_name'].apply(lambda x: place2city_place(x))
trip['destination_code'] = trip['destination_name'].apply(lambda x: place2code(x))


trip[['destination_state', 'destination_city', 'destination_place', 'destination_code']]
```

	destination_state	destination_city	destination_place	destination_code
0	uttar pradesh	kanpur	central	6
1	karnataka	doddablpur	chikadpp	d
2	haryana	gurgaon	bilaspur	hb
3	maharashtra	mumbai	mirard	ip
4	karnataka	sandur	wrdn1dpp	d
...	...	...	...	...
14782	punjab	chandigarh	mehmdpur	h
14783	haryana	faridabad	blbgarh	dc
14784	uttar pradesh	kanpur	govndngr	dc
14785	tamil nadu	tirchchndr	shnmgprm	d
14786	karnataka	sandur	wrdn1dpp	d

14787 rows × 4 columns

```
trip['source_state'] = trip['source_name'].apply(lambda x: place2state(x))
trip['source_city'] = trip['source_name'].apply(lambda x: place2city(x))
trip['source_place'] = trip['source_name'].apply(lambda x: place2city_place(x))
trip['source_code'] = trip['source_name'].apply(lambda x: place2code(x))


trip[['source_state','source_city', 'source_place', 'source_code']]
```

	source_state	source_city	source_place	source_code
0	Uttar Pradesh	Kanpur	Central	6
1	Karnataka	Doddablpur	ChikaDPP	D
2	Haryana	Gurgaon	Bilaspur	HB
3	Maharashtra	Mumbai Hub	Mumbai	none
4	Karnataka	Bellary	Bellary	none
...	...	...	...	...
14782	Punjab	Chandigarh	Mehmdpur	H
14783	Haryana	FBD	Balabhgarh	DPC
14784	Uttar Pradesh	Kanpur	GovndNgr	DC
14785	Tamil Nadu	Tirunelveli	VdkkuSrt	I
14786	Karnataka	Sandur	WrdN1DPP	D

14787 rows × 4 columns

```
trip['trip_creation_time'] = pd.to_datetime(trip['trip_creation_time'])

trip['trip_year'] = trip['trip_creation_time'].dt.year
trip['trip_month'] = trip['trip_creation_time'].dt.month
trip['trip_hour'] = trip['trip_creation_time'].dt.hour
trip['trip_day'] = trip['trip_creation_time'].dt.day
trip['trip_week'] = trip['trip_creation_time'].dt.isocalendar().week
trip['trip_dayofweek'] = trip['trip_creation_time'].dt.dayofweek


trip[['trip_year', 'trip_month', 'trip_hour', 'trip_day', 'trip_week', 'trip_dayofweek']]
```

	trip_year	trip_month	trip_hour	trip_day	trip_week	trip_dayofweek
0	2018	9	0	12	37	2
1	2018	9	0	12	37	2
2	2018	9	0	12	37	2
3	2018	9	0	12	37	2
4	2018	9	0	12	37	2
...	...	...	...	...	...	...
14782	2018	10	23	3	40	2
14783	2018	10	23	3	40	2
14784	2018	10	23	3	40	2
14785	2018	10	23	3	40	2
14786	2018	10	23	3	40	2

14787 rows × 6 columns

```
trip.head()
```

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_cente
0	training	2018-09-12 00:00:16.535741	thanos::sroute:d7c989ba-a29b-4a0b-b2f4-288cdc6...	FTL	trip-153671041653548748	IND209304AA
1	training	2018-09-12 00:00:22.886430	thanos::sroute:3a1b0ab2-bb0b-4c53-8c59-eb2a2c0...	Carting	trip-153671042288605164	IND561203AA
2	training	2018-09-12 00:00:33.691250	thanos::sroute:de5e208e-7641-45e6-8100-4d9fb1e...	FTL	trip-153671043369099517	IND000000AC
3	training	2018-09-12 00:01:00.113710	thanos::sroute:f0176492-a679-4597-8332-bbd1c7f...	Carting	trip-153671046011330457	IND400072AA
4	training	2018-09-12 00:02:09.740725	thanos::sroute:d9f07b12-65e0-4f3b-bec8-df06134...	FTL	trip-153671052974046625	IND583101AA

5 rows × 32 columns

```
trip['destination_state'].value_counts(ascending=False)
```

```
maharashtra               2561
karnataka                 2294
haryana                   1640
tamil nadu                1084
uttar pradesh              805
telangana                  784
gujarat                    734
west bengal                697
delhi                      657
punjab                     617
rajasthan                  550
andhra pradesh             442
bihar                      367
madhya pradesh             350
kerala                     270
assam                      232
jharkhand                  181
uttarakhand                122
orissa                     119
chandigarh                  65
goa                         52
chhattisgarh                43
himachal pradesh            42
arunachal pradesh           25
jammu & kashmir             20
dadra and nagar haveli      17
meghalaya                    8
mizoram                      6
nagaland                     1
tripura                      1
daman & diu                  1
Name: destination_state, dtype: int64
```

```
trip['destination_city'].value_counts(ascending= False)
```

```
bengaluru     1088
mumbai         966
gurgaon        877
delhi          554
bangalore      551
              ...
chapra           1
shamshabad       1
kullu            1
oriyur           1
lunawada         1
Name: destination_city, Length: 855, dtype: int64
```

```
trip['source_state'].value_counts()
```

```
Maharashtra               2714
Karnataka                 2143
Haryana                   1823
Tamil Nadu                1039
Telangana                  784
Uttar Pradesh              760
Gujarat                    750
Delhi                      725
```

```
West Bengal               665
Punjab                    536
Rajasthan                 514
Andhra Pradesh            435
Bihar                     351
Madhya Pradesh            318
Kerala                    289
Assam                     268
Jharkhand                 160
Uttarakhand               114
Orissa                    107
Chandigarh                 93
Goa                        65
Chhattisgarh               43
Himachal Pradesh           34
Jammu & Kashmir            17
Dadra and Nagar Haveli     15
Pondicherry                12
Nagaland                    5
Arunachal Pradesh           4
Mizoram                     4
Name: source_state, dtype: int64
```

```
trip['source_city'].value_counts()
```

```
Gurgaon        1128
Bengaluru      1052
Bhiwandi        697
Mumbai          654
Bangalore       648
               ...
Dhaka             1
Ukkadagatri       1
Sultana           1
Banka             1
Sandur            1
Name: source_city, Length: 731, dtype: int64
```

```
trip[['source_city', 'destination_city']].value_counts(ascending= False).reset_index()
```

	source_city	destination_city	0
0	Bengaluru	bengaluru	581
1	Bangalore	bengaluru	455
2	Hyderabad	hyderabad	355
3	Mumbai	mumbai	339
4	Bhiwandi	mumbai	332
...	...	...	...
1746	Bokaro	bokaro	1
1747	Mainaguri	coochbehar	1
1748	Boisar	dahanu	1
1749	Mainaguri	mathabhang	1
1750	Junagadh	veraval	1

1751 rows × 3 columns

```
trip['destination_center'].value_counts()
```

```
IND000000ACB    821
IND562132AAA    548
IND421302AAG    403
IND560099AAB    342
IND501359AAE    280
               ...
IND442101AAB      1
IND679576AAA      1
IND412210AAC      1
IND798601AAA      1
IND389230AAB      1
Name: destination_center, Length: 1035, dtype: int64
```

```
trip[trip['destination_center'] == 'IND000000ACB']
```

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_c
2	training	2018-09-12 00:00:33.691250	thanos::sroute:de5e208e-7641-45e6-8100-4d9fb1e...	FTL	trip-153671043369099517	IND0000
32	training	2018-09-12 00:41:16.117700	thanos::sroute:074b1ab2-c84c-42a1-a662-098ded0...	Carting	trip-153671287611744830	IND1210
41	training	2018-09-12 00:46:57.104787	thanos::sroute:951d77aa-4725-4c4e-882d-42acc35...	FTL	trip-153671321710455800	IND4213
44	training	2018-09-12 00:49:56.561006	thanos::sroute:a8c7e9f0-237d-47a1-9024-a48afeb...	FTL	trip-153671339656068455	IND3020
104	training	2018-09-12 02:30:20.747941	thanos::sroute:e3e7c92f-55a9-4ecb-a4a7-919bbb7...	Carting	trip-153671942074761472	IND1310
...	...	...	...	...	...	
14688	test	2018-10-03 22:13:55.442334	thanos::sroute:1d942309-bf28-4b95-9e29-9d03ca1...	Carting	trip-153860483544204900	IND1210
14698	test	2018-10-03 22:26:10.593755	thanos::sroute:df0ea6fc-ce7d-40d0-8ca9-2737694...	Carting	trip-153860557059347623	IND1100
14730	test	2018-10-03 23:04:39.165766	thanos::sroute:61cf7ae5-03bf-4202-9047-8df9f38...	Carting	trip-153860787916550213	IND1100
14733	test	2018-10-03 23:12:21.256232	thanos::sroute:074b1ab2-c84c-42a1-a662-098ded0...	Carting	trip-153860834125598161	IND1210
14739	test	2018-10-03 23:19:54.394130	thanos::sroute:105d3425-e2d7-4f05-ad0c-157dc94...	FTL	trip-153860879439383883	IND0000

821 rows × 32 columns

```
trip['trip_year'].value_counts()
```

```
2018    14787
Name: trip_year, dtype: int64
```

```
trip['trip_month'].value_counts()
```

```
9     13011
10     1776
Name: trip_month, dtype: int64
```

```
trip['trip_week'].value_counts(ascending= False)
```

```
38    5001
39    4402
37    3608
40    1776
Name: trip_week, dtype: Int64
```

```
trip['actual_time'] - trip['osrm_time']
```

```
0         845.0
1          75.0
2        1607.0
3          44.0
4         224.0
          ...
14782      21.0
14783       9.0
14784     234.0
14785      85.0
14786     207.0
Length: 14787, dtype: float64
```

```
trip['trip_month'].value_counts()
```

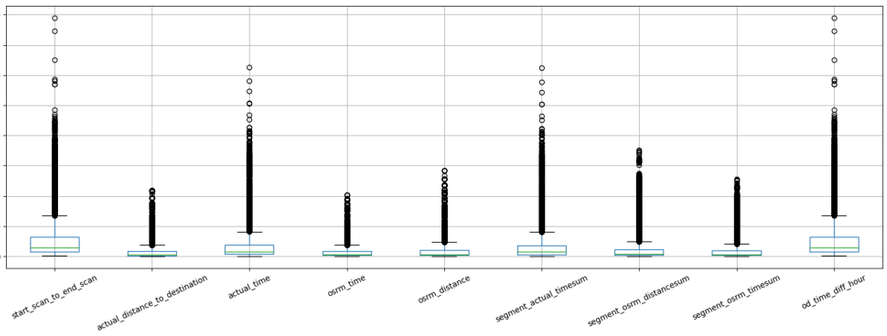
```
9     11172
10     1551
Name: trip_month, dtype: int64
```

Outlier Detection

```
num_columns = ['start_scan_to_end_scan','actual_distance_to_destination','actual_time','osrm_time', 'osrm_distance','segment_actual_timesum','segment_osrm_
```

```
trip[num_columns].boxplot(rot= 25,figsize= (20,6))
```

```
<Axes: >
```




```
# Quantile Method to handle outliers

q1 = trip[num_columns].quantile(0.25)
q3 = trip[num_columns].quantile(0.75)

IQR = q3 - q1


IQR
```

```
start_scan_to_end_scan            483.000000
actual_distance_to_destination    140.814159
actual_time                       300.000000
osrm_time                         139.000000
osrm_distance                     175.887300
segment_actual_timesum            298.000000
segment_osrm_distancesum          183.981750
segment_osrm_timesum              154.000000
od_time_diff_hour                 483.839201
dtype: float64
```

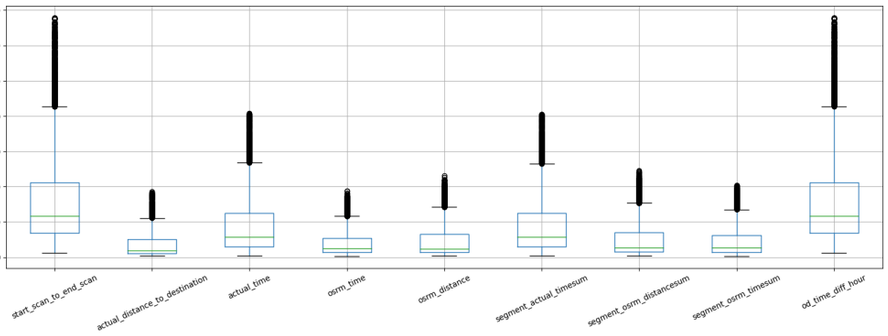
```
trip = trip[-((trip[num_columns] < (q1 - 1.5 *IQR)) | (trip[num_columns] > (q3 + 1.5 * IQR))).any(axis = 1)]

trip = trip.reset_index(drop= True)

trip.head(10)
```

	data	trip_creation_time	route_schedule_uuid	route_type	trip_uuid	source_cente
0	training	2018-09-12 00:00:22.886430	thanos::sroute:3a1b0ab2-bb0b-4c53-8c59-eb2a2c0...	Carting	trip-153671042288605164	IND561203AA
1	training	2018-09-12 00:01:00.113710	thanos::sroute:f0176492-a679-4597-8332-bbd1c7f...	Carting	trip-153671046011330457	IND400072AA
2	training	2018-09-12 00:02:09.740725	thanos::sroute:d9f07b12-65e0-4f3b-bec8-df06134...	FTL	trip-153671052974046625	IND583101AA
3	training	2018-09-12 00:02:34.161600	thanos::sroute:9bf03170-d0a2-4a3f-aa4d-9aaab3d...	Carting	trip-153671055416136166	IND600056AA
4	training	2018-09-12 00:04:22.011653	thanos::sroute:a97698cc-846e-41a7-916b-88b1741...	Carting	trip-153671066201138152	IND600044AA
5	training	2018-09-12 00:04:28.263977	thanos::sroute:d5b71ae9-a11a-4f52-bcb7-274b65e...	Carting	trip-153671066826362165	IND560043AA
6	training	2018-09-12 00:05:40.333071	thanos::sroute:a0e60427-16ad-4b17-b3b0-6a06643...	Carting	trip-153671074033284934	IND395009AA
7	training	2018-09-12 00:06:39.565253	thanos::sroute:a10888ff-f794-41e1-9b7a-7f62ef6...	Carting	trip-153671079956500691	IND110024AA
8	training	2018-09-12 00:08:29.805514	thanos::sroute:580c788b-ff17-4c1b-9bbd-c59e7b0...	Carting	trip-153671090980523004	IND412105AA
9	training	2018-09-12 00:11:40.783923	thanos::sroute:c2ee580f-f4b2-4fa5-98ab-0c5b327...	Carting	trip-153671110078355292	IND121004AA

10 rows × 32 columns

```
trip[num_columns].boxplot(figsize= (20,6), rot= 25)
```

<Axes: >



Handling categorical variables

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

```
Carting    8812
FTL        3911
Name: route_type, dtype: int64
```

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

Column Normalization /Column Standardization

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

```
scaler = StandardScaler()
scaler.fit(trip[num_columns])
```

▾ StandardScaler
StandardScaler()

```
trip[num_columns] = scaler.transform(trip[num_columns])
```

```
trip[num_columns]
```

	start_scan_to_end_scan	actual_distance_to_destination	actual_time	osrm_time	osrm_dista
0	-0.548546	0.012060	-0.217856	-0.144341	-0.073
1	-0.861602	-0.765152	-0.749015	-0.877085	-0.804
2	1.552838	0.764988	1.034163	0.533102	0.614
3	-0.513328	-0.662169	-0.736369	-0.766482	-0.710
4	-0.869428	-0.877197	-0.970332	-0.904736	-0.890
...	...	...	...	...	
12718	-0.247231	-0.201970	-0.597255	-0.227293	-0.204
12719	-1.018130	-0.788207	-0.989302	-0.918561	-0.844
12720	0.394533	-0.466688	0.661086	-0.420848	-0.366
12721	0.104957	0.865940	0.547267	1.390274	0.886
12722	0.128436	-0.086534	0.616823	-0.144341	-0.124

12723 rows × 9 columns

```
trip[num_columns].describe(include ='all')
```

	start_scan_to_end_scan	actual_distance_to_destination	actual_time	osrm_time	osrm_d
count	1.272300e+04	1.272300e+04	1.272300e+04	1.272300e+04	1.272
mean	-1.619566e-17	-7.371818e-17	-8.041983e-17	4.467769e-17	3.797
std	1.000039e+00	1.000039e+00	1.000039e+00	1.000039e+00	1.000
min	-1.162918e+00	-8.785574e-01	-1.065181e+00	-1.001514e+00	-9.229
25%	-7.207269e-01	-7.065920e-01	-7.363685e-01	-7.111809e-01	-7.077
50%	-3.411472e-01	-4.689012e-01	-4.012322e-01	-3.931975e-01	-4.836
75%	4.023595e-01	4.073375e-01	4.650634e-01	4.224989e-01	4.419
max	4.049455e+00	4.178358e+00	4.031419e+00	4.113871e+00	4.150

Insights:

1. This analysis shows that their is a significant difference between the OSRM and actual data, not only for the time difference but the traffic-wise difference as well.
2. The data shows that the most number of orders are coming from Maharashtra, but the overall city-wise data shows that Bengaluru, Karnataka has most orders from a single city, followed by Mumbai, Maharashtra. Also, IND000000ACB is the busiest center. Bengaluru and Mumbai are the busiest corridors.
3. The data has 8812 Carting type orders where the FTL type orders are 3911. So, the Carting trips are more popular choice.
4. The data shows that the traffic is from 2018 in the months of September and October, especially for week 38 and 39 being the busiest.

Recommendation:

1. As this is the data for just 2 months, we can not conclude this but the we have a very strong presence in the Maharashtra, Karnataka, Haryana and Tamilnadu Sectors. We can still make progress in rest of the sectors.
2. The state-wise analysis shows, we have major order traffic share from Maharashtra and neighboring states. So, there should be opportunity to make a strong presence in the eastern corridor of Maharashtra.
3. The data shows that the Indian Festive season coming in the months of September and October are very productive.

Start coding or generate with AI.

Double-click (or enter) to edit

Double-click (or enter) to edit

+ Code + Text

Double-click (or enter) to edit

Start coding or generate with AI.

Start coding or generate with AI.