Assignment-Uber

January 23, 2025

1 Assignment on Uber dataset

Load the dataset

[4]:
```
import pandas as pd
ud=pd.read_csv('Uber.csv')
```

Display basic info about dataset

[5]:
```
ud.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1156 entries, 0 to 1155
Data columns (total 7 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   START_DATE*  1156 non-null   object 
 1   END_DATE*    1155 non-null   object 
 2   CATEGORY*    1155 non-null   object 
 3   START*       1155 non-null   object 
 4   STOP*        1155 non-null   object 
 5   MILES*       1156 non-null   float64
 6   PURPOSE*     653 non-null    object 
dtypes: float64(1), object(6)
memory usage: 63.3+ KB
```

Check for missing values

[6]:
```
mising_values=ud.isnull()
mising_values
```

[6]:

	START_DATE*	END_DATE*	CATEGORY*	START*	STOP*	MILES*	PURPOSE*
0	False	False	False	False	False	False	False
1	False	False	False	False	False	False	True
2	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False
…	…	…	…	…	…	…	…
1151	False	False	False	False	False	False	False

```
1152        False      False      False   False  False   False     False
1153        False      False      False   False  False   False     False
1154        False      False      False   False  False   False     False
1155        False       True       True    True   True   False      True

[1156 rows x 7 columns]
```

[10]:
```
ud.isnull().sum()
```

[10]:
```
START_DATE*      0
END_DATE*        1
CATEGORY*        1
START*           1
STOP*            1
MILES*           0
PURPOSE*       503
dtype: int64
```

Drop rows with missing values

[11]:
```
ud.dropna()
```

[11]:
```
           START_DATE*         END_DATE* CATEGORY*            START*  \
0       1/1/2016 21:11    1/1/2016 21:17  Business       Fort Pierce
2       1/2/2016 20:25    1/2/2016 20:38  Business       Fort Pierce
3       1/5/2016 17:31    1/5/2016 17:45  Business       Fort Pierce
4       1/6/2016 14:42    1/6/2016 15:49  Business       Fort Pierce
5       1/6/2016 17:15    1/6/2016 17:19  Business   West Palm Beach
…                  …                 …         …                …
1150   12/31/2016 1:07   12/31/2016 1:14  Business           Kar?chi
1151  12/31/2016 13:24  12/31/2016 13:42  Business           Kar?chi
1152  12/31/2016 15:03  12/31/2016 15:38  Business  Unknown Location
1153  12/31/2016 21:32  12/31/2016 21:50  Business        Katunayake
1154  12/31/2016 22:08  12/31/2016 23:51  Business           Gampaha

                  STOP*  MILES*         PURPOSE*
0          Fort Pierce     5.1   Meal/Entertain
2          Fort Pierce     4.8  Errand/Supplies
3          Fort Pierce     4.7          Meeting
4       West Palm Beach    63.7   Customer Visit
5       West Palm Beach     4.3   Meal/Entertain
…                  …     …              …
1150           Kar?chi     0.7          Meeting
1151  Unknown Location     3.9   Temporary Site
1152  Unknown Location    16.2          Meeting
1153           Gampaha     6.4   Temporary Site
1154         Ilukwatta    48.2   Temporary Site
```

[653 rows x 7 columns]

fill missing values (propose column with unknown value)

[13]:
```
ud.fillna('Null')
```

[13]:
```
            START_DATE*         END_DATE* CATEGORY*            START*  \
0       1/1/2016 21:11    1/1/2016 21:17  Business       Fort Pierce
1        1/2/2016 1:25     1/2/2016 1:37  Business       Fort Pierce
2       1/2/2016 20:25    1/2/2016 20:38  Business       Fort Pierce
3       1/5/2016 17:31    1/5/2016 17:45  Business       Fort Pierce
4       1/6/2016 14:42    1/6/2016 15:49  Business       Fort Pierce
…                  …                 …       …                 …
1151  12/31/2016 13:24  12/31/2016 13:42  Business           Kar?chi
1152  12/31/2016 15:03  12/31/2016 15:38  Business  Unknown Location
1153  12/31/2016 21:32  12/31/2016 21:50  Business        Katunayake
1154  12/31/2016 22:08  12/31/2016 23:51  Business           Gampaha
1155            Totals              Null      Null              Null

                  STOP*   MILES*         PURPOSE*
0          Fort Pierce      5.1   Meal/Entertain
1          Fort Pierce      5.0             Null
2          Fort Pierce      4.8  Errand/Supplies
3          Fort Pierce      4.7          Meeting
4       West Palm Beach     63.7   Customer Visit
…                  …      …              …
1151  Unknown Location      3.9   Temporary Site
1152  Unknown Location     16.2          Meeting
1153           Gampaha      6.4   Temporary Site
1154         Ilukwatta     48.2   Temporary Site
1155              Null  12204.7             Null

[1156 rows x 7 columns]
```

Check and remove duplicates

[14]:
```
ud.duplicated()
```

[14]:
```
0       False
1       False
2       False
3       False
4       False
        …
1151    False
1152    False
1153    False
1154    False
```

```
1155    False
Length: 1156, dtype: bool
```

[15]:
```
ud.drop_duplicates()
```

[15]:
```
            START_DATE*         END_DATE* CATEGORY*            START*  \
0       1/1/2016 21:11    1/1/2016 21:17  Business       Fort Pierce
1        1/2/2016 1:25     1/2/2016 1:37  Business       Fort Pierce
2       1/2/2016 20:25    1/2/2016 20:38  Business       Fort Pierce
3       1/5/2016 17:31    1/5/2016 17:45  Business       Fort Pierce
4       1/6/2016 14:42    1/6/2016 15:49  Business       Fort Pierce
…                  …                 …       …                 …
1151  12/31/2016 13:24  12/31/2016 13:42  Business           Kar?chi
1152  12/31/2016 15:03  12/31/2016 15:38  Business  Unknown Location
1153  12/31/2016 21:32  12/31/2016 21:50  Business        Katunayake
1154  12/31/2016 22:08  12/31/2016 23:51  Business           Gampaha
1155            Totals               NaN       NaN               NaN

                  STOP*   MILES*         PURPOSE*
0          Fort Pierce      5.1   Meal/Entertain
1          Fort Pierce      5.0              NaN
2          Fort Pierce      4.8  Errand/Supplies
3          Fort Pierce      4.7          Meeting
4      West Palm Beach     63.7   Customer Visit
…                  …      …              …
1151  Unknown Location      3.9   Temporary Site
1152  Unknown Location     16.2          Meeting
1153           Gampaha      6.4   Temporary Site
1154         Ilukwatta     48.2   Temporary Site
1155               NaN  12204.7              NaN

[1155 rows x 7 columns]
```

Convert START_DATE and END_DATE to datetime

[16]:
```
ud['START_DATE*'] = pd.to_datetime(ud['START_DATE*'], errors='coerce')
ud.dtypes
```

[16]:
```
START_DATE*    datetime64[ns]
END_DATE*              object
CATEGORY*              object
START*                 object
STOP*                  object
MILES*                float64
PURPOSE*               object
dtype: object
```

[17]:
```
ud['END_DATE*'] = pd.to_datetime(ud['END_DATE*'], errors='coerce')
ud.dtypes
```

[17]:
```
START_DATE*    datetime64[ns]
END_DATE*      datetime64[ns]
CATEGORY*              object
START*                 object
STOP*                  object
MILES*                float64
PURPOSE*               object
dtype: object
```

Total number of rides per category:

[18]:
```
rides=ud.groupby('CATEGORY*').size()
rides
```

[18]:
```
CATEGORY*
Business    1078
Personal      77
dtype: int64
```

Total miles traveled for each purpose:

[20]:
```
total_miles=ud.groupby('PURPOSE*')['MILES*'].sum()
total_miles
```

[20]:
```
PURPOSE*
Airport/Travel       16.5
Between Offices     197.0
Charity ($)          15.1
Commute             180.2
Customer Visit     2089.5
Errand/Supplies     508.0
Meal/Entertain      911.7
Meeting            2851.3
Moving               18.2
Temporary Site      523.7
Name: MILES*, dtype: float64
```

Average distance for business vs. personal rides:

[23]:
```
filtered=ud[ud['CATEGORY*'].isin(['Business','Personal'])]
avg=filtered.groupby('CATEGORY*')['MILES*'].sum()
avg
```

[23]:
```
CATEGORY*
Business    11487.0
Personal      717.7
```

```
Name: MILES*, dtype: float64
```

Add a column for cost estimation (assuming $2 per mile):

[22]:
```
ud['COST_ESTIMATION'] = ud['MILES*'] * 2
print(ud[['MILES*', 'COST_ESTIMATION']].head())
```

```
   MILES*  COST_ESTIMATION
0     5.1             10.2
1     5.0             10.0
2     4.8              9.6
3     4.7              9.4
4    63.7            127.4
```

Filter rides longer than 50 miles:

[24]:
```
mile=ud[ud['MILES*']>50]
mile
```

[24]:

	START_DATE*	END_DATE*	CATEGORY*	START*	\
4	2016-01-06 14:42:00	2016-01-06 15:49:00	Business	Fort Pierce	
232	2016-03-17 12:52:00	2016-03-17 15:11:00	Business	Austin	
251	2016-03-19 19:33:00	2016-03-19 20:39:00	Business	Galveston	
268	2016-03-25 13:24:00	2016-03-25 16:22:00	Business	Cary	
269	2016-03-25 16:52:00	2016-03-25 22:22:00	Business	Latta	
270	2016-03-25 22:54:00	2016-03-26 01:39:00	Business	Jacksonville	
295	2016-04-02 12:21:00	2016-04-02 14:47:00	Business	Kissimmee	
296	2016-04-02 16:57:00	2016-04-02 18:09:00	Business	Daytona Beach	
297	2016-04-02 19:38:00	2016-04-02 22:36:00	Business	Jacksonville	
298	2016-04-02 23:11:00	2016-04-03 01:34:00	Business	Ridgeland	
299	2016-04-03 02:00:00	2016-04-03 04:16:00	Business	Florence	
546	2016-07-14 16:39:00	2016-07-14 20:05:00	Business	Morrisville	
559	2016-07-17 12:20:00	2016-07-17 15:25:00	Personal	Boone	
707	2016-08-24 13:01:00	2016-08-24 15:25:00	Business	Unknown Location	
710	2016-08-25 17:19:00	2016-08-25 19:20:00	Business	Unknown Location	
726	2016-08-27 14:01:00	2016-08-27 15:44:00	Business	Lahore	
727	2016-08-27 16:15:00	2016-08-27 19:13:00	Business	Unknown Location	
751	2016-09-06 17:49:00	2016-09-06 17:49:00	Business	Unknown Location	
776	2016-09-27 21:01:00	2016-09-28 02:37:00	Business	Unknown Location	
788	2016-10-06 17:23:00	2016-10-06 17:40:00	Business	R?walpindi	
869	2016-10-28 15:53:00	2016-10-28 17:59:00	Business	Cary	
870	2016-10-28 18:13:00	2016-10-28 20:07:00	Business	Winston Salem	
871	2016-10-28 20:13:00	2016-10-28 22:00:00	Business	Asheville	
873	2016-10-29 17:13:00	2016-10-29 19:19:00	Business	Hayesville	
880	2016-10-30 13:24:00	2016-10-30 14:37:00	Business	Bryson City	
881	2016-10-30 15:22:00	2016-10-30 18:23:00	Business	Asheville	
1088	2016-12-21 20:56:00	2016-12-21 23:42:00	Business	Rawalpindi	
1155	NaT	NaT	NaN	NaN	

	STOP*	MILES*	PURPOSE*	COST_ESTIMATION
4	West Palm Beach	63.7	Customer Visit	127.4
232	Katy	136.0	Customer Visit	272.0
251	Houston	57.0	Customer Visit	114.0
268	Latta	144.0	Customer Visit	288.0
269	Jacksonville	310.3	Customer Visit	620.6
270	Kissimmee	201.0	Meeting	402.0
295	Daytona Beach	77.3	Customer Visit	154.6
296	Jacksonville	80.5	Customer Visit	161.0
297	Ridgeland	174.2	Customer Visit	348.4
298	Florence	144.0	Meeting	288.0
299	Cary	159.3	Meeting	318.6
546	Banner Elk	195.3	NaN	390.6
559	Cary	180.2	Commute	360.4
707	Unknown Location	96.2	NaN	192.4
710	Unknown Location	50.4	NaN	100.8
726	Unknown Location	86.6	NaN	173.2
727	Unknown Location	156.9	NaN	313.8
751	Unknown Location	69.1	NaN	138.2
776	Unknown Location	195.6	NaN	391.2
788	Unknown Location	112.6	NaN	225.2
869	Winston Salem	107.0	Meeting	214.0
870	Asheville	133.6	Meeting	267.2
871	Topton	91.8	Meeting	183.6
873	Topton	75.7	NaN	151.4
880	Asheville	68.4	NaN	136.8
881	Mebane	195.9	NaN	391.8
1088	Unknown Location	103.0	Meeting	206.0
1155	NaN	12204.7	NaN	24409.4

Filter by specific purpose (e.g., meetings):

[25]:
```
purpose=ud[ud['PURPOSE*']=='Meeting']
purpose
```

[25]:

	START_DATE*	END_DATE*	CATEGORY*	START*	\
3	2016-01-05 17:31:00	2016-01-05 17:45:00	Business	Fort Pierce	
6	2016-01-06 17:30:00	2016-01-06 17:35:00	Business	West Palm Beach	
7	2016-01-07 13:27:00	2016-01-07 13:33:00	Business	Cary	
8	2016-01-10 08:05:00	2016-01-10 08:25:00	Business	Cary	
10	2016-01-10 15:08:00	2016-01-10 15:51:00	Business	New York	
…	…	…	…	…	
1142	2016-12-29 20:15:00	2016-12-29 20:45:00	Business	Kar?chi	
1144	2016-12-29 23:14:00	2016-12-29 23:47:00	Business	Unknown Location	
1148	2016-12-30 16:45:00	2016-12-30 17:08:00	Business	Kar?chi	
1150	2016-12-31 01:07:00	2016-12-31 01:14:00	Business	Kar?chi	
1152	2016-12-31 15:03:00	2016-12-31 15:38:00	Business	Unknown Location	

```
                  STOP*  MILES* PURPOSE*  COST_ESTIMATION
3           Fort Pierce     4.7  Meeting              9.4
6            Palm Beach     7.1  Meeting             14.2
7                  Cary     0.8  Meeting              1.6
8           Morrisville     8.3  Meeting             16.6
10               Queens    10.8  Meeting             21.6
…                   …      …       …                …
1142            Kar?chi     7.2  Meeting             14.4
1144            Kar?chi    12.9  Meeting             25.8
1148            Kar?chi     4.6  Meeting              9.2
1150            Kar?chi     0.7  Meeting              1.4
1152  Unknown Location     16.2  Meeting             32.4

[187 rows x 8 columns]
```

What is the total number of business trips versus personal trips?

[26]:
```
business_trips = ud[ud['CATEGORY*'] == 'Business'].shape[0]
personal_trips = ud[ud['CATEGORY*'] == 'Personal'].shape[0]
print(business_trips)
print(personal_trips)
```

```
1078
77
```

What percentage of trips are business versus personal?

[27]:
```
total_trips=business_trips+personal_trips
bus_per = (business_trips / total_trips) * 100
bus_per
```

[27]:
```
93.33333333333333
```

[28]:
```
per_per = (personal_trips / total_trips) * 100
per_per
```

[28]:
```
6.666666666666667
```

[ ]:
```

```