

```
In [1]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

```
In [2]: dataset = pd.read_csv("UberDataset.csv")
```

```
In [3]: dataset
```

```
Out[3]:
```

	START_DATE	END_DATE	CATEGORY	START	STOP	MILES	PURPOSE
0	01-01-2016 21:11	01-01-2016 21:17	Business	Fort Pierce	Fort Pierce	5.1	Meal/Entertain
1	01-02-2016 01:25	01-02-2016 01:37	Business	Fort Pierce	Fort Pierce	5.0	NaN
2	01-02-2016 20:25	01-02-2016 20:38	Business	Fort Pierce	Fort Pierce	4.8	Errand/Supplies
3	01-05-2016 17:31	01-05-2016 17:45	Business	Fort Pierce	Fort Pierce	4.7	Meeting
4	01-06-2016 14:42	01-06-2016 15:49	Business	Fort Pierce	West Palm Beach	63.7	Customer Visit
...
1151	12/31/2016 13:24	12/31/2016 13:42	Business	Kar?chi	Unknown Location	3.9	Temporary Site
1152	12/31/2016 15:03	12/31/2016 15:38	Business	Unknown Location	Unknown Location	16.2	Meeting
1153	12/31/2016 21:32	12/31/2016 21:50	Business	Katunayake	Gampaha	6.4	Temporary Site
1154	12/31/2016 22:08	12/31/2016 23:51	Business	Gampaha	Ilukwatta	48.2	Temporary Site
1155	Totals	NaN	NaN	NaN	NaN	12204.7	NaN

1156 rows × 7 columns

```
In [4]: dataset.shape
```

```
Out[4]: (1156, 7)
```

In [5]: `dataset.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
```

Data Preprocessing

In [6]: `dataset['PURPOSE'].fillna("NOT", inplace=True)`

In [7]: `dataset.head()`

Out[7]:

	START_DATE	END_DATE	CATEGORY	START	STOP	MILES	PURPOSE
0	01-01-2016 21:11	01-01-2016 21:17	Business	Fort Pierce	Fort Pierce	5.1	Meal/Entertain
1	01-02-2016 01:25	01-02-2016 01:37	Business	Fort Pierce	Fort Pierce	5.0	NOT
2	01-02-2016 20:25	01-02-2016 20:38	Business	Fort Pierce	Fort Pierce	4.8	Errand/Supplies
3	01-05-2016 17:31	01-05-2016 17:45	Business	Fort Pierce	Fort Pierce	4.7	Meeting
4	01-06-2016 14:42	01-06-2016 15:49	Business	Fort Pierce	West Palm Beach	63.7	Customer Visit

In [8]: `dataset['START_DATE']=pd.to_datetime(dataset['START_DATE'], errors='coerce')`
`dataset['END_DATE']=pd.to_datetime(dataset['END_DATE'], errors='coerce')`

In [9]: `dataset.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  421 non-null    datetime64[ns]
1   END_DATE    420 non-null    datetime64[ns]
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     1156 non-null   object
dtypes: datetime64[ns](2), float64(1), object(4)
memory usage: 63.3+ KB
```

In [10]: `from datetime import datetime`

```
dataset['date'] = pd.DatetimeIndex(dataset['START_DATE']).date
```

In [11]: `dataset.head()`

Out[11]:

	START_DATE	END_DATE	CATEGORY	START	STOP	MILES	PURPOSE	date
0	2016-01-01 21:11:00	2016-01-01 21:17:00	Business	Fort Pierce	Fort Pierce	5.1	Meal/Entertain	2016- 01-01
1	2016-01-02 01:25:00	2016-01-02 01:37:00	Business	Fort Pierce	Fort Pierce	5.0	NOT	2016- 01-02
2	2016-01-02 20:25:00	2016-01-02 20:38:00	Business	Fort Pierce	Fort Pierce	4.8	Errand/Supplies	2016- 01-02
3	2016-01-05 17:31:00	2016-01-05 17:45:00	Business	Fort Pierce	Fort Pierce	4.7	Meeting	2016- 01-05
4	2016-01-06 14:42:00	2016-01-06 15:49:00	Business	Fort Pierce	West Palm Beach	63.7	Customer Visit	2016- 01-06

In [12]: `from datetime import datetime`

```
dataset['date'] = pd.DatetimeIndex(dataset['START_DATE']).date
dataset['time'] = pd.DatetimeIndex(dataset['START_DATE']).hour
```

In [13]: `dataset.head()`

Out[13]:

	START_DATE	END_DATE	CATEGORY	START	STOP	MILES	PURPOSE	date	time
0	2016-01-01 21:11:00	2016-01-01 21:17:00	Business	Fort Pierce	Fort Pierce	5.1	Meal/Entertain	2016- 01-01	21.0
1	2016-01-02 01:25:00	2016-01-02 01:37:00	Business	Fort Pierce	Fort Pierce	5.0	NOT	2016- 01-02	1.0
2	2016-01-02 20:25:00	2016-01-02 20:38:00	Business	Fort Pierce	Fort Pierce	4.8	Errand/Supplies	2016- 01-02	20.0
3	2016-01-05 17:31:00	2016-01-05 17:45:00	Business	Fort Pierce	Fort Pierce	4.7	Meeting	2016- 01-05	17.0
4	2016-01-06 14:42:00	2016-01-06 15:49:00	Business	Fort Pierce	West Palm Beach	63.7	Customer Visit	2016- 01-06	14.0

In [14]: `dataset['day-night']=pd.cut(x=dataset['time'], bins=[0,10,15,19,24], labels= [`

In [15]: `dataset.head()`

Out[15]:

	START_DATE	END_DATE	CATEGORY	START	STOP	MILES	PURPOSE	date	time
0	2016-01-01 21:11:00	2016-01-01 21:17:00	Business	Fort Pierce	Fort Pierce	5.1	Meal/Entertain	2016- 01-01	21.0
1	2016-01-02 01:25:00	2016-01-02 01:37:00	Business	Fort Pierce	Fort Pierce	5.0	NOT	2016- 01-02	1.0
2	2016-01-02 20:25:00	2016-01-02 20:38:00	Business	Fort Pierce	Fort Pierce	4.8	Errand/Supplies	2016- 01-02	20.0
3	2016-01-05 17:31:00	2016-01-05 17:45:00	Business	Fort Pierce	Fort Pierce	4.7	Meeting	2016- 01-05	17.0
4	2016-01-06 14:42:00	2016-01-06 15:49:00	Business	Fort Pierce	West Palm Beach	63.7	Customer Visit	2016- 01-06	14.0



In [16]: `dataset.dropna(inplace=True)`

In [17]: `dataset.shape`

Out[17]: (413, 10)

Data Visualization

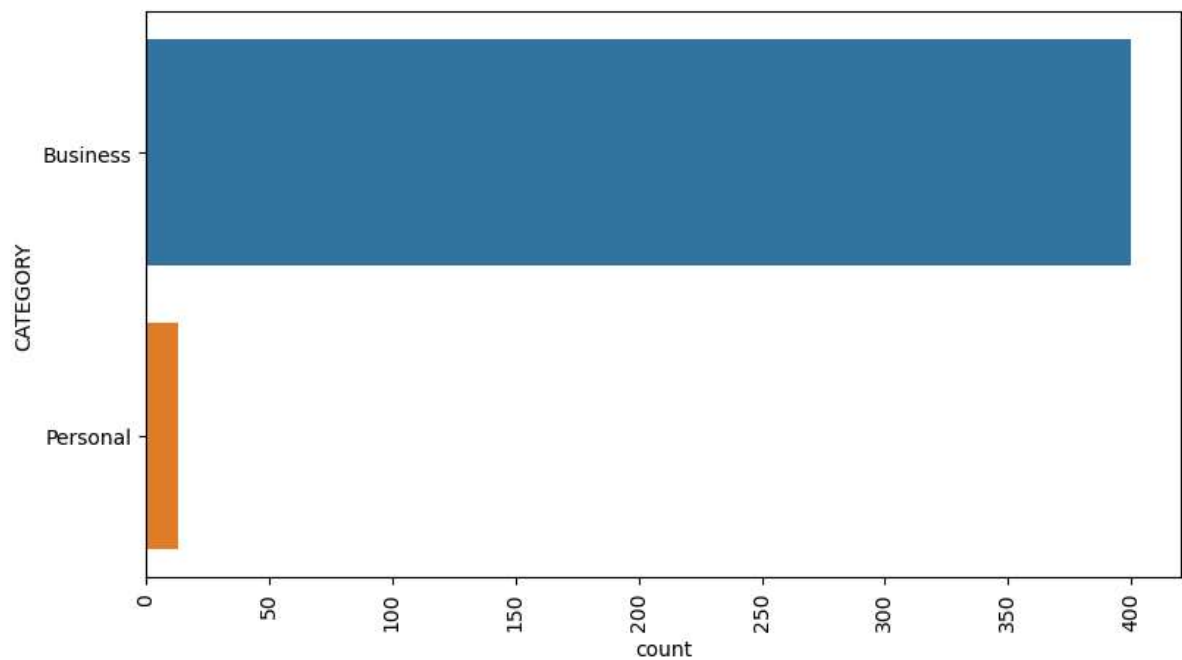
```
In [18]: plt.figure(figsize=(20,5))

plt.subplot(1,2,1)

sns.countplot(y='CATEGORY', data=dataset)

plt.xticks(rotation =90)
```

```
Out[18]: (array([ 0., 50., 100., 150., 200., 250., 300., 350., 400., 450.]),
 [Text(0.0, 0, '0'),
  Text(50.0, 0, '50'),
  Text(100.0, 0, '100'),
  Text(150.0, 0, '150'),
  Text(200.0, 0, '200'),
  Text(250.0, 0, '250'),
  Text(300.0, 0, '300'),
  Text(350.0, 0, '350'),
  Text(400.0, 0, '400'),
  Text(450.0, 0, '450')])
```



Conclusion: Business Category do people book the most Uber rides.

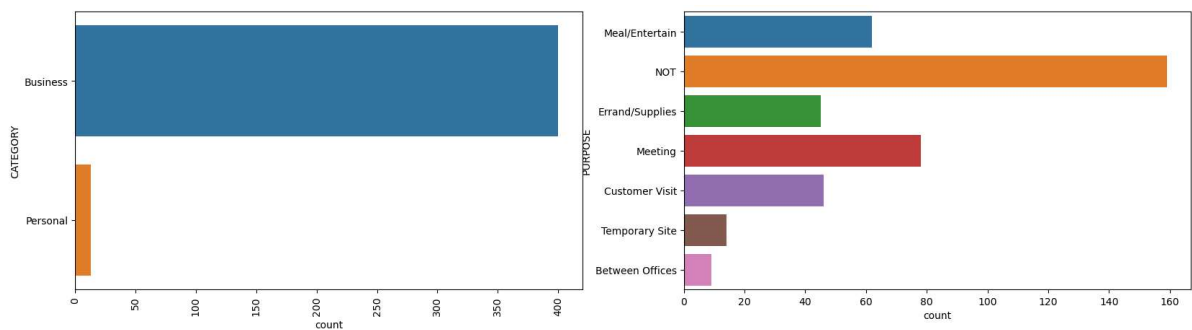
```
In [19]: plt.figure(figsize=(20,5))

plt.subplot(1,2,1)

sns.countplot(y='CATEGORY', data=dataset)

plt.xticks(rotation =90)
plt.subplot(1,2,2)
sns.countplot(y='PURPOSE', data=dataset)
```

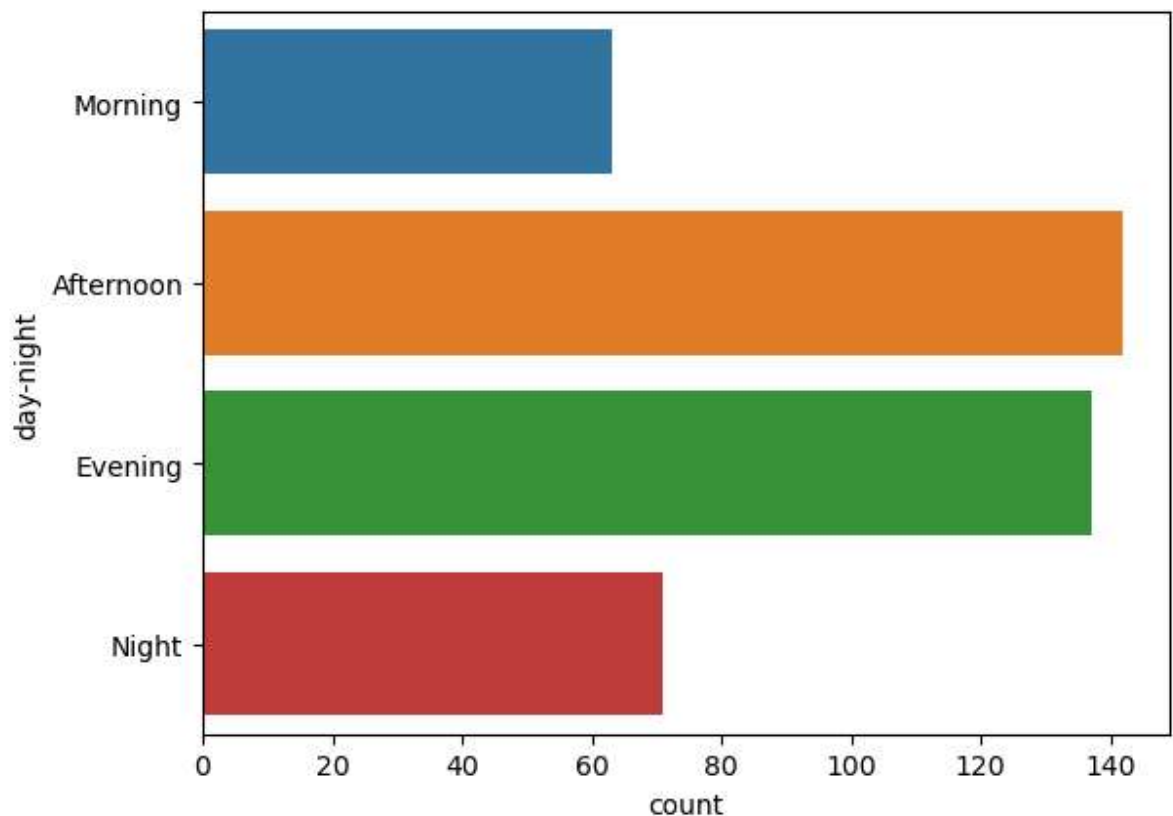
Out[19]: <Axes: xlabel='count', ylabel='PURPOSE'>



Conclusion: For Meeting purpose do people book Uber rides the most.

```
In [20]: sns.countplot(y='day-night', data=dataset)
```

Out[20]: <Axes: xlabel='count', ylabel='day-night'>



Conclusion: At Afternoon do people book cabs the most from Uber.

In [21]: dataset.head()

Out[21]:

	START_DATE	END_DATE	CATEGORY	START	STOP	MILES	PURPOSE	date	time
0	2016-01-01 21:11:00	2016-01-01 21:17:00	Business	Fort Pierce	Fort Pierce	5.1	Meal/Entertain	2016- 01-01	21.0
1	2016-01-02 01:25:00	2016-01-02 01:37:00	Business	Fort Pierce	Fort Pierce	5.0	NOT	2016- 01-02	1.0
2	2016-01-02 20:25:00	2016-01-02 20:38:00	Business	Fort Pierce	Fort Pierce	4.8	Errand/Supplies	2016- 01-02	20.0
3	2016-01-05 17:31:00	2016-01-05 17:45:00	Business	Fort Pierce	Fort Pierce	4.7	Meeting	2016- 01-05	17.0
4	2016-01-06 14:42:00	2016-01-06 15:49:00	Business	Fort Pierce	West Palm Beach	63.7	Customer Visit	2016- 01-06	14.0

```
In [22]: dataset['MONTH']=pd.DatetimeIndex(dataset['START_DATE']).month
month_label= {1.0: 'Jan', 2.0: 'Feb', 3.0: 'Mar', 4.0: 'April',
              5.0: 'May', 6.0: 'June', 7.0: 'July', 8.0: 'Aug',
              9.0: 'Sep', 10.0: 'Oct', 11.0: 'Nov', 12.0: 'Dec'}
dataset["MONTH"]=dataset.MONTH.map(month_label)
mon= dataset.MONTH.value_counts(sort=False)
```

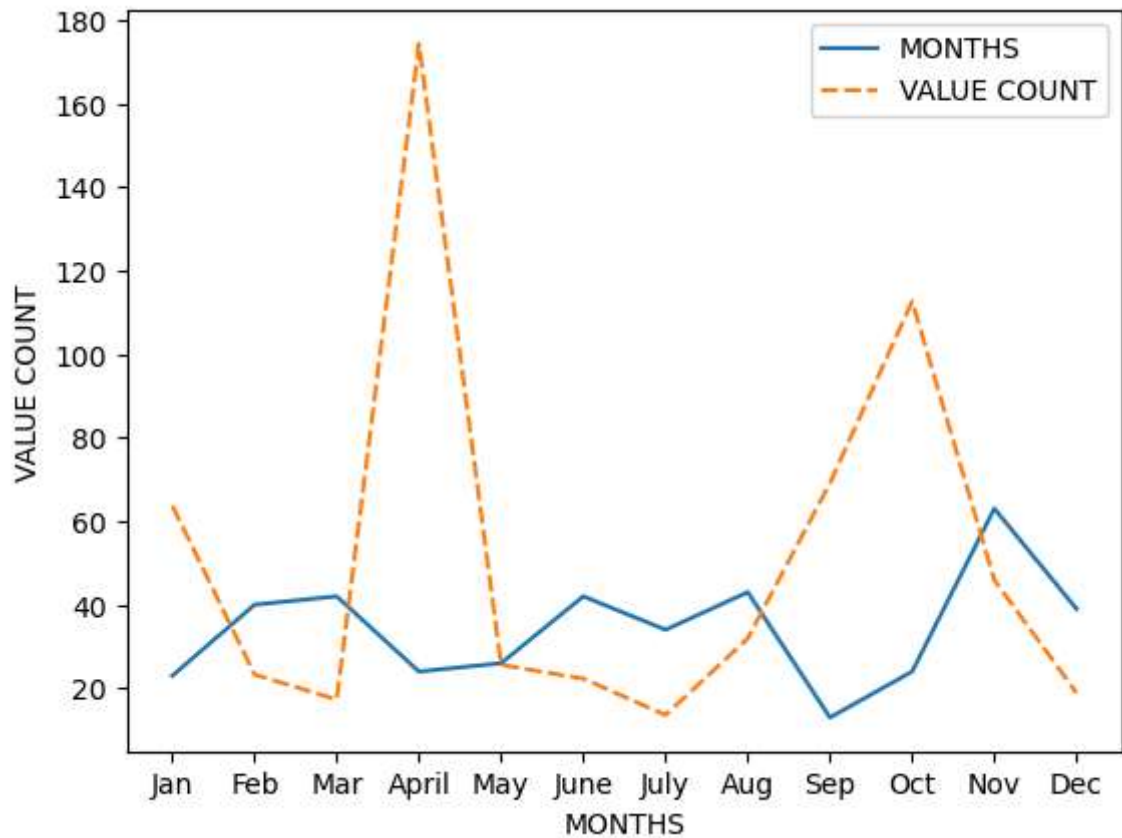
In [23]: dataset.head()

Out[23]:

	START_DATE	END_DATE	CATEGORY	START	STOP	MILES	PURPOSE	date	time
0	2016-01-01 21:11:00	2016-01-01 21:17:00	Business	Fort Pierce	Fort Pierce	5.1	Meal/Entertain	2016- 01-01	21.0
1	2016-01-02 01:25:00	2016-01-02 01:37:00	Business	Fort Pierce	Fort Pierce	5.0	NOT	2016- 01-02	1.0
2	2016-01-02 20:25:00	2016-01-02 20:38:00	Business	Fort Pierce	Fort Pierce	4.8	Errand/Supplies	2016- 01-02	20.0
3	2016-01-05 17:31:00	2016-01-05 17:45:00	Business	Fort Pierce	Fort Pierce	4.7	Meeting	2016- 01-05	17.0
4	2016-01-06 14:42:00	2016-01-06 15:49:00	Business	Fort Pierce	West Palm Beach	63.7	Customer Visit	2016- 01-06	14.0

```
In [24]: df=pd.DataFrame({  
    "MONTHS": mon.values,  
    "VALUE COUNT": dataset.groupby('MONTH', sort=False)['MILES'].max()  
    })  
  
p=sns.lineplot(data=df)  
p.set(xlabel="MONTHS", ylabel="VALUE COUNT")
```

```
Out[24]: [Text(0.5, 0, 'MONTHS'), Text(0, 0.5, 'VALUE COUNT')]
```



Conclusion: In November, December and January do People book Uber rides less frequency

In [25]: `dataset.head()`

Out[25]:

	START_DATE	END_DATE	CATEGORY	START	STOP	MILES	PURPOSE	date	time
0	2016-01-01 21:11:00	2016-01-01 21:17:00	Business	Fort Pierce	Fort Pierce	5.1	Meal/Entertain	2016- 01-01	21.0
1	2016-01-02 01:25:00	2016-01-02 01:37:00	Business	Fort Pierce	Fort Pierce	5.0	NOT	2016- 01-02	1.0
2	2016-01-02 20:25:00	2016-01-02 20:38:00	Business	Fort Pierce	Fort Pierce	4.8	Errand/Supplies	2016- 01-02	20.0
3	2016-01-05 17:31:00	2016-01-05 17:45:00	Business	Fort Pierce	Fort Pierce	4.7	Meeting	2016- 01-05	17.0
4	2016-01-06 14:42:00	2016-01-06 15:49:00	Business	Fort Pierce	West Palm Beach	63.7	Customer Visit	2016- 01-06	14.0

In [26]: `dataset['DAY']=dataset.START_DATE.dt.weekday`

```
day_label={
    0:'Mon', 1:'Tues', 2:'Wed', 3:'Thur', 4:'Fri', 5:'Sat', 6:'Sun'}
dataset['DAY']= dataset['DAY'].map(day_label)
```

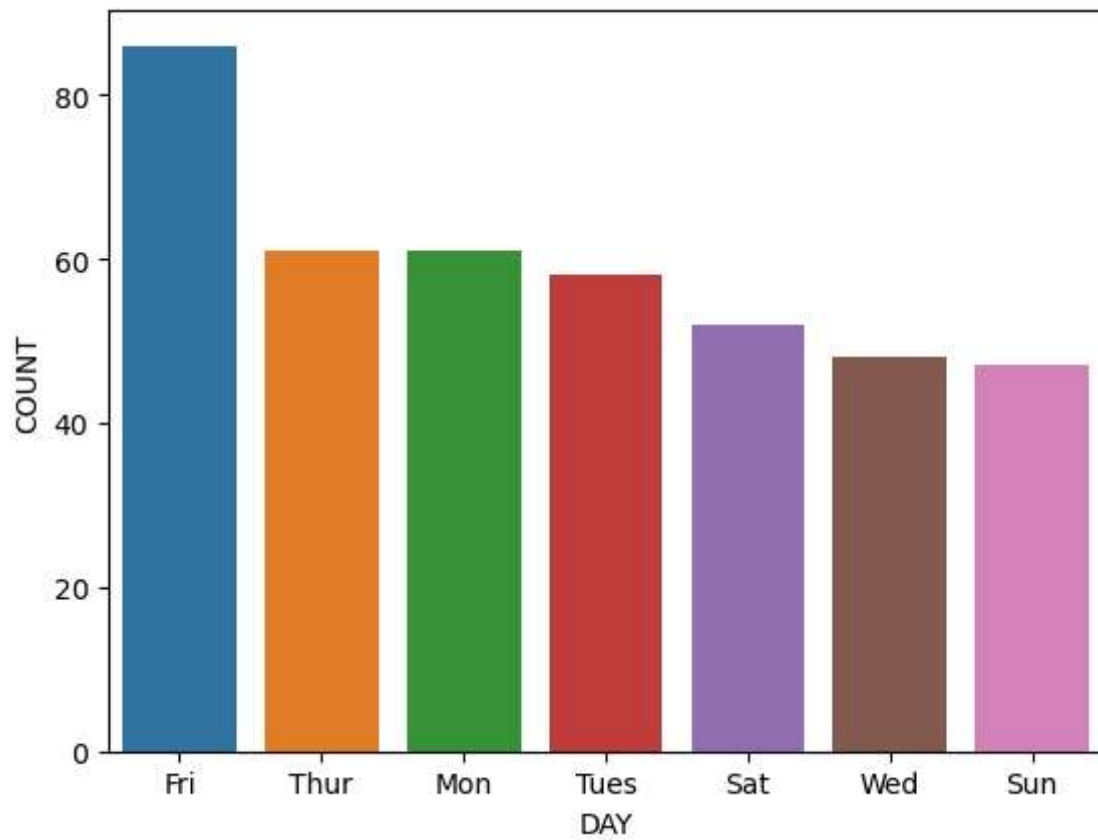
In [27]: `dataset.head()`

Out[27]:

	START_DATE	END_DATE	CATEGORY	START	STOP	MILES	PURPOSE	date	time
0	2016-01-01 21:11:00	2016-01-01 21:17:00	Business	Fort Pierce	Fort Pierce	5.1	Meal/Entertain	2016- 01-01	21.0
1	2016-01-02 01:25:00	2016-01-02 01:37:00	Business	Fort Pierce	Fort Pierce	5.0	NOT	2016- 01-02	1.0
2	2016-01-02 20:25:00	2016-01-02 20:38:00	Business	Fort Pierce	Fort Pierce	4.8	Errand/Supplies	2016- 01-02	20.0
3	2016-01-05 17:31:00	2016-01-05 17:45:00	Business	Fort Pierce	Fort Pierce	4.7	Meeting	2016- 01-05	17.0
4	2016-01-06 14:42:00	2016-01-06 15:49:00	Business	Fort Pierce	West Palm Beach	63.7	Customer Visit	2016- 01-06	14.0

```
In [28]: day_label=dataset.DAY.value_counts()  
sns.barplot(x=day_label.index, y=day_label)  
plt.xlabel('DAY')  
plt.ylabel('COUNT')
```

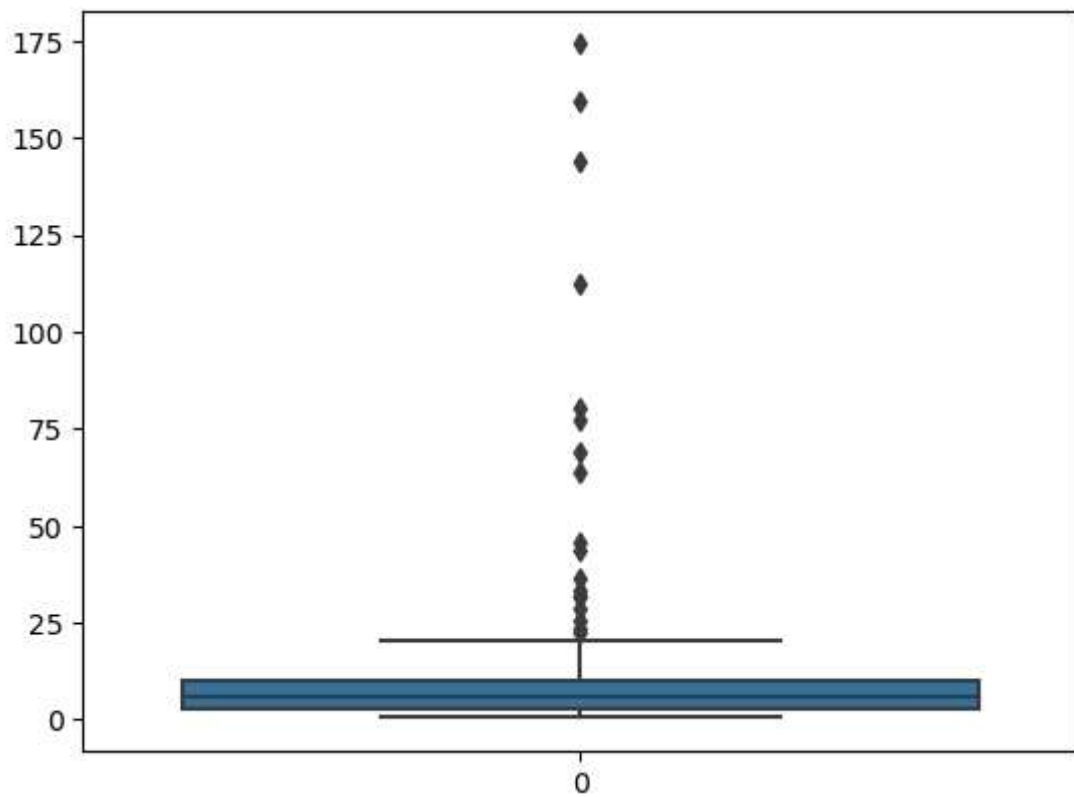
```
Out[28]: Text(0, 0.5, 'COUNT')
```



Conclusion: On Friday do People book Uber rides the most

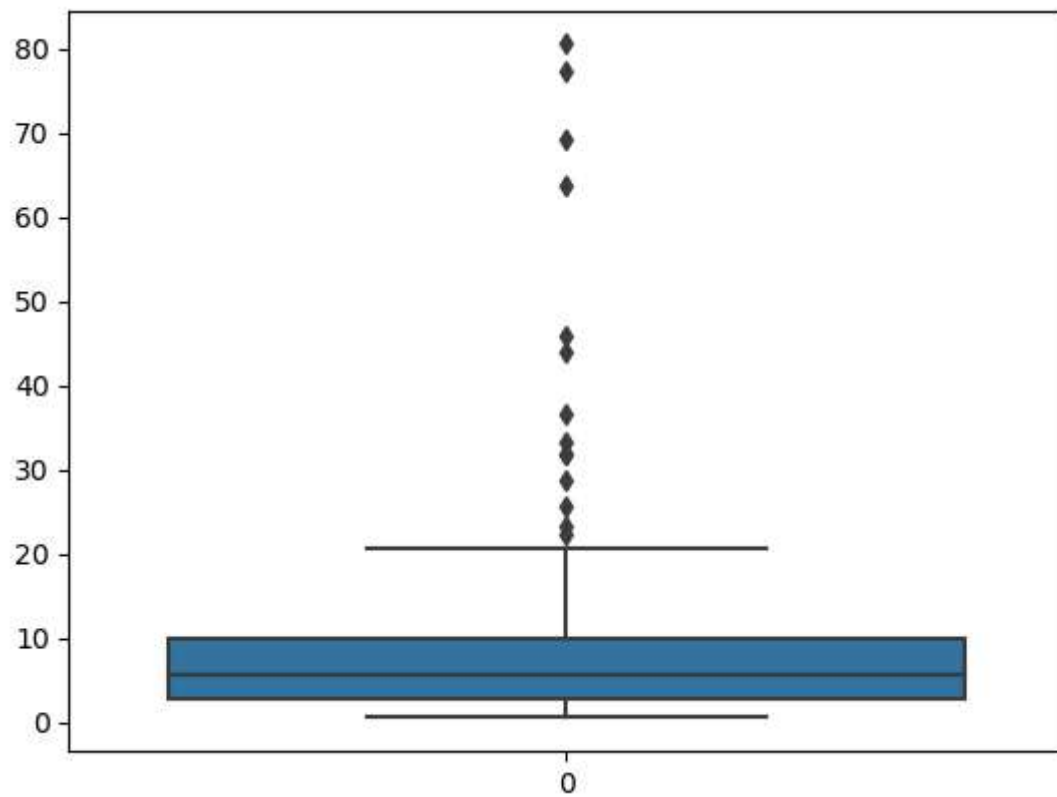
```
In [29]: sns.boxplot(dataset['MILES'])
```

```
Out[29]: <Axes: >
```



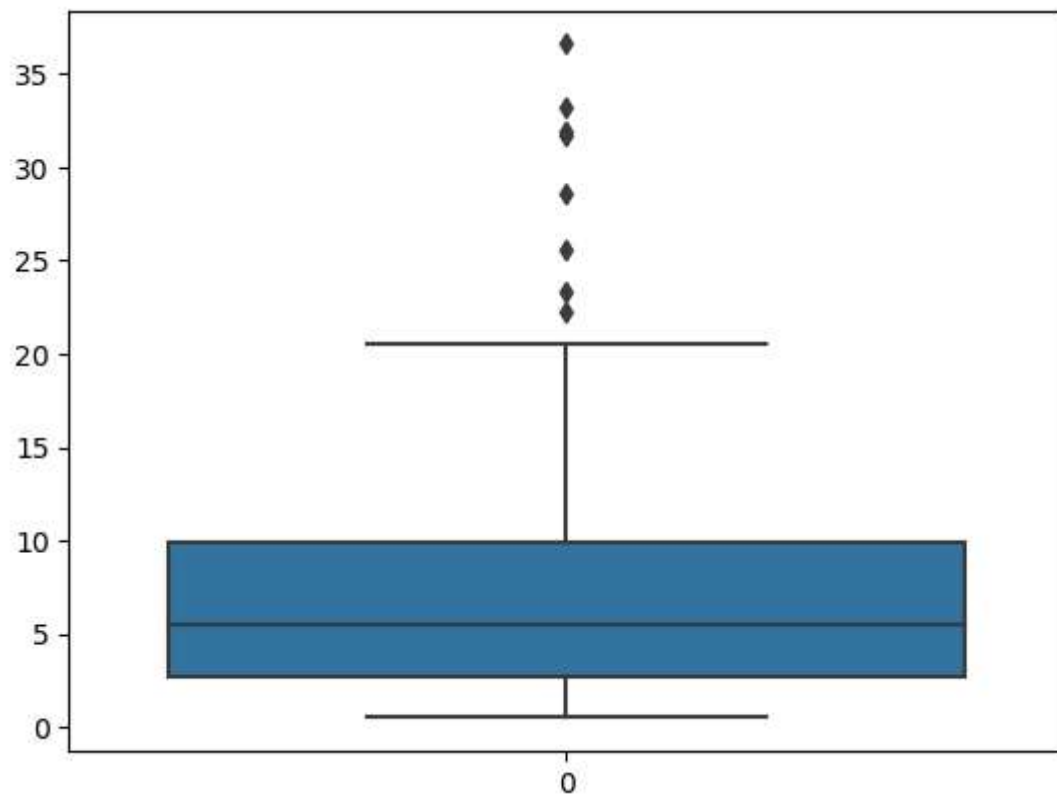
```
In [32]: sns.boxplot(dataset[dataset['MILES'] < 100]['MILES'])
```

```
Out[32]: <Axes: >
```



```
In [33]: sns.boxplot(dataset[dataset['MILES'] < 40]['MILES'])
```

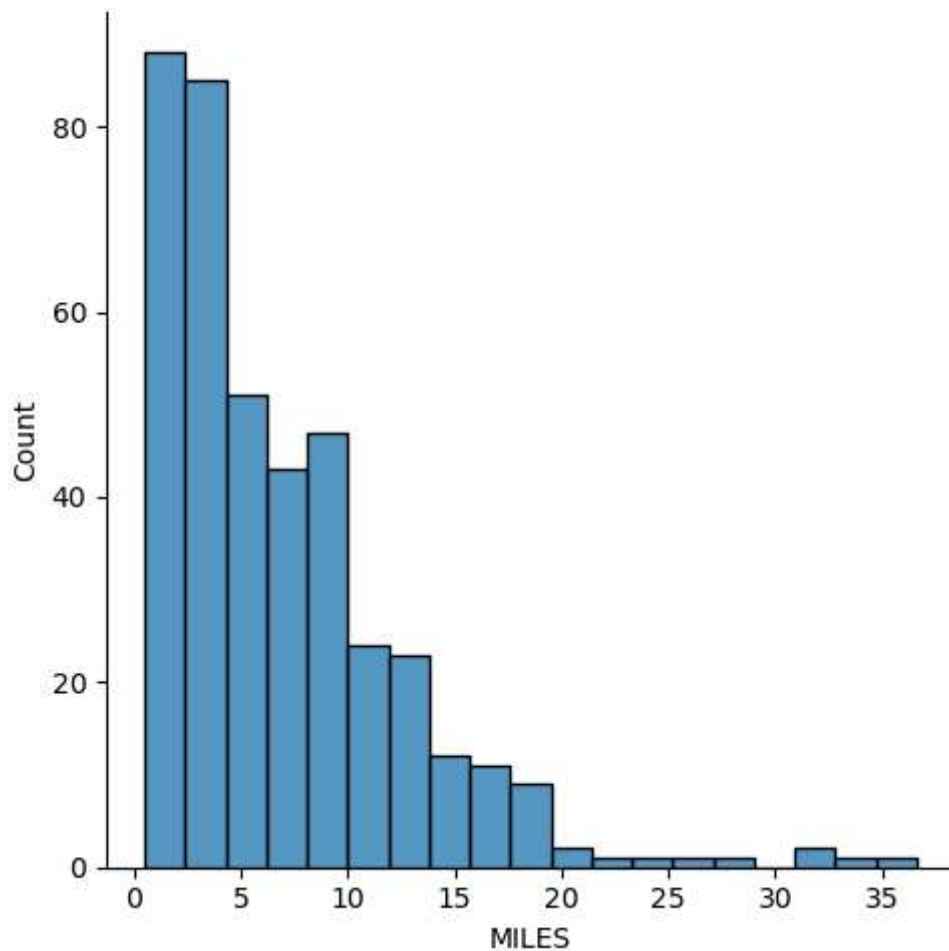
```
Out[33]: <Axes: >
```



```
In [35]: sns.displot(dataset[dataset['MILES'] < 40]['MILES'])
```

```
C:\Users\win 10\anaconda3\Lib\site-packages\seaborn\axisgrid.py:118: UserWarning: The figure layout has changed to tight
  self._figure.tight_layout(*args, **kwargs)
```

```
Out[35]: <seaborn.axisgrid.FacetGrid at 0x2505bbe3150>
```



Conclusion: 0 to 20 Miles do People Usually book a cab for through Uber

```
In [ ]:
```