Uber Rides Data analysis using Python

This project analyzes Uber rides data to uncover insights such as peak ride times, ride distances, and day-night travel patterns.

We use python libraries such as Pandas, Matplotlib and seaborn for data manipulation and Visualization.

Importing Required Libraries

We import required **Python** libraries such as Pandas for Data Manipulation and Matplotlib, Seaborn for Data Visualization.

In [2]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

Loading The Dataset

We load the Uber Dataset from a CSV File. It contains columns like Start Date, End Date, Miles traveled and Purpose.

In [3]:
```
df = pd.read_csv(r'C:\Users\HP\Desktop\Python Jupyter Project\Uber Rides Data An
```

In [4]:
```
df.head()
```

Out[4]:

	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

In [5]:
```
df.shape
```

Out[5]:
```
(1156, 7)
```

In [6]:
```
df.info
```

Out[6]:

```
<bound method DataFrame.info of             START_DATE          END_DATE  CATEG
ORY              START  \
0     01-01-2016 21:11  01-01-2016 21:17  Business       Fort Pierce
1     01-02-2016 01:25  01-02-2016 01:37  Business       Fort Pierce
2     01-02-2016 20:25  01-02-2016 20:38  Business       Fort Pierce
3     01-05-2016 17:31  01-05-2016 17:45  Business       Fort Pierce
4     01-06-2016 14:42  01-06-2016 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

[1156 rows x 7 columns]>
```

In [7]:

```
pd.isnull(df)
```

Out[7]:

	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 × 7 columns

In [8]:

```
pd.isnull(df).sum()
```

```
Out[8]: START_DATE      0
        END_DATE        1
        CATEGORY        1
        START           1
        STOP            1
        MILES           0
        PURPOSE       503
        dtype: int64
```

Now we drop null values and irrelevant columns to clean the dataset for analyzing.

```
In [9]: df.dropna(inplace = True)
```

```
In [10]: df.shape
```

```
Out[10]: (653, 7)
```

```
In [11]: df.drop_duplicates(inplace=True)
```

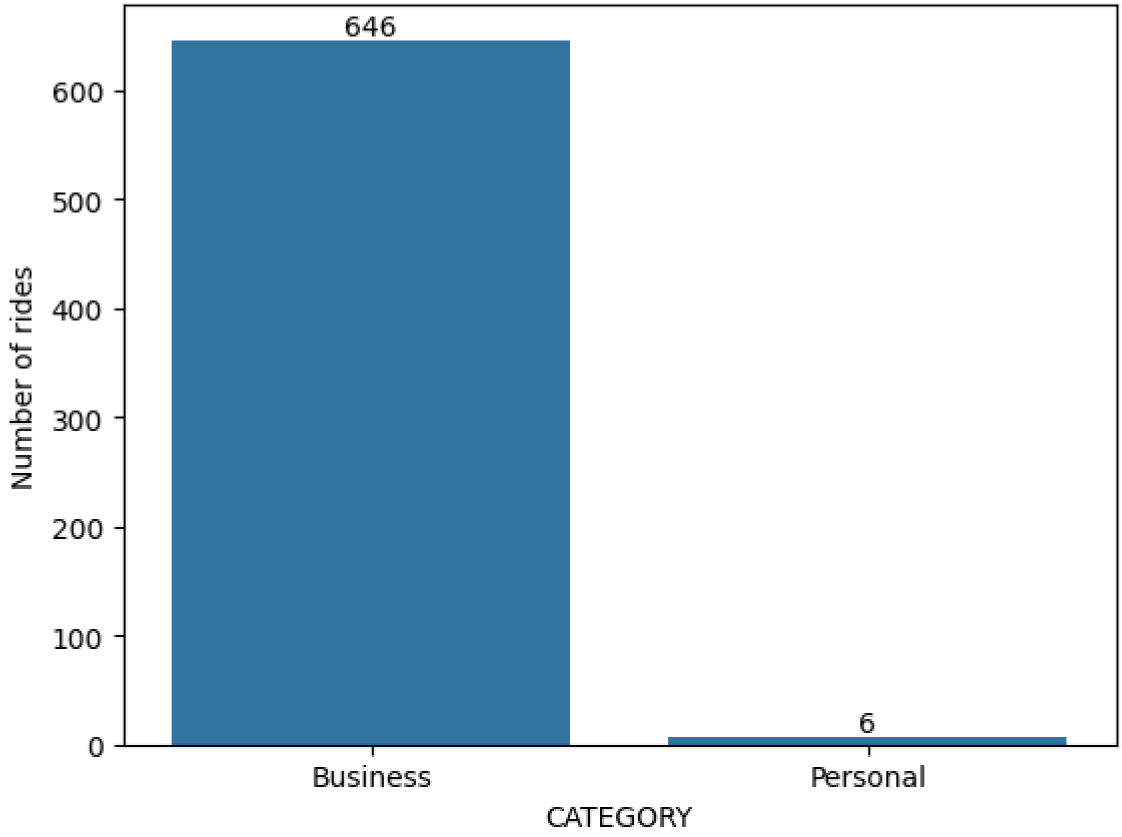
```
In [12]: df.shape
```

```
Out[12]: (652, 7)
```

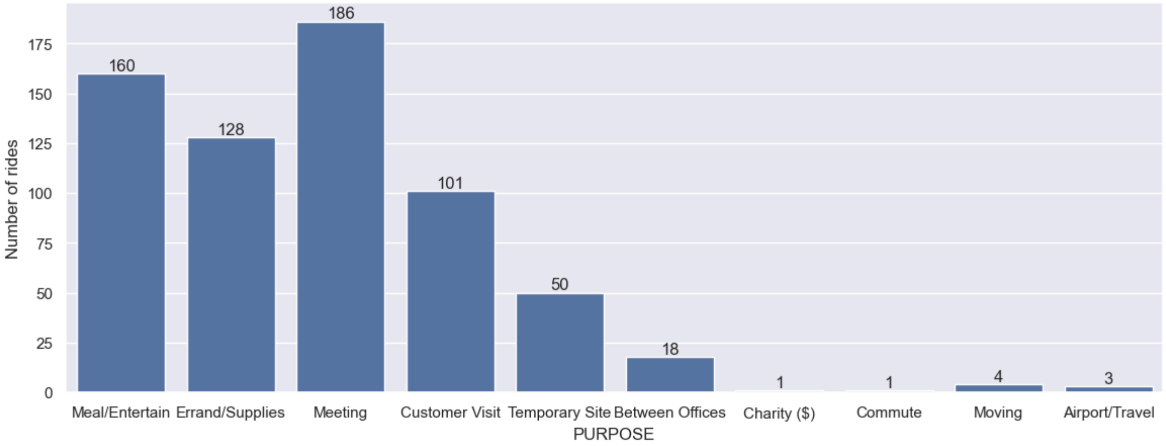
```
In [13]: df.columns
```

```
Out[13]: Index(['START_DATE', 'END_DATE', 'CATEGORY', 'START', 'STOP', 'MILES',
                'PURPOSE'],
               dtype='object')
```

```
In [14]: ax = sns.countplot(x = 'CATEGORY', data=df)
         plt.ylabel("Number of rides")

         for bars in ax.containers:
             ax.bar_label(bars)
```



In [29]:
```
sns.set(rc={'figure.figsize':(14,5)})
ax = sns.countplot(x = 'PURPOSE', data=df)
plt.ylabel("Number of rides")

for bars in ax.containers:
    ax.bar_label(bars)
```



In [16]:
```
df['START_DATE'] = pd.to_datetime(df['START_DATE'], errors = 'coerce')
df['END_DATE'] = pd.to_datetime(df['END_DATE'], errors = 'coerce')
```

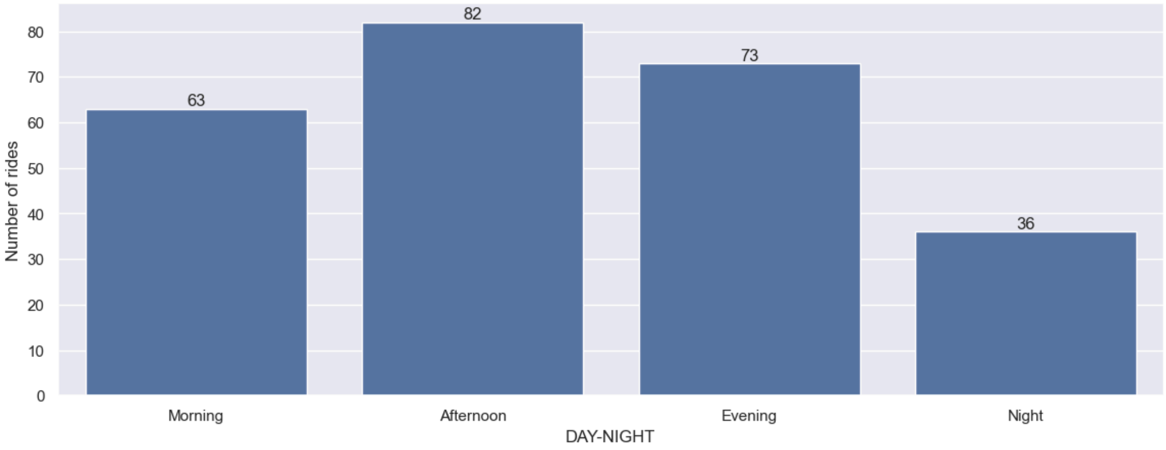
Now we categorize ride times into Morning, afternoon, evening and night to find when users travel the most.

In [17]:
```
from datetime import datetime

df['DATE'] = pd.DatetimeIndex(df['START_DATE']).date
df['TIME'] = pd.DatetimeIndex(df['START_DATE']).hour
```

```
df['DAY-NIGHT'] = pd.cut(x = df['TIME'], bins = [0, 12, 16, 20, 24], labels = ['
```

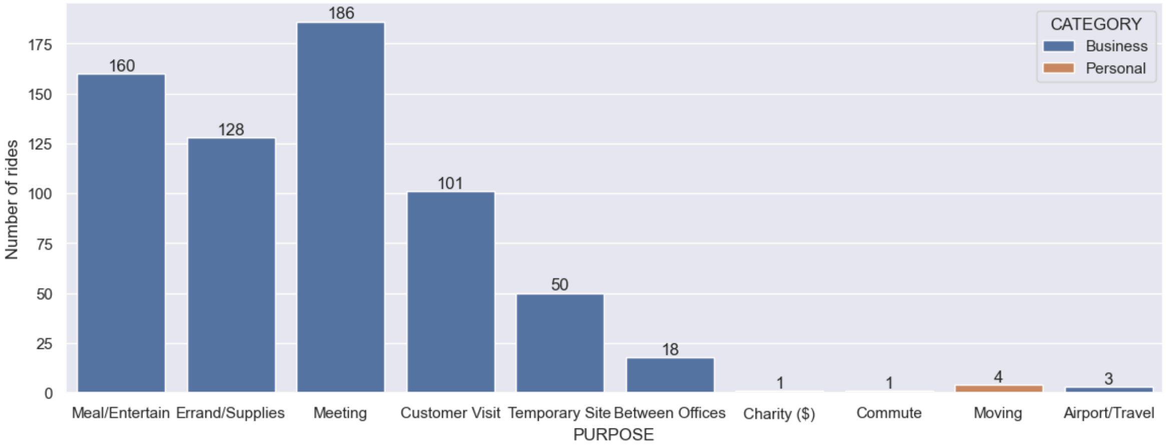
In [18]:

```
ax = sns.countplot(x = 'DAY-NIGHT', data=df)
plt.ylabel("Number of rides")

for bars in ax.containers:
    ax.bar_label(bars)
```



In [19]:

```
ax = sns.countplot(x = 'PURPOSE', hue = 'CATEGORY', data = df)

plt.ylabel("Number of rides")

for bars in ax.containers:
    ax.bar_label(bars)
```



Insights from the above count plots:

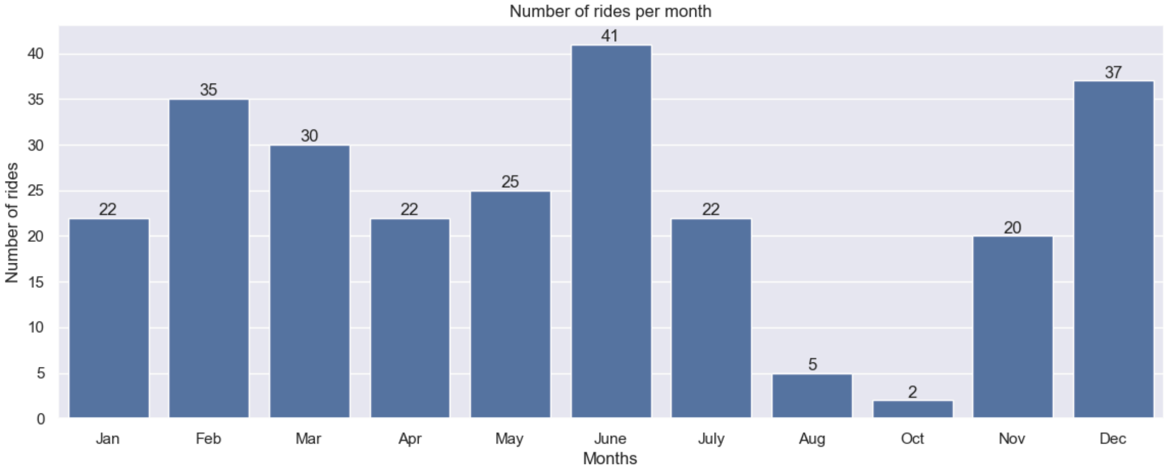
1. Most of the cabs are book for the **business purpose**.
2. Most of the people book cabs for the **Meeting and Meal/Entertainment Purpose**.
3. Most number of cabs are book in the **Afternoon (from 12 PM to 4 PM)**.

In [20]:

```
df['Month'] = pd.DatetimeIndex(df['START_DATE']).month
month_label = {1.0:'Jan', 2.0:'Feb', 3.0:'Mar', 4.0:'Apr', 5.0:'May', 6.0:'June'

df['Month'] = df.Month.map(month_label)
```

In [21]:

```
ax = sns.countplot(x = 'Month', data=df)
plt.title("Number of rides per month")
plt.xlabel("Months")
plt.ylabel("Number of rides")
# plt.show()

for bars in ax.containers:
    ax.bar_label(bars)
```



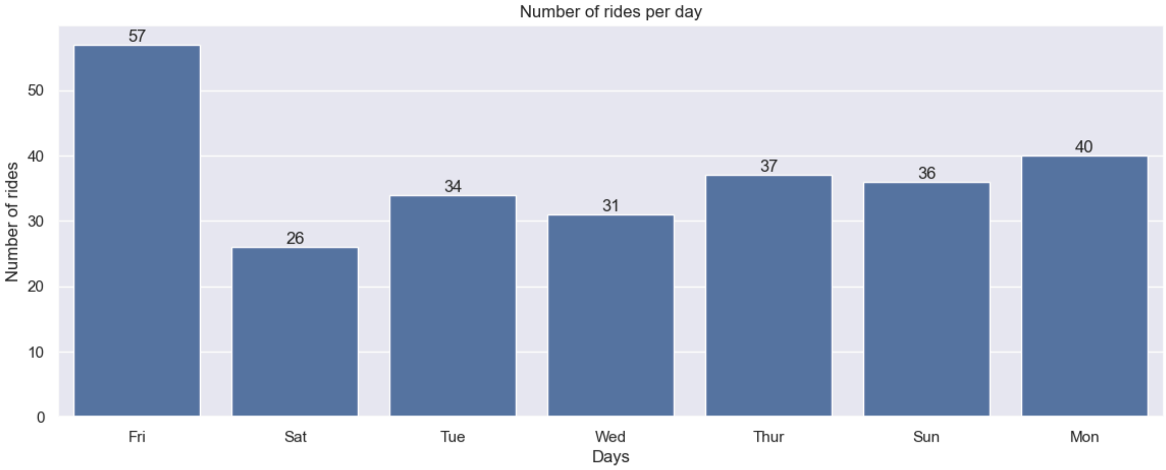
From the above graph, we can see that most of the cabs are book in the month of **June, Dec and Feb**.

In [22]:

```
df['Days'] = pd.DatetimeIndex(df['START_DATE']).weekday
# df['Days'] = df.START_DATE.dt.weekday

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

In [23]:

```
ax = sns.countplot(x = 'Days', data = df)
plt.title("Number of rides per day")
plt.xlabel("Days")
plt.ylabel("Number of rides")

for bars in ax.containers:
    ax.bar_label(bars)
```



From the above graph, we clearly see that most of the cabs are book on **Fridays**.

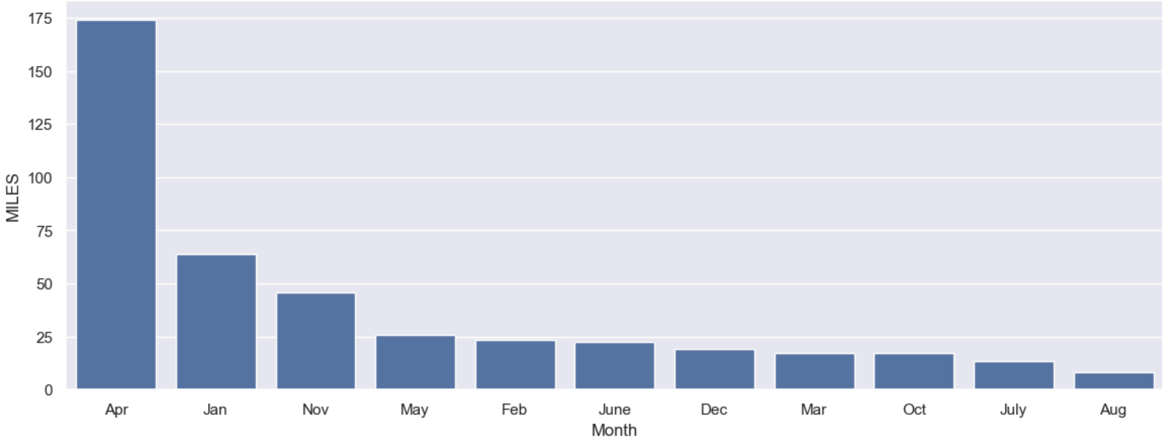
In [24]:

```
df.groupby('Month')['MILES'].max()
```

Out[24]:
```
Month
Apr     174.2
Aug       8.4
Dec      18.9
Feb      23.3
Jan      63.7
July     13.3
June     22.3
Mar      17.3
May      25.6
Nov      45.9
Oct      17.1
Name: MILES, dtype: float64
```

In [25]:
```
mile_dis = df.groupby('Month', as_index=False)['MILES'].max().sort_values(by='MI

sns.barplot(x = 'Month', y = 'MILES', data = mile_dis)
```

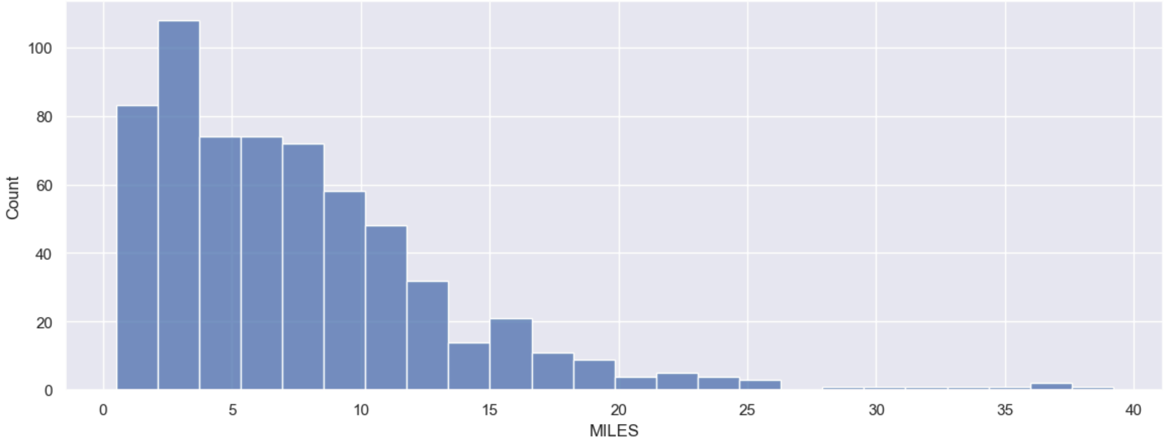
Out[25]:
```
<Axes: xlabel='Month', ylabel='MILES'>
```



Here we can see that **April** month had the ride with the longest distance.

In [26]:
```
sns.histplot(df[df['MILES'] < 40]['MILES'])
```

Out[26]:
```
<Axes: xlabel='MILES', ylabel='Count'>
```



Insights from the above plots:

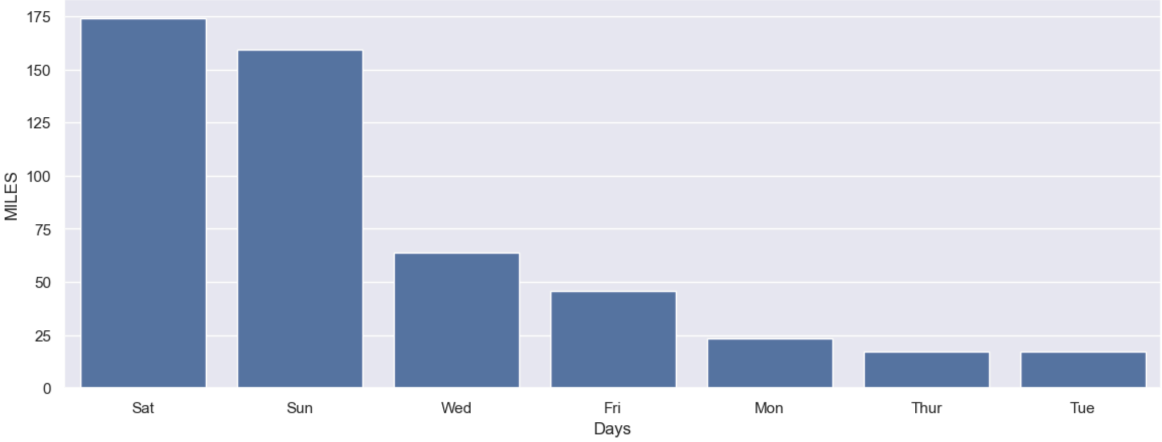
1. Majorly people chooses cabs for the distance of 0-20 miles.
2. For the distance more than 20 miles, cab counts is nearly negligible.

In [27]:
```
df.groupby(['Days'])['MILES'].max()
```

Out[27]:
```
Days
Fri      45.9
Mon      23.3
Sat     174.2
Sun     159.3
Thur     17.3
Tue      17.1
Wed      63.7
Name: MILES, dtype: float64
```

In [28]:
```
day_miles = df.groupby(['Days'], as_index=False)['MILES'].max().sort_values(by='

sns.barplot(x = 'Days', y = 'MILES', data=day_miles)
```

Out[28]:
```
<Axes: xlabel='Days', ylabel='MILES'>
```



Here we can see that cabs are travelled most of the distance on **Weekends**.

Key Insights & Conclusion

- **Peak ride hours** are during Afternoon and Evening.

- **Peak ride days** are during Fridays.

- **Longest Ride** typically happen during the month of April and on the Weekends.

- **Most number of cab** book for the distance of 0-20 Miles, for Business category, and for Meal/Entertainment Purpose.