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

In [3]:
```
dataset = pd.read_csv("Uber Data.csv")
dataset.head()
```

Out[3]:

	START_DATE	END_DATE	CATEGORY	START	STOP	MILES	PURPOSE
0	01 January 2016	01 January 2016	Business	Fort Pierce	Fort Pierce	5.1	Meal/Entertain
1	01 February 2016	01 February 2016	Business	Fort Pierce	Fort Pierce	4.8	Errand/Supplies
2	01 May 2016	01 May 2016	Business	Fort Pierce	Fort Pierce	4.7	Meeting
3	01 June 2016	01 June 2016	Business	Fort Pierce	West Palm Beach	63.7	Customer Visit
4	01 June 2016	01 June 2016	Business	West Palm Beach	West Palm Beach	4.3	Meal/Entertain

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

Out[4]:
```
(653, 7)
```

In [5]:
```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 653 entries, 0 to 652
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   START_DATE  653 non-null    object 
 1   END_DATE    653 non-null    object 
 2   CATEGORY    653 non-null    object 
 3   START       653 non-null    object 
 4   STOP        653 non-null    object 
 5   MILES       653 non-null    float64
 6   PURPOSE     653 non-null    object 
dtypes: float64(1), object(6)
memory usage: 35.8+ KB
```

In [8]:
```
dataset.dropna()
```

Out[8]:

	START_DATE	END_DATE	CATEGORY	START	STOP	MILES	PURPOSE
0	01 January 2016	01 January 2016	Business	Fort Pierce	Fort Pierce	5.1	Meal/Entertain
1	01 February 2016	01 February 2016	Business	Fort Pierce	Fort Pierce	4.8	Errand/Supplies
2	01 May 2016	01 May 2016	Business	Fort Pierce	Fort Pierce	4.7	Meeting
3	01 June 2016	01 June 2016	Business	Fort Pierce	West Palm Beach	63.7	Customer Visit
4	01 June 2016	01 June 2016	Business	West Palm Beach	West Palm Beach	4.3	Meal/Entertain
...	...	...	...	...	...	...	...
648	12/31/2016 1:07	12/31/2016 1:14	Business	Kar?chi	Kar?chi	0.7	Meeting
649	12/31/2016 13:24	12/31/2016 13:42	Business	Kar?chi	Unknown Location	3.9	Temporary Site
650	12/31/2016 15:03	12/31/2016 15:38	Business	Unknown Location	Unknown Location	16.2	Meeting
651	12/31/2016 21:32	12/31/2016 21:50	Business	Katunayake	Gampaha	6.4	Temporary Site
652	12/31/2016 22:08	12/31/2016 23:51	Business	Gampaha	Ilukwatta	48.2	Temporary Site

653 rows × 7 columns

In [9]:
```
print(dataset.to_string())
```

```
           START_DATE           END_DATE  CATEGORY                    
START                         STOP  MILES          PURPOSE                
0      01 January 2016    01 January 2016  Business                  Fort
Pierce                  Fort Pierce    5.1   Meal/Entertain              
1     01 February 2016   01 February 2016  Business                  Fort
Pierce                  Fort Pierce    4.8  Errand/Supplies              
2          01 May 2016        01 May 2016  Business                  Fort
Pierce                  Fort Pierce    4.7         Meeting               
3         01 June 2016       01 June 2016  Business                  Fort
Pierce             West Palm Beach   63.7   Customer Visit              
4         01 June 2016       01 June 2016  Business             West Pal
m Beach             West Palm Beach    4.3   Meal/Entertain              
5         01 June 2016       01 June 2016  Business             West Pal
m Beach                  Palm Beach    7.1         Meeting               
6         01 July 2016       01 July 2016  Business                    
Cary                        Cary    0.8         Meeting               
7      01 October 2016    01 October 2016  Business                    
Cary                 Morrisville    8.3         Meeting               
8      01 October 2016    01 October 2016  Business                    
Jamaica                      New York   16.5   Customer Visit
```

In [10]:
```
dataset['START_DATE'] = pd.to_datetime(dataset['START_DATE'],
                                       errors='coerce')
dataset['END_DATE'] = pd.to_datetime(dataset['END_DATE'],
                                     errors='coerce')
```

In [11]:
```
from datetime import datetime

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

#changing into categories of day and night
dataset['day-night'] = pd.cut(x=dataset['time'],
                              bins = [0,10,15,19,24],
                              labels = ['Morning','Afternoon','Evening','Nigh
```

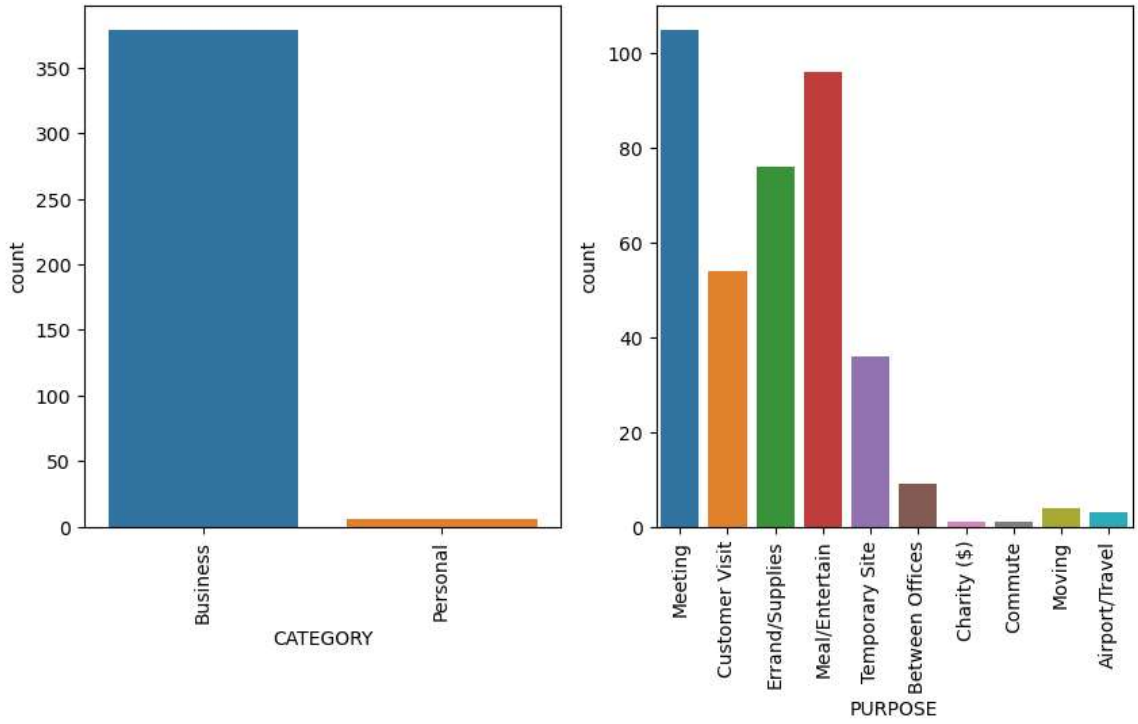
In [12]:
```
dataset.dropna(inplace=True)
```

In [13]:
```
dataset.drop_duplicates(inplace=True)
```

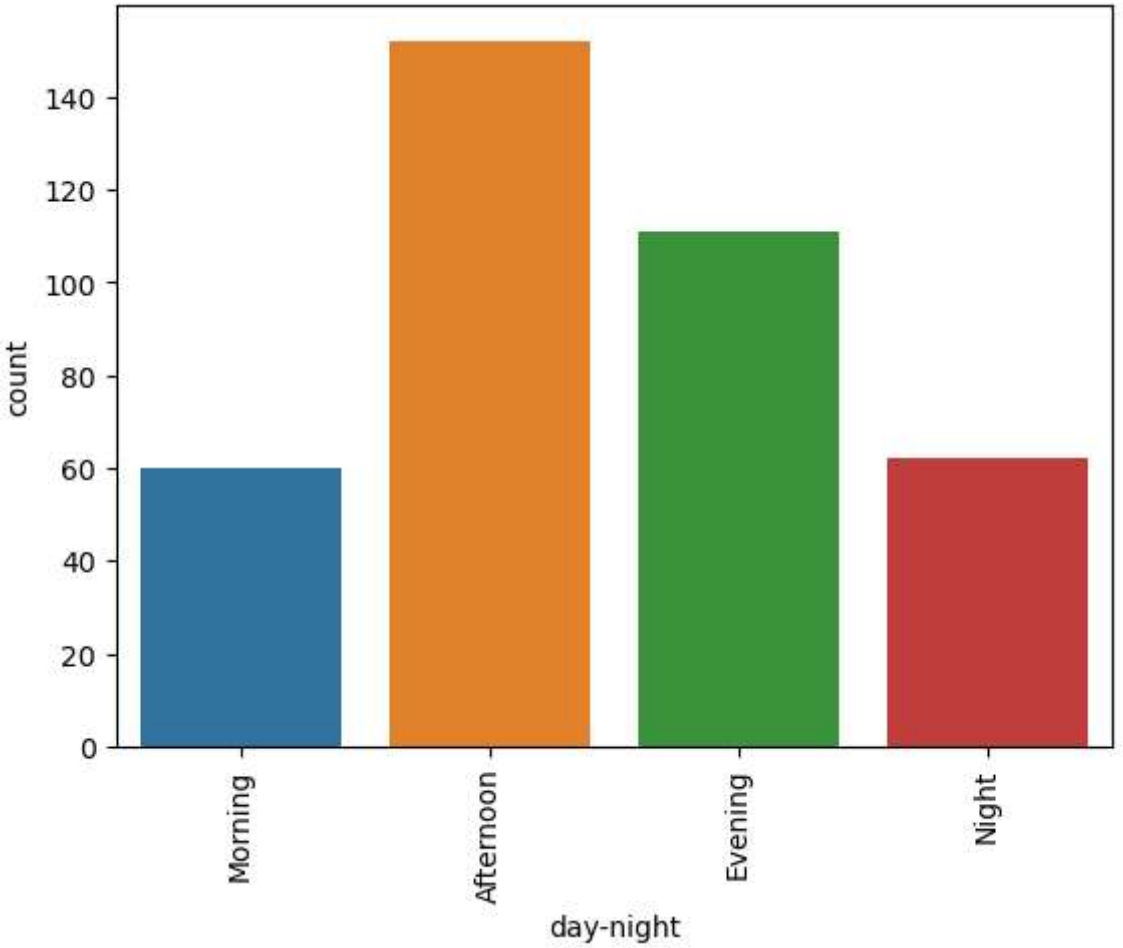
In [15]:
```
obj = (dataset.dtypes == 'object')
object_cols = list(obj[obj].index)

unique_values = {}
for col in object_cols:
  unique_values[col] = dataset[col].unique().size
unique_values
```

Out[15]:
```
{'CATEGORY': 2, 'START': 87, 'STOP': 91, 'PURPOSE': 10, 'date': 130}
```

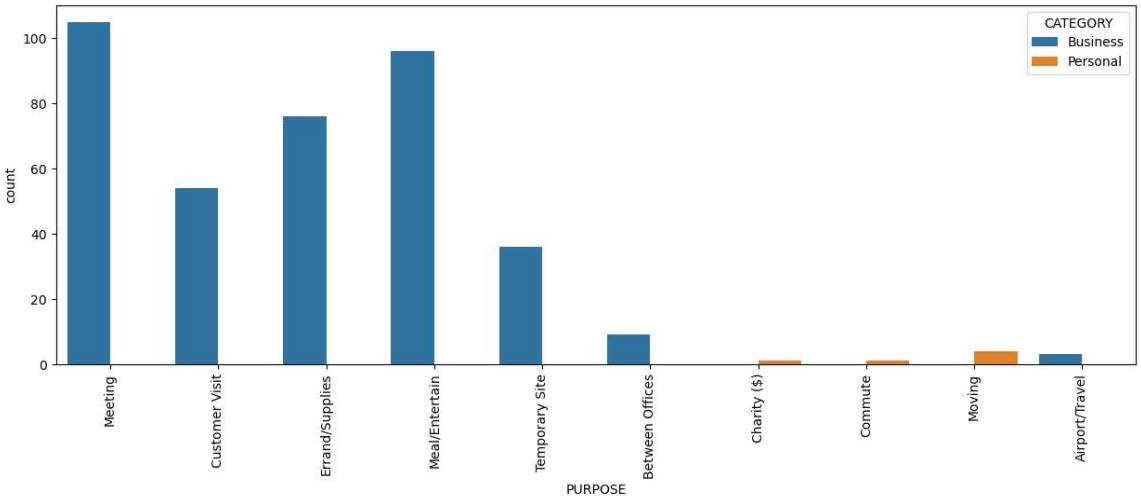
In [17]:
```
plt.figure(figsize=(10, 5))

plt.subplot(1, 2, 1)
sns.countplot(data=dataset, x='CATEGORY')
plt.xticks(rotation=90)

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

plt.show()
```



In [23]:
```
sns.countplot(data=dataset, x='day-night')
plt.xticks(rotation=90)
plt.show()
```

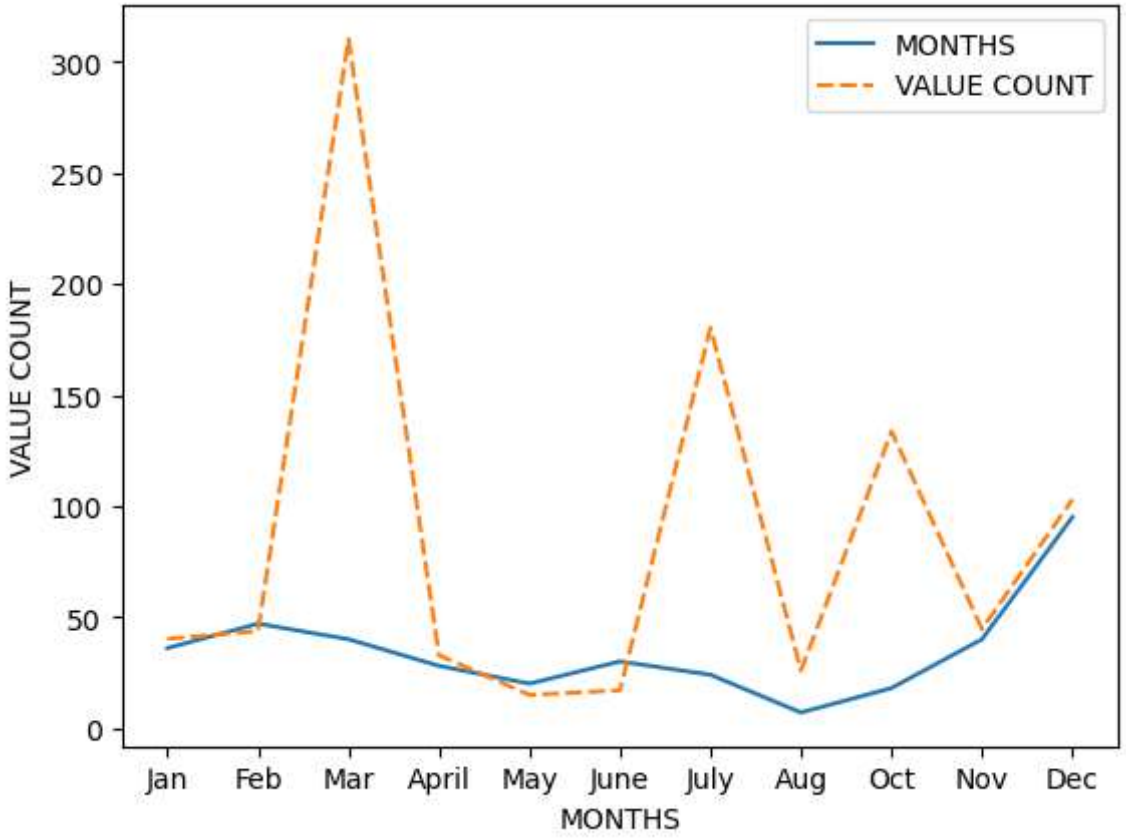


In [24]:
```
plt.figure(figsize=(15, 5))
sns.countplot(data=dataset, x='PURPOSE', hue='CATEGORY')
plt.xticks(rotation=90)
plt.show()
```



In [29]:
```
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)

# Month total rides count vs Month ride max count
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")
```

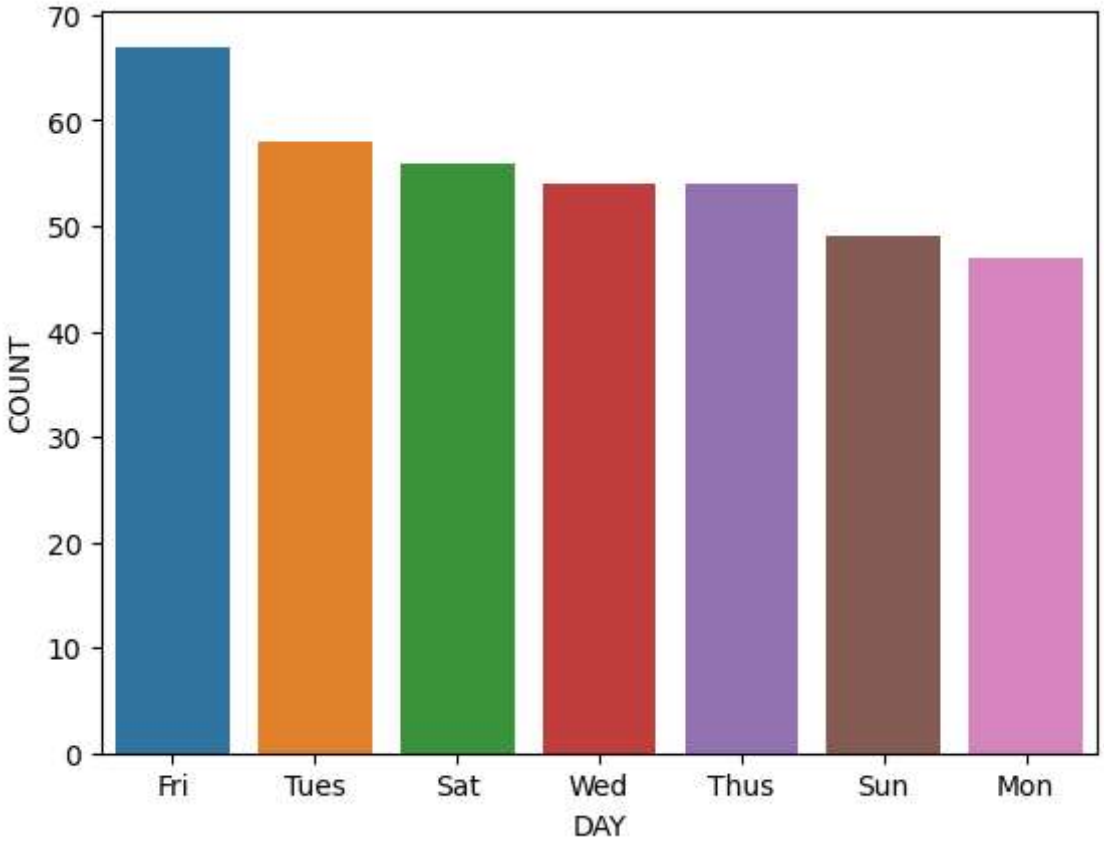
Out[29]:
```
[Text(0.5, 0, 'MONTHS'), Text(0, 0.5, 'VALUE COUNT')]
```



In [30]:
```
dataset['DAY'] = dataset.START_DATE.dt.weekday
day_label = {
    0: 'Mon', 1: 'Tues', 2: 'Wed', 3: 'Thus', 4: 'Fri', 5: 'Sat', 6: 'Sun'
}
dataset['DAY'] = dataset['DAY'].map(day_label)
```

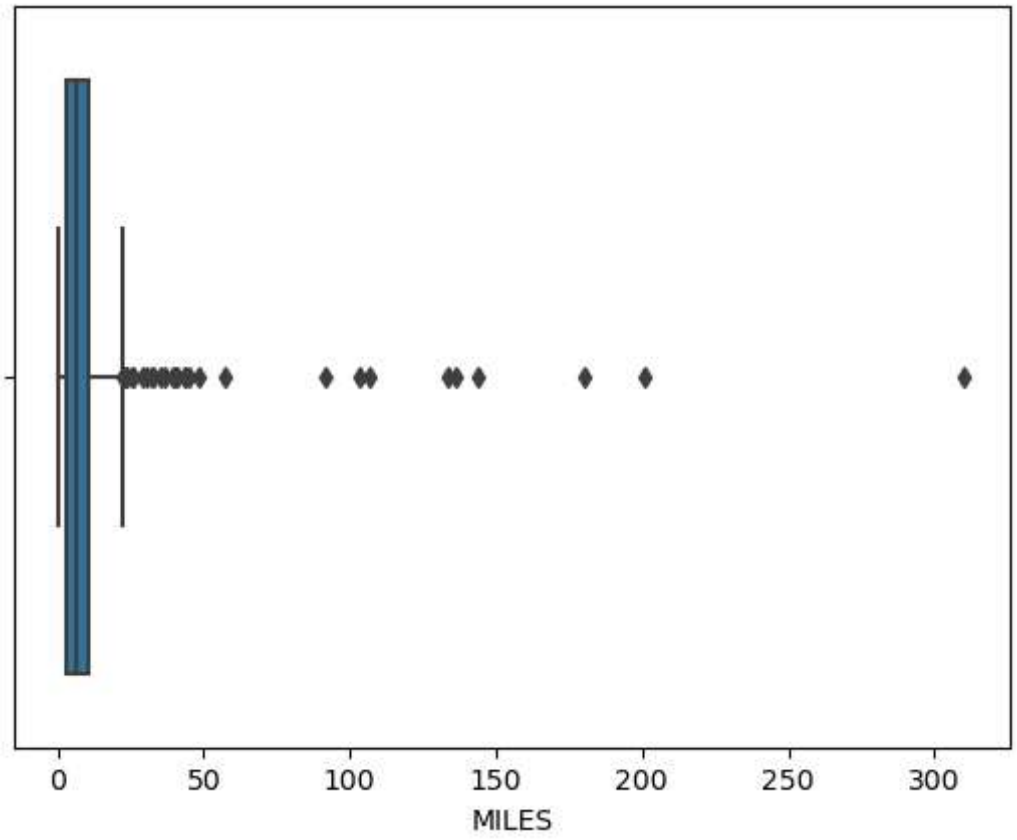
In [31]:
```
day_label = dataset.DAY.value_counts()
sns.barplot(x=day_label.index, y=day_label);
plt.xlabel('DAY')
plt.ylabel('COUNT')
```

Out[31]:
```
Text(0, 0.5, 'COUNT')
```



In [34]:
```
sns.boxplot(data=dataset, x='MILES')
```

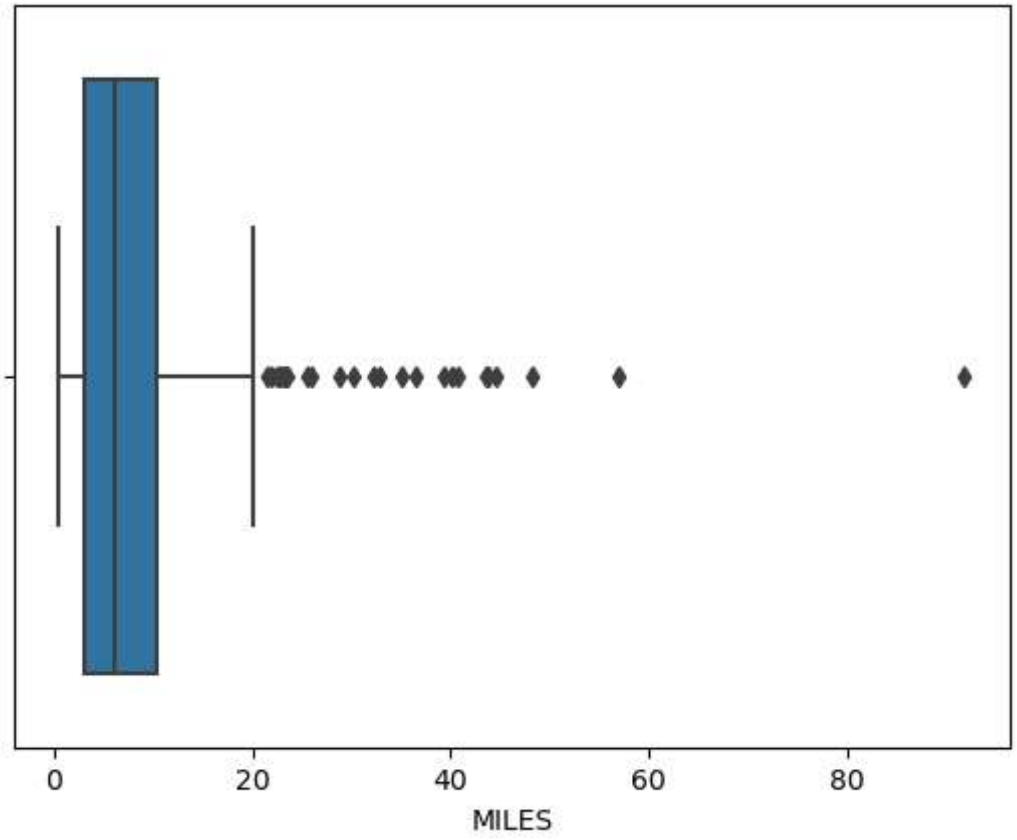
Out[34]:
```
<Axes: xlabel='MILES'>
```



In [44]:
```
# Verify column names
print(dataset.columns)

# Check and convert data type if needed
dataset['MILES'] = pd.to_numeric(dataset['MILES'], errors='coerce')

# Reset the index if necessary
# dataset.reset_index(drop=True, inplace=True)

# Create the boxplot
sns.boxplot(data=dataset[dataset['MILES'] < 100], x='MILES')
```

```
Index(['START_DATE', 'END_DATE', 'CATEGORY', 'START', 'STOP', 'MILES',
       'PURPOSE', 'date', 'time', 'day-night', 'MONTH', 'DAY'],
      dtype='object')
```

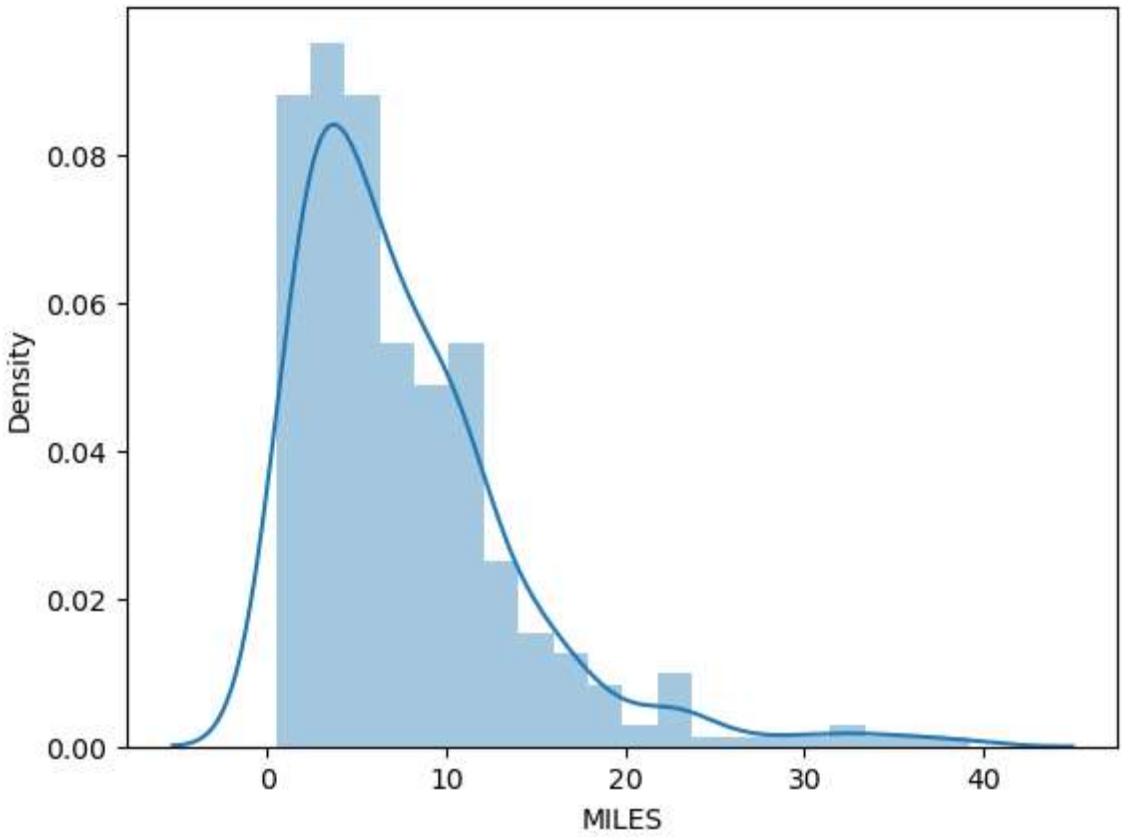
Out[44]: <Axes: xlabel='MILES'>



In [45]:
```
sns.distplot(dataset[dataset['MILES']<40]['MILES'])
```

```
C:\Users\HP\AppData\Local\Temp\ipykernel_11324\615779499.py:1: UserWarnin
g:

`distplot` is a deprecated function and will be removed in seaborn v0.14.
0.

Please adapt your code to use either `displot` (a figure-level function wi
th
similar flexibility) or `histplot` (an axes-level function for histogram
s).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751 (https://
gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751)

  sns.distplot(dataset[dataset['MILES']<40]['MILES'])
```

Out[45]: <Axes: xlabel='MILES', ylabel='Density'>



In [ ]:
```

```