ML LAB 1

Create a generic segregation of any business scenario data into training and testingpart with 70-30% proportions and analyze missing values. Further statistically summarize the data also.

In [1]:
```
import pandas as pd
import numpy as np
df=pd.read_csv('E:/DS/Datasets/daily-bike-share.csv')
```

In [2]:
```
df.columns
```

Out[2]:
```
Index(['day', 'mnth', 'year', 'season', 'holiday', 'weekday', 'workingday',
       'weathersit', 'temp', 'atemp', 'hum', 'windspeed', 'rentals'],
      dtype='object')
```

In [3]:
```
df.shape
```

Out[3]:
```
(731, 13)
```

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

Out[4]:

	day	mnth	year	season	holiday	weekday	workingday	wea
count	731.000000	731.000000	731.000000	731.000000	731.000000	731.000000	731.000000	731.0
mean	15.738714	6.519836	2011.500684	2.496580	0.028728	2.997264	0.683995	1.3
std	8.809949	3.451913	0.500342	1.110807	0.167155	2.004787	0.465233	0.5
min	1.000000	1.000000	2011.000000	1.000000	0.000000	0.000000	0.000000	1.0
25%	8.000000	4.000000	2011.000000	2.000000	0.000000	1.000000	0.000000	1.0
50%	16.000000	7.000000	2012.000000	3.000000	0.000000	3.000000	1.000000	1.0
75%	23.000000	10.000000	2012.000000	3.000000	0.000000	5.000000	1.000000	2.0
max	31.000000	12.000000	2012.000000	4.000000	1.000000	6.000000	1.000000	3.0

In [5]:
```
df.isna().sum()
```

Out[5]:
```
day           0
mnth          0
year          0
season        0
holiday       0
weekday       0
workingday    0
weathersit    0
temp          0
atemp         0
hum           0
windspeed     0
rentals       0
dtype: int64
```

In [6]:
```
from sklearn.model_selection import train_test_split
```

In [7]:
```
training,testing=train_test_split(df,test_size=0.30,random_state=24)
```

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

Out[10]:
```
(511, 13)
```

In [11]:
```
testing.shape
```

Out[11]:
```
(220, 13)
```

In [13]:
```
training.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 511 entries, 338 to 418
Data columns (total 13 columns):
day           511 non-null int64
mnth          511 non-null int64
year          511 non-null int64
season        511 non-null int64
holiday       511 non-null int64
weekday       511 non-null int64
workingday    511 non-null int64
weathersit    511 non-null int64
temp          511 non-null float64
atemp         511 non-null float64
hum           511 non-null float64
windspeed     511 non-null float64
rentals       511 non-null int64
dtypes: float64(4), int64(9)
memory usage: 55.9 KB
```

In [14]:
```
testing.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 220 entries, 307 to 278
Data columns (total 13 columns):
day           220 non-null int64
mnth          220 non-null int64
year          220 non-null int64
season        220 non-null int64
holiday       220 non-null int64
weekday       220 non-null int64
workingday    220 non-null int64
weathersit    220 non-null int64
temp          220 non-null float64
atemp         220 non-null float64
hum           220 non-null float64
windspeed     220 non-null float64
rentals       220 non-null int64
dtypes: float64(4), int64(9)
memory usage: 24.1 KB
```

Interpretation:

The daily bike share data has been statistically described & split into 70%-30% proportion

In [ ]:
```

```