DC Bikeshare

2011-2012 Dataset

Dataset Availability

- Dataset can be found on UCI Machine Learning Repository
- or
- Kaggle [Active Competition]
- Data has daily dataset & hourly dataset

Response Variable

Choice of:

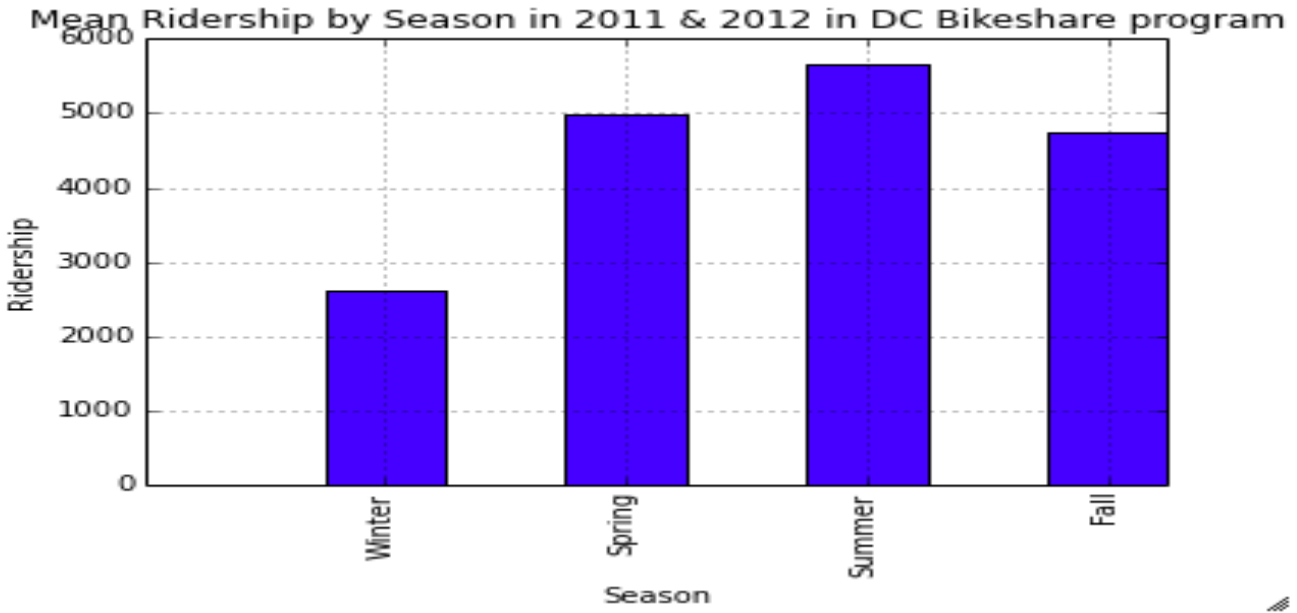
- Casual Ridership
- Registered Ridership

Explanatory Variables

- Season
- Date/Time/Hour
- Holiday/working Day/Weekday
- Weather [1-4 categorical]
- Temperatures
- Humidity
- Windspeed

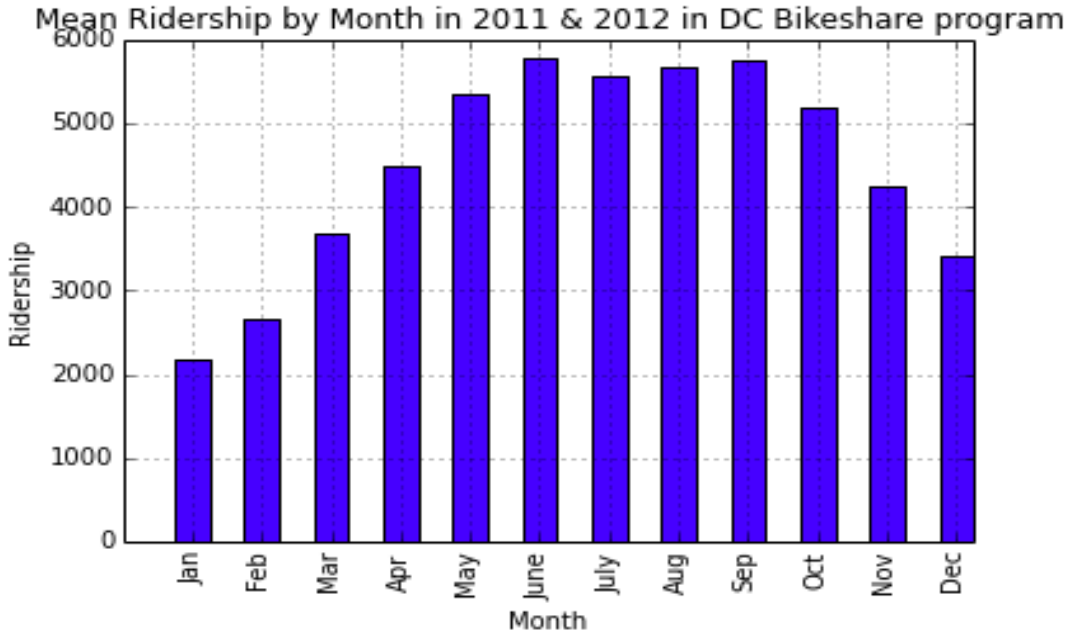
Ridership by Season

Shows summer as peak riding season: make sense



Ridership by Month

Makes sense that Summer months are highest



Weathersit Categorical Variable [1-4]

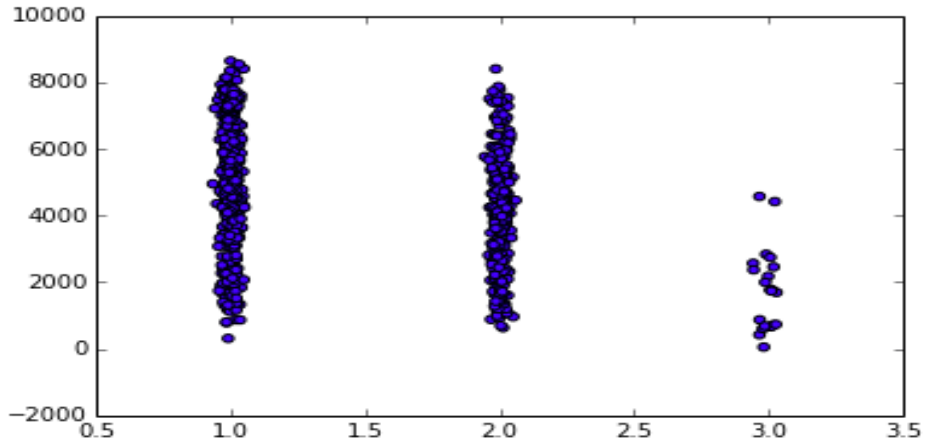
- 1: Clear, Few clouds, Partly cloudy, Partly cloudy
- 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
- 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
- 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

WeatherSit Categorical Variable

Shows no riding when weather at level 4

```
jitter(bike4['weathersit'], bike4['cnt'])
```

```
<matplotlib.collections.PathCollection at 0x10d86f250>
```



Explanatory Variable: Temperature

count 731.000000
mean 0.495385
std 0.183051
min 0.059130
25% 0.337083
50% 0.498333
75% 0.655417
max 0.861667
dtype: float64

Explanatory Variable: Temperature

Covariance between Total Daily Ridership vs. Daily Average temperature was:

np.cov(bike4['temp'], bike4['cnt'])[0][1]:

222.51470045305516 [what does this mean?]

Scatter of temp vs. ridership

Linear relationship with temperature: this is something that can be modelled with regression

```
In [125]: plt.scatter(bike4['temp'], bike4['cnt'])
Out[125]: <matplotlib.collections.PathCollection at 0x10c7a6610>
```