Exploratory Data Analysis

Understanding your data

Contents

- Exploratory Data Analysis
  - Understanding data!
- Interactive notebooks
  - Jupyter
- Data loading & processing
  - Pandas
- Data Visualization
  - Seaborn

Exploratory Data Analysis

- Understanding data
- Spotting patterns & relations
- Machine Learning “Preconditions”

Jupyter notebooks

- Interactive (python) shell
- Code, story & charts

This is a markdown cell used for documenation

The above cell was written as: "### This is a markdown cell", followed by Shift+Enter

In [12]:
```
import math
print "This is a code cell... and Pi is = ", math.pi
```

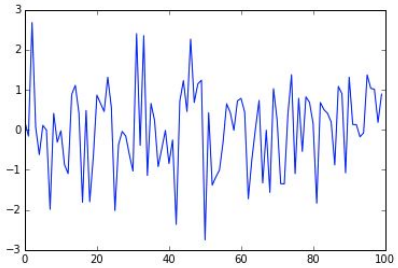
```
This is a code cell... and Pi is =  3.14159265359
```

In [13]:
```
# to enable inline graphs, etc.
%pylab inline

plot(randn(100))
```

```
Populating the interactive namespace from numpy and matplotlib
```

Out[13]:
```
[<matplotlib.lines.Line2D at 0x7fe2a2496cd0>]
```



Pandas - Data loading

In [1]:
```
import pandas as pd

from matplotlib import pyplot as plt
%matplotlib inline
```

First we load the weather data:

In [2]:
```
df_weather = pd.read_csv('weather.csv', index_col='Date', parse_dates=['Date'])
df_weather.head()
```

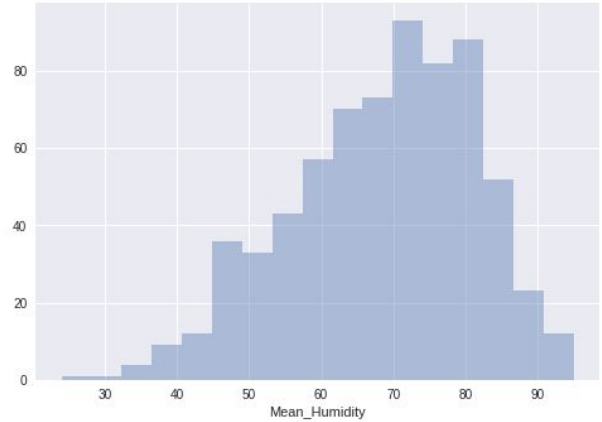
Out[2]:

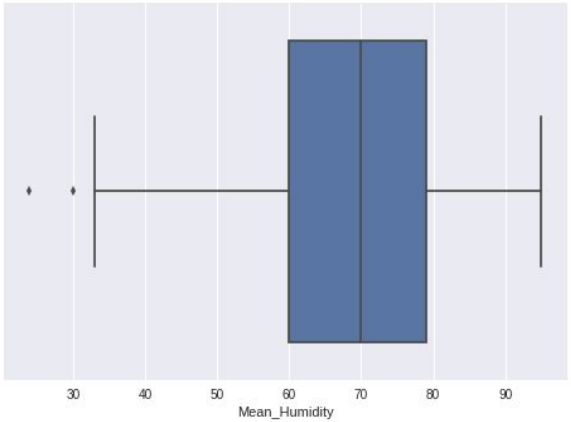
	Max_Temperature_F	Mean_Temperature_F	Min_TemperatureF	Max_Dew_Point_F	MeanDew_Point_F	Mi
Date						
2014-10-13	71	62.0	54	55	51	46
2014-10-14	63	59.0	55	52	51	50
2014-10-15	62	58.0	54	53	50	46
2014-10-16	71	61.0	52	49	46	42
2014-10-17	64	60.0	57	55	51	41

Pandas - Data processing

Seaborn - Data Visualization

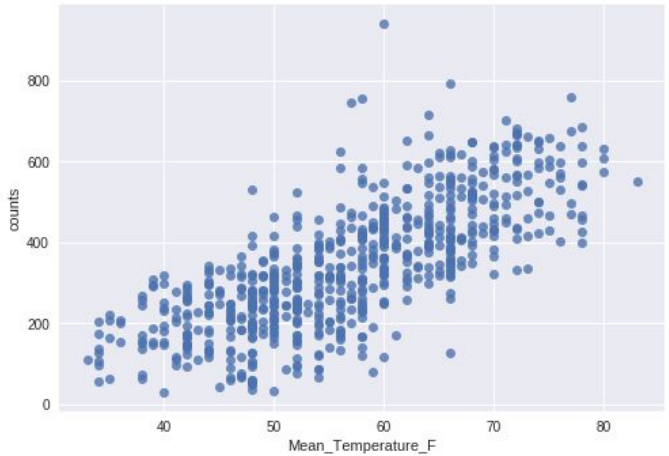
- Plot one variable
  - Histogram & boxplot

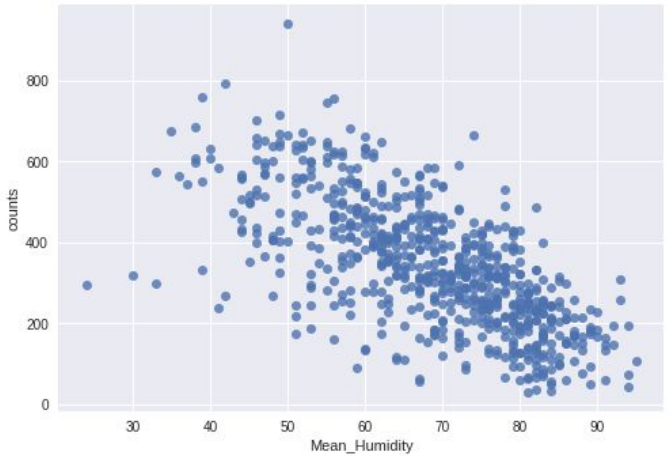




Seaborn - Data Visualization

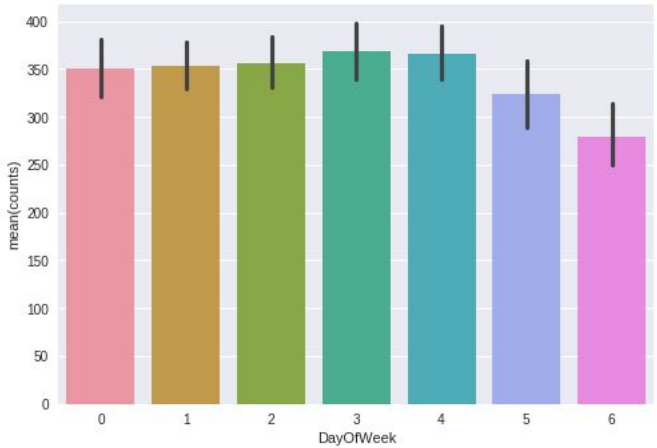
- Plot relations between (numerical) variables
  - Scatterplot

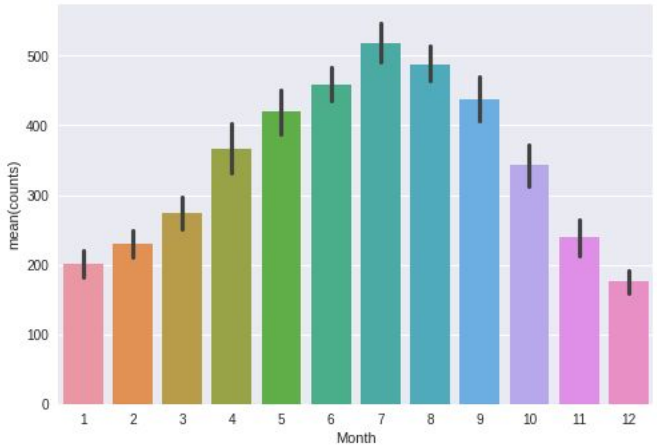




Seaborn - Data Visualization

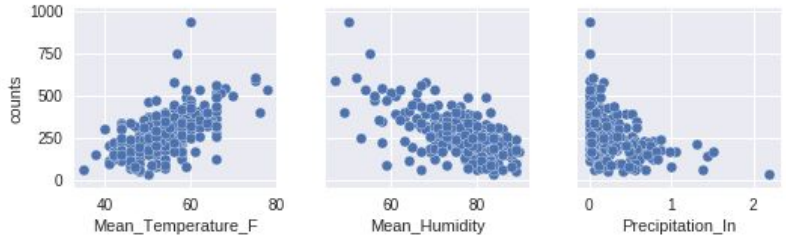
- Plot relations between (categorical) variables
  - Barplot & boxplot

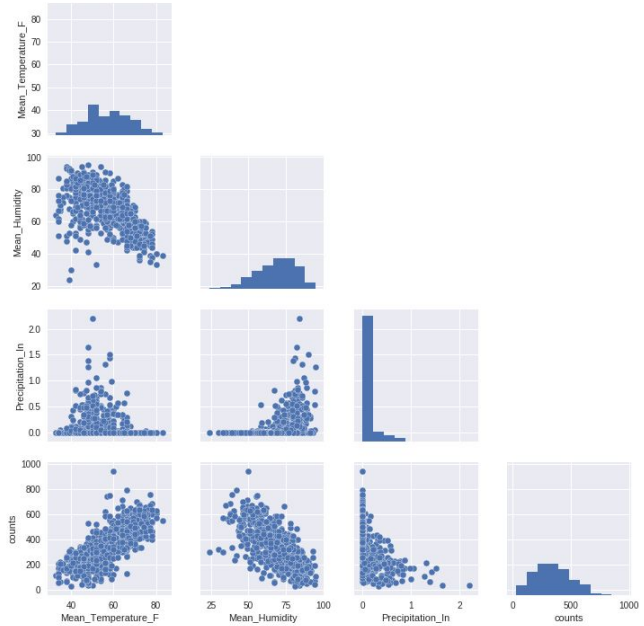




Seaborn - Data Visualization

- Pairplot - plot many relations





Example

- Data Analysis Titanic Survivors

Questions ?

- Let’s start hacking !