**[Assignment06: Python dataset exploration](https://mymasonportal.gmu.edu/webapps/assignment/uploadAssignment?content_id=_8335369_1&course_id=_347779_1&group_id=&mode=view)-Code**

**Part 1: Dataset Exploration**

import pandas as pd

import matplotlib.pyplot as plt

import os

os.getcwd()

os.chdir('C:/Users/Abhishek/Desktop/AIT 580/Assignments/Assignment 06')

hurridate = pd.read\_csv("hurricanes.csv", sep="|")

hurridate

list(hurridate)

1. **Summary Statistics:**

**Month:**

hurridate['Month'].describe()

**Highest Category:**

hurridate['Highest\_Category'].describe()

**Central pressure mb:**

hurridate['Central\_Pressure\_mb'].describe()

**Max Winds KT**:

hurridate['Max\_Winds\_kt'].describe()

hurridate['Highest\_Category'].head()

hurridate.head()

hurridate.tail()

hurridate.info()

hurridate.groupby('Month')

**Visualizations:**

**Month:**

plt.style.use('ggplot')

hurridate.groupby(['Month']).Month.count().plot.bar(legend=True)

plt.show()

**Highest Category:**

plt.style.use('ggplot')

hurridate.groupby(['Highest\_Category']).Highest\_Category.count().plot.bar(legend=True)

plt.show()

**Central pressure mb:**

hurridate['Central\_Pressure\_mb'].plot(kind='line', legend=True)

**Max Winds KT**:

hurridate['Max\_Winds\_kt'].plot(kind='line', legend=True)

1. **Relationship:**

plt.scatter(hurridate['Highest\_Category'],hurridate['Max\_Winds\_kt'])

plt.scatter(hurridate['Central\_Pressure\_mb'],hurridate['Max\_Winds\_kt'])

1. **Missing Data:**

Missing data are very crucial in a dataset and needs to be properly handled as they impact the insights that can be perceived from the model. From our dataset we can see that Central\_Pressure\_mb, Max\_Winds\_kt and Name has missing values which can either be removed or can be imputed. Since, we have a large number almost 1/3rd of the values missing for Max\_Winds\_kt it is better to replace(impute) with either mean, median or the mode. Also, as only 4 Central\_Pressure\_mb values are missing we can ignore those rows. Deleting or imputing large number of rows would definitely impact the output of the dataset and the dataset can be biased.

hurridate['Central\_Pressure\_mb'].fillna(value=hurridate['Central\_Pressure\_mb'].mean())

hurridate.dropna()

pd.isnull(hurridate).any()

hurridate.info()

**Part 2: Web/HTML Screen Scraping**

**(a)Extracting and creating csv dataset:**

import urllib

from bs4 import BeautifulSoup

import requests

import pandas as pd

filename = "top500list1.csv"

f = open(filename, "w")

headers = "Rank, Site, System, Cores, Rmax (TFlop/s), Rpeak (TFlop/s), Power (kW)\n"

f.write(headers)

for i in range(1,6):

r = "https://www.top500.org/list/2018/06/?page={}".format(i)

url=requests.get(r)

soup = BeautifulSoup(url.content, 'html.parser')

for record in soup.findAll('tr'):

tbltxt = ""

for data in record.findAll('td'):

tbltxt = tbltxt.replace("\n","") + data.text.replace(",","") + ","

print(tbltxt)

f.write(tbltxt[0:-1]+"\n")

f.close()

import pandas as pd

import matplotlib.pyplot as plt

import os

os.getcwd()

os.chdir('C:/Users/Abhishek/Desktop/')

toplist = pd.read\_csv("top500list1.csv",encoding='latin1')

toplist

list(toplist)

**Summary statistics for Cores, RMax, RPeak, and Power**

toplist[' Cores'].describe()

toplist[' Rmax (TFlop/s)'].describe()

toplist[' Rpeak (TFlop/s)'].describe()

toplist[' Power (kW)'].describe()

toplist.head()

toplist.tail()

toplist.info()

**Display and Explaining Relationship:**

plt.scatter(toplist[' Cores'],toplist[' Rpeak (TFlop/s)'])

plt.scatter(toplist[' Cores'],toplist[' Power (kW)'])

1. **Summary Statistics for Country Data:**

os.getcwd()

os.chdir('C:/Users/Abhishek/Desktop/')

list1 = pd.read\_csv("top500list1.csv", encoding='latin1')

list1['Country'].describe()

plt.style.use('ggplot')

list1.groupby(['Country']).Country.count().plot.bar(legend=True)

plt.show()

list1['Country'].describe()