

## Contents

Date Preparation and Exploration using python.....	1
Reading a Titanic dataset from a CSV file .....	1
2. Detecting missing values.....	2
3.Imputing missing values.....	3
4. Exploring and visualizing data .....	4
Bar Plot which shows how many passengers survived and how many perished .....	4
Bar Plot which shows how many passesngers travelled by first class, second class and thrid class.....	4
e) Stacked barplot's to find out.....	6
Visualization:.....	8
Descriptive and Inferential Statistics: .....	9

## Date Preparation and Exploration using python

Reading a Titanic dataset from a CSV file

```
import os
```

```
import pandas as pd
```

```
from pandas import DataFrame as df
```

```
os.getcwd()
```

```
filename = 'titanic_data_1.csv'
```

```
data = pd.read_csv(filename)
```

```
print (data.head)
```

1

```
In [24]: data[0]
```

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...)	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Arhem)	female	14.0	1	0	237736	30.0700	NaN	C

```
In [25]: data['Age']
```

```
Out[25]: 0    22.0
```

## 2. Detecting missing values

Missing values reduce the representativeness of the sample, and furthermore, might distort inferences about the population.

How many missing values are there in age Attribute of Titanic dataset



titanic data.csv

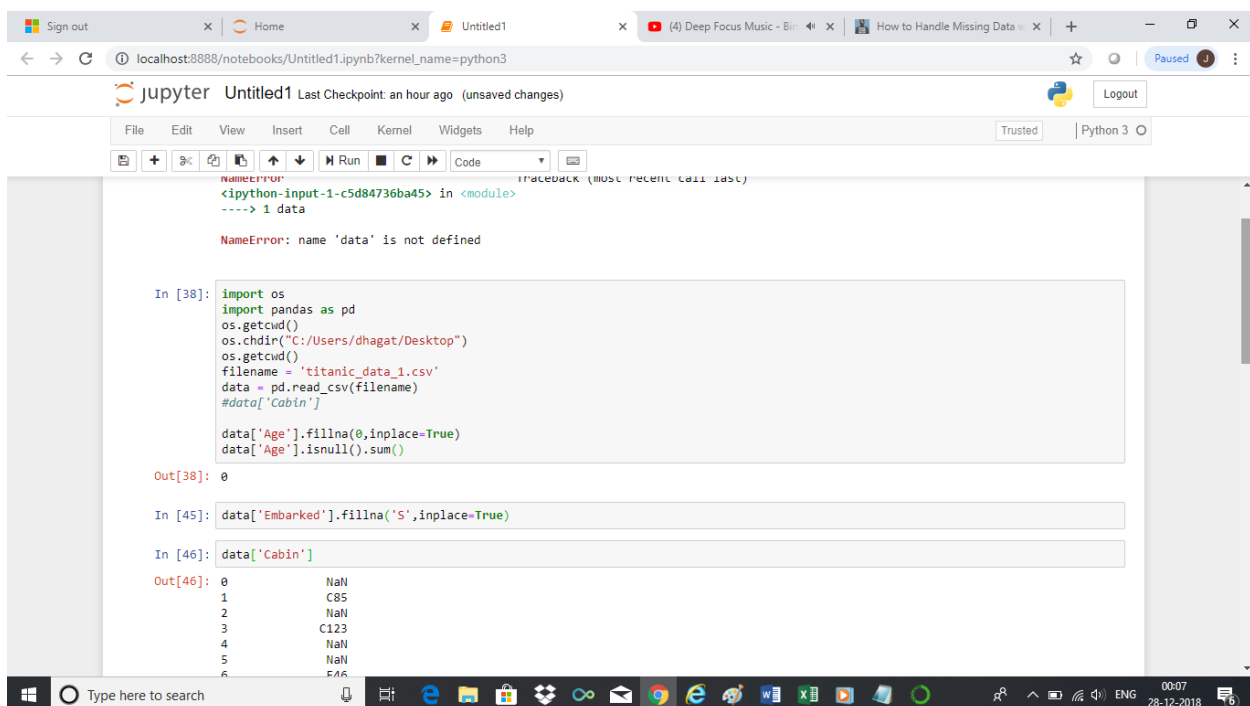
```
In [38]: data.isna().sum()
```

```
Out[38]: PassengerId      0
Survived      0
Pclass        0
Name          0
Sex           0
Age          177
SibSp         0
Parch         0
Ticket        0
Fare          0
Cabin        687
Embarked      2
dtype: int64
```

### 3.Imputing missing values

After detecting the number of missing values within each attribute, we have to impute the missing values since they might have a significant effect on the conclusions that can be drawn from the data.

Assign the missing value to the most likely port, which is Southampton



The screenshot shows a Jupyter Notebook window with the following content:

```
NAMEERROR
Traceback (most recent call last)
<ipython-input-1-c5d84736ba45> in <module>
----> 1 data

NameError: name 'data' is not defined
```

```
In [38]: import os
import pandas as pd
os.getcwd()
os.chdir("C:/Users/dhagat/Desktop")
os.getcwd()
filename = 'titanic_data_1.csv'
data = pd.read_csv(filename)
#data['Cabin']

data['Age'].fillna(0,inplace=True)
data['Age'].isnull().sum()

Out[38]: 0

In [45]: data['Embarked'].fillna('S',inplace=True)

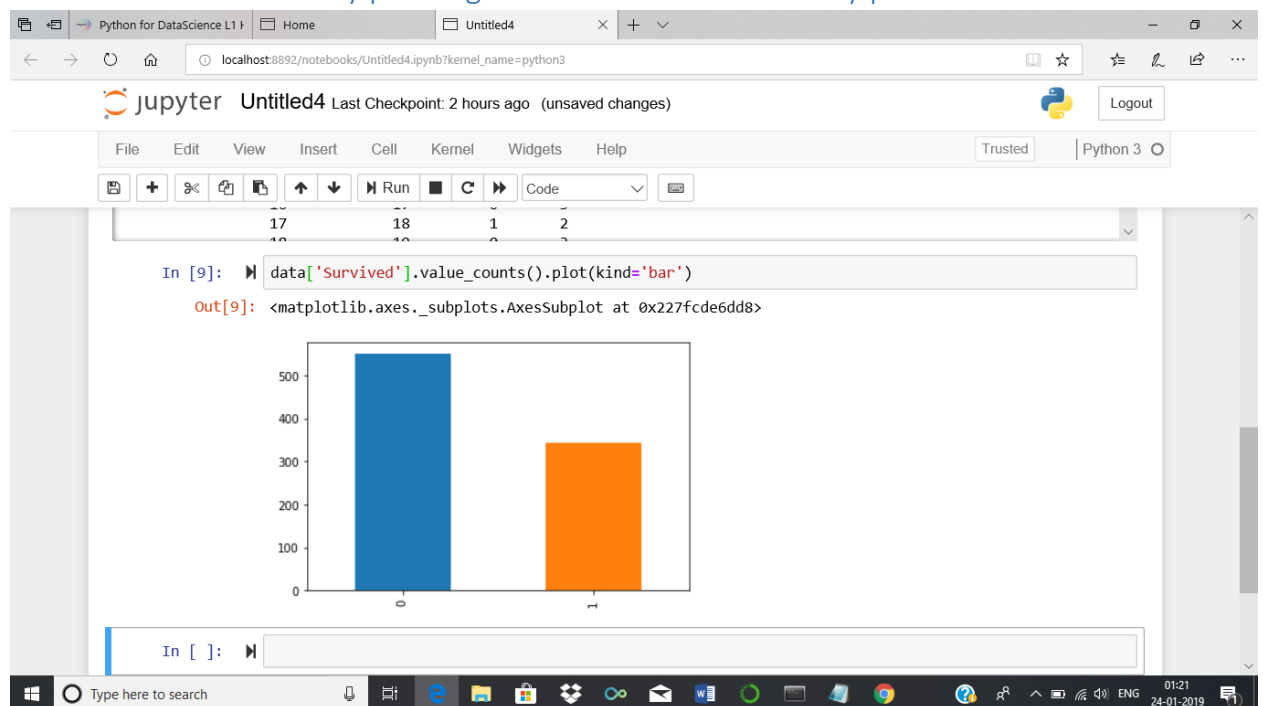
In [46]: data['Cabin']

Out[46]: 0      NaN
1      C85
2      NaN
3     C123
4      NaN
5      NaN
6     NaN
```

## 4. Exploring and visualizing data

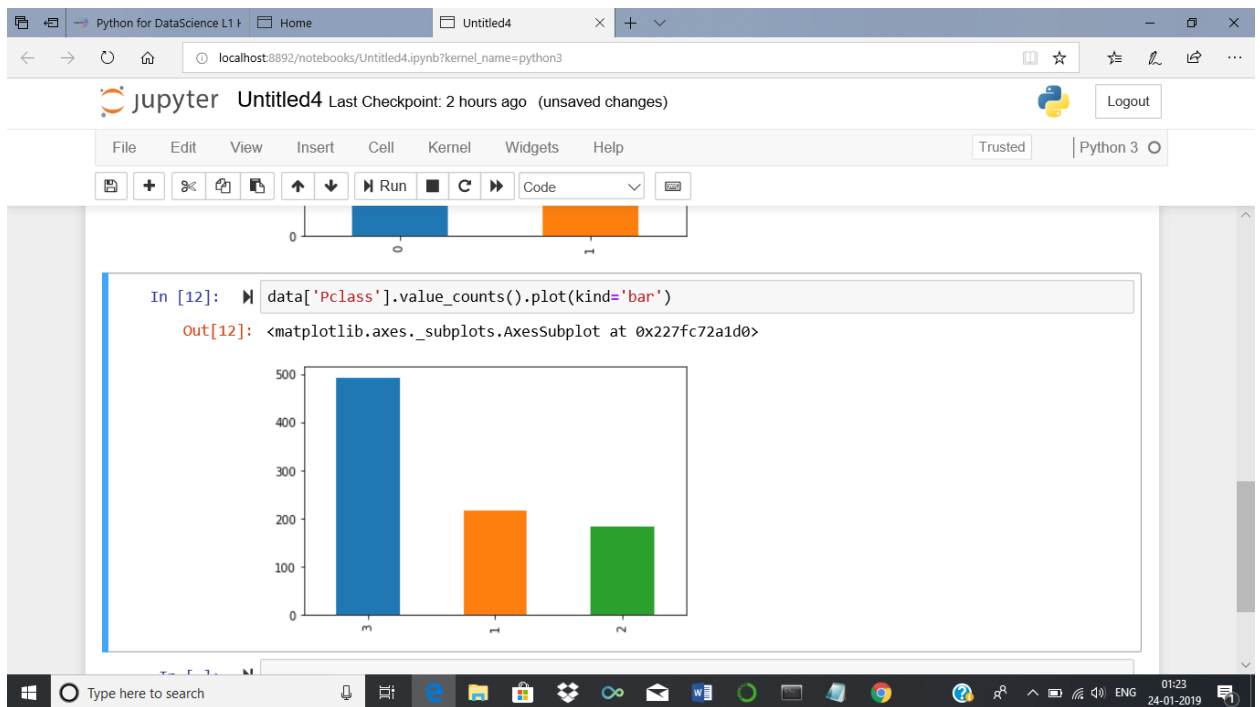
After imputing the missing values, one should perform an exploratory analysis, which involves using a visualization plot and an aggregation method to summarize the data characteristics. The result helps the user gain a better understanding of the data in use.

Bar Plot which shows how many passengers survived and how many perished

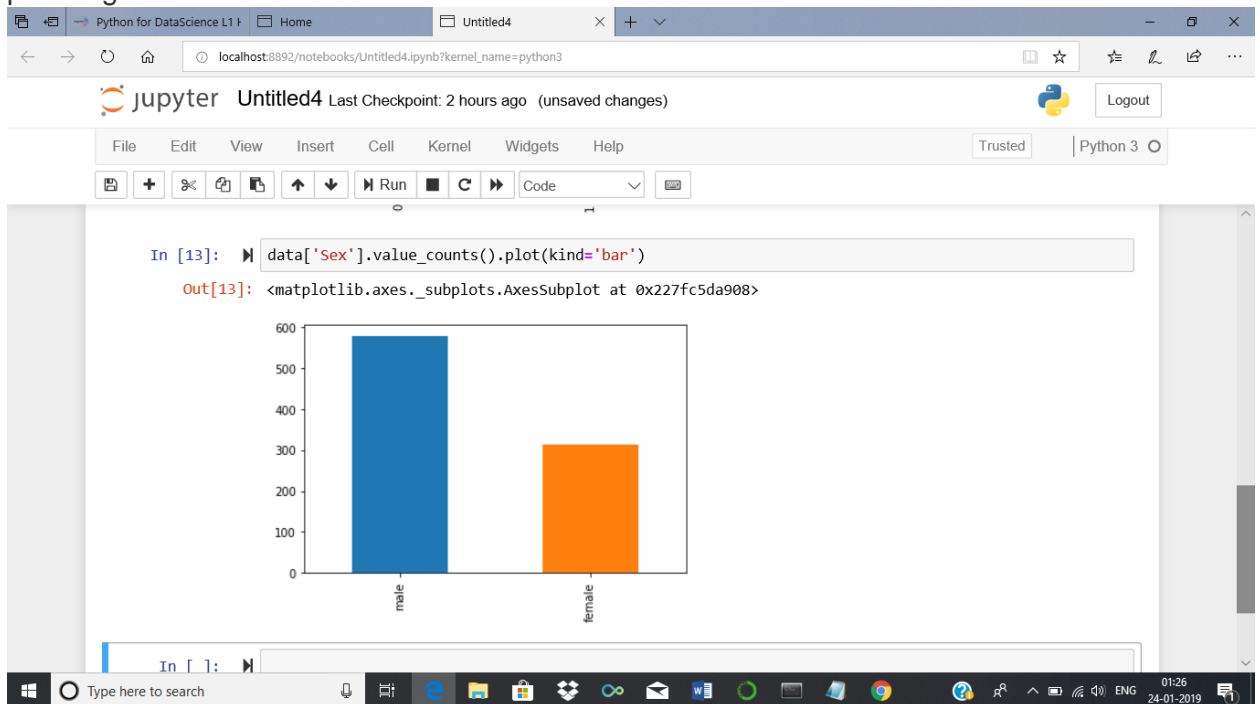


a)

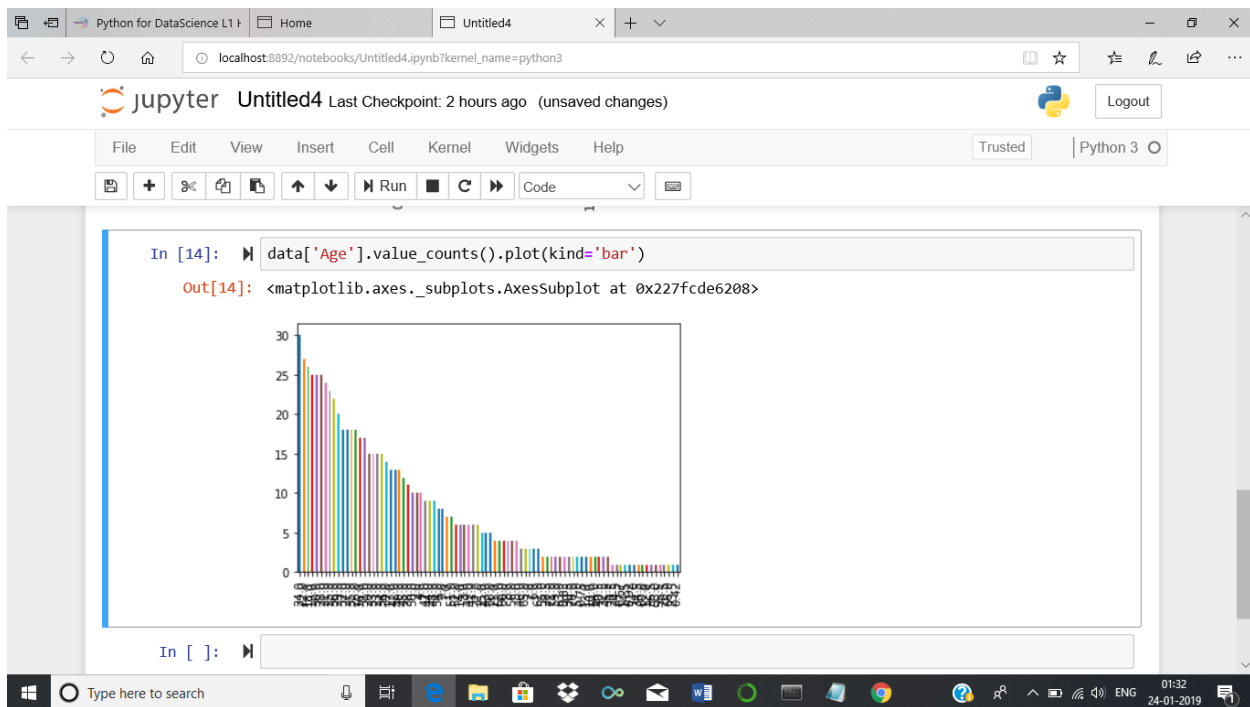
Bar Plot which shows how many passengers travelled by first class, second class and third class



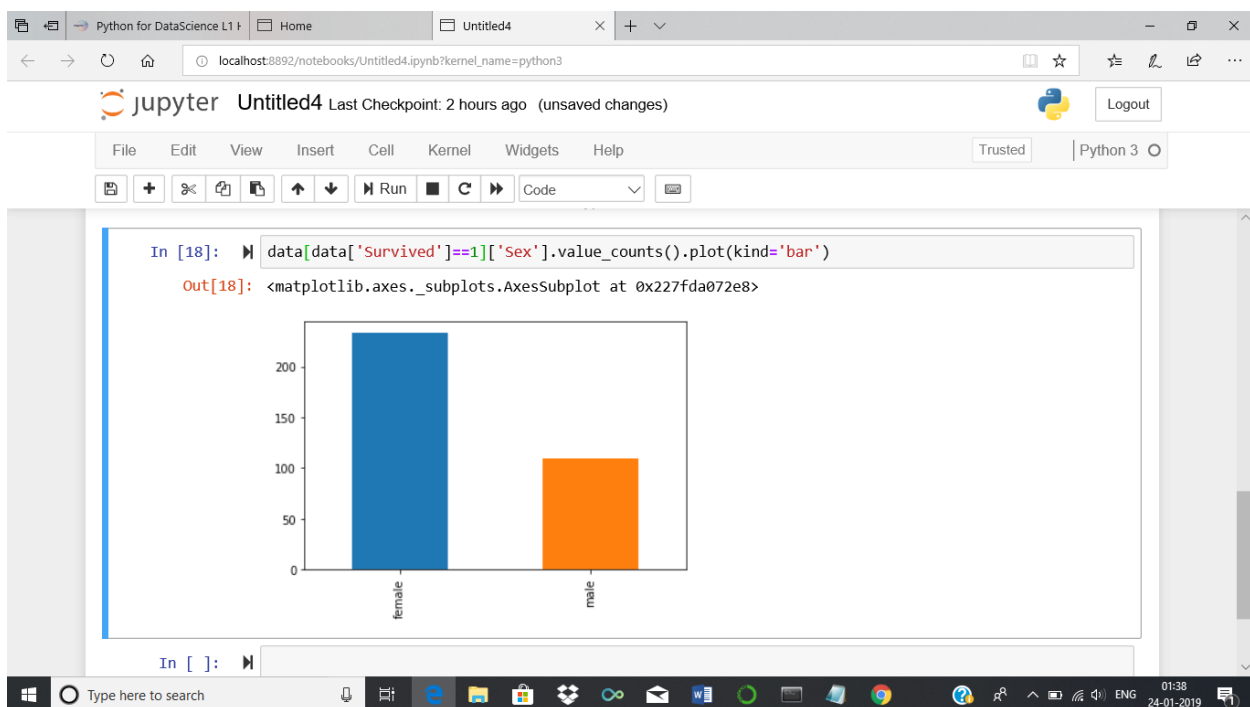
b) Barplot which shows how many are male passengers and how many are female passengers



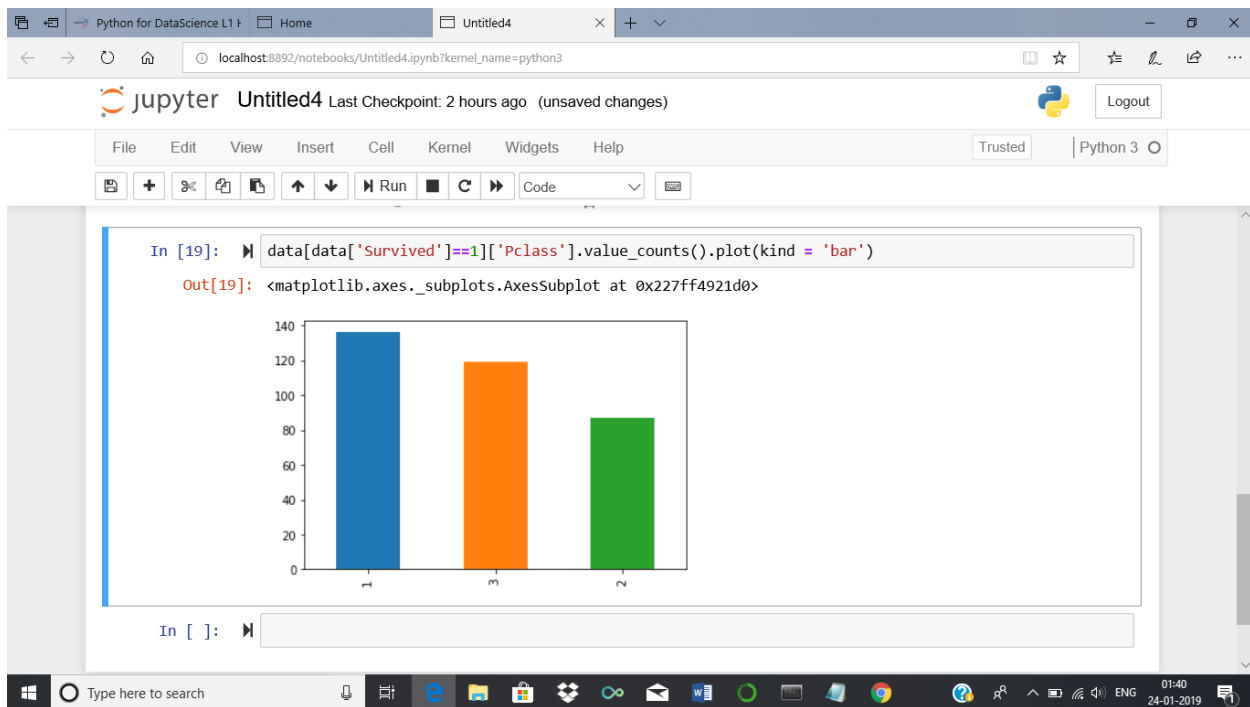
c) plot histogram of different ages



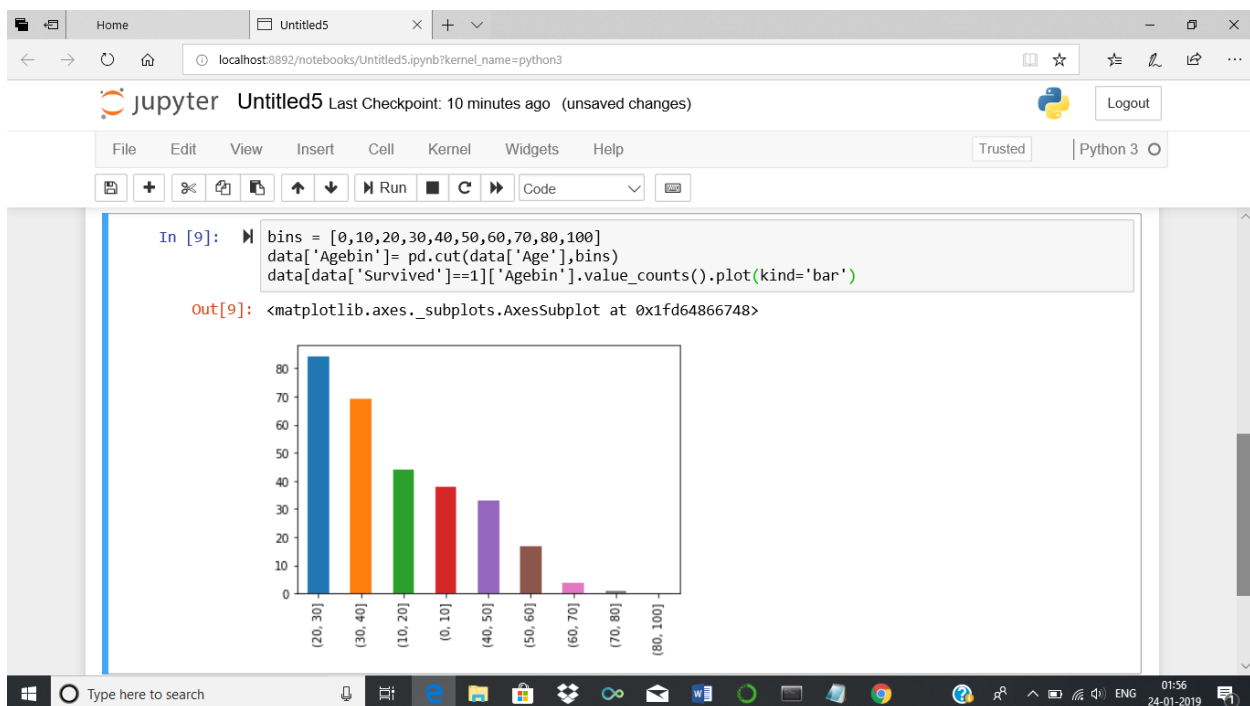
e) Stacked barplot's to find out  
-which gender is more likely to perish during shipwrecks



-Passenger survival by class



d) Box Plot which shows passenger survival by age



## Visualization:

5. Read from sample superstore xl file into pandas dataframe and perform below operations

- a) Display Subcategory wise sum of profit
- b) Exclude Office Furniture SubCategory
- c) Sort SubCategory in Desc order
- d) Categorywise sum of profit in pie chart
- e) Line Chart yearwise sum of profit
- f) Display Top 10 most profitable customers
- g) scatter plot between profit and sales

6. Create dept dataframe and emp dataframe with suitable data and perform inner , leftouter,RightOuter and FullOuter Joins based on common column Deptno.

Dept Data Frame will have

Deptno

Dname

Loc

Emp DataFrame will have below columns

Deptno

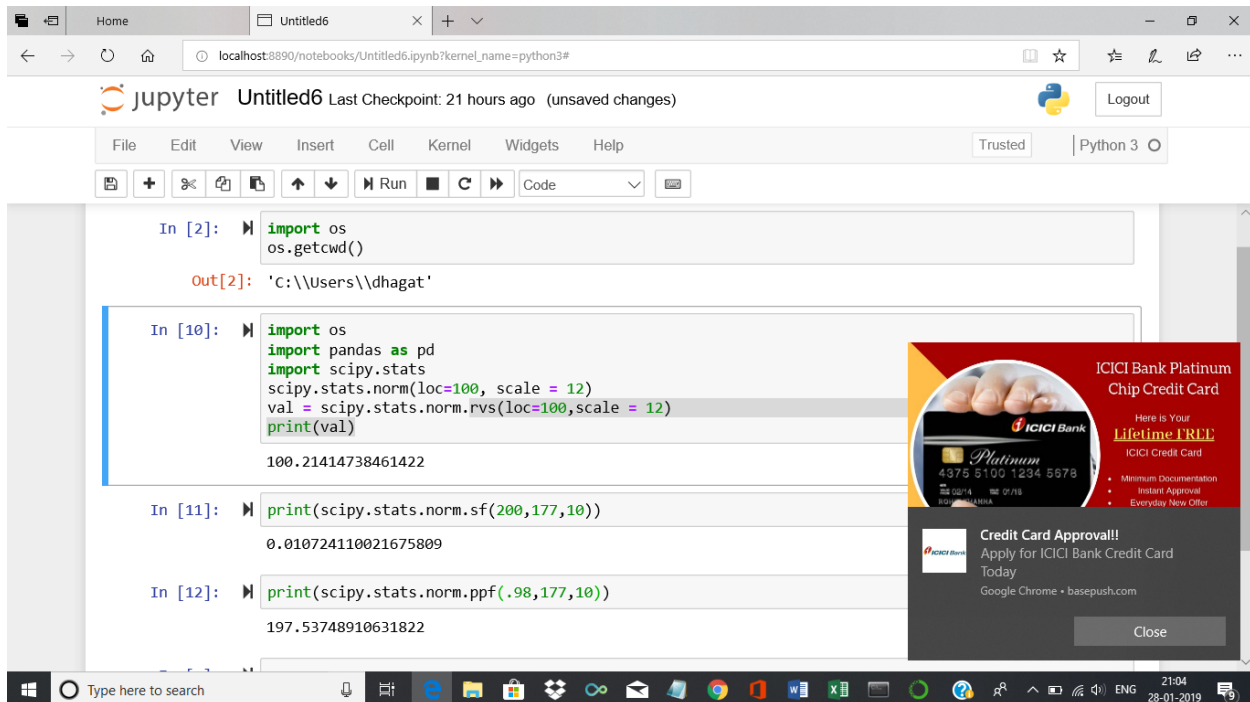
Eno

Sal



## Descriptive and Inferential Statistics:

7. Suppose the height of men in the United Kingdom is known to be normally distributed with a mean of 177 centimeters and a standard deviation of 10 centimeters. If you were to select a man from the United Kingdom population at random, what is the probability that he would be more than 200 centimeters tall?



```
In [2]: import os
os.getcwd()

Out[2]: 'C:\\Users\\dhagat'

In [10]: import os
import pandas as pd
import scipy.stats
scipy.stats.norm(loc=100, scale = 12)
val = scipy.stats.norm.rvs(loc=100, scale = 12)
print(val)

100.21414738461422

In [11]: print(scipy.stats.norm.sf(200,177,10))

0.010724110021675809

In [12]: print(scipy.stats.norm.pdf(.98,177,10))

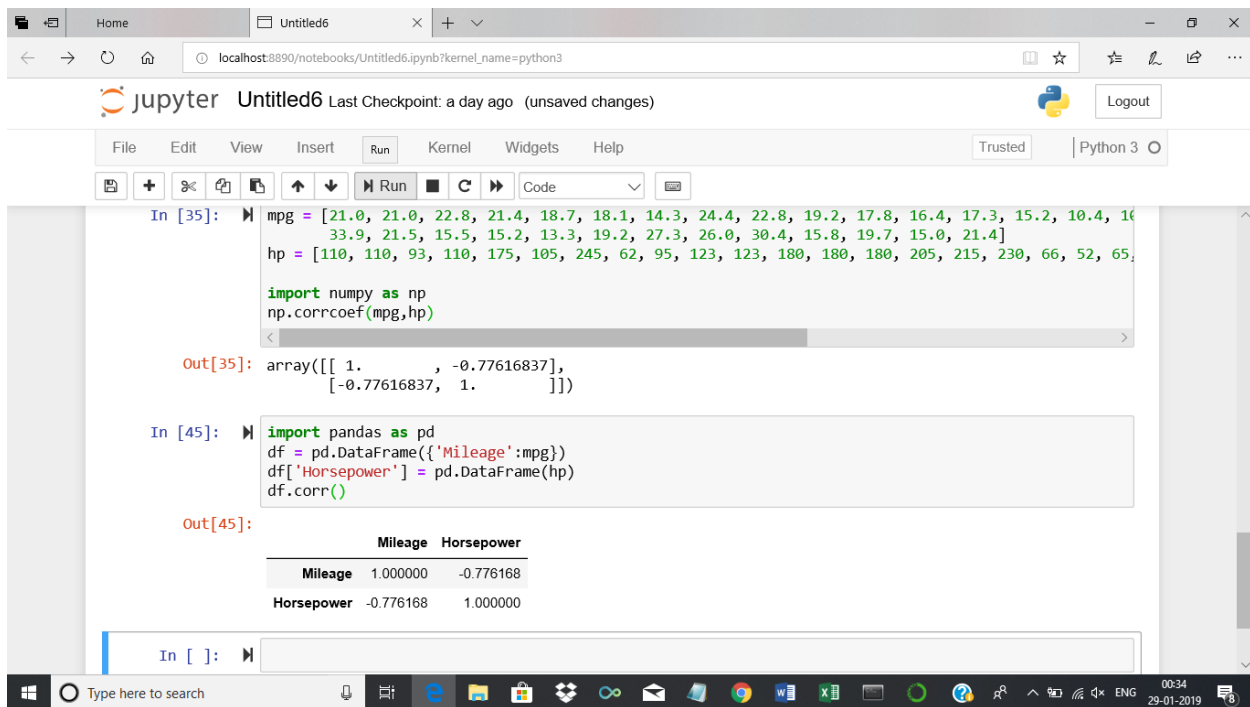
197.53748910631822
```

8. Let's take the mileage and horsepower of various cars and see if there is a relation between the two.

mpg = [21.0, 21.0, 22.8, 21.4, 18.7, 18.1, 14.3, 24.4, 22.8, 19.2, 17.8, 16.4, 17.3, 15.2, 10.4, 10.4, 14.7, 32.4, 30.4,

33.9, 21.5, 15.5, 15.2, 13.3, 19.2, 27.3, 26.0, 30.4, 15.8, 19.7, 15.0, 21.4]

hp = [110, 110, 93, 110, 175, 105, 245, 62, 95, 123, 123, 180, 180, 180, 205, 215, 230, 66, 52, 65, 97, 150, 150, 245, 175, 66, 91, 113, 264, 175, 335, 109]



9. Perform T-test on two classes that are given a mathematics test and have 10 students in each class. Determine if 2 distributions are identical or not.

```
class1_score = np.array([45.0, 40.0, 49.0, 52.0, 54.0, 64.0, 36.0, 41.0, 42.0, 34.0])
```

```
class2_score = np.array([75.0, 85.0, 53.0, 70.0, 72.0, 93.0, 61.0, 65.0, 65.0, 72.0])
```

10. The mean score of the mathematics exam at a national level is 60 marks and the standard deviation is 3 marks. The mean marks of a class are 53. The null hypothesis is that the mean marks of the class are similar to the national average. Test this Hypothesis using Z – Test.

11. Calculate Pearson correlation coefficient between Girth and Volume in trees dataset( trees



trees.csv

csv file) Pl mention what you draw from correlation coefficient

12. Suppose that you want to perform a hypothesis test to help determine whether the correlation between tree girth and tree volume is statistically significant.

Perform a two-sided test of the Pearson's product moment correlation between tree girth and volume at the 5% significance level,

## Machine Learning Algorithms



Automobile price  
data\_Raw.csv

### 12. Use Automobile price data Raw csv file :

- Split data 80% to train 20% for test
- predict price for 20% test data
- Determine R-Squared value

13. **The Pima Indians Diabetes Binary Classification dataset** csv file contains all of the data of female patients of the same age belonging to Pima Indian heritage. The data includes medical data, such as glucose and insulin levels, as well as lifestyle factors of the patients. The



Pima Indians  
Diabetes Binary Clas

columns in the dataset are as follows:

- Number of times pregnant
- Plasma glucose concentration of 2 hours in an oral glucose tolerance test
- Diastolic blood pressure (mm Hg)
- Triceps skin fold thickness (mm)
- 2-hour serum insulin (mu U/ml)
- Body mass index (weight in kg/(height in m)<sup>2</sup>)
- Diabetes pedigree function
- Age (years)
- Class variable (0 or 1)
- The last column is the target variable or class variable that takes the value 0 or 1, where 1 is positive or affected by diabetes and 0 means that the patient is not affected.
- **You have to build models that could predict whether a patient has diabetes or tests positive or not using logistic regression**

14. Use hotel.csv file and Show how to cluster hotel location data with K-means Clustering. That is perform K-means clustering on hotel location data to identify whether the hotels are located in the same district.



hotel.csv

Using k-means cluster location data for 3 clusters.