CIS-5810 Fall’17 Python Project on

**By: Nikeeta Akbari & Abhishek Shah**

**Guided By: Dr. S. Balan**



**U.S. CARDIOVASCULAR DISEASE MORTALITY RATES**

**A. DATASET URL**

<https://data.opendatasoft.com/explore/dataset/us-cardiovascular-disease-mortality-rates-by-county-1980-2014@public/>

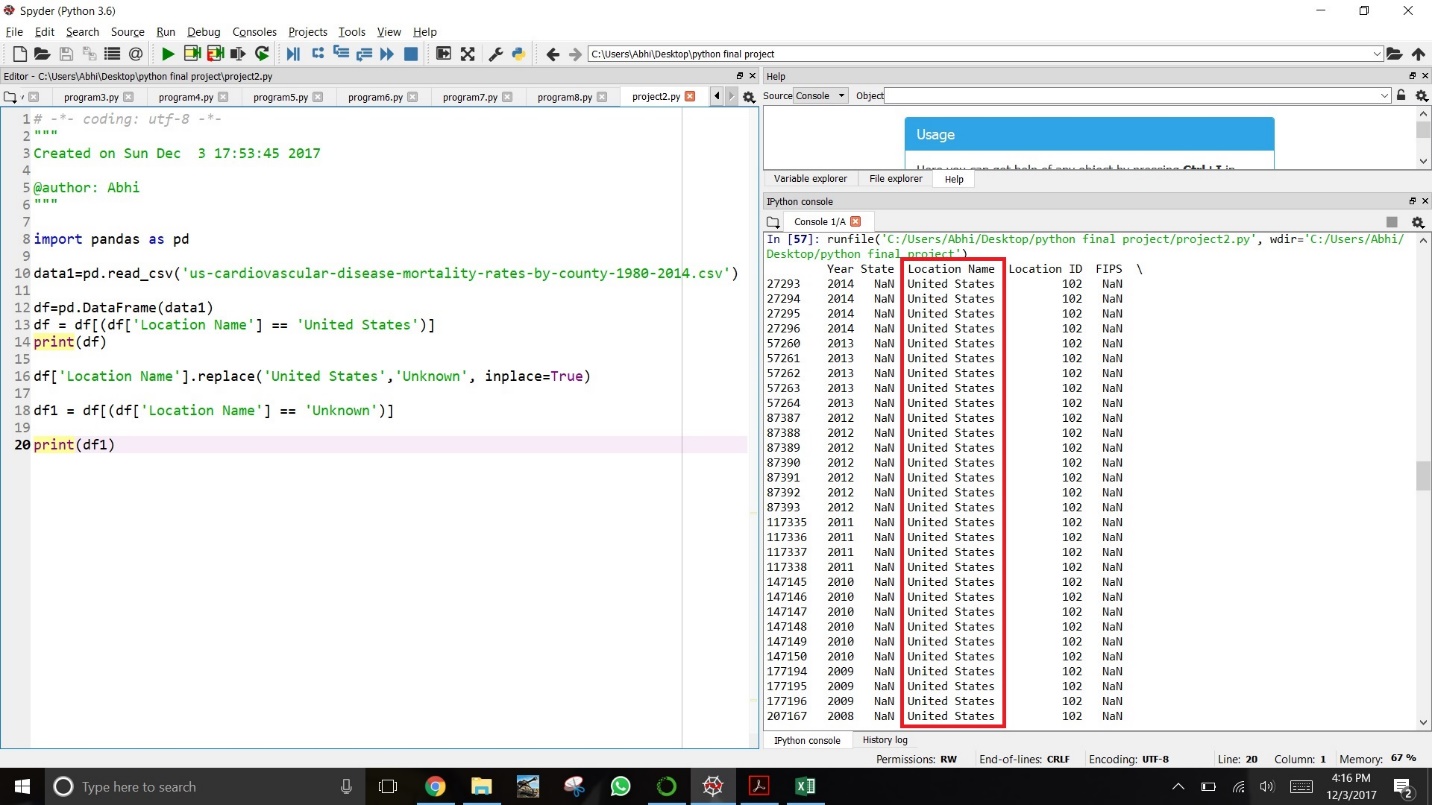
**DATASET FORMAT:**Comma Separated Values (CSV)

**DATASET DESCRIPTION:**

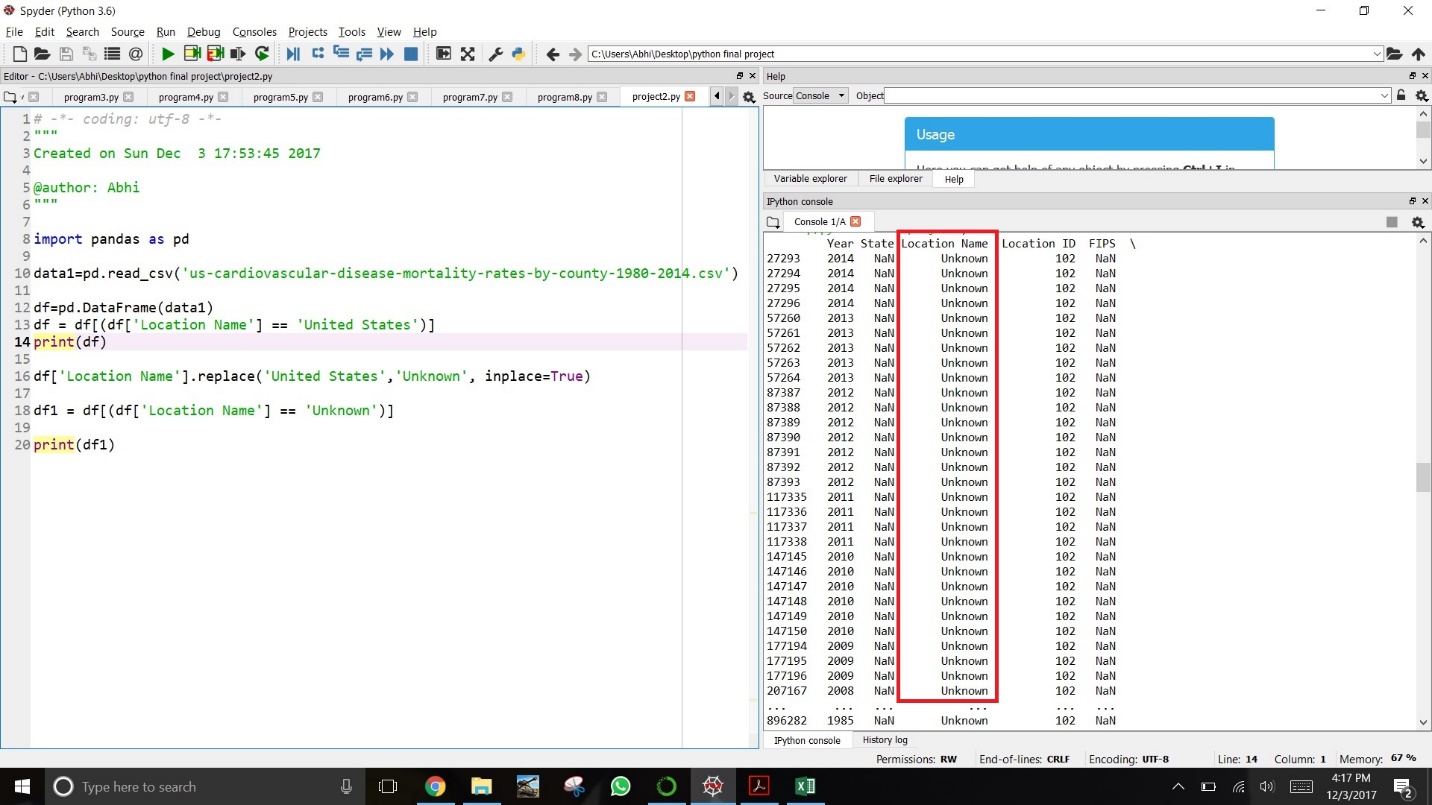
|  |  |
| --- | --- |
| **Data Field** | **Description** |
| Year | The year in which the death occurred |
| State | State initials of US |
| Location Name | Name of the Counties |
| Location ID | County ID |
| FIPS | 5-Digit Federal Information Processing Standards Code which uniquely identified Counties |
| Cause Name | Name of the type of Cardiovascular Disease |
| Cause ID | Disease ID |
| Sex | Male/Female |
| Sex ID | 1 for Male, 2 for Female |
| Upper bound | Upper bound of number of deaths per 100,000 |
| Lower bound | Lower bound of number of deaths per 100,000 |
| Value | Average of number of deaths per 100,000 |
| Measure Name | All deaths |

**B. DATA CLEANING:**

1.) Used **replace()** function to replace inappropriate values in column ‘Location Name’ (Replaced value ‘United States’ in Location name column with value ‘Unknown’, as Location Name column represents County Name)

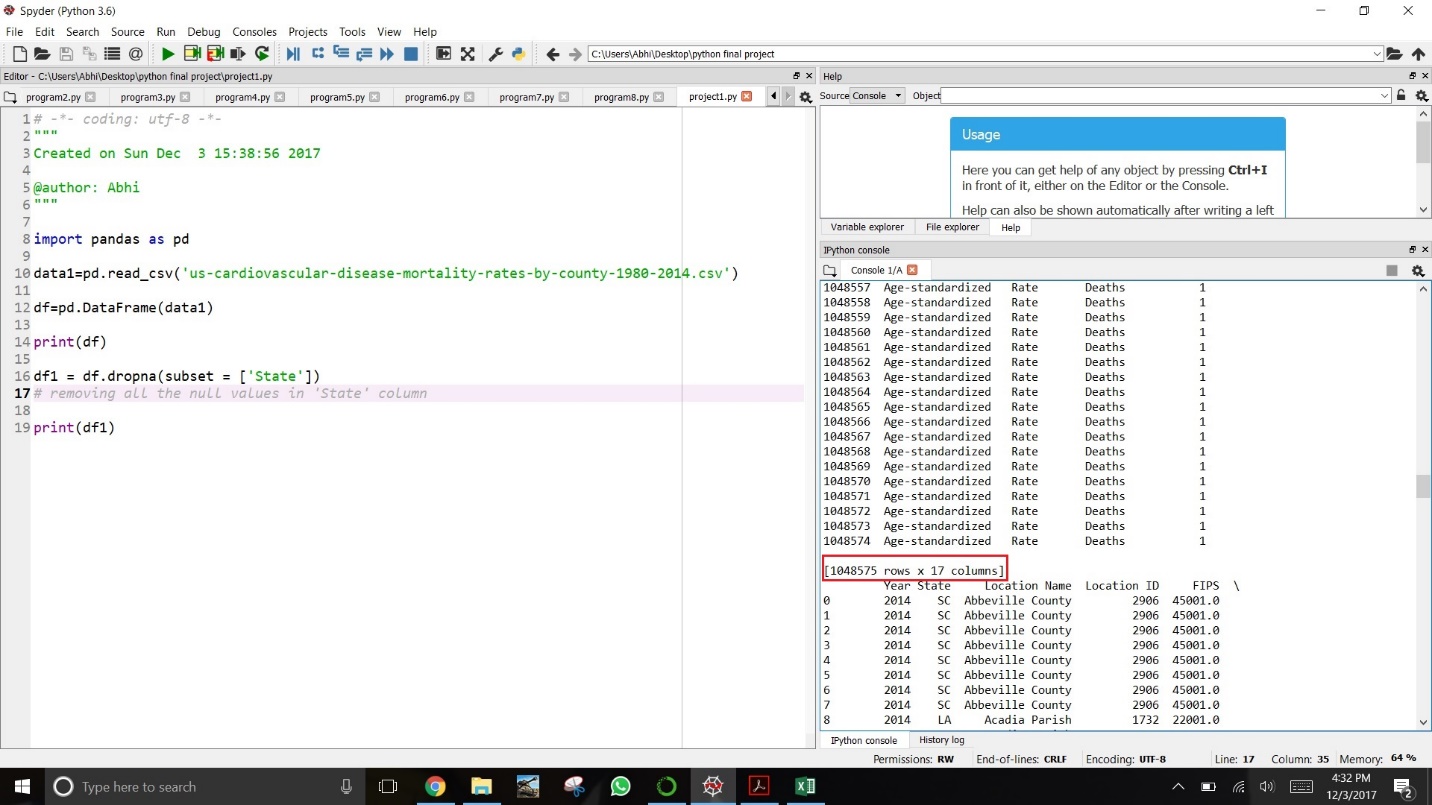


Screenshot 1: Before replacing values

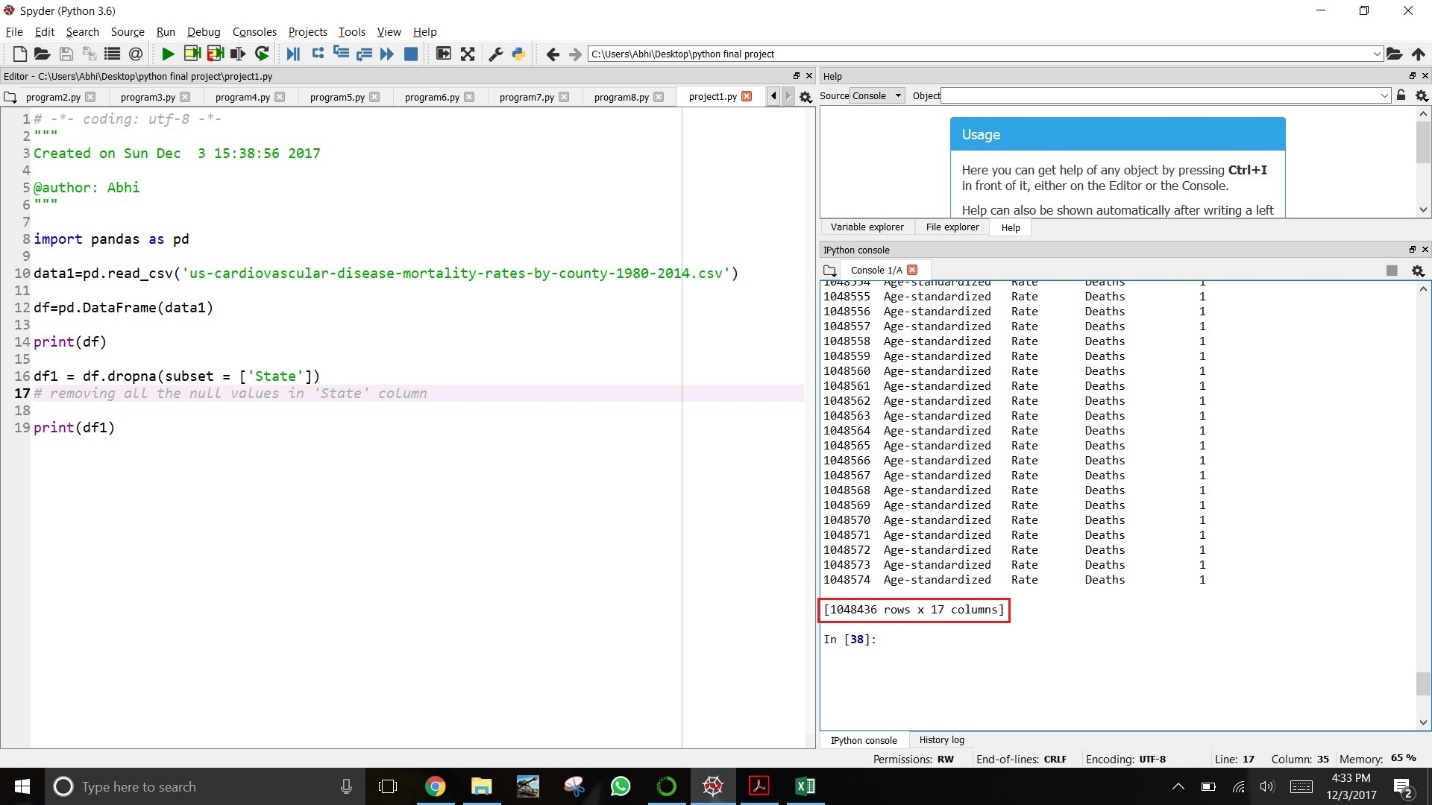


Screenshot 2: After Replacing values

2.) Eliminating **null values** from column ‘State’ using **dropna()** function

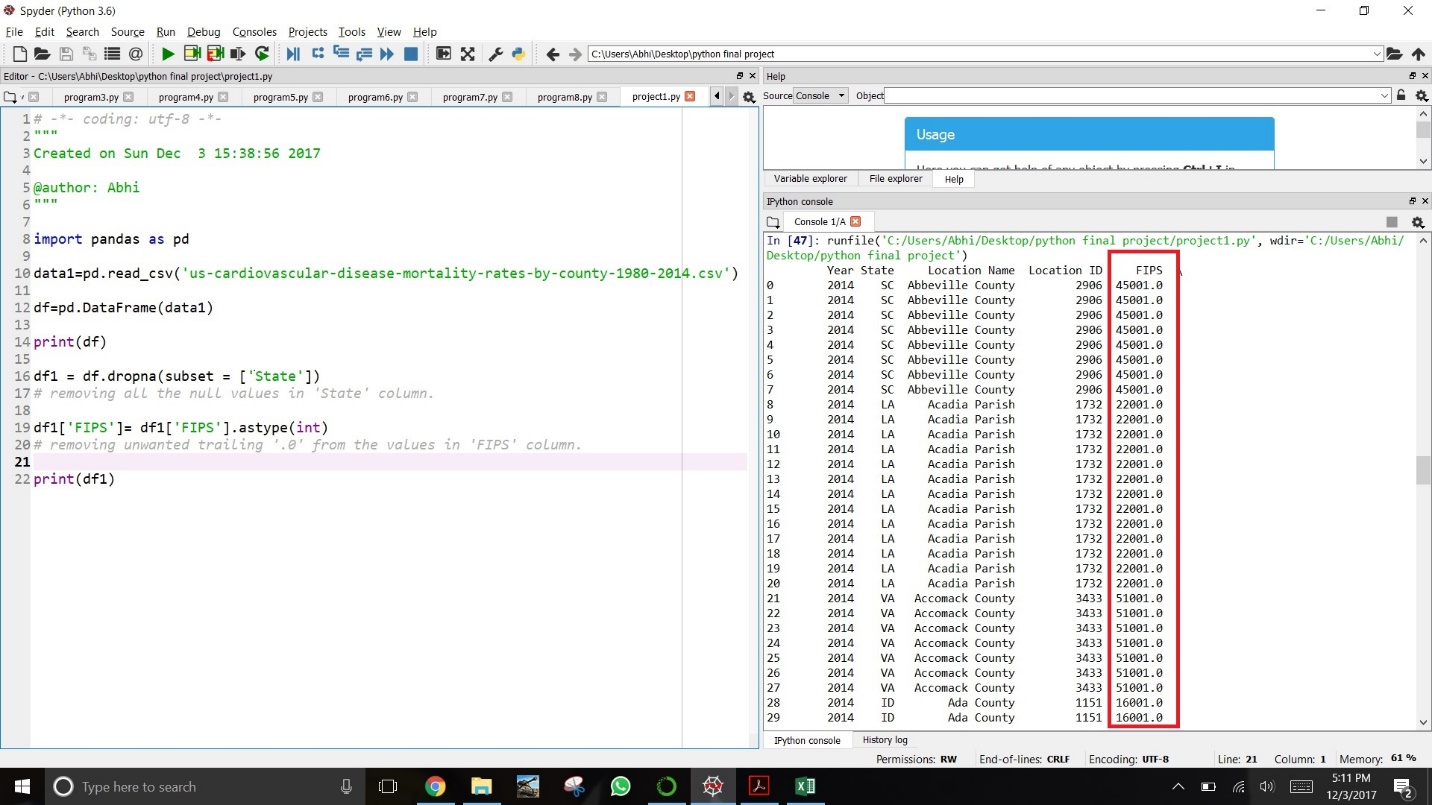


Screenshot 3: Before Removing **Null Values**, it shows 1048575 number of Records

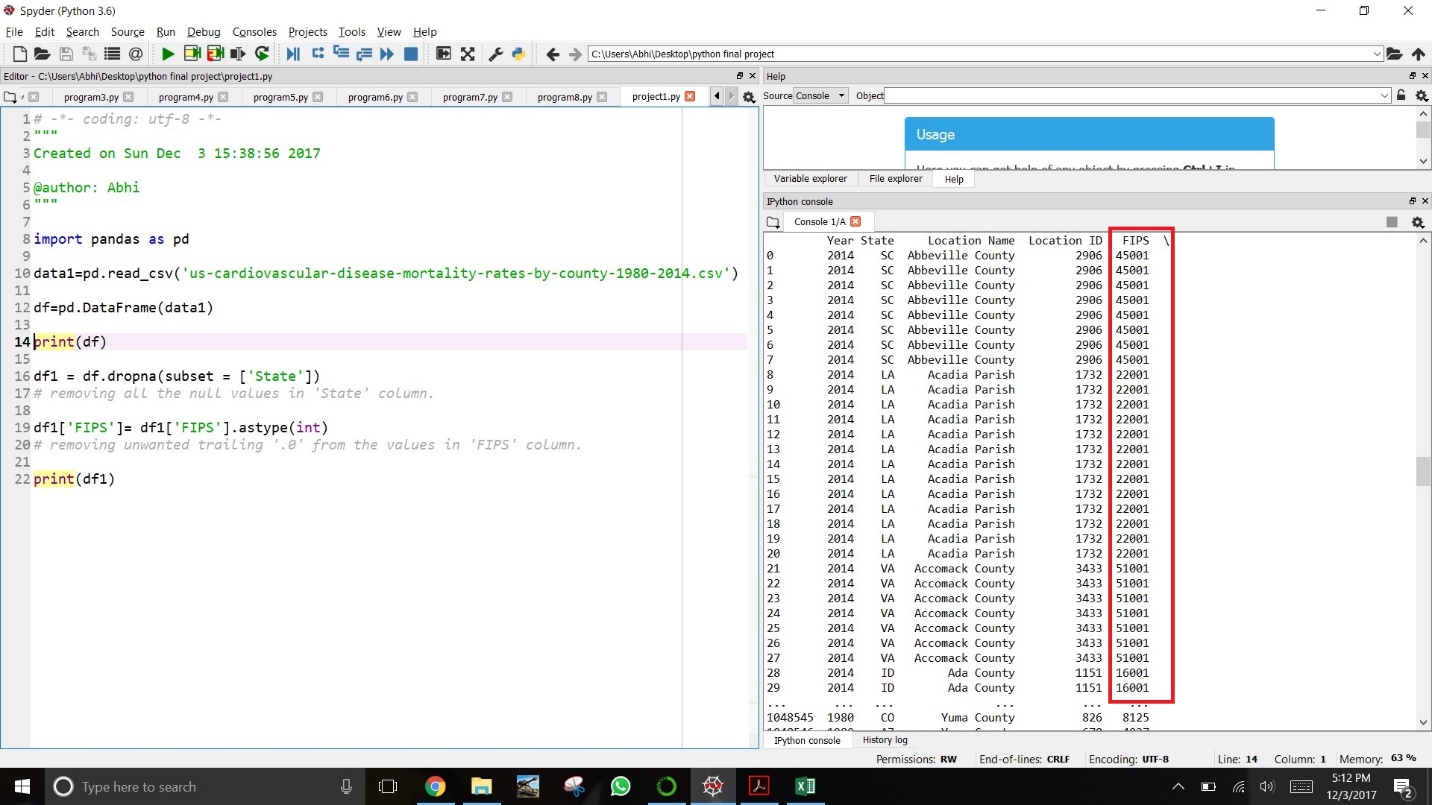


Screenshot 4: After Removing **Null Values**, it shows 1048436 number of Records

3.) Removing unwanted trailing zeros from column ‘FIPS’ using **astype()** function

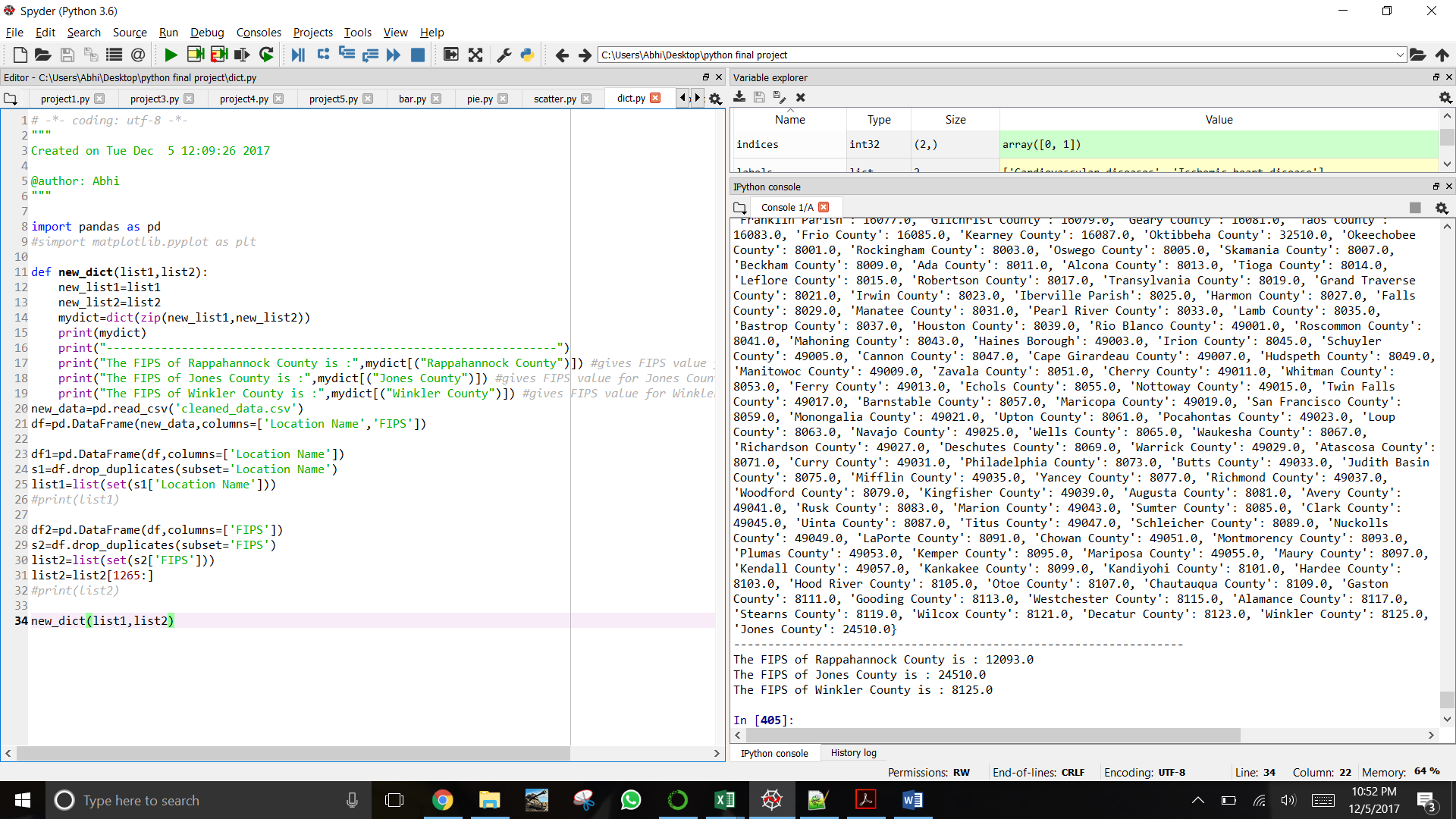


Screenshot 5: Before removing trailing .0s from column ‘FIPS’



Screenshot 6: After removing trailing .0s from column ‘FIPS’

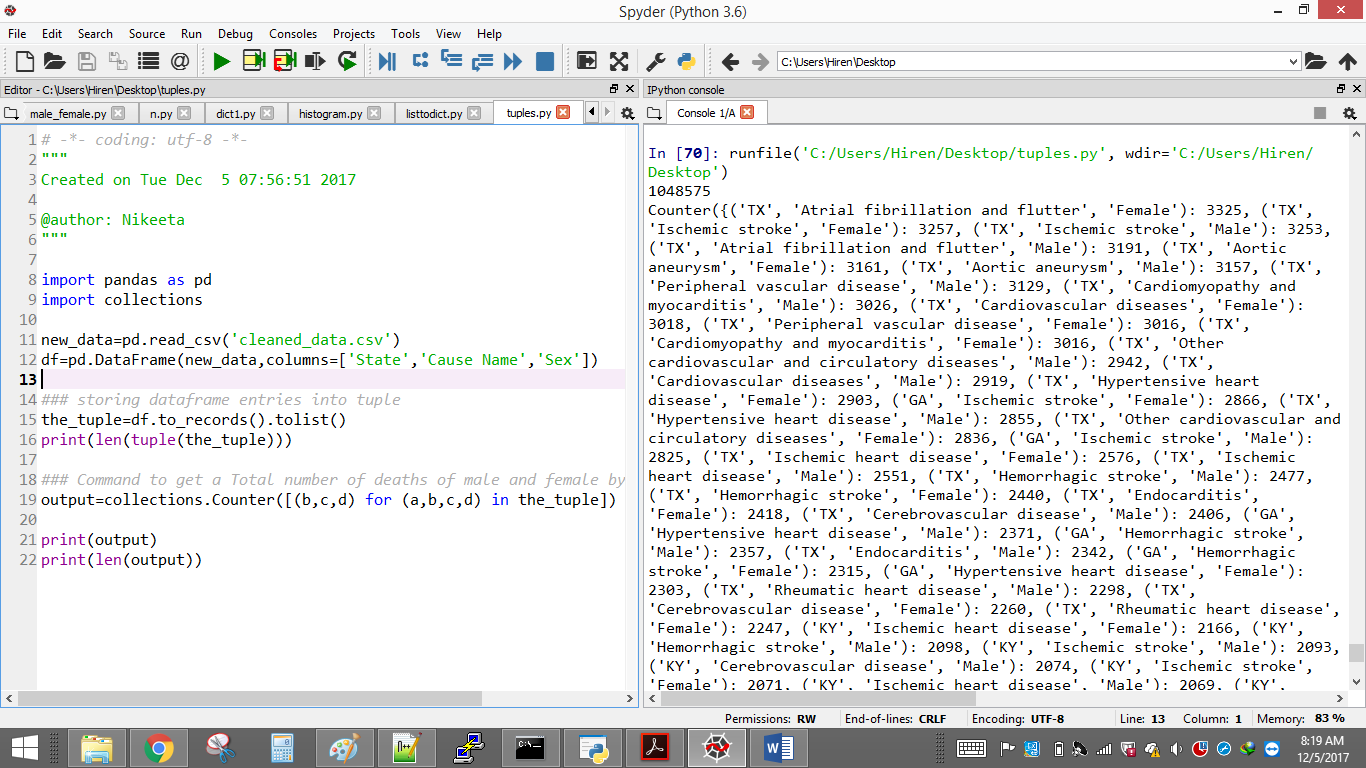
**USE OF LISTS AND DICTIONARIES:**



Screenshot 7: Shows use of **Lists, Dictionary and user defined Functions**

**Description:**

We used **list** to get a specific column from a dataset and then we created a **user defined function** which is used to create **dictionary**. This dictionary has ‘Location Name’ as its key and ‘FIPS’ as its value. The output shows key-value pair of Dictionary as well as retrieve value from a specific key. This is an example of Lists, Dictionary and Function in a simple manner. We have continued the use of the same in our further analysis and visualizations.

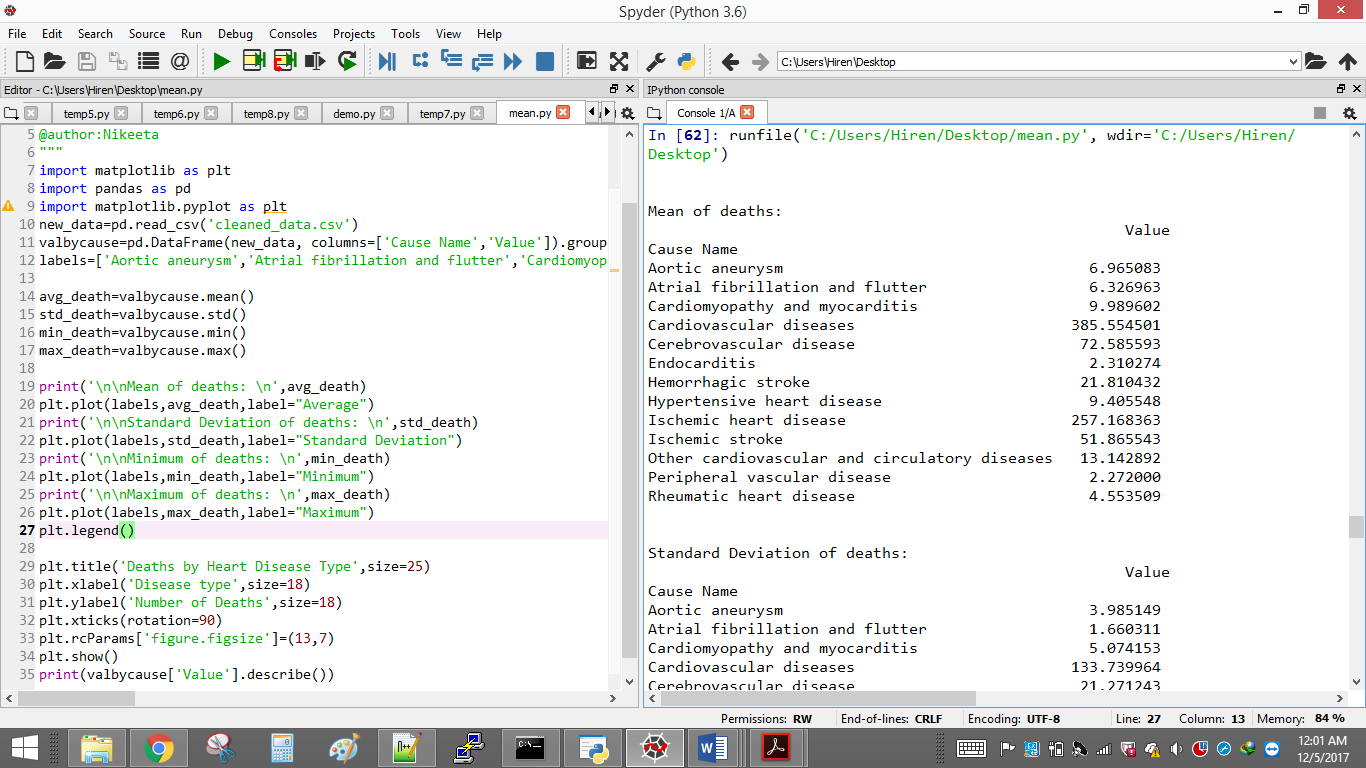


Screenshot 8: Shows use of **Data Frames and Tuples**

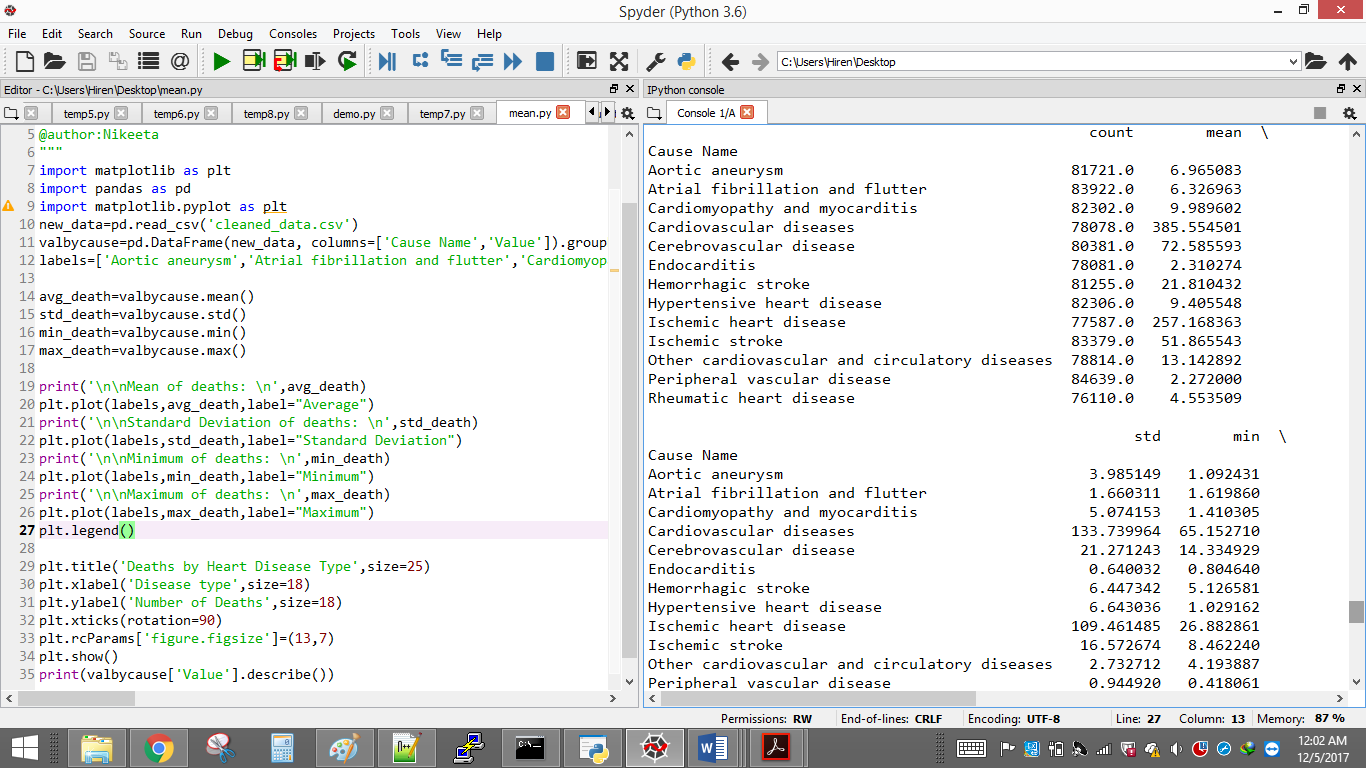
**Description:**

Here, we used **Data frame** to fetch a specific column from a dataset. This data frame is further used to create a **Tuple**. Tuple has columns State, Cause Name and Sex which gives us unique occurrences using these 3 columns.

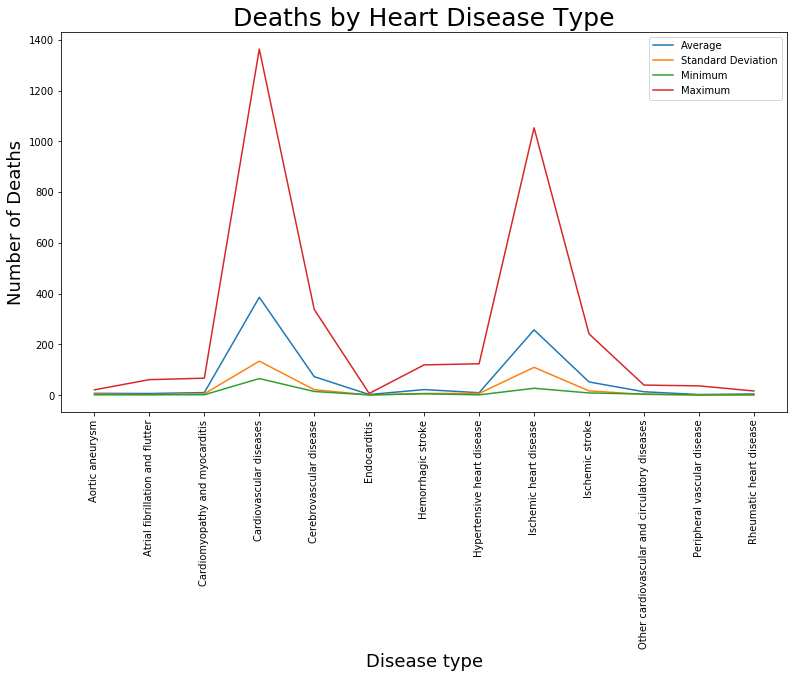
**C. SUMMARY STATISTICS:**

****

Screenshot 9: Summary Statistics showing **Mean** and **Standard Deviation** of Deaths

****

Screenshot 10: Summary Statistics using **describe()** function to show count, mean, std, min, max for Deaths for all Disease



Screenshot 11: Line graph showing Disease Type vs Number of Deaths

**Description:**

Above line chart shows 4 line-graphs according to the death rate statistical data for 13 heart disease type, where Blue, Yellow, Green and Red shows Average, Standard Deviation, Minimum and Maximum death values for 13 different disease type respectively. It clearly shows that Number of Deaths due to Cardiovascular diseases and Ischemic heart disease are significantly high compared to other heart diseases.

**D. ANALYSIS & VISUALIZATIONS:**

**Question 1. How is the Female and Male death trend analyzed year by year in United States?**

**Code:**   
import pandas as pd  
import matplotlib.pyplot as plt

new\_data=pd.read\_csv('cleaned\_data.csv')

df=pd.DataFrame(new\_data, columns=['Year','Sex','Value'])

#Adding Values for Male & Female for each Year

s=df.groupby(['Year','Sex']).sum()

print(s)

#Shows Bar chart

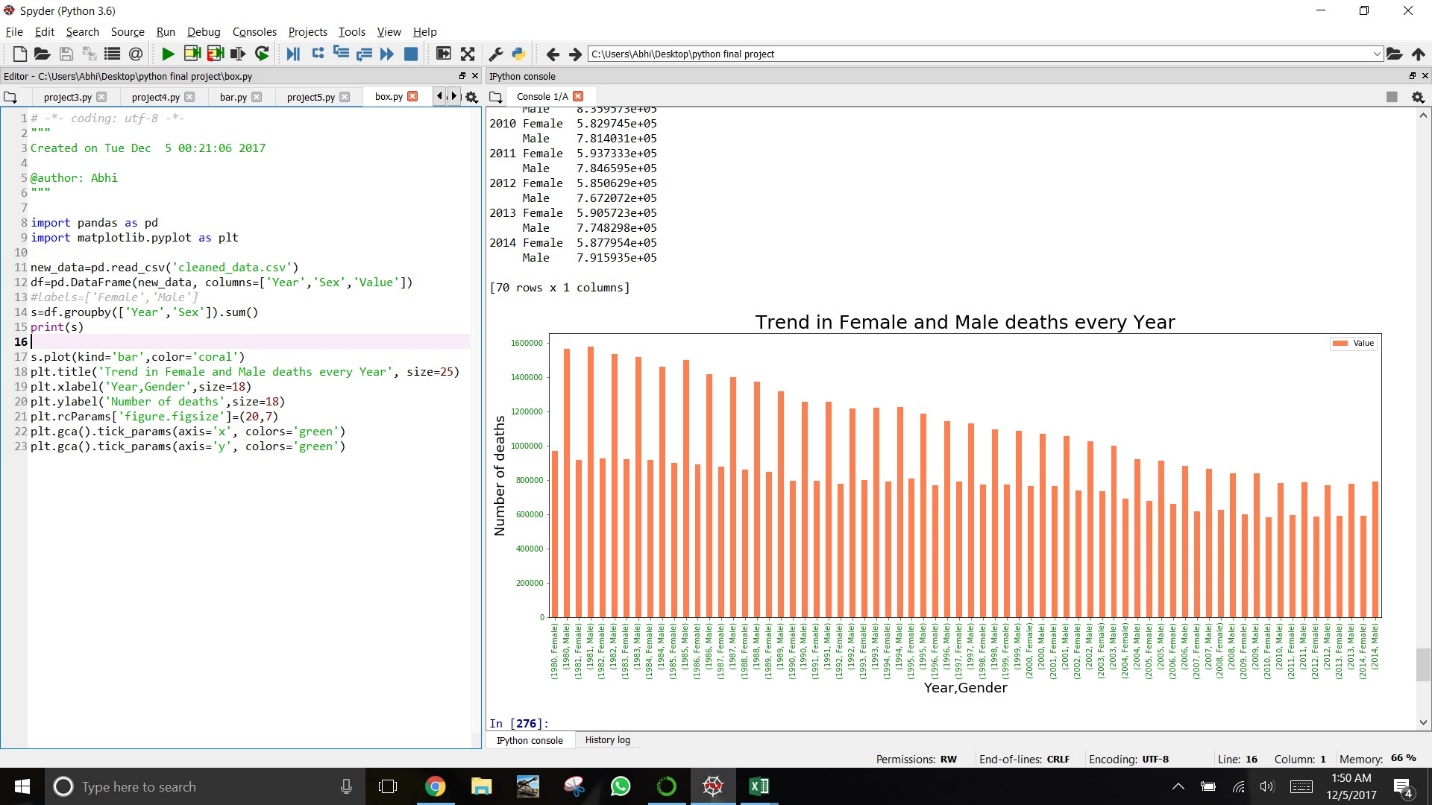
s.plot(kind='bar',color='Coral')

plt.title('Number of Female and Male deaths every Year', size=25)

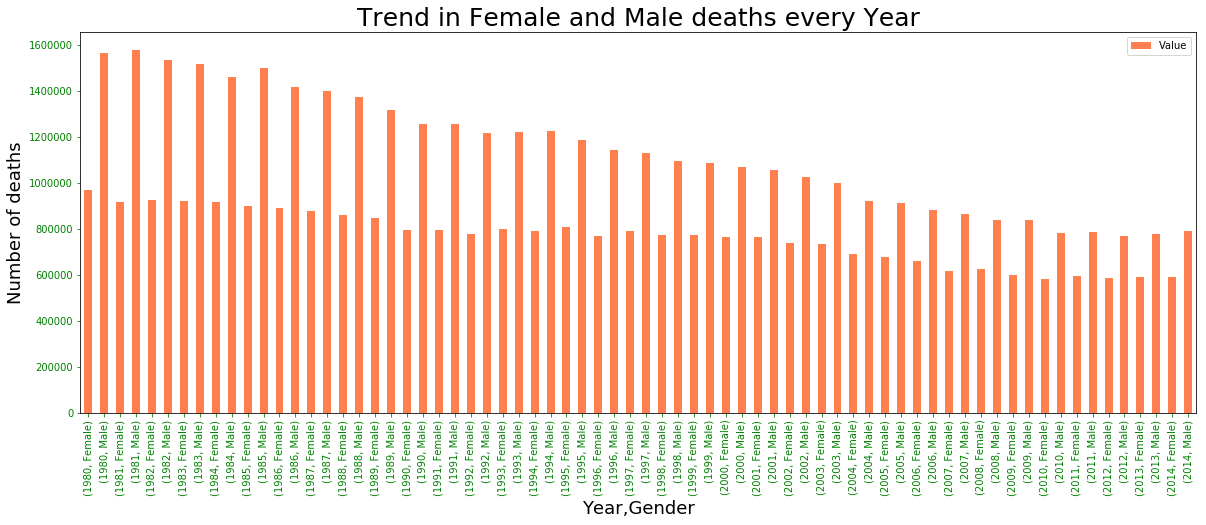
plt.xlabel('Year,Gender',size=18)

plt.ylabel('Number of deaths',size=18)

plt.rcParams['figure.figsize']=(25,7)

**Output**

Screenshot 12: Summary statistic of Number of deaths by Year and Gender



Screenshot 13: **Bar chart** showing the trend of Number of Deaths for Male and Female every Year

**Description:**

The above **Bar chart** shows the number of Male and Female deaths every year. We analyzed that Male deaths compared to Female deaths were relatively high (slightly more than 1.5x) in 1980. Looking at the trend from 1980 to 2014, we can see that Male deaths reduced linearly from 1980 to 2014 whereas Female deaths slightly reduced from 1980 to 2014 but not as linearly as Male. Thus, the medical institutes or hospitals should note this changes or trends for their treatment processes gender-wise.

**Question 2. Which State has high death rate due to Cardio vascular disease for year 2014?**

**Code**

import pandas as pd

import matplotlib.pyplot as plt

new\_data=pd.read\_csv('cleaned\_data.csv')

df=pd.DataFrame(new\_data,columns=['Year','State','Cause Name','Value'])

df1= df[(df['Year']==2014) & (df['Cause Name']=='Cardiovascular diseases')]

#print(df1)

df2=pd.DataFrame(df1,columns=['State','Value'])

#print(df2)

df3=pd.DataFrame(df2,columns=['State'])

s1=df3.sort\_values('State',ascending=True).drop\_duplicates(subset='State')

#print(s1)

list1=list(set(s1['State'])) #using list for labeling the pie plot.

list1=sorted(list1) #sorting the list alphbetically.

s2=df2.groupby(['State']).sum()

#prints total number of deaths state-wise for year 2014

print(s2)

#plot the pie chart

plt.pie(s2, labels=list1, autopct='%1.1f%%', startangle=140)

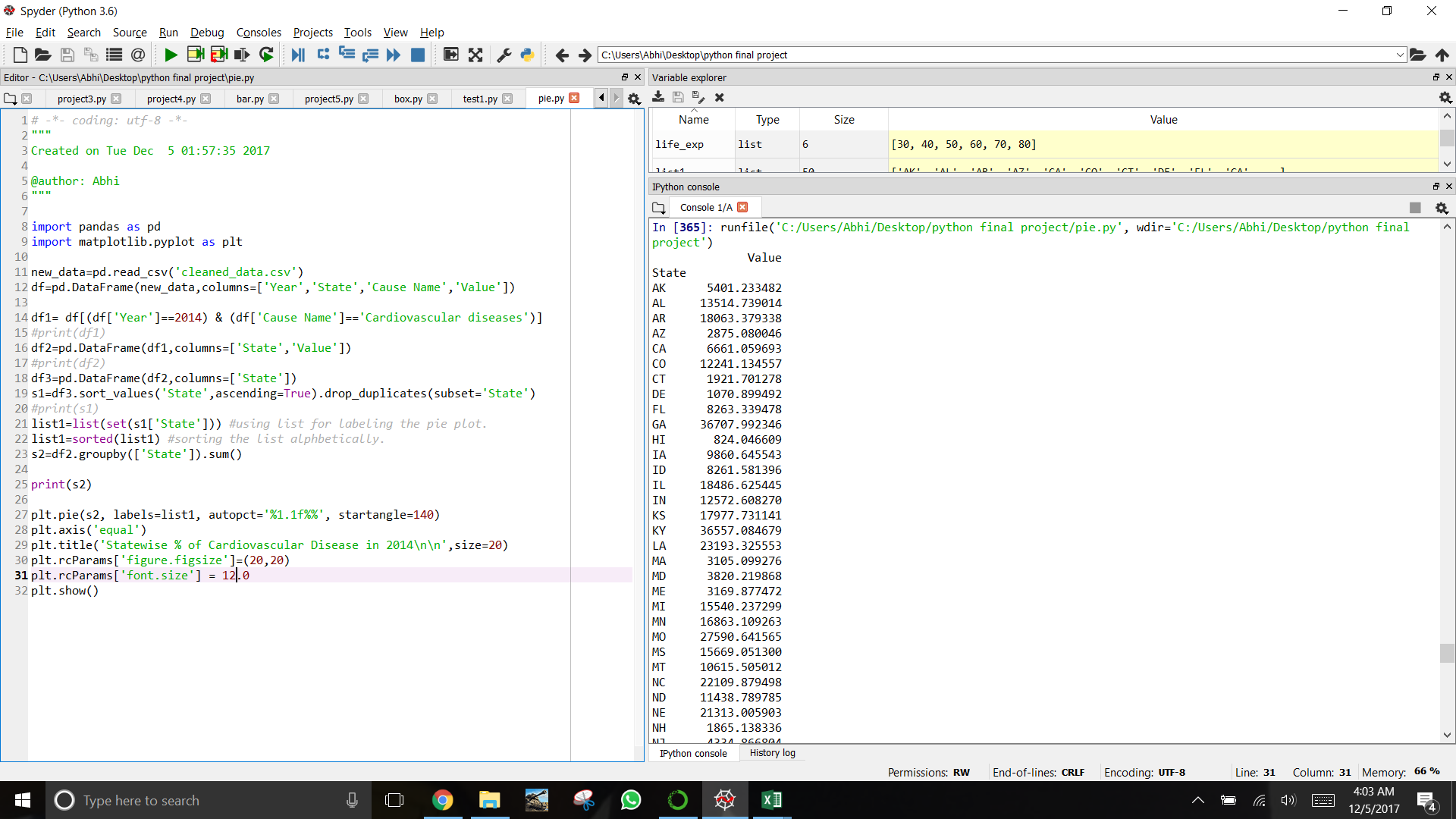
plt.axis('equal')

plt.title('Statewise % of Cardiovascular Disease in 2014\n\n',size=25)

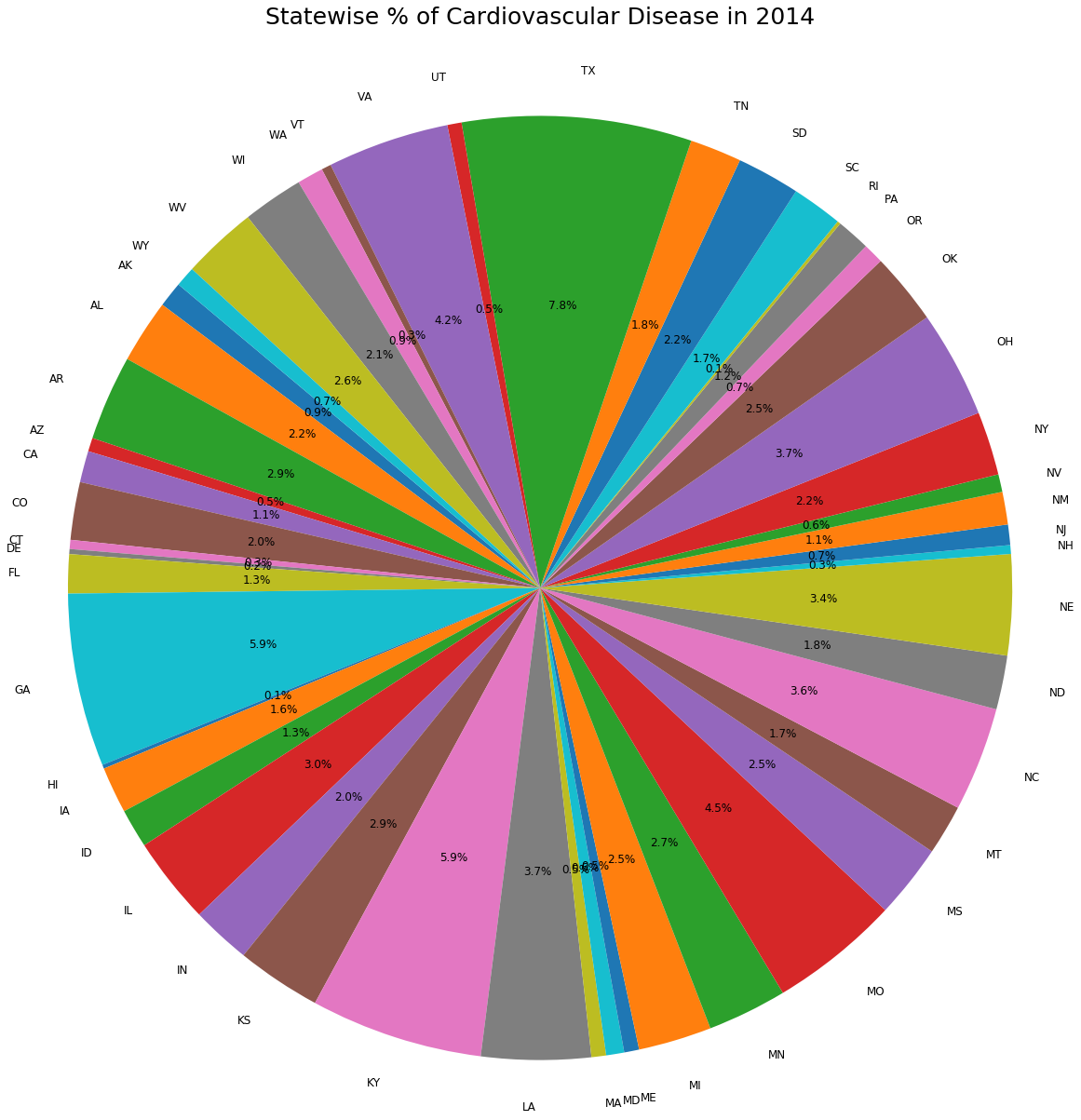
plt.rcParams['figure.figsize']=(20,20)

plt.rcParams['font.size'] = 12.0

plt.show()

**Output**  
  


Screenshot 14: Snapshot showing the total number of deaths state-wise due to Cardio vascular disease



Screenshot 15: **Pie Chart** showing State-wise percentage of deaths due to Cardiovascular disease in 2014

**Description:**  
We have chosen the **Pie Chart** because we thought to analyze a specific disease type i.e. Cardiovascular disease for each State in year 2014. We see that Texas, Georgia, Kentucky, Missouri and Virginia are the top 5 states that has high death rates due to cardio vascular disease compared to other states. If we notice, these states belong to the South-Eastern part of the United States of America. A location represents its people’s lifestyle. Climate, food habits, etc. can be the reason behind the cause of this disease in South-east US. This data might help medical researchers to analyze deeper in this trend.

**Question 3. How many people died due to cardio vascular disease in year 2014 compared to start of data accumulation (Year 1980)? How much is the variation from one state to another (or one county to another in 1980 and 2014?**

**Code**

import pandas as pd

import matplotlib.pyplot as plt

new\_data=pd.read\_csv('cleaned\_data.csv')

df=pd.DataFrame(new\_data,columns=['Year','State','Value','Cause Name','Sex'])

### creating a dataframe for 1980 year

df1 = df[(((df['Cause Name'] == ' Ischemic heart disease ')) & (df['Year'] == 1980))]

### creating a dataframe for 2014 year

df2 = df[(((df['Cause Name'] == ' Ischemic heart disease’)) & (df['Year'] == 2014))]

### Appending more data frames and store it into temp csv files

df2 = df2.append(df1)

df2.to\_csv('temp.csv')

## read a temp.csv for further analysis

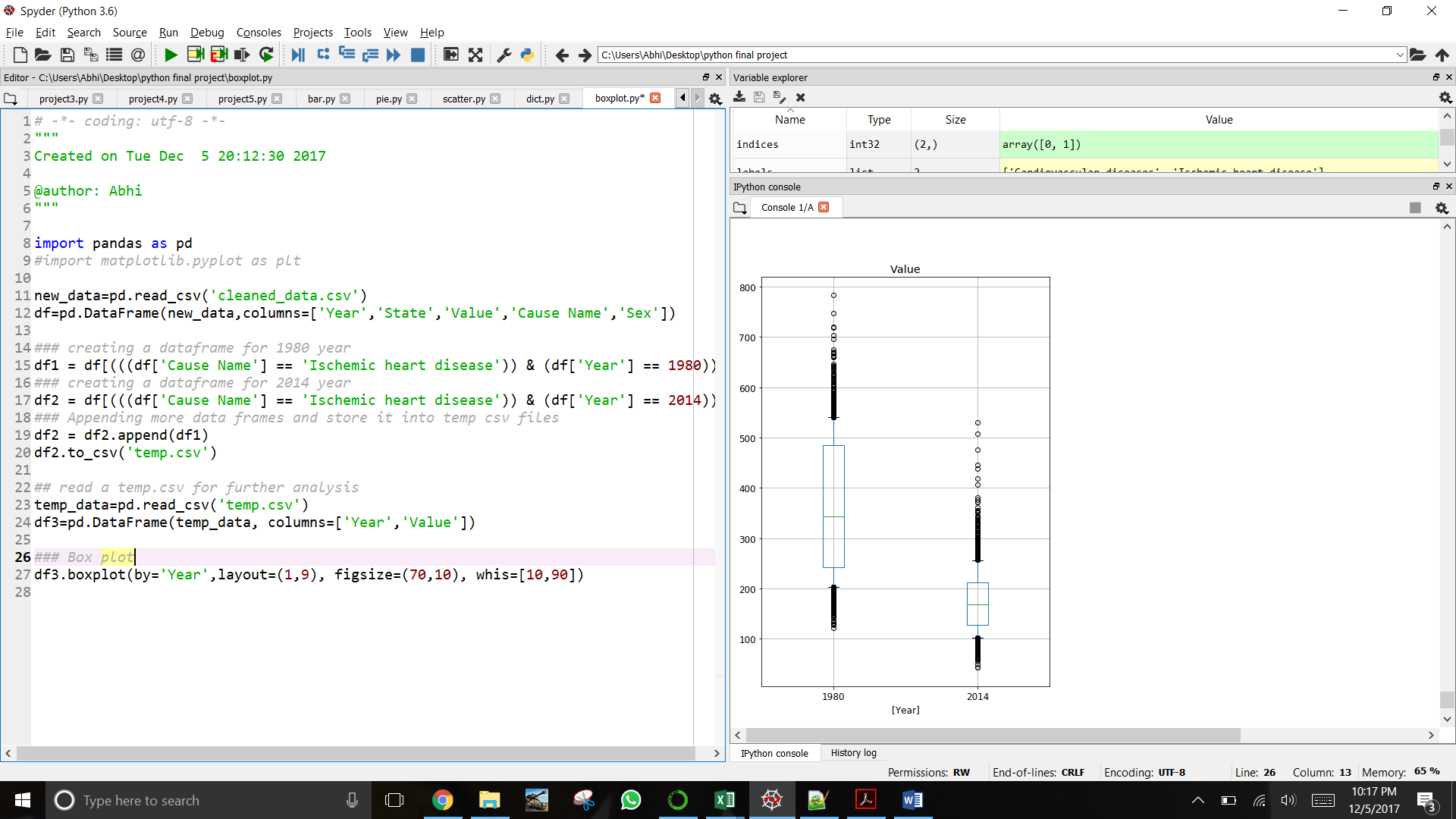
temp\_data=pd.read\_csv('temp.csv')

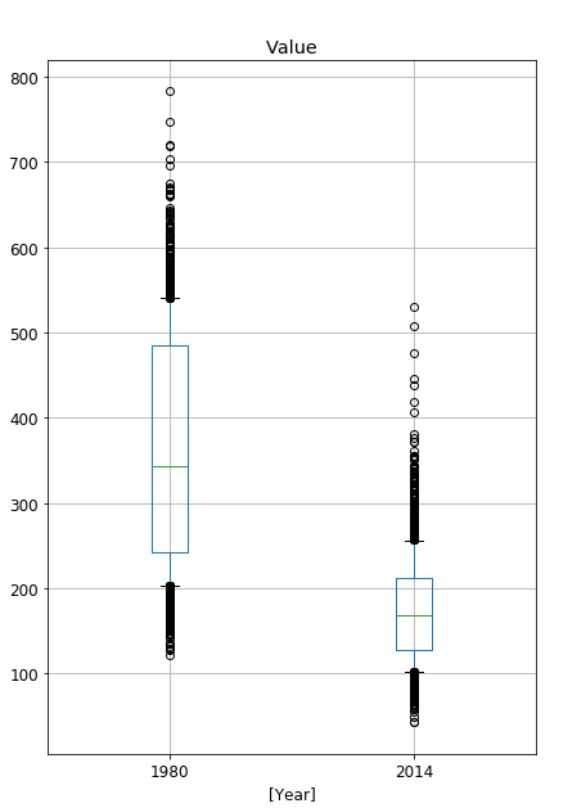
df3=pd.DataFrame(temp\_data, columns=['Year','Value'])

### Box plot

df3.boxplot(by='Year',layout=(1,9), figsize=(70,10), whis=[10,90])

**Output**

  
Screenshot 16: Snapshot showing the upper and lower bound of the deaths due to Ischemic Heart Disease



Screenshot 17: **Box Plot** showing range of Death Rate due to Ischemic Heart disease comparing years 1980 and 2014

**Description:**

Above Box Plot shows total deaths due to Ischemic Heart disease for year 1980 and 2014. We used .to\_csv() function to create a new **file** from an existing dataset. It clearly shows that death value was almost 1.5x times in year 2014 as compared to 1980. In addition, it also shows death values for Male and Female. In Chart, Male death value shows in upper bound and Female death value shows in lower bound, this represents that death value is high for male as compared to female for Ischemic Heart disease. We can also say that irrespective of disease type death rate due to heart disease is high in Male.

**CONCLUSION:**

Heart disease remains the leading cause of death in the U.S., accounting for almost 1 in every 4 deaths, and affecting significantly more men than women. Death rate due to Cardio vascular disease and ischemic heart disease are high compared to other disease type. High death rate because of heart disease might be because of multiple factors like high cholesterol level, fast foods, busy life schedule, stress, unavailable medical facilities, etc. Science has been making a good progress in medical field but still one must eat a healthy diet and do a regular exercise to reduce probability of getting impacted by heart disease. We hope that our analysis helps and acknowledges doctors, researchers, hospitals and medical institutions to reduce this count of deaths due to different types of Cardiovascular diseases in United States. Number of death due to heart disease decreases significantly from 1980 to 2014, but it is still a long way to go.