

**CIS 5810: Healthcare Information System**

**Fall 2017**

**Pregnancy Risk Assessment Monitoring System Data Analysis**

**using Python**

Submitted to: Dr. Shilpa Balan

**Submitted By:**

**Narendra Mali CIN:305848717**

1. **Dataset URL:**

<https://catalog.data.gov/dataset/cdc-pramstat-data-for-2011-8b9bf>

**Data Format & Size:**

Format: .CSV

Size: 155MB

**Data Description:**

Dataset contains information for 26 states in United States and this information is based on 200 diverse types of questions. Data also covers different topics such as prenatal care, diabetes, HIV, type of delivery (vaginal or Caesarean) etc. This dataset includes information for the year 2011 only. This dataset is comprehensive data which also gives information about Confidence limit which helps us to get biological statistics. The data is organized by location (national and state) class, topic and data source. The data can be viewed by temporal trends and stratified by age group, and race/ethnicity.

Dataset has column named question and from this column there are several questions which are highly valuable for data analysis. Some of the questions can determine cases of physical abuse, alcohol consumption during pregnancy, teenage pregnancy, health of the infant after the delivery and many more. This dataset also provides the response that was checked by the mother herself during her visit to hospital. This dataset can also help us determine the mothers who are taking the benefits of Insurance, Medicaid and Medicare services. Family planning is important, and this dataset helps us understand that there are many mothers who visits hospitals specially to discuss parenthood, family length and number of years of difference between two births.

Weight of the baby is another crucial part of post and pre-pregnancy to make sure that baby brought to this world is healthy. This is a collaborative effort of parent and doctor who make sure that mother is fed properly so that baby and mother both are healthy post-delivery. This dataset gives us information about baby weight post pregnancy. Also, it gives information of mother post-delivery. PRAMS topics include contraception, breastfeeding, mental health, morbidity, obesity, preconception health, pregnancy history, prenatal-care, sleep behavior, smoke exposure, stress, tobacco use, WIC, Medicaid, infant health, and unintended pregnancy. There are more than 500,000 rows and 28 columns of comprehensive information regarding pregnant women which can lead to many meaningful insides.

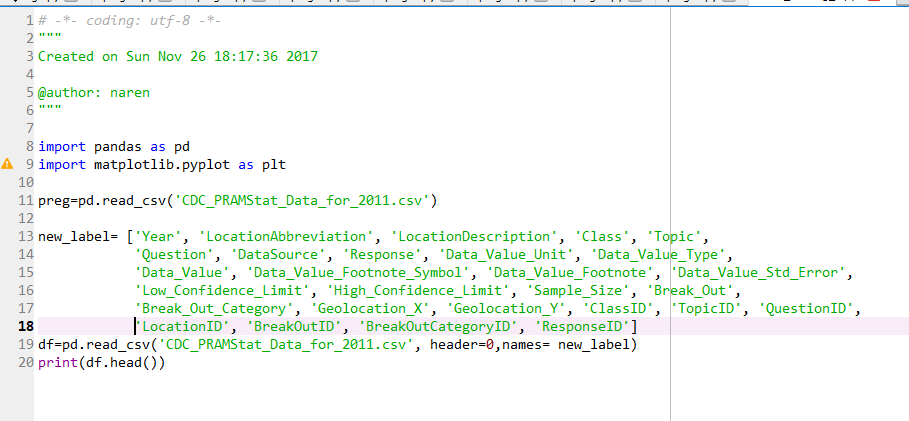
1. **Data Cleaning**
2. **Column names were in abbreviation format**

Names of the column were changed to increase the understanding of the data.

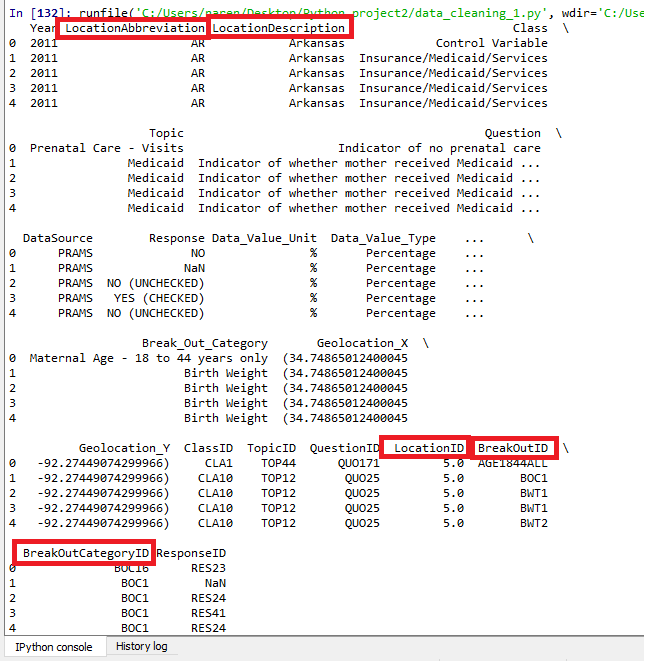
Before cleaning: Location Abbreviation was LocationAbbr and Location Description was LocationDesc etc.

**Code: [Tools used: “List”]**

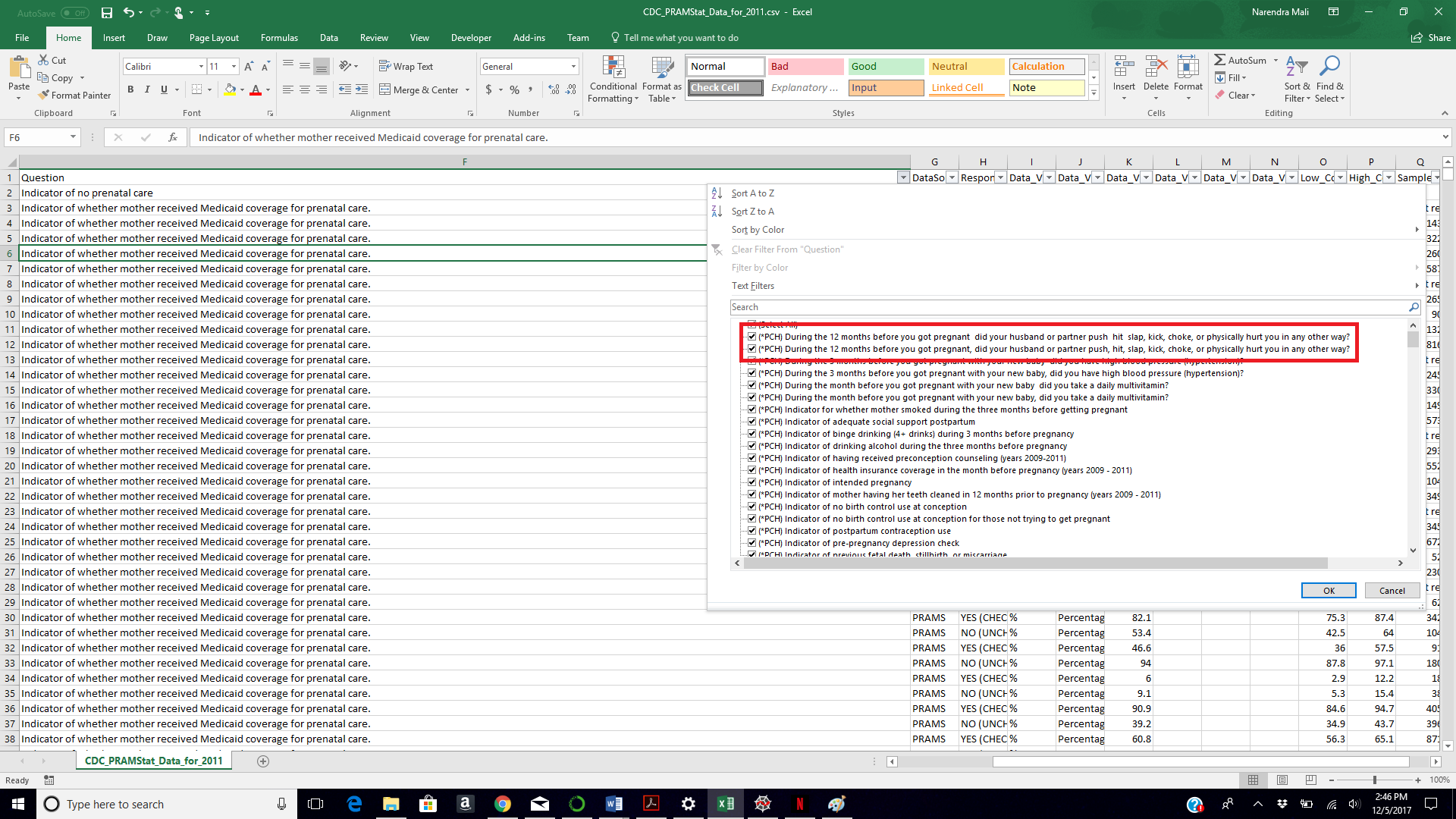


**Output:**



1. **In the question column, one question was repeated because it had no punctuation hence while querying the dataset output was getting calculated wrongly.**

Before cleaning:



**[Tools used: “File” & “String”]**

Question was getting calculated twice even though the question was same. Only difference was the second question was missing punctuation. By saving the file through python File was used and string function Replace was used to find and replace the repeated question.

(Cannot paste the screenshot since the question to be replaced was long. Hence, pasting the final code)

**Code:**

import pandas as pd

import matplotlib.pyplot as plt

preg=pd.read\_csv('CDC\_PRAMStat\_Data\_for\_2011.csv')

preg = preg.fillna('')

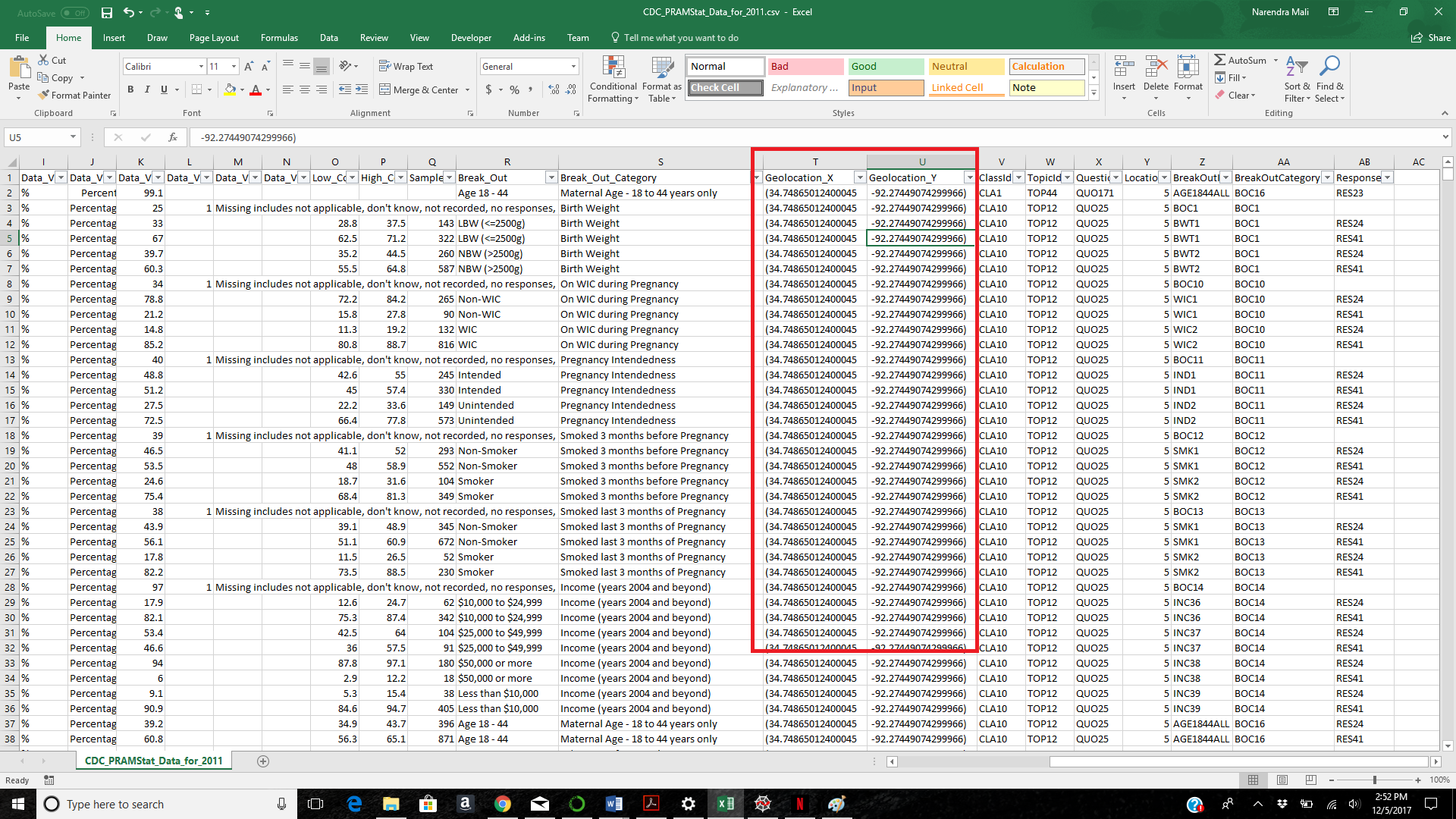
preg['Question'].replace('(\*PCH) During the 12 months before you got pregnant did your husband or partner push hit slap, kick, choke, or physically hurt you in any other way?','(\*PCH) During the 12 months before you got pregnant, did your husband or partner push, hit, slap, kick, choke, or physically hurt you in any other way?', inplace=True)

print(preg.head())

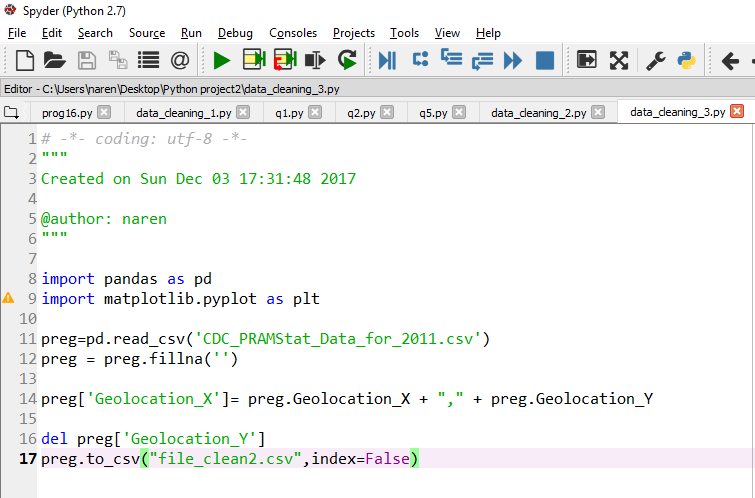
preg.to\_csv("file\_clean.csv",index=False)

1. **Geolocation were divided into two columns i.e. X-coordinate and Y-coordinate. Joined the two columns to create a single column. Also removed ‘NA’**

**Before cleaning:**



**Code:**

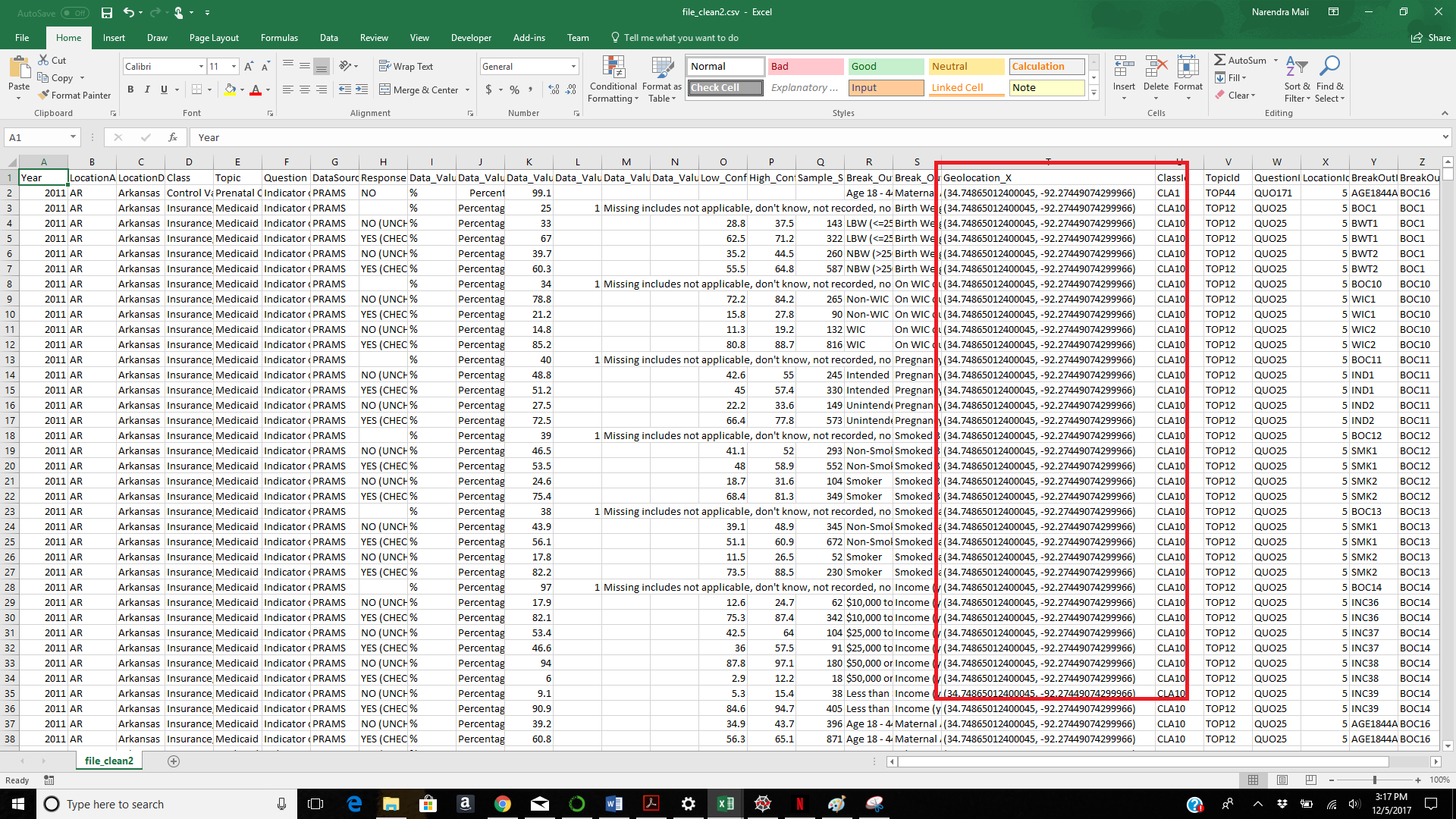


**Output:**

New instance of the file generated



**Deleted Column Geolocation\_Y** and joined the column and edited existing column Geolocation\_X.



1. **Summary Statistics:**

**Code for Summary Statistics: (dropped NA’s and used Pandas Data frame)**

import pandas as pd

import matplotlib.pyplot as plt

preg=pd.read\_csv('CDC\_PRAMStat\_Data\_for\_2011.csv')

preg.dropna()

print('minimum' )

print(preg[('High\_Confidence\_Limit')].min())

print("maximum")

print(preg['High\_Confidence\_Limit'].max())

print("mean")

print(preg['High\_Confidence\_Limit'].mean())

print('Standard Deviation')

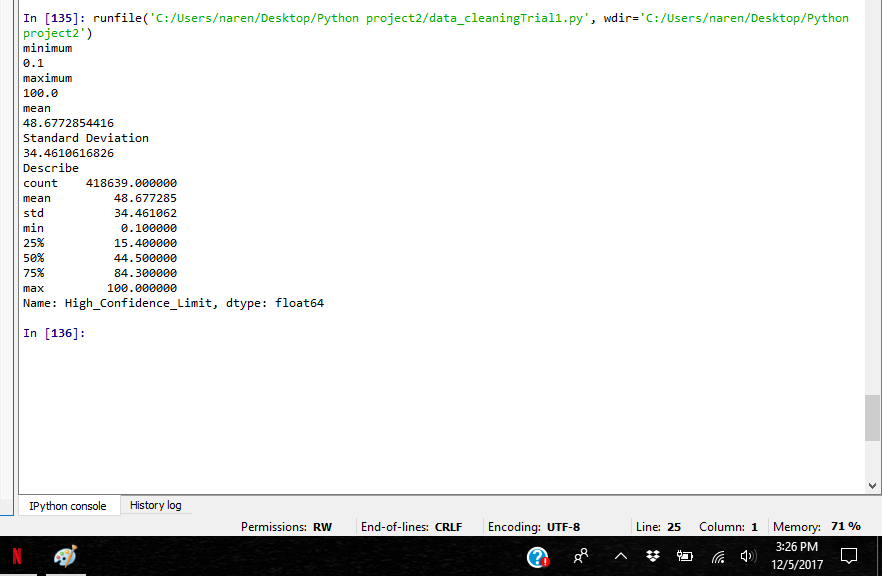
print(preg['High\_Confidence\_Limit'].std())

print('Describe')

print(preg['High\_Confidence\_Limit'].describe())

preg.to\_csv("file\_clean45.csv",index=False)

**Output:**



1. **Analysis & Visualizations**

**Question 1) How many women across United States said that they were physically abused during pregnancy?**

**[Tools Used: Tuple, Pie chart]**

**Code:**

import pandas as pd

import matplotlib.pyplot as plt

preg=pd.read\_csv('file\_clean2.csv')

preg = preg.fillna('')

topic1= (preg.Topic == 'Abuse - Physical')

intersection = preg[topic1]

explode = (0.1, 0, 0, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)

plot1 = intersection['LocationDesc'].value\_counts()

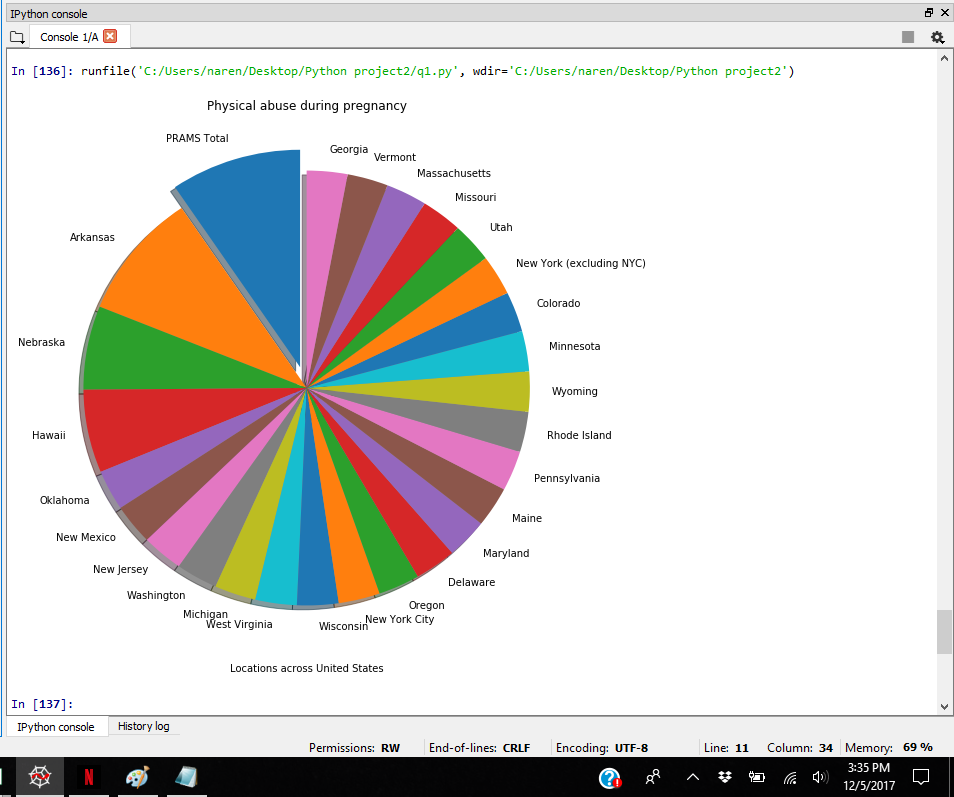
plot1.plot.pie(figsize=(10,10), shadow=True, startangle=90, explode=explode)

plt.title('Physical abuse during pregnancy')

plt.xlabel('Locations across United States')

plt.ylabel('')

**Visualization:**



In the United States, there were more than 3.5 million women are pregnant in the year 2011. It is important to go through the process of their pregnancy and hence this question of physical abuse. As per the analysis, most pregnant women in the state of Arkansas are physically abused during their pregnancy. As per the pie chart above, PRAMS total is the category with highest total which represents that more than 450 women have opted opt to list their living state.

Least amount of physical abused cases was observed were in state of Georgia. As per the statistics in USA according to CDC is 1 in 6 pregnant woman goes through situation of physical abuse. (Dimes, 2017) Above analysis supports this outcome since total, out of (almost) 600,000 entries 6799 pregnant women went through problem of physical abuse during their course of pregnancy.

**Question 2) State wise distribution of women who came in the hospital and stated that they smoked 3 months before pregnancy.**

**[Tools Used: “Functions”, Bar graph]**

**Code:**

import pandas as pd

import matplotlib.pyplot as plt

preg=pd.read\_csv(‘file\_clean2.csv’)

preg = preg.fillna('')

def format\_text(preg):

topic1= (preg.Break\_Out == 'Smoker') & (preg.Response == 'YES') | (preg.Response == 'YES (CHECKED)')

intersection = preg[topic1]

#plot1 = intersection['Topic'].value\_counts().unstack()

plot1 = intersection.groupby('LocationDesc').Class.value\_counts().unstack()

plot2 = plot1.sort\_values(by='Delivery', ascending= False)[:5]

plot2.plot(kind='bar',colormap='winter', figsize=(16,6)

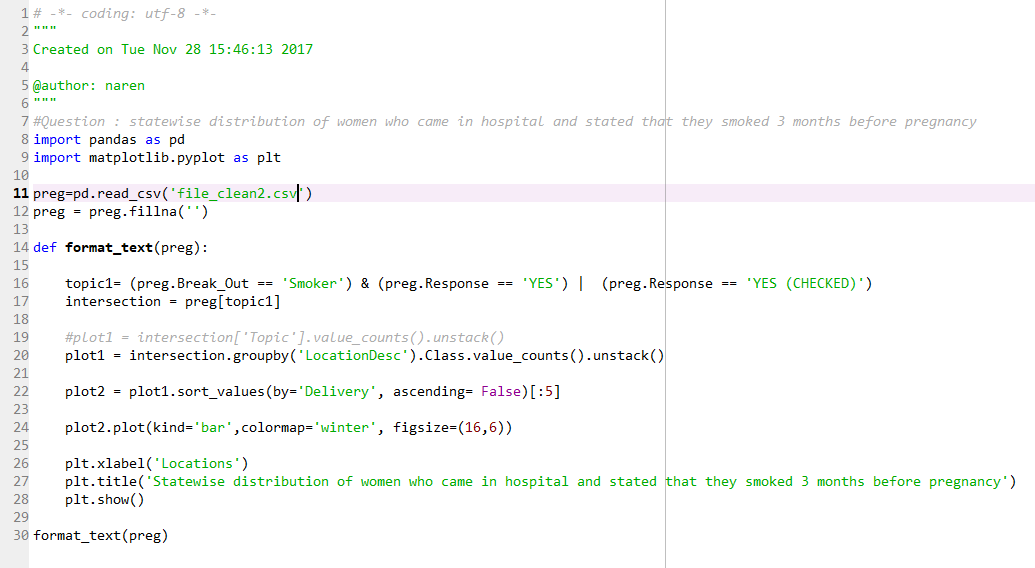
plt.xlabel('Locations')

plt.title('Statewise distribution of women who came in hospital and stated that they smoked 3 months before pregnancy')

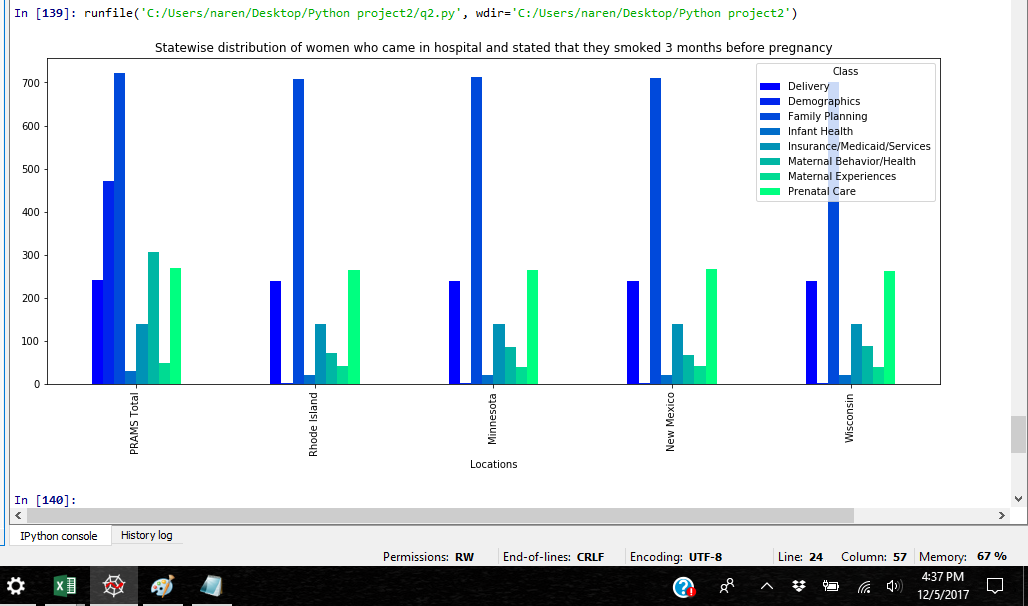
plt.show()

format\_text(preg)

**Code Screenshot:**



**Visualization:**



Smoking is a big issue in the United States. 21 percentage of all adult’s smokes in USA. (CDC, 2017) Smoking for an average human is dangerous but smoking during pregnancy is dangerous for you as well as for your child. Even knowing this possibility there are women who did smoked during their pregnancy back in 2011. As per the analysis from the state of Rhode Island(RI) most pregnant women were found smoking during their first three months of pregnancy. Minnesota, New Mexico and Wisconsin are the states followed by Rhode Island. These are the 4 states where women were smoked even though they were pregnant at that time.

Again, category named PRAMS total came first which shows that there are women who participated in this survey but did not mentioned their location. This number is still there specifically for PRAMS because number for different states can go higher if those women are included in the analysis. Data that was analyzed was taken when any woman visited the hospital for reasons such as Delivery, family planning, infant health, prenatal care etc.

**Question 3) How many women got pregnant before the age of 18 years and consumed alcohol or tobacco during their pregnancy in the United States?**

**[Tools Used: ‘Area chart’, ‘Pandas Data frame’]**

**Code:**

import pandas as pd

import matplotlib.pyplot as plt

preg=pd.read\_csv('file\_clean2.csv')

preg = preg.fillna('')

topic1= (preg.Topic == 'Alcohol Use') | (preg.Topic == 'Tobacco Use') & (preg.Break\_Out == 'Age < 18')

intersection = preg[topic1]

plot1 = intersection.groupby('LocationDesc').Topic.value\_counts().unstack()

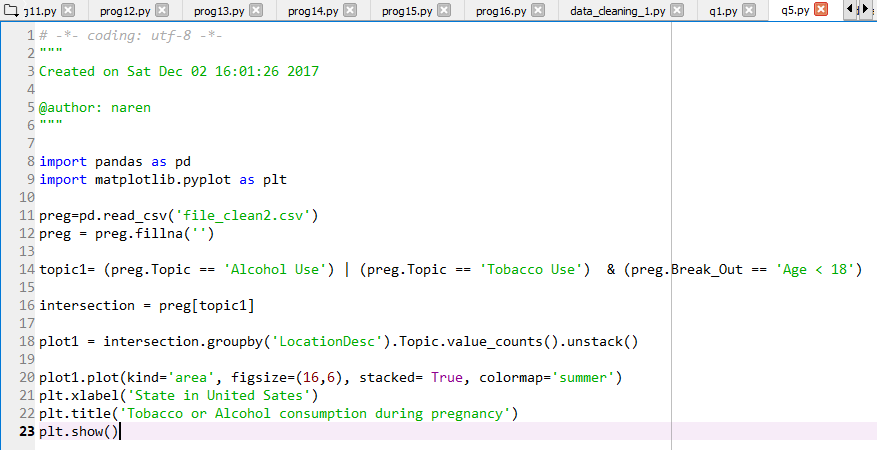
plot1.plot(kind='area', figsize=(16,6), stacked= True, colormap='summer')

plt.xlabel('State in United Sates')

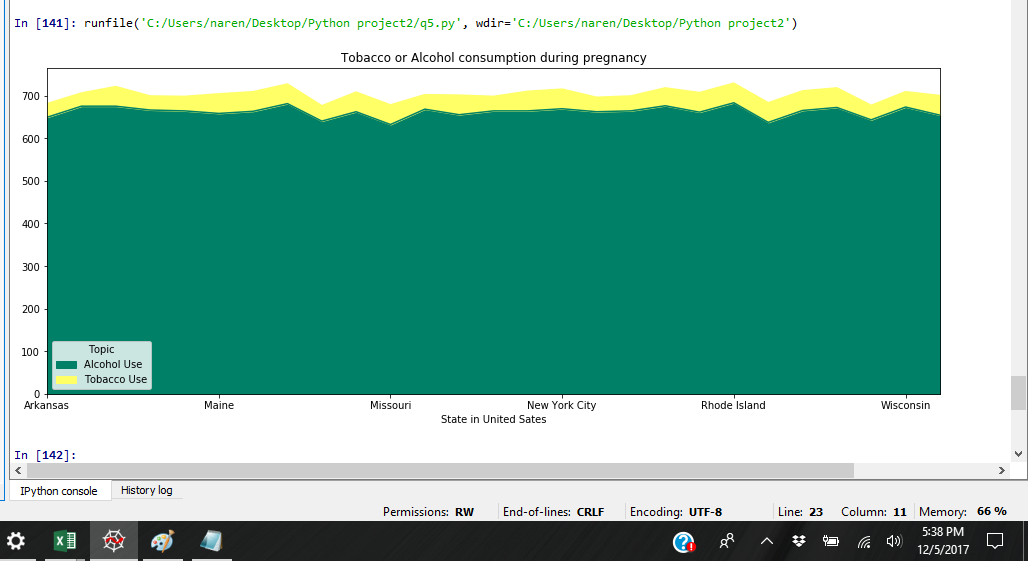
plt.title('Tobacco or Alcohol consumption during pregnancy')

plt.show()

**Code Screenshot:**



**Visualization**:



Teenage pregnancy is a big issue in the United States. Every year government allocates more than $50 million to avoid teenage pregnancy. (Youth, 2017) More than teenage pregnancy, most crucial issue here is the health of the child. Since, mother who is delivering the child is under age, child is in danger of many things. To add more to that danger, the pregnant women who consume alcohol or tobacco during their pregnancy. Above visualization represents the fact that there are more than 600 women/girls who are getting pregnant at the age less than 18 years old and consuming either alcohol or tobacco during pregnancy. States which has highest number of cases are Maine, Missouri and New York City. Considering that there are more than 600 girls found in just one New York city who consumed either tobacco or alcohol is scary. Public awareness is required for teenagers as there are so many cases of teenage pregnancy occurring in different states in USA.

Alcohol or tobacco consumption should be reduced to lower percentage than current situation. When a woman drinks alcohol while she is pregnant, the alcohol goes to the baby through her bloodstream. Drinking alcohol during pregnancy can cause fetal alcohol syndrome (FAS), a serious condition that can affect a child throughout life. (Health, 2002) This is the independent study presented by US National library of medicine.

# References

CDC. *Burden of Tobacco Use in the U.S.* 20 06 2017. Web article. 5 12 2017. <https://www.cdc.gov/tobacco/campaign/tips/resources/data/cigarette-smoking-in-united-states.html>.

Dimes, March of. *Abuse during pregnancy*. 2017. Web article. 30 11 2017. <https://www.marchofdimes.org/pregnancy/abuse-during-pregnancy.aspx>.

Health, Paediatr Child. *Fetal alcohol syndrome: What you should know about drinking during pregnancy*. 7 3 2002. web article. 5 12 2017. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2794811/>.

*PRAMS*. 30 08 2017. WEB ARTICLE. 12 11 2017. <https://www.cdc.gov/prams/index.htm>.

Youth. *New Teen Pregnancy Prevention Funding Opportunities*. 2017. web article. 5 12 2017. <https://youth.gov/feature-article/new-teen-pregnancy-prevention-funding-opportunities>.

**All code:**

**Data cleaning 1): Column names were in abbreviation format**

import pandas as pd

import matplotlib.pyplot as plt

preg=pd.read\_csv('CDC\_PRAMStat\_Data\_for\_2011.csv')

preg.dropna()

new\_label= ['Year', 'LocationAbbreviation', 'LocationDescription', 'Class', 'Topic',

'Question', 'DataSource', 'Response', 'Data\_Value\_Unit', 'Data\_Value\_Type',

'Data\_Value', 'Data\_Value\_Footnote\_Symbol', 'Data\_Value\_Footnote', 'Data\_Value\_Std\_Error',

'Low\_Confidence\_Limit', 'High\_Confidence\_Limit', 'Sample\_Size', 'Break\_Out',

'Break\_Out\_Category', 'Geolocation\_X', 'Geolocation\_Y', 'ClassID', 'TopicID', 'QuestionID',

'LocationID', 'BreakOutID', 'BreakOutCategoryID', 'ResponseID']

df=pd.read\_csv('CDC\_PRAMStat\_Data\_for\_2011.csv', header=0,names= new\_label)

print(df.head())

**Data Cleaning 2): In the question column, one question was repeated because it had no punctuation hence while querying the dataset output was getting calculated wrongly.**

import pandas as pd

import matplotlib.pyplot as plt

preg=pd.read\_csv('CDC\_PRAMStat\_Data\_for\_2011.csv')

preg = preg.fillna('')

preg['Question'].replace('(\*PCH) During the 12 months before you got pregnant did your husband or partner push hit slap, kick, choke, or physically hurt you in any other way?','(\*PCH) During the 12 months before you got pregnant, did your husband or partner push, hit, slap, kick, choke, or physically hurt you in any other way?', inplace=True)

print(preg.head())

#preg.to\_csv("file\_clean.csv",index=False)

**Data Cleaning 3): Geolocation were divided into two columns i.e. X-coordinate and Y-coordinate. Joined the two columns to create a single column. Also removed ‘NA’**

import pandas as pd

import matplotlib.pyplot as plt

preg=pd.read\_csv('CDC\_PRAMStat\_Data\_for\_2011.csv')

preg = preg.fillna('')

preg['Geolocation\_X']= preg.Geolocation\_X + "," + preg.Geolocation\_Y

del preg['Geolocation\_Y']

preg.to\_csv("file\_clean2.csv",index=False)

**#Code for Summary Statistics: (dropped NA’s and used Pandas Data frame)**

import pandas as pd

import matplotlib.pyplot as plt

preg=pd.read\_csv('CDC\_PRAMStat\_Data\_for\_2011.csv')

preg.dropna()

print('minimum' )

print(preg[('High\_Confidence\_Limit')].min())

print("maximum")

print(preg['High\_Confidence\_Limit'].max())

print("mean")

print(preg['High\_Confidence\_Limit'].mean())

print('Standard Deviation')

print(preg['High\_Confidence\_Limit'].std())

print('Describe')

print(preg['High\_Confidence\_Limit'].describe())

preg.to\_csv("file\_clean45.csv",index=False)

**Question1) How many women across United States said that they were physically abused during pregnancy?**

#Locationwise physical abuse during pregnancy using Tuple

import pandas as pd

import matplotlib.pyplot as plt

preg=pd.read\_csv('file\_clean2.csv')

preg = preg.fillna('')

topic1= (preg.Topic == 'Abuse - Physical')

intersection = preg[topic1]

explode = (0.1, 0, 0, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)

plot1 = intersection['LocationDesc'].value\_counts()

plot1.plot.pie(figsize=(10,10), shadow=True, startangle=90, explode=explode)

plt.title('Physical abuse during pregnancy')

plt.xlabel('Locations across United States')

plt.ylabel('')

**Question 2) State wise distribution of women who came in the hospital and stated that they smoked 3 months before pregnancy.**

#Question : statewise distribution of women who came in hospital and stated that they smoked 3 months before pregnancy

import pandas as pd

import matplotlib.pyplot as plt

preg=pd.read\_csv('file\_clean2.csv')

preg = preg.fillna('')

def format\_text(preg):

topic1= (preg.Break\_Out == 'Smoker') & (preg.Response == 'YES') | (preg.Response == 'YES (CHECKED)')

intersection = preg[topic1]

#plot1 = intersection['Topic'].value\_counts().unstack()

plot1 = intersection.groupby('LocationDesc').Class.value\_counts().unstack()

plot2 = plot1.sort\_values(by='Delivery', ascending= False)[:5]

plot2.plot(kind='bar',colormap='winter', figsize=(16,6))

plt.xlabel('Locations')

plt.title('Statewise distribution of women who came in hospital and stated that they smoked 3 months before pregnancy')

plt.show()

format\_text(preg)

**Question 3) How many women got pregnant before the age of 18 years and consumed alcohol or tobacco during their pregnancy in the United States?**

import pandas as pd

import matplotlib.pyplot as plt

preg=pd.read\_csv('file\_clean2.csv')

preg = preg.fillna('')

topic1= (preg.Topic == 'Alcohol Use') | (preg.Topic == 'Tobacco Use') & (preg.Break\_Out == 'Age < 18')

intersection = preg[topic1]

plot1 = intersection.groupby('LocationDesc').Topic.value\_counts().unstack()

plot1.plot(kind='area', figsize=(16,6), stacked= True, colormap='summer')

plt.xlabel('State in United Sates')

plt.title('Tobacco or Alcohol consumption during pregnancy')

plt.show()