Week08 Assignment

Exploratory Data Analysis with Python (pandas)

Venkata Sai Pallavi Pallapolu

Saint louis university

Course code: ORES-5160: data management

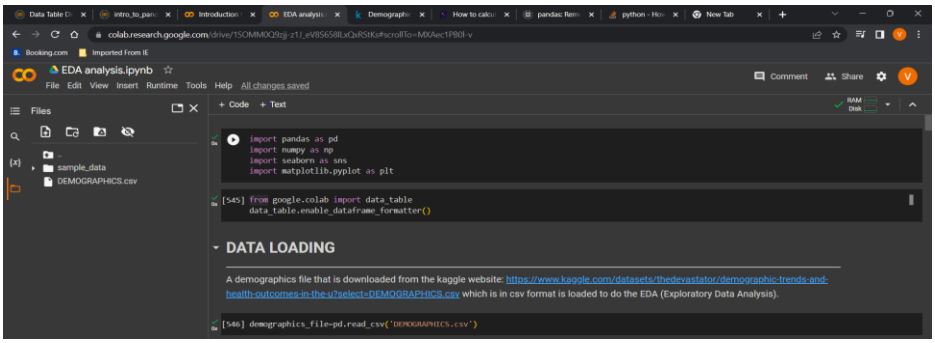
Prof. Jason Eden

October 26th, 2023

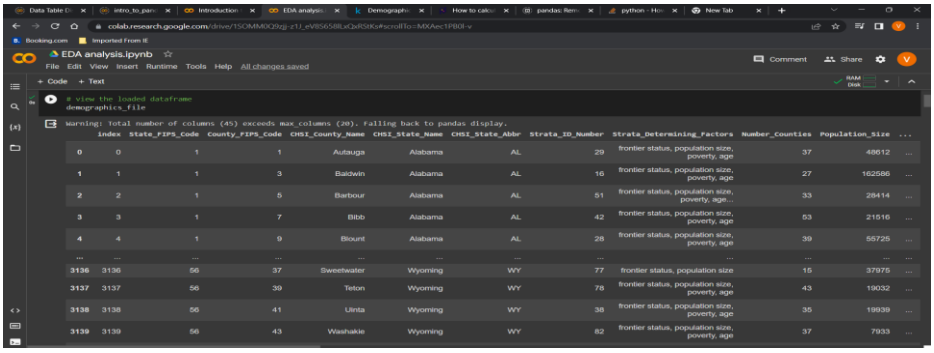
RESEARCH REPORT ON EXPLORATORY DATA ANALYSIS WITH PYTHON (PANDAS):

For Identifying outliers, trends and removing data outliers the Exploratorty data analysis is performed by following the steps like: loading the data, imputation of the data, identifying the outliers, data visualization and generating insights.

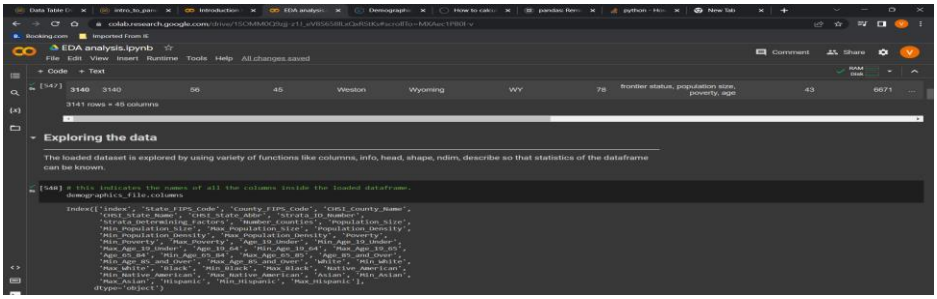
1. **Data loading:** demographics dataset is downloaded on local computer as a csv (comma-separated value) file and then loaded into the google colab platform by clicking on the upload option inside the file icon. Then, to perform the Exploratory Data Analysis on the loaded dataset pandas, NumPy, Seaborn, and Matplotlib libraries are imported in the code. The code is as follows:



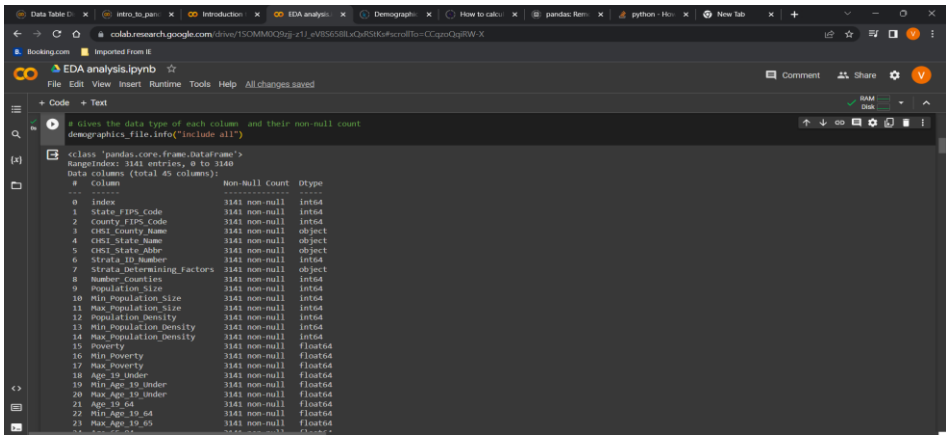
The loaded data Frame looks something like the picture below indicating all the series (columns) and their values. The bottom of the output cell indicates the total number of rows and columns which is 3141 rows and 45 columns in this case.



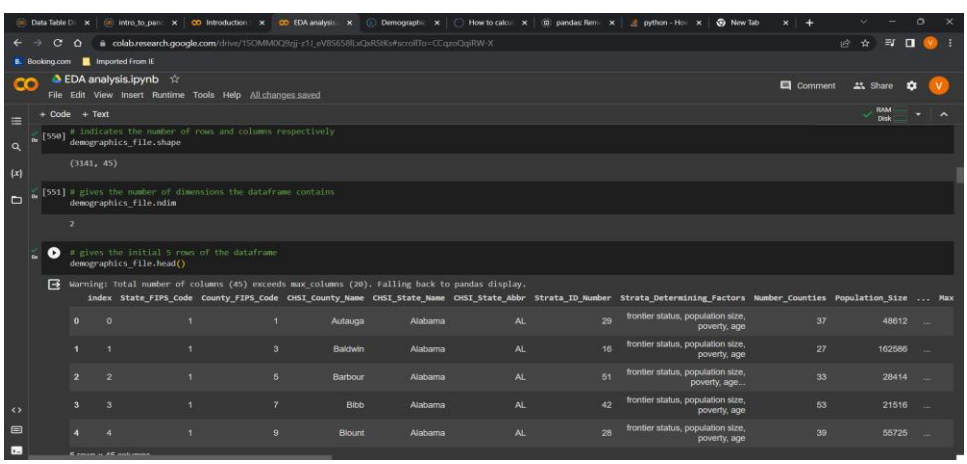
2. **Exploring the Data:** To know what the loaded dataframe contains, it is explored by using several functions:



The above picture indicates the code by using the function columns () where the output shows the names of all the series.

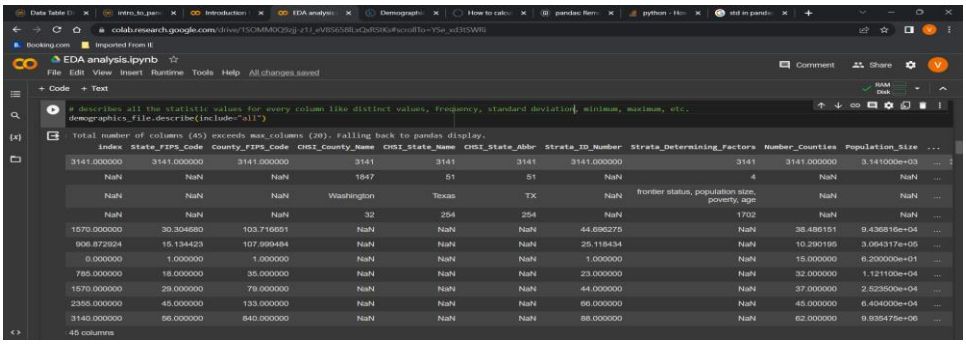


The above picture indicates the info function which gives an output including all the series names and their data types along with number of non-null count. Here, if a column shows unmatched datatype, then the datatype of that column can be changed to appropriate datatype by using dataframe.astype function () (for a numeric to category typecasting) or dataframe.apply(pd.to_datatype, errors='coerce') (for category or object to numeric data typecasting). When typecasting object to numeric datatype invalid numbers are converted to NaN when errors= 'coerce' is used (1). This dataframe has all appropriate column datatypes so changing the datatypes is not needed.

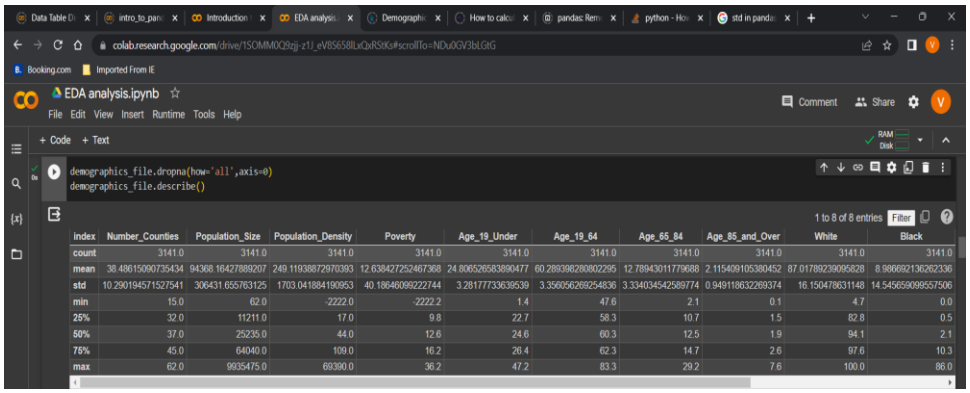


The above picture indicates different types of functions to get:

- The number of rows and columns: by using dataframe.shape() function
- Dimensions: by using dataframe.ndim() function- gives the number of dimensions in the dataframe
- Initial rows of the dataframe: by using dataframe.head() function- produces the first 5 rows of dataframe.



This above picture indicates the usage of describe () function to produce the output showing the statistics of each series that includes the count, distinct values, top value (if it's a category), frequency, mean, standard deviation, minimum value, 25th percent value, 50th percent value, 75th percent value, maximum value.



index	Number_Counties	Population_Size	Population_Density	Poverty	Age_19_Under	Age_19_64	Age_65_84	Age_85_and_Over	White	Black
count	3141.0	3141.0	3141.0	3141.0	3141.0	3141.0	3141.0	3141.0	3141.0	3141.0
mean	38.48615090735434	94368.16427889207	249.11938872970393	12.638427252467368	24.806526583890477	60.28939828080229 5	12.789430117796688	2.115409105380452	87.01789239095828	8.986689213626233 6
std	10.290194571527541	306431.655763125	1703.041884190953	40.18646099222744	3.28177733639539	3.356056269254836	3.334034542589774	0.949118632269374	16.150478631148	14.545659099557506
min	15.0	62.0	-2222.0	-2222.2	1.4	47.6	2.1	0.1	4.7	0.0
25%	32.0	11211.0	17.0	9.8	22.7	58.3	10.7	1.5	82.8	0.5
50%	37.0	25235.0	44.0	12.6	24.6	60.3	12.5	1.9	94.1	2.1
75%	45.0	64040.0	109.0	16.2	26.4	62.3	14.7	2.6	97.6	10.3
max	62.0	9935475.0	69390.0	36.2	47.2	83.3	29.2	7.6	100.0	86.0

The above picture indicates the usage of dropna() function. This is a function used to remove not a number (NaN) values, null values, or missing values. If any missing values are present in the dataframe to remove them, this function has been used. Also, this code indicates ‘axis=0’, which means rows that is to remove any row which has missing values (Axis=1 for columns).

```
#to know the unique states and their names and number of unique counties in each state
print(len(demographics_file.CHSI_State_Name.unique()))
print(demographics_file.CHSI_State_Name.unique())
print(len(demographics_file.CHSI_County_Name.unique()))
```

```
51
['Alabama' 'Alaska' 'Arizona' 'Arkansas' 'California' 'Colorado'
 'Connecticut' 'Delaware' 'District of Columbia' 'Florida' 'Georgia'
 'Hawaii' 'Idaho' 'Illinois' 'Indiana' 'Iowa' 'Kansas' 'Kentucky'
 'Louisiana' 'Maine' 'Maryland' 'Massachusetts' 'Michigan' 'Minnesota'
 'Mississippi' 'Missouri' 'Montana' 'Nebraska' 'Nevada' 'New Hampshire'
 'New Jersey' 'New Mexico' 'New York' 'North Carolina' 'North Dakota'
 'Ohio' 'Oklahoma' 'Oregon' 'Pennsylvania' 'Rhode Island' 'South Carolina'
 'South Dakota' 'Tennessee' 'Texas' 'Utah' 'Vermont' 'Virginia'
 'Washington' 'West Virginia' 'Wisconsin' 'Wyoming']
1847
```

The skewness of the dataset can be known by using skew () function. The dataframe that has been loaded is highly skewed as the columns indicated the skewness as 27.56 as highest and –54.80 as least so this dataset can be thought as highly skewed one (2).

```
skew=demographics_file.skew(axis=0)
print(skew)
```

```
Number_Counties        0.382221
Population_Size       15.273352
Population_Density    27.566985
Poverty              -54.805956
Age_19_Under           0.724751
Age_19_64              0.382587
Age_65_84              0.530835
Age_85_and_Over        1.155048
White                 -1.979011
Black                  2.244748
Native_American        7.862679
Asian                 10.770755
Hispanic               3.543323
dtype: float64
<ipython-input-681-4608871029a>:1: FutureWarning: The default value of numeric_only in DataFrame.skew is deprecated. In a future version, it will default to False. In addition, specif
  skew=demographics_file.skew(axis=0)
```

The above picture also indicates a code to print the distinct number of states from the dataframe, their names, and the number of distinct counties in those states.

```
# to know the relationship between population size and poverty
demographics_file[["Population_Size","Poverty"]].describe()
```

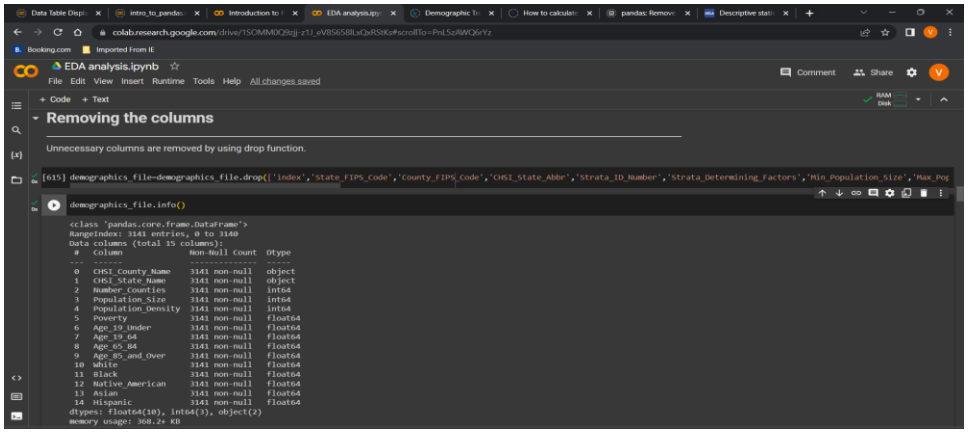
1 to 8 of 8 entries

index	Population_Size	Poverty
count	3141.0	3141.0
mean	94368.16427889207	12.638427252467368
std	306431.6557763125	40.18646099222744
min	62.0	-2222.2
25%	11211.0	9.8
50%	25235.0	12.6
75%	64040.0	16.2
max	9835475.0	36.2

Show 25 per page

The above picture shows the relationship between population size and poverty. That is for an average of 94368.16 population the poverty rate is 12.63. It also shows other descriptive statistical values.

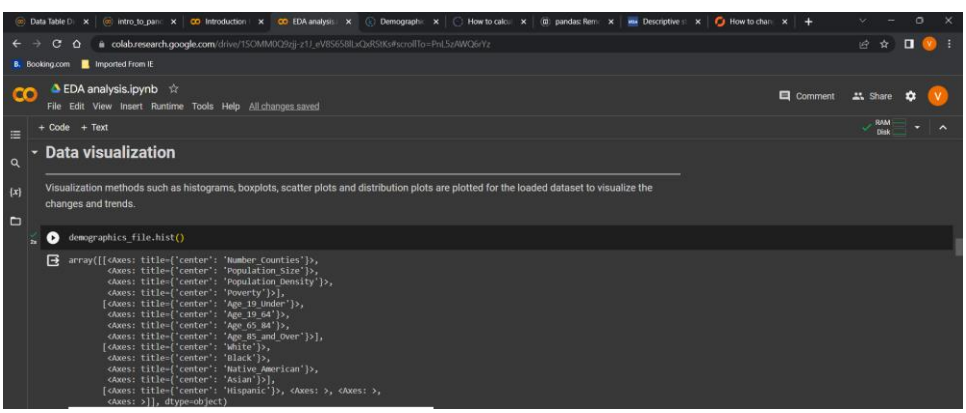
Removing the unnecessary columns from the dataframe:

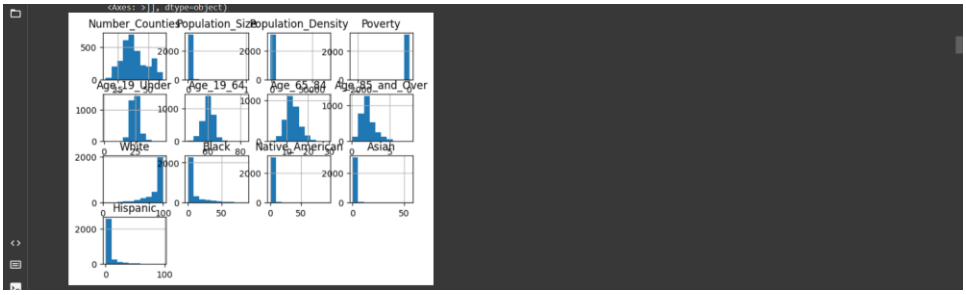


This above picture indicates the names of series and their datatypes after the removal of unnecessary columns.

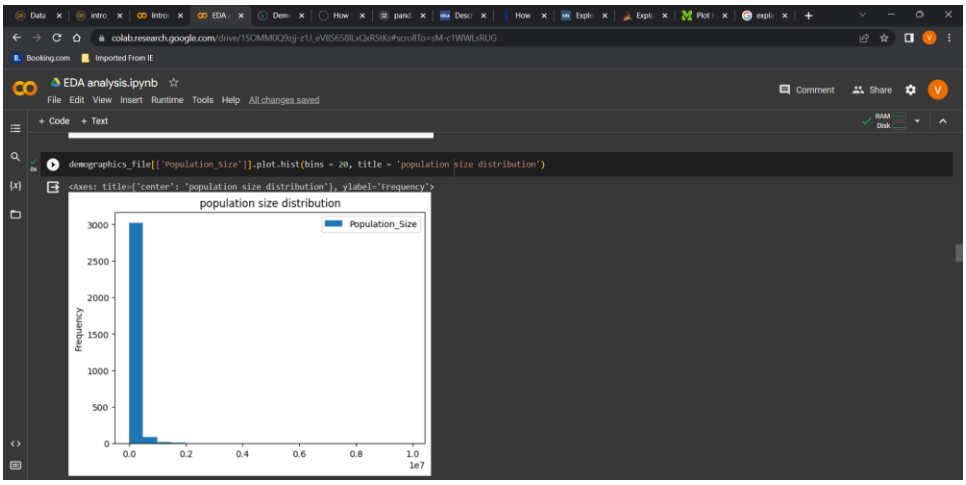
3. **Data visualization:** Visualizations can provide valuable insights into your data that might not be apparent from looking at the raw data alone (3). The various data visualization methods used in pandas for EDA are histograms, boxplots, scatterplots, pie charts, distribution charts and many more.

- HISTOGRAMS: Histograms are a great way to visualize the distribution of your data. They can provide insights into the central tendency, variability, and skewness of the data. These histograms divide the numerical values into bins and count the number of observations that fall into each bin (4).

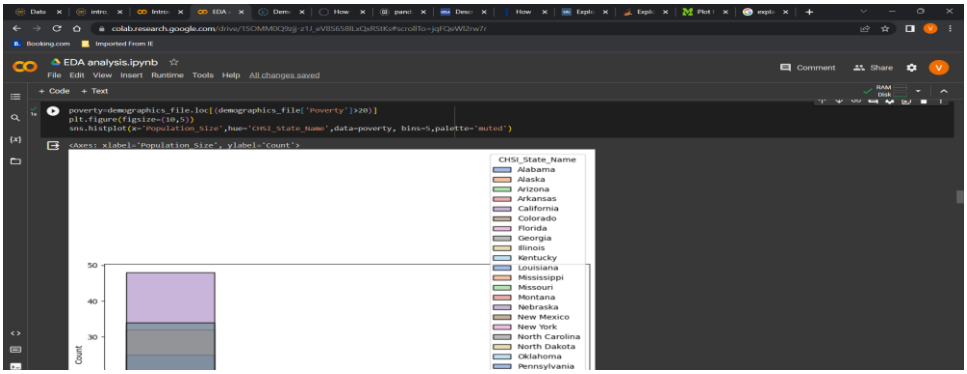


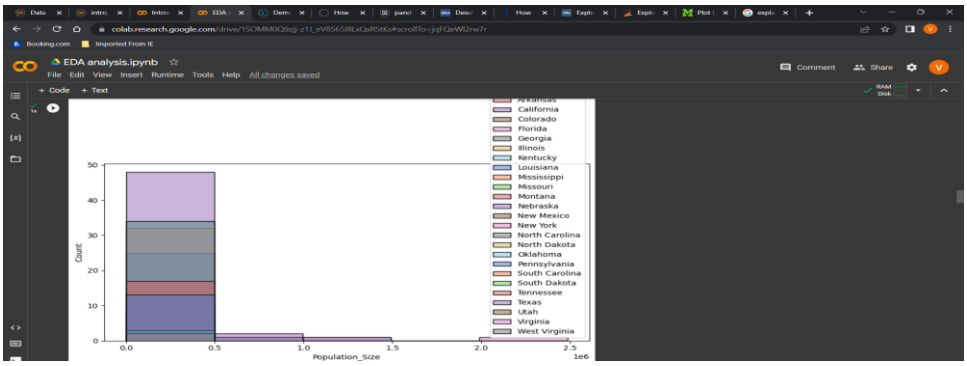


The above pictures indicate the histograms for all the useful columns in the dataframe by writing the code dataframe.hist() using hist () function.



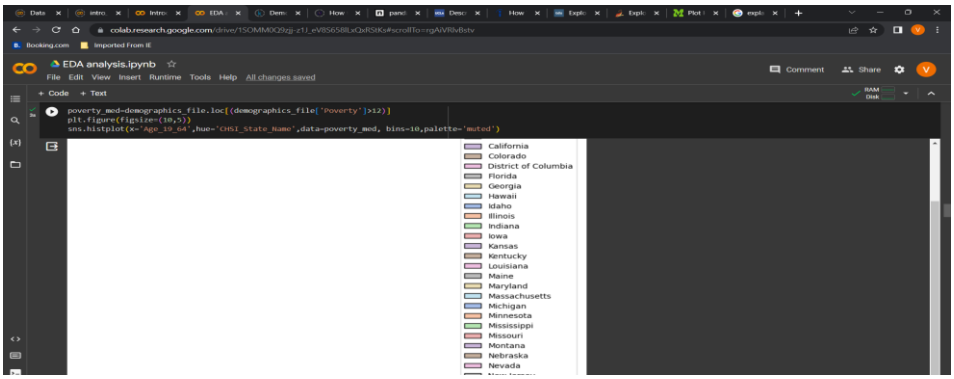
The above picture indicates the population size distribution that is the data that has been aggregated based on different population size intervals. The height of each bar represents the frequency of occurrences within that range. This histogram shows the asymmetric distribution suggesting highly skewness and uneven spread. From this histogram, population size with range 0.005 -0.1 can be identified as the most frequently occurring range (central tendency).

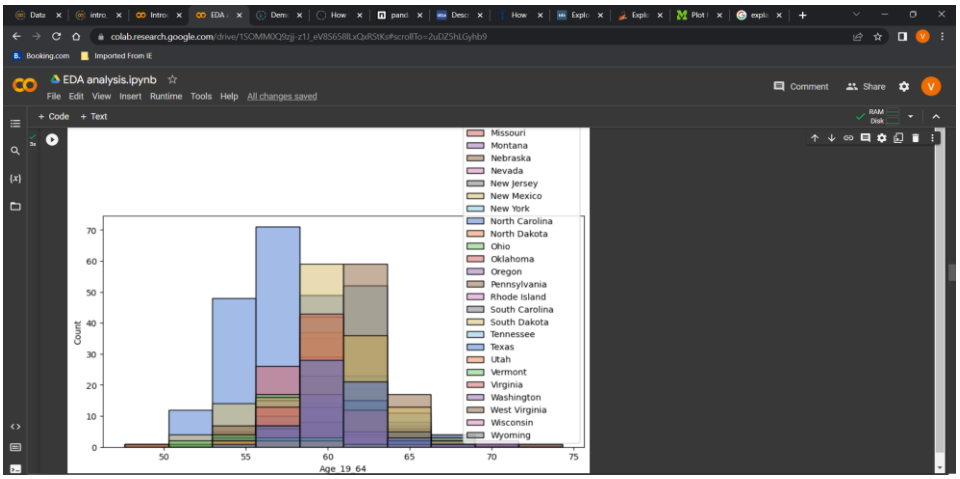




The above pictures produce the histogram that shows the distribution of population size with poverty greater than 20 categorized by different states by using seaborn library. X-axis represents the population size with state name as hue that means histogram will display different colors for each state and y-axis represents the count of counties falling within each population size range and the number of intervals or bins are 5 for this histogram. By looking at the histogram, it is clear

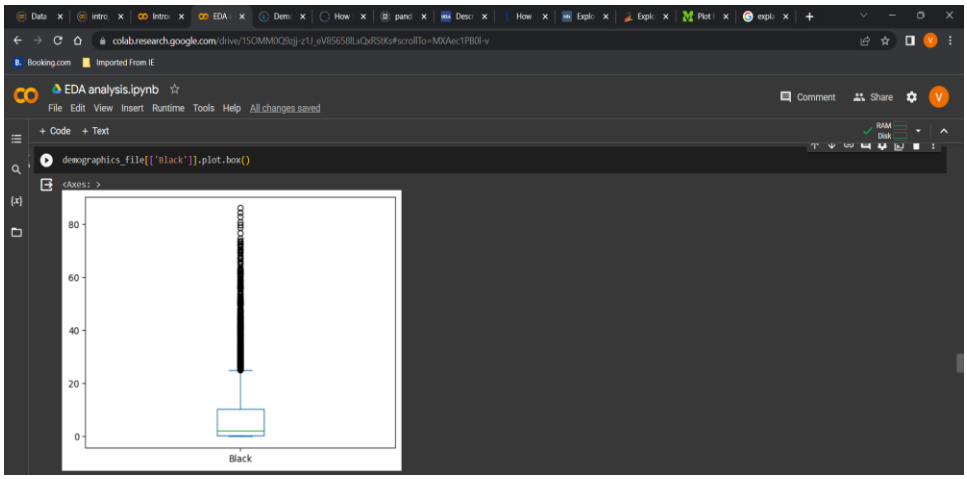
that ‘Nebraska’ has a greater number of counties with poverty greater than 20 with population size under 500000.



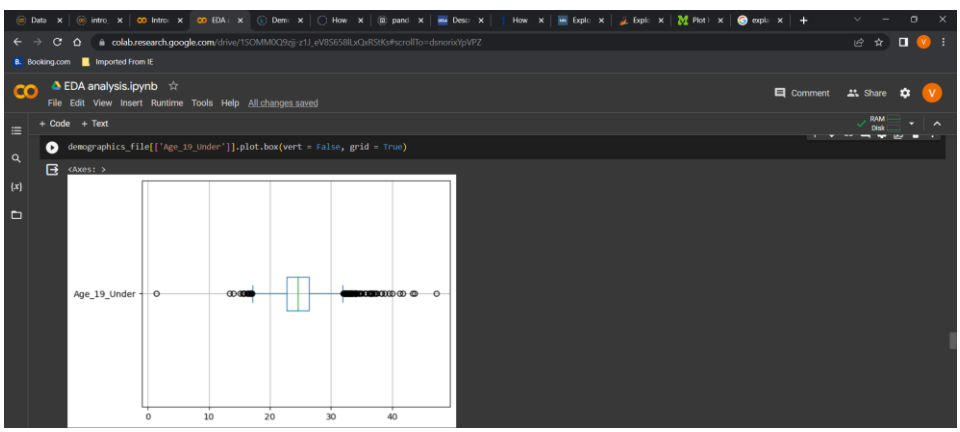


The above pictures indicate the distribution of age 19-64 people with poverty greater than 12 categorized by different states. The code also used a filter condition that is poverty greater than 12 and replaced it in the data. In the code palette=’muted’ indicates the color palette. From the histogram, it can be said that ‘Alabama’, ‘Texas’ have the highest number of people in age groups 19-64 .

- **Boxplots:** boxplots can tell you about outliers and what their values are. They can also tell you if a data is symmetrical, how tightly the data is grouped and if and how the data is skewed (5).

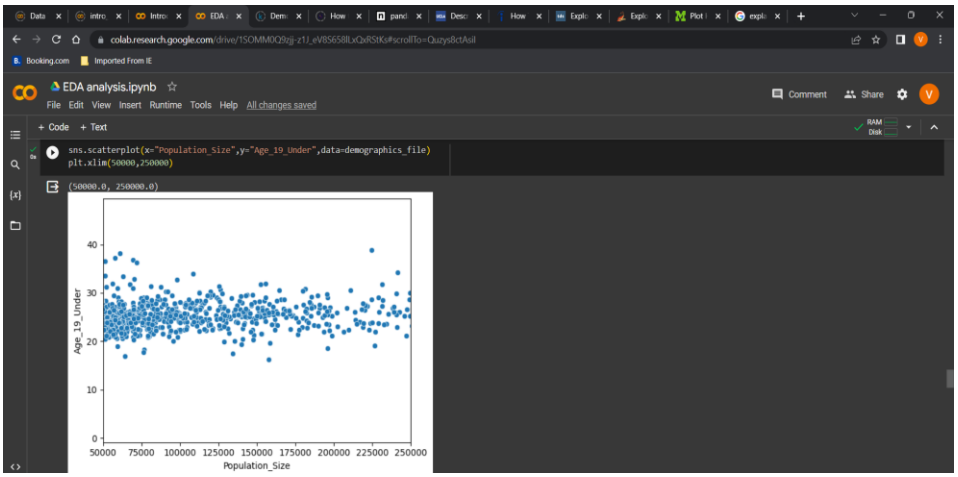


The above picture shows the code by using plot.box() function to get the distribution, central tendency and spread of 'Black' variable (race) from demographics file that is the count of population within a state. The line in the middle of the box indicates the median value of the data. The bottom of the box represents the first quartile, and the top indicates the third quartile while the box indicates the interquartile range. The points beyond the black line (Whiskers) are considered as outliers.

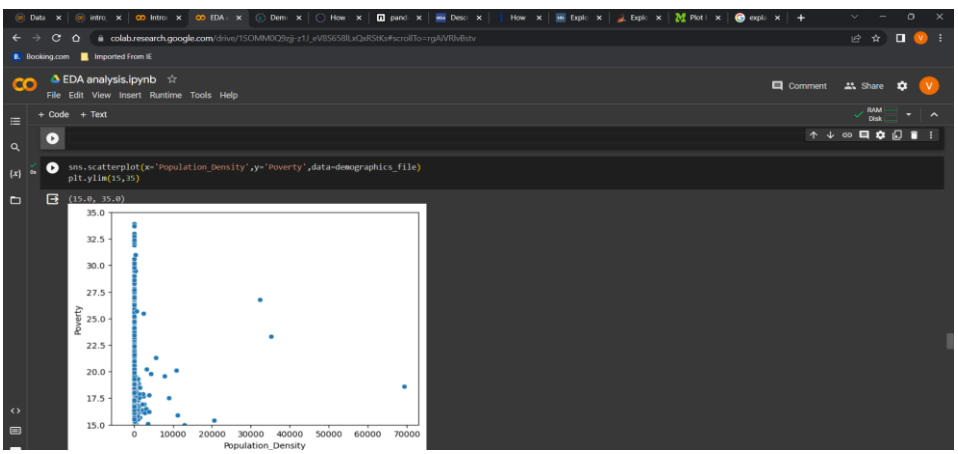


The above picture represents another format for box plot (6) to represent the plot in horizontal direction by using plot.box(vert=false and grid=true) meaning the plot should be in grid format not vertical. The box plot that resulted contains the distribution of 'age 19 and under' variable from demographics file, its median, minimum value, maximum value and outliers. From the boxplot, it is clear that there are a median of 23 people in each county under age 19.

- **Scatter plots:** A scatter plot uses dots to represent values for two different numeric variables. Each dot represents a single tree; each point's horizontal position indicates that tree's diameter (in centimeters) and the vertical position indicates that tree's height (in meters) (7).

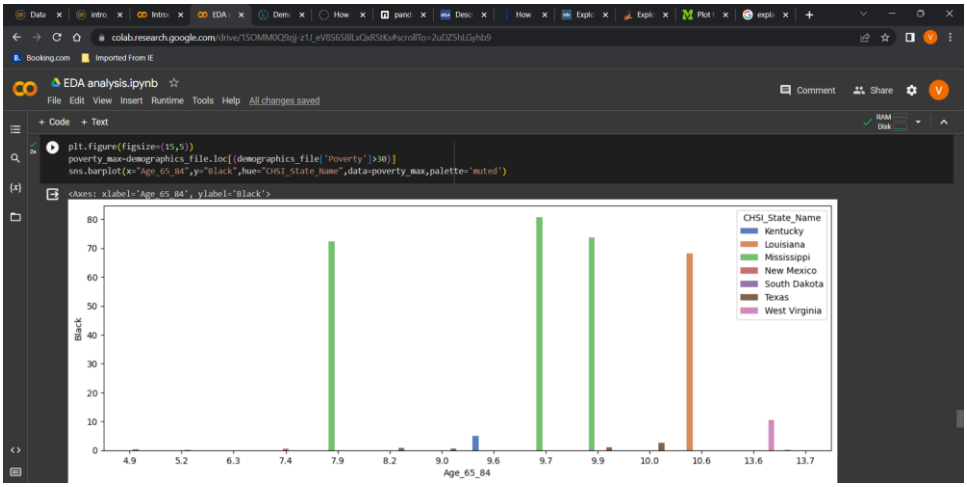


The above picture indicates the scatter plot between two variables they are: ‘age 19 under’ and ‘population size’. The code is written by using seaborn library and gives x-axis a parameter: population size and on y-axis: age 19 under is represented. And also, the limit for population size is set between 50000 and 250000. 20-30 range in age 19 under is the frequent range for population size under age 19 for this scatter plot.



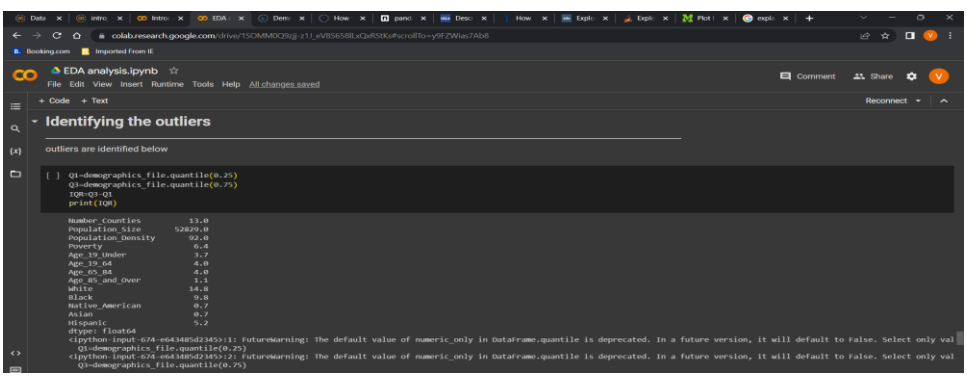
The above picture indicates the scatter plot which on x-axis have population density of various states living in 15-35 range of poverty and on y-axis poverty is represented. It also has a limit of 15 to 35. This scatterplot explains the relationship between poverty and population density.

- **Bar plots**: A bar plot or bar chart is a graph that represents the category of data with rectangular bars with lengths and heights that is proportional to the values which they represent (8).

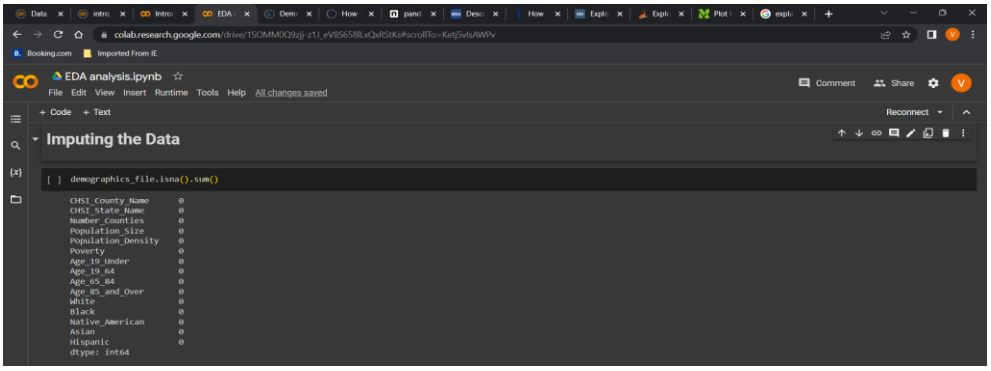


The above picture indicates the bar plot by using matplotlib, seaborn libraries which indicates the age 65-84 on x-axis and black column on y-axis and hue is state name the condition here is poverty greater than 30. Distribution of black with age 65-84 from various states having poverty more than 30 is indicated in this bar chart.

4. **Identifying the outliers:** The outliers can be identified by calculating the inter quartile range of the demographics file which is the difference between third quartile and first quartile. The below picture indicates the interquartile range for all the columns and their values are as follows:



5. **Data imputation:** In statistics, imputation refers to the process of substituting alternate values for missing data. It is known as "unit imputation" when replacing a data point and "item imputation" when replacing a data point's component (8).



This picture indicates that there are no null values in the data by using the function .isna().sum() so, there is no need to replace the null values.

6.**Gathering insights:** on performing the Exploratory Data Analysis, this dataset has provided with some insights like understanding the distribution and characteristics of counties within different states, poverty rate of various states with different population size, different ethnic composition like white, black, native American, Hispanic, Asian to understand the ethnic diversity and distribution across various states and also proportions of various age groups and their distribution.

Conclusion:

By performing the above-mentioned steps such as data loading, exploring the data, data visualization, data imputation, identifying the outliers, and gathering insights, valuable insights into the demographic and socioeconomic characteristics of the population across various states and counties are drawn. The study revealed trends in population size, population density, age distribution, poverty rates, and ethnic diversity. Several visualizations, such as histograms, boxplots, scatter plots, and bar plots aided in the identification of patterns, outliers, and potential correlations in the data.

This EDA provides a full overview of the demographics dataset's important aspects and informs possibilities for more specialized study and decision-making procedure.

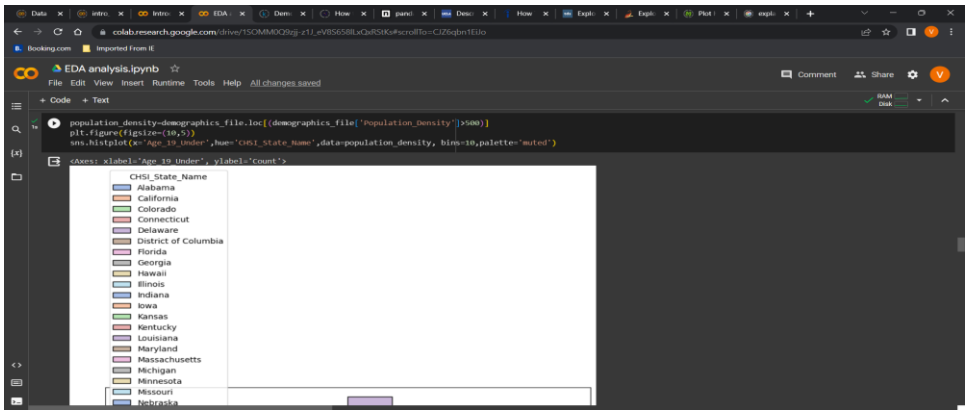
References:

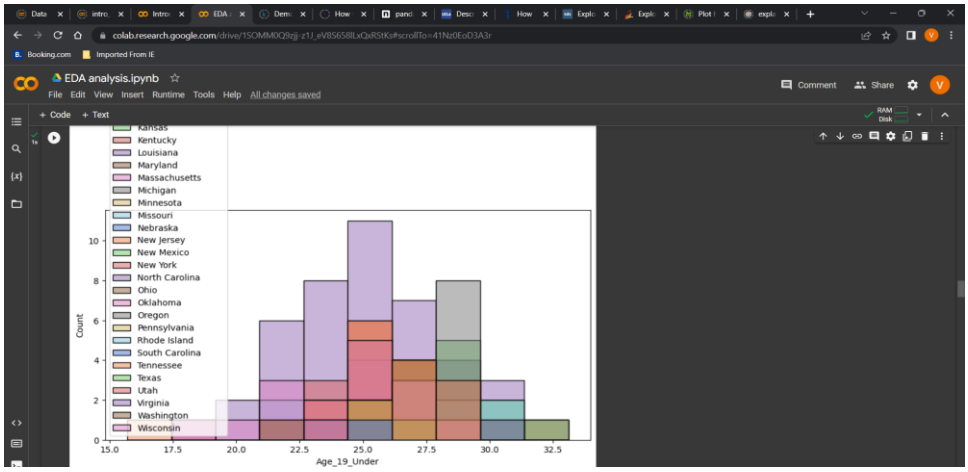
1. https://saturncloud.io/blog/pandas-tips-change-column-type/
2. https://medium.com/@atanudan/kurtosis-skew-function-in-pandas-aa63d72e20de#:~:text=skewness()%20function%20in%20pandas,present%20in%20the%20DataFrame%20object.
3. https://docs.kanaries.net/articles/exploratory-data-analysis-python-pandas
4. https://mode.com/example-gallery/python_histogram/#:~:text=A%20histogram%20divides%20the%20values,of%20values%20within%20a%20variable.
5. https://builtin.com/data-science/boxplot
6. https://towardsdatascience.com/exploratory-data-analysis-eda-visualization-using-pandas-ca5a04271607

7. https://chartio.com/learn/charts/what-is-a-scatter-plot/#:~:text=What%20is%20a%20scatter%20plot,to%20observe%20relationships%20between%20variables.
8. https://www.simplilearn.com/data-imputation-article#:~:text=ProgramExplore%20Program-,What%20Is%20Data%20Imputation%3F,from%20a%20dataset%20each%20time.

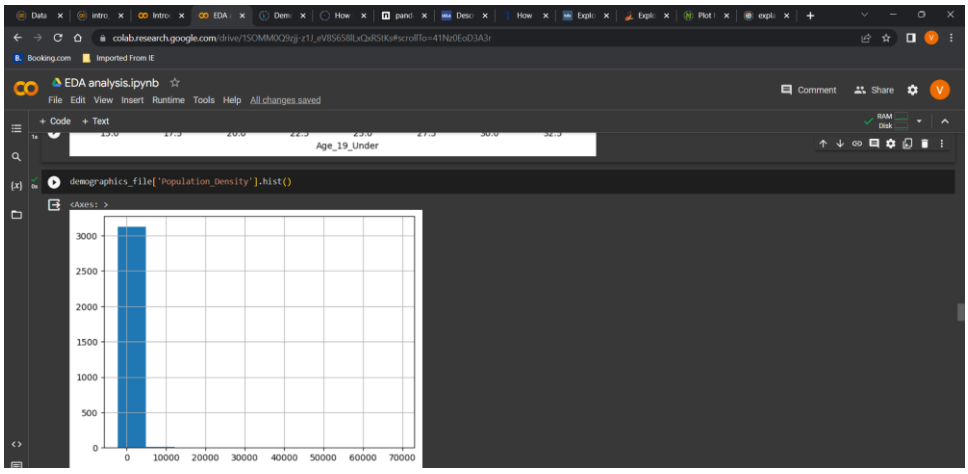
APPENDIX:

- Histogram indicating the distribution of age 19 under with population density greater than 500 over different states.

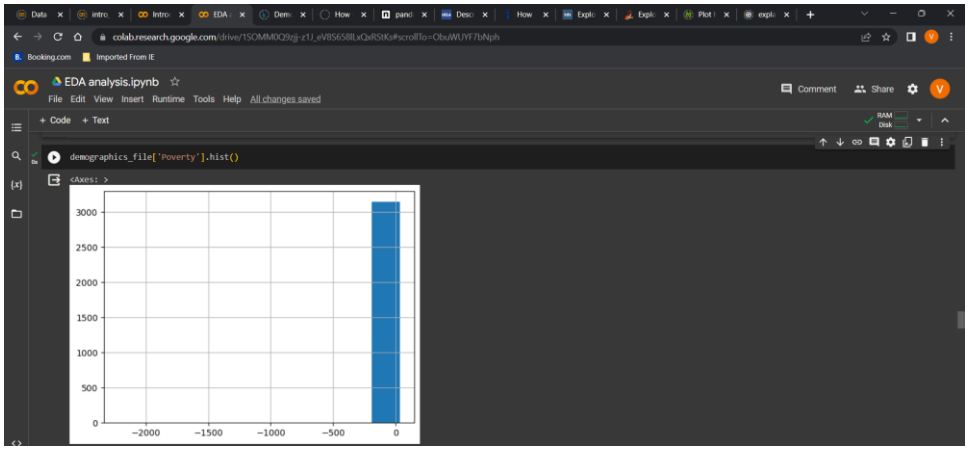




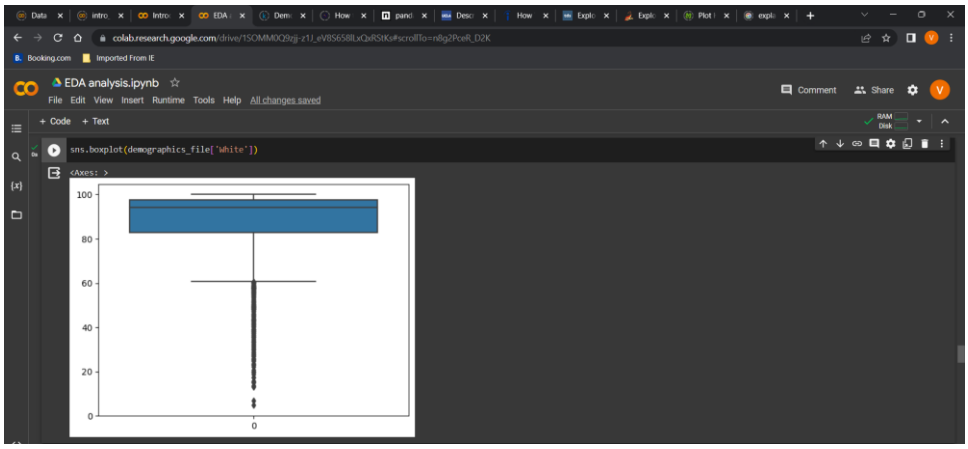
- Histogram indicating the distribution of population density



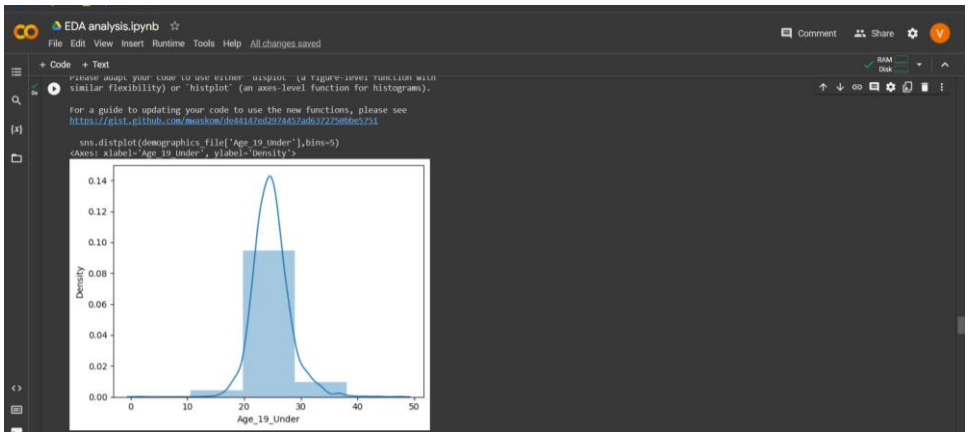
- Histogram indicating the distribution of poverty



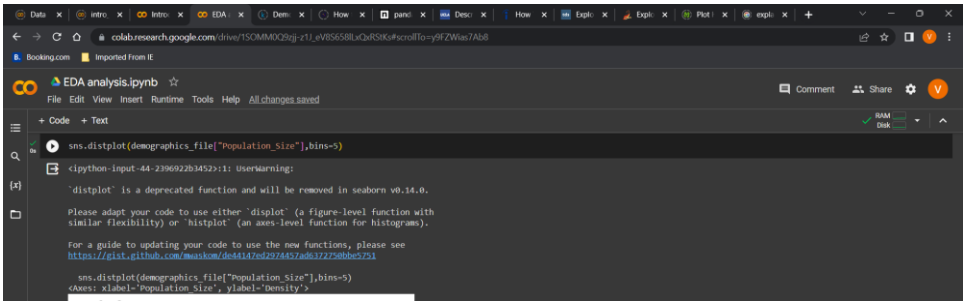
- Box plot indicating the distribution of white column from dataframe

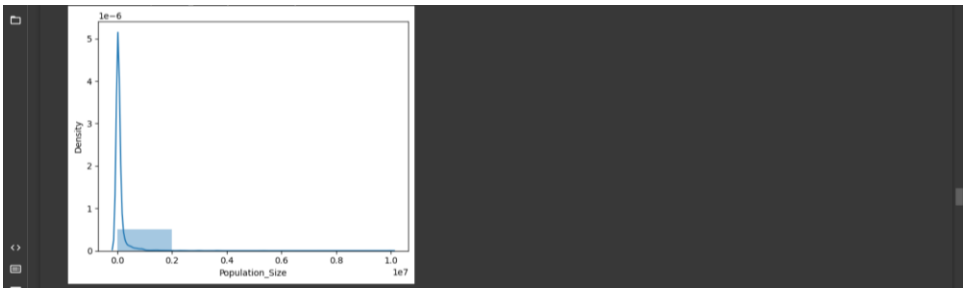


- Distribution plot indicating the distribution of age 19 under

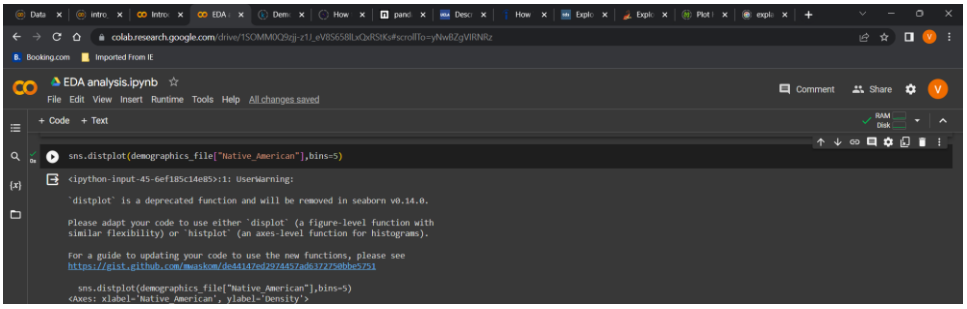


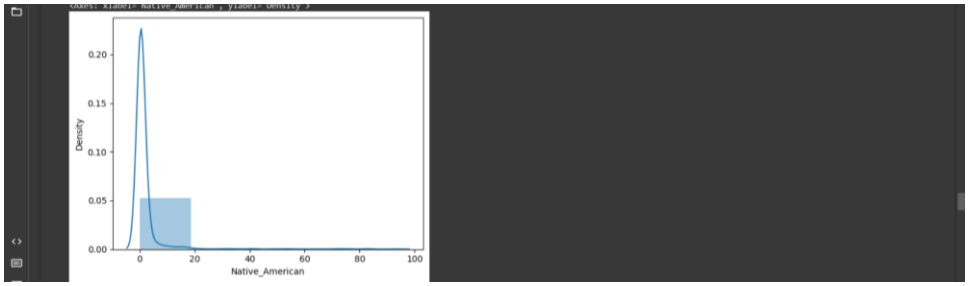
- Distribution plot indicating the distribution of population size column



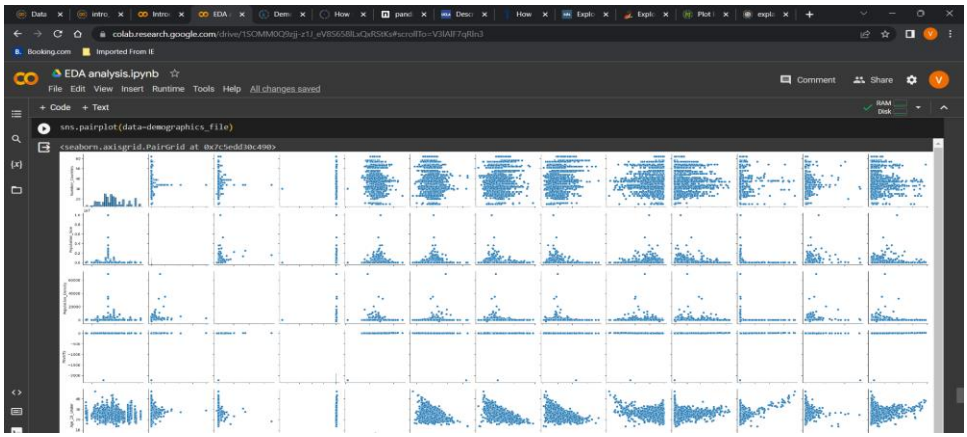


- Distribution plot indicating the distribution of native american column from dataframe

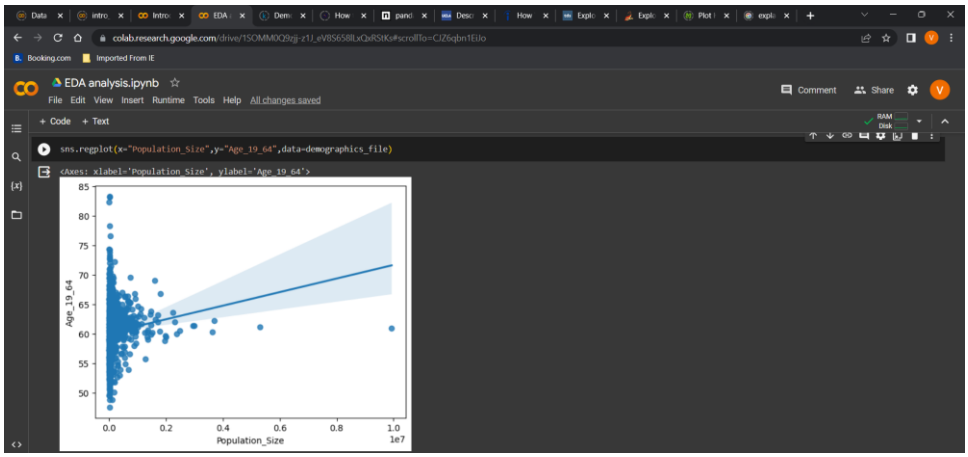




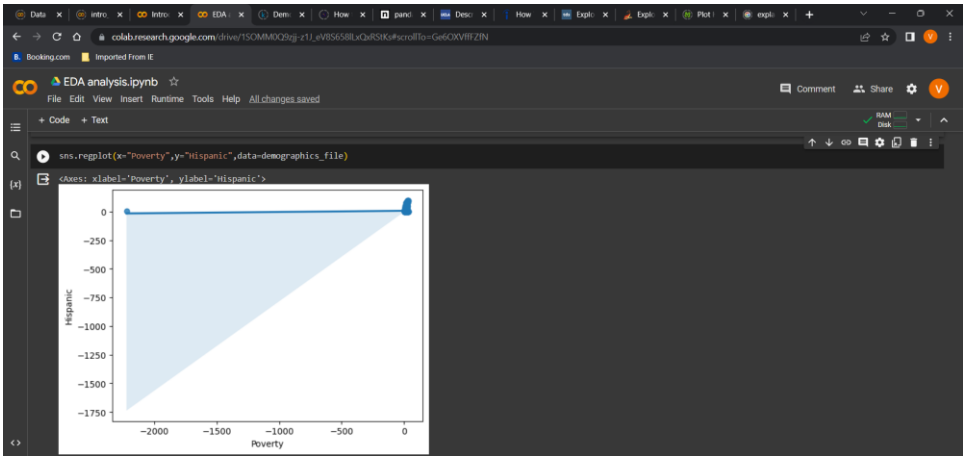
- Pairplot of all the columns



- Regression plot of population size and age 19-64 on x-axis and y-axis respectively



- Regression plot for distribution of poverty and hispanic on x-axis and y-axis respectively



- Correlation plot for population size, age 19-64 and poverty

EDA analysis.ipynb

File Edit View Insert Runtime Tools Help All changes saved

the below method is referred from this website https://towardsdatascience.com/exploratory-data-analysis-eda-visualization-using-pandas-ca5a04271607

```
[54] corr = demographics_file[['Population_Size','Age_19_64','Poverty']].corr()
     corr.style.background_gradient(cmap='BrBG').set_precision(2)
```

```
<ipython-input-54-38758c9ce841>:2: FutureWarning: this method is deprecated in favour of `Styler.format(precision=..)`
  corr.style.background_gradient(cmap='BrBG').set_precision(2)
```

	Population_Size	Age_19_64	Poverty
Population_Size	1.00	0.10	-0.00
Age_19_64	0.10	1.00	-0.14
Poverty	-0.00	-0.14	1.00