

COVID-19 US County JHU Data & Demographics

Introduction:

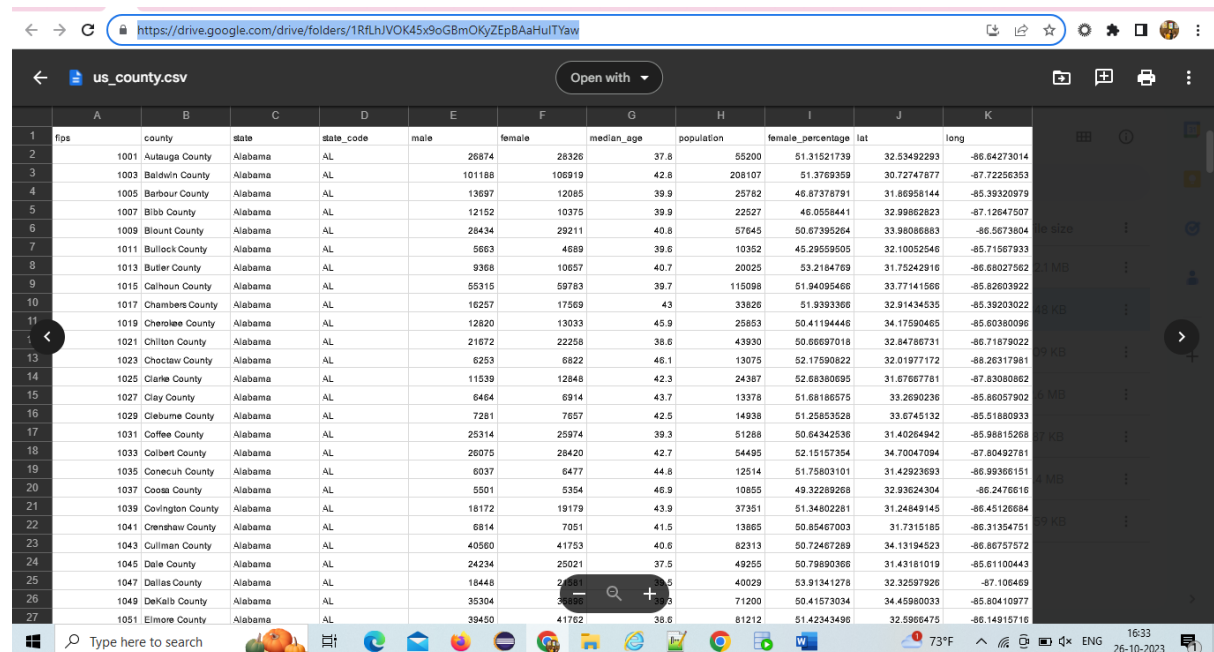
The United States of America has recently, had the most reported COVID-19 cases and this dataset that I have taken gives a piece of detailed information about the country, state, male, female, age group, and demographics information such as latitude and longitude. To perform this research, I used this dataset.

DATASET LINK:

<https://drive.google.com/drive/folders/1RfLhJVOK45x9oGBmOKyZEpBAaHuITYaw>

US_COUNTY.CSV

The main objective of this analysis is to find out the patterns within the dataset to get a further understanding of the data. I also wanted to leverage it to choose a machine algorithm for predicting the survival rate of patients during the period of COVID-19.



	A	B	C	D	E	F	G	H	I	J	K
1	fips	county	state	state_code	male	female	median_age	population	female_percentage	lat	long
2	1001	Autauga County	Alabama	AL	26874	28326	37.8	55200	51.31521739	32.53492293	-86.64273014
3	1003	Baldwin County	Alabama	AL	101188	106919	42.8	208107	51.3769359	30.72747877	-87.72256353
4	1005	Barbour County	Alabama	AL	13697	12085	39.9	25782	46.87378791	31.86958144	-85.39320979
5	1007	Bibb County	Alabama	AL	12152	10375	39.9	22527	46.0558441	32.99862623	-87.12647507
6	1009	Blount County	Alabama	AL	28434	29211	40.8	57645	50.87395264	33.98066883	-86.56738004
7	1011	Bullock County	Alabama	AL	5663	4689	39.6	10352	45.29559505	32.10052548	-85.71587933
8	1013	Butler County	Alabama	AL	9368	10657	40.7	20025	53.2184769	31.75242916	-86.66027562
9	1015	Calhoun County	Alabama	AL	55315	59783	39.7	115098	51.94095486	33.77141566	-85.82603922
10	1017	Chambers County	Alabama	AL	16257	17569	43	33826	51.9393366	32.91434535	-85.39203022
11	1019	Cherokee County	Alabama	AL	12820	13033	45.9	25853	50.41194448	34.17590465	-85.60380096
12	1021	Chilton County	Alabama	AL	21672	22256	38.6	43930	50.66697018	32.84786731	-86.71878022
13	1023	Choctaw County	Alabama	AL	6253	6822	46.1	13075	52.17590822	32.01977172	-88.26317981
14	1025	Clarke County	Alabama	AL	11539	12846	42.3	24387	52.68306095	31.87667781	-87.83080862
15	1027	Clay County	Alabama	AL	6464	6914	43.7	13378	51.68186575	33.2690236	-85.86057902
16	1029	Cleburne County	Alabama	AL	7281	7657	42.5	14938	51.25853528	33.6745132	-85.51880933
17	1031	Coffee County	Alabama	AL	25314	25974	39.3	51288	50.64342536	31.40264942	-85.98815288
18	1033	Colbert County	Alabama	AL	26075	28420	42.7	54495	52.15157354	34.70047094	-87.80492781
19	1035	Conecuh County	Alabama	AL	6037	6477	44.8	12514	51.75803101	31.42923693	-86.99366151
20	1037	Cook County	Alabama	AL	5501	5554	46.9	10855	49.32289268	32.93624304	-86.2478616
21	1039	Covington County	Alabama	AL	18172	19179	43.9	37351	51.34802281	31.24849145	-86.45126684
22	1041	Crenshaw County	Alabama	AL	6814	7051	41.5	13865	50.85467003	31.7315185	-86.31354751
23	1043	Cullman County	Alabama	AL	40560	41753	40.6	82313	50.72467289	34.13194523	-86.86757572
24	1045	Dale County	Alabama	AL	24234	25021	37.5	49255	50.79890366	31.43181019	-85.61100443
25	1047	Dallas County	Alabama	AL	14448	15445	40.5	40029	53.91341278	32.32597926	-87.106469
26	1049	DeKalb County	Alabama	AL	35304	36775	39.3	71200	50.41573034	34.45980033	-85.80410977
27	1051	Elmore County	Alabama	AL	39450	41762	38.6	81212	51.42343496	32.5986475	-86.14915716

The dataset consists of demographic information population information (Such as male and female rates) and age information.

Data attributes: Fips, County, State, State code, male, female, median age, population, female_percentage, lat, long.

So totally my dataset has 3220 rows * 11 columns with no null values. The columns have a title/heading, which makes them readable.

	A	B	C	D	E	F	G	H	I	J	K
	fips	county	state	state_code	male	female	median_age	population	female_percent	lat	long
2	1001	Autauga County	Alabama	AL	26874	28326	37.8	55200	51.31521739	32.53492293	-86.64273014
3	1003	Baldwin County	Alabama	AL	101188	106919	42.8	208107	51.3769359	30.72747877	-87.72256353
4	1005	Barbour County	Alabama	AL	13697	12085	39.9	25782	46.87378791	31.86958144	-85.39320979
5	1007	Bibb County	Alabama	AL	12152	10375	39.9	22527	46.0558441	32.99862823	-87.12647507
6	1009	Blount County	Alabama	AL	28434	29211	40.8	57645	50.67395264	33.98086883	-86.5673804
7	1011	Bullock County	Alabama	AL	5663	4689	39.6	10352	45.29559505	32.10052546	-85.71567933
8	1013	Butler County	Alabama	AL	9368	10657	40.7	20025	53.2184769	31.75242916	-86.68027562
9	1015	Calhoun County	Alabama	AL	55315	59783	39.7	115098	51.94095466	33.77141566	-85.82603922
10	1017	Chambers Coun	Alabama	AL	16257	17569	43	33826	51.9393366	32.91434535	-85.39203022
11	1019	Cherokee Count	Alabama	AL	12820	13033	45.9	25853	50.41194446	34.17590465	-85.60380096
12	1021	Chilton County	Alabama	AL	21672	22258	38.6	43930	50.66697018	32.84786731	-86.71879022
13	1023	Choctaw County	Alabama	AL	6253	6822	46.1	13075	52.17590822	32.01977172	-88.26317981
14	1025	Clarke County	Alabama	AL	11539	12848	42.3	24387	52.68380695	31.67667781	-87.83080862
15	1027	Clay County	Alabama	AL	6464	6914	43.7	13378	51.68186575	33.2690236	-85.86057902
16	1029	Cleburne County	Alabama	AL	7281	7657	42.5	14938	51.25863528	33.6745132	-85.51880933
17	1031	Coffee County	Alabama	AL	25314	25974	39.3	51288	50.64342536	31.40264942	-85.98815268
18	1033	Colbert County	Alabama	AL	26075	28420	42.7	54495	52.15157354	34.70047094	-87.80492781
19	1035	Conecuh County	Alabama	AL	6037	6477	44.8	12514	51.75803101	31.42923693	-86.99366151
20	1037	Coosa County	Alabama	AL	5501	5354	46.9	10855	49.32289268	32.93624304	-86.2476616
21	1039	Covington Count	Alabama	AL	18172	19179	43.9	37351	51.34802281	31.24849145	-86.45126684

Observations of the dataset:

- It has all the states in the United States of America.
- The data includes patients whose ages range from 30 to 60.
- The data also contains fips code, latitude, and longitude details for easy understanding of the location details.

Dataset and Code Description:

This data contains the total population, male and female.

Explanation: This code helps us to know the total count of males from different states.

```
[10] print(data_frame["male"].value_counts)

<bound method IndexOpsMixin.value_counts of 0      26874
1      101188
2      13697
3      12152
4      28434
...
3215    25580
3216     4332
3217    11169
3218    16541
3219    17475
Name: male, Length: 3220, dtype: int64>
```

Explanation: This code helps us to know the total count of females from different states.

```

[11] print(data_frame["female"].value_counts)

<bound method IndexOpsMixin.value_counts of 0      28326
1      106919
2      12085
3      10375
4      29211
...
3215    27791
3216     4439
3217    11824
3218    17608
3219    18964
Name: female, Length: 3220, dtype: int64>

```

Explanation: This code helps us to know the total count of population from different state

```

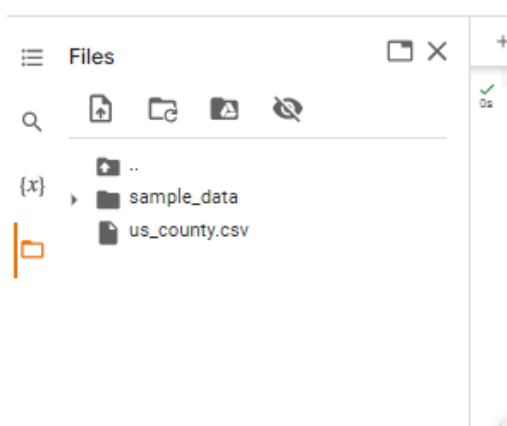
print(data_frame['population'].value_counts)

<bound method IndexOpsMixin.value_counts of 0      55200
1      208107
2      25782
3      22527
4      57645
...
3215    53371
3216     8771
3217    22993
3218    34149
3219    36439
Name: population, Length: 3220, dtype: int64>

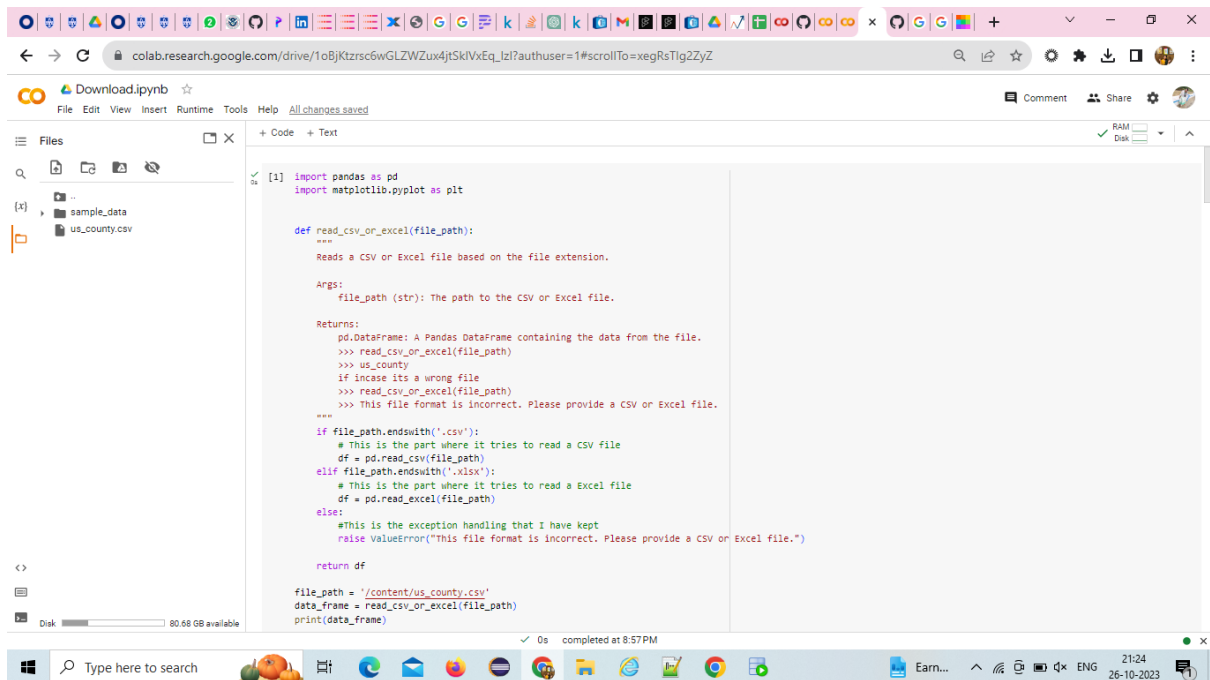
```

Important note:

Before performing this code, we need to down the dataset and upload it in the Google Colab environment.



Code: This code helps me to read a CSV or Excel file in order to due EDA



```
[1] import pandas as pd
import matplotlib.pyplot as plt

def read_csv_or_excel(file_path):
    """
    Reads a CSV or Excel file based on the file extension.

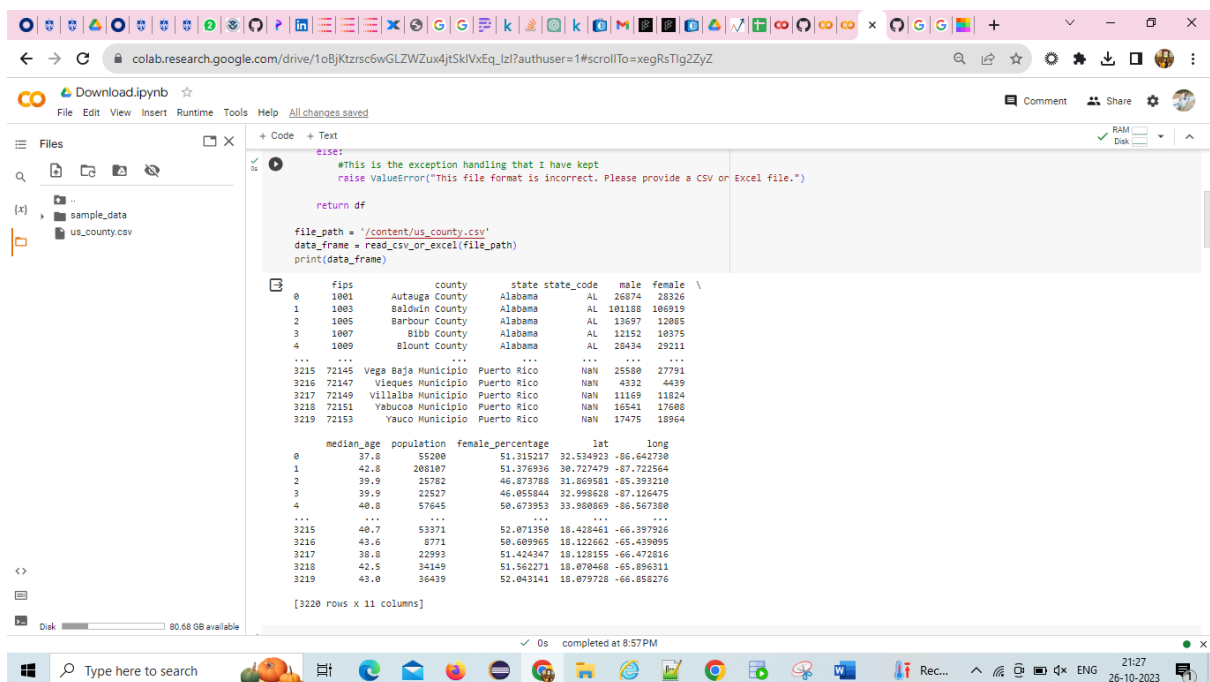
    Args:
        file_path (str): The path to the CSV or Excel file.

    Returns:
        pd.DataFrame: A Pandas DataFrame containing the data from the file.
    """
    df = pd.read_csv(file_path)
    if isinstance(df, pd.DataFrame):
        return df
    elif file_path.endswith('.xlsx'):
        df = pd.read_excel(file_path)
    else:
        raise ValueError("This file format is incorrect. Please provide a CSV or Excel file.")

    return df

file_path = '/content/us_county.csv'
data_frame = read_csv_or_excel(file_path)
print(data_frame)
```

Output:



```
else:
    #This is the exception handling that I have kept
    raise ValueError("This file format is incorrect. Please provide a CSV or Excel file.")

return df

file_path = '/content/us_county.csv'
data_frame = read_csv_or_excel(file_path)
print(data_frame)
```

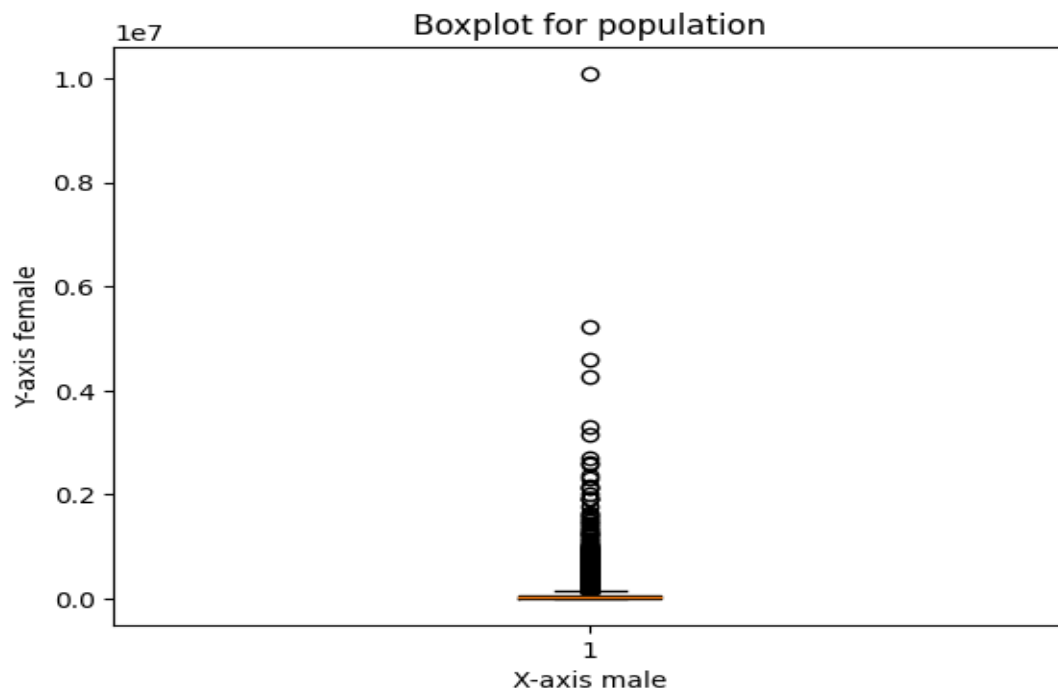
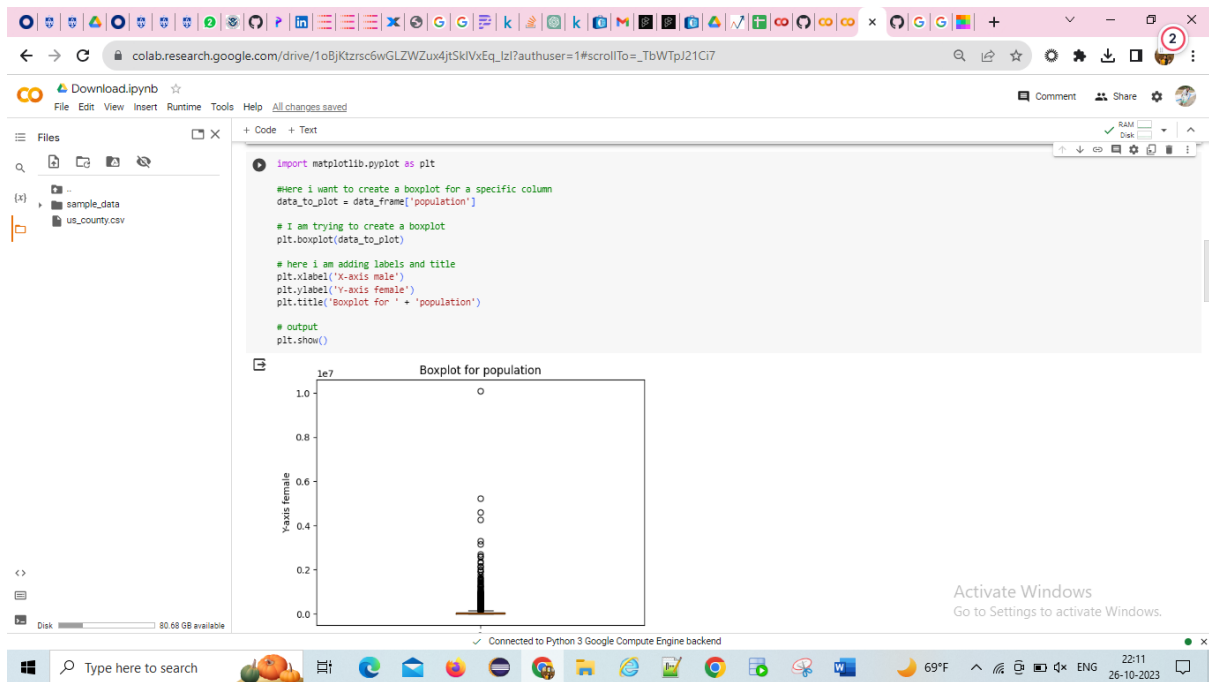
fips	county	state	state_code	male	female
0	1001	Autauga County	Alabama	26874	28326
1	1003	Barbour County	Alabama	10118	10619
2	1005	Barbour County	Alabama	13697	12085
3	1007	Bibb County	Alabama	12152	10375
4	1009	Blount County	Alabama	28434	29211
...
3215	72145	Vega Baja Municipio	Puerto Rico	NAN	25580
3216	72147	Vieques Municipio	Puerto Rico	NAN	4332
3217	72149	Villalba Municipio	Puerto Rico	NAN	11169
3218	72151	Yabucoa Municipio	Puerto Rico	NAN	16541
3219	72153	Vauco Municipio	Puerto Rico	NAN	17475

median_age	population	female_percentage	lat	long
0	37.8	55280	51.315217	32.534923
1	42.8	208107	51.376595	30.727479
2	39.9	25782	46.873788	31.669581
3	39.9	22527	46.855844	32.998628
4	40.8	57645	50.673953	33.980869
...
3215	40.7	53371	52.071350	18.428461
3216	43.6	8771	50.609965	18.122662
3217	38.8	22993	51.424347	18.128155
3218	42.5	34149	51.562271	18.079458
3219	43.0	36439	52.043141	18.079728

[3220 rows x 11 columns]

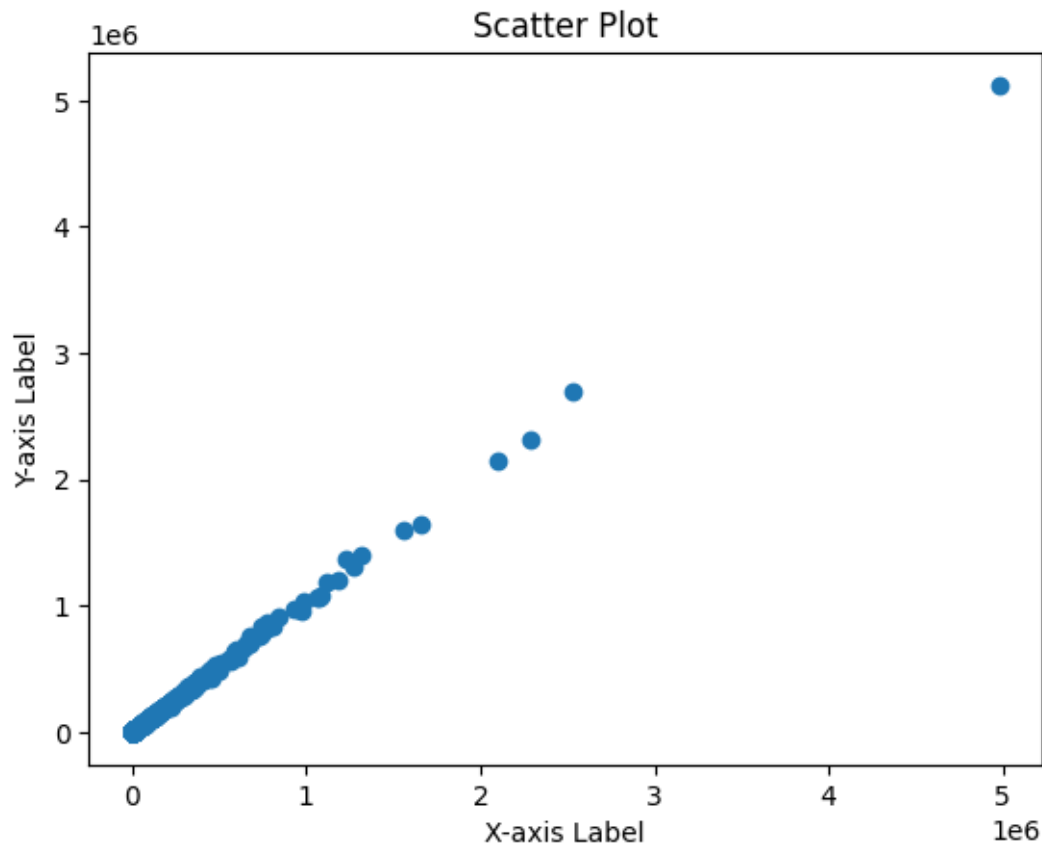
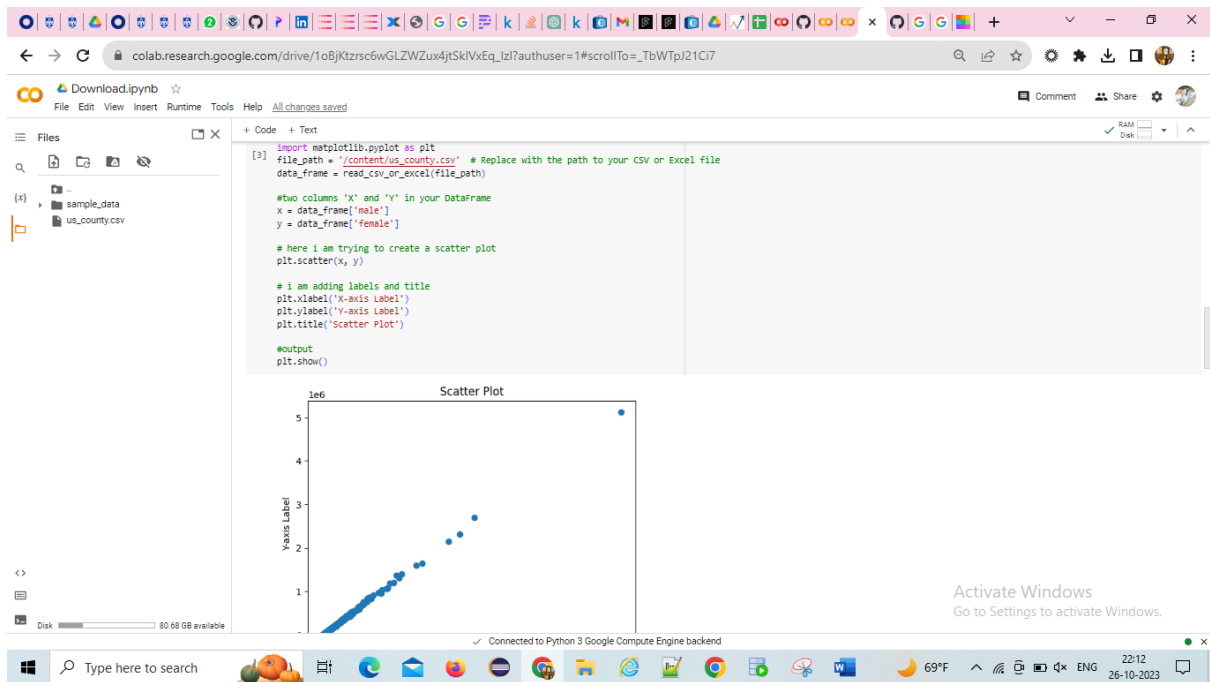
Boxplot Graph:

This graph shows a clear understanding of the male and female ratio



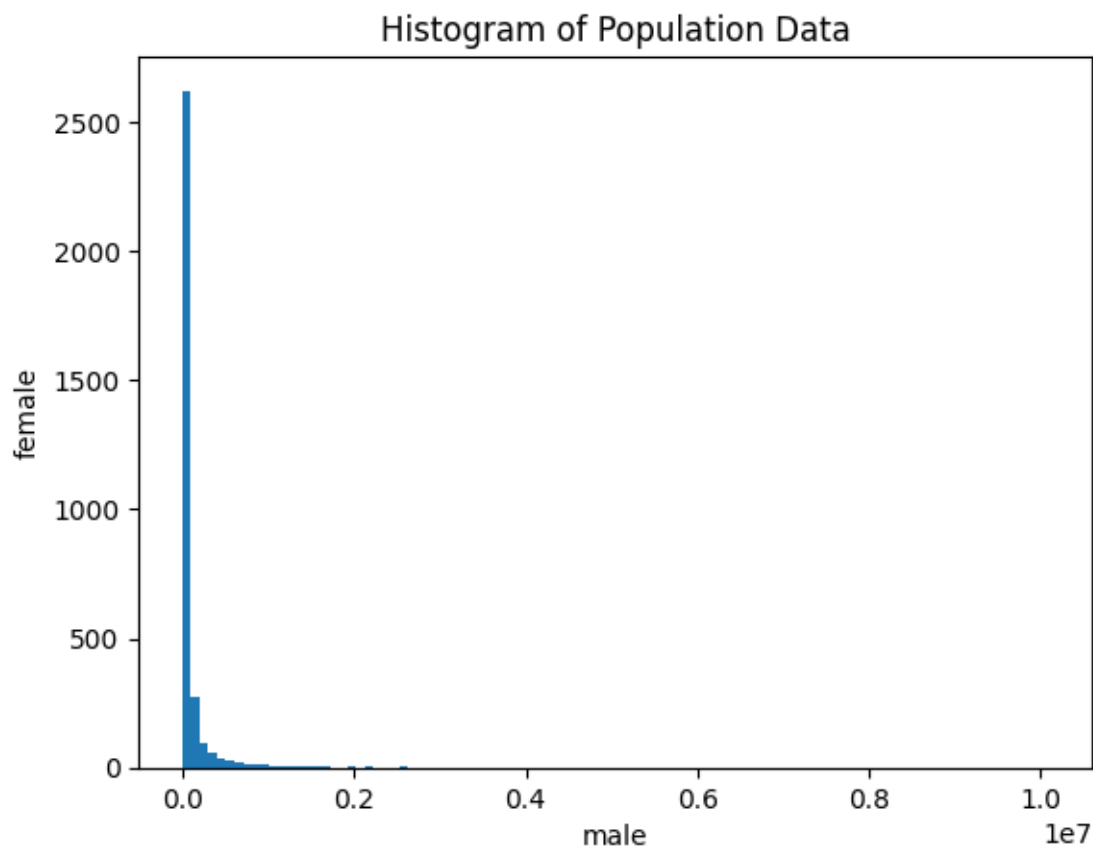
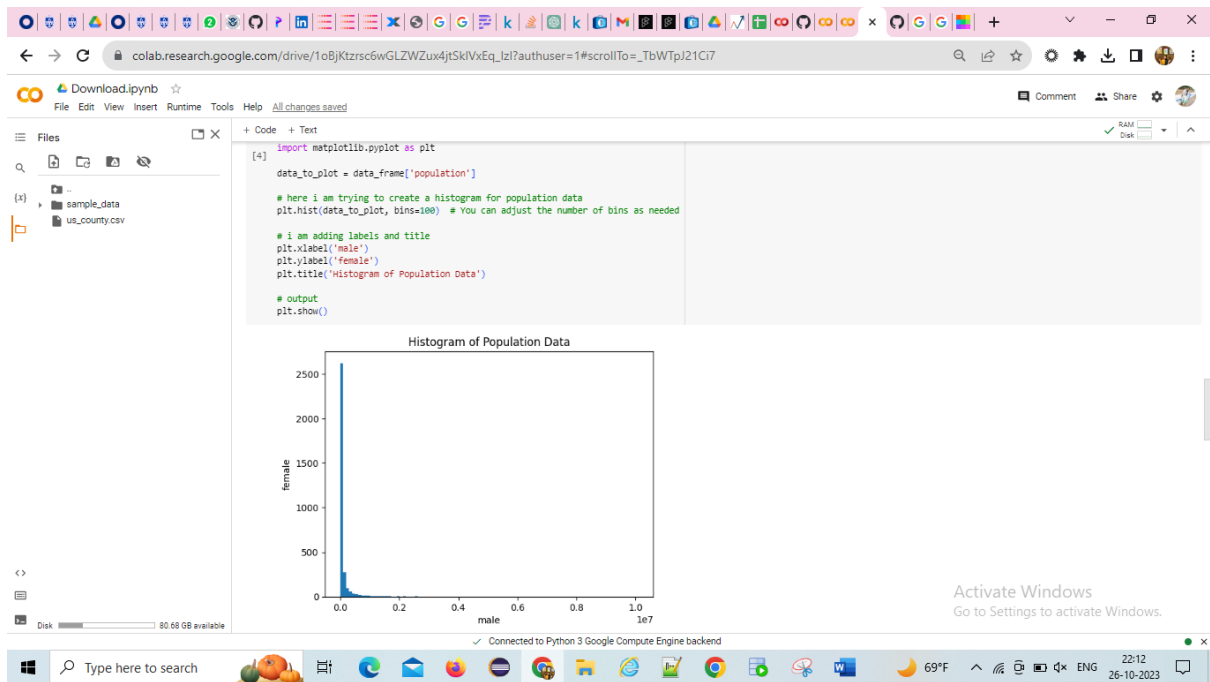
Scatterplot:

This graph shows a clear understanding of the male and female ratio.



Histogram:

This graph shows a clear understanding of the male and female ratio



Important Links:

Dataset Link:

<https://drive.google.com/drive/folders/1RfLhJVOK45x9oGBmOKyZEpBAaHuITYaw>

<https://docs.google.com/spreadsheets/d/1OVgcN0T2npE5nRc9RTND8tUP9znStHVZJwMrOthtqDo/edit#gid=1650272371>

GitHub Link:

<https://github.com/santhiya-hds5210/ORES-5160-EDA>

Drive Link:

https://drive.google.com/drive/folders/1W8AiXxbgTYK-HOXSPKjee9qGdj_Ar1O

Appendix:

- https://www.google.com/search?q=what+is+eda+in+data+science&oq=what+is+EDA+inn&gs_lcrp=EgZjaHJvbWUqCQgBEAAYDRiABDIICAAQRRg5MgkIARAAGA0YgAQyCQgCEAAYDRiABDIJCAMQABgNGIAEMgkIBBAAGA0YgAQyCQgFEAAYDRiABDIJCAYQABgNGIAEMgkIBxAAGA0YgAQyCQgIEAAYDRiABDIJCAkQABgNGIAE0gEJMTE4MjhqMGo3qAIAAsAIA&sourceid=chrome&ie=UTF-8
- https://www.kaggle.com/datasets/headsortails/covid19-us-county-jhu-data-demographics?select=us_county.csv
- <https://stackoverflow.com/questions/18039057/pandas-parser-cparsererror-error-tokenizing-data>
- <https://chat.openai.com/c/8da6a9dc-bee7-4983-9bf9-7530b2178d31>
- <https://www.kaggle.com/code/masoudfaramarzi/basics-of-accesing-data-from-urls-using-pandas>
- <https://www.forefront.ai/app/chat/new>
- https://www.numbeo.com/quality-of-life/rankings_by_country.jsp
- <https://www.analyticsvidhya.com/blog/2022/03/exploratory-data-analysis-with-an-example/>
- <https://docs.google.com/spreadsheets/d/1OVgcN0T2npE5nRc9RTND8tUP9znStHVZJwMrOthtqDo/edit#gid=1650272371>
- <https://canvas.slu.edu/courses/45377/assignments/343230>
- https://colab.research.google.com/drive/1Yr_FH_rjTCW7741e1rArixu4ZWl02FGC#scrollTo=ZflbVsMyiqOI
- <https://github.com/santhiya-hds5210/ORES-5160-EDA>
- https://www.google.com/search?q=scatter+plot&oq=scatter&gs_lcrp=EgZjaHJvbWUqDQgBEAAYgwEYsQMYgAQyDwgAEEUyORiDARixAxiABDINCAEQABiDARixAxiABDIKCAIQABixAxiABDINCAMQABiDARixAxiABDINCAQQABiDARixAxiABDIKCAUQABixAxiABDINCAYQABiDARixAxiABDIHCAcQABiABDIKCAgQABixAxiABDINCAkQABiDARixAxiABNIBCDMzOTdqMGo3qAIAAsAIA&sourceid=chrome&ie=UTF-8
- https://www.google.com/search?q=boxplot&oq=boxpl&gs_lcrp=EgZjaHJvbWUqDAGBEAAYQxiAxiKBTIGCAAQRRg5MgwIARAAGEMYsQMYgUyDwgCEAAYQxiDARixAxiKBTIKCAMQABixAxiABDIJCAQQABhDGloFMgclBRAAGIAEMgkIBhAAGEMYgUyCQgHEAAYQxiKBTIJCAgQABhDGloFMgclCRAAGIAE0gEIMzEwNmowajeoAgCwAgA&sourceid=chrome&ie=UTF-8