Bala_Stage_1_Report

February 14, 2022

1 Project Stage I: Data and Project Understanding

1.1 Enrichment Datasets for COVID-19 NC(North Carolina) State

1.2 Census Demographic ACS

1.3 Task 2

1.3.1 Team:

- Create a team notebook to read in the COVID-19 data (cases, deaths, and population) using pandas and display the dataframe in a notebook.
- Merge all the three variables (cases, deaths, and population) to create a super COVID-19 datafame. Export it to a .csv format. ### Individual:
- Calculate COVID-19 data trends for the last week of the data. Are the cases increasing, decreasing, or stable? Each student chooses a state to analyze.
- Each student member creates notebooks to read the Enrichment data and displays them in a notebook.
- Each student member performs initial merges with the COVID-19 data using the variables in the Enrichment data.

We have imported all required libraries for Team task, Covid-19 data analysis

[136]:

```
USCovid19DF = pd.read_csv("..\\..\\Data\\US_Covid_19_Dataset.
 ↪csv")#,parse_dates=['ReportDate'], index_col = "ReportDate"
USCensusDemoDF = pd.read_csv("..\\..\\Data\\Census_Demographic_ACS.csv",␣
 ↪low_memory=False)
```

We have downloaded latest data from provided website *Census Demographic ACS* for below datasets & Team prepared Covid19 Dataset. Latest US County level data available for all states

- Combined Covid19 Confirmed, Death & Population Dataset
- US Census Demographic ACS dataset

Verified US census Demographics data set

- It has 3221 Observations
- Total 358 columns
- Firstrow with US Census code columns and 2nd row has actual column Names

1.3.2 Calculate COVID-19 data trends for the last week of the data. Are the cases increasing, decreasing, or stable? Each student chooses a state to analyze.

I have filtered NC State from combined COVID19 data set and arranged by ReportDate latest values.

- It contains 75,851 rows and 9 columns

As per problem statement, I have taken latest one week NC Covid data data, which shows 808 rows and 9 columns from Feb 3rd to Feb 10th 2022

[143]:

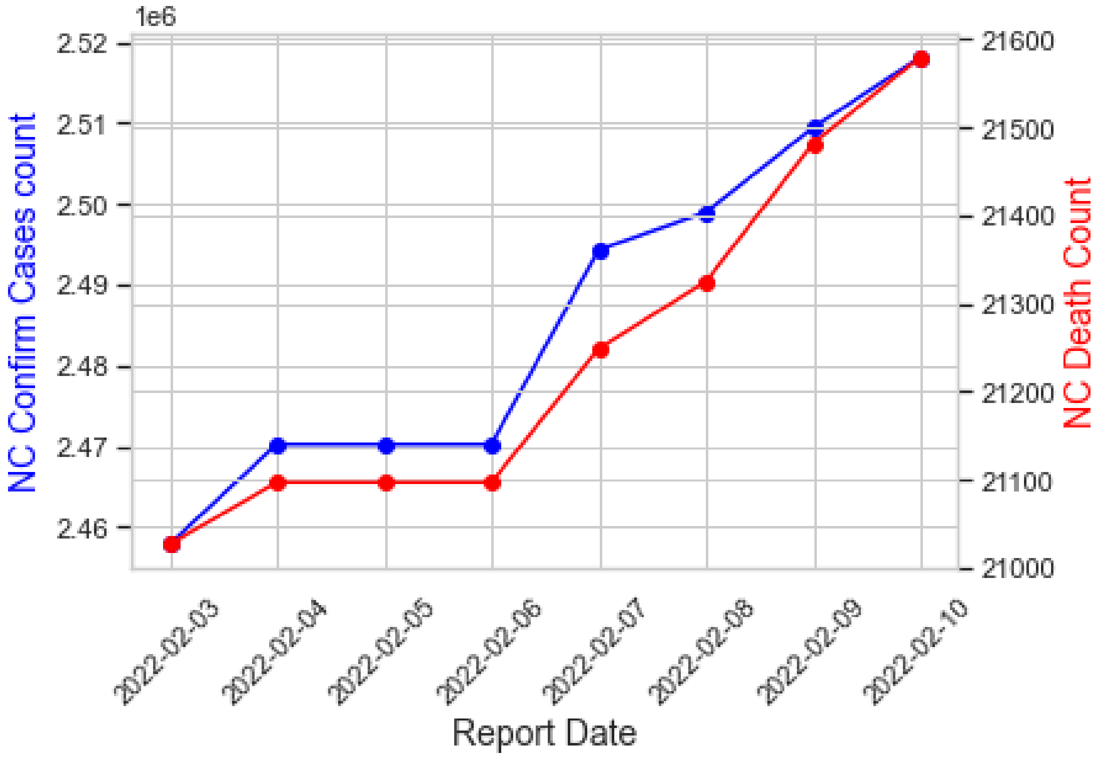
```
[143]:   ReportDate  StateFIPS State  TotalConfirmCasesCount  TotalDeathCount  \
       7 2022-02-10         37    NC                 2518195            21580
       6 2022-02-09         37    NC                 2509470            21482
       5 2022-02-08         37    NC                 2498957            21325
       4 2022-02-07         37    NC                 2494309            21249
       3 2022-02-06         37    NC                 2470242            21097
       2 2022-02-05         37    NC                 2470242            21097
       1 2022-02-04         37    NC                 2470242            21097
       0 2022-02-03         37    NC                 2457857            21027

          TotalPopulation
       7         10488084
       6         10488084
       5         10488084
       4         10488084
       3         10488084
       2         10488084
       1         10488084
       0         10488084
```

In order to see the NC Covid Cases trends by daywise

- I have aggregated data based on Report Date, StateFIPS & State
- I got above aggregate table by report date descending order
- It shows that day to day its Death count increased and during the weekend numbers are same , may not be reported on Feb 4th, 5th & 6th.
- Daily Death count average 100 its increased for population 10.4Million.

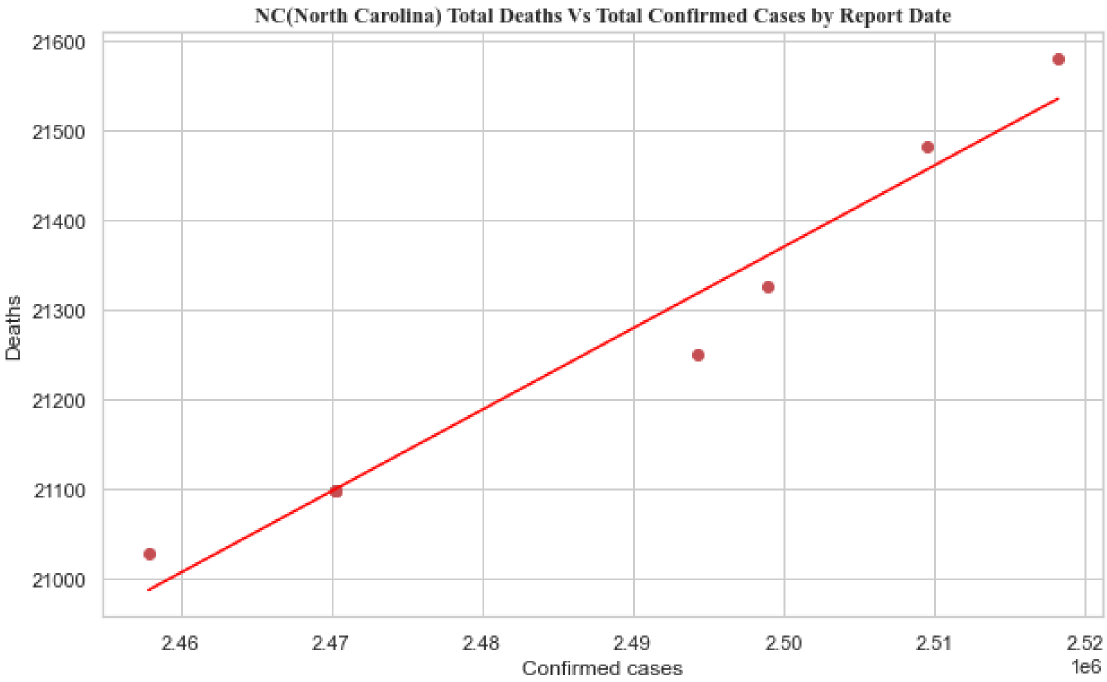
[145]:



Plotted visualization between NC Covid Confirmed cases vs Death Count for 10.4M population.

- Observed that both Cases & Deaths are correlated
- Weekends from Feb 4th to 6th cases/deaths not reported.
- Both Cases and deaths are increased from Feb 3rd to 10th on avg 100.

[146]:



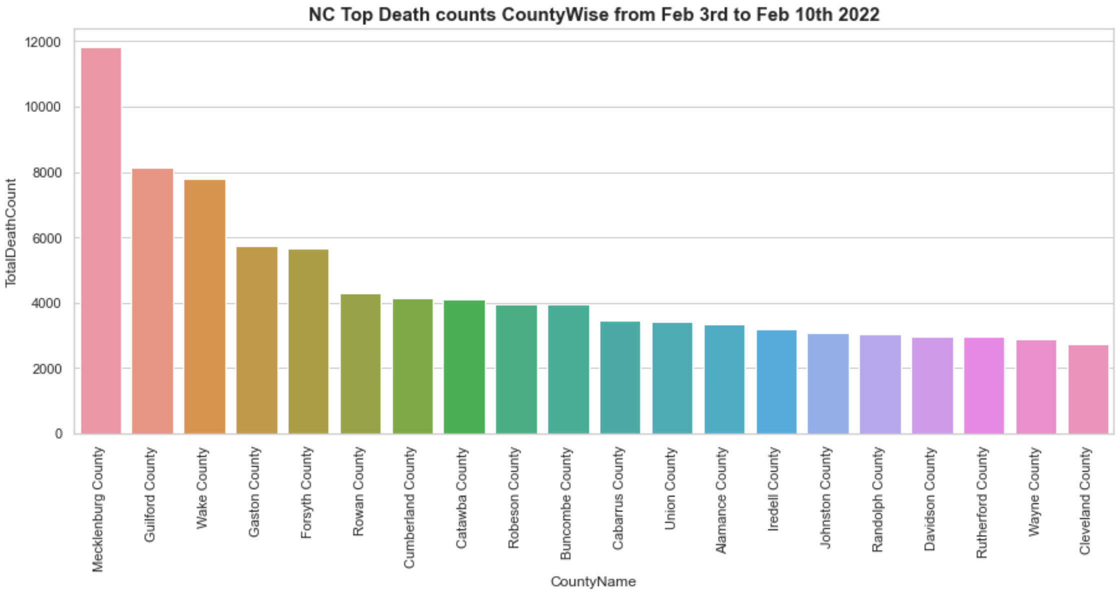
Plotted correlation between NC Confirmed cases vs Deaths

- I see both are highly correlated
- Both Cases & Deaths are increasing linearly.
- For every increase of 10K Confirmed Cases 100 deaths increasing.
- This means for every 10K Covid19 cases there are 100 deaths approximately happening.

Even for further Weekly analysis at county level instead of State

- Aggregated all Confirm cases & deaths by County for latest week data
- There are 101 counties reported data from Feb 3rd to 10th.
- Arrannged dataset by Total Death count descending order
- Observed that Top Coivd19 Deaths happend in county Mecklenburg and then Guilford, Wake, Gaston, Forsyth counties in the order

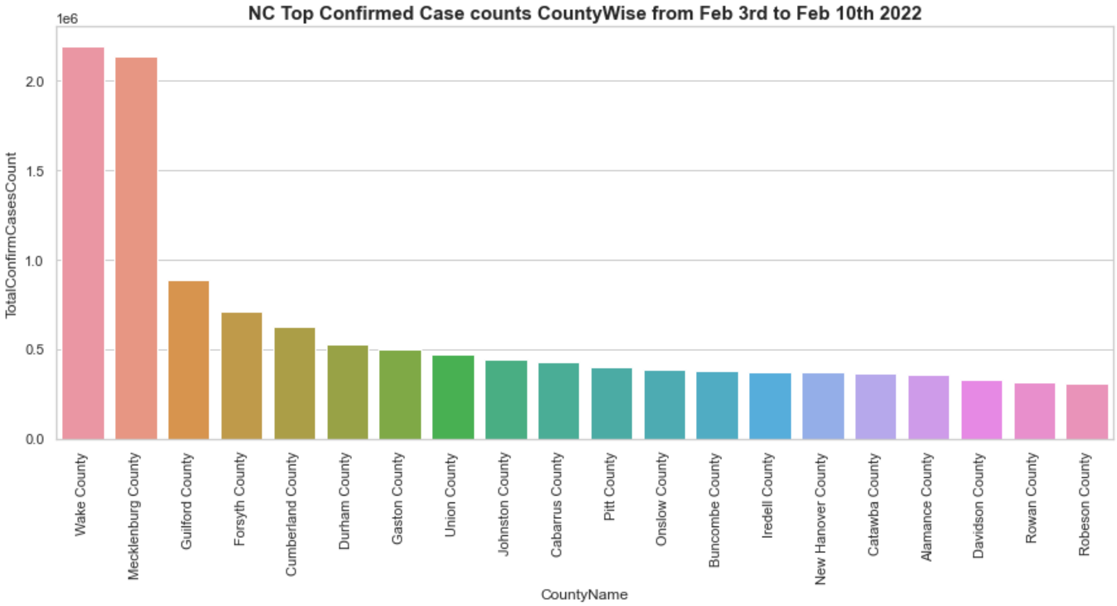
```
[148]:
```



Plotted NC Covid19 last available week data analysis

- I have taken Seaborn to plot this and took top 20 counties to show
- Since its comparision taken Barplot by County wise Total Death Count.
- Mecklenburg county shows 11,813 deaths, Guilford county 8116, Wake county 7775 are the highest in the order.

[149]:



Aggregated & Plotted NC Covid19 last available week data analysis

- Aggregated data and rearranged by Confirmed Cases Counts
- I have taken Seaborn to plot this and took top 20 counties to show
- Since its comparision taken Barplot by County wise Total Death Count.
- Wake county shows 2.19Million Cases, Mecklenburg county 2.13M, Guilford county 881K are the highest in the order.
- Guilford County Covid Cases are less compared to other counties but deaths are more.

1.3.3 Each student member creates notebooks to read the Enrichment data and displays them in a notebook.

1.3.4 Enrichment Datasets for COVID-19

1.3.5 Census Demographic ACS

US Census Demographic data import

- Imported US Census Demographic data again to USCensusDemoDF
- Observe that 3220 rows and 358 columns exists

Formatted the enrichment dataset

- Split the Geographic Area Name column into two different columns(County Name and state)
- Rename some of the columns
- Display the new dataset

We have merged US Covid19 Combined NC dataset with Enrichment Data US Census Demographic Set by using below steps.

- Taken all common columns between data sets. i.e. StateName, CountyName
- Also to avoid missing data due to joins, we have used Left join on enrichment data.
- So merged based on above columns and verified data, they are all looks good

```
[155]: USCovid19NCCensusDemoDF.shape
```

```
[155]: (75851, 367)
```

Combined US Covid19 data with Census Geographic data shows 75,851 observations and 367 columns and looks good. I may need to drop unnessesary columns from Census Demo data.