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 Employment Dataset

1.4 Task 2

1.4.1 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)
USEmpStCntyDF = pd.read_excel("..\\..\\Data\\Employment_Dataset.xlsx",␣
↪sheet_name='US_St_Cn_MSA', engine='openpyxl')
USEmpPRVIDF = pd.read_excel("..\\..\\Data\\Employment_Dataset.xlsx",␣
↪sheet_name='US_PR_VI', engine='openpyxl')
```

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
- Employment Dataset with 2 sheets

Exported Python Dataframe data into csv file and stored under Data directory named: US_Covid_19_Dataset.csv & modified

- Converted Population column datatype from Float to Int
- Converted ReportDate datatype from Object to Date datatype

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

US Employment data has 62,790 rows and 21 columns based on first sheet under exported dataset and all are non null columns.Its wide dataset as well

US Employment data has 1,546 rows and 21 columns based on first sheet under exported dataset and all are non null columns.Its wide dataset as well

1.4.2 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]:

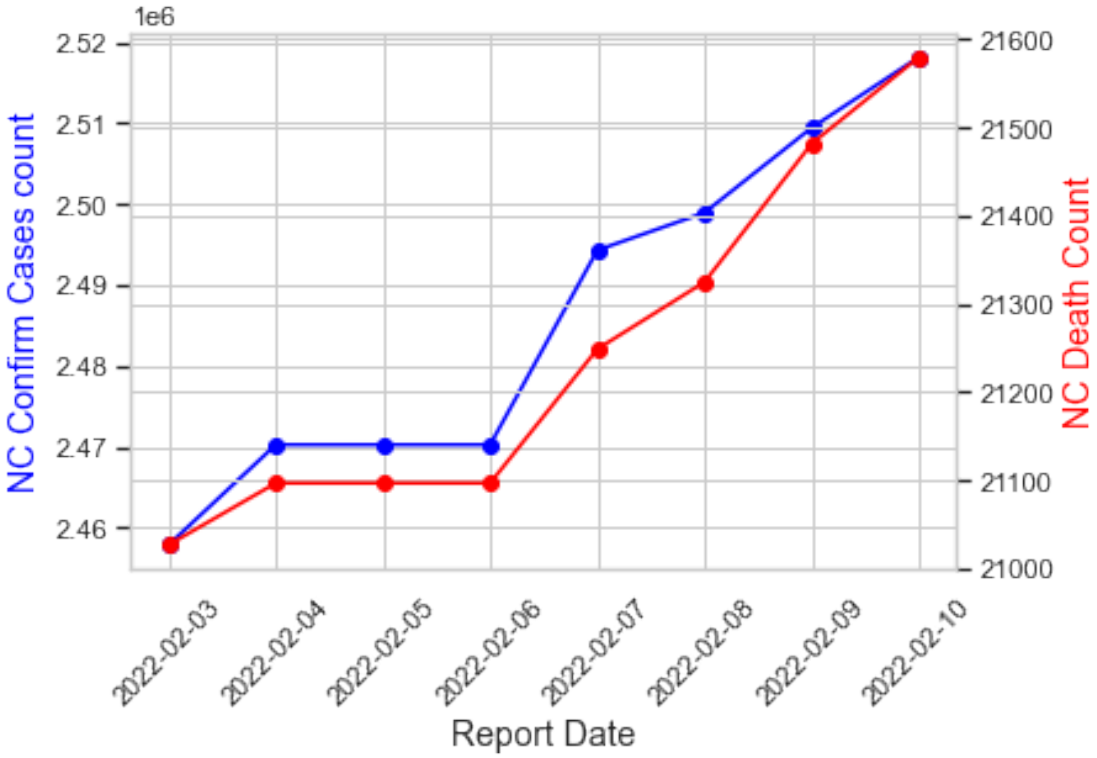
[143]:

	ReportDate	StateFIPS	State	TotalConfirmCasesCount	TotalDeathCount	TotalPopulation
7	2022-02-10	37	NC	2518195	21580	10488084
6	2022-02-09	37	NC	2509470	21482	10488084
5	2022-02-08	37	NC	2498957	21325	10488084
4	2022-02-07	37	NC	2494309	21249	10488084
3	2022-02-06	37	NC	2470242	21097	10488084
2	2022-02-05	37	NC	2470242	21097	10488084
1	2022-02-04	37	NC	2470242	21097	10488084
0	2022-02-03	37	NC	2457857	21027	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.

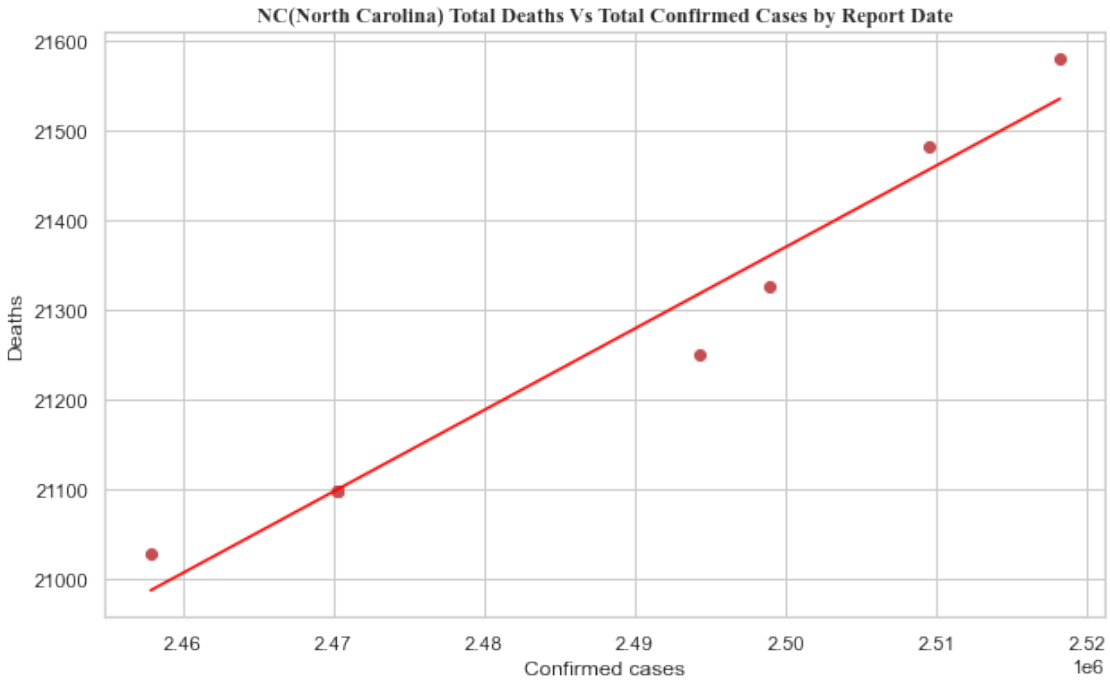
[145]:



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.

[147]:

[147]:

	CountyFIPS	CountyName	State	TotalConfirmCasesCount	TotalDeathCount	TotalPopulation
60	37119	Mecklenburg County	NC	2130740	11813	8882848
41	37081	Guilford County	NC	881389	8116	4297392
92	37183	Wake County	NC	2190201		
36	37071	Gaston County	NC	498497		
34	37067	Forsyth County	NC	705963		
..	…	…	…	…		
48	37095	Hyde County	NC	9974		
3	37005	Alleghany County	NC	22042		
15	37029	Camden County	NC	13007		
89	37177	Tyrrell County	NC	7416		
0	0	Statewide Unallocated	NC	11038		

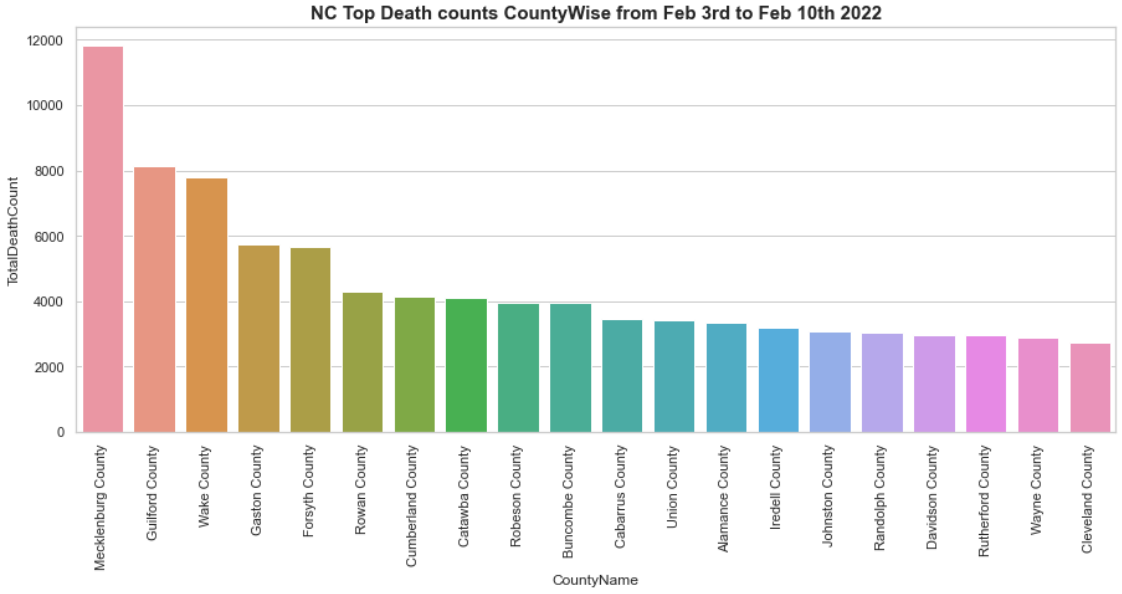
```
92              7775            8894088
36              5725            1796232
34              5663            3058360
..               …                 …
48               104              39496
3                 88              89096
15                72              86936
89                56              32128
0                  0                  0

[101 rows x 6 columns]
```

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

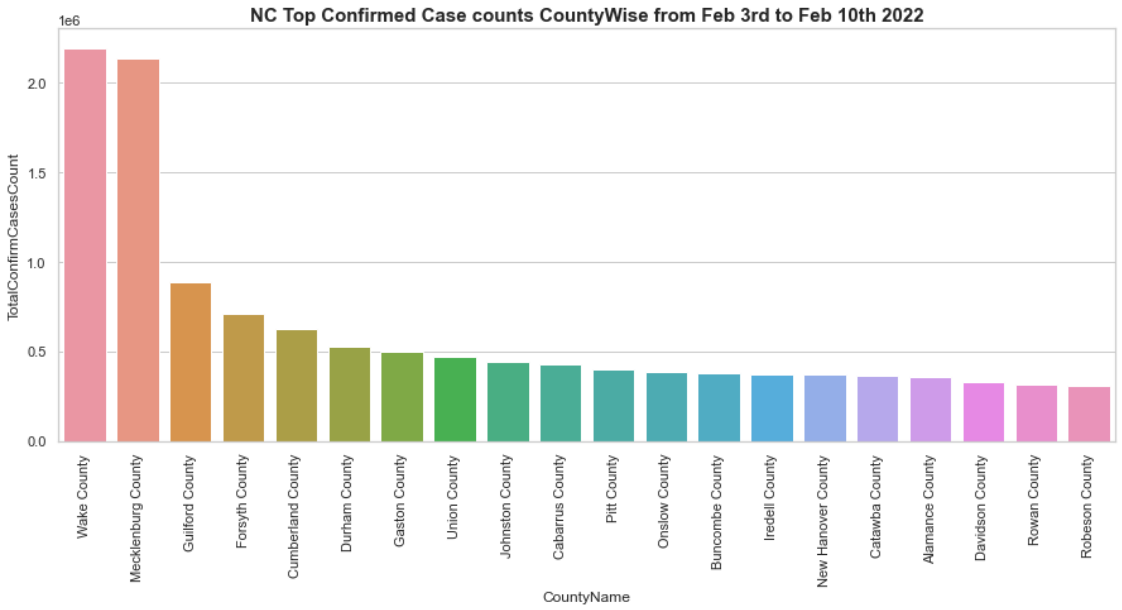
[148]:



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.

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

1.4.4 Enrichment Datasets for COVID-19

1.4.5 Census Demographic ACS

[150]:
```
USCensusDemoDF = pd.read_csv("..\\..\\Data\\Census_Demographic_ACS.csv",␣
↪header=1, low_memory=False)
USCensusDemoDF.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3220 entries, 0 to 3219
Columns: 358 entries, Estimate!!SEX AND AGE!!Total population to Geographic Area
Name
dtypes: float64(140), int64(153), object(65)
memory usage: 8.8+ MB
```

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)
```

[156]:
```
USCovid19NCCensusDemoDF.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 75851 entries, 0 to 75850
Columns: 367 entries, ReportDate to Geographic Area Name
dtypes: Int64(1), datetime64[ns](1), float64(293), int64(4), object(68)
memory usage: 213.0+ MB
```

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.