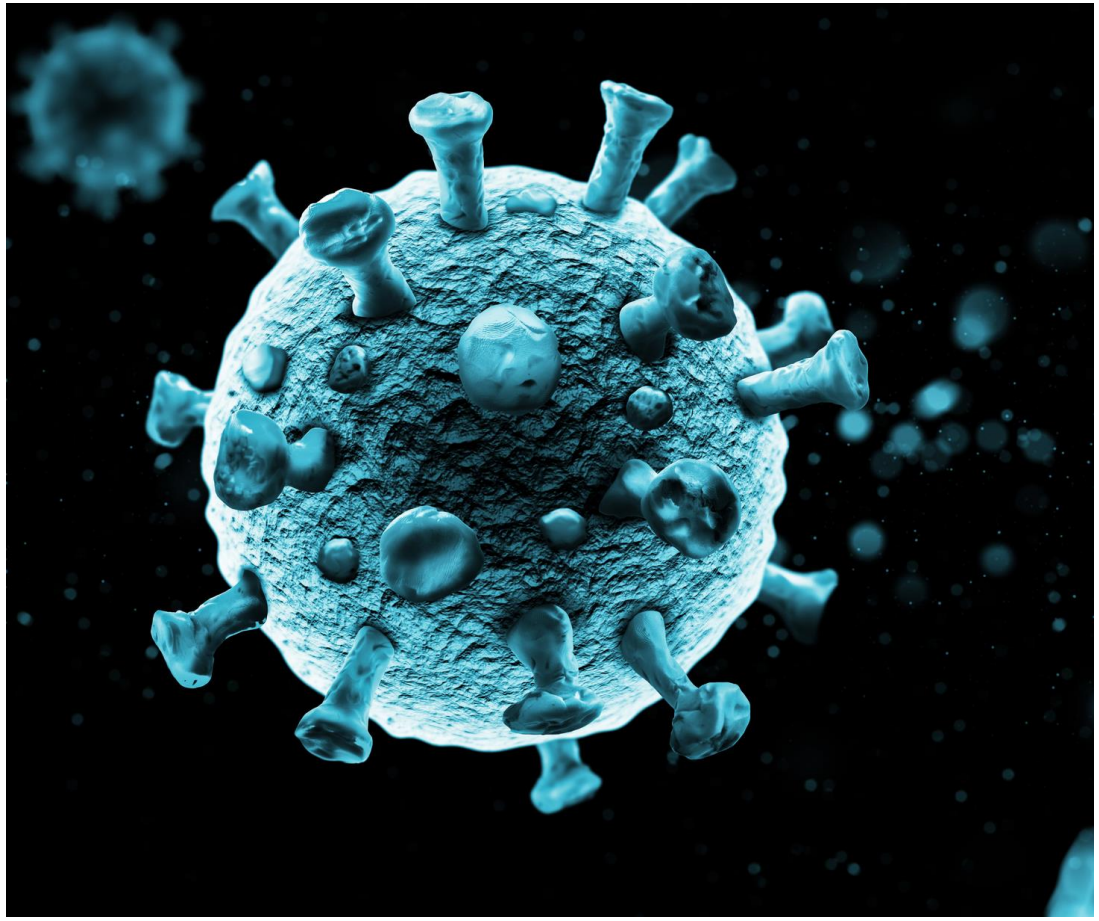
Group 5

COVID-19 Data Analysis

Sara, Amna, Jennifer, Natalie, Kannan

Hypothesis:

- Counties with more people, different social and economic factors, a variety of housing situations, and certain political views had a quicker and bigger spread of COVID-19 during the pandemic.
- **more people** in areas that are living close together, like cities, probably had a harder time stopping the spread.
- **social and economic** factors like income, job types, and access to healthcare might have affected how the virus spread.
- **housing situations** like crowded or multi-generational households could see more infections because it's harder to keep distance from each other
- **political views** in the sense of public health measures (like masks and social distancing) could have influenced how the virus spread in certain areas.



Findings/output

We noticed that many counties had zero cases and deaths in the early months of 2020. From January to early March 2020, there were very few reported cases

This makes sense because covid 19 hadn't spread widely in the U.S. during that time, especially in rural counties.

As you can see in the data....

```
confirmed cases data:
   countyFIPS            County Name State  StateFIPS  2020-01-22  2020-01-23  \
0           0  Statewide Unallocated    AL          1           0           0
1        1001         Autauga County     AL          1           0           0
2        1003         Baldwin County     AL          1           0           0
3        1005         Barbour County     AL          1           0           0
4        1007            Bibb County     AL          1           0           0

   2020-01-24  2020-01-25  2020-01-26  2020-01-27  ...  2023-07-14  \
0           0           0           0           0  ...           0
1           0           0           0           0  ...       19913
2           0           0           0           0  ...       70521
3           0           0           0           0  ...        7582
4           0           0           0           0  ...        8149

   2023-07-15  2023-07-16  2023-07-17  2023-07-18  2023-07-19  2023-07-20  \
0           0           0           0           0           0           0
1       19913       19913       19913       19913       19913       19913
2       70521       70521       70521       70521       70521       70521
3        7582        7582        7582        7582        7582        7582
4        8149        8149        8149        8149        8149        8149

   2023-07-21  2023-07-22  2023-07-23
0           0           0           0
1       19913       19913       19913
2       70521       70521       70521
3        7582        7582        7582
4        8149        8149        8149

[5 rows x 1269 columns]
```

Variable Dictionary

- Variable: countyFIPS
- Data Type: int64
- Description: A unique identifier for each county.
-
- Variable: County Name
- Data Type: object (string)
- Description: The name of the county.
-
- Variable: State
- Data Type: object (string)
- Description: The state where the county is located.
-
- Variable: Population
- Data Type: int64

Presidential dataset:

1. Loaded and filtered the presidential election dataset for New York.
2. Made a data dictionary to describe the key variables:
    - **state:** U.S. state (new york).
    - **county**: The county within the state
    - **current_votes:** The number of votes tallied so far
    - **total_votes:** The total expected votes in the county
    - **percent:** The percentage of votes that was counted

3. loaded the covid data set for confirmed cases, deaths, and population for the chosen state New York.

4. filtered and reshaped the data because my data I encountered an issue where the County Name values didn't match due to formatting like extra spaces and lower and uppercase inconsistencies and to fix this I took out spaces and standardized the case (lowercase) in the County Name columns in both datasets using .melt, this helps for easier future findings and for new york (focused on 2020 data). (explained more on the report)

5. calculated covid case trends for the last week in new york.

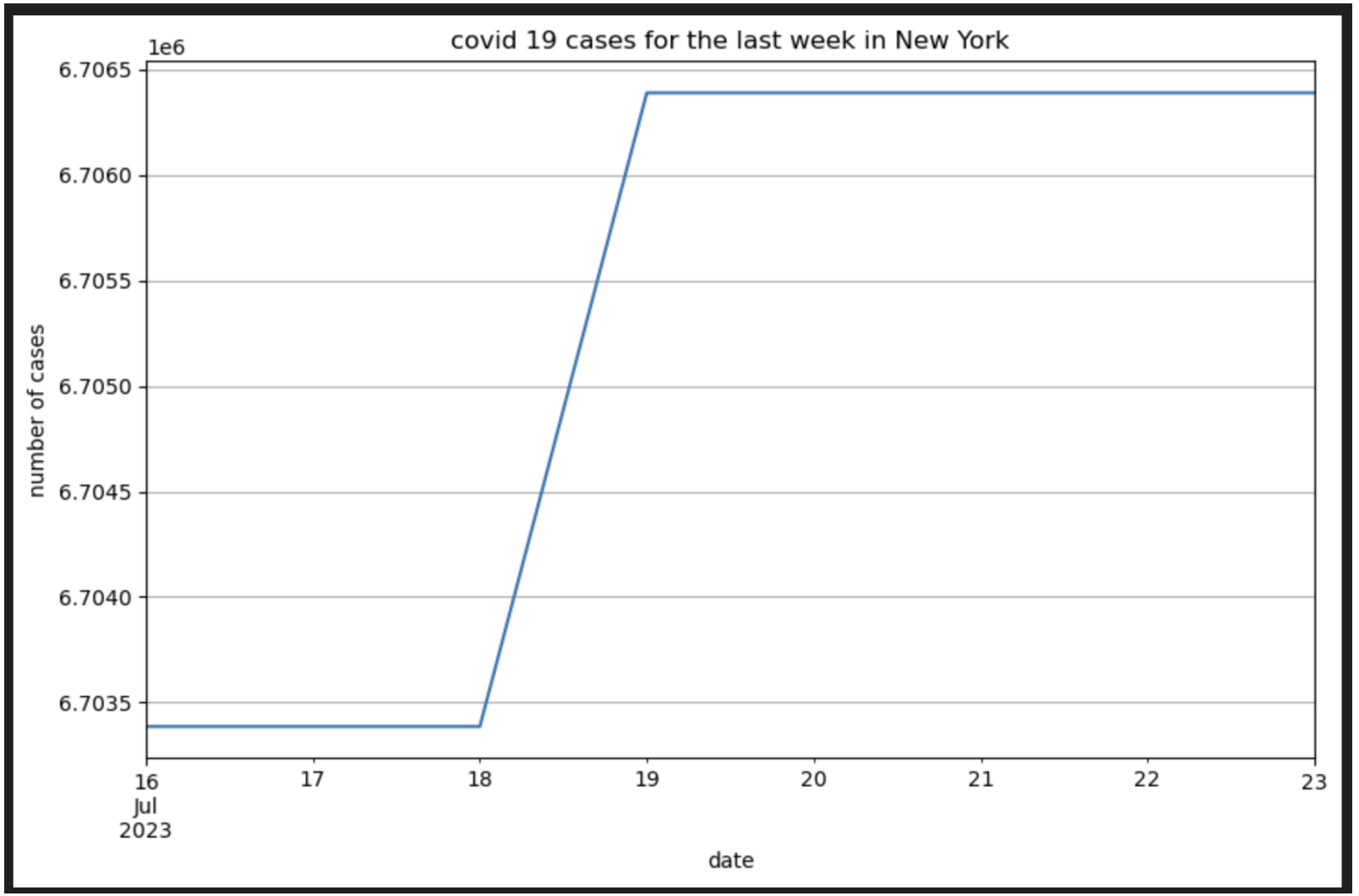
6. made a plot to see if the cases are increasing, decreasing, or stable.

Initial hypothesis to prove: A county's political leaning, by the presidential candidate or voter turnout, may influence covid 19 trends. So, counties with higher voter turnout or specific political preferences can show different case patterns. In easier words, counties that supported different political candidates or had different levels of voter turnout could show patterns in how covid 19 cases developed over time, like higher/lower infection rates or the differences in how cases increased/decreased.

- Test it through regression analysis and look at the p value and r squared value to determine the relationship between voter turnout and covid 19 cases

Output findings:

The plot showed a sudden increase in covid cases in New York on July 18, 2023. And then after the spike the number of cases remained stable for the rest of the week this could be because of multiple reasons one reason could be a delay with reporting the data



```
Variable: state
  Data Type: object (string)
  Description: The U.S. state (New York in this dataset).

Variable: county
  Data Type: object (string)
  Description: The county within the state of New York.

Variable: current_votes
  Data Type: int64
  Description: The number of votes tallied in the county so far.

Variable: total_votes
  Data Type: int64
  Description: The total number of votes expected in the county.

Variable: percent
  Data Type: int64
  Description: The percentage of votes counted for the county.
```

Census Demographic ACS

Hypothesis:

There will be a strong relationship between the demographic composition of a county and the spread of COVID-19.

Specifically, counties with a higher population density, a larger percentage of elderly residents, or a larger percentage of minority populations may affect COVID-19 case counts and death rates.

COVID-19 and ACS Data Analysis: Alabama

1. **Loaded and filtered** the COVID-19 dataset for Alabama.

2. **Created a data dictionary** to describe the key variables from the ACS dataset**:**

-GEO_ID: Unique geographic identifier (county-level FIPS code).

-NAME: Name of the county and state.

-DP05_0001E: Total population estimate for each county.

-DP05_0018E: Median age estimate of the population.

-DP05_0037PE: Percent of the population that identifies as White.

-DP05_0071PE: Percent of the population that identifies as Hispanic or Latino.

3. Loaded the COVID-19 datasets for confirmed cases, deaths, and population for Alabama.

4. Filtered and reshaped the data for easier analysis, focusing on Alabama in 2020.

5. Calculated COVID-19 trends for the last week in Alabama.

Findings:

Trend: COVID-19 cases in Alabama over the last 7 days of the data provided were stable.

Summary Statistics:

Mean cases: 4264.714285714285

Median cases: 4265.0

Standard deviation: 0.7559289460184544

First few rows of covid data in alabama displayed

COVID-19 Deaths:

	countyFIPS	County Name	State	StateFIPS	2020-01-22	2020-01-23	2020-01-24	2020-01-25	2020-01-26	2020-01-27	...	2023-07-14	2023-07-15	2023-07-16	2023-07-17	2023-07-18	2023-07-19	2023-07-20	2023-07-21	2023-07-22	20 07
0	0	Statewide Unallocated	AL	1	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	
1	1001	Autauga County	AL	1	0	0	0	0	0	0	...	235	235	235	235	235	235	235	235	235	
2	1003	Baldwin County	AL	1	0	0	0	0	0	0	...	731	731	731	731	731	731	731	731	731	
3	1005	Barbour County	AL	1	0	0	0	0	0	0	...	104	104	104	104	104	104	104	104	104	
4	1007	Bibb County	AL	1	0	0	0	0	0	0	...	111	111	111	111	111	111	111	111	111	

5 rows × 1269 columns

COVID-19 Cases:

	countyFIPS	County Name	State	StateFIPS	2020-01-22	2020-01-23	2020-01-24	2020-01-25	2020-01-26	2020-01-27	...	2023-07-14	2023-07-15	2023-07-16	2023-07-17	2023-07-18	2023-07-19	2023-07-20	2023-07-21	2023-07-22
0	0	Statewide Unallocated	AL	1	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
1	1001	Autauga County	AL	1	0	0	0	0	0	0	...	19913	19913	19913	19913	19913	19913	19913	19913	19913
2	1003	Baldwin County	AL	1	0	0	0	0	0	0	...	70521	70521	70521	70521	70521	70521	70521	70521	70521
3	1005	Barbour County	AL	1	0	0	0	0	0	0	...	7582	7582	7582	7582	7582	7582	7582	7582	7582
4	1007	Bibb County	AL	1	0	0	0	0	0	0	...	8149	8149	8149	8149	8149	8149	8149	8149	8149

5 rows × 1269 columns

COVID-19 Population:

	countyFIPS	County Name	State	population
0	0	Statewide Unallocated	AL	0
1	1001	Autauga County	AL	55869
2	1003	Baldwin County	AL	223234
3	1005	Barbour County	AL	24686
4	1007	Bibb County	AL	22394

Merged data sets:

ACS Demographic Data:

	GEO_ID	NAME	DP05_0001E	DP05_0001M	DP05_0002E	DP05_0002M	DP05_0003E	DP05_0003M	DP05_0004E	DP05_0004M	...	DP05_
0	Geography	Geographic Area Name	Estimate!!SEX AND AGE!!Total population	Margin of Error!!SEX AND AGE!!Total population	Estimate!!SEX AND AGE!!Total population!!Male	Margin of Error!!SEX AND AGE!!Total population...	Estimate!!SEX AND AGE!!Total population!!Female	Margin of Error!!SEX AND AGE!!Total population...	Estimate!!SEX AND AGE!!Total population!!Sex r...	Margin of Error!!SEX AND AGE!!Total population...	...	Percer Error!!H OF
1	0100000US	United States	326569308	*****	160818530	7991	165750778	8011	97.0	0.1	...	

2 rows × 359 columns

The Employment Data

Hypothesis Questions:

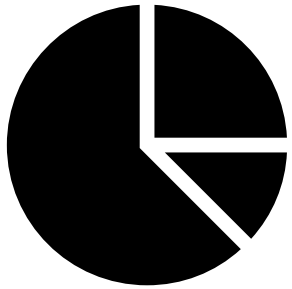
Is there a correlation between a higher covid case/death rate and employment rates?

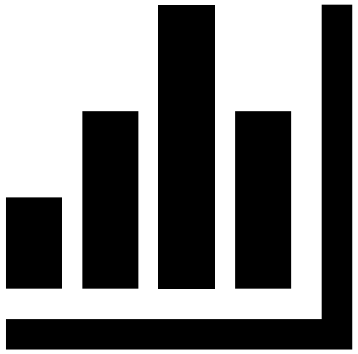
Does the type of employment correlate with covid cases & death rates?

Would lower employment lead to lower cases?

Extracting Relevant Data:

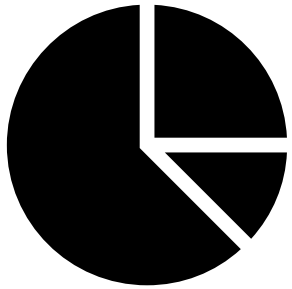
- Year
- Qtr
- Area Type
- St Name
- Area Ownership
- Industry
- Annual Average Status Code
- Annual Average Establishment Count
- Annual Average Employment
- Annual Total Wages
- Annual Average Weekly Wage
- Annual Average Pay
- Employment Location Quotient Relative to U.S.
- Total Wage Location Quotient Relative to U.S.

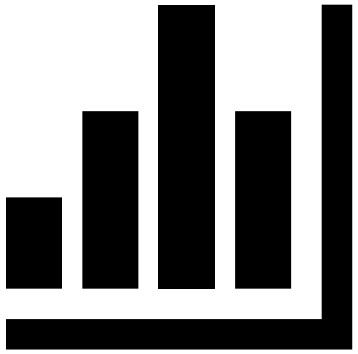




The Most Relevant Data:

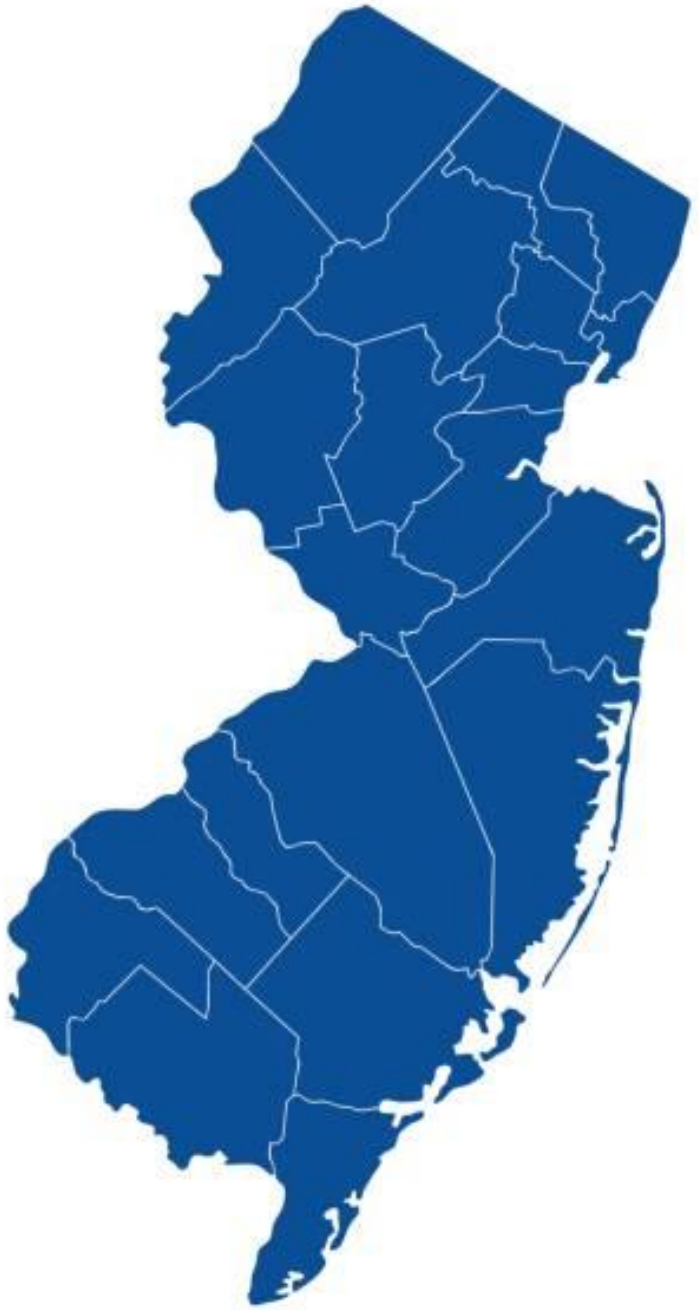
- Year
- Qtr
- Area Type
- St Name
- Area Ownership
- Industry
- Annual Average Status Code
- Annual Average Establishment Count
- ***Annual Average Employment***
- Annual Total Wages
- Annual Average Weekly Wage
- Annual Average Pay
- Employment Location Quotient Relative to U.S.
- Total Wage Location Quotient Relative to U.S.





Annual Average Employment

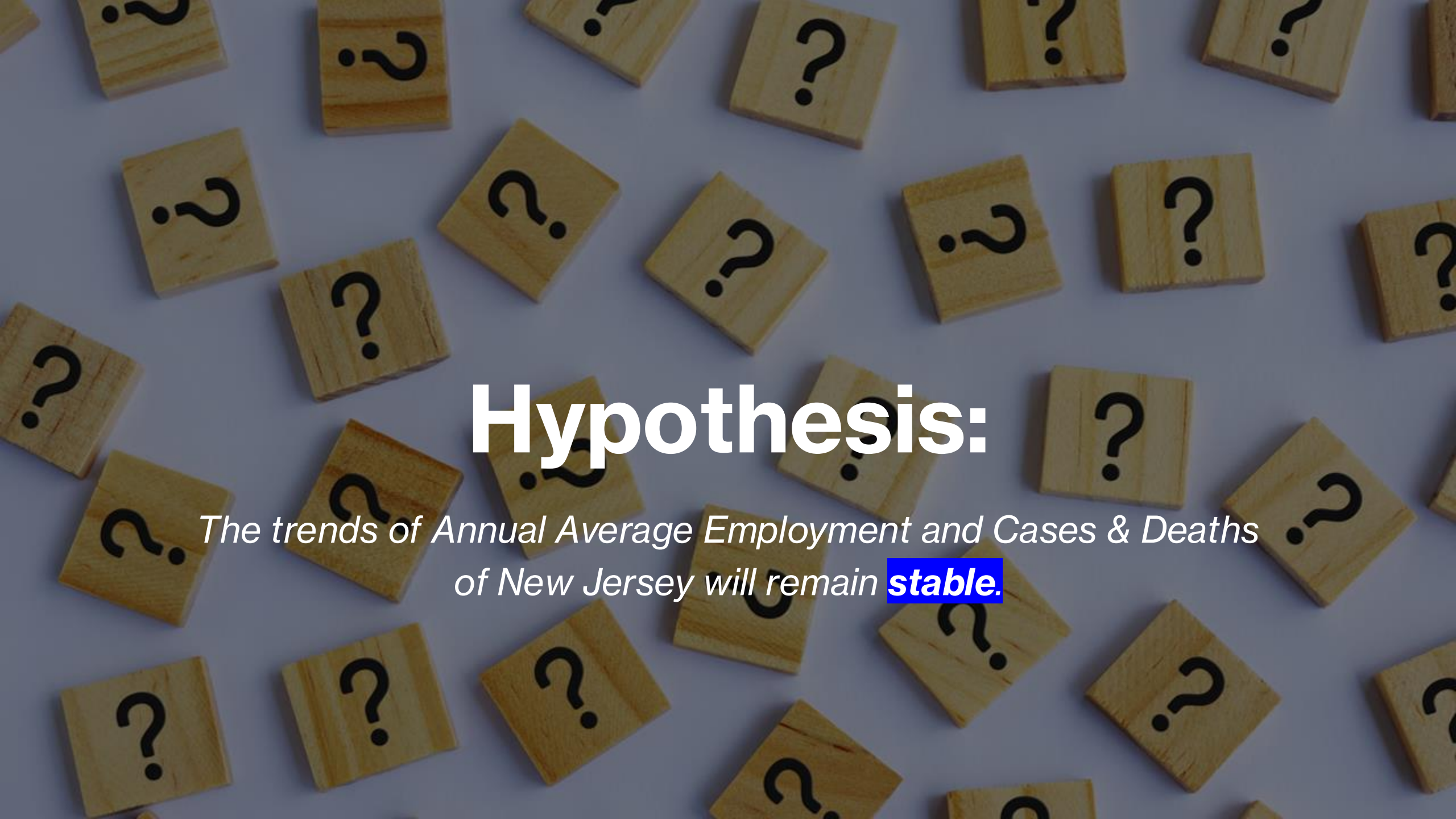
Using this data...



Evaluating the Data:

When evaluating the last week of available data:

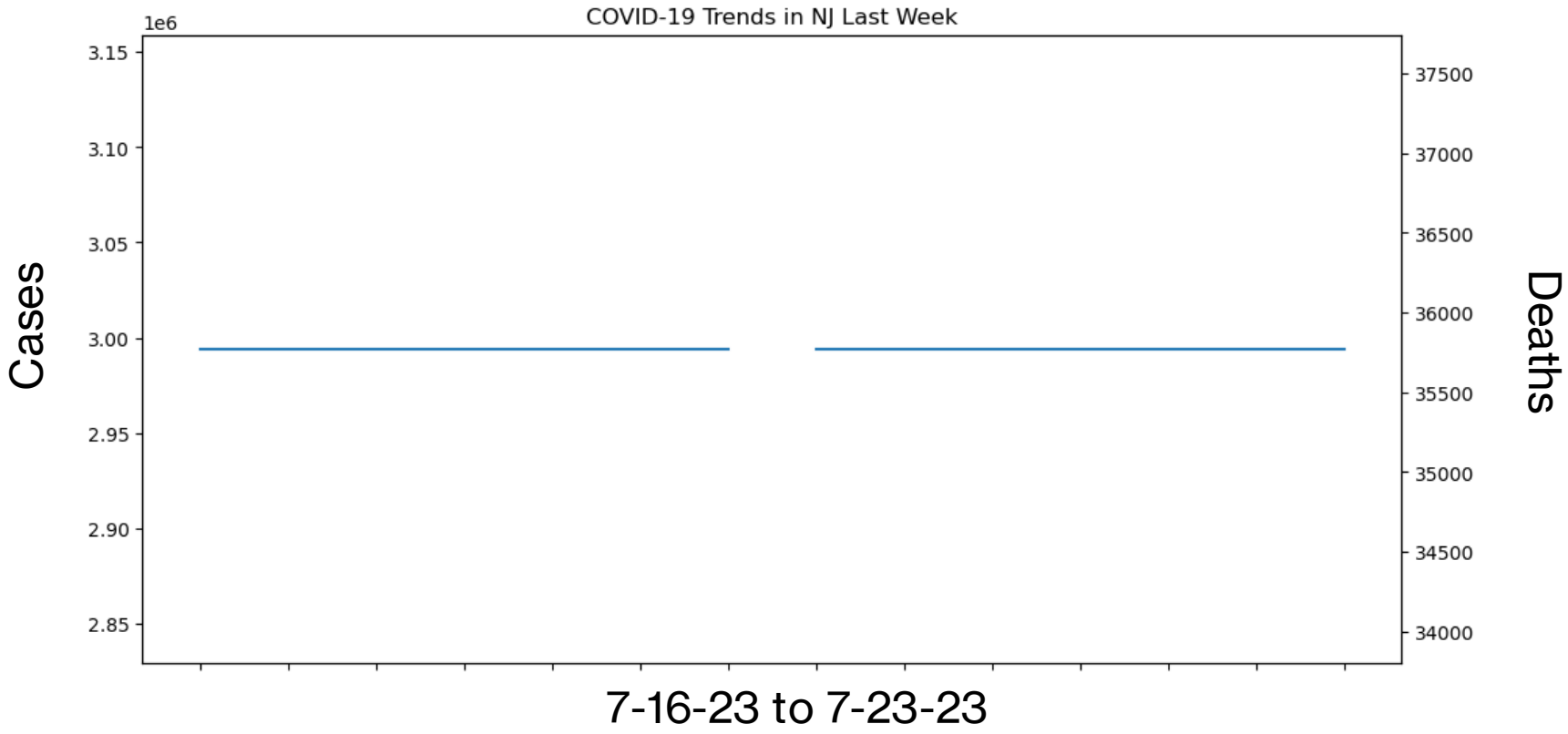
- Timeframe: July 16th - 23rd of 2023
- After the peak of the initial COVID-19 outbreak



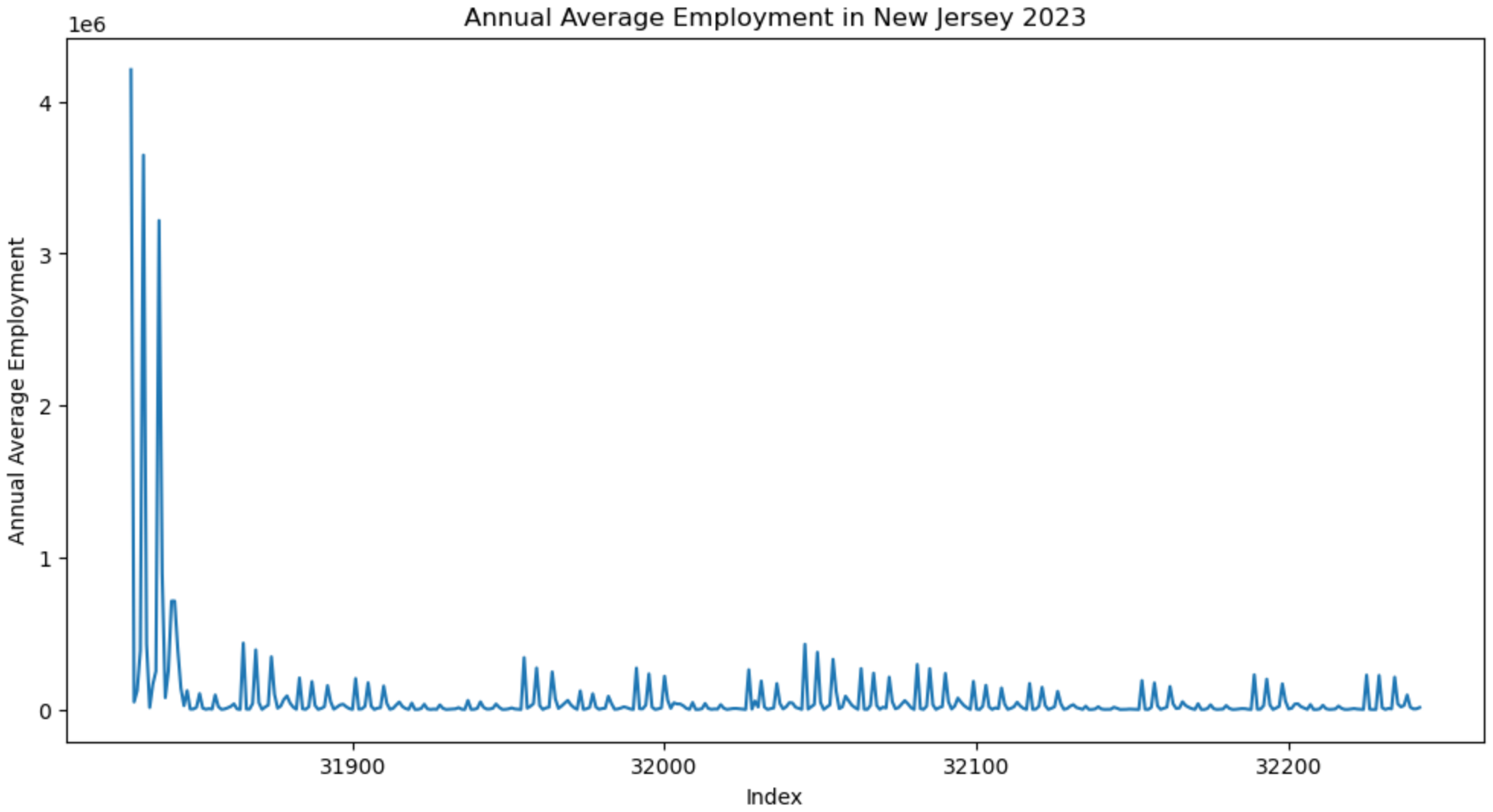
Hypothesis:

*The trends of Annual Average Employment and Cases & Deaths of New Jersey will remain **stable**.*

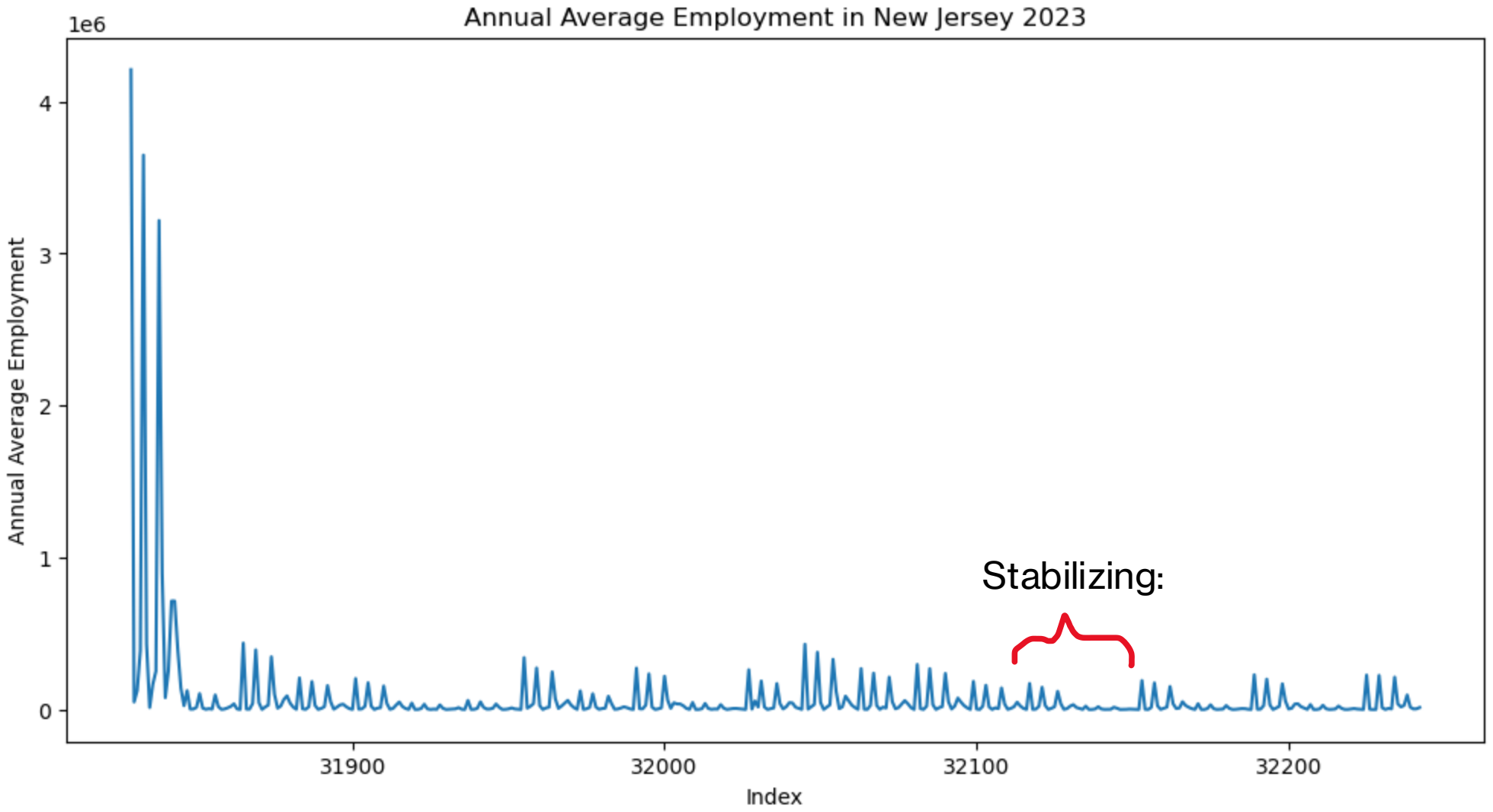
Graphing Deaths & Cases



Graphing Average Annual Employment



Graphing Average Annual Employment



Both Datasets are:

Consistent!

- As shown by both graphs, the data for Average Annual Employment, and the Cases & Deaths of New Jersey during July of 2023, are consistent.

Employment Data Analysis

- **Objective:**
- Examine recent trends in COVID-19 cases and deaths in North Carolina.
- Compare these trends with employment data from the same year (2023).

- **Data Used:**
- COVID-19 statistics: cases, deaths, and population figures.
- Employment figures for 2023.

Data Loading and Filtering :

- **Data Source:** Quarterly Census of Employment and Wages
- QCEW Data Files

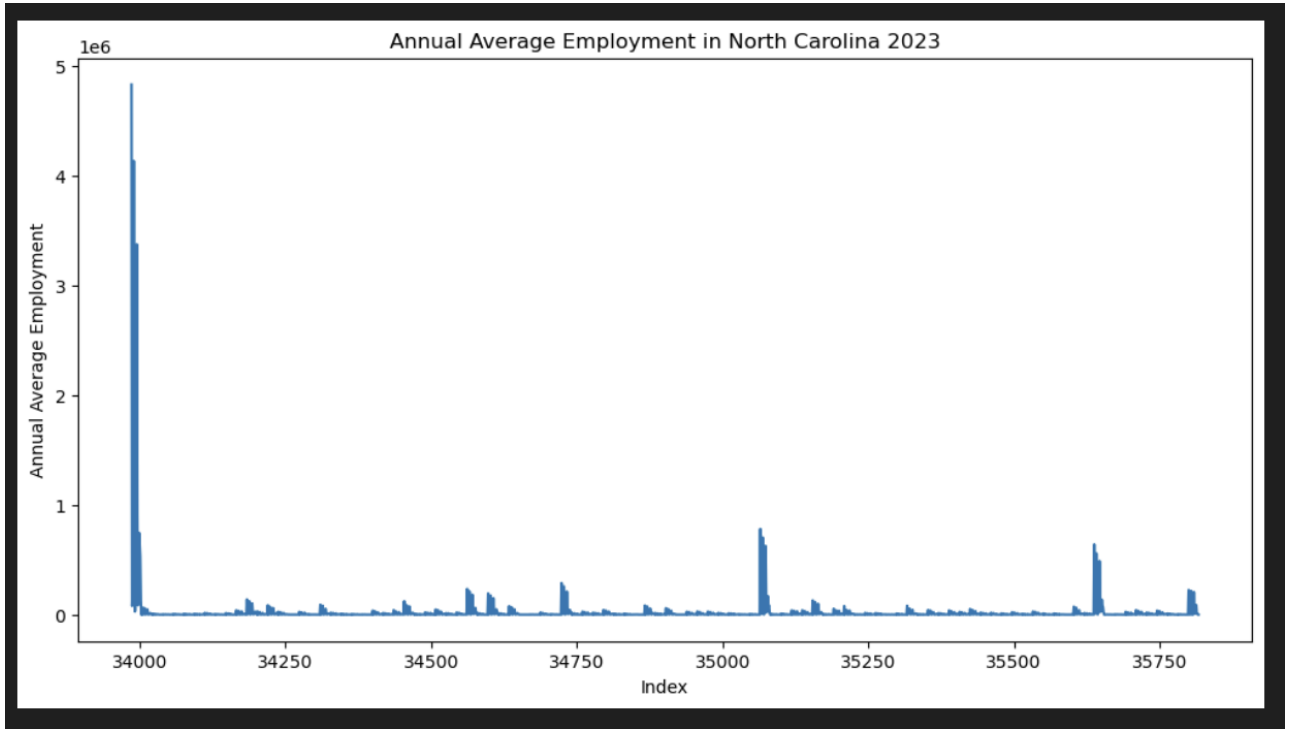
- **Filter Applied:** Extracted data for North Carolina.

```
# Calculate total cases, deaths, and population for the last week
nc_last_week['total_cases_last_week'] = nc_last_week[last_week_cases].sum(axis=1)
nc_last_week['total_deaths_last_week'] = nc_last_week[last_week_deaths].sum(axis=1)
nc_last_week['total_population_last_week'] = nc_last_week[last_week_population].sum(axis=1)

# Sum cases and deaths by day
daily_cases = nc_last_week[last_week_cases].sum(axis=0)
daily_deaths = nc_last_week[last_week_deaths].sum(axis=0)

# Display the result
print(nc_last_week.head())
```

```
      countyFIPS        County Name State  2023-07-17_cases  2023-07-18_cases  \
1890       37001   Alamance County     NC             62441             62441
1891       37003  Alexander County     NC             12844             12844
1892       37005  Alleghany County     NC              3810              3810
1893       37007      Anson County     NC              8549              8549
1894       37009       Ashe County     NC              7692              7692

      2023-07-19_cases  2023-07-20_cases  2023-07-21_cases  2023-07-22_cases  \
1890             62441             62441             62441             62441
1891             12844             12844             12844             12844
1892              3810              3810              3810              3810
1893              8549              8549              8549              8549
1894              7692              7692              7692              7692

      2023-07-23_cases  ...  2023-07-18_deaths  2023-07-19_deaths  \
1890             62441  ...                585                585
1891             12844  ...                161                161
1892              3810  ...                 23                 23
1893              8549  ...                114                114
1894              7692  ...                 96                 96

      2023-07-20_deaths  2023-07-21_deaths  2023-07-22_deaths  \
1890                585                585                585
1891                161                161                161
1892                 23                 23                 23
...
1893                798                              24446
1894                672                              27203

[5 rows x 21 columns]
```

Plotting Employment Trends:

- **Description:**
- Annual Average Employment in North Carolina for 2023.
- **Findings:** Trends observed in the latter half of 2023.



ACS Housing, Social, and Economic

Main Objective:

- Examine and interpret the social characteristics that may have occurred throughout the period of Covid-19.

My Hypothesis Questions:

- Does socioeconomic status affect the rate/severity of COVID-19?
- Do household sizes affect the rate of Covid-19?

Some relevant data:

- State
- County
- Geography
- County_FIPS

Covid-19 and Florida:

```
import pandas as pd
cases_data = pd.read_csv('covid_confirmed_usafacts.csv')
cases_df = pd.DataFrame(cases_data)

cases_df[cases_df.loc[:, 'State'] == 'FL']
```

Statewide Unallocated = **Stable**

										...									
0	Statewide Unallocated	FL	12	0	0	0	0	0	0	...	20363	20363	20363	20363	20363	20363	20363	20363	203

In this segment of code, we can see how I compare the confirmed covid-data set that we were given and then we compare it with the state that I chose, Florida.

When doing this, I found that it was consistent for Florida since the "statewide unallocated" stated the same number which was "20363."

Merging the Data:

- By merging the data, now I can analyze different counties and measure the correlation between COVID-19 case rates and severity.
- We can identify if different levels of income/education experienced different types of COVID-19 rates/outcomes.
- Now we could also identify if larger amounts in household sizes had increasingly more COVID-19 transmission rates.

```
import pandas as pd

data_df['countyFIPS'] = data_df['countyFIPS'].astype(int)

covid_data = pd.read_csv('final_merged_data.csv')
covid_data_df = pd.DataFrame(covid_data)

merged_data = pd.merge(data_df, cases_df, on='countyFIPS', how='outer')
merged_data.head()
```