**REPORT OF COVID19 DATA ANALYSIS PROJECT STAGE ONE**
**(Data and Project Understanding)**

## 1. Group One Members

The names of the group members are:

I. Neetha Ravva
II. Kol Herget
III. Kevin Hayes
IV. Ayodeji Iwayemi

## 2. Describing the Dataset and Data Type - Variable Dictionary

The description of the various data sets are presented in Table I, Table II, Table III, and Table IV. Table I contains the description of the Population dataset. It comprises the following fields: CountyFIPS, County Name, State, and Population. Table II represents the description of the COVID19 cases dataset. The dataset comprises the number of infections with covid 19 virus. The table has the following fields: CountyFIPS, County Name, State, StateFIPS, and Dates from 1/22/2020 till 7/23/2023

In a similar vein, Table III represents the variable description of COVID19 death dataset. The Covid19 deaths dataset comprises the number of deaths arising from covid 19 infections. The table has the following fields: CountyFIPS, County Name, State, StateFIPS, and Dates from 1/22/2020 till 7/23/2023.

Table IV indicates the description of ACS Social, Economic, and Housing

**Table I: Description of Population Dataset**

| Name | Definition | Data Type | Possible Values | Required? |
|---|---|---|---|---|
| CountyFIPS | Unique County ID (unknown counties have an ID of 0) | Integer | 0,1001,1003,1041,1027 | Yes |
| County Name | Name of County | Text | Baldwin County, Blount County, Conecuh County, Coffee County, Dallas County | Yes |
| State | Name of State | Text | AL, CA, CO, GA, FL | Yes |
| Population | Number of population. | Integer | 292256,156714, 16116, 132235, 149910 | Yes |

**Table II: Description of COVID19 Cases Dataset**

| Name | Definition | Data Type | Possible Values | Required? |
|---|---|---|---|---|
| CountyFIPS | Unique County ID (unknown counties have an ID of 0) | Integer | 1001, 1003, 1011, 1013, 56045 | YES |
| County Name | Name of County | Text | Crook County, Vilas County, Weston County, Goshen County, Park County | YES |
| State | Name of State | Text | TX, PA, WI, TN, SD | YES |
| StateFIPS | FIPS Code for a State | Text | 01, 02, 03, 04, 56 | YES |
| Date | Number of cases per day from 1/22/2020 to 7/23/2023 | Integer | 0, 1, 2, 3, 4,5 | YES |

**Table III: Description of COVID19 Deaths Dataset**

| Name | Definition | Data Type | Possible Values | Required? |
|---|---|---|---|---|
| countyFIPS | Unique County ID (unknown counties have an ID of 0) | Integer | 0, 1001, 1003, 51047, 56045 | Yes |
| County Name | Name of County | Text | Wake County, Orange County | Yes |
| State | Name of State | Text | TX, PA, WI, TN, SD | Yes |
| StateFIPS | State ID | Text | 01, 02, 10, 12, 56 | Yes |
| Deaths (all date columns) | Number of Death per day from 1/22/2020 to 7/23/2023 | Integer | 0, 26, 1310, 6472 | Yes |

**Table IV: Description of ACS Social, Economic, and Housing - Kevin Hayes**

| Name | Definition | Data Type | Possible Values | Required? |
|---|---|---|---|---|
| Geography | A unique number associated with each county.<br><br>This will not be used because I am only looking at the US, so the names of states and counties are sufficient for identification. | Text | 0500000US01003<br>0500000US01015<br>0500000US01043<br>0500000US01049<br>0500000US01051<br>0500000US01055<br>0500000US01069 | Yes. |
| Geographic Area Name | The name of the county, followed by a comma and the full name of its state. | Text | Baldwin County, Alabama<br>Calhoun County, Alabama<br>Cullman County, Alabama<br>DeKalb County, Alabama<br>Elmore County, Alabama<br>Etowah County, Alabama | Yes |
| Estimate | The estimated value of a given label | Number (Either Integer or Float. Depends on row) | 327,167,439, 97.0, 38.2, 52,423,114, 79.9 | Yes |
| Margin of Error | Absolute expected deviation from a given estimate. | Number (Either Integer or Float. Depends on row) | 27,812, 27,815, 0.1, 0.2, 34,697 | No |
| Percent | Percentage of a given sublabel. (Note: base level labels have either their initial value or (X) for this. This will be fixed in data cleaning) | Float | 49.2,<br>50.8,<br>6.0,<br>6.1,<br>6.5, | No |
| Percent | relative expected | float | 0.4 | No |

| | | | | |
|---|---|---|---|---|
| Margin of Error | deviation from a given estimate (Note: labels where this does not apply have either their initial value or (X) for this. This will be fixed in data cleaning) | | 0.3<br>0.6<br>0.9<br>0.5 | |

- We can join this to our main dataframe by taking a counties state and name and using that to generate its Geographic Area Name, which we can use as a foreign key to map between the two tables.

- All of these are followed by !!<label>!!<sublabel> to denote where the value applies to. These labels are SEX AND AGE, RACE, Race alone or in combination with one or more other races, HISPANIC OR LATINO AND RACE, and CITIZEN, VOTING AGE POPULATION. The sublabels refer to various races or qualifiers to the main label. In practice I plan to break up each label into its own dataframe.

- This enrichment data can help analyze the spread of COVID-19 by allowing us to see if areas with people of given ages or races were more or less likely to have people infected with or killed by COVID-19.

- Potential hypothesis questions:
    - Is there a correlation between mean age in a county and the number of COVID-19 deaths in said county?
    - Is there a correlation between mean age in a county and the number of COVID-19 infections in said county?
    - Is there a correlation between the percentage of people of age 85 or over in a county and the number of COVID-19 deaths in said county?
    - Is there a correlation between the percentage of people in a county that are hispanic or latino and deaths from COVID-19?

**Table V: Description of Employment Variables**

| Name | Definition | Data Type | Possible Values | Required? |
|---|---|---|---|---|
| area_fips | County FIPS | Text | 05123, 01041, 56045 | Yes |
| own_code | 1-character Ownership code | Text | 0, 1, 2, 3, 4, 5, 8, 9 | No |
| industry_code | NAICS industry code | Text | 49111, 92, 522298 | No |
| agglvl_code | Code for aggregation level | Text | 74, 77, 78 | No |
| size_code | Establishment size code | Text | 0 | No |
| year | Census year | Text | 2020, 2021, 2022 | Yes |
| qtr | Fiscal quarter (always A for annual) | Text | A | No |
| disclosure_code | 1-character disclosure code | Text | “”(blank), N | No |
| area_title | County name (based on area_fips) | Text | St. Francis County, Arkansas; Crenshaw County, Alabama; Weston County, Wyoming | No |
| own_title | Type of ownership (based on own_code) | Text | Federal Government, State Government, Private | No |
| industry_title | Industry title (based on industry_code) | Text | NAICS 49111 Postal service; NAICS 92 Public administration; NAICS 522298 All other nondepository credit intermediation | No |
| agglvl_title | Title of level of aggregation (based on agglvl_code) | Text | County, NAICS Sector -- by ownership sector; County, NAICS | No |

| | | | | |
|---|---|---|---|---|
| | | | 5-digit -- by ownership sector; County, NAICS<br>6-digit -- by ownership sector | |
| size_title | Size title (based on size_code) | Text | All establishment sizes | No |
| annual_avg_estabs | Annual average of quarterly establishment counts for a given year* | Numeric | 385, 109, 254 | No |
| annual_avg_emplvl | Annual average of monthly employment levels for a given year* | Numeric | 4756, 1231, 2149 | Yes |
| total_annual_wages | Total wages for the fiscal year | Numeric | 217856056, 82305381, 101598481 | Yes |
| taxable_annual_wages | Sum of the four quarterly total taxable wage totals for a given year* | Numeric | 36984510, 10206971, 39354459 | Yes |
| annual_contributions | Sum of the four quarterly contribution totals for a given year* | Numeric | 567653, 100344, 946941 | No |
| annual_avg_wkly_wage | Average weekly wage based on the 12-monthly employment levels and total annual wage levels* | Numeric | 881, 1286, 909 | Yes |
| avg_annual_pay | Average annual pay based on employment and wage levels for a given year* | Numeric | 45809, 66865, 47270 | Yes |
| lq_disclosure_code | 1-character location-quotient disclosure code* | Text | “”(blank), N | No |

| | | | | |
|---|---|---|---|---|
| lq_annual_avg_estabs | Location quotient of annual average establishment count relative to the U.S. (Rounded to hundredths place)* | Numeric | 1 | No |
| lq_annual_avg_emplvl | Location quotient of annual average employment relative to the U.S. (Rounded to hundredths place)* | Numeric | 1 | No |
| lq_total_annual_wages | Location quotient of total annual wages relative to the U.S. (Rounded to hundredths place)* | Numeric | 1 | No |
| lq_taxable_annual_wages | Location quotient of taxable annual wages relative to the U.S. (Rounded to hundredths place)* | Numeric | 1 | No |
| lq_annual_contributions | Location quotient of total annual contributions relative to the U.S. (Rounded to hundredths place)* | Numeric | 1 | No |
| lq_annual_avg_wkly_wage | Location quotient of annual average weekly wage relative to the U.S. (Rounded to hundredths place)* | Numeric | 1 | No |
| lq_avg_annual_pay | Location quotient of annual average pay relative to the U.S. (Rounded to hundredths place)* | Numeric | 1 | No |
| oty_disclosure_code | 1-character over-the-year disclosure code (either ' '(blank) or 'N' not | Text | “”(blank), N | No |

| | | | | |
|---|---|---|---|---|
| | disclosed)* | | | |
| oty_annual_avg_estabs_chg | Over-the-year change in annual average establishments for a given year* | Numeric | 15, -8 | |
| oty_annual_avg_estabs_pct_chg | Over-the-year percent change in annual average establishments for a given year (Rounded to the tenths place)* | Numeric | 4.1, -6.8, -3.1 | No |
| oty_annual_avg_emplvl_chg | Over-the-year change in annual average employment for a given year* | Numeric | 103, -58, -17 | No |
| oty_annual_avg_emplvl_pct_chg | Over-the-year percent change in annual average employment for a given year (Rounded to the tenths place)* | Numeric | 2.2, -4.5, -0.8 | Yes |
| oty_total_annual_wages_chg | Over-the-year change in the total annual wages for a given year* | Numeric | 3927403, 3743942, 2850032 | No |
| oty_total_annual_wages_pct_chg | Over-the-year percent change in total annual wages for a given year (Rounded to the tenths place)* | Numeric | 1.8, 4.8, 2.9 | No |
| oty_taxable_annual_wages_chg | Over-the-year change in taxable annual wages for a given year* | Numeric | 66827, -603070, 3088805 | No |
| oty_taxable_annual_wages_pct_chg | Over-the-year percent change in taxable annual wages for a given year (Rounded to the tenths place)* | Numeric | 0.2, -5.6, 8.5 | No |

| | | | | |
|---|---|---|---|---|
| oty_annual_contributions_chg | Over-the-year change in annual contributions for a given year* | Numeric | 19847, -16026, 143365 | No |
| oty_annual_contributions_pct_chg | Over-the-year percent change in annual contributions for a given year (Rounded to the tenths place)* | Numeric | 3.6, -13.8, 17.8 | No |
| oty_annual_avg_wkly_wage_chg | Over-the-year change in annual average weekly wage for a given year* | Numeric | -3, 114, 32 | No |
| oty_annual_avg_wkly_wage_pct_chg | Over-the-year percent change in annual average weekly wage for a given year (Rounded to the tenths place)* | Numeric | -0.3, 9.7, 3.6 | Yes |
| oty_avg_annual_pay_chg | Over-the-year change in average annual pay for a given year* | Numeric | -165, 5933, 1676 | No |
| oty_avg_annual_pay_pct_chg | Over-the-year percent change in average annual pay for a given year (Rounded to the tenths place)* | Numeric | -0.4, 9.7, 3.7 | Yes |

* This definition is taken from the U.S. Department of Labor Statistics QCEW annual field layout

- We can use the area_FIPS variable to merge data with the COVID dataset.
- We could look at the correlation between Employment levels and wages and COVID infection and death rates
- Hypothesis questions:
    - Is there a correlation between employment levels and COVID infection/death rates?
    - Is there a correlation between wages and COVID infection/death rates?

**Table VI : Description of Presidential Dataset**

| Name | Definition | Data Type | Possible Values | Required? |
|---|---|---|---|---|
| State | Name of State | Text | Florida, Geaogia Hawaii, Idaho, Illinois | Yes |
| County | Name of county | Text | Kent County, Ward 2, Duval County, Hardee County, Highlands County | Yes |
| Candidate | Name of Contestant | Text | Joe Biden, Donald Trump, Jo Jorgensen, Howie Hawkins | Yes |
| Party | Name of Political Party | Text | DEM, REP, LIB, GRN | Yes |
| Total-Vote | Vote Count | Integer | 420, 1282, 1003, 259, 56682 | Yes |
| Won | Winner in the Polls | Boolean | True, False | Yes |

- Presidential dataset is the table containing the outcomes of the 2020 presidential poll in the USA. This dataset provides detailed information on the 2020 USA presidential election results. The table has six fields namely State, County, Candidate, Political Party, Total votes per candidate, and the winner per county.
- To join the two tables, a common variable that exists in both tables is essential. In this case, the common variables are "State" and "County" since they are present in both the presidential results table and the COVID-19 cases table. So, the variable that can be used to join the two tables would be a combination of "State" and "County". When performing the join operation, the rows in the presidential results table will be matched with the corresponding rows in the COVID-19 cases table based on the values in these two columns.
- Hypothesis

- Is there a correlation between the political affiliation and the spread of Covid-19 cases/deaths
- Is there a correlation between the victory of a candidate and the spread of the covid-19 cases/deaths
- Is there a correlation between total votes cast and the COVID 19 cases/deaths