Investigate whether US states (California) with higher population have a higher Covid-19 rate

Kim Ngoc Nguyen

Singelyn Graduate School of Business at Cal Poly Pomona

GBA 6070: Programming Foundation for Business Analytics

Dr. Fadi Batarseh

December 3, 2023

**FINAL PROJECT**

Analytical question: Investigate whether US states (California) with higher population have a higher Covid-19 rate.

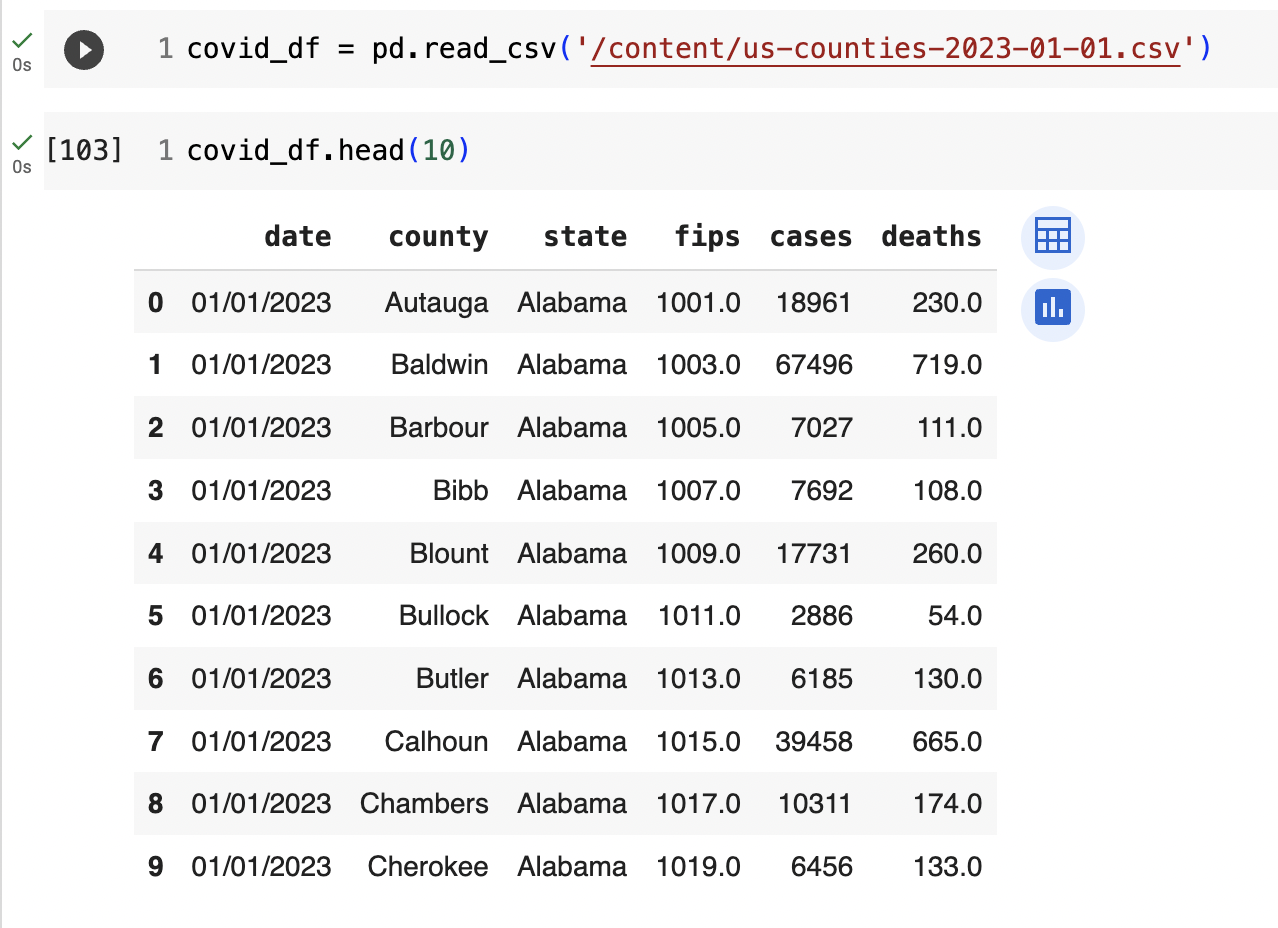
**I. Introduction**

In this study, I conduct an investigation into the relationship between population size and covid19 rates in different states across the United States. The dataset chosen contains detailed data on confirmed cases, population statistics, and relevant variables. My goal is to see if states with larger populations have higher covid19 rates and whether in that state, counties which have higher population records attach to a higher covid rate. This investigation is critical for informing public health policies and resource allocation, as well as providing insights into the pandemic's regional dynamics and contributing to effective mitigation strategies. The datasets are extracted from the official site of the NYTimes and are made public by the federal government.

**II. Methodology**

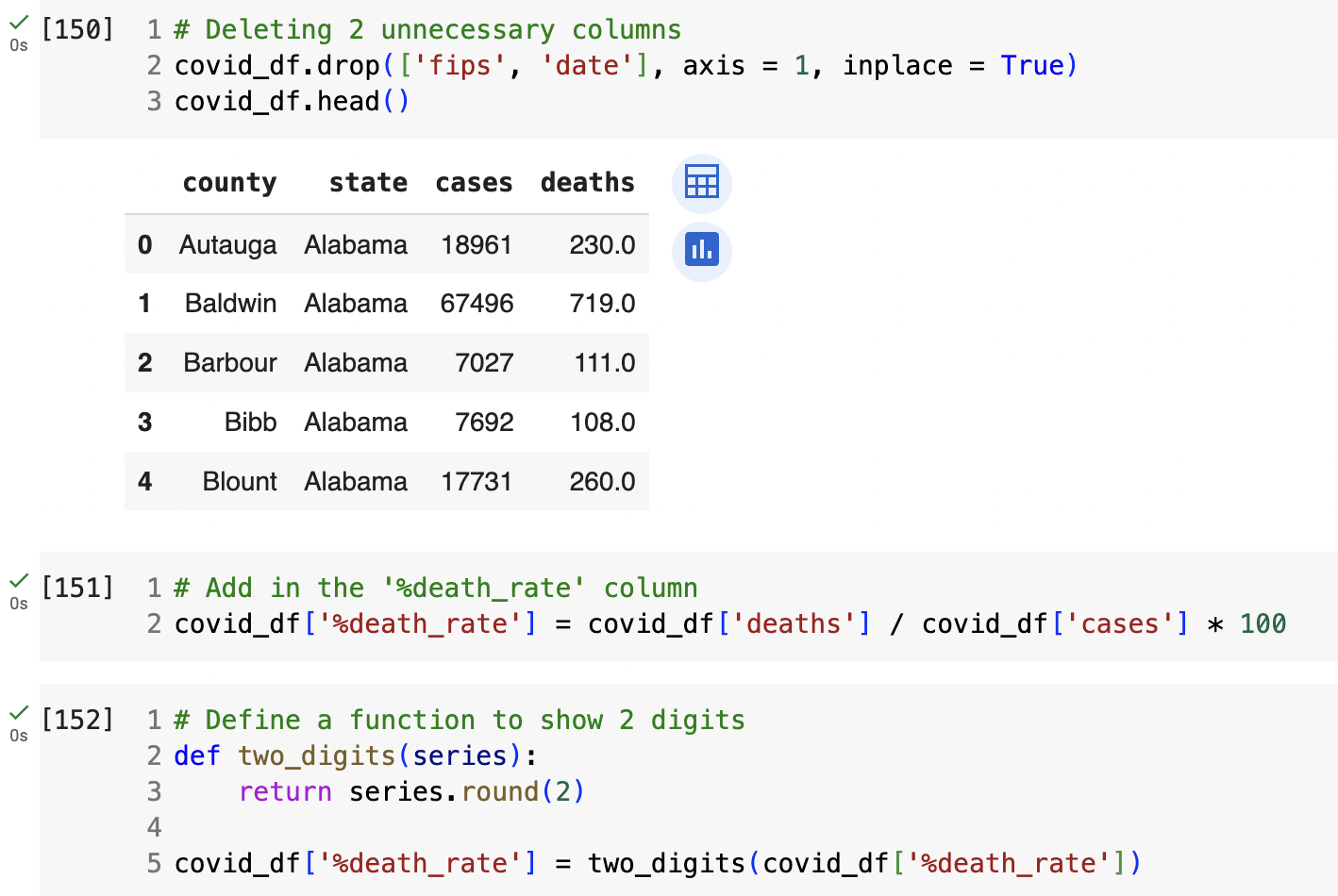
The project will be represented following the flow of the coding workspace.

The first chosen dataset if about number of covid19 cases and deaths recorded in the US from the very first period till 2023-01-01.



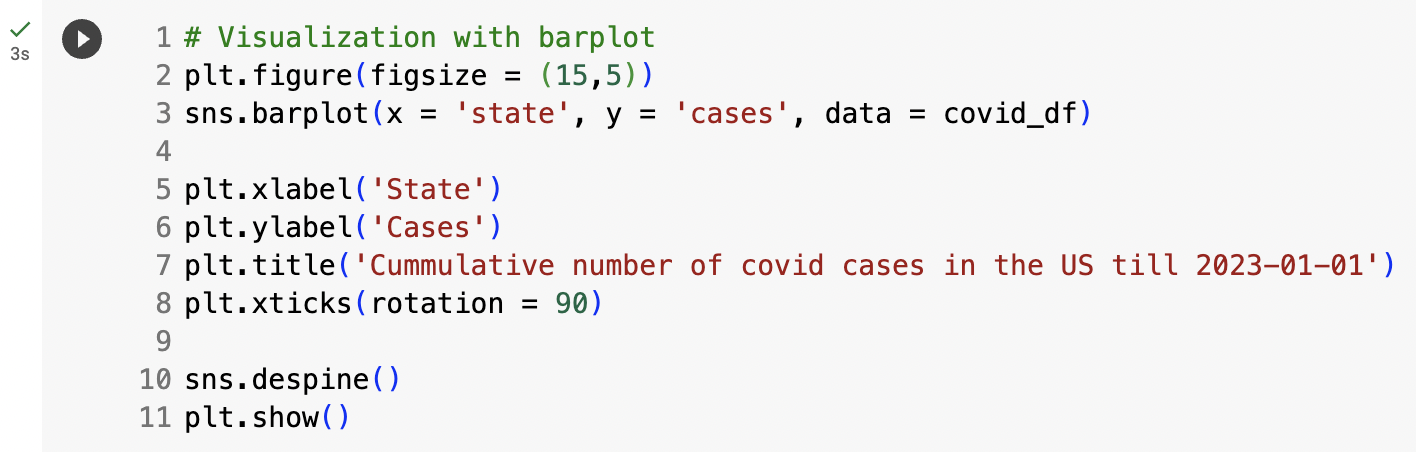
The “covid\_df” dataset consist of 3255 rows of every county in every state with covid19 records. The dataset includes the date, county name, state name in which exact number of cases and deaths are calculated cumulatively. With this information, I can calculate the rate based on number of cases easily.

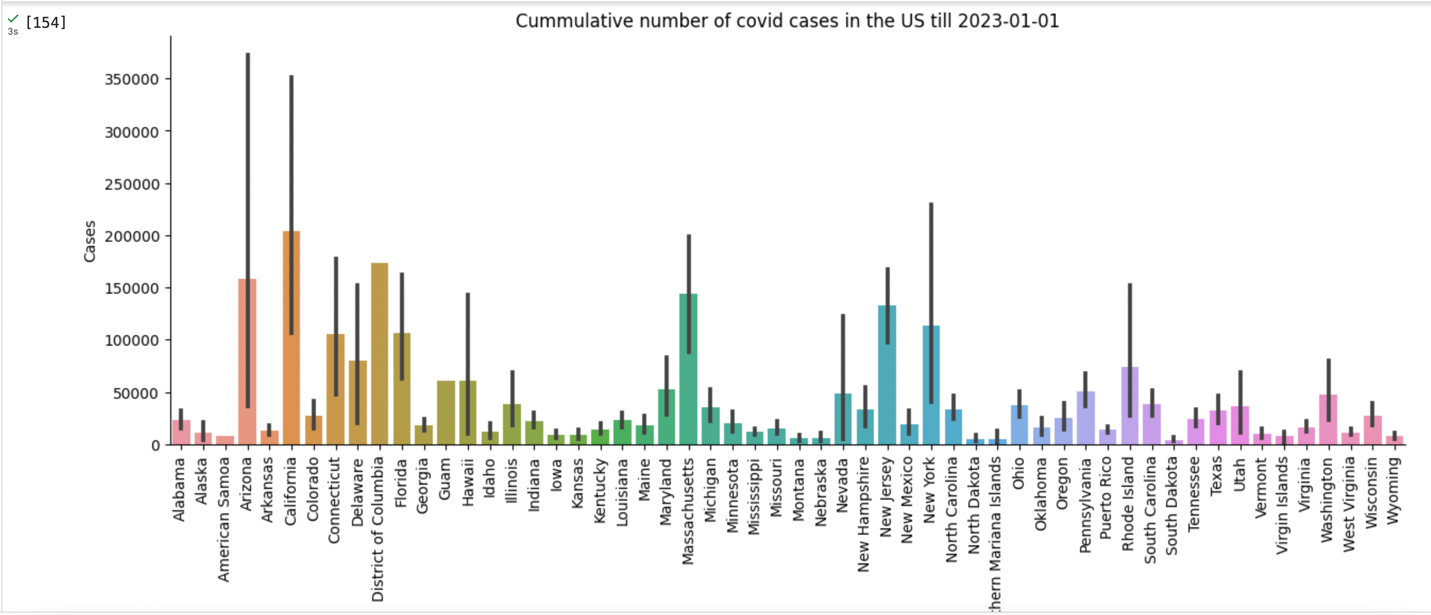
After reading the dataset, there were some unnecessary data so cleaning the dataframe with dropping out some unnecessary columns is the next step.



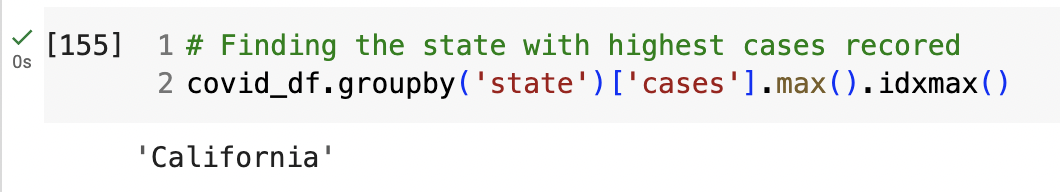
In the line [152], a function is defined to round up calculation to show only 2 digits. Since analysis will definitely require plenty calculation, having a to-go function will cut off some time and make the workspace clearer. The provided code defines a function named two\_digits that rounds the values in a given pandas series to two decimal places. Subsequently, the code applies this function to the '%death\_rate' column in a DataFrame called covid\_df. The result is stored in a new column named '%death\_rate', ensuring that the death rate values are now rounded to two digits.

After cleaning and obtaining vital information for comparing data, I moved on with graph visualization to show the most imaginable insights of the covid19 situation in the US. The bar plot illustrates all US states with exact number of cases recorded from the beginning of the covid19 pandemic till January 1st, 2023 using the “barplot” function from the seaborn library.

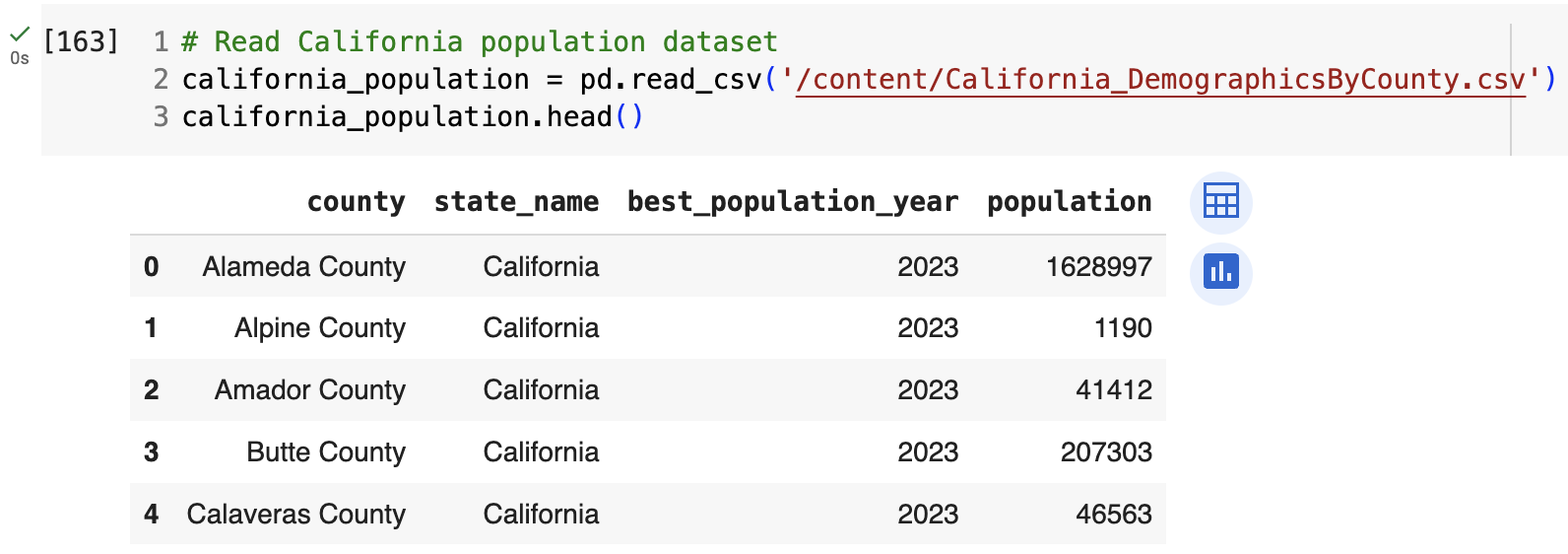
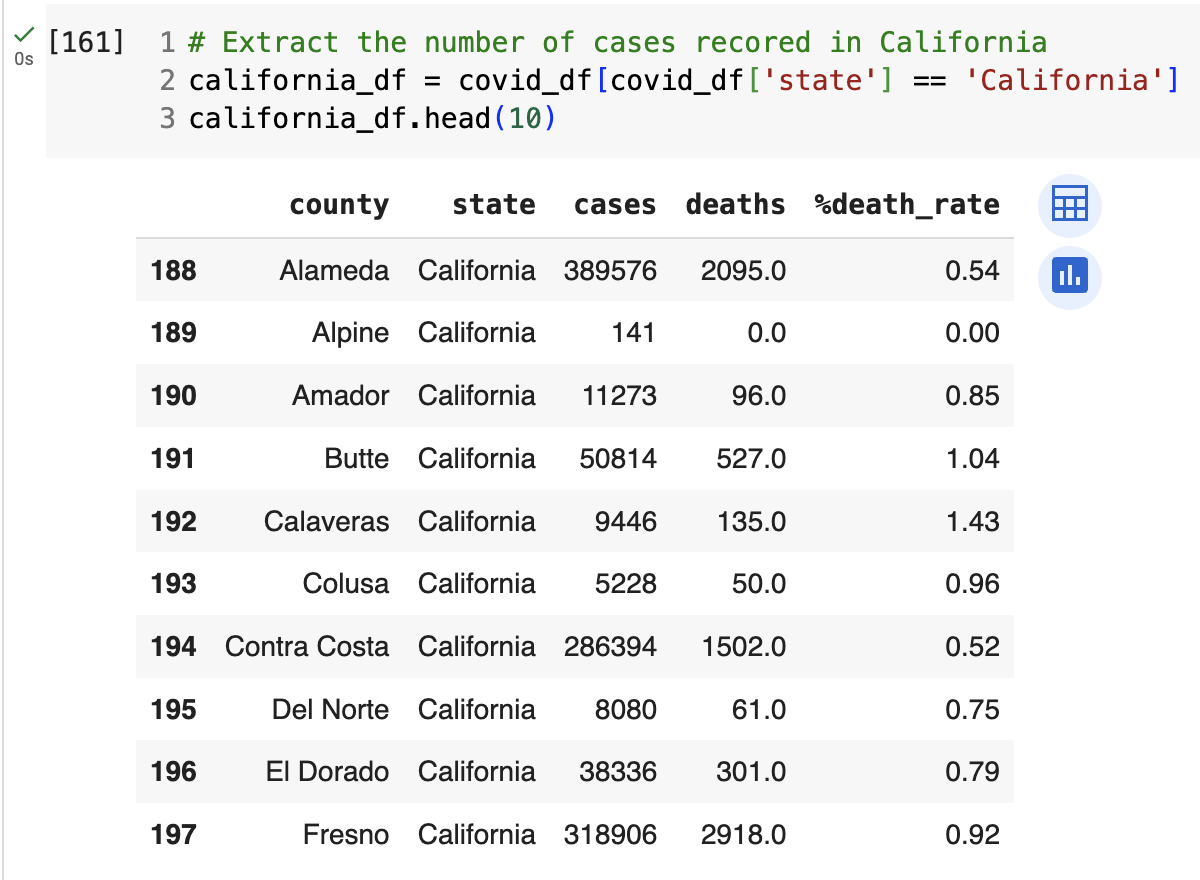




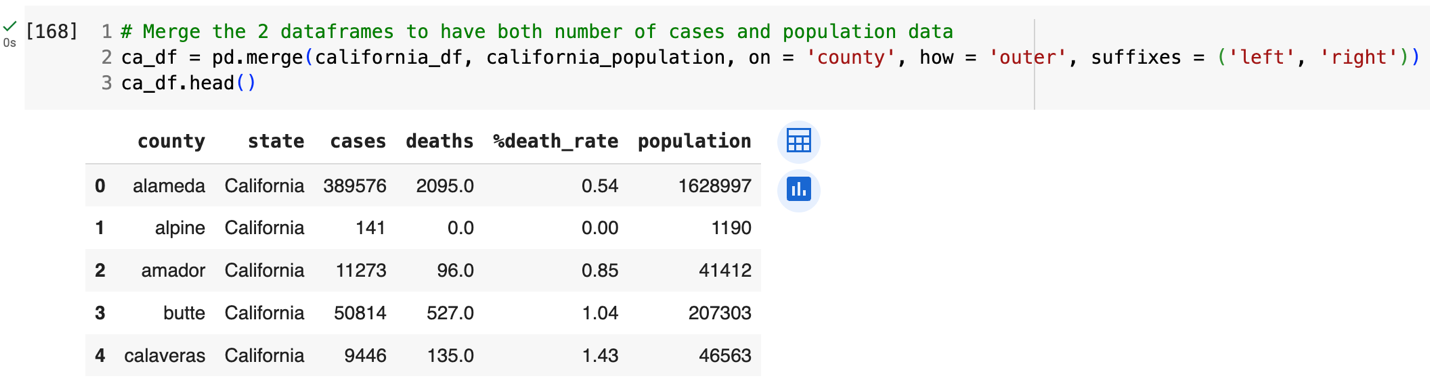
This barplot effectively conveys the distribution of covid cases across different states, allowing for quick comparisons and identification of states with higher case numbers. A quick glance at the graph shows that California posses highest records in terms of covid cases. To double-check the result, I used another code snippet to find the state with the highest recorded number of covid19 cases in the DataFrame covid\_df. With the ‘group by’ technique, only ‘cases’ is calculated and ‘idxmax’ allows to find the index associated with the maximum value.



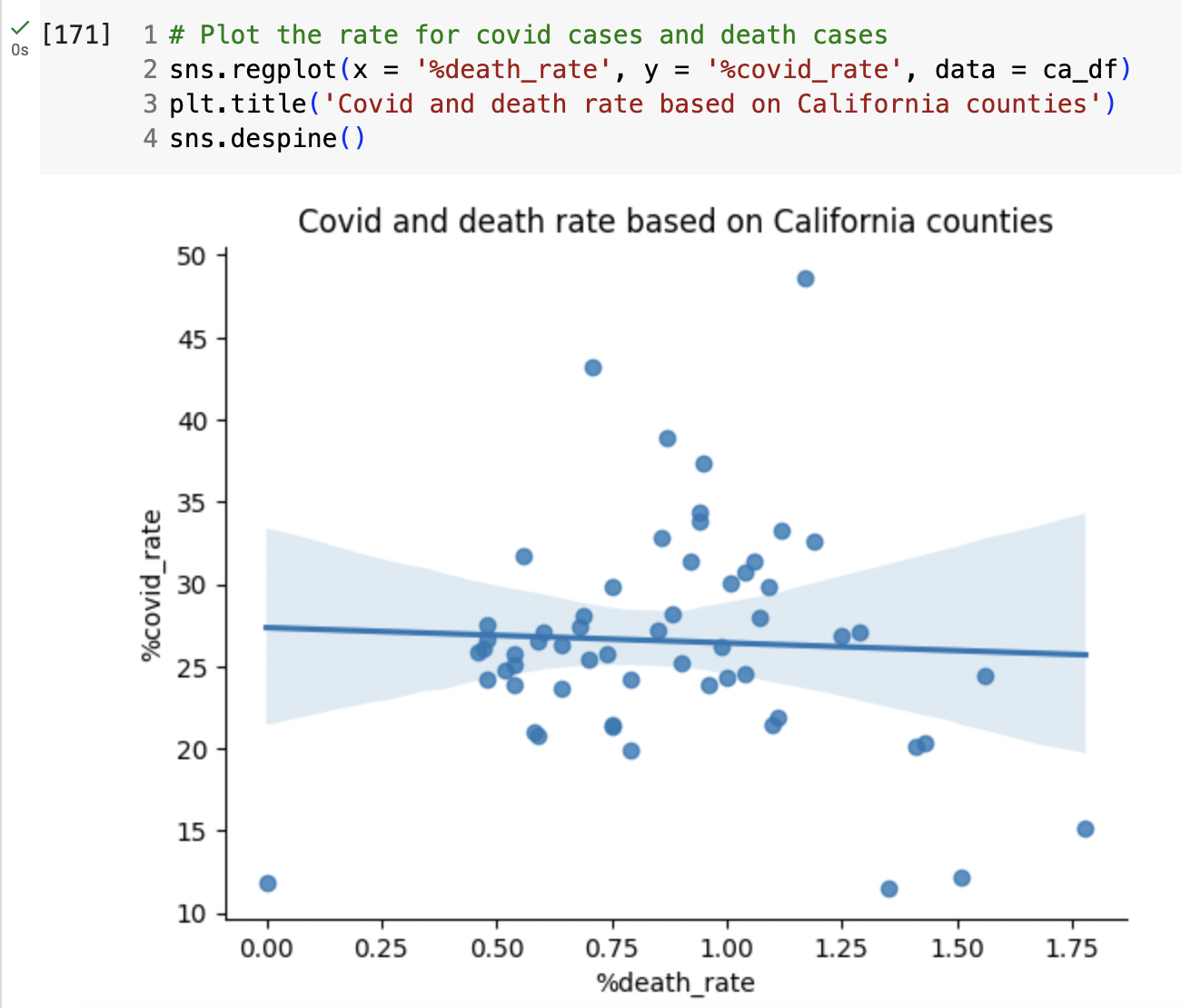
With this result, I decided to insert the California population dataset to study more on whether the higher number of state citizens possess higher rate of covid 19 disease.



First, the California-related data is extracted from the original dataframe and then the California population dataset was inserted based on county to match with the ‘california\_df’. It is then followed by the cleaning and screening data.

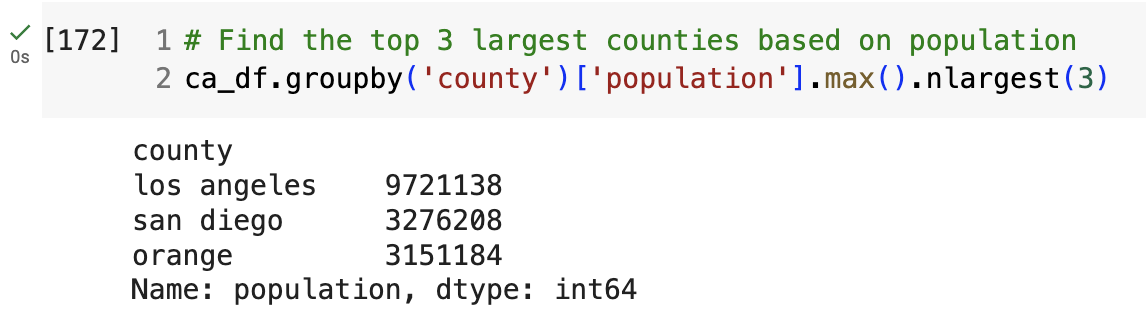


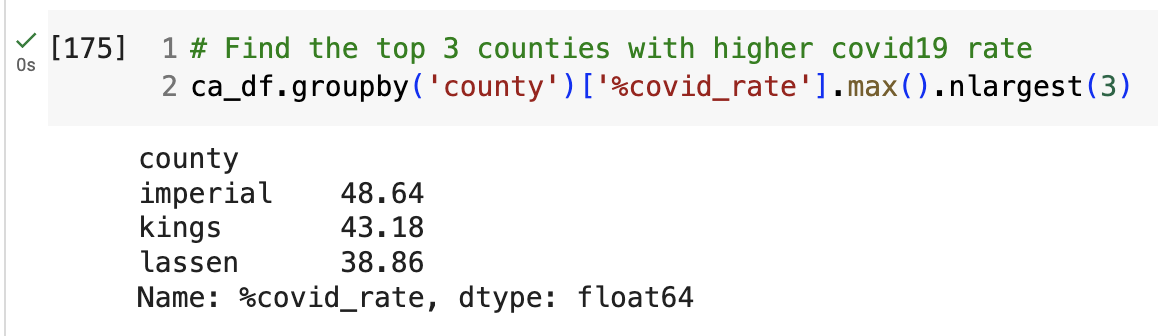
Merging the 2 dataframes using pandas outer ‘merge’ function, stored in the new ‘ca\_df’ with all 2 former dataset information.

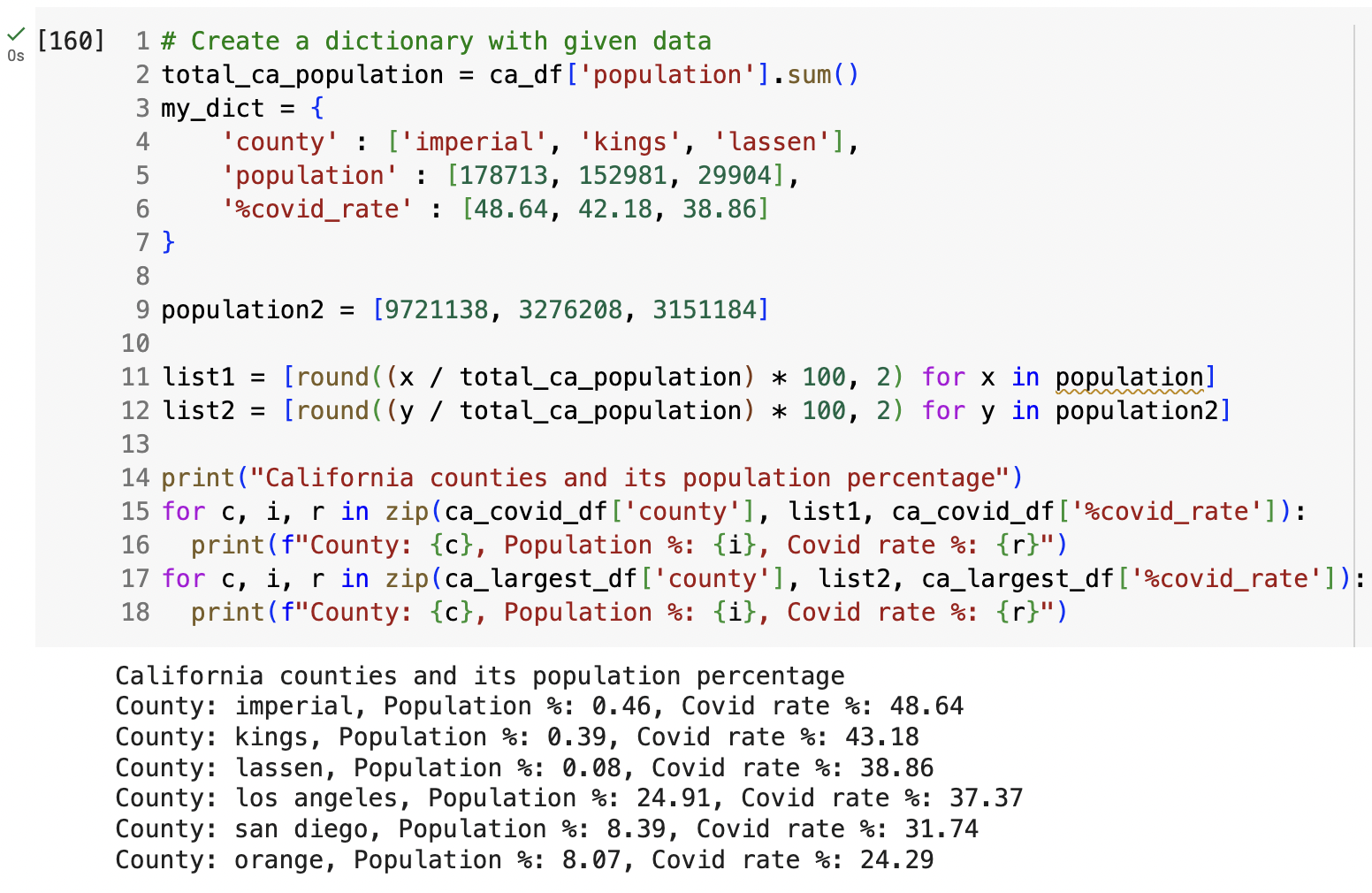


In this stage, the scatter plot is used to represent the trends and patterns between the number of cases and deaths recorded in California. The central trend line indicates the average relationship or correlation between the percent death rate and the percent COVID rate. With the flat trend line, the two variables may have a weak correlation.

To get more information about the relationship of number of cases and population data, some information on counties with highest population and counties with highest covid19 rate are extracted.

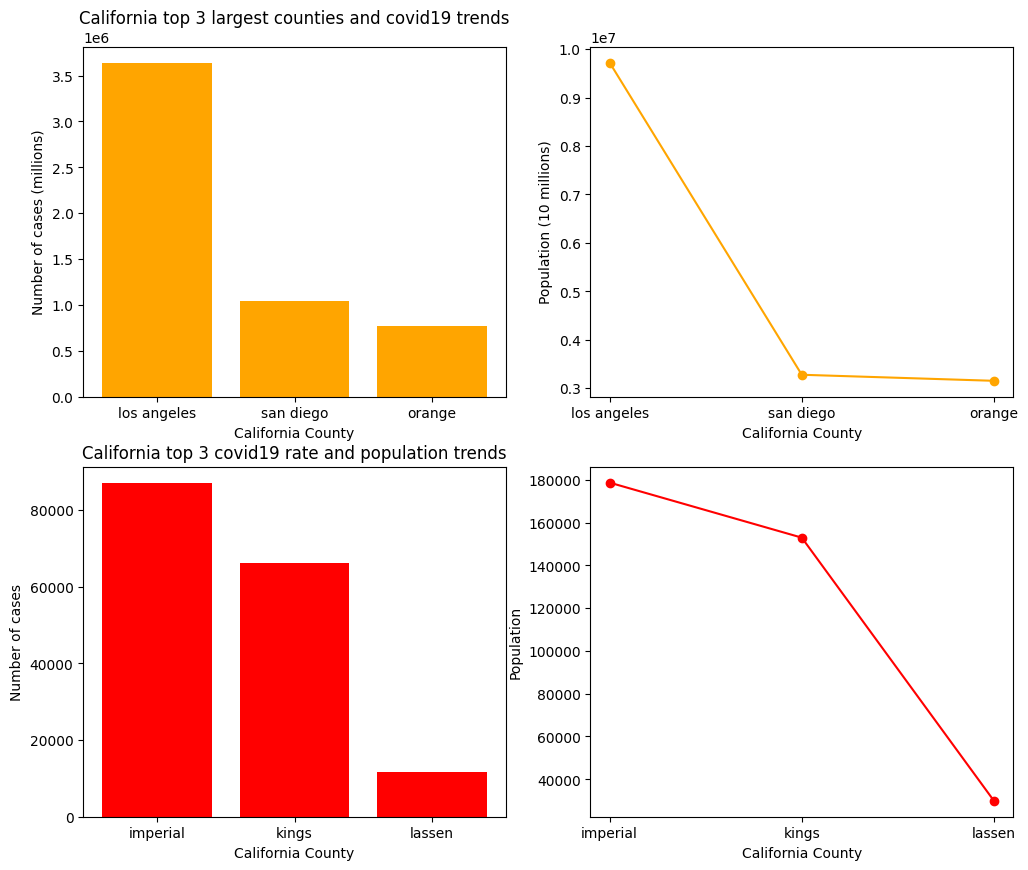






Here, dictionaries store related data in Python code, lists hold county populations, and list comprehensions efficiently calculate population percentages. The added list to make calculation easier and quicker than extracting or directing its part in the dataframe. While the ‘for’ loops with the ‘zip()’ function iterate over data sets to output county-specific covid and population rates, rounding numbers and printing them with ‘f-strings’. These constructs offer a structured, readable, and efficient method of manipulating and presenting data. The printouts show necessary insights on population and number of cases only.

**III. Analysis and Findings**



This data suggests that smaller counties can have high covid rate percentages due to a variety of factors such as localized outbreaks, testing rates, or reporting practices. Meanwhile, despite having lower covid rate percentages, larger counties like Los Angeles represent a significant number of cases in absolute terms due to their larger populations. These findings emphasize the importance of taking into account both the rate of covid19 cases and population size when assessing the pandemic's impact on different regions.

**IV. Conclusion**

After investigating into the US in general and California state in specific, covid19 rates vary across California counties and do not appear to be directly related to county population percentages. Covid-19 rates are significantly higher in counties with smaller populations, such as Imperial, Kings, and Lassen, when compared to population percentages. While Los Angeles, the most populous county, has a lower covid19 rate relative to its population percentage, indicating a better-than-expected outcome given the dense population. To comprehend the factors influencing the spread in these counties, more research is necessary. Population density, public health regulations, socioeconomic status, and access to healthcare are a few examples of such variables.

**V. Reflection**

During the studying of US and California covid and population trends, I realized some important values. Firstly, the power of visualization makes the numbers, calculations and data readable and imaginable. Efficient data visualization made it easier to identify patterns and gain insights quickly, which is why it's crucial for both presenting and analyzing large amounts of complex data. Secondly, the disparities in covid19 impact were found in localized county data, highlighting the significance of granular analysis for focused public health initiatives.

**VI. References**

US states 2023 covid dataset: The New York Times. (2023). *COVID-19 Data in the United States*. GitHub. Retrieved from <https://github.com/nytimes/covid-19-data>

California population dataset: California Demographics. (n.d.). Demographics by City, County or Zip Reports. Retrieved from <https://www.california-demographics.com/demographics_reports>