DSC540 T302 Project

Dipika Sharma
Bellevue University

Project Subject Area

It's been more then 2 years since we first heard about the coronavirus. Even with all the facilities available today we are not able to completely get rid of it and still it is a big news. As my term project I want to analyze the covid data to get answer to questions. I will perform the deep analysis on California state covid data to see if vaccination reducing the number of covid cases, which county is most effected by covid in California, which county doing well and others.

Data Sources

Below are the different data sources I will use while working on my project.

1. Flat File:
    - Using the HealthData government site to see the count of coronavirus cases, number of deaths in California state by county. https://healthdata.gov/browse?tags=hhs+covid-19
    - Also, I am using the IRS site to get the average income of each county of California. https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-zip-code-data-soi
2. API:
    - Using the CHHS government site to get the data APIs, this includes information about the vaccination in California state by zip code.
    - https://data.chhs.ca.gov/dataset/covid-19-vaccine-progress-dashboard-data-by-zip-code/resource/235174a5-f5b6-4759-8ab9-76191bfea324
3. Website:
    - GeoNames website gives us all the county name and postal code of California state.
    - https://www.geonames.org/postalcode-search.html?q=&country=US&adminCode1=CA

Relationships

I will be studying the covid data for California state counties, so I tried to find all the data sources which gives information based on the counties and zip code. The csv files which I got it from HealthData government website has covid data for all the counties of California. Along with Covid data I am also using the csv files from IRS site which will give me the average income of each county of California state. I think it would be interesting to see the trends of covid vaccination in lower income county and in higher income county.
I am using geonames website to get all the counties and its postal code of California state. CHHS government site will give me data APIs to get the vaccination information of California state by zip code.

All the data sources are related to each other with county name or postal codes. Before cleaning or processing these data sources for analysis I will add the county name or postal code to each data set using the geonames website.

Project Steps and Challenges

I will start the project by defining its objective. If I have the question to begin with, it would be easier for me to decide the project steps. I will divide the project into small steps as it will make easier for me to track the project status and tackle the issues.
Different steps as follow:

- Decide project objective, what I want to achieve at the end of the project.
- Do I have the correct data sources to achieve the project objective?
- Cleaning and preparing of the data. In this step I will see if data has any null values or values which do not make any sense or look for any outliers in dataset. Do I have to do any formatting of data before analyzing the data?
- Perform data analysis and exploration to detect the different patterns and trends in the dataset. I will be able to get useful insights of the data in this step which will helps me learn about the behavior of data in different cases.
- Present the information so that even the non-technical person can understand it.
- Create a final report which shows answer to every questions. And see if we achieved the project objective or not.

I am using the website geonames to get the county and postal code of California state. It would be challenging to get the information from this website and save it to csv file so that I can use it easily. Also, I understand the most important step in order to get the accurate analysis is to use the correct and clean data. And since I am using the data available online my next challenge it to see if data is good and does not have any junk values.