Theodore Koby-Hercsky

September 24, 2021

DSC540-T301

**Milestone One**

1. **CSV File - Provisional COVID-19 Deaths by Sex and Age** 
   1. **Website:** <https://data.cdc.gov/NCHS/Provisional-COVID-19-Deaths-by-Sex-and-Age/9bhg-hcku>
   2. **Description –** In the csv file that is provided in the CDC website link attached above it goes over the deaths that are involving the coronavirus disease. Along with the death counts for pneumonia and any influenza that have been reported. Along with the variables such as state, sex, and age that will be used to help tie the three data sets together. As this data set will allow me to determine the number of deaths by state for each gender and the age of the individuals that dies by either Corona virus, pneumonia, influenza, and or other causes.
2. **API - COVID-19 Vaccination Trends in the United States, National and Jurisdictional** 
   1. **Website:** <https://dev.socrata.com/foundry/data.cdc.gov/rh2h-3yt2>
   2. **CDC Website:** <https://data.cdc.gov/Vaccinations/COVID-19-Vaccination-Trends-in-the-United-States-N/rh2h-3yt2>
   3. **API Endpoint:** <https://data.cdc.gov/resource/rh2h-3yt2.json>
   4. **Description –** The API website and endpoint provided above has a detailed data set on the vaccination trends that are currently going on in the United States. Which goes over the overall number of Covid-19 Vaccinations that have been administered in the United States and the jurisdictional partner clinics, retail pharmacies, long term care facilities, dialysis centers, and or any other facilities that are administering the vaccine. This API will give us variables such as the date, week of administering, and location that will tie the number of vaccines administered to our other data frames that go into detail on the death rate in each state for each date.
3. **Website - Social Vulnerability Index Vaccine Hesitancy for COVID-19: County and local estimates**
   1. **Website:** <https://data.cdc.gov/Vaccinations/Social-Vulnerability-Index/ypqf-r5qs>
   2. **Data Source:** <https://data.cdc.gov/Vaccinations/Vaccine-Hesitancy-for-COVID-19-County-and-local-es/q9mh-h2tw>
   3. **Description** – In this data set that is stored in a table on the CDCs website it goes over the percentage of adults in the United States that unsure, probably not, and definitely not going to get the vaccine. While I will use the state variable to tie to my other data frames to determine which states had the highest hesitancy towards the vaccine and see if that played a part in an increase in death rates due to the Covid-19 virus.
   4. **Screenshot of data stored in a table from the CDC Website:**

Table

Description automatically generated

**Relationship Between Data Sets & What is Needed to Accomplish All Five Milestones**

In the three data sets above the variables that they have in common with each other are the State that they are in and the date that the data was taken. While our CSV file also gives us extra variables such as the sex of the individual along with their age. I feel as if those variables are also useful as it allows us to determine the age of individuals that have died from this virus. While also utilizing all three data sets to determine if the hesitancy of some adults in certain states caused the number of Covid-19 deaths to rise. While the data set that focuses on vaccination rates will tell me if the states with a higher rate of vaccination had a lower impact on death rates and if the hesitancy of unvaccinated adults was lower as well.

During this semester I will need to clean each data set and update the variable names so that they match each other before I join the three data sets together. When taking my first looks at these data sets, I already see that some of the data will need to be condensed and trim off any excess data that is not needed for my analysis. While I believe the most important variable across all data sets will be the states variable as this will allow me to connect each data set together and determine the number of deaths, vaccination rate, and the percentage of hesitancy in getting the vaccine has played during this pandemic. In the final milestone it will be interesting to see the visualizations and what is determined when all three data sets are merged and brought together to see which states are better off and which one’s individuals that are vaccinated would not want to live in.

**Week 5: What is Tabula? What is it used for?**

Great post I was working on a post on Tabula as well, but I saw you posted it, so I’ll just add some extra facts I found lol. As this is a tool that can be used for Windows or Mac that allows you to liberate data tables that are locked inside PDF files. While also having a source from GitHub that can be seen in your article attached. But by using this tool and simple interface you can extract the data into a CSV or Microsoft Excel spreadsheet. While Tabula is noted to be used by news organizations such as The New York Times and The Times of London to power investigative reporting.

Reference: <https://tabula.technology/>

**Week 5: What is fuzzy matching? What are some challenges you might face when using fuzzy matching?**

Fuzzy matching like the name sounds like it allows you to identify non-exact matches for a target item that you are searching for. Which is known as a foundation stone for many search engine frameworks. That allows you to get relevant search results even if there was a typo in the query or a different verbal tense. As there might be many algorithms that can be used for fuzzy searching on text data the Levenshtein Distance is primarily used for fuzzy string matching.

**Challenges with fuzzy matching**

While a challenge one might face with fuzzy matching is that it messes with the relevance. While the Damerau-Levenshtein algorithm is most common for user misspellings. That can also include a significant level of false positives that stain from when we use a language with an average of just five letters per word. But in the article attached below you can see more examples:

Reference: <https://blog.couchbase.com/fuzzy-matching/>