ETL Final Report

My project is based around finding the amount the United States invested in the health department over the years and looking at it with outbreaks of virus or diseases to see how well each event was handled. This will be compared to the number of cases on the COVID-19 and their funding within the US.

**Extract**

I selected National Health Expenditure data, Ebola data, and HIV/AIDS data, COVID-19 cases, and COVID-19 funds.

The sources are listed here.

NHE – <https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NationalHealthAccountsStateHealthAccountsResidence>

Ebola – <https://data.humdata.org/dataset/ebola-cases-2014>

HIV/AIDS – <https://data.world/chhs/2269414f-024a-47fa-b055-ee3b28dda27e/workspace/file?filename=hiv-aids-cases-2.xlsx>

All three are originally CSV files.

COVID-19 Cases – <https://covid19api.com/>

This is an API.

COVID-19 Funds – <https://data.cdc.gov/Administrative/Provider-Relief-Fund-COVID-19-High-Impact-Payments/b58h-s9zx/data>

This is a JSON.

**Transform**

For the NHE data, since I am only looking for data in the US, I filtered the data to only the United States region. I took out the “Code”, “Group”, “Region Number”, “State Name”, and “Average Annual Percent Growth” columns because they were not necessary columns after filtering the data to just the United States. I want numbers in the United States as a whole, so other numbers are not relevant. I also reset the index because it was based off of the original dataset. Finally, I renamed the columns, so it will be easier to query.

For the Ebola data, I filtered by US data only in the “Country” column because I only need data on the US. Then, I located only the total data in the rows the dates when they were recorded. I also reset the index because it was based off of the original dataset.

For the HIV/AIDS data, it was not a table that was too detailed, so not much was needed to be done. I removed the “Category” and “Group” columns because I only want the counts of the cases, rather than a whole breakdown.

For the COVID-19 cases, there was no API key needed, so I was able to access it easily. I query to look for the countries, total confirmed cases, total deaths and total recovered. I chose total deaths and total recovered to see if having higher funds will contribute in the survival rate of infected people. I created a list, looped through the API, and appended the data needed.

For the COVID-19 funds, the stored data was a little bit messier. They were stored in lists and displayed through the website on a table, that had each class encrypted. Luckily, they had an option for JSON, so I utilized this. When I appended the data from the JSON file, most of the data were ready to use, except for the actual data that mattered, the funds. They had unnecessary spacing, a $ sign and commas, so they had to be removed before I convert it from a string to a float.

**Load**

I chose the data table from the National Health Expenditure because they had data on past funds in the health department. I wanted to use that to link to past outbreaks for possible correlations. I chose Ebola because it was an outbreak, similar to COVID-19 because it was well-contained. I chose HIV/AIDS because it was an outbreak, that was rather huge, but is still present to this today, which may happen with COVID-19. Finally, I chose COVID-19 cases and funds so that it’ll be compared to past funds.