Emmanuel Fabre

Aishah Newson

6 March 2019

ETL Project Report

For the ETL Project, we decided to focus on data that looked at various suicide rates. One of our datasets was a CSV file on Kaggle. It looked at suicide rates in countries around the world and broke them into categories that focused on age, gender, year, population size, number of suicides and even GDP. The second dataset we utilized came from the CDC website. It had a narrower focus with a break down by US states, but also included gender, age, total count of suicides and various years.

The process for cleaning and transforming the data to depict what we wanted was different for each. On the CDC website, it requires a session ID unique to each query result. This made it impossible for BeautifulSoup to scrape the data. Due to this, we saved a copy of the query result as a local HTML file and then used BeautifulSoup to scrape that instead. However, during this process, we encountered an encoding error when parsing it. So once again, we decided to save the same dataset as a plain HTML file. From there, we were able to use the Pandas read\_html function and converted the HTML table to a list of DataFrame objects. We then printed a list of the DataFrame columns which returned very messy column headers and columns filled with NaNs. In order to clean this up, we renamed the column headers and dropped any extraneous columns. As for the Kaggle dataset, it was slightly more straight forward. As mentioned previously, there was more information in the table than what we wanted to work with. So, we utilized Pandas library to call specific functions in order to drop specific columns, filter for a certain time frame and adjusted for the years to be written in the same format so that we could join on that column.

Lastly, we needed to figure how we were going to connect the datasets and what database we were going to store the information in. Ultimately, we decided to use MySQL for this portion of the project. We did have a somewhat hard time trying to figure out what would be the best way to put all of our information into this database. At one point in time we thought we might join on age which is when we realized our datasets were formatted slightly different, we considered appending the datasets so that we didn’t lose any of the information we had formatted, and we also thought about joining them on the year. In the end, we decided to join them on year because that seemed to be the option that would give the end-user the best overall picture When we created our database, one of the interesting things that we noticed is that the data differed in its reporting of information. Specifically, the total number of deaths were not equal in the datasets. This could have potentially been because each dataset defined what was a suicide differently.

While this project might have caused us a few minor hiccups along the way, our final database provides anyone who wants to study suicide rates a clean picture of data, pulled from multiple sources and has a broad look at the topic. We set up our end-user to be able to easily understand what is in front of them and to be able to manipulate however they may see best fit for their needs.