Project Proposal

Samantha Rodriguez

2/28/2022

There are different trends, health wise, depending on where a person lives. This could be demographic or location wise. To look further into what exact trend is occurring I will be looking at nine different types of sicknesses.

- Asthma
- Carbon Monoxide Poising
- COPD
- Drinking Water Quality
- Food Safety
- Heart Disease
- Heat Stress Illness
- Infectious Diseases
- Mortality

The data will be collected from Arizona Environmental Public Health Tracking (EPHT) Explorer. This website is a Graphical User Interface (GUI) with filters to look at the different statistics of each sickness. The data is collected from different hospitals in Arizona. I will be specifically looking at county based information. To look more at the demographic information in the counties I then moved to collected information from the Census Database. I will be collecting county based data for the American Community Service (ACS), which has yearly data rather than every ten years. This will be able to match up better with the EPHT data. The ACS only has information with the top ten populated counties.

- Apache
- Cochise
- Coconino
- Maricopa
- Mohave
- Navajo
- Pima
- Pinal
- Yavapi
- Yuma

This will exclude five counties in Arizona, but I feel that the EPHT data is more accurate for the top ten counties.

My goals for this project is to use new resources, to myself, to collect the data from the two separate sources listed above. For the EPHT data, I will be web scrapping the data through Python. Once the data is collected it will transferred to R studio for cleaning and organization. For the ACS data, I will be collecting it through File Transfer Protocol (FTP) calls made in the Anaconda terminal. Once the data is collected it will be moved to pgAdmin, a PostgreSQL application. There will be separate tables in one database

holding the data. It will be cleaned and organized in the application as well. Then, from R studio it will be connected to the database in order to interact with the EPHT data. For both sources, the data will be collected from 2005 to 2019.