Capstone Project 1: Data Wrangling

For the data wrangling portion, I divided my work into seven steps. For the first six steps, I am reading in and cleaning six different datasets from 2016 county level data found from the web. In the final step, I combine the six datasets together:

1. prescription_rates: opioid prescription rates
2. education
3. population
4. poverty
5. unemployment
6. death rates from drug overdose
7. merge data together on county code (FIPS code)

1. Prescription rates:

   I obtained the US County Opioid Prescribing Rates 2016 data from the following website:

   - https://www.cdc.gov/drugoverdose/maps/rxcounty2016.html

   This website has a table containing the relevant data. I scraped the table using BeautifulSoup, saved the scraped data into a txt file, and then read the txt file into a pandas dataframe.

   Missing values:

   > There were 181 missing values for prescription rates out of 3143 counties, which is 5.6% of the data. Since this is a small percentage, I decided to remove the missing rows.

2. Education:

   I obtained the US County level 2016 education data from the following website:

   - https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/

   The website provided an Excel file for the data. I downloaded the file using curl command from Jupyter Notebook. Then, I read the Excel file into python, removed comments from the file, and selected the relevant 2016 columns.

   Missing values:

   > There were 10 rows with missing values out of 3283 total counties included in the dataset. Since 10 out of 3283 is only 0.3% of the data, I decided to remove the rows with missing data.

3. Population:

   I obtained the US County level 2016 population data from the following website:

   - https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/

   The website provided an Excel file for the data. I downloaded the file using curl command from Jupyter Notebook. Then, I read the Excel file into python, removed comments from the file, and selected the relevant 2016 columns.

   Missing Values:

   > There were 10 rows with missing values out of 3283 total counties included in the dataset. Since 10 out of 3283 is only 0.3% of the data, I decided to remove the rows with missing data.

4. Poverty Estimates:

   I obtained the US County level 2016 population data from the following website:

   - https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/

   The website provided an Excel file for the data. I downloaded the file using curl command from Jupyter Notebook. Then, I read the Excel file into python, removed comments from the file, and selected the relevant 2016 columns. I also removed the columns that contained information for only 52 counties.

   Missing Values:

   > There was one row with missing values out of 3283 total counties included in the dataset. Since one out of 3283 is only 0.03% of the data, I decided to remove the rows with missing data.

5. Unemployment Rate:

   I obtained the US County level 2016 population data from the following website:

   - https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/

   The website provided an Excel file for the data. I downloaded the file using curl command from Jupyter Notebook. Then, I read the Excel file into python, removed comments from the file, and selected the relevant 2016 columns. I also removed the columns that contained information for only 52 counties.

   Missing Values:

There were 82 rows with missing values out of 3275 total counties included in the dataset. Since 82 out of 3275 is only 2.5% of the data, I decided to remove the rows with missing data.

6. Drug Overdose Death Rate in 2016:

   I obtained the data from the following website:

   https://wonder.cdc.gov/controller/saved/D77/D39F042

   The link above contains a saved query from the CDC website that will generate a txt file with the relevant data. I've also saved a copy of the txt file generated from this saved query on github for anyone to access:

   https://raw.githubusercontent.com/danielhanbitlee/Springboard/master/capstone_project/drug_overdose_death_opioid_2016.txt

   I downloaded this file using curl from Jupyter Notebook. Then, edited the file to remove the comments and only select the relevant data. I read the file into a pandas dataframe. When reading the dataset into a dataframe, I forced all the rows with "Suppressed", "Missing", or "Unreliable" values to na. "Suppressed" means that the county had 0 - 9 deaths per 100,000 people due to opioid and narcotics overdose in 2016. Then, I removed the double quotes from certain columns and converted certain columns to numeric data type. From the 3147 county information provided from the data, only 423 rows did not have missing crude death rate for the county. However, since I am trying to predict the crude death rate on the county level, I decided to remove all the rows with missing data and only keep 423 rows.

7. Merge the Six Datasets Together:

   Using pandas, I merged the six datasets together on FIPS county code. In the process, I dropped redundant columns. The final dataset contains 423 rows and 57 columns.