Parisa Farahani

Project stage I

Approximately 30 years ago, physicians and patients were informed that taking opioid pain relievers may not cause addition. As a result, opioid dispensing rate started to gradually increase mainly from 2006. Opioid prescription rates reached its peak in 2012. Such an increase resulted widespread use and misuse of opioid pain relievers and opioid related mortality rates went up in certain areas across the US. Therefore, this started to become an alarming health concern across the US.

Data since can identify how different elements or factors are contributing to the opioid related mortality rate. This way, socioeconomic factors that contribute most can be identified to reduce death rate in US counties with most alarming opioid related mortality rates.

- Outline how the datasets can be merged together and the common variables.

The three main datasets provide some information on US opioid dispensing rate, cause of death, and health related factors at county level. In other words, these datasets provide county specific information. We should be able to merge them based on US county names or code as common variable.

"County Code" column of first dataset (Underlying Cause of Death-County-2019.txt), **"5-digit FIPS Code"** column of second dataset (County_Health_Ranking.csv), and **"FIPS"** column of the third dataset (2019-Opioid_Rate.csv) can be used as common variable for merging three datasets. These three column provide county specific code or county FIPS count. We should convert them to integer (it is in string format in first dataset) and rename these fields properly (e.g., name them "County FIPS") to be able to merge the three datasets.

- **M2.1** Study prior research in the area. (20 pts)

I first studied the provided paper [1] titled "Correlation of Opioid Mortality with Prescriptions and Social Determinants: A Cross-sectional Study of Medicare Enrollees". In this paper, the impact of opioid dispensing rates and socioeconomic factors on opioid-related mortality rates are evaluated. The link between dispensing rates and mortality rates is evaluated for various group of practitioners (e.g., emergency medicine, family medicine, internal medicine, and physician assistants). Among different groups of physicians, emergency medicine practitioners dispensing rates has highest correlation with mortality rate. I assume that people who went to emergency room were in acute pain and there has been no way but to take available pain killers to alleviate the pain. This correlation however, is minimal for other categories of practitioners specifically for family medicine practitioners. Family medicine is usually the first point of contact in most cases including routine visits but not the emergency visit. This may explain the lower opioid prescription administrated by a family medicine compared to other categories of physicians.

Later the authors have identified important demographic factors that correlates well with opioid dispensing rates. These factors do not necessarily impact opioid related mortality rates. Proportion of white, black, and male population are among those factors that impact dispensing (prescription) rates. Proportion of black population is negatively correlated with prescription rates. It is expected for minority to usually live in slightly more deprived areas. Such areas as will be discussed later has higher rate of opioid prescription and higher risk of overdose related death rate. These facts together may not really be compatible. I am assuming that these facts are presented based on physicians office locations not patients neighborhood.

According to Grigoras et al. (2018) [1], the first demographic factor (percentage of white population) as well as poverty rate are important parameters that positively correlate with opioid mortality rates.

Other researchers who cited provided paper [1] or are somehow related to this work have also identified poverty related indices to have an impact on opioid related mortality rate. Kurani et al. (2019) [2] evaluated the rates of opioid prescription and opioid related mortality rate between 2012 and 2017. They found the two rate to change in opposite direction. In other words, it is an unfortunate fact that mortality rate has increased even with reduction in opioid dispensing rates. Kurani et al. (2019) found risk of administrating and poisoning of opioid related pain relievers to be higher by %72 and %36 in most deprived areas compared to the least deprived areas. The parameter used in Kurani et al. (2019) was Area Deprivation Index (ADI) at county level.

Stoicea et al. (2019) reviewed a list of studies that researched the relation between demographics factors (ethnicity, culture, gender, religion) and opioid accessibility, abuse and overdose. According to this study, deprived communities are at higher risk of drug abuse or misuse. New strategies to manage opioid crisis should be in place in such communities to respond to opioid crisis in near future.

Madras (2017) sees the opioid epidemic crisis a supply issue. Some patients particularly those suffering from cancer related side effects start to use illegal form of opioid or low cost heroin after reduction of prescription opioid supplies. He considers physicians responsibility as a key to reduce opioid overdoses death rate. Training current physicians, adding more physicians to under staffed hot spots, and screening prescription by substance abuse specialist may be a good starting point to combat current crisis.

References:

[1] Grigoras, C.A., Karanika, S., Velmahos, E. et al. Correlation of Opioid Mortality with Prescriptions and Social Determinants: A Cross-sectional Study of Medicare Enrollees. Drugs 78, 111–121 (2018). https://doi.org/10.1007/s40265-017-0846-6

[2] Kurani S, McCoy RG, Inselman J, et al. Place, poverty and prescriptions: a cross-sectional study using Area Deprivation Index to assess opioid use and drug-poisoning mortality in the USA from 2012 to 2017 BMJ Open 2020;10:e035376. doi: 10.1136/bmjopen-2019-035376

[3] Stoicea N, Costa A, Periel L, Uribe A, Weaver T, Bergese SD. Current perspectives on the opioid crisis in the US healthcare system: A comprehensive literature review. Medicine (Baltimore). 2019 May;98(20):e15425. doi: 10.1097/MD.0000000000015425. PMID: 31096439; PMCID: PMC6531094.

[4] Madras BK. The Surge of Opioid Use, Addiction, and Overdoses: Responsibility and Response of the US Health Care System. *JAMA Psychiatry.* 2017;74(5):441–442. doi:10.1001/jamapsychiatry.2017.0163

M2.2 Each student member of the team selects 10 variables they think that are important from the available dataset. (20 pts)

- **Deliverable**
- Prepare a data dictionary (data and datatype - variable dictionary. https://urldefense.com/v3/__https://analystanswers.com/what-is-a-data-dictionary-a-simple-thorough-overview/__;!!JTSHVUr6R1OOzg!PuQ1_c5o64T89fDQ3UEtBmkyfn0EZO89p710KzEbj55Z6JjwguQN1ghxQwnr6j7RogbGWh7t3LzP4psP0xOdgWN5tOU$) of the selected variables

Name	Definition	Correlation	Type	Possible Values	
				Min	Max
Reading proficiency	Ability to read text	0.30	Float	0.11	0.73
Life expectancy	Number of years one expect to live	0.30	Float	66.88	94.66
Older adults living alone	Older adults not married living on their own	0.32	Float	0.23	0.34
Premature death	Death that occurs before the average age of death	0.34	Float	2610.69	22123.7
Injury hospitalizations	People admitted to hospital due to serious injury from accident or falling down or etc.	0.34	Float	174.30	857.50

Cancer incidence	new cancers occurring in the population per year,	0.47	Float	330.30	584.3
Hate crimes	Violent crime motivated by prejudice on the basis of race, religion, sexual orientation, or other grounds.	0.51	Float	0.38	7.078
Total male population	Male population in county	0.53	Integer	2176.0	461457.0
Communicable disease	A disease that is spread from one person to another through a variety of ways	0.54	Float	425.31	2053.39
Drug arrests	All arrest related to drug abuse or drug dealing	0.56	Float	4.0	4489.0

- Include justification of why you think the variables are important.

We want the evaluate the impact of socioeconomic factors on opioid related death rate and opioid dispensing rate. In super_df, Norm_Deaths and Opiod_Dispensing_Rate represent opioid related death rate and opioid dispensing rate, respectively. Therefore, from super_df, I will first select **Norm_Deaths** and **Opiod_Dispensing_Rate** as the two dependent variables. I will select 10 more variables to see how these variables interact with the two dependent variables.

The first few columns of the super_df dataset describe county information. Therefore, I'll skip those columns for now because we don't want to perform evaluation at regional level (state level) for at least now. Accordingly, I focus on column "N" to "TT" of the super-df. In these columns, there are usually five variants for each variable or column. For example, **Low birthweight** is expressed as Low birthweight **raw value**, Low birthweight **numerator**, Low birthweight **denominator**, Low birthweight **CI low**, and Low birthweight **CI high**. The raw value is the ratio of numerator to denominator. The CI low and CI high are just the lower band and higher band of the raw value. Therefore, it may sound reasonable to only select columns that express raw values in our analysis.

I selected columns that contain "raw" in their header first. This resulted in selecting 105 columns. Next, I evaluated the correlation between Norm_Deaths column and these 105 columns. I sorted these columns based on the absolute value of the correlation coefficients. Then, I looked at those with correlation coefficients above 0.30. This criteria filtered out some other columns. Among remaining columns, some appear to represent similar variable. When there was a group of columns that represented similar variable, I selected the one that is more generic. For example, male or female population in various age group were important factors but I only considered total male population raw value. As another example, death due to injury, motor vehicle crash occupancy, or Fall fatalities 65+ are all related to injury hospitalizations.

Below is the code to pre select a long list of potentially import socioeconomic factors:

```
# import required libraries
import pandas as pd
import numpy as np
# read super_df
df = pd.read_csv('super_df.csv')
# select columns that contain "raw" in their header in addition to
"Norm_Deaths" as the most important variable
select = [item for item in df.columns if 'raw' in item] +
['Norm_Deaths']
# calculate correlation coefficients between all selected columns and
select those calculated for "Norm_Deaths" column
ndf = np.abs(df[select].corr()['Norm_Deaths'])
# drop nan values from correlation matrix
ndf = ndf.dropna()
# sort correlation coefficients and save it as a csv file
ndf.sort_values().to_csv('important_socioeconomic_factors_preselected.c
sv')
```