Group 14 Project Prototype

Part 1: Introduction and Research Questions

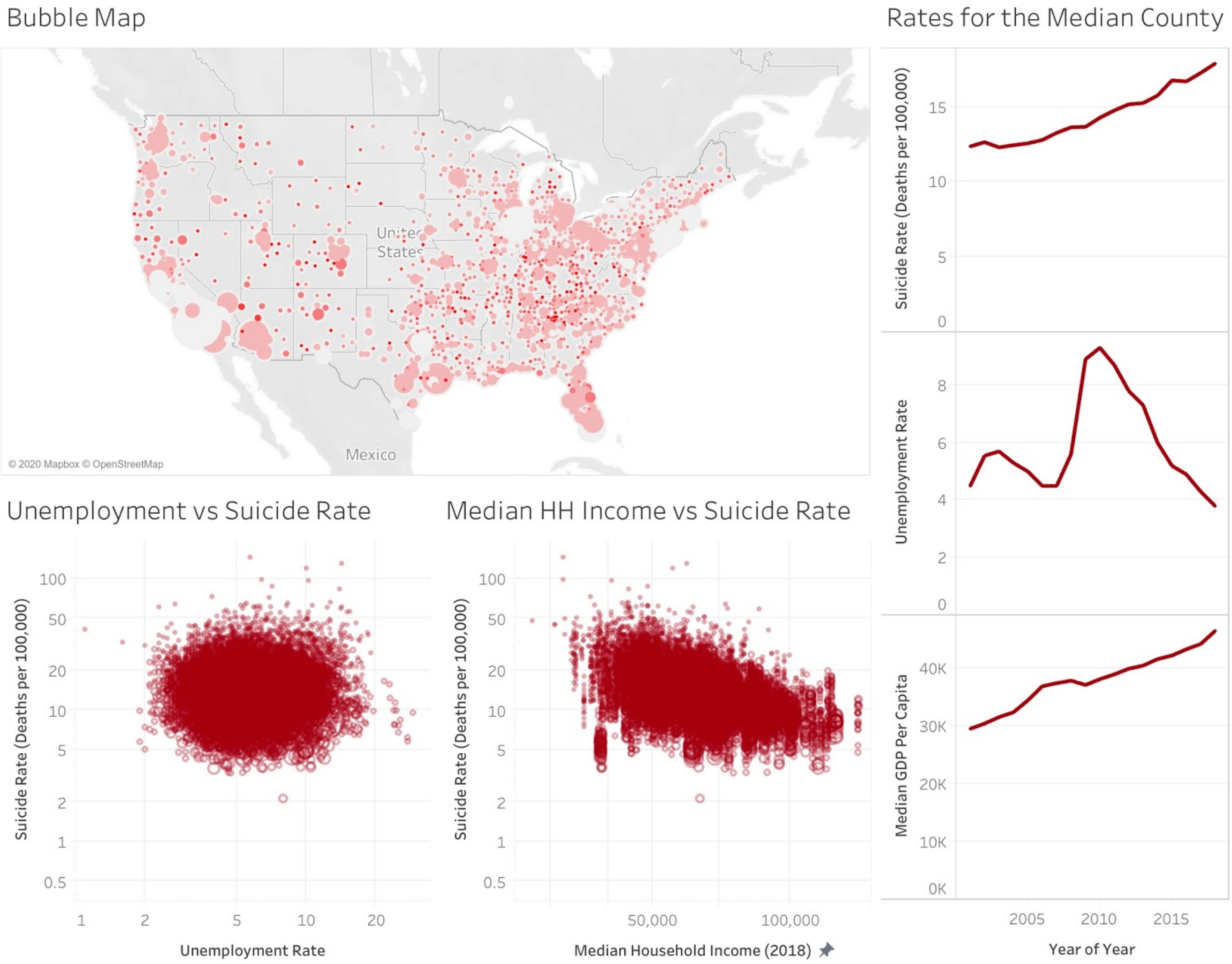
In recent years, mental health awareness has become a prominent focal point in the mainstream media and public psyche. Celebrities, media outlets, and large corporations have pledged to large financial contributions, as well as increasing attention, in an effort to help combat mental illness. In its recovery from the great recession in 2008, the US as a whole has enjoyed a period of economic prosperity by many common metrics such as GDP, unemployment rates, and interest rates that have many analysts and media outlets believing that a recession is in our near future. Following the universal outbreak and spread of COVID-19, the world is now experiencing its greatest pandemic since 1918; moreover, March 9th, 12, and 16th experienced record point drops in the Dow Jones Industrial Average, resulting in massive financial loss across the country. To date, unemployment has become a reality for over 6 million Americans. In this troubling time, mental health can become a serious concern.

Our hypothesis is that a region's mental health is dramatically impacted by its financial status; thus, we are looking to explore the following research question in more depth: What impact does a region's economic prosperity have on the population's mental health? Given that mental health is quite challenging to quantify, especially on a large scale, we will use suicide data as a proxy for mental health. To help define a region's economic status, we will use datasets including employment/unemployment rates, average income, and GDP. Both the mental health and economic proxies have thorough, public micro-datasets available through the United State government. Nonetheless, many of these datasets will require wrangling in order to become manageable. In the end, we hope to use the economic proxies to quantify economic wellbeing and determine how significantly it correlates with mental health.

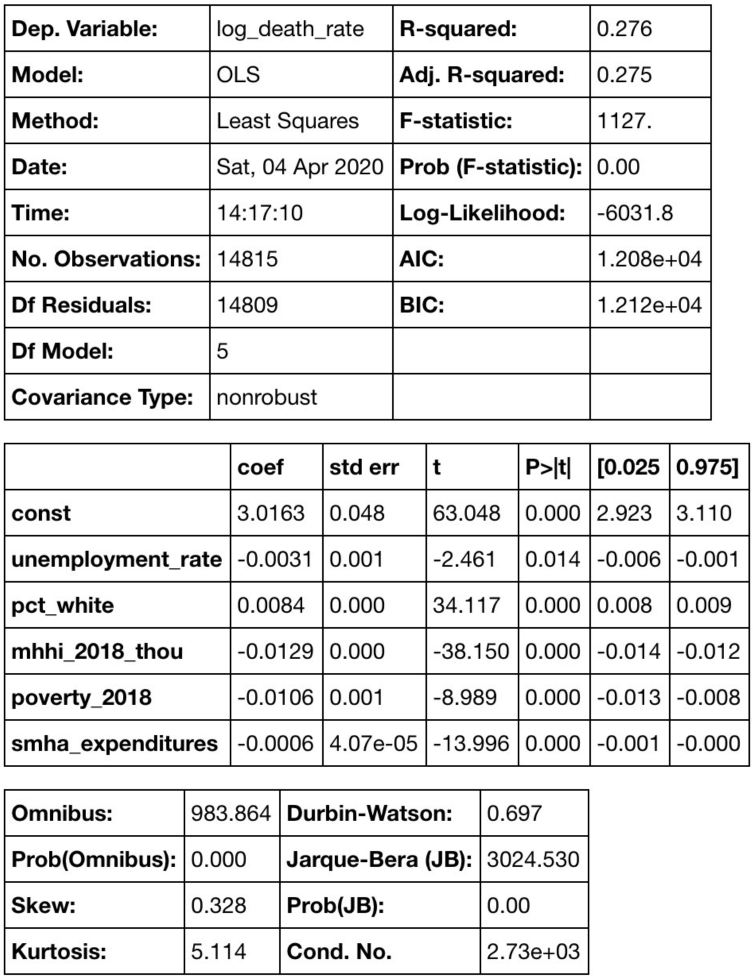
Part 2: Preliminary Results and Methods

To begin this project, we first started by creating a GitHub repository where all our team members could share all project materials with one another in a collaborative environment. We then downloaded several different datasets from various government agencies, which we then cleaned in the Jupyter notebook platform. After properly cleaning our data, we then moved to merging the key information from our separate datasets into one master dataset that we would be able to utilize for data analysis. Our final dataset included county-level data on suicide rates, GDP per capita, median household income, unemployment and poverty rates, racial demographics, and state-level data on per capita mental health agency spending.

We then used the Tableau software to explore the patterns in our datasets and to graph key findings from our observations. This led us to find and report potentially significant findings from our data that we plan to explore in greater detail in our next steps in our projects. We displayed our initial findings in a dashboard in Tableau shown below.



From this data, we noticed a few important trends. The first is that unemployment seemingly has low correlation with suicide rates in America. However, the median household income of counties clearly seemed to have a negative correlation with suicide rates. Said candidly, the relationship that we seemed to have found that when counties had a higher median household income, they tended to have lower suicide rates than counties that have a lower median household income. We thought this relationship was greatly intriguing to explore our question at hand so we thought to explore it more with GDP per Capita data and look at the relationship between that economic metric and suicide rates. Looking at these statistics (Top right of the dashboard), we noticed an almost growth rate in both of the metrics in the same time frame. We found this be rather intriguing especially in light of finding of relationship between the median household income and suicide rates. Because of this, we decided to create a linear regression model in the Jupyter notebook to see what variables and how significant the data that we were using was to suicide rates. We chose to take a log-transformation of suicide rates due to a significant right skew in the data. The output of the multiple linear regression on log suicide rates is shown below.



Dep. Variable:	log_death_rate	**R-squared:**	0.276
Model:	OLS	**Adj. R-squared:**	0.275
Method:	Least Squares	**F-statistic:**	1127.
Date:	Sat, 04 Apr 2020	**Prob (F-statistic):**	0.00
Time:	14:17:10	**Log-Likelihood:**	-6031.8
No. Observations:	14815	**AIC:**	1.208e+04
Df Residuals:	14809	**BIC:**	1.212e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	3.0163	0.048	63.048	0.000	2.923	3.110
unemployment_rate	-0.0031	0.001	-2.461	0.014	-0.006	-0.001
pct_white	0.0084	0.000	34.117	0.000	0.008	0.009
mhhi_2018_thou	-0.0129	0.000	-38.150	0.000	-0.014	-0.012
poverty_2018	-0.0106	0.001	-8.989	0.000	-0.013	-0.008
smha_expenditures	-0.0006	4.07e-05	-13.996	0.000	-0.001	-0.000

Omnibus:	983.864	**Durbin-Watson:**	0.697
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	3024.530
Skew:	0.328	**Prob(JB):**	0.00
Kurtosis:	5.114	**Cond. No.**	2.73e+03

From this model, we were able to obtain an adjusted R^2 value of 0.275 which means that the five variables that we used in our model (unemployment rate, rate of white Americans vs non-white American, median household income, poverty rates, and state mental health agency spending) explains about 30 percent of the variation in suicide rate. Furthemore, each of the variables have a statistically significant relationship with suicide rate. While we believe that this is a rather significant finding, we plan to explore other models, possibly including interaction effects, or utilizing hierarchical modeling, with year as a random effect.

Part 3: Reflection and Next Steps

Although we believe that our initial data cleaning and analysis has allowed us to investigate patterns and further understand how economic performance impacts mental health, there are 3 core ways that we can improve our project.

First, upon doing initial exploratory data analysis, we found that the CDC, our source of suicide rate data, censors data from counties with a small number of suicides in a given year in order to protect the privacy of those individuals. This could bias our data in several ways. As an example, when considering the relationship between GDP and suicide rates, we found a very strong negative relationship between GDP and suicide rates; however, this relationship is apparent partially becasue smaller counties with low suicide rates are inherently censored from the data. When looking at GDP per capita versus suicide rates, the trend is much more muted as a result. If we had not noticed this practice by the CDC, this would have drastically altered our findings and could have potentially invalidated our project. Therefore, with this experience in mind, we want to dive deeper into our data sources and make sure that they are not censoring data, as well as incorporate techniques which can reduce or eliminate the bias introduced by censored data.

Next, we want to try to find additional relevant data that may allow us to fill some of the gaps in our current model. Some examples of possible data sources include other pertinent demographic data such as educational attainment or age.

Lastly, we want to explore some more complicated models, including using transformations of our explanatory variable and interaction effects. We will also utilize cross-validation in order to test whether we are overfitting our models. These techniques will help us find a model that we believe more accurately addresses our core research problem.