Assignment: Homework 2 Report

Name: Horace Fung

Instructor: Prof. Simonoff

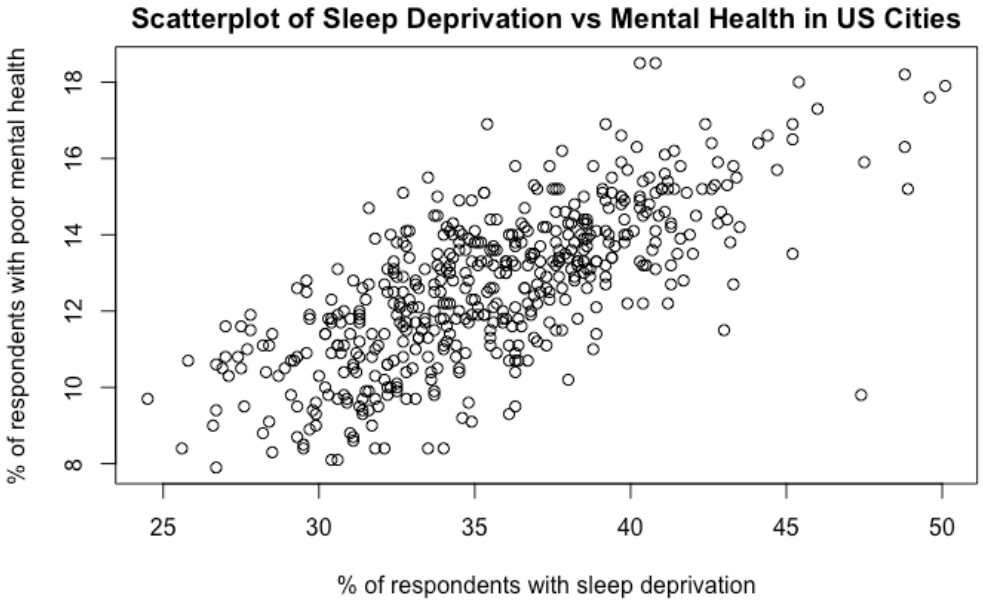
Course: Regression and Multivariate Data Analysis

The Relationship Between Lack of Sleep and Mental Health in American Cities

In the United States, mental health and wellbeing is a growing concern. The Mental Health America organization estimates that 1 in 5 American adults have a mental health condition (Mental Health America). Among young Americans, the rate of severe depression has grown from 5.9% to 8.2% over a 5-year period (Mental Health America). Since the variety of mental health conditions is large— and their severity and symptoms can vary dramatically—it is difficult to identify broad variables that potentially affect mental health. This report will simply examine the association between American cities that sleep less and their mental health. The idea is that sleep deprivation affects psychological state and can potentially expose people to greater "negative thinking and emotional vulnerability" (Harvard Health Publishing). Conversely, poor mental health may also lead to less sleep. A key clarification here, however, is that the observations are on a city level, not individual. Therefore, the analysis is not a direct examination of an individual's average sleep duration and their mental health. If an individual reports getting sufficient sleep but suffering poor mental health, the data does not match this to one observation because it only sees the city's aggregate response. Hence, this report is looking at the association between the general sleeping habits and the general mental health of cities. A simple regression can model this:

$$\begin{array}{c}\text{\% of respondents that}\\ \text{report poor mental health}\\ \text{in a city}\end{array} = \beta_0 + \beta_1 \times \begin{array}{c}\text{\% of respondents that}\\ \text{report sleep deprivation}\\ \text{in a city}\end{array} + \text{random error}$$

The data is from 500 Cities Project, a collaboration between Centers for Disease Control and Prevention (CDC) and The Robert Wood Johnson Foundation. It contains three measure categories— Unhealthy Behaviour, Health Outcomes and Prevention— for each city. This report examines the measure sleep deprivation under Unhealthy Behaviour and the measure poor mental health under Health Outcome. Sleep deprivation is defined as the percentage of respondents "sleeping less than 7 hours among adults aged ≥ 18 years on average" and poor mental health is defined as the percentage of respondents "aged ≥ 18 years who report 14 or more days during the past 30 days during which their mental health was not good". The results are self-reported. The sample in this report contains 500 observations and each one represents a city in the United States. Below is a scatterplot of the two variables:



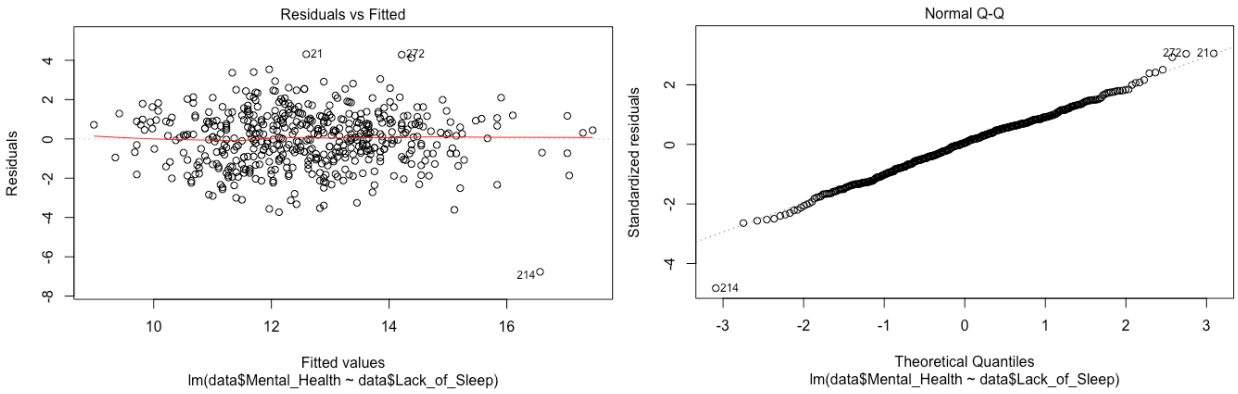
The relationship seems to be positive, which suggests cities with more sleep deprivation tend to have more people with poor mental health. There is a clear unusual observation at (47.4%, 9.8%). That observation is Honolulu. It reports a very high percentage of sleep deprivation but a

low percentage of poor mental health. Below is the summary for a least square regression that fits sleep deprivation as the predictor and poor mental health as the target:

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.8716     0.5191   1.679   0.0938 .
data$Lack_of_Sleep  0.3311     0.0145  22.827   <2e-16 ***
---

Residual standard error: 1.413 on 498 degrees of freedom
Multiple R-squared:  0.5113,      Adjusted R-squared:  0.5103
F-statistic: 521.1 on 1 and 498 DF,  p-value: < 2.2e-16
```

The model has an R^2 of 51.13%. For health policy data, that seems to be a decent fit. The F-statistic is very significant and the null hypothesis H_0: $\beta_1 = 0$ can be rejected, which suggests that the sleep deprivation variable improves the predictive power of the model. For a simple regression, the p-values and conclusion of the F-test and t-test are the same. The intercept coefficient states that for a city with no sleep deprived resident, the estimated expected percentage of people with poor mental health is 0.8716%. The coefficient, however, has a somewhat high p-value and does not seem significant. The slope coefficient states that a one percentage point change in the percentage of people sleep deprived is associated with a 0.3311 percentage point change in the percentage of people with poor mental health in a U.S. city. The slope coefficient is positive and very significant, but the value is small. 0.3311% translates to 682 individuals for the average sized city in this sample.

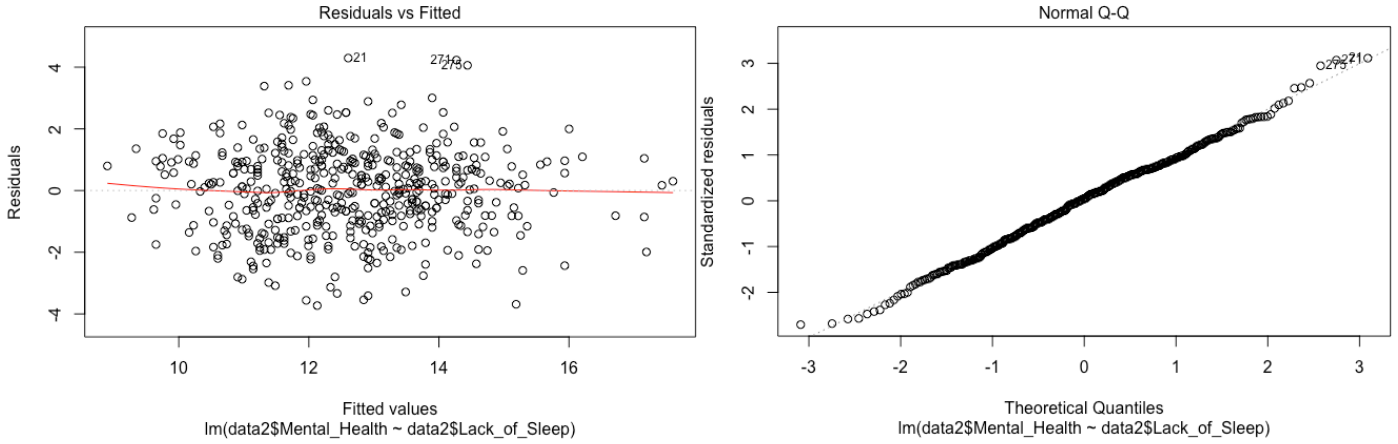
The initial scatterplot showed that some of the observations seem unusual, most notably Honolulu, and may violate the assumptions of a linear regression. The relevant diagnostic plots are presented below:



The residuals vs fitted values plot shows that the residuals are quite random and spread out about the zero line. There could be a cone shape between the fitted value = 10 and fitted value = 12 region, which would violate the constant variance assumption, but there are fewer observations on both ends of the fitted values axis which makes it difficult to interpret. The residual for observation 214, Honolulu, is abnormally negative which indicates an outlier— that the observed response value y is surprising given the expected value of y. On the normal probability plot, majority of the observations lie on the straight line. This occurs when the observed residuals are close to the expected residuals and indicates the residuals are normally distributed. Again, Honolulu is an outlier on this plot. Honolulu's unusually low value on the poor mental health response variable is possibly due to low levels of other unhealthy behaviours, which may be associated to mental health. For example, Honolulu reports much lower values for percentage of respondents that are smokers, physically inactive and obese compare to the other top ten most sleep deprived cities. Further multivariate analysis is required but that is beyond the scope of this first report. After removing Honolulu, the new regression output is:

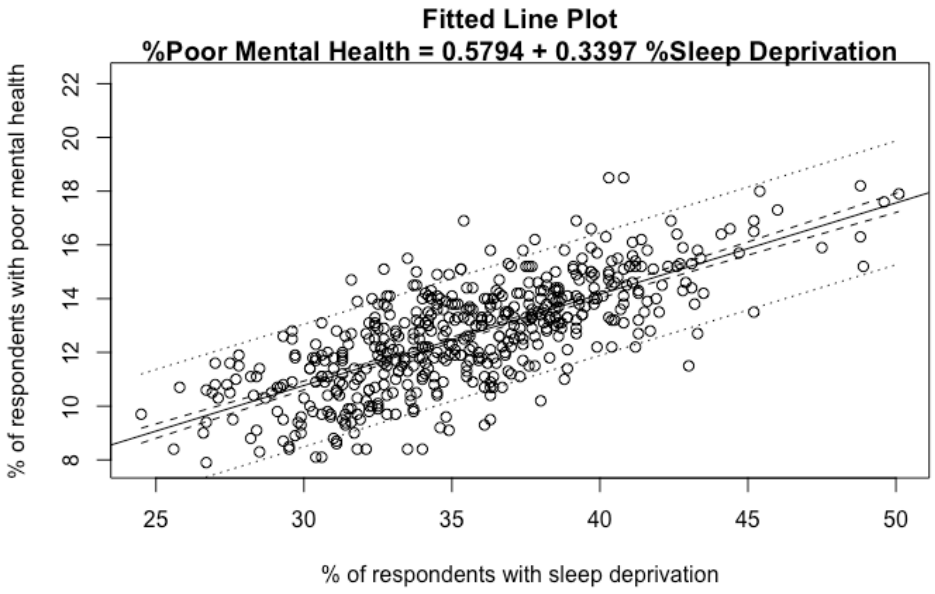
```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.57936    0.51072   1.134    0.257
data2$Lack_of_Sleep  0.33971    0.01428  23.787   <2e-16 ***
---

Residual standard error: 1.381 on 497 degrees of freedom
Multiple R-squared:  0.5324,      Adjusted R-squared:  0.5314
F-statistic: 565.8 on 1 and 497 DF,  p-value: < 2.2e-16
```

The outlier does not seem to be influential. R^2 increased by only 2.11%, the F-statistic is still highly significant, and the slope coefficient is similar. The only major change is that the intercept is now clearly not significant, as its p-value increased from 0.094 to 0.257. The new diagnostic plots:



With Honolulu removed, there is no clear unusual observations on both the residual vs fitted plot and the normal probability plot. It seems like the assumptions for a linear regression holds.

The confidence and prediction interval can provide information on the variability of the model's prediction and assess the usefulness of its predictions. Below is a confidence and prediction interval plot:



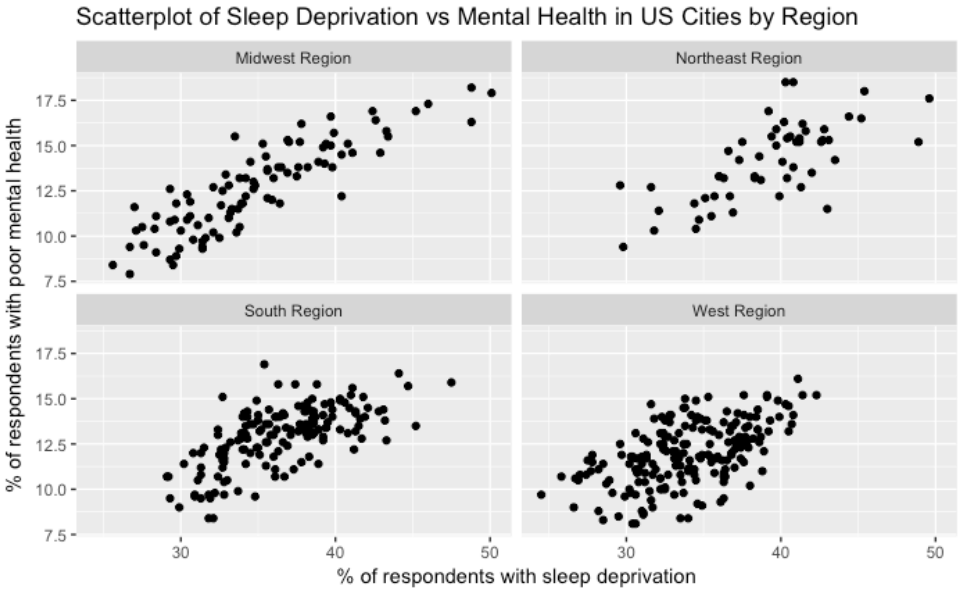
The prediction interval is much larger than the confidence interval, which implies that the inherent variability of the response variable, σ^2, contributed to a large part of the wide prediction interval. Taking New York City as an example, where 39.9% of respondents report sleep deprivation, the regression outputs the following confidence and prediction intervals at 95% confidence:

```
fit     PI(lwr)   PI(upr)
14.1339 11.4153 16.85249

        CI(lwr)   CI(upr)
        13.96066 14.30714
```

The confidence interval is (13.960, 14.307), which means that for all cities with 39.9% of people experiencing sleep deprivation, the model predicts that the average percentage of people experiencing poor mental health would lie within this interval. The prediction interval is (11.415,

16.852), which means for one city with 39.9% of people experiencing sleep deprivation, the prediction of the percentage of people experiencing poor mental health in that particular city lies within this interval. The prediction interval seems too wide and raises concerns over the precision of the model's predictions.

Lastly, the outlier Honolulu suggested that other factors may be associated with mental health in American cities. One way to examine this without delving into multiple regression is to separate the cities into segments that have similar characteristics like culture, income, diet and more. The U.S. Census Bureau divides the country into four regions; Northeast, Midwest, South and West. Historically, the regions were only based on the location of major drainage basins (U.S. Census Bureau). Later, the Census Bureau performed analysis to group the states into new homogenous groups based on factors like socio-economic similarity. They found four regions with similar positions as the current census regions, but many states near the border of one region switched to another (U.S. Census Bureau). Although the current census regions are imperfect, they may implicitly contain some information that is unique to each group and affects our relationship between sleep deprivation and mental health. Below is a scatterplot for each region:



In all four regions, there is a positive relationship. The Northeast has fewer observations, which makes interpretations less generalizable. The South and West regions have a greater spread in the response variable. That suggests sleep deprivation may not explain poor mental health well in these two regions. In the Midwest, the relationship seems much closer to a linear fit and the linear model may be more useful here. Below are the regression outputs:

Midwest

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          -1.29681    0.83059  -1.561    0.122
midwest$Lack_of_Sleep  0.40000    0.02355  16.982   <2e-16 ***

Residual standard error: 1.2 on 91 degrees of freedom
Multiple R-squared:  0.7601,      Adjusted R-squared:  0.7575
F-statistic: 288.4 on 1 and 91 DF,  p-value: < 2.2e-16

---------------------------------------------------------------------------
```

Northeast

```
Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              0.18443    1.99003   0.093    0.927
northeast$Lack_of_Sleep  0.35596    0.05054   7.044 4.22e-09 ***

Residual standard error: 1.553 on 52 degrees of freedom
Multiple R-squared:  0.4883,      Adjusted R-squared:  0.4784
F-statistic: 49.61 on 1 and 52 DF,  p-value: 4.216e-09
```

South

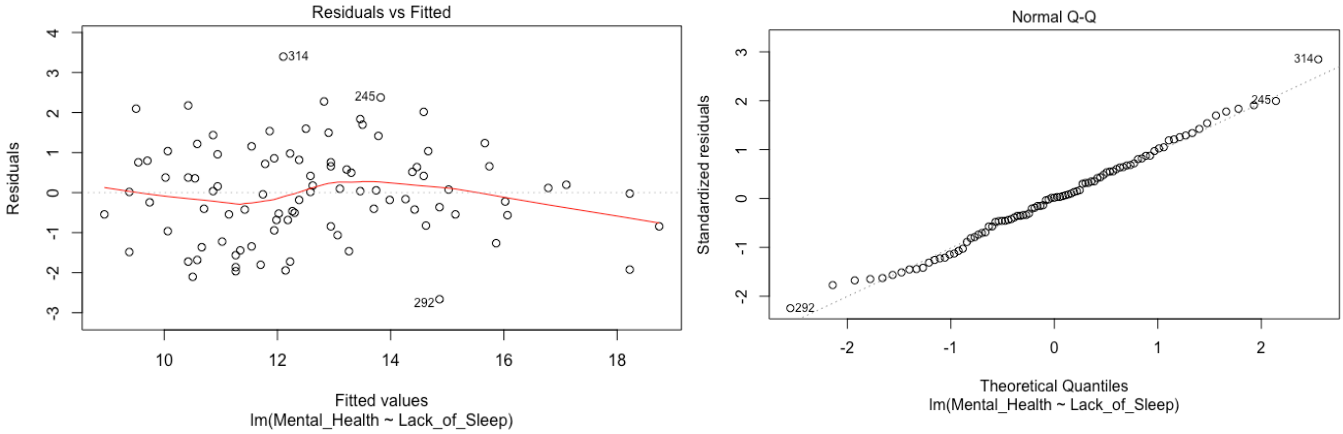
```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.93337    1.02810   1.881   0.0619 .
south$Lack_of_Sleep  0.30167    0.02803  10.762   <2e-16 ***

Residual standard error: 1.27 on 156 degrees of freedom
Multiple R-squared:  0.4261,      Adjusted R-squared:  0.4224
F-statistic: 115.8 on 1 and 156 DF,  p-value: < 2.2e-16
```

West

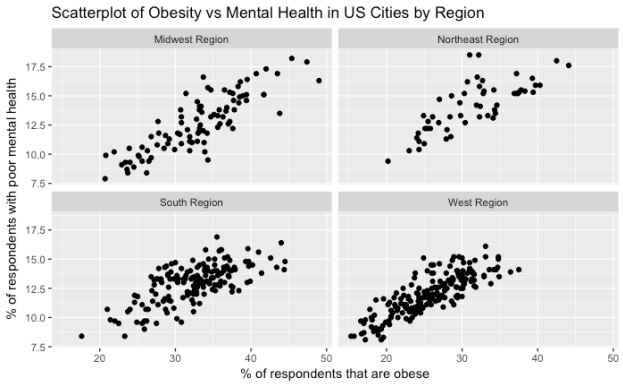
```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.80504    1.01357   2.767   0.0062 **
west$Lack_of_Sleep  0.26985    0.02968   9.091   <2e-16 ***

Residual standard error: 1.449 on 192 degrees of freedom
Multiple R-squared:  0.3009,      Adjusted R-squared:  0.2973
F-statistic: 82.64 on 1 and 192 DF,  p-value: < 2.2e-16
```

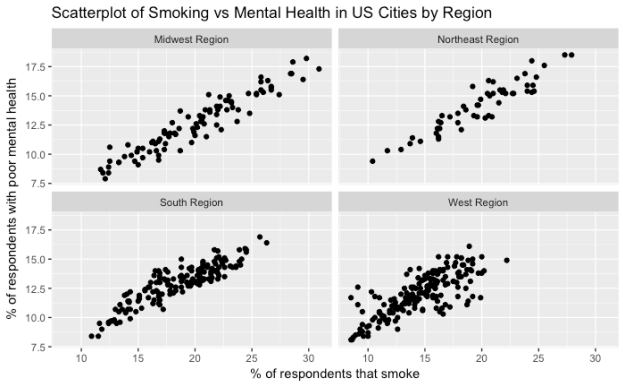
The F-test for all four regions are still highly significant, with the Northeast relatively less significant. The R^2 for Northeast, South and West are all below the R^2 of the model fitted on the entire dataset. For the Midwest, however, the R^2 is much higher at 76%. Furthermore, the Midwest has the largest slope coefficient at 0.40%, which means that a one percentage point change on the sleep deprivation variable is associated with a greater percentage point change in poor mental health response variable than any other regions. This, however, does not necessarily

mean the best simple regression model is one restricted to the Midwest. Below are the diagnostic plots for the Midwest model:



The residuals vs fitted plot appears to be mostly random. The cloud of points seem to narrow a bit towards the larger fitted values, which suggests heteroscedasticity, but there are also fewer observations there. On the normal probability plot, the observations mostly lie on a straight line, which means the residuals are close to normally distributed. The 95% confidence and prediction for a Midwest city like Chicago, where the predictor is 35.9%, are CI (12.811, 13.315) and PI (10.666, 15.460). The prediction interval is still quite wide which suggests that the predictions are not precise. And despite higher R^2, there is no clear explanation for why the fit is better for Midwest cities. Below are three other predictor variables vs poor mental health, segmented by region:





Just looking at the scatter plots, both obesity and smoking seems to have a fairly strong linear relationship in the Midwest, much more than the other regions (except for smoking in the Northeast which also seems to have a good linear fit). Therefore, it is possible that geographical information is not useful at all. Rather, selecting the Midwest is simply selecting a subsample where other predictors are strongly correlated with sleep deprivation, and these other predictors are the true factors of mental health. Hence, the high R^2 could be deceiving. Further multivariate analysis is required to examine whether region is truly informative with regards to the strength of the relationship between sleep deprivation and mental health. There is not enough evidence to believe that building four simple regression models based on region will reflect the true relationship and generalize for new data; predictors like smoking is not bounded by a region's border and a change in smoking habits of each region can occur in new data. Hence, the best simple regression model at this moment is the full sample regression with Honolulu removed.

This report demonstrates that there is a small but statistically significant association between the prevalence of sleep deprivation and poor mental health in American cities. This knowledge could be used to support further investigations into sleep duration and mental health, as well as confounding factors like length of working hours, commute time etc. that could further

explain the relationship. Due to the small slope coefficient, however, the model implies that policies only targeting sleep deprivation may not reduce mental health issues by much. To better understand the factors associated with poor mental health in American cities, a multiple regression may be a better tool.

Data:

1) 500 Cities: Local Data for Better Health, 2018 Release | Chronic Disease and Health Promotion Data & Indicators." *Centers for Disease Control and Prevention, Centers for Disease Control and Prevention*, chronicdata.cdc.gov/500-Cities/500-Cities-Local-Data-for-Better-Health-2018-relea/6vp6-wxuq.

2) *United States Census Bureau*, www2.census.gov/programs-surveys/popest/geographies/2017/.

Citations:

1) "The State of Mental Health in America 2018." *Mental Health America*, 31 Oct. 2018, www.mentalhealthamerica.net/issues/state-mental-health-america-2018.

2) Harvard Health Publishing. "Sleep and Mental Health." *Harvard Health Blog*, Harvard Health Publishing, 19 June 2018, www.health.harvard.edu/newsletter_article/sleep-and-mental-health.

3) United States Census Bureau, "Geographic Areas Reference Manual Chapter 6", *United States Census Bureau*, https://www2.census.gov/geo/pdfs/reference/GARM/Ch6GARM.pdf