Project Proposal Revised

due October 20, 2021 by 11:59 PM

Isabelle Xiong and Maya Ghanem

10/17/2021

Load Data

```
X500_cities_diabetes <- read_excel("~/ Stats 198 Project/data/500_Cities_diabetes.xlsx")
X500_cities_health_insurance <- read_excel("~/ Stats 198 Project/data/500_cities_health_insurance.xlsx")
X500_cities_heart_diseases <- read_excel("~/ Stats 198 Project/data/500_cities_heart_diseases.xlsx")
X500_cities_mental_health <- read_excel("~/ Stats 198 Project/data/500_cities_mental_health.xlsx")
```

Introduction and Data, including Research Questions

America, unlike other developed countries, does not have a universal healthcare program, meaning that not all individuals are granted free health insurance by the government. In America, Health insurance is purchased in the private marketplace or provided by the government only to certain groups, such as pregnant, low income, elderly, families with children. Having lived in Canada, where all citizens can apply for a public health insurance for free which grants them access to quality healthcare, I've never had to think of what it would be like without access to healthcare. In doing this research, I wanted to understand the effects of a lack of health insurance in the US, by finding the correlation between not having health insurance and the likelihood of contracting different diseases in 500 largest US cities.

Research Question:

Do cities with higher percentages of uninsured people have poorer mental health outcomes, higher rates of diabetes, and higher rates of heart disease?

Glimpse

We randomly selected 50 cities from the original CDC dataset as our sample.

```
glimpse(final_CDC_data)
```

```
## Rows: 50
## Columns: 8
## $ CityName               <chr> "Mobile", "Birmingham", "Montgomery", "Alhambra~
## $ percent_mental_health  <dbl> 15.2, 20.7, 10.7, 11.0, 12.4, 15.6, 12.1, 11.7,~
## $ percent_lack_insurance <dbl> 16.5, 26.2, 10.3, 16.3, 11.1, 20.3, 12.0, 9.7, ~
## $ percent_diabetes       <dbl> 10.1, 20.5, 8.8, 12.0, 6.5, 7.3, 7.6, 10.8, 4.9~
## $ StateDesc              <chr> "Alabama", "Alabama", "Alabama", "California", ~
## $ GeoLocation            <chr> "(30.6718491643, -88.1996782045)", "(33.5532050~
## $ UniqueID               <chr> "0150000-01097003703", "0107000-01073001902", "~
## $ percent_heart_disease  <dbl> 6.2, 8.7, 4.5, 5.1, 3.0, 2.9, 2.7, 5.9, 2.5, 5.~
```

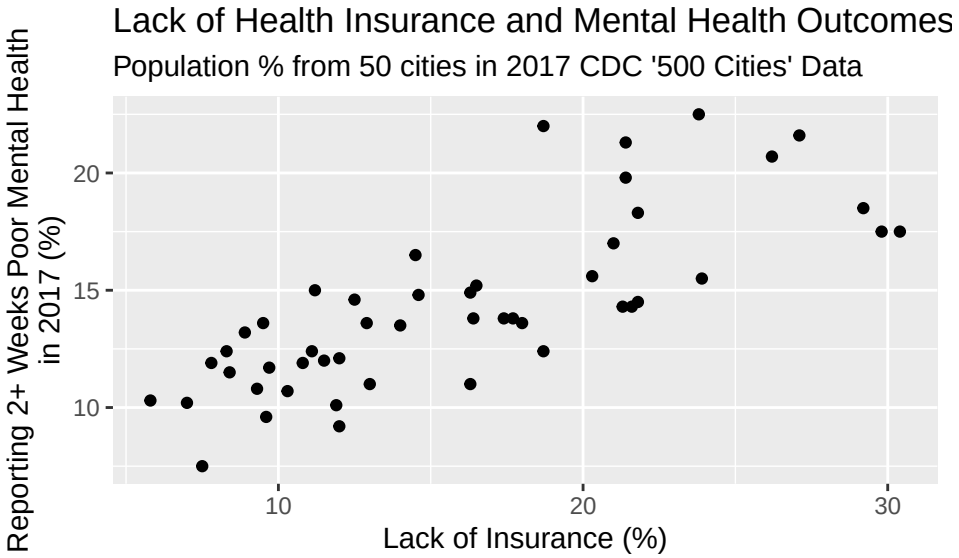
Data Analysis Plan

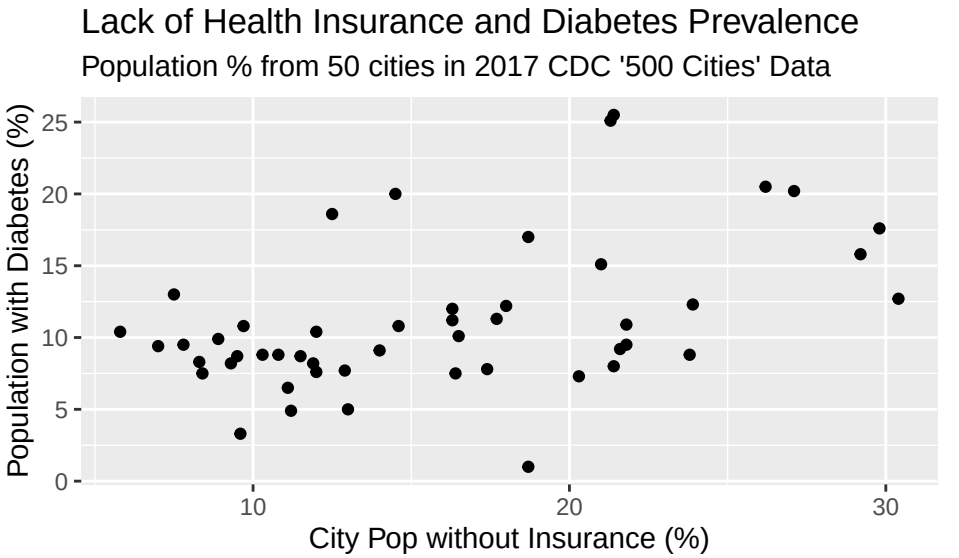
Explanatory and Response Variables

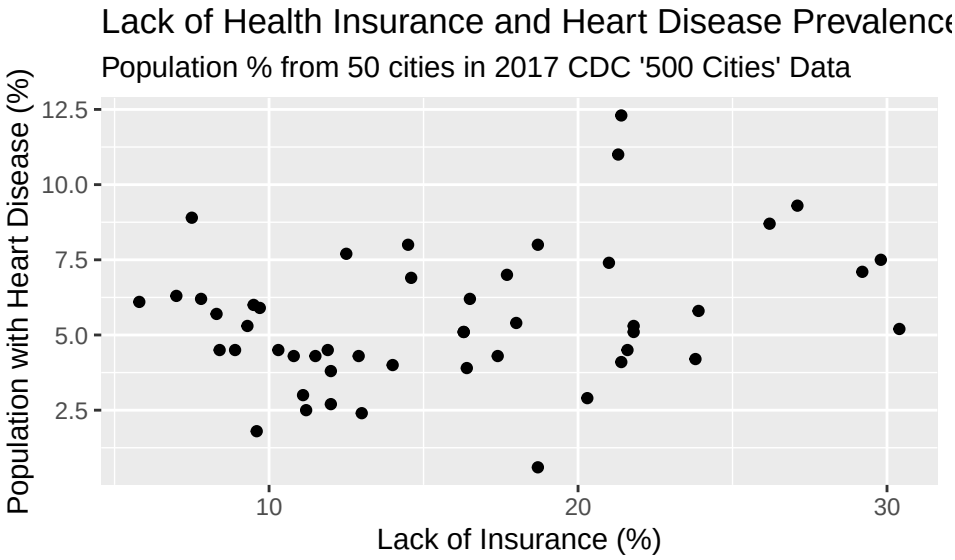
Explanatory Variable: percentage of uninsured people in each city

Response Variable: percentage of people in each city who reported poor mental health for more than two weeks in 2017, percentage of people in each city diagnosed with diabetes, percentage of people in each city with coronary heart disease

Initial Visualizations







Based on the visuals above, the strongest correlation seems to be between Lack of Health Insurance and Mental Health Outcomes. All three graphs depict positive correlations. There is large spread and variation in the data for all visuals.

Statistical Tests

We will first run three t-tests to determine if there is any significance in the relationship between the percentage of uninsured people and prevalence of poor mental health outcomes, diabetes, and heart disease (respectively). We will then run three regression models to further understand the association between the explanatory variable and each response variable. Through the regression model, we hope to quantify the relationship between the percentage of uninsured people and prevalence of poor mental health outcomes, diabetes, and heart disease. We will also determine the significance of the regression by understanding the t-statistics of the coefficients in each regression.

References

https://chronicdata.cdc.gov/500-Cities-Places/500-Cities-Local-Data-for-Better-Health-2019-relea/6vp6-wxuq