Group1：COVID-1

Qingyuan Guo

Xianru Liu

Di Qi

Dawei Wang

Initial Exploratory Analysis

The raw dataset we had was from a health and social science dataset. We made some basic cleanings from the original dataset to make it suitable for our goal. To get the dataset we use for this project, two datasets were merged, some variables were dropped and for our convenience of further analysis, some content of categorical variables substituted. Thus, we get a heart disease dataset that contains 23 variables and 3198 observations.

The dataset contains information directly describing the counties of the United States. It consists of several socioeconomic variables and two death rates that originally collected from multiple resources including the Bureau of Labor Statistics, National Center for Chronic Disease Prevention and Health Promotion, US Census Population Estimates, etc.

There are some missing values in some variables, such as:
“health_pct_adult_smoking”,“health_pct_excessive_drinking”,
“health__air_pollution_particulate_matter”. In total, we have 3198 samples and we deleted the records which contain missing values. We have 1069 samples in the end. There is a lot of missing data in those variables, at first, we tried to use the mean, but after comparing the mean and just deleting all of the NA, we think it is a good way to delete the NA. If we choose to replace them, it will influence a lot in our data analysis, so we just delete all of the NA, thus, we can keep the quality of our dataset.

There are 19 numerical variables histograms, and 3 categorical histograms. After using log transformation in variables:

econ__pct_unemployment

econ_pct_uninsured_children

demo_birth_rate_per_1k

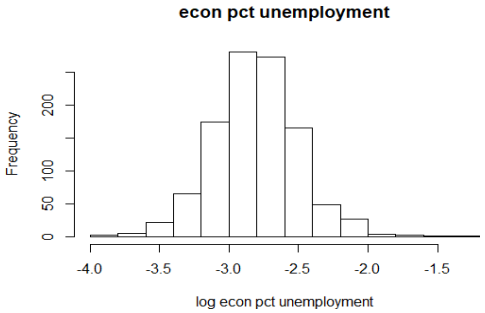
health_homicides_per_100k

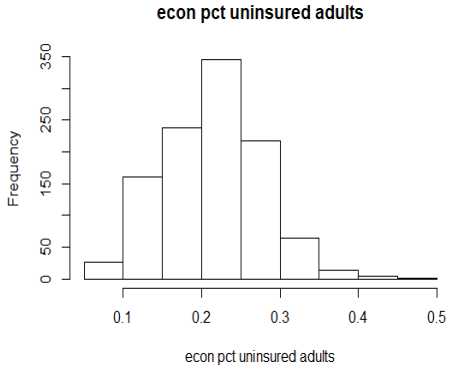
health_motor_vehicle_crash_deaths_per_100k

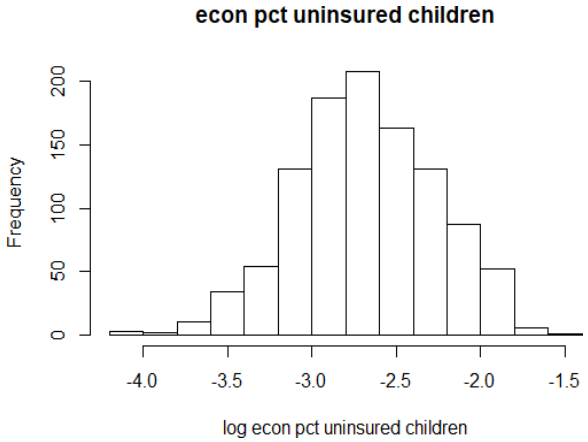
health_pop_per_dentist

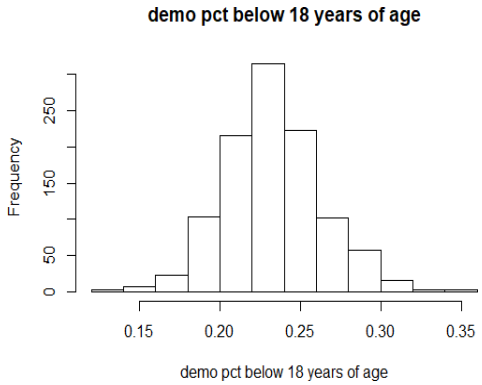
health_pop_per_primary_care_physician

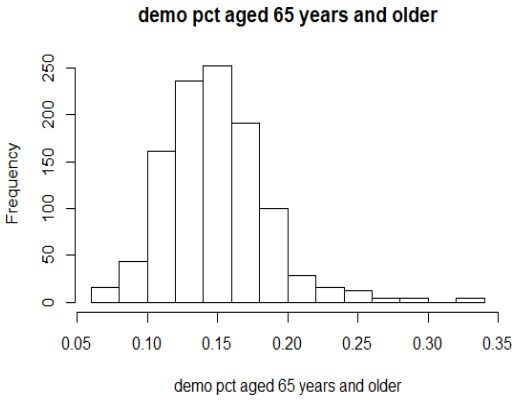
we thought their histograms became much better. We also use sqrt transformation in variables demo_death_rate_per_1k and health_pct_diabetes, and then they become better as well.

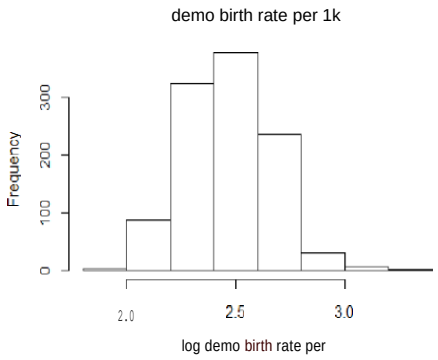


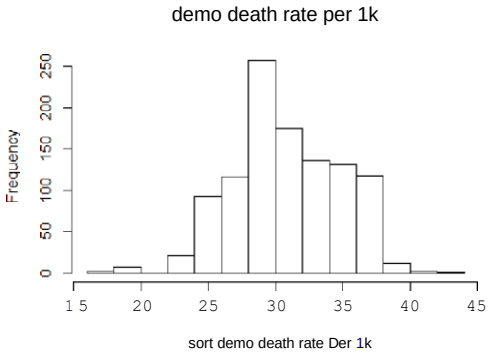


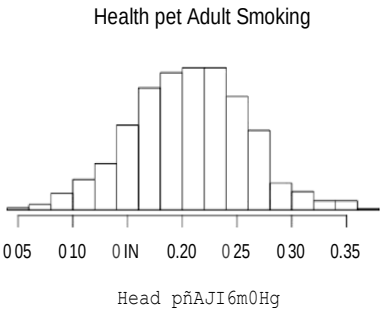


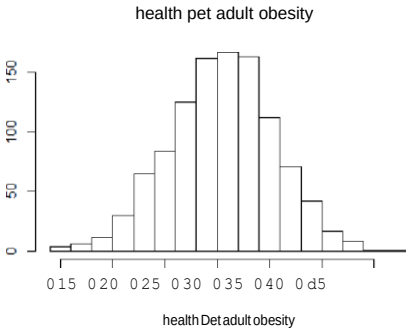


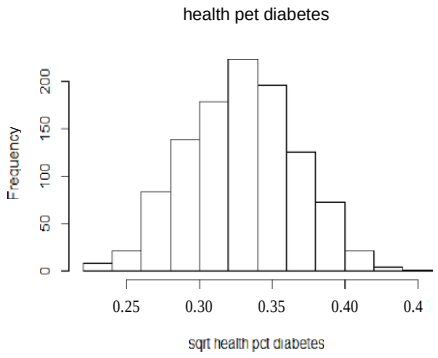


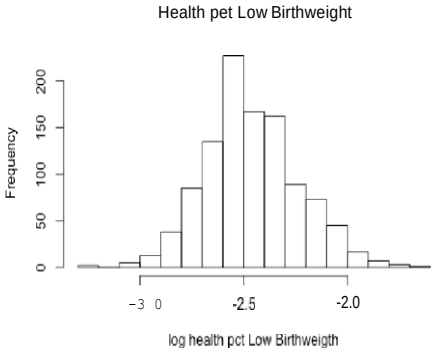

demo birth rate per 1k
Frequency
0
100
200
300
2.0
2.5
3.0
log demo birth rate per

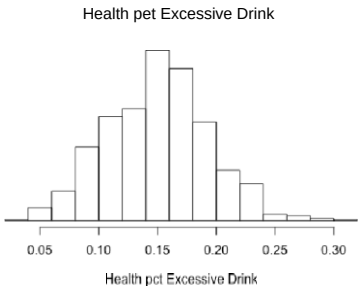

demo death rate per 1k
Frequency
0
50
100
150
200
250
15
20
25
30
35
40
45
sort demo death rate Der 1k

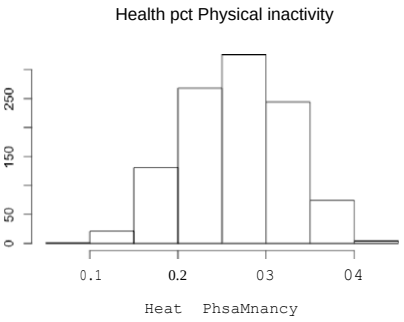

Health pet Adult Smoking
0 05
0 10
0 IN
0.20
0 25
0 30
0.35
Head pñAJI6m0Hg

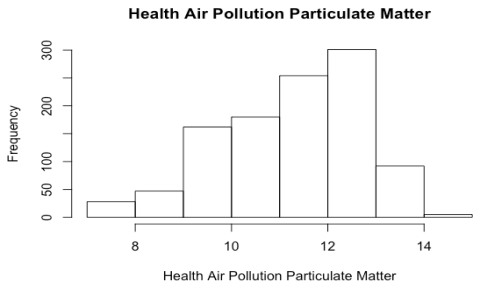

health pet adult obesity
0
50
100
150
015
020
025
030
035
040
0d5
health Det adult obesity

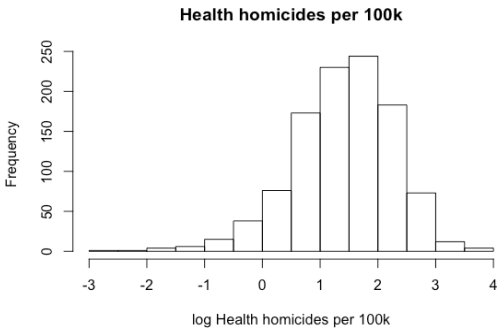

health pet diabetes
Frequency
0
50
100
150
200
0.25
0.30
0.35
0.40
0.4
sqrt health pct diabetes

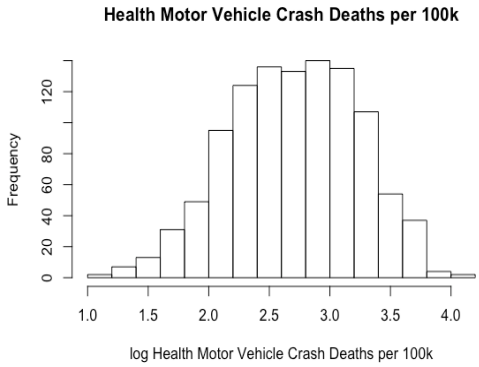

Health pet Low Birthweight
Frequency
0
50
100
150
200
-3 0
-2.5
-2.0
log health pct Low Birthweigth

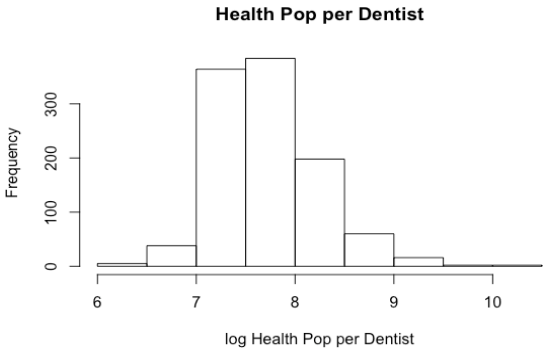

Health pet Excessive Drink
0.05
0.10
0.15
0.20
0.25
0.30
Health pct Excessive Drink

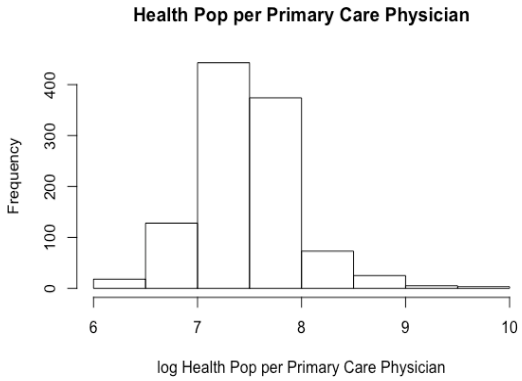

Health pct Physical inactivity
0
50
150
250
0.1
0.2
03
04
Heat PhsaMnancy







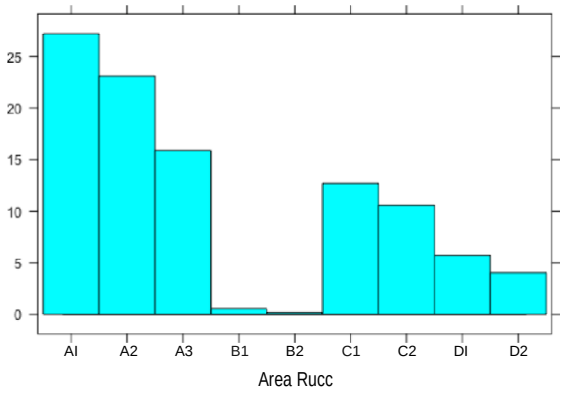




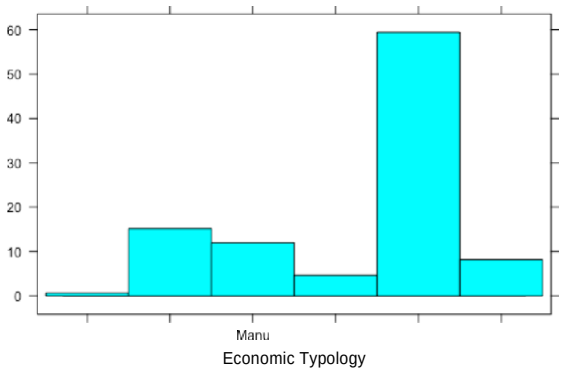
Categorical Variable Histogram:

According to the plot, in the Area Rucc plot, the population number in B1 and B2 (Nonmetro-completed rural or less than 2,500 urban population, adjacent to a metro area, Nonmetro-completed rural or less than 2,500 urban population, not adjacent to a metro area) is the least compared to other sub-variables. The Nonsp(non-specialized) is the main economic typology for different counties which means the residents do not rely on the other five economic typologies. Moreover, the M1 and M2(Large-in a metro area with at least 1 million residents or more, Small-in a metro area with fewer than 1 million residents) are the most impacted for area urban.

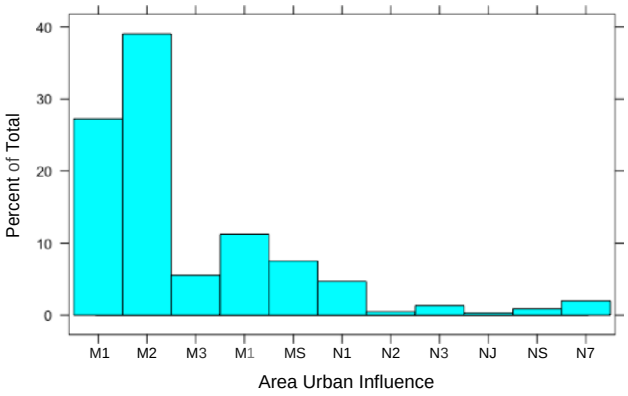
Area Rucc

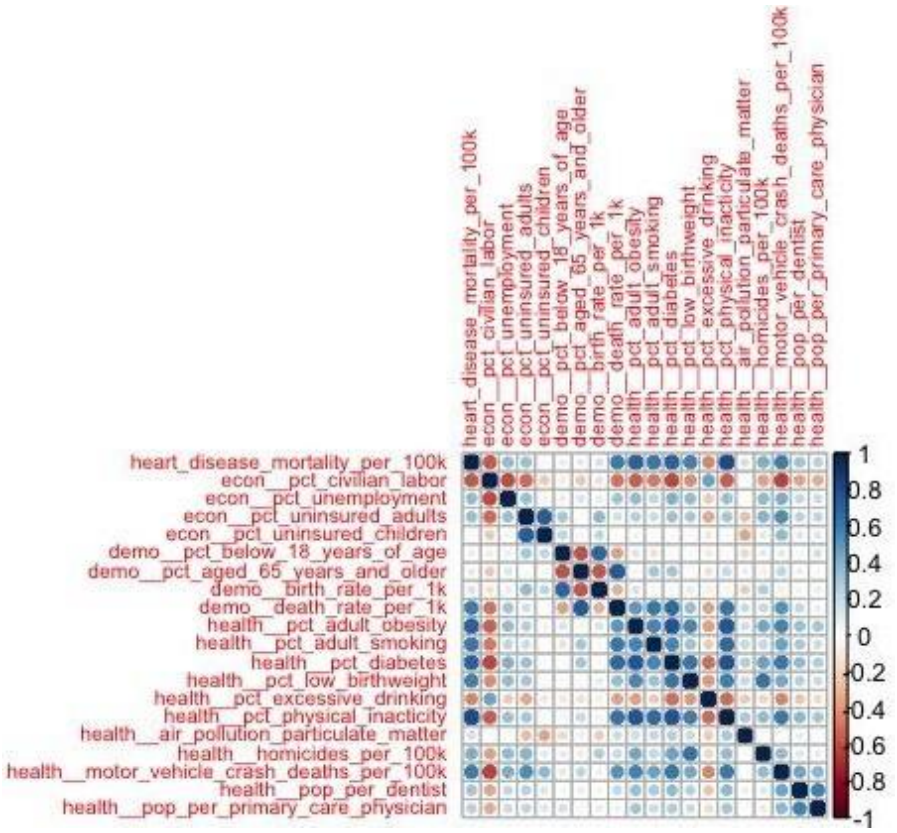


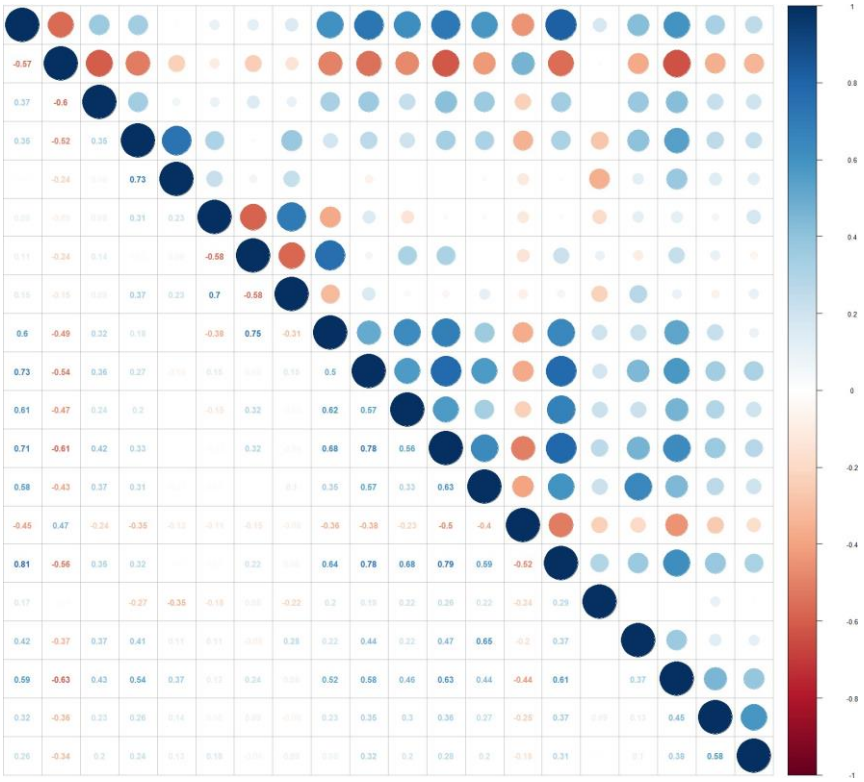
Economic Typology



Area Urban Influence







The difference between these two plots is with the variable name and without the variable name which shows more clearly for the relationship with the response variable. Based on the correlation plot, we see that adult obesity, health pct diabetes, the death rate per 1k, adult smoking, low birth weight, and physical inactivity are highly correlated to heart disease

mortality per 100k. All of the independent variables are having a positive trend with the dependent variable.

Moreover, we are mainly interested in those two parameters which are heart_disease_mortality_per_100k and demo_death_ rate_ per_ 1k.

These two variables are both mortality rates. One is the general mortality rate and the other is the amount of death caused by heart disease, and we are interested in the latter one most. For doing this, we will try demographic features, economic indicators, local information, and health indicators to make the prediction and analysis of the hidden information.

Appendix: The Data Dictionary

Name of Variable	Type of variable	Description
heart_*disease*_mortality_*per*_100k	numeric	Amount of death due to heart disease per 100k people
Economic Indicators		
econ_*pct*civilian_labor	numeric	Civilian labor force, annual average, as the percent of the population
econ_*pct*unemployment	numeric	Unemployment, annual average, as the percent of the population
econ_pct*uninsured*adults	numeric	Percent of adults without health insurance
econ_pct*uninsured*children	numeric	Percent of children without health insurance
econ_*economic*typology	categorical	County Typology Codes classify all U.S. counties according to six mutually exclusive categories of economic dependence and six overlapping categories of policy-relevant themes.
Demographics Information		
demo_*pct*below*18*years*of*age	numeric	Percent of population that is below 18 years of age
demo_*pct*aged*65*years*and*older	numeric	Percent of population that is aged 65 years or older
demo_*birth*rate*per*1k	numeric	Births per 1,000 of population
demo_*death*rate*per*1k	numeric	Deaths per 1,000 of population
Health Indicators		
health_*pct*adult_obesity	numeric	Percent of adults who meet the clinical definition of obese
health_*pct*adult_smoking	numeric	Percent of adults who smoke
health_*pct*diabetes	numeric	Percent of population with diabetes
health_*pct*low_birthweight	numeric	Percent of babies born with low birth weight
health_*pct*excessive_drinking	numeric	Percent of the adult population that engages in excessive consumption of alcohol
health_*pct*physical_inacticity	numeric	Percent of the adult population that is physically inactive
health_*air*pollution*particulate*matter	numeric	Fine particulate matter in µg/m³
health_*homicides*per_100k	numeric	Deaths by homicide per 100,000 population
health_*motor*vehicle*crash*deaths*per*100k	numeric	Deaths by motor vehicle crash per 100,000 population
health_*pop*per_dentist	numeric	Population per dentist
health_*pop*per*primary*care_p	numeric	Population per Primary Care Physician

hysician ***Information About The County***		
Area_Rucc	categorical	Rural-Urban Continuum Codes form a classification scheme that distinguishes metropolitan counties by the population size of their metro area and nonmetropolitan counties by the degree of urbanization and adjacency to a metro area. The official Office of Management and Budget (OMB) metro and nonmetro categories have been subdivided into three metro and six nonmetro categories. Each county in the U.S. is assigned one of the 9 codes
Area_*Urban*Influence	categorical	Urban Influence Codes form a classification scheme that distinguishes metropolitan counties by population size of their metro area and nonmetropolitan counties by the size of the largest city or town and proximity to metro and micropolitan areas

Annotation

A1	**Metro - Counties in metro areas of 1 million population or more**
A2	**Metro - Counties in metro areas of 250,000 to 1 million population**
A3	**Metro - Counties in metro areas of fewer than 250,000 population**
B1	**Nonmetro - Completely rural or less than 2,500 urban population, adjacent to a metro area**
B2	**Nonmetro - Completely rural or less than 2,500 urban population, not adjacent to a metro area**
C1	**Nonmetro - Urban population of 20,000 or more, adjacent to a metro area**
C2	**Nonmetro - Urban population of 2,500 to 19,999, adjacent to a metro area**
D1	**Nonmetro - Urban population of 20,000 or more, not adjacent to a metro area**
D2	**Nonmetro - Urban population of 2,500 to 19,999, not adjacent to a metro area**
N1	**Noncore adjacent to a small metro with town of at least 2,500 residents**
N2	**Noncore adjacent to a small metro and does not contain a town of at least 2,500 residents**
N3	**Noncore adjacent to micro area and contains a town of 2,500-19,999 residents**
N4	**Noncore adjacent to micro area and does not contain a town of at least 2,500 residents**
N5	**Noncore not adjacent to a metro/micro area and contains a town of 2,500 or more residents**
N6	**Noncore not adjacent to a metro/micro area and does not contain a town of at least**
N7	**Noncore adjacent to a large metro area**
M1	**Large-in a metro area with at least 1 million residents or more**
M2	**Small-in a metro area with fewer than 1 million residents**
M3	**Micropolitan adjacent to a large metro area**
M4	**Micropolitan adjacent to a small metro area**
M5	**Micropolitan not adjacent to a metro area**
Farm	**Farm-dependent**
Fed	**Federal/State government-dependent**
Manu	**Manufacturing-dependent**
Mining	**Mining-dependent**
Nonsp	**Nonspecialized**
Recr	**Recreation**

Dataset:

https://www.kaggle.com/nandvard/microsoft-data-science-capstone

R Code:

1.
```
dt<- na.omit(hd)
```

2.
```
> dt = na.omit(hd)
> hist(dt$econ pct_civilian_labor,main="econ pct civilian labor",xlab="econ pct civilian labor")
> hist(log(dt$econ pct_unemployment),main="econ pct unemployment",xlab="log econ pct unemployment")
> hist(dt$econ pct_uninsured_adults,main="econ pct uninsured adults",xlab="econ pct uninsured adults")
> hist(log(dt$econ pct_uninsured_children),main="econ pct uninsured children",xlab="log econ pct uninsured children")
> hist(dt$demo pct_below_18_years_of_age,main="demo pct below 18 years of age",xlab="demo pct below 18 years of age")
> hist(dt$demo pct_aged_65_years_and_older,main="demo pct aged 65 years and older",xlab="demo pct aged 65 years and older")
> hist(log(dt$demo birth_rate_per_1k),breaks=10,main="demo birth rate per 1k",xlab="log demo birth rate per 1k")
> hist(sqrt(dt$demo death_rate_per_1k),main="demo death rate per 1k",xlab="sqrt demo death rate per 1k")
> hist(dt$health pct_adult_obesity,main="health pct adult obesity",xlab="health pct adult obesity")
> hist(sqrt(dt$health pct_diabetes),main="health pct diabetes",xlab="sqrt health pct diabetes")
```

```
library(lattice)
library(tidyverse)
hist(log(dt$health_pct_low_birthweight), main = "Health pct Low Birthweight", xlab = "log health pct Low Birthweigth")
```

```
hist(dt$health_pct_excessive_drinking, main= "Health pct Excessive Drink", xlab = "Health pct Excessive Drink")
hist(dt$health_pct_physical_inacticity, main = "Health pct Physical Inacticity", xlab = "Health pct Physical Inacticity")
hist(dt$health_air_pollution_particulate_matter, breaks = 10, main = "Health Air Pollution Particulate Matter", xlab = "Health Air Pollution Particulate Matter")
hist(log(dt$health_homicides_per_100k), main = "Health homicides per 100k", xlab = "log Health homicides per 100k")
hist(log(dt$health_motor_vehicle_crash_deaths_per_100k), main = "Health Motor Vehicle Crash Deaths per 100k", xlab = "log Health Motor Vehicle Crash Deaths per 100k")
hist(log(dt$health_pop_per_dentist), main = "Health Pop per Dentist", xlab = "log Health Pop per Dentist")
hist(log(dt$health_pop_per_primary_care_physician), main = "Health Pop per Primary Care Physician", xlab = "log Health Pop per Primary Care Physician")
hist(dt$health_pct_adult_smoking, main = "Health pct Adult Smoking", xlab = "Health pct Adult Smoking")
histogram(~ Area_Rucc, data = dt, main = "Area Rucc", xlab = "Area Rucc")
histogram(~ Econ_Economic_typology, data = dt, main = "Economic Typology", xlab = "Economic Typology")
histogram(~ Area_Urban_Influence, data = dt, main = "Area Urban Influence", xlab = "Area Urban Influence")
dt$heart_disease_mortality_per_100k<- as.numeric(dt$heart_disease_mortality_per_100k)


#3
library(corrplot)
c<- cor(dt[2:21],)
corrplot(c,method = "circle",tl.pos = NULL, tl.cex = 0.65)

corrplot(c,type="upper",tl.cex = 0.3,tl.pos="n")

corrplot(c,add=TRUE, type="lower", method="number",diag=FALSE,tl.pos="n", cl.pos="n")
```