SDS348 Project

Seungchul Yeom

3/21/2021

Data Imports

Introduction

I was curious about the relationship between the income and Suicide rate, so I brought three data from County health ranking website and applied into this project. I brought several types of income such as income inequality and 20th percentile income. Joining all three data sets will make too much variables that are not necessary, so they will be deleted by manipulating the data. The variables that I consider to include for the final data are County, Number of Death, Percentile of income

```
library(readxl)
IncomeMed <- read_excel("IncomeMed.xlsx")
```

```
## New names:
## * `` -> ...2
```

```
IncomeMed
```

```
## # A tibble: 254 x 9
##    County   ...2  `County Value` `Error Margin`  AIAN Asian Black Hispanic White
##    <chr>    <lgl>          <dbl> <chr>          <dbl> <dbl> <dbl>    <dbl> <dbl>
##  1 Anderson NA             46000 $40,700-51,200 28100 62500 26500    42000 47400
##  2 Andrews  NA             84900 $81,800-88,100    NA    NA 36500    73900 75500
##  3 Angelina NA             46700 $42,100-51,200 50300 83500 29000    42200 53100
##  4 Aransas  NA             46900 $41,000-52,900    NA    NA    NA    37800 47800
##  5 Archer   NA             61200 $53,500-68,900    NA    NA    NA    33900 67800
##  6 Armstro~ NA             57200 $49,800-64,600    NA    NA    NA       NA    NA
##  7 Atascosa NA             50600 $44,900-56,300    NA    NA    NA    49000 66900
##  8 Austin   NA             59900 $54,600-65,300    NA    NA 29300    60100 78400
##  9 Bailey   NA             45100 $39,500-50,600    NA    NA    NA    36900 64200
## 10 Bandera  NA             53000 $47,900-58,100    NA    NA 75500    56900 53700
## # ... with 244 more rows
```

```
Suicide <- read_excel("Suicide.xlsx")
```

```
## New names:
## * `` -> ...2
```

```
Suicide
```

```
## # A tibble: 254 x 11
##    County  ...2  `# Deaths` `County Value` `Error Margin (Ag~ `Crude Rate`  AIAN
##    <chr>   <lgl>      <dbl>          <dbl> <chr>                      <dbl> <dbl>
##  1 Anders~ NA            80             25 20-31                         28    NA
##  2 Andrews NA            18             21 12754                         20    NA
##  3 Angeli~ NA            64             14 44518                         15    NA
##  4 Aransas NA            32             21 14-31                         26    NA
##  5 Archer  NA            11             23 15646                         25    NA
##  6 Armstr~ NA            NA             NA <NA>                          NA    NA
##  7 Atasco~ NA            39             16 44552                         16    NA
##  8 Austin  NA            37             24 17-34                         25    NA
##  9 Bailey  NA            NA             NA <NA>                          NA    NA
## 10 Bandera NA            25             18 44529                         23    NA
## # ... with 244 more rows, and 4 more variables: Asian <dbl>, Black <dbl>,
## #   Hispanic <dbl>, White <dbl>
```

```
IncomeInequal <- read_excel("IncomeInequal.xlsx")
```

```
## New names:
## * `` -> ...2
```

```
IncomeInequal
```

```
## # A tibble: 254 x 6
##    County  ...2  `80th Percentile ~ `20th Percentile~ `County Value` `Z-Score`
##    <chr>   <lgl>              <dbl>             <dbl>          <dbl> <chr>
##  1 Anders~ NA                 82461             19270            4.3 -0.6
##  2 Andrews NA                132125             30987            4.3 -0.62
##  3 Angeli~ NA                 88122             19834            4.4 -0.36
##  4 Aransas NA                102962             19557            5.3 0.81
##  5 Archer  NA                116886             26571            4.4 -0.43
##  6 Armstr~ NA                124750             34021            3.7 -1.47
##  7 Atasco~ NA                105098             23811            4.4 -0.41
##  8 Austin  NA                117074             25103            4.7 -0.05
##  9 Bailey  NA                 91974             25580            3.6 -1.57
## 10 Bandera NA                104633             24256            4.3 -0.5500000~
## # ... with 244 more rows
```

Description

Imported total 3 data from the County Health Rankings websites about Median household income, Suicides, and Income inequality.

Tidy Data

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
#Opt out unnecessary variables. Ethnicity variables had lack of informations (Too many NAs)
IncomeMed = IncomeMed %>%
  select(-...2, -AIAN, -Asian, -Black, -Hispanic,
         - White)

Suicide = Suicide %>%
  select(-...2, -AIAN, -Asian, -Black, -Hispanic,
         - White, -`Crude Rate`) %>%
  select(-`Error Margin (Age-adjusted)`)

IncomeInequal = IncomeInequal %>%
  select(-...2, -`County Value`, -"Z-Score")

head(IncomeMed)
```

```
## # A tibble: 6 x 3
##   County    `County Value` `Error Margin`
##   <chr>              <dbl> <chr>
## 1 Anderson           46000 $40,700-51,200
## 2 Andrews            84900 $81,800-88,100
## 3 Angelina           46700 $42,100-51,200
## 4 Aransas            46900 $41,000-52,900
## 5 Archer             61200 $53,500-68,900
## 6 Armstrong          57200 $49,800-64,600
```

```
head(Suicide)
```

```
## # A tibble: 6 x 3
##   County    `# Deaths` `County Value`
##   <chr>          <dbl>          <dbl>
## 1 Anderson          80             25
## 2 Andrews           18             21
## 3 Angelina          64             14
## 4 Aransas           32             21
## 5 Archer            11             23
## 6 Armstrong         NA             NA
```

```
head(IncomeInequal)
```

```
## # A tibble: 6 x 3
##   County    `80th Percentile Income` `20th Percentile Income`
##   <chr>                        <dbl>                    <dbl>
## 1 Anderson                     82461                    19270
## 2 Andrews                     132125                    30987
## 3 Angelina                     88122                    19834
## 4 Aransas                     102962                    19557
## 5 Archer                      116886                    26571
## 6 Armstrong                   124750                    34021
```

The data is already tidy because it fulfills three assumptions which are:

1. Each variable have its own column. 2. Each observation have its own row. 3. Each value have its own cell.

Join/Merge

```
#Join all 3 separate data sources into a single dataset.
Income_Suicide = left_join(Suicide, IncomeMed,
                           by = "County")
IncomeInequal_Suicide = left_join(Income_Suicide,
                                  IncomeInequal,
                                  by = "County")
library(tidyr)
#Change colnames
IncomeInequal_Suicide = IncomeInequal_Suicide %>%
  rename(Suicide_Rate = "County Value.x",
         "Household_Income ($)" = "County Value.y",
         Death_Num = "# Deaths")

head(IncomeInequal_Suicide)
```

```
## # A tibble: 6 x 7
##   County Death_Num Suicide_Rate `Household_Inco~ `Error Margin` `80th Percentil~
##   <chr>      <dbl>        <dbl>            <dbl> <chr>                     <dbl>
## 1 Ander~        80           25            46000 $40,700-51,200            82461
## 2 Andre~        18           21            84900 $81,800-88,100           132125
## 3 Angel~        64           14            46700 $42,100-51,200            88122
## 4 Arans~        32           21            46900 $41,000-52,900           102962
## 5 Archer        11           23            61200 $53,500-68,900           116886
## 6 Armst~        NA           NA            57200 $49,800-64,600           124750
## # ... with 1 more variable: 20th Percentile Income <dbl>
```

Discussion *The variable County is the common categorical variable among all three data. Also, changed the colnames to understandable names. The size of the this joined and cleaned data has 7 variables and 159 observations.*

Summary Statistics

```
IncomeInequal_Suicide = IncomeInequal_Suicide %>%
  mutate(Inequality = `80th Percentile Income`
         - `20th Percentile Income`) %>%
  arrange(Suicide_Rate) %>%
  select(-`Error Margin`) %>%
  filter(!is.na(Suicide_Rate)) %>%
  mutate(State = "Texas") #create a new variable
```

```
#Create 10 summary statistics
stat = IncomeInequal_Suicide %>%
  group_by(State) %>%
  summarize(Mean_Suicide =
              mean(IncomeInequal_Suicide$Suicide_Rate),
            Sd_Suicide =
              sd(IncomeInequal_Suicide$Suicide_Rate),
            Min_Suicide =
              min(IncomeInequal_Suicide$Suicide_Rate),
            Max_Suicide =
              max(IncomeInequal_Suicide$Suicide_Rate),
            Med_Suicide =
              median(IncomeInequal_Suicide$Suicide_Rate),
            Num_County =
              n(),
            Var_Suicide =
              var(IncomeInequal_Suicide$Suicide_Rate),
            Mad_Suicide =
              mad(IncomeInequal_Suicide$Suicide_Rate),
            IQR_Suicide =
              IQR(IncomeInequal_Suicide$Suicide_Rate))

stat
```

```
## # A tibble: 1 x 10
##   State Mean_Suicide Sd_Suicide Min_Suicide Max_Suicide Med_Suicide Num_County
## * <chr>        <dbl>      <dbl>       <dbl>       <dbl>       <dbl>      <int>
## 1 Texas         17.7       5.40           5          33          17        159
## # ... with 3 more variables: Var_Suicide <dbl>, Mad_Suicide <dbl>,
## #   IQR_Suicide <dbl>
```

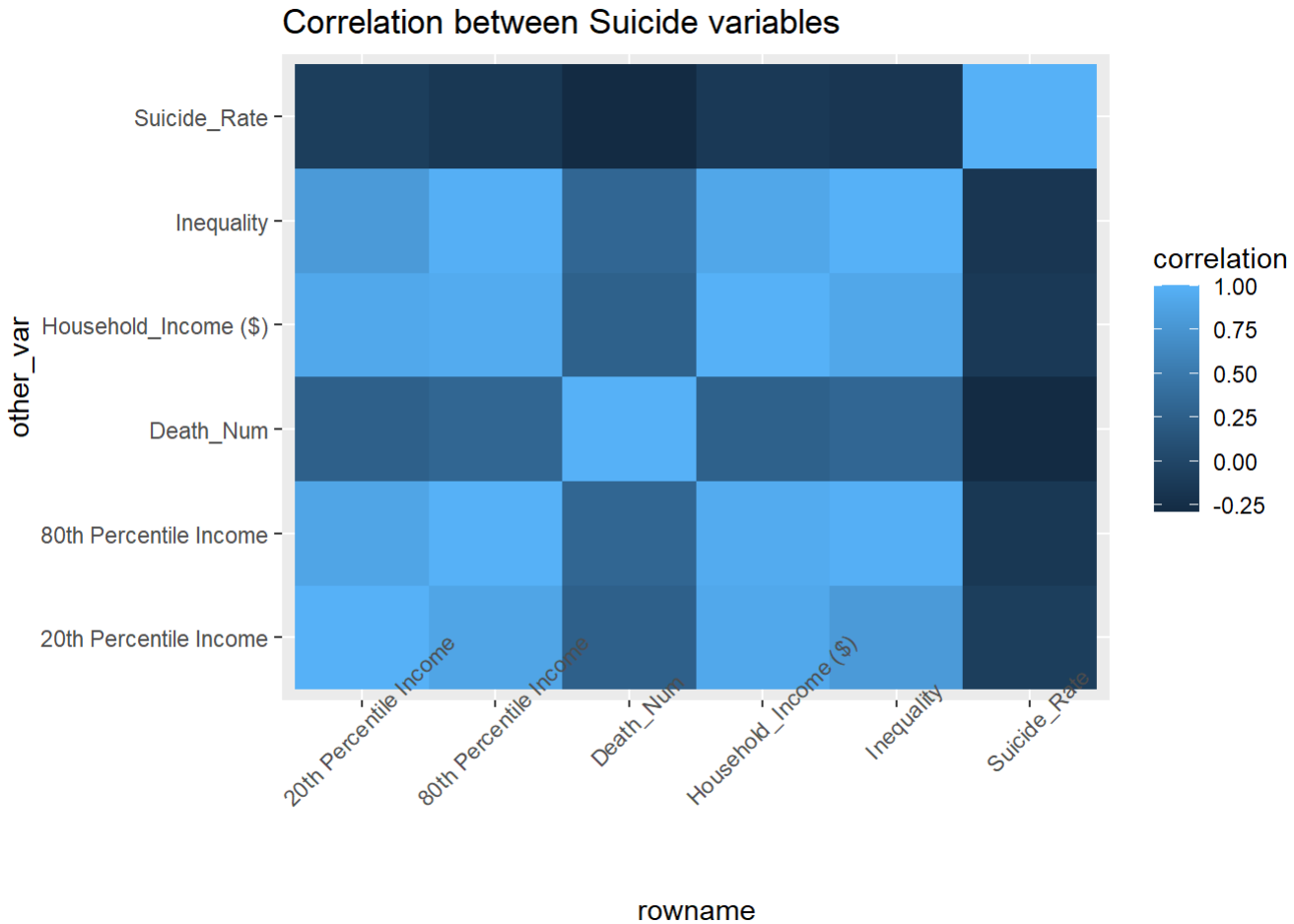
```
quantile(IncomeInequal_Suicide$Suicide_Rate)
```

```
##   0%  25%  50%  75% 100%
##    5   14   17   21   33
```

Discussion *Suicide rate mean, std, min, max, med, number of county, and quantile were 17.7 %, 5.4 %, 5 %, 33 %, 17 %, and 159, respectively.*

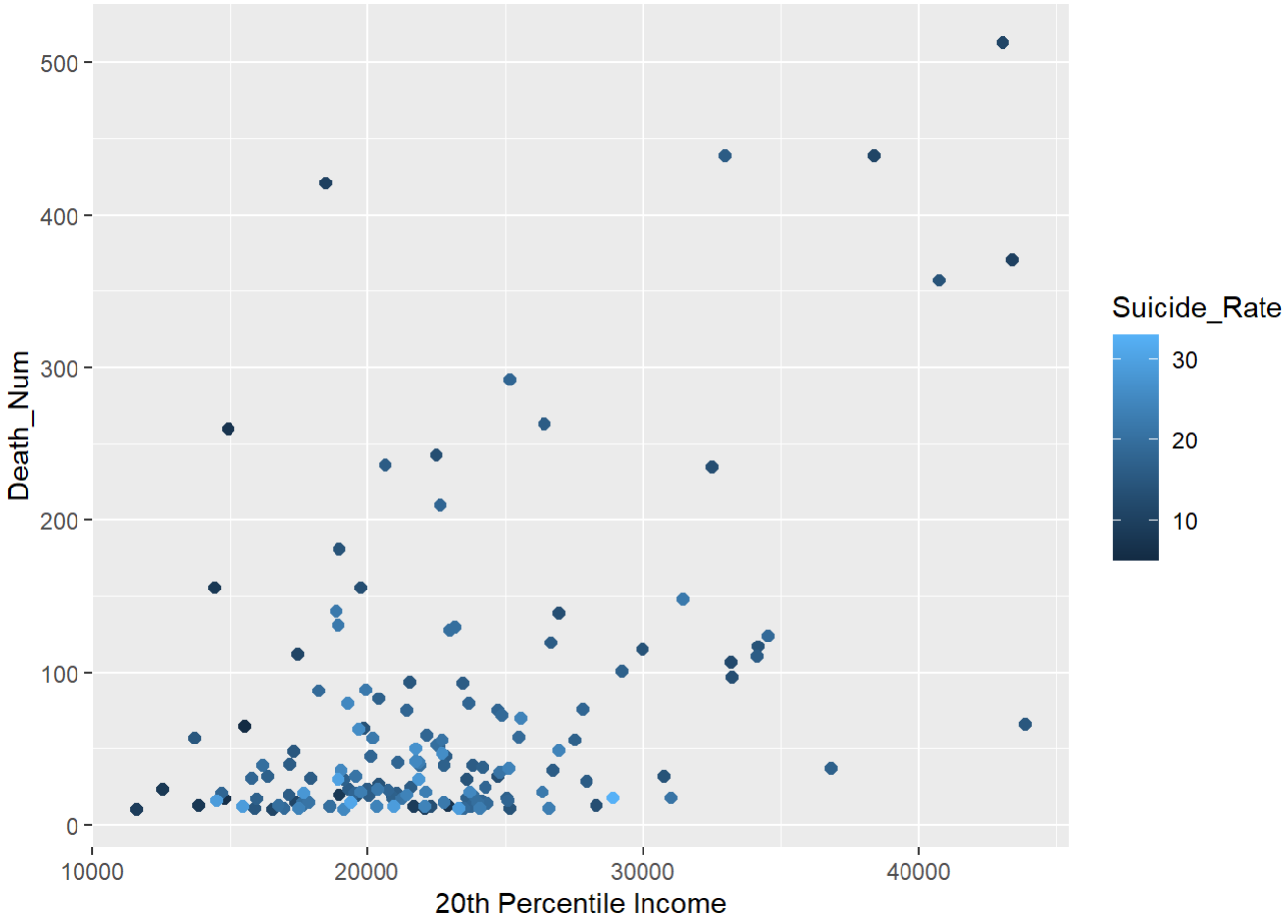
Visualizations

```
#Visualizations
library(ggplot2)
library(tibble)
#Heatmap
Correlation = IncomeInequal_Suicide %>%
  select_if(is.numeric)
cor(Correlation, use = "pairwise.complete.obs") %>%
  as.data.frame %>%
  rownames_to_column %>%
  pivot_longer(-1, names_to = "other_var",
               values_to = "correlation") %>%
  ggplot(aes(rowname, other_var, fill = correlation)) +
  geom_tile() +
  theme(axis.text.x = element_text(angle = 45)) +
  ggtitle(label = "Correlation between Suicide variables")
```



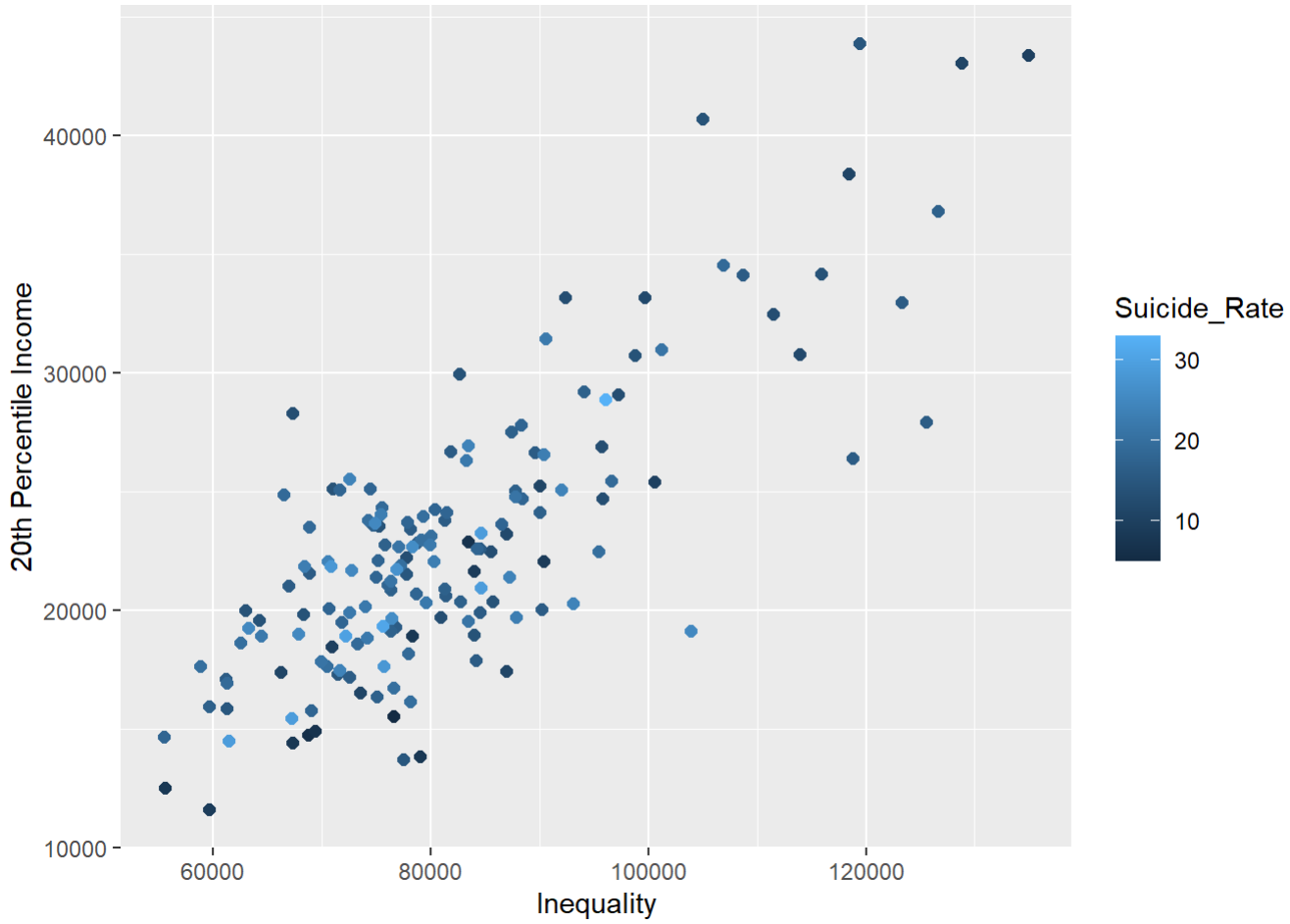
According to this heatmap, the correlation between Suicide rate and any other variable are really low. On the other hand, high correlations are observed among the variables that are related to income. The interesting finding of this heatmap was decent correlation amount with number of death (Suicide).

```
#Graph 2
IncomeInequal_Suicide %>%
  filter(Death_Num <= 700) %>%
  ggplot(aes(x = `20th Percentile Income`,
             y = Death_Num,
             color = Suicide_Rate)) +
  geom_point(size = 2)
```



20th percentile income, death (suicide) number, and suicide rate are the numeric variables that are compared each other. It is hard to determine that the counties with low income and low death number has whether low or high suicide rate. However, counties with high 20th percentile income and death (suicide) number has lower suicide rate compared to other data points. The data points with light blue color are crowded in the area of low 20th percentile income and low death (suicide) number.

```
#Graph3
IncomeInequal_Suicide %>%
  ggplot(aes(x = Inequality,
             y = `20th Percentile Income`,
             color = Suicide_Rate)) +
  geom_point(size = 2)
```



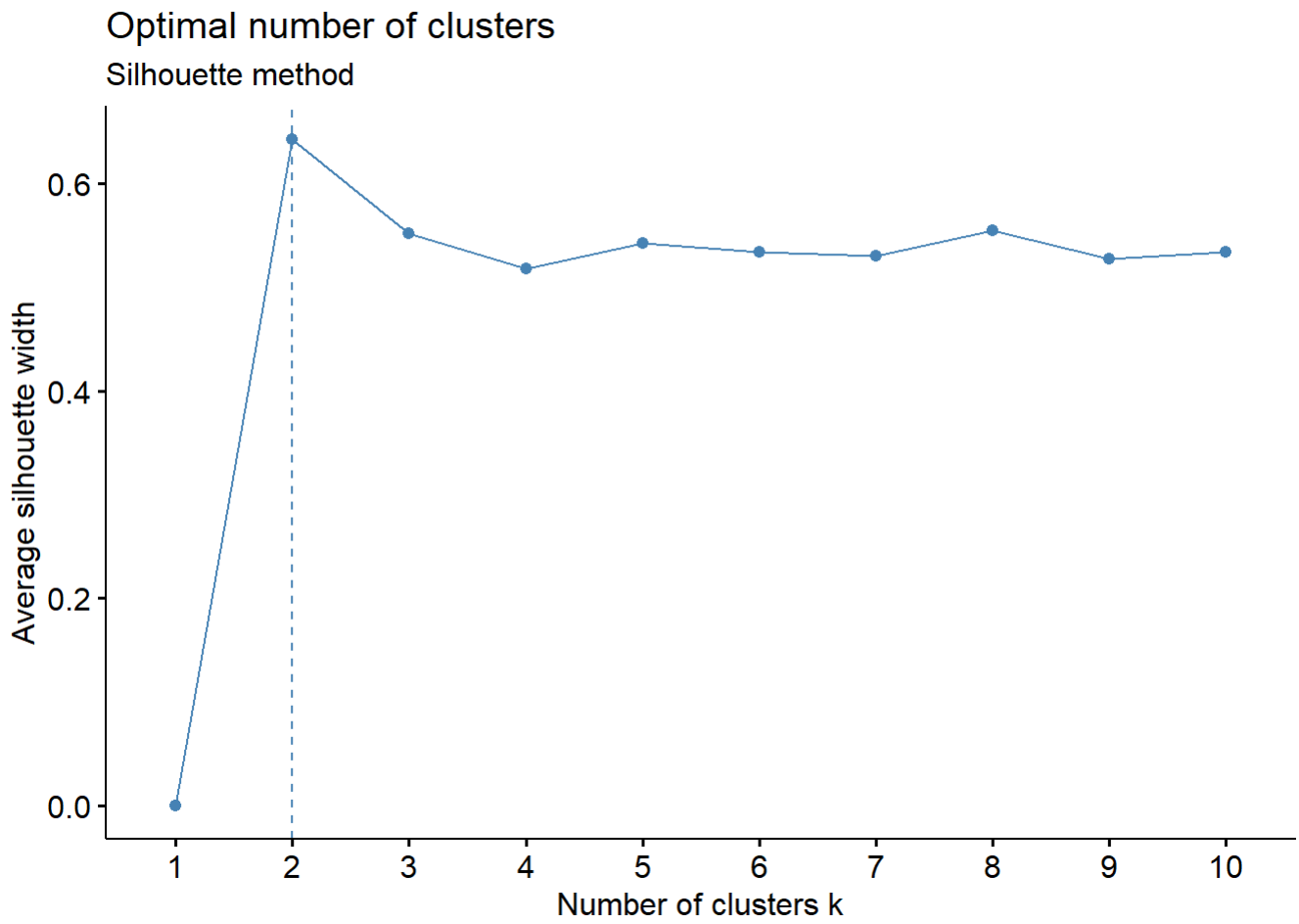
Income inequality is the difference between 80th percentile income and 20th percentile income. As inequality and 20th percentile income increase, the number of light blue data points that indicate high suicide rate are decreased. I anticipated that the trend between income inequality and suicide rate would be negative, but the correlation value and this graph showed that those two variables are not quiet related.

##Dimensionality Reduction

```
library(factoextra)
```

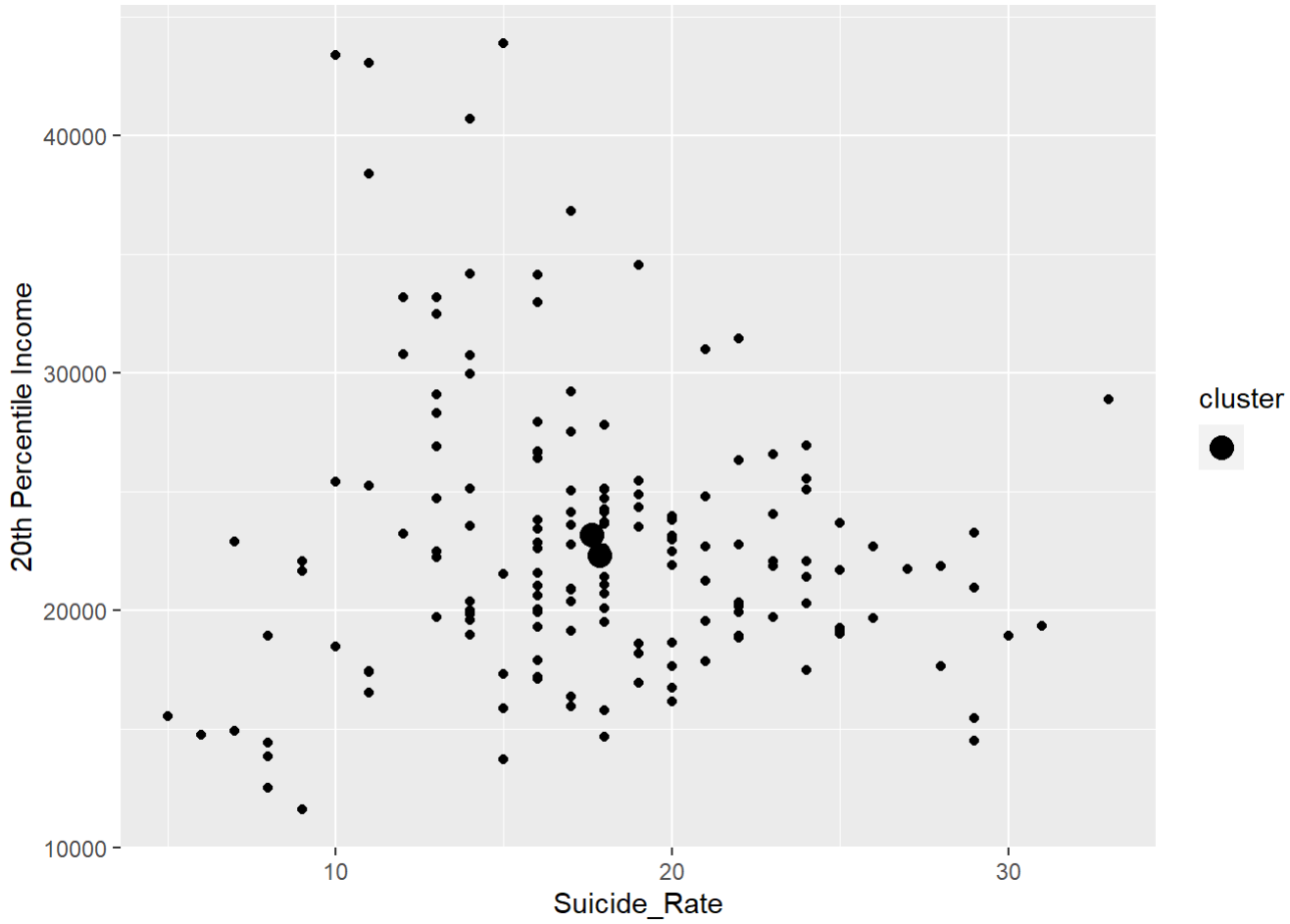
```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(cluster)
clust_data = IncomeInequal_Suicide %>%
  select(Suicide_Rate, `20th Percentile Income`)

fviz_nbclust(clust_data, kmeans, method =
               "silhouette") +
  labs(subtitle = "Silhouette method")
```

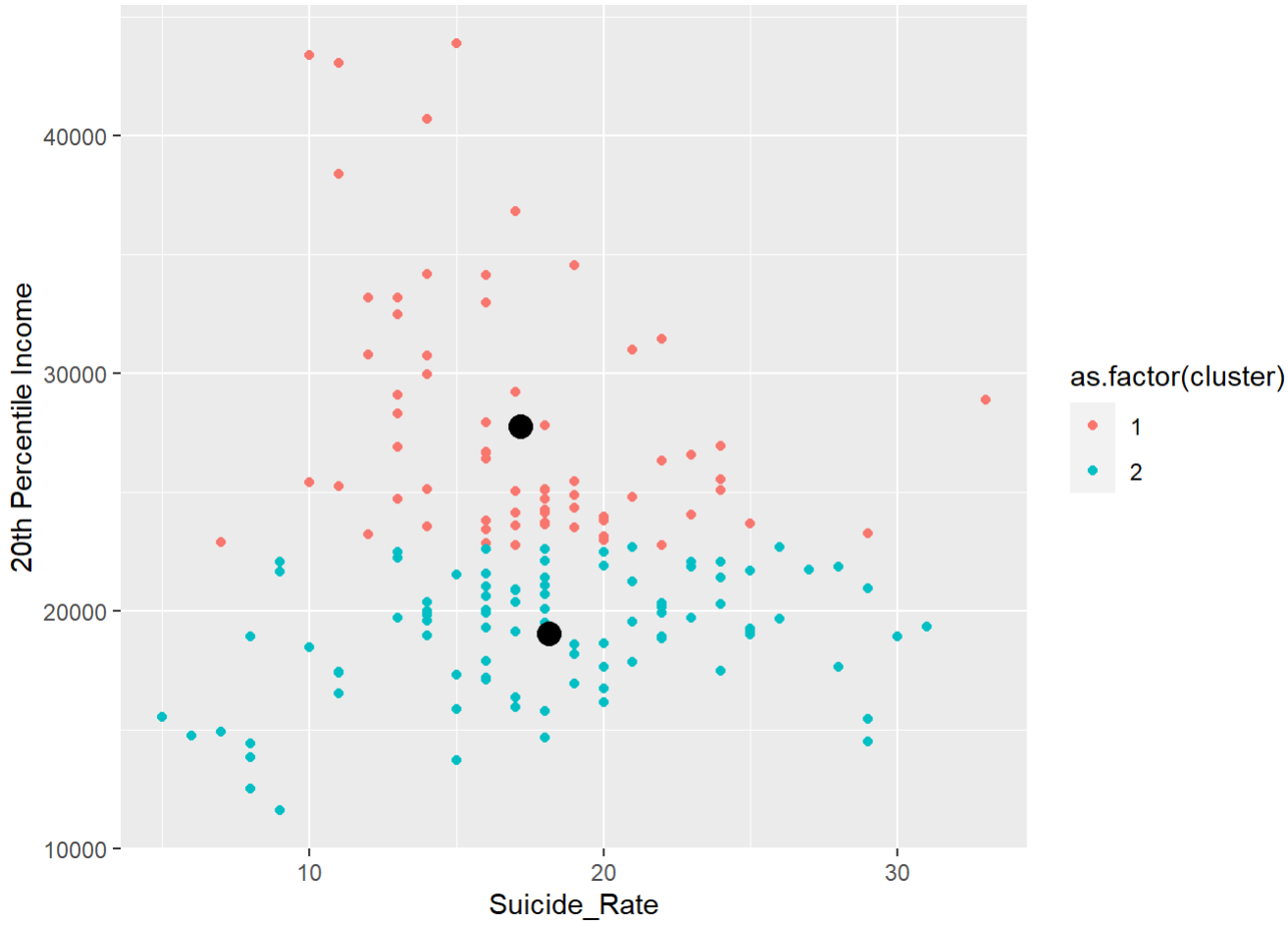


I used the silhouette Method to determine the optimal number of clusters. The optimal number of clusters value k is the maximum value which was 2 in this case.

```
# Initialize random means
centers = IncomeInequal_Suicide %>%
  mutate(cluster = sample(c('1','2'), 159, replace = T)) %>%
  group_by(cluster) %>%
  summarize(Suicide_Rate = mean(Suicide_Rate),
            `20th Percentile Income` =
              mean(`20th Percentile Income`))

# Plot centers
ggplot(IncomeInequal_Suicide) +
  geom_point(aes(Suicide_Rate, `20th Percentile Income`)) +
  geom_point(data = centers, aes(Suicide_Rate,
                                 `20th Percentile Income`, fill = ""),
             color = "black",
             size = 4) +
  scale_fill_manual(name = "cluster", values = "black")
```




```
iter = IncomeInequal_Suicide %>%
  mutate(dist1 = sqrt((Suicide_Rate - centers$Suicide_Rate[1])^2 + (`20th Percentile Income` - c
enters$`20th Percentile Income`[1])^2),
         dist2 = sqrt((Suicide_Rate - centers$Suicide_Rate[2])^2 + (`20th Percentile Income` - c
enters$`20th Percentile Income`[2])^2)) %>%
  rowwise() %>%
  mutate(cluster = which.min(c(dist1,dist2))) %>%
  ungroup()

#New Centers
Centers = iter %>%
  group_by(cluster) %>%
  summarize(Suicide_Rate = mean(Suicide_Rate),
            `20th Percentile Income` =
              mean(`20th Percentile Income`))

#visualize
ggplot(iter) +
  geom_point(aes(Suicide_Rate, `20th Percentile Income`, color = as.factor(cluster))) +
  geom_point(data = Centers, aes(Suicide_Rate,
                                 `20th Percentile Income`), color = "black", size = 4)
```




```
iter %>%
  mutate(gmeanSR = mean(Suicide_Rate),
         gmean2P = mean(`20th Percentile Income`)) %>%
  group_by(gmean2P, gmeanSR, cluster) %>%
  summarize(meanSR = mean(Suicide_Rate),
            mean2P = mean(`20th Percentile Income`),
            WSS = sum((Suicide_Rate - meanSR)^2 +
                      (`20th Percentile Income` -                          mean2P)^2),
            BSS = sum((meanSR-gmeanSR)^2 +
                      (mean2P-gmean2P)^2)) %>%
  ungroup() %>%
  summarize(BSS=sum(BSS), WSS=sum(WSS))
```

```
## `summarise()` has grouped output by 'gmean2P', 'gmeanSR'. You can override using the `.groups
` argument.
```

```
## # A tibble: 1 x 2
##           BSS         WSS
##         <dbl>       <dbl>
## 1 2964238658. 2502742825.
```

BSS means the sum of distances between the centroids and the total sample mean multipled by the number of points within each cluster, and WSS means the sum of distances between the points and the corresponding centroids for each cluster.

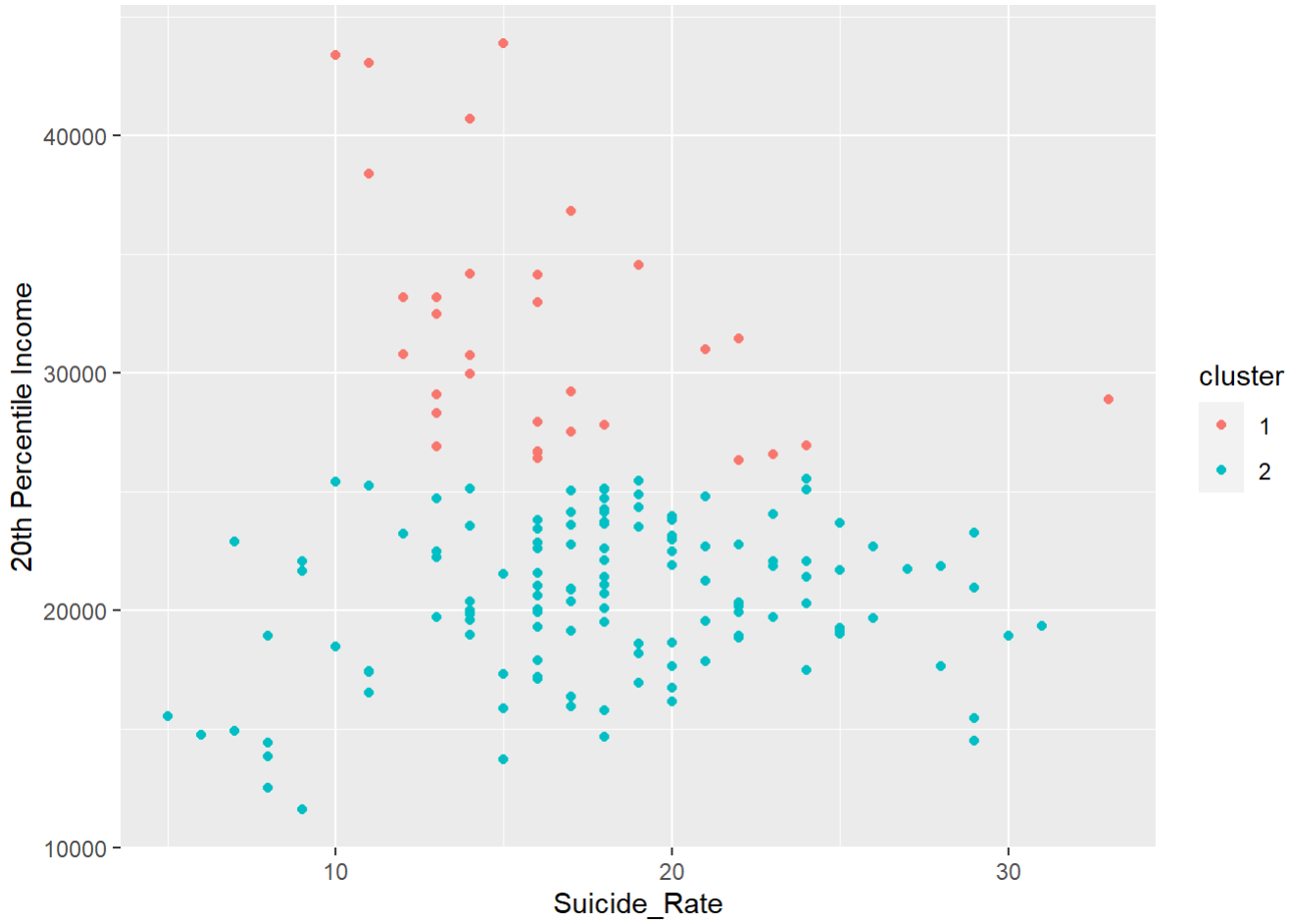
```
kmeans_data = clust_data %>%
  kmeans(2)

kmeans_data
```

```
## K-means clustering with 2 clusters of sizes 32, 127
##
## Cluster means:
##   Suicide_Rate 20th Percentile Income
## 1     16.28125               31876.94
## 2     18.10236               20474.71
##
## Clustering vector:
##   [1] 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 1 2 1 2 2 2 1 1 1 2 1 1 2 2 1 2 1 2 2 2 2 1
##  [38] 2 1 2 2 1 1 2 2 2 1 2 2 1 1 1 2 1 2 1 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [75] 1 1 2 2 2 1 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2
## [112] 2 2 2 2 2 2 1 2 2 2 2 2 2 2 1 2 2 2 1 2 1 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2
## [149] 2 2 2 2 2 2 2 2 2 2 1
##
## Within cluster sum of squares by cluster:
## [1]  855080232 1288857250
##  (between_SS / total_SS =  60.8 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

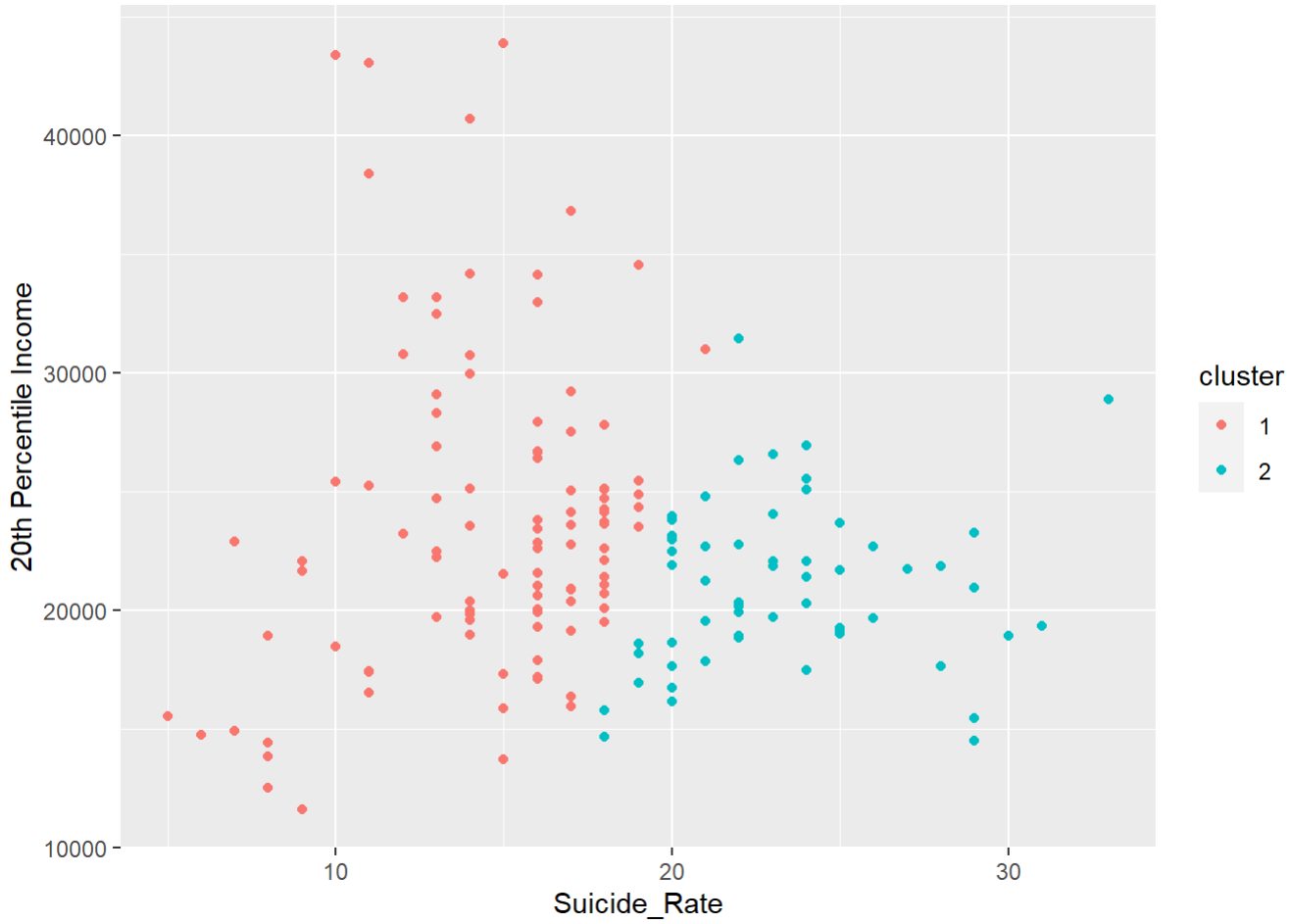
```
kmeans_data$size
```

```
## [1]  32 127
```

```
kmeansclust = clust_data %>%
  mutate(cluster = as.factor(kmeans_data$cluster))

kmeansclust %>%
  ggplot(aes(Suicide_Rate, `20th Percentile Income`, color = cluster)) +
  geom_point()
```




```
kmeans_data2 = clust_data %>%
  scale %>%
  kmeans(2)

clust_data %>%
  mutate(cluster = as.factor(kmeans_data2$cluster)) %>%
  ggplot(aes(Suicide_Rate, `20th Percentile Income`, color = cluster)) +
  geom_point()
```



Kmeans Description

*Each data point is connected to the nearest centroid according to the squared Euclidean distance, and the centroids are recomputed by the mean of the data points in the centroid's cluster. These steps are looped until a certain criteria is fulfilled.

PCA data

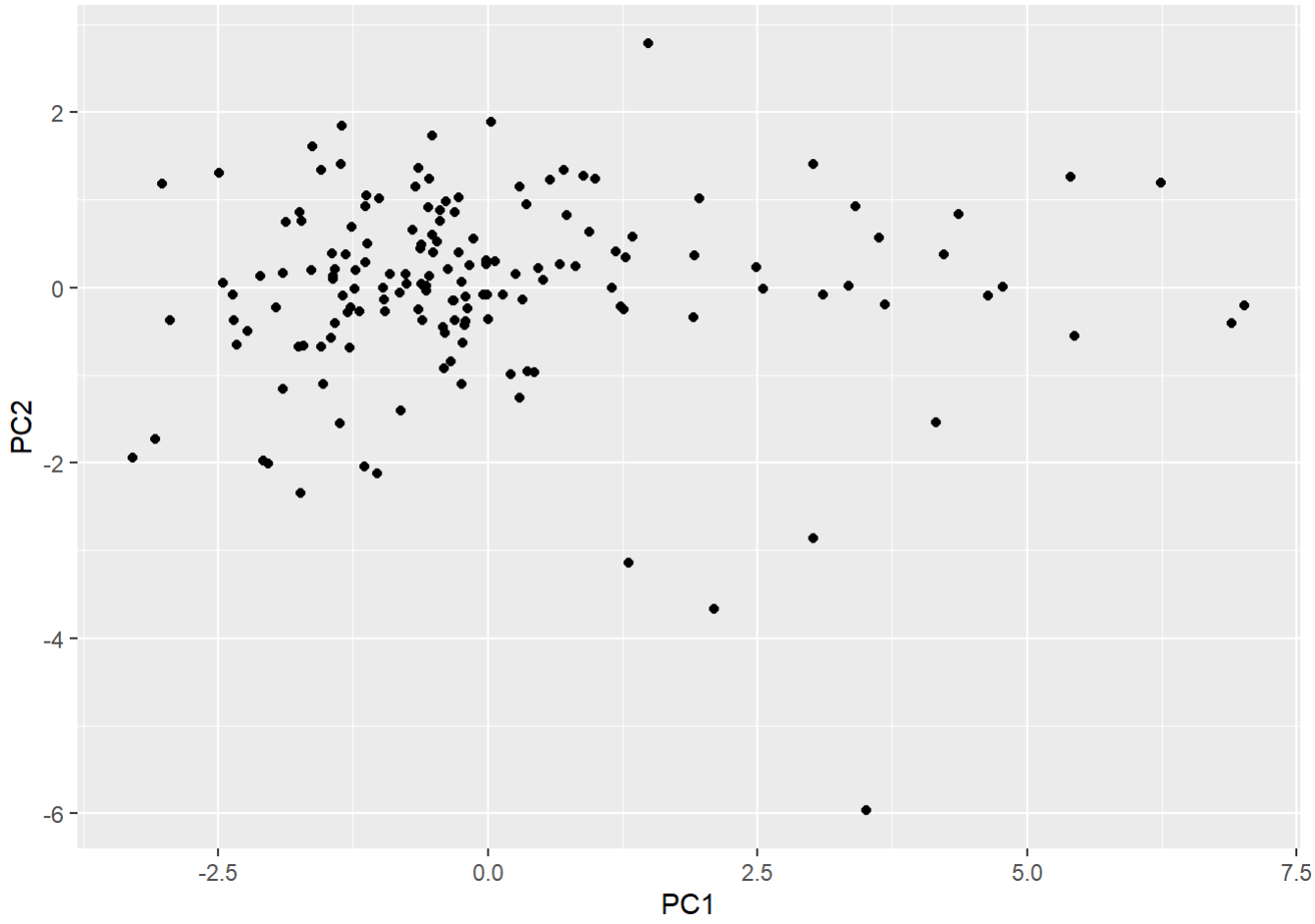
```
PCA_Data = IncomeInequal_Suicide %>%
  select(-County, -State)

PCA = PCA_Data %>%
  scale() %>%
  prcomp()
```

Visualization

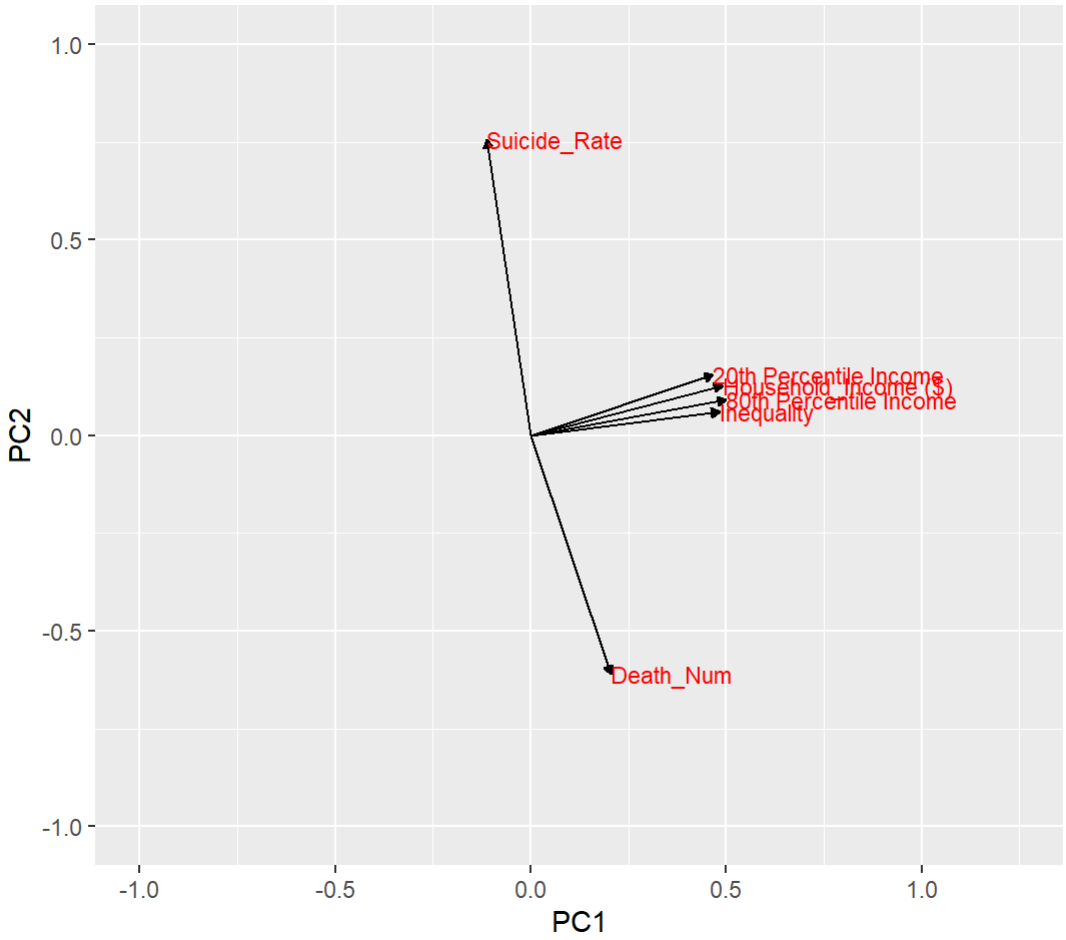
```
pca_data = as.data.frame(PCA$x)
head(pca_data)
```

```
##          PC1       PC2       PC3        PC4           PC5           PC6
## 1 -1.1399841 -2.038637 1.5195548  0.5026180  0.1322478557 -3.330669e-16
## 2 -2.0831915 -1.971795 1.4481396  0.2875760 -0.2128280825  1.110223e-16
## 3 -1.7354495 -2.349468 0.6519918  0.2360195  0.0006767287  4.996004e-16
## 4  0.2934942 -1.256229 1.5650551 -0.0229835 -0.0452493117 -5.551115e-16
## 5 -2.0363389 -2.009627 0.8176275  0.2172927  0.0413022490  4.440892e-16
## 6 -1.3707236 -1.552288 1.2973920  0.8825435  0.0640667281 -3.885781e-16
```

```
ggplot(pca_data, aes(x = PC1, y = PC2)) +
  geom_point()
```




```
rotation_data <- data.frame(
  PCA$rotation,
  variable = row.names(PCA$rotation))

arrow_style <- arrow(length = unit(0.05, "inches"),
                     type = "closed")

ggplot(rotation_data) +
  geom_segment(aes(xend = PC1, yend = PC2), x = 0, y = 0, arrow = arrow_style) +
  geom_text(aes(x = PC1, y = PC2, label = variable), hjust = 0, size = 3, color = "red") +
  xlim(-1., 1.25) +
  ylim(-1., 1.) +
  coord_fixed()
```

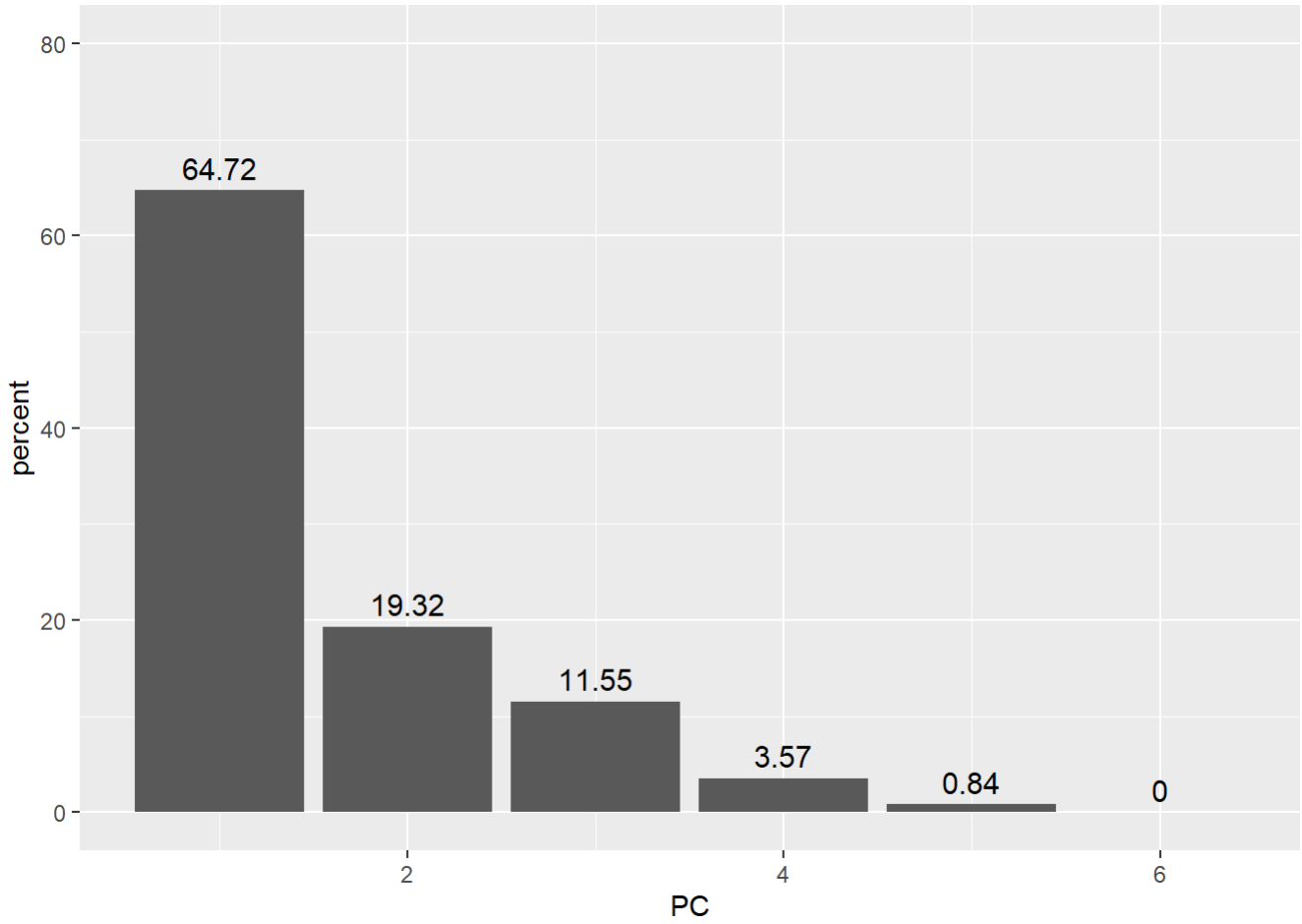


#Percent

```
#percent equation
percent <- 100* (PCA$sdev^2 / sum(PCA$sdev^2))
percent
```

```
## [1] 6.471620e+01 1.932266e+01 1.154867e+01 3.570294e+00 8.421753e-01
## [6] 4.403888e-30
```

```
perc_data <- data.frame(percent = percent,
                        PC = 1:length(percent))

ggplot(perc_data, aes(x = PC, y = percent)) +
  geom_col() +
  geom_text(aes(label = round(percent, 2)), size = 4, vjust = -0.5) +
  ylim(0, 80)
```



##Reference *https://www.countyhealthrankings.org/app/texas/2020/measure/factors/63/data (https://www.countyhealthrankings.org/app/texas/2020/measure/factors/63/data) https://www.countyhealthrankings.org/app/texas/2020/measure/factors/161/data (https://www.countyhealthrankings.org/app/texas/2020/measure/factors/161/data) https://www.countyhealthrankings.org/app/texas/2020/measure/factors/44/map (https://www.countyhealthrankings.org/app/texas/2020/measure/factors/44/map)*