Does Unemployment lead to Suicide?

Julia Capelli

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

1.Introduction

Introduction: I choose to measure the correlation between Unemployment and the Suicide Rate in New York State. The impact of the COVID-19 pandemic has had a serious effect on the workforce and millions are finding themselves without jobs. Unemployment has been shown to be linked to many different unhealthy health-related behaviors, such as damaging mental and emotional health. Since I am from the State of New York I wanted to explore the link between unemployment and suicide rates. For my project, I have decided to grab my data from www.Countyhealthrankings.org. Unemployment is the percentage of someone who is 16 years or older that does not have a job but is seeking work. The Unemployment data was acquired from the Local Area Unemployment Statistics (LAUS) program of the Bureau of Labor Statistics. There are many different methods that were used to acquire this data including, 1) a signal-plus-noise time-series model for states, a building block approach, and disaggregation procedures. The data set acquired for Unemployment Rates used data from 2002 to 2018. It consisted of 5 Variables which are County, #Unemployed, Labor Force, County Value, and Z-score. The county value is stated as a percentage which was calculated by taking the total number of people in the civilian labor force who are ages 16 and older within a certain county and divided it by the total number of people in the civilian labor force who are ages 16 and then multiplying it by 100. It's important to recognize that these values are obtained by only using data from the labor force and not the total population since not everyone is actively seeking work. Calculating the county value is important because not all County has the same population, so it corrects for population variation. There were some limitations in the data which are not including people who are not in the workforce because they were discouraged or can't find work that suits their Salary preference. Z-score is an important indicator for an indication of how far each county *is away- either above or below- from the mean unemployment rate. Calculating a Z-score is important because it is a method that can be used to compare different Counties to New York States' average unemployment. The data for New York State County Suicide rates were taken from countyheathrankings.org but was provided by NCHS and National Vital Statistic Systems. They are drawn from the vital registration systems which are operated by each county's jurisdictions who are responsible for documenting all life events. The New York suicidal data used in this project was measured from 2014-2018. In this project, I choose to use Suicide as an important marker for mental health for the county's population. Suicide is generally characterized as death from self-inflicted injury. This data set contained 4 variables used which are County, #Deaths, County Value, and Crude Rate per 100,000 population. The Crude Suicide rate is the number of deaths occurring among the population of a given geographical area during a given year. Crude Rate is an extremely important variable because it takes the total events that happened over the year and is divided by the average number of people who are at risk during a specific period of time. Crude Rates allow counties with different populations to compare their suicide data. The County Value was calculated by taking the number of suicidal deaths in a county and dividing it by the total aggregate population of a county. One thing that I found interesting was that all suicide deaths were counted in the county of residence, not the county in which the suicide occurred. This is Tidy Data because each variable is its own column and each observation is its own row.*

2. Input the Data Set

```
library(readxl)
New_York_Suicide_Rates_<- read_excel("New York Suicide Rates .xlsx")
NY_Unemployment_Rates_ <-read_excel("NY Unemployment Rates .xlsx")
```

3. Join the Datasets

```
library(dplyr)
# Apply full join on dataset x with y,  with the key variable
NY_DATA <- New_York_Suicide_Rates_

NY_DATA <- NY_DATA %>%
  full_join(NY_Unemployment_Rates_, by = "County")
```

*For this project I decided to merge both of the Data sets together using Full Join. Both Data sets had the same number of rows, and they only had one variable that was the same. This clean Data set contained 61 observations and 5 8 variables Full join was able to successfully join all of the new records into a new table. Since there were no cases with NA values and both data sets contained the same number of observations none of the cases were dropped *

4. Filter

```
library(dplyr)
#Find the mean of the Suicide County Value
mean(NY_DATA$`Suicide County Value`)
```

```
## [1] 11.57377
```

```
#Find the mean of the Unemployment County Value
mean(NY_DATA$`Unemployment County Value`)*100
```

```
## [1] 4.452459
```

```
# Filter Multiple Criteria. Find the Value of the Counties that have both unem
NY_DATA %>%
filter(`Suicide County Value`>=11.57377, `Unemployment County Value`>=0.044524
```

```
## # A tibble: 19 x 8
##    County `# Suicide Deat… `Suicide County… `Crude Rate` `# Unemployed`
##    <chr>              <dbl>            <dbl>        <dbl>          <dbl>
##  1 Alleg…                36               15           15           1096
##  2 Broome               119               12           12           4115
##  3 Catta…                63               17           16           1820
##  4 Chaut…                91               14           14           2766
##  5 Chemu…                52               12           12           1649
##  6 Chena…                33               12           14           1047
##  7 Cortl…                29               12           12           1165
##  8 Delaw…                41               16           18            928
##  9 Fulton                38               14           14           1166
## 10 Greene                42               17           18            935
## 11 Herki…                45               14           14           1374
## 12 Jeffe…                70               12           12           2510
## 13 Lewis                 27               20           20            647
## 14 Niaga…               159               14           15           5197
## 15 Oswego                93               15           16           2892
## 16 Schoh…                23                15             15              712
## 17 Schuy…                17                20             19              423
## 18 Steub…                71                14             15             2106
## 19 Wyomi…                34                14             17              838
## # … with 3 more variables: `Labor Force` <dbl>, `Unemployment County
## #   Value` <dbl>, `Z-Score` <dbl>
```

Filter was used in this section to first calculate the mean for both the variables "Suicide County Value" and "Unemployment County Value". The mean for "Suicide County Value" was 11.57377 indivudals and the mean for "Unemployment County Value" was 4.452459 individuals. The means were then used to determine which counties had both high suicide and unemployment rates. There were a total of 19 Counties out of 61 that had both high "Suicide County Value" and "Unemployment County Value" rates. The top five counties were Allegany, Broome, Cattaraugus, Chemung, and Chenango. This indicates that only 31% of the counties have high values for both suicide and unemployment

5. Arrange Counties from High to Low

```
#Arrange Using Suicide County Value
NY_DATA %>%
  arrange(desc(NY_DATA$'Suicide County Value'))
```

```
## # A tibble: 61 x 8
##    County `# Suicide Deat… `Suicide County… `Crude Rate` `# Unemployed`
##    <chr>             <dbl>            <dbl>        <dbl>          <dbl>
##  1 Lewis                27               20           20            647
##  2 Schuy…               17               20           19            423
##  3 Catta…               63               17           16           1820
##  4 Greene               42               17           18            935
##  5 Delaw…               41               16           18            928
##  6 Alleg…               36               15           15           1096
##  7 Colum…               51               15           17           1043
##  8 Oswego               93               15           16           2892
##  9 Schoh…               23               15           15            712
## 10 Chaut…               91               14           14           2766
## # … with 51 more rows, and 3 more variables: `Labor Force` <dbl>, `Unemploy
## #   County Value` <dbl>, `Z-Score` <dbl>
```

```
#Arrange Using Unemployment County Value
NY_DATA %>%
  arrange(desc(NY_DATA$'Unemployment County Value'))
```

```
## # A tibble: 61 x 8
##    County `# Suicide Deat… `Suicide County… `Crude Rate` `# Unemployed`
##    <chr>              <dbl>            <dbl>        <dbl>          <dbl>
##  1 Bronx                396                5            5          34319
##  2 Alleg…                36               15           15           1096
##  3 Jeffe…                70               12           12           2510
##  4 St. L…                54                9           10           2448
##  5 Lewis                 27               20           20            647
##  6 Oswego                93               15           16           2892
##  7 Catta…                63               17           16           1820
##  8 Frank…                31               11           12           1014
##  9 Niaga…               159               14           15           5197
## 10 Cortl…                29               12           12           1165
## # … with 51 more rows, and 3 more variables: `Labor Force` <dbl>, `Unemploy
## #   County Value` <dbl>, `Z-Score` <dbl>
```

```
# Sort by County, Suicide County Value and Unemployment County Value. This is
NY_DATA %>%
  arrange((NY_DATA$'Suicide County Value'), (NY_DATA$'Unemployment County Valu
```

```
## # A tibble: 61 x 8
##    County `# Suicide Deat… `Suicide County… `Crude Rate` `# Unemployed`
##    <chr>              <dbl>            <dbl>        <dbl>          <dbl>
##  1 Kings                692                5            5          51220
##  2 Bronx                396                5            5          34319
##  3 Queens               738                6            6          41766
##  4 Nassau               471                7            7          25027
##  5 New Y…               668                7            8          33750
##  6 Rockl…               111                7            7           5775
##  7 Westc…               350                7            7          18828
##  8 Richm…               169                7            7           9112
##  9 Putnam                43                8            9           1886
## 10 Orange               159                8            8           7155
## # … with 51 more rows, and 3 more variables: `Labor Force` <dbl>, `Unemploy
## #   County Value` <dbl>, `Z-Score` <dbl>
```

```
#Arrange Using # of Suicide Deaths
NY_DATA %>%
  arrange(desc(NY_DATA$'# Suicide Deaths'))
```

```
## # A tibble: 61 x 8
##    County `# Suicide Deat… `Suicide County… `Crude Rate` `# Unemployed`
##    <chr>              <dbl>            <dbl>      <dbl>          <dbl>
##  1 Queens              738                6          6          41766
##  2 Kings               692                5          5          51220
##  3 Suffo…              678                9          9          29952
##  4 New Y…              668                7          8          33750
##  5 Erie                530               11         11          19574
##  6 Nassau              471                7          7          25027
##  7 Bronx               396                5          5          34319
##  8 Monroe              367               10         10          15459
##  9 Westc…              350                7          7          18828
## 10 Onond…              242               10         10           8826
## # … with 51 more rows, and 3 more variables: `Labor Force` <dbl>, `Unemploy
## #   County Value` <dbl>, `Z-Score` <dbl>
```

```
#Arrange Using # Unemployed
NY_DATA %>%
  arrange(desc(NY_DATA$'# Unemployed'))
```

```
## # A tibble: 61 x 8
##    County `# Suicide Deat… `Suicide County… `Crude Rate` `# Unemployed`
##    <chr>              <dbl>            <dbl>      <dbl>          <dbl>
##  1 Kings               692                5          5          51220
##  2 Queens              738                6          6          41766
##  3 Bronx               396                5          5          34319
##  4 New Y…              668                7          8          33750
##  5 Suffo…              678                9          9          29952
##  6 Nassau              471                7          7          25027
##  7 Erie                530               11         11          19574
##  8 Westc…              350                7          7          18828
##  9 Monroe              367               10         10          15459
## 10 Richm…              169                7          7           9112
## # … with 51 more rows, and 3 more variables: `Labor Force` <dbl>, `Unemploy
## #   County Value` <dbl>, `Z-Score` <dbl>
```

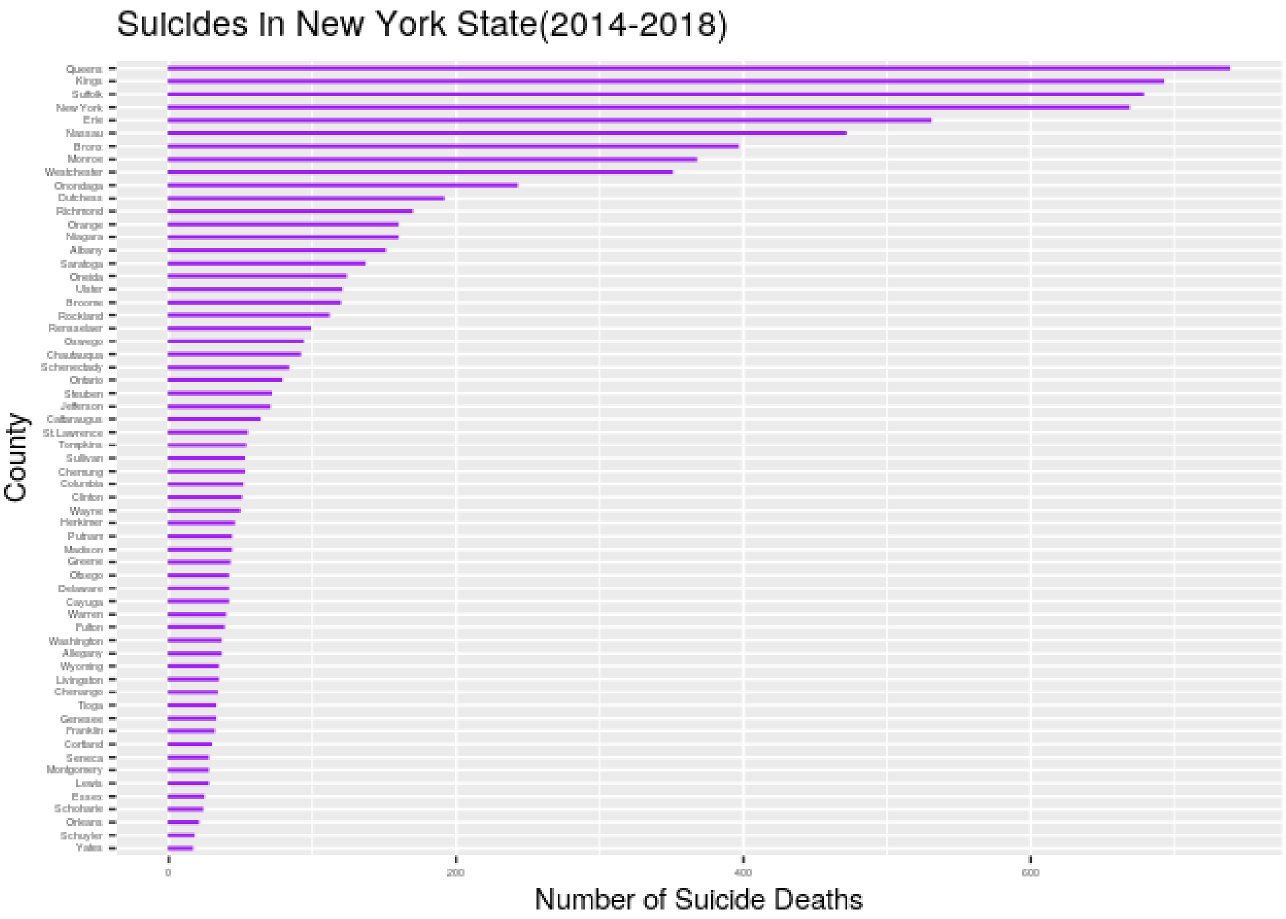
```
# Sort by County, #Suicide Deaths and # Unemployed. This is from greatest to l
NY_DATA %>%
  arrange(desc(NY_DATA$'# Suicide Deaths'), desc(NY_DATA$'# Unemployed'))
```

```
## # A tibble: 61 x 8
##    County `# Suicide Deat… `Suicide County… `Crude Rate` `# Unemployed`
##    <chr>              <dbl>            <dbl>      <dbl>          <dbl>
##  1 Queens              738                6          6          41766
##  2 Kings               692                5          5          51220
##  3 Suffo…              678                9          9          29952
##  4 New Y…              668                7          8          33750
##  5 Erie                530               11          11          19574
##  6 Nassau              471                7           7          25027
##  7 Bronx               396                5           5          34319
##  8 Monroe              367               10          10          15459
##  9 Westc…              350                7           7          18828
## 10 Onond…              242               10          10           8826
## # … with 51 more rows, and 3 more variables: `Labor Force` <dbl>, `Unemploy
## #   County Value` <dbl>, `Z-Score` <dbl>
```

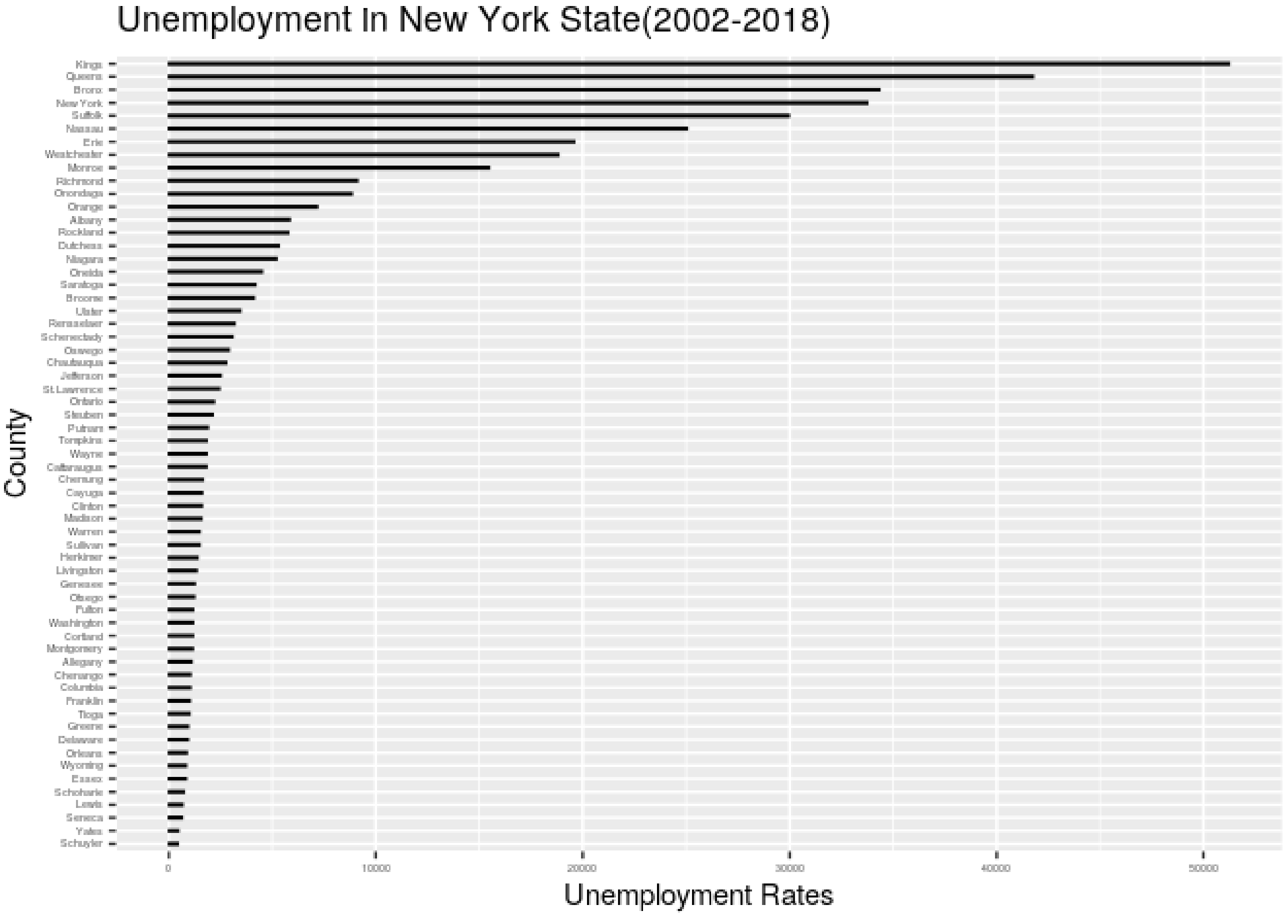
The function arrange was used to see which counties had the highest, “# Suicide Deaths”,“# Unemployed”,“Suicide County Value”, and “Unemployment County Value”. The results show that Queens and Kings County both have the highest number of people who are unemployed and they have the highest suicide rate. However, Lewis County has the highest Suicide County Value and the Bronx has the highest unemployment county value. These results are extremley interesting because there seems to a huge different between the county value and the number of people who are facing unemployment and suicidel in each county.

6. Distinguish Differences among the different Variables

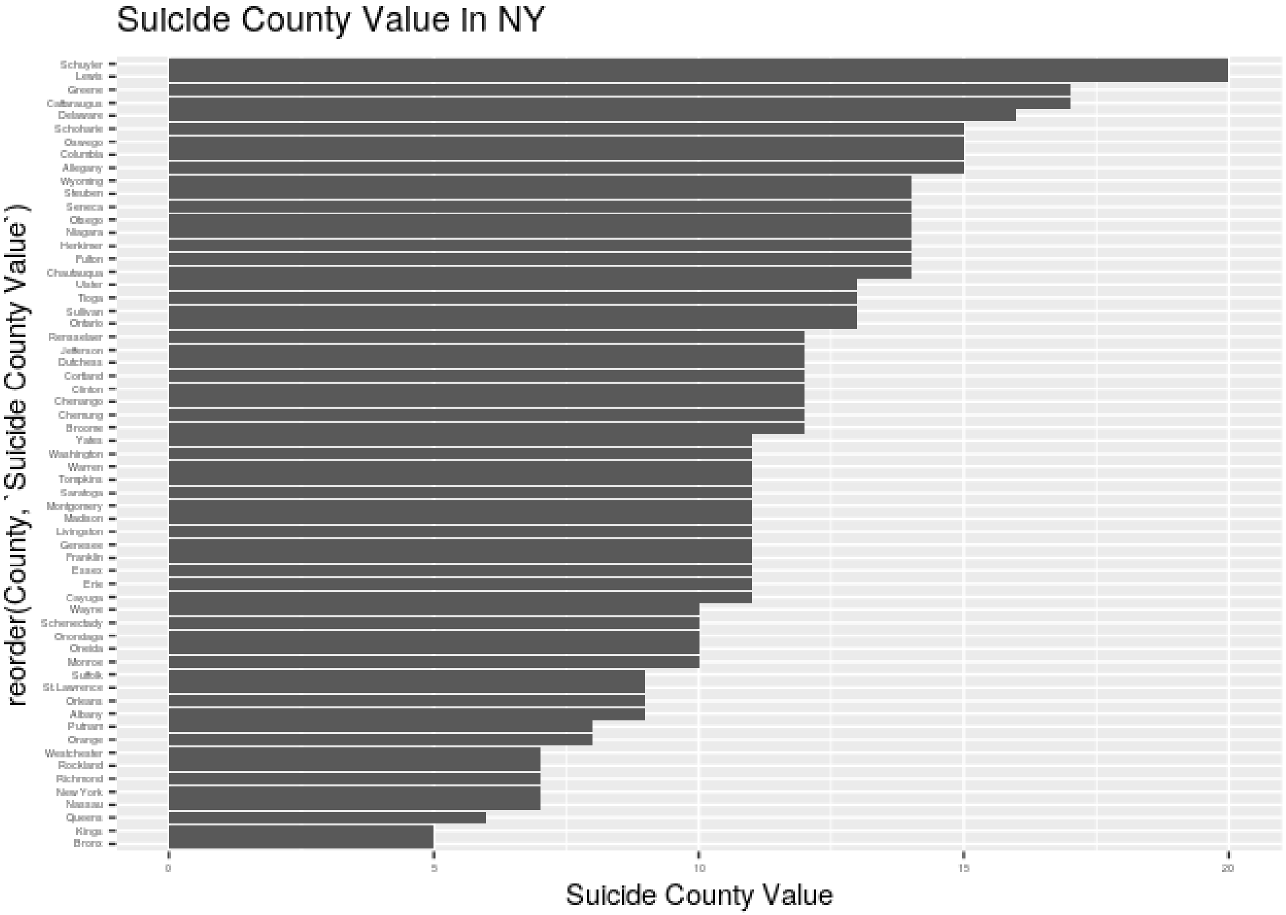
```
library(ggplot2)
# Filter by County, #Unemployed and #Suicide Deaths, arrange by descending Cou
chain1 <- NY_DATA %>%
  select(1,2,5) %>%
  arrange(desc(County))

# Represent the first ten rows of the data with ggplot and reorder the cities
chain1 %>%
  slice(1:61) %>%
  ggplot(aes(x=reorder(County,`# Suicide Deaths`), y=`# Suicide Deaths`))+
  ggtitle("Suicides in New York State(2014-2018)")+
  geom_bar(stat="identity", width = 0.1, color="purple") + coord_flip()+
   theme(axis.text = element_text(size = 4))+
  xlab("County")+
  ylab("Number of Suicide Deaths")
```



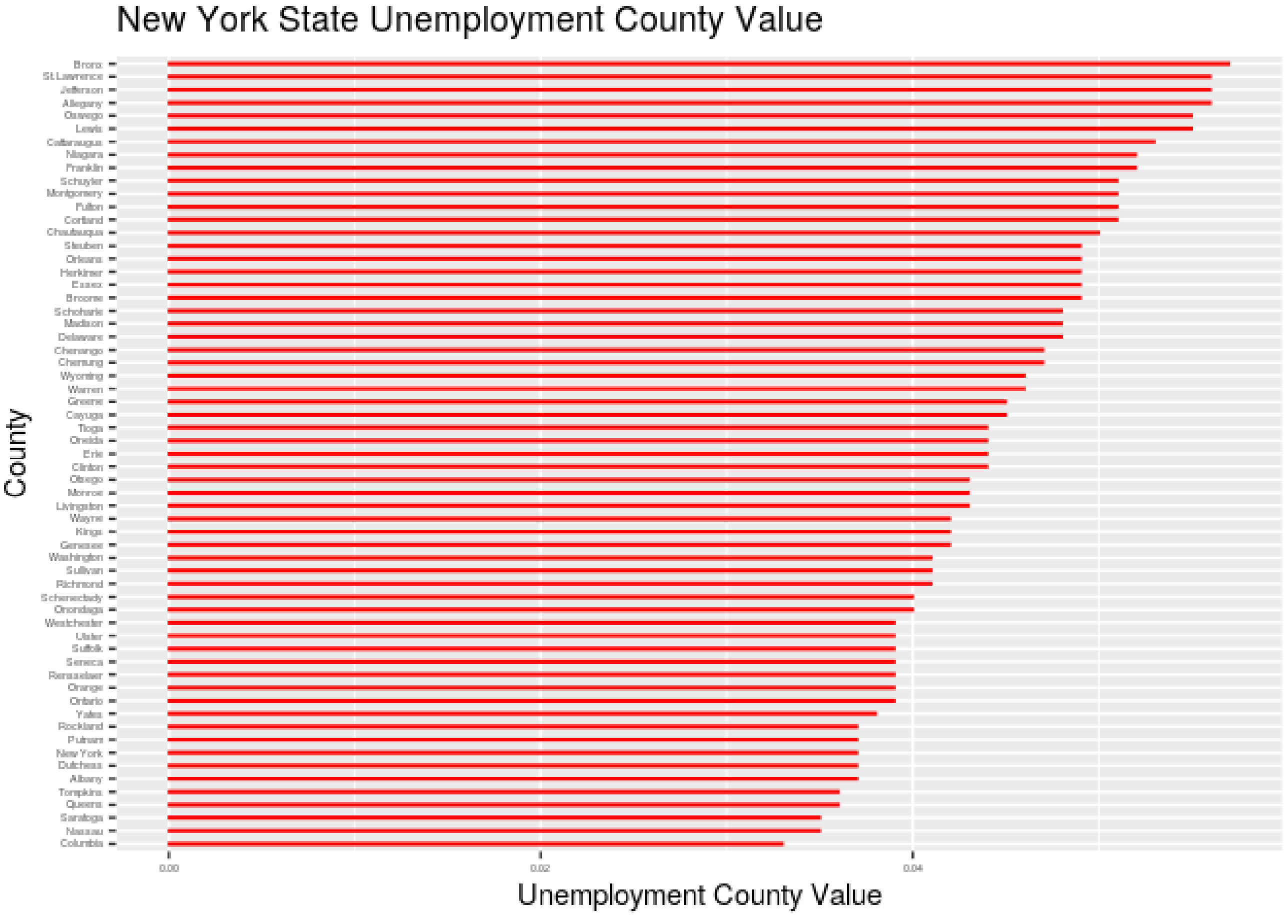
```
# Represent the first ten rows of the data with ggplot and reorder the cities
chain1 %>%
  slice(1:61) %>%
  ggplot(aes(x=reorder(County,`# Unemployed`), y=`# Unemployed`)) +
  ggtitle("Unemployment in New York State(2002–2018)")+
  geom_bar(stat="identity", width = 0.1, color="black") + coord_flip()+
   theme(axis.text = element_text(size = 4))+
  xlab("County")+
  ylab("Unemployment Rates")
```



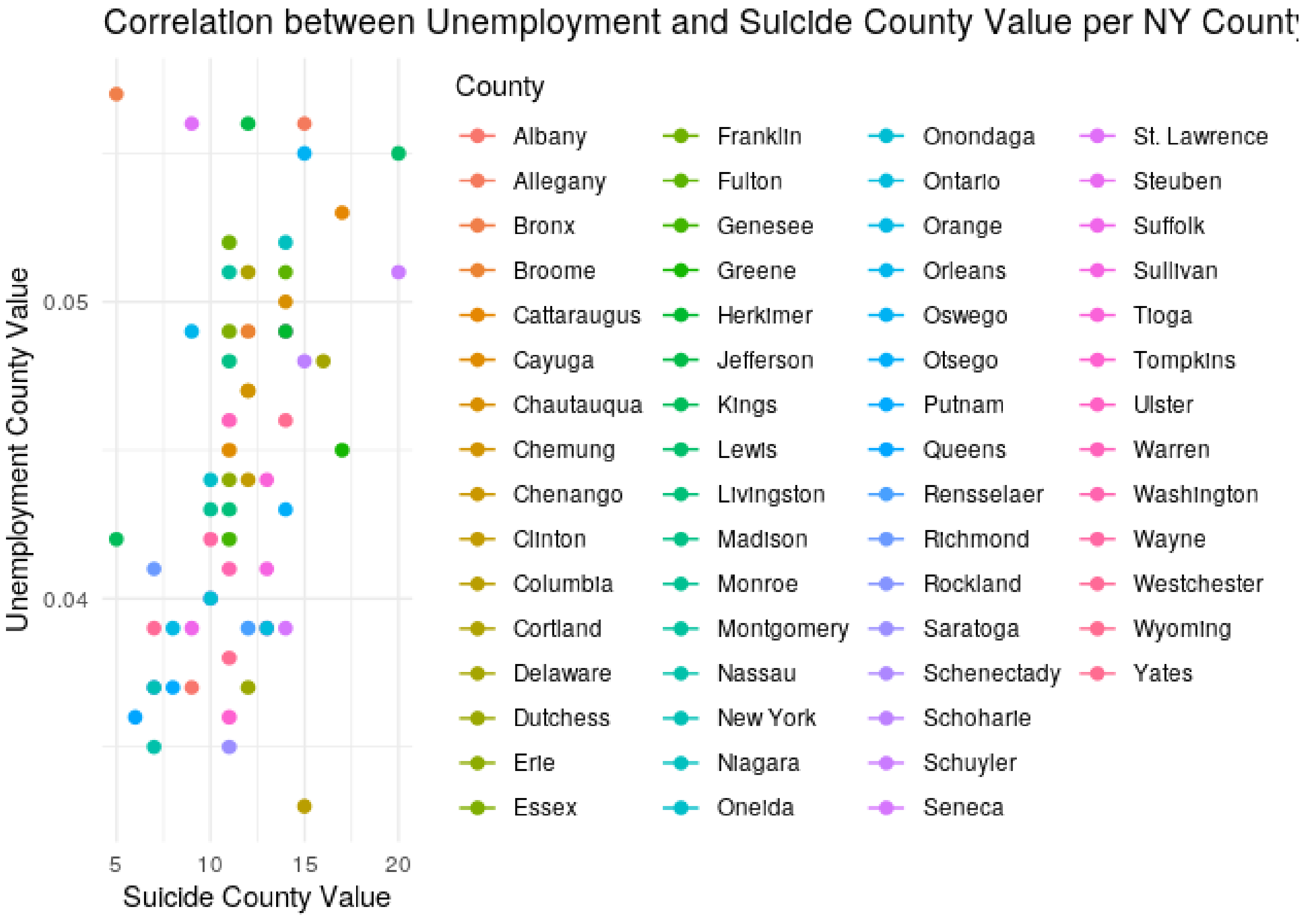
```
# Filter by County, and both Suicide and County Value
chain2 <- NY_DATA %>%
  select(1,3,7) %>%
  arrange(desc(County))

# Represent the first 61 rows by County Value of Suicide Rates
chain2 %>%
  slice(1:61) %>%
  ggplot(aes(x=reorder(County, `Suicide County Value`), y= `Suicide County Val
    theme(axis.text = element_text(size = 4))+
  ggtitle("Suicide County Value in NY")
```



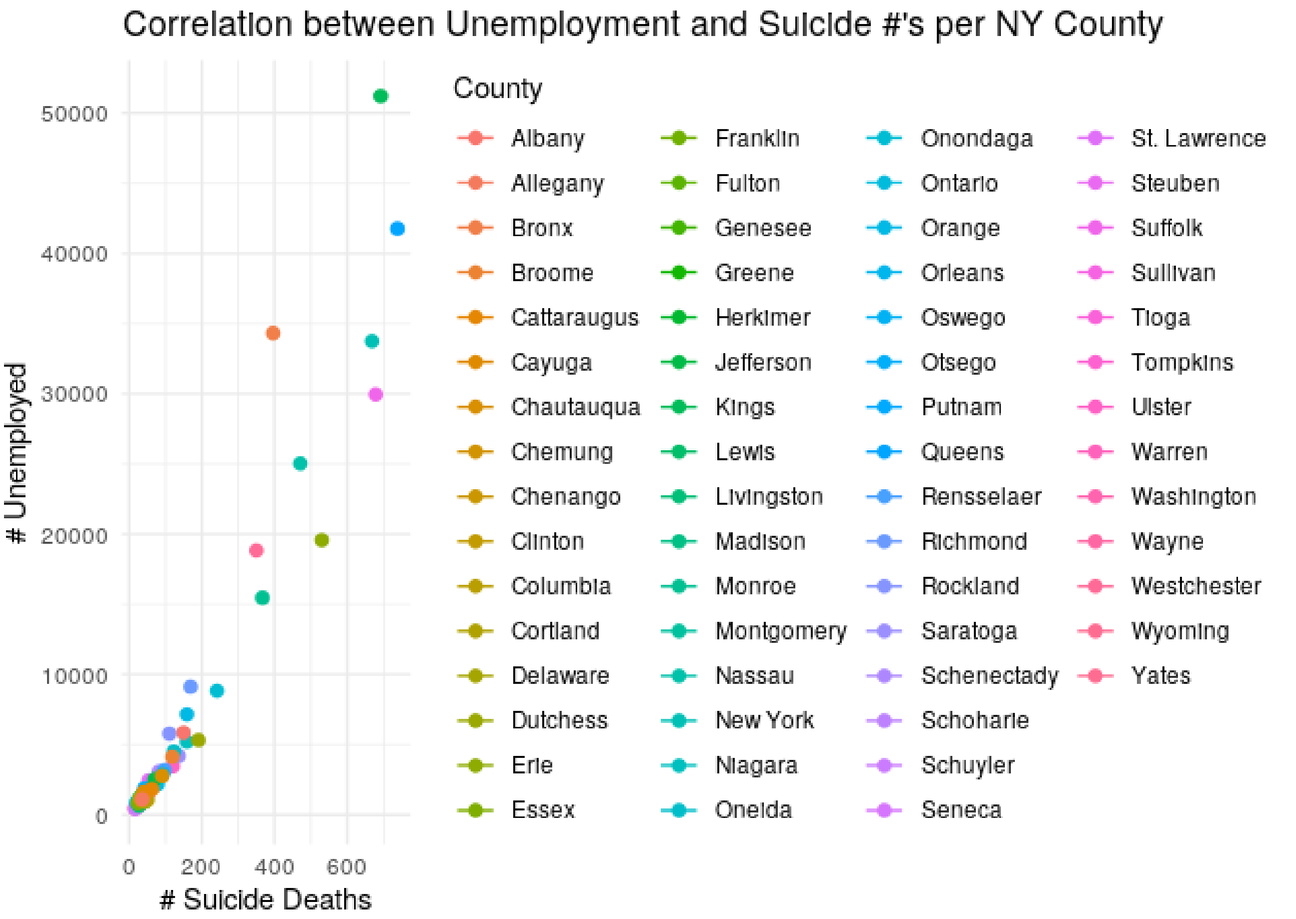
```
#Represent the first 61 rows by County Value of Unemployment
chain2 %>%
  slice(1:61) %>%
  ggplot(aes(x=reorder(County,`Unemployment County Value`), y= `Unemployment C
   theme(axis.text = element_text(size = 4))+
  ggtitle("New York State Unemployment County Value")+
  xlab("County")+
  ylab("Unemployment County Value")
```



```
#Represent the Correlation Between Unemployment and Suicide County Value
ggplot(data = chain2, aes(x = `Suicide County Value`, y = `Unemployment County
theme_minimal()+
  ggtitle("Correlation between Unemployment and Suicide County Value per NY Co
```



```
#Represent the Correlation Between Unemployment and Suicide Numbers
ggplot(data = chain1, aes(x = `# Suicide Deaths`, y = `# Unemployed`, color =
theme_minimal()+
  ggtitle("Correlation between Unemployment and Suicide #'s per NY County ")
```

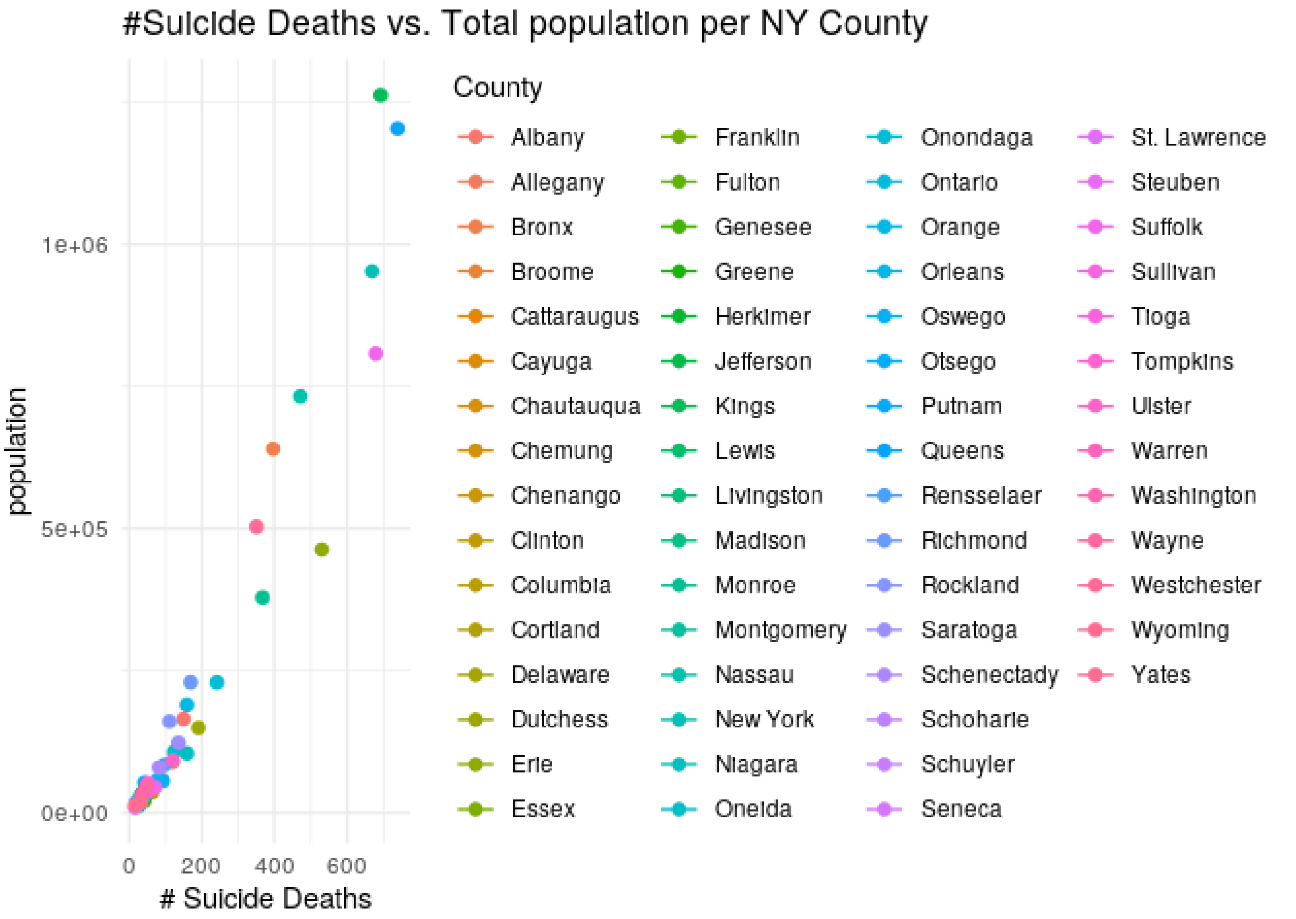


In order to further investigate the variation in County values and number of individuals whom are facing unemployment and suicide I decided to use the arrange function to make a chart to better visualize it. According to the “\#Unemployed” and “\#Suicides” graphs, Bronx and Kings County have the highest Sucides and Unemployment numbers. However, according the the County Value Graphs there is no pattern between County unemployment and Suicide since they have different state that rank at the top. Furthermore, the correlation graphs indicate that “\# Suicide Deaths” is directly correlated “\# Unemployed”. As Suicde deaths increase so do the amounto f unemployes individuals. The Correalation graphs for County value shows no relationship between “Suicide County Value”, and “Unemployment County Value”.

7. Mutate to create a new variable

```
#create new Variable in order to figure out population per County
NY_DATA_pop <- NY_DATA
NY_DATA_pop <- NY_DATA_pop %>%
  mutate(population=NY_DATA_pop$`Labor Force`+ NY_DATA_pop$`# Unemployed`)
```

```
# Use the default theme_minimal()
ggplot(data = NY_DATA_pop, aes(x = `# Suicide Deaths`, y = `population`, color
  geom_point(size=2) + geom_line() +
theme_minimal()+
  ggtitle("#Suicide Deaths vs. Total population per NY County")
```



This graph indicates that there is a positive relationship between Suicide rates and larger population sizes. While it is hard to interpret which Counties have both high population and suicide rates there is clearly a trend that goes in a positive direction, so as the y-value increases so does the x-value

8. Summarize

```
library(kableExtra)
library(tidyverse)
library(dplyr)
library(ggplot2)
library(fBasics)
library(knitr)
#Create a summary table
NY_DATA_NUM <-NY_DATA

#Remove the Categorical Value
NY_DATA_NUM<-select(NY_DATA_NUM, -County)
```

```
#Create a Summary Table for NY DATA
basicStats(NY_DATA_NUM)[c("Mean", "Stdev", "Median", "Minimum", "Maximum", "nc
  kbl(caption = "Summary of New York Data") %>%
kable_classic(full_width = F, html_font = "Cambria")
```

Summary of New York Data

	X..Suicide.Deaths	Suicide.County.Value	Crude.Rate	X..Unemploye
Mean	138.1639	11.573770	12.098361	6.450656e+0
Stdev	185.0459	3.201349	3.335191	1.092102e+0
Median	52.0000	11.000000	12.000000	1.823000e+0
Minimum	16.0000	5.000000	5.000000	4.230000e+0
Maximum	738.0000	20.000000	20.000000	5.122000e+0
nobs	61.0000	61.000000	61.000000	6.100000e+0
1. Quartile	36.0000	10.000000	10.000000	1.155000e+0
3. Quartile	136.0000	14.000000	14.000000	5.197000e+0
Variance	34241.9727	10.248634	11.123497	1.192686e+0
Sum	8428.0000	706.000000	738.000000	3.934900e+0

```
 #Find the mean #of Suicide Deaths, and the mean #unemployed
NY_DATA %>%
 summarize(`# Suicide Deaths` = mean(`# Suicide Deaths`), `# Unemployed` = mea
```

```
## # A tibble: 1 x 2
##   `# Suicide Deaths` `# Unemployed`
##                <dbl>          <dbl>
## 1               138.          6451.
```

```
# Create a new variable under different conditions with case_when()
NY_DATA_Conditions <- NY_DATA
  NY_DATA_Conditions <- NY_DATA_Conditions %>%
  mutate(Rate_Suicide = case_when(NY_DATA$`# Suicide Deaths`>136 ~ "high",
                                  36<=NY_DATA$`# Suicide Deaths` & NY_DATA$`#
                                  NY_DATA$`# Suicide Deaths`<36 ~ "low"))

# Create a new variable under different conditions with case_when()
  NY_DATA_Conditions <- NY_DATA_Conditions %>%
  mutate(Rate_Unemployed = case_when(NY_DATA$`# Unemployed`>5197 ~ "high",
                                  1155 <=NY_DATA$`# Unemployed` & NY_DATA$`# U
                                  NY_DATA$`# Unemployed`<1155  ~ "low"))

#Find summaries of Sub Group of high, medium and low Unemployed numbers.
NY_DATA_Conditions %>%
  group_by(`Rate_Unemployed`) %>%
  summarize(mean_rate=mean(`# Unemployed`),mean_suicide=mean(`# Suicide Deaths
  kbl(caption = "Mean Averages basd on Unemployment Groups") %>%
kable_classic(full_width = F, html_font = "Cambria")
```

Mean Averages basd on Unemployment Gr

Rate_Unemployed	mean_rate	mean_suicide	mean_SuicideCounty	mean_crude
high	20794.1333	394.13333	9	8.20000
high	20794.1333	394.13333	5	8.20000
high	20794.1333	394.13333	12	8.20000
high	20794.1333	394.13333	11	8.20000
high	20794.1333	394.13333	5	8.20000
high	20794.1333	394.13333	10	8.20000
high	20794.1333	394.13333	7	8.20000

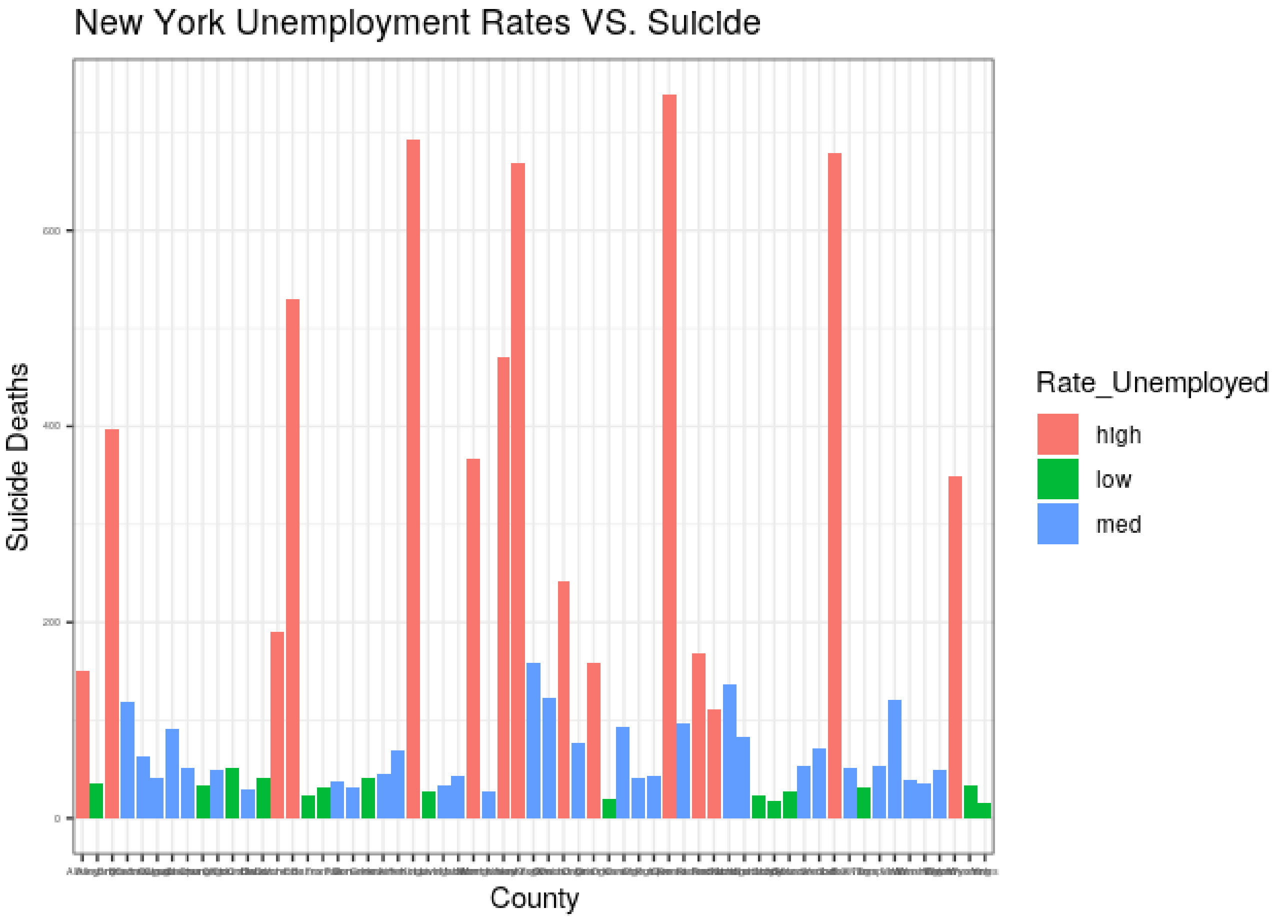
Rate_Unemployed	mean_rate	mean_suicide	mean_SuicideCounty	mean_crude
high	20794.1333	394.13333	7	8.20000
high	20794.1333	394.13333	10	8.20000
high	20794.1333	394.13333	8	8.20000
high	20794.1333	394.13333	6	8.20000
high	20794.1333	394.13333	7	8.20000
high	20794.1333	394.13333	7	8.20000
high	20794.1333	394.13333	9	8.20000
high	20794.1333	394.13333	7	8.20000
low	830.3333	30.26667	15	15.33333
low	830.3333	30.26667	12	15.33333
low	830.3333	30.26667	15	15.33333
low	830.3333	30.26667	16	15.33333
low	830.3333	30.26667	11	15.33333
low	830.3333	30.26667	11	15.33333
low	830.3333	30.26667	17	15.33333
low	830.3333	30.26667	20	15.33333
low	830.3333	30.26667	9	15.33333

Rate_Unemployed	mean_rate	mean_suicide	mean_SuicideCounty	mean_crude
low	830.3333	30.26667	13	15.33333
low	830.3333	30.26667	20	15.33333
low	830.3333	30.26667	14	15.33333
low	830.3333	30.26667	13	15.33333
low	830.3333	30.26667	14	15.33333
low	830.3333	30.26667	11	15.33333
med	2229.7742	66.51613	12	12.41935
med	2229.7742	66.51613	17	12.41935
med	2229.7742	66.51613	11	12.41935
med	2229.7742	66.51613	14	12.41935
med	2229.7742	66.51613	12	12.41935
med	2229.7742	66.51613	12	12.41935
med	2229.7742	66.51613	12	12.41935
med	2229.7742	66.51613	14	12.41935
med	2229.7742	66.51613	11	12.41935
med	2229.7742	66.51613	14	12.41935
med	2229.7742	66.51613	12	12.41935

Rate_Unemployed	mean_rate	mean_suicide	mean_SuicideCounty	mean_crude
med	2229.7742	66.51613	11	12.41935
med	2229.7742	66.51613	11	12.41935
med	2229.7742	66.51613	11	12.41935
med	2229.7742	66.51613	14	12.41935
med	2229.7742	66.51613	10	12.41935
med	2229.7742	66.51613	13	12.41935
med	2229.7742	66.51613	15	12.41935
med	2229.7742	66.51613	14	12.41935
med	2229.7742	66.51613	8	12.41935
med	2229.7742	66.51613	12	12.41935
med	2229.7742	66.51613	11	12.41935
med	2229.7742	66.51613	10	12.41935
med	2229.7742	66.51613	9	12.41935
med	2229.7742	66.51613	14	12.41935
med	2229.7742	66.51613	13	12.41935
med	2229.7742	66.51613	11	12.41935
med	2229.7742	66.51613	13	12.41935

Rate_Unemployed	mean_rate	mean_suicide	mean_SuicideCounty	mean_crude
med	2229.7742	66.51613	11	12.41935
med	2229.7742	66.51613	11	12.41935
med	2229.7742	66.51613	10	12.41935

```r
#Graph
NY_DATA_Conditions %>%
  slice(1:61)%>%
  ggplot(aes(County,`# Suicide Deaths`, fill= Rate_Unemployed)) +
  geom_bar(position = "dodge", stat = "identity") +
  theme_bw()+
  theme(axis.text = element_text(size = 4))+
  ggtitle("New York Unemployment Rates VS. Suicide")+
  xlab("County")+
  ylab("Suicide Deaths")
```



An interesting finding from the graph is that the county's with the highest suicide rates are the ones that are experiencing the highest amount of unemployment. However, there seems to be a lot of variation among the counties with medium to low unemployment rates. Additionally, the summary statistics tables show that there are a total of 8428 people who have committed suicide from 2014-2018 and 3,934,900 people who have been unemployed between the years 2002-2018. I was surprised to see that the mean for \# of Suicides is 138.1639 since I believe this is a large value.

9. Correlation Matrix/ HEAT MAP

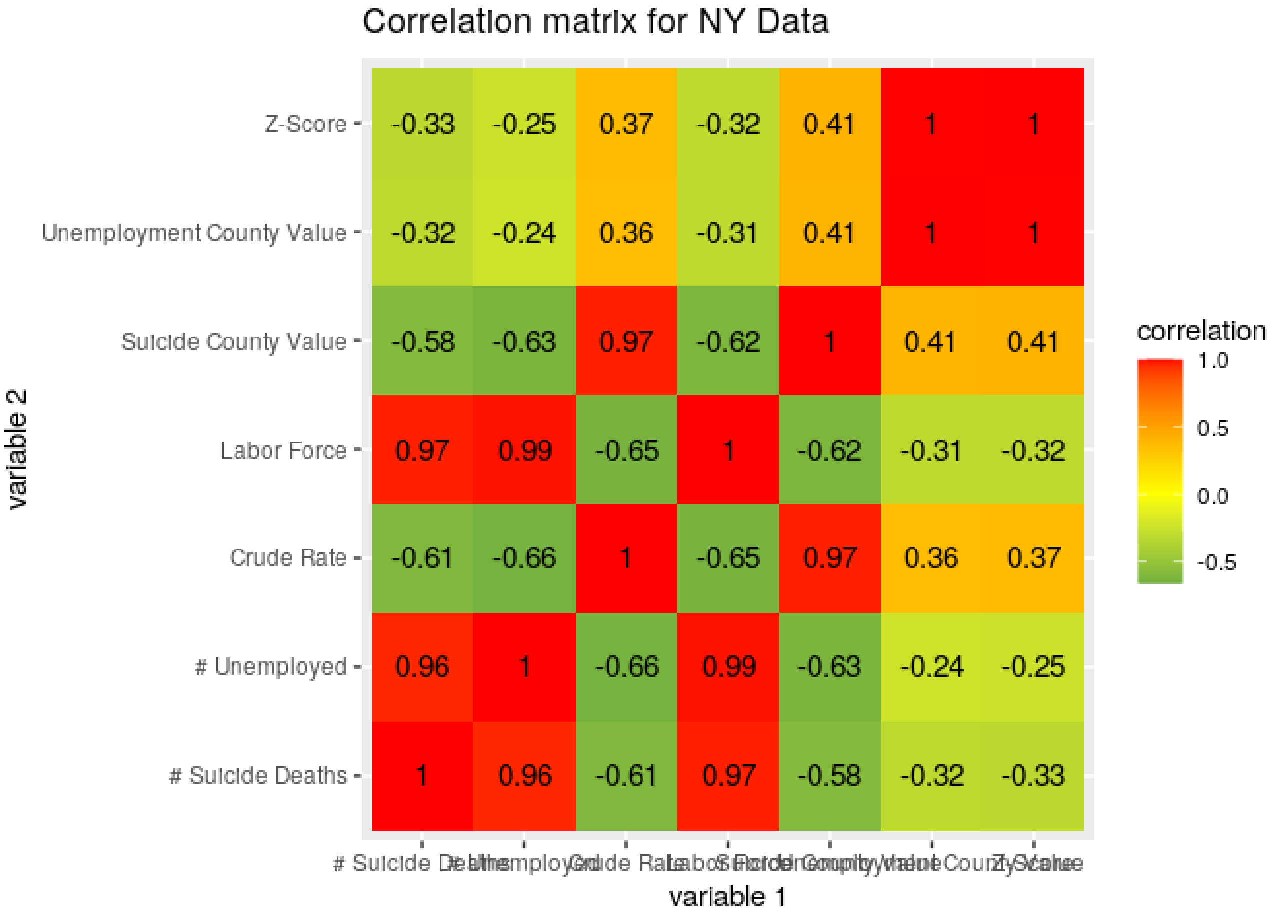
```
library("tidyr")
library("ggplot2")
# Find the correlation between two variables
cor(NY_DATA$`Suicide County Value`,NY_DATA$`Unemployment County Value`, use =
```

```
## [1] 0.407211
```

```
# Build a correlation matrix between all numeric variables
NY_DATA_NUM <- NY_DATA %>%
  select_if(is.numeric)
cor(NY_DATA_NUM, use = "pairwise.complete.obs")
```

```
##                           # Suicide Deaths Suicide County Value Crude Rate
## # Suicide Deaths                 1.0000000           -0.5806005 -0.6088066
## Suicide County Value            -0.5806005            1.0000000  0.9702336
## Crude Rate                      -0.6088066            0.9702336  1.0000000
## # Unemployed                     0.9586605           -0.6294585 -0.6565785
## Labor Force                      0.9706114           -0.6229070 -0.6477176
## Unemployment County Value       -0.3200109            0.4072110  0.3618588
## Z-Score                         -0.3272544            0.4149503  0.3679974
##                           # Unemployed Labor Force Unemployment County Valu
## # Suicide Deaths             0.9586605   0.9706114                -0.320010
## Suicide County Value        -0.6294585  -0.6229070                 0.407211
## Crude Rate                  -0.6565785  -0.6477176                 0.361858
## # Unemployed                 1.0000000   0.9887149                -0.239391
## Labor Force                  0.9887149   1.0000000                -0.311988
## Unemployment County Value   -0.2393913  -0.3119889                 1.000000
## Z-Score                     -0.2455637  -0.3174200                 0.999020
##                              Z-Score
## # Suicide Deaths          -0.3272544
## Suicide County Value       0.4149503
## Crude Rate                 0.3679974
## # Unemployed              -0.2455637
## Labor Force               -0.3174200
## Unemployment County Value  0.9990207
## Z-Score                    1.0000000
```

```
#We need to shape the correlation matrix as a date frame then we use manipulat
# Make it pretty using a heatmap with geom_tile!
cor(NY_DATA_NUM, use = "pairwise.complete.obs") %>%
  # Save as a data frame
  as.data.frame %>%
  # Convert row names to an explicit variable
  rownames_to_column %>%
  # Pivot so that all correlations appear in the same column
  pivot_longer(-1, names_to = "other_var", values_to = "correlation") %>%
  ggplot(aes(rowname, other_var, fill=correlation)) +
  # Heatmap with geom_tile
  geom_tile() +
  # Change the scale to make the middle appear neutra. Netural or zero is whit
  scale_fill_gradient2(low="springgreen4",mid="yellow",high="red") +
  # Overlay values
  geom_text(aes(label = round(correlation,2)), color = "black", size = 4) +
  # Give title and labels
  labs(title = "Correlation matrix for NY Data", x = "variable 1", y = "variat
```



	# Suicide Deaths	# Unemployed	Crude Rate	Labor Force	Suicide County Value	Unemployment County Value	Z-Score
Z-Score	-0.33	-0.25	0.37	-0.32	0.41	1	1
Unemployment County Value	-0.32	-0.24	0.36	-0.31	0.41	1	1
Suicide County Value	-0.58	-0.63	0.97	-0.62	1	0.41	0.41
Labor Force	0.97	0.99	-0.65	1	-0.62	-0.31	-0.32
Crude Rate	-0.61	-0.66	1	-0.65	0.97	0.36	0.37
# Unemployed	0.96	1	-0.66	0.99	-0.63	-0.24	-0.25
# Suicide Deaths	1	0.96	-0.61	0.97	-0.58	-0.32	-0.33

According to the heat map the variables “Unemployment” and “Labor Force have an extremely high correlation of 0.99. Another Variable that has a high correlation of 0.96 are the variables”\# Suicide Deaths" and “unemployed. On the other hand, the variables”Unemployment County Value" and “Suicide County Value” have an average correlation of 0.41.

10. PCA

```
# Prepare data for PCA and run PCA
pca_NY <- NY_DATA %>%
  # Remove categorical variables
  select(-County) %>%
  # Scale to 0 mean and unit variance (standardize)
  scale() %>%
  #preform PCA
  prcomp()

# Results from PCA
names(pca_NY)
```

```
## [1] "sdev"     "rotation" "center"   "scale"    "x"
```

```
# Visualize the results and get the variance on each principle component
pca_NY$sdev^2
```

```
## [1] 4.4969570085 1.6188872879 0.8091508344 0.0395285319 0.0271060643
## [6] 0.0075073458 0.0008629272
```

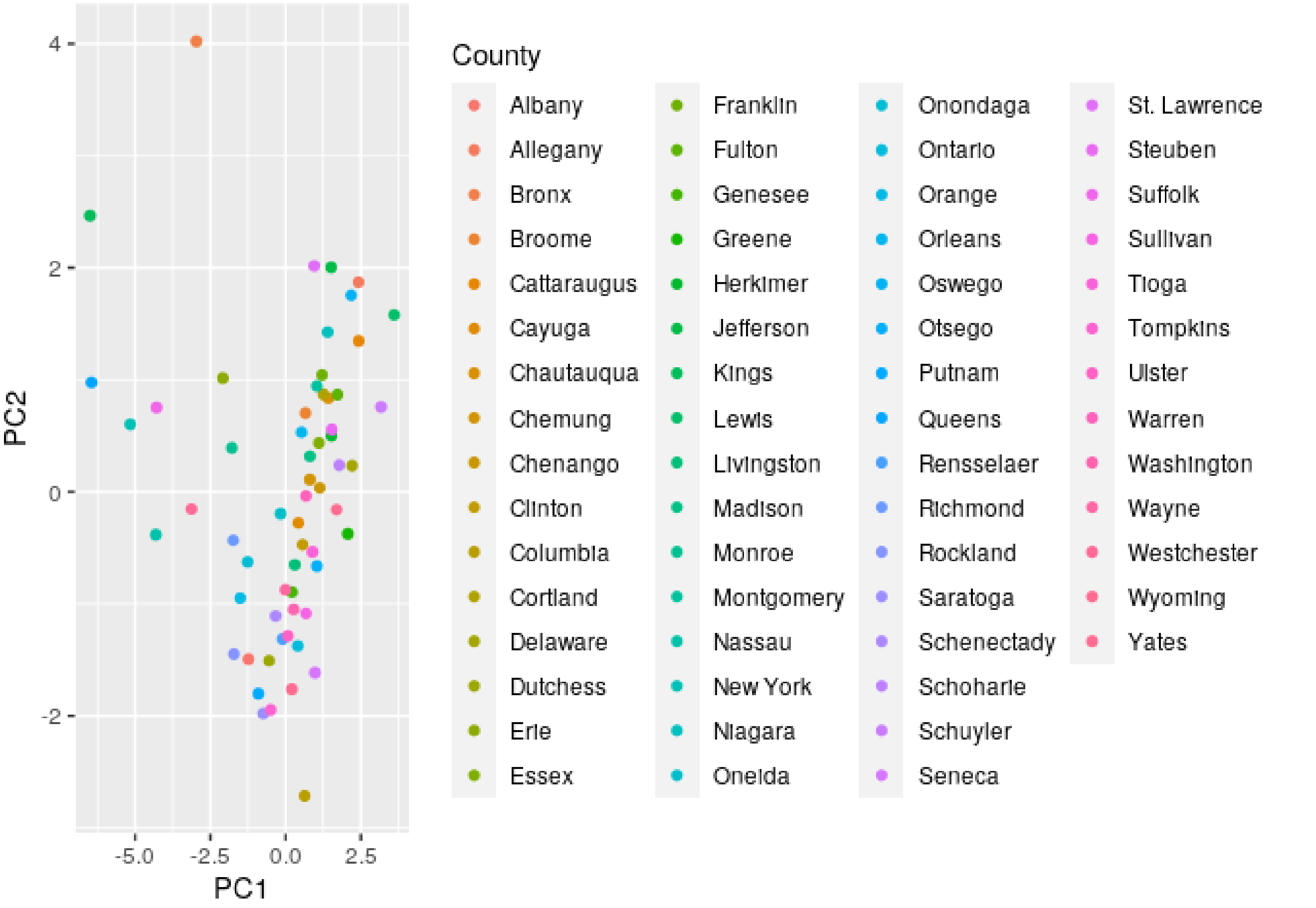
```
# Visualize the rotated data, in order to change the coordinate system. A rota
head(pca_NY$x)
```

```
##             PC1        PC2         PC3           PC4           PC5
## [1,] -1.2332060 -1.4946474  0.44783119  3.615397e-02 -0.0810691092 -0.01235
## [2,]  2.4184109  1.8715027 -0.09693041  2.859454e-02  0.0765139638  0.07469
## [3,] -2.9662768  4.0205054  1.43706027 -3.925517e-01  0.0003283254 -0.42326
## [4,]  0.6551090  0.7025776  0.31751325  1.707082e-01  0.0535972779 -0.01412
## [5,]  2.4245562  1.3464201 -0.86427843  8.292399e-02  0.3112503319  0.01479
## [6,]  0.4215321 -0.2785281  0.73580608  5.739968e-05  0.1254237772  0.01423
##              PC7
## [1,]  0.02248613
## [2,] -0.01467418
## [3,]  0.02021225
## [4,]  0.04176744
## [4,]  0.04176744
## [5,] -0.02920794
## [6,]  0.03995462
```

```
# Add the information about the different groups back into PCA data
pca_data_NY <- data.frame(pca_NY$x, County = NY_DATA$County)
head(pca_data_NY)
```

```
##          PC1        PC2         PC3           PC4           PC5         PC6
## 1 -1.2332060 -1.4946474  0.44783119  3.615397e-02 -0.0810691092 -0.01235572
## 2  2.4184109  1.8715027 -0.09693041  2.859454e-02  0.0765139638  0.07469366
## 3 -2.9662768  4.0205054  1.43706027 -3.925517e-01  0.0003283254 -0.42326335
## 4  0.6551090  0.7025776  0.31751325  1.707082e-01  0.0535972779 -0.01412238
## 5  2.4245562  1.3464201 -0.86427843  8.292399e-02  0.3112503319  0.01479398
## 6  0.4215321 -0.2785281  0.73580608  5.739968e-05  0.1254237772  0.01423776
##           PC7      County
## 1  0.02248613      Albany
## 2 -0.01467418    Allegany
## 3  0.02021225       Bronx
## 4  0.04176744      Broome
## 5 -0.02920794 Cattaraugus
## 6  0.03995462      Cayuga
```

```
# Plot the data according to the new coordinate system: PC1 and PC2
ggplot(pca_data_NY, aes(x = PC1, y = PC2, color = County)) +
  geom_point()
```

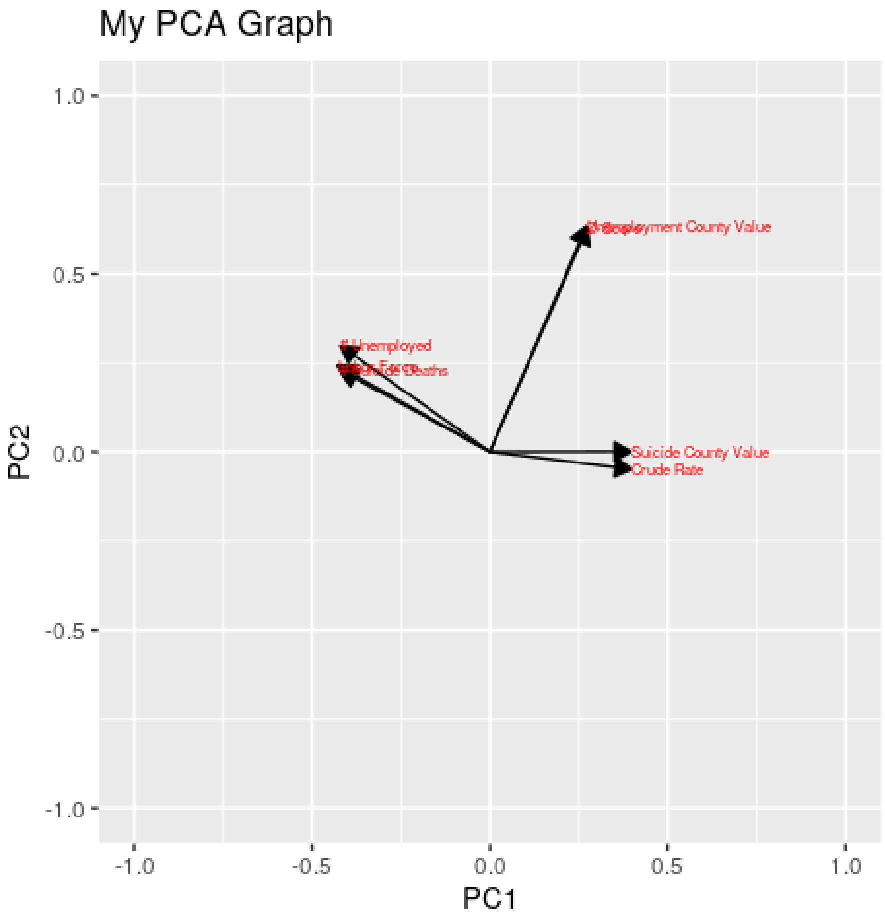


```
# Take a look at the rotation matrix
pca_NY$rotation
```

```
##                                 PC1           PC2        PC3         PC4
## # Suicide Deaths         -0.4196015  0.2286684856 -0.3492205  0.77534954
## Suicide County Value      0.3961174  0.0007729327 -0.5890773  0.12547623
## Crude Rate                0.3980663 -0.0488399317 -0.5769323 -0.20644068
## # Unemployed             -0.4200284  0.2978862055 -0.2496599 -0.47920400
## Labor Force              -0.4285557  0.2425691624 -0.2945078 -0.32988277
## Unemployment County Value  0.2709728  0.6331411150  0.1582083  0.04298416
## Z-Score                   0.2739685  0.6299850204  0.1552009  0.01281787
##                                   PC5          PC6         PC7
## # Suicide Deaths         -0.200687195 -0.085619529 -0.03028725
## Suicide County Value      0.691743724 -0.031751128  0.02857665
## Crude Rate               -0.680523216  0.005772885 -0.02332542
## # Unemployed              0.085006901 -0.659764035 -0.01893881
## Labor Force               0.086387050  0.743209907  0.04603314
## Unemployment County Value -0.058272252 -0.005444852  0.70384889
## Z-Score                   0.004344229  0.062838219 -0.70699434
```

```
#Look to see how much variation each sample accounts for
pca.var <-pca_NY$sdev^2
view(pca.var)
```

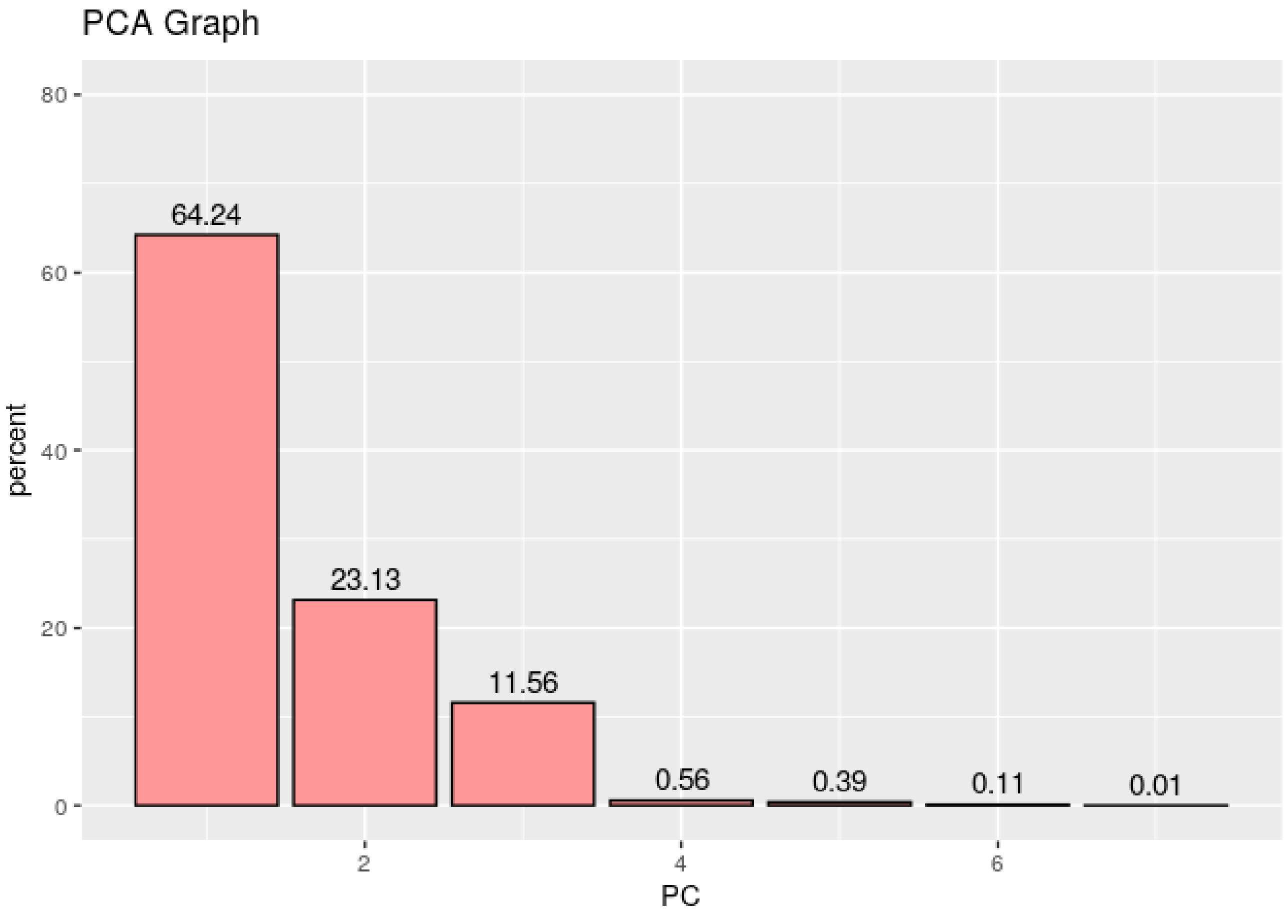
```
# Save the rotation matrix in a data frame
rotation_data <- data.frame(
  pca_NY$rotation,
  variable = row.names(pca_NY$rotation))

# Define an arrow style
arrow_style <- arrow(length = unit(0.1, "inches"), type = "closed")

# Plot the contribution of variables to PCs using geom_segment() for arrows ar
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 = 2, colc
  xlim(-1., 1) +
  ylim(-1., 1.) +
  coord_fixed()+
  ggtitle("My PCA Graph")
```



```
# Determine the percentage of variance explained by each component with sdev
percent <- 100* (pca_NY$sdev^2 / sum(pca_NY$sdev^2))
percent
```

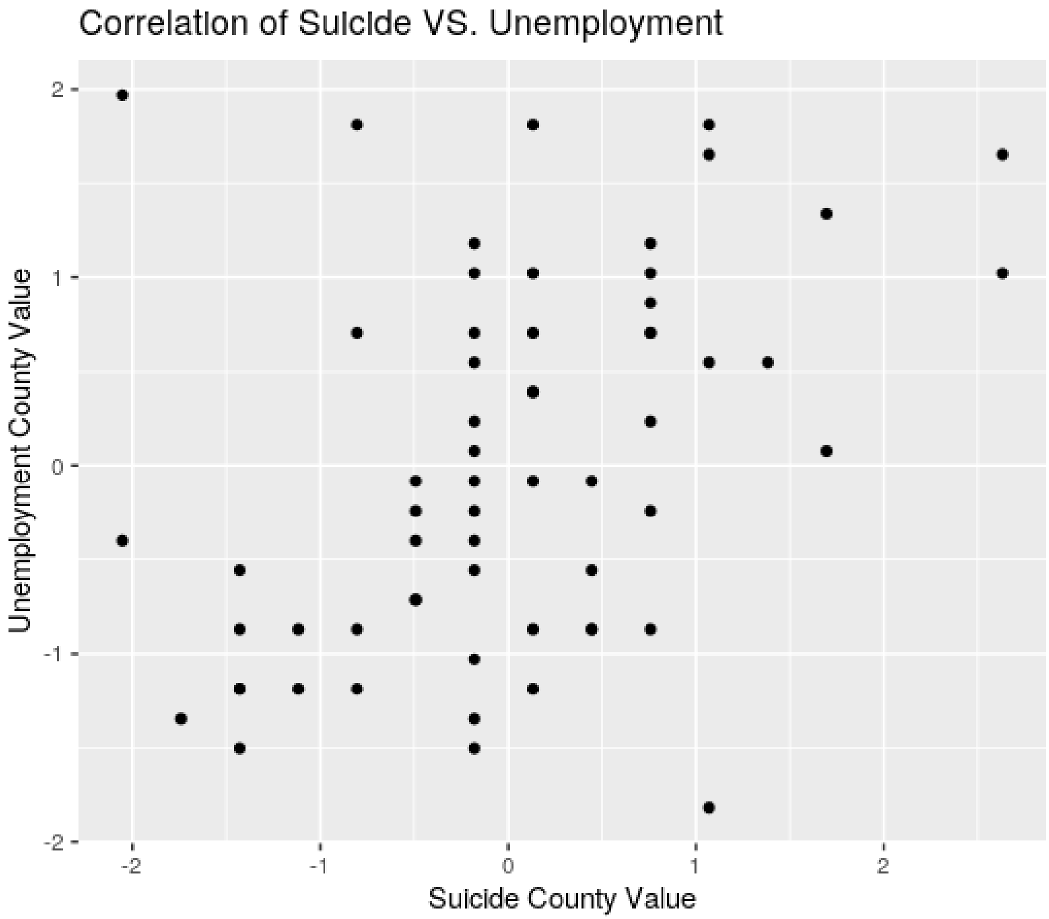
```
## [1] 64.24224298 23.12696126 11.55929763  0.56469331  0.38722949  0.1072478
## [7]  0.01232753
```

```
# Visualize the percentage of variance explained by each component
perc_data_NY <- data.frame(percent = percent, PC = 1:length(percent))
ggplot(perc_data_NY, aes(x = PC, y = percent)) +
  geom_col(fill="#FF9999", colour="black") +
  geom_text(aes(label = round(percent, 2)), size = 4, vjust = -0.5) +
  ylim(0, 80) +
ggtitle("PCA Graph")
```



\#\# 11.PCA continued

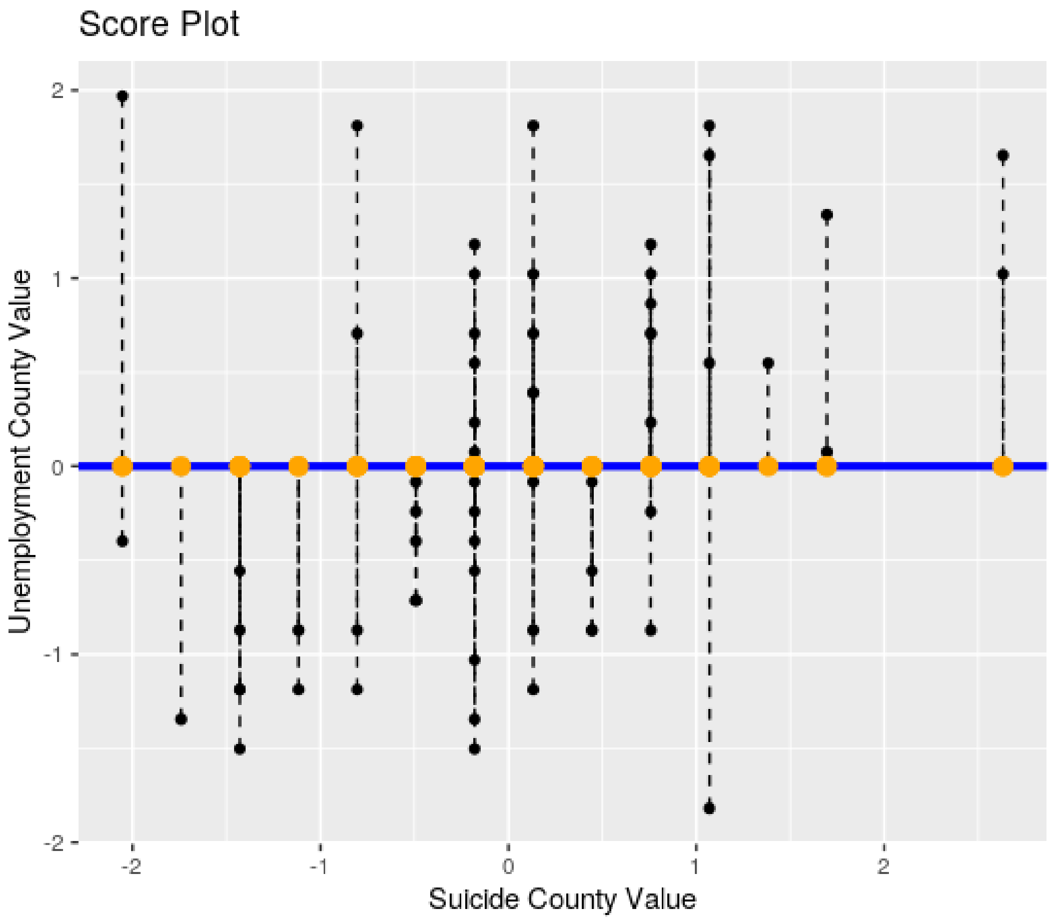
```
library(ggplot2)
library(tidyverse)
library(cluster)
NY_DATA_NUM <- NY_DATA

NY_DATA_NUM <- NY_DATA_NUM %>%
  select(-County) %>%
  scale () %>%
  as.data.frame

# Prepare the data frame with only numeric variables
NY_DATA_NUM %>%
  ggplot(aes(x =  `Suicide County Value`,  y =  `Unemployment County Value`))
  geom_point() +
  coord_fixed()+
  ggtitle("Correlation of Suicide VS. Unemployment")
```



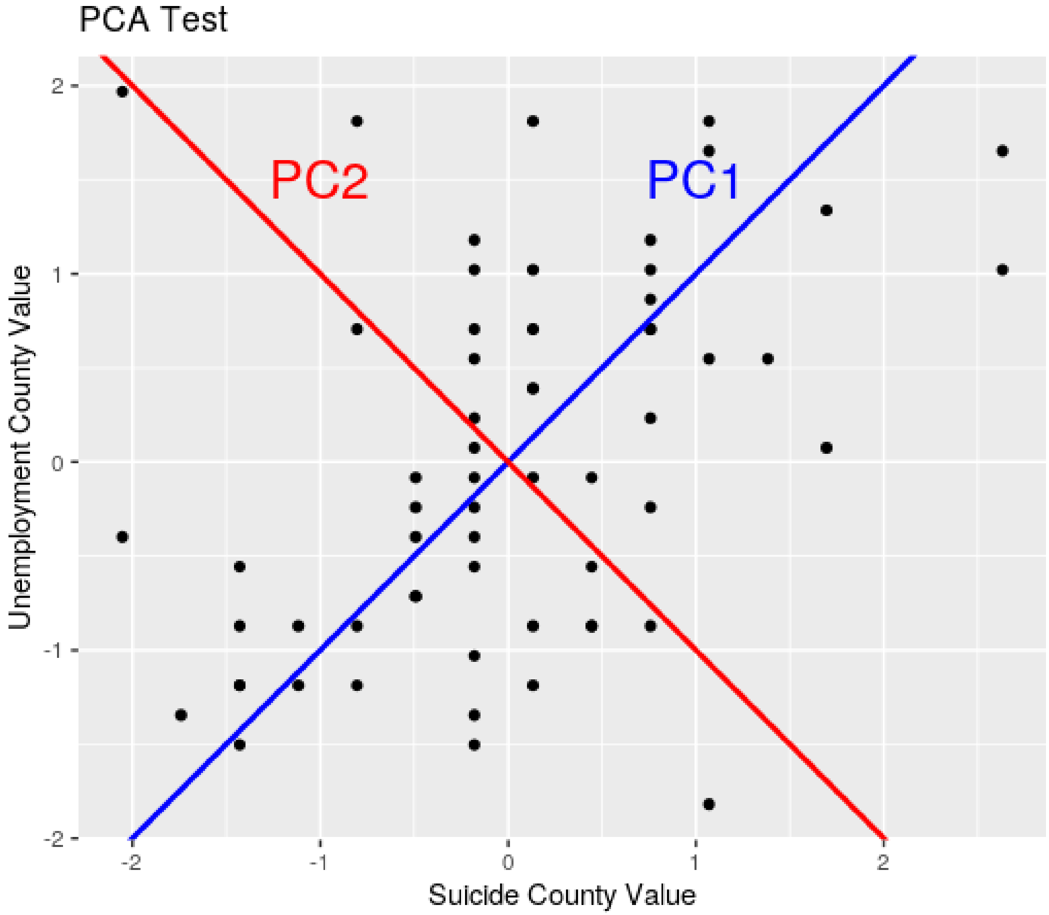
```
# Calculate total variance of Suicide  and Weight
var(NY_DATA_NUM$`Suicide County Value`) + var(NY_DATA_NUM$`Unemployment County
```

```
## [1] 2
```

```
#projection for R
ggplot(NY_DATA_NUM,  aes(x = `Suicide County Value`,  y =  `Unemployment Count
  geom_point() +
  geom_segment(aes(xend = `Suicide County Value`,  yend =  0),  lty = 2) +
  geom_hline(yintercept = 0, color = "blue", lwd = 1.5) +
  geom_point(aes(y =  0),  colour =  "orange", size = 3) + coord_fixed()+
  ggtitle("Score Plot")
```



```
# Set up PCA by hand (math explanation in lecture)
xy =  cbind(NY_DATA_NUM$`Suicide County Value`,  NY_DATA_NUM$`Unemployment Cou
svda =  svd(xy)
pc =  xy %*% svda$v[,  1] %*% t(svda$v[,  1])
bp =  svda$v[2,  1] / svda$v[1,  1]
ap =  mean(pc[,  2]) - bp * mean(pc[,  1])


#Plot a Scree plot too look at the relationship among the two variables.
ggplot(NY_DATA_NUM, aes(`Suicide County Value`,`Unemployment County Value`)) +
 geom_abline(intercept =  ap,  slope =  bp,  col =  "blue",  lwd =  1) + coord
   annotate(geom = "text", 1, 1.5, col = "blue", label = "PC1", size = 7)+
 geom_abline(intercept =  ap,  slope =  -bp,  col =  "red",  lwd =  1) +
 annotate(geom = "text", -1, 1.5, col = "red", label = "PC2", size = 7)+
  ggtitle("PCA Test")
```



We are using a PCA test in order to select for variation and look for maximum variance in order to create an optimum component. The goal of a PCA test is to show how all of the samples are related to each other. The samples in our data matrix must be the columns and the variables are the rows. Since there are 7 variables there are 7 principle components. After plotting the rotated data we can see that most county's in New York state have overlapping variance. However, since there are many county's it's hard to identify which counties have the most variance. Another result of the PCA indicates that the variables "Suicide County Value", "Crude Rate", "\#Unemployed", "Labor Force", and \#Suicide Deaths" Contribute to PC1. The variables "Z-score" and "Unemployment County Value" contribute to PC2. Principal 1 component accounts for 65% of the variance, and PC1 accounts for 23.13% of the variance in the data so together they account for 87.37% of the variance. With both PC1 and PC2 we are able to explain 87.37% of the variance in the data which means that we will explain the maximum variance by using only PC1 and PC2. We then decided to look at the variance between the two groups of "Unemployment county value" and "Suicide County Value". These two groups have a variance value of 2 which is large enough to distinguish the groups. The results from the PCA test shows that "Suicide County Value" and "Unemployment County Value" seem to have a negative relationship. As Unemployment levels increase Suicide Rates seem to decrease.

12.Conclusion

*After analyzing my results I can not conclude that there is a relationship between Suicide and Unemployment in New York State. The County Value which adjusts for the differences in population size among the counties shows no relationship. This County Value is very reliable since it accounts for differences in population size. However, if one was to solely look at the number of unemployment and amount of suicide in a given county, there seems to be a direct relationship therefore its important to make sure that one is looking at reliable variables. The results from this experiment were surprising since I expected there to a be relationship between Suicide and Unemployment in each New York County. This project clearly demonstrates a relationship between population size and suicide rates, this may explain why more suburban areas typically experience more Suicides yearly. There were some limitations with this data which include, the data not be collected throughout the same years. *

13.References

1. Arndt S, Acion L, Caspers K, Blood P. How reliable are county and regional health rankings? Prev Sci. 2013;14:497–502. doi: 10.1007/s11121-012-0320-3. [PubMed] [CrossRef] [Google Scholar] 2. https://www.countyhealthrankings.org/app/new-york/2020/measure/factors/23/data?sort=sc-0 *3. 7. Centers for Disease Control and Prevention (CDC) Behavioral risk factor surveillance – selected states, 1986. MMWR Morb Mortal Wkly Rep. 1987;36:252–4. [PubMed] [Google Scholar]*

```
##                                         sysname
##                                         "Linux"
##                                         release
##                            "4.15.0-142-generic"
##                                         version
## "#146-Ubuntu SMP Tue Apr 13 01:11:19 UTC 2021"
##                                        nodename
##                    "educcomp02.ccbb.utexas.edu"
##                                         machine
##                                        "x86_64"
##                                           login
##                                       "unknown"
##                                            user
##                                       "jcc5625"
##                                  effective_user
##                                       "jcc5625"
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.