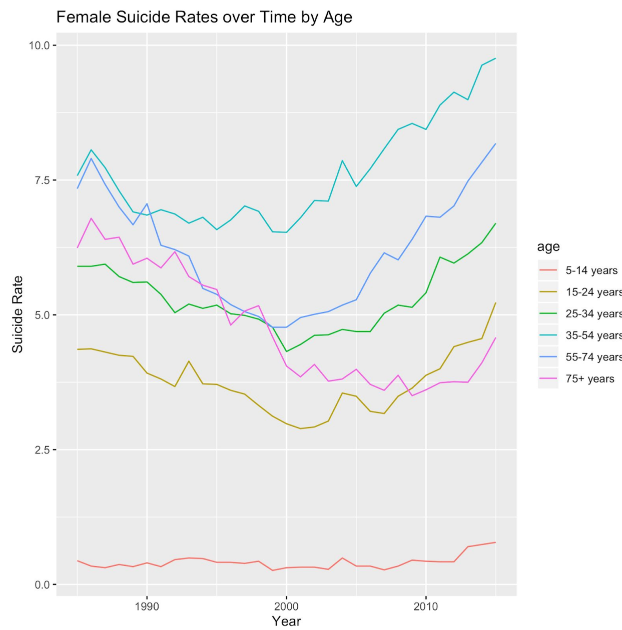
IEOR 142: Final Report

Group 10

Dorothy Leung, Fanice Nyatigo, Haibin Lim, Kana Mishra, Ashleigh Purvis

Motivation & Impact

The rate of people committing suicide every year in the United States is continually increasing under different circumstances and factors. The rate for women has been increasing for all age groups for the last 20 years, and the rate for men[1] for the last 15 years. Our final project focuses on predicting suicide rates for every 100,000 people in the United States by age group and gender over time. We are interested in learning the impact of various socio-economic factors in relation to suicide rates, in order to better ascertain which effects brought by the society and economy have the most impact on people prone to committing suicide. As a result, we hope to give more informed recommendations for public health interventions on significant factors based on our findings.



To accomplish this, the factors we decided to analyze were population size, HDI[2] (Human Development Index), GDP per capita, age group, gender, year, the depression percentage by gender, and the drug related death rate by age and gender. The GDP per capita and HDI are demonstrative of how much the government is investing in its people, with the population representing how large of a body is being governed and how the limited resources are being divided. The age group and gender are indicative of what individuals might be facing or experiencing in their lives (school for younger populations, work for middle aged, retirement for old age, and other life events). The year is used to find the trend over time. We were able to gather the depression rate over time by gender and average drug related death rates for every 100,000 people by age and gender over time from two additional data sources. We reasoned that the depression rate would give us a bit more insight into how individuals might be feeling on average, which we believe has a strong correlation with not just suicides, but also thoughts and attempts. The drug related death rate shed light on what individuals might be intaking to help cope with depression or other medical and non-medical issues. Overdosing can also be an indicator of a drug related death that was a suicide.

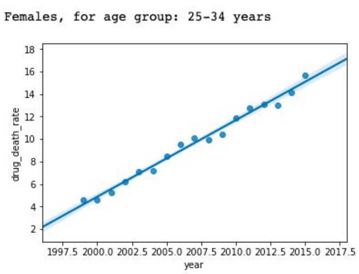
With our findings, we hope to be able to recommend changes be made to things like HDI and GDP per capita in order to help suicide rates decrease. In the best case, we would be able to see suicide rates decreasing and verify which factors are the most significant. This way, we could suggest that whatever changes caused the decreased be continued. The ultimate goal is to help save lives in the long run, whether it be for someone who has barely lived their life, lived most of it, or even anywhere in between. With more accessible open source data, our models could even be applied to other countries to help reduce suicide rates there. We could also further the impact by considering other features such as education level, food quality, and stress levels for all age groups by gender and country per year.

[1] The chart for males can be found in the appendix.
[2] For more information about HDI, visit http://hdr.undp.org/en/content/human-development-index-hdi.

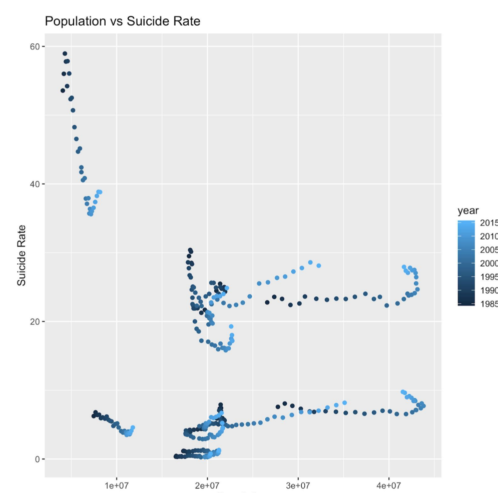
Data[3]

The data used in this study was originally collected from Kaggle[4] with the original topic of "find signals correlated to increased suicide rates among different cohorts globally." We extended the scope of our analysis by focusing on patterns particularly in the US of suicide rates by year, country, age group, etc. To fill in the NA values for the Kaggle dataset, we first made sure that the data was growing in a linear trend, and then we were able to interpolate by averaging the previous and next existing data points to fill the middle missing values.

To better analyze the overall impact of factors, we searched for additional data to add to our features. We found data by gender for depression[5], which is the most common mental illness that people suffer before having a suicide attempt and is also often undiagnosed and untreated[6]. We were also able to find data for drug related death rates[7] by gender and age group per year. In order to merge the depression data which was not at the same level of granularity, we had to fill all age groups of the gender for the year with the same value, which could lead to limitations in finding patterns in the data. There were also some issues with NA values since the data we found did not fully span across the years we had from the Kaggle dataset. To fill in the NA values from the new datasets we performed EDA to study the general trend of the interval wrapping the missing values. We looked at each gender and age group pairing individually, and calculated best fit polynomials for each one. Using these best fit predictions (courtesy of numpy.polyfit[8]), we were able to backcast[9] values to fill the early years we could not find data for. We chose not to overfit and most of these were filled in using linear or constant models, based on what our EDA showed us. However, some age groups seemed to follow no coherent trends, so we decided to simply average the next existing 5 years of values to find the value for that year.

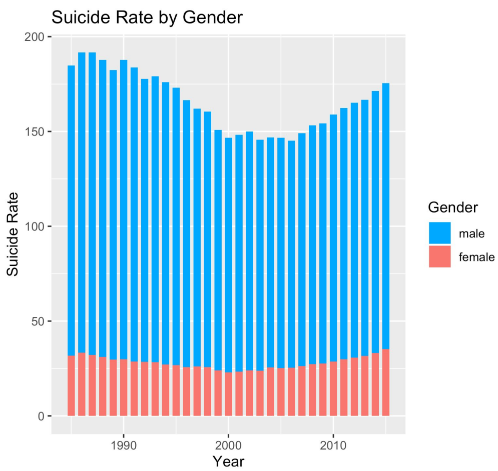


EDA: Upon exploring the data[10] through our preliminary analysis, we found that the range of the suicide rate per 100,000 people ranged from 26% to about 59%. The sex most likely to commit suicide is male, and the generation with the highest suicide rate is the Silent generation. Interestingly, the age group that contains the highest rate was 75 and over (according to our data, this age group is mainly encompassed within the Silent and G.I. generations). We also discovered that population and suicide rate are slightly negatively correlated: suicide rates are highest in smaller populations. Suicide rates also seem to fluctuate the most around population of sizes 2 billion, which was an interesting find and may allude to a certain area with this size population having a particularly high suicide rate – we could further investigate this if we had additional data pertaining to individual regions within the U.S. We also found that there is a positive correlation between population and drug death rate of about 68%.



---

[3] Note: We changed our project completely a few weeks ago, so we had to start over with a new data set.
[4] Data obtained from https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016.
[5] https://www.statista.com/statistics/979898/percentage-of-people-with-depression-us-by-gender/.
[6] https://afsp.org/about-suicide/risk-factors-and-warning-signs/.
[7] https://www.cdc.gov/nchs/hus/contents2018.htm#Table_008.
[8] https://docs.scipy.org/doc/numpy/reference/generated/numpy.polyfit.html.
[9] https://stats.stackexchange.com/questions/59271/what-is-the-proper-name-for-a-backward-forecast.
[10] Final version of our dataset: https://github.com/kanam12/ieor142finalproject/blob/master/us_suicides_merged_no_na.csv.



There is a correlation of about 29% between suicide rate and drug death rate.

Lastly, we determined that while the suicide rate declined from the late 1990s until about 2009, it has since increased and has steadily been increasing in recent years. In order to show the alarming increase of commited suicides the chart to the left shows the trend over time colored by gender. This further affirms the social magnitude of our project and the potential impact it can have.

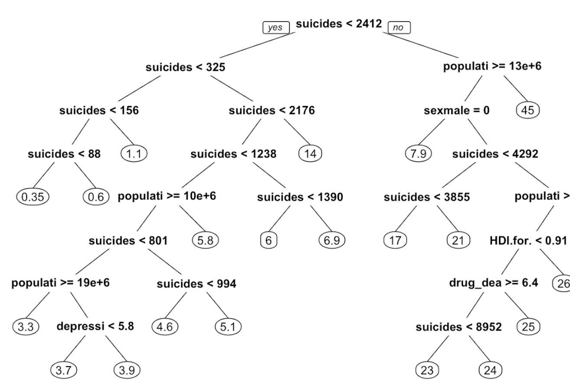
Our final dataframe after merging and cleaning contains 14 columns and 372 rows. The original Kaggle dataset had 12 columns to which we added the 2 more mentioned earlier.

Analytics Models

We focused on implementing our models with linear regression, CART, random forest, boosting, and time series. We also provided methods of cross validation to train models for an in-depth analysis. The metrics we are using are RMSE and out-of-sample R^2 across the different models. Assuming our variables are independent and identically distributed, we performed a normalized random 70:30 split for training:test set ratio.

We initially used a baseline approach that always predicted the average suicide rate of 13.82 and obtained a fairly low OSR^2 of 0.00003278162. We realized this naive baseline approach is not the right indication for us to reference our results for other upcoming models. Since suicides rate is a continuous variable, we opted to first start with a linear regression model to explore the relationship between suicides rate and some of the other variables. The first exploratory model included all the independent variables, after which we computed the Variance Inflation Factor (VIF) to determine multicollinearity[11]. From the linear model equation, it appears that the number of suicides, rate of death from drugs and generations X (part of intercept), Boomers, Silent and G.I all have a positive linear relationship with the rate of suicides. Generations Z, millenials and the population have a negative linear relationship. After 5 training rounds, the final linear regression model had an R^2 value of 0.8014 and an OSR^2 value of 0.7357. This shows that the model did not overfit too much because the model performed decently on the test set, relatively similar to the results of the training set. With CART, random forest, boosting, and time series, we are hoping to get even better results.

For CART, we built a regression tree using a 10-fold cross validation method with a tuning parameter, `cp` that ranged from 0 to 0.1, with a step size of 0.001. We obtained the highest R^2 with the lowest RMSE at a `cp` value of 0. The regression tree yields the following results, as summarized in the diagram. The reason we picked CART as one of our models is because we do not need to make any implicit assumptions about the underlying relationships of the features we are using, allowing us to capture nonlinear trends. From our EDA, we found that depression seemed to be nonlinearly correlated with suicide rates, seeming to increase only when the average depression rates were high or



[11] See appendix for details about VIF procedure and linear equation.

nonexistent and decreasing otherwise. Our goal was for CART to be able to capture and describe these nuances. As a result, our CART model had an R^2 of 0.9398040 and an OSR^2 of 0.8777108.

Along with CART, we created a random forest model in search for a possible model similar to CART but with a better OSR^2. We first created a tree with 5-fold cross validation with `mtry` values ranging from 1 to 5, and obtained 5 as the optimal fitting `mtry`. The final model for this `mtry` value gave us an OSR^2 of 0.9645385. To explore further, we trained another set of random forest with the same 5-fold cross validation except with `mtry` values ranging from 1 to 10. From this training set, we obtained best tuned `mtry` of 10. Our final model on the test set with `mtry` = 10 had an OSR^2 of 0.9928406 which was higher than that of CART.

The next model we ran was a boosting model under a gaussian distribution with `n.trees` = 1000, `shrinkage` = 0.001, and `interaction.depth` = 2. However, this model surprisingly underperformed and only had an OSR^2 of 0.7628486. Some of the most influential features were suicides_no, population, and sex. We also tried running different versions of this gradient boosting model with cross validation, but for some reason, all of those attempts[12] resulted in an OSR^2 of -0.2342707.

The last models we tried was the time series models: random walk, auto-regressive (AR), and random forest. The random walk model had an R^2 of 0.9965203. However, this seemed to be unreliable for predicting drastic fluctuations. The one term AR model had an R^2 of 0.9967 and an OSR^2 of 0.9097873. The two term AR model had a slightly higher R^2 of 0.9969. Both the two term AR model and RF model performed similarly[13], with the two term AR model having a slightly higher OSR^2 of 0.9090414 as opposed to the RF model's OSR^2 value of 0.8945664.

We also tried to run a neural net model, but unfortunately ran into some issues that we (with the help of the professor) nor Google were able to resolve. We decided to not run unsupervised models like k-means since we were trying to predict continuous data instead of cluster things into groups or classify things. For this same reason, we also did not run confidence intervals for metrics such as accuracy, TPR, and FPR since those metrics did not apply to our project.

Models	R2	OSR2
Baseline	-	3.27816e-05
Linear Regression	0.8014	0.7357
CART	0.939804	0.877711
RF (mtry = 5)	0.973565	0.964538
RF (mtry = 10)	0.986766	0.992841
Boosting	-	0.762849
Time Series (Random Walk)	0.99652	-
Time Series (one term AR)	0.9967	0.909787
Time Series (two term AR)	0.9969	0.909041
Time Series (RF)	-	0.894566

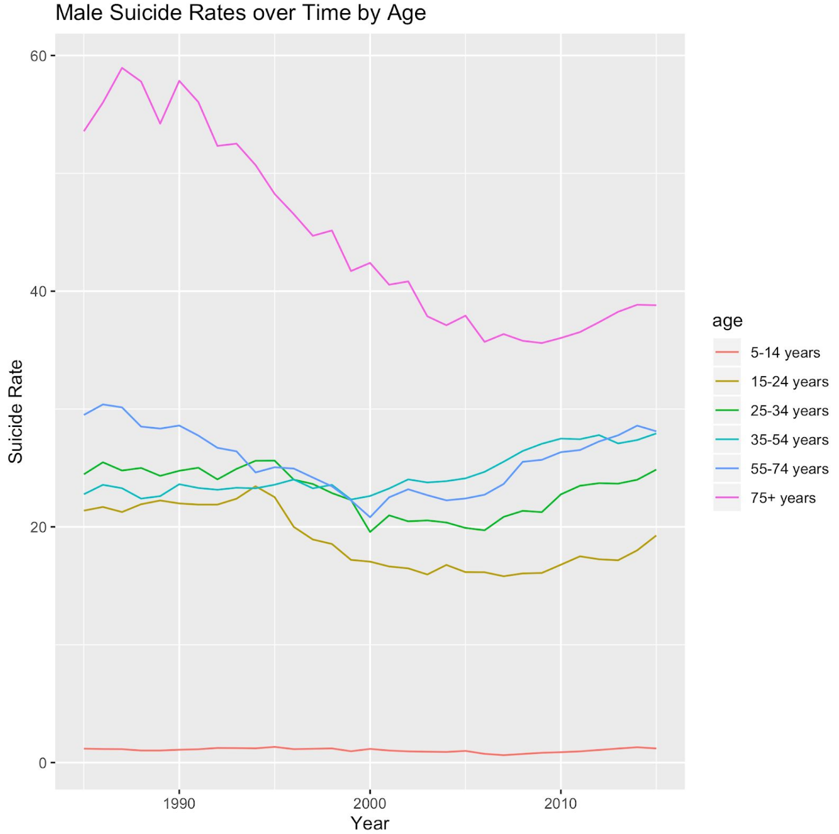
Above is a chart detailing all of the different models we tried along with their performances[14]. The final model we would propose based on our OSR^2 values would be the random forest with `mtry` = 10. We could try to cross validate the random forest model more and potentially use a larger test set to gain more confidence in its performance, or we could also use a one or two term AR model.

[12] The code is in the cartandrf.Rmd file but is commented out at the end.
[13] See appendix for the plot with both models.
[14] The code for all our models can be found on our GitHub.

Appendix

All of the project code can be found at https://github.com/kanam12/ieor142finalproject. Below are the elements from the footnotes, and following this is the pdf version of the Rmd file for our EDA.

1. Male Suicide Rate Trend over Time by Age (ref: page 1, Motivation & Impact)

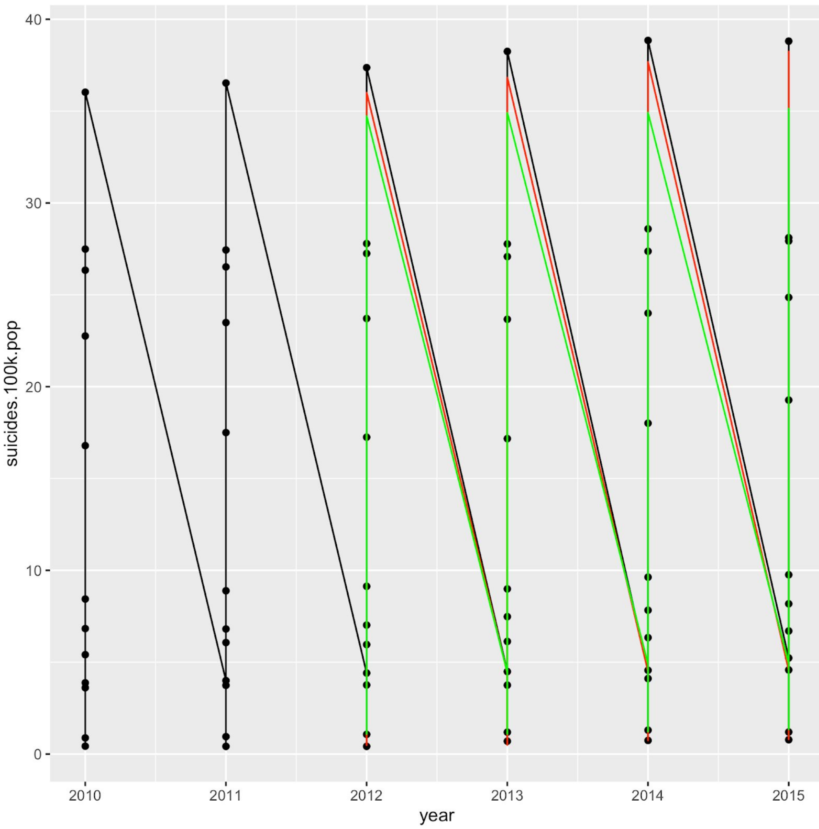


3. See https://github.com/kanam12/ieor142finalproject/tree/master/old for the old dataset we were working with for previous rendition of the project. (ref: page 2, Data)

11. Not surprisingly, `gdp_per_capita ($)` had the highest VIF value (`gdp_per_capita ($)` = `gdp_for_year ($)`/`population`), so it was removed and the model re-trained. We sequentially removed variables with the highest VIF and re-run the models until we were left with variables having little linear correlation with each other (VIF less than or equal to 5). The p-values of the variables removed (p-values greater than our cut-off of 0.05) were also not significant, and their removal only changed the model's R^2 by 0.0024. Next, we re-trained the model by sequentially removing variables with the highest p-value above our 0.05 cut-off value, until we remained with those whose p-value was statistically significant. As with the VIF case, we constantly monitored the R^2 value to ensure that it did not decrease significantly (the R^2 only dropped by 0.0025). After 5 training rounds, the final linear regression model has an R^2 value of 0.8014 and upon testing it on the test set, an OSR^2 of 0.7357. The resulting equation is as follows:

`suicides/100k pop` = 3.815e+01 + 2.110e-03 `suicides_no` - 7.623e-07 `population` + 5.602e+00 `generationBoomers` + 3.905e+00 `generationSilent` + 9.855e+00 `generationG.I. Generation` -1.682e+00 `generationMillenials` - 4.237e+00 `generationGeneration Z` - 3.834e+00 `depression_percentage` + 2.253e-01 `drug_death_rate`

13. The red line represents the two term AR model and the green represents the RF model, with the black line representing the data.



EDA IEOR 142, Final Project

3033342158

December 16, 2019

```
#install.packages("Rcpp")
#install.packages("purrr")
#install.packages("dplyr")

library(softImpute)
```

```
## Warning: package 'softImpute' was built under R version 3.5.3
```

```
## Loading required package: Matrix
```

```
## Loaded softImpute 1.4
```

```
library(gridExtra, verbose=FALSE, warn.conflicts=FALSE, quietly=TRUE)
```

```
## Warning: package 'gridExtra' was built under R version 3.5.3
```

```
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 3.5.3
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:gridExtra':
##
##     combine
```

```
library(ranger)
```

```
## Warning: package 'ranger' was built under R version 3.5.3
```

```
##
## Attaching package: 'ranger'
```

```
## The following object is masked from 'package:randomForest':
##
##     importance
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.5.3
```

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

```
## The following object is masked from 'package:randomForest':
##
##     combine
```

```
## The following object is masked from 'package:gridExtra':
##
##     combine
```

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

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

```
library(reshape2)
library("caTools")
```

```
## Warning: package 'caTools' was built under R version 3.5.3
```

```
library(ROCR)
```

```
## Warning: package 'ROCR' was built under R version 3.5.3
```

```
## Loading required package: gplots
```

```
## Warning: package 'gplots' was built under R version 3.5.3
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following object is masked from 'package:randomForest':
##
##     margin
```

```
restore = list(repr.plot.width=8, repr.plot.height=3)
PALETTE = c("#00A9FF", "#F8766D", "#7CAE00", "#C77CFF", "#CD9600", "#00BE67", "#FF61CC", "#00BFC
4")
theme.x_axis_only = theme(axis.title.y=element_blank(), axis.text.y=element_blank(), axis.ticks.
y=element_blank(), panel.grid.major.y=element_blank(), panel.grid.minor.y=element_blank())
theme.no_legend = theme(legend.position="none")
theme.legend_title = theme(legend.title=element_text(size=7))
data <- read.csv("us_suicides_merged_no_na.csv")
```

EDA

Our cleaned and merged data consists of 372 observations and 14 variables.

```
nrow(data)
```

```
## [1] 372
```

```
ncol(data)
```

```
## [1] 14
```

```
#First look at the first 6 and last 6 observations of our data
head(data)
```

```
##         country year    sex         age suicides_no population
## 1 United States 1985 female 15-24 years         854   19589000
## 2 United States 1985   male 15-24 years        4267   19962000
## 3 United States 1985 female 25-34 years        1242   21041000
## 4 United States 1985   male 25-34 years        5134   20986000
## 5 United States 1985 female 35-54 years        2105   27763000
## 6 United States 1985   male 35-54 years        6053   26589000
##   suicides.100k.pop      country.year HDI.for.year gdp_for_year....
## 1              4.36 United States1985        0.841     4.346734e+12
## 2             21.38 United States1985        0.841     4.346734e+12
## 3              5.90 United States1985        0.841     4.346734e+12
## 4             24.46 United States1985        0.841     4.346734e+12
## 5              7.58 United States1985        0.841     4.346734e+12
## 6             22.77 United States1985        0.841     4.346734e+12
##   gdp_per_capita....   generation depression_percentage drug_death_rate
## 1              19693 Generation X              6.519361         0.00000
## 2              19693 Generation X              3.520442         0.00000
## 3              19693      Boomers              6.519361         0.00000
## 4              19693      Boomers              3.520442         0.00000
## 5              19693       Silent              6.519361         0.00000
## 6              19693       Silent              3.520442        10.69853
```

```
tail(data)
```

```
##           country year    sex         age suicides_no population
## 367 United States 2015 female  5-14 years         158   20342901
## 368 United States 2015   male  5-14 years         255   21273987
## 369 United States 2015 female 55-74 years        2872   35115610
## 370 United States 2015   male 55-74 years        9068   32264697
## 371 United States 2015 female   75+ years         540   11778666
## 372 United States 2015   male   75+ years        3171    8171136
##     suicides.100k.pop      country.year HDI.for.year gdp_for_year....
## 367              0.78 United States2015         0.92     1.812071e+13
## 368              1.20 United States2015         0.92     1.812071e+13
## 369              8.18 United States2015         0.92     1.812071e+13
## 370             28.11 United States2015         0.92     1.812071e+13
## 371              4.58 United States2015         0.92     1.812071e+13
## 372             38.81 United States2015         0.92     1.812071e+13
##     gdp_per_capita....   generation depression_percentage drug_death_rate
## 367              60387 Generation Z                  6.03             0.3
## 368              60387 Generation Z                  3.51             0.2
## 369              60387      Boomers                  6.03            23.7
## 370              60387      Boomers                  3.51            34.7
## 371              60387       Silent                  6.03             7.4
## 372              60387       Silent                  3.51             8.9
```

The dataset contains 31 unique years from 1985 to 2015, the suicide rate per 100k has a variance of 175.0296.

```
unique(sort(data$year))
```

```
##  [1] 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998
## [15] 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012
## [29] 2013 2014 2015
```

```
length(unique(sort(data$year)))
```

```
## [1] 31
```

```
var(data$suicides.100k.pop)
```

```
## [1] 175.0296
```

```
#table(data$suicides.100k.pop) / length(data$suicides.100k.pop)) # relative frequencies
as.numeric(names(table(data$suicides.100k.pop))[table(data$suicides.100k.pop) == max(table(data
$suicides.100k.pop))]) # mode for suicide rate
```

```
## [1] 0.34
```

```
as.numeric(names(table(data$gdp_per_capita....))[table(data$gdp_per_capita....) == max(table(dat
a$gdp_per_capita....))]) # mode for gdp per capita
```

```
##  [1] 19693 20588 21631 23103 24654 26004 26503 27760 28891 30375 31518
## [12] 32928 34644 36164 38072 39218 40018 40845 42468 44867 47423 49666
## [23] 50563 51585 51989 52128 53452 55170 56520 58531 60387
```

```
range(data$suicides.100k.pop)
```

```
## [1]  0.26 58.95
```

```
data[data$suicides.100k.pop == min(data$suicides.100k.pop), ]
```

```
##           country year    sex        age suicides_no population
## 175 United States 1999 female 5-14 years          50   19275566
##     suicides.100k.pop      country.year HDI.for.year gdp_for_year....
## 175              0.26 United States1999        0.885     9.660624e+12
##     gdp_per_capita.... generation depression_percentage drug_death_rate
## 175              38072 Millenials                  5.92             0.1
```

```
data[data$suicides.100k.pop == max(data$suicides.100k.pop), ]
```

```
##          country year  sex       age suicides_no population
## 36 United States 1987 male 75+ years        2532    4295000
##    suicides.100k.pop      country.year HDI.for.year gdp_for_year....
## 36             58.95 United States1987         0.85     4.870217e+12
##    gdp_per_capita....      generation depression_percentage
## 36              21631 G.I. Generation               3.51864
##    drug_death_rate
## 36        7.466624
```

Investigating Suicide rate and Sex

There are 186 males and 186 females. There is also 62 records for every age range provided in the data. The data seems to be split evenly thus far except for the generation variable.Generation X has the highest amount of records and Generation Z has the least. The Suicide rate had a decline from about the late 1990’s to the mid 2000’s but has been steadily increasing since around the year 2008.

```
table(data$sex)
```

```
## 
## female   male 
##    186    186
```

```
table(data$age)
```

```
##
## 15-24 years 25-34 years 35-54 years  5-14 years 55-74 years   75+ years
##          62          62          62          62          62          62
```

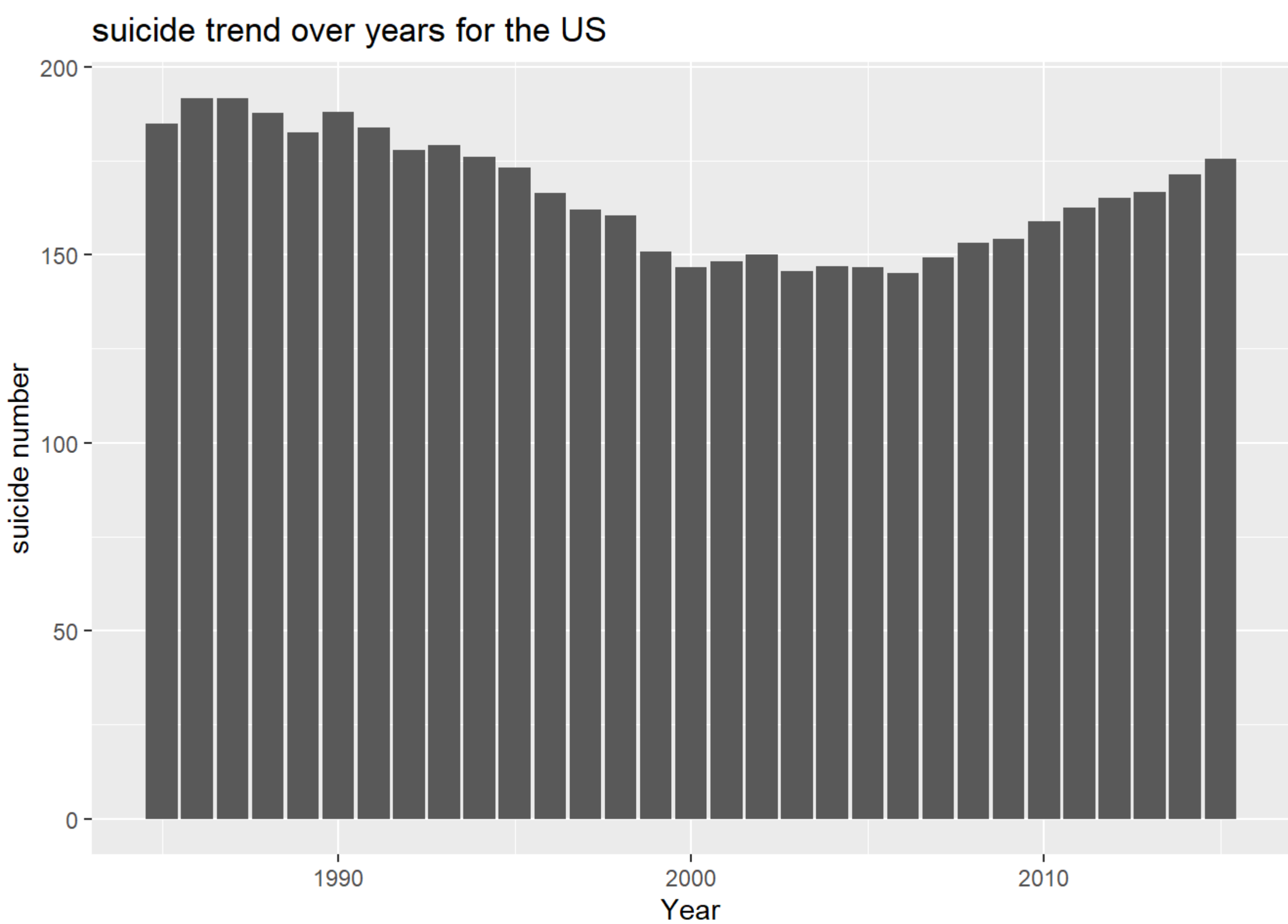
```
table(data$generation)
```

```
##
##         Boomers G.I. Generation     Generation X     Generation Z
##              68              44               88               18
##      Millenials          Silent
##              72              82
```

```
max(table(data$generation))
```

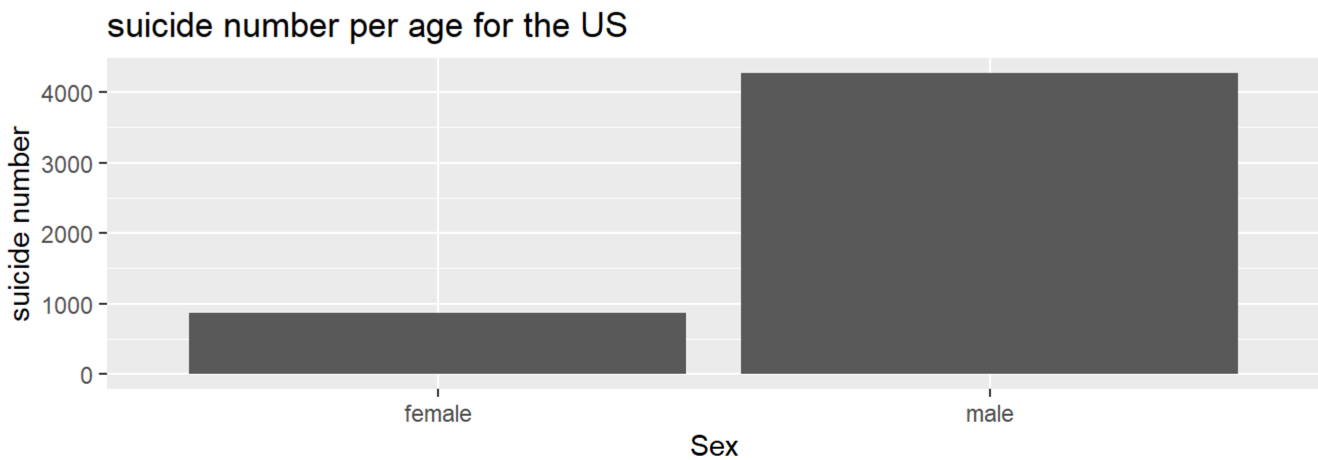
```
## [1] 88
```

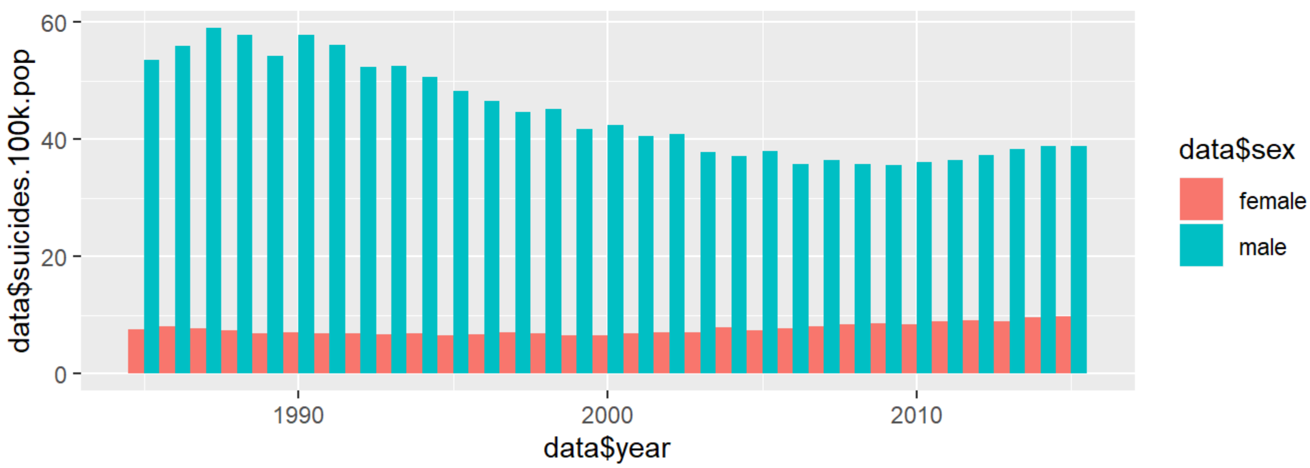
```
ggplot(data)  + ggtitle("suicide trend over years for the US") +
geom_col(aes(x=data$year, y=data$suicides.100k.pop)) + xlab("Year") + ylab("suicide number")
```



```
p1 = ggplot(data)  + ggtitle("suicide number per age for the US") +
geom_col(aes(x=data$sex, y=data$suicides.100k.pop)) + xlab("Sex") + ylab("suicide number")


p2= ggplot(data, aes(x=data$year, y=data$suicides.100k.pop, fill=data$sex), xlab("Year"), ylab(
"Suicide Rate")) +
   geom_bar(stat="identity", width=1, position = "dodge")
grid.arrange(p1, p2, nrow=2, ncol = 1.2)
```



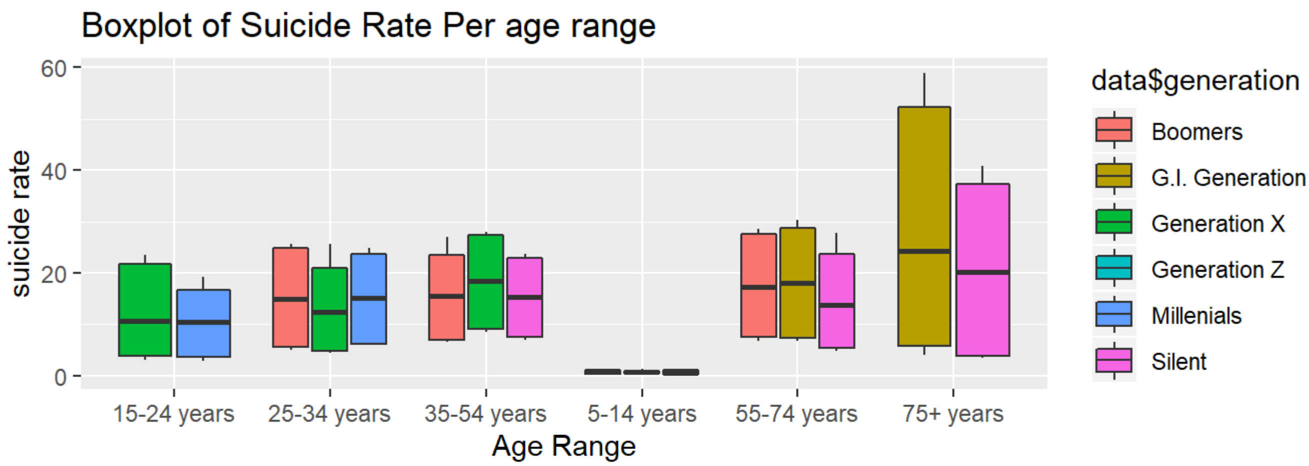


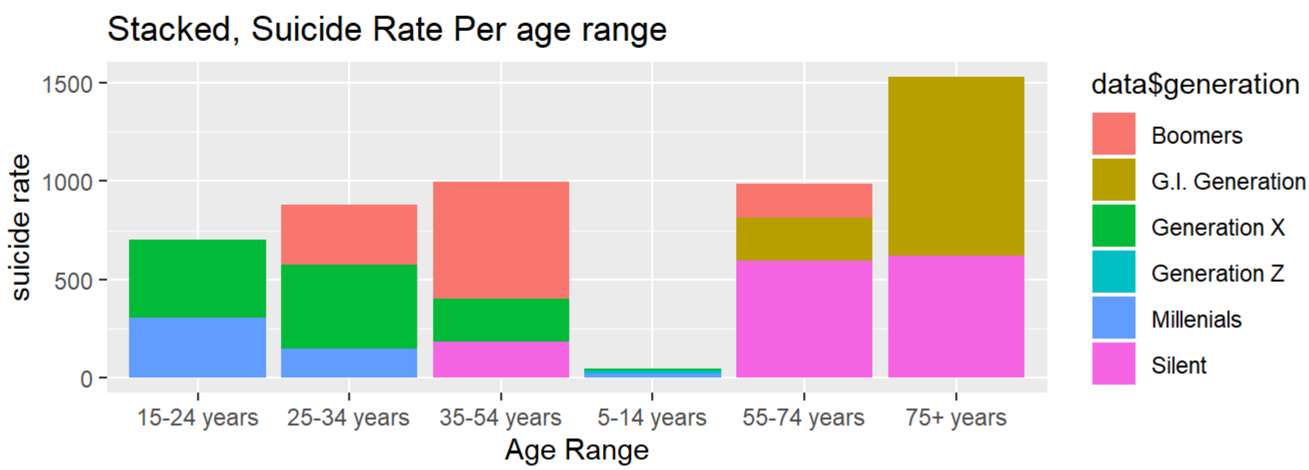
Investigating Suicide rate and Age

Suicide rates are highest among indivduals in the age group 75+(This is mostly people considered to be from the G.I. generation(1901-1924) and Silent generation(1925-1945)) and the lowest rates occur in the age group 5-14(generation X and generation Z).

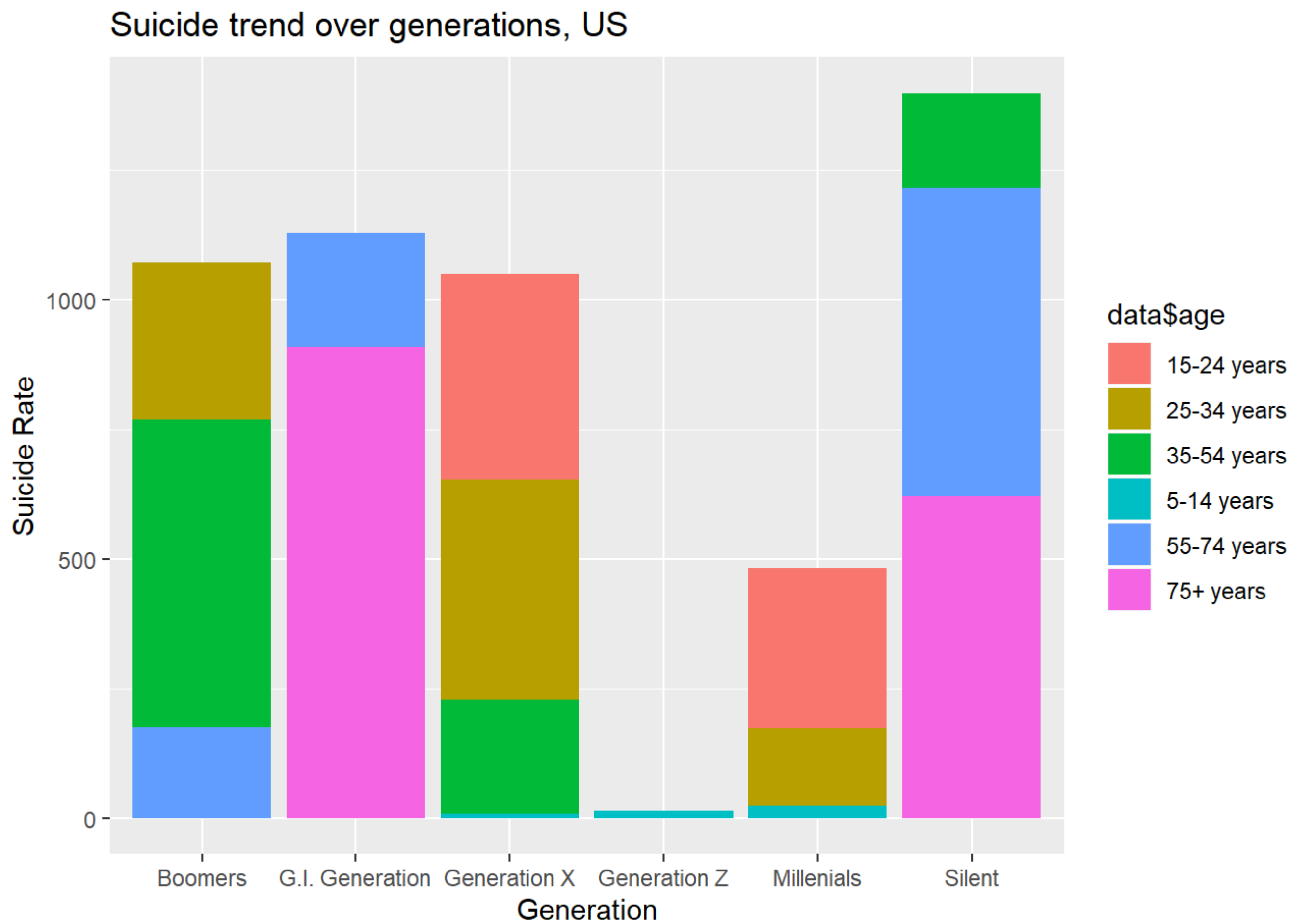
```
p3 = ggplot(data) + ggtitle("Boxplot of Suicide Rate Per age range") + geom_boxplot(aes(x= data
$age, y=data$suicides.100k.pop, fill = data$generation)) +
xlab("Age Range") + ylab("suicide rate")

p4 = ggplot(data) + ggtitle("Stacked, Suicide Rate Per age range") + geom_col(aes(x= data$age, y
=data$suicides.100k.pop, fill = data$generation)) +
xlab("Age Range") + ylab("suicide rate")
grid.arrange(p3, p4, nrow=2, ncol = 1.1)
```






```
ggplot(data)  + ggtitle("Suicide trend over generations, US") +
geom_col(aes(x= data$generation, y= data$suicides.100k.pop, fill=data$age), position="stack") +
 xlab("Generation") + ylab("Suicide Rate")
```

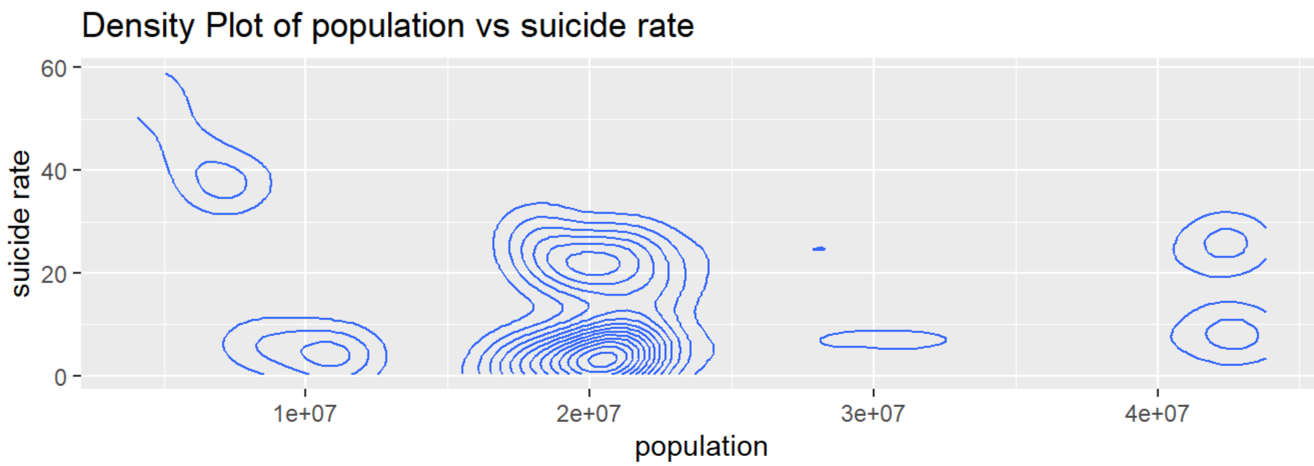


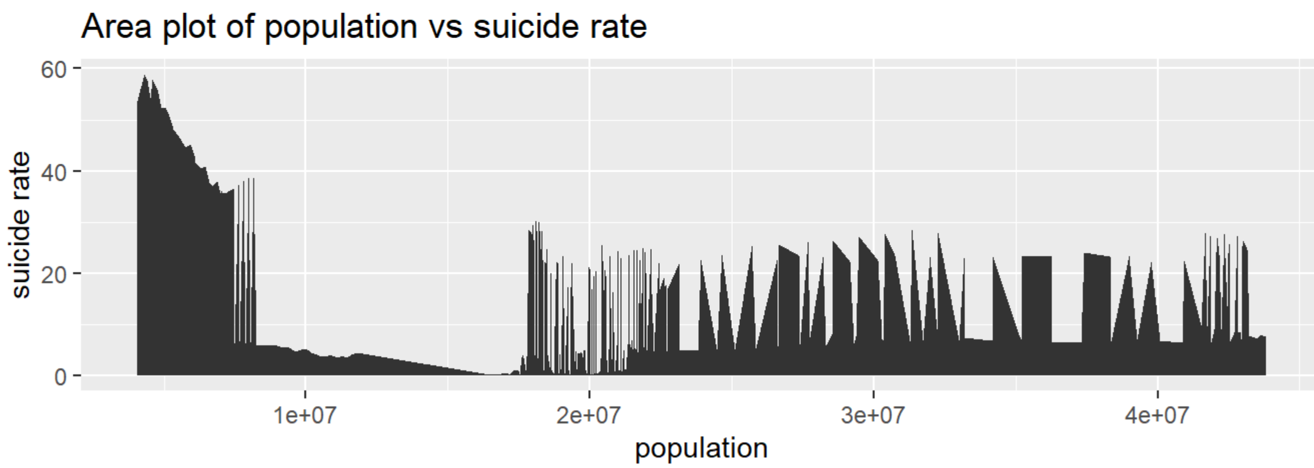
Investigating Suicide rate and Population

```
p5 = ggplot(data)  + ggtitle("Density Plot of population vs suicide rate") +
geom_density_2d(aes(x=data$population, y=data$suicides.100k.pop)) + xlab("population") + ylab("s
uicide rate")

p6 = ggplot(data)  + ggtitle("Area plot of population vs suicide rate") +
geom_area(aes(x=data$population, y=data$suicides.100k.pop)) + xlab("population") + ylab("suicide
rate")

grid.arrange(p5, p6, nrow=2, ncol = 1.1)
```






```
var(data$population)
```

```
## [1] 8.92766e+13
```

```
cor(data$suicides.100k.pop, data$population)
```

```
## [1] -0.1703968
```

```
summary(data$population)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  4064000 18185450 20375469 21650611 22616944 43805214
```

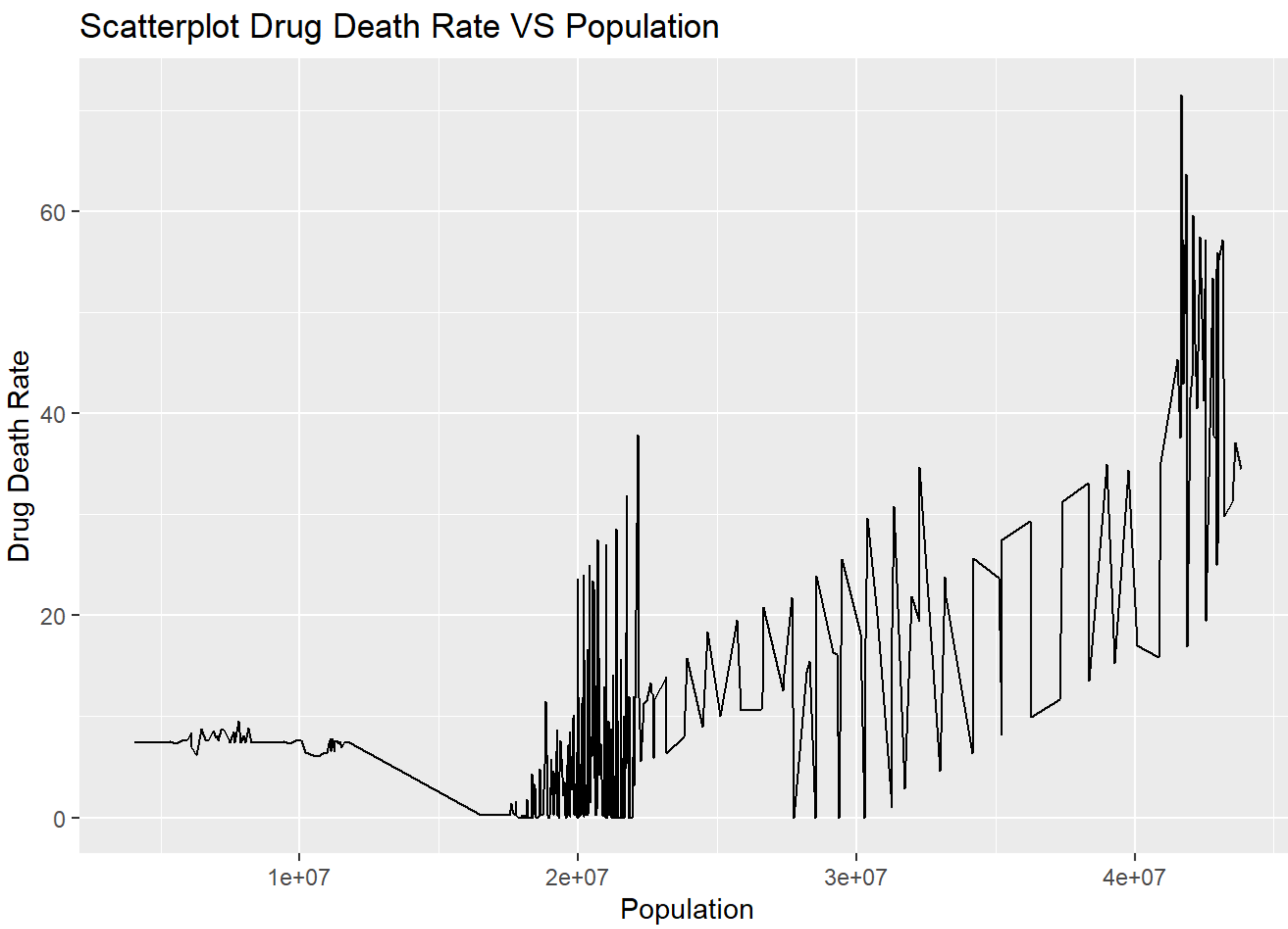
```
cor(data$population, data$depression_percentage)
```

```
## [1] 0.05065976
```

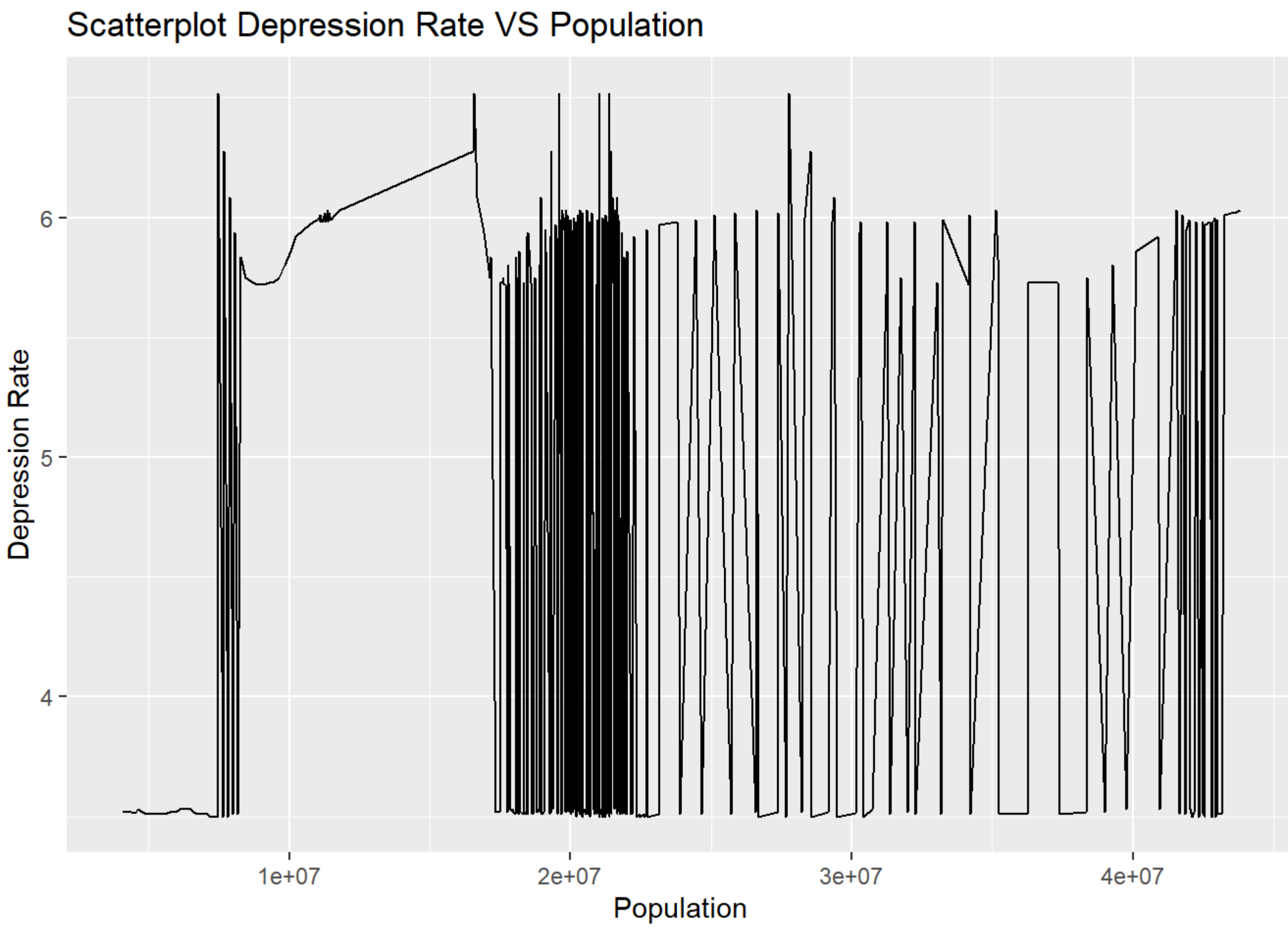
```
cor(data$population, data$drug_death_rate)
```

```
## [1] 0.6774055
```

```
ggplot(data) + geom_line(aes(x = data$population, y = data$drug_death_rate)) + xlab("Population"
) +ggtitle("Scatterplot Drug Death Rate VS Population") + ylab("Drug Death Rate")
```




```
ggplot(data) + geom_line(aes(x = data$population, y = data$depression_percentage)) + xlab("Popul
ation") +ggtitle("Scatterplot Depression Rate VS Population") + ylab("Depression Rate")
```



###Investigating Suicide rate and HDI for year

```
var(data$HDI.for.year) #Very low variance for HDI year to year
```

```
## [1] 0.0005165123
```

```
cor(data$suicides.100k.pop, data$HDI.for.year) #Barley negatively correlated
```

```
## [1] -0.06456609
```

```
cor(data$population, data$HDI.for.year)
```

```
## [1] 0.2177246
```

```
cor(data$gdp_per_capita...., data$HDI.for.year)#Sanity check: has a positive correlation
```

```
## [1] 0.9853092
```

```
cor(data$depression_percentage, data$HDI.for.year)
```

```
## [1] -0.0009472623
```

```
cor(data$drug_death_rate, data$HDI.for.year) # correlation: 0.4429688 somewhat positivley correl
ated
```

```
## [1] 0.4429688
```

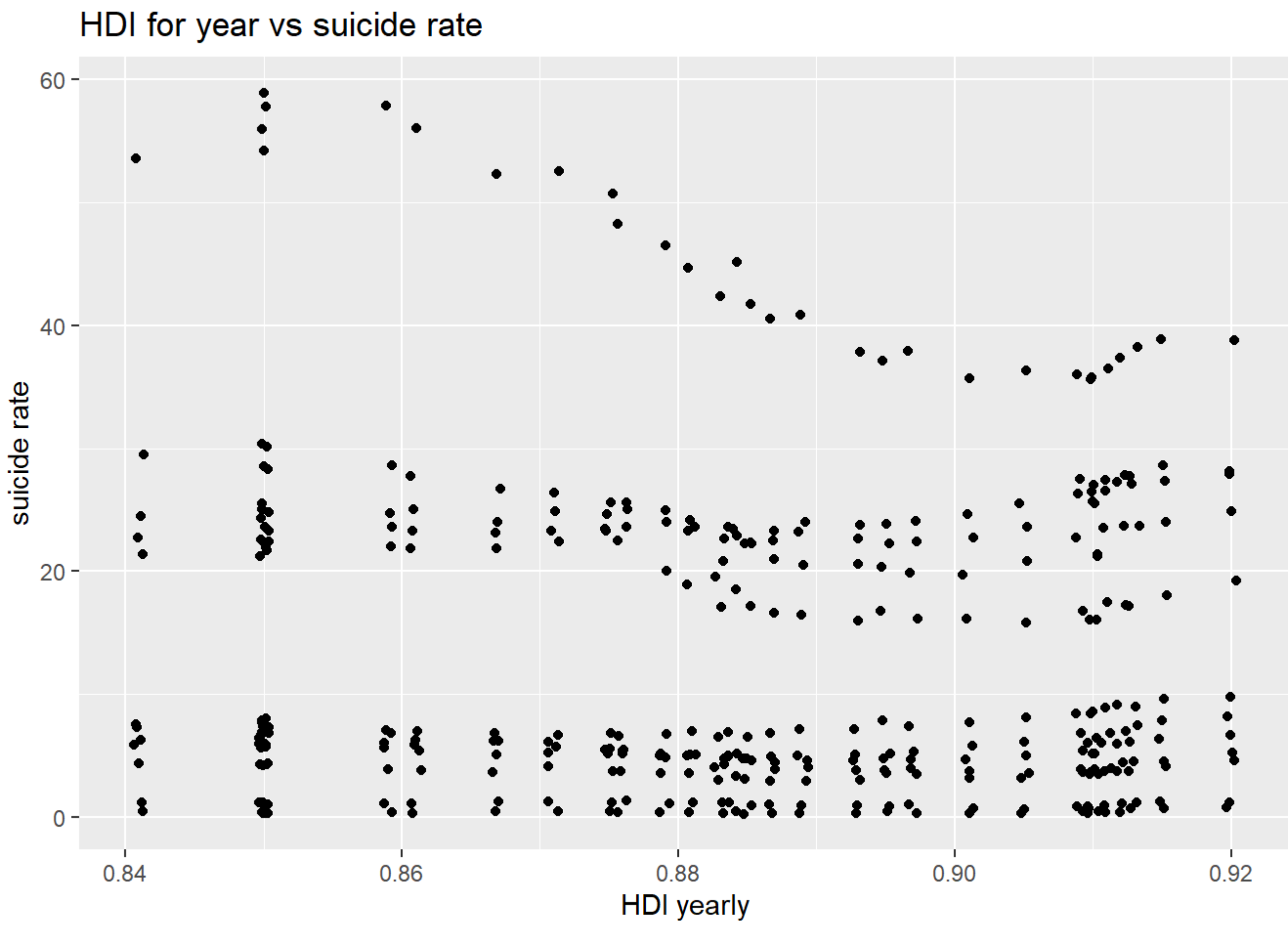
```
summary(data$HDI.for.year)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.8410  0.8670  0.8850  0.8848  0.9090  0.9200
```

```
as.numeric(names(table(data$HDI.for.year))[table(data$HDI.for.year) == max(table(data$HDI.for.ye
ar))]) # mode for HDI
```

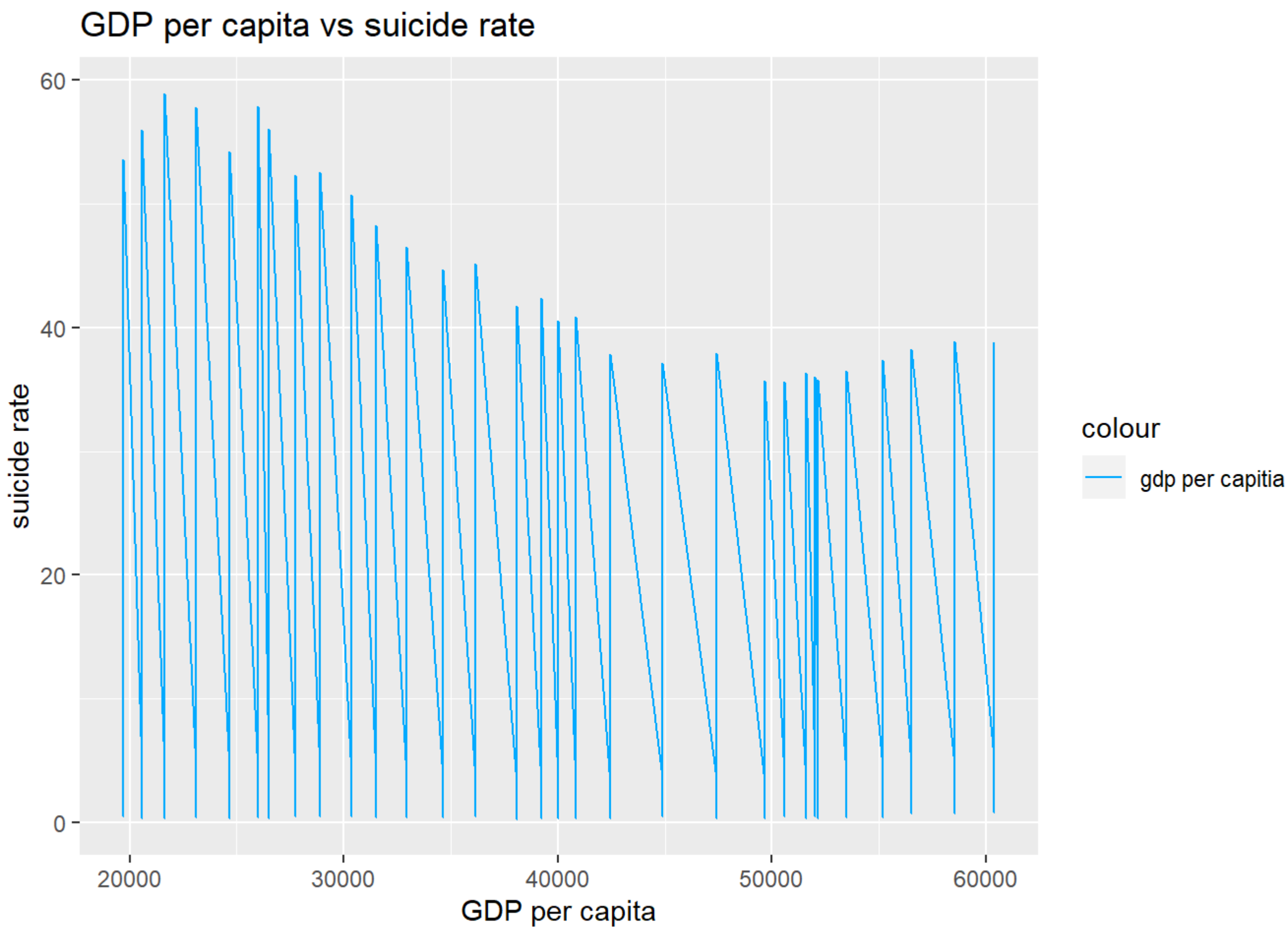
```
## [1] 0.85
```

```
ggplot(data)  + ggtitle("HDI for year vs suicide rate") +
geom_jitter(aes(x=data$HDI.for.year, y=data$suicides.100k.pop)) + xlab("HDI yearly") + ylab("sui
cide rate")
```



###Investigating Suicide rate and GDP per capita

```
ggplot(data)  + ggtitle("GDP per capita vs suicide rate") +
geom_line(aes(x=data$gdp_per_capita...., y=data$suicides.100k.pop, color = "gdp per capitia")) +
xlab("GDP per capita") + ylab("suicide rate") + scale_color_manual(values=PALETTE[1:3])
```




```
var(data$suicides.100k.pop, data$gdp_per_capita....)
```

```
## [1] -9979.495
```

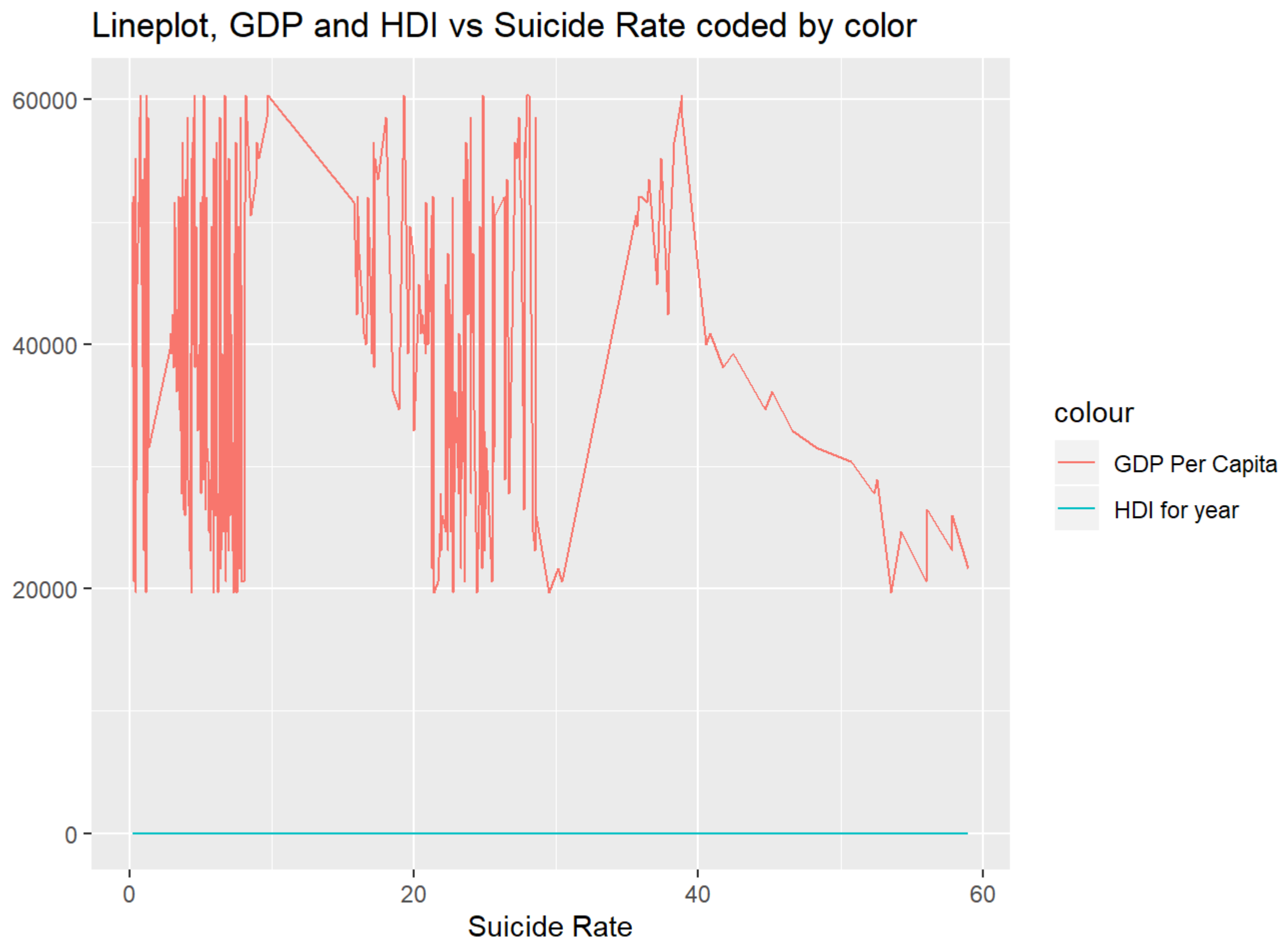
```
cor(data$suicides.100k.pop, data$gdp_per_capita....)
```

```
## [1] -0.0611568
```

```
summary(data$suicides.100k.pop, data$gdp_per_capita....)
```

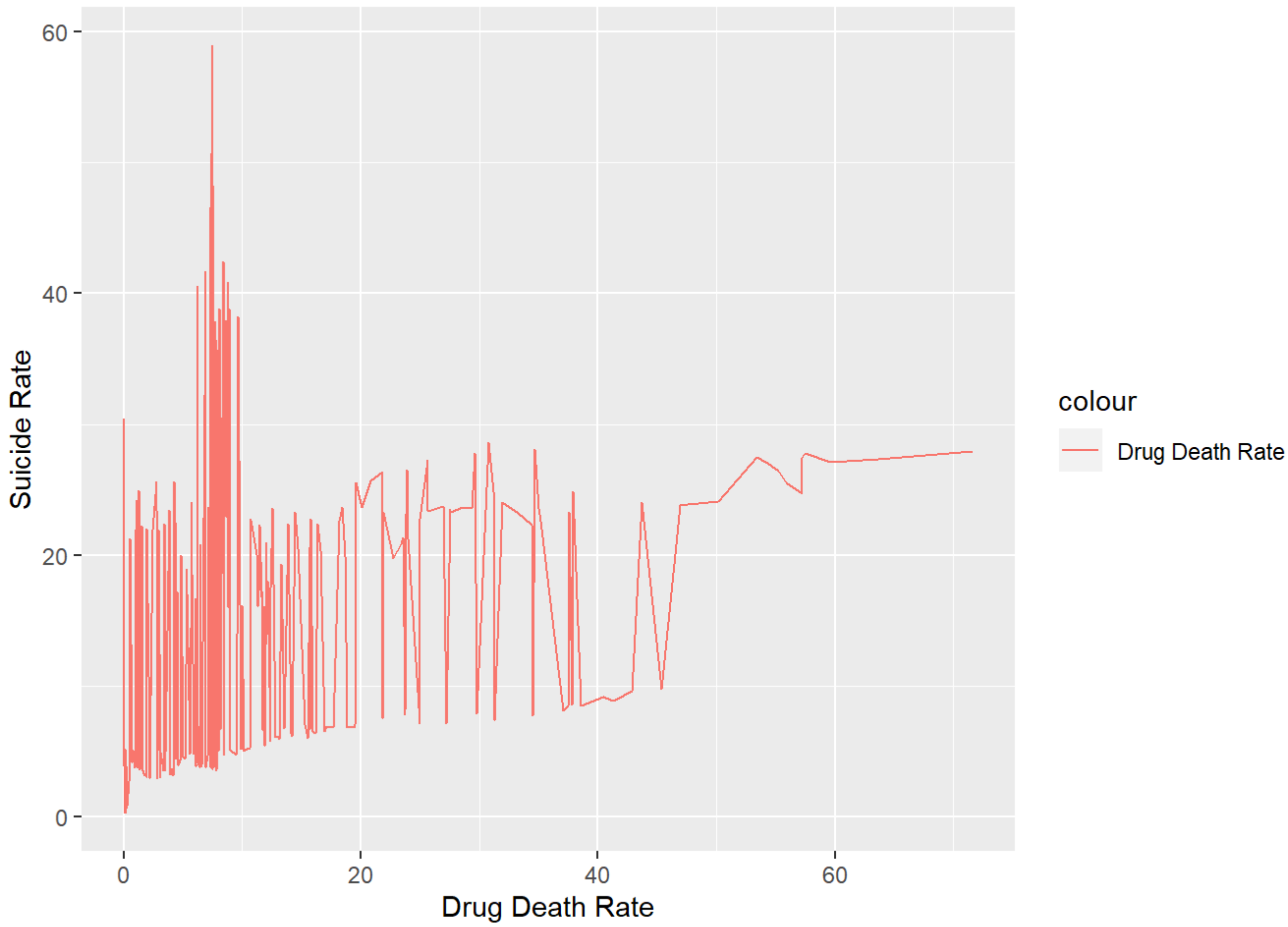
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.260   3.973   6.890  13.820  23.305  58.950
```

```
ggplot(data) + ylab("") + xlab("Suicide Rate")+ ggtitle("Lineplot, GDP and HDI vs Suicide Rate c
oded by color") +
      geom_line(aes(x=data$suicides.100k.pop, y= data$HDI.for.year, color = "HDI for year")) +
     geom_line(aes(x=data$suicides.100k.pop, y= data$gdp_per_capita...., color = "GDP Per Capit
a"))
```



Investigating Suicide rate and the drug death rate

```
ggplot(data)+ geom_line(aes(x= data$drug_death_rate, y= data$suicides.100k.pop, color = "Drug De
ath Rate")) +
ylab("Suicide Rate") + xlab("Drug Death Rate")
```




```
var(data$suicides.100k.pop, data$drug_death_rate)
```

```
## [1] 51.17861
```

```
cor(data$suicides.100k.pop, data$drug_death_rate)
```

```
## [1] 0.2891455
```

```
cor(data$population, data$drug_death_rate)
```

```
## [1] 0.6774055
```

```
cov(data$suicides.100k.pop, data$drug_death_rate)
```

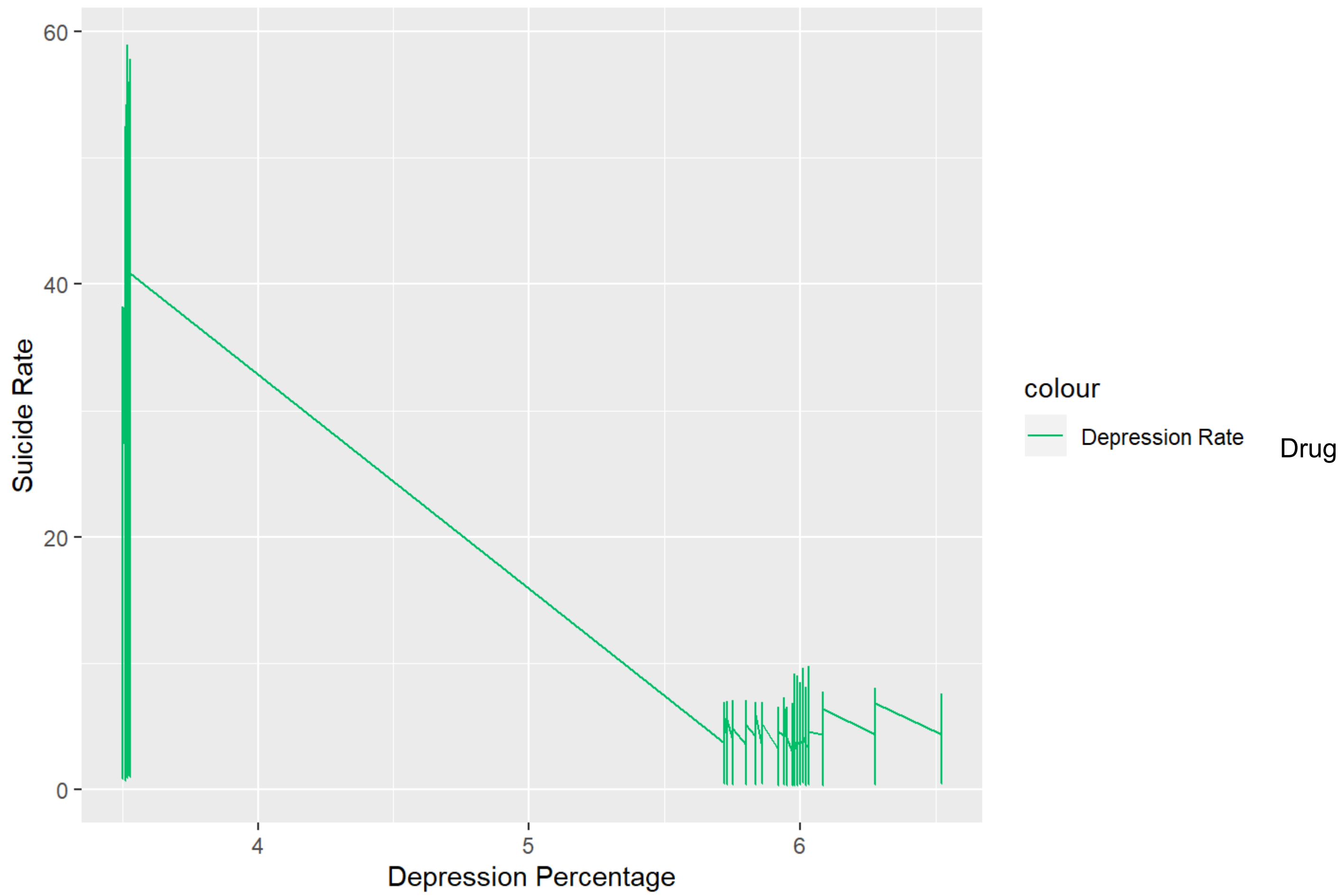
```
## [1] 51.17861
```

```
summary(data$drug_death_rate)
```

```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.00    0.20    6.55   10.08   12.03   71.60
```

Investigating Suicide rate and the Depression Percentage

```
ggplot(data)+ geom_line(aes(x= data$depression_percentage, y= data$suicides.100k.pop, color = "D
epression Rate")) +
ylab("Suicide Rate") + xlab("Depression Percentage") + scale_color_manual(values=PALETTE[6])
```



death rate and depression rate are slightly negitively correlated as well as suicide rates and drepression percentages.

```
var(data$suicides.100k.pop, data$depression_percentage)
```

```
## [1] -11.12504
```

```
cor(data$suicides.100k.pop, data$depression_percentage)
```

```
## [1] -0.6878586
```

```
cor(data$depression_percentage, data$drug_death_rate)
```

```
## [1] -0.1956575
```

```
cov(data$suicides.100k.pop, data$depression_percentage)
```

```
## [1] -11.12504
```

```
summary(data$depression_percentage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   3.500   3.510   4.625   4.729   5.980   6.519
```

```
ggplot(data) + ylab("") + ggtitle("Lineplot, of Depression Rate, and Drug Death Rate v.s Suicide Rate") +
     geom_line(aes(x= data$suicides.100k.pop, y= data$depression_percentage, lty="Dashed", color = "Depression")) +
     geom_line(aes(x=data$suicides.100k.pop, y= data$drug_death_rate, lty="Solid", color = "Drug Death")) +
     #geom_line(aes(x=data$suicides.100k.pop, y= data$population, lty="x9", color = "Population")) +
     scale_linetype_manual(values=c("solid","longdash")) + xlab("Suicide Rate")
```