Suicide and Smartphones, edX Harvard Data Science Capstone 2

Adam Ausperk

1/5/2019

Executive Summary

Suicide rates have increased dramatically since 1999, especially amongst teens and those in their mid-20s, according to the CDC (1). In contrast, McKeown, et. al reported in 2006 that from roughly 1989 to 2002 the suicide rate for the age group of 15-24 significantly declined (2).

The Pew Research Center (PRC) has compiled and published survey data representing cellphone and smartphone use over this general timeframe, amongst other data points (3). One of the roles of the Pew Research Center is to document cultural changes in the United States of America (U.S.). In 2012, the PRC began to report ownership rates of smartphones. It is possible that smartphone use is the most significant cause of the rapid increase in suicide amongst the 10-24 age group that was reported in the 2019 study by the CDC.

The focus of this report is on the 15-24 age group for two reasons. The first of these is that this is the only youth age group that is directory comparable between the 1970-2002 report by McKeown, et. al and published in the American Journal of Public Health and the 2019 study reported directly by the CDC. The second reason that the 15-24 age group was chosen to be studied as representative of U.S. youth suicide is that fortunately, there is a small number of youth in the 10-14 age group that commits suicide. However, their suicide rates have also increased dramatically in the smartphone era.

Suicide is strongly associated with severe mental illness. In 2016, the National Alliance on Mental Illness published an article that stated that "90% of children and adolescents who die by suicide live with a mental health condition" (4).

It seems strongly plausible then that suicide is in fact a strong indicator of the overall quality of the mental health of a population, with the exception of war-torn countries. In a country with peace and security within its borders, with no imminent risk of violent death or personal intimidation, there is no functional and reasonable reason for an individual to kill oneself. It is then argued in this report that the suicide rate in the United States is primarily representative of the mental health of the population. A "copycat" effect for suicides is acknowledged as being a probable minor contributor to any increase in suicide rate amongst a cultural cohort.

Data has been obtained that will be used to attempt to find a cause to the recent increase in suicide rates. To not restrict the analysis to smartphones only, data has also been obtained for cellphone use (simple cellphones before smartphones), computer and internet ownership in the home before cell phones and smartphones were available to the U.S. public, and stock market data (which attempts to signify changes in the economic environment).

Economic recessions have many different and usable definitions, in this report the economic environment is determined to be represented by the average annual stock market close amount of the Dow, the S & P 500, and the NASDAQ stock market indexes. The stock market indexes are used although it is known that negative GDP or GNP growth for two quarters is the "normal" way to define a recession.

Regarding the sources of data used in this report, suicide rates were obtained from the CDC, as well as from the well cited study by McKeown (et. al). All suicide rates used in this report then have the CDC as their source, have virtually identical defnitions as to how they were tabulated, are age adjusted, and represent suicides per 100,000 people. Cell phone use rates from 1999 and forward were compiled from multiple publications from the Pew Research Center. Pre-phone era computer and internet use rates were obtained from the U.S. Census (5). Stock market close data was compiled from Yahoo Finance (6).

The objective of this report is to see how significant the hypothesis is that smart phones are the cause of the recent increase in suicide rates in the United States. If smart phones are not the cause, what is the cause?

Methods/Analysis

This report begins with the already compiled data in both "wide" and "tidy" format.This data has already been cleaned and processed. The steps taken to clean and process this data are not included in this report because of the high number of sources from which it was obtained. This is to assist in the brevity and readability of this report. A separate report can be accessed which lists the steps for cleaning and compiling these separate reports of data.

```r
# installing the required packages

if(!require(readxl)) install.packages("readxl", repos = "http://cran.us.r-project.org")
if(!require(tidyverse)) install.packages("tidyverse", repos = "http://cran.us.r-project.org")
if(!require(caret)) install.packages("caret", repos = "http://cran.us.r-project.org")
if(!require(gridExtra)) install.packages("gridExtra", repos = "http://cran.us.r-project.org")

# importing the wide dataset and creating the tidy dataset
getwd()
```

```
## [1] "C:/Users/aausp/Documents/R/Capstone 2 - GitHub/Upload"
```

```r
wide <- read_excel("wide_1-7-20.xlsx", col_types = "numeric")
wide <- round(wide,2)
tidy <- gather(wide, "Category", "Rate", -1, na.rm = TRUE)
tidy$Category <- as.character(tidy$Category)
tidy$Rate <- as.numeric(tidy$Rate)
str(tidy)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    270 obs. of  3 variables:
##  $ Year    : num  2000 2001 2002 2003 2004 ...
##  $ Category: chr  "15-19_rate" "15-19_rate" "15-19_rate" "15-19_rate" ...
##  $ Rate    : num  8 7.9 7.3 7.2 8.1 7.5 7.1 6.7 7.2 7.5 ...
```

```r
str(wide)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    50 obs. of  14 variables:
##  $ Year           : num  1970 1971 1972 1973 1974 ...
##  $ 15-19_rate     : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ 20-24_rate     : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ 10-14_rate     : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Owns_Cellphone : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Owns_Smartphone: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ 15-24          : num  8.8 9.4 10.2 10.6 10.9 11.8 11.7 13.6 12.4 12.7 ...
##  $ 25-44          : num  15.4 15.3 15.7 15.6 16.2 16.8 16.1 17.3 16.3 16.2 ...
##  $ 45-64          : num  20.6 20.1 20.1 19.9 19.6 20.1 19.6 19.1 17.6 16.8 ...
##  $ >65            : num  20.8 21.3 21 20.3 19.3 19.1 19.8 20.3 19.9 19.1 ...
##  $ Avg_SM_Use     : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Markets        : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Comp_athome    : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Internet_athome: num  NA NA NA NA NA NA NA NA NA NA ...
```

```r
# the tidy dataset

head(tidy, n = 10)
```

```
## # A tibble: 10 x 3
##     Year Category    Rate
##    <dbl> <chr>      <dbl>
##  1  2000 15-19_rate   8
##  2  2001 15-19_rate   7.9
##  3  2002 15-19_rate   7.3
##  4  2003 15-19_rate   7.2
##  5  2004 15-19_rate   8.1
##  6  2005 15-19_rate   7.5
##  7  2006 15-19_rate   7.1
##  8  2007 15-19_rate   6.7
##  9  2008 15-19_rate   7.2
## 10  2009 15-19_rate   7.5
```

```r
colnames(tidy)
```

```
## [1] "Year"     "Category" "Rate"
```

```r
dim(tidy)
```

```
## [1] 270   3
```

```r
# the wide dataset

head(wide, n = 10)
```

```
## # A tibble: 10 x 14
##     Year `15-19_rate` `20-24_rate` `10-14_rate` Owns_Cellphone
##    <dbl>        <dbl>        <dbl>        <dbl>          <dbl>
##  1  1970           NA           NA           NA             NA
##  2  1971           NA           NA           NA             NA
##  3  1972           NA           NA           NA             NA
##  4  1973           NA           NA           NA             NA
##  5  1974           NA           NA           NA             NA
##  6  1975           NA           NA           NA             NA
##  7  1976           NA           NA           NA             NA
##  8  1977           NA           NA           NA             NA
##  9  1978           NA           NA           NA             NA
## 10  1979           NA           NA           NA             NA
## # ... with 9 more variables: Owns_Smartphone <dbl>, `15-24` <dbl>,
## #   `25-44` <dbl>, `45-64` <dbl>, `>65` <dbl>, Avg_SM_Use <dbl>,
## #   Markets <dbl>, Comp_athome <dbl>, Internet_athome <dbl>
```

```r
colnames(wide)
```

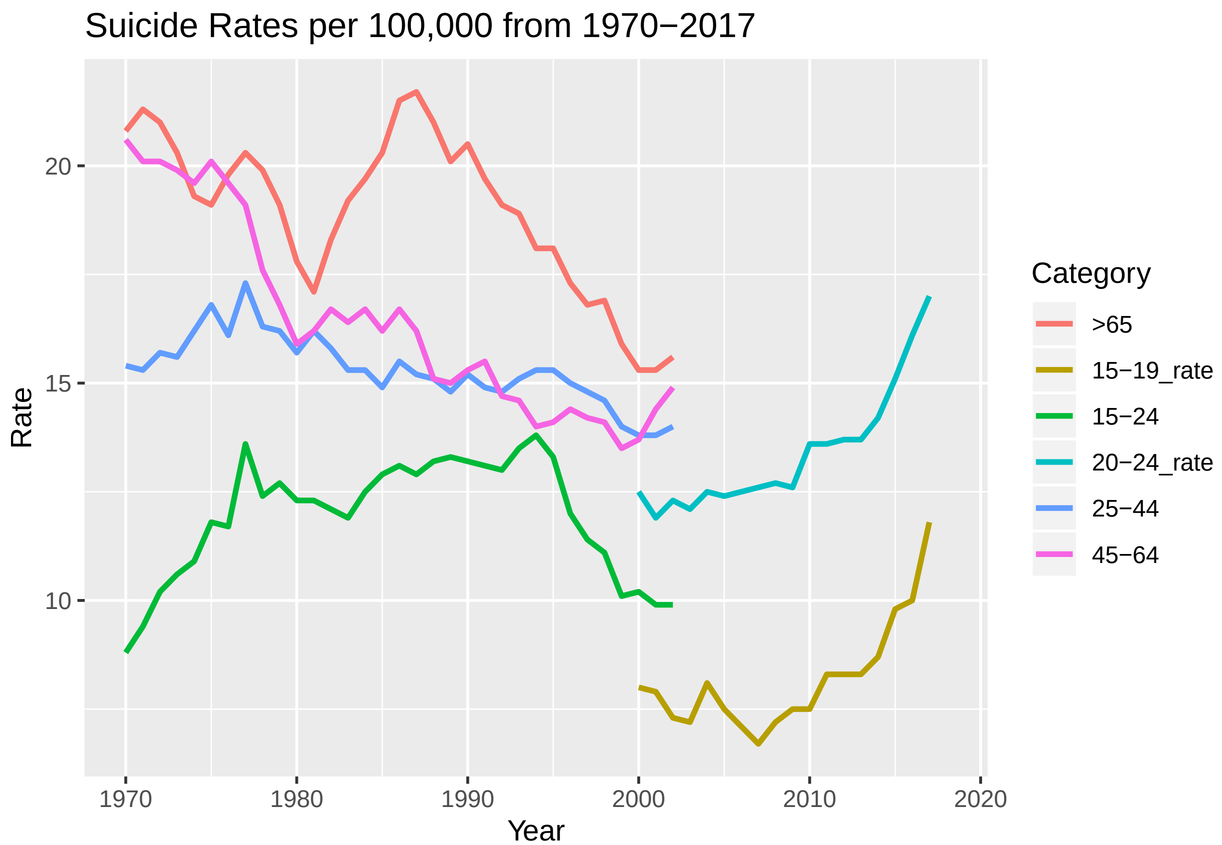
```
##  [1] "Year"            "15-19_rate"      "20-24_rate"
##  [4] "10-14_rate"      "Owns_Cellphone"  "Owns_Smartphone"
##  [7] "15-24"           "25-44"           "45-64"
## [10] ">65"             "Avg_SM_Use"      "Markets"
## [13] "Comp_athome"     "Internet_athome"
```

```r
dim(wide)
```

```
## [1] 50 14
```

Visualizing Suicide Rates Over Time and Discussion

```
# Plotting suicide rates to begin analysis

a_plot <- tidy %>%
  filter(Category %in% c("15-19_rate", "20-24_rate", "15-24", "25-44", "45-64", ">65")) %>%
           ggplot(aes(x = Year, y = Rate, color = Category)) +
           geom_line(size = 1) +
  ggtitle("Suicide Rates per 100,000 from 1970-2017") +
  xlim(1970,2018)
a_plot
```



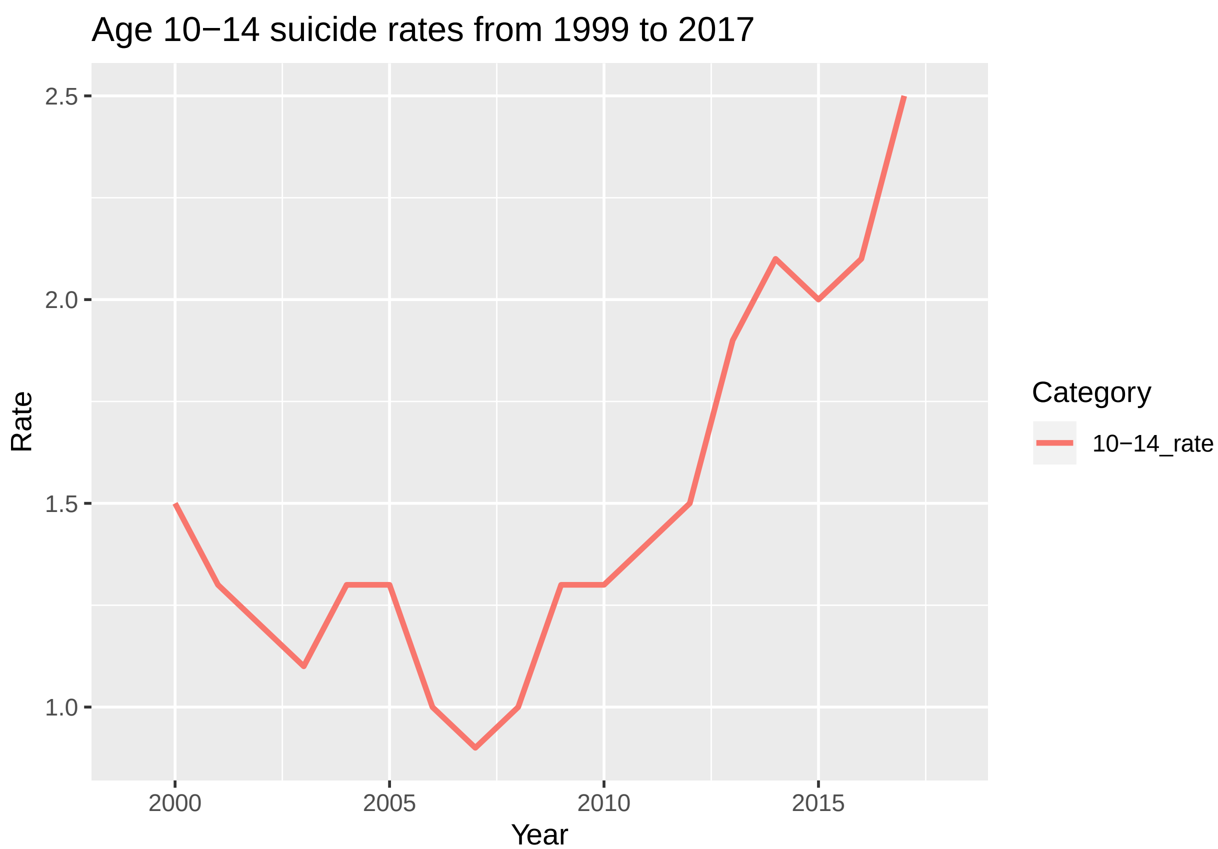
As can be seen in the plot above, the suicide rates for all age groups was trending down beginning around 1994. The most significant ages groups that trended down in rate from 1994 to 2002 was the 15-24 and >65 age groups. It should be noted that in the 2006 study by McKeown, et. al (that reported rates from 1970-2002) and the 2019 study by the CDC used the same source for the suicide rates reported and a virtually identical definition.

The categories in the plot above names "15-19_rate" and "20-24_rate" for the 1999 to 2017 data are both included to be comparable to the 1970-2002 15-24 age group.

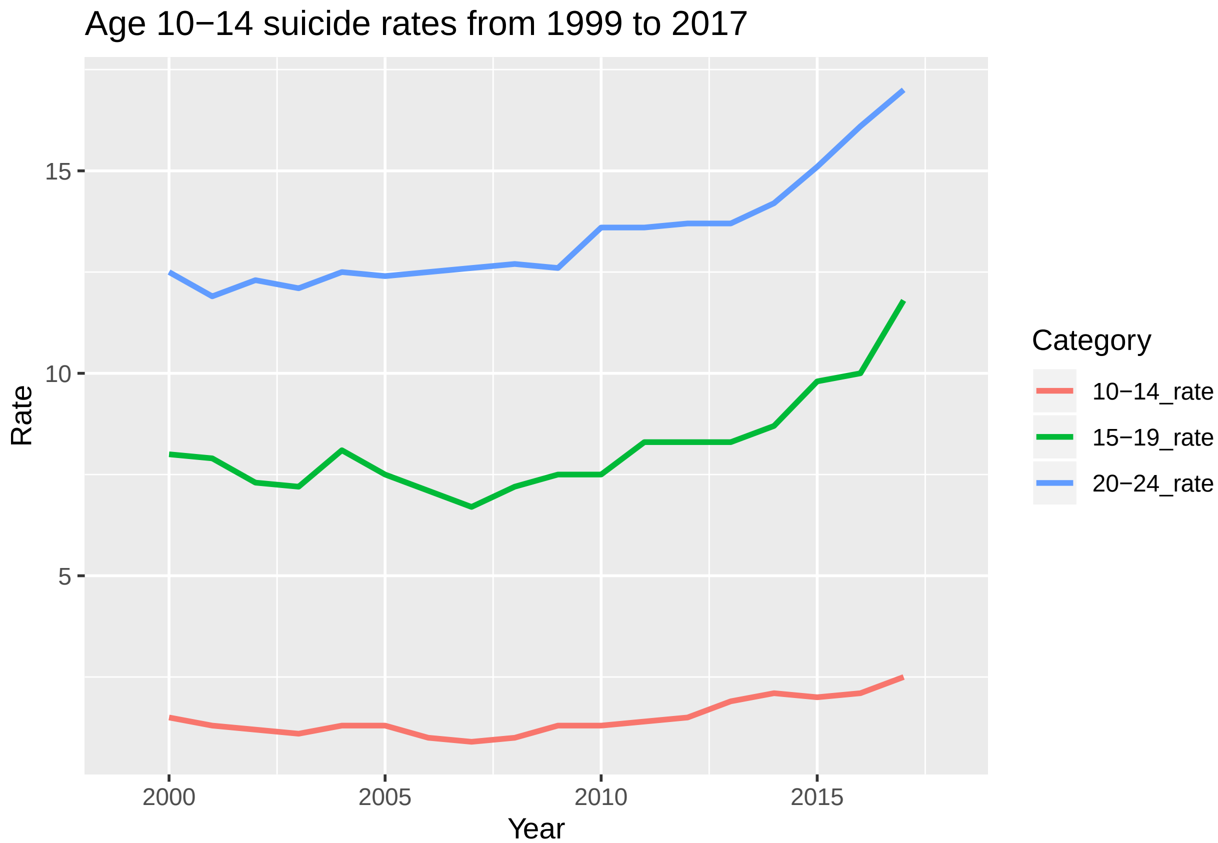
When focusing on the 15-24 age group, the plot above also shows the increase in suicide rates per 100,000 people beginning at around 2002 for the 20-24 age group and beginning at around 2007 for the 15-19 age group. Both of these age groups have a much sharper increase in suicide rates beginning in 2012-2013.

The first plot above which contains illustrates that the two sets of suicide rate data are comparable to each other, as the mean of the 1999-2017 data seems to be very similar to the 15-24 age group rate that was provided as a part of the 1970-2002 analysis.

```r
# plotting just the 10-14 rate

a_1plot <- tidy %>%
  filter(Category %in% c("10-14_rate")) %>%
           ggplot(., aes(x = Year, y = Rate, color = Category)) +
           geom_line(size = 1) +
  ggtitle("Age 10-14 suicide rates from 1999 to 2017") +
  xlim(1999,2018)
a_1plot
```




```r
# plotting the 10-14, 15-19, and 20-24 suicide rates post 1999

a_2plot <- tidy %>%
  filter(Category %in% c("10-14_rate", "15-19_rate", "20-24_rate")) %>%
           ggplot(., aes(x = Year, y = Rate, color = Category)) +
           geom_line(size = 1) +
  ggtitle("Age 10-14 suicide rates from 1999 to 2017") +
  xlim(1999,2018)
a_2plot
```



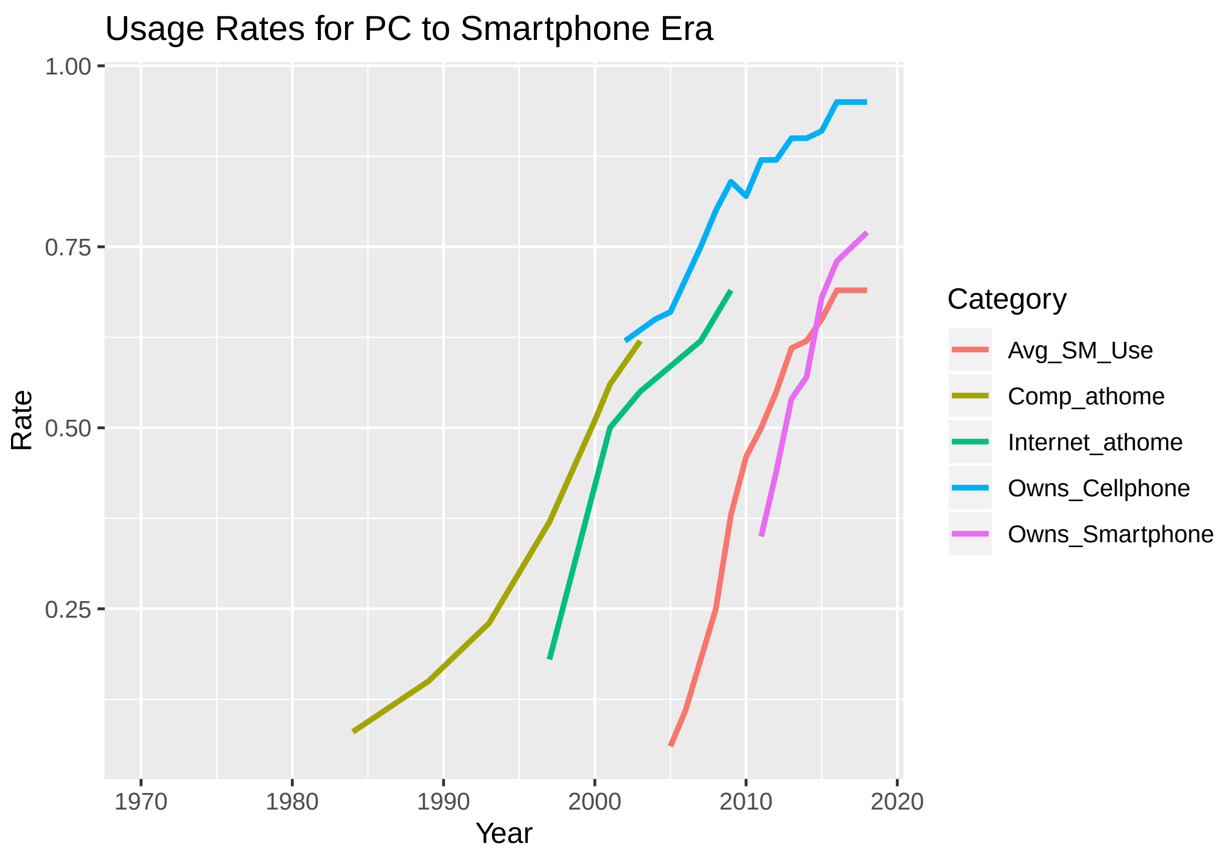
The "10-14_rate" has been included separately in the plots above to allow the reader to review their trend and timing in only this subsection of this report. As noted in the Executive Summary section, the 10-14 rate will not be analyzed in this study because it does not have a data comparison in the 2006 McKeown, et. al study. As can be seen from the plots, the rate itself is much lower than the 15-19 and 20-24 rate. But it has increased steadily beginning in 2010 from its average looking back to 2000, with a substantial increase beginning in 2010 and a sharp increase beginning in 2014.

Analyzing Computer, Internet, Social Media, Cell Phone, and Smart Phone Use as Potential Independent Causes Separately

Overview

As noted in the Executive Summary section, data was obtained that separately describes the rate of computer and internet use in U.S. households, social media use, and cell and smart phone ownership. These will be plotted below, followed by a discussion. The hypothesis that smart phone ownership is the cause of the increase in youth suicide rates implies that electronics themselves may also cause suicide. To isolate smart phone use from other types of electronic/media use, all of the data in the plot below is necessary.

```
# plotting usage rates for all electronic media types

b_plot <- tidy %>%
  filter(Category %in% c("Owns_Cellphone", "Owns_Smartphone", "Avg_SM_Use", "Comp_athome", "Internet_at
           ggplot(aes(x = Year, y = Rate, color = Category)) +
           geom_line(size = 1) +
  ggtitle("Usage Rates for PC to Smartphone Era") +
  xlim(1970, 2018)
b_plot
```



The plot above illustrates that computer use in U.S. homes began in about 1980 or so (extrapolating the brown line backwards in time), and reached 62% by around 2003. The rate of internet use at home was about 17% in 1997, increasing to 65% in 2009. The increase in internet use in the home was faster than the increase of computer use in the home.

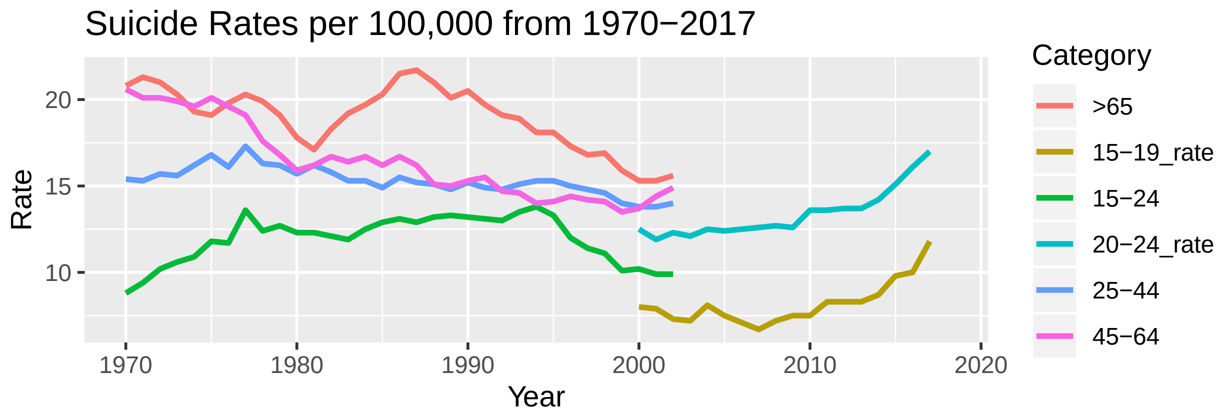
Regarding mobile phones, cell phones (not internet based smart phones) were owned by 62% of adults in 2002. Internet based phones (deemed "smart phones" for this analysis) were adopted at about 37% in 2011, increasing to 80% in 2019. The PRC definition of cell phones includes smart phones after smart phones were available to the public. The overall mobile phone ownership rate in the U.S. is at about 95% in 2019 (blue line for cell phone ownership).

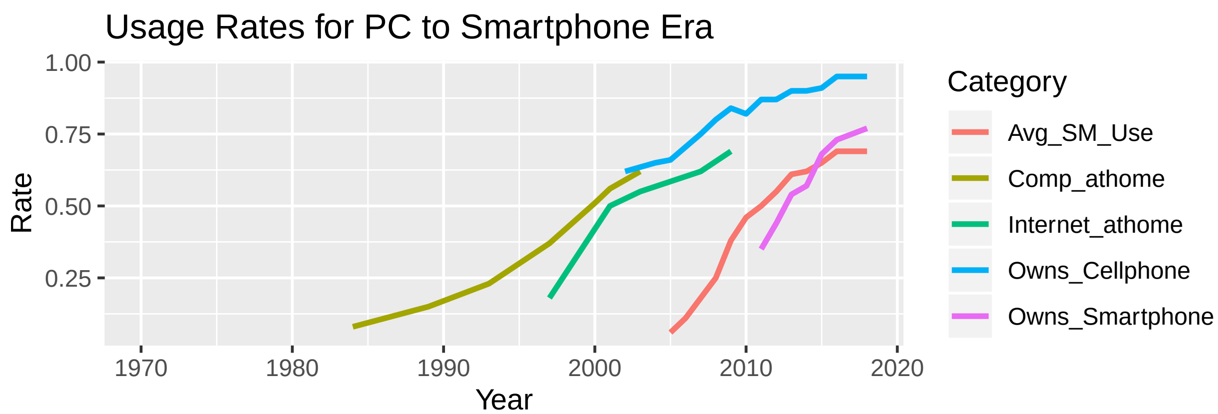
Social media use (with providers such as Twitter, Facebook, etc.) started in 2005, with about 5% of Americans using at least one of these services. By 2019, around 72% of Americans were using at least one of these services.

Overall comparative analysis of these time series plots

The two plots above will now be compared to each other side by side.

```
# arranging plots for comparison

grid.arrange(a_plot,b_plot)
```





The plots above illustrate that if any electronic media is the primary cause of the increase in suicide rates amongst the youth of the United States, that social media and smartphones are the most significant reason.

The reason for that statement is that suicide rates began to steadily increase for the 15-24 age group around the year 2005. In 2005, computer and internet use in the home were already established. Cell phone use was already established as well in 2005 in about 65% of the population.

Isolate each potential cause and run linear regression for each for 15-24 age group

This study will focus on the 15-24 age group. The 15-24 age group includes the most suicides of the 10-24 age group. Fortunately, few youths in the 10-14 age group commit suicide in the U.S. For the purposes of this study then, the 15-24 age group will be considered to be representative of the youth suicide trends in the U.S.

This section will isolate each potential electronic media related cause of the suicide rate of the 15-24 age group. This differs from the above discussion by including the suicide rate for U.S. children that commit suicide while 15-14 years of age.

Combining the 2 15-24 age group datasets to one

```
# combining the 2 15-24 vectors to one

ratepost99 <- (wide$`15-19_rate` + wide$`20-24_rate`)/2
wide <- cbind(wide, ratepost99)

a <- wide$`15-24`
a
```

```
##  [1]  8.8  9.4 10.2 10.6 10.9 11.8 11.7 13.6 12.4 12.7 12.3 12.3 12.1 11.9
```

```
## [15] 12.5 12.9 13.1 12.9 13.2 13.3 13.2 13.1 13.0 13.5 13.8 13.3 12.0 11.4
## [29] 11.1 10.1 10.2  9.9  9.9   NA   NA   NA   NA   NA   NA   NA   NA   NA
## [43]   NA   NA   NA   NA   NA   NA   NA   NA
```

```
b <- wide$`ratepost99`
b
```

```
##  [1]    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
## [12]    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
## [23]    NA    NA    NA    NA    NA    NA    NA    NA 10.25  9.90  9.80
## [34]  9.65 10.30  9.95  9.80  9.65  9.95 10.05 10.55 10.95 11.00 11.00
## [45] 11.45 12.45 13.05 14.40    NA    NA
```

```
c <- data.frame(a,b)
c[['d']] <- rowMeans(c,na.rm=TRUE)
c
```

```
##       a     b      d
## 1   8.8    NA  8.800
## 2   9.4    NA  9.400
## 3  10.2    NA 10.200
## 4  10.6    NA 10.600
## 5  10.9    NA 10.900
## 6  11.8    NA 11.800
## 7  11.7    NA 11.700
## 8  13.6    NA 13.600
## 9  12.4    NA 12.400
## 10 12.7    NA 12.700
## 11 12.3    NA 12.300
## 12 12.3    NA 12.300
## 13 12.1    NA 12.100
## 14 11.9    NA 11.900
## 15 12.5    NA 12.500
## 16 12.9    NA 12.900
## 17 13.1    NA 13.100
## 18 12.9    NA 12.900
## 19 13.2    NA 13.200
## 20 13.3    NA 13.300
## 21 13.2    NA 13.200
## 22 13.1    NA 13.100
## 23 13.0    NA 13.000
## 24 13.5    NA 13.500
## 25 13.8    NA 13.800
## 26 13.3    NA 13.300
## 27 12.0    NA 12.000
## 28 11.4    NA 11.400
## 29 11.1    NA 11.100
## 30 10.1    NA 10.100
## 31 10.2 10.25 10.225
## 32  9.9  9.90  9.900
## 33  9.9  9.80  9.850
## 34   NA  9.65  9.650
## 35   NA 10.30 10.300
## 36   NA  9.95  9.950
## 37   NA  9.80  9.800
## 38   NA  9.65  9.650
```

```
## 39   NA  9.95  9.950
## 40   NA 10.05 10.050
## 41   NA 10.55 10.550
## 42   NA 10.95 10.950
## 43   NA 11.00 11.000
## 44   NA 11.00 11.000
## 45   NA 11.45 11.450
## 46   NA 12.45 12.450
## 47   NA 13.05 13.050
## 48   NA 14.40 14.400
## 49   NA    NA    NaN
## 50   NA    NA    NaN
```

```r
c <- as.tibble(c[,3])
colnames(c) <- c("All_15_24_Rates")
wide <- cbind(wide, c)
dim(wide)
```

```
## [1] 50 16
```

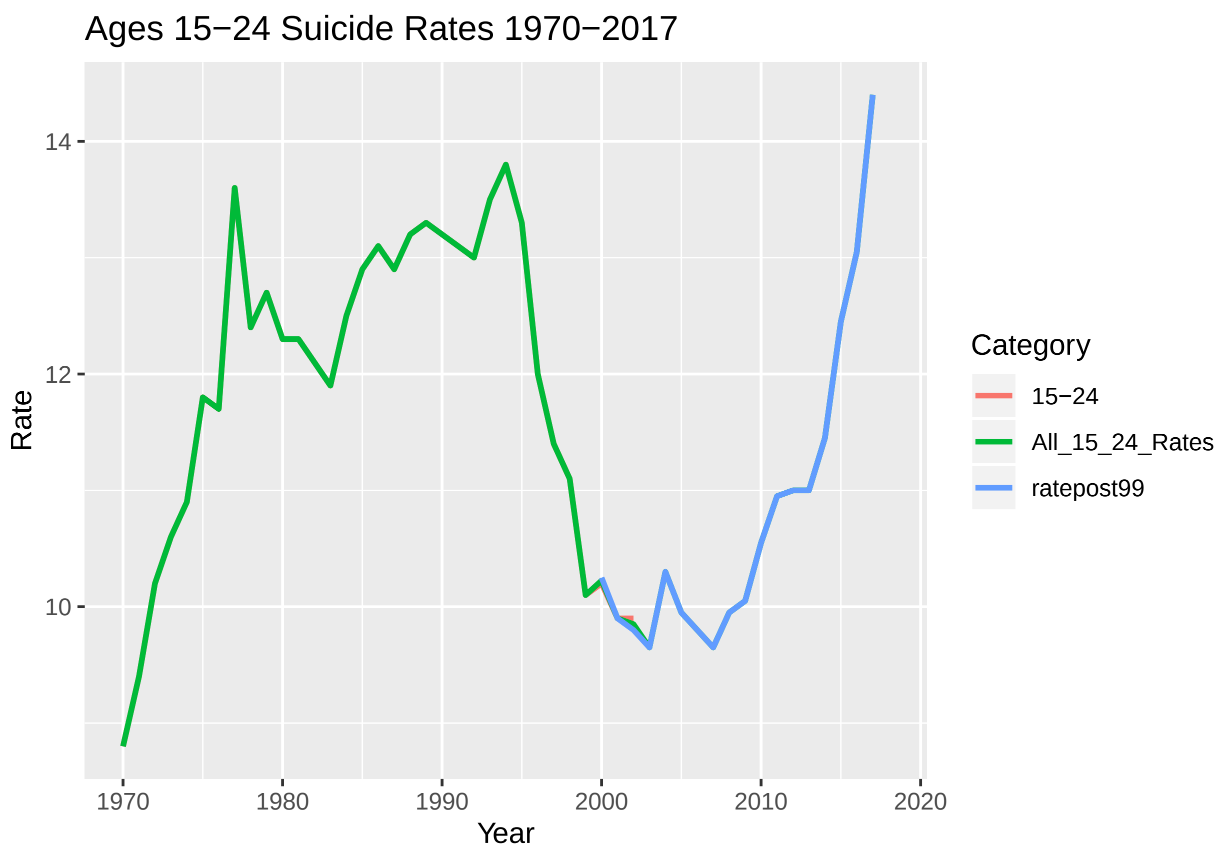
```r
wide$All_15_24_Rates
```

```
##  [1]  8.800  9.400 10.200 10.600 10.900 11.800 11.700 13.600 12.400 12.700
## [11] 12.300 12.300 12.100 11.900 12.500 12.900 13.100 12.900 13.200 13.300
## [21] 13.200 13.100 13.000 13.500 13.800 13.300 12.000 11.400 11.100 10.100
## [31] 10.225  9.900  9.850  9.650 10.300  9.950  9.800  9.650  9.950 10.050
## [41] 10.550 10.950 11.000 11.000 11.450 12.450 13.050 14.400    NaN    NaN
```

```r
# adding the new column to the tidy dataset

tidy <- gather(wide, "Category","Rate", -1, na.rm=TRUE)
colnames(wide)
```

```
##  [1] "Year"             "15-19_rate"       "20-24_rate"
##  [4] "10-14_rate"       "Owns_Cellphone"   "Owns_Smartphone"
##  [7] "15-24"            "25-44"            "45-64"
## [10] ">65"              "Avg_SM_Use"       "Markets"
## [13] "Comp_athome"      "Internet_athome"  "ratepost99"
## [16] "All_15_24_Rates"
```

```r
# plotting the new column with the 2 original 15-24 columns

d_plot <- tidy %>%
  filter(Category %in% c("15-24", "ratepost99", "All_15_24_Rates")) %>%
          ggplot(aes(x = Year, y = Rate, color = Category)) +
          geom_line(size = 1) +
  ggtitle("Ages 15-24 Suicide Rates 1970-2017") +
  xlim(1970, 2018)
d_plot
```



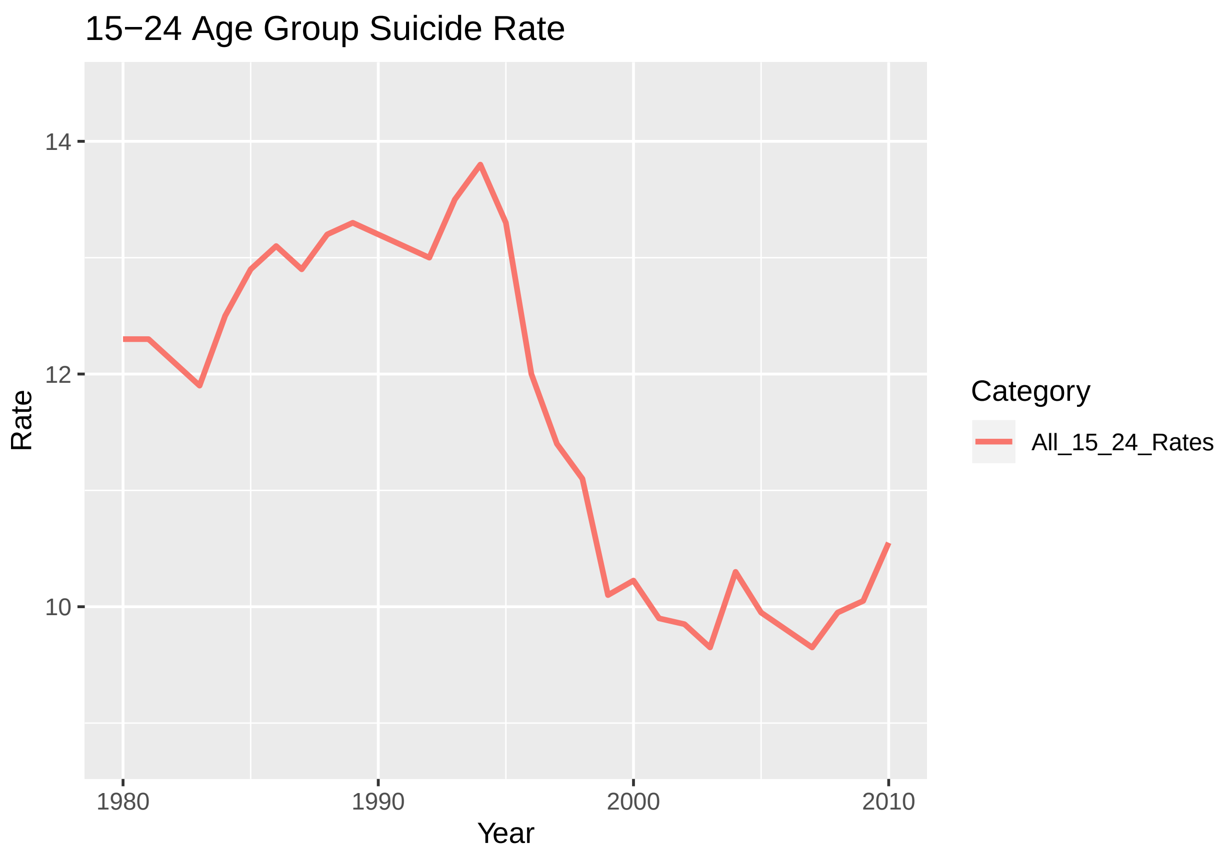
Isolate individual variables as potential causes

Computer Ownership in U.S. Households

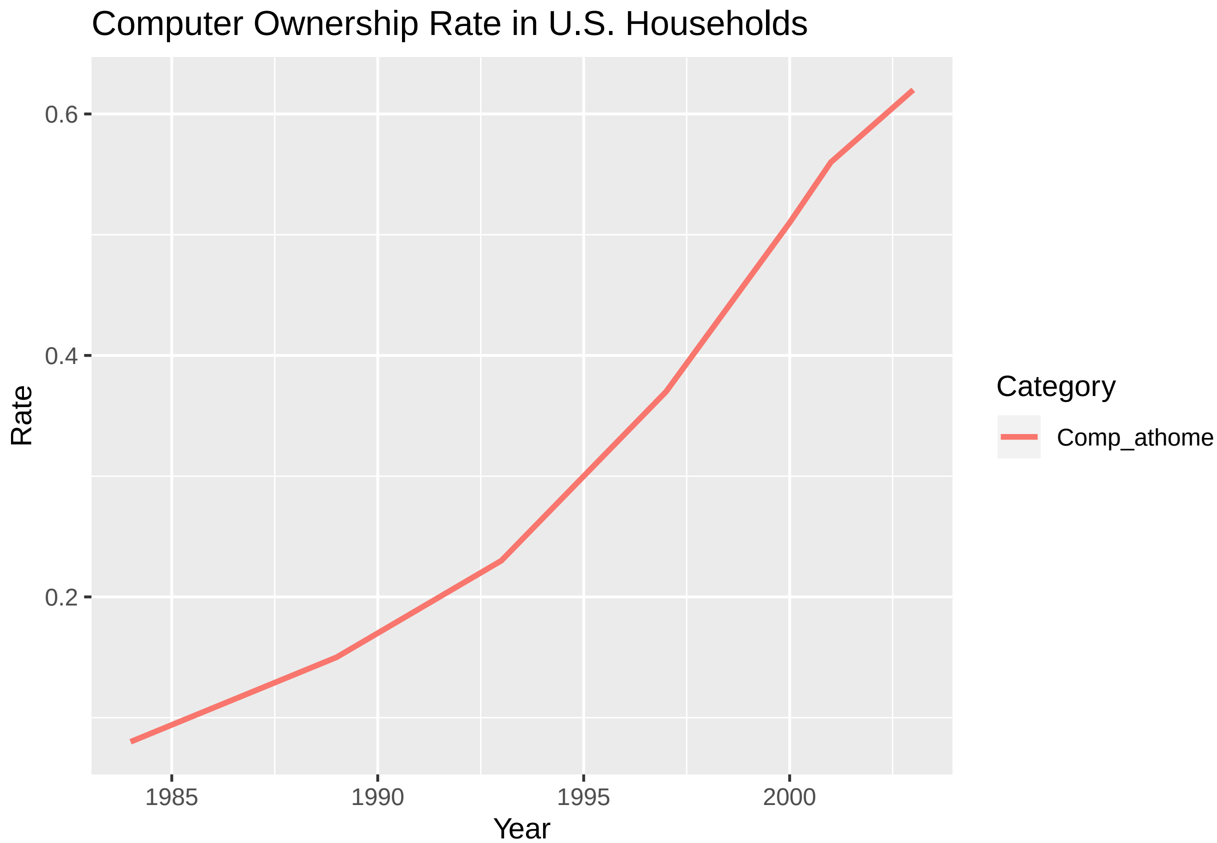
```
# calculating correlation for home computer ownership and 15-24 suicide rate

cor(wide$All_15_24_Rates,wide$Comp_athome, use = "pairwise.complete.obs")
```

```
## [1] -0.9243424
```

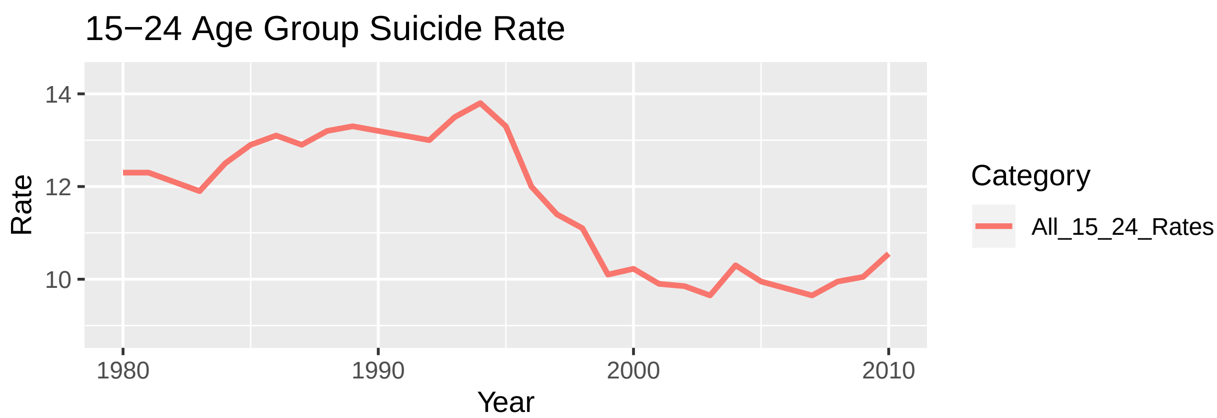
```
# plotting just 15-24 S.R.

e_plot <- tidy %>%
  filter(Category %in% c("All_15_24_Rates")) %>%
           ggplot(aes(x = Year, y = Rate, color = Category)) +
           geom_line(size = 1) +
  ggtitle("15-24 Age Group Suicide Rate") +
  xlim(1980,2010)
e_plot
```

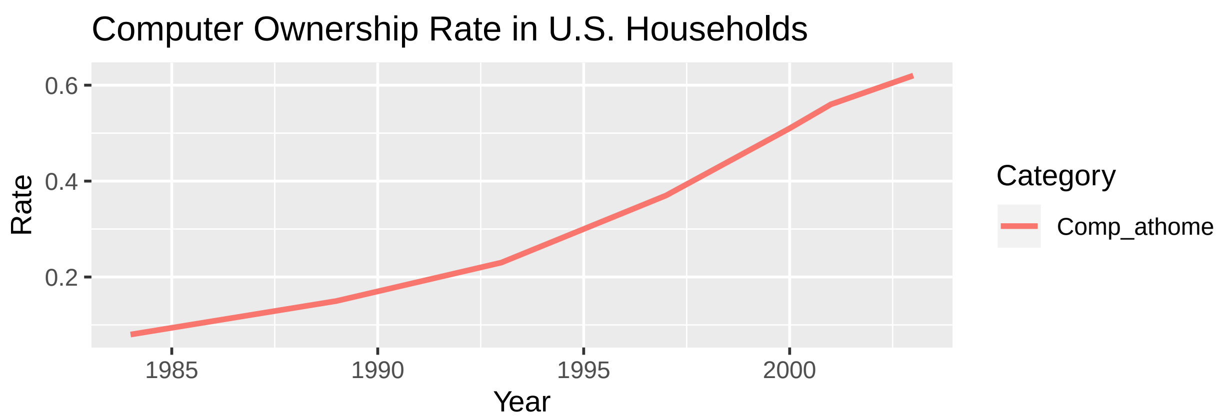



```
# plotting just computer ownership

f_plot <- tidy %>%
  filter(Category %in% c("Comp_athome")) %>%
          ggplot(aes(x = Year, y = Rate, color = Category)) +
          geom_line(size = 1) +
  ggtitle("Computer Ownership Rate in U.S. Households")
f_plot
```




```
# arranging plots for comparison

grid.arrange(e_plot, f_plot)
```






```r
# computing linear regression statistics for these variables

fit1 <- wide %>% lm(All_15_24_Rates ~ Comp_athome, data = .)
summary(fit1)
```

```
##
## Call:
## lm(formula = All_15_24_Rates ~ Comp_athome, data = .)
##
## Residuals:
##         15         20         24         28         31         32         34
## -0.981406  0.314838  1.081975 -0.025537 -0.208048 -0.178587 -0.003235
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  14.0485     0.5366  26.181 1.52e-06 ***
## Comp_athome  -7.0892     1.3087  -5.417   0.0029 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6795 on 5 degrees of freedom
##   (43 observations deleted due to missingness)
## Multiple R-squared:  0.8544, Adjusted R-squared:  0.8253
## F-statistic: 29.34 on 1 and 5 DF,  p-value: 0.002902
```

Results of Computers at Home Analysis

There is a negative correlation of -.92 for these two variables. Correspondingly, there is a negative slope of -7.09 for the regression calculation. Home ownership of computers arrived at 50% saturation in the United States in the year 2000. As computer ownership increased, 15-24 S.R. decreased.

It is notable to some degree that at this 50% saturation point the fall in 15-24 S.R. stoppped in 1999. The beginning year of this analysis, 1970, was at 8.8 per 100,000, when in 2000, the rate was 10.225. The drop in S.R. then could have traveled lower.

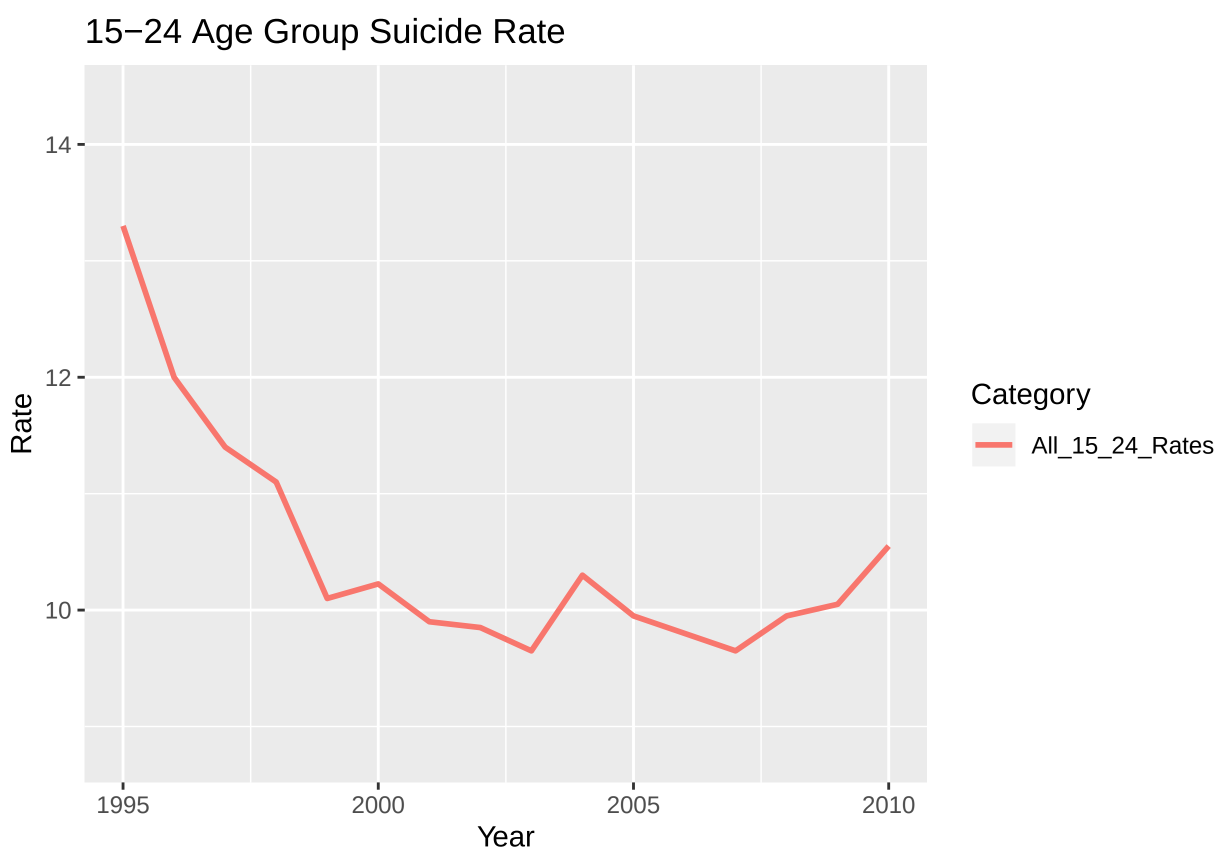
With this data, it is not at all a suggestion that could be made that computer ownership (with no internet access) caused a significant increasing trend in youth suicide.

Internet access in U.S. Households

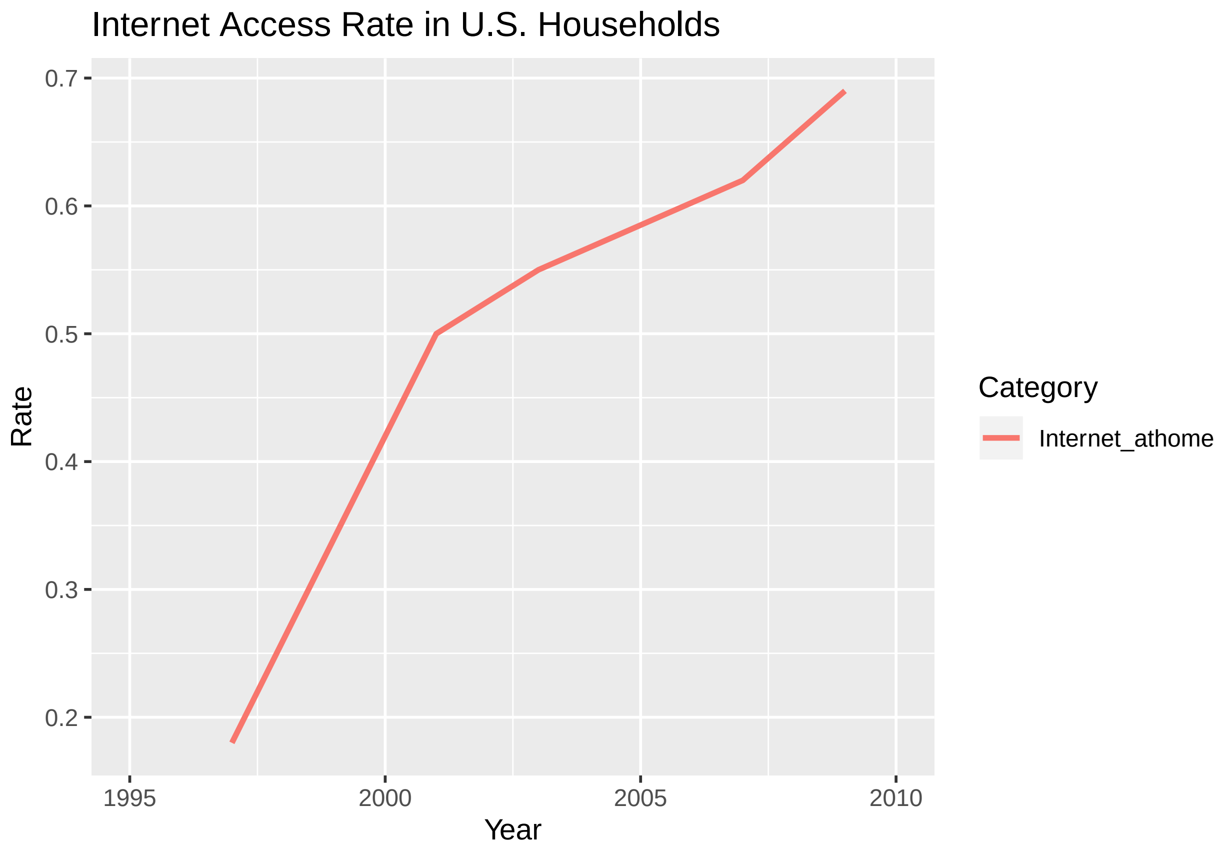
```
# calculating correlation for household internet access rate and 15-24 suicide rate

cor(wide$All_15_24_Rates,wide$Internet_athome, use = "pairwise.complete.obs")
```

```
## [1] -0.8676035
```

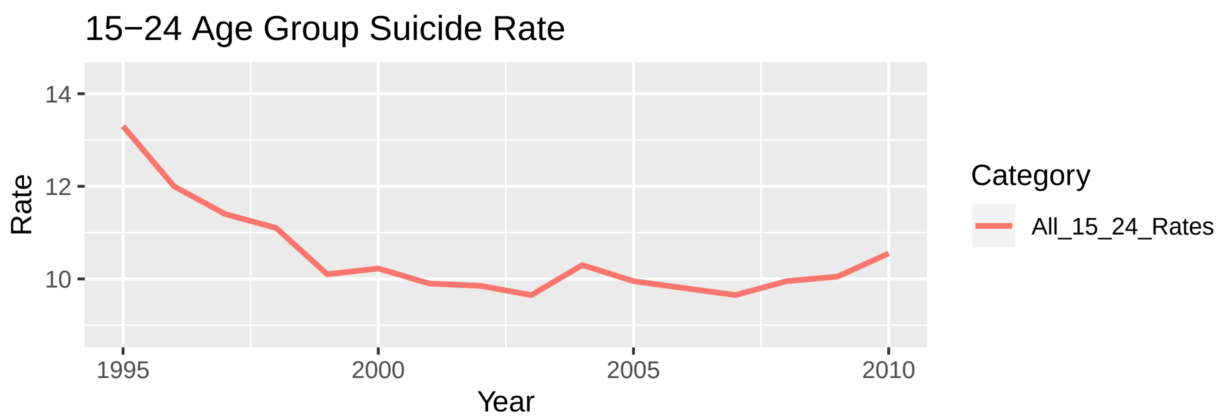
```
# plotting just 15-24 S.R.

g_plot <- tidy %>%
  filter(Category %in% c("All_15_24_Rates")) %>%
         ggplot(aes(x = Year, y = Rate, color = Category)) +
         geom_line(size = 1) +
  ggtitle("15-24 Age Group Suicide Rate") +
  xlim(1995,2010)
g_plot
```

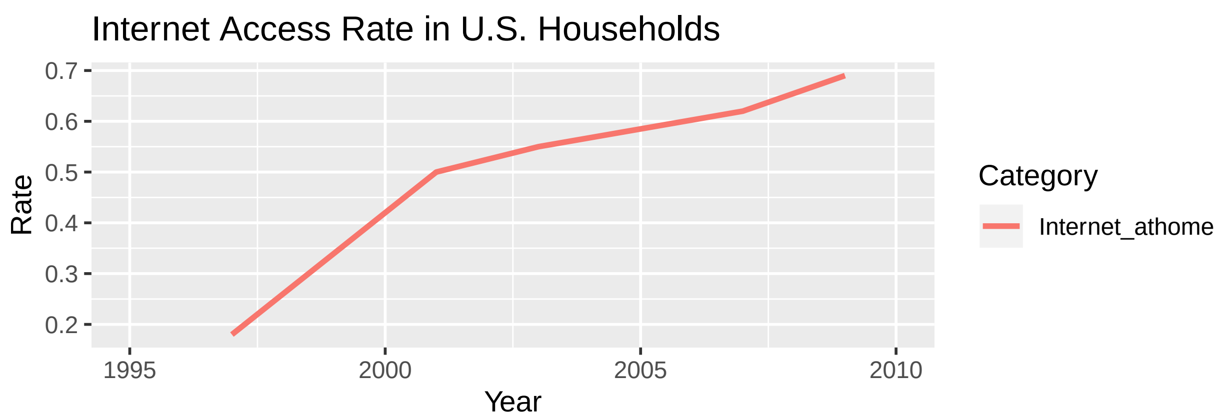


```r
# plotting just household internet access rate

h_plot <- tidy %>%
  filter(Category %in% c("Internet_athome")) %>%
         ggplot(aes(x = Year, y = Rate, color = Category)) +
         geom_line(size = 1) +
  ggtitle("Internet Access Rate in U.S. Households") +
  xlim(1995,2010)
h_plot
```




```r
# arranging plots for comparison

grid.arrange(g_plot, h_plot)
```






```r
# computing linear regression statistics for these variables

fit2 <- wide %>% lm(All_15_24_Rates ~ Internet_athome, data = .)
summary(fit2)
```

```
##
## Call:
## lm(formula = All_15_24_Rates ~ Internet_athome, data = .)
##
## Residuals:
##       28       31       32       34       38       40
##  0.26457 -0.15244 -0.22478 -0.31686 -0.09579  0.52529
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      11.7039     0.4705   24.87 1.55e-05 ***
## Internet_athome  -3.1583     0.9051   -3.49   0.0251 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3638 on 4 degrees of freedom
##   (44 observations deleted due to missingness)
## Multiple R-squared:  0.7527, Adjusted R-squared:  0.6909
## F-statistic: 12.18 on 1 and 4 DF,  p-value: 0.02513
```

Results of internet at home analysis

There is a negative correlation of -.87 for these two variables. Correspondingly, there is a slope of -3.1 for the lm (regression) calculation. Home internet use reached 50% saturation in the United States in the year 2002.

The 15-24 S.R. was at a baseline and did not vary much up or down from around 1998 to 2009, although what appears in this indivdual graph to be be beginning of a consistent upward trend begins in about 2006.

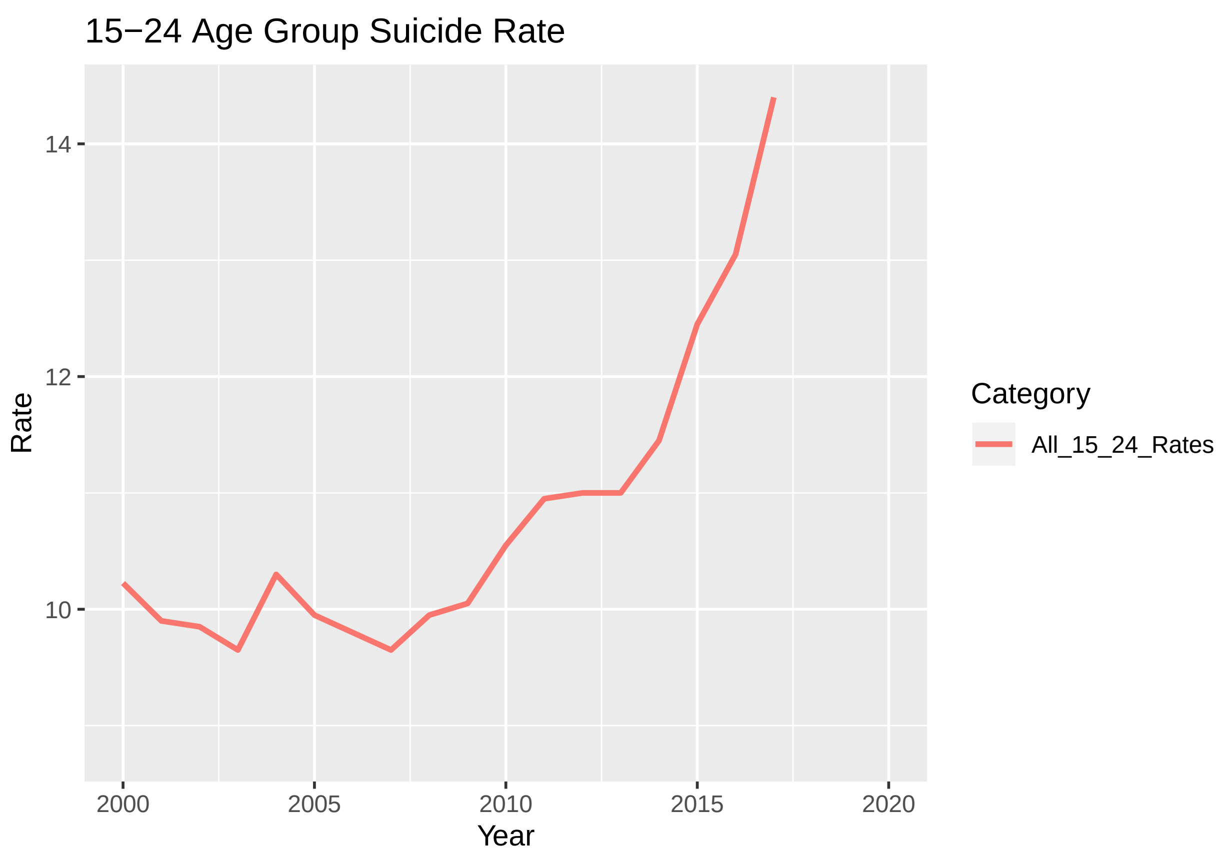
It can be concluded with resonable certainty that internet access at home did not cause an uptick in youth suicide rates.

Cell Phone Ownership Analysis

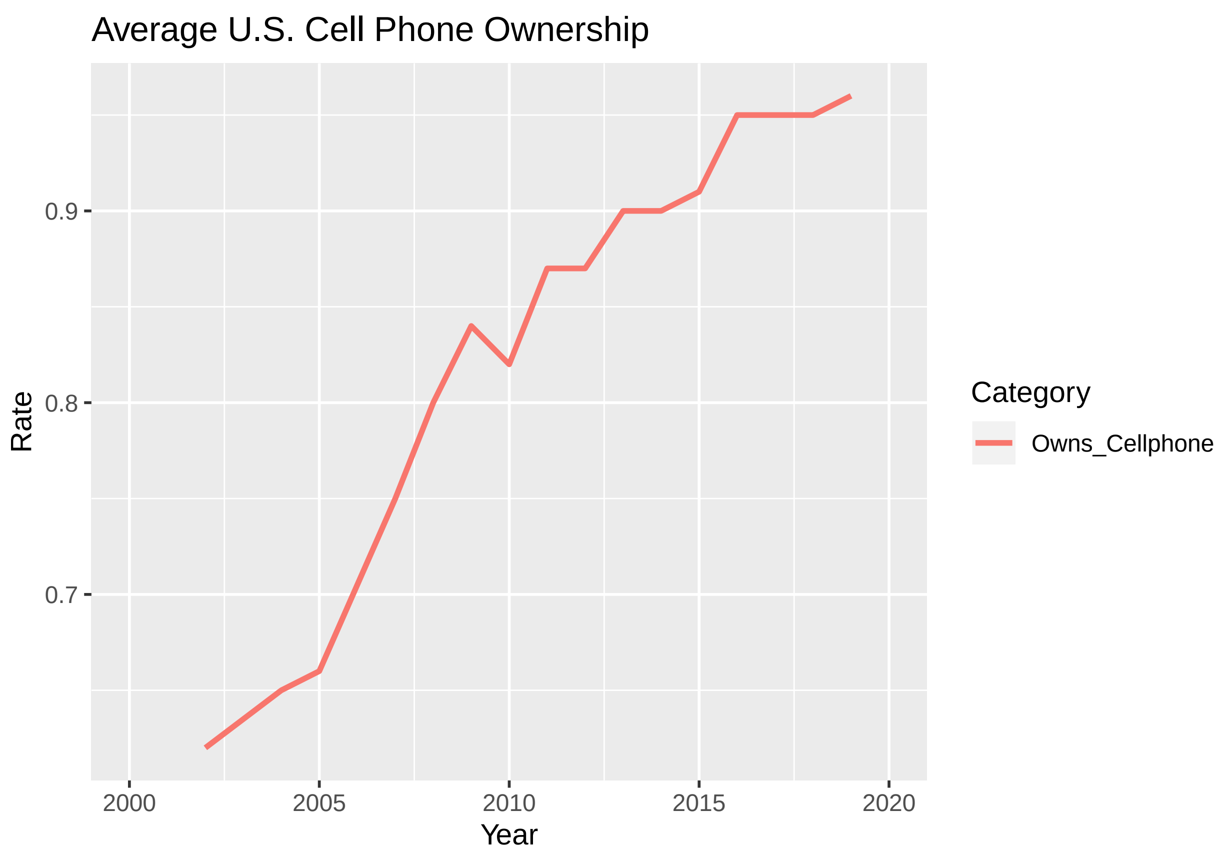
```
# calculating correlation for cell phone ownership and 15-24 suicide rate

cor(wide$`All_15_24_Rates`,wide$Owns_Cellphone, use = "pairwise.complete.obs")
```

```
## [1] 0.7539628
```

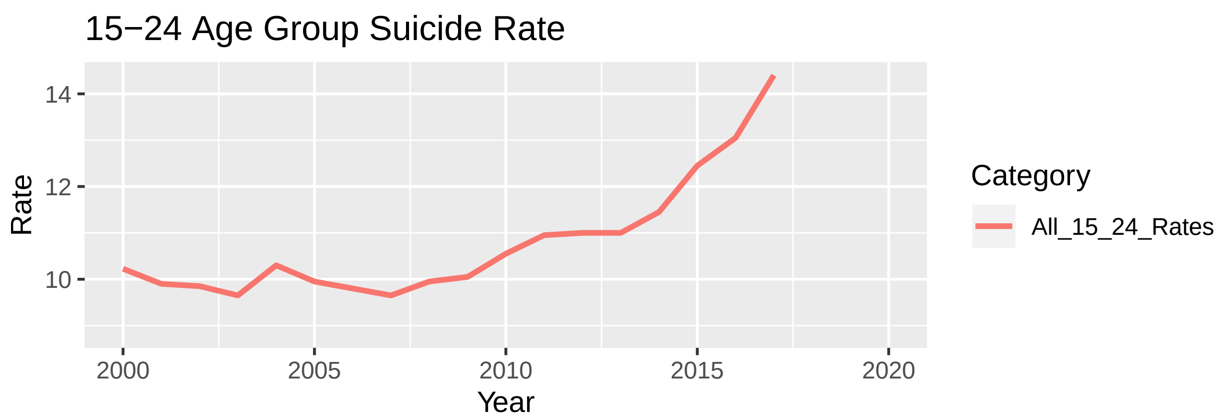
```
# plotting just 15-24 S.R.

i_plot <- tidy %>%
  filter(Category %in% c("All_15_24_Rates")) %>%
          ggplot(aes(x = Year, y = Rate, color = Category)) +
          geom_line(size = 1) +
  ggtitle("15-24 Age Group Suicide Rate") +
  xlim(2000,2020)
i_plot
```

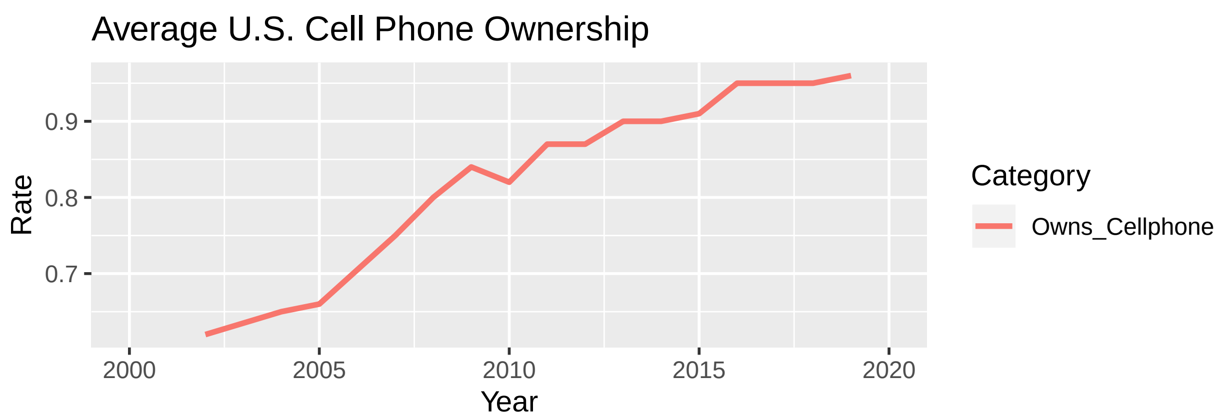


```r
# plotting just adult average cell phone ownership

j_plot <- tidy %>%
  filter(Category %in% c("Owns_Cellphone")) %>%
          ggplot(aes(x = Year, y = Rate, color = Category)) +
          geom_line(size = 1) +
  ggtitle("Average U.S. Cell Phone Ownership") +
  xlim(2000,2020)
j_plot
```




```r
# arranging plots for comparison

grid.arrange(i_plot, j_plot)
```






```r
# computing linear regression statistics for these variables

fit3 <- wide %>% lm(All_15_24_Rates ~ Owns_Cellphone, data = .)
summary(fit3)
```

```
##
## Call:
## lm(formula = All_15_24_Rates ~ Owns_Cellphone, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9444 -0.4249 -0.2096  0.4343  1.2663
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.967      1.541   3.223  0.00811 **
## Owns_Cellphone    7.176      1.885   3.807  0.00291 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7101 on 11 degrees of freedom
##   (37 observations deleted due to missingness)
## Multiple R-squared:  0.5685, Adjusted R-squared:  0.5292
## F-statistic: 14.49 on 1 and 11 DF,  p-value: 0.00291
```

Results of cell phone ownership analysis

There is a positive correlation of .75 for these two variables. Correspondingly, there is a positive slope of 7.1 with a p-value of .002 for the lm calculation. Cell phone ownership arrived at about a 65% saturation level in the United States in the year 2004.

It should be noted that the Pew Research Center data used for cell phone ownership includes the ownership of smart phones (which were owned by 50% of the population in 2012). Before smart phones (with mobile internet access), cell phones were mostly just used for calls and later, for calls and texting.

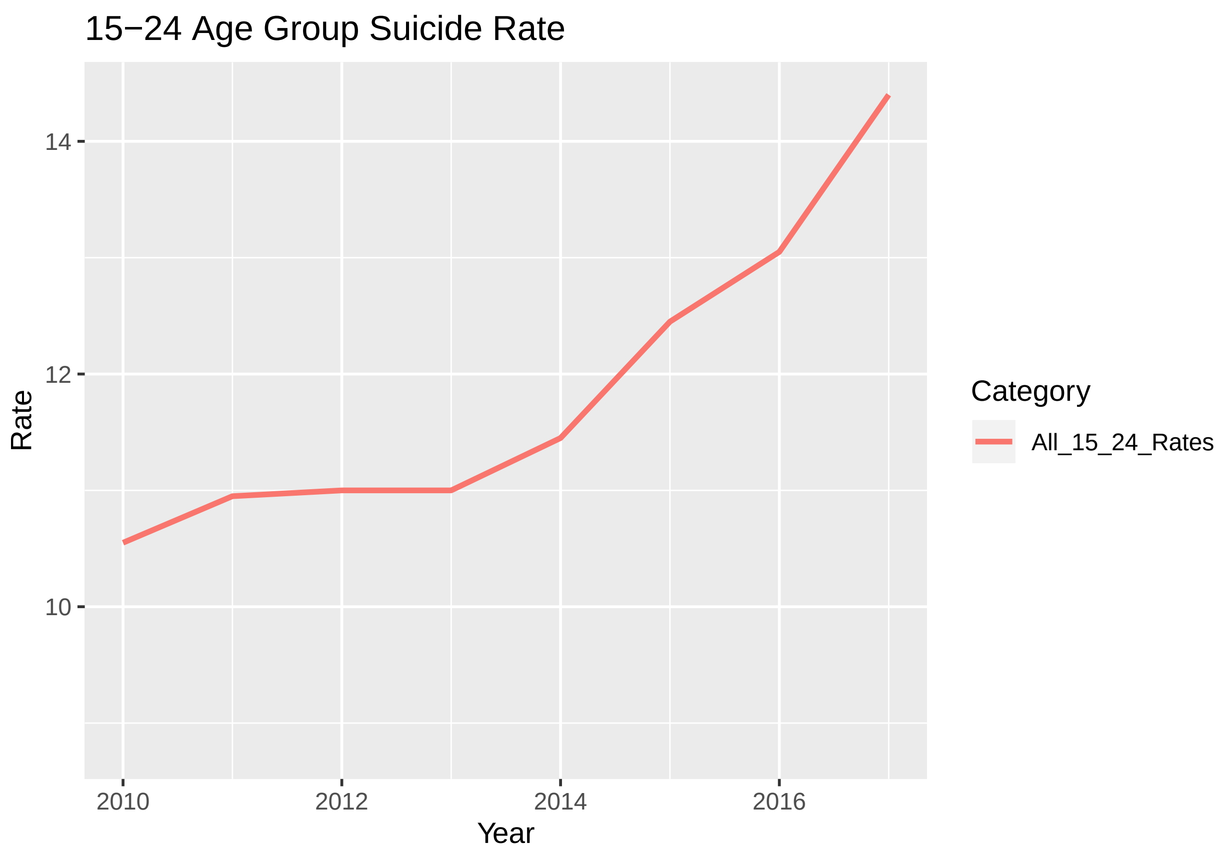
The consistent increase in 15-24 age group suicide rate began around 2008. The .75 correlation between cell phone ownership and suicide rate indicates that owning cell phones may be a significant contributor to the increse in 15-24 S.R.

Smart Phone Ownership Analysis

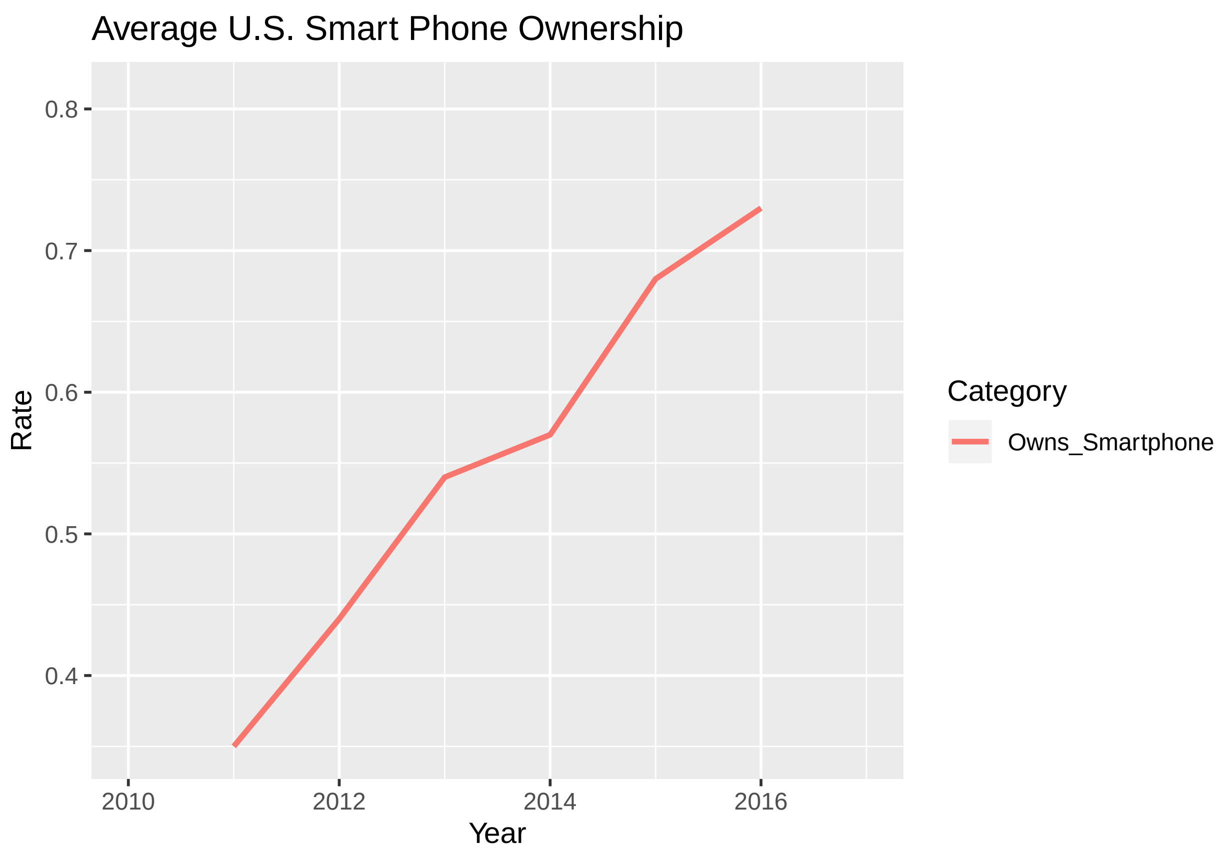
```
# calculating correlation for smart phone ownership and 15-24 suicide rate

cor(wide$All_15_24_Rates,wide$Owns_Smartphone, use = "pairwise.complete.obs")
```

```
## [1] 0.8956838
```

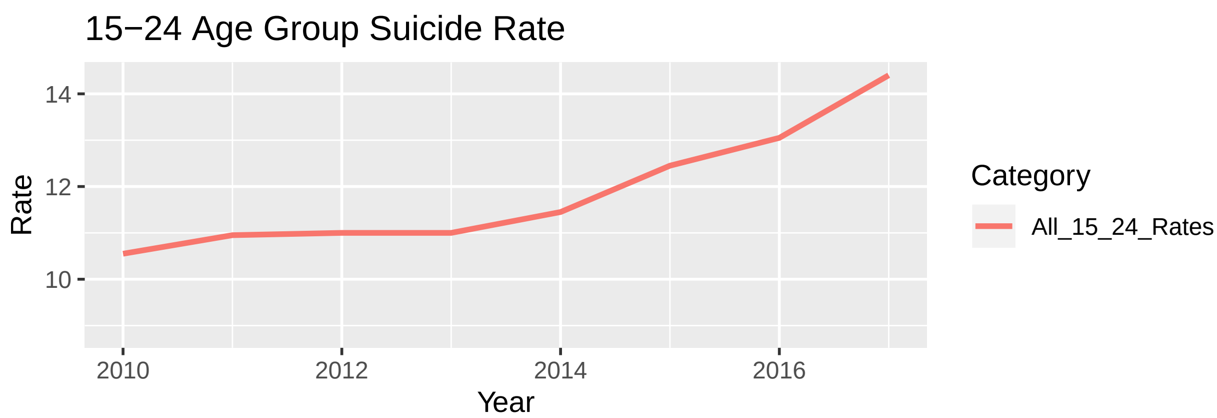
```
# plotting just 15-24 S.R.

k_plot <- tidy %>%
  filter(Category %in% c("All_15_24_Rates")) %>%
           ggplot(aes(x = Year, y = Rate, color = Category)) +
           geom_line(size = 1) +
  ggtitle("15-24 Age Group Suicide Rate") +
  xlim(2010,2017)
k_plot
```

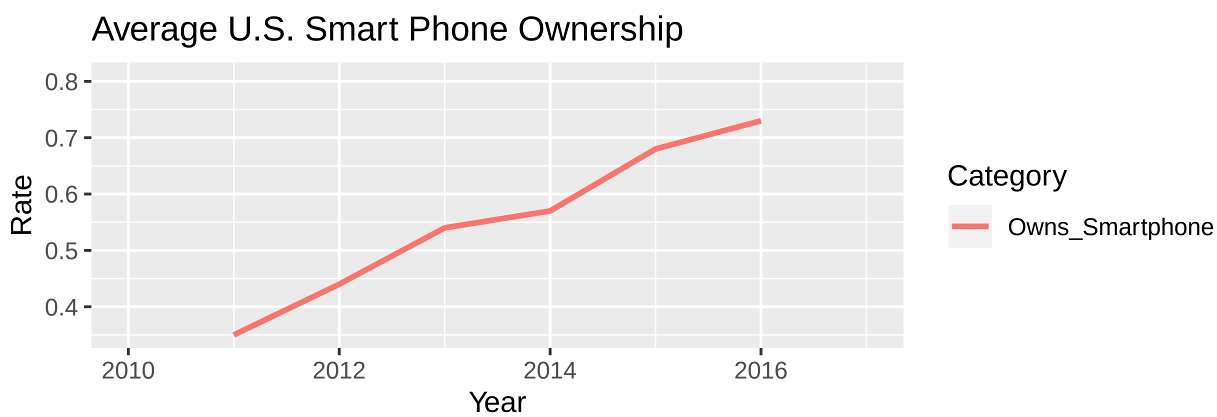



```r
# plotting just adult average smart phone ownership

l_plot <- tidy %>%
  filter(Category %in% c("Owns_Smartphone")) %>%
           ggplot(aes(x = Year, y = Rate, color = Category)) +
           geom_line(size = 1) +
  ggtitle("Average U.S. Smart Phone Ownership") +
  xlim(2010,2017)
l_plot
```




```
# arranging plots for comparison

grid.arrange(k_plot, l_plot)
```






```r
# computing linear regression statistics for these variables

fit4 <- wide %>% lm(`All_15_24_Rates` ~ Owns_Smartphone, data = .)
summary(fit4)
```

```
##
## Call:
## lm(formula = All_15_24_Rates ~ Owns_Smartphone, data = .)
##
## Residuals:
##       42       43       44       45       46       47
##  0.42825 -0.02527 -0.58473 -0.30257  0.08202  0.40229
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       8.5636     0.7872  10.878 0.000405 ***
## Owns_Smartphone   5.5946     1.3888   4.028 0.015755 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4433 on 4 degrees of freedom
##   (44 observations deleted due to missingness)
## Multiple R-squared:  0.8022, Adjusted R-squared:  0.7528
## F-statistic: 16.23 on 1 and 4 DF,  p-value: 0.01576
```

Results of smart phone ownership analysis

There is a positive correlation of .90 for these two variables. Correspondingly, there is a positive slope of 5.6 for the lm calculation with a p-value of .015.

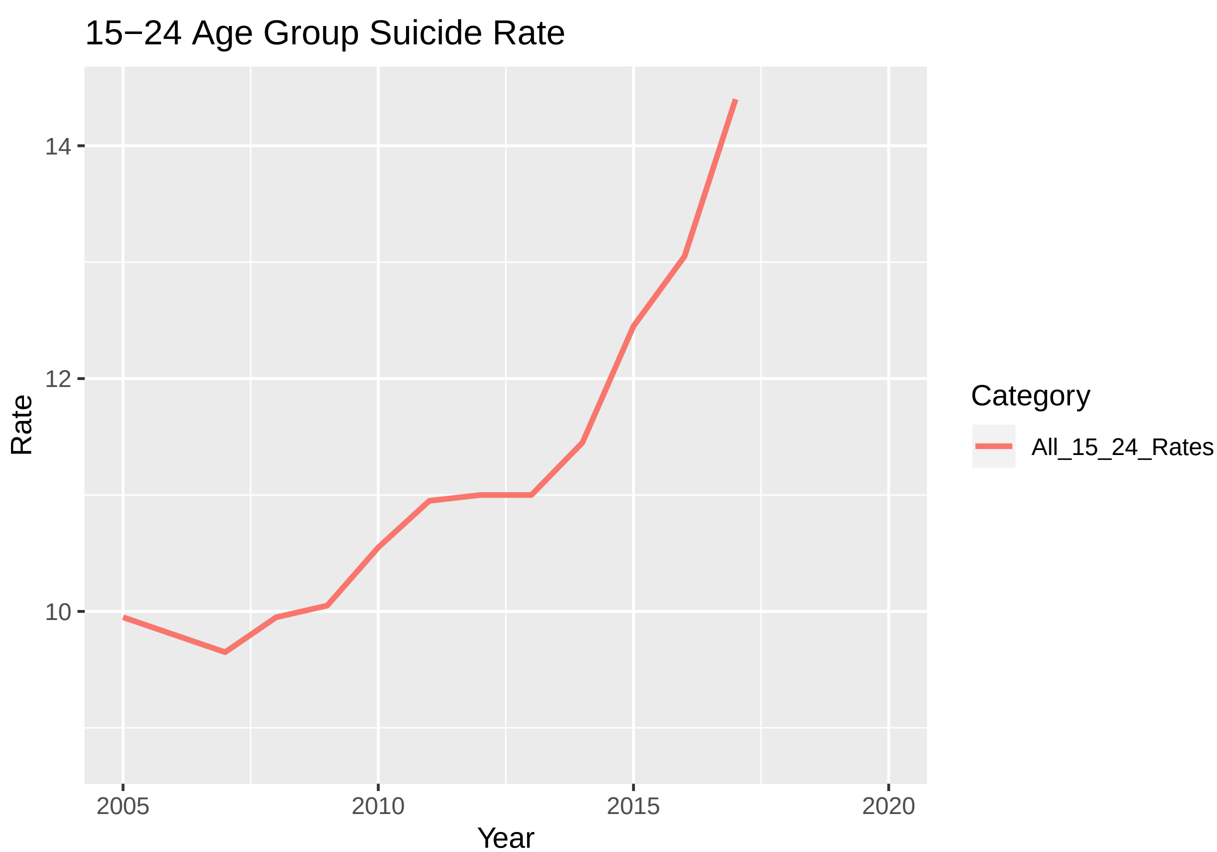
There seems to be a strong indication then that smart phones are a significant, if not the leading cause of the increase in the 15-24 age group S.R.

Social Media Use analysis

```
# calculating correlation for social media use and 15-24 suicide rate

cor(wide$All_15_24_Rates,wide$Avg_SM_Use, use = "pairwise.complete.obs")
```

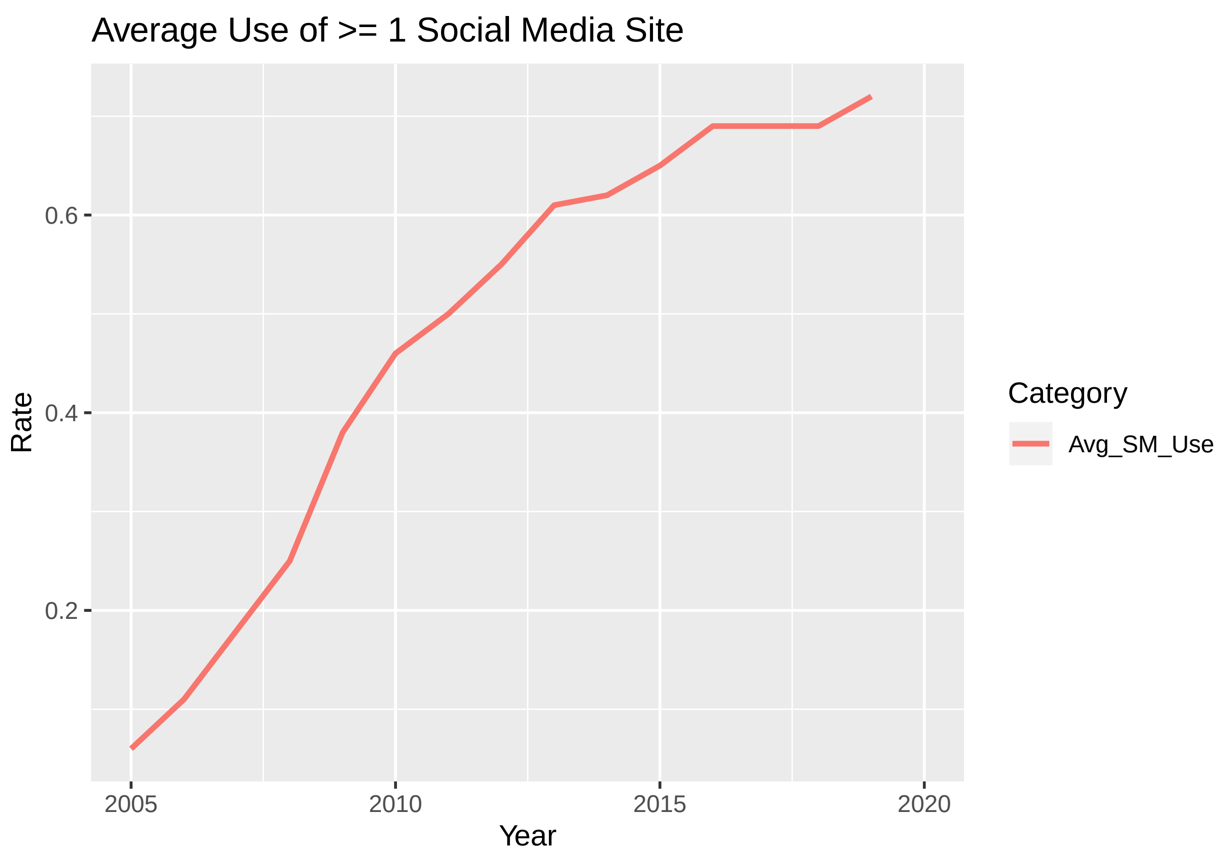
```
## [1] 0.8395684
```

```
# plotting just 15-24 S.R.

m_plot <- tidy %>%
  filter(Category %in% c("All_15_24_Rates")) %>%
           ggplot(aes(x = Year, y = Rate, color = Category)) +
           geom_line(size = 1) +
  ggtitle("15-24 Age Group Suicide Rate") +
  xlim(2005,2020)
m_plot
```

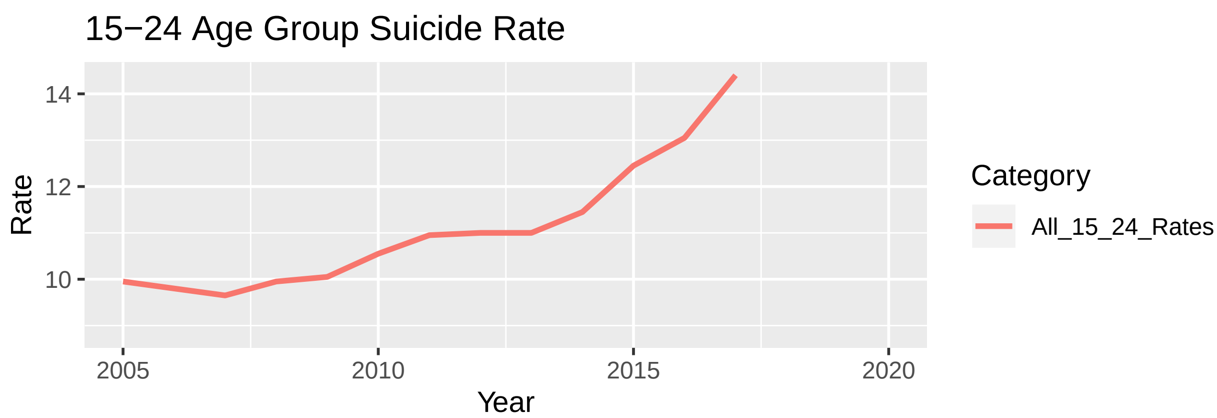


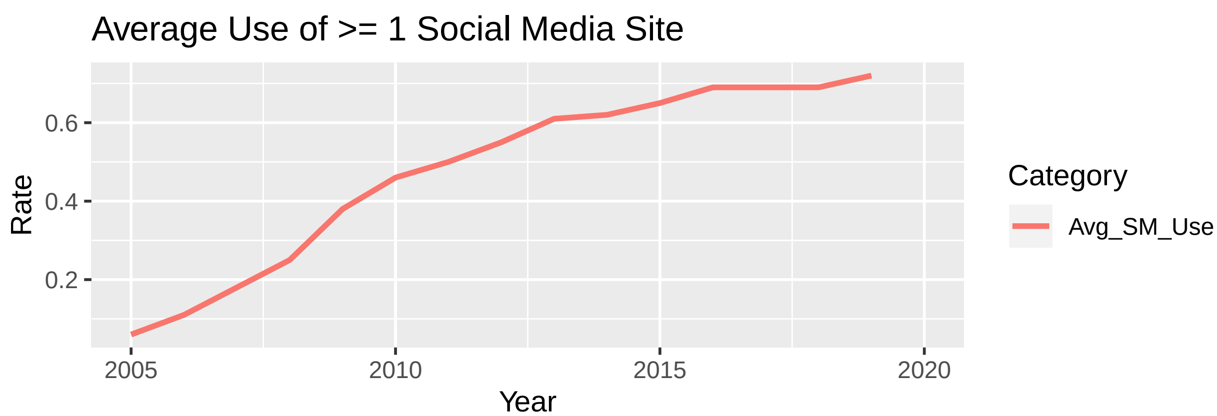

```
# plotting just social media use
```

```r
n_plot <- tidy %>%
  filter(Category %in% c("Avg_SM_Use")) %>%
           ggplot(aes(x = Year, y = Rate, color = Category)) +
           geom_line(size = 1) +
  ggtitle("Average Use of >= 1 Social Media Site") +
  xlim(2005,2020)
n_plot
```




```r
# arranging plots for comparison

grid.arrange(m_plot, n_plot)
```






```r
# computing linear regression statistics for these variables

fit5 <- wide %>% lm(All_15_24_Rates ~ Avg_SM_Use, data = .)
summary(fit5)
```

```
##
## Call:
## lm(formula = All_15_24_Rates ~ Avg_SM_Use, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6183 -0.4020 -0.1949  0.4071  1.1202
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.1220     0.4301  21.208 5.41e-09 ***
## Avg_SM_Use    4.0692     0.8777   4.636  0.00123 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6061 on 9 degrees of freedom
##   (39 observations deleted due to missingness)
## Multiple R-squared:  0.7049, Adjusted R-squared:  0.6721
## F-statistic:  21.5 on 1 and 9 DF,  p-value: 0.001226
```

Results of social media use analysis

There is a positive correlation of .84 for these two variables. Correspondingly, there is a positive slope of 4.01 for the lm calculation with a p-value of .001. Social media use reached 50% saturation in around 2011.

It is noteworthy that social media use reached 50% saturation in 2012, as that was the year in which the suicide rate began to steadily and sharply increase for the 15-24 age group. 2012 is also the year that smart phones were adopted by 50% of the U.S. population.

Social media use on a stand alone basis can be considered a strong contributor to the suicide rate of the 15-24 U.S. age group. This is subject to a confounding effect that likely exists with the use of smart phones.

Markets analysis

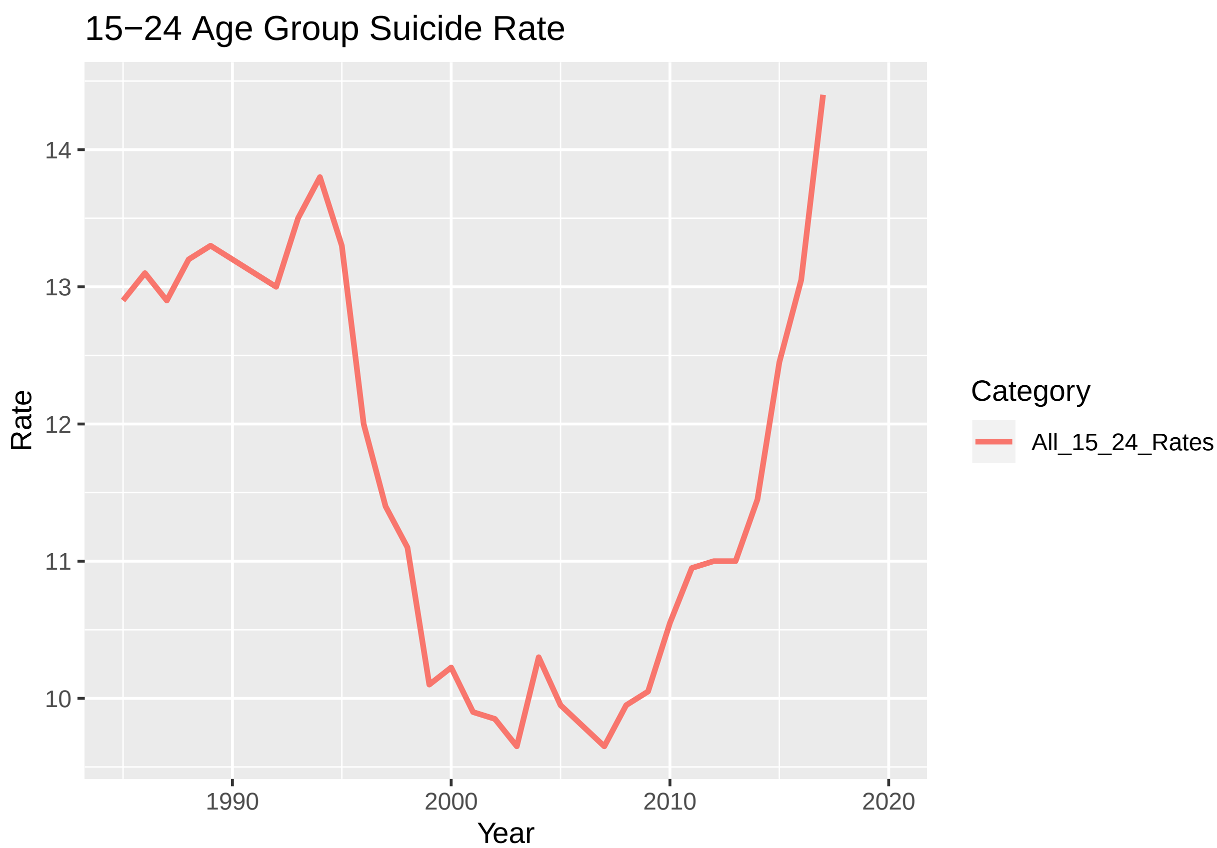
It is a proper control to include an analysis of the economic environment in the U.S. with the electronic media analysis. The market data used below was obtained from Yahoo Finance. It represents the average stock market close amount by year for the three major U.S. stock indexes. The indexes used were the Dow, the S&P 500, and the NASDAQ. A simple average was taken for each monthly close amount for these three indexes, and then these averaged monthly close amounts were themselves averaged by year.

This analysis will begin in 1985, when Yahoo Finance provides data for all three of these indexes.

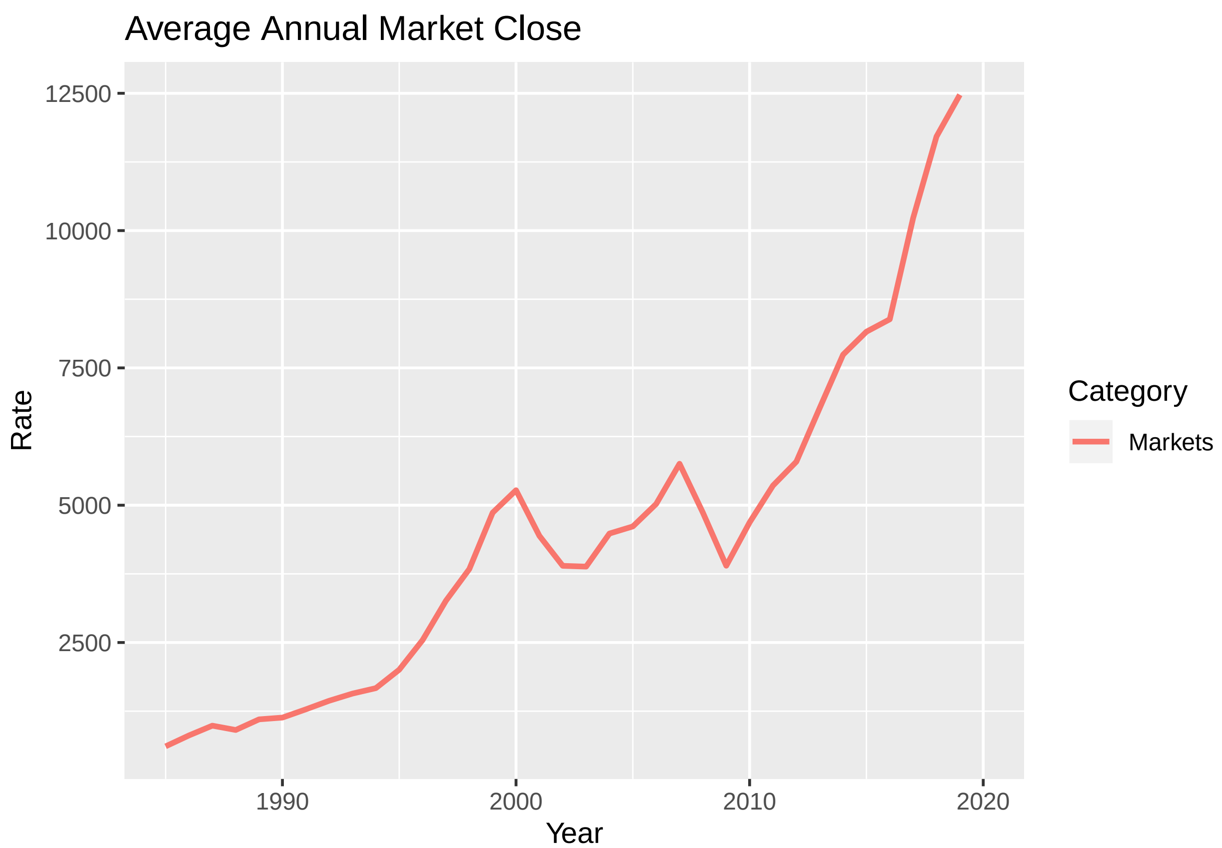
```r
# calculating correlation for economic variation and 15-24 suicide rate

wide2 <- wide %>% filter(Year >= 1985)
cor(wide2$All_15_24_Rates,wide2$Markets, use = "pairwise.complete.obs")
```

```
## [1] -0.3134168
```

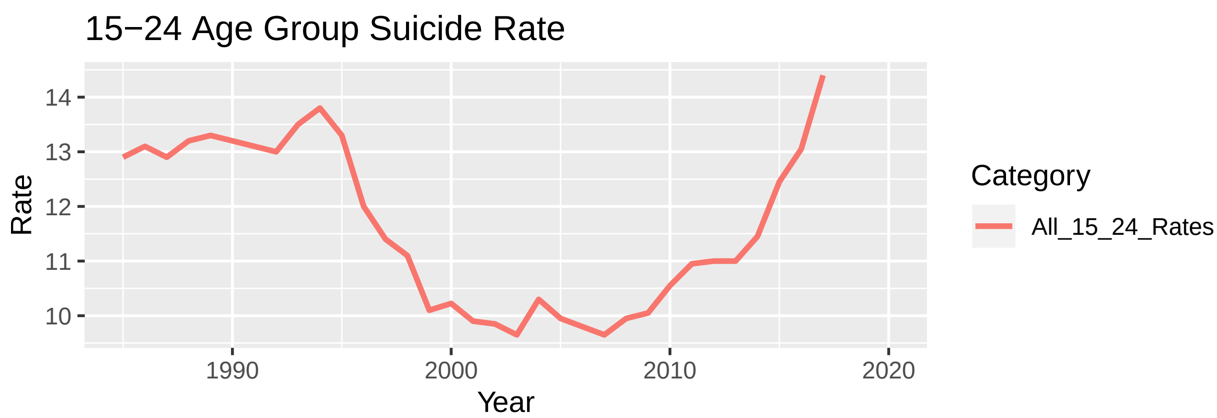
```r
# plotting just 15-24 S.R.

tidy2 <- gather(wide2, Category, Rate, -1, na.rm = TRUE)

q_plot <- tidy2 %>%
  filter(Category %in% c("All_15_24_Rates")) %>%
           ggplot(aes(x = Year, y = Rate, color = Category)) +
           geom_line(size = 1) +
  ggtitle("15-24 Age Group Suicide Rate") +
  xlim(1985,2020)
q_plot
```




```
# plotting just market variation

r_plot <- tidy2 %>%
  filter(Category %in% c("Markets")) %>%
         ggplot(aes(x = Year, y = Rate, color = Category)) +
         geom_line(size = 1) +
  ggtitle("Average Annual Market Close") +
  xlim(1985,2020)
r_plot
```




```
# arranging plots for comparison

grid.arrange(q_plot, r_plot)
```




```r
# computing linear regression statistics for these variables

fit6 <- wide2 %>% lm(All_15_24_Rates ~ Markets, data = .)
summary(fit6)
```

```
##
## Call:
## lm(formula = All_15_24_Rates ~ Markets, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0068 -1.3706 -0.1093  0.9792  3.9467
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.3927574  0.4824517  25.687   <2e-16 ***
## Markets     -0.0001896  0.0001032  -1.838   0.0757 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.457 on 31 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.09823,    Adjusted R-squared:  0.06914
## F-statistic: 3.377 on 1 and 31 DF,  p-value: 0.07572
```

Results of Markets Analysis

There is a negative correlation of -.31 when comparing the Markets data to the 15-24 S.R. data. The slope of the regression line is roughly zero, indicating that markets have no influence postively or negatively on the 15-24 S.R.

It is notable that the "dot-com bubble", which was the term used for the fast rate of market expansion illustrated by the sharp increase in market close from 1995 to 2000, is joined by a very sharp decrease in the 15-24 S.R. This may indicate that the market increase did play a role in the suicide rate decrease during that 5 year time frame.

Finally, it is notable that the "dot-com" market crash of 2000-2002 does not visually appear much less severe than the 2007-2009 "great recession" caused by the sub-prime housing bubble. During 2000-2002, although the markets went into the dot-com crash recession period, the 15-24 S.R. decreased.

Assessment of above 6 parameter analysis

It seems that there are clear indications that cell phones, smart phones, and social media are strongly correlated with the sharp increase in suicide rate. It also appears to be true that computer and internet use at home, as well as the state of the U.S. economy, do not have a causative effect for increasing youth suicide, represented above by the 15-24 S.R.

An argument could be made that the commonly termed "Great Recession" of 2008 played more of a role in the mental health of the youth of the United States than the markets normally have from 1970 and forward. It was to a degree a more severe recession than the 2000 "dot-com" crash. Therefore, multiple regression analysis will be performed with these two variables.

The "owns cell phone" and "markets" data will be used from 2005 forward for this analysis to allow for the influence of the 2007-2009 "great recession" to be included in the analysis. It should be noted that the correlation and linear regression for cell phone ownership to 15-24 S.R. was less than the correlation for smart phone ownership, so the influence of the multiple regression calculated below will slightly overstate the market impact and understate the impact of smart internet connected phones replacing "old-style" cell phones.

```r
# running linear regression analysis for the two predictor variables

fit1 <- wide2 %>% filter(Year >= 2005) %>% lm(All_15_24_Rates ~ Owns_Cellphone + Markets, data = .)
summary(fit1)
```

```
##
## Call:
## lm(formula = All_15_24_Rates ~ Owns_Cellphone + Markets, data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68530 -0.37527  0.03328  0.29760  0.59857
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.1010881  1.7203608   2.384   0.0443 *
## Owns_Cellphone 5.1340968  2.5857841   1.986   0.0823 .
## Markets        0.0004142  0.0001385   2.990   0.0173 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4892 on 8 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.8344, Adjusted R-squared:  0.793
```

```
## F-statistic: 20.15 on 2 and 8 DF,  p-value: 0.0007525
```

The above multiple regression analyis attributes all of the increase in 15-24 S.R. to cell phone ownership, with the market impact being zero for the time frame of 2005-2020.

Fitting GLM machine learning model

Machine Learning Analysis for Market Impact

A machine learning analysis is provided to help to illustrate the lack of predictive power of the economic environment to 15-24 S.R. Using a generalized linear regression algorithm, an attempt to predict the suicide rate using markets and cell phone ownership is provided.

The GLM prediction is first made, and then it is rounded to the nearest integer. The actual "test set" 15-24 S.R. is also rounded to the nearest integer, and then an assessment of the accuracy of the prediction is made.

```
# Create data frame with just the columns needed

data_glm <- wide[,c(1,12,16)] %>% drop_na()

dim(data_glm)
```

```
## [1] 33  3
```

```
# Create data partitions with 75% of data as train set and 25% of data as test set

set.seed(1)
test_index <- createDataPartition(data_glm$Year, times = 1, p = .75, list = FALSE)
train_set <- data_glm[test_index,]
test_set <- data_glm[-test_index,]
train_set
```

```
##    Year Markets All_15_24_Rates
## 17 1986     808           13.10
## 18 1987     986           12.90
## 19 1988     907           13.20
## 20 1989    1101           13.30
## 21 1990    1133           13.20
## 23 1992    1439           13.00
## 24 1993    1570           13.50
## 25 1994    1669           13.80
## 26 1995    2005           13.30
## 27 1996    2544           12.00
## 29 1998    3837           11.10
## 30 1999    4864           10.10
## 32 2001    4443            9.90
## 34 2003    3882            9.65
## 35 2004    4484           10.30
## 36 2005    4613            9.95
## 37 2006    5023            9.80
## 38 2007    5755            9.65
## 40 2009    3898           10.05
## 41 2010    4687           10.55
## 42 2011    5358           10.95
## 44 2013    6770           11.00
## 45 2014    7741           11.45
## 47 2016    8385           13.05
```

```
## 48 2017   10230           14.40
```

```r
test_set
```

```
##    Year Markets All_15_24_Rates
## 16 1985     609          12.900
## 22 1991    1283          13.100
## 28 1997    3261          11.400
## 31 2000    5273          10.225
## 33 2002    3896           9.850
## 39 2008    4863           9.950
## 43 2012    5792          11.000
## 46 2015    8159          12.450
```

```r
# run GLM machine learning algorithm

glm_fit <- train_set %>%
  glm(All_15_24_Rates ~ Markets, data = ., family = "gaussian")
glm_fit
```

```
##
## Call:  glm(formula = All_15_24_Rates ~ Markets, family = "gaussian",
##     data = .)
##
## Coefficients:
## (Intercept)      Markets
##   12.446314    -0.000183
##
## Degrees of Freedom: 24 Total (i.e. Null);  23 Residual
## Null Deviance:       59.9
## Residual Deviance: 54.64     AIC: 96.49
```

```r
y_hat <- predict(glm_fit, newdata = test_set, type = "response")
y_hat
```

```
##       16       22       28       31       33       39       43       46
## 12.33487 12.21153 11.84956 11.48137 11.73336 11.55640 11.38640 10.95324
```

```r
# round y hat to nearest integer

y_hat <- round(y_hat, 0)
y_hat
```

```
## 16 22 28 31 33 39 43 46
## 12 12 12 11 12 12 11 11
```

```r
# round the test set to the nearest integer

actual_rates <- round(test_set$All_15_24_Rates, 0)
actual_rates
```

```
## [1] 13 13 11 10 10 10 11 12
```

```r
# calculate the number predicted accurately by the GLM model

mean(y_hat == actual_rates)
```

```
## [1] 0.125
```

Results of Markets Only Analysis

The economic environment is again proven as being a non-predictor of U.S. youth suicide rates, with an accurate prediction occuring only 12.5% of the time in the test set when the predicted and the actual suicide rates were rounded to the nearest integer.

Results

This study has provided is compelling if not conclusive evidence that the combination of smart phones and social media are the principal cause of the recent uptick in youth suicide rates. The overall separate component analysis shows a graded effect in which internet based phones cause more suicides than pre-internet enabled cell phones. The LM analysis which combined the economic environment of the U.S. indicates that the stock market (deemed representative of the economic environment) had no effect on suicide rates, and that cell phones were the sole cause of the increase from 2005-2017. The glm machine learning analysis to test the predictive power of markets alone supported the evidence that very little to no causative relationship exist between the economic environment and youth suicide.

Conclusion

It appears to be apparent that youth suicide in the U.S. has increased because of the increased use of smartphones and social media.

On a "macro" scale, there could be several reasons for youth suicide changes. They are cultural changes, war, economic conditions, and substance abuse. The cultural change of the introduction and widespread use of smartphones and social media causes behavior that often leads to isolation and social "shaming" on social media sites. Both isolation and shaming have been suggested as strong causes of mental health problems in the past, with isolation being a cause of mental disturbance for all age groups and shaming a cause of mental disturbance for youth age groups.

A factor other than isolation and shaming that should not be overlooked is the new concept of "remote electronic friendship" that only has existed since smartphones were created. For all of human history, friendship has been an interpersonal endeavor. Before smartphones, phones were simply used to organize physical meeting places for friends. Further, phones themselves were invented in the late 1800s. Before the late 1800s, humans only had interpersonal social interaction. Smart phones are then a forced new adaptation for humans. Another note that may impact suicide rates is the constant barrage of sensational advertising that exists in smartphones. All of these impacts should be rigorously studied.

Smartphone use is one of the most significant cultural changes in human history. Modern human civilization goes back 5 - 10 thousand years. It is only in the last 10 years that smartphones have been a part of human society. Our generation is then the first generation ever that had the opportunity to use smartphones to communicate with other people, or to use as a method of finding information and/or solve problems.

The use of these devices has increased very rapidly since their general inception in 2012. The youngest generations now use these devices many hours during the day, evening, and even overnight. Other studies commonly state that U.S. youth uses these devices more than the rest of the adult U.S. population.

It is prudent that the social and psychological effect of this new development in technology be scrutinized heavily early in its lifecycle. This study supports the conclusion that it harms the mental health of those who use these devices heavily. As this study helped to illustrate, these devices are being used in somewhat of an addictive pattern, by all age groups. They provide a means of escape from real life interpersonal challenges. Chemical dependency also provides this means of emotional escape. The author of this report then hopes that others may join in a comprehensive examination of the effect of the regular use of these devices by all ages.

This study is limited with the small sample size of years to study with cell/smart phone use. Further studies may wish to expand into subgroups such as race and geographic location with significantly differing smartphone/social media use rates. Further studies may also wish to use other sources that record electronic use by age group on a macro scale, and not just the sources cited by this report.

Another opportunity for a future study related to this topic stems from the fact that young people are known as being reckless with poor decision making skills. It is possible that with more electronic media use, that there is less physical (external) reckless behavior is occuring amoungst youth. The death rate for this age group with the cause of being an accident may be decreasing at the same time that the suicide rate is increasing. This could potentially lead to a more stable trend when considering the overall youth death rate.

References

1. “Death Rates Due to Suicide and Homicide Among Persons Aged 10-24:United States, 2000-2017” (Curtin, Heron). NCHS Data Brief, No. 352, October 2019.
2. “US Suicide Rates by Age Group”, 1970-2002 (McKeown, Cuffe, Schulz). American Joural of Public Health, October, 2006, Vol. 96, No. 10.
3. Pew Research Center - https://www.pewresearch.org/internet/fact-sheet/mobile/
4. “Suicide Rate Correlations You May Not Have Known About” (Greenstein). National Alliance on Mental Illness, Sep. 9, 2016.
5. U.S. Census Bureau, Current Population Survey, October 1984, 1989, 1993, 1997, 2000, 2001, 2003, 2007, 2009. “https://www.census.gov/prod/2013pubs/p20-569.pdf”
6. Yahoo Finance. S&P market index link - https://finance.yahoo.com/quote/%5EGSPC/history/ Dow link - https://finance.yahoo.com/quote/%5EDJI/ NASDAQ link - https://finance.yahoo.com/quote/%5Eixic/history/

author email - adam.ausperk@tri-c.edu…email with inquiries