DATA 606 FINAL PROJECT - PROPOSAL -

Tage N Singh

2021-04-11

```
output:
  prettydoc::html_pretty:
    theme: architect
    highlight: github
```

```
library(readr)
library(dplyr)
```

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

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

```
## The following objects are masked from 'package:base':
##
```

```
##     intersect, setdiff, setequal, union
```

```
library(kableExtra)
```

```
## Warning: package 'kableExtra' was built under R version 4.0.5
```

```
##
## Attaching package: 'kableExtra'
```

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

```
library(ggplot2)
library(sm)
```

```
## Warning: package 'sm' was built under R version 4.0.5
```

```
## Package 'sm', version 2.2-5.6: type help(sm) for summary information
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

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

Data Preparation

```
# load data

masks <- data.frame(read.csv(file = "https://raw.githubusercontent.com/tagensingh/SPS-DATA606-FI

US <- data.frame(read.csv(file = "https://raw.githubusercontent.com/tagensingh/SPS-DATA606-FP/ma

states <- data.frame(read.csv(file = "https://raw.githubusercontent.com/tagensingh/SPS-DATA606-

counties <- data.frame(read.csv(file = "https://raw.githubusercontent.com/tagensingh/SPS-DATA60
```

Research question

You should phrase your research question in a way that matches up with the scope of inference your dataset allows for.

What is the impact of mask usage on COVID-19 cases and deaths in the USA

COVID-19 is the most devastating modern pandemic since the AIDS pandemic.

Currently we have recorded 135 million cases with 2.9 million deaths worldwide.

Currently the USA have recorded 31.2 million cases with 561 thousand deaths.

Our study will examine the data and relationship (correlation) between mask usage, cases and deaths at national, state and county level.

Cases

What are the cases, and how many are there?

The cases are contained in 2 main datasets

The mask usage dataset contain county records from a NY Times survey done by survey firm Dynata of 250,000 responses between July 2nd and July 14th 2020, the dataset contain 3142 records.

The counties dataset contain daily cases and deaths for each county in the USA from 01/20/2020 to 04/04/2021, the dataset contain 1189856 records.

The “US” and “STATES” datasets are supplemental for reference.

Data collection

Describe the method of data collection.

The data was downloaded from Kaggle and then uploaded to Github. I am reading the .csv from github raw form.

Type of study

What type of study is this (observational/experiment)?

This is an observational study analyzing data collected by NY times and Dynata.

Data Source

If you collected the data, state self-collected. If not, provide a citation/link.

The data was sourced from kaggle.com, it was compiled by the NY Times

Data from The New York Times, based on reports from state and local health agencies.

https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html

Dependent Variable

What is the response variable? Is it quantitative or qualitative?

In this study the response variables are the COVID-19 cases and deaths in the USA, measured at the county level. This is a quantitative variable.It is contained in the us_counties dataset.

Independent Variable

You should have two independent variables, one quantitative and one qualitative.

In this study the independent variables are the COVID-19 mask usage in the USA, measured at the county level. This is a quantitative variable contained in the mask_use_by_county dataset.

Relevant summary statistics

Provide summary statistics for each the variables. Also include appropriate visualizations related to your research question (e.g. scatter plot, boxplots, etc). This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

```
## The summary for the masks dataset

summary(masks)
```

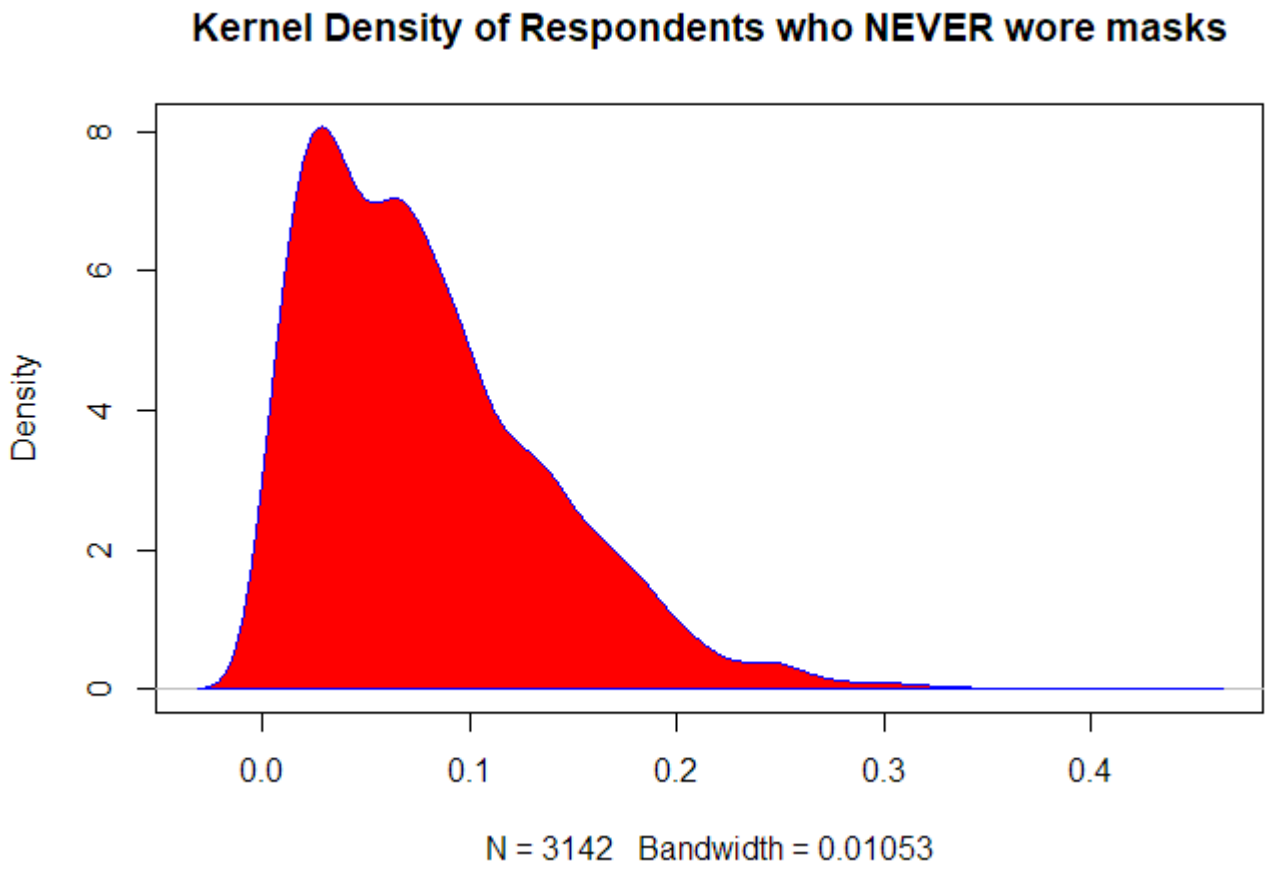
```
##     COUNTYFP         NEVER             RARELY          SOMETIMES
##  Min.   : 1001   Min.   :0.00000   Min.   :0.00000   Min.   :0.0010
##  1st Qu.:18178   1st Qu.:0.03400   1st Qu.:0.04000   1st Qu.:0.0790
##  Median :29176   Median :0.06800   Median :0.07300   Median :0.1150
##  Mean   :30384   Mean   :0.07994   Mean   :0.08292   Mean   :0.1213
##  3rd Qu.:45081   3rd Qu.:0.11300   3rd Qu.:0.11500   3rd Qu.:0.1560
##  Max.   :56045   Max.   :0.43200   Max.   :0.38400   Max.   :0.4220
##    FREQUENTLY         ALWAYS
##  Min.   :0.0290   Min.   :0.1150
##  1st Qu.:0.1640   1st Qu.:0.3932
##  Median :0.2040   Median :0.4970
##  Mean   :0.2077   Mean   :0.5081
##  3rd Qu.:0.2470   3rd Qu.:0.6138
##  Max.   :0.5490   Max.   :0.8890
```

```
## The summary for the respondents who NEVER wore a mask - masks$NEVER  field

summary(masks$NEVER)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.03400 0.06800 0.07994 0.11300 0.43200
```

```
n <- density(masks$NEVER) # returns the density data
plot(n, main="Kernel Density of Respondents who NEVER wore masks")
polygon(n, col="red", border="blue")
```

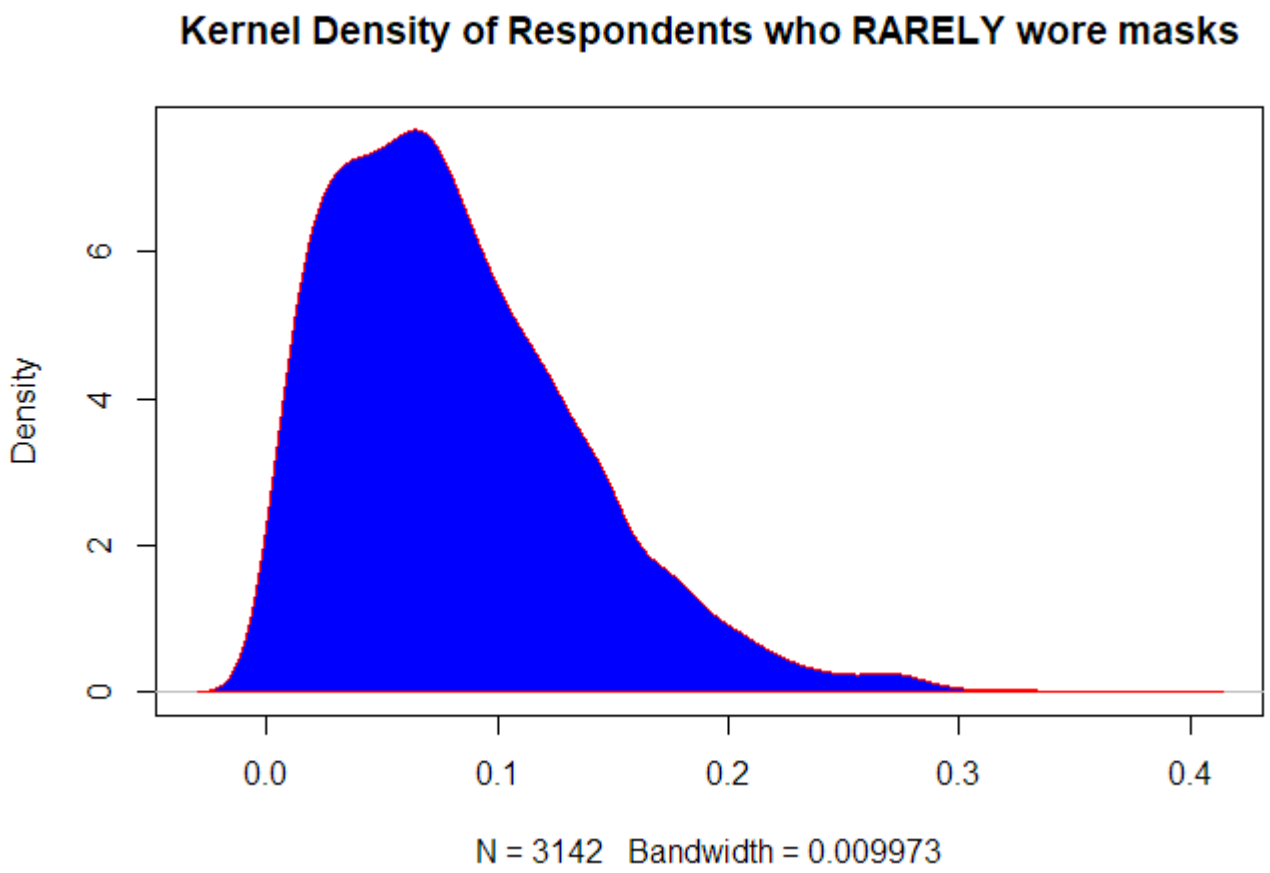


```
## The summary for the respondents who RARELY wore a mask - masks$RARELY  field

summary(masks$RARELY)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.04000 0.07300 0.08292 0.11500 0.38400
```

```
r <- density(masks$RARELY) # returns the density data
plot(r, main="Kernel Density of Respondents who RARELY wore masks")
polygon(r, col="blue", border="red")
```

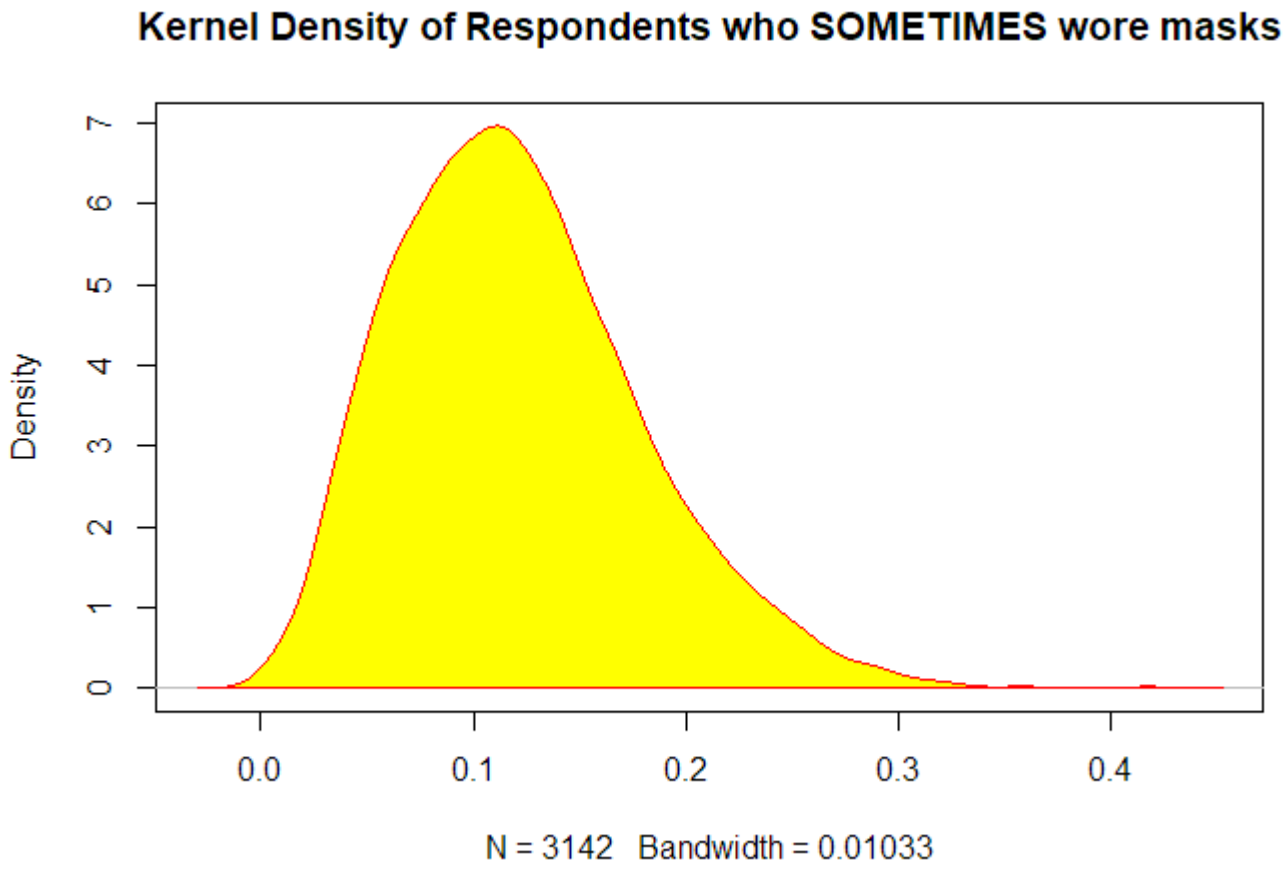



```
## The summary for the respondents who SOMETIMES wore a mask - masks$SOMETIMES  field

summary(masks$SOMETIMES)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0010  0.0790  0.1150  0.1213  0.1560  0.4220
```

```
s <- density(masks$SOMETIMES) # returns the density data
plot(s, main="Kernel Density of Respondents who SOMETIMES wore masks")
polygon(s, col="yellow", border="red")
```

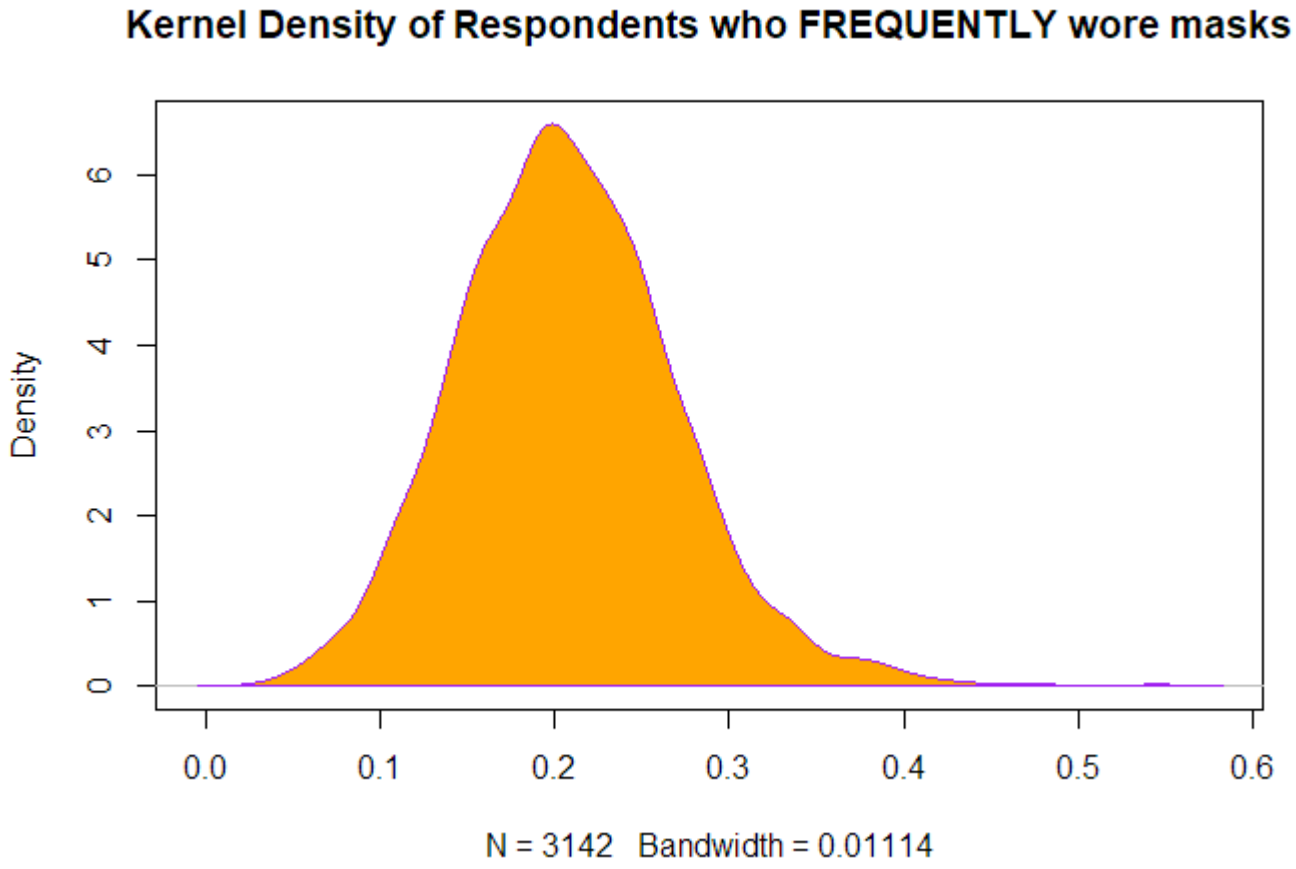



```
## The summary for the respondents who FREQUENTLY wore a mask - masks$FREQUENTLY  field

summary(masks$FREQUENTLY)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0290  0.1640  0.2040  0.2077  0.2470  0.5490
```

```
f <- density(masks$FREQUENTLY) # returns the density data
plot(f, main="Kernel Density of Respondents who FREQUENTLY wore masks")
polygon(f, col="orange", border="purple")
```

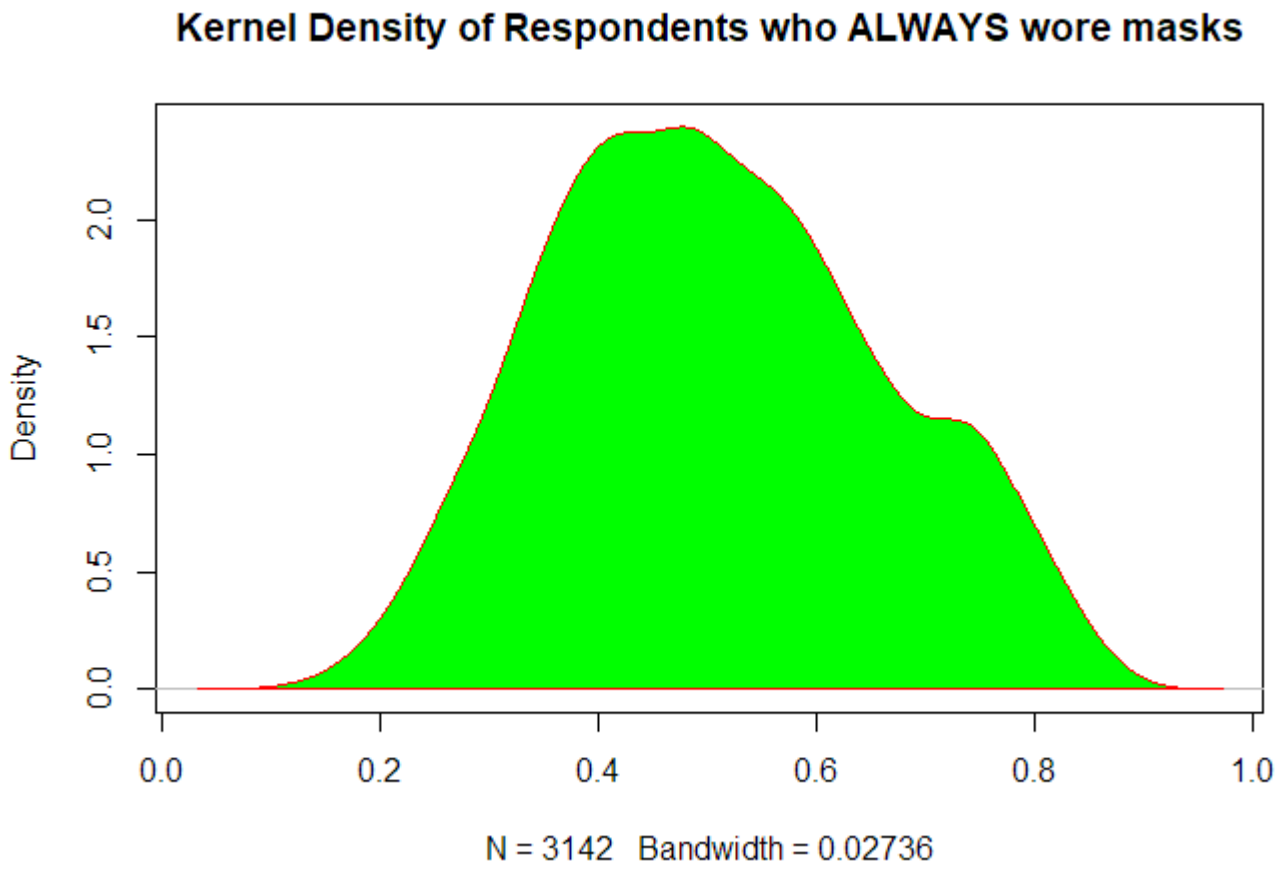


```
## The summary for the respondents who ALWAYS wore a mask - masks$ALWAYS  field

summary(masks$ALWAYS)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1150  0.3932  0.4970  0.5081  0.6138  0.8890
```

```
a <- density(masks$ALWAYS) # returns the density data
plot(a, main="Kernel Density of Respondents who ALWAYS wore masks")
polygon(a, col="green", border="red")
```



PROPOSAL CONCLUSION

This scope of this project and the datasets used allow for a fairly indepth analysis of the COVID-19 Pandemic. The brief analysis of the masks dataset above is just a small subset "tibble" of the insights we hope to draw out. This project is providing a learning platform for my R programming and while the process is sometimes painstaking, it is providing a soild foundation and rewarding experience as I create a tangible product with real data.