Eric Holberg 115 Final Project 2021

Eric Holberg

6/14/2021

1.) Describe the dataset and why you selected it for this project.

- The dataset that I choose was the COVID-19 dataset. I wanted to do this dataset because it has had a huge impact on the world. The functioning of our society has changed dramatically during this pandemic. Looking into the data that we had on COVID-19 seemed like it would be interesting and provide me with greater understanding of what the world has gone through.

2.) Describe any processing problems you identified with the data and how you overcame those issues.

- One of the issues that I ran into was that all of the data was cumulative for one of the datasets that contained deaths and cases for every county in the US. The other dataset that I am using is a survey of mask use that was done in June 2020. I wanted to limit the data that I was using in the cases and deaths dataset to a time frame that was close to the time that the survey was taken. Another issue was merging the two datasets so that the mask survey data was with the cases and deaths dataset.

3.) Describe your 'Big Question' and why the data is a good choice to answer it.

- The 'Big Question' that I chose was to see how mask use impacted death rates in the US. The reason that this is a good dataset for answering this question is it contains all of the cases and deaths in the US. There is also a dataset that has a survey for how likely someone is to wear a mask.

4.) Describe the results of your exploratory analysis and what preliminary conclusions you were able to draw based on this analysis.

- The results of my exploratory analysis was a little lackluster. It was at this point that I realized that a lot more data was going to be need to reach the conclusions that I wanted to reach. For instance a population dataset would be needed to see the impact of masks on infection rate.

5.) Describe how you selected the methodology for your analysis of the big question and the pros and cons of that method and any alternative methods you considered.

- The methodology for my analysis was cluster analysis. The analysis was able to show that the data was grouped into two clusters. The survey results and the death and case results. There was no linear relationship between these two variables. This was a classification problem as half of the data was a survey.

6.) Describe your final conclusions based on your analysis and support them with analytics on your dataset.

- Unfortunately the conclusion that I reached was that to make meaningful prediction for this I would need to use a non-linear analysis which is beyond my current skill set. Another conclusion that i reached was that more data would be needed. A population dataset would be essential to make limited predictions. More data about population density and transmissibility would also be needed. It is clear to me that these relationships exist even from my limited analysis. Take the map plots of deaths per case and total deaths. King county where the highest number of deaths occurred was not the highest death per case. With more data these relationships could be explored.

7.) Describe any additional analyses that you would have liked to carry out and any additional data that would have been needed in order to extend your analysis.

- One issue that I ran into is that the data was not linear related so I would like to do a non linear analysis to build a model. More data would also be need to be able to draw meaningful conclusions, one instance of this would be a population dataset to be able to see the effect of masks on infection rate. I know that modeling infection rates can get very complicated using epidemiological models.

```
library("ggplot2", "pdflatex")
library("cluster")

require(ggplot2)
require(GGally)
```

```
## Loading required package: GGally
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
require(gridExtra)
```

```
## Loading required package: gridExtra
```

```
require(tidyr)
```

```
## Loading required package: tidyr
```

```
require(Rmisc)
```

```
## Loading required package: Rmisc
```

```
## Loading required package: lattice
```

```
## Loading required package: plyr
```

```
require(corrplot)
```

```
## Loading required package: corrplot
```

```
## corrplot 0.89 loaded
```

```
require(lubridate)
```

```
## Loading required package: lubridate
```

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

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

```
require(tibble)
```

```
## Loading required package: tibble
```

```
require(readr)
```

```
## Loading required package: readr
```

```
require(data.table)
```

```
## Loading required package: data.table
```

```
##
## Attaching package: 'data.table'

## The following objects are masked from 'package:lubridate':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday, week,
##     yday, year
```

```
require(factoextra)
```

```
## Loading required package: factoextra

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

```
require(usmap)
```

```
## Loading required package: usmap
```

```
require(dplyr)
```

```
## Loading required package: dplyr

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:data.table':
##
##     between, first, last

## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## 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
```

```
county_deaths <-read.csv("us_county_deaths1.csv")

mask_use <- read.csv("mask_use.csv")
```

```
maskdeath <- county_deaths
#rename variables
names(mask_use)[1] <- "fips"
names(maskdeath)[1]<- "date"
#combine the datasets that i wanted to use
maskdeath <- merge(maskdeath, mask_use, by.x = c("fips"), by.y = c("fips"))
#format the dates to be more useful
maskdeath$date <- as.Date(maskdeath$date, format = "%m/%d/%Y")
```

- One of the issues that I struggled with was where to limit my data. Both CSV's are for every county in the US and the us_counties.csv has data from the beginning of the pandemic to now. I chose to limit the data to Washington and to the dates that were a month before and after when the mask survey was done. The dates were chosen because that is when I thought that peoples responses to the survey would closely match reality.

```
#For ease of changing the dates in the future
date.begin <- "2020-05-01"
date.end <- "2020-07-31"

maskdeath <- maskdeath %>%
  select(fips, date, county, state, cases, deaths, NEVER, RARELY, SOMETIMES,
         FREQUENTLY, ALWAYS) %>%
  filter(date >= date.begin & date <= date.end)

maskdeath <-maskdeath[order(maskdeath$fips, maskdeath$date),]

#Narrow results to WA based on fips
maskdeath <- maskdeath %>%
  select(fips, date, county, state, cases, deaths, NEVER, RARELY, SOMETIMES,
         FREQUENTLY, ALWAYS) %>%
  dplyr::filter(fips >= "5300" & fips <= "5399")

wamaskuse <- mask_use %>%
  select_all() %>%
  dplyr::filter(fips >= "5300" & fips <= "5399")
```

- Adding columns that showed breakdowns of variables per day and cumulative for the time period needed to be done to accurately draw conclusions. The deaths column was cumulative from the beginning of the pandemic, as was cases.

```
#This creates a deaths per day column
maskdeath <- maskdeath %>% dplyr::group_by(fips) %>%
  dplyr::mutate(deaths.day = deaths- lag(deaths))

maskdeath <- maskdeath %>% dplyr::group_by(fips) %>%
  dplyr::mutate(cases.day = cases - lag(cases))
#Gets rid of NA values
maskdeath <- maskdeath %>% mutate_if(is.numeric, replace_na, 0)
```

```
## `mutate_if()` ignored the following grouping variables:
## Column `fips`
```

```
#Creates a column that is a cumulative column of deaths per day
maskdeath <- maskdeath %>% dplyr::group_by(fips) %>%
  dplyr::mutate(cdeath = cumsum(deaths.day))

maskdeath <- maskdeath %>% dplyr::group_by(fips) %>%
  dplyr::mutate(ccases = cumsum(cases.day))

#Create a column that is a death per case
maskdeath <- maskdeath %>% dplyr::group_by(fips) %>%
  dplyr::mutate(deaths.per.case = round(cdeath / ccases, digits = 3))

#Make sure the NaN and infinite values are 0 so they don't
#interfere with calculations
maskdeath$deaths.per.case[is.nan(maskdeath$deaths.per.case)] <- 0
maskdeath$deaths.per.case[is.infinite(maskdeath$deaths.per.case)] <- 0
```

```
maskdeath %>% dplyr::ungroup(fips)
```

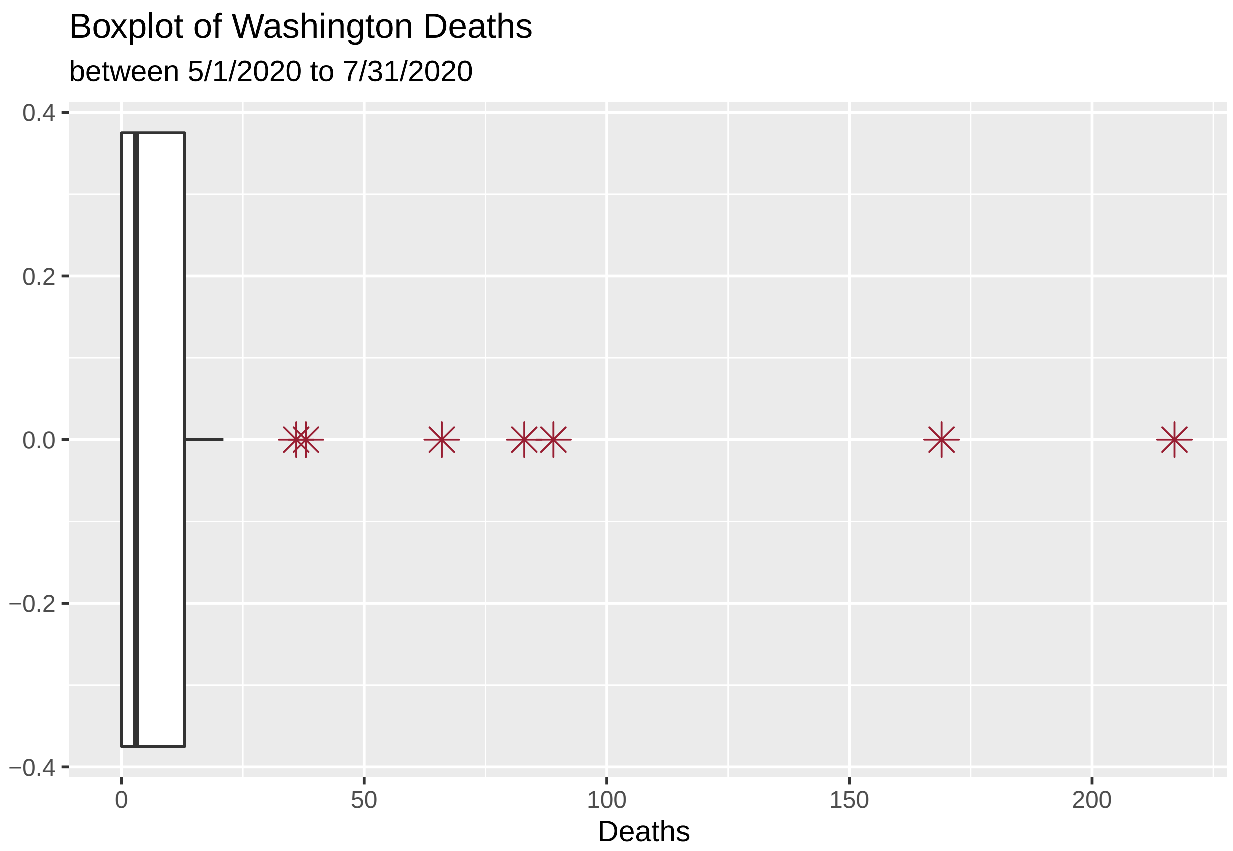
```
## # A tibble: 3,516 x 16
##     fips date       county state  cases deaths NEVER RARELY SOMETIMES FREQUENTLY
##    <int> <date>     <chr>  <chr>  <dbl>  <dbl> <dbl>  <dbl>     <dbl>      <dbl>
##  1 53001 2020-05-01 Adams  Washi~    47      0 0.077  0.051     0.072      0.076
##  2 53001 2020-05-02 Adams  Washi~    48      0 0.077  0.051     0.072      0.076
##  3 53001 2020-05-03 Adams  Washi~    48      0 0.077  0.051     0.072      0.076
##  4 53001 2020-05-04 Adams  Washi~    48      0 0.077  0.051     0.072      0.076
##  5 53001 2020-05-05 Adams  Washi~    48      0 0.077  0.051     0.072      0.076
##  6 53001 2020-05-06 Adams  Washi~    49      0 0.077  0.051     0.072      0.076
##  7 53001 2020-05-07 Adams  Washi~    49      0 0.077  0.051     0.072      0.076
##  8 53001 2020-05-08 Adams  Washi~    49      0 0.077  0.051     0.072      0.076
##  9 53001 2020-05-09 Adams  Washi~    49      0 0.077  0.051     0.072      0.076
## 10 53001 2020-05-10 Adams  Washi~    49      0 0.077  0.051     0.072      0.076
## # ... with 3,506 more rows, and 6 more variables: ALWAYS <dbl>,
## #   deaths.day <dbl>, cases.day <dbl>, cdeath <dbl>, ccases <dbl>,
## #   deaths.per.case <dbl>
```

```
#Create a "Summary" of the dates so the cumulatives of the variables would be
#available without using the entire dataset
wadeathc <- maskdeath %>%
  select_all %>%
  dplyr::filter(date == date.end)
```

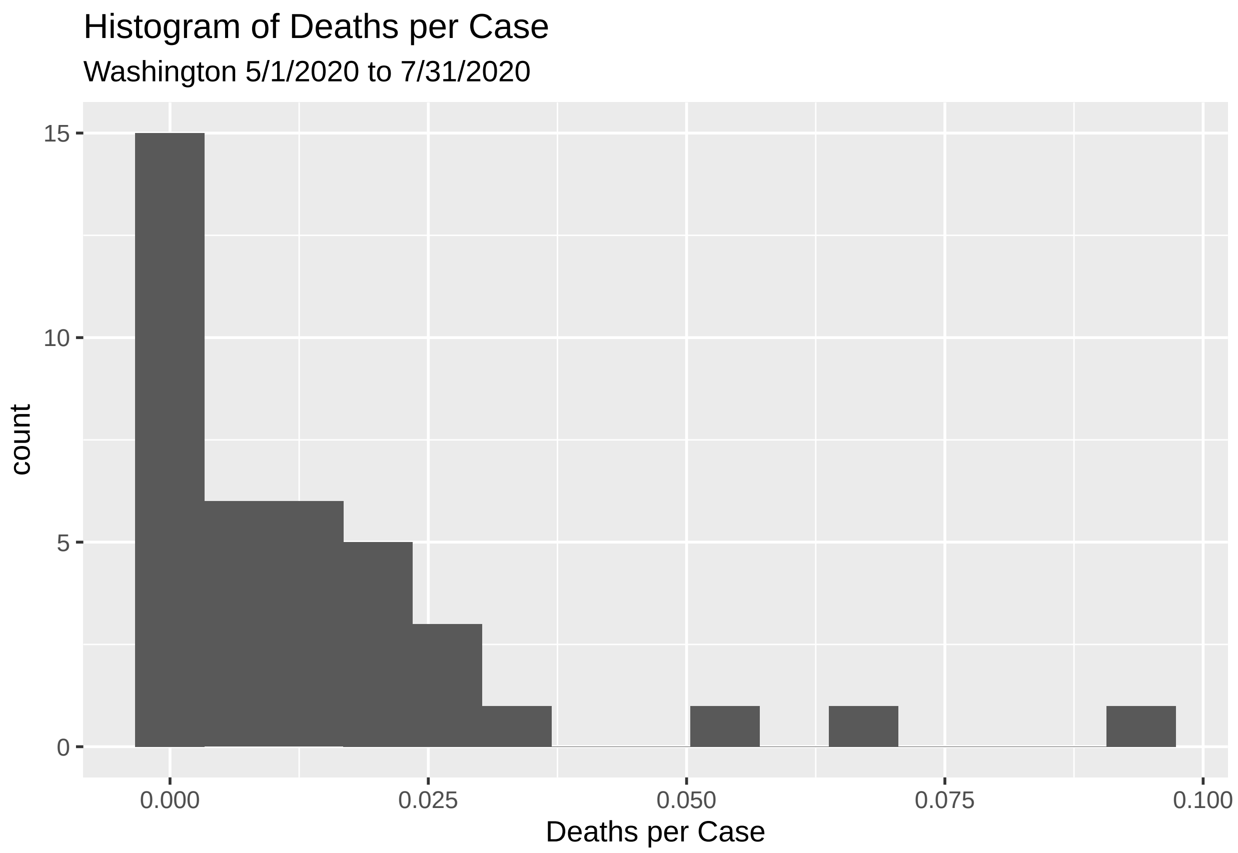
```
write.csv(maskdeath, "maskdeath13")
```

- I really wanted to find a relationship between this data. A boxplot of the deaths showed that most of the deaths were contained in a narrow band with outliers on the high end. Elimination of those outliers didn't make sense in this case because they were really the interesting data points. A histogram of the deaths per case to see what kind of distribution there was, which turned out to be skewed right. The barpot shows that the there are negitive death values that are deaths that got reclassified to another cause of death.I then used the pairs plot to see if there was any relationships that I didn't think to do and I didn't see anything that jumped out at me.

```
ggplot(wadeathc, aes(cdeath)) +
  geom_boxplot(outlier.colour="#981e32", outlier.shape=8,outlier.size=4)+
  labs(title="Boxplot of Washington Deaths",
       subtitle = "between 5/1/2020 to 7/31/2020", x="Deaths ")
```

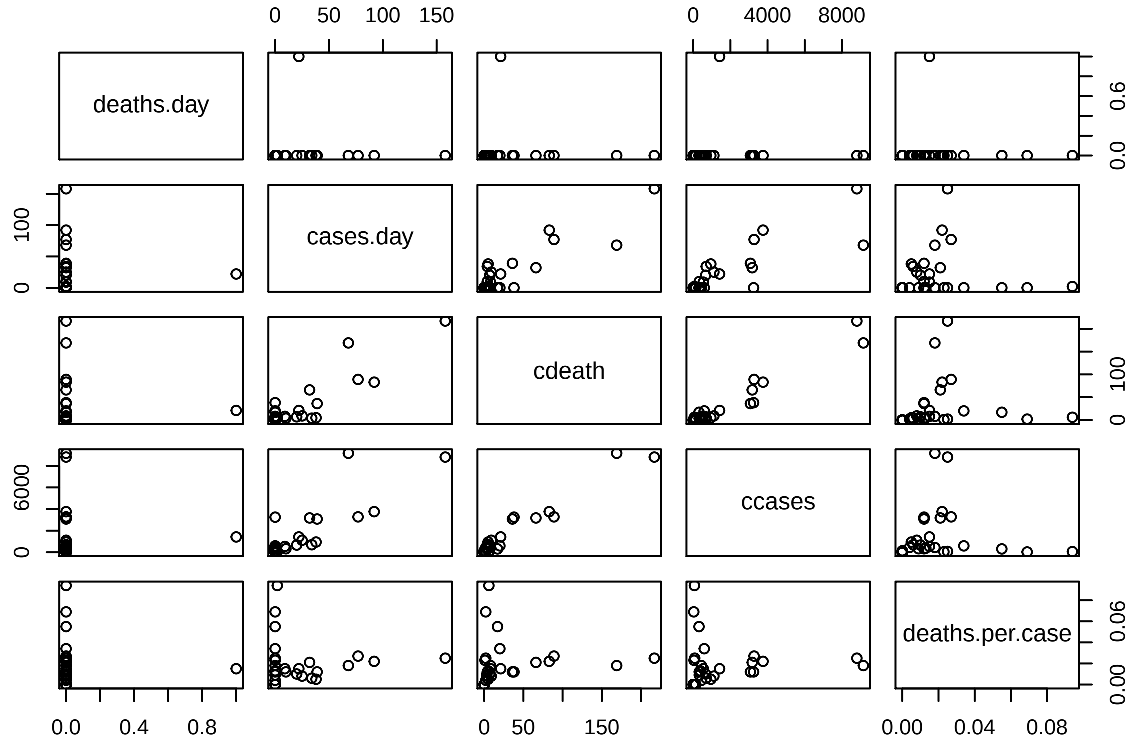


```
ggplot(wadeathc, aes(deaths.per.case))+
  geom_histogram(bins = 15)+
  labs(title = "Histogram of Deaths per Case",
       subtitle = "Washington 5/1/2020 to 7/31/2020",x = "Deaths per Case")
```

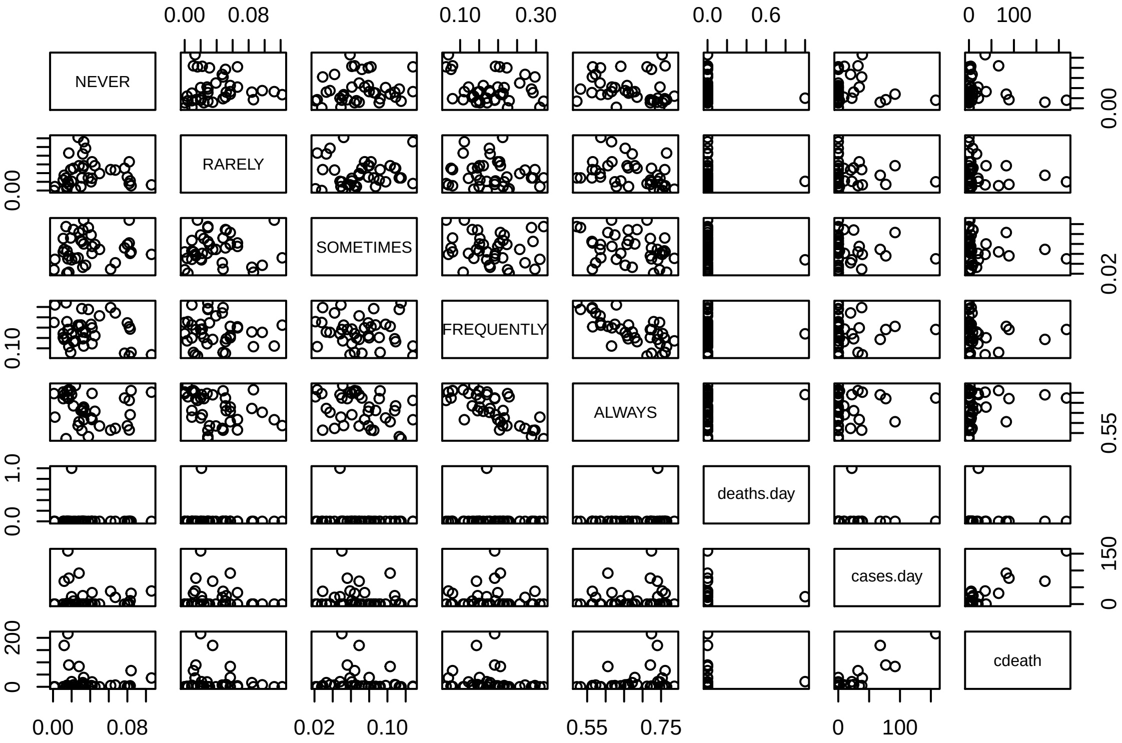


```r
p1 <- ggplot(maskdeath, aes(date, deaths))+
    geom_point(color = "orange")+
    labs(title="Scatterplot of Washington Deaths",
       subtitle = "between 5/1/2020 to 7/31/2020", x="Date", y="Deaths")


p2 <- ggplot(maskdeath, aes(deaths.day))+
  geom_bar()+
  labs(title = "Barplot of Deaths per Day", x = "Deaths per Day")

p3 <- ggplot(maskdeath, aes(date, deaths.per.case, color = ALWAYS))+
  geom_point()

p4 <- ggplot(maskdeath, aes(date, deaths.per.case, color = FREQUENTLY))+
  geom_point()


p5 <-ggplot(maskdeath, aes(date, deaths.per.case, color = SOMETIMES))+
  geom_point()

p6 <-ggplot(maskdeath, aes(date, deaths.per.case, color = RARELY))+
  geom_point()

p7 <-ggplot(maskdeath, aes(date, deaths.per.case, color = NEVER))+
  geom_point()
```

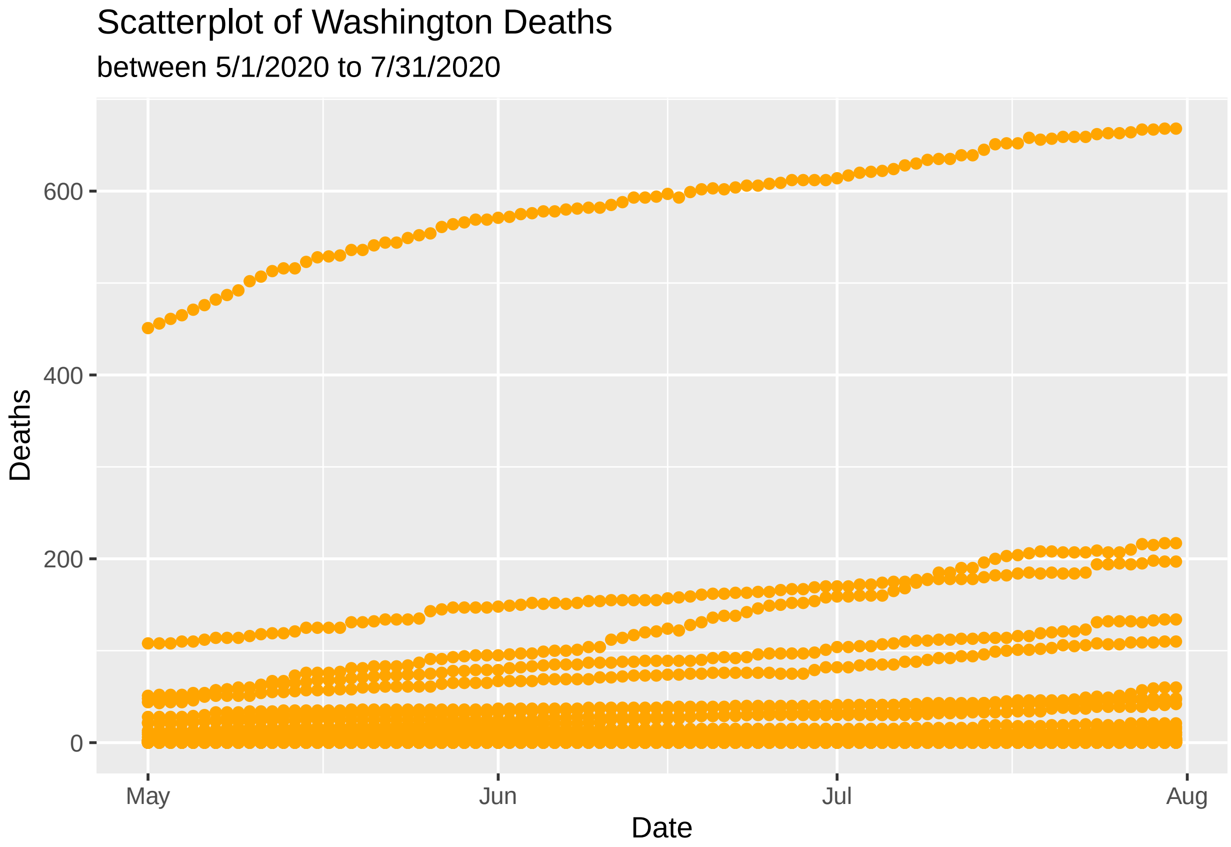
```
p8 <-ggplot(maskdeath, aes(date, deaths.per.case, color = fips))+
  geom_point()
pairs(wadeathc[12:16])
```



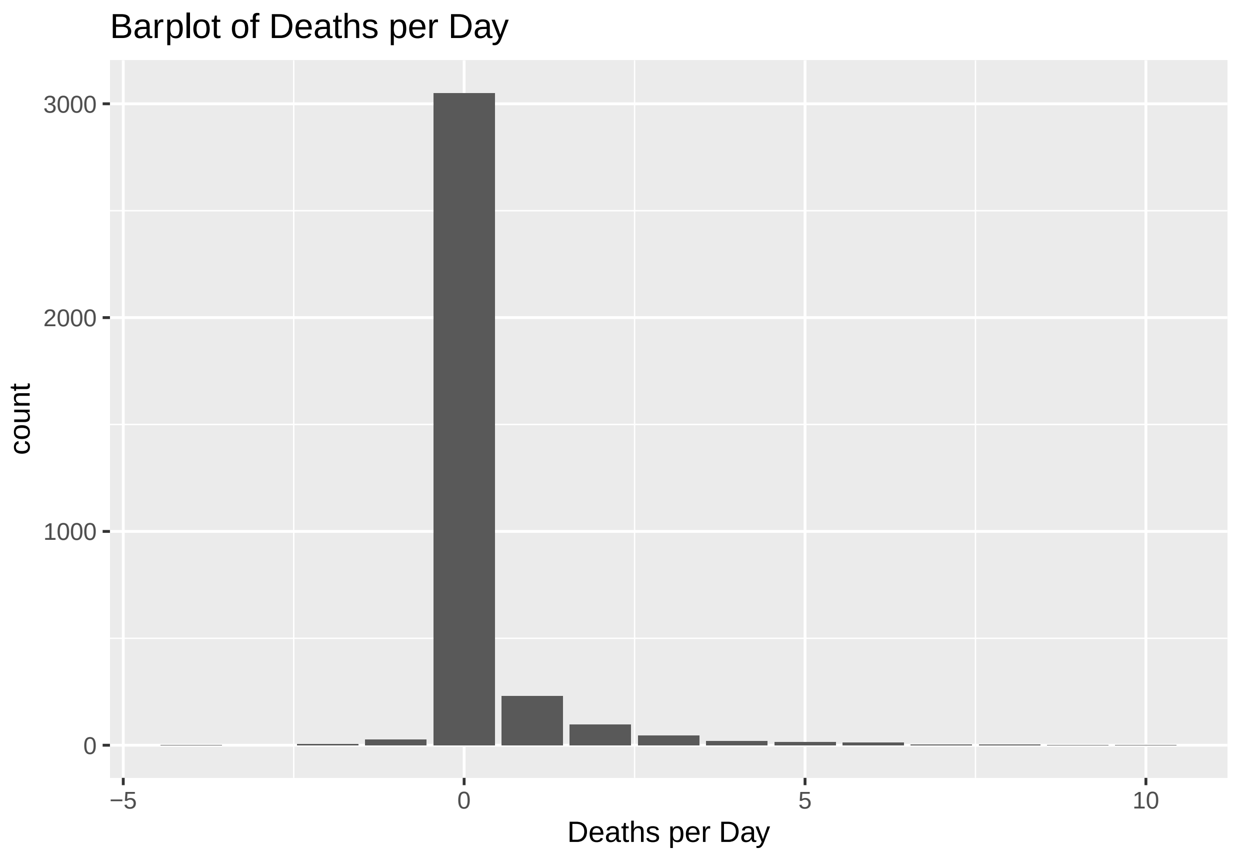
```
pairs(wadeathc[7:14])
```


0.00
0.08
0.10
0.30
0.0
0.6
0
100
NEVER
0.00
RARELY
0.00
SOMETIMES
0.02
FREQUENTLY
0.10
ALWAYS
0.55
1.0
0.0
deaths.day
cases.day
150
0
200
0
cdeath
0.00
0.08
0.02
0.10
0.55
0.75
0
100

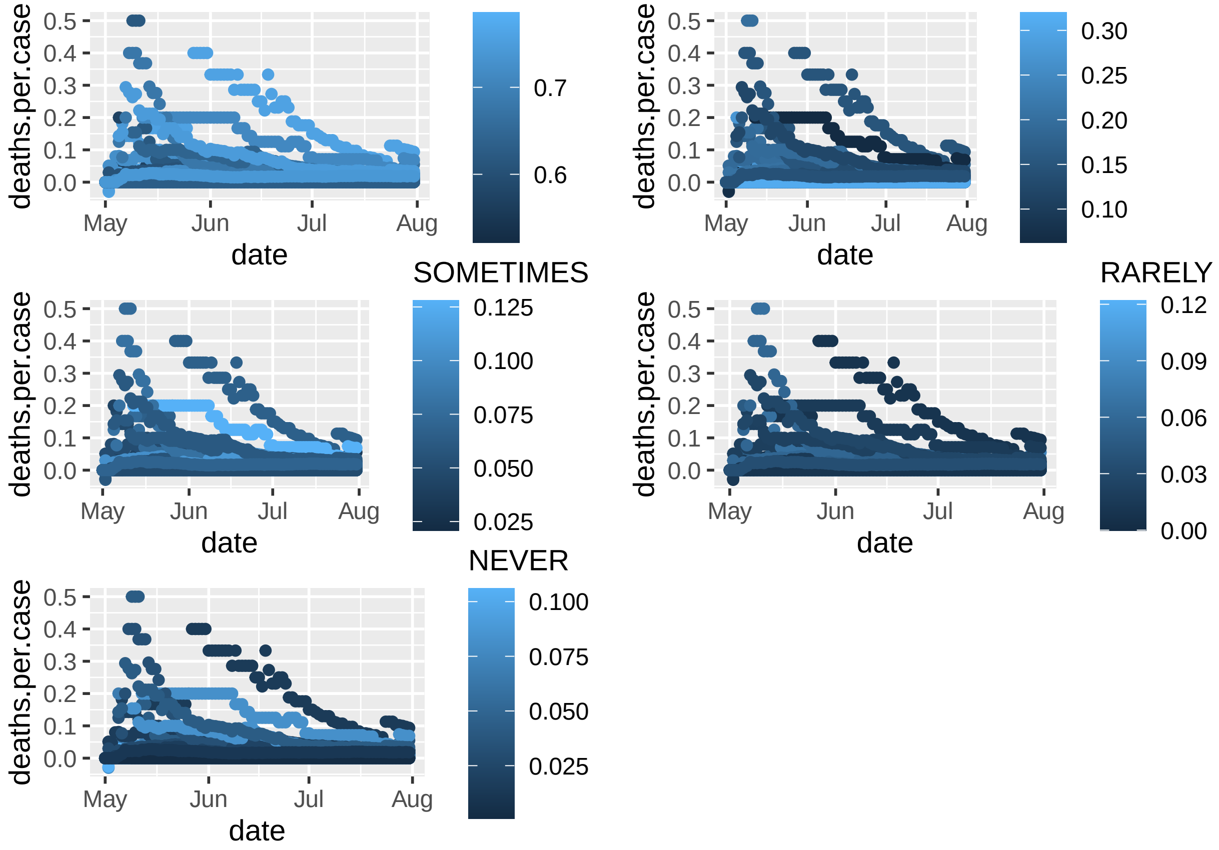
```
p1
```


Scatterplot of Washington Deaths
between 5/1/2020 to 7/31/2020
Deaths
600
400
200
0
May
Jun
Jul
Aug
Date

```
p2
```

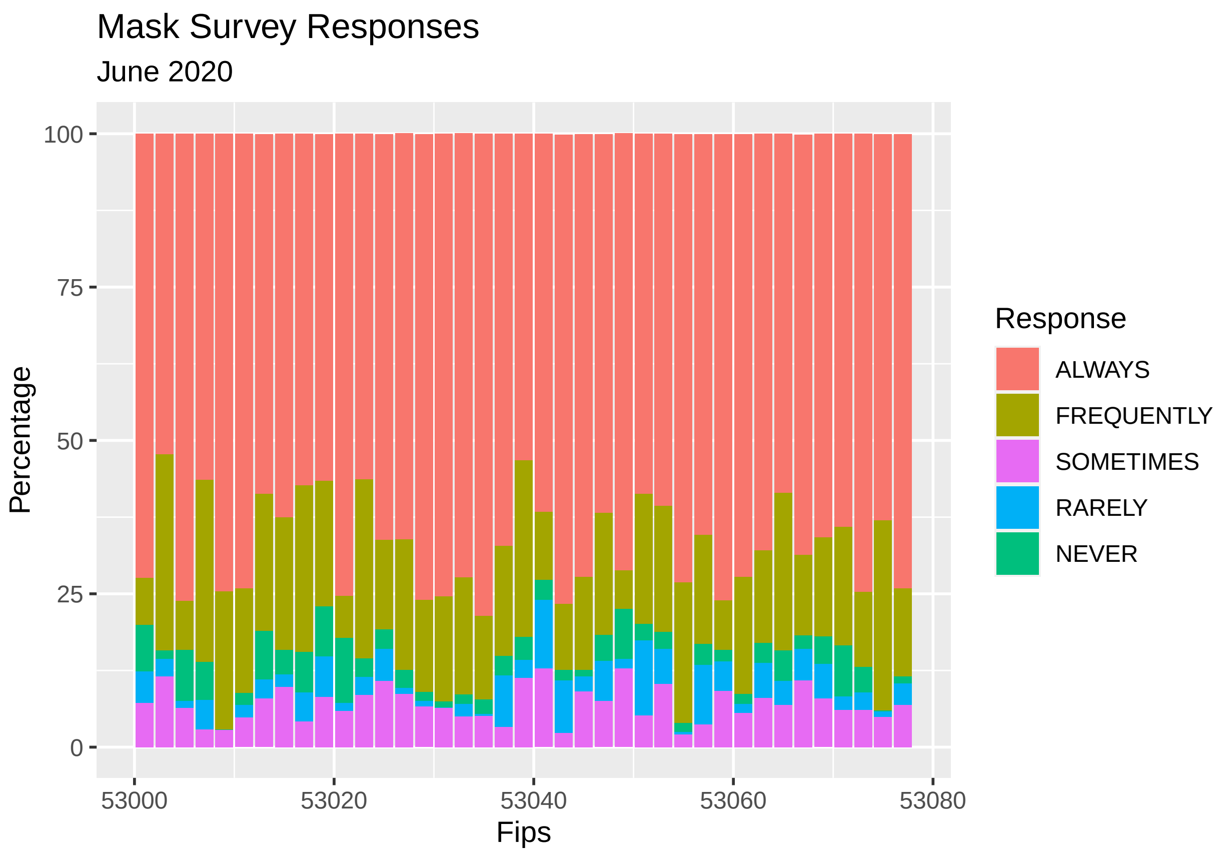


```
grid.arrange(p3,p4,p5,p6,p7, ncol=2)
```

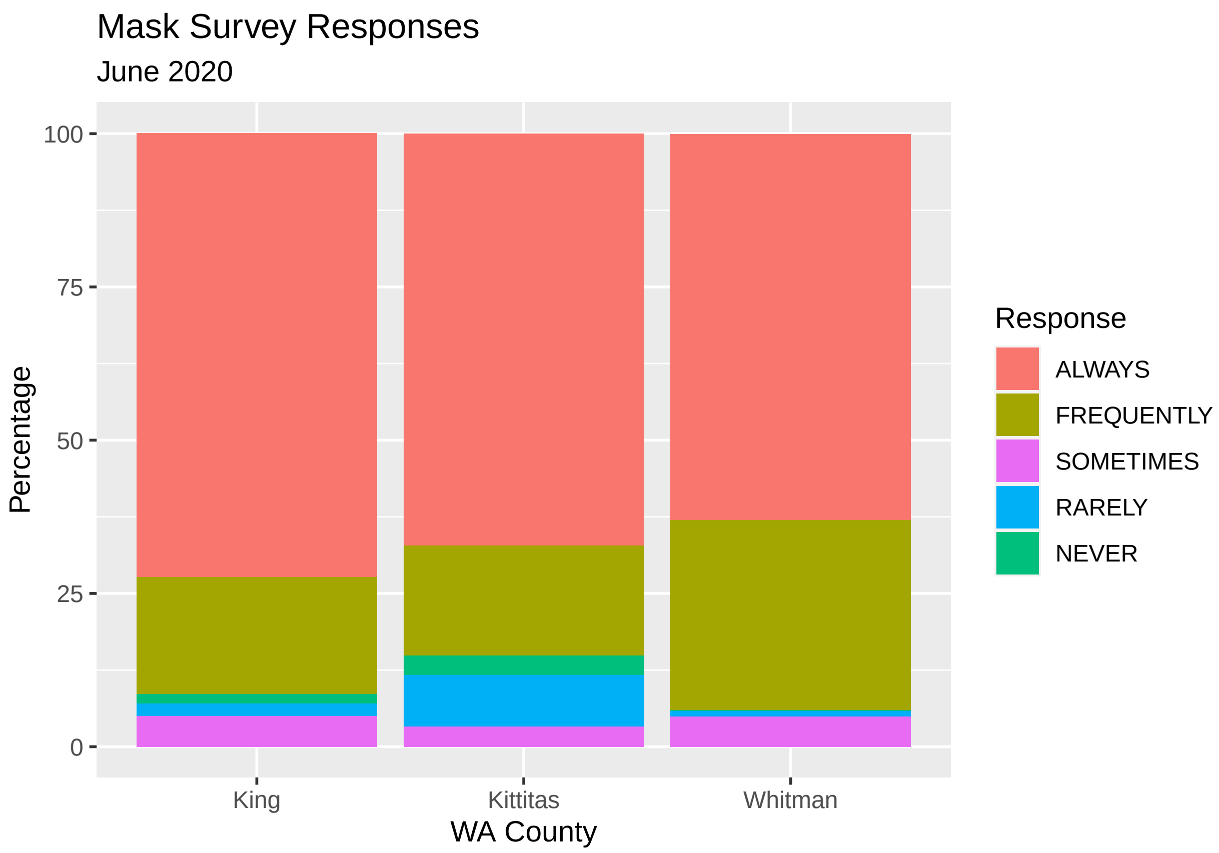


- The stacked bar chart shows the responses for the mask survey based upon fips codes

```
wamask<- wamaskuse %>% gather(Response, Percentage, NEVER:ALWAYS)
# Transform this data in %
wamask$Percentage <- wamask$Percentage * 100



response.order <- c("ALWAYS", "FREQUENTLY","SOMETIMES","RARELY","NEVER")

ggplot(wamask, aes(x = fips, y = Percentage))+
  geom_col(aes(fill = Response))+
  scale_fill_discrete(breaks = response.order)+
  labs(title = "Mask Survey Responses", subtitle = "June 2020", x = "Fips")
```

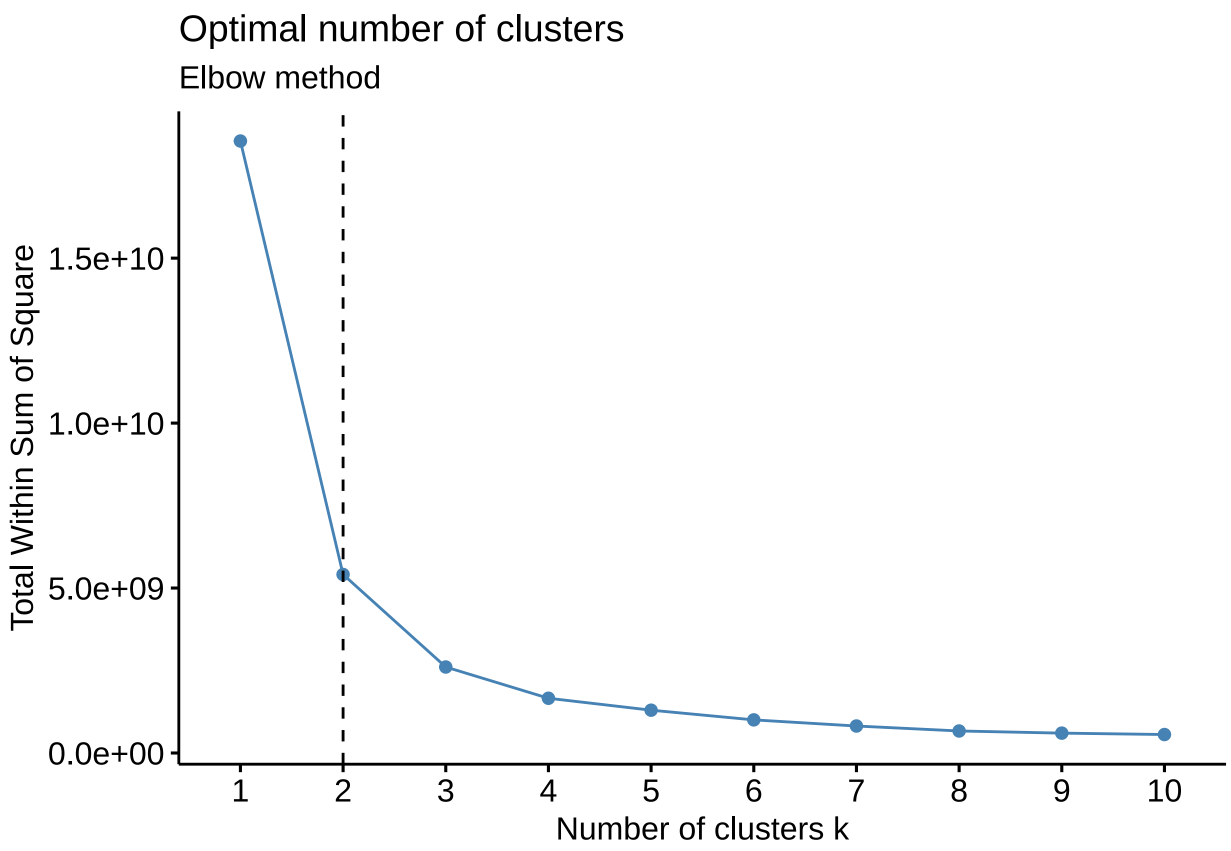


```
wamask.1 <- wamask %>%
  select_all() %>%
  dplyr::filter(wamask$fips == "53033"| wamask$fips =="53075"|
                wamask$fips =="53037")
wamask.r <- wamask.1

wamask.r$fips[wamask.r$fips == "53033"] <- "King"
wamask.r$fips[wamask.r$fips == "53037"] <- "Kittitas"
wamask.r$fips[wamask.r$fips == "53075"] <- "Whitman"

ggplot(wamask.r, aes(x = fips, y = Percentage))+
  geom_col(aes(fill = Response))+
  scale_fill_discrete(breaks = response.order)+
  labs(title = "Mask Survey Responses", subtitle = "June 2020", x = "WA County")
```

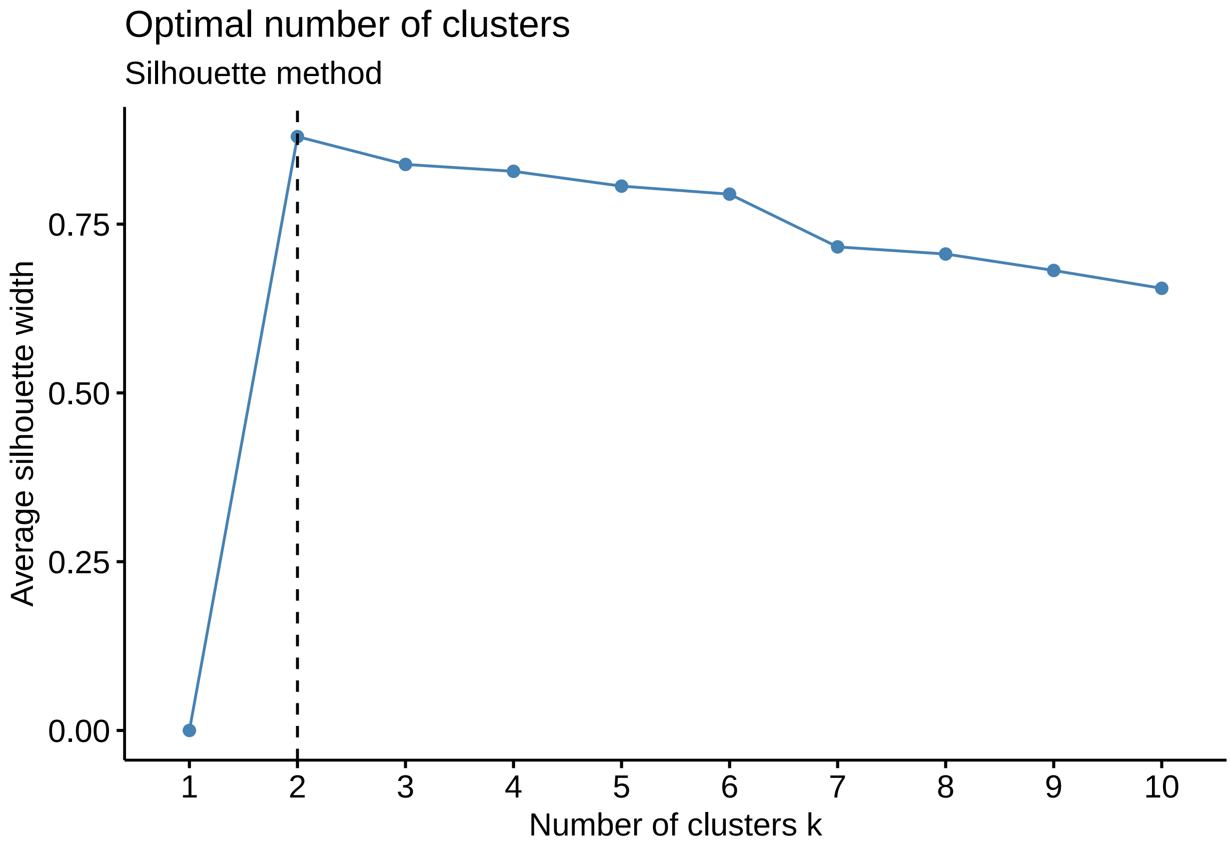


- The cluster analysis shows that the optimum number of clusters should be two. The first cluster is the mask results survey and the second cluster is the cumulative deaths and cumulative cases.

```
set.seed(54)
#Remove non-numeric for cluster analysis
clusterD <- maskdeath[c(5:11,14:15)]
cluster.maskdeath <- kmeans(clusterD,2,nstart = 50)

#Determine the optimum number of clusters
fviz_nbclust(clusterD, kmeans, method = "wss") +
    geom_vline(xintercept = 2, linetype = 2)+
  labs(subtitle = "Elbow method")
```



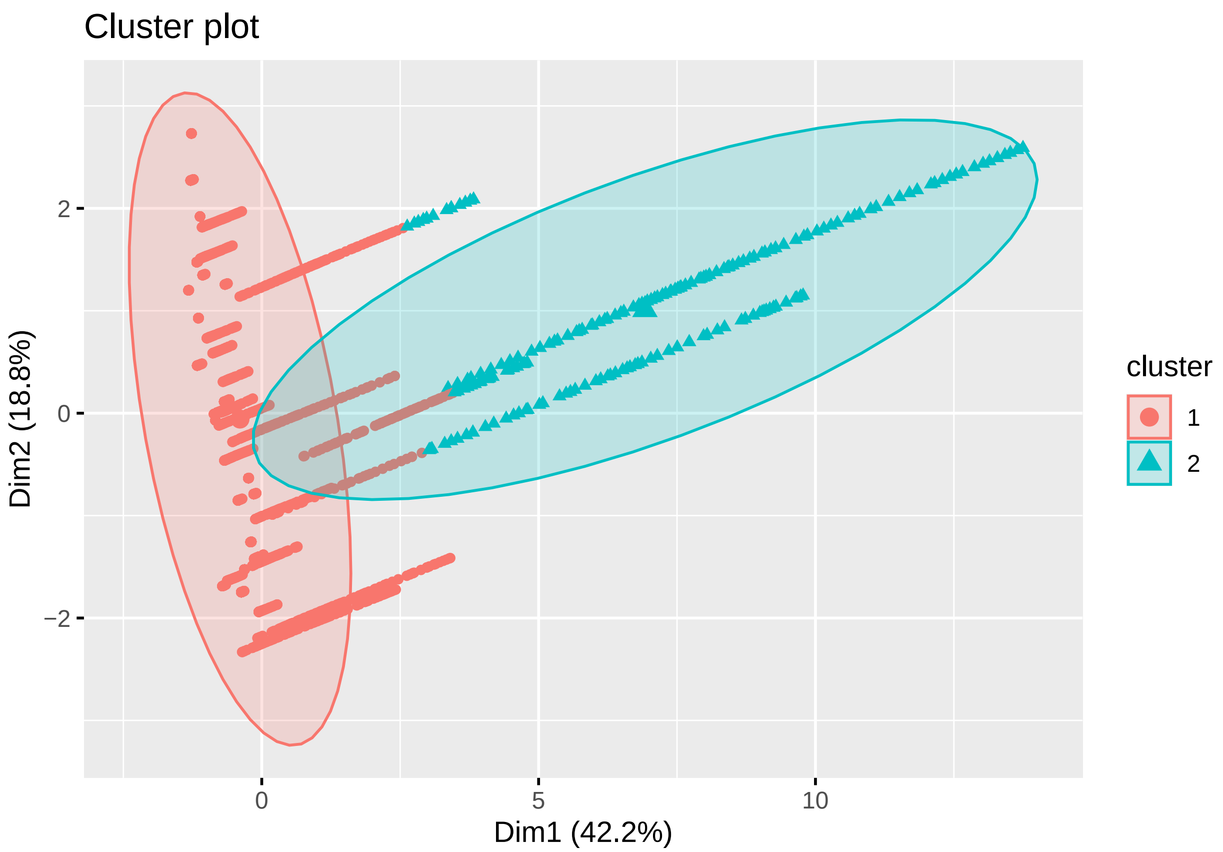
```
fviz_nbclust(clusterD, kmeans, method = "silhouette") +
    geom_vline(xintercept = 2, linetype = 2)+
  labs(subtitle = "Silhouette method")
```



```
kresults <- kmeans(clusterD, center=2, nstart = 50)
str(kresults)
```

```
## List of 9
##  $ cluster     : int [1:3516] 1 1 1 1 1 1 1 1 1 1 ...
##  $ centers     : num [1:2, 1:9] 4.09e+02 8.04e+03 1.33e+01 3.68e+02 3.93e-02 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:2] "1" "2"
##   .. ..$ : chr [1:9] "cases" "deaths" "NEVER" "RARELY" ...
##  $ totss       : num 1.86e+10
##  $ withinss    : num [1:2] 2.83e+09 2.58e+09
##  $ tot.withinss: num 5.41e+09
##  $ betweenss   : num 1.31e+10
##  $ size        : int [1:2] 3327 189
##  $ iter        : int 1
##  $ ifault      : int 0
##  - attr(*, "class")= chr "kmeans"
```

```
fviz_cluster(kresults, clusterD, geom = c("point"),ellipse.type = "norm")
```



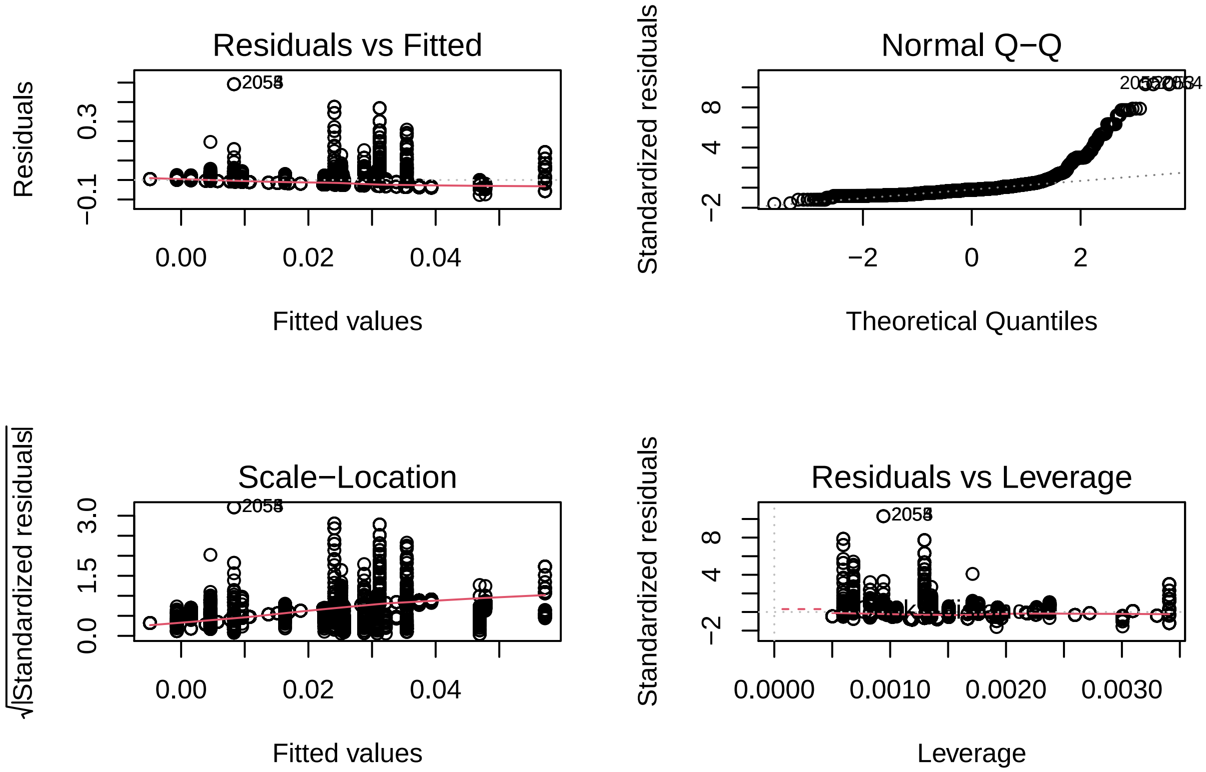
- The data that I used did not have a linear relationship. The lack of linearity is shown by the quantile vs quantile plot that should be linear but is non-linear.

```
mlinear <- lm(deaths.per.case ~ ALWAYS + FREQUENTLY + SOMETIMES + RARELY +
                NEVER, maskdeath)
summary(mlinear)
```

```
##
## Call:
## lm(formula = deaths.per.case ~ ALWAYS + FREQUENTLY + SOMETIMES +
##     RARELY + NEVER, data = maskdeath)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.07692 -0.02292 -0.01030  0.00488  0.49169
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -5.061      1.414  -3.580 0.000348 ***
## ALWAYS         5.121      1.414   3.620 0.000299 ***
## FREQUENTLY     4.955      1.414   3.505 0.000462 ***
## SOMETIMES      5.178      1.411   3.669 0.000247 ***
## RARELY         4.854      1.420   3.419 0.000636 ***
## NEVER          5.116      1.407   3.636 0.000280 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

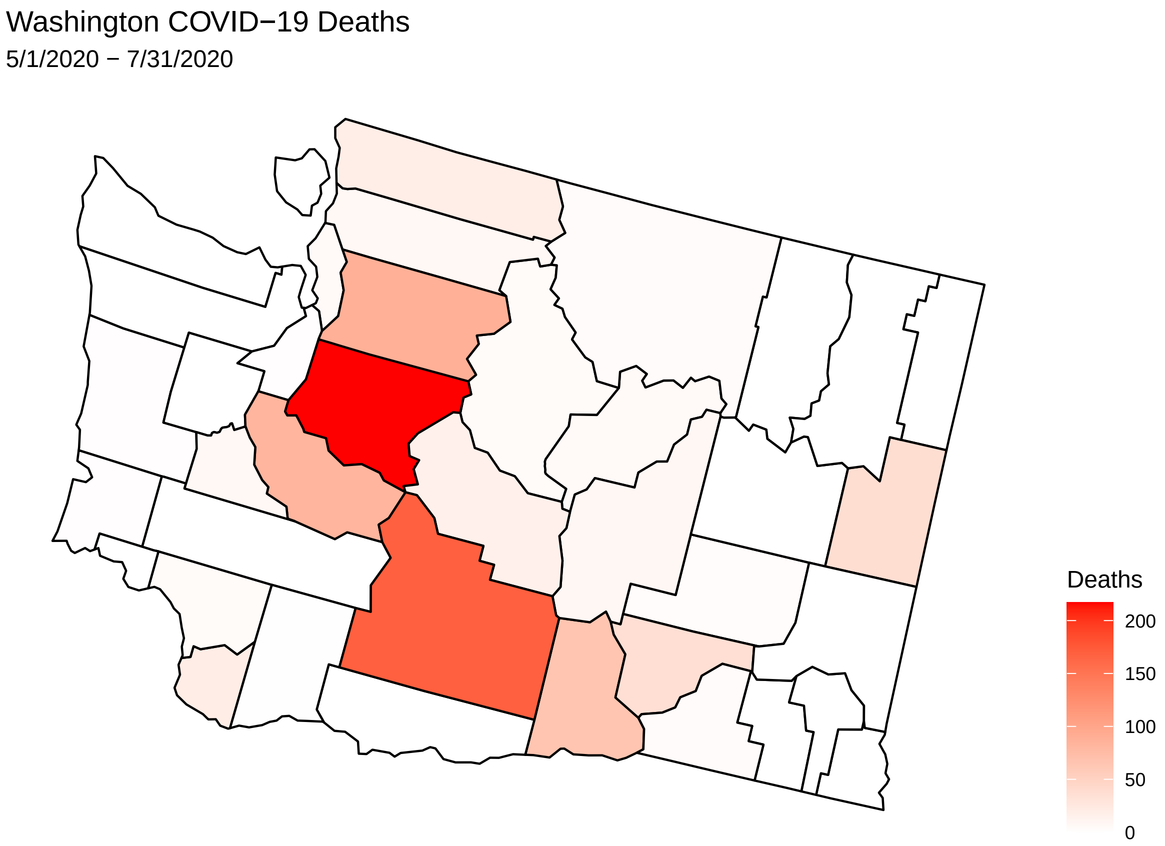
```
## Residual standard error: 0.04776 on 3510 degrees of freedom
## Multiple R-squared:  0.07966,    Adjusted R-squared:  0.07835
## F-statistic: 60.76 on 5 and 3510 DF,  p-value: < 2.2e-16
```

```r
par(mfrow=c(2,2))
plot(mlinear)
```

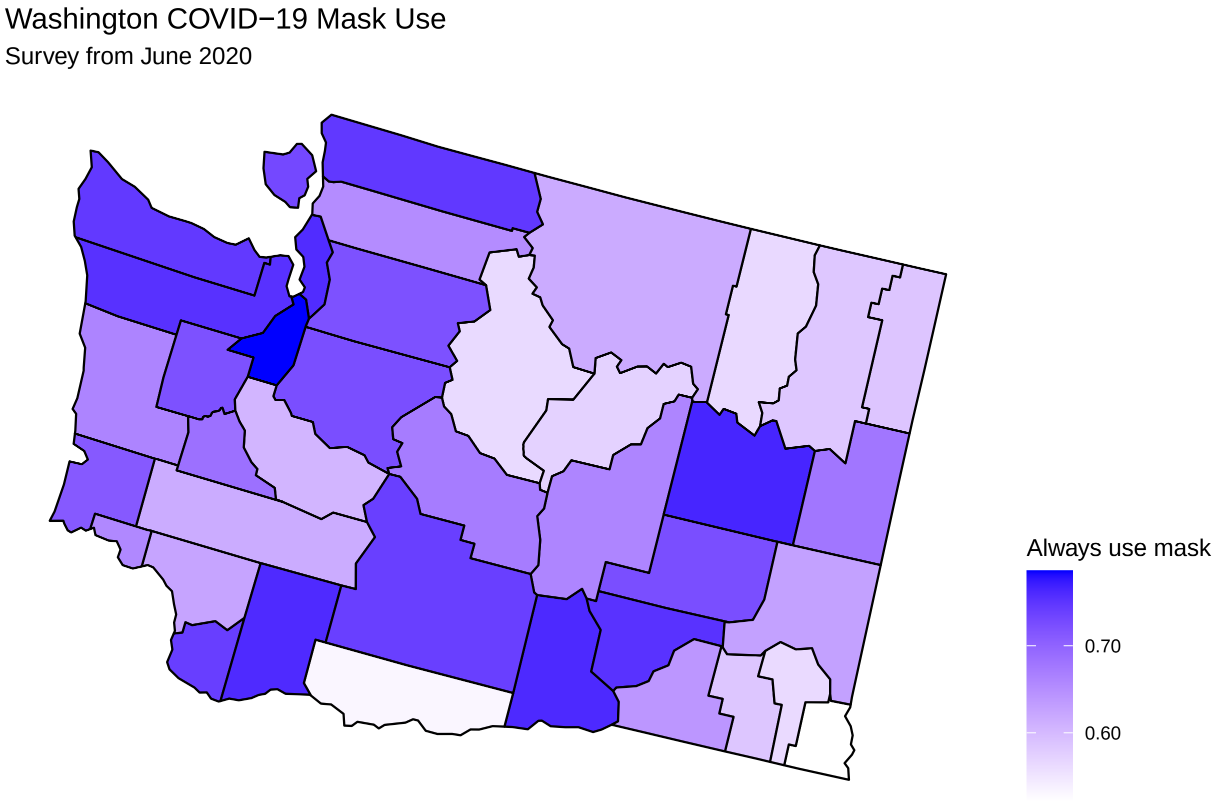


- The data in map form as it would be easier to understand. This also allows for geographical relationships to be shown. Even without more data it is clear that some areas were differently effected by the pandemic.

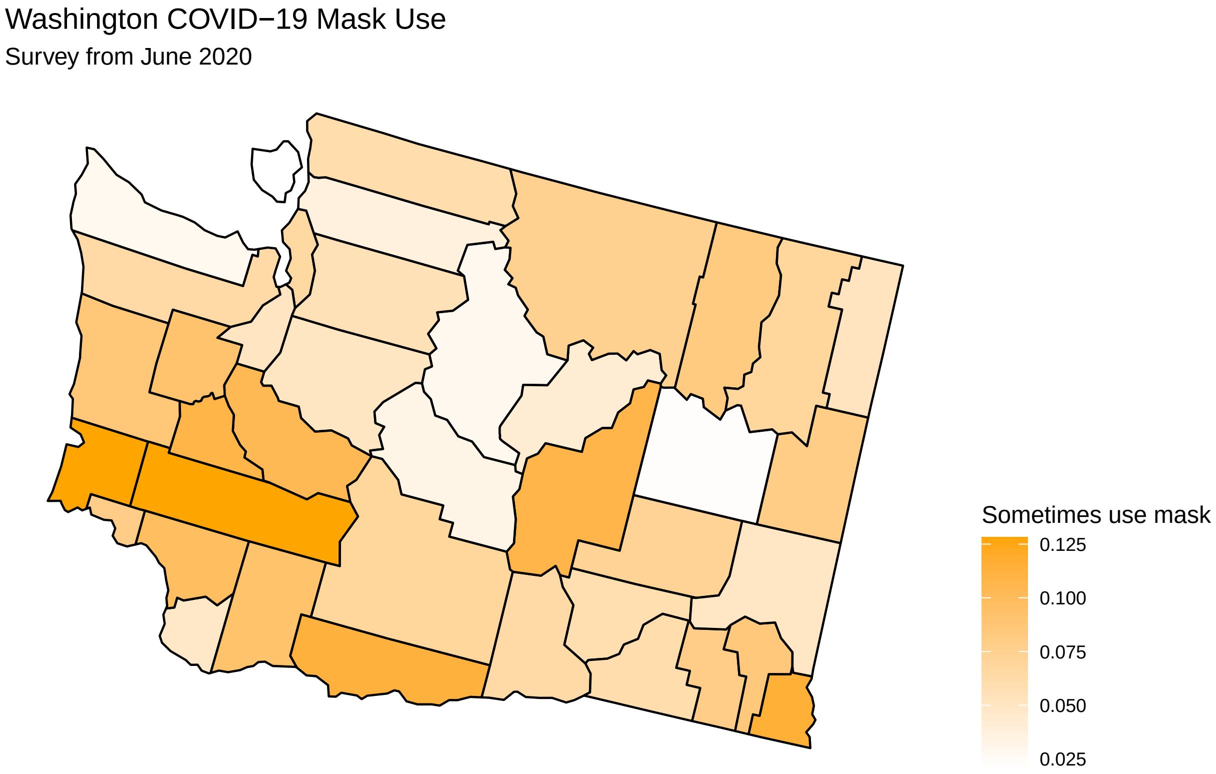
```r
usmap::plot_usmap(data = wadeathc, values = "cdeath", "counties",
                  include = c("WA"), color = "black")+
  scale_fill_continuous(low = "white", high = "red", name = "Deaths",
                        label = scales::comma) +
  labs(title = "Washington COVID-19 Deaths",
       subtitle = "5/1/2020 - 7/31/2020") +
  theme(legend.position = "right")
```



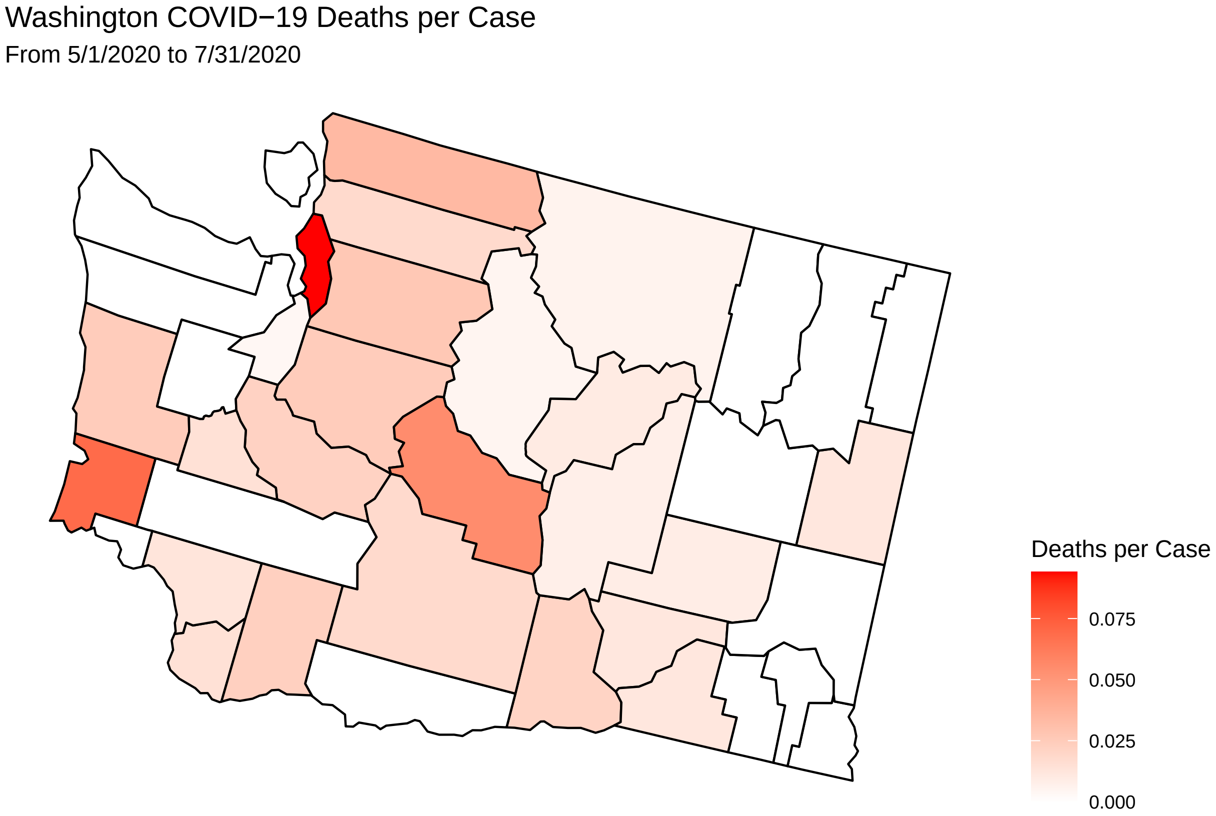
```
usmap::plot_usmap(data = wadeathc, values = "ALWAYS", "counties",
                  include = c("WA"), color = "black")+
  scale_fill_continuous(low = "white", high = "blue",
                        name = "Always use mask", label = scales::comma) +
  labs(title = "Washington COVID-19 Mask Use",
       subtitle = "Survey from June 2020") +
  theme(legend.position = "right")
```



```
usmap::plot_usmap(data = wadeathc, values = "SOMETIMES", "counties",
                  include = c("WA"), color = "black")+
  scale_fill_continuous(low = "white", high = "orange",
                        name = "Sometimes use mask", label = scales::comma) +
  labs(title = "Washington COVID-19 Mask Use",
       subtitle = "Survey from June 2020") +
  theme(legend.position = "right")
```



```
usmap::plot_usmap(data = wadeathc, values = "deaths.per.case", "counties",
                  include = c("WA"), color = "black")+
  scale_fill_gradient2(low=("blue"), mid="white", high=("red"),
                       name = "Deaths per Case", label = scales::comma) +
  labs(title = "Washington COVID-19 Deaths per Case",
       subtitle = "From 5/1/2020 to 7/31/2020") +
  theme(legend.position = "right")
```


Washington COVID−19 Deaths per Case
From 5/1/2020 to 7/31/2020
Deaths per Case
0.075
0.050
0.025
0.000