Week 4 homework assignment

Anubha Nagar

Due Date (22nd September)

DSI-EDA

Professor Michael Shepherd

STEPS:

(1) Read the following article (it is for data purposes!)

NPR "CHART: The Relationship Between Seeing Discrimination And Voting For Trump" - link here - even more information on the data here (2) Your goal is to recreate the graphic titled "Perceptions Of Discrimination Track Closely With Voting Against Trump." Data for analysis is in the "week4-hw-data.csv" file. Note that this data is from a different version of the PRRI survey, and the results won't match the NPR graph exactly.

(3) Graphic replication: You should:

1. Create an .rmd file to show your analysis. Write 1 sentence for each block of code, explaining what you do in that line of code.
2. Identify and prepare the variables of interest (you may need to group, summarize, or rename variables to reflect the NPR chart.)
3. Next recreate the graphic using ggplot. Recreate everything *except*:

- the labels that are on California and Wyoming.
- the colors (you can choose your own or use the same as the article)

4. This means your plot should have the same labels on the axes, gridlines, etc.
5. Is there anything misleading about the graphic? Why or why not?
6. You can give your colleagues "hints" but do *not* give them the code for reproducing the graphic. That breaks our honor code since I asked you not to do it!

(4) What do you think about this graphic? Do you think this relationship exists? Why or why not, in your own words?

(5) Make one additional graphic of your choice using this data. Write 3-5 sentences in clear, plain language about what the graphic illustrates about the data.

Part 1

Loading the data

```
library(readr)
week4_hw_data <- read_csv("week4-hw-data.csv")
```

```
## New names:
## * `` -> ...1
```

```
## Rows: 51 Columns: 4
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): state
## dbl (3): ...1, trump, discrim
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
View(week4_hw_data)
```

Loading the packages

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v dplyr   1.0.7
## v tibble  3.1.3     v stringr 1.4.0
## v tidyr   1.1.3     v forcats 0.5.1
## v purrr   0.3.4
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
```

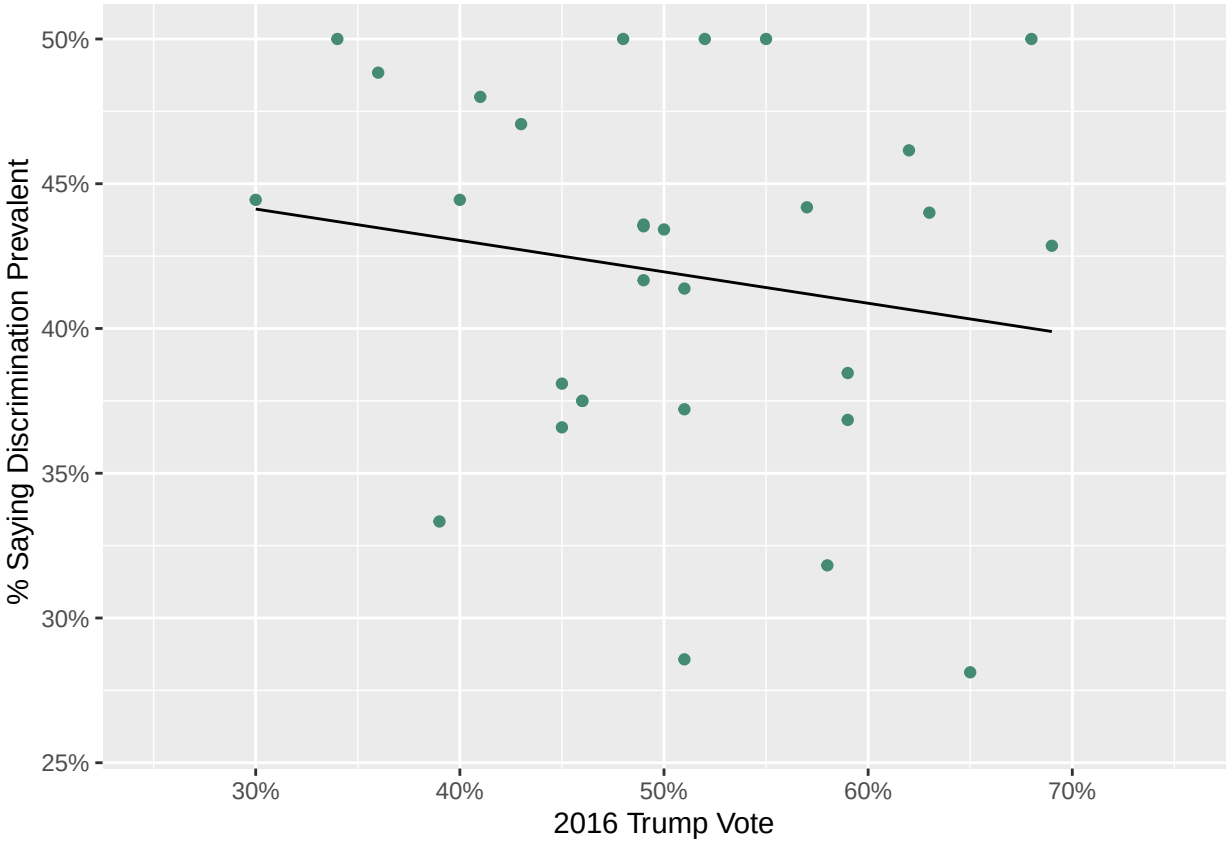
Plotting a graph to show trump voters vs discrimination, add color to the graph, add percentage to the x and y axis and make a line to show the correlation between the two variables.

```
ggplot(week4_hw_data, aes(x = trump, y = discrim))+
  geom_point(color = "aquamarine4", fill = "aquamarine4")+
  scale_x_continuous(name = "2016 Trump Vote", limits = c(0.25, 0.75, 0.5), labels = function(x){paste0
  scale_y_continuous(name = " % Saying Discrimination Prevalent", limits = c(0.26, 0.5, 0.2), labels = 
  geom_smooth(method='lm', col='black', size=0.5, se = F)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 21 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 21 rows containing missing values (geom_point).
```

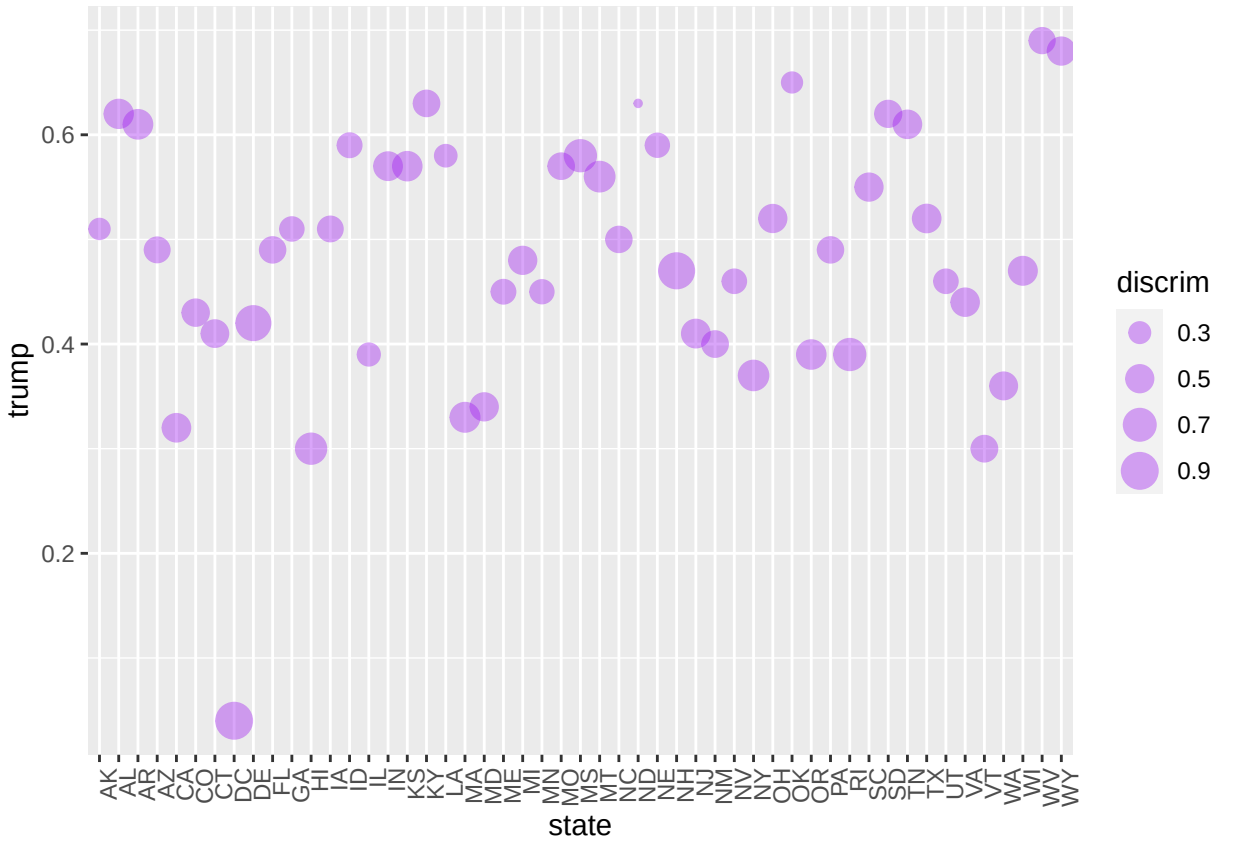


Q: *Relationship shown in graph. Is it misleading?*

The graph shows that there is a negative correlation between the votes for trump and the discrimination. Whereas if you see the points having high trump voters also have discrimination above 45 % (which is a high discrimination rate considering everyone voted for trump thinking there will be less discrimination). This is what is misleading in this graph.

Q : *Make one additional graphic of your choice using this data. Write 3-5 sentences in clear, plain language about what the graphic illustrates about the data.*

```
ggplot(week4_hw_data, aes(x = state, y = trump, size = discrim))+geom_point(col = "purple", alpha = 0.4
```



In this graph, I have plotted the number of votes for trump for each state. Furthermore each point's size shows the amount of discrimination in that state. District of Columbia (DC) has lowest Trump voters and West Virginia (WV) has highest number of Trump voters.Ironically, District of Columbia has very high discrimination rate as well (90%) which proves their decision. North Dakota(ND) has high trump votes and lowest discrimination ($< 30\%$).

Part 2

```
library(readr)
library(ggplot2)
```

```
library(readr)
winequality_red <- read_delim("winequality-red.csv",
    delim = ";", escape_double = FALSE, trim_ws = TRUE)
```

```
## Rows: 1599 Columns: 12
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ";"
## dbl (12): fixed acidity, volatile acidity, citric acid, residual sugar, chlo...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

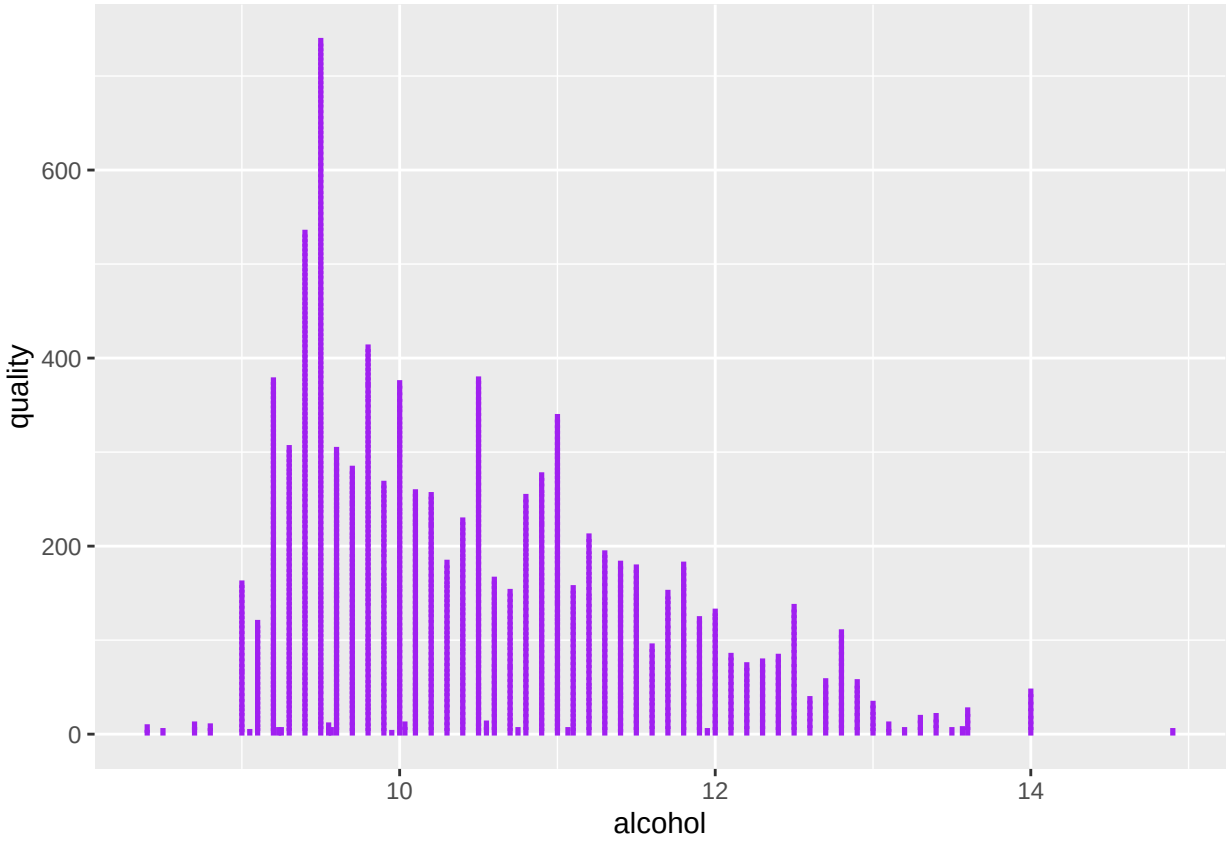
```
View(winequality_red)
head(winequality_red)
```

```
## # A tibble: 6 x 12
##   `fixed acidity` `volatile acidity` `citric acid` `residual sugar` chlorides
##             <dbl>              <dbl>         <dbl>            <dbl>     <dbl>
## 1             7.4               0.7           0                 1.9     0.076
## 2             7.8               0.88          0                 2.6     0.098
## 3             7.8               0.76          0.04              2.3     0.092
## 4            11.2               0.28          0.56              1.9     0.075
## 5             7.4               0.7           0                 1.9     0.076
## 6             7.4               0.66          0                 1.8     0.075
## # ... with 7 more variables: free sulfur dioxide <dbl>,
## #   total sulfur dioxide <dbl>, density <dbl>, pH <dbl>, sulphates <dbl>,
## #   alcohol <dbl>, quality <dbl>
```

Previously, we explored data about wine.

1. Make a barplot with the wine data to explore our original question about the relationship between alcohol content and quality. Why might a simple bar plot be misleading here? (Read a bit about geom_bar first, if needed)

```
g1 <- ggplot(winequality_red, aes(x=alcohol, y=quality)) + geom_bar(stat="identity", color = "purple", i
g1
```



Misleading Graph:

Bar plot, *geom_bar()* cannot be used in this case as it finds the number of cases (i.e sums up all quality values for each alcohol content). We want the the height of the graphs to represent each value of quality and hence we should use a column plot in this case using *geom_col()*

---

2. A lot of the hardwork in learning R skills is learning how to read helpfiles and use stackoverflow on your own. I showed you a glimpse of information about themes in class. Use the code below to make additional changes. This might take a bit of digging and reading about these graphical components online. Specifically, can you figure out how to (try to complete 2 out of 4):

- get rid of panel border completely and keep the grid lines?
- put the legend on the top or bottom?
- capitalize the legend name appropriately?
- add units for alcohol content and wine quality (if applicable)?

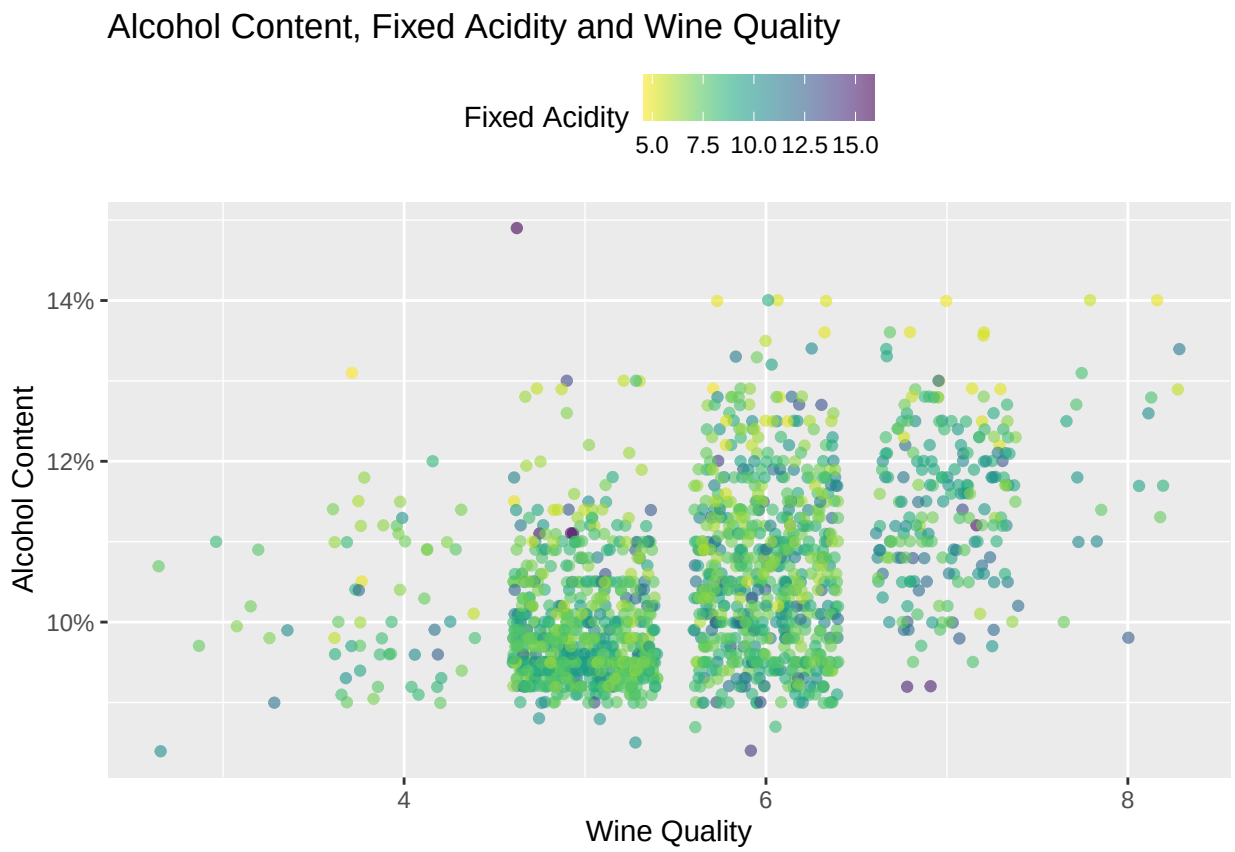
```
library(viridis)
```

```
## Loading required package: viridisLite
```

```
library(RColorBrewer)

ggplot(winequality_red, aes(x=quality, y= alcohol)) +
```

```
    geom_jitter(aes(col=`fixed acidity`)) +
    scale_color_viridis(option = "D", direction = -1, alpha = .6) +
    labs(title="Alcohol Content, Fixed Acidity and Wine Quality", y="Alcohol Content", x="Wine Quality
```



---

2. Explain in 1-3 sentences what we can learn from our wine graphic produced in class.

Answer:

- The data falls in 6 clusters, most prominent clusters are between 5 and 6.
- In the graph we see where all the three parameters converge at these dense areas (between 5 & 6 in wine quality , between 10% & 12% in alcohol content, between 6 & 8 for acidity).
- Lower acidity and higher alcohol content
- Heavy acidity and lower alcohol content (except for one outlier).
- People like to buy lower alcohol content but of more quantity (between 9% and 11% alcohol content)