Project Group 5: Sport Popularity vs. Trump Support

Cal Hale, Harrison Thompson, Atticus Patrick

11/20/2019

Introduction

This data contains the percentage of Trump voters in the 2016 presidential election and the approximate percentage of major sport searches on Google for the NFL, MLB, NHL, NBA, NASCAR, college basketball and college football. This data is taken from the 207 US Census Designated Market Areas (DMAs) of the United States. The data is sourced from Google and the official voting percentages in the 2016 election. There could be sampling bias in that the DMAs are not equal in population, which could lead certain areas having a disproportinate effect on the results. This is an example of an observational study. There doesn't appear to be any bias in the questions or measurements because the data is solely from google searches and voting polls. This is interesting to us and should be to the class because there may be certain types of sports that draw people with certain political views, and it's interesting to see which sports are related to being supportive of Trump. To clean this data we needed to first clean the "%" signs out of the excel file and turn each number into decimal form. R did this for us when we uploaded it as an excel file. Also we needed to remove the top row in the excel file in order for the data to be "tidy." We also had to filter each location into two new columns: state and region (code shown below).

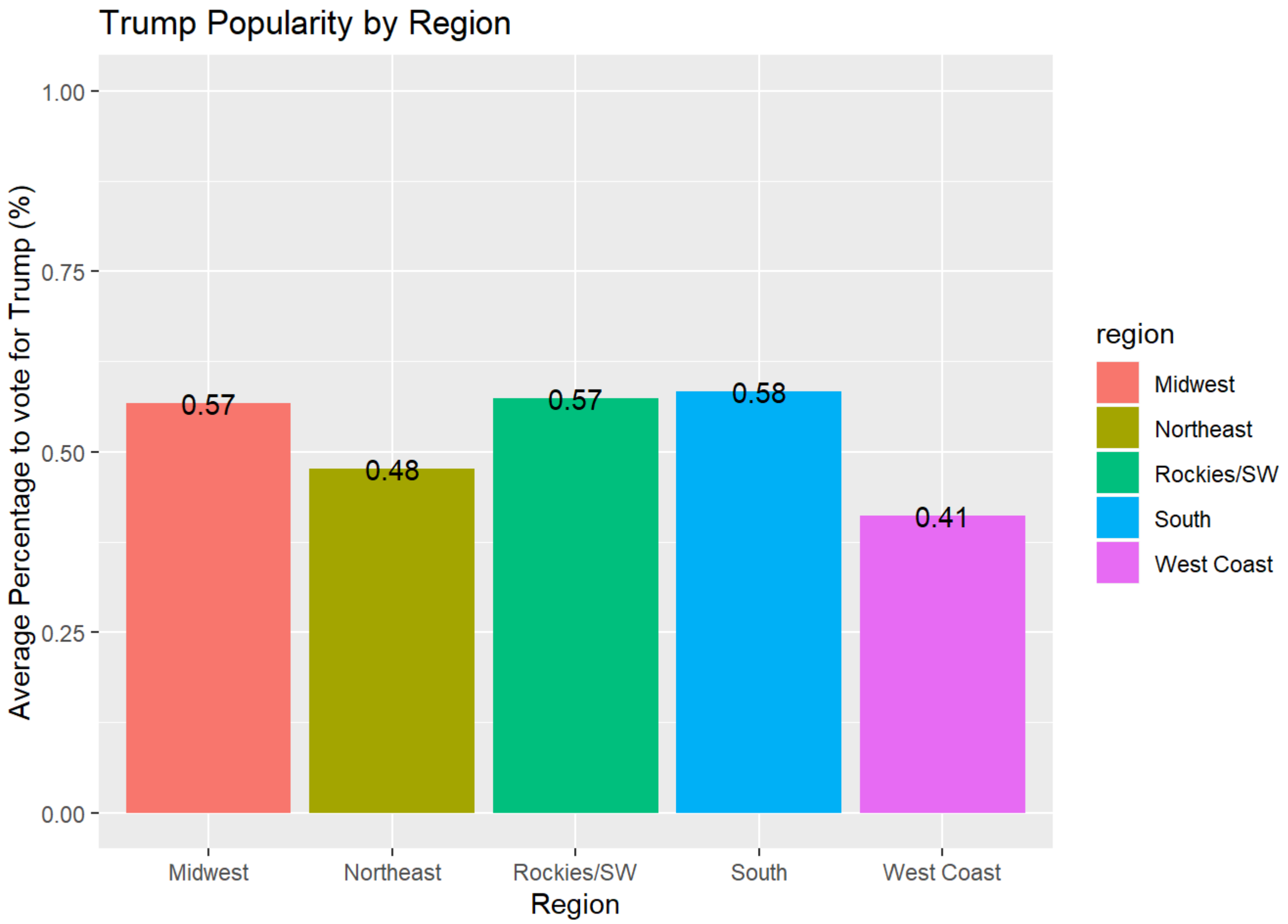
Setting up States and Regions - Data Cleaning

```
library(tidyverse)
library(readxl)
library(dplyr)
library(scales)
n <- read_excel("NFL Excel.xlsx")
n$state <-  word(n$DMA, -1)
View(n)
names(n)[9] <- "TrumpP"
n$region <- ifelse(n$state =="MA"|n$state =="NY"|n$state == "VT"|n$state =="NH"|
                     n$state =="ME"|n$state =="CT"|n$state =="RI"|n$state =="PA"|
                     n$state =="NJ"|n$state =="DE"|n$state =="MD"|n$state =="DC"|
                     n$state=="MD)", "Northeast", NA)
n$region <- ifelse(n$state == "FL"|n$state == "GA"|n$state =="AL"|n$state =="MS"|
                     n$state =="LA"|n$state =="TN"|n$state =="KY"|n$state =="AR"|
                     n$state =="NC"|n$state =="SC"|n$state =="VA"|n$state =="WV"|
                     n$state =="TN-VA","South", n$region)
n$region <- ifelse(n$state == "OH"|n$state =="IN"|n$state =="IL"|n$state =="MI"|
                     n$state =="WI"|n$state =="MN"|n$state =="ND"|n$state =="SD"|
                     n$state =="IA"|n$state =="NE"|n$state =="MO"|n$state =="KS",
                   "Midwest", n$region)
n$region <- ifelse(n$state =="TX"|n$state =="OK"|n$state =="NM"|n$state =="AZ"|
                     n$state =="UT"|n$state =="CO"|n$state =="WY"|n$state =="MT"|
                     n$state =="ID", "Rockies/SW", n$region)
n$region <- ifelse(n$state =="CA"|n$state =="NV"|n$state =="OR"|n$state =="WA"|
                     n$state =="AK"|n$state =="HI", "West Coast", n$region)
```

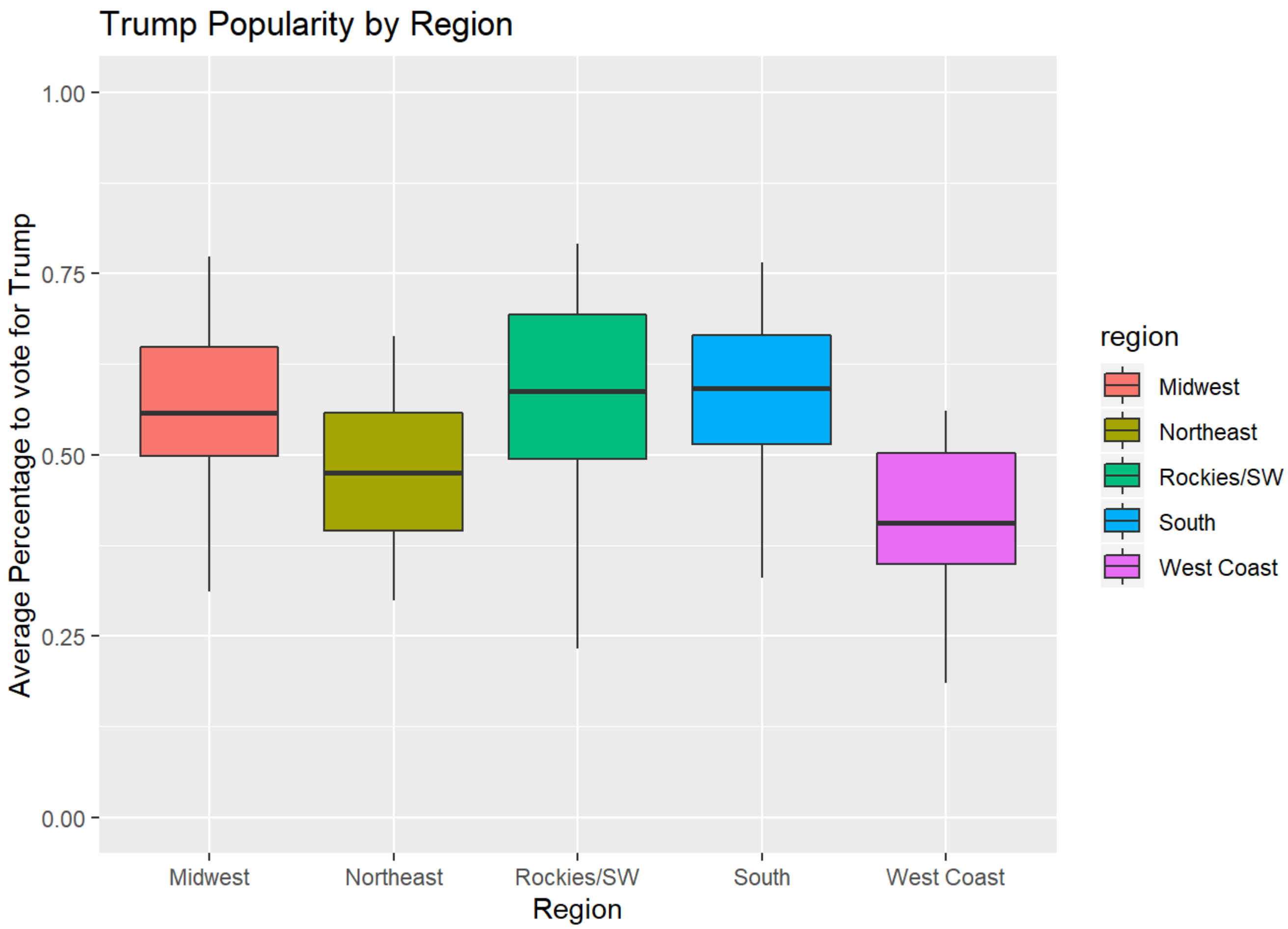
Data Analysis

Trump by Region

```
TP <- n %>% group_by(region) %>% summarise(TrumpAvg = mean(TrumpP))
View(TP)

ggplot(data = TP, mapping = aes(x = region, y = TrumpAvg, fill = region)) +
  geom_bar(stat='identity') + ylim(c(0, 1)) + geom_text(aes(label=round(TrumpAvg,digits = 2))) +
  labs(title = 'Trump Popularity by Region',
       x = 'Region',
       y = 'Average Percentage to vote for Trump (%)')
```



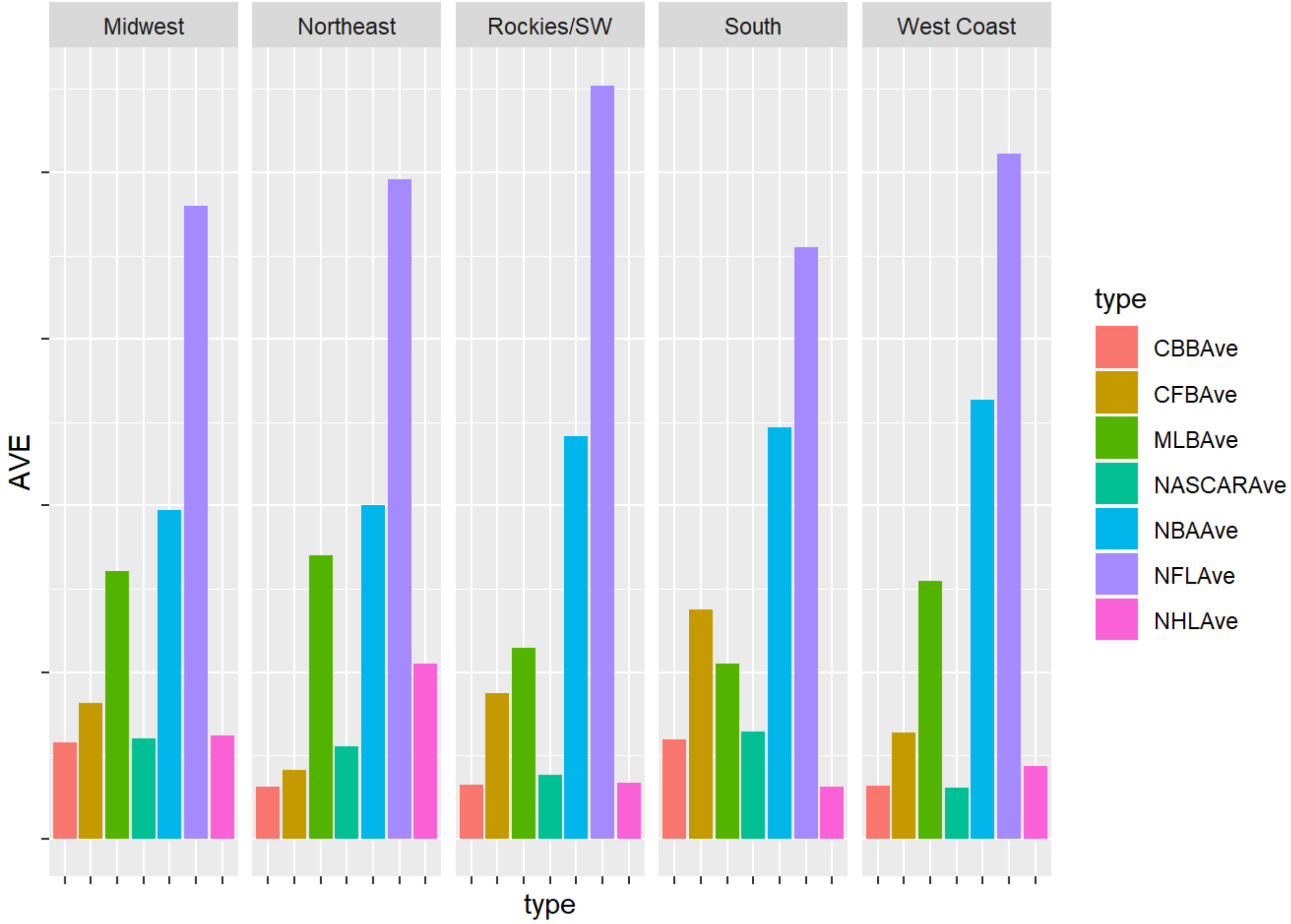
```
ggplot(data = n, mapping = aes(x = region,y = TrumpP, fill = region)) +
  geom_boxplot() + ylim(c(0, 1)) +
  labs(title = 'Trump Popularity by Region',
        x = 'Region',
        y = 'Average Percentage to vote for Trump')
```



Based on these two graphs we can see that South, Rockies/SW, and the Midwest have stronger support for Trump. This will be interesting to keep in mind when considering the sports that are inherently popular in each area (such as NASCAR in the south and NHL in the north). Trump has a lower popularity in the Northeast and the West Coast.

Sport by Region

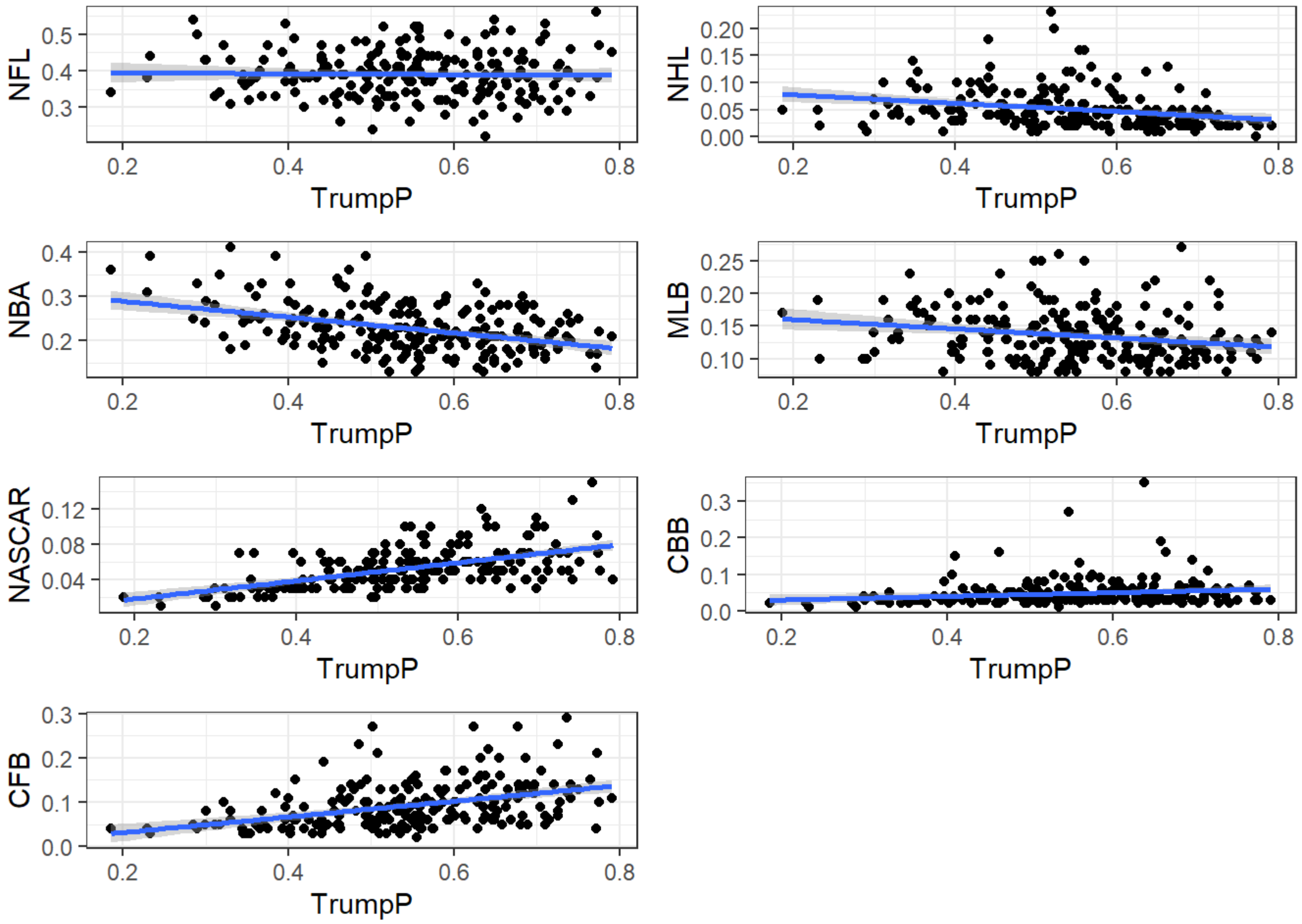
```
RS <- n  %>% group_by(region) %>% summarise(NFLAve = mean(NFL),
                                           NBAAve = mean(NBA),
                                           MLBAve = mean(MLB),
                                           NHLAve = mean(NHL),
                                           NASCARAve = mean(NASCAR),
                                           CFBAve = mean(CFB),
                                           CBBAve = mean(CBB))

RS_tidy <- RS %>%
  gather(key = type, value = AVE, -region)
View(RS_tidy)

ggplot(data = RS_tidy, mapping = aes(x = type, y = AVE, fill = type))+
  geom_bar(stat = 'identity') +
  theme(axis.text = element_blank())+
  facet_grid(. ~ region)
```



This graph is important because it not only shows which sport is the most popular, but which sport is the most popular in each area of the country. It appears that the NFL is by far the most popular sport in the U.S. followed by the NBA and MLB. This is important when considering this data in comparison to Trump's popularity because the more popular sports will carry more weight when considering their relationship to Trump's popularity.

Sport by Trump

```r
library(gridExtra)

p1 <- ggplot(data = n,
       mapping = aes(x = TrumpP, y = NFL))+
  geom_point()+
  geom_smooth(method = 'lm')+
  theme_bw()

p2 <- ggplot(data = n,
       mapping = aes(x = TrumpP, y = NHL))+
  geom_point()+
  geom_smooth(method = 'lm')+
  theme_bw()

p3 <- ggplot(data = n,
       mapping = aes(x = TrumpP, y = NBA))+
  geom_point()+
  geom_smooth(method = 'lm')+
  theme_bw()

p4 <- ggplot(data = n,
       mapping = aes(x = TrumpP, y = MLB))+
  geom_point()+
  geom_smooth(method = 'lm')+
  theme_bw()

p5 <- ggplot(data = n,
       mapping = aes(x = TrumpP, y = NASCAR))+
  geom_point()+
  geom_smooth(method = 'lm')+
  theme_bw()


p6 <- ggplot(data = n,
       mapping = aes(x = TrumpP, y = CBB))+
  geom_point()+
  geom_smooth(method = 'lm')+
  theme_bw()

p7 <- ggplot(data = n,
       mapping = aes(x = TrumpP, y = CFB))+
  geom_point()+
  geom_smooth(method = 'lm')+
  theme_bw()

grid.arrange(p1, p2, p3, p4, p5, p6, p7, nrow = 4)
```



##

Correlations ### NFL by Trump

```
cor.test(n$TrumpP, n$NFL)
```

```
##
##  Pearson's product-moment correlation
##
## data:  n$TrumpP and n$NFL
## t = -0.32433, df = 205, p-value = 0.746
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1585266  0.1140762
## sample estimates:
##         cor
## -0.02264612
```

NHL by Trump

```
cor.test(n$TrumpP, n$NHL)
```

```
## 
##  Pearson's product-moment correlation
## 
## data:  n$TrumpP and n$NHL
## t = -3.8925, df = 205, p-value = 0.0001341
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3849429 -0.1306487
## sample estimates:
##        cor 
## -0.2623446
```

NBA by Trump

```
cor.test(n$TrumpP, n$NBA)
```

```
## 
##  Pearson's product-moment correlation
## 
## data:  n$TrumpP and n$NBA
## t = -6.1755, df = 205, p-value = 3.485e-09
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5051347 -0.2745020
## sample estimates:
##        cor 
## -0.3960465
```

MLB by Trump

```
cor.test(n$TrumpP, n$MLB)
```

```
## 
##  Pearson's product-moment correlation
## 
## data:  n$TrumpP and n$MLB
## t = -3.1609, df = 205, p-value = 0.001811
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.34189265 -0.08160242
## sample estimates:
##        cor 
## -0.2155735
```

NASCAR by Trump

```
cor.test(n$TrumpP, n$NASCAR)
```

```
##
##  Pearson's product-moment correlation
##
## data:  n$TrumpP and n$NASCAR
## t = 9.4508, df = 205, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4481857 0.6392340
## sample estimates:
##      cor
## 0.550886
```

CBB by Trump

```
cor.test(n$TrumpP, n$CBB)
```

```
##
##  Pearson's product-moment correlation
##
## data:  n$TrumpP and n$CBB
## t = 2.3541, df = 205, p-value = 0.01951
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.02645413 0.29214531
## sample estimates:
##       cor
## 0.1622389
```

CFB by Trump

```
cor.test(n$TrumpP, n$CFB)
```

```
##
##  Pearson's product-moment correlation
##
## data:  n$TrumpP and n$CFB
## t = 6.5928, df = 205, p-value = 3.588e-10
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2989305 0.5246937
## sample estimates:
##       cor
## 0.4182505
```

After looking at these 7 graphs and comparing each sport’s popularity to trump voting percentages, it is clear there are a few strong relationships in our data. The graph NBA by TrumpP shows a strong, negative, linear relationship, meaning Trump is less popular where the NBA is more popular. Additionally it has a strong, negative correlation coefficient of -0.396. This makes sense because the NBA draws from more urban, liberal areas of the US. On the other hand, the graph NASCAR by Trump shows a strong, positive, linear relationship, indicating that Trump is more popular where NASCAR is more popular. It also has a strong, positive correlation coefficient of 0.551. This is expected because NASCAR is generally in the south where Trump’s support is higher. NFL by Trump shows no real relationship between Trump support and NFL popularity and has a very small, negative correlation coefficient of -0.023, which makes sense because it is probably the most popular American sport, drawing in people from a very diverse range of political views.

Preliminary Conclusions

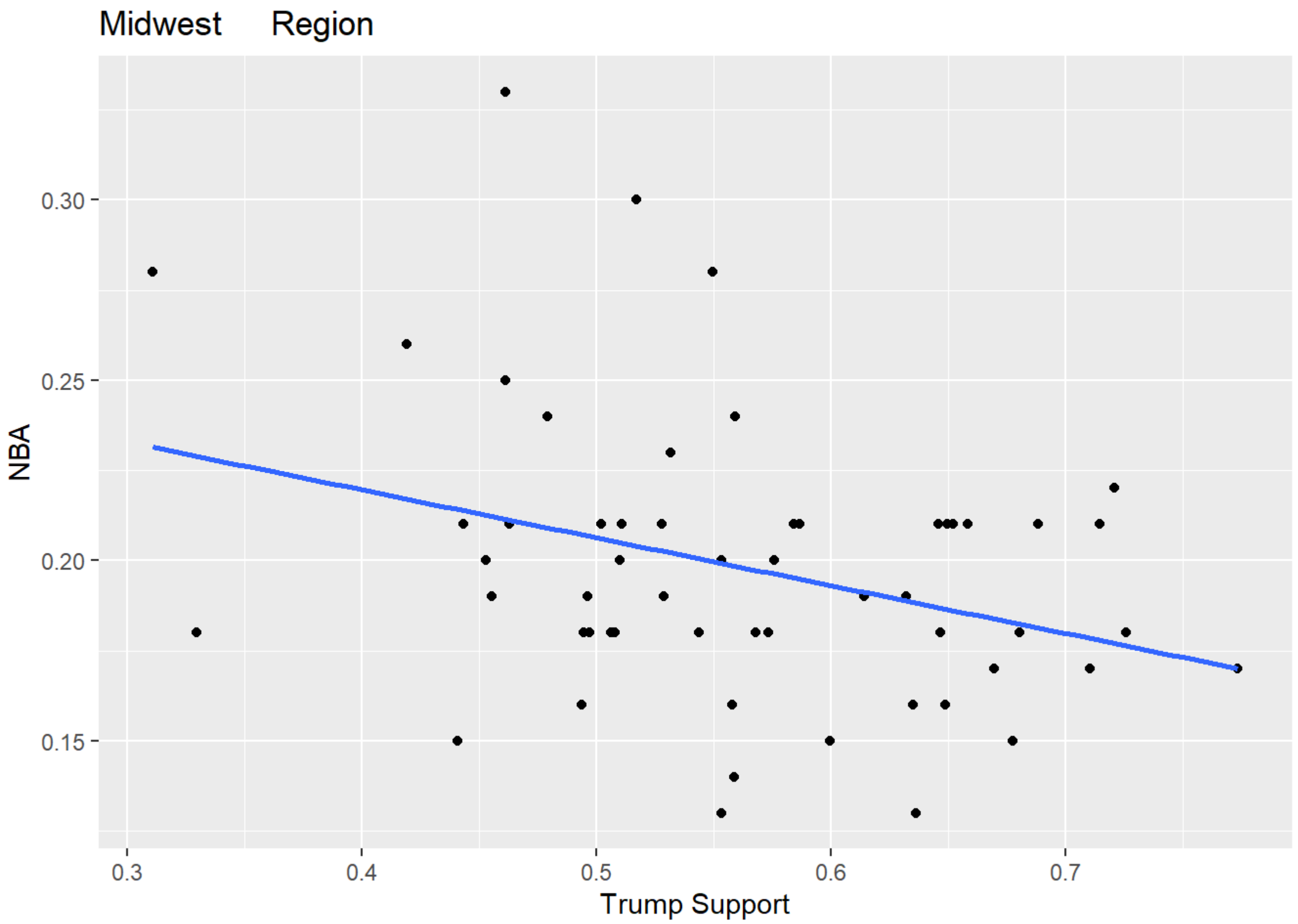
After considering each relationship between Trump and sport popularity in the different regions of the country, a few patterns and associations are clear. First, it is apparent that people who vote for Trump are more likely to be fans of NASCAR and college football and less likely to be fans of the NBA, NHL and MLB. It’s clear that it is important to consider where each sport is popular as well, so because the South has a higher general popularity for Trump, it makes sense that NASCAR popularity has the strongest association with being a Trump supporter (because NASCAR mostly takes place in the south). Overall it is very interesting to see each different association and relationship between each sport and Trump’s popularity.

Additional Application

For the final application of our data, we took a deeper dive into the geographic distributions of the two sports fandoms that had the greatest positive and negative correlation with Trump support, NASCAR and the NBA respectively. We chose to look at NBA support on the West Coast (where it was the highest) and in the Midwest (where it was the lowest) and NASCAR support in the South (where it was the highest) and in the Rockies/SW region and in the Northeast (where it was the lowest). For each region, we made a scatterplot in which each point represents a DMA. Those plots, in addition to the correlation coefficients and standard deviations are displayed below.

NBA in the Midwest

```
MW <- n %>% filter(region == 'Midwest')

ggplot(data = MW, mapping = aes(x = TrumpP, y = NBA )) + geom_point()+
  geom_smooth(method = 'lm', se = FALSE) + labs(title = "Midwest     Region", x = "Trump Suppor
t")
```



```
cor.test(MW$TrumpP, MW$NBA)
```

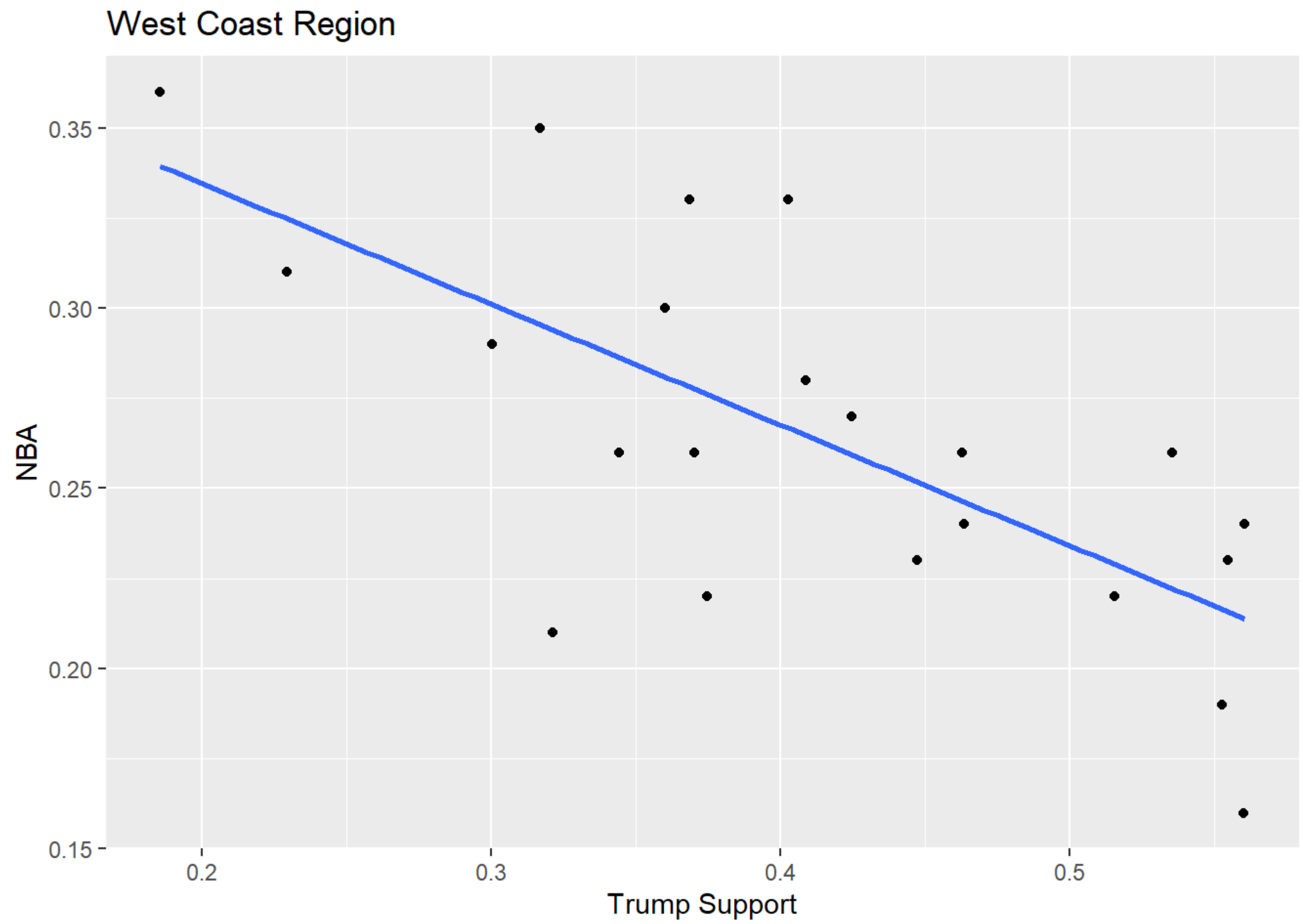
```
##
##  Pearson's product-moment correlation
##
## data:  MW$TrumpP and MW$NBA
## t = -2.7036, df = 56, p-value = 0.009066
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.54982741 -0.08933216
## sample estimates:
##        cor
## -0.3397876
```

```
sd(MW$TrumpP, MW$NBA)
```

```
## [1] 0.09995257
```

NBA on the West Coast

```
WC <- n %>% filter(region == 'West Coast')

ggplot(data = WC, mapping = aes(x = TrumpP, y = NBA )) + geom_point()+
  geom_smooth(method = 'lm', se = FALSE) + labs(title = "West Coast Region",
                                                x = "Trump Support")
```



```
cor.test(WC$TrumpP, WC$NBA)
```

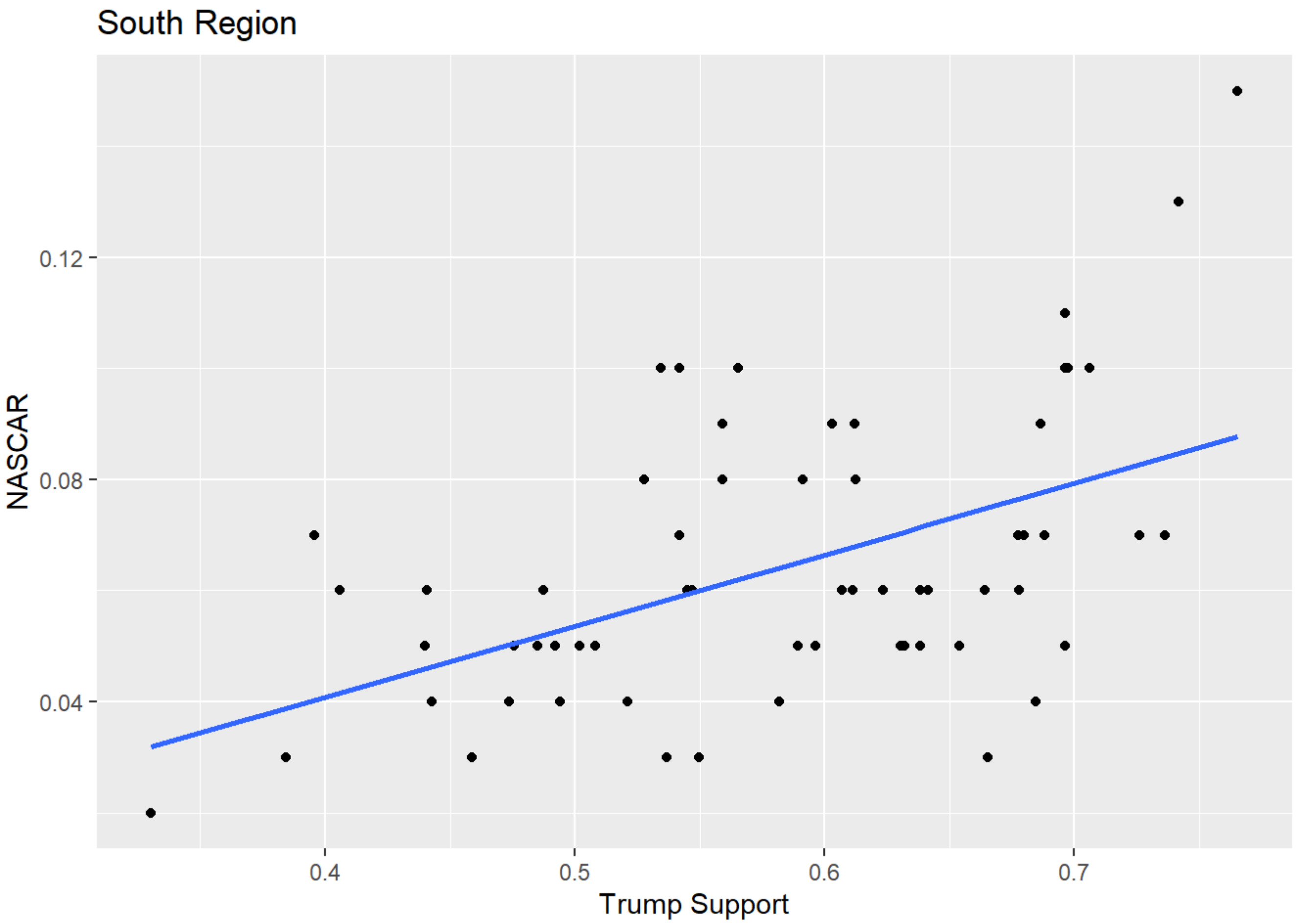
```
##
##  Pearson's product-moment correlation
##
## data:  WC$TrumpP and WC$NBA
## t = -4.337, df = 20, p-value = 0.00032
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8641456 -0.3886373
## sample estimates:
##        cor
## -0.6961754
```

```
sd(WC$TrumpP, WC$NBA)
```

```
## [1] 0.107645
```

NASCAR in the South

```
SO <- n %>% filter(region == 'South')

ggplot(data = SO, mapping = aes(x = TrumpP, y = NASCAR )) + geom_point()+
  geom_smooth(method = 'lm', se = FALSE) + labs(title = "South Region",
                                                x = "Trump Support")
```



```
cor.test(SO$TrumpP, SO$NASCAR)
```

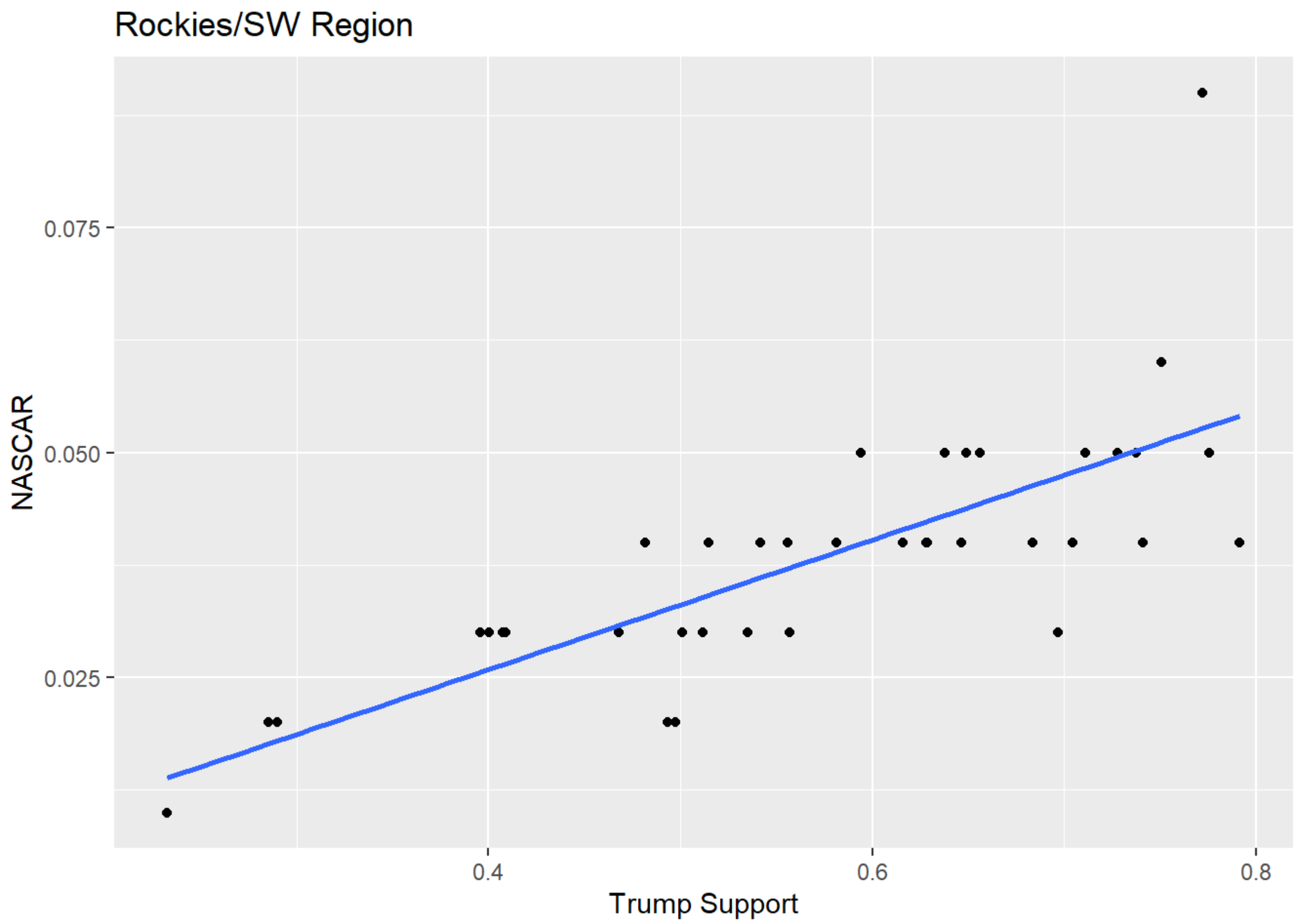
```
##
##  Pearson's product-moment correlation
##
## data:  SO$TrumpP and SO$NASCAR
## t = 4.5159, df = 61, p-value = 2.944e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2885768 0.6657516
## sample estimates:
##       cor
## 0.5005527
```

```
sd(SO$TrumpP, SO$NASCAR)
```

```
## [1] 0.09882505
```

NASCAR in the Rockies/SW

```
RSW <- n %>% filter(region == 'Rockies/SW')

ggplot(data = RSW, mapping = aes(x = TrumpP, y = NASCAR )) + geom_point()+
  geom_smooth(method = 'lm', se = FALSE) + labs(title = "Rockies/SW Region",
                                                x = "Trump Support")
```



```
cor.test(RSW$TrumpP, RSW$NASCAR)
```

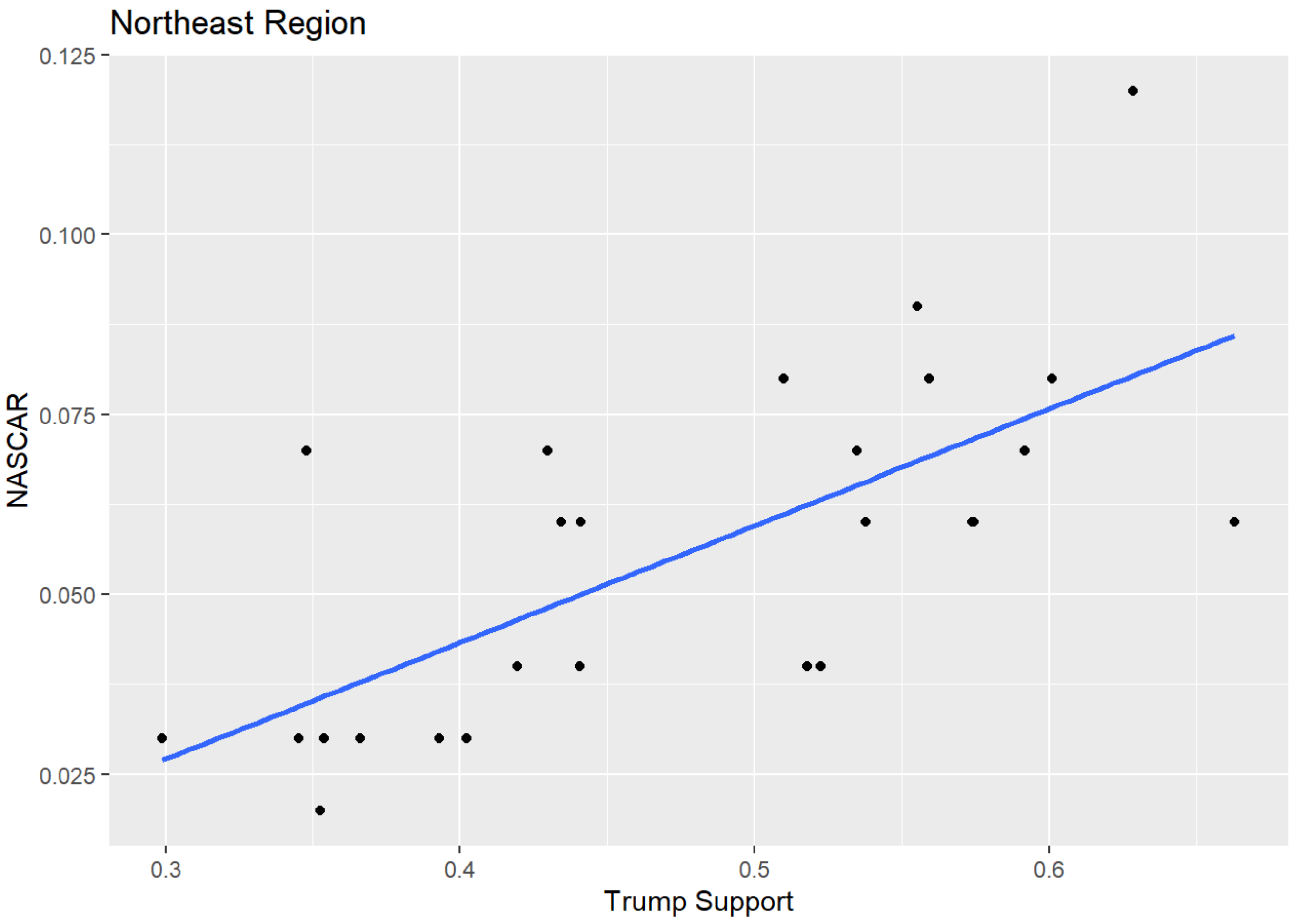
```
##
##  Pearson's product-moment correlation
##
## data:  RSW$TrumpP and RSW$NASCAR
## t = 6.8478, df = 36, p-value = 5.203e-08
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5693434 0.8640575
## sample estimates:
##       cor
## 0.7521316
```

```
sd(RSW$TrumpP, RSW$NASCAR)
```

```
## [1] 0.1446796
```

NASCAR in the Northeast

```
NE <- n %>% filter(region == 'Northeast')

ggplot(data = NE, mapping = aes(x = TrumpP, y = NASCAR )) + geom_point()+
  geom_smooth(method = 'lm', se = FALSE) + labs(title = "Northeast Region", x = "Trump Support")
```



```
cor.test(NE$TrumpP, NE$NASCAR)
```

```
##
##  Pearson's product-moment correlation
##
## data:  NE$TrumpP and NE$NASCAR
## t = 4.7379, df = 24, p-value = 8.086e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4212871 0.8528782
## sample estimates:
##       cor
## 0.695193
```

```
sd(NE$TrumpP, NE$NASCAR)
```

```
## [1] 0.1022177
```

This further analysis of the data found that while the NBA already has a significant negative correlation with Trump support nationwide (-0.396), the effect is magnified on the West Coast, where the correlation is -0.696 with a standard deviation of 0.108. We can therefore say with certainty that the correlation between NBA fandom and Trump support is magnified on the West Coast. This may be due to the concentration of West Coast NBA fans in liberal urban centers.

In the Midwest, where the NBA entertains the least popularity of any region, the correlation is slightly weaker (-0.340) than the national average, although the nationwide correlation is within the standard deviation (0.100).

For the correlations between Trump support and NASCAR fandom, we looked at the South (where NASCAR fandom is high) and the Northeast and Rockies/SW regions (where NASCAR fandom is low). Here, we found the opposite effect:

The correlation between Trump support and NASCAR fandom was much higher in the regions with fewer NASCAR fans as a whole (0.752 with a standard deviation of 0.145 in the Rockies/SW and 0.695 with a standard deviation of 0.102 in the Northeast) than the national average (0.551). Meanwhile, while in the South where NASCAR is relatively popular, the correlation of 0.501 with a standard deviation of 0.099 was weaker than the national average. We interpreted this to mean that in the south, NASCAR is a more popular sport among everyone, regardless of whether they live in a liberal or conservative area, while in the Northeast and Rockies/SW, NASCAR watching is a cultural phenomenom limited to more conservative DMAs.

Limitations/Recommendations

One major limitation inherent to this study is that the popularity of each sport is based solely on google searches. While it is fairly accurate as to showing the popularity of each sport, it does not describe the full picture of how popular different sports are in these different areas. For example, it could be more accurate to use additional data such as TV viewership of each sport or tickets bought to sporting events in these areas as well. Another inherent limitation is that it is solely showing the percentage of people voting for Trump and therefore we can’t ultimately decide which political views align with which sport – because there are really only two main candidates in a presidential election, people must disregard some of their views and pick one candidate or the other; we can’t simply assume all Trump voters are strictly right wing and all non-Trump voters are strictly left wing. Another possible limitation is that each area has a different population, which could be misleading about the amount of google searches in areas with higher populations compared to other areas with smaller populations. In future work, it may be useful to look at TV viewership of these different sports to take a different look at sport popularity in each area.