# DATA 607 Project 2

*Dilip Ganesan*

*03/05/2017*

## Data 607 Project 2

Choose any three of the wide datasets identified in the Week 6 Discussion items. You may use your own dataset please dont use my Sample Post dataset since that was used in your Week 6 assignment For each of the three chosen datasets:

## Data set 1: 2016 Election Analysis

***Since most data sets are related to states and counties, i have tried to use maps and graphics to make it colorful***

***Ref: Inspired by Dilip post = "Election Data"***

***Data Source: Kaggle***

## Abstract

***2016 general election proved all pollsters wrong and it was one of the biggest upsets in modern day political history. Having said that there were few states which was predicted very badly in general election, so figure out why we are going to see how those states performed in Primary Elections between two candidates Bernie and Donald Trump.***

***1. Wanted to see how white non-college educated voters voted for Bernie Sanders and Donald Trump.***

***2. Wanted to see how white women voters voted for Bernie Sanders and Donald Trump.***

***3. Wanted to see how overalpping counties of voters voted for Bernie Sanders and Donald Trump using MAPS.***

## Data Load

```
primaryresults=fread('primary_results.csv')
primaryresults=data.frame(primaryresults)
#head(primaryresults)

demographics=fread('county_facts.csv')
```

```
## Warning in fread("county_facts.csv"): Some columns have been read as type
## 'integer64' but package bit64 isn't loaded. Those columns will display as
## strange looking floating point data. There is no need to reload the data.
## Just require(bit64) to obtain the integer64 print method and print the data
## again.
```

```
demographics=data.frame(demographics)
#head(demographics)
```

## Data Munging.

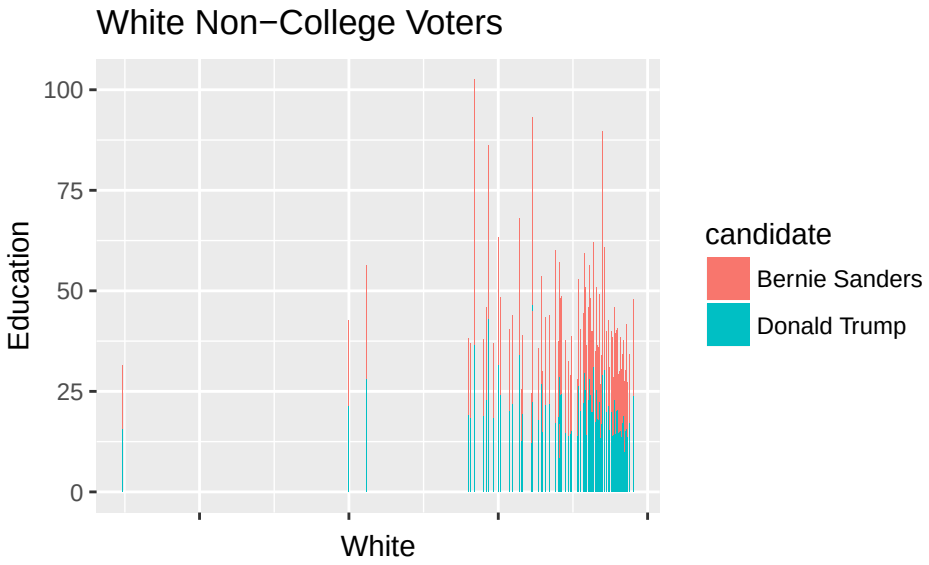
```
primaryvotes = primaryresults %>%
            filter(candidate %in% c("Bernie Sanders","Donald Trump")) %>%
            filter(state_abbreviation %in% c("MI","WI"))

demographics %<>%
    filter(state_abbreviation %in% c("MI","WI")) %>%
    select(state_abbreviation = state_abbreviation, county = area_name,
           income = INC110213, hispanic = RHI725214, female = SEX255214,
           white= RHI825214, college = EDU685213, density = POP060210) %>%
    mutate(county = gsub(" County", "", county))


votes = inner_join(primaryvotes, demographics, by = c("state_abbreviation","county"))

knitr::kable(votes[1:5,], caption="Bernie vs Donald")
```

Table 1: Bernie vs Donald

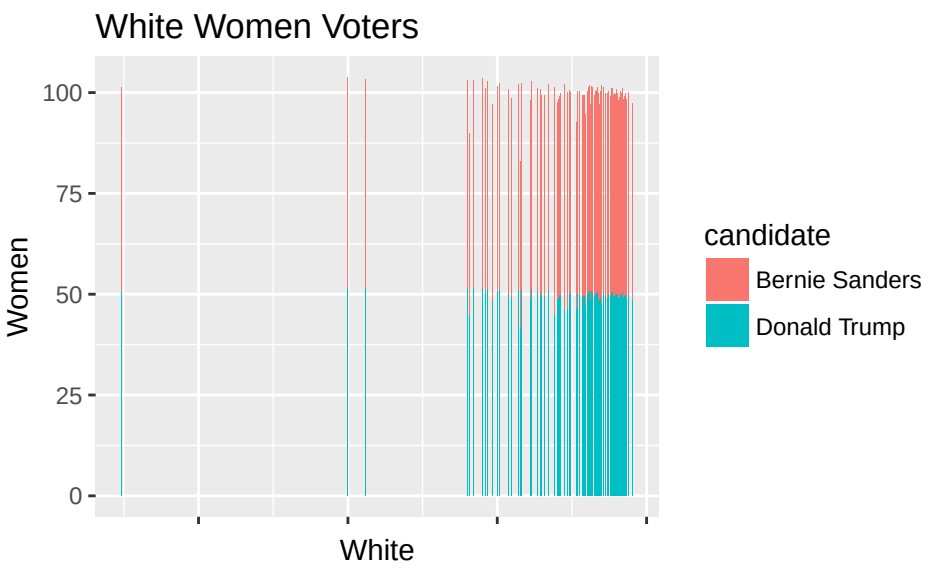
| state | state_abbreviation | county | fips | party | candidate | votes | fraction_votes | income | hispan |
|---|---|---|---|---|---|---|---|---|---|
| Michigan | MI | Alcona | 26001 | Democrat | Bernie Sanders | 455 | 0.479 | 37189 | 1 |
| Michigan | MI | Alger | 26003 | Democrat | Bernie Sanders | 622 | 0.600 | 37586 | 1 |
| Michigan | MI | Allegan | 26005 | Democrat | Bernie Sanders | 5545 | 0.603 | 52061 | 7 |
| Michigan | MI | Alpena | 26007 | Democrat | Bernie Sanders | 1347 | 0.541 | 38016 | 1 |
| Michigan | MI | Antrim | 26009 | Democrat | Bernie Sanders | 1491 | 0.621 | 45362 | 2 |

## Data Plots

```
# Among Less Educated White Folks.
ggplot(votes) + geom_bar(aes(white , college, fill=candidate), stat="summary", fun.y="mean") + labs(x="
```



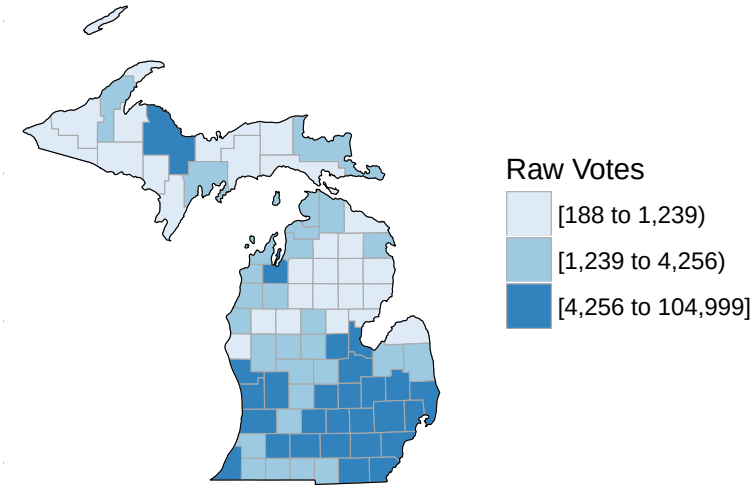

```
# Among White Women.
ggplot(votes) + geom_bar(aes(white , female, fill=candidate), stat="summary", fun.y="mean") + labs(x="W
```




### County Map with Vote Share.

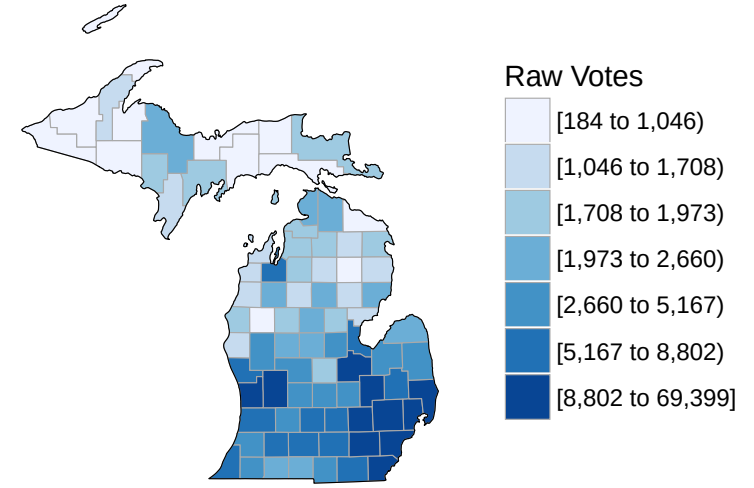
```
# Wanted to see the counties of Donald Trump and Bernie Sanders in the Map.

BerniePrimary=votes %>%
            filter(candidate %in% c("Bernie Sanders"))

BerniePrimary$region=BerniePrimary$fips
BerniePrimary$value= BerniePrimary$votes


TrumpPrimary=votes %>%
            filter(candidate %in% c("Donald Trump"))

TrumpPrimary$region=TrumpPrimary$fips
TrumpPrimary$value= TrumpPrimary$votes

# Michigan Counties

mi_choro_bs = county_choropleth(BerniePrimary, state_zoom="michigan", legend = "Raw Votes", num_colors=3
  ggtitle("Bernie Sanders in Michigan Dem Primary") +
  coord_map()
mi_choro_bs
```

Bernie Sanders in Michigan Dem Primary



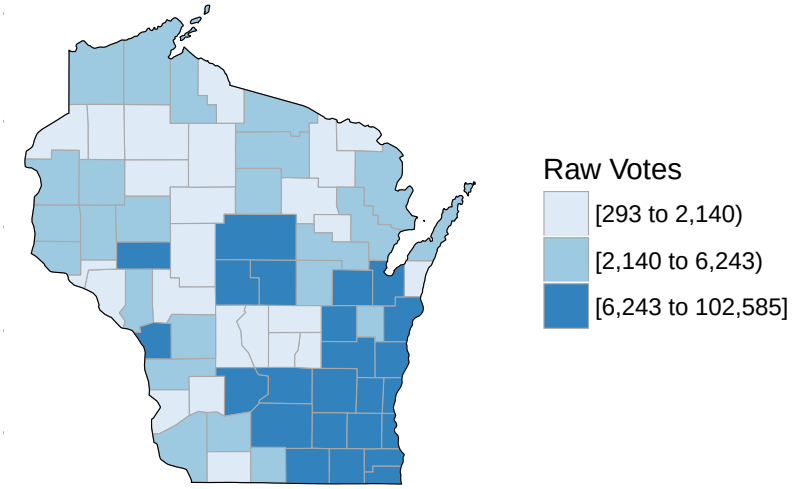

```
d_mi_choro_bs = county_choropleth(TrumpPrimary, state_zoom="michigan", legend = "Raw Votes", num_colors=
  ggtitle("Donald Trump in Michigan Rep Primary") +
  coord_map()
d_mi_choro_bs
```

Donald Trump in Michigan Rep Primary



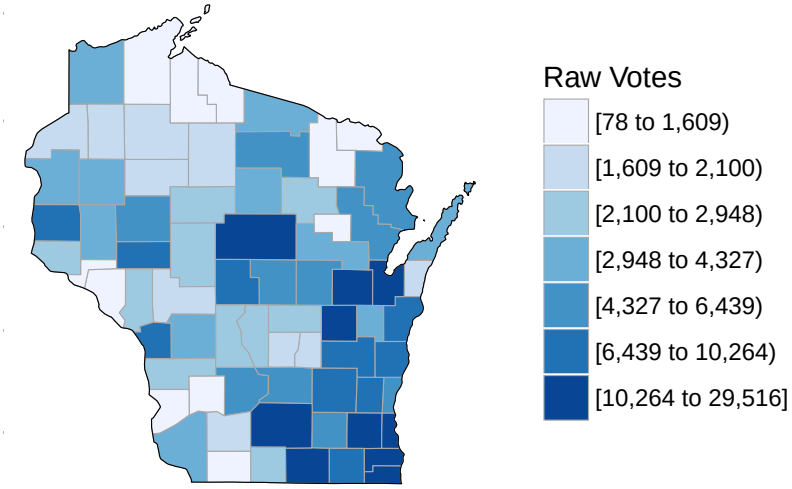

```
## Wisconsin Counties

wi_choro_bs = county_choropleth(BerniePrimary, state_zoom="wisconsin", legend = "Raw Votes", num_colors=
  ggtitle("Bernie Sanders in Wisconsin Dem Primary") +
  coord_map()
wi_choro_bs
```

Bernie Sanders in Wisconsin Dem Primary




```
d_wi_choro_bs = county_choropleth(TrumpPrimary, state_zoom="wisconsin", legend = "Raw Votes", num_colors
  ggtitle("Donald Trump in Wisconsin Rep Primary") +
  coord_map()
d_wi_choro_bs
```

Donald Trump in Wisconsin Rep Primary




## Analysis for Data 1.

***Of all the states why i selected the states of Michigan and Wisconsin, because a) they were so called blue wall states b) Hillary lost the primaries in these two states.***

***From the analysis it is clear that white working class non-college educated voters and white women were favoring Bernie.***

***The counties which Trump has maximum vote share in republican primaries are also the counties Bernie Sanders has performed very well in democratic primaries.***

***Having seen a disconnect with voters, still finding it surprising why team Hillary did not send her to campaign in these two states which she eventually lost in general election.***

## Data set 2: IRS Data

***Ref: Inspired by Pavan Akula post = "IRS Data"***

***Data Source: IRS***

## Abstract

IRS data set contains the data of files returned by Individuals from all through the USA .

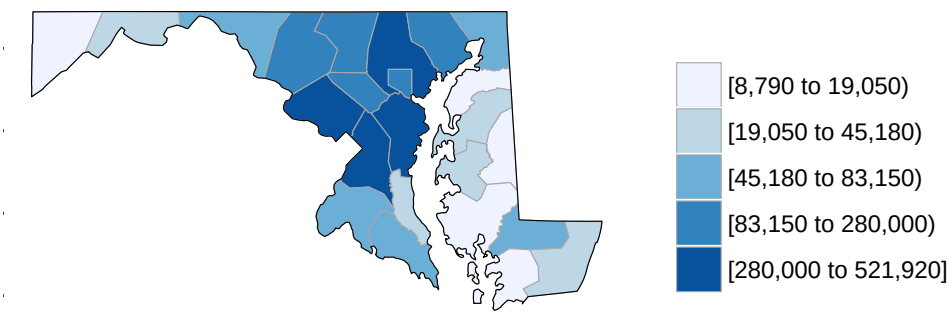
Per Pavan's post would like to do analysis on how in my home state of Maryland the tax filing has happened and also in the states in and around my home state.

## Data Load

```
irsdata=fread('irsdataset.csv')
irsdata=data.frame(irsdata)
```

## Data Munging

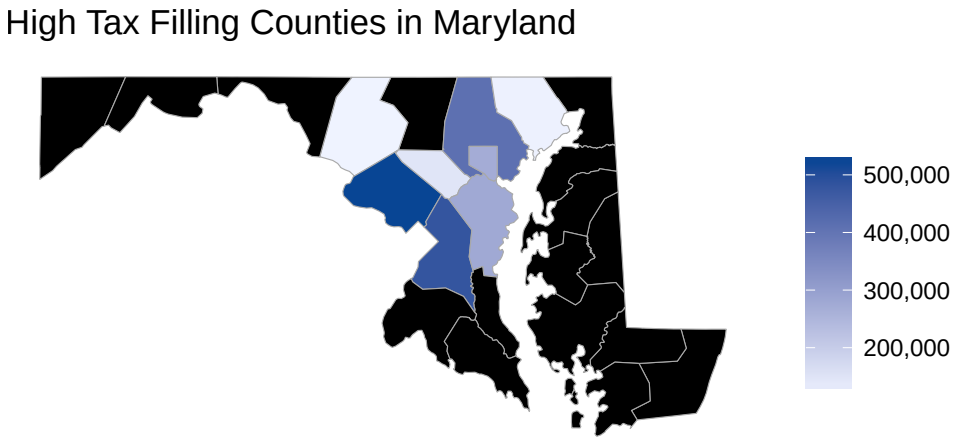
```
tristatedata = irsdata %>%
            filter(STATE %in% c("MD")) %>%
            filter(COUNTYFIPS != 0) %>%
            select(state= STATE, filling = N1,cnty= COUNTYNAME ,fips=COUNTYFIPS,statefips=STATEFIPS)

tristatedata$fips=str_pad(tristatedata$fips,3,pad = "0")
tristatedata$newfips=paste(tristatedata$statefips,tristatedata$fips,sep = "")


tristatedata$region=as.numeric(tristatedata$newfips)
tristatedata$value= tristatedata$filling

mntg = county_choropleth(tristatedata,title = "Tax Filling In State of Maryland Counties", state_zoom="
mntg
```



```
# More tax paying counties in Maryland.
hightaxcountie = tristatedata %>%
                 filter(tristatedata$filling > 120000 )

mntg2 = county_choropleth(hightaxcountie,title = "High Tax Filling Counties in Maryland", county_zoom =
```

```
## Warning in self$bind(): The following regions were missing and are being
## set to NA: 24011, 24013, 24023, 24043, 24047, 24015, 24045, 24001, 24009,
## 24017, 24019, 24029, 24035, 24037, 24039, 24041
```

```
mntg2
```




### Analysis for Data 2

***For this analysis I took my home state of Maryland how taxes are being paid across counties.***

***Did some plotting of counties and tax payer info and figured that the county i live has more tax payers :)***

---

### Data set 3: NYC Data

***Ref: Inspired by Dubar post = "NYC Data Set"***

***Data Source:New York Open Data***

### Abstract

*** Decided to analysis the population growth in the city of newyork and map it to the counties***

### Data Load

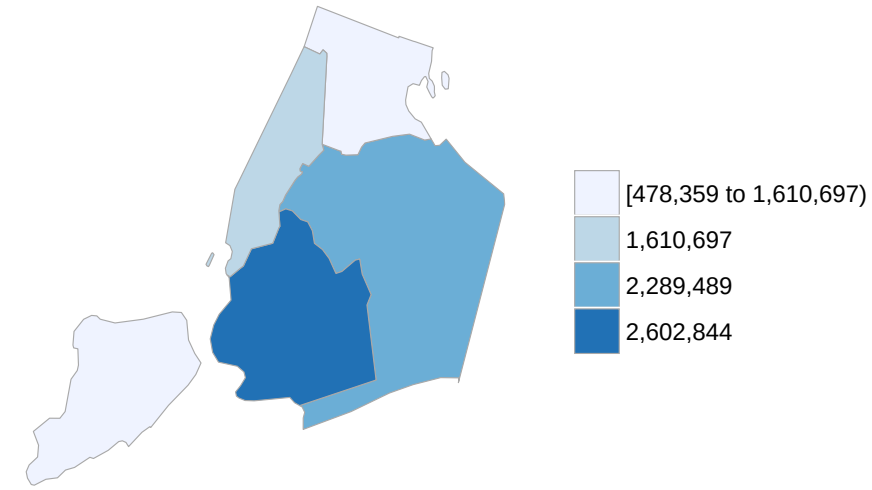
```
nypopulation=fread('https://data.cityofnewyork.us/api/views/97pn-acdf/rows.csv?accessType=DOWNLOAD',hea
nypopulation=data.frame(nypopulation)
```

### Data Munging

```
nypopulation = nypopulation %>%
                filter(Borough != 'NYC Total')

nypopulation = nypopulation %>%
                filter(Age == 'Total')
```

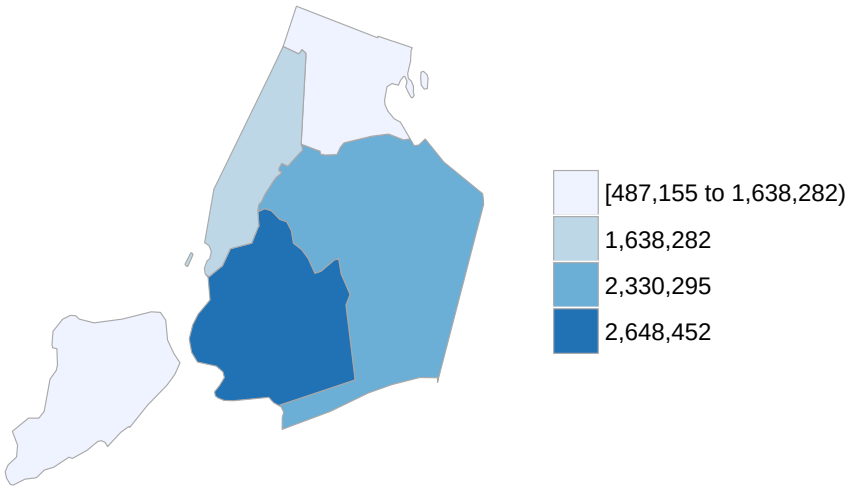
### Analysis for Data 3

```
data(county.regions)
nyc_county_names = c("kings", "bronx", "new york", "queens", "richmond")

nyc_county_fips = county.regions %>%
  filter(state.name == "new york" & county.name %in% nyc_county_names)

nypopulation$Age = c("36005","36047","36061", "36081", "36085")


# Population in 2015
nypopulation$region=as.numeric(nypopulation$Age)
nypopulation$value=nypopulation$X2015

county_choropleth(nypopulation, title = "Population of Counties in New York City in 2015", num_colors
```



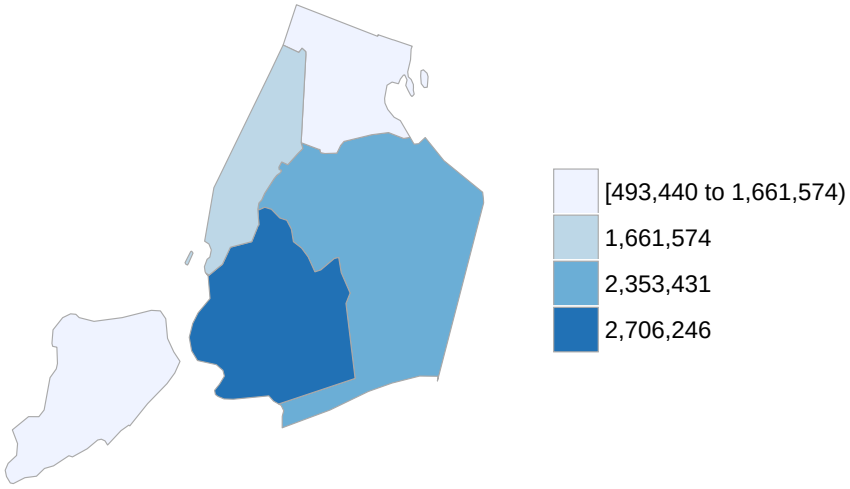

```
# Population in 2020
nypopulation$region=as.numeric(nypopulation$Age)
nypopulation$value=nypopulation$X2020

county_choropleth(nypopulation, title = "Population of Counties in New York City in 2020", num_colors
```

Population of Counties in New York City in 2020




```
# Population in 2025
nypopulation$region=as.numeric(nypopulation$Age)
nypopulation$value=nypopulation$X2025

county_choropleth(nypopulation, title = "Population of Counties in New York City in 2025", num_colors
```

Population of Counties in New York City in 2025




## References :

*** r-bloggers for usage of choroplethr.***

***https://rdrr.io/cran/choroplethr/***