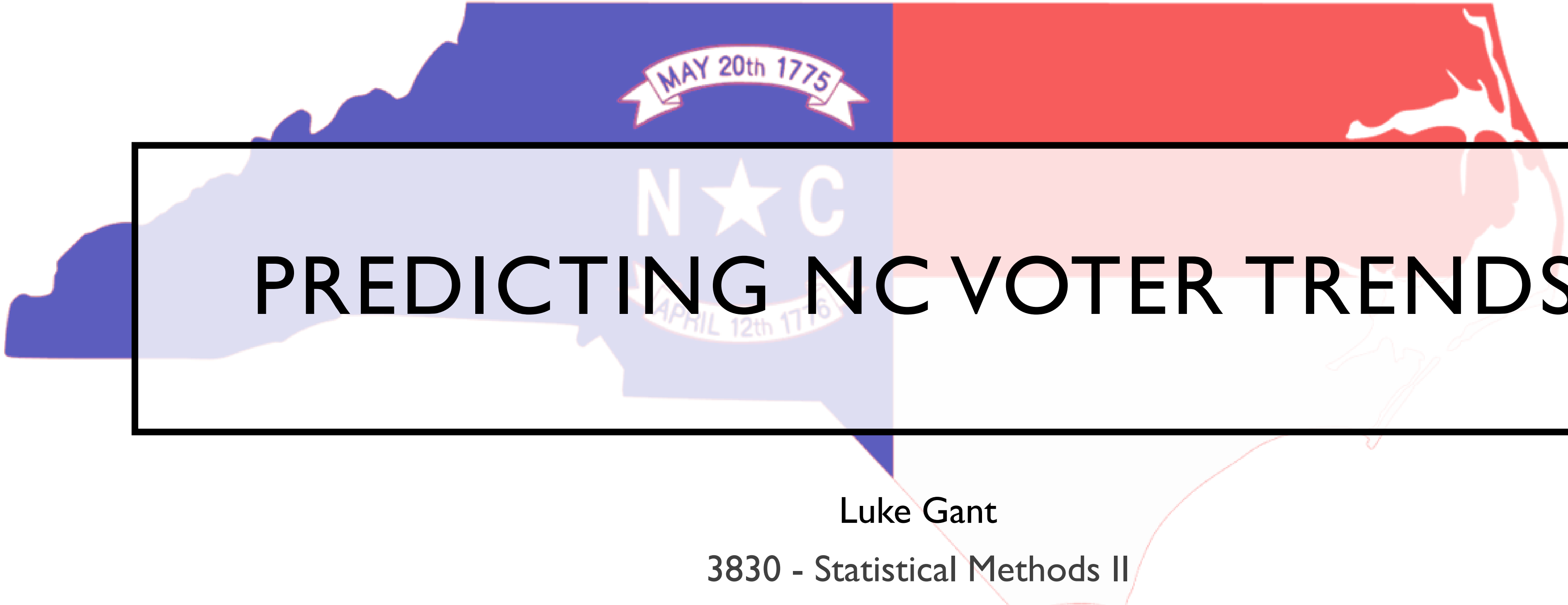


PREDICTING NC VOTER TRENDS

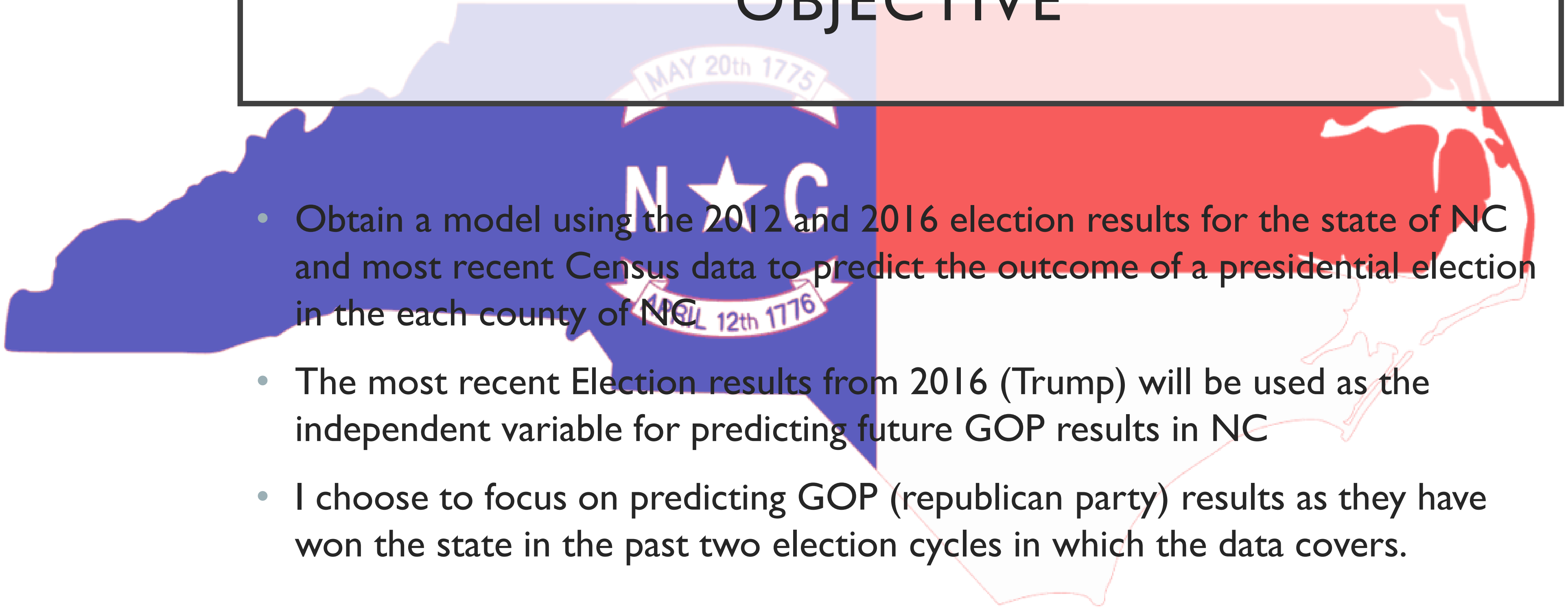
Luke Gant

3830 - Statistical Methods II

Appalachian state University

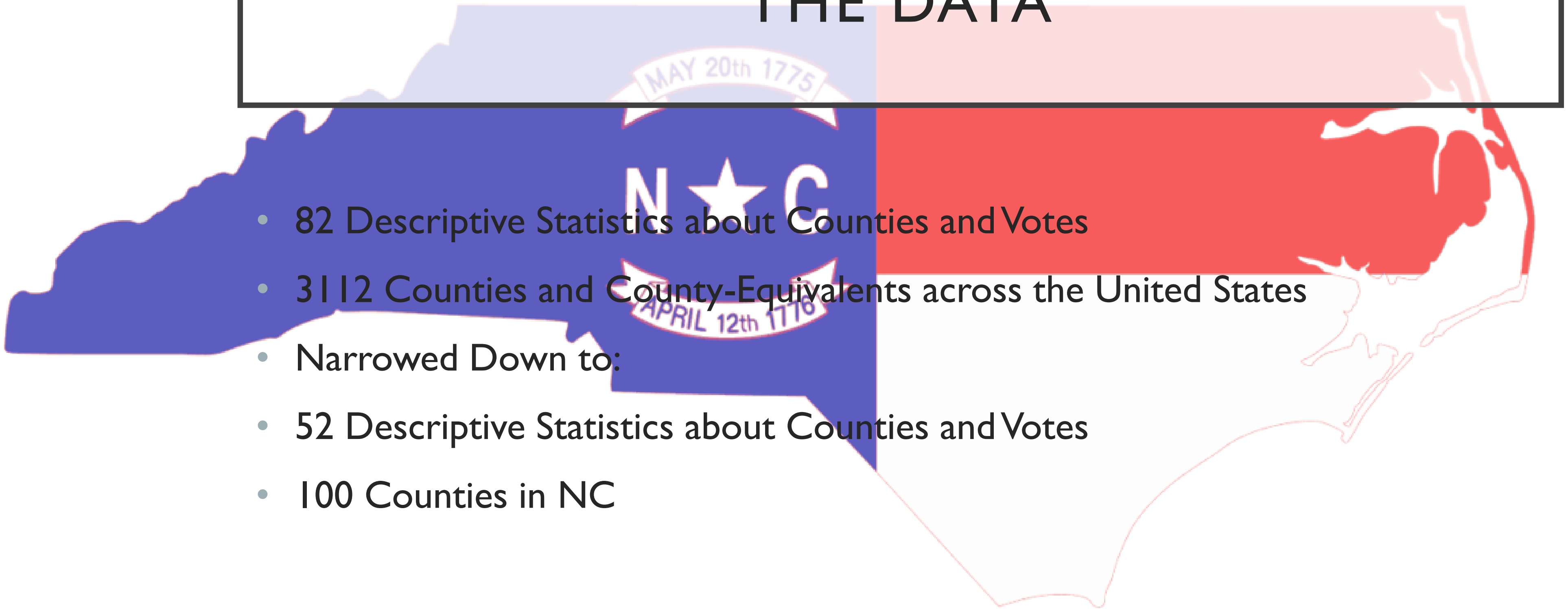
December 3, 2018

OBJECTIVE



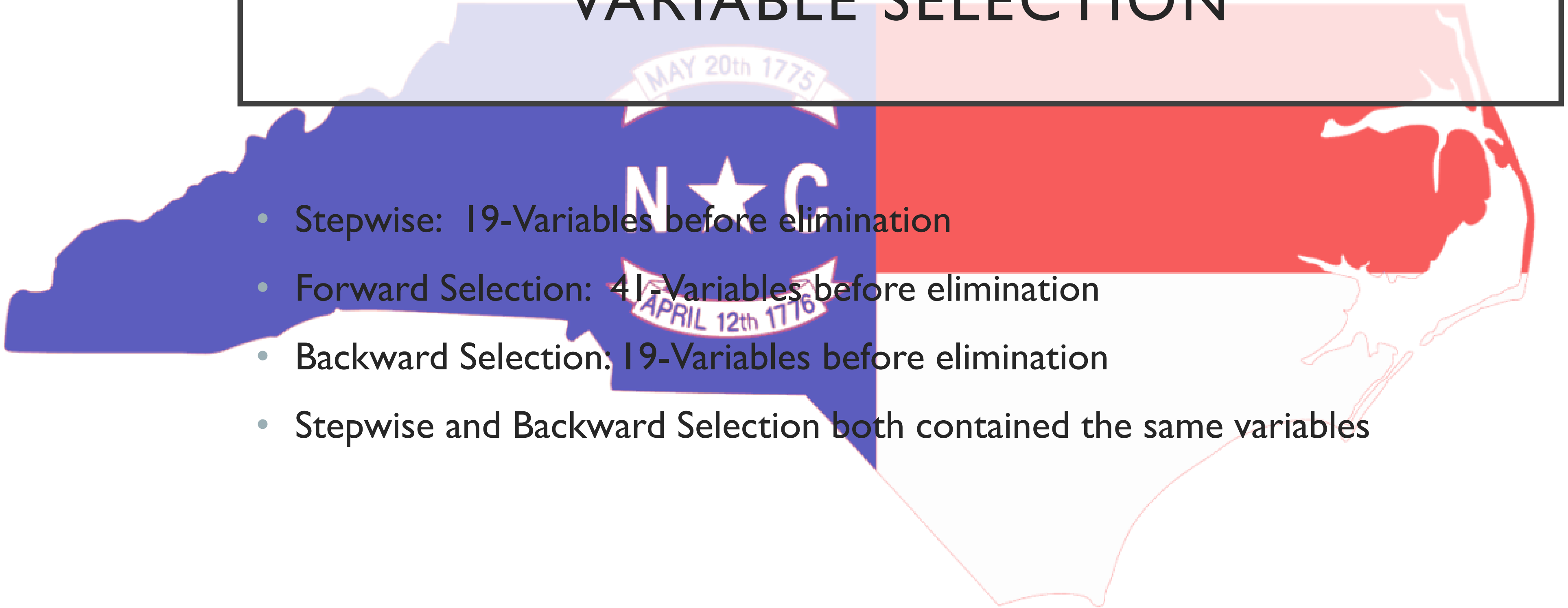
- Obtain a model using the 2012 and 2016 election results for the state of NC and most recent Census data to predict the outcome of a presidential election in the each county of NC
- The most recent Election results from 2016 (Trump) will be used as the independent variable for predicting future GOP results in NC
- I choose to focus on predicting GOP (republican party) results as they have won the state in the past two election cycles in which the data covers.

THE DATA



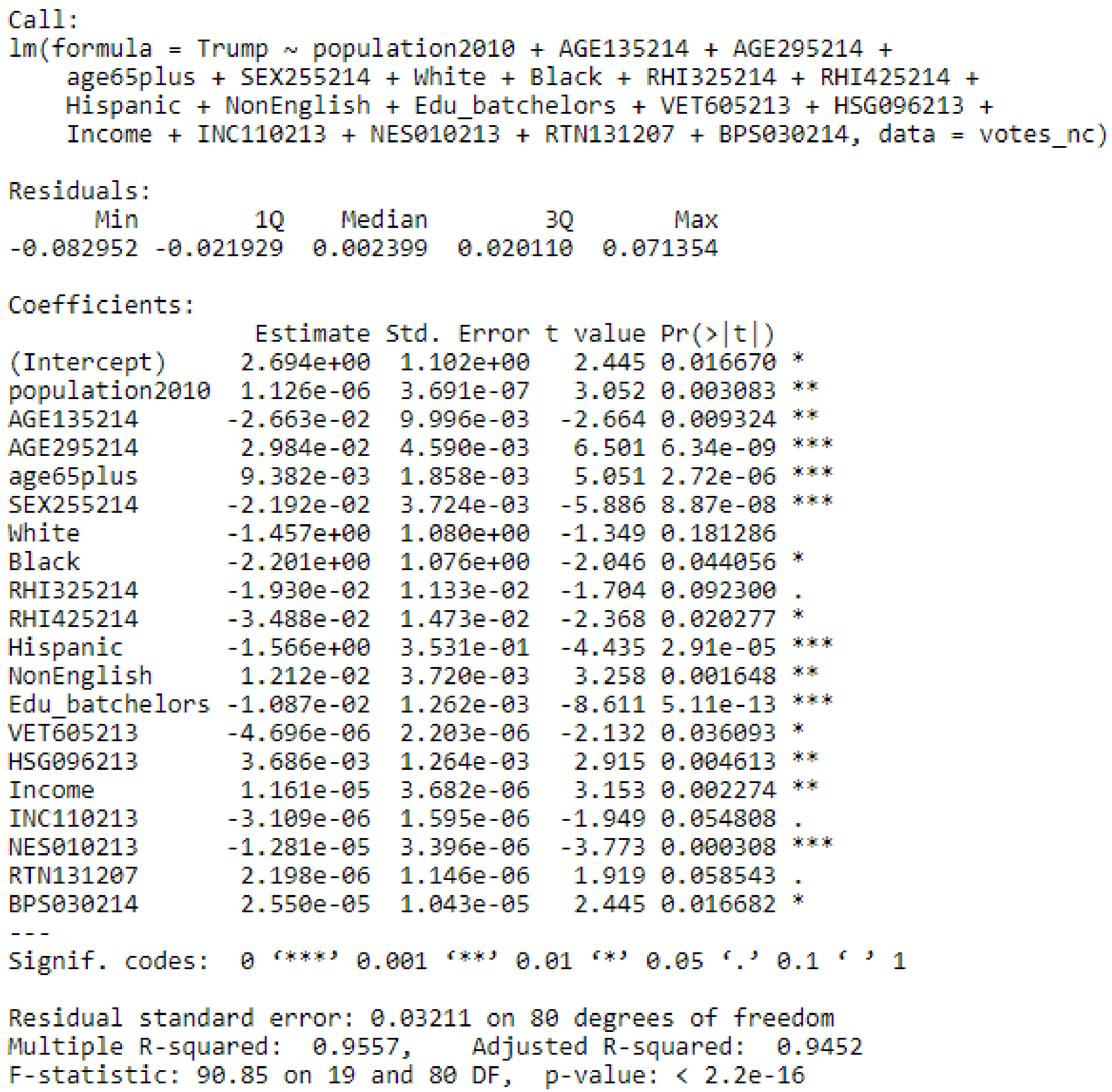
- 82 Descriptive Statistics about Counties and Votes
- 3112 Counties and County-Equivalents across the United States
- Narrowed Down to:
- 52 Descriptive Statistics about Counties and Votes
- 100 Counties in NC

VARIABLE SELECTION



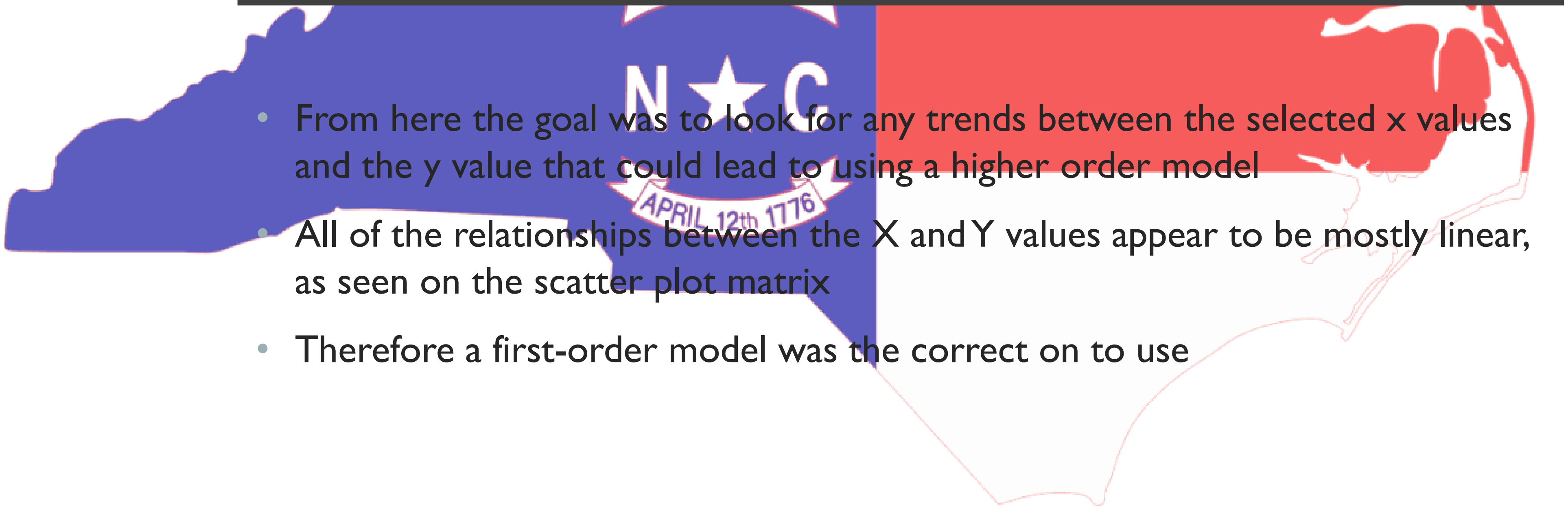
- Stepwise: 19-Variables before elimination
- Forward Selection: 41-Variables before elimination
- Backward Selection: 19-Variables before elimination
- Stepwise and Backward Selection both contained the same variables

VARIABLE SELECTION

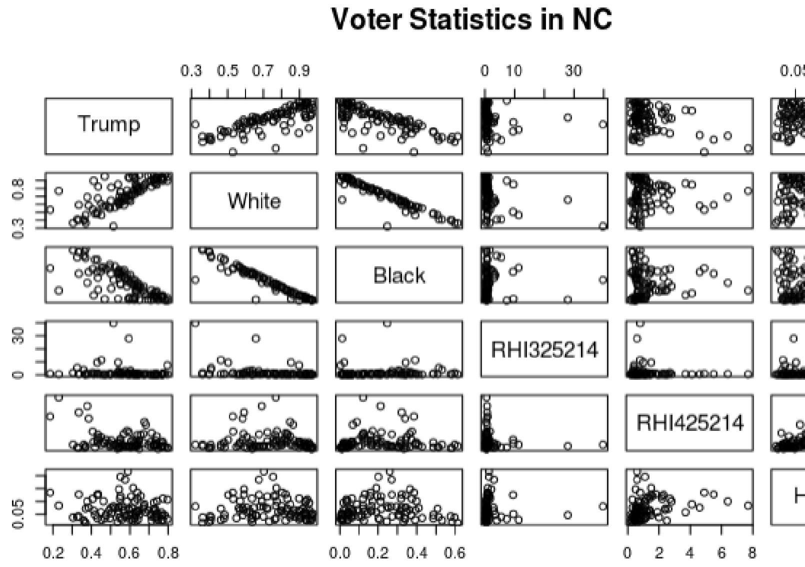


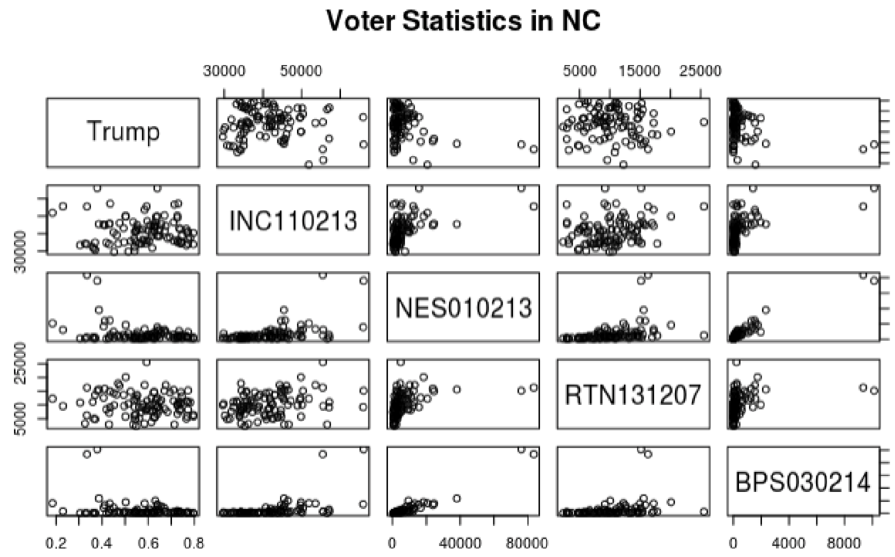
```
Call:
lm(formula = Trump ~ population2010 + AGE135214 + AGE295214 + 
    age65plus + SEX255214 + White + Black + RHI325214 + RHI425214 + 
    Hispanic + NonEnglish + Edu_batchelors + VET605213 + HSG096213 + 
    Income + INC110213 + NES010213 + RTN131207 + BPS030214, data = votes_nc)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.082952 -0.021929  0.002399  0.020110  0.071354 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     2.694e+00  1.102e+00   2.445 0.016670 *  
population2010  1.126e-06  3.691e-07   3.052 0.003083 ** 
AGE135214      -2.663e-02  9.996e-03  -2.664 0.009324 ** 
AGE295214       2.984e-02  4.590e-03   6.501 6.34e-09 ***
age65plus       9.382e-03  1.858e-03   5.051 2.72e-06 ***
SEX255214      -2.192e-02  3.724e-03  -5.886 8.87e-08 ***
White          -1.457e+00  1.080e+00  -1.349 0.181286    
Black          -2.201e+00  1.076e+00  -2.046 0.044056 *  
RHI325214      -1.930e-02  1.133e-02  -1.704 0.092300 .  
RHI425214      -3.488e-02  1.473e-02  -2.368 0.020277 *  
Hispanic       -1.566e+00  3.531e-01  -4.435 2.91e-05 ***
NonEnglish      1.212e-02  3.720e-03   3.258 0.001648 ** 
Edu_batchelors -1.087e-02  1.262e-03  -8.611 5.11e-13 ***
VET605213      -4.696e-06  2.203e-06  -2.132 0.036093 *  
HSG096213       3.686e-03  1.264e-03   2.915 0.004613 ** 
Income          1.161e-05  3.682e-06   3.153 0.002274 ** 
INC110213      -3.109e-06  1.595e-06  -1.949 0.054808 .  
NES010213      -1.281e-05  3.396e-06  -3.773 0.000308 ***
RTN131207       2.198e-06  1.146e-06   1.919 0.058543 .  
BPS030214       2.550e-05  1.043e-05   2.445 0.016682 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03211 on 80 degrees of freedom
Multiple R-squared:  0.9557,	Adjusted R-squared:  0.9452 
F-statistic: 90.85 on 19 and 80 DF,  p-value: < 2.2e-16
```

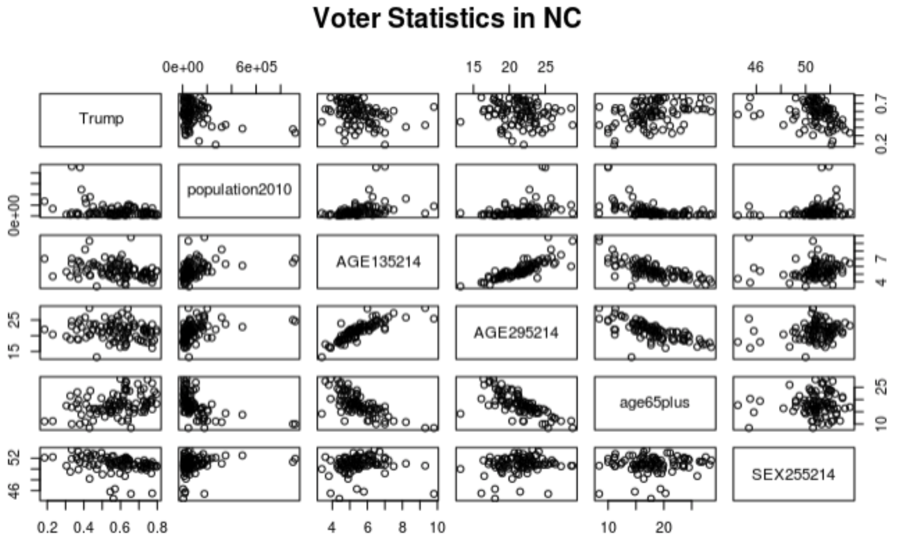
LOOKING FOR TRENDS

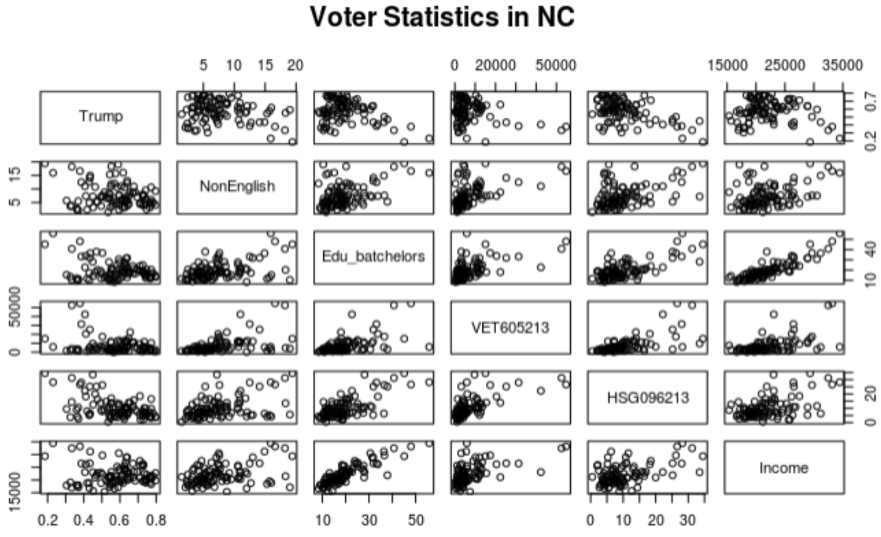


- From here the goal was to look for any trends between the selected x values and the y value that could lead to using a higher order model
- All of the relationships between the X and Y values appear to be mostly linear, as seen on the scatter plot matrix
- Therefore a first-order model was the correct on to use


Voter Statistics in NC
Trump
White
Black
RHI325214
RHI425214
Hispanic
0.3 0.5 0.7 0.9
0 10 30
0.05 0.15
0.2 0.7
0.3 0.8
0.0 0.5
0 30
0 4 8
0.05
0.2 0.4 0.6 0.8
0.0 0.2 0.4 0.6
0 2 4 6 8

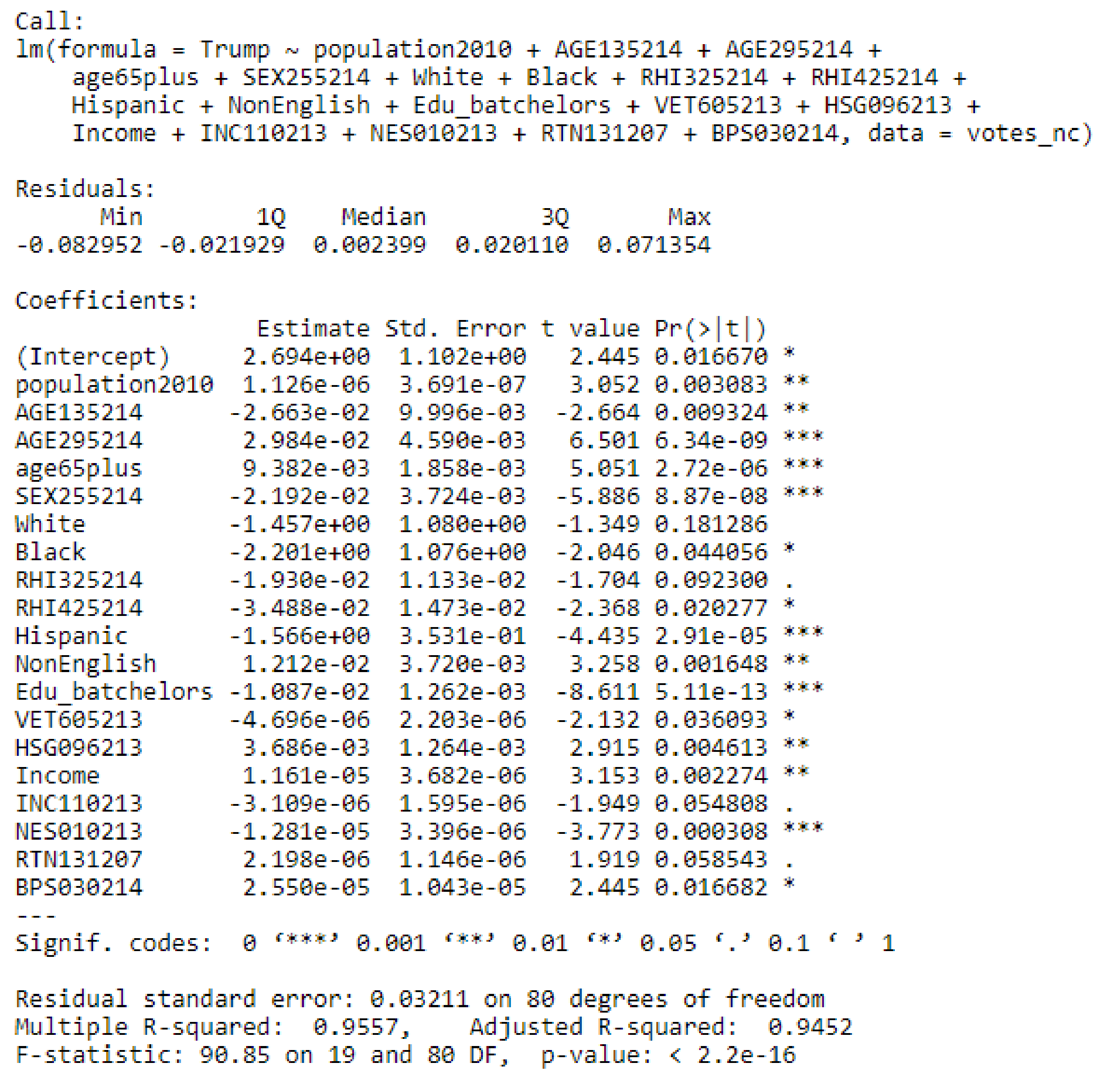

Voter Statistics in NC
Trump
INC110213
NES010213
RTN131207
BPS030214
30000 50000
5000 15000 25000
0.2 0.6
30000
0 60000
5000 25000
0 6000
0.2 0.4 0.6 0.8
0 40000 80000
0 4000 8000


Voter Statistics in NC
Trump
population2010
AGE135214
AGE295214
age65plus
SEX255214
0e+00 6e+05
15 20 25
46 50
0.2 0.7
0e+00
4 7
15 25
10 25
46 52
0.2 0.4 0.6 0.8
4 6 8 10
10 20


Voter Statistics in NC
Trump
NonEnglish
Edu_batchelors
VET605213
HSG096213
Income
5 10 15 20
0 20000 50000
15000 25000 35000
0.2 0.7
5 15
10 40
0 50000
0 20
15000
0.2 0.4 0.6 0.8
10 30 50
0 10 20 30

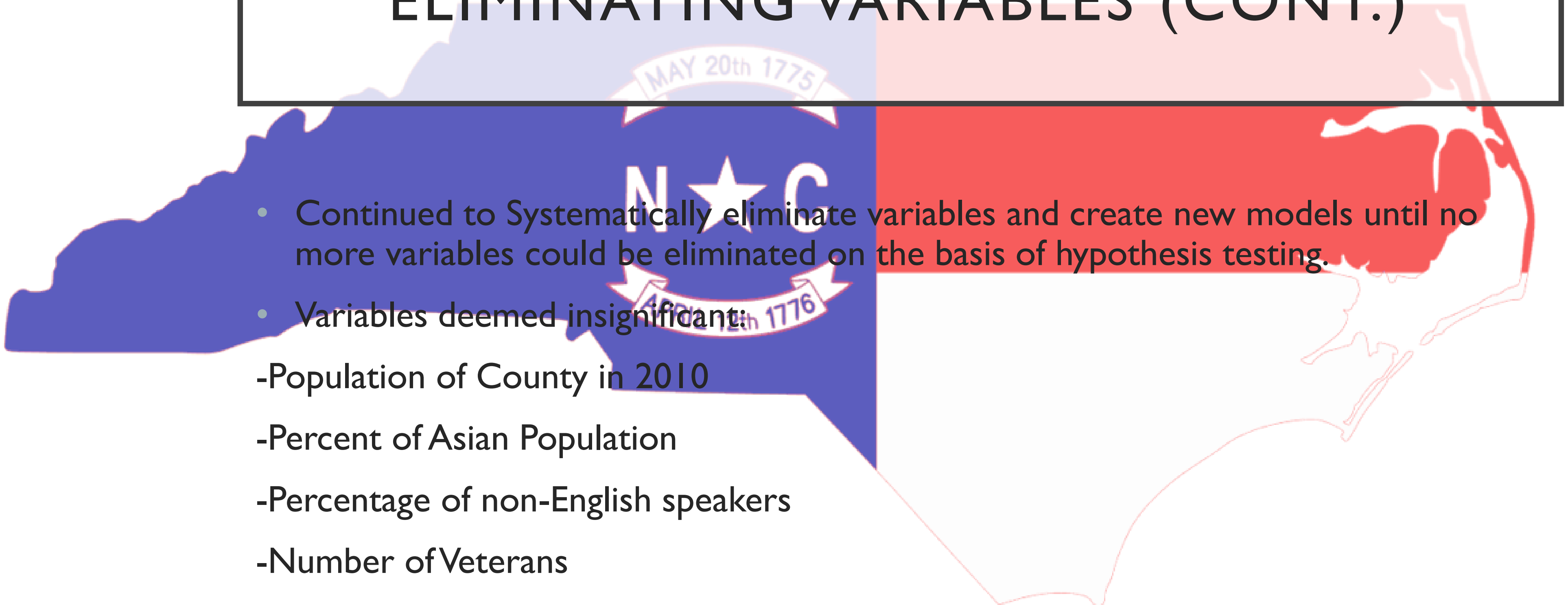
ELIMINATING VARIABLES

- Started With 19 Variables
- Used hypothesis testing at alpha level = .05
- After Hypothesis Testing Eliminated 4 Variables
- Percentage of population that is white, Percentage of the population that is American Indian, Median Household Income, and Retail Sales Per Capita.
- Ending with 15 Variables



```
Call:
lm(formula = Trump ~ population2010 + AGE135214 + AGE295214 +
    age65plus + SEX255214 + White + Black + RHI325214 + RHI425214 +
    Hispanic + NonEnglish + Edu_batchelors + VET605213 + HSG096213 +
    Income + INC110213 + NES010213 + RTN131207 + BPS030214, data = votes_nc)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.082952 -0.021929  0.002399  0.020110  0.071354 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     2.694e+00  1.102e+00   2.445 0.016670 *  
population2010  1.126e-06  3.691e-07   3.052 0.003083 ** 
AGE135214      -2.663e-02  9.996e-03  -2.664 0.009324 ** 
AGE295214       2.984e-02  4.590e-03   6.501 6.34e-09 ***
age65plus       9.382e-03  1.858e-03   5.051 2.72e-06 ***
SEX255214      -2.192e-02  3.724e-03  -5.886 8.87e-08 ***
White          -1.457e+00  1.080e+00  -1.349 0.181286    
Black          -2.201e+00  1.076e+00  -2.046 0.044056 *  
RHI325214      -1.930e-02  1.133e-02  -1.704 0.092300 .  
RHI425214      -3.488e-02  1.473e-02  -2.368 0.020277 *  
Hispanic       -1.566e+00  3.531e-01  -4.435 2.91e-05 ***
NonEnglish      1.212e-02  3.720e-03   3.258 0.001648 ** 
Edu_batchelors -1.087e-02  1.262e-03  -8.611 5.11e-13 ***
VET605213      -4.696e-06  2.203e-06  -2.132 0.036093 *  
HSG096213       3.686e-03  1.264e-03   2.915 0.004613 ** 
Income          1.161e-05  3.682e-06   3.153 0.002274 ** 
INC110213      -3.109e-06  1.595e-06  -1.949 0.054808 .  
NES010213      -1.281e-05  3.396e-06  -3.773 0.000308 ***
RTN131207       2.198e-06  1.146e-06   1.919 0.058543 .  
BPS030214       2.550e-05  1.043e-05   2.445 0.016682 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03211 on 80 degrees of freedom
Multiple R-squared:  0.9557,	Adjusted R-squared:  0.9452 
F-statistic: 90.85 on 19 and 80 DF,  p-value: < 2.2e-16
```

ELIMINATING VARIABLES (CONT.)



- Continued to Systematically eliminate variables and create new models until no more variables could be eliminated on the basis of hypothesis testing.
- Variables deemed insignificant:

-Population of County in 2010

-Percent of Asian Population

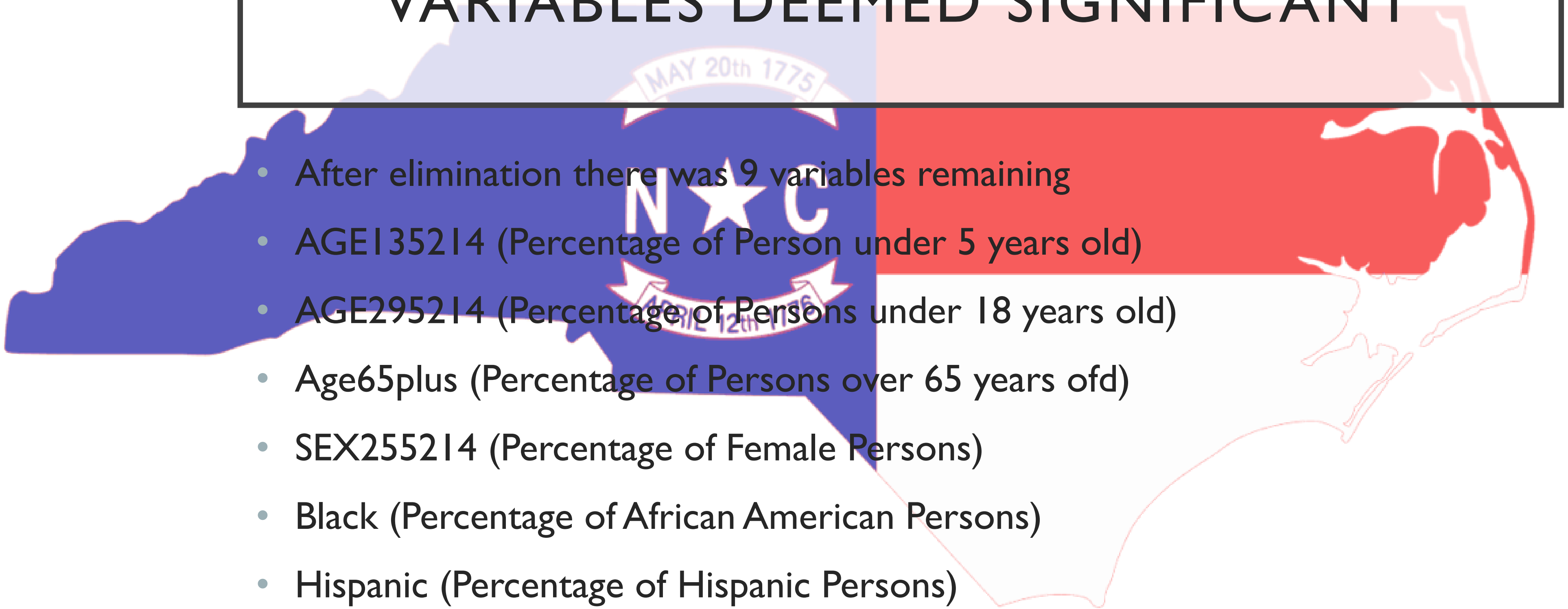
-Percentage of non-English speakers

-Number of Veterans

-Number of Non-employer Establishments (Business that has no employees)

-Number of building permits

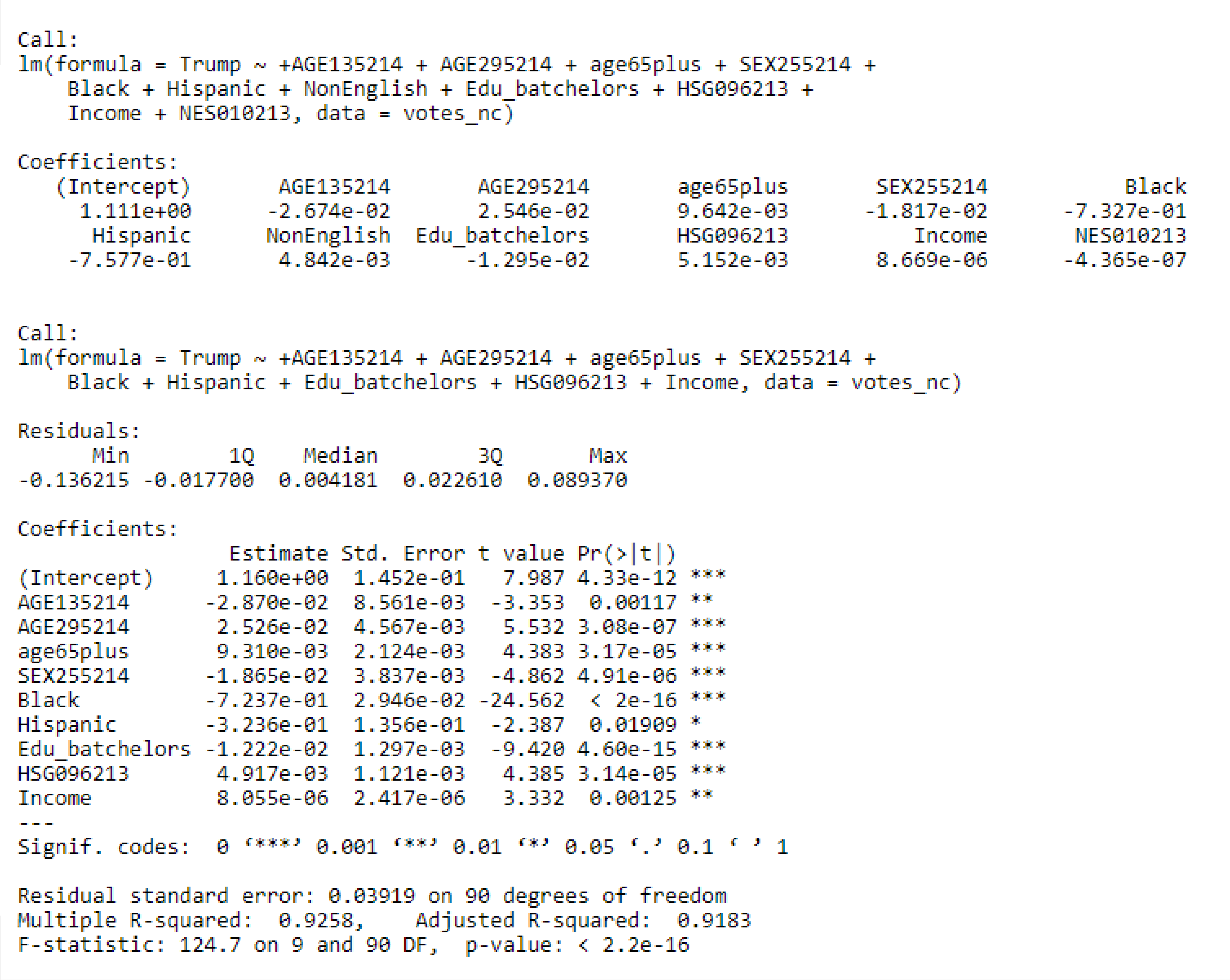
VARIABLES DEEMED SIGNIFICANT



- After elimination there was 9 variables remaining
- AGE135214 (Percentage of Person under 5 years old)
- AGE295214 (Percentage of Persons under 18 years old)
- Age65plus (Percentage of Persons over 65 years ofd)
- SEX255214 (Percentage of Female Persons)
- Black (Percentage of African American Persons)
- Hispanic (Percentage of Hispanic Persons)
- Edu_bachelors (Percentage of Person over 25 with a bachelor's degree or higher)
- HSG096213 (Number of Housing units in multi-unit strictures, percent)
- Income (Per capita money income in past 12 months)

REDUCED MODEL

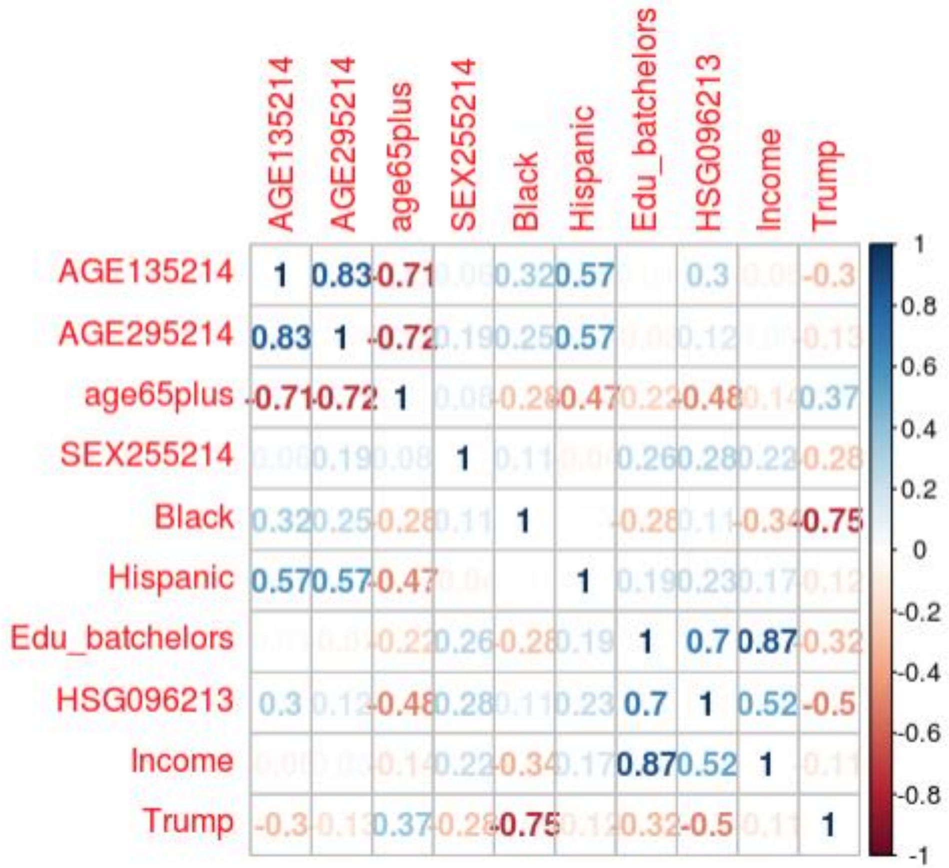
- Notes from the Summary
- From this model it seems that the most important variable is the Percentage of African Americans in the county



```
Call:
lm(formula = Trump ~ +AGE135214 + AGE295214 + age65plus + SEX255214 + 
    Black + Hispanic + NonEnglish + Edu_batchelors + HSG096213 + 
    Income + NES010213, data = votes_nc)

Coefficients:
   (Intercept)       AGE135214       AGE295214       age65plus       SEX255214           Black  
     1.111e+00      -2.674e-02       2.546e-02       9.642e-03      -1.817e-02      -7.327e-01  
      Hispanic      NonEnglish  Edu_batchelors       HSG096213          Income       NES010213  
    -7.577e-01       4.842e-03      -1.295e-02       5.152e-03       8.669e-06      -4.365e-07  


Call:
lm(formula = Trump ~ +AGE135214 + AGE295214 + age65plus + SEX255214 + 
    Black + Hispanic + Edu_batchelors + HSG096213 + Income, data = votes_nc)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.136215 -0.017700  0.004181  0.022610  0.089370 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     1.160e+00  1.452e-01   7.987 4.33e-12 ***
AGE135214      -2.870e-02  8.561e-03  -3.353  0.00117 ** 
AGE295214       2.526e-02  4.567e-03   5.532 3.08e-07 ***
age65plus       9.310e-03  2.124e-03   4.383 3.17e-05 ***
SEX255214      -1.865e-02  3.837e-03  -4.862 4.91e-06 ***
Black          -7.237e-01  2.946e-02 -24.562  < 2e-16 ***
Hispanic       -3.236e-01  1.356e-01  -2.387  0.01909 *  
Edu_batchelors -1.222e-02  1.297e-03  -9.420 4.60e-15 ***
HSG096213       4.917e-03  1.121e-03   4.385 3.14e-05 ***
Income          8.055e-06  2.417e-06   3.332  0.00125 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03919 on 90 degrees of freedom
Multiple R-squared:  0.9258,	Adjusted R-squared:  0.9183 
F-statistic: 124.7 on 9 and 90 DF,  p-value: < 2.2e-16
```

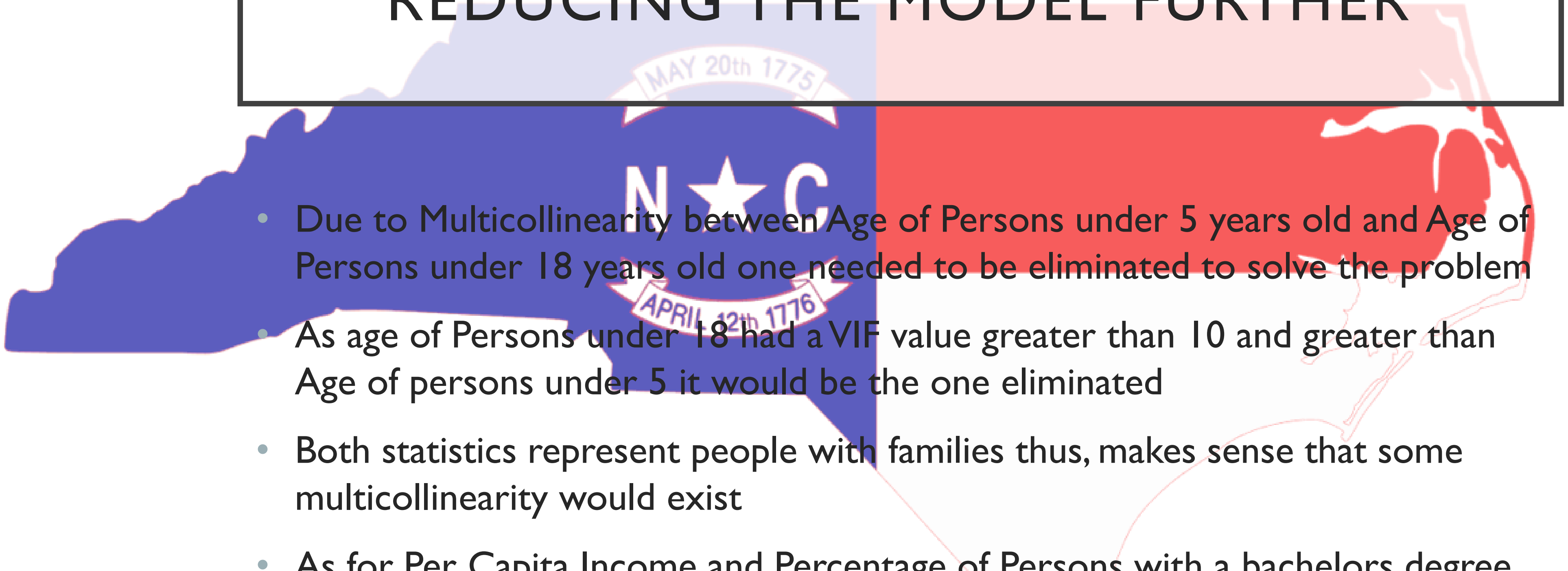
TESTING MODEL DOWNFALLS

- Next I tested the model for downfalls such as multicollinearity
- There appeared to be some multicollinearity between the values AGE135214 and AGE 295214
- There was also multicollinearity Between Income and Edu_Bachelors
- Would need to test for VIF to see what variables to elminate

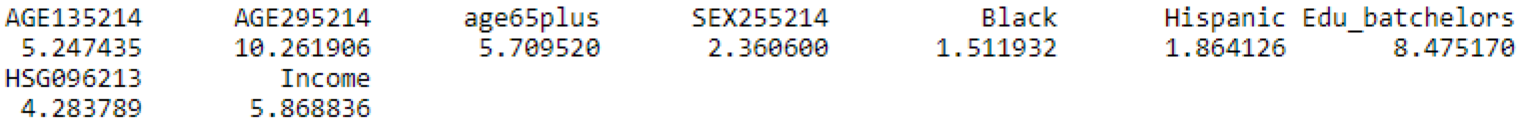


	AGE135214	AGE295214	age65plus	SEX255214	Black	Hispanic	Edu_batchelors	HSG096213	Income	Trump
AGE135214	1	0.83	-0.71	0.06	0.32	0.57		0.3	0.05	-0.3
AGE295214	0.83	1	-0.72	0.19	0.25	0.57	[illegible]	0.12	[illegible]	-0.13
age65plus	-0.71	-0.72	1	0.08	-0.28	-0.47	-0.22	-0.48	-0.14	0.37
SEX255214	0.06	0.19	0.08	1	0.11	[illegible]	0.26	0.28	0.22	-0.28
Black	0.32	0.25	-0.28	0.11	1		-0.28	0.11	-0.34	-0.75
Hispanic	0.57	0.57	-0.47	[illegible]		1	0.19	0.23	0.17	-0.12
Edu_batchelors		[illegible]	-0.22	0.26	-0.28	0.19	1	0.7	0.87	-0.32
HSG096213	0.3	0.12	-0.48	0.28	0.11	0.23	0.7	1	0.52	-0.5
Income	0.05	[illegible]	-0.14	0.22	-0.34	0.17	0.87	0.52	1	-0.11
Trump	-0.3	-0.13	0.37	-0.28	-0.75	-0.12	-0.32	-0.5	-0.11	1

REDUCING THE MODEL FURTHER



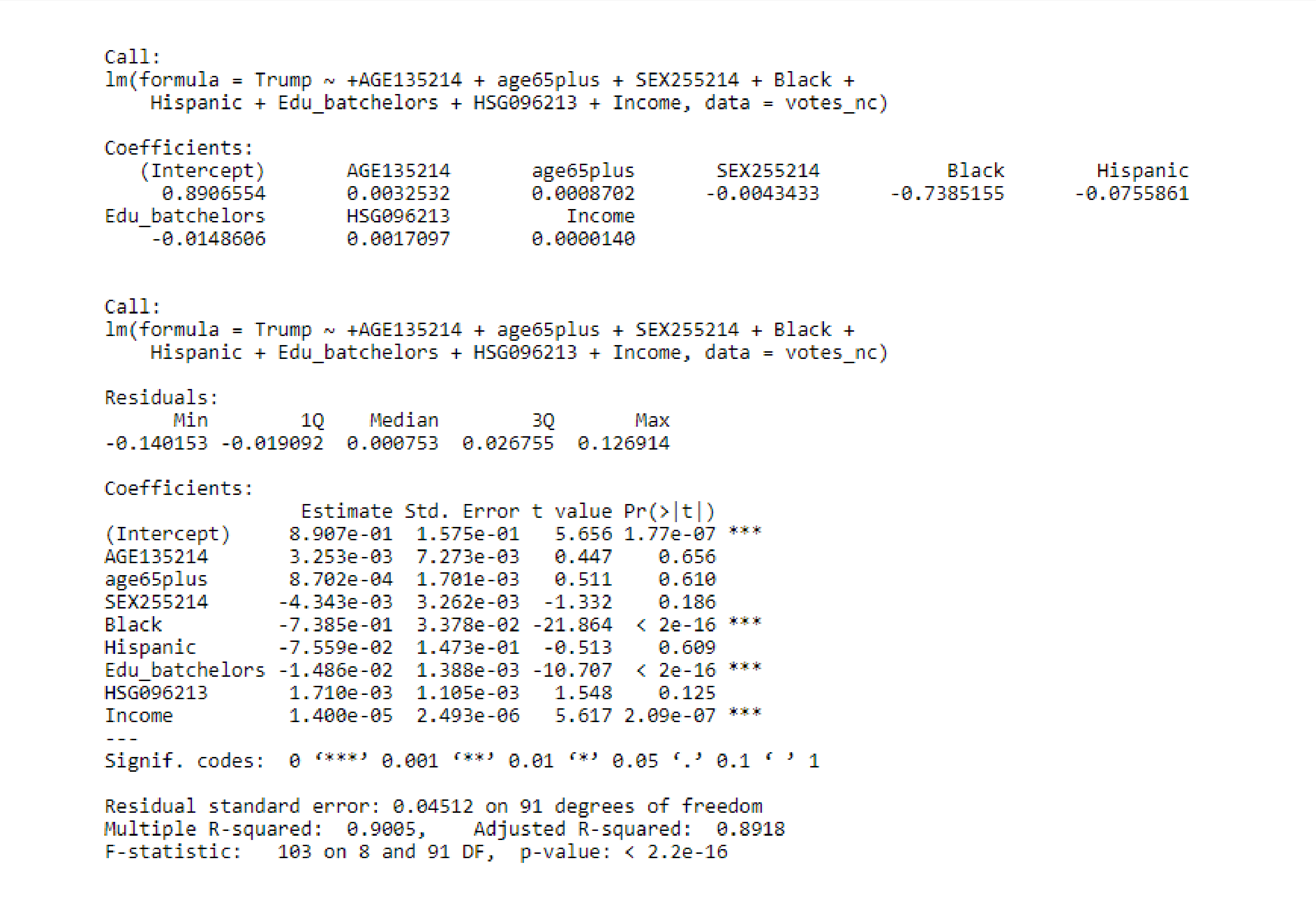
- Due to Multicollinearity between Age of Persons under 5 years old and Age of Persons under 18 years old one needed to be eliminated to solve the problem
- As age of Persons under 18 had a VIF value greater than 10 and greater than Age of persons under 5 it would be the one eliminated
- Both statistics represent people with families thus, makes sense that some multicollinearity would exist
- As for Per Capita Income and Percentage of Persons with a bachelors degree Percentage of Persons with a bachelors degree has the higher VIF value so it was selected for elimination



```
AGE135214     AGE295214     age65plus     SEX255214         Black      Hispanic Edu_batchelors
 5.247435     10.261906      5.709520      2.360600      1.511932      1.864126       8.475170
HSG096213        Income
 4.283789      5.868836
```

MODEL AFTER TESTING

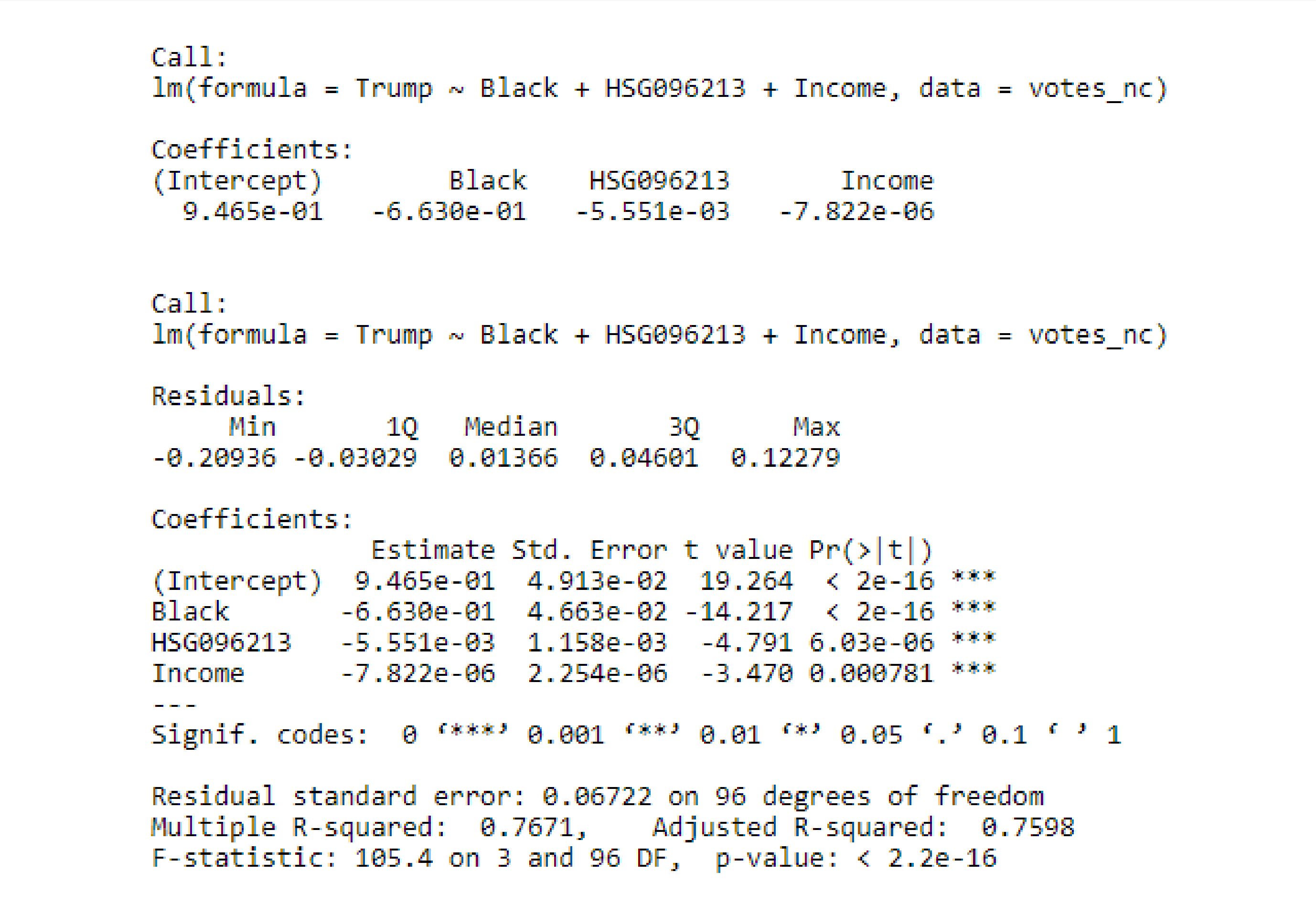
- After elimination the variables to solve multicollinearity the model was then ran again giving the output shown
- As seen some p-values forsome variables are higher than in the previous model
- Hypothesis Testing at a significance level of .05 was used again to eliminate insignificant variables



```
Call:
lm(formula = Trump ~ +AGE135214 + age65plus + SEX255214 + Black +
    Hispanic + Edu_batchelors + HSG096213 + Income, data = votes_nc)

Coefficients:
   (Intercept)       AGE135214       age65plus       SEX255214           Black        Hispanic
     0.8906554       0.0032532       0.0008702      -0.0043433      -0.7385155      -0.0755861
Edu_batchelors       HSG096213          Income
    -0.0148606       0.0017097       0.0000140


Call:
lm(formula = Trump ~ +AGE135214 + age65plus + SEX255214 + Black +
    Hispanic + Edu_batchelors + HSG096213 + Income, data = votes_nc)

Residuals:
      Min        1Q    Median        3Q       Max
-0.140153 -0.019092  0.000753  0.026755  0.126914

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     8.907e-01  1.575e-01   5.656 1.77e-07 ***
AGE135214       3.253e-03  7.273e-03   0.447    0.656
age65plus       8.702e-04  1.701e-03   0.511    0.610
SEX255214      -4.343e-03  3.262e-03  -1.332    0.186
Black          -7.385e-01  3.378e-02 -21.864  < 2e-16 ***
Hispanic       -7.559e-02  1.473e-01  -0.513    0.609
Edu_batchelors -1.486e-02  1.388e-03 -10.707  < 2e-16 ***
HSG096213       1.710e-03  1.105e-03   1.548    0.125
Income          1.400e-05  2.493e-06   5.617 2.09e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04512 on 91 degrees of freedom
Multiple R-squared:  0.9005,	Adjusted R-squared:  0.8918
F-statistic:   103 on 8 and 91 DF,  p-value: < 2.2e-16
```

MODEL AFTER REDUCING FOR MULTICOLLINEARITY

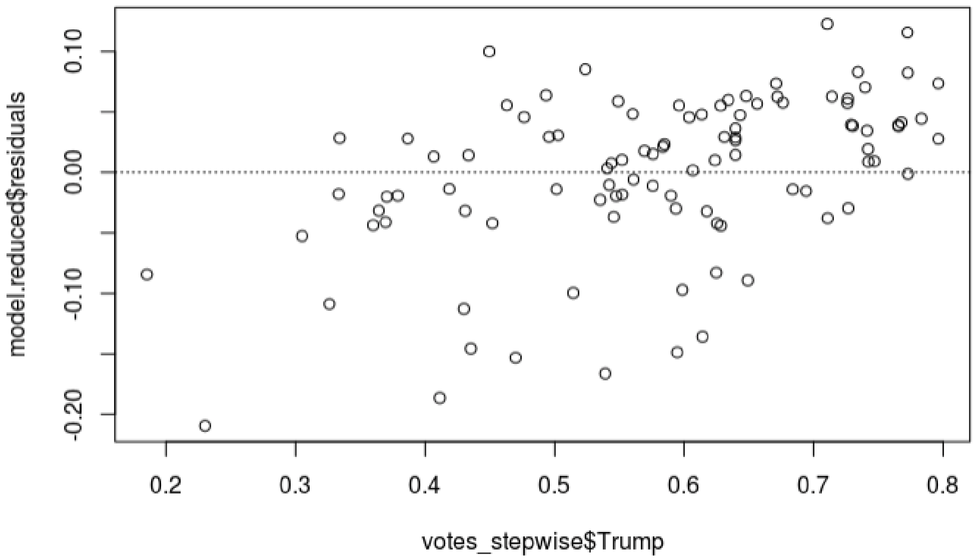
- This model has been tested for model downfalls and had all insignificant variables removed
- This model appears to be statistically significant
- F = 105.4
- P-value = 2.2e-16
- $R^2 adjusted$ = .7598



```
Call:
lm(formula = Trump ~ Black + HSG096213 + Income, data = votes_nc)

Coefficients:
(Intercept)        Black    HSG096213       Income  
  9.465e-01   -6.630e-01   -5.551e-03   -7.822e-06  


Call:
lm(formula = Trump ~ Black + HSG096213 + Income, data = votes_nc)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.20936 -0.03029  0.01366  0.04601  0.12279 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  9.465e-01  4.913e-02  19.264  < 2e-16 ***
Black       -6.630e-01  4.663e-02 -14.217  < 2e-16 ***
HSG096213   -5.551e-03  1.158e-03  -4.791 6.03e-06 ***
Income      -7.822e-06  2.254e-06  -3.470 0.000781 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06722 on 96 degrees of freedom
Multiple R-squared:  0.7671,	Adjusted R-squared:  0.7598 
F-statistic: 105.4 on 3 and 96 DF,  p-value: < 2.2e-16
```

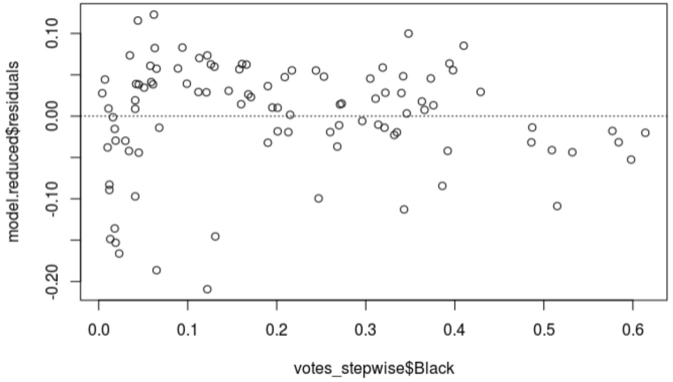
TESTING FOR TRANSITIONS

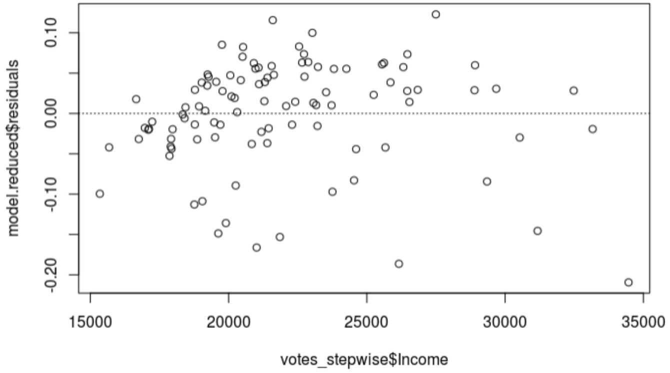
- As seen in the graph in the top left there are no need for transitions to occur on the y-value as the residual plot follows all normal assumptions

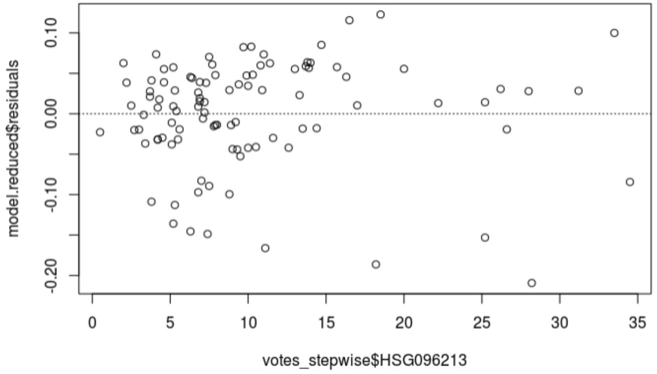


TRANSITIONS IN THE X-VARIABLES

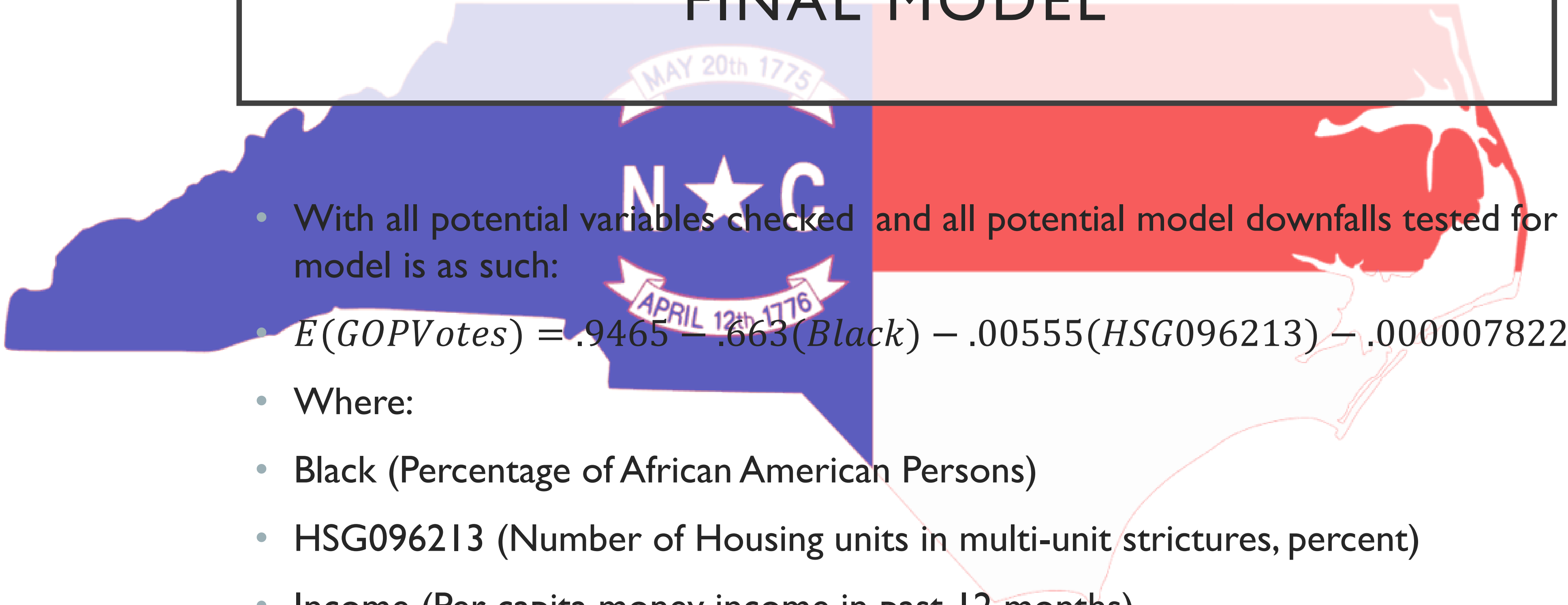
- As seen in the graphs below there is no need for transitions to occur in any of the X-values, as all residual plots seem to indicate a nearly normal model.





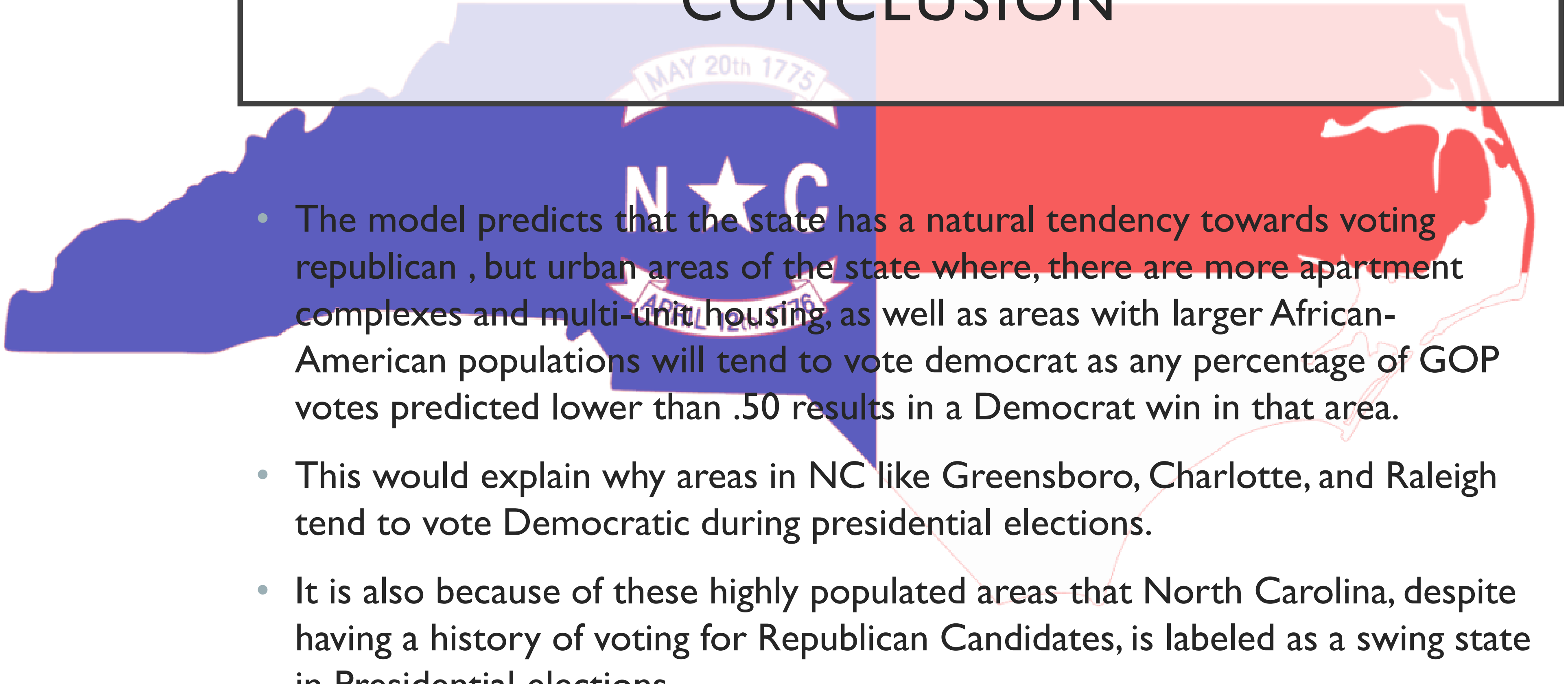


FINAL MODEL



- With all potential variables checked and all potential model downfalls tested for the final model is as such:
- $E(GOPVotes) = .9465 - .663(Black) - .00555(HSG096213) - .000007822(Income)$
- Where:
- Black (Percentage of African American Persons)
- HSG096213 (Number of Housing units in multi-unit strictures, percent)
- Income (Per capita money income in past 12 months)

CONCLUSION



- The model predicts that the state has a natural tendency towards voting republican , but urban areas of the state where, there are more apartment complexes and multi-unit housing, as well as areas with larger African-American populations will tend to vote democrat as any percentage of GOP votes predicted lower than .50 results in a Democrat win in that area.
- This would explain why areas in NC like Greensboro, Charlotte, and Raleigh tend to vote Democratic during presidential elections.
- It is also because of these highly populated areas that North Carolina, despite having a history of voting for Republican Candidates, is labeled as a swing state in Presidential elections.

QUESTIONS?

THANK YOU