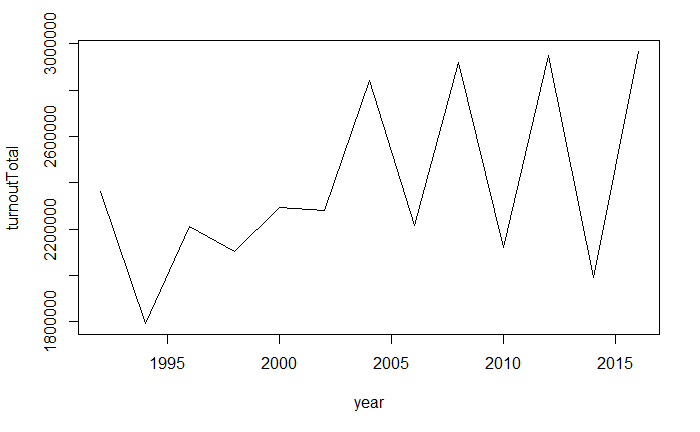
1. On 2016:

4120 precinct

87 counties

8 district

1. Presidential year vs. non presidential year



1. Difference between: Congress District, County and precinct

A congress district contains several precincts.

A county also contains several precincts.

However, Congress District and County has no subset (contain) relationship

1. Why we are using County instead of precinct

Most data set we can find is based on County. We cannot ignore these data.

Instead, we are using County as the smallest unit when we training our model.

1. How we dealing with the ‘One county multi Congress District’ Problem?
   1. Only few county are shared by different district (9 counties)
   2. We will divide these county by the historical registered voter @ 7 am, assign them into different district.
2. What variables we are looking?
   1. Unemployment rate
3. Why?
   1. According to our discover, unemployment rate won’t help us tell the difference (of party share)
4. Beside the dataset, what else we are considering?
   1. Polling data
   2. Minnesota google search frequency
   3. Candidate opinion
   4. Candidate experience, age, race and gender.
5. Why they are important or not important?
   1. Polling data directly shows how people love each candidate.
   2. Search frequency is combined with candidate political opinion to c.
6. Why we are not using it or why we are using it?
   1. We are not using polling data because we decide to only use historical data

And

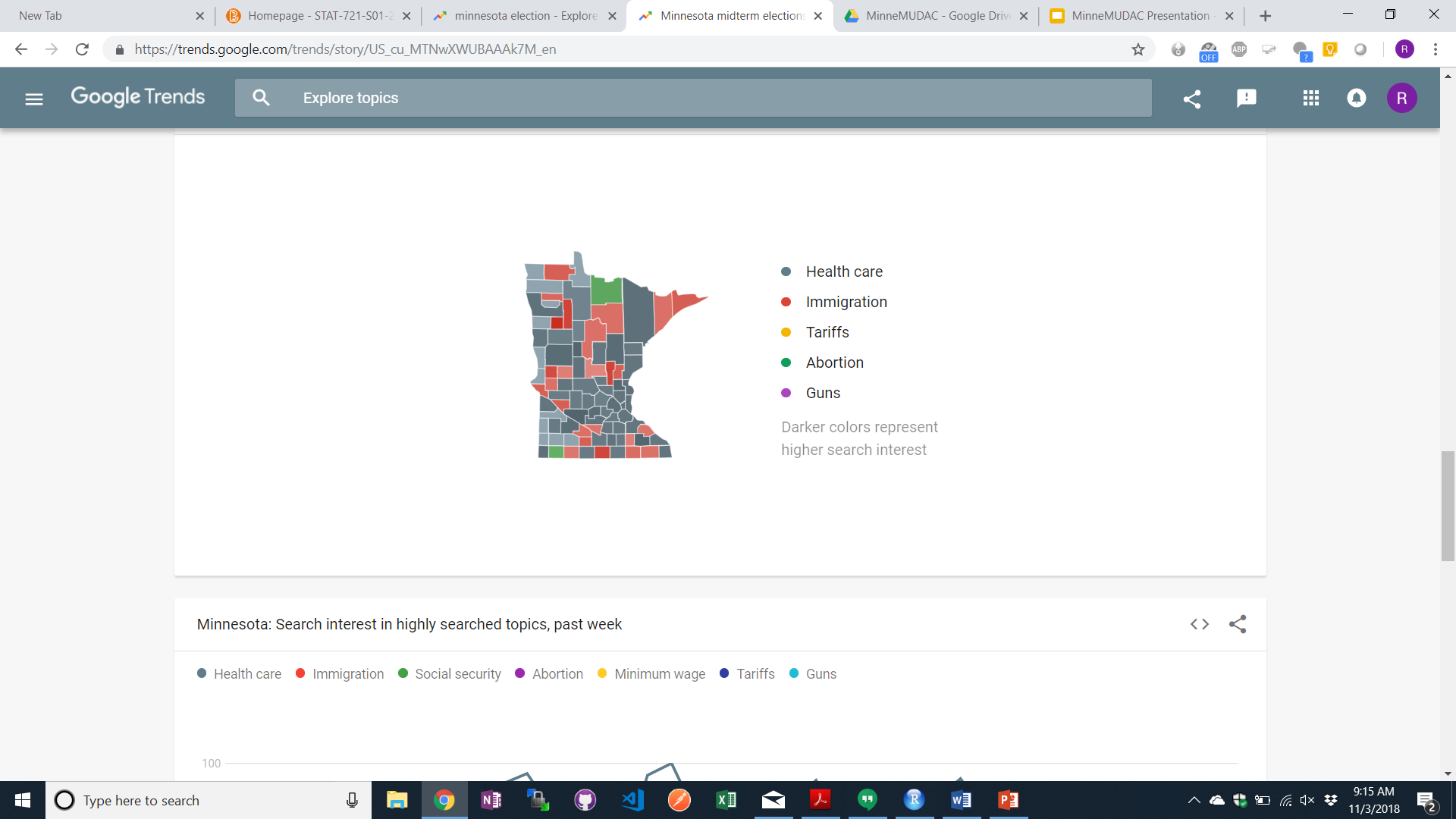
We are not able to get the polling data in county level

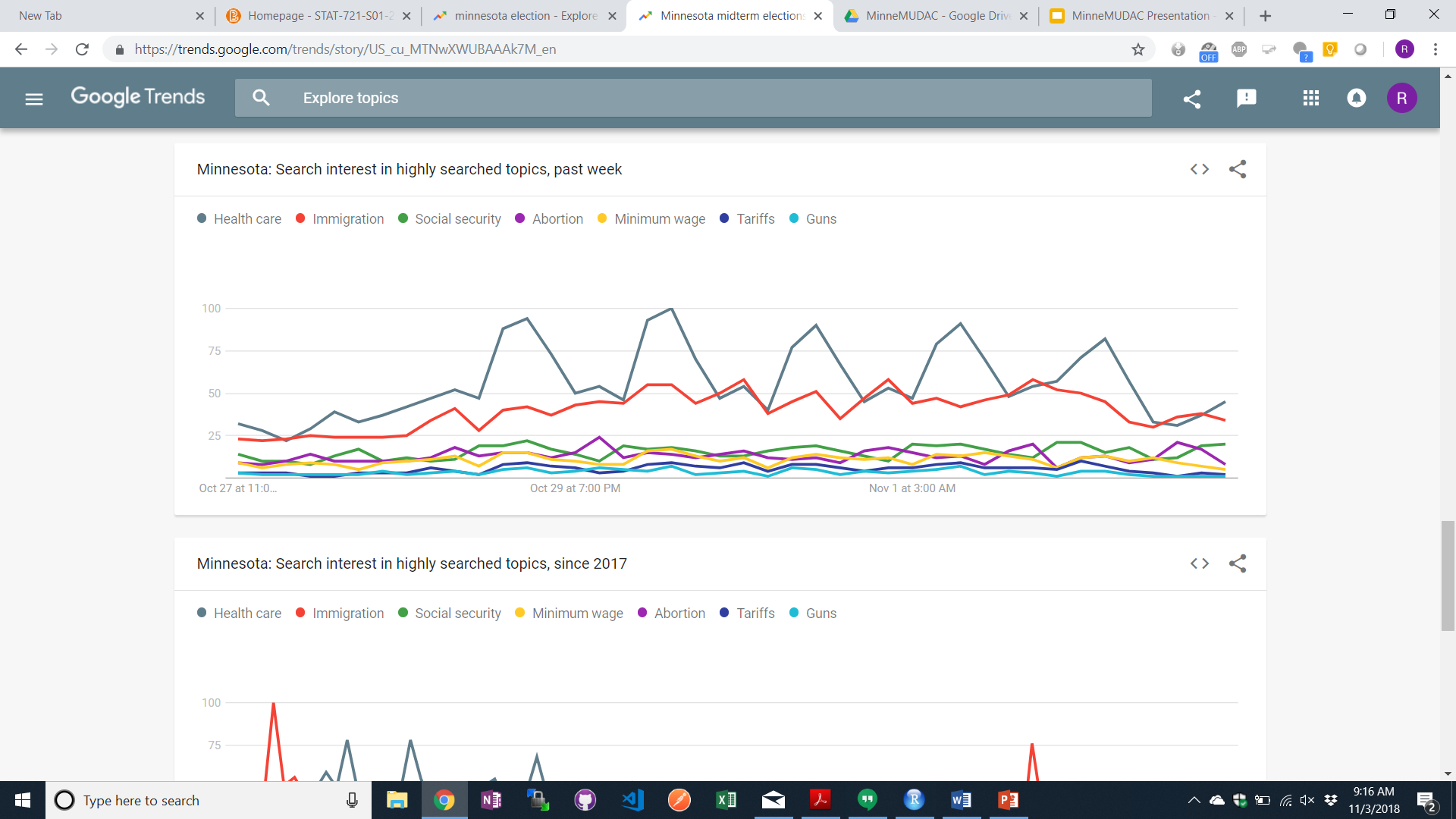
And

Polling data are bias since they may representing some party

* 1. We use google trends to find the topic Minnesota people are interested in

And we found Health Care and Immigration





* 1. We viewed the news of each candidate, they did some statement regarding to these topic. But none of them are hot.

1. Modeling:
   1. Linear model
   2. Negative binomial
   3. Random Forest
2. Workflow:
   1. **i. Data Explorer  
      ii. Adding important variable**
   2. **i. Eliminate high correlated variables  
      ii. Using Random Forest to select the important variables**
   3. **i. Using Linear regression, Negative binomial and Random forest  
      ii. Checking the error rate of each model, using cross validation**