# The problem statement

If during the early part of 2014 we had to predict the upcoming Lok Sabha Elections, what would be the predictive model?

-Identify data sources.

-Identifying the right technique/techniques to model that data

-Make predictions, validate and update your model.

# The various references for the data sources that we used as our final dataset to build the Predictive Model

* <http://eci.nic.in/eci/eci.html>
* <http://www.indiavotes.com/>
* <http://www.myneta.info/>
* <http://planningcommission.nic.in/data/datatable/>

# What factors we think can determine winning elections?

1. To train the model, we need historical data of past Lok Sabha elections. Hence, we started from the year 2004. For 2014 Lok Sabha Election Prediction model, we decided to take the 2004 and 2009 Lok Sabha election data.
2. We focused on taking data at the constituency level because every constituency will represent one seat in the parliament
3. We took the winning candidate details for the year 2004 and 2009 at the Constituency level.
4. Dependent feature is Party which will have values/labels as ‘INC’, ‘BJP’, ‘Others’
5. Independent features are Type, Turnout, Gender, Age, PrevParty (which was the previous ruling Party), GSDP (Gross State Domestic Product) growth rate, Candidate criminal cases, Candidate education, Candidate Assets, Candidate Liabilities, GrowthRateAgri (State-wise Growth Rate of Agriculture & Allied Sector in India)
6. We collected winning netas information on criminal cases, education, assets and liabilities. Apart from the Gender and Age. Script written to clean the special character and added to our base dataset consisting of Constituency, Party, State , Type, Gender, Age, Turnout
7. Economic indicators are GSDP and GrowthRateAgri
8. Turnout is Total Votes Polled / Total electors

# Challenges in Data Collection

* The data scattered across different sources and not available in csv/excel format except the past election results from Election Commission website.
* Complete automation script not possible as joining on winning candidate name or constituency will not be completely successful because of spelling mismatch for candidate name or constituency across different data sources. For example, Pondicherry vs Puducherry, Aruku vs Araku etc
* Between, 2004 and 2009 many constituencies have changed. 1999 and 2004 more similar where similar constituencies details can be obtained. Similarly for 2004 and 2009 the constituencies are pretty much similar. This created challenge in obtaining the PrevParty details for the year 2009 from previous election 2004
* Could not find the Religious related population data for the year 2004, 2009 and 2014 which could have been an important predictor

# Exploratory Data Analysis

### Average turnout for BJP in the year 2004, 2009 and 2014

### 

We can see decrease in the Turnout for BJP from 2004 to 2009 and then increase in 2014

### Average turnout for INC in the year 2004, 2009 and 2014

### 

Increase in Turnout for INC from 2004 to 2009 to 2014

### Average turnout for Others in the year 2004, 2009 and 2014

### 

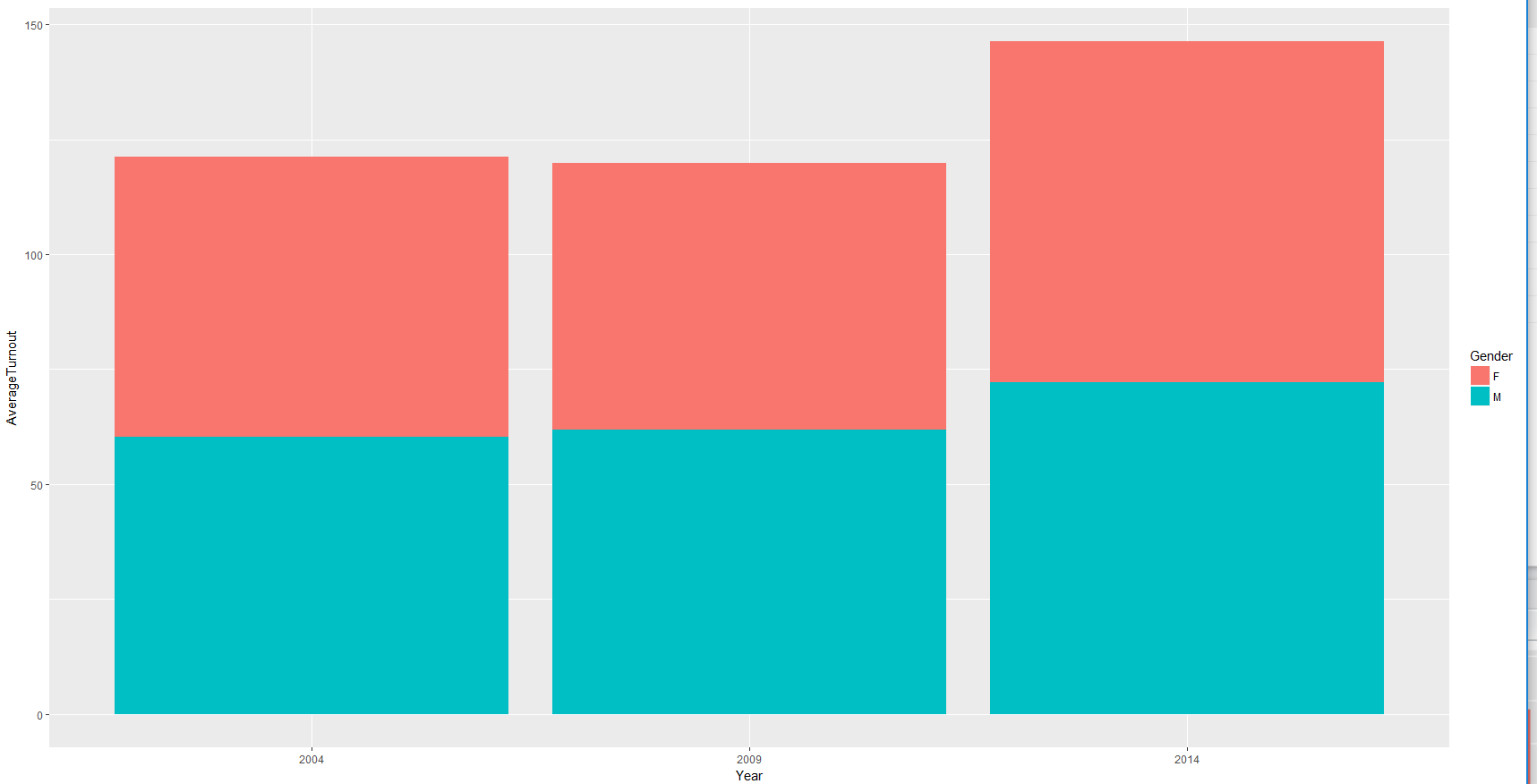
### Average turnout for BJP in the year 2004, 2009 and 2014 based on Gender

### 

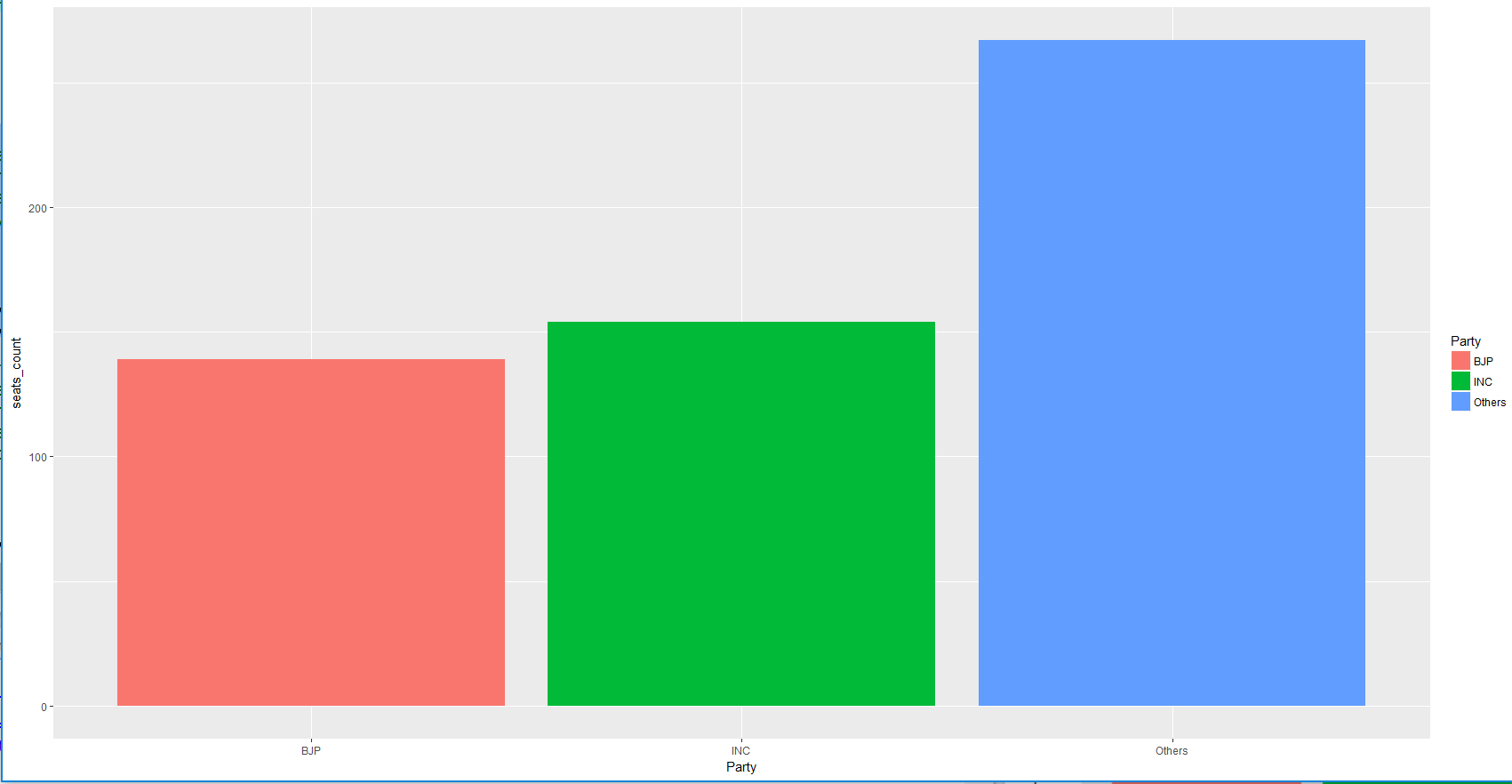
### Average turnout for INC in the year 2004, 2009 and 2014 based on Gender

### 

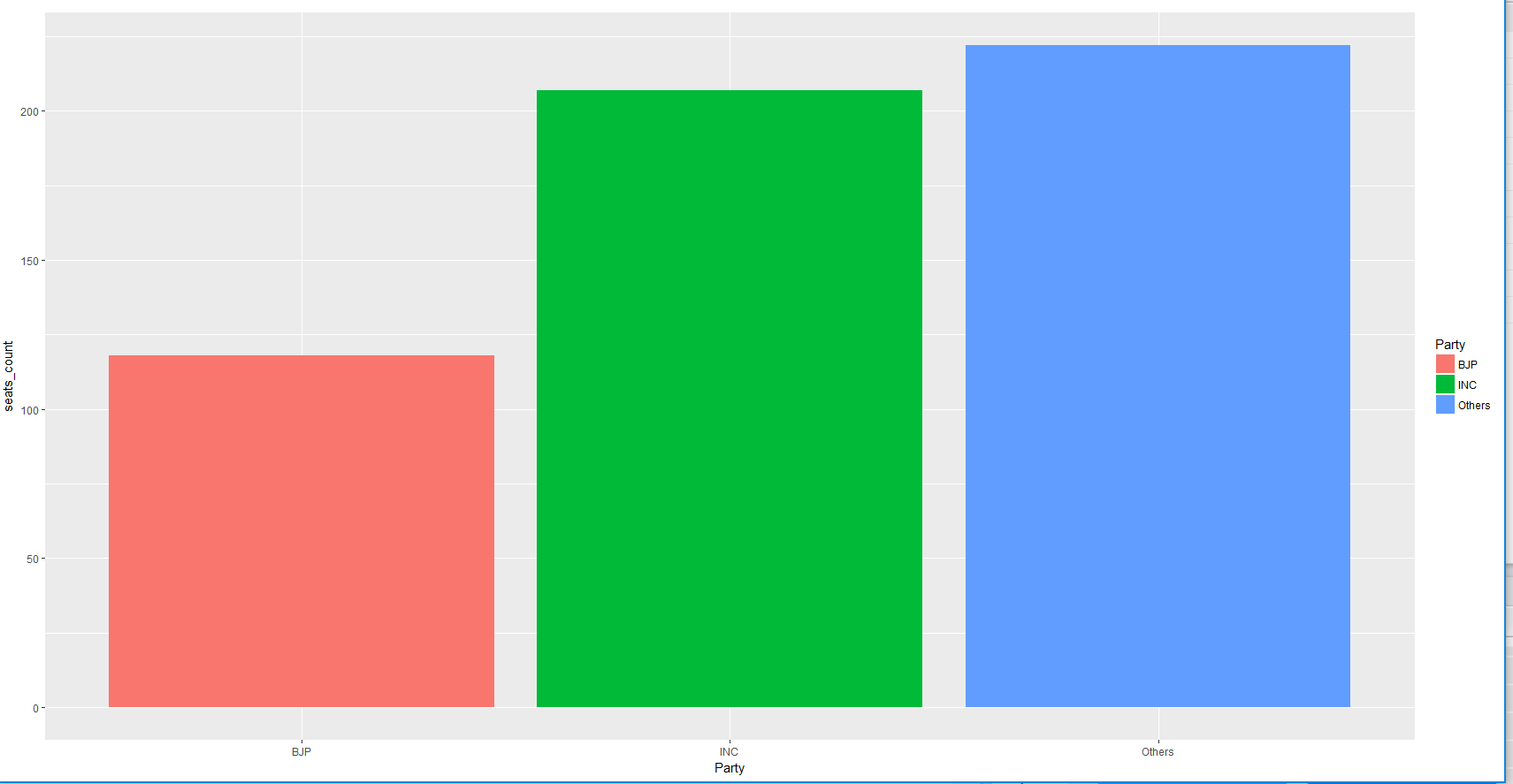
### Average turnout for Others in the year 2004, 2009 and 2014 based on Gender



### 2004 and Total Seats for BJP, INC and Others

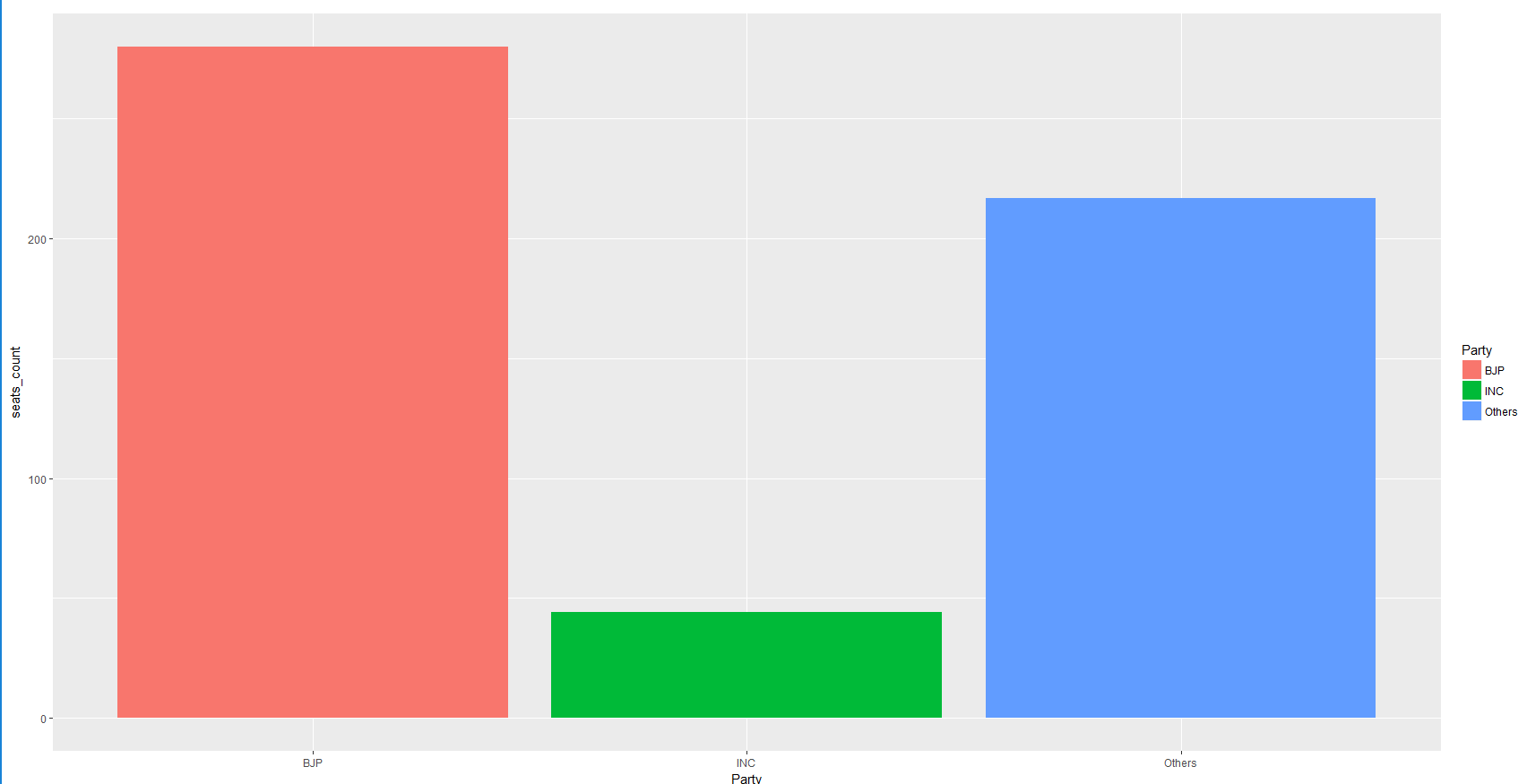


### 2009 and Total Seats for BJP, INC and Others



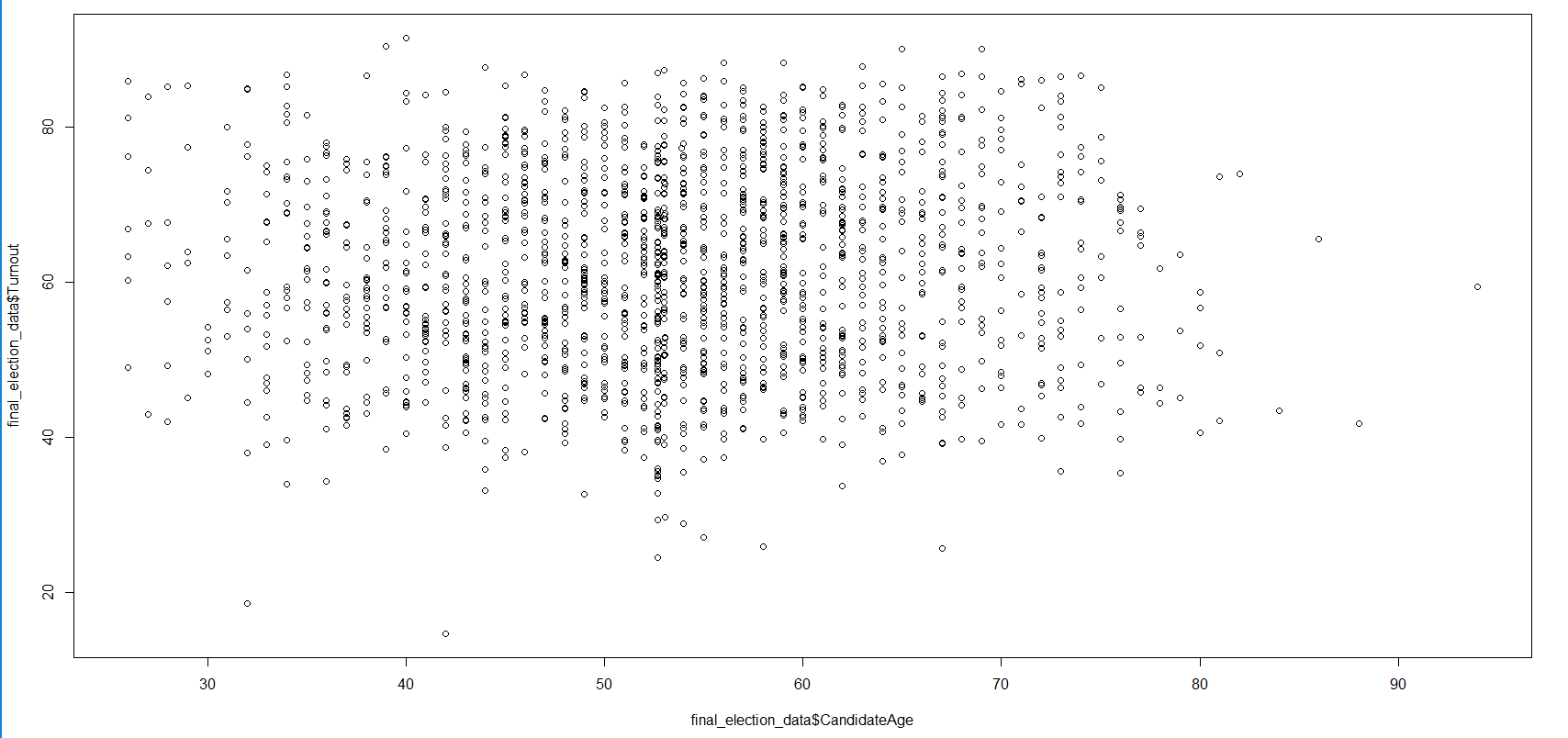
### 2014 and Total Seats for BJP, INC and Others

BJP with the highest number of seats



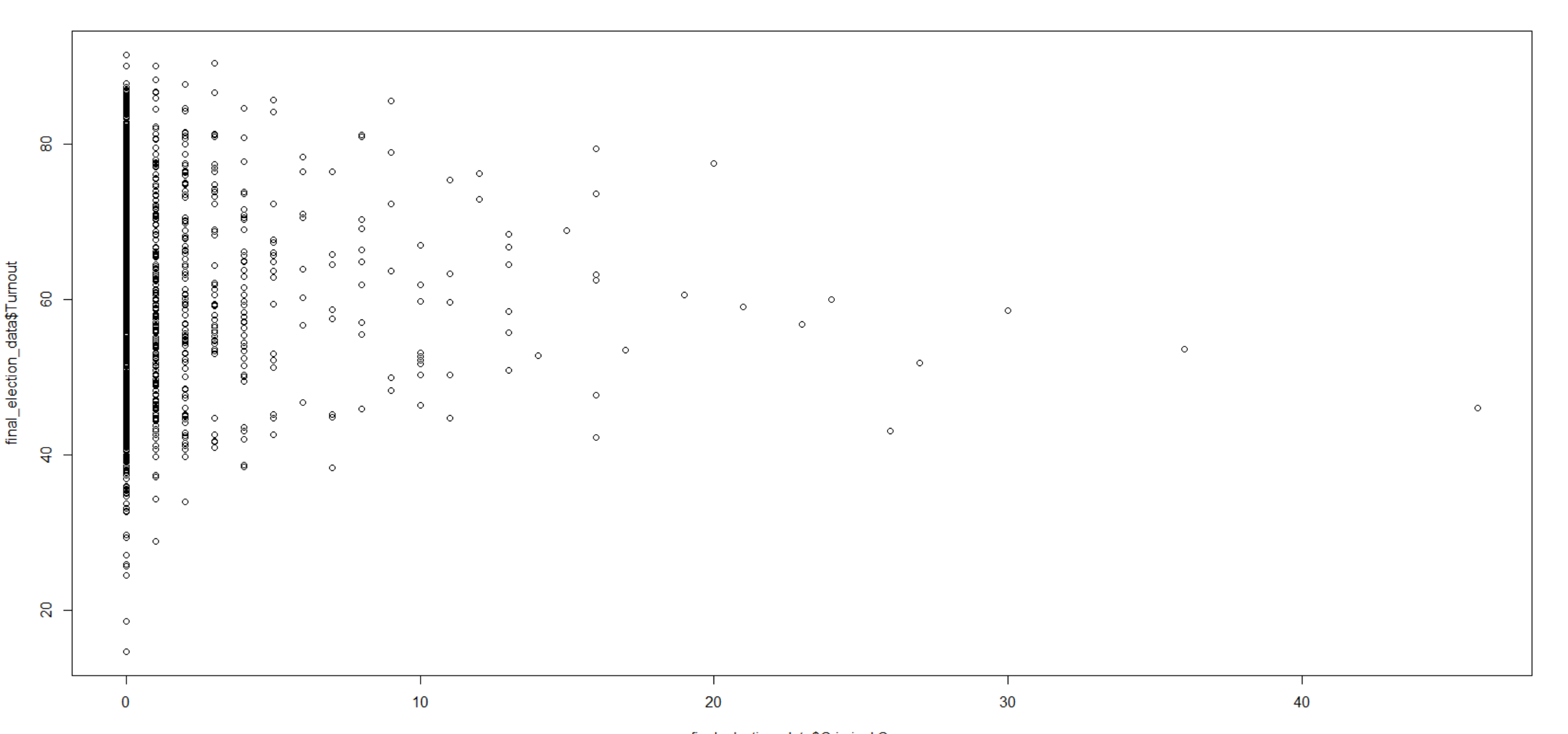
### Turnout versus Candidate Age

Datapoints are all scattered



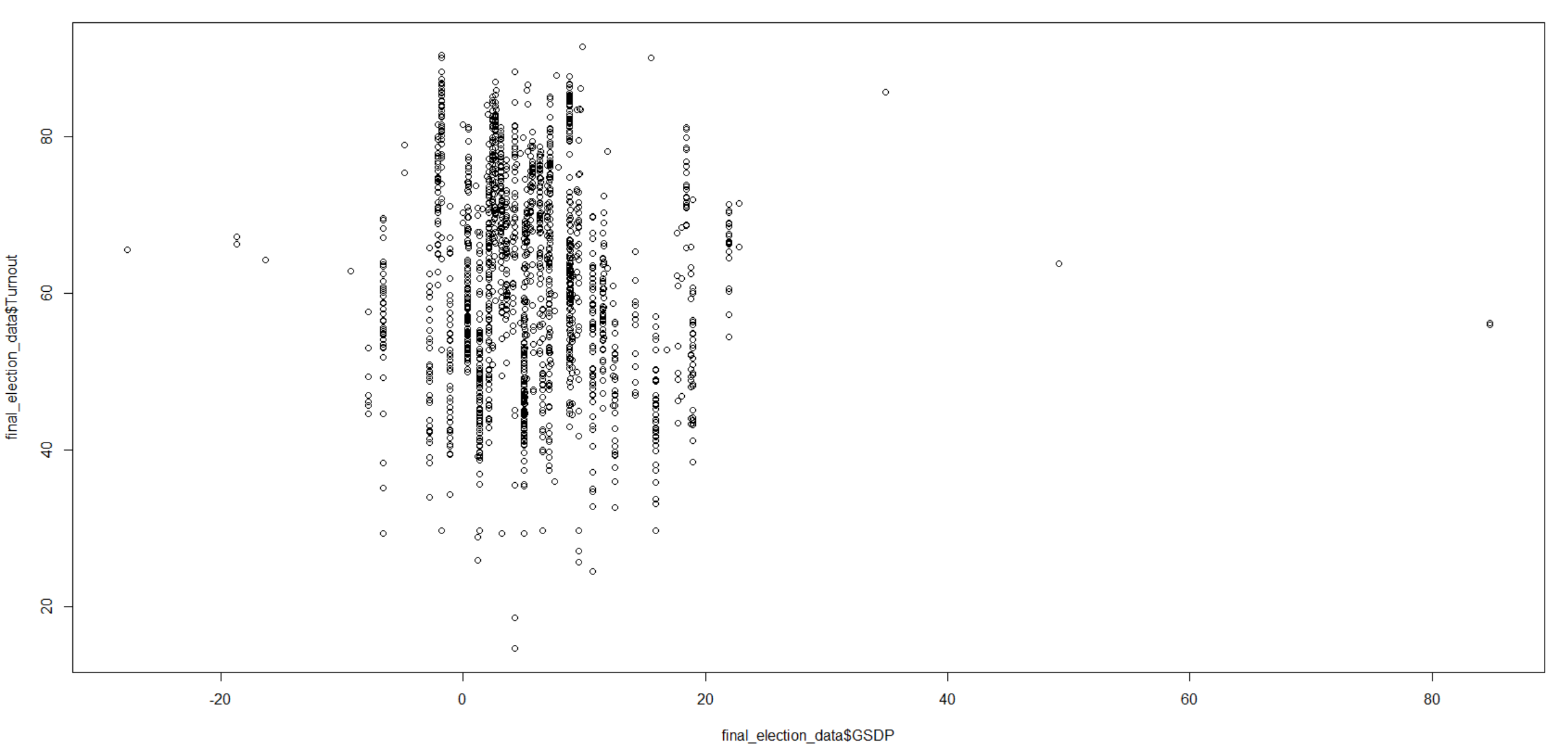
### Turnout versus Criminal Cases

With high criminal cases, the Turnout surely decreases



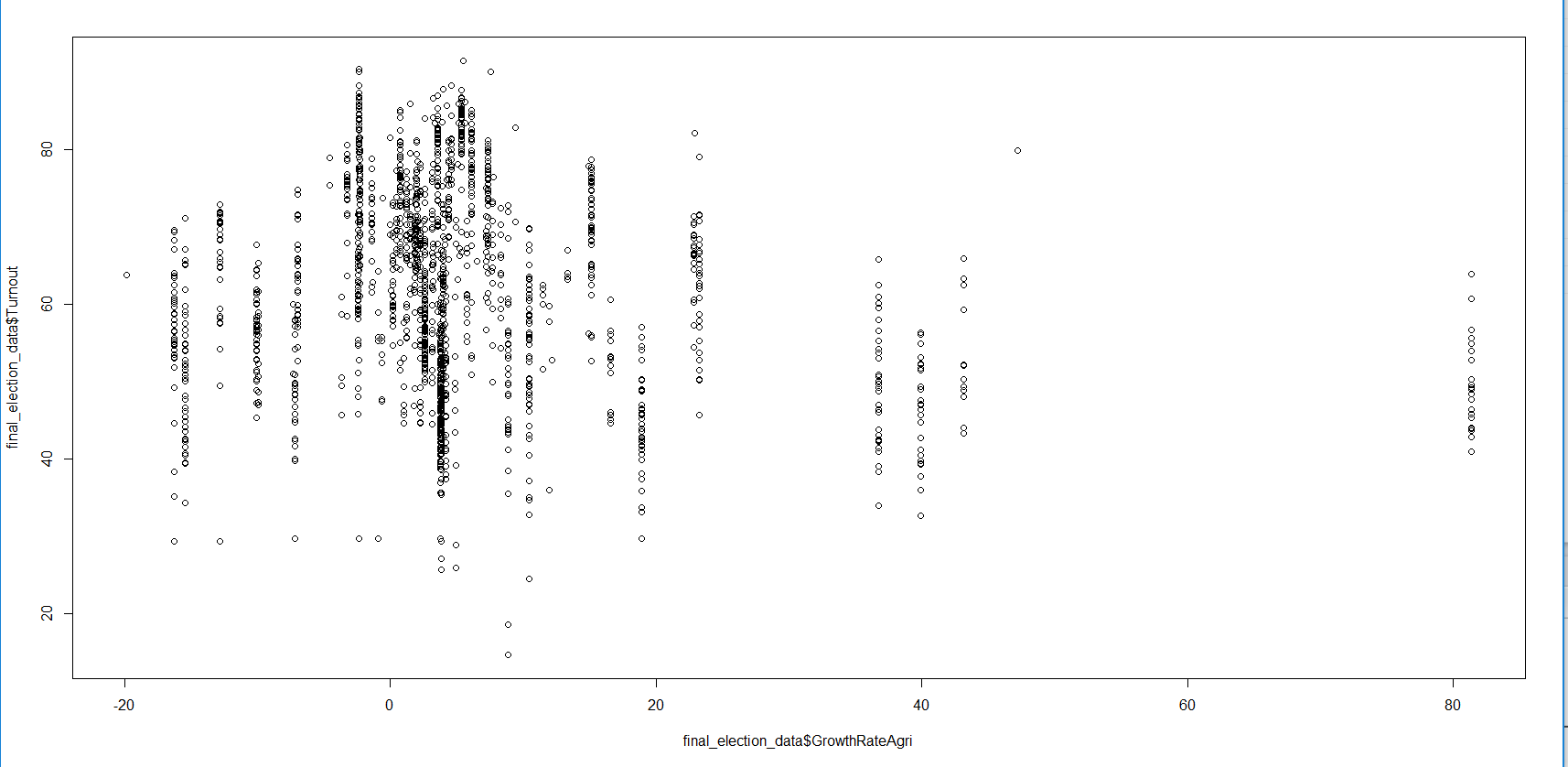
### Turnout versus GSDP

The data points are concentrated in one region

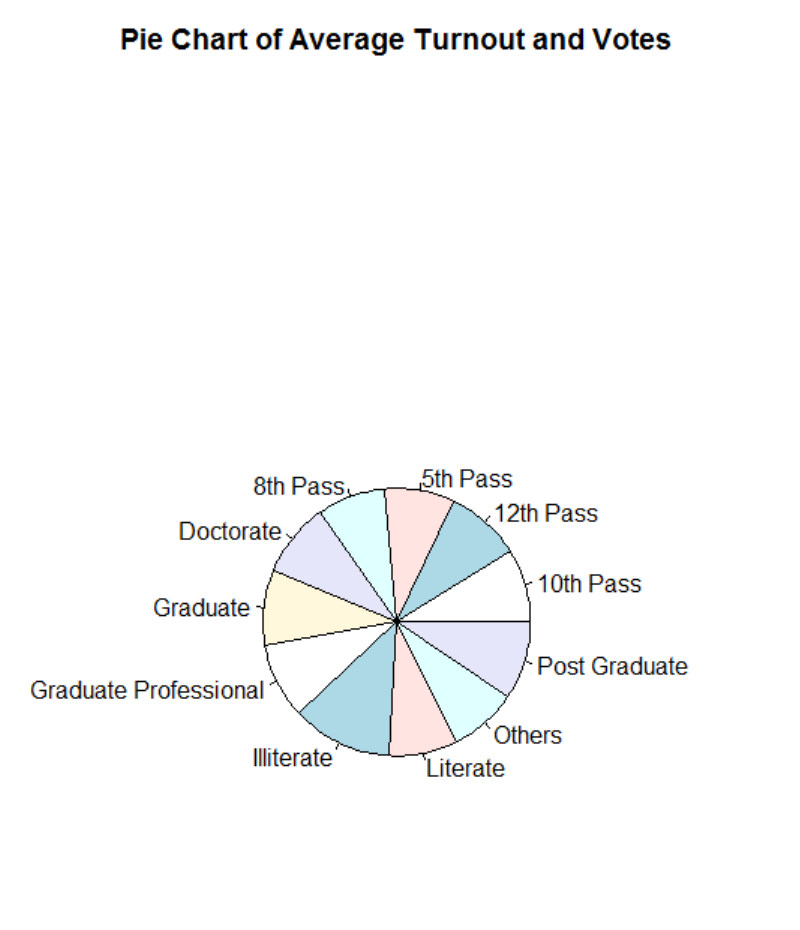


### Turnout versus GrowthRateAgri

Data points are all scattered



### Average Turnout across different Education Levels



# The Predictive Model

We wanted to build and train a model that will predict the winning Party based on the independent predictors as Type, Turnout, Gender, PrevParty, GSDP, Criminal.Case, Education, Total.Assets, Liabilities, CandidateAge, GrowthRateAgri

We wanted to build a very basic classification model with the Neural Network. Why Neural Network? We did not want to select model that assumes some underline distribution like Logistic Regression. Our focus was more on Accuracy then Interpretation.

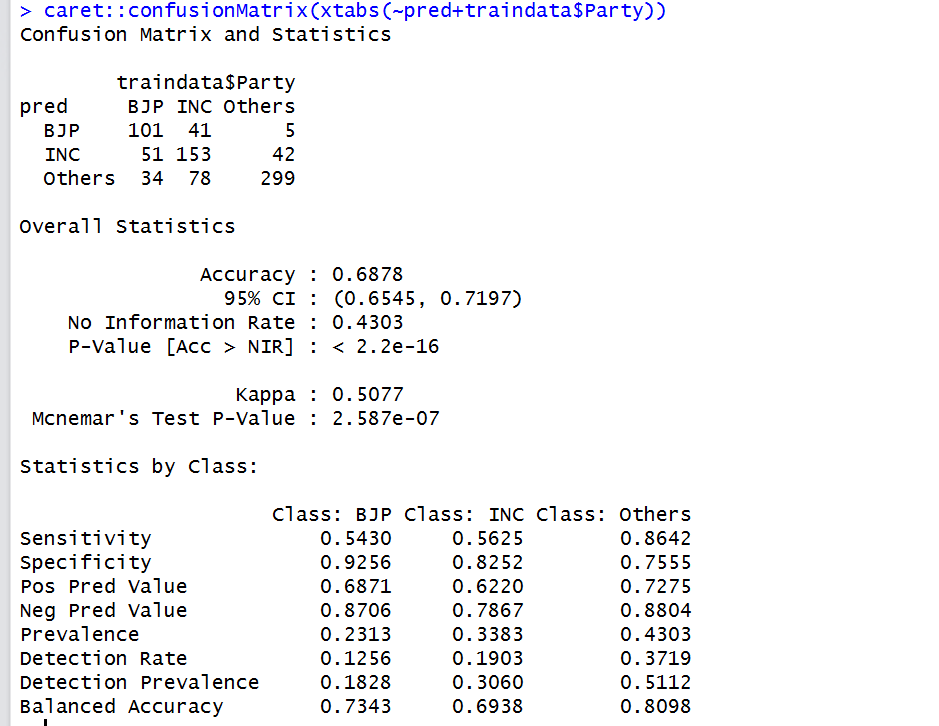
We did trial run with 1 hidden layer comprising of 3/5/10/15 neurons as well as 2 hidden layers with (5,3) / (3,5) / (10,5)/(5,10) architecture. In all these runs we found that 2 hidden layers with 10 and 5 neurons suffice our need.

# The Thought Process behind building the Basic Predictive Model for Classification

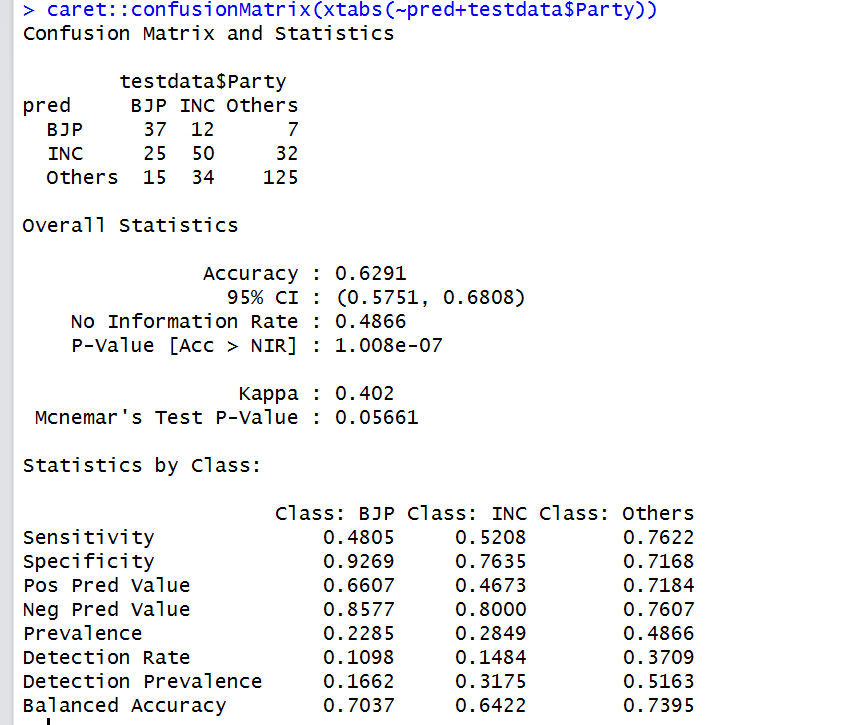
* The Party dependent variable will have three labels INC/BJP/Others. All the parties other than INC and BJP are clubbed in “Others”
* Will combine the 2004 and 2009 election data in order to train the model.
* Will split the data into 70% and 30%.
* 70% we will use for training the model
* 30% we will use for validating our model
* 2014 election data will be the unseen data for testing the model.
* Once we train and validate the model, we will run the model on the test data
* Will record the Accuracy and Sensitivity and Specificity at each stage of model building that is training/validating/testing.

# Model Prediction Results

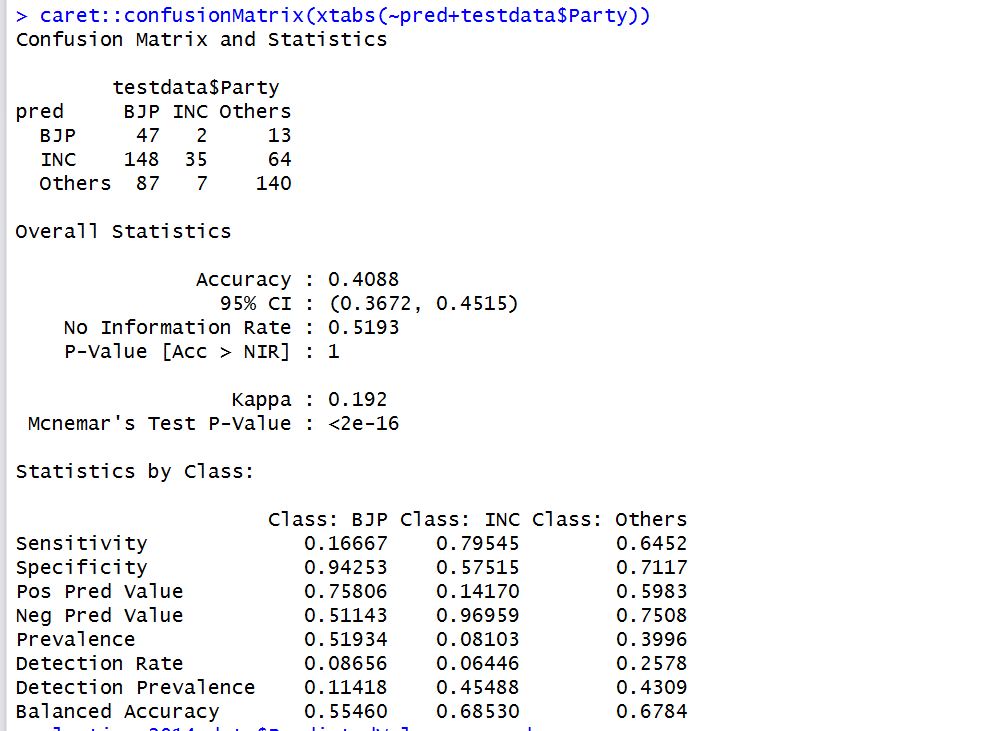
### Model Building Stage



### Model Validation Stage



### Model Testing Stage



# Conclusions

The model cannot be used as it is failing to predict the BJP label correctly. 2004 and 2009 was INC majority hence the model is trained and validated on that data. The 2014 election result was BJP majority hence the model fails to predict the label correctly for the BJP label….Can we improve upon the model?

# Next Steps….

* We can add more features to the model to improve the accuracy
* We can text mine the Party manifestos for BJP, INC to get the top key words in the year 2004, 2009 and 2014 to build a new feature for the model.
* We can build a new feature on the approval rating of the Prime Minister candidate for the year 2004, 2009 and 2014
* We can do twitter analysis to understand the sentiments on Party or Prime Minister candidates to do text mining and build new feature to add to our model. Getting twitter data for the past years not possible but we can do this going forward and keep storing the data to further improve the model in the future.
* We can use the opinion poll or exit poll to come with some useful feature that can improve the accuracy
* We can add religious wise population which will certainly be an important predictor that will influence the prediction.