SUMMER TRAINING REPORT

On

ELECTION ANALYTICA

Submitted to Guru Gobind Singh Indraprastha University, Delhi
(India)
in partial fulfilment of the requirement for the award of the degree of

B.TECH
in

COMPUTER SCIENCE AND ENGINEERING

Submitted By:-

PRANJAL RASTOGI

Roll. No:- 05215002716

DEPTT. OF COMPUTER SCIENCE AND ENGINEERING
MAHARAJA SURAJMAL INSTITUTE OF TECHNOLOGY,
NEW DELHI-110058

OCTOBER 2018

CONTENTS

ACKNOWLEDGEMENT

A research work owes its success from commencement to completion, to the people in love with researchers at various stages. Let me in this page express my gratitude to all those who helped us in various stage of this study. First, I would like to express my sincere gratitude indebtedness to **Dr. KOYEL DATTA** (HOD, Department of Computer Science And Engineering, Maharaja Surajmal Institute of Technology, New Delhi) for allowing me to undergo the summer internship of 60 days at **AAM AADMI PARTY.**

I am grateful to our guide **Mr. ARJUN TIWARI**, for the help provided in completion of the project, which was assigned to me. Without his friendly help and guidance it was difficult to develop this project.

Last but not least, I pay my sincere thanks and gratitude to all the Staff Members of **AAM AADMI PARTY** for their support and for making our training valuable and fruitful.

Submitted By:

Name- Pranjal Rastogi

Batch- CSE-2 (Morning)

Enrollment Number- 05215002716

CANDIDATE'S DECLARATION

I, **PRANJAL RASTOGI** Roll No **05215002716**, B.Tech (Semester- 5th) of the Maharaja Surajmal Institute of Technology, New Delhi hereby declare that the Internship Report entitled **"ELECTION ANALYTICA"** is an original work and data provided in the study is authentic to the best of my knowledge. This report has not been submitted to any other Institute for the award of any other degree.

PRANJAL RASTOGI

(05215002716)

Place: New Delhi
Date: 03/10/2018

1. Abstract

The project ‘Election Analytica’ is about analyzing previous Vidhan Sabha election results for particular constituencies for the state of Haryana, India. The analysis includes judging various parameters such as literacy rate, rural population percentage, employability rate etc with voting pattern for top paties/independent candidates for each constituency in each respective election and to compare them over parameters such as time period, election outcomes to conclude final results.

The analysis consists of various steps including:

- Collection of data from trusted sources,
- Wrangling of data in proper format
- Cleaning of data
- Visualisation of data over parameters
- Deriving conclusions and results

The visualisation part of analysis consists of comparing the data with various parameters that can affect the elections or vote percentage for a candidate. The parameters are compared by drawing plots such as line graphs, bar graphs, pie-charts etc. over time period. Finally after keen observation of plots etc, the conclusions were derived from the comparison..
The conclusion consists of determination of floating and fixed vote percentage, effect of caste bias and local leadership over election results for each constituency or polling stations. Also, analysis includes identification of top candidates/parties over the years hat may have significant impact on upcoming 2019 elections and analyzing the change in vote percentage for them, over the years with various parameters. Finally some constituencies were pinpointed where Aam Aadmi Party might have higher probability of winning without any existential voter’s base.

2. Organization Introduction

Aam Aadmi Party is an Indian political party, formally launched on 26 November 2012, and is currently the ruling party of the National Capital Territory of Delhi. The party was born by the anti-corruption movement that started in April 2011, in the backdrop of big ticket corruption scandals. A group of activists from varied backgrounds had come together to demand the enactment of the Jan Lokpal Bill. A Jan Lokpal was directly detrimental to the interests of the political parties and their leaders. Thus a section of activists decided to join politics and hence, the birth of Aam Aadmi Party.

The party made its electoral debut in the 2013 Delhi Legislative Assembly election, where it emerged as the second-largest party, winning 28 of the 70 seats and formed a minority government with conditional support from the Indian National Congress. A significant part of its agenda was to quickly introduce the Jan Lokpal bill in the National Capital Territory of Delhi. When it became clear after the election that the other major parties would not support this bill, the AAP government resigned after being in power for 49 days. In the 2015 Delhi Legislative Assembly election, AAP won 67 of the 70 seats in the assembly and formed a majority government.

Data Analytics Department

The AAP internship cell was founded in 2017, with data analytics department established in early May, 2018 with a view of fighting elections differently from years' old exhausted approach. The aim of department was to analyse the data related to elections and produce results in terms of facts consisting of how, when and where party should spend it's valuable assets to contest elections. The first task given to department is regarding upcoming Vidhan Sabha elections in state of Haryana, India. The task comprises of pinpointing the constituencies and provide an effective approach for each and every booth so as to maximise the impact as well as probability of winning that particular booth while exhausting minimal of assets and resources.

3. Election Analytica

The project comes under the banner of data analysis, in which the data regarding election results, census and other miscellaneous data is formatted, cleaned and visualised together to derive results/conclusions.

4. Data Analysis

Data analysis is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making.
It is the process of examining data sets in order to draw conclusions about the information they contain,. It consists of qualitative and quantitative techniques and processes used to enhance productivity. Data is extracted and categorized to identify and analyze behavioral data and patterns, and techniques vary according to organizational requirements.
The process consists of applying statistical practices to organize, represent, describe, evaluate, and interpret data. It helps to take informed decisions for data, improve efficiency, increase profits and achieve organisational goals. To do this effectively it is important to first clearly understand for what purpose one is conducting the analysis. The purpose of data analytics in simple word is make best out of waste.

4.1 Process

The process of Data Analysis refers to breaking a whole into its separate components for individual examination. The process consists of obtaining raw data and converting it into information useful for decision-making by users. Data is collected and analyzed to answer questions, test hypotheses or disprove theories.
There are several phases that can be distinguished, described below. The phases are iterative, in that feedback from later phases may result in additional work in earlier phases.

4.1.1 Data requirements

The data is necessary as inputs to the analysis, which is specified based upon the requirements of those directing the analysis or customers. The general type of entity upon which the data will be collected is referred to as an experimental unit (e.g., a person or population of people or related activity). Specific variables regarding a population (e.g., age and economic status, caste) may be specified and obtained. Data so obtained and analysed may be numerical or categorical.

The data required for the project includes

- Outcomes of previous vidhan sabha elections,
- Population statistics such as literacy rate, rural population percentage, employment rate
- Vote percentage per polling booth
- Information or history about Major parties/candidates

4.1.2 Data collection

Data is collected from a variety of sources. The data can be collected from on-ground surveys or can also be collected from other trusted sources conducting such experiments. The data may also be collected from sensors in the environment, such as traffic cameras, satellites, recording devices, etc. It may also be obtained through interviews, downloads from online sources, or reading documentation.

The Data is collected from various trusted sources such as

- Census report for 1991, 2001, 2011
- Information from Chief election officer of Haryana,
- Reports from website of Election Commision of India
- Information about constituency and candidates from various miscellaneous resources

4.1.3 Data wrangling

Data initially obtained must be processed or organised for analysis. It is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate for analysis. For instance, these may involve placing data into rows and columns in a table format (i.e., structured data) for further analysis, such as within a spreadsheet or statistical software.

The data wrangling for project was done with help of various tools and APIs including Adobe Acrobat, ‘pdf tables api’, and other conversion tools. Manual data entry was also done to fill some separate information.

4.1.4 Data cleaning

Once processed and organised, the data may be incomplete, contain duplicates, or contain errors. The need for data cleaning will arise from problems in the way that data is entered and stored or is in form after wrangling.

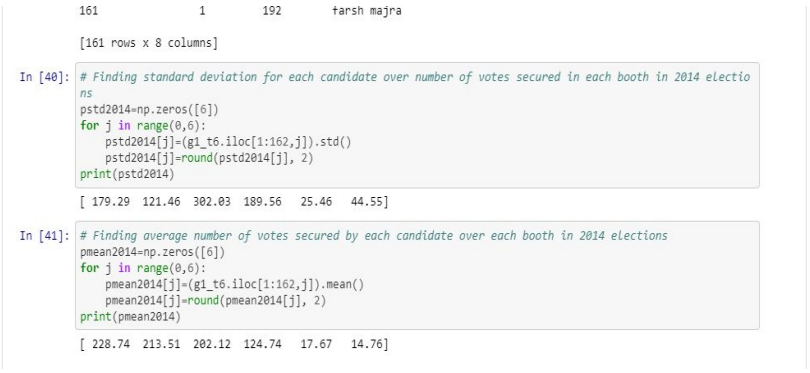
Data cleaning is the process of preventing and correcting these errors. Common tasks include record matching, identifying inaccuracy of data, overall quality of existing data, deduplication, and column segmentation. Such data problems can also be identified through a variety of analytical techniques, such as textual data spell checkers can be used to lessen the amount of mistyped words. There are several types of data cleaning that depend on the type of data such as filling of incomplete information or correcting the type of information such as, NaN quantity can occur many times while wrangling of data and can create problem in analysis. The problem can be solved by writing a simple script which can replace NaN values with a default value.

```
import pandas as pd
import numpy as np

# for column
df['column'] = df['column'].replace(np.nan, 0)

# for whole dataframe
df = df.replace(np.nan, 0)

# inplace
df.replace(np.nan, 0, inplace=True)
```

The script written above replaces the nan values with the mentioned value.

4.1.5 Exploratory data analysis

Once the data is cleaned, it can be analyzed. Analysts may apply a variety of techniques referred to as exploratory data analysis to begin understanding the messages contained in the data. The process of exploration may result in additional data cleaning or additional requests for data, so these activities may be iterative in nature. Descriptive statistics, such as the average or median, may be generated to help understand the data.



```
161              1        192       tarsh majra

[161 rows x 8 columns]

In [40]: # Finding standard deviation for each candidate over number of votes secured in each booth in 2014 electio
         ns
         pstd2014=np.zeros([6])
         for j in range(0,6):
             pstd2014[j]=(g1_t6.iloc[1:162,j]).std()
             pstd2014[j]=round(pstd2014[j], 2)
         print(pstd2014)

         [ 179.29  121.46  302.03  189.56   25.46   44.55]

In [41]: # Finding average number of votes secured by each candidate over each booth in 2014 elections
         pmean2014=np.zeros([6])
         for j in range(0,6):
             pmean2014[j]=(g1_t6.iloc[1:162,j]).mean()
             pmean2014[j]=round(pmean2014[j], 2)
         print(pmean2014)

         [ 228.74  213.51  202.12  124.74   17.67   14.76]
```

Figure 4 (i)

The code snippet above is used to calculate mean and standard deviation of a set of values using functions such as mean(), std().

4.1.6 Data visualization

Data visualization comes under exploratory data analysis and is used to examine the data in graphical format via plots, to obtain additional insight regarding the messages within the data

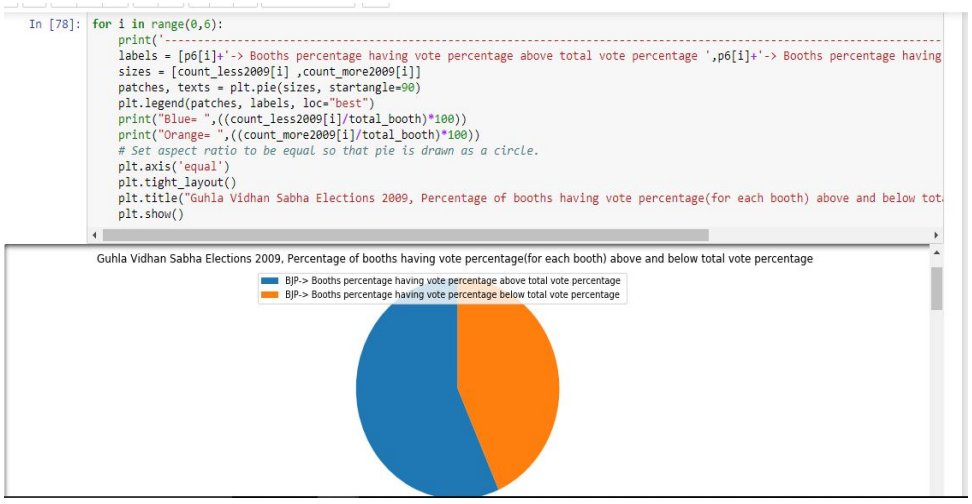


Figure 4(ii)

The above plot is Pie-Chart denoting the percentage of polling booths where BJP have above and below average number of votes per booth.

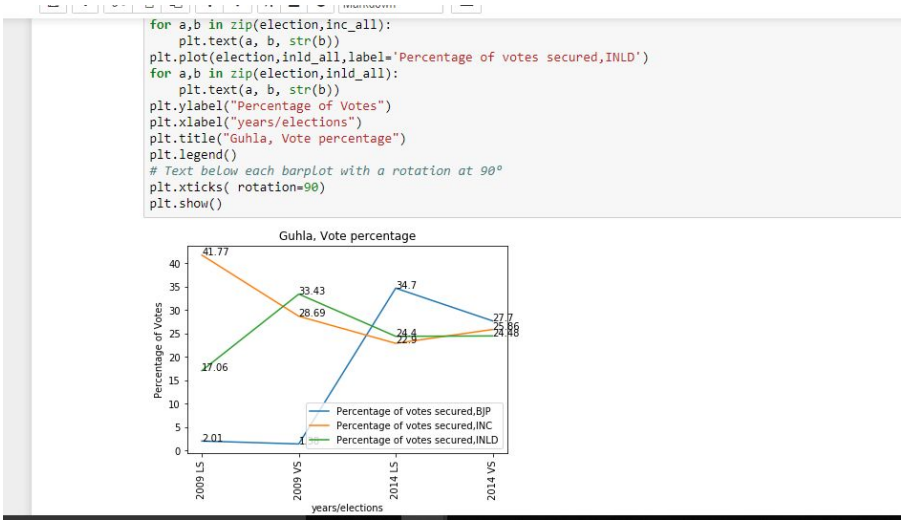


Figure 4(iii)

The above graph plot is comparing the Lok Sabha vote percentage of major parties including BJP, INC and INLD with vidhan sabha election results in Haryana for past 2 elections. It can be concluded easily that, INLD hold a very important position in Vidhan Sabha elections but are not a major threat to BJP, INC in Lok Sabha Elections where both parties share a big percentage of votes, about 25% each.

4.1.7 Modeling and algorithms

Mathematical formulas or models called algorithms may be applied to the data to identify relationships among the variables, such as correlation or causation. In general terms, models may be developed to evaluate a particular variable in the data based on other variables in the data, with some residual error depending on model accuracy i.e., we can infer that Data = Model + Error.

Inferential statistics includes techniques to measure relationships between particular variables. For example, regression analysis may be used to model whether a change in independent variable X explains the variation in dependent variable Y.

It may be described as Y = aX + b + error, where the model is designed such that a and b minimize the error when the model predicts Y for a given range of values of X.

```
In [113]: y_predict=lm.predict(x_train)
          y_predict

Out[113]: array([[ 23.17599057],
                 [ 34.58778302],
                 [ 43.71721698],
                 [ 55.12900943]])

In [114]: lm.score(x_train,y_train)

Out[114]: 0.98657056232454832
```

This is the R^2 score of our model. This is the percentage of variance of the predictions.

```
In [115]: lm.intercept_

Out[115]: array([ 4619.84599057])

In [116]: lm.coef_

Out[116]: array([[-2.28235849]])

In [117]: x_year=x_train
          x_year[0]=2019
          for i in range(1,4):
              x_year[i]=(x_year[i-1])+5
          x_year

Out[117]: array([[2019],
                 [2024],
                 [2029],
                 [2034]], dtype=int64)

In [118]: inld_train=lm.predict(x_year)
          inld_train

Out[118]: array([[ 11.76419811],
```

Figure 4(iv)

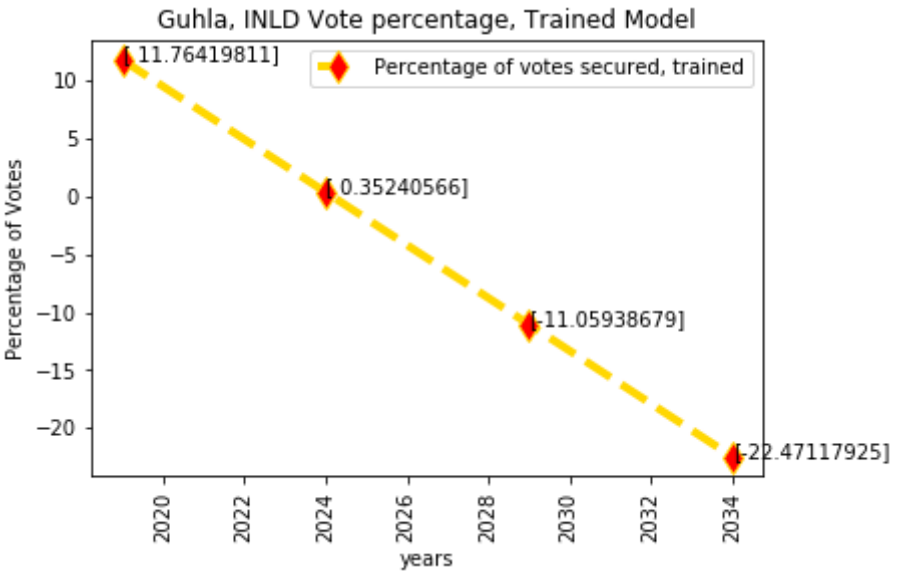


Figure 4(v)

In above picture linear regression is applied, but the prediction can not be accurately made because of lack of data points which algorithm can use to model the trend in data and make prediction.

5. Technologies Used

5.1 Language: Python

Python is an interpreted high-level programming language for general-purpose programming. Released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace. It provides constructs that enable clear programming on both small and large scales.

Python features a dynamic type system i.e., verifying the type safety of a program at runtime and automatic memory management. It supports multiple programming paradigms, including object-oriented, functional and procedural, and has a large and comprehensive standard library.

5.2 Tools:

- Anaconda- Anaconda is a free and open source distribution of the Python and R programming languages for data science and machine learning related applications, that aims to simplify package management and deployment. Package versions are managed by the package management system conda. It consists of a desktop graphical user interface (GUI) included in Anaconda distribution that allows users to launch applications and manage conda packages, environments and channels without using command-line commands. Various applications are available by default in Navigator
- Jupyter Notebook: It is a web-based interactive computational environment for creating Jupyter notebooks documents.Jupyter Notebook can connect to many kernels. A Jupyter kernel is a program responsible for handling various types of request (code execution, code completions, inspection), and providing a reply.
- IDLE: IDLE (short for integrated development environment or integrated development and learning environment) is an integrated development environment for Python. IDLE is intended to be a simple IDE and suitable for beginners, especially in an educational environment. To that end, it is cross-platform, and avoids feature clutter.
- Adobe Acrobat: It consists of a family of application software and Web services developed by Adobe Systems to view, create, manipulate, print and manage files in Portable Document Format. The tool was used in data wrangling step to convert PDF files into more appropriate CSV/Excel files.

5.3 Libraries:

- Pandas: It is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. The library supports features such as :
  DataFrame object for data manipulation with integrated indexing,
  Tools for reading and writing data between in-memory data structures and different file formats,
  Data alignment and integrated handling of missing data,
  Reshaping and pivoting of data sets, label-based slicing, fancy indexing,
  Data structure column insertion and deletion.
- NumPy: NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. The core functionality of NumPy is its "ndarray", for *n*-dimensional array, data structure. But it has a limitation as inserting or appending entries to an array is not as trivially possible as it is with Python lists.
- Matplotlib: Matplotlib is a Python 2D plotting library for the Python programming language and its numerical mathematics extension NumPy. It produces publication quality figures in a variety of hard copy formats and interactive environments across platforms.

```
In [66]: barWidth = 0.9
plt.bar(p6,boothwin_perc2014 , width = barWidth, label='Percentage of booths won')
for a,b in zip(p6,boothwin_perc2014):
    plt.text(a, b, str(b))
plt.ylabel("Percentage of Booths Won")
plt.xlabel("Parties/Candidates")
plt.title("Guhla 2014 Elections, Booths winning percentage")
plt.legend()
# Text below each barplot with a rotation at 90°
plt.xticks( rotation=90)
plt.show()
```

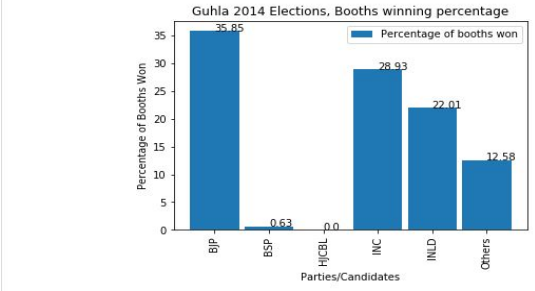


Figure 5(i)

Using matplotlib to generate a bar graph from the data

Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter notebook, web application servers, and graphical user interface toolkits. Various types of plots that can be generated using this library. A few of them are, line graphs, bar graphs, scatter plots can be made using this library.

- Scikit-learn: Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms.

5.4 API: It stands for application programming interface, and is a set of subroutine definitions, communication protocols, and tools for building software. In general terms, it is a set of clearly defined methods of communication among various components.

- Pdf-tables api: It's a simple web based API, so can be called from any programming language. For the project, the api was used in data wrangling process to successfully converting tables in pdf format to required format

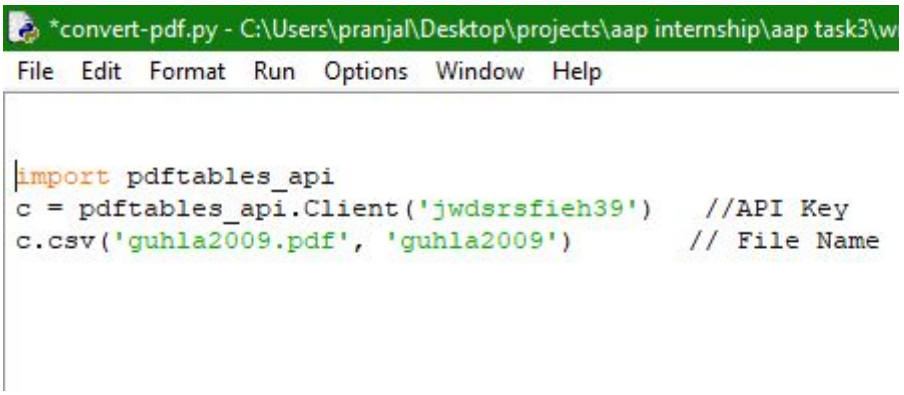


Figure 5(ii)

The above screenshot shows the code for using the API to convert pdf files into required format. Only a valid key is required to generate the file in required format. The API is versatile and can be used to convert PDF documents into appropriate format to be used in various languages like C++, Java, PHP, R etc.

6. Project Analysis

The project consists of various graphs plotted for better understanding of topic and driving results. The analysis was done for 3 Vidhan Sabha constituencies in Haryana namely, 'Kaithal', 'Guhla', 'Kalayat' in 'Kurukshetra' Lok Sabha Constituency. A few of the plots, along with derived conclusions are presented as follows:-

- Elections were held just after Lok Sabha Elections in which BJP won clearly under leadership of Shri Narendra Modi. It's effect can be seen clearly in these elections as BJP clinched the victory. Even Though INLD and INC have strong roots here and managed to get substantial number of votes, but the effect of Lok Sabha elections helped BJP to clinch victory. The Margin was not so much with a mere difference of 4000 votes among top 3 parties.

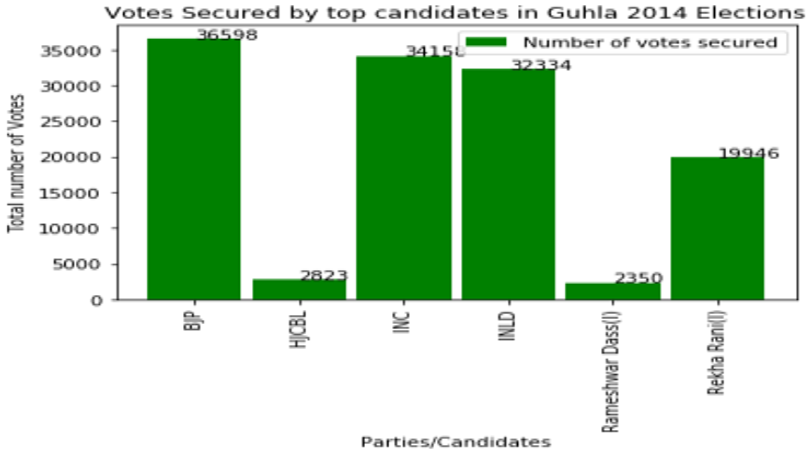


Figure 6(i)

- Elections were held just after Lok Sabha Elections,2009 in which INC formed alliance and restored power. INLD and INC have strong roots and voters base here and managed to get substantial number of votes. More than 50% votes have been clinched by these 2 parties. Others also got many votes showing that no other candidate has strong influence here. The Margin was not so much with a mere difference of 5000 votes between top 2 parties.

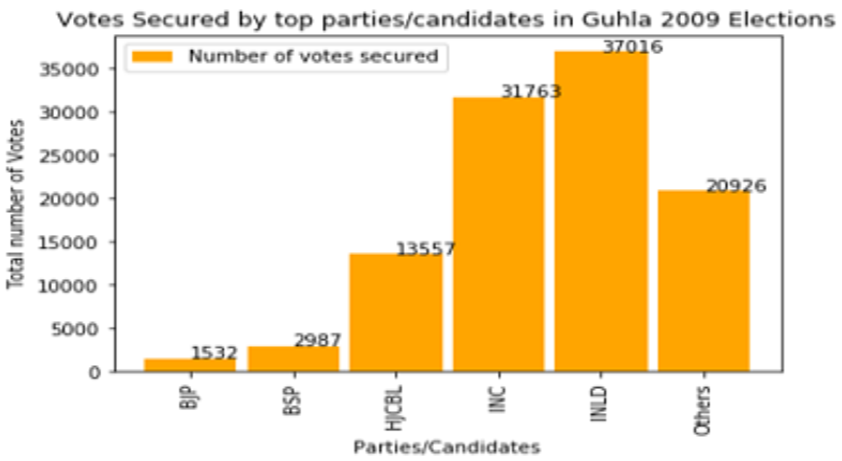


Figure 6(ii)

- Elections were held just after Lok Sabha Elections. Bhai Jai Prakash won 2014 elections but did not contested 2009 elections showing his strong leadership. INLD and INC have strong voter's base and managed to get substantial number of votes, which account for more than 50%. No other candidate was able to get substantial votes.

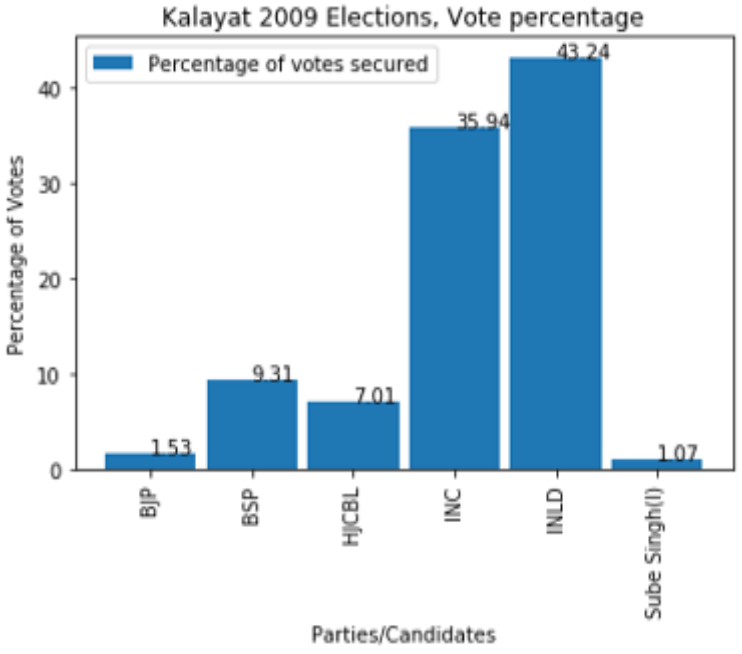


Figure 6(iii)

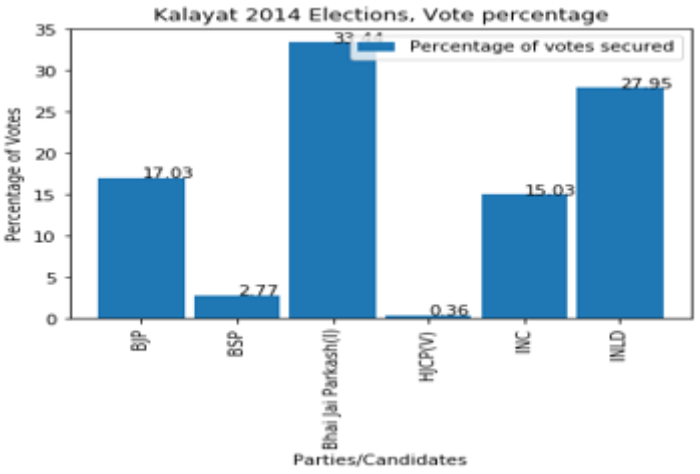


Figure 6(iv)

- Elections were held just after Lok Sabha Elections. INC won 2005 elections and INLD won 2000 elections showing their strong leadership .INLD and INC have strong roots here and managed to get substantial number of votes, which account for more than 50%. No other candidate was able to get substantial votes. But others got high number of vote percentage in 2000, showing that besides INLD no other party had strong base there.

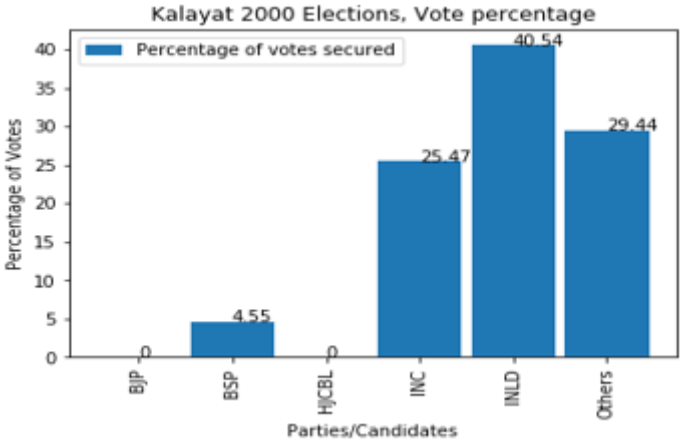


Figure 6 (v)

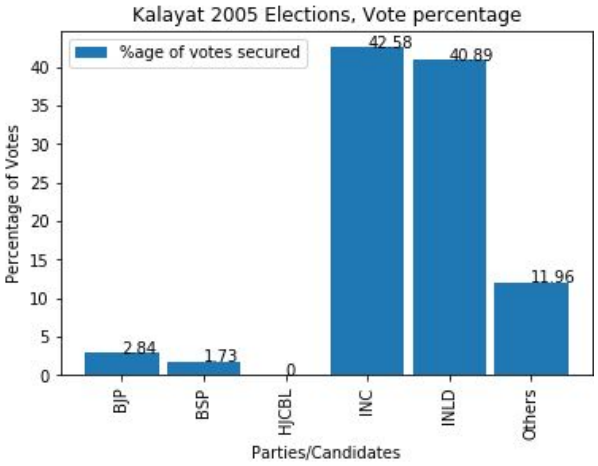


Figure 6 (vi)

- BJP vote percentage pattern shows that BJP didn't have strong influence their but in 2014 because of strong wave for BJP among people, they managed to secure 17% of votes. BSP on other hand has similar pattern securing between 1-4 %votes with an exception in 2009

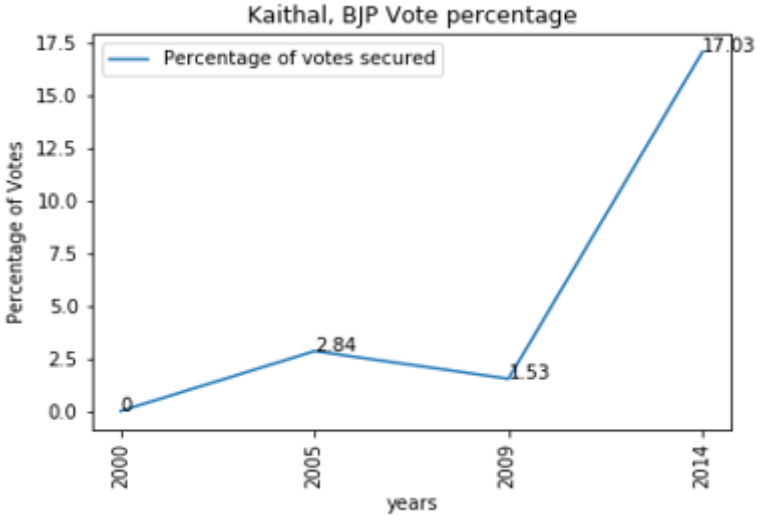


Figure 6 (vii)

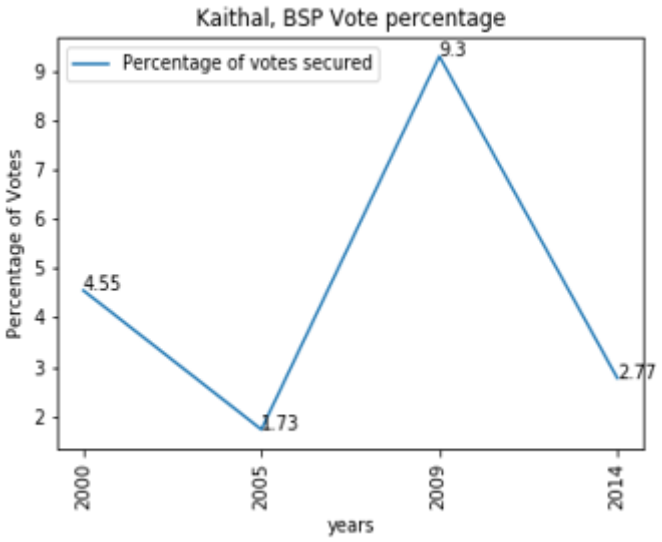


Figure 6 (viii)

- A low standard deviation indicates that the data points tend to be close to the mean of the set, while a high standard deviation indicates that the data points are spread out over a wider range of values. INLD in 2014 has an exceptionally high standard deviation while BSP has low standard deviation comparatively. Rest have intermediate standard deviation in between 100-150.

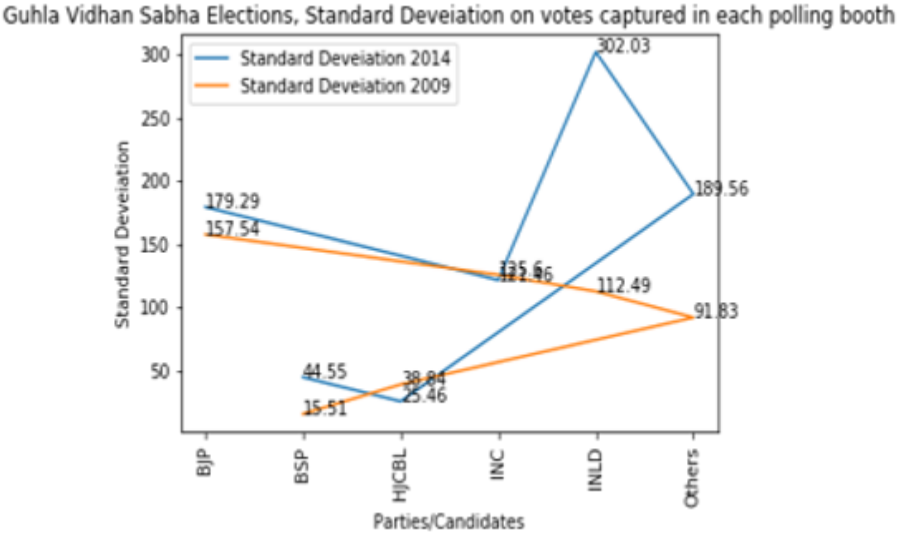
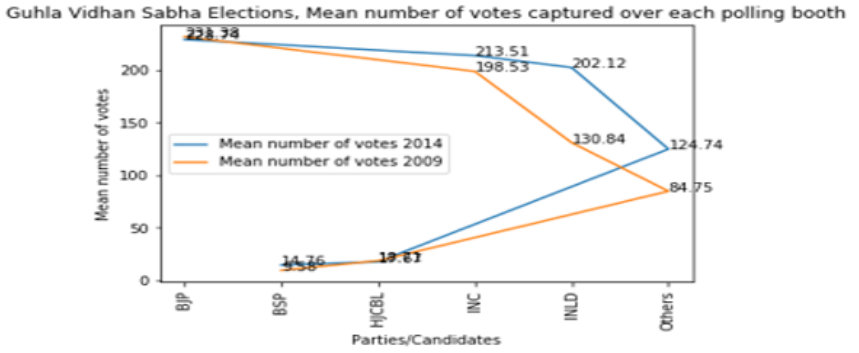


Figure 6 (ix)

- The plot shows the mean number of votes captured at each polling booth by major parties. The approach should be made to win each and every polling booth and this analysis will help to point out booths where more work is required.



\

Figure 6 (x)

- BJP and INC in 2014 have an exceptionally high number of booths above mean votes. Others and HJCBL also secured 34.59% of votes above mean. This shows that analysis is similar to votes won by each party . While BSP got lowest number of booths, equal to 11.95%.

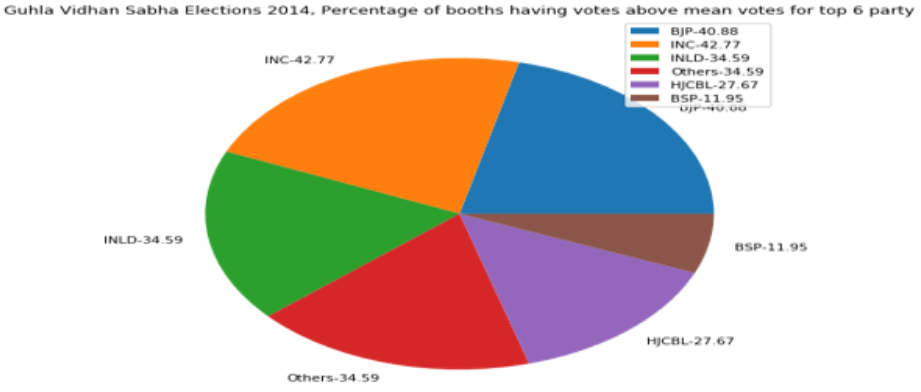


Figure 6 (xi)

- INLD and INC in 2014 have an exceptionally high number of booths above mean votes. Bhai Jai Prakash leads with 46.84% of booths. BJP and HJCBL also secured around 36% of votes above mean. This shows that analysis is similar to votes won by each party . While BSP got lowest number of booths, equal to 11%.

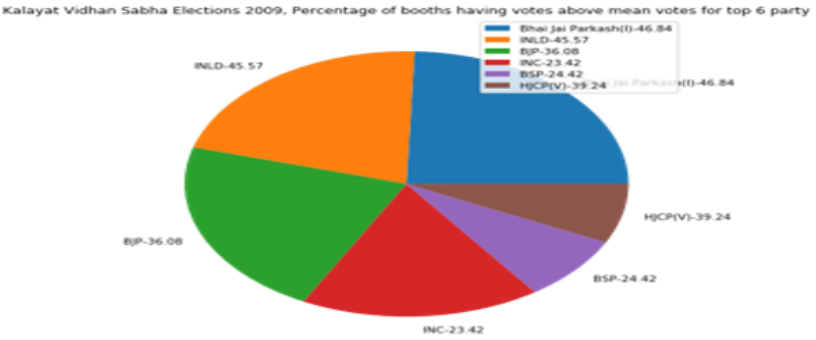


Figure 6 (xii)

- BJP leads and INC is on second each year showing their fixed voter base in the constituency. The number varies a lot for INLD and BSP showing their floating voters are substantial in numbers.

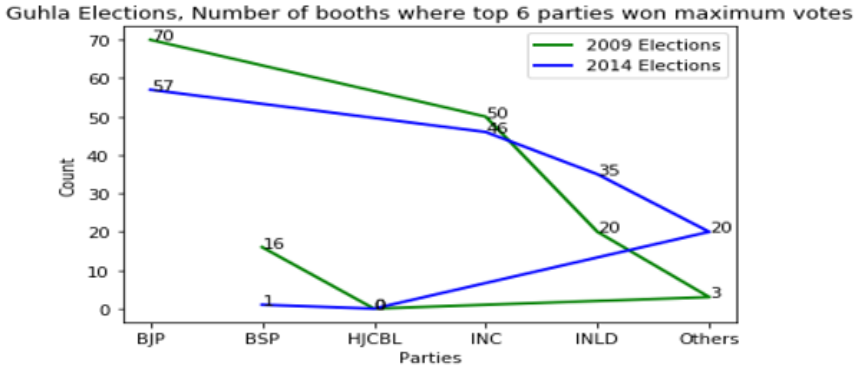


Figure 6 (xiii)

INC won highest number of booths as well as their booth win percentage is very high but INLD even though got large vote percentage could no secure lots of booths in terms of winning. BJP also got more vote percentage but less booth win percentage. This shows that INC has strong voter's base.

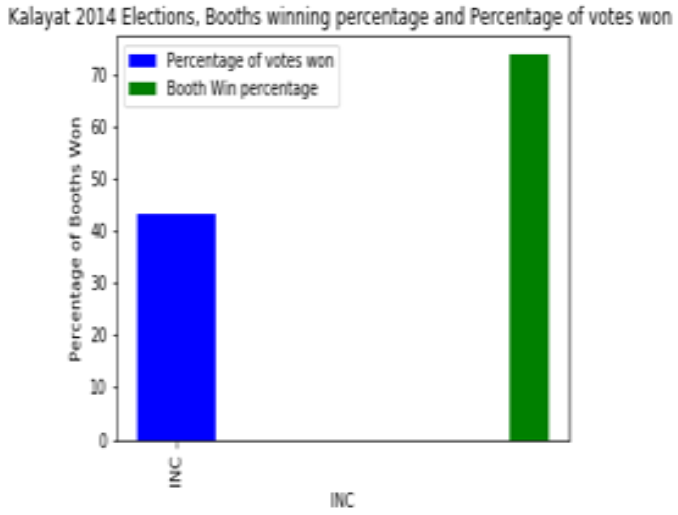


Figure 6 (xiv)

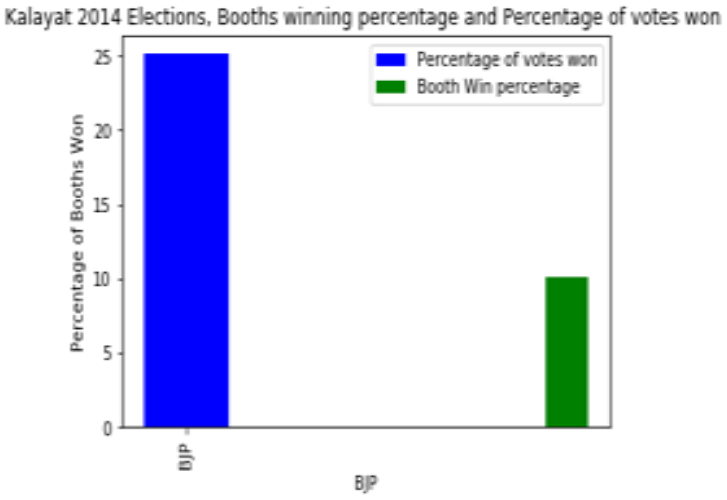


Figure 6 (xv)

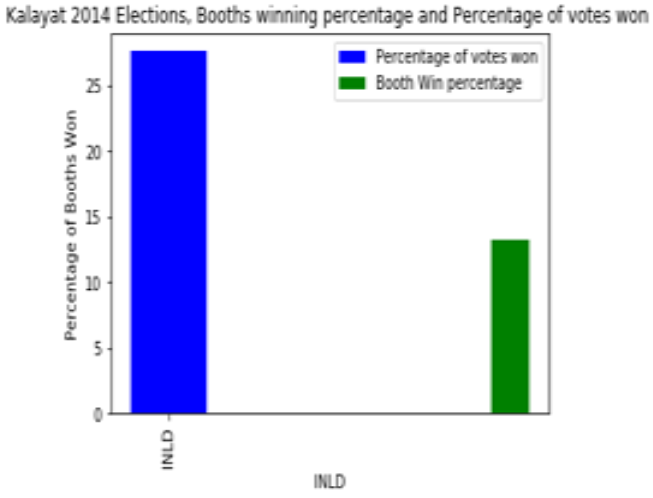


Figure 6 (xvi)

- Bhai Jai Prakash won the elections which can be inferred from this graph as he has more than 50 % of booths in which he secured votes above his average votes per booth. But the number of booths in which he secured votes below his average votes per booth is also very high

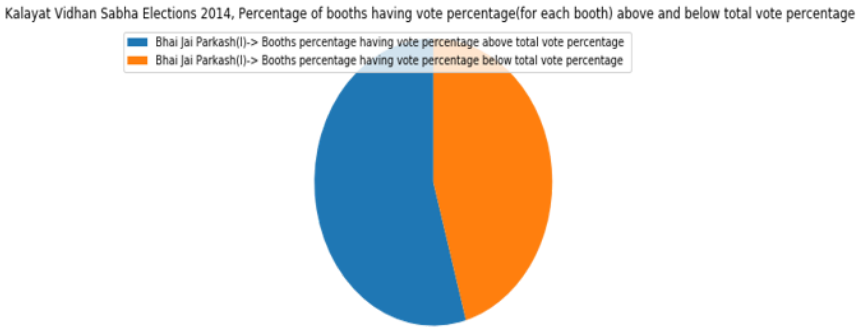


Figure 6 (xvii)

- Most of the parties have won 50 % of votes in about 16 booths showing consistency in how people have voted. The only difference is in 10-20% of votes which can make parties win or lose.

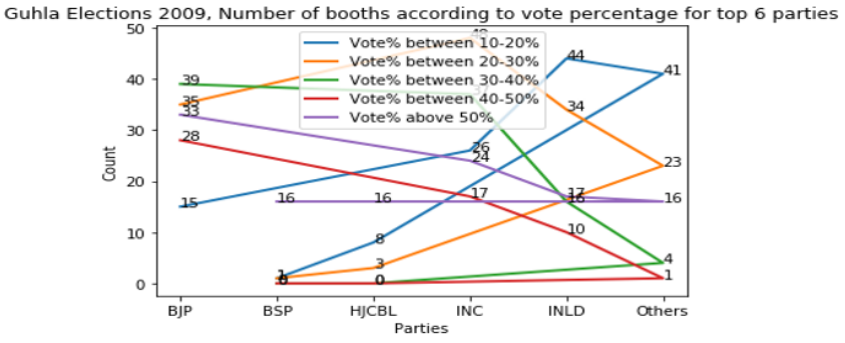


Figure 6 (xviii)

- The vote percentage is becoming clustered as years increases, showing that now parties cannot secure votes on old ways of caste bias and leader's impression so easily. The effect can be seen clearly as literacy rate is increasing over the years and people are becoming more aware and now vote with their clear understanding.

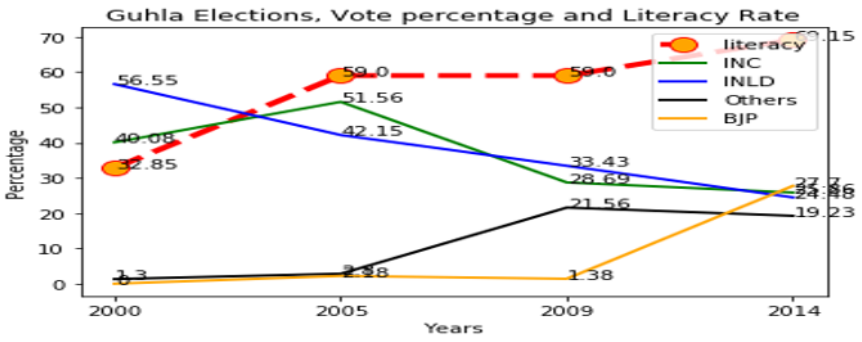


Figure 6 (xix)

- The vote percentage is becoming clustered as years increases, but there is negligible effect on percentage Rural Population.

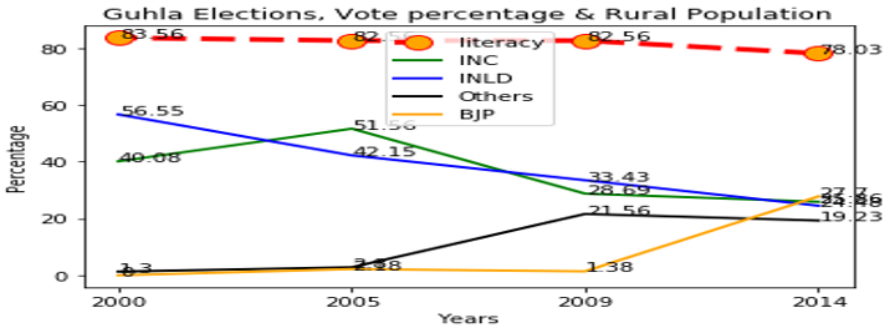


Figure 6 (xx)

- As can be analysed from screenshot, the vote percentage for major parties is becoming identical as time is increasing, but there is very little effect on employment rate showing that the scope of employment opportunities has not changed much in the constituency being analysed, over the years.

  Thus, it can be concluded that the population is voting for alternatives in search of better employment opportunities and hence the vote percentage for parties have become identical over the years.

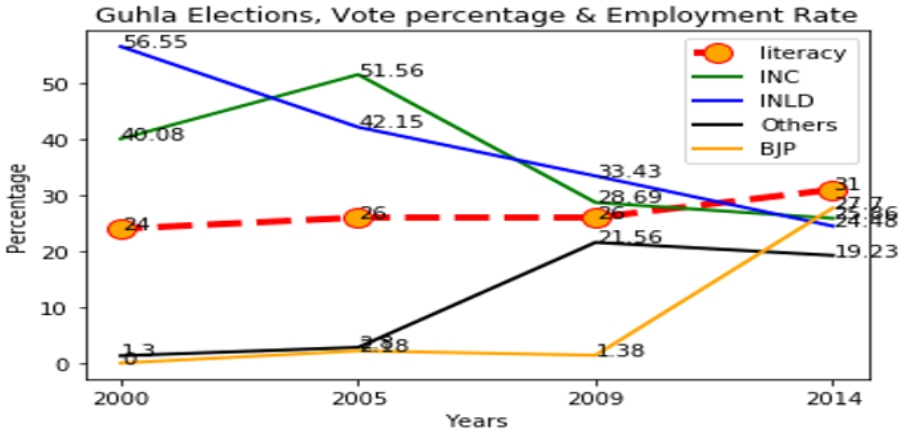


Figure 6 (xxi)

- The analysis shows that now major parties get similar percentage of votes instead of one party getting majority of votes. This can be understood as a by-product of increasing literacy rate and employment.

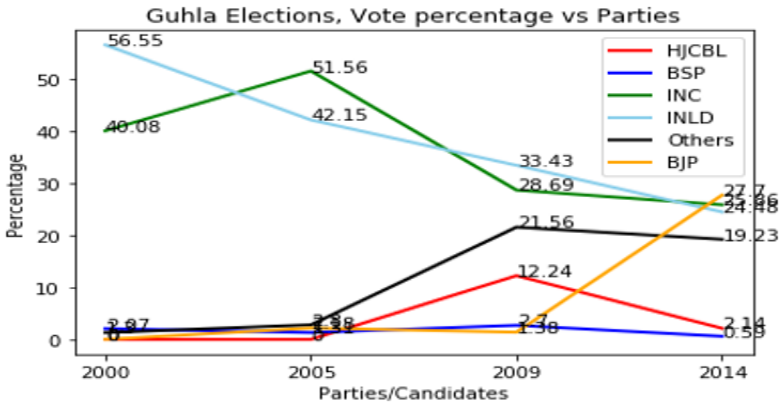


Figure 6 (xxii)

Conclusion

The results were concluded with help of the visual plots for various parameters and prediction made. For the 3 constituencies, we can conclude

- Each had a floating vote share of more than 30%, (marked with green color). It means that there is lack of very strong voter's base and voting patterns can elections can be influenced easily using a correct approach and targeting a correct set of population.

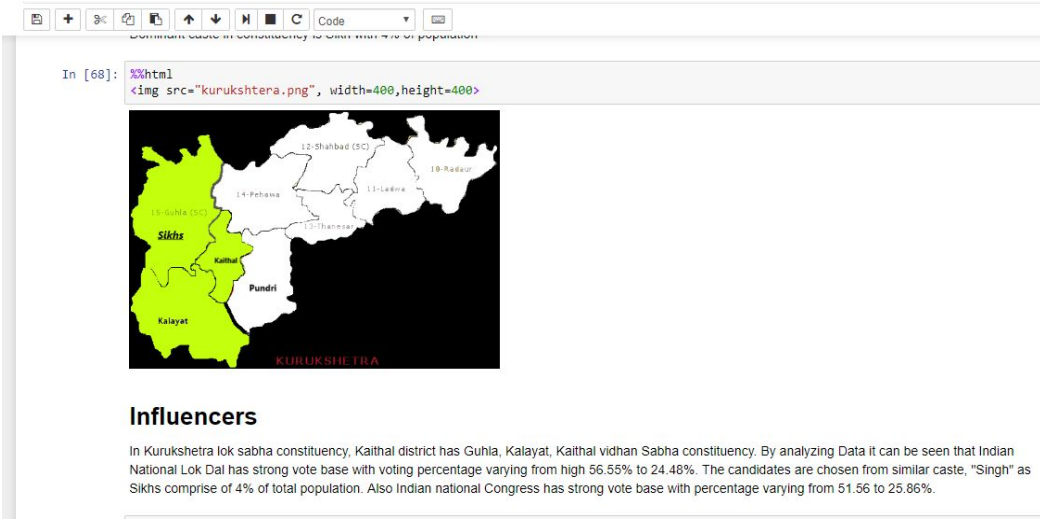


Figure 7 (i)

- It can be seen that INLD (Indian National Lok Dal) has a great influence in voting percentage in these 3 constituencies, and thus can be said to have a strong influencer/leader.
- With rise in Literacy rate over the years, as can be seen in screenshot below, the voting percentage has become more and more similar for each party, as compared to in previous times, when a certain party enjoyed a higher vote percentage because of old and existing voter base. The plot shows that majority of votes are no more cast on basis caste and economic status completely, but literate population now cast its vote on basis of logical thinking as per their choice.

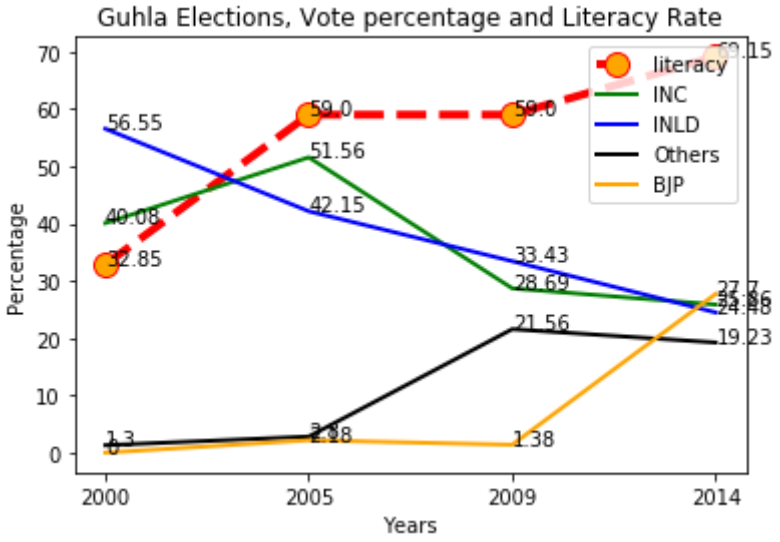


Figure 7(ii)

- If people are targeted with correct agendas, and alliances are formed carefully Aam Aadmi Party has a very good probability of winning this area. But still, party has to keep in mind, the caste factor, which still plays a major role in determining the voting pattern and caste of candidate contesting election as well. The candidate should be from similar caste, ('Singh') that forms the majority of population, as will be followed by all other parties.

Bibliography

1. ***Statistical Report on General Election , 2014 to the Legislative Assembly***

2. ***Statistical Report on General Election , 2009 to the Legislative Assembly***

3. ***Statistical Report on Elections 2004***

4. ***Statistical Report on Elections 1999***

5. ***Python Data Science Handbook, O'Reilly Media, 2017***

6. ***Joel Grus, Data Science From Scratch***

7. ***Chief Electoral Officer Haryana Data (2012)***
    a. < http://ceoharyana.nic.in/>

8. ***District wise map of India***
    <https://www.mapsofindia.com/assemblypolls/haryana/>

9. ***Election statistics for state of Haryana***
    a. <https://www.wikipedia.org/electionstatistics/haryana>

10. ***Census Statistics for state of Haryana***
    a. <https://www.census2011.co.in/census/state/haryana.html>

11. ***Numpy Documentation 2018***
    a. < https://docs.scipy.org/doc/>

12. ***Pandas Documentation 2018***
    a. < https://pandas.pydata.org/pandas-docs/stable/>