DEMOGRAPHIC PROFILING (EDA)

Under the mentorship of:
Dr. Ritika Wason
(Associate Professor)

Submitted by: Harsimran Singh
Roll No.: 01611604416

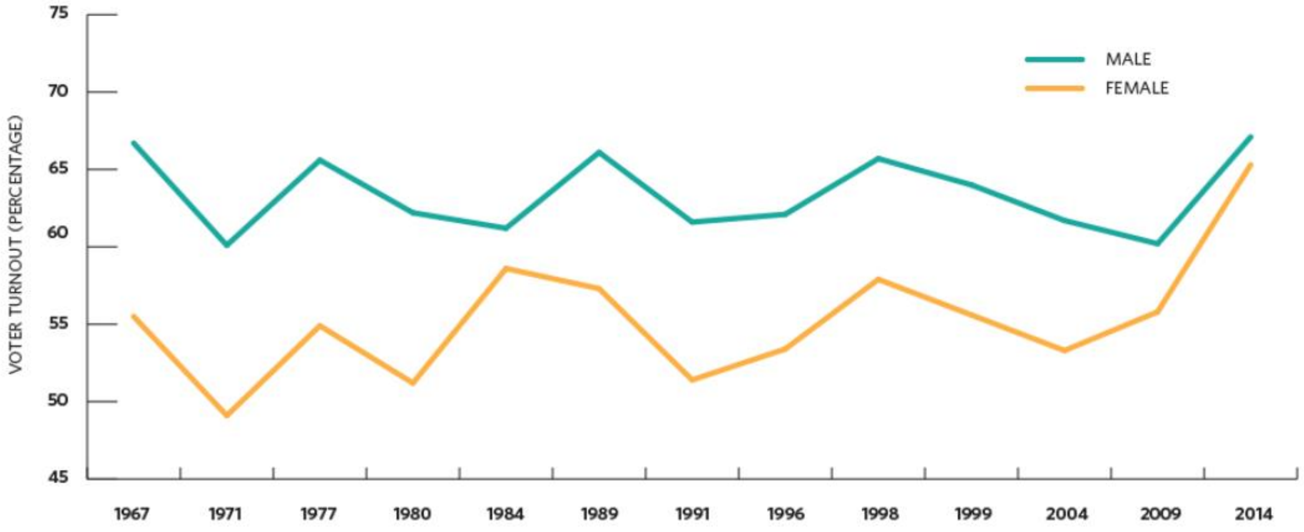
Content

- Company profile
- Objective
- Description
- Technology used
- Data flow diagram
- Screenshots
- Conclusion
- Limitations
- Future scope
- References

Company profile

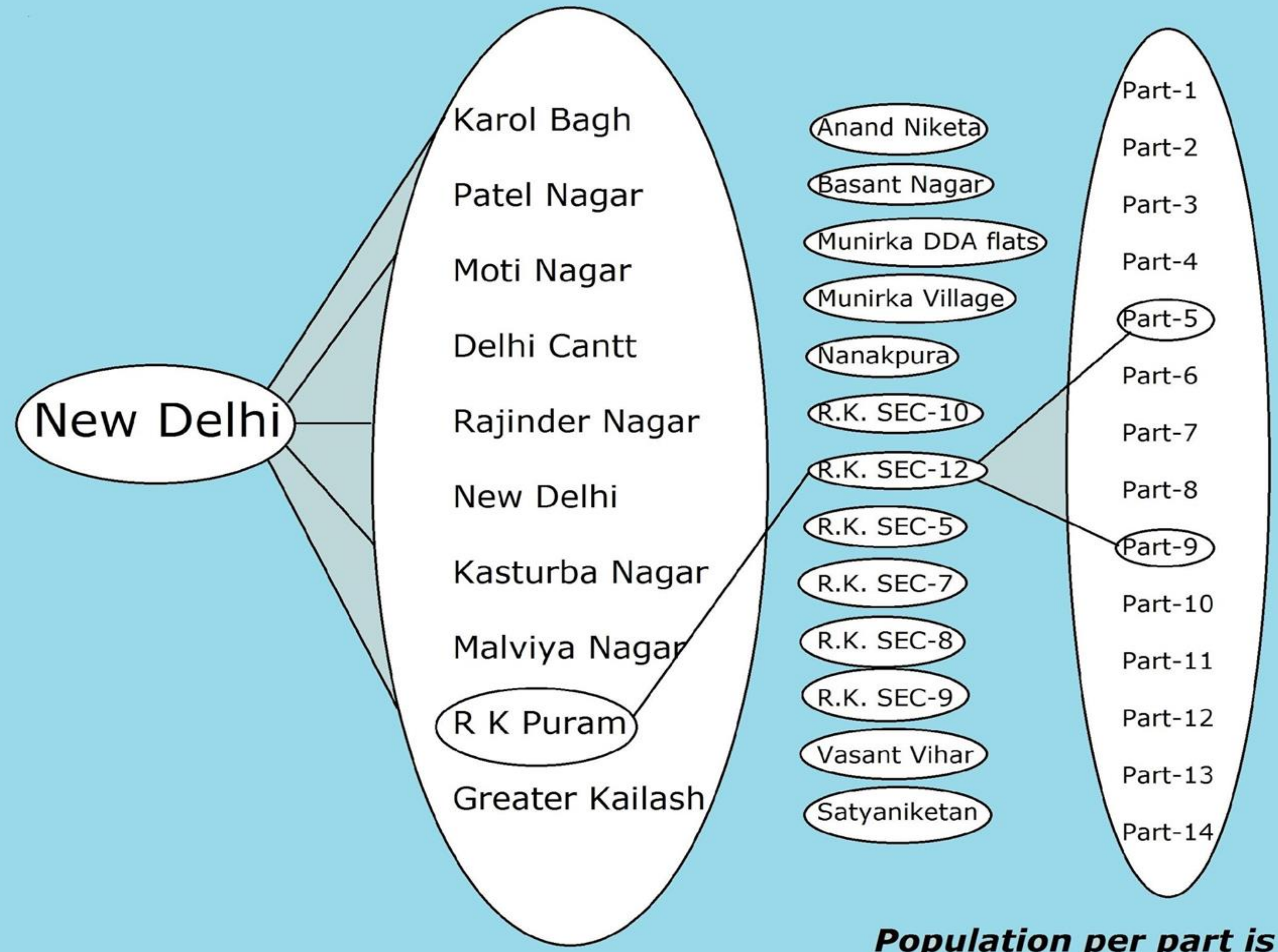
- Company name: **Aam Aadmi Party**
- Aam Aadmi Party (AAP), is an Indian political party, formally launched on 26 November 2012, and is currently the ruling party of the National Capital Territory of Delhi.
- It deals in various profiles like HR, Public Speaking, Data Analytics and Machine Learning.

Objectives

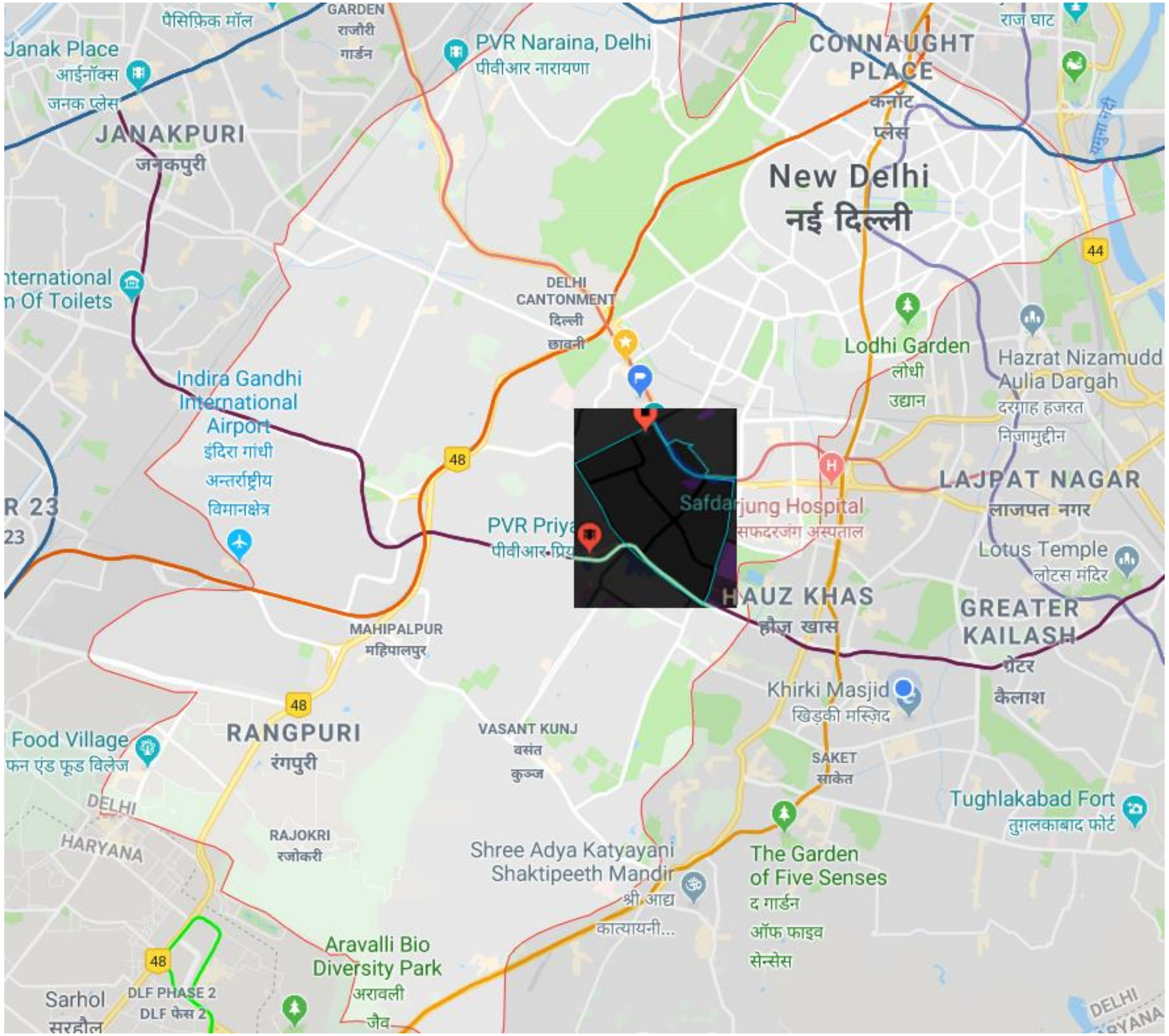


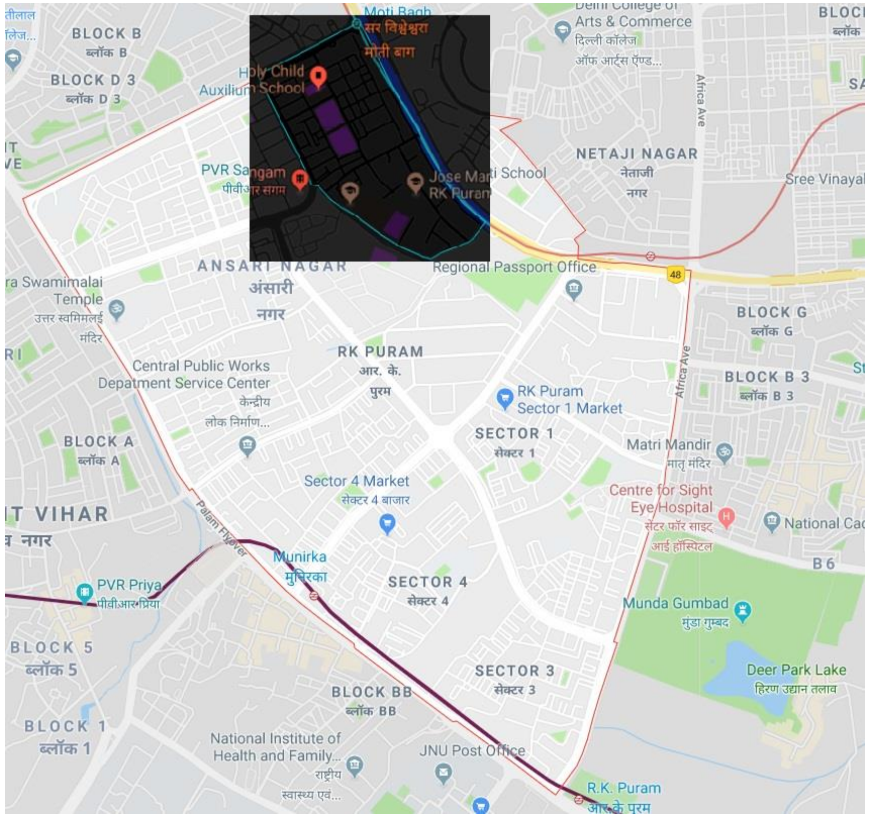
- Percentage variation in gender ratio, young female voters and their correlation with the affluence of an area
- Variation in affluence within one constituency

- Variations between smallest possible population groups. i.e. at block level
- We saw a dramatic increase in female voter turnout in 2014 Lok Sabha election
- We explored the significance of female votes.

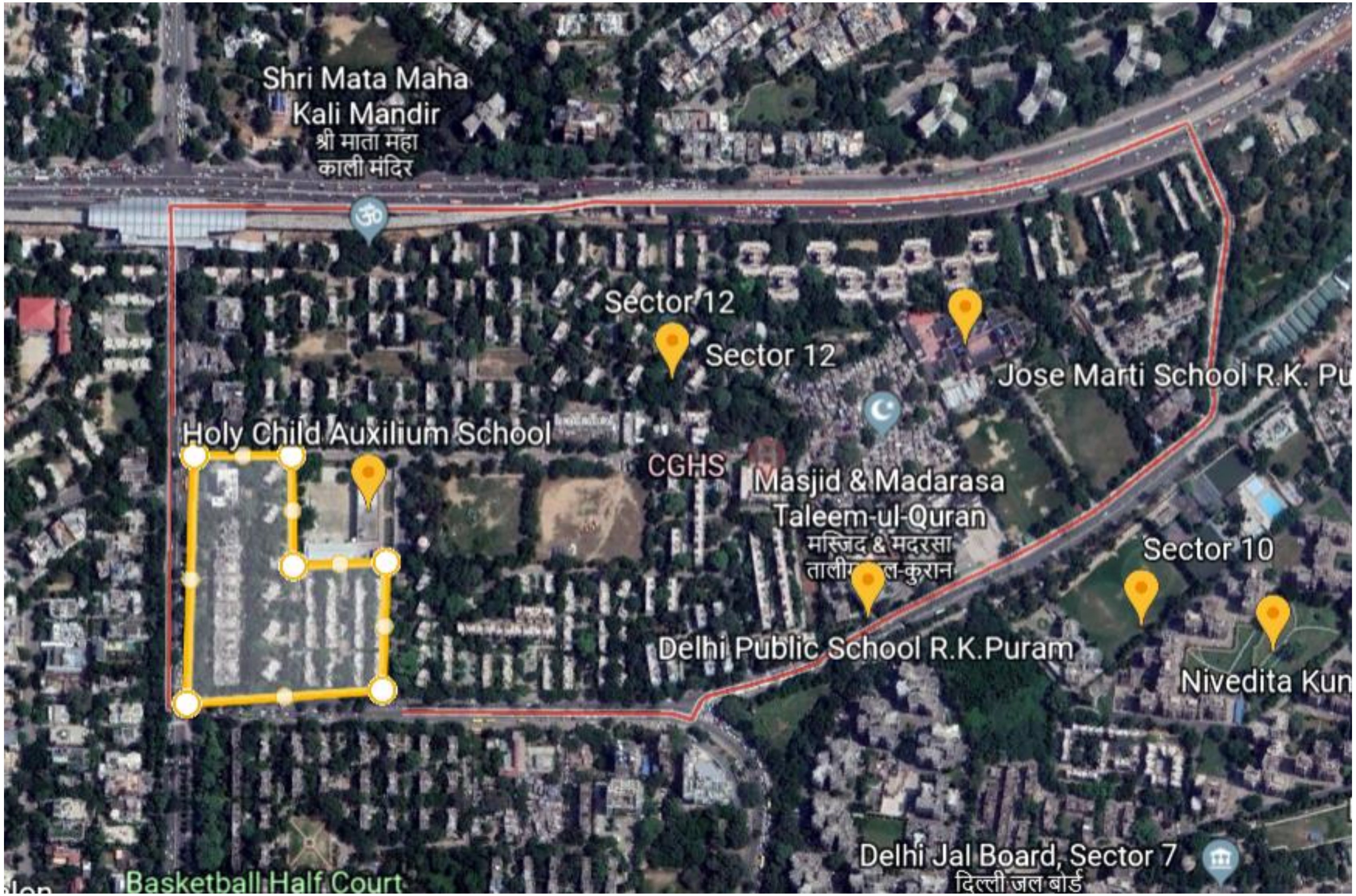


Population per part is approx 1200





Geographic Scale of Investigation



Part no. wise geographic scale of a constituency area

Software Requirements

- Python 3.7 with various libraries
- Libraries:
  - Numpy 1.16.2
  - Matplotlib 3.0.3 and Seaborn 0.9.0
  - Scikit-learn 0.20.3
  - Text Recognition (OCR), OpenCV 3.4.5 and Tesseract 3.05.02
- Jupyter notebook
- Anaconda 2018.12
- Excel 2016

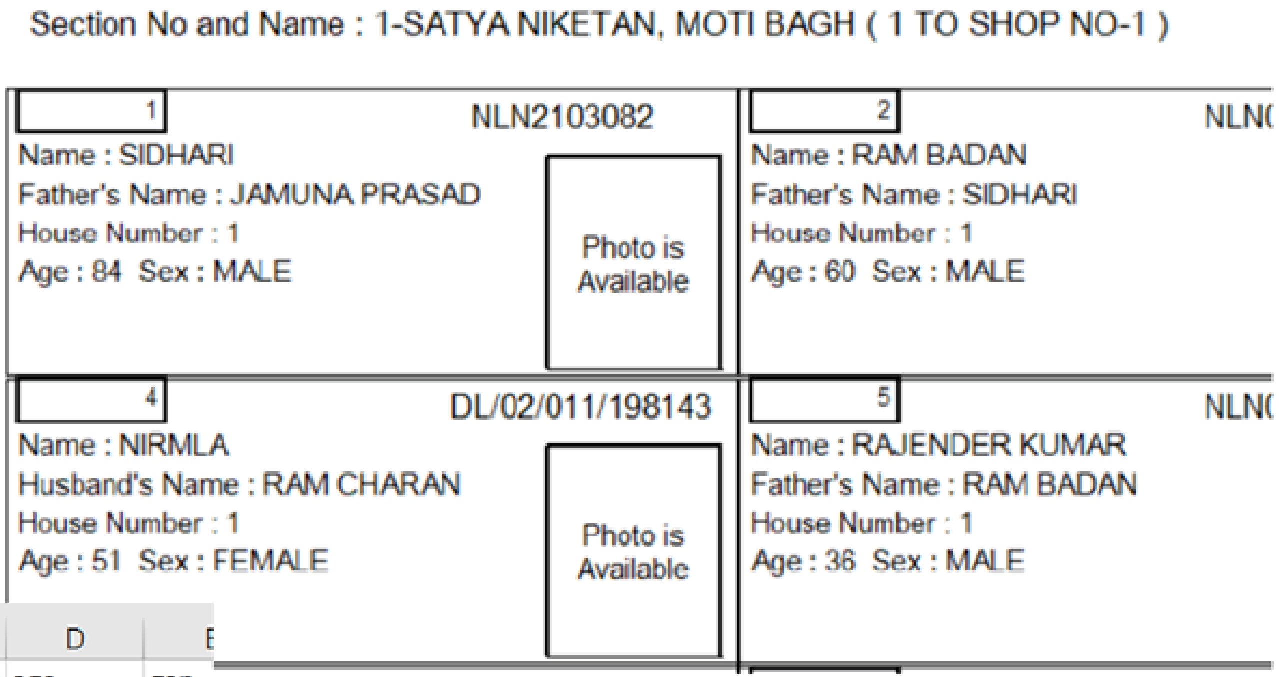
Minimum Hardware Requirements

- OS: Windows 7 or higher, Linux
- RAM: 4GB
- HDD: 512GB

Data Collection

- We extracted electoral roll using image processing. Source of this data is election commission’s website.
- Collection of features was done through web scraping.
- Optical character recognition techniques were used in image processing and geolocation API was used in web scraping.

Data Extraction



Section No and Name : 1-SATYA NIKETAN, MOTI BAGH (1 TO SHOP NO-1)

1 NLN2103082
Name : SIDHARI
Father's Name : JAMUNA PRASAD
House Number : 1
Age : 84 Sex : MALE
Photo is Available

2 NLN(
Name : RAM BADAN
Father's Name : SIDHARI
House Number : 1
Age : 60 Sex : MALE

4 DL/02/011/198143
Name : NIRMLA
Husband's Name : RAM CHARAN
House Number : 1
Age : 51 Sex : FEMALE
Photo is Available

5 NLN(
Name : RAJENDER KUMAR
Father's Name : RAM BADAN
House Number : 1
Age : 36 Sex : MALE

Image to excel

Voter data

↓

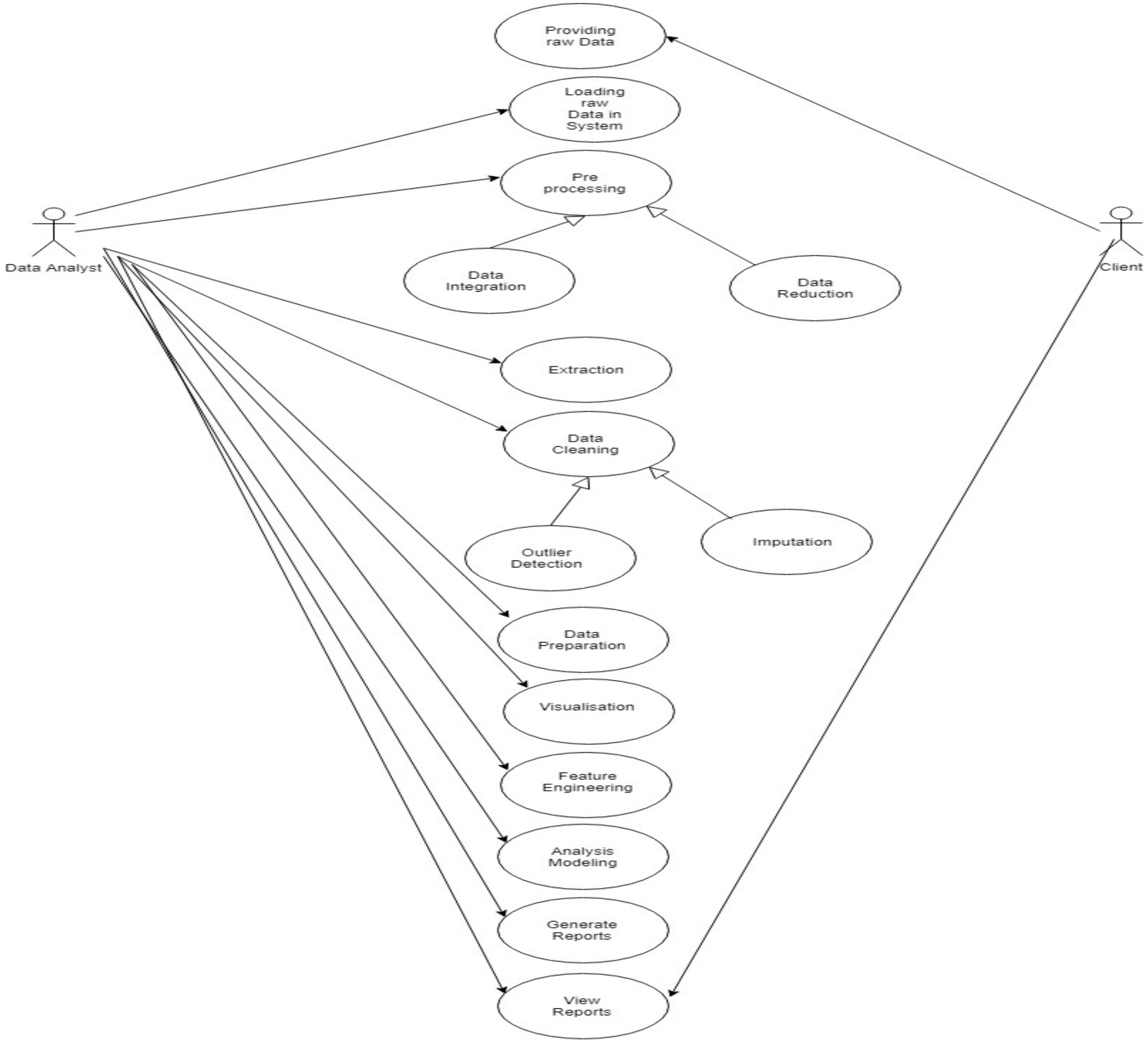
	A	B	C	D	E
1	name	gaurdian	g_relation	age	gen
2	Sidhani	Jamuna Prasad	F	84	M
3	Ram Badan	Sidhari	F	60	M
4	Ram Charan	Sidhari	F	56	M
5	Nirmala	Ram Charan	H	51	F
6	Rajender Kumar	Ram Badan	F	36	M

Feature Engineering

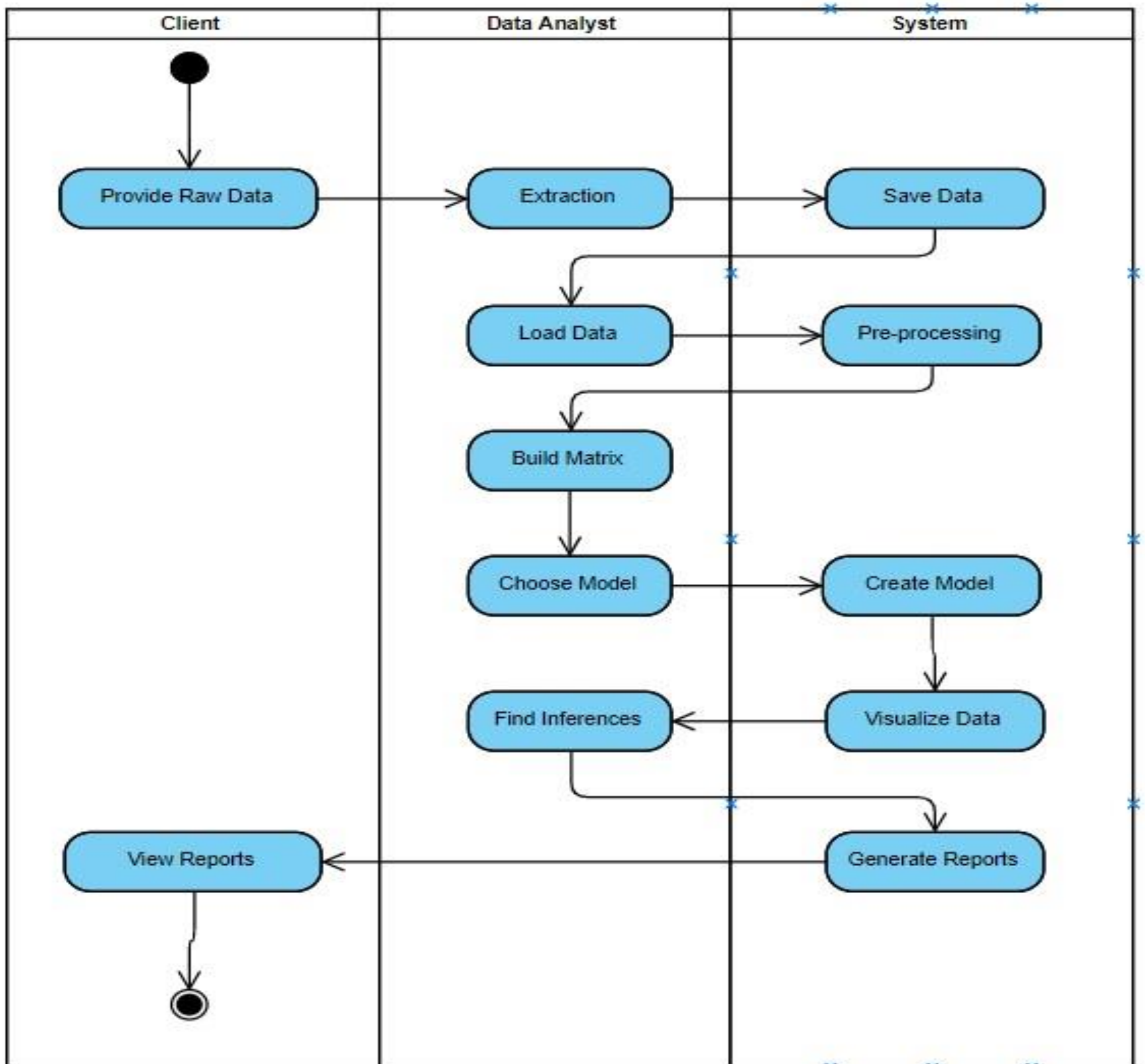
Coffee, salon, property rates and centroid distance

A	B	C	D	E	F	G	H	I
	Locality	partno1	Male	Female	Coffee.shc	Salon	buy.per.sq	m_dist5
9	Munirka D	156	394	400	4	2	12000	0.28
12	Munirka D	152	428	437	4	2	12000	0.28
18	Munirka D	155	481	482	4	2	12000	0.28
42	Munirka D	153	415	459	4	2	12000	0.28
66	Munirka D	154	471	528	4	2	12000	0.28
1	Munirka Vi	146	554	401	4	2	5598	0.6
19	Munirka Vi	151	757	501	4	2	5598	0.6

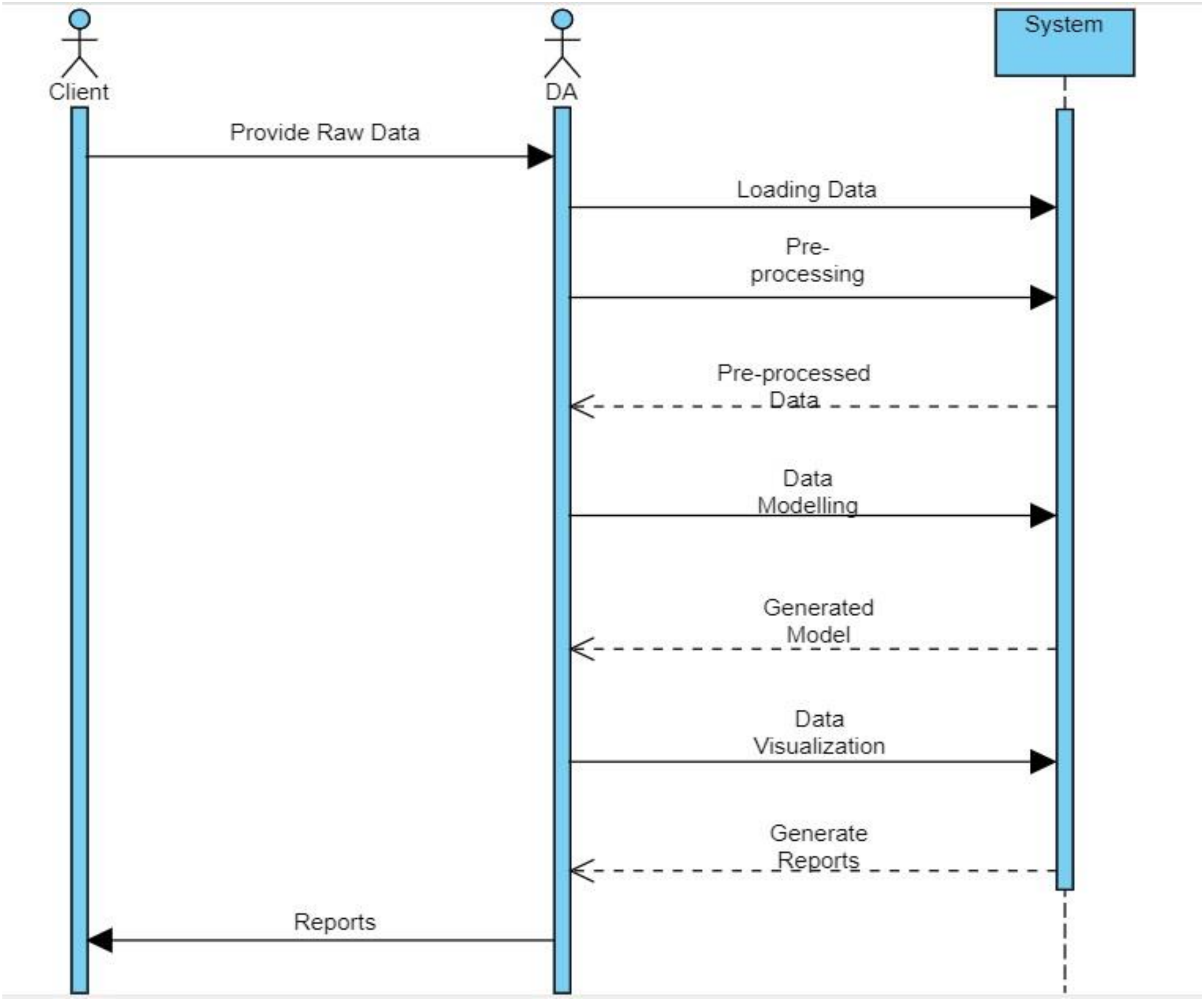
Use Case Diagram



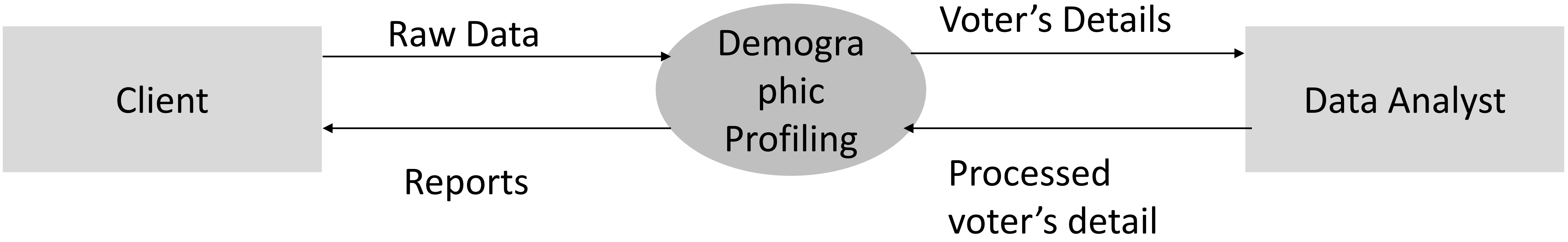
Activity Diagram



Sequence Diagram

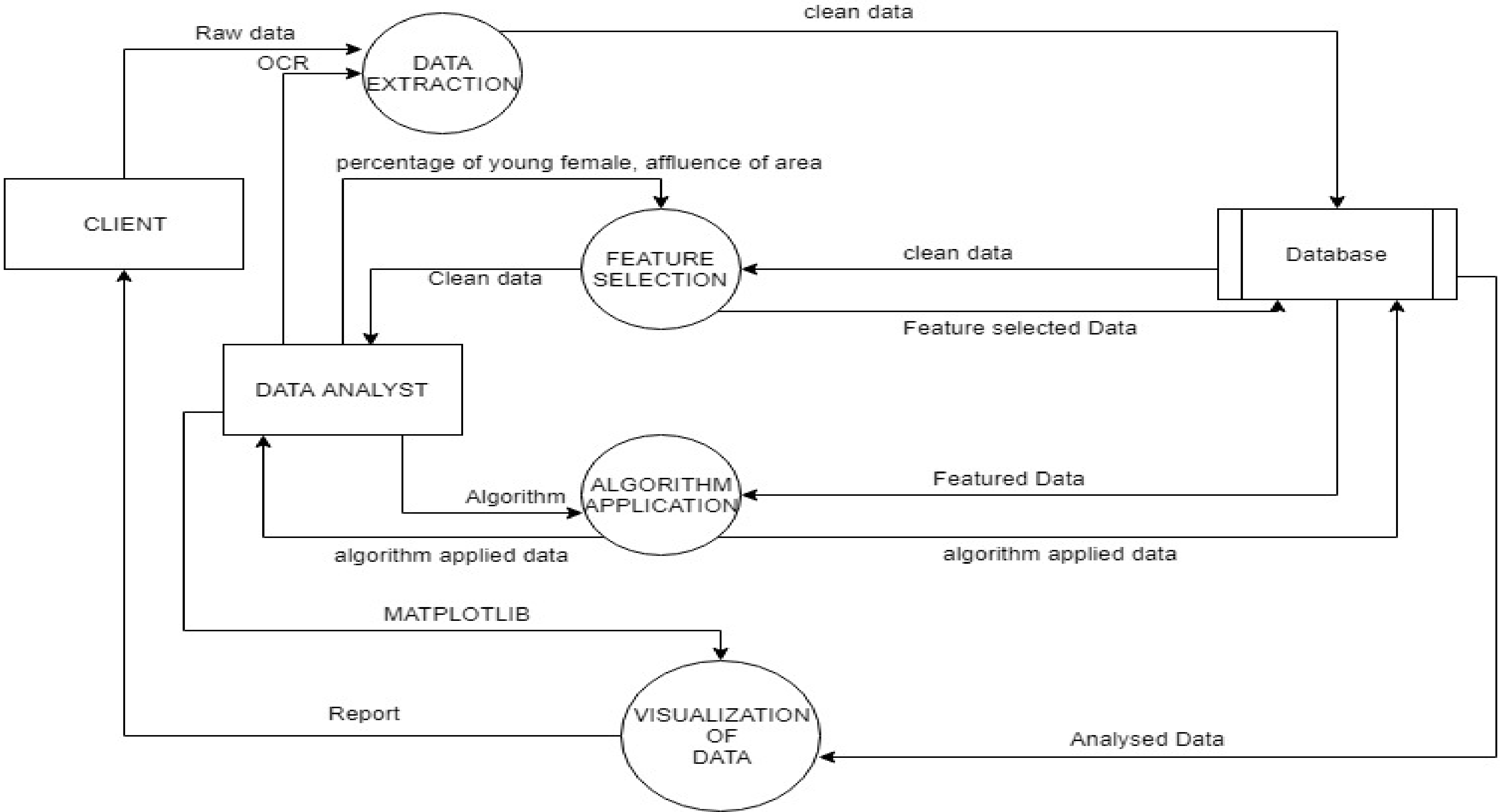


DFD: Level 0

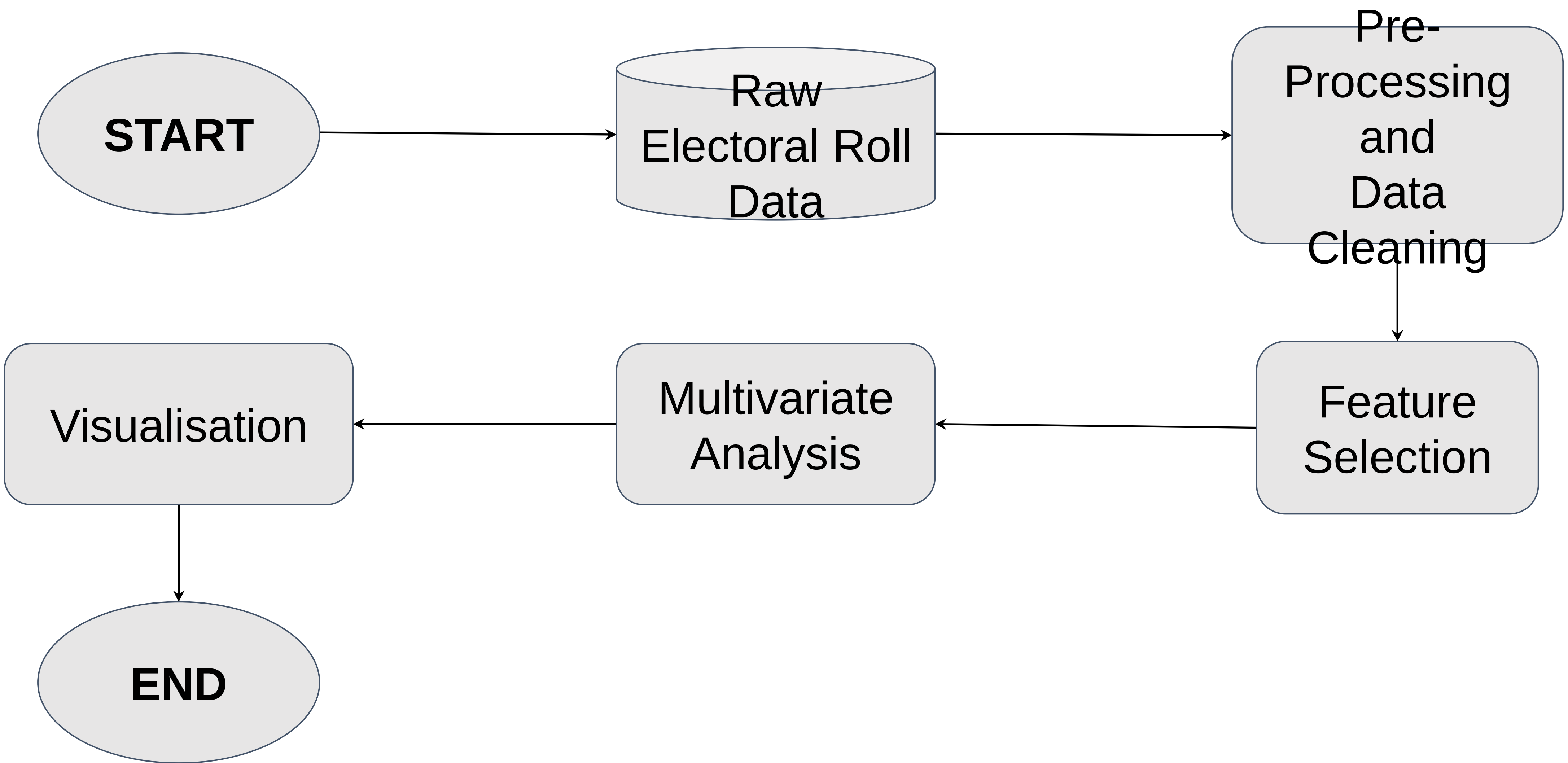


Raw Data
Voter's Details
Client
Demogra
phic
Profiling
Data Analyst
Reports
Processed
voter's detail

DFD: Level 1


clean data
Raw data
OCR
DATA
EXTRACTION
percentage of young female, affluence of area
CLIENT
FEATURE
SELECTION
clean data
Database
Clean data
Feature selected Data
DATA ANALYST
Featured Data
Algorithm
ALGORITHM
APPLICATION
algorithm applied data
algorithm applied data
MATPLOTLIB
Report
VISUALIZATION
OF
DATA
Analysed Data

Workflow Diagram



Feature Selection:

- Features were selected such that their combination could give us a good measure of affluence.
- Those coffees and salons were selected which would be visited by people from affluent areas.
- We set a benchmark in terms of cost and distance from a locality.
- Distance was set keeping in mind the geo-locations of coffee-shops and salons.
- For e.g.: Most coffee shops are found near a market place and a market is visited by people living within 1 km area distance.
- Those salons were selected which are within 400m of area distance.
- Cost was selected such that it can give a good approximation about purchasing power
- Property rates were the last piece of the puzzle. As there could be a coffee-shop located which is closer to both an affluent area and an area with high population density
- Coffee shops and salons also helped to distinguish one affluent area from another.

Analysis Procedure

- We did a *Multivariate analysis* through which we assigned a value to each ***Part no. of a constituency.***

- Then, we found a centroid of all the Part nos. and did our analysis by measuring the Mahalanobis distance from the centroid.

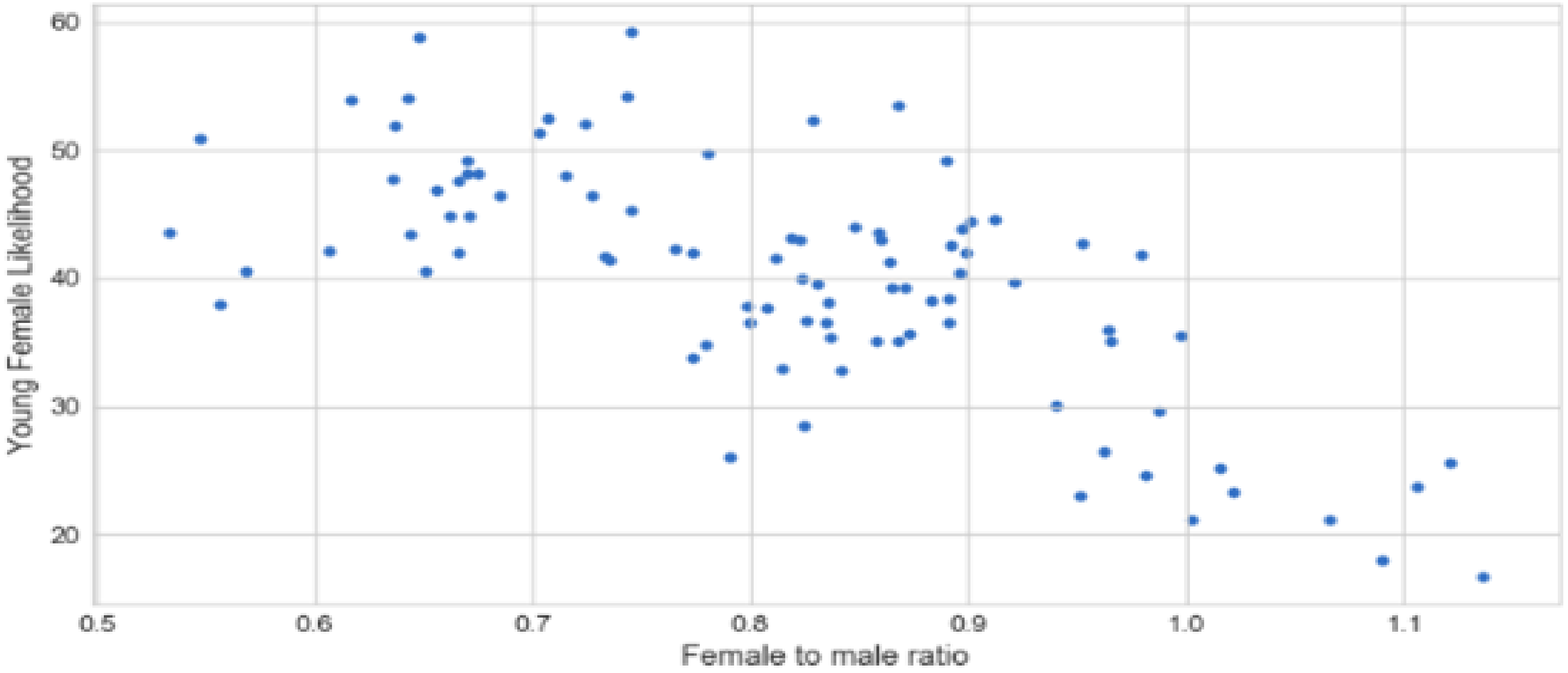
- Through electoral roll data, we calculated gender ratio, young female and married percentage, young population percentage.

- We analyzed the variations in above calculated values with respect to affluence

Variance between a constituency and its sub parts

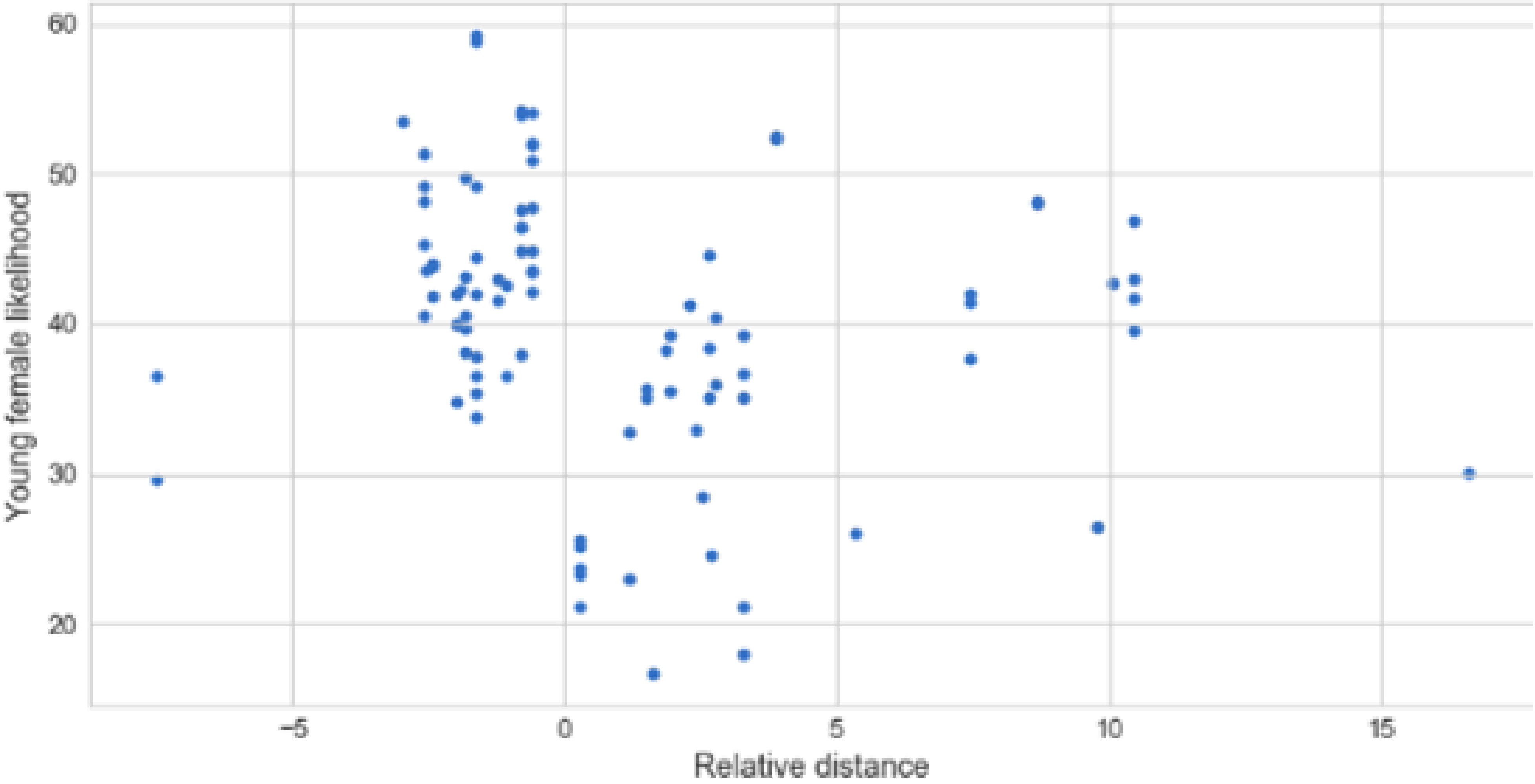
- Karol Bagh has gender ratio(Female to Male) of 0.83 whereas it varies in the range(0.4 to 1.09) for its subparts
- The likelihood of a female voter being young is 35% in Karol Bagh whereas it lies in the range(23% to 51.6%) for its subparts
- R K Puram has gender ratio(Female to Male) of 0.7 where as it varies in the range(0.51 to 1.2) for its subparts.
- The likelihood of a female voter being young is 40.4% in R K Puram whereas it lies in the range(21% to 60%) for its subparts.

Rk puram : Young female likelihood vs Female to male ratio.



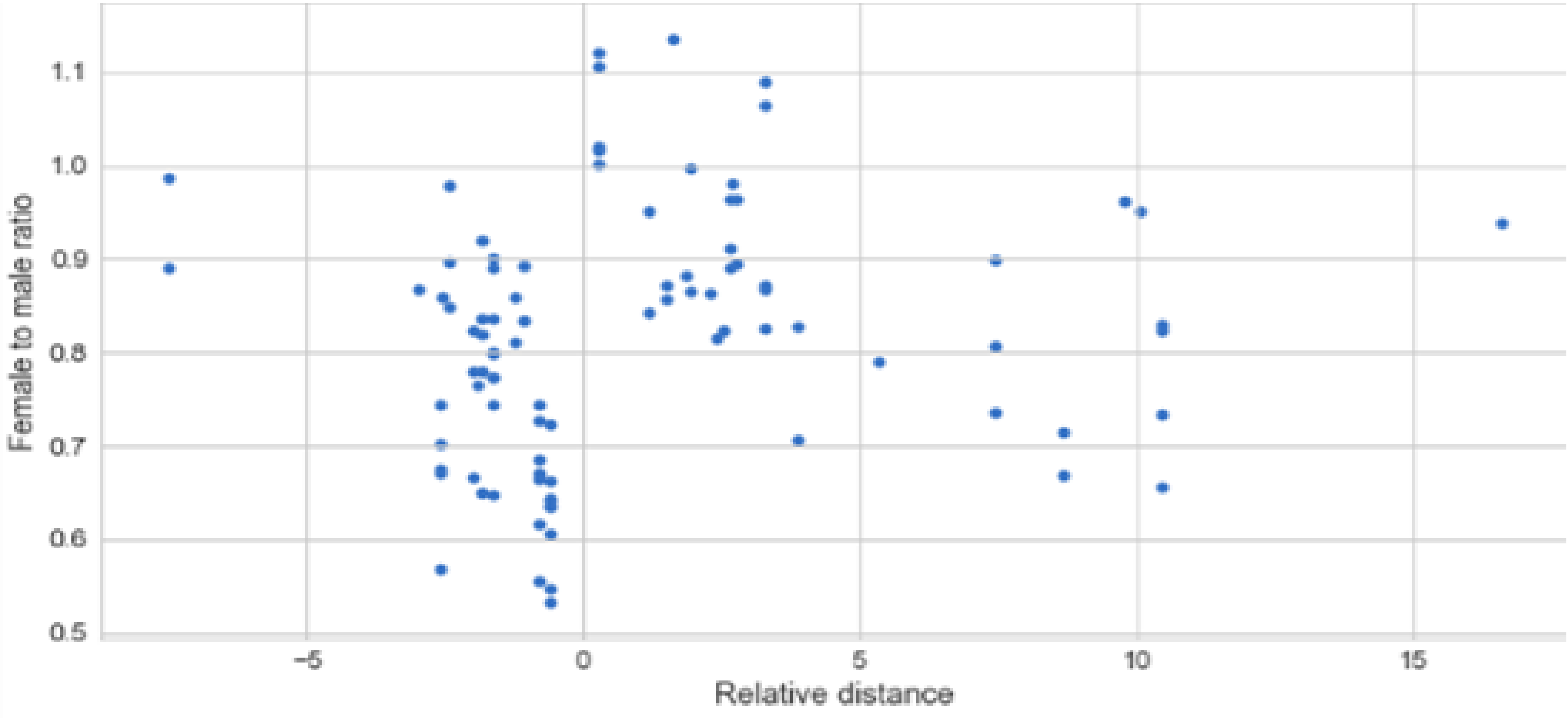
Young female likelihood decreases as female to male ratio increases.

Rk Puram: Young female likelihood vs Relative Distance.



There are more number of young females near the centroid that means near munirka dda flats.

Rk Puram: Female to male ratio vs Relative Distance.



Females are more near centroid that means females are more near munirka dda flats.

Conclusion

- It can be used to study **Voters** base in the constituency and target them accordingly(ward level micro targeting.
- Formulate strategies to target the **younger audience digitally** effectively.
- Many localities have **juggi** near them which can be identified by data and accordingly targeted for building **narratives**.
- Ward wise narrative can be formed according to the data such as % of Female and Male.
- Areas with higher no. of **married women** can be targeted with policies and narratives keeping in view with the **family** orientation.
- We may derive future health, education and family policies keeping in view of our analysis done, so as more voters can be attracted as we can **extrapolate** central gist of families in an area; bases upon there current **female trends**.

Limitations

The only limitation of the model is that it only works in this time-space only.
It may or may not work after 10 years or as the space changes.
The model may not work outside a specific region.

Future Prospects

- We can further extend this project by conducting ***inter-cluster analysis*** (between two or more different constituencies, for eg: R.K. Puram and Rajinder Nagar).
- Currently we have only conducted ***intra-cluster analysis*** (within one constituency).
- We can predict features of other constituencies by determining the similarity of their regions based on the affluence of that region.

References

- https://www.tutorialspoint.com/python/
- https://www.scrapehero.com/web-scraping-tutorials/
- https://www.stackoverflow.com/
- https://www.youtube.com/watch?v=mKxFfjNyj3c
- Christopher M. Bishop, “Pattern Recognition and Machine Learning”, Springer, 2006
- Jiawei Han, Micheline Kamber and Jian Pei, “Data Mining Techniques and Concepts”, Morgan Kaufmann Publishers, 3rd Ed., 2012
- Allen B. Downey, “Think Stats: Probability and Statistics for Programmers”, Green Tea Press, 2011

THANK YOU