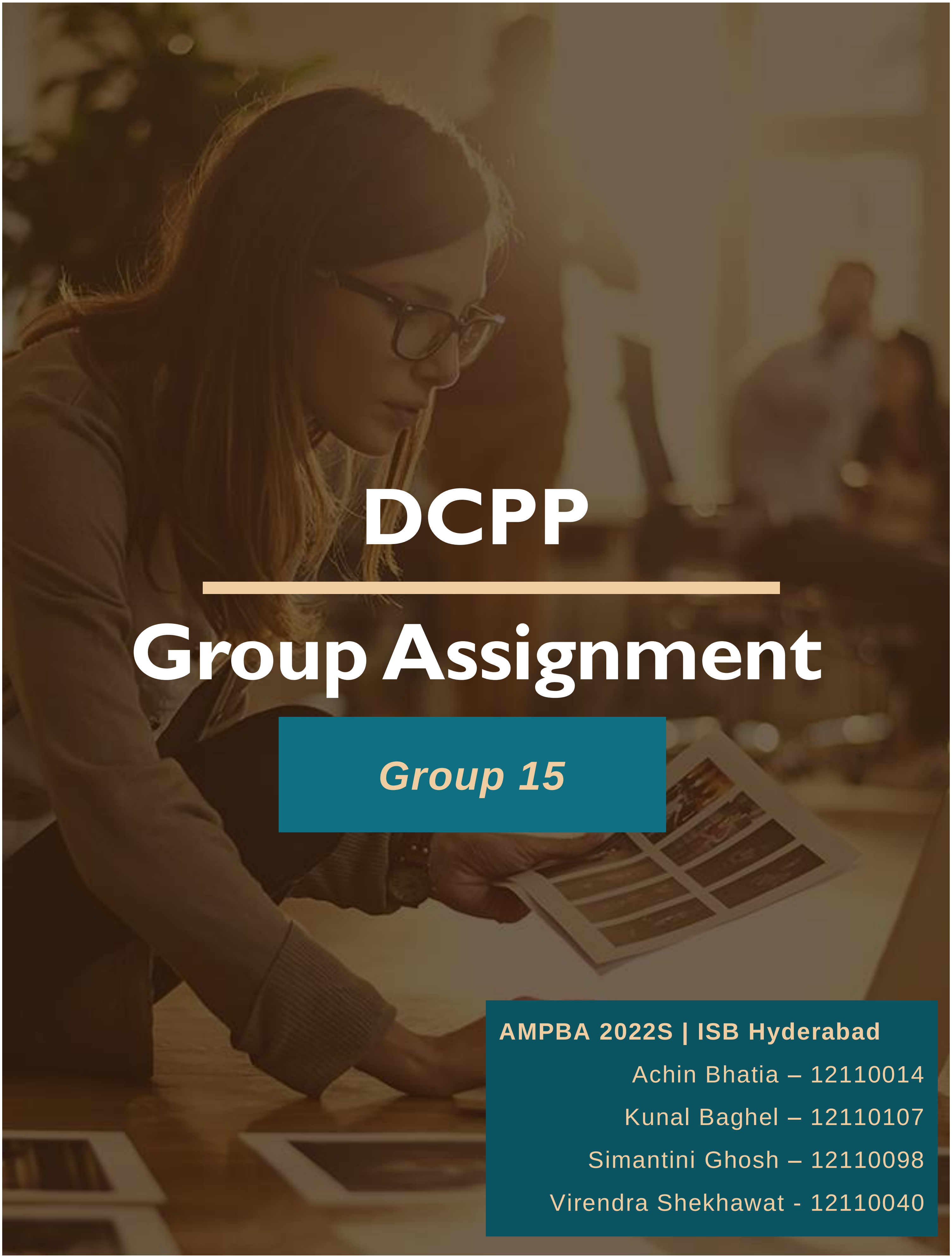


DCPP

Group Assignment

Group 15

AMPBA 2022S | ISB Hyderabad

Achin Bhatia – 12110014

Kunal Baghel – 12110107

Simantini Ghosh – 12110098

Virendra Shekhawat - 12110040

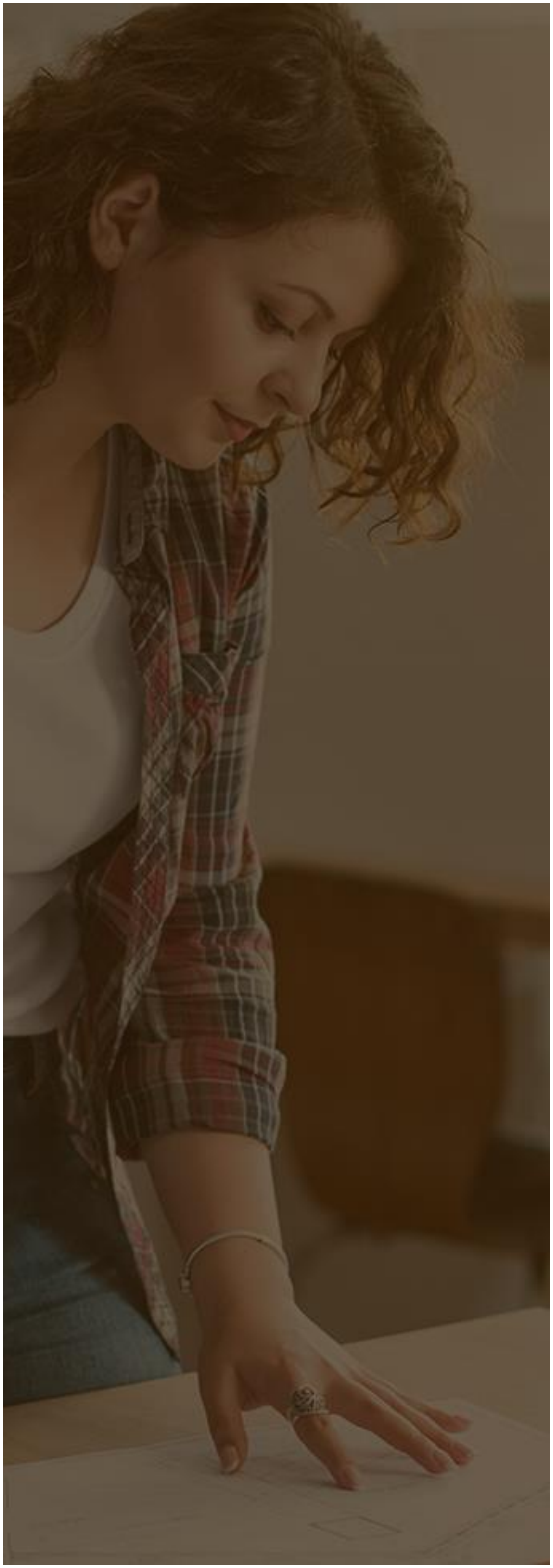


TABLE OF CONTENTS

1. EXECUTIVE SUMMARY

While it's a cliché that 'data is the new oil', it is running the global engine currently. However, just like crude oil, data needs extraction and refining. In our case, a group of four non-coders sat together and enjoyed every bit of the BeautifulSoup that was served to us.
Our analysis of data on Members of Parliament revealed that average Indian MP is a married senior citizen with 2 kids (contrary to popular belief) and one in 2 MPs is from either BJP or Congress. Moreover, most of the MPs are engaged in Agriculture or Social Work.

PROBLEM STATEMENT

Ideally, data should be available in structured format, ready to do EDA on which we can strengthen our hypothesis & decision-making abilities. The data available was highly unstructured. Mostly available on Govt. website only but because of unstructured coding, it is very hard to decipher. Even the data that is available, is in various formats making it harder to understand for Python/ or any other coding language.
As a result, it takes a large amount of time to scrap, refine & structure the data before actual analysis can be done. Also, in terms of skills, the challenge was behemoth. A particular software was not enough to do the scraping & refinement of the data.

PROPOSED SOLUTION

Our group seeks to resolve the problem by using Selenium & BeautifulSoup for data scraping & Python for EDA. These tools will allow us to scrap the data in a meaningful format making it easy to understand & use in the public domain through powerful visualization techniques.

BRIEF UNDERSTANDING OF THE CHALLENGES

While the understanding on requirement was clear, the biggest challenge we faced was that none of us was from coding background.
We initially started with a different domain ('Awards and Recognition') however the data was in multiple unstructured formats rendering it almost impossible to extract data. The only way to extract that data was to do web slicing for each website separately – this would have been a tough ask. Hence, we consulted Prof. Vasudeva Varma and TA Shreya Singireddy and changed the domain to Member of Parliaments.

- There are multiple unstructured formats in the HTML data.
- This makes it impossible to use a single code that can run through all HTMLs.
- The code used to extract data was taking a lot of time and was often getting timed out.

2. CHOSEN DOMAIN AND SEED SOURCES

Our domain is Members of Parliament across all years since independence. For this we have gathered Lok Sabha members information from official Loksabha website.

We believe politics (and politicians) in India is a tea-time discussion BUT there is very few analysis carried out on the same. We think a lot of inferences can be derived from these datasets and can be used to take further actions.

Moreover, there are many notions that people have for politicians, including but not limited to:

- Average age of an MP is between 40 to 60
- Most of the MPs are married
- Almost all MPs have 4 kids
- Congress is the all-time biggest party

While some of these notions are indeed true, others are not. These are explained in further sections of the report.

3. STRUCTURED AND UNSTRUCTURED SOURCES

STRUCTURED DATA

We were provided with a Seed data file with neatly organized columns including Name, Party, Constituency and Lok Sabha experience. This sheet also contained links to the profiles of these MPs.

UNSTRUCTURED DATA

We extracted links and from those links we gathered unstructured data using web scraping.

PUBLIC AND PRIVATE DATA

All the data being utilized for the current exercise is in open domain and can be freely accessed.

There are other datasets which could add more value (property value of MPs, salary of MPs, cases against MPs etc.) but these datasets are out of purview for the current analysis carried.

4. DOWNLOAD/ CRAWL/ COLLECT DATA

BELOW ARE SOME OF THE METHODS UTILIZED

- BeautifulSoup
- HTML Parser
- Python libraries

CHALLENGES

- None of us is from coding background so this was a real test for us
- There are multiple unstructured formats in the HTML data
- This makes it impossible to use a single code that can run through the all HTMLs
- The code used to extract data was taking a lot of time and was often getting timed out

HOW DID WE HANDLE/ SOLUTIONS IMPLEMENTED

- We took help from different sources including
    - Google (geeks4geeks, stackoverflow etc.)
    - Batchmates – we asked few of our batchmates who helped us in correcting our code
- We created loops to run through different unstructured formats
- We ran the code in pieces (0 to 40, 40 to 80 etc.) and converted them into excel (so that we do not need to do this time and again)

5. CONVERT DATA

- We started with an excel file containing hyperlinks
- We utilized BeautifulSoup for extracting data
- We converted data thus extracted into excel
- Lastly, we used this excel to carry out further analysis

6. DATA CLEANING/PRE-PROCESSING

Since most of the datasets were in different formats and with disparate content, we had to align all of them into a desirable format.

TECHNIQUES USED FOR EDA/ DATA CLEANING/ PRE-PROCESSING

- We **filtered** data where values were not present (in Constituency column)
- We **dropped** columns which we considered as not adding value to our analysis
- For couple of columns, we extracted numbers from a combination of numbers and text
- For few columns we **replaced NaN values** as zero – this was done where we understood that NaN would mean zero (like No. of Sons and No. of Daughters)
- We **changed data format** for few columns where it should have been number instead of string
- We changed values that did not make sense or were erroneous. For instance, in Date of Birth, there were few values dating up to 2029 – this was happening as 1929 was being treated as 2029. We adjusted such values (which made sure that MPs age is more than 20 at the least) by subtracting 100 years from the values. Simply put, we **changed DOB in 20xx values to 19xx** (except 2000)
- For few columns the values were repeating in different cases, we aligned them (example: Marital status: ‘Married’ and ‘MARRIED’)
- For some columns we **merged values** that convey similar meaning (example: Single and Unmarried in ‘Marital Status’ and Independent and Independents in ‘Party Name’ columns)
- For many columns, the text contained a mix of different values like Constituency contained information about Constituency, Caste and State. We tried various methods to bifurcate this information but were unsuccessful

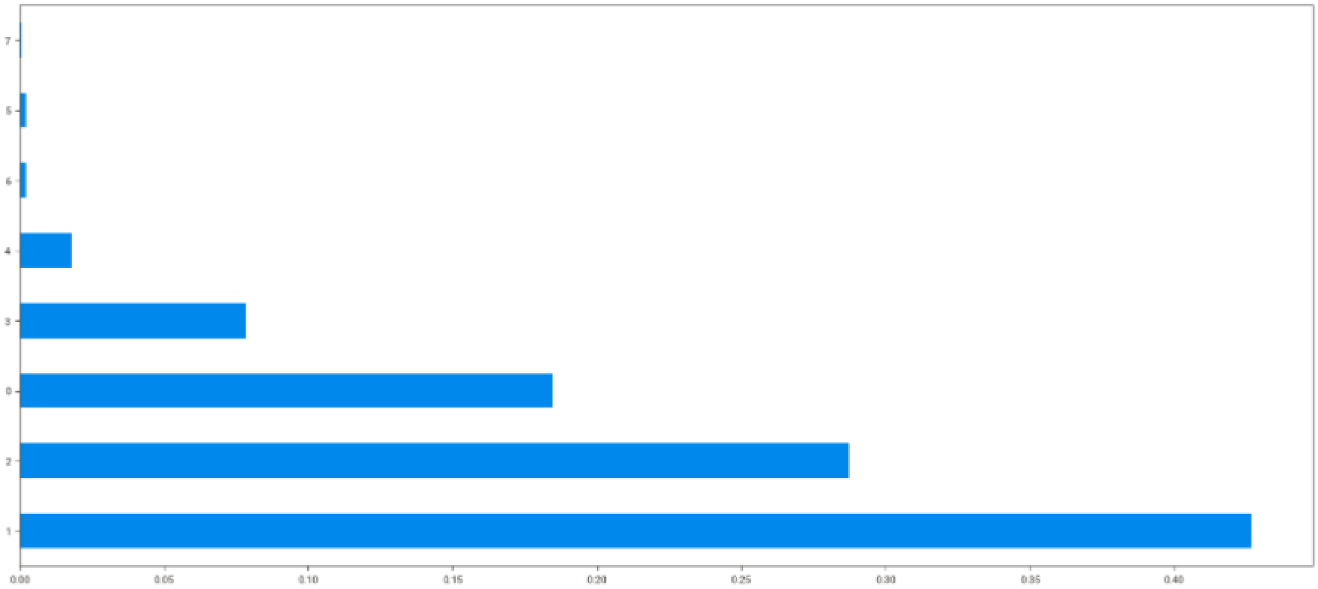
EXPLORATORY DATA ANALYSIS OF MEMBERS OF PARLIAMENT DATASET

OBSERVATIONS OF INITIAL DATASET:

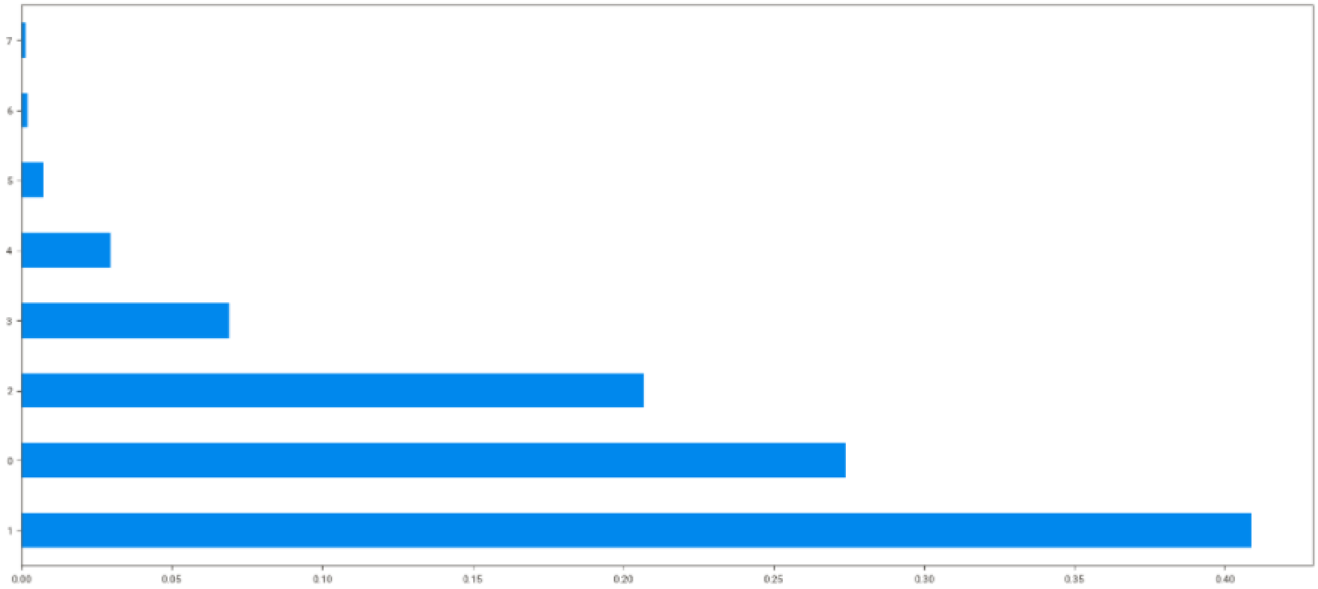
- Initial no. of attributes (Columns) was 20
- Most of the values are categorical
- We observed that variable ‘Constituency’ has "." as values. We will filter our data accordingly.
- Few columns seemed not adding any value
  - Unnamed: 0 - this looks like serial number
  - Email - we cannot use this for any analysis
  - Unnamed: 5 - has only NaN values
  - Positions Held - has only NaN values

ANALYSIS FOR NUMBER OF SONS AND DAUGHTERS:

- No. of Sons and No. of Daughters has NaN values
- Also, few values have text embedded, we read only the first character which gives us the number
- We will convert its datatype to integer to carry out calculation and analysis

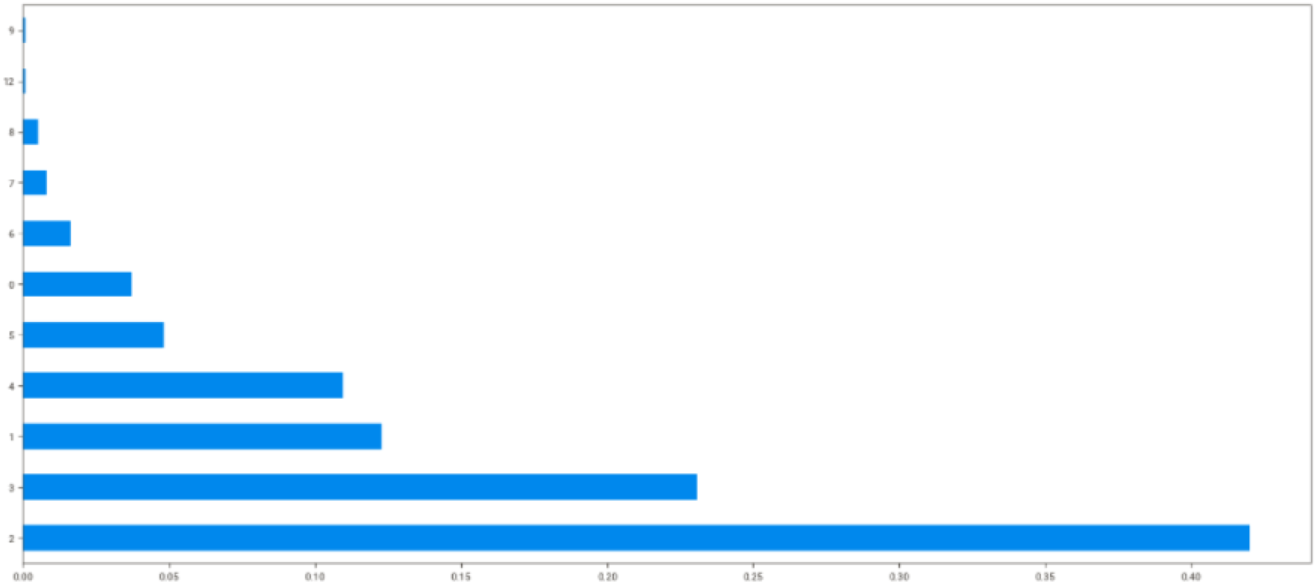


  - ○ Sons per MP: average is 1.3, median is 1.0
  - ○ Almost 3/4th of the MPs have 1 or 2 sons
  - ○ Around 18% have no sons
  - ○ Very few MPs have more than 3 sons



- Daughters per MP: average is 1.2, median is 1.0
- Almost 2/3rd of the MPs have 1 or 2 daughters
- Very few MPs have more than 3 daughters
- ~27% MPs have no daughters

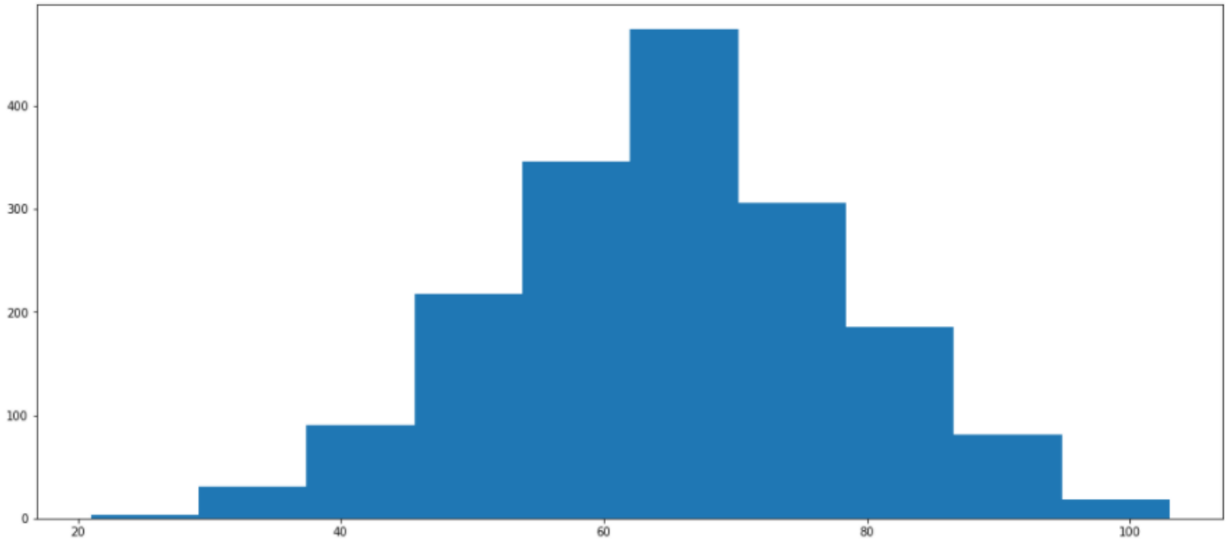
- We added these columns to arrive at Number of Kids

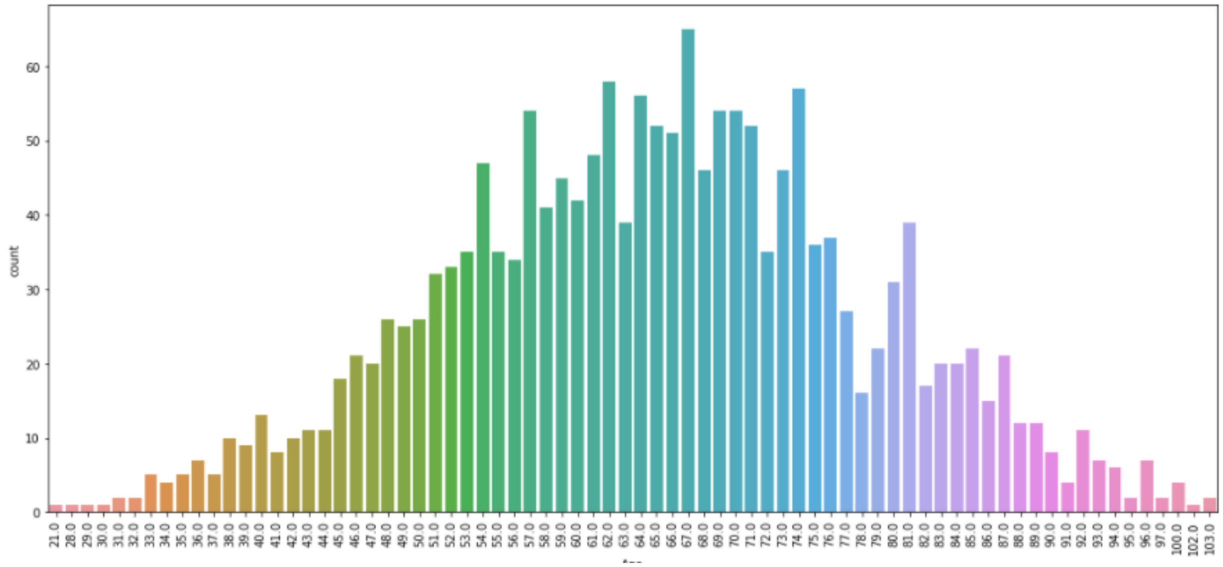


- Kids per MP: average is 2.5, median is 2.0
- Maximum number of Children for any MP is: 12
- MP with maximum no. of children is: Mohale, Shri Punnulal
- Around 47% MPs have 2 kids and around 12% have just one kid
- 4% MPs have no daughters
- Almost half of the MPs have more than 2 kids

ANALYSIS OF AGE OF MPS:

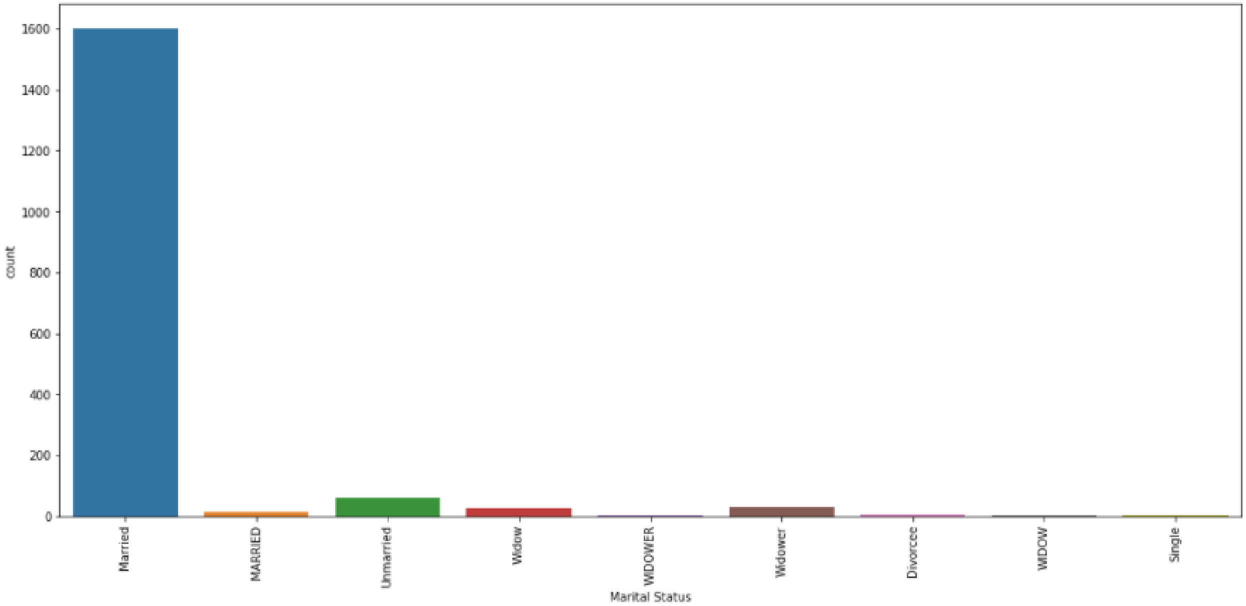
- We observed that few values in ‘Date of Birth’ column are incorrect. This is because data of birth of some of the MPs is in early 20th Century
- We assume that MPs have to be at least 20 years (though we know it should be 25 at the least), we will subtract 100 years from people born in or after 2000
- Average age of an MP is: 65.0 years.
- Median age of an MP is: 65.0 years.
- Oldest MP's age is: 103.0
- Oldest MP is Karunakaran, Shri K.
- Youngest MP is Madhavi, Smt. Goddeti
- Here is the distribution of age of MPs:



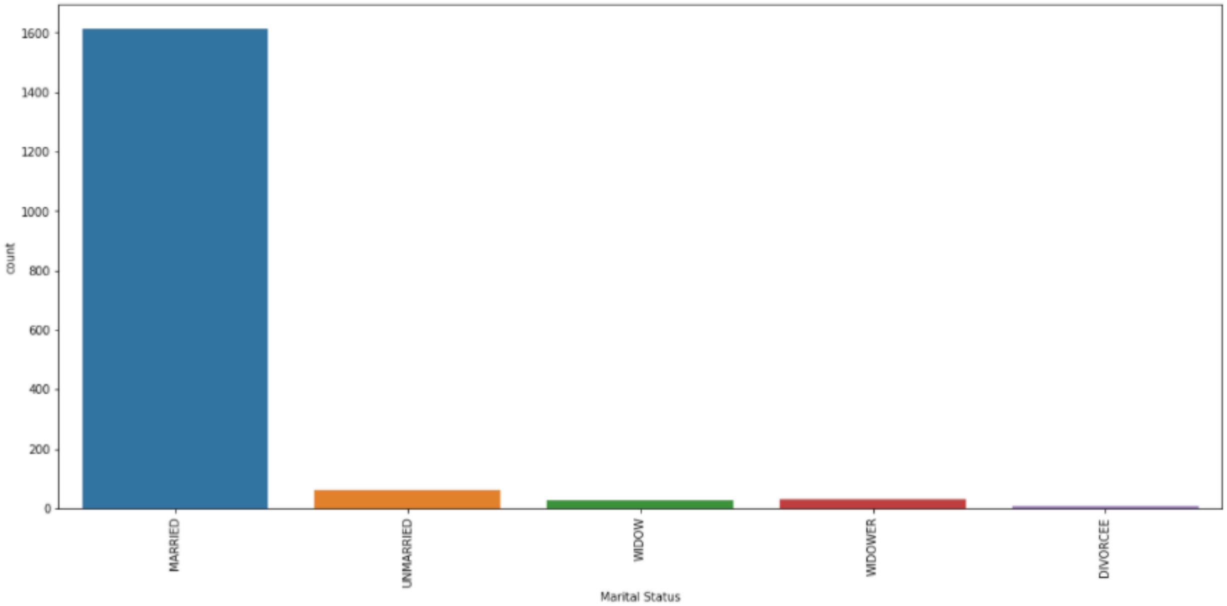


ANALYSIS OF MARITAL STATUS OF MPS:

- We observed that most of the categories are repeated ('Married' and 'MARRIED', 'Widow' and 'WIDOW' etc.)

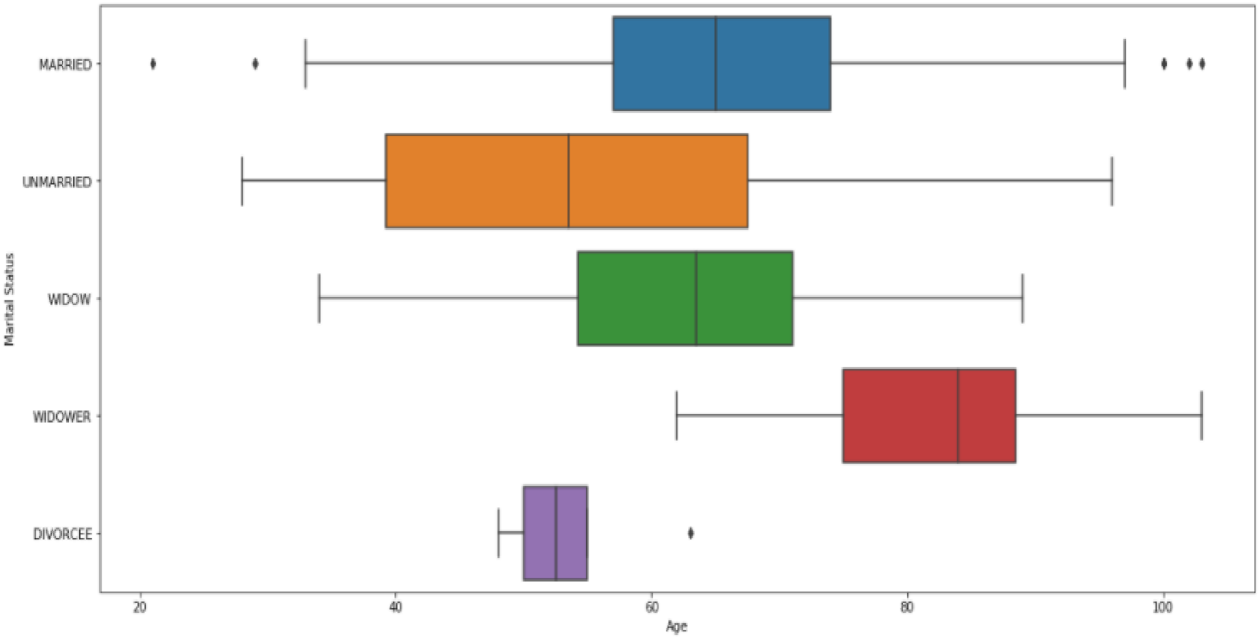


  - Let's keep only one of the formats (uppercase)
  - There is only one value for 'SINGLE', we will merge this with 'UNMARRIED'



COMPARISON OF MARITAL STATUS WITH AGE:

- Among married MPs, there are outliers. This could be because of early marriage for few and/ or data update missing for older MPs
- Almost 93% of the MPs are married while just under 4% are unmarried



ANALYSIS OF PROFESSION:

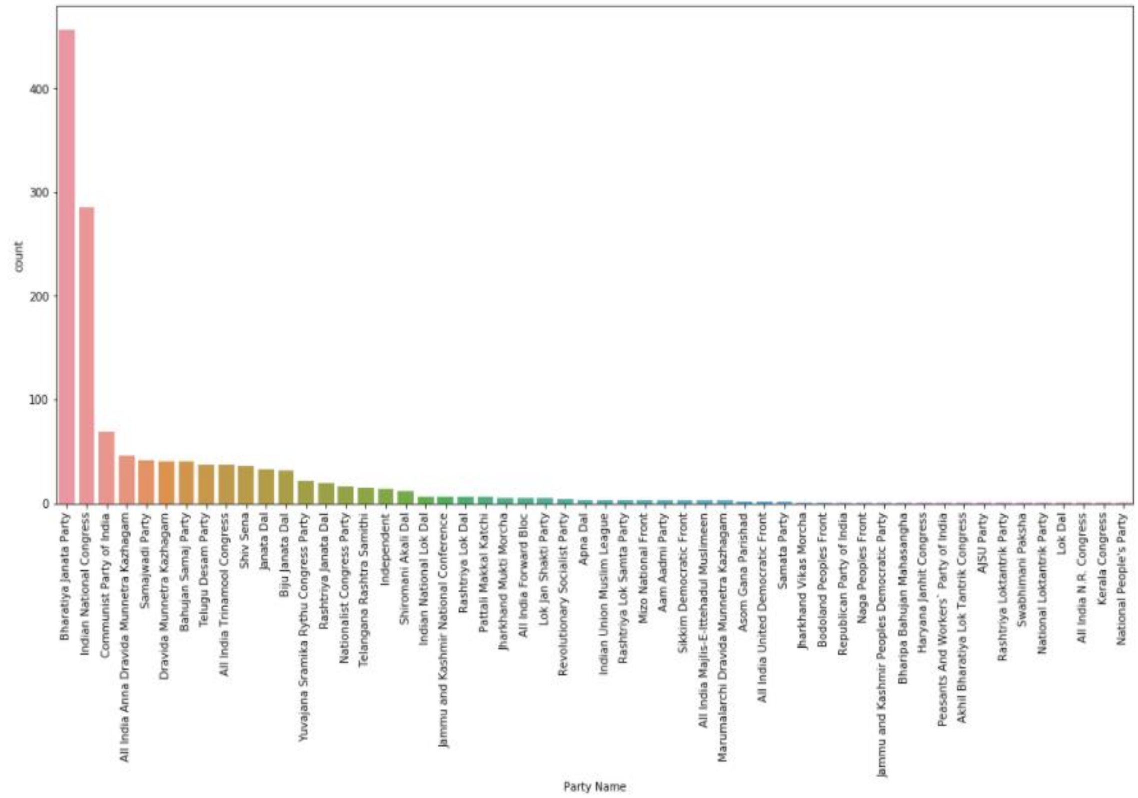
- Agriculture is the most common profession amongst the MPs with count of 227
- Social worker is the 2nd most common profession amongst the MPs with count of 148

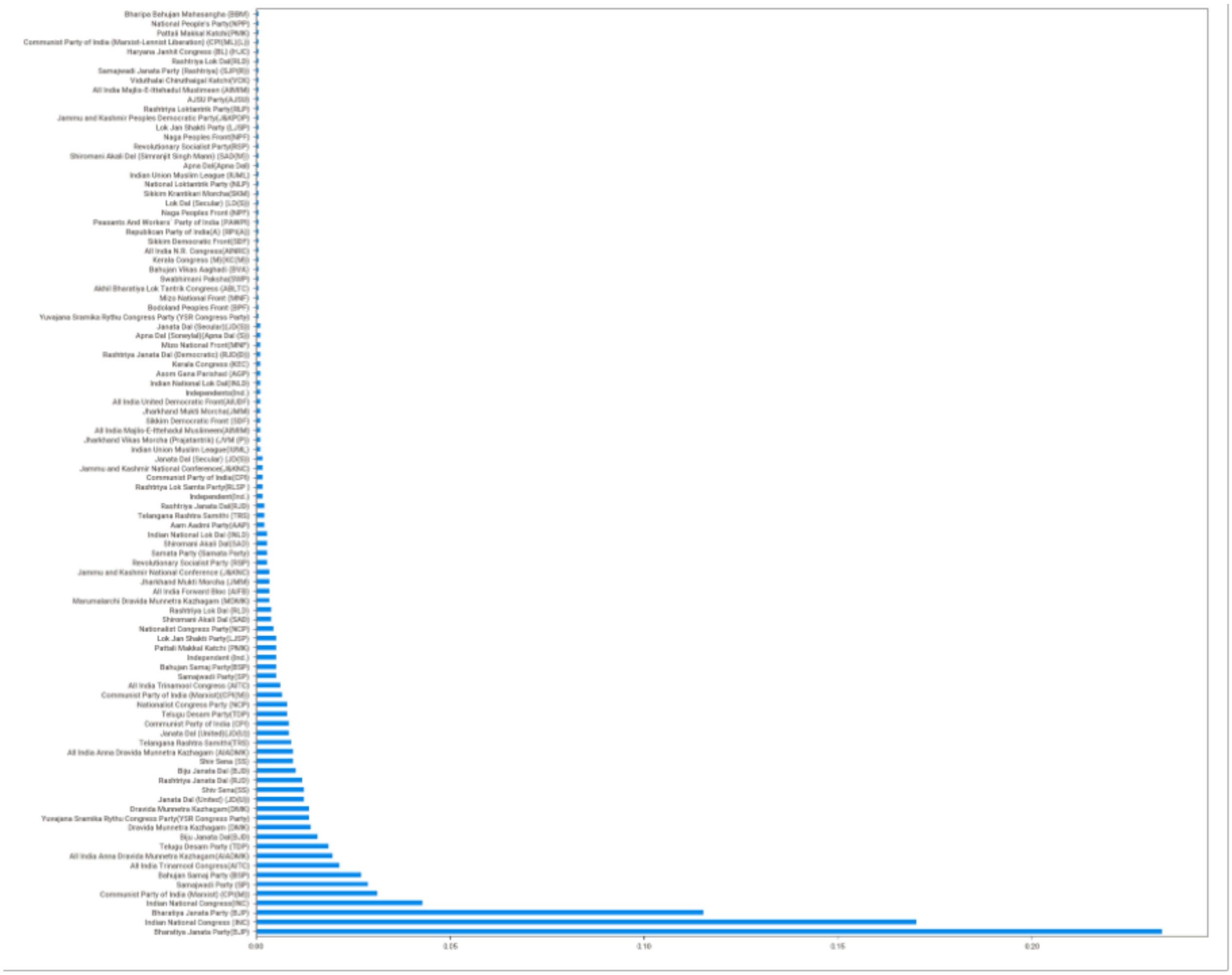
NULL VALUES IN THE DATASET:

- We have dropped data with null value below 2%

ANALYSIS FOR PARTY COLUMN NAME:

- We observed that each party's name is repeated in two formats: one in which acronym of the party within brackets is right after the text and another one where there is a space between the two.
- Also, categories Independents and Independent should be same, thus we have combined both
- BJP seems to be having the highest number of MPs at 34% followed by Congress (21%) and CPI (5%)
- Most of the parties have less than 3.5% presence among MPs across the years




Bharipa Bahujan Mahasangha (BBM)
National People's Party(NPP)
Pattali Makkal Katchi(PMK)
Communist Party of India (Marxist-Lennist Liberation) (CPI(ML)(L))
Haryana Janhit Congress (BL) (HJC)
Rashtriya Lok Dal(RLD)
Samajwadi Janata Party (Rashtriya) (SJP(R))
Viduthalai Chiruthaigal Katchi(VCK)
All India Majlis-E-Ittehadul Muslimeen (AIMIM)
AJSU Party(AJSU)
Rashtriya Loktantrik Party(RLP)
Jammu and Kashmir Peoples Democratic Party(J&KPDP)
Lok Jan Shakti Party (LJSP)
Naga Peoples Front(NPF)
Revolutionary Socialist Party(RSP)
Shiromani Akali Dal (Simranjit Singh Mann) (SAD(M))
Apna Dal(Apna Dal)
Indian Union Muslim League (IUML)
National Loktantrik Party (NLP)
Sikkim Krantikari Morcha(SKM)
Lok Dal (Secular) (LD(S))
Naga Peoples Front (NPF)
Peasants And Workers` Party of India (PAWPI)
Republican Party of India(A) (RPI(A))
Sikkim Democratic Front(SDF)
All India N.R. Congress(AINRC)
Kerala Congress (M)(KC(M))
Bahujan Vikas Aaghadi (BVA)
Swabhimani Paksha(SWP)
Akhil Bharatiya Lok Tantrik Congress (ABLTC)
Mizo National Front (MNF)
Bodoland Peoples Front (BPF)
Yuvajana Sramika Rythu Congress Party (YSR Congress Party)
Janata Dal (Secular)(JD(S))
Apna Dal (Soneylal)(Apna Dal (S))
Mizo National Front(MNF)
Rashtriya Janata Dal (Democratic) (RJD(D))
Kerala Congress (KEC)
Asom Gana Parishad (AGP)
Indian National Lok Dal(INLD)
Independents(Ind.)
All India United Democratic Front(AIUDF)
Jharkhand Mukti Morcha(JMM)
Sikkim Democratic Front (SDF)
All India Majlis-E-Ittehadul Muslimeen(AIMIM)
Jharkhand Vikas Morcha (Prajatantrik) (JVM (P))
Indian Union Muslim League(IUML)
Janata Dal (Secular) (JD(S))
Jammu and Kashmir National Conference(J&KNC)
Communist Party of India(CPI)
Rashtriya Lok Samta Party(RLSP)
Independent(Ind.)
Rashtriya Janata Dal(RJD)
Telangana Rashtra Samithi (TRS)
Aam Aadmi Party(AAP)
Indian National Lok Dal (INLD)
Shiromani Akali Dal(SAD)
Samata Party (Samata Party)
Revolutionary Socialist Party (RSP)
Jammu and Kashmir National Conference (J&KNC)
Jharkhand Mukti Morcha (JMM)
All India Forward Bloc (AIFB)
Marumalarchi Dravida Munnetra Kazhagam (MDMK)
Rashtriya Lok Dal (RLD)
Shiromani Akali Dal (SAD)
Nationalist Congress Party(NCP)
Lok Jan Shakti Party(LJSP)
Pattali Makkal Katchi (PMK)
Independent (Ind.)
Bahujan Samaj Party(BSP)
Samajwadi Party(SP)
All India Trinamool Congress (AITC)
Communist Party of India (Marxist)(CPI(M))
Nationalist Congress Party (NCP)
Telugu Desam Party(TDP)
Communist Party of India (CPI)
Janata Dal (United)(JD(U))
Telangana Rashtra Samithi(TRS)
All India Anna Dravida Munnetra Kazhagam (AIADMK)
Shiv Sena (SS)
Biju Janata Dal (BJD)
Rashtriya Janata Dal (RJD)
Shiv Sena(SS)
Janata Dal (United) (JD(U))
Dravida Munnetra Kazhagam(DMK)
Yuvajana Sramika Rythu Congress Party(YSR Congress Party)
Dravida Munnetra Kazhagam (DMK)
Biju Janata Dal(BJD)
Telugu Desam Party (TDP)
All India Anna Dravida Munnetra Kazhagam(AIADMK)
All India Trinamool Congress(AITC)
Bahujan Samaj Party (BSP)
Samajwadi Party (SP)
Communist Party of India (Marxist) (CPI(M))
Indian National Congress(INC)
Bharatiya Janata Party (BJP)
Indian National Congress (INC)
Bharatiya Janata Party(BJP)
0.00
0.05
0.10
0.15
0.20

PLACE OF BIRTH FOR MPS:

- We tried a number of different ways to extract the first word out of this field BUT we could not get it.
- Additionally, we tried to analyse ‘Constituency’ field but due to inconsistent unstructured data we were not able to extract State and City name for each MP.

CONCLUSION:

- We learnt how to do web scraping
- We cleaned the dataset
- We figured out that datasets don't just need cleaning but feature engineering as well
- We understood how to utilize visualizations
- We made inferences from the dataset which might help us in doing further analysis
- Average age of an MP in India is 65 years
- Most of the MPs are married
- Majority of the MPs have 1 or 2 kids
- Every 2nd MP is from BJP or Congress – two of the biggest political parties
- Agriculture and Social Work are dominant professions for the MPs

7. STRATEGY TO ENHANCE DATA

Considering the data is factual and not opinion or experienced based, direct crowd sourcing for data would not work in our case.

However, we tried to gather public opinion regarding few attributes, a quick summary is as below:

QUESTION ASKED	RESPONSE
Which party would have most no. of MPs in India post-independence?	64% Congress, 12% BJP, 14% No idea
What would be average age of an MP?	8% below 40, 63% 40 to 60, 28% 60+
What would be % of female MPs since independence?	78% below 25%, 22% (25-45%), 0% (45%+)
What would be % of MPs with at least graduation?	65% less than 10%, 32% (10-25%), 3% (25%+)
What % of MPs would be married?	90% (> 90%), 10% (less than 50%)
What would be average number of kids per MP?	4
What would be maximum number of kids for an MP?	12
- Who would be the MP with maximum no of kids?	Lalu Prasad Yadav

8. REFERENCES AND SOURCES

Loksabha

Wikipedia

geeks4geeks

stackoverflow

analyticsindiamag

towardsdatascience

medium

ISB LMS

ISB