UPSC CSE Prelims Exam Trend Report by Davies Obiekea

Description

This task saw the provision of past years' questions of the UPSC CSE Prelims examination from 2013 to 2023 and required a comprehensive analysis and presentation of findings after processes of Data Cleaning and Preparation, Trend Analysis, Visualization and Predictive analysis had been carried out.

Data Cleaning and Preparation

Python was employed in my analysis. I imported the necessary libraries for data processing and preparation. After the .csv file that had the data required for analysis was loaded, I checked to see what it contained. In my analysis, I realized that there 1100 data points and all data types were Object. I also noted the special characters and inconsistencies in the writing format which I thought might have some effect on the outcome of the predictions to be carried out on the data.

I began by first structuring the data. I stripped the text of leading or trailing white spaces, reduced all the texts to lower case for consistency and removed new line characters from the data. I standardized the subjects as there were repetitions, checked for missing and duplicated data-which by the way, were none.

I categorized questions based on topics and subjects and created a column titled 'Topic' in the process to aid with the analysis.

Trend Analysis

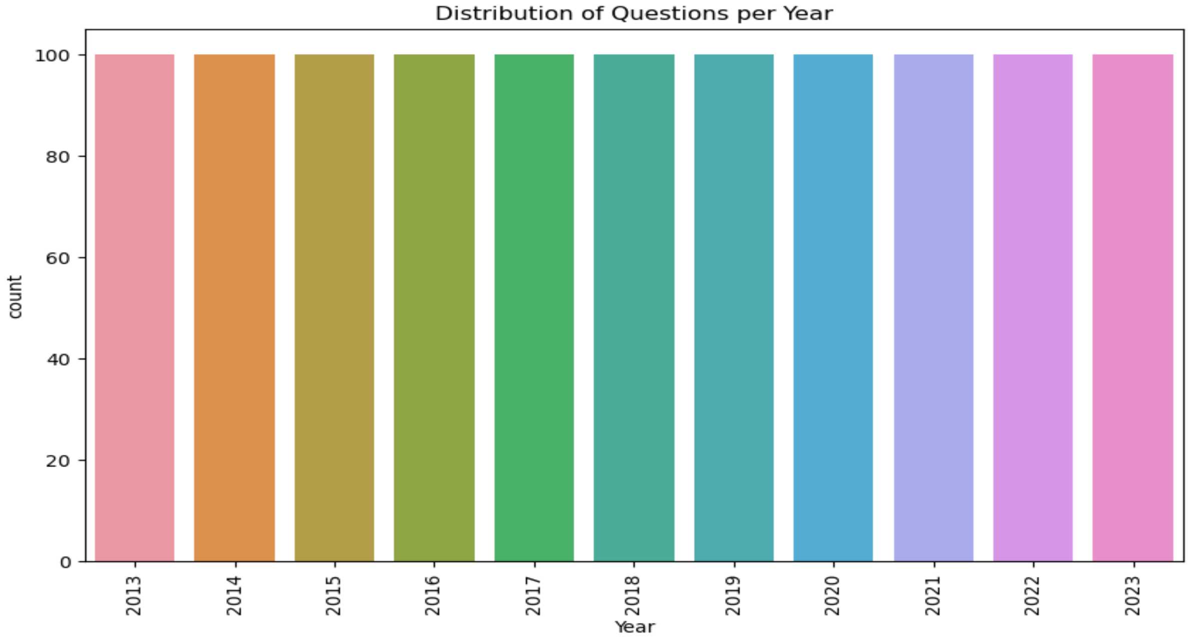
Using value counts and groupby from the pandas library, I noted that at 100 question per year, the subject_counts, which was a distribution of the various questions by subject from 2013 to 2023 are as follows:

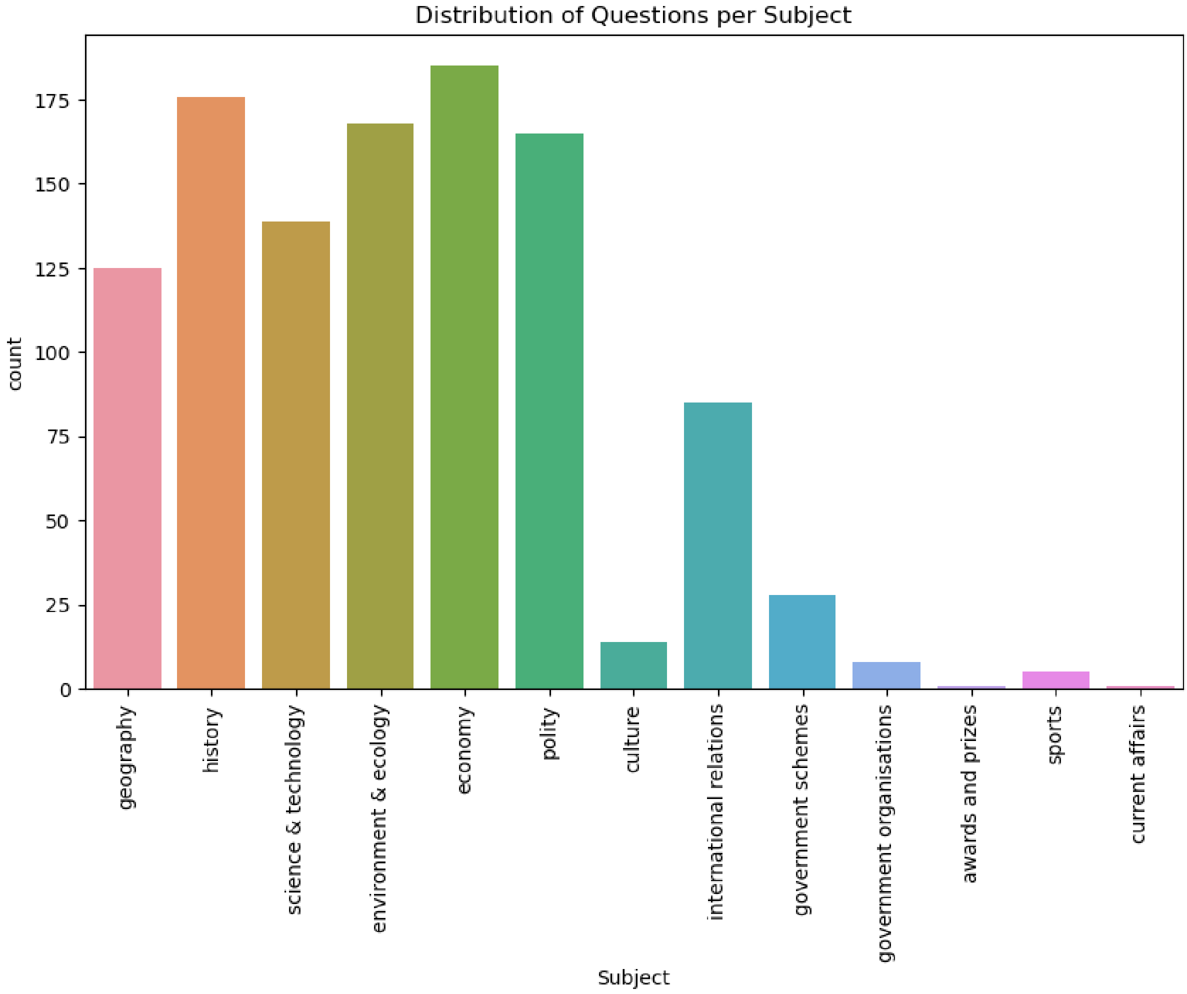
```
economy                     185
history                     176
environment & ecology       168
polity                      165
science & technology        139
geography                   125
international relations      85
government schemes           28
culture                      14
government organisations      8
sports                        5
awards and prizes             1
current affairs               1
```

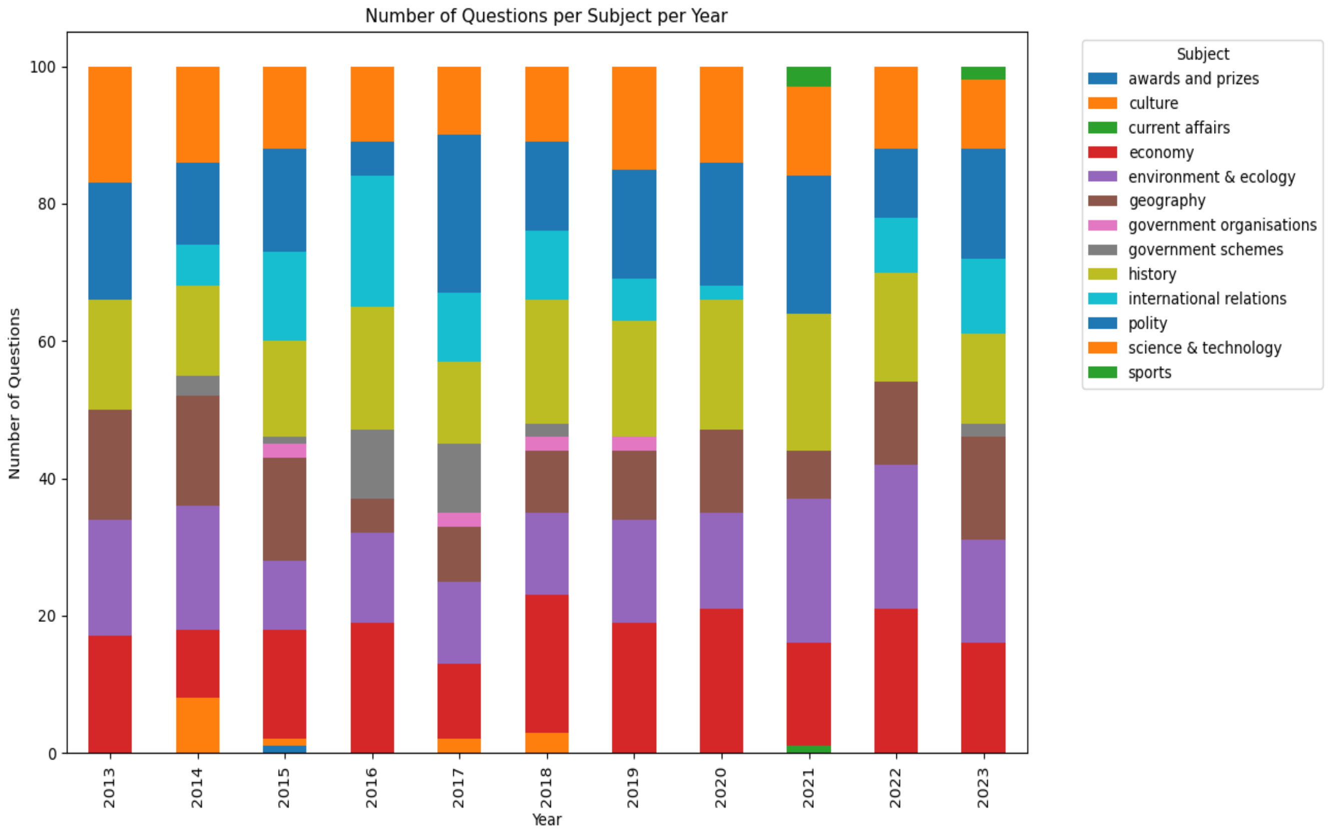
The analysis showed that over the years Economy, History, Environment and Ecology, Polity, Science and Technology and Geography were with the most questions in the distribution with Awards and Prizes and Current Affairs with the least.

Bar Charts/Count Plots

Find below Visualizations to further confirm the trends;





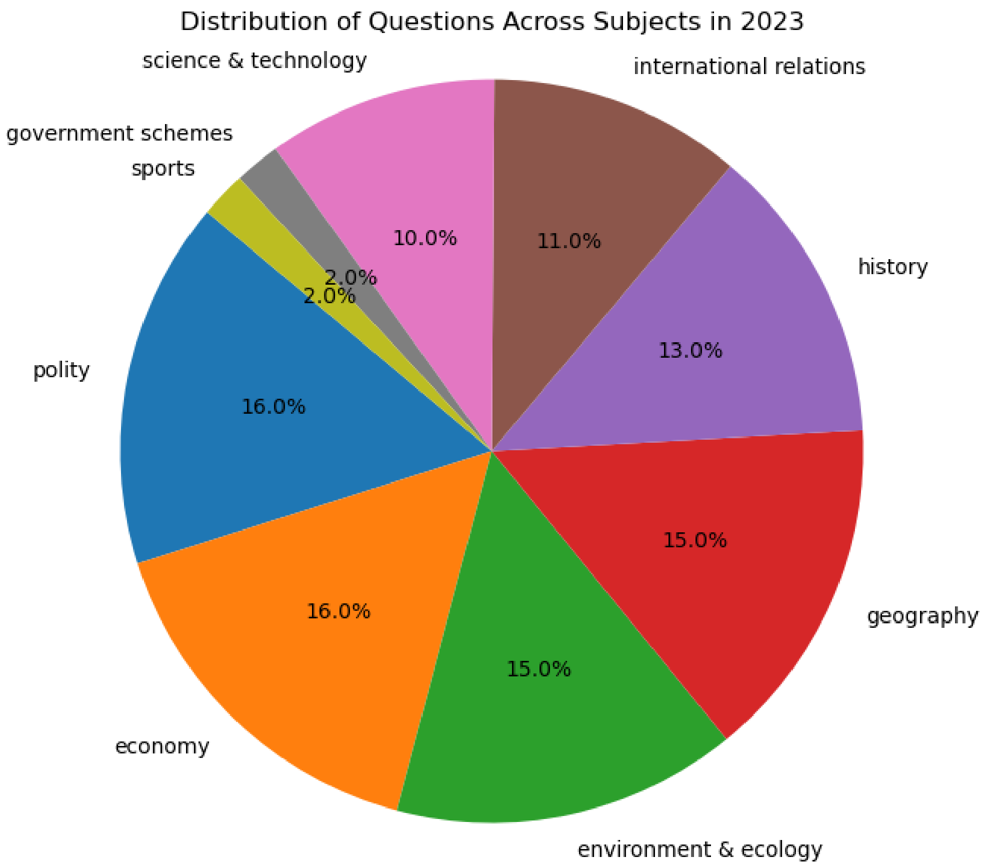


This plot above, represents Number of Questions per Year per Subject. It would be seen, that as calculated from the value counts and groupby operations, Economy, History, Environment and Ecology among others lead the pack of subjects from which questions administered for the exams.

Pie Chart

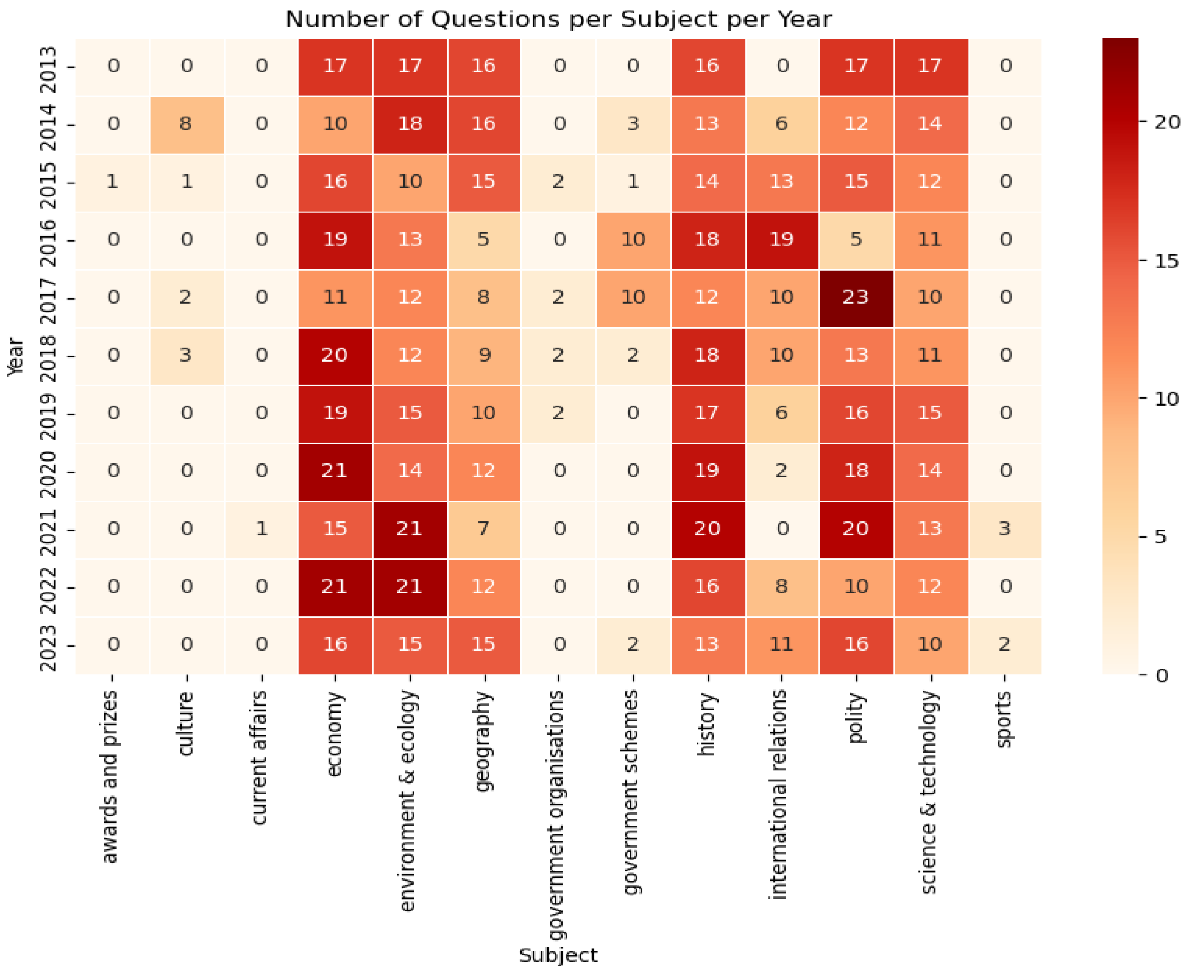
I also plotted a pie chart to portray the distribution of the questions per subject in the year 2023. The top 6 (as I would like to refer to them) Economy and Polity had 16% slice each, Environment and Ecology, and Geography had 15% each and History followed with 13% totalling about 75% of the distribution.

See Pie Chart below;



Heat Map

The heat map also revealed support of the historical trends with heavier intensity (darker red) on the topics/subjects with higher distribution. For example, in 2017, Polity had the highest distribution from 2013 to 2023 with value 23.



Year	awards and prizes	culture	current affairs	economy	environment & ecology	geography	government organisations	government schemes	history	international relations	polity	science & technology	sports
2013	0	0	0	17	17	16	0	0	16	0	17	17	0
2014	0	8	0	10	18	16	0	3	13	6	12	14	0
2015	1	1	0	16	10	15	2	1	14	13	15	12	0
2016	0	0	0	19	13	5	0	10	18	19	5	11	0
2017	0	2	0	11	12	8	2	10	12	10	23	10	0
2018	0	3	0	20	12	9	2	2	18	10	13	11	0
2019	0	0	0	19	15	10	2	0	17	6	16	15	0
2020	0	0	0	21	14	12	0	0	19	2	18	14	0
2021	0	0	1	15	21	7	0	0	20	0	20	13	3
2022	0	0	0	21	21	12	0	0	16	8	10	12	0
2023	0	0	0	16	15	15	0	2	13	11	16	10	2

Building the Model

To build the model, I carried out feature extraction using the Term Frequency-Inverse Document Frequency (Tfidf) Vectorizer. This converted my independent feature 'Question' which were texts to arrays with number representation. I chose 'Subject' as my 'Target' variable seeing as the task was to forecast distribution of questions along subject lines.

The data was split into training and test using 75% for training and 25% for test. I determined that the problem was a classification type and so I used the Random Forest and Naïve Bayes Classifiers to train the data and for model prediction.

The classification report generated scored accuracy of both models as follows;
Random Forest Classifier: 60%
Naïve Bayes Classifier: 71%

This score was acceptable to me so I went ahead to forecast potential distribution of questions in 2024.

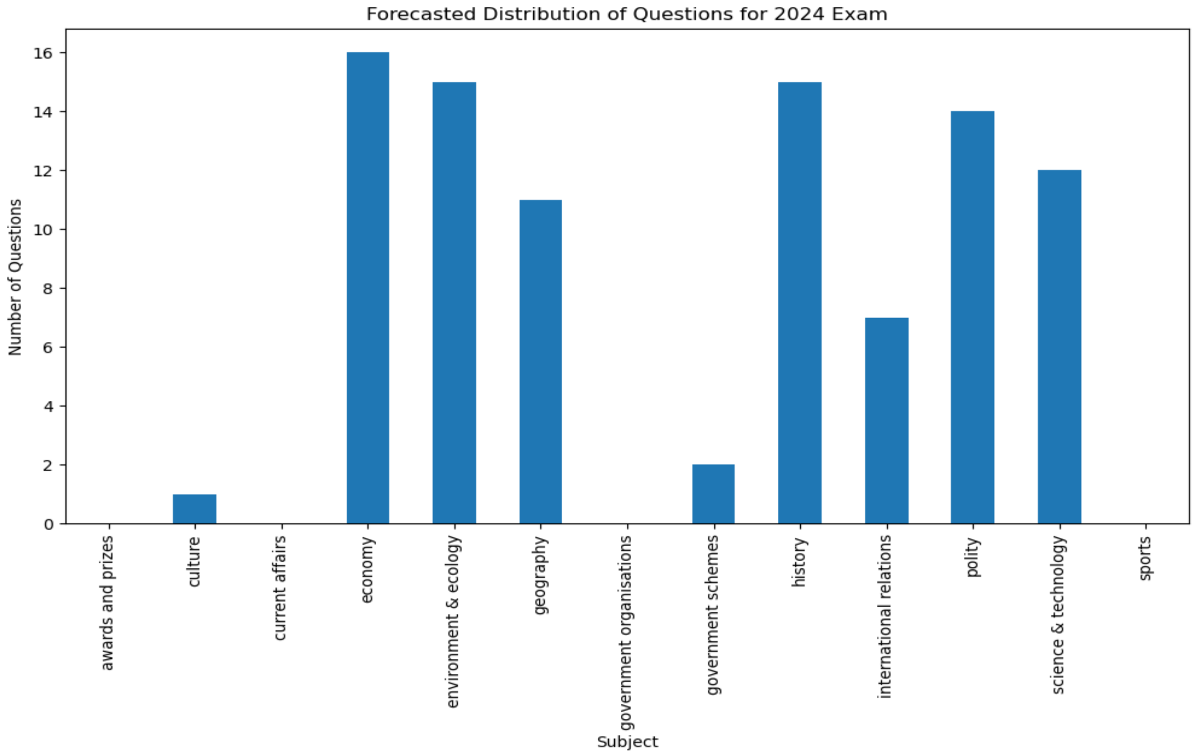
I began by calculating average distribution which was the mean of the distribution of questions per subject per year as derived during the trend analysis. Then I used it to calculate the probabilities which was average distribution divided by the sum of the average distribution. The results are as follows;

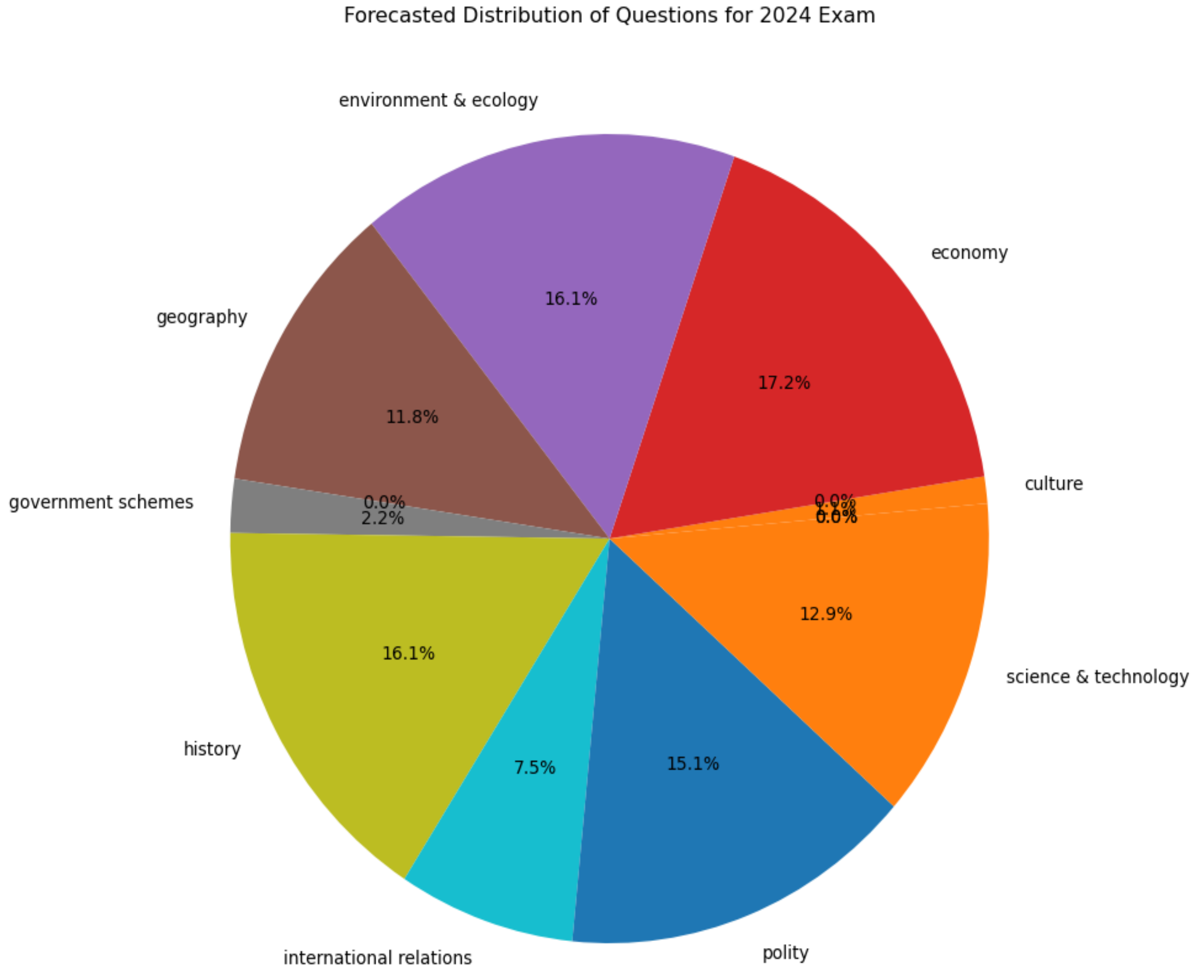
```
Subject
awards and prizes          0.000909
culture                    0.012727
current affairs            0.000909
economy                    0.168182
environment & ecology      0.152727
geography                  0.113636
government organisations   0.007273
government schemes         0.025455
history                    0.160000
international relations    0.077273
polity                     0.150000
science & technology       0.126364
sports                     0.004545
```

The expected number of questions for 2024 was marked at 100 following historical trends. I multiplied the probabilities by this 100 which provided the potential distribution of questions for 2024 as follows;

```
Subject
awards and prizes            0
culture                      1
current affairs              0
economy                     16
environment & ecology       15
geography                   11
government organisations     0
government schemes           2
history                     15
international relations      7
polity                      14
science & technology        12
sports                       0
```

Again, as seen in past trends the top six topics recorded higher distribution which is represented in the Bar and Pie Charts below;


Forecasted Distribution of Questions for 2024 Exam


Forecasted Distribution of Questions for 2024 Exam

The charts above visualizes the forecasted distribution, making it easy to understand the proportional representation of each subject.

The mean and standard deviations which are statistical means to justify the predictions were calculated as follows;

```
Mean distribution of questions per subject:
 Subject
awards and prizes            0.090909
culture                      1.272727
current affairs              0.090909
economy                     16.818182
environment & ecology       15.272727
geography                   11.363636
government organisations     0.727273
government schemes           2.545455
history                     16.000000
international relations      7.727273
polity                      15.000000
science & technology        12.636364
sports                       0.454545

Standard deviation of distribution per subject:
 Subject
awards and prizes           0.301511
culture                     2.453198
current affairs             0.301511
economy                     3.736795
environment & ecology       3.635682
geography                   3.854160
government organisations    1.009050
government schemes          3.830500
history                     2.683282
international relations     5.780846
polity                      4.919350
science & technology        2.203303
sports                      1.035725
```

The mean represents the average number of questions per subject over the years. The standard deviation indicates the variability or spread of the number of questions per subject.

Topics with low standard deviation (e.g., Awards and Prizes, Current Affairs) have more consistent question counts and are very few, in this case 1. Topics with higher standard deviation (e.g., Economy, Polity) exhibit more variability. They could vary significantly in number.

I compared the 2024 forecast to the Historical average by concatenating the series. These are the results;

```
Comparison of Forecasted and Historical Average Distribution:
                           Forecast 2024  Historical Average
Subject
awards and prizes                      0            0.090909
culture                                1            1.272727
current affairs                        0            0.090909
economy                               16           16.818182
environment & ecology                 15           15.272727
geography                             11           11.363636
government organisations               0            0.727273
government schemes                     2            2.545455
history                               15           16.000000
international relations                7            7.727273
polity                                14           15.000000
science & technology                  12           12.636364
sports                                 0            0.454545
```

Comparing the forecasted distribution with the historical average helps validate the predictions.

As can be seen in the final comparison between the 2024 Forecast and the Historical average (which is an aggregated average of distribution of questions over the last ten years), historical trends do align with our predictions. Conclusively, both models- The RandomForest Classifier and the Naive Bayes model, at 60% and 71% accuracy respectively performed well.

The analysis of mean and standard deviation helps in understanding the consistency and variability, providing a solid foundation for the forecast.