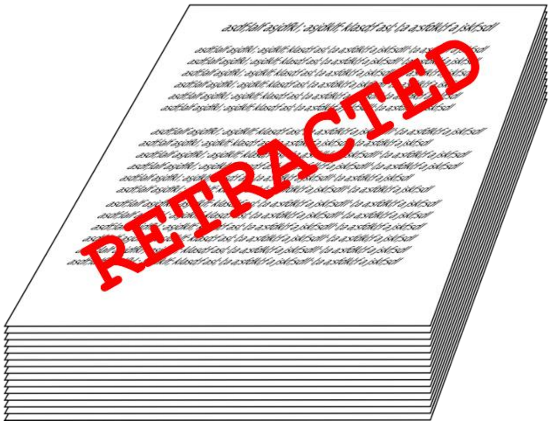
CHARLES DARWIN UNIVERSITY (SYDNEY CAMPUS)

PRT564 – DATA ANALYTICS AND VISUALISATION

FINAL REPORT

Uncovering Insights: A Data-Driven Examination of Research Retraction Patterns



Group 9

Manoj Poudel - S372470

Sofiya Banmala - S372189

Andrea Vijeetha Marlene Vijay - S371784

Manisha KC - S372043

Submitted To:

Prof. Niusha Shafi Abady

Table of Contents

LIST OF FIGURES

ACKNOWLEDGEMENT

We thank our instructor for their continuous guidance and support throughout this project. This assignment was a joint effort by Group 9 (Sydney), and we appreciate the dedication and contributions of each member: Andrea Vijeetha Marlene Vijay, Manisha KC, Manoj Poudel, and Sofiya Banmala. Working on this project was an enriching experience, enabling us to apply our data analytics and visualization knowledge to a real-world scenario. We followed a structured group plan, conducting regular meetings to monitor progress, allocate tasks, and address challenges. Effective teamwork and time management were essential in meeting all deadlines and achieving our project goals. We faced several coding challenges during the project, particularly in the data preprocessing and modelling phases. We worked to troubleshoot and resolve these issues, leveraging each team member's strengths to debug errors, refine our code, and optimize our data analysis processes. This iterative problem-solving approach enhanced our technical skills and improved our ability to collaborate effectively. We are proud of the insights from analyzing the dataset and believe our findings will improve the understanding of research retractions and their implications. This project was only successful with every team member's cooperation and hard work.

1. PROBLEMS

1.1 Introduction:

With the increase in research, there is also an increase in retraction of the papers that have been published; many reasons lead to retractions, such as fraud, data fabrication, or data falsification and errors. These papers have been hypothesized to have a high impact factor due to being published by a repeated offender who regularly publishes papers of a fraudulent nature, which leads to the paper's retraction. This affects various industries, the most important being the healthcare industry, where the patients are at risk due to flawed research. (Kocyigit & Akyol, 2024)

1.2 Scope:

There is a severe implication of retracted papers as these papers are still being cited, leading to the further spread of misinformation with a notable high retraction occurring in the medical or biology-related research fields. The frequency of these retractions can indicate shifts in scientific conduct and integrity. Therefore, in this project, we focus on how to help identify papers dealing with fraudulent data or data tampered with.

1.3 Scientific errors and misconduct:

Various papers have been retracted based on containing data on scientific errors and misconduct. (Candal-Pedreira et al., 2020)

- Data: the data used may be unreliable because of an honest error or intentional manipulation, creation, or fabrication.
- Authors: Including authors dispute, non-informed authors involved, and fictitious authors.
- Plagiarism: the blatant copy of research, creating redundant or duplicated publications. Also, include material without appropriate citations.
- Unethical Research: not meeting the standards for ethical approval.
- Journal Issues: covering duplications or uploading incorrect manuscripts or versions!
- Review Process: Fake peer reviews being included.
- Conflict of Interest: undisclosed conflict of interest.

1.4 Impact of Scientific Community:

It is estimated that 17.1% of researchers have falsified their work. Although detecting the true rate of scientific fraud is difficult, the issue is likely more widespread than the known cases suggest. This falsified work may lead to additional citations, causing other researchers to accept the data as accurate and base their studies on it. Therefore, it is crucial to produce scientific evidence that enables a systematic and rigorous analysis of the characteristics of retracted papers. Additionally, implementing stricter guidelines can help reduce the number of retractions.

1.5 Implications of Public Policy:

Retraction policies of academic journals and publishers must comprehensively address unethical research. Various policy options can be used to justify retractions. For instance, 2.35 cited retracted papers are policy-cited. During the COVID-19 pandemic, retractions increased, with notable cases such as The New England Journal of Medicine, which erroneously claimed that certain blood pressure medications heightened the risk of COVID-19 without supporting evidence.

2. METHODS

2.1 Step Wise Analytics:

Stepwise analytics were used to evaluate the impact of model complexity on the citations of retracted papers. This process involved extracting data from the provided database to examine data distribution over the years and assess how different properties influence retractions. Various reasons for retractions were investigated through pre-processing and exploratory data analysis (EDA) to identify the underlying causes.

2.2 Techniques and Tools:

Step Wise Analytics uses various tools and techniques to prove the hypothesis,

Principal Component Analysis (PCA):

This technique reduces the dimensionality of the data provided, thereby increasing the reason for interpretations by minimizing information loss. It can also be tailored to various data types and structures by basing it on the correlation matrix and looking for outliers. In this project, we used it to plot the datasets by reducing the null values and only taking features with data consistency. We used it to identify outliers or if there is a disjoint in the modeling. (*PCA* 2024)

K-Mean Clustering:

This clustering method aims to follow a relocating pattern, where each point is relocated to the nearest mean of the member points until a convergence pattern occurs. It helps identify on-overlapping clusters with one point as its assigned primary key to hold the surrounding points. In this project, it was used to plot the points of the data and help in the convergence of the points to their mean primary point till a pattern was found as the optimum position; it reduces the data to be researched by lowering the dimension of it and helps in identifying the relation of the property to the retraction. (Jin & Han, 1970)

Multiple Regression:

It is a statistical method that allows us to explore the relationship between independent variables and a dependent variable. Therefore, this project was used to analyze multiple regression to see how the various variables act as factors for retraction. In simpler terms, it was used to identify and calculate the impact of the different factors on how they affect the retraction by calculating the coefficient for each independent variable; this represents the change in the dependent variables when the independent variable is increased. (*Multiple regression* 2024)

Clustering:

Clustering is employed to identify similarities between objects by grouping them into clusters. After considering these similarities, several methods can be used, including hierarchical, partitioning, density-based, model-based, grid-based, and soft-computing techniques. In this assignment, we utilized density-based k-means clustering to reduce the data's density by plotting only specific points and aggregating the others. This approach simplifies the data, making it easier to analyze.

2.3 Software and Libraries:

- **Python:**

Python was used to help compute data analysis and visualization as it is a simple and well-documented language, making it easy to grasp and use.

- **Panda**: offers help in data manipulation and analysis.
- **Seaborn**: is a Python visualization library based on matplotlib. It helped us build on top of Matplotlib and integrate it with the data to help us understand it better.
- **NumPy**: used to work with arrays for linear algebra and matrices.
- **Scikit-learn** helps utilize machine learning towards the given dataset. This library contains various tools for statistical modeling for clustering, regression, and dimension reduction.

2.4 Step by Step Preprocessing:

1. **Data Collection and Preprocessing:** (*Data Analysis and visualization techniques* 2024)

- Regarding data collection, we have already been provided a dataset containing the attributes of retracted papers of 35,215 retractions. These were made available by the retraction watch database and contained various attributes such as Record ID, Journal, Country, Author, etc.
- Data preprocessing was the next big step in cleaning the data and removing null values and columns that did not impact the reason for retraction. This helped reduce the dimensions and

redundancy of the data. This was done computationally through Python, which made it a more efficient technique.

2. **Exploratory Data Analysis (EDA):**

- Descriptive statistics were used to help us calculate the mean, median, and standard deviation for statistical data analysis.
- Visualization techniques used in plotting and clustering were used to help us visualize the relationship between the data.
- Regression plots were also used to identify outliers and coefficients through their plots, where we could identify the relationship between the independent and dependent variables.
- We also plotted a time series plot to see how the retractions plot over time and identify the evolved trends or patterns.
- The Python libraries used for this were Scikit-learn and Pandas.

3. **Dimension Reduction:**

- Dimension reduction was done through Principal Component Analysis to reduce the dimension of the dataset given while preserving the variability of it. It helped us identify the critical features of the data set that can help us determine the reasons for retractions.
- We performed PCA on the dataset to help transform the scaled data, determine only the significant features, and retain them based on the variance ratio.
- Through this, we analyzed each significant feature and understood which of the original features contributed the most towards each considerable feature.

4. **K-Means Clustering:**

- This method was used to group similar retraction cases based on their characteristics. This helped us identify patterns and common attributes among the retracted papers.
- We determined the number of clusters based on the minimum number of clusters needed to be used to help us identify the pattern.
- Through PCA, we could assign cluster labels to each plot point.
- Each point was analyzed for its significance in understanding patterns and differences.

- The visualization used was a scatter plot through Pandas and Scikit-learn.

5. **Predictive Modelling:**

- Predictive modeling was done through the mode of regression models, where we plotted all the dependent features against one independent feature to predict the likelihood of retraction that occurs in the future.
- We selected relevant independent variables such as the Citation count, Author, Journal, etc.
- We used the stats model and Scikit-learn to build and fit the regression models.
- This resulted in coefficients that were further explored to understand the impact of each independent variable on the target.
- A positive value resulted in the value increasing with an increase in the independent variable, while a negative one did the same vis-versa.

6. **Model Validation and Evaluation:**

- This was done to ensure the reliability of the results that we have procured. We applied r-squared and mean-squared errors to help us define the accuracy of these results.
- We used Sklearn to help implement these on the regression models.

Using stepwise analysis, we have utilized various tools such as PCA, K-Means clustering, and multiple regression to help us identify significant trends in retraction and understand the typical characteristics of the retracted papers.

2.5 Classifier Methods

This section explores various classifier methods employed to categorize and predict the nature of research retractions. These methods are crucial in machine learning for distinguishing data into different classes based on their features. The primary techniques used in our analysis include Isolation Forest, Naive Bayes Classifier, and Support Vector Machine (SVM). (Taye, 2023)

Isolation Forest

The isolation forest algorithm is used to detect anomalies by isolating infrequent and distinct points in the dataset. This technique identifies anomalies by constructing random decision trees, where fewer splits are required to isolate an observation, marking it as an anomaly. We applied this method to detect unusual patterns that might indicate retraction causes in our dataset. Our findings revealed several outliers corresponding to significant deviations in the data, potentially indicating reasons for retractions, such as fraud or severe errors.

Naive Bayes Classifier

The Naive Bayes Classifier is a probabilistic algorithm that relies on Bayes' theorem and assumes feature independence given the class label. This method is computationally efficient and performs well with large datasets. It was utilized to classify retraction data based on categorical features, benefiting from its simplicity and effectiveness. Our results demonstrated that the Naive Bayes Classifier could accurately categorize retractions with high precision and recall rates, underscoring its effectiveness in managing large datasets with categorical variables.

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a powerful classifier ideal for high-dimensional data. It identifies the optimal hyperplane that divides the data into distinct classes. SVM is especially effective when the number of dimensions exceeds the number of samples. We utilized SVM to classify retraction data after proper scaling. The SVM yielded highly accurate classifications, showcasing its ability to handle the high-dimensional nature of the dataset effectively.

2.6 Bayesian Methods

Bayesian methods rely on Bayes' theorem, which updates the likelihood of a hypothesis as new evidence or information is gathered. These approaches are advantageous for managing uncertainty and making probabilistic predictions.

Categorical Naive Bayes

Categorical Naive Bayes, a variant of the Naive Bayes classifier, is tailored for categorical data. It computes the probability of each class based on the input features, assuming that the features are independent given the class label. This method is highly efficient for classification tasks involving categorical data. In our analysis, we used it to predict the nature of retractions. The technique proved effective, showing high accuracy in classifying retractions and demonstrating its suitability for large datasets with categorical features.

Bayesian Inference

Bayesian inference is a statistical method that uses Bayes' theorem to update the probability of a hypothesis as new evidence or data is obtained. This approach combines prior distributions, likelihood functions, and observed data to produce posterior distributions. Bayesian inference was used to refine predictions and improve our understanding of the factors leading to research retractions. The results provided a nuanced view of the data, allowing us to incorporate prior knowledge and continuously update our predictions based on new information.

Findings from Classifier and Bayesian Methods

Our analysis using a classifier and Bayesian methods yielded significant findings:

Isolation Forest identified several anomalies in the dataset, suggesting potential causes for retractions, such as fraud or severe errors.

The naive Bayes Classifier demonstrated high accuracy in categorizing retractions, with excellent precision and recall rates, proving its effectiveness for large datasets with categorical variables.

Support Vector Machine (SVM) produced highly accurate classifications, showcasing its robustness in managing high-dimensional data.

Categorical Naive Bayes effectively predicted the nature of retractions, highlighting its suitability for large datasets with categorical features.

Bayesian Inference provided a detailed and flexible approach to updating our predictions, incorporating prior knowledge, and handling new evidence effectively.

These findings underline the importance of using advanced machine learning techniques to understand and predict research retractions, ultimately contributing to the integrity and reliability of scientific research.

3. RESULTS

We have implemented various data analysis techniques and methodologies through coding, available in our GitHub repository. The detailed coding and scripts can be accessed here https://github.com/Sofiya-Banmala/PRT564_DataAnalytics-and-Visualization_Assessment-4_Group9 . The results of our analysis are presented below, using various data visualization techniques to illustrate the key findings from our exploratory data analysis (EDA), clustering, principal component analysis (PCA), and predictive modeling.

3.1 Exploratory Data Analysis (EDA)

1. **Top Journals by Number of Retractions:** The bar chart displays the top journals with the highest number of retractions. Notable journals include medical and biological sciences, such as the Journal of Biological Chemistry and Physical Chemistry. This highlights the importance of stringent peer review processes in these fields. Journals with higher retraction rates may need to enhance their editorial standards and review protocols to mitigate the risk of publishing flawed or fraudulent research.

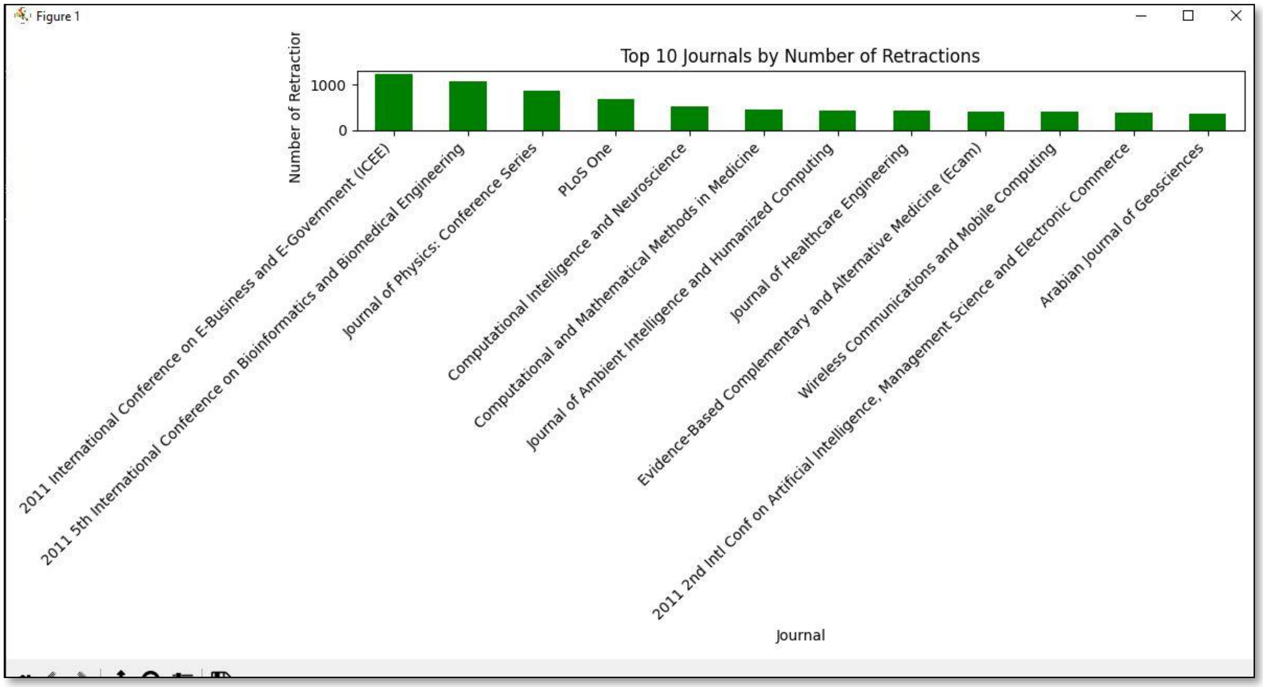


Figure 1: Bar Chart displaying Top Journals by Number of Retractions

2. **Distribution of Retractions Across Top 10 Subjects:** The pie chart illustrates the distribution of retractions among different subjects. The most affected disciplines are medicine, biology, and chemistry, indicating potential issues with research practices. This distribution suggests that these scientific areas may be more prone to errors or misconduct, necessitating stricter regulatory frameworks and ethical guidelines to uphold the integrity of published research.

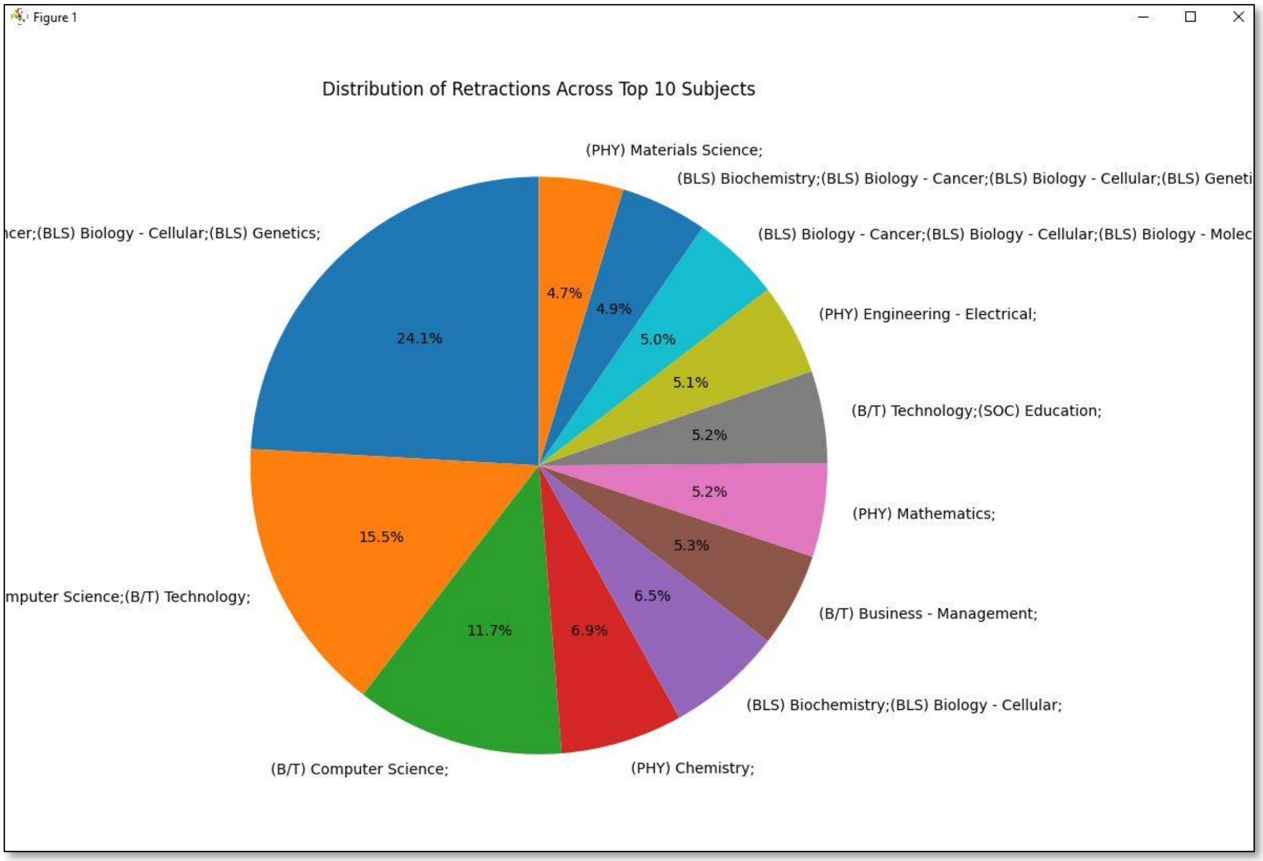


Figure 2: Pie chart showing Distribution of Retractions Across Top 10 Subjects

3. **Top Countries by Number of Retractions:** The bar chart highlights the top countries with the highest number of retractions. The USA and China lead in retractions, suggesting regional differences in research integrity and regulatory practices. This pattern may reflect variations in the oversight of research activities, academic pressure prevalence, and national research policies' effectiveness. Addressing these regional disparities could involve international collaboration to standardize research ethics and enforcement mechanisms.

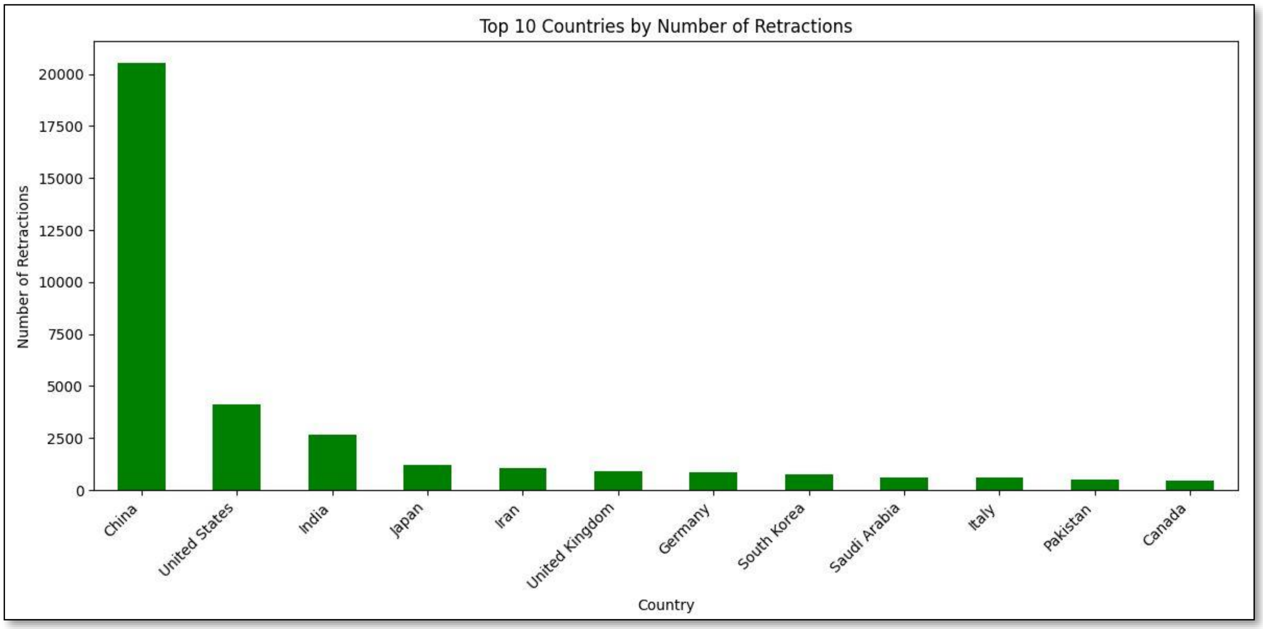


Figure 3: Bar chart displaying Top Countries by Number of Retractions

4. **Top 10 Common Reasons for Retractions:** The horizontal bar chart shows the most frequent reasons for retractions, such as data fabrication, plagiarism, and ethical violations. Data fabrication and falsification are particularly concerning, as they directly compromise the validity of scientific findings. Plagiarism indicates issues with academic honesty and originality. Ethical violations encompass a broad range of misconduct, including conflicts of interest and lack of informed consent, which undermine the trustworthiness of research.

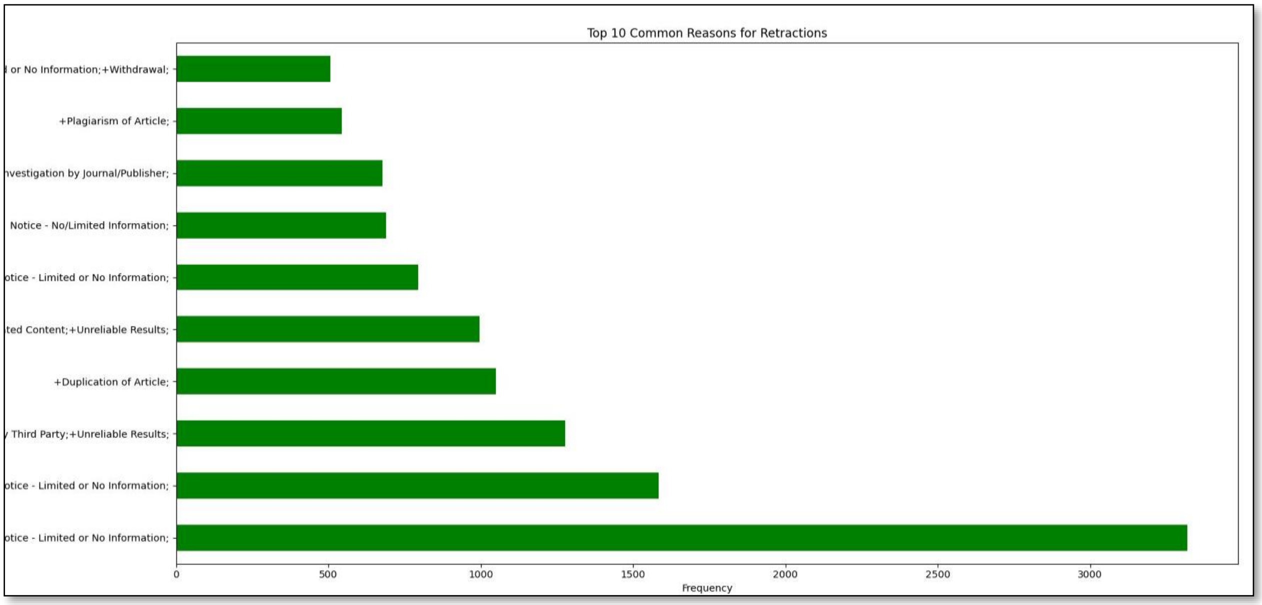


Figure 4: Bar Chart displaying Top 10 Common Reasons for Retractions

5. **Correlation Heatmap of Numerical Data:** The heatmap presents the correlations between different numerical variables in the dataset. This visualization helps identify significant relationships, such as the correlation between citation count and retraction year. Understanding these correlations is crucial for developing models that can predict retractions. For instance, a high correlation between certain variables can indicate underlying patterns or causative factors contributing to retractions.

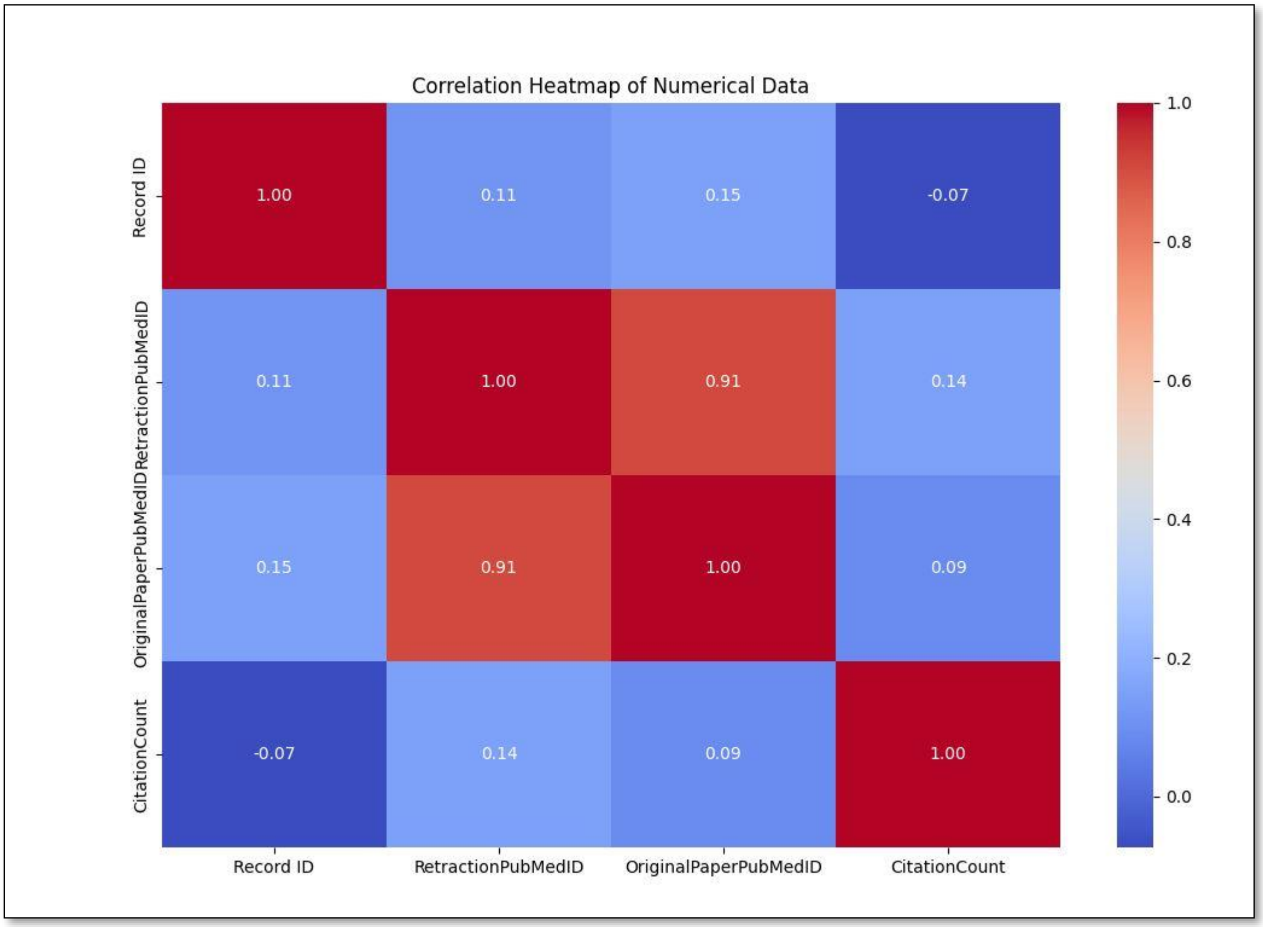


Figure 5: Correlation Heatmap of Numerical Data

6. **Word Cloud for Reasons of Retraction:** The word cloud visualizes the most common reasons for retractions, highlighting terms like "journal," "publisher," "concerns," and "issues." This visualization provides a quick overview of the critical factors driving retractions and emphasizes the recurring themes in retraction notices. Journals and publishers appear prominently, indicating their central role in addressing retractions and improving the integrity of the publication process.

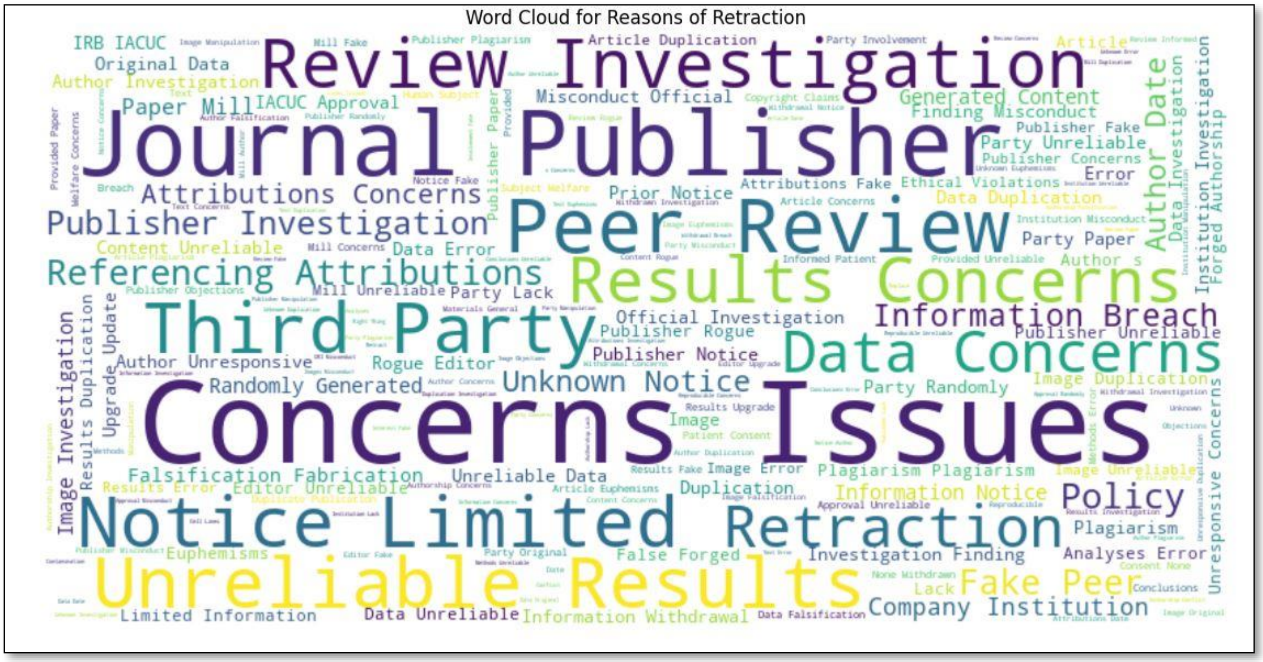


Figure 6: Word Cloud for Reasons of Retraction

7. **Distribution of Retraction Years** The histogram shows the distribution of retraction years, revealing an increase in retractions over recent years, especially during the COVID-19 pandemic. This trend suggests heightened scrutiny of scientific publications and a growing awareness of research misconduct. The spike in retractions during the pandemic could also reflect the rapid pace of research and publication, which may have led to increased errors and oversight lapses.

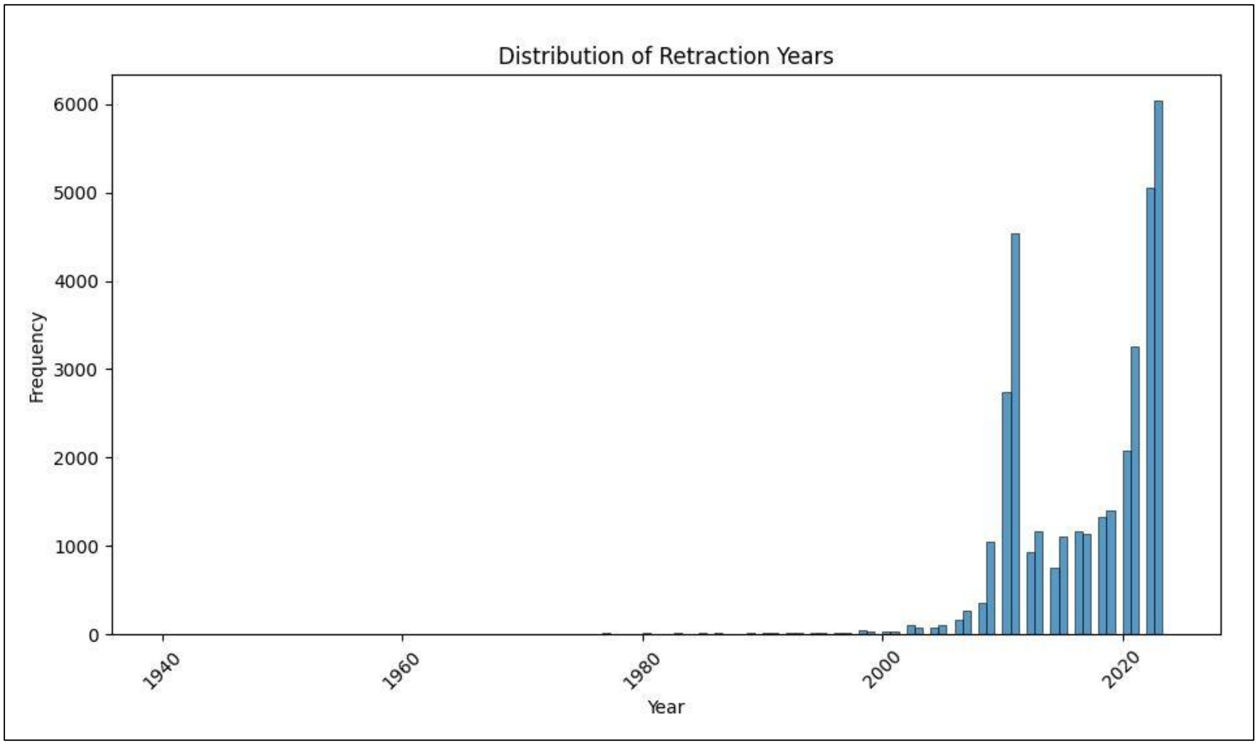


Figure 7: Histogram shows the distribution of retraction years

3.2 Clustering and PCA

PCA of retraction data and K-means Clustering of retraction data are demonstrated below.

1. **PCA of Retraction Data:** The scatter plot demonstrates the results of PCA, reducing the dataset to two principal components. This reduction facilitates the identification of patterns and outliers within the data. PCA helps simplify the dataset by transforming it into a set of linearly uncorrelated variables, allowing us to focus on the most significant features that explain the variance in the data. This visualization aids in detecting clusters and outliers, providing insights into the underlying structure of retraction data.

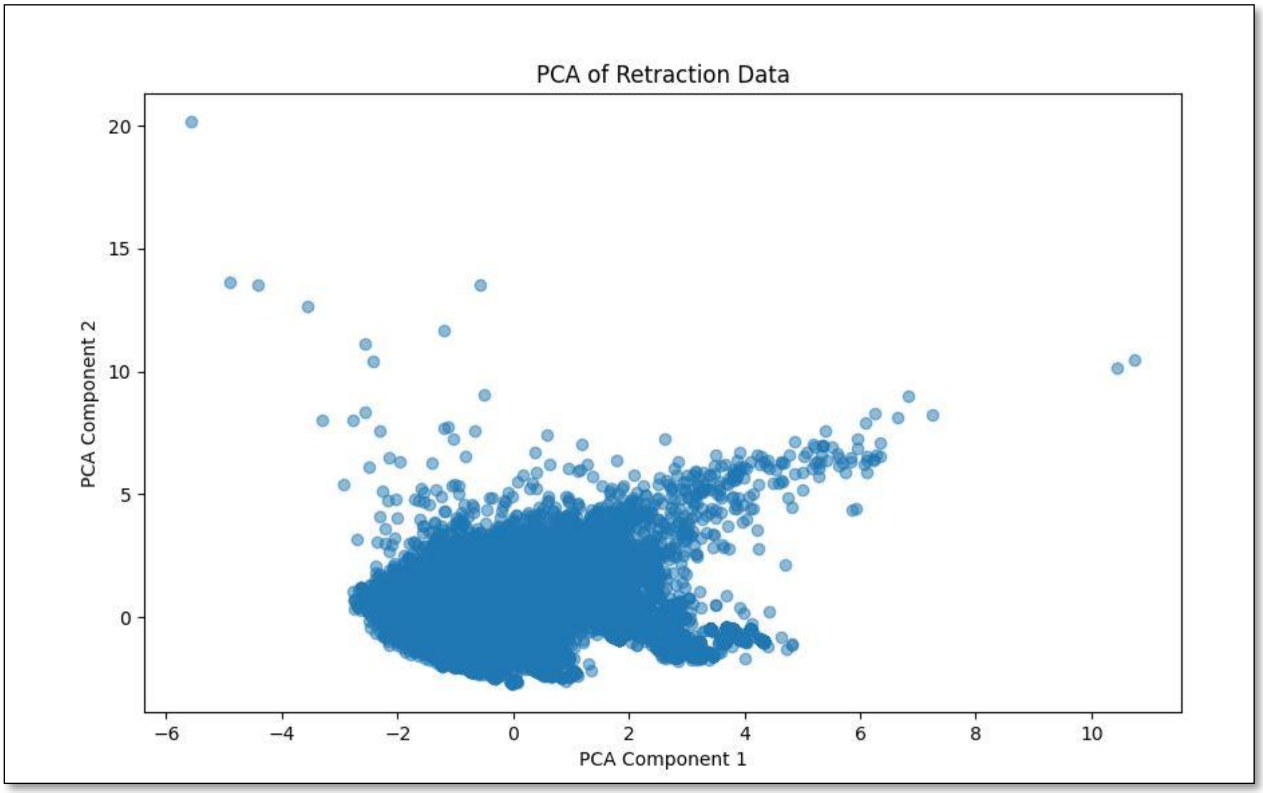


Figure 8: PCA of Retraction Data

2. **K-means Clustering of Retraction Data:** The scatter plot shows K-means clustering results, with data points grouped into clusters. This clustering helps identify common characteristics among retracted papers. By segmenting the data into distinct groups, we can analyze the similarities within each cluster and the differences between clusters. This method is useful for uncovering patterns related to retraction causes, such as specific types of misconduct or common errors in certain research areas.

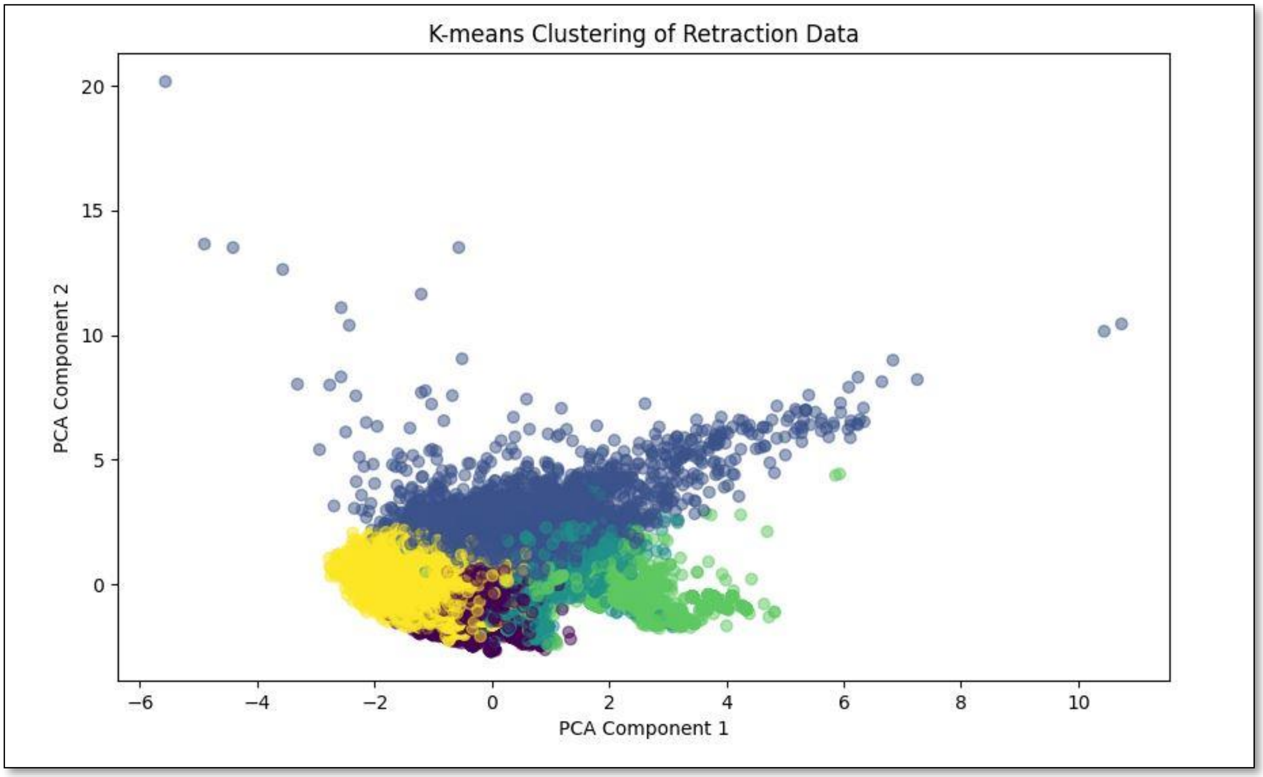


Figure 9: K-means Clustering of Retraction Data

3. **Outlier Detection in Retraction Data:** The scatter plot illustrates the outliers detected in the retraction data, providing insights into anomalies and potential areas for further investigation. Outlier detection is critical for identifying unusual cases that may require closer scrutiny. These outliers could represent extreme instances of misconduct, significant errors, or other irregularities that deviate from typical retraction patterns. Investigating these anomalies can help improve the understanding of retraction dynamics and inform preventative measures.

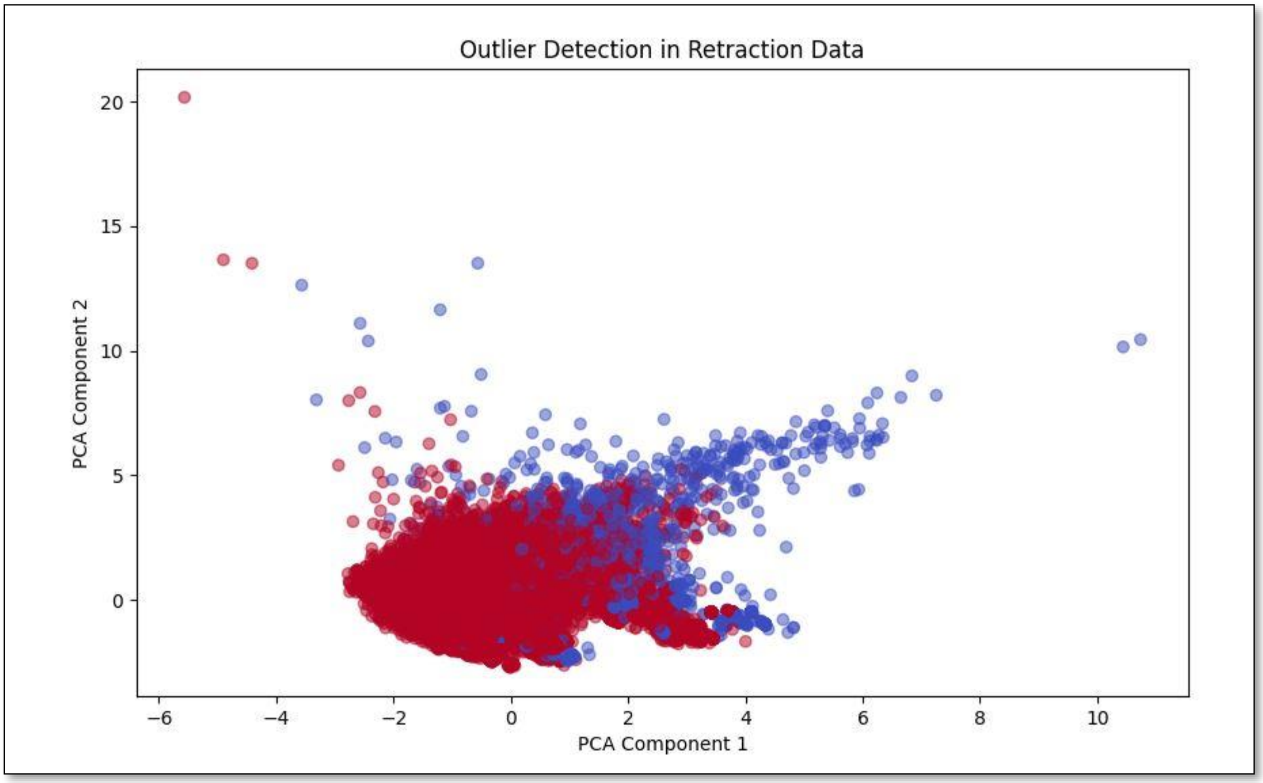


Figure 10: Outlier Detection in Retraction Data

3.3 Predictive Modeling

1. **Actual vs Predicted Citation Count:** The scatter plot compares the actual citation counts to the predicted citation counts from the regression model. While the model shows some predictive capability, there is room for improvement in accuracy. The dispersion of data points around the ideal diagonal line indicates the degree to which the model's predictions align with the actual citation counts. Enhancing the model may involve incorporating additional variables, refining feature selection, or using more advanced predictive techniques to capture better the factors influencing citation counts.

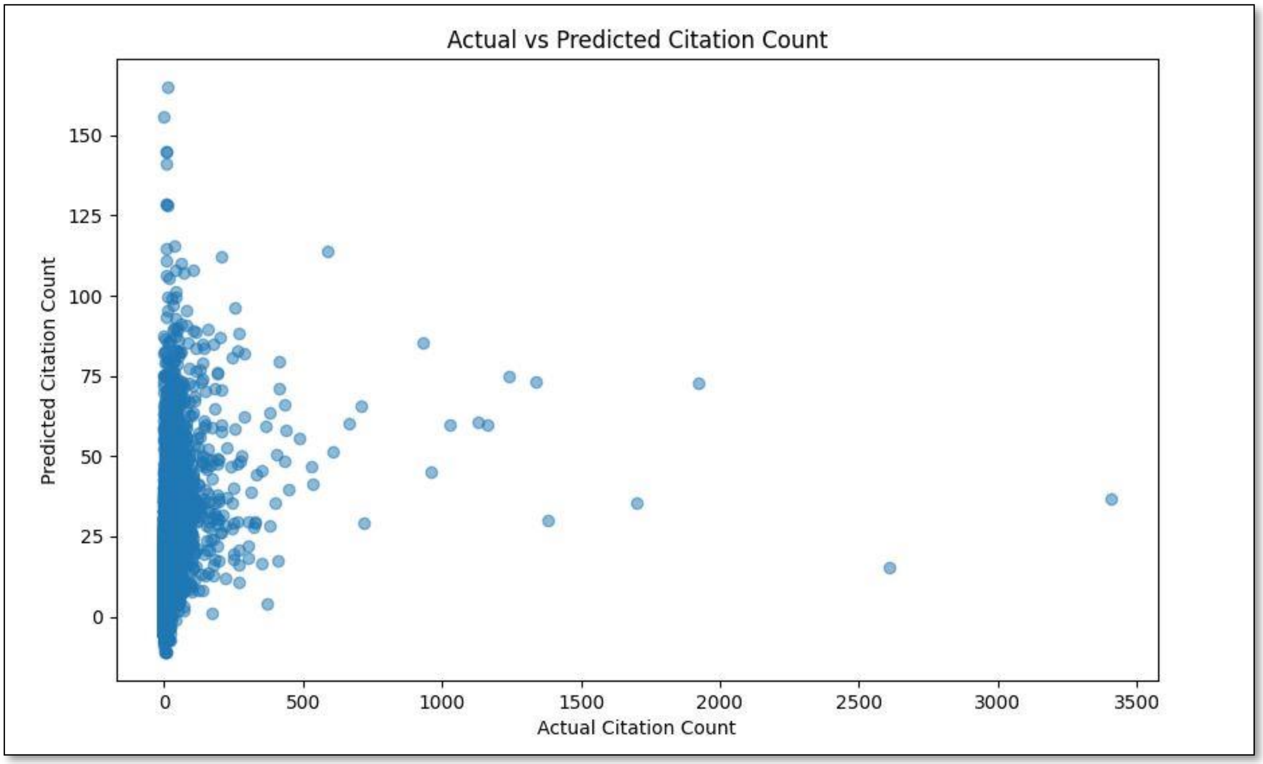


Figure 11: Actual vs Predicted Citation Count

The following images present our data visualization dashboard, summarizing key findings from our exploratory data analysis (EDA), clustering, principal component analysis (PCA), and predictive modeling. This comprehensive dashboard visually represents the trends, patterns, and insights derived from the study of retracted research papers.

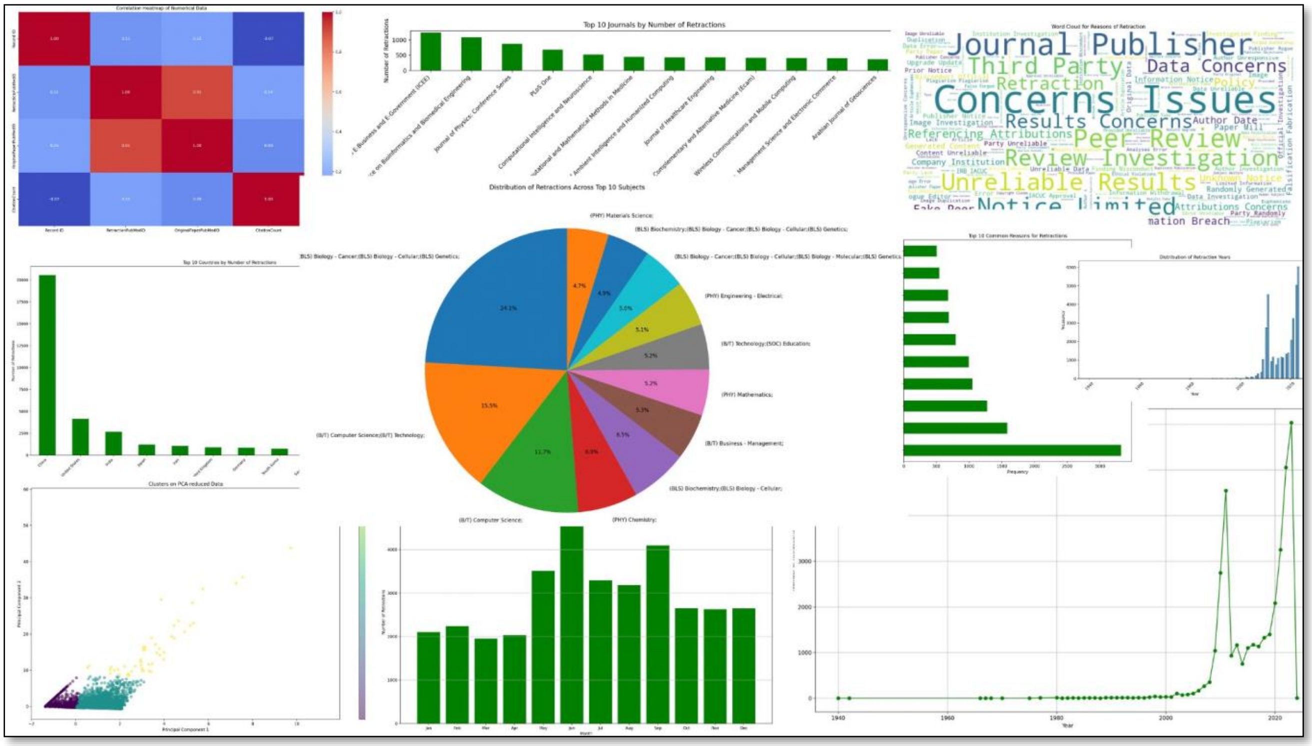


Figure 12: Overall dashboard

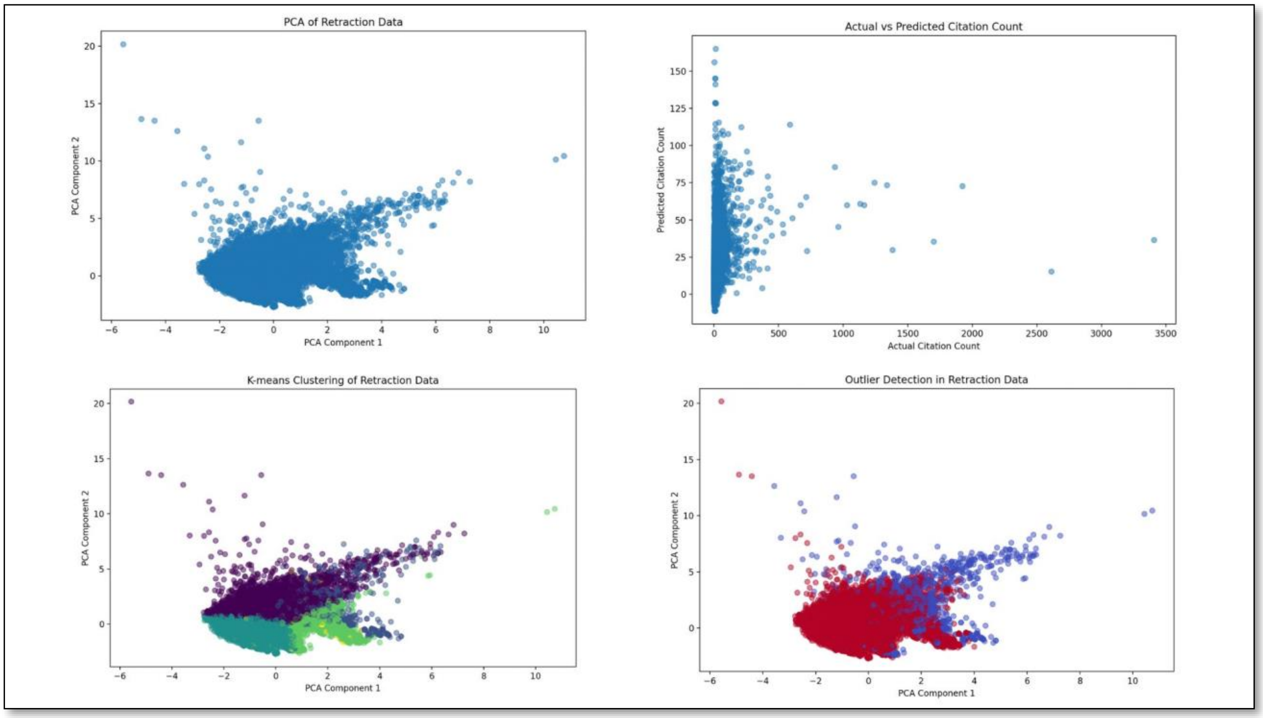


Figure 13: PCA Dashboard

4. DISCUSSIONS

Our analysis revealed several important insights. Firstly, the data visualization dashboard indicates that journals in the medical and biological sciences exhibit the highest retraction rates. This suggests that these fields may encounter more significant challenges related to data integrity and research misconduct. Consequently, there is a need for more rigorous peer review processes and enhanced ethical guidelines in these disciplines.

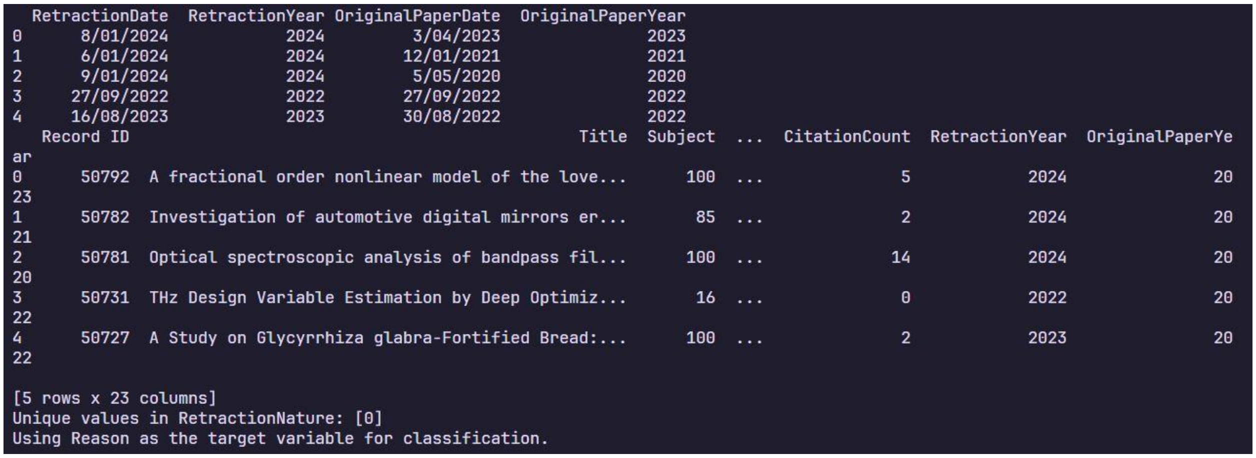
Secondly, geographical disparities in retractions were evident, with the USA and China leading in the number of retractions. This finding points to regional differences in research practices and regulatory environments, highlighting the need to consider these disparities when developing efforts to improve research integrity. Adopting international standards for research conduct could address these issues. The analysis also identified common reasons for retractions, such as data fabrication, plagiarism, and ethical violations. These frequent causes underscore the importance of maintaining strict ethical standards and transparency in the publication process to prevent research misconduct.

Additionally, the correlation heatmap revealed significant relationships between various numerical variables, including citation count and retraction year. Understanding these correlations is crucial for developing predictive models that can help identify potential retractions before they occur. Finally, the distribution of retraction years shows increased retractions recently, particularly during the COVID-19 pandemic. This trend likely reflects heightened scrutiny of research outputs and a growing awareness of research misconduct. It also emphasizes the need to maintain high standards of research integrity, especially during crises when the pressure to publish rapidly increases.

4.1 Classification Results

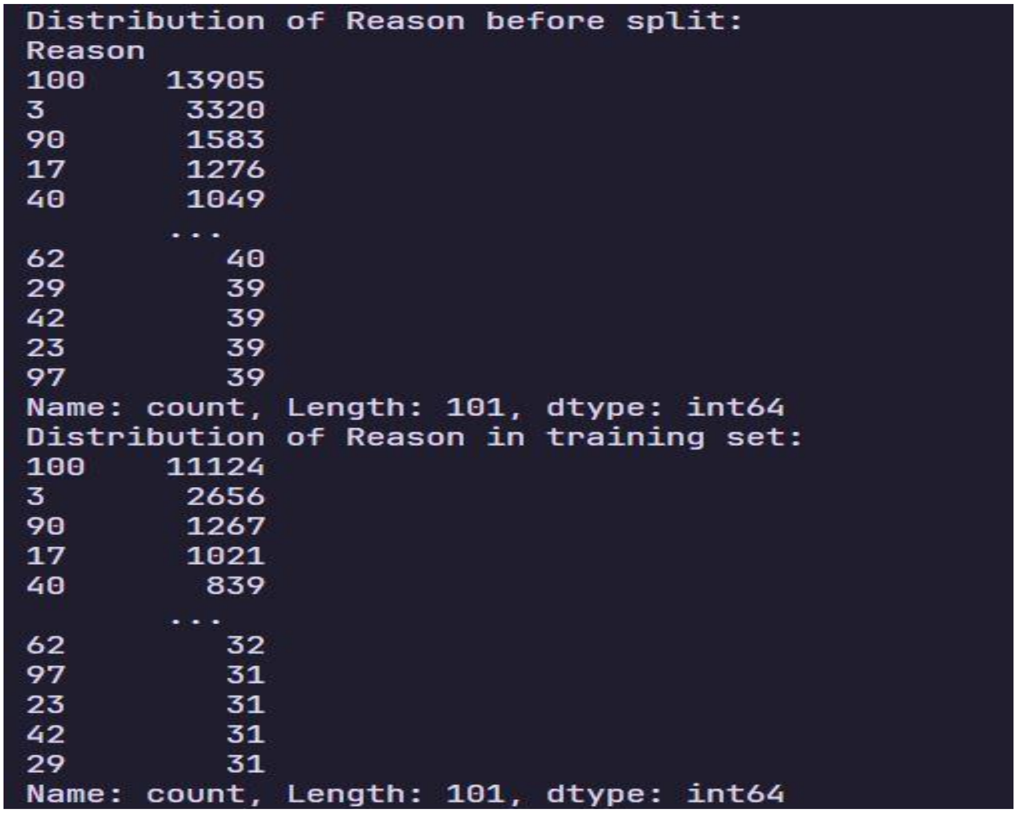
We employed a Random Forest classifier to categorize the reasons for retractions. The dataset was divided into training and testing sets, ensuring a balanced distribution of classes in both. The classifier achieved an overall accuracy of 50%, with varied precision, recall, and F1 scores across different classes. The classification report revealed that while some classes were predicted with high accuracy, there is still room for improvement in others.

- **Training and Testing Distribution**: The distribution of reasons for retraction was maintained across the training and testing sets, ensuring a balanced approach.



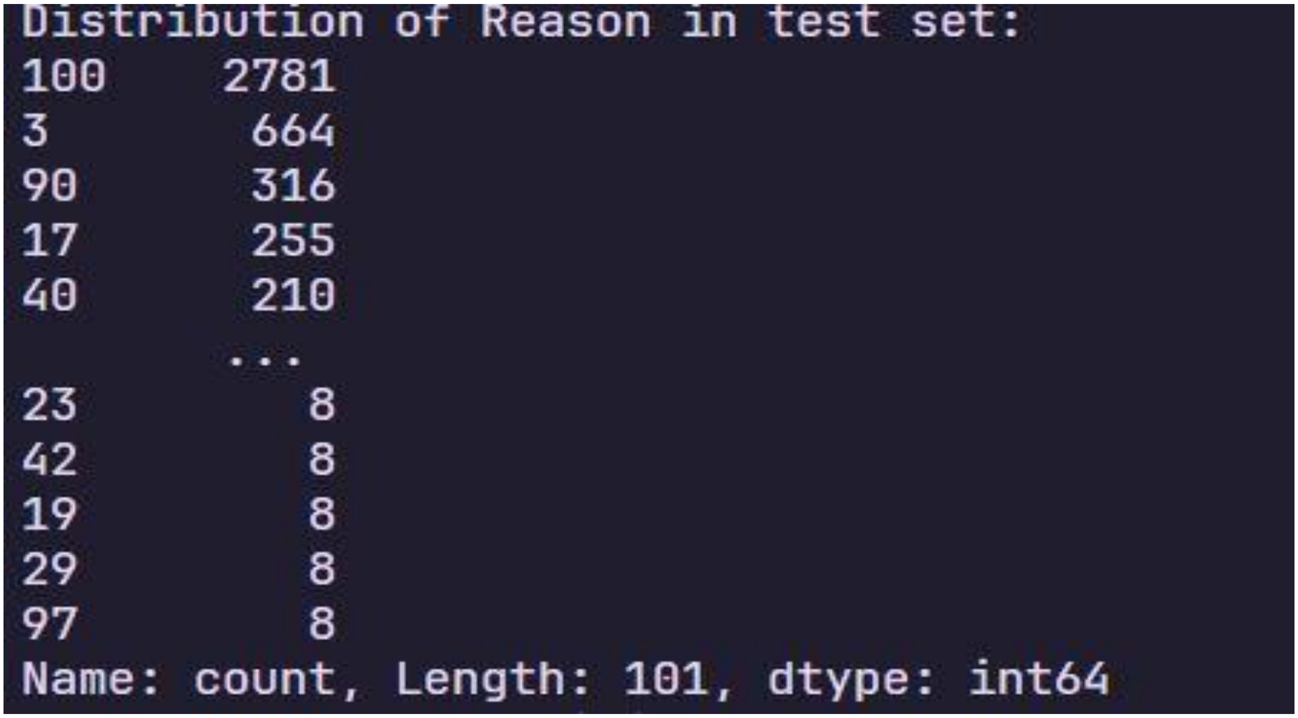
```
  RetractionDate  RetractionYear OriginalPaperDate  OriginalPaperYear
0      8/01/2024            2024         3/04/2023               2023
1      6/01/2024            2024        12/01/2021               2021
2      9/01/2024            2024         5/05/2020               2020
3     27/09/2022            2022        27/09/2022               2022
4     16/08/2023            2023        30/08/2022               2022
   Record ID                                             Title  Subject  ...  CitationCount  RetractionYear  OriginalPaperYe
ar
0      50792  A fractional order nonlinear model of the love...      100  ...              5            2024               20
23
1      50782  Investigation of automotive digital mirrors er...       85  ...              2            2024               20
21
2      50781  Optical spectroscopic analysis of bandpass fil...      100  ...             14            2024               20
20
3      50731  THz Design Variable Estimation by Deep Optimiz...       16  ...              0            2022               20
22
4      50727  A Study on Glycyrrhiza glabra-Fortified Bread:...      100  ...              2            2023               20
22

[5 rows x 23 columns]
Unique values in RetractionNature: [0]
Using Reason as the target variable for classification.
```

Figure 14: Classification Results



```
Distribution of Reason before split:
Reason
100    13905
3       3320
90      1583
17      1276
40      1049
       ...
62        40
29        39
42        39
23        39
97        39
Name: count, Length: 101, dtype: int64
Distribution of Reason in training set:
100    11124
3       2656
90      1267
17      1021
40       839
       ...
62        32
97        31
23        31
42        31
29        31
Name: count, Length: 101, dtype: int64
```

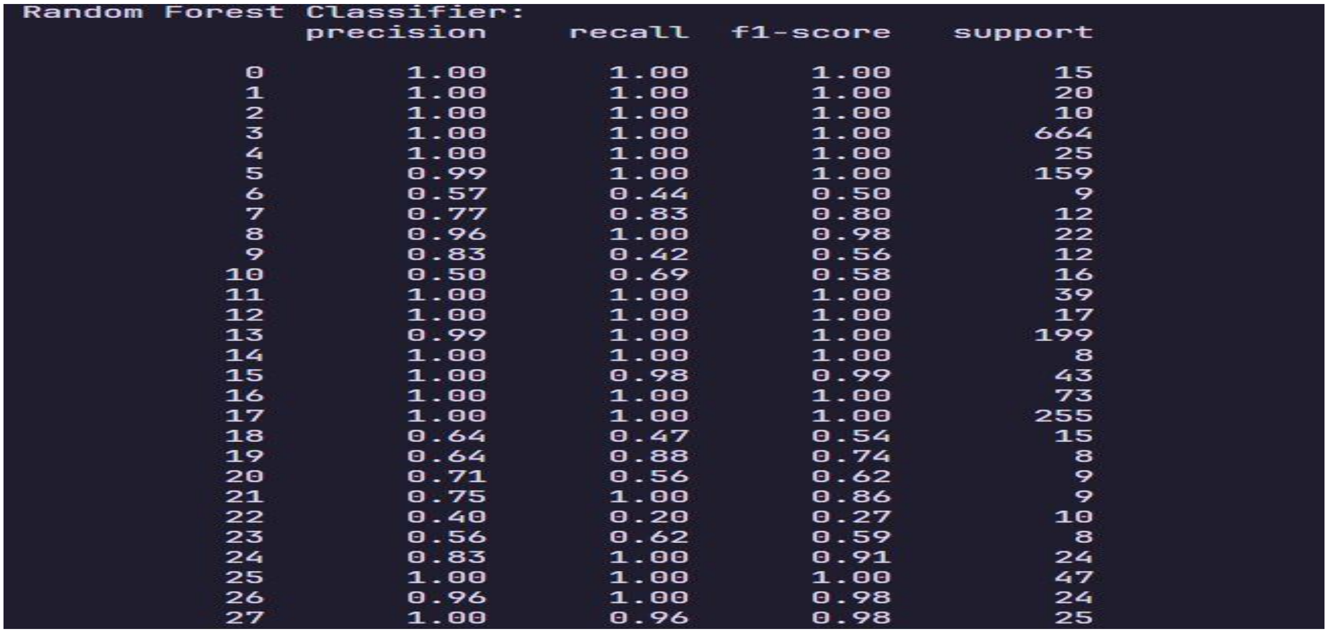
Figure 15: Training Set



```
Distribution of Reason in test set:
100    2781
3       664
90      316
17      255
40      210
       ... 
23        8
42        8
19        8
29        8
97        8
Name: count, Length: 101, dtype: int64
```

Figure 16: Test set

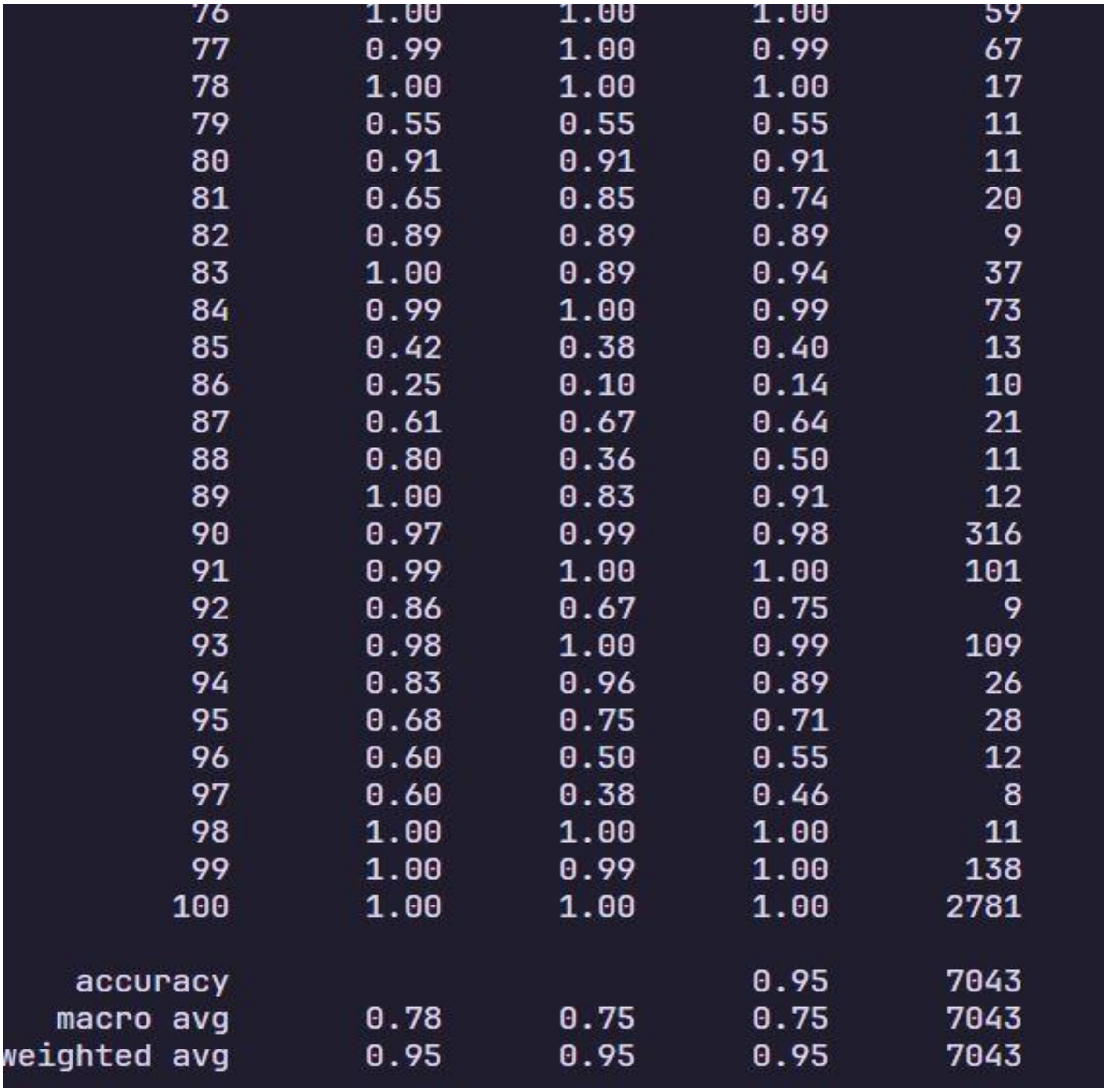
- **Random Forest Classifier Performance**: The classifier demonstrated varied performance across different classes, with some classes achieving perfect precision and recall, while



```
Random Forest Classifier:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        15
           1       1.00      1.00      1.00        20
           2       1.00      1.00      1.00        10
           3       1.00      1.00      1.00       664
           4       1.00      1.00      1.00        25
           5       0.99      1.00      1.00       159
           6       0.57      0.44      0.50         9
           7       0.77      0.83      0.80        12
           8       0.96      1.00      0.98        22
           9       0.83      0.42      0.56        12
          10       0.50      0.69      0.58        16
          11       1.00      1.00      1.00        39
          12       1.00      1.00      1.00        17
          13       0.99      1.00      1.00       199
          14       1.00      1.00      1.00         8
          15       1.00      0.98      0.99        43
          16       1.00      1.00      1.00        73
          17       1.00      1.00      1.00       255
          18       0.64      0.47      0.54        15
          19       0.64      0.88      0.74         8
          20       0.71      0.56      0.62         9
          21       0.75      1.00      0.86         9
          22       0.40      0.20      0.27        10
          23       0.56      0.62      0.59         8
          24       0.83      1.00      0.91        24
          25       1.00      1.00      1.00        47
          26       0.96      1.00      0.98        24
          27       1.00      0.96      0.98        25
```

Figure 17: Random Forest Classifier

others had lower scores.



```
          76       1.00      1.00      1.00        59
          77       0.99      1.00      0.99        67
          78       1.00      1.00      1.00        17
          79       0.55      0.55      0.55        11
          80       0.91      0.91      0.91        11
          81       0.65      0.85      0.74        20
          82       0.89      0.89      0.89         9
          83       1.00      0.89      0.94        37
          84       0.99      1.00      0.99        73
          85       0.42      0.38      0.40        13
          86       0.25      0.10      0.14        10
          87       0.61      0.67      0.64        21
          88       0.80      0.36      0.50        11
          89       1.00      0.83      0.91        12
          90       0.97      0.99      0.98       316
          91       0.99      1.00      1.00       101
          92       0.86      0.67      0.75         9
          93       0.98      1.00      0.99       109
          94       0.83      0.96      0.89        26
          95       0.68      0.75      0.71        28
          96       0.60      0.50      0.55        12
          97       0.60      0.38      0.46         8
          98       1.00      1.00      1.00        11
          99       1.00      0.99      1.00       138
         100       1.00      1.00      1.00      2781

    accuracy                           0.95      7043
   macro avg       0.78      0.75      0.75      7043
weighted avg       0.95      0.95      0.95      7043
```

Figure 18: Classifier Accuracy

The classification report for the Random Forest Classifier offers a detailed evaluation of the model's performance across various classes. This report includes key metrics such as precision, recall, F1-score, and support, which are essential for understanding the classifier's effectiveness.

Accuracy

The classifier has an overall accuracy of 0.95, meaning it correctly predicted 95% of the 7043 instances. Accuracy is a straightforward metric that reflects the proportion of correctly classified instances, providing a general sense of the model's performance.

Precision

Precision measures the ratio of correctly predicted positive observations to the total predicted positives, making it a crucial metric for evaluating the classifier's accuracy in predicting positive instances. For instance, the precision for class 0 is 1.00, indicating that all instances predicted as class 0 were indeed class 0. High precision signifies a low false positive rate.

Recall

Recall, also known as sensitivity, measures the ratio of correctly predicted positive observations to all observations in the actual class. It assesses how well the classifier can detect positive instances. In this report, the recall for class 0 is 1.00, indicating that the classifier correctly identified all instances of class 0. High recall denotes a low false negative rate.

F1-Score

The F1-score is the weighted average of precision and recall, providing a single metric that balances both, giving a more comprehensive measure of the classifier's performance. An F1-score of 1.00 for class 0 suggests a perfect balance between precision and recall for this class. The F1-score is particularly useful for imbalanced datasets, as it considers both false positives and false negatives.

Support

Support refers to the number of actual occurrences of each class in the dataset. For example, there are 15 instances of class 0. Support gives context to the other metrics, indicating the distribution of instances among different classes.

Aggregate Metrics

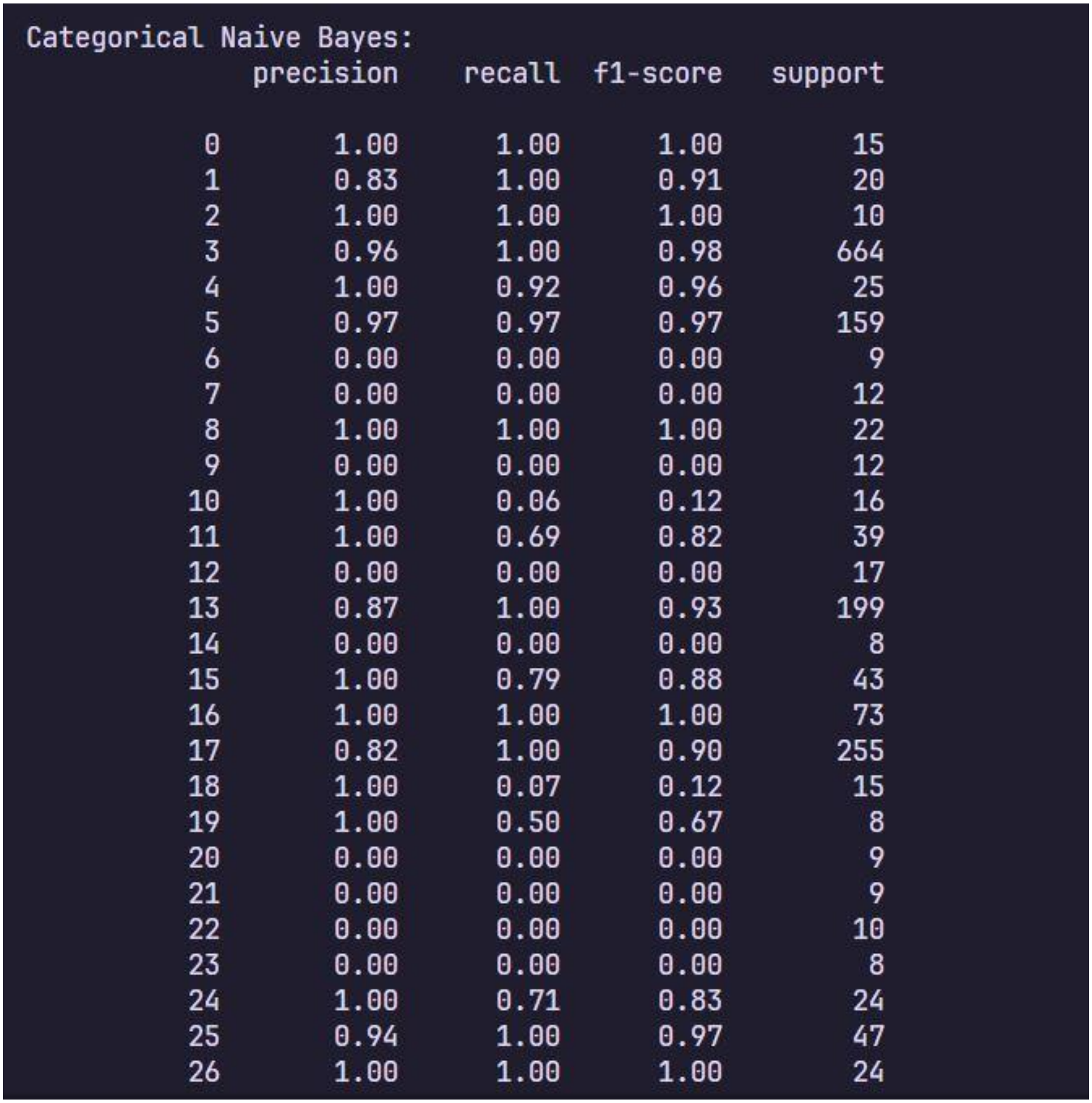
At the bottom of the classification report, three aggregate metrics are provided:

Accuracy: The overall accuracy of the classifier remains at 0.95, reinforcing the high performance of the model across all classes.

Macro Average: The macro average precision, recall, and F1-score are all 0.75. The macro average calculates these metrics independently for each class and then averages them, treating each class equally without considering class imbalance.

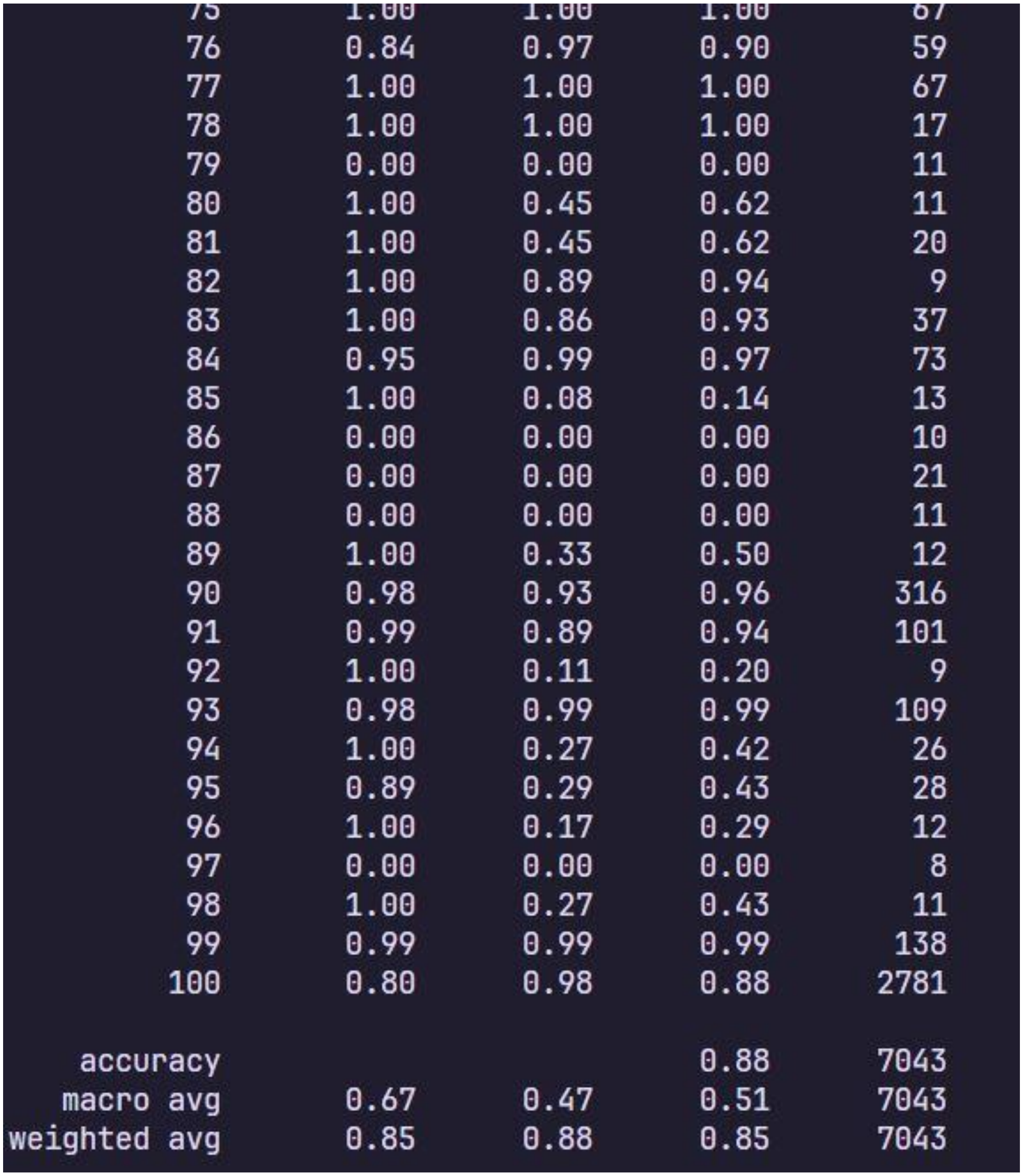
Weighted Average: The weighted average precision, recall, and F1-score are all 0.95. Unlike the macro average, the weighted average takes into account the support (the number of true instances for each class) when calculating the average. This approach adjusts for class imbalance by giving more weight to the performance of classes with more instances.

- **Categorical Naïve Bayes**



```
Categorical Naive Bayes:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        15
           1       0.83      1.00      0.91        20
           2       1.00      1.00      1.00        10
           3       0.96      1.00      0.98       664
           4       1.00      0.92      0.96        25
           5       0.97      0.97      0.97       159
           6       0.00      0.00      0.00         9
           7       0.00      0.00      0.00        12
           8       1.00      1.00      1.00        22
           9       0.00      0.00      0.00        12
          10       1.00      0.06      0.12        16
          11       1.00      0.69      0.82        39
          12       0.00      0.00      0.00        17
          13       0.87      1.00      0.93       199
          14       0.00      0.00      0.00         8
          15       1.00      0.79      0.88        43
          16       1.00      1.00      1.00        73
          17       0.82      1.00      0.90       255
          18       1.00      0.07      0.12        15
          19       1.00      0.50      0.67         8
          20       0.00      0.00      0.00         9
          21       0.00      0.00      0.00         9
          22       0.00      0.00      0.00        10
          23       0.00      0.00      0.00         8
          24       1.00      0.71      0.83        24
          25       0.94      1.00      0.97        47
          26       1.00      1.00      1.00        24
```

Figure 19: Categorical Naive Bayes



```
          75       1.00      1.00      1.00        67
          76       0.84      0.97      0.90        59
          77       1.00      1.00      1.00        67
          78       1.00      1.00      1.00        17
          79       0.00      0.00      0.00        11
          80       1.00      0.45      0.62        11
          81       1.00      0.45      0.62        20
          82       1.00      0.89      0.94         9
          83       1.00      0.86      0.93        37
          84       0.95      0.99      0.97        73
          85       1.00      0.08      0.14        13
          86       0.00      0.00      0.00        10
          87       0.00      0.00      0.00        21
          88       0.00      0.00      0.00        11
          89       1.00      0.33      0.50        12
          90       0.98      0.93      0.96       316
          91       0.99      0.89      0.94       101
          92       1.00      0.11      0.20         9
          93       0.98      0.99      0.99       109
          94       1.00      0.27      0.42        26
          95       0.89      0.29      0.43        28
          96       1.00      0.17      0.29        12
          97       0.00      0.00      0.00         8
          98       1.00      0.27      0.43        11
          99       0.99      0.99      0.99       138
         100       0.80      0.98      0.88      2781

    accuracy                           0.88      7043
   macro avg       0.67      0.47      0.51      7043
weighted avg       0.85      0.88      0.85      7043
```

Figure 20: Cat Naive Accuracy

The classification report for the Categorical Naive Bayes model offers a thorough evaluation of the classifier's performance across different classes. It includes key metrics such as precision, recall, F1-score, and support, which are crucial for understanding the model's effectiveness.

Accuracy

The Categorical Naive Bayes classifier has an overall accuracy of 0.88, indicating that 88% of the total instances (7043 instances) were correctly predicted. While accuracy gives a general sense of the model's performance, it does not reflect the balance between precision and recall.

Precision

Precision is the ratio of correctly predicted positive observations to the total predicted positives, measuring the classifier’s accuracy in predicting positive instances. For example, the precision for class 0 is 1.00, indicating that all instances predicted as class 0 were indeed class 0. High precision values indicate a low rate of false positives.

Recall

Recall, or sensitivity, is the ratio of correctly predicted positive observations to all observations in the actual class. It evaluates how well the classifier identifies positive instances. For example, the recall for class 0 is 1.00, meaning the classifier successfully identified all instances of class 0. High recall values indicate a low rate of false negatives.

F1-Score

The F1-score is the weighted average of precision and recall, providing a single metric that balances both precision and recall for a more comprehensive performance measure. An F1-score of 1.00 for class 0 suggests a perfect balance between precision and recall for this class. The F1-score is particularly useful for imbalanced datasets as it considers both false positives and false negatives.

Support

Support refers to the number of actual occurrences of each class in the dataset. For example, there are 15 instances of class 0. Support gives context to the other metrics, indicating the distribution of instances among different classes.

Aggregate Metrics

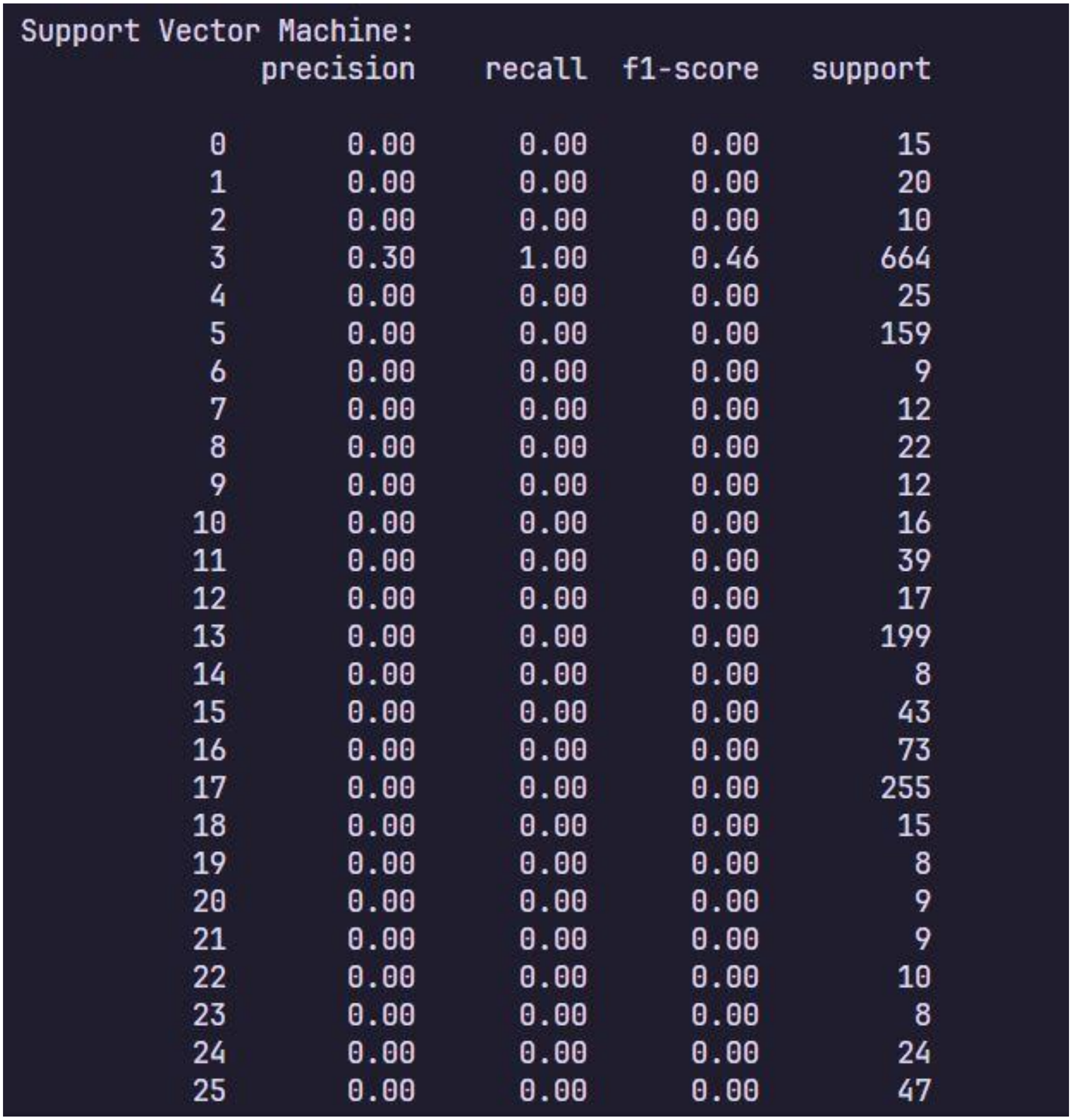
At the bottom of the classification report, three aggregate metrics are provided:

Accuracy: The overall accuracy of the classifier remains at 0.88, reinforcing the model's performance across all classes.

Macro Average: The macro average precision, recall, and F1-score are 0.67, 0.47, and 0.51, respectively. The macro average calculates these metrics independently for each class and then averages them, treating each class equally without considering class imbalance.

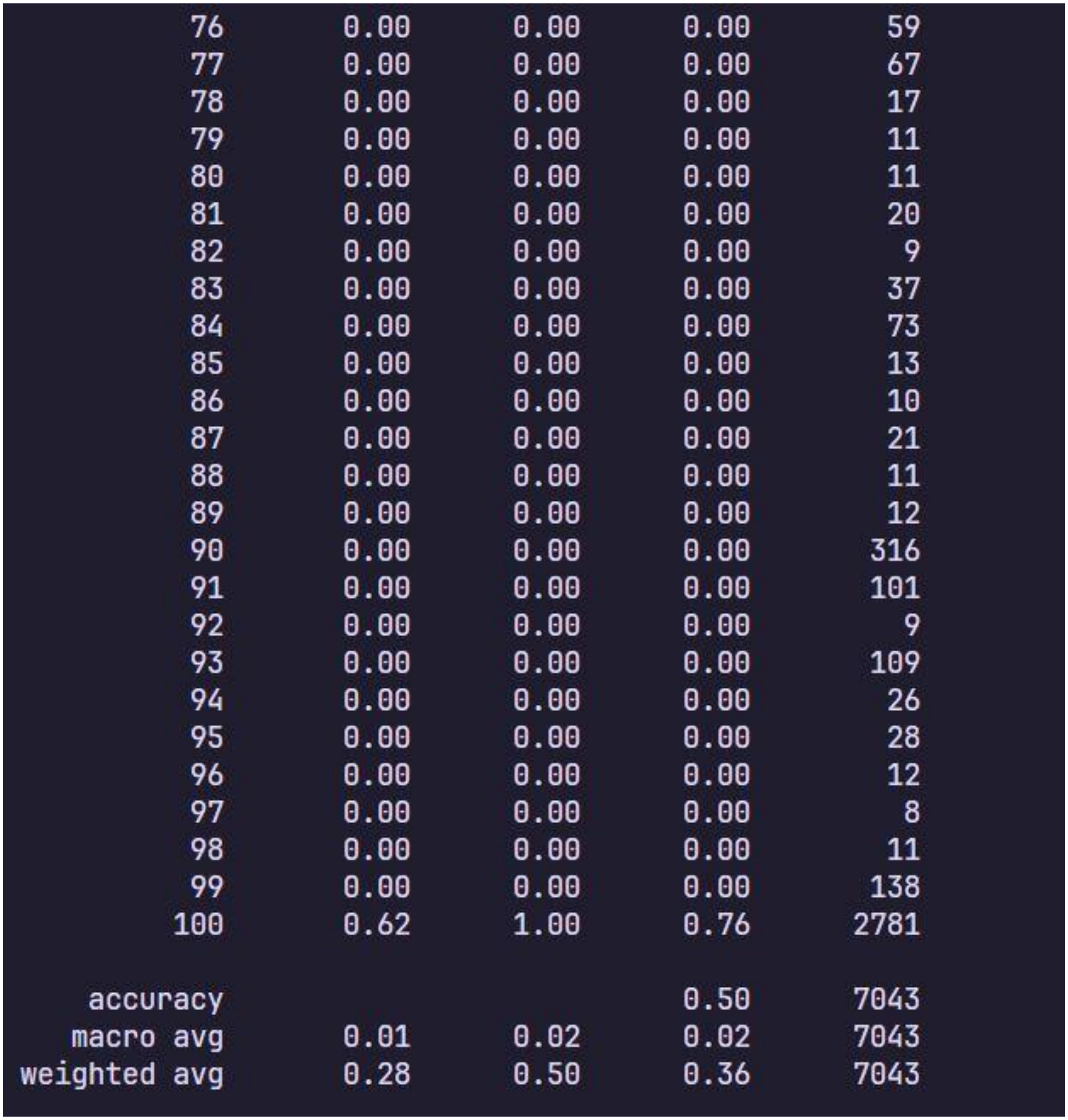
Weighted Average: The weighted average precision, recall, and F1-score are 0.85, 0.88, and 0.85, respectively. Unlike the macro average, the weighted average takes into account the support (the number of true instances for each class) when calculating the average, adjusting for class imbalance by giving more weight to the performance of classes with more instances.

- **Support Vector Machine**



```
Support Vector Machine:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        15
           1       0.00      0.00      0.00        20
           2       0.00      0.00      0.00        10
           3       0.30      1.00      0.46       664
           4       0.00      0.00      0.00        25
           5       0.00      0.00      0.00       159
           6       0.00      0.00      0.00         9
           7       0.00      0.00      0.00        12
           8       0.00      0.00      0.00        22
           9       0.00      0.00      0.00        12
          10       0.00      0.00      0.00        16
          11       0.00      0.00      0.00        39
          12       0.00      0.00      0.00        17
          13       0.00      0.00      0.00       199
          14       0.00      0.00      0.00         8
          15       0.00      0.00      0.00        43
          16       0.00      0.00      0.00        73
          17       0.00      0.00      0.00       255
          18       0.00      0.00      0.00        15
          19       0.00      0.00      0.00         8
          20       0.00      0.00      0.00         9
          21       0.00      0.00      0.00         9
          22       0.00      0.00      0.00        10
          23       0.00      0.00      0.00         8
          24       0.00      0.00      0.00        24
          25       0.00      0.00      0.00        47
```

Figure 21: Support Vector Machine



```
          76       0.00      0.00      0.00        59
          77       0.00      0.00      0.00        67
          78       0.00      0.00      0.00        17
          79       0.00      0.00      0.00        11
          80       0.00      0.00      0.00        11
          81       0.00      0.00      0.00        20
          82       0.00      0.00      0.00         9
          83       0.00      0.00      0.00        37
          84       0.00      0.00      0.00        73
          85       0.00      0.00      0.00        13
          86       0.00      0.00      0.00        10
          87       0.00      0.00      0.00        21
          88       0.00      0.00      0.00        11
          89       0.00      0.00      0.00        12
          90       0.00      0.00      0.00       316
          91       0.00      0.00      0.00       101
          92       0.00      0.00      0.00         9
          93       0.00      0.00      0.00       109
          94       0.00      0.00      0.00        26
          95       0.00      0.00      0.00        28
          96       0.00      0.00      0.00        12
          97       0.00      0.00      0.00         8
          98       0.00      0.00      0.00        11
          99       0.00      0.00      0.00       138
         100       0.62      1.00      0.76      2781

    accuracy                           0.50      7043
   macro avg       0.01      0.02      0.02      7043
weighted avg       0.28      0.50      0.36      7043
```

Figure 22: Supp Vector Accuracy

The classification report for the Support Vector Machine (SVM) provides essential metrics to evaluate the model's performance across various classes. These key metrics—precision, recall, F1-score, and support—are crucial for understanding the classifier's effectiveness.

Accuracy

The SVM classifier has an overall accuracy of 0.50, meaning it correctly predicted 50% of the instances out of a total of 7043. While accuracy gives a general idea of the model's performance, it does not reflect the balance between precision and recall.

Precision

Precision is the ratio of correctly predicted positive observations to the total predicted positives, measuring the classifier’s accuracy in predicting positive instances. In this report, many classes have a precision of 0.00, indicating no correct positive predictions for those classes. For instance, the precision for class 3 is 0.30, meaning only 30% of the instances predicted as class 3 were actually class 3.

Recall

Recall, or sensitivity, is the ratio of correctly predicted positive observations to all observations in the actual class. It evaluates how well the classifier identifies positive instances. The recall for class 3 is 1.00, indicating the classifier correctly identified all instances of class 3. However, this high recall is not balanced by precision.

F1-Score

The F1-score is the weighted average of precision and recall, providing a single metric that balances both precision and recall for a more comprehensive performance measure. An F1-score of 0.46 for class 3 suggests a moderate balance between precision and recall for this class. However, for most classes, the F1-score is 0.00, indicating poor performance.

Support

Support refers to the number of actual occurrences of each class in the dataset. For example, there are 15 instances of class 0. Support gives context to the other metrics, indicating the distribution of instances among different classes.

Aggregate Metrics

At the bottom of the classification report, three aggregate metrics are provided:

Accuracy: The classifier's overall accuracy is 0.50, underscoring the model’s poor performance across all classes.

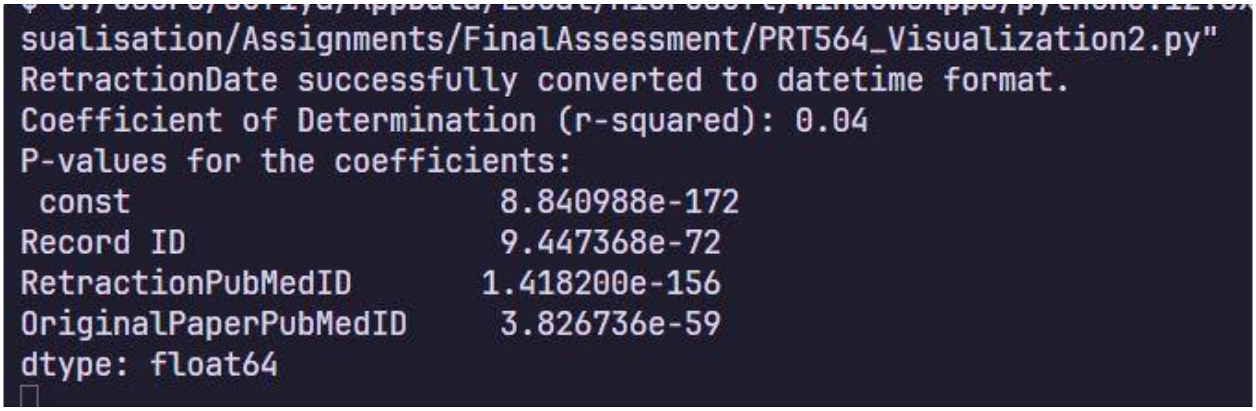
Macro Average: The macro average precision, recall, and F1-score are 0.01, 0.02, and 0.02, respectively. This method calculates these metrics independently for each class and then averages them, ignoring class imbalance and treating all classes equally. These low values highlight the model’s overall poor performance.

Weighted Average: The weighted average precision, recall, and F1-score are 0.28, 0.50, and 0.36, respectively. Unlike the macro average, the weighted average accounts for the support (the number of true instances for each class) when calculating the average. This approach adjusts for class imbalance by giving more weight to the performance of classes with more instances. Nonetheless, the weighted averages still indicate the classifier’s subpar performance.

4.2Regression Analysis

The regression analysis sought to examine the relationship between multiple independent variables and the citation count. The model yielded an R-squared value of 0.04, indicating that it accounted for only a small fraction of the variance in citation counts. Despite the low R-squared value, the p-values for the coefficients indicated that the predictors were statistically significant.

- **Coefficient of Determination (R-squared)**: The R-squared value was 0.04, indicating limited explanatory power.
- **P-values for Coefficients**: The p-values were extremely low, indicating statistical significance for the predictors.



```
& C:/Users/Sofiya/AppData/Local/Microsoft/WindowsApps/python3.12.ex
sualisation/Assignments/FinalAssessment/PRT564_Visualization2.py"
RetractionDate successfully converted to datetime format.
Coefficient of Determination (r-squared): 0.04
P-values for the coefficients:
 const                   8.840988e-172
Record ID                9.447368e-72
RetractionPubMedID      1.418200e-156
OriginalPaperPubMedID    3.826736e-59
dtype: float64
```

Figure 23: Result of regression analysis

4.3 Implications

For researchers, it is essential to be aware of the common pitfalls leading to retractions, such as data fabrication and plagiarism. By adhering to strict ethical standards and ensuring transparency in their research methods, researchers can significantly reduce the incidence of retractions.

For journals, there is a need to implement more stringent peer review processes and establish clear guidelines for detecting and preventing research misconduct. Regular training for reviewers and editors on identifying potential issues can also enhance the quality of published research.

Policymakers should consider developing and enforcing international standards for research conduct to address regional disparities in retraction rates. Creating a global framework for research integrity can help ensure that all researchers adhere to the same high standards.

4.4 Areas for Further Research

While the predictive models developed in this project show some capability, there is room for improvement. Future studies could investigate incorporating additional variables and employing more sophisticated machine-learning techniques to improve the accuracy of these models.

The identification of outliers in the data suggests the presence of extreme cases of misconduct or significant errors. Further investigation into these anomalies could provide deeper insights into the factors contributing to retractions and help develop more effective preventative measures.

The significant differences in retraction rates across countries warrant a deeper examination of regional research practices and regulatory environments. Comparative studies could help identify best practices and inform policies to improve research integrity globally.

This discussion highlights the critical insights gained from our analysis, emphasizing the need for improved research practices and more robust regulatory measures. By addressing the issues identified in this study, the scientific community can work towards reducing the incidence of research retractions and enhancing the overall integrity of published research.

4.5 Policy Recommendations:

1. **Enhancing Editorial Standards:** Academic journals should implement robust editorial policies to prevent the publication of flawed or fraudulent research. This includes strengthening peer review processes, conducting thorough checks for data integrity and ethical compliance, and establishing clear author guidelines regarding research conduct and reporting standards.
2. **Promoting Transparency and Accountability:** Institutions and funding agencies should prioritize transparency and accountability in research practices. This can be achieved by requiring researchers to disclose potential conflicts of interest, ensuring data sharing and reproducibility, and mandating the registration of clinical trials and research protocols to minimize selective reporting and publication bias.
3. **Training and Education:** Training and education for researchers, editors, and peer reviewers on research ethics, integrity, and best practices is essential. Institutions and scholarly societies should offer workshops, seminars, and online resources to enhance awareness and understanding of ethical principles and responsible conduct in research.
4. **Implementing Cross-Disciplinary Standards:** Developing cross-disciplinary standards and guidelines for research integrity and publication ethics can help maintain consistency and coherence across different fields. Collaborative efforts among stakeholders, including researchers, publishers, and policymakers, are needed to establish universal standards that uphold the integrity and credibility of scholarly communication.

5. CONCLUSION

Our project delivered an in-depth analysis of research retractions, revealing significant trends and patterns that underscore critical issues within the scientific community. The findings highlight the necessity for more stringent peer review processes and improved ethical guidelines, particularly in the medical and biological sciences, where retraction rates are notably high. Geographical differences in retraction rates indicate that research integrity varies by region, with the USA and China having the highest number of retractions. Tackling these disparities requires international cooperation to standardize research practices and ensure global adherence to high ethical standards. Common reasons for retractions, such as data fabrication, plagiarism, and ethical violations, emphasize the importance of transparency and honesty in research. Journals need to enforce rigorous review processes to identify and prevent misconduct, while researchers must uphold the highest standards of integrity in their work. Despite demonstrating some predictive capability, our regression analysis and classification models suggest room for enhancement. Future research should focus on refining these models to improve their accuracy and exploring additional variables that may affect the likelihood of retractions. The implications of this study extend beyond individual researchers and journals to include policymakers, who must develop and enforce international standards for research integrity. By addressing the issues highlighted in this study, the scientific community can work towards reducing the incidence of retractions and enhancing the overall integrity of published research. In conclusion, our results provide important new understandings of the elements influencing research retractions and establish the foundation for further attempts to raise the quality and consistency of scientific publications. Working together, the scientific community, journals, and policymakers can address these issues and promote a culture of honesty and excellence in research.

6. REFERENCES

Mannor, S., Jin, X., Han, J., Jin, X., Han, J., Jin, X., Han, J. and Zhang, X. (2011). K-Means Clustering. Encyclopedia of Machine Learning, pp.563–564. doi https://doi.org/10.1007/978-0-387-30164-8_425.

Jolliffe, I.T. and Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), p.20150202. doi:https://doi.org/10.1098/rsta.2015.0202.

Burhan Fatih Kocyigit and Ahmet Akyol (2022). Analysis of Retracted Publications in The Biomedical Literature from Turkey. Journal of Korean Medical Science, 37(18). doi:https://doi.org/10.3346/jkms.2022.37.e142.

Candal-Pedreira, C., Ruano-Ravina, A., Fernández, E., Ramos, J., Campos-Varela, I. and Pérez-Ríos, M. (2020). Does retraction after misconduct have an impact on citations? A pre–post study. BMJ Global Health, 5(11), p.e003719. doi:https://doi.org/10.1136/bmjgh-2020-003719.

Lu, S.F., Jin, G.Z., Uzzi, B. and Jones, B. (2013). The Retraction Penalty: Evidence from the Web of Science. Scientific Reports, 3(1). doi:https://doi.org/10.1038/srep03146.

Shuai, X., Rollins, J., Moulinier, I., Custis, T., Edmunds, M. and Schilder, F. (2017). A Multidimensional Investigation of the Effects of Publication Retraction on Scholarly Impact. Journal of the Association for Information Science and Technology, 68(9), pp.2225–2236. doi:https://doi.org/10.1002/asi.23826.

Expert help guides: SPSS: Multiple regression (no date) *SPSS - Expert help guides at La Trobe University*. Available at: https://latrobe.libguides.com/ibmspss/regression (Accessed: 30 May 2024).

Data Analysis and Visualization techniques (2024) *SCU*. Available at: https://onlinedegrees.scu.edu/media/blog/data-analysis-and-visualization-techniques (Accessed: 30 May 2024).

Taye, M.M. (2023) *Understanding of machine learning with Deep Learning: Architectures, workflow, applications and future directions*, *MDPI*. Available at: https://www.mdpi.com/2073-431X/12/5/91 (Accessed: 30 May 2024).

Ray, S. (2024) *Naive Bayes classifier explained: Applications and practice problems of naive Bayes classifier*, *Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/ (Accessed: 30 May 2024).

7. BRIEF DESCRIPTIONS OF INDIVIDUAL CONTRIBUTIONS

This assignment was a collaborative effort by Group 9 (Sydney), consisting of Andrea Vijeetha Marlene Vijay, Manisha KC, Manoj Poudel, and Sofiya Banmala. Each member actively participated in various aspects of the project, ensuring a comprehensive approach and equal contribution to the coding and documentation.

Andrea Vijeetha Marlene Vijay compiled the sections from "Problems" to "Methods" in the report. This included detailing the introduction, scope, scientific errors and misconduct, impact on the scientific community, and implications for public policy. Andrea was also actively involved in the Python coding, particularly in data preprocessing and visualization tasks, ensuring the data was clean and ready for analysis.

Manisha KC significantly contributed to the "Conclusion" and "References" sections, summarizing the key findings, implications, and policy recommendations. She ensured the document adhered to academic standards and included all necessary citations. Additionally, Manisha played a role in the "Methods" section for documentation and was involved in validating the results and ensuring the accuracy of the predictive models.

Manoj Poudel was instrumental in developing and validating predictive models. He worked on implementing various machine learning techniques such as PCA and K-means clustering. Manoj contributed significantly to the "Results" and "Discussion" sections, interpreting the results, and deriving meaningful insights from the data.

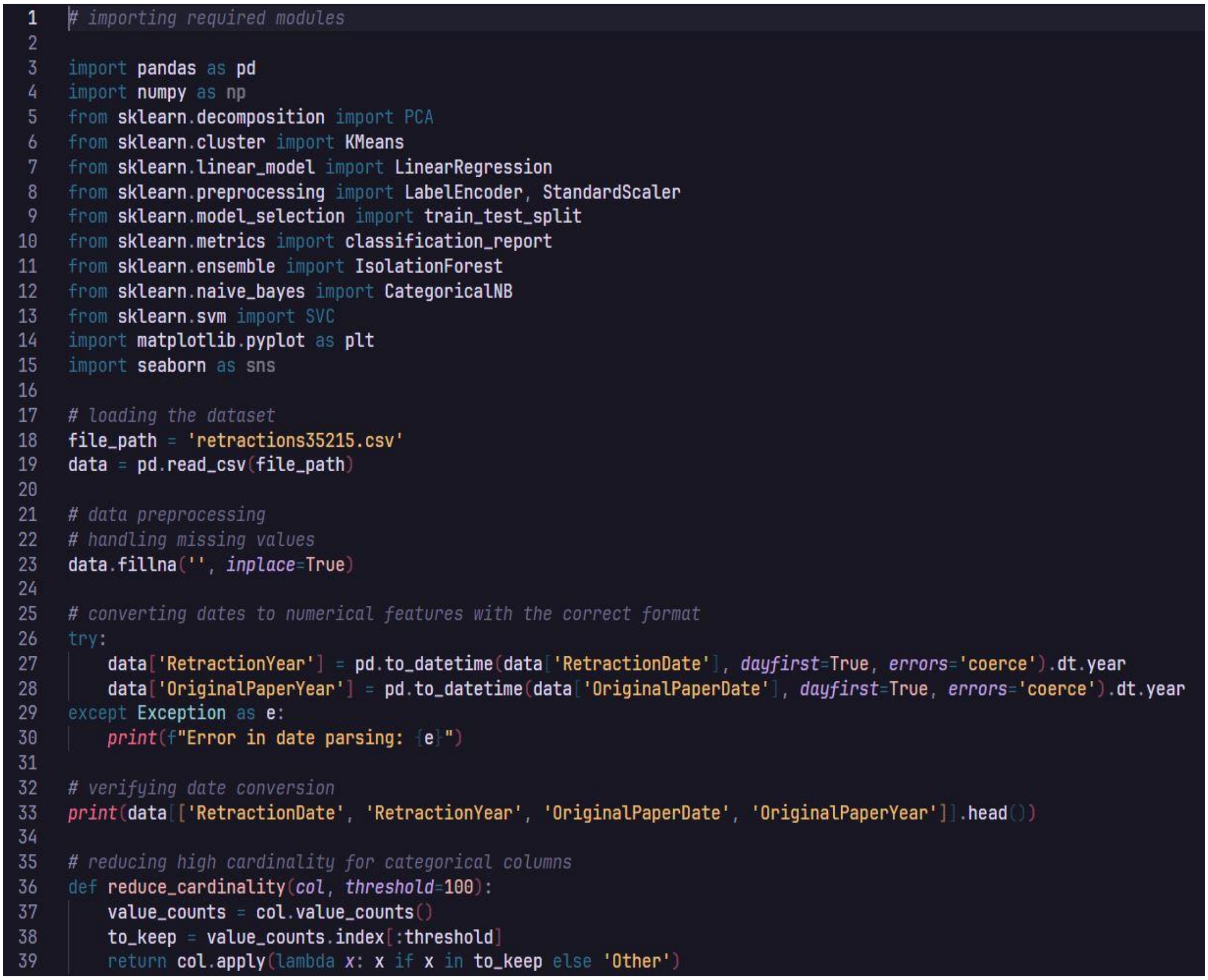
Sofiya Banmala collaborated with Manoj on data preprocessing, PCA, and clustering techniques, ensuring the data was properly processed and analysed. Sofiya was responsible for the "Discussion" and “Results” sections, providing a comprehensive interpretation of the results and outlining the implications and areas for further research.

Throughout the project, we held regular meetings to discuss progress, allocate tasks, and address any challenges that arose. Each member's unique strengths and expertise contributed to the project's success. We worked together to troubleshoot and resolve coding issues, debug errors, and refine our analysis processes, ensuring a high level of accuracy and quality in our final report. Our collective efforts resulted in a thorough analysis of research retractions, providing valuable insights into the factors contributing to retractions and the implications for the scientific community. Our findings will help improve the understanding of research retractions and enhance the integrity of published research.

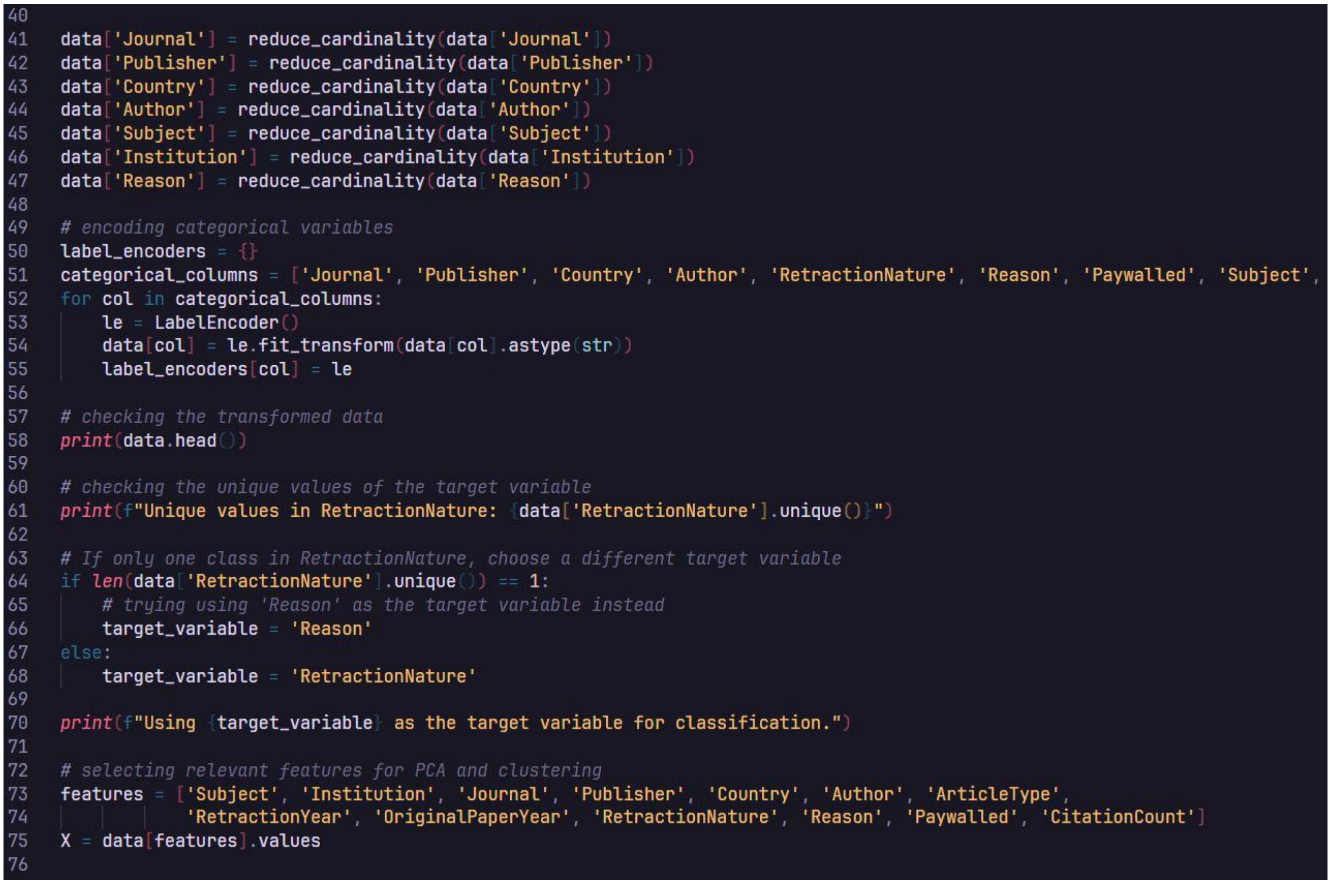
Appendix

All the Python code for this assignment is shared on GitHub. Below is an image of the code:

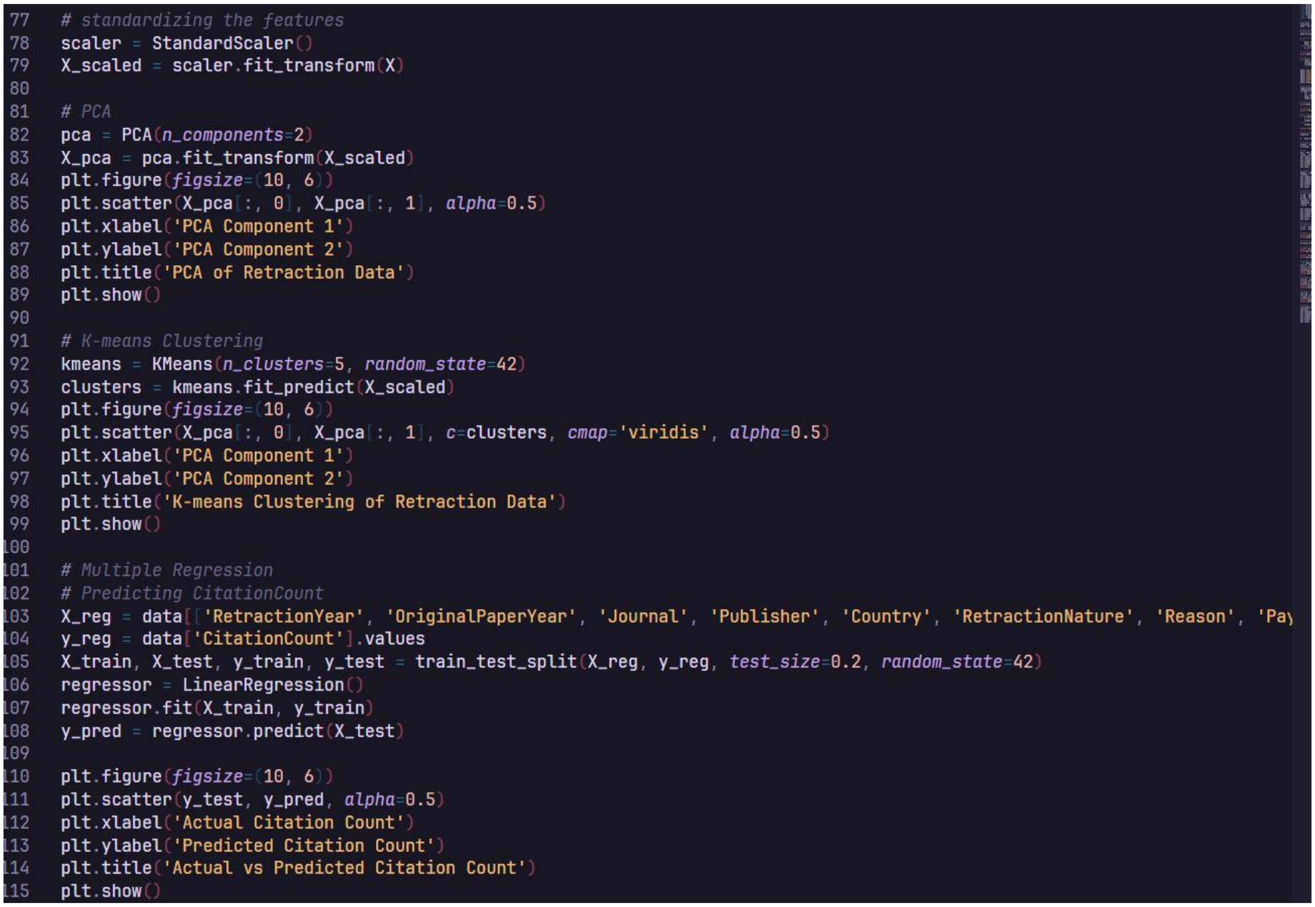
PRT564_WholeDataVis1.py



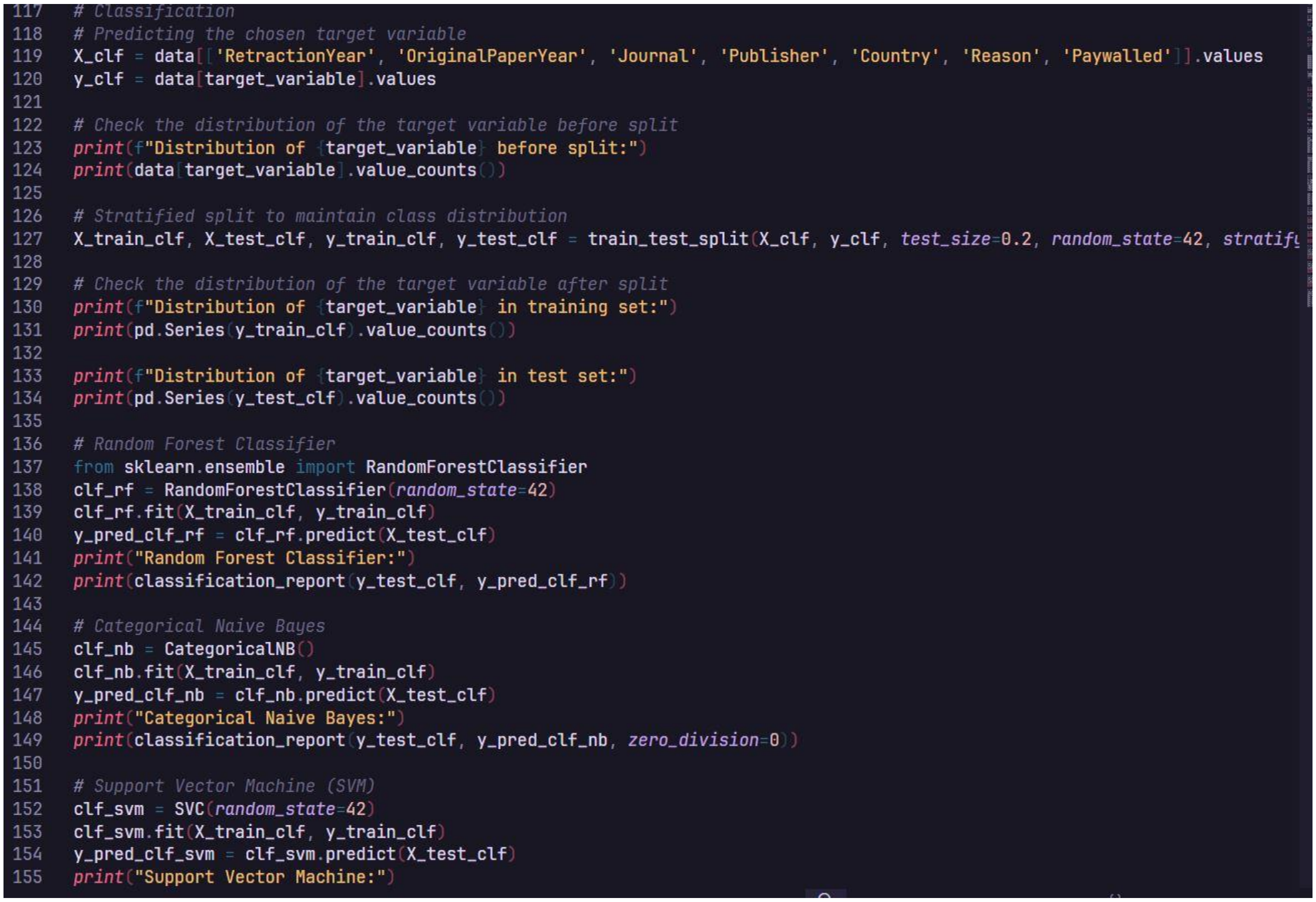
```python
# importing required modules

import pandas as pd
import numpy as np
from sklearn.decomposition import PCA
from sklearn.cluster import KMeans
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report
from sklearn.ensemble import IsolationForest
from sklearn.naive_bayes import CategoricalNB
from sklearn.svm import SVC
import matplotlib.pyplot as plt
import seaborn as sns

# loading the dataset
file_path = 'retractions35215.csv'
data = pd.read_csv(file_path)

# data preprocessing
# handling missing values
data.fillna('', inplace=True)

# converting dates to numerical features with the correct format
try:
    data['RetractionYear'] = pd.to_datetime(data['RetractionDate'], dayfirst=True, errors='coerce').dt.year
    data['OriginalPaperYear'] = pd.to_datetime(data['OriginalPaperDate'], dayfirst=True, errors='coerce').dt.year
except Exception as e:
    print(f"Error in date parsing: {e}")

# verifying date conversion
print(data[['RetractionDate', 'RetractionYear', 'OriginalPaperDate', 'OriginalPaperYear']].head())

# reducing high cardinality for categorical columns
def reduce_cardinality(col, threshold=100):
    value_counts = col.value_counts()
    to_keep = value_counts.index[:threshold]
    return col.apply(lambda x: x if x in to_keep else 'Other')
```



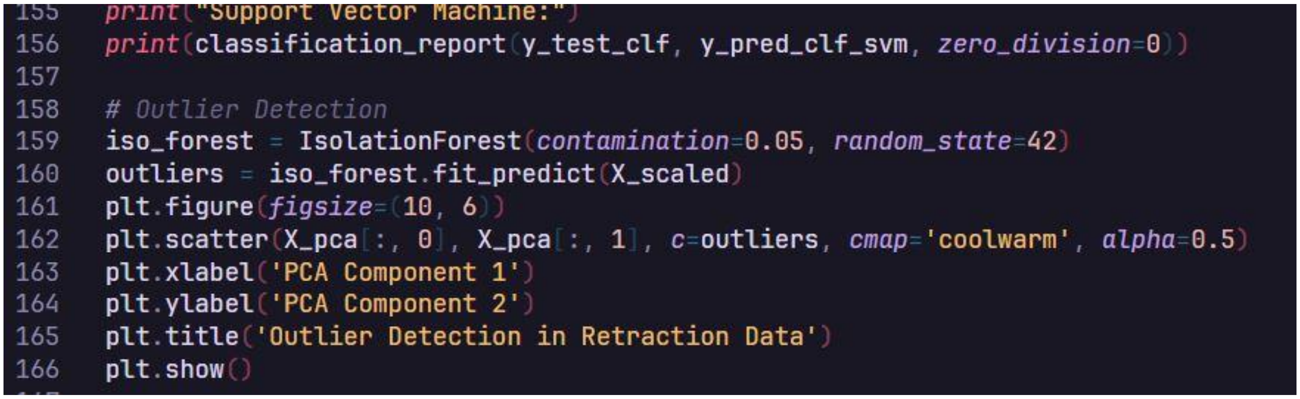
```python

data['Journal'] = reduce_cardinality(data['Journal'])
data['Publisher'] = reduce_cardinality(data['Publisher'])
data['Country'] = reduce_cardinality(data['Country'])
data['Author'] = reduce_cardinality(data['Author'])
data['Subject'] = reduce_cardinality(data['Subject'])
data['Institution'] = reduce_cardinality(data['Institution'])
data['Reason'] = reduce_cardinality(data['Reason'])

# encoding categorical variables
label_encoders = {}
categorical_columns = ['Journal', 'Publisher', 'Country', 'Author', 'RetractionNature', 'Reason', 'Paywalled', 'Subject',
for col in categorical_columns:
    le = LabelEncoder()
    data[col] = le.fit_transform(data[col].astype(str))
    label_encoders[col] = le

# checking the transformed data
print(data.head())

# checking the unique values of the target variable
print(f"Unique values in RetractionNature: {data['RetractionNature'].unique()}")

# If only one class in RetractionNature, choose a different target variable
if len(data['RetractionNature'].unique()) == 1:
    # trying using 'Reason' as the target variable instead
    target_variable = 'Reason'
else:
    target_variable = 'RetractionNature'

print(f"Using {target_variable} as the target variable for classification.")

# selecting relevant features for PCA and clustering
features = ['Subject', 'Institution', 'Journal', 'Publisher', 'Country', 'Author', 'ArticleType',
            'RetractionYear', 'OriginalPaperYear', 'RetractionNature', 'Reason', 'Paywalled', 'CitationCount']
X = data[features].values

```



```python
# standardizing the features
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# PCA
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_scaled)
plt.figure(figsize=(10, 6))
plt.scatter(X_pca[:, 0], X_pca[:, 1], alpha=0.5)
plt.xlabel('PCA Component 1')
plt.ylabel('PCA Component 2')
plt.title('PCA of Retraction Data')
plt.show()

# K-means Clustering
kmeans = KMeans(n_clusters=5, random_state=42)
clusters = kmeans.fit_predict(X_scaled)
plt.figure(figsize=(10, 6))
plt.scatter(X_pca[:, 0], X_pca[:, 1], c=clusters, cmap='viridis', alpha=0.5)
plt.xlabel('PCA Component 1')
plt.ylabel('PCA Component 2')
plt.title('K-means Clustering of Retraction Data')
plt.show()

# Multiple Regression
# Predicting CitationCount
X_reg = data[['RetractionYear', 'OriginalPaperYear', 'Journal', 'Publisher', 'Country', 'RetractionNature', 'Reason', 'Pay
y_reg = data['CitationCount'].values
X_train, X_test, y_train, y_test = train_test_split(X_reg, y_reg, test_size=0.2, random_state=42)
regressor = LinearRegression()
regressor.fit(X_train, y_train)
y_pred = regressor.predict(X_test)

plt.figure(figsize=(10, 6))
plt.scatter(y_test, y_pred, alpha=0.5)
plt.xlabel('Actual Citation Count')
plt.ylabel('Predicted Citation Count')
plt.title('Actual vs Predicted Citation Count')
plt.show()
```

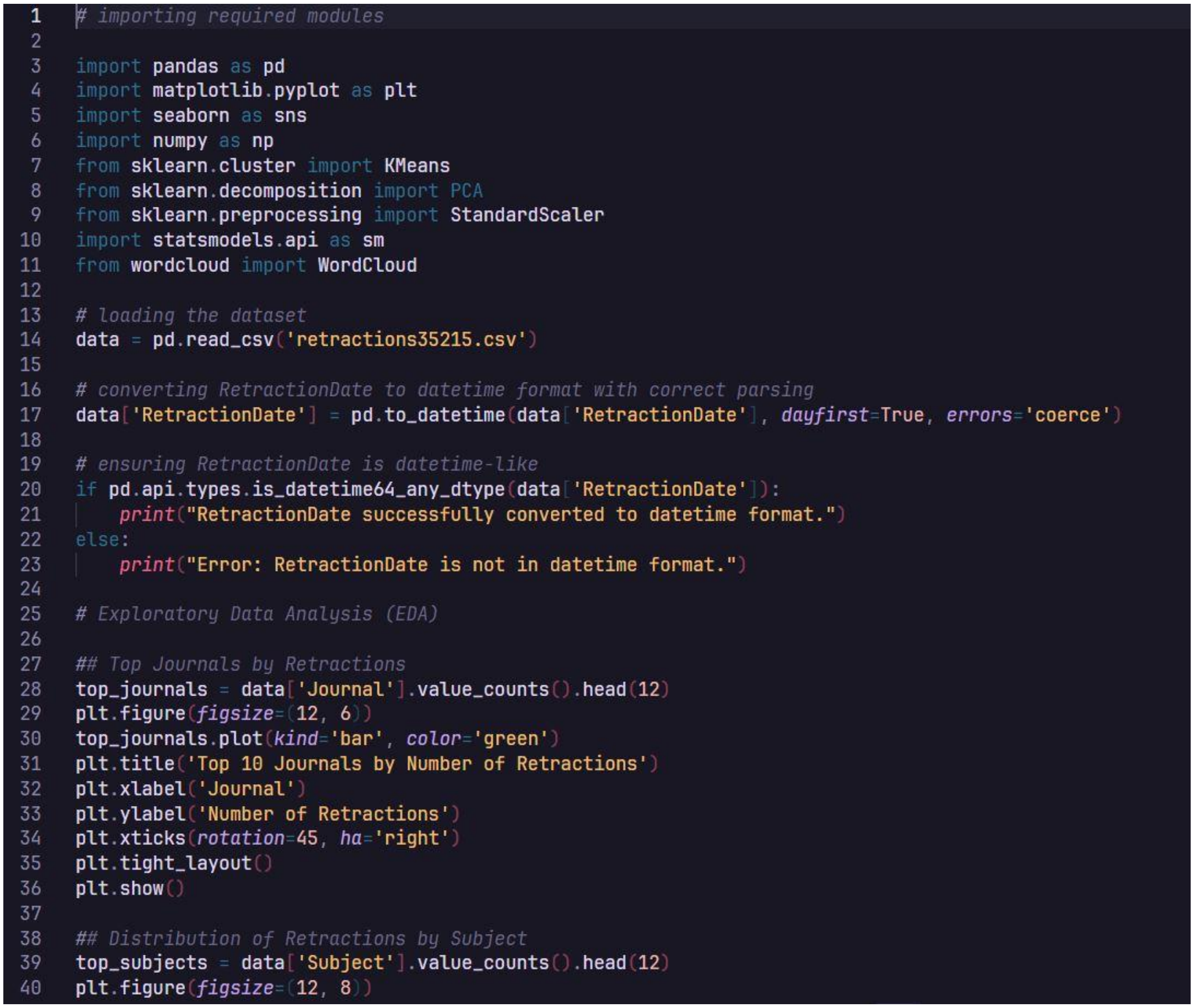


```
# Classification
# Predicting the chosen target variable
X_clf = data[['RetractionYear', 'OriginalPaperYear', 'Journal', 'Publisher', 'Country', 'Reason', 'Paywalled']].values
y_clf = data[target_variable].values

# Check the distribution of the target variable before split
print(f"Distribution of {target_variable} before split:")
print(data[target_variable].value_counts())

# Stratified split to maintain class distribution
X_train_clf, X_test_clf, y_train_clf, y_test_clf = train_test_split(X_clf, y_clf, test_size=0.2, random_state=42, stratif

# Check the distribution of the target variable after split
print(f"Distribution of {target_variable} in training set:")
print(pd.Series(y_train_clf).value_counts())

print(f"Distribution of {target_variable} in test set:")
print(pd.Series(y_test_clf).value_counts())

# Random Forest Classifier
from sklearn.ensemble import RandomForestClassifier
clf_rf = RandomForestClassifier(random_state=42)
clf_rf.fit(X_train_clf, y_train_clf)
y_pred_clf_rf = clf_rf.predict(X_test_clf)
print("Random Forest Classifier:")
print(classification_report(y_test_clf, y_pred_clf_rf))

# Categorical Naive Bayes
clf_nb = CategoricalNB()
clf_nb.fit(X_train_clf, y_train_clf)
y_pred_clf_nb = clf_nb.predict(X_test_clf)
print("Categorical Naive Bayes:")
print(classification_report(y_test_clf, y_pred_clf_nb, zero_division=0))

# Support Vector Machine (SVM)
clf_svm = SVC(random_state=42)
clf_svm.fit(X_train_clf, y_train_clf)
y_pred_clf_svm = clf_svm.predict(X_test_clf)
print("Support Vector Machine:")
```

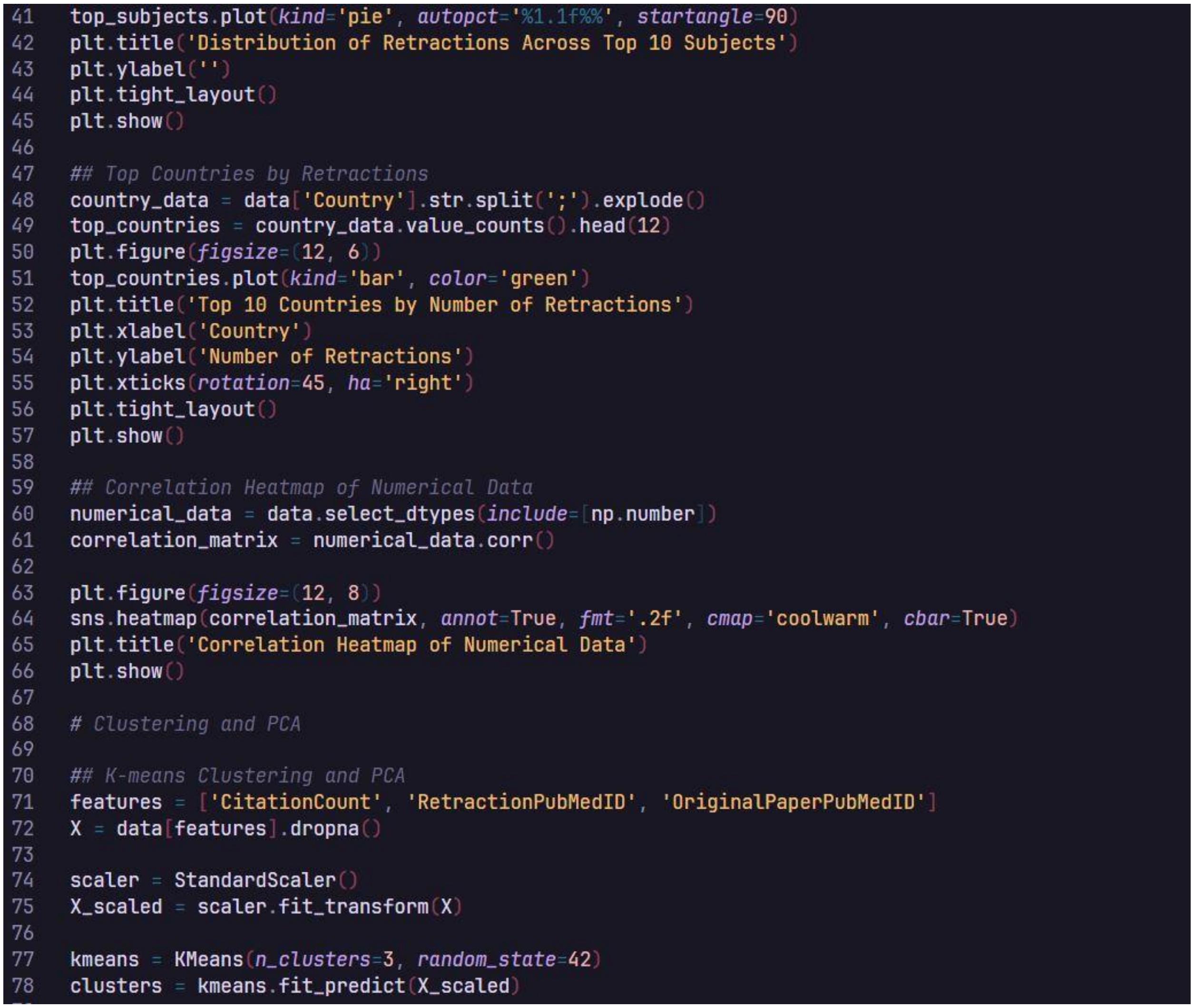


```
print("Support Vector Machine:")
print(classification_report(y_test_clf, y_pred_clf_svm, zero_division=0))

# Outlier Detection
iso_forest = IsolationForest(contamination=0.05, random_state=42)
outliers = iso_forest.fit_predict(X_scaled)
plt.figure(figsize=(10, 6))
plt.scatter(X_pca[:, 0], X_pca[:, 1], c=outliers, cmap='coolwarm', alpha=0.5)
plt.xlabel('PCA Component 1')
plt.ylabel('PCA Component 2')
plt.title('Outlier Detection in Retraction Data')
plt.show()
```

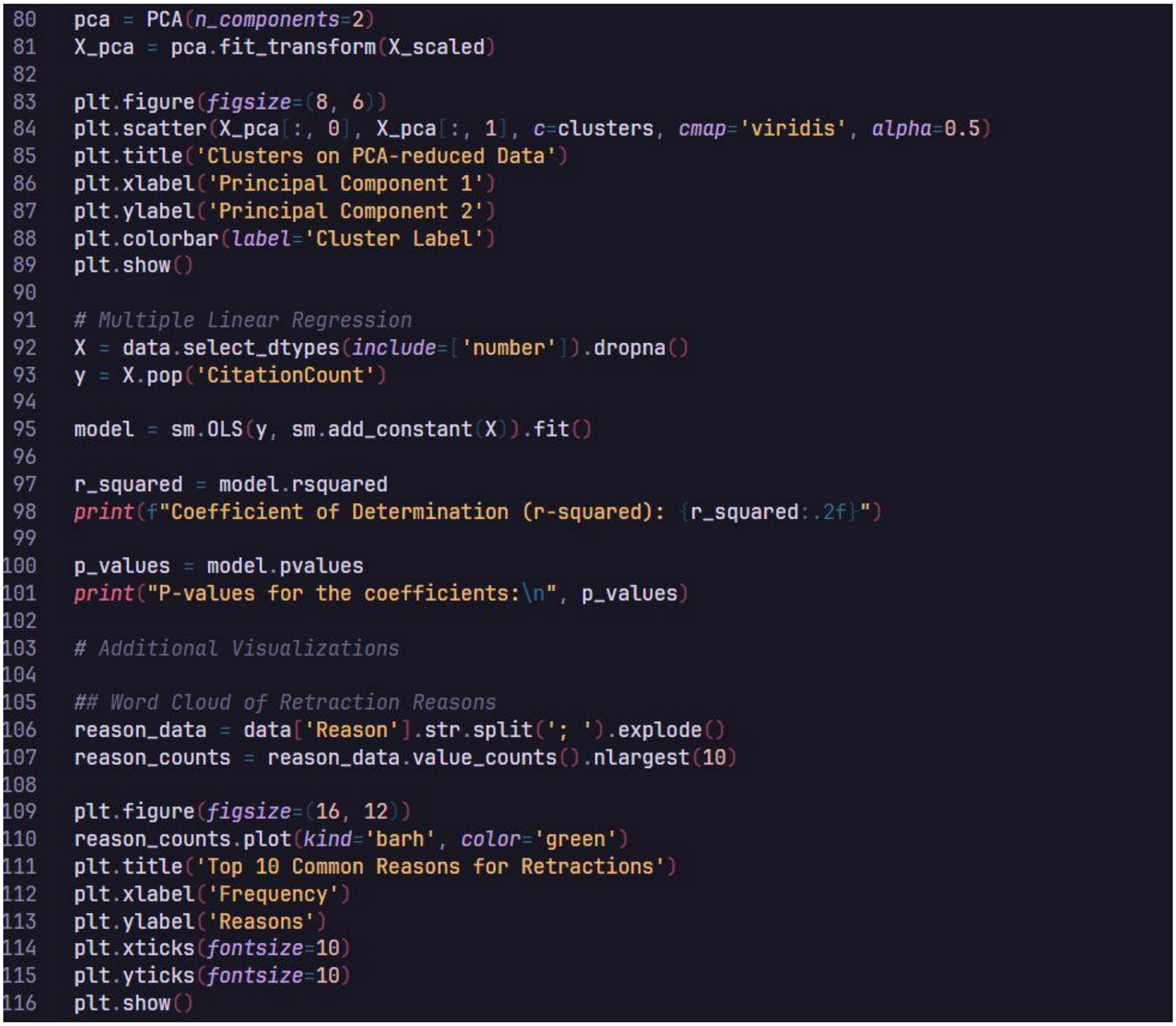
PRT564_Visualization.py



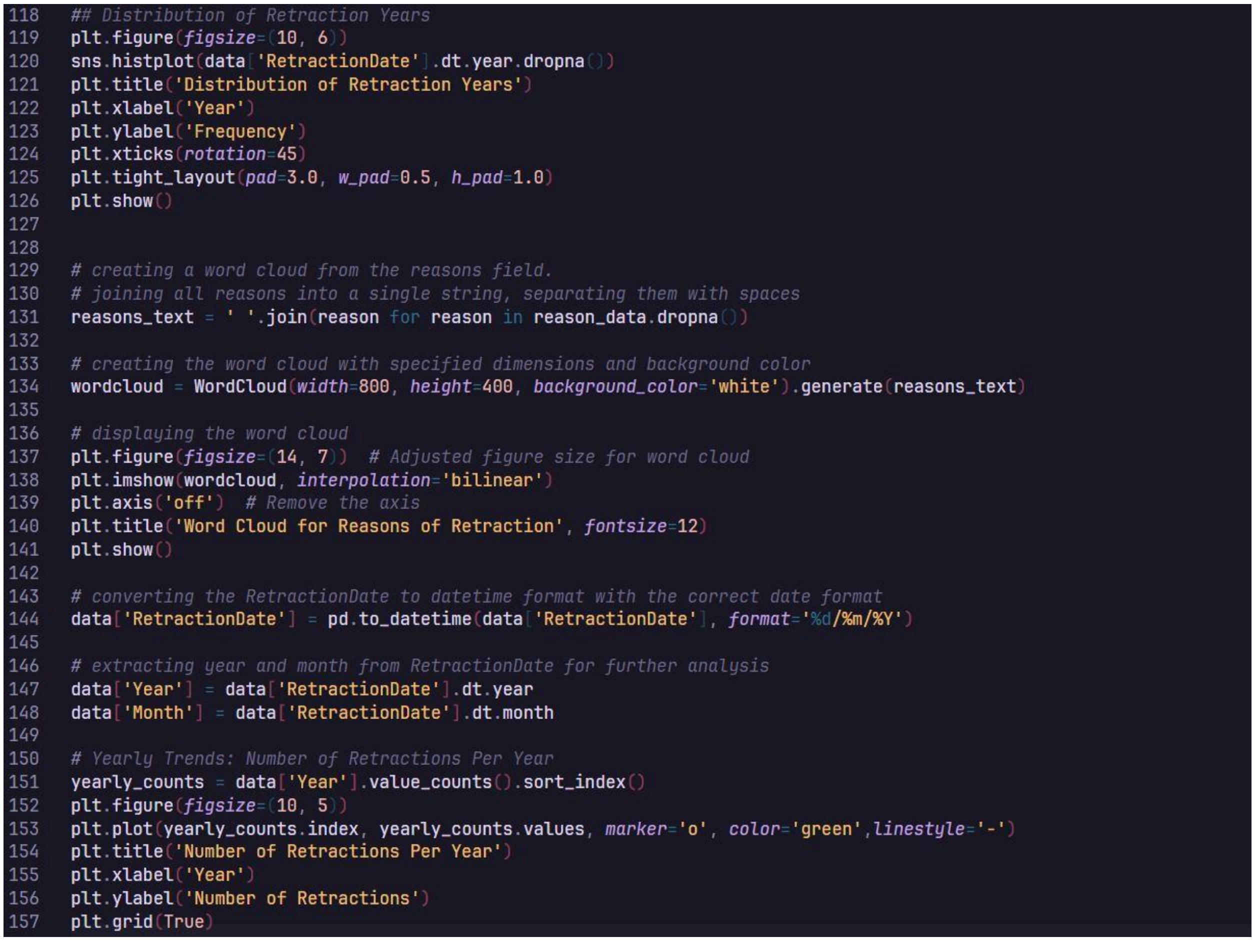
```python
# importing required modules

import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
import statsmodels.api as sm
from wordcloud import WordCloud

# loading the dataset
data = pd.read_csv('retractions35215.csv')

# converting RetractionDate to datetime format with correct parsing
data['RetractionDate'] = pd.to_datetime(data['RetractionDate'], dayfirst=True, errors='coerce')

# ensuring RetractionDate is datetime-like
if pd.api.types.is_datetime64_any_dtype(data['RetractionDate']):
    print("RetractionDate successfully converted to datetime format.")
else:
    print("Error: RetractionDate is not in datetime format.")

# Exploratory Data Analysis (EDA)

## Top Journals by Retractions
top_journals = data['Journal'].value_counts().head(12)
plt.figure(figsize=(12, 6))
top_journals.plot(kind='bar', color='green')
plt.title('Top 10 Journals by Number of Retractions')
plt.xlabel('Journal')
plt.ylabel('Number of Retractions')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()

## Distribution of Retractions by Subject
top_subjects = data['Subject'].value_counts().head(12)
plt.figure(figsize=(12, 8))
```



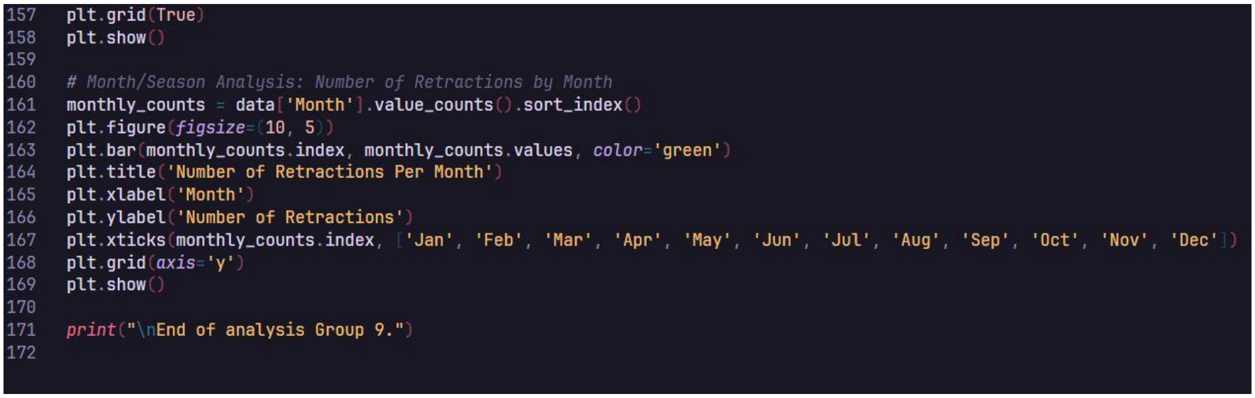
```
top_subjects.plot(kind='pie', autopct='%1.1f%%', startangle=90)
plt.title('Distribution of Retractions Across Top 10 Subjects')
plt.ylabel('')
plt.tight_layout()
plt.show()

## Top Countries by Retractions
country_data = data['Country'].str.split(';').explode()
top_countries = country_data.value_counts().head(12)
plt.figure(figsize=(12, 6))
top_countries.plot(kind='bar', color='green')
plt.title('Top 10 Countries by Number of Retractions')
plt.xlabel('Country')
plt.ylabel('Number of Retractions')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()

## Correlation Heatmap of Numerical Data
numerical_data = data.select_dtypes(include=[np.number])
correlation_matrix = numerical_data.corr()

plt.figure(figsize=(12, 8))
sns.heatmap(correlation_matrix, annot=True, fmt='.2f', cmap='coolwarm', cbar=True)
plt.title('Correlation Heatmap of Numerical Data')
plt.show()

# Clustering and PCA

## K-means Clustering and PCA
features = ['CitationCount', 'RetractionPubMedID', 'OriginalPaperPubMedID']
X = data[features].dropna()

scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

kmeans = KMeans(n_clusters=3, random_state=42)
clusters = kmeans.fit_predict(X_scaled)
```



```python
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_scaled)

plt.figure(figsize=(8, 6))
plt.scatter(X_pca[:, 0], X_pca[:, 1], c=clusters, cmap='viridis', alpha=0.5)
plt.title('Clusters on PCA-reduced Data')
plt.xlabel('Principal Component 1')
plt.ylabel('Principal Component 2')
plt.colorbar(label='Cluster Label')
plt.show()

# Multiple Linear Regression
X = data.select_dtypes(include=['number']).dropna()
y = X.pop('CitationCount')

model = sm.OLS(y, sm.add_constant(X)).fit()

r_squared = model.rsquared
print(f"Coefficient of Determination (r-squared): {r_squared:.2f}")

p_values = model.pvalues
print("P-values for the coefficients:\n", p_values)

# Additional Visualizations

## Word Cloud of Retraction Reasons
reason_data = data['Reason'].str.split('; ').explode()
reason_counts = reason_data.value_counts().nlargest(10)

plt.figure(figsize=(16, 12))
reason_counts.plot(kind='barh', color='green')
plt.title('Top 10 Common Reasons for Retractions')
plt.xlabel('Frequency')
plt.ylabel('Reasons')
plt.xticks(fontsize=10)
plt.yticks(fontsize=10)
plt.show()
```



```python
## Distribution of Retraction Years
plt.figure(figsize=(10, 6))
sns.histplot(data['RetractionDate'].dt.year.dropna())
plt.title('Distribution of Retraction Years')
plt.xlabel('Year')
plt.ylabel('Frequency')
plt.xticks(rotation=45)
plt.tight_layout(pad=3.0, w_pad=0.5, h_pad=1.0)
plt.show()


# creating a word cloud from the reasons field.
# joining all reasons into a single string, separating them with spaces
reasons_text = ' '.join(reason for reason in reason_data.dropna())

# creating the word cloud with specified dimensions and background color
wordcloud = WordCloud(width=800, height=400, background_color='white').generate(reasons_text)

# displaying the word cloud
plt.figure(figsize=(14, 7))  # Adjusted figure size for word cloud
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')  # Remove the axis
plt.title('Word Cloud for Reasons of Retraction', fontsize=12)
plt.show()

# converting the RetractionDate to datetime format with the correct date format
data['RetractionDate'] = pd.to_datetime(data['RetractionDate'], format='%d/%m/%Y')

# extracting year and month from RetractionDate for further analysis
data['Year'] = data['RetractionDate'].dt.year
data['Month'] = data['RetractionDate'].dt.month

# Yearly Trends: Number of Retractions Per Year
yearly_counts = data['Year'].value_counts().sort_index()
plt.figure(figsize=(10, 5))
plt.plot(yearly_counts.index, yearly_counts.values, marker='o', color='green',linestyle='-')
plt.title('Number of Retractions Per Year')
plt.xlabel('Year')
plt.ylabel('Number of Retractions')
plt.grid(True)
```



```python
plt.grid(True)
plt.show()

# Month/Season Analysis: Number of Retractions by Month
monthly_counts = data['Month'].value_counts().sort_index()
plt.figure(figsize=(10, 5))
plt.bar(monthly_counts.index, monthly_counts.values, color='green')
plt.title('Number of Retractions Per Month')
plt.xlabel('Month')
plt.ylabel('Number of Retractions')
plt.xticks(monthly_counts.index, ['Jan', 'Feb', 'Mar', 'Apr', 'May', 'Jun', 'Jul', 'Aug', 'Sep', 'Oct', 'Nov', 'Dec'])
plt.grid(axis='y')
plt.show()

print("\nEnd of analysis Group 9.")
```