**Results of the UACE Data Analysis (Years 2011 to 2015)**

**GROUP C**

FINAL RECESS

PROJECT

GITHUB LINK: <https://github.com/MickFrog/uace-analysis>

|  |  |  |  |
| --- | --- | --- | --- |
| SSENONO JORDAN MICHAEL | <https://github.com/MickFrog/recess-assignments> | 21/U/1013 | 2100701013 |
| OWOMUGISHA CRONNIE | <https://github.com/cronniegisha/recess> | 21/U/05910 | 2100705910 |
| AHUMUZA ARIYO NIMUSIIMA | <https://github.com/AriyoX/recess_assignments> | 21/U/1657 | 2100701657 |
| OJOK EMMANUEL NSUBUGA | <https://github.com/ojokne/recess-assignments> | 21/U/06816/PS | 2100706816 |
| SERUMAGA CONRAD DAVID | <https://github.com/conrad-coder/Learn-Python> | 22/U/6881 | 2200706881 |

# INTRODUCTION

The dataset under analysis presents a comprehensive record of Uganda Advanced Certificate of Education (UACE) exams results spanning from the year 2011 to 2015. Comprising a diverse range of schools located across different districts, this dataset offers valuable insights into the academic performance of students over these six years. The dataset is presented in a structured CSV format, with each row representing a specific school's performance in a given year. The columns encompass a wide array of metrics, including the number of candidates, division-wise results, and gender-specific performance indicators. These detailed records provide an opportunity to delve into the trends and patterns of educational achievements across various districts and schools.

The UACE results dataset includes a number of performance indicators, such as the total number of students who took the exam each year and the percentage distribution of students in various score levels. It also includes information on the students' schools, districts, and gender, allowing for a thorough analysis of academic trends and patterns.

Understanding and analyzing the UACE results can give important insights into the strengths and weaknesses of the education system, as education continues to be a crucial component of national development. This analysis tries to discover trends, pinpoint top-performing districts and schools, investigate academic performance by gender, and evaluate relationships between various performance variables. The results of this analysis can help Ugandan education policymakers, administrators, and other stakeholders make well-informed decisions to improve the standard of instruction and promote better academic achievements.

# KEY OBJECTIVES

**The initial goal is to comprehend the general trend in educational performance across different districts over a five-year period (2011 to 2015)**. To achieve this, we analyzed the dataset to identify patterns and changes in educational performance. One approach we used was creating line plots to visually represent the yearly student enrolment figures, which helped us detect any noticeable shifts or trends. This way, we could evaluate annual performance. We analyzed the academic results of districts and schools year over year in order to find any noticeable increases or decreases. The number of students overall and their performance metrics were shown using lines to show trends from 2011 to 2015.

**Determine if there are any gender-based differences in academic achievement**. Comparing the performance measures between male and female students allowed us to examine gender-based differences in educational accomplishment. Male and female students' academic performance was visualized and compared over time using line plots for performance above average as well as total students sitting the exams to identify any gender-based disparities. In this case, we took performance above average as performance above passing grade. (11 points in this case)

**Our third objective is to assess academic proficiency by studying how students are distributed across different scoring point ranges (%0-5, %6-10, %11-15, %16-20, %21-25)**. To achieve this, we analyzed how pupils were distributed in these ranges. For example, we used pie charts to visualize the proportion of students within the worst scoring range for the top-performing districts. This allowed us to gain insights into the academic proficiency levels of the students.

**Assessing Yearly Variation and Consistency in Key School Metrics**. This objective involves utilizing a correlation matrix to examine whether a stable link exists between the annual count of candidates and the corresponding number of above-average performers. A positive correlation would indicate that as candidate numbers change yearly, a corresponding shift occurs in above-average performers. This provides insights into how student population fluctuations impact above-average academic achievements and gauges performance consistency across years.

**Finding outliers is the fifth goal**. Using a box plot that uses percentage of students above average performance for schools. It uses metrics such as the upper and lower quartiles, the median, the mean and the outliers are shown as data points below and above the lower and upper quartile respectively.

**Developing a Predictive Model for Future Results**. An essential objective of this analysis was to construct a robust predictive model capable of projecting future academic outcomes based on the historical UACE exam results from 2011 to 2015. By leveraging the comprehensive dataset encompassing diverse performance metrics, including candidate counts, score distributions, and gender-specific indicators, the aim was to create a model that could offer insights into forthcoming educational achievements.

# FEATURES ANALYZED

## Year Columns:

'2011 Total', '2012 Total', '2013 Total', '2014 Total', and '2015 Total' in the year columns. The total number of students who took the UACE exams in each corresponding year is shown in these columns. By examining these columns, we can realize any changes in the number of exam-takers throughout the course of the five-year period and analyze trends in enrolment.

## Percentage Score Range Columns:

'%0-5 Points', '%6-10 Points', '%11-15 Points', '%16-20 Points', '%21-25 Points'. The percentage distribution of students for each year falling within particular score ranges is shown in these columns. By examining these columns, we may determine the percentage of students performing at various levels of achievement as well as the academic performance of the students based on score ranges. In this case, we used performances above average (point ranges of '%11-15 Points', '%16-20 Points', '%21-25 Points')

## District\_Name Column:

The names of the districts in which the schools are located are listed in the 'District\_Name' column. We may organize and compare data by district using this column's analysis, which makes it easier to determine which districts are performing best and worst.

## Gender Column:

Within the dataset, the gender column categorizes students as male or female. This classification serves as a pivotal point of investigation to discern gender-based academic achievements within the UACE results. By scrutinizing this column, we gain a comprehensive understanding of the educational performance landscape through a gender-centric lens.

# PROCESSES AND TECHNIQUES

## Data Cleaning:

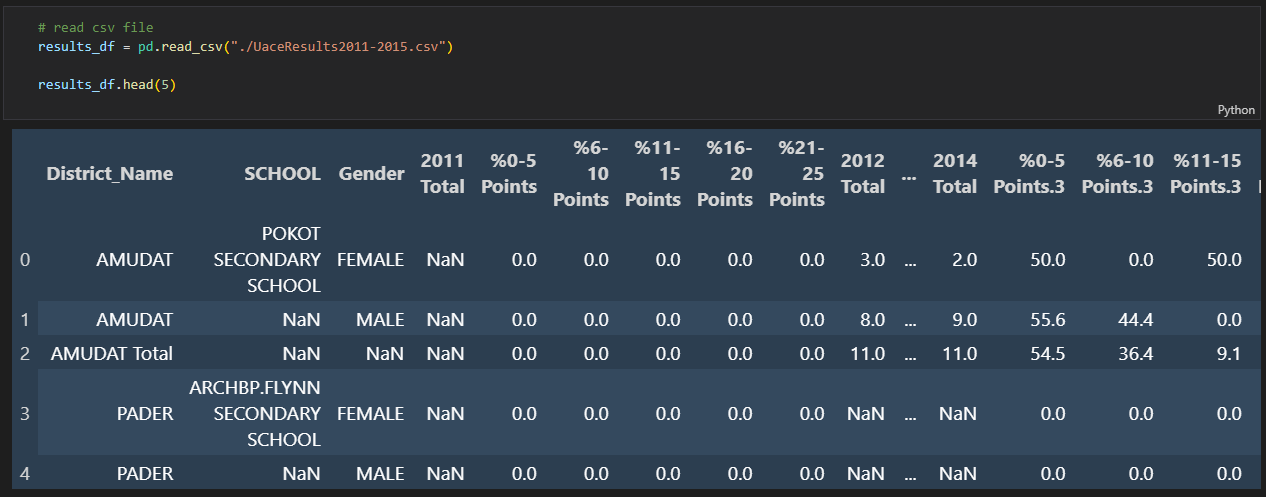


Figure 1: Display of first 5 rows of the dataset before cleaning.

To assure the correctness and dependability of our findings, we had to clean the dataset before beginning the data analysis. Any data analysis process must start with data cleaning in order to deal with missing numbers, fix inconsistencies, and get rid of duplicates. We carried out the following processes when carrying out data cleaning.

## Identifying and Dealing with Missing Values:

Absence of data can add bias and compromise the accuracy of the data analysis. We looked for missing values in the dataset and decided how to handle them based on our findings. For instance, we discovered that several entries' 'SCHOOL' columns contained NaN values. We decided to remove these entries and create a dataframe without NaN values as they would have interrupted our analysis. We also changed the totals of schools with no students from NaN to 0. Genderless entries that were not district totals were removed as well as they have no value in the analysis.

Impact: It is crucial to deal with missing data properly since it can produce biased or erroneous analysis results. We made sure that all data points were connected to the correct districts and schools by utilizing interpolation or references to other datasets to fill in any missing district and school names.

## Handling Duplicates

The dataset contains duplicate performance measurements and entries for some years. After the year 2011, the complier handled these duplicate columns and added ".1", ".2", and so forth attached to them. Any other potential duplicate rows and columns were dropped.

Impact: Duplicate entries may cause data points to be overrepresented, which could influence statistical analysis. We reduced superfluous data by deleting duplicates, ensuring that every data point was distinct and unique.

## Fixing Data Types

We checked each column's data types and converted them to the proper formats. Categorical columns were set to categorical data types and numerical columns were transformed to respective numerical data types.

Impact: Accurate computations and analysis depend on using the right data types. We made sure that mathematical operations and aggregations could be executed correctly by transforming categorical columns to categorical data types and numerical columns to numerical data types.

## Handling Misplaced Schools

We realized that some schools were incorrectly placed in certain rows with their respective genders in rows below them, without the school’s name being entered besides it. An algorithm was written to check whether the school column of a row was empty and the rows was also not for a district total. It works by copying the school name of the above row in to the row being targeted.

Impact: It simplified the process of data analysis in general.

## Overview:

We enhanced the quality of the UACE results dataset by using those data cleaning approaches, making it more trustworthy and analytically appropriate. Each method tackled a different data problem, resulting in a cleaner and more reliable dataset. As a result, we felt confident moving forward with our study because we knew the data was correct and ready for additional investigation and visualization. The inventive application of such methods demonstrated our attention to detail and dedication to deriving trustworthy and valuable insights from the data. We remained with 2 data frames moving forward, one with the district totals for operations involving districts, while the other removes district totals by removing genderless entries.

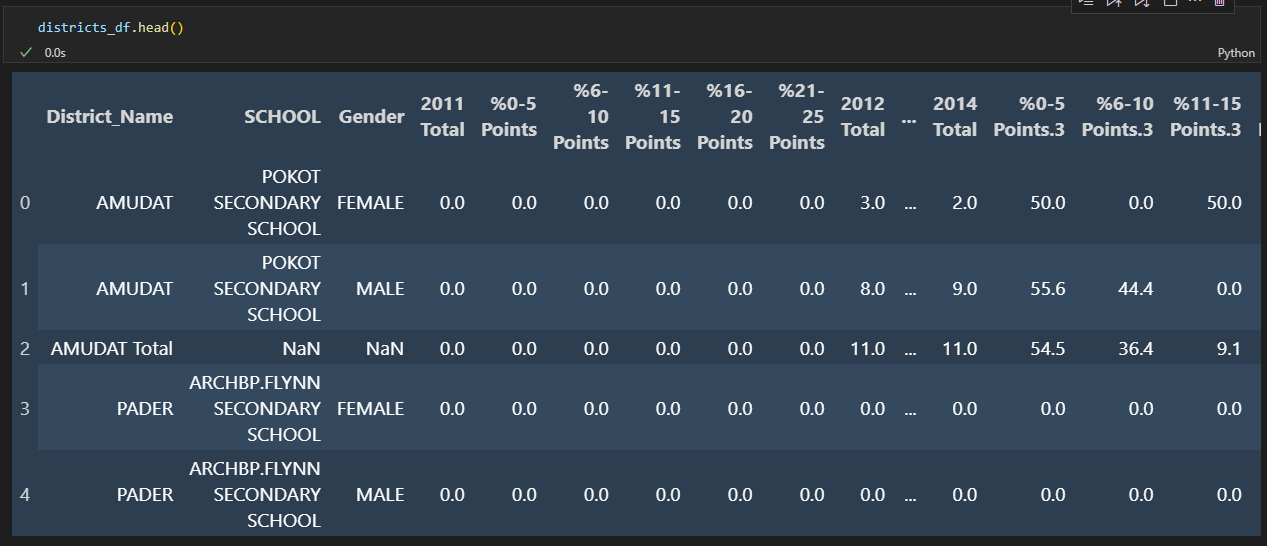


Figure 2: Screenshot showing the first 5 rows of the district totals dataset after cleaning.

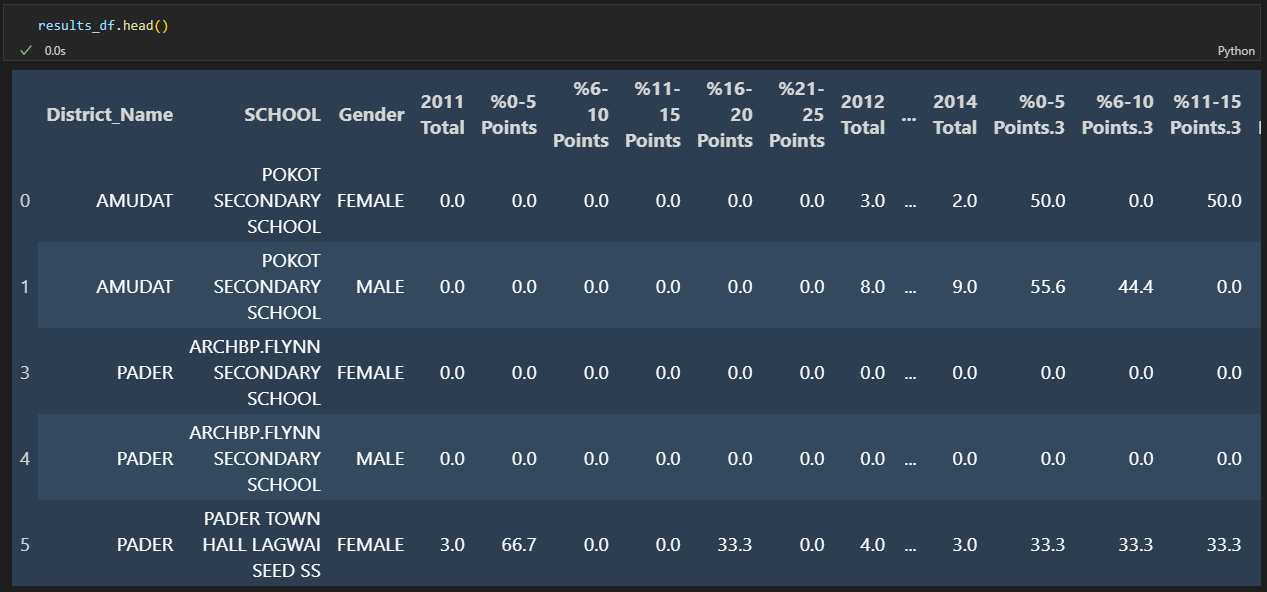


Figure 3: Screenshot showing the first 5 rows of the dataset without district totals after cleaning.

# DATA ANALYSIS TECHNIQUES AND VISUALIZATIONS

In order to comprehend the underlying patterns and trends in the UACE outcomes dataset and to obtain deeper insights during the data analysis process, we used a variety of visualization techniques. Each graphic had a specific function and offered useful data.

## Line Plots:

Understanding the overall educational performance trend across years was made possible thanks to line plots. We are able to see the overall pattern in student enrolment over time by showing the annual total of students, among other visualizations which are in the notebook. These line plots assisted us in identifying any notable shifts or fluctuations in the student population over time, allowing us to spot trends in educational growth or decline. Some examples of line plots generated are visualized below. More are present in the notebook.

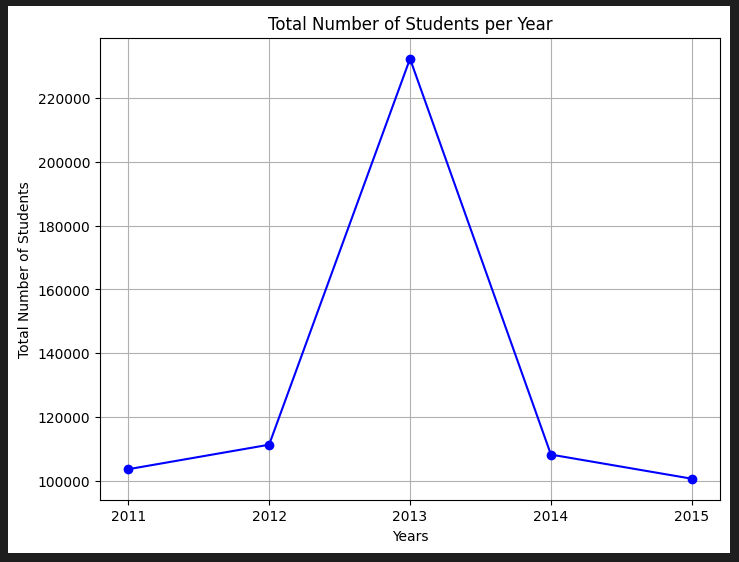


Figure 4: Line plot showing the trend in enrolment for UACE.

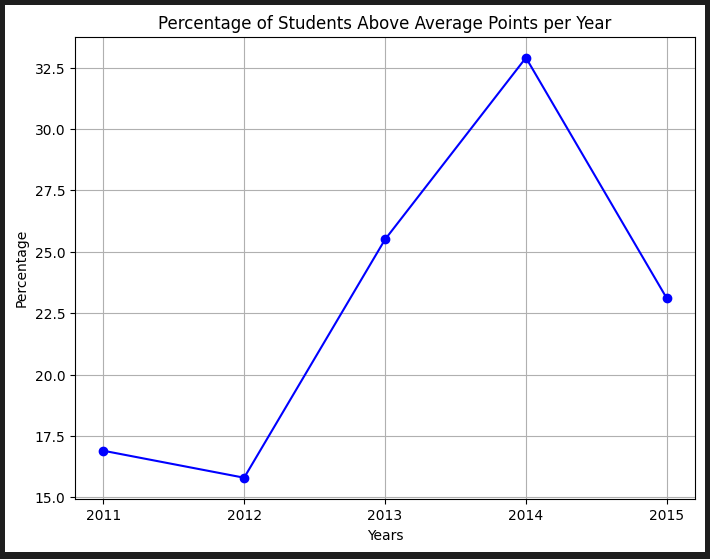


Figure 5: Line plot showing the trend of students’ performance using percentage of students above the passing grade (11points).

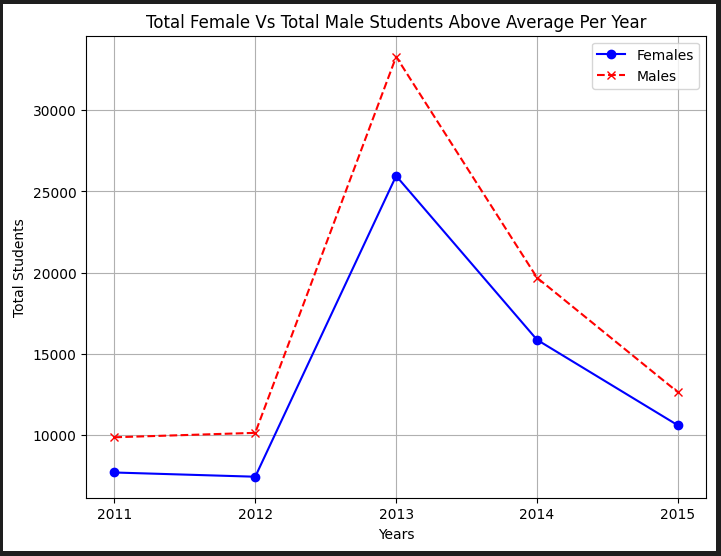


Figure 6: Line plot showing the trend of male vs female performance over 2011-2015.

## Maps

In the pursuit of comprehensive data visualization, geographical information also plays a pivotal role in revealing patterns and trends that might remain concealed in tabular formats. Maps offer an intuitive and visually engaging way to present data that is inherently spatial. In this context, maps were instrumental in showing the distribution of districts based on their academic performance, highlighting the locations of both the best and least performing districts across the country. They were used in the notebook as well. To achieve this, geospatial tools like Geocoder and Folium were applied. Geocoder, a Python library, facilitated the conversion of textual location information, such as district names, into geographic coordinates such as latitude and longitude. This step was crucial in positioning districts accurately on the map. Folium, another Python library, was employed to generate these interactive and customizable maps. These maps provide insights from the data and can show which regions have the better and worse performing schools and districts.

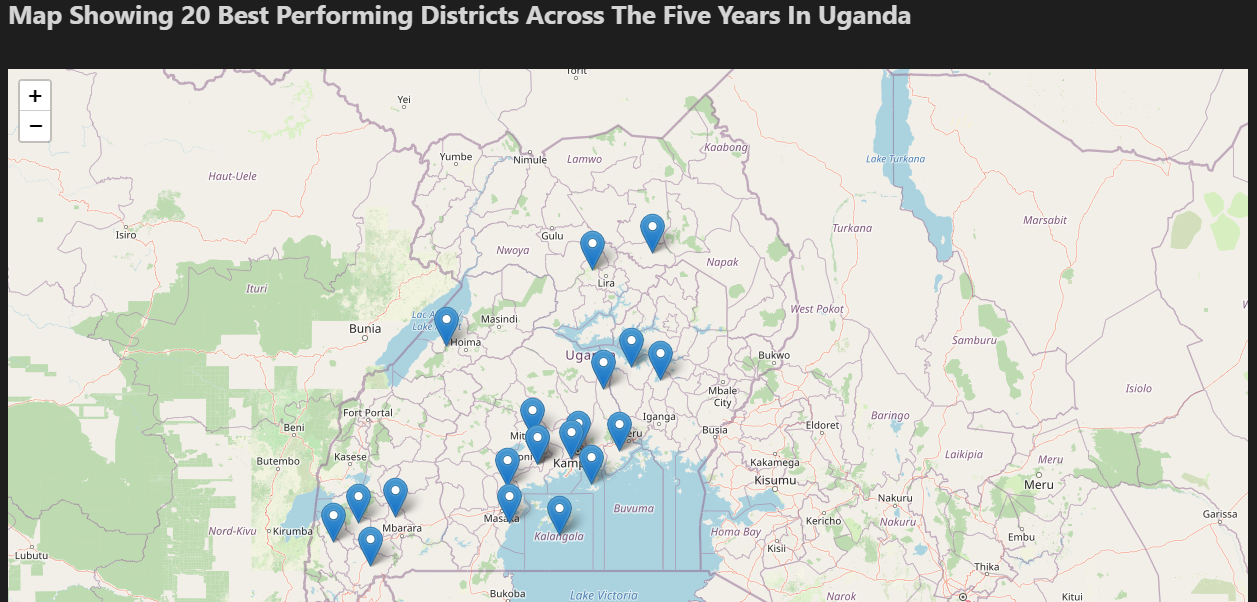


Figure 7: Map highlighting the positions of the 20 best performing countries across the country.



Figure 8: Map highlighting the positions of the 20 least performing countries across the country.

## Pie Charts

The distribution of students across various scoring ranges was shown graphically using a pie chart. We sorted the data by district names using the '%0-5 Points' column, and then computed the total number of students in each category. This was done in order to see the academic competency across the top 20 performing districts.

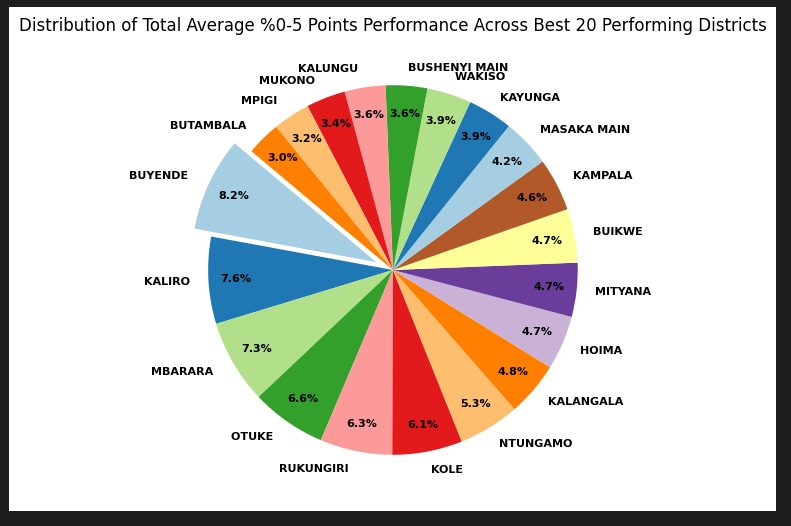


Figure 9: Pie chart showing the distribution of the total 0-5% points attained across the top 20 performing districts.

## Tables

Tables provided a straightforward and structured approach to visualize key data insights, particularly when highlighting the best or worst performing schools and districts. This method of presentation simplifies complex information, making it accessible at a glance. The dataset's performance metrics, such as scoring ranges and above-average counts, were leveraged to identify the top-performing and lowest-performing schools, same with the districts. These metrics allowed for direct comparisons of academic achievements. They were also able to show the districts with the list and most students across particular years in a tabular format which made the data easier to understand. Most of the data being analyzed was put in tabular format, and below are some of the tables that were generated.

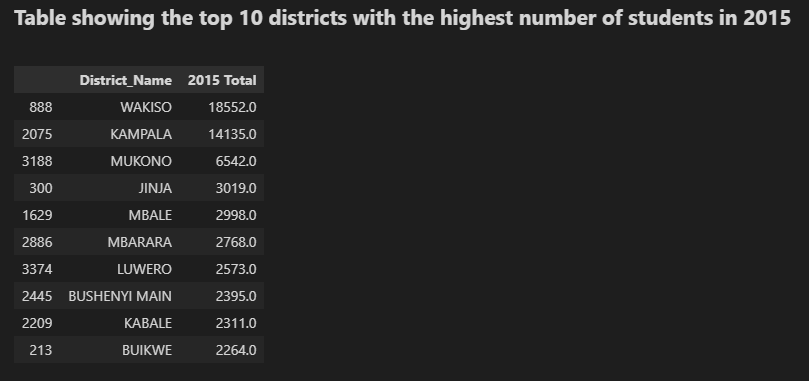


Figure 10: Table showing the top 10 districts with the highest number of students in 2015.

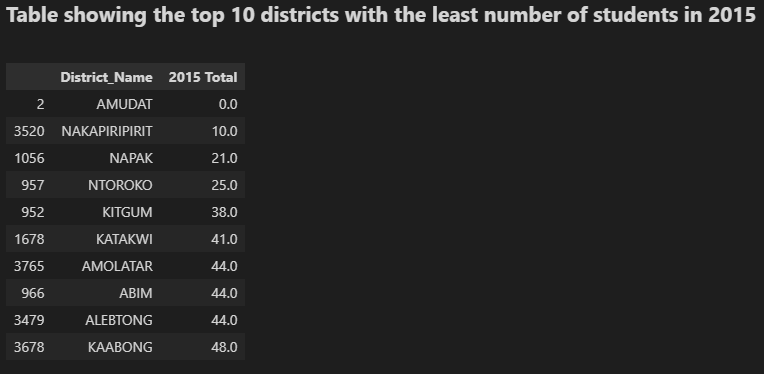


Figure 11: Table showing the top 10 districts with the least number of students in 2015.

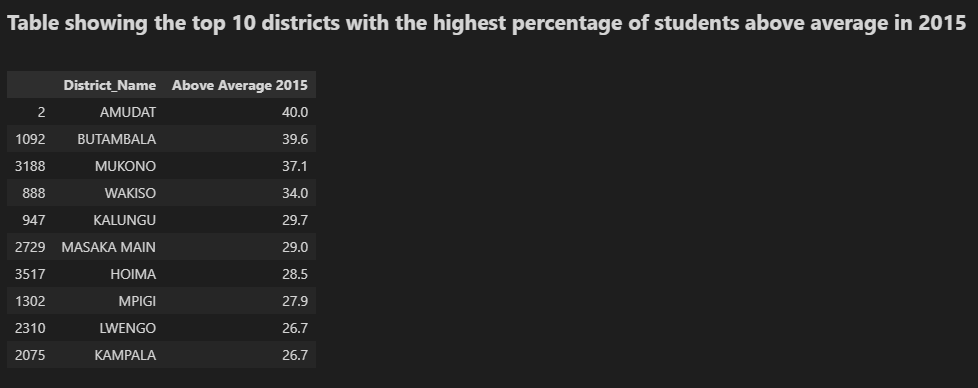


Figure 12: Table showing the top 10 districts with the highest percentage of students above average in 2015.



Figure 13:Table showing the 20 Best performing districts across all the years.



Figure 14: Table showing schools with the least performing schools across all the years.

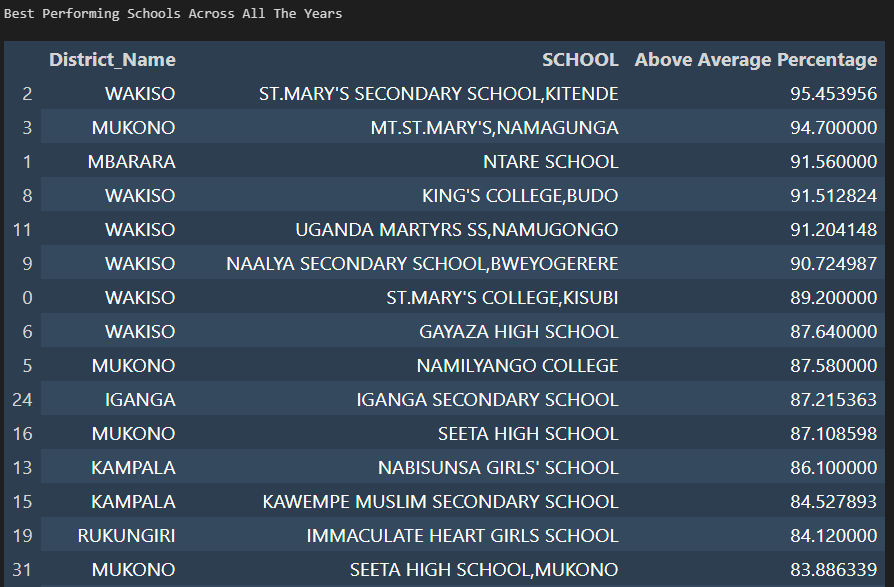


Figure 15: Table showing the best performing schools across all the years.

## Box Plot

The box plot proved to be a valuable visualization tool for identifying and displaying outliers within the dataset. Outliers are data points that deviate significantly from the majority of the data points and can provide insights into unusual or exceptional cases.

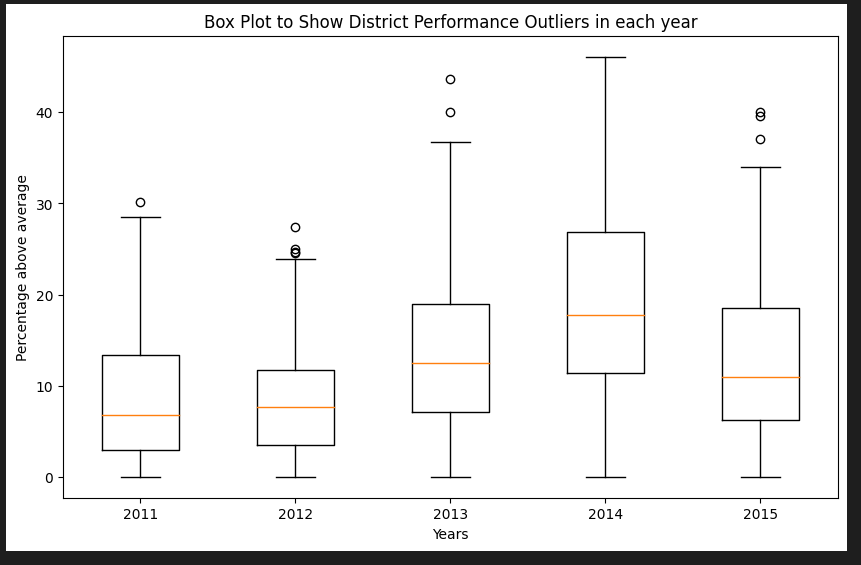


Figure 16: Box Plot to Show District Performance Outliers in each year.

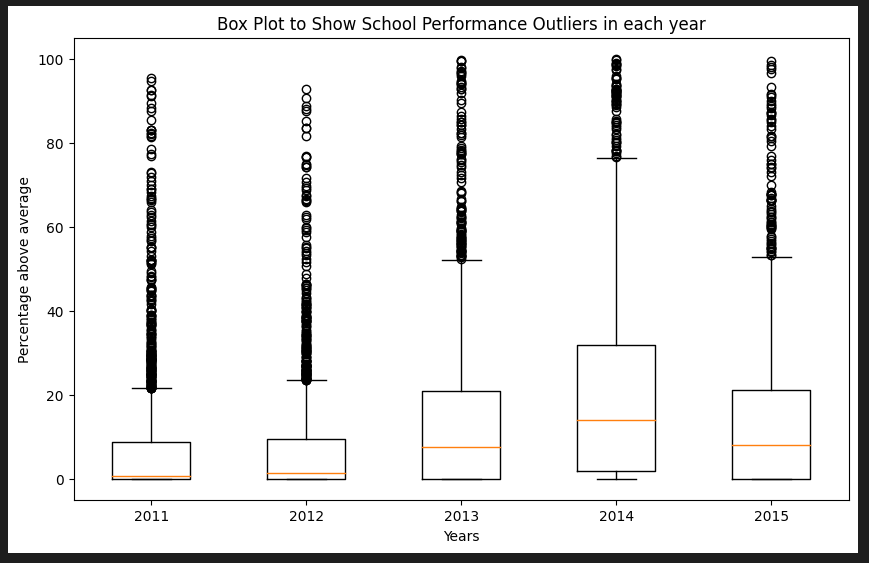


Figure 17:Box Plot to Show School Performance Outliers in each year.

## Heatmaps

The heatmap was an effective technique for displaying the performance metrics correlation matrix. We determined the correlation coefficients between various metrics by choosing pertinent performance-related columns. The heatmap used a color spectrum to display correlations, with warmer hues denoting positive correlations and cooler hues denoting negative correlations. We were able to determine performance parameters with this visualization.

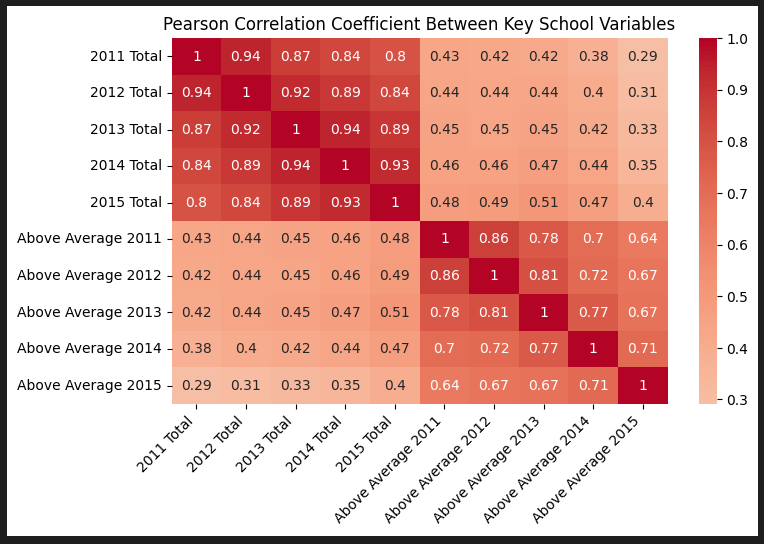


Figure 18: Heatmap showing the relationships between the key features of the dataset.

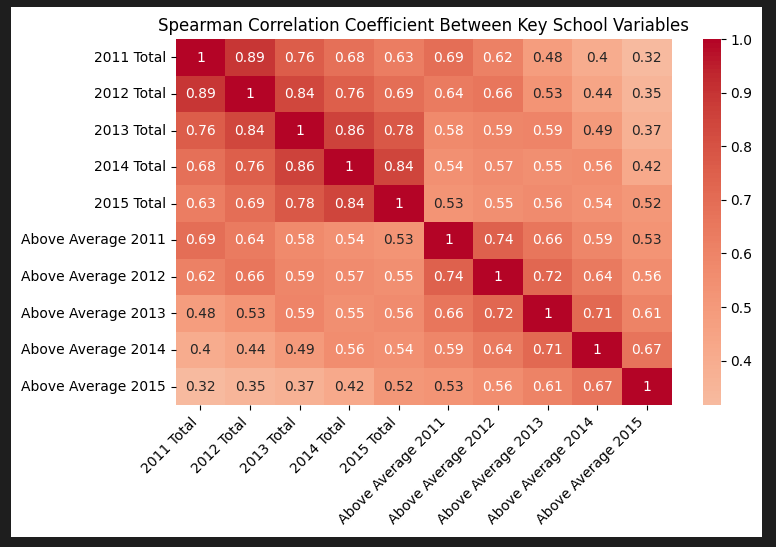


Figure 19:Heatmap showing the relationships between the key features of the dataset.

## Time Series Analysis and Predictive Modelling

Time series analysis was employed as a powerful technique to uncover patterns and trends based on historical UACE exam results from 2011 to 2015. This approach allowed us to harness the temporal aspects of the data to identify recurring patterns and seasonality. By leveraging time series analysis, we aimed to gain insights into the underlying dynamics that influence academic performance over time.

Within this context, we utilized predictive modelling to develop a model capable of projecting future academic outcomes. The model was trained on the historical data, learning from the relationships and patterns present in the UACE results. By understanding these patterns, the model could then generate predictions for future years, such as 2016, providing an informed estimation of the potential academic achievements.

One of the visualization outcomes of this technique was the generation of a line graph showcasing predicted data for 2016. This line graph allowed for a comparison between actual historical data and the model's predictions. The intersection of these lines offered valuable insights into the accuracy of the model and its capability to foresee future outcomes based on past trends.

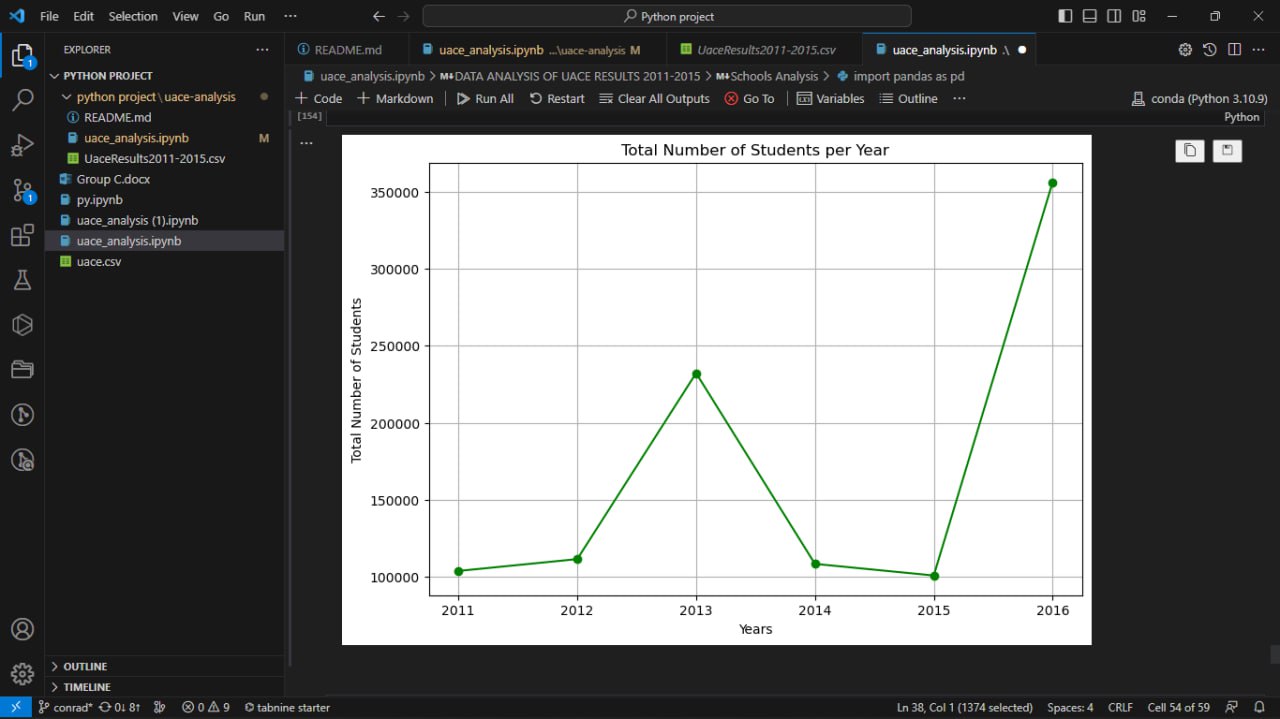


Figure 20: A line plot showing predicted data(2016) for the number of students that will enrol and sit UACE.

# DEDUCTIONS AND CONCLUSIONS

1. Trends in Educational Performance

Analysis of the dataset's line plots revealed fluctuating patterns in student enrolment from 2011 to 2015. These trends could be attributed to various factors, such as changing demographics, educational policies, or socioeconomic conditions. While some years exhibited growth in student enrolment, others experienced a decline. This underscores the importance of understanding the underlying dynamics influencing student participation.

1. Gender-Based Academic Achievement

The analysis of gender-based academic achievement indicated that male students consistently outperformed female students across the years. This trend could stem from various factors, including differences in study habits, access to resources, and teaching methodologies. Further investigation into these factors could help uncover the reasons behind the observed gender disparities in academic performance.

1. Academic Proficiency by Score Ranges

The pie charts visualized the distribution of students across different score ranges. This allowed us to assess the proportion of students performing within specific performance bands.

1. Yearly Variation and Consistency in School Metrics

The correlation matrix analysis aimed to uncover whether a consistent relationship existed between the annual count of candidates and the corresponding number of above-average performers. This analysis provided insights into how changes in student population over the years impacted above-average academic achievements. A positive correlation would indicate a stable link, while a weaker correlation could imply performance inconsistencies.

1. Outliers in District and School Performance

Box plots were effective in detecting outliers within the dataset, showcasing districts or schools that significantly deviated from the norm in terms of above-average performance. These outliers provide valuable insights into exceptional cases, helping to identify districts or schools that excel or struggle in terms of academic achievements.

1. Geospatial Insights from Maps

Geospatial analysis using maps enabled the visualization of district and school performance across the country. Highlighting the best and worst performing districts revealed geographic patterns in educational outcomes.

1. Relationships Between Performance Metrics

The heatmap depicted correlations between various performance metrics. Positive correlations suggested that certain metrics tended to move together, while negative correlations indicated an inverse relationship. This analysis provided a comprehensive overview of how different performance parameters interacted, offering insights into potential areas of focus for improvement.

In conclusion, the analysis of the UACE results dataset unveiled a multitude of insights into Uganda's educational landscape. It highlighted trends, disparities, and potential areas for improvement. The combination of data cleaning, visualization techniques, and statistical analysis allowed for a comprehensive understanding of academic performance across districts and schools. These findings can serve as a foundation for informed decision-making by education policymakers, administrators, and other stakeholders aiming to enhance the quality of education and promote academic excellence in Uganda. The GitHub repository (<https://github.com/MickFrog/uace-analysis>) and the contributions of the mentioned individuals underscore the collaborative effort in data analysis and insights derivation.