MAKERERE  UNIVERSITY

SCHOOL OF COMPUTING AND INFORMATION TECHNOLOGY

FINAL REPORT DOCUMENT

UACE RESULTS 2011-2015

DATA SCIENCE PROJECT BY GROUP D

GitHub Link:
https://github.com/Muhangi2/DataScience_UACE_Analysis

Name	Registration Number	Student Number
Prajapati Heli Anilkumar	21/X/20194/PS	2100720194
Nankunda Josephine	21/U/0843	2100700843
Mugoya Moses	21/U/10447/PS	2100710447
Muhangi Elioda	22/U/6342	2200706342

INTRODUCTION ABOUT THE DATASET

The dataset shows the district and school performance in the Uganda Advanced Certificate of Education (UACE) administered by the Uganda National Examinations Board (UNEB) from 2011 to 2015.

The performance is indicated as the percentage of candidates who scored in each range of points that include 0-5, 6-10, 11-15, 16-20, 21-25. Gender breakdown is also indicated per school.

We provide some analysis to try and get a better understanding of the results. The results are listed by school and show the performance in terms of how many students passed in a particular division.

DataFrame before cleaning

	District_Name	SCHOOL	Gender	2011 Total	%0-5 Points	%6-10 Points	%11-15 Points	%16-20 Points	%21-25 Points	2012 Total	...
0	AMUDAT	POKOT SECONDARY SCHOOL	FEMALE	NaN	0.0	0.0	0.0	0.0	0.0	3.0	...
1	AMUDAT	NaN	MALE	NaN	0.0	0.0	0.0	0.0	0.0	8.0	...
2	AMUDAT Total	NaN	NaN	NaN	0.0	0.0	0.0	0.0	0.0	11.0	...
3	PADER	ARCHBP.FLYNN SECONDARY SCHOOL	FEMALE	NaN	0.0	0.0	0.0	0.0	0.0	NaN	...
4	PADER	NaN	MALE	NaN	0.0	0.0	0.0	0.0	0.0	NaN	...

Features to be analyzed

1. Importing libraries

We started by importing the libraries like pandas, matplotlib, NumPy, seaborn, etc. as they include external code modules into your Python program to perform tasks.

2. Data Cleaning

Then we did data cleaning which is the process of Handling Missing Values, Removing Duplicates, Standardizing and Formatting, Correcting Inaccuracies, etc.

3. Columns analyzed

Gender – It contained some wrong data and null values so we first filled the null values with Total and then defined a list defining correct values [male, female, Total] and dropped all the others that aren't in the list.

2015_Total – It was of type object so we converted it to float.

School – The column contained null values so we filled the Null values by just the previous value using the function fillna.

District_Name - It had some rows containing total which don't qualify as districts so we extracted them and created a new csv to store those rows.

4. Analysis on each year (2011 TO 2015)

Make analysis on specific years according to performance at district level, performance at school level, performance at gender basis and performance at grade level of points.

5. General performances in the range of 5-years

Here we shall majorly be focusing on the trends in: Rank, performance at gender basis, performance at district level, performance at grade level of points, total students who sat per year, check for how many Schools Are in District, performance at year level, map for presentation for the top performing school.

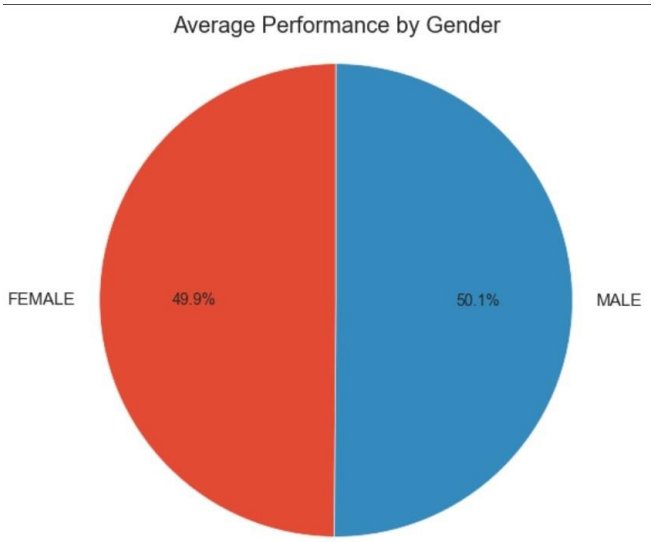
Data after cleaning

	District_Name	SCHOOL	Gender	2011_Total	%0-5_Points	%6-10_Points	%11-15_Points	%16-20_Points	%21-25_Points	2012_Total
0	AMUDAT	POKOT_SECONDARY_SCHOOL	FEMALE	0.0	0.0	0.0	0.0	0.0	0.0	3.0
1	AMUDAT	POKOT_SECONDARY_SCHOOL	MALE	0.0	0.0	0.0	0.0	0.0	0.0	8.0
3	PADER	ARCHBP.FLYNN_SECONDARY_SCHOOL	FEMALE	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	PADER	ARCHBP.FLYNN_SECONDARY_SCHOOL	MALE	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	PADER	PADER_TOWN_HALL_LAGWAI_SEED_SS	FEMALE	3.0	66.7	0.0	0.0	33.3	0.0	4.0
6	PADER	PADER_TOWN_HALL_LAGWAI_SEED_SS	MALE	17.0	41.2	52.9	0.0	5.9	0.0	22.0
7	PADER	PAJULE_SECONDARY_SCHOOL	FEMALE	0.0	0.0	0.0	0.0	0.0	0.0	5.0
8	PADER	PAJULE_SECONDARY_SCHOOL	MALE	0.0	0.0	0.0	0.0	0.0	0.0	19.0
10	SERERE	KAMOD_SECONDARY_SCHOOL	FEMALE	0.0	0.0	0.0	0.0	0.0	0.0	0.0

OBJECTIVES

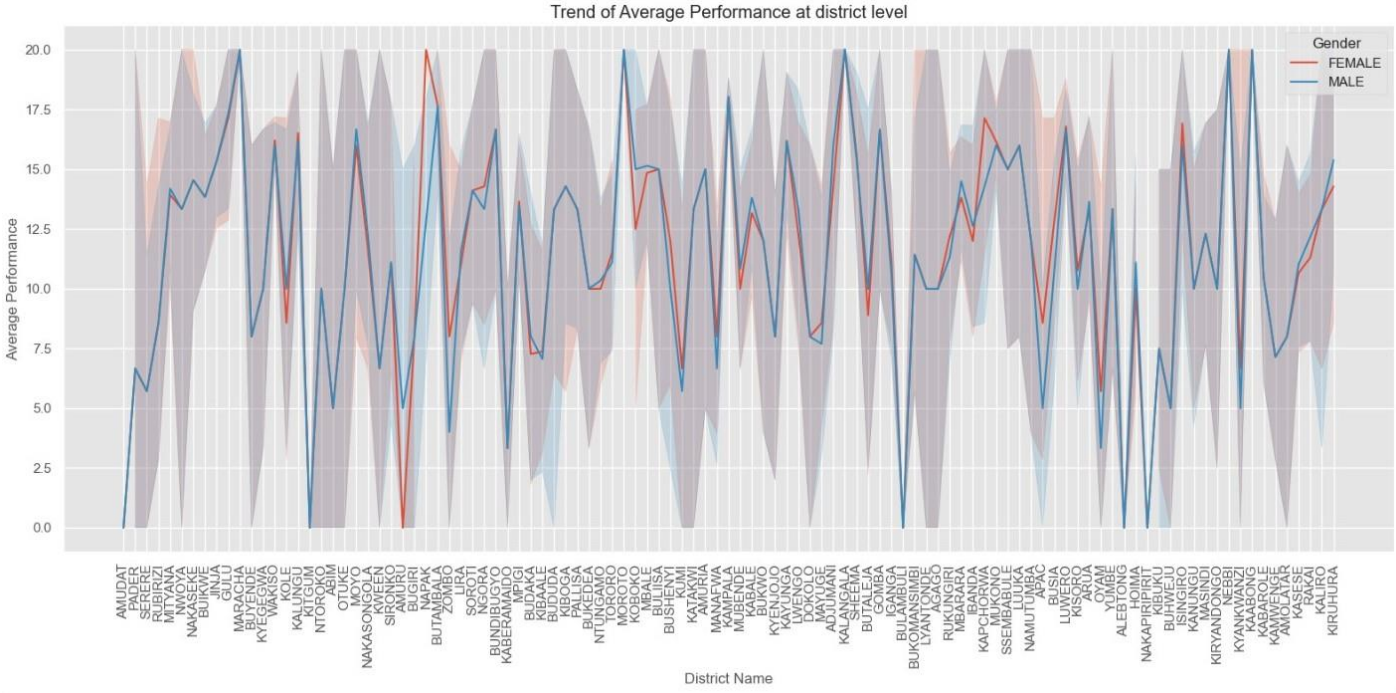
✓ Performance at gender basis

The objective is to analyze the academic performance of students based on gender using a pie chart. It helps identify any areas where specific gender groups might need additional support.



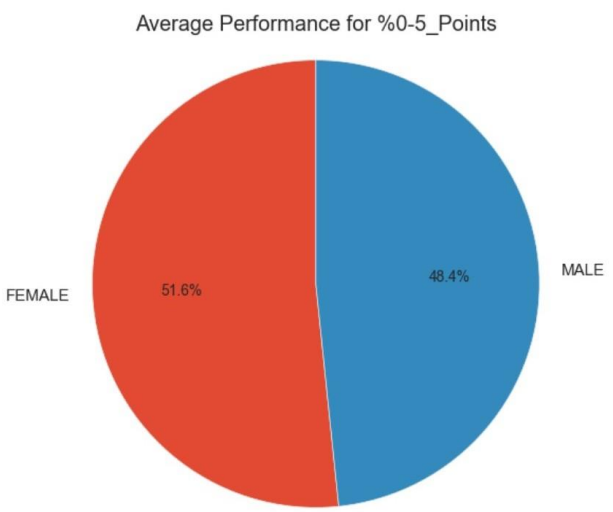
✓ Performance at district level

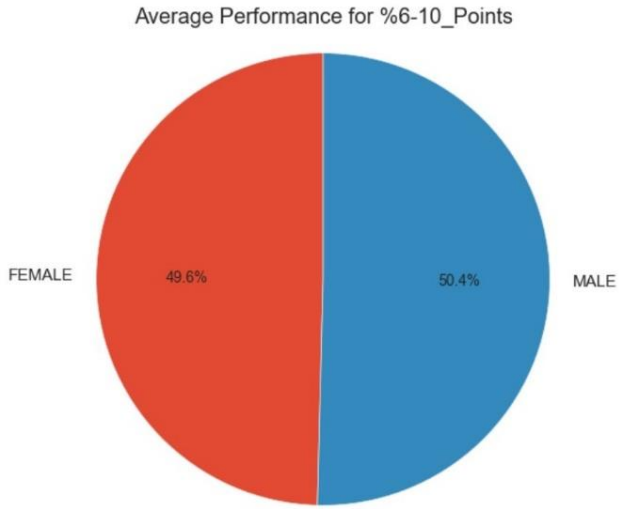
Explore the trends of Evaluating the academic performance of schools within a specific district or region. District-level analysis can provide insights into educational quality across schools in the same area. It helps education authorities make informed decisions about resource allocation and improvement strategies.

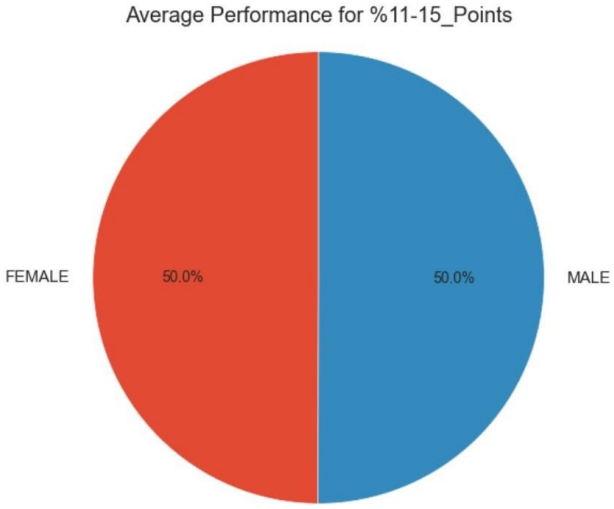


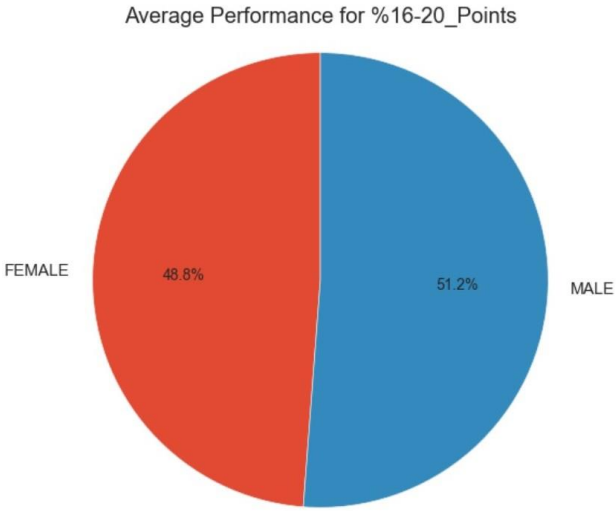
✓ Performance at grade level of points

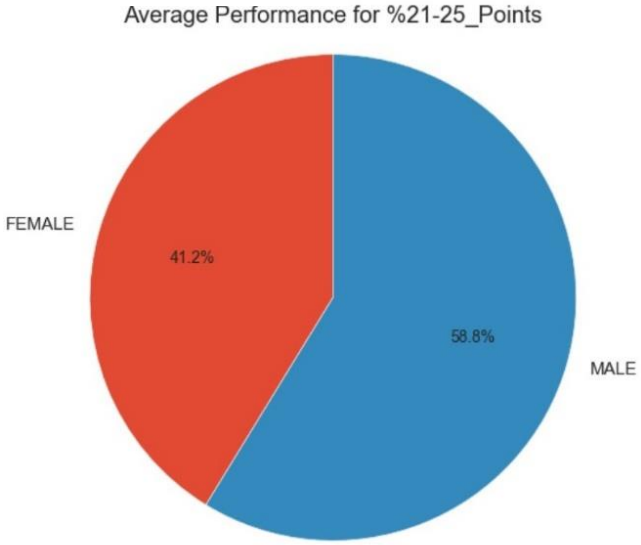
Examining how students perform at different grade levels, based on point ranges. It helps educators understand which grade levels need more attention or targeted interventions.





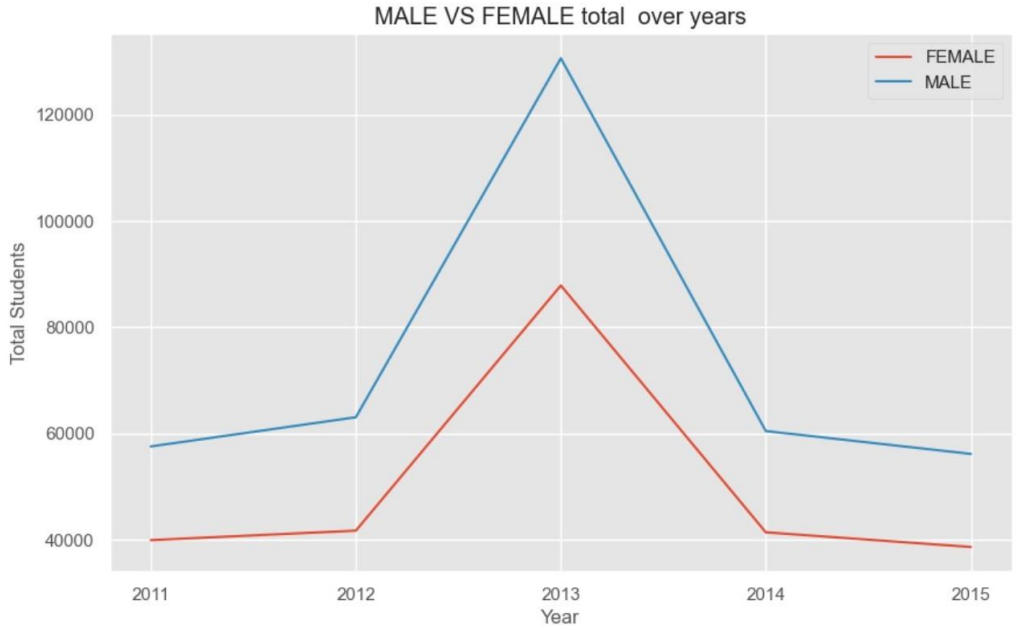






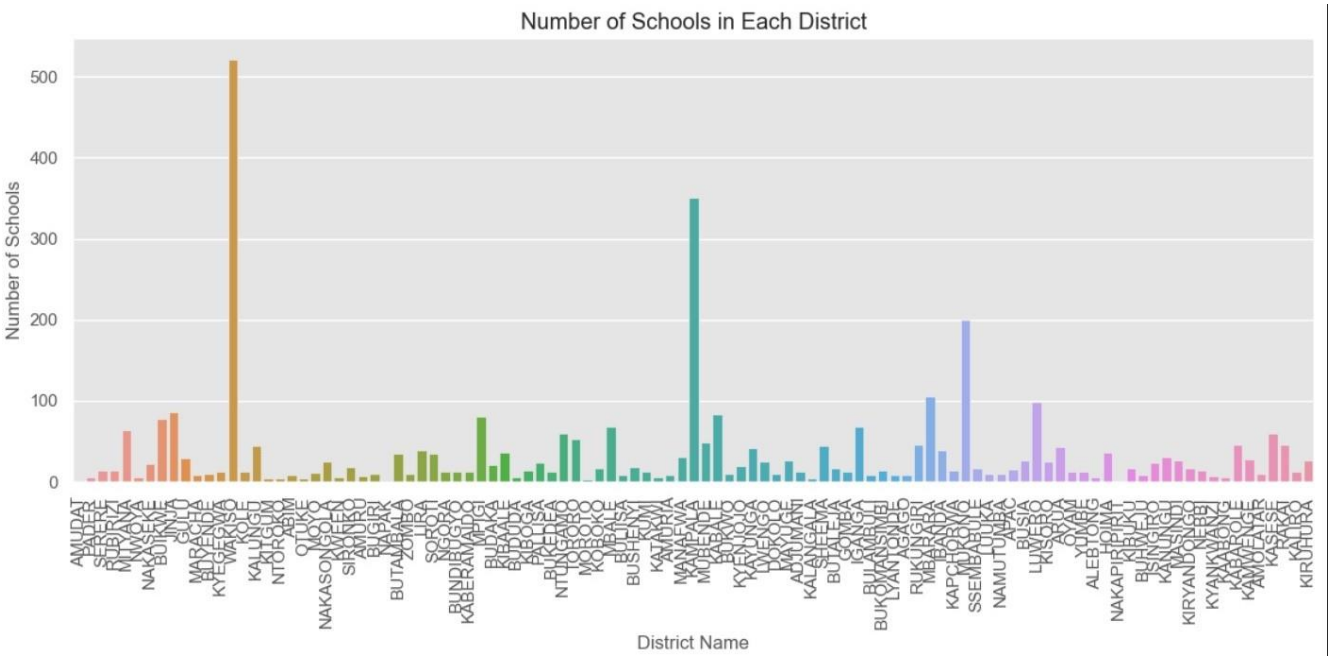
✓ Total students who sat per year

Checking the number of students who took exams in different years. This provides context for understanding performance data and making policy decisions.



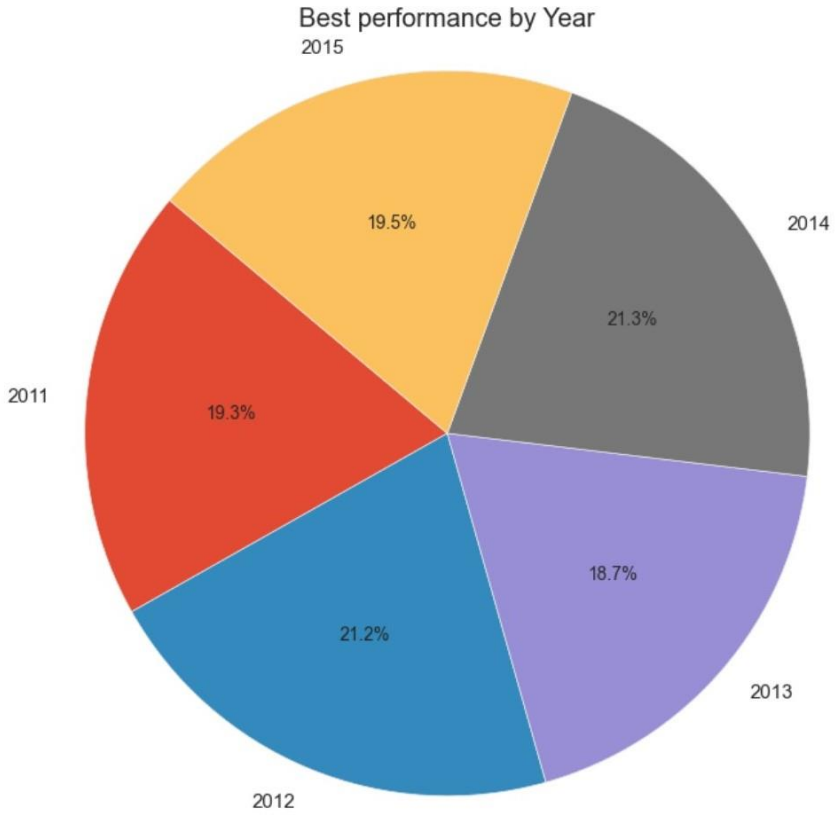
✓ Check For How many Schools are In Are District

Checking the number of schools in each district. This can help officials understand the educational landscape and distribution of schools. It's essential for effective district-level planning and resource allocation.



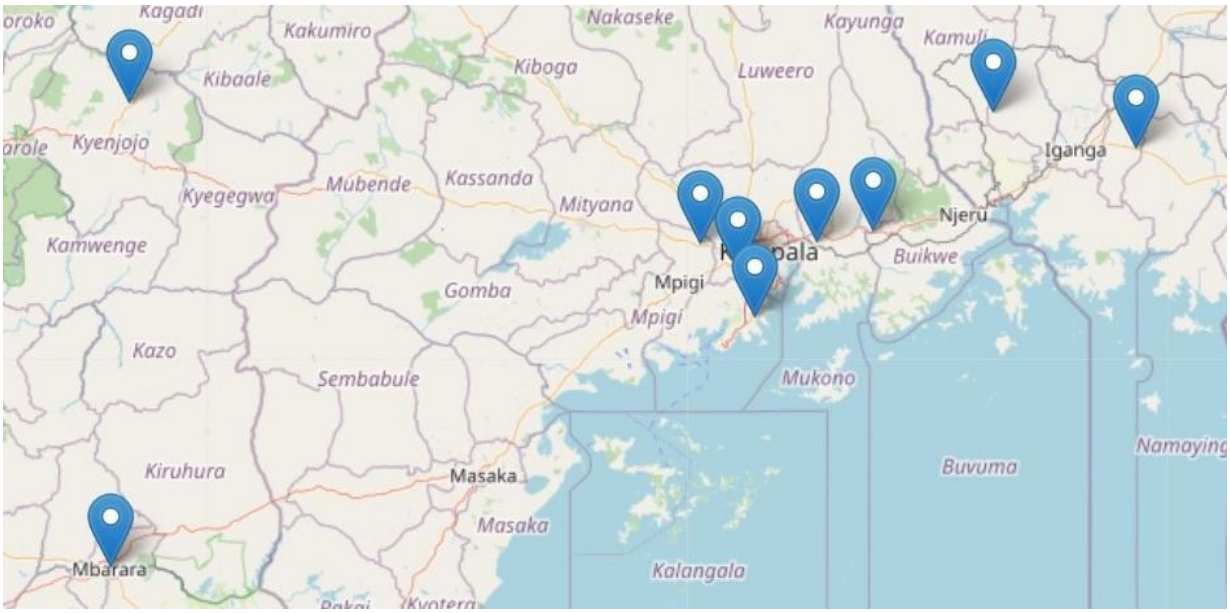
✓ Performance at year level

Evaluating academic performance year by year. This can show whether there are improvements or declines over time. It helps identify long-term trends and the impact of changes in education policies.



✓ Map for presentation for the top performing school

Creating a visual map that displays the locations of top-performing schools. This can provide a geographical context to educational data. It helps stakeholders easily identify high-performing schools in specific areas.



✓ Predictive analysis

We started by:

I. **Dataset Preparation:** It starts by selecting three main features from the df_all DataFrame: "District_Name", "SCHOOL", and "Gender" and will be used to predict the target variables.

II. **Data Encoding:** The categorical features "District_Name", "SCHOOL", and "Gender" are encoded using LabelEncoder. This transforms the string values into numerical values, which are required by most machine learning models.

III. **Target Data Preparation:** The target variable y is constructed from various columns of the df_all DataFrame. These columns represent different grade percentage categories.

IV. **MultiTaskLasso Model:** This model performs feature selection and helps to identify the most influential features for predicting the target variables.

V. **Feature Importance Visualization:** To create a horizontal bar plot showing the feature importance based on the coefficients of the MultiTaskLasso model. This plot helps visualize which features are more influential in predicting the target variables.

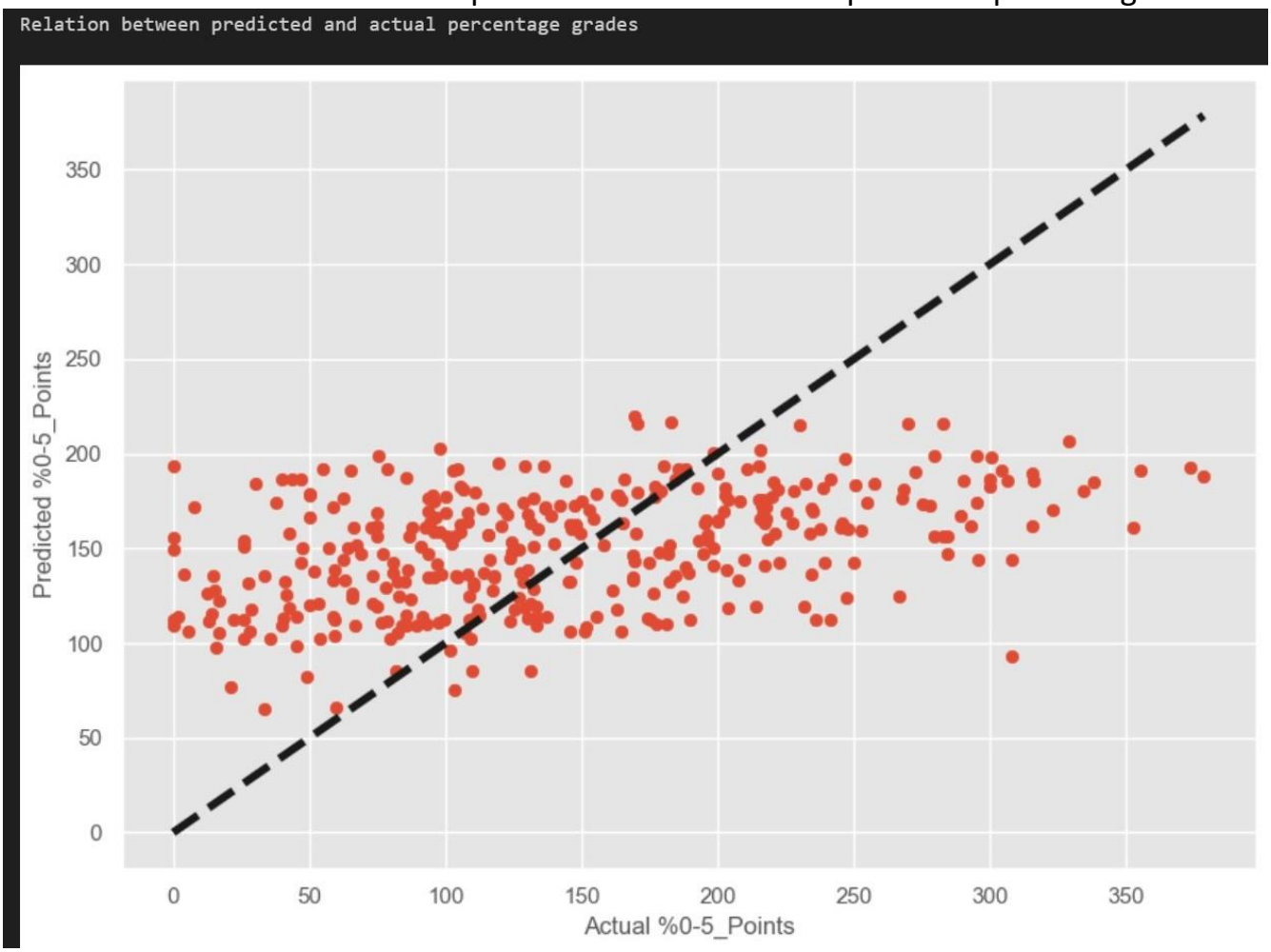
VI. **Feature Removal:** The feature "Gender" is observed to have low importance in the previous visualization and is dropped from the feature dataset X.

VII. **MultiRegression Ridge Model:** The code uses the MultiOutputRegressor from scikit-learn to create a multi-output regression model. The model is trained using the training data (xtrain and ytrain).

VIII. **Model Evaluation:** The training score of the model is calculated and printed to the console. For each target variable, the mean squared error (MSE)

between the predicted and actual values is calculated and printed to the console.

IX. **Visualization of Predictions:** For each target variable, a scatter plot is created to visualize the relationship between the actual and predicted percentages.



CONCLUSIONS

- In conclusion we discovered that the best performing year was 2014 where male students performed better than the female students.
- The year 2013 had the most students with male students above 120,000 students and female students about 85,000 students.
- The district Wakiso had the greatest number of schools i.e., above 500 schools.
- Most male students had average performance in 21-25 points compared to female students.
- The male students performed better than the female students in all years.