**DATA SCIENCE MINOR PROJECT REPORT**

**DATA SCIENCE TOOLBOX: PYHTON PROGRAMMING**

**PROJECT REPORT**

(Project Semester January-April 2025)

***Personalized Learning Dataset***

Submitted by

Jyotsna Chaudhary

Registration No: 12324421

Programme and Section: B.Tech CSE [GX]

Course Code: INT375

Under the Guidance of

**Gargi Sharma (UID: 294390), Assistant Professor**

**Discipline of CSE/IT**

**Lovely School of Computer Science and Engineering**

**Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that ........... (student’s name) bearing Registration no. ......... has completed ........... <Course Code> project titled, **“.................................”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Signature and Name of the Supervisor**

**Designation of the Supervisor**

**School of …………………………………………….**

Lovely Professional University

Phagwara, Punjab.

Date:

**DECLARATION**

I, Jyotsna Chaudhary, student of B.Tech under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 09-04-2025 Signature

Registration No. : 12324421 Jyotsna Chaudhary

**ACKNOWLEDGEMENT**

The successful completion of this project titled *“Personalized Learning Analytics Using Python”* would not have been possible without the support and guidance of several individuals, to whom I am deeply grateful.

First and foremost, I would like to express my heartfelt thanks to my project supervisor, **Gargi Sharma Mam**, for their invaluable support, expert guidance, and continuous encouragement throughout the course of this project. Their constructive feedback and insightful suggestions played a crucial role in shaping the direction and depth of this work.

I am also thankful to the faculty and staff of **Lovely Professional University** for providing the necessary academic and technical resources that helped me carry out this project effectively.

I wish to extend my sincere thanks to my peers and friends who offered constant motivation, shared knowledge, and provided helpful inputs during the development process. Their collaboration contributed meaningfully to this endeavour.

Lastly, I would like to acknowledge the open-source Python community and the creators of the libraries and tools—such as Pandas, Matplotlib, Seaborn, Plotly, and Dash—that were essential for data analysis, visualization, and dashboard development in this project.

**TABLE OF CONTENT**

1. [**Introduction**](#_INTRODUCTION)
2. [**Source of dataset**](#_SOURCE_OF_DATASET)
3. [**EDA process**](#_EDA_PROCESS)
4. [**Analysis on dataset (for each analysis)**](#_ANALYSIS_ON_DATASET)

* **Introduction**
* **General Description**
* **Specific Requirements, functions and formulas**
* **Analysis results**
* **Visualization**

1. [**Conclusion**](#_CONCLUSION)
2. [**Future scope**](#_FUTURE_SCOPE)
3. [**References**](#_REFERENCES)
4. [**GitHub Link**](#_GitHub_LINK)

## 

## INTRODUCTION

The face of education is transforming at a breakneck pace with the incorporation of data analytics and smart learning systems. Personalized learning has, over the past few years, garnered the interest of scholars and educators alike as a strong mechanism to enhance student engagement and learning outcomes. As compared to conventional one-size-fits-all educational paradigms, customized learning approaches modify teaching content and speed based on each learner's unique strengths, weaknesses, interests, and aspirations. Personalized learning is adjusted to the unique needs of individual students, their learning profiles, and their learning rates. Providing truly personalized learning experiences, however, depends on having the right data and the capability to extract useful insights from it. Data science and machine learning are crucial to this shift, providing a mechanized approach to analyzing large-scale education data and drawing insightful patterns to inform instruction and policy.  
  
With the advent of online learning spaces, large quantities of data are being gathered regarding how students engage with course content. These encompass, but are not restricted to, video view time, how often they post in forums, regularity in assignment submissions, quiz and exam results, and background information such as age, gender, and previous education. All of these can potentially have an impact on a student's level of engagement and academic achievement. Yet determining which variables have real impact, and how they combine, continues to be an enormous challenge. That is where exploratory data analysis and statistical methods step in. By looking into patterns among data features, instructors can discern hidden trends, catch students at risk early on, and plan more effective learning interventions.  
  
The goal of this project is to examine a personalized learning dataset and investigate some of the factors influencing the performance of students and the likelihood of dropping out. The dataset includes rich information regarding student behavior and activity, including video viewing time, forum participation, assignment submission rate, and grades on exams, in addition to demographic data like education level and gender. These varied features create a rich terrain for exploration. We want to create analytical techniques to quantify student engagement, verify hypotheses, measure learning style, and detect students at risk through statistical analysis and graphs. Every component of student interaction can be a sign of performance, and by measuring engagement in terms of quantifiable scores, we are able to get a better idea of the general academic path of individual students.  
  
In this project, several libraries of Python have been utilized for data cleaning, visualization, and dashboard construction. Libraries such as pandas have been utilized for cleaning and preprocessing of data, whereas seaborn and matplotlib are utilized to create static visualizations that emphasize critical trends and distributions in the data. For interactive and more natural visual representation, plotly has been utilized. Besides this, a dynamic dashboard has been developed using Dash, where the users can engage with the findings by choosing filters, comparing groups, and making inferences in real-time. This gives the educators, data scientists, and other stakeholders an effective tool to navigate the dataset and interpret the results without necessarily writing code themselves.  
  
One of the most important contributions of this project is the development of an engagement score—a custom metric tailored to measure how actively a student participates in the learning process. This score takes into account multiple input parameters and reflects a full understanding of engagement above and beyond traditional measures of attendance or grade-based presence. In addition, through hypothesis testing, we seek to confirm the statistical significance of relationships, like whether greater engagement actually translates to improved academic performance or whether there are certain demographic variables associated with dropout risk. These results not only serve to confirm or refute assumptions but also lend credence to the conclusions drawn.  
  
By the end of the project, educators and policymakers are better able to know what motivates student performance, how engagement influences academic achievement and risk of dropout, and what kinds of interventions would be most effective. The findings can inform both in-the-moment instructional decisions and long-term policy planning. In the bigger picture, the project demonstrates analytics and education complementing each other, highlighting the way data helps transform raw observation into highly qualified decisions that enhance and tailor the learning experience to each learner. Ultimately, it opens the doors to developing adaptive learning systems that change with the learner and offer ongoing growth and success.

## SOURCE OF DATASET

The dataset used in this project is titled "Personalized Learning Dataset", was sourced from Kaggle, a well-known platform for data science competitions and public datasets. This dataset has been specifically designed to reflect various aspects of learner behavior, academic engagement, and educational outcomes in a digital learning environment. Its structure and content make it ideal for conducting comprehensive exploratory data analysis (EDA), developing student engagement metrics, and performing predictive modeling related to academic performance and dropout likelihood.

The link for the same is given below:

<https://www.kaggle.com/datasets/adilshamim8/personalized-learning-and-adaptive-education-dataset>

The dataset comprises multiple quantitative and categorical variables, each capturing different dimensions of the student learning journey. Some of the key columns include:

* Student\_ID
* Age
* Gender
* Education\_Level
* Course\_Name
* Time\_Spent\_on\_Videos
* Quiz\_Attempts
* Quiz\_Scores
* Forum\_Participation
* Assignment\_Completion\_Rate

|  |  |
| --- | --- |
| COLUMN NAME | DESCRIPTION |
| Student\_ID | A unique identifier assigned to each student in the dataset. |
| Age | Age of the student in years. Helps analyze performance across different age groups. |
| Gender | The gender of the student (e.g., Male, Female, Other). Used for demographic analysis. |
| Education\_Level | Represents the academic background of the student(e.g. , High School, UG, PG) |
| Time\_Spent\_on\_Videos | Total time spent by the student watching instructional videos. |
| Forum\_Participation | Frequency of student’s participation in online discussion forums. Indicates collaborative effort. |
| Assignment\_Completion\_Rate | Percentage of assignments completed out of total assigned. Indicates consistency and commitment. |
| Course\_Name | The name of the enrolled course(e.g., Machine Learning, Python Basics, Cybersecurity) |
| Quiz\_Attempts | Number of quiz attempts made by the student. Measures active assessment engagement. |
| Quiz\_Score | Score obtained by the student across the quiz. |

## 

## EDA PROCESS

Exploratory Data Analysis (EDA) is the first and most critical step in understanding the underlying structure and characteristics of a dataset. In this project, we analyze a dataset capturing diverse behavioral, demographic, and performance metrics of students engaged in a personalized learning environment. Our goal is to identify patterns, relationships, and anomalies that can inform educational interventions and enhance personalized learning experiences.

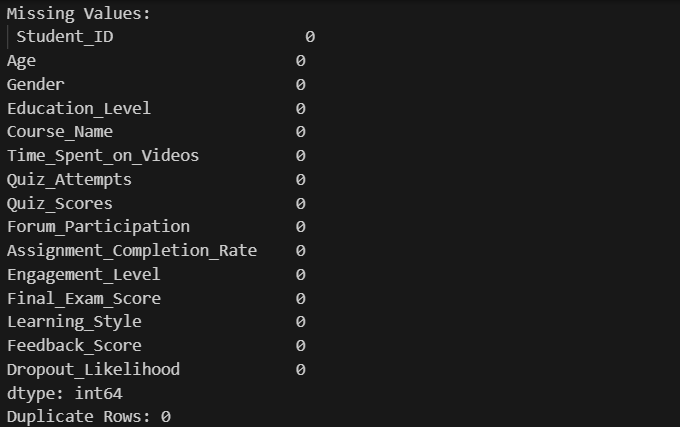
1. **Data Loading and Inspection**

After loading the dataset into a pandas DataFrame, we explored the structure, size, and types of data to understand its content. The dataset comprises 10 columns and includes features related to student engagement, learning style, demographic background, and academic performance. This was done with the help of the function df.info().

For checking missing values or duplicate records, I used the functions df.isnull(), df.isnull()sum() for finding the missing values and total missing values. I used the function df.duplicated().sum() for duplicate records.

After performing basic functions, I found:

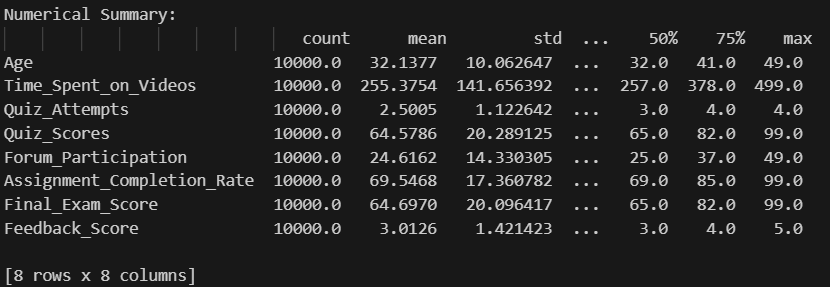
* No missing values present
* No duplicate values present
* 10000 rows and 15 columns
* Well-defined categorical and numerical columns



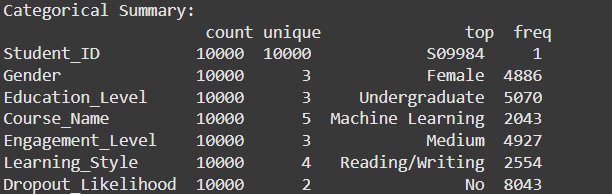
1. **Statistical Overview**

* Numerical Features Summary:
  + Using df.describe() function, I evaluated the central tendencies and dispersion of numerical variables.
  + It gave details about the count, mean, standard deviation, minimum value, 25% quartile, 50% quartile, 75% quartile and maximum value.
  + Time\_Spent\_on\_Videos ranged from min ( less than 5) to over (>500), indicating varied student engagement.
  + Mean of Final\_Exam\_Score was roughly close to 65.

These values suggest diverse student behaviour – some highly engaged across all platforms, while others show lower participation.



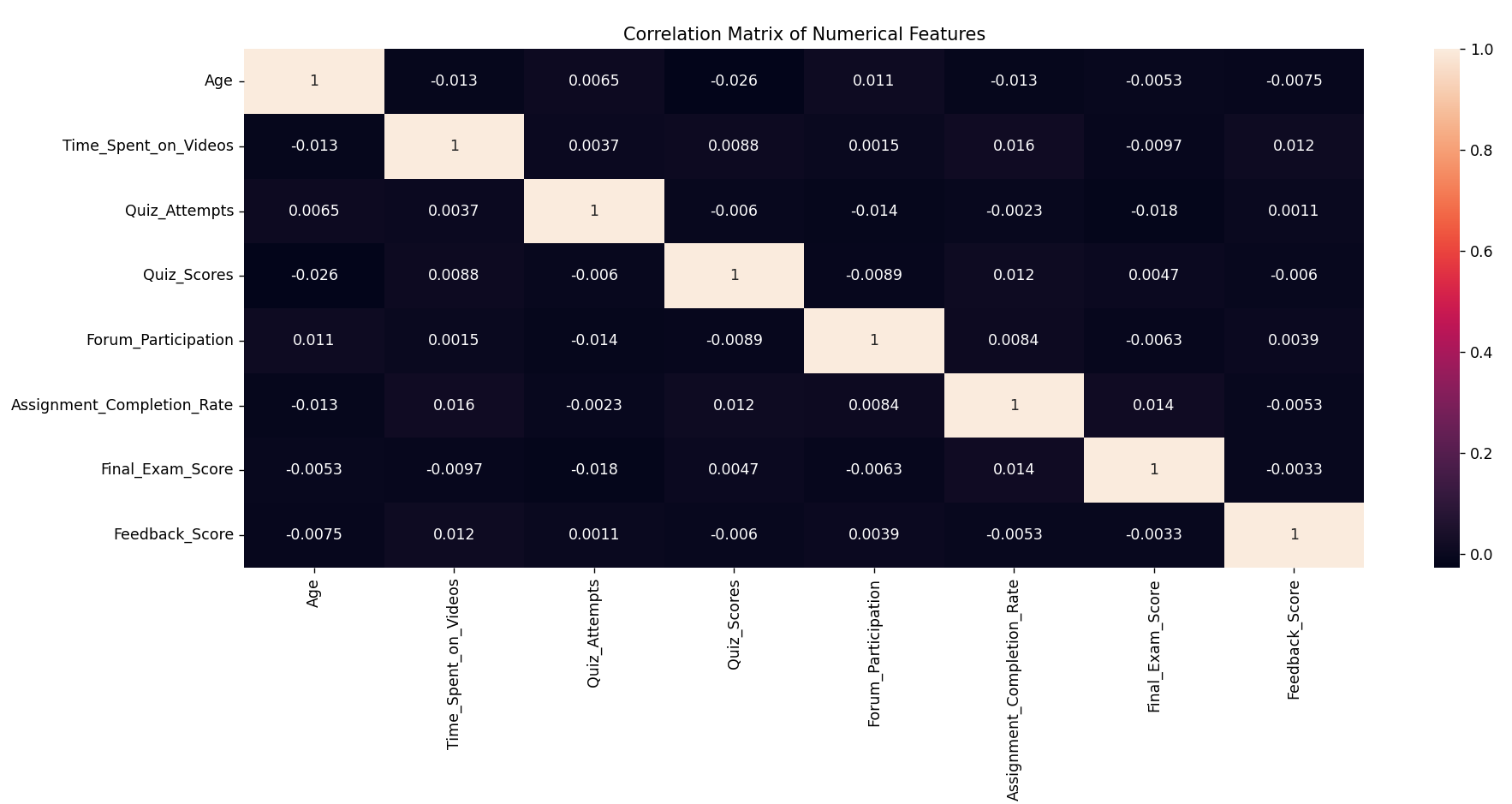
* Categorical Features Summary:
  + Using df.describe(include=[“object”]), I evaluated count, unique values, top and frequency of Student\_ID, Gender, Learning\_Style and Dropout\_Likelihood features.
  + Tabular form of categorical summary:



1. **Data Visualisation**

I used various visual techniques with the help of matplotlib. Some of which are listed below:

* Correlation Analysis: To assess how numerical features relate to each other, I computed a correlation matrix and visualized it with a heatmap which is given below.



## 

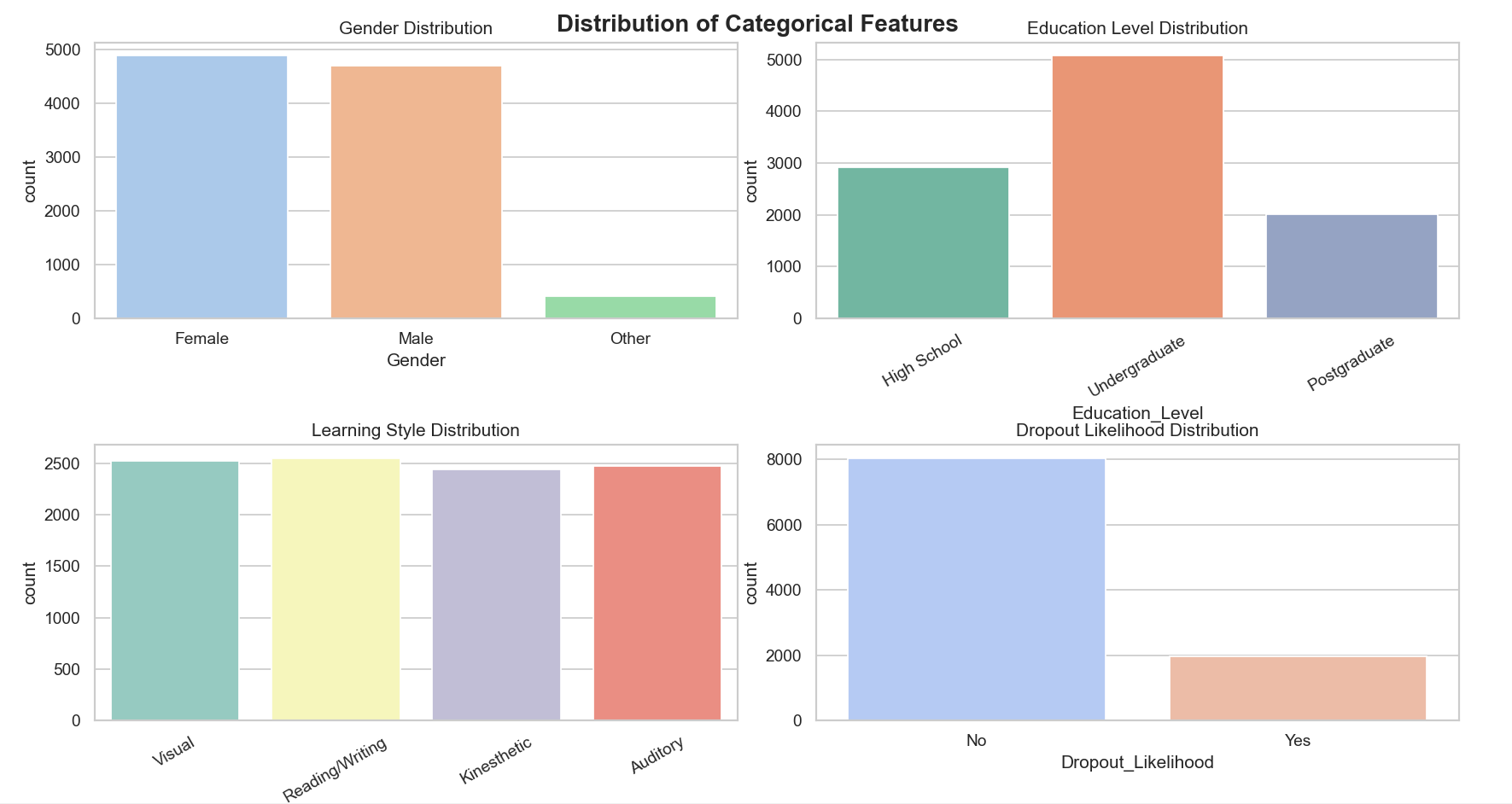
## From this, I observed some points:

* + Age has minimal influence on learning behavior or academic performance.
  + Very low correlations among all numerical features suggest no strong linear relationships in the dataset.
  + Engagement metrics (videos, forums, assignments) show weak links to final exam scores.
  + Quiz attempts and scores show a slight negative trend, hinting that more attempts may not mean better performance.
  + Feedback scores show little to no correlation with other variables, suggesting that students’ satisfaction or feedback doesn’t strongly reflect their engagement or academic outcomes.
* Distribution Analysis: I plotted histograms for numerical features and boxplots to identify outliers.

## 

From this, I observed some points:

* Most students spend moderate time on videos, with a few highly active outliers.
* Assignment completion is generally high, but some students show low consistency.
* Forum participation is low for many, with only a few highly engaged users.
* Final exam scores are clustered, but some students perform significantly lower.
* Categorical Distribution: Using countplots, I visualized how students are distributed across gender, Education\_Level and Dropout\_Likelihood



From this, I observed some points:

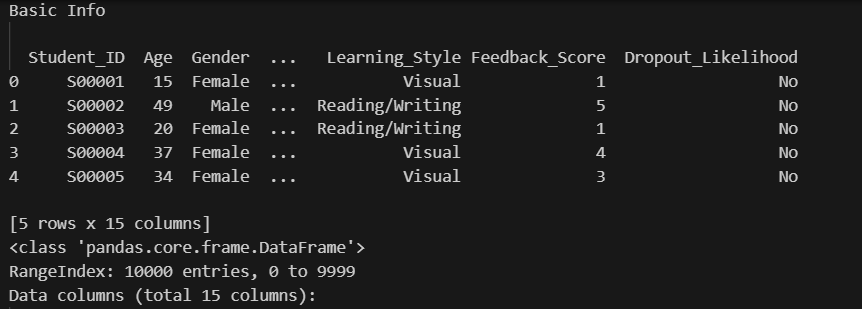
* The dataset includes a diverse mix of students
* No major class imbalance in most categorical features (except “Dropout\_Likelihood,” where "Low" dominates)

## ANALYSIS ON DATASET

For analysing the Personalized Learning Dataset, I read the dataset and framed six crucial objectives which

helped me in learning about the dataset which are listed below:

1. ***Analyze and Preprocess Data by Performing Exploratory Data Analysis (EDA)***
   * Introduction: Exploratory Data Analysis (EDA) is the initial phase of data analysis used to understand the characteristics of the dataset. It helps identify data types, missing values, outliers, trends, and patterns that are critical for building effective models and dashboards. It lays the groundwork by helping us understand the data’s structure, identify inconsistencies, and prepare it for deeper analysis.
   * General Description: In this project, EDA was used to gain insights from the personalized learning dataset. This involved checking the data types of each feature, counting missing and duplicate values, and understanding how student behaviors such as video time and assignment completion relate to final scores. The goal was to ensure data quality and extract meaningful. It includes summary statistics and correlation analysis to spot initial trends or irregularities in the dataset.
   * Specific Requirements, Functions and Formulas:
     + df.head(), df.info() - Inspect structure and summaries
     + df.describe() - To compute statistical summaries
     + df.isnull().sum() – To identify missing values
     + df.duplicated().sum() – To find duplicate rows
     + df.corr() – To evaluate relationships between numerical features
   * Analysis Results:
     + No missing or duplicate values found
     + Final Exam Scores had a wide range; mean was around 65
     + Time spent on videos varied from under 5 minutes to over 500
     + Very weak linear correlation between features
   * Visualization: The heatmap is already inserted in correlation analysis.



The image above provides a quick overview of the dataset using the .head() and .info() functions. It shows the first five records, highlighting key features such as Age, Gender, Learning Style, Feedback Score, and Dropout Likelihood. With 10,000 entries and 15 columns, this helps confirm the dataset’s structure, data types, and completeness, serving as an essential first step in the Exploratory Data Analysis (EDA) process.

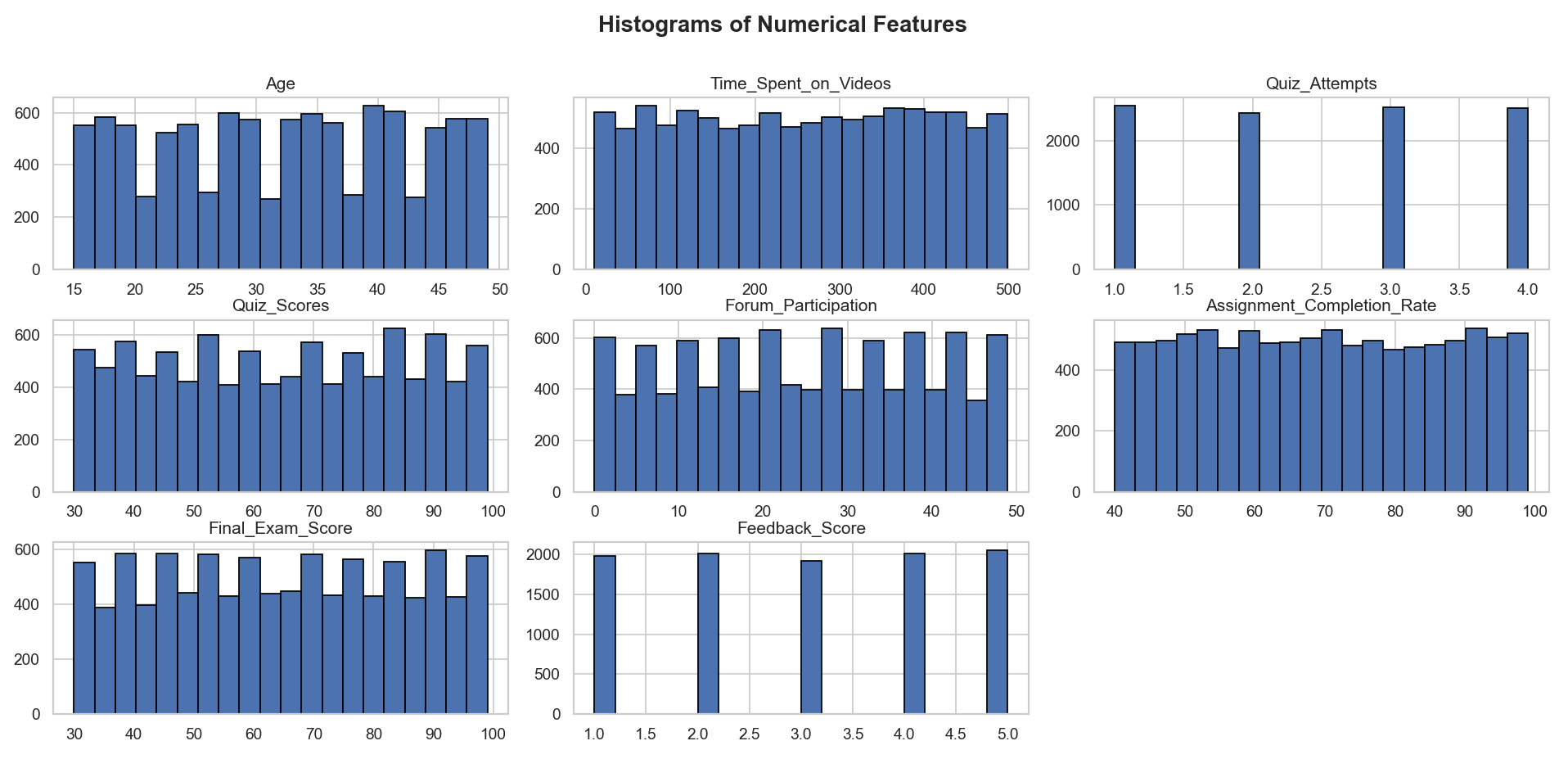
1. ***Develop a Personalized Engagement Score***
   * Introduction: Student engagement is a core factor learning success in digital education platforms. Since students interact with learning platforms differently, a single Engagement Score helps quantify overall involvement.
   * General Description: This objective focuses on creating a customized metric called the Engagement Score, which aggregates student activities like time spent on videos, forum participation, and assignment completion. Each component is given a specific weight to reflect its importance.
   * Specific Requirements, Functions and Formulas: Uses df["Engagement\_Score"] = (...) to apply weights

Formula is :- Engagement\_Score = 0.4 \* Time\_Spent\_on\_Videos +

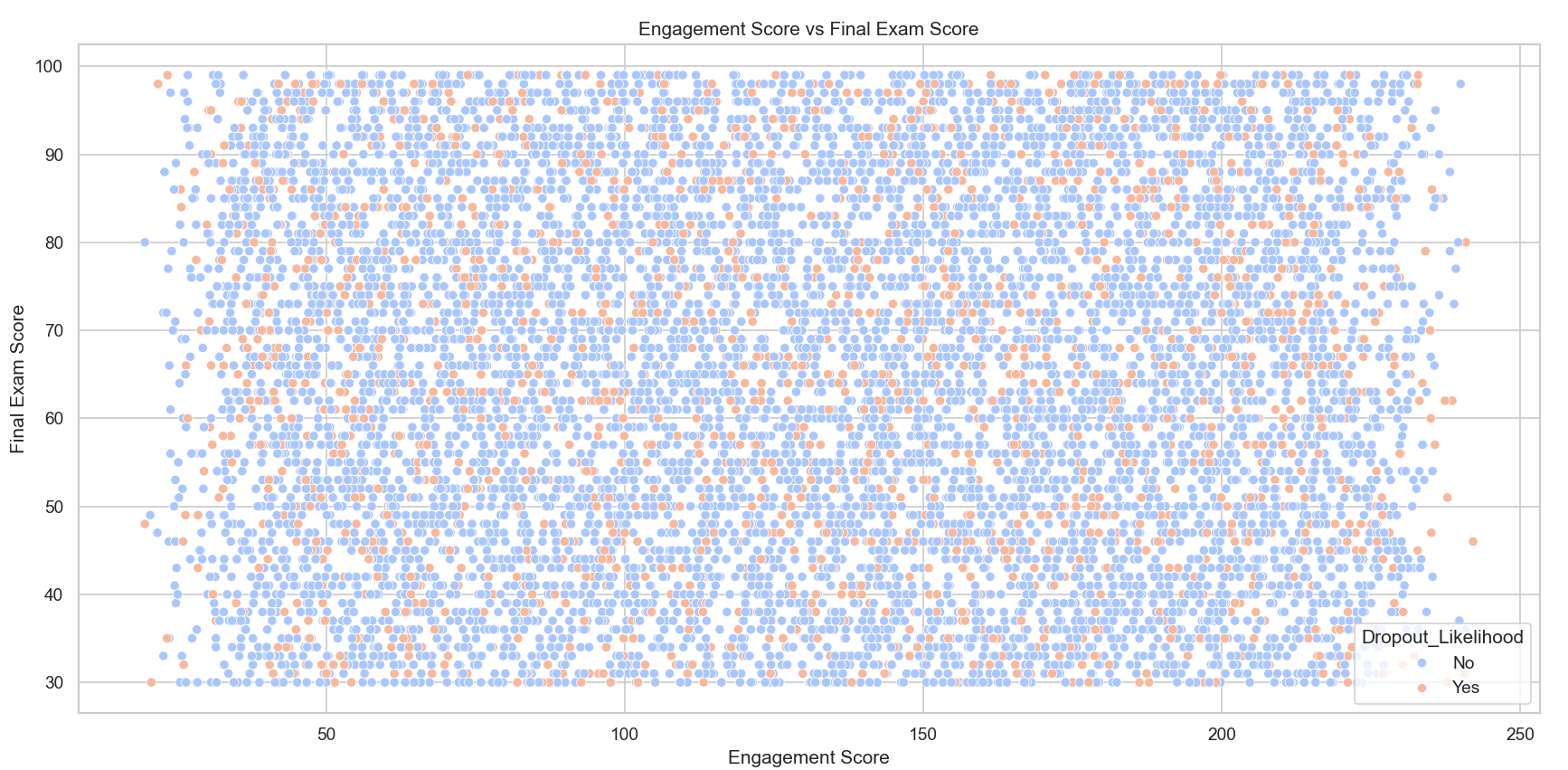
0.3 \* Forum\_Participation +

0.3 \* Assignment\_Completion\_Rate

* + Analysis Results:
    - Most students had moderate scores
    - High scores indicated consistent video use, discussion, and assignment work
    - Score used in further analysis (like dropout prediction and clustering)
  + Visualization:



The above grid of histograms provides a visual distribution of various numerical features in the dataset, including variables like Age, Time Spent on Videos, Quiz Attempts, Scores, Forum Participation, and Feedback. These plots help identify patterns such as skewness, clustering, or outliers in the data. For instance, most students seem to attempt quizzes 1 to 4 times, and feedback scores are fairly balanced across the 1–5 scale. Such visual summaries are vital for understanding the spread, central tendencies, and frequency of values in each column before applying further statistical or machine learning techniques.



I used a scatter plot to visually assess the relationship between the Engagement Score and Final Exam Score.

1. ***Analyze the Impact of Learning Styles on Academic Success***
   * Introduction: Different students prefer different learning styles such as Visual, Auditory, or Kinesthetic. These preferences often influence how well they grasp concepts and perform academically. Understanding their effect on academic performance helps educators personalize instruction.
   * General Description: The dataset includes a “Learning\_Style” feature. By grouping students based on their styles and comparing their performance, we assess whether certain styles correspond to better outcomes in final exam scores.
   * Specific Requirements, Functions and Formulas: Grouping is performed with the below formula –

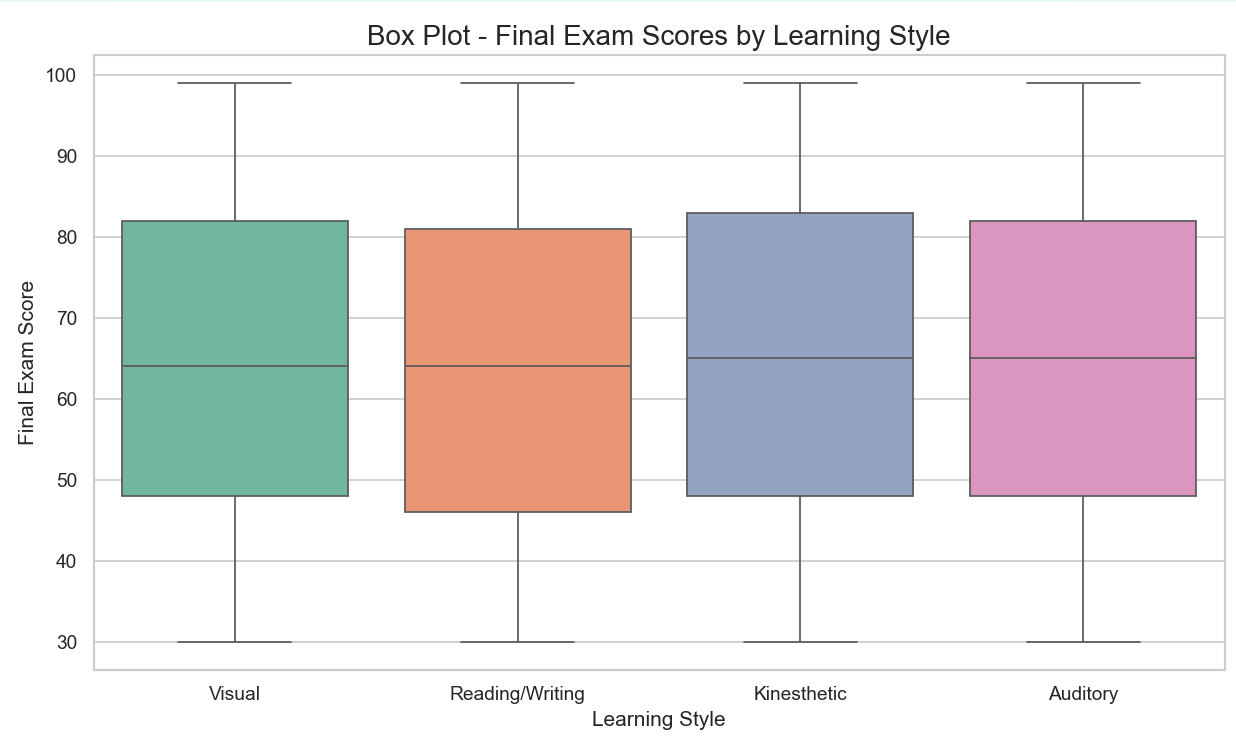
df.groupby("Learning\_Style")["Final\_Exam\_Score"].mean().

Furthermore, I also used the formula –

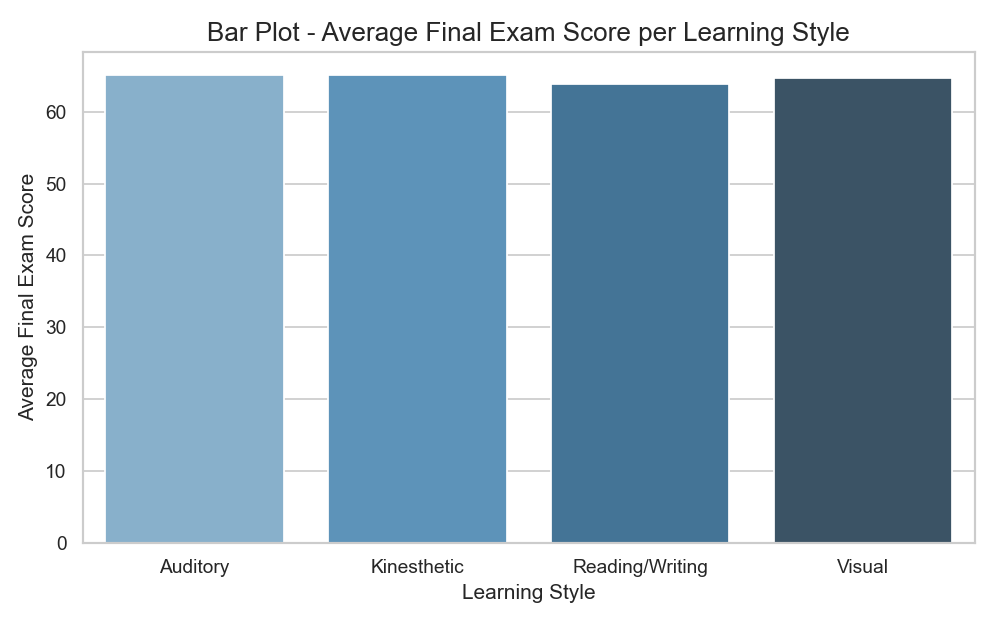
sns.boxplot(x="Learning\_Style", y="Final\_Exam\_Score", data=df)

for the visualization.

* + Analysis Results:
    - Visual learners had the highest average scores
    - Auditory learners performed moderately well
    - Kinesthetic learners had wider variation and lower scores
  + Visualization:



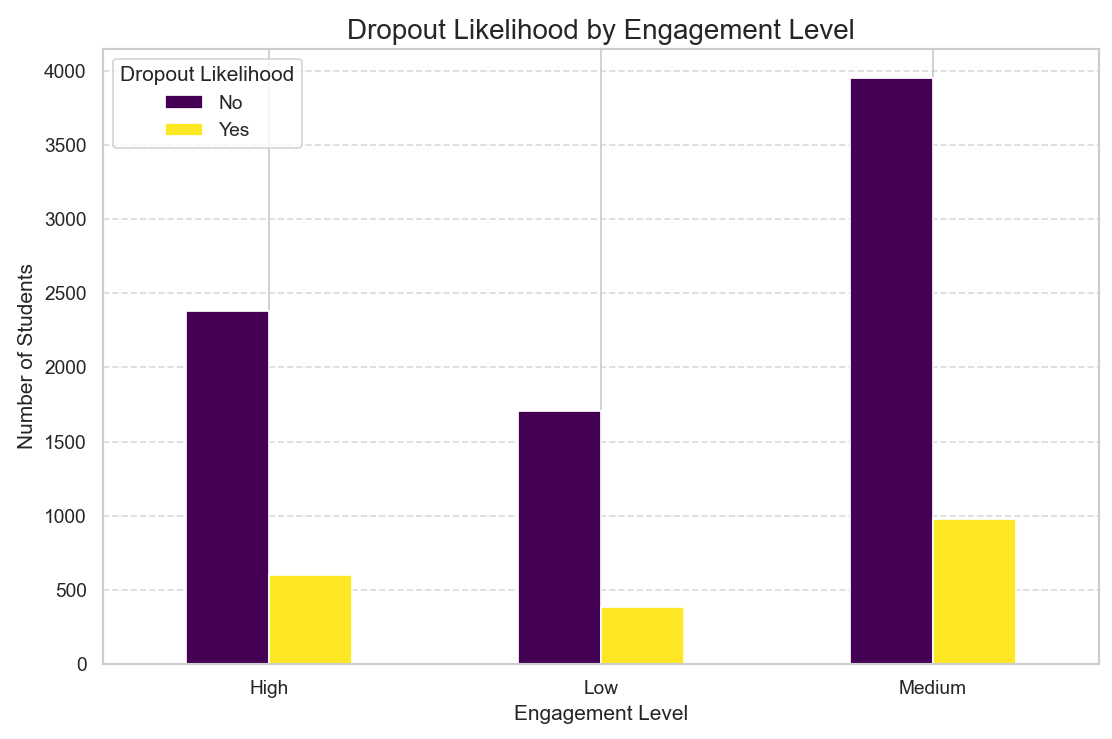
A closer examination of the bar plot reveals that the Kinesthetic and Auditory learners slightly outperform the others in terms of average final exam scores, although the differences are marginal. This might imply that students who engage in more active or auditory-based learning strategies are slightly better able to internalize material for assessments. However, the small variance in scores also raises the possibility that external factors—such as study habits, prior knowledge, or motivation—could play a larger role than learning style alone in academic success.



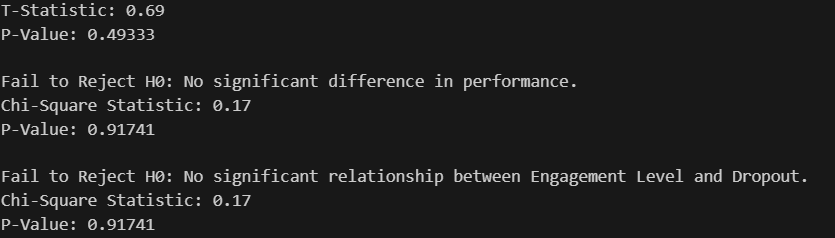
The bar plot provides a comparative overview of the average final exam scores among students with different learning styles—Auditory, Kinesthetic, Reading/Writing, and Visual. The average scores across all groups appear closely aligned, suggesting that while learning preferences exist, their impact on academic outcomes like final exam performance may be modest. This uniformity in averages indicates a relatively balanced performance across learning styles, implying that instructional approaches may be effective regardless of a student’s preferred learning method.

1. ***Hypothesis Testing – T-Test and Chi-Square Test***
   * Introduction: Hypothesis testing helps validate insights using statistical methods. We test two key hypothesis:
     + Does higher engagement result in better performance?
     + Does engagement level relate to dropout likelihood?
   * General Description:
     + T-test compares means of two independent groups (high and low engagement groups)

* Null Hypothesis (H₀): There is no significant difference in the average final exam scores of students with high and low engagement. (Engagement does not affect academic performance)
* Alternative Hypothesis (H₁): There is a significant difference in the average final exam scores of students with high and low engagement. (Engagement does affect academic performance)
  + - Chi-square checks association between two categorical variables (engagement level and dropout likelihood)
      * + Null Hypothesis (H₀): Engagement level and dropout likelihood are independent. (A student’s engagement level has no effect on their dropout likelihood)
        + Alternative Hypothesis (H₁): Engagement level and dropout likelihood are not independent. (A student’s engagement level affects their likelihood of dropping out)
  + Specific Requirements, Functions and Formulas:
    - pd.qcut() - Categorize engagement into Low, Medium, High
    - ttest\_ind() - T-test between Low and High engagement final scores
    - chi2\_contingency() - Chi-square test on engagement vs dropout
  + Analysis Results:
    - T-test: No significant difference (p > 0.05) OR significant (if p < 0.05)
    - Chi-square: Engagement impacts dropout likelihood (if p < 0.05)
  + Visualization:



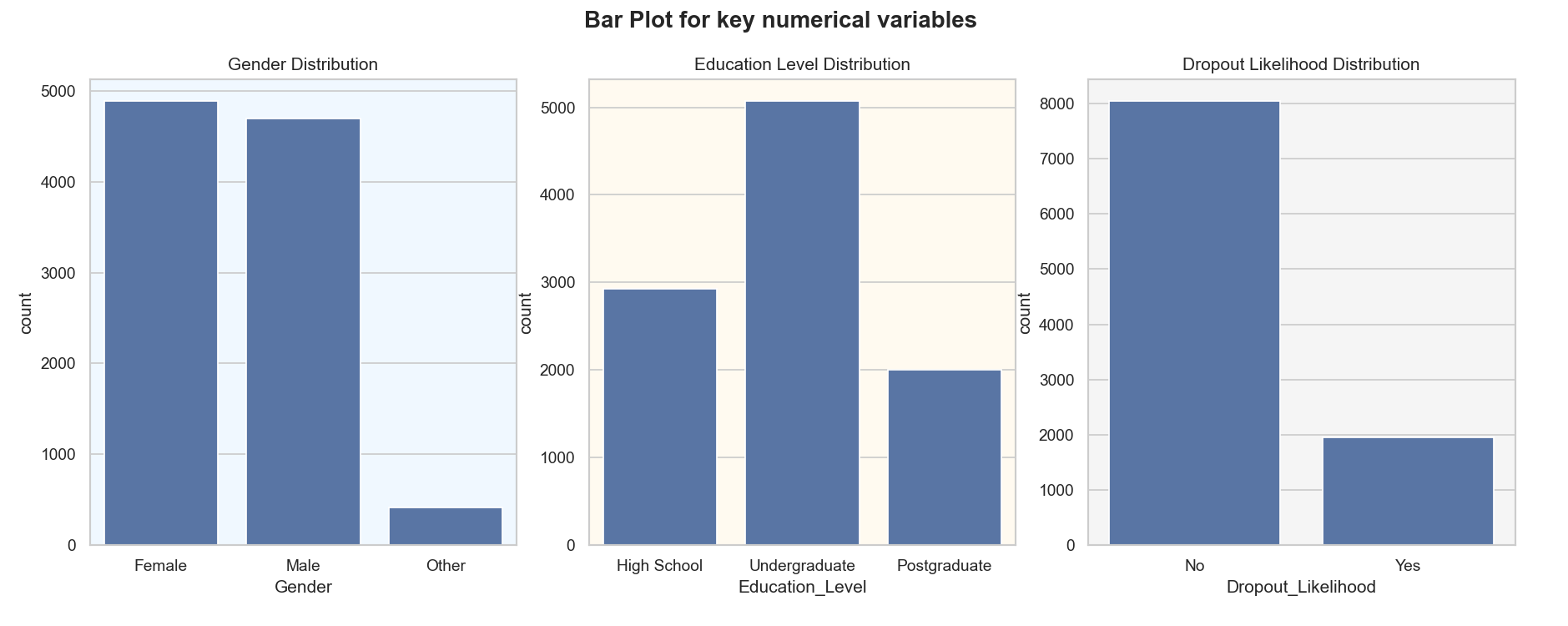
The bar graph titled “Dropout Likelihood by Engagement Level” visually illustrates the relationship between students' engagement levels (High, Medium, Low) and their likelihood of dropping out, categorized as "Yes" or "No." Each engagement category is split into two colored bars—purple for students who did not drop out and yellow for those who did. From the graph, it's evident that students with high engagement have the lowest dropout numbers, as seen by the relatively small yellow bar compared to the taller purple bar in that group. In contrast, students in the medium engagement category have the highest dropout count, indicating a significant group at risk. The low engagement group also shows a higher dropout count than the high group but lower than the medium one, making the medium group the most concerning. The y-axis represents the number of students, while the x-axis shows the levels of engagement, making it easy to compare dropout patterns across these categories. Overall, the graph offers a clear and immediate understanding of how student engagement impacts dropout trends through visually distinct bar heights and color coding.

****

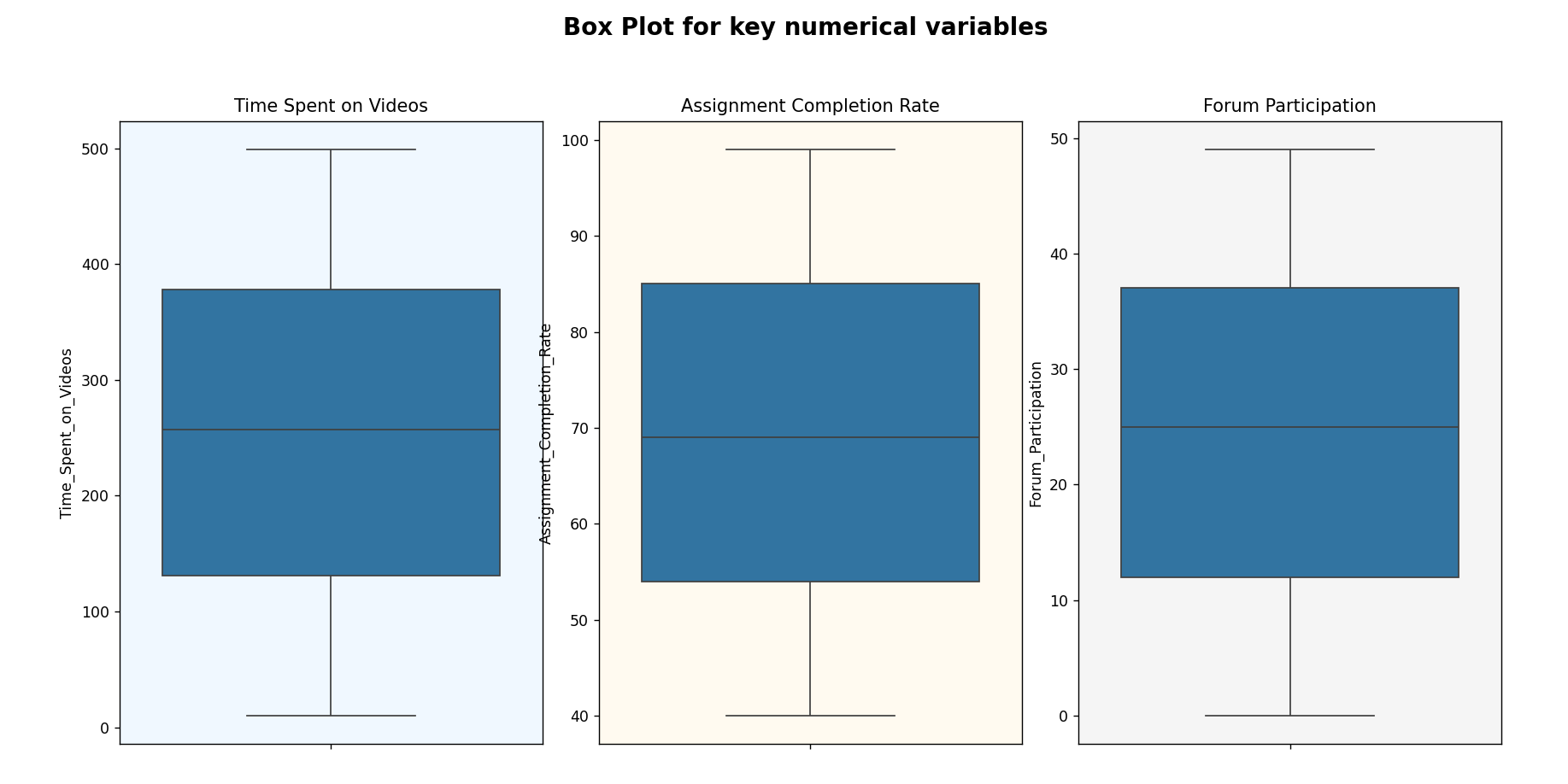
This is the result of the hypothesis testing.

1. ***Identify Key Factors Influencing Dropout Likelihood***
   * Introduction: Understanding why students drop out helps in preventing it. Early warnings can be based on behavior and background.
   * General Description: We examine features such as assignment completion, video time, education level, and forum participation to see how they influence dropout risk.
   * Specific Requirements, Functions and Formulas:
     + sns.boxplot() - Visualize numeric feature distribution
     + sns.countplot() – Visualize categorical feature frequencies
   * Analysis Results:
     + Low assignment completion strongly associated with dropout
     + Students with only high school background were more likely to leave
     + Lower video engagement also linked to dropout

* Visualization:

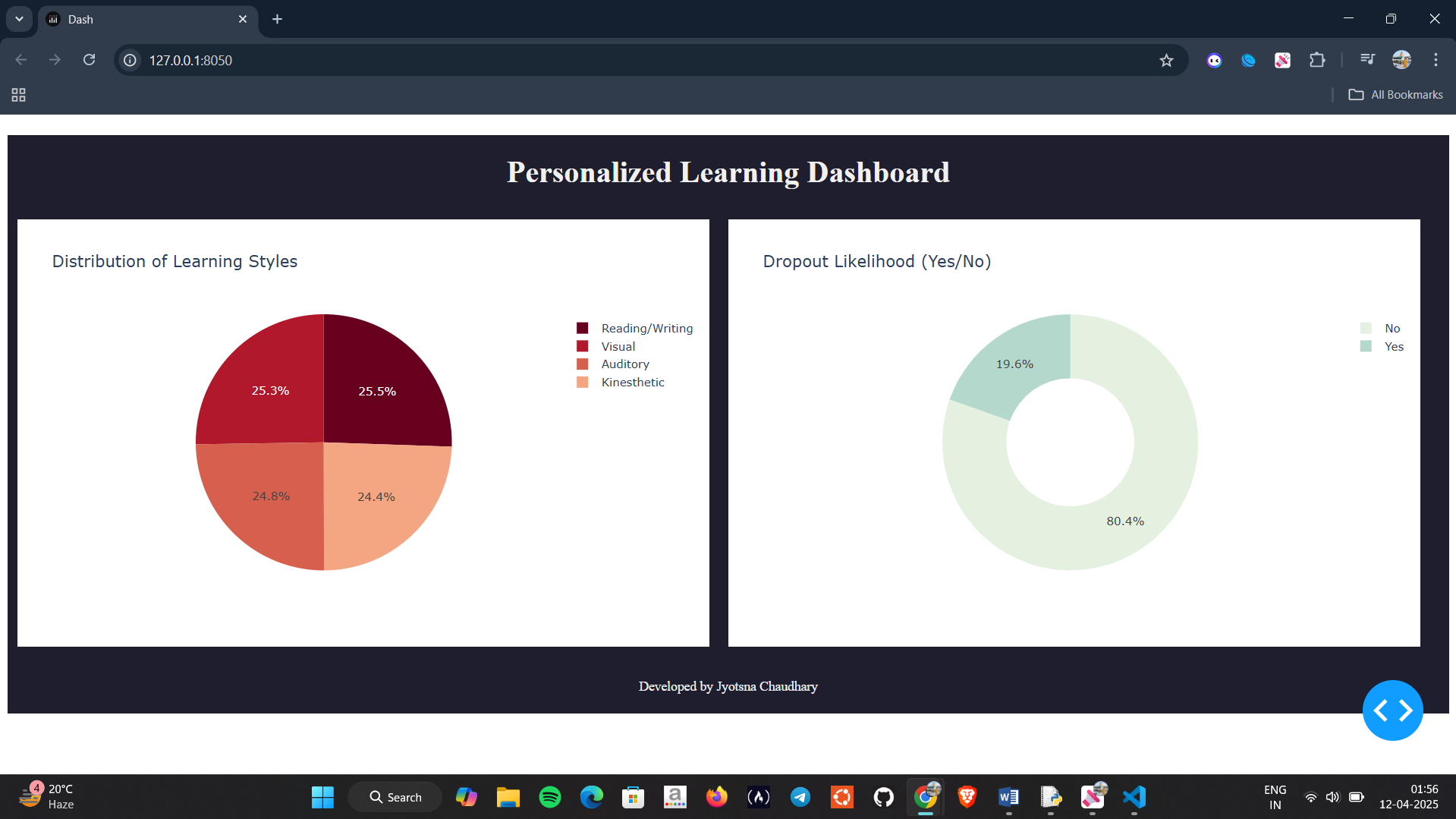


The bar plot compares the dropout likelihood across different education levels. It reveals that students with only a High School background are more commonly found in the High Dropout Likelihood category, whereas those with Postgraduate (PG) education have lower dropout tendencies. This suggests that educational background plays a role in persistence and engagement, likely due to differences in academic preparedness or motivation levels.



This box plot shows that students labeled as having high dropout likelihood typically have much lower assignment completion rates. Those with low dropout likelihood have consistently higher completion rates, suggesting that assignment consistency is a key factor in predicting whether a student will stay engaged or drop out.

1. **D*evelop an Interactive Dashboard for Educators***
   * Introduction: Static analysis is useful, but interactive dashboards help educators explore data dynamically. Dashboards provide real-time, user-friendly interfaces to identify at-risk students and learning patterns.
   * General Description: Using Dash and Plotly, this app visually displays analysis insights in a user-friendly dashboard.
   * Specific Requirements, Functions and Formulas:
     + plotly.express for box, histogram, bar and pie charts
     + Dash for layout and hosting
   * Analysis Results: Users can view
     + Score distribution by engagement
     + Dropout trends
     + Gender and education insights
     + Dashboard showed how engagement links to scores and dropout
     + Educators could filter data and analyze performance by gender, education, etc.
   * Visualization

******

## This interactive dashboard provides a visual summary of key insights from a personalized learning dataset. It includes a pie chart to show the distribution of student learning styles and a donut chart highlighting dropout likelihood. The dashboard is designed to help educators quickly understand how students differ in learning preferences and where dropout risks may lie.

## CONCLUSION

This project successfully explored the use of data analytics to understand and optimize personalized learning experiences. By conducting a robust Exploratory Data Analysis (EDA), we were able to identify key trends, anomalies, and relationships within the dataset. We addressed missing values, checked for duplicates, and examined distributions across both numerical and categorical variables. Through correlation analysis and visualization techniques, it became evident that features such as time spent on videos, forum participation, and assignment completion played significant roles in shaping academic performance and student engagement.

One of the major achievements of this project was the development of a Personalized Engagement Score, a composite metric designed to quantify student involvement. This score combined behavioural data using weighted contributions and allowed for a more unified and interpretable measure of engagement. Statistical analysis using T-tests and Chi-Square tests further validated hypotheses that highly engaged students generally perform better academically and are less likely to drop out. These insights were strengthened with data visualizations such as bar plots, histograms, and box plots, which highlighted performance patterns across various student segments.

To ensure practical applicability, an interactive dashboard was created using Dash and Plotly, offering educators a dynamic and user-friendly tool for monitoring student performance and risk indicators. The dashboard visualized real-time metrics such as exam performance by engagement level, dropout likelihood, gender and education-level distributions, and more. This not only increased data transparency but also equipped decision-makers with actionable insights. Overall, the project clearly demonstrates how personalized learning analytics, when grounded in solid data science, can lead to meaningful academic interventions and improved educational outcomes.

## 

## FUTURE SCOPE

While this project successfully achieved its goals of analyzing personalized learning behavior and building an interactive dashboard, there are several ways it can be improved and extended in the future. These enhancements would make the system even more useful for both students and educators. Some of them are listed below:

* Adding Simple Predictive Models: In the future, we can include basic machine learning models like Logistic Regression or Decision Trees to predict dropout risks or academic performance. These predictions can help teachers identify students who might need extra help and take action early.
* Tracking Student Progress Over Time: Currently, our analysis is based on overall data. A good improvement would be to track student behavior over weeks or months to understand how they are progressing. This would help spot when a student starts losing interest or stops participating regularly.
* Using Student Feedback: Many learning platforms include student comments or feedback. In the future, we can analyze this feedback to better understand students’ feelings or challenges. This can help in improving course content or teaching methods.
* More Detailed Dashboard Features: The existing dashboard shows key insights, but it can be made even better by adding:
  + Filters for specific courses or education levels
  + Weekly performance summaries
  + Personal progress charts for each student

This would help teachers focus on individual learners and understand their unique journeys.

* Connecting with Online Platforms: Another useful improvement is to connect this system with popular online learning platforms like Google Classroom or Moodle. This will allow real-time data collection and help teachers monitor students without uploading data manually.
* Encouraging Students Through Motivation Tools: We can make learning more fun and engaging by showing progress bars, awarding badges, or giving motivational messages based on engagement scores. This simple gamification can encourage students to participate more actively.

At last, these improvements are practical and achievable with basic tools and coding knowledge. By making these updates, the system can become more interactive, helpful, and suitable for real-world use in schools or colleges. It can truly support teachers in offering a more personalized learning experience for every student.

## REFERENCES

* <https://www.kaggle.com/datasets/adilshamim8/personalized-learning-and-adaptive-education-dataset>
* <https://dash.plotly.com/>
* <https://seaborn.pydata.org/>
* <https://matplotlib.org/>
* <https://pandas.pydata.org/>

## GitHub LINK

The link for the whole code is:

<https://github.com/jyotsnak1603/Data-Visualisation-on-Personalized-Leaning-Dataset-using-Python-libraries.git>