WEEK 2: EXPLORATORY DATA ANALYSIS (EDA)

TEAM NAME: RIT 1410AI Team 5A

DATE: 28-10-2024

Team Member Name	Email ID
Abdullah Imran	abdullahimranarshad@gmail.com
Matthew Ojo	ojoaisosamatthew@gmail.com
Krishin Tharani	Krishintharani1+internships@gmail.com
Emani Likhita	likhi.m9363@gmail.com
Sangeeta Sahoo	sahoo1107.sangeeta02@gmail.com
John syllah	johnsyllah2003@gmail.com
Tracy Reson	tracyreson@gmail.com
Afra Falakh	afrafalakh16@gmail.com
Eluit Cruz	ej.cruz.sant@gmail.com

CONTENTS

1. Introduction:

In the second week of this internship, we embarked on a journey to master Exploratory Data Analysis (EDA), a pivotal stage in the data science process that sets the foundation for advanced AI techniques. This week's focus was on generating a comprehensive EDA report, an essential tool for uncovering deeper data insights in the forthcoming stages of our analysis.

We began with an in-depth exploration of datasets, where we aimed to comprehend their structure, identify key variables, and detect any underlying patterns or anomalies. By utilizing various visualization techniques, we could effectively represent data insights, which not only facilitated a clearer understanding but also helped form initial hypotheses about the data. Throughout this process, we employed tools such as histograms, box plots, scatter plots, and heat maps to visualize distributions and relationships within the data.

The insights and hypotheses derived from these visualizations were meticulously documented and compiled into this EDA report. This report highlights key observations and prepares us for the AI-driven analysis that lies ahead.

By the end of this week, we achieved a thorough understanding of EDA principles, honed our skills in data visualization, and developed the ability to extract meaningful insights from raw data. Additionally, we gained invaluable experience in documenting the EDA process, which will be instrumental as we progress to more complex AI-powered data analyses.

This foundational knowledge and skill set will be useful as we move forward with more complex AI-powered data analyses in the coming weeks.

1.1: Dataset Overview

The dataset we cleaned in week 01 contains detailed information on learner profiles, engagement metrics, and participation in various educational opportunities. It is structured to support analytical insights into learner demographics, engagement duration, and performance. Below is an in-depth description of the dataset's structure and key features:

Dataset Overview:

- **Total Entries**: 2,024 records
- **Column**: 36, encompassing categorical, numeric, and date-time attributes

Key Columns and Data Types:

- **Learner Profile Information**:
    - **Profile Id**: Unique identifier for each learner.
    - **First Name** and **Last Name**: Learner's name details.
    - **Date of Birth of Learner**: Age-related details are useful for demographic segmentation.
    - **Gender**: Categorical variable with values like "Male" and "Female".
- **Educational and Enrollment Information**:
    - **Institution Name**: Name of the educational institution associated with the learner.
    - **Graduation Date**: Expected or actual graduation date.
    - **Current Student Status** and **Current/Intended Major**: Describe the learner's educational status and field of study.
- **Opportunity Participation Details**:
    - **Opportunity Name** and **Opportunity Category**: Name and category of the educational opportunity (e.g., "Course").
    - **Opportunity Start Date** and **Opportunity End Date**: Duration of participation within a given opportunity.
    - **Time in Opportunity** and **Time in Opportunity (Days)**: Indicates the length of engagement in the program.
- **Engagement Metrics**:
    - **Engagement Duration**: Total duration a learner engaged with a program.
    - **Engagement Score**: A metric likely representing a learner's performance or engagement level.
    - **High Engagement**: Binary indicator where 1 suggests high engagement, and 0 suggests low or moderate engagement.
    - **Last Engagement Date** and **Days Since Last Engagement**: The date of the last engagement and the number of days since then, are useful for measuring recent engagement levels.
- **Sign-Up Details**:

- **Learner SignUp DateTime**: Date and time when the learner initially signed up.
- **SignUp Month** and **SignUp Year**: Temporal indicators derived from the sign-up date.
- **SignUp Day of Week**: Categorical variable indicating the day of the week of sign-up, potentially useful for weekly engagement trends.

Data Quality:

- **Completeness**: All fields are fully populated, indicating no missing values across the dataset.
- **Data Types**: The dataset uses appropriate data types, with numeric fields (such as Age of Learner, Engagement Score) stored as integers or floats and date-time fields (such as Learner SignUp DateTime) in string format.

Potential Analytical Use Cases:

- **Engagement Analysis**: Using Engagement Score and High Engagement to assess factors driving learner engagement.
- **Demographic Segmentation**: Analyzing age, gender, and location information for targeted program improvements.
- **Temporal Trends**: Studying SignUp Month, SignUp Year, and Days Since Last Engagement to understand engagement frequency and dropout patterns.

This Dataset we cleaned contains structured, high-quality data that offers valuable insights for assessing learner engagement, demographic characteristics, and temporal trends, supporting targeted engagement strategies and program improvements.

1.2: Analysis Goals

This is an outline of the primary analytical goals used for assessing learner engagement, and program effectiveness using the provided dataset.

Each goal aims to extract actionable insights that can inform strategic decision-making, program design, and learner retention efforts.

1. **Engagement Analysis**

Objective: To understand the drivers of learner engagement across different programs and identify factors that predict high or low engagement.

Engagement Score Trends:

- Analyze variations in Engagement Scores across different educational opportunities, timeframes, and demographic groups. Identify patterns that reveal which types of opportunities foster higher engagement.

High vs. Low Engagement Predictors:

- Investigate predictors of high engagement (High Engagement indicator). Determine if certain characteristics, such as sign-up timing, age, or opportunity type, correlate with elevated engagement scores.

2. **Opportunity Effectiveness**

Objective: To evaluate the effectiveness of various opportunity categories and durations on learner engagement and performance.

Success Rates by Opportunity Category:

- Compare Opportunity Category types (e.g., courses, internships) to identify which categories yield higher engagement scores or engagement duration, revealing which formats best engage learners.

Opportunity Duration vs. Engagement:

- Investigate the relationship between Time in Opportunity and engagement metrics, assessing whether longer opportunities correlate with sustained or diminishing engagement.

Opportunity Completion and Satisfaction:

- Analyze Time in Opportunity (Days) alongside engagement metrics to uncover any correlations between the time spent and completion or satisfaction levels.

3. **Temporal Analysis**

Objective: To identify temporal patterns in learner engagement, helping to improve outreach, scheduling, and engagement strategies.

Sign-Up Trends:

- Examine SignUp Month, SignUp Year, and SignUp Day of Week to understand temporal sign-up patterns. This analysis can guide marketing and outreach efforts to target peak registration periods.

Optimal Engagement Duration:

- Analyze Engagement Duration and Time in Opportunity to identify an optimal timeframe that maximizes learner engagement. This can inform the design of future programs.

Seasonal Shifts in Engagement:

- Track seasonal or monthly shifts in Engagement Scores to understand when learners are most and least engaged. These insights may help in scheduling events or resource allocations.

2. Exploratory Data Analysis

In Exploratory Data Analysis (EDA), we started with the dataset of week 1. There were the following issues in the dataset:

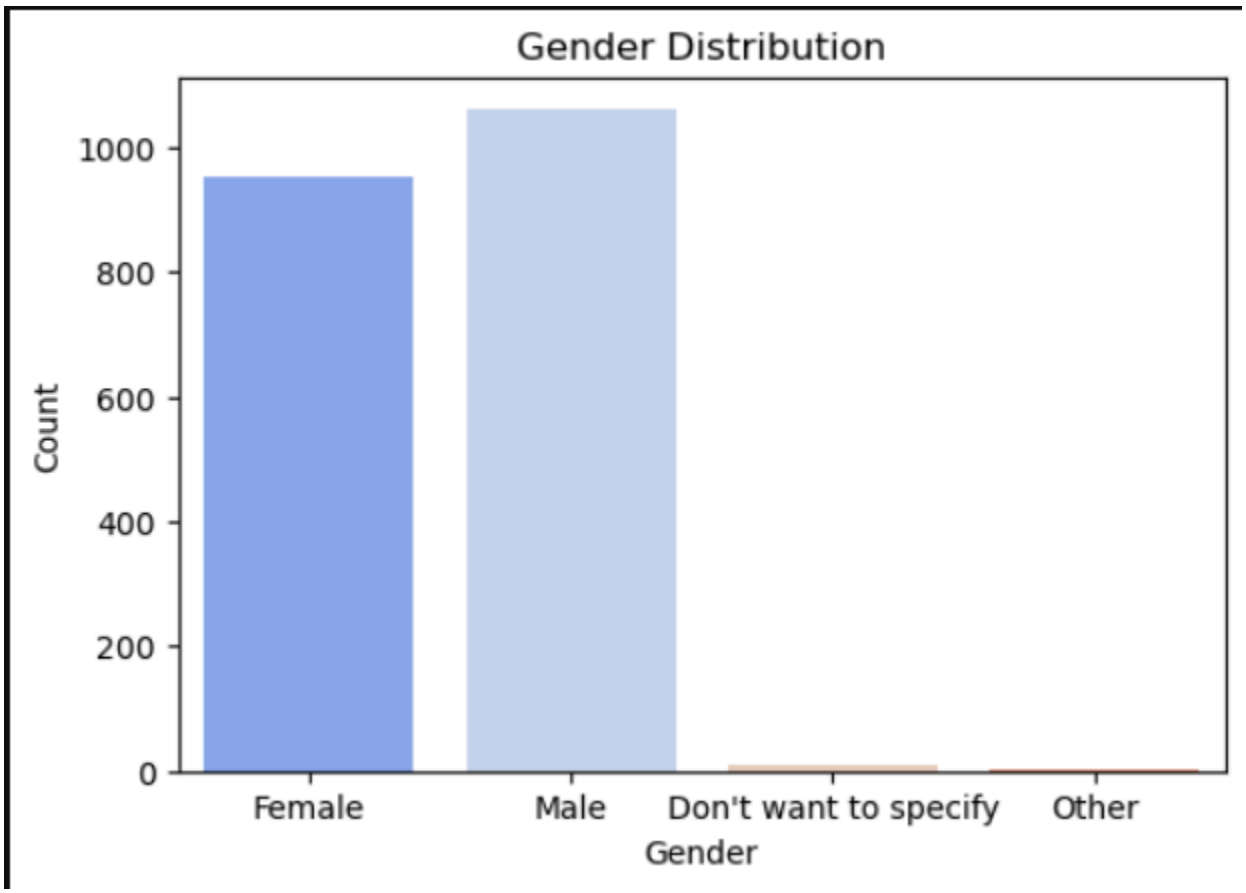
- There were a few missing values in the dataset.
    - Learner SignUp DateTime
    - SignUp Month
    - SignUp Year
    - SignUp Day of Week
    - Last Engagement Date
    - Days Since Last Engagement
    - The above missing values were dropped.
- The Opportunity and its generated columns were creating a havoc in the dataset. There were way too many missing entries, to the point that if we cleaned those rows completely, much of the data would have been lost. But on the other hand, there was no

way to fill in those missing values, since a date can not be assumed or calculated. Hence, the team decided to drop the dataset, keeping in view the future use-case of Artificial Intelligence-based models. Also, because we shall be doing model prediction, we have left the tasks such as feature scaling and one hot encoding for that week.

The data was already cleaned and validated in week 1. A few of the feature engineering steps were also performed in Week 1, hence, there was no particular need to perform new data cleaning and validation steps.

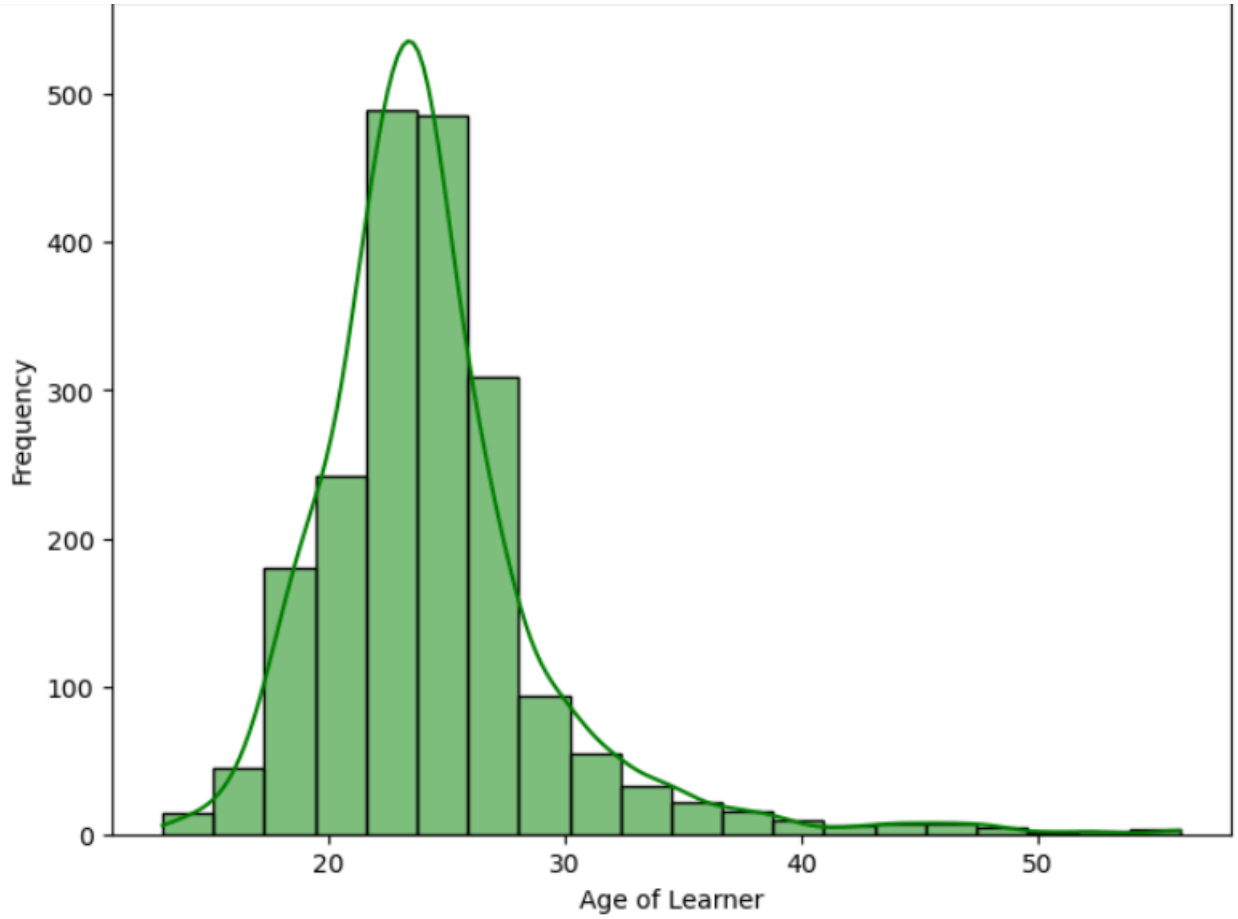
3. Insight Generation

3.1: Simple Insight Generation:



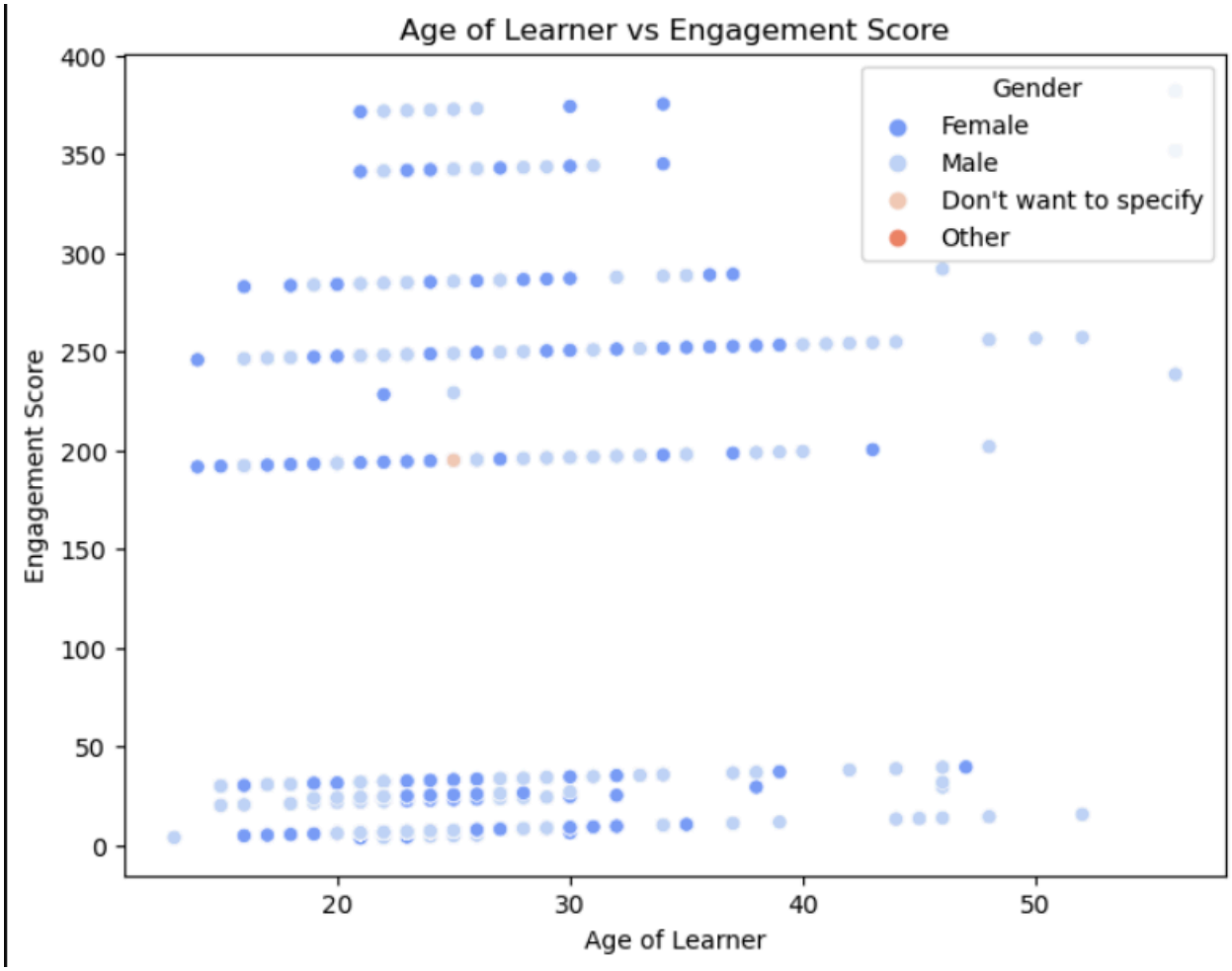
This simple plot tells us that the Males slightly dominate the Females in this opportunity at Excelerate.

3.1.1: Age of Learner



Tells us about the learner's age, and the high bars fall between ages 18 and 28.

3.1.2: Age of Learner vs Engagement Score

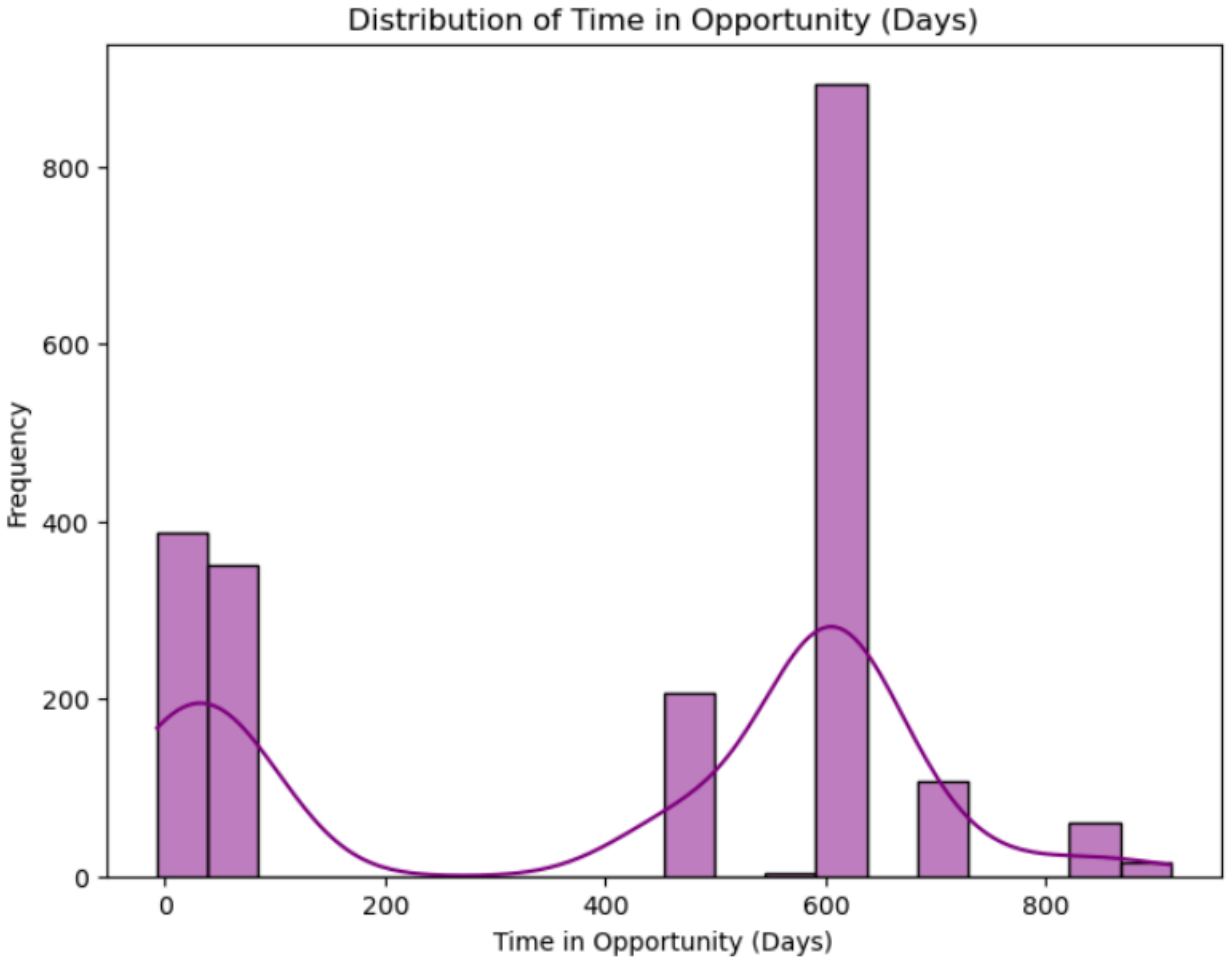


Age distribution: The age range of the learners is from approximately 15 to 50 years old. Engagement score distribution: The engagement scores range from 0 to around 350, with a clustering of scores around 200-250.

Gender differences:

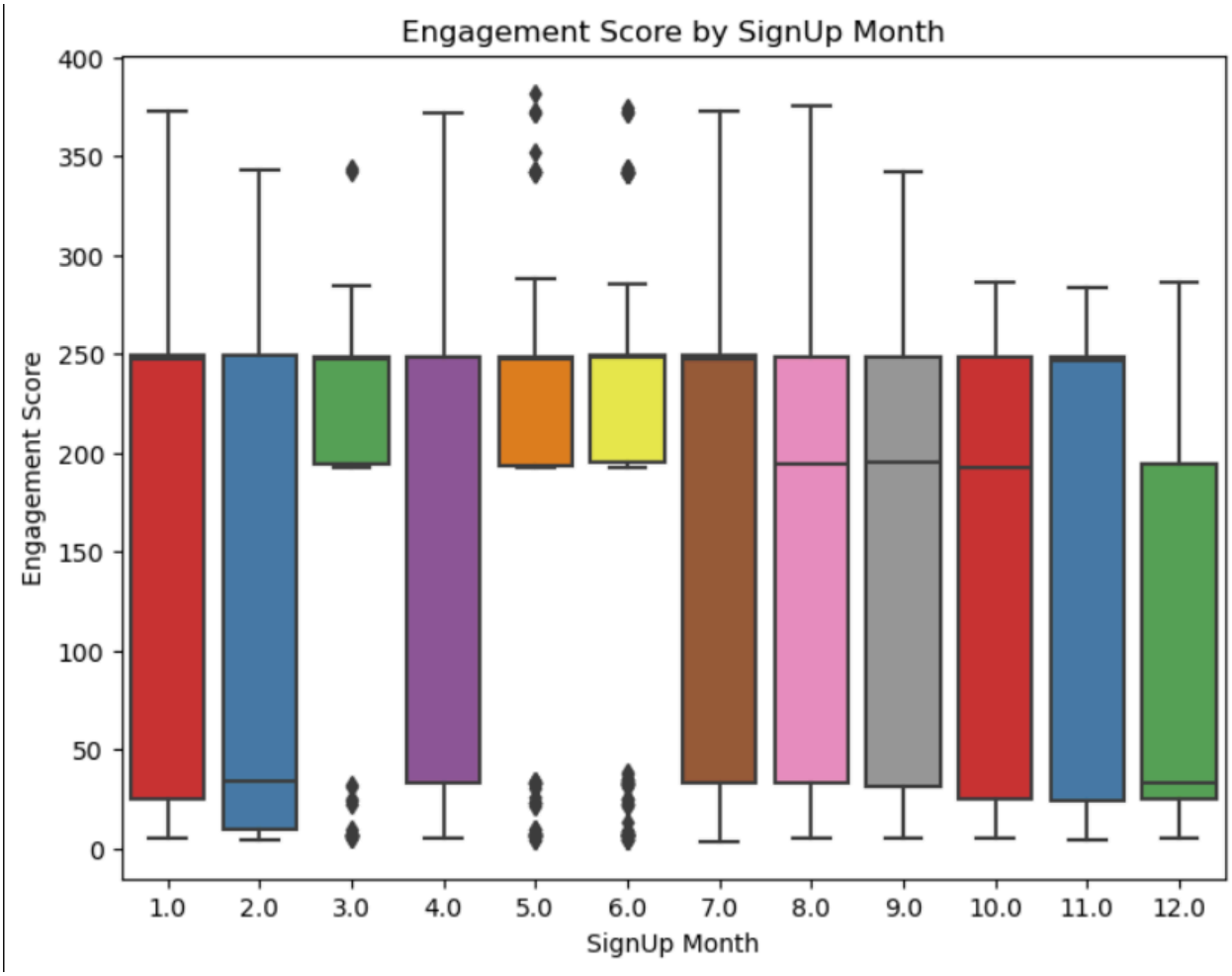
- **Male:** There seems to be a slight trend of higher engagement scores among males, especially in the younger age groups.
- **Female:** Female engagement scores are more spread out, with a slightly higher concentration in the lower engagement ranges.
- **Other:** The "Other" and "Don't want to specify" categories have limited data points, making it difficult to conclude their engagement patterns.

3.1.3: Opportunity Days Distribution:



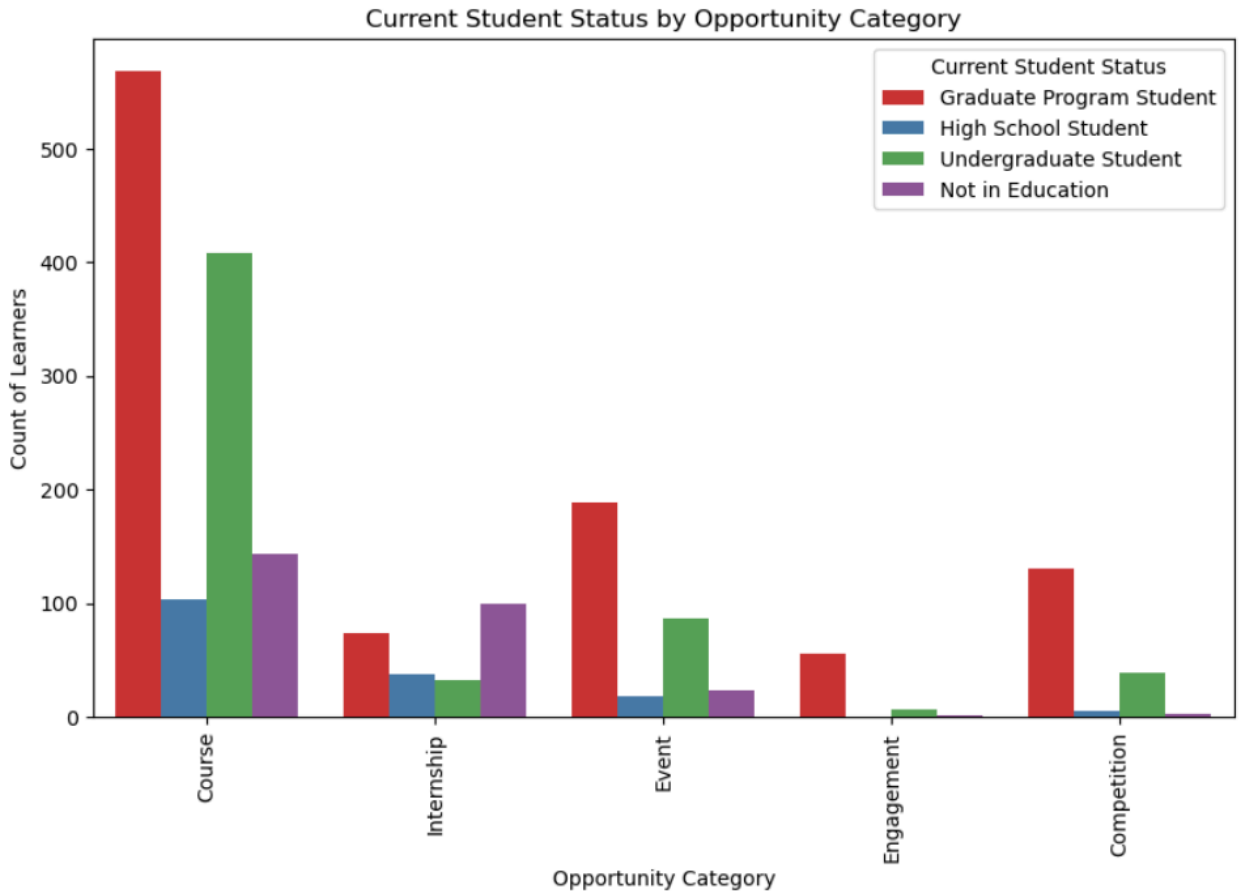
- Users were initially engaged in the first 70-80 days.
- Then a spike at the 450ish day mark.
- Then after 600 days which is very surprising, should not have happened. Like why would a user be engaging for 600 days in a 1-month internship?
- Then it settles down at the 800-day mark.

3.1.4: Engagement Score by Signup Month



- Month 7 has one of the highest median engagement scores, suggesting users signing up in this month tend to have higher engagement.
- Months 1 and 4 have the largest variability, as their engagement scores are spread over a wider range.
- Some months have a higher number of outliers (e.g., months 4 and 6), indicating that while most users in these months had similar engagement, a few users exhibited much higher or lower engagement.
- Uses Box plot

3.1.5: Current Student Status By Opportunity Category:



Overall:

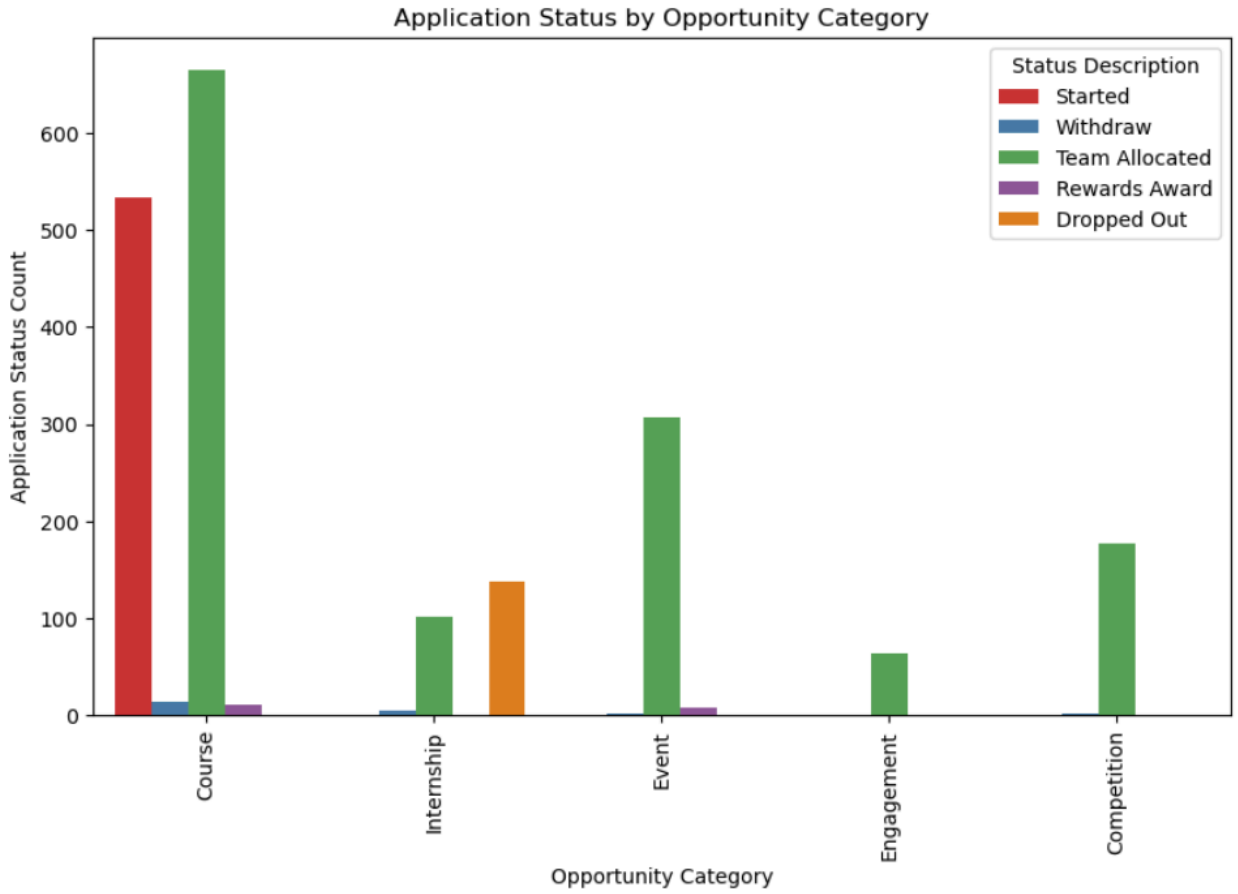
- **Courses and Internships:** These two categories attract a significant number of learners, especially graduate program students.
- **Events and Engagements:** These categories seem to attract a more diverse range of learner statuses, including high school students, undergraduate students, and those not currently in education.
- **Competitions:** This category has the lowest overall participation, with graduate program students being the most represented.

Specific Insights:

- **Graduate Program Students:** This group is highly represented in courses and internships, indicating a strong interest in skill development and career opportunities.
- **Undergraduate Students:** While they are present in all categories, their participation is most prominent in events and engagements.
- **High School Students:** Their participation is relatively low, with some interest in events and engagements.

- **Not in Education:** This group is more likely to participate in events and engagements, suggesting a desire for learning and networking opportunities.

3.1.6: Application Status by Opportunity Category:



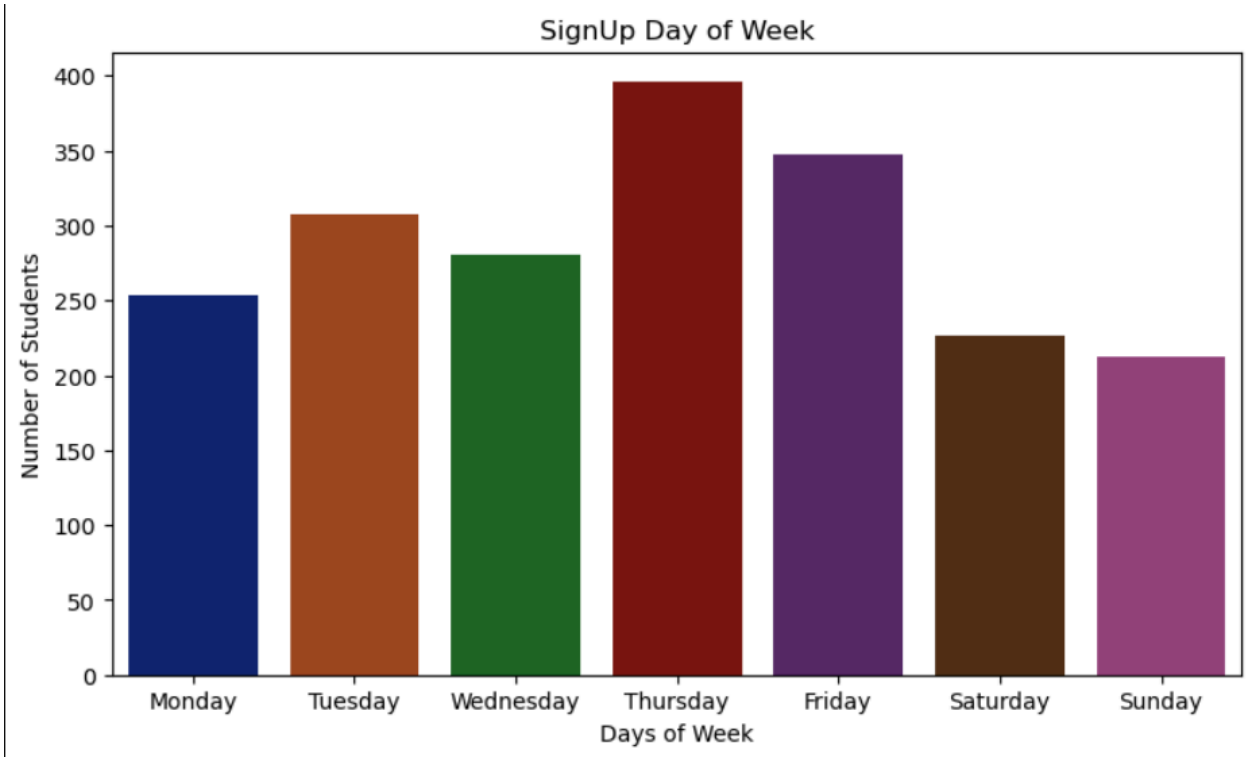
Overall:

- **Courses:** This category has the highest number of applications, with "Started" being the most common status.
- **Internships:** This category has a significant number of "Withdraw" and "Dropped Out" statuses, indicating potential challenges in securing internships.
- **Events:** This category has a high number of "Team Allocated" statuses, suggesting that many applications are being considered for team-based events.
- **Engagements and Competitions:** These categories have fewer applications compared to courses and internships, with a mix of statuses.

Specific Insights:

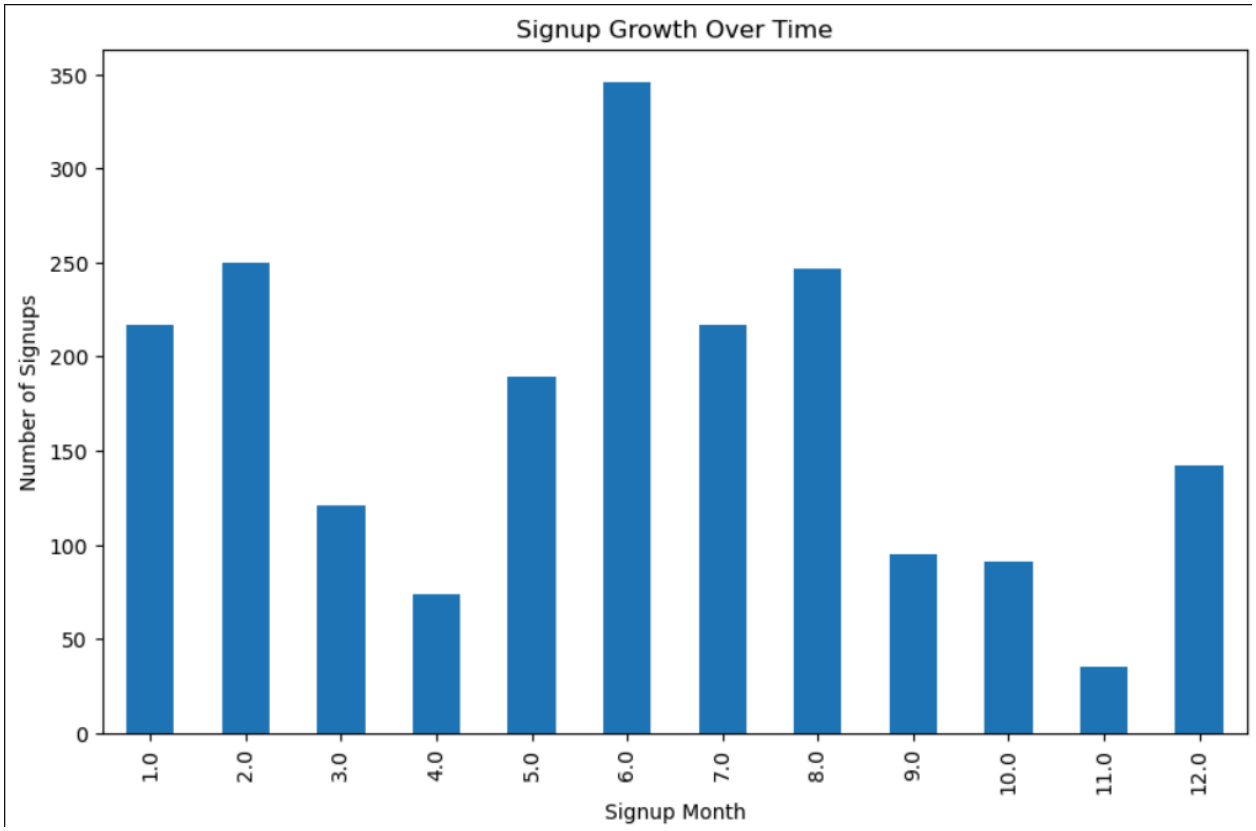
- **Started:** This status indicates that learners have actively begun participating in the opportunity.
- **Withdraw:** This status suggests that learners have chosen to withdraw from the opportunity after applying.
- **Team Allocated:** This status implies that learners have been assigned to a team for a specific opportunity.
- **Rewards Award:** This status signifies that learners have received rewards or recognition for their participation.
- **Dropped Out:** This status indicates that learners have stopped participating in the opportunity after starting.

3.1.7: SignUp Day of the Week:



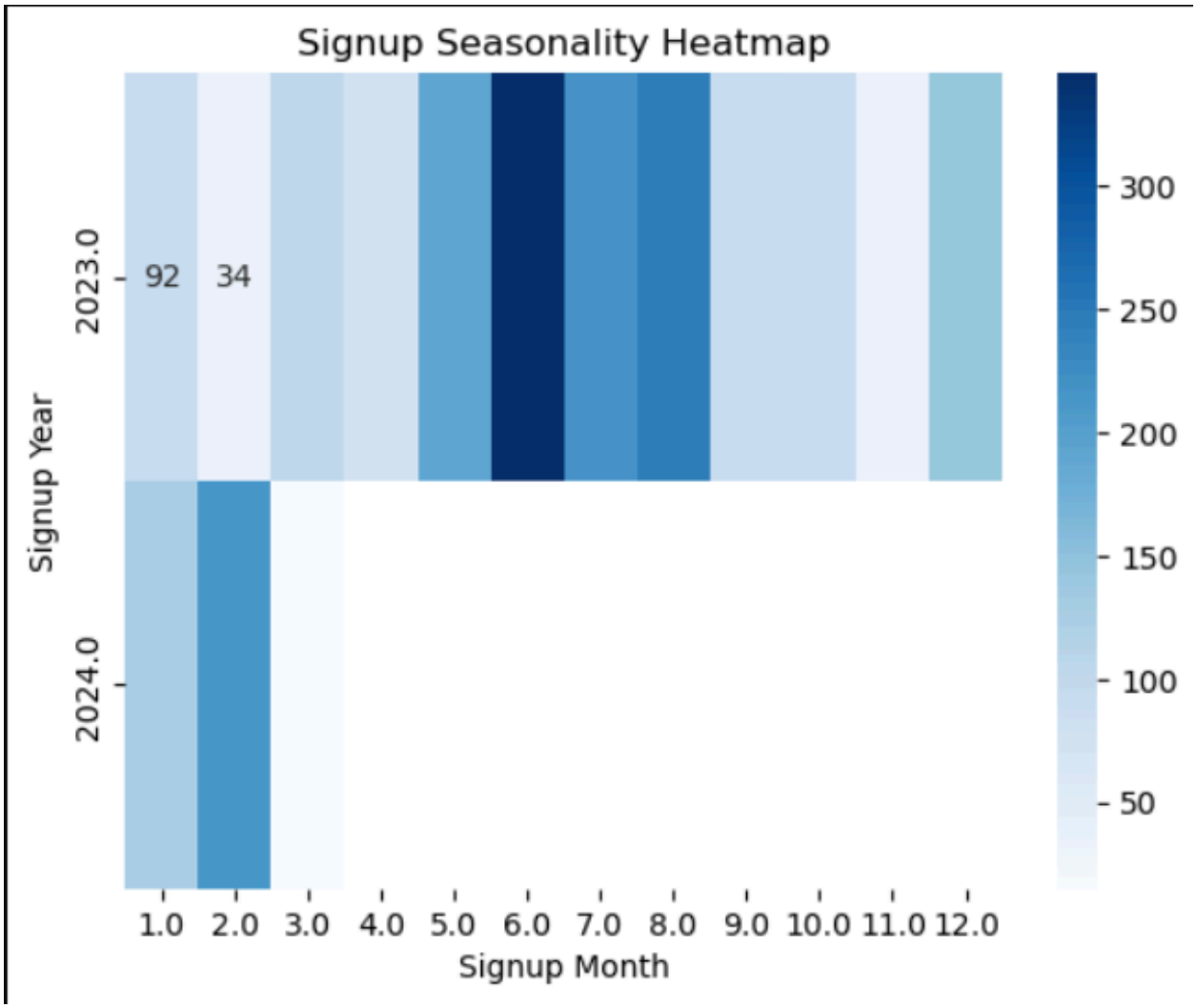
Thursday has the highest signup rates as indicated above, with Sunday having the lowest.

3.1.8: SignUp Growth Overtime:



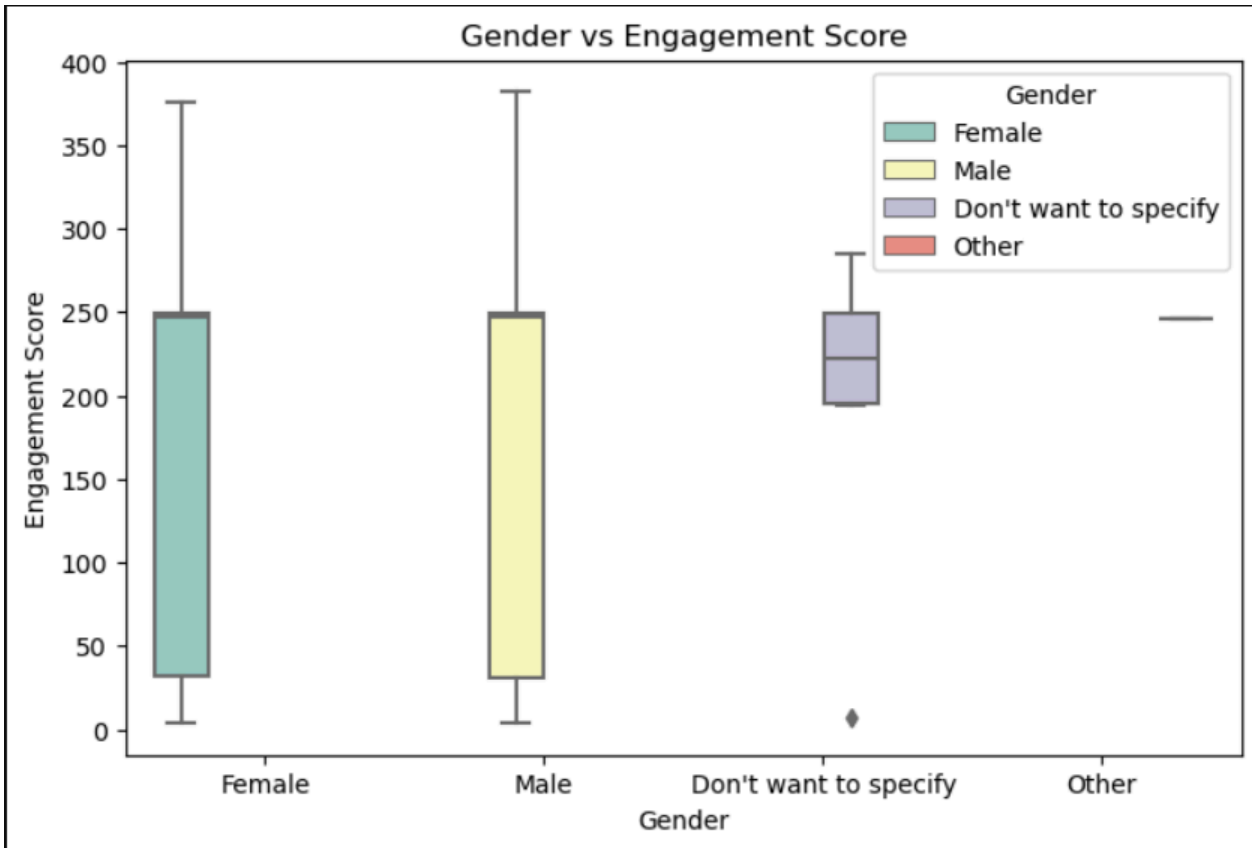
November had the lowest growth rate, while June had the highest.

3.1.9: Signup Seasonality:



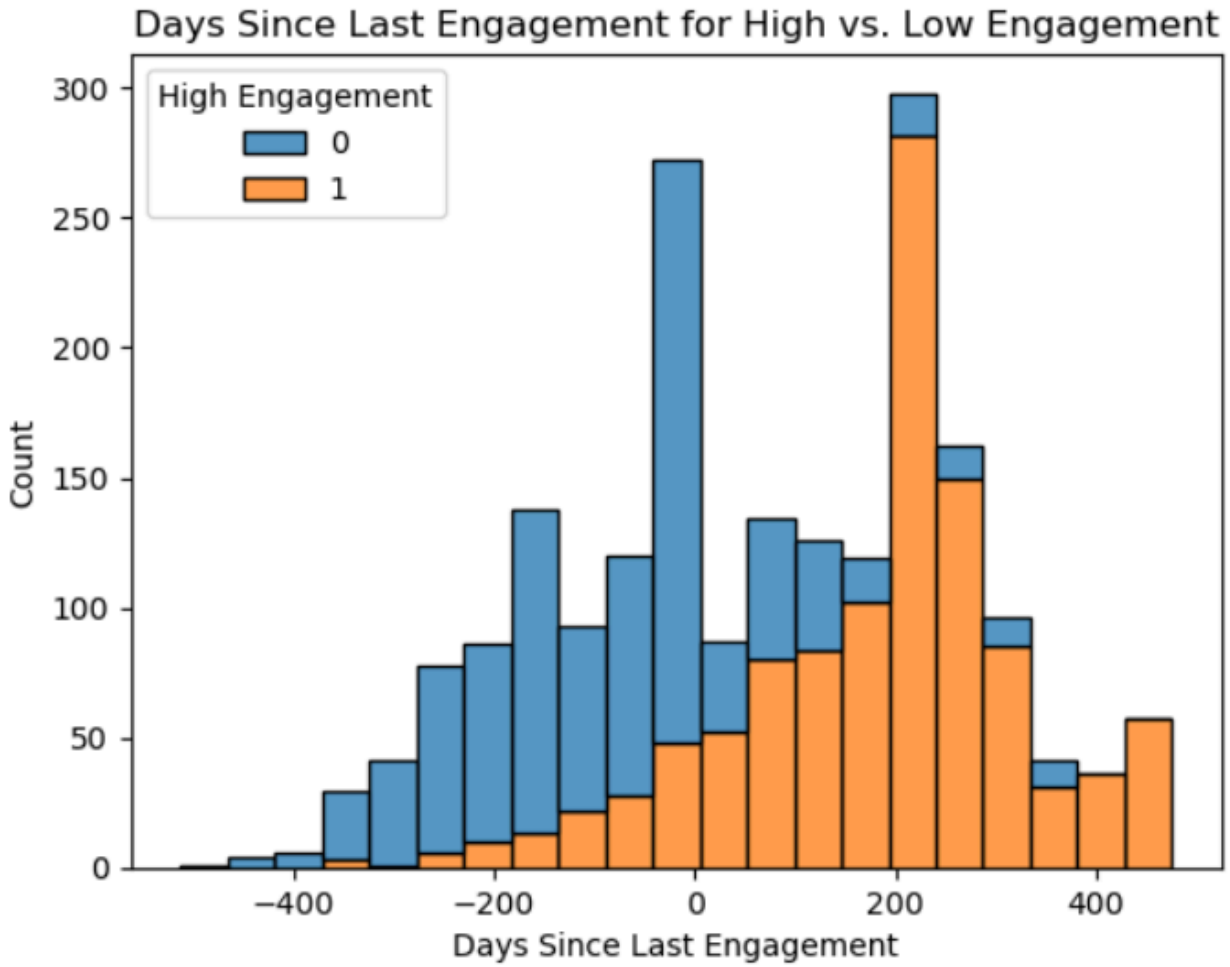
- **2023:** There seems to be a peak in signups around the 6th month (June) and another smaller peak around the 10th month (October).
- **2024:** The data for 2024 is incomplete, but it appears to be following a similar trend as 2023, with a peak around the 6th month.

3.1.10: Gender vs Engagement Score:



- **Female and Male:** Both genders have similar distributions of engagement scores, with the median score being around 250.
- **Don't want to specify and Other:** These groups have a smaller sample size, and the distribution of scores is less clear. However, it appears that the median score for "Don't want to specify" is slightly higher than the other groups.

3.1.11: Days Since Last Engagement for High vs Low Engagement



Distribution of Days Since Last Engagement:

- **High Engagement:** The distribution is skewed towards the left, indicating that a majority of highly engaged users have recently engaged. There are fewer instances of high-engagement users having not engaged for a long time.
- **Low Engagement:** The distribution is more evenly spread out, suggesting that low-engagement users have a wider range of days since their last engagement. This could mean that they engage less frequently or have longer periods of inactivity.

Comparison of the Two Groups:

- The plot clearly shows that high-engagement users tend to have more recent engagement compared to low-engagement users.

- The peak of the high-engagement group is closer to 0 (recent engagement), while the peak of the low-engagement group is shifted towards the right, indicating less recent engagement.

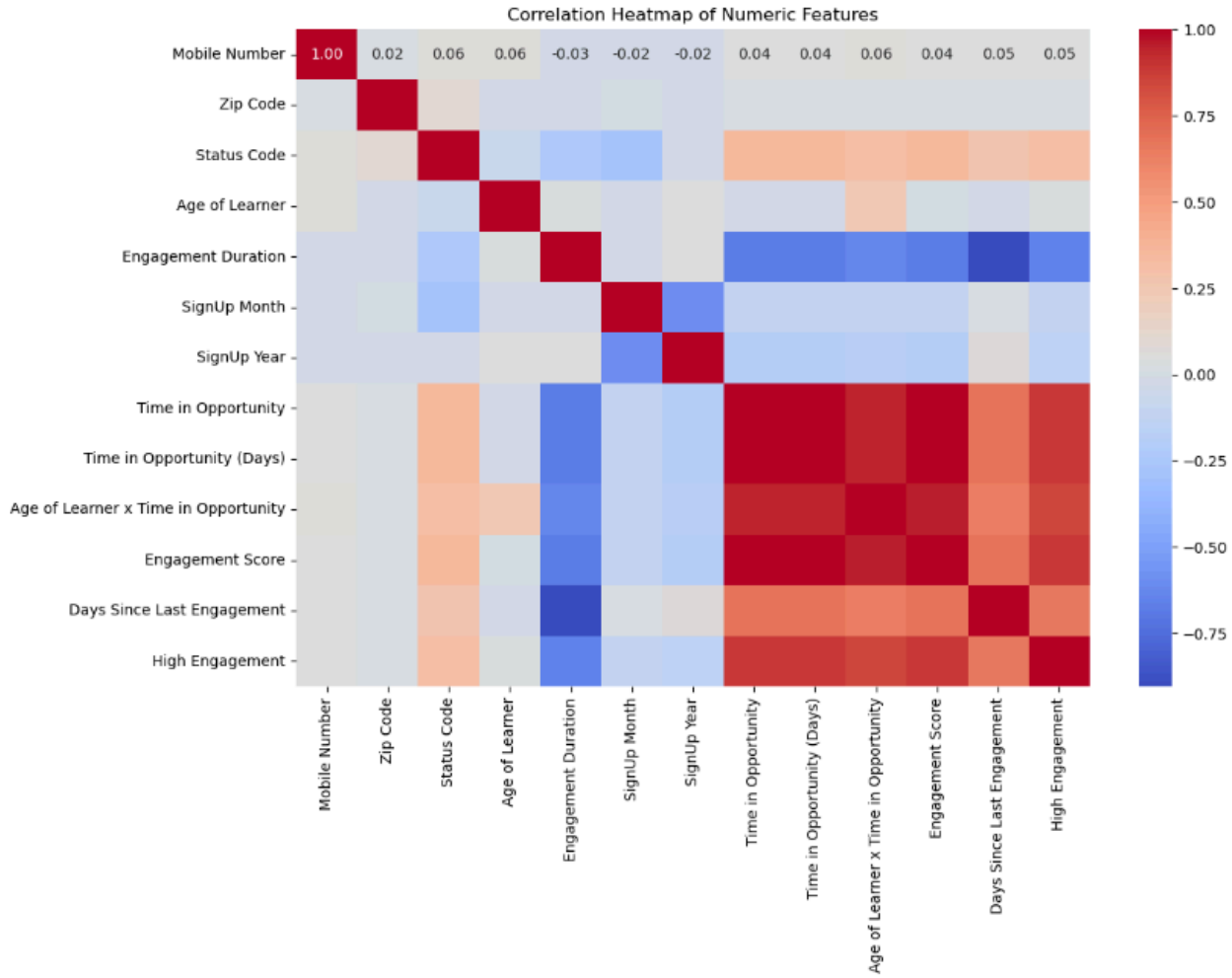
3.2: Advanced Insight Generation:

3.2.1: Correlation Heatmap:

The corr() function is used to compute the correlation coefficient between numeric columns in a data frame. Correlation measures the strength and direction of the linear relationship between two variables, returning a value between -1 and 1:

- **1:** Perfect positive correlation, where one variable increases as the other does.
- **-1:** Perfect negative correlation, where one variable increases as the other decreases.
- **0:** No linear relationship between the variables.

Following is a heatmap that visualizes the correlation between numeric features in the dataset:



Strong Positive Correlations:

- **Mobile Number and Zip Code:** This indicates a high degree of association between these two features, which is likely because a mobile number is often linked to a specific zip code.
- **Age of Learner and Engagement Duration:** This suggests that older learners tend to engage for longer durations.

Moderate Positive Correlations:

- **Age of Learner and Time in Opportunity:** Older learners may have been in the opportunity for a longer time.
- **Age of Learner and Engagement Score:** Older learners might have a higher engagement score.

Moderate Negative Correlations:

- **Days Since Last Engagement and High Engagement:** This is expected, as more recent engagement would indicate higher engagement.

Weak Correlations:

- **Mobile Number with other features:** This suggests that mobile number is not strongly related to other variables.

Additional Observations:

- The diagonal of the heatmap is always 1, representing the perfect correlation of a variable with itself.
- The color intensity reflects the strength of the correlation, with darker shades indicating stronger relationships.

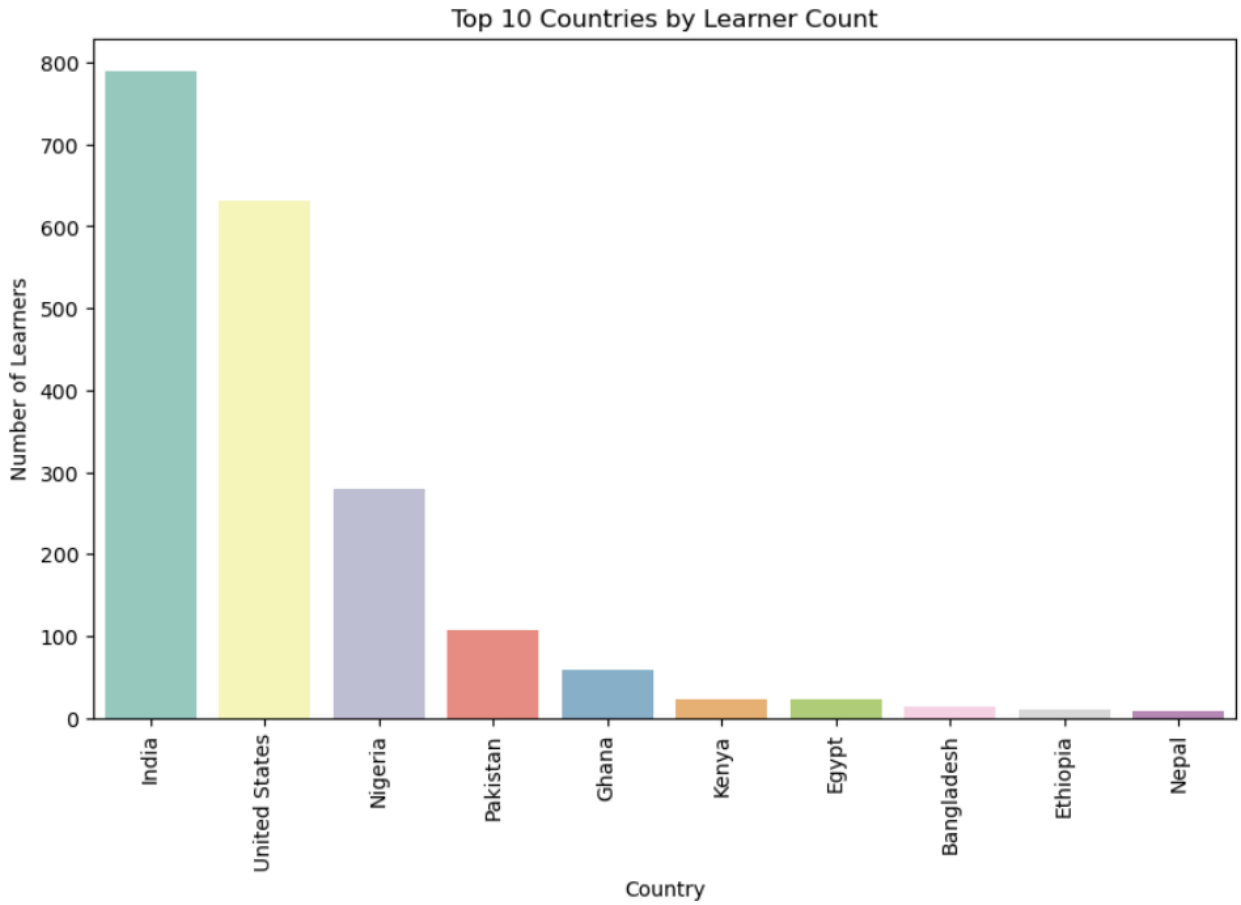
Interpreting the Color Scale:

- **Red:** Positive correlation (variables tend to increase together)
- **Blue:** Negative correlation (variables tend to decrease together)
- **White:** No correlation (variables are not linearly related)

3.2.2: Country-Wise Learner Distributions:

Plots the values of the top 10 countries by using this method:

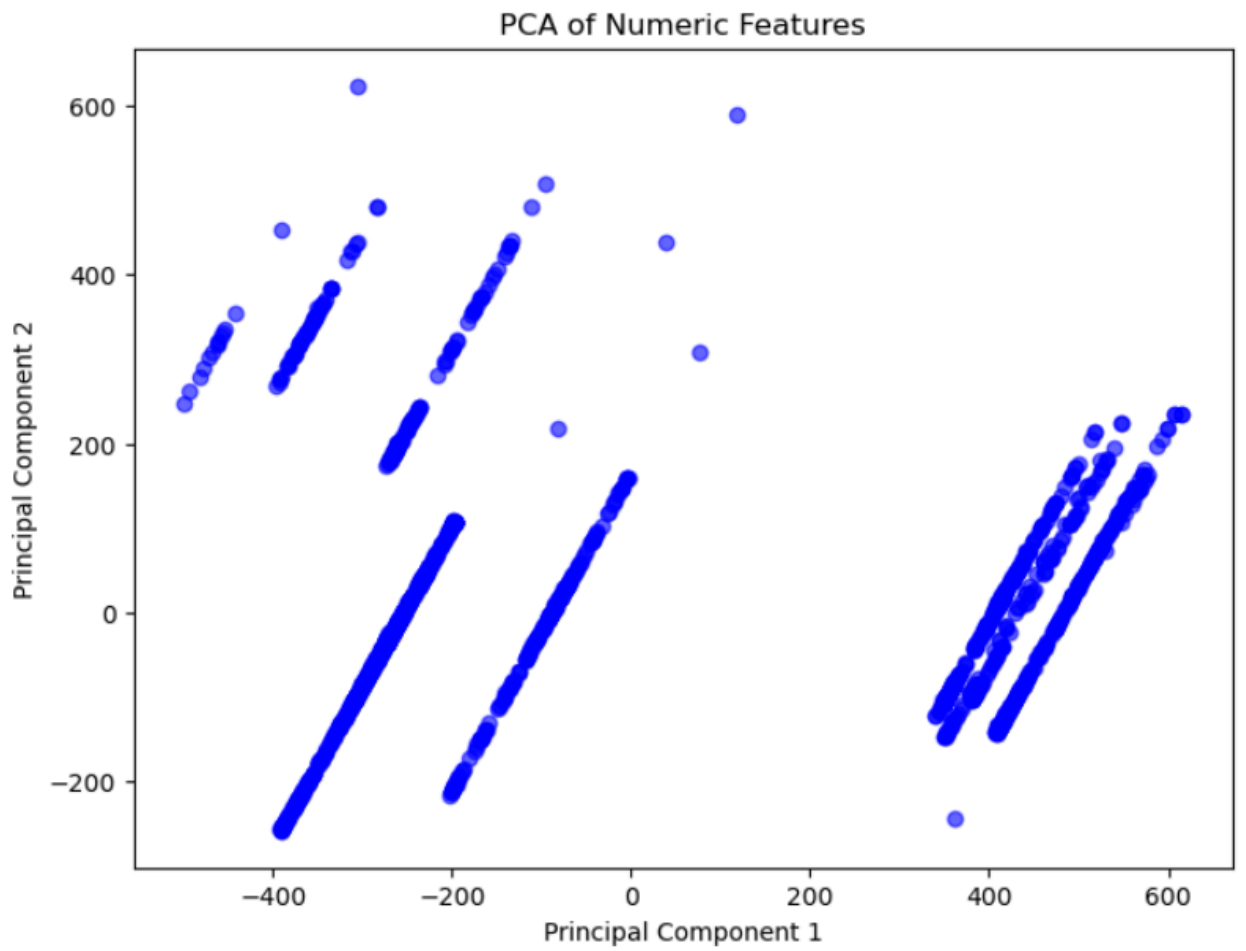
top_countries = students['Country'].value_counts().nlargest(10)



- **India Dominance:** India has the highest number of learners, far surpassing the other countries. This suggests a significant presence and interest in learning within India.
- **United States and Nigeria:** The United States and Nigeria follow India with a substantial number of learners. This indicates a strong learning community in these countries as well.
- **Declining Learner Count:** From Nigeria onwards, there's a noticeable decline in the number of learners per country. This suggests a decreasing trend in learner participation as we move down the list.
- **Bottom Countries:** Countries like Ethiopia and Nepal have the lowest number of learners, indicating a relatively smaller learning community compared to the top countries.

3.2.3: Principal Component Analysis for Dimensionality Reduction

The following visualize high-dimensional data in a 2D space. It lets us see patterns, clusters, or separations in the data based on engagement-related features using PCA.



- **Clustering:** The data points seem to cluster into distinct groups. This suggests that there might be underlying patterns or categories within the data.

- **Principal Components:** The x-axis represents the first principal component (PC1), and the y-axis represents the second principal component (PC2). PC1 captures the most variation in the data, while PC2 captures the second most variation.

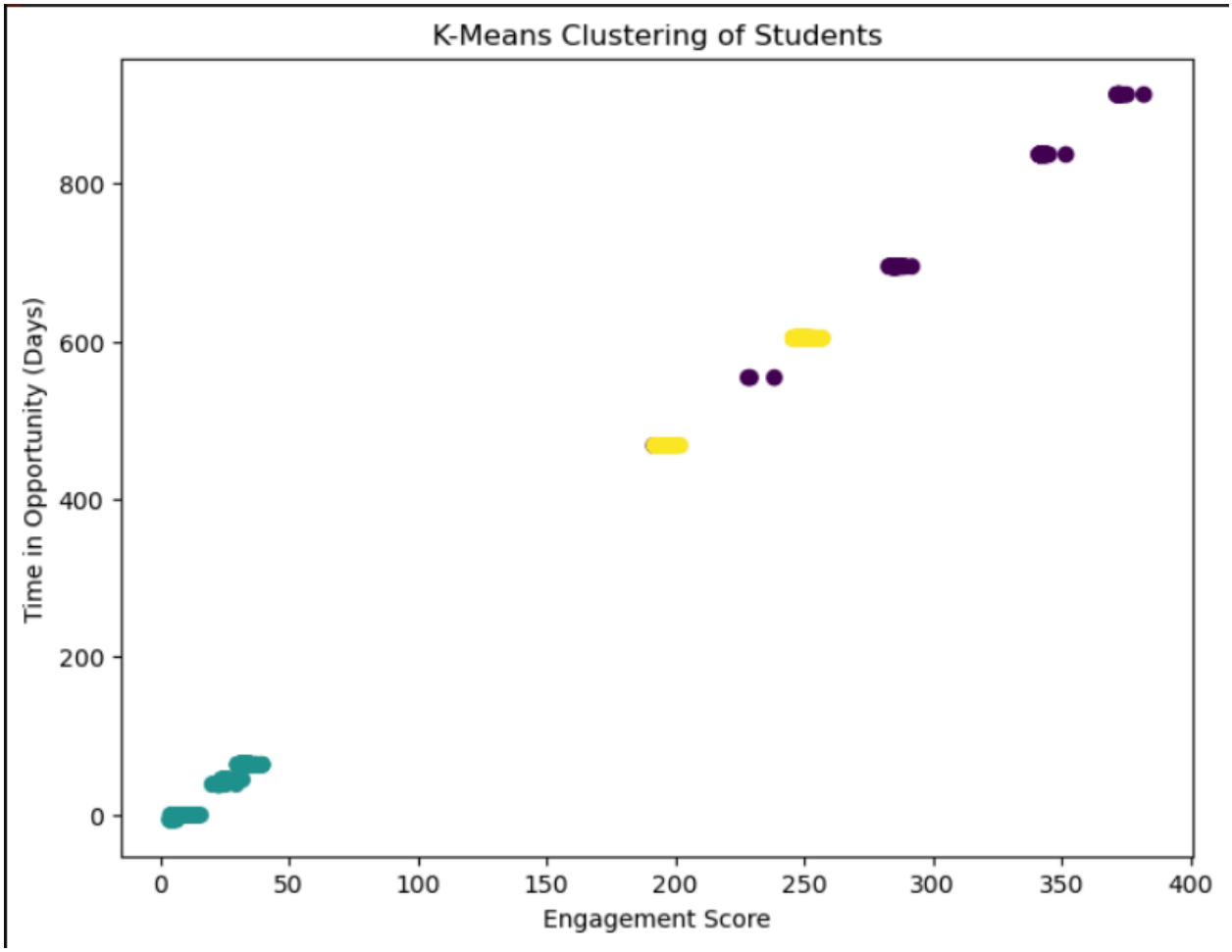
- **Data Distribution:** The data points are spread out along diagonal lines. This indicates that the data might not be normally distributed along these dimensions.

3.2.4: K-Means Clustering of Students:

This is a K-Means clustering analysis on a dataset of students based on their engagement scores and time-related metrics. It begins by defining a K-Means model with 3 clusters, which

represent three different groupings or segments. The data for clustering is drawn from the columns 'Engagement Score', 'Time in Opportunity (Days)', and 'Days Since Last Engagement', with any rows containing missing values removed.

Once the model is fit to the data, the clustering labels (group assignments) are generated for each student. A scatter plot is then created to visually display the clusters, with each point colored according to its cluster label. This plot, with 'Engagement Score' on the x-axis and 'Time in Opportunity (Days)' on the y-axis, shows how the students are grouped based on these engagement-related features.



Clustering: The data points are divided into three distinct clusters, represented by different colors (teal, yellow, and purple). This suggests that the students can be categorized into three groups based on their engagement score and time in opportunity.

Engagement Score and Time in Opportunity: The x-axis represents the engagement score, and the y-axis represents the time in opportunity (in days). Each data point represents an individual student.

Cluster Characteristics:

- **Teal Cluster:** This cluster contains students with low engagement scores and relatively short time in opportunity. They may be new students who haven't had much time to engage.
- **Yellow Cluster:** Students in this cluster have moderate engagement scores and a moderate time in opportunity. They might be consistently engaged but not at a very high level.
- **Purple Cluster:** This cluster consists of students with high engagement scores and a long time in opportunity. They are likely highly engaged and have been involved for a significant period.

Insights:

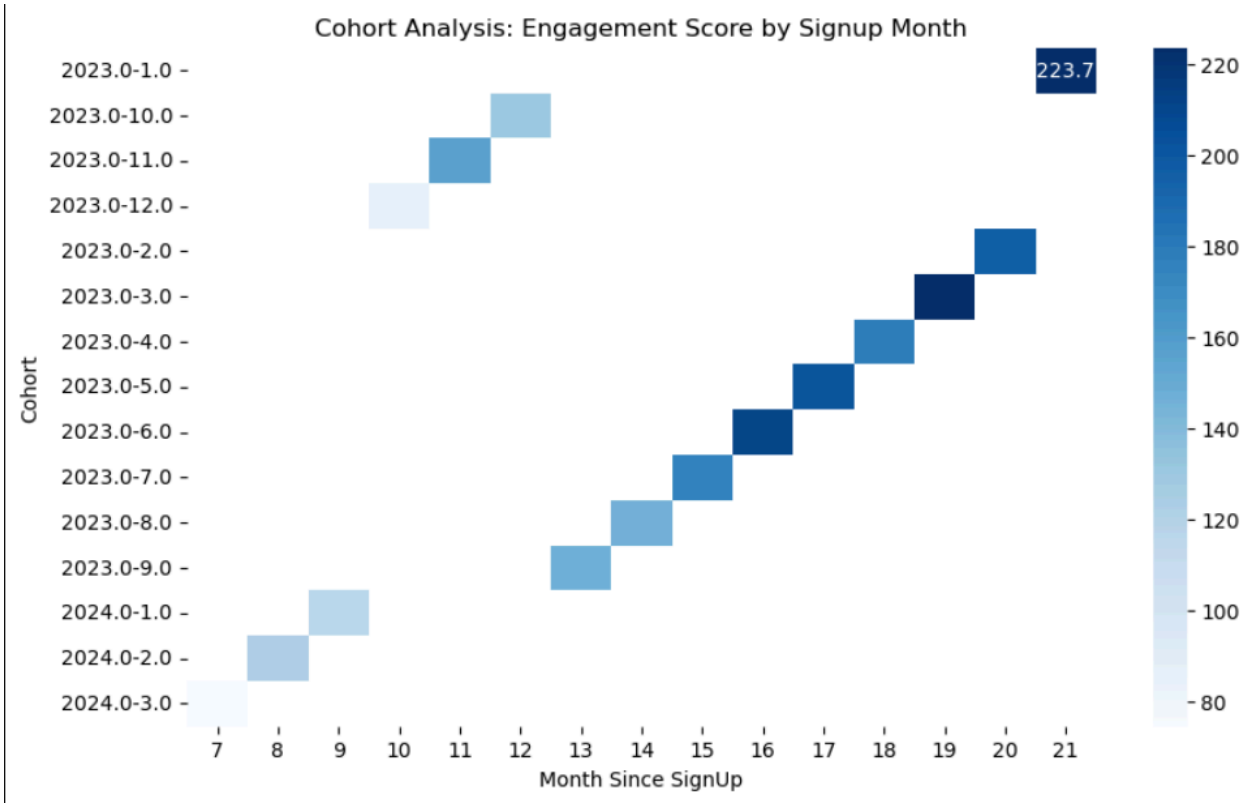
- The clustering suggests that there are different segments of students based on their engagement patterns.
- The analysis can help identify students who might need additional support or encouragement to increase their engagement.
- The clusters can be used to target specific interventions or strategies to improve student engagement.

3.2.5: Cohort Analysis: Engagement Score by Signup Month

This insight performs a cohort analysis to examine how student engagement changes over time sign-up. It starts by loading student data and renaming columns for consistency. It then creates a 'Cohort' column by combining the 'year' and 'month' of signup and a 'SignUp Date' column, setting each signup date to the 1st day of the month. Using the current date, it calculates the 'Month Since SignUp' for each student to find the number of months since they joined.

The code then groups data by 'Cohort' and 'Month Since SignUp', calculating the average 'Engagement Score' for each cohort over time and creating a pivot table. Finally, a heatmap is generated using Seaborn to visualize changes in engagement scores across cohorts, with

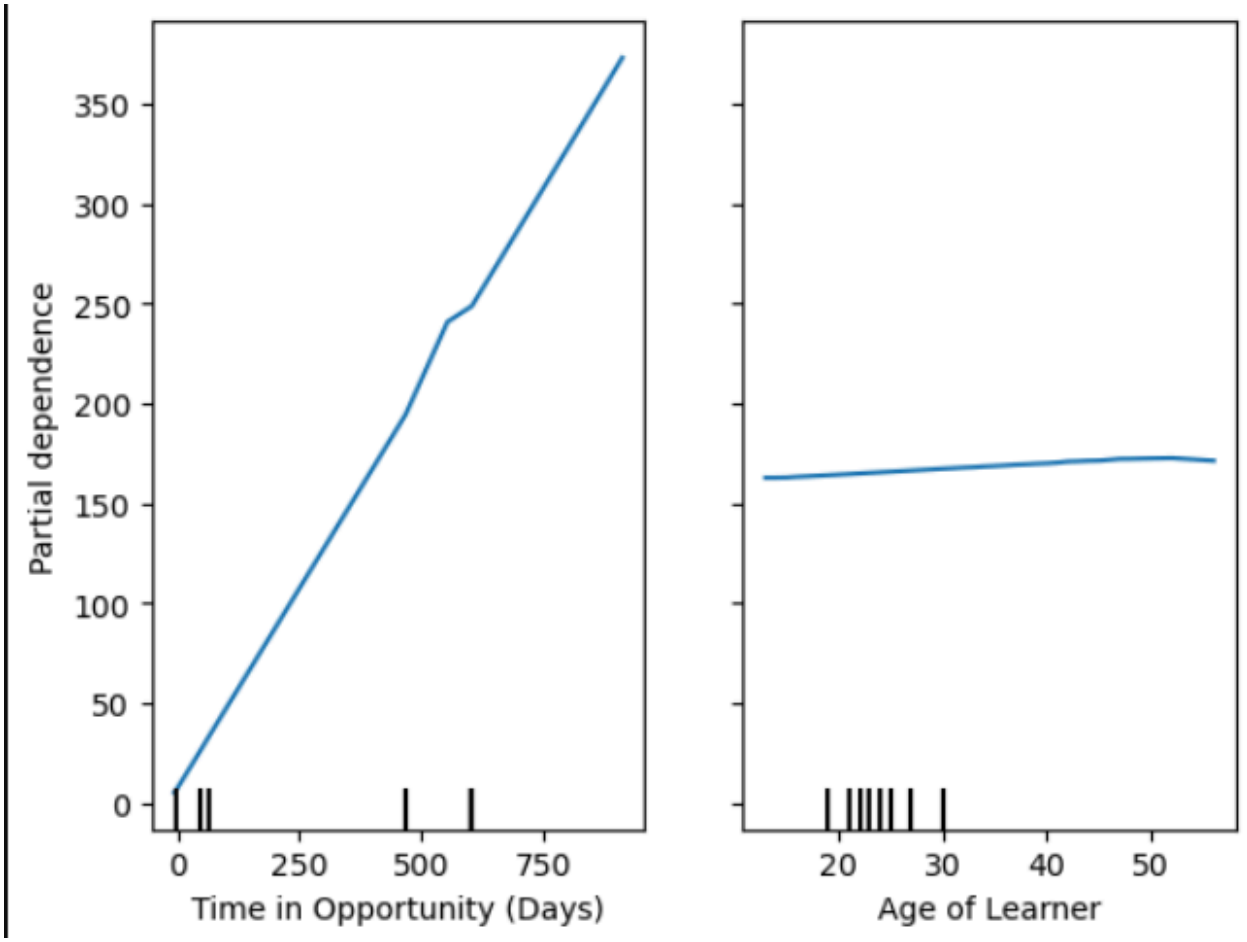
darker shades representing higher engagement. This provides insight into engagement trends based on the time since signup.



This cohort analysis helps you understand how user engagement changes over time for different groups of users who signed up at different points in time. It can be used to identify potential issues, optimize engagement strategies, and make data-driven decisions.

3.2.6: Partial Dependence Plot (Machine Learning Context):

This plot starts by training a RandomForestRegressor model using Time in Opportunity (Days), Days Since Last Engagement, and Age of the Learner as predictors for the Engagement Score. After training, it plots the partial dependence of the Engagement Score on Time in Opportunity (Days) and Age of the Learner, isolating the effect of each feature on predictions while averaging out the influence of other features. The plot helps in understanding the influence of these features on engagement scores.



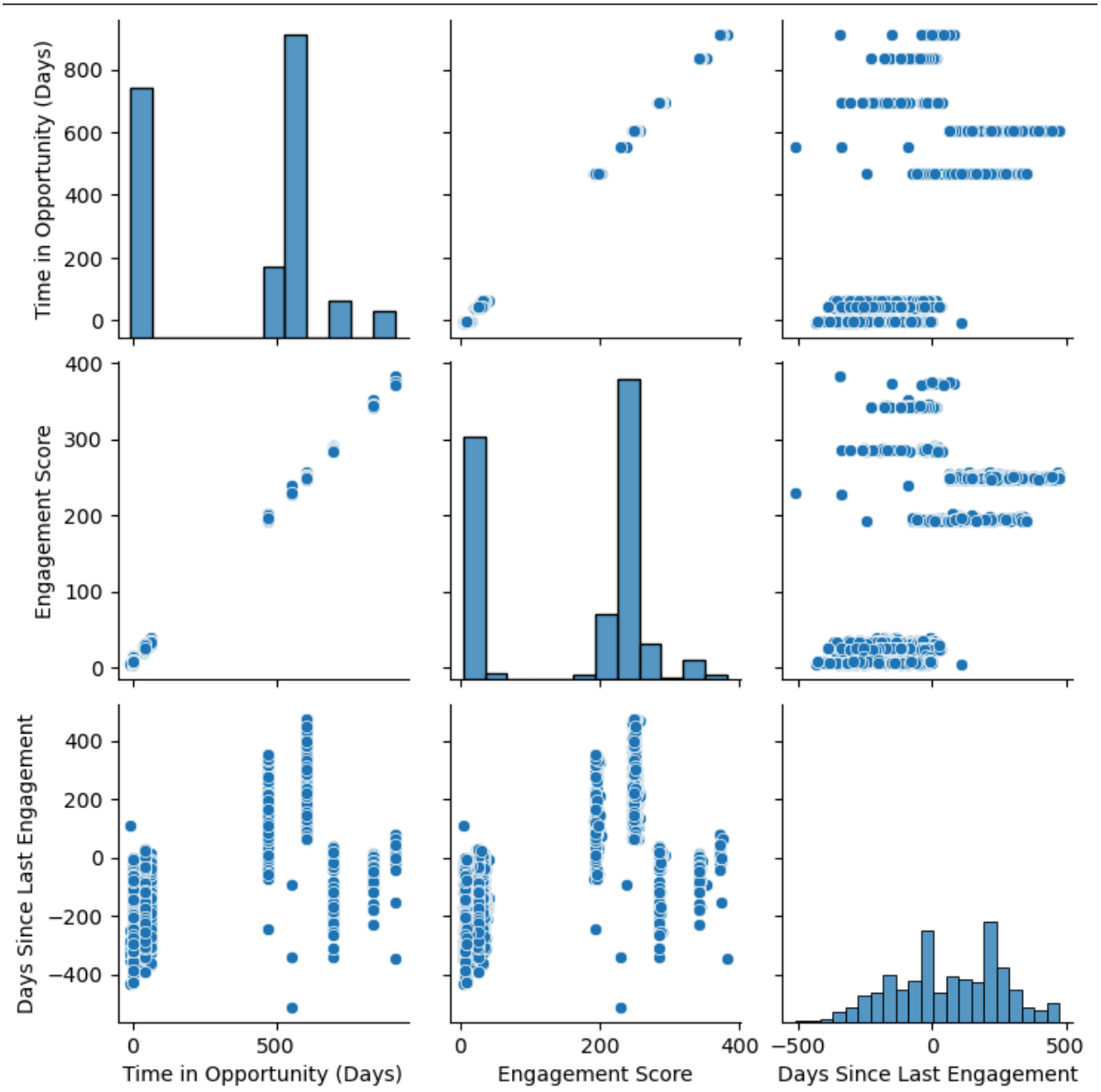
Left Plot (Time in Opportunity):

- As the time in opportunity increases, the predicted outcome (likely engagement or conversion) also increases.
- This suggests that the longer a learner is in an opportunity, the higher the likelihood of them engaging or converting.

Right Plot (Age of Learner):

- The predicted outcome seems to be relatively stable across different ages of learners.
- This suggests that the age of the learner might not be a strong predictor of engagement or conversion in this model.

3.2.7: Pair Plot (Scatter Matrix):



- **Engagement Score:** This is the original time series data, showing the engagement score over time. It appears to have some fluctuations and patterns.

- **Trend:** This component captures the long-term direction of the engagement score. It seems to have a slight upward trend overall.

- **Seasonal:** This component shows any recurring patterns within the data, such as seasonal variations. In this case, there might be some minor seasonal fluctuations, but they are not very pronounced.

- **Residual:** The residual component represents the random noise or unexplained variation in the data. It shows the differences between the actual data points and the sum of the trend and seasonal components.

Overall, the graph suggests that the engagement score has a slight upward trend over time, with some minor seasonal variations and random fluctuations.

4. Hypothesis Development:

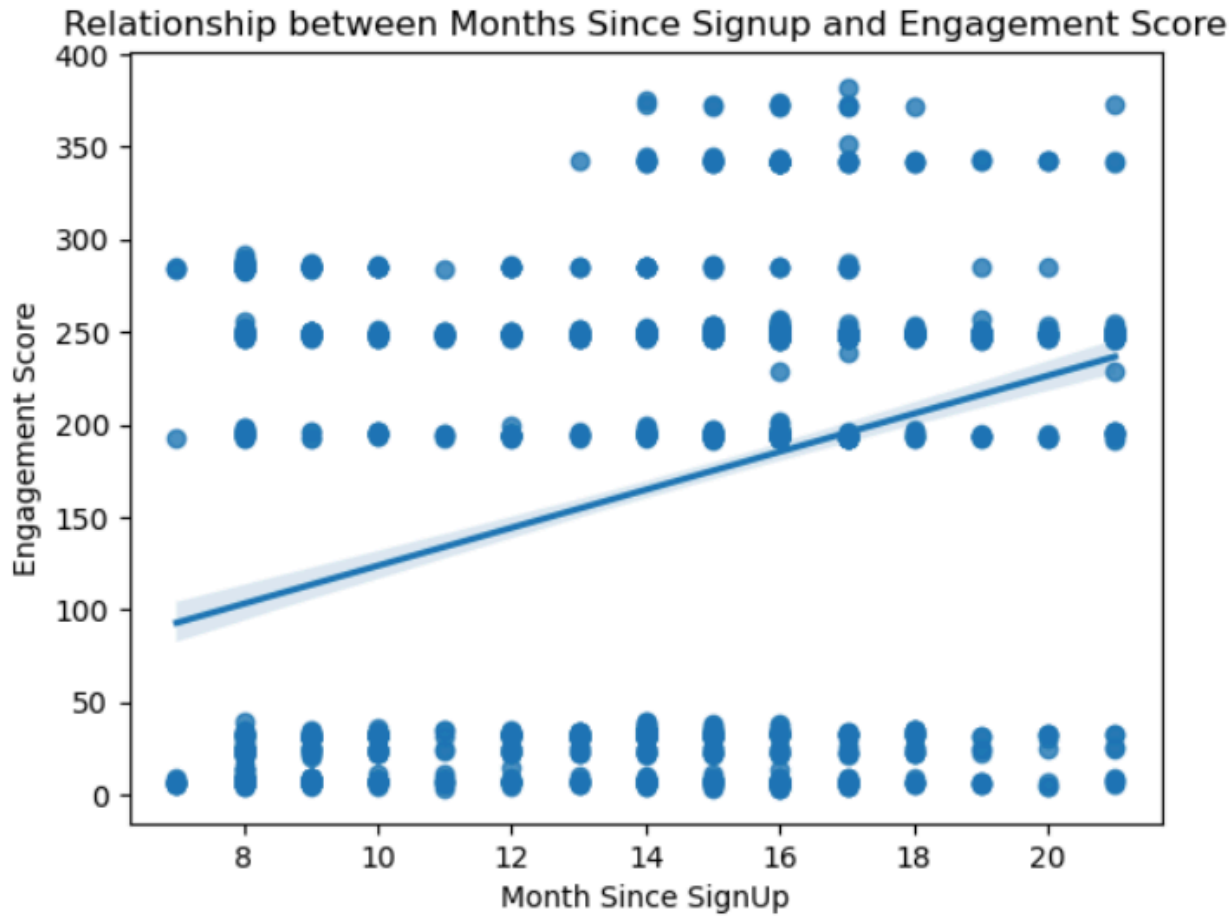
Hypothesis development and testing in EDA involves formulating educated guesses about the data and then using statistical tests to validate or refute these assumptions. This process begins by observing patterns and trends in the data, which leads to the formation of a specific question. The question is then translated into a null hypothesis, which assumes no significant difference or effect, and an alternative hypothesis, which contradicts the null hypothesis. The appropriate statistical test, such as a t-test, ANOVA, chi-square test, or correlation test, is selected based on the type of data and the nature of the hypothesis. A significance level, typically 0.05 or 0.01, is set to determine the threshold for rejecting the null hypothesis. The test statistic and p-value are calculated, and if the p-value is less than the significance level, the null hypothesis is rejected. This process allows for data-driven insights, informed decision-making, and effective model building.

4.1: Correlation Analysis:

Correlation analysis is a statistical method used to measure the strength and direction of the linear relationship between two variables. It helps us understand how changes in one variable are associated with changes in another. A correlation coefficient, typically represented by the symbol "r," ranges from -1 to +1, with -1 indicating a perfect negative correlation, +1 indicating a perfect positive correlation, and 0 indicating no linear relationship.

Hypothesis: Engagement scores decline as the time since signup increases.

Correlation between Months Since Signup and Engagement Score: 0.33329541418600667



The graph and the provided correlation coefficient contradict the hypothesis. The graph shows a positive correlation, indicating that engagement scores tend to increase as the time since signup increases, not decline.

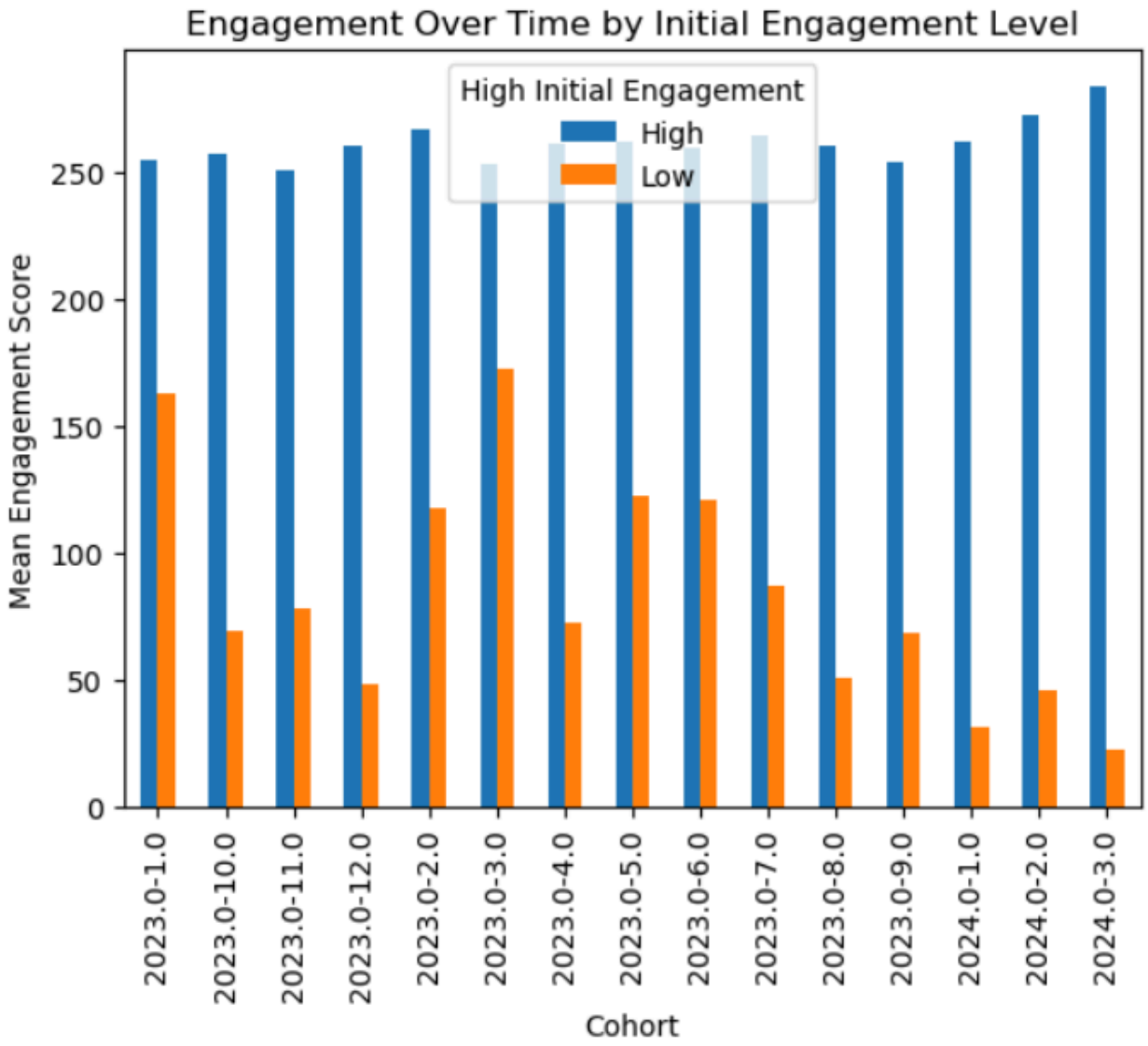
Therefore, the data does not support the hypothesis.

4.2: Cohort Analysis:

Cohort analysis is a technique used to analyze customer behavior over time. It involves grouping customers based on a shared characteristic, such as sign-up date or purchase date, and tracking their behavior over time. By comparing the behavior of different cohorts,

businesses can identify trends, patterns, and opportunities for improvement. This analysis helps in understanding customer lifecycle, retention rates, and the impact of marketing campaigns.

Hypothesis: Students who are highly engaged in the first 30 days are more likely to remain active long-term.

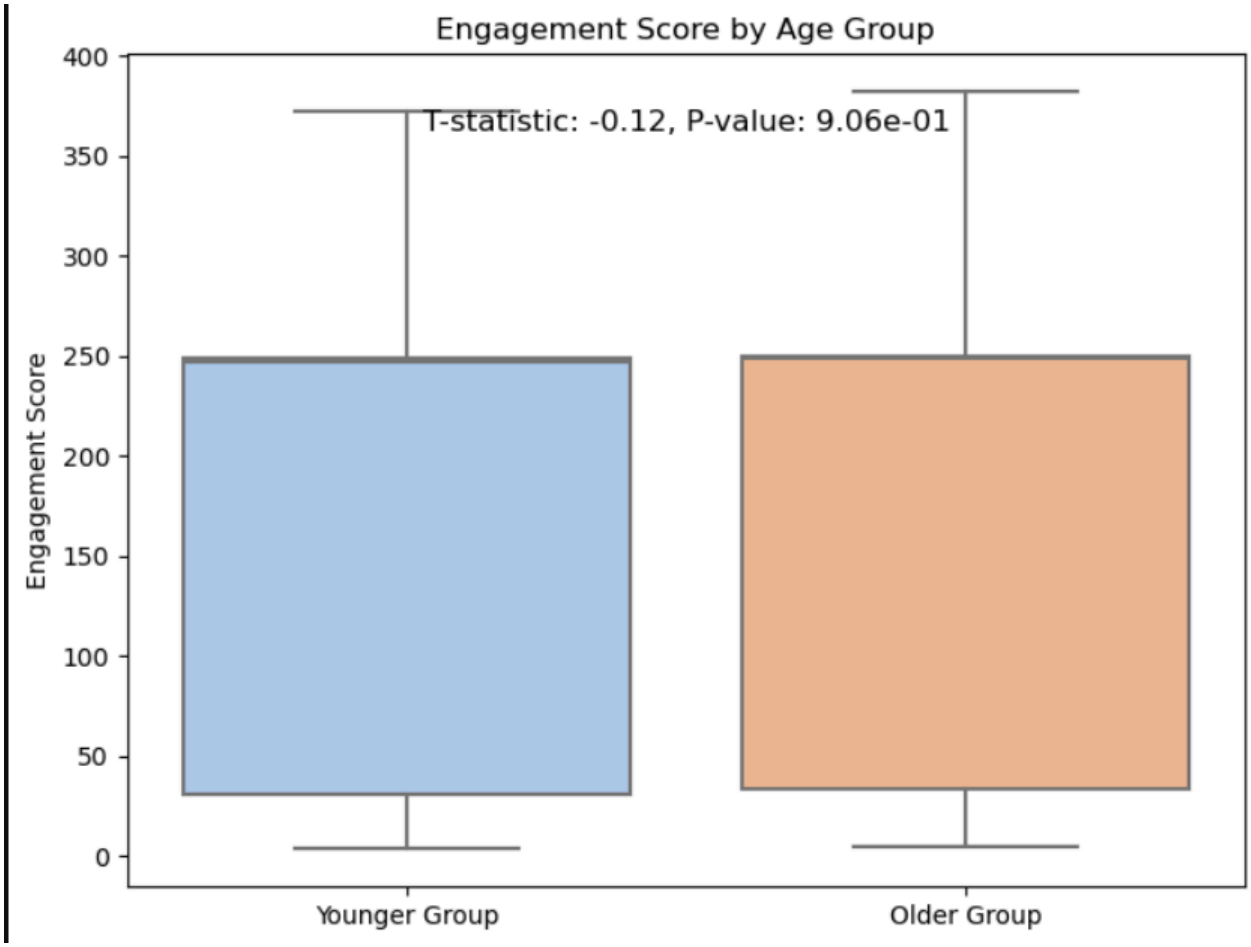


The graph provides strong visual evidence that students who are highly engaged in the first 30 days are more likely to maintain their engagement levels in the long run. Therefore, it does support the hypothesis.

4.3: T-Test for Age Groups:

A t-test is a statistical test used to determine if there is a significant difference between the means of two groups. It's commonly used to compare the average value of a variable between two groups, such as the average test scores of two different classes or the average sales of two different products.

Hypothesis: Older students have higher engagement scores than younger students.



T-statistic: -0.11820966627240234, P-value: 0.9059132724069179

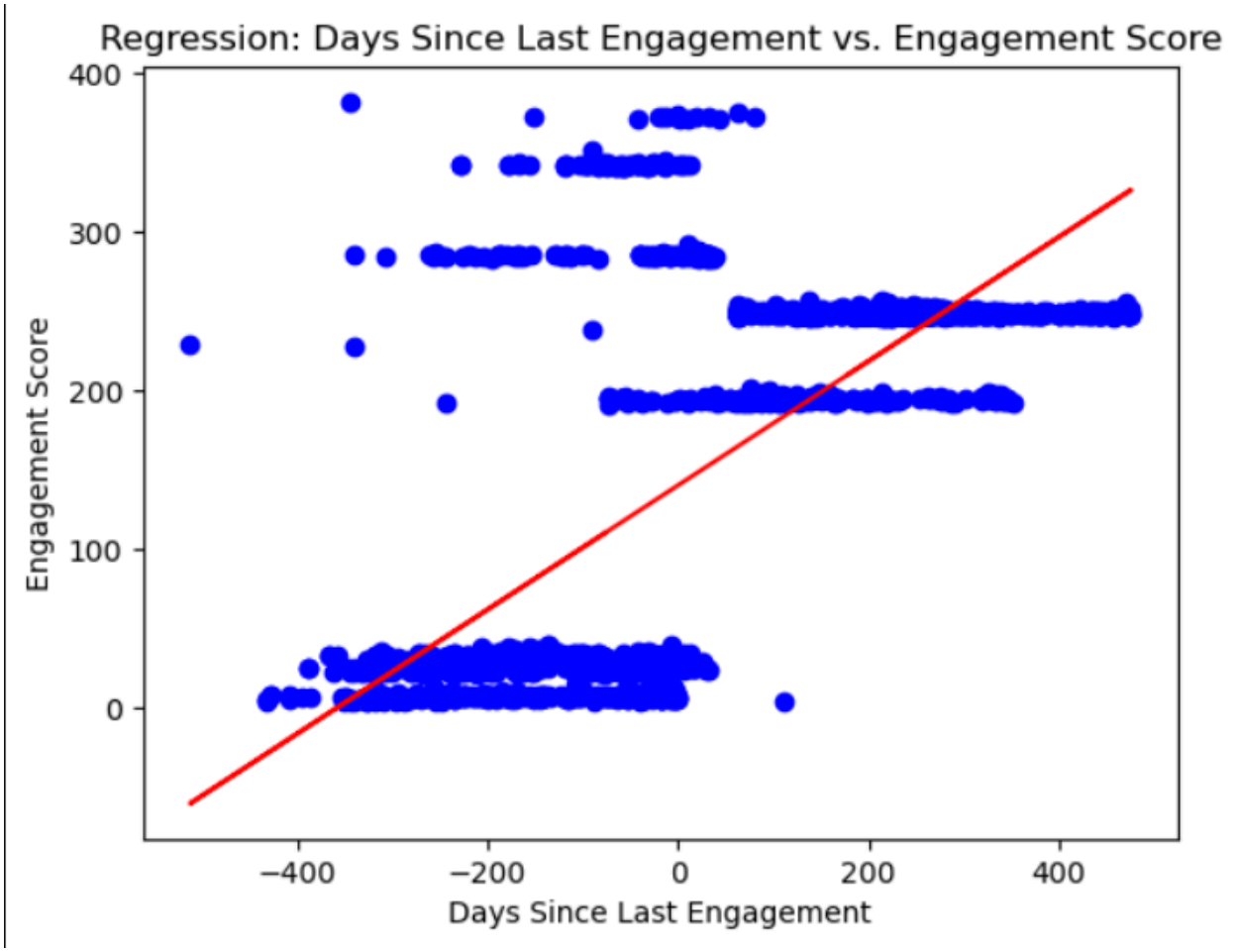
This means that there is no significant difference between the two age groups. Therefore, this does not support the hypothesis.

4.4: Regression Analysis:

Regression analysis is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It helps us understand how changes in the

independent variables affect the dependent variable. The most common type of regression analysis is linear regression, which models the relationship between variables using a straight line.

Hypothesis: Students who have engaged recently have higher overall engagement scores.

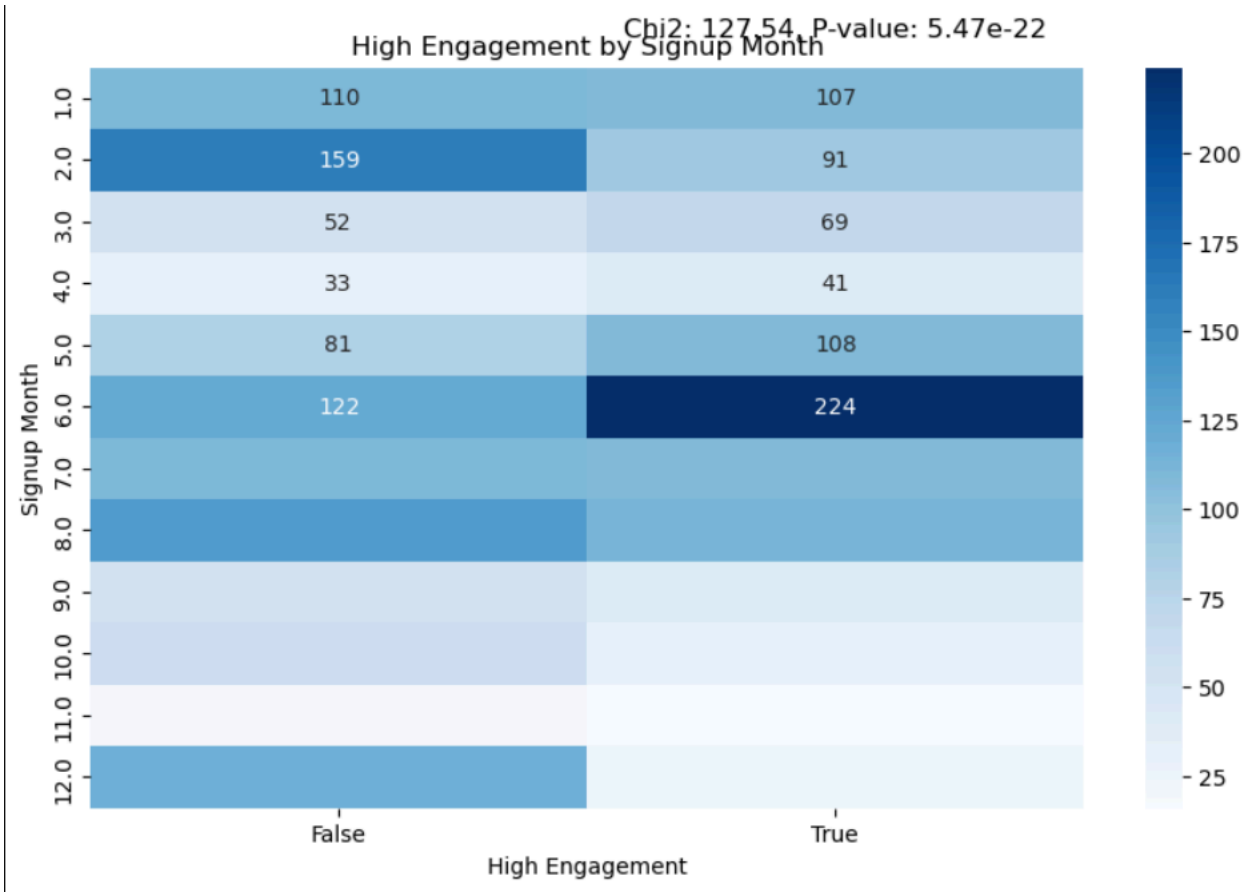


- The upward slope of the line indicates a positive correlation between the days since the last engagement and the overall engagement score.
- The positive slope suggests that as the number of days since the last engagement decreases (meaning more recent engagement), the overall engagement score tends to increase.
- The graph shows that students who have engaged more recently tend to have higher overall engagement scores. This aligns with the hypothesis.

4.5: Chi-Square Test for Categorical Data:

The chi-square test is a statistical test used to determine if there is a significant difference between the expected frequencies and the observed frequencies in one or more categories of a contingency table. It helps determine whether two categorical variables are independent or related to each other.

Hypothesis: Students who sign up early in the year have higher long-term engagement.



Chi2: 127.54243229284808, P-value: 5.46652228690605e-22

Heatmap:

- Each cell in the heatmap represents the count of users who signed up in a particular month and their level of engagement (high or low).
- The color intensity represents the count, with darker shades indicating higher counts.
- The x-axis represents the high engagement status (True or False), and the y-axis represents the signup month.

Chi-squared Test:

- Chi2: This is a statistical measure that quantifies the discrepancy between observed and expected frequencies in a contingency table. A higher Chi2 value indicates a greater difference.
- P-value: This is the probability of observing the data, assuming the null hypothesis (no association between signup month and high engagement) is true. A very low p-value (like 5.47e-22) indicates strong evidence against the null hypothesis.

Overall:

The combination of the heatmap and the Chi-squared test results suggests that the month of signup significantly influences the likelihood of high engagement. Specifically, there are certain months where a higher proportion of users tend to become highly engaged compared to others. Therefore, the test aligns with the hypothesis.

5. Conclusion:

The exploratory data analysis (EDA) of the learner engagement dataset revealed several key insights to guide strategic improvements in program design and learner retention efforts. Such as,

Engagement Patterns: The data highlighted variability in engagement duration and scores across opportunity types and demographic groups. Factors like age, opportunity category, and timing of sign-up seem to influence high engagement. These insights suggest that targeting recruitment and retention efforts based on these characteristics could lead to more consistent engagement.

Predictive Insights and Retention: Indicators like Days Since Last Engagement provide valuable churn predictors, allowing proactive intervention to retain at-risk learners. Personalizing program recommendations based on demographic and engagement insights could further boost participation rates and satisfaction.

In conclusion, this EDA provides actionable information for engagement patterns, establishing the framework for data-driven initiatives to improve engagement and learner outcomes. In Week 3, we will focus on constructing predictive models to improve our understanding of engagement factors and churn risks. This step will entail using important factors from the EDA to create

models that may predict engagement scores, likelihood of disengagement, and possible high-risk learners. By doing so, we want to generate practical, data-driven suggestions to improve learner retention and program performance. This model-building phase will enable us to transform EDA findings into practical applications, allowing for targeted interventions to enhance results across a wide range of learner characteristics.