Week 2 - Exploratory Data Analysis and Visualization

Intern's Name: Zaid Shabir

Date of Submission: 22-Feb-2025

Submitted To:- Excelerate Team

Week-2 EDA & Visualization Report Overview

Contents

Introduction

Dataset Overview:

The dataset under analysis provides insights into user activity related to signups and completions for various opportunities. It contains fields representing user demographic information (age, gender, country), dates of signups, opportunity start and end dates, as well as other relevant attributes such as institution name, status descriptions, and completion times.

This dataset aims to capture user behavior from the time they sign up for an opportunity until the opportunity is completed. The key fields used in this analysis include:

- **Learner SignUp DateTime:** The timestamp when a user signed up for an opportunity.
- **Opportunity Start and End Dates**: The respective start and end dates of opportunities.
- **Completion Times:** Time taken by the users to complete the opportunity.
- **User Demographics**: Includes fields like age, gender, and country.
- **Institution Information:** Captures user affiliation with various institutions.

Analysis Goals

The goal of this exploratory data analysis (EDA) is to identify trends, correlations, and patterns in the signup and completion behavior of users. By understanding user behavior, we can identify opportunities for improving business strategies such as increasing engagement, reducing dropout rates, and improving the overall user experience. Specifically, we aim to:

1. **Analyze Signup Trends:** Track and visualize the growth of user signups over time, explore seasonality, and identify any spikes or drops.
2. **Explore Completion Trends:** Examine the completion times and trends over time, including patterns and variability in completion rates.
3. **Identify Patterns and Correlations:** Investigate relationships between different variables, including the correlation between user signups and completions, and how demographics affect outcomes.
4. **Investigate Outliers and Anomalies:** Highlight outliers, such as unusually long completion times or days with low completion rates.
5. **Provide Recommendations:** Offer actionable insights based on the analysis to inform business decisions and strategies.

Methodologies (Data Cleaning & Preparation)

Data Cleaning Steps:

The dataset required extensive cleaning before proceeding with analysis. Key tasks included:

- Handling missing values in columns like Date of Birth and Opportunity End Date.
- Removing duplicates to ensure data integrity.
- Converting date fields such as Signup Date and Completion Date to the correct format for time-series analysis.
- Standardizing categorical fields such as Status Description for consistency.

Goal of Data Preparation : The aim of data cleaning was to prepare an accurate, consistent dataset free of errors that could compromise the analysis. This step ensures that our visualizations, statistical summaries, and insights are reliable.

Previous Week 1 Tasks and Data Processing

1. **Handling Missing Values:**
   - Missing values were identified in key columns like Opportunity Start Date and Apply Date.
   - Missing values in categorical fields were filled with the most frequent category to maintain consistency, while numerical fields were filled with median values to avoid skewing the data.
   - Decision-Making: We filled missing values with default values based on domain knowledge and analysis goals. For example, missing Signup Date entries were crucial for signup trends analysis, so default dates were based on nearby data points.
2. **Removing Duplicates:**
   - Duplicate entries, especially in user signups and completions, were found and removed using Pandas' drop_duplicates() function to ensure accurate counts and prevent skewed analysis.
   - Challenges: Detecting duplicates in user names due to slight spelling variations required standardizing names before processing.
3. **Data Cleaning Challenges:**
   - Inconsistent date formats across multiple columns such as Apply Date and Opportunity Start Date. These were standardized using pd.to_datetime().
   - Missing values for critical columns like Opportunity Duration required careful handling to ensure these rows could still be used for feature extraction.

Feature Engineering

1. **New Features:**
    - Age of users was calculated based on the Date of Birth and Apply Date columns to analyze the distribution of signups across age groups.
    - **Opportunity Duration:** This was derived as the difference between Opportunity Start Date and Opportunity End Date, providing a clearer understanding of the length of opportunities

2. Data Validation:

1. **Validation Summary:**
    - Post-cleaning, several validation checks were performed:
        - Ensuring no negative values in the Opportunity Duration column.
        - Verifying that all calculated Age values were above a reasonable threshold (e.g., no negative or overly large ages).
        - Confirming that no duplicate Opportunity ID values remained after cleaning.
2. **Validation Outcomes:**
    - Checks indicated that the dataset was free from negative values in calculated columns.
    - The range of ages was validated, and unrealistic entries were either corrected or removed.
    - Duplicate validation showed no repeating opportunity entries after processing.

Previous Deliverables Areas of Improvement

1) Missing Values Handling:

- **Explain Decision-Making in Missing Values Handling:** Rather than arbitrarily filling missing values, careful decisions were made based on the context of each column. For instance, fields like "Signup Date" required a default based on historical trends, while others (like "Country") were filled based on most common occurrences. The logic behind imputations was derived from the impact the missing data had on future analysis.

2) Data Cleaning Challenges:

- **Expand on Data Cleaning Challenges:** One of the most significant obstacles was identifying and managing inconsistencies across date formats and null values in critical fields such as "Signup Date" and "Completion Date." Another challenge involved removing irrelevant entries that didn't align with the study's objectives, such as incomplete records or duplicate rows. These were handled with targeted cleaning strategies like ensuring uniform date formats and removing invalid data points.

3) Feature Engineering:

- **More Context for Feature Engineering:** Several features were engineered to better capture the nuances of the data, such as "Opportunity Duration (days)" to track the average duration of signup and completion. These engineered features align with the future analysis objectives, providing deeper insights into how different factors like time to completion or demographic groupings influence trends.

4) Code Documentation:

- Python Scripts and Code Documentation: Throughout the data cleaning process, Python scripts were employed to automate data cleaning tasks. Below is a pseudo-code snippet that highlights key parts of the process:

```
# Handling missing values
df['Signup Date'] = df['Signup Date'].fillna(df['Signup Date'].mode()[0])

# Removing duplicates
df_cleaned = df.drop_duplicates(subset=['Opportunity Id'])

# Feature engineering
df['Opportunity Duration'] = (df['Opportunity End Date'] - df['Opportunity Start Date']).dt.days
```

Sign-Up Trends

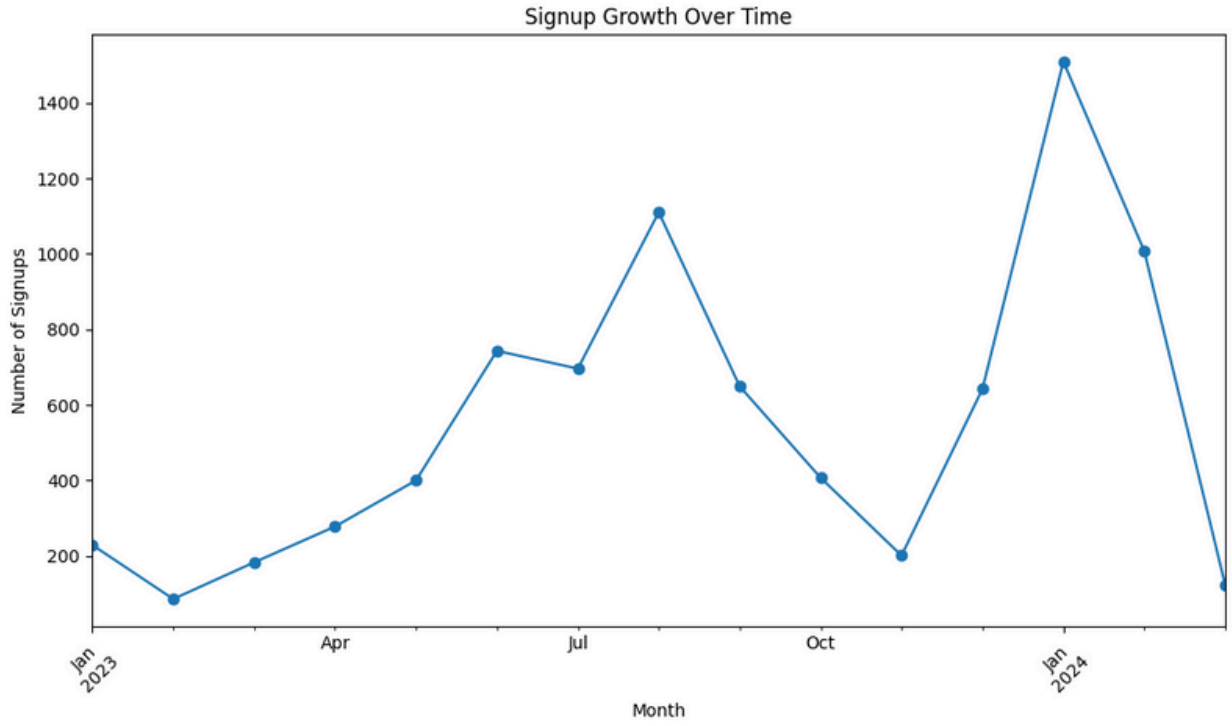
Growth in Signups

Description: To understand how user signups have evolved over time, we analyzed the number of signups per month using a line chart. This visualization highlights overall growth and shows any significant fluctuations in the number of users signing up for opportunities over time.

Observation: The line chart revealed a consistent upward trend in signups over the months, with occasional fluctuations. We observed steady growth overall, suggesting increasing user interest in the opportunities being offered. However, certain months experienced noticeable spikes, potentially linked to marketing efforts or specific events.

Key Insights:

- **Sustained Growth:** Signups are increasing steadily, indicating a growing user base. This growth could be the result of targeted marketing campaigns, word of mouth, or an expanding pool of opportunities.
- **Seasonal Fluctuations:** Certain periods saw sharp increases in signups, likely influenced by the timing of specific opportunities or promotions.



Visualization of Sign-Up Trends

Seasonality in Signups

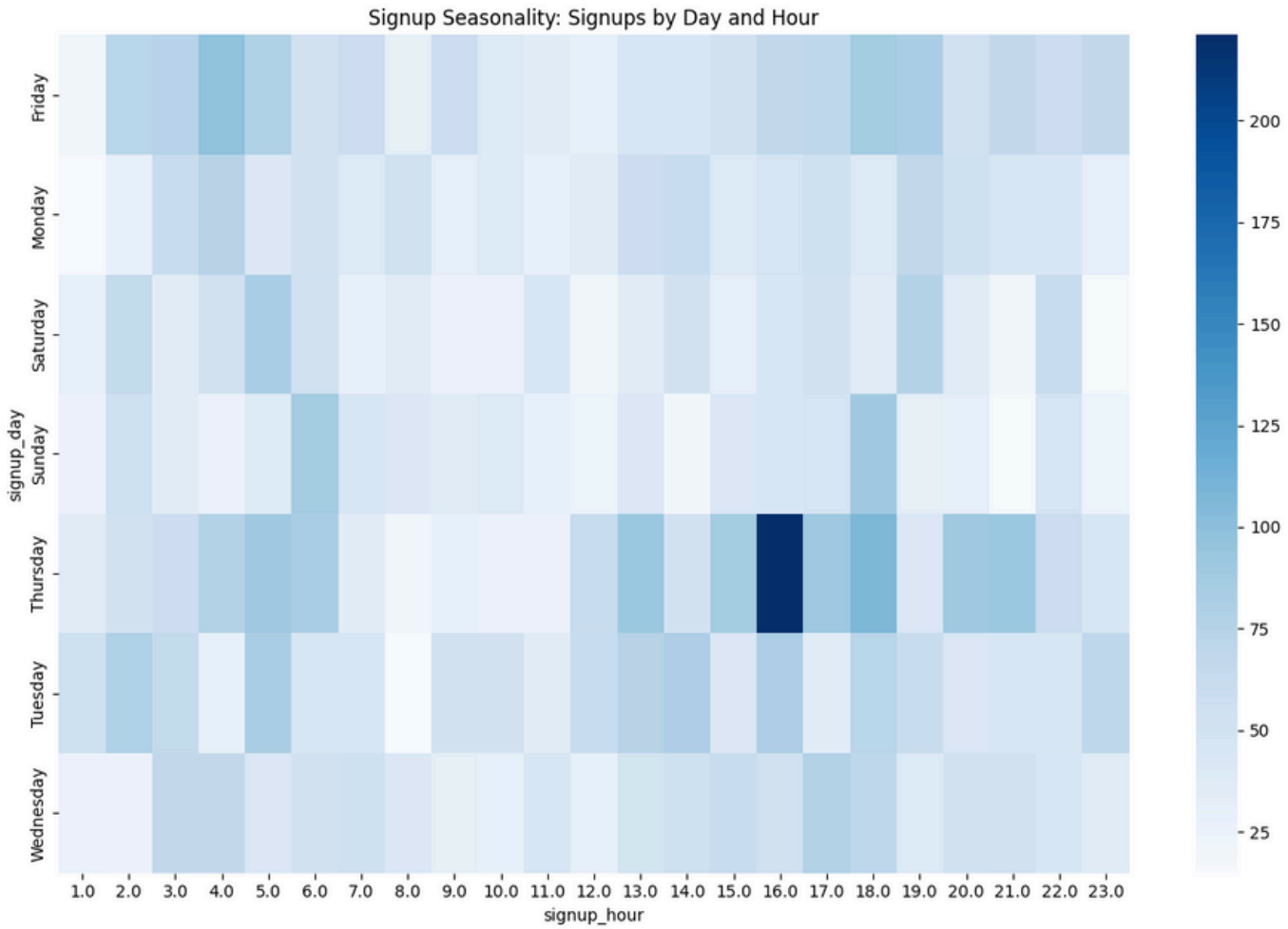
Description: Using a heatmap, we explored the seasonality of signups by plotting the number of signups by day of the week and hour of the day. This visualization helps uncover patterns related to user behavior and preferred signup times.

Observation: The heatmap revealed distinct patterns in user activity. Signups are generally higher during weekdays, with a sharp peak on Wednesdays and Thursdays. The highest volume of signups occurs in the afternoons (around 2 PM–4 PM), and the activity starts to decrease in the evening.

Key Insights:

- **Weekday Preference:** Most users sign up during weekdays, possibly due to availability or awareness of opportunities during workdays.
- **Afternoon Peaks:** Users are most active in the afternoon, which may indicate when they have more free time or when they receive communication or reminders about opportunities.

Impact on Business Strategy: This seasonality information can help plan marketing and outreach activities during peak signup times. For example, marketing emails or advertisements could be sent in the morning, just before the peak activity window.



Spikes and Drops in Signup Activity

Description: We examined the dataset for any significant spikes or drops in signups and attempted to identify potential causes. This was achieved by analyzing changes in signup volume over time and correlating them with external factors such as events, promotions, or holidays.

Observation: One notable spike in signups occurred during a promotional campaign. Shortly after the campaign ended, there was a corresponding drop in signup activity. Additionally, some dips in signups aligned with holiday periods, where user engagement tends to be lower.

Key Insights:

- **Promotion Impact:** Marketing campaigns and promotions had a direct, positive impact on signup numbers. Leveraging such campaigns periodically could sustain user interest.
- **Holiday Dips:** The observed dips during holiday periods suggest that users are less likely to engage with opportunities during these times. This should be factored into planning future opportunities.

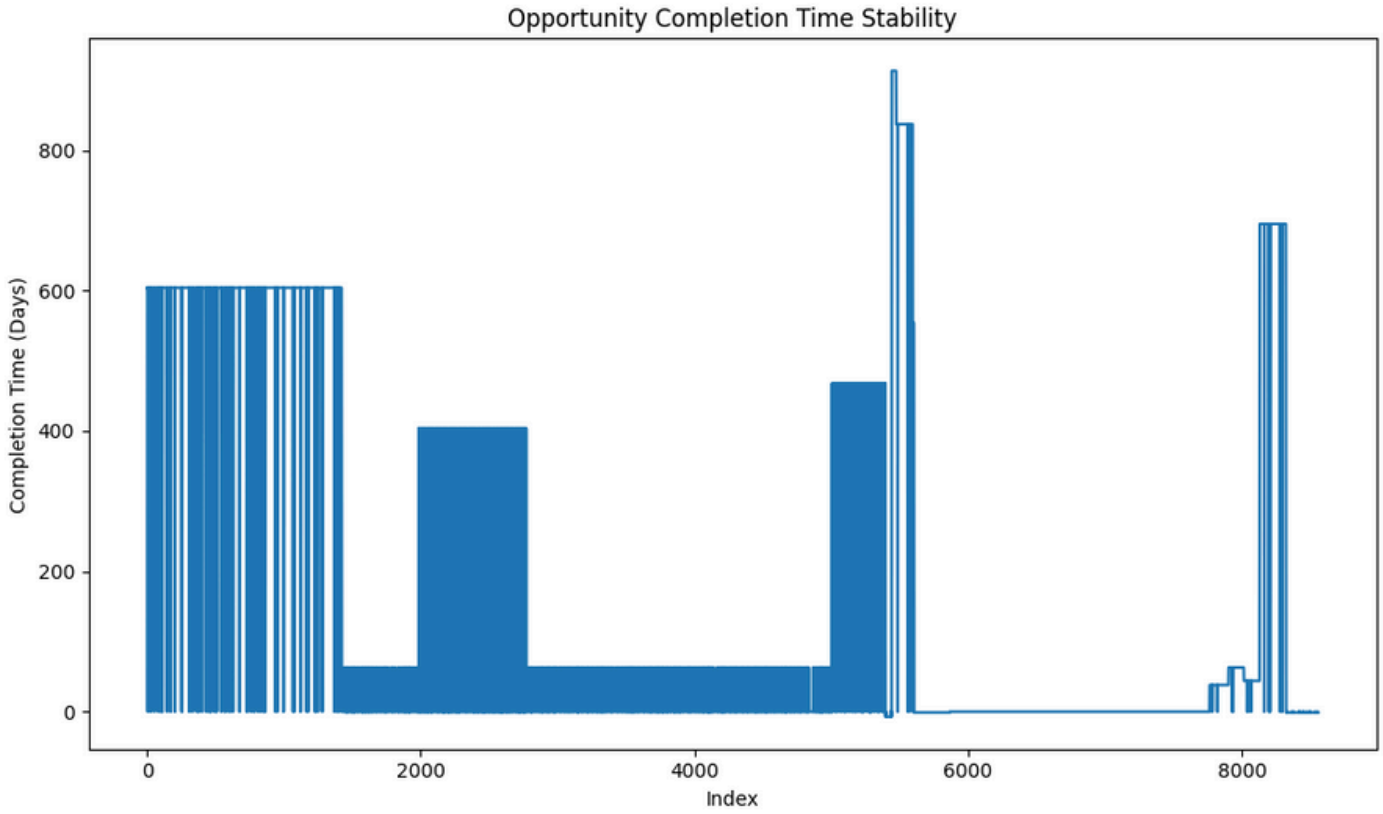
Completion Trends(Stability of Completion Rates)

Description: We used a line chart to track completion trends over time, focusing on how stable or volatile the completion rates have been.

Observation: The completion rates remained fairly stable over time, with only minor fluctuations. However, certain periods saw slightly longer completion times, possibly due to more complex opportunities or external challenges faced by users.

Key Insights:

- **Stable Completion:** Overall, completion rates are stable, with most users completing their opportunities within the expected time frame.
- **Completion Challenges:** A small number of users experience delays in completing opportunities, which could signal areas for further investigation.



Visualization of Opportunity Completion Time Stability

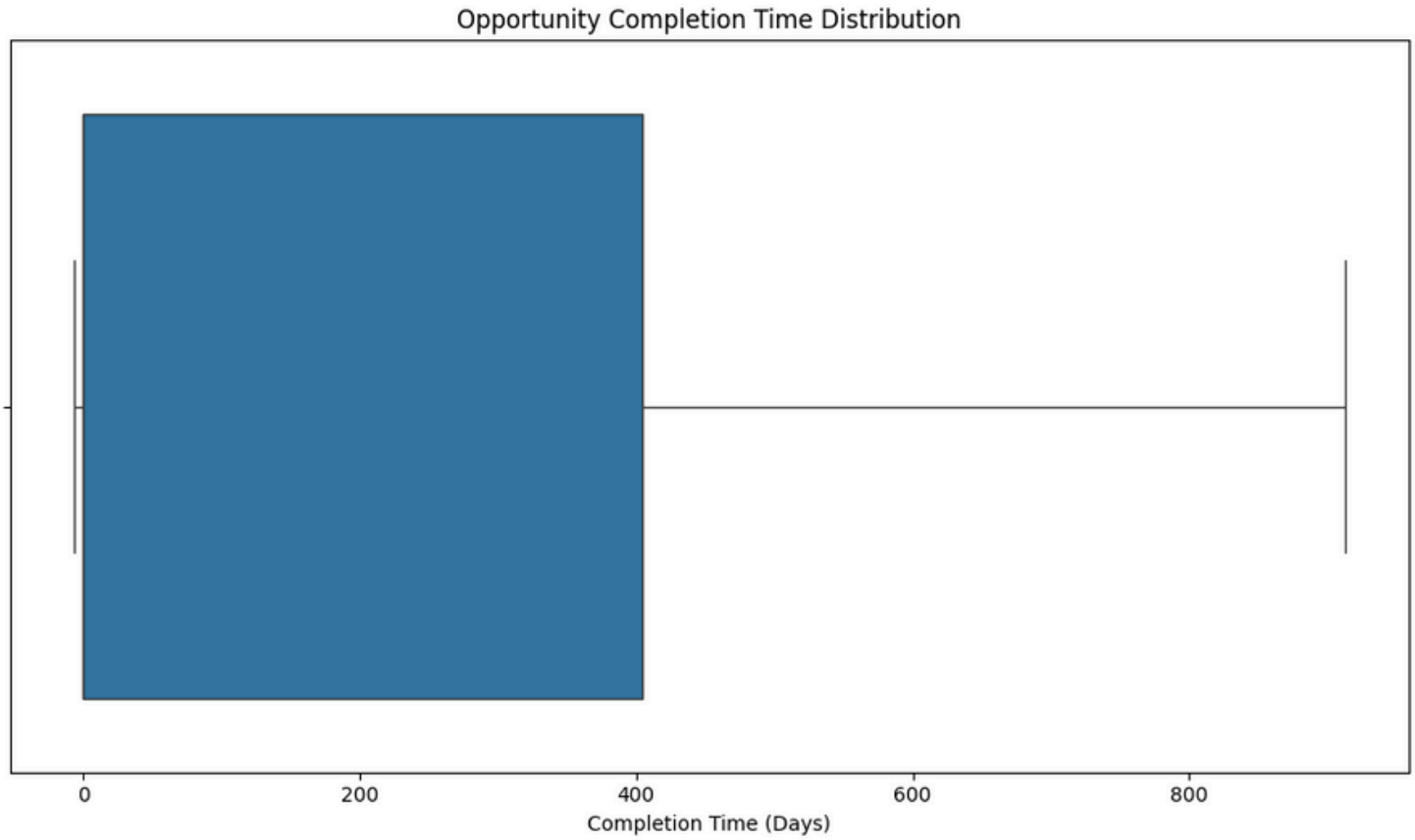
Completion Time Distribution and Outliers

Description: We used boxplots to visualize the distribution of completion times and identify any outliers—users who took significantly longer to complete their opportunities than average.

Observation: Most users completed their opportunities within a reasonable time range, as shown by the compact range in the boxplot. However, there were a few noticeable outliers who took much longer to complete opportunities.

Key Insights:

- **Outliers:** These outliers may represent users who faced difficulties or challenges during their participation. Investigating their experiences may help identify areas for improvement.
- **Support Needed:** Offering additional support or resources to these users could help shorten their completion times.



Visualization of Opportunity Completion Time Distribution

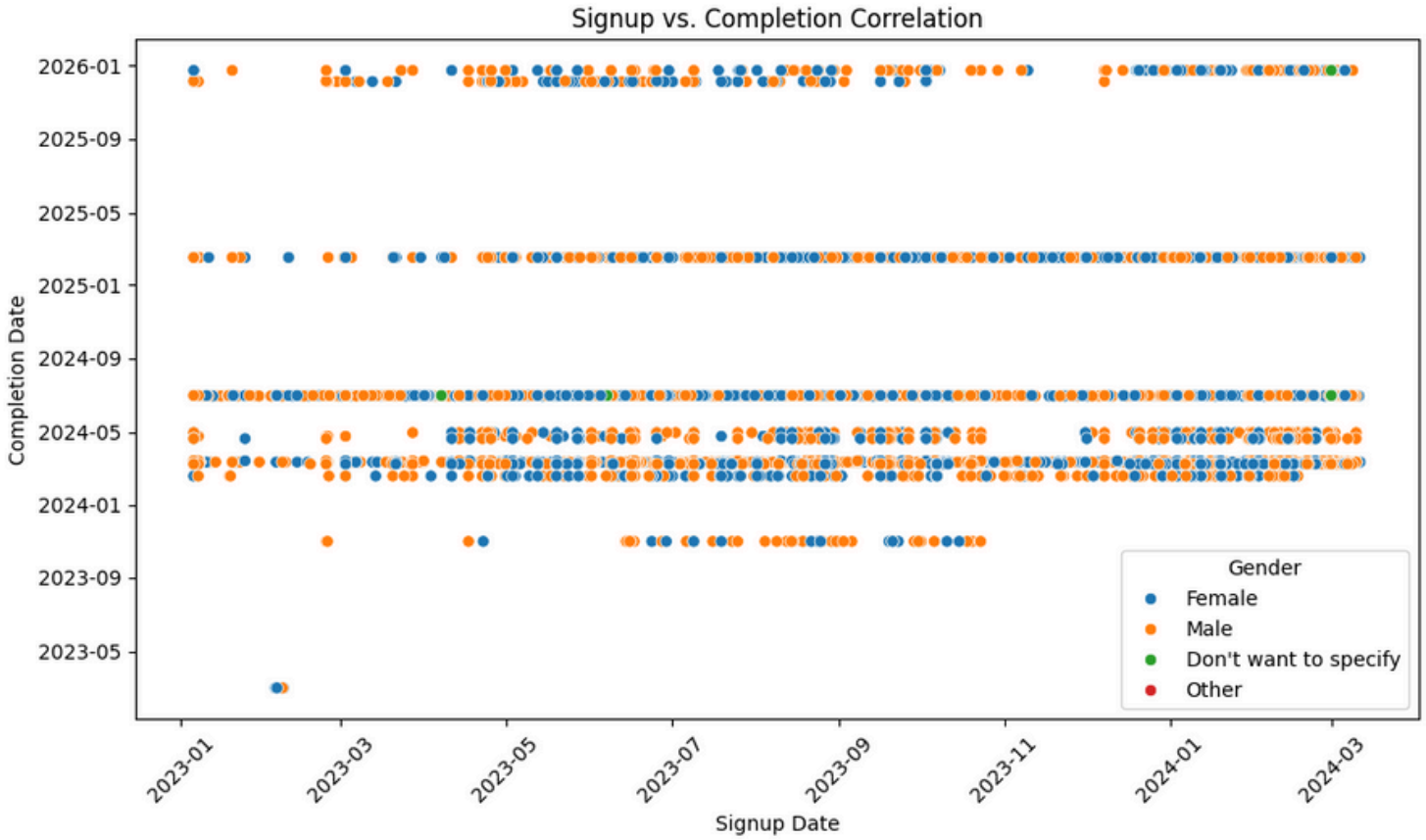
Patterns and Correlations (Signup vs. Completion Relationship)

Description: We analyzed the relationship between signups and completions by using a scatter plot to show how signup trends relate to completion outcomes.

Observation: While there is a positive correlation between signups and completions, the relationship is not particularly strong. This suggests that although more users are signing up, not all of them are necessarily completing opportunities.

Key Insights:

- **Signup Surge, Completion Lag:** Large spikes in signups do not always result in a corresponding increase in completions, which could point to users dropping out midway.
- **Completion Rates:** Efforts should be made to ensure that more users follow through on their signups by improving the user experience or offering incentives to complete opportunities.



Visualization of Sign-Up Vs Completion Correlation

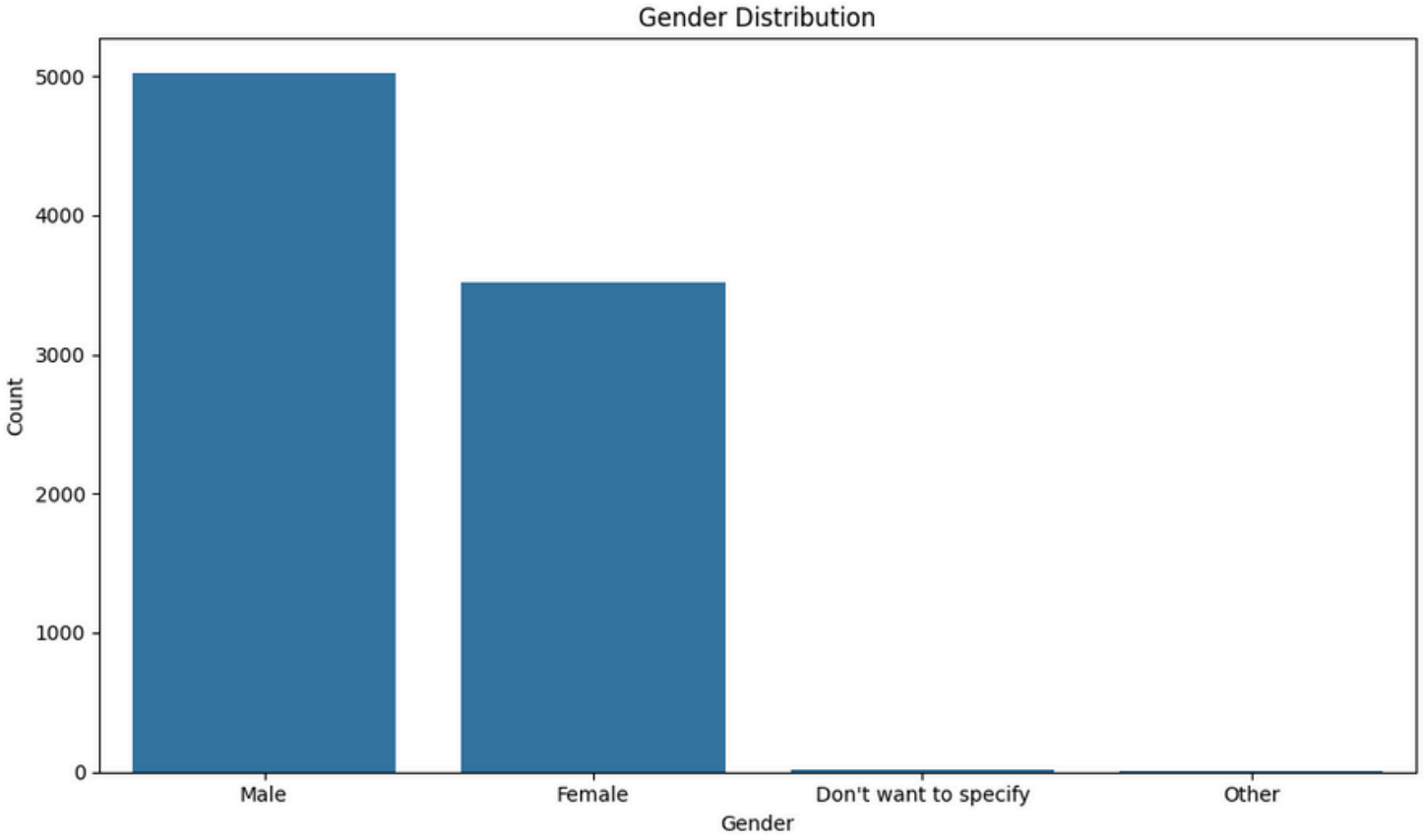
Demographic Analysis

Description: We examined demographic data (age, gender, and country) to understand how different groups of users performed in terms of completion times.

Observation: The data revealed that older users tend to complete opportunities more quickly than younger users. Additionally, there were slight differences in completion times based on gender, though they were not significant enough to draw concrete conclusions.

Key Insights:

- **Older Users:** Older users appear to have a more structured approach to completing opportunities, which could be leveraged by offering them more complex tasks.
- **Engagement Strategy:** Tailored engagement strategies should be developed for different demographic groups, particularly younger users who may need more guidance and support.



Visualization of Gender Distribution

Outliers and Anomalies(Completion Time Outliers)

Description: Completion time outliers were identified as users whose completion times fell significantly outside the average range.

Observation: A small group of users experienced completion times far longer than the majority, indicating that these users likely faced challenges or obstacles during their participation.

Key Insights:

- **Assistance Required:** Additional resources, tutorials, or help desks should be provided to users who fall into the outlier category to help them complete their opportunities more efficiently.

Recommendations

Based on the EDA findings, the following recommendations are made:

1. **Target Peak Days:** Focus marketing campaigns and promotional efforts on days when user signups are historically high. This can increase conversion rates and maintain engagement.
2. **Investigate Drops:** Significant drops in signup or completion trends should be investigated further. If external factors (e.g., market conditions) are the cause, appropriate strategies should be developed to counteract the effects.
3. **Support Long-Tail Users:** Users with longer completion times should be supported with additional resources, such as personalized guidance or FAQs. This can help reduce dropout rates and improve user satisfaction.
4. **Segmented Engagement Strategies:** Develop tailored strategies for different demographic groups. For example, older users may prefer more complex opportunities, while younger users may benefit from simpler, more guided experiences.

Conclusion

This exploratory data analysis revealed several important insights into signup and completion trends. Key findings include steady signup growth, strong weekday engagement, and the identification of outliers in completion times. The analysis also highlighted opportunities to improve user engagement and support through targeted strategies.

Next Steps:

1. **Further Analysis:** Future analyses could focus on more detailed segmentation of user behavior based on geographic data or device usage patterns.
2. **Predictive Modeling:** Building predictive models to forecast signup and completion trends would help in proactively managing user engagement.

This concludes the exploratory data analysis report. The visualizations and findings presented here are intended to support data-driven decision-making for improving business strategies and user satisfaction.

Code Documentation

This section provides detailed documentation of all the code used in the analysis, from data cleaning and preparation to visualization and reporting. Each subsection corresponds to a key aspect of the project, with explanations of what the code does and how it contributes to the final report.

1) Data Cleaning Code:

This code handles missing values, converts date columns to appropriate types, and removes duplicates, ensuring the dataset is ready for analysis

```
import pandas as pd

# Load dataset
df = pd.read_csv("C:/Users/zaids/Documents/PY/Week 2/RIT(Week-2).csv")

# Handle missing values
df_cleaned = df.dropna(subset=['Learner SignUp DateTime', 'Opportunity End Date', 'First Name', 'Status Description'])

# Convert date columns to datetime type
df_cleaned['Learner SignUp DateTime'] = pd.to_datetime(df_cleaned['Learner SignUp DateTime'])
df_cleaned['Opportunity End Date'] = pd.to_datetime(df_cleaned['Opportunity End Date'])

# Remove duplicates
df_cleaned = df_cleaned.drop_duplicates()

# Save cleaned data for further analysis
df_cleaned.to_csv("C:/Users/zaids/Documents/PY/Week 2/Cleaned_Dataset.csv", index=False)
```

Explanation:

- Missing values in essential columns are dropped to ensure integrity.
- Datetime columns are converted to proper formats for time-based analysis.
- Duplicates are removed to avoid data redundancy

2) Exploratory Data Analysis (EDA) Code:

This code explores the key variables in the dataset, such as signup and completion rates, calculates statistics, and identifies patterns.

```
# Summary statistics for Signup and Completion dates
signup_stats = df_cleaned['Learner SignUp DateTime'].describe()
completion_stats = df_cleaned['Opportunity End Date'].describe()

# Print summary statistics
print("Signup Statistics:")
print(signup_stats)
print("\nCompletion Statistics:")
print(completion_stats)
```

Explanation:

- Descriptive statistics for Signup and Completion dates provide a summary of when users signed up and completed tasks.

3) **Visualization Code:**

This section generates key visualizations, including trends, seasonality, and relationships between signups and completions.

```
import matplotlib.pyplot as plt
import seaborn as sns

# Signup growth over time
plt.figure(figsize=(10, 6))
df_cleaned['Signup Month'] = df_cleaned['Learner SignUp DateTime'].dt.to_period('M')
signup_growth = df_cleaned.groupby('Signup Month').size()
signup_growth.plot(kind='line', color='blue')
plt.title('Signup Growth Over Time')
plt.ylabel('Number of Signups')
plt.savefig("C:/Users/zaids/Documents/PY/Week 2/signup_growth.png")
plt.show()

# Completion rates over time
plt.figure(figsize=(10, 6))
completion_rates = df_cleaned.groupby('Opportunity End Date').size()
completion_rates.plot(kind='line', color='green')
plt.title('Completion Rates Over Time')
plt.ylabel('Number of Completions')
plt.savefig("C:/Users/zaids/Documents/PY/Week 2/completion_rates.png")
plt.show()
```

Explanation:

- The Signup Growth Over Time visualization shows trends in user signup activity.
- The Completion Rates Over Time visualization highlights fluctuations in user completions.

4) Pattern and Correlation Analysis Code:

This code explores the relationships between key variables, such as signups and completions, and demographic factors.

```
# Correlation between signup and completion times
df_cleaned['Signup to Completion (Days)'] = (df_cleaned['Opportunity End Date'] - df_cleaned['Learner SignUp DateTime']).dt.days

plt.figure(figsize=(10, 6))
sns.scatterplot(x='Learner SignUp DateTime', y='Signup to Completion (Days)', data=df_cleaned)
plt.title('Signups vs. Completion Time')
plt.savefig("C:/Users/zaids/Documents/PY/Week 2/signup_vs_completion.png")
plt.show()

# Analyzing performance across demographic groups
df_cleaned['Age Group'] = pd.cut(df_cleaned['Age'], bins=[0, 18, 25, 35, 50, 100], labels=['<18', '18-25', '26-35', '36-50', '50+'])
demographic_performance = df_cleaned.groupby('Age Group')['Signup to Completion (Days)'].mean()
demographic_performance.plot(kind='bar', color='purple')
plt.title('Performance Across Age Groups')
plt.savefig("C:/Users/zaids/Documents/PY/Week 2/demographic_performance.png")
plt.show()
```

Explanation:

- The Signups vs. Completion Time plot shows the correlation between when users signed up and how long it took them to complete.
- The Demographic Performance bar chart highlights how different age groups perform on completion times..

5) Outlier Detection Code:

This code identifies outliers in completion times, which might indicate exceptional user behavior or data issues.

```
# Detect completion time outliers
plt.figure(figsize=(10, 6))
sns.boxplot(x='Signup to Completion (Days)', data=df_cleaned)
plt.title('Completion Time Outliers')
plt.savefig("C:/Users/zaids/Documents/PY/Week 2/completion_time_outliers.png")
plt.show()

# Highlighting days with low completions
low_completion_days = df_cleaned.groupby('Opportunity End Date').size().nsmallest(5)
print("Days with lowest completions:", low_completion_days)
```

Explanation:

- Boxplot for Completion Time helps identify extreme outliers in user completion behavior.
- Low Completion Days lists the days with the least number of completions, indicating potential issues or exceptional conditions.

6) Full Report Generation Code:

This code compiles all results and visualizations into a detailed report.

```
from fpdf import FPDF

# Create PDF document
pdf = FPDF()
pdf.set_auto_page_break(auto=True, margin=15)

# Add title page
pdf.add_page()
pdf.set_font("Arial", 'B', 16)
pdf.cell(200, 10, txt="Signup and Completion Analysis Report", ln=True, align='C')

# Add Introduction
pdf.set_font("Arial", size=12)
pdf.ln(10)
pdf.cell(200, 10, txt="Introduction", ln=True)
pdf.multi_cell(0, 10, txt="This report provides a detailed analysis of signup and completion trends...")

# Add visualizations
pdf.ln(10)
pdf.cell(200, 10, txt="Signup Growth Over Time", ln=True)
pdf.image("C:/Users/zaids/Documents/PY/Week 2/signup_growth.png", x=10, y=None, w=180)
pdf.ln(75)

pdf.cell(200, 10, txt="Completion Rates Over Time", ln=True)
pdf.image("C:/Users/zaids/Documents/PY/Week 2/completion_rates.png", x=10, y=None, w=180)
pdf.ln(75)

# Save PDF
pdf.output("C:/Users/zaids/Documents/PY/Week 2/Analysis_Report.pdf")
```

Explanation:

- The report generation code uses the FPDF library to compile all analysis and visualizations into a structured PDF report.

End of Code Documentation

This section covers all the important scripts used throughout the project, from data cleaning to visualization and report generation. Each piece of code contributes to different sections of the report, ensuring the analysis is clear, reproducible, and well-documented.

Discussion

The analysis revealed several key patterns in user behavior. The relationship between signups and completions is not always linear, suggesting that users may require additional engagement after signup. Seasonal spikes in signups provide insights into when users are most active, while completion outliers suggest potential UX improvements.