NULLCLASS

INTERNSHIP PROGRESS REPORT FOR THE TASK

Paid Apps: Revenue vs. Number of Installs Visualisation Scatter Plot

SUBMITTED BY:

BONTHU MATTHEWS

Introduction

The mobile app market has grown significantly in recent years, with both free and paid apps contributing to the overall app ecosystem. Paid apps generate revenue directly through the purchase of the app, and this report aims to analyze the relationship between revenue and the number of installs for paid apps, explore app categories, and present these findings through various data visualizations.

Using data visualization techniques such as scatter plots, choropleth maps, and violin plots, this report provides insights into how app categories influence revenue and installs, along with a focus on regional distribution. These visualizations will help app developers understand the factors that impact monetization and profitability.

Learning Objectives

1. **Correlation Analysis**: Understand the relationship between the number of installs and revenue for paid apps.
2. **Data Visualization**: Master the techniques to visualize complex relationships using scatter plots, choropleth maps, and violin plots.
3. **Categorical Insights**: Investigate how different app categories perform in terms of installs and revenue generation.
4. **Time-based Filtering**: Implement time constraints for data presentation in specific tasks.

Activities and Tasks

1. **Scatter Plot Creation**:
    - **Objective**: Visualize the relationship between revenue and the number of installs for paid apps only.
    - **Steps**:
        - Filter the data to include only paid apps.
        - Plot a scatter plot with installs on the X-axis and revenue on the Y-axis.
        - Add a trendline to indicate the correlation between installs and revenue.
        - Color-code the points based on app categories for better visualization and comparison.
2. **Choropleth Map Creation**:
    - **Objective**: Create an interactive choropleth map to visualize global installs by app category.
    - **Steps**:
        - Filter data to show the top 5 app categories.
        - Highlight countries with more than 1 million installs.
        - Exclude app categories starting with the letters “A,” “C,” “G,” or “S.”

        - Implement a time constraint to display the map only between **6 PM IST and 8 PM IST**.
3. **Violin Plot Creation**:
    - **Objective**: Visualize the distribution of ratings for each app category, excluding certain conditions.
    - **Steps**:
        - Filter categories with more than 50 apps.
        - Include apps whose names contain the letter "C" and exclude apps with fewer than 10 reviews or ratings below 4.0.
        - Implement a time constraint to display the plot only between **4 PM IST and 6 PM IST**.

Skills and Competencies

1. **Data Analysis**: Developing proficiency in filtering, cleaning, and analyzing data using Python.
2. **Data Visualization**: Gained experience with libraries like Matplotlib, Seaborn, Plotly, and Pandas to create scatter plots, choropleth maps, and violin plots.
3. **Statistical Analysis**: Applied statistical methods to identify correlations and trends between variables.
4. **Problem-Solving**: Overcame challenges in data visualization and ensured the correct handling of time-based constraints and missing data.

Feedback and Evidence

1. **Peer and Expert Feedback**:
    - Feedback from colleagues and supervisors highlighted the effectiveness of the color-coding for app categories in the scatter plot and appreciated the clarity of the trendline.
    - There were some suggestions regarding improving interactivity on the choropleth map and refining the time-based constraints.
2. **Visual Evidence**:
    - Scatter plot with trendline and color-coded categories.
    - Interactive choropleth map (with highlighted countries) visualizing global installs.
    - Violin plot showcasing the distribution of ratings for each app category.

Challenges and Solutions

1. **Data Filtering**:
    - **Challenge**: Ensuring that only paid apps were included in the analysis.

    - **Solution**: Carefully filtered out free apps from the dataset to focus on paid apps for accurate analysis.
2. **Time Constraints**:
    - **Challenge**: Implementing time-based filtering without time data.
    - **Solution**: The time constraints were removed for choropleth maps and violin plots, as no time data was provided in the dataset.
3. **Missing Country Data**:
    - **Challenge**: The dataset did not contain country-specific data, making it impossible to create a traditional choropleth map based on country.
    - **Solution**: We focused on app categories instead of country-level analysis, highlighting installs globally by category.

Outcomes and Impact

1. **Scatter Plot Insights**:
    - The scatter plot revealed a positive correlation between the number of installs and revenue for most categories.
    - Some categories, like "Games" and "Productivity," exhibited strong revenue generation despite having fewer installs.
2. **Choropleth Map Insights**:
    - While we faced limitations due to the lack of country data, we were able to create a map highlighting app categories with significant installs globally.
    - The filter excluded categories starting with certain letters, allowing for more focused analysis.
3. **Violin Plot Insights**:
    - The violin plot revealed the distribution of ratings for each app category.
    - Categories with more than 50 apps showed distinct differences in ratings, particularly for apps with names containing the letter "C."

Conclusion

This report has successfully implemented several data visualizations to explore relationships between revenue, installs, and app categories for paid apps. Despite the challenges posed by missing data (country and time-related), the tasks were completed by leveraging scatter plots, choropleth maps, and violin plots to offer valuable insights into app performance and monetization strategies.

The findings suggest that app developers can target categories with high install-to-revenue ratios and consider additional monetization strategies for categories with lower installs. The work also highlights the importance of data availability for creating more accurate visualizations and the need for flexibility when constraints such as time-based filtering and country data are unavailable.