Google PlayStore App

PROJECT AT UNIFIED MENTOR

About Dataset

Context
While many public datasets (on Kaggle and the like) provide Apple App Store data, there are not many counterpart datasets available for Google Play Store apps anywhere on the web. On digging deeper, I found out that the iTunes App Store page deploys a nicely indexed appendix-like structure to allow for simple and easy web scraping. On the other hand, Google Play Store uses sophisticated modern-day techniques (like dynamic page load) using JQuery making scraping more challenging.

Content
Each app (row) has values for catergory, rating, size, and more.
Acknowledgements
This information is scraped from the Google Play Store. This app information would not be available without it.

Inspiration
The Play Store apps data has enormous potential to drive app-making businesses to success. Actionable insights can be drawn for developers to work on and capture the Android market

What to do

Objective

The goal of this project is to analyze the characteristics of apps on the Google Play
Store, including their ratings, reviews, sizes, installation counts, and more. Theanalysis will help identify trends, outliers, and patterns in the app market

Data Cleaning

You may need to clean the data by handling missing values, converting data types,and removing duplicates

Dataset Description

Context:

- This dataset provides information on various apps available in the Google Play Store. Collecting data from this platform is challenging due to dynamic page load techniques, unlike the easier-to-scrape Apple App Store.

Content:

- Each app entry includes attributes like category, rating, size, type, price, content rating, and genres.
- Acknowledgment:
- The dataset was sourced from web scraping the Google Play Store.

Project Overview

This project analyzes data from the Google Play Store, providing insights into app categories, ratings, reviews, and various features available on the platform. The dataset includes multiple attributes such as app category, rating, reviews, size, installs, price, and content rating

- Objective: Load the dataset into a DataFrame for initial examination.

Process:

- Imported necessary libraries: pandas, numpy, matplotlib, seaborn, sweetviz, ydata_profiling.
- Read the dataset file googleplaystore.csv and displayed a preview to verify successful ingestion.
- Showing details like name, category, rating, reviews, size, installs, etc.

Data Preprocessing

- Goal: Clean the data by handling missing values, transforming data types, and preparing it for analysis.

Steps:

- Checked for null values and handled them as necessary.
- Converted data types, especially for numeric fields like Reviews, Installs, and Price.
- Standardized formats for features like Size and Last Updated to make them suitable for analysis.

Exploratory Data Analysis (EDA)

- **Objective:** Generate visual insights and identify key patterns and relationships within the dataset.

  Techniques Used:
- Univariate Analysis: Analyzed individual features, including app ratings and install counts, to understand general distributions.
- Bivariate Analysis: Explored relationships between variables, such as how app category correlates with the number of installs or average rating.
- **Visualization Tools:**
  - Plots using matplotlib and seaborn to illustrate distribution and relationships.
  - Sweetviz and ydata profiling for automated report generation to summarize dataset characteristics.

Analysis Summary

Key Findings:

- Top Categories: Most popular app categories based on install count and user ratings.
- Rating Trends: Analysis of app ratings across different categories and their distribution.
- App Sizes: Insights on how app size varies across categories and its potential impact on installs and ratings.
- Pricing Patterns: Free vs. paid app comparison, observing trends in installs and ratings across price tiers.

Project Summary

OVERVIEW OF THE DATA

The first step is to get a feel for the data's structure. In the Google Play Store dataset, we have information about apps, including their category, rating, number of reviews, size, and installs. The second dataset focuses on user reviews and the sentiment (positive, negative, or neutral).

WHAT WE LEARNED

The datasets are large and contain valuable information about both apps and user feedback. No major structural issues were found in this step.

Project Summary

CLEANING THE DATA

Before analysis, we clean the data by checking for missing values, ensuring the correct data types, and removing any duplicate entries.

WHAT WE LEARNED

Some apps are missing ratings, and we need to decide how to handle these missing values.

- The "Installs" and "Price" columns need to be cleaned and converted to numeric formats for proper analysis.

SUMMARY STATISTICS

We check basic statistics like the average rating, number of reviews, and app prices

WHAT WE LEARNED

- Most apps have high ratings (around 4.0 and above).
- The number of reviews is uneven, with a few apps having millions of reviews and many having very few.
-Most apps are free, but the few paid apps vary widely in price.

LOOKING AT RATINGS

We analyze the distribution of app ratings to identify any patterns or trends.

WHAT WE LEARNED

Most apps are rated highly, but some have notably lower ratings. These apps might have issues that require attention.
- A histogram of ratings shows that most apps fall into the 4.0+ rating range.

EXPLORING APP CATEGORIES

Next, we examine the distribution of apps across different categories (e.g., games, productivity, family).

WHAT WE LEARNED

Certain categories like "Games" and "Family" have many apps, but these categories differ in terms of popularity and ratings.
- Visualizing this with bar charts or pie charts shows the dominance of certain app categories.

FINDING RELATIONSHIPS BETWEEN FEATURES

We explore whether there are correlations between key features like installs, reviews, app size, and ratings.

WHAT WE LEARNED

Popular apps (with more installs) tend to have more reviews and higher ratings.
- A correlation heatmap helps visualize the strength of these relationships.

EXPLORING USER SENTIMENT (FROM REVIEWS)

The second dataset contains user reviews, labeled as positive, negative, or neutral. This gives us deeper insights into user opinions beyond just ratings.

WHAT WE LEARNED

Most reviews are positive, but there are some negative ones that could point to areas of improvement for certain apps.

- A word cloud or bar plot can help visualize the most common words or sentiment counts.

COMPARING RATINGS AND REVIEWS

We analyze whether the sentiment of user reviews matches the app ratings.

WHAT WE LEARNED

Apps with higher ratings tend to have more positive reviews, but there are exceptions where highly-rated apps still receive negative feedback.

- A scatter plot can show how closely ratings and review sentiments align.

KEY TAKEAWAYS

HIGH RATINGS	Most apps are rated highly, but ratings alone might not reveal all issues
CATEGORY DIFFERENCES	Popular categories like "Games" and "Family" have a large number of apps and installs.
USER REVIEWS	Positive reviews are common, but negative reviews provide valuable feedback for app improvements
RELATIONSHIPS	Popular apps often have better ratings and more reviews, indicating user engagement

In summary, the EDA gives us a broad understanding of the Google Play Store landscape and helps highlight areas where app developers could improve based on user feedback.

Thank you for Watching!

Created by : Rajat Singh