Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

1. Omkar Sargar – omkar0418@gmail.com
2. Utkarsh Shrivastava – utkarshs75@gmail.com

Contributor Roles:

1. **Omkar Dadasaheb Sargar**:
   - Data Wrangling
   - Basic explorations like (head, tail, describe)
   - Handling missing and null value
   - Exploratory data analysis of some columns (‘Ratings’, ‘Reviews’, 'Size', 'Installs', 'Price', ‘Revenue’, ‘Category’, ‘Content_Rating’)
   - Description and plotting of categorical data from multiple columns
   - Presentation, Technical Documentation

2. **Utkarsh Shrivastava:**
   - Basic Explorations (head, tail, describe, columns, shape, index)
   - Outlier detection and handling missing values or null values
   - Exploratory data analysis of some columns (‘Ratings’, ‘Reviews’, 'Size', 'Installs', 'Price', ‘Revenue’, ‘Category’, ‘Last_Updated’)
   - Creation and plotting of new columns and features through various columns
   - Presentation, Technical Documentation

Please paste the GitHub Repo link.

GitHub Link: ->

1. Omkar Dadasaheb Sargar– https://github.com/Omkar-184/Exploratory_Data_Analysis/blob/master/Play_Store_App_Analysis/PlayStore Analysis.ipynb
2. Utkarsh Shrivastava – https://github.com/utkarshs75/PlayStore-Data-Analysis/blob/main/Play_Store_App_review_Analysis.ipynb

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

Our task at hand was to perform the exploratory data analysis of google play store data as well as user review data on apps and provide relevant business insights based on Analysis report. The first sheet of the playstore data contained 13 columns and described various details and aspects of over 10k apps from 33 different categories. The second sheet contained multiple user reviews from 1074 apps.

First, we looked at play store data and performed data wrangling over the dataset. We cleaned the data, converted some column data to numeric type as required and possible; handled missing values, replacing with median for numeric type and with mode for object type data.

Next, we looked at popularity of apps using installs, Categories and reviews columns. We also plotted the categorical description of apps by category. We plotted the correlation heatmap of columns and plotted it but found no significant correlation except between installs and reviews columns. We also plotted the revenue generating app distributions to find ways to boost revenue.
For popularity of apps, we also used the reviews table to find sentiment polarity and filter using subjectivity to discover key areas users focus upon, and plotted multiple graphs as we found relevant. Rating boxplot was used to analyze important categories for developers and users.

Conclusions:

- Most popular categories amongst developers include Family, Game and Tools; while the users prefer Games, Communication, Social and Productivity above others, followed by Tools and Family.
- Paid apps occupy 7.4 % of the market share while 92.6% of the apps are free.
- Through top 10 most installed and most reviewed/popular apps, we can generate revenue by increasing advertisement in them. We can direct users towards our revenue generating apps. Also, paid versions of such free apps can also be introduced including some exclusive features.
- With least popular and least installed free apps, we probably don't need to invest resources in them, so they can be pushed towards decommissioning. If the installs are high with bad ratings, then those apps need to be improved.
- To cover a broader user base, from the most profitable apps, we can conclude that more apps with low price and high profitability should be promoted.
- The Rating plot seems to be a skewed normal distribution left skewed around average 4.2 stars.
- The sentiment distribution shows the degree of polarity of the distributions. In general, we find 63.6% of reviews to be positive, 25% negative and 11.4% neutral.
- The boxplot of rating distribution can be used to check for differences between categories. No. of ratings & reviews signifies the level with which the user preferred the app. Categories with less outliers and a more balanced distribution and more installs can be a much safer investment than others.

Please paste the drive link to your deliverables folder. Ensure that this folder consists of the project Colab notebook, project presentation and video.

Drive Link:
https://drive.google.com/drive/folders/1VtgwAREcpePymwlbOGmRvxi2Vxmhp54M?usp=sharing