Exploratory Data Analysis(EDA) with PySpark on Google Playstore

Introduction

This project is a journey of analyzing various apps found on the Google play store with the help of PySpark.

Data Content

Dataset: Dataset is downloaded from Kaggle
It consists of 28 columns as mentioned below and 450796 rows.

App Name, App Id, Category, Rating, Rating Count, Installs, Minimum Installs, Free, Price, Currency, Size, Minimum Android, Developer Id, Developer Website Developer Email, Released, Last update, Privacy Policy, Content Rating, Ad Supported, In app purchases, Editor Choice, Summary, Reviews, Android version Text, Developer, Developer Address, Developer Internal ID, Version

Motivation

- Being an everyday phone user, it is interesting to take the real time application information and drive the insights on installations. Looking forward to learn more concepts in Pyspark to implement this project in a better way.

Data Preparation

- The data mostly appears to be clean.
- There are some special characters in the data which needs to be cleaned.

Goal with EDA

- Exploratory Analysis and Visualization
- Top categories in the play store?
- Top apps contains the highest number of installations?
- Ratings structure of the Apps?
- Distribution of App Sizes?
- What % of Apps are updating in a regular basis?
- Free Vs Paid Apps?
- Paid Apps and its top earnings calculated based on installations?

Thank You!