**1/ Project Overview**

This project project involved comprehensive data cleaning, transformation, and analysis using pyspark running on the databricks community version. In this pyspark project, I did a comprehensive analysis of the Android app market by comparing over ten thousand apps in Google Play across different categories to answer the following questions:

* Find out Top 10 reviews given to the apps
* Top 10 install apps and distribution of type (free/paid)
* Category wise distribution of installed apps
* Top paid apps
* Top paid rating apps

**2/ About dataset**

Google Play Store data set which can be found on Kaggle

[Google Play Store data (kaggle.com)](https://www.kaggle.com/datasets/yassershrief/goggle-play-data)

Mobile apps are everywhere. They are easy to create and can be lucrative. Because of these two factors, more and more apps are being developed. The data for this project was scraped from the\* Google Play \*website [<https://play.google.com/store/apps?hl=en> .]() While there are many popular datasets for Apple App Store, there aren't many for Google Play apps, which is partially due to the increased difficulty in scraping the latter as compared to the former. The data files are as follows:

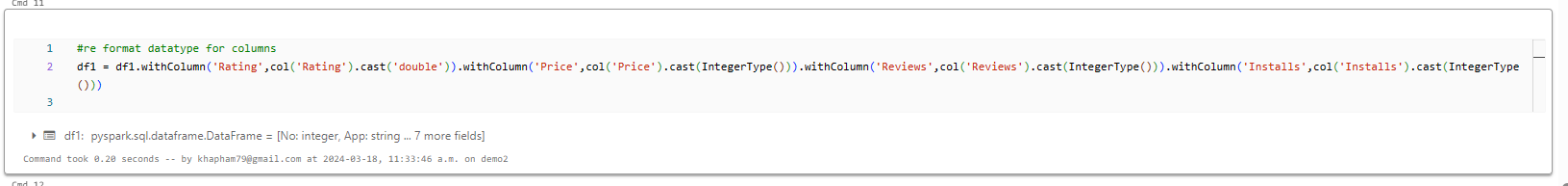
* apps.csv: contains all the details of the applications on Google Play. There are 13 features that describe a given app.
* user\_reviews.csv: contains 100 reviews for each app, most helpful first. The text in each review has been pre-processed and attributed with three new features: Sentiment (Positive, Negative or Neutral), Sentiment Polarity and Sentiment Subjectivity.

**3/ Notes**

Data cleaning is required for the 2 columns, otherwise the data in these 2 columns will be converted to either zero or null.

* Install: to remove the none digit off the data
* Price: to remove the $ sign off the data





**4/ Results**

Find out Top 10 reviews given to the apps

A screenshot of a computer

Description automatically generated

Top 10 install apps and distribution of type (free/paid)

A screenshot of a computer

Description automatically generated

Category wise distribution of installed apps

A screenshot of a computer

Description automatically generated

Top paid apps

A screenshot of a computer

Description automatically generated

Top paid rating apps

A screenshot of a computer

Description automatically generated