# Analyzing Google Play Apps to identify most popular app sector

# ETL Project

## **Team members**

|  |
| --- |
| Fenny Patel, Justin Ho, Nikki Ryan, Parisa Charkhgary, Sidney Bowe |

## **Extract**

|  |
| --- |
| About the Datasets The three datasets that we used for the purpose of this project, gives any app-developing business an enormous potential to success. By working on these datasets we can study Android market. These datasets are chosen from Kaggle.   * Google Playstore Dataset is a CSV file with 1,048,576 rows and 24 columns, the data was collected in June 2021 * Extended google playstore user review has 64,296 rows and 33 columns * Extended google playstore has 64296 rows and 23 columns |

|  |  |  |
| --- | --- | --- |
| File Name | Source | Format |
| extended\_googleplaystore.csv | kaggle.com  Google Play Store Apps Extended  Based on the original dataset 'Google Play Store Apps' with more features | CSV |
| extended\_googleplaystore\_user\_reviews.csv | kaggle.com  Google Play Store Apps Reviews (+110K Comment)  Web scraped data of over 110k reviews from different genres of Apps. | CSV |
| Google-Playstore.csv | kaggle.com  Google Play Store Apps  Google Play Store App data of 2.3 Million+ applications. | CSV |

## **Transformation**

|  |
| --- |
| We would like to study these datasets in more depth to find out the about App’s ratings, reviews, price, etc. Therefore we took steps to clean data by removing duplicates so we would work with more accurate data, removing columns that are not useful for the purpose of this analysis and work only with data from which useful insights can be drawn, and renaming columns. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| File Name | Data Frame Name | Rejected data | Kept Columns | Renamed Columns |
| extended\_googleplaystore.csv | clean\_neo  (removed 15 columns) | None | App, Genres (categorical), Rating, Reviews, Installs, Type, Price, Content Rating | Genres (categorical)' changed to 'genre  All other columns changed to lower case  Spaces on columns names replaced with \_ |
| extended\_googleplaystore\_user\_reviews.csv | clean\_neo\_reviews (removed 30 columns) | dropna | app, translated\_review, original\_sentiment | All columns names changed to lower case. Spaces on columns names replaced with \_ |
| Google-Playstore.csv | clean\_gauthamp (removed 14 columns) |  | app\_name, category, rating, rating\_count, installs, free, price, content\_rating,  ad\_supported, in\_app\_purchases | All columns names changed to lower case. Spaces on columns names replaced with \_ |

## **Load**

|  |  |  |  |
| --- | --- | --- | --- |
| Finale DB | Table |  |  |
| PostgreSQL using SQLAlchemy | neomatrix |  |  |
| PostgreSQL using SQLAlchemy | user\_reviews |  |  |
| PostgreSQL using SQLAlchemy | gauthamp |  |  |

## **Analysis**

The data has been prepared to review GoogleApps ratings to …