UTM

UNIVERSITI TEKNOLOGI MALAYSIA

FACULTY OF ENGINEERING

SCHOOL OF COMPUTING

SEMESTER 1/20202021

SCSP3213: Business Intelligence

LECTURER: Dr. Noorfa Hazlinna Mustafa

SECTION: 02

TITLE: AA Project Proposal (PlayStore Applications)

Name: Aiman Shaquan Bin Mohd Yusof

Matric No.: A18CS0021

Data Background

This data set is about android applications in Google Playstore. The data holds 2 tables in which one of the tables includes the details of Google PlayStore apps such as name, category, rating, installs, price, genres and etc. Another one of the tables consist of the user's review of the app such as comments, sentiment, polarity and subjectivity.

Proper cleaning, analysis and visualizations are crucial for Google to be able to see trending applications or games, genres, a highly legitimate functioning apps and also in providing all possible recommendations to the users.

Information of Data Source

The data set has 2 .csv files which are googleplaystore.csv and googleplaystore_user_reviews.csv

- googleplaystore.csv (contains the data of the applications in the Google Playstore)
  - App – the applications' name
  - Category – the applications' category
  - Rating – ratio and summed ratings of user's reviews
  - Reviews – review counts of the application
  - Size – the memory storage for users to download
  - Installs – counts of the application's installations
  - Type – the application's type either Free or Paid
  - Price – the price for the paid type applications
  - Content Rating – which targetted category ages for the applications
  - Genres – the application's genre and an app can belong to multiple genres
  - Last Update – application's last update date by the developers
  - Current Version – application's current version
  - Android Ver – minimum android version for the application to operate

- googleplaystore_user_reviews.csv (contains the data of the reviews of the applications)
  - App – the applications' name that the review is on
  - Translated_Review – the user's comment or text review
  - Sentiment – the evaluation of the review whether it's positive or negative
  - Sentiment_Polarity – the sentiment polarity score in numeric value
  - Sentiment_Subjectivity – the sentiment subjectivity value to the applications

Link for the dataset: https://www.kaggle.com/lava18/google-play-store-apps

ETL Methods

cleaning, selection, join.

File Sources	**Transformation**	**Component/Implementation**
googleplaystore.csv googleplaystore_user_reviews.csv	Generate id for each app and id for each review	tMap: generate random unique UUID for each
googleplaystore.csv	Remove or fill null values with appropriate values	tMap: fill the null value with default value Relational.ISNULL(column)?default value:column
googleplaystore_user_reviews.csv	Transform nan or null string into real null	tMap: fill the null value with real null Relational.ISNULL(column)?””:column
googleplaystore.csv	Change update dates to real date format	tMap: joining the time dimension by TalendDate.addDate(TalendDate.parseDate)
googleplaystore.csv googleplaystore_user_reviews.csv	Merge the table by app id or app name	tMap and tUniqueRow

Database Schema

- Star schema

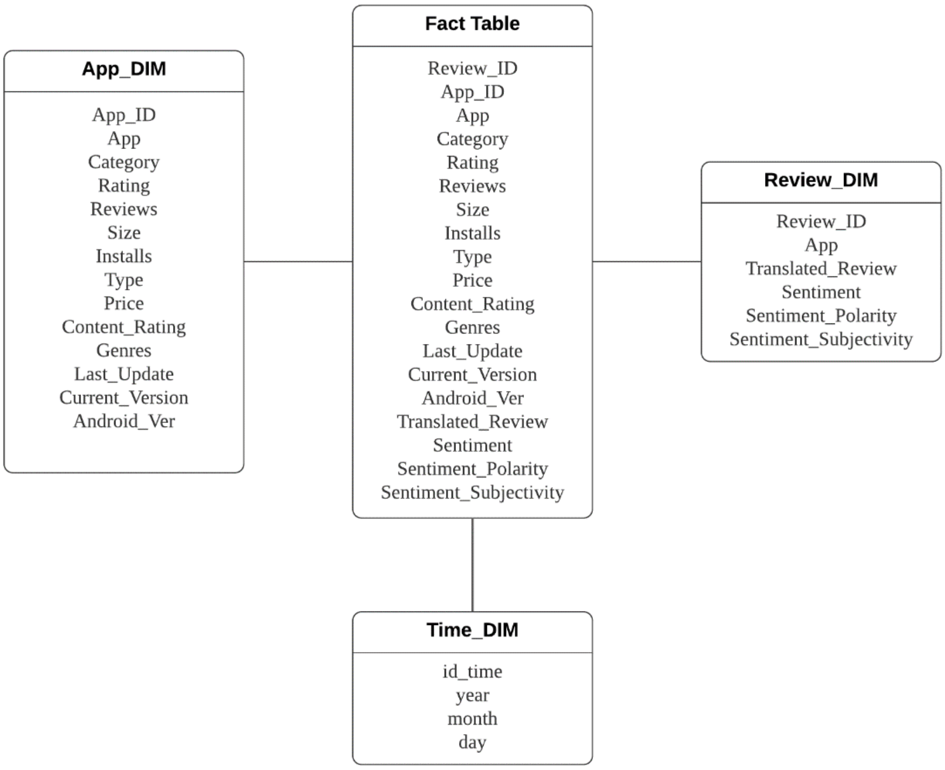

Chart and Dashboard Design

Charts:

- Line chart of the genre and update date
- Pie chart of the percentage of type and number installs
- Bar chart of rating and by app category and type apps
- Stacked barchart of total installs and total paid/free apps
- Box plot for paid/free apps and average rating
- Treemap of genre and review

Dashboard:

- Dashboard that'd focus on applications installs by using line chart of the genre and update date, pie chart of the percentage of type and number installs, and stacked barchart of total installs and total paid/free apps.

- Dashboard that'd focus on applications review by using bar chart of rating and by app category and type apps, Box plot for paid/free apps and average rating, and treemap of genre and review.