CS-418 Project Proposal

Project Idea : Google Play Store Apps
Problem Statement: Identify few of the parameters that can affect or determine the rating of your application on Google Play Store based on the existing dataset that is present on Kaggle.

Brief Description of Data Science Solution:
1. **Regression**: Identify variables(predictor variables) that are important for a higher rating on Google Play Store. We can use Multiple Linear Regression to identify how directly/inversely proportional a variable can be to the Play Store Rating. The variables are category, size, number of reviews, number of installs, pricing,etc.
2. **Clustering**: Based on this unsupervised machine learning method, we can identify and group similar apps from the dataset and thus determine what variables contribute towards the similarity.

Alternate:
3. **Sentiment Analysis**: Based on the second dataset, we can read the reviews and predict if the review was positive or negative.

Data Source: https://www.kaggle.com/lava18/google-play-store-apps
This data source has two datasets:
1. **googleplaystore.csv** containing the app name, the category of the app, ratings, number of reviews, app size, pricing and other variables to identify and describe an app.
2. **googleplaystore_user_reviews.csv** which contains the user reviews for each app, and also an attribute named sentiment to identify if the review was positive or not.