Dataset

Google Play Store Apps

- Source: https://www.kaggle.com/lava18/google-play-store-apps
- googleplaystore_user_reviews.csv (64295 rows x 5 columns)

	App	Translated_Review	Sentiment	Sentiment_Polarity	Sentiment_Subjectivity
0	10 Best Foods for You	I like eat delicious food. That's I'm cooking ...	Positive	1.00	0.533333
1	10 Best Foods for You	This help eating healthy exercise regular basis	Positive	0.25	0.288462
2	10 Best Foods for You	NaN	NaN	NaN	NaN
3	10 Best Foods for You	Works great especially going grocery store	Positive	0.40	0.875000

- googleplaystore.csv (10841 rows x 13 columns)

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up

EDA Findings

1. Data quality/Clean:

- Missing values (drop)
- Duplicated rows (drop)
- Column naming (convert to Snake_case)
- Data type (convert String to Numeric/int, Numeric to float)
- Unused columns (drop)
- Non-letter characters in Translated_Reviews (replace with space)
- Upper case (convert to lower case)
- Irrelevant words (remove)

3. Variable relationships

- Installs & Reviews: Weak correlation
- Sentiment_Subjectivity & Rating: Weak correlation
- Sentiment_Polarity & Reviews: Negative correlation
- Sentiment_Polarity & Rating: Weak correlation
- Rating & Reviews: Weak correlation

2. Common features:

- Column: App
- Reorganize googleplaystore_user_reviews with unique list of App names.
- Merge
- Inner join
- Left on: new_googleplaystore_user_reviews.App
- Right on: googleplaystore.App

4. Data distribution:

- Highly skewed distribution: Rating, Reviews, Installs
- Moderately skewed distribution: Sentiment_Subjectivity
- Approximately symmetric distribution: Sentiment_Polarity
- Installs attribute has a lot of outliers.

5. Common trends:

- Most of the apps belong to the Family Category.
- DATING category has the min rating and EVENTS has the max rating.
- Most of the apps belongs to the content rating of Everyone and are freely available.

Research Questions

- Can we predict App Ratings for Apps on the Google Play Store? (Supervised Learning - Classification)
- What categories of Apps that a marketing strategy should targeted on the Google Play Store Apps? (Unsupervised Learning - Clustering)
- Can we use NLP to train the Google Play Store dataset to predict reviews? (NLP - Syntactic Analysis + Semantic Analysis)