STAT 350: Project Dataset Description

The data is modified from a Kaggle dataset Google Play Store Apps dataset. While many public datasets (on Kaggle and the like) provide Apple App Store data, there are not many counterpart datasets available for Google Play Store apps anywhere on the web. The iTunes App Store page deploys a nicely indexed appendix-like structure to allow for simple and easy web scraping. On the other hand, Google Play Store uses sophisticated modern-day techniques (like dynamic page load) using JQuery making scraping more challenging.

The Play Store apps data has enormous potential to drive app-making businesses to success. Actionable insights can be drawn for developers to work on and capture the Android market. The dataset contains the 'most relevant' reviews for each app, and each review text/comment has been pre-processed and attributed with several sentimental features. On the other hand, details of the applications on Google Play such as number of reviews and categories are also available. We want to look at the relationship between app rating, comments and other useful features.

Variable Name	Dataset Column	Description
Id	1	Id.
AgeReviewer	2	Self-reported age of the reviewer
GenderReviewer	3	Self-reported gender of the reviewer
RevRating	4	Rating on 1~5 scale from the review author, adjusted according to his/her rating history.
RevType	5	'Long' if the review has more than 15 words, otherwise 'short'.
RevLen	6	Review length in characters.
WordCount	7	Review word count.
AvgWordLen	8	Average word length in the review.
SentiG2	9	Sentiment on a positive/negative scale, generated from Google API.
SentiG3	10	Sentiment on a positive/negative/neutral scale, generated from Google API.
SentiGPol	11	Sentiment polarization score with 1 being totally positive, -1 being totally negative, generated from Google API.
SentiGSub	12	Sentiment subjectivity score with 1 being totally subjective, 0 being totally objective, generated from Google API.
SentiR2	13	Sentiment on a positive/negative scale, generated from R SentimentAnalysis package.
SentiR3	14	Sentiment on a positive/negative/neutral scale, generated from R SentimentAnalysis package.

SentiRPol	15	Sentiment polarization score generated from R SentimentAnalysis package.
App	16	App name.
AppCat	17	App category.
AppRating	18	Average app rating on 1~5 scale.
nReviews	19	Number of reviews for the app.
nInstalls	20	Number of installs for each app.
ContRating	21	App content rating, everyone/ everyone10+/teen/mature
lRevLen	22	Natural log transformation of RevLen $\ln(\mathbf{RevLen} + \mathbf{1})$

Disclaimer: the original datasets were modified and changed to suit the purposes of STAT 350. No serious conclusions should be drawn from the analysis.