

STAT 350: Project Dataset Description

The data was collected from the Google Play Store in July 2026 using an automated collection tool. For roughly 640 apps across four categories (games, family, dating, and finance), we gathered the app's store details together with a sample of its recent user reviews. The raw file is AppRatingFALL2026.csv; the cleaned analysis file appRating_data_clean.csv is produced from it in Computer Assignment 1.

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. Each review text has been pre-processed and attributed with several sentiment features scored by two independent sentiment-analysis algorithms, and app-level details such as the number of reviews, installs, and content rating are also included. We want to look at the relationship between app rating, reviews, and other useful features. Review text itself is not included in the dataset.

Variable Name	Dataset Column	Description
Id	1	Unique Identifier of the reviewer
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 their rating history.
RevType	5	'Long' if the review has 15 or more 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 class on a positive/negative scale, from the first sentiment algorithm (G).
SentiG3	10	Sentiment class on a positive/negative/neutral scale, from the first sentiment algorithm (G).
SentiGPol	11	Sentiment polarity score with 1 being totally positive, -1 being totally negative, from the first sentiment algorithm (G).
SentiGSub	12	Sentiment subjectivity score with 1 being totally subjective, 0 being totally objective, from the first sentiment algorithm (G).
SentiR2	13	Sentiment class on a positive/negative scale, from the second sentiment algorithm (R).
SentiR3	14	Sentiment class on a positive/negative/neutral scale, from the second sentiment algorithm (R).
SentiRPol	15	Sentiment polarity score from the second sentiment algorithm (R).
App	16	App name.

AppCat	17	App category.
nReviews	18	Number of ratings the app has received in the store.
nInstalls	19	Number of installs for the app (ordered categories such as 10000+).
ContRating	20	App content rating: Everyone / Everyone 10+ / Teen / Mature.
lRevLen	21	Natural log transformation of review length: $\log(\text{RevLen} + 1)$.
AppRating	22	Average app rating on 1~5 scale as displayed in the store.
ThumbsUp	23	Number of users who marked the review as helpful (thumbs-up count).
DaysSinceReview	24	Days between the review date and the dataset snapshot date (2026-07-18).
AppFree	25	Whether the app is 'Free' or 'Paid'.
Price	26	App price in US dollars (0 for free apps).

Disclaimer: the dataset was modified and some variables were simulated or adjusted to suit the purposes of STAT 350. No serious conclusions should be drawn from the analysis.