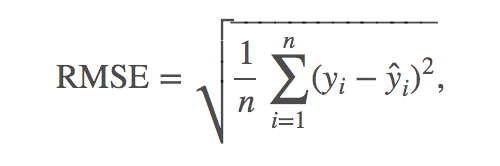
### **Root Mean Squared Error (RMSE)**

Submissions are scored on the root mean squared error. RMSE is defined as:



where y hat is the predicted revenue for a customer and y is the natural log of the actual revenue value.

What files do I need?

You will need to download train.csv and test.csv. These contain the data necessary to make predictions for each fullVisitorIdlisted in sample\_submission.csv.

You can also access this data via BigQuery, using the example notebook provided as a way to get started. The data in train.csv is contained in the BigQuery ga\_train\_set dataset, and the data in test.csv is contained in the ga\_test\_set dataset, under the kaggle-public-datasets project, accessible through Kernels. In those BigQuery datasets, each day's worth of data is a separate table for more efficient EDA / download.

All information below pertains to the data in both CSV and BigQuery format.

What should I expect the data format to be?

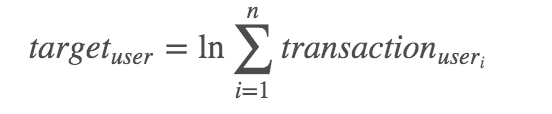
Both train.csv and test.csv contain the columns listed under Data Fields. Each row in the dataset is one visit to the store. Because we are predicting the log of the total revenue per user, be aware that not all rows in test.csv will correspond to a row in the submission, but all unique fullVisitorIds will correspond to a row in the submission.

*IMPORTANT: Due to the formatting of fullVisitorId you must load the Id's as strings in order for all Id's to be properly unique!*

There are multiple columns which contain JSON blobs of varying depth. In one of those JSON columns, totals, the sub-column transactionRevenue contains the revenue information we are trying to predict. This sub-column exists only for the training data.

What am I predicting?

We are predicting the natural log of the sum of all transactions per user. For every user in the test set, the target is:



File Descriptions

* train.csv - the training set - contains the same data as the BigQuery rstudio\_train\_set.
* test.csv - the test set - contains the same data as the BigQuery rstudio\_test\_set.
* sampleSubmission.csv - a sample submission file in the correct format. Contains all fullVisitorIds in test.csv.

Data Fields

* *fullVisitorId*- A unique identifier for each user of the Google Merchandise Store.
* *channelGrouping* - The channel via which the user came to the Store.
* *date* - The date on which the user visited the Store.
* *device* - The specifications for the device used to access the Store.
* *geoNetwork* - This section contains information about the geography of the user.
* *sessionId* - A unique identifier for this visit to the store.
* *socialEngagementType* - Engagement type, either "Socially Engaged" or "Not Socially Engaged".
* *totals* - This section contains aggregate values across the session.
* *trafficSource* - This section contains information about the Traffic Source from which the session originated.
* *visitId* - An identifier for this session. This is part of the value usually stored as the \_utmb cookie. This is only unique to the user. For a completely unique ID, you should use a combination of fullVisitorId and visitId.
* *visitNumber* - The session number for this user. If this is the first session, then this is set to 1.
* *visitStartTime* - The timestamp (expressed as POSIX time).

Removed Data Fields

Some fields were censored to remove target leakage. The major censored fields are listed below.

* *hits* - This row and nested fields are populated for any and all types of hits. Provides a record of all page visits.
* *customDimensions* - This section contains any user-level or session-level custom dimensions that are set for a session. This is a repeated field and has an entry for each dimension that is set.
* *totals* - Multiple sub-columns were removed from the totals field.