Final Report: Group 8 (Jaimie, Steven, Raj)

Extract:

Our original data source was 4 csv files from Kaggle. [Link](https://www.kaggle.com/cjgdev/formula-1-race-data-19502017?select=constructors.csv)

We used four csv files:

1. constructorsResults.csv
2. constructorStandings.csv
3. constructors.csv
4. races.csv

The data populating the csv came from APIs and was pulled at the end of the 2017 season. [Link](http://ergast.com/mrd/)

Transform:

Cleaning that was done was removing columns that contained (NaN) values. We also renamed some columns to reflect out database column names.

In the races.csv we had a column labeled time. We did not fill in the empty rows because in the documentation from the APIs the data is continually being collected. We did not pull from the API and add to the data because of time constraints.

Load:

The final database was labeled Formula1 and was built in Postgresql. Folder DB Objects on GitHub contains our .sql table creation. We chose to make most values not null to allow for accurate reporting. We chose to go with a relational database because we were working with tables and it would ultimately be easier to query data.

We chose to create a table for each csv to keep the information together and maintain the integrity of the data. We did not want to combine tables for this reason. We structured the database so that each table has a primary key with some tables having multiple foreign keys for easy access during querying.