**ETL Project Report**

**Super Bowl Analysis**

Group Members: Kirby Miller, Kerry Wehner, Ashwini Kamat

Objective:

The purpose of this project is to perform the Extract-Transform-Load (ETL) process by extracting and reading original files, cleaning the datasets, and loading the final database. With Super Bowl LV coming up in the next week, we felt inspired to explore various aspects of past Super Bowl games (1967-2020).

Data Cleanup and Analysis:

Extraction:

* The data are all formatted as CSV files and have been downloaded from Kaggle.

[TV, halftime shows, and the Big Game Dataset | Kaggle](https://www.kaggle.com/amithasanshuvo/tv-halftime-shows-and-the-big-game-dataset?select=tv.csv)

[Superbowl History 1967 - 2020 | Kaggle](https://www.kaggle.com/timoboz/superbowl-history-1967-2020)

* Tables Used:

super\_bowls.csv

Columns: Year, super\_bowl, venue, city, state, attendance, team\_winner, winning\_pts, qb\_winner\_1, qb\_winner\_2, coach\_winner, team\_loser, losing\_pts, qb\_loser\_1, qb\_loser\_2, coach\_loser, combined\_pts, difference\_pts

Name\_MVP.csv

Columns: Year, Super\_bowl, Winner, Winner\_Pts, Loser, Loser\_Pts, MVP, Stadium, City, State

tv.csv

Columns: super\_bowl, network, avg\_us\_viewers, total\_us\_viewers, rating\_household, share\_household, rating\_18\_49, share\_18\_49, ad\_cost

Transformation:

* Each of the .csv files is read into a Pandas data frame in Jupyter notebook
* Created Tables:

super\_bowls.csv

Columns: Year, team\_winner, team\_loser, attendance, qb\_winner, qb\_loser, coach\_winner, coach\_loser, combined\_pts, difference\_pts

Name\_MVP.csv

Columns: Super\_bowl, MVP, Year

tv.csv

Columns: super\_bowl, network

* Limited and renamed the columns (listed above) for data standardization
* Identified primary key

Loading:

* Loaded data using PostgreSQL
  + Used PostgreSQL because of use of foreign keys and familiarity of SQL
* Created an entity relationship diagram
* Final database: Super\_Bowl\_Analysis