Group 3: Joshua Cohen, Nam Luu, Renata Sant'Anna and Sarah Raisian

ETL Project Proposal

Proposal:
We decided on two datasets that provide Jeopardy questions. We want to extract them from two tsv files, transform to drop irrelevant data columns, clean any missing data, and load into a PostgreSQL relational database.

Dataset:
https://www.kaggle.com/prondeau/350000-jeopardy-questions/version/2#master_season1-35.tsv

Dataset 1: Kids Jeopardy
Number of Rows: 20,800
Dataset 2: Master Season 1-35
Number of Rows: 349,641
Features(for both datasets): Round, Value (Points, 0 is in Round 3 because they place their own bets), daily_double (Y/N), category, comments, answer, question, air_date, and notes.

Cleaning:
Drop comments column, drop daily_double, drop notes, add binary column 0 for Master, 1 for Teen/Kids.

Hypotheses:
Categories will be easily distinguishable between Master and Kids.
Question difficulty will seem easier for Kids and more approachable.

Obstacles: value column had to be renamed (reserved name)

SQL Code:

```
CREATE TABLE jeopardy_table (
id SERIAL PRIMARY KEY,
round INT,
category TEXT,
answer TEXT,
question TEXT,
air_date DATE,
kids_jeopardy INT,
point_value INT);
```

Group 3: Joshua Cohen, Nam Luu, Renata Sant'Anna and Sarah Raisian

ETL Project Report

The ***Extract part*** was done by choosing the dataset about questions and answers from Jeopardy game show, seasons 1 through 35 from kids/teen and adults. It was obtained from Kaggle, on this following link: https://www.kaggle.com/prondeau/350000-jeopardy-questions/version/2#master_season1-35.tsv

The data was formatted in two *.tsv* files, one corresponding to the kids/teen and the other to the adults (named master). The kids/teen dataset has data from Feb/1987 to June/2019 and has a total of 20,800 observations. The adults/master dataset has data from Sep/1984 to July/2019 and has a total of 349,641 observations.

Both datasets contained information of ***Round***, which represents the three rounds participants compete: 1 for Single Jeopardy, 2 for Double, 3 for Final; ***Value***, which corresponds to the value of a correct answer, ex.: 100, 200, 400, 600, etc; ***Daily Double***, which is a yes or no variable related to whether the participant got the secret question that could potentially double its earnings; ***Category***, which corresponds to the category classification; ***Question***, the question asked; ***Answer***, the correct answer; ***Comments***, extra information on category, ***Air Date***, date the show aired; ***Notes***, records if it was a tournament or special match.

After exploring the database, we came across the ***Transformation part*** that we realized the ***Comments*** and ***Notes*** columns were basically empty on both datasets and we decided to drop them. We also decided that ***Value and Daily Double*** wouldn't be necessary information when combining the two datasets. So we also decided to drop those columns.

In order for us to distinguish between Kids/Teen questions and answers and adults, we created a dummy variable - a new column called ***kids_jeopardy*** - that takes value 0 when it's a kids/teen game and takes value 1 when it's an adult (master) game.

For the last part, ***Load part***, we had created two different dataframes in *pandas* containing the same name and number of columns. So we merged the datasets in one dataframe in pandas, using Jupyter notebook script and used the connection in PGAdmin 4 SQL in order to create a combined dataset that includes the kids/teen and adults in one unique dataframe.