**Jeopardy versus the World… or just other random factors**

**Extract, Transform, Load (ETL) Project**

GW Data Analytics and Data Visualization Boot Camp, Instructor: Dan Wood

Author: Melinda Wiley

15 November 2018

Task is to identify several datasets, transform identified sets, and load them into a database. The initial datasets and primary focus of this project was based around Jeopardy Questions and Answer datasets; Jeopardy Questions and Answers during the reign of Alex Trebek (1984 – 2018). Is there any correlation between Jeopardy categories, questions, and/or answers and facial hair, divorce rate, marriage rate, leading cause of death, world population, or Ted Talks? Unlikely but let’s check.

**Extraction:**

Searched dataset focused sites for information along with google searches to find additional supporting documentation that I would be able to manipulate/transform. In this, the realization that without a strong project idea and foundation, scope creep, project deviation, and project wander can derail a short deadline project.

Datasets from extraction for transformation are the following: 2x jeopardy, 1x facial hair, 2x marriage and divorce rates, 2x population growth, 1x Ted Talks. Primary grouping category will be time.

**Sources:**

1. Jeopardy questions and answers dataset 1, jeopardy\_1.csv, <https://data.world/sya/200000-jeopardy-questions>
2. Jeopardy questions and answers dataset 2, jeopardy\_2.csv, <https://www.kaggle.com/sanzgiri/jeopardy-questions>
3. Facial hair over the years, University\_Graduate\_Facial\_Hair\_Styles.xlm, <https://www.reddit.com/r/dataisbeautiful/comments/64q73v/university_graduate_facial_hair_styles_18982008_oc/dg44iay/?st=joivc1il&sh=9fb47495>
4. Divorce rate and marriage rate, national\_marriage\_divorce\_rates\_00-16.xlm, <https://www.cdc.gov/nchs/nvss/marriage-divorce.htm>
5. Populations Growth, WPP2017\_POP\_F01\_1\_TOTAL\_POPULATION\_BOTH\_SEXES.xlm, <https://population.un.org/wpp/Download/Standard/Population/>
6. Additionally scraped for projected population growth from <http://www.worldometers.info/world-population/world-population-projections/>
7. Ted Talks, TED\_Talks\_by\_ID\_plus-transcripts-and-LIWC-and-MFT-plus-views.csv, <https://data.world/owentemple/ted-talks-complete-list>

**Transform:**

Data cleaning and transformation. All data was read into a jupyter notebook. Data was formatted in .xlm, .csv, and table from a website.

I found two Jeopardy datasets in .csv. Data for Jeopardy was very limited or more difficult to find than initially expected. Renamed and dropped columns in both datasets to match each other. Jeopardy\_1 was the larger dataset. Jeopardy\_2 was merged into Jeopardy\_1 for its Jeopardy/Double Jeopardy data with a left join creating completed\_jeopardy.

Facial Hair data (1899-2008) was converted from excel to csv. Alex Trebek shaved off his iconic mustache in 2001. He grew it back in 2014 and grew a beard in 2018. Unfortunately, the facial hair dataset only runs to 2008.

Marriage and Divorce rates data (2000-2016) was implemented in excel since the sets were very small and was converted to a csv from an excel.

World Population data comes from two data sources in excel. Set1 date range 1950-2015, set2 2018-2100. Columns were renamed and dropped to match each other then sets were merged by concatenation and create a mega csv.

Ted Talks renamed year filmed to Year.

**Loaded:**

Cleaned collections were loaded into a MongoDB final database.

NoSQL, non-relational database was chosen for the easy of handling large volume semi-structured data. Additionally, MongoDB versatility. Lastly, I needed the practice working with a NoSQL db.

The type of final production database to load the data into (relational or non-relational).

The final tables or collections that will be used in the production database.

Jeopardy is just a syndicated game show. No correlations were found and it can be determined that Jeopardy is not a major trend decision maker ….but not sure about Ted Talks.