Assignment - Information Seeking

**IMDB 5000 Movie Dataset**

Sun, Chuan (2016). *IMDB 5000 Movie Dataset* [Data Set]. Retrieved from <https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset>

**License**: Database is released under Open Database License

The license states that the user is allowed to share and use the database, produce works from the database, and adapt the database. The limitations that apply are that the user must attribute any public use of the database, offer the adapted database under the same license (if database is modified), and if redistributing a restricted version then the user must also redistribute an unrestricted version.

For the exact terms of the license, please refer to this link: <http://opendatacommons.org/licenses/odbl/1.0/>

**Description of Data Set**

This data set is interesting because it provides a means to delve deep into the world of film and analyze the architecture behind every award winner. The data set contains 28 columns, ranging from basics such as title, language, genre, to more insightful data like the number of Facebook likes a movie has. It will be exciting to find if patterns really do exist between some of the factors. Is there a positive correlation between budget and IMDB rating? Or is that idea just a preconceived notion? I could compare a few movies to jump to a conclusion, but comparing 5000 movies would produce a much more accurate and trustworthy result. This and many more interesting questions can be answered using this data set.

A link to the data set can be found here: <https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset>

**Potential Data Users and Decision Makers**

The main potential data user category would be film personnel, starting from the producers. They will want to know how elements such as choice of director, budget amount, and release date influenced the success of the movie (Success is measured by the gross income). Casting directors would be interested in the number of Facebook likes of the entire cast, and the number of Facebook likes for the first 3 main characters, and how this reflects on the outcome. The marketing and advertising manager would like to see if there is a connection between the number of faces on the poster and the popularity of the movie.

Movie critiques and movie buffs would also find the results of this data very useful, as it can help them in writing reviews about the movie.

**3 Questions that the Data Set May Answer:**

1. How can we predict the success of a movie before it is released, using the data presented in the data set? Can we use past data, ex. Cast, director, genre to predict the box office hit of a new movie?
2. What is the relationship between budget and IMDB rating? Does a higher budget translate into a higher gross outcome?
3. How does the release date affect the gross income? Do movies released on a national holiday generate more or less income? What season is the best to release a movie and expect a big turnout?

**House Sales Prediction**

Harl, F (2016). *House Sales in King County, USA* [Data Set]. Retrieved from https://www.kaggle.com/harlfoxem/housesalesprediction

**License:** Database is released under CC0: Public Domain License

The license states that the creator has dedicated this database to the public domain by waiving his/her right to a copyright. The user can copy, modify, distribute and perform on the data set, even for commercial purposes, without asking permission. When using or citing the work, do not imply endorsement by the author.

For the exact terms of the license, please refer to this link: <https://creativecommons.org/publicdomain/zero/1.0/legalcode>

**Description of Data Set**

This data set provides details on the houses that have been sold in King County from May 2014 – May 2015. I find this data set intriguing because it acts as a looking glass into the real estate world, to unfurl mysteries such as why two very similar houses vastly differ in price. The data set contains columns such as the number of bedrooms, square feet, latitude/longitude, the year it was built, and less quantifiable details such as whether or not the house has a good view. Other ordinal data types are expressed as a range of values from 0-5, i.e. ‘Condition’, to represent the quality of the housing structure. This data can be used to predict the sale price for houses on the market in King County.

A link to the dataset can be found here:

<https://www.kaggle.com/harlfoxem/housesalesprediction>

**Potential Data Users and Decision Makers**

One potential data-user and decision maker for this data set would be prospective buyers in King County. This data set will help them filter through houses and navigate through the tedious task of house hunting. They could strike a better bargain with the broker if they are aware of the selling prices of similar houses. On the other side of the argument, brokers and sellers would also be potential data-users. Economists would also be interested in the findings of this data, to predict the status and stability of the housing market in King County.

**3 Questions that the Data Set May Answer:**

1. What will be the estimated sale price of a certain house in King County?
2. How do sale prices differ for houses that are part of the same neighborhood/locality? What is the range of prices, and what factors causes this range of difference? The latitude/longitude can be used to map the houses, to estimate the boundaries of the neighborhood.
3. What is the benefit of renovating your house? How do the prices differ for 2 houses built around the same time, with similar physical characteristics (ex. Number of bedrooms, bathrooms, size of house), where one has been renovated, and the other one has not? Do not take into account features such as ‘condition’, which could be affected by the renovation.

Note: there are uncertainties on the authorship of this data set.

**World Food Facts**

Open Food Facts (2016). World Food Facts [Data Set]. Retrieved from https://www.kaggle.com/openfoodfacts/world-food-facts

**License:** Database is released under Open Database License

The license states that the user is allowed to share and use the database, produce works from the database, and adapt the database. The limitations that apply are that the user must attribute any public use of the database, offer the adapted database under the same license (if database is modified), and if redistributing a restricted version then the user must also redistribute an unrestricted version.

For the exact terms of the license, please refer to this link: <http://opendatacommons.org/licenses/odbl/1.0/>

**Description of Data Set**

This data set is fascinating because it is a vast repository of food facts from almost every country in the world. A multitude of attributes have been captured for each item on the list, spanning from the country of manufacture, the type of packaging, to the amount of vitamin D present per 100 gram of the food item. Unlike the other two data sets I have collected for this assignment, this data set comprises trace data. The data was loaded by volunteers, along with a picture of the product and its nutrition facts, as a means of verifying the data. It will be a distinctive experience to work with data that is meant to be analyzed.

A link to the data set can be found here:

<https://www.kaggle.com/openfoodfacts/world-food-facts>

**Potential Data Users and Decision Makers**

One potential data-user for this data set is nutritionists. This data set is laden with all the various minerals, vitamins, and other nutrients, making it a data warehouse for nutritionists. Other potential users are doctors, fitness experts, food manufacturers and the general public.

Last but not least, this data set would be interesting to data science students.

**3 Questions that the Data Set May Answer:**

1. What one food item is consumed by the most number of countries? Do the ingredients of that item vary with each country?
2. Which country consumes the most sugar? Which food items contain the most sugar?
3. What type of food do Europeans eat? How do the food habits differ for each European country?