**INFM600: Info-Seeking Assignment**

**Data Set 1:**

Chuan Sun, Kaggle (2016), *IMDb 5000 Movie Data Set* (Version 1) [movie\_metadata.csv]. Retrieved from <https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset/version/1>

**License:**

Database released under [Open Database License](http://opendatacommons.org/licenses/odbl/1.0/), individual contents under [Database Contents License](http://opendatacommons.org/licenses/dbcl/1.0/)

**Overview of the data:**

IMDb has now become the world’s most popular source for movie rating and reviewing. The ratings given to movies/shows in this site are considered to be accurate most of the times. A movie is rated by the audience based on various factors like the movie cast, music quality, direction and so on. IMDb also has a feature in which a movie can be rated even before its release, going by the posters, trailers that have been released. It is interesting to know about the correlation among these various factors which finally affect the rating score. The data set contains information regarding 28 such variables for 5043 movies. For example, one of the topics of research can be whether a movie gets a higher rating if the director is popular (which will be reflected by the number of Facebook likes he/she has).

**Potential data users and decision makers:**

The data collected can be used by a wide range of users. Some of them are as follows:

* The audience can use it for finding the higher rated movies, movies that have got more critical reviews or even movies in which their favorite actor has acted in.
* Movie production teams can study the data to find what factors influence the movie ratings and what the audience like.
* IMDb themselves can use this data to test if their rating system is good enough and if it actually reflects the overall quality of the movie.

**Some questions this dataset can answer:**

Following are some questions that can be addressed by studying this data set:

* Are the ratings affected by the popularity of the director/actors? The popularity can be measured using the number of Facebook likes that they have got, which is available in the data.
* What are the genres that people prefer the most? This can be concluded using either the gross earnings of the movie or the ratings available in the data.
* Do high budget movies generally get high ratings or earn significantly better than the others?
* Are the movie ratings consistent with the gross earnings? That is, do the earnings of the movie increase with its ratings or are there discrepancies in this?

**Data Set 2:**

City of San Francisco (2016), *San Francisco Flood Health Vulnerability* [San\_Francisco\_Flood\_Health\_Vulnerability.csv]. Retrieved from <https://data.sfgov.org/Health-and-Social-Services/San-Francisco-Flood-Health-Vulnerability/cne3-h93g>

**License:**

Dataset released under [ODC Public Domain Dedication and License](http://opendatacommons.org/licenses/pddl/1.0/).

**Overview of the data:**

Climate changes across the world might be sudden or gradual. Such climate changes sometimes lead to unforeseen circumstances such as floods and storms. While it is important to adopt preventive measures (like reducing gas emissions) to avoid such natural disasters in the future, it is equally important to study the impact such disasters have on public health. This dataset aims to provide details regarding the various populations living in San Francisco, that are vulnerable to health hazards after or during a flood. The data is collected and categorized for various census block groups and neighborhoods across the city. The dataset contains information regarding geographical attributes of the area (like elevation above sea level and so on), information about the various populations (children, elders, poor and so on) and also information about the various common disorders (asthma, diabetes and so on) that are present in the blocks.

**Potential data users and decision makers:**

The data collected can be used by a wide range of users. Some of them are as follows:

* Public health officials can use this data to determine what populations (i.e. children, elderly people) of any block, which recently witnessed a natural disaster, are likely to be badly affected. For example, if it is a winter storm or a hurricane, then the most vulnerable might be the homeless population.
* Storm Prediction Centre can use the data regarding geographical and climatic attributes like sea level and precipitation to predict a storm in any particular block.
* Government can use this data to understand the health risks associated with the possible disasters the area is prone to, and take preventive measures (or plan infrastructure accordingly) in advance to minimize damage to public health.

**Some questions this dataset can answer:**

Following are some questions that can be addressed by studying this data set:

* Is there any particular health disorder that is significantly more than others in any block?
* How do the health disorders vary according to the housing quality of any particular block?
* The Flood Health Vulnerability Index of every block, which is calculated by proportionately adding up the socioeconomic, demographic, housing and health indicators along with the exposure indicators.
* Which blocks/neighborhood in the city has the highest Flood Health Vulnerability Index values?

**Data Set 3:**

Highways England (2016), *Planned road works on the Highway Agency Road Network* [highways\_agency\_planned\_roadworks.zip]. Retrieved from <https://data.gov.uk/dataset/highways_agency_planned_roadworks>

**License:**

Dataset licensed under [OGL Open Government License](http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/).

**Overview of the data:**

Roadworks along highways can sometimes be dangerous or irritating for drivers, but are necessary nevertheless. It is hence very useful and safe to have the data about planned roadworks and constructions along highways beforehand itself, so that commuters can plan their journey’s accordingly, without sudden and unexpected delays. This dataset consists of descriptions, addresses, expected delays and other relevant data about the roadworks being done on the highways. The data is being stored in XML files. Each file contains data about the roadworks happening on the highways for the next fifteen days after the day the file was created. Other important data in the dataset includes details like if the whole lane would be blocked, if the work was an emergency work or a planned one and so on.

**Potential data users and decision makers:**

The data collected can be used by a wide range of users. Some of them are as follows:

* Regular commuters can use this data to find if there are any roadworks or constructions going on in their routes and plan their journeys accordingly.
* The Highway Agency officials can use the data to track the work being done on the highways. The data contains details regarding the proposed end date, which allows them to assess if the work is being implemented in time.
* Hospitals and Ambulance services can use the data to avoid routes which get them delayed in cases of emergencies
* Traffic Control authorities can use the data to plan backup routes in advance, by considering the expected delays in the case of traffic congestions.

**Some questions this dataset can answer:**

Following are some questions that can be addressed by studying this data set:

* When is a particular roadwork started and when is the tentative end date?
* What kind of traffic management is being done for the roadworks? (i.e. is the carriageway being closed, is the whole lane being closed and so on)
* How many roadworks or constructions will be taking more than fifteen days, and from when are they due.
* Which area/locality has the most number of roadworks going on the whole over the next fortnight.
* How many of the roadworks going on are planned works, and how many are being done due to an emergency.