**PROJECT #2 - ETL Project Report**

**Members - Abla, Alex, Jordy, Viral**

World Happiness Indicators

**Logic**

This dataset collection provides information that can be used to study which socio-economic factors affect happiness levels in a country or vice-versa. By joining Arable Land, Military Expenses, and Charges of the use of intellectual property from World Bank Data Catalog and World Happiness Data from Kaggle datasets, the end user is provided with information that they can use for analysis and decision making, comparing various happiness indicators to the World Happiness Database. Military expenditure was chosen because it may affect happiness by increasing citizen’s sense of security. Arable land was chosen because the amount of natural areas in a country may affect stress levels or the aesthetic appeal of the areas citizens live in. Charges of the use of intellectual property was chosen because this may affect the rate of technological innovation in a country which may affect the happiness by improving citizen’s lives.

**The ETL Process**

(E)xtract

Our data was retrieved from Kaggle and the World Bank – Data Catalog.

* <https://data.worldbank.org/indicator/AG.LND.ARBL.ZS?view=chart>
* <https://www.kaggle.com/unsdsn/world-happiness/version/2>
* <https://data.worldbank.org/indicator/BM.GSR.ROYL.CD?view=chart>
* <https://data.worldbank.org/indicator/MS.MIL.XPND.ZS?view=chart>

Kaggle’s "World Happiness Report" dataset included CSV files with data from 2015, 2016, and 2017. For simplicity we decided to focus on one year, and selected the 2016 dataset. These files provided data on each country’s happiness score. This score is made up of several factors, which were are also provided. These include: GDP per capita, family, life expectancy, freedom, government corruption, generosity, and dystopia score. World Bank Data Catalog provided data on a litany of development indicators for every country in the world. Each dataset could be downloaded as CSV, Excel, or XML files. We chose to use datasets on charges for the use of intellectual property (CSV), military expenditure as a percent of GDP (Excel), and arable land as a percent of total land area (CSV).

(T)ransform

CSV and Excel files provided on Kaggle and World Bank Data Catalog required cleaning and joining. MySQL, Pandas and Jupyter Notebook were used for that process. Table columns were dropped and renamed in all the 4 tables included in the project. Data was cleaned by dropping NAs and we scanned each column to make sure that there were similarities. In couple of datasets, there were columns with special characters as parentheses, which was manually cleaned to make the data runnable. The Primary key was set for each table in MySQL for all the tables. The column types were also set at MySQL. During the transformation stage of an XML dataset, we encountered some issues with parsing. To save time we decided to move forward without it. Given that we had four sources, we did not see much of an impact with the integrity of the ETL.

(L)oad

The relational database MySQL was used to load our data. That process was pretty smooth because all of the 4 tables were around couple of hundreds. Each member had their own files of python, jupyter and csv/excel. Therefore, we had to share these files to one member to combine them in a single database. Therefore, from four separate group of files, we ended up with one combined group of files. Depending on the tables we had to use different joins to find the matching link.

**Production Database Final Tables**

The final data table was created through a series of joins on the primary key “Country Name” and “Country Code”. We found one main database that covered world happiness through specific socio-economic indicators. As a team, we searched additional sources with other indicators for happiness. Each member selected factors that they feel strong towards and individually cleaned, transformed and imported to their own SQL databases. We then send the SQL file, Jupyter Notebook File, and CSV/Excel file to one member for the final presentation. The final table has a “Happiness Score” that will be used to see the relation and integrity of the additional columns.

The results include the following columns: “Country”, “Country Code”, “Happiness Rank”, “Happiness Score”, “Arable Land”, “Economy GDP Per Capita”, “Government expenditure on education, total (% of GDP)”, “Health Life Expectancy”, “Income Group”, “Intellectual Property Charges”, and “Military Expenditure”, “Trust Government Corruption”.

With the teamwork of Python, Jupyter Notebook, and Pandas we successfully imported the data into MySQL. We chose MySQL because our sources were similar and each source had common information.