**ETL DATA PROJECT –Group2 – TECHNICAL REPORT**

**1) Extract:** your original data sources and how the data was formatted (CSV, JSON, pgAdmin 4, etc).

Our group used Kaggle to identify data sources for this project. We began by finding data from the popular “World Happiness Report (WHR)” (<https://www.kaggle.com/unsdsn/world-happiness>). While our original dataset was shared by the authors of the report, upon further investigation we found other similar Kaggle sites with larger datasets, including additional variables, for the World Happiness project (<https://www.kaggle.com/anamvillalpando/world-happiness-ranking>). We verified that these were indeed correct by reviewing the findings from the 2020 WHR ( <https://happiness-report.s3.amazonaws.com/2020/WHR20.pdf> ). The 2020 WHR features data collected from the 2019 Gallup World Poll, a survey that assesses an individual’s perceptions of their life quality and social environment. Of central interest, participating countries are ranked according to the “happiness” of their citizens. We decided to combine the WHR data with population statistics from the World Bank, including total population and population categorized by age and by those who are employed (<http://wdi.worldbank.org/table/2.1>). Both data files were available as excel files for download.

**2) Transform**: what data cleaning or transformation was required

There were a number of transformations needed for the two datasets, which we planned to merge based on country name. Prior to importing the data, we reviewed the excel files and needed to adjust the World Bank file as it was formatted as a table. We adjusted the formatting for import so that the first row reflected variable column headings. We also noticed that the two datasets had a different number of records, with the World Bank file having more data than the WHR file. In particular, there were a number of records that would need to be deleted and some that needed renaming to be consistent with the WHR. Lastly, we added an index column to each file and assigned each record an id number.

We imported the data into 2 data frames via pandas. We adjusted each data frame by renaming and deleting a few columns, in particular ensuring that the country variable was the same in each. We also renamed a number of the countries in the World Bank data frame so that they could be consistent with the WHR data and merge easily via a join. We then did a left join on the two tables so that the final table only contained the World Bank data for those countries that participated in the World Happiness Project (n=153).

**3) Load**: the final database, tables/collections, and why this was chosen.

To be sure we could complete the final load step of this process, we added code at the beginning of our file to establish a connection with our MySQL database via AWS. As part of this, we also established a config file to import our AWS credentials and the name of database schema (“happiness\_bank”) that would house our new merged data table. After the data was loaded and merged via pandas in the extract and transform steps, we created a database schema in MySQL called “happiness\_bank” for our new merged data table. We then wrote the code to push the merged data table (also renamed happiness\_bank) to our MySQL happiness\_bank database schema (and the AWS cloud). The end of our file confirms that our data was successfully loaded to our MySQL happiness\_bank database as we display the first 50 rows of the happiness\_bank table. We chose these data as they were interesting and could make for fun visualizations and analysis in the future.