First, I decided visually to observe and delete any data that won’t give me any

insight. For example, in Global\_Alcohol\_Consuption\_by\_WHO dataset the following

column names ‘DisaggregatingDimension2’, DisaggregatingDimension2ValueCode’,

‘DisaggregatingDimension3’, ‘DisaggregatingDimension3ValueCode’ and ‘Comments’

don’t contain any data. I didn’t find any use for them. So, I deleted the five

columns. Whatever can be deleted manually should be deleted because it will make

it more efficient and easier for R to process.

In the second dataset named ‘HappinessAlcoholConsumption\_Kaggle’, I noticed

spelling mistakes. In column name ‘Hemisphere’, the word ‘north’ spelled as

‘noth’. It’s a mistake so I changed it too ‘north’.

In the third dataset named ‘lifeexpectancy-verbose\_Kaggle’, I filled out all

missing data for income group classification code, countries income group

display and world bank income code display. Corrected region for South Sudan as

African, instead of Eastern Mediterranean region. Additionally, after doing some

research online, classified South Sudan as low-income country instead of

middle-income category.

After manually cleaning the data, I decided to import all three

datasets and use code to check and clean the remaining data.

At this point I don’t know how to do advanced data manipulation. I need to learn advanced functions in the ‘dplyr’ package for data transformation. Another part that I don’t know at this point is the machine learning techniques. I would like to explore all the machine learning packages in ‘caret’, ‘randomForest’, and ‘xgboost’ to implement predictive models. Additionally, I need to gain proficiency in using ‘ggplot2’ package for creating advanced visualization to uncover patterns and insights of the data. Also, I need to learn how and which additional variables can be created to better capture important parts of the data and improve model predictive performance. Finally, I need to learn methods that will help me to analyze time-series data so I can better understand and observe trends and seasonality.

For my research project I would like to leverage machine learning techniques because it can provide good insights into the data and better predictive performance to answer various questions. With machine learning I can select relevant features for my model. Also, I can create new feature variables that might enhance the performance of predictive models.

I am planning to leverage different models such as regression models to explore linear relationships between different variables. Classification models like logistic regression could be used to classify data into categories to predict outcomes. Clustering techniques like K-means could be used to identify trends and then group all the data into similar categories.

I need to learn more about advanced data manipulation, visualization techniques, and machine learning methods. At this point it’s not easy identify the insights in the data that are immediately self-evident. However, using Exploratory Data Analysis (EDA) will help me to understand the patterns of the data. With the use of methods like summary() function to get summary statistics overview of each variable and using ‘ggplot2’ library package for data visualization I will identify patterns, outliers and relationships between variables.

Finally, I have additional questions such as.

What additional datasets can I leverage to enhance my analysis?

Which specific machine learning techniques would be best to implement that will improve my analysis?

What additional variables can I create that were not present in the datasets so that I can better capture important insights of the data and improve predictive power of my model?