7.1 Project 2: Project Draft/Milestone 3

Soukhna Wade  
Bellevue, USA

FALL 2020  
GitHub Portfolio URL: <https://souwadegit.github.io/>

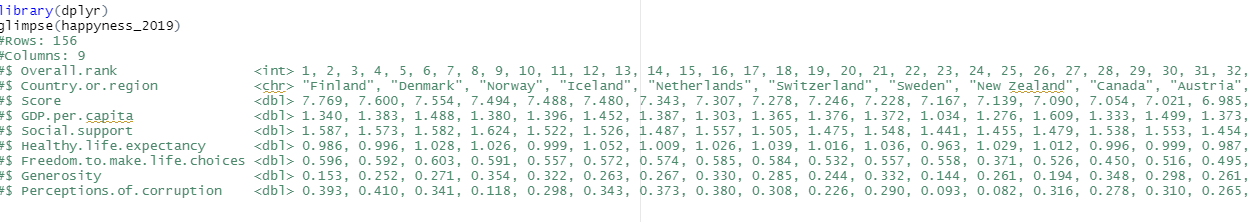
**DSC680 Applied Data Science 10/18/2020**

**7.1 Project 2: Project Draft/Milestone 3**

**About the Dataset**

**Data Sources: World Happiness Report - The dataset (world-happiness/2019.csv)**

**Link:** <https://www.kaggle.com/pinarkaya/world-happiness-eda-visualization-ml#2019-Data>



**Background**

Like the previous project, the dataset is under the Kaggle website. The World Happiness Report is a landmark survey of the state of global happiness. The dataset (World Happiness 2019), which ranks 155 countries by their happiness levels. Since the 1960s, scientific disciplines have researched happiness, to determine how humans can live happier lives. The scientific pursuit of positive emotion and happiness is the pillar of positive psychology, first proposed in 1998 by Martin E. P. Seligman.

* **Introduction**

The topic is about the world happiness from all the continent. For the first project, the dataset that I have chosen is the happiness 2019 dataset. This dataset provides the happiness score and happiness rank of 155 countries around the world based on some factors such as social support, health life expectancy, freedom to make life choices, generosity, and perception of corruption. The Sum of the value of these factors gives us the happiness score and the higher the happiness score, the lower the happiness rank. As a result, the higher value of each of these above seven factors mean the level of happiness is higher. Therefore, it is possible to define the meaning of these factors as the scale to which these factors lead to happiness. **Problem statement**

The purpose of this project is to find out which factors are more important to live a happier life. As a result, people and countries can focus on the more significant factors to achieve a higher happiness level. The goal of choosing this work is to find out which factors are more important to live a happier life. We also will implement several machine learning algorithms to predict the happiness score and compare the result to discover which algorithm works better for this specific dataset. In this case, it is evident to use the Random Forest method.

**Methods**

**World Happiness Understanding:**

In 2019, the World Happiness ranks 155 countries by their happiness levels. The report continues to gain global recognition as governments, organizations, and civil society increasingly uses happiness indicators to inform their policy-making decisions. Leading experts across fields (economics, psychology, survey analysis, national statistics, health, public policy, and more) describe how measurements of well-being can be used effectively to assess the progress of nations. The reports review the state of happiness in the world today and reveal how the new science of happiness explains personal and national variations in happiness.

**Data Understanding:**

Understanding the data available for our analysis is important before any modeling can be performed. This involved identifying those factors most likely to influence happiness. The dataset displays 156 rows or variables with 9 attributes or columns. I review the data to identify what attributes will be necessary and what data manipulation needs to be carried out before going through exploratory analysis, prediction modeling and random forest. For example, I am going to change the columns’ names and drop some of the columns. Python and R are the two programming languages I am going to use for this project, but I do not exclude Tableau and Power BI if need it.

**Exploratory Data Analysis and Data Preparation**

For this first project, I am going to focus on the following three different areas such as EDA, visualization, prediction, and random forest. For this project, I plan to implement Random Forest machine learning algorithms to predict the happiness score and compare the result to discover which algorithm works better for these specific datasets I have chosen.

From the data check, there were no missing values. After some preliminary exploration by using Python (Jupiter Notebook), these are the features we chose to include in this first project study:

* + Overall Rank (happiness Rank)
  + Country or region: Name of countries
  + Score: Happiness measurement on a scale of 0 to 10
  + GPA per capita (economy): Value of all final goods and services produced within a nation each
  + Social Support (Family): Importance of having a family
  + Health Life Expectancy(health): Importance of health and amount of time people expect to live
  + Freedom: Importance of freedom in each country
  + Generosity: The quality of being kind and generous
  + Perception of corruption(trust) in a government

**Questions arise from the dataset are:**

1. What year is the dataset?

This data is of year 2019.

1. How is the happiness score is distributed?

**The happiness score has values above 2.86 and below 7.77. So, there is no single country which has a happiness score above 8..**

1. What means overall rank?

It means happiness Rank.

1. What are the factors that contributed to the word happiness?

The factors are economy, social support, and health play the most significant role in contributing to happiness.

1. What makes people in a country happy?

Health is number one follows by meaningful work, positive thinking, and the ability to forgive.

1. What are the five first countries with high GDP?

The five first countries with high GDP are: Qatar, Luxemburg, Singapore, United Arab Emirates, and Kuwait.

1. Why generosity is important?

Generosity is important because it reduces stress, supports one's immune system and enhances one's sense of purpose. So, what is it about generosity that makes it so vital to a happy and healthy life? First, it is important to note that the form of generosity that most benefits us is not measured in a dollar amount or a physical gain.

1. What are the first five countries with high generosity?

The first five countries with high generosity: Myanmar, Indonesia, Haiti, Malta, and Kenya.

1. What means “Freedom to make life choices”?

It is the national average of responses to the question “Are you satisfied or dissatisfied with your freedom to choose what you do with your life?”

1. **Is corruption impact Happiness Score?**

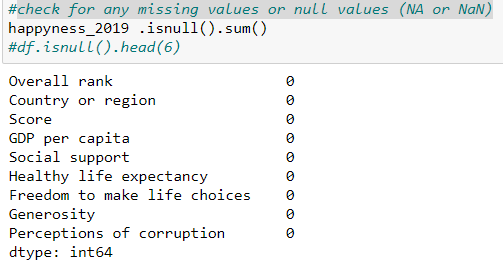
**Corruption is a very big problem for the Word happiness.**

1. What countries or regions rank the highest in overall happiness and each of the six factors contributing to happiness?

* **Cleaning Dataset**

Data cleansing or data cleaning is the process of detecting and correcting corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate, or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. Data cleansing may be performed interactively with data wrangling tools, or as batch processing through scripting. First, I load the required libraries and import my dataset using Pandas to check for missing or null values. Then I check for any missing values or null values (NA or NaN). In the following lines I used Pandas to determine if there are any missing values in the dataset (happiness\_2019).

Example:

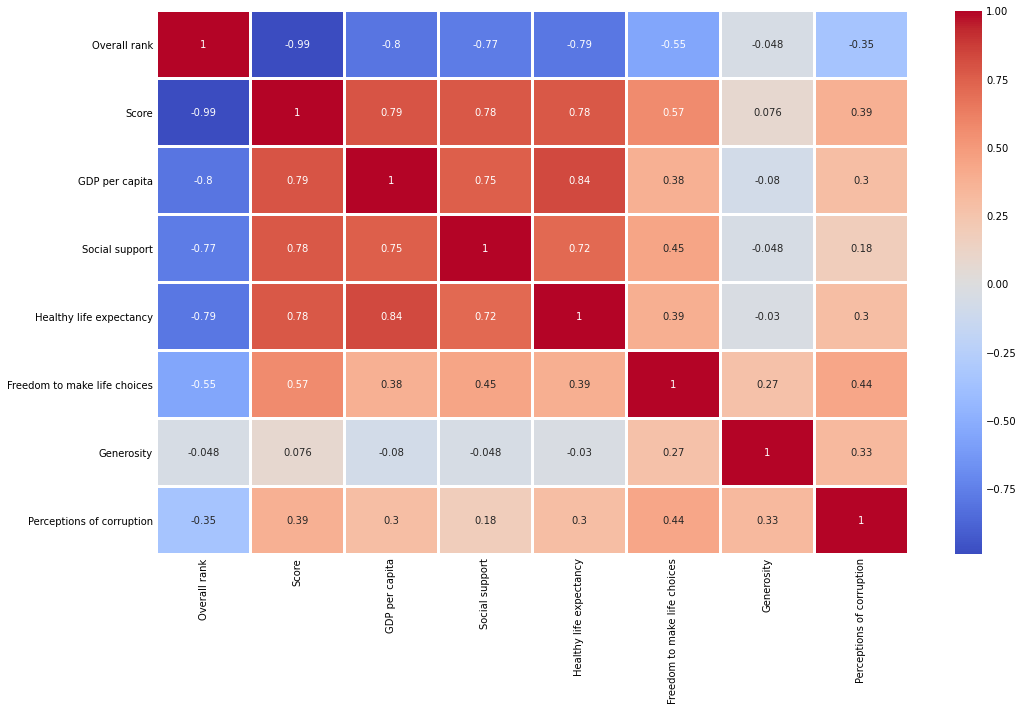


As you can see, in the below line that the dataset is cleaned. So, there is no need to clean it again.

* **Visualization**

visualization's purpose is the communication of data. Visualization transforms from the invisible to the visible. In this section, we use different variables to determine their correlation.

* + **Correlation plot between entire dataset variables**

****

Note that there is no correlation between two variables when correlation is zero or near zero and the color is gray. The darkest red means there is a perfect positive correlation, while the darkest blue means there is a perfect negative correlation.

The lower the overall rank, the higher the happiness score, and the higher the other five factors that contribute to happiness.

The above correlation plot shows that the economy, life expectancy, and family play the most significant role in contributing to happiness. Trust and generosity have the lowest impact on happiness scores.

* **Random forest**

The Random Forest is one of the most effective machine learning models for predictive analytics, making it an industrial workhorse for machine learning.

Random Forest is known as an ensemble machine learning technique that involves the creation of hundreds of decision tree models. These hundreds of models are useful to label or score new data by evaluating each of the decision trees and then determining the outcome based on the majority result from all the decision trees. It is a technique that can combine many ideas to create an accurate model that can learn from past data.



In conclusion, according to the 2019 World Happiness Report, negative feelings are rising around the world. The United States is particularly hard-hit with an “epidemic of addictions.” Also, we cannot exclude the pandemic that is still around the continents and hit hard in the USA. The report also shows a widening happiness gap, with some people reporting much more well-being, and others show much less within each country.

**References:**

1. https://www.kaggle.com/PromptCloudHQ/world-happiness-report-2019/activity
2. https://worldhappiness.report/archive/
3. https://www.kaggle.com/unsdsn/world-happiness
4. https://www.kaggle.com/dgtech/world-happiness-with-basic-visualization-and-eda
5. https://www.kaggle.com/pinarkaya/world-happiness-eda-visualization-ml/data#Linear-Regression
6. https://www.kaggle.com/javadzabihi/happiness-2017-visualization-prediction/report
7. https://www.kaggle.com/roshansharma/world-happiness-report?select=2017.csv
8. https://www.kaggle.com/pinarkaya/world-happiness-eda-visualization-ml#Random-Forest
9. https://www.kaggle.com/dhanyajothimani/basic-visualization-and-clustering-in-python/data
10. https://towardsdatascience.com/understanding-random-forest-58381e0602d2
11. https://turi.com/learn/userguide/supervised-learning/random\_forest\_regression.html
12. Statistical Appendix for Chapter 2 of the World Happiness Report 2020 John F. Helliwell, Haifang Huang, Shun Wang, and Max Norton February 29, 2020