4.4 Project 1: Project presentation/Milestone 4

Soukhna Wade

Bellevue, USA

FALL 2020

GitHub Portfolio URL: https://souwadegit.github.io/

DSC680 Applied Data Science **09/27/2020**

4.1 Project 1: Presentation/Milestone 4

About the Dataset

Data Sources: World Happiness Report

https://www.kaggle.com/javadzabihi/happiness-2017-visualizationprediction/report?select=2017.csv

Background

The dataset is under the Kaggle website. The World Happiness Report is a landmark survey of the state of global happiness. There are four reports published: the first in 2012, the second in 2013, the third in 2015, and the fourth in 2016. The dataset (World Happiness 2017), which ranks 155 countries by their happiness levels. It was released at the United Nations at an event celebrating the International Day of Happiness on March 20th.

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**

  I found the datasets from the website called Kaggle, and the topic is about the world happiness from all the continent. For the first project, the dataset that I have chosen is the

happiness 2017 dataset. This dataset provides the happiness score and happiness rank of 155 countries around the world based on seven factors. These factors are involving family, life expectancy, economy, generosity, freedom, trust in government, and dystopia residual. The Sum of the value of these seven 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. Here, we talked about the term dystopia, which is the opposite of utopia and has the lowest happiness level. Dystopia will be considered as a reference for other countries to show how far they are from being the poorest country regarding happiness level.

Dystopia is a world in which everything is imperfect, and everything goes wrong. Dystopian literature shows us a nightmarish image about what might happen to the world soon. Usually the main themes of dystopian works are rebellion, oppression, revolutions, wars, overpopulation, and disasters. On the other hand, utopia is a perfect world – exactly opposite of dystopia.

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. As a result, people and countries can focus on the more significant factors to achieve a higher happiness level. 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.

Methods

World Happiness Understanding:

In 2017, 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 155 rows or variables with 12 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 and prediction modeling. 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, and prediction. For the next project, I plan to implement several 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:

- Country: Name of countries
- Happiness. Rank: Rank of the country based on the Happiness Score
- Happiness Score: Happiness measurement on a scale of 0 to 10
- Economy: Value of all final goods and services produced within a nation each
- Family: Importance of having a family
- Life Expectancy: 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
- Trust: Perception of corruption in a government
- Dystopia Residual: Plays as a reference
- Continents

Questions arise from the dataset are:

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

   When it comes to happiness, there are many factors that affect it. Money, family and friends, health, freedom, and generosity are just the five main factors. Other factors include corruption, education, and weather.
2. What makes people in a country happy?

Close behind are meaningful work, positive thinking, and the ability to forgive. What does not seem to make people happy are money, material possessions, intelligence, education, age, gender, or attractiveness. In rough order of importance, here are the factors that make us happy and what you can do to increase happiness in your life.

3. What is Dystopia?

   A dystopia is a community or society that is undesirable or frightening. It is an antonym of utopia, a term that was coined by Sir Thomas More and figures as the title of his best known work, published in 1516, which created a blueprint for an ideal society with minimal crime, violence and poverty.

4. What are the residuals?

   In statistical models, a residual is the difference between the observed value and the mean value that the model predicts for that observation. Residual values are especially useful in regression and ANOVA procedures because they indicate the extent to which a model accounts for the variation in the observed data. It is a deviation from an observed value of any sample set from its estimated value. Put more precisely, a residual (e) is the difference observed in the predicted function value ($\hat{y}$) and the final observed value (y) of a dependent variable

5. What is the RMSE?

   The root-mean-square deviation (RMSD) or root-mean-square error (RMSE) is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed. The RMSD represents the

square root of the second sample moment of the differences between predicted values and observed values or the quadratic mean of these differences.

6. What would be the predicted coefficients?

```
coeffecients = pd.DataFrame(lm.coef_,X.columns)
coeffecients.columns = ['Coeffecient']
coeffecients
```

	Coeffecient
Economy	1.000016
Family	0.999884
Life.Expectancy	1.000109
Freedom	1.000070
Generosity	1.000102
Trust	0.999772
Dystopia.Residual	0.999935

7. What countries or regions rank the highest in overall happiness and each of the six factors contributing to happiness? Answer: Finland
8. What are the Ranks top ten?

   Finland, the Netherlands, Canada, New Zealand, Australia, and Sweden
9. How did country ranks or scores change between the 2015 and 2016 as well as the 2016 and 2017 reports?
10. Did any country experience a significant increase or decrease in happiness?

    South Sudan
11. Does happiness impact an organization?

12. The employees' level of happiness impacts the potential of success for an organization

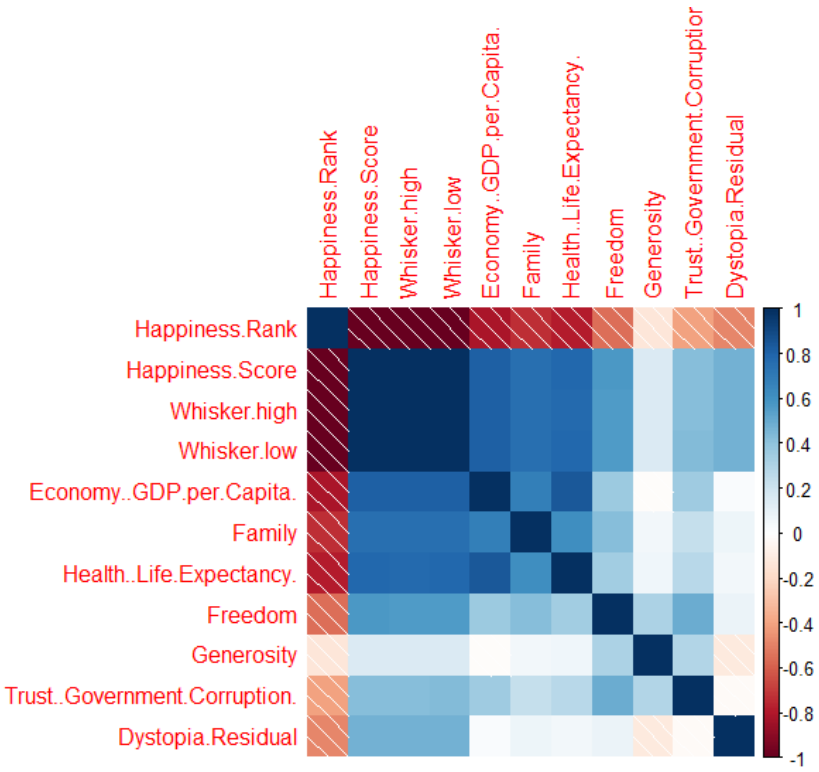
- # Cleaning Dataset

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 my dataset using Pandas to check for missing or null values. For instance, you can find in the below line that the dataset is cleaned. So, there is no need to clean it again. In the following example I used Pandas to determine if there are any missing values in the dataset (happiness_2017).

- # Visualization

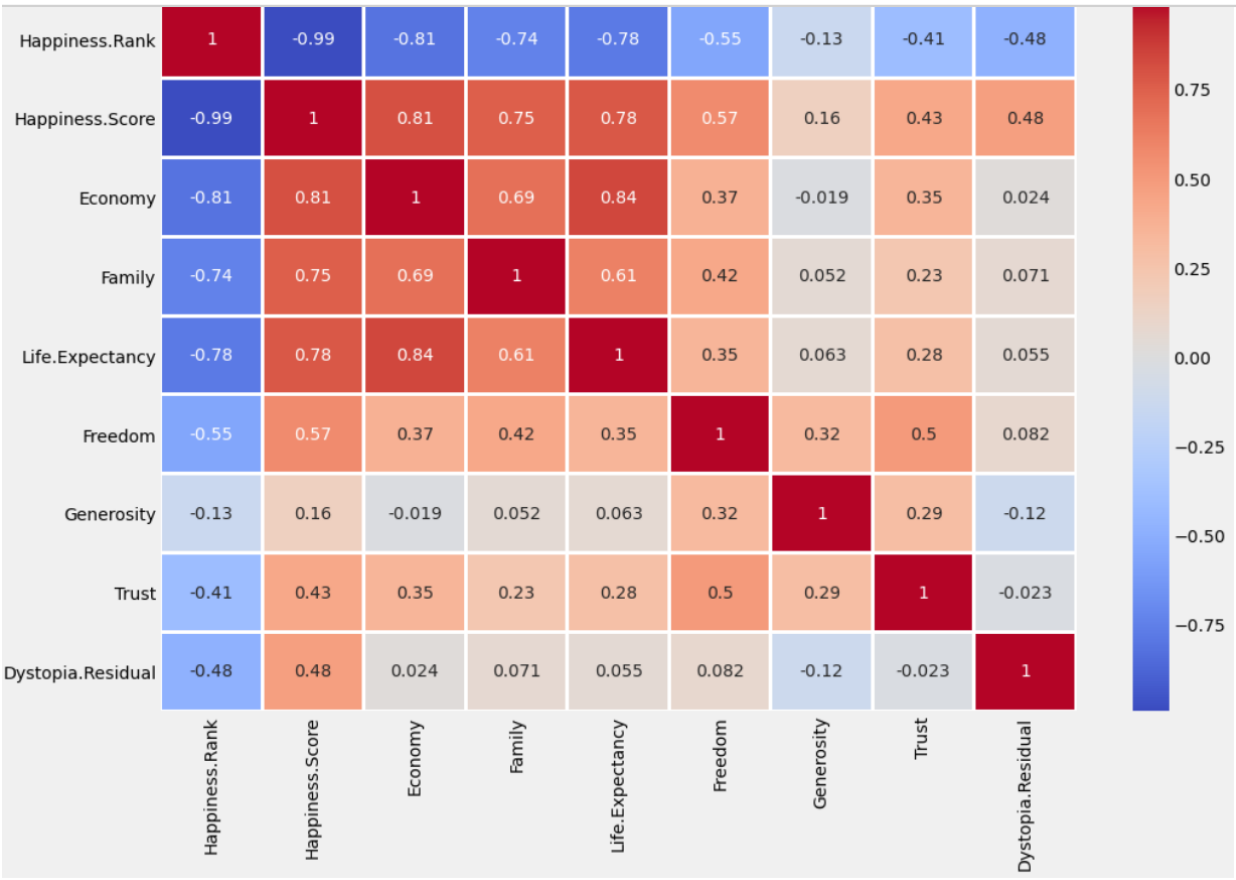
visualization's purpose is the communication of data. Visualization transforms from the invisible to the visible. In this section, we use different variables to determine their correlation.

1. **Correlation plot between numerical variables**



Note that there is an inverse correlation between the "Happiness Rank" and all the other numerical variables. The lower the happiness rank, the higher the happiness score, and the higher the other seven factors that contribute to happiness.

2. **Correlation after removing two variables (Whisker. High and Whisker. Low)**



The map visualizes the correlation values between happiness scores and the factors that contribute to happiness scores. It shows that there is a direct positive correlation between the happiness score of a country's economy, family, and health/ life expectancy.

The above heatmap of correlation among the variables displays the color palette in the side represents the amount of correlation among the variables. Therefore, lighter shade represents a high correlation. You can see that the happiness score correlated with the economy, family, and life expectancy. It is least correlated with generosity.

3. **Correlation plot after removing happiness Rank:**

By creating the following correlation without using Happiness Rank, we found 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.

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.

- # Prediction

I implement several machine learning algorithms to predict happiness scores. First, I split our dataset into training and test set. The dependent variable is happiness score, and the independent variables are economy, family, life expectancy, freedom, generosity, trust, and dystopia residual.

```
coeffecients
```

	Coeffecient
Economy	1.000016
Family	0.999884
Life.Expectancy	1.000109
Freedom	1.000070
Generosity	1.000102
Trust	0.999772
Dystopia.Residual	0.999935

The above result shows that there is a positive correlation. This indicates that when the predictor variable increases, the response variable will also increase.

Conclusion:

The summary shows that all independent variables have a significant impact, and the adjusted R squared is one. There is a linear correlation between dependent and independent variables. Also, the sum of the independent variables is equal to the dependent variable which is the happiness score. This is the justification for having an adjusted R squared equal to one. As a result, multiple Linear Regression will predict happiness scores with 100 % accuracy.

References:

1. https://worldhappiness.report/archive/
2. https://www.kaggle.com/unsdsn/world-happiness
3. https://www.kaggle.com/pinarkaya/world-happiness-eda-visualization-ml/data#Linear-Regression
4. https://www.kaggle.com/javadzabihi/happiness-2017-visualization-prediction/report
5. https://www.kaggle.com/roshansharma/world-happiness-report?select=2017.csv
6. https://www.kaggle.com/pinarkaya/world-happiness-eda-visualization-ml#Random-Forest
7. https://www.kaggle.com/dhanyajothimani/basic-visualization-and-clustering-in-python/data
8. https://towardsdatascience.com/understanding-random-forest-58381e0602d2
9. https://en.wikipedia.org/wiki/The_Happiness_Hypothesis
10. https://www.theworldcounts.com/happiness/the-definition-of-happiness-in-psychology
11. Statistical Appendix for Chapter 2 of the World Happiness Report 2020 John F. Helliwell, Haifang Huang, Shun Wang, and Max Norton February 29, 2020

https://www.cnn.com/travel/article/worlds-happiest-countries-united-nations-2017/index.html

https://www.cnn.com/travel/article/worlds-happiest-countries-united-nations-2017/index.html