­­Celine Cherian: Data Engineer

Miguel Mendoza: Data Scientist

7/23/2020

Data Preparation and Analysis

World Happiness Report Datasets: <https://www.kaggle.com/unsdsn/world-happiness>

Data Preprocessing for Machine Learning: <https://www.kaggle.com/getting-started/167283>

[**Skip to Celine’s Initial “Preprocessing” below, Click here!**](#_vaip2x2kh38s)

## **Celine’s Observations:**

* 2015 Data is 159 rows by 12 columns - [159, 12]
  + Headings: Country, Region, Happiness Rank, Happiness Score, **\***Standard Error, Economy (GDP per Capita), Family, Health (Life Expectancy), Freedom, Trust (Government Corruption), Generosity, Dystopia Residual
* 2016 is [158, 13]
  + Headings: Country, Region, Happiness Rank, Happiness Score, **\***Lower Confidence Interval, **\***Upper Confidence Interval, Economy (GDP per Capita), Family, Health (Life Expectancy), Freedom, Trust (Government Corruption), Generosity, Dystopia Residual
  + ***Standard Error*** column is gone and replaced with ***Lower/Upper Confidence Interval*** *columns* from 2015 to 2016 (don’t know what confidence represents)
* 2017 is [156, 12]
  + Headings: Country, Happiness Rank, Happiness Score, **\***Whisker.high, **\***Whisker.low, Economy (GDP per Capita), Family, Health (Life Expectancy), Freedom, **\***Generosity, **\***Trust (Government Corruption), Dystopia Residual
  + ***Region*** column is gone, ***Trust (Government Corruption)*** and ***Generosity*** columns have switched. Also, ***Lower/Upper Confidence Interval*** columns got replaced with ***Whisker.high/Whisker.low*** (don’t know what whisker represents) from 2016 to 2017.
* 2018 is [157, 9]
  + Headings: Overall Rank, Country or Region, Score, GDP per Capita, Social Support, Healthy Life Expectancy, Freedom to make life choices, Generosity, Perceptions of corruption
  + ***Happiness Rank*** is replaced with ***Overall Rank***, ***Country or Region*** column combines ***Country and Region***, ***Happiness Score*** is replaced with ***Score*** column, ***Economy (GDP per Capita)*** renamed to ***GDP per Capita***, ***Family*** is replaced with ***Social Support***, ***Health (Life expectancy)*** is replaced with ***Healthy Life Expectancy***, ***Freedom*** is replaced with ***Freedom to make life choices***, ***Trust (Government Corruption)*** is replaced with ***Perceptions of corruption***. ***Dystopia Residual*** and ***whisker.high/whisker.low*** columns are gone.
* 2019 is [157, 9]
  + Headings and heading replacements that are needed are the same as 2018.

## **Celine’s Initial Preprocessing (to make sure all datasets are consistent and mergeable):**

* Change the headings of all datasets to: **Country or Region, Happiness Rank, Happiness Score, Economy (GDP per Capita), Social Support, Health (Life Expectancy), Freedom to make life choices, Generosity, Perceptions of Corruption (this is 9 columns).** You do so by (look below to see needed changes):
  + 2015 and 2016 changes:
    - Remove “region”, “standard error”, and “dystopia residual” columns in feature selection (or in code)
      * **2016**: Remove “region”, “lower confidence interval”, “upper confidence interval”, and “dystopia residual” columns in feature selection (or in code)
    - Switch “Trust (Government Corruption)” and “Generosity” columns (or make sure you refer to them properly in code)
    - Rename “Trust (Government Corruption)” to “Perceptions of Corruption”, “Country” to “Country or Region”, “Family” to "Social Support”, and “Freedom” to “Freedom to make life choices”
  + 2017 changes:
    - Remove “whisker.high”, “whisker.low”, and “dystopia residual” columns in feature selection (or in code)
    - Rename “Trust (Government Corruption)” to “Perceptions of Corruption”, “Country” to “Country or Region”, “Family” to "Social Support”, and “Freedom” to “Freedom to make life choices”
  + 2018 and 2019 changes:
    - Switch “Country or Region” and “Overall rank” columns (or make sure you refer to them properly in code)
    - Rename “Overall rank” to “Happiness Rank”, “Score” to “Happiness Score”, “GDP per Capita” to “Economy (GDP per Capita)”, and “Healthy Life Expectancy” to “Health (Life Expectancy)”
* Only **NaN values** I expect to show are data from the “Country or Region” column since the data is all names of countries/regions and would be object type. I also see no missing values (currently)

## **How to visually display data after datasets are finally consistent:**

* Need to do scatter plot of the data and multiple linear regression to find any outliers and possibly remove to increase accuracy.
* May want to visually display the data as a map or output as a selection so users can select which region and which country data they want to see
* Display a bar graph that compares countries within a region of their data (like who has the highest happiness rank in the region)
* May also have a way for users to see who has the highest happiness rank in the world

# **Hypothesis:**

We will be applying multilinear regression on the data with “Happiness Score” as our target variable and Economy, Social Support, Health, Freedom, Generosity, and Corruption as our features. First, we will use sklearn to split our data into validation and training data. We will be using gradient descent as our optimization technique, along with Ridge and LASSO modifiers to improve optimization. We will compare Ridge and LASSO and select the one that produces the best fit using sklearn’s mean absolute error function to compare our prediction data with the validation data.