DSC 520 Final Project Step 3

Taylor Imhof

11/19/2021

Project Narrative

The overall objective of this EDA project was to identify which factors contribute most to the likelihood of being happy. As noted in the first step, happiness is innately as subjective state of being. As such, finding statistical factors that contribute to being happy is a tough endeavor. However, after searching around for potential datasets, I found a few that I thought would help me answer this question.

To frame the project, I acted as though I was an advisor hired by an imaginary country's policy makers to help them determine what measures could be put in place to help increase their population's overall happiness. To simulate a business understanding, I was directed with an overarching question of, "what societal factors contribute most to a person's perceived level of happiness?" I anticipated being able to identify relationships between my selected explanatory and target variables.

Project Methodolgy

Data Sets The first step after the simulated business understanding was finding an appropriate data set. I wound up finding a few useful data sets, two of which I ultimately wound up merging into a combined data set. The first data set was derived from the findings of the U.N.'s World Happiness Report (WHR). Metrics contained in this data set included a ladder score (target variable) and various societal conditions such as GDP (economic prosperity), life expectancy, social support, trust in government, as well as generosity.

The other data set I decided to merge with the WHR was a dataset containing results from the Human Development Index (HDI). The dataset contained similar columns (life expectancy and GDP) but had additional columns like GNI, expected years of education, and HDI that I felt would provide additional insights.

The final data set I decided to include in this project was a data set containing the results of a survey conducted by the person who maintains the data set. Metrics included were general lifestyle information such as fruits/veggies consumed, stress levels, places visited, sense of achievement, and social support amongst others. I felt that in addition to the global scope of the first two data set, this one could provide insight as to what people could be doing at an individual level to increase their level of happiness.

All of my data sets were retrived from the data sets section of Kaggle.com

Dataset 1 Link World Happiness Report

Dataset 2 Link Human Development Index

Dataset 3 Link Lifestyle and Wellness Data

Data Import and Cleaning After reading in the data using the built-in read.csv() function for all of my data sets. Then I used the str() function to get a feel for how the data in each set were codified to see if I needed to change any data types. I also did not like how some of the column were named, so I changed them right away. There were a few columns whose data types needed to be converted, and I achieved this by using the as.numeric() function. I was also able to practice replacing NA values with imputed mean values, which was something I found difficult when learning about it during the course. There was also a column in the lifestyle data set that contained string values for male and female. I was able to convert these to binary values, in case I wanted to do analysis on them such as logistic regression.

The next thing I did was combine the WHR and HDI data frames on each other using the 'country' column to match the observations. There was also a column that should have been numeric, but since it contained commas, I had to use gsub() to remove them, and then cast to numbers using as.numeric().

Dealing with NA/Unwanted Values The next step was to account for missing or null values. I was able to use a slick implementation of the map_df() function from the purrr package to iterate across all of my data frames and view the count of missing values. I opted to impute the mean for all of the missing values, as I felt it would not impact the accuracy of my planned analysis. The final thing I did to get my final data sets was drop a few columns that I did not feel would be useful for the project.

Data Understanding and Visualization The next step was to get a feel for the distributions of my variables. There were two target variables that I identified in the previous step (WORK_LIFE_BALANCE_SCORE and Score) that I checked for normality. I found it quite interesting that the distribution of the WORK_LIFE_BALANCE_SCORE was almost perfectly normal.

To identify potential predictors, I ran a correlation test via the cor() function, passing in an argument for Pearson's r. I was able to select four variables to focus on for predicting the target score for my combined data frame. I then graphed histograms for these four values to get an understanding for their underlying distributions. I messed around with performing some transformations on some of the skewed data, such as the log and square root transformations.

I performed the same correlation test on the lifestyle data frame and was again able to come up with four potential predictors with strong correlation coefficients. After plotting their respective histograms and identifying potential outliers or variance, it was time to test for relationships.

Scatter Plots and Trend Analysis For each of my selected explanatory variables, I created a scatter plot against each dataframes respective target variable. A nice feature of ggplot is the ability to fit a regression line via geom_smooth(). As I had expected, for all of my selected features, I found indications of significant positive relationships.

Interpreting the Results Referring back to the driving data science question of which factors contribute most to an increased level of perceived happiness, I was able to identify and test several conditions for statistical significance. As an advisor to the policy makers, I would be sure to present my findings in a manner that would be most effective at explaining the results.

Summarizing my findings from the combined data frame of the WHR and HDI, I found that the strongest indications of a populations happiness include a strong sense of social support, economic prosperity (indicated by GDP metric), and life expectancy. As such, I would advice the policy makers to invest heavily in developing strong social programs, especially with regard to health care. Moreover, implementing programs that help the population prosper economically will likely result in a happier populace.

Summarizing my findings from the lifestyle data frame, I found that the strongest life factors that result in higher scores of work/life balance include places visited, sense of achievement, supporting others, and time for passions. As such, I would adivse the policy makers in investing in programs that allow citizens to travel to different locations for service projects. Not only would this likely increase the support and travel metrics, people will likely discover passions in the new locations as well as finding a sense of achievement at the end of the service projects.