Statistical Data Analysis Report

Introduction

This capstone project seeks to identify additional indicators of a country's World Happiness Score. The factors investigated in this report are: gender disparities (development and inequality), economic freedom and population density. Outlined below, are the steps taken during the statistical analysis of the 3 datasets. Please see the accompanying jupyter notebook for a detailed overview of the code, calculations and plots referenced below.

Collinearity

It is important to check for any collinearity between the feature variables to ensure that the regression coefficient is uniquely determined. Any variables with a correlation greater than 0.8 would be undesirable. From the heat map and table below it can be determined that collinearity is not an issue with this dataset.

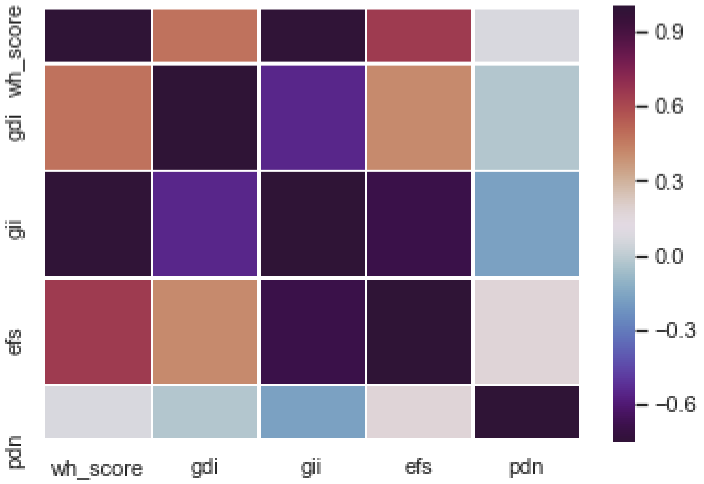

Figure 1: Correlation heat map

	GDI	GII	EFS	PDN
GDI	1.00	-0.546	0.416	-0.0215
GII	-0.546	1.00	-0.686	-0.167
EFS	0.416	-0.686	1.00	0.175
PDN	-0.022	-0.167	0.175	1.00

Exploration

During the data storytelling portion of this project, it was noted that several of the variables' distributions were skewed and did not have a normal distribution. Population density also had some extreme values which could affect the calculation of a Pearson correlation coefficient. For these reasons, this statistical analysis used Spearman's rank correlation coefficient. It is robust enough to handle extreme values and scenarios where one or both of the variables are not normally distributed. Figure 1 details the correlation coefficient, slope and intercept calculated for each predictor.

Based on the initial statistical calculations, there appears to be a strong negative relationship between World Happiness and Gender Inequality. Economic Freedom and Gender Development also have relatively significant relationships between World Happiness scores. There does not appear to be a significant correlation between Population Density and World Happiness.

Predictor	R	Slope	Intercept
Gender Development	0.479	5.67	0.128
Gender Inequality	-0.751	-4.31	6.94
Economic Freedom	0.416	0.07	1.01
Population Density	-0.022	0.00	5.43

Figure 2: Correlation Coefficient, Slope and Intercept Table

The null hypothesis for all four predictors is: there is no correlation between them and world happiness scores. Using the Spearman correlation coefficient as my test statistic, I concluded that a permutation test would be the best choice to simulate the data. I conducted a permutation on the predictor variables and left the world happiness scores fixed to generate a new set of data. This uses all the data and eliminates any correlation because the happiness scores and predictor value pairs are shuffled. The data was replicated 10,000 times and a correlation coefficient was calculated for each. The figures below display the distribution of correlation coefficients for the replicate data.

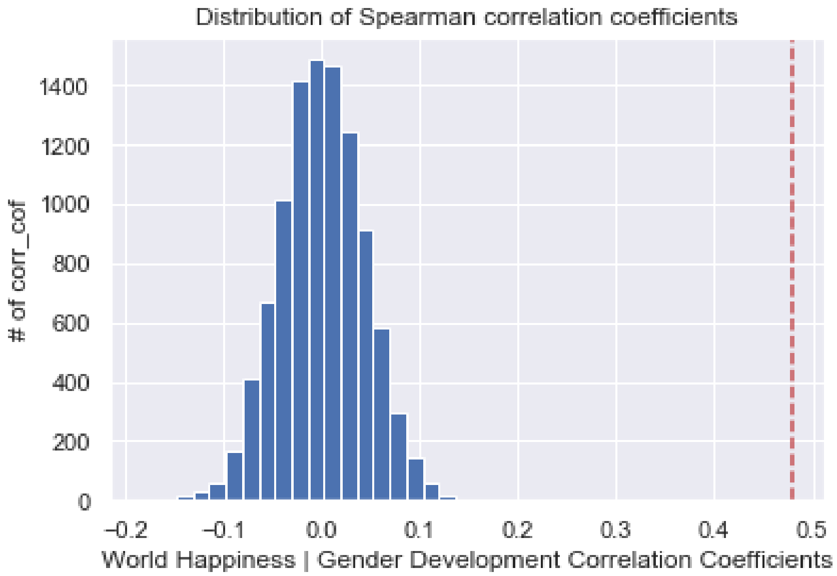

Figure 3: Distribution of Gender Development correlation coefficients

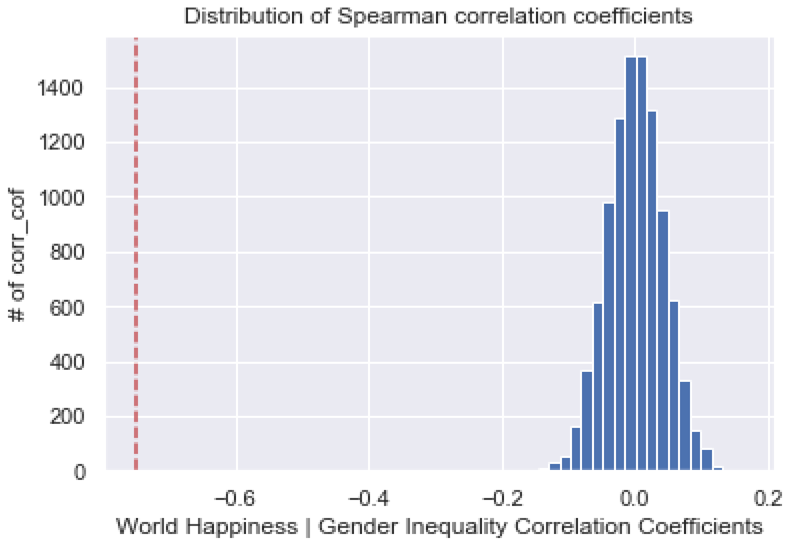

Figure 4: Distribution of Gender Inequality correlation coefficients

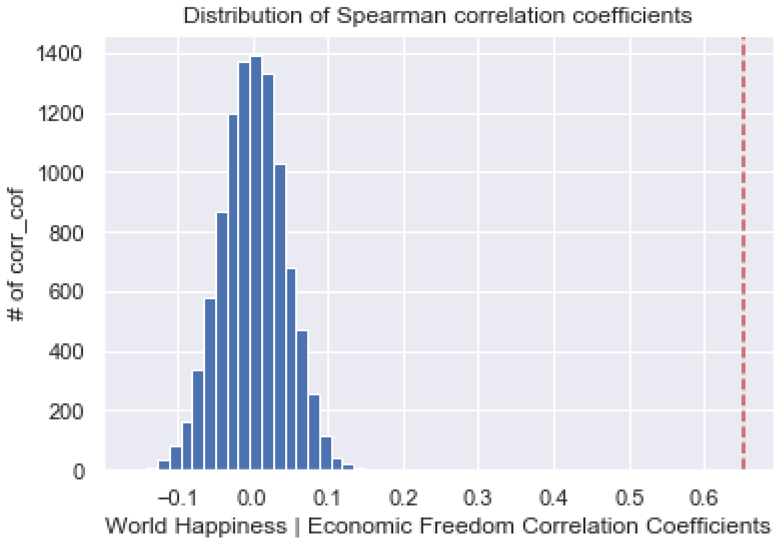

Figure 5: Distribution of Economic Freedom correlation coefficients

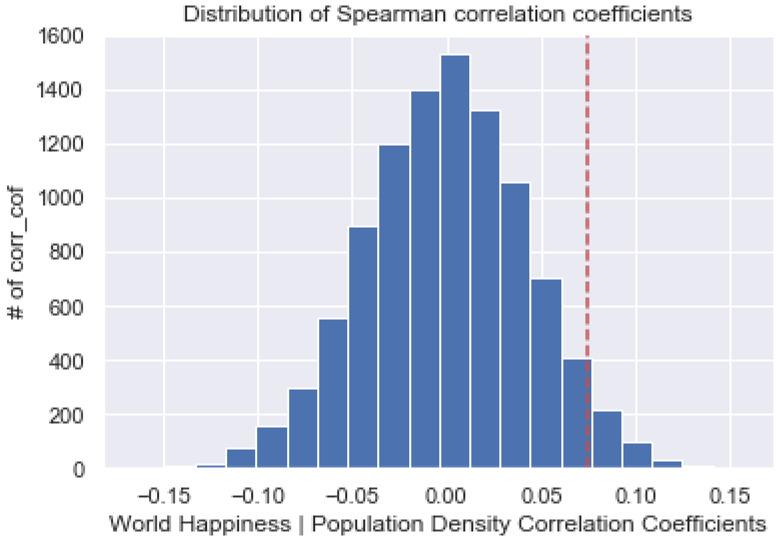

Figure 6: Distribution of Population Density correlation coefficients

Conclusion

The p-value in the Population Density test was above 0.01 and therefore we will accept H_0. The tests for Gender Development, Gender Inequality and Economic Freedom had a p-value of 0.0 which is statistically significant and we can reject H_0. We should be able to utilize these three factors to make predictions concerning a country's happiness score.