Final Report:
Predictors for World Happiness Score

Introduction

Objective

The World Happiness Report uses 6 factors: GPD Per Capita, Life Expectancy, Social Support, Freedom, Generosity and Corruption to explain inter-country variances in happiness. This report investigates if there are additional measures that can be used to explain those differences. Using data from the World Bank, United Nations and The Heritage Foundation that contain statistics on a myriad of topics from gender inequality to population density to economic freedom, the goal of this project is to identify additional indicators for happiness.

Target Audience

The more factors we have to better explain variations in happiness, the more effective toolkit governments, politicians and organizations have to make more informed policy decisions.

Data Source

World Happiness

The Sustainable Development Solutions Network (SDSN) has published the World Happiness Report yearly since 2012. The scores and rankings for the countries in the report were based on data from the Gallup World Poll.

Gender Development and Inequality

The United Nations Development Programme produces a report that includes two datasets with statistics on the disparities between the two genders based on various socioeconomic factors..

Population Density

The dataset was sourced and extracted from the World Bank. The full dataset has over 50 years of data from 155 countries.

Economic Freedom

The data from 186 countries was distributed by The Heritage Foundation. The scores are based on 12 freedoms (i.e. property rights, business freedom, monetary freedom, etc.).

Data Wrangling

Raw Data and Content Structure

All datasets were in either CSV or XLS format. Each row in every dataset contains scores, factors, and identifying information from a unique country. The columns included for each dataset are:

<u>World Happiness (Figure 2.2 tab)</u>

- Country
- Happiness score
- Whisker-high
- Whisker-low
- Explained by: GDP per capita
- Explained by: Social support
- Explained by: Healthy life expectancy
- Explained by: Freedom to make life choices
- Explained by: Generosity
- Explained by:
- Perceptions of corruption
- Dystopia (2.33) + residual

<u>Gender Development</u>

- HDI Rank (2018)
- Country
- 1995
- 2000
- 2005
- 2010
- 2011
- 2012
- 2013
- 2014
- 2015
- 2016
- 2017
- 2018

<u>Gender Inequality</u>

- HDI Rank (2018)
- Country
- 1995
- 2000
- 2005
- 2010
- 2011
- 2012
- 2013
- 2014
- 2015
- 2016
- 2017
- 2018

Population Density

- Country Name
- Country Code
- Indicator Name
- Indicator Code
- 1960
- 1961
- 1962
- 1963
- 1964
- 1965
- 1966
- 1967
- 1968
- 1969
- 1970
- 1971
- 1972
- 1973
- 1974
- 1975
- 1976
- 1977
- 1978
- 1979
- 1980
- 1981
- 1982
- 1983
- 1984
- 1985
- 1986
- 1987
- 1988
- 1989
- 1990
- 1991
- 1992
- 1993
- 1994
- 1995
- 1996
- 1997
- 1998
- 1999
- 2000
- 2001
- 2002
- 2003
- 2004
- 2005
- 2006
- 2007
- 2008
- 2009
- 2010
- 2011
- 2012
- 2013
- 2014
- 2015
- 2016
- 2017
- 2018

Economic Freedom

- CountryID
- Country Name
- WEBNAME
- Region
- World Rank
- Region Rank
- 2019 Score
- Property Rights
- Judicial Effectiveness
- Government Integrity
- Tax Burden
- Govt Spending
- Fiscal Health
- Business Freedom
- Labor Freedom
- Monetary Freedom
- Trade Freedom

- Investment Freedom
- Financial Freedom
- Tariff Rate (%)
- Income Tax Rate (%)
- Corporate Tax Rate (%)
- Tax Burden % of GDP
- Govt Expenditure % of GDP
- Country
- Population (Millions)
- GDP (Billions, PPP)
- GDP Growth Rate (%)
- 5 Year GDP Growth Rate (%)
- GDP per Capita (PPP)
- Unemployment (%)
- Inflation (%)
- FDI Inflow (Millions)
- Public Debt (% of GDP)

Data Wrangling Steps

Note: Please reference my github account for a detailed overview of data wrangling code and steps (http://bit.ly/33try77)

1. Download Data

   All of the datasets were downloaded from their respective site and stored in my Capstone Project directory.

2. Raw Data Inspection

   Since the majority of my data was sourced from UN based sources, it is relatively organized and in a logical and tidy format. There are column headers and each country is represented by a row. Since I am mostly concerned with the countries and their corresponding score for years 2015-2018, much of the columns in each dataset I won't be using.

3. Import Data

   For XLS files, I used Pandas read_excel function with the sheet_name and usecols parameter to load the datasets as a dataframe from a specific sheet and the specific columns that I needed. CSV files were imported as dataframes using the read_csv function in Pandas.

4. Inspect in Jupyter

   Using the head, tail, and info methods I can see that datasets have differing numbers of countries and null values in several cells.

5. Cleaning and Merging Data
   a. Since I need to merge the dataframes based on Country, I wanted to ensure that there were no leading or trailing spaces in the Country columns that would complicate combining the dataframes. I used the strip() method to accomplish this.

b. I also wanted to rename all columns so that there are no spaces in the name to facilitate filtering and slicing
c. The world happiness and economic freedom datasets had to be downloaded individually by year. I wanted to combine all the years for the respective feature into one single dataset to make concatenation easier.
d. In both the World Happiness and Economic Freedom datasets, a few of the country names varied by year. For example in 2015 Trinidad and Tobago was spelled with “&” instead of “and”. I updated the handful of countries in each dataset with the same naming convention.
e. Once the country names were consistent, I merged the years for the respective features together. Each dataframe had a Country column and 4 additional columns for the score for each year.
f. I removed all empty rows using the .drop() method.
g. Since I knew that I wanted to do some analysis based on region and create a map of the scores, I added ISO A3 codes and regions to the world happiness dataset since I planned to merge the other scores to that one. I did this importing the ISO A3 spreadsheet as a dataframe and merging it to my world happiness dataframe.
h. The gender development and gender inequality datasets had multiple empty columns. I used the .dropna() method to remove those. Additionally, several empty cells were annotated with “..”. I used the .replace() method to substitute NaN in those cells.
i. For my purposes, the population density dataset did not need any cleaning or merging at this point. The columns were named appropriately so that I could merge the ones I needed with the rest of my features.

6. Combining Dataframes
   a. I created a function to do the following for each year:
      i. I sliced the 'Country' column in each dataset I needed to merge so that I can compare and remove any countries that are not in the corresponding datasets.
      ii. I chained the concat() and drop_duplicates() methods to review the countries that have no match in the World Happiness dataframes. I also used the keys parameter to create a hierarchical index that allowed me to easily inspect the dataframe and identify which index was associated with their respective dataframe.
      iii. After inspection, I created a list that explicitly identifies the locations of all country names in the dataframe that need to be changed to match the names in the World Happiness dataframe. I have a function in dwfunctions.py to update the countries in the dataframe to match the format of the World Happiness dataframe.
      iv. Next, I removed the countries that do not match in both dataframes by slicing the dataframe using the multiindex and the drop() method.
      v. I then merged the dataframes on the Country column for each feature.

        vi. Lastly, I added a year column

   b. Finally once all the data was merged by year, I concatenated all the years into one dataframe
   c. I used the info method to ensure columns had no empty values. If there were, I removed that row.

7. Export

I exported the cleaned and combined dataframe using Pandas to_csv and setting the index parameter to False, so that when I import the file for further analysis it does not include an extra column

Exploratory Analysis

Note: Please see the project's github account for a detailed overview of the visual analysis including interactive maps (https://bit.ly/39cSdGF)

Data analysis was composed of various visualizations and inferential statistics. The purpose was to identify correlations between the 4 factors listed above and world happiness.

Visual Analysis

Gender Development | Economic Freedom | World Happiness

Grouping the data by region it was noted that Southern Asia, Northern Africa and Sub-Saharan Africa have the lowest average World Happiness, Gender Development Scores and Economic Freedom Scores. This is also observed in the maps plotted for each score.

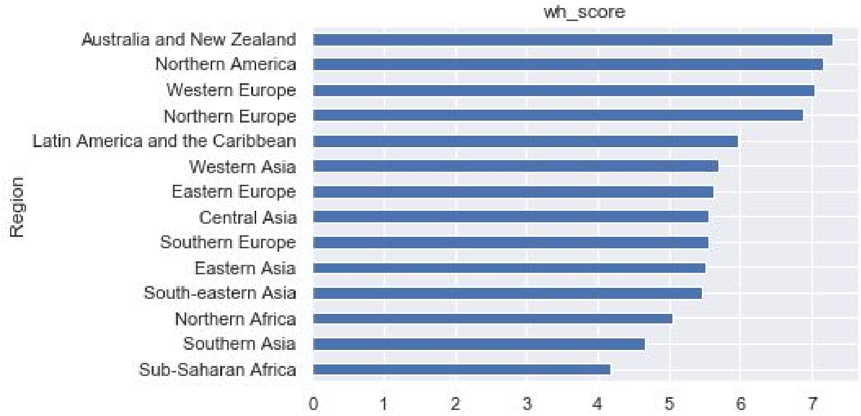


Figure 1: World Happiness Average by Region

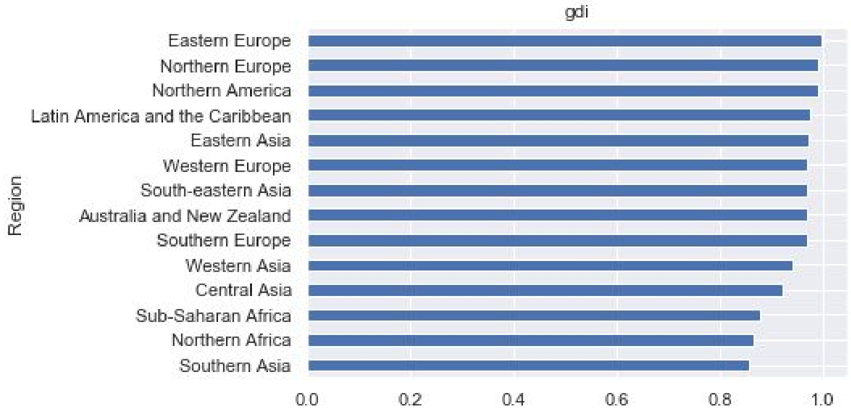


Figure 2: Gender Development Average by Region

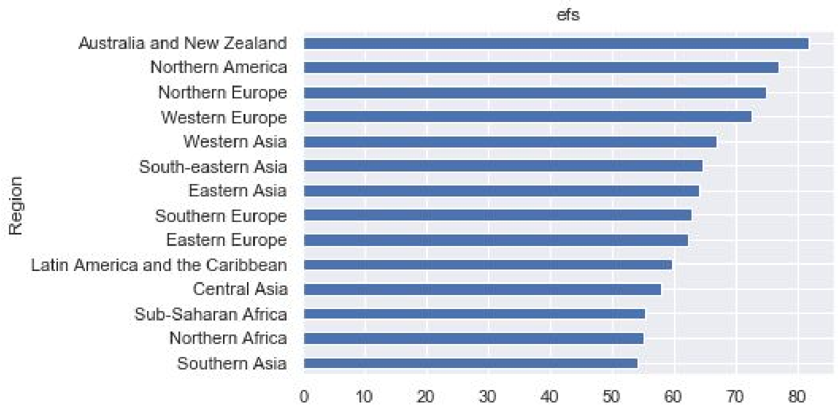


Figure 3: Economic Development Average by Region

Gender Inequality | World Happiness

Visualizing the data showed that countries with higher world happiness scores tended to have lower gender inequality scores, pointing towards a negative correlation between the two.

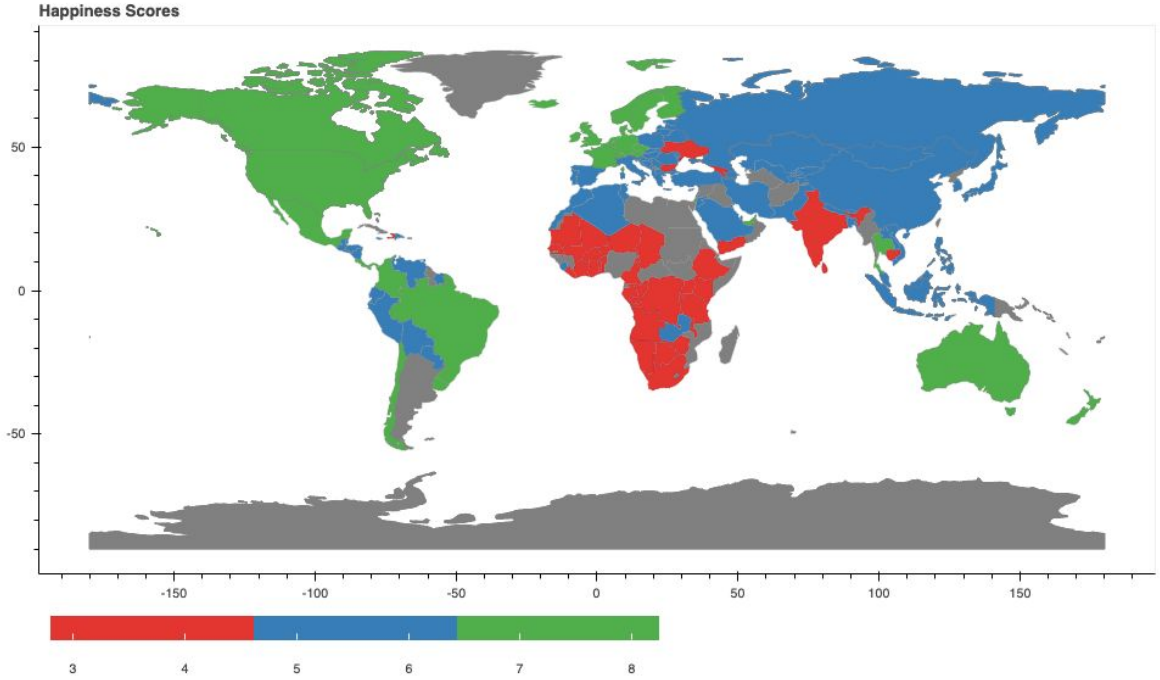


Figure 4: Map of World Happiness Scores

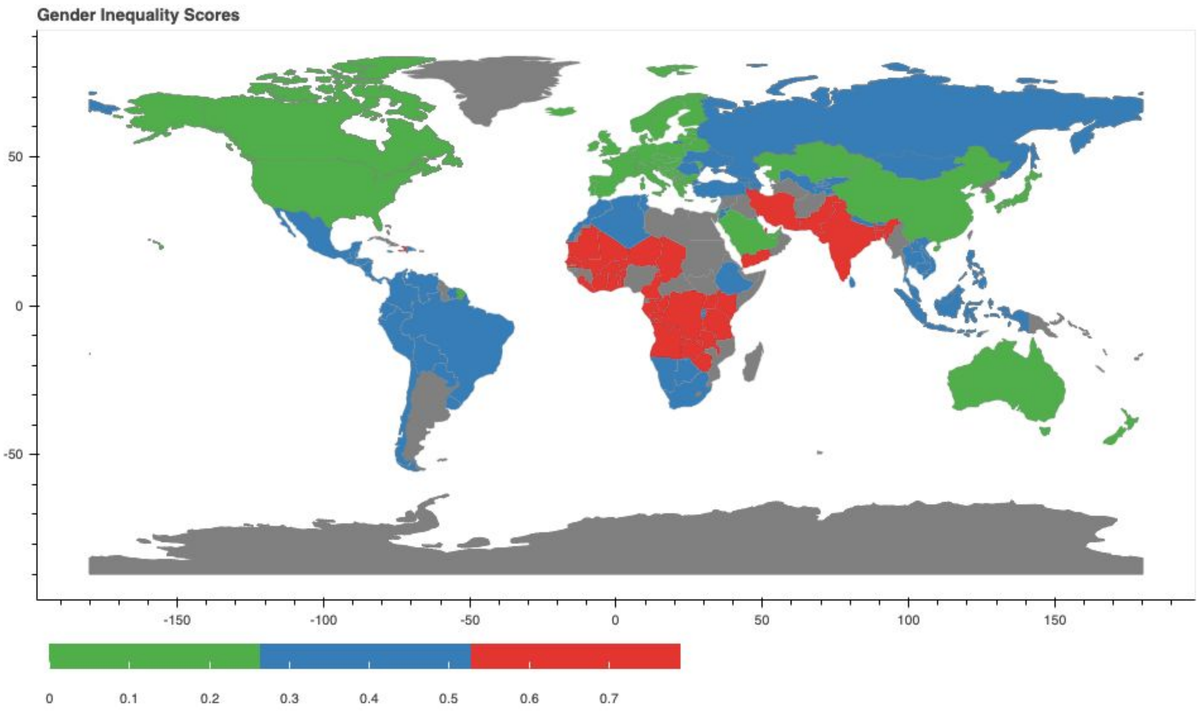


Figure 5: Map of Gender Inequality Scores

<u>Population Density | World Happiness</u>

Population has a large range of values with a few extreme values that skew the data. Although the regions with the highest happiness scores had lower population densities, visualizations did not uncover much correlation between the two datasets.

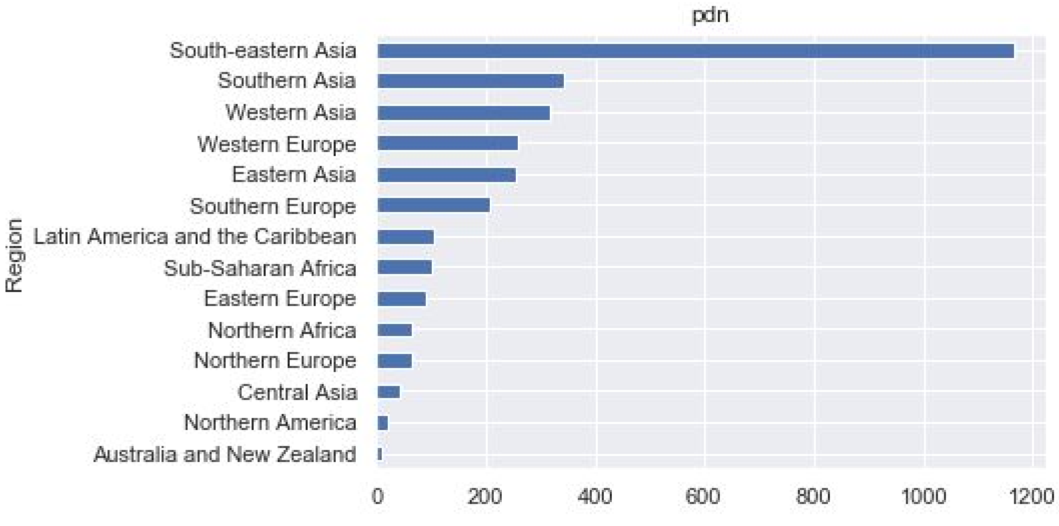


Figure 6: Population Density Average by Region

Distribution of Features

World Happiness Scores and Economic Freedom Scores have fairly normal distributions. Gender Development Scores have a left-skewed distribution with more countries' scores are higher. Gender Inequality Scores are bimodal with one peak around 0.1 and another around 0.5. Population density has some major outliers and a right-skewed distribution.

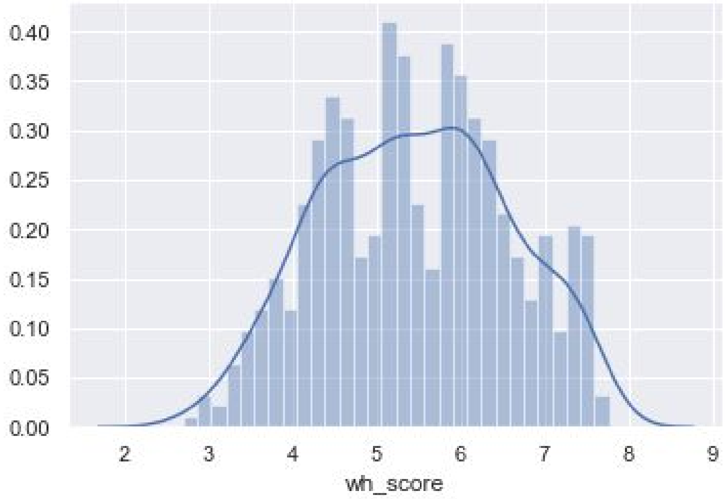


Figure 7: Distribution of World Happiness Scores

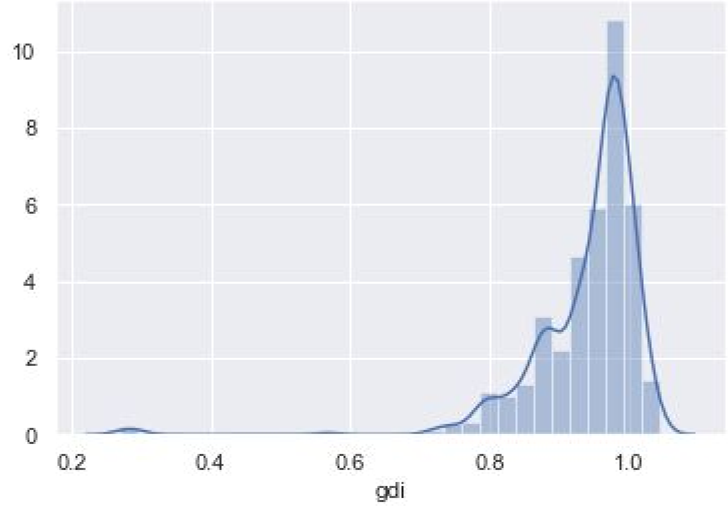


Figure 8: Distribution of Gender Development Scores

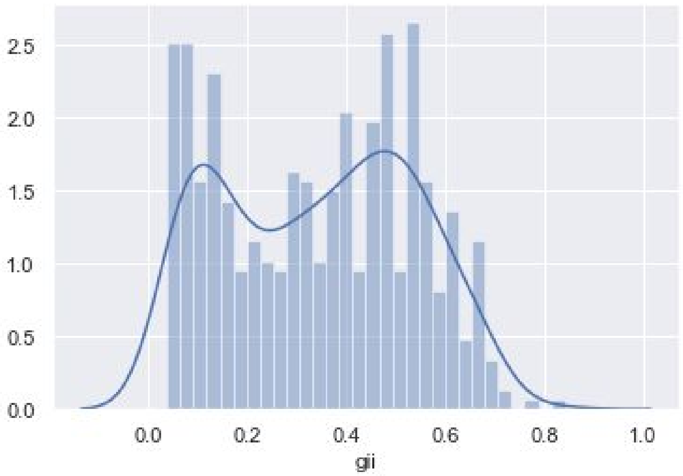


Figure 9: Distribution of Gender Inequality Scores

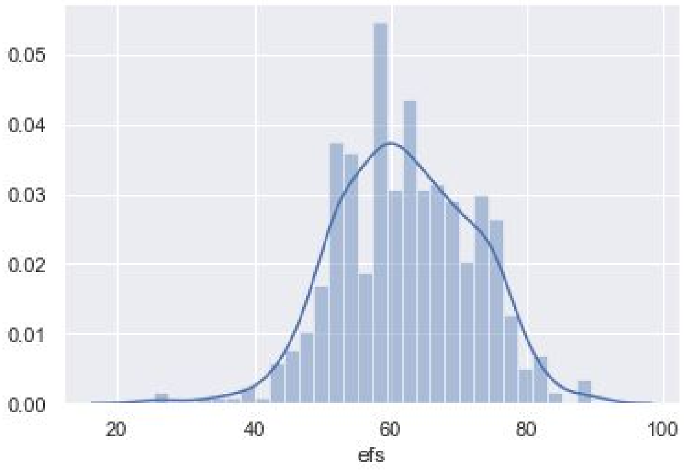


Figure 10: Distribution of Economic Freedom Scores

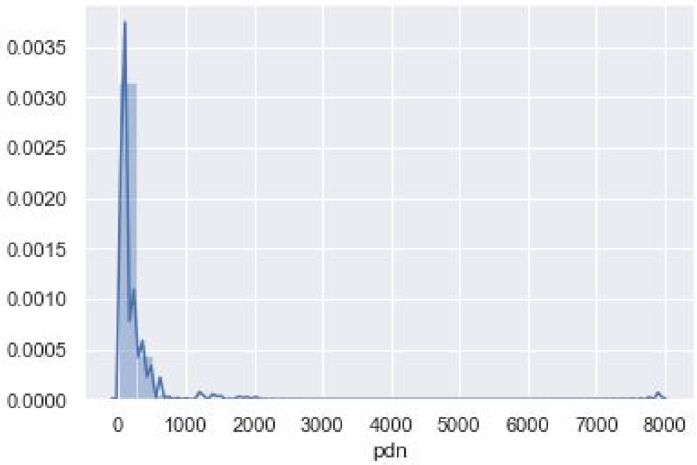


Figure 11: Distribution of Population Density

Statistical Analysis

Note: Please see the accompanying jupyter notebook for a detailed overview of the code, calculations and plots referenced below (https://bit.ly/2JcjdeP)

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.

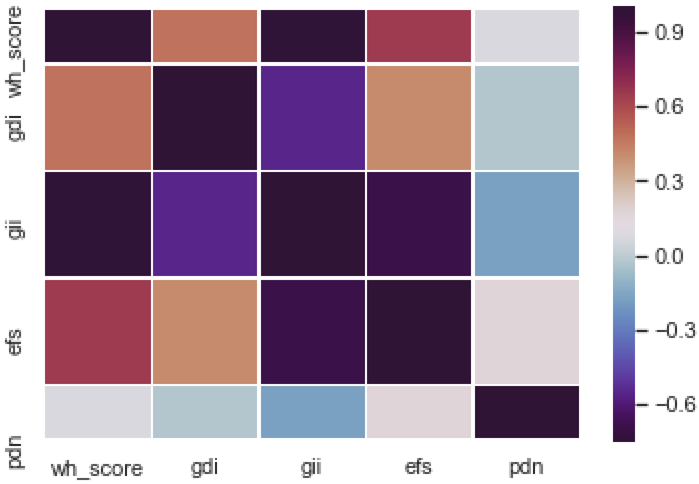


Figure 12: Correlation heat map

	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 13: 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.

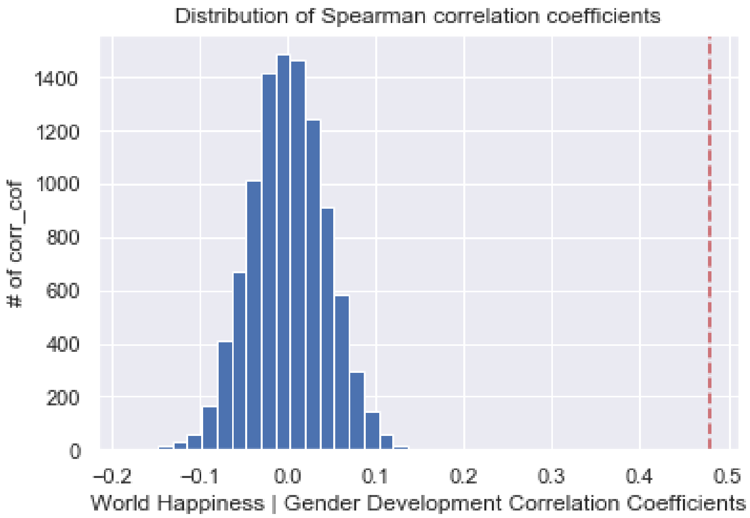


Figure 14: Distribution of Gender Development correlation coefficients

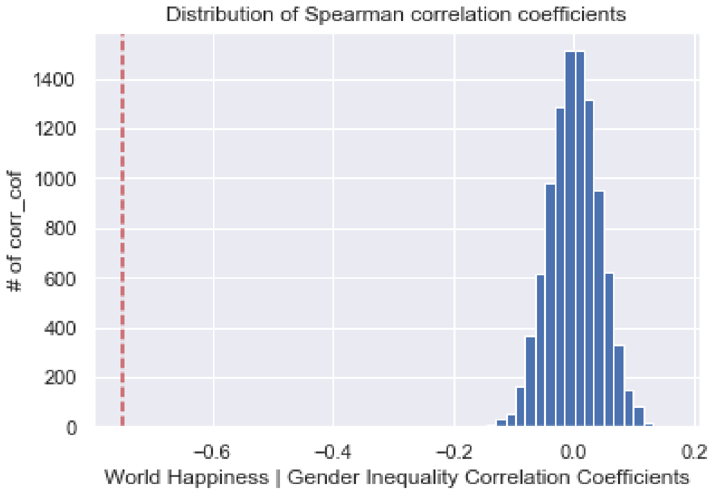


Figure 15: Distribution of Gender Inequality correlation coefficients

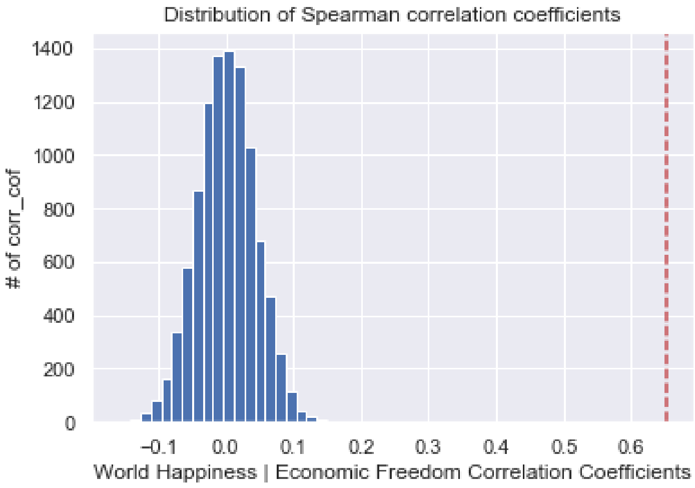


Figure 16: Distribution of Economic Freedom correlation coefficients

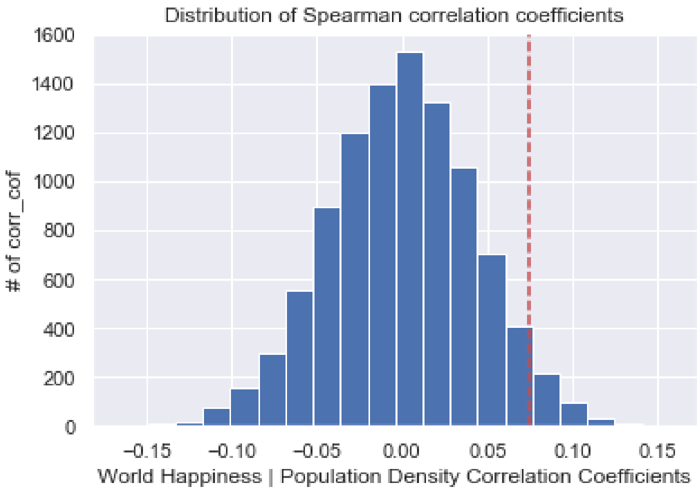


Figure 17: Distribution of Population Density correlation coefficients

Final Analysis

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.

In-Depth Analysis

Note: Please see the accompanying jupyter notebook for a detailed overview of the code referenced below (https://bit.ly/2UhvViO)

Preparing the Dataset

During EDA, it was noted countries of the same geographical regions had similar happiness scores. It would be advantageous to include regions in the feature set of the regression models. The regions included in the dataset are:

- Southern Europe
- Northern Africa
- Sub-Saharan Africa
- Western Asia
- Australia and New Zealand
- Western Europe
- Southern Asia
- Eastern Europe
- Latin America and the Caribbean
- South-eastern Asia
- Northern America
- Eastern Asia
- Northern Europe
- Central Asia

Linear regression models do not allow categorical data. In order to include regions, one hot encoding was utilized. Pandas .get_dummies() method was applied to create a column for each region and a value of 1 if the country was in the region and 0 if it was not.

	Country	Region	iso_a3	wh_score	gdi	gii	efs	pdn	year	region_1	region_2	region_3
0	Albania	Southern Europe	ALB	4.655	0.97	0.256	65.650	105.135	2015	1	0	0
1	Algeria	Northern Africa	DZA	6.355	0.858	0.442	48.881	16.680	2015	0	1	0
2	Angola	Sub-Saharan Africa	AGO	3.865	0.838	0.575	47.88	22.366	2015	0	0	1

Figure 18: Truncated example of dataset after One Hot Encoding

Feature Selection

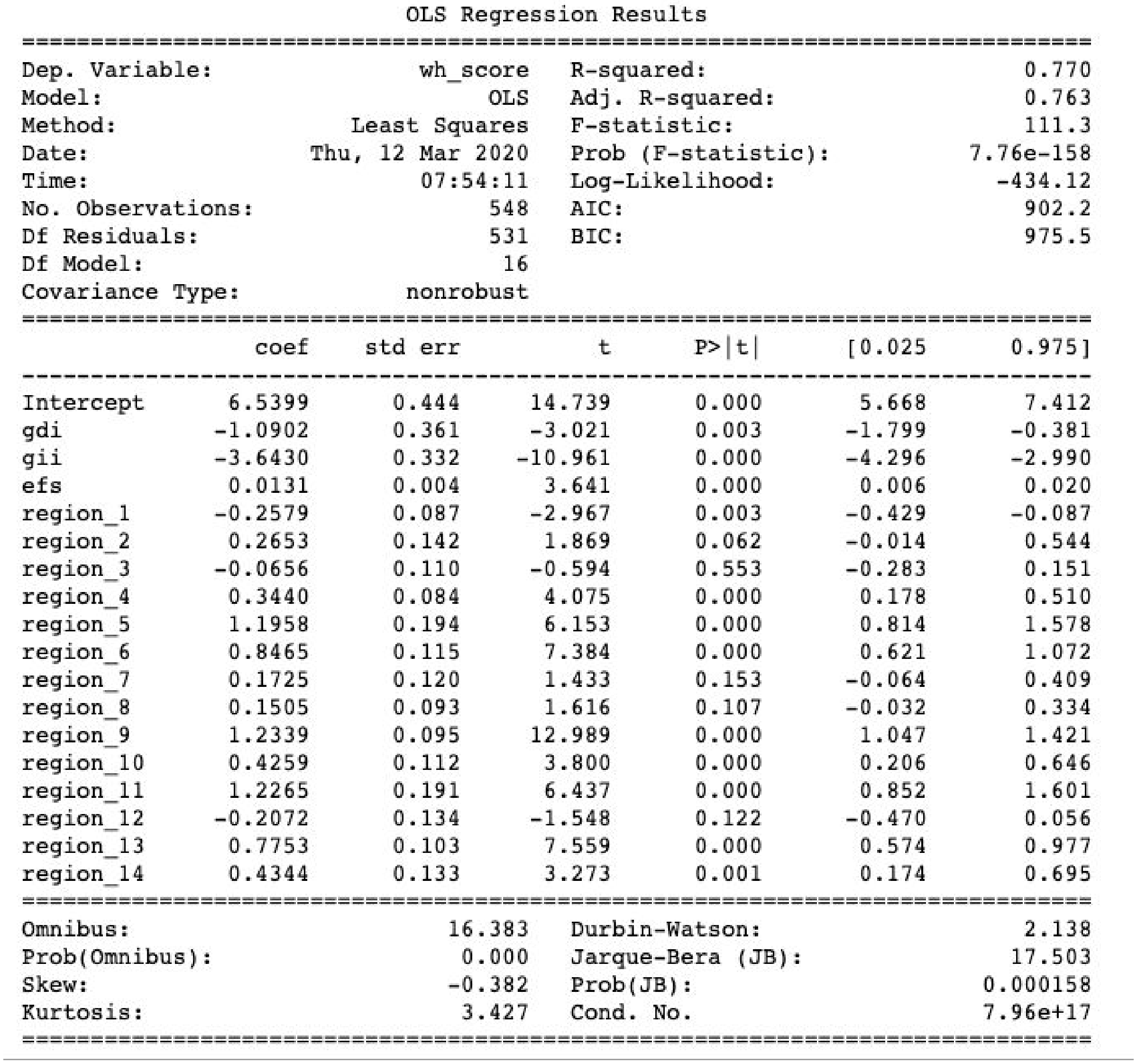
Python module statsmodels was used to determine which features to include in the models. The summary method provides result statistics for each target/feature combination:

- R2: a measure of the proportion of the variance for the target variable that can be explained by the feature(s). A higher R2 is better
- P-value: the probability of obtaining test results at least as extreme as the ones observed. A p-value of less than 0.05 provides confidence that a feature has statistical significance
- AIC: provides an estimate on the quality of the model relative to the other models. A lower AIC indicates a better fitting model.

Based on the statistics above, it appears it would be beneficial to remove population density from the feature set. I chose not to include population density in the model because when grouped with the other features, it's p-value was higher than 0.05. Additionally, the AIC is higher, albeit not by much, than the model that omits population density. The features that will be included in the models are: gender development, gender inequality, economic freedom, and regions. Since the regions are a categorical feature, their p-value is not evaluated for significance. Another interesting observation is that the coef for gdi is negative when grouped with the other features.

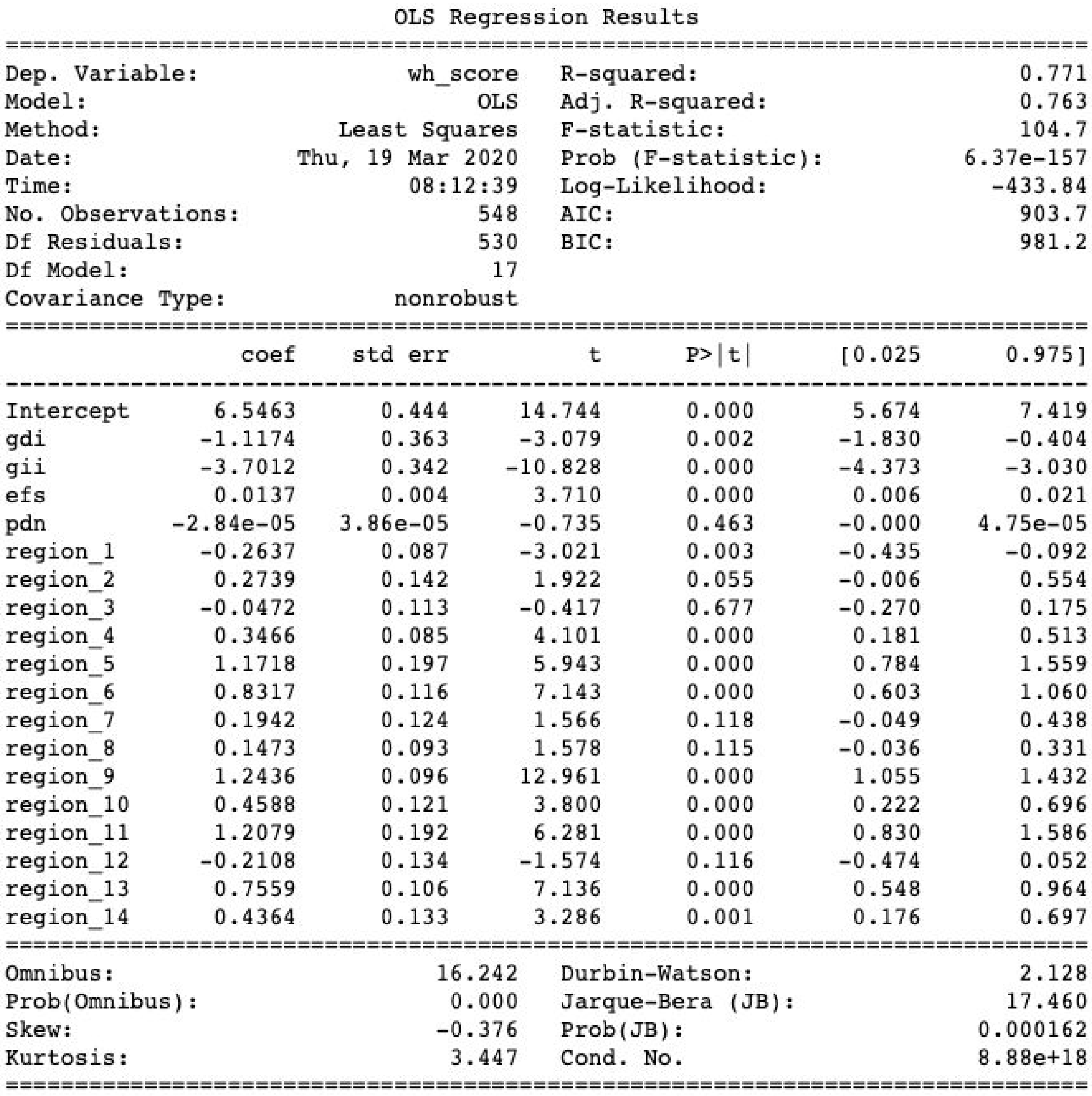
	wh_score ~ gdi	**wh_score ~ gii**	**wh_score ~ efs**	**wh_score ~ pdn**	**wh_score ~ region_x**	**wh_score ~ gdi+gii+efs+ region_x**
R2	0.188	0.557	0.402	0.009	0.68	0.77
Adjusted R2	0.186	0.556	0.401	0.007	0.672	0.763
F-statistic	126.1	686.2	367.3	4.912	87.26	111.3
AIC	1565	1232	1397	1674	1078	902.2
coef	5.6736	-4.3129	0.0716	0.0001		
std err	0.505	0.165	0.004	6.75E-05		
P>|t|	0	0	0	0.027		

Figure 19: Key result statistics for target/feature combinations



```
                            OLS Regression Results
==============================================================================
Dep. Variable:               wh_score   R-squared:                       0.770
Model:                            OLS   Adj. R-squared:                  0.763
Method:                 Least Squares   F-statistic:                     111.3
Date:                Thu, 12 Mar 2020   Prob (F-statistic):         7.76e-158
Time:                        07:54:11   Log-Likelihood:                -434.12
No. Observations:                 548   AIC:                             902.2
Df Residuals:                     531   BIC:                             975.5
Df Model:                          16
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      6.5399      0.444     14.739      0.000       5.668       7.412
gdi           -1.0902      0.361     -3.021      0.003      -1.799      -0.381
gii           -3.6430      0.332    -10.961      0.000      -4.296      -2.990
efs            0.0131      0.004      3.641      0.000       0.006       0.020
region_1      -0.2579      0.087     -2.967      0.003      -0.429      -0.087
region_2       0.2653      0.142      1.869      0.062      -0.014       0.544
region_3      -0.0656      0.110     -0.594      0.553      -0.283       0.151
region_4       0.3440      0.084      4.075      0.000       0.178       0.510
region_5       1.1958      0.194      6.153      0.000       0.814       1.578
region_6       0.8465      0.115      7.384      0.000       0.621       1.072
region_7       0.1725      0.120      1.433      0.153      -0.064       0.409
region_8       0.1505      0.093      1.616      0.107      -0.032       0.334
region_9       1.2339      0.095     12.989      0.000       1.047       1.421
region_10      0.4259      0.112      3.800      0.000       0.206       0.646
region_11      1.2265      0.191      6.437      0.000       0.852       1.601
region_12     -0.2072      0.134     -1.548      0.122      -0.470       0.056
region_13      0.7753      0.103      7.559      0.000       0.574       0.977
region_14      0.4344      0.133      3.273      0.001       0.174       0.695
==============================================================================
Omnibus:                       16.383   Durbin-Watson:                   2.138
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               17.503
Skew:                          -0.382   Prob(JB):                     0.000158
Kurtosis:                       3.427   Cond. No.                     7.96e+17
==============================================================================
```

Figure 20: Statsmodel summary results for selected model features (excluding Population Density)



```
                            OLS Regression Results
==============================================================================
Dep. Variable:               wh_score   R-squared:                       0.771
Model:                            OLS   Adj. R-squared:                  0.763
Method:                 Least Squares   F-statistic:                     104.7
Date:                Thu, 19 Mar 2020   Prob (F-statistic):          6.37e-157
Time:                        08:12:39   Log-Likelihood:                -433.84
No. Observations:                 548   AIC:                             903.7
Df Residuals:                     530   BIC:                             981.2
Df Model:                          17
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      6.5463      0.444     14.744      0.000       5.674       7.419
gdi           -1.1174      0.363     -3.079      0.002      -1.830      -0.404
gii           -3.7012      0.342    -10.828      0.000      -4.373      -3.030
efs            0.0137      0.004      3.710      0.000       0.006       0.021
pdn        -2.84e-05   3.86e-05     -0.735      0.463      -0.000    4.75e-05
region_1      -0.2637      0.087     -3.021      0.003      -0.435      -0.092
region_2       0.2739      0.142      1.922      0.055      -0.006       0.554
region_3      -0.0472      0.113     -0.417      0.677      -0.270       0.175
region_4       0.3466      0.085      4.101      0.000       0.181       0.513
region_5       1.1718      0.197      5.943      0.000       0.784       1.559
region_6       0.8317      0.116      7.143      0.000       0.603       1.060
region_7       0.1942      0.124      1.566      0.118      -0.049       0.438
region_8       0.1473      0.093      1.578      0.115      -0.036       0.331
region_9       1.2436      0.096     12.961      0.000       1.055       1.432
region_10      0.4588      0.121      3.800      0.000       0.222       0.696
region_11      1.2079      0.192      6.281      0.000       0.830       1.586
region_12     -0.2108      0.134     -1.574      0.116      -0.474       0.052
region_13      0.7559      0.106      7.136      0.000       0.548       0.964
region_14      0.4364      0.133      3.286      0.001       0.176       0.697
==============================================================================
Omnibus:                       16.242   Durbin-Watson:                   2.128
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               17.460
Skew:                          -0.376   Prob(JB):                     0.000162
Kurtosis:                       3.447   Cond. No.                     8.88e+18
==============================================================================
```

Figure 21: Statsmodel summary results for selected model features (including Population Density)

Fit Diagnostics

Statsmodel was used again to obtain the residuals and plot a residual and probability plot. The residual plot is fairly symmetrically distributed around 0. Both the probability plot and the residual histogram show that the data is normally distributed.   Therefore, a linear regression model appears to be a good fit for this data.

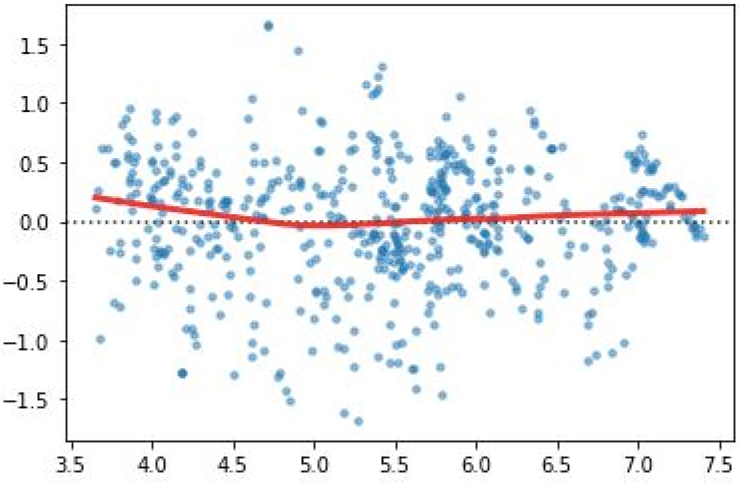


Figure 22: Residual plot: fitted values vs. residuals

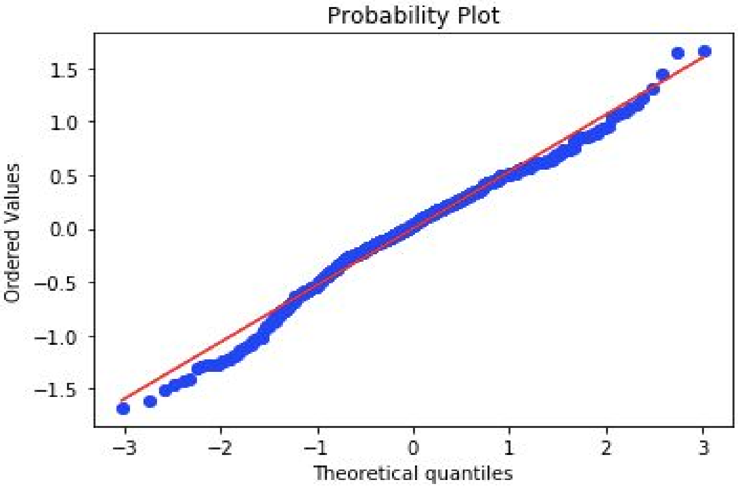


Figure 23: Probability plot

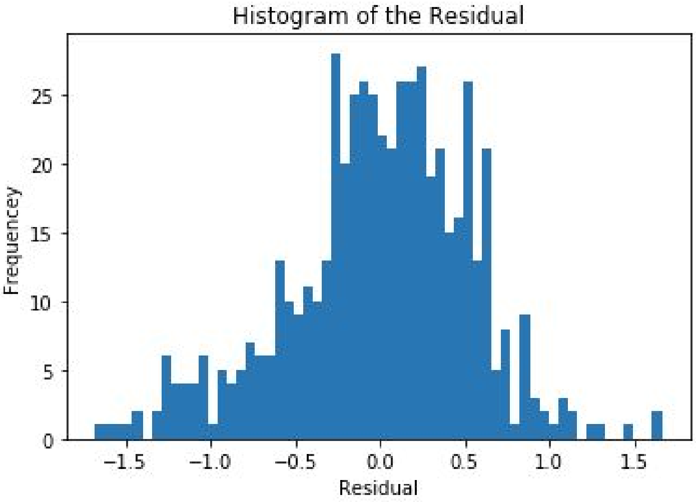
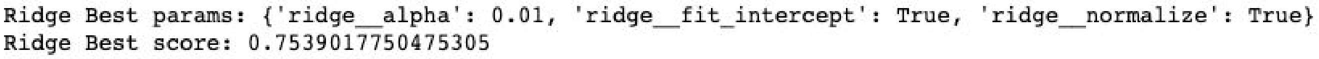


Figure 24: Histogram of the Residual

Modeling

I used the train_test_split() function to split my data into a training set and a testing set. 5 models were applied to the training data: Linear Regression, Ridge Regression, Lasso Regression, K-Nearest Neighbor Regressor, and Random Forest Regressor. Hyper-parameters' performance was selected and cross-validated using GridSearchCV. R2, mean squared error, and root mean squared error where the metrics used to evaluate the regression models against the training dataset and the testing dataset as well as check for overfitting.



```
Ridge Best params: {'ridge__alpha': 0.01, 'ridge__fit_intercept': True, 'ridge__normalize': True}
Ridge Best score: 0.7539017750475305
```

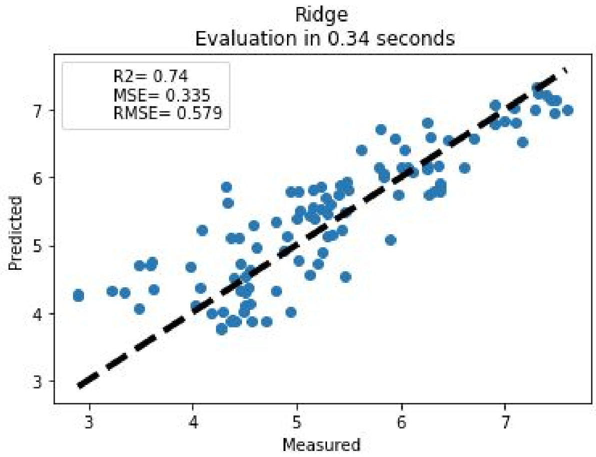
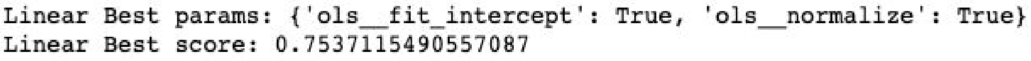


Figure 25: Ridge Regression Predicted vs. Actual Plot



```
Linear Best params: {'ols__fit_intercept': True, 'ols__normalize': True}
Linear Best score: 0.7537115490557087
```

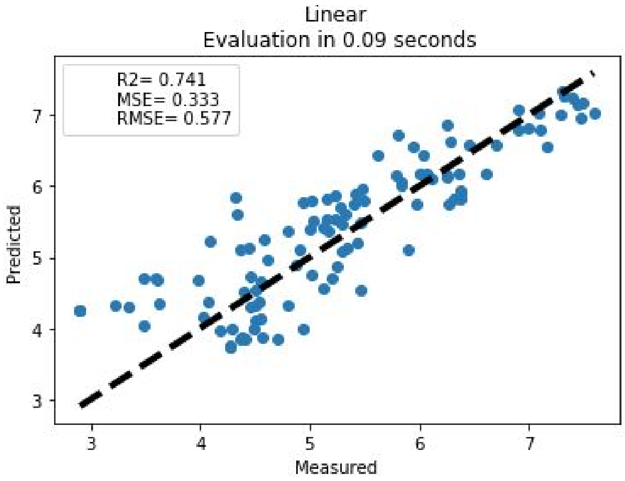
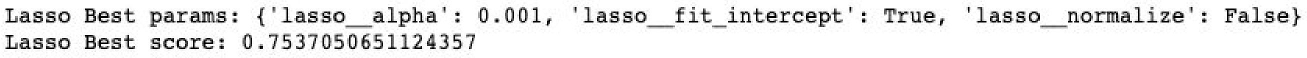


Figure 26: Linear Regression Predicted vs. Actual Plot



```
Lasso Best params: {'lasso__alpha': 0.001, 'lasso__fit_intercept': True, 'lasso__normalize': False}
Lasso Best score: 0.7537050651124357
```

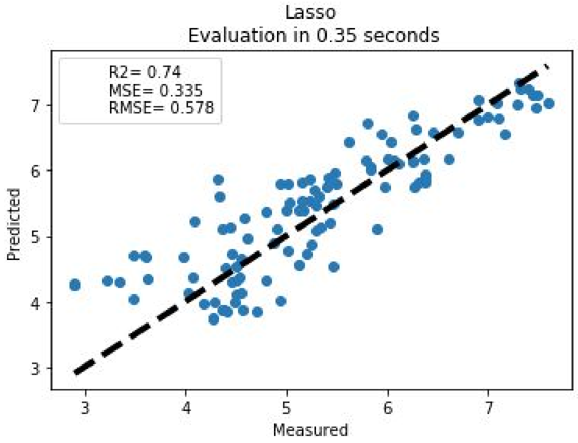
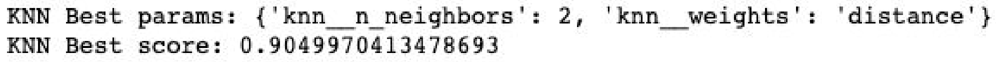


Figure 27: Lasso Regression Predicted vs. Actual Plot



```
KNN Best params: {'knn__n_neighbors': 2, 'knn__weights': 'distance'}
KNN Best score: 0.9049970413478693
```

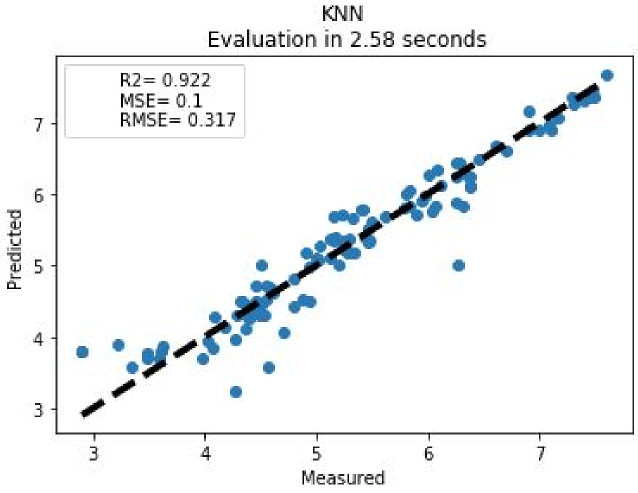


Figure 28: K-Nearest Neighbors Regressor Predicted vs. Actual Plot

```
RandomForest Best params: {'rf__criterion': 'mse', 'rf__max_depth': 90, 'rf__n_estimators': 70}
RandomForest Best score: 0.8757049990150947
```

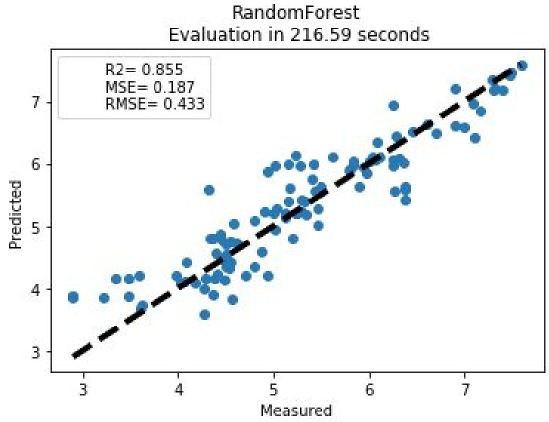


Figure 29: Random Forest Regressor Predicted vs. Actual Plot

Model Analysis

In all of the models, the delta between the R2 of the training data and the R2 of the testing data is less than 10%, therefore overfitting is not an issue. Random Forest performed well, but it’s execution time is significantly higher than the other models and not ideal from a business perspective.

Conclusion

K-Nearest Neighbors Regressor model was selected for this dataset.
The model was selected based on the following criteria:

1. No issue with overfitting
2. Highest performance metric (R2)
3. Relatively low level of complexity