Happiness:
And its Effects Over Time

Alex Arnold
Sophie Knight
Samantha Lane
Rachel Podemski

Introduction:

Everyone's happiness changes over time. Everyone has different variables that impact their happiness. So, how did the World Happiness Report turn happiness into a quantitative value and how did this variable of happiness change over time? We wanted to investigate this variable of happiness and determine how it has changed over time and what variables impacted this measured value.

Inspiration:

Buddha said it best, "There is no path to happiness. Happiness is the path." Throughout this project we tried to understand happiness and its features based on how the World Happiness Report broke down its reporting on the Happiness Score. It was intriguing to investigate whether or not happiness could be viewed as a standard metric or if other factors can play a part in affecting it, which led us to further examine weather characteristics.

Most relevantly, the coronavirus and the global pandemic has affected our happiness personally, so what can we anticipate for next year's score? People and countries have certainly gone through turmoil before, but what will its future effect on happiness be?

Hypotheses:

When investigating happiness, we came up with a couple hypothesis: Would happiness increase over time, and does weather positively affect happiness? Since the years in our dataset predict the previous years Happiness Score (e.g. 2020 Happiness Score is based off of 2019 data), we predict that the dataset for next year, 2021, will show a decrease in happiness.

Sources of Data:

We collected our data from a variety of different data sources including Kaggle, the Open Weather API, and Google Maps API. Our initial data set was sourced from Kaggle and it included 6 separate CSV's that contained information from the World Happiness Report. They collected data on seven different variables including: Economy, Social Support, Health, Freedom, Trust, Generosity, and Residuals. Each of these variables were added up and the sum was reported as the overall Happiness Score for each country.

We also used the Google Maps API to pull the latitude and longitude for each country that was included in the 2020 World Happiness Report. This allowed us to then pull the Temperature, Feels Like, Pressure, Humidity, Wind Speed, and Cloudiness data from the Open Weather API in order to cross examine the World Happiness Report with other variables.

Data Cleaning and Exploration:

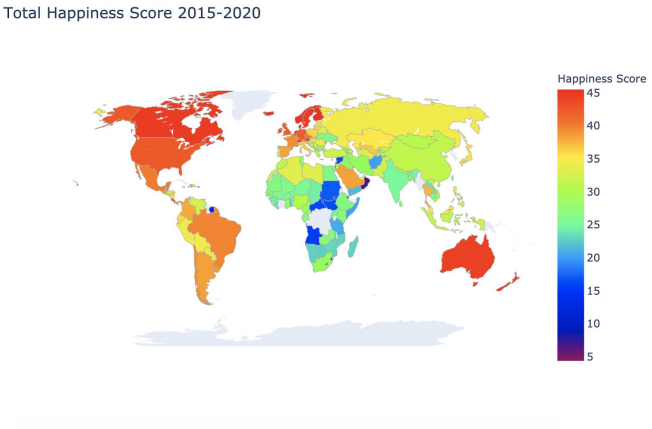
Before we could begin our analysis we had to clean and organize our data. We started out by importing each individual year's CSV to a dataframe to compare the different column names. After doing this we had to format and rename the different columns for each year's

dataframe for consistency. We also had to make sure all of the different variables were in a consistent format across each year.

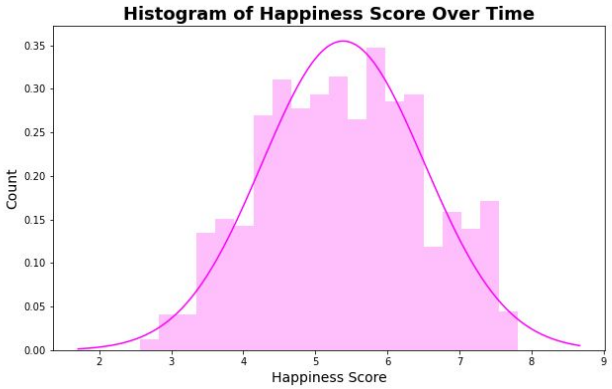
Once each year's dataframe was clean and consistent, we were able to merge all six dataframes into one large dataframe. We appended each of the CSV's on top of each other, creating a vertical dataframe, and added a separate column to include the year each line was from. This merged dataframe allowed us to easily compare the data across each year.

Analysis:

We first wanted to know who is happy? Figure 1 displays the world chloropeth map illustrating the Happiness Score from 2015-2020. While a limitation is that not every country reported its Happiness Score each year, this representation gave us a nice visual of the happiness spectrum across countries.



After analyzing who was happy, we analyzed the Happiness Score over time. The histogram shows a normal distribution of these scores over the 2015-2020 timeframe highlighted by the overlay. The violin plot of the Happiness Score goes on to emphasize this distribution for each year in addition to showing the slight increase in scores.

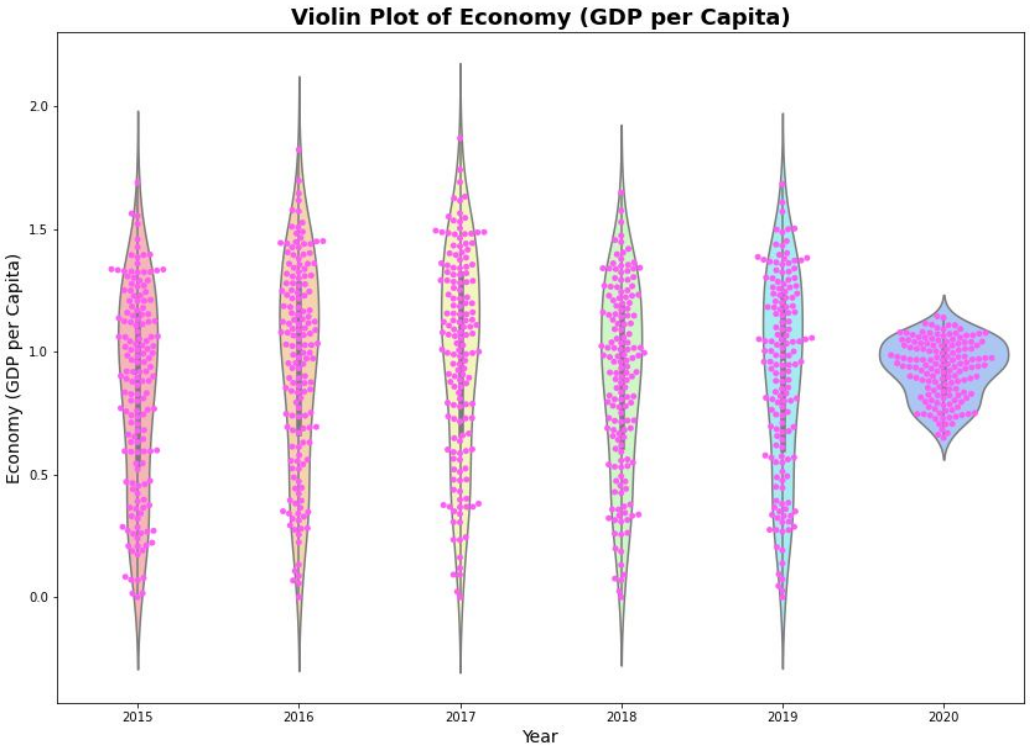


ANOVA: Economy (GDP per Capita):

After investigating the ANOVA for overall Happiness Score, we ran it with all of the different variables that made up the Happiness Score. In table 1 we can see that all of the variables were statistically significant. It is important to note that for the variable Economy, the statistically significant result was fragile. By looking at the series of means in figure 4, the year 2015 is a very slight outlier compared to the rest of the years. Despite how inconsequential this

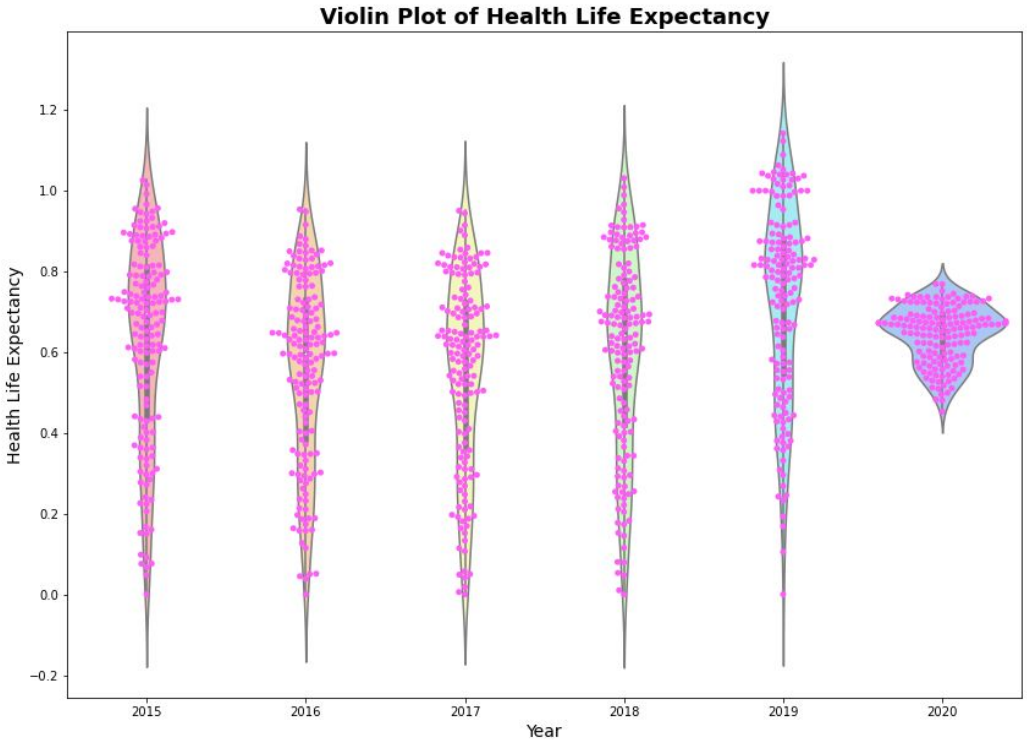
may appear, we found that after removing this year from the test, there was no longer overall statistical significance for Economy.

Variable	P-Value
Happiness Score	0.9522
Economy (GDP per Capita)	0.0163
Social Support	4.16E-07
Health Life Expectancy	1.95E-12
Freedom	3.40E-13
Trust (Government Corruption)	2.37E-30
Generosity	1.39E-08



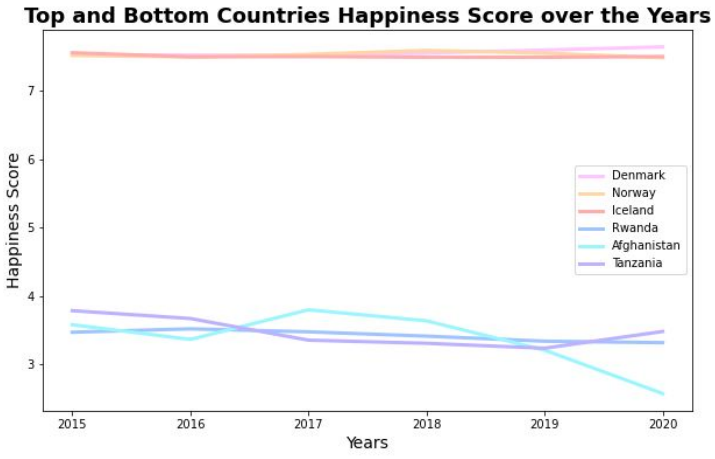
ANOVA: Health Life Expectancy:

Through ANOVA testing we found that another significant factor of overall happiness was the Health Life Expectancy variable. We also tried manipulating the test by removing the year 2017 in figure 5, as it was the outlier for this variable, to see if it would cause the same result as the Economy test. Once it was removed, the data was still statistically significant.

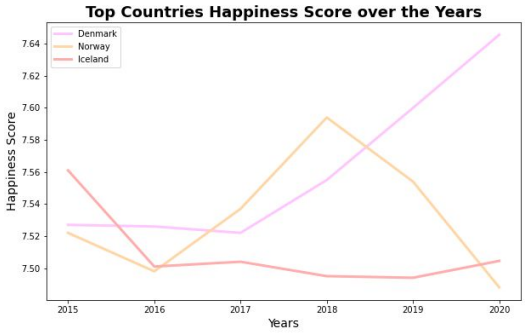


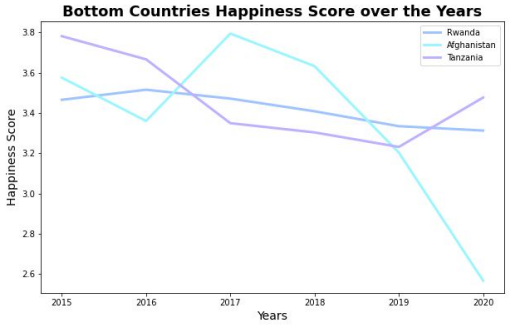
Top & Bottom Countries:

After investigating the difference between different variables across the years we decided to dive a little deeper and look into the major difference between countries that were repeatedly in the top and bottom 10% based on their Happiness Score. We found three total countries that were in the top 10% every year and three countries that were in the bottom 10% every year. The top three countries were Denmark, Norway and Iceland. The bottom three countries were Rwanda, Afghanistan, and Tanzania. We then graphed each variable to see how each of these six

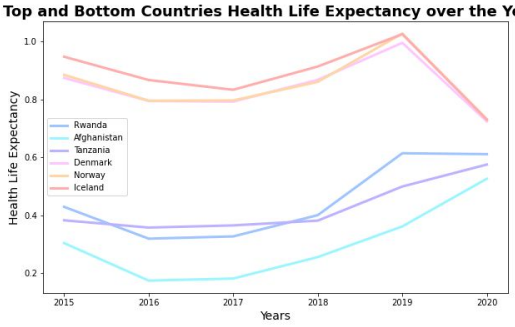


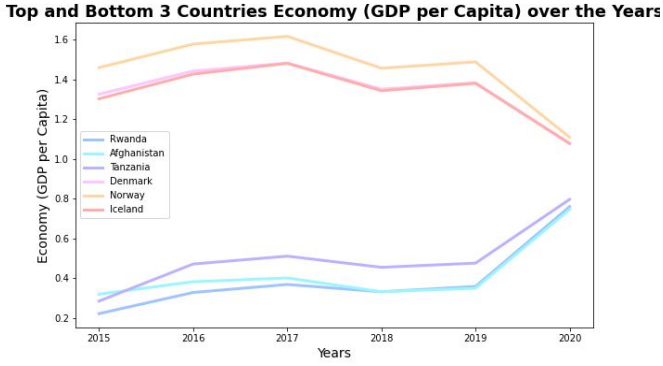
countries changed over the years and to determine if there was any pattern we could visualize. As you can see in figure 6 there is a visual difference between the Happiness Score for the top three countries and bottom three countries. It was difficult to visualize if there was any consistent variation across all 6 of the countries when they were graphed together, so we also graphed them separately in figure 7 and figure 8. In figure 7 it is easy to see that each top country certainly had variation over the years, but they all did not increase or decrease at the same time. Similarly, in figure 8, the bottom countries also varied across the years.





We also were able to compare each of the individual variables that were summed into the Happiness Score and how they varied based on the top and bottom countries. As you can see in figure 9 it is very apparent that all three of the top countries' Health and Life Expectancy dropped from 2019 to 2020 whereas the bottom three countries have risen from 2019 to 2020. It is also worth noting that these values are quickly approaching each other, yet the bottom three countries remained in the bottom 10% and the top three countries remained in the top 10% based on Happiness Rank. Similarly figure 10 shows how these six countries fared in terms of economy over the six years included in the study. Again, as the years go on, the top three countries are slowly decreasing whereas the bottom three continue to rise. This analysis of how the top and bottom countries changed over time lead us to dig deeper into our data to see what variable correlated most with Happiness Score.





T-tests:

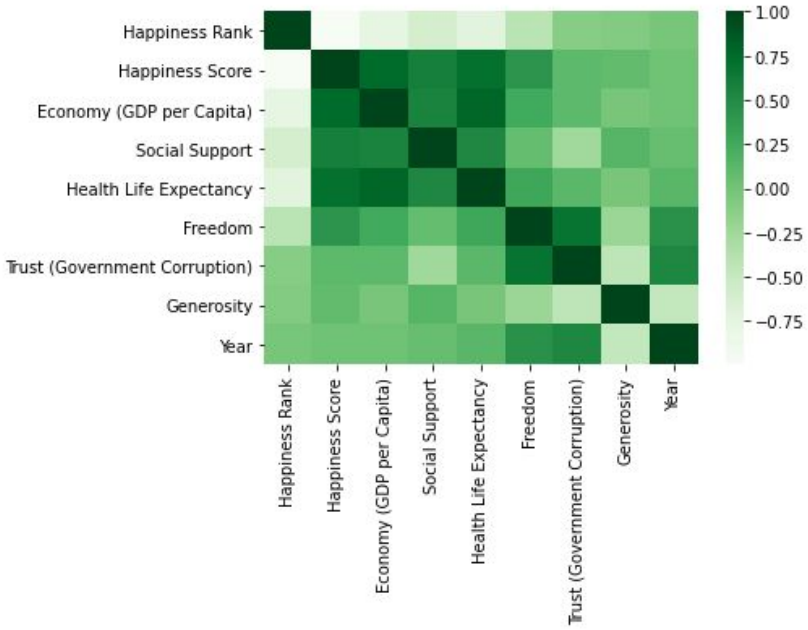
We decided to then perform t-tests on the overall Happiness Score, and the two significant variables, Health and Economy. We did this by comparing the top 10 countries Happiness Score of 2015 to 2020, and then ran the same test for the bottom 10 countries. When looking at Happiness Score, both tests produced p-values greater than .05. Therefore the changes over time are not statistically significant, and we rejected the alternative hypothesis, accepting the null.

When running the test for Economy and Health, these tests all produced p-values less than .05. Therefore the changes were statistically significant, rejecting the null hypothesis and accepting the alternative. The result of each test can be seen below, in table 2.

P-Values	Happiness Score	Economy (GDP per Capita)	Health Life Expectancy
Top 10 Countries	0.6576	1.73E-8	5.94E-10
Bottom 10 Countries	0.6796	5.43E-7	0.0005

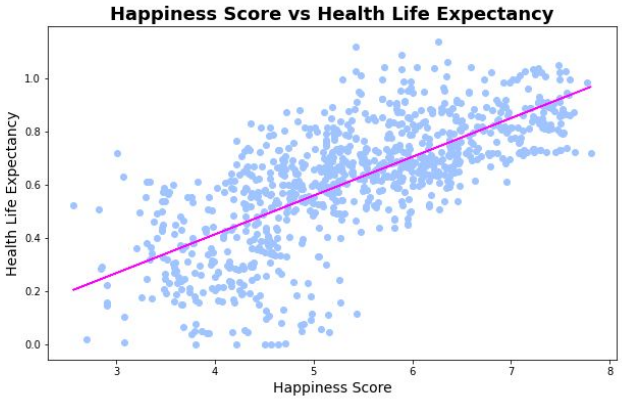
Correlations:

Next we wanted to look at correlation of Happiness Score to each individual variable. As we can see here in figure 11, Economy and Health Life Expectancy have the strongest correlations.



From there we decided to look into these variables further to see how strong the correlations are.We started off by creating scatter plots for each variable to Happiness Score, and then ran linregress. For both of these figures we can now visually see the positive

correlations they have. Next, to see just how strong the correlations were, we ran stats models. For figure 12 we got an r-value of .575, and for figure 13, we got an r-value of .510. Since both r-values were above .5, they have a strong positive correlation. We also ran the stats models on the other variables. All of them had a positive r-value that was below .5 and therefore have a weak correlation with Happiness Score as can be seen by the results in the Appendix.



Dep. Variable:	Economy (GDP per Capita)	**R-squared:**	0.575
Model:	OLS	**Adj. R-squared:**	0.574
Method:	Least Squares	**F-statistic:**	1259.
Date:	Sat, 14 Nov 2020	**Prob (F-statistic):**	3.37e-175
Time:	14:37:25	**Log-Likelihood:**	-6.6648
No. Observations:	934	**AIC:**	17.33
Df Residuals:	932	**BIC:**	27.01
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.4421	0.039	-11.300	0.000	-0.519	-0.365
Happiness Score	0.2520	0.007	35.485	0.000	0.238	0.266

Omnibus:	6.608	**Durbin-Watson:**	1.712
Prob(Omnibus):	0.037	**Jarque-Bera (JB):**	6.532
Skew:	-0.202	**Prob(JB):**	0.0382
Kurtosis:	3.064	**Cond. No.**	27.9

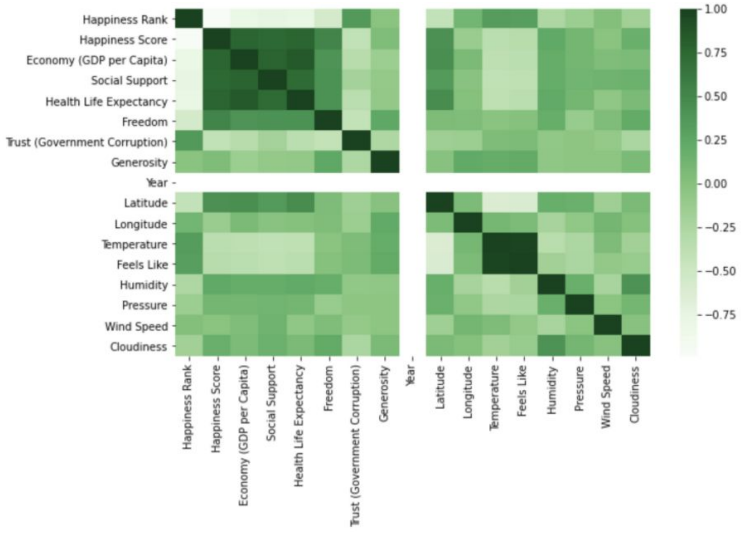
Dep. Variable:	Health Life Expectancy	**R-squared:**	0.510
Model:	OLS	**Adj. R-squared:**	0.510
Method:	Least Squares	**F-statistic:**	971.1
Date:	Sat, 14 Nov 2020	**Prob (F-statistic):**	1.21e-146
Time:	14:37:25	**Log-Likelihood:**	384.28
No. Observations:	934	**AIC:**	-764.6
Df Residuals:	932	**BIC:**	-754.9
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.1677	0.026	-6.516	0.000	-0.218	-0.117
Happiness Score	0.1456	0.005	31.162	0.000	0.136	0.155

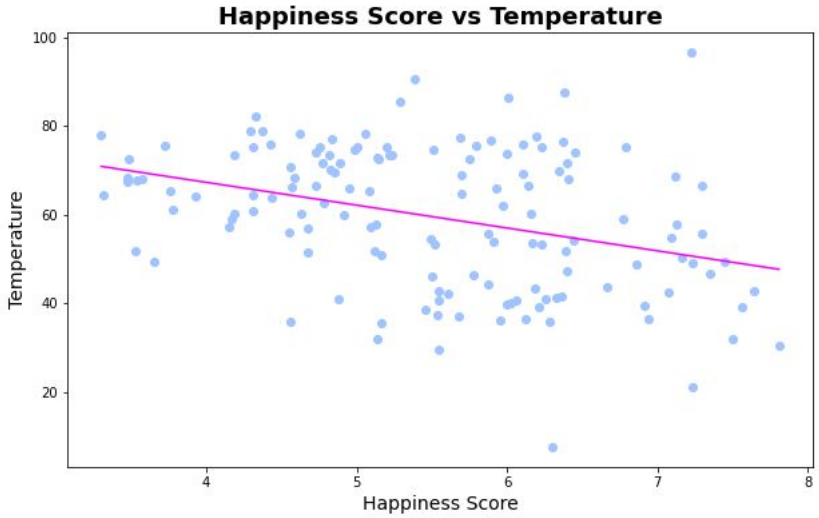
Omnibus:	24.331	**Durbin-Watson:**	1.496
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	30.649
Skew:	-0.299	**Prob(JB):**	2.21e-07
Kurtosis:	3.656	**Cond. No.**	27.9

Correlations with Weather:

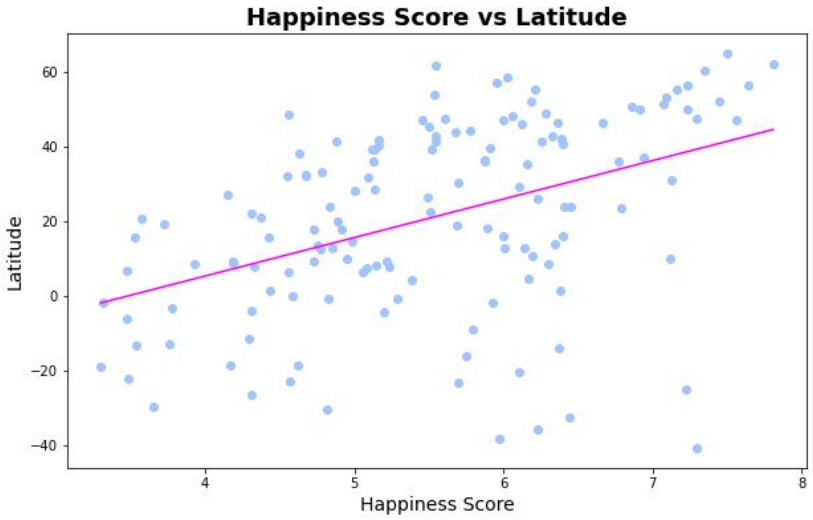
As mentioned above we wanted to cross examine our data with an outside data source. We combined the weather data for each country in the 2020 World Happiness Report Data set and compared it with each variable that was included in the 2020 World Happiness Report. As you can see in figure 14 there are not very many strong correlations between any of the weather variables and the World Happiness Report variables. When looking at the heat map you can see a slight positive correlation between Latitude and Happiness Score and a slight negative correlation between Temperature and Happiness Score.



In order to investigate these correlations deeper, we were able to create a scatter plot and linregress models for each weather variable with their Happiness Score. After reviewing each R-value it became apparent none of the weather variables had a strong correlation with Happiness Score. We decided to include the two variables with the strongest correlations. As you can see in figure 15 there is a weak relationship between Happiness Score and Temperature. The calculated r-values for the simple linregress is 0.44 which shows that there is a weak positive correlation between Happiness Score and Latitude.



Similarly, the relationship between Happiness Score vs Latitude as you can see in figure 16 shows a slight negative correlation. The calculated r-values for the simple linregress is -0.35 which shows that there is a weak negative correlation between Happiness Score and Temperature.



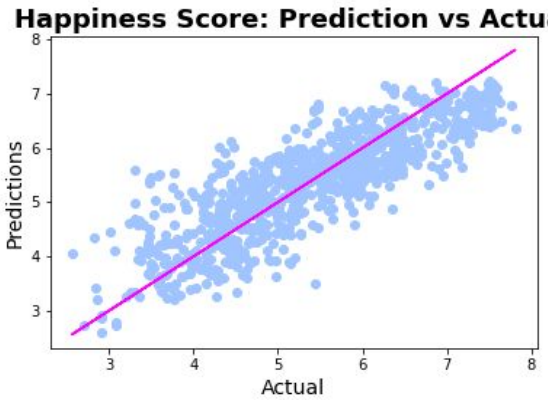
Prediction Model:

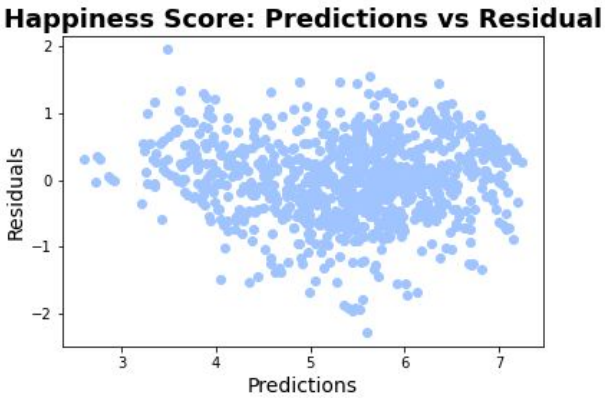
For our Prediction Model we started off by running the stats models on all of the variables from the World Happiness Report versus Happiness Score. This gave us an Adjusted R-squared of .714, as seen in table 5. As mentioned before, the Happiness Score is a sum of all of the variables, so it was expected that the r-squared would be high. We also looked at the p-values for each variable, and all were at zero, or very close to it. From this information we gather that none of the variables are having a negative effect on correlation.

Dep. Variable:	Happiness Score	**R-squared:**	0.716
Model:	OLS	**Adj. R-squared:**	0.714
Method:	Least Squares	**F-statistic:**	390.1
Date:	Sat, 14 Nov 2020	**Prob (F-statistic):**	1.44e-249
Time:	14:37:37	**Log-Likelihood:**	-846.18
No. Observations:	934	**AIC:**	1706.
Df Residuals:	927	**BIC:**	1740.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	2.0161	0.084	24.034	0.000	1.852	2.181
Economy (GDP per Capita)	1.2495	0.090	13.900	0.000	1.073	1.426
Social Support	0.6589	0.084	7.855	0.000	0.494	0.824
Health Life Expectancy	1.0139	0.143	7.076	0.000	0.733	1.295
Freedom	1.8870	0.147	12.874	0.000	1.599	2.175
Trust (Government Corruption)	-0.4468	0.129	-3.456	0.001	-0.700	-0.193
Generosity	0.7356	0.143	5.127	0.000	0.454	1.017

Omnibus:	23.312	**Durbin-Watson:**	1.205
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	25.811
Skew:	-0.339	**Prob(JB):**	2.48e-06
Kurtosis:	3.450	**Cond. No.**	18.8





From there we created two scatter plots comparing the Predicted versus Actual Happiness Score, and the Predicted versus Residual Happiness Score. In figure 17, we can see that the lower Happiness Scores were over predicted because most of the plots are above the regression line, while the higher Happiness Scores were under predicted because all of the plots are below the regression line. The predictive modeling works best on a Happiness Score that falls between four and seven. These plots are pretty evenly distributed above and below the regression line, or on it. Figure 18, shows that Happiness Scores between four and seven have the most points that are closest or on zero. This further implies that our predictive modeling works well for this Happiness Score range.

Prediction Model with Weather:

Using our correlation heat map we used a multivariate regression prediction model to investigate if multiple weather variables could predict Happiness Score. We determined that the "best" model included Latitude, Humidity, and Cloudiness. When we attempted to include any other weather variables the Adjusted R-squared value actually decreased. In table 6 it is important to point out the Adjusted R-squared value of 0.210 and different p-values for each variable included.

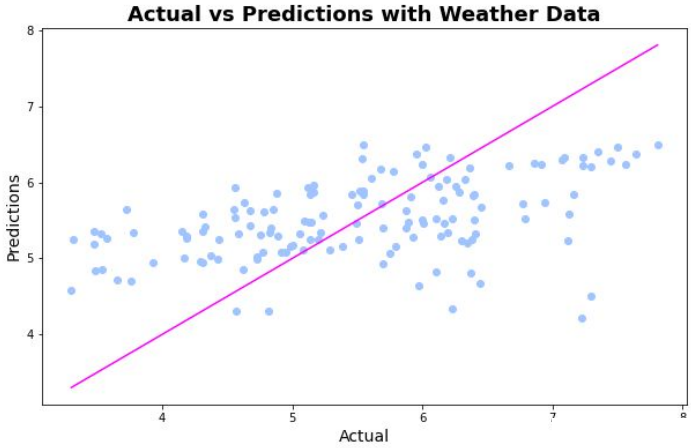
Dep. Variable:	Happiness Score	**R-squared:**	0.227
Model:	OLS	**Adj. R-squared:**	0.210
Method:	Least Squares	**F-statistic:**	13.30
Date:	Sat, 14 Nov 2020	**Prob (F-statistic):**	1.15e-07
Time:	14:59:17	**Log-Likelihood:**	-191.95
No. Observations:	140	**AIC:**	391.9
Df Residuals:	136	**BIC:**	403.7
Df Model:	3		
Covariance Type:	nonrobust		

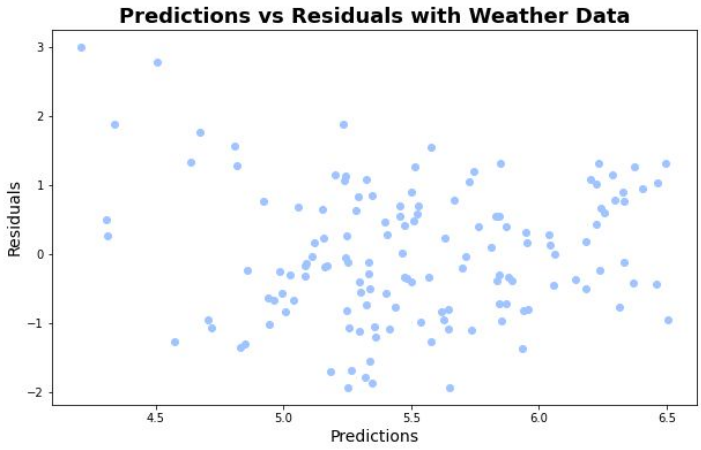
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	4.5600	0.289	15.805	0.000	3.989	5.131
Latitude	0.0174	0.003	5.321	0.000	0.011	0.024
Humidity	0.0062	0.004	1.509	0.134	-0.002	0.014
Cloudiness	0.0026	0.002	1.165	0.246	-0.002	0.007

Omnibus:	2.416	**Durbin-Watson:**	0.501
Prob(Omnibus):	0.299	**Jarque-Bera (JB):**	2.256
Skew:	0.311	**Prob(JB):**	0.324
Kurtosis:	2.970	**Cond. No.**	344.

Overall, we determined that our prediction model including the weather variables would only be accurate about 21% of the time, which indicated that our regression model is not the best. In figure 20 it is easy to see that our predictions do not cluster around the

actual values (pink line) and there is a large discrepancy between the two values. This is also visualized in figure 21 where the residual values do not cluster around the 0 value, further proving that our regression model cannot be validated.





Conclusion:

For our hypotheses we can now conclude that we were correct that happiness has increased over time, weather does not positively affect happiness, and that our ability to predict happiness is questionable because it is dependent upon the variables.

Limitations:

There were a few limitations that we came across in our exploration of this dataset. These were in part due to our own lack of mastery over the method, but also due to the data itself. The dataset was limited because it was trying to quantify a feeling that everyone defines differently. While it may at first seem like the responses to the questions given were prone to subjectivity, we instead believe that the questions asked were. The dataset was also not uniform across all 6 years. Some countries were only reported for one year, and we didn't use the raw dataset for this reason.

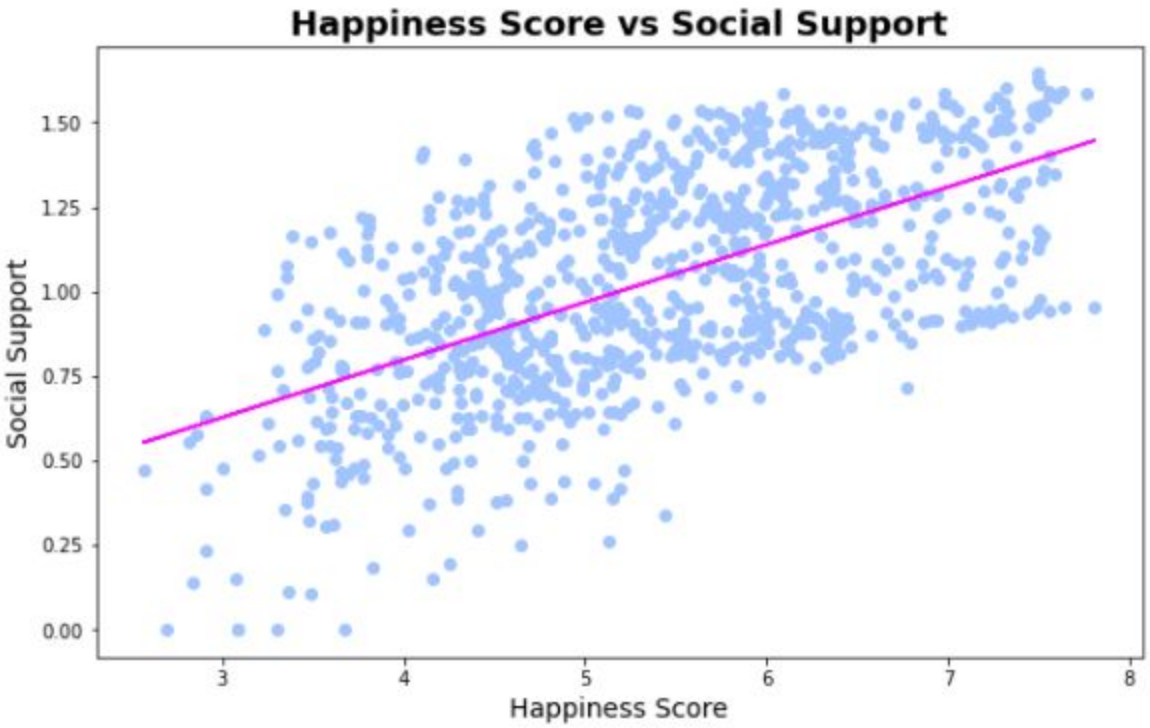
Future Work:

In the future we would like to expand our weather data by accumulating it over the same 2015-2020 time period and see if there really is any correlation between weather and happiness.. We would also bring in other factors such as education, housing, etc. to see if there is any connection between them and Happiness Scores.

References:

- **Country Codes:** https://raw.githubusercontent.com/plotly/datasets/master/2014_world_gdp_with_codes.csv
- **Google Maps API:** https://cloud.google.com/maps-platform/
- **Inspiration:**
- https://au.toluna.com/thumbs/8116344/There-Is-No-Path-To-Happiness,-Happiness-Is-The-Path
- https://hms.harvard.edu/coronavirus
- **Kaggle Notebooks:**
    - https://www.kaggle.com/alexisbcook/interactive-maps
    - https://www.kaggle.com/mathurinache/world-happiness-report
    - https://www.kaggle.com/londeen/world-happiness-report-2020
- **Open Weather API:** https://openweathermap.org/api
- **World Happiness Report:** https://worldhappiness.report/

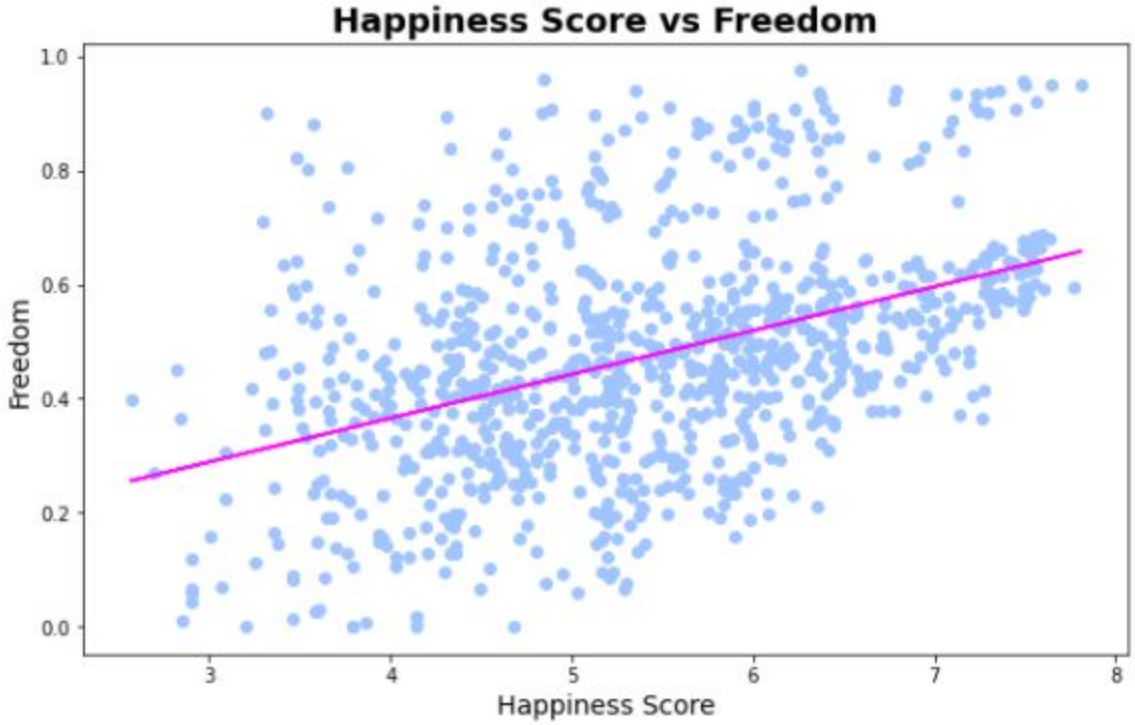
Appendix



Dep. Variable:	Social Support	**R-squared:**	0.356
Model:	OLS	**Adj. R-squared:**	0.355
Method:	Least Squares	**F-statistic:**	514.8
Date:	Sat, 14 Nov 2020	**Prob (F-statistic):**	4.36e-91
Time:	14:37:25	**Log-Likelihood:**	-58.885
No. Observations:	934	**AIC:**	121.8
Df Residuals:	932	**BIC:**	131.4
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	0.1156	0.041	2.793	0.005	0.034	0.197
Happiness Score	0.1704	0.008	22.689	0.000	0.156	0.185

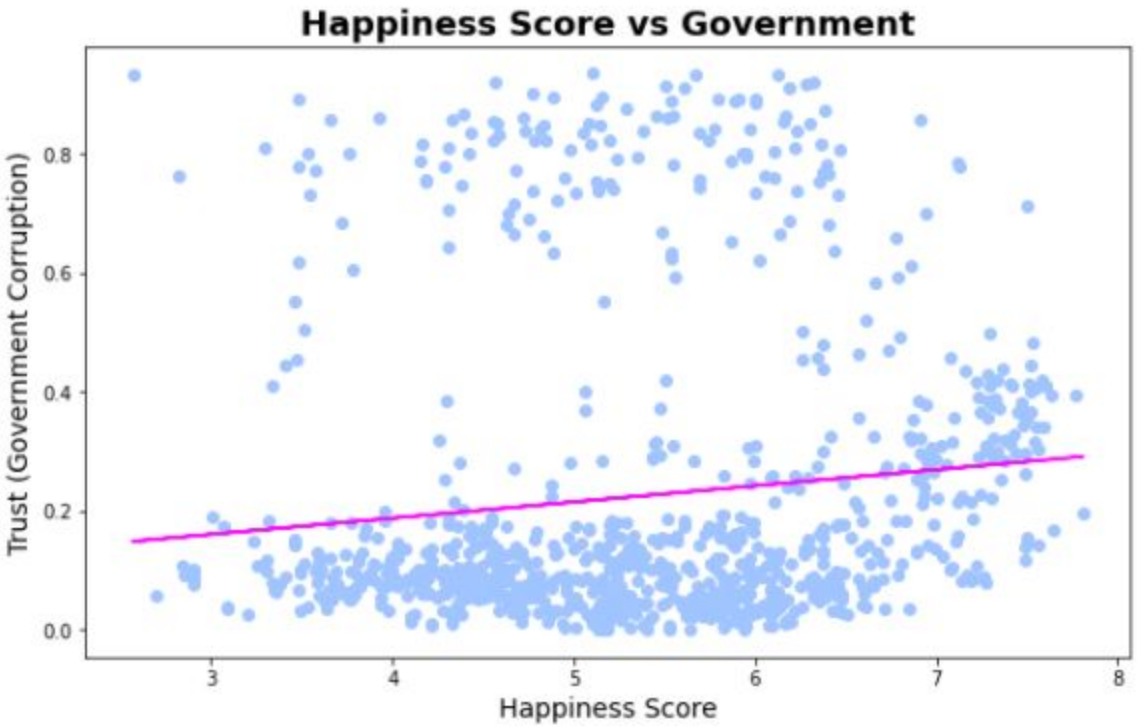
Omnibus:	34.036	**Durbin-Watson:**	0.925
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	19.327
Skew:	-0.186	**Prob(JB):**	6.35e-05
Kurtosis:	2.401	**Cond. No.**	27.9



Dep. Variable:	Freedom	**R-squared:**	0.183
Model:	OLS	**Adj. R-squared:**	0.182
Method:	Least Squares	**F-statistic:**	209.1
Date:	Sat, 14 Nov 2020	**Prob (F-statistic):**	6.62e-43
Time:	14:37:25	**Log-Likelihood:**	263.77
No. Observations:	934	**AIC:**	-523.5
Df Residuals:	932	**BIC:**	-513.9
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	0.0576	0.029	1.967	0.049	0.000	0.115
Happiness Score	0.0769	0.005	14.460	0.000	0.066	0.087

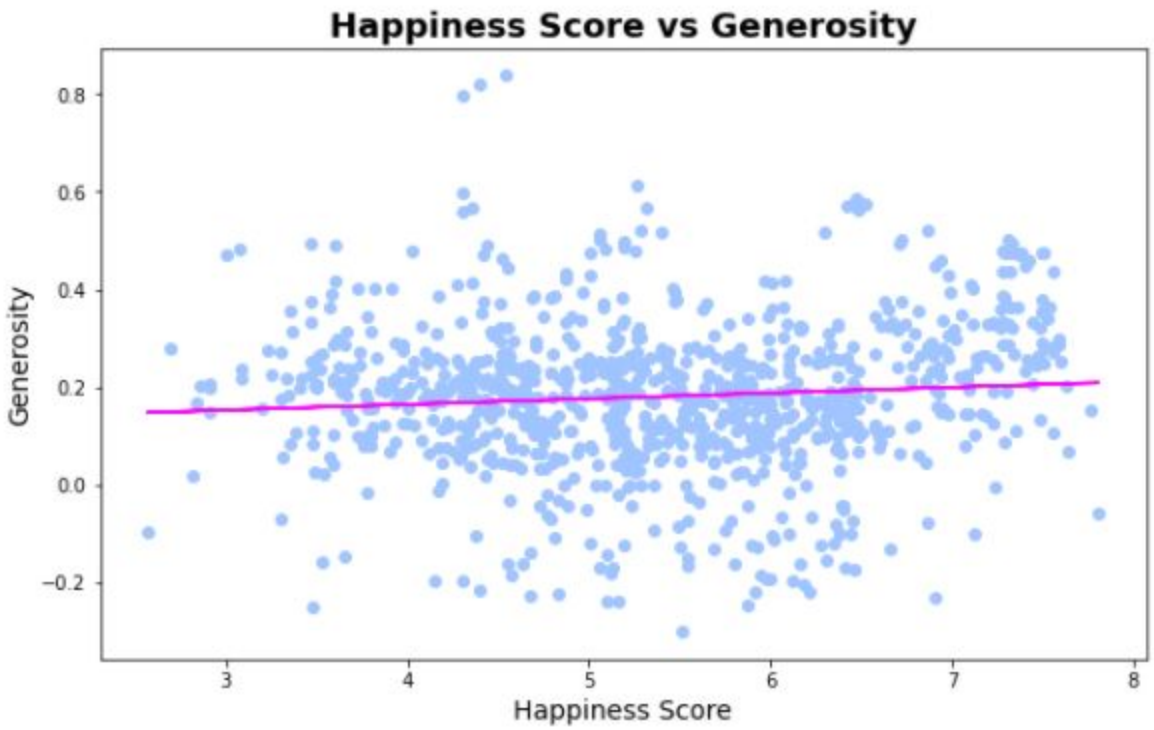
Omnibus:	35.311	**Durbin-Watson:**	0.905
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	38.636
Skew:	0.498	**Prob(JB):**	4.08e-09
Kurtosis:	3.033	**Cond. No.**	27.9



Dep. Variable:	Trust (Government Corruption)	**R-squared:**	0.014
Model:	OLS	**Adj. R-squared:**	0.013
Method:	Least Squares	**F-statistic:**	13.65
Date:	Sat, 14 Nov 2020	**Prob (F-statistic):**	0.000233
Time:	14:37:26	**Log-Likelihood:**	-41.497
No. Observations:	934	**AIC:**	86.99
Df Residuals:	932	**BIC:**	96.67
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	0.0781	0.041	1.923	0.055	-0.002	0.158
Happiness Score	0.0272	0.007	3.694	0.000	0.013	0.042

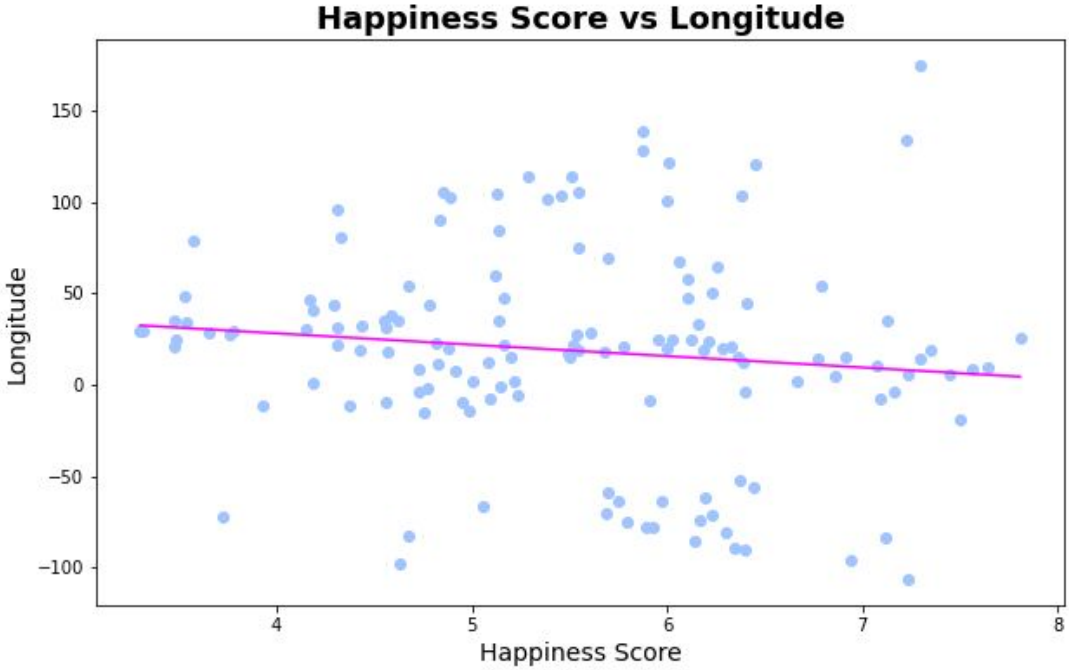
Omnibus:	246.829	**Durbin-Watson:**	0.283
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	472.757
Skew:	1.615	**Prob(JB):**	2.20e-103
Kurtosis:	4.311	**Cond. No.**	27.9

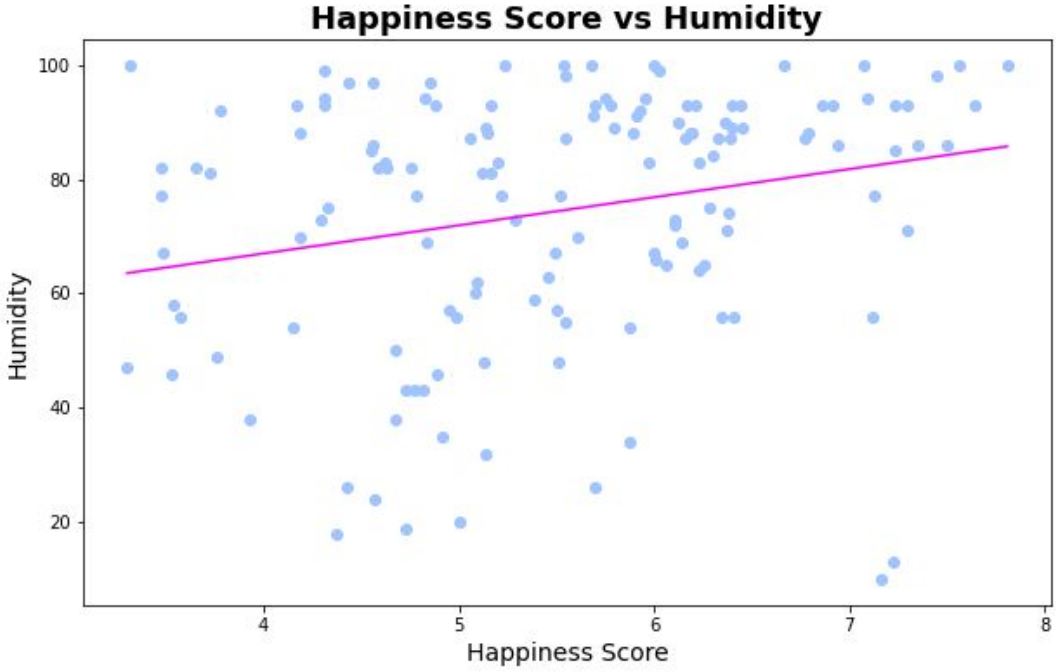


Dep. Variable:	Generosity	**R-squared:**	0.007
Model:	OLS	**Adj. R-squared:**	0.006
Method:	Least Squares	**F-statistic:**	6.814
Date:	Sat, 14 Nov 2020	**Prob (F-statistic):**	0.00919
Time:	14:37:26	**Log-Likelihood:**	425.58
No. Observations:	934	**AIC:**	-847.2
Df Residuals:	932	**BIC:**	-837.5
Df Model:	1		
Covariance Type:	nonrobust		

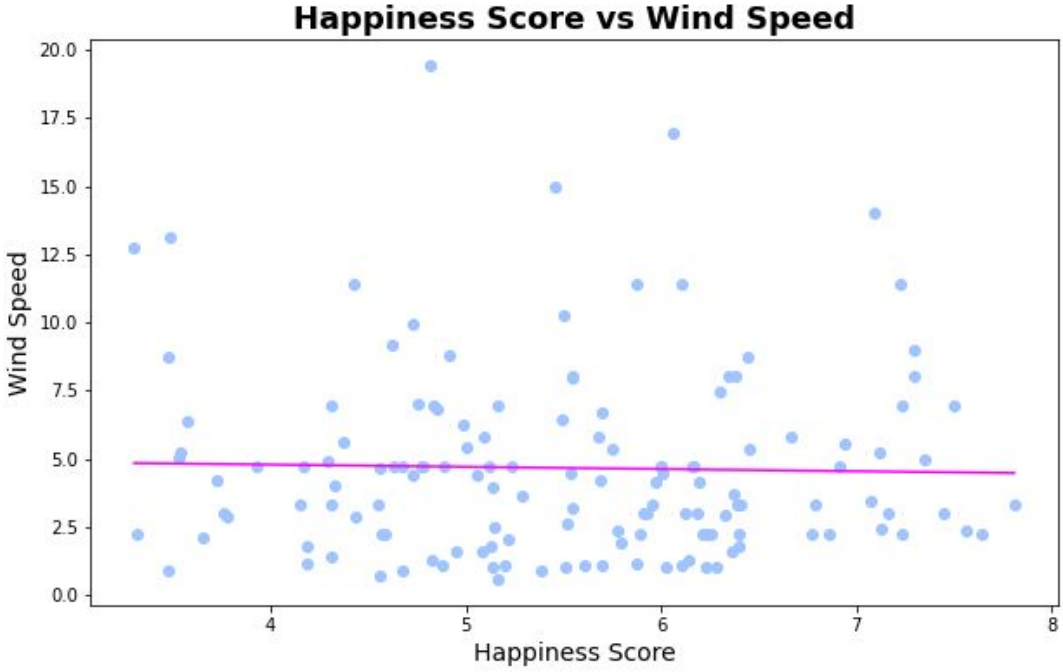
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	0.1175	0.025	4.770	0.000	0.069	0.166
Happiness Score	0.0117	0.004	2.610	0.009	0.003	0.020

Omnibus:	28.411	**Durbin-Watson:**	1.258
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	67.327
Skew:	0.025	**Prob(JB):**	2.40e-15
Kurtosis:	4.314	**Cond. No.**	27.9


Happiness Score vs Longitude
Longitude
Happiness Score
150
100
50
0
−50
−100
4
5
6
7
8


Happiness Score vs Humidity
Humidity
Happiness Score
100
80
60
40
20
4
5
6
7
8


Happiness Score vs Cloudiness
Cloudiness
Happiness Score
0
20
40
60
80
100
4
5
6
7
8


Happiness Score vs Wind Speed
Wind Speed
Happiness Score
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
4
5
6
7
8