WORLD HAPPINESS REPORT
FINAL PROJECT IST652

Rohini Shrivastava
Alejandro Mora
Rebecca Burnett

INTRODUCTION:

The word ‘happiness’ is used to categorize the mental states of a person. It is related to positive emotions like contentment to euphoria. It is hard to quantify exactly how people feel happiness and what factors cause these mood changes. However, through research and dedication there have been studies done to see how each country can try to appeal to its citizens.

The World Happiness Report is a report that ranks 156 countries based on the citizens' responses. It uses various factors, such as GDP, Mortality, etc to help show the influences these variables may have. All these factors are ranked against each other to show which countries are rated the happiest and which are rated the least happiest. The World Happiness report takes a sample of 2000-3000 people per country and asks them to rank their happiness on a scale of 1 to 10, with 10 being the happiest.

BUSINESS QUESTIONS:

A few questions have been created in order to fully understand happiness and how it has changed over time between countries.

1. Who is the happiest? Who is the least happy?
2. What is the strongest variable on happiness? What is the least influential variable?
3. How has COVID impacted the variables associated with happiness?
4. What are people’s overall sentiment in regards to COVID? What about the report itself?
5. How have people’s feelings changed between 2019 and 2021?

METHODOLOGY:

Data Import:

Three CSV files were downloaded from WorldHappinessReport.com. These CSV files were then read into a pandas dataframe. Tweets were pulled from Twitter using *Twint,* a twitter scraping tool that lets users pull tweets without using the API. The tweets tagged with ‘#covid’, “#WorldHappinessReport’, ‘#2021’.’#2019’ were pulled and put into a pandas dataframe.

Data Cleaning:

The 2021 world happiness report had 20 columns while the 2019 report only had 9 columns. The excess columns were dropped from the 2021 dataframe and the remaining columns were renamed to align more closely with the foundational column names. In the 2019 dataframe there was an extra unnecessary “overall rank” column that was also dropped to reduce the file down to more similarities. The 2019 data did not contain regional data therefore to compare the North America and ANZ region we had to create that region by grouping those variable rows into their own named dataframe for comparison.

The tweets pulled from twitter were placed into a pandas dataframe. Only tweets written in English were kept. Within the tweets, words that contained the hashtag were removed. Examples of this would be words like ‘COVID19’, ‘WorldHappiness’, ‘Report’, etc. Along with removing related words, any links inside the tweets were removed as well.

DATA ANALYSIS:

Initial Data Exploration:

To begin answering some of the easier questions, the top 10 highest countries happiness scores were selected and grouped with some additional variables that might begin to assist in understanding correlations to the happiest countries. We compared 2019 alongside 2021 to determine if there were any significant changes to metrics pre and post covid identification.

2019

```
1  happiest2019
```

Country or region	Healthy life expectancy	Freedom to make life choices	Generosity	
Finland	0.986	0.596	0.153	7.769
Denmark	0.996	0.592	0.252	7.600
Norway	1.028	0.603	0.271	7.554
Iceland	1.026	0.591	0.354	7.494
Netherlands	0.999	0.557	0.322	7.488
Switzerland	1.052	0.572	0.263	7.480
Sweden	1.009	0.574	0.267	7.343
New Zealand	1.026	0.585	0.330	7.307
Canada	1.039	0.584	0.285	7.278
Austria	1.016	0.532	0.244	7.246

Name: Score, dtype: float64

2021

```
1  happiest2021
```

Country name	Healthy life expectancy	Freedom to make life choices	Generosity	
Finland	0.741	0.691	0.124	7.842
Denmark	0.763	0.686	0.208	7.620
Switzerland	0.816	0.653	0.204	7.571
Iceland	0.772	0.698	0.293	7.554
Netherlands	0.753	0.647	0.302	7.464
Norway	0.782	0.703	0.249	7.392
Sweden	0.763	0.685	0.244	7.363
Luxembourg	0.760	0.639	0.166	7.324
New Zealand	0.785	0.665	0.276	7.277
Austria	0.782	0.640	0.215	7.268

Name: Ladder score, dtype: float64

Finland takes the win both years along with many other European countries. What is also noticed is that the happiness score and the freedom to make life choices both seem to slightly increase in 2021. It turns out that Finland actually never enacted a nationwide lockdown and it reacted quickly to only the region where the first reported cases existed and locked that region down in mid-March strictly for 1 month and then loosened it’s criteria the more and more each month thereafter. It seems contradictory that freedom to make life choices would increase post government lockdowns of any sort but perhaps after a lockdown the gradual reintroduction of freedom renews appreciation of even the most simple freedoms taken for granted.

On the other hand, it appears that the healthy life expectancy and generosity scores have declined since 2019. Covid very likely would be a contributing factor to the life expectancy and perhaps generosity declined scores can be attributed to lockdowns, social distancing, and facemasks.

The next question is also easily determined by evaluating the bottom 10 countries to see who is the least happy in 2019 and 2021 and how those might look comparatively.

2019

```
1 leasthappy2019
```

Country or region	Healthy life expectancy	Freedom to make life choices	Generosity	
South Sudan	0.295	0.010	0.202	2.853
Central African Republic	0.105	0.225	0.235	3.083
Afghanistan	0.361	0.000	0.158	3.203
Tanzania	0.499	0.417	0.276	3.231
Rwanda	0.614	0.555	0.217	3.334
Yemen	0.463	0.143	0.108	3.380
Malawi	0.495	0.443	0.218	3.410
Syria	0.440	0.013	0.331	3.462
Botswana	0.538	0.455	0.025	3.488
Haiti	0.449	0.026	0.419	3.597

Name: Score, dtype: float64

2021

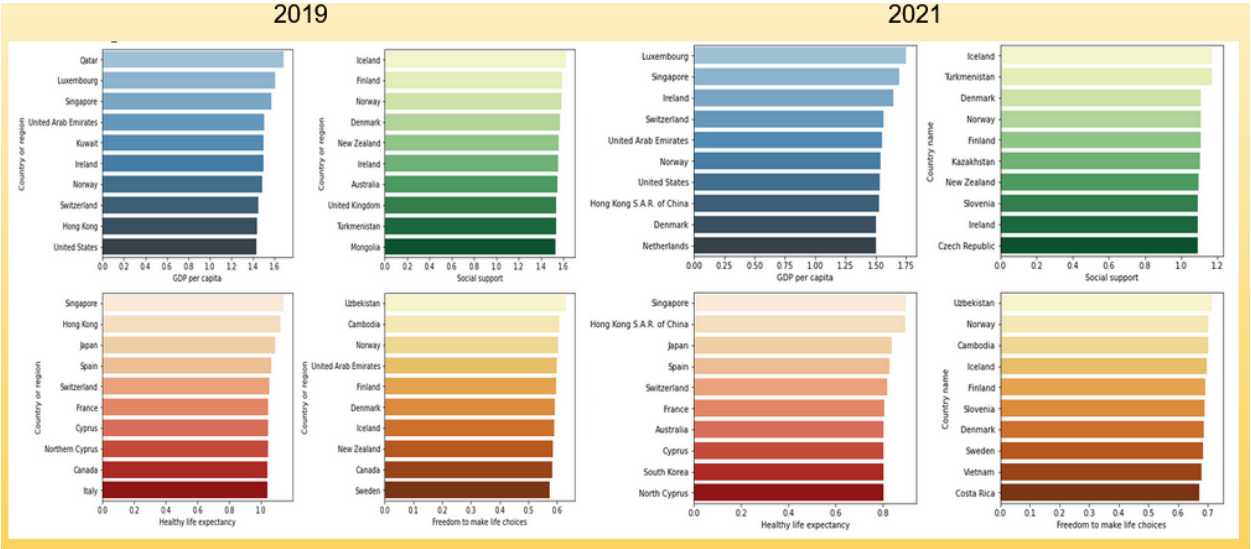
```
1 leasthappy2021
```

Country name	Healthy life expectancy	Freedom to make life choices	Generosity	
Afghanistan	0.126	0.000	0.122	2.523
Zimbabwe	0.243	0.359	0.157	3.145
Rwanda	0.407	0.627	0.227	3.415
Botswana	0.340	0.539	0.027	3.467
Lesotho	0.007	0.405	0.103	3.512
Malawi	0.298	0.484	0.213	3.600
Haiti	0.227	0.257	0.463	3.615
Tanzania	0.300	0.549	0.307	3.623
Yemen	0.272	0.268	0.092	3.658
Burundi	0.155	0.298	0.172	3.775

Name: Ladder score, dtype: float64

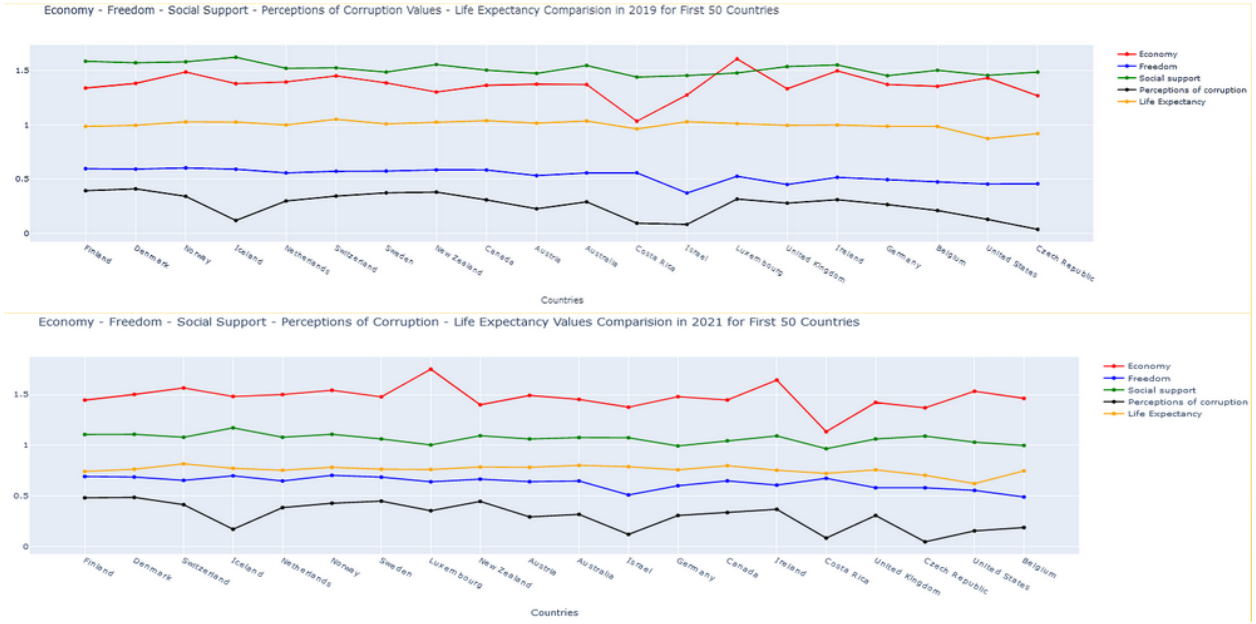
Just as the most happy analysis, the least happy proved to be consistent. The happiness rating and the freedom to make life choices have slightly increased (with the exception of Afghanistan). Healthy life expectancy and generosity have also declined since 2019.

After looking at the bottom and the top and noticing some score change trends, some graphs were made to compare the top countries by some variable scores side by side to validate if these changes appear to be consistent without only looking at the rank.

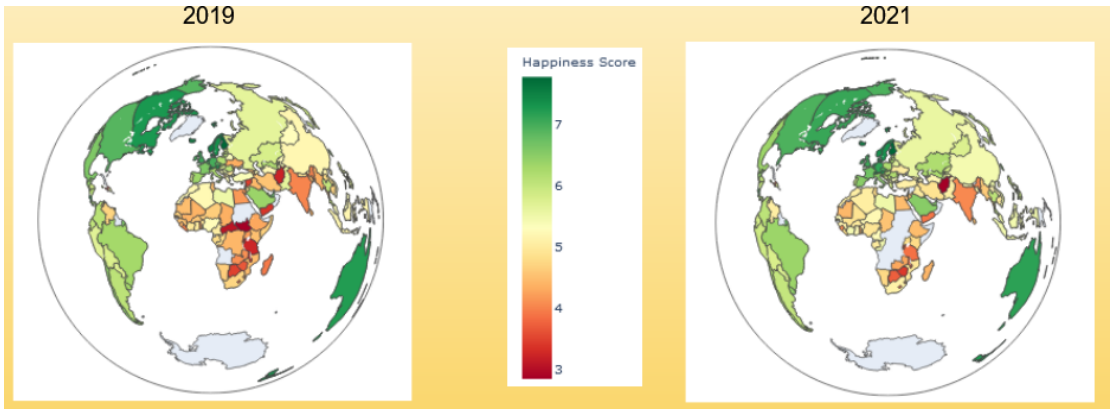


From the chart above, it can be seen that although the graphs appear to be similar, the scale does change along the x-axis. GDP per capita and freedom to make life choices increase in 2021 while healthy life expectancy and social support both decline in 2021.

Below, another visualization was created aimed in showing how the top 20 countries' variables compare between years and the decline in social support and life expectancy are clear as they both move down the y-axis measurement scale while overall it seems economy, freedom, and perceptions of corruption did not vary significantly.



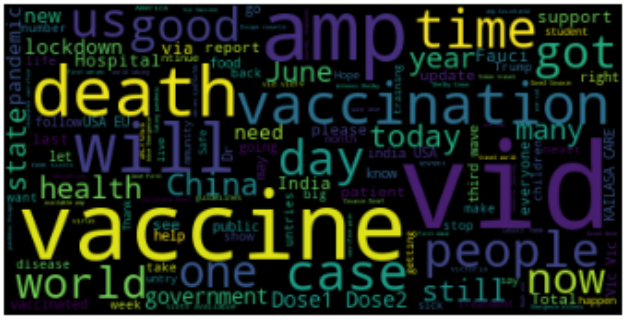
A quick global side by side shows relatively easily that happiness scores seem to slightly increase from 2019 to 2021. This change is most visible by comparing the orange country colors in Africa.



For the tweets gathered, a word cloud was created for each hashtag. This made it easier to understand what the most common words used were.

For '#covid', the most common words were vaccine, Death, good, India, government. These most used words show that covid has impacted many but there positive feelings coming

through, possibly due to the vaccine being distributed. It also discusses government involvement with the pandemic.

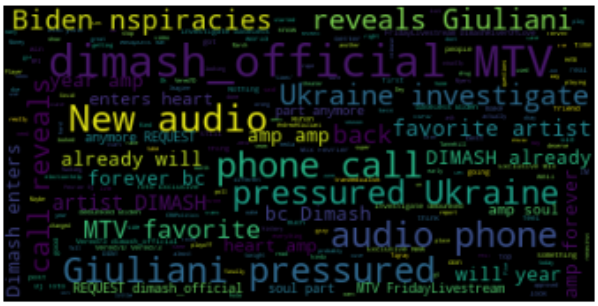


covid

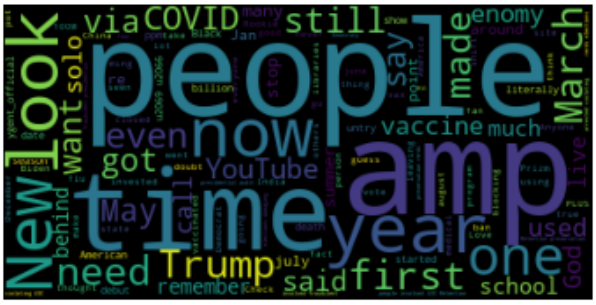
When using #worldhappinessreport, we saw Finland mentioned a lot. Both in 2019 and 2021 this was the happiest country listed. Other countries were also listed, but the interesting finding was the work ‘pandemic’. People may be discussing how the pandemic affected the reports or how the numbers changed.

World happiness report

Both word clouds for #2019 and #2021 had a lot of political elements to it, specifically American politics. Words like Trump were common between the two years, however 2021 focused more on words like COVID. Since the COVID19 pandemic didn’t occur until 2020, it's understandable no one would be referencing it in 2019.



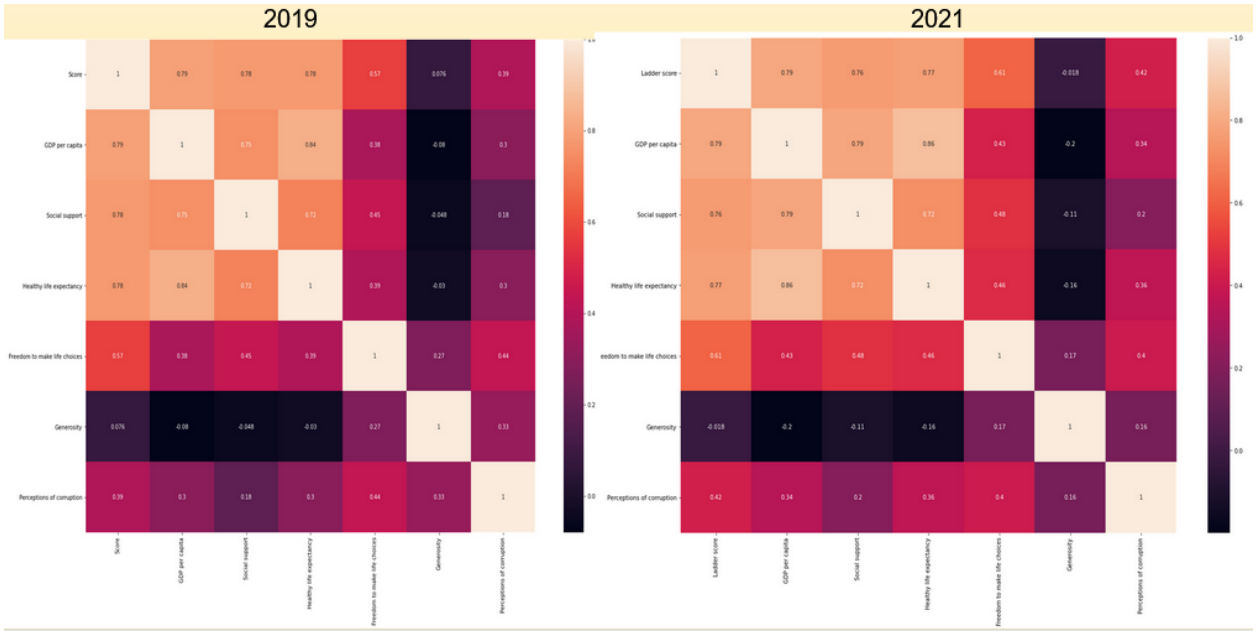
2019



2021

Correlation:

Correlation matrices were produced in order to show statistical comparisons between variable assumptions from the EDA above.



The above correlation matrix comparison shows:

- No change: GDP
- Increased strength: social support, healthy life expectancy, generosity went from negative to positive
- Decreased strength: freedom to make life choices, perceptions of corruption

Linear Regression:

With having already looked at correlation plots to measure coefficients of how close some variables are to Ladder scores, or happiness scores, some models can be made to help further analyze the variables and see which ones are most or least significant to happiness. A linear

regression model can be useful as we can try to predict a score based on multiple variables and compare how much weight they contribute to the equation of happiness. These models are to help further analyze which variables are worth looking at when comparing country happiness, they will not be looked at to solve for a specific country's happiness. Two packages will be used to see which linear regression model is best to use and to see which one gives the most information to help with our questions. The packages are Scikit-Learn and Scipy with Statsmodels; the questions that will be further answered will be:

-What is the strongest variable of happiness?

-What is the least influential variable?

This will be a continuation of looking at the correlation plots to help back up the findings of significant variables.

In order to see what variables are worth putting into a regression equation, the correlation plots can be viewed. The variable of the Ladder score will be seen as the dependent variable and all the other variables will be seen as independent variables to see the correlation between them. There are five variables that show a decent correlation to Ladder score:

-Logged GDP per Capita(0.79)

-Social Support(0.76)

-Healthy Life Expectancy(0.77)

-Freedom to Make Life Choices(0.61)

-Perceptions of Corruption(0.42)

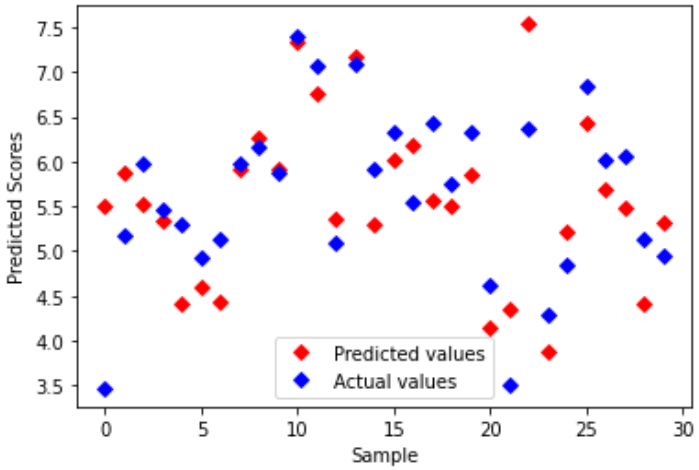
All these variables show decent correlation coefficients and can be assessed in the models. These variables can be put into their own dataframe to help with ease of model building between both packages. The data is now split into training and test sets for the models using a train_test_split() function in SKLearn. This splits the data into the specified 80/20 percent split(80% of data is in the training set, 20% is in the testing data set). Once the data is randomly split up, the sets are split again into sets containing all the independent variables and sets of "labels", or in this case, the Ladder scores. This is to train the set on the variables(X_train) and test the accuracy on the labels(Y_train); the testing sets are to see how they can predict against the test sets' variables(X_test) and compare with the actual scores in the testing set's label set(Y_test).

First, using SKlearn and its LinearRegression() function, the first model can be created. The model takes the arguments of fitting the X_train and Y_train data sets to create the model on the training data. SKlearn has the option to see model accuracy and display a accuracy score, for this data set the model created an accuracy of 0.7861 accuracy. This is fairly good for a predictive model with this little amount of predictive variables. Using the predict() function next we can just get a picture of what this model can predict based on these variables. SKLearn also provides functions to calculate some metrics created by the model. These metrics are able to be calculated based on the model from SKLearn:

```
Mean Absolute Error: 0.4981
Mean Squared Error: 0.4073
```

```
Root Mean Squared Error: 0.6382
```

A plot can also be created using Matplotlib and the comparison between predicted values and actual values can be seen; this is based off of the data from test data that was predicted based on the already fitted model.



Moving onto another package Scipy using statsmodels, a different type of linear regression can be created and different specifics from the model can be viewed.
When using all the variables used in the model before, a model can be created by just using the data frame without having to split the data itself.

OLS Regression Results

Dep. Variable:	Ladder_score	R-squared:	0.754
Model:	OLS	Adj. R-squared:	0.745
Method:	Least Squares	F-statistic:	87.50
Date:	Mon, 07 Jun 2021	Prob (F-statistic):	9.60e-42
Time:	01:44:06	Log-Likelihood:	-117.17
No. Observations:	149	AIC:	246.3
Df Residuals:	143	BIC:	264.4
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.4114	0.181	13.334	0.000	2.054	2.769
GDP_per_capita	0.7554	0.246	3.075	0.003	0.270	1.241
Social_support	1.1119	0.296	3.751	0.000	0.526	1.698
Healthy_life_expectancy	0.9334	0.423	2.207	0.029	0.098	1.769
Freedom_to_make_life_choices	1.7531	0.397	4.413	0.000	0.968	2.538

Perceptions_of_corruption	1.0466	0.447	2.340	0.021	0.162	1.931

In this type of model created, the coefficient values and significance of each variable can be viewed to see what variables are contributing more than others in the equation.
If a standard alpha value of 0.05 were to be set for the boundary of significance, all variables have a p-value of below 0.05 which shows that they are all significant in this test. All these variables are significant and show that they are correlated to the variable of Ladder Score. Instead of an accuracy score with this model, the OLS function allows for the viewing of the R-squared value, this shows how well the movement of an independent variable moves with the dependent. This model shows an R-squared of 0.754 which shows a good comparison of the independent and dependent variables.

With all this being said, another model can be made to test and see if a better R-squared value can be obtained and also compare AIC values. Taking out the highest of the five variables based on p-values within the last model, we can be left with three variables to review; GDP_per_capita, Social_support, and Freedom_to_make_life_choices. This is just to see if a more significant model can be created using the same data with less predictors.

OLS Regression Results

Dep. Variable:	Ladder_score	R-squared:	0.733
Model:	OLS	Adj. R-squared:	0.728
Method:	Least Squares	F-statistic:	133.0
Date:	Mon, 07 Jun 2021	Prob (F-statistic):	1.95e-41
Time:	01:44:20	Log-Likelihood:	-123.05
No. Observations:	149	AIC:	254.1
Df Residuals:	145	BIC:	266.1
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.3878	0.186	12.809	0.000	2.019	2.756
GDP_per_capita	1.2482	0.185	6.762	0.000	0.883	1.613
Social_support	1.0398	0.297	3.499	0.001	0.452	1.627
Freedom_to_make_life_choices	2.2064	0.383	5.757	0.000	1.449	2.964

This model shows to have a little lower of an R-squared value and also a higher AIC value. For AIC, the lower value is better when comparing models of the same data. This just shows that taking out the variables takes away from the significance of the model.

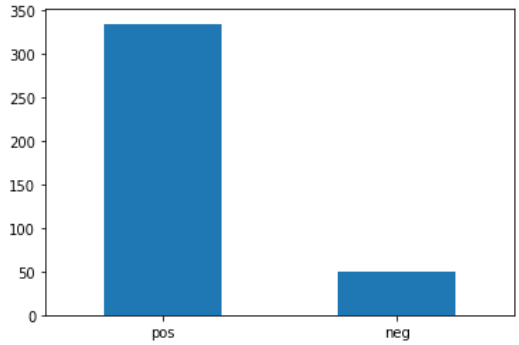
When looking at both types of models(Sklearn and statsmodels), both have advantages and downfalls. Statsmodels allow the ability to see how significant each variable is to the equation in the model. For this purpose, it is best to use this type of model. The question is which variables

show to have the least and most significance to happiness score. Starting with the correlation plot, the one variable that was not included into the linear regressions was Generosity, this variable showed little to no correlation from the beginning so it would be considered the least valuable variable. After the creation of the models, Social Support showed to have a really low p-value, decently low standard error compared to the other variables, strong correlation to Ladder score, and the second highest coefficient Beta weight in the model. With this, it can be said that Social Support can be considered the strongest variable when predicting and solving for happiness score.

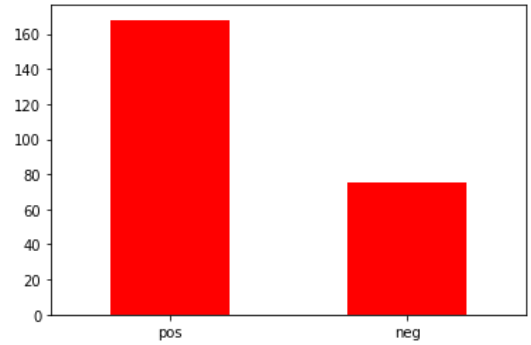
Sentiment Analysis:

VADER (Valence Aware Dictionary for Sentiment Reasoning) was used to perform sentiment analysis on the tweets. This is a model that is sensitive to polarity and intensity of text. In order to run the sentiment analysis using VADER, the NLTK package was installed and imported into the python script.

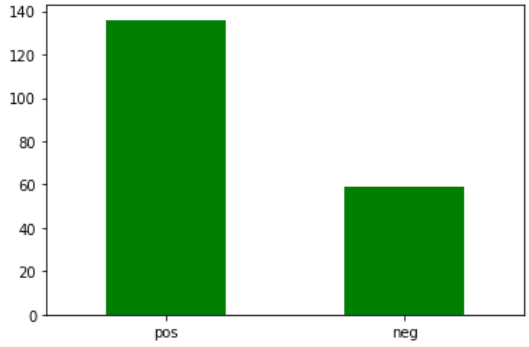
For all four of the hashtags searched, the tweets had more of a positive sentiment than negative. The closest in distribution was for 2021 with a 90 positive 55 negative difference. The greatest difference was the World Happiness Report with 330 positive and 50 negative tweets.



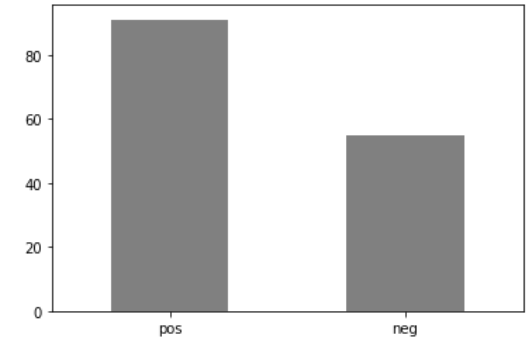
World Happiness Report



COVID



2019



2021

CONCLUSIONS:

Looking at the question of what country is the happiest for 2019 and 2020, we can say that it is Finland based on the happiness score. Based on the same scores, Afghanistan in 2021 and South Sudan of 2019 showed to have the lowest scores from the rest.

After analyzing the variables correlated to Ladder score, or happiness score, there can be an association between some variables and the scores. In our questions before analyzing the data, we were trying to find the strongest and least influential variables towards the scores. After correlation plots and linear regression models, the least influential variable is the Generosity variable. This variable lacks any significant production nor deduction of happiness score so we can say that is most likely the least influential based on correlation. For the strongest variable, we can say that it is most likely the Social Support having a really good correlation towards the happiness score and influence on the linear regression models. With that being said, it is surprising for the least valuable as Generosity seems like it would go a long way for happiness.

Overall it appears happiness itself has increased along with the variables GDP and freedom to make life choices, even with the COVID pandemic. Healthy life expectancy, social support, and generosity have declined as we expected. COVID has caused an increase in the number of deaths around the world, and forced people into lock down in some areas. We can see how life expectancy and social support have declined through this, however it was interesting to see freedom to make life choices increased.

Looking at the word graphs, a lot of thoughts about COVID are about vaccines and death. Looking at the sentiment report however it shows these thoughts are more positive than negative. Using both information together we can assume that the number of deaths are decreasing and the vaccines are giving people more hope, causing these higher positive tweets.

With the World Happiness Report we expected more positive sentiment and that is what we saw. Finland, being the most happy country in the world, popped up a lot in our searches. The ratio between positive and negative tweets was the highest out of all the searches we made.

Between 2019 and 2021 people seem to be less political in their thoughts. 2019 contained almost exclusively political thoughts. 2021 has a few mentioning it, but overall it seems more focused on the future, with words like “look” and “time” being popular. 2021 also seems to be more negative than 2019. While both categories had more positive tweets than negative, 2021 had a smaller gap between the number of each compared to 2019.

FUTURE STEPS:

Bring in several more years of world happiness data to build a better understanding of trends over time and also strengthen our model prediction capabilities. Using only 2019 and 2021 gave us a very limited view on happiness. As there was no report for 2020 generated, we don’t have a solid understanding of what baseline to use either.

The tweets that were used for analysis were all written in English. Through this there is definite bias on what people are discussing, as it gives all of us an understanding of our topics from a very Western viewpoint. Along with this, only the most recent 500 tweets were taken. This also adds a lot of bias since they may be focused on only a few recent happenings. It would be better to grab the most liked/viewed tweets or focus only on tweets generated around the

relevant dates. For 2019, we would grab tweets made back then. For COVID we can discuss trends from 2020 when it started to now. And for the World Happiness Report we can compare thoughts from the time the report was released in each of the years.