Aidan Wall, Ishaan Chainani, Matt Gold

DATASET:

As of now, we will be using the world happiness report data for our project. There is a chance we instead use spotify data, because we all like music and would like to dig deeper into a product we use so much.

2021 data Column Description:

- 'Country name': name of country
- 'Regional indicator' : what region country is in
- 'Ladder score': ladder survey with the best possible life for them being a 10, and the worst possible life being a 0
- 'Standard error of ladder score': standard error of ^
- 'Upperwhisker': Lower Confidence Interval of the Happiness Score
- 'Lowerwhisker': Upper Confidence Interval of the Happiness Score
- 'Logged GDP per capita': logged GDP per person
- 'Social support': The extent to which social support contributed to calculation of the Happiness Score
- 'Healthy life expectancy': The extent to which Life expectancy contributed to the calculation of the Happiness Score
- 'Freedom to make life choices': The extent to which Freedom contributed to the calculation of the Happiness Score.
- 'Generosity': The extent to which generosity contributed to calculation of the Happiness Score
- 'Perceptions of corruption': The extent to which Perception of Corruption contributes to Happiness Score
- 'Ladder score in Dystopia': The extent to which Dystopia Residual contributed to the calculation of the Happiness Score, scored as a ladder from 1-10.

Historical data columns: ['Country name', 'year', 'Life Ladder', 'Log GDP per capita', 'Social support', 'Healthy life expectancy at birth', 'Freedom to make life choices', 'Generosity', 'Perceptions of corruption', 'Positive affect', 'Negative affect']

These questions can be about the relationships between variables, or how well one thing can predict another, clustering...etc, but note that in your final project you must use at least **1 supervised learning model** (includes both regression and classification models), **1 clustering model, and 1 instance of dimensionality reduction** (PCA or LASSO), so keep that in mind when creating questions. You can use more than one of these for a single question (e.g. using PCA and then doing linear regression on the components).

Aidan:

1. What are main factors that contribute to a country's happiness and which is the most important
2. How do some happiness factors affect other factors?
3. Is there any correlation between happiness and region of the world? How much?
4. Between regions of the world, do regions have stronger/weaker or interesting correlation to other variables (happiness, corruption, etc.
5. Are there any signs of happiness being subjective to the individual? Some individuals are happy with no social support, or generosity, etc.

Matt:

1. Are there differences in happiness between wealthy and poor countries, and what factors affect this.
2. Are there any discrepancies in people saying they're happy when data should say they're not (what are some reasons for this/can it be cultural)
3. Are certain cultures/groups pre decided if happy or sad even if their circumstances change(will always answer happy even if things get worse)
4. How has solely GDP affected worldwide happiness, if a country has exponential happiness growth is it caused by GDP?
5. How is each factor influenced by region?

Ishaan:

1. Looking at the top 5 wealthiest countries, besides wealth, how do they differ from each other on other factors and how does that determine happiness? Or is it just wealth.
2. Look at the correlation between GDP and corruption, does it occur in poorer, or richer countries, or both?
3. Might be hard to tell without adding additional data, but does language spoken have any impact on the happiness of individuals? (say this because I listened to a podcast on how different languages shape your thoughts and attitudes and would be interesting to dive deeper into this.
4. Is the model accurate as determined by accuracy score, silhouette score, confusion matrix, or whatever measurement is best. Is it more accurate for some countries/regions than others?
5. Maybe not able to tell, but food definitely has a large influence on people's happiness, are people over eating, under eating, what are you eating, and does this affect your happiness.
6. How would the model perform if GDP was not considered? Or any variable with money and just look at social factors (freedom to make choices, generosity, etc.)

- a) describe the analysis you're planning (include details like whether you're using standardization, regularization, model validation, distance/similarity metrics, how you'll choose clusters or hyperparameters, which variables you're using...etc)
    - Use these variables as predictors, then use this model and get feature importance etc

- b) explain **why** this analysis and the choices you described above are good and explicitly **how** these methods will answer the question.
    - Which variable most impactful this method will do that
- c) describe **two** ggplot data visualizations you'll use to support your answers (graphs must be in ggplot, the ONLY exception is a dendrogram for HAC).

- For Part I, #3 You should come up with the plan together and EACH submit a copy of the same plan. You need to answer 3 questions **per member** (so a group of 2 would need to answer 6 questions, a group of 3 would need to answer 9).

PART 1 #3

1) What are main factors that contribute to a country's happiness and which is the most important
    a) First I will z-score all the variables to make sure they are on the same scale. I will then use KFold Cross Validation on my data to reduce bias and make sure I am not leaving any countries out from my training and testing data (in TTS you do not use all the data in either. Then I will do either PCA or look at the coefficients in my linear regression model to see which has the most importance. I will be able to compare them to each other because I have z-scored the variable. If I use PCA then I will create a scree plot and either use the elbow method or the percentage method to determine how many features to use in my model. I might also use the LASSO model to remove any non important variables because it automatically sets those variables to 0. LASSO will remove small, unimportant effects from the model and allow me to grab and compare coefficients of the remaining variables to compare their strength (since they are z scored, the coefficients are more comparable).
    b) I will use regression because my output variable is continuous. I will use KFold cross validation because unlike TTS, this allows me to use all the data when training my model, and because I don’t have large amounts of data, this seems like a better option. After looking at the coefficients of this model, if any of them seem of interest to look at, either large or small, I will do PCA to determine if I want to keep all of these variables in my model. If some features are unimportant then they will be removed and our model will be revised.
    c) I will then create a bar chart showing the importance of each variable to visualize their relationship to each other. If I decide to do PCA then I will make a scree plot to determine how many features to keep in my model. I can also plot the expected values vs actual values to look at the error of my model.
2) Are there any signs of happiness being subjective to the individual and that some individuals are happy with less social support, or generosity, money, etc. and are just more genuinely happy?

a) I will use a clustering algorithm to compare happiness to other factors such as GDP, 'Social support', 'Healthy life expectancy at birth', 'Freedom to make life choices', 'Generosity', 'Perceptions of corruption', to see if there are any distinct clusters and if there are groups of people who are happy with more of something, less of something or anywhere in between. I do not know which clustering algorithm I will be using, because I have to look at the data in a graph before making a decision on which to use and how many clusters to make etc. If I use HAC I will use the euclidean distance metric and if I use DBSCAN I will use the elbow method to determine the EPS. I will probably have a min samples that is not that high because there are not that many different data points, but will play around with specific values that give me the highest silhouette score.

b) In order to see if there are trends that might not be visible, and see if there are differences within groups of different people on how many resources people have we have to use a clustering algorithm. This will help visualize the effects between different variables and happiness, and will hopefully show us some interesting trends.

c) I will show graphs comparing the different variables and their relationships to happiness, and see if there are any distinct groups. I will then apply the clustering algorithm and show the clusters that the algorithm has identified.

3) Look at the correlation between GDP and corruption, does it occur in poorer, or richer countries, or both, or in some regions?

a) I will use the z-scored data to look at a scatterplot graph comparing GDP and corruption, and then add a trend line to see if there is any obvious correlation between GDP and corruption. I will also make a correlation matrix comparing all variables and their effect on each other so I will be able to see what variables cause an increase or decrease in corruption. I will also graph the data by region and see if some regions have higher perception of corruption, and if that is caused by higher/lower other variables.

b) Doing this will show me the correlation between GDP and corruption. Between plotting the graph and the coefficient of the correlation matrix I will be able to tell if there is a correlation between GDP and corruption or corruption and any other variables.

c) I will make a scatter plot, with GDP as the X variable and Perception of Corruption as the Y variable. This will show me if there is an increase of perception of corruption as GDP increases or its vice versa. I will also make a correlation matrix but graph and color it to show the impacts each variable has on each other.

d)

4) Is each model accurate as determined by accuracy score, silhouette score, confusion matrix, or whatever measurement is best. Is it more accurate for some countries/regions than others?

a) For each model I run I will look at model score, accuracy score, R2, silhouette score, MSE or whatever measurement seems to be best depending on the situation. I will also use my model to test data separating countries by region, to see if it does better or worse depending on the region. I also might divide countries into 3 tiers by GDP, poor, wealthy, and middle. If the model does better/worse at predicting any of these values in the different tiers, I will dive deeper into why this is, and if for each of the different tiers they have different factors influencing their happiness

  - i) For ex: wealthy countries might not have to worry about GDP as much, so will they place importance on other features like social ones?

- b) I think this is the best way to validate my model, by comparing all the different accuracy measurements, depending on which ones are applicable. I will compare the different accuracy metrics, and see if my model under or over performs on some groups or areas.
- c) For this question I will graph the different accuracy scores/r2 values on a barchart to show which model is the most accurate and describe why each model is performing the way it is.

5) How would the model perform if GDP/any variable associated with money was not considered? (only look at freedom to make choices, generosity, etc.)
   - a) Build another model (either linear regression, without GDP and then see how it performs. I will probably create another linear regression, using KFold cross validation and then look at the R2 score to see how it performed. I will also look at how the coefficients have changed, and which are the most impactful without GDP. I will create a df with predicted and actual values to use later to graph.
   - b) This is the best way to do it because I want to see how things are when money is not a factor. Is it still possible to predict Happiness without money and is that the most important factor?
   - c) I will show the same graphs as above, showing the coefficients for each variable (minus GDP). I will also plot the expected values vs actual values for my model to show if the model is performing well.
6) Use the best model we have found from above to predict data for 2021(completely new unseen data so see how well it truly performs)
   - a) Historical and 2021 data have the same predictors and columns, so I will use the data model created from historical data to try and predict happiness scores for 2021 data. I will have to do all previous steps of model validation and finding the best model before completing this step. I will then compare the predicted values for the 2021 data and the predicted 2021 data. I will then plot this on a graph, comparing the expected values vs the predicted, and probably look at the MSE to see how well the model did.
   - b) I think this is a good way to tell how good the model worked. It is training on completely historical data, then being used to predict unseen new data. If it is not accurate, it will tell us more about our model and about the data as a whole.
   - c) I will make a graph comparing predicted values to actual values, and see how our model performed and if there are any trends. Maybe it tells us if there is overfitting. I will also make a graph comparing happiness over time, and see if there has been a hit in 2021 due to the pandemic or if this tells me anything interesting about my data.

7) Because we have 2020 and 2021 data (2021 is the new unseen data) now has the coronavirus pandemic affected overall happiness and some of the factors within each country?
    a) I will build another model again, probably a linear regression but only looking at the unseen current data. I will use KFold to train this new model. I will also build a correlation matrix looking at all variables. I will look to see what the data tells me about how happiness has changed in the last year, was there a shift in monetary vs societal/health factors that mattered to happiness? I will compare the values from my previous model, and see if some values are lower and higher than current ones
    b) If GDP is still the biggest factor, I wonder what this tells us about humans as a whole, that even when all of our lives are at risk, is money still the most important thing in our life, Or does money affect the other factors such as health. I will compare the numbers from last year, the coefficients and the values from the correlation matrix and think this is the best way to analyze if humans have shifted their behavior.
    c) I will graph the coefficients of the model comparing last year and this year to see how they have changed, probably a bar part overlaying each year. I will also build a correlation matrix but use coloring the make it known what variables have largest impact on Happiness
    d)

8) Are certain cultures/groups pre decided if happy or sad even if their circumstances change(will always answer happy even if things get worse)
    a) Are there any countries or regions where, although the individual scores of the factors decrease, their happiness does not increase? I will look within different regions, if the average score of individual variables is increasing/decreasing but the happiness score does not reflect that. I will do this by plotting all the variables and how they have tracked over time, while overlaying how the happiness is doing over time. I will look at how the graph is trending at the end and if it shows me anything interesting.
    b) I think this is the best way to answer this question, because some people might have a firm stance of happiness, and it is built into their culture, and that this happiness will not be affected by outside things. In order to answer this I have to look at different countries and regions, and see how their happiness compared to other variables has changed over time.
    c) I will show this by graphing the trendlines for each variable for a specific country or region tonight and overlap that with a trendline for happiness. I will do this for as many regions or countries to try and see if this trend is a thing.

9) What was the most challenging part of this assignment/ what are some expected things you didn't know you were going to have to deal with that are interesting?
    a) This is a pretty vague question but I know as I dive deeper into this data I will uncover some things that I did not expect to and will write about them here. Some challenges that I predict to see are the division of regions and looking at more specific subsets of the data and trying to find patterns.

b) I think it is important to look at what was challenging for us and why it was challenging. This will help us learn from what we could have done better and what we can do better next time

c) I will make some graphs hopefully showing what I was trying to do even if it may be unfinished, just to show what I was trying to make