World Happiness Report

Author:

Xingyuan Zhao (Student ID: 2039394, xingyz5@uw.edu)

Mariana Li Chen (Student ID: 2021951, Email: yq0522@uw.edu)

Wanjia Ruan (Student ID: 2038420, Email: wruan2@uw.edu)

Summary of Questions and Results

Q1: What is the relationship between different social indexes and Happiness scores?

The Happiness score measures life satisfaction and other domains such as health, social support, freedom and economy (GDP). We discovered the correlation between Happiness score which is the dependent variable and each of the social indexes which are independent variables by using the OLS methodology, histograms and scatter plots.

Having a look at the histogram, we can deduce that happiness scores seem to be normally distributed each year and tend to vary only slightly between years. This suggests that most people tend to be content with their lives from year to year. The slight variations in the data could be due to minor changes in economic conditions or other external factors such as the COVID situation. Moreover, according to the four scatterplots, we can observe that all the four social indexes are positively correlated with the happiness score.

Q2: How is the geographic distribution of Happiness scores around the world?

We explored the dataset and mainly focused on the Happiness score for each country from the years 2018 to 2021. We constructed an interactive map plot to visually show the average happiness score of each country during that period.

Based on the map, European and North American countries have higher average happiness scores as the color appears to be brighter and yellowish. The following countries are from the South American region, with average scores of 6.2 to 5.8 (with some exceptions). Asian countries have average scores ranging from 6.1 to 5.1 (with some exceptions). The continent of Africa has average scores of 5.4 and below.

Finland has the highest average score, around 7.83054566375, and Afghanistan has the lowest average score, around 2.5346975325.

Q3: Predict the Happiness score using Economy(GDP), Social support, Health(Life Exp.), and Freedom.

We will mainly focus on each of the social indexes in the dataset and use the Multilinear Regression Model and logistic regression model to obtain the coefficients which allows us to discover the relationship among the variables and generate equations for future calculation.

The linear regression model gives us an equation to calculate the happiness score: '-3.70041 + 0.29422 x GDP_log + 2.8898 x social + 0.03666 x life_expectancy + 2.24249 x freedom'.

With respect to marginal effects, the logistic regression model indicates that social support is the one that contributes most to achieving a score above average, followed by freedom, GDP(log), and life expectancy. Since life expectancy is a relatively large number, a single change may contribute less to the outcome than other factors. Additionally, the country's continent determines whether the score will be above average or not. The chances of the country being higher than the average score are lower in Asia than in any other continent, followed by Europe, South America, North America, and Oceania.

Motivation

Happiness score is an important tool used by many researchers and policy makers to better understand each country's population's Happiness, well being, equality and many other aspects. This data helps them to implement more effective policies and encourage positive transformation that helps improve the society and country as a whole.

Everyone is an important part of the society and maintaining a higher Happiness score means that people are more satisfied with their lives. This helps create a better living environment for everyone such as reducing crime rates or suicide. It might also lead to superior work performances that increase productivity and might boost a country's GDP. It also encourages more cooperative behaviors such as volunteering, charity or simply people are more likely to help each other.

We are also part of the society and we understand that many people suffer from social, economic and mental pressures and some even health issues. That's why we believe that Happiness is extremely important. However, we can't imagine how many people in this world are still suffering and looking forward to reaching their Happiness. That's why we decided to focus on this topic which helps us to understand more about how people feel, to raise awareness of the importance of the Happiness index and encourage positive changes.

Dataset

The Dataset Link

World Happiness Report dataset:

https://worldhappiness.report/ed/2022/#appendices-and-data

(See Data for Table 2.1)

World Map by Countries:

https://www.naturalearthdata.com/downloads/110m-cultural-vectors/110m-admin-0-countries/ (Download Countries)

Testing datasets:

one.csv and two.csv (both are GitHub links)

All files are also available in our GitHub repository data folder.

More detailed information about those two datasets, please see Appendix.

Description of dataset

World Happiness Report dataset

The dataset we chose is from the *World Happiness Report*, it contains all statistical data GallUp gathered from 2006 to 2021 from 166 countries or regions.

The main focus for our project is the Happiness Score (Ladder Score), which is a national average of the responses to the main life evaluation question asked in the Gallup World Poll (GWP). The Happiness Score is calculated based on the following factors: GDP per capita, Healthy (Life Expectancy), Social support, Freedom to make life choices, Generosity, Corruption Perception, Residual error. Because of the inconsistency of the data, the data we used only include the four factors of GDP per capita, Social support, Healthy (Life Expectancy) and Freedom as social indexes from 2018- 2021 that focus on the COVID-19 impact on happiness.

World Map by countries

The geodataframe we used is from Natural Earth. A total of 258 countries are represented in this dataset, and it contains a great deal of information regarding each of these countries/regions. Countries distinguish between metropolitan (homeland) and independent and semi-independent portions of sovereign states.

We will primarily use the variables 'NAME', 'SUBUNIT', ‘CONTINENT’, and 'geometry' for mapping the geographical distribution of happiness scores in this dataset by merging it with the previous dataset. Since we are practicing joining these variables with the happiness score dataset, we used 'SUBUNIT' as the key instead of 'NAME' (A more intuitive choice). SUBUNIT matched more columns with the previous dataset, resulting in a reduction in the amount of non-values that were wasted.

Method

1. **Preparation**
    - Clean data and drop the NA values, select the column `Ladder Score` (renamed into `score`) and pick four indexes from 2018-2021 for calculating.
    - Join the previous dataset with world map GeoDataFrame (`world.shp` from `world.zip`) to get a new GeoDataFrame for each country’s Happiness score from 2018-2021 and drop NA values. (GeoDataFrame: `world_data`)

2. **What is the relationship between different social indexes and Happiness scores?**
    - Clean the datasets and drop the NA values (if have), select the column of Economy(GDP), Social Support, Health, Freedom, and Happiness Score from 2018-2022 for calculating.
    - Using histogram and scatter plot to show the relationship between these social indexes and Happiness scores.
    - This method would be the realization of how we are dealing with the multiple datasets and make the plot for analyzing how social index affects the Happiness score.
3. **How is the geographic distribution of Happiness scores around the world each year?**
    - Calculated the average score for each country using `groupby()` using `world_data` and make a copy of it with selected columns (save as `yeardata`)
    - Filter the `world_data` with map needed columns ('SUBUNIT' and 'geometry')
    - Join those two DataFrame together as an GeoDataFrame
    - Preset the GeoDataFrame using `from_features` for plotting
    - Use the `plotly` to make the map plot of each country's average Happiness score from 2018-2021.
    - Update the layout and functionality of the plot.
    - This would be the reflection of the first challenge goal of modifying the multiple data sets and using the new library to make the data interactive and visualized.
4. **Predict the Happiness score using Economy**(**GDP**), **social support, Health**(**Life Exp.**), **and Freedom.**
    - Clean the datasets and drop the NA values (if have), select the column of Economy(GDP), Social Support, Health, Freedom, and Happiness Score from 2018-2021 for calculating.

- Split the data into a training set and a test set.
- Using the mean() function to get the average number of the world happiness score.
- Using smf.logit() function to create a logistic regression model and use the get_margeff() function to get the table for the marginal effects for each variable.
- Using the Scikit-learn library, recreate the logistic regression model and create the confusion matrix based on the result.
- Calculate the accuracy, precision, recall, F-score, and MSE based on the logistic regression model and the confusion matrix.
- Using the Scikit-learn library, create the linearRegression Economy(GDP), Social Support, Health, Freedom, and Happiness Score as labels.
- Calculate the corresponding coefficient value usingScikit-learn library based on a multilinear regression model.
- Using the coefficient and intercept create the whole equation of happiness score and print it in the terminal.
  i. Y = β0 + β1 × Economy(GDP) + β2 × Health (Life expectancy) + β3 × Freedom + β4 × Social Support + ε
  ii. Y = Happiness Score
  B0 = intercept
  β1 = Economy(GDP)
  β2 = Health (Life expectancy)
  β3 = Freedom
  β4 = Social Support
- Use the result of the multilinear regression model to explain each factor and then use the predict function to calculate the predicted value for each factor.
- Calculate the MSE to tell the accuracy of the model predicting.

    - This would be an example of how we deal with the challenge goals of machine learning that will predict the Happiness scores based on four social indexes mentioned in the report.

Results

Note (Because the numerical values of our model are based on the results of prediction, each person's prediction results will be different every time they run the model. This leads to slight differences in our coefficients, intercepts, confusion matrices, and the results you get when you run the model. Our analysis results are based solely on the data we predict ourselves.)

Histograms and Scatter Plots

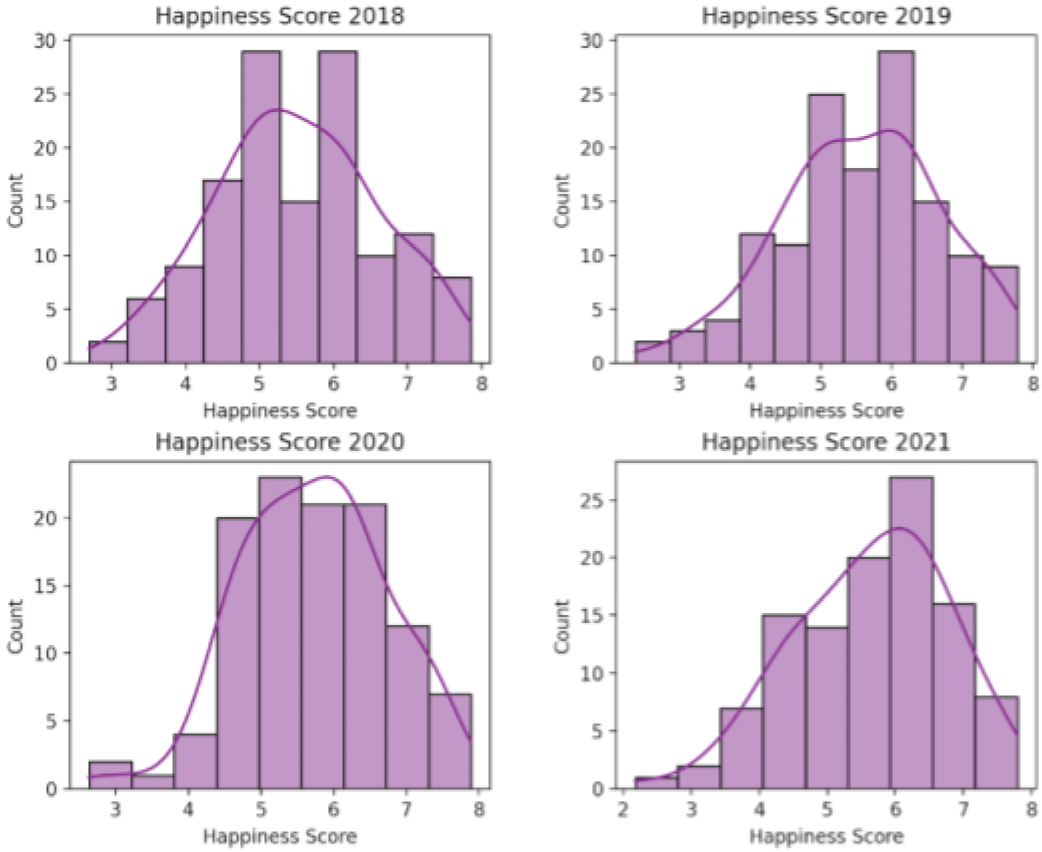


Figure 1: Distribution Histogram of Happiness Score (2018 - 2021)

Looking closely at the four histograms for each year between 2018 to 2021. We can observe that from the years 2018 to 2020, the histograms look more normally distributed and in 2021 the histogram looks more left skewed because its long tail is in the negative direction on a number line. This also reflects that from the year 2018 to 2021, the population happiness level tends to be more positive, people are more satisfied with their life in many aspects including the social index analyzed in the scatter plots below.

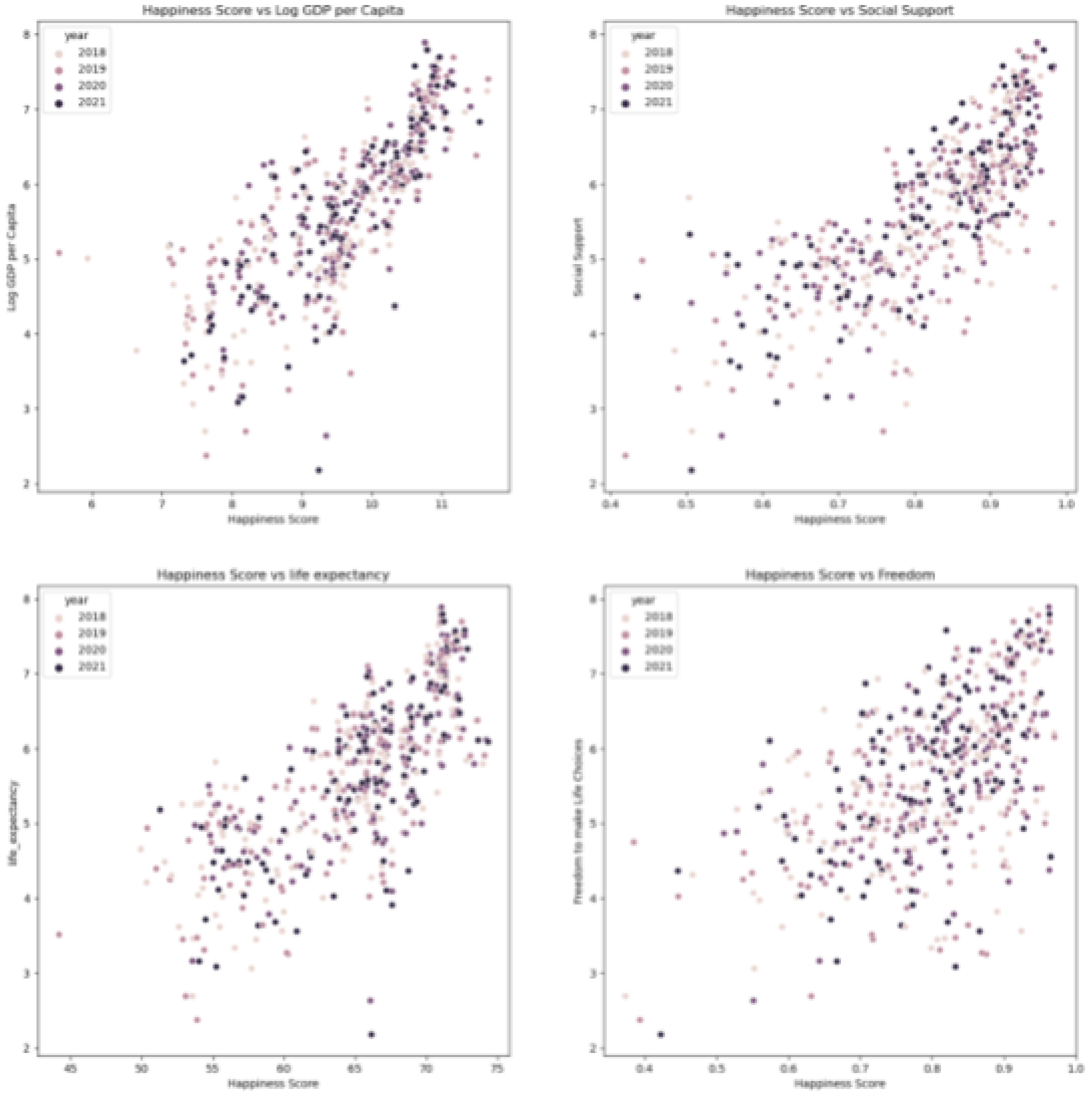


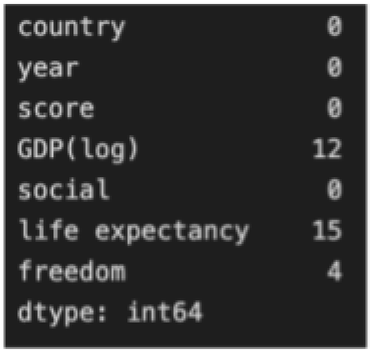
Figure 2: Scatter Plot for Four Social Indexes and Happiness Score

There are four scatter plots which show the correlation between each of the social index with the happiness score. After observing the four graphs we can conclude that all the four social indexes including Log GDP per Capita, Social Support, Life Expectancy and Freedom have positive correlations with the Happiness Score. This also means that all the independent variables in the analysis have an effect on the dependent variable. Therefore, happiness is influenced by lots of factors in life. Having these correlations in mind, policy makers can develop or change the policies based on these correlated factors to enhance population happiness.

Map

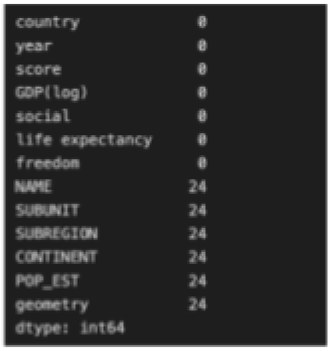
In order to determine the distribution of happiness scores across countries, we combined the happiness score dataset with the geodataframe that retains all columns in happiness score and only stores geometry and subunits for plotting using the function `join_data`.

Please note that the happiness score dataset does not contain all countries around the world, and that not all countries have a happiness score every year. As a result, there will be null values before and after the joining of those two datasets. Prior to joining the dataset, the happiness dataset contained 15 rows of null data for life expectancy, 12 rows of null data for GDP(log), and four rows for freedom. We decided to drop those values since they are not significant in comparison to the whole dataset (shape: (521, 7)). When the dataset is joined to the geodataframe, 24 additional rows of data appear to not match each other and are also dropped for the same reason. (See Figure 1 and 2 below). We then calculated the average score for each country for plot.



```
country              0
year                 0
score                0
GDP(log)            12
social               0
life expectancy     15
freedom              4
dtype: int64
```

Figure 3: Sum of null values before join



```
country            0
year               0
score              0
GDP(log)           0
social             0
life expectancy    0
freedom            0
NAME              24
SUBUNIT           24
SUBREGION         24
CONTINENT         24
POP_EST           24
geometry          24
dtype: int64
```

Figure 4: Sum of null values after join

Having joined the data and converted the data into a geodataframe, we used px.choropleth_mapbox to plot an interactive map with crs equal to WGS84 and mapbox_style equal to 'carto-positron' in order to ensure that the world geo-information is used as the base map. This map is colored gradiently to indicate the distribution of scores (see Figure 3).

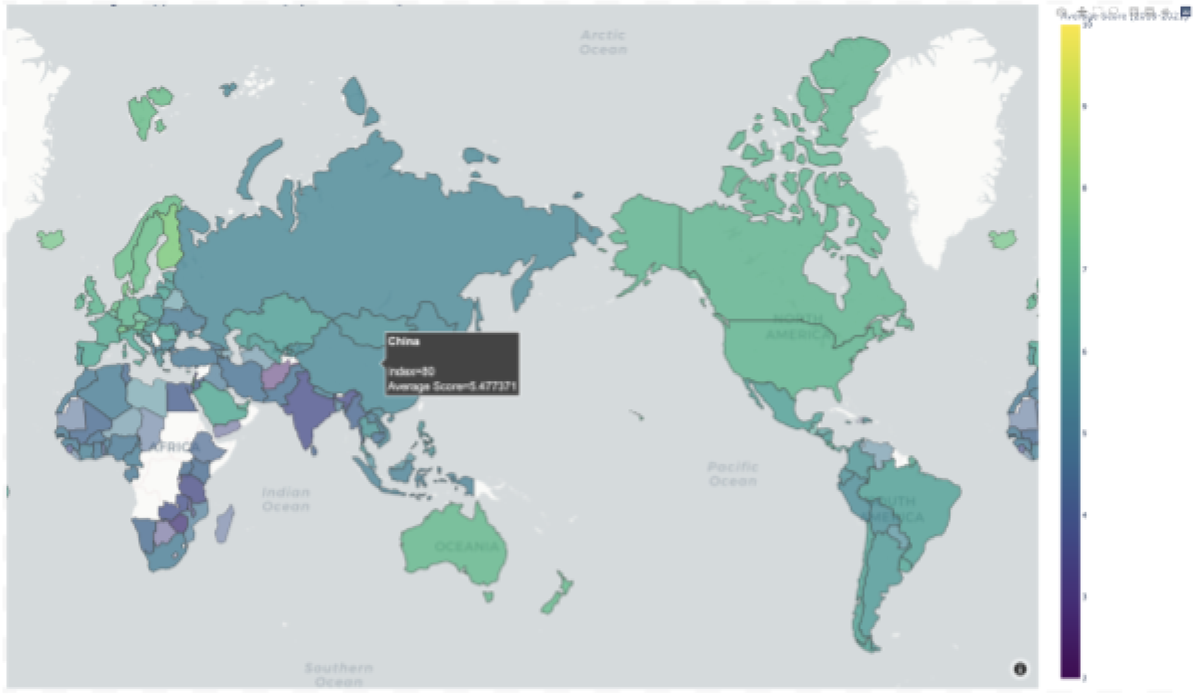


Figure 5: An interactive map showing the average happiness score around the world

From the map, we see the 'happiest' country is 'Finland' with average score 7.83 and the least happy country is 'Afghanistan' with average score 2.53.

Finland is famous for its well-developed living systems (medical, education, welfare etc.) which lead to relatively high life expectancy (range: 70 - 72) and social support values (range: 0.93 - 0.97). Also, freedom to make choices (range: 0.92 - 0.96) is high as people tend to have high commitment towards gender equality and inclusiveness with low levels of income inequality (high GDP_Log around 10.7).

Afghanistan, on the other hand, consistently ranks as the least happy country in the world. This is primarily due to the war and violence that occurs in the country. The war has resulted in a low life expectancy (range: 50-54) among the population. Economic insecurity, as reflected by a low GDP_Log (range: 7.3 - 7.6), causes people to feel despair and hopelessness. A frequent change of political power also results in a low social support score (range: 0.41 - 0.55), as people have limited access to basic infrastructure, education, and housing. People are also less able to make free decisions as a result of discrimination and social inequality (0.37 - 0.71).

Generally, yellowish-green colored regions are found in Europe (See Figure 4), North America (See Figure 5), and Oceania (See Figure 6), which indicate a higher average happiness score. According to these four indexes, western countries tend to have higher values in GDP, Life expectancy, Freedom, and Social Support. The standard of living in those countries is higher because they are experiencing higher levels of economic development. The provision of social support could also result in a more comprehensive approach to social warfare, including healthcare, education, and housing policies. These countries tend to be more open, with a higher level of civil liberties, freedom of speech, and human rights. Those values all contribute to a feeling of happiness for each individual.

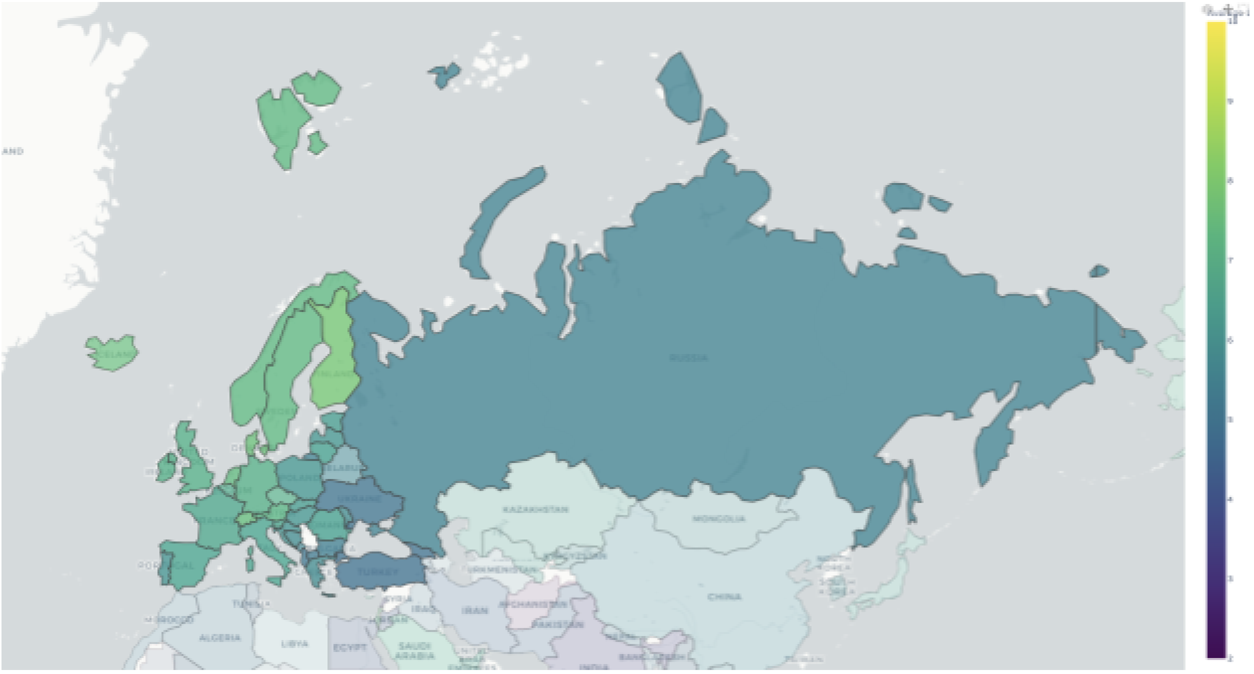


Figure 6: An interactive map showing the average happiness score for European countries

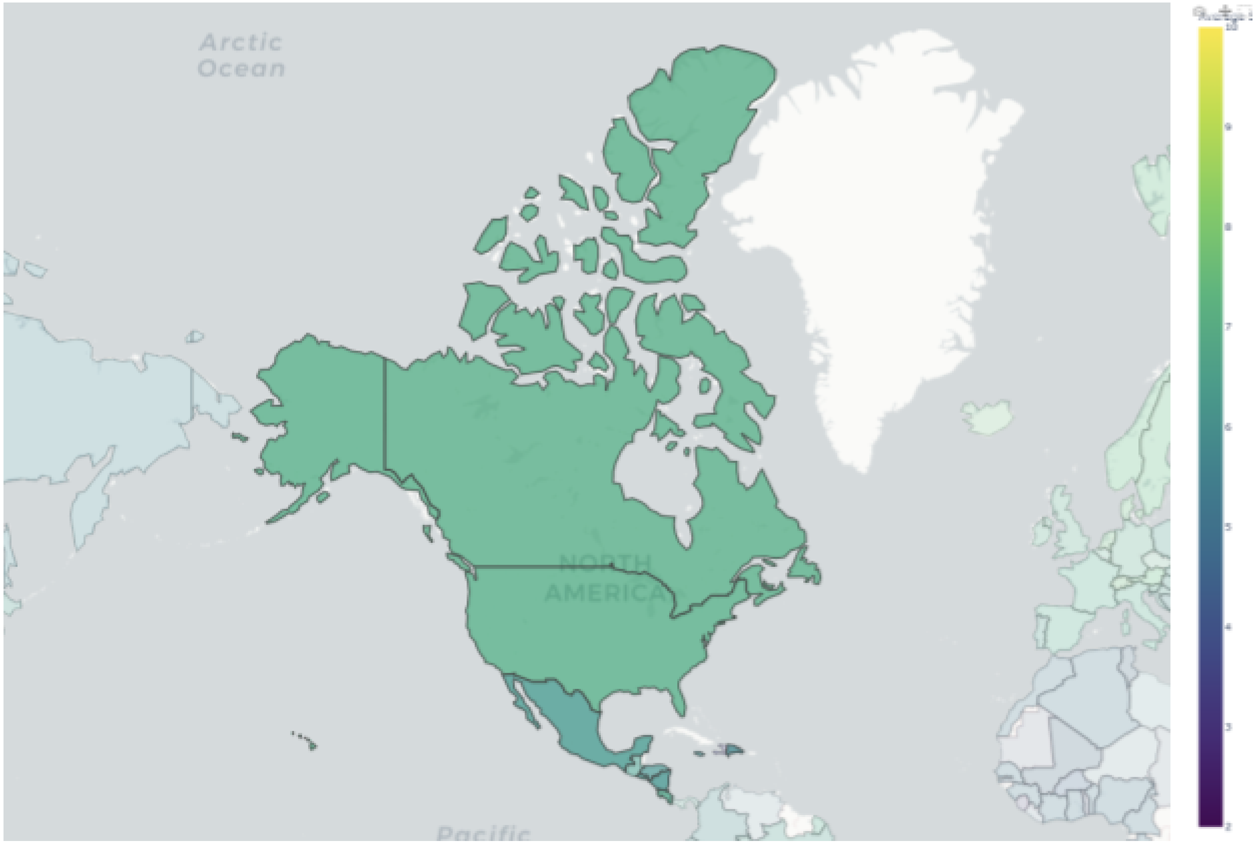


Figure 7: An interactive map showing the average happiness score for North American countries

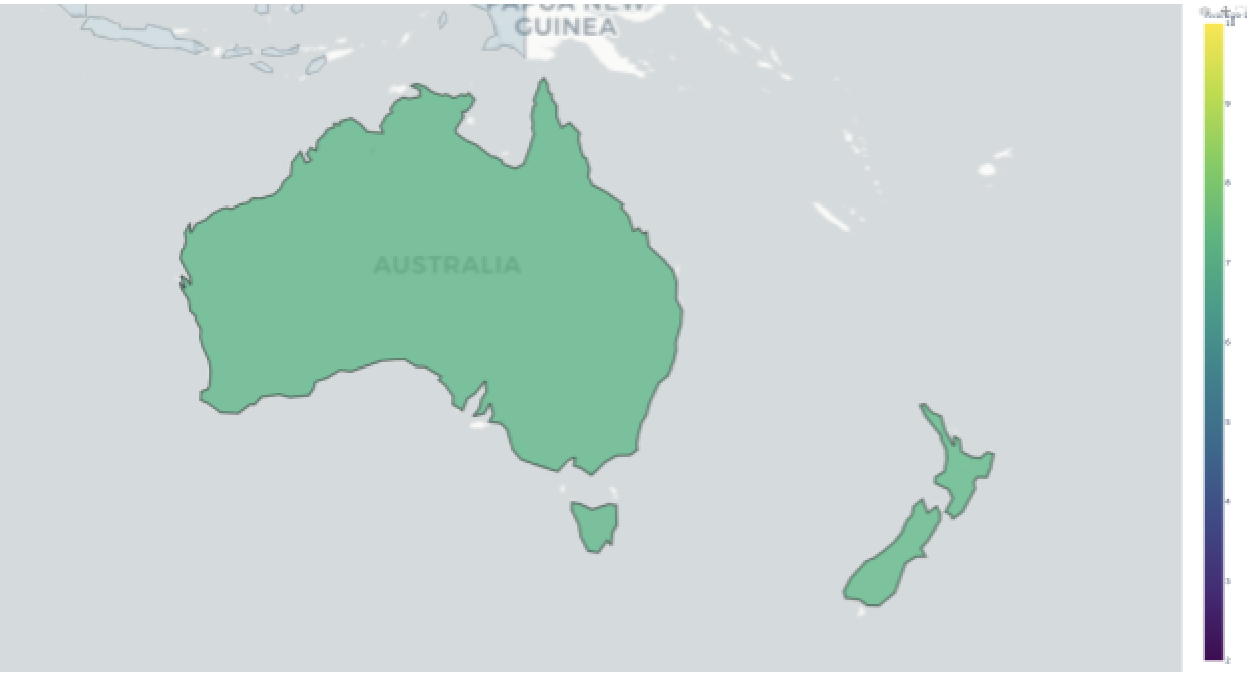


Figure 8: An interactive map showing the average happiness score for Oceania countries

After these three continents, South American (See Figure 7) and Asian countries (See Figure 8) tend to be ranked in the middle. This may be due to the fact that the majority of those countries are still considered developing countries, which makes it difficult to rank higher because of economic factors and health factors. In addition, some of those countries are experiencing political instability or are undergoing war, which makes them less likely to have high levels of freedom and social support.

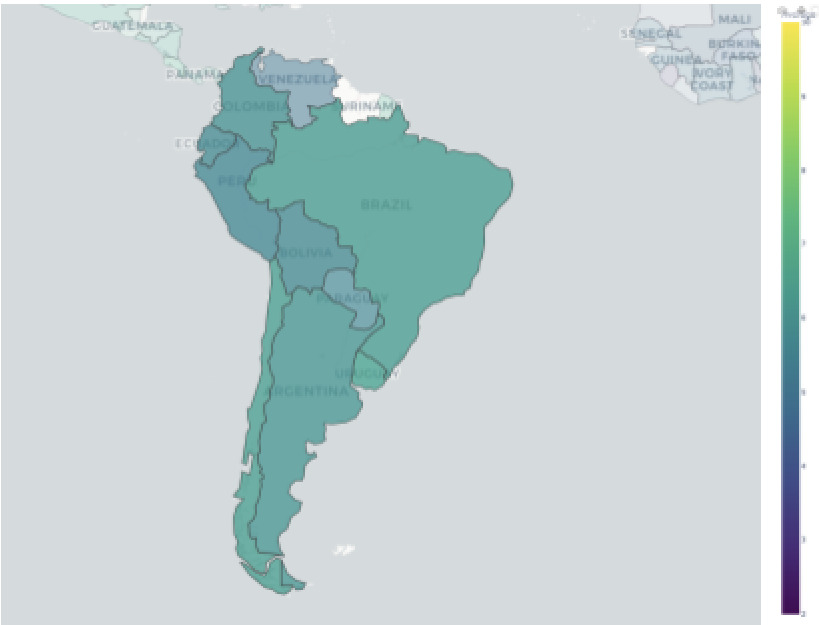


Figure 9: An interactive map showing the average happiness score for South American countries

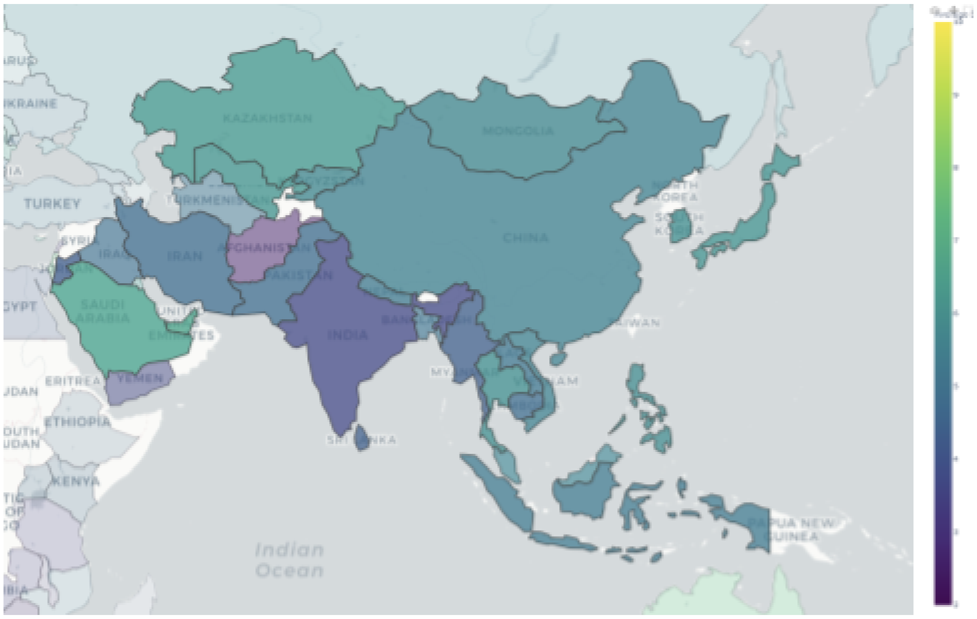


Figure 10: An interactive map showing the average happiness score for Asian countries

According to the map, African countries ranked the lowest (See Figure 9). These countries are less developed due to their low economic and global status. Life expectancy is reduced due to poor access to clean water, the medical system, and adequate food. The discrimination, violence, and war in some of those countries contribute to their low level of social support and freedom. This results in a high rate of poverty and lack of access to education. These factors further hinder development and contribute to the cycle of poverty that is less likely to feel happy.

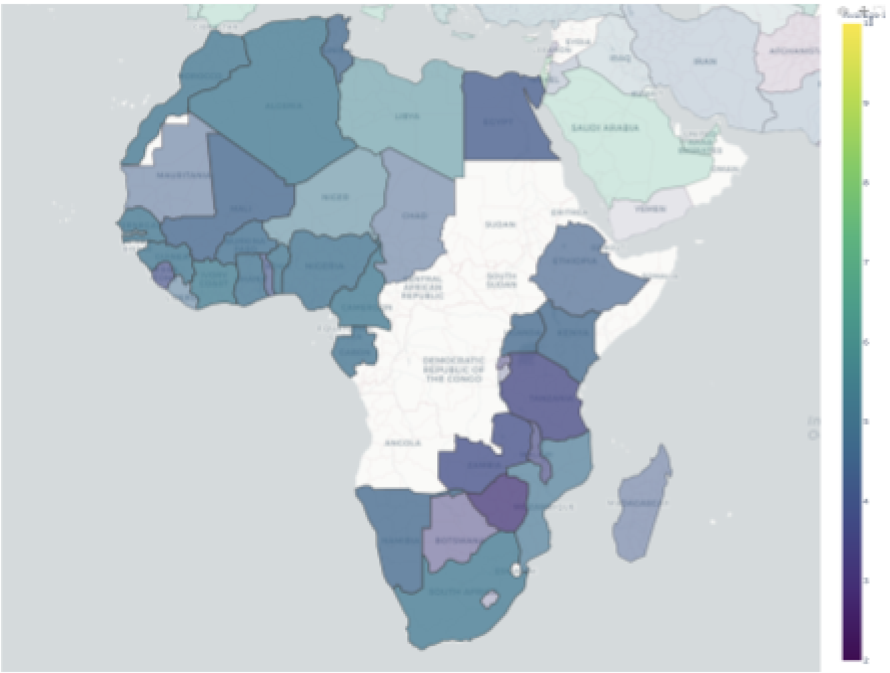


Figure 11: An interactive map showing the average happiness score for Afrian countries

It should be noted, however, that the four indexes that the happiness report evaluates are primarily based on western definitions of happiness. There may be bias in the results due to cultural and belief differences regarding happiness. Therefore, it is important to consider the report's findings with a critical eye and to recognize its limitations. It is also important to recognize that happiness is subjective and can vary from person to person and culture to culture.

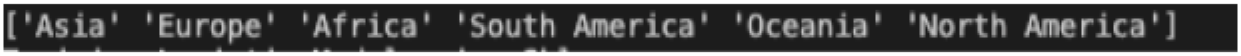
In summary, the interactive map provides an overview of the geographical distribution of the world happiness score. The continent-based understanding provides a sketch that can be used to calculate marginal effects as well as predict future scores using logistic and linear models.

Machine Learning

1. Logistic Regression Model

For the third question of predicting the happiness score using the social indexes (GDP, Social support, Health, and Freedom. For logistic regression, the outcome should be binary (Yes/No, True/False), so we use the mean() method to get the average happiness score for the whole world, and use the logistic regression model to check whether the happiness score is above the average score or lower than the average score.

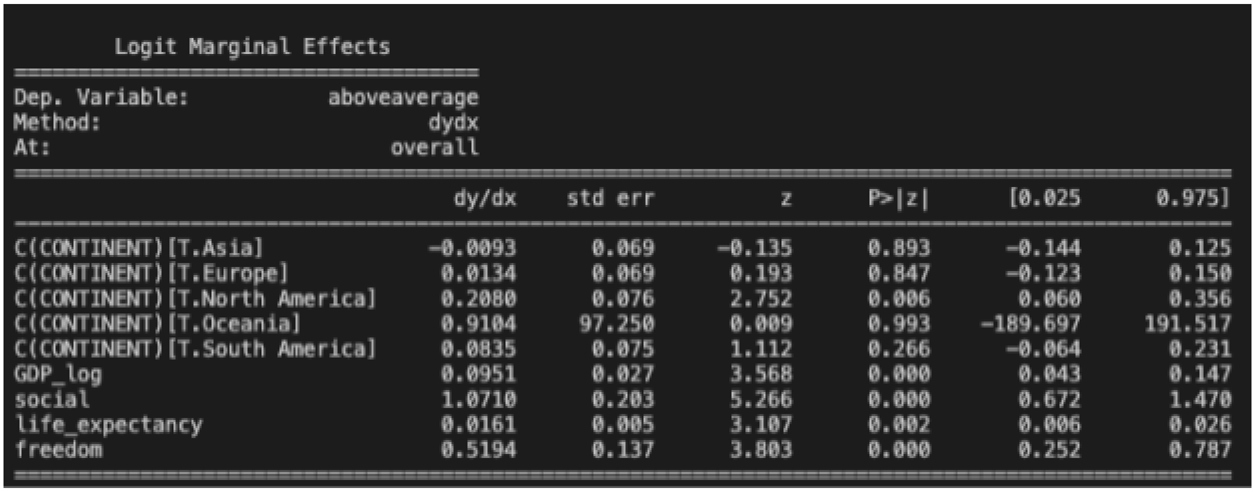
We first choose to use smf.logit() to construct the logic regression because we choose to use the marginal effect to analyze the different indices in our regression model.We also add the column “CONTINENT” inside the regression model and make it into an categorical variable. We used get_dummies() to change it into the categorical variable. The column of “CONTINENT” has 6 unique variable which are



```
['Asia' 'Europe' 'Africa' 'South America' 'Oceania' 'North America']
```

Figure 12: The unique variable in the column “CONTINENT”

Based on the Summery() of Marginal Effect we can see that the reference category is Africa.



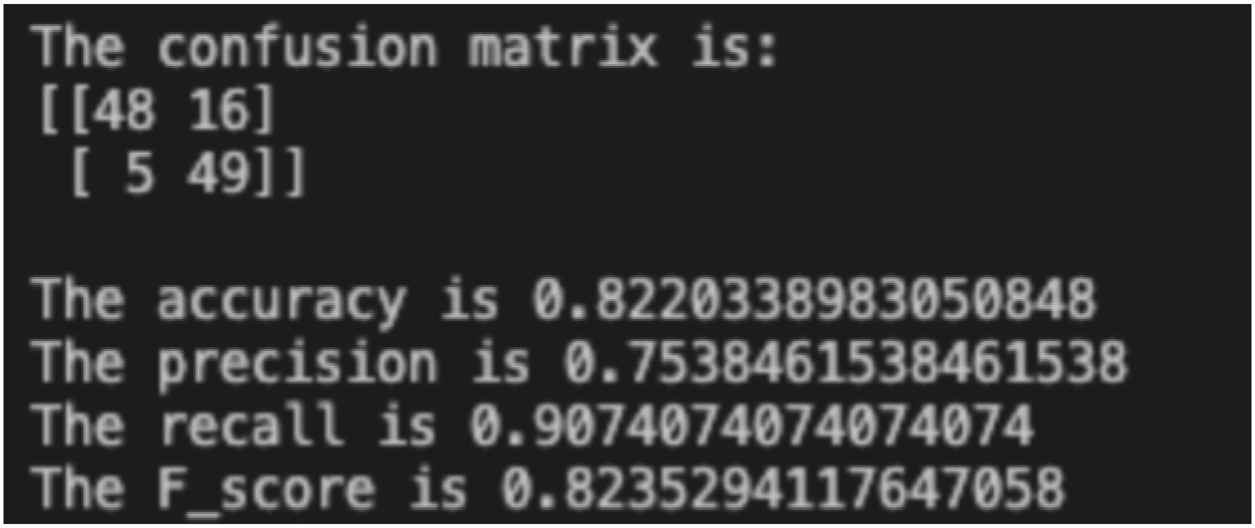
```
        Logit Marginal Effects
=======================================
Dep. Variable:           aboveaverage
Method:                          dydx
At:                           overall
==================================================================================================
                                   dy/dx    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
C(CONTINENT)[T.Asia]             -0.0093      0.069     -0.135      0.893      -0.144       0.125
C(CONTINENT)[T.Europe]            0.0134      0.069      0.193      0.847      -0.123       0.150
C(CONTINENT)[T.North America]     0.2080      0.076      2.752      0.006       0.060       0.356
C(CONTINENT)[T.Oceania]           0.9104     97.250      0.009      0.993    -189.697     191.517
C(CONTINENT)[T.South America]     0.0835      0.075      1.112      0.266      -0.064       0.231
GDP_log                           0.0951      0.027      3.568      0.000       0.043       0.147
social                            1.0710      0.203      5.266      0.000       0.672       1.470
life_expectancy                   0.0161      0.005      3.107      0.002       0.006       0.026
freedom                           0.5194      0.137      3.803      0.000       0.252       0.787
==================================================================================================
```

Figure 13: The logistic regression Marginal Effect

From the summary provided above, we can see the variable of "C(CONTINENT)[T.North America] ", "GDP_log", "social", ""life_expectancy", and "freedom"are statistically significant as the p value is less than 0.05.

- The result shows there is a -0.93% probability more likely for Asia to have a happiness score above average level for the whole world compared to Africa.
- The result shows there is a 1.34% probability more likely for Europe to have a happiness score above average level for the whole world compared to Africa.
- The result shows there is a 20.890% probability more likely for North America to have a happiness score above average level for the whole world compared to Africa.
- The result shows there is a 91.04% probability more likely for Oceania to have a happiness score above average level for the whole world compared to Africa.
- The result shows there is a 8.35% probability more likely for South America to have a happiness score above average level for the whole world compared to Africa.
- The result shows there is a 9.51% probability more likely to have a happiness score above average for the whole world when the GDP increased by 1 unit.
- The result shows there is a 107.10% probability more likely to have a happiness score above average for the whole world when the social index increased by 1 unit.
- The result shows there is a 1.61% probability more likely to have a happiness score above average for the whole world when the life expectancy (Health index) increased by 1 unit.
- The result shows there is a 51.94% probability more likely to have a happiness score above average for the whole world when the freedom index increased by 1 unit.

We then built the same logistic regression model with the Scikit-learn library by using the same variables. We also divided it into a training set and a test set to calculate the prediction of the confusion matrix, precision, accuracy, F-score, and the MSE of this model.

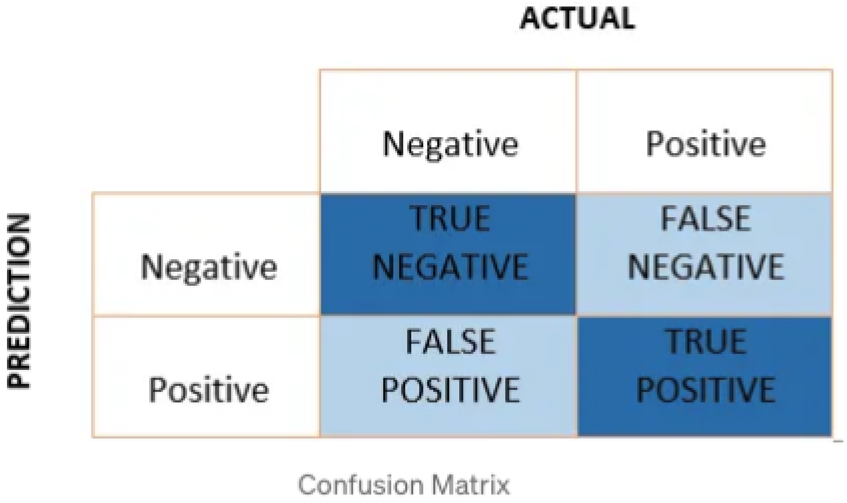


```
The confusion matrix is:
[[48 16]
 [ 5 49]]

The accuracy is 0.8220338983050848
The precision is 0.7538461538461538
The recall is 0.9074074074074074
The F_score is 0.8235294117647058
```

Figure 14: Results for logistic regression model

Confusion matrix is a matrix of size 2×2 for binary classification with actual values on one axis and predicted on another.



		ACTUAL	
		Negative	Positive
PREDICTION	Negative	TRUE NEGATIVE	FALSE NEGATIVE
	Positive	FALSE POSITIVE	TRUE POSITIVE

Confusion Matrix

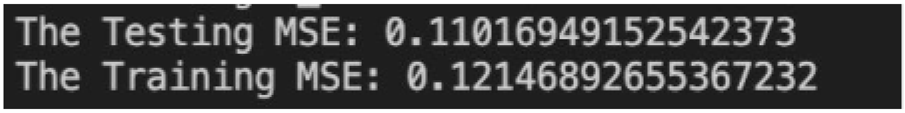
Figure 15: Example of confusion matrix

Accuracy (ACC) is calculated as the number of all correct predictions divided by the total number of the dataset. The best accuracy is 1.0, whereas the worst is 0.0. Based on the result we can see that the Accuracy for the confusion matrix is 0.82203 which means we have around82% of all values predicted correctly.

Precision is what percentage is truly positive, out of all the positives predicted. The precision value lies between 0 and 1. Based on the result above we can ensure that the precision for this model is 0.75384, which means 75% of all positive values are correctly labeled.

Recall tells us about how well the model identifies true positives. Based on the result above we can see that the recall for the model is 0.90740. Which means 90% of results are predicted correctly.

F-score is the harmonic mean(average) of the precision and recall. It will be high only when both accuracy and recall are high. It indicates the balance between precision & recall in the system. In the result above we can see that the F-score is 0.82352, which means 82.35% of data perform well on an imbalanced dataset.



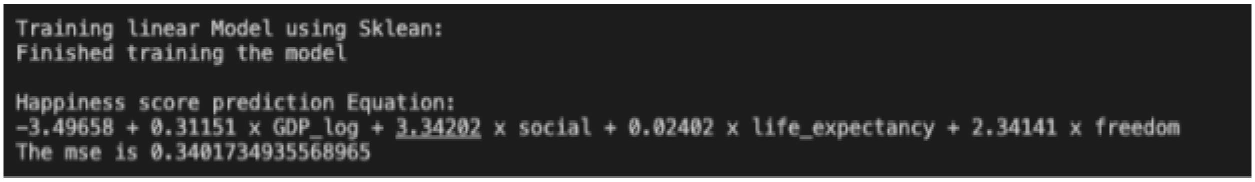
```
The Testing MSE: 0.11016949152542373
The Training MSE: 0.12146892655367232
```

Figure 16: The MSE for Logistic regression for Training Set and Test Set

MSE is calculated by taking the average of the square of the difference between the original and predicted values of the data. Since we split the data into a training set and a test set, we have two MSE for each model. The MSE for Testing set is 0.11016, which means there are 11.016% of data that are not predicted correctly from the original data. The MSE for Training set is 0.12146, which means there are 12.146% of data that are not predicted correctly from the original data.

2. **Linear Regression Model**

On the basis of the logistic regression model, we also used the Scikit-learn library to perform a multilinear regression model. We also divided the model into a training set and a test set. This allowed us to obtain the coefficients and intercept for each variable. Additionally, we were able to calculate the mean squared error (MSE) to assess the predicted accuracy of the model.



```
Training linear Model using Sklean:
Finished training the model

Happiness score prediction Equation:
-3.49658 + 0.31151 x GDP_log + 3.34202 x social + 0.02402 x life_expectancy + 2.34141 x freedom
The mse is 0.3401734935568965
```

Figure 17: The result of linear regression model and MSE for the model

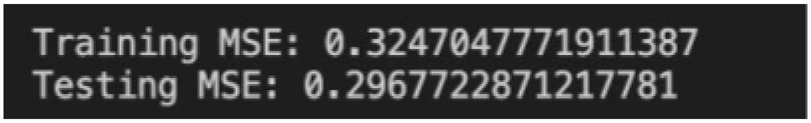
The intercept of -3.49658 means when GDP_log is 0, social index is 0, the life expectancy is 9, and freedom is 0, the happiness score would be -3.49658.

The index of GDP_log is 0.31151 which means those with one unit increase in GDP_log will let the happiness score increase 0.31151.

The index of social is 3.34202 which means those with one unit increase in social index will let the happiness score increase3.34202.

The index of life_expectancy is 0.02402 means those with one unit increase in life expectancy will let the happiness score increase 0.02402.

The index of freedom is 2.34141 means those with one unit increase in freedom will let the happiness score increase 2.34141.



```
Training MSE: 0.3247047771911387
Testing MSE: 0.2967722871217781
```

Figure 18: Linear regression model MSE for Training set and Testing Set

MSE is calculated by taking the average of the square of the difference between the original and predicted values of the data.

Since we split the data into training sets and the test set, Hence, the MSE for the Training model is 0.32470, which means there are 32.47% of data that are not predicted correctly from the original data. The MSE for the Testing model is 0.29677, which means there are 29.677% of data that are not predicted correctly from the original data.

Impacts and Limitations

The logistic regression model and the subsequent multilinear regression model provide valuable insights into the variables that have a significant impact on happiness scores. The results show that the variables of GDP, social index, life expectancy, and freedom have a positive impact on happiness scores. Specifically, an increase in GDP, social index, life expectancy, and freedom is likely to increase the probability of having a happiness score above the average level for the whole world. Additionally, the model shows that the continents of North America and Oceania have the highest probability of having a happiness score above the average level, whereas Africa has the lowest probability. These results have important implications for policymakers and stakeholders in various fields, as they suggest that investing in economic development, social welfare, health care, and human rights may improve happiness levels.

Furthermore, the accuracy, precision, recall, F-score, and MSE of the logistic regression model and the multilinear regression model provide a comprehensive evaluation of the models' predictive power. The accuracy of 82.20%, precision of 75.38%, and recall of 90.74% indicate that the models are effective in predicting happiness scores based on the selected variables. However, the MSE of 0.17796 and 0.34017 in the logistic regression and multilinear regression models, respectively, suggests that the models have room for improvement in accurately predicting happiness scores. These results highlight the need for further research to refine the models and consider additional variables that may affect happiness levels. Overall, the models provide valuable insights and tools for understanding and predicting happiness scores, which can inform policies and decision-making in various fields.

Based on the column ‘CONTINENT’, we only used data from Asia, Africa, Europe, North America, South America, and Oceania. However, we did not include data from other continents like Australia and Antarctica in our model, so our model cannot be applied to a broader range of regions. People from other regions (Asia, Africa, Europe, North America, South America, and Oceania) cannot use our results because our predictions are not based on the situation in that region.

Challenge Goals

Machine Learning

This project was intended to predict Happiness scores based on four social indexes mentioned in the report and evaluate those models. Testing data will be separated from training data and used for training the model, while testing data will be used for validation. We used a multi-linear regression model and two logistic regression models that were constructed via smf and Scikit-learn.

The linear regression model provides us with an equation that could be used to predict future happiness scores. Additionally, the mean squared error is provided as a means of evaluating the accuracy of the model.

Scikit-learn-based logistic regression models can also be used to train the data and make predictions. This model will be evaluated using a confusion matrix that will assist in calculating accuracy, precision, recall, and F1 score.

Marginal effects measures provide a general understanding of how different social indices can affect happiness scores, and identify which ones have the greatest impact. This can then be used to develop policies and social interventions to increase the overall happiness of the country.

New Library

In order to go beyond what we learned in class and make a better representation of the map. We used the new library `plotly` for Advanced/Interactive visualizations. `plotly` is not only capable of coloring the country according to its happiness score, but it also enables users to zoom in and out to customize the country boundaries. Furthermore, users are able to hover over the map and see the corresponding name and value calculated for the region. With 'plotly', we are able to create more visually appealing and interactive maps which can be used to effectively convey information and highlight important points. This allows users to better understand the data and explore the data in more detail.

In order to better understand the dataset, we used a new statistical library called statsmodels (smf) for the machine learning component of the project. For the analysis, we used the smf's logistic regression model and calculated marginal effects from data. By doing this, we are able to understand how different data changes impact the score to a greater or lesser degree, and not only make predictions. We were also able to identify the main drivers of the score and build a model to explain the relationship between each of these drivers and the score.

Work Plan Evaluation

1. **Loading and Combing data Expected: 3 Hours/Actual: 4 Hours**
   - Expected: Create a new dataframe that stores all dataframes together. Cleaning, converting, and storing the data in a format that will make it easier to use in the future. There is a formatting issue with the data for 2022 that should be corrected in all columns and rows. It should be converted into the same format as other dataframes.
   - Actual: It took approximately four hours to complete this part since we were changing our dataset once to better match our challenge goal. Finding the right key to reduce `NaN` values for joining the geodataframe with the happiness data takes some time.
2. **Plotting Expected: 2 Hours/Actual: 2.5 Hours**
   - Expected: Filter and select the Happiness score and one of the social indexes to be plotted. Plot a scatter plot based on the data in selected two columns, showing us the relationship between the Happiness score and the social index. Do statistical analysis based on those graphs we draw above.
   - Actual: It took our team approximately 2.5 hours to complete this task. The decision regarding whether to plot those scatter plots into a single figure takes some time. Moreover, we have added another histogram here to analyze the overall pattern of the happiness score. Also, we moved the analysis part into the end.
3. **Geospatial Plot Expected: 3 Hours/Actual: 5 Hours**
   - Expected: Merge a geospatial dataset that contains a geometry column for countries in the world with a Happiness score dataset. Background: Plot a map of the world using color = #EEEEEE. Then, overlay the color of each country according to the Happiness score.Repeat the above steps for each year.
   - Actual: Since our dataset was altered once and while processing, we noticed that the score maps for each year are quite similar to each other since countries are stable in their development and there are no major world events

that affect the score. Due to this, we have decided to use average scores for each country between 2018 and 2021 and upgrade the visualization to an interactive plot using `plotly`. After we have completed the coding phase, we moved to the analysis phase. Due to changes made to the geospatial plot, the process took approximately 5 hours.

4. **Machine Learning and Prediction Expected: 5 Hours/Actual: 7 Hours**
   - Expected: Separate the data into two categories: 20% for testing and 80% for training. The model will be trained using training data (OLS is expected, but alternative models may be tested to find the most accurate model by comparing the accuracy score). The model is used to predict the training data label and calculate its accuracy score. Analyze statistically the coefficients of different social indices.
   - Actual: As a result, this part takes approximately seven hours to complete. Although we expected to use OLS to generate a multilinear regression model, it turned out to be too simple to implement. Consequently, we decided to keep the model, but also generate two logistic regression models for the purpose of 1. Calculation of accuracy using a confusion matrix 2. Calculation of marginal effects. These two logistic regressions were generated using two different libraries, which resulted in varying outcomes. Adding a new model as well as learning how the model works takes approximately two hours additional.
5. **Report Writing Expected: 2 Hours/Actual: 4 Hours**
   - Expected: The analysis should be conducted as necessary, with the objective of identifying how different social factors affect the Happiness score and determining an equation that predicts Happiness over time.
   - Actual: Some parts of the written part appear to take longer than expected. The results part of the report requires extensive analysis in order to understand the data we collected. This part of the procedure is also longer because we put all analyses at the end. In total, 4 hours were spent writing the report.

Testing

We will use the `one.csv` and `two.csv` dataset for testing. All the testing codes can be found in the `testing.py` which contains three different testing functions for checking the correctness of the certain functions implemented in “deliverable.py”. Please note that the function is one-hot encoding, we assume the used data is using the same column names to correctly perform the testing.

The first function `test_clean_data` uses assert_equals function to test the shape of the cleaned data, the data tested are `one.csv` and `two.csv` and were imported in the main method that can be found at the end of the testing file. If either of these assertions fails, the function will raise an AssertionError indicating that the test has failed.

The second function, `test_join_data` uses assert_equals function and mainly tests the shape of two GeoDataFrames that join the score dataset with `world.shp`, named one_gpd and two_gpd. If either of these assertions fails, the function will raise an AssertionError indicating that the test has failed.

The last function, `test_split_data` also uses assert_equals function that verifies if the `split_data` function correctly splits two GeoDataFrames `one_gpd` and `two_gpd` into training and testing sets for logistic and linear regression models. This testing function mainly checks for the features and labels variables we created in `deliverable.py` and also tests on their number to see if it splits correctly. The first part of the function tests `one_gpd`, it uses assert_equal function to verify that the logistic features have 9 columns, the logistic labels are equal to the `above` column of `one_gpd`, the linear features have 4 columns, and the linear labels are equal to the 'score' column of `one_gpd`. Then, it splits the linear features and labels into training and testing sets using train_test_split and checks that the number of samples in the training and testing sets are as expected. Then, it performs similar tests with `two_gpd`. If either of these assertions fails, the function will raise an AssertionError indicating that the test has failed.

Collaboration

New Library

We used `plotly` and `statsmodels` for our project, documentations help a lot.

Help with debugging

During the coding process, we have faced lots of errors and the following sites help us to deal with those issues: plotly crs issue, linear regression summary issue.

Acknowledgement

Sincerely thanks to our TA Vatsal Chandel for his guidance, assistance, and invaluable feedback throughout the project. His availability and encouragement supports us to achieve our goal.

No other people other than course staff helped this project.

Appendix

Happiness Dataset Information

- **shape:** `(2089, 12)`
- **Column names:**

  ```
  Index(['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',
         'Confidence in national government'],
        dtype='object')
  ```

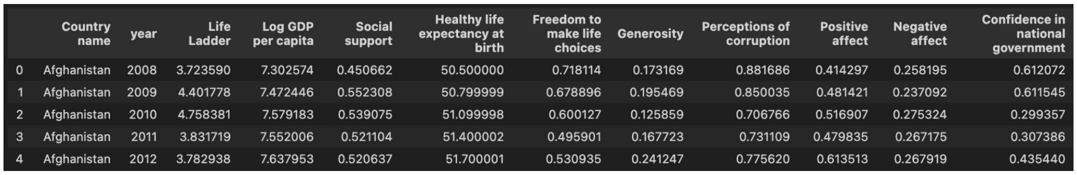
- **NaN values:**

  ```
  Country name                          0
  year                                  0
  Life Ladder                           0
  Log GDP per capita                   27
  Social support                       13
  Healthy life expectancy at birth     58
  Freedom to make life choices         32
  Generosity                           80
  Perceptions of corruption           113
  Positive affect                      24
  Negative affect                      16
  Confidence in national government   216
  dtype: int64
  ```

- **Types:**

  ```
  Country name                          object
  year                                   int64
  Life Ladder                          float64
  Log GDP per capita                   float64
  Social support                       float64
  Healthy life expectancy at birth     float64
  Freedom to make life choices         float64
  Generosity                           float64
  Perceptions of corruption            float64
  Positive affect                      float64
  Negative affect                      float64
  Confidence in national government    float64
  dtype: object
  ```

- **DataFrame.head(5)**



	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	Confidence in national government
0	Afghanistan	2008	3.723590	7.302574	0.450662	50.500000	0.718114	0.173169	0.881686	0.414297	0.258195	0.612072
1	Afghanistan	2009	4.401778	7.472446	0.552308	50.799999	0.678896	0.195469	0.850035	0.481421	0.237092	0.611545
2	Afghanistan	2010	4.758381	7.579183	0.539075	51.099998	0.600127	0.125859	0.706766	0.516907	0.275324	0.299357
3	Afghanistan	2011	3.831719	7.552006	0.521104	51.400002	0.495901	0.167723	0.731109	0.479835	0.267175	0.307386
4	Afghanistan	2012	3.782938	7.637953	0.520637	51.700001	0.530935	0.241247	0.775620	0.613513	0.267919	0.435440

World Dataset Information

- **shape:** (177, 169)
- **Column names:**

```
Index(['featurecla', 'scalerank', 'LABELRANK', 'SOVEREIGNT', 'SOV_A3',
       'ADM0_DIF', 'LEVEL', 'TYPE', 'TLC', 'ADMIN', 'ADM0_A3', 'GEOU_DIF',
       'GEOUNIT', 'GU_A3', 'SU_DIF', 'SUBUNIT', 'SU_A3', 'BRK_DIFF',
       'NAME', 'NAME_LONG', 'BRK_A3', 'BRK_NAME', 'BRK_GROUP', 'ABBREV',
       'POSTAL', 'FORMAL_EN', 'FORMAL_FR', 'NAME_CIAWF', 'NOTE_ADM0',
       'NOTE_BRK', 'NAME_SORT', 'NAME_ALT', 'MAPCOLOR7', 'MAPCOLOR8',
       'MAPCOLOR9', 'MAPCOLOR13', 'POP_EST', 'POP_RANK', 'POP_YEAR',
       'GDP_MD', 'GDP_YEAR', 'ECONOMY', 'INCOME_GRP', 'FIPS_10', 'ISO_A2',
       'ISO_A2_EH', 'ISO_A3', 'ISO_A3_EH', 'ISO_N3', 'ISO_N3_EH', 'UN_A3',
       'WB_A2', 'WB_A3', 'WOE_ID', 'WOE_ID_EH', 'WOE_NOTE', 'ADM0_ISO',
       'ADM0_DIFF', 'ADM0_TLC', 'ADM0_A3_US', 'ADM0_A3_FR', 'ADM0_A3_RU',
       'ADM0_A3_ES', 'ADM0_A3_CN', 'ADM0_A3_TW', 'ADM0_A3_IN',
       'ADM0_A3_NP', 'ADM0_A3_PK', 'ADM0_A3_DE', 'ADM0_A3_GB',
       'ADM0_A3_BR', 'ADM0_A3_IL', 'ADM0_A3_PS', 'ADM0_A3_SA',
       'ADM0_A3_EG', 'ADM0_A3_MA', 'ADM0_A3_PT', 'ADM0_A3_AR',
       'ADM0_A3_JP', 'ADM0_A3_KO', 'ADM0_A3_VN', 'ADM0_A3_TR',
       'ADM0_A3_ID', 'ADM0_A3_PL', 'ADM0_A3_GR', 'ADM0_A3_IT',
       'ADM0_A3_NL', 'ADM0_A3_SE', 'ADM0_A3_BD', 'ADM0_A3_UA',
       'ADM0_A3_UN', 'ADM0_A3_WB', 'CONTINENT', 'REGION_UN', 'SUBREGION',
       'REGION_WB', 'NAME_LEN', 'LONG_LEN', 'ABBREV_LEN', 'TINY',
       'HOMEPART', 'MIN_ZOOM', 'MIN_LABEL', 'MAX_LABEL', 'LABEL_X',
       'LABEL_Y', 'NE_ID', 'WIKIDATAID', 'NAME_AR', 'NAME_BN', 'NAME_DE',
       'NAME_EN', 'NAME_ES', 'NAME_FA', 'NAME_FR', 'NAME_EL', 'NAME_HE',
       'NAME_HI', 'NAME_HU', 'NAME_ID', 'NAME_IT', 'NAME_JA', 'NAME_KO',
       'NAME_NL', 'NAME_PL', 'NAME_PT', 'NAME_RU', 'NAME_SV', 'NAME_TR',
       'NAME_UK', 'NAME_UR', 'NAME_VI', 'NAME_ZH', 'NAME_ZHT',
       'FCLASS_ISO', 'TLC_DIFF', 'FCLASS_TLC', 'FCLASS_US', 'FCLASS_FR',
       'FCLASS_RU', 'FCLASS_ES', 'FCLASS_CN', 'FCLASS_TW', 'FCLASS_IN',
       'FCLASS_NP', 'FCLASS_PK', 'FCLASS_DE', 'FCLASS_GB', 'FCLASS_BR',
       'FCLASS_IL', 'FCLASS_PS', 'FCLASS_SA', 'FCLASS_EG', 'FCLASS_MA',
       'FCLASS_PT', 'FCLASS_AR', 'FCLASS_JP', 'FCLASS_KO', 'FCLASS_VN',
       'FCLASS_TR', 'FCLASS_ID', 'FCLASS_PL', 'FCLASS_GR', 'FCLASS_IT',
       'FCLASS_NL', 'FCLASS_SE', 'FCLASS_BD', 'FCLASS_UA', 'geometry'],
      dtype=object, length=169)
```

- **NaN Values (only extract columns have NaN values):**

```
TLC             1                            FCLASS_IN     173
BRK_GROUP     177                            FCLASS_NP     173
FORMAL_EN       3                            FCLASS_PK     172
FORMAL_FR     172                            FCLASS_DE     174
NAME_CIAWF      5                            FCLASS_GB     174
NOTE_ADM0     169                            FCLASS_BR     174
NOTE_BRK      170                            FCLASS_IL     174
NAME_ALT      173                            FCLASS_PS     172
ADM0_DIFF     173                            FCLASS_SA     171
TLC_DIFF      173                            FCLASS_EG     173
FCLASS_US     174                            FCLASS_MA     172
FCLASS_FR     173                            FCLASS_PT     174
FCLASS_RU     174                            FCLASS_AR     173
FCLASS_ES     174                            FCLASS_JP     174
FCLASS_CN     173                            FCLASS_KO     174
FCLASS_TW     172                            FCLASS_VN     174
```

```
FCLASS_TR     173
FCLASS_ID     172
FCLASS_PL     173
FCLASS_GR     174
FCLASS_IT     174
FCLASS_NL     173
FCLASS_SE     174
FCLASS_BD     171
FCLASS_UA     174
dtype: int64
```

- **Types (Only extract useful columns' type)**

```
SUBUNIT         object
NAME            object
POP_EST        float64
POP_RANK         int64
POP_YEAR         int64
GDP_MD           int64
GDP_YEAR         int64
ECONOMY         object
CONTINENT       object
REGION_UN       object
SUBREGION       object
REGION_WB       object
NAME_LEN         int64
LONG_LEN         int64
ABBREV_LEN       int64
TINY             int64
NAME_EN         object
geometry      geometry
dtype: object
```

- **geometry.sample(2)**

```
POLYGON ((105.21878 14.27321, 104.28142 14.41674, 102.98842 14.22572, 102.34810
13.39425, 102.58493 12.18659, 101.68716 12.64574, 100.83181 12.62708, 100.97847
13.41272, 100.09780 13.40686, 100.01873 12.30700, 99.47892 10.84637, 99.15377
9.96306, 99.22240 9.23926, 99.87383 9.20786, 100.27965 8.29515, 100.45927
7.42957, 101.01733 6.85687, 101.62308 6.74062, 102.14119 6.22164, 101.81428
5.81081, 101.15422 5.69138, 101.07552 6.20487, 100.25960 6.64282, 100.08576
6.46449, 99.69069 6.84821, 99.51964 7.34345, 98.98825 7.90799, 98.50379
8.38231, 98.33966 7.79451, 98.15001 8.35001, 98.25915 8.97392, 98.55355
9.93296, 99.03812 10.96055, 99.58729 11.89276, 99.19635 12.80475, 99.21201
13.26929, 99.09776 13.82750, 98.43082 14.62203, 98.19207 15.12370, 98.53738
15.30850, 98.90335 16.17782, 98.49376 16.83784, 97.85912 17.56795, 97.37590
18.44544, 97.79778 18.62708, 98.25372 19.70820, 98.95968 19.75298, 99.54331
20.18660, 100.11599 20.41785, 100.54888 20.10924, 100.60629 19.50834, 101.28201
19.46258, 101.03593 18.40893, 101.05955 17.51250, 102.11359 18.10910, 102.41300
17.93278, 102.99871 17.96169, 103.20019 18.30963, 103.95648 18.24095, 104.71695
17.42886, 104.77932 16.44186, 105.58904 15.57032, 105.54434 14.72393, 105.21878
14.27321))

POLYGON ((-77.56960 18.49053, -76.89662 18.40087, -76.36536 18.16070, -76.19966
17.88687, -76.90256 17.86824, -77.20634 17.70112, -77.76602 17.86160, -78.33772
18.22597, -78.21773 18.45453, -77.79736 18.52422, -77.56960 18.49053))
```