**Jaxon Bundy & Musaab Bargicho**

**Team Project #1 – Discover New Data and Analyze the Results**

**(100 points)**

**Please note that this team assignment must be completed and submitted electronically on Canvas by 11:59pm of the due date. Create a separate Word document to summarize your findings. If you submit your team’s Excel file, make sure to include on your comments on the Word document. Please submit only ONE Word document per team.**

Carefully analyze the data and variables in your dataset, as you’ll be asked to organize and visualize some of the variables. Note that this is an open-ended set of questions. To make this project engaging to you and your team, you will be determining the topic of interest, the data set and the variables you will be using.

You and your team will need to perform appropriate analysis in Excel and write a short report summarizing your findings for the sample data and supporting your conclusions with the tables and charts performed in Excel. When writing a report, copy all supporting tables and graphs, when applicable, from Excel or StatCrunch to a .doc file. Pay attention to the formatting. Please note that with the online submission, the final report should be submitted in either .doc or .pdf format. You may submit both Doc and Excel files, however only the Word document will be graded. I may refer to your Excel file if I have questions on how certain charts/tables were created. Please do NOT include the original data in your Word report. **Failure to abide by these rules may result in up to a 10% grade reduction at the instructor’s discretion**. Your report should address the following:

**1)** (6 pts) Clarify your problem statement, that is, explain the topic of interest your team is investigating.

**We are analyzing the happiness scores to find out what factors contribute to the happiness scores of people around the world. The dataset contains several variables such as the GDP, life expectancy, freedom, generosity, and dystopia residuals with the happiness score from each country. We want to investigate the relationship and impact these variables have on the happiness score.**

**2)** (4 pts) Classify all variables as qualitative or quantitative. For qualitative variables, define the scale (nominal or ordinal). Present the results in a tabular format if preferred.

| **Data Name** | **Type** |
| --- | --- |
| Happiness Score | Qualitative (Ordinal) |
| Freedom | Qualitative (Ordinal) |
| Trust(Government Corruption) | Qualitative(Ordinal) |
| Health(Life Expectancy) | Quantitative |
| Generosity | Qualitative(Ordinal) |
| Dystopia Residual | Quantitative |
| Economy (GDP per capita) | Quantitative |

**3)** (8 pts) Use appropriate descriptive statistics to summarize six variables of interest. Comment on your findings by including a short paragraph.

**Happiness Score: This is a scale from 0 to 10 on how people evaluate their own lives. This is measured by a survey where people are asked to rate their happiness and well beings.**

**Freedom: A scale that represents the ability to make life choices freely. Averages the response to the question “Are you satisfied or dissatisfied with your freedom to choose what you do with your life?” from GWP.**

**Trust (Government Corruption): This measures the national average of the response to the questions “Is corruption widespread throughout the government or not” and “Is corruption widespread within businesses or not?”, both of which require only a 1 (not corrupt or 0 (corrupt) for answers.**

**Health: This is the average number of years someone is expected to live based on various factors such as age, health, gender, etc. Life expectancy is calculated by analyzing mortality rates over a specific period of time and this determines the average number of years a person is expected to live based on these rates.**

**Generosity: The national average of the response to the question “Have you donated money to a charity in the past month?”.**

**Dystopia Residual: The fictional low benchmark used to describe the worst possible country in 6 major categories. This is used to make sure all countries are represented on the positive end of the scale.**

**4)** **a)** (6 pts) Construct a frequency and relative frequency distributions for the three quantitative variables. (Hint: Create the classes using a meaningful width, you may choose to deviate from the 2k rule).

**ECONOMY (GDP)**

| **Range** | **Frequency** | **Relative Frequency** |
| --- | --- | --- |
| **0 - 0.5** | 28 | 18% |
| **0.5 - 1.0** | 48 | 31% |
| **1.0 - 1.5** | 72 | 46% |
| **>1.5** | 9 | 6% |
| **Total** | 157 | 100% |

**Health (Life Expectancy)**

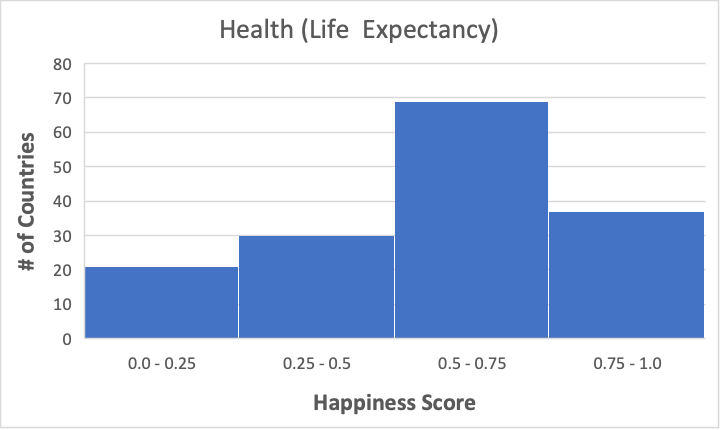
| **Range** | **Frequency** | **Relative Frequency** |
| --- | --- | --- |
| **0.0 - 0.25** | 21 | 13% |
| **0.25 - 0.5** | 30 | 19% |
| **0.5 - 0.75** | 69 | 44% |
| **0.75 - 1.0** | 37 | 24% |
| **Total** | 157 | 100% |

**Dystopia Residual**

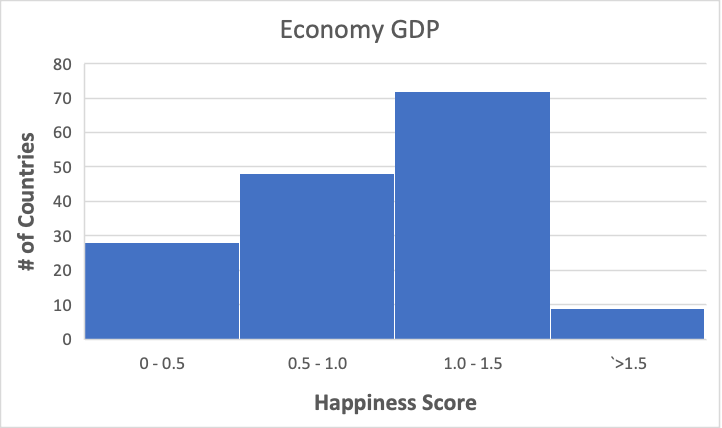
| **Range** | **Frequency** | **Relative Frequency** |
| --- | --- | --- |
| **0.0 - 1.0** | 5 | 3% |
| **1.0 - 2.0** | 34 | 22% |
| **2.0 - 3.0** | 101 | 64% |
| **>3.0** | 17 | 11% |
| **Total** | 157 | 100% |

**b)** (6 pts) Construct histograms for the variables in part a). Comment on the shape of the distribution in each case. Be sure to properly label the axes and include the titles.

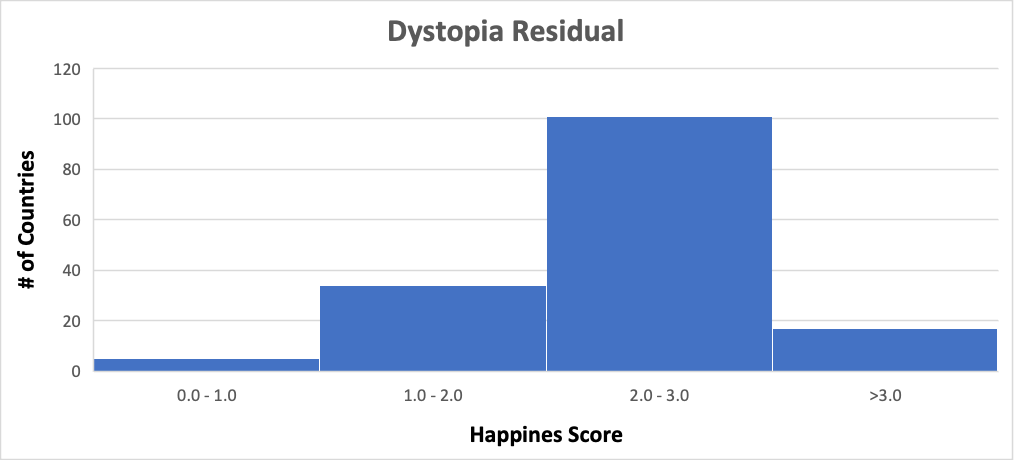
Health: This graph is not symmetric, and is left-skewed



Economy(GDP): This graph is not symmetric, and is left-skewed.



Dystopia Residual: This graph is not symmetric, and is left-skewed.

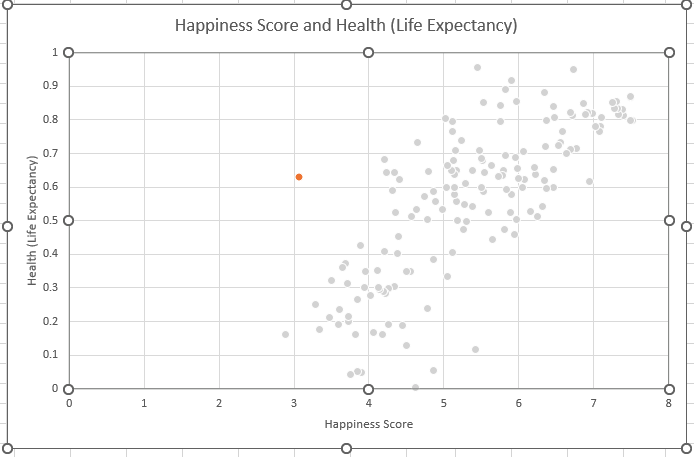


**5)** Examine the relationship between two variables of interest.  
 **a)** (4 pts) Prepare a contingency table for the two variables of interest. Please note that quantitative variables need to be grouped by classes. Comment on your findings from the table (up to 3 sentences).

| Happiness Score ->    Life Expectancy v | 0.0 - 5.0 | 5.0 - 10.0 |
| --- | --- | --- |
| 0.0 - 0.2 | 15 | 1 |
| 0.2 - 0.4 | 24 | 1 |
| 0.4 - 0.6 | 12 | 26 |
| 0.6 - 0.8 | 7 | 44 |
| 0.8 - 1.0 | 0 | 27 |

As Life Expectancy increases, the ratio of high happiness scores to low happiness scores increases. Lower health scores can be correlated with low happiness scores, and vice versa.

**b)** (4 pts) Make a scatter plot to explore the relationship between the two variables. Label both axes. What can you conclude?

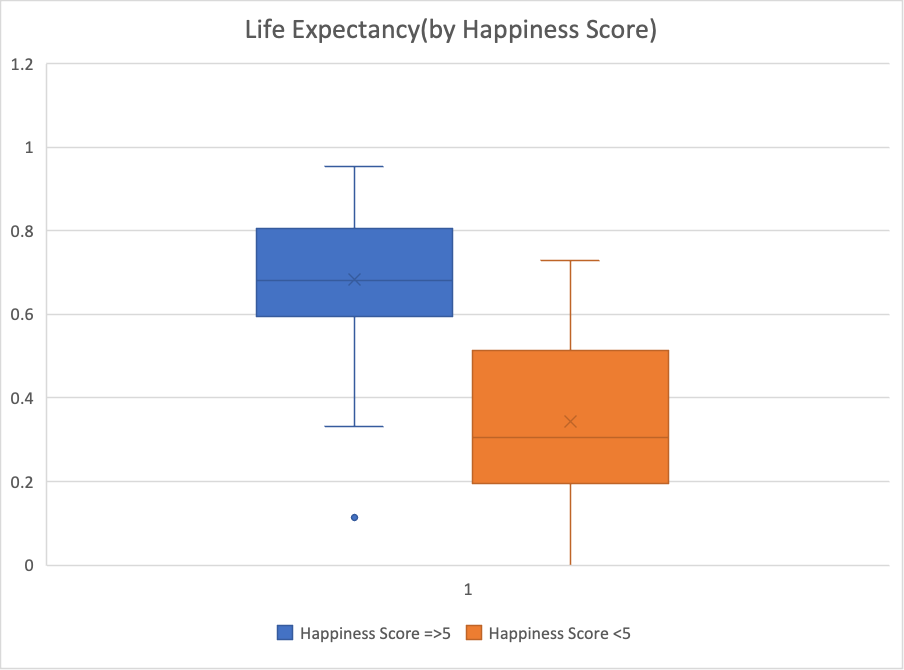


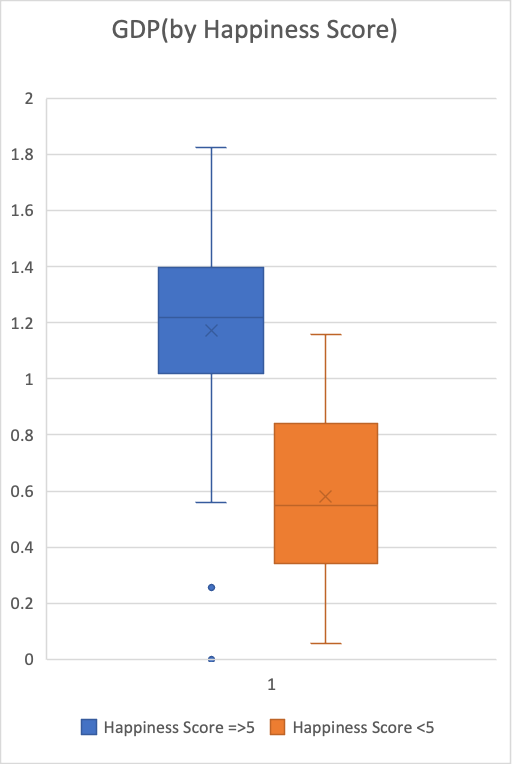
**As the happiness score increases the life expectancy increases.**

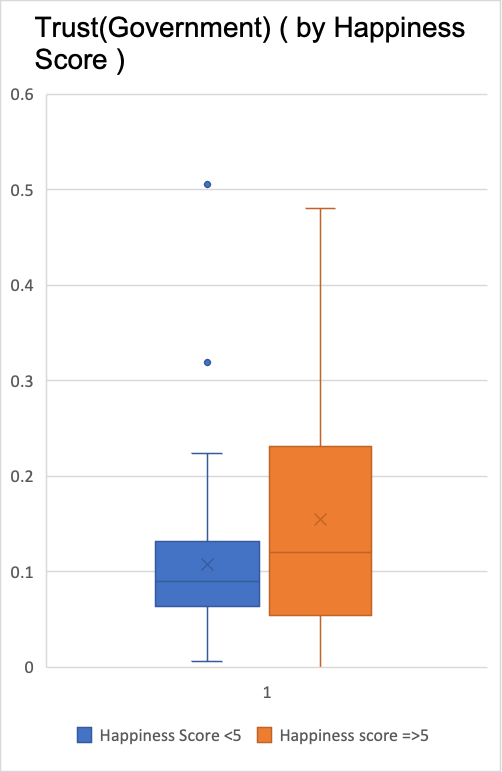
**c)** (4 pts) Compute and interpret the sample coefficient of correlation between the variables in the previous part.

**0.7654**

**This indicates a strong linear correlation between the two data sets, meaning a strong connection between happiness and life expectancy.**

**6)** (6 pts) Create boxplots for the three variables using a qualitative variable to split them (for instance, compare spending between male and female shoppers). Use the boxplots to compare the shape and spread for two distributions. Please note that for each variable the boxplots for the qualitative variable should be displayed on the same chart. For boxplots, you may either use Excel or Statcrunch.  






(2 pts) Are there any outliers? Comment based on the results in a).

**In the box plot of countries with a happiness score of less than 5 and their level of trust in the government, there were some that fell outside the range of the typical values, which are outliers. Similarly, in the box plot of countries with a happiness score greater than 5 and their GDP, there were some outliers with lower GDP values.**

**7)** (10 pts) Describe the type of sampling that was done to create your dataset. Justify your choice by both logical reasoning and by going back to the original documentation and citing it directly. This might require you to dig a few layers down (for example, datasets from Kaggle might have the original documentation, but some will only cite where the data originated from).

**The data was collected by surveying sample populations with a number of questions and averaged for each country. Roughly 1000 people were surveyed per country, with a few outliers in the sample size. Participants were asked to give answers in the form of binary (yes or no questions) as well as rankings (0-10). All of these answers were then averaged into scores, and ranked accordingly.**

**Choose 2 quantitative variables from your dataset for the following question:**

**8)** **a)** (2 pts) Calculate the mean for each variable.

**Economy(GDP): 0.9539**

**Health(Life Expectancy): 0.5576**

**b)** (2 pts) Calculate the critical value (aka critical score) for each variable.

**-1.96 & 1.96 for both**

**c)** (4 pts) Calculate a 95% confidence interval for the mean for each variable using the appropriate calculation.

**Economy(GDP): 0.889 & 1.018**

**Health(Life Expectancy): 0.522 & 0.593**

**d)** (4 pts) Summarize your findings from parts a and c in one sentence per variable.

**We can be 95% sure that the true population mean of the Economy scores is between 0.889 & 1.018. We can be sure that the population mean for health score is between 0.522 & 0.593.**  
 **e)** (4 pts) Does the Central Limit Theorem allow the analysis from parts a through d to work? Why or why not? (Spend at most 3 sentences to describe your answer)

**The Central Limit Theorem would allow the analysis to work, because the sample population size is greater than 30. CLM is used primarily for large sample sizes, which we are using. Our sample size of 157 countries is well above this threshold.**

**Choose 1 quantitative variable from your dataset for the following question:**

**9)** Imagine a recent article covering your variable of interest has claimed the population mean is 20% higher than your sample mean. For example, you are using voter age and find that your data shows the average age is 35. Now, the news article has claimed that average voter age is actually 42.  
 **a)** (2 pt) Report the claimed population mean.

**They are claiming that the Life expectancy average is .6691. Our data shows the average life expectancy is .5576**

**b)** (2 pt) State the null hypothesis (in plain English).

**The life expectancy mean is .6691**  
 **c)** (2 pt) State the alternative hypothesis (in plain English).

**The life expectancy mean is not .6691**

**d)** (4 pts) State both the null and alternative hypotheses in a mathematical formula.

**H0: μ = 0.6691**

**H1: μ ≠ 0.6691**

**e)** (2 pt) Pick an alpha level.

**0.10** **f)** (2 pts) Calculate and report the appropriate critical value (aka critical score).

**-1.644 & 1.644** **g)** (4 pts) Calculate and report the appropriate test statistic.

**-3.386** **h)** (4 pts) State your conclusion in “statistics speak”.

**The test statistic does not fall within the critical values so it rejects the null hypothesis.**

**i)** (4 pts) State your conclusion in plain English.

**If the test statistic does not fall within the critical values, it means that the evidence we have is not strong enough to support the alternative hypothesis**