COMPSCI 216 Project Report
Ben L. Goldfried, Sammy M. Korol, Andrew K. Dawson, Ezra F. Odio

Part 1: Introduction and Research Question

There has been increasing political polarization regarding government funding for the military. An argument from the right is that we need to heavily invest in the military for national security purposes. The left's counter argument is that governments should be spending more of the military budget on programs that more directly affect their citizens – such as aiding the homeless or increasing medicare. These programs directly affect the well-being of citizens, but also a country's national security also casts its own influence over a population's sentiments. Thus, we hypothesize that military spending should have an effect on the population's level of happiness. In our project, we are first going to investigate whether the amount of money a government spends on their military is correlated to the overall happiness of its citizens, and if that correlation is positive or negative.

Moreover, a population's happiness relies on numerous factors, such as perceptions of government corruption, widespread generosity, family, life expectancy, and more. Military spending has often been linked to government corruption throughout history and can positively or negatively change the population's trust in its government. Military spending also has an indirect impact on health programs, as explained above. Moreover, the sense of national security from military spending may prove correlated to the sentiments of freedom. Thus, military spending has possible correlations to these contributing facets of happiness: government trust, life expectancy, and freedom. Analyzing these correlations will reveal the impact military spending has on specific facets of a population's happiness.

Research Question: Does military spending as a percentage of GDP have a positive, negative, or insignificant correlation with a country's overall happiness? Does military spending correlate to more specific factors of happiness, such as freedom, health expectancy, and government trust? Are those factors of happiness strongly correlated with overall happiness or are they more inelastic?

This question is **relevant** because it examines the impact of a nation's military budget on citizens, which is pertinent to our topic of interest. The conclusions found can provide an important insight on the debate between the left and right wings. By finding the relationship between military spending and sentiments of freedom, life expectancy, and government trust, we are able to further develop the accuracy of each side's arguments, revealing the positive or negative impacts military spending has on these facets. This question provides an important insight on military spending's effect on a nation's people. An understanding of these effects can have a positive impact on broader society. We will also be able to analyze the results to see which factors of happiness do not fluctuate as much when overall happiness increases and which factors of happiness are more elastic and likely to be subject to influence.

Part 2: Data Sources

For our project, we will be looking at the following three datasets from Kaggle:
1: "Happiness and Corruption 2015-2020"
Type: csv, Kaggle Ownership: Elias Turk, License: CC0: Public Domain

This dataset provides information about the population's overall happiness by country by aggregating all valid data for countries with existent data from the year 2015 to 2020. The dataset provides specific columns to outline the factors that contributed to the overall happiness score. This dataset will be immensely useful in comparing the overall happiness and its contributing factors between countries. By analyzing this data, we can possibly find correlations in the happiness scores with increases in military spending.
2:"Military Spending of Countries (1960-2019)"
Type: csv, Kaggle Ownership: Nitin Singh, License: Creative Commons Attribution 4.0 International License [(CC BY 4.0)](CC BY 4.0)

Consolidated from The World Bank Database, this dataset provides the military expenditure (in USD) of over 200 countries from 1960-2019. We will be able to analyze this dataset alongside the "Happiness and Corruption 2015-2020" dataset, analyzing the countries with large or small military expenditures and merging the datasets to find any correlation to overall happiness.

3: "[World GDP (GDP, GDP per capita, and annual growths](#)"
Type: csv, Kaggle Ownership: Ozgur Cem Tas, License: CC0 – Public Domain

This dataset contains the overall GDP per country from 1960-2020. We will utilize this dataset along with the "Military Spending of Countries (1960-2019)" dataset to calculate the percentage of military expenditure relative to GDP. This calculation will allow us to analyze military expenditure relative to overall GDP to account for differences in the size and economy of countries. This dataset will be easier to work with when compared to the one marked in our proposal, as the years are the columns rather than the rows, like our military spending dataset.

Data Preparation/Wrangling (also explains how our results were generated for Part 4):

Data preparation was needed in order to put the data in a usable/consistent format. First, because the data sets included data over many years, it was necessary to remove entries from years that did not appear in every data set. This meant that all data from years other than 2015, 2016, 2017, and 2018 had to be removed. We also had to remove null values, because some countries did not have reports for those exact years. Thus, we had to remove all rows that contained *NaN* values.

From among these four years, we utilized *mean* across the columns to find the average GDP and military expenditure for each country over the years 2015-2018.

Next, the data sets had to be merged so that all the data was in one data set. An inner join was used to ensure that the resulting data set only included entries related to countries that appeared within every original data set. We merged the two data sets on the country names, so that in the resulting data set, each row had a country that appeared in both datasets with its associated GDP and military expenditure. We then used those two columns to create a new column for each country, which showed the percentage of GDP dedicated towards the military expenditure.

For the happiness score dataframe, we utilized *groupby* and *aggregate* to find the average happiness scores of the countries between 2015-2018. This data set was different from the other two, because the years were values in a column, whereas in the other dataset, the years were the columns themselves. Thus, we had to drop the rows that had a year outside of 2015-2018 in the year column, and then we did *groupby* on the country names. We then utilized *aggregate mean* in conjunction with *groupby* to take the mean happiness over those 4 years for each country.

Finally, we merged the average happiness datatable with the military expenditure as % of GDP table, using another inner join to match the country names.

Once this was completed, the data was finally in a usable format, because each country was listed with its associated average military expenditure as % of GDP and its average happiness score, spanning 2015-2018.

Part 3: What modules are we using?

Module 5: Statistical Inference

We used this model to determine whether there was sufficient evidence that countries that spend a higher percent of their GDP on the military have a different happiness score than countries that spend a lower percent of their GDP on the military. Our **justification** is that we needed to perform a hypothesis test in order to answer this question, as we had to understand whether the results were due to random chance or not. The **concept** we used to answer this question was a two-sided t-test because we were trying to determine whether there was a difference between two groups of data. We used this module during the data analysis and final report **stages**.

Module 6: Combining Data

We used this module to combine our three datasets in order to use the observations from all datasets in our analysis. Our **justification** is that we needed to combine the datasets because our questions can only be answered when our datasets are used in conjunction with one another. Though explained more in depth above, the **concepts** we used were merging, mean values, groupby, and aggregating, as well as cleaning the data and clearing null values. This module allowed us to combine our datasets together, and be able to view the necessary associated values for each country. We used this module in the data cleaning and data investigation **stages**.

Module 8: Visualization
We used this module to create charts to better communicate our findings to our audience. Our **justification** for using this module is that, oftentimes, it is easier for the audience to understand data when it is presented in a visualization rather than just a table of numbers. The **concept** we used to create our visualizations was the Seaborn library in Python. This library allowed us to create a number of charts, including a scatter plot and a box plot. We used this module during the data analysis and final report **stages**, in order to visualize the observe correlations or alternatively, the lack of correlations.

Module 9: Prediction & Supervised Machine Learning
We used this model to measure the correlation between the different factors of happiness and military spending. Our **justification** is that we needed to use this module to understand the relationship between the different factors and how each one affected the others. The **concepts** we used were measuring correlation between the military spending data and the happiness data, utilizing the r^2 value and MSE. To do this, we used linear regression, as we wished to understand the correlation between different factors. With these values, we were able to quantify correlation, seeing if happiness or any contributing factor had a correlation to military spending. We also utilized this module to show and explain the correlation between each factor of happiness and overall happiness. We used this module in the data analysis **stage**.
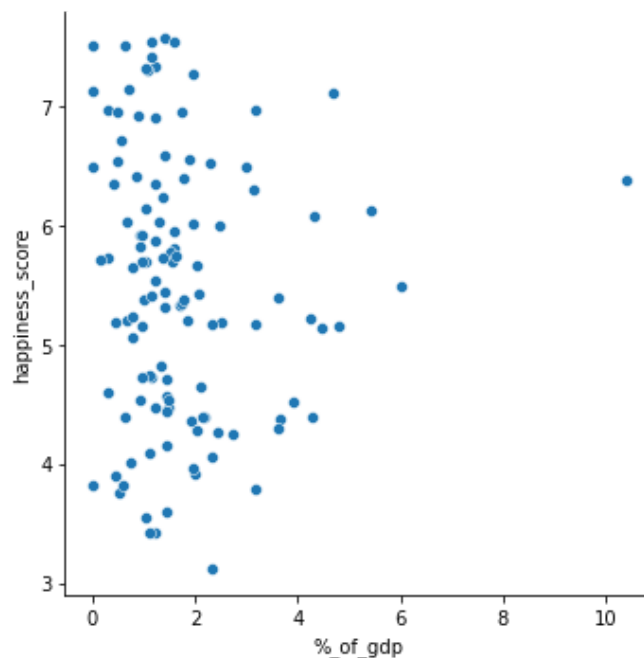
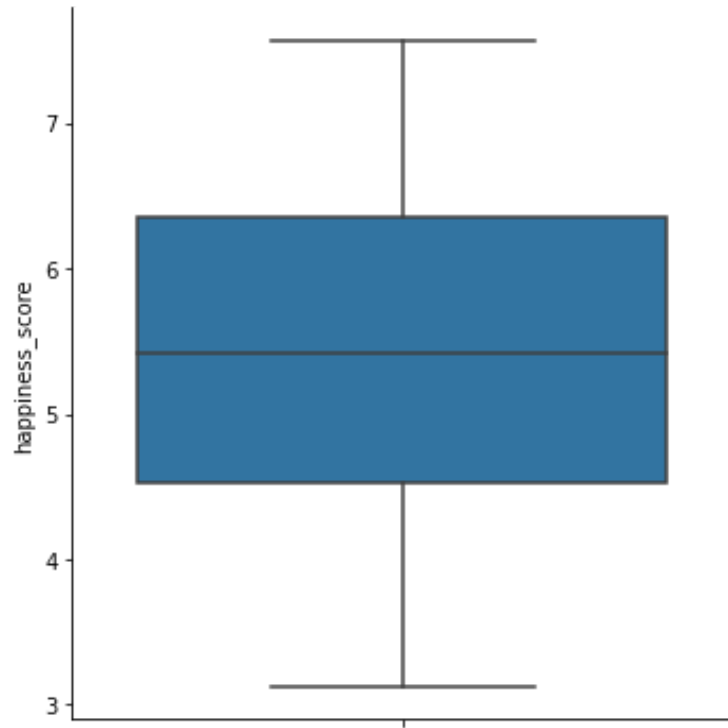## Part 4: Results and Methods

Figure 1

Figure 2

% of GDP Spent on Military	r^2	MSE
Happiness	0.0024	1.2546
Freedom	0.0316	0.0192
Corruption	0.0038	0.0107
Health	0.0009	0.0510

Figure 3
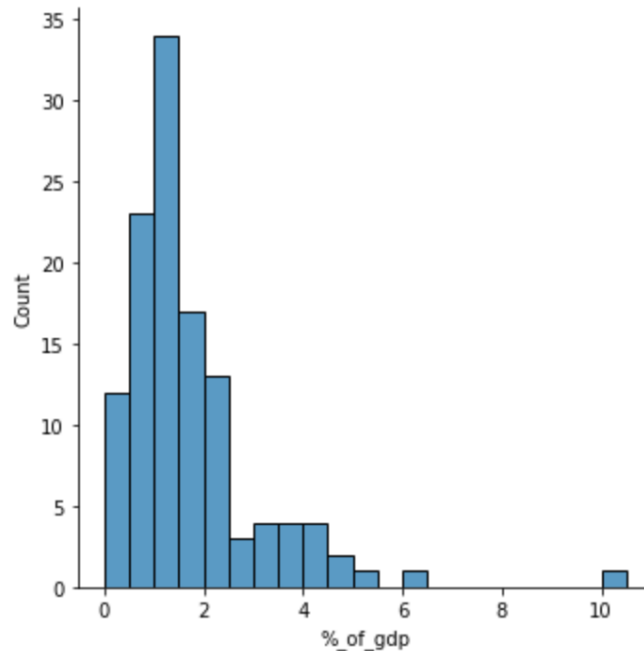
Figure 4

Correlation to Happiness	r^2	MSE
Freedom	0.3352	0.0132
Health	0.6420	0.0183
Corruption	0.2317	0.0083

Figure 5

In order to read the full coding methods, click the following link to re-read the "Data Preparation/Wrangling" section. In summary, we began by consolidating the data from across the three data sets into one data set that we could use to then answer our questions. We started by filtering out all data that was not from 2015, 2016, 2017, or 2018. Then, we averaged the data from across these years for each country to get the data for each country for average GDP, average military expenditure, and average happiness scores. Finally, the data sets were merged with the country name as the key, and we had our final data set that we used to answer our research question. We utilized **Figure 4** to further analyze this final data set which demonstrates that the majority of countries spend between 0-2% of their GDP on the military.

To determine the relationship between percent of GDP spent on military and a country's happiness, freedom, corruption, and health, we created linear regression models to predict these values based on military spending. We had to make four different linear models, each with a different target for the regression: one with overall happiness as the target, one with freedom as the target, one with government trust as the target, and one with health as the target. The r^2 scores and mean squared error (MSE) for each of these linear regression models can be seen in **Figure 3**. Our primary question focused on determining if there was a correlation between military spending (as a percentage of GDP) and overall country's happiness. The linear model where the overall happiness was the target produced an r^2 score of 0.0024, revealting that there is no correlation between a country's happiness and military spending.

This conclusion is supported by **Figure 1**, which also indicates that there does not seem to be a relationship between military spending and happiness score. In order to further bolster this claim about a lack of correlation, we utilized a hypothesis test. The null hypothesis was that the 25% of countries that spend the most on their military have the same happiness score as the 25% of countries that spend the least in terms of percentage of GDP. We chose this null hypothesis because of the fairly even distribution of countries' military expenditure, as shown in **Figure 2**, so this null hypothesis would be able to verify a lack of correlation with happiness. A p-value of 0.08 was calculated, meaning that we fail to reject this null hypothesis. Our failure to reject this null hypothesis lends support to our discoveries about the lack of correlation, since we cannot reject the claim that countries with higher percentiles of military spending have the same happiness as countries with lower percentiles of military spending.

Our question then expands to a more specific category, where we wanted to analyze the specific factors of happiness (freedom, health, and government trust) and see if we could derive any correlation to military spending. As explained above, we accomplished this by changing the overall happiness target of the models to the specific factors of health, freedom, and government trust. Looking at the r^2 value for each of these more specific linear regression models, the results remained quite close to 0, revealing similarly very low correlations between each factor of happiness and the military spending.

Last, we analyzed the relationship between factors of happiness and overall happiness in order to discover which factors had the highest correlation with overall happiness as seen in **Figure 5**. We discovered that each factor had a positive correlation with overall happiness. Furthermore, we found that health had the highest r^2 at 0.64. The freedom factor had the second highest correlation, with an r^2 score of 0.33. The most inelastic factor proved to be corruption, with an r^2 score of 0.23. Thus, we find that as overall happiness increases, the health expectancy factor typically increases significantly, while the corruption factor remains more constant. These conclusions reveal that health expectancy is a more elastic factor and is more likely to be influenced. Alternatively, the corruption factor is more firmly set, with minimal changes when overall happiness is increased.

In order to access all of our code and the data sets we used, click on this link:
https://duke.box.com/s/6j2kcisk9twk4mqcnn88likap7vrhnuy

Our three data sets are listed as "militaryexpenditure.csv", "WorldHappiness_Corruption_2015_2020.csv", and "gdp.csv." Our code file is called "216_Final_Project.ipynb"

Part 5: Limitations and Future Work

While we feel our project provides important insights that help us answer our research question, it is important to understand that there are some key limitations to our findings. Arguably, the most important limitation has to deal with the actual research question itself. The question of what a country's overall happiness is is extremely nuanced, and one could easily make the argument that this cannot be expressed as a single number like it is in our data set. Therefore, a critic may make the argument that our report is over-simplified.

Furthermore, when running a scientific experiment, we usually try to limit outside factors to try and create a controlled environment for our experiment. By the nature of our question and the complexities of the world, it is probable that there are underlying factors contributing to our results that we do not know about. The number of countries in the world, and thus in our data set, is relatively small, meaning that certain hidden factors could greatly impact our findings.

There are many possibilities for future work related to our report. For example, as mentioned above, our investigation could look at much more specific factors. As measurements such as overall

happiness are broad and hard to quantify, one would be wise to narrow down the question and data to be much more specific before taking any sort of action based on the results. Furthermore, as the military spends money on many different things, future reports could investigate different categories of military spending and their impact on countries.

Part 6: Conclusion

We successfully aggregated the data and were able to discern the desired correlation between all of our variables. Our primary question focused on determining if there was a correlation between military spending (as a percentage of GDP) and a country's happiness. After aggregating all the data, we produced a final data set of every country, which included their average military spending percentage over 2015-2018 and the average happiness scores over 2015-2018. Through our linear models, we did not find any correlation between military spending and overall happiness. We also found 0 correlation between military spending and any sub-factors of happiness.

Our secondary question was focused on determining the strength of the correlation between overall happiness and each of its sub-factors. We found that health expectancy is the largest contributor to happiness, sentiments of corruption are the smallest contributor, and freedom is the medium contributor.