Henri Einroos
Mia Melanie Moor

PROJECT E11: KAGGLE-HUMAN-FREEDOM-AND-SUICIDE REPORT

Task 2. Business understanding

1. Business goals:

1.1. Background:

Globally medicine has started to reach the point at which we can treat many acute illnesses extremely effectively and that is an amazing accomplishment, but now we are left facing other dark aspects of human suffering on the physical side. While progress on chronic illness remains limited, an even more neglected and historically taboo domain demands attention: mental health.

Studying mental health is challenging, particularly on a large epidemiological scale, as its outcomes, good or bad mental health, are difficult to quantify. One stark exception, however, is suicide - a proxy even more taboo than mental health itself.

Suicide has been present in every recorded society, dating back thousands of years before the birth of Christ. But what has driven, and continues to drive, people to end their own lives? Suicide rates vary widely between countries. Even after accounting for outliers such as war-torn regions, the rankings seem inconsistent, with affluent Western nations interspersed among countries with fewer resources and vastly different governance structures. What could then be the driving forces behind this killer? This is the question we are going to attempt to tackle with our project.

1.2. Goals:

While we do not have a specific client commissioning this study, we believe that analyzing the relationship between mental health and what we consider the best proxy metric - suicide, would provide valuable insights. This analysis could be particularly beneficial to ministries of social affairs in various governments and mental health advocacy groups, shedding light on the nature of suicide and its potential risk factors on a national level.

Additionally, since we are utilizing data provided by the Cato Institute, an important secondary goal would be to test the robustness of their metrics. By investigating whether meaningful causal relationships exist between their metrics and real-world outcomes such as suicide rates, we hope to lend further legitimacy to their methodologies and the standards they have established.

Then use the findings to create a predictive model that could help identify early warning signs of increased suicide risk within specific populations or contexts.

1.3. Success criteria:

Successfully identify the top 10 features out of the nearly 40 metrics provided by the Freedom Index that have the strongest correlation or predictive power for suicide rates and highlight factor combinations that may be linked to mortality. If possible to create a model to predict the amounts of suicides for a country when given their freedom index that would have an accuracy north of at least 70%. Validate the models performance using techniques such as cross-validation or out-of-sample testing.

Henri Einroos
Mia Melanie Moor

2. Assessing our situation:

2.1. Inventory of resources:
For the project we will be using our two personal laptops for analysis and report writing, Jupyter Notebook running Python for data analysis and visualization. We will be sourcing our data from The Human Freedom Index Kaggle database and from the Cato institutes website for human freedom annual reports that are not already in the Kaggle database, and that overlap with the years we have data for in our other Kaggle database with data on global suicide rates.
During the describing of the data phase we have decided to also use some data from the Fraser institute's Economic freedom index, also freely available on the internet.

2.2. Requirements, assumptions, and constraints:
The schedule will largely rely on the homeworks and deadlines given to us in the Introduction to Data Science course. The data we are using is free and in the public domain but we must remember to declare as much as possible from where we got our data originally. The analysis will be performed using Python, utilizing libraries such as pandas, NumPy, and scikit-learn for statistical and machine learning tasks.

2.3. Risks and contingencies:
The possible risks that could affect the project are time management and allotment, to tackle we will produce a schedule and attempt to stick to it to the best of our current ability. Also unforeseen circumstances like illness could affect the completion of our project, then one team member would have to take over the workload of the other member as well. If there happens to be a technical problem with our computers, we would need to possibly work on the computers in the Delta library for example.

2.4.Terminology:
HFI (Human Freedom Index) - A composite metric that measures personal, civil, and economic freedoms across nations.
Procedural Justice (column from HFI) - The average of three indicators: Right to life and security (evaluates violations during arrests or searches), due process and rights of the accused (assesses presumption of innocence, humane treatment, and access to evidence), and freedom from arbitrary interference (includes violations like unauthorized wiretapping).
Internal Organized Conflict (column from HFI) - measure the extent to which war or armed conflict with internal aggressors impinges on personal freedom in observed countries, based on the Global Peace Index's qualitative assessments.
HDI (Human Development Index (column from suicide rates)) - a summary measure of average achievement in key dimensions of human development: a long and healthy life, being knowledgeable and having a decent standard of living.

PCA Analysis (Principal Component Analysis): A statistical technique used to reduce the dimensionality of datasets while preserving as much variability as possible, often useful for identifying key predictors in large datasets.
SVL Model: Refers to a supervised learning model used in machine learning for predictive tasks (e.g., regression or classification).

2.5. Costs and benefits:
The cost of this project is basically only our time, work and effort, we do not necessarily invest any money in this. The benefit of this project would be a good grade and passing the course, if we do well and put a lot of effort into this project.

3. Defining our data-analysis and machine learning goals:

3.1. Goals:
The goal is to see if there are significant correlations between The Human Freedom Index and suicide rates and to find the most important factors linked to suicide, exploring patterns among countries to understand trends better.
Then we also want to try creating a machine learning model that predicts suicide rates based on data from the Human Freedom Index and some from the Economic Freedom annual reports, aiming for decent accuracy (70%).

3.2. Data-mining success criteria:
Success means identifying key factors influencing suicide rates, on which we could build a model with acceptable accuracy. Then creating clear visuals and reports that could help explain the findings to others and putting that info onto a concise poster that we successfully present in the poster session, also effectively addressing questions from the audience.

Task 3. Data understanding

1. Gathering data

1.1. Outline data requirements
The required data is available online on the Kaggle website and on the Cato institute website, which are both free to access and use. All files have different file format options: CSV is the preferred option but a few years of HFI data is only available in the XLSX format.
The few choice data points that we are using from the Fraser institute's annual reports are only given in a XLSX format.

1.2. Verify data availability
The data for the years not on Kaggle is available also on the Cato institute website and looking around it is possible to find many more years worth of data than is on the most popular human freedom database on Kaggle.

1.3. Define selection criteria

Since with time the HFI has added more features to its reports, we will only be working with the features that are present in all of our selected years under study which should be the overlapping years 2010-2016 for both of our data sources. We will be discarding all data for all countries directly engaged in a major nationwide conflict or war during our study, since it could be a major confounding factor which could overshadow the impact of real metrics for the given country.

2. Describing data

   The suicide database: values for almost all countries from 1986-2016. Some years for some countries are missing. The values for each year are separated by sex and by the 5 age ranges (`5-14 years, 25-34 years, 35-54 years, 55-74 years, 75+ years`) and the total gross number of people in that age and sex category in that country in that year.

   The human freedom index: has many more features than is good to list here but the main categories are things like: Legal Protection and Security, Specific Personal Freedoms and economic freedom. Under these categories there are a lot of different features each named and rated on a scale from 0-10 0 being the lowest possible score and 10 the highest. All countries are again graded separately each on their own row. The data from the Cato Institute's XLSX files does not include all of the economic metrics. Upon investigating their methodology, we found that for the economic freedom component of the HFI, they have used the Economic Freedom Index produced by the Fraser Institute. After reviewing the data behind the Economic Freedom Index, we have decided not to use all of its features. Instead, we will focus on 'sub-summaries' of different categories that may be useful, such as regulation, soundness of money, and size of government, along with a few other summary metrics. These are all also graded on a scale from 0 to 10.

3. Exploring data

   The HFI dataset consists of 40 columns, with each representing different aspects of freedom or related societal measures. The number of rows is different from year to year, but it is usually around 150. The rows correspond to different countries, while columns include metrics such as civil liberties, political rights, freedom of expression, economic freedom, and more. In the Human Freedom Index all of the columns go from a scale of 0-10.

   The suicide rate dataset has 12 columns representing the age group, sex and the count of suicides in those categories, and a few other factors, like population, HDI and GDP. There are a lot of rows, because different from the HFI, each country does not have one row, but multiple, for all the age groups. There is not a particular scale for the values, because each column represents very different things which can vary a lot from country to country.

4. Verifying data quality
   The data seems to be good enough to support our goals. We performed basic data checks and verified that there are no major formatting issues, such as inconsistent data types or corrupted values. There seems to be some missing data and although the amount is not excessive, it does require attention to ensure that it does not significantly impact the quality of our analysis. We will address this by removing rows that have too many missing values to maintain the integrity of our results.

Task 4. Planning our project

List of tasks:

1. Begin by exploring the datasets to understand their structure, contents, and any potential issues. Clean the data to remove inconsistencies. Identify and drop rows or columns that could negatively impact the analysis, such as countries involved in ongoing wars. Align the datasets in terms of years and features.
2. Conduct preliminary statistical analyses, to summarize the data and identify trends. Apply techniques like Principal Component Analysis (PCA) to uncover the most important features and simplify the dataset for further analysis.
3. Use clustering techniques like K-Means Clustering to group countries based on similarities in their Human Freedom Index metrics and suicide rates. This will help reveal patterns and relationships, such as which combinations of freedom metrics are associated with higher or lower suicide rates.
3. Build a machine learning model to predict suicide rates using the cleaned and processed data. Experiment with different algorithms (e.g., regression, random forests, etc) and fine-tune the model to achieve acceptable accuracy, ideally above 70%.
4. Analyze feature importance from the predictive model to determine the metrics most closely associated with suicide rates. Use this data to show which factors might influence mental health outcomes the most and outline the most important features.
5. Design an awesome poster to showcase the project's methodology, key findings, and conclusions. Ensure it includes clear visuals such as graphs and charts (but only the most valuable ones), along with concise text explaining the main takeaways. Prepare to present the poster and answer questions during the session.