PROJECT W4: KAGGLE-HUMAN FREEDOM

Rudolf Stenar Saluoks, Kristina Lillo, Killu Allik

TASK 2. BUSINESS UNDERSTANDING

1. Identifying business goals

Background

In a democratic country, it is really important that all citizens have good knowledge about what kind of decisions are beneficial for the country. In recent years, there is a big rise of populist parties and fake news, which means that citizens could be informed better, so they would know how to decide what is really right and what is not. Nowadays countries are so different and people are mobile. People are travelling or even moving to different countries, but what they usually do not know are differences in aspects of freedom, which affect their lives every day.

Business goals

We want to give a better overview to the citizens of democratic countries that how different aspects of freedom are related to GDP growth, life expectancy, etc. So that they can make more informed decisions when voting.

Business success criteria

Our analysis has been successful if someone tells us that our analysis was interesting and educating to read. Also, if people actually understand how aspects of freedom are affecting their lives and decisions.

2. Assessing our situation

Inventory of resources

We are 3 students, who have to contribute to the project about 30 hours each. We have possibility to consult with our course mates and lectors. We have data about human freedom related subjects by countries by the years of 2008 to 2016 and also information about GDP, life expectancy, etc. We can use Jupyter Notebook for analyzing our data.

Requirements, assumptions, and constraints

4.12.2019 - Choosing which indicators to focus on in Dataset I.

6.12.2019 - Preparing Dataset I and Dataset II.

8.12.2019 - Finding correlations between all indicators.

8.12.2019 - Comparing 5-10 countries with the highest and lowest freedom rating.

11.12.2019 - Finding out, which aspects of freedom are related to economic growth, population growth and so on.

15.12.2019 - Making and introducing our final poster.

All datasets are available and public. No restrictions for us.

Risks and contingencies

1. End of semester -> have a lot to do in different courses and have deadlines for other homeworks. Solution: try to plan our time as good as possible.
2. Different contributions from team members -> Solution: motivate each other and ask if there are any problems.
3. Sickness -> Solution: other team members must work harder to get project done by deadline.

Terminology

Human freedom index - The Human Freedom Index presents the state of human freedom in the world based on a broad measure that encompasses personal, civil, and economic freedom. Human freedom is a social concept that recognizes the dignity of individuals and is defined here as negative liberty or the absence of coercive constraint. Because freedom is inherently valuable and plays a role in human progress, it is worth measuring carefully. The Human Freedom Index is a resource that can help to more objectively observe relationships between freedom and other social and economic phenomena, as well as the ways in which the various dimensions of freedom interact with one another.

Costs and benefits

Costs: poster and printing -> for us actually free. Benefits: lots of good information to others.

3. Defining your data-mining goals

Data-mining goals

Our goal is to paint a broad but reasonably accurate picture of the extent of overall freedom in the world. A larger purpose is to more carefully explore what we mean by freedom and to better understand its relationship to any number of other social and economic phenomena. We will try to find correlations between indicators and find out how different aspects of freedom affect people. Also, we compare 5-10 countries with the highest and lowest freedom rating to show the contrast between countries.

Data-mining success criteria

Our analysis has been successful if we find strong connections and patterns between highest freedom rating countries and the same thing about countries of lowest freedom rating. If we can show, that these certain aspects of freedom are affecting decisions, then in the future people can be more subjective about countries and decisions they will make.

TASK 3. DATA UNDERSTANDING

1. Gathering data

Outline data requirements (necessary data)

Dataset I.

The dataset contains 79 distinct indicators of personal and economic freedom for 162 countries. Covered areas are: Rule of Law; Security and Safety; Movement; Religion; Association, Assembly, and Civil Society; Expression and Information; Identity and Relationships; Size of Government; Legal System and Property Rights; Access to Sound Money; Freedom to Trade Internationally and Regulation of Credit, Labor, and Business. Updated up to 2016.

Dataset II.

The dataset should contain information about the same 162 countries in the following areas: population, average income, area and GDP. (2008 - 2016)

Dataset III

Should contain info about the Human Freedom Index, such as score, rank and quartile. (2008 - 2016)

Verify data availability

All datasets are available.

Define selection criteria

Dataset I [1] Remove columns that we won't use. Removed columns: pf_identity_sex_male, pf_identity_sex_female. (Data from 2008 - 2016)

Dataset II [2]. We will filter out the necessary 162 countries. Used columns: total population, male and female population, land area, population density, expense, revenue, gross domestic income (GDI).

Dataset III [3]. Only extract columns hf_score, hf_rank, hf_quartile. (Data from 2008 - 2016)

2. Describing data

Dataset I.

Source: hfi_cc_2018.csv

Description and columns: Can be found on the Kaggle page [1]. The data is from 2008 to 2016.

Dataset II.

Source: Dataset2.csv (Filtered from the original data [2], currently has data for 2016).

General description: Contains info about population, GDI, expense and revenue of 162 countries, the data is from 2008 to 2016.

Columns

- (String) **Country Name**
- (Decimal) **Expense** (% of GDP)
- (Decimal) **Gross domestic income**
- (Integer) **Land area** *km^2*
- (Decimal) **Population density** *people per km^2 of land area*
- (Integer) **Population, female**
- (Integer) **Population, male**
- (Integer) **Population, total**
- (Decimal) **Revenue, excluding grants**

- (Decimal) **Life expectancy at birth, total (years)**
- (Decimal) **Military expenditure** (% of GDP)

Dataset III.

Source: hf.csv (Filtered from [3], currently has data for 2016).

General description: Contains the human freedom index score, ranking and quartile for 162 countries (2008 - 2016).

Columns

- (String) **ISO_code** *ISO code of country*
- (String) **countries** *Name of country*
- (Decimal) **hf_score** *Human freedom index*
- (Integer) **hf_rank** *Current ranking among the countries*
- (Integer) **hf_quartile** *Quartile of freedom*

3. Exploring data

Dataset I. There are quite a lot of NaN values, but they are mostly related to specific countries and columns. Many of the columns have values on the edge of the range (etc 10 or 0, when the range is from 0 to 10). Data in the later years is clearer than in the earlier years.

Dataset II. This data has been chosen from the initial dataset. By giving a look at the data, we think that this data should be the one we will try to predict from the Dataset I and Dataset III.

Dataset III. Data with just the score and rank of the country in human and economic freedom. These scores are calculated by the scores of the Dataset I.

4. Verifying data quality

Dataset III has no missing values and can be fully used. Dataset II and I however, have multiple missing values in different columns. This could be because these datasets have a lot of countries and columns that we actually don't need. So we have two possibilities to clean that data, we can either exclude countries with missing values or exclude columns that have missing values. If needed, we can also remove some of the earlier years of the data. Dataset I has missing values because some countries were added later on, the data from 2016 has no missing values, we will take this into account when using this dataset.

TASK 4. PLANNING (5 tasks, hours per team member)

1) Choosing which indicators to focus on in Dataset I. (Each member 1 hour)

2) Preparing Dataset I and Dataset II. (Rudolf 10 hours and Killu 10 hours)

- ❏ Task 2.1. Removing countries from Dataset II that we are definitely not going to use.
- ❏ Task 2.2. Removing indicators from Dataset I that we are not going to use.
- ❏ Task 2.3 Adding additional indicators to Dataset II that could be interesting / come in handy.

3) Finding correlations between all indicators. (Kristina 10 hours, Killu 5 hours)

- ❏ Task 3.1. Bring forward any interesting correlations.
- ❏ Task 3.2. Report pairs that are the most positively correlated and negatively correlated.

4) Comparing 5-10 countries with the highest and lowest freedom rating. (Kristina 5 hours, Killu 10 hours)

- ❏ Task 4.1. Try to find any anomalies in the data.
- ❏ Task 4.2. Analyze and report anomalies that do not fit into these characteristics.

5) Finding out, which aspects of freedom are related to economic growth, population growth and so on. (Kristina 10 hours, Rudolf 10 hours)

6) Making and introducing our final poster. (Kristina 5 hours, Killu 5 hours, Rudolf 10 hours)