D13: KAGGLE-World Happiness Report

Team: Laura Raudsepp, Anna-Helena Salurand, Raul Niit

Task 1. Setting up

Link to project Github repository: https://github.com/raulniit/IDS2020

Task 2. Business understanding

- Identifying your business goals
    - Background

      Most countries in the world are striving towards a better future for its people - ultimately, their goal is to provide the population with a happy life. Achieving it is a complex and time-consuming mission. Governments use several tools: pass new legislation or decide on state budget allocation. Which area of life or problem in the society to focus on often depends on the political beliefs of the ruling parties.

      Effectiveness of policies and funding is often assessed on a micro-scale, looking into whether the specific problem got solved and to what extent. However, do governments and policymakers zoom out to macro-level often enough to understand whether efforts are being invested into the most impactful areas in terms of making people's lives happier?

      Understanding the universal drivers of human happiness would help to ensure that we are investing our resources in the most important areas of life in terms of happiness.

    - Business goals

      Although the task clearly asks to state a clearer goal than "making the world a better place", the goal of this project is along those lines. More specifically, we want to explain which society-describing factors are in correlation with human happiness.

      The grand goal of such analysis would be for governments to make better decisions in resource allocation to drive happiness levels. A more modest goal would be to provide the project team and the readers of the report insights into what are the common denominators between countries that tend to have people

leading a happier life. By doing so, the aim of this project is to raise the awareness among the readers of the project results in the following areas:

- *which fields of life or factors (among those that are impacted by humans) are in correlation with human happiness and should thus be preferred when choosing the field of live for future career or investment decisions,*
- *Which country-describing factors should be considered when choosing a place abroad for travelling or relocating, if one wants to lead a happy life.*

- Business success criteria

  The project team should find at least three factors describing states or societies, that are in correlation with the world happiness score, and three factors, that are not correlated significantly with the happiness score.

- Assessing your situation

  - Inventory of resources

    Human resources: *project team consists of three members: Laura (2nd year Informatics BSc), Anna (2nd year Actuarial and Financial Engineering MSc) and Raul (3rd year Mathematical Statistics BSc). There are four instructors for this course, who can be consulted.*

    Hardware and software: *we can use Python and we have personal computers in use.*

    Data and insights: *the project is based on a Kaggle competition, the dataset on world happiness scores is provided. There are numerous public Notebooks on this topic which we can read up on to gain more insights into the topic and data. Furthermore, there are several public databases available to source data about factors describing countries that we can test for correlation with happiness.*

  - Requirements, assumptions, and constraints

    We are required to do at least 30 hours of work per student for this project and get at least 50% of points for the project to pass the course. We need to submit our source code, poster and 3-minute video by noon of Dec 14. We need to participate in the online event for presenting and discussing projects with fellow students and instructors.

  - Risks and contingencies

    There are no major risks or contingencies that the team has identified, besides the general Covid-situation in Estonia, which might affect our lives in unexpected

ways.

- Terminology

  - World happiness score - *countries' happiness level indicator from the World Happiness Report.*
  - Correlation - *a connection between two variables, measured from -1 to 1. (absolute value shows the strength and sign shows the direction of the relationship).*
  - Regression model - *a model that describes the relationship between a dependent and independent variable(s).*

- Costs and benefits

  The main cost of this project is the time spent on it by the project team. On the benefit side, we have new insights into the topic of happiness, and new learnings in the field of data science. As we have chosen to embark on this project journey, we have assessed the learnings we can get out of this project to be worth the time spent.

- Defining your data-mining goals

  - Data-mining goals

    The data-mining goal of the project is to train a regression model explaining the happiness score based on at least three factors that are in correlation with it. Furthermore, we aim to illustrate happiness score data and the final model with visualisations for easier comprehension of the topic and project outcomes.

  - Data-mining success criteria

    To assess the regression model, we can use criteria such as RMSE to choose the best model out of those we train in the process. We can do significance tests to assess whether the independent variables in the regression model are significant.

    Assessing the success of the visualisations can be done during the project presentation day based on the feedback from the audience.

Task 3. Data understanding

- Gathering data

- Outline data requirements

  To address our data mining goals we would need
  1) data on the happiness score and subscores for many countries (World Happiness Report)
  2) some additional data that correlates with the happiness scores (GDP per Capita).
  Because our goal is to find at least 3 factors that correlate with the happiness score and 3 factors that don't, then we will need to find more additional datasets along the way to use in this project.
  These datasets should be in some reasonable format like .csv or .xlsx.

- Verify data availability

  All the required datasets are available and have been downloaded, in case the data becomes unavailable online for some reason.

- Define selection criteria

  The data sources for this project are:
  1. *World happiness report acquired from Kaggle. All of the data that is available for every year (2015-2019) and every country (155+) will be used.*
  2. *Projections for world countries GDP per capita acquired from USDA. Only the projections for years 2015-2019 will be used.*

- Describing data

  ***Kaggle** is a platform for numerous open datasets, which can be used for data science. The world happiness report data (https://www.kaggle.com/unsdsn/world-happiness) consists of 5 .csv files for every year from 2015 to 2019 and for 155+ countries. In each of those files rows represent countries and columns represent various attributes: country name, happiness rank, happiness score and scores for different sub-parts of the calculated happiness score. In our project we will only use those sub-parts which are reported for all the years, which are health, economy, freedom, generosity and trust in the government.*

  ***USDA** (United States Department of Agriculture) has economic projections data for all the world countries. The Projections for world countries GDP per capita dataset (https://www.ers.usda.gov/data-products/international-macroeconomic-data-set).is a .xls file, where rows represent countries, columns represent years and the cells show the projected GDP per capita for a specific country and for a specific year. Although the data*

is given for years 2011-2031, we will only use the data from 2015 to 2019, which match the happiness report data.

- Exploring data

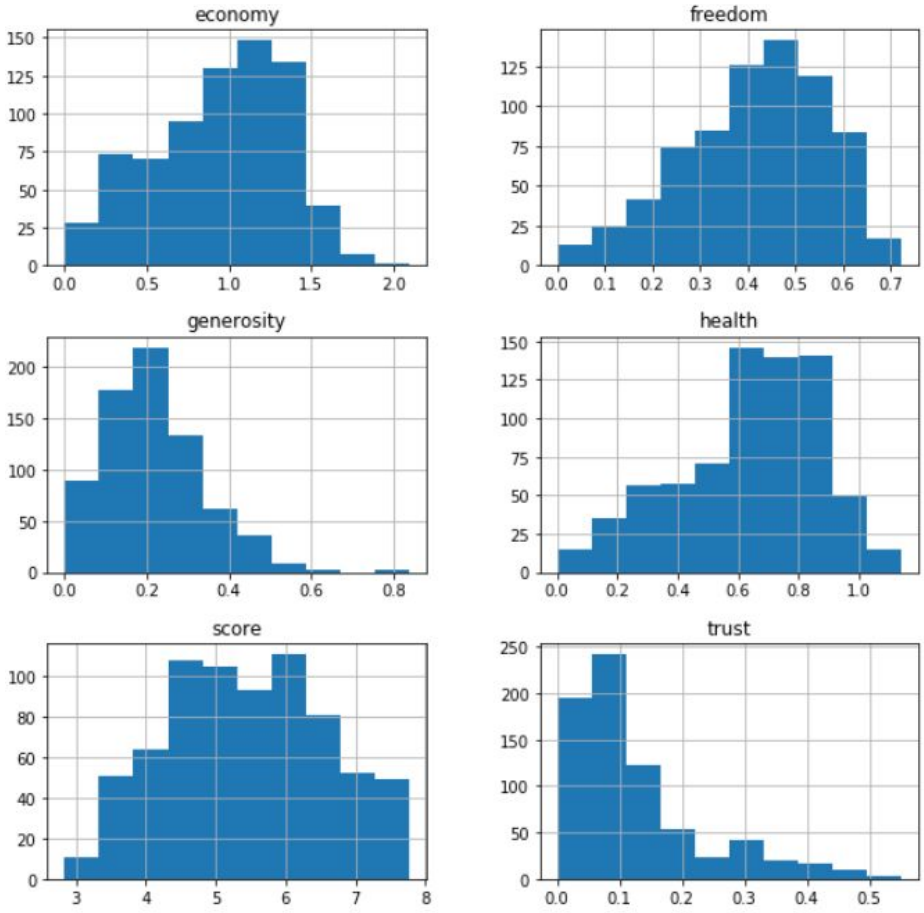

Figure 1. Histograms of World Happiness Report attributes

For exploring the distribution of the happiness report attributes, histograms were plotted (Figure 1). As we can see the dependent variable "score" seems to be from normal distribution, while all the sub-parts of the score are more skewed in either direction.

	score	economy	health	freedom	trust	generosity
count	725.000000	725.000000	725.000000	725.000000	725.000000	725.000000
mean	5.442050	0.943891	0.635260	0.415346	0.124678	0.219106
std	1.118515	0.397239	0.235546	0.148180	0.107150	0.124204
min	2.839000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	4.553000	0.657000	0.499000	0.312000	0.053000	0.128890
50%	5.430000	1.007610	0.669000	0.434000	0.089000	0.201870
75%	6.293000	1.252785	0.814550	0.531640	0.155000	0.282000
max	7.769000	2.096000	1.141000	0.724000	0.551910	0.838075

Figure 2. Properties of World Happiness Report attributes

Figure 2 shows the range, mean, standard deviation and quantiles of the happiness report attributes. Some aspects to consider in terms of data quality are the 0 values of some of the sub-scores (are these possible or just missing values?) and also the high standard deviation compared to the mean for "trust" and "generosity".

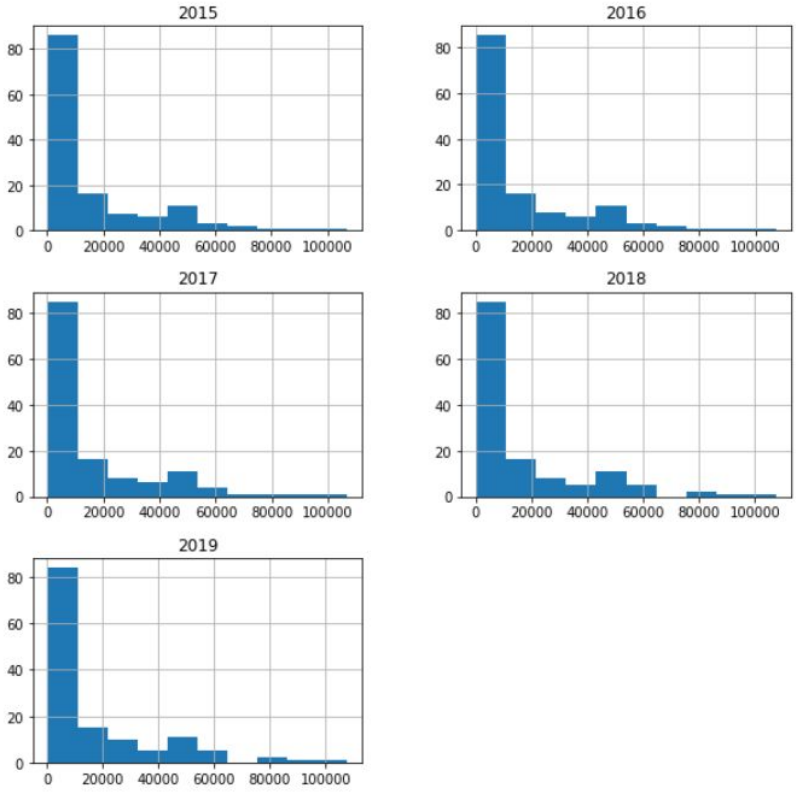

Figure 3. GDP per capita histograms

Histograms of the GDP per capita data seem to indicate that the statistic forms an exponential distribution. This also confirms that the economic sub-score from the World

Happiness Report dataset is not purely based on GDP per capita. Data quality-wise we have some missing values for some of the countries from the World Happiness Report dataset.

- Verifying data quality

  Both of the datasets used in this project do not indicate any major quality issues. The World Happiness Report dataset has one missing value, but because the data expands over five years, we can replace the missing value with imputation methods.

  The biggest problem is that some countries are not present in both of the datasets. The proportion of these countries is not very large (10/145), so one solution would be to find the model without these values. Another solution could be to find the data from some other sources, for example we could find GDP per capita from certain countries' official homepages.

  Overall our data seems to be of good enough quality that we can conduct our analysis and reach our goals.

Task 4. Planning your project

(Make a detailed plan of your project with a list of tasks.There should be at least 5 tasks. Specify how many hours each team member is going to contribute to each task.)

First two are basically done.

1. UNDERSTANDING. Finding the purpose and defining business objectives for the problem that needs to be tackled.
   First phase of any data analytics project is understanding the business that our project is part of. This is essential for ensuring projects' success. In order to have direction and purpose, we have to identify a clear objective of what we are gonna do with the data, and what specific questions we want to answer.
   ~2 hour per person

2. OBTAIN. Gathering data from relevant sources.
   Mixing and merging data from as many data sources as possible is what makes a data project great. We have already downloaded our datasets from Kaggle, but we also decided to use some open data from the Internet so we could compare even more aspects of the data.
   ~1-2 hours per person

3. SCRUB. Exploring and cleaning our data to formats that the machine understands.
   The main part of this task is to link everything together to achieve our original goal.One crucially important element of data preparation not to overlook is data cleaning (handling the missing values and correcting typos).
   ~3 hour per person.

4. EXPLORE. Enriching our dataset and getting the most value of it.
   This task should start by joining all our different sources/datasets to narrow our data down to the essential features (selecting important features and constructing more meaningful ones). Also finding significant patterns and trends using statistical methods so we could form hypotheses by visually analyzing the data.
   ~12 hours per person

5. MODELLING. Constructing models to predict and forecast
   Building helpful visualisations -> getting insights and predicting future trends.
   This means training machine learning models, evaluating their performance and using them to make predictions. Using different statistical modeling methods and determine which is the best for our data. For example using regression models.
   ~6 hours per person

6. INTERPRET. Proving its effectiveness and putting the results into good use.
   For presenting our work we have to make a poster explaining motivation, objectives, used methods, presenting main results in an understandable way. Being able to tell a story with our data will help explain the value of our findings.
   ~4 hours per person

List the methods and tools that you plan to use. Add any comments about the tasks that you think are important to clarify.

Tools:

- Our main tool is Python.
- At the end we're probably gonna use some data visualization tool (Tableau or something else).

Methods:

- Regression models
- Tests for significance