PROJECT PROPOSAL

Understanding the Indicators of World Happiness

Kristen Goh and Jonathan Goh
Woz U & Bethel School of Technology
January 9, 2021

Executive Summary

World happiness is an increasingly important topic for policymaking and health behavior programs. This project aims to provide additional analysis on happiness scores of different countries around the world and explore other significant influencers including health, socioeconomic, and lifestyles factors. Using open-source datasets provided on Kaggle and the World Bank Data Repository, the team will analyze the base dataset from the World Health Report tying it together with data from the World Health Organization (WHO) and the World Bank.

This project is planned to be completed within six-weeks culminating in a presentation provided live via Zoom and a PowerPoint published on GitHub. The project plan and final product will be approved by the Data Science Program Chair and Instructor at Woz U.

This document will lay out the objectives, background information on the datasets, functional requirements, and provide a timeline for the different deliverables associated with this project. The planned deadline is February 14, 2021.

Business Objectives

- Identify how happiness differs by country and region.
- Identify the socioeconomic and lifestyle indicators that influence happiness.
- Identify health metrics that influence happiness.
- Explore how happiness scores have changed over time.

Background

The World Happiness Report was first published in 2012 and continued annually. It was first released at a United National event. The report ranks 155 countries by their happiness levels and looks at other indicators to help inform policy-making decisions. One of the most important sources of data utilized for the World Happiness Report comes from the Gallup World Poll.

In addition to the World Happiness Report data, we will also utilize the data from the World Health 2020 Kaggle dataset. The data includes health metrics for countries recognized by the World Health Organizations (WHO). The dataset includes data on chronic, acute, and infectious diseases scores.

This project aims to provide further analysis utilizing the country happiness scores along with socioeconomic, health, and lifestyle indicators. We will investigate these other potential influencers of happiness and their impact over time. As a result of the project,

we hope to provide insights to other significant factors associated with happiness and the quality of life for residents living in the different countries around the world.

Scope

This project is about the happiness score of each country in the world between 2015-2020. This project will also explore the data by identifying top and bottom scores over time. Furthermore, the project will use other datasets to discover other associations between happiness and socioeconomic, lifestyle, and health factors.

This project will not explore county or city level happiness scores or factors.

Functional requirements

Data Exploration: The team will familiarize themselves with the dataset and background of the World Happiness Report. They should have a good understanding of what each column means, and how the values are measured. They will brainstorm on questions to ask, and what they might gather from the dataset. Then, they will identify the proper functions to create models, predictions, etc.

Data Wrangling: The downloaded dataset should be cleaned, sorted and made ready for analyzing. Columns and unusable columns should be removed. The data types for each column should also be converted to a usable format as necessary for analysis. Majority of data wrangling will occur in Python.

Data Analysis: The team will test the assumptions for the statistical analysis and run the ANOVA and Multiple Linear Regression tests. Graphs and box plots will be created to understand the distribution and relationship of the variables. Majority of the data analysis will occur in R Studio.

Data Visualization: In this phase, with a comprehensive understanding and learned insights from the analysis, they will work on visualizing the findings. Tableau and Excel will be utilized for graphing and creating interactive dashboards. The visuals will be exported into a PowerPoint slideshow.

Presentation: The team will present their findings in PowerPoint presentation and publish the report on Github.

Data Sets

1. World Happiness Report up to 2020 (*linked here*) - Kaggle
    - World Happiness Report began in 2012 by the United Nations and is a landmark survey of the stats of global happiness of about 150 countries.

2. World Development Indicators (*linked here*) – World Bank Data Repository
    - Sourced from World Bank's collection of annual development indicators of over 217 economies.

3. World Health Statistics (*linked here*) - Kaggle
    - Data collected by the World Health Organization (WHO) consisting of health metrics.

Tool and Software Needed

- Team Communications and Task Reminder
    - Trello
    - Slack
    - Outlook
- Data Analysis and Visualization Programs
    - R Studio
    - Python
    - Excel
    - Tableau
- Data Storage
    - GitHub
    - Google Drive
- Password protected Laptop or Desktop Computers

Statistical Tests

- ANOVA
- Multiple Linear Regression

Personnel requirements

The team will work closely for this project to create a report and presentation on their findings. The team will problem-solve and check-in on work progress. During weekly huddles, they will review the past week workload and plan out the next week. They will take turns being the Scrum Master. Communications will be primarily completed on Slack and project planning on tello using the KanBan scheduling system.

Once a week, the team will meet with their instructor and Project Champion. The weekly meetings will be utilized to report on progress, ask questions, and seek guidance for the next steps.

They may also consult with their coding mentor for additional support.

Delivery schedule

Week 1: January 10, 2021

- Select dataset(s)
- Develop evaluation questions
- Import dataset into preferred software to begin data wrangling. Any unnecessary columns should be removed
- Research about the topic

Week 2: January 17, 2021

- Set up source control through GitHub
- Explore the dataset and ask questions
- Brainstorm possible correlations and determine if the data is normally distributed
- Brainstorm potential predictive models
- Visualize the data to see if there are any interesting findings in Tableau

Week 3: January 24, 2021

- Modeling/Optimization (ANOVA, and Multiple Linear Regression)
- Visualize the data to see if there are any interesting findings in Tableau

Week 4: January 31, 2021

- Review and validate findings from the previous week
- Draw insights/conclusions

Week 5: February 7, 2021

- Compile findings into a PowerPoint slideshow
- Review draft of presentation with instructor to ensure that the presentation is clear and logical
- Work on the style and layout of the presentation

Week 6: February 14, 2021

- Make final touches to the Power Point presentation
- Practice presenting
- Present findings via Zoom and publish final product to GitHub

Other requirements

All programs used should be free of charge though the team may decide to use a paid service, such as a more advanced version of Tableau.

Assumptions

The software programs and platforms we will use should be available, up-to-date, and not broken.

Limitations

This project does not have additional budget for other advanced software programs. The main programs to be utilized for statistical analysis are mainly open-sourced.

This project is planned to be completed in roughly a six-week period. Three main datasets will be utilized from Kaggle. Limited research on happiness and health will be completed within this time-period. The project members have an academic background in healthcare.

Risks

The risks that may arise are such as natural disasters, power outages, family emergencies or broken software/hardware. These unforeseen circumstances may cause delay. Any issues will be brought up with the instructor in a timely manner.

The project is not dependent on securing funding. Datasets are provided on open-source websites. All laptops and programs are incurred by the team members themselves.