***Preliminary Project Proposal***

***DSE6311***

***Basics***

For this week Alisha is taking the role as team lead and Annelise will be taking the role of recorder. Since we are a team of two we will be sharing the role as Spokesperson. The title for our preliminary project is Global Perspectives on Happiness: Investigating Quality of Life Determinants Across Nations

***Background & Question***

The research question of interest is: What factors contribute to the quality of life as a determinant of happiness in various countries? This question fills the need to incorporate better determinants for a happier life. It provides awareness of which factors influence the quality of life, which is correlated to happiness. This information could be used preventatively. If living in a country with high stress factors negatively influencing quality of life - need to improve factors so a high quality of life is recognized globally.

This question is important to support everyone's background. Quality of life measures are important statistics to ensure a healthy and prosperous society. This question is not novel as this is a very sought after topic. Though it is a well known topic, it is important to add different layers and perspectives to the research. Global happiness is a topic that could help a multitude of overarching themes. Some of the happiest countries have some of the highest global peace ratings (<https://www.bbntimes.com/society/happiest-and-safest-countries-in-the-world>).

Our hypothesis for this project is: Happiness is affected by quality of life in developing nations because of higher disease rates and lower GDP. In addition to this our prediction is: People in developing nations would have lower happiness levels due to a lower quality of life. People in developed nations have a higher quality of life because of less diseases and more access to resources.

**Data & Analysis**

The data sets chosen were “world happiness report” and a “health and demographics” dataset. They contain measures of health (presence of diseases such as measles, HIV/Aids etc), financial measures (GDP, income level ect), which contribute to both quality of life and the definition of happiness as outlined by this report. These will be the predictor variables in the model. Happiness will be the response variable for high quality of life. The happiness data points will be recoded to whole numbers for efficiency when creating a model. The data is segregated by country and year to further help analysis, as emphasis on whether the nation is developed or not. We chose these data sets as we wanted to take a deeper dive into various health and socioeconomic factors that affect happiness and were able to get this information by using two datasets. The first dataset is called “World Happiness Report” (<https://www.kaggle.com/datasets/unsdsn/world-happiness>) and the second dataset being used is “Health and Demographics Dataset” (<https://www.kaggle.com/datasets/uom190346a/health-and-demographics-dataset>). The following table summarizes the variables being used in the model.

| Data Set | Column | Description | Other Notes |
| --- | --- | --- | --- |
| Life Expectancy | Country |  | Variable on which data sets will be combined |
| Status | developing/ developed | Socio-economic Variables |
| GDP | Gross domestic product (economic output of the country) |
| Income composition of resources | Index reflecting the income distribution and access to resources in the country |
| Population |  |
| Schooling | The average number of years of schooling received by the population |
| Adult Mortality | Probability of dying between ages of 15 - 60 per 1000 people | Health Variables |
| Infant Deaths | Number of infant deaths per 1,000 live births |
| Alcohol | Average alcohol consumption in liters per capita |
| BMI | Average body mass index of population |
| HIV/ AIDS | Prevalence of HIV/AIDS as a percentage of the population |
| Happiness | Country |  | Variable on which data sets will be combined |
| Happiness score | Surveyed metric of happiness from 0-10 |  |
| Family | Score of impact of family on happiness |  |
| Freedom | Score of impact of freedom on happiness |  |
| Trust (Government corruption) | Score of perception of government corruption as an impact to happiness |  |
| Generosity | Score of extent to which generosity contributed to happiness |  |

*Table 1. Variables in both datasets being used including descriptions and overarching categories.*

To begin performing an in depth analysis, the data will first need to be stripped and cleaned. The two datasets will be merged based on the “country” variable, which is present in both data sets. Only the data columns mentioned in table 1 will be kept and the rest will be deleted. The data will then be checked for consistency and completeness. Any data points with missing data will either be filled in through an average aggregation for that country or be left blank. A preliminary analysis will be done to visualize the data at hand to ensure countries are being represented uniformly across different metrics.

Additional exploratory data analysis and multivariate regression analysis will be conducted to identify key determinants that strongly influence happiness. Our tentative analysis plan includes a supervised learning methodology. This approach involves training a model on labeled data, where the outcomes (in this case, happiness levels) are known, to make predictions or classifications on new data for future analysis. Supervised learning is suitable for our research question as we aim to predict happiness levels based on various predictor variables. Specifically, we will use predictive modeling techniques to analyze the relationship between quality of life determinants and happiness levels across different countries. This model will allow us to assess the extent to which socio-economic and health factors contribute to variations in happiness scores.

A potential pitfall we have thought about is getting lost in the meanings of our data question. For example, we are defining the quality of life variable through happiness which also needs to be defined. We will know if our question is answered by looking at the happiness scores broken down by determinants. Seeing whether negative determinants are higher in developed or developing countries. We will also know whether or not our hypothesis is supported by looking at the direction (developed countries have a lower quality of life and happiness) will be supported through data analysis and visualization.

***Technical Details***

For this project the language we plan on coding in is R. When thinking about additional resources doing some background research into determining whether a country is developed or not along with a greater understanding of what factors make a country developed. Here is the link to our to your GitHub repo:

<https://github.com/loboannelise99/DSE6311>