Explain your Data (3 pages) and answer the following questions:

Why do you use these datasets?

- **Life Expectancy Dataset:** We chose this dataset because it contains health data of most of the countries of the world. It also contains economic data, so it was interesting to see the correlation between economy and health data. One of the more important reasons for why we chose this dataset was because it contained a country column that contained a lot of countries similar to the Happiness dataset. This country column was able to act like a "primary key". Lastly, this dataset contained data from the 2000's to 2015, and our research question primarily focused on data from 2015.
- **Happiness Dataset:** We chose this dataset because it contains information on happiness. To be specific, this dataset looked at multiple factors related to happiness such as social support, corruption perception, freedom, and more. Also, this dataset contained a country column similar to the Life Expectancy Data, so the country column in the Happiness Dataset was able to act similarly to a "foreign key" to connect our data to the Life Expectancy Dataset. Lastly, this dataset contained updated data for 2015, which was perfect because the life expectancy dataset also contained data from 2015. We wanted our data to be consistent so it was very important that both datasets had data from the same year.

Where do your datasets originate from?

- **Life Expectancy Dataset:** We got this data from Kaggle, but the dataset for Life Expectancy originated from the Global Health Observatory (GHO) data repository under the World Health Organization.

- **Happiness Dataset:** We got his data from Kaggle, but it was originally from the Gallup World Poll.

Who collected the data?

- **Life Expectancy Dataset:** The health data from the data set was collected by the World Health Organization. Meanwhile, the economic data was collected by the United Nations.
- **Happiness Dataset:** This data was primarily collected by the Gallup World Poll, but some of the data was collected from World Health Organization, World Development Indicators, and other other published journal articles. The data was then prepared and researched through the help of these universities: Sustainable Development (Columbia University), Centre of Economic Performance (London School of Economics), Vancouver School of Economics (University of British Columbia), Wellbeing Research Centre (University of Oxford), and the Helping and Happiness Lab (Simon Fraser University).

The data usage permissions: are the data freely available for everyone? Are there certain restrictions on who can access data, or for what purpose it can be used (e.g. non-commercial only)?

- **Life Expectancy Dataset:** The data is publicly available meaning that anyone can access it. The purpose of this data is for health data analysis.
- **Happiness Dataset:** This data is publically available. The main purpose of this data is to help the Sustainable Development Solutions Network of the United Nations, but it could be used by anyone.

What are the datasets about? What is the population and what is the sample?

- **Life Expectancy Dataset:** This dataset focuses primarily on immunization, mortality, economic, social, and other health related factors for most countries in the world. I wasn't able to track down the original datasets from WHO (because they have so much data), but I'm assuming the population is all the countries looked at in the dataset. As of sample, since I couldn't find the original dataset's info, I am unable to supply anything on sample.
- **Happiness Dataset:** This data set focuses on multiple factors that relate to happiness on an international scale. Contents include happiness rank, happiness score, economy (GDP), family, health (life expectancy), freedom, trust (gov corruption), generosity, and more. The population of this dataset includes most countries of the world. The sample of this dataset aims to have 2000+ people per country to have a large enough sample size to reduce random sampling errors. It has a 95% confidence interval.

Discuss any data quality issues you encountered.

- **Life Expectancy Dataset:** A data quality issue that we encountered in this dataset was that there were a few NA values. Also something to note according to the author of this dataset on Kaggle was that less known countries that didn't have enough information were excluded from the dataset such as Tonga, Togo, and Cabo Verde.
- **Happiness Dataset:** Some countries with this dataset didn't align with the Life Expectancy data, so those countries were excluded. Some things to note from the World Happiness Report, which is the website that includes all the information about this data, is that there were a few out of date data (data not from updated to 2015) at the time when they were collecting and creating this dataset. For these data, they mixed other related up-to-date data with mathematical calculations to infer these values.

Life Expectancy (WHO)

- Basic info:
    - From: Global Health Observatory (GHO) data repository under World Health Organization (WHO)
        - Health data: WHO data repository
        - Economic data: United Nations
    - Public
    - Purpose: health data analysis
    - Meta data:
        - Health status and other factors for all countries (193 countries)
    - Release Date: 2015
        - Data from 2000 - 2015
    - Data merge through R
    - Countries without enough data → removed from data frame
        - Vanuatu, Tongo, Tofo, Cabo Verde

Happiness Data

- Survey measure
    - From: Gallup World Poll (GWP)
    - Release Date: December 26, 2014
        - Covers years from 2005 - 2014
    - National avg response to question of life evaluations
    - Analogy of question: “Please imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?
        - Name of measure: “Cantril life ladder” or “Life ladder”
    - Sample size:
        - Use most recent years → up-to-date
        - Combine data from (2012 - 2014) → sample size large enough to reduce random sampling errors → 95% confidence interval
        - 2000+ people per country
- Additional info:
    - Public work

    - Presented every year to United Nations on March 20 (International Day of Happiness)
    - Publisher: Sustainable Development Solutions Network
    - Preparations/Research help:
        - Sustainable Development (Columbia University)
        - Centre of Economic Performance (London School of Economics)
        - Vancouver School of Economics (University of British Columbia)
        - Wellbeing Research Centre (University of Oxford)
        - Helping and Happiness Lab (Simon Fraser University)
- Measure techniques:
    - GDP per capita
        - Date: November 6, 2014
        - From: World Development Indicators (WDI)
        - Exceptions: Taiwan, Syria, Argentina
            - Used earlier released data (before Nov 6, 2014) → adjusted levels by factor of 1.17
                - Why 1.17?
                    - (US GDP per capita [2011 prices])/(US GDP per capita [2005 prices])
            - Looked at GDP per capita time series forecast from OECD Economic Outlook
    - Corruption Perception
        - Technique:
            - National average of survey (GWP)
                - 2 questions:
                    - "Is corruption widespread throughout the government or not"
                    - "Is corruption widespread within businesses or not?"
                - Overall perception = avg of 2 responses
                    - If gov corruption answer missing → use ONLY business corruption answer
                    - Corruption perception (national level) = avg response of overall reception at individual level
    - Healthy Life Expectancy (HLE)
        - Measure: healthy life expectancy at birth
        - From: calculated by authors based on data from World Health Organization (WHO), World Development Indicators (WDI), and statistics published in journal articles
        - Exceptions:

    - Hong Kong, Puerto Rico, Kosovo
    - Use mathematical estimates
- Social support (having someone to count on in times of trouble)
  - Measure: binary response (either 0 or 1)
  - Question: “If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?”
- Freedom
  - Measure: national avg of responses
  - Question: “Are you satisfied or dissatisfied with your freedom to choose what you do with your life?”
- Generosity
  - Measure: residual of regressing national avg of response
  - Question: “Have you donated money to a charity in the past month?” on GDP per capita
- Positive affect
  - Measure: avg of 3 positive effect measures (GWP)
    - Happiness:
      - Question: “Did you experience the following feelings during A LOT OF THE DAY yesterday? How about Happiness?”
    - Laugh:
      - Question: “Did you smile or laugh a lot yesterday?”
    - Enjoyment:
      - Question: “Did you experience the following feelings during A LOT OF THE DAY yesterday? How about Enjoyment?”
- Negative effect
  - Measure: avg of 3 negative effect measure (GWP)
    - Worry:
      - Question: “Did you experience the following feelings during A LOT OF THE DAY yesterday? How about Worry?”
    - Sadness:
      - Question: “Did you experience the following feelings during A LOT OF THE DAY yesterday? How about Sadness?”
    - Anger:
      - Question: “Did you experience the following feelings during A LOT OF THE DAY yesterday? How about Anger?”

- Dystopia:
    - Imaginary country that has world’s least happy people (opposite of utopia)
    - Purpose: establish benchmark
    - Measure: lowest scores for each six key variables
        - World’s lowest incomes
        - Lowest life expectancy
        - Lowest generosity
        - Most corruption
        - Least freedom
        - Least social support