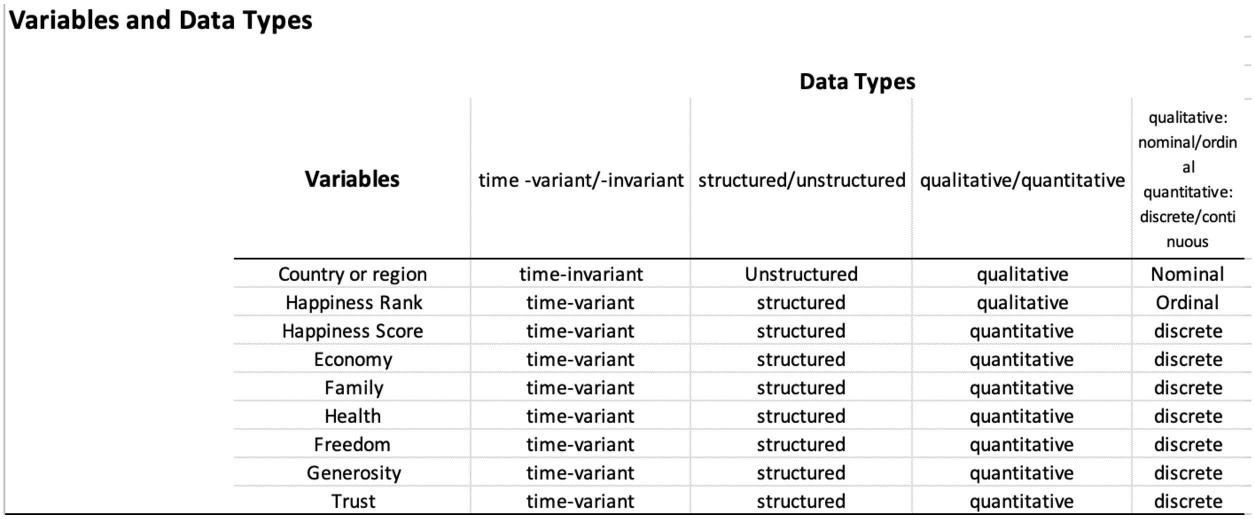
6.1 Sourcing Open Data
World Happiness Report 2015-2019

Data Source:

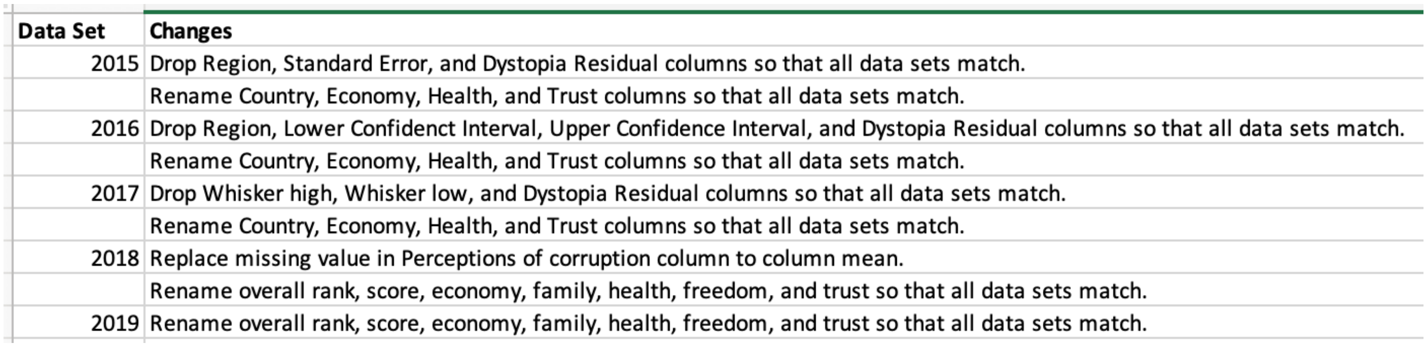
- The topic I have chosen for achievement 6 is the World Happiness Report 2015-2019 which is a landmark survey of the state of global happiness. The data used is external open data that comes from the Gallup World Poll and is based on answers to the main life evaluation questions. The same questions are asked every time in the same way. This makes it possible to trend data form year to year and make direct country comparisons. Telephone surveys and face to face interviews are used to collect the data. Telephone surveys take about 30 minutes while face to face interviews can be an hour. The samples are weighted, probability based, and nationally representative of the resident population aged 15 and older.
- Data variables: overall rank, country or region, score, GDP per capita, social support, healthy life expectancy, freedom to make life choices, generosity, and perceptions of corruption.
- I chose this data set for a few reasons. The data comes from a reliable source and holds the information needed to create a successful project. I also find the information to be very interesting. I am eager to learn what variables effect the happiness score most in each country and how each country differs from year to year.

Data Profile:



Variables and Data Types

	Data Types			
Variables	time -variant/-invariant	structured/unstructured	qualitative/quantitative	qualitative: nominal/ordinal quantitative: discrete/continuous
Country or region	time-invariant	Unstructured	qualitative	Nominal
Happiness Rank	time-variant	structured	qualitative	Ordinal
Happiness Score	time-variant	structured	quantitative	discrete
Economy	time-variant	structured	quantitative	discrete
Family	time-variant	structured	quantitative	discrete
Health	time-variant	structured	quantitative	discrete
Freedom	time-variant	structured	quantitative	discrete
Generosity	time-variant	structured	quantitative	discrete
Trust	time-variant	structured	quantitative	discrete



Data Set	Changes
2015	Drop Region, Standard Error, and Dystopia Residual columns so that all data sets match.
	Rename Country, Economy, Health, and Trust columns so that all data sets match.
2016	Drop Region, Lower Confidenct Interval, Upper Confidence Interval, and Dystopia Residual columns so that all data sets match.
	Rename Country, Economy, Health, and Trust columns so that all data sets match.
2017	Drop Whisker high, Whisker low, and Dystopia Residual columns so that all data sets match.
	Rename Country, Economy, Health, and Trust columns so that all data sets match.
2018	Replace missing value in Perceptions of corruption column to column mean.
	Rename overall rank, score, economy, family, health, freedom, and trust so that all data sets match.
2019	Rename overall rank, score, economy, family, health, freedom, and trust so that all data sets match.

Consider Limitations and Ethics:

- With telephone interviews, random phone numbers are selected. It is possible that individuals did not answer.
- The sampling may not accurately represent the entire country.
- Those in places without telephone access and deemed dangerous for face-to-face interviews were not counted for.
- Collection bias, sample bias, and exclusion bias may occur.

Define Questions to Explore:

- What countries or regions rank the highest in overall happiness?
- What countries or regions rank the highest in all six factors contributing to happiness?
- Does the happiness rank of each country change over the years?
- What factors contributed to the change in country rank if any?
- How did the countries scores change over time?
- Did any countries experience a significant increase or decrease in happiness scores?
- What factors contributed to the change in country happiness scores if any?