6.1 Sourcing Open Data

Inhaltsverzeichnis

Objective

- The aim of the project is to show how the Happiness Score has developed and changed globally between 2015 and 2023.
- It will show which countries have the highest happiness score, which have the lowest, and whether the same countries hold the same positions over the entire time period, as well as which specific factors determine the level of the happiness score.
- The project will also provide a forecast of the Happiness Score of the different countries for the next year.
- The main project questions and hypotheses are described in the final section of this document.

Data sets

- Since a new World Happiness data set is collected each year by Gallup World Poll (GWP), we have as a base nine different data sets that were transfomed into one data set during the initial data preparation/cleaning process.

Initial Data sets:

- World Happiness Data 2015
- World Happiness Data 2016
- World Happiness Data 2017
- World Happiness Data 2018
- World Happiness Data 2019
- World Happiness Data 2020
- World Happiness Data 2021
- World Happiness Data 2022
- World Happiness Data 2023

Transformed Data set:
World Happiness Data 2015 – 2023
The final transformed and cleaned data set has 15 Columns and 1470 rows.

Data source for all initial Data sets:

- The Data source ist he same for all initial data sets. The data stems from Kaggle.
    - Link (Data for 2015-2022): https://www.kaggle.com/datasets/mathurinache/world-happiness-report?select=2022.csv
    - Link (Data for 2023): https://www.kaggle.com/datasets/ajaypalsinghlo/world-happiness-report-2023

- But the data provided there come from the annual Happiness reports since 2012. The data that come from the Gallup World Poll (for more information see the Gallup World Poll methodology).

Data Collection

The rankings are based on answers to the main life evaluation question asked in the poll. This is called the Cantril ladder: it asks respondents to think of a ladder, with the best possible life for them being a 10, and the worst possible life being a 0. They are then asked to rate their own current lives on that 0 to 10 scale. The rankings are from nationally representative samples, for the years 2013-2023. They are based entirely on the survey scores, using the Gallup weights to make the estimates representative. The typical annual sample is 1,000 people for every country. However, there are many countries that have not had annual surveys.

Data Limitations, Ethics & Biases

Limitations
The data comes from an established worldwide annual survey, so we can speak of a reliable source here. However, the data used was generated by Kaggle members, so the table names and the selection of columns per dataset vary somewhat. Thus, we have missing values, especially for the year 2022.
We also don't have data for every year for every country since many countries don't do the survey annually.

Ethical issue: The data contains no personal Information or indiscrimnates variable,so no PLA Security is required.

Biases: I decided to use the interpolation method to fill up missing values. This may be a potential bias, since for example, if you interpolate a linear trend between two points, you're assuming that the data follows a linear trend between those points. This may not always be the case.

Data Relevance

The used data sets are of high importance for our project, since we can only show the course of the Happiness Score geographically through all years.

Data Content

Columns for the initial Data sets 2015-2023

Note: If I only removed points from Columns Names in Data Cleaning, then I did not list them separately here in this overview. (E.G.Trust.Goverment.Corruption)

Column	Definition	Included in which data sets?
Country	The country which was took part In the survey.	2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023
Region	The region a country is part of.	2015, 2016, 2017
Regional indicator	The former called column “Regionis called Regional Indicator in 2020 & 2021.	2020, 2021
Happiness Rank	The Rank (1 to 157) given, connected to the assessment of Happiness Score of a country.	2015, 2016, 2017
Overall Rank	The former called column “Happiness Rank” called Overall Rank in 2018 & 2019.	2018, 2019
RANK	The former called column “Happiness Rank” called Overall Rank in 2022.	2022
Happiness Score	Assessment of Happiness of a country based on all survey data on a scale of 1 to 10.	2015, 2016, 2017, 2018, 2019, 2022
Ladder Score	The former called column “Happiness Score” called Ladder Score in 2020, 2021 & 2023.	2020, 2021, 2023
Standard Error	Measure of how much an observed parameter - such as the mean or median - in a sample deviates on average from the true parameter of the population.	2015
Standard Error of Ladder Score	The former called column “Standard Error” called Standard Error of Ladder Score since 2020.	2020, 2021, 2023
Economy (GDP per Capita)	GDP per capita is in terms of purchasing power parity (PPP) adjusted to constant 2011 international dollara, talen from the World Development Indicators released by the Worldbank on November 28, 2019. Since 2020 the equation uses the natural log of GPD per capita.	2015, 2016, 2017
GDP per Capita	The former called column “Economy (GDP per Capita)” is called GDP per Capita in 2018 & 2019.	2018, 2019

Logged GDP per Capita	The former called column "Economy (GDP per Capita)" is called Logged GDP per Capita in since 2020.	2020, 2021, 2023
Family	Assessment of Family (or having someone to count on in times of trouble) is the national average of the binary responses (either 0 or 1) to the 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?"	2015, 2016, 2017
Social support	The former called column "Family" is called social support since 2018.	2018, 2019, 2020, 2021, 2023
Health (Life Expectancy)	Assessment of the healthy life expectancy at birth of a country on a scale of 1 to 10 <u>Relevant for the data from 2015:</u> The GWP adopted the following strategy to construct healthy life expectancy at birth for other country-years: first they generated the ratio of healthy life expectancy to life expectancy in 2007 for countries with both data, and assigned countries with missing data the ratio of world average of healthy life expectancy over life expectancy; then they applied the ratio to other years (i.e. 2005, 2006, and 2008-12) to generate the healthy life expectancy data.	2015, 2016, 2017, 2018, 2019, 2020, 2021
Freedom	Assessment of freedom of a country on a scale of 1 to 10, it is the national average of responses to the question "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?"	2015, 2016, 2017
Freedom to make life choices	The former called column "Freedom" is called Freedom to make life choices since 2018.	2018, 2019, 2020, 2023
Trust (Government Corruption)	The column represents the average of answers to two questions: "Is corruption widespread throughout the government or not" and "Is corruption widespread within businesses or not?" Coefficients are reported with robust standard errors clustered by country in parentheses. ***, **, * and + indicate significance at the 0.1, 1, 5 and 10% levels respectively.	2015, 2016, 2017
Perception of Corruption	The former called column "Trust (Government Corruption)" is called Perception of Corruption to make life choices since 2018.	2018, 2019, 2020, 2021, 2023

Generosity	The column represents the residual of regressing national average of response to the question “Have you donated money to a charity in the past month?” on GDP per capita.	2015, 2016, 2017, , 2018, 2019, 2020, 2021, 2023
Dystopia Residual	The residuals, or unexplained components, differ for each country, reflecting the extent to which the six variables either over- or under-explain average life evaluations.	2015, 2016, 2017, 2020, 2021, 2022, 2023
Lower Confidence Interval	The confidence intervals, as shown by the horizontal lines at the right-hand end of the country bars, show the range of values within which there is a 95% likelihood of the population mean being located. These are useful to readers wishing to see whether countries differ significantly in the average life evaluations.	2016
Whisker low	The former called column “Lower Confidence Interval” is called Whisker low since 2017.	2017, 2020, 2021, 2022, 2023
Upper Confidence Interval	The confidence intervals, as shown by the horizontal lines at the right-hand end of the country bars, show the range of values within which there is a 95% likelihood of the population mean being located. These are useful to readers wishing to see whether countries differ significantly in the average life evaluations.	2016
Whisker high	The former called column “Upper Confidence Interval” is called Whisker high since 2017.	2017, 2020, 2021, 2022, 2023
Ladder score in Dystopia	For the Gallup analysis a hypothetical country called “Dystopia” created. It has values equal to the worlds lowest national averages. The Dystopia score is calculated based on these lowest national averages in relative to the individual country averages.	2020, 2021, 2022, 2023
Explained by: Log GDP per capita	Share of the Dystopia score that can be explained by the log GDP per capita.	2020, 2021, 2022, 2023
Explained by: Social support	Share of the Dystopia score that can be explained by the social support.	2020, 2021, 2022, 2023
Explained by: Healthy life expectancy	Share of the Dystopia score that can be explained by theHealthy life expectancy.	2020, 2021, 2022, 2023, 2023
Explained by: Freedom to make life choices	Share of the Dystopia score that can be explained by Freedom to make life choices.	2020, 2021, 2022, 2023
Explained by: Generosity	Share of the Dystopia score that can be explained by generosity.	2020, 2021, 2022, 2023

Explained by: Perceptions of corruptio	Share of the Dystopia score that can be explained by perception of corruption.	2020,2021, 2022, 2023
Year		2015, 2016

Columns for the transformed Data set, includes Data from all data sets (2015-2023) plus the column "Year" to be able to connect the data to the specific years.

Column	Definition
Country	The country which was took part In the survey.
Region	The region a country is part of.
Happiness Rank	The Rank (1 to 157) given, connected to the assessment of Happiness Score of a country.
Happiness Score	Assessment of Happiness of a country based on all survey data on a scale of 1 to 10.
Standard Error	Measure of how much an observed parameter - such as the mean or median - in a sample deviates on average from the true parameter of the population.
Economy (GDP per Capita)	GDP per capita is in terms of purchasing power parity (PPP) adjusted to constant 2011 international dollara, talen from the World Development Indicators released by the Worldbank on November 28, 2019. Since 2020 the equation uses the natural log of GPD per capita.
Social support	Assessment of Family (or having someone to count on in times of trouble) is the national average of the binary responses (either 0 or 1) to the 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?"
Health (Life Expectancy)	Assessment of the healthy life expectancy at birth of a country on a scale of 1 to 10 <u>Relevant for the data from 2015:</u> The GWP adopted the following strategy to construct healthy life expectancy at birth for other country-years: first they generated the ratio of healthy life expectancy to life expectancy in 2007 for countries with both data, and assigned countries with missing data the ratio of world average of healthy life expectancy over life expectancy; then they applied the ratio to other years (i.e. 2005, 2006, and 2008-12) to generate the healthy life expectancy data.

Freedom	Assessment of freedom of a country on a scale of 1 to 10, it is the national average of responses to the question “Are you satisfied or dissatisfied with your freedom to choose what you do with your life?”
Trust (Government Corruption)	The column represents the average of answers to two questions: “Is corruption widespread throughout the government or not” and “Is corruption widespread within businesses or not?” Coefficients are reported with robust standard errors clustered by country in parentheses. ***, **, * and + indicate significance at the 0.1, 1, 5 and 10% levels respectively.
Generosity	The column represents the residual of regressing national average of response to the question “Have you donated money to a charity in the past month?” on GDP per capita.
Lower Confidence Interval	The confidence intervals, as shown by the horizontal lines at the right-hand end of the country bars, show the range of values within which there is a 95% likelihood of the population mean being located. These are useful to readers wishing to see whether countries differ significantly in the average life evaluations.
Upper Confidence Interval	The confidence intervals, as shown by the horizontal lines at the right-hand end of the country bars, show the range of values within which there is a 95% likelihood of the population mean being located. These are useful to readers wishing to see whether countries differ significantly in the average life evaluations.
Year	The year in which the individual data was collected.

Data Profile: Data understanding for the transformed Data set

Column	Qualitative/Quantitative	Discrete/Continous	Nominal/Ordinal/Binal
Country	Qualitative	/	Nominal
Region	Qualitative	/	Nominal
Happiness Score	Quantitative	Continous	Ordinal
Happiness Rank	Quantitative	Discrete	Ordinal
Standard Error	Quantitative	Continous	Nominal
Economy (GDP per Capita)	Quantitative	Continous	Ordinal
Social support	Quantitative	Continous	Ordinal
Health (Life Expectancy)	Quantitative	Continous	Ordinal
Freedom	Quantitative	Continous	Ordinal
Trust (Government Corruption)	Quantitative	Continous	Ordinal

Generosity	Quantitative	Continous	Ordinal
Upper Confidence Interval	Quantitative	Continous	Ordinal
Lower Confidence Interval	Quantitative	Continous	Ordinal
Year	Qualitative	Continous	nominal

Data Cleaning & Wrangeling

- Only the colums, that were cleaned/change will be listed here.
  I began the data cleaning by working through every initial Data set to establish the same name for the same column across the data sets and to establish a region and year column in order to connect all datasets in the end.
- I started to cleaning the data in terms of checking the consistency, finding missing values and duplicates only after I transformed the initial data sets into one data set. The process is documented in the Phython scripts.

Data set 2015

Problem	Description	Solution
Column-name “Family”	The column name “Family” is not same as in the other data sets.	I changed the name of the column to “Social Support”
No Year Column	Since my goal was to create a dataset with all data from the different years, I needed a Year column.	I added a column “Year” to the dataframe, where every value is 2015.

Data set 2016

Problem	Description	Solution
Column-name “Family”	The column name “Family” is not same as in the other data sets.	I changed the name of the column to “Social Support”
No Year Column	Since my goal was to create a dataset with all data from the different years, I needed a Year column.	I added a column “Year” to the dataframe, where every value is 2016.

Data set 2017

Problem	Description	Solution
Column-name “Family”	The column name “Family” is not same as in the other data sets.	I changed the name of the column to “Social Support”

Column-name “Happiness.Rank”	The column has a Dot in the name.	I changed the name of the column to “Happiness Rank”.
Column-name “Happiness.Score”	The column has a Dot in the name.	I changed the name of the column to “Happiness Score”.
Column-name “Trust..Government.Corruption”	The column has Dots in the name.	I changed the name of the column to “Trust (Government Corruption)”.
Column-name “Whisker.low”	The column name is not same as in the other data sets.	I changed the name of the column to “Lower Confidence Interval”.
Column-name “Whisker-high”	The column name is not same as in the other data sets.	I changed the name of the column to “Upper Confidence Interval”.
Column-name “Economy..GDP.per.capita”	The column name is not same as in the other data sets.	I changed the name of the column to “Economy (GDP per capita)”
Column-name “Dystopia.Residual”	The column has a Dot in the name.	I changed the name of the column to “Dystopia Residual”.
Column-name “Health..Life.Expactancy”	The column has Dots in the name.	I changed the name of the column to “Health (Life Expectancy”.
No year Column	Since my goal was to create a dataset with all data from the different years, I needed a Year column.	I added a column “Year” to the dataframe, where every value is 2017.
No Region column	Since my goal was to create a dataset with regional data that could also be grouped in region, I needed to add a region column.	I added a column “Region” to the dataframe, by first creating a subset with the region and country column from the data set 2015 and then merged this subset with the the data set 2017 with the key being the country column.

Data set 2018

Problem	Description	Solution
Column-name “Social support”	**The column name “Social support” has a “s” instead of an “S”.**	**I changed the name of the column to “Social Support”**
Column-name “Overall Rank”	**The column name is not same as in the other data sets.**	**I changed the name of the column to “**Happiness Rank**”.**
Column-name “Score”	**The column name is not same as in the other data sets.**	**I changed the name of the column to “**Happiness Score**”.**
Column-name “Perception of Corruption”	**The column name is not same as in the other data sets.**	**I changed the name of the column to “**Trust (Government Corruption)**”.**
Column-name “GDP per capita”	**The column name is not same as in the other data sets.**	**I changed the name of the column to “Economy (GDP per capita)”**
Column-name “Country or region”	**The column name is not same as in the other data sets and only includes Countries.**	**I changed the name of the column to “Country”.**
Column-name “Healthy life expectancy”	**The column name is not same as in the other data sets.**	**I changed the name of the column to “Health (Life Expectancy”.**
Column-name “Freedom to make life choices”	**The column name is not same as in the other data sets.**	**I changed the name of the column to “Freedom”.**
No year Column	**Since my goal was to create a dataset with all data from the different years, I needed a Year column.**	**I added a column “Year” to the dataframe, where every value is 2018.**
No Region column	**Since my goal was to create a dataset with regional data that could also be grouped in region, I needed to add a region column.**	**I added a column “Region” to the dataframe, by first creating a subset with the region and country column from the data set 2015 and then merged this subset with the the data set 2018 with the key being the country column.**

Data set 2019

Problem	Description	Solution
Column-name “Social support”	The column name “Social support” has a “s” instead of an “S”.	I changed the name of the column to “Social Support”
Column-name “Overall Rank”	The column name is not same as in the other data sets.	I changed the name of the column to “Happiness Rank”.
Column-name “Score”	The column name is not same as in the other data sets.	I changed the name of the column to “Happiness Score”.
Column-name “Perception of Corruption”	The column name is not same as in the other data sets.	I changed the name of the column to “Trust (Government Corruption)”.
Column-name “GDP per capita”	The column name is not same as in the other data sets.	I changed the name of the column to “Economy (GDP per capita)”
Column-name “Country or region”	The column name is not same as in the other data sets and only includes Countries.	I changed the name of the column to “Country”.
Column-name “Healthy life expectancy”	The column name is not same as in the other data sets.	I changed the name of the column to “Health (Life Expectancy”.
Column-name “Freedom to make life choices”	The column name is not same as in the other data sets.	I changed the name of the column to “Freedom”.
No year Column	Since my goal was to create a dataset with all data from the different years, I needed a Year column.	I added a column “Year” to the dataframe, where every value is 2019.
No Region column	Since my goal was to create a dataset with regional data that could also be grouped in region, I needed to add a region column.	I added a column “Region” to the dataframe, by first creating a subset with the region and country column from the data set 2015 and then merged this subset with the the data set 2019 with the key being the country column.

Data set 2020

Problem	Description	Solution
Column-name “Social support”	The column name “Social support” has a “s” instead of an “S”.	I changed the name of the column to “Social Support”
Column-name “Ladder Score”	The column name is not same as in the other data sets.	I changed the name of the column to “Happiness Score”.
Column-name “Perception of Corruption”	The column name is not same as in the other data sets.	I changed the name of the column to “Trust (Government Corruption)”.
Column-name “Logged GDP per capita”	The column name is not same as in the other data sets.	I changed the name of the column to “Economy (GDP per capita)”
Column-name “Country name”	The column name is not same as in the other data sets and only includes Countries.	I changed the name of the column to “Country”.
Column-name “Healthy life expectancy”	The column name is not same as in the other data sets.	I changed the name of the column to “Health (Life Expectancy”.
Column-name “Freedom to make life choices”	The column name is not same as in the other data sets.	I changed the name of the column to “Freedom”.
Column-name “Dystopia + Residual”	The column has a plus in the name.	I changed the name of the column to “Dystopia Residual”.
Column-name “Ladder Score in Dystopia”	The column name is not same as in the other data sets.	I changed the name of the column to “Dystopia Score”.
Column-name “Regional Indicator”	The column name is not same as in the other data sets.	I changed the name of the column to “Region”.
Column-name “lowerwhisker”	The column name is not same as in the other data sets.	I changed the name of the column to “Lower Confidence Interval”.
Column-name “upperwhisker”	The column name is not same as in the other data sets.	I changed the name of the column to “Upper Confidence Interval”.
No year Column	Since my goal was to create a dataset with all data from the different years, I needed a Year column.	I added a column “Year” to the dataframe, where every value is 2020.
No “Happiness Rank” Column	Since my goal was to create a dataset with all data from the different years that also includes the Rank-column, I	I added the column “Happiness Rank” to the dataframe by adding the “Happiness Rank” Column

	needed to add the year column.	from the data set 2019. I didn't need to create a subset, since I only needed the numbers from 1 to 157 and since the datset was already sorted based on the ranking I also didn't need a key to establish this column.

Data set 2021

Problem	Description	Solution
Column-name "Social support"	The column name "Social support" has a "s" instead of an "S".	I changed the name of the column to "Social Support"
Column-name "Ladder Score"	The column name is not same as in the other data sets.	I changed the name of the column to "Happiness Score".
Column-name "Perception of Corruption"	The column name is not same as in the other data sets.	I changed the name of the column to "Trust (Government Corruption)".
Column-name "Logged GDP per capita"	The column name is not same as in the other data sets.	I changed the name of the column to "Economy (GDP per capita)"
Column-name "Country name"	The column name is not same as in the other data sets and only includes Countries.	I changed the name of the column to "Country".
Column-name "Healthy life expectancy"	The column name is not same as in the other data sets.	I changed the name of the column to "Health (Life Expectancy".
Column-name "Freedom to make life choices"	The column name is not same as in the other data sets.	I changed the name of the column to "Freedom".
Column-name "Dystopia + Residual"	The column has a plus in the name.	I changed the name of the column to "Dystopia Residual".
Column-name "Ladder Score in Dystopia"	The column name is not same as in the other data sets.	I changed the name of the column to "Dystopia Score".
Column-name "Regional Indicator"	The column name is not same as in the other data sets.	I changed the name of the column to "Region".

Column-name “lowerwhisker”	The column name is not same as in the other data sets.	I changed the name of the column to “Lower Confidence Interval”.
Column-name “upperwhisker”	The column name is not same as in the other data sets.	I changed the name of the column to “Upper Confidence Interval”.
No year Column	Since my goal was to create a dataset with all data from the different years, I needed a Year column.	I added a column “Year” to the dataframe, where every value is 2021.
No “Happiness Rank” Column	Since my goal was to create a dataset with all data from the different years that also includes the Rank-column, I needed to add the year column.	I added the column “Happiness Rank” to the dataframe by adding the “Happiness Rank” Column from the data set 2019. I didn’t need to create a subset, since I only needed the numbers from 1 to 157 and since the datset was already sorted based on the ranking I also didn’t need a key to establish this column.

Data set 2022

Problem	Description	Solution
Column-name “RANK”	The column name is not same as in the other data sets.	I changed the name of the column to “Happiness Rank”
Column-name “Happiness score”	The column name “Social support” has a “s” instead of an “S”.	I changed the name of the column to “Happiness Score”.
Column-name “Dystopia + Residual”	The column has a plus in the name.	I changed the name of the column to “Dystopia Residual”.
Column-name “Whisker-low”	The column name is not same as in the other data sets.	I changed the name of the column to “Lower Confidence Interval”.
Column-name “Whisker-high”	The column name is not same as in the other data sets.	I changed the name of the column to “Upper Confidence Interval”.
No year Column	Since my goal was to create a dataset with all data from the different years, I needed a Year column.	I added a column “Year” to the dataframe, where every value is 2022.
No Region column	Since my goal was to create a dataset with regional data	I added a column “Region” to the dataframe, by first

	that could also be grouped in region, I needed to add a region column.	creating a subset with the region and country column from the data set 2015 and then merged this subset with the the data set 2022 with the key being the country column.

Data set 2023

Problem	Description	Solution
Column-name “Social support”	The column name “Social support” has a “s” instead of an “S”.	I changed the name of the column to “Social Support”
Column-name “Ladder Score”	The column name is not same as in the other data sets.	I changed the name of the column to “Happiness Score”.
Column-name “Perception of Corruption”	The column name is not same as in the other data sets.	I changed the name of the column to “Trust (Government Corruption)”.
Column-name “Logged GDP per capita”	The column name is not same as in the other data sets.	I changed the name of the column to “Economy (GDP per capita)”
Column-name “Country name”	The column name is not same as in the other data sets and only includes Countries.	I changed the name of the column to “Country”.
Column-name “Healthy life expectancy”	The column name is not same as in the other data sets.	I changed the name of the column to “Health (Life Expectancy”.
Column-name “Freedom to make life choices”	The column name is not same as in the other data sets.	I changed the name of the column to “Freedom”.
Column-name “Dystopia + Residual”	The column has a plus in the name.	I changed the name of the column to “Dystopia Residual”.
Column-name “Ladder Score in Dystopia”	The column name is not same as in the other data sets.	I changed the name of the column to “Dystopia Score”.
Column-name “Regional Indicator”	The column name is not same as in the other data sets.	I changed the name of the column to “Region”.

Column-name “lowerwhisker”	The column name is not same as in the other data sets.	I changed the name of the column to “Lower Confidence Interval”.
Column-name “upperwhisker”	The column name is not same as in the other data sets.	I changed the name of the column to “Upper Confidence Interval”.
No year Column	Since my goal was to create a dataset with all data from the different years, I needed a Year column.	I added a column “Year” to the dataframe, where every value is 2019.
No “Happiness Rank” Column	Since my goal was to create a dataset with all data from the different years that also includes the Rank-column, I needed to add the year column.	I added the column “Happiness Rank” to the dataframe by adding the “Happiness Rank” Column from the data set 2019. I didn’t need to create a subset, since I only needed the numbers from 1 to 157 and since the datset was already sorted based on the ranking I also didn’t need a key to establish this column.
No Region column	Since my goal was to create a dataset with regional data that could also be grouped in region, I needed to add a region column.	I added a column “Region” to the dataframe, by first creating a subset with the region and country column from the data set 2015 and then merged this subset with the the data set 2022 with the key being the country column.

Transformed Data set (2015-2023)

Problem	Description	Solution
Unnecessary columns	Since I want wo focus on the Happiness Score and the 6 Main Factors to impact the score, I don’t need the Dystopia related data.	I dropped all Dystopia related data. That’s includes the following columns: Unnamed:0, merge column,Explained by: Log GDP per capita, Explained by: Social support, Explained by: Healthy life expectancy, Explained by: Freedom to make life choices, Explained by: Generosity, Explained by: Perceptions of corruption, Explained by: GDP per capita, Distopia score and Dystopia Residual.

Wrong Data types	Some of the columns had the wrong data type. For example year was identified as “int64” or the “Lower confidence Interval” was identified as “object.	I changed the data type of the following columns: Year (to str) Lower Confidence Interval (to float) Upper Confidence Interval (to float) Happiness Rank (to str) Happiness Score (to float)
Missing values	Since some years had slightly other data for example no data for the Lower Confidence Interval or no Standard Error, I had some missing values. Especially the year 2022 was critical, since I didn’t had data for the main factors of the happiness score.	I decided to use interpolation to inpute my data. Since most data was missing from the year 2022 I sorted my data first by country and then by year to the observation for every country for 2022 was surrounded by the data from 2021 and 2023 for the individual data. In That way the data for 202 was filled with country specific data and with data based on near years
Duplicates	I found one duplicate for the Country Afghanistan for the year 2015.	I droped the duplicate.

List of questions to explore and hypotheses

Questions:

- How did the Happiness score change over the years?
- Which countries have the highest or lowest Happiness score?
- Have always the same countries the highest or lowest Happiness score?
- Which of the main factors has the highest impact on the Happiness score?
- Can we see the Corona-Impact in the Happiness Score in the years 2020-2022?

Hypotheses:

- The happiness score is globally lower for the years 2020 to 2022 (Corona Impact).
- The countries with the highest and lowest Happiness score stay mostly the same.
- The countries with the highest Happiness score are located in europe.