Data Source

Raw Data

- 2015.csv
- 2016.csv
- 2017.csv
- 2018.csv
- 2019.csv
- 2020.csv
- 2021.csv
- 2022.csv

Cleaned Data

- whr_merged.csv

Data Description

The World Happiness Record is external data. The World Happiness Report is not run by a single organization or individual. It is a collaborative effort involving several organizations and researchers from around the world. The report is typically edited by a group of leading experts in the field of happiness and well-being, and it is often published by the Sustainable Development Solutions Network (SDSN), which is a global initiative for the United Nations.

Variables included in this dataset are Country, Region, Happiness Rank, Happiness Score, Economy, Health, Freedom, Trust, Generosity, and Year.

The data provided spans from 2015 to 2022.

It looks at things like how people feel about their lives, their emotions, and factors like having friends and being able to make choices.

The report helps us understand what makes people in different countries happy and can be used to make policies to improve people's well-being.

Kaggle Link

Defining Questions to Explore

Clarifying question:

- Which region ranks highest, and conversely, which region ranks lowest?

Funneling question:

- Has there been a change in the overall trend?
- Is the economy a significant contributing factor to these rankings?

Clarifying question:

- How does Germany's happiness rank compare to that of other Western European countries?

Funneling question:

- What are the primary factors that contribute to Germany's happiness ranking?

- According to Germans, which happiness factors do they feel are lacking or need improvement?

Data Limitation

The World Happiness Report, which measures how happy people are in different countries, can have some biases that affect its accuracy.

Self-reporting bias: People might not always tell the truth about their happiness in surveys, or they might answer in a way they think is expected.
Sample bias: The people surveyed might not represent everyone in a country. Some groups might be left out.
Cultural bias: Different cultures see happiness differently, so it's hard to compare.

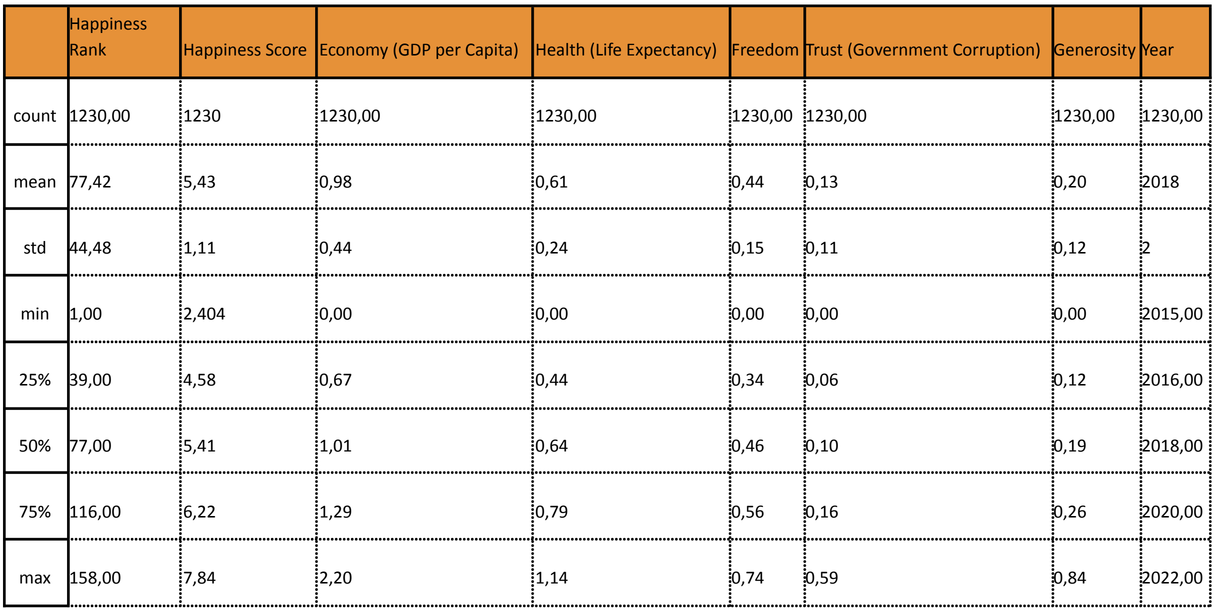
Additionally, economic and social conditions, as well as political and social stability might affect the overall result.

Data Profile

The dataset contains 1230 rows and 11 columns.

Data Types

Variables	Time-variant / -invariant	Structured / Unstructured	Qualitative / Quantitative	Qualitative: Nominal / Ordinal Quantitative: Discrete / Continuous
Country	Time Invariant	Structured	Qualitative	Nominal
Region	Time Invariant	Structured	Qualitative	Nominal
Happiness Rank	Time Invariant	Structured	Quantitative	Discrete
Happiness Score	Time Invariant	Structured	Quantitative	Discrete
Economy (GDP per Capita)	Time Invariant	Structured	Quantitative	Discrete
Health (Life Expectancy)	Time Invariant	Structured	Quantitative	Discrete
Freedom	Time Invariant	Structured	Quantitative	Discrete
Trust (Government Corruption)	Time Invariant	Structured	Quantitative	Discrete
Generosity	Time Invariant	Structured	Quantitative	Discrete
Year	Time variant	Structured	Quantitative	Discrete



	Happiness Rank	Happiness Score	Economy (GDP per Capita)	Health (Life Expectancy)	Freedom	Trust (Government Corruption)	Generosity	Year
count	1230,00	1230	1230,00	1230,00	1230,00	1230,00	1230,00	1230,00
mean	77,42	5,43	0,98	0,61	0,44	0,13	0,20	2018
std	44,48	1,11	0,44	0,24	0,15	0,11	0,12	2
min	1,00	2,404	0,00	0,00	0,00	0,00	0,00	2015,00
25%	39,00	4,58	0,67	0,44	0,34	0,06	0,12	2016,00
50%	77,00	5,41	1,01	0,64	0,46	0,10	0,19	2018,00
75%	116,00	6,22	1,29	0,79	0,56	0,16	0,26	2020,00
max	158,00	7,84	2,20	1,14	0,74	0,59	0,84	2022,00

Data cleaning and wrangling history

2015.csv

- Dropped column ‘Standard Error’
- Derived a new column ‘Year’ and filled it with 2015
- No missing values
- No duplicate values

2016.csv

- Dropped 'Lower Confidence Interval' and 'Upper Confidence Interval' columns
- Derived a new column ‘Year’ and filled it with 2016
- No missing values
- No duplicate values

2017.csv

- Renamed columns on 2017 dataset (*old* : *new*)
    - 'Happiness.Rank' : 'Happiness Rank'
    - 'Happiness.Score' : 'Happiness Score'
    - 'Economy..GDP.per.Capita.' : 'Economy (GDP per Capita)
    - 'Health..Life.Expectancy.' : 'Health (Life Expectancy)'
    - 'Trust..Government.Corruption.' : 'Trust (Government Corruption)'
    - 'Dystopia.Residual' : 'Dystopia Residual'
- Dropped ' Whisker.high' and 'Whisker.low' columns
- Derived a new column ‘Region’ and assigned them according to data from 2016.csv
- Derived a new column ‘Year’ and filled it with 2016

- No missing values
- No duplicate values

2018.csv

- Renamed columns on 2018 dataset (*old* : *new*)
    - 'Overall rank' : 'Happiness Rank'
    - 'Country or region' : 'Country'
    - 'Score' : 'Happiness Score'
    - 'Healthy life expectancy' : 'Health (Life Expectancy)
    - 'Perceptions of corruption' : 'Trust (Government Corruption)'
    - 'GDP per capita' : 'Economy (GDP per Capita)'
    - 'Freedom to make life choices' : 'Freedom'
- Derived region from region lists
- One missing value from ‘Trust (Government Corruption)’
  The Trust score was missing from a row which country is United Arab Emirates.
  I imputed the value from last year, 0.32449
- Derived a year column and filled with 2018
- No duplicate values

2019.csv

- Renamed columns on 2018 dataset (*old* : *new*)
    - 'Overall rank' : 'Happiness Rank'
    - 'Country or region' : 'Country'
    - 'Score' : 'Happiness Score'
    - 'Healthy life expectancy' : 'Health (Life Expectancy)
    - 'Perceptions of corruption' : 'Trust (Government Corruption)'
    - 'GDP per capita' : 'Economy (GDP per Capita)'
    - 'Freedom to make life choices' : 'Freedom'
- Derived region from region lists
- Derived a new column 'Year' and fill it with the value 2019
- No missing values
- No duplicate values

2020.csv

- Dropped the following columns
    - 'Standard error of ladder score'
    - 'upperwhisker'
    - 'lowerwhisker'
    - 'Ladder score in Dystopia' ,'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'
- Renamed columns on 2019 datasett (*old* : *new*)
    - 'Country name' : 'Country'
    - ‘Regional indicator' : 'Region

    - 'Ladder score' : 'Happiness Score'
    - 'Healthy life expectancy' : 'Health (Life Expectancy)'
    - 'Perceptions of corruption' : 'Trust (Government Corruption)'
    - 'Logged GDP per capita' : 'Economy (GDP per
    - 'Dystopia + residual' : 'Dystopia Residual'
    - 'Freedom to make life choices' : 'Freedom'}
- Derived a rank column based on 'Happiness Score'
- Derived a year column and filled with 2020
- No missing values
- No duplicate values

2021.csv

- Dropped the following columns
    - 'Standard error of ladder score'
    - 'upperwhisker
    - 'lowerwhisker'
    - 'Ladder score in Dystopia' ,'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'
- Renamed columns on 2019 datasett (*old* : *new*)
    - 'Country name' : 'Country'
    - ‘Regional indicator' : 'Region
    - 'Ladder score' : 'Happiness Score'
    - 'Healthy life expectancy' : 'Health (Life Expectancy)'
    - 'Perceptions of corruption' : 'Trust (Government Corruption)'
    - 'Logged GDP per capita' : 'Economy (GDP per
    - 'Dystopia + residual' : 'Dystopia Residual'
    - 'Freedom to make life choices' : 'Freedom'}
- Derived a rank column based on 'Happiness Score'
- Derived a year column and filled with 2021
- No missing values
- No duplicate values

2022.csv

- dropped the following columns
    - 'Whisker-high'
    - 'Whisker-low'
- Renamed the following columns (*old* : *new*)
    - {'RANK' : 'Happiness Rank'
    - 'Dystopia (1.83) + residual' : 'Dystopia Residual'
- Derived region from region lists
- Renamed 'Eswatini, Kingdom of' to 'Swaziland'
- Some countries contained an asterisk symbol so they were gotten rid of.

- One value was missing and found out that row 147 was a placeholder (everything was NaN) and therefore gotten rid of.
- No duplicate values
- Derived a year column and filled with 2022

whr_merged.csv

- All cleaned data was merged using python’s CONCAT function.
- The variables included in the dataset are
    - Country
    - Region
    - Happiness Rank
    - Happiness Score
    - Economy
    - Health
    - Freedom
    - Trust
    - Generosity
    - Year
- Country names are region names were altered for consistency