Final Project – World Happiness Report

My chosen theme is an extension of one of my previous projects. (ti4hful/DataVisualisation (github.com))

I used the databases of the World Happiness Report between 2015 and 2022.
[Repository for this project - ti4hful/FINAL (github.com)]

1. SRS documentation

1. Introduction

1.1 Purpose
My chosen theme is an extension of one of my previous projects. (ti4hful/DataVisualisation (github.com))

I used the databases of the World Happiness Report between 2015 and 2022.

During my analyses, I chose the 2015 and 2021 databases.

The SRS below includes details of the data analysis of the World Happiness 2015-2022 report.

1.2 Scope

The system is designed to analyze and display data from the World Happiness Report 2015 to 2022, Python for data management, Pandas for data analysis, SQL for data storage and Power BI for reporting.

2. Functional Requirements

2.1 Data Acquisition

2.1.1 Description

The system needs to retrieve the World Happiness Report data sets for the years 2015-2022.

2.1.2 Implementation

Language: Python
Python scripts will be used:
Pandas
Numpy
Matplotlib
Seaborn
Pyodbc
MySQL connector
SQL Alchemy
Mcolors

2.2 Data Storage

2.2.1 Description

The system must store the processed data in a relational database for efficient querying.

2.2.2 Implementation

SQL

2.3 Data Analysis

2.3.1 Description

The system must perform data analysis on the World Happiness Report datasets to derive meaningful insights.

2.3.2 Implementation

Using Pandas for data processing and analysis in Python.

2.4 Reporting

2.4.1 Description

The system must create interactive reports and visualizations based on the analyzed data.

2.4.2 Implementation

Power BI: Create Power BI dashboards to display key happiness metrics and trends.

3. Non-Functional Requirements

3.1 Performance

3.1.1 Description

The system must efficiently manage large data sets.

3.1.2 Implementation

Optimizing Python scripts and SQL queries for performance.

4. Constraints

4.1 Technical

4.1.1 Description

The system must be implemented using Python, Pandas, SQL, and Power BI.

4.2 Time

4.2.1 Description

The project schedule:

Project Selection

1. **Choose a Project**

World Happiness Report up to 2022 (kaggle.com)

2. **Project Proposal**

The aim of the project is to make a comparative analysis based on the World Happiness Reports, what changes took place between 2015-2022, mainly in the indicators of the top 15 countries. I have already examined what the data showed in 2019 within the framework of a small-scale project, but this database makes it possible to examine the change even within a single country, since 2015.

Data Gathering and Cleaning

Day 1: Data Source Selection

Day 1: Data Import and Database Setup

- Download the chosen dataset and set up a SQL database, document the data source, download location, and any necessary credentials.

Time required for completion: 6h

ER diagram

Day 2: Data Cleaning and Preprocessing

- Write python queries to clean the data, document the data cleaning process, highlighting issues and steps taken.

Data Analysis and Visualization

Day 2: Data Analysis Kick-off

- Begin your data analysis using Python. Document the Python libraries and tools you use.

Day 3: Dashboard Development

- Start creating the initial version of your Power BI or Tableau dashboard, document the design decisions and initial visualization components.

Time required for completion: 5h

Day 3: Exploratory Data Analysis (EDA)

- Conduct in-depth EDA to uncover trends, patterns, and insights in the data, document interesting findings and prepare to integrate them into your dashboard.

Time required for completion: 5h

Final Day: Project Submission

Day 4-5

Project Report Compilation & Presentation Preparation

Time required for completion: 6h

Visualizations and Interpretations

1. Bar chart **- >** To visualize the top 5 happiest countries
2. **Scatter plot - >** To explore the relationship between GDP and happiness score
3. **Correlation heatmap - >** To visualize the correlations between different factors
4. **Pairplot - >** To visualize relationships between multiple variables in this dataset
5. **Box plot - >** To provides information about the distribution of the happiness scores, including the median, quartiles, and potential outliers
6. **Bar chart - >** To visualize the generosity of different countries
7. **Histogram - >** to visualize the distribution of GDP
8. **Radar chart ->** to compare multiple factors for each country in a single chart
9. **Bubble plot - >** To visualize three variables simultaneously

These visualizations provide different insights into the World Happiness Report datasets, including relationships between factors, distributions of happiness scores, generosity rankings, and distributions of GDP per capita.

Relationships

Country Entity
Attributes: CountryID (Primary Key), CountryName
Relationships: One-to-Many with other entities (e.g., GDP, Life Expectancy)

Region Entity
Attributes: RegionID (Primary Key), RegionName
Relationships: One-to-Many with Country

Happiness Entity
Attributes: HappinessID (Primary Key), CountryID (Foreign Key), RegionID (Foreign Key), HappinessRank, HappinessScore, StandardError
Relationships: Many-to-One with Country and Region

Economy Entity
Attributes: EconomyID (Primary Key), CountryID (Foreign Key), GDPPerCapita
Relationships: Many-to-One with Country

Family Entity
Attributes: FamilyID (Primary Key), CountryID (Foreign Key), FamilyScore

Relationships: Many-to-One with Country
Health Entity
Attributes: HealthID (Primary Key), CountryID (Foreign Key), LifeExpectancy
Relationships: Many-to-One with Country

2. Data Import and Database Setup

- > World Happiness Report up to 2022 (kaggle.com)

3. ER diagram

- > Draw.io
[FINAL/WHR_ER.drawio.png at main · ti4hful/FINAL (github.com)]

4. Data Cleaning and Preprocessing

w Python, using Pandas, mysql.connector and sqlalchemy
[full code : FINAL/WHR at main · ti4hful/FINAL (github.com)]

- Imported Libraries
- Create Database Engine
- Read CSV Files into DataFrames
- Write DataFrames to MySQL Tables
- Print the Shape of the DataFrames, with the first 4 rows
- Prints the shapes of multiple DataFrames, to provide informations about the number of rows and columns, in each of the datasets for the years
- Adding a new column 'Year' to each DataFrames
- Explore column names, and converting them into a Python list
- **In y2015 Column Renaming**
  'Family' to 'Family (Social Support)'
- **In y2016 Column Renaming**
  renamed the column:
  'Family' to 'Family (Social Support)'
- **In y2017 Column Renaming**
  'Happiness.Rank' to 'Happiness Rank'
  'Happiness.Score' to 'Happiness Score'
  'Economy..GDP.per.Capita.' to 'Economy (GDP per Capita)'
  'Family' to 'Family (Social Support)'
  'Health..Life.Expectancy.' to 'Health (Life Expectancy)'
  'Trust..Government.Corruption.' to 'Trust (Government Corruption)'
  Left Merge
  between the data_2017 df and a subset of columns from the data_2015 df, specifically the "Country" and "Region" columns. And it's filling any missing values in the "Region" column with a hyphen ("-").

- **In y2018 Column Renaming**
  'Overall rank' to 'Happiness Rank'
  'Country or region' to 'Country'
  'Score' to 'Happiness Score'
  'GDP per capita' to 'Economy (GDP per Capita)'
  'Social support' to 'Family (Social Support)'
  'Healthy life expectancy' to 'Health (Life Expectancy)'
  'Freedom to make life choices' to 'Freedom'
  'Perceptions of corruption' to 'Trust (Government Corruption)'
- **In y2019 Column Renaming**
  'Overall rank' to 'Happiness Rank'
  'Country or region' to 'Country'
  'Score' to 'Happiness Score'
  'GDP per capita' to 'Economy (GDP per Capita)'
  'Social support' to 'Family (Social Support)'
  'Healthy life expectancy' to 'Health (Life Expectancy)'
  'Freedom to make life choices' to 'Freedom'
  'Perceptions of corruption' to 'Trust (Government Corruption)'
  Left merge and fill missing values
- **In y2020 Column Renaming**
  'Country name' is renamed to 'Country'
  'Regional indicator' is renamed to 'Region'
  'Ladder score' is renamed to 'Happiness Score'
  'Explained by: Social support' is renamed to 'Family (Social Support)'
  'Explained by: Healthy life expectancy' is renamed to 'Health (Life Expectancy)'
  'Explained by: Freedom to make life choices' is renamed to 'Freedom'
  'Explained by: Perceptions of corruption' is renamed to 'Trust (Government Corruption)'
  'Explained by: Log GDP per capita' is renamed to 'Economy (GDP per Capita)'
  'Explained by: Generosity' is renamed to 'Generosity'
- **In y2021 Column Renaming**
  'Country name' is renamed to 'Country'
  'Regional indicator' is renamed to 'Region'
  'Ladder score' is renamed to 'Happiness Score'
  'Explained by: Social support' is renamed to 'Family (Social Support)'
  'Explained by: Healthy life expectancy' is renamed to 'Health (Life Expectancy)'
  'Explained by: Freedom to make life choices' is renamed to 'Freedom'
  'Explained by: Perceptions of corruption' is renamed to 'Trust (Government Corruption)'
  'Explained by: Log GDP per capita' is renamed to 'Economy (GDP per Capita)'
  'Explained by: Generosity' is renamed to 'Generosity'
  Remove duplicates and ensure that only the last occurrence of each duplicated column is retained.
  Adding a new column named 'Happiness Rank' to the df data_2021
- **In y2022**
  Merging the df data_2022 with a subset of the data_2015 df that includes the columns "Country" and "Region" based on the common column "Country."
  Fills any missing values in the "Region" column with a hyphen ("-")
  Column Renaming
  'RANK' is renamed to 'Happiness Rank'
  'Happiness score' is renamed to 'Happiness Score'
  'Explained by: GDP per capita' is renamed to 'Economy (GDP per Capita)'
  'Explained by: Social support' is renamed to 'Family (Social Support)'

'Explained by: Healthy life expectancy' is renamed to 'Health (Life Expectancy)'
'Explained by: Freedom to make life choices' is renamed to 'Freedom'
'Explained by: Generosity' is renamed to 'Generosity'
'Explained by: Perceptions of corruption' is renamed to 'Trust (Government Corruption)'

- **Retrieved the list of column names in each of the dfs**
- **Extracted the columns from each DF**
- **Find the common columns**
- **Converted the set of common columns to a list**
- **Created a list what is contains subsets of df-s for the years 15-22, this list contains only the columns specified in the 'common_cols'**
- **Created an empty DF**
- **Stack vertically the df-s from the dfs list into a single df**
- **Retrieved the dimensions**
- **'dropna' method to remove any rows containing missing values**
- **Check dimensions again**
- **Save the cleaned DF to a CSV**

5. Data Analysis and Visualization

w Matplotlib, Pandas, Numpy, Matplotlib.colors, Seaborn,

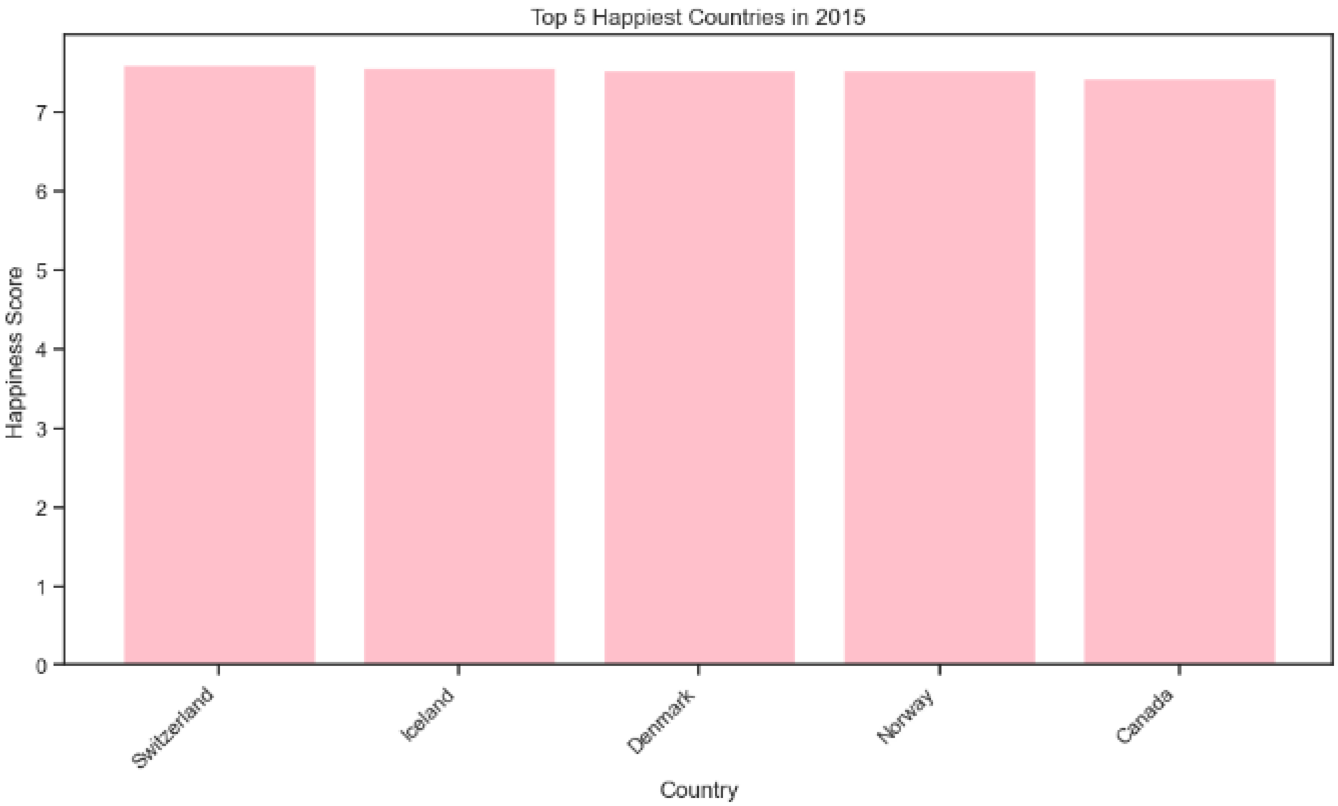
- **# Import Libraries**

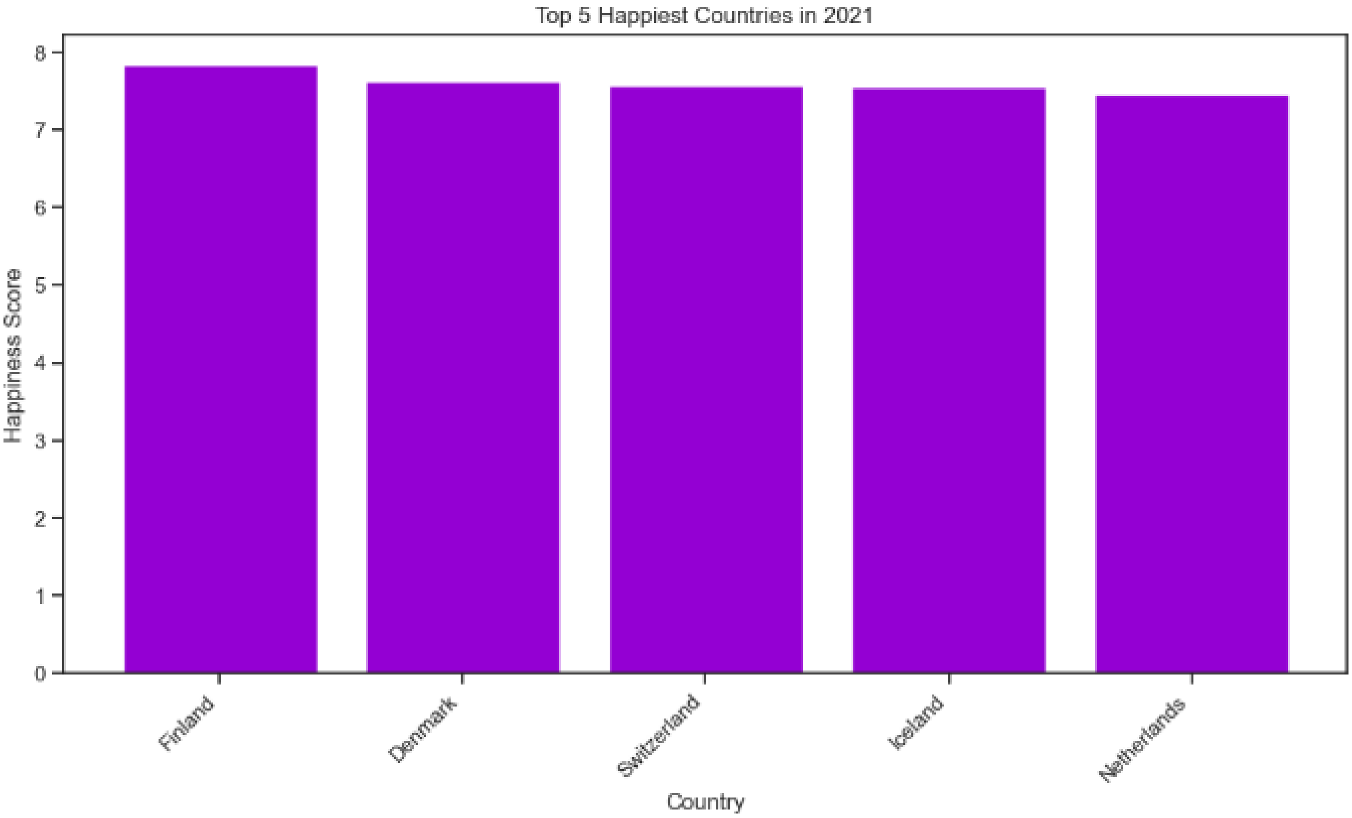
```
import matplotlib.pyplot as plt
import pandas as pd
import matplotlib.colors as mcolors
import numpy as np
import seaborn as sns
```

- **# Load the cleaned CSV file into a new DataFrame**
- **# Print the data types of columns in the DataFrame**

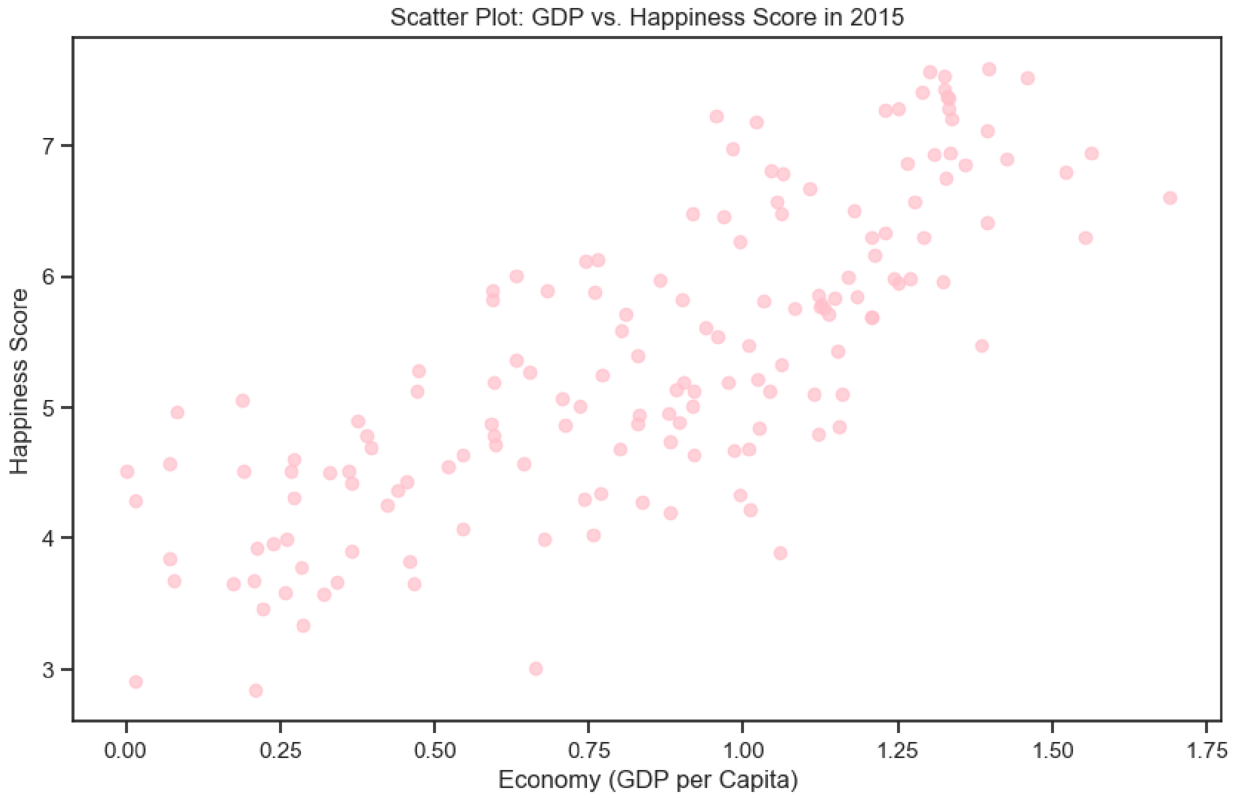
5.1. Visualizations

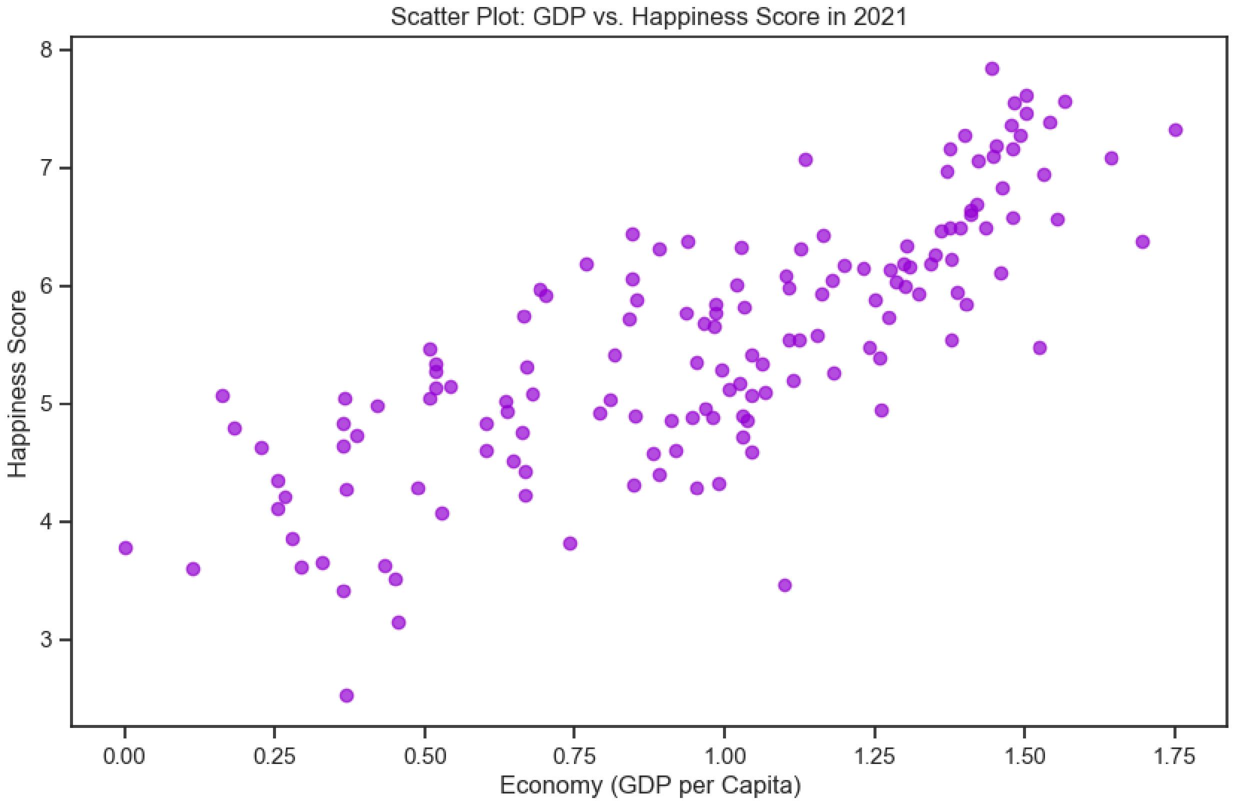
5.1.1. Bar chart - > To visualize the top 5 happiest countries in 2015 and 2021



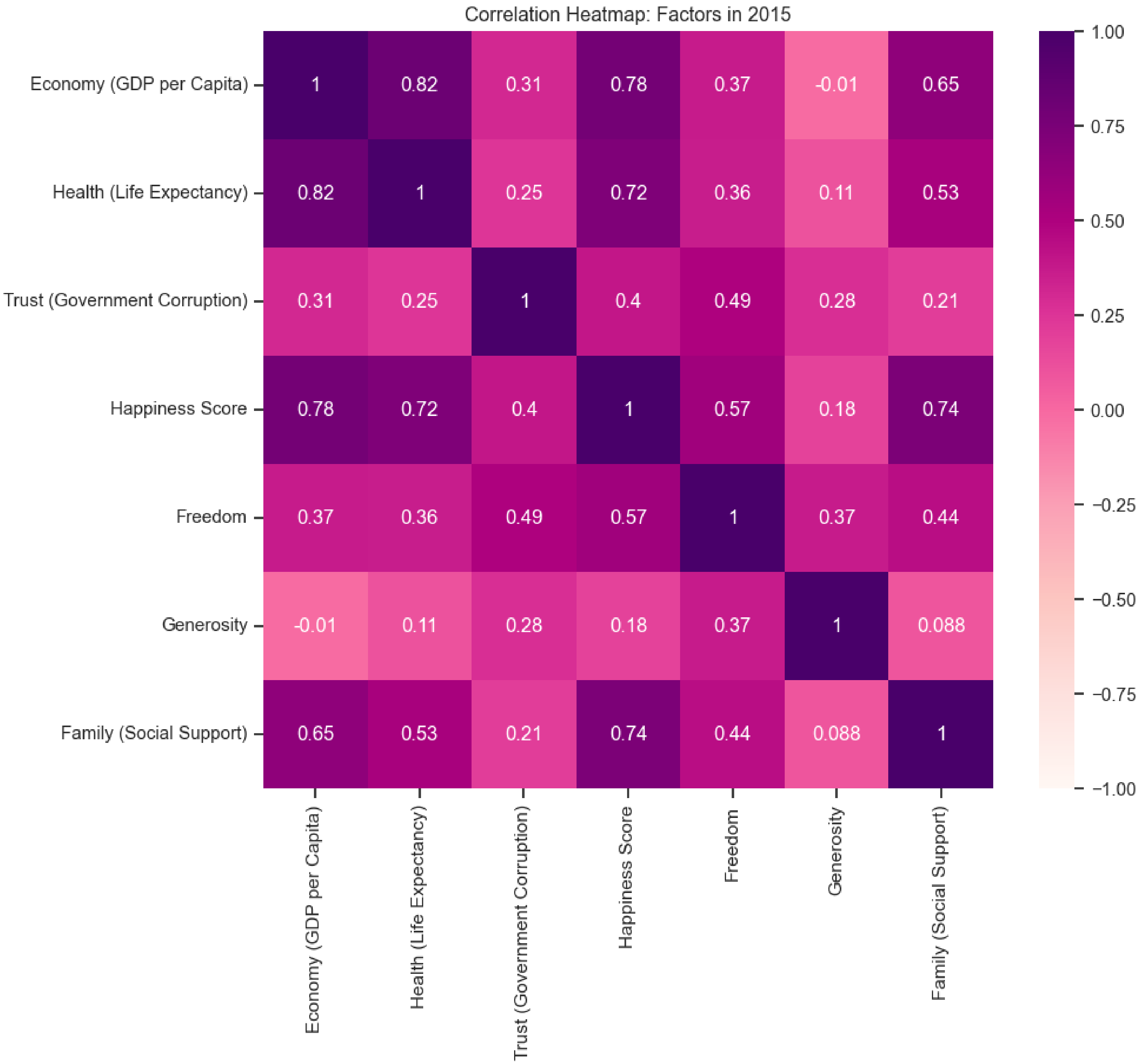


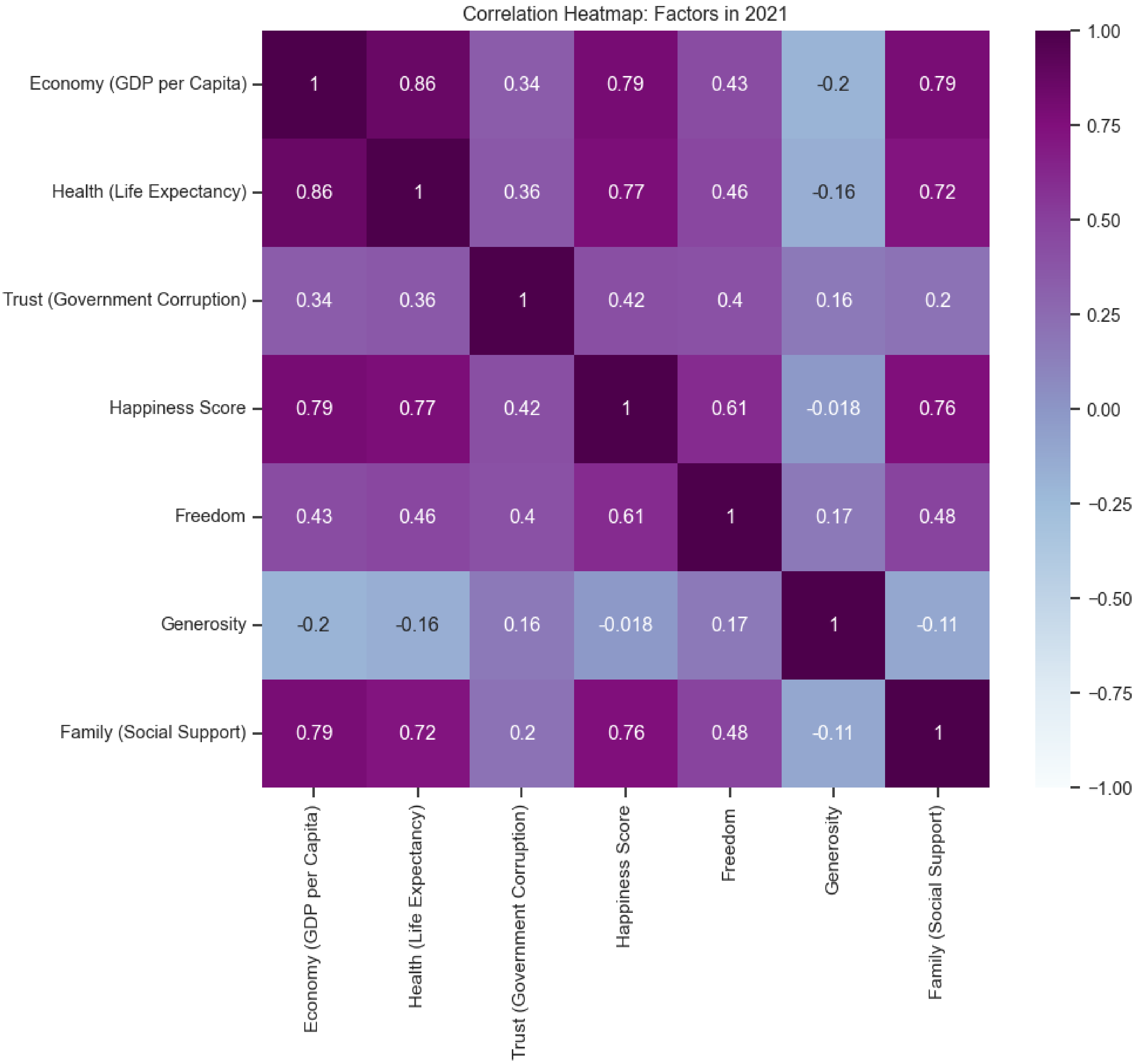
5.1.2. Scatter plot - > To explore the relationship between GDP and happiness



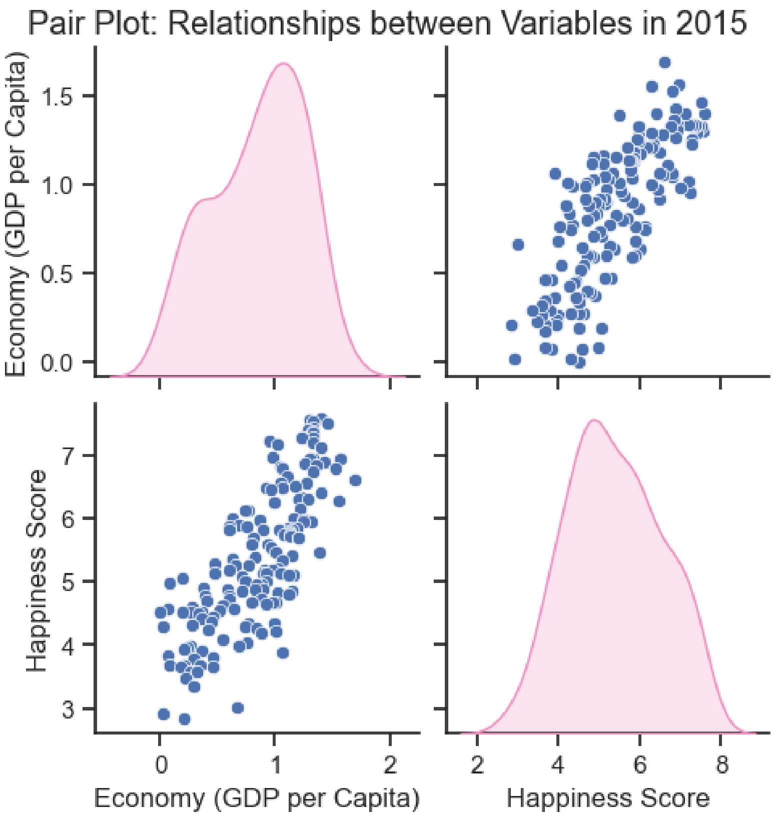


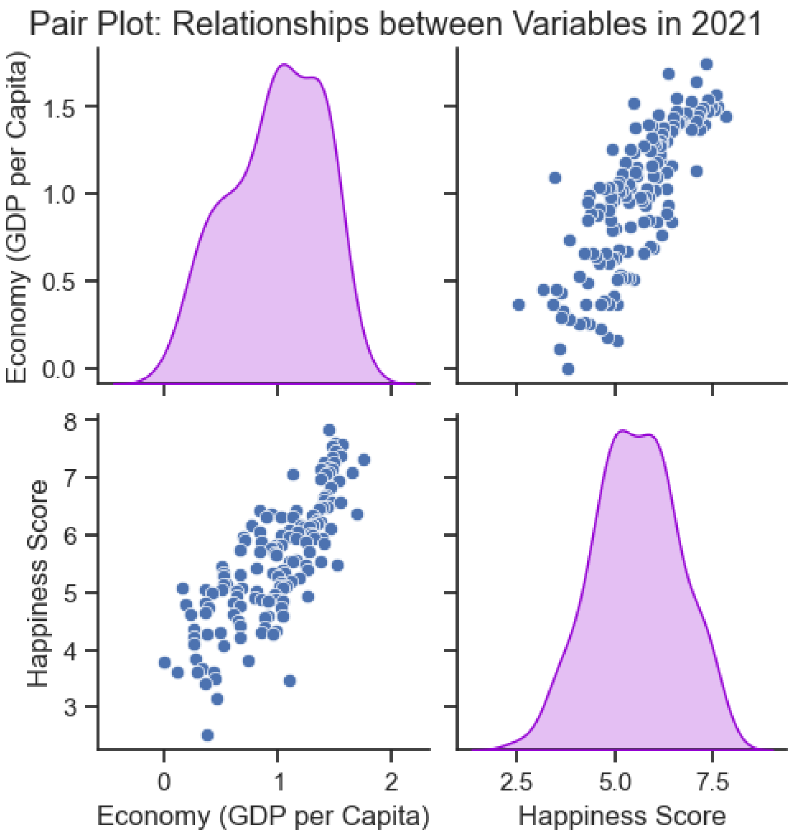
5.1.3. Correlation heatmap - > To visualize the correlations between different factors



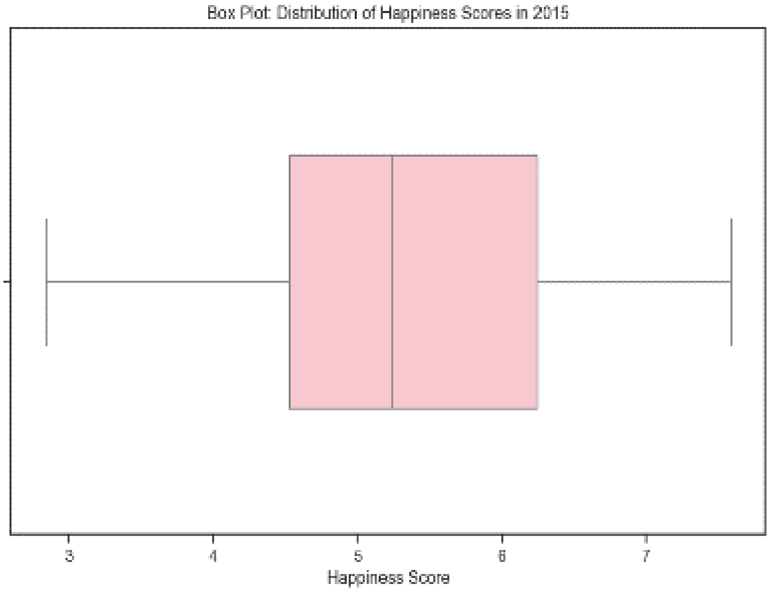

Correlation Heatmap: Factors in 2021
Economy (GDP per Capita)
1
0.86
0.34
0.79
0.43
-0.2
0.79
Health (Life Expectancy)
0.86
1
0.36
0.77
0.46
-0.16
0.72
Trust (Government Corruption)
0.34
0.36
1
0.42
0.4
0.16
0.2
Happiness Score
0.79
0.77
0.42
1
0.61
-0.018
0.76
Freedom
0.43
0.46
0.4
0.61
1
0.17
0.48
Generosity
-0.2
-0.16
0.16
-0.018
0.17
1
-0.11
Family (Social Support)
0.79
0.72
0.2
0.76
0.48
-0.11
1
Economy (GDP per Capita)
Health (Life Expectancy)
Trust (Government Corruption)
Happiness Score
Freedom
Generosity
Family (Social Support)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

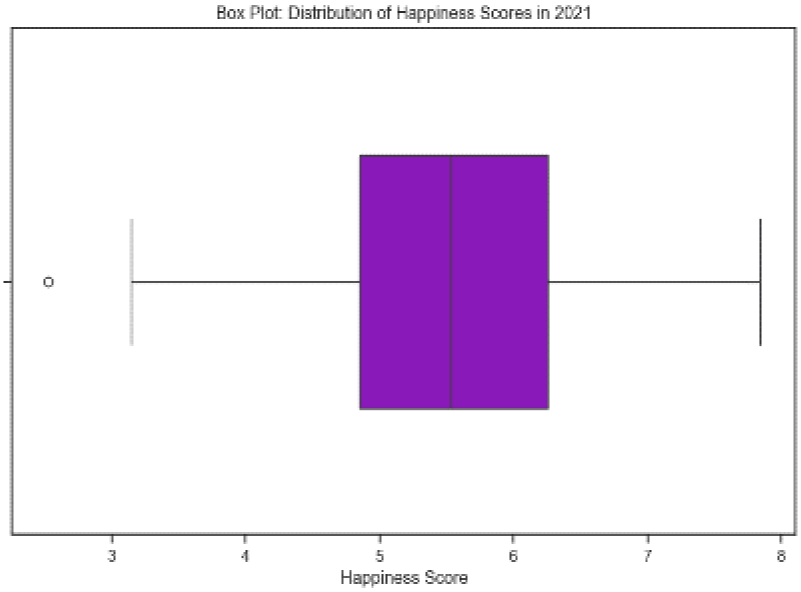
5.1.4. PairPlot - > To visualize relationships between multiple variables in this dataset


Pair Plot: Relationships between Variables in 2015

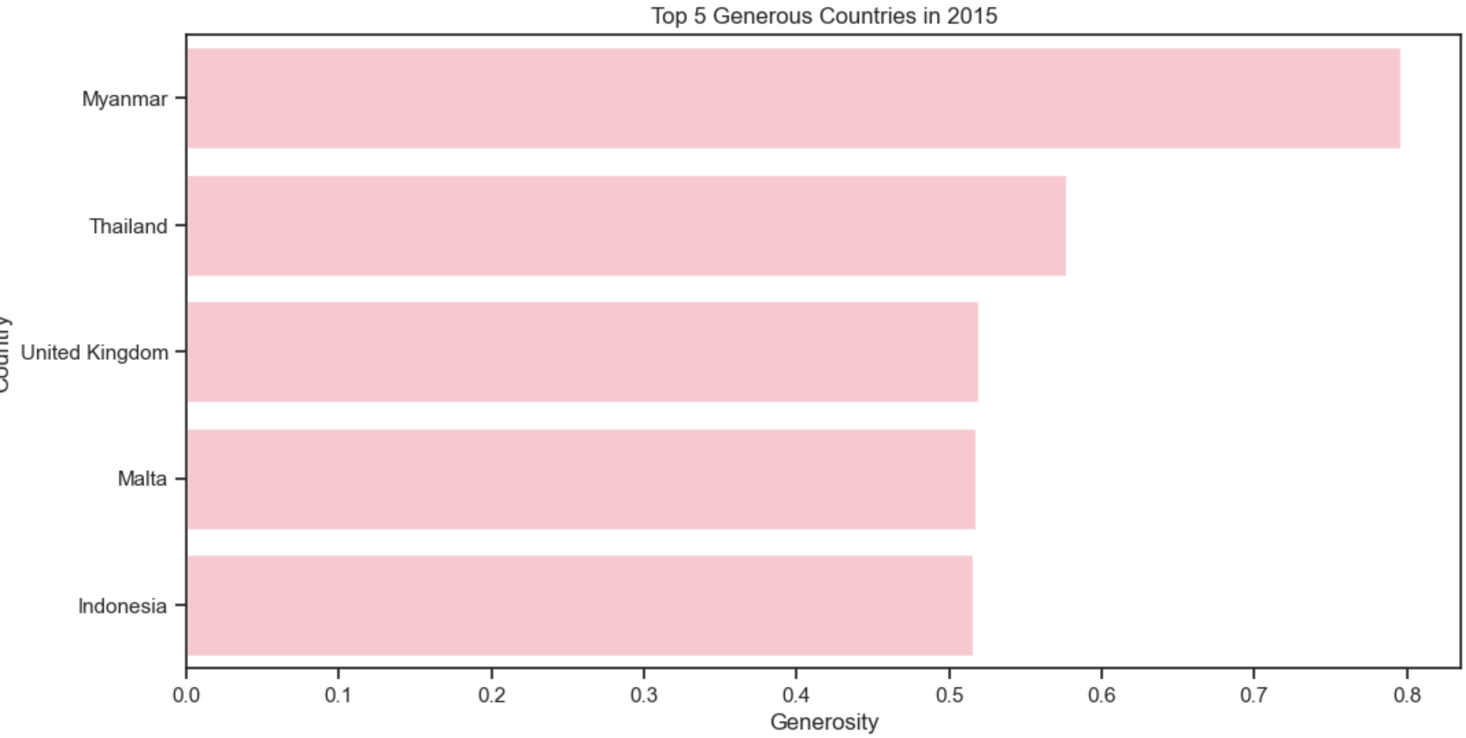

Pair Plot: Relationships between Variables in 2021

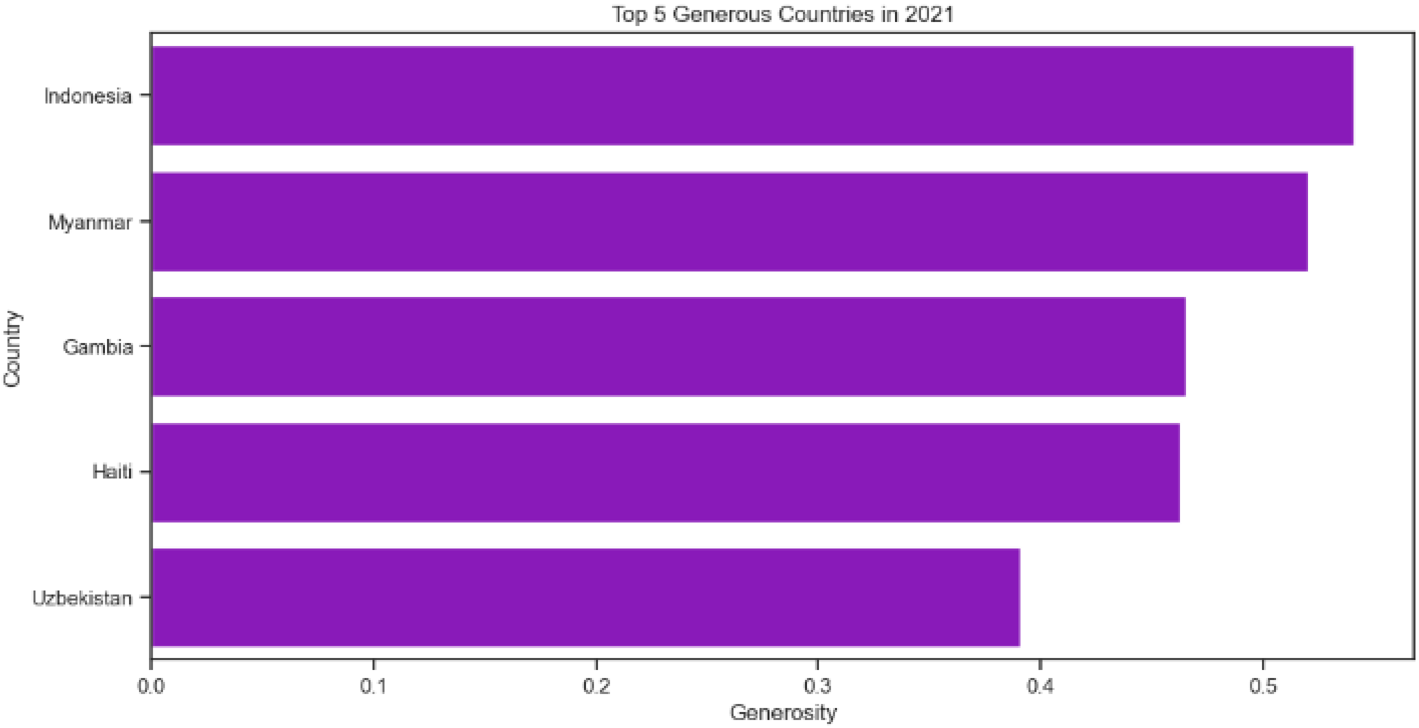
5.1.5. Box plot - > To provides information about the distribution of the happiness scores, including the median, quartiles, and potential outliers


Box Plot: Distribution of Happiness Scores in 2015

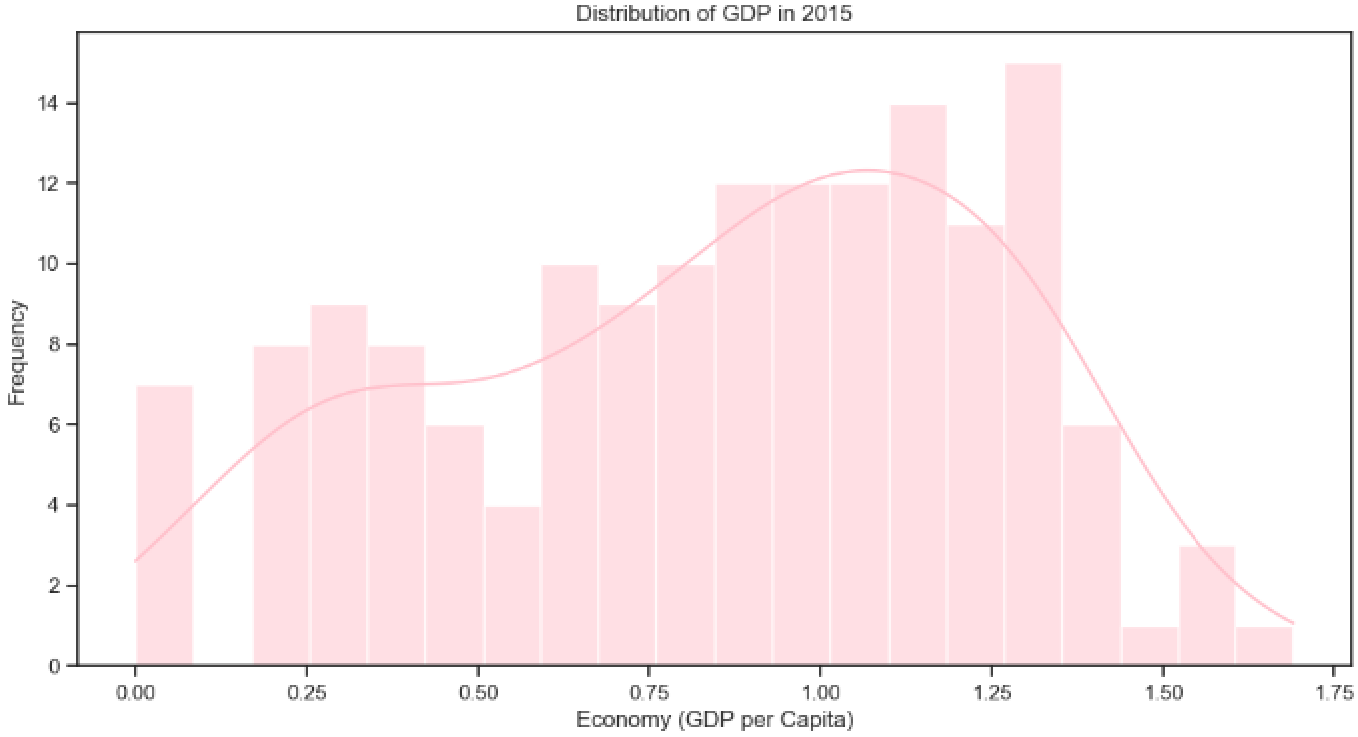

Box Plot: Distribution of Happiness Scores in 2021

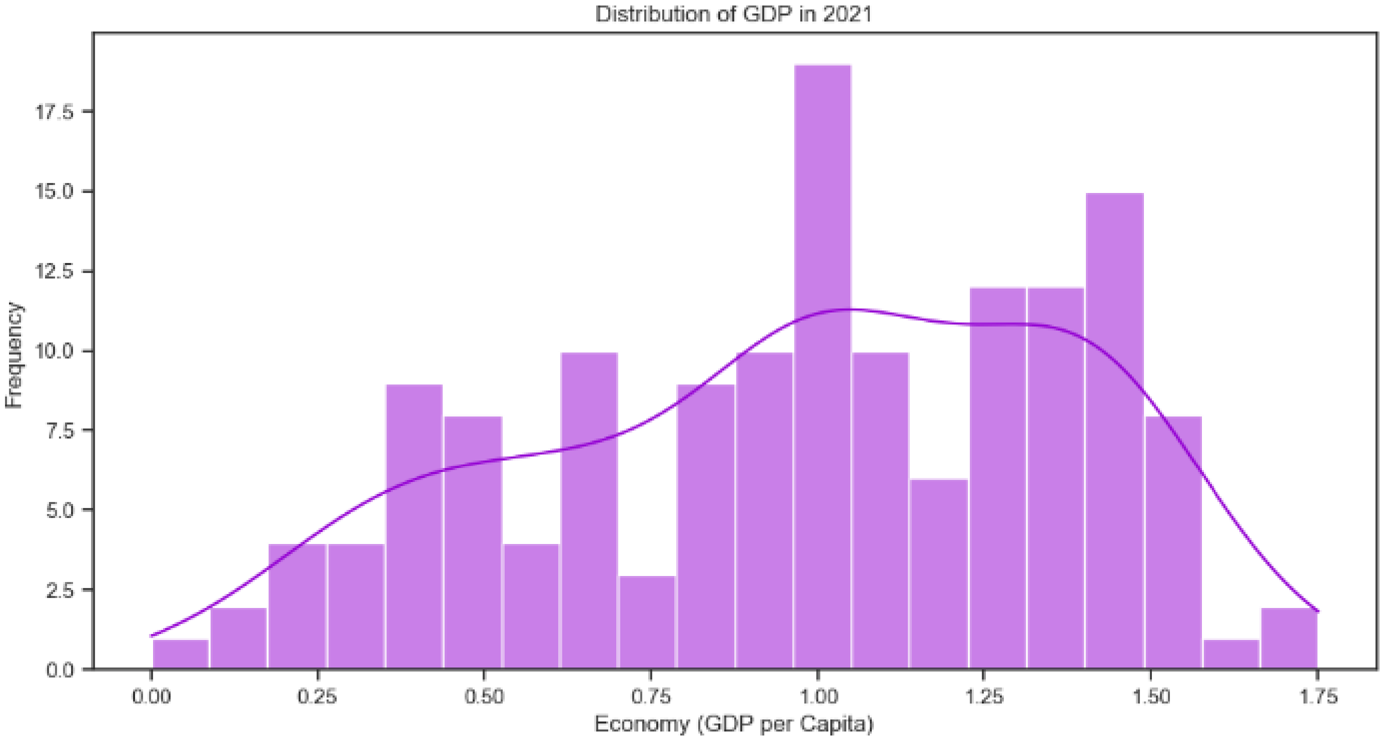
5.1.6. Bar chart - > To visualize the generosity of different countries



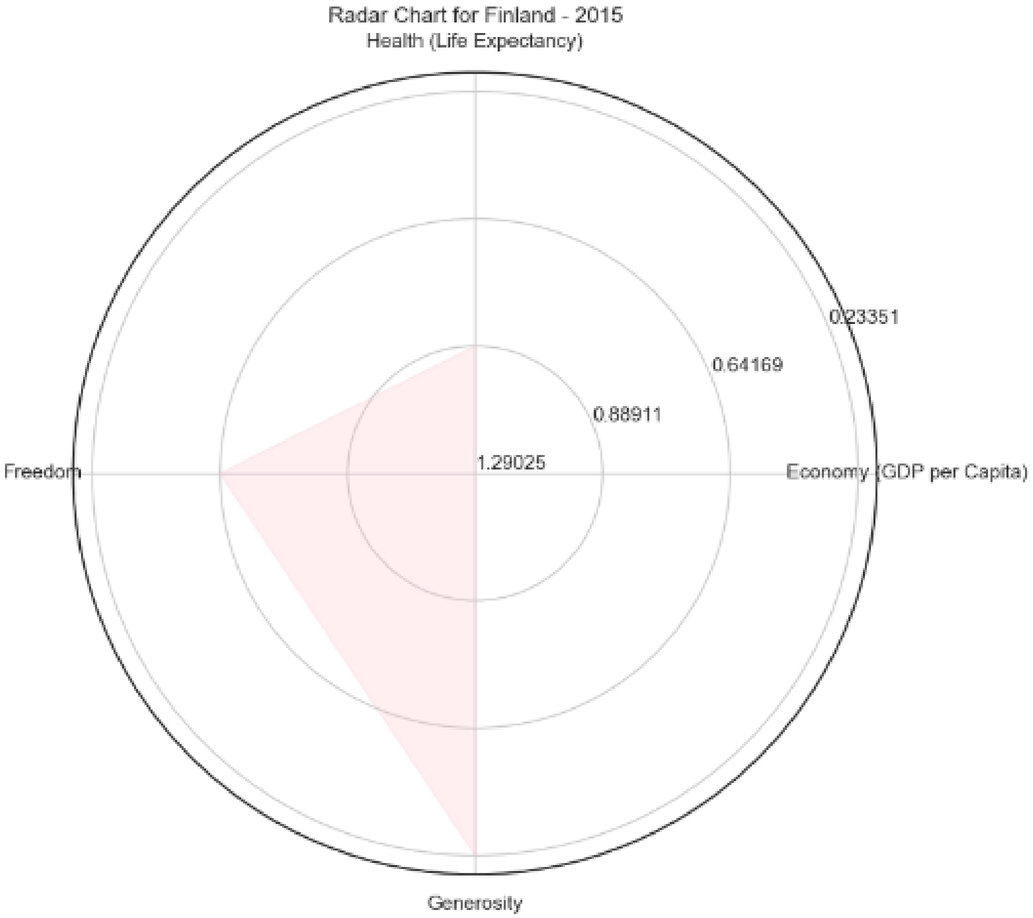


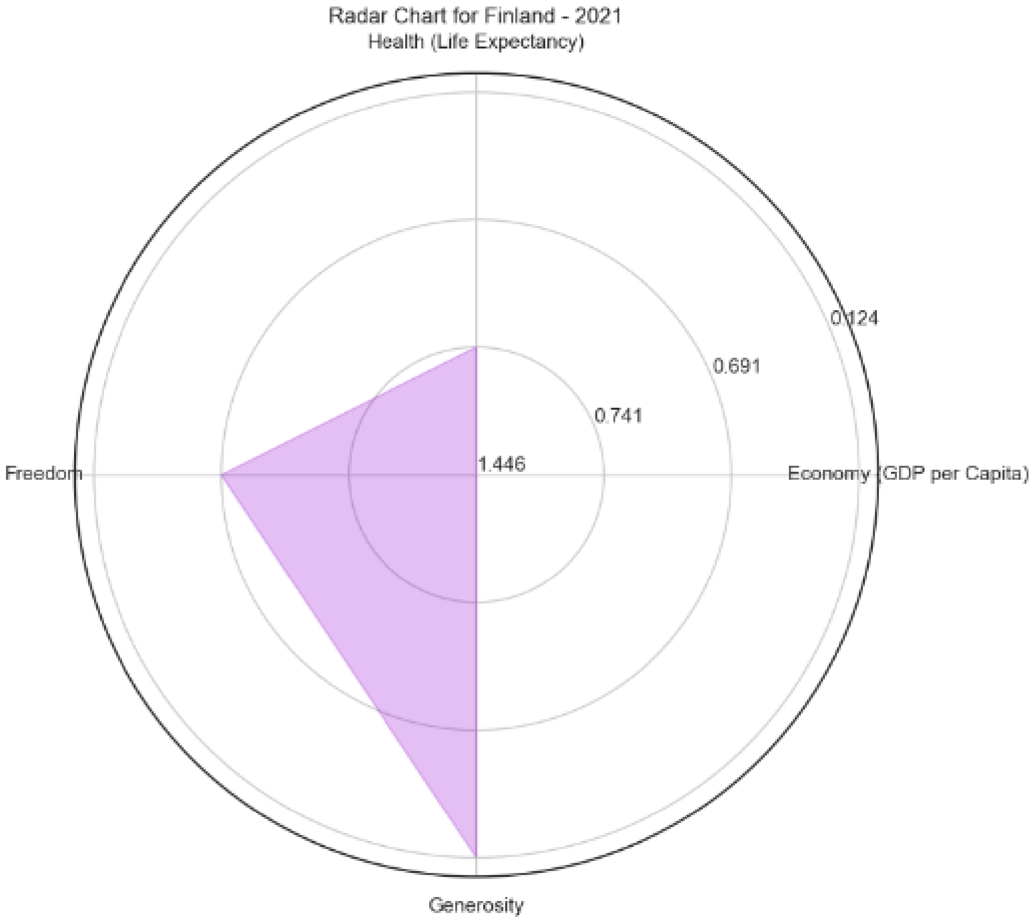
5.1.7. Histogram - > To visualize the distribution of GDP



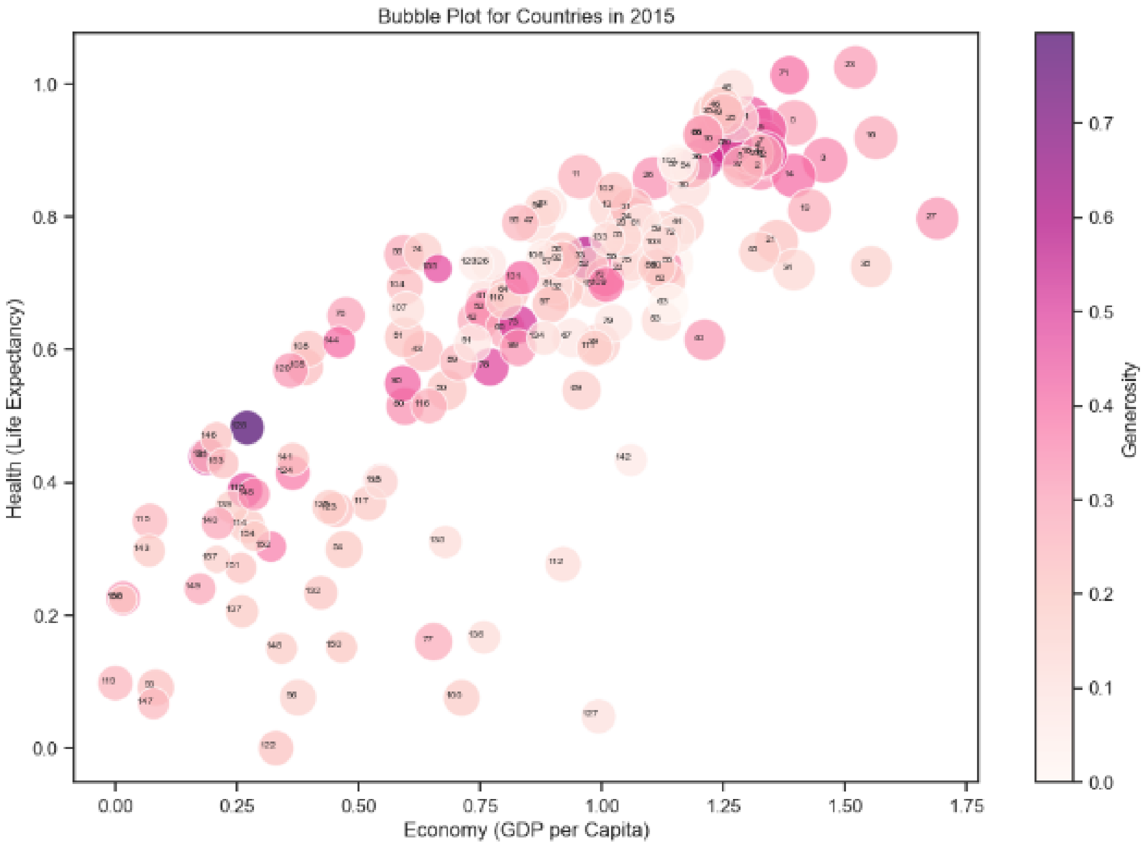


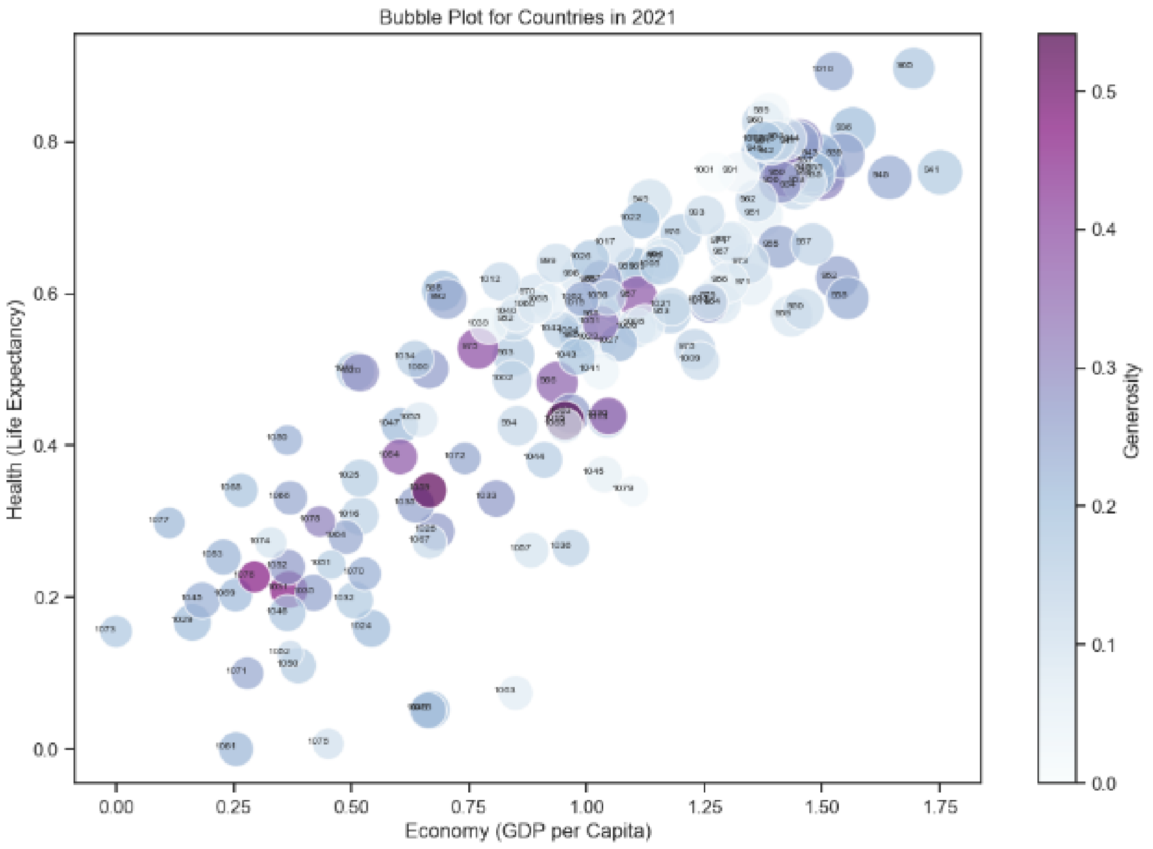
5.1.8. Radar chart - > To compare multiple factors for Finland in a single chart





5.1.9. Bubble plot - > To visualize three variables simultaneously





6. Dashboard Development

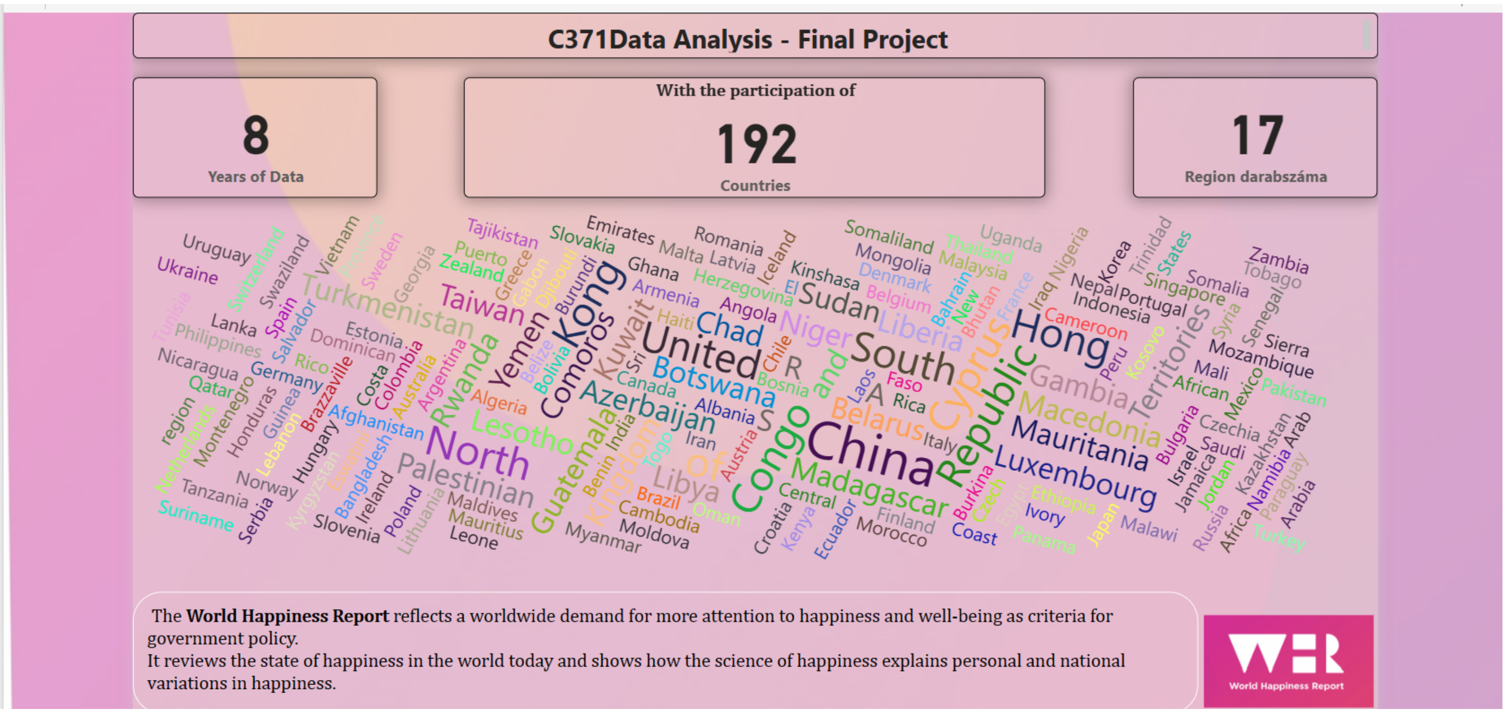
6.1. Power BI

[Power BI | FINAL/whr.pbix at main · ti4hful/FINAL (github.com)]

# Connect to Data
# Load Data
# Create a New Report Page
# Drag and Drop Fields
# Sort and Limit Data
# Data Filters
# Format and Customize
# 'Get data' w MySQL, load whr cleaned_data, 2015-2022
# From AppSource, search for extensions:
WordCloud, Zebra BI Charts, Zebra BI Tables 6.6, Drilldown Choropleth, Table Heatmap 3.5.0, Multi Info Cards, Timeline Storyteller ver.2.0.5, KPI Grid by MAQ Software, Venn Diagram by MAQ Software, Rotating Chart by MAQ Software, Rotating Tile by MAQ Software, Globe Data Bars, Horizontal Bullet Chart with Label
# Created a custom theme, with background, text format, borders, shadows, font, WHR logo and style creds [Home | The World Happiness Report]

6.1.1. WorldHappinessReport

- Text box
- Cards
- WordCloud
- WHR Logo



C371Data Analysis - Final Project

8
Years of Data

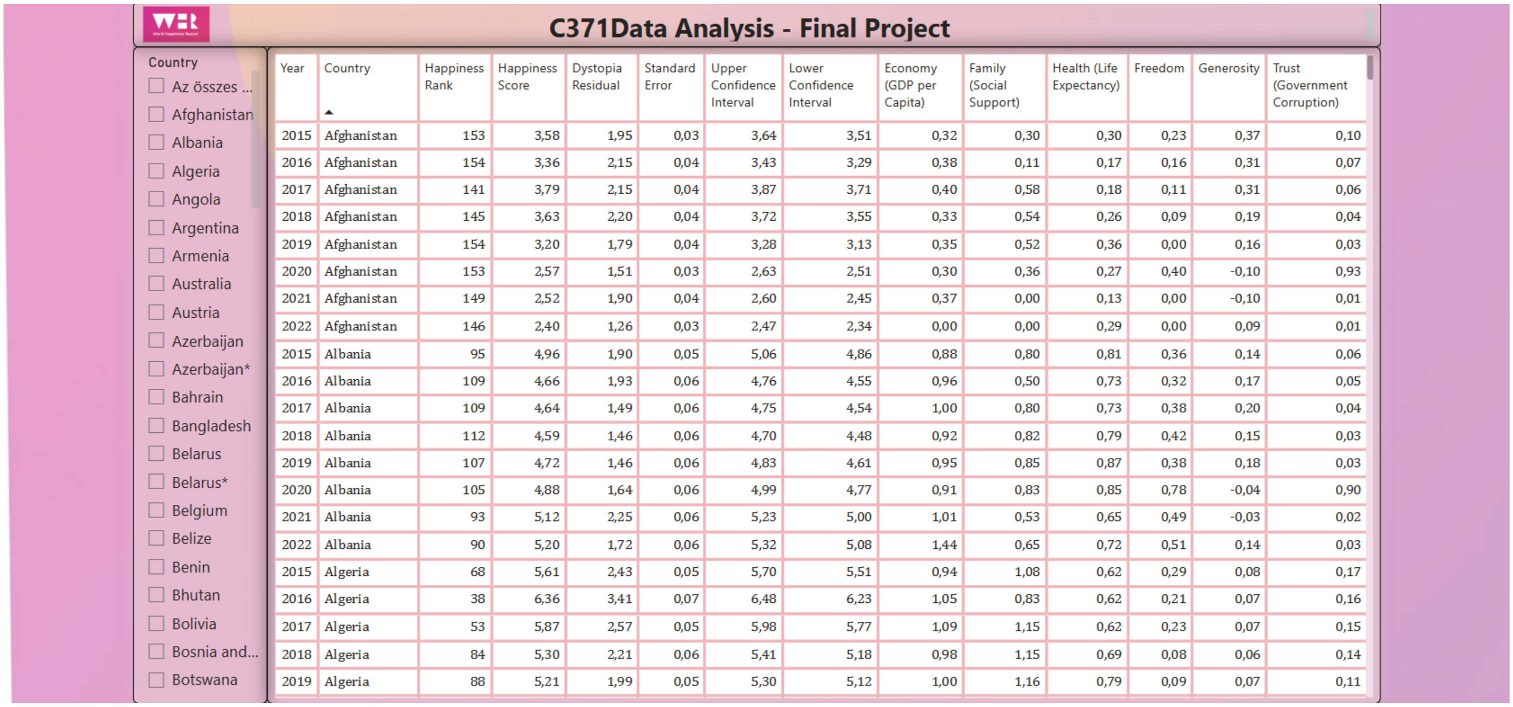
With the participation of
192
Countries

17
Region darabszáma

The **World Happiness Report** reflects a worldwide demand for more attention to happiness and well-being as criteria for government policy.
It reviews the state of happiness in the world today and shows how the science of happiness explains personal and national variations in happiness.

6.1.2. Summarized

- Text box
- Slicer
- Table



C371Data Analysis - Final Project

Country

- □ Az összes ...
- □ Afghanistan
- □ Albania
- □ Algeria
- □ Angola
- □ Argentina
- □ Armenia
- □ Australia
- □ Austria
- □ Azerbaijan
- □ Azerbaijan*
- □ Bahrain
- □ Bangladesh
- □ Belarus
- □ Belarus*
- □ Belgium
- □ Belize
- □ Benin
- □ Bhutan
- □ Bolivia
- □ Bosnia and...
- □ Botswana

Year	Country ▲	Happiness Rank	Happiness Score	Dystopia Residual	Standard Error	Upper Confidence Interval	Lower Confidence Interval	Economy (GDP per Capita)	Family (Social Support)	Health (Life Expectancy)	Freedom	Generosity	Trust (Government Corruption)
2015	Afghanistan	153	3,58	1,95	0,03	3,64	3,51	0,32	0,30	0,30	0,23	0,37	0,10
2016	Afghanistan	154	3,36	2,15	0,04	3,43	3,29	0,38	0,11	0,17	0,16	0,31	0,07
2017	Afghanistan	141	3,79	2,15	0,04	3,87	3,71	0,40	0,58	0,18	0,11	0,31	0,06
2018	Afghanistan	145	3,63	2,20	0,04	3,72	3,55	0,33	0,54	0,26	0,09	0,19	0,04
2019	Afghanistan	154	3,20	1,79	0,04	3,28	3,13	0,35	0,52	0,36	0,00	0,16	0,03
2020	Afghanistan	153	2,57	1,51	0,03	2,63	2,51	0,30	0,36	0,27	0,40	-0,10	0,93
2021	Afghanistan	149	2,52	1,90	0,04	2,60	2,45	0,37	0,00	0,13	0,00	-0,10	0,01
2022	Afghanistan	146	2,40	1,26	0,03	2,47	2,34	0,00	0,00	0,29	0,00	0,09	0,01
2015	Albania	95	4,96	1,90	0,05	5,06	4,86	0,88	0,80	0,81	0,36	0,14	0,06
2016	Albania	109	4,66	1,93	0,06	4,76	4,55	0,96	0,50	0,73	0,32	0,17	0,05
2017	Albania	109	4,64	1,49	0,06	4,75	4,54	1,00	0,80	0,73	0,38	0,20	0,04
2018	Albania	112	4,59	1,46	0,06	4,70	4,48	0,92	0,82	0,79	0,42	0,15	0,03
2019	Albania	107	4,72	1,46	0,06	4,83	4,61	0,95	0,85	0,87	0,38	0,18	0,03
2020	Albania	105	4,88	1,64	0,06	4,99	4,77	0,91	0,83	0,85	0,78	-0,04	0,90
2021	Albania	93	5,12	2,25	0,06	5,23	5,00	1,01	0,53	0,65	0,49	-0,03	0,02
2022	Albania	90	5,20	1,72	0,06	5,32	5,08	1,44	0,65	0,72	0,51	0,14	0,03
2015	Algeria	68	5,61	2,43	0,05	5,70	5,51	0,94	1,08	0,62	0,29	0,08	0,17
2016	Algeria	38	6,36	3,41	0,07	6,48	6,23	1,05	0,83	0,62	0,21	0,07	0,16
2017	Algeria	53	5,87	2,57	0,05	5,98	5,77	1,09	1,15	0,62	0,23	0,07	0,15
2018	Algeria	84	5,30	2,21	0,06	5,41	5,18	0,98	1,15	0,69	0,08	0,06	0,14
2019	Algeria	88	5,21	1,99	0,05	5,30	5,12	1,00	1,16	0,79	0,09	0,07	0,11

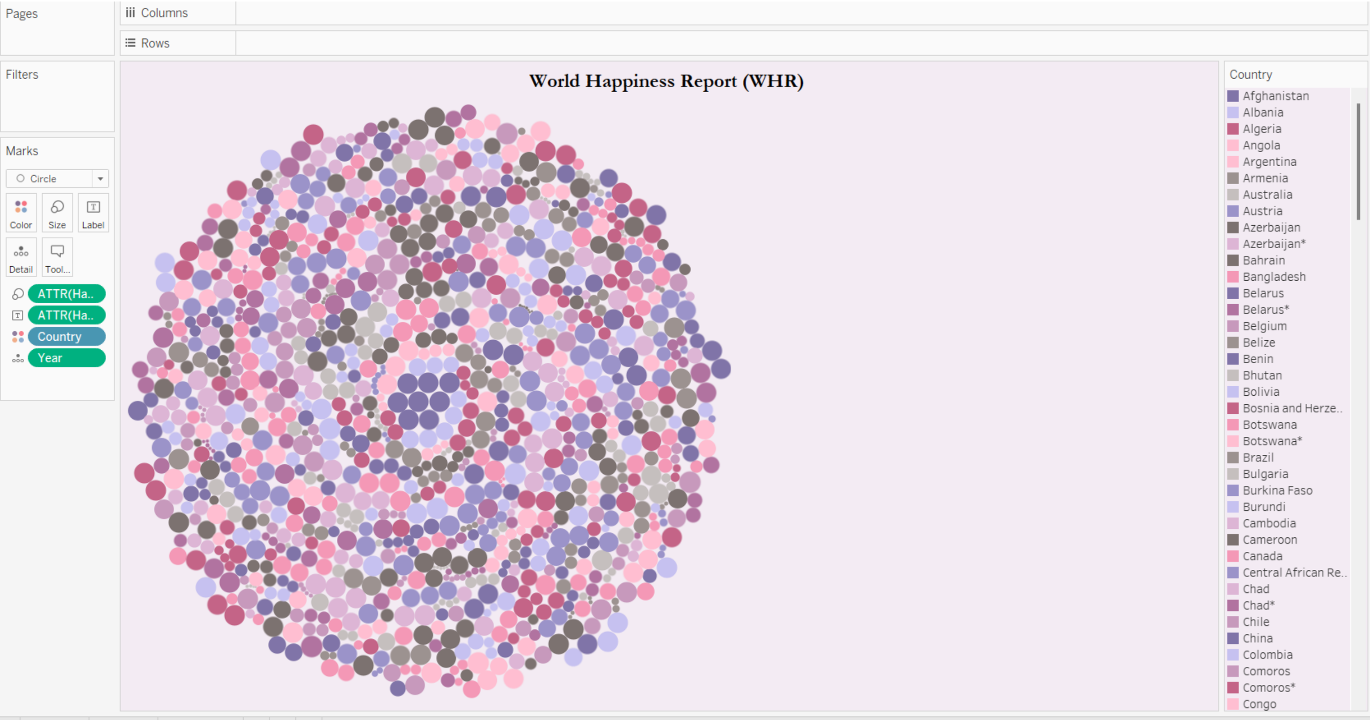
6.2. Tableau

[WorldHappinessReport_c371 | Tableau Public |

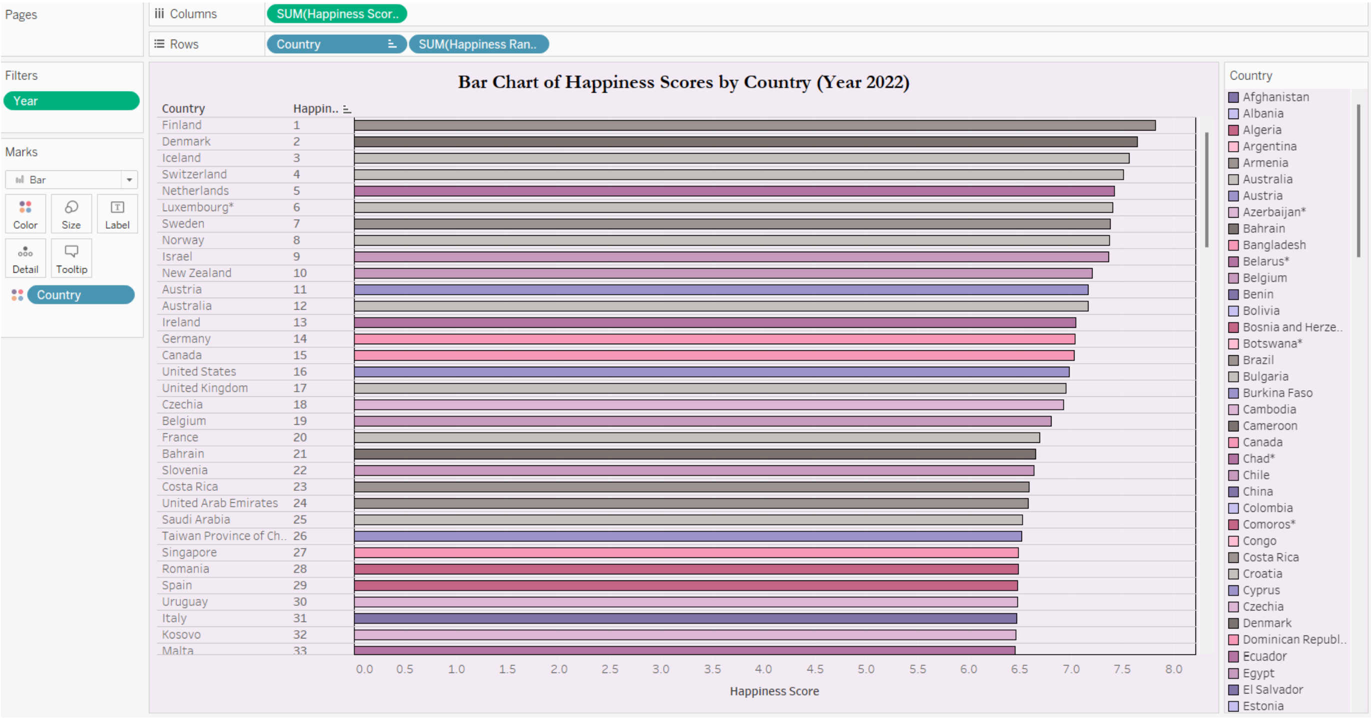
FINAL/WorldHappinessReport_c371.twbx at main · ti4hful/FINAL (github.com)]

# Connect to Data
# Create a New Worksheet
# Drag and Drop Fields
# Sort and Limit Data
# Color / Tooltip
# Filter Data

6.2.1. WHR

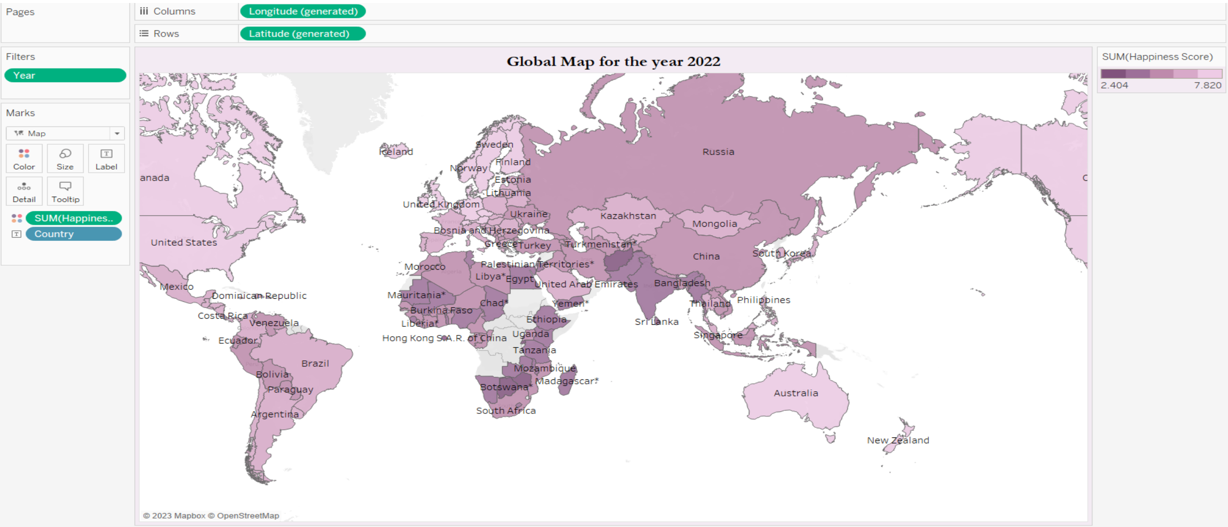


6.2.2. 2022_HS



Country	Happin..
Finland	1
Denmark	2
Iceland	3
Switzerland	4
Netherlands	5
Luxembourg*	6
Sweden	7
Norway	8
Israel	9
New Zealand	10
Austria	11
Australia	12
Ireland	13
Germany	14
Canada	15
United States	16
United Kingdom	17
Czechia	18
Belgium	19
France	20
Bahrain	21
Slovenia	22
Costa Rica	23
United Arab Emirates	24
Saudi Arabia	25
Taiwan Province of Ch..	26
Singapore	27
Romania	28
Spain	29
Uruguay	30
Italy	31
Kosovo	32
Malta	33

6.2.3. 2022_Map



Pages

Columns: Longitude (generated)

Rows: Latitude (generated)

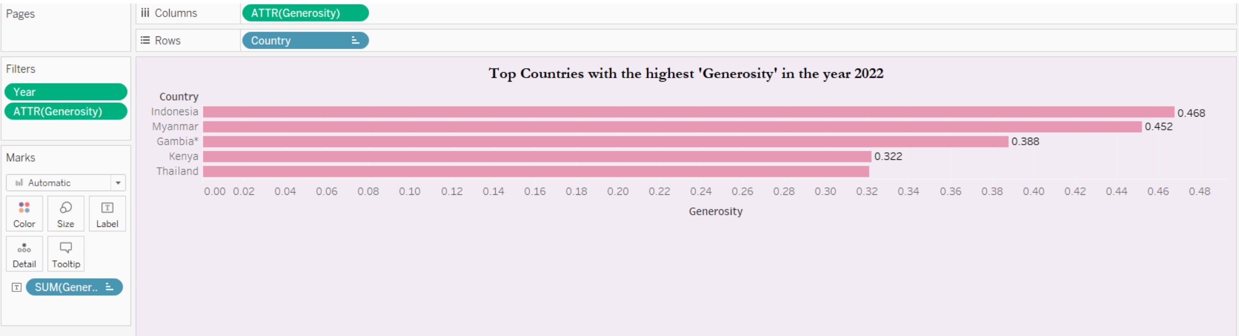
Filters: Year

Marks: Map

SUM(Happiness Score)

Country

6.2.4. 2022_TOP



Pages

Columns: ATTR(Generosity)

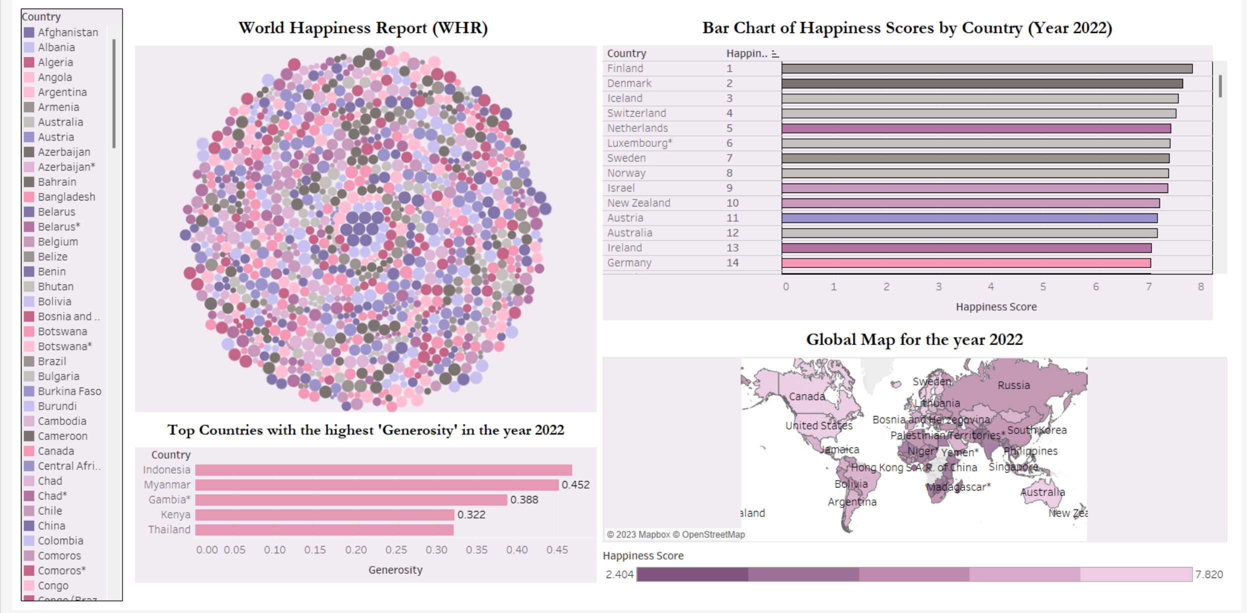
Rows: Country

Filters: Year, ATTR(Generosity)

Marks: Automatic

SUM(Gener..

6.3. Dashboard



Summary

The World Happiness Report, from 2012 published by the Sustainable Development Solutions Network, is based on Gallup World Poll data and supported by various organizations.
It was written and made by a group of independent experts, and the views expressed do not necessarily reflect the views of any United Nations organization, agency, or program.
The report is backed by Fondazione Ernesto Illy, illycaffè, Davines Group, Wall's, The Blue Chip Foundation, The Happier Way Foundation, and The Regeneration Society Foundation.
[source: About | The World Happiness Report]

Understanding happiness determinants is crucial for policymakers, governments, and organizations to improve citizens' quality of life. By focusing on the factors that truly matter, policies and interventions can be designed to enhance nation well-being.
However, happiness is a complex concept that cannot be fully captured by quantitative data alone.
A comprehensive understanding requires a multidisciplinary approach considering cultural, social, and psychological aspects.

As a sociologist specializing in media and quantitative methods, I think it was very exciting and appropriate to work on my World Happiness Report project.