Group Project – Exploratory Data Analysis, Statistical Visualization,

Classification, Cluster and Sentiment Analysis

Analyzing Global Happiness Trends (2018): What Factors Influence a Country's Happiness Ranking and How Have They Changed Over Time?

DATA 605 (Spring 2025) Actionable Visualization and Analytics

Group:

Bola Lateef (30260164)

Gautham Nagaraj (30273439)

Prichelle Lal (30261130)

Prince Eworitsemoghan Oloma (30263726)

June 15th, 2025

Section A: Introduction

This project explores the key factors influencing the World Happiness Index and examines how these factors have contributed to changes in the happiness rankings of countries from 2018 to 2024. The goal is to understand which variables, such as economic strength, climate, healthcare, and social well-being most strongly correlate with national happiness, as measured by the World Happiness Report. Our analysis includes exploratory data analysis, statistical visualization, classification, clustering, and sentiment analysis to provide a comprehensive understanding of happiness trends across countries and over time.

The motivation for this project was sparked by a video titled *"World's Happiest Countries: Explained"* (2018), which argues that economic power is not the strongest driver of happiness. Instead, it emphasizes social support, mental well-being, and other non-economic factors. This insight prompted us to investigate what variables shape happiness across the globe and how sentiment in public discourse aligns with the empirical data. We selected the year 2018 as a baseline to match the video's timeline and complement our sentiment analysis of YouTube comments.

Additionally, the video mentions that many of the world's happiest countries are Nordic or European nations (see Fig 1.1) with colder climates, raising interesting questions about the role of environmental factors in shaping national well-being.

Rank ↑	Country ↑	Life evaluation (3-year average) ↑	Change (since 2012) ↑
1	Finland	7.769	0.380
2	Denmark	7.600	−0.093
3	Norway	7.554	−0.101
5	Netherlands	7.488	−0.024
6	Switzerland	7.480	−0.170
7	Sweden	7.343	−0.137
8	New Zealand	7.307	0.086
9	Canada	7.278	−0.199
10	Austria	7.246	−0.123

Fig 1.1: Top 10 Happiest countries in the world

This observation led us to further explore climate as a potential predictor of happiness, alongside traditional economic and social indicators. By analyzing both 2018 and 2024 data, we aim to uncover how and why happiness rankings have shifted over time and whether new factors have emerged as more influential in recent years.

Through this multi-faceted approach combining quantitative analysis with sentiment-based insights, we aim to identify meaningful patterns that not only explain happiness trends but could also inform global policies focused on improving well-being beyond economic growth.

Section B: Analysis Questions

The purpose of this project is to investigate factors influencing the World Happiness Index using Exploratory Data Analysis, Statistical Visualization, Classification, Clustering, and Sentiment Analysis. The analysis aims to provide meaningful insights into the underlying drivers of national well-being and how these factors have evolved over time. Specifically, this project is guided by five key objectives:

1. What is the average climate of the countries at the top and bottom of the World Happiness Index?
2. How does economy fair with the happiness index?
3. What is a more influential factor, country's GDP or GDP per capita?
4. How have the rankings changed from 2018 to 2024?
5. How does the perception of corruption impact the happiness index of a country?

Section C: Data Sourcing and Justification

To investigate the factors influencing the World Happiness Index and analyze changes between 2018 and 2024, we sourced multiple datasets, both quantitative and qualitative that align with our research objectives and chosen analytical approaches.

1. World Happiness Report Data (Primary Dataset)

The main dataset used in this project was obtained from the Gallup World Poll via the World Happiness Report. This dataset provides annual happiness scores and rankings for countries based on survey results collected from 2011 to 2024. We selected this dataset because it is the original source used in global happiness studies and includes a rich set of variables directly aligned with our analysis questions. These variables include GDP per capita, social support, healthy life expectancy, freedom, generosity, and perceptions of corruption, all of which contribute to a country's ladder score (happiness index).

The columns are as follows:

- Year: The data consists of happiness index of countries from 2011 to 2024
- Rank: The rank of countries in the list
- Country name: Name of the country, a categorical variable
- Ladder score: Happiness index of the country
- upperwhisker: Upper-bound CI of the score
- lowerwhisker: Lower-bound CI of the score

- Explained by: Log GDP per capita: how GDP per capita explains the happiness index
- Explained by: Social support: how social support per capita explains the happiness index
- Explained by: Healthy life expectancy: how healthy life expectancy explains the happiness index
- Explained by: Freedom to make life choices: how freedom explains the happiness index
- Explained by: Generosity: how generosity explains the happiness index
- Explained by: Perceptions of corruption: how perception of corruption explains the happiness index
- Dystopia + residual: The deviance from the ladder score

This dataset contains 1,969 rows and 13 columns and meets the minimum data requirements for quantitative analysis. It allows us to perform classification, clustering, correlation, and time-series comparison.

2. Climate Data

To explore the potential influence of climate on happiness (as raised in the video analysis), we obtained average temperature data from the World Bank Climate Knowledge Portal (CCKP). This dataset includes the annual average temperature of countries from 1901 onward. It is publicly available under the World Bank's Open Data Policy.

We filtered this dataset to extract temperature data for 2018 and 2024 to align with the happiness dataset. This allows us to compare climate conditions in the top- and bottom-ranked countries and assess any correlations using EDA and cluster analysis.

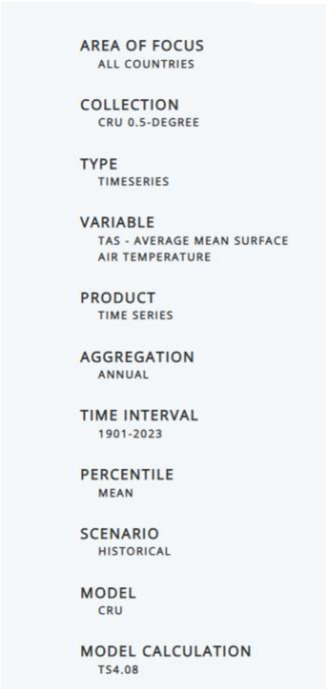

Fig 3.1: Filters applied to obtain climate data

3. GDP and GDP per Capita Data

To supplement and validate the economic indicators from the happiness dataset, we downloaded GDP and GDP per capita data for 2018 from the World Bank Data Catalog. These datasets are also licensed under the Creative Commons Attribution 4.0 International License (CC-BY 4.0), allowing ethical reuse and redistribution with proper citation.

We selected these external datasets to ensure accuracy and consistency in economic variables and allow for more granular analysis of how overall GDP compares with GDP per capita in predicting happiness.

4. Sentiment Analysis Data (Qualitative Component)

For the qualitative portion of our project, we performed sentiment analysis on user comments from the YouTube video titled "World's Happiest Countries: Explained" (published in 2018). This provides public opinion and emotional context related to happiness, economy, and quality of life, offering a complementary lens to the numerical data.

Lastly, all datasets used in this project are publicly available under open data licenses and comply with ethical standards for academic use.

Section D: Data Cleaning

Given the use of multiple datasets, Power BI was utilized to transform the data and establish relationships between them using the Country name as the common key.

1. World Happiness Report Data

To streamline the analysis, data prior to 2018 was excluded by filtering the dataset to retain only rows from 2018 onward.

Dirty Attributes:

a. Year, Rank, Country Name, Ladder Score
   - *Explanation* – The year attribute must be cleaned as there are a lot of missing values, especially for data prior to 2018. Since we are missing a large amount of data for some years, replacing the data is not feasible. Therefore, we opt to delete these columns (2011-2017). This will help with our analysis as our sentiment analysis video is focused on 2018 data only, reducing noise from prior years.
   - *Dirty Data* –

	0
Year	0
Rank	0
Country name	0
Ladder score	0
upperwhisker	1094
lowerwhisker	1094
Explained by: Log GDP per capita	1097
Explained by: Social support	1097
Explained by: Healthy life expectancy	1099
Explained by: Freedom to make life choices	1098
Explained by: Generosity	1097
Explained by: Perceptions of corruption	1098
Dystopia + residual	1101

Fig 4.1: Missing Values

- *Clean Data –*

Year	Rank	Country	Ladder score
2018	1	Finland	7.769
2018	2	Denmark	7.6
2018	3	Norway	7.554
2018	4	Iceland	7.494
2018	5	Netherlands	7.488
2018	6	Switzerland	7.48
2018	7	Sweden	7.343
2018	8	New Zealand	7.307
2018	9	Canada	7.278
2018	10	Austria	7.246
2018	11	Australia	7.228
2018	12	Costa Rica	7.167
2018	13	Israel	7.139
2018	14	Luxembourg	7.09
2018	15	United Kingd	7.054
2018	16	Ireland	7.021
2018	17	Germany	6.985
2018	18	Belgium	6.923
2018	19	United States	6.892
2018	20	Czechia	6.852
2018	21	United Arab E	6.825

Fig 4.2: Countries ranked from dataset 'Worlds' Happiest Countries'

Once these columns have been removed - "filtering" the years to only include 2018 to 2023 data, we can check for missing values.

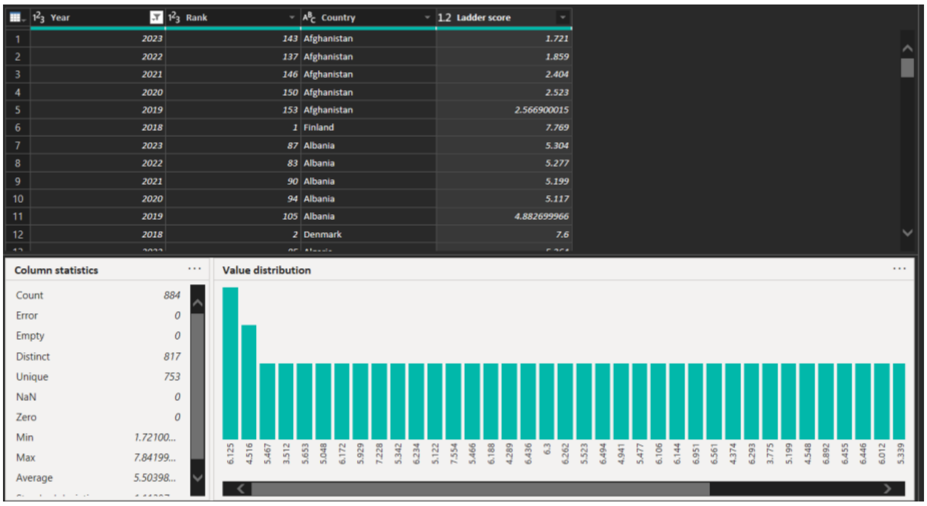

Fig 4.3: Column profile of happiness index from dataset 'Worlds' Happiest Countries'

There were no missing values in either of these columns, therefore this data is now clean.

Clean Attributes

a. Upper whisker, Lower whisker, GDP, Social Support, Healthy Life Expectancy, Freedom to make life choices, Generosity, Perceptions of Corruption, Dystopian & Residual
   - *Cleaning Techniques (standard checks)* - The dataset passed all standard cleaning checks for the selected attributes. There were no duplicate rows for the same country and year, and all values were present and in the correct numeric format. Each attribute fell within expected, plausible ranges such as non-negative values for GDP and Life Expectancy. Although some values like Generosity were very low or zero, these were deemed to reflect actual conditions rather than errors. Overall, the data is clean and ready for analysis without requiring further adjustments.
   - *Why this data is already clean* – The selected attributes are complete, numeric, and free of formatting or structural issues. No encoding or naming inconsistencies were found, and there are no missing or duplicate entries. Since the values are within reasonable ranges for socioeconomic indicators, this portion of the dataset is considered clean and ready for analysis.

2. GDP by Country & GDP per Capita Datasets

Cleaning the second dataset 'GDP by Country': There a lot of missing values for GDP for different countries across the year.

Dirty attributes

a. Country Name in GDP by Country
   - *Explanation* – We identified that many Countries have missing values in this dataset as well. Since there are many countries, and only a limited number were evaluated in our sentiment analysis we check to see if any countries from our sentiment analysis have missing data. In Fig 4.4 below, the countries with

missing values are not in the list that we need for our analysis. So instead of replacing them we can remove these rows.

- *Dirty Data –*

```
rows_with_nan = df2[df2.isnull().any(axis=1)]

countries_with_empty_values = rows_with_nan['Country Name'].tolist()

print("Countries with empty values:")
countries_with_empty_values
```

```
Countries with empty values in the second dataframe:
['American Samoa',
 'Bhutan',
 'Cuba',
 'Eritrea',
 'Gibraltar',
 'Greenland',
 'Guam',
 'Isle of Man',
 'Not classified',
 'Lebanon',
 'Liechtenstein',
 'St. Martin (French part)',
 'Northern Mariana Islands',
 'New Caledonia',
 "Korea, Dem. People's Rep.",
 'French Polynesia',
 'San Marino',
 'South Sudan',
 'Syrian Arab Republic',
 'Tonga',
 'Venezuela, RB',
 'British Virgin Islands',
 'Virgin Islands (U.S.)',
 'Yemen, Rep.']
```

Fig 4.4: List of countries with missing data from dataset 'GDP by Country'

```
= Table.Skip(#"Renamed Columns",1)
```

	e	1.2 2018	1.2 2019	1.2 2020	1.2 2021	1.2 2022	1.2 2023
1		3276184358	3395798883	2481857123	2929446578	3279343544	3648573136
2		1.01231E+12	1.00972E+12	9.33392E+11	1.08575E+12	1.19142E+12	1.24547E+12
3		18053222687	18799444490	19955929052	14259995441	14497243872	17233051620
4		7.70495E+11	8.26484E+11	7.89802E+11	8.49312E+11	8.83974E+11	7.99106E+11
5		79450688259	70897962732	48501561204	66505129988	1.044E+11	84824654482
6		15379509892	15585105131	15241458745	18032010564	19017242586	23547179830
7		3218419897	3155149348	2891001149	3324647529	3380612573	3785067332
8		2.86735E+12	2.89903E+12	2.56278E+12	2.96924E+12	3.60981E+12	3.51242E+12
9		4.27049E+11	4.1799E+11	3.49473E+11	4.15179E+11	5.02732E+11	5.1413E+11
10		5.2482E+11	4.47755E+11	3.85741E+11	4.86564E+11	6.3279E+11	6.46075E+11
11		12457940695	13619290539	12641698583	13878908629	19513506553	24085749592
12		639000000	647000000	721000000	750000000	871000000	null
13		1661529630	1725351852	1410796296	1601366667	1867733333	2033085185
14		1.42781E+12	1.39272E+12	1.32841E+12	1.55674E+12	1.69086E+12	1.72806E+12
15		4.52582E+11	4.42984E+11	4.34398E+11	4.80467E+11	4.71774E+11	5.11685E+11
16		47112470052	48174235294	42693000000	54825411765	78807470588	72356176471
17		2667182200	2576518880	2649680261	2775798698	3338722828	2642161669
18		5.42639E+11	5.36726E+11	5.29694E+11	5.98494E+11	5.93439E+11	6.44783E+11
19		14262408080	14391686309	15651545209	17690083520	17401746309	19676049076
20		15890066221	16032813503	17725010533	19697516284	18820218691	20324617845
21		3.21363E+11	3.51232E+11	3.73979E+11	4.16272E+11	4.60132E+11	4.37415E+11
22		66289266349	68631726289	70716821235	84417951261	90642851836	1.02408E+11
23		39567978723	40446808511	35837632979	40840212766	46680398936	46079867021
24		12615800000	13016200000	9958200000	11368900000	13136400000	14338500000
25		20484053869	20482608755	20226036564	23672712242	24534663636	27514782476
26		60031173808	64410122847	61371755326	69673747132	73775179925	71857382746

Fig 4.5: Columns displaying the spread of data across the years from dataset 'GDP by Country'

We can unpivot these columns and rename the column to year to match the format of the other datasets.

- *Clean data –*

	Country	Year	1.2 GDP
1	Aruba	01-01-2018	3276184358
2	Aruba	01-01-2019	3395798883
3	Aruba	01-01-2020	2481857123
4	Aruba	01-01-2021	2929446578
5	Aruba	01-01-2022	3279343544
6	Aruba	01-01-2023	3648573136
7	Africa Eastern and Southern	01-01-2018	1.01231E+12
8	Africa Eastern and Southern	01-01-2019	1.00972E+12
9	Africa Eastern and Southern	01-01-2020	9.33392E+11
10	Africa Eastern and Southern	01-01-2021	1.08575E+12
11	Africa Eastern and Southern	01-01-2022	1.19142E+12
12	Africa Eastern and Southern	01-01-2023	1.24547E+12
13	Afghanistan	01-01-2018	18053222687
14	Afghanistan	01-01-2019	18799444490
15	Afghanistan	01-01-2020	19955929052
16	Afghanistan	01-01-2021	14259995441
17	Afghanistan	01-01-2022	14497243872
18	Afghanistan	01-01-2023	17233051620
19	Africa Western and Central	01-01-2018	7.70495E+11
20	Africa Western and Central	01-01-2019	8.26484E+11
21	Africa Western and Central	01-01-2020	7.89802E+11
22	Africa Western and Central	01-01-2021	8.49312E+11
23	Africa Western and Central	01-01-2022	8.83974E+11
24	Africa Western and Central	01-01-2023	7.99106E+11
25	Angola	01-01-2018	79450688259
26	Angola	01-01-2019	70897962732
27	Angola	01-01-2020	48501561204

Fig 4.6: Un-pivoted columns and changed data type to year from dataset 'GDP by Country'

In the above step the columns 2018 – 2023 were un-pivoted, renamed to year and changed data type to year. The other columns 'Country code', 'Indicator name', 'Indicator code' were removed as they are not required for our analysis. The GDP column was filtered to remove missing data from the list of countries which are not part of the analysis as shown in the image of the Python code above (Fig 4.4).

b. Date in GDP per Capita
   - *Explanation* – The next dataset 'GDP per Capita' will follow the same steps as the dataset 'GDP by Country'. This includes unpivoting the year columns (2018–2023), renaming the column to "Year," and converting the data type to numeric or date format. Unnecessary columns such as 'Country Code', 'Indicator Name', and 'Indicator Code' will be removed, and any rows with missing GDP per capita values or irrelevant countries will be filtered out.
   - *Dirty Data –*

Column2	Column3	Column4	Column5	Column6	Column7
Country Code	Indicator Name	Indicator Code	2018	2019	2020
ABW	GDP per capita growth (annual %)	NY.GDP.PCAP.KD.ZG	2.234427843	-2.496548796	-25.79322999
AFE	GDP per capita growth (annual %)	NY.GDP.PCAP.KD.ZG	-0.102513344	-0.549574951	-5.45140964
AFG	GDP per capita growth (annual %)	NY.GDP.PCAP.KD.ZG	-1.713742693	0.856294729	-5.382514642
AFW	GDP per capita growth (annual %)	NY.GDP.PCAP.KD.ZG	0.271791785	0.744453271	-3.341151985
AGO	GDP per capita growth (annual %)	NY.GDP.PCAP.KD.ZG	-4.665969917	-4.010017337	-8.672076174
ALB	GDP per capita growth (annual %)	NY.GDP.PCAP.KD.ZG	4.276311551	2.498289041	-2.756939782
AND	GDP per capita growth (annual %)	NY.GDP.PCAP.KD.ZG	-0.302119304	0.265352821	-12.22383849
ARB	GDP per capita growth (annual %)	NY.GDP.PCAP.KD.ZG	0.82172505	-0.31022907	-6.694720125
ARE	GDP per capita growth (annual %)	NY.GDP.PCAP.KD.ZG	-0.024508823	0.047746022	-4.504667628
ARG	GDP per capita growth (annual %)	NY.GDP.PCAP.KD.ZG	-3.415536839	-2.695067235	-10.33610969
ARM	GDP per capita growth (annual %)	NY.GDP.PCAP.KD.ZG	5.568501179	7.836084387	-7.168664527
ASM	GDP per capita growth (annual %)	NY.GDP.PCAP.KD.ZG	4.038507121	0.897584681	5.351787426
ATG	GDP per capita growth (annual %)	NY.GDP.PCAP.KD.ZG	6.414456144	2.606180505	-19.30610797
AUS	GDP per capita growth (annual %)	NY.GDP.PCAP.KD.ZG	1.350866132	0.673067166	-1.343977433
AUT	GDP per capita growth (annual %)	NY.GDP.PCAP.KD.ZG	1.986262691	1.303503434	-6.706393805
AZE	GDP per capita growth (annual %)	NY.GDP.PCAP.KD.ZG	0.624486168	1.635850481	-4.952701739
BDI	GDP per capita growth (annual %)	NY.GDP.PCAP.KD.ZG	-1.411807623	-1.476334832	-2.548980787
BEL	GDP per capita growth (annual %)	NY.GDP.PCAP.KD.ZG	1.415289473	1.890719342	-5.202440099
BEN	GDP per capita growth (annual %)	NY.GDP.PCAP.KD.ZG	3.737015543	3.982113266	1.120202648
BFA	GDP per capita growth (annual %)	NY.GDP.PCAP.KD.ZG	3.767726783	3.243918662	-0.443420002
BGD	GDP per capita growth (annual %)	NY.GDP.PCAP.KD.ZG	6.441974947	6.972643998	2.586478166
BGR	GDP per capita growth (annual %)	NY.GDP.PCAP.KD.ZG	3.292633847	4.522003262	-2.633051094
BHR	GDP per capita growth (annual %)	NY.GDP.PCAP.KD.ZG	1.918495729	3.382740869	-5.173074009
BHS	GDP per capita growth (annual %)	NY.GDP.PCAP.KD.ZG	2.046594728	-1.846636147	-21.65927426
BIH	GDP per capita growth (annual %)	NY.GDP.PCAP.KD.ZG	5.105248784	4.162604855	-1.657508895

Fig 4.7 GDP per capita pivoted columns

- *Clean Data –*

	Country	Date	GDP Per Capita
1	Aruba	01-01-2018	2.234427843
2	Aruba	01-01-2019	-2.496548796
3	Aruba	01-01-2020	-25.79322999
4	Aruba	01-01-2021	25.15496385
5	Aruba	01-01-2022	8.912308079
6	Aruba	01-01-2023	4.21613214
7	Africa Eastern and Southern	01-01-2018	-0.102513344
8	Africa Eastern and Southern	01-01-2019	-0.549574951
9	Africa Eastern and Southern	01-01-2020	-5.45140964
10	Africa Eastern and Southern	01-01-2021	1.842086727
11	Africa Eastern and Southern	01-01-2022	0.903488101
12	Africa Eastern and Southern	01-01-2023	-0.226928405
13	Afghanistan	01-01-2018	-1.713742693
14	Afghanistan	01-01-2019	0.856294729
15	Afghanistan	01-01-2020	-5.382514642
16	Afghanistan	01-01-2021	-22.58448187
17	Afghanistan	01-01-2022	-7.576668912
18	Afghanistan	01-01-2023	0.540655554
19	Africa Western and Central	01-01-2018	0.271791785
20	Africa Western and Central	01-01-2019	0.744453271
21	Africa Western and Central	01-01-2020	-3.341151985
22	Africa Western and Central	01-01-2021	1.60719669
23	Africa Western and Central	01-01-2022	1.397033814
24	Africa Western and Central	01-01-2023	0.917670772
25	Angola	01-01-2018	-4.665969917
26	Angola	01-01-2019	-4.010017337
27	Angola	01-01-2020	-8.672076174

Fig 4.8: Un-pivoted columns and changed data type to year from dataset 'GDP per Capita'

Clean Attributes

a. Country code, country anme, indicator name, year
   - *Cleaning Techniques (standard checks)* - Standard data cleaning checks were performed to ensure the dataset's integrity. No duplicate rows were found for any country-indicator pair, and all key attributes had complete values. String fields such as country and indicator names were consistently formatted, and UTF-8 encoding was verified to prevent character issues. While some GDP values appear in scientific notation, they remain accurate and usable, though they may be reformatted for easier interpretation in analysis or visualization.
   - *Why this data is already clean* – This dataset appears to be sourced from a reputable international database (World Bank), which maintains high data

quality standards. The structural format is consistent, variables follow standard conventions (ISO country codes, GDP indicators), and column headers are meaningful. No syntactical or formatting errors were found in the key identifying attributes, indicating the data is clean and ready for use in time series analysis or economic comparison.

3. Climate Data

a. Date

- *Explanation* – to align with our above datasets, we transpose all the date columns in year-month form into one column called Date. This will make for easy merging and reliability in findings when using the other datasets alongside Climate Data
- *Dirty Data* –

code	Country	2018-07	2019-07	2020-07	2021-07	2022-07
ABW	Aruba (Neth.)	28.95	29.33	29.34	29.04	
AFG	Afghanistan	14.18	13.8	13.19	14.34	
AGO	Angola	21.66	21.76	21.72	21.73	
AIA	Anguilla (U.K.)	27.7	28.08	28.04	27.96	
ALA	Finland	6.94	6.92	7.82	6.21	
ALB	Albania	13.49	13.47	13.17	12.84	
AND	Andorra	9.24	8.98	9.19	8.33	
ARE	United Arab Emirates	28.85	28.4	28.2	29.02	
ARG	Argentina	15.27	15.22	15.55	15.42	
ARM	Armenia	9.38	8.6	8.29	8.98	
ASM	American Samoa (U.S.)	27.33	27.61	27.52	27.34	
ATF	French Southern and Antarctic Lands (Fr.)	4.3	4.32	4.21	4.27	
ATG	Antigua and Barbuda	27.15	27.5	27.51	27.4	
AUS	Australia	22.52	22.8	22.53	22.04	
AUT	Austria	8.63	8.45	8.34	7.3	
AZE	Azerbaijan	14.11	13.71	13.33	13.96	
BDI	Burundi	20.38	20.63	20.43	20.51	
BEL	Belgium	11.72	11.43	11.87	10.45	
BEN	Benin	28.37	28.15	28.45	28.74	
BES	Bonaire, Sint Eustatius and Saba	27.44	27.8	27.78	27.68	
BFA	Burkina Faso	29.63	29.31	29.68	30.04	
BGD	Bangladesh	25.29	25.64	25.4	25.74	
BGR	Bulgaria	12.26	12.69	12.48	11.8	
BHR	Bahrain	28.8	28.34	28.14	29.16	
BHS	The Bahamas	25.87	26.36	26.41	25.93	
BIH	Bosnia and Herzegovina	11.54	11.61	11.06	10.41	

Fig 4.9: Climate data across the years for different countries from dataset 'Climate data'

- **Clean data –**

	Country	Date	Temperature °C
1	Aruba (Neth.)	01-07-2018	28.95
2	Aruba (Neth.)	01-07-2019	29.33
3	Aruba (Neth.)	01-07-2020	29.34
4	Aruba (Neth.)	01-07-2021	29.04
5	Aruba (Neth.)	01-07-2022	28.81
6	Aruba (Neth.)	01-07-2023	29.39
7	Afghanistan	01-07-2018	14.18
8	Afghanistan	01-07-2019	13.8
9	Afghanistan	01-07-2020	13.19
10	Afghanistan	01-07-2021	14.34
11	Afghanistan	01-07-2022	14.51
12	Afghanistan	01-07-2023	14.67
13	Angola	01-07-2018	21.66
14	Angola	01-07-2019	21.76
15	Angola	01-07-2020	21.72
16	Angola	01-07-2021	21.73
17	Angola	01-07-2022	21.62
18	Angola	01-07-2023	21.51
19	Anguilla (U.K.)	01-07-2018	27.7
20	Anguilla (U.K.)	01-07-2019	28.08
21	Anguilla (U.K.)	01-07-2020	28.04
22	Anguilla (U.K.)	01-07-2021	27.96
23	Anguilla (U.K.)	01-07-2022	27.73
24	Anguilla (U.K.)	01-07-2023	28.35
25	Finland	01-07-2018	6.94
26	Finland	01-07-2019	6.92
27	Finland	01-07-2020	7.82

Fig 4.10: Un-pivoted columns and changed data type to year from dataset 'Climate data'

Clean Attributes

a. Code, Name
   - Cleaning Techniques (standard checks) - Checked for duplicates, consistent formatting, extra spaces, formatting consistency, and missing values
   - Why this data is already clean – All entries are consistent 3-letter ISO country codes with no missing or duplicate values. Each value is standardized and uniquely identifies a country or territory. Additionally, all country/region names are present, properly capitalized, and consistently formatted. No nulls or extraneous characters were detected.

Now that all the data has been cleaned, we can create relationships for the dataset based on the country name.

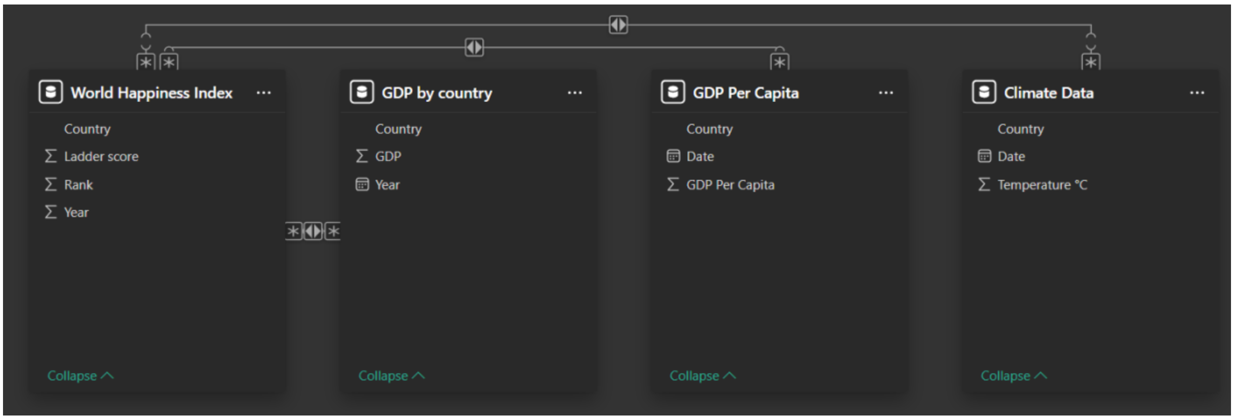

Fig 4.11: Creating Many-to-Many relationships between multiple tables based on column country

It is important to create this relationship as the other datasets will help explain the relationship between World Happiness index and other factors. With this, we conclude the data cleaning/pre-processing part. We can now move on to visualizing the data.

Section E: Visualizations & Findings

The first 3 analysis questions can be answered using the Power BI dashboard created. To answer these questions our focus will be on the data from 2018 as it aligns with the findings from the Sentiment Analysis.

1. **What is the average climate of the countries at the top and bottom of the list of world's happiest countries?**
   - *Visualization –*

Country	Rank	Ladder score	Average of Temperature °C
Finland	1	7.77	5.01
Denmark	2	7.60	9.61
Norway	3	7.55	2.62
Iceland	4	7.49	2.04
Netherlands	5	7.49	11.35
Switzerland	6	7.48	7.48
Sweden	7	7.34	3.70
New Zealand	8	7.31	11.10
Canada	9	7.28	-3.97
Austria	10	7.25	8.37

Fig 5.1 Top Countries

- *Explanation* – Ranking the countries in ascending order we obtain results for the world's happiest countries and their associated ladder score/happiness index. The climate and geographical location of these countries lie very close to each other.
- *Findings* – The first image represents the top 10 of the world's happiest countries and their associated ladder score and temperature(average). The average is taken across the year for the year 2018, we can see a pattern here where these countries do not exceed the average temperature of 11°C. This is a key finding that was mentioned in the video – stating that the happier countries have a relatively cooler climate.

Although some of these countries have extreme climate conditions the maximum temperature does not cross a threshold compared to the list of countries at the bottom of the list, shown below. Now let us look at the countries on the bottom of the list:

- *Visualization* –

Country	Rank	Ladder score	Average of Temperature °C
South Sudan	156	2.85	27.80
Central African Republic	155	3.08	25.58
Afghanistan	154	3.20	14.12
Tanzania	153	3.23	22.98
Rwanda	152	3.33	19.23
Malawi	150	3.41	22.74
Botswana	148	3.49	22.16
Haiti	147	3.60	25.27
Zimbabwe	146	3.66	22.02
Burundi	145	3.78	20.54

Fig 5.2 Bottom Countries

- *Explanation* – When the countries are sorted in descending order, based on rank; the average temperature is much higher compared to the countries on top of the list. There is no similar region among these countries as opposed to the geographical pattern from the previous findings. The lowest rank is 156, this is due to many countries simply not having enough data through surveys to be published in the world happiness index report.
- *Findings* – The pattern is clear; these countries are at the bottom of the list and have a much higher temperature throughout the year averaging more than 22°C in most cases. However, it is not a good idea to single out temperature as the only factor, though there is an undeniable effect on the happiness index/ladder score.

2. **How does economy fair with the happiness index?**
   - Visualization –

Country	Average GDP (Trillions)	Ladder score	Rank	Sum of GDP Per Capita
United States	23.49 T	6.89	19	11.84
China	16.06 T	5.19	93	30.50
Japan	4.78 T	5.89	58	3.26
Germany	4.16 T	6.99	17	2.02
India	3.05 T	4.02	140	23.54
United Kingdom	3.01 T	7.05	15	3.06
France	2.83 T	6.59	24	4.58
Italy	2.10 T	6.22	36	9.29
Canada	1.91 T	7.28	9	0.63
Brazil	1.84 T	6.30	32	7.28
Russian Federation	1.83 T	5.65	68	10.46
Australia	1.52 T	7.23	11	6.51
Spain	1.44 T	6.35	30	5.22
Mexico	1.38 T	6.60	23	1.13

Fig 5.3 Economic Indicators & Happiness

- Explanation – The table presents data for several countries, showing their Average GDP (in trillions), Happiness Index (Ladder Score), Happiness Rank, and Sum of GDP Per Capita. This comparative view helps evaluate whether higher economic indicators correspond to greater national happiness.
- *Findings* – The visualization displays the countries with the highest GDP, averaged across the year; 2018. None of the countries with the highest GDP make it to the top of the list, with one exception being Canada. It has been argued that the economy is not a good predictor when it comes to measuring the world happiness index.

- *Additional Visualizations –*

Country	Average GDP (Trillions)	Ladder score	Rank
Malawi	0.01 T	3.41	150
Madagascar	0.01 T	3.93	143
Togo	0.01 T	4.09	139
Chad	0.01 T	4.35	132
Sierra Leone	0.01 T	4.37	129
Mauritania	0.01 T	4.49	122
Niger	0.01 T	4.63	114
Namibia	0.01 T	4.64	113
Somalia	0.01 T	4.67	112
North Macedonia	0.01 T	5.27	84
Tajikistan	0.01 T	5.47	74
Montenegro	0.01 T	5.52	73
Mauritius	0.01 T	5.89	57
Kosovo	0.01 T	6.10	46

Fig 5.4 Average GDP in Trillions

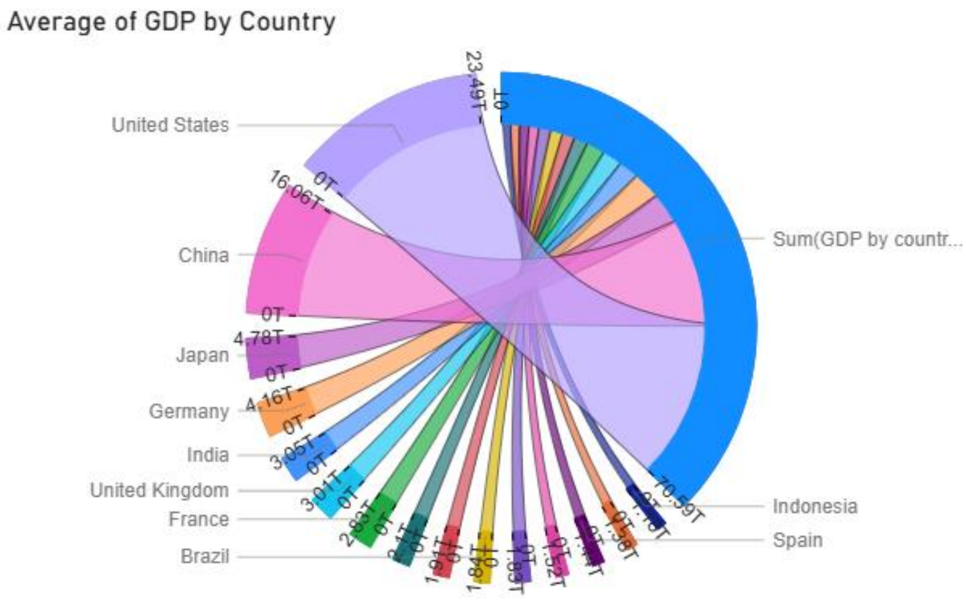

Fig 5.5 Average GDP by Country

- *Additional Explanation* – The pie chart titled "Average of GDP by Country" visually represents the share of average GDP contributed by each country in the dataset. Each slice corresponds to a country, and the size of the slice reflects its proportion of the total combined GDP (in trillions). The chart highlights global economic distribution among major economies.
    - The United States has the largest segment, with an average GDP of 23.49 trillion, followed by China **at** 16.06 trillion, together dominating a

significant portion of the pie. Other countries like Japan, Germany, India, and the United Kingdom occupy moderately sized segments, while the remaining countries hold much smaller proportions of the global GDP represented in this chart.

- *Additional Findings* – The countries with the lowest GDP do not perform well on the ranking either, simply proving that economy is not a great way to measure the happiness index. The numbers may not provide useful information as it is measured in trillions. But we can some European countries with a low GDP outperform their African counterpart, in terms of rank, with almost equal GDPs.

3. **What is a more influential factor, country's GDP or GDP per capita**
    - *Visualization – Tree Map*

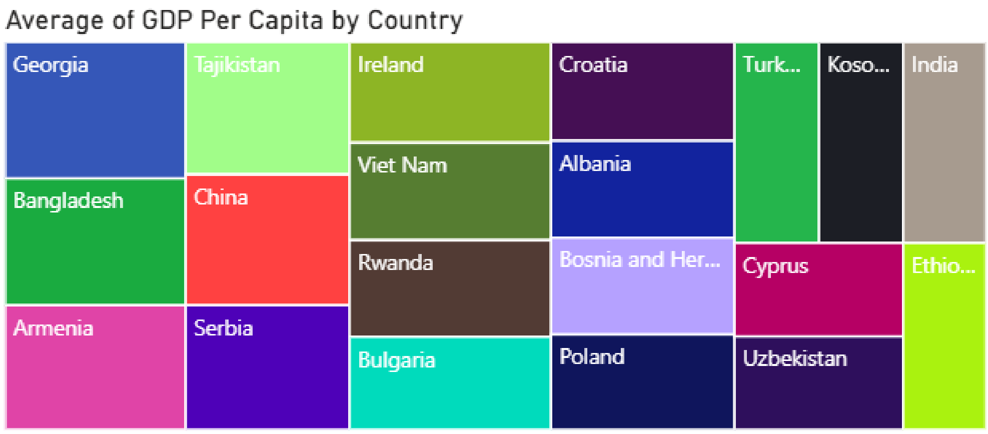

Fig 5.6 Average GDP Capita by Country

- *Explanation* – The tree map was used to visually compare the relative influence of GDP and GDP per capita by displaying the contribution of each within the same space. Each square represents a country, and the size and color intensity reflect the strength or weight of the GDP-related factor in the happiness index. This format allows for quick visual comparisons between countries and highlights whether GDP or GDP per capita better aligns with higher happiness scores. The colorful layout makes patterns easier to detect than with raw numbers, supporting the exploration of whether GDP per capita offers a more meaningful perspective than total GDP when assessing well-being.
- *Findings* – We see a similar pattern here, none of these countries make it to the top of the list. GDP per capita is the total value of a country's finished goods and services (gross domestic product) divided by its total population (per capita). Gross domestic product (GDP) per capita is often considered an indicator of a country's standard of living; however, this is inaccurate because GDP per capita is not a measure of personal income. Measures of personal income include average wage, real income, median income, disposable income and GNI per capita. There are variety of factors that come into play while measuring the happiness index as although a countries GDP and GDP per capita may be good, the distribution of wealth and spending power plays a bigger factor. This cannot be measured in a dataset by rather through research which will be included in our findings.

4. **How have the rankings changed from 2018 to 2024?**
    - *Visualization –*

Country	Rank_2018	Rank_2024	RankChange
Afghanistan	145	146	-1
Denmark	3	2	1
Finland	1	1	0
Iran	106	122	-16
Kosovo	67	37	30
Lebanon	88	137	-49
Lithuania	50	19	31
Serbia	70	37	33
Sri Lanka	120	127	-7
Venezuela	102	133	-31

Fig 5.7 2018 vs 2024 rankings

- *Explanation* – a table-style visualization with four columns — Country, Rank 2018, Rank 2024, and Difference — was used. This layout allows for a direct, side-by-side comparison of each country's position in 2018 versus 2024. The Difference column clearly quantifies how much each country's rank has changed, and the use of conditional color formatting (a gradient from red to green) highlights the direction and magnitude of that change: Dark green represents major positive shifts in Rank, Red cells highlight significant declines, and Light shades indicate minimal change. This colour formatting enhances the ability to visually interpret performance across countries.
- *Findings* –

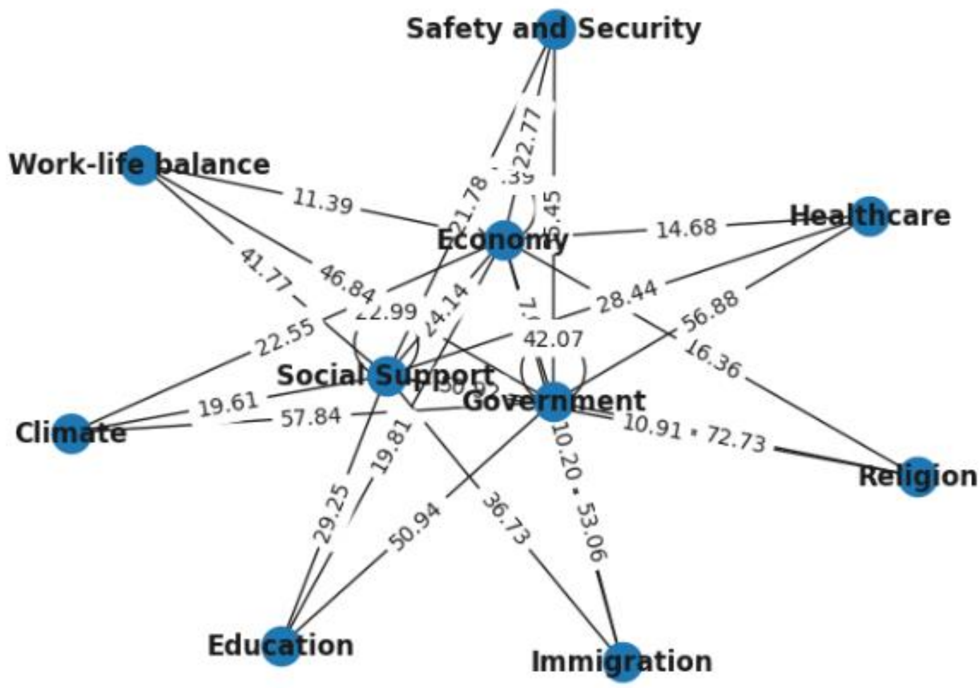

Fig 5.8 Variable Centrality

The analysis of happiness rankings from 2018 to 2024 reveals notable shifts among countries. Serbia, Lithuania, and Kosovo showed the most significant improvements, each climbing 30 or more ranks — likely reflecting positive changes in governance, health, or social conditions. In contrast, Lebanon, Venezuela, and Iran experienced steep declines, possibly driven by political unrest, economic hardship, or conflict. Some countries saw little to no change: Finland maintained its top position, Denmark remained near the top with a slight rise, and Afghanistan continued to rank at the bottom, with only a marginal decline. These patterns highlight both progress and persistent challenges in global well-being. The analysis reveals dynamic shifts in global happiness rankings analysed by 10 sampled countries. While some nations have made remarkable progress, others have experienced notable setbacks, reflecting the influence of economic, political, and social conditions. Nordic countries like Finland and Denmark demonstrate persistent excellence in citizen satisfaction, while crisis-affected regions show declining trends. This comparison underscores the complexity of happiness, influenced by more than just economic output but also including governance, public trust, and quality of life.

- Countries like Canada noticed a decline in the ranking possibly due to high inflation and immigration, as explained by the degree of centrality from the findings of the sentiment analysis. The people believe in decrease of economy, social support and confidence in the government leading to a steady downward shift in the happiness index.

5. **How does the perception of corruption impact the happiness index of a country?**
   - *Visualization –*

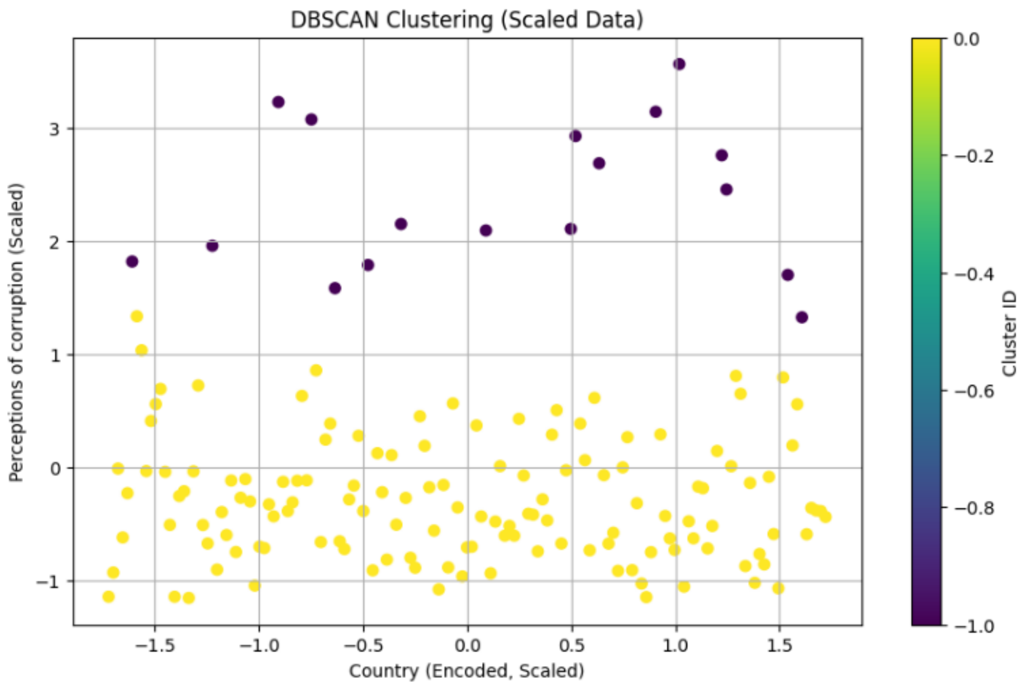

Fig 5.9 Clusters and Corruption in Countries

   - *Explanation* – A DBScan clustering was visualized using the columns 'Country name' and 'Explained by perception of corruption'. The columns may be misleading as one may think of it as how corrupted do the people think a country is, it is rather how much of the happiness can be explained using the 'perception' of corruption. From the visualization we get two clusters; one with very low Explained by perception of corruption and the other with a high Explained by perception of corruption which is almost considered as outliers by the DBScan algorithm. As the country name is categorical it is encoded using one-hot encoding and displayed as above.

- *Findings* – Looking at the countries included in the lower cluster:

Country name	Explained by: Perceptions of corruption
Afghanistan	0.001226
Albania	0.025361
Algeria	0.129191
Argentina	0.060415
Armenia	0.104618
...	...
Venezuela	0.063625
Viet Nam	0.089848
Yemen	0.087352
Zambia	0.086705
Zimbabwe	0.080929

Fig 5.10 Overview of Corruption score

These countries have a lower score indicating that the people believe that their government is indeed corrupted.

Now we look at the countries on the upper cluster:Fig

Australia	0.335996
Canada	0.351702
Denmark	0.495260
Finland	0.477857
Germany	0.309362
Hong Kong SAR of China	0.332485
Ireland	0.373433
Luxembourg	0.367041
Netherlands	0.368570
New Zealand	0.461268
Norway	0.434101
Rwanda	0.485542
Singapore	0.533162
Sweden	0.442066
Switzerland	0.407946
United Kingdom	0.322602
Uzbekistan	0.280260

Fig 5.11 Upper Cluster Country Overview

It is no surprise that some these countries are the highest ranked when it comes to the happiness index. To reiterate, the high score does not mean that the government is corrupted but rather that the people perceive the opposite. These scores are so high that they are detected as outliers.

<u>Conclusion</u>

The project explores the key factors influencing the World Happiness Index from 2018 to 2024, aiming to understand what drives changes in national happiness rankings. Drawing from datasets on GDP, GDP per capita, climate, and public perception, alongside sentiment analysis from a 2018 YouTube video, the project investigates whether economic indicators or social factors are more influential. One of the most notable findings is the link between climate and happiness: countries with cooler average temperatures (around 11°C), particularly Nordic nations, consistently rank among the happiest. Conversely, countries with warmer climates often fall to the bottom of the rankings. The analysis also reveals that GDP and GDP per capita do not strongly correlate with happiness. High-GDP countries like the U.S. and China are not necessarily the happiest, and GDP per capita, while slightly more aligned, still fails to capture the impact of income distribution and social well-being. A comparison of rankings between 2018 and 2024 shows significant movement, with Serbia, Lithuania, and Kosovo improving markedly, while Lebanon, Venezuela, and Iran declined, largely due to political and economic instability. Trust in governance also emerged as a critical factor. Through clustering analysis, the study found that countries perceived as having low corruption tended to score higher on the happiness index. Sentiment analysis further supported these findings, reflecting public concern around economic inequality and governance in countries experiencing a drop in happiness. Overall, the report concludes that while economic metrics are relevant, social trust, climate, and perceived corruption play a more decisive role in shaping a nation's happiness.

In conclusion, the report affirms that global happiness is influenced by a complex interplay of variables. While economic factors like GDP and GDP per capita matter, they are far from sufficient on their own. Climate, perception of corruption, social support, freedom, and trust in governance emerge as more powerful indicators. These findings suggest that policymakers seeking to improve national happiness should prioritize social equity, institutional trust, and citizen well-being rather than focusing solely on economic growth. The report offers a compelling case for adopting a broader, more human-centered approach to understanding and improving happiness on a global scale.

Recommendations

Lastly, based on our analysis and findings we have four recommendations:

1. Prioritize Governance and Public Trust to Boost Happiness
    - Countries with low perceived corruption consistently rank higher on the happiness index. Governments should invest in transparent institutions, accountability mechanisms, and community engagement to build public trust. Policies that improve citizens' confidence in their leadership can significantly elevate national well-being, even in the absence of high GDP.
2. Focus more on well-being than economic factors
    - The findings clearly show that high GDP or GDP per capita does not guarantee happiness. Policymakers should focus on equitable wealth distribution, improved social services, and mental health initiatives. Emphasizing citizen well-being rather than just economic indicators will lead to more meaningful and sustained improvements in happiness rankings.

3. Greater number of community well-being initiatives, especially in countries with warmer climates
    - Top-ranking countries tend to have cooler climates, but the key takeaway is that these nations also foster strong social systems and resilience**.** Rather than trying to replicate geographic conditions, countries in warmer climates should invest in climate adaptation strategies**,** community well-being programs**, and** infrastructure that enhances quality of life regardless of temperature.
4. Keep track of Public Sentiment to analyze and actively make changes
    - Changes in happiness rankings (e.g., Canada's decline) reflect broader public concerns like inflation, social support, and governance. Governments should routinely analyze citizen sentiment, using tools like surveys or sentiment analysis, to identify and act on early warning signs of societal dissatisfaction. This proactive approach supports long-term happiness and stability.