ASSIGNMENT 05
9-09-2025

WEB SCRAPPING:

Web Scraping means extracting data directly from a website (like tables, lists, or structured information) and then loading it into Power BI for analysis and visualization.

Web: contains the data, it is an interface or a server page we storing the data

Scrapping: it is nothing but an extracting.

- Internet source is necessary for web scrapping, without internet we cannot fetch the information.
- when Data is loading from an any data collective (Excel, SQL what ever it may be) should be in a tabular format, if it is another format we need to convert into a tabular format.
- In web scrapping also data should be in a tabular format only. Then that web page can be upload into a power bi.

Advantage: the data will be dynamic. It is not necessary to do again & again Re-aploading the data, because the data will automatically updated.

Disadvantage: internet should be compulsory.

How Web Scraping works in Power BI:

1. Go to Power BI Desktop → Home → Get Data → Web.
2. Enter the URL of the website you want to extract data from.
3. Power BI will scan the webpage and show you available tables/lists that can be imported.
4. Select the required table(s) → Load or Transform in Power Query.
5. Apply cleaning, transformations, and modelling.
6. Use the data to build visuals, dashboards, and reports.

For this web Scrapping concept we have taken data from chrome i.e., **World Happiness Report.**

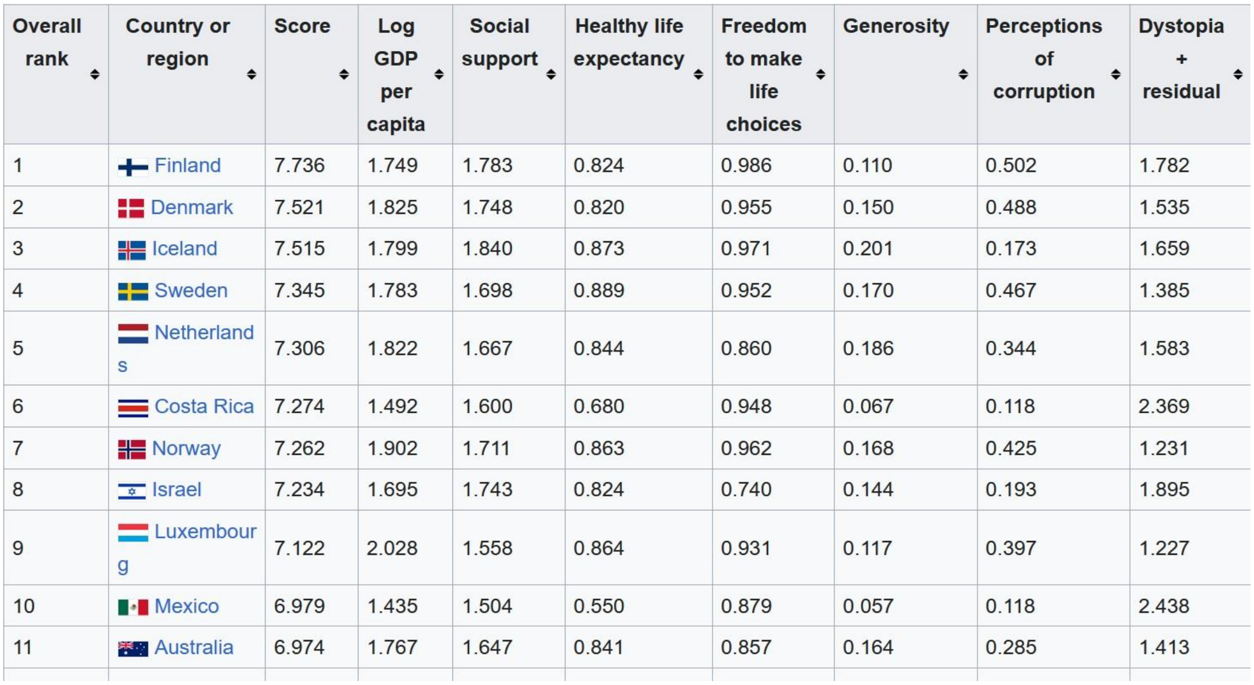
The World Happiness Report is a publication that contains articles and rankings of national happiness, based on respondent ratings of their own lives, which the report also correlates with various quality of life factors.

Data is collected from people in over 150 countries. Each variable measured reveals a populated-weighted average score on a scale running from 0 to 10 that is tracked over time and compared against other countries. These variables currently include:

- real GDP per capita
- social support
- healthy life expectancy
- freedom to make life choices
- generosity
- perceptions of corruption

Each country is also compared against a hypothetical nation called Dystopia. Dystopia represents the lowest national averages for each key variable and is, along with residual error, used as a regression benchmark. The six metrics are used to explain the estimated extent to which each of these factors contribute to increasing life satisfaction when compared to the hypothetical nation of Dystopia, but they themselves do not have an effect on the total score reported for each country.

Data as a table format:



Overall rank	Country or region	Score	Log GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Dystopia + residual
1	Finland	7.736	1.749	1.783	0.824	0.986	0.110	0.502	1.782
2	Denmark	7.521	1.825	1.748	0.820	0.955	0.150	0.488	1.535
3	Iceland	7.515	1.799	1.840	0.873	0.971	0.201	0.173	1.659
4	Sweden	7.345	1.783	1.698	0.889	0.952	0.170	0.467	1.385
5	Netherlands	7.306	1.822	1.667	0.844	0.860	0.186	0.344	1.583
6	Costa Rica	7.274	1.492	1.600	0.680	0.948	0.067	0.118	2.369
7	Norway	7.262	1.902	1.711	0.863	0.962	0.168	0.425	1.231
8	Israel	7.234	1.695	1.743	0.824	0.740	0.144	0.193	1.895
9	Luxembourg	7.122	2.028	1.558	0.864	0.931	0.117	0.397	1.227
10	Mexico	6.979	1.435	1.504	0.550	0.879	0.057	0.118	2.438
11	Australia	6.974	1.767	1.647	0.841	0.857	0.164	0.285	1.413

In the above figure there are more number of numerical data columns and one categorical data column. When there are multiple numerical values we need to use the method **co-relation** relationship between them. For identifying the corelation between two numerical value we need to use the chart. **i.e., Scatter Plot chart**

Co-relation:

Correlation is a statistical measure that describes the extent of a linear relationship between two variables. It indicates how closely two variables move together, either in the same direction (positive correlation) or in opposite directions (negative correlation). A correlation coefficient, a value between -1 and 1, quantifies this relationship, where 1 signifies a perfect positive correlation, -1 a perfect negative correlation, and 0 no correlation. It is crucial to remember that correlation does not imply causation; just because two variables are correlated does not mean one causes the other.

Types of Correlation

- Positive Correlation:

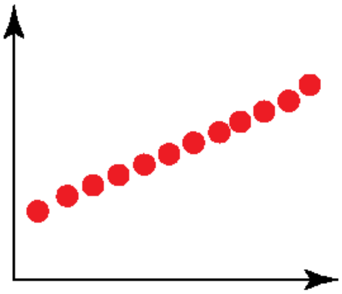
Both variables change in the same direction. For example, as a baby's length increases, its weight also tends to increase.

- Negative Correlation:

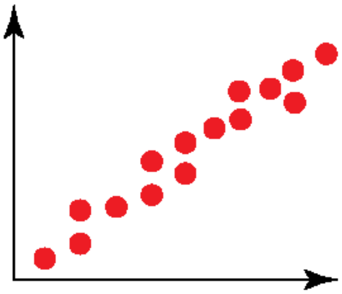
The variables change in opposite directions. For instance, as elevation increases, air pressure decreases.

- No Correlation:

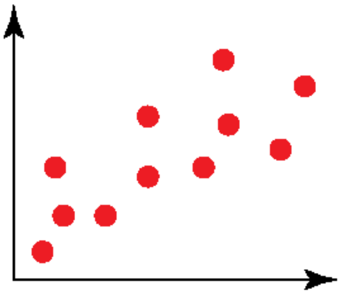
There is no discernible linear relationship between the variables. Their values do not change in a consistent way.



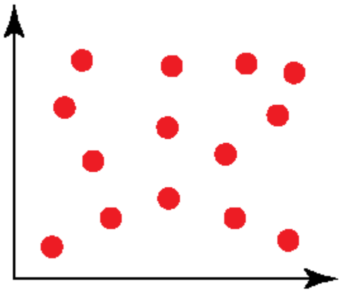
Perfect
Positive
Correlation



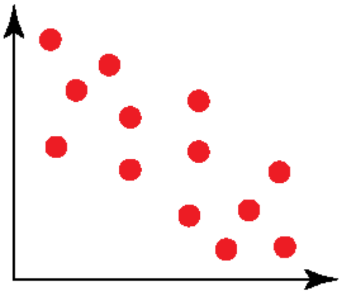
Strong
Positive
Correlation



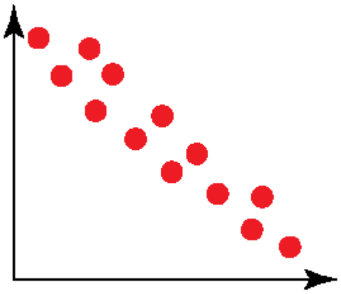
Weak
Positive
Correlation



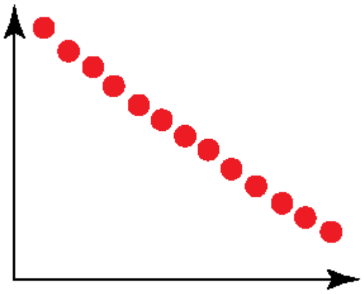
No
Correlation



Weak
Negative
Correlation



Strong
Negative
Correlation



Perfect
Negative
Correlation

About Columns and their relations and how they calculated for each column:

Overall Rank:

- What it is: The position of the country in the global happiness ranking.
- How calculated:
  Sort all countries by their Score (highest → lowest).
  The country with the highest score gets Rank 1.

Country or Region:

- Just the name of the country/region being measured.
- No calculation here.

Score (Happiness Score)

- What it is: The final "Happiness Index" of each country.
- How calculated:

Score=Log GDP per capita+Social support+Healthy life expectancy+Freedom to make life choices+ Generosity+ Perceptions of corruption+ (Dystopia + residual)

This is the sum of contributions from 6 main factors + a baseline value (Dystopia + residual).

Log GDP per capita

- What it is: Measures standard of living using income per person, adjusted by purchasing power (PPP).
- How calculated:

Log GDP per capita= ln (PPP-adjusted GDP per person)

- Then it is normalized and scaled between 0–2 for comparability.

Social Support

- What it is: Measures whether people have family/friends to rely on in times of trouble.
- How calculated:
  Taken from the Gallup World Poll question:
  "If you were in trouble, do you have relatives or friends you can count on?"
- Value is the proportion of people answering "Yes", averaged across the country (range: 0–2 after scaling).

Healthy Life Expectancy

- What it is: Life expectancy at birth, adjusted for quality of health.
- How calculated:
  Based on WHO and UN health data.
  Countries with longer, healthier lives score higher.
- Scaled to fit the happiness score contribution (0–1 scale approx).

Freedom to Make Life Choices

- What it is: Captures people's sense of freedom in life decisions.
- How calculated:
  Survey-based (Gallup): "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?"
  Proportion of "satisfied" responses, rescaled (0–1).

Generosity

- What it is: Reflects how much people donate/help others.
- How calculated:
  Gallup survey: "Have you donated money to a charity in the past month?"
  Adjusted for income levels → produces a relative generosity score (0–1).

Perceptions of Corruption

- What it is: Measures how much corruption people believe exists in government/business.
- How calculated:
  Gallup survey: "Is corruption widespread throughout the government/businesses?"
  Value = reverse-coded so that higher values = less corruption.
  Range: ~0–0.5.

Dystopia + Residual

- What it is: A baseline that ensures scores remain positive and comparable.
- Dystopia = a hypothetical worst-case country (lowest observed values in all factors).
- Residual = error term that captures differences not explained by the six main factors.
- How calculated:
  Dystopia + Residual= Score−

(GDP + Social support + Life expectancy + Freedom + Generosity + Corrup