CMSC178DA

*Assignment 1*

# Retrieval of Multivariate Dataset and Performing Descriptive Multivariate Analysis

Submitted by:

Peladas, Daenielle Rai

Section B (TF 10:30 AM - 12:00 PM)

# The Multivariate Dataset

---

## 🔗World Happiness Report 2024

### Context

The World Happiness Report is a landmark survey of the state of global happiness. The report continues to gain global recognition as governments, organizations and civil society increasingly use happiness indicators to inform their policy-making decisions. Leading experts across fields – economics, psychology, survey analysis, national statistics, health, public policy and more – describe how measurements of well-being can be used effectively to assess the progress of nations. The reports review the state of happiness in the world today and show how the new science of happiness explains personal and national variations in happiness.

### Content

The happiness scores and rankings use data from the Gallup World Poll . The columns following the happiness score estimate the extent to which each of six factors – economic production, social support, life expectancy, freedom, absence of corruption, and generosity – contribute to making life evaluations higher in each country than they are in Dystopia, a hypothetical country that has values equal to the world's lowest national averages for each of the six factors. They have no impact on the total score reported for each country, but they do explain why some countries rank higher than others.

*Context and content was derived word for word from the Kaggle page description.*

# World Happiness Report Dataset Attributes

---

| ATTRIBUTE/VARIABLE | DESCRIPTION |
|---|---|
| Country Name | Name of country. |
| Ladder Score | A metric measured in 2024 by asking the sampled people the question: 'How would you rate your happiness on a scale of 0 to 10 where 10 is the happiest?' |
| Upper Whisker | Upper Confidence Interval of the Happiness Score. |
| Lower Whisker | Lower Confidence Interval of the Happiness Score. |
| GDP per Capita | Level of economic output that contributes to a country's happiness score. |
| Social Support | Strength of social relationships that contributes to a country's happiness score. |
| Health Life Expectancy | Average life expectancy at birth that contributes to a country's happiness score. |
| Freedom to Make Life Choices | Level of personal freedom that contributes to a country's happiness score. |
| Generosity | Prevalence of charitable behavior that contributes to a country's happiness score. |
| Perceptions of Corruption | Perception of corruption in government and business that contributes to a country's happiness score. |
| Dystopia | The theoretical happiness level of a hypothetical country with the lowest possible scores on all six factors used in the index. |

# Multivariate Data Analysis for Central Tendency and Dispersion

---

## Descriptive Summaries

| ATTRIBUTE | MEAN | MEDIAN | VARIANCE | STANDARD DEVIATION | COEFFICIENT OF VARIATION |
|---|---|---|---|---|---|
| Ladder Score | 5.52758042 | 5.785 | 1.370577147 | 1.17071651 | 21.17954731 |
| Upper Whisker | 5.641174825 | 5.895 | 1.334043821 | 1.155008148 | 20.47460296 |
| Lower Whisker | 5.413972028 | 5.674 | 1.409284394 | 1.187132846 | 21.92720686 |
| GDP per Capita | 1.378807143 | 1.4315 | 0.1807085884 | 0.425098328 | 30.83087654 |
| Social Support | 1.134328571 | 1.2375 | 0.1111003085 | 0.3333171291 | 29.38453085 |
| Health Life Expectancy | 0.5208857143 | 0.5495 | 0.02719944008 | 0.1649225275 | 31.661941 |
| Freedom to Make Life Choices | 0.6206214286 | 0.641 | 0.02640358947 | 0.1624918135 | 26.18211458 |
| Generosity | 0.1462714286 | 0.1365 | 0.005393623638 | 0.07344129382 | 50.20891266 |
| Perceptions of Corruption | 0.1541214286 | 0.1205 | 0.01593607867 | 0.1262381823 | 81.90826122 |
| Dystopia | 1.575914286 | 1.6445 | 0.2888616473 | 0.5374585075 | 34.10455203 |

# Narrative Text

### Import library

Import the `pandas` library for data manipulation and analysis.

```python
import pandas as pd
```

### Load dataset to dataframe

Define the `path` containing the World Happiness Report (WHR) 2024 dataset and use `read_csv` to load the said dataset to a pandas DataFrame labeled `df`.

```python
path = "C:/Users/daeni/Desktop/LOVE/Academics/CMSC178DA/Assignment/Retrieval of Multivariate Dataset and Performing Descriptive Multivariate Analysis/World Happiness Report/WHR2024.csv"
df = pd.read_csv(path)
```

### Get dataset attributes

Utilize the `df.columns` attribute to gain a clear understanding of the dataset's structure. This provides a list containing the original column names.

```python
df.columns
```

```
Index(['Country name', 'Ladder score', 'upperwhisker', 'lowerwhisker',
       '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',
       'Dystopia + residual'],
      dtype='object')
```

### Update WHR2024's `df` columns

For improved readability, these names are then renamed and stored in a separate list called `headers`. Finally, the headers list is used to update the column names within the DataFrame itself.

```python
headers = ['Country Name', 'Ladder Score', 'Upper Whisker', 'Lower Whisker',
'GDP per Capita', 'Social Support', 'Health Life Expectancy', 'Freedom to
Make Life Choices', 'Generosity', 'Perceptions of Corruption', 'Dystopia']
df.columns = headers
```

### Create new DataFrame `df_attr` containing WHR2024's attributes

Create a new DataFrame called `df_attr` to contain the columns `Attributes/Variable` for the attributes of `df` and `Description` for their corresponding descriptions. First, a list called `columns` was made to contain the headers `df_attr`. Then, `df_attr` was initialized as a DataFrame containing the just created columns.

```python
columns = ["Attribute/Variable", "Description"]
df_attr = pd.DataFrame(columns=columns)
```

### Set the content of the columns in `df_attr`

Insert the columns from `df` into the `Attribute/Variable` column of `df_attr` DataFrame by setting its values accordingly.

```python
df_attr["Attribute/Variable"] = df.columns
```

As for the description of each attribute, they were added manually since the original Kaggle dataset did not contain any. The description is contained in a list called `desc` and is then placed into the `Description` column.

```python
desc = ["Name of country.", "A metric measured in 2024 by asking the sampled
people the question: 'How would you rate your happiness on a scale of 0 to
10 where 10 is the happiest?'", "Upper Confidence Interval of the Happiness
```

```
Score.", "Lower Confidence Interval of the Happiness Score.", "Level of
economic output that contributes to a country's happiness score.", "Strength
of social relationships that contributes to a country's happiness score.",
"Average life expectancy at birth that contributes to a country's happiness
score.", "Level of personal freedom that contributes to a country's
happiness score.", "Prevalence of charitable behavior that contributes to a
country's happiness score.", "Perception of corruption in government and
business that contributes to a country's happiness score.", "The theoretical
happiness level of a hypothetical country with the lowest possible scores on
all six factors used in the index."]

df_attr["Description"] = desc

df_attr
```

| | Attribute/Variable | Description |
|---|---|---|
| **0** | Country Name | Name of country. |
| **1** | Ladder Score | A metric measured in 2024 by asking the sampled people the question: 'How would you rate your happiness on a scale of 0 to 10 where 10 is the happiest?' |
| **2** | Upper Whisker | Upper Confidence Interval of the Happiness Score. |
| **3** | Lower Whisker | Lower Confidence Interval of the Happiness Score. |
| **4** | GDP per Capita | Level of economic output that contributes to a country's happiness score. |
| **5** | Social Support | Strength of social relationships that contributes to a country's happiness score. |
| **6** | Health Life Expectancy | Average life expectancy at birth that contributes to a country's happiness score. |
| **7** | Freedom to Make Life Choices | Level of personal freedom that contributes to a country's happiness score. |
| **8** | Generosity | Prevalence of charitable behavior that contributes to a country's happiness score. |
| **9** | Perceptions of Corruption | Perception of corruption in government and business that contributes to a country's happiness score. |
| **10** | Dystopia | The theoretical happiness level of a hypothetical country with the lowest possible scores on all six factors used in the index. |

### Export `df_attr` to an Excel spreadsheet

Using the `to_excel()` function, the `df_attr` DataFrame is then converted into spreadsheet format for later reference. The excel file can be found through this link.

```python
df_attr.to_excel("AttributeDescription.xlsx")
```

### Calculate summary statistics of the WHR2024 `df` DataFrame

Using the `describe()` function from `pandas`, the summary statistics of all numerical variables are then obtained easily. This is then stored in a variable called `desc_stats`.

```python
desc_stats = df.describe()
desc_stats
```

| | Ladder Score | Upper Whisker | Lower Whisker | GDP per Capita | Social Support | Health Life Expectancy | Freedom to Make Life Choices | Generosity | Perceptions of Corruption | Dystopia |
|---|---|---|---|---|---|---|---|---|---|---|
| **count** | 143.000000 | 143.000000 | 143.000000 | 140.000000 | 140.000000 | 140.000000 | 140.000000 | 140.000000 | 140.000000 | 140.000000 |
| **mean** | 5.527580 | 5.641175 | 5.413972 | 1.378807 | 1.134329 | 0.520886 | 0.620621 | 0.146271 | 0.154121 | 1.575914 |
| **std** | 1.170717 | 1.155008 | 1.187133 | 0.425098 | 0.333317 | 0.164923 | 0.162492 | 0.073441 | 0.126238 | 0.537459 |
| **min** | 1.721000 | 1.775000 | 1.667000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | -0.073000 |
| **25%** | 4.726000 | 4.845500 | 4.606000 | 1.077750 | 0.921750 | 0.398000 | 0.527500 | 0.091000 | 0.068750 | 1.308250 |
| **50%** | 5.785000 | 5.895000 | 5.674000 | 1.431500 | 1.237500 | 0.549500 | 0.641000 | 0.136500 | 0.120500 | 1.644500 |
| **75%** | 6.416000 | 6.507500 | 6.319000 | 1.741500 | 1.383250 | 0.648500 | 0.736000 | 0.192500 | 0.193750 | 1.881750 |
| **max** | 7.741000 | 7.815000 | 7.667000 | 2.141000 | 1.617000 | 0.857000 | 0.863000 | 0.401000 | 0.575000 | 2.998000 |

### Swap rows and columns of `desc_stats`

While the `describe()` function provides a good overview of the data, it doesn't include statistics like variance or coefficient of variation. To simplify calculations of these additional statistics, we can transpose the `desc_stats` DataFrame. This makes it easier to reference specific statistics for each column during calculations.

```python
desc_stats = desc_stats.T
```

### Removing and rearranging the columns in `desc_stats`

To filter the columns in the `desc_stats` DataFrame to the desired statistical measures, the current columns of the descriptive statistics are first obtained through getting the `desc_stats columns` attribute.

```python
desc_stats.columns
```

```
Index(['count', 'mean', 'std', 'min', '25%', '50%', '75%', 'max'], dtype='object')
```

A list named `cols_to_remove` containing the names of unwanted columns was created. Then, the `drop()` function was used on the DataFrame, specifying `cols_to_remove` as an argument. This efficiently removes the unwanted columns, resulting in a cleaner DataFrame for further analysis.

```python
cols_to_remove = ['count', 'min', '25%', '75%', 'max']
desc_stats = desc_stats.drop(columns=cols_to_remove)
```

The remaining columns are then rearranged following the provided set from the assignment by creating a subset of the columns into the correct arrangement being `mean`, `50%`, and `std` accordingly. From further reflection, this could have been done directly to subset the `desc_stats` DataFrame to include only the wanted columns. Thus, removing and rearranging the DataFrame to fit the assignment needs.

```python
desc_stats = desc_stats[["mean", "50%", "std"]]
```

**Deriving and inserting new values into `desc_stats`**

Since the standard deviation ($\sigma$) is just the square root of the variance, the variance must have the following formula:

$$\sigma^2$$

The DataFrame is updated with a new column named `variance` using the `insert()` function. This column is inserted at the second position and holds the squared `std` of the existing data.

```python
desc_stats.insert(loc=2, column="variance", value=desc_stats["std"]**2)
```

The same flow is applied to insert the value for the coefficient of variation in the DataFrame which can be derived from the following formula:

$$cv = \sigma/\bar{x} \cdot 100$$

```python
desc_stats.insert(loc=4, column="coef",
value=(desc_stats["std"]/desc_stats["mean"])*100)
```

### Updating `desc_stats` headers for readability

For improved readability, the column names of desc_stats are renamed and stored in a variable called `desc_stats_headers`. Finally, the list is used to update the column names within the DataFrame itself.

```python
desc_stats_headers = ["Mean", "Median", "Variance", "Standard Deviation",
"Coefficient of Variation"]
desc_stats.columns = desc_stats_headers
desc_stats
```

### Export `desc_stats` to an Excel spreadsheet

Using the `to_excel()` function, the `desc_stats` DataFrame is then converted into spreadsheet format for later reference. The excel file can be found through this link.

```python
desc_stats.to_excel('DescriptiveSummaryWHR2024.xlsx')
desc_stats
```

| | Mean | Median | Variance | Standard Deviation | Coefficient of Variation |
|---|---|---|---|---|---|
| **Ladder Score** | 5.527580 | 5.7850 | 1.370577 | 1.170717 | 21.179547 |
| **Upper Whisker** | 5.641175 | 5.8950 | 1.334044 | 1.155008 | 20.474603 |
| **Lower Whisker** | 5.413972 | 5.6740 | 1.409284 | 1.187133 | 21.927207 |
| **GDP per Capita** | 1.378807 | 1.4315 | 0.180709 | 0.425098 | 30.830877 |
| **Social Support** | 1.134329 | 1.2375 | 0.111100 | 0.333317 | 29.384531 |
| **Health Life Expectancy** | 0.520886 | 0.5495 | 0.027199 | 0.164923 | 31.661941 |
| **Freedom to Make Life Choices** | 0.620621 | 0.6410 | 0.026404 | 0.162492 | 26.182115 |
| **Generosity** | 0.146271 | 0.1365 | 0.005394 | 0.073441 | 50.208913 |
| **Perceptions of Corruption** | 0.154121 | 0.1205 | 0.015936 | 0.126238 | 81.908261 |
| **Dystopia** | 1.575914 | 1.6445 | 0.288862 | 0.537459 | 34.104552 |

# Interpretation

---

While all the coefficients of variation (CV) are greater than 1, indicating high variability across most attributes, the standard deviation helps us pinpoint the specific attributes with the most significant variations. These top three are:

1. Lower Whisker ($\sigma = 1.187133$)
2. Ladder Score ($\sigma = 1.170717$)
3. Upper Whisker ($\sigma = 1.155008$)

Even though the CVs suggest high overall variability, the standard deviations reveal a few attributes with surprising consistency. The top three most consistent attributes, based on standard deviation, are:

1. Generosity ($\sigma = 0.073441$)
2. Perceptions of Corruption ($\sigma = 0.126238$)
3. Freedom to Make Life Choices ($\sigma = 0.162492$)

Standard deviation serves as a measure of how spread out the data points are around the mean. A large standard deviation signifies high variance, meaning data points are widely dispersed. Conversely, a low standard deviation indicates low variance, with data points clustered tightly around the mean. The coefficient of variation (cv) complements this analysis by providing a relative measure of variability. A CV greater than 1 suggests the standard deviation is larger than the mean, implying high variability relative to the average value.

In this dataset, all CVs being greater than 1 tells us that the standard deviation is consistently larger than the mean for each attribute. This can be attributed to several factors stemming from the dataset's nature: a worldwide report on general happiness across diverse attributes. The large population and inherently diverse sample contribute significantly to the observed variations in the statistical summary. Additionally, the potential subjectivity in the collected values, where fluctuations are likely, can further influence the results.