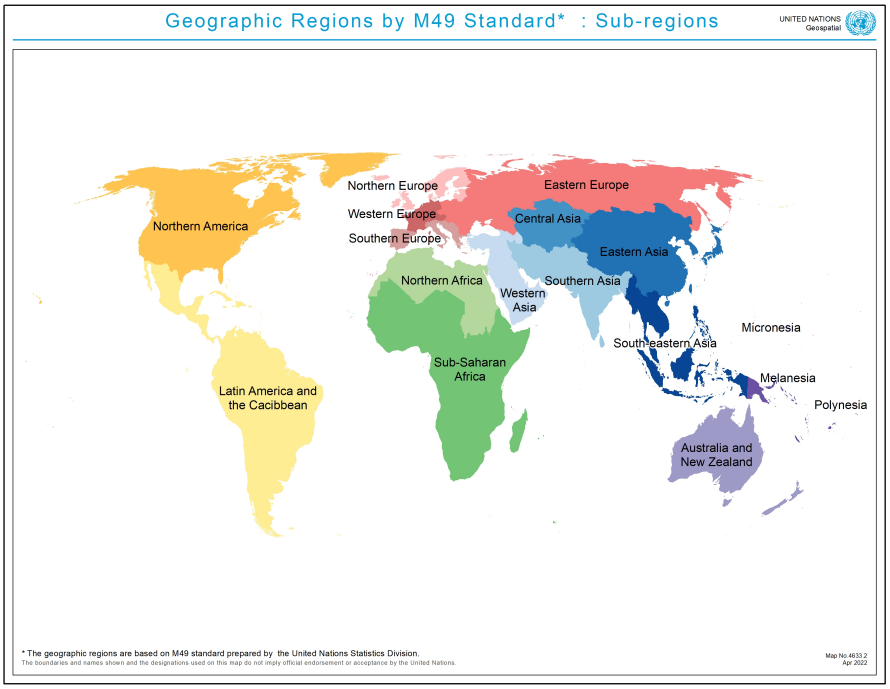
Descriptive Statistics, fall 2024

MAT-203:00010 and MAT-203:00020

Every year since 2012 the UN release the World Happiness Report (WHR). The report of this year can be downloaded here.

In these assessments you are going to explore the data used to create such report. The data is of free access, and can be downloaded here. The dataset is in the section `World Happiness Report 2024, Data for Table 2.1`. There is also a statistical appendix in the section `World Happiness Report 2024, Statistical Appendix 1 for Chapter 2`, explaining the meaning of each variable in the dataset. For your convenience, both documents are available also in the repository of the course, being `DataWhr2024.csv` and `AppendixWhr2024.pdf` respectively.

On the other hand, the UN M49 is a standard used by the UN to aggregate countries and territories in 6 regions–Africa, Antarctica, Americas, Asia, Europe, and Oceania–which can be split in 17 sub-regions, as it is shown in the next figure.



You can find more information about the UN M49 here, and download it here. For your convenience, the table with the relation between the countries and territories, and the sub-regions and regions of the UN M49, can be found in the repository of the course, namely `UnM49.csv`.

Even while both assessments, MAT-203:00010 and MAT-203:00020, use the same data sets, they are different assessments. Thus, **you have to create two different reports, one for each assessment. The reports should be in PDF. Any report presented in a different**

format, including Jupyter Notebooks (.ipynb), LibreOffice Writer (.odt), and Microsoft Word (.docx) **will not be evaluated**. Your name and student ID should appear on the first page of the reports.

MAT-203:00010

Download the datasets `DataWhr2024.csv` and `UnM49.csv`, the appendix `AppendixWhr2024.pdf`, and the notebook `Whr2024Descriptive.ipynb`. The first cells of the notebook merge the datasets corresponding with the WHR 2024 and the UN M49 (you might need to modify the directory used to read the corresponding documents).

1. (4 points) Read the first section of the appendix. Explain in your own words and briefly the variables `Life Ladder` and `Positive affect`.

   Use the data corresponding to the year 2023 for the next bullets.

2. (4 points) Execute the next instruction

   ```
   Dat2023.describe()
   ```

   What does it do?

3. (4 points) Based on the variable `Life Ladder`, which is the happiest country in the world? which is the unhappiest?

4. (4 points) What is the median of the healthy life expectancy at birth?

5. (4 points) What is the first quartile of the negative affect?

6. (40 points) Compute the following descriptive statistics for the variable `Life Ladder`.

   (a) (8 points) Compute the quartiles, and present them in a table similar to the next one.

	Value
Q_1	
Q_2	
Q_3	

   (b) (8 points) Compute the next statistics of central tendency: median, mean, geometric mean, harmonic mean, trimmed mean, interquartile mean, winsorized mean. For the trimmed mean and the winsorized mean, you have freedom to determine the parameters used as thresholds. Present them in a table similar to the next one, indicating with 1 if the statistic is robust and 0 if it is not.

	Value	Robust
Median		
Mean		
Geometric Mean		
Harmonic Mean		
Trimmed Mean		
Interquartile Mean		
Winsorized Mean		

   Do you observe any difference between robust and non-robust statistics?

(c) (8 points) Let s_1 and s_2 be the sample standard deviation with zero degrees of freedom and one degree of freedom, respectively. Compute the next statistics of dispersion: s_1, s_2, range, IQR, MAD, AAD. Compute their value after correcting for the bias, and indicate if they are robust or not. Present your results in a table similar to the next one.

	Value	Bias corrected	Robust
s_1			
s_2			
Range			
IQR			
MAD			
AAD			

Do you observe any difference between robust and non-robust statistics?

(d) (8 points) Compute the coefficient of skewness based on central moments, g_1, and k-statistics, G_1. Present them in a table similar to the next one.

	Value
g_1	
G_1	

Interpret these statistics.

(e) (8 points) Compute the coefficient of kurtosis based on central moments, g_2, and k-statistics, G_2. Compute also the excess of kurtosis. Present them in a table similar to the next one.

	Value	Excess kurtosis
g_2		
G_2		

Interpret these statistic.

7. (40 points) Choose a country whose name starts with the same letter that your name, and choose a variable different of `Life Ladder`. Repeat the instructions of the bullet (6.) for the variable that you chose and the subregion of the country that you chose. Include in the report the name of such country.

MAT-203:00020

Download the datasets `DataWhr2024.csv` and `UnM49.csv`, the appendix `AppendixWhr2024.pdf`, and the notebook `Whr2024Descriptive.ipynb`. The first cells of the notebook merge the datasets corresponding with the WHR 2024 and the UN M49 (you might need to modify the directory used to read the corresponding documents).

1. (10 points) What do the next instructions do?

```
Variable = 'Life Ladder'
DatThroughTime = Dat[['year', 'Continent', Variable]].groupby(['year', '
                                        Continent']).mean().reset_index()
```

   Present the result in an appropriate graph, comment what you observe.

   Use the data corresponding to the year 2023 for the next bullets.

2. (10 points) Present a scatter plot between `Life Ladder` (in the vertical axis), and `Social Support` (in the horizontal axis). Use the function `regplot` of the `seaborn` library to add a regression line. What do you observe?

3. (20 points) Present in a chart the 5 happiest countries in the world and the 5 unhappiest. It must be possible to know the value (at least an approximate value) for the variable `Life Laddder` from your plot.

4. (60 points) Choose a country whose name starts with the same letter that your name, and choose a variable different of `Life Ladder`. Present a kernel density estimator (KDE) for the variable that you chose and the subregion of the country that you chose. Include in the report the name of such country.

   Let `Mu` and `Sigma` be the mean and any unbiased statistic for the standard deviation, for such variable and subregion. Add to the same graph a prediction interval, whose limits can be calculated as

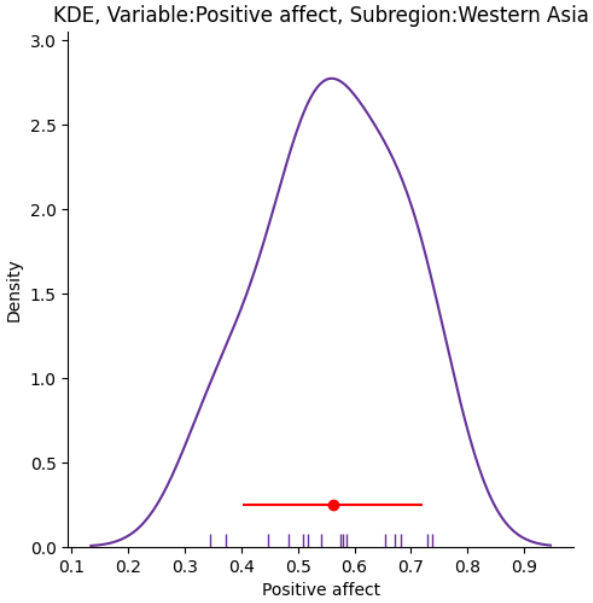
```
LowerLimit = norm.ppf(ALPHA/2, Mu, (np.sqrt(1+1/N))*Sigma)
UpperLimit = norm.ppf(1-ALPHA/2, Mu, (np.sqrt(1+1/N))*Sigma)
```

   where `N` is the number of observations in the data set.

   You have freedom to choose the level `ALPHA` that you wish. Show on the prediction interval the value of `Mu`.

   Show the observed values of the variable as a "rug" on the horizontal axis.

   Your plot should look similar to the next one.


KDE, Variable:Positive affect, Subregion:Western Asia
Density
Positive affect
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9