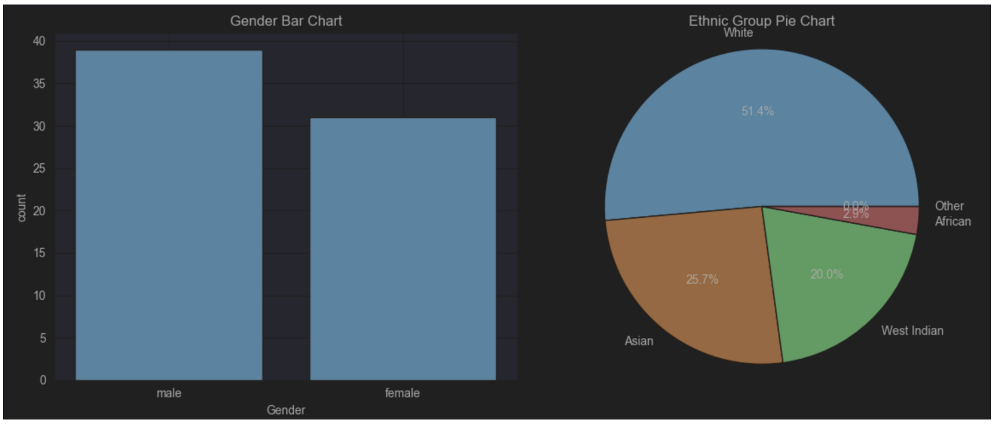
Part One: Statistical Analysis

Exercise 1.1

a) Make plot of gender in Bar chart and ethnic group in pie diagram.



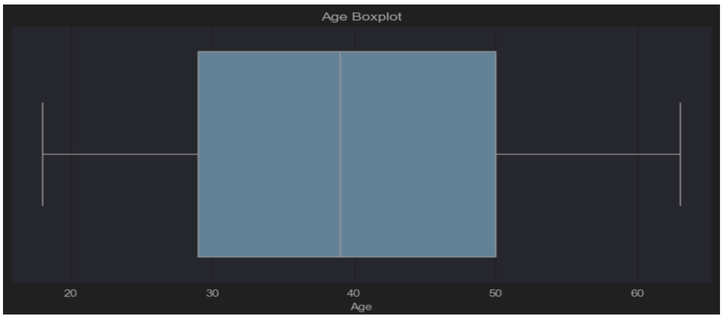
From the gender bar chart, it's observable that there are slightly more males (around thirty-seven individuals) than females (around thirty-one individuals) represented in the dataset.

From the ethnic group pie diagram, it's possible to observe that around half of the population sample has White ethnicity. One fourth is of Asian ethnicity and one fifth of West Indian ethnicity. While African ethnicity is two percent represented and other ethnicities are not represented.

b) Make a five number summary (max, min, median, the first and third quartiles) of Age then a box-plot.

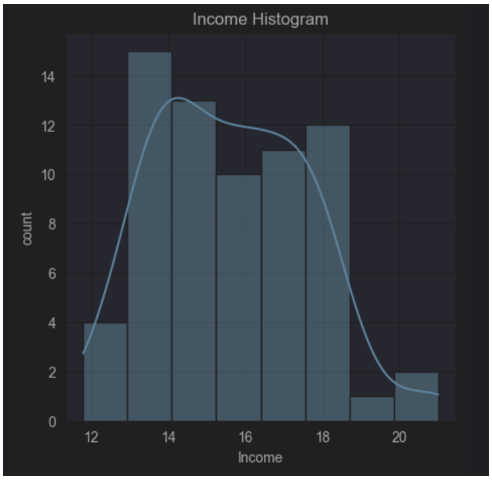
The five number summary for the age feature in the dataset is: maximum age is 63 years; minimum age is 18 years; median age is 39 years; first quartile (25%) of age distribution is 29 years; third quartile (75%) of age distribution is 50 years.

Bellow follows the boxplot of age, where the five numbers mentioned above are represented visually.



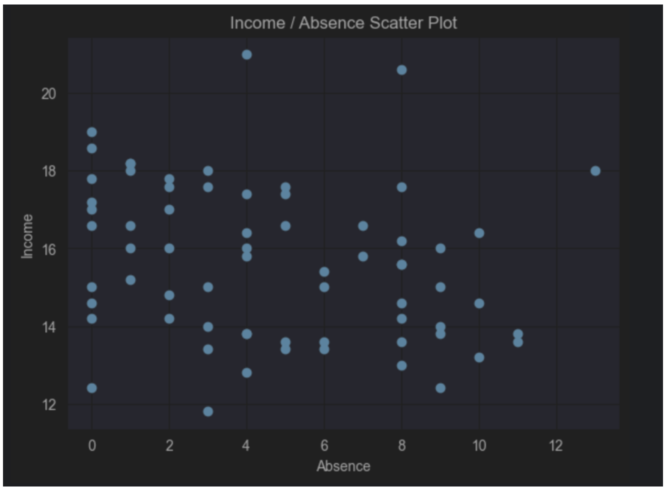
c) Find the mean and standard deviation of income, and as well as histogram of it.

The mean of gross annual income before tax in 1000£ is 15.6382 (rounded in four decimals), and standard deviation is 1.9812 (rounded in four decimals). Bellow is the histogram of the income feature. It is possible to observe that it doesn't follow a normal distribution, having a high concentrated group around 14 and another around 17.



Exercise 1.2 Consider the variables income and absence.

a) Make scatter plot to visualize the relationship between them



From the scatter plot of income and absence, it is possible to infer that there is a weak negative relationship between the variables, since there is a decrease in absence as the income increases.
There is also a bigger concentration of data points in the lower income from zero to five and less concentration in the higher income from nine to ten, meaning that at a lower income the variation of absence tends to be bigger.

b) Find the simple regression model where income is dependent variable and absence is independent variable. What is your determination coefficient?

The determination coefficient is 0.0624 (rounded to 4 decimals) which means that only six percent of the variance in income can be explained by absence. Meaning that absence is not a strong predictor of income, confirming what was observed in the previous exercise.

Exercise 1.3 Study the multiple regression model with satis as dependent variable and commit, autonom, income, skill, rated quality, age, years as independent variables.

a) Which variables among them do NOT have any significant impact on satis?

From the statistical significance table generated from fitting the multiple regression model presented below, it is possible to conclude that the independent variables quality, age and years do not have a significant impact on job satisfaction. Since their p-value is above 0.05, is not possible to reject the null hypothesis that there is no relationship between an independent and a dependent variable.

Const	0.013147
Commit	0.000051
Autonom	0.000053
Income	0.010946
Skill	0.003559
Qual	0.155518
Age	0.735361
Years	0.660962

b) Find a simpler multiple regression model with satis as dependent variable by deleting all those non-impact variables.

A simpler multiple regression model was made with job satisfaction as dependent variable and commitment, autonomy, income and skill as independent variables and it was observed that all the p-values stayed low, as observed in the table below. Meaning that the statistical significance between the independent and dependent variable remained significant.

Const	0.000018
Commit	0.000005
Autonom	0.000281
Income	0.003689

Exercise 1.4 Find confidence interval of job satisfaction and also confidence interval of difference in job satisfaction between men and women.

The confidence interval of average job satisfaction with 95% confidence level is between 10.0386 and 11.6379 (rounded to fours decimals). This means that we can be 95% confident that the true average job satisfaction for the entire population lies between the aforementioned results.

The confidence interval of the difference between the average job satisfaction between males and females with 95% confidence level falls between 1.7727 and -1.3002 (rounded to four decimals). This interval gives a range within which the true difference in average job satisfaction between men and women likely lies.
Since the interval contains zero, there is no statistically significant difference in job satisfaction between men and women at the chosen confidence level.

Exercise 1.5 Using the Mann-Whitney-Wilcoxon test to see if there is any significance in skill between man and woman, and compare the result with confidence interval for the difference.

The resulting p-value of the Mann-Whitney-Wilcoxon test to compare if there is any significance difference in the distributions of skill between man and woman was 0.1173 (rounded to four decimals). Since the p-value is greater than 0.05, we fail to reject the null hypothesis that the sample distributions are equal.

The confidence interval between the difference of skill between man and woman falls between 2.2044 and -1.3756 (rounded to four decimals). Since the interval contains zero, there is no statistically significant difference in job skill between men and women at the 95% confidence level.

With these two results we can infer that the distribution and the average skill between woman and man is similar.

Exercise 1.6 Using the Kruskal-Wallis' test to see if there is any significance in absence among ethnic group, and compare the result with One-Way ANOVA analysis.

The resulting p-value of the Kruskal-Wallis test to see if there is any significant difference in the median of absence among ethnic groups was 0.3324 (rounded to 4 decimals). Since the p-value for the test is greater than 0.05, we cannot reject the null hypothesis that the median absence is the same for all ethnic groups.

The resulting p-value of the One-Way ANOVA test to see if there is any significant difference in the average of absence among ethnic groups was 0.3357 (rounded to 4 decimals). Since the p-value for the test is greater than 0.05, we cannot reject the null hypothesis that the means of all the ethnic groups are equal.

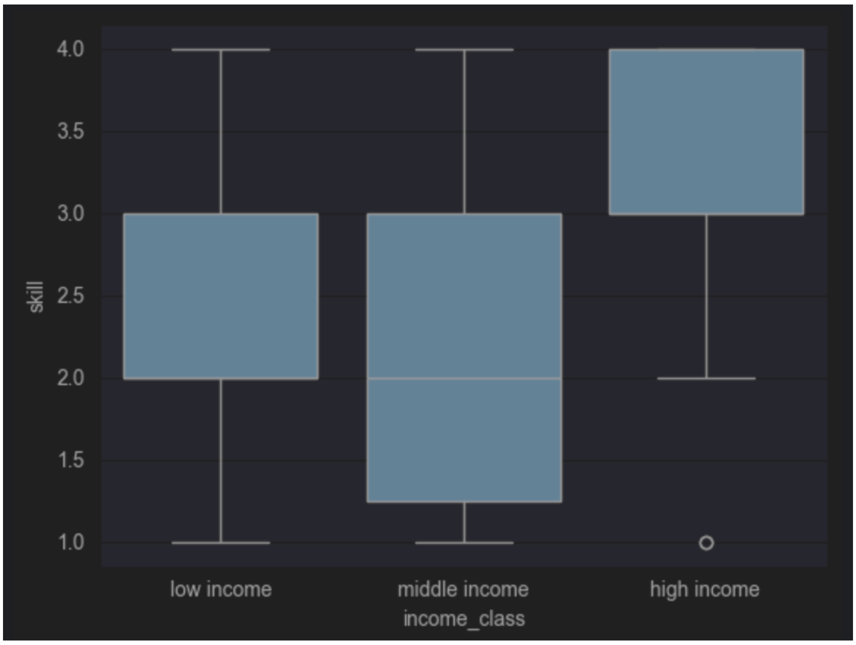
For the above-mentioned tests, the ethnicity “others” was not considered since it was empty.

Exercise 1.7 Re-code the variable “income” into “income_class “ with proper choice of limits of class classification thereafter investigate if there is any significant relationship between income_class and skill.

The variable income was binned into: low income from the minimum income to the first quartile (25%); middle income from the first quartile to the third quartile (75%); and high income from the third quartile to the maximum value.

Performing the Kruskal-Wallis’ test to see if there is any significant difference between the average skill between the different income classes aforementioned, the resulting p-value was 0.0212 (rounded to four decimals). Since the p-value for this test is less than 0.05, we reject the null hypothesis that the means of all the income classes groups are equal.

We can also observe in the boxplot below that the average of the classes of income differ from each other. Thus, confirming the previous conclusion of the test.



From these observations we can conclude that the relationship between income class and skill is generally positive, since there is an increase in skill as the income class increases. Though, as observed in the boxplot, between low-income and middle-income class there is no apparent difference in skill.

Part Two: Statistical Analysis on International Energy Agency Dataset

The chose dataset was “Monthly Electricity Production in GWh [2010-2022]” from the International Energy Agency (IEA) found in kaggle: https://www.kaggle.com/datasets/ccanb23/iea-monthly-electricity-statistics

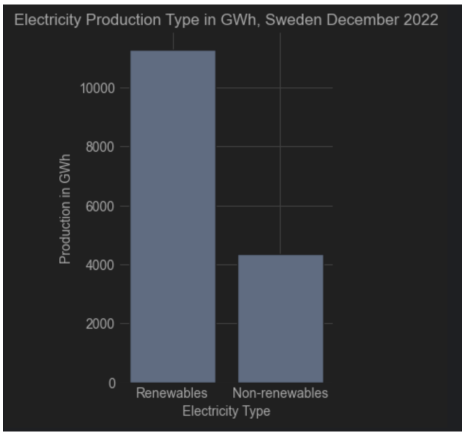
The data includes information about energy production in various countries on a monthly basis from 2010 to 2022. The energy production is measured in gigawatt-hours (GWh) and covers a range of energy products including hydro, wind, solar, geothermal, nuclear, fossil fuels, and others.

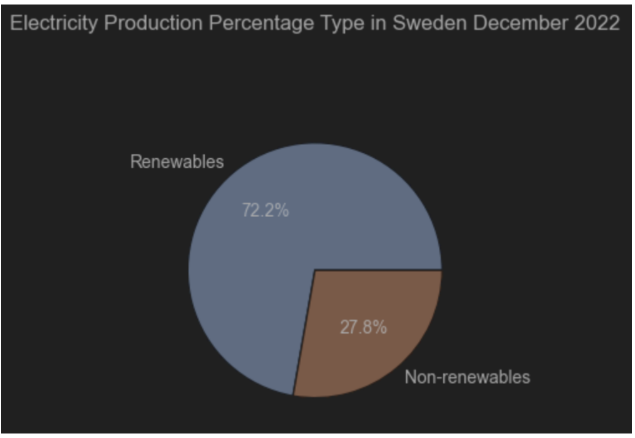
The dataset columns include:

- COUNTRY: Name of the country
- CODE_TIME: A code that represents the month and year (e.g., JAN2010 for January 2010)
- TIME: The month and year in a more human-readable format (e.g., January 2010)
- YEAR: The year of the data point
- MONTH: The month of the data point as a number (1-12)
- MONTH_NAME: The month of the data point as a string (e.g., January)
- PRODUCT: The type of energy product - Hydro, Wind, Solar, Geothermal, Other renewables, Nuclear, Total combustible fuels, Coal, Oil, Natural gas, Combustible renewables, Other combustible non-renewables, Not specified, Net electricity production, Total imports, Total exports, Electricity supplied, Used for pumped storage, Distribution losses, Final consumption, Electricity trade, Renewables, Non-renewables, Others, Other renewables aggregated, Low carbon, Fossil fuels
- VALUE: The amount of electricity generated in gigawatt-hours (GWh)
- DISPLAY_ORDER: The order in which the products should be displayed
- yearToDate: The amount of electricity generated for the current year up to the current month in GWh
- previousYearToDate: The amount of electricity generated for the previous year up to the current month in GWh
- share: The share of the product in the total electricity generation for the country in decimal format

Exercise 2.1 Descriptive statistics analysis for at least two qualitative and quantitative variables.

The first statistical analysis conveyed was the type of energy production (renewable or non-renewable) in Sweden in December 2022. As shown in the graphs bellow the energy production from renewable energy sources was much bigger, at more than 10 000 GW/h, around 72% of total production, than from non-renewables, at around 4 000 GW/h, representing 28%, in December 2022.





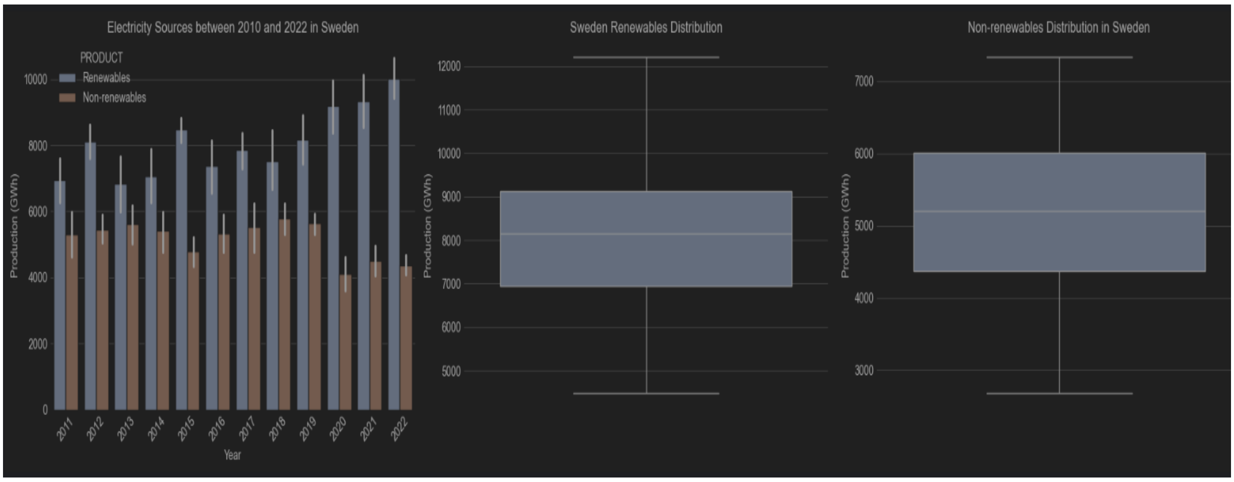
We can further compare these variables by exploring the statistical overview of renewables and non-renewables between the period of 2011 and 2022, as presented in the tables below and in the boxplot.

It is possible to see that renewable energy dominates in absolute production with an average of 8078 GW/h and range between 4487 and 12203 GW/h but is more unstable with a standard deviation of 1599 GW/h. While non-renewable energy production has an average of 5155 GW/h with a range between 2683 and 7329 GW/h and is more stable with a standard deviation of 1076 GW/h.

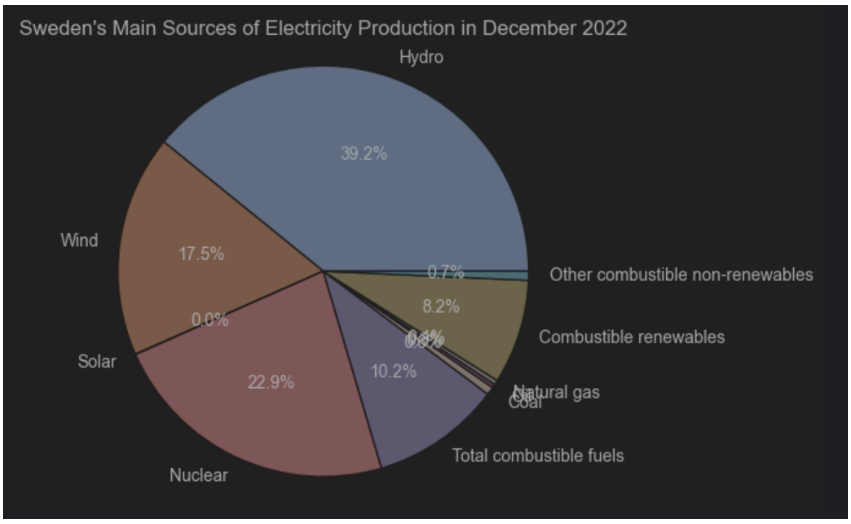
STATISTICAL OVERVIEW RENEWABLE ELECTRICITY (2011-2022)	(GW/H)
COUNT	144
MEAN	8078
STD	1599
MIN	4487
25%	6943
50%	8150
75%	9128
MAX	12203

STATISTICAL OVERVIEW NON-RENEWABLES ELECTRICITY (2011-2022)	(GW/H)
COUNT	144
MEAN	5155
STD	1076
MIN	2683
25%	4375
50%	5203
75%	6002
MAX	7329

It is also possible to infer that since 2011 the renewable energy production in Sweden has been tending to increase while the non-renewable production has been steadily decreasing, as shown in the bar plot below.



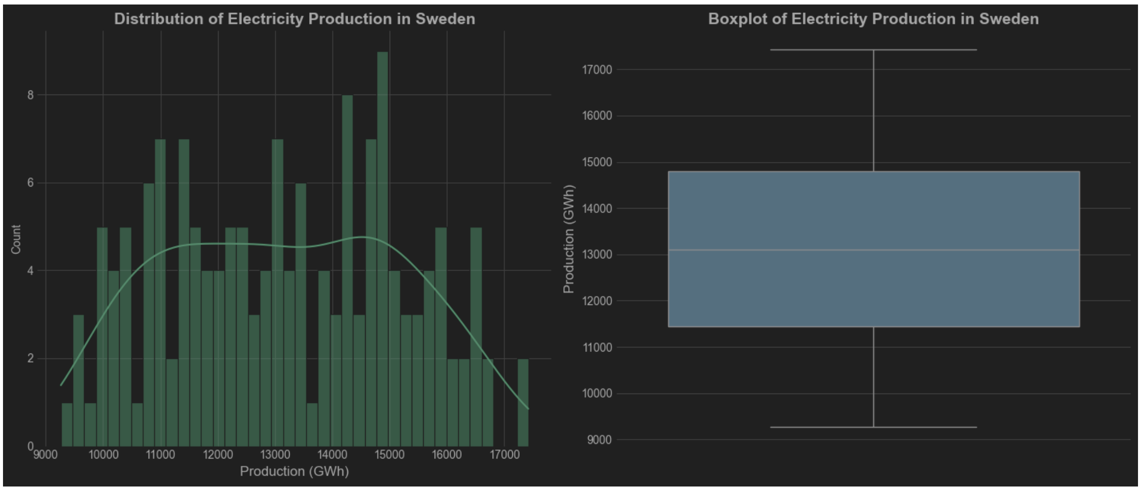
Another variable analyzed was the main sources of energy produced in Sweden. As show in the chart below, the biggest production of energy is hydro generated, at around 39%, the second biggest is nuclear, at around 23%, thirdly is wind, at around 17.5%, followed by combustible fuels at around 10%, then combustible renewables at around 8%. The other power sources are below 1%.



We can conclude that Sweden’s renewable production volumes are significant, reflecting its commitment to clean energy, particularly hydro, nuclear and wind power.

The last variable analyzed was the overall energy production in GW/h in Sweden from 2011 to 2022. The average electricity produced was 13146 GW/h, during the aforementioned time period, with a range between 9264 and 17421 GW/h. The standard deviation of electricity production was 2043 GW/h. As shown in the table and boxplot below.
The distribution of electricity production does not follow a normal distribution since there are two concentrations around the 11000 GW/h and 15000 GW/h, as observable in the histogram below.

STATISTICAL OVERVIEW ELECTRICITY PRODUCTION IN SWEDEN (2011-2022)	(GW/H)
COUNT	156
MEAN	13146
STD	2043
MIN	9264
25%	11442
50%	13106
75%	14799
MAX	17421



Exercise 2.2 Confidence interval for one quantitative variable; Confidence interval for difference between two groups.

The confidence interval for net energy production in Sweden from 2011 to 2022 with 95% confidence level is between 12822.5 and 13468.9 GW/h (rounded to one decimal). This means that we can be 95% confident that the true average net energy production for the entire population lies between the aforementioned results.

The confidence interval between the difference of energy production of renewables and non-renewables falls between 2924.3 and 2921.8 GW/h (rounded to one decimal). Since the interval is far above zero, there is a statistically significant difference in energy production between renewables and non-renewables at the 95% confidence level.

Exercise 2.3 Carry out a T-test to check if the difference in characteristics between two groups is significant, or conduct an ANOVA to see if all groups have the same mean value in some characteristics.

The resulting p-value of the T-test to see if there is any significant difference in the average of renewable energy production in Denmark and Sweden between 2011 and 2022 was 2.36 e-141. Since the p-value for the test is much smaller than 0.05, we reject the null hypothesis that the means of renewable energy production in Denmark and Sweden are equal.

Exercise 2.4 Non-parametric test for same variable as in 3) and even compare the conclusion(s) with ANOVA

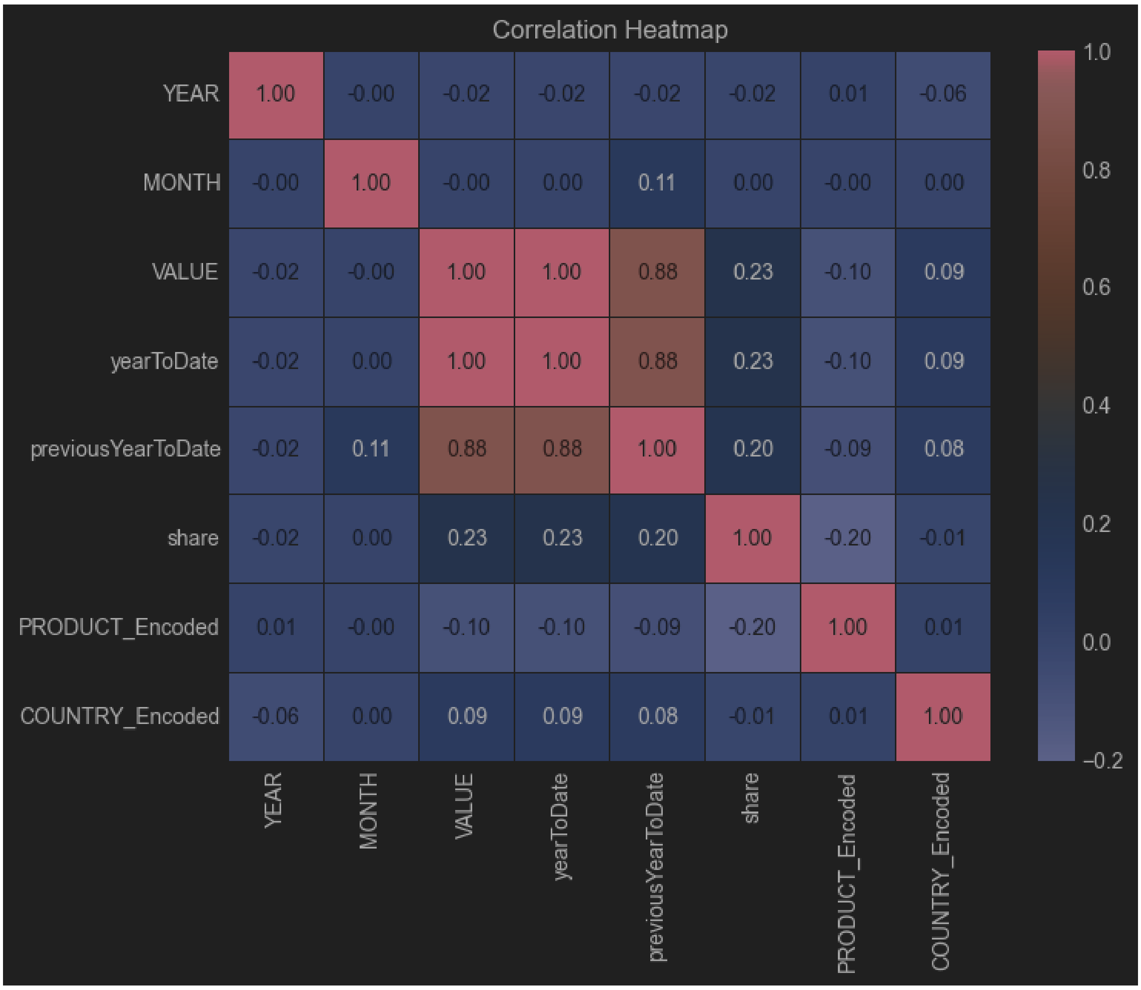
The resulting p-value of the Mann-Whitney-Wilcoxon test to compare if there is any significance difference in the distributions of renewable energy production in Denmark and Sweden between 2011 and 2022 was 1.29 e-50. Since the p-value is much smaller than 0.05, we reject the null hypothesis that the sample distributions are equal.

Both values are much smaller than the 0.05 significance threshold, giving proof that Denmark and Sweden have significantly different renewable energy production patterns between 2011 and 2022.

Exercise 2.5 Correlation analysis, thereafter, identify the strongest correlation and statistically not significant relation(s)

For this study the categorical variables product and country were label encoded to provide a better correlation overview of all the features. The other categorical values were ignored as was the display order since it is not relevant for this study.
As seen in the correlation heatmap bellow, Value and YearToDate show perfect correlation (1.00) with each other, indicating they likely contain the same information or one is derived from the other.

previousYearToDate and YearToDate variables show a strong correlation (0.88) as expected. While share and VALUE/yearToDate show a weak positive correlation (0.23), suggesting some relationship between the share metric and energy production values.



	YEAR	MONTH	VALUE	yearToDate	previousYearToDate	share	PRODUCT_Encoded	COUNTRY_Encoded
YEAR	1.00	-0.00	-0.02	-0.02	-0.02	-0.02	0.01	-0.06
MONTH	-0.00	1.00	-0.00	0.00	0.11	0.00	-0.00	0.00
VALUE	-0.02	-0.00	1.00	1.00	0.88	0.23	-0.10	0.09
yearToDate	-0.02	0.00	1.00	1.00	0.88	0.23	-0.10	0.09
previousYearToDate	-0.02	0.11	0.88	0.88	1.00	0.20	-0.09	0.08
share	-0.02	0.00	0.23	0.23	0.20	1.00	-0.20	-0.01
PRODUCT_Encoded	0.01	-0.00	-0.10	-0.10	-0.09	-0.20	1.00	0.01
COUNTRY_Encoded	-0.06	0.00	0.09	0.09	0.08	-0.01	0.01	1.00

Exercise 2.6 Make a linear multiple regression analysis

For this linear multiple regression analysis, the dependent variable analyzed was value, representing the amount of electricity generated in gigawatt-hours (GWh), with the independent variables being: year, month, previous Year To Date, share, product and country.
Since the Year To Date variable, through the correlation matrix, showed that it likely contains the same information as value, this feature was ignored.

As for the overall model performance there is a strong predictive power ($R^2 = 0.780$). This means that the model explains 78% of the variance in global energy production (value). The significant predictors in the model were: previousYearToDate (coef = 0.1192, t = 612.313), month (coef = -2,524.19, t = -72.514), share (coef = 10,220, t = 35.544), country (coef = 127.89, t = 15.822) and product (coef = -125.66, t = -8.115).