VIS-Assignment 4

Name: Botond Jenő Kovács
Student number: 12323327

1. Data, users, and tasks (25 points)

Out of the first pair, I chose the activists from WWF as one of the user groups. These activists are deeply committed to environmental conservation and are associated with WWF, an international organization focused on biodiversity conservation and addressing global environmental challenges. As they are preparing for a meeting where multiple country leaders will be present, their aim is to show them that climate change has different effects in different countries, and they should probably come up with a project to help the countries most in need. For this, they look at the problem from three different points of view:
Their primary task is to investigate the relative impact of climate change on natural habitats across various regions worldwide. This task involves analyzing climate change indicators such as natural disasters, sea level rise, and changes in forest cover to assess their impact on ecosystems and biodiversity. They also plan to investigate if there are some countries or regions where the mean surface temperature has increased significantly more than the average. The third perspective is to identify the countries or regions where the ratio of renewable energy sources to non-renewable energy sources is the lowest.

Out of the second pair, I have chosen the UN Commission on Climate Change.
This commission consists of relatively young people with a new, fresh, data-driven mindset. Currently, their goal is to prove that the number of natural disasters is increasing over time, and they want to establish a connection between these occurrences and the actions of governments.
Naturally, their primary task is to demonstrate which countries are most affected by climate change to facilitate debate on strategies.
Additionally, they are tasked with exploring connections between the frequency and severity of natural disasters and the amount of expenses allocated to activities that have a negative impact on the environment.
Furthermore, they aim to identify correlations between natural disaster occurrences and expenses allocated to activities that have a positive impact on the environment, such as renewable energy investments or conservation efforts.

Datasets:
To address these tasks, I will utilize the following datasets:
14) Trend in number of climate-related natural disasters from 1980-2022.
This is necessary for all tasks 1.1, 2.1, 2.2 and 2.3.
In this dataset we have data for each country and for each type of disaster, as well as the total number of them, and for each year the exact number of the occurrences.
We have 214 countries in the dataset, which is of course enough.
Data is missing from a lot of fields, but as there are no 0s recorded, it is reasonable to assume that the missing values are all zeros. Also, there are countries, where not all the disaster types are listed, but after a sanity check it can be assumed that it is because such a disaster has never happened in that given country.

An interesting fact to note is that the highest number of natural is disaster is almost in every year in the United States, for example in 2022 there were 25 ones there, while the second highest value was 14 in that year (Colombia).
There are also records of countries, which do not exist anymore: the Soviet Union, the German Democratic Republic, and the German Federal Republic, which will need to be handled manually. Or another solution is to only look at the data starting from 1992, after all these countries were transformed.

7) Environmental Taxes from 1995-2021.
This is necessary for the task 2.2.
In this dataset we have data for each country and for each type of taxes, as well as the total sum of them (Indicator=Environmental Taxes), and for each year the exact number of the occurrences. Also, for most of countries we have these taxes as exact numbers, and as percentages of their GDP.
Unfortunately for many countries the data is only reported since 2015, so we possibly must reduce our timeframe from 1995-2021 to 215-2021 for this task.
For some countries, we do not have all types of taxes, it is reasonable to only consider the sum of the taxes and drop all countries for which it is not available.
We have empty values as well, which we should not consider as being 0.
As an example, in 2021 the highest environmental tax compared to the GDP was in Croatia, 3.79%.

8) Environmental Protection Expenditures from 1995-2022.
This is necessary for the task 2.3.
The structure of the dataset is like the previous one. We have several types of spendings for each country in each year, both in their own currency and as the percentage of their GDP, however we do not have a sum of them.
Unfortunately, we have a lot of missing data, especially before 2014, but also for some countries in total, so we must both shrink the timeframe and the number of countries to investigate.
There are 0 values, which seem to be consistent, hence I accept them as actually being 0, not missing.

10) Renewable Energy from 2000-2022
This is necessary for the task 1.3.
In this dataset we have an instance on a country-energy type-measure type level, where measure type is either the electricity generation or the electricity's installed capacity. We have columns for years starting in 2000, ending in 2022, but in 2022 we do not have data on the electricity generation.
We also have some other missing values, but mostly before 2013. Thus when creating the plots it is considerable to only look at the data starting in 2013.
What is especially useful for us is that we have aggregated data on continent level, which is option to plot instead of plotting every country. Especially that it also includes data in Europe split into South, West, North and East regions.

24) Annual Surface Temperature Change from 1961-2023
This is necessary for the task 1.2.
This might be the simplest and most complete dataset of the ones I have selected.
We have one row for each country, and continents and a worldwide one, all in all 236 rows. Out of these 236 we have 162 rows which do not have any missing values.
It is interesting to note that in the sixties it was common to have negative values, but unfortunately in the latest year (2023) we only have positives, the number of negative values decreases as we move forward in time.

After looking at the selected datasets I am confident that I will be able to assist the UN commission by a dashboard which helps them identify regions more in need than others from all the three perspectives they asked for. The data needs to be manipulated, because we have the years as different columns, it would be useful to merge them into one column, and we also have to work with some missing values, but I am confident because by shrinking the time interval, so we have only an insignificant amount of missing values, we will still have more than enough instances.

By analyzing these datasets, I also aim to provide evidence supporting the WWF activists' claims regarding the increasing frequency of natural disasters and their correlation with environmental policies and expenditures. My data selection is sufficient for this goal, as we have data for all countries on the number of natural disasters, environmental taxes, which account for spendings resulting in negative impact and environmental protection expenditures, which account for spendings resulting in positive impact.

2. Task abstraction (15 points)

Task 1.1 - Investigate the relative impact of climate change on natural habitats across various regions worldwide.
For this task first we have to decide the granulation, do we want to look at each countries, or just regions, continents?
The next step is to plot for example the number of natural disasters in each of these regions.
I would abstract into tasks the following way:
Filter: Find data subset that satisfies the conditions, i.e. has all the relevant data in the given years and also choose the granulation level
Compute Derived Value: Add the number of disasters in the selected countries for each year
Find Extremum: Find some records with extreme values and try to identify its cause

Task 1.2 – Anomaly detection in surface temperature increases
Next, we are tasked to identify countries/continents where the surface temperature increases are the highest. It is a standard maximum values search task.
I would abstract into tasks the following way:
Filter: Find data subset that satisfies the conditions, i.e. has all the relevant data in the given countries/regions and years
Sort: rank the data according to the most recent surface temperature increase
Find Anomalies: fin the highest values

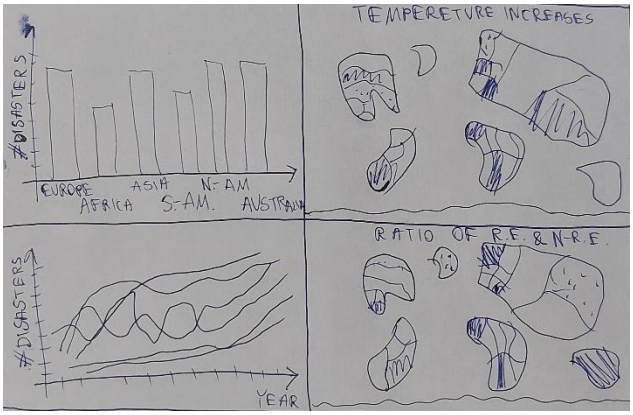
Task 1.3 – Anomaly detection in ratio of renewable energy sources to non-renewable energy
Task 1.3 is really similar to 1.2, but now we are looking for minimums, and we have to calculate our metrics, so it is less straightforward.
Filter: Find data subset that satisfies the conditions, i.e. has all the relevant data in the given countries/regions and years
Compute Derived Value: calculate the most recent ratio of renewable energy sources and non-renewable energy sources
Sort: rank the data according to this ratio
Find Anomalies: fin the lowest values

Task 2.1 – Demonstrate which countries are affected most by climate change to debate strategies.
For this task I have chosen to look at the natural disaster dataset, thus first we must select the countries and years, in which we have appropriate data. Afterwards we need plots which helps us understand how the number of disasters changed over the years in different countries, or regions. We can also use the fact that we have data on different disaster types, not only their total number, thus we can use it as a filter.
I would abstract into tasks the following way:
Filter: Find data subset that satisfies the conditions, i.e. has all the relevant data in the given countries and years
Compute Derived Value: Add the number of disasters in the selected countries for each year
Find Extremum: Find some records with extreme values and try to identify its cause
Determine Range: Try to find a subset of countries, for which we can see a relevant increase

Task 2.2 – Explore connections between the frequency of natural disasters and the amount of negative expenses.
After looking at the data, I am aware that choosing the timeframe and group of countries with sufficient data is challenging for this task. However, afterwards our task is relatively simple, as this connection can be investigated on a scatter plot, each point representing a country, the number of disasters and one type of expenses are the axis, and we can use the year as a filter.
I would abstract into tasks the following way:
Filter: Find data subset that satisfies the conditions, i.e. has all the relevant data in the given countries and years
Compute Derived Value: Identify a spendings, or group of spendings which will be relevant for the plot
Find Extremum: Find some records with extreme values and try to identify its cause
Determine Range: We have to choose a range of years we want to look at
Correlate: Identify correlations between the spendings and number of disasters in the selected time range

Task 2.3 – Explore connections between the frequency of natural disasters and the amount of positive expenses.
This task is absolutely similar to the previous one, the only difference is basically one the axis being the governmental spendings (positive expenses). Another layer of difficulty is deciding if we want to use one specific spendings, or their sum (here we do not have it automatically, but it is straightforward to calculate).
I would abstract into tasks the following way:
Filter: Find data subset that satisfies the conditions, i.e. has all the relevant data in the given countries and years
Compute Derived Value: Identify a spendings, or group of spendings which will be relevant for the plot
Find Extremum: Find some records with extreme values and try to identify its cause
Determine Range: We have to choose a range of years we want to look at
Correlate: Identify correlations between the spendings and number of disasters in the selected time range

3. Functionality and usability of the dashboards (10 points each per dashboard = 40 points)



My first dashboard will consist of 4 views. The views on the left are for task 1.1, the upper right view is for task 1.2 and the bottom right is for task 1.3.
The Upper left view is a bar plot, where each bar represents a continent, and the height of the bar represents the number of natural disasters in the most recent year for the given continent.
This view has an interactive function with the remaining ones. Upon choosing some of these bars, the other views will only show the selected continent(s).
I have chosen a bar plot, because it is one if not the most common way to visualize one measure for different groups, and the bars can have their own colors, which helps to match them with the second plot.
The second plot, the bottom left one is a line plot, which is basically an expansion of the first view.
For each (selected) continent we have a line plot, which shows the number of natural disasters over the years, thus making it easy to identify which continent suffers the most.
On the right we have two similar views, both is a colored world map.
The upper one showing us the rise in temperature in the most recent year, the bottom one showing the ratio of renewable energy and non-renewable energy.
I have chosen a world map for these tasks because I wanted it to be able to show all countries, and having a world map is the most efficient way to do so.

For the second dashboard my first view is a world map, where the color will indicate one measure, such as the number of natural disasters in the most recent year, but its main functionality will be interaction. By choosing on this a map a country or a group of countries, the other plots will change to show only these selected countries.

For task 1.1 I have chosen a barplot, which shows the number of natural disasters in each year. If multiple countries are selected, these numbers are summed up, so without selecting anything, it shows the number of natural disasters over the years worldwide. I found this plot to be the most efficient to show the change of the number of disasters over the years.

For tasks 1.2 and 1.3 I plan to create scatter plots, the y axis being the number of disasters, the x axis being the respective spendings. By filtering for a group of countries, the plot would show only the points of those countries. I have chosen scatterplots, because they are a good option to show correlation, either negative or positive, between two metrics.

I also plan to add additional features on the type of spendings, so the users can check multiple things. It is possible that there is a correlation, but only with one type of the spendings, which could not be identified if we can only look at the sum of spendings.

4. Reflection (15 points)

Dashboard 1

Also, as mentioned above I try to include a link between the views based on the regions selected on the first view, I am confident that I will be able to have the data on continent level in every dataset. This first view is an overview view, where one can compare the continents in the most recent year in an instant, while the following ones are detail views.

The line plot is expanding the first view, as now we can look at the different years at the same time, not just the most recent one.

The plots for tasks 1.2 and 1.3 will provide an easy way for the user to find the extreme values in both cases.

As I have selected datasets with only a few features each, I had no issues with the number of attributes.

One significant tradeoff is as I did not want to have a dropdown menu for choosing a year, everywhere only the most recent year's data is shown, expect for the line plot.

Dashboard 2
Also, as mentioned above the link between the views is based on the countries selected on the map view. The other views will automatically show only these selected countries. Additionally, its color will represent the number of natural disasters in the most recent year, hence this is an overview view, while the following ones are detail views.
With my second plot the users will be easily identify how the number of natural disasters in a certain group of countries changed over the years, however a limitation is that they cannot have a look at multiple countries' separately at the same time. On the other views they will be able to identify the correlations, but if they want to look at many countries, the view will be overwhelming, hard to distinguish between the different countries.
I decided to include only a portion of the attributes in the data, this is how I tried to find the most effective visual encodings. This way I will have a low data-density, but my views will be still relevant.

5. Conclusion (5 points)

As of visual encodings I can confidently state that with these dashboards I will be able to avoid having lie factors at all. The axes are aligned correctly, and labels will be shown everywhere. This way I will also omit any kind of distortions.
The charts also will not show any unnecessary grids, hence my chartjunk is minimal. These plots also do not require plenty of colors, I will work only with one color's shades, so data-ink ratio will be high, but vividness is present via this color, probably blue.
Data density can be high on my second last plots of my second dashboard, but after filtering for a portion of the countries, this issue is also resolved.
On dashboard 1 the data-ink ratio will be lower, as we will use less colors than for dashboard 2. However there is some repetition, the upper left view can seem redundant, as all the information there is incorporated in the bottom left plot, but its main functionality lies in that it is easier to filter for some continents this way.
A minimal repetition might be present on dashboard 2, as the map will show the number of natural disasters, and the second view is also about it, but I would say that this amount of repetition is still acceptable, it is a tradeoff between having vividness, as I would prefer the map not to be black and white.
I cannot state anything about proximity before starting the actual visualization, but I think I will not have any issues with it.