Faculdade de Ciências da Universidade de Lisboa

Master's Degree in Data Science

Data Visualization (2022/2023)

2nd Project - Data Visualization using Power BI Desktop

Group VD 02

André Dias 59452, Cláudia Afonso 36273, Nihan Ahat 61010, Tiago Rodrigues 49593

4th of January 2023

Table of Contents

I. Introduction

The goal of this project is to use the PowerBI software to assemble interactive visualisation dashboards for finding relations between different variables of a given dataset.

To fulfil this goal, the work was divided into the following tasks: (1) data acquisition by searching for compelling data in the PORDATA database (https://www.pordata.pt), (2) data pre-processing, be it by only selecting a portion of the data or by making calculations on the data to enable further exploration, (3) evaluation of the best types of graphical representations and assembly of dashboards, (4) assessment of relationships between data derived from the produced dashboards, and (5) identification of limitations of the PowerBI software felt throughout the development of the project.

The default data for this project is related to the Public Education System and was supplied by Direção-Geral de Estatísticas da Educação e Ciência (DGEEC). From the provided dataset, which came in an excel (.xlsx) format, the variables “Urban Area Typology” (TIPAU), “NUTS II” and “NUTS III” were used. A greater focus was given to other variables that were not in the provided dataset, which were acquired through PORDATA. The first three extra variables were mandatory and another three were arbitrarily chosen. The extra variables were: (1) Total resident population (according to 2021 Census) [1]; (2) Population density (according to the 2021 Census) [2]; (3) Unemployment registered at the public employment office (2021) [3]; (4) Total crimes registered by the police (2020) [4]; (5) Unemployed population by education level (2021) [5]; (6) Average monthly income by education level (2019) [6].

II. Description of data visualization techniques in PowerBI

Power BI offers a wide range of data visualisation techniques that allow the representation of data in a clear and meaningful way. Data visualisation tools make it easy to see and comprehend data trends, outliers, and patterns by employing visual elements like maps, charts, and graphs. Some of the techniques that Power BI offers include cards, ArcGIS, stacked bar chart, radar chart, tornado chart, scatter chart, donut chart, line and stacked column chart, slicers and tables.

Cards are a simple visualisation that displays a single value or metric. These are useful for showing a data summary or highlighting a critical value [7].

ArcGIS Map is a geographical information system platform that allows the creation and analysis of spatial data. It is used in a variety of settings, including government, natural resources, and defence, to create maps, analyse data, and make informed decisions [8].

Stacked bar chart is a graph that shows the relationship between two or more variables, with rectangular bars stacked on top of each other. Each bar represents a different category, and the height of the bar reflects the size of the value within that category [9].

Radar chart is a representation that shows relationships between multivariate data on a two-dimensional chart. They are useful for seeing which variables score high or low within a dataset, making them ideal for displaying performance [10].

Tornado chart is a graphical representation of sensitivity analysis that is used to compare the relative importance of different variables. It is a useful tool for identifying the factors that have the greatest impact on a particular outcome [11].

Scatter chart, also known as a scatter plot, is a graphical representation of two or more sets of data points that shows the relationship between variables. It is used to plot data points on a horizontal and vertical axis to show the relationship between variables [12].

Donut chart is a graphical representation of data that uses a circular area to show the proportions of different data points relative to the whole. It is similar to a pie chart, but it has a hole in the centre, making it easier to compare the sizes of different data points [13].

Line and stacked column chart is a graph that combines a line chart with a stacked column chart. It is used to show how two or more variables change over a period of time, with the stacked column chart showing the contribution of different components to the total amount. The line can be used to represent an extra variable [14].

Slicers are interactive filters that allow for a quick and easy way to filter data in the dashboard. They are useful for narrowing down the data displayed in the dashboard and for exploring different subsets of data [15].

Tables are a type of visualisation that displays data in rows and columns. They are useful for showing raw data and for creating a basic overview of the data set [16].

An example of each of the mentioned representations can be seen in Figure 1.

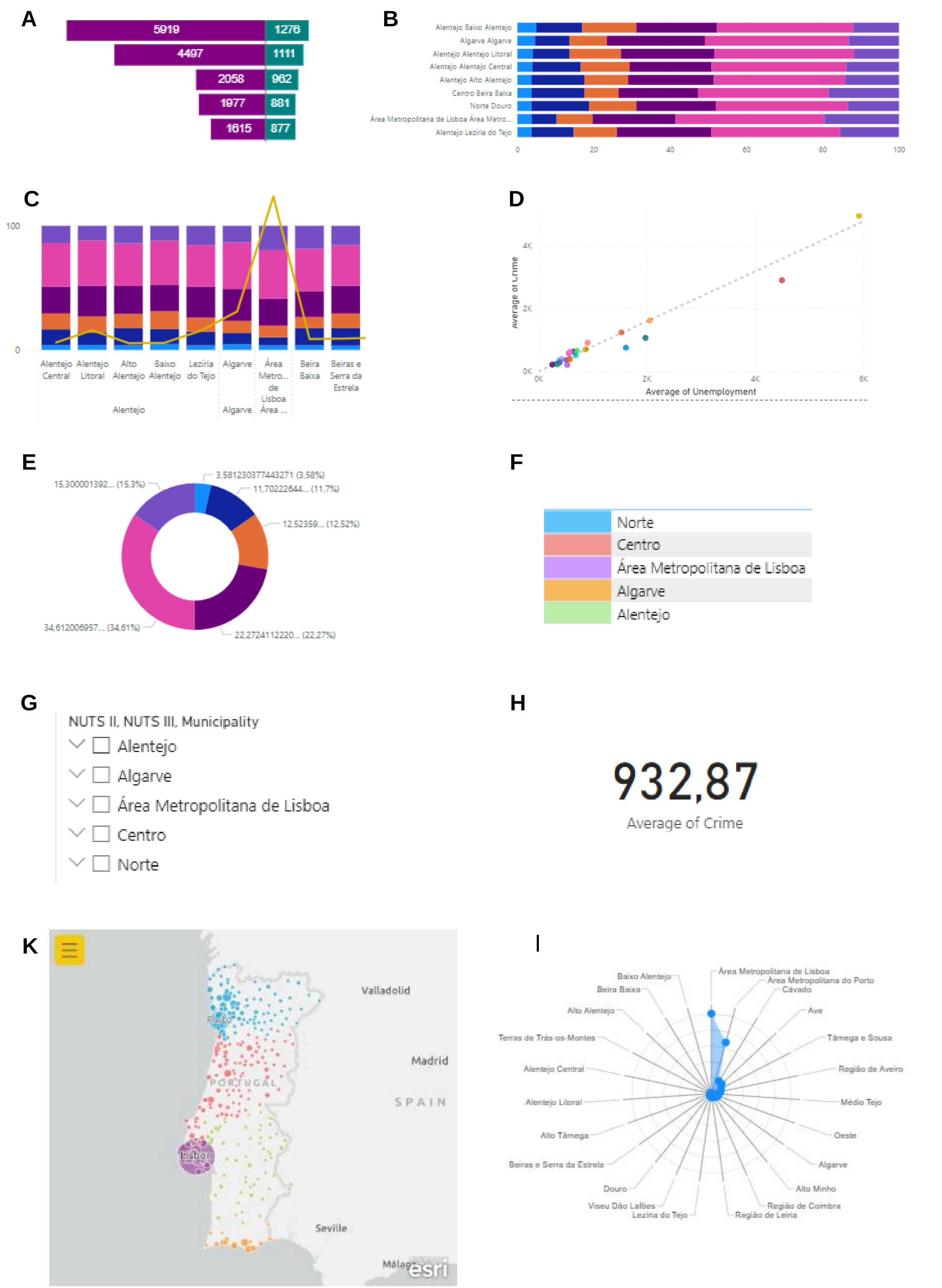

Figure 1 – Examples of the myriad of representations that PowerBI offers. (A) – Tornado chart; (B) – Stacked bar chart; (C) – Line and stacked column chart; (D) – Scatter chart; (E) – Donut chart; (F) – Table; (G) – Slicer; (H) – Card; (K) – ArcGIS map; (L) – Radar chart.

III. Data acquisition and pre-processing

As previously mentioned, two types of datasets were used: one which was provided and the other acquired through the online PORDATA database.

To obtain data from this platform, relevant information can be sought with the search bar of its website using specific keywords. Since municipality data was desired, the “Municípios” tab was chosen. In this tab, all information regarding the keyword that existed on a municipality basis can be explored. After selecting the intended data, the latter was downloaded by exporting it to an excel file (.xlsx) through the “Exportar para Excel” option.

After data acquisition of the beforementioned extra variables, all of them underwent some type of data pre-processing.

At this stage, the first step consisted in the removal of all the unnecessary data in each excel file. Despite the ability to select the variable data for a chosen year, the PORDATA database would always aggregate it with the information pertaining to the first recorded year as well. Thus, for variables “total resident population” and “population density”, information concerning the year 1960 was removed. For variables “unemployment”, “crime”, “unemployed by education level” and “income”, information regarding the years 1997, 1993, 1981 and 1985 were removed, respectively.

Furthermore, each of these excel files contained information regarding not only the municipalities, but also NUTS III, NUTS II and NUTS I. As such, this information was removed from each excel file, to avoid conflicts downstream.

To facilitate the importation of the data to PowerBI, a table was created for each dataset, through the “Insert Table” tool in the “Insert” tab of Excel.

For the creation of ArcGIS graphical representations, a new column termed “Location” was created for each excel file to assist georeferencing. This column consisted of the “Regions” field concatenated with the string “, Portugal”. To achieve this, the “CONCAT” function in excel was used, uniting each municipality cell with a cell that contained the beforementioned string. This step was also performed for the provided excel file of the Public Education System.

Following this, all the data was imported into PowerBI and each table connected with the Public Education System through the “Location” column. This was achieved by traversing to the “Model” tab in PowerBI, interacting with the “Manage Relationships” button, creating a new relationship, and selecting two “Location” columns from different tables. A snippet of the relationship creation window can be seen in Figure 2.

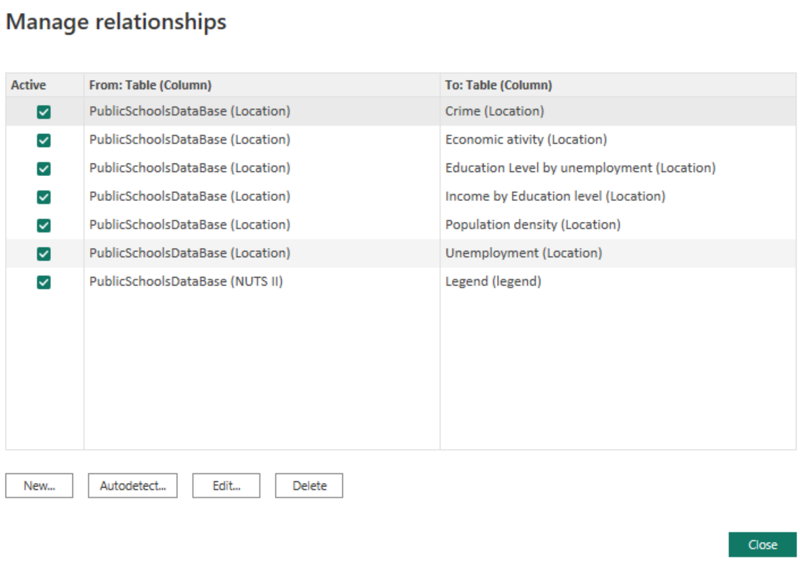

Figure 2 – Snippet of the “Manage relationships” window in PowerBI.

For the variable “unemployed by education level”, the percentage of unemployed individuals by education level for each municipality was calculated within PowerBI. To accomplish this, the “Transform Data” button was selected in the “Home” tab. Then, in the “Add Column” tab, the “Standard” button was pressed after choosing one of the columns to perform operations on. Firstly, the “Divide” option was chosen, and the “Enter a value” option was changed to “Use values in a column”, on which the “Total” column was used. To obtain a percentage, the newly created column was then multiplied by 100, using the “Multiply” option, instead of “Divide”. This process was performed for every education category in the mentioned variable. In the end, the percentage of unemployed individuals by education level was obtained for every municipality.

For the ArcGIS map representations, a custom legend was created due to limitations of its automatically generated legend. To do so, a new excel file was created containing the NUTS II field and a given colour value, similar to the colouring that would be applied by the ArcGIS representation. This table was then imported into PowerBI and, in each ArcGIS map, a table representation was created with its data. In the “Format” tab of the representation, the background colour was changed according to its “colour value” field, by selecting the field value in the format style. The font colour was changed through the same process to hide the text in the background colour. In the end, the font colour of the headers was also changed, but this time to plain white, to conceal it on the background of their respective cells. A relationship between the finished table and the Public Education System dataset was created through the “NUTS II” column.

IV. Interpretation of the visualizations created

An important part of visualisation is the creation of graphical representations to help summarise the data in an informative and simpler way. As such, interpreting these representations is extremely important to understand trends, outliers, and relationships between variables. Initially, graphical representations for the beforementioned variables were created individually for each, to understand their behaviour. Following individual assessment of all the variables, relationships between pairs of variables were subsequently investigated.

In the forthcoming topics, both the individual and pairwise dashboards will be presented and discussed.

IV.1. Descriptive analysis of each individual variable

<u>IV.1.1. Total resident population</u>

The first dashboard was designed to analyse the variable "total resident population" using the following visualization techniques:

- A card that indicates the total number of residents in continental Portugal. Here, it is possible to see that approximately 10 million people were residing in the country in the year 2021.
- An ArcGIS Map that shows the resident population in Portugal. The bubble size depicts the total residents in each municipality, while their colouring represents the respective NUTS II region. Here, it is possible to see that most individuals reside in the North and in the Lisbon Metropolitan Area (AML).
- A stacked bar chart that shows the values of the resident population for NUTS II, NUTS III and municipality, coloured according to NUTS III (the one being visualized in the dashboard is the municipality). Analysing this graph, it is possible to see that the municipality with the highest resident population is Lisbon followed by Sintra, while the municipalities with the lowest values are Barrancos and Alvito.
- Another stacked bar chart which shows the resident population according to the urban area typology (TIPAU). Here, it is possible to visualize that most of the population resides in predominantly urban areas (APU, "Áreas Predominantemente Urbanas").
- A radar chart that represents the number of residents in Portugal according to the NUTS III region. Here, it is possible to observe that the AML has the highest number of residents, with the Porto Metropolitan Area (AMP) following closely behind. In contrast, Beira Baixa and Alto Tâmega display the lowest number of residents in this country.
- A slicer that allows the information concerning the number of residents in Portugal to be interactively filtered according to the NUTS II, NUTS III regions and municipalities of this country.

The information above was introduced in the first dashboard and the end result is shown in Figure 3.

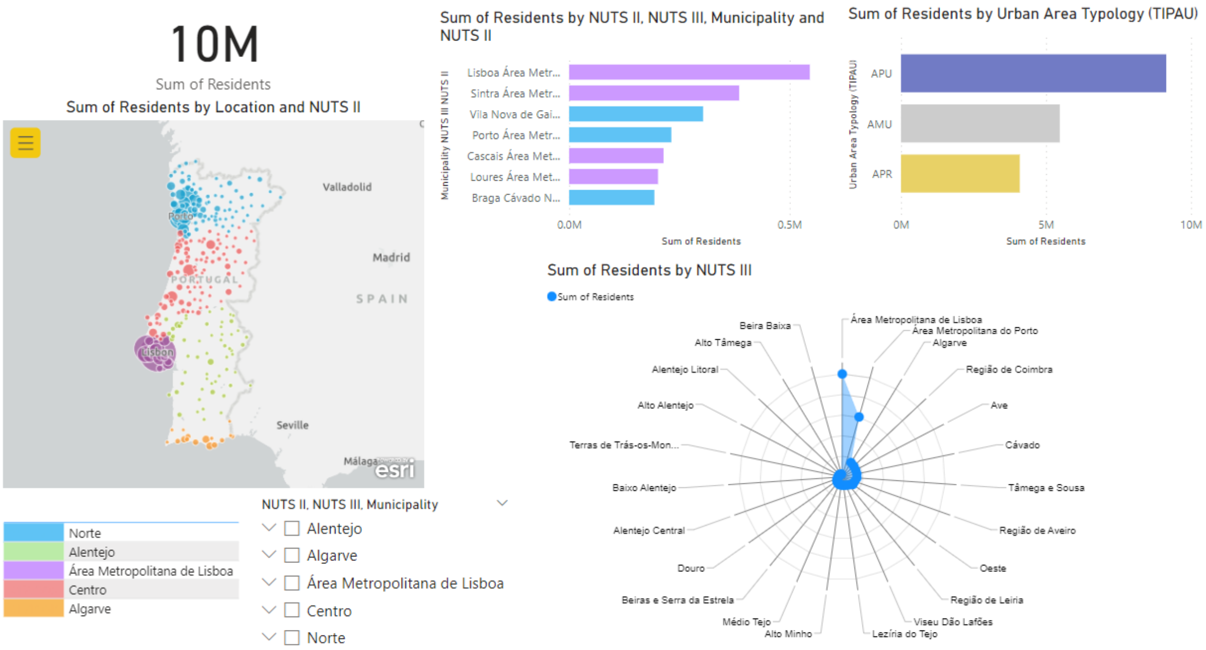

Figure 3 – Assembled dashboard regarding the variable “total resident population”.

IV.1.2. Resident population density

Similarly to the dashboard for “total resident population”, the one for “population density” contains the same data visualizations techniques, changing only the variable in question:

- A card that indicates the average population density per km^2 in continental Portugal. Here, it is possible to see that approximately 292 inhabitants per km^2 were residing in the country in the year 2021.
- An ArcGIS Map that shows the average population density per km^2 in Portugal. The bubble size depicts the average population density in each municipality, while their colouring represents the respective NUTS II region. Here, it is possible to see that most individuals reside in the North and in the AML.
- A stacked bar chart that shows the values of the average population density for NUTS II, NUTS III, and municipality coloured according to NUTS III (the one being visualized in the dashboard is the municipalities). Analysing this graph, it is possible to see that the municipality with the highest average population density is Amadora, followed by Porto. The municipalities with the lowest average population density are Alcoutim and Mértola.
- Another stacked bar chart which shows the average population density according to TIPAU. Here, it is possible to visualize that most of the population resides in APU.
- A radar chart that represents the average population density per km^2 in Portugal according to the NUTS III region. Here, it is possible to observe that AML has the highest average population density, with AMP following closely behind. In contrast, Beira Baixa and Alto Tâmega display the lowest average population densities in Portugal.

- A slicer that allows the information concerning the average population density per km² in Portugal to be interactively filtered according to the NUTS II, NUTS III regions and municipalities of this country.

The information above was introduced in the second dashboard and the end result is shown in Figure 4.

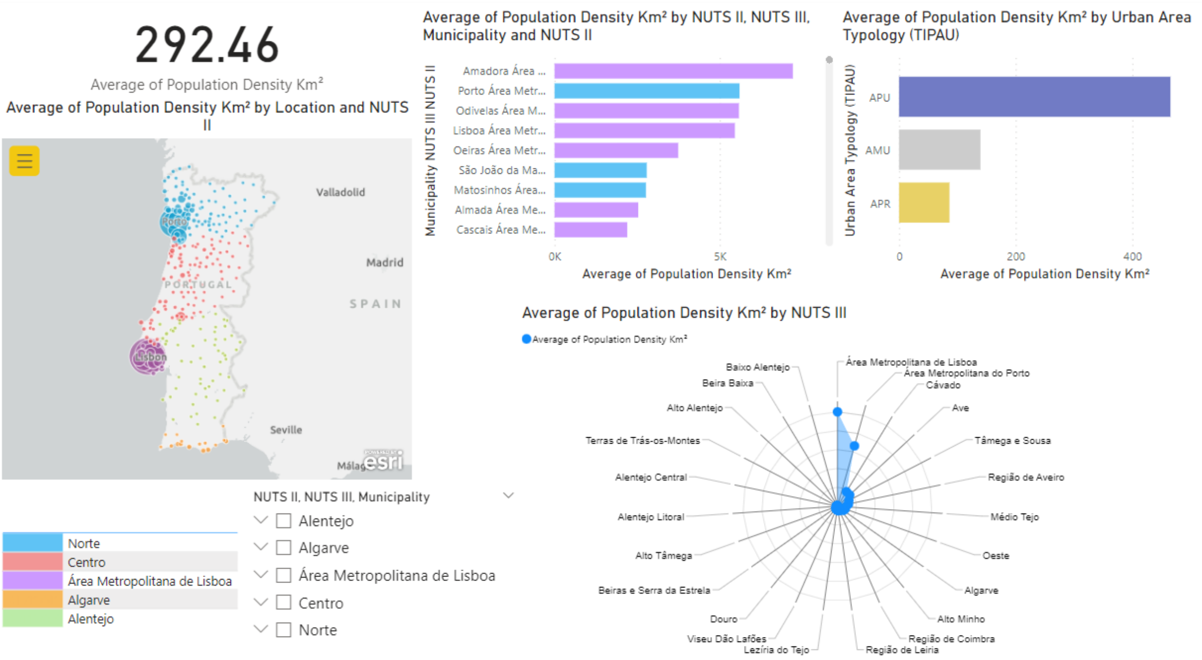

Figure 4 – Assembled dashboard regarding the variable “population density”.

IV.1.3. Unemployment registered at the public employment office

The third dashboard of this project consisted in a set of graphical representations to visualise the variable termed “unemployment”, which is related to the number of unemployed individuals registered at the public employment office in the year 2021. To this end, the following visualisation techniques were created in PowerBI:

- A card that indicates the total unemployment in continental Portugal. Here, it is possible to observe that approximately 361420 unemployed individuals were registered in this country in the year 2021.
- An ArcGIS Map that shows the total unemployment in Portugal. The bubble size depicts the total unemployment in each municipality, while their colouring represents the respective NUTS II region. Here, it is possible to see that most unemployed individuals reside in the North and in the AML.

- A stacked bar chart that shows the unemployment values for NUTS II, NUTS III and municipality, coloured according to NUTS III (the one being visualized in the dashboard is the municipalities). Analysing this graph, it is possible to observe that the municipality with the highest unemployment is AML followed by Vila Nova de Gaia, while the municipalities with the lowest values are Arronches and Alcoutim.
- Another stacked bar chart which shows the total unemployment according to TIPAU. Here. it is possible to visualize that most unemployed individuals reside in APU.
- A radar chart that represents the total unemployment in Portugal according to the NUTS III region. Here, it is possible to observe that AML has the highest unemployment, with AMP following closely behind. In contrast, Beira Baixa and Alto Tâmega display the lowest unemployment values in this country.
- A slicer that allows the information concerning the unemployment in Portugal to be interactively filtered according to the NUTS II, NUTS III regions and municipalities of this country.

The information above was introduced in the third dashboard and the end result is shown in Figure 5.

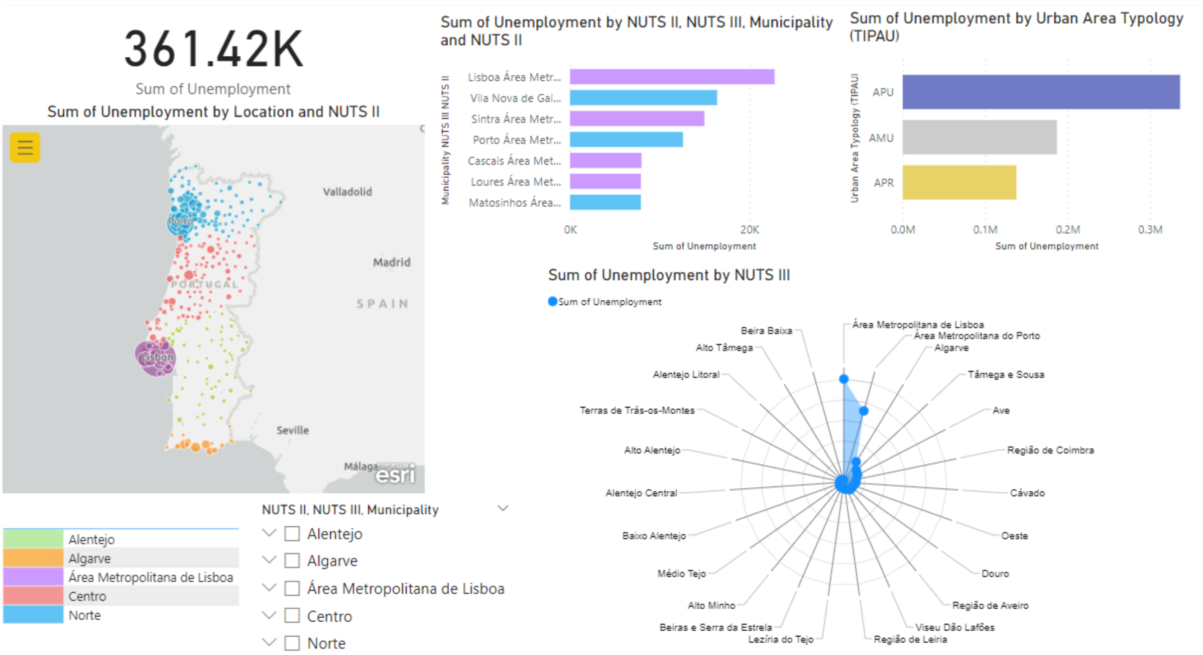

Figure 5 – Assembled dashboard regarding the variable "unemployment".

IV.1.4. Total crimes registered by the police

The fourth dashboard of this project consisted in a set of graphical representations to visualise the variable termed "Crime", which is related to the number of crimes registered by the police in the year 2020. To this end, the following visualisation techniques were created in PowerBI:

- A card that indicates the total number of crimes registered in Portugal. Here, it is possible to see that approximately 287 thousand crimes were registered by the police in the year 2020.
- An ArcGIS Map that shows the geographical distribution of crimes registered in Portugal. The bubble size depicts the total number of crimes registered for a particular municipality, while their colouring represents the respective NUTS II region. Here, it is possible to see that most crimes are registered in the North and AML.
- A stacked bar chart that illustrates the total number of crimes registered in Portugal according to their municipality. Here, it is possible to see that from the eight municipalities with the highest values of registered crime, six of these belong to AML while the other two belong to AMP. The municipalities that display the highest and the lowest crime registered are Lisboa and Arronches, respectively.
- Another stacked bar chart which portrays the total number of crimes according to TIPAU. Here, it is possible to observe that approximately half of the crimes in Portugal are registered in APU, with the remaining half being registered in both mediumly urban (AMU, “Áreas Mediamente Urbanas”) and predominantly rural (APR, “Áreas Predominantemente Rurais”) ones.
- A radar chart that also represents the number of crimes but according to the NUTS III region of where these were registered in Portugal. Here, it is possible to observe that AML has the highest number of registered crimes, with AMP following closely behind. In contrast, Beira Baixa and Alto Tâmega display the lowest number of registered crimes in this country.
- A slicer that allows the information concerning the number of crimes in Portugal to be interactively filtered according to the NUTS II, NUTS III regions and municipality of this country.

The information above was introduced in the fourth dashboard and the final result can be seen in Figure 6.

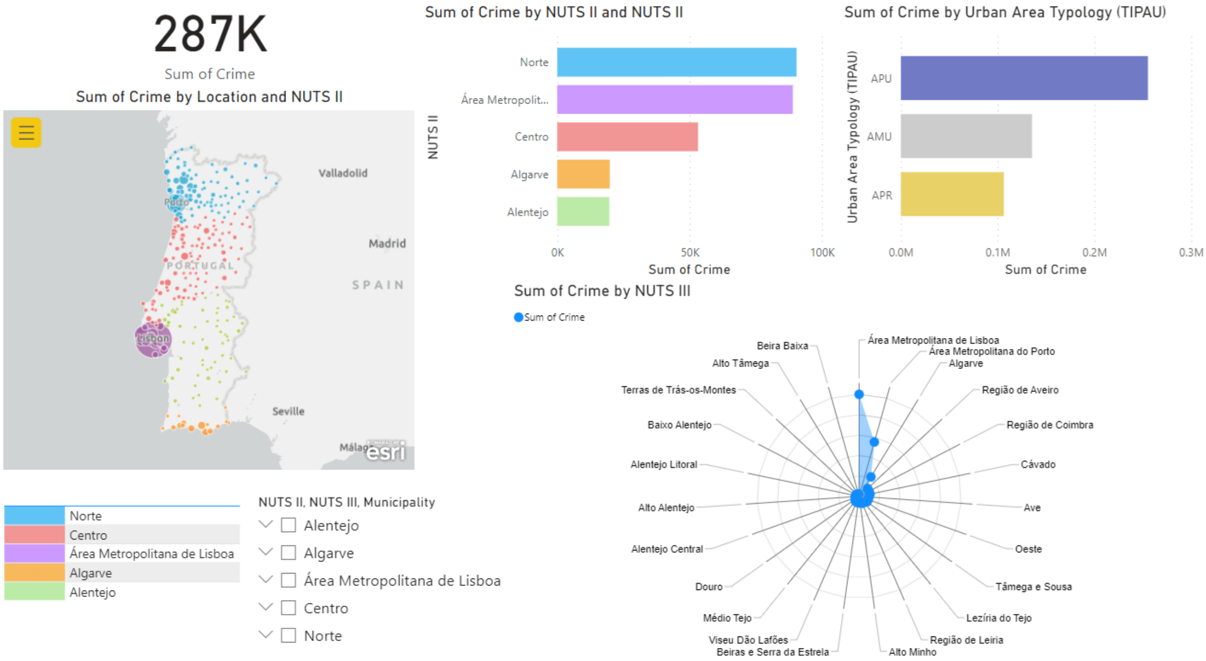

Figure 6 - Assembled dashboard regarding the variable “crime”.

IV.1.5. Unemployed population by education level

The fifth dashboard of this project consisted in a set of graphical representations to visualise the variable termed “unemployed by education level”, which is related to the number of unemployed individuals by their education level in the year 2021. To this end, the following visualisation techniques were created in PowerBI:

- Six cards which consisted of the average percentage of unemployed individuals for each educational level (“Less 1st cycle”, “1st cycle”, “2nd cycle”, “3rd cycle”, “Secondary”, and “Higher”). Here, it is possible to see that for each educational level, the percentage of unemployed individuals was 3.58% for “Less 1st cycle”; 11.70% for “1st cycle”; 12.52% for “2nd cycle”; 22.27% for “3rd cycle”; 34.61% for “Secondary”; and 15.30% for “Higher”.
- A stacked bar chart that shows the values of the average percentage of unemployed individuals for each educational level for NUTS II, NUTS III and municipality, coloured according to NUTS III (the one being visualized in the dashboard is the NUTS II). From this graph, it was possible to visualize that Baixo Alentejo, Algarve and Alentejo Litoral are the regions that show the highest percentage of unemployed individuals with less than the 1st cycle. Meanwhile, the regions of Aveiro, Coimbra and AML show the highest percentage of unemployed individuals with higher education.
- Another stacked bar chart which displays the average percentage of unemployed individuals for each educational level according to TIPAU. From this analysis, it is possible to see that most of the educational levels of unemployed individuals are overall equally spread out across every urban area typology.

- The donut chart shows the average percentage of unemployed individuals for each educational level. From the donut chart is possible to conclude that 35.65% of the unemployed people have graduated from high school. Almost half of the unemployed population has an educational level that is below high school, and only 15.22% have completed higher education.
- A slicer that allows the information concerning the average percentage of education level in Portugal to be interactively filtered according to the NUTS II, NUTS III regions and municipality of this country.

The information above was introduced in the fifth dashboard and the end result is shown in Figure 7.

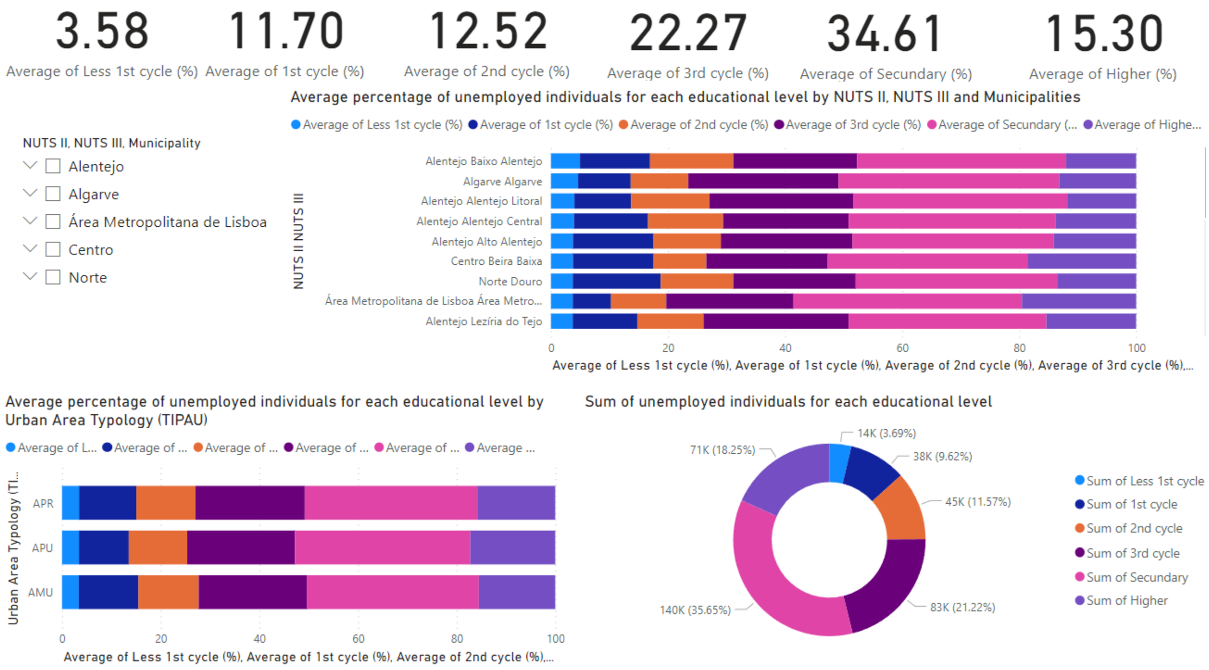

Figure 7 – Assembled dashboard regarding the variable “unemployed by education level”.

IV.1.6. Average monthly income by education level

The sixth dashboard of this project consisted in a set of graphical representations to visualise the variable termed “income”, which indicated the average monthly earnings of employees according to their education levels. In this dashboard, the following visualisation techniques were used:

- A card that indicates the average monthly income in Portugal. Here, it is possible to see that an employee in Portugal earns on average 999.23 euros per month.

- An ArcGIS Map that shows the geographical distribution of the average monthly income in Portugal. The bubble size depicts the average monthly income of employees for a particular municipality, while the colouring represents their respective NUTS II region. Here, it is possible to see that the average monthly income is higher in certain cities of the AML (such as Lisbon, Oeiras and Alcochete) and Alentejo (such as Sines) than in the rest of the country.
- A stacked bar chart that illustrates the average monthly income in Portugal according to their NUTS II, NUTS III regions and municipality (the one being visualized in the dashboard is the municipalities). Here, it is possible to conclude that from the 8 municipalities that have the highest monthly income five of them are located in AML. The highest earning municipality is Alcochete (2011.5 euros) from AML while the lowest earning is Celorico de Basto (794.4 euros) from the north of this country.
- Another stacked bar chart that portrays the average monthly income in Portugal according to TIPAU. Here, it was possible to observe that the average monthly income is higher in APU in comparison with AMU and APR. However, this difference does not appear to be very significant. Furthermore, it is also possible to conclude that the average monthly income in APU is higher than the national average of 999.23 euros while in AMU and APR its lower.
- A radar chart that also represents the monthly average income but according to the NUTS III region. Here, it is possible to observe that AML shows the highest value for this variable, with Alentejo Litoral following closely behind. Thus, an employee earns, on average, a higher salary working in AML than anywhere else in the country. In contrast, Tâmega e Sousa and Ave display the lowest values for average monthly income. The difference in monthly income between the highest and lowest earning regions is approximately 400 euros.
- A donut chart that displays the proportion of employees and the average monthly income according to education level. Here, it is possible to see that individuals who completed higher education represent 25.89% of total employees and earn on average the highest salary at 1475.22 euros. In contrast, individuals who did not complete the 1st cycle represent 11.09% of total employees and earn on average the lowest salary at 632.02 euros.
- A slicer that allows the information concerning the average monthly income of employees in Portugal to be interactively filtered according to the NUTS II, NUTS III regions and municipality of this country.

The information above was introduced in the sixth dashboard and the end result is shown in Figure 8.

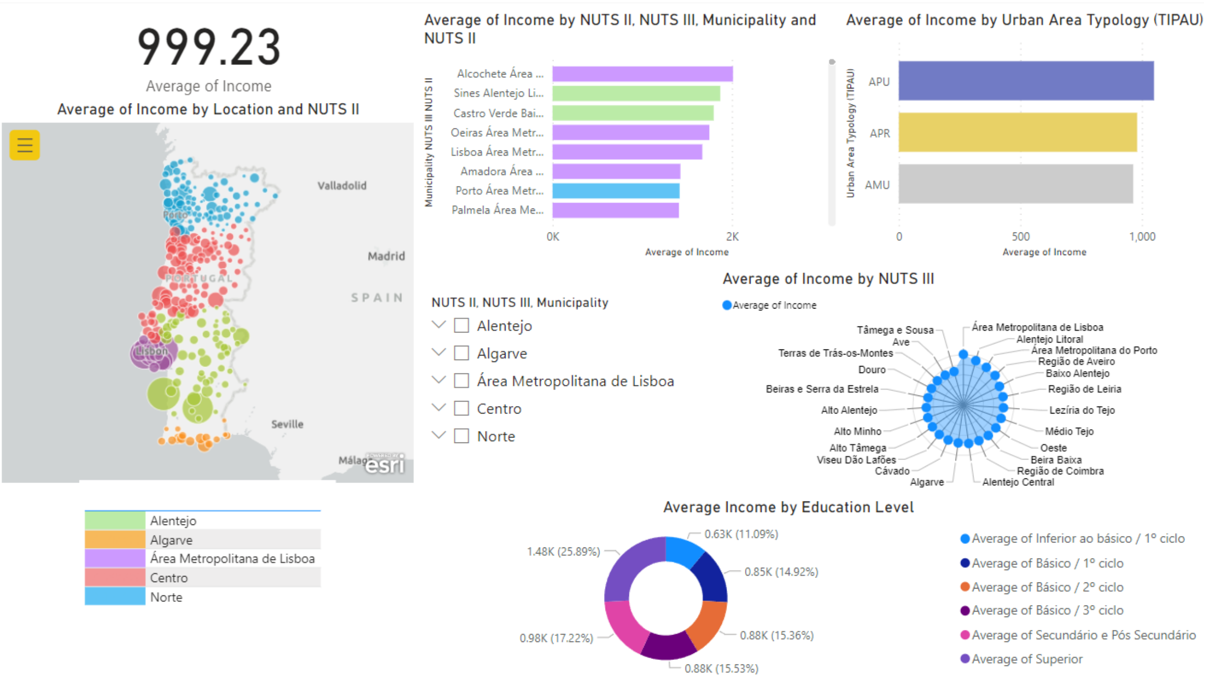

Figure 8 – Assembled dashboard regarding the variable “income”.

IV.2. Descriptive analysis between pairs of variables

<u>IV.2.1. Unemployment registered at the public employment office and average monthly income by education level</u>

The seventh dashboard of this project consisted in a set of graphical representations to visualise the variables termed “unemployment” and “income”. To this end, the following visualisation elements were created in PowerBI:

- Two cards, the one on the left indicating the total number of unemployed people registered at the public employment office and the one on the right showing the average monthly income in Portugal.
- An ArcGIS Map that shows the geographical distribution of average monthly income in Portugal. The bubble size depicts the average monthly income for a particular location, while their colouring represents the respective number of unemployed people (with brown indicating lower values and yellow higher values for this variable). Here, it is possible to observe that a positive correlation exists between higher monthly income values and greater unemployment levels in the AML and AMP. However, in other areas of the country this positive correlation between the variables does not exist. For instance, in Sines the monthly income is quite high, but its unemployment levels are low.
- A stacked bar chart that portrays the average unemployment and average monthly income in Portugal according to the TIPAU. Here, it is possible to see that while a significant difference in monthly income does not exist between the three typologies, a greater level of unemployment is registered in the APU in comparison with AMU and APR.

- A tornado chart that represents total unemployment levels and the average monthly income in Portugal according to the various NUTS III regions of this country. Here, it is again possible to observe that in the metropolitan areas of Lisbon and Porto, higher unemployment levels are associated with slightly higher monthly incomes. However, this positive correlation does not extend to other areas of the country. For instance, Aveiro displays the third highest monthly income (1091 euros) of Portugal, but its number of unemployed people is approximately 10 times lower than in Lisbon (where the monthly salary is on average 1276 euros).
- A slicer that allows the information concerning the average monthly salary and the total number of unemployed people in Portugal to be interactively filtered according to the NUTS II, NUTS III regions and municipality of this country.

The information above was introduced in the seventh dashboard and the end result is shown in Figure 9.

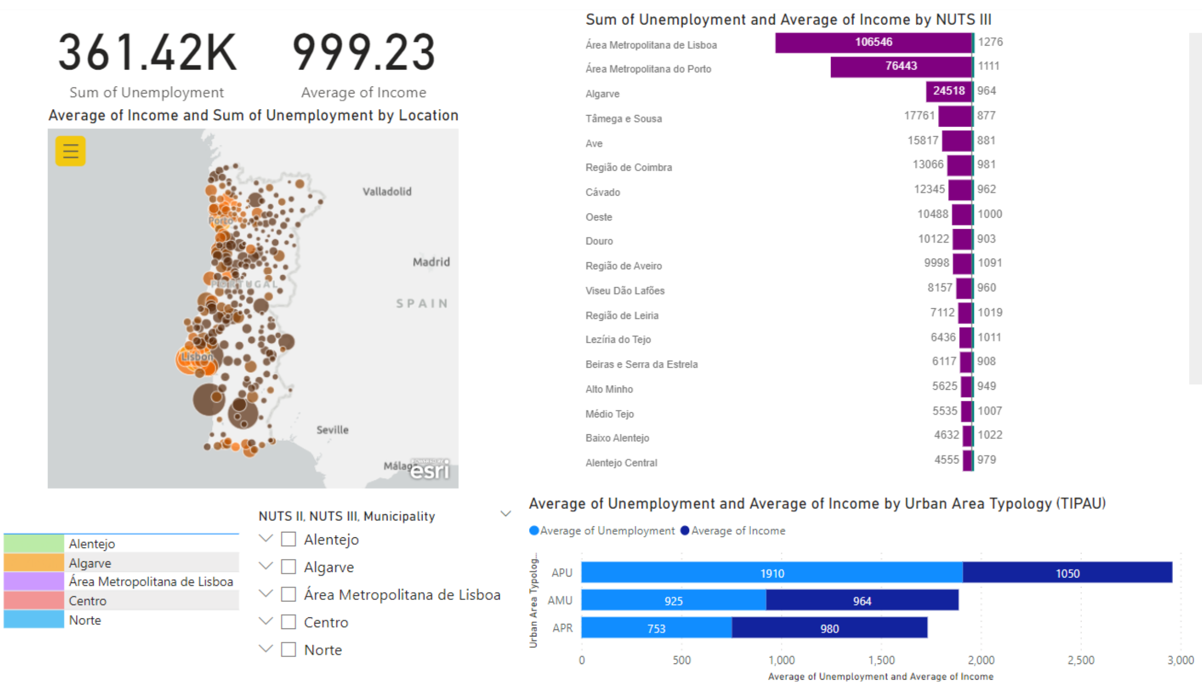

Figure 9 - Assembled dashboard regarding the comparison between the variables “unemployment” and “income”.

<u>IV.2.2. Unemployed registered at the public employment office and total crimes registered by the police</u>

The eighth dashboard of this project consisted in a set of graphical representations to visualise the variables termed “unemployment” and “crime”. To this end, the following visualisation elements were created in PowerBI:

- Two cards, the one the left with information concerning the total number of unemployed people registered at the public employment office and one on the right with information regarding the total number of crimes registered by the police in Portugal.
- An ArcGIS Map that shows the geographical distribution of unemployment compared with crime in Portugal. The bubble size depicts the total crime committed for a particular location, while their colouring represents the respective number of unemployed people (with brown indicating lower values and yellow higher values for this variable). It is possible to observe that a positive correlation exists between higher unemployment values and higher crime rates. This is particularly visible in the areas of AML, AMP and Faro. For the rest of the country, typically areas that show a lower number of unemployed individuals also have fewer registered crimes.
- A stacked bar chart that portrays the total unemployment and total crimes registered in Portugal according to the TIPAU. Here, it is possible to see that both variables display significantly higher values in APU when compared with AMU and APR.
- A tornado chart that represents the total number of unemployed individuals on the left and the total crimes registered on the right according to the various NUTS III regions in Portugal. Here, it is possible to observe that higher unemployment values are associated with greater number of registered crimes. In addition, decreased unemployment values are also associated with lower numbers of registered crimes. This is the first indication that both “unemployment” and “income” are positively correlated.
- A scatter chart that represents the total number of unemployed individuals on the x-axis and the total number of registered crimes on the y-axis. Here, it is possible to see that there is a clear linear trend (almost performing a 45º degree) between both variables. This corroborates the previous observation that both variables display a positive correlation.
- A slicer that allows the information concerning the total number of unemployed individuals and the total of crimes registered in Portugal to be interactively filtered according to the NUTS II, NUTS III regions and municipalities of this country.

The information above was introduced in the eighth dashboard and the end result is shown in Figure 10.

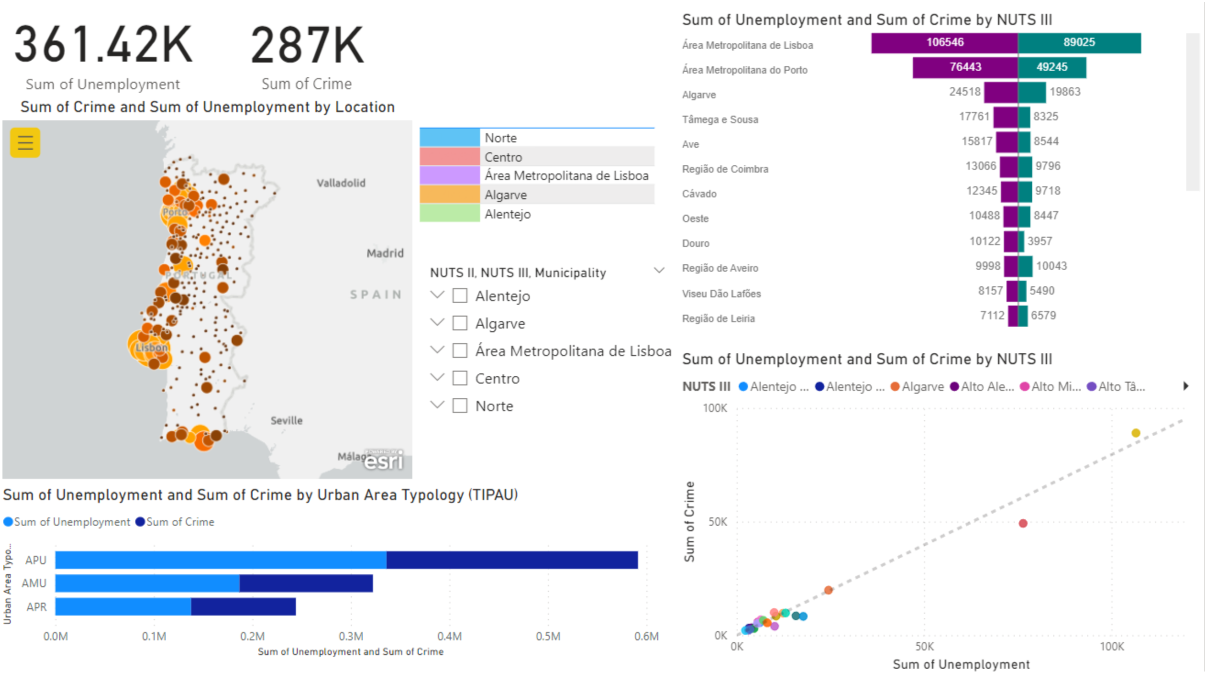

Figure 10 - Assembled dashboard regarding the comparison between the variables “unemployment” and “crime”.

IV.2.3. Resident population density and average monthly income by education level

The ninth dashboard of this project consisted in a set of graphical representations to visualise the variables termed “population density” and “income”. To accomplish this, the following visualisation elements were created in PowerBI:

- Two cards, the one on the left indicating the average population density and the one on the right showing the average monthly income in Portugal.
- An ArcGIS Map that shows the geographical distribution of population density in Portugal according to the average monthly income. The bubble size depicts the population density for a particular municipality, while their colouring represents the average monthly income (with brown indicating lower values and yellow higher values for this variable). Here, it is possible to see that in AML and AMP, a positive correlation exists between greater population density and higher income.
- A stacked bar chart that portrays the population density and average monthly income in Portugal according to TIPAU. Here, it is possible to observe that while the monthly income does not appear to vary between the three typologies, a significantly greater population density is observed in APU.
- A tornado chart that represents the population density and average monthly income in Portugal according to the various NUTS III regions of this country. Here, it is again possible to see that the highest monthly incomes are associated with the more densely populated AML and AMP. However, this positive correlation does not seem to extend to other areas of the country. For instance, in Aveiro the monthly income is practically the same as the one in Porto, but the population density of the former is approximately 5 times lower than that of the latter.

- A slicer that allows the information concerning the population density and average monthly income in Portugal to be interactively filtered according to the NUTS II, NUTS III regions and municipality of the country.

The information above was introduced in the ninth dashboard and the end result is shown in Figure 11.

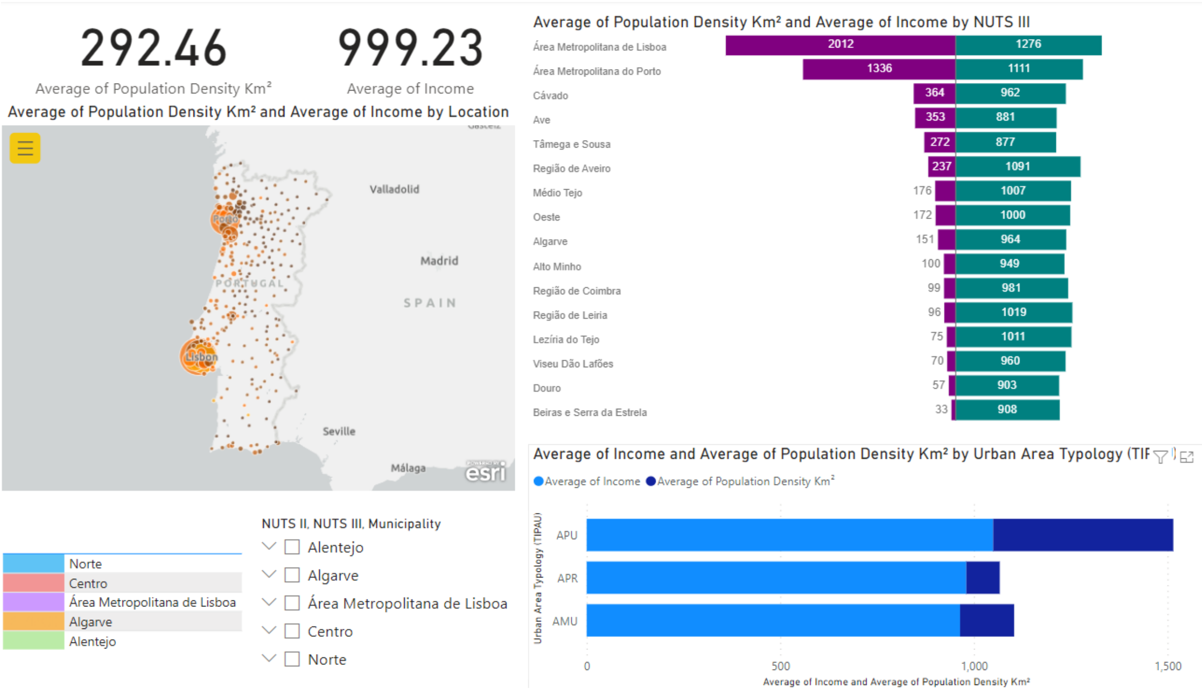

Figure 11 – Assembled dashboard regarding the “population density” and “income”.

IV.2.4. Total crimes registered by the police and average monthly income by education level

The tenth dashboard of this project consisted in a set of graphical representations to visualise the variables termed “crime” and “income”. To this end, the following visualisation elements were created in PowerBI:

- Two cards, the one on the left indicating the total number of crimes reported in Portugal and the one on the right showing the average monthly income in the country.
- An ArcGIS Map that shows the geographical distribution of crimes registered in Portugal according to the average monthly income. The bubble size depicts the total number of crimes registered for a particular municipality, while their colouring represents the respective average monthly income (with brown indicating lower values and yellow higher values for this variable). Here, it is possible to observe that a positive correlation between higher income and greater number of registered crimes exists in the AML and AMP.

- A stacked bar chart that portrays the average monthly income and the average number of crimes registered in Portugal according to TIPAU. Here, it is possible to see that most crimes are registered in APU, even though significant differences in monthly incomes do not exist between urban and rural areas.
- A tornado chart that represents the average monthly income and the average number of crimes registered in Portugal according to the various NUTS III regions of this country. Here, it is again possible to observe that a positive correlation between the monthly income and the number of registered crimes exists in AML and AMP. However, a correlation between these two variables does not appear to occur in other areas. For instance, in Alentejo Litoral the average monthly income is almost the same as the one in AML, but the average number of crimes registered in the former is approximately 8 times lower than that of the latter.
- A scatter chart that depicts on the x-axis the average monthly income and on the y-axis the average number of crimes according to the NUTS III regions, displayed by varying colours. Here, it is again possible to see that apart from AML and AMP, where a positive correlation exists between higher income and greater crime levels, these two variables do not appear to be correlated in other regions of the country.
- A slicer that allows the information concerning the average monthly income and the crime levels in Portugal to be interactively filtered according to the NUTS II, NUTS III regions and municipality of the country.

The information above was introduced in the tenth dashboard and the end result is shown in Figure 12.

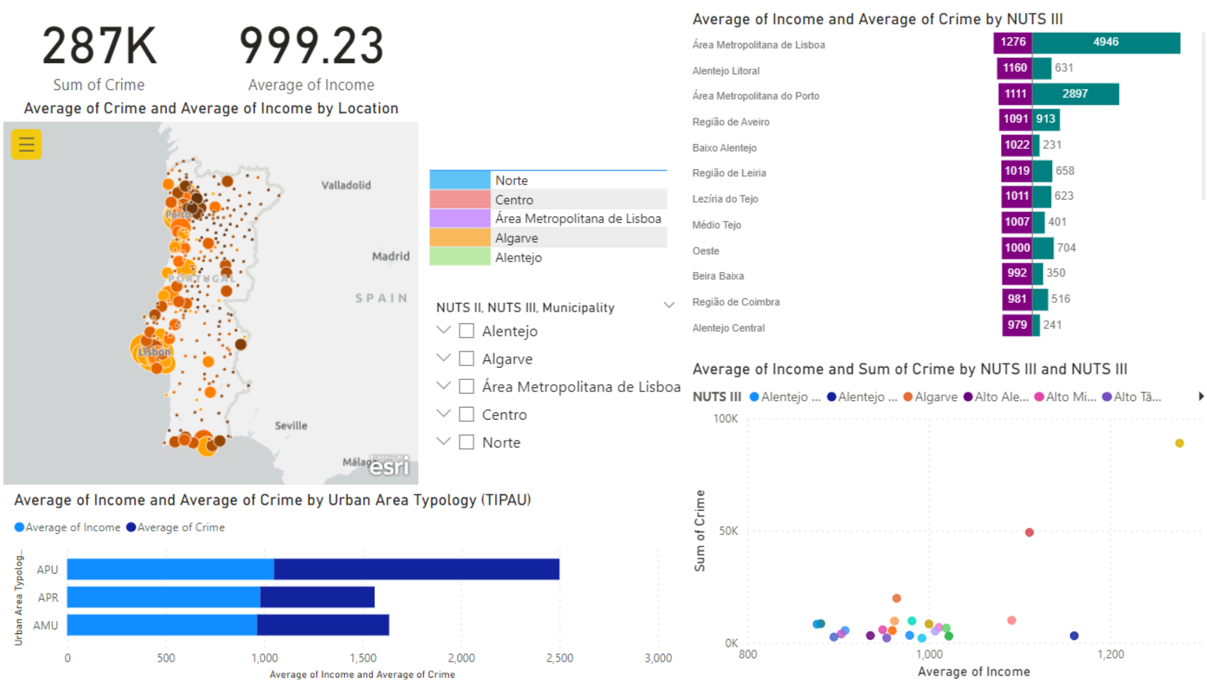

Figure 12 – Assembled dashboard regarding the variables “crime” and “income”.

IV.2.5. Total crimes registered by the police and unemployed population by education level

The eleventh dashboard of this project consisted in a set of graphical representations to visualise the variables termed “crime” and “unemployed by education level”. To this end, the following visualisation elements were created in PowerBI:

- Seven cards, with six of them pertaining to the average percentages of unemployed individuals by their respective education levels, and one concerning the total number of crimes registered in Portugal.
- A line-stacked bar chart where the line represents the crimes registered in each NUTS III region and the stacked columns correspond to the average percentages of unemployed individuals by their education level. Here, it is possible to conclude that both variables are not correlated. Since AML shows the highest number of registered crimes (89025) and the largest percentage of individuals with a higher education (19.52%), one could assume that these variables are correlated. However, there are several regions where this correlation does not occur. For instance, Aveiro has a high percentage of unemployed individuals with higher education (21.29%), but the number of registered crimes (10043) is approximately 10 times lower than in AML.
- A slicer that allows the information concerning the average education level of individuals and the total of crimes registered in Portugal to be interactively filtered according to the NUTS II, NUTS III regions and municipalities of the country.

The information above was introduced in the eleventh dashboard and the end result is shown in Figure 13.

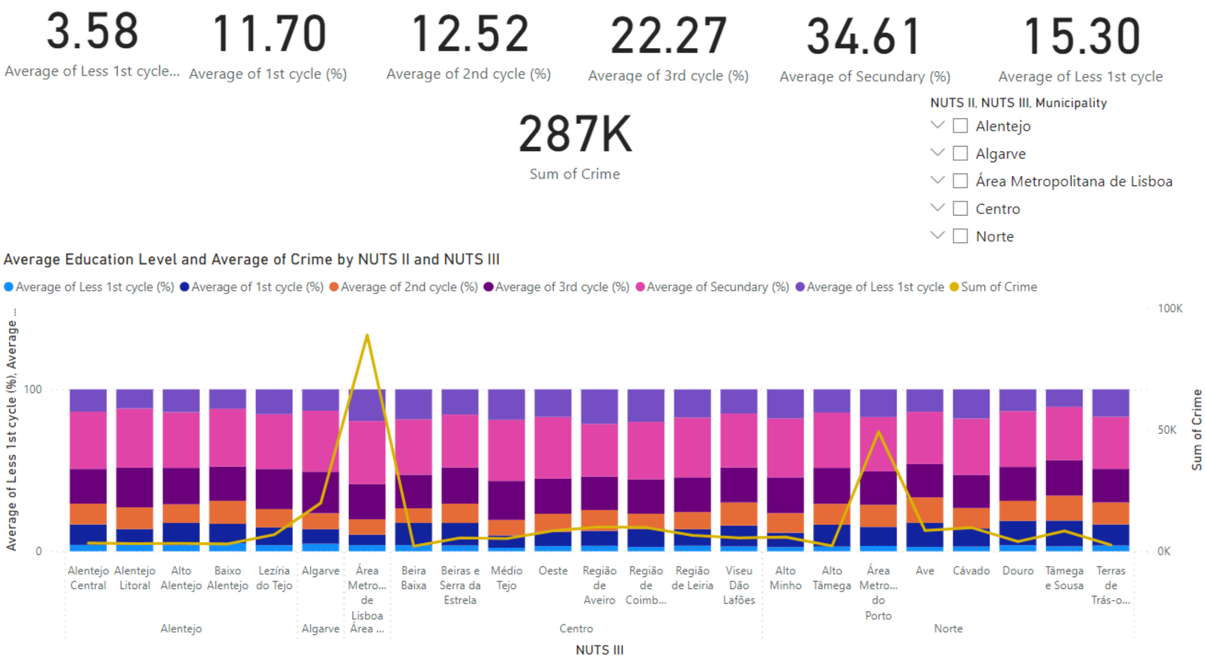

Figure 13 – Assembled dashboard regarding the variables “crime” and “unemployed by education level”.

V. Limitations of the PowerBI software

In this section, the limitations and inferred shortcomings felt throughout the development of this project using the PowerBI software will be identified.

The first limitation relates to the excessive effort required to create relationships between different datasets in PowerBI. Instead of allowing users to create relationships between distinct datasheets in a single attempt, this process had to be repeated several times for each intended relationship. Furthermore, the auto-detect feature of PowerBI that allows relationships to be created automatically between equally named columns did not work. Lastly, the manual selection of columns from which relationships between distinct tables are formed could be more interactive and easily perceived. As it stands, the selection of these columns is not intuitive and confirmation of selected columns is visualized through a simple grey highlight. Implementing an option box to select the column and then showing the selected column would be far more intuitive.

A second limitation concerns the process of performing calculations on data columns. For some of the columns, percentages were calculated using the data manipulation tool, but the whole process was a lot more complex than it would have been using Excel. While this could be due to our limited knowledge regarding this tool, data manipulation seems far more cumbersome, requiring a division operation to create a new column with the resulting values, and then a multiplication to create another column with the desired output. In Excel the two operations can be done simultaneously in a single column without much effort.

A third limitation pertains to the representations that allow the visualization of more than one variable at a time. For instance, in ArcGIS, two variables can be assigned according to the size and colour of the map bubbles. When trying to alternate between the assignment of two variables to either bubble size or colour, it is first necessary to delete one variable from the “size” or “colour” field entirely or change it to another random field. While this limitation is easy to overcome, it became frustrating to do so since many ArcGIS representations were used in this project.

Still regarding the ArcGIS representation, despite it being quite useful and visually pleasant, its respective legends were not user-friendly. First, the automatically generated legend takes too much space, which complicates the creation of dashboards containing many graphical representations. Second, this legend does not even appear in the final dashboard and the user needs to directly interact with it whenever necessary to inspect it. To overcome this, a manually generated legend was created, taking a lot of time and effort to make it presentable. Another limitation seen in ArcGIS representations was that values associated with the variables assigned to size and colour of the map bubbles were sometimes wrong. After exploring this issue, it was found that this could occur when using equally named variables.

Stability problems were also felt during this project, with the PowerBI software crashing several times while making graphical representations in a dashboard already populated with five or six representations. This resulted in a slight loss of previously performed work, which was tiresome. Furthermore, for dashboards with several representations, the response time of the software also saw a hit even when using high-powered computers. As such, it is safe to assume that the software could be optimized for stability and speed.

Not so much a limitation but more of a critic of the software, despite the existence of numerous default representations, some of these are not significantly different from one another. This makes experimentation more tedious than necessary and some of these representations could be traded for others that are only obtainable through manual acquisition, such as radar charts or tornado charts.

To add to the felt limitations, others were also searched in the web from the PowerBI userbase [17]–[21]. Some commonly described limitations were file size limitations and its inability to deal with large datasets; complex software with a steep learning curve; difficulties in dealing with more than one relationship between tables; limited data cleaning solutions; restricted sharing of data.

VI. Conclusions

From the developed work it was possible to assess the behaviour of the chosen variables and investigate relationships between them.

Regarding each individual variable, the constant NUTS III region that constantly displayed higher values for variables “total resident population, “population density”, “unemployment”, “crime” and “income” was AML. Apart from “income”, the AMP was constantly a runner-up for the remaining variables. An interesting find was that Alentejo Litoral was the NUTS III region with the second highest average monthly income, above AMP. It was also possible to conclude that most unemployed individuals only completed secondary education (34.6%). Expectedly, the higher the education level of an individual, the higher its average monthly income. When analysing TIPAU, it was possible to conclude that APU registered the highest values for variables “total resident population, “population density”, “unemployment”, “crime” and “income”. APR and AMU had a slightly lower average monthly income, a slightly higher percentage of unemployed individuals with only the first cycle and a lower percentage of unemployed individuals with a higher education degree.

Following individual evaluation of each variable, pairwise relationships between them were assessed. For “crime” and “income”, it was possible to conclude that no clear correlation between these two variables exists. On the other hand, when considering “crime” and “unemployment”, a positive correlation between these two was observed, with a higher number of crimes being registered in areas with more unemployed individuals. Regarding “population density” and “income”, no correlations were seen. The number of crimes also did not appear to be correlated to neither the average monthly income nor the education level of the unemployed population.

Overall, the PowerBI software clearly showed its ability to make varied graphical representations capable of summarizing and highlighting the most relevant information. Despite being an extremely powerful software, some limitations were felt during the project. Some of these were: unintuitive interface for relationship creation; lacklustre data processing; inability to format legends in some representations; and most important of all, stability problems and astonishingly slow response time when dealing with several graphical representations in multiple dashboards.

VII. References

[1] “Municípios: População residente por concelho por sexo: Censos | Pordata.” https://www.pordata.pt/municipios/populacao+residente+segundo+os+censos+total+e+por+sexo-17 (accessed Jan. 03, 2023).

[2] “Municípios: Densidade populacional por concelho: Censos | Pordata.” https://www.pordata.pt/municipios/densidade+populacional+segundo+os+censos-591 (accessed Jan. 01, 2023).

[3] “Municípios: Desempregados inscritos nos centros de emprego e de formação profissional (média anual): total e por tipo de desemprego | Pordata.” https://www.pordata.pt/municipios/desempregados+inscritos+nos+centros+de+emprego+e+de+formacao+profissional+(media+anual)+total+e+por+tipo+de+desemprego-229 (accessed Jan. 01, 2023).

[4] “Municípios: Crimes registados pelas polícias: total e por tipo de crime | Pordata.” https://www.pordata.pt/municipios/crimes+registados+pelas+policias+total+e+por+tipo+de+crime-599 (accessed Jan. 01, 2023).

[5] “Municípios: População desempregada segundo os Censos: total e por nível de escolaridade completo | Pordata.” https://www.pordata.pt/municipios/populacao+desempregada+segundo+os+censos+total+e+por+nivel+de+escolaridade+completo-150 (accessed Jan. 01, 2023).

[6] “Municípios: Ganho médio mensal dos trabalhadores por conta de outrem: total e por nível de escolaridade | Pordata.” https://www.pordata.pt/municipios/ganho+medio+mensal+dos+trabalhadores+por+conta+de+outrem+total+e+por+nivel+de+escolaridade-288 (accessed Jan. 01, 2023).

[7] “Card visualizations (large number tiles) - Power BI | Microsoft Learn.” https://learn.microsoft.com/en-us/power-bi/visuals/power-bi-visualization-card?tabs=powerbi-desktop (accessed Jan. 03, 2023).

[8] “Create an ArcGIS map in Power BI - Power BI | Microsoft Learn.” https://learn.microsoft.com/en-us/power-bi/visuals/power-bi-visualizations-arcgis (accessed Jan. 03, 2023).

[9] “Basic area chart - Power BI | Microsoft Learn.” https://learn.microsoft.com/en-us/power-bi/visuals/power-bi-visualization-basic-area-chart (accessed Jan. 03, 2023).

[10] “GitHub - microsoft/PowerBI-visuals-RadarChart: Repo for Power BI Radar Chart custom visual.” https://github.com/Microsoft/powerbi-visuals-radarchart/ (accessed Jan. 04, 2023).

[11] “GitHub - microsoft/PowerBI-visuals-Tornado: Repo for Power BI Tornado custom visual.” https://github.com/Microsoft/powerbi-visuals-tornado/ (accessed Jan. 04, 2023).

[12] “Scatter, bubble, and dot plot charts in Power BI - Power BI | Microsoft Learn.” https://learn.microsoft.com/en-us/power-bi/visuals/power-bi-visualization-scatter?tabs=powerbi-desktop (accessed Jan. 03, 2023).

[13] "Doughnut charts in Power BI - Power BI | Microsoft Learn." https://learn.microsoft.com/en-us/power-bi/visuals/power-bi-visualization-doughnut-charts?tabs=powerbi-desktop (accessed Jan. 03, 2023).

[14] "Combo chart in Power BI - Power BI | Microsoft Learn." https://learn.microsoft.com/en-us/power-bi/visuals/power-bi-visualization-combo-chart?tabs=powerbi-desktop (accessed Jan. 03, 2023).

[15] "Slicers in Power BI - Power BI | Microsoft Learn." https://learn.microsoft.com/en-us/power-bi/visuals/power-bi-visualization-slicers?tabs=powerbi-desktop (accessed Jan. 03, 2023).

[16] "Table visualizations in Power BI reports and dashboards - Power BI | Microsoft Learn." https://learn.microsoft.com/en-us/power-bi/visuals/power-bi-visualization-tables?tabs=powerbi-desktop (accessed Jan. 03, 2023).

[17] "Disadvantages and Limitations of Power BI - Microsoft Power BI Community." https://community.powerbi.com/t5/Desktop/Disadvantages-and-Limitations-of-Power-BI/m-p/397939 (accessed Jan. 03, 2023).

[18] "What are Power BI disadvantages and limitations?" https://www.knowledgehut.com/blog/business-intelligence-and-visualization/power-bi-limitations (accessed Jan. 03, 2023).

[19] "Limitation of Microsoft Power BI: Top Limitation listed - EPCGroup." https://www.epcgroup.net/limitations-of-power-bi/ (accessed Jan. 03, 2023).

[20] "Limitations Of Power Bi - DataScienceCentral.com." https://www.datasciencecentral.com/limitations-of-power-bi/ (accessed Jan. 03, 2023).

[21] "Pros and Cons of Power BI - The Bright & the Dull side of visualization suite - DataFlair." https://data-flair.training/blogs/power-bi-advantages-and-disadvantages/ (accessed Jan. 03, 2023).