Crime In Portugal: A Data Analysis for a Security Company

Index:

* Introduction:
  + Dataset
  + List of questions
* Data Analysis:
  + Data Preprocessing
  + Exploratory Data Analysis
  + Creating the Visualizations
* Results and Findings
* Conclusions and Recommendations

Introduction:

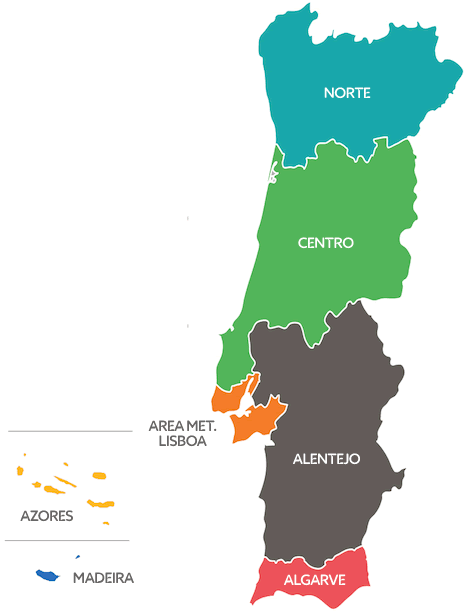
In an increasingly interconnected world, security has become a top priority for both individuals and businesses. For a company in the security alarm industry, gaining insights into the local crime landscape is crucial. This project is driven by the necessity for a security alarm company to understand the crime situation in Portugal as it establishes their presence in this Mediterranean country. The company's vision extends to three major cities—Lisbon, Porto, and Algarve—where they intend to open offices. However, they question whether these are the optimal locations to open the new offices?

Figure - Regions of Portugal.

**Dataset**

To achieve this goal, I used a Kaggle dataset that has the registered number of crimes in the country of Portugal by different locations and from the year 2011 to the 2019.

The dataset, sourced from Kaggle, provides a comprehensive overview of crime statistics in Portugal spanning from 2011 to 2019. This comprehensive dataset comprises of 166 columns and 312 rows, it offers a wealth of data points that paint a detailed picture of crime across different regions (figure 1). The columns are Territory, Region, Intermunicipallity, Municipallity and the subsequent columns are the number of crimes committed separated by different type of crimes and years (figure 2). The first four are text datatype values and the rest are numbers type.

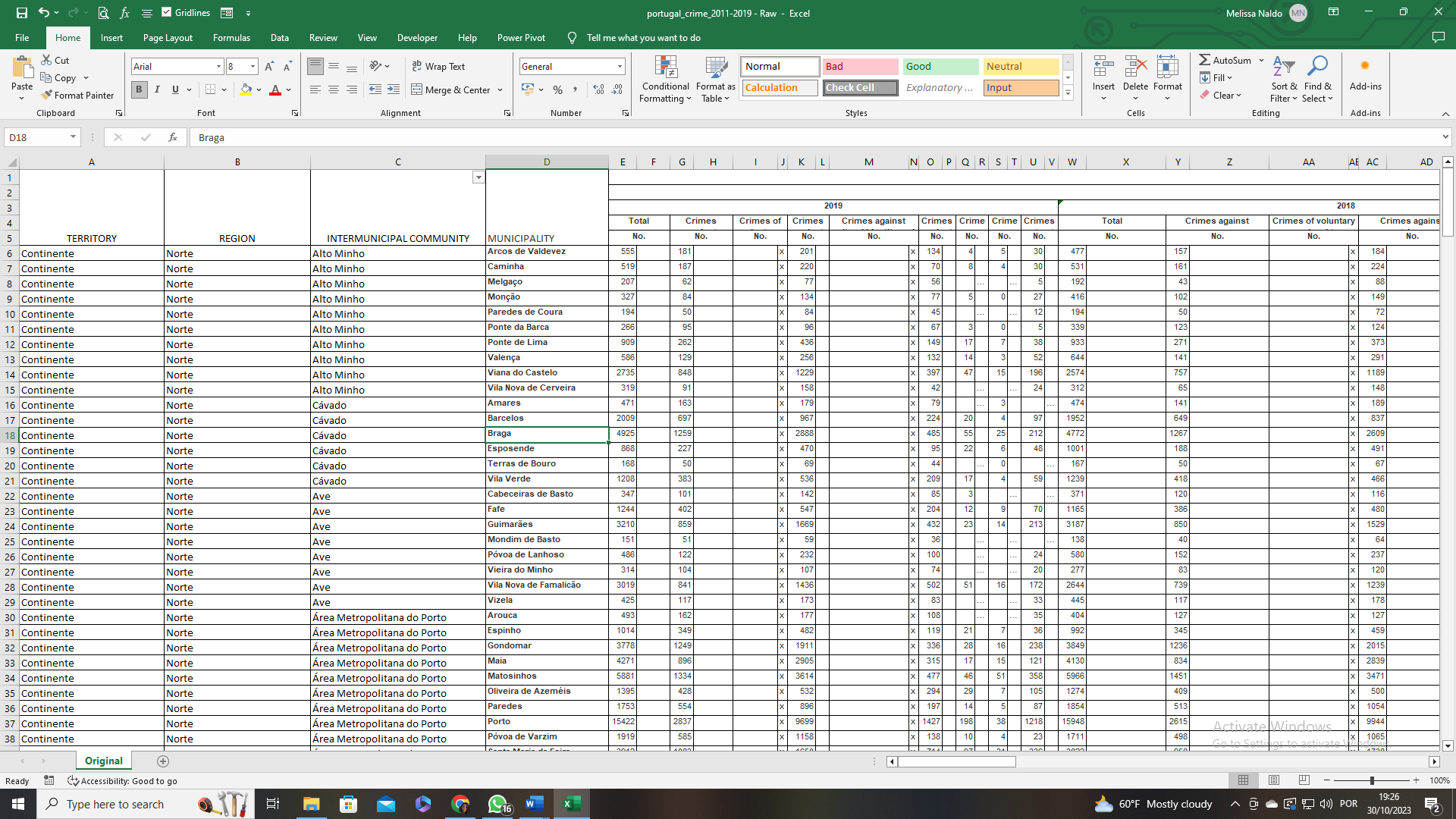


Figure - original dataset (link to original dataset in github).

The dataset can be accessed through the following link: https://www.kaggle.com/datasets/sameerkulkarni91/portugal-crime-statistics-20112019?select=portugal\_consolidated\_2011-2019.xlsx

**List of questions**

This analysis is guided by a set of pertinent questions that aims to uncover the underlying patterns in crime data and inform the company's strategic decisions:

1. How many crimes were reported in Portugal from 2011 to 2019?
2. What is the extent of crimes against patrimony within the same period?
3. How do crime rates vary across regions, and what is the average crime rate in each region?
4. How has crime evolved over the years?
5. What are the top 10 municipalities with the highest crime rates?
6. What are the top 3 locations to open the offices that would be most beneficial for the company?

Data Analysis:

**Data Preprocessing**

After an initial examination of the dataset, several preprocessing steps were necessary to ensure its suitability for analysis. The following actions were taken:

***Column and row removal:***

No duplicate rows were identified in the dataset. Columns with no meaningful information, marked with "x" or "-", were removed. The "Total" column was also eliminated as it did not contribute to the analysis. Subsequently, the data was transformed by combining the years column with the crime type, creating a single header for these columns (figure 3).

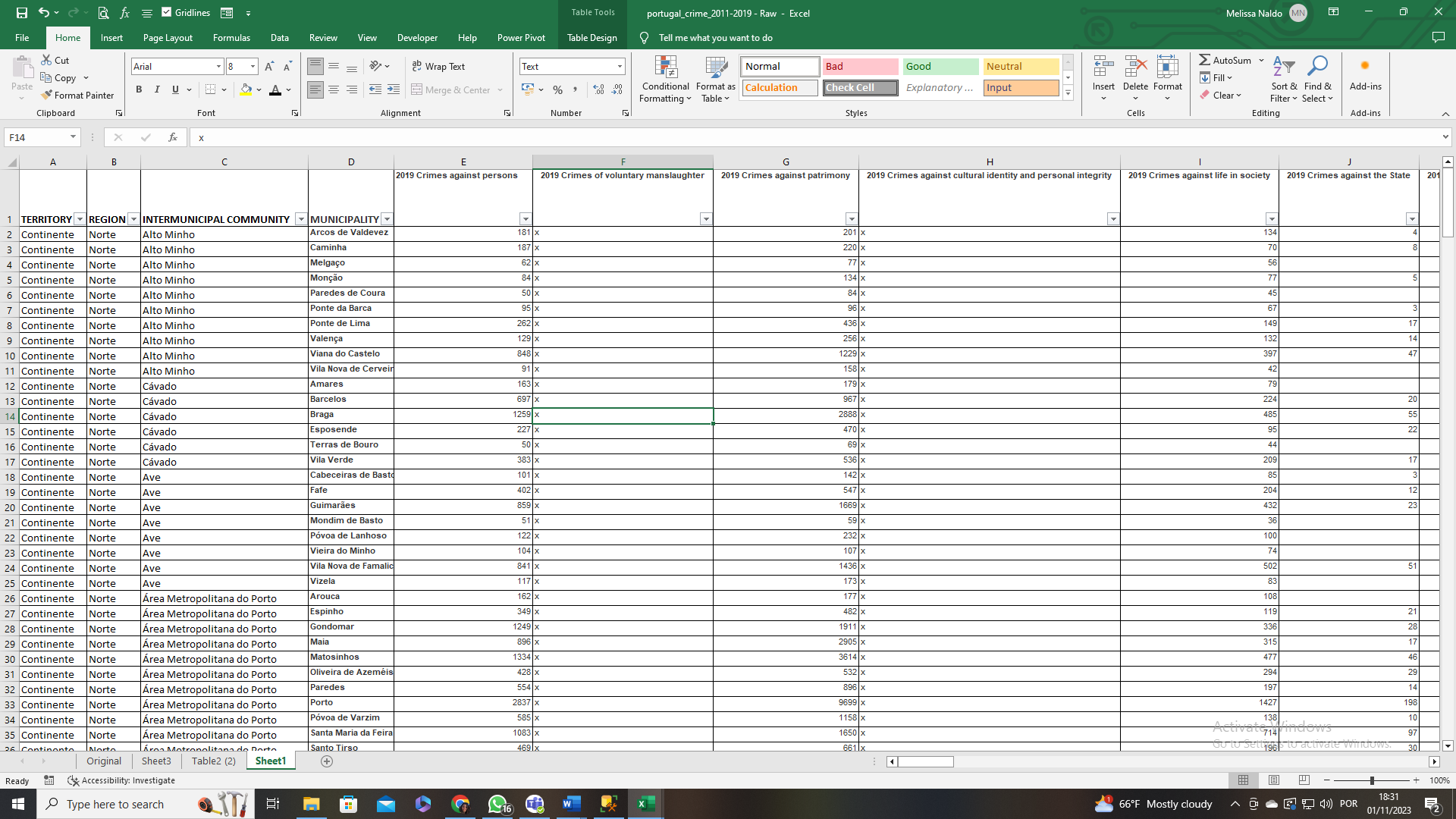


Figure - Dataset with single headers.

***Unpivoting Columns:***

After finishing that part of the process, I employed the Power Query Editor to help me unpivot my columns, resulting in a more structured dataset. Next, I used the Left(), Len() and the Right() functions to separate the years from the type of crimes. With this, I now had created distinct columns for crime types and years, leading to the following image (figure 4):

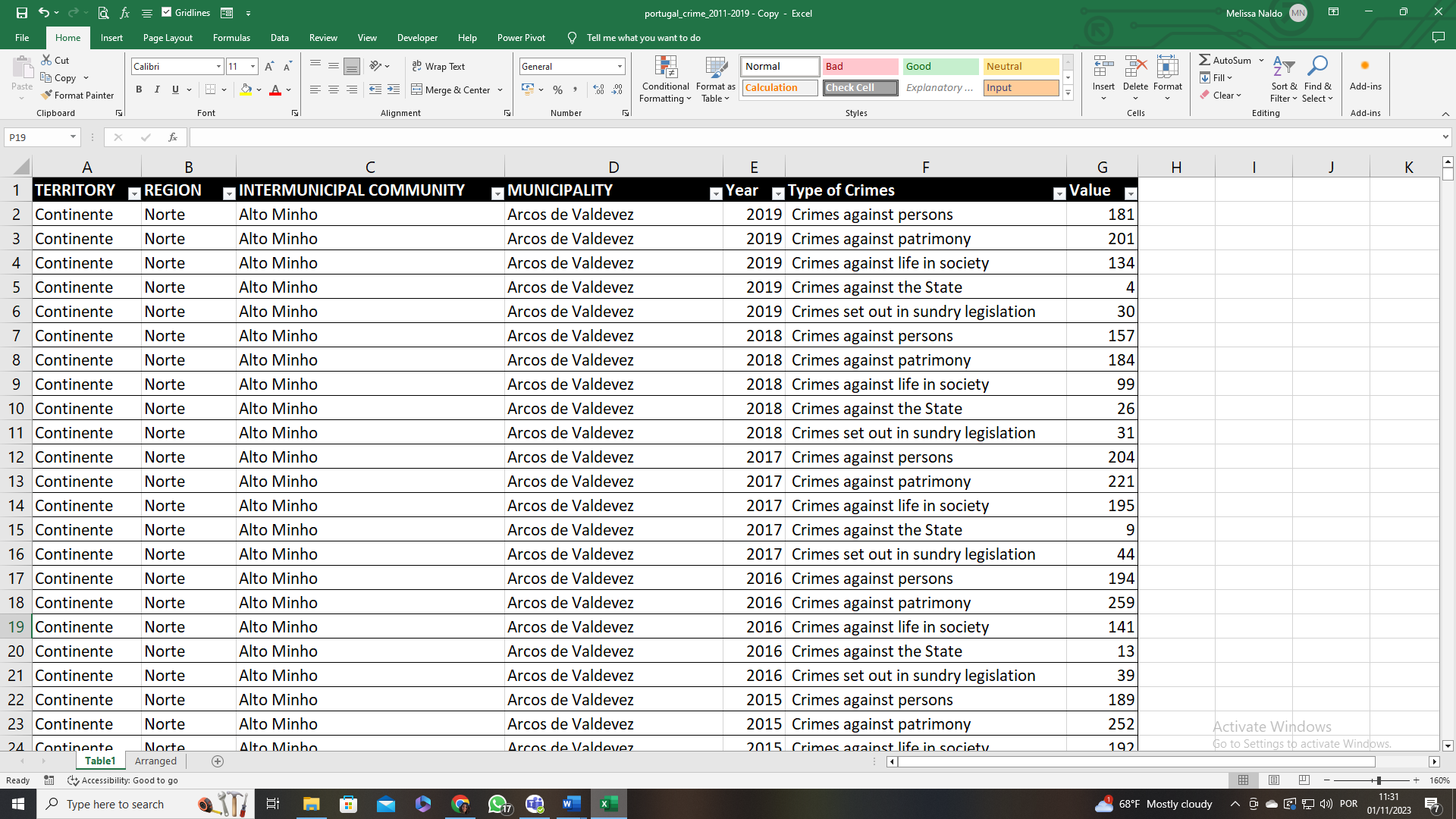


Figure - Dataset after unpivot columns.

***Data Standardization:***

To ensure uniformity, text values were standardized. Furthermore, misspellings were corrected and some text values were altered to enhance precision (figure 5):

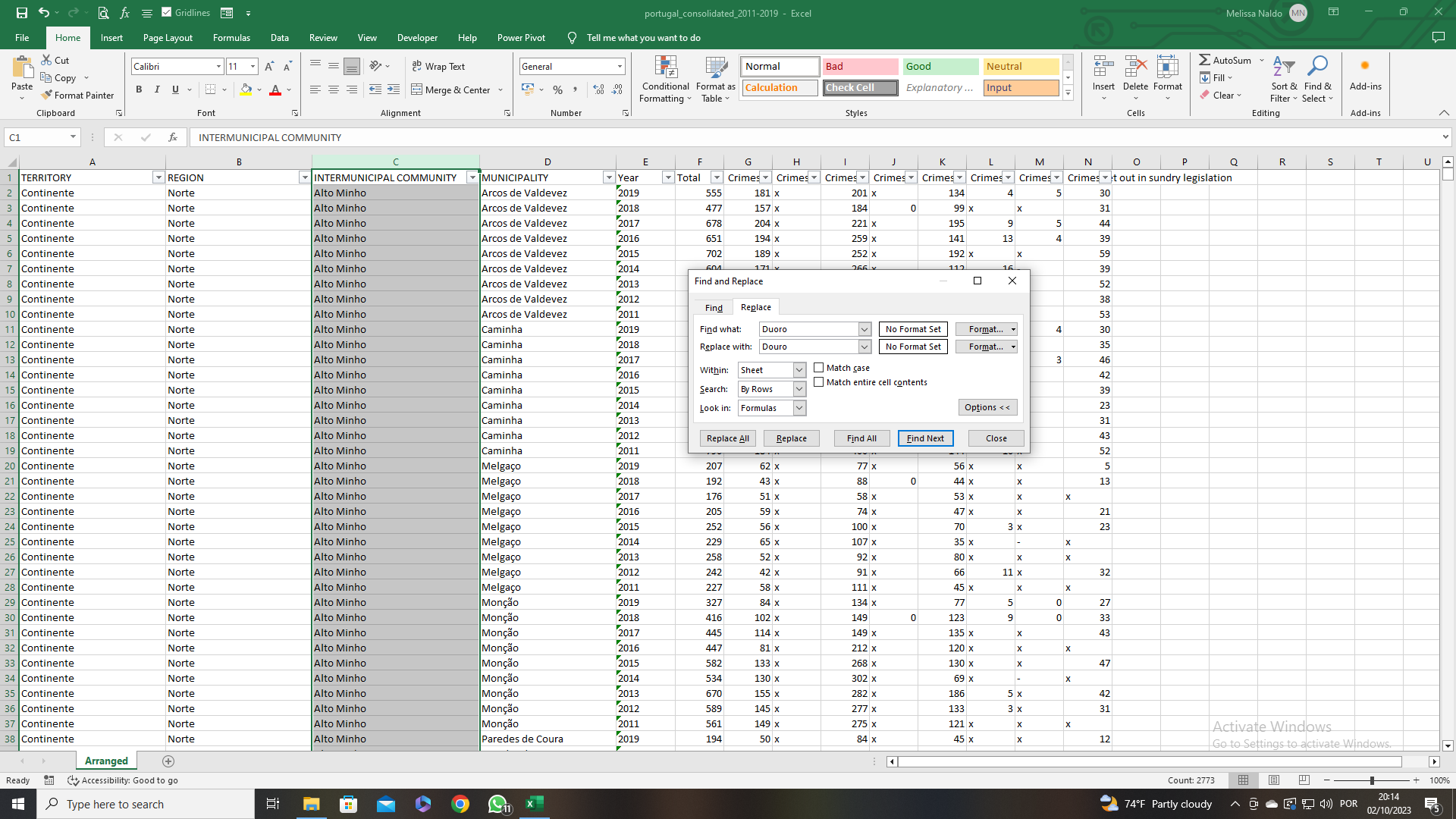
* "Região Autónoma da Madeira" to "RA Madeira";
* "Região Autónoma dos Açores" to "RA Açores";
* "Área Metropolitana de Lisboa" to "AM Lisboa";
* "Área Metropolitana do Porto" to "AM Porto";
* "Duoro" to "Douro".

Figure - Find & Replace feature.

***Handling Missing Values:***

Several “Types of Crime” had missing values. These categories were "voluntary manslaughter," "cultural identity & personal integrity," "pet animals," "life in society," "the State," and "set out in sundry legislation." The first three columns had extensive missing data and were removed as they did not contribute significantly to the project's objectives. For the remaining columns, the missing values were replaced with the mean values, which were similar to the mode and median, and these values were: 148 for "crime against life in society," 26 for "crimes against the State," and 96 for "crimes set out in sundry legislation."

***Column Name Adjustment:***

With this, I finalized the cleaning part of the analysis by changing my column names (figure 6).

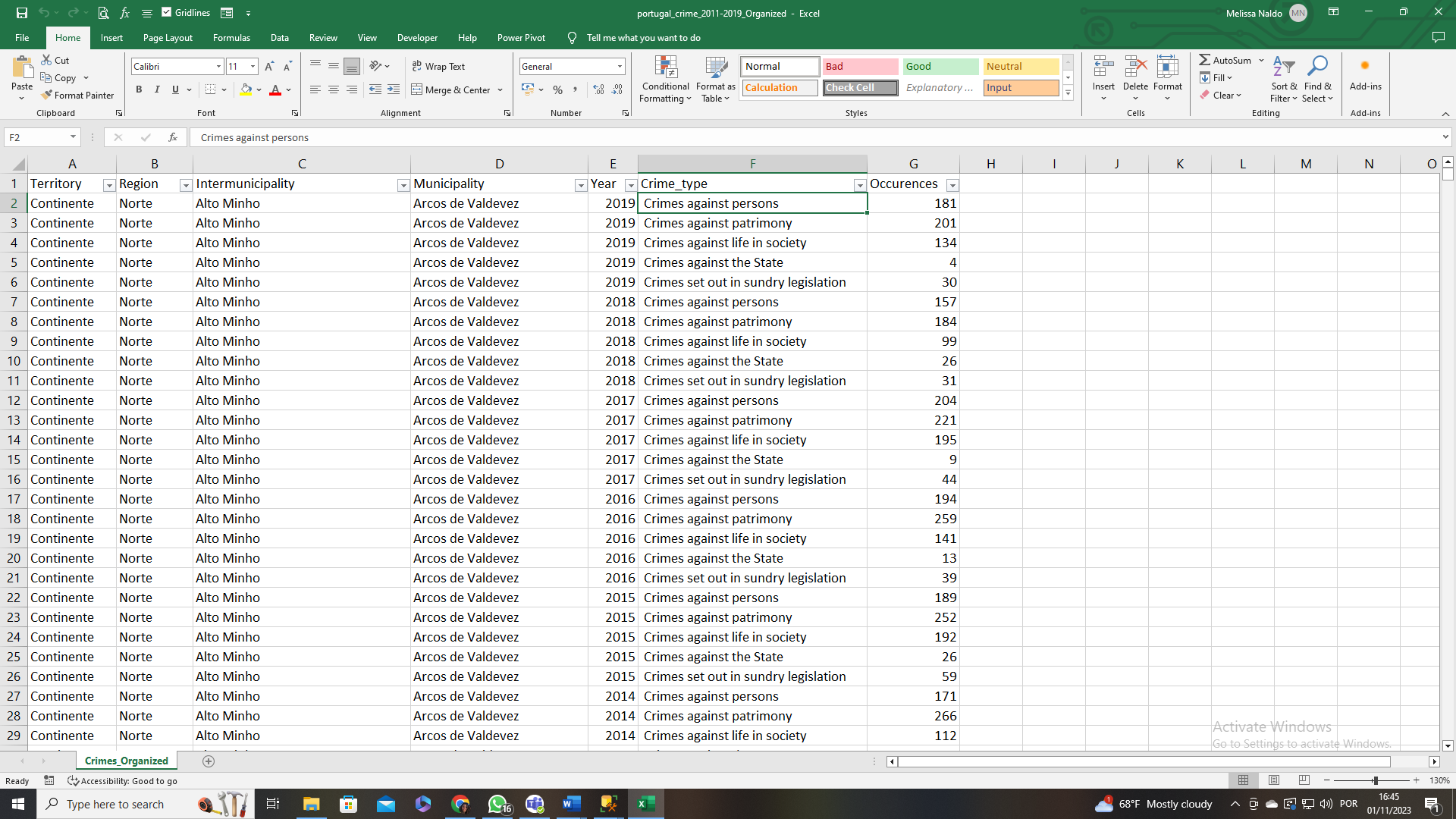


Figure - Cleaned dataset.

You can access the initial dataset here(https://github.com/Melissa-Naldo/Crime\_In\_Portugal/blob/main/portugal\_crime\_2011-2019%20-%20Raw.xlsx) and the cleaned dataset here(https://github.com/Melissa-Naldo/Crime\_In\_Portugal/blob/main/portugal\_crime\_2011-2019\_Organized.xlsx) for reference.

Exploratory Data Analysis:

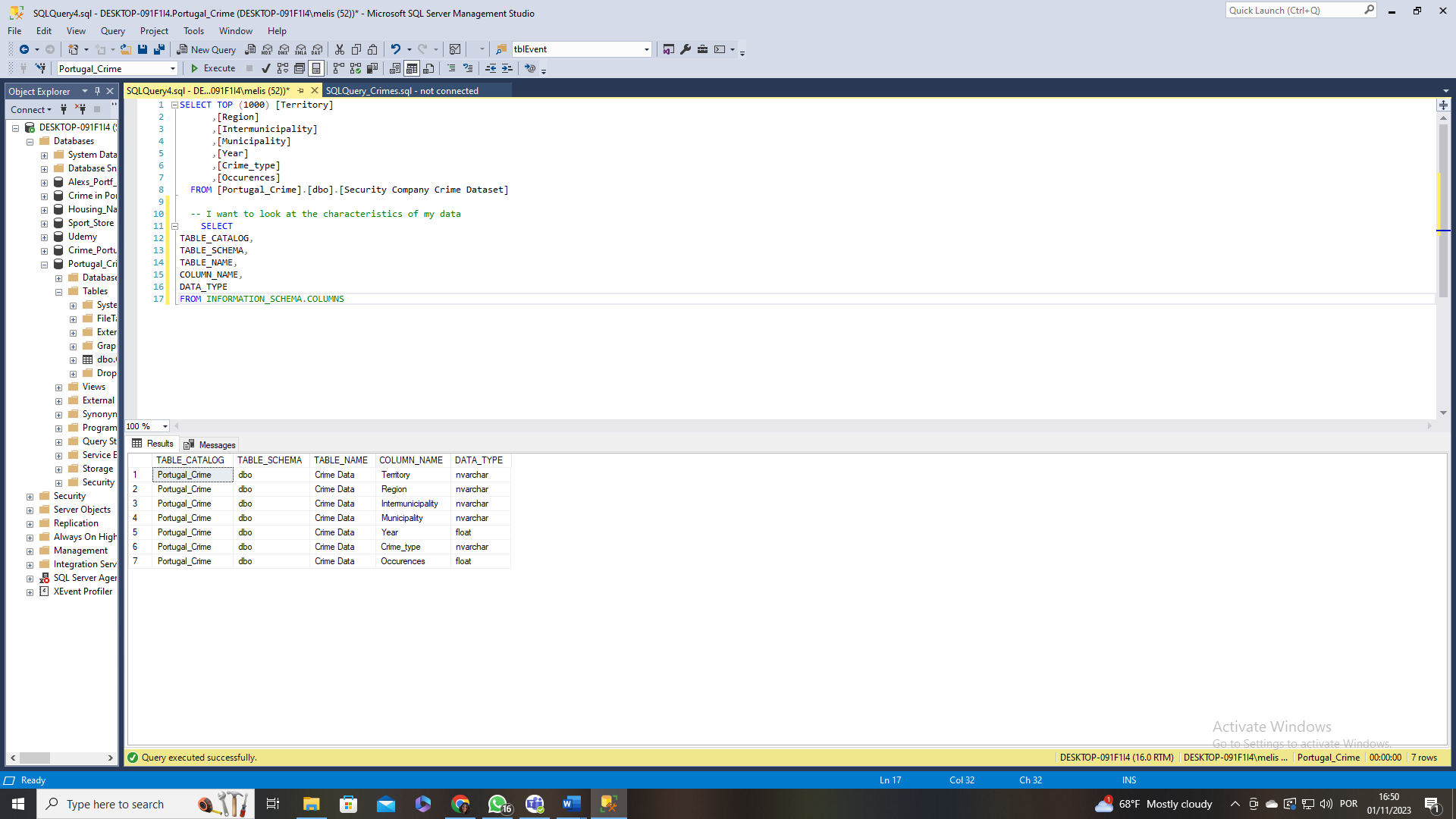
This data analysis project focuses on a security company seeking to understand crime values in Portugal, especially crimes against patrimony (includes house burglaries). You can find my SQL query file on: https://github.com/Melissa-Naldo/Crime\_In\_Portugal/blob/main/SQL\_Crime.sql.

Figure - Different columns and their data type.

So, the first thing I did was look at my table to understand my data (figure 7). By using:

SELECT

TABLE\_CATALOG,

TABLE\_SCHEMA,

TABLE\_NAME,

COLUMN\_NAME,

DATA\_TYPE

FROM INFORMATION\_SCHEMA.COLUMNS

The dataset primarily consists of nvarchar columns, except for the "Year" and "Occurrences," which are of float type.

**General data analysis:**

To gain a general view before focusing on the project's objectives, various queries were performed:

* Total amount of crimes in Portugal from 2011 to 2019

Select Sum(Occurences) as Total\_Crimes, round(AVG(Occurences), 2) as Avg\_Crimes

From [dbo].[Crime Data]

* The amount of crime per region

Select Region, SUM(Occurences) AS Sum\_Crimes\_per\_Region, round(AVG(Occurences), 2) AS AVG\_Crimes\_per\_Region, Rank() OVER (Order by SUM(Occurences) desc) as Rank

From [dbo].[Crime Data]

Group by Region

Order by AVG(Occurences) desc

* The amount of crime per year

Select Year, Sum(Occurences) as Total\_Crimes\_Year, lag(Sum(Occurences)) Over (Order by Year) as Pre\_Year\_Crime, ((((Sum(Occurences))-(lag(Sum(Occurences)) OVER (Order by Year)))/(lag(Sum(Occurences)) OVER (Order by Year)))\*100) as Percentage\_Difference

From [dbo].[Crime Data]

Group by Year

Order by Year

Select Year, Sum(Occurences) as Total\_Crimes\_Year, lag(Sum(Occurences)) Over (Order by Year) as Pre\_Year\_Crime, ((((Sum(Occurences))-(lag(Sum(Occurences)) OVER (Order by Year)))/(lag(Sum(Occurences)) OVER (Order by Year)))\*100) as Percentage\_Difference

From [dbo].[Crime Data]

Where Year in (2011, 2016, 2017, 2019)

Group by Year

Order by Year

I found it essential to delve into the evolution of total crime across various regions and over the years. Upon analysing the comprehensive data obtained from this inquiry, I chose to focus on specific regions, namely Norte, AM Lisboa, Algarve—regions where the company plans to open offices. Additionally, I included Centro, a region that exhibited notably high crime values. The subsequent query was used:

Select Region, Year, SUM(Occurences) AS Sum\_Crimes\_per\_Region, round(AVG(Occurences), 2) AS AVG\_Crimes

From [dbo].[Crime Data]

Where Region In ('Norte', 'AM Lisboa', 'Algarve', 'Centro')

Group by Region, Year

Order by Sum(Occurences) desc

**Crime against patrimony:**

After looking at the whole data, I decided to take a further step and analyse my data in a different view. I wanted to see these similar values, but applied to just my crime against patrimony. With this in mind I made queries to find the following:

* My total crimes against patrimony

Select Crime\_type, Sum(Occurences) as Total\_Patrimony\_Crimes, round(AVG(Occurences), 2) as Avg\_Crimes

From [dbo].[Crime Data]

Group by Crime\_type

* The crime against patrimony since 2011 to 2019

Select Year, Sum(Occurences) as Total\_Patrimony\_Crimes, (lag(Sum(Occurences)) OVER (Order by Year)) as Pre\_Year\_TotalCrime, ((((Sum(Occurences))-(lag(Sum(Occurences)) OVER (Order by Year)))/(lag(Sum(Occurences)) OVER (Order by Year)))\*100) as Percentage\_Difference

From [dbo].[Crime Data]

Where Crime\_type = ' Crimes against patrimony '

Group by Year

Order by Year

Select Year, Sum(Occurences) as Total\_Patrimony\_Crimes, (lag(Sum(Occurences)) OVER (Order by Year)) as Pre\_Year\_TotalCrime, ((((Sum(Occurences))-(lag(Sum(Occurences)) OVER (Order by Year)))/(lag(Sum(Occurences)) OVER (Order by Year)))\*100) as Percentage\_Difference

From [dbo].[Crime Data]

Where Crime\_type = ' Crimes against patrimony ' and Year in (2011, 2019)

Group by Year

Order by Year

* My top four regions with most crime against patrimony

Select Region, Sum(Occurences) as Total\_Patrimony\_Crimes, round(AVG(Occurences), 2) as Avg\_Patrimony\_Crimes

From [dbo].[Crime Data]

Where Crime\_type = ' Crimes against patrimony '

Group by Region

Order by SUM(Occurences) desc

* Evolution of crimes against patrimony in the top 4 regions with most crime along the 9 years.

Select Region, Year, SUM(Occurences) AS Sum\_Patrimony\_Crimes, round(AVG(Occurences), 2) AS AVG\_Crimes\_patrimony

From [dbo].[Crime Data]

Where Crime\_type = ' Crimes against patrimony '

Group by Region, Year

Having Region In ('Norte', 'AM Lisboa', 'Algarve', 'Centro')

Order by Sum(Occurences) desc

* Main municipalities with the most crime in 2019:

Select Year,

Region,

Municipality,

Occurences

From [dbo].[Crime Data]

Where Crime\_type = ' Crimes against patrimony 'AND Year=2019

Order by Occurences desc

**Data correction:**

However, taking a closer look I can see that my average values for the total amount of crime and for the crime types are extremely low. This happened because SQL is dividing my total value with all my values from the different locations. Hence, I need to find out my average value. I used the following queries to do such:

----Total

--Correcting the overall average

With Total\_crime\_avg as(Select

Year,

Sum(Occurences) as total

From [dbo].[Crime Data]

Group by Year

)

Select round(avg(total), 2) as avg\_crimes\_per\_year

From Total\_crime\_avg

-- Correcting the average per Region

Select Region, Count(Distinct(Municipality)) as Nº\_municipalities, SUM(Occurences) AS Sum\_Crimes, round((SUM(Occurences)/Count(Distinct(Municipality))), 2) AS AVG\_Crimes

From [dbo].[Crime Data]

Group by Region

Order by round((SUM(Occurences)/Count(Distinct(Municipality))), 2) desc

-----Patrimony - I'm going to do the same correction but focused on my patrimony crime

--Correcting the overall patrimony crime average

With Patrimony\_crime\_avg as (Select

Year,

Sum(Occurences) as total

From [dbo].[Crime Data]

Where Crime\_type = ' Crimes against patrimony '

Group by Year

)

Select round(avg(total), 2) as avg\_patrimony\_crimes

From Patrimony\_crime\_avg

--Correcting my overall average per region

Select Region, Count(Distinct(Municipality)) as Nº\_municipalities, SUM(Occurences) AS Sum\_Crimes, round((SUM(Occurences)/Count(Distinct(Municipality))), 2) AS AVG\_Crimes

From [dbo].[Crime Data]

Where Crime\_type = ' Crimes against patrimony '

Group by Region

Order by round((SUM(Occurences)/Count(Distinct(Municipality))), 2) desc

-- Correcting the average per region and year

Select Year, Region, Count(Distinct(Municipality)) as Nº\_municipalities, SUM(Occurences) AS Sum\_Patrimony\_Crimes, round((SUM(Occurences)/Count(Distinct(Municipality))), 2) AS AVG\_Patrimony\_Crimes

From [dbo].[Crime Data]

Where Crime\_type = ' Crimes against patrimony '

Group by Region, Year

Having Region In ('Norte', 'AM Lisboa', 'Algarve', 'Centro')

Order by round((SUM(Occurences)/Count(Distinct(Municipality))), 2) desc

After finding my correct values I decided to create tables to store important data, so afterwards I could access it in the visualization phase of the project. With this I created my Updated\_Crime\_Average table with the corrected total average of crimes and the average of crimes against patrimony. I used the following query to achieve this objective:

Create Table Updated\_Crime\_Average (

Avg\_Total decimal(8, 2),

Avg\_Patrimony decimal(8, 2))

Insert into Updated\_Crime\_Average Values

(348254.72, 189197.11)

Select \*

From Updated\_Crime\_Average

I wanted to create a table with my correct crime average for the different regions and I decided to manually create this table in object explorer and manually edit the rows to input my values (figure 8).

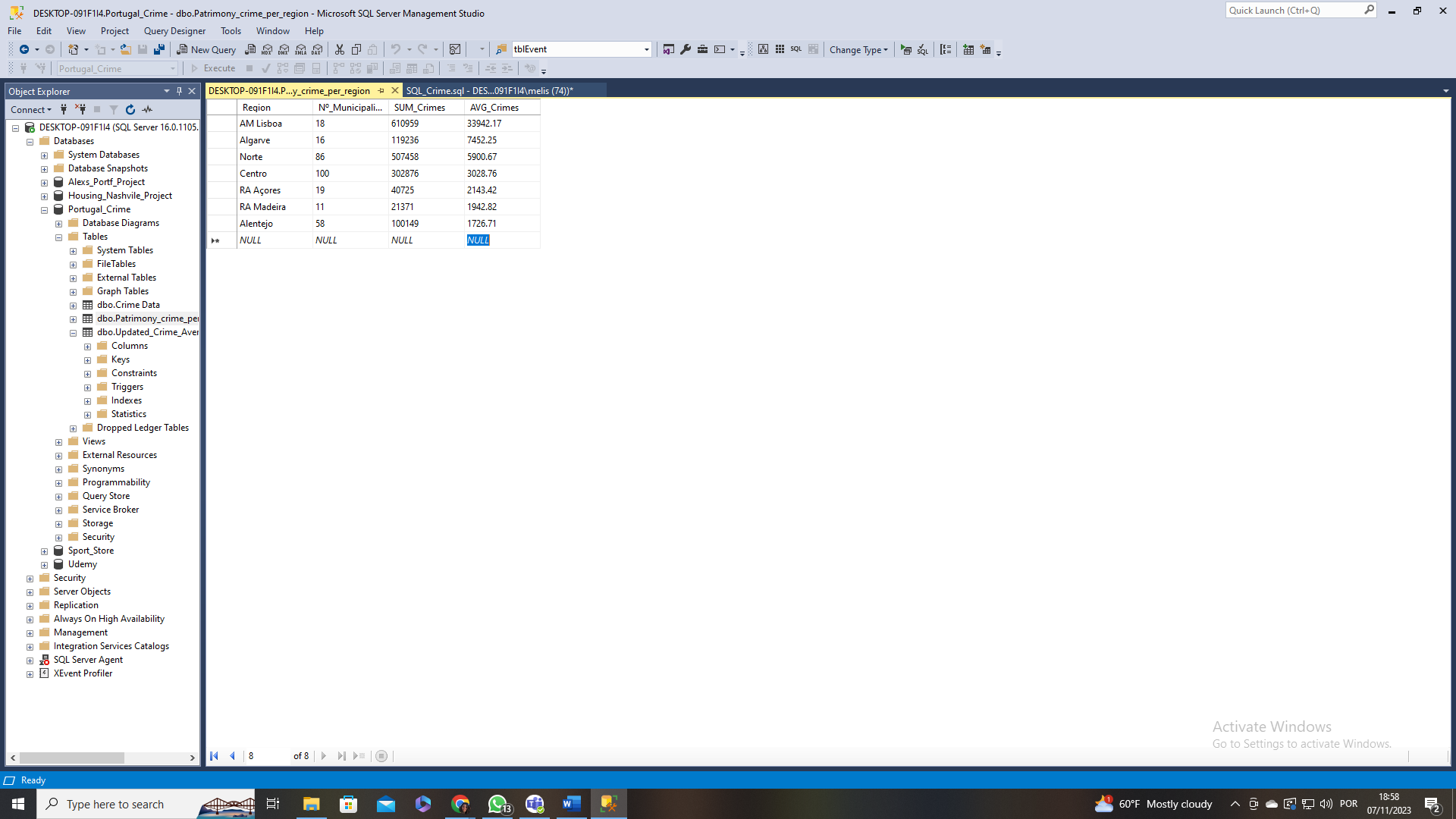


Figure - Table with updated average values for the different regions.

Creating Visualizations

In creating visualizations, I opted to organize the findings into two distinct pages on the Power BI dashboard. The first page aims to offer a comprehensive understanding of the evolution of crime rates from 2011 to 2019, while the second page directly addresses the primary goal of the project.

The data was imported from SQL to Power BI, incorporating the three main tables along with an additional table containing the coordinates for different regions in Portugal. Ensuring correct relationships between these tables, particularly using the "Region" column as a key connection, was crucial.

**Visualizations:**

Visualizations were created using a variety of elements, such as:

* Cards: Utilized to showcase crucial information.
* Line Graphs: Employed to illustrate the evolution of data across the years.
* Bar Charts: Utilized to highlight regions with the highest crime rates.
* Map Charts: Utilized to pinpoint specific locations.

**Dashboard:**

The final report is encapsulated in a Power BI dashboard, encompassing key findings and insights. This dashboard is divided into two pages, each dedicated to a specific aspect of the analysis (figure 9 and 10). You can access this file here: https://github.com/Melissa-Naldo/Crime\_In\_Portugal/blob/main/Crime\_Report.pbix

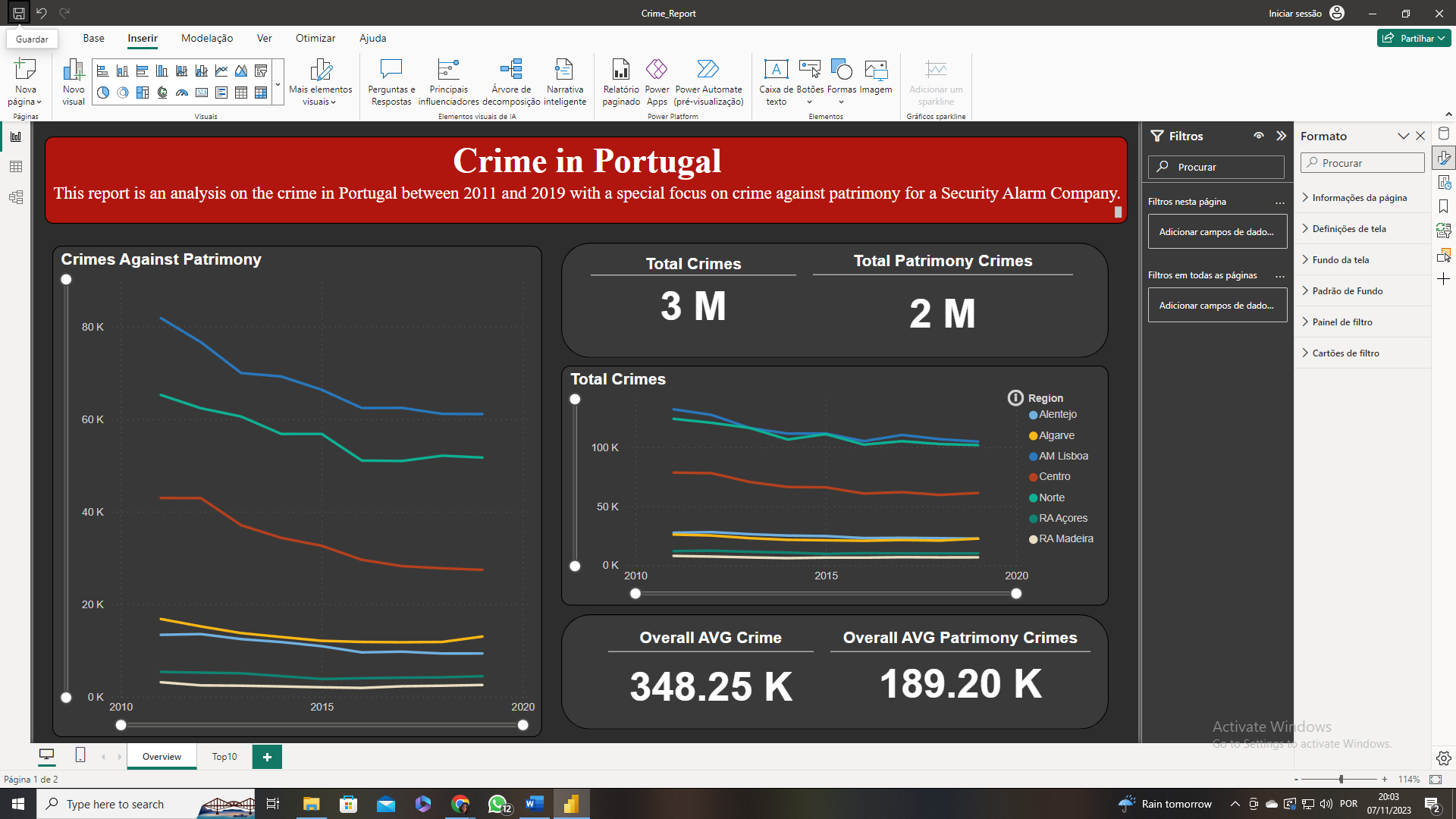


Figure 9 - First page of dashboard crime in Portugal.

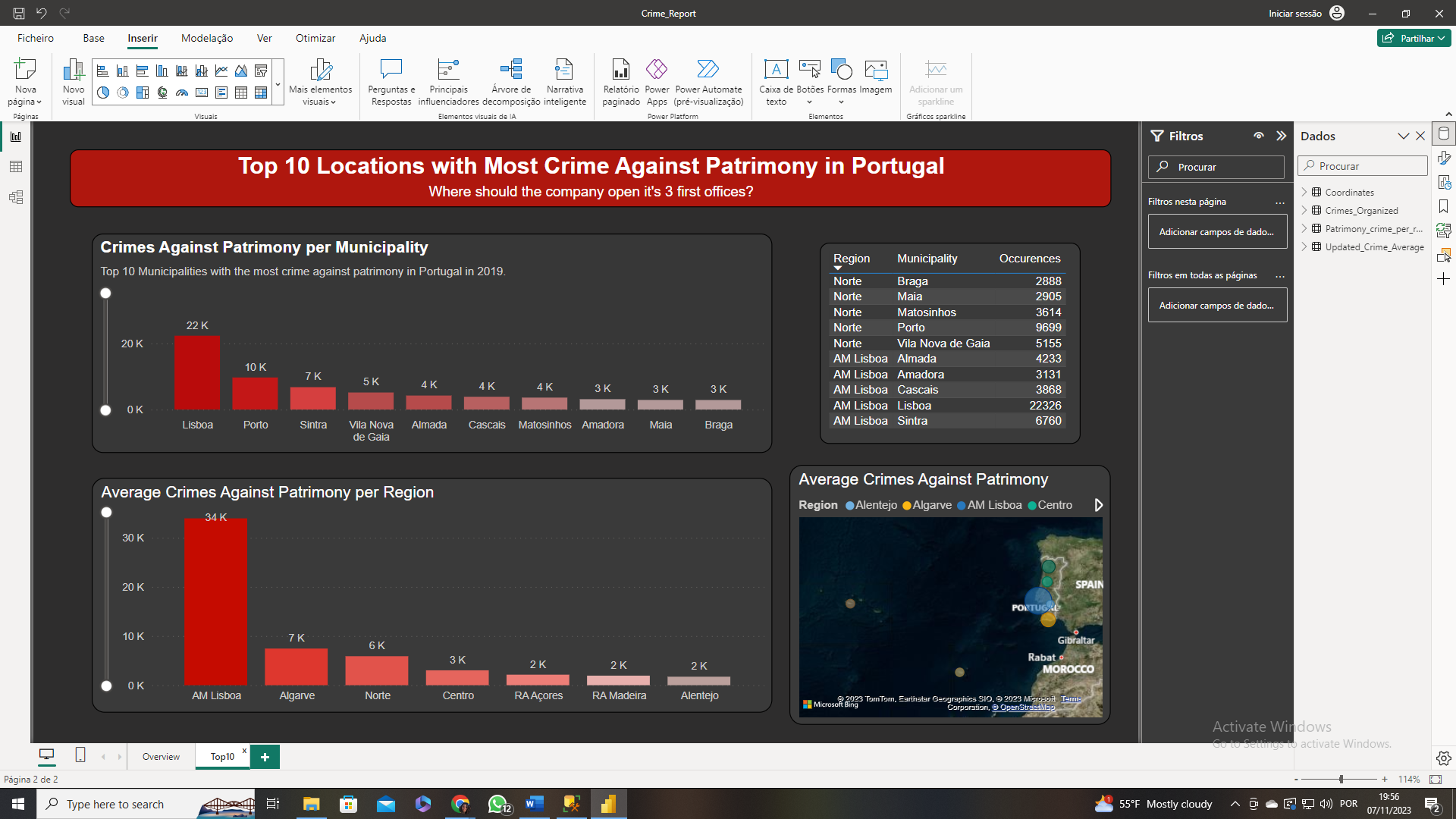


Figure 10 - Second page of dashboard crime in Portugal.

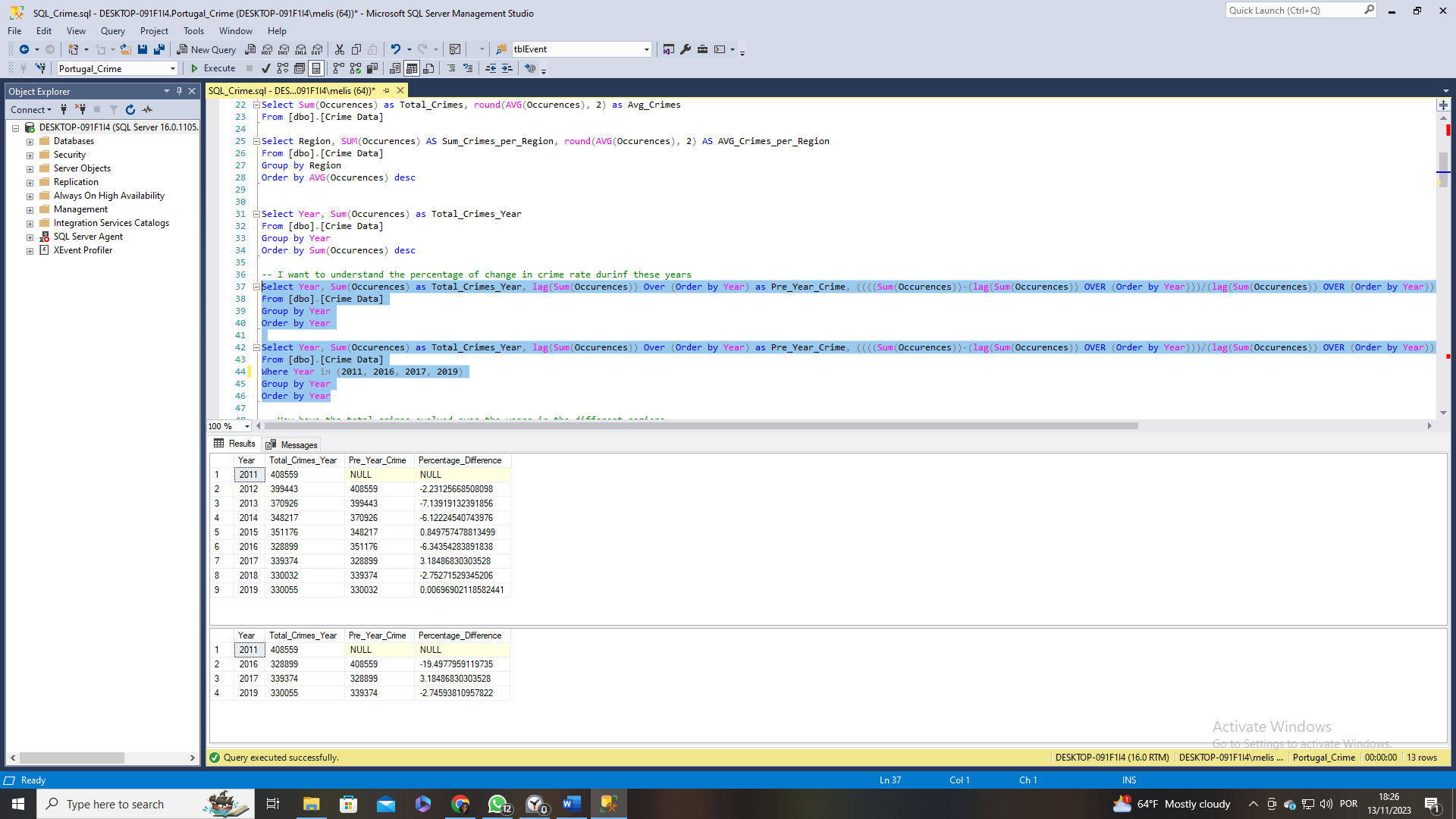
Results and Findings

Figure 11 - Total value of crimes per year and its percentage evolution.

After analysing my clean dataset, I was able to attain significant insights. In a broad perspective, the total number of reported crimes in Portugal between 2011 and 2019 amounts to 3 206 681, with an annual average of 356 297,89.

However, adding the time component to these values it was observed that the total crime in Portugal started with a high value (408 559) and began descending till it reached 328 899 in 2016 (a decrease of 19.49%), before rising again, hitting a peak in 2017 with 339 374 total crimes (a 3.18% increase). After 2017, the crime rate once again began to decline until 2019. Interestingly, after 2015, the total crime values consistently remained below the average. The image to the right shows the total amount of crimes per year and its evolution.

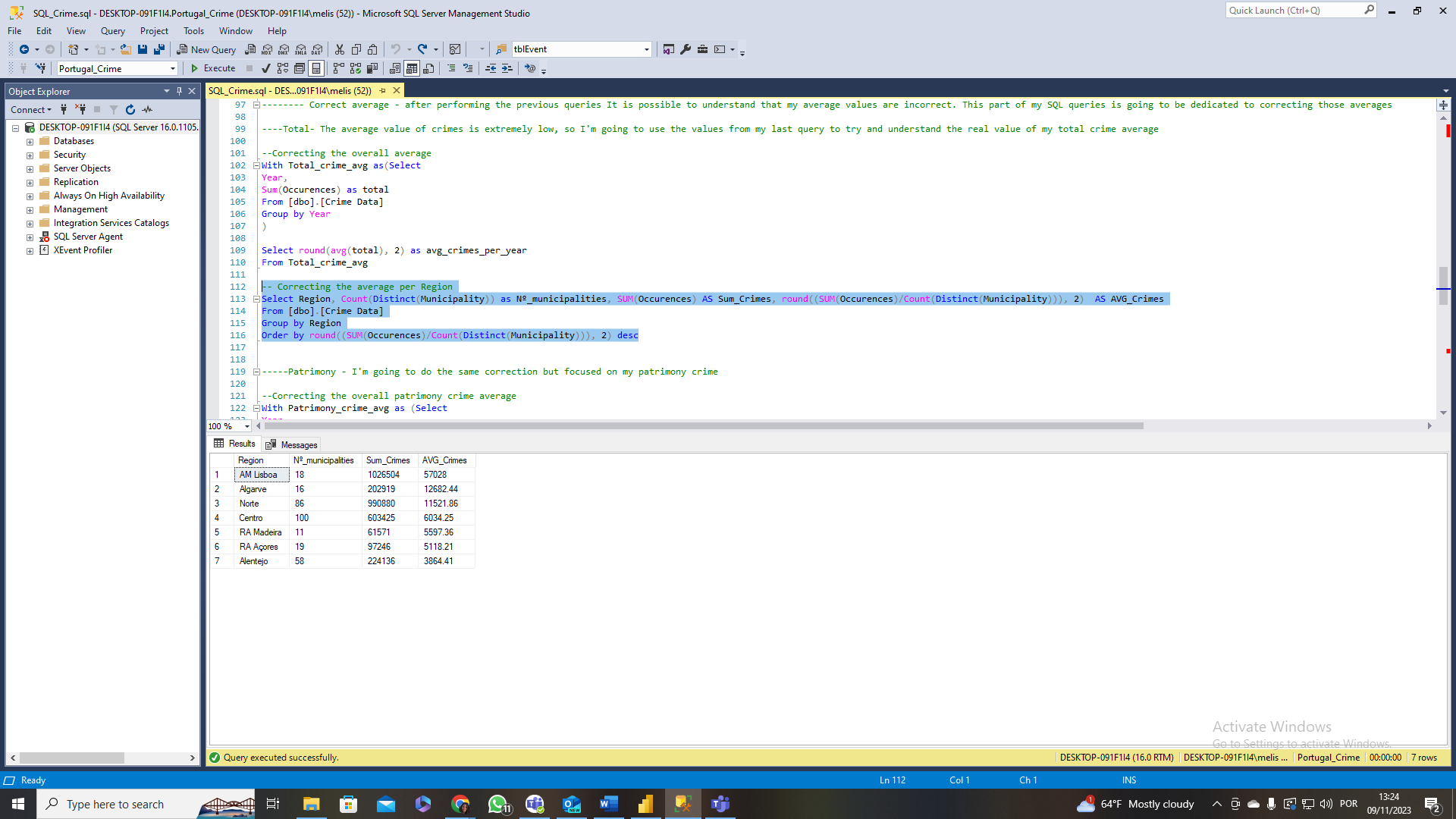
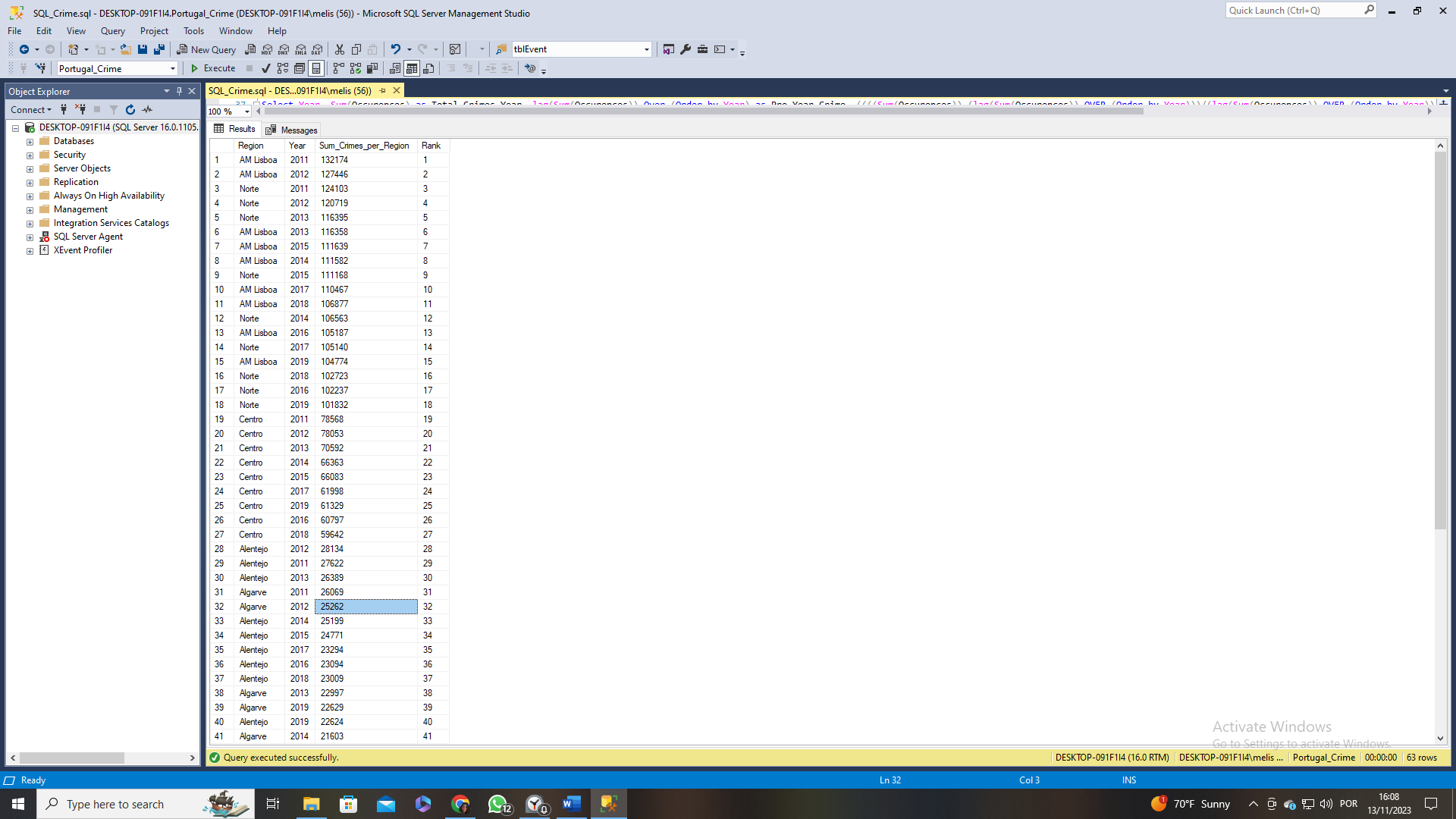
 Portugal has 7 regions, together they account for the values previously mention, but the top 4 locations with highest average crime are AM Lisboa, Algarve, Norte and Centro (represented in figure 13). However, I a step further was taken into this analysis by querying in SQL so it would return the rank of the total amount of crime per year and per region. This allows for the understanding that AM Lisboa is the region with most crime, having in 2011 registered its highest value and in 2012 its 2nd highest; afterwards, in 3rd place we have Norte with a total of 124103 in 2011; Centro comes in 19th with 78568 in 2011 as well; and Algarve only appears in 32nd place with 25262. We can see the first 20 ranks in figure 12.

Figure 12 - Rank of the sum of crimes per region and per year.

Figure 13 – Amount of crime per region.

After gaining a general view of my data, I am instigated to look into my type of crimes and see how much crime against patrimony exists in Portugal (figure 12). Looking at the data, represented in figure 14, we can see that crime against patrimony is the most committed in Portugal. Notably, over half of the total crimes fall under the category of crimes against patrimony, totalling 1 702 774, with an average of 189 197,11 crimes per year.

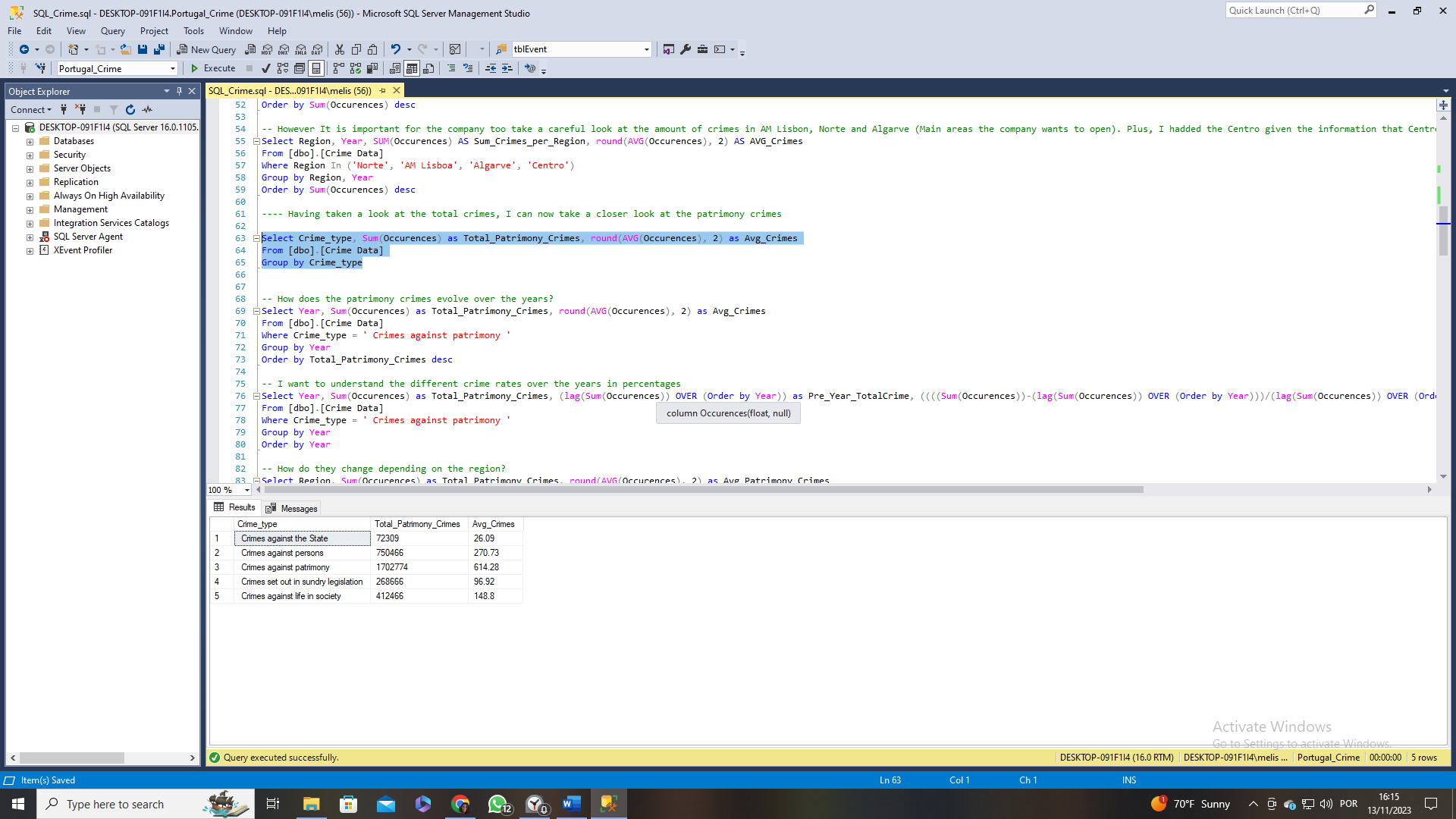


Figure 14 - Types of crime in Portugal.

However, crime against patrimony has been steadily decreasing over the years (figure 15), starting with a value of 228 704 in 2011 and finishing with a value of 169 548 in 2019, a value that translates to a 25.86% reduction. Notably, it is possible to see that crimes against patrimony from 2016 to 2019 are beneath average. There is a possibility that this event is due to the increasing presence of home security systems but other factors also have influence in these values.

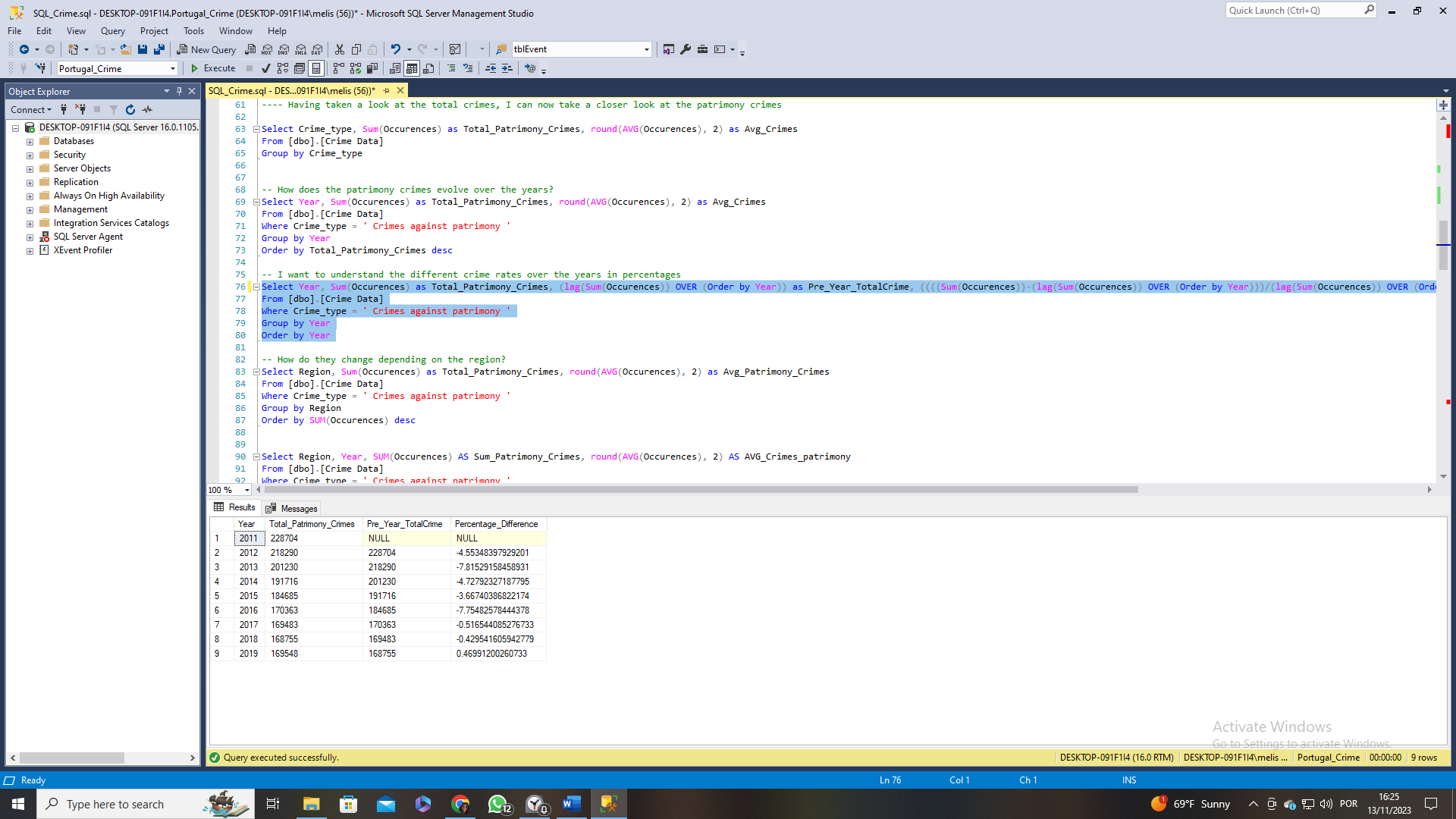


Figure 15 - Total value of crimes against patrimony between 2011 and 2019 and its percentage evolution.

Looking more specifically into which regions and municipalities have the most crime against patrimony, using SQL query I was able to obtain the information represented in figure 16 and 17. The top 4 regions with most crime against patrimony are AM Lisboa, Algarve, Norte and Centro. Nonetheless, looking at specific locations, using the municipality column, it's interesting to note that in 2019 the municipalities with the highest crime rates belong to AM Lisboa, then Norte, Algarve and then Centro.

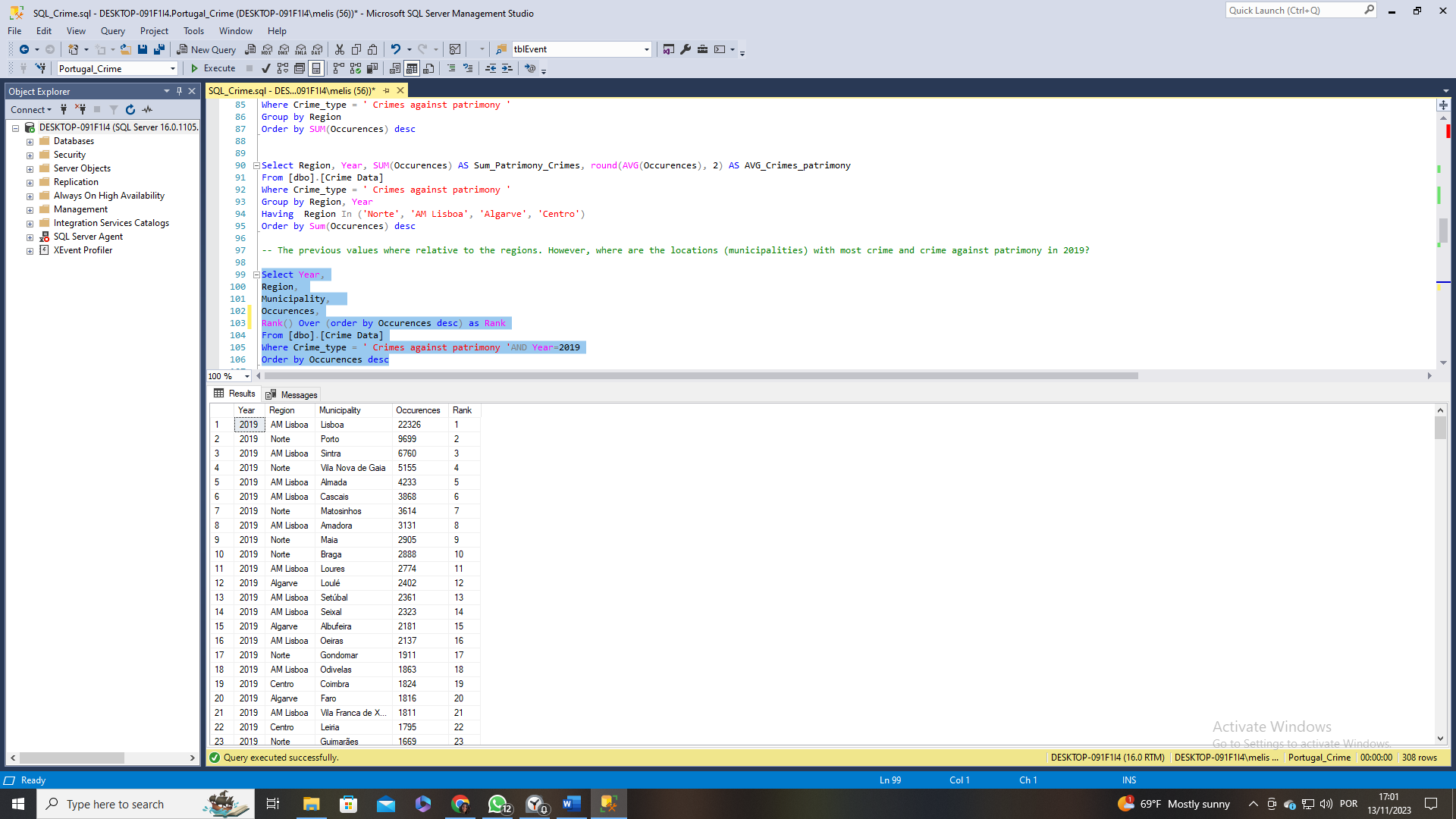
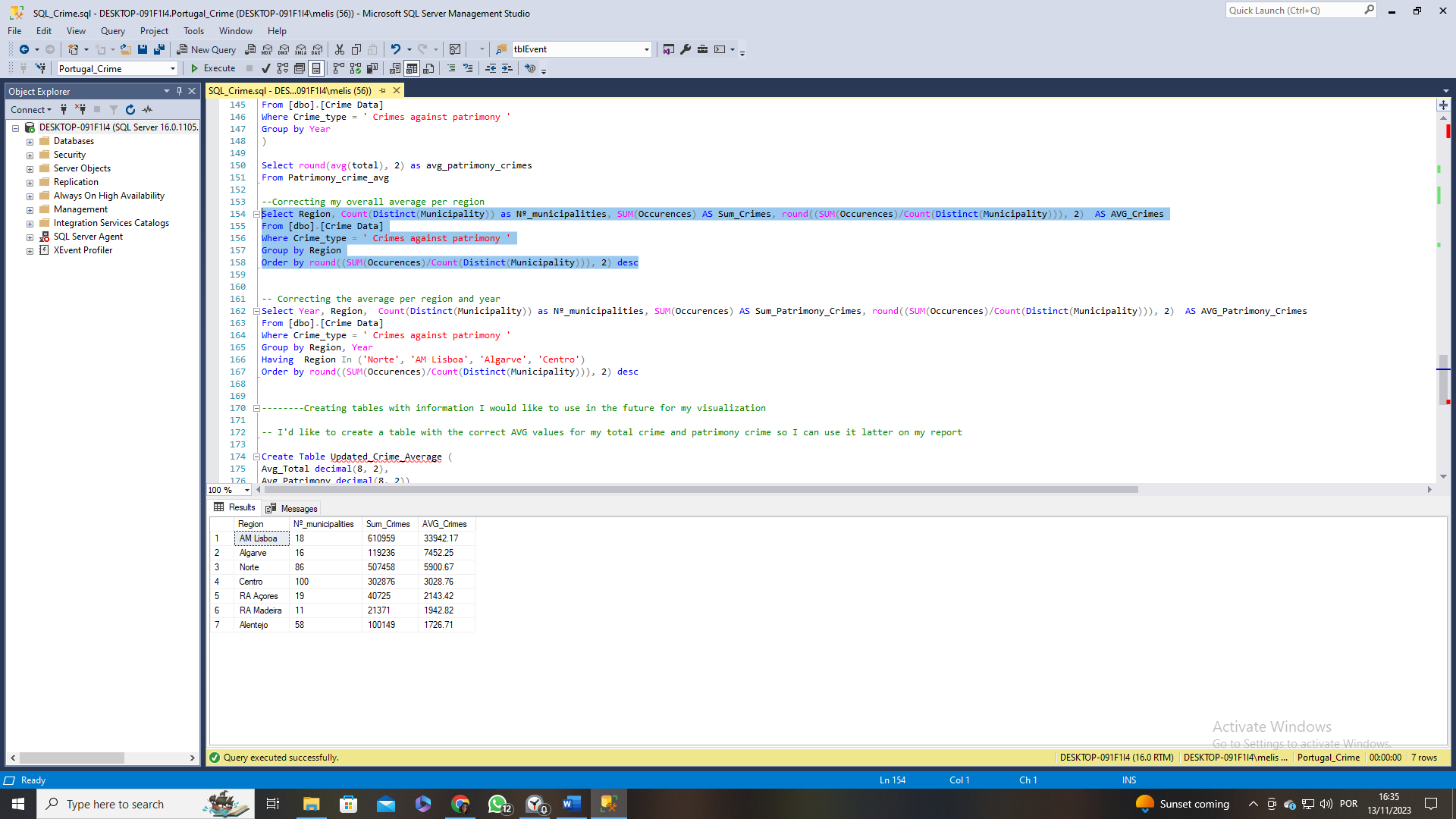


Figure 16 - Top 20 municipalities with most crime against patrimony in 2019.

Figure 17 - Values of crime against patrimony per Portuguese region.

Conclusions and Recommendations

The objective of this project was to determine the optimal regions in Portugal for a home alarm security company to establish its first offices. With a focus on three major cities—Lisbon (AM Lisboa), Porto (Norte), and Algarve—the analysis delved into understanding the general crime condition of Portugal and, more specifically, crimes against patrimony (house-related crimes).

After my analysis process, we could see that between 2011 and 2019 there were 3 206 681 crimes report with an average of 356 298 per year. Of all those crimes, more than half were crimes against patrimony with a value of 1 702 774 and this crime category had an average of 189 197 crimes per year. The overall crime had a decrease 19.5% from 2011 to 2016, still in the following years an increase in crime of 0.35% took place. However, the crime against patrimony had a significant decrease of 25.86%, in spite of the increase of total crime from 2016 to 2019.

Focusing even more on my different regions, I know that the ones with most crime and crime against patrimony are AM Lisboa, Norte, Algarve and Centro. The following figure represents the crime values of these different regions:

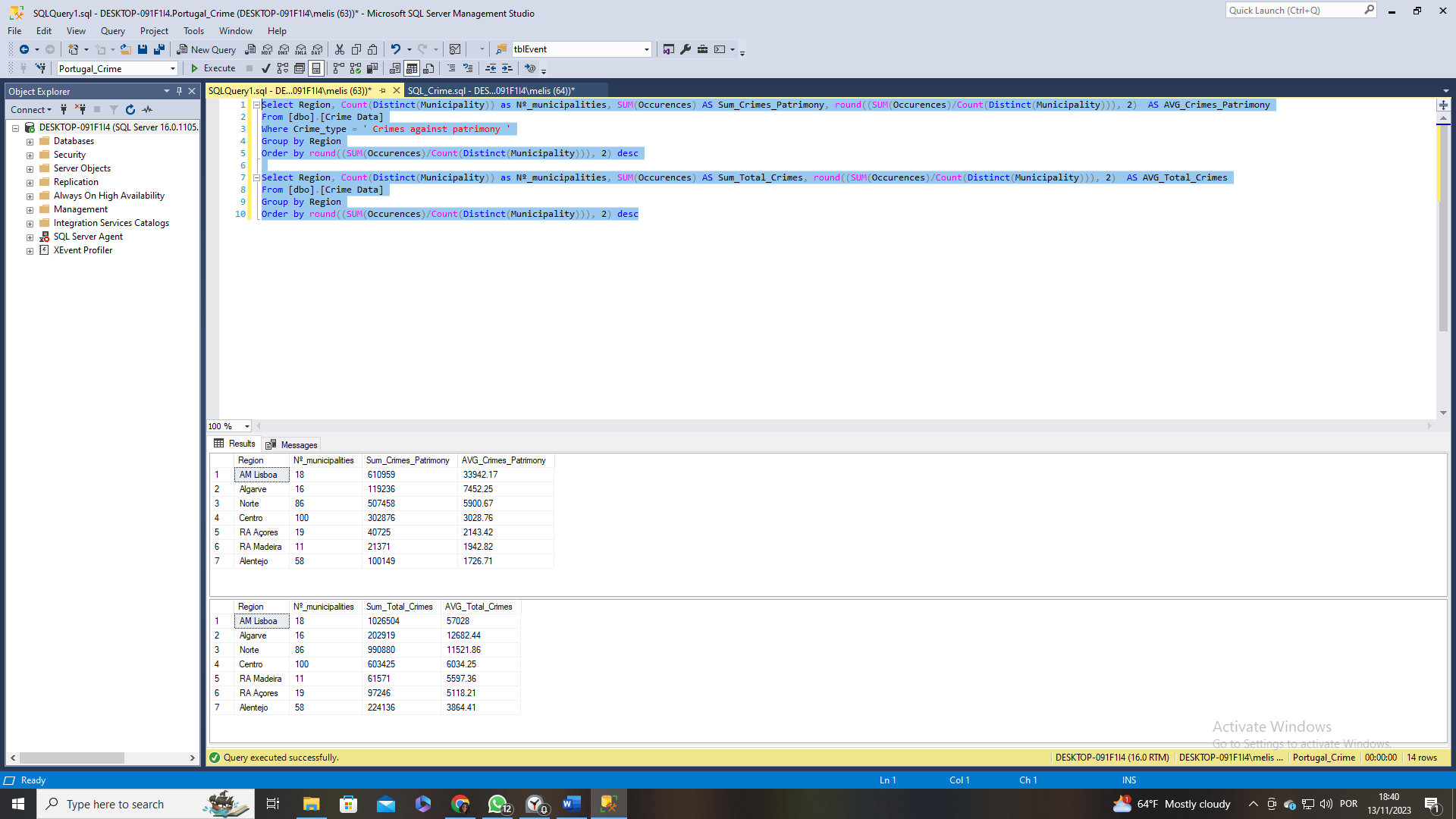


Figure - Comparison of total crime and crimes against patrimony between the top 4 regions.

If we focus on our average values, then we would rate Lisbon as the main location to open the offices and then Algarve and Norte. This would exclude the Centro region which has a significantly higher total crime value than Algarve, in spite of that we know that Algarve is significantly smaller with only 16 municipalities in comparison to Centro with 100 municipalities.

Going even further into this analysis by taking a closer look at the municipalities, it’s known that the top 10 municipalities with most crime belong to AM Lisboa and Porto. As for the Algarve region, the municipality with most crime is ranked number 12 on the list of municipalities with most crime against patrimony.

Based on the analysis, AM Lisboa, followed by Norte and Algarve, emerges as the most favourable locations for opening offices. These locations would be most beneficial because they exhibit the highest average crime values. While the analysis points towards specific regions, it's crucial to adapt recommendations based on real-time data, local policies, and the evolving nature of crime. A comprehensive understanding of the population dynamics would further refine the evaluation.