What influences the Real Estate Prices in Portugal

Hugo Martins
90727
hugo.r.f.martins@tecnico.ulisboa.pt

Sara Ferreira
93756
sara.c.ferreira@tecnico.ulisboa.pt

Pedro Leitão
90764
pedro.de.leitao@tecnico.ulisboa.pt

ABSTRACT

This report presents a project done for the Information Visualization course at Instituto Superior Técnico. We chose to explore the theme regarding the Portuguese Real Estate Market and give an overview on what influences the prices of Portuguese properties.

For the development of this project, we used programming languages such as HTML, JavaScript, CSS and the d3.js framework.

INTRODUCTION

The problem domain that we are going to address in our project is the Portuguese Real Estate Market and how it is influenced by the quality of life in a certain location. This is relevant to know what are the best locations to buy or rent a certain property and what are the characteristics of the properties that we can obtain with a certain budget.

With the constant growth of the Real Estate Market, not just nationally but internationally, it is important that we know well all the information about the Real Estate Market in our country to be aware of where and how are the best properties to obtain or to sell.

Nowadays, there are several tools that filter and show the different properties depending on their price and their location. However, our project makes it possible to filter the different property prices, either for selling or for renting, from each Portuguese district and verify how the different characteristics of the properties vary with the different price ranges and how those prices vary with the attributes, such as the security, the population density or the purchase power, of the zone where they are located. This makes it possible to answer the questions that we had defined in Checkpoint I:

- **Question 1:** How is the quality of life in a particular region?
- **Question 2:** In which regions would it be easier to pay the rent?
- **Question 3:** Within a certain budget, what type of commodities can you expect from properties in a specific location?
- **Question 4:** How does the population density of a certain location influence the search of properties in that location?
- **Question 5:** How does the safety rate in each district influence the prices of the properties?

Before doing this visualization, we needed to search for property ads on a specific website, and search on the internet about a certain location in order to know not just about the property we searched for, but also about the location where it is placed and therefore answer those questions. With our visualization, we can answer those questions in an easier way by filtering and analysing the idioms.

RELATED WORK

On the Internet there are visualizations that analyze how the average selling prices in Portugal varied over the years by region [1]. There are also papers that analyze the most important factors that determine housing prices where they say that those relate to each property's location and surrounding area as well as the dwelling itself, and explain each factor in depth [2]. Other than that, we weren't able to find any visualization or paper that could relate all the factors that affect house prices within Portugal's domain.

To design the idioms presented in this project we looked up examples done in d3.js [3] and saw which type of charts were able to answer our questions best.

THE DATA

We faced some challenges regarding our data, since not all the information that we needed was in a single dataset or website, making us have to merge and process data from various sources. We started by obtaining our original data from different websites. We got the data about the ads of the properties that were for selling or renting in Portugal from a dataset in the website "Kaggle", that we downloaded as a csv file. For the data about the locations itself, either districts, councils or parishes, we got it from datasets from the website "Instituto Nacional de Estatística", one about the criminality rates, one about the purchase power, one about the population and the other one about the area, and we downloaded all of them as csv files. We faced some challenges related to those four datasets obtained from "Instituto Nacional de Estatística", such as the semicolon that was present in both the end of every line and in the beginning of the attributes line. Another challenge that we faced related to those datasets was with the "Location" attribute, since its values started by a certain code before the name of the location. We had to remove that code, extracting the locations names from that, and then associate each parish to a certain county and to a certain district.

After obtaining our data, we processed it by using the Pandas Python library. We started by filtering out some irrelevant attributes, such as the rates of specific types of crimes from the criminal rate dataset and the male and female populations from the population dataset. Then we proceed to clean the data in our five original datasets, by discarding every item that had at least one missing value. We found the outliers by computing the Interquartile Range (IQR) formula, ending up finding outliers in all our original datasets. In the dataset about the property ads, obtained from "Kaggle", we discarded all the items that presented outliers, while in the other datasets, obtained from "Instituto Nacional de Estatística", we maintained almost every item presenting an outlier, since those outlier values weren't differing a lot from the rest in order to distort the visualization, discarding only an outlier in the rent price of a property in Évora. We then had to remove the duplicates from the property ads dataset, sorting the dataset by its "Locations" attribute and then removing every duplicate value.

Given the multitude of original datasets, we had to correlate these different sources in order to get our final data ready to be used in our InfoVis. We started this by merging our four cleaned datasets about the characteristics of every location, by comparing their "Location" attribute.

After having merged our different datasets about the locations information, we created a derived measure, called "PopulationDensity", that had the values of the population divided by the values of the area. We also created another derived measure, called "TotalCrimes", that was the values of the criminal rates multiplied by the values of the population divided by 1000. We also created the "Security", which is a rate to show the security levels of a district, which we computed by subtracting 1000 with the "CrimeRate" value and then dividing it by 10. We then removed the duplicates from our produced dataset about the locations information, by sorting the dataset by the "PopulationDensity" attribute in descending order and then keeping only an item for each "Location". With the "PopulationDensity" computed, we detected another outlier that would distort our visualization, which was the "PopulationDensity" value in Almada, which is a parish from Setúbal.

With the cleaned dataset about the property ads we created derived measures for this dataset, the "AverageSellPrice" and "AverageRentPrice", by separating the dataset in two different data sources, one with the items with "Sell" as its "AdsType" value and the other with the items with "Rent" as that value, and then computing the mean of their "Price" value.

We ended up creating five different datasets to be used in our InfoVis, one about the average characteristics of the properties in each district, other with the properties from every parish and its respective district and three datasets with the informations about every location, one for the parishes, one for the councils and the other one for the districts. We also had a geojson file of the Portuguese councils and a json file of the Portuguese districts for the creation of the Choropleth Map.

In the properties dataset there were three categories for AdsType, in which we only considered "Rent" and "Sell" because "Vacation" wasn't relevant in our domain.

The data that we thought we would get but we ended up not finding was not necessarily data but specifically a geojson for every parish of Portugal. While we actually found one, it was impossible to use since topojson wouldn't recognize it. Some other data we thought we could find with relative ease but ended up not being able to find was the ICL (index cost of living) to all the Portuguese districts, as well as the industrialization index of a region, with the number of industries being an indicator for the job availability in an area.

The compromises that we made were using different data to represent the index cost of living (rent prices, sell prices) that we collected from the dataset, and we weren't able to implement the zoom on the map for the drill down. There are some scalability issues we noticed with the size of the datasets, especially when loading the dataset to make the box plots, since there were many entries and it was slowing down the rendering. If the dataset was to be bigger we're sure there would be some more problems as well.

VISUALIZATION

Overall Description

The layout of our visualization (Figure 1.) contains five idioms:

- A **Choropleth Map**, that shows the Portuguese districts and their average values of either the sell prices, rent prices, security rates, purchase power or population density. It also displays a tooltip that shows all that information when we hover over a specific district.

- A **Histogram**, that shows the price distribution, either for selling or for renting.
- Three **Boxplots**, for showing the distribution of property components, one for the number of bedrooms, one for the number of bathrooms and the other for the property area, all from properties in a specific price range.
- A **Radar Chart**, where we can compare different districts, according to their security rates, average sell prices, average rent prices, purchase power and population density.
- A **Line Chart**, that shows the evolution of either the sell or the rent price depending on different factors such as the security rate, the purchase power or the population density.

We chose white for the background color of our visualization, with our idioms being all inside borders with an either white or light grey background. We decided that those colors match the best with the overall look and feel of our visualization. We chose Helvetica Sans-serif for the text font, due to its worldwide usage popularity, which makes the users connect easily with it. To represent our five different attributes, we used a different color for each one. For this effect we used ColorBrewer [4] to select them. The selected colors were Cinnabar (red) for the average sell price, Green Blue (blue) for the average rent price, Green Pigment (green) for the security rate, Flame (orange) for the population density and Blue Violet Crayola (violet) for the purchase power. We also used five different colors to represent the different districts selected which were Neon Green (green), Deep Cerise (pink), Rebecca Purple (purple), Chocolate Web (orange) and Burnt Orange (brown).

The Choropleth Map fills the left half of our InfoVis, with the other four idioms being in the right half. The Choropleth Map also works as a filtering tool, which makes us able to choose specific districts, which changes the whole visualization, since the other four idioms start showing the values related to those selected districts, instead of the national values. There is also an options box in the Choropleth Map, with Global Filters, that change all the visualization, showing either the sell or the rent prices in all the idioms, and Map specific filters, that just change the values in the Choropleth Map, which are represented by the color saturation.

In the Line Chart, we also have a dropdown menu where it is possible to change the values of the x axis (the y axis is changed by the Global Filters menu in the Choropleth Map).

In the Histogram we also have a brushing technique that selects a specific range of price values, changing the data represented in the Boxplots showing only the data of the properties within that budget.

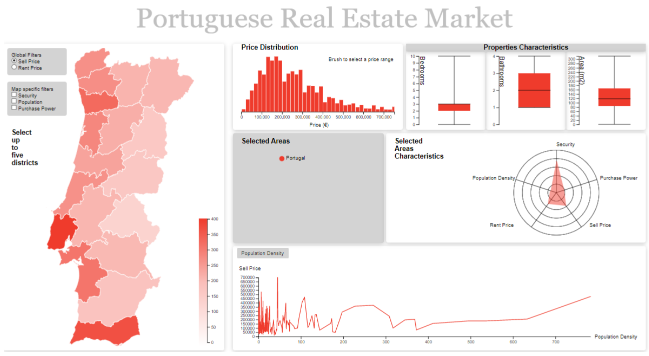


Figure 1. Overview of the layout of our visualization

Selected Areas Characteristics

To see the characteristics of Portugal or of the districts that we selected in order to find out their quality of life and compare with other districts, we implemented a Radar Chart (Figure 2).

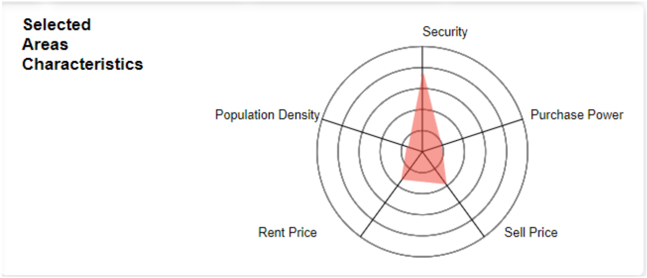


Figure 2. Radar Chart

We chose the Radar Chart for this goal since it's a good idiom for comparisons, which makes it useful to compare the different districts that we selected in the Choropleth Map. Each one of the five axes represents each one of the attributes, which are the Security Rate, the Purchase Power, the Sell Price, the Rent Price and the Population Density.

Each one of the selected districts is represented with different colors in the radar chart, which are the same colors as in the strokes in the choropleth and in the lines in the Line Chart for each selected district.

Price Distribution

In order to see the price distribution of either the Portuguese properties or the properties from the selected districts, we implemented a Histogram (Figure 3).

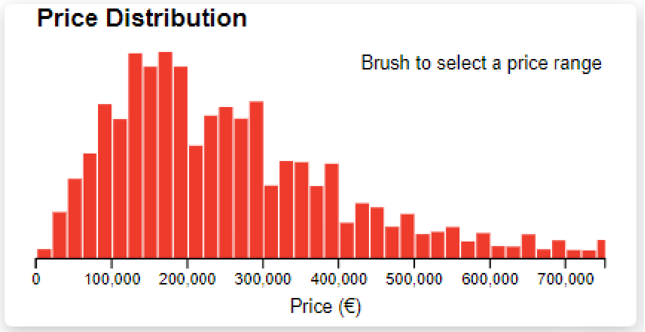


Figure 3. Histogram

We chose the Histogram for this goal since it's a good idiom to show the number of items in a specific range of values. In our visualization, we show the number of properties either from the selected districts or from all the Portuguese properties (if no district is selected) for each range of prices, which would be either sell or rent prices, depending on what was selected in the Global Filters menu.

This is useful, for example, to check what are the districts in which it would be easier to pay the house rent, for example.

The Histogram also presents a brushing technique to select a specific range of price values, showing in the Boxplots only the values of the properties with the price between the selected range. When no area has been selected the data shown switches to all price ranges (Figure 4.).

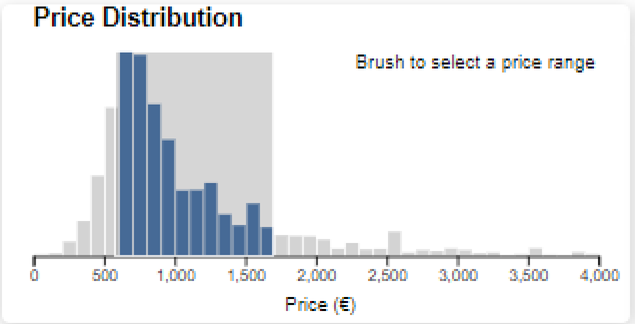


Figure 4. The Histogram's brushing tool

Properties Characteristics

In order to know what type of commodities we can expect from a property in a specific location given a certain budget, we implemented three Boxplots (Figure 5.).

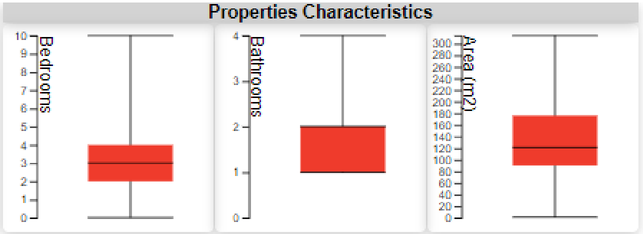


Figure 5. The three Boxplots

We chose the Boxplots for this goal, since they are a useful idiom to show the distribution of a specific attribute and therefore check the most common values of that attribute, thanks to the lower quartile, median and upper quartile values. This idiom is also good at spotting outliers (values lower than the lower whisker or higher than the higher whisker) and lesser probable values (values between the lower whisker and the lower quartile or between the higher quartile and the higher whisker).

The three different Boxplots represent the number of Bedrooms, the number of Bathrooms and Property Area.

The values present in the Boxplots vary depending on the selected districts and the selected price interval in the Histogram (showing only the values of properties with its price inside the selected price interval).

Due to the nature of the data, the first and third quartile and median may overlay so we opted not to draw lines for the first and third quartiles there, because we could already know what they are from the rectangle.

Influence of a certain attribute in the price of the properties

In order to know if and how a certain attribute, like the population density for example, influence the prices of the properties in Portugal or in certain districts, we implemented a Line Chart (Figure 6.).

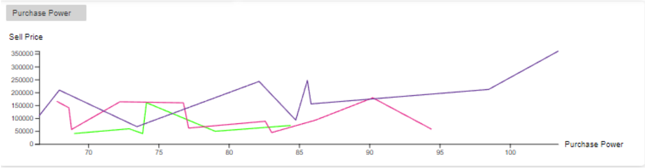


Figure 6. Line Chart

We chose the Line Chart for this goal since it's a useful idiom to show how the values of a certain attribute change depending on the other variable, thanks to the slope of the line.

Our Line Chart has a Dropdown Menu that allows us to choose the attribute represented by the x axis, between the Security Rate, the Population Density or the Purchase Power (Figure 7.).

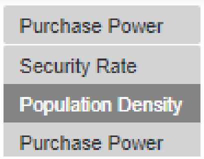


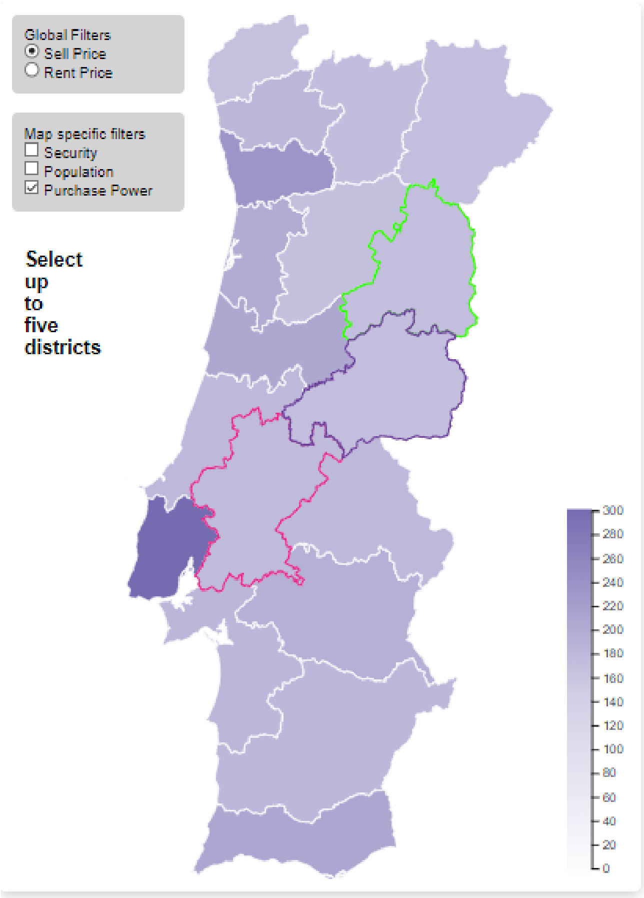
Figure 7. Line Chart drop down menu

The Line Chart shows a line representing the change of either the sell or the rent prices (given the option that we choose in the Global Filters menu), that are represented in the y axis, depending on the security rate, the population density or the purchase power (depending on the value that we choose in the Dropdown Menu), that are represented in the x axis.

If more than one district is selected in the Choropleth Map, the Line Chart shows a line for each selected district, each one of them with a different color, which is the same color as the one representing the same district in the Radar Chart, as well as the one in the outline of that district's area in the Choropleth Map.

The Portuguese Map

In order to represent all the Portuguese districts, as well as their information, and being able to select the ones that we want to analyse, we implemented a Choropleth Map (Figure 8.).



Global Filters
● Sell Price
○ Rent Price

Map specific filters
☐ Security
☐ Population
☑ Purchase Power

Select up to five districts

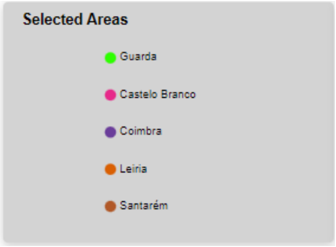
Figure 8. Choropleth Map

The Choropleth Map occupies the left half of our InfoVis and has the highest influence in the whole visualization. We chose the Choropleth Map for this goal since it's a good area-based idiom to represent geographical data.

Our Choropleth Map shows the Portuguese map with all its continental districts represented. It shows the average values of either the properties' sell price, their rent price, the security rate, the population density or the purchase power. The color saturation in every district's area represents the value of one of those attributes, that can be chosen with the Global Filters or with the Map specific filters. To see all the information about a specific district, we can hover the mouse over the district we want to analyse and that district stays red while we are hovering over that area, showing a tooltip with its information. The Choropleth Map also displays a scale where we can see what color saturation is associated with which values.

We can select specific districts to filter the values in the whole visualization by clicking on that district. When we click on it, the district stays with its original color (before the hovering) such that we can still verify its value represented by the color saturation, but with an outline that has a different color for each selected district.

In order to know what districts are selected and what color represents them, we have implemented a legend, at the left of the Radar Chart, that states that information (Figure 9.).



Selected Areas

● Guarda
● Castelo Branco
● Coimbra
● Leiria
● Santarém

Figure 9. The Selected Areas legend

Rationale

For the visual encodings, the channels that we used for the Choropleth Map were the color saturation, since it's a magnitude channel and so it should be used for ordered attributes such as the five attributes that we can choose to be represented on the map, and the position in the map to represent the districts. For the Histogram, we chose the x position to represent the price. For the Boxplots, we used the positions to represent their attributes (the bedrooms, bathrooms and area). For the Radar Chart, we used each axis position to represent each one of its five attributes, since they are all ordered attributes. For the Line Chart, we used the y position to represent either the rent or sell prices (depending on the chosen attribute), and the x position to represent either the security rate, the population density or the purchase power, since those three attributes are ordered attributes that could influence the price's value.

Based on those decisions, we designed our first sketch (Figure 10.):

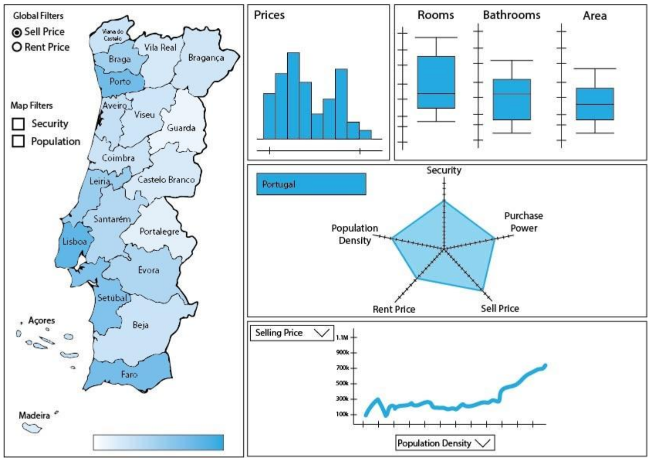


Figure 10. First Sketch

From our First Sketch (Figure 10.) to our First Prototype (Figure 11.), we removed the Dropdown Menu for choosing the y axis attribute, since that attribute could be changed with the Global Filters options.

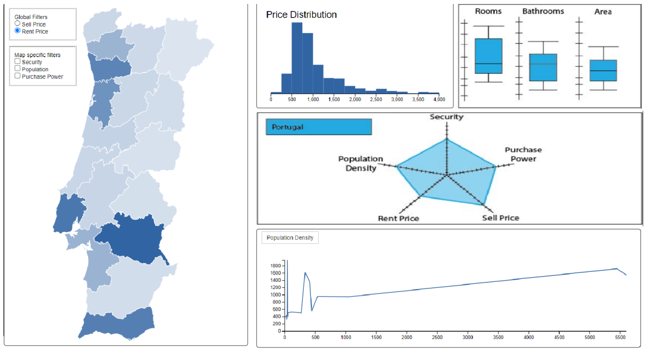


Figure 11. The layout of our First Prototype

From our First Prototype (Figure 11.) to the Final Version of our Prototype (Figure 1.), we changed various things in our visualization.

In the Choropleth Map, we learned that we couldn't see the value represented by the color saturation in a selected district because of its area becoming red and so we represented the selected district with a colored outline (which color is different for each selected district) and the previous color (before hovering) as the area color, instead of having all the area with the red color, and included a scale representing the color saturation values.

We replaced the Bar Chart from the First Prototype with an Histogram, since we learned that the Histogram is an idiom more suited to represent the count of elements in a certain range of values than the Bar Chart.

For the Line Chart and the Radar Chart we implemented the possibility of representing more than one district, with different lines with different colors for the Line Chart and different areas with different colors for the Radar Chart. In the Line Chart, we also changed the x axis, in order that its minimum is the minimum value of that attribute from the selected districts, instead of zero.

We included a legend, at the left of the Radar Chart, that states the selected districts and their respective colors.

We also chose to fixate the scales domain in the Histogram, all Boxplots and Radar Chart so that when there are districts and price ranges selected it's easier to establish comparisons between them.

Demonstrate the Potential

With our visualization we can answer the questions we have previously created.

Question 1: How is the quality of life in a particular region?
To answer this question we can select a certain district in the Choropleth Map (Figure 8.) and see the Radar Chart representing the selected district (Figure 12.).

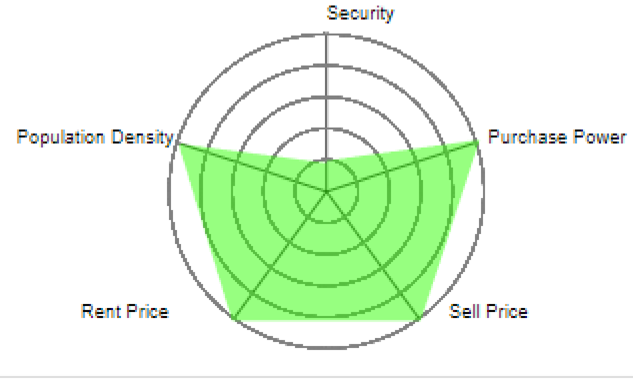


Figure 12. The Radar Chart showing Lisbon's quality of life

Question 2: In which regions would it be easier to pay the house rent?
To answer this question we need to compare the Rent Price and Purchase Power in each district. The best way to do so is by comparing these two attributes in the Radar Chart in all districts. It's easier to pay the rent when the value in the Purchase Power axis is greater than the value of the Rent Price (Figure 13). We can also choose the "Rent Price" in the Global Filters and "Purchase Power" in the Maps Filters and observe the color saturation in the Choropleth Map (Figure 8.), compare saturations and hover over each district's area to confirm its values.

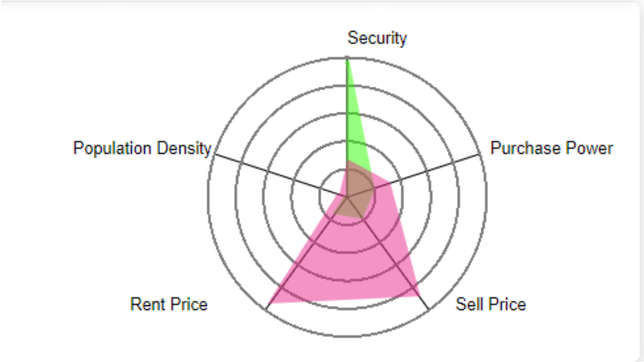


Figure 13. The Radar Chart showing the Purchase Power and Rent Price of Faro and Évora

Question 3: Within a certain budget, what type of commodities can you expect from a property in a specific location?
To answer this question we can choose the district in the Choropleth Map (Figure 8.), select a price interval for our budget (Figure 4.) and check the values in the Boxplots (Figure 14.).

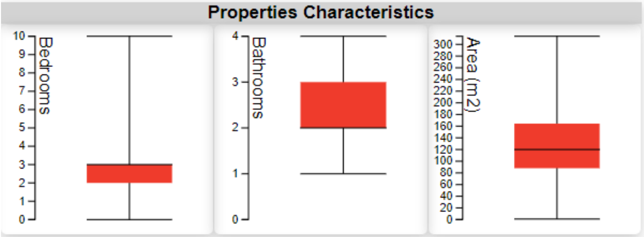


Figure 14. The Boxplots showing what type of commodities I can expect from properties in Lisbon between 300000€ and 500000€

Question 4: How does the population density of a certain location influence the search of properties in that location?
To answer this question we can choose a certain district in the Choropleth Map (Figure 8.) and select the population density in the Dropdown Menu of the Line Chart, and analyse the Line Chart (Figure 15.).

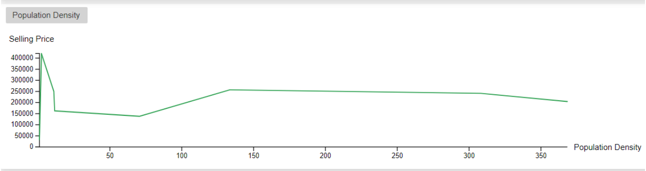


Figure 15. The influence of the population density in the search of properties in Braga

Question 5: How does the security rate in each district influence the prices of the properties?
To answer this question we can select the security rate in the Dropdown Menu of the Line Chart, and analyse the Line Chart (Figure 16.).

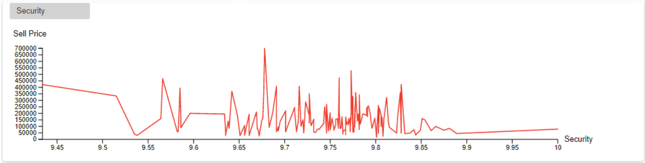


Figure 16. The influence of the security rate in the properties' price

Unexpected insights
Our visualization provided some golden unexpected insights, such as, for example, when comparing in the Histogram the sell prices in each 100,000 range we observe that there are fewer properties in the first 20,000 and 60,000 to 80,000 than in the other ranges.

Other interesting and unexpected insight that we discovered by analysing the Histogram is that the majority of the properties for renting have their rent price between 500€ and 1000€.

With the Histogram and Boxplots we observed that there are properties with no bedrooms for sale for 640,000€.

We also found another interesting and unexpected insight by analysing the Choropleth Map, which is the fact that Portalegre is the only continental district in Portugal that has an average property sell price lower than 100000€.

IMPLEMENTATION DETAILS

To implement this infoVis, we made some decisions prior to developing the visualization.

Since waiting ruins the user experience, we decided to compute as much as possible in the beginning of the interaction, and store all the scales etc. to be used upon request.

Another decision we took was that instead of iterating the districts dataset each time when using the map, we would append the information to the map json itself, with a python script beforehand. This also greatly reduces the overhead of the rendering.

We created two Javascript files, one called behavior.js and one called classes.js, which we'll explain in greater detail next. We created our idioms in the behavior.js file, creating every idiom in a different function that has an attribute called "data" for the dataset that it uses and an attribute called "update" to know if the view was changed or if it's still in the initial phase. To implement the interactions between the views, we call those functions for creating the views with the update argument set as true. Our other Javascript file is the classes.js and is used for keeping all the information that we need in other parts of code, for example if we need to change a color scale in our choropleth, those are pre-calculated to avoid overhead while using the dashboard, and kept in the Choropleth instance, to be used whenever they are needed. This happens as well with the selected filters, for example, since they are kept in the Choropleth (as well as the Visualization) instance and are accessed to do the rendering on the other idioms.

In the histogram, to trigger the brushing it is necessary to start brushing, this will call a brushed function that will get the selected coordinates, translate them to the price's scale and create the box plots.

We defined the visual attributes of our view in a CSS file, the styles.css, with different CSS classes for each div in order to keep a consistent visualization layout.

For each idiom we took inspiration from other online examples, but had to adapt them so that they supported the links between the views and changed the axis.

CONCLUSION & FUTURE WORK

With this project we learned many data processing and visualization techniques. The visualization part made us solidify the concepts learned in the classes and the data processing part made us reinforce our previous experience with python and the Pandas library. By implementing our visualization, we improved our HTML, CSS and Javascript skills and acquired a new skill with the D3.js library.

With our visualization, we were able to address all our questions and even get new insights on data that we were not expecting.

If we had one more month to do this, in order to enrich our solution, we would implement a zoom tool in the Choropleth Map that would make us able to drill down the data from districts to parishes and therefore select the data from a specific parish to be shown in the visualization. We would also find a way to fix the scalability issues we had when computing the data in the boxplots and histogram without compromising its purpose. One good idea would be to make all the sections of the histogram pre-computed and then we wouldn't have to determine them in run-time.

REFERENCES

1. Lalaine C. Clemendo. 2021. *Portugal's house price rises moderating.* https://www.globalpropertyguide.com/Europe/Portugal/Price-History
2. Dirk Wittowsky, Josje Hoekveld, Janina Welsch, and Michael Steier. 2018. *Residential housing prices: impact of housing characteristics, accessibility and neighbouring apartments – a case study of Dortmund, Germany.* https://www.tandfonline.com/doi/full/10.1080/21650020.2019.1704429
3. https://www.d3-graph-gallery.com/, last accessed in November 16, 2021
4. https://colorbrewer2.org/, last accessed in November 16, 2021