Master Degree Program in
Data Science and Advanced Analytics

Data Visualization
Final Project

Ana Carolina Ottavi, 20220541
Carolina Bezerra, 20220392
Carolina Confraria, 20220711
Daniella Camilato, 20221641

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

April, 2023

1. Introduction

Within the scope of data visualization, a project was proposed where the students' ability to use visualization concepts and techniques to transform data into a meaningful interactive visualization would be tested. The private luxury real estate in Lisbon is very attractive for foreign customers that look for high quality of life, security and are willing to invest a high amount in a luxury property. This results in significant social and economic consequences for people without the same financial conditions that have an impact on a lot of people and families, which is why this is a relevant topic. In this project, we intend to gain in-depth knowledge on real estate and understand how the price of housing for the luxury segment differs from the average segment. Subsequently, we found a dataset that met our interests, whose name is kc_house_data.csv.

The problem consists of a real estate company that required our team to elaborate an interactive dashboard of the luxury real estate in King County, Washington, between May 2014 and May 2015. UpRealeEtate, a Florida-based company that is a specialist in the luxury real estate sector, is dedicated to acquisitions and property development with the intention of selling or renting the property. After 10 years of experience, UpRealEstate decided to expand their area of operation to King County, located in the state of Washington. King County is a desirable location to invest in luxurious real estate because of its robust economy, high standard of living, low supply, and high demand. The presence of major corporations, a prime location, and the scarcity of available land all contribute to a competitive market with high property values.

The real estate market must be thoroughly evaluated to decide whether expanding would be a sensible choice. The informed assessment of the decision-makers, who are looking to increase their portfolio of luxury properties, will be built on the findings of this study. Hence, we looked at the dataset of house sale prices in King County, which covers properties sold between May 2014 and May 2015.

2. Dataset Description

The dataset is structured with 21613 items and 21 attributes. There are three categorical attributes (waterfront, view, and condition); the remaining features are quantitative. A brief description of the attributes is provided in the following table (***Table 1***):

Variable	Type	Description
id	int64	Unique ID for each home sold
date	object	Date of the home sale
price	float64	Price of each home sold
bedrooms	int64	Number of bedrooms
bathrooms	float64	Number of bathrooms, where .5 accounts for a room with a toilet but no shower
sqft_living	int64	Square footage of the apartments interior living space
sqft_lot	int64	Square footage of the land space
floors	float64	Number of floors
waterfront	int64	A dummy variable for whether the apartment was overlooking the waterfront or not
view	int64	An index from 0 to 4 of how good the view of the property was
condition	int64	An index from 1 to 5 on the condition of the apartment
grade	int64	An index from 1-13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level.
sqft_above	int64	The square footage of the interior housing space that is above ground level
sqft_basement	int64	The square footage of the interior housing space that is below ground level
yr_built	int64	The year the house was initially built
yr_renovated	int64	The year of the house's last renovation
zipcode	int64	What zip code area the house is in
lat	float64	Latitude
long	float64	Longitude
sqft_living15	int64	The square footage of interior housing living space for the nearest 15 neighbors
sqft_lot15	int64	The square footage of the land lots of the nearest 15 neighbors

3. Methods

At the top, the dashboard shows a line chart comparing “price” and “yr_builty”. Uprealeste plans to acquire recently constructed properties, which, compared to older ones, offer lower maintenance and repair costs. Therefore, considering newly constructed properties, we can conclude through the visualization that there was a time period between 2000 and 2010 where the average price was between $500k and $600k. The previously mentioned filters are the target, which both allow a lower repair and maintenance investment from UpRealEstate and a lower property price. The dataframe is first divided into groups according to the year of construction, with the mean price of each group determined. By year of construction, the generated dataframe is arranged in ascending order. The newly generated dataframe is then utilized with Plotly Express to produce a line chart. The year of construction is shown on the x-axis, while the average cost of houses built in that year is shown on the y-axis.

The next visualization explores the relationship between “Location” and “Price”. In order to evaluate where there were more properties that would be of interest to UpRealEstate and which areas should be targeted, we decided to develop a Scatter MapBox from which it was possible to observe that most of the luxury properties are located along the waterfront. We started by creating a price range filter for a map that shows where properties are located according to their corresponding price. The filter lets the Dash user choose a range of prices between $1 million and $10 million by employing the dcc.RangeSlider component from the Dash framework. This maximum ceiling is defined in order to filter properties with an excessive “price” value - these aren’t the properties UpRealEstate is looking to acquire. The data shown on the scatter plot is then filtered based on the chosen price range. In the next step, we designed a scatter plot on a mapbox to illustrate the association between the geographic location of the houses considered in this dataset for King County and their corresponding prices and square footage of the apartment's interior living space. The map's size, boundaries, center, and zoom level were customized manually. The latitude and longitude coordinates of each property are represented as points on the scatter plot, which displays the locations of the properties on a Mapbox map with an “open-street-map” style. In order to further highlight the properties with higher monetary values, we also altered the size of the bubbles, which grew in proportion to price.

The third visualization, right next to the mapbox, shows a scatter plot that compares the existence of Bedrooms/Bathrooms per price and square footage of the apartments interior living space. An essential factor in evaluating a property's pricing and investment potential is the number of bedrooms and bathrooms, as it's assumed that the more bedrooms a property has, the more expensive it will be. The default bar chart, however, shows that in our sample this is not the case; rather, the majority of homes, including the priciest ones, have fewer than five bedrooms. As for the feature bathrooms, the pricier a property is, the more bathrooms it has. Using Plotly Express, the code generates a scatter plot, with the size of each point denoting the amount of sqft_living, to show the relationship between the number of bathrooms (and bedrooms) and the price of homes. The final scatter plot is shown inside an HTML div element that has a title and a dropdown menu, so users may choose to plot either bedrooms or bathrooms on the x-axis. The dropdown's selected value is kept in a dcc. Dropdown object.

At the bottom, we explored the relationship between “waterfront” and “view” according to “price”. The default graph illustrates the relationship between the attributes "waterfront" and "price." In the dropdown, the use of "x-dropdown" allows the user to choose between the variables "waterfront" and "view," both of which are relevant since they can have a significant impact on the overall value and appeal of a property to those looking to buy or rent a property. For the callback, the use of "update_graph" updates the x-axis of the plot based on the feature specified by the user. This implies that if "waterfront" is chosen, the x-axis will have the categories "Waterfront" and "Non-Waterfront", whereas "view" will show categories for the five quality levels of "view".

For the last visualization, there is a bar plot that reports the comparison between the mean “price” and the corresponding “condition”. This visualization allows us to assess the sum of prices per condition using a bar chart. UpRealEstate is not interested in properties in precarious conditions, considering that such would require higher investment costs for the necessary renovations. Every month that the properties sit vacant is a month of losses, and these losses will affect the overall profit margin of UpRealEstate. To our surprise, the majority of the properties are in an intermediate level of condition, requiring minimal investment costs, which meets the business plan of the company. The bar chart displays the average cost of properties according to their condition. A lambda function is used to change the "condition" column, giving each level of the initial numeric scale a category value. "Low," "Medium," and "High" are the ensuing categories, respectively. The newly developed "condition_" column is used to group the "price" column, and the mean value of each group is calculated to create a new dataframe. The resulting dataframe is fed into the plotly express bar chart, whose mean "price" column is represented by the y-axis and whose x-axis corresponds to the column labeled "condition_".

4. Results and Discussion

Encoding:

This visualization, known as a "Temporal Line Chart" or a "Time Series Plot", is used to show the changes in a variable over time and connect the data points with a line. In the line chart, the mark is a line that represents each data point of the mean price of properties built in a specific year. The line chart consists of one quantitative value attribute and one ordered key attribute expressed in a mix of vertical and horizontal positions, where the channel will control the appearance of a mark in the chart. The encoded express value attribute has an aligned vertical position and point marks separated into horizontal regions by the key attribute. The chart consists of one quantitative value attribute and one ordered key attribute.

A scatter plot on a Mapbox map is an example of a "Geo-Positional Encoding" technique (Munzner, 2015). For the spatial scatterplot visualization, latitude and longitude were used to place each data point on the map as "channels". The size and color of the marker were also used as channels to encode additional characteristics. The spatial scatterplot is color encoded, with blue indicating a higher "price", white indicating a middle "price" and red indicating a lower "price" value. To represent different types of data, markers can be customized with a variety of colors, sizes, and symbols.

A scatter plot is a sort of visualization that falls under the category of "Positional Encoding" techniques (Munzner, 2015). The scatter plot includes quantitative attributes, in which each dot symbolizes a single property. The color, once again, is the main visual encoding, portraying different values for “price” in accordance to the range of colors previously explored in the spatial scatterplot.

A bar chart is a sort of visualization that falls under the category of "Length Encoding" techniques (Munzner, 2015). The three bar chart visualizations developed at the bottom of the dashboard are based on a quantitative attribute on the y-axis and a qualitative attribute on the y-axis.

Filtering:

Users can modify the criteria and study the data in various ways by using interactive filtering tools like sliders or dropdown menus. This way, the user has more control over the visualizations and can concentrate on particular features of the data. Sliders are frequently used to select or emphasize data in the context of data visualization based on a numerical value. In the spatial scatterplot, we used a slider to display the "price" of the different properties.

A dropdown is a feature of a user interface that enables users to choose an item from a menu of choices. Dropdown menus are frequently used in the context of data visualization to filter or group data based on a categorical variable. Dropdowns were used to group properties by different qualitative attributes. In the scatter plot, data was filtered so that the default scatter plot had the feature "bathrooms" in the x-axis and the second scatter plot in the same visualization had the feature "bedrooms" in the x-axis. As for the Bar Chart in the bottom left, the default plot had the feature "waterfront" in the x-axis and the other plot had the feature "view" in the x-axis.

Conclusion:

In the scope of data visualization, we have accomplished the design of a dashboard where the users are capable of understanding the concepts of the business while interacting with visualizations developed in the Plotly and Dash frameworks. It is possible to have an overview of the luxury real estate in King County by analyzing the graphic explorations with the price of properties and attributes such as bedrooms, bathrooms, view, waterfront, and condition. We also implemented filters, which allow users to interact with the visualizations. There are some limitations in our project, such as not including enough attributes to explore more patterns in this dataset. In the future, it is desired to apply machine learning to predict the price of properties and develop a dynamic and stylized dashboard.

The corresponding GitHub repository can be accessed through this link:
https://github.com/AnaOttavi/Data_Visualization_Project

References

https://www.thestreet.com/personal-finance/how-much-luxury-property-costs-largest-us-cities#gid=ci0278369750002668&pid=9-washington-dc-sh

https://www.seattlemet.com/news-and-city-life/2016/03/2015-real-estate-loses-its-mind

https://www.idealista.pt/news/imobiliario/habitacao/2022/12/20/55316-sentimos-cada-vez-mais-o-interesse-na-compra-de-casas-de-luxo

Munzner, Tamara. Visualization Analysis & Design. Boca Raton, Crc Press, Taylor & Francis Group, 2015.

OpenAI. (2021). ChatGPT [Computer software]. https://openai.com/

“Add Border around Div HTML Component in Dash (Python).” Stack Overflow, stackoverflow.com/questions/55358280/add-border-around-div-html-component-in-dash-python. Accessed 6 Apr. 2023.

“How to Create a Sidebar in Dash Python.” Stack Overflow, stackoverflow.com/questions/75027100/how-to-create-a-sidebar-in-dash-python. Accessed 5 Apr. 2023.

Gonçalves, Frederico. ““Sentimos Cada Vez Mais O Interesse Na Compra de Casas de Luxo” — Idealista/News.” Www.idealista.pt, 20 Dec. 2022, www.idealista.pt/news/imobiliario/habitacao/2022/12/20/55316-sentimos-cada-vez-mais-o-interesse-na-compra-de-casas-de-luxo. Accessed 1 Apr. 2023.

Luvsandorj, Zolzaya. “DASH101 — Part 2: Prettify Dash Dashboard with CSS and Python.” Medium, 16 Mar. 2022, towardsdatascience.com/dash101-part-2-prettify-dash-dashboard-with-css-and-python-3866c069a3b6. Accessed 5 Apr. 2023.