7CCSMSDV – Simulation and Data Visualization of House Prices

Sayaka Bhandari
K20012633

ACADEMIC HONESTY AND INTEGRITY

I agree to abide by the expectations as to my conduct, as described in the academic honesty and integrity statement.

1 Part 1. Analytics

1.1 Exploratory Research Questions And Reasoning

The section below contains two additional exploratory research questions beyond Q1. These questions will require access to the Housing pricing dataset provided by the Office for National Statistics (ONS)[1] through Kings College London (KCL) with the addition of the House Price Statistics dataset from Data.gov [2] to fill in any data missing from ONS dataset. The rationale behind these research question choices is explained along with the questions. The chosen track is Track 1, and the primary research question is as follows:

Q1. Analyze the development of house prices over time. Are there any detectable trends?
For this question, the focus will be on the dataset provided by the Office for National Statistics (ONS), which focuses on the UK population from 1992 - 2023. The sub-question for this section could be how regional variations in house price trends over time compare to the overall UK trends. This allows us to understand the housing market trends and make informed decisions.
Q2. Analyze the average advance amount for other dwellings compared to new dwellings and how this trend evolves. Are there any noticeable trends?
This question builds on the initial question but focuses on a deeper analysis of house prices over time. In particular, comparing the average advance amount for other dwellings with new dwellings can reveal trends in mortgage lending patterns and investment preferences in different properties. This allows us to get an insight into the financial dynamic of property purchases.
Q3. Explore housing affordability by analyzing the relationship between house prices and the average recorded income of first-time buyers and former owner-occupiers across different regions in the UK. Are there any discernible patterns?
This question thoroughly examines the evolution of house price dynamics by examining various buyer segments' affordability and financial characteristics across different regions. They build upon each other to delve into different aspects of the housing market, offering insight into trends, disparities, and potential factors influencing housing prices.

1.2 Assessing the Appropriateness of Data Types and Dataset

1.2.1 Dataset for Q1

Q1 uses time-series data to see the development of housing prices over the period. Therefore, the table data provided by ONS on the Keats page will be used. We will be looking at data from 1993 to 2023 across all regions of the UK, including the UK. To simplify, we will only look at the 'all dwelling price' attribute, which is quantitative and discreet for all regions from 1993 to 2023. To simplify further, there needs to be further manipulation of the prices provided as they are divided into Q1. Q2, Q3 and Q4. Since all the prices are integers, an aggregation needs to be made to calculate the average house price for each year will be calculated, resulting in intervals rather than discrete data. This allows us to see the changes and development of housing prices over different regions and in the UK from 1993 to 2023.

1.2.2 Dataset for Q2

Q2 expands on the first question by providing a deeper analysis of house prices. This will also require time-series data as we investigate trends over the period. We will use the same ONS dataset on the Keats page for this question. However, this time, from the category of dwelling types, we will be focusing on 'new dwelling' and 'other dwelling' and their attribute of 'average advance' for each category. Again, for simplicity, we will calculate the average for each year. Therefore, we will be working with quantitative, discrete data between 1993- 2023 across the different regions in the UK.

1.2.3 Dataset for Q3

Q3 explores more categories from the ONS dataset, such as 'types of buyers' and how their 'average recorded income' attribute relates to the housing prices (measured using average dwelling price attribute for each category of buyers) across different regions. In addition to the ONS dataset, we will use House Price Statistics by Data.gov [2]. The House Price Statistics dataset provides categorical and quantitative data regarding average prices for first-time and former buyers across various regions in the UK until January 2024. This provides us with more recent data than the one provided by ONS, which ends in quarter 4 of 2023. Therefore, combining both datasets will allow us to see and do a more recent analysis to answer Q3. We will be working with quantitative and categorical data to answer this question.

1.3 Correlation between Dataset

All the datasets contain a common field of years and regions, which could be used to correlate all the datasets. However, the dataset provided by ONS finishes in December 2023, whereas the dataset provided by Data Gov has more recent data for most regions across the UK, which could cause discrepancies. Furthermore, even though both datasets contain a field of years, the data provided for the years are different. ONS provides yearly data into a quarterly subset, which is a discrete data type. In contrast, the dataset provided by data.gov has a continuous monthly period of data for each attribute, which can cause discrepancies between the correlation of data.
However, since we use the ONS dataset to answer all our exploratory research questions, we see the correlation between regions and different attributes being analyzed and investigated in each question. Furthermore, to answer Q2 and Q3, we use a directly proportional dataset, as both look at the dwelling category types to investigate any noticeable trends over time across different regions. Therefore, they can be seen following similar patterns even though they look at different attributes for that category.

2 Part 2. Design and Discussion

Most of the design below allows users not to see certain aspects they might not be as interested in. For example, the design used for

Q1, Q2, and Q3 allows users to view information for the regions they are interested in, allowing them to discard any irrelevant information. Furthermore, all designs provide users with some level of interactivity to provide further information and offer more transparency. The designs were created with a few common concepts of visual channels, such as color and size, that allow for effective data visualization and ensure it's simple and easy to understand to enhance user experiences. These concepts are:

Simplicity: All designs were created to prioritize visual simplicity so users are not overwhelmed with excessive information. Therefore, one of the key commandments of "Less is more" in data visualization is followed throughout the design planning [3]. This ensures users of all levels can easily understand the information provided, so only a 2-D diagram is employed to answer all exploratory questions. Furthermore, all designs provide features for users to interact with and access more information if required.
Colour: Colour channel supports pre-attentive features and causes elements to "pop-out" from display to grab users' attention. The color was selected using a variety of hues as well as research on the color scheme for colorblindness to verify its accessibility [4]
Size: As all our datasets contain quantitative data, size is an effective channel to use to visualize quantitative data to help users perceive differences or relationships in the data. This allows for the data to be read more easily.

2.1 Design of visualization to answer Question 1

The First question explores a visualization approach suitable for time-series data to display the information throughout 1993-2023. Therefore, designs such as scatter plots, line charts, heat maps, bar charts, and area charts were explored. The next step was to investigate the design approach suitable for categorical data to display different regions in the UK. Therefore, line charts, area charts, bar charts, pie charts, and chord diagrams were looked into, and it was found that line charts and bar charts are the most suitable designs to display one or more groups' numeric values change over time. Therefore, the initial idea was to use a stacked area graph, combining line and bar charts [5]. As shown in Figure 1 in the appendix. The visualization shows the trend of average house prices over 1993-2023 years by regions in the UK. It displays the average price for all dwellings on the y-axis and the year timeline on the x-axis. Initially, only the average house prices for the UK over time are displayed. However, the user can click on the UK area to see more zoomed-in data by country. Therefore, users are displayed with data for England, Northern Ireland, Wales, and Scotland, which allows users to compare the development of house prices between different countries in the UK. Users can get further insight by clicking on England to see the average house prices for all its regions within England. Then, the user can reset the graph, taking them to the first stage of data visualization, where only UK information is provided. However, during the design phase, the problem with this approach was discovered: it was too crowded with too much information, plus too many colors were used, which could be a hindrance to colorblind people, and lastly, it does not meet our concept of simplicity. Therefore, the line graph was used as a final version, as it was minimalist, easy to read and understand, and user-friendly for all people. The user can filter out the area they are interested in and see the development of house prices from 1993 to 2023 for that region. Also, when the user changes to a different region, only the line changes, and the scale remains the same so users can compare the differences in the region. Furthermore, users can hoover around the line to get an exact value, so they are not misled. This is shown in Figure 2.

2.2 Design of visualization to answer Question 2

The second question also requires us to explore a visualization approach suitable for time-series data to display the information throughout 2010-2023 and for categorical data to display different regions in the UK. However, this time, there was another set of categories, 'types of dwelling,' and we had to display data for its attribute of 'new dwelling' and 'other dwelling.' Therefore, to keep things simple and interactive for users, two different visualizations were used to display the two categories of information. Our main aim in question 2 was to investigate the evolution of the average advance trend for other dwellings compared to new dwellings. Therefore, a group bar chart was used to display the average advance prices for both categories of dwelling. The x-axis represents the years, the y-axis represents the average advance prices of dwelling, and two bars of different colors represent the new dwelling and other dwelling attributes. The bars follow the Gestalt Rules of proximity to show that two bars closed together represent data for one year, the other two bars close together represent the next year, and so on. Furthermore, the choropleth map is used to show the regions in the UK. Various hues were used to show different regions while making it accessible to colorblind people [4]. When users hover their mouse around the chart, the clickable region gets darker shaded, and the non-clickable area gets lighter shade; also, the region for that section slightly pops out, providing users with a snapshot of information, such as the average advance price for Dwellings in 2023, for that region. The saturation of the hues is used to enhance user experience. Furthermore, the split of information is used for simplicity, as providing all the information about the UK and all its regions in one chart could overwhelm users. The user can click on the area in the map, and the bar chart gets updated with the data for that region. Therefore, users focus on the trends in that specific region. The breakdown of information into two different visualization approaches is for simplicity, allowing users easier readability. The diagram is shown in Figure 3 in the appendix.

2.3 Design of visualization to answer Question 3

Last question requires us to explore a visualization approach suitable for time-series data and multiple categories of data, such as UK regions and buyers. Furthermore, we are comparing two quantitative data sets, average recorded income, and average dwelling price, for all the categories from 2010 to 2023. Therefore, the first idea was to create a dual-axis line chart, which can be used to compare the two related variables over the same period. In this case, the 'average recorded income' would be on one axis, and the 'average dwelling price' would be on the other. The users have a slider to interact with and see the different timeline information. Furthermore, for regions, the idea is to use geographical dimension, where the user can click on the area to update the line chart, similar to question 2. However, this approach did not seem user-friendly, as it required users to interact too much to see the information. Therefore, a bubble chart was used, as shown in Figure 4 in the appendix. A bubble chart allows for comparison between three variables to compare the average recorded income and average dwelling price for the period. Users can click to choose which region they wish to view the information for. The year is represented in the x-axis, the average recorded income is defined in the y-axis, and the bubble area or size corresponds to average dwelling price values. For simplicity purposes and to not overwhelm users, the number of points to plot each year is limited; since the buyer category has fewer attributes than the region category, the buyer category is used for the points, so only first and former buyer points are displayed, which are color-coded to show the distinction. Therefore, the region is provided separately for the user to click and interact with to see the information for the relevant region.

3 IMPLEMENTATION

The data supplied by ONS had to be manually aggregated to provide the user with a user-friendly visual approach. The initial raw data contained many irrelevant and unnecessary data, plus the partition of the average "all dwelling price" was split into quarters, which were aggregated to find the average for the year. This will provide users with more accessible snapshots and data readability in a line graph. The cleaned-up data could be found in the github, which was used to read the data in our code. Therefore, from the table dataset provided, table 11 was used. However, since the quarter 1 data was missing in 1992, that year was omitted, and only data from 1993 to 2023 was used. Data provided for "Great Britain" and combined data for "England and Wales" were also omitted as they served no value in answering Exploratory Research Question 1. Therefore, the average for each year was calculated using the "AVERAGE=()" function provided by Excel for every other country and region we were looking for. You can find the detailed process of cleaning data in this excel. Once The average value for "all dwellings was obtained for each country and regions from 1993 to 2023, the file was converted as CSV file and committed to

To successfully implement the line graph and to provide the users with interactivity, multiple existing d3 visualization implementations were used to build this graph successfully. One of the key concepts to creating this graph was to aim for simplicity and accessibility so the users are not overwhelmed with too much information and are also suitable for color-blind people. The idea to filter out line graphs by country and region was taken from D3.js Graph Gallery [6] to use the dropdown to filter areas. Furthermore, the hoovering of the mouse online in the line graph was used to show the exact value in the line graph used from the D3.js Graph Gallery and the cursor shows the exact value[7].

REFERENCES

[1] ONS. *House price data: quarterly tables*, 2024. 1

[2] Data Gov. *House Price Statistics*, 2024. 1

[3] Forbes. *Five Key Commandments Of Data Visualization*, 2023. 2

[4] Ivan Kilin. *The best charts for color blind viewers*. Datylon, 2022. 2

[5] Mike Yi. *A complete guide to area charts*. Atlassian, 2024. 2

[6] d3 graph gallery. *Line plot with dropdown to filter group in d3.js*, 2018. 3

[7] d3 graph gallery. *Line chart with cursor showing exact value*, 2018. 3

Appendix

Q1 Stack Area Graph Design

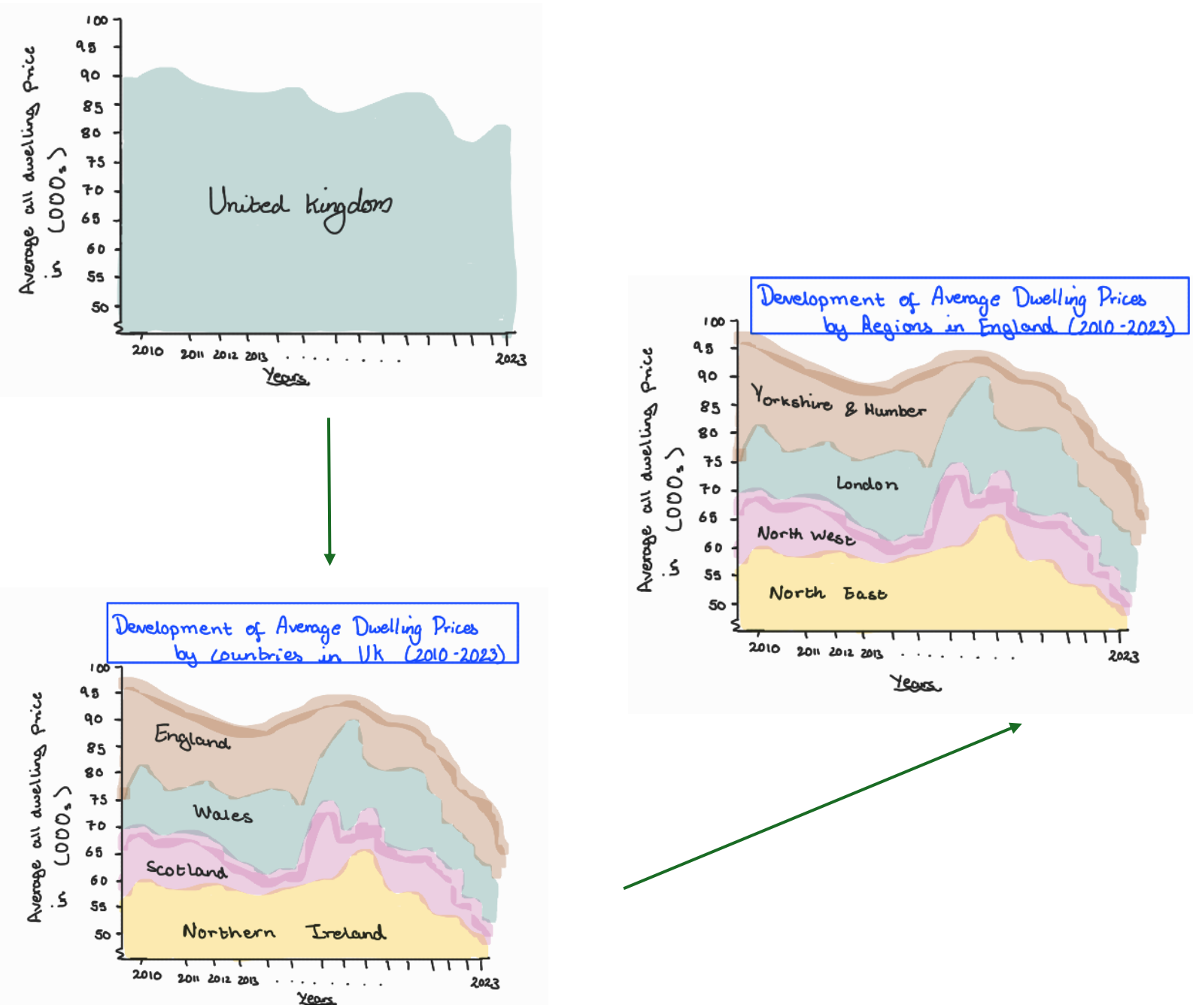

Figure 1 Stack Area Graph to Display Development of Average House Prices

Q1 Line Graph Design

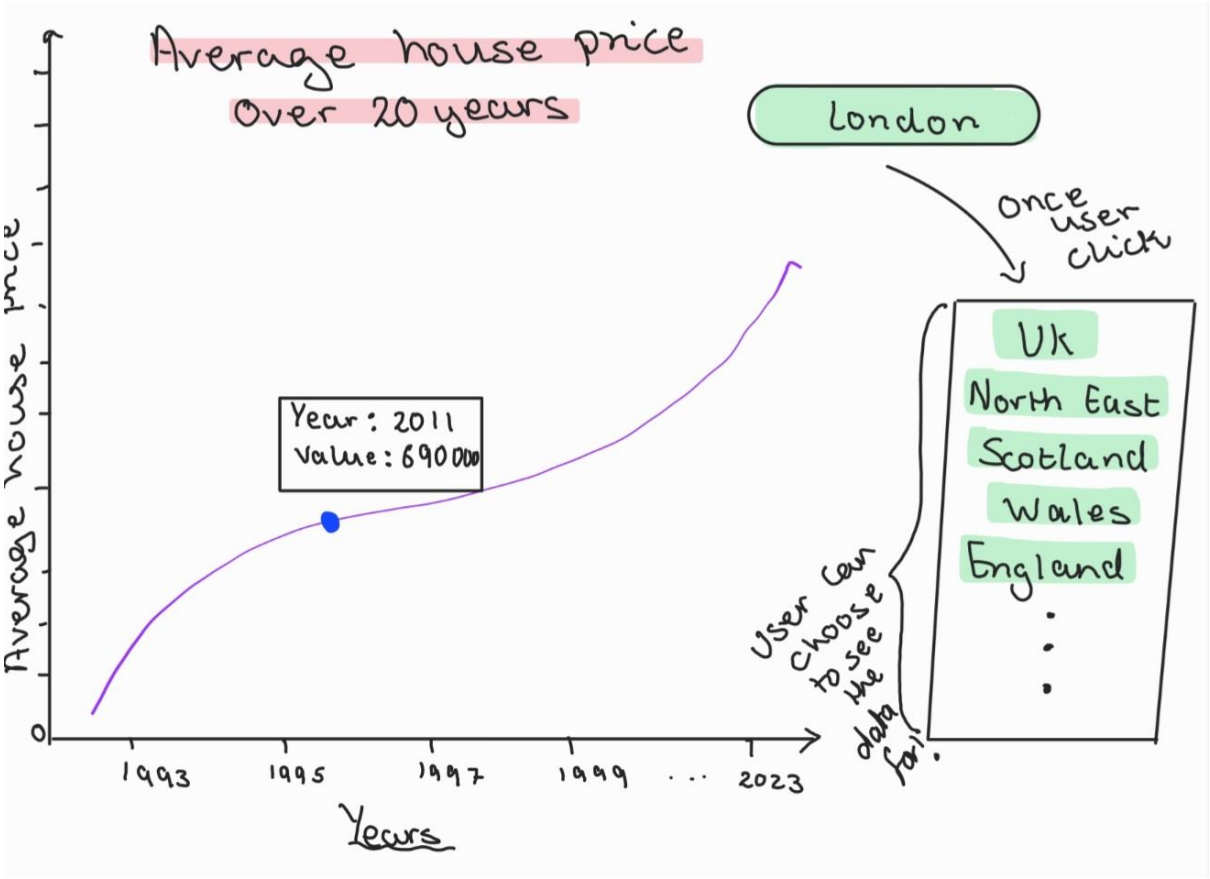

Figure 2 Line Graph Design to Display Development of Average House Price Across Region

Q2 Grouped Bar Chart and Choropleth Map to answer Q2.

This map was used from UKmap360.

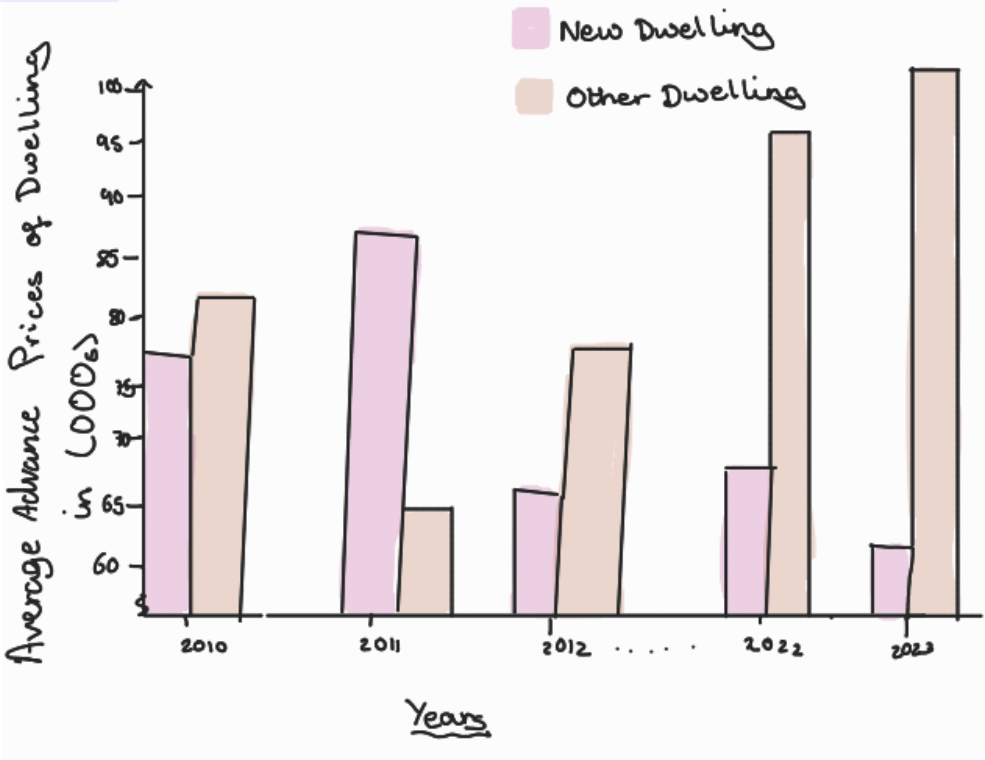

Figure 3 Grouped Bar Chart and Choropleth Map to show Average Advance Price for New and Other Dwelling

Q2 BUBBLE GRAPH TO ANSWER Q3.

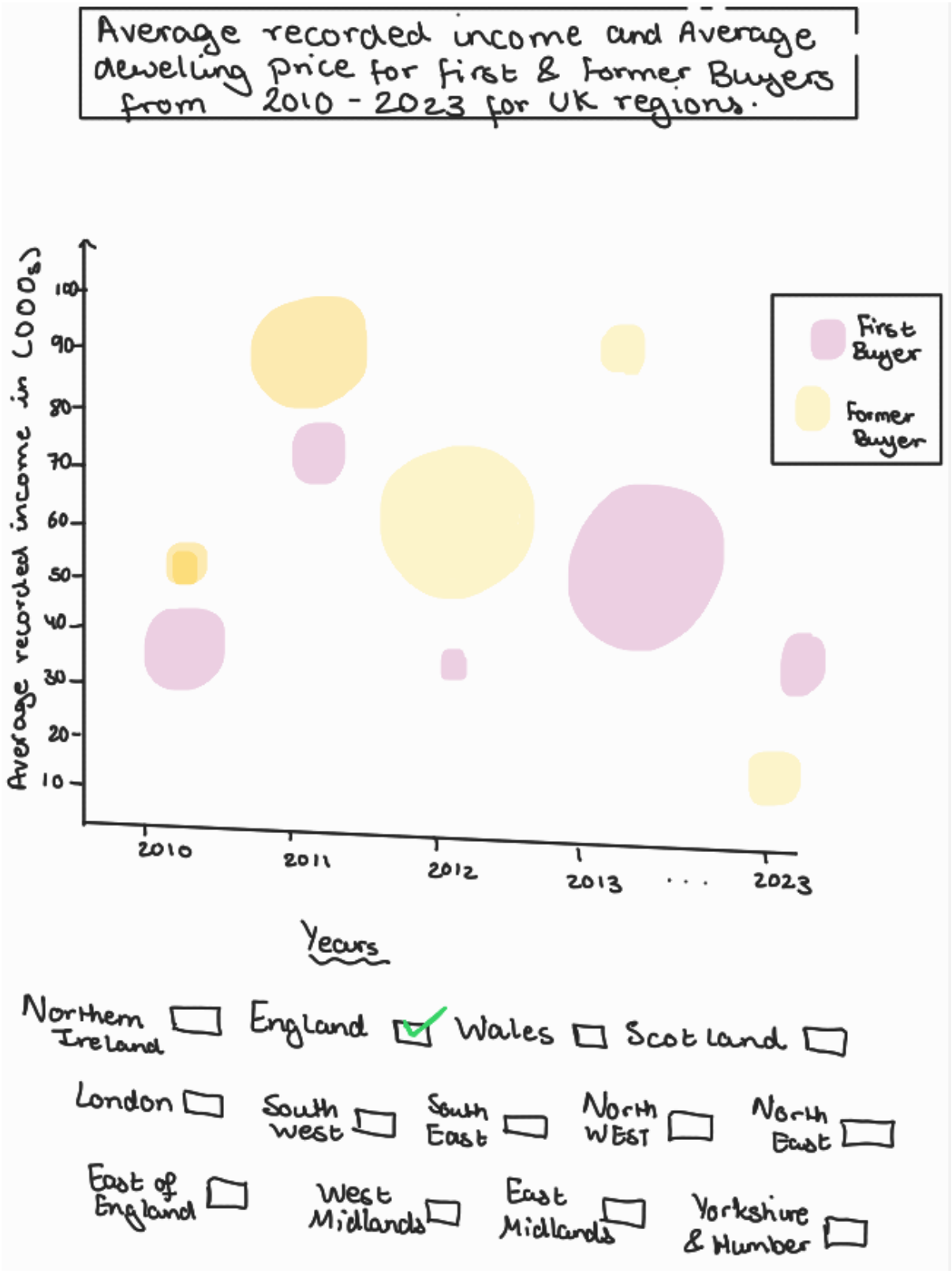

Figure 4 Bubble Graph to Display Income and Dwelling Price for First and Former Buyers

Other Details and References

- **Implementation of the Figure 2 can be accessed in codepen:**
  - https://codepen.io/Sayaka-the-reactor/pen/oNOMvBK
- **THE DATA USED IN THE CODE CAN BE ACCESSED IN GITHUB:**
  - https://github.com/Saya32/Average-House-Price-Graph/blob/main/updated_data.csv
- **THE PROCESSING OF DATA FOR THE FINAL CODE CAN BE SEEN IN THIS EXCEL:**
  - https://docs.google.com/spreadsheets/d/1WM0ryv8PvtIshVqcnooOwuA3yuYH7mC1/edit?usp=drive_link&ouid=104958396338485689163&rtpof=true&sd=true
- **FOR THE FINAL IMPLEMENTATION OF THE CODE BELOW TEMPLATE AND RESOURCES WERE USED:**
  - https://d3-graph-gallery.com/graph/line_cursor.html
  - https://d3-graph-gallery.com/graph/line_filter.html
- **DATA USED FOR IMPLEMENTATION WERE PROVIDED BY ONS:**
  - https://www.ons.gov.uk/economy/inflationandpriceindices/datasets/housepriceindexmonthlyquarterlytables1to19

Acknowledgment of AI

Grammarly was used to check the spelling and grammar of the report.