**Objectives**

As a single Singaporean with a normal financial background, one of my main priorities for the future includes housing. For most Singaporeans, public housing remains the most affordable option, and from there, we can further choose between built-to-order (BTO) and resale flats, among other things. However, I have always wondered if the various dimensions of a HDB flat do indeed influence its eventual pricing. With that in mind, I went to data.gov.sg to look for a relevant dataset, and found one that could aid me in answering the following:

* What is the general trend of resale prices over the years?
* Do the size of the house and flat type influence the price?
* Does the size of the houses change over the years?
* Does location affect resale price?
* Do older houses or newer houses end up on the resale market more often?
* Does the remaining lease affect the price?

While the project was personally motivated, some of the professional use cases for it could include helping young couples with financial planning for their futures; specifically regarding how much it could potentially cost for public housing when they are ready to make that purchase.

**Related Work**

I had only looked around official sources for relevant datasets, and have yet to search for projects that might explore similar questions. Hence, there is no academic or professional inspiration for this project apart from curiosity.

**Data**

The [dataset](https://data.gov.sg/dataset/resale-flat-prices) was downloaded in CSV form from the Singapore Government’s official data resource portal, data.gov.sg. Within the folder, there exist five CSV files containing HDB resale information from five periods: 1990-1999, 2000-2012, 2012-2014, 2015-2016, and 2017-present time. Just to note that the 2012 data in the 2000-2012 dataset only consists of information from January to February of that year.

Given the large chunk of data available for this project, I decided to only use data spanning the years 2012-present because I felt that the inflation would have been gradual, and that 10 years would be a more than sufficient range to illustrate my findings.

From initial inspection, there did not seem to be null data or any requirement to clean the data further, from the first few rows, but after taking a closer look, I realized that not all of the CSV files had a column for remaining lease, and for the ones that did, the data was stored in inconsistent formats. Furthermore, seeing as I intended to use data for the last 10 years, I would have to join at least three dataframes together, and ensure that the data types remained consistent in the combined dataframe. Additionally, I was made aware that I would have to implement some form of encoding for at least one of the columns; namely the flat type. Data processing was carried out using Python in a Jupyter Notebook before illustrating the visualizations on Tableau.

**Exploratory Data Analysis**

Once the various datasets have been loaded, I used the info, describe and correlation functions on the respective dataframes to get an initial understanding of the data. From this, I confirmed that the datasets had varying numbers of columns. After also ensuring that the lease commencement year values were integers, I could then calculate the remaining lease using lambda functions. The correlation and describe functions did not provide me with much insights as the raw and un-cleaned data only had two feature columns worth of numerical values.

In the 2017-present CSV file, the remaining lease years were stored in “X years Y months” format, so I decided to strip the values and just keep the number of years as the remaining lease, before converting those values to integers as well. To add remaining lease to the dataframes that were lacking that column, I used the pandas map function to pass in a lambda function that would do the calculation for me. With the lease for any HDB flat being 99 years, I added that value to the respective lease years, and subtracted the year of record from that figure to derive the remaining lease.

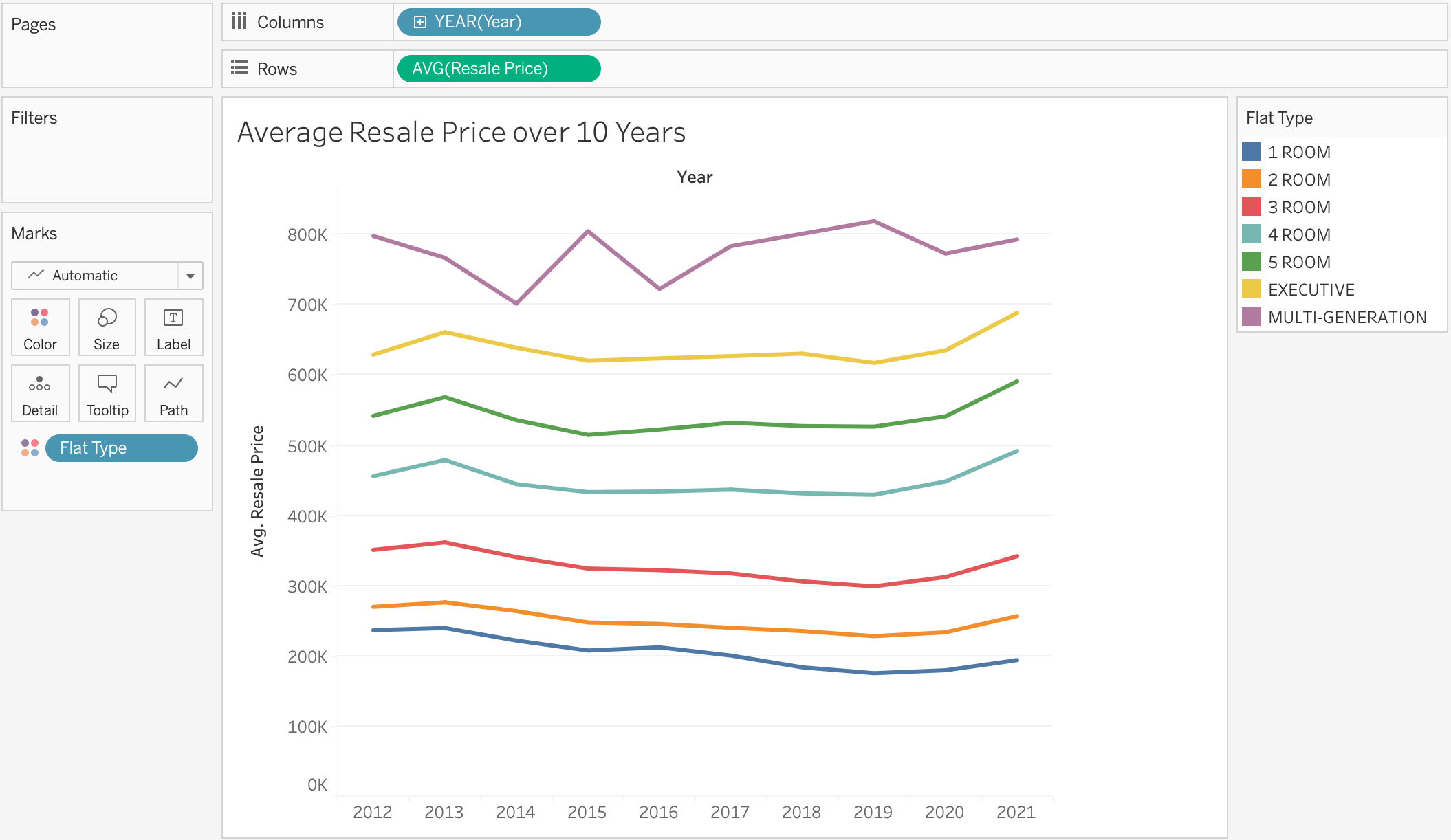
Once I was satisfied with the consistency of the datasets, and that the number of columns was the same throughout the datasets I required, I was then ready to concatenate the tables into a consolidated dataframe. After extracting the 2012 dates from the 2000-2012 dataset and storing the chosen values into a new dataframe, I created a new 2012 dataframe by concatenating it with the wider 2012 dataframe. Finally, I performed one more concatenation on the remaining dataframes to produce the master dataframe.

From the concatenated dataframe, I proceeded to encode the six flat types, otherwise known as the number of rooms in each flat, from numbers 1-6, which should also reflect the expected size and price tier of each flat type based on initial research and word of mouth. Once the new column was created, I considered encoding the flat model as well but after hypothesizing that the number of rooms should also reflect the type of flat model, I decided to leave it as is. Encoding was also considered for the level range of each flat, but after reviewing the data, there did not seem to be a proper way to encode the values given the overlap in most of the levels throughout the dataset.

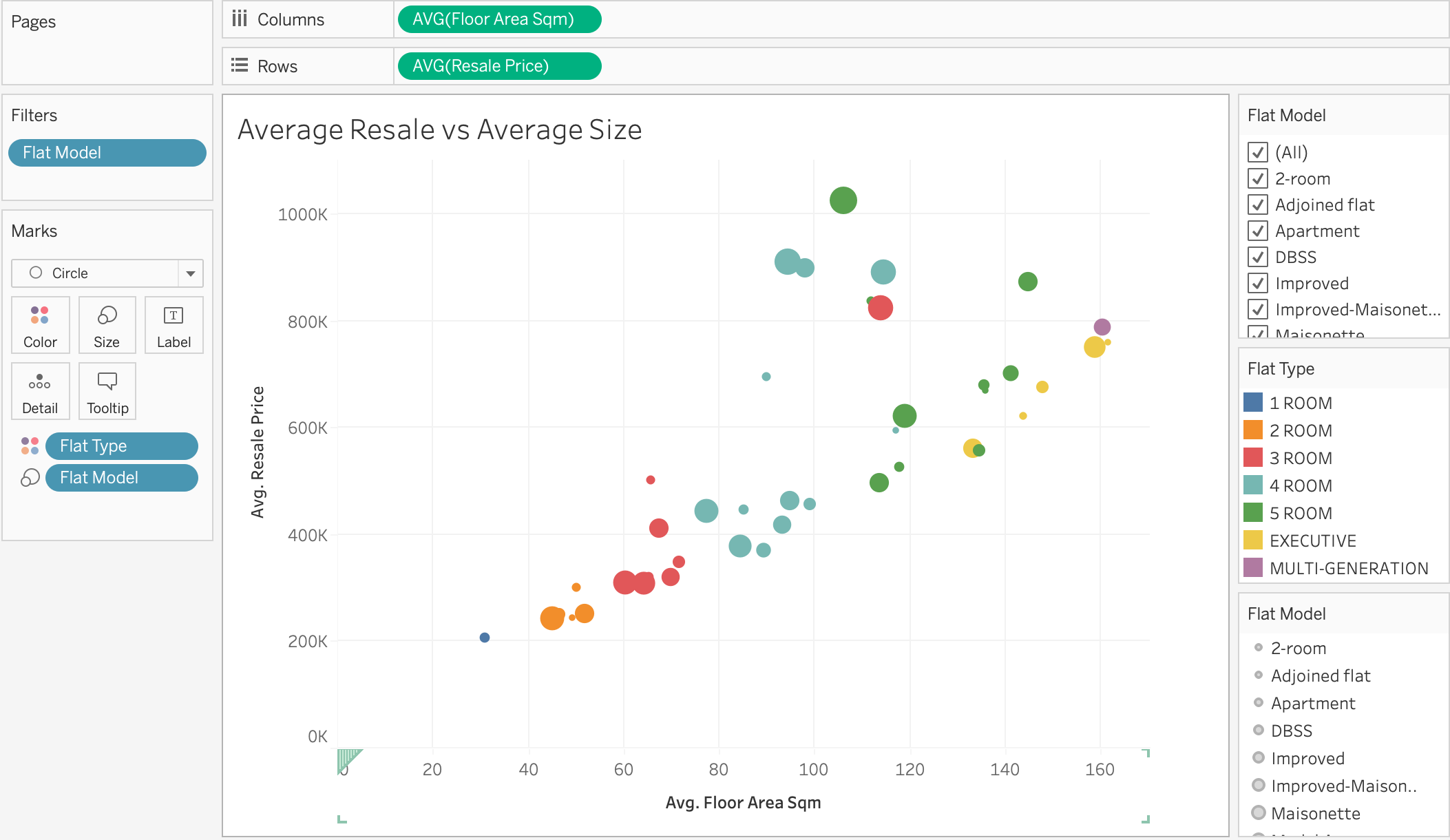
Once I was satisfied with the level of processing implemented, I checked for correlation one more time before the visualization phase. Surprisingly, there did not seem to be particularly strong correlations between the most obvious feature columns; the size of each house and the number of rooms, while for other features such as remaining lease, the correlation with resale price was not very strong as. It would appear that the correlations and patterns would show themselves better once the features have been grouped together against the resale price in the illustrations using Tableau. Using Seaborn and Matplotlib, I also worked on basic visualizations such as a heatmap to make the correlation more apparent, using bars to plot resale price against flat type and flat model in two separate charts, and using a boxplot to find if there were outlier values in the resale price.

**Visualization**

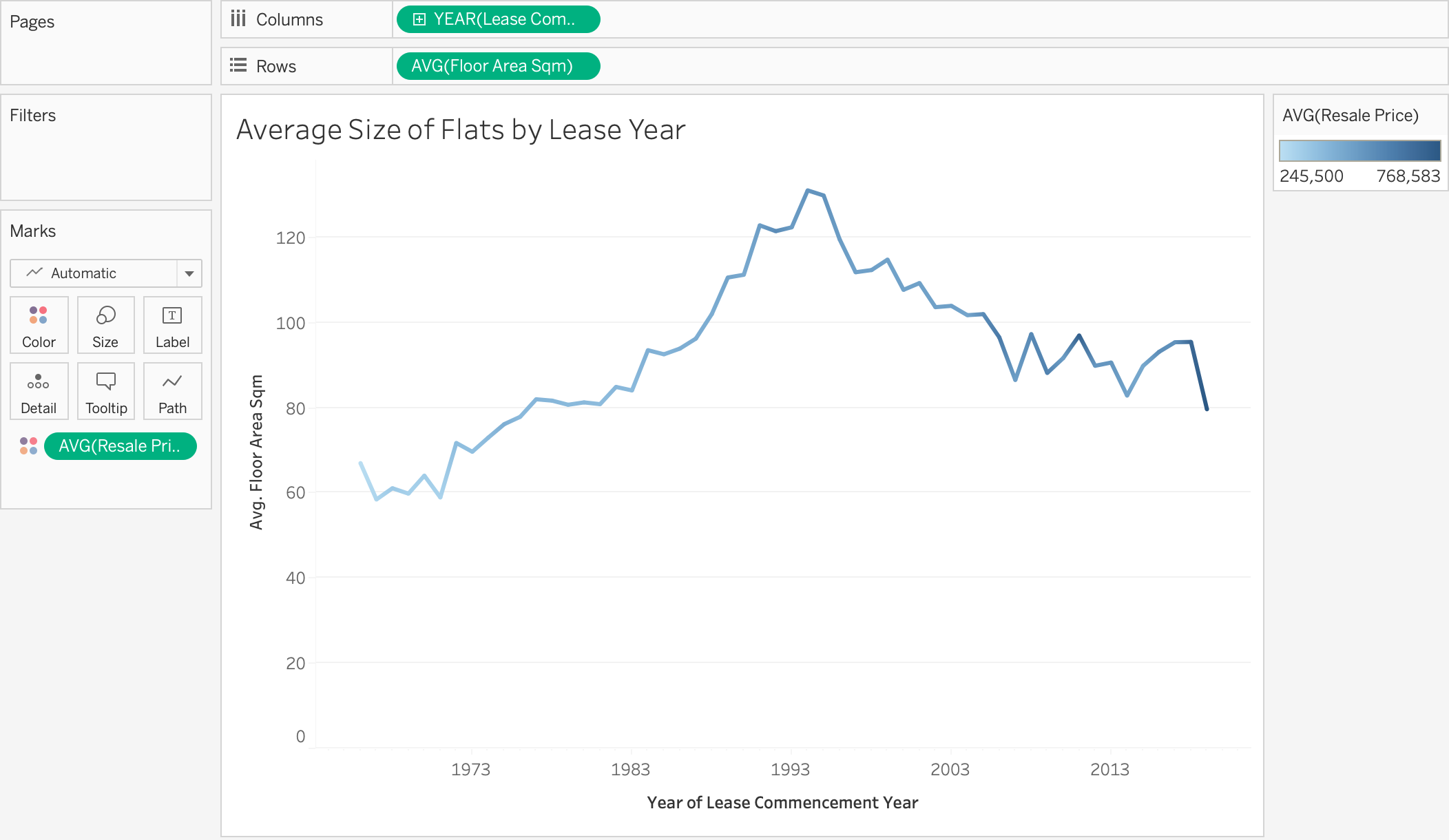
The following visualizations were chosen for their simplicity in conveying information, and for providing a more visually appealing way for users to discern data. Filters were also added where necessary in the event that toggling through features was required. The Tableau workbook consists of six sheets and a dashboard containing four sheets.



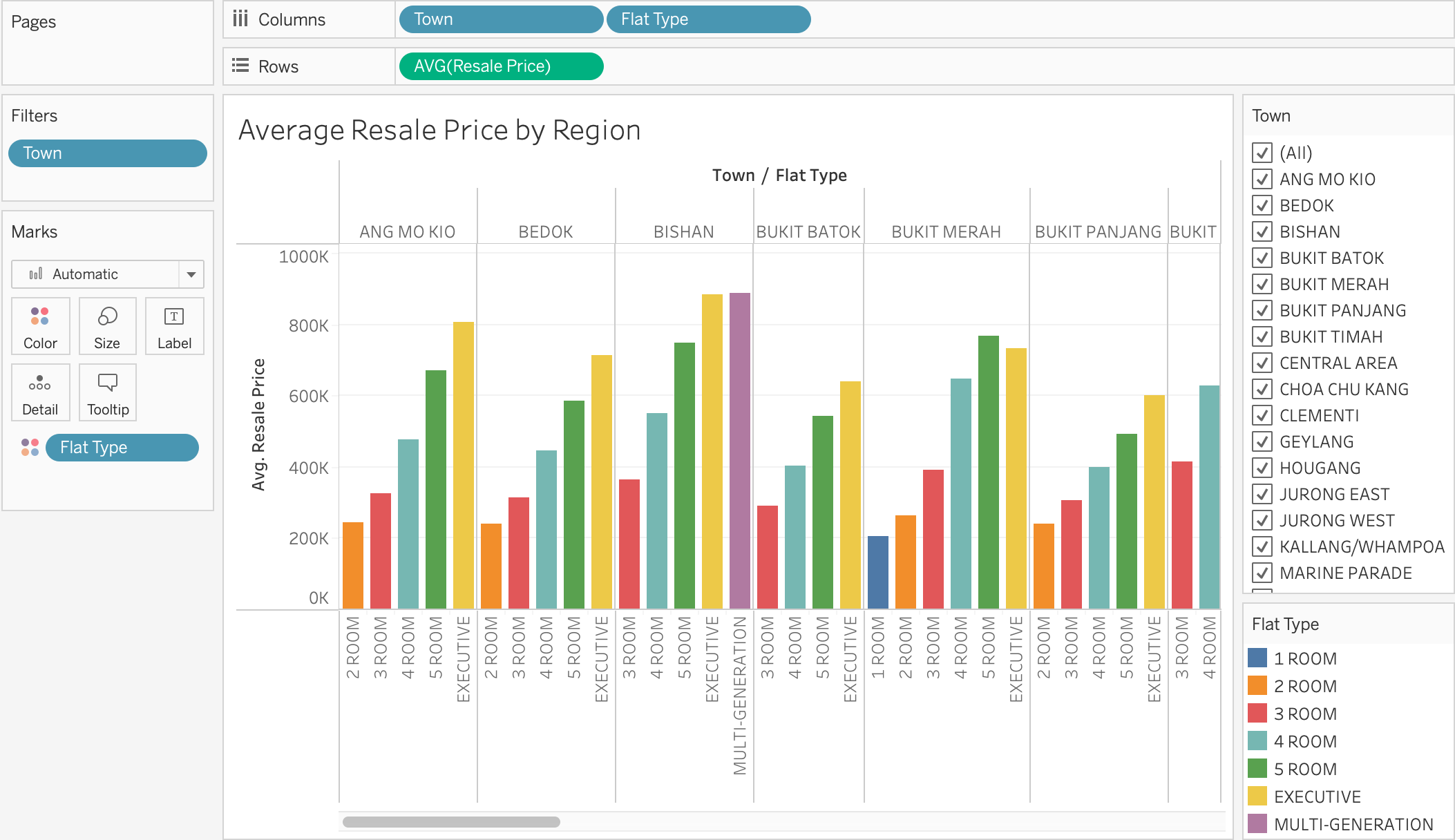
With the first visualization, I wanted to find the trend of resale prices over 10 years using a simple line chart. To my surprise, multi-generation flats had the highest average price but was the only one out of the six categories to have frequent changes in average price over the years. Additionally, where there was a perceived rise in average pricing in 2013 across the other flat types, multi-generation flats suffered a dip instead, and conversely, spiked in average resale price around 2014 where the average resale price for the other flat types was in a slight decline. The price of all flat types eventually rose a little towards 2021.



The next visualization was a scatter chart, plotting the average size of the flats against the average resale price, with the flat type as color and flat model as the size indicator. Through this visualization, I found that despite not having the biggest size and not being in the highest tier of flats, a 5-room flat of Type S2 fetched the highest average resale price. I also set the flat model as a filter in case the user would like to toggle the rather large number of categories.



Using a line chart once more, this visualization helps to check the average size of flats throughout the lease commencement years. The average resale price has also been included as the color indicator. Through this, we can see that flats with the highest average size were from the mid 1990s, while flats with the latest lease commencement years have the highest average price. This ties in with what I have been told by various other locals; that the newer HDB flats are built smaller and cost more, again, possibly due to inflation, and potentially limited real estate across the nation.



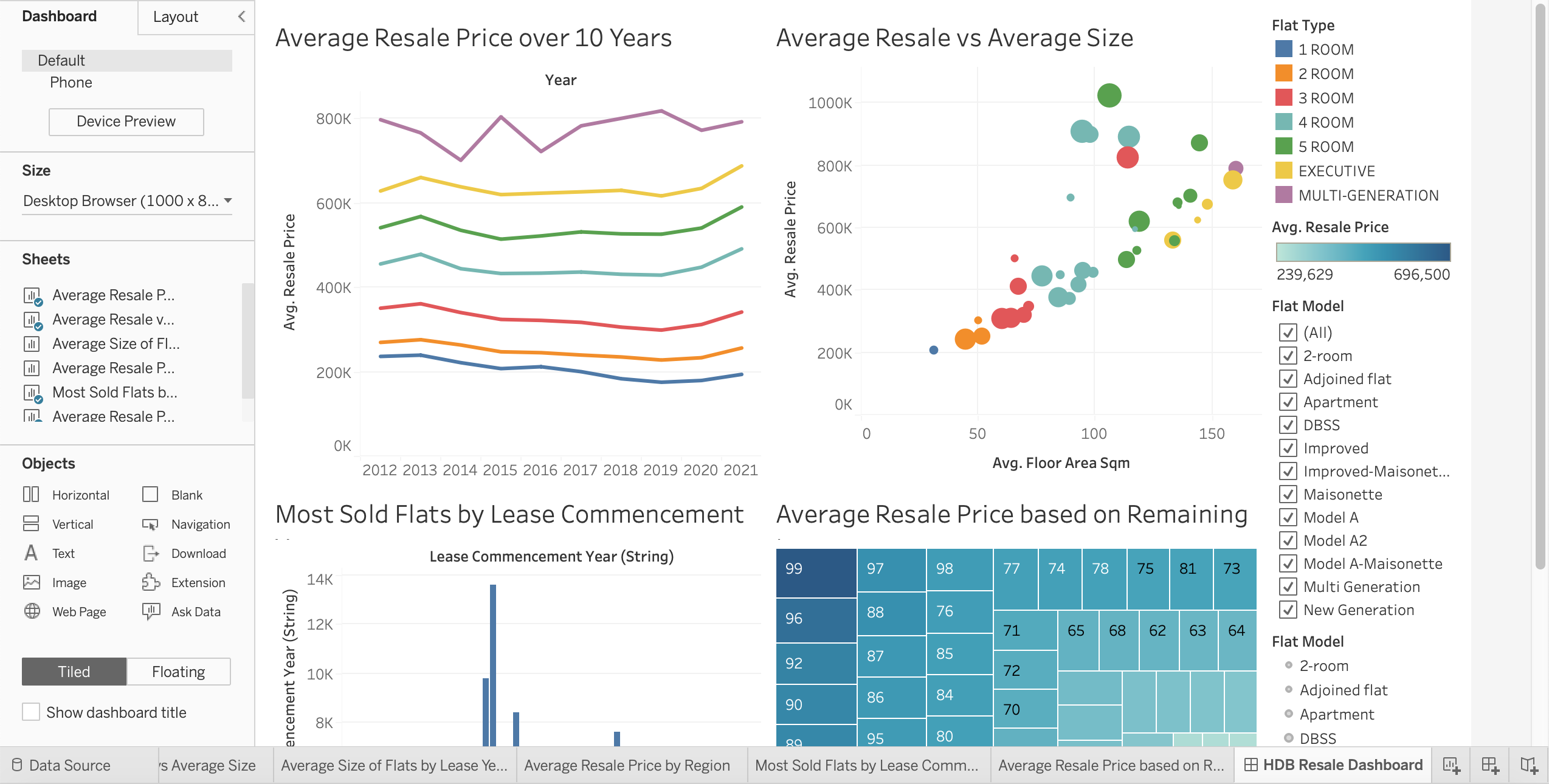
With a bar chart this time, I was able to visualize the relationship between regions in Singapore, as well as the flat type against the average resale price of flats. This chart shows that region, along with flat type, does indeed affect the price of a flat. Using flat type again as the various categories denoted by color, the user will be able to better differentiate the various types of flats. Additionally, I have included a filter option for region should the user want to toggle between regions since the list is quite big.



Going back to the bar chart, I was also curious to see which lease year yielded the highest sales in the dataset. Keeping the visualization simple and using just two features this time, we can see that flats leased from the year 1985 have been sold the most times in this dataset. This insight, along with the chart for average size of flats by lease year, confirms what I have always heard; that the bigger houses, peaking in size in the mid-90s, are more in demand than the newer flats.



Utilizing a treemap, I was able to illustrate the average price with regards to the number of years remaining on a lease. As expected, results show that newly minted flats fetched the highest average resale price.



I picked the average resale price over 10 years, average resale price against average flat size, most flats sold by lease year, and average resale by remaining lease to display in the dashboard because the visualizations remained relatively unchanged, and were not stretched or squashed too much. From this project, I can conclude that the price of flats will not only continue to increase over the next few years, but that the average sizes will not be as big as they were 20 years ago, and that factors such as region, influence resale price as well. On a side note, I was unable to change the axis names via the edit axes window – the editor was blank when I clicked on it – and so decided to change the pill names instead.