Data Visualization Final Project

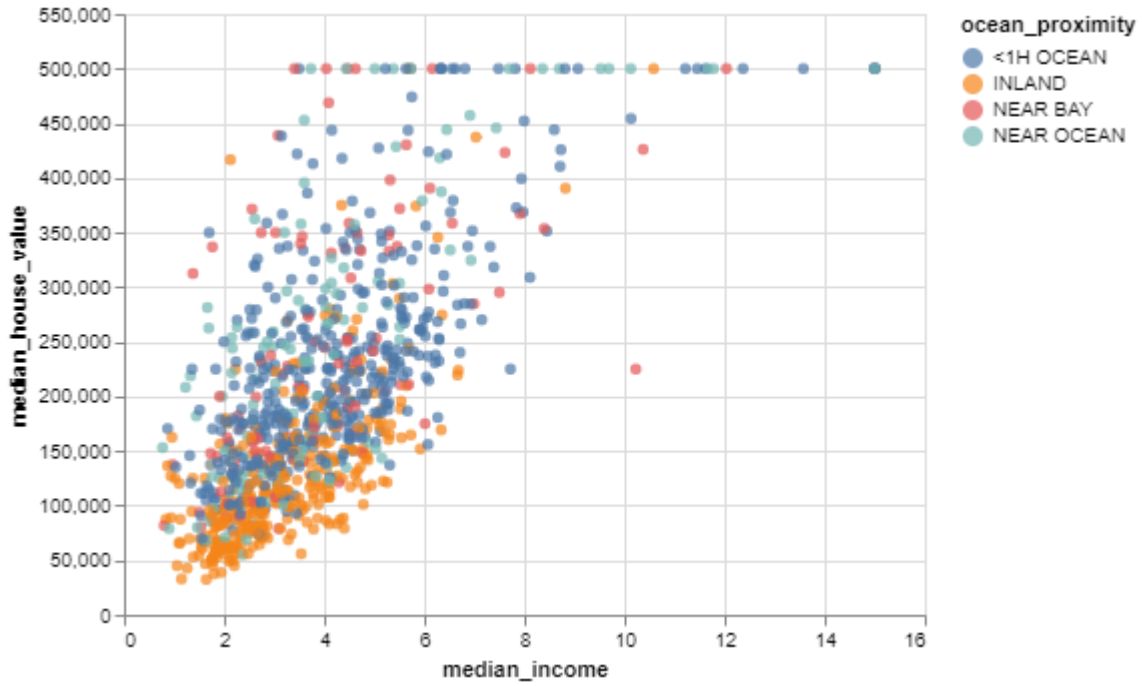
For my visualization project I chose to work with a California Real Estate source that I found via Kaggle, and a sourcing that was originally published for use in the *Statistics and Probability Letters*. I thought this source would be perfect as I am a resident of California and I found it interesting to further examine the state of housing in California. The dataset included information block by block in California for nearly 20,000 blocks, but for sake of use with Altair and run speeds, I took a random sample of 1,000 homes to use in this analysis. The dataset included variables based on location, median home age on the block, the number of bedrooms on the block, the population on the block along with the number of households, the median income of the area and our final key variable of interest in the median home value on the block. These variables can be used to see associations with home value in California. One of my primary goals with this tool was to create a visual that allowed users and analysts to quickly compare trends in these variables on the impact of home prices in California. I wanted a way to incorporate all aspects of these variables in my design to have a comprehensive tool, so with location based data I began thinking of maps, but with most other variables being continuous, it seemed scatterplots were a great option. One of my main goals was to combine these main aspects with some form of interactivity to create an effective tool. The primary tasks to achieve this main goal included creating high quality and easy to follow scatterplots, adding usable maps, and introducing interactivity where possible to create ease of use.

My first iterations involved creating a map of the points with the California state borders as a background. I used the longitude and latitude data to plot the points on the map and colored the points by their location attribute: near the ocean, near the bay area, less than one hour drive to the ocean, and inland. A key part of this task was to create groups that had differentiating colors so that we could identify the groups.
Figure 1: California Mapping by location

My next tasks included creating scatter plots based upon the various continuous variables and the home price. I took each variable and included it in scatter plots that were again

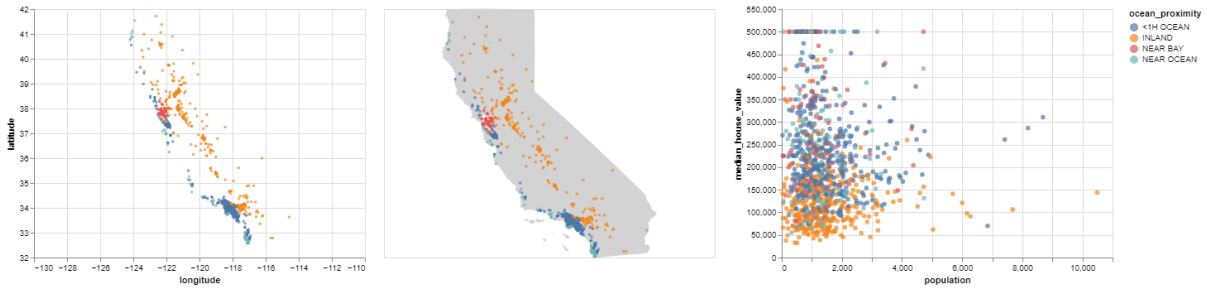
colored by the ocean proximity variable. Figure 2 shows one of the scatter plots produced between income and home value.

Figure 2: Scatter plot of home value and income



After evaluating these steps, I wanted users to be able to select regions on the map and display those selected points in the following scatterplots. Since location plays such a role in home prices, it was essential to include it in our results and trend analysis with other variables. In my next iteration I added functionality for selections combined with the scatterplots. Because Altair does not allow selection based on geographic maps, I approached the issue by creating two map images, one with the blocks plotted on a graph where selections were made, and then these selections would be displayed on a second frame with the map for further location accuracy in selections. In the future, Altair stated they would add selections on geographic objects, which would improve this product. Then the following frames displayed the scatterplots from the selection.

Figure 3: Second iteration, adding interactivity



In the above images, selections can be made for further location based analysis. Additionally, more scatter plots are produced, but this image is simplified to fit in with pdf formatting.

Now that I had improved on my graphs, I discussed the project with a student from my undergraduate program and several family members. My evaluation process involved leaving my family members and friends with the tool and a brief description of the data. I

would let them get used to the tool before asking them to compare trends in home prices across different regions of California as a test of the tool. This allowed me to analyze and evaluate the tool with a journaling study, where I could take notes on how they use the program and their feedback. In general, they quickly achieved their goals and had only a few recommendations. They recommended polishing titles, axes, and adding zoom interactivity to allow for better selections if someone wanted to visualize a specific region. I found this evaluation process to be really helpful, as it introduced a set of eyes to my tool that I had not noticed after working on it diligently. In the future, I will be sure to utilize evaluation and peer review techniques we discussed in this course. I found that my preliminary models were close to being finished, with a few adjustments from my evaluation. My polished model is displayed in figure 4.

Figure 4: Polished Model with Zoom interaction

The above image displays the zoom capability with selection in frame 1, with the overview of the selection displayed in frame 2 in the state of California. The scatter plot is polished and includes only the selected points. Only one scatter plot is shown again, to allow formatting in the pdf. A brief analysis of the data shows that higher population and higher incomes lead to increased home value, along with increased prices in coastal regions and near cities like San Francisco, Los Angeles, and San Diego.

Key Elements in my Final Approach:

- Including coloring by ocean proximity
- Mapping that included a zoomed, select frame and a second frame showing the overview map.
- Following frames including our continuous variables to properly investigate trends by our desired area.
- The option to investigate multiple variables
- Key use of position, color, and mapping

This tool allowed for users to quickly analyze trends by specific regions and consider several variables at once. I believe using mapping techniques along with selections allowed the tool to quickly incorporate the importance of location into the analysis and investigate how that pairs with the other variables. One portion of this mapping that could use future improvement is having only one mapping frame, as we introduced two since Altair does not allow interactivity with geographic objects. Future improvements can be made as updates occur, or using another software package. Another improvement I could make is having only one third frame, and allowing selections for the explanatory variable, instead of having all the graphs output at once. I did think that having all the graphs out at once did allow for the

quick synthesizing of trends in the home price data so I kept it as is for now. Future work could create solutions in this area. Another improvement could be using a document such as markdown that allows interactions to be investigated in a document like this, as a pdf does not allow interaction. One of the greatest aspects of this project was the evaluation phase. There were a few minor issues with my project that required only a quick fix, but I had overlooked them since I was so involved with the design. Being able to watch others use the tools for a purpose I designed allowed me to quickly spot any issues and gain valuable feedback and insight. Overall, I thought my tool was quite effective for fully synthesizing and analyzing all the possible data and variables using one interactive tool. Thanks for joining me today and I hope you enjoyed my visual!