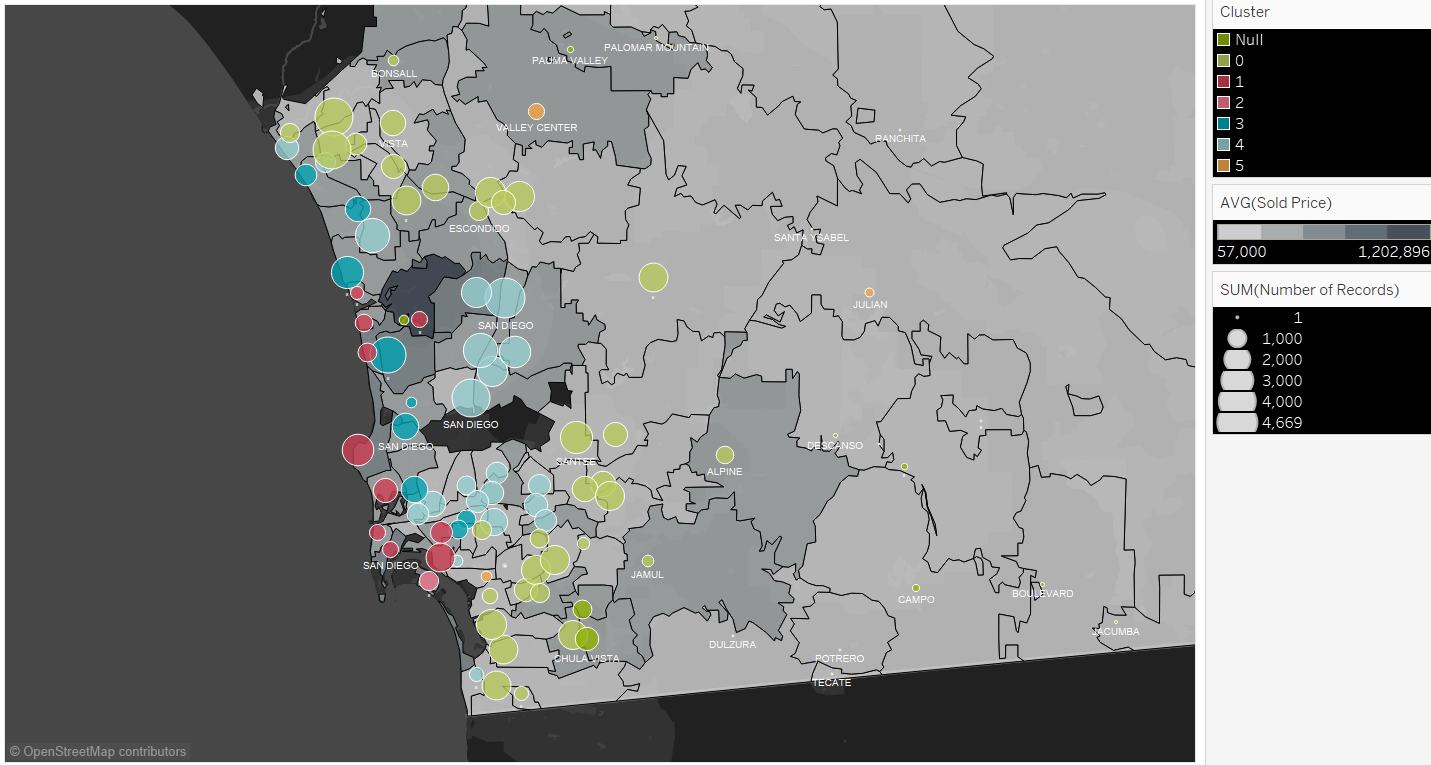
**San Diego House Price Predicting**

**Wen Yan, Mengting Wang, Salah Ahmad, Xia Song**

**Major Findings and Audience**

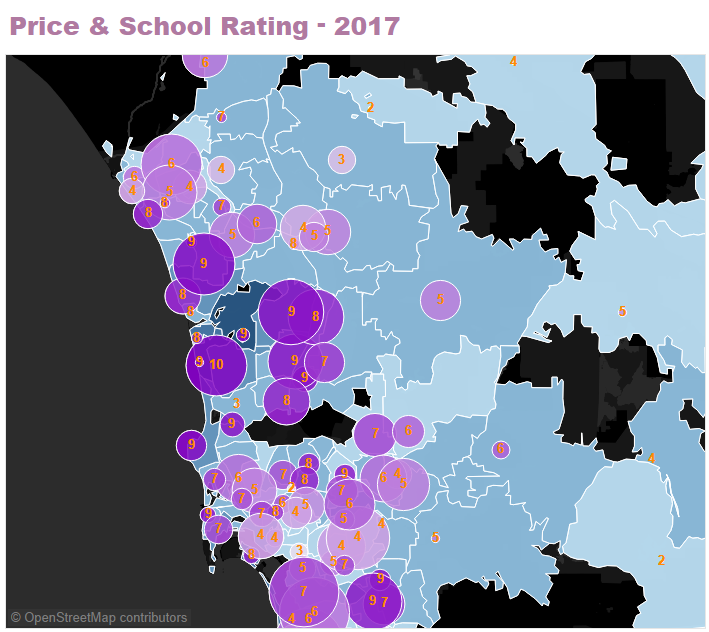
Part 1- Price Distribution over Geographical Clusters



Audience- Investors, Sellers

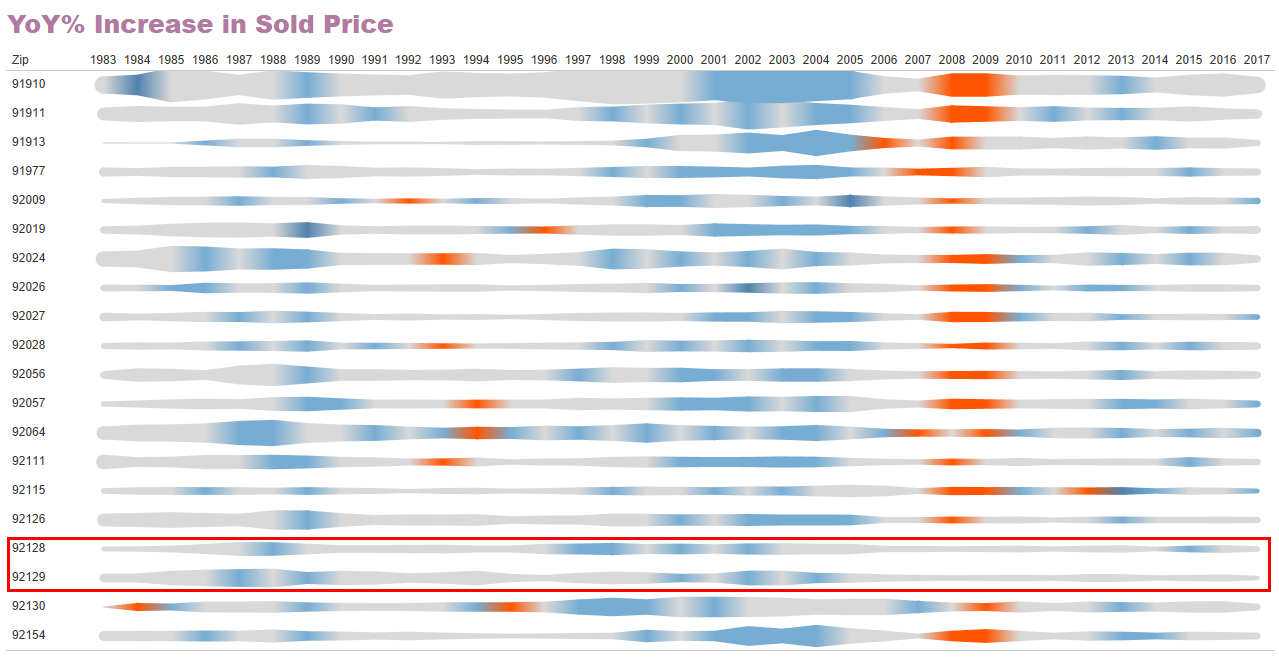
This visualization is a representation of the whole county sales. Sold price is the background color. The darker the color is, the higher the price. Size of the bubble represents the number of transactions. Bubbles show our clustering. I.e. notice the costal red cluster (Central and south San Diego)

Part 2- Zip price YoY% change (for potential investment opportunities)



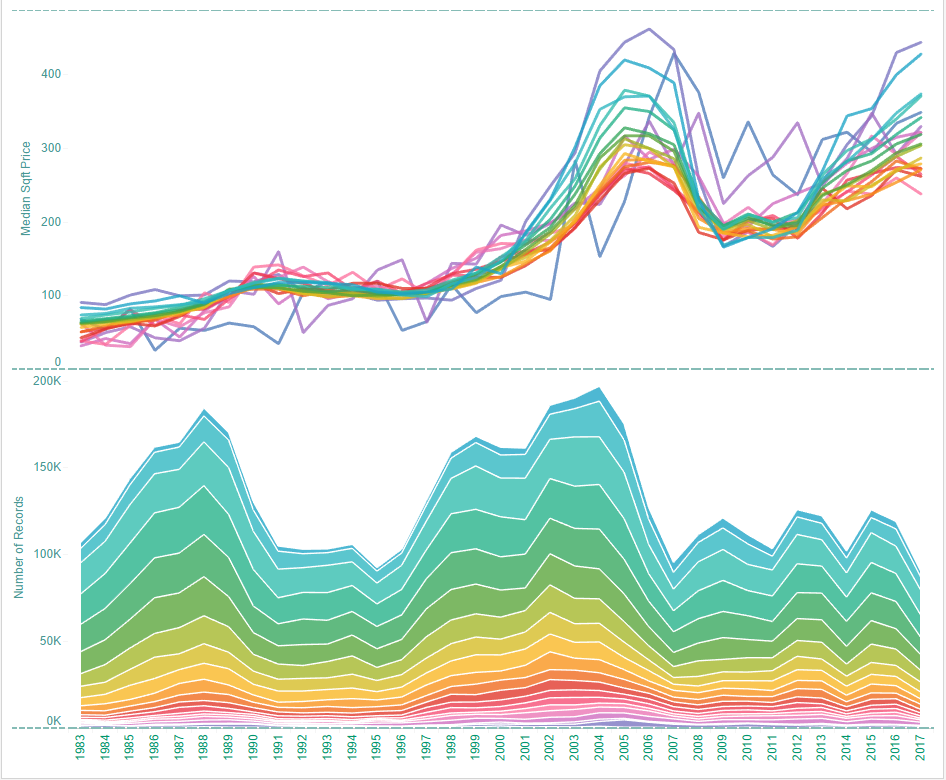
Audience- Investors, Sellers

Notice the size of the bubble represents the number of transactions. The darker the bubble is, the higher the price. There is an obvious correlation between the school rating and the price of the house.



Audience- Investors, Sellers

This is YoY analysis from 1983 till 2017. Red means significant decrease in price, gray is lowest and blue is medium. Notice the two zip codes highlighted (92128, 92129) are the most two stable zipcodes in the county. Others follow similar rhythm with slight variation in behaviour.

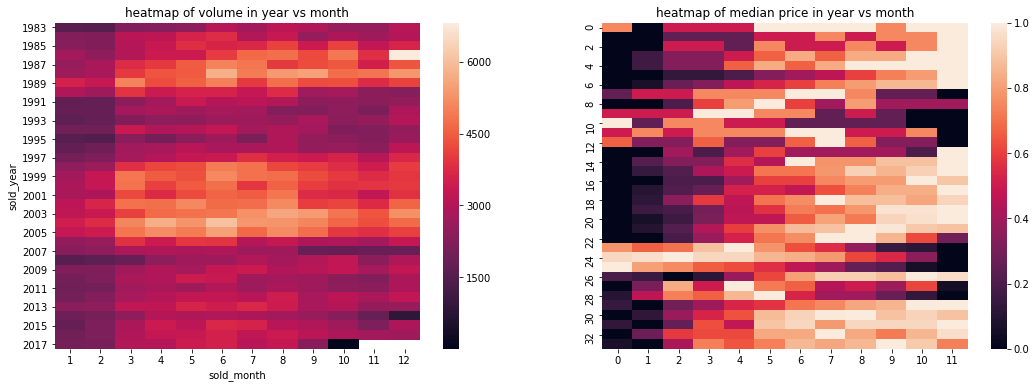


Part 3- Timing of investment/seasonality

For seasonality patterns, we looked into both volume and median price trend, heatmap is the most efficient way to show the pattern, as in below figure, the two axis are representing year and month separately.

For volume(figure on left), we can see the volume increased dramatically during 2 year segments, roughly 1987~1989 and 1999~2005 when the market bloomed. And volume is consistently at bottom around Jan. and Feb every year and at peak around Jun. Jul and Aug.

For median price, the absolute price difference between months is not big, the heatmap(figure on right) was showing normalized data otherwise the variance between different month is small and hard to see. We can see clearly that price is lower during winter time(black area) although it’s not consistent on beginning from which month. In most case, price fell to bottom in Jan. and Feb., but there’re 2 year segments that it fell to bottom since Nov. or Dec. It doesn’t seem to be coincidence that the 2 segments were aligned with the period of market crisis.



**How to Determine What to Present in Findings?**

Throughout the entire project, we have been developed many visualizations that served multiple purposes, such as EDA, examining outliers, evaluating residuals, and presenting results. We want to focus on the most important and interesting findings to present in the final presentation, in order to tell a compelling story of SD housing market, as well as demonstrating an intuitive tool for audience to explore on their own. To achieve that, we categorize our visualizations by identifying the audiences-- investors and sellers, as well as individuals who are interested in knowing the housing market in San Diego. With this target set, we lay out the visualizations by showing the trend/seasonality to provide a basic understanding, presenting features by geographic locations to provide a zoom-in on popular features, and showing price changes (appreciation vs. depreciation) by zip codes to give an idea on which are the areas that appreciate/depreciate the most throughout history.

**Techniques and Tools**

Tableau is one of the main tools that we’re using to produce majority of the visualizations and to perform most of the EDA work.

**EDA:**

Tableau was used throughout the entire course of project for plotting trends and detecting anomalies. These visuals are produced by zips, by sold year, with targets of features such as sold price, sqft price, number of bathrooms (&bedrooms), sqft, etc. We strategically use statistical charts such as cluster charts, box-plots or scatter plots to pre-process data, identify structural and skewness of data for a given geographic or time related features. As we connect Tableau datasource to PostgreSQL, we are able to clean identified outliers and anomalies and refresh Tableau dashboard in real time to see the improvements, which increased efficiency.

We also aligned visualization techniques along with our research purpose and strategically applied multiple visual channels to maximize the information. For example, in a choropleth map, we use colors on the map to indicate average selling price for each zip code; using size of bubbles to indicate number of transactions within the zip codes, and with colors of bubbles showing average school rating of zip codes. Within one visualization, we are able to communicate multiple layers of information to audience. Filters and automations are also used in the meantime for easy maneuver throughout the visual, especially for audiences who are interested in specific areas of given features.

**Residual Analysis:**

As mentioned earlier, since we connect Tableau datasource to PostgreSQL, it enables real time data refresh for better model performance tracking purposes. We effectively plot residuals and have them spread over choropleth map for clean and easy identification of “problematic” areas in terms of price range or geographic allocations.

**Communicating Modeling results:**

Last but not the least, we use this tool to communicate our modeling results in terms of trending and clusterings. However, one thing that we noticed is that it is relatively difficult to control multi-dimensional information through Tableau, especially when we are trying to plot trending of our modeling estimates vs. historical actuals-- it is doable but the visual does not communicate the information in quite an efficient way as we were hoping for. Therefore, we also developed website to show trending and seasonality in a more effective and cleaner manner.

**Web Based Visualization**

A prototype web application is developed for both user exploration and our EDA and debugging purpose. It gives the flexibility to achieve any visualization since we have full control of the visualization coding. It’s much faster since it can be optimized at various levels, eg data aggregation in database, web server, or javascript front-end. It’s easy to scale to support many users’ interaction in real time, eg most basic setup allows hundreds access at same time.

Javascript is the main front-end coding language used with the help of various libraries including D3, C3, datagrid, leaflet etc. Ruby on Rails is the web framework which is very good with ORM connecting to relational database eg Postgresql as what we used.

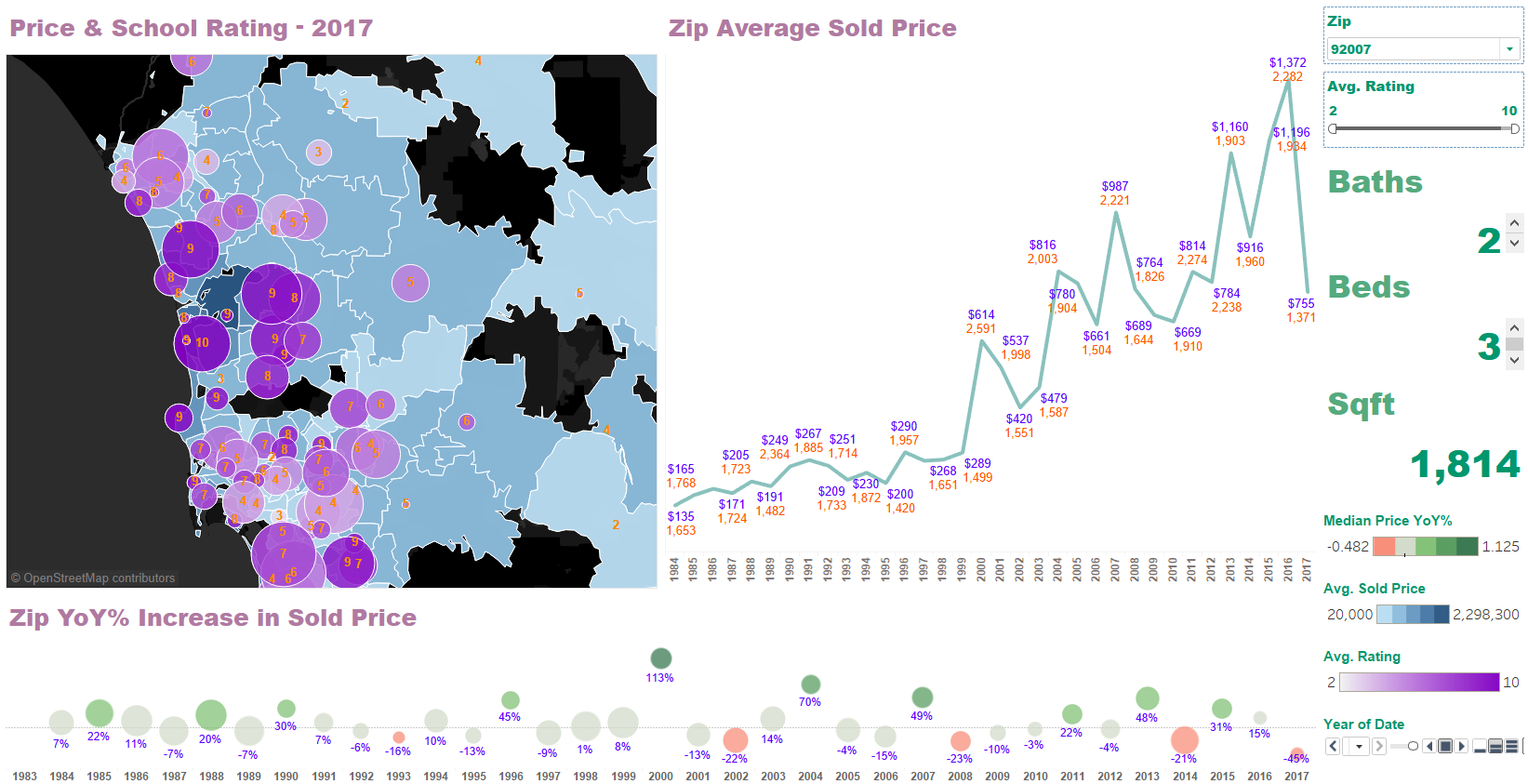
We have 4 dashboard pages planned, house searching, market statistics, EDA results and modeling for price estimation. The EDA and modeling dashboards are more like static pages that demonstrate the findings and research results from us, and house searching and statistics dashboards are provided to user for house hunting and as a tool for their own EDA.

**Visualizations and Other Products**

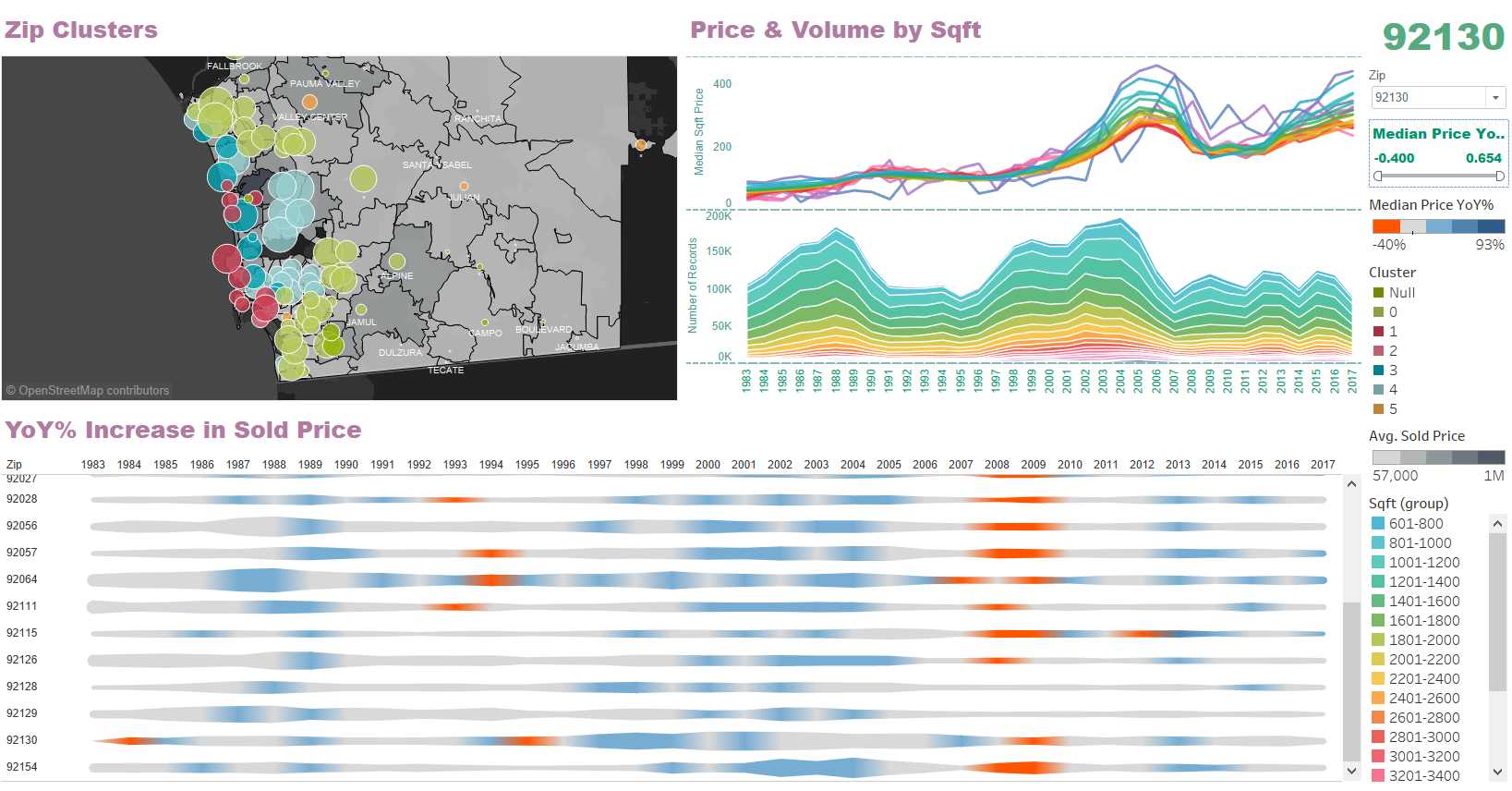
* **Tableau Dashboards**

Throughout the entire course of work, we developed numerous versions of Tableau visuals and dashboards for multiple uses as mentioned above. We ended up selecting two dashboards as our final product to demo in our presentation. These two dashboards included our findings on EDA, partial of our modeling results (another part will be included in web based visualization), and can be used as a tool for audience who are potential investors, or simply are interested in San Diego housing market.

Dashboard 1 below is a typical tool for audience to maneuver and get a good understanding of housing market information by different zip codes. In this dashboard, audience can effectively select zip code of interest, dashboard will return results on a zip code’s average sold price, zip code year-over-year price increase and number of transactions, and number of bedroom, bathroom and square footage. Average school rating is organized and distributed on a choropleth map, where the background color shows average sold price of zips, bubble sizes indicates number of transactions and bubble color shows school ratings.

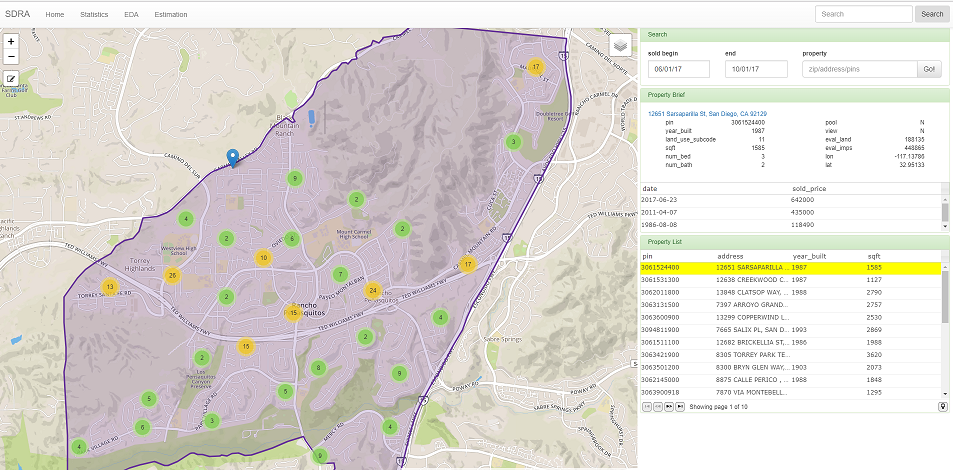
****

Dashboard 2 below illustrates our zip clustering results in the top left corner. Not surprisingly, clusters clearly indicates cutoffs on coastal, inland and mountain areas. In this dashboard we also include zip code price vs. sqft. Interestingly, the smaller the properties are, the quicker the price increases and decrease; and vise versa for bigger properties. Last but not the least, we enabled filter on YoY price changes for zips. Users can easily select desired YoY increasing rate to see which zip codes meet the criteria and overall trending throughout the entire time span.

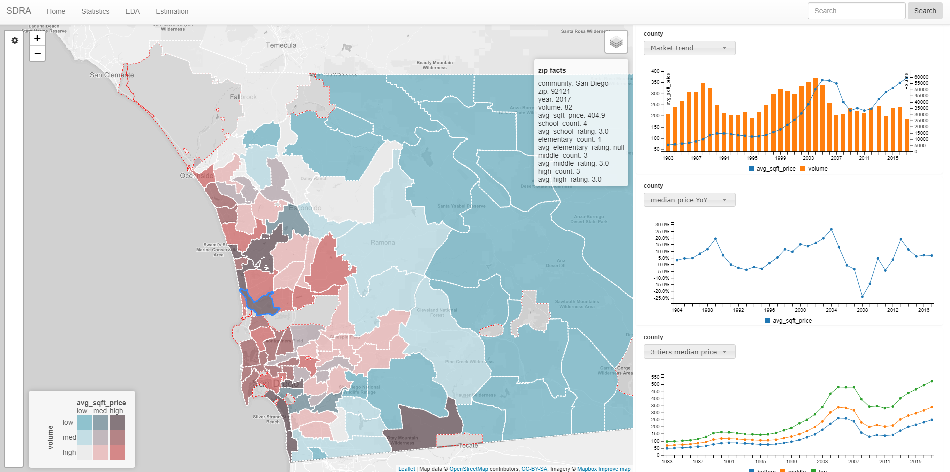
****

* **Web Dashboards**

‘House Search’ page tries to mimic basic search functions that any realtor website provides. It’s also used for our debug purpose for our project. It allows various searching criteria and drawing a scope on map to select houses or transactions during a time window.



‘Market statistics’ provides various choropleth encoding with up to 4 market statistics indicators and the customizable trending charts on right side.



**Team Contribution:**

* **Wen:** implemented web visualization for prototype application, visualization for seasonality.
* **Mengting:** developed Tableau visualization charts and dashboards for findings and interactive data exploration tools
* **Salah:** Review sessions, analysis and some report writing
* **Xia:** Preparing final report writing