**San Diego House Value Analysis**

Advisor: Ilkay Altintas, Volkan Vural

Team Members: Wen Yan, Mengting Wang, Salah Ahmad, Xia Song

**Abstract**

San Diego housing market is on the rise. Investors, buyers and sellers are all seeking insights and clues related to the housing prices. This effort explores various housing data such as Sales, Property Characteristics, School, Foreclosure, etc... We aim to provide insights and visualizations for findings. Our solution provides easy to follow and interactive visualization showing the house prices in conjunction with various other dimensions such as location, year built and size of the house. We also provided historical trending illustrating the cyclical nature of the market .We studied and provided insights about sales seasonality in conjunction with various attributes such as location, prices and size.

**1. Introduction and Question Formulation**

* The challenge that leads to the problem: What contributes to the house prices? There are so many attributes related to the house itself such as size, year\_built, etc.. and others related to location, school district, etc..What are these features that are really driving a difference in housing price? Are houses sold the market higher or lower than how they are supposed to be?.
* The ingredients to form the data science problem:
  + Various data such as property characteristics, sales history, etc…
  + Data Integration for easy and consistent processing.
  + Data exploration
  + Descriptive analytics and data mining to find and produce insights for investors, buyers and sellers.
  + Predictive analytics as an attempt to give a value for properties.
  + Communicate findings and insights through interactive and intuitive visualization.
* The questions to the data science process:

### What are major factors; except global economic environment; that impact the housing prices in San Diego county under current market?

### In short-term uptrend/downtrend market, what type of houses by location, profile (attributes) will appreciate/depreciate faster than others?

### In long-term (several decades), will certain houses or areas appreciate more than others?

### Can we evaluate the house price based on current average price movement in its area?

**2. Team Roles and Responsibilities**

The team is consisted of multiple key functions and responsibilities. Each team member is responsible for one to two responsibilities, and make themselves available for supporting other functions when necessary. The key functions include:

Project Managing-- Mengting is the project manager for the project and also performs visualizations throughout the entire project. She organizes meetings within the team and with advisors, creates plans with team members for each step and to ensure goals are accomplished on time.

Data collection, database construction, integration and querying-- Salah is the key contact on this area. He plays an essential role in getting the data integrated and dumped into flat files. On another front, he connected with data source provider-- San Diego County-- to ensure all our questions related to data is answered and clarified, as well as following up to obtain new data whenever needed.

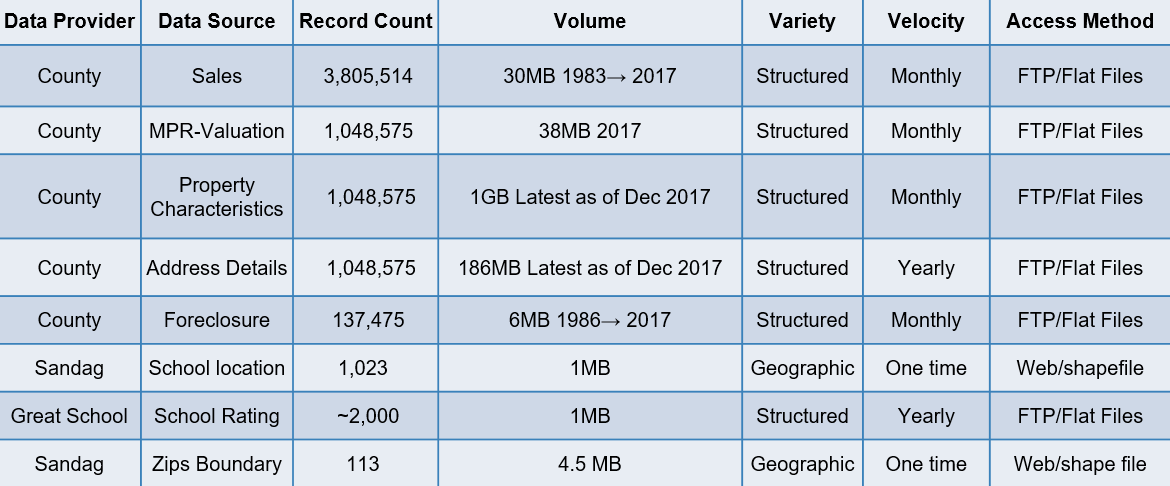
Feature engineering and general data engineering function-- Wen is taking responsibility for feature engineering role as well as majority of all other data engineering related tasks. He is the key person for the entire technique solutions, architecture, model evaluations and general research design. He also contributes in developing web based visual as a part of final demo.

Modeling-- Xia is mainly focusing on modeling tasks. She performed research on selection of models, model tuning and parameter selections.

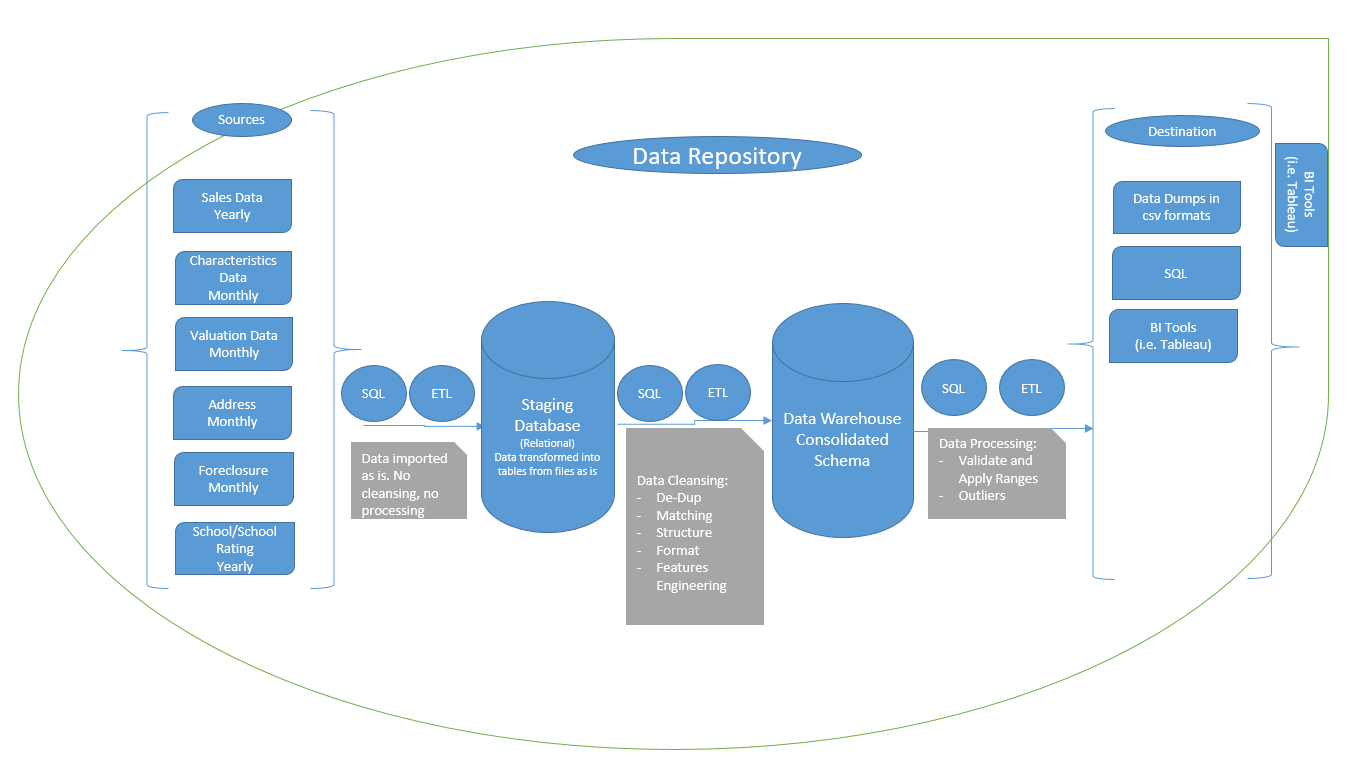
Visualization-- Mengting is leading this role throughout the project. She developed visualization tools and dashboards which served EDA process, model evaluation, residual analysis and communicating final results. She is also developing tools for final presentation demo.

**3. Data Acquisition**

**Data Sources and Collection:**

****

**Data pipeline:**



**Data Deployment:**

* Used RDS on AWS cloud for storage.
* Also available Offline on postgres.

**4. Data Preparation**

**Data Quality Issues:**

* Invalid content/data types: Some fields are supposed to have numeric values, but some of their related records came with explainable texts (Year\_built is an example of that).
* Data structure:An example of that is the year-built field. It comes with 2 digits so when you have 09, it could be 1909 or 2009.
* Formatting: Some fields are combined together. The address field was not split into Add1, Add2, City, State and Zip in some of the feeds.
* Missing values: Many of the valuable fields didn’t have values for the whole dataset. Example of that is the view for the property.

**Data Transformation and Integration for analytics:**

* Data Cleansing: For invalid content and when it is safe, our SQL/ETL altered the data to fit the file layouts and our integrated schema.
* Data Formatting: For consistency and easy querying, all data from all sources were formatted to meet our integrated schema layout.
* Data Finding: In some cases, we had to apply logic and go to other sources to decide about the correct of the records. Example of that is how we handled the 2 digits year\_built. We had to go and look if there are sales or tax bills in these records in question prior to 2000 or not. If there was, then we counside them build in the 19xx.
* Data Matching: We applied basics string matching to link the data sources with each others. Matching Parcels and names are examples of that.
* Data Modeling/Reconstructing: From flat to staging to the DW/integrated schema, we made the data lives in easy to query and process set of tables.

**Pre-Processing Methods:**

* Data sources identification and Acquisition.
* Data moved to staging as is.
* Data being cleansed, formatted and restructured.
* Features Engineering.
* Remodeling the data for easier querying.
* Queries built, optimized and performance tuning concepts were applied such as materialized queries and indexes.
* Data dump into flat files for analytics.

**Features Management and Selection:**

* Features were reviewed from each source with domain experts to discover meaning and relevance for our analysis.
* Features were categorized into 3 groups:
  + Important and needed for our analysis: These were used in our analysis till they prove to be not needed or they don’t add any value.
  + Potentially needed for our analysis: Examined to be used in the analysis and modeling.
  + Not needed for our analysis: Excluded from our analysis and modeling.
* Feature Engineering for missing and others that require extra processing. Examples are:
  + Longitude and altitude: We used postgis encoding extension
  + School Distance from the property.
  + Average rating of closest 3 schools
  + Aggregated features such as zip average sqft price.

**5. Analysis Methods**

* **How did you identify methods to perform preliminary analysis of your data? What was significant about using these methods? How did they influence the design of the next steps of the project and define your data science question(s) further?**

Data analysis includes different processes, for example, data inspecting, data cleaning, data transforming and modeling data, and all these processes are served for discovering useful information to support project purpose. The original data in its raw form is not meaningful for the objective purpose. The major purpose of data analysis is to seek concepts, patterns, themes and insights (Patton, 2002). Our datasets include properties characteristics and transaction history and these data provides multiple scans of property and their appraisal for different years from San Diego County data house.

Based on our project objectives and data status, a preliminary statistical analysis and data visualization were conducted to perform our data preliminary data analysis, and statistical data analysis include arithmetic mean method and correlation method. Through preliminary analysis, it was found that mean house price analysis significantly help us to detect house price changing patterns over time and house price difference in different spatial scale, and the correlation between different house characteristics with mean house price help use to determine the important features. The data visualization provide direct insight for us to understand the house price changing trend over time and spatial gradient.

Based on our preliminary analysis found that data features needed to be modified or combined in some way in order to be suitable for regression model analysis. In addition, we also need to ignore some data items that are correlated with each other in ways that could affect the regression analysis.

* **What was involved in applying analysis techniques to your data? Why?**

During conducting data analysis, missing data and outliers data processing are involved. Missing values and outliers are frequently encountered while collecting and processing data. The presence of missing values lead to a smaller sample size, and it can also produce biased results (Kwak and Kim, 2017). As a part of the pretreatment process, missing data are either ignored in favor of simplicity or replaced with the substituted values estimated with statistical method.

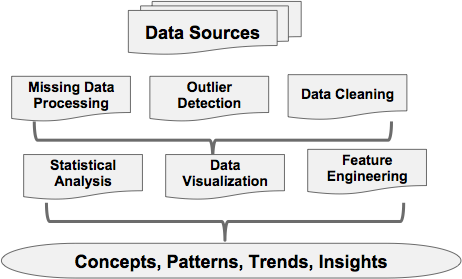
Feature engineering is another techniques involved in our data analysis. Through analysis we realized that more featured maybe useful for our regression model to predict future house price. At first we found that square feet, number of bedroom, number of bathroom, house view, pool, average square feet in zip code level, average square feet price in zip level and average sold price in zip level are most important features for our model, however, house age, distance to school and score of school can also affect house price.

* **Which basic analysis techniques you used and why?**

Three major data analysis techniques were used in this project.

1. Statistical analysis was used to detect house price trend along historical and spatial scale.
2. Data visualization provided direct insight for us to understand the house price changing trend along historical and spatial gradient.
3. Feature engineering help us to determine important features for our regression model.

* **Analytical workflow**



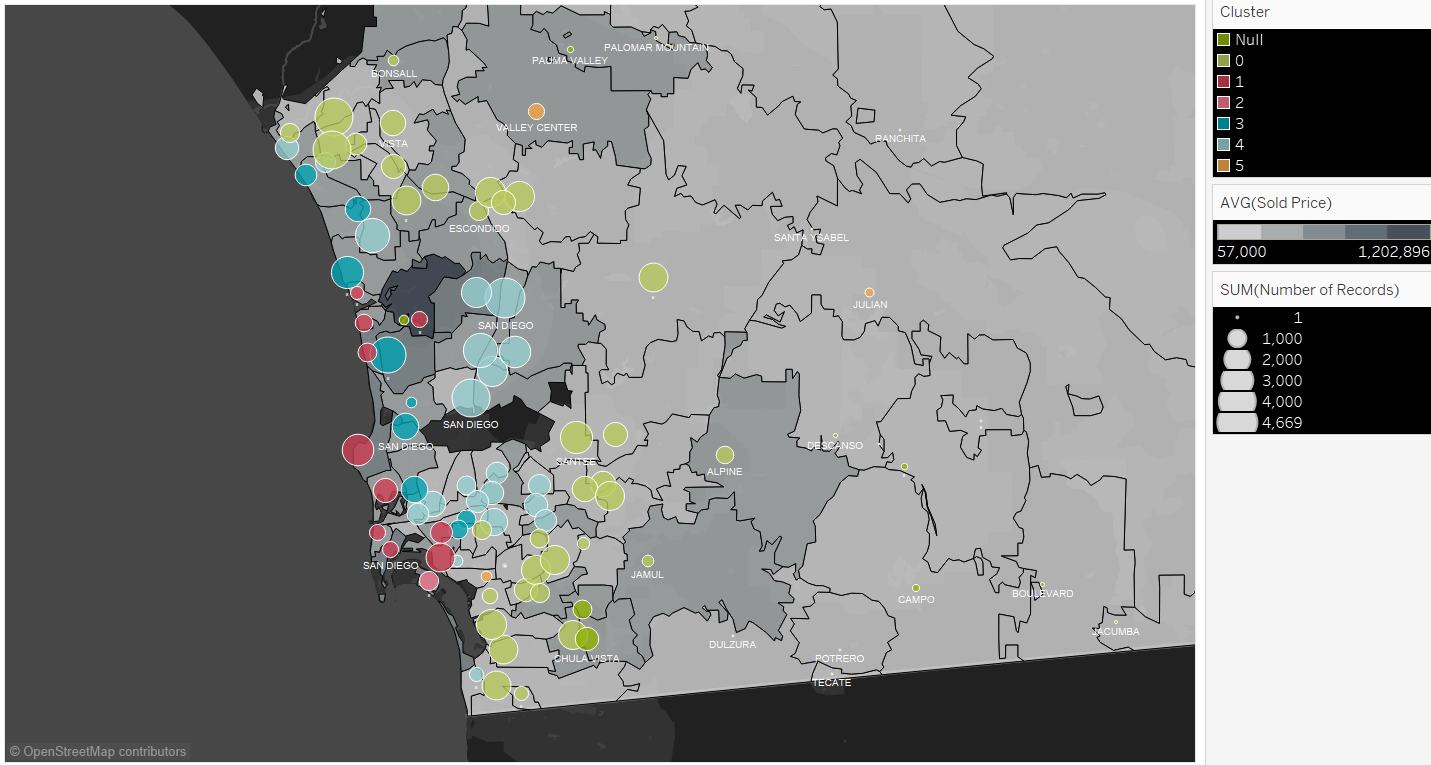
* Set up for your processing environment

**6. Findings and Reporting**

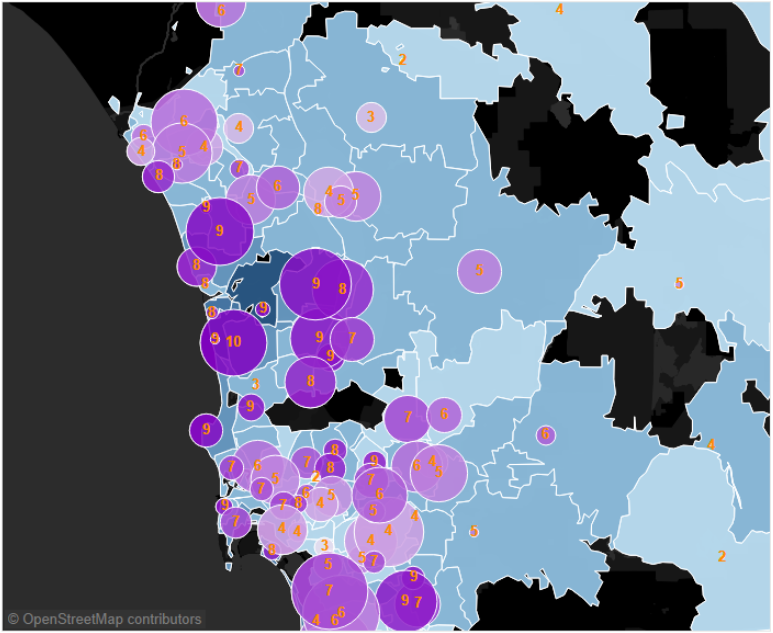
* **What are the findings?**

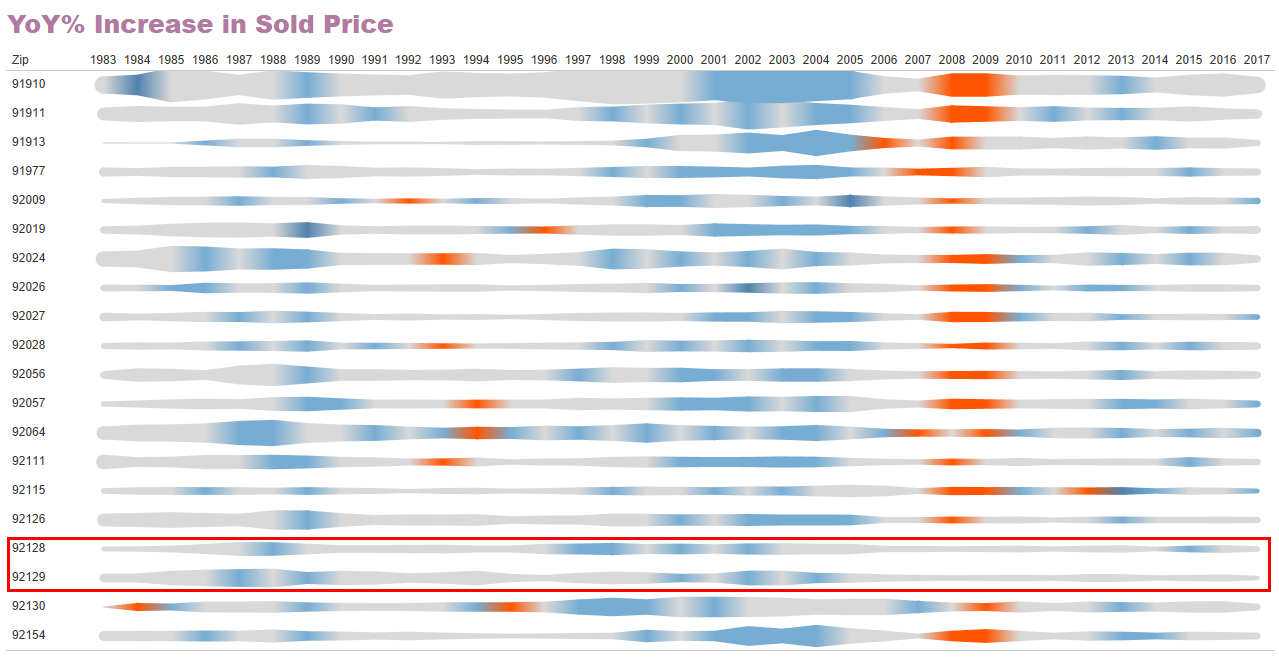
We’re driven by the interests of major audience throughout the project, so that we organized our findings in the sense and logics that people would be interested in gaining deep understanding of the SD housing market. With that, we categorized major findings into three steps, with each step deeper down than the previous.

We started by introducing the overall price distribution of SD housing market. We carried out visualization technique to map our observations of price over entire county through geographical manner through choropleth map. Instead of simply putting descriptive information on the map, we are also presenting our data engineering result in terms of zip clustering. In the map below, each color represents a cluster, which was calculated based on price behavior of each zip code over the past years. Size of the bubble represents number of transactions occurred in each zip and background of map (in grey) shows average sold price of each zip. The clustering results clearly demonstrate a clear cut between coastal area, inland area and mountain area, which confirms that in San Diego county, there is a clear difference in price behavior among coastal, inland and mountainous area.

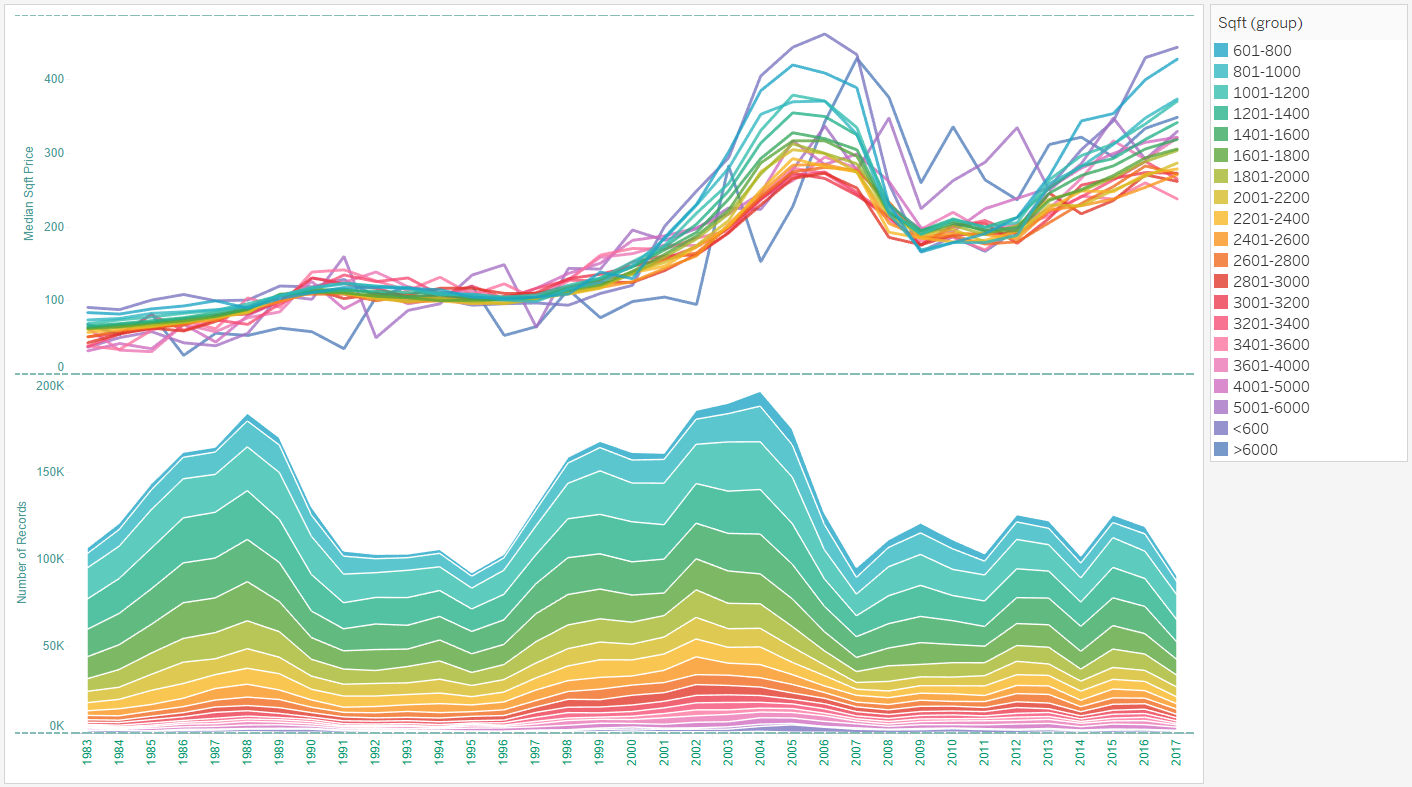


Taking one step further, considering investment opportunities, one may wonder which area in SD has the biggest potential to increase. This consideration definitely involve price increasing patterns among different zips, and also school rating should also be taken consideration as a major factor. We identified most popular areas in terms of number of transactions, highest school ratings and high year over year price increase for investment consideration. For example, 92130 (Carmel Valley) being one of the most popular area with highest school rating, started to boom since 1984; 91910, 91911 and 91913 (Chula Vista) had been popular since mid 1980s, but obviously with higher school rating, 91913 gets more and more popular in recent years and price has gone up since late 1990s; 92127 (east bound of Rancho Santa Fe), 92128 (Poway), 92129 (Rancho Penasquitos) has been popular with high school ratings but more friendly price are also good choices for investment. Most importantly, among these 3 areas, 92128 and 92129 are stronger in resistance of the past economic shock-- in YoY % of price change chart below, it is very clear that among the top price increasing zips, these two zips are the only ones that didn’t decrease dramatically as others during 2008-2009 economic shock, which makes them a safer choice for long term investors.





Another interesting finding is related to the price versus square footage. Referring to the graph below: each color represents a sqft bracket, the smaller the property is, the faster the price appreciates over time when the market is hot; however it also depreciates faster too when market slows down; on the other hand, the larger the property is, the slower the price appreciates when market is hot, but it also depreciates slower when market slows down. So in general, smaller properties are good for short term investment in the hot market; larger properties are safer investment for longer term investment.

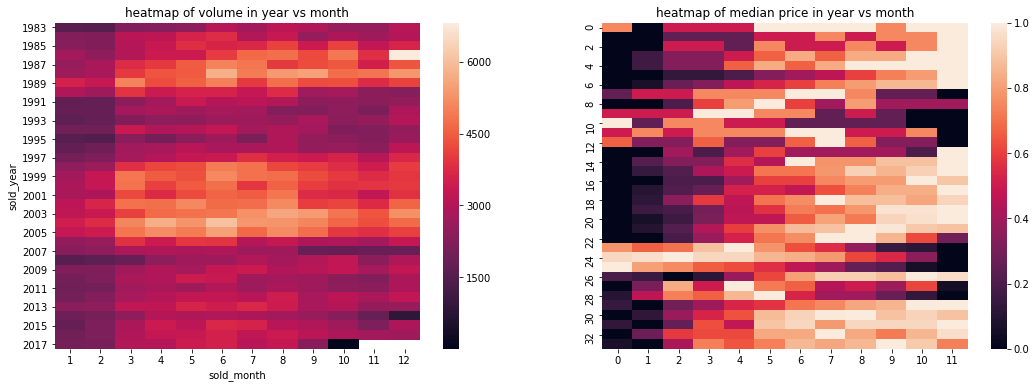


Last but not the least, we also looked into timing of investment by analyzing seasonality.

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 are 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 did you determine what to present in reporting your 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.

* **What techniques and tools did you use to communicate your results?**

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**

As we try to mimic what a realtor company need provides, that is public usage like browing, searching as well as analytics for market insight, 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.

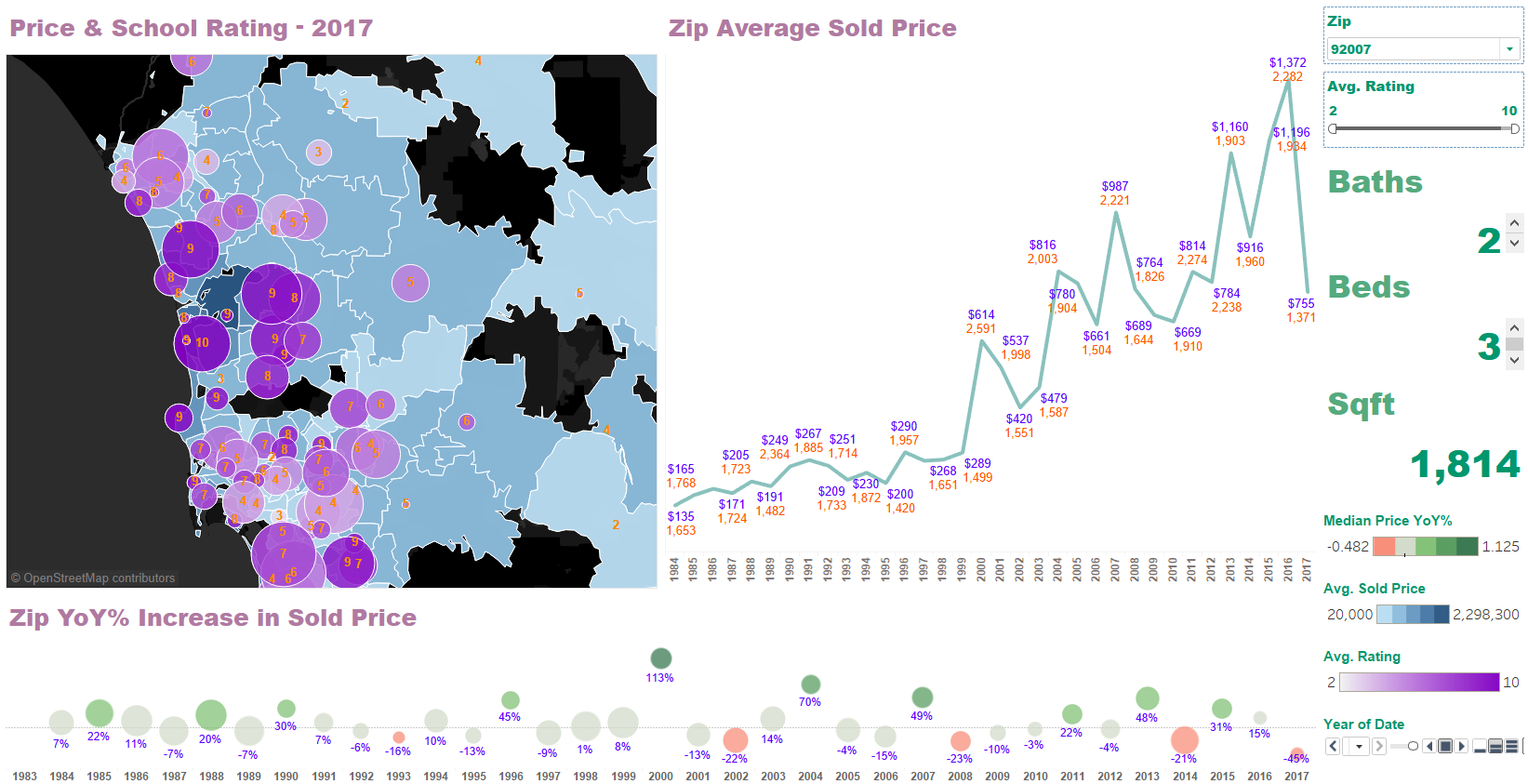
We have 4 dashboard pages planned, house searching, market statistics, EDA results and modeling for price estimation. The EDA and modeling dashboards are more like static pages that demonstrate the findings and research results from us, and house searching and statistics dashboards are provided to user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

* **Visualizations and other reportable products of your findings, e.g., reporting dashboard.**

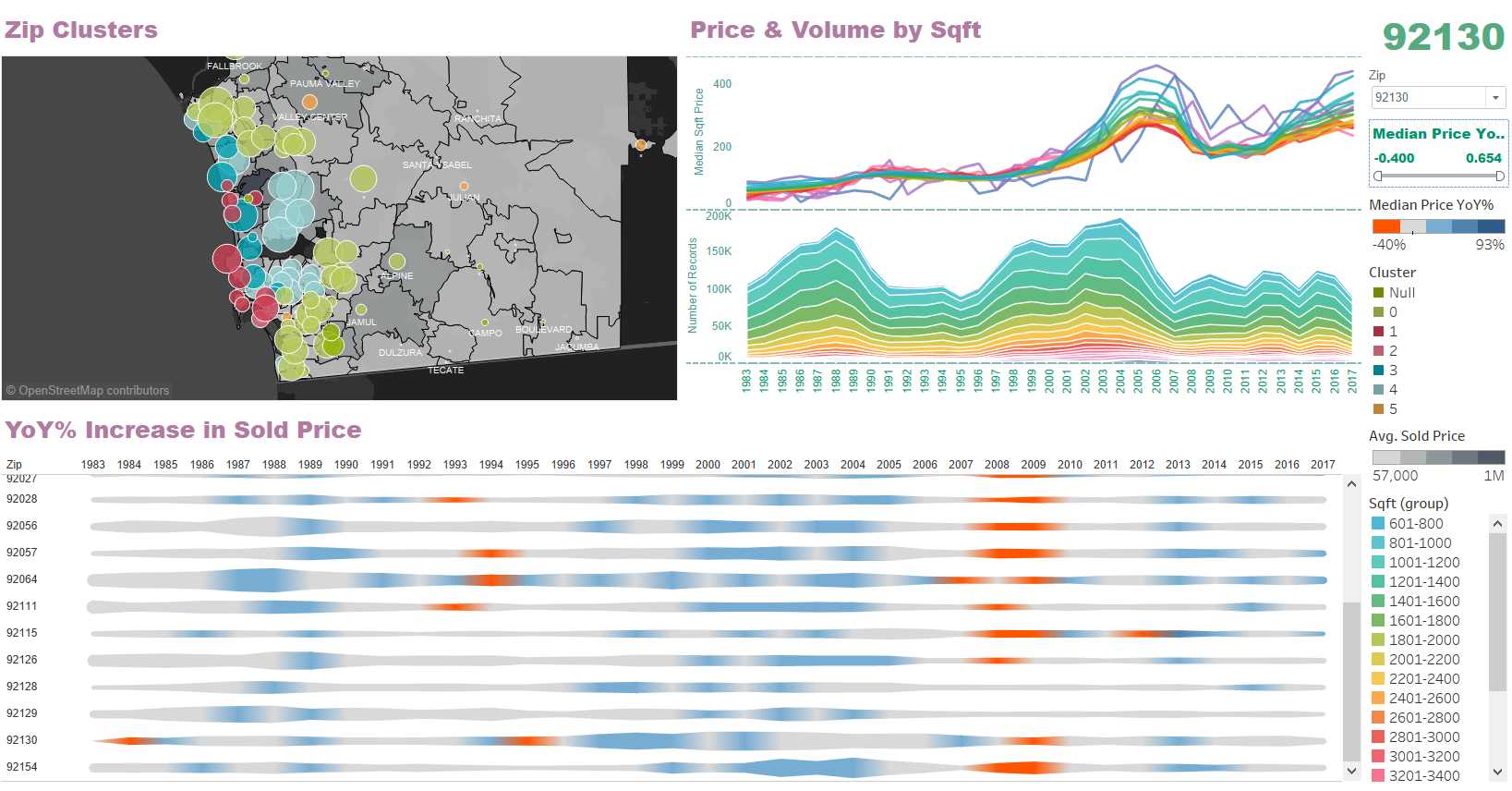
**Tableau Dashboards**

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.

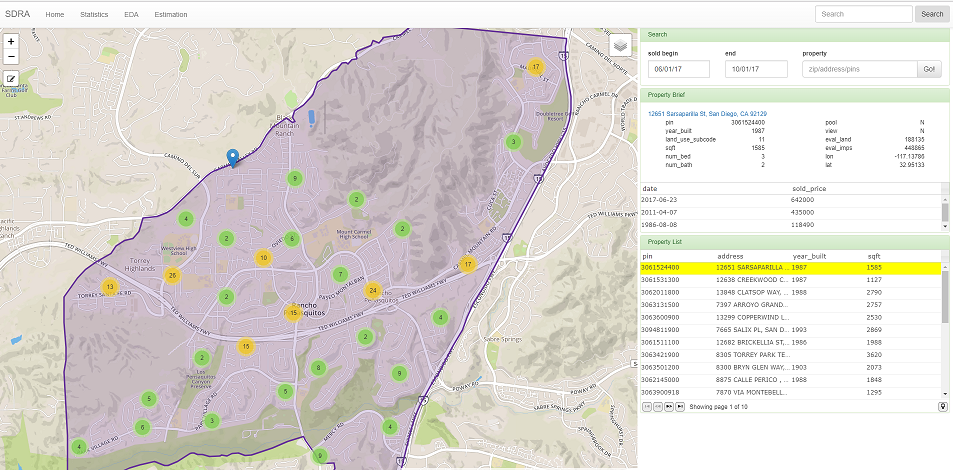
****

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.

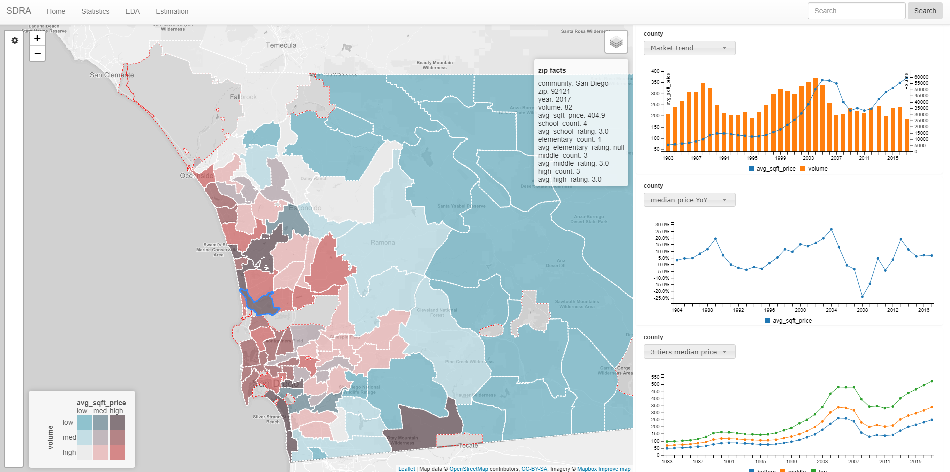
****

**Web Dashboards**

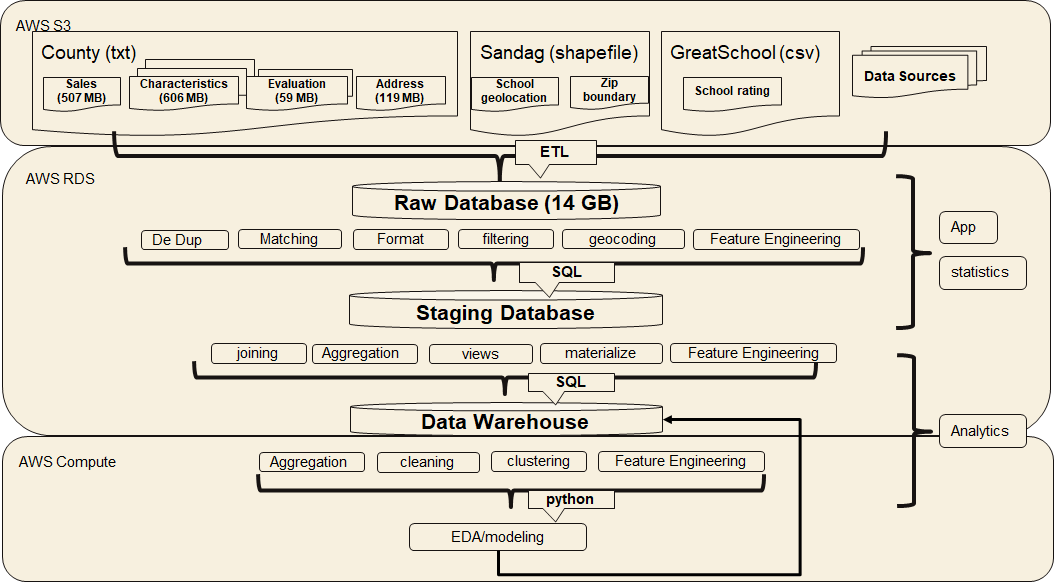
‘House Search’ page tries to mimic basic search functions that any realtor website provides. It’s also used for our debug purpose for our project. It allows various searching criteria and drawing a scope on map to select houses or transactions during a time window.



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

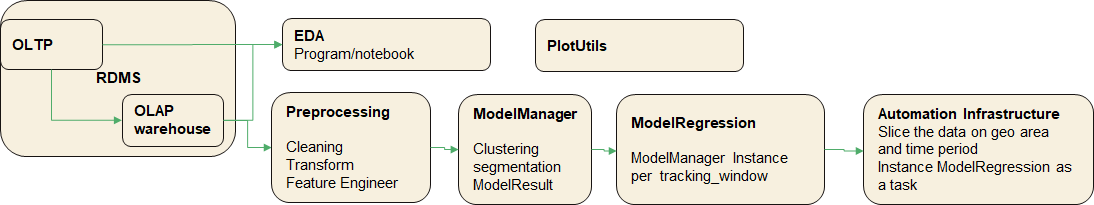


**7. Solution Architecture, Performance and Evaluation**



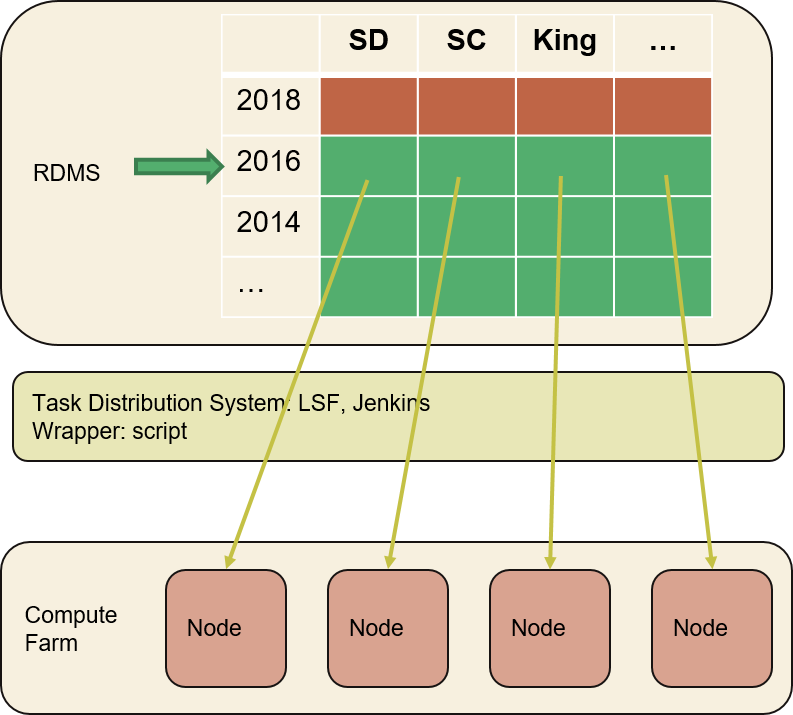
Our solution architecture tried to mimic what a realtor company could come up with. It was designed to well support both public usage like browsing, searching and internal data engineering and analytical tasks. It’s scalable regarding to data expansion, query capability and computation.

As described in previous section, the data we used are well structured and relational, SQL database is heavily relied on for storage, query and various statistical computation. Potentially, the architecture need be extended to use hybrid data storage to support images and texts, but as long as structured data is the main data source, relational database is good choice for this project.



For data processing pipeline, a significant part is done on database server which provides various builtin functions for analytical tasks like sampling, aggregation, histogram and is good in robustness, scalability and intermediate result caching. For other part, python coding with modules numpy, pandas and sklearn are sufficient for analytical and modeling tasks we targeted, and matplotlib, seaborn used for visualization.

For the computation requirement, the scalability we need for the modeling is multi-tasking rather than distributed computing. We explained in more details of modeling strategy in previous section.



For the application, 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.

**Performance measure**

Speed is not so critical for modeling and other analytics tasks of our project, as long as it’s not too slow to impact productivity, eg single model training should be finished in minutes(eg 12 months data training and 4 months data testing). And for full regression on whole historic dataset, it’s acceptable to take hours. Right now one full regression takes ~ 1 hour to finish San Diego 35 years worth data. We need make sure the node number deployed in compute farm are linear to the data amount that is scaled, and the total hours doesn’t change much.

Modeling accuracy is measured in metrics RSME normalized by sold price. Accuracy need be monitored to guarantee its accuracy doesn’t decay over time. It’s tracked along time, recorded in database and monitored through dashboard.

**Scalability**

As the benefit of building whole architecture on AWS, every part is good for scalability. Eg. RDS Database server can be scaled up to 16TB storage with 30K IOPS throughput, 64 vCPU and 488GB mem. EC2 compute nodes can be allocated and released quickly through their SDK or control dashboard. As described in previous section, a multi-task system like Jenkins can be used for task scheduling and distribution.

**Budget management**

Usage and budget planning and monitoring is another benefit that AWS provides. The usage is charged on demand and AWS building blocks all support auto scaling that dynamically adjusts the capacity required for usage. Well, the static allocation of the resources can result in lower rate which require good planning. The practical way is to use auto scaling for certain period, and understand the usage requirement, then switch to static allocation.

**8. Conclusions**

Through this project, we provides a comprehensive understanding of San Diego county housing market and a proactive way for user to explore the market by themselves based on their own need and interest. And for predictive analysis, we were modeling the price history for each property with various direct and engineered features, as well as various modeling strategy like segmentation with clustering and walk-forward validation process.

The solution architecture we built well supported both analytical tasks and web application as prototype product and potential scalability is considered and addressed.

As the analysis result, we gave guidance by visualization of the overall market trending and difference between different geographic areas, clustering by price and volume. We identified some interesting patterns in seasonality which suggests timing for buying and selling, also the appreciation difference caused by schools and other house characteristics. The estimation of house price history from our modeling gave insight of potentials for each house and is the foundation for price forecast.

**References**

* Patton MQ. Qualitative research and evaluation methods. 3rd Sage Publications; Thousand Oaks, CA: 2002.
* Sang KK, Jong HK. Statistical data preparation: management of missing values and outliers. Korean J Anesthesiol, 2017.

**Appendices**

A. DSE MAS Knowledge Applied to the Project

The skills and knowledge are summarized and categorized in the courses we’ve taken for DSE MAS program.

**Python programming**

1. python is main language for coding and OOP design principle is adopted in the whole framework.
2. Modules heavily used include numpy, pandas, sklearn, sqlalchemy, matplotlib, seaborn

**SQL Database**

1. basic SQL queries like subquery, joining, grouping & aggregation, partition etc
2. Advanced SQL like materialized view, indexing, fuzzy searching, statistics operations eg sampling, histogram
3. PostGIS extension including geocoding, geo query, data conversion from shapefile, geojson.

**Data Integration**

1. ETL for data extraction and conversion from raw data.
2. Schema matching and mapping when integrate and join multiple source.

**Probability and Statistics**

1. statistics indices like mean, variance, standard deviation, covariance, IQR
2. Distribution, hypothesis and confidence interval

**Machine Learning**

1. Unsupervised learning: PCA, clustering
2. Supervised learning: regression model

**Data Visualization**

1. Various visual encoding and idiom like line/bar/stream chart, choropleth and their effectiveness and efficiency
2. Colormap for different type of data, linked view, interaction.
3. Tools and libraries: matplotlib, tableau, d3, leaflet stacks

**B. Data and Software Archive for Reproducibility**

**Project Links**

1. **dataset**

It includes 4 parts: county, sandag, greatschool, addresses\_to\_geocode.

**Google drive:** <https://drive.google.com/drive/u/0/folders/1UctEtPBpXk399xiBc-2IIgg3R1e8glYo>

**AWS S3:** <https://s3.console.aws.amazon.com/s3/buckets/dse-cohort3-group8/data/?region=us-east-1&tab=overview>

**Dataset description**: <https://github.com/alexyanw/dse_capstone/tree/master/data>

1. **Report, notebooks and source code.**

All source code and notebooks and report are uploaded to github project, the project folder is well structured, refer to README.

<https://github.com/alexyanw/dse_capstone>

Python notebooks: <https://github.com/alexyanw/dse_capstone/tree/master/notebooks>

Python source code: <https://github.com/alexyanw/dse_capstone/tree/master/src>

Report and presentation: <https://drive.google.com/drive/u/0/folders/1SaLhY15xsh7c7zMdb9mikPP3QqZ39eKF>

**Tools Set Up**

1. **Python and Jupyter Notebook**

Follow readme on <https://github.com/alexyanw/dse_capstone> to install the necessary python module.

1. **PostgresSQL**

PostgreSQL 9.6 and below extensions installed

CREATE EXTENSION postgis;

CREATE EXTENSION fuzzystrmatch; --needed for postgis\_tiger\_geocoder

CREATE EXTENSION address\_standardizer;

CREATE EXTENSION address\_standardizer\_data\_us;

CREATE EXTENSION postgis\_topology;

CREATE EXTENSION postgis\_tiger\_geocoder;

Follow the instructions in database folder to import the data.

1. **Tableau Workbook**

Tableau workbook is located at the path below named “Final.twb”; all datasource needed are saved in the same location in a zip file called “data.zip”:

<https://drive.google.com/drive/u/1/folders/18_uR5MxTFf5_dj7giTZ65iO2TAwSE1y0>

Once opened, please connect datasource to all files that is in the data.zip container. The workbook should open up correctly.