San Diego House Price Predicting

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

### **Raw Data Sources**

Two different datasets are employed by this project, one is from redfin website and another dataset is from San Diego County data house. Both datasets include property’s characteristics and transaction history. In addition, dataset from San Diego county also include county’s appraisal values for each property by years. Because redin data is scraped from website, it is one snapshot data on 09/2017, however, county data provides multiple scans of property and their appraisal for different years.

Table 1. Data source description

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset Sources | Data Linkage | Destination Pipeline | Scripts | Data Size | Records |
| County - Property Characteristics (PAR) | https://drive.google.com/drive/u/0/folders/1kf9h\_wCDUWtVUFd2VwWGhYkE8VocDkvk | Characteristics (PAR) | SQL Server ETL packages | 1 GB/month  We will probably have 15 years worth of data load | 1.3 million/file |
| County - Sales | https://drive.google.com/drive/u/0/folders/1kf9h\_wCDUWtVUFd2VwWGhYkE8VocDkvk | Sales | SQL Server ETL packages | 30 MB/month. Data available since 1983 | 2.5 millions records total |
| County - MPR | https://drive.google.com/drive/u/0/folders/1kf9h\_wCDUWtVUFd2VwWGhYkE8VocDkvk | (Property Value) MPR | SQL Server ETL packages | 38 MB/month  We will probably have 15 years worth of data load | 1.3 millions |
| County - Full and Detailed Address | https://drive.google.com/drive/u/0/folders/1kf9h\_wCDUWtVUFd2VwWGhYkE8VocDkvk | Clean data set for property addressess | SQL Server ETL packages | 186 MB | 1.4 millions |
| Redfin | <https://drive.google.com/drive/u/0/folders/0BwxgHUbb55PqV25mRWZRcmg4QlU> | redfin\_properties/redfin\_transactions | <https://drive.google.com/drive/u/0/folders/1-kjh9d60E1Cy7DBccSVQ2lf7kye9Txyw> | 200MB txt | 722867 properties 2434534 transaction events |

### **Data Exploration, Cleaning, Wrangling and Engineering**

### **Data Exploration Summary:**

#### **Redfin data quality issues**

1. ‘redfin\_properties’ table has several duplicate columns to be removed.
2. ‘redfin\_properties’ missed several key features like schools.
3. ‘redfin\_properties’ missed values of different features (Fig 1).
4. ‘redfin\_transactions’ misses transaction records between 1980 and 1987 (Fig 2)

#### **County data quality issues**

1. ‘County\_properties’ missed geographical features: city/zip as well as other features like hoa/mello roos, which are still under investigation/pursuing of data.
2. Still work on selecting features from 5 pages 250+ property characteristics. 10 explored features show good quality though.
3. More scans of PAR/MPR are necessary for previous years.
4. ‘County\_transactions’ doesn’t have listing/pending events as redfin provides

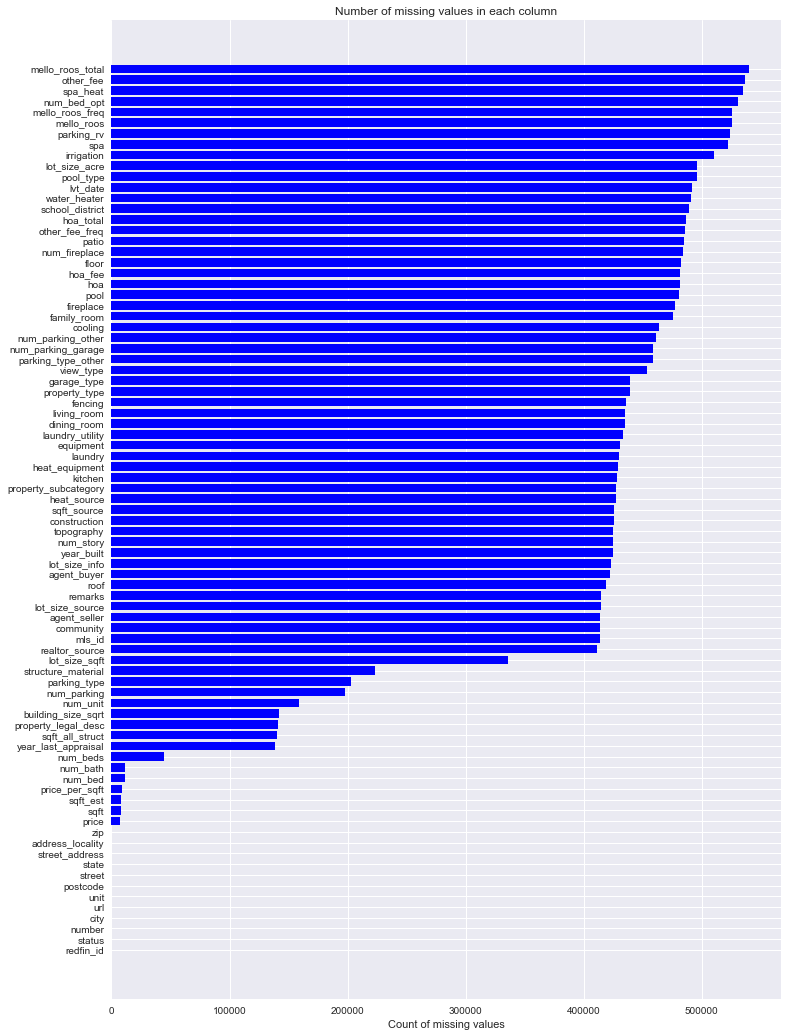
****

Fig 1. Missing Values Statistics of Different Features

#### **Transaction distribution over years**

In order to better understand house transaction history in San Diego area, average selling price and average price per square feet were plotted from 1945 to 2017 (Fig 2). From 1996 to 2006, there is a strong increasing trend were presented.

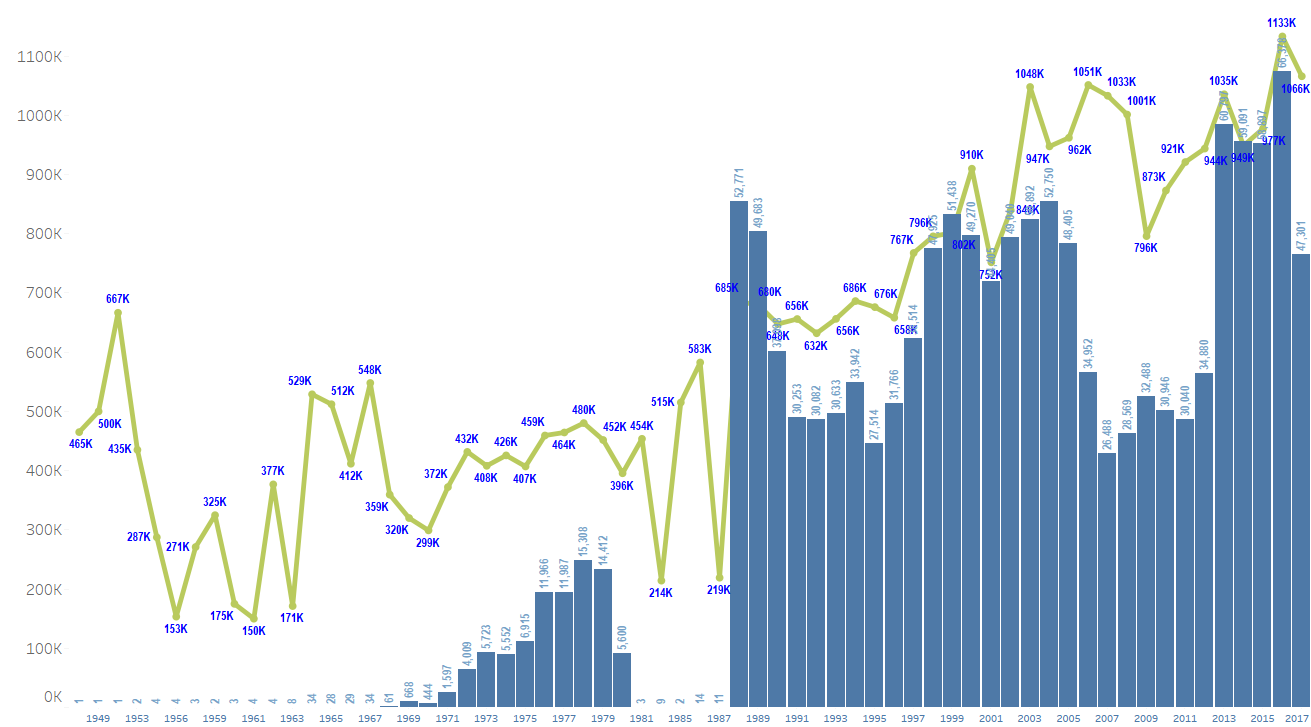


Fig 2. Historical house price time series-- record missing in 1981-1987 time frame

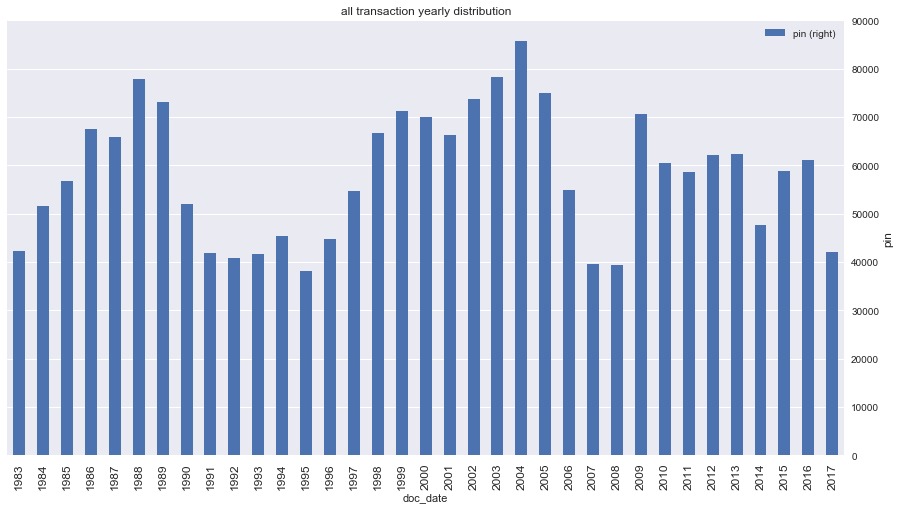


Fig 3 county transactions yearly distribution

#### **House unit price(dollar per sqft.) trend over years**

House unit price (dollar per sqft) of different cities and house unit price of different size from 1987 to 2017. During these thirty years, except the economic crisis year 2008, house price has an obvious increasing trend. Our analysis demonstrated that current house price is already exceeded last peak price value which occured in 2008 (Fig 4). From figure 2, we can derived that the price increasing rate of cheaper properties is larger than expensive properties, however, when the market is down, those properties tend to drop quicker than others.

**Sale volume and unit price viz on map**

House location is one major factor impacting house price. House price distribution map were demonstrated in Fig 5. Generally, the house close to coastal area is more expensive than the house which is far away beach zone.

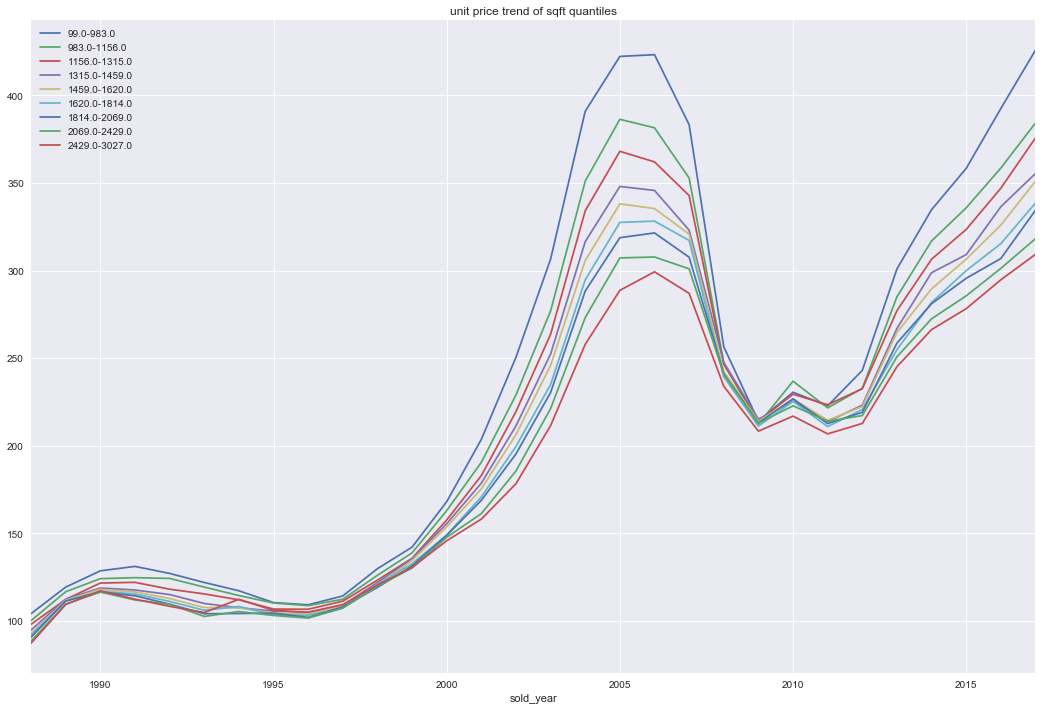
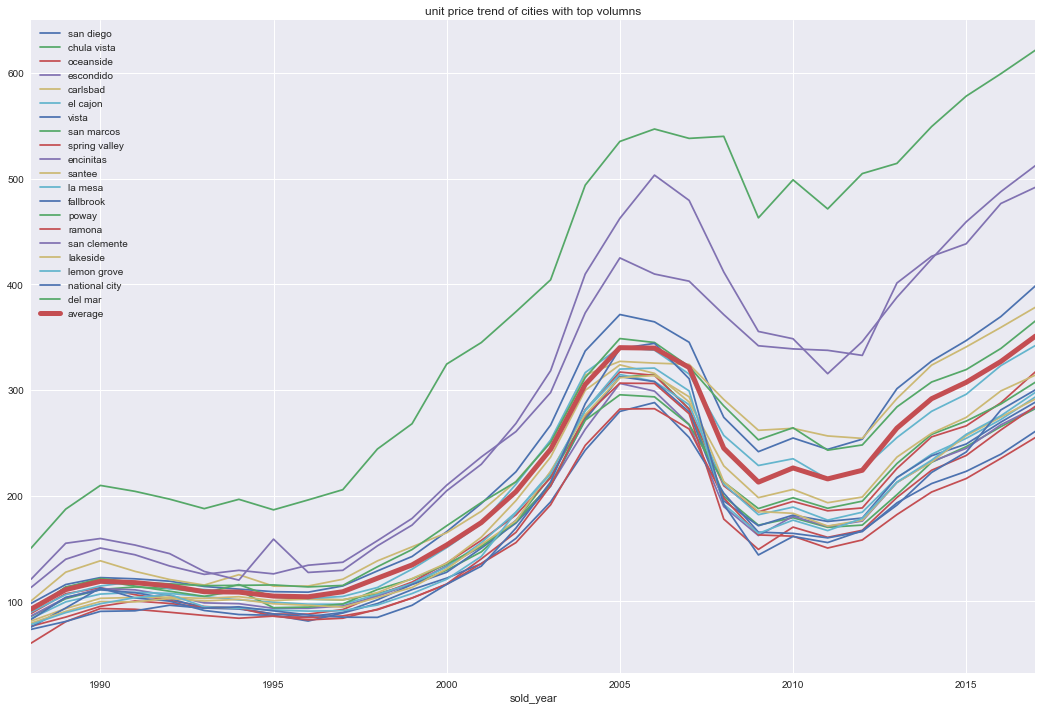


Fig 4. House unit price in different cities with top volumes (left), and house unit price of different house price (right)

#### **Feature correlation to unit price**

We analyzed correlation between unite price and other property features (Fig 6). Among 13 different property features, current land, land values, current imps, improvement values, sold year, units, addition area and improvement are all positively correlated with house unit price, however, usable square feet, bathroom, square feet, total living area and bedrooms are negatively correlated with house unit price.

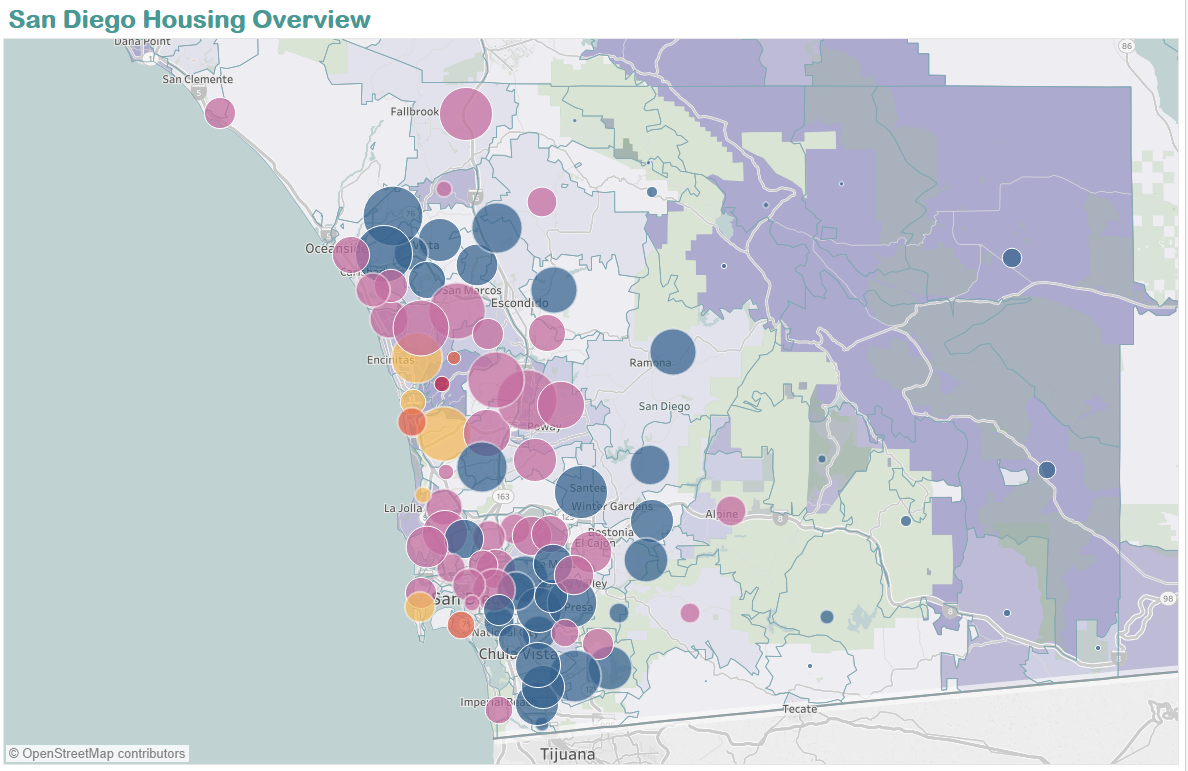


Fig 5. Map view where bubble size showing number of house sold; color shows price level (red being as the most expensive; blue being as the least expensive; purple being at middle)

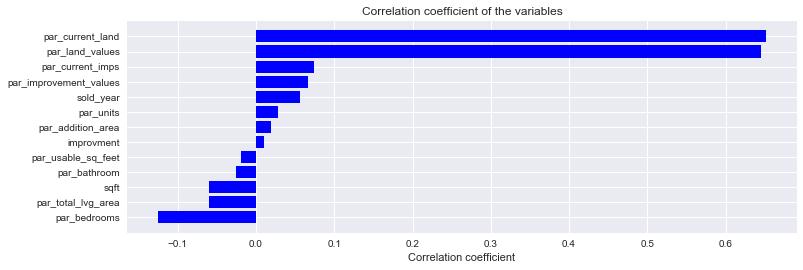


Fig 6. Correlation between property attributes and unit price

### Data Preprocessing Approach:

Data will be processed from the source to staging area and then into a Data Warehouse.

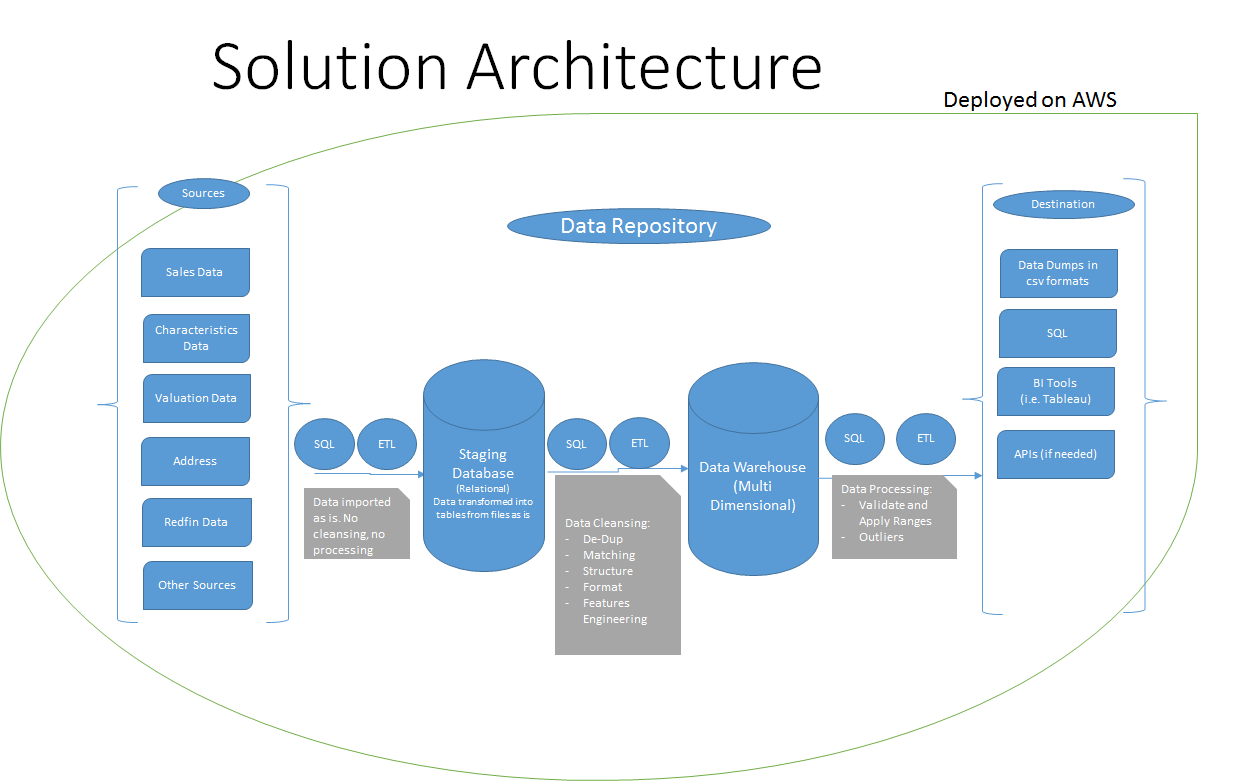
1. Source to staging: Transformation will move the data from flat files or APIs as is to the staging area. No cleaning or processing.
2. From staging to Data Warehouse, data will be cleansed, processed and reconstructed. Outliers will be removed and filtration will be done as part of querying the Data Warehouse.

#### **Approach for storing processed and/or integrated data**

1. Data will be initially loaded into staging area from all sources as is.
2. Data will be processed from the staging area into the DW. The DW will follow the dimensional modeling for each querying and better performance.
3. Data Dictionary will be created to illustrate definition, processing, rules and types of data at all stages.

#### **ETL Process:**

1. Duplicates: The ETLs will remove duplicates and decide about the master record(s) to transform based on configurations and rules.
2. Matching: The ETLs will match records from the sources and make sure they are consistent.
3. Missing value:
   1. The ETLs will assign the missing values based on predefined rules.
   2. For some others, they will be logged for manual reviewing and decisioning.
4. Lookup tables: Predefined values will be defined as lookup tables and will have references through PK/FK representation.



#### **Features Engineering:**

### Several engineered features need be created, ongoing effort as in <https://docs.google.com/spreadsheets/d/1PUASGQKurgiyRMCb1oxgO51ZZ_OQ9u3YErJ3N2RYQLU/edit#gid=331228961>

1. Select usable attributes by checking missing/invalid value percentage, then rank the attributes by the correlation with target variable (unit price etc), our understanding as well as domain knowledge from expert.
2. Each dataset of redfin and county data provides certain valuable features, we try to merge them with common keys(eg parcel id or address). unfinished

#### **Summary of feature sets**

In order to build static model to simulate house unit price different features are derived from our data sources.

1. Direct features(Examples ...)

sqft, num\_bed,num\_bath, city, zip, year\_built, school district, lot\_size, mello\_roos, hoa, num\_parking, num\_story, property\_type, current\_land, current\_imps

1. Engineered features

listing\_days, times\_relisting, street\_supply, street\_hotness, dist\_school, avg\_price\_street

#### **Approach for Data Access**

#### The following are the means to access the data:

1. Data dump into common format such as .csv.
2. Direct query to the DW using SQL or Python.
3. BI Tool such as Tableau.
4. API whenever is needed.

#### **Data Pipeline**

1. More scans of county MPR/PAR data are expected, need be preprocessed and integrated to central database as well as local copy release.
2. Other dataset like census etc may be needed during the project.

#### **Set up for your data environment**

#### A central database will be used to host integrated data

* 1. AWS RDS will be used as host.
  2. Try to refresh the central database whenever we integrate any new data.

1. Local copy
   1. Flat csv files are dumped after initial processing with ETL and copied to everyone, SQL importing script is delivered together.
   2. Everyone can load csv to either local database or tableau.

### Bullets for each team member’s individual contributions in Step 2

Wen: data exploration on redfin data, investigated geocoding with PostGIS.

Mengting: data exploration, report writing

Salah:

* Obtained additional historial sales and address dataset.
* Store the data into staging table.
* Run some queries.
* Participating on the report - Data Processing Approach Section.

Xia: writing capstone report, data exploratory

### Any major updates to Step 1 as a result of data pipeline step

1. Design data processing and storage scheme
2. House property and house transaction Data exploratory
3. Got second data source (house property and house transaction data from San Diego)
4. Synthesis and understand all the data sources