

George Drakulic

[gdrakulic@gmail.com](mailto:gdrakulic@gmail.com)

<https://www.linkedin.com/in/drakulic>

<https://github.com/drakulic/Capstone>

# REAL ESTATE ANALYSIS

SAN FRANCISCO

DISTRICTS 7, 8, AND 10

2013 NOVEMBER 12 – 2016 NOVEMBER 08

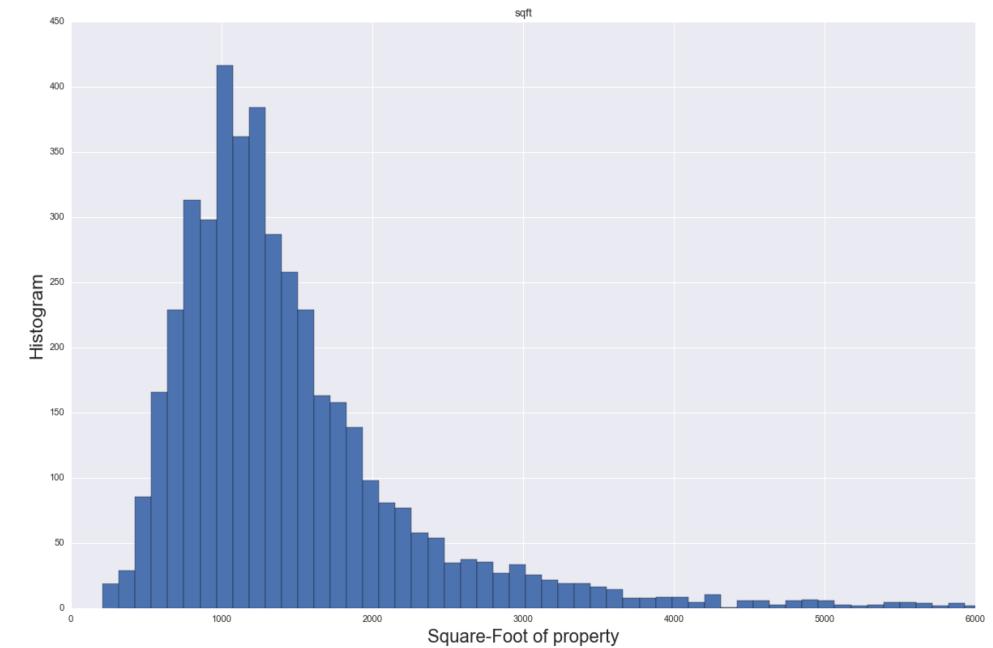
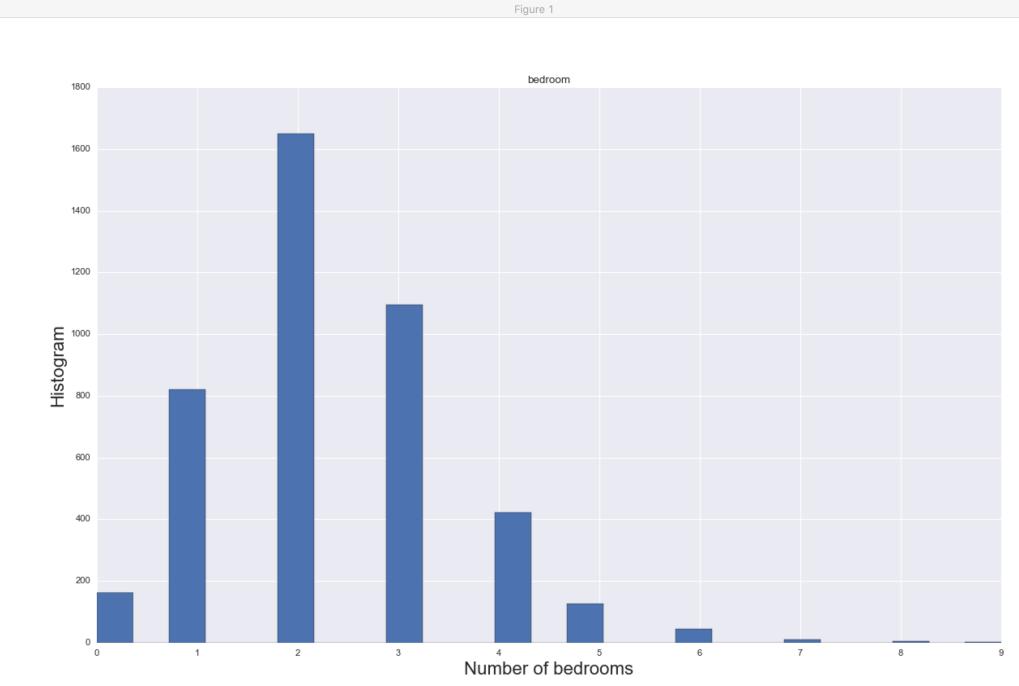
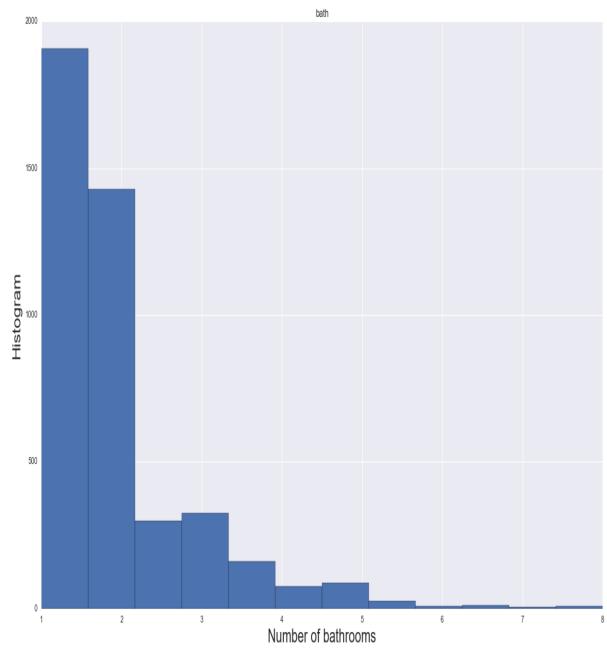
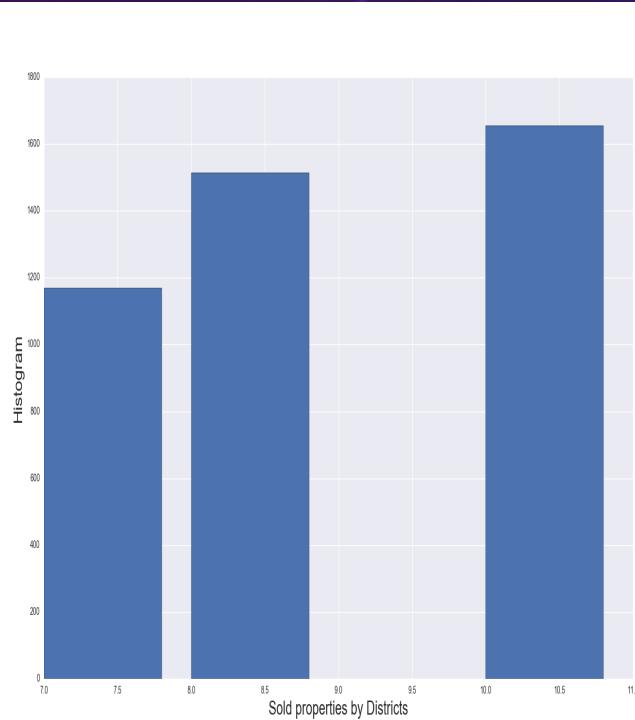
# DISTRICTS 7, 8, AND 10

4339 DATA POINTS, 35 FEATURES

|                   |                      |
|-------------------|----------------------|
| Bedroom           | Financial District   |
| Bath              | Hunters Point        |
| Parking           | Little Hollywood     |
| Sqft              | Marina               |
| home_own_ass      | Mission Terrace      |
| day_on_market     | Nob Hill             |
| single_f_h        | North Beach          |
| Condo             | North Waterfront     |
| dist_no           | Outer Mission        |
| sold_year         | Pacific Heights      |
| sold_month        | Portola              |
| Bayview Heights   | Presidio Heights     |
| Bayview           | Russian Hill         |
| Candlestick Point | Silver Terrace       |
| Cow Hollow        | Telegraph Hill       |
| Crocker Amazon    | Van Ness/Civic Cente |
| Downtown          | Visitacion Valley    |
| Excelsior         |                      |

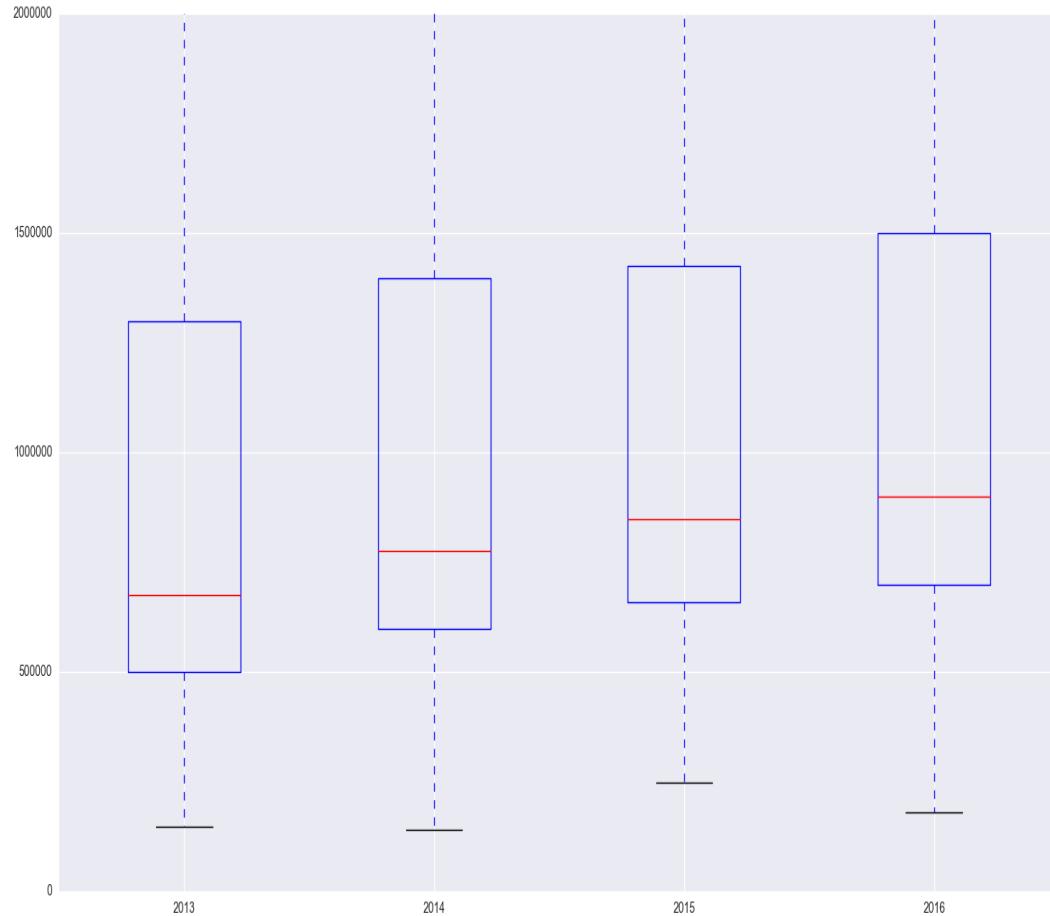


# EXPLORATORY DATA ANALYSIS

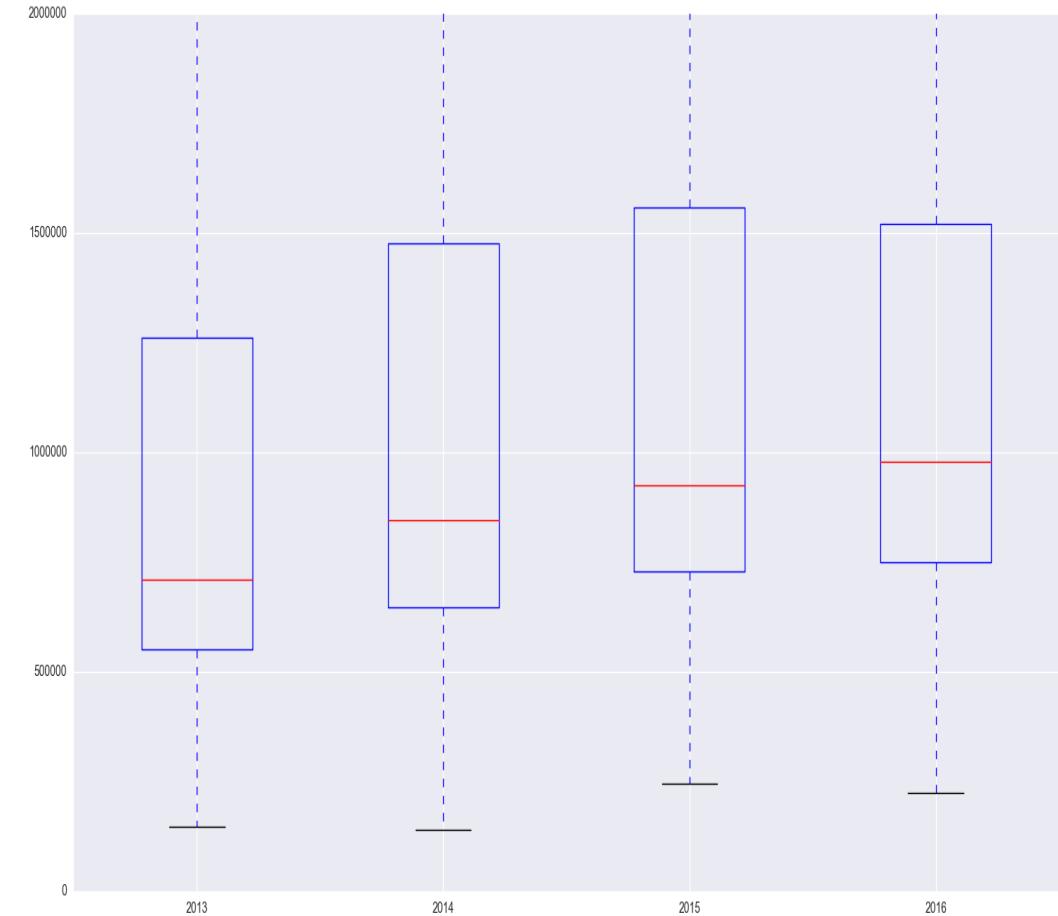


# EXPLORATORY DATA ANALYSIS

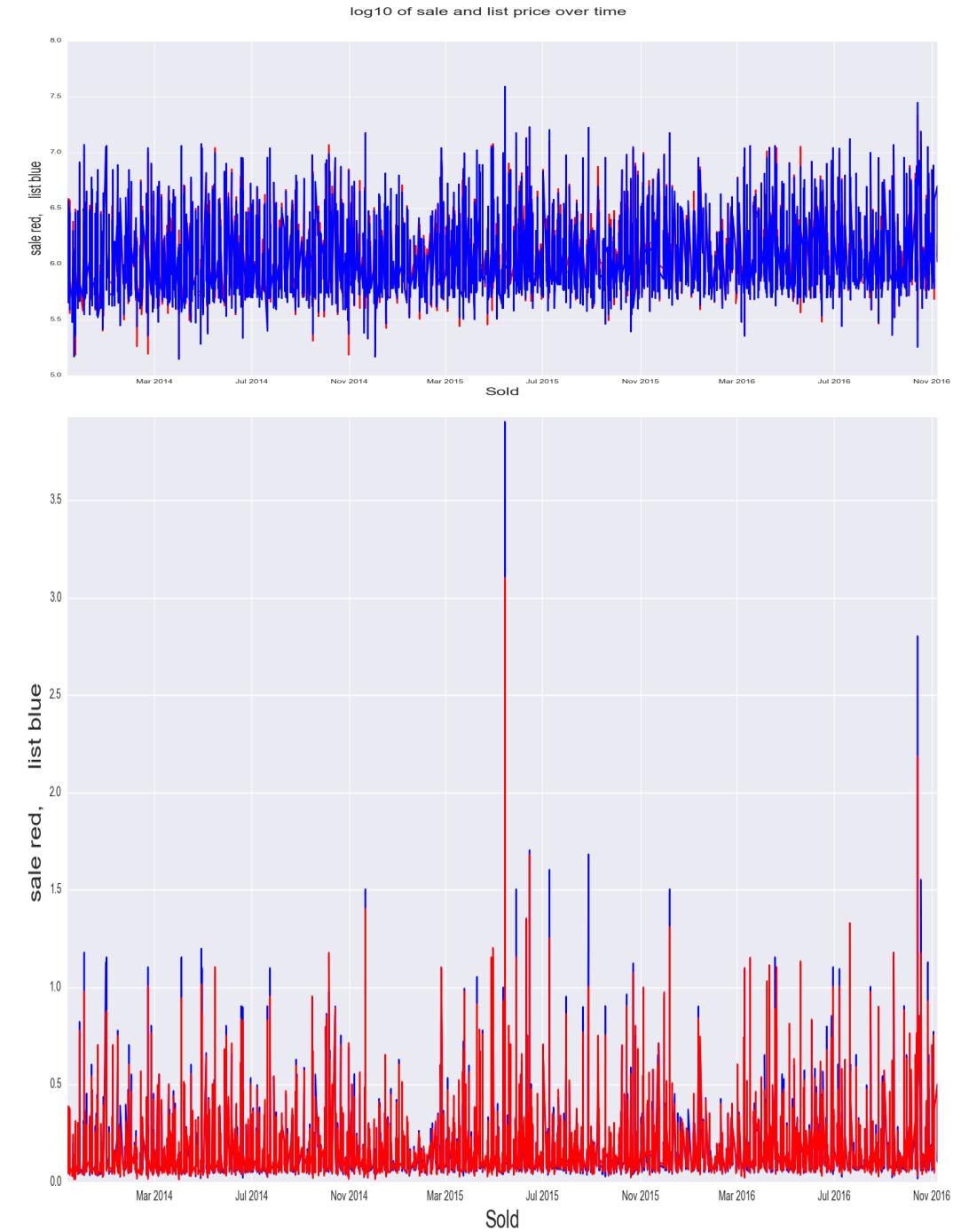
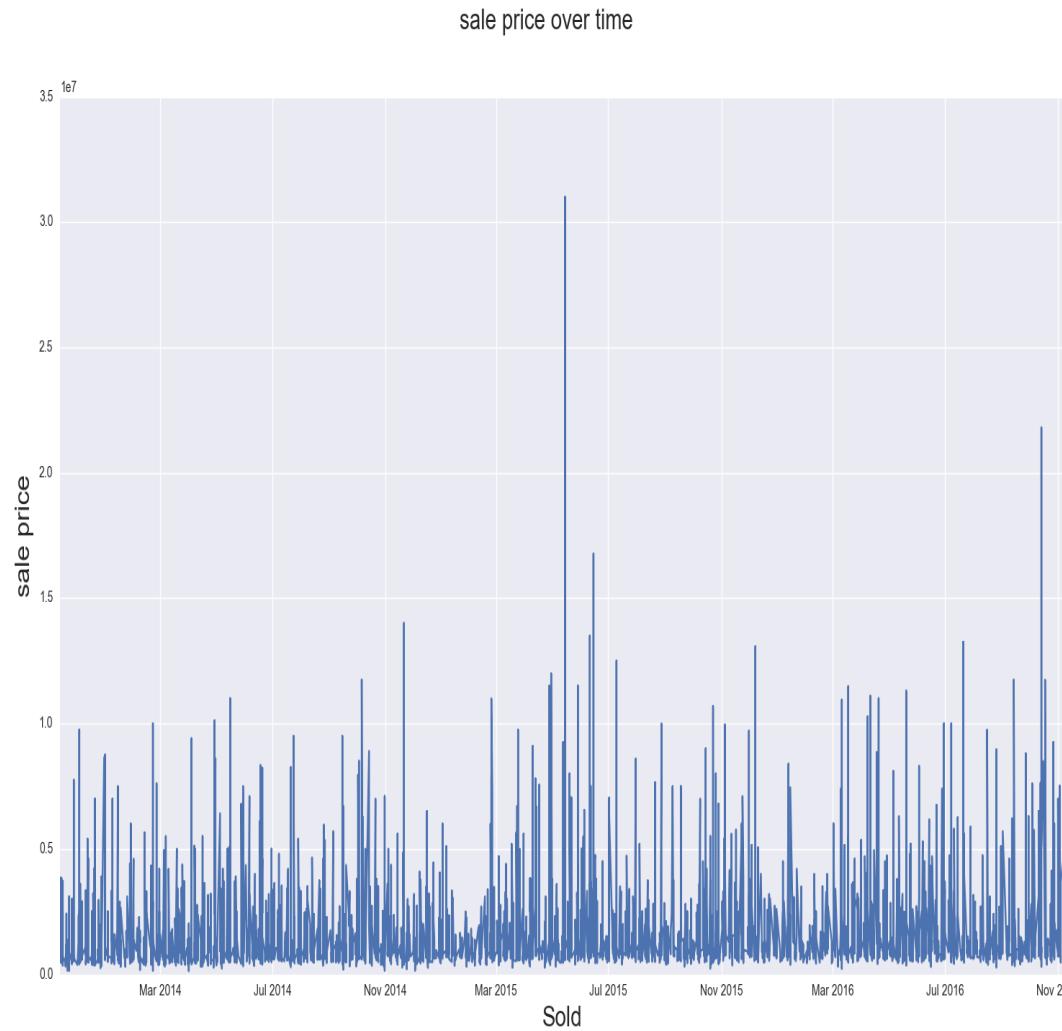
List price by year



Sale price by year

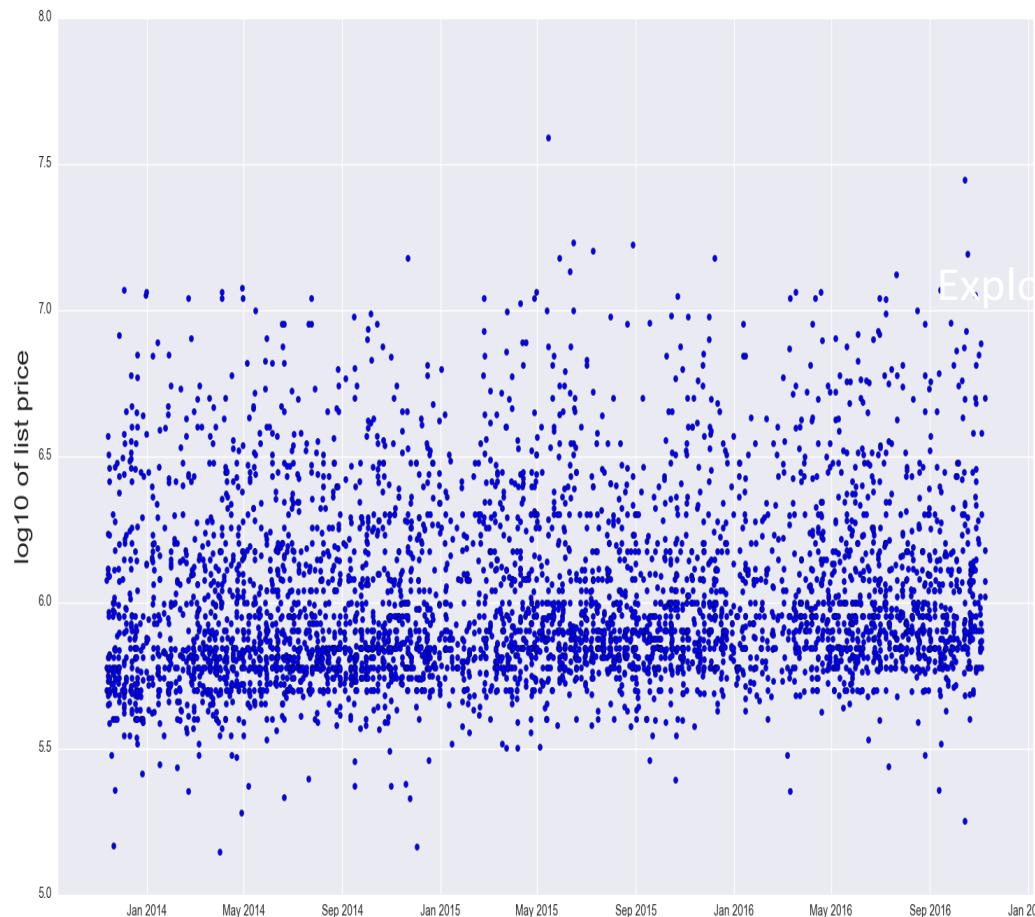


# EXPLORATORY DATA ANALYSIS

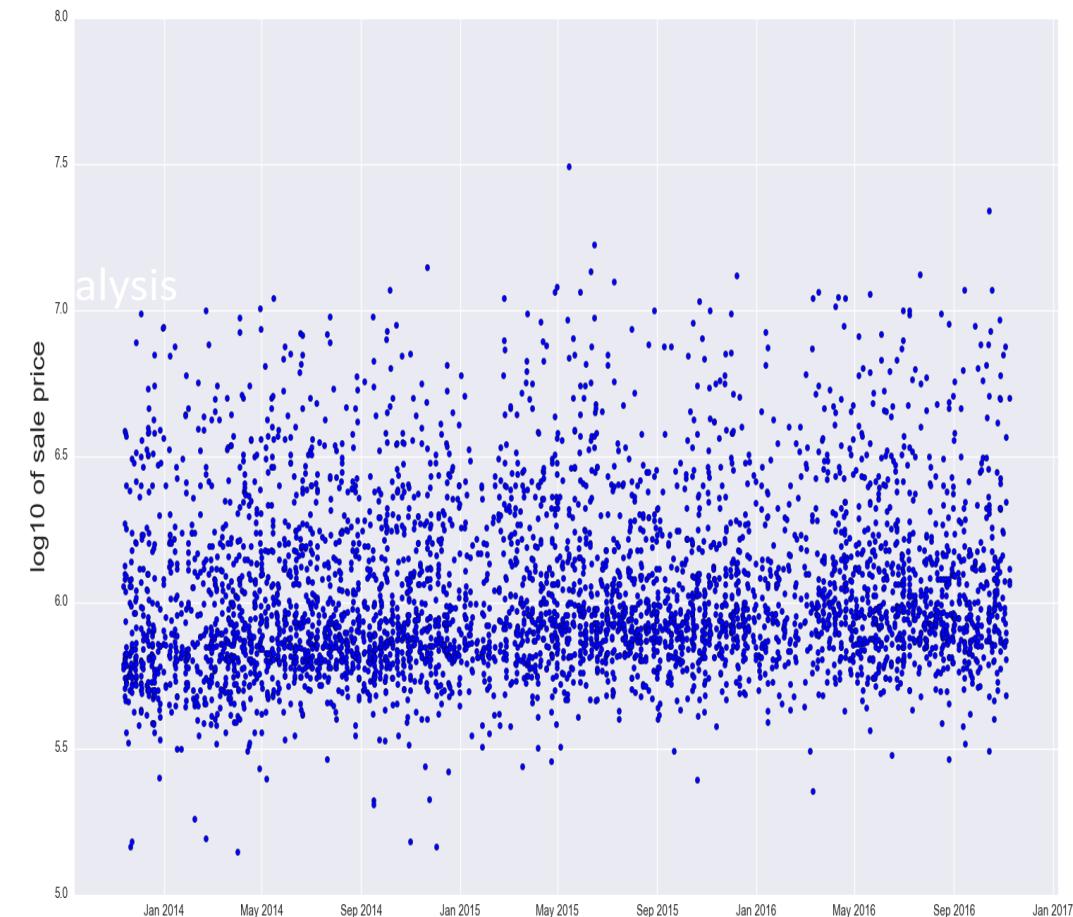


# EXPLORATORY DATA ANALYSIS

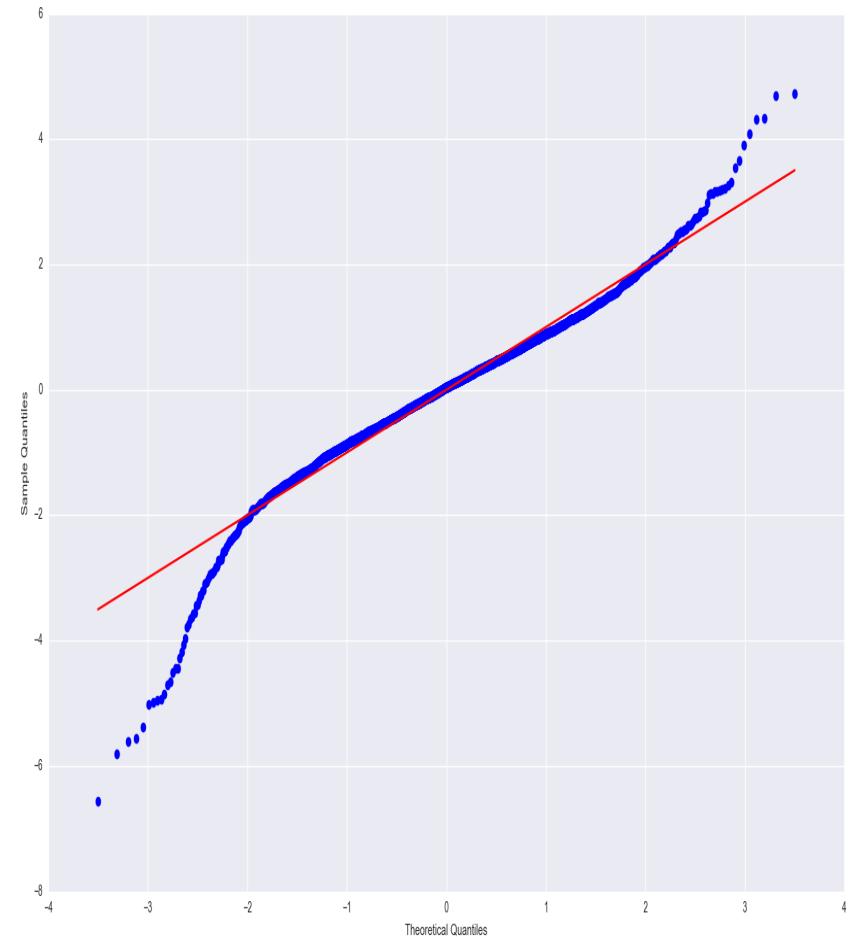
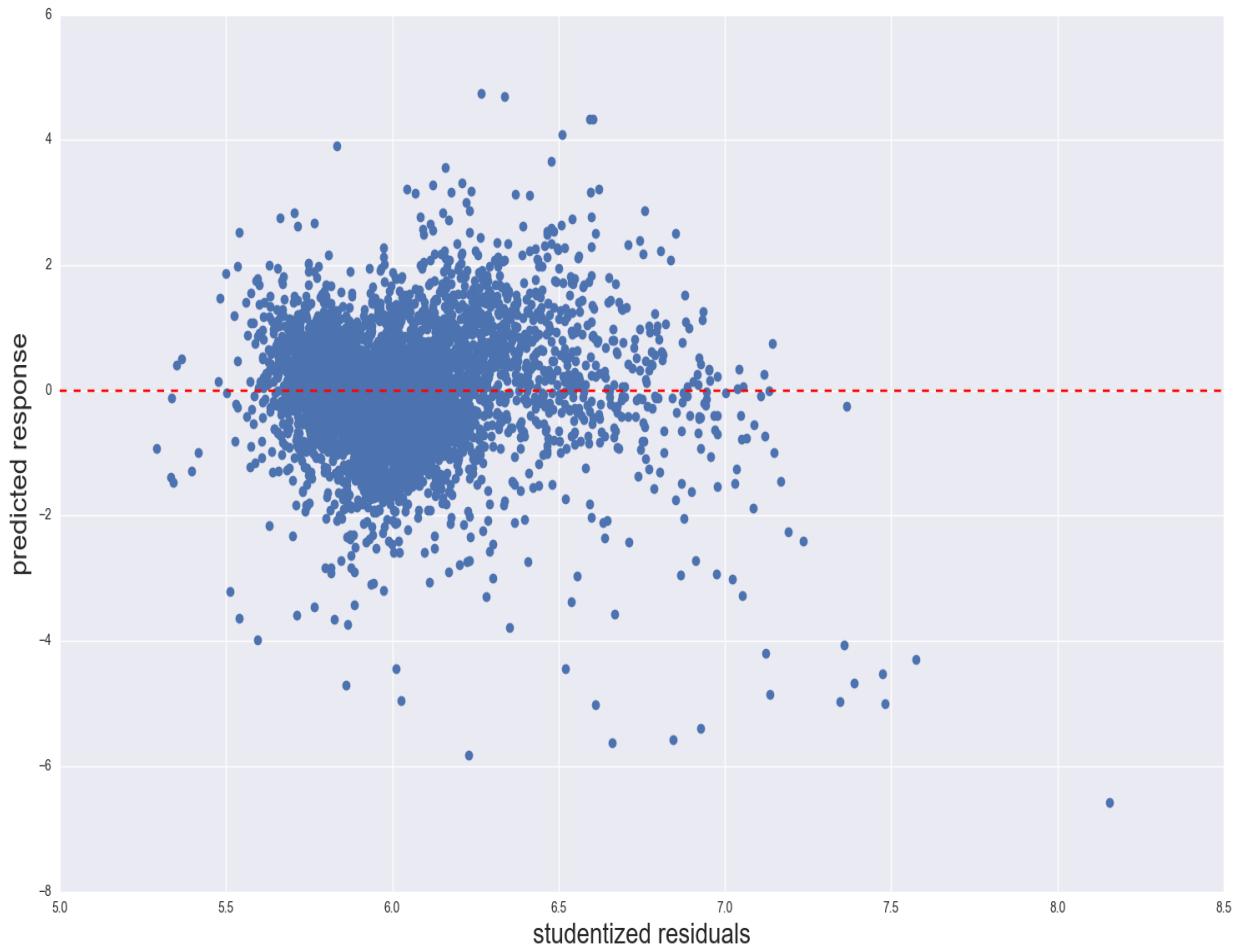
list price over time



sale price over time



# EXPLORATORY DATA ANALYSIS



# LINEAR REGRESSION WITH LIST PRICE

TRAIN VARIANCE SCORE: 0.98

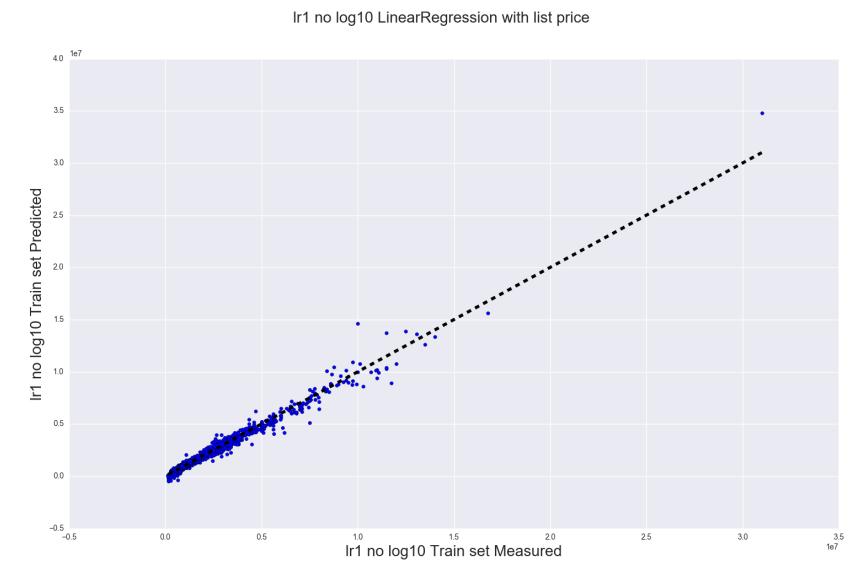
TEST VARIANCE SCORE: 0.96

VALIDATION VARIANCE SCORE: 0.98

model with year month columns, no log10:

## OLS Regression Results

|                   |                  |                    |           |
|-------------------|------------------|--------------------|-----------|
| Dep. Variable:    | sale_price       | R-squared:         | 0.978     |
| Model:            | OLS              | Adj. R-squared:    | 0.978     |
| Method:           | Least Squares    | F-statistic:       | 5925.     |
| Date:             | Tue, 29 Nov 2016 | Prob (F-statistic) | 0.00      |
| Time:             | 12:25:23         | Log-Likelihood     | -59955.   |
| No. Observations: | 4339             | AIC:               | 1.200e+05 |
| Df Residuals:     | 4305             | BIC:               | 1.202e+05 |
| Df Model:         | 33               |                    |           |



Cross-Validated Predictions

11/29/16

8

## LINEAR REGRESSION WITH LIST PRICE WITH LOG10 OF PRICES

TRAIN VARIANCE SCORE: 0.98

TEST VARIANCE SCORE: 0.96

VALIDATION VARIANCE SCORE: 0.98

model with log10 list price:

### OLS Regression Results

| Dep. Variable:    | np.log10(sale_price) | R-squared:         | 0.980      |
|-------------------|----------------------|--------------------|------------|
| Model:            | OLS                  | Adj. R-squared:    | 0.980      |
| Method:           | Least Squares        | F-statistic:       | 6302.      |
| Date:             | Tue, 29 Nov 2016     | Prob (F-statistic) | 0.00       |
| Time:             | 12:25:23             | Log-Likelihood     | 7569.7     |
| No. Observations: | 4339                 | AIC:               | -1.507e+04 |
| Df Residuals:     | 4305                 | BIC:               | -1.485e+04 |
| Df Model:         | 33                   |                    |            |

### Cross-Validated Predictions

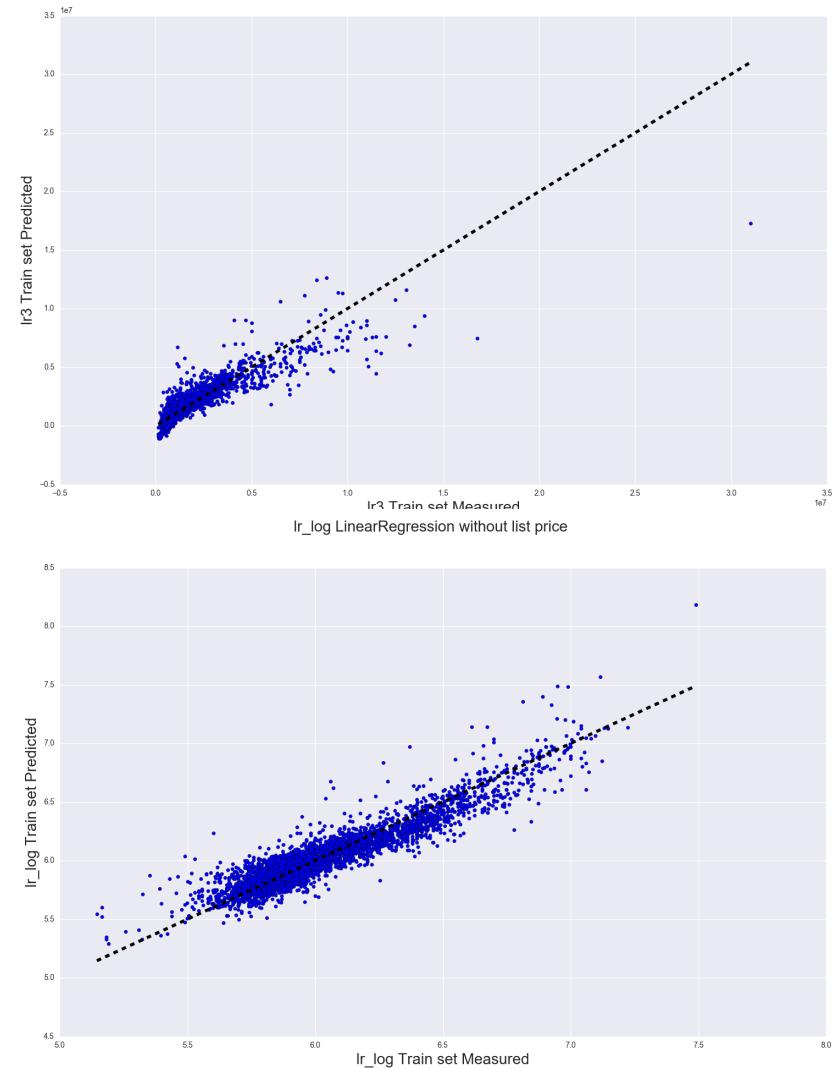
# LINEAR REGRESSION WITHOUT LIST PRICE

model with log10 list price:

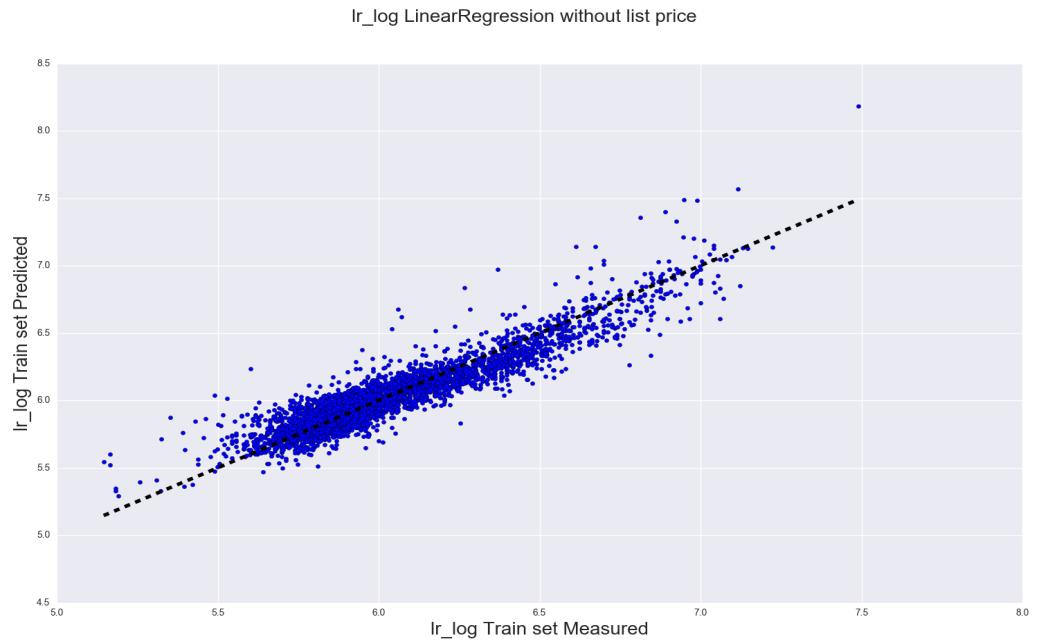
## OLS Regression Results

|                                     |                    |        |
|-------------------------------------|--------------------|--------|
| Dep. Variable: np.log10(sale_price) | R-squared:         | 0.867  |
| Model: OLS                          | Adj. R-squared:    | 0.866  |
| Method: Least Squares               | F-statistic:       | 880.6  |
| Date: Tue, 29 Nov 2016              | Prob (F-statistic) | 0.00   |
| Time: 12:25:23                      | Log-Likelihood     | 3496.9 |
| No. Observations: 4339              | AIC: -6928.        |        |
| Df Residuals: 4305                  | BIC: -6717.        |        |
| Df Model: 33                        |                    |        |

## Cross-Validated Predictions



# LINEAR REGRESSION WITHOUT LIST PRICE, WITH LOG10 SALE PRICE



| Date       | Square-Foot | Sale-price | Predicted price |
|------------|-------------|------------|-----------------|
| 2016-08-16 | 1050        | 980,000    | 1,113,973.86    |
| 2016-08-16 | 1312        | 850,000    | 619,429.93      |
| 2016-08-16 | 700         | 917,500    | 1,207,900.75    |
| 2016-08-16 | 760         | 670,000    | 602,939.08      |
| 2016-08-16 | 4617        | 9,750,000  | 5,256,578.14    |
| 2016-08-17 | 550         | 699,000    | 951,734.09      |
| 2016-08-17 | 527         | 385,000    | 338,296.01      |
| 2016-08-17 | 1000        | 775,000    | 639,240.18      |
| 2016-08-17 | 1210        | 1,050,000  | 824,167.77      |
| 2016-08-17 | 975         | 790,000    | 625,026.51      |
| 2016-08-17 | 1140        | 1,100,000  | 973,377.38      |
| 2016-08-18 | 1290        | 755,000    | 883,667.87      |
| 2016-08-18 | 1841        | 1,850,000  | 2,044,310.57    |
| 2016-08-18 | 1350        | 1,475,000  | 1,382,372.23    |
| 2016-08-18 | 460         | 760,000    | 820,469.61      |
| 2016-08-19 | 825         | 595,000    | 530,407.16      |
| 2016-08-19 | 1300        | 925,000    | 716,916.79      |

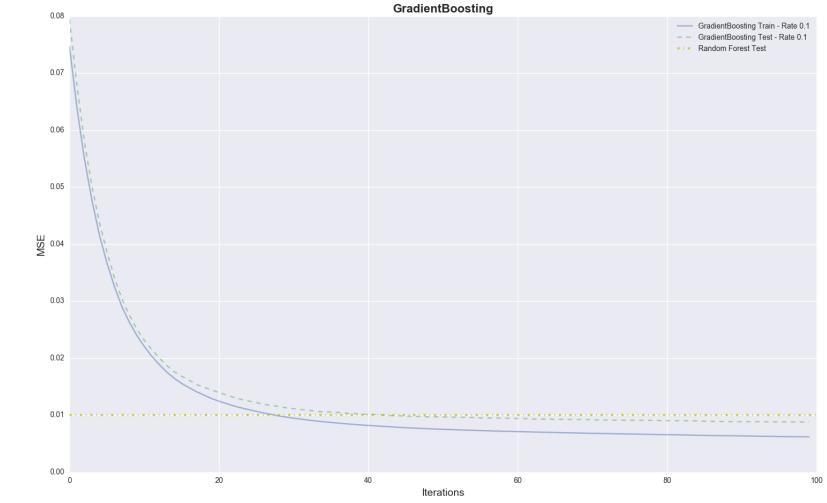
Train Variance score: 0.87  
Test Variance score: 0.87

11/30/16

11

# CHOOSING THE FINAL MODEL:

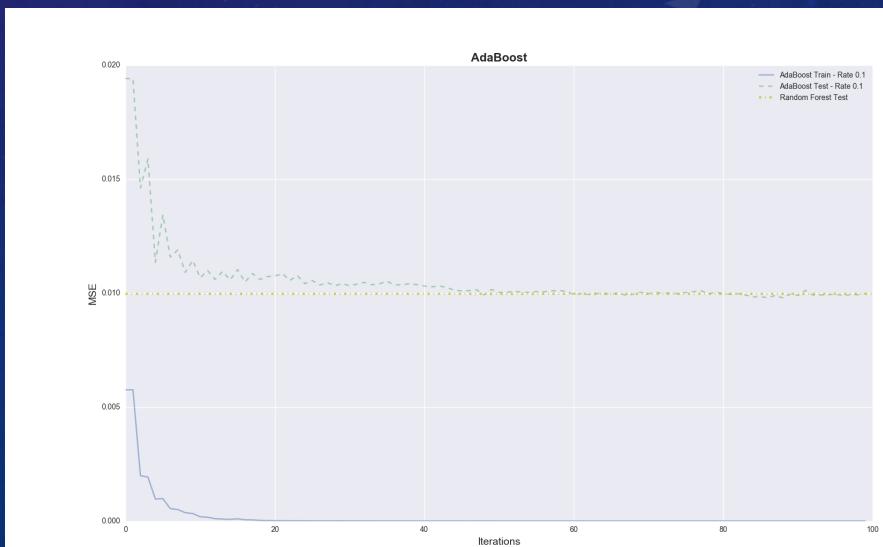
- Random Forest Regressor
- Gradient Boosting Regressor
- Ada Boosting Regressor



RandomForestRegressor Train CV | MSE: 0.008 | R2: 0.904

GradientBoostingRegressor Train CV | MSE: 0.008 | R2: 0.906

AdaBoostRegressor Train CV | MSE: 0.009 | R2: 0.901



# CHOOSING THE FINAL MODEL: GRID SEARCH FOR BEST ESTIMATORS

Grid search RandomForestRegressor

Fitting 3 folds for each of 324 candidates, totalling 972

Done 52 tasks | elapsed: 2.1s

Done 352 tasks | elapsed: 14.0s

Done 657 tasks | elapsed: 28.2s

Done 972 out of 972 | elapsed: 45.5s finished

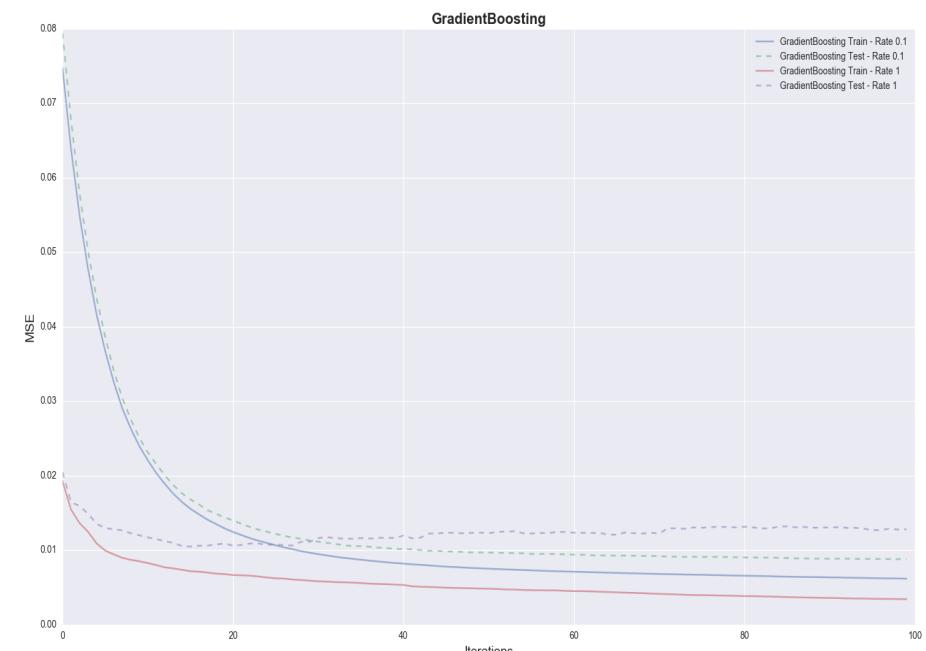
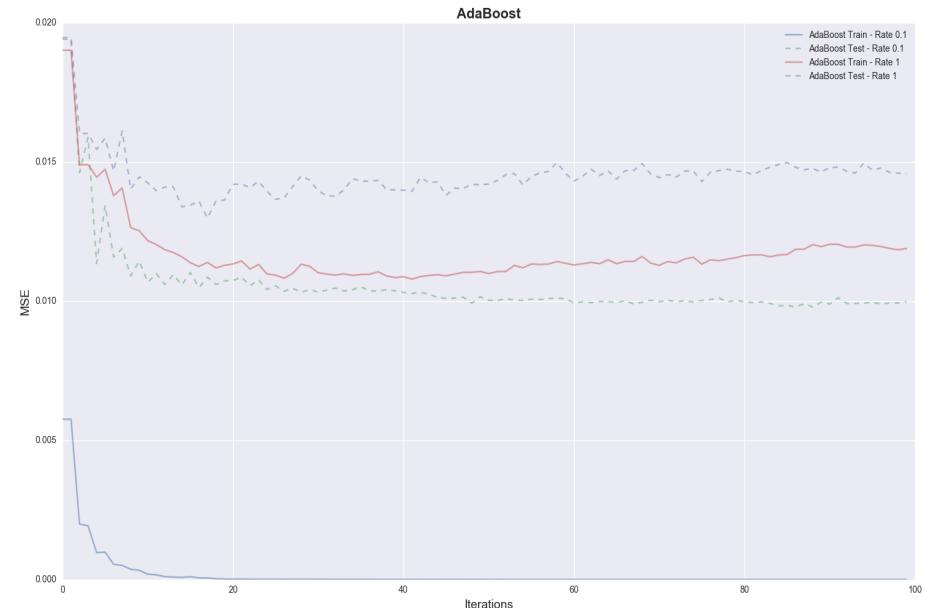
Grid search GradientBoostingRegressor

Fitting 3 folds for each of 96 candidates, totalling 288

Done 34 tasks | elapsed: 6.0s

Done 184 tasks | elapsed: 44.4s

Done 288 out of 288 | elapsed: 1.2min finished



# FINAL MODEL: GRADIENT BOOSTING REGRESSOR

- GradientBoostingRegressor Train CV | MSE: 0.007 | R2: 0.919
- RandomForestRegressor Train CV | MSE: 0.007 | R2: 0.914
- Gradient Boost Test MSE: 0.00806594315088
- Gradient Boost Test R2: 0.90601555473
- Random Forest Test MSE: 0.00834159580026
- Random Forest Test R2: 0.900919571981
- Gradient Boost Test MSE: 0.00806594315088
- Gradient Boost Test R2: 0.90601555473

GradientBoostingRegressor best estimator:

```
GradientBoostingRegressor(alpha=0.9, criterion='friedman_mse', init=None, learning_rate=0.02, loss='ls', max_depth=6,  
max_features=0.3, max_leaf_nodes=None, min_impurity_split=1e-07, min_samples_leaf=5, min_samples_split=2,  
min_weight_fraction_leaf=0.0, n_estimators=500, presort='auto', random_state=1, subsample=1.0, verbose=0,  
warm_start=False)
```

# GRADIENT BOOSTING REGRESSOR

Final model without list price, GradientBoostingRegressor

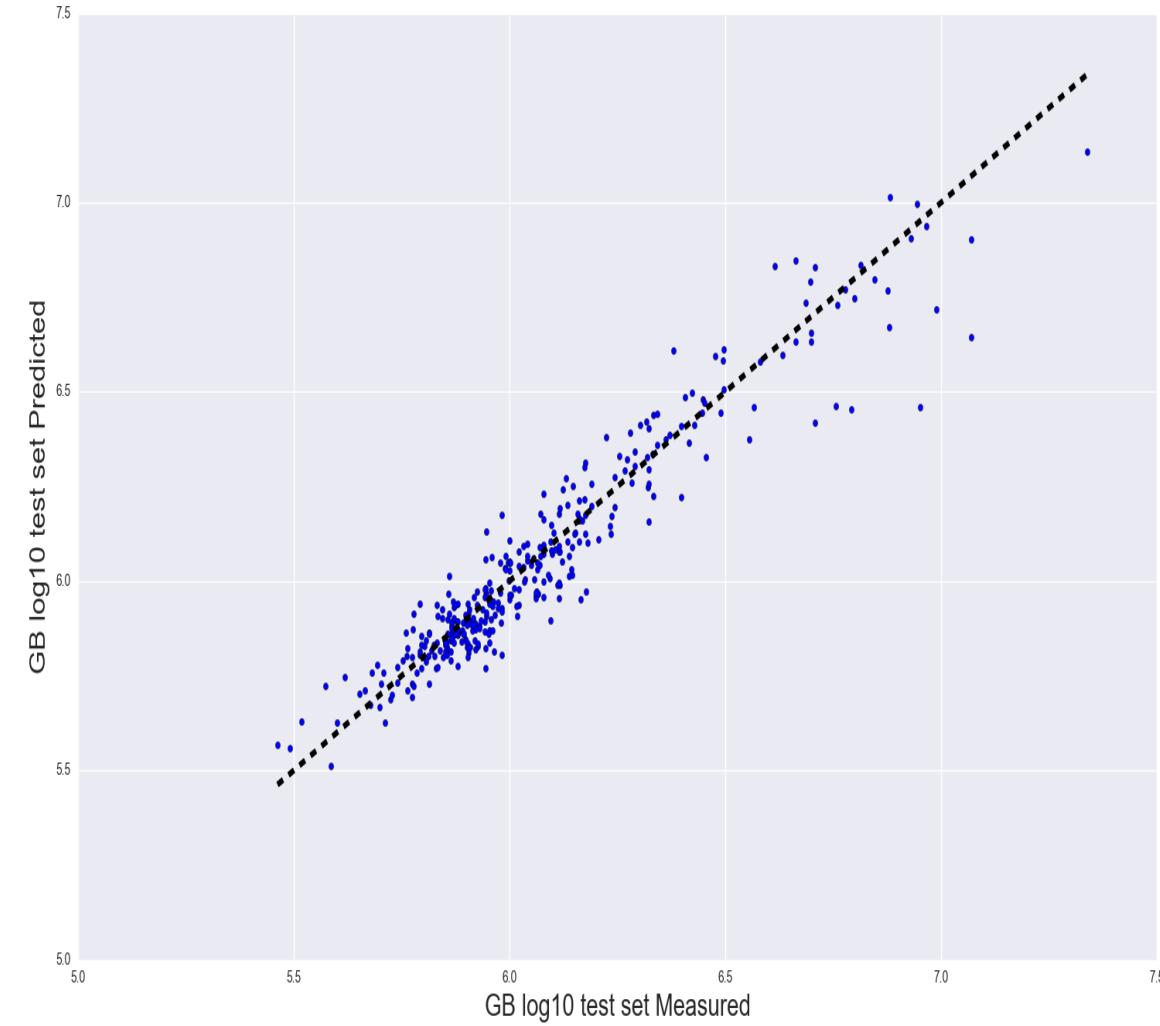
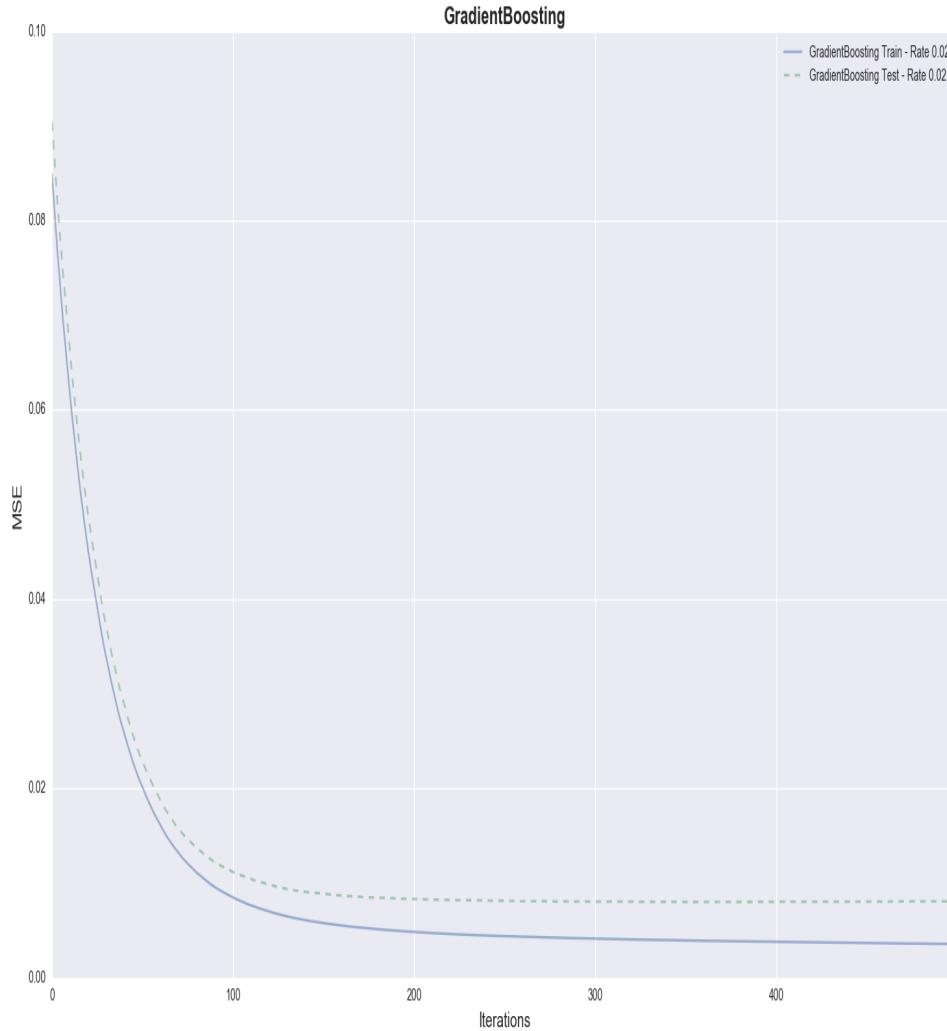
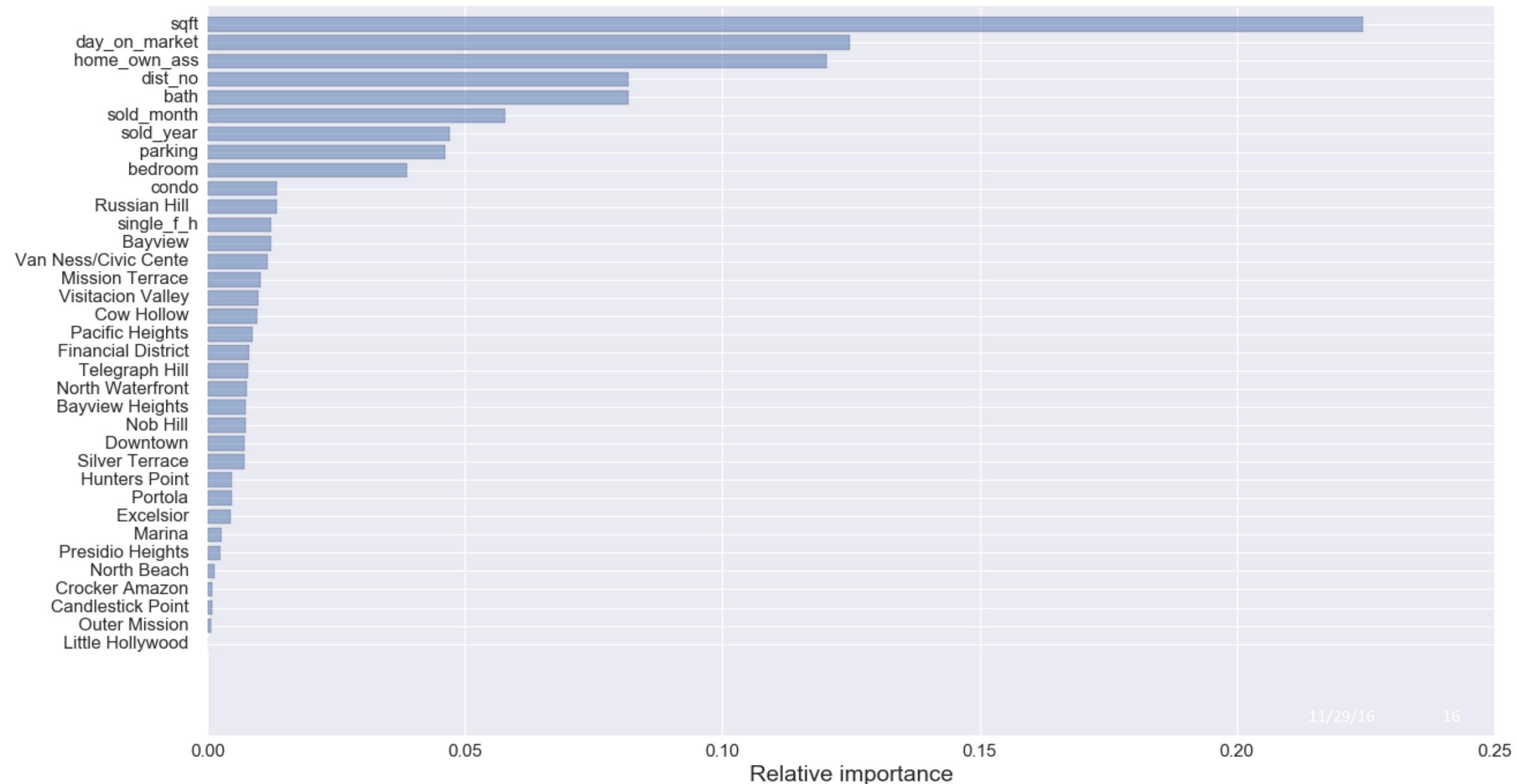
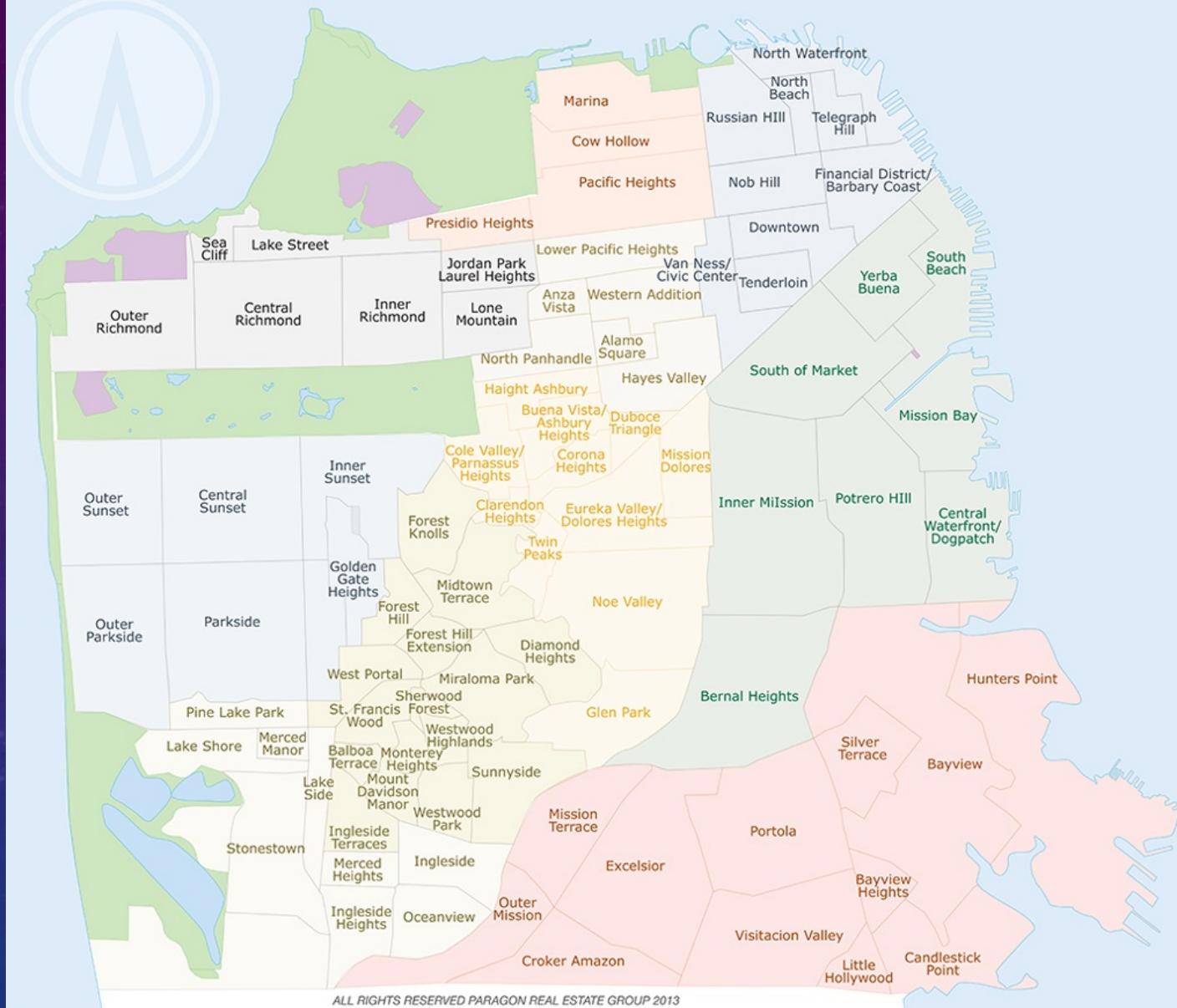


Figure 1



# PREDICTIONS:

| Date       | Square-Foot | Sale-price | Predicted price |
|------------|-------------|------------|-----------------|
| 2016-08-16 | 1050        | 980,000    | 1,167,352.68    |
| 2016-08-16 | 1312        | 850,000    | 763,698.09      |
| 2016-08-16 | 700         | 917,500    | 885,745.47      |
| 2016-08-16 | 760         | 670,000    | 634,854.98      |
| 2016-08-16 | 4617        | 9,750,000  | 5,233,383.73    |
| 2016-08-17 | 550         | 699,000    | 800,931.58      |
| 2016-08-17 | 527         | 385,000    | 326,196.34      |
| 2016-08-17 | 1000        | 775,000    | 691,178.71      |
| 2016-08-17 | 1210        | 1,050,000  | 950,538.13      |
| 2016-08-17 | 975         | 790,000    | 704,674.36      |
| 2016-08-17 | 1140        | 1,100,000  | 1,256,932.34    |
| 2016-08-18 | 1290        | 755,000    | 781,256.91      |
| 2016-08-18 | 1841        | 1,850,000  | 1,955,437.64    |
| 2016-08-18 | 1350        | 1,475,000  | 1,450,500.51    |



# CONCLUSION AND FUTURE WORK

|                                  |              |
|----------------------------------|--------------|
| test set sale price avg:         | 1,642,855.04 |
| test set with list price avg:    | 1,655,398.35 |
| diff:                            | -12,543.31   |
| test set without list price avg: | 1,607,460.92 |
| diff:                            | 35,394.12    |

- More data is needed: individual sales across the City, for all districts for at least 5 years. That data would allow us to analyze time series better, and the model selection through grid-search and cross-validated estimator would give us more reliable results.