Home Sales Analysis (I) - Exploratory Analysis

Outline

1. Data Cleanup
2. Pattern detection through Visualization
3. Correlations
    - between response and predictor variables
    - between predictor variables
4. Linear Model Selection
    a. Subset Selection
    - Best subset
    - Stepwise selection
    - Hybrid approaches
    b. Shrinkage
    c. Dimension Reduction
5. Choosing the Optimal Model
    a. Cp, AIC, BIC, Adjusted R^2
    b. Validation / Cross Validation
6. Assessing Model Accuracy
    a. Residual Analysis
    b. Regression Results
    c. Potential Problems
        - Non linearity of the response-predictor relationship
        - Correlation of error terms
        - Non-constant variance of error terms
        - Outliers
        - High Leverage points
        - Collinearity

Introduction

The data set of interest in this analysis is for homes listed for sale in the various zip code of Colorado. We want to develop a linear regression model to model the home price as the dependent variable against 168 odd independent variables. Linear model is selected primarily for interpretability.

We note that there are 2 variables representing the home price, namely, *curr.list.price* and *sold.price*. List price is determined by home owner/agent based on a number of factors pertinent to the market. We develop a model to determine the variables influencing the *curr.list.price*. In other words, *curr.list.price* is treated as the dependent variable. We note, *sold.price* is highly correlated to *curr.list.price* and is well explained by the *curr.list.price* in a regression model as shown in appendix.

Data / Quality

Data
* The data comes from 2 zip codes in Colorado
* It contains both quantitative (SqFt, Bed, Bath) and qualitative variables (Style, Bsmt Fin)
* Quantitative variables are both discrete (Bed, Bath) and continuous (SqFT).

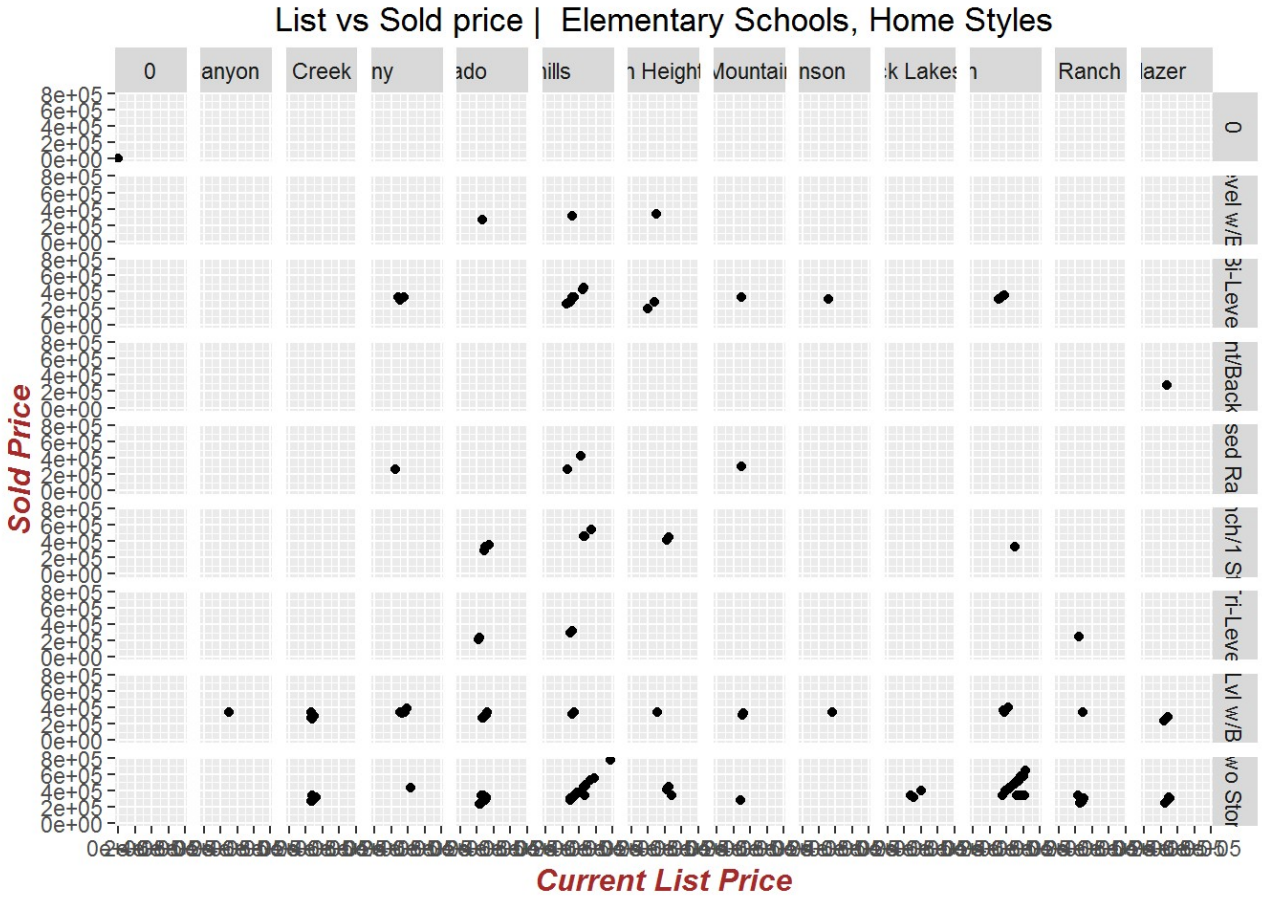
Quality
* There are many missing values both in qualitative and quantiative variables. Regressions models cannot be run in presence of missing values.
* Some columns had a large number of missing values. In this case there is no point in trying to impute the values. Hence, we deleted the column. The rationale is that if these columns are really meaningful then source data must be enhanced or cleaned as approriate by domain expert and provided as input. * For columns which had fewer missing values.
+ If they were quantitaive values, then we imputed them with column averages.
+ If qualitative we again deleted the column as there is no meaningful way to impute qualitiative values in most cases. Imputation of missing values with column average
* Finally, there were columns which needed to converted to numeric format, $ sign had to removed from currency columns etc.

All Data cleanup is done directly in the source csv file. Ideally, it should be done programmatically.

Pattern detection through Visualizations

Patterns in Home prices

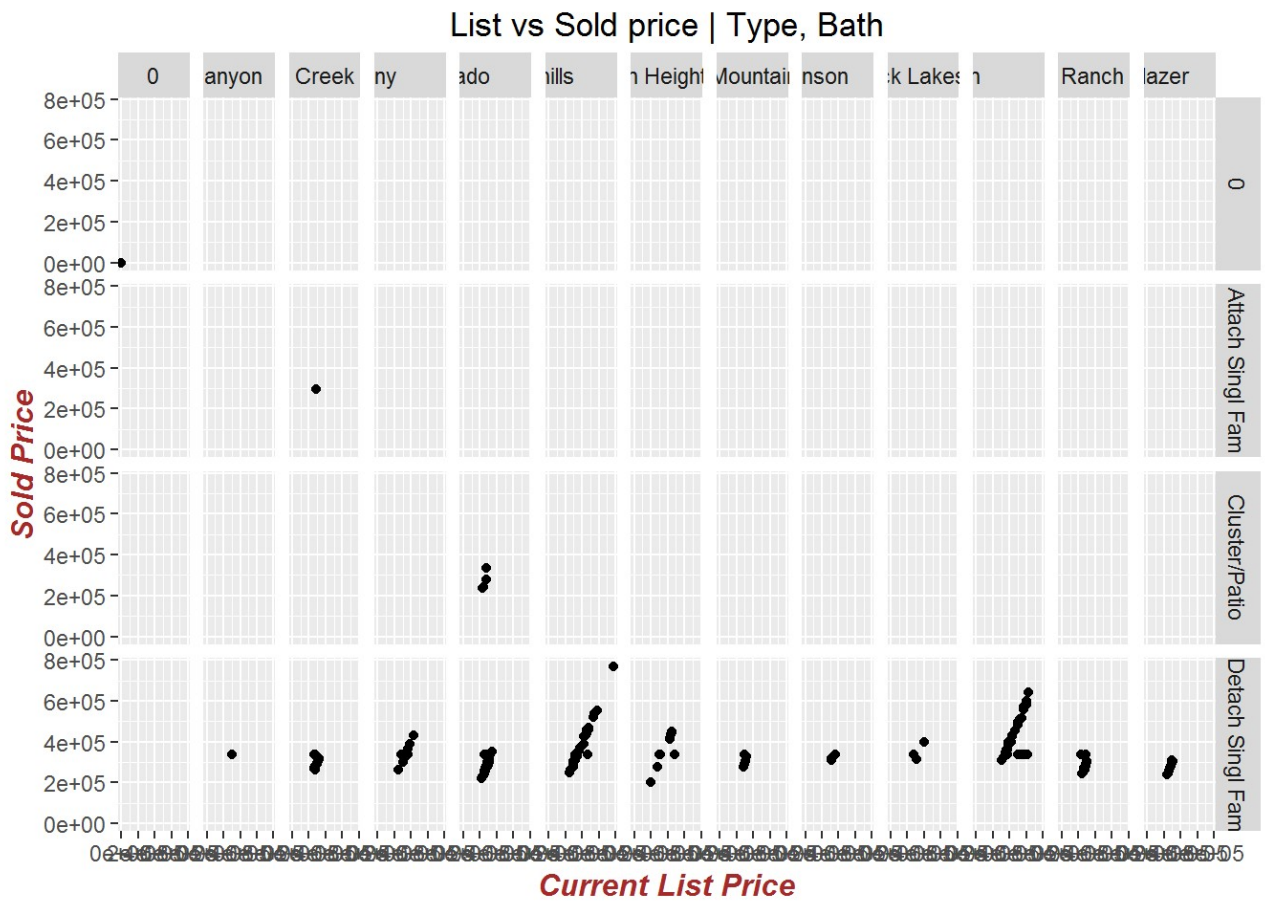
with Elementary Schools & Home Styles



Observations

- Eldarado elementary seems to be the most active school in 80129 zip code.
- Two story & Tri Level w/bsmt in high demand in Eldarado, Coyote, Saddle Ranch elementaries across price points.

with Elementary Schools & Home Types



Observations

- detached ‘Single Family’ homes are hot favorites in 80129

with Beds & Baths

Observations

- Homes with 3bed/3bath most sought after
- 3 Bed/4 Bath - price seems to flatten out at higher list prices.

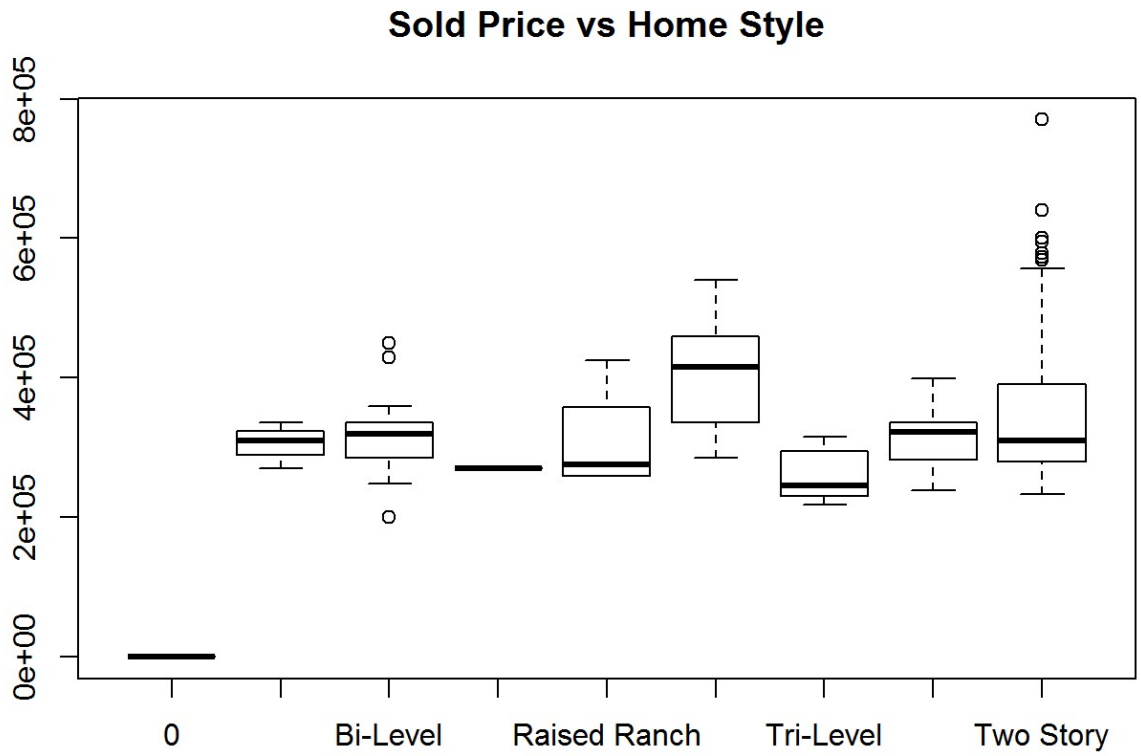
####Schools
Why is the Eldorado
seeing so much activity
with respect to Home
Sales?

The ranking data is from ‘greatschool.com’ website.

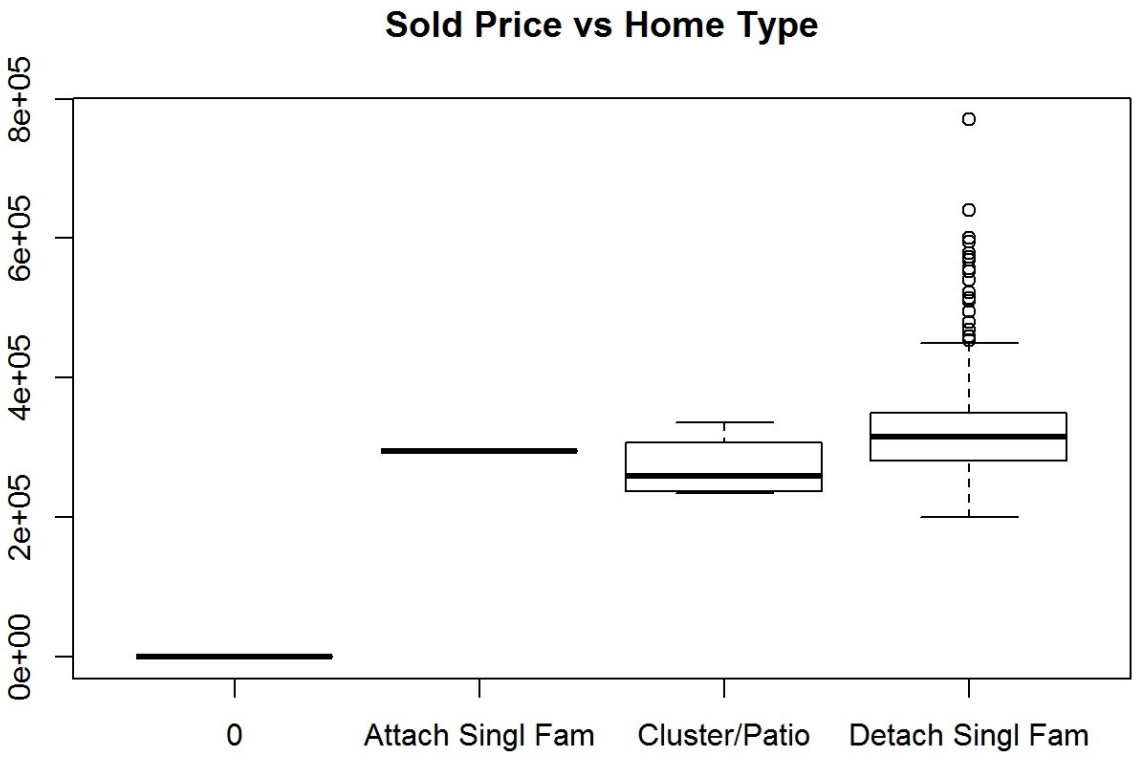
School	Rank
Bear Canyon	9
Cayote Creek	9
Eldorado	8
Saddle Ranch	10
Trail Blazer	7

Impact of (some) Qualitative Variables on Home prices

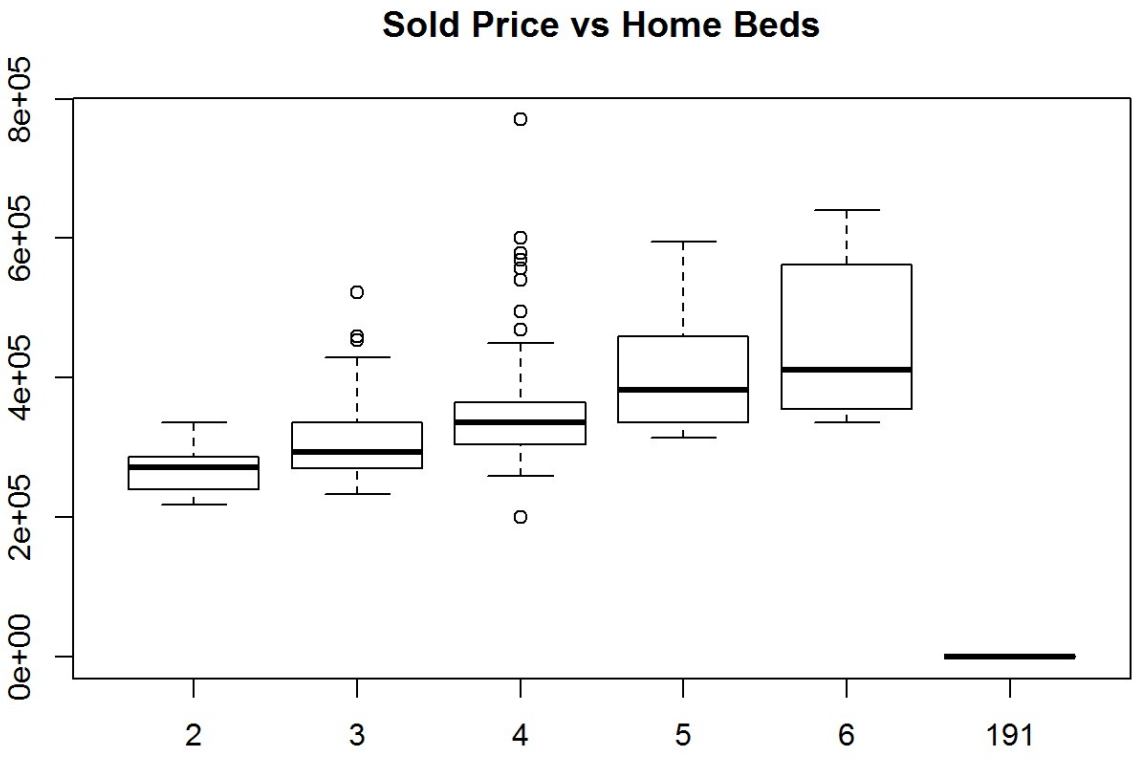
Home Styles

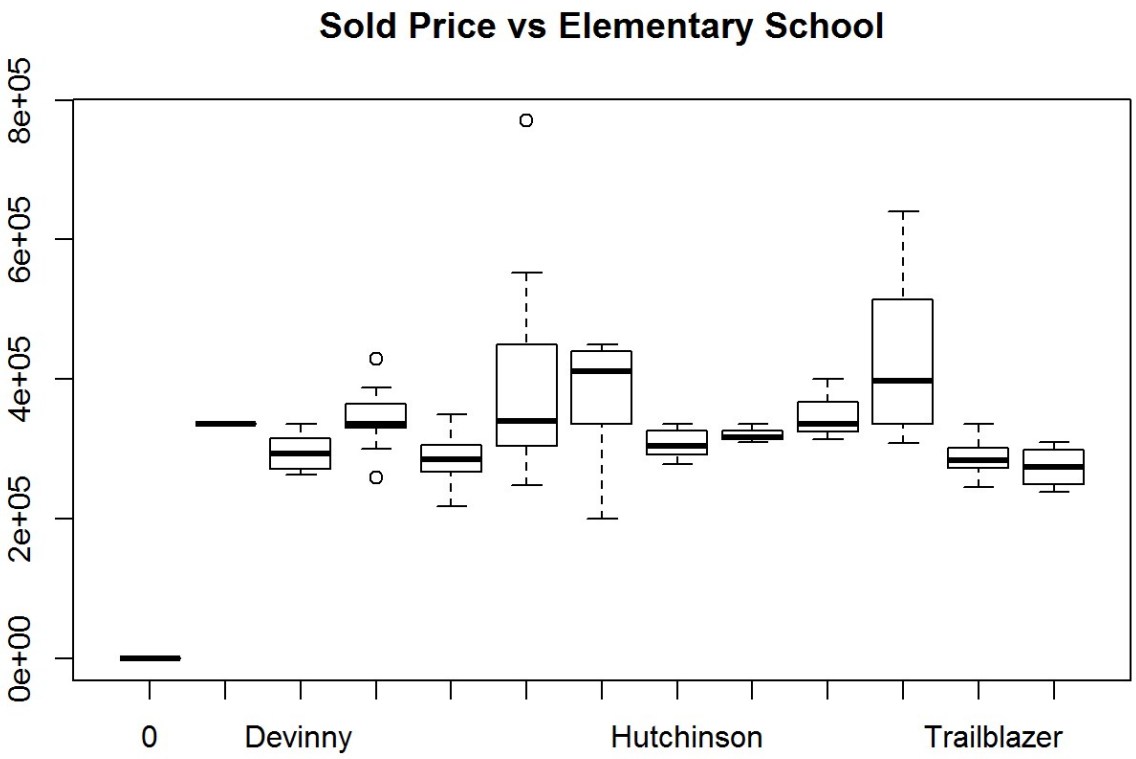


Home Types


Sold Price vs Home Type
0e+00
2e+05
4e+05
6e+05
8e+05
0
Attach Singl Fam
Cluster/Patio
Detach Singl Fam

Bedrooms





We examine the correlations of the *sale.price* with other numerical variables in the data set. We notice, *sold.price*

- has a high correlation with the list price
- has a postive correlation with PSF, area, number of Bedrooms and Bathrooms.
- has a negative correlation with variables such as PSF.Fin, Current and Total DOM.

Correlations

```
##      Sold.Price Curr.List.Price            Beds           Baths
##      1.00000000      0.90440827     -0.23656682     -0.24568139
##            SqFt             PSF         PSF.Fin        Curr.DOM
##      0.71291958      0.09706055     -0.18560925      0.26754970
##       Total.DOM
##      0.24386815
```