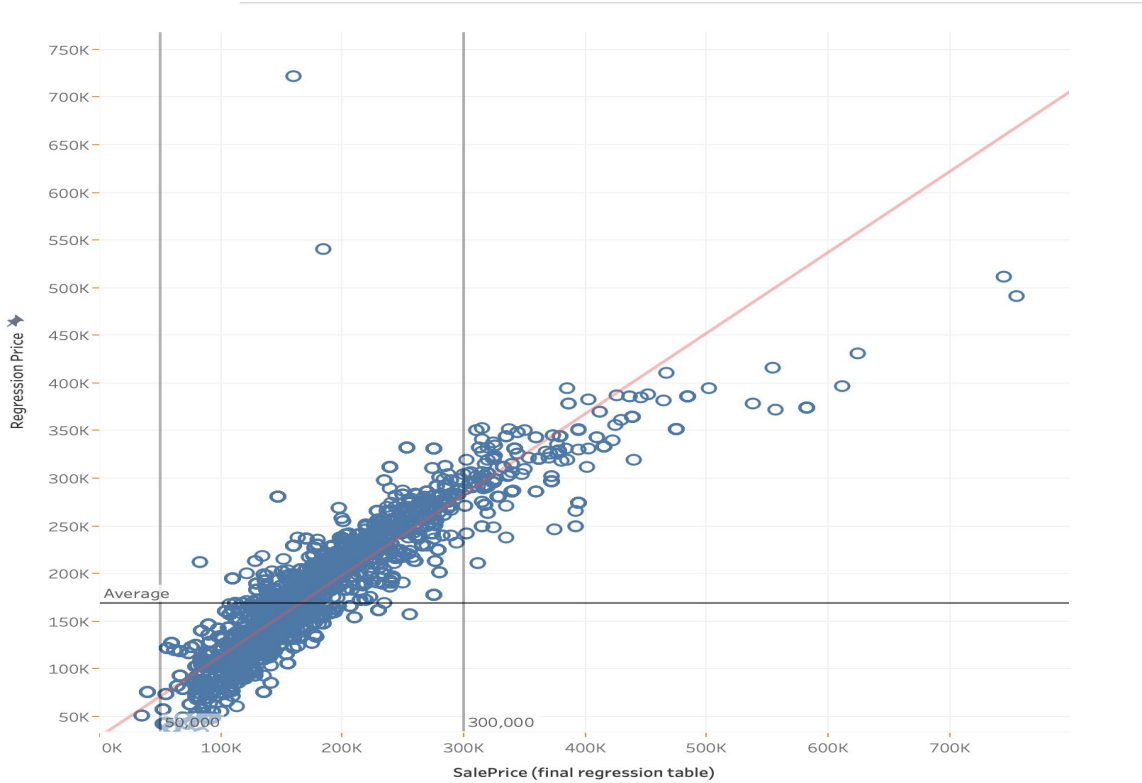
Exploration & Prediction of House Sale Prices using regression modelling – Technical Presentation

Date: 01/04/2019

Presenter: Prahlad Lama

With an average accuracy of 80% (plus or minus $20,000), an improved pricing estimate can be utilised by the real estate agents to improve sales pricing, and potentially commission yield estimates.

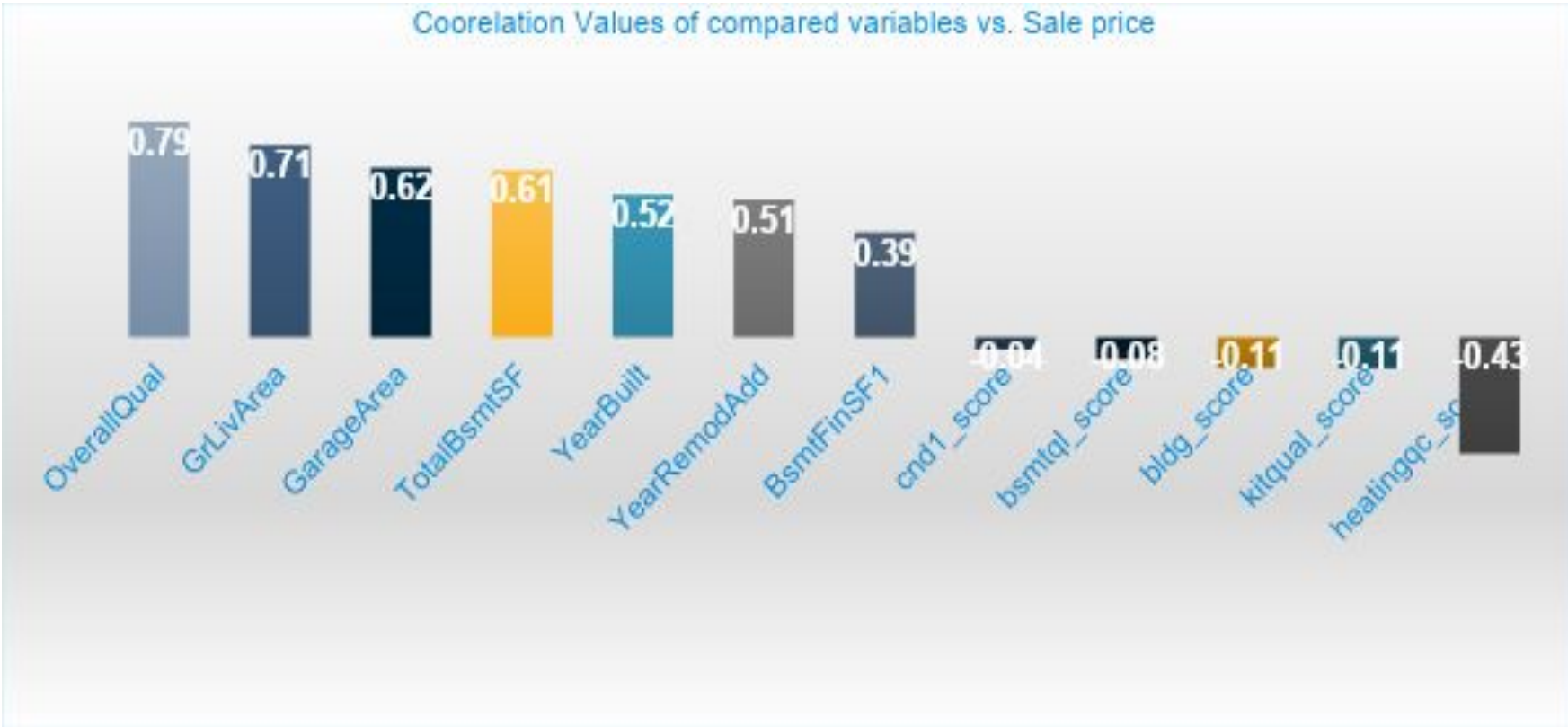


1) **Improved Accuracy -** Between Sales Price of $50k -$300k, we can make sales estimate of a house in Ames,Iowa using this Regression model with 80 % accuracy rate
2) **Predictor Variables** - House feature like Overall Quality, Living Area above Ground and Garage Area are the top 3 important predictors of house price in Ames,Iowa area.

1) Beyond these points ,sales estimate accuracy is likely to deviate heavily especially after $400k and predicted price is less than the actual price majority of the time with bigger margin of error.

Source: Ames,Iowa HousingRecords

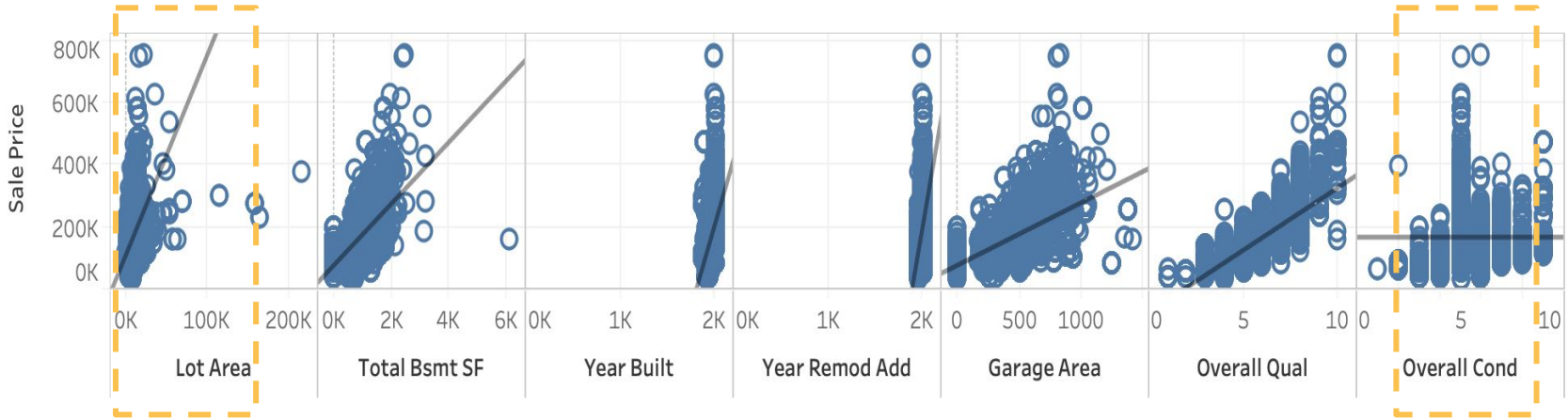
The top 3 variables with the most predictive strength are Overall Quality, Ground above Living Area and Garage Area which are further reinforced by the correlation bar chart below.



Correlation of categorical dummy variables like building style (bldg_score),kitchen quality(kitqual_score), radial distance from main highway(cnd1_score), basement quality(bsmtql_score) show weak correlation with Sale Price

OverallQual: Overall quality of the house **bldg_score**: building **style kitqual_score**: kitchen quality **cnd1_score** :radial distance from main highway
GrLIvArea :Ground Above Living Area
TotalBsmtSF:Total Basement Surface Area

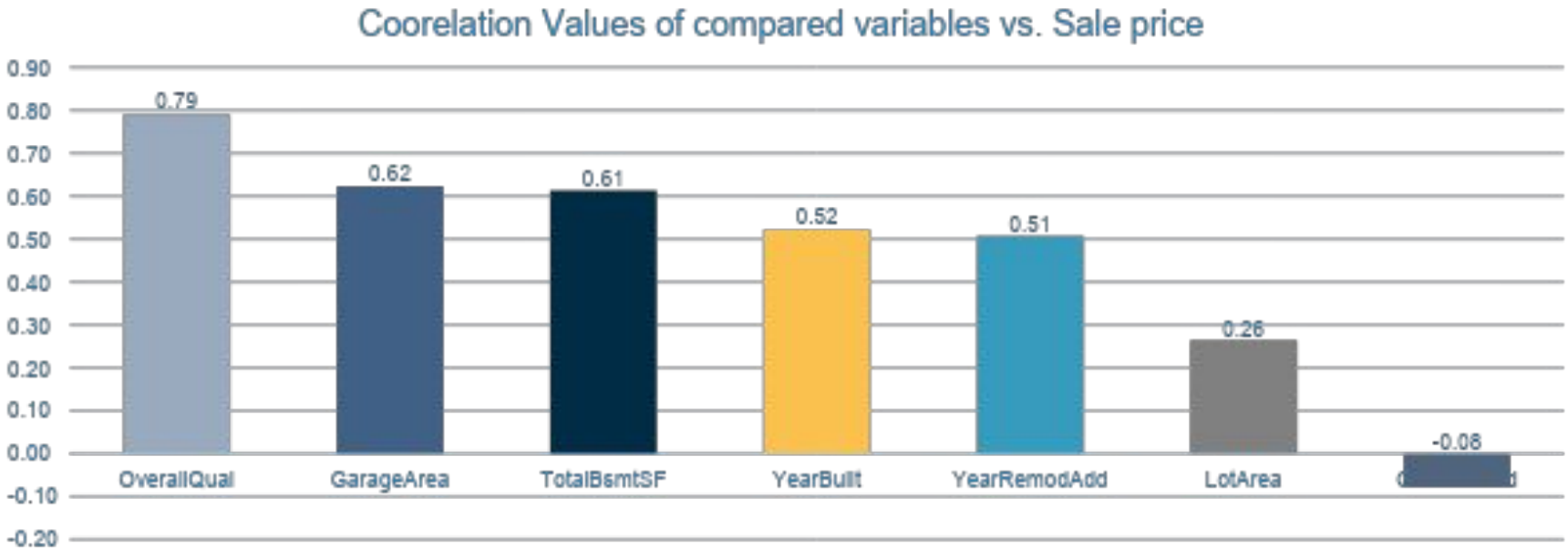
Variables exhibiting strong linear relationships were chosen whilst those with weaker relationships were avoided.



1) We Chose numerical features of house from our data: Lot Area, Overall Quality, Overall Condition ,Year Built, Year Remodeled ,Total Basement Surface Area, Garage Area to see if they are strong predictors of housing prices.
2) As we can see from the scatter plots, Overall Condition of the house and Lot Area don't show strong linearity with Sales Price ,so we will avoid these variables in our model .

Source: Ames,Iowa Housing Records Data

We carefully selected core variables using the Pearson correlations as our measurement for what variables are significant or insignificant.

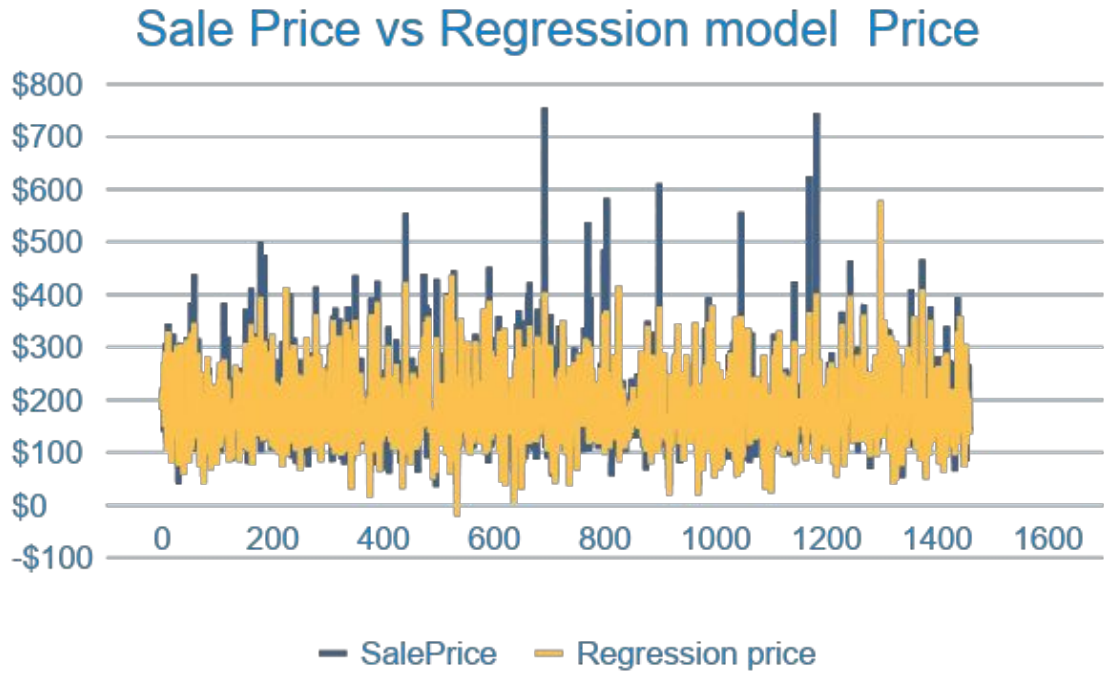


1) We Chose numerical features of house from our data: Lot Area, Overall Quality, Overall Condition ,Year Built, Year Remodeled ,Total Basement Surface Area, Garage Area to see if they are strong predictors of housing prices.
2) As we can further confirm from the Pearson Coefficient, Overall Condition of the house and Lot Area, show weak correlation with Sales Price ,so we will avoid these variables in our model .

Source: Ames,Iowa Housing Records

From using a multivariate regression equation, we noted a 0.73 Adjusted R square value,and achieved improved accuracy in sales estimate

Regression Statistics	
Multiple R	0.856976005
R Square	0.734407874
Adjusted R Square	0.731834663
Standard Error	41139.0481
Observations	1460

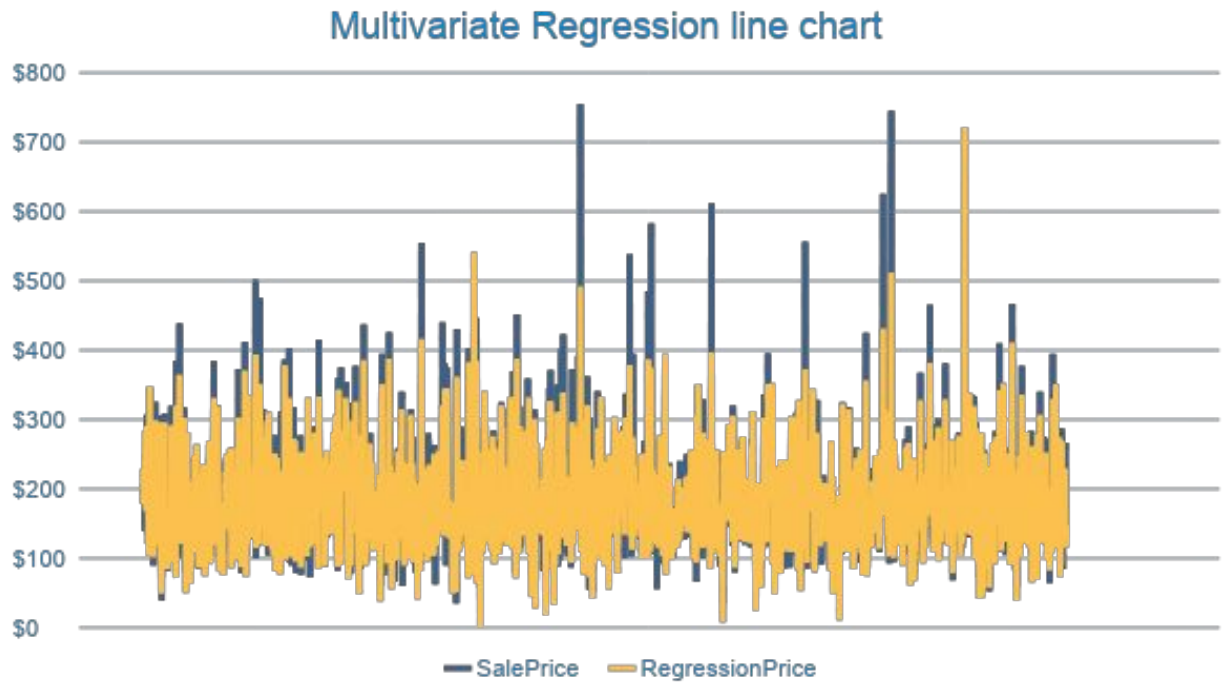


1) We Chose numerical features of house from our data: Overall Quality, Overall Condition ,Year Built, Year Remodeled ,Total Basement Surface Area, Garage Area and transformed categorical variables : neighborhood(ngh_score)building style (bldg_score),kitchen quality(kitqual_score), radial distance from main highway(cnd1_score), basement quality(bsmtql_score) ,sale condition (salecond_score)to see if they are strong predictors of housing prices.
2) As we can see from plots, Our model prediction price varies largely over sales price above $ 400k .
3) From Excel calculation ,we found out that the model Adjusted R square value is 0.73 , and variables like neighborhood ,sale condition, and basement condition were statistically insignificant (p>=0.05)so further testing with other variables is required to improve this value in order to make more accurate prediction.

Source: Ames,Iowa Housing Records
Housing prices data.xls

This was further improved by inclusion of variables in our model to achieve 80% accuracy in sales estimate between prices $50k-$300K

Regression Statistics	
Multiple R	0.89563203
R Square	0.802156733
Adjusted R Square	0.800516015
Standard Error	35481.90564
Observations	1460



1) We added numerical features of house from our data: Ground Above Living Area(GrLivArea),First Floor Basement Surface Area(BsmtFinSF1), and removed statistically insignificant categorical variables and tested different regression model and finally came up with above model.

2) As we can see from plots, Our model can accurately predict sales price between $50k -$300k with 80% accuracy.

3) From Excel calculation ,found out that the model Adjusted R square value increase to 0.80 ,so more accurate prediction was achieved.

Source: Ames,Iowa Housing Records
Housing prices data.xls