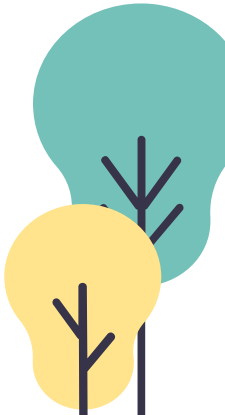
HOUSE PRICE PREDICTION

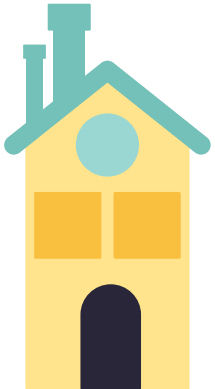
Team 5
Site Bai, Smitha Kannanaikkal,
Ellie Park, Joy Chen

TABLE OF CONTENTS



CONTEXT

- Google reported the search “When is the housing market going to crash?” had spiked 2,450% in the past month.
- The real estate market is booming right now and is breaking historic records in April, as home prices rose 21% YoY and the median home-sale price soared nationwide
- With the housing market evolving rapidly, it is important to understand the fluctuation in market trends



PROJECT QUESTION

KEY QUESTIONS

- Our team would like to focus on two objectives
  - First: For Sellers
    - We want to learn different attributes that impact the sale price and be able to boost the price when budget is limited.
  - Second: For Buyers
    - By forecasting the count of house sales, we enable customers looking to purchase houses to have an idea when there will be more options listed, and also estimate the price based on supply and demand.
- Our dependent variable is sales price & sales volume for time series models

DATA GATHERING

DATA GATHERING

Our Dataset was gathered from Kaggle & was part of the Kaggle contest

Dataset has 1,457 observations with 80 variables

It includes variables related to features of the houses including home condition, kitchen, porch, fireplace, basement, size of each floor, etc.

It also includes sales related information including date sold, sale type, sale condition, etc.

METHODOLOGY

METHODOLOGY

STEP 1

Data cleaning, data transformation

STEP 2

Feature selection, training/ testing data splitting

STEP 3

Running multiple models including regression tree, random forest and XGBoost

STEP 4

Qualitative research to better understand our data

STEP 5

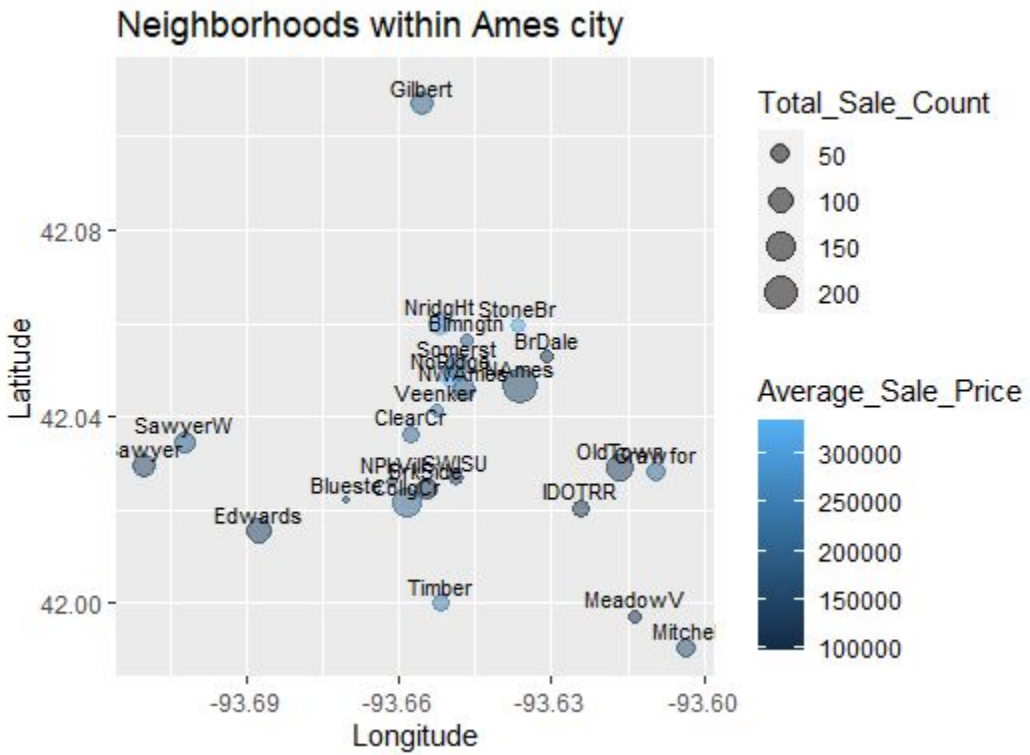
Business implication and insight

DATA Visualization

SALES PRICE

median price around 180K dollar

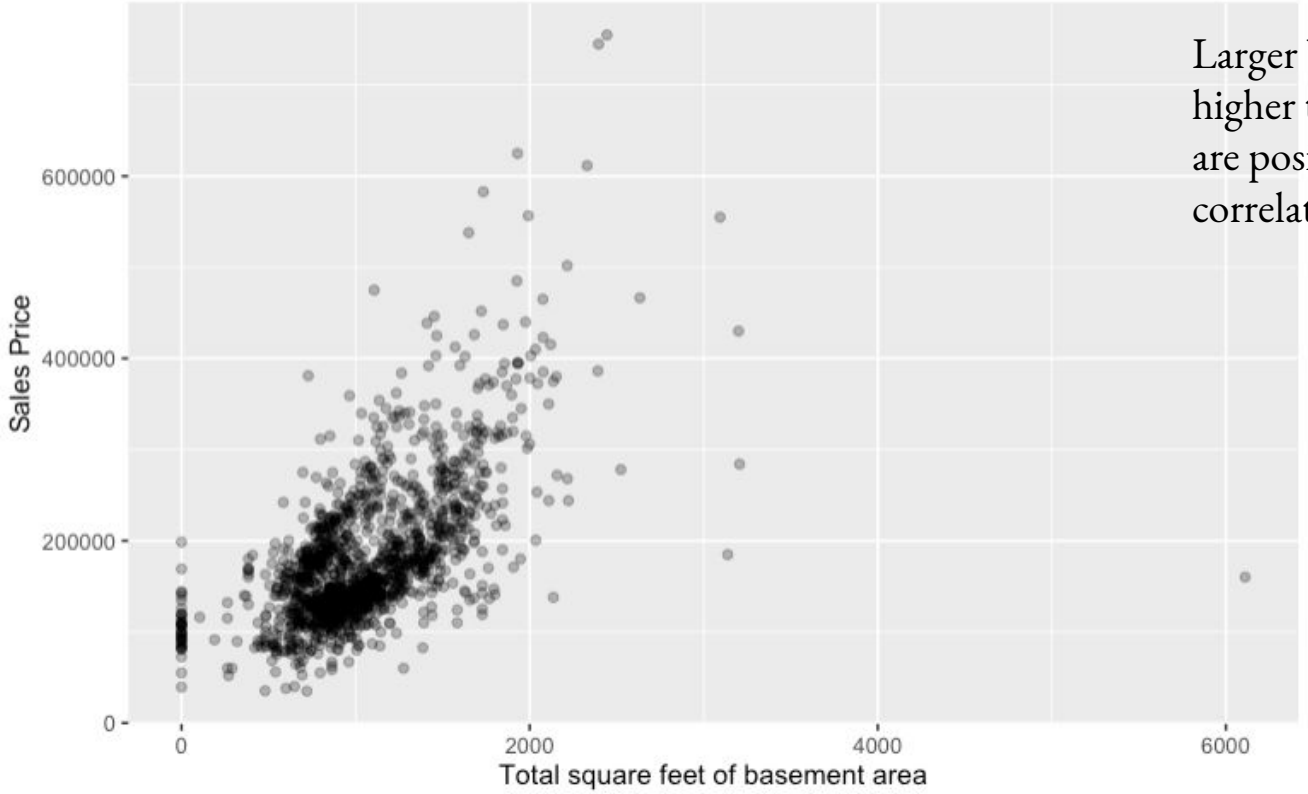
NEIGHBORHOODS



large sale happening where the average sale price is lower

SALES PRICE & MATERIAL

BASEMENT AREA & SALES PRICE



Larger basement area, higher the price. They are positively correlated.

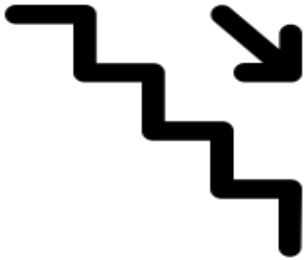
ANALYSIS
For Sellers

REGRESSION

	All in Linear Regression	Backward Feature Selection
Number of Attributes	125	82
R-squared	90.81%	90.68%
Test RMSE	28308.45	28344.37

Features impact on the Sales Price

	Neighborhood Stone Brook	Has Basement Yes	Heating Quality Condition Poor
House Sale Price Prediction	↑ 20%	↑ 12%	↓ 29.5%



Features impact on the Sales Price

	Neighborhood Stone Brook	Has Basement Yes	Heating Quality Condition Poor
House Sale Price Prediction	↑ 20%	↑ 12%	↓ 29.5%
Potential Cause	high rating schools in this area	Means more space, also according to the National Association of Realtors, a finished basement returns about 69 percent of the investment.	Iowa is cold in winter

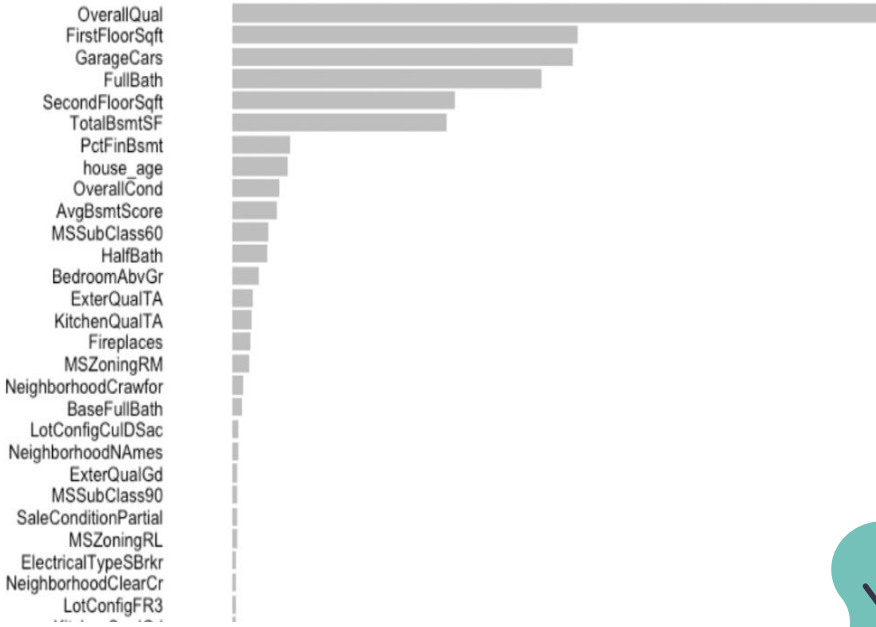
DECISION TREE/ RANDOM FOREST/ XGBOOST

smallest

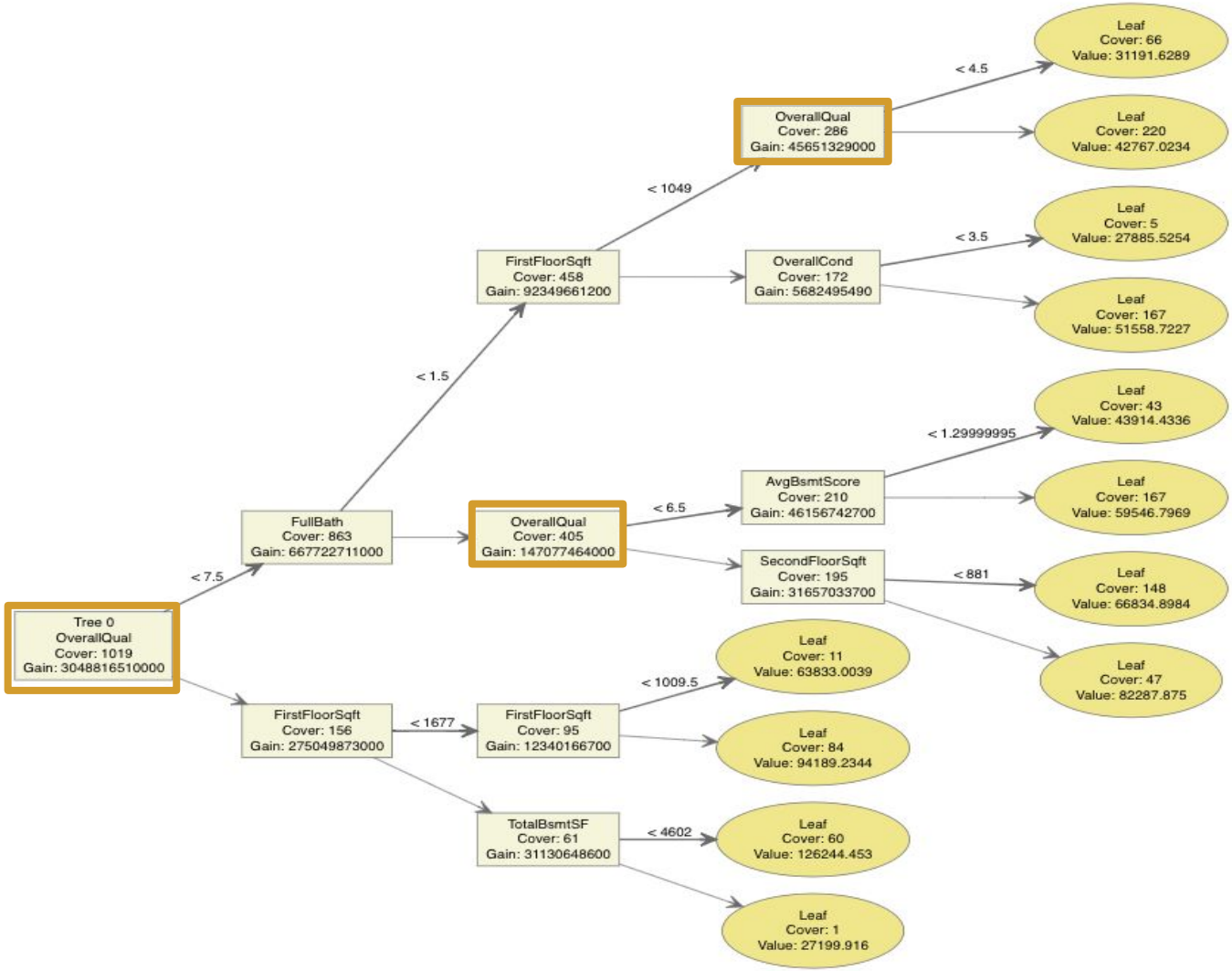
		Regression tree	Random Forest	XGBoost
RMSE	Training data	38413.77	19754.89	**10023.13**
	Testing data	45617.69	35995.96	**35812.75** (20% of average sales price)

Our regression model has an RMSE of 28K, we think here the regression is better than the trees is probably because our dataset is rather small.

Significant Coefficients From XGBOOST

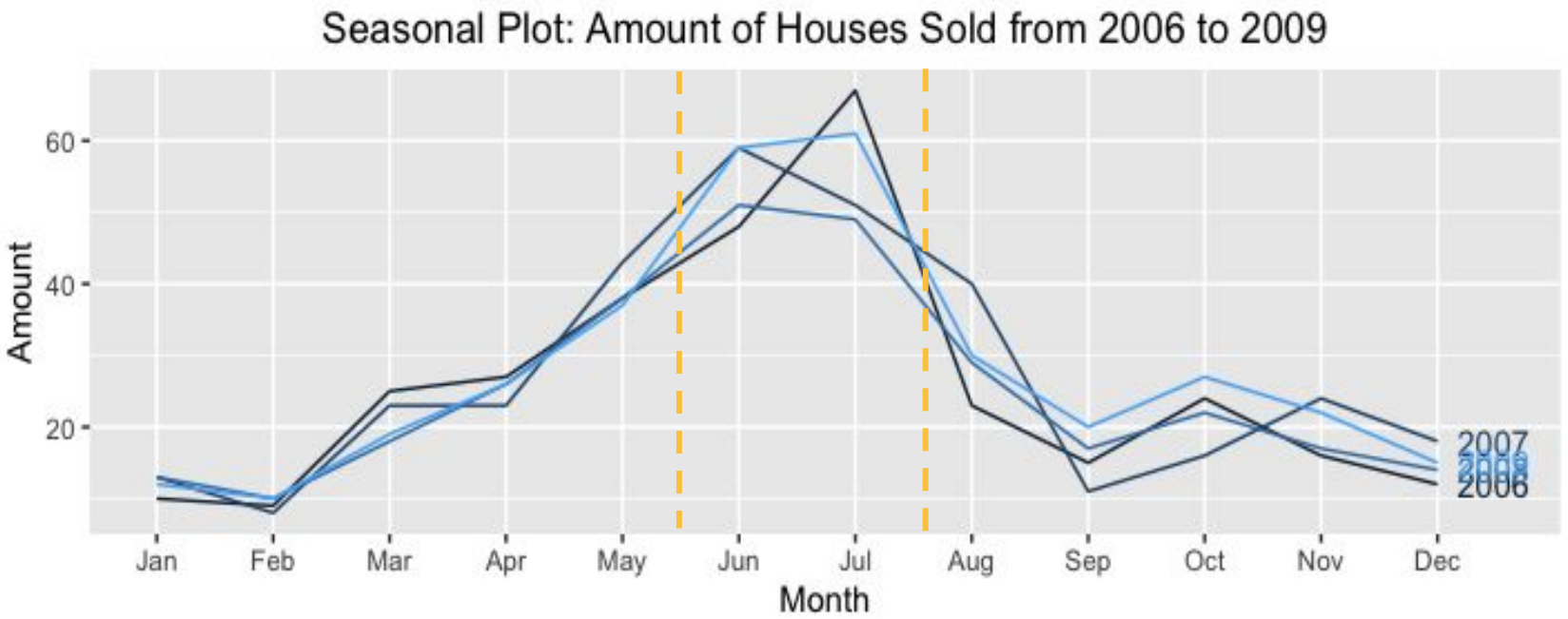


Only 10 out of 82 attributes are significant in affecting the sales price, and “Overall Quality” is the most important.


Tree 0
OverallQual
Cover: 1019
Gain: 3048816510000
< 7.5
FullBath
Cover: 863
Gain: 667722711000
< 1.5
FirstFloorSqft
Cover: 458
Gain: 92349661200
< 1049
OverallQual
Cover: 286
Gain: 45651329000
< 4.5
Leaf
Cover: 66
Value: 31191.6289
Leaf
Cover: 220
Value: 42767.0234
OverallCond
Cover: 172
Gain: 5682495490
< 3.5
Leaf
Cover: 5
Value: 27885.5254
Leaf
Cover: 167
Value: 51558.7227
OverallQual
Cover: 405
Gain: 147077464000
< 6.5
AvgBsmtScore
Cover: 210
Gain: 46156742700
< 1.29999995
Leaf
Cover: 43
Value: 43914.4336
Leaf
Cover: 167
Value: 59546.7969
SecondFloorSqft
Cover: 195
Gain: 31657033700
< 881
Leaf
Cover: 148
Value: 66834.8984
Leaf
Cover: 47
Value: 82287.875
FirstFloorSqft
Cover: 156
Gain: 275049873000
< 1677
FirstFloorSqft
Cover: 95
Gain: 12340166700
< 1009.5
Leaf
Cover: 11
Value: 63833.0039
Leaf
Cover: 84
Value: 94189.2344
TotalBsmtSF
Cover: 61
Gain: 31130648600
< 4602
Leaf
Cover: 60
Value: 126244.453
Leaf
Cover: 1
Value: 27199.916

ANALYSIS
For Buyers

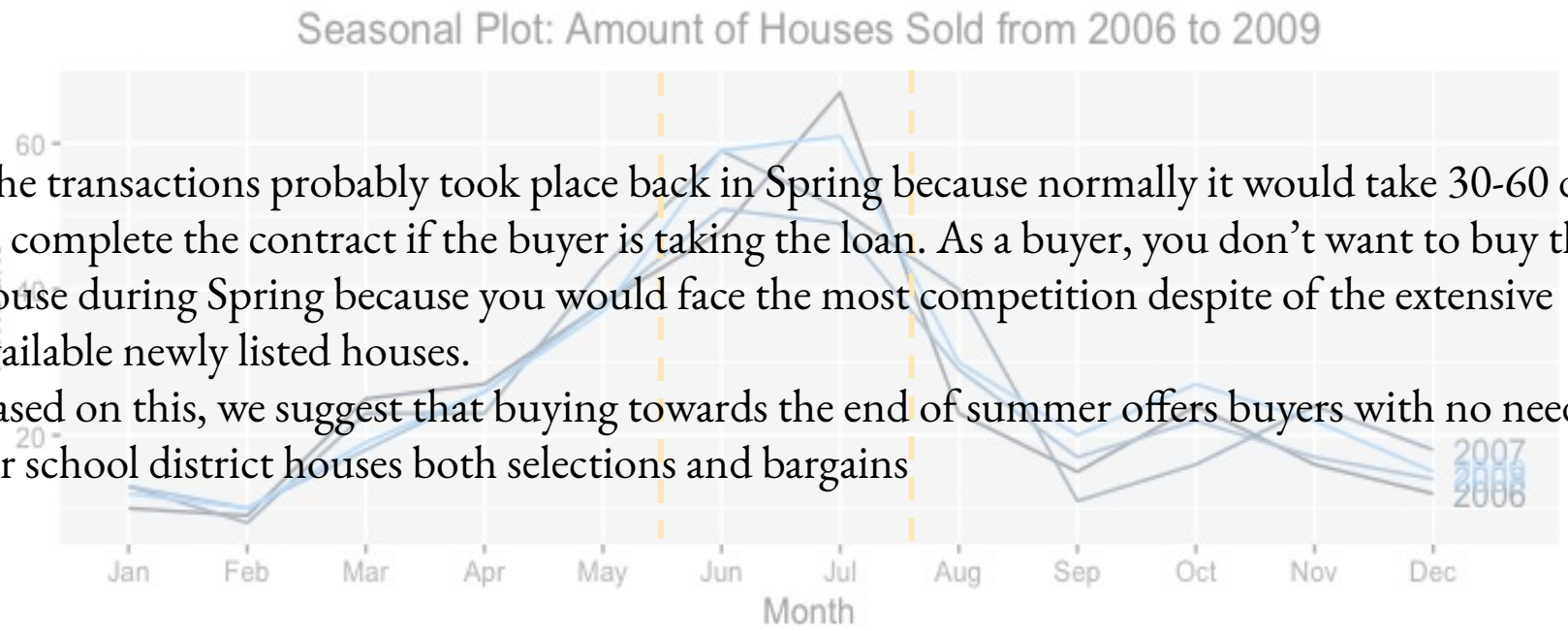
TIME SERIES



Spike between May and July

Potential reason: Moving closer to school before school starts

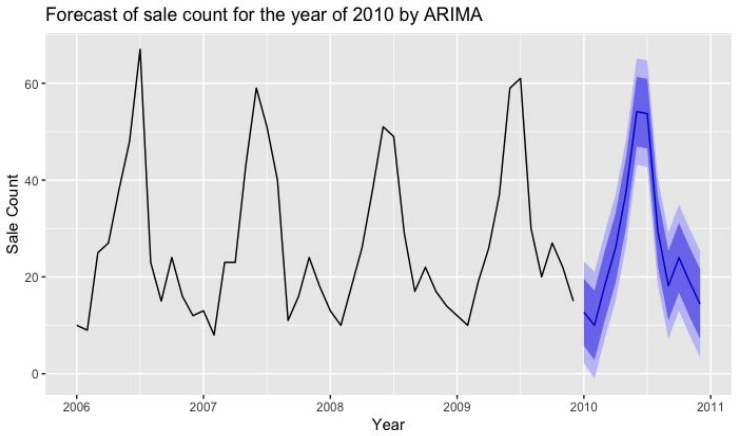
TIME SERIES



The transactions probably took place back in Spring because normally it would take 30-60 days to complete the contract if the buyer is taking the loan. As a buyer, you don't want to buy the house during Spring because you would face the most competition despite of the extensive available newly listed houses.

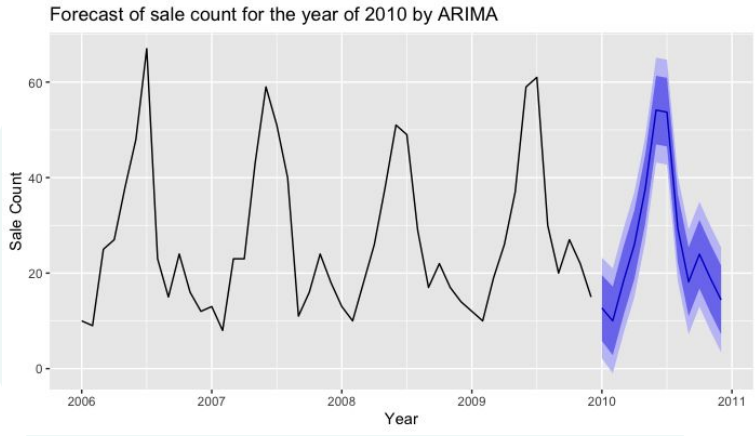
Based on this, we suggest that buying towards the end of summer offers buyers with no need for school district houses both selections and bargains

FORECAST



ETS(M,N,A) PREDICTION

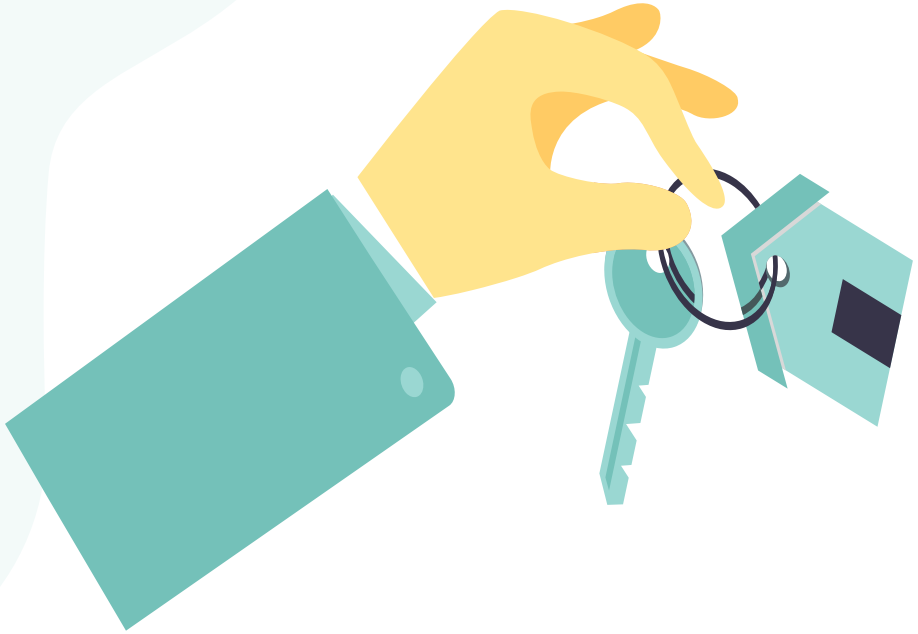
RMSE: 20.87



ARIMA PREDICTION

RMSE: 20.42

IMPLICATIONS



ANALYSIS & INSIGHTS

For sellers:

- Location matters.
- Polish the basement if there is one.
- Install heating system will enable sellers to ask higher price.

For buyers:

From late May to mid July, there is a greater demand for houses with sales peaking

- To avoid competition, buyers should look for houses after summer months so they have greater bargain power

LIMITATIONS

Dataset

- Our dataset has limited number of observations
- Is restricted to Iowa, thus, findings may not be applied to other states

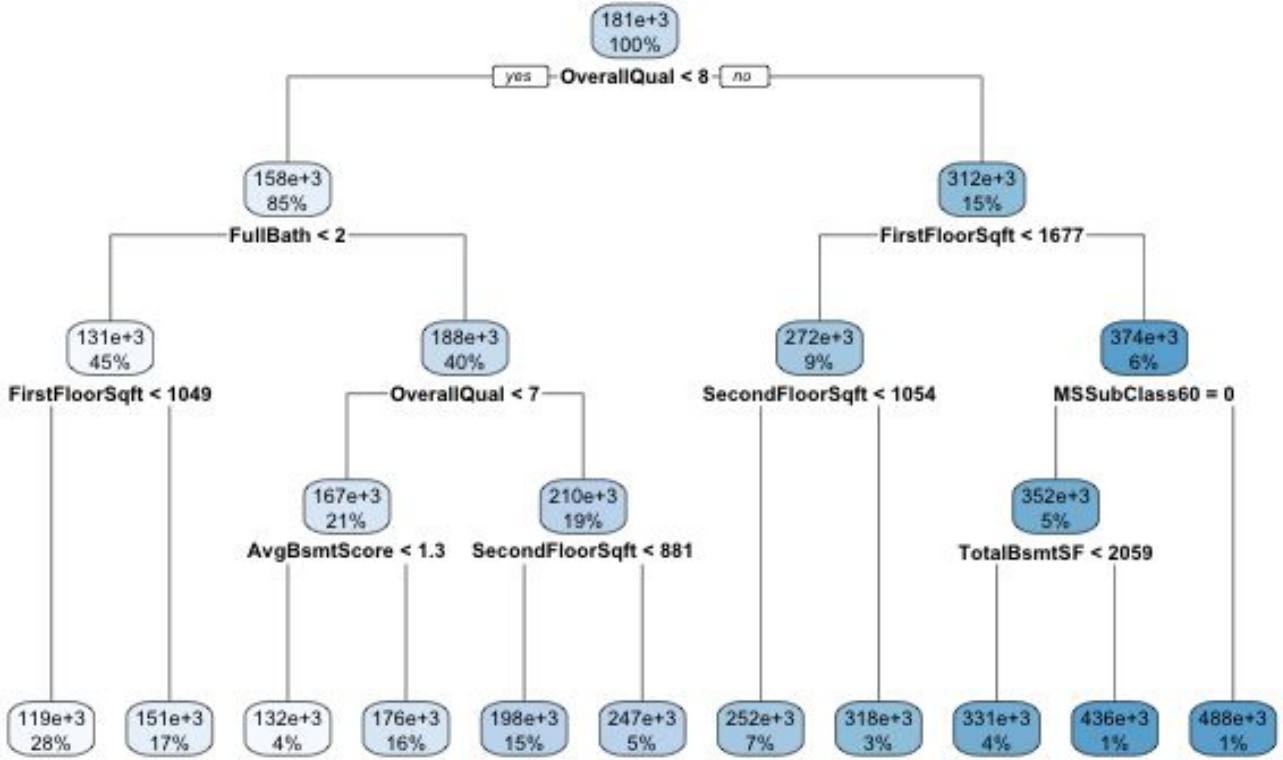
APPENDIX

STEP 1	Data cleaning: Manually go over 80 variables such as checking correlation between variables, take out or combined data with very few instances, take log for skewed numeric data.
STEP 2	Split data into training and testing (70:30).
STEP 3	Run multiple models including regression tree, random forest and XGBoost
STEP 4	Time series analysis.
STEP 5	Do further qualitative research to better understand our results.
STEP 6	Business implication and insight

Decision Tree Result

Decision Tree

```
library(rpart)
rpart(
 formula = SalePrice ~ .,
 data    = train,
 method  = "anova",
 control = list(cp = 0,
                minsplit = 13
                maxdepth = 12,
                xval = 10))
```

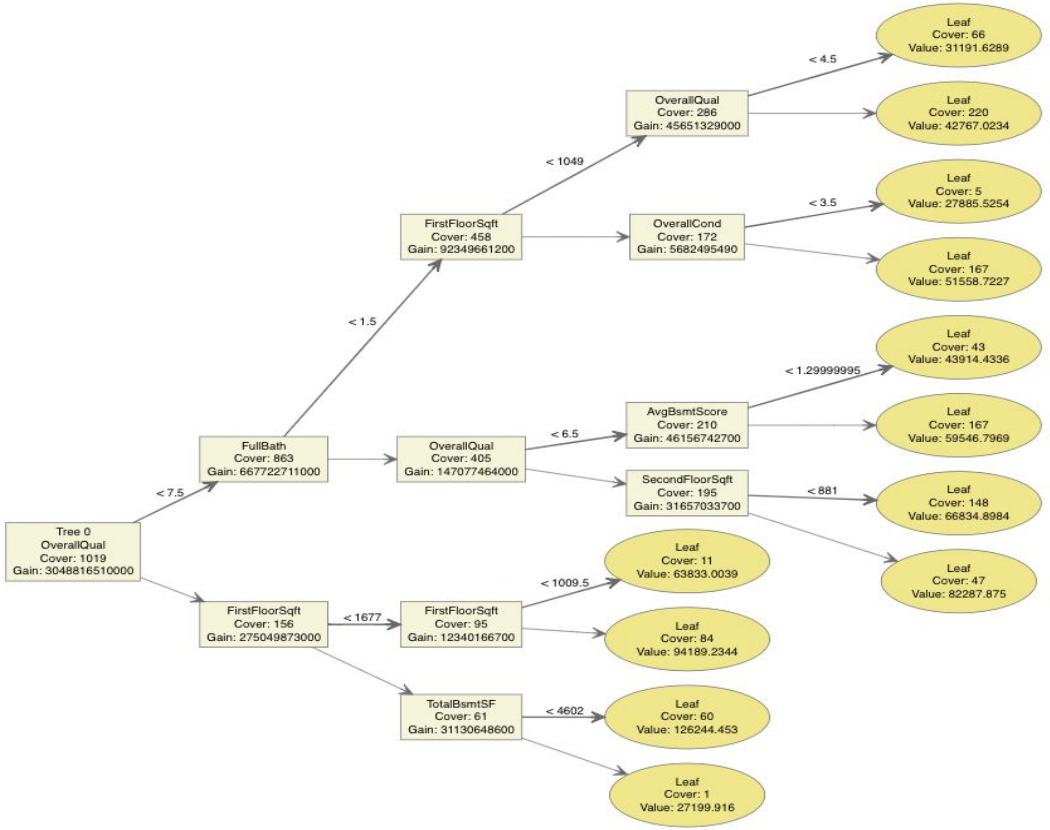


XGBoosting Result

XGBOOST

```
library(xgboost)

 xgboost(data = xgb_train ,
         label = SalePrice
         max_depth = 4,
         eta = 0.34,
         nrounds = 50,
         nthread = 2,
objective =
"reg:squarederror")  )
```



THANK YOU