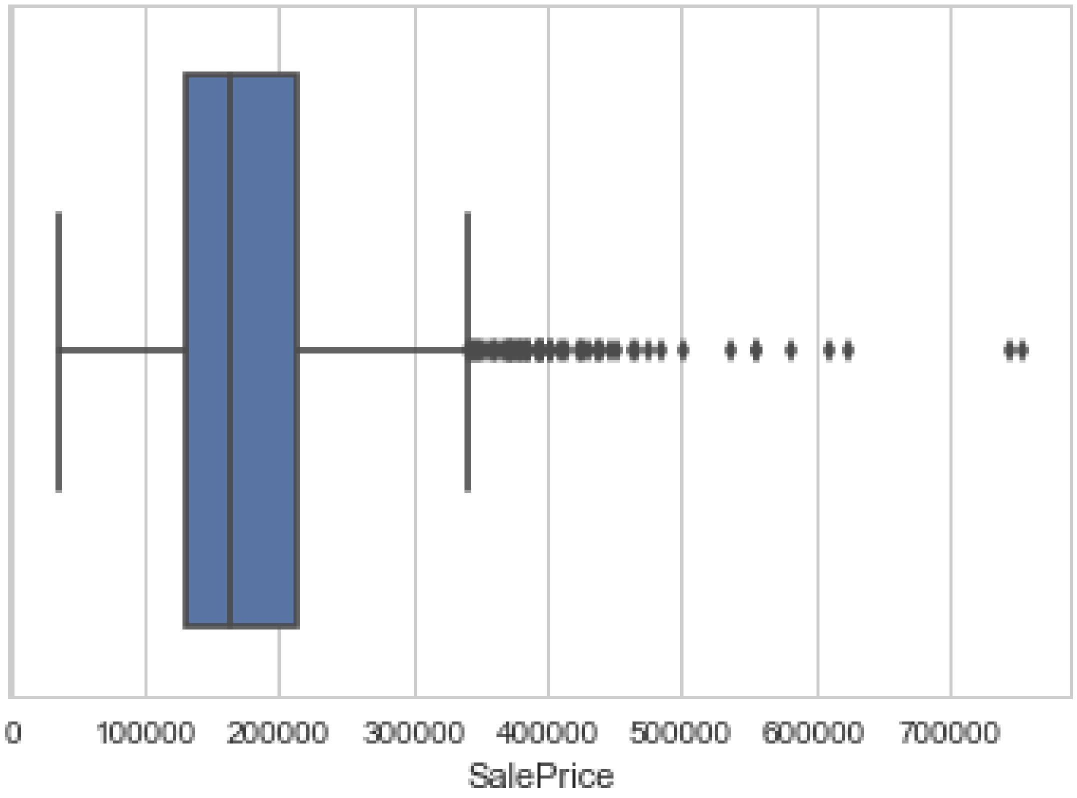
Ames housing Data

- In our data, the SalePrice is the target variable. My intuition, based off of the "Location, Location, Location" saying, was that categorical features such as the neighborhood and proximity to key roads, will be more important.




```
count       1460.000000
mean      180921.195890
std        79442.502883
min        34900.000000
25%       129975.000000
50%       163000.000000
75%       214000.000000
max       755000.000000
Name: SalePrice, dtype: float64
```

Data Cleaning

- There were 81 features, 43 of which were non-numerical.
- One of the numerically encoded variables was in fact categorical.

20	1-STORY 1946 & NEWER ALL STYLES
30	1-STORY 1945 & OLDER
40	1-STORY W/FINISHED ATTIC ALL AGES
45	1-1/2 STORY - UNFINISHED ALL AGES
50	1-1/2 STORY FINISHED ALL AGES
60	2-STORY 1946 & NEWER
70	2-STORY 1945 & OLDER
75	2-1/2 STORY ALL AGES
80	SPLIT OR MULTI-LEVEL
85	SPLIT FOYER
90	DUPLEX - ALL STYLES AND AGES
120	1-STORY PUD (Planned Unit Development) - 1946 & NEWER
150	1-1/2 STORY PUD - ALL AGES
160	2-STORY PUD - 1946 & NEWER
180	PUD - MULTILEVEL - INCL SPLIT LEV/FOYER
190	2 FAMILY CONVERSION - ALL STYLES AND AGES

Data Cleaning

- Many of the variables were subjective quality ratings, which could be ordered and ranked. However, some of these were open to interpretation:

```
BsmtFinType1: Rating of basement finished area

       GLQ     Good Living Quarters
       ALQ     Average Living Quarters
       BLQ     Below Average Living Quarters
       Rec     Average Rec Room
       LwQ     Low Quality
       Unf     Unfinshed
       NA      No Basement
```

- I considered this to be ordered, but reasonable people could disagree.

Random Forest Model

- For the actual modeling, I started with a Random Forest, which we know from class to have the greatest predictive power.
- Random Forest also has the advantage of being able to calculate feature importance.
- Oob score: 0.835710226929
- RMSE: 31879.720433631563
- Tuning the max features brought this down to

around 28,399.

OverallQual	0.057863
GrLivArea	0.045550
TotalBsmtSF	0.037886
GarageArea	0.036898
GarageYrBlt	0.030827
GarageCars	0.030787
BsmtFinSF1	0.029528
BsmtQual	0.028606
1stFlrSF	0.028328
2ndFlrSF	0.025593

Linear Regression

- That done, the problem with random forest is that it is less interpretable. It would be more useful to a homeowner to have a set of betas to know how to improve their house so as to increase the value.
- RMSE with all features comes out to 30,117.

Linear Regression

- Problems- intercepts and coefficients:

```
-1225960.64085
RoofMatl_ClyTile     -596458.719545
Condition2_PosN      -198145.225736
MiscFeature_TenC      -75761.407736
Exterior2nd_Other     -34333.722350
MSSubClass_180        -33806.499449
```

- Some of these are clear outliers- e.g. there is only one house with a tennis court.
- RMSE is 33,581.

Linear Regression

- Do another regression, but with only the "important" features as determined in the Random Forest. Resulting coefficients:

-94742.2859095	
GrLivArea	-32.522775
GarageYrBlt	-14.786447
TotalBsmtSF	-7.777781
GarageArea	4.344599
BsmtFinSF1	19.761692
2ndFlrSF	73.263248
1stFlrSF	100.924641
BsmtQual	11682.282470
OverallQual	20931.886587
GarageCars	21785.110861

- These are more sensible, but also less actionable. It is difficult to add square footage to your house, easier to pave a driveway.
- RMSE is 35,546.