Distributed Computing Project Report

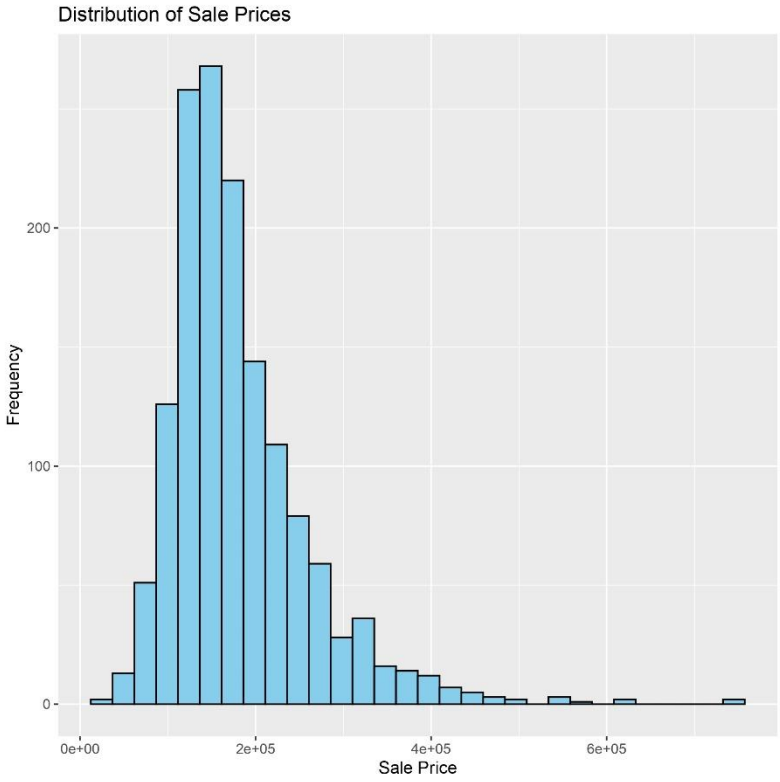
Team 44

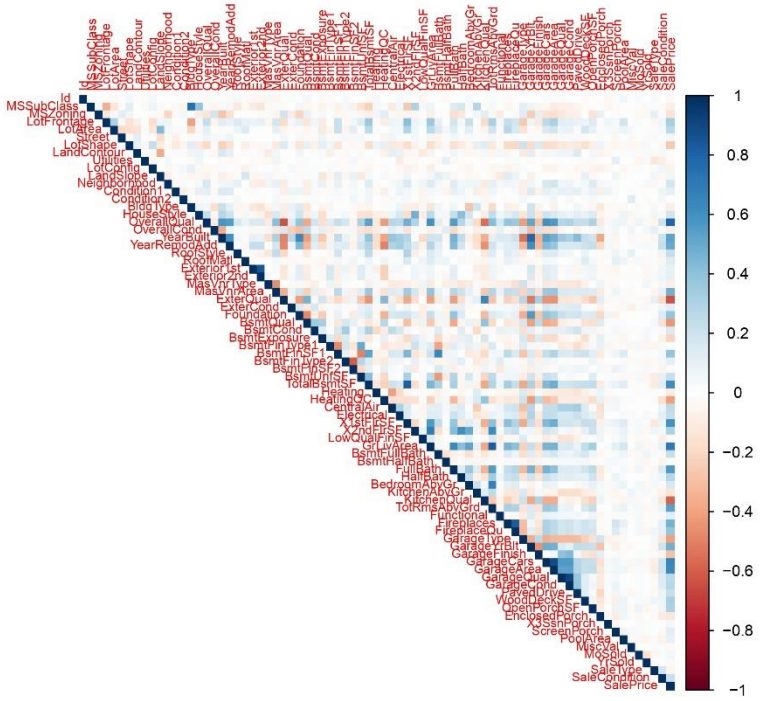
عمر عبد الرحمن فتحي إبراهيم	20191700413
بدر محمد أشرف شفيق جويلي	20191700179
شهاب إيهاب أحمد طلعت	20191700315

Preprocessing & Feature Selection:

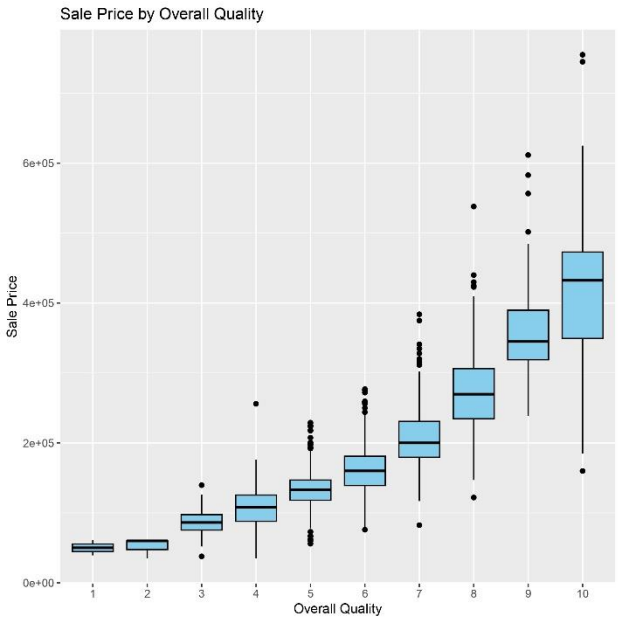
1. Train and test data are combined into one data frame to apply preprocessing once then split again at the end.
2. Dropped any column having more than 50% of it as NA.
3. Filled any remaining NA(s) with:
    a. Mean for numeric values.
    b. Mode for non-numeric values.
4. Applied factorization on the non-numeric columns.
5. Tried to remove values with correlation <0.2 but it didn't improve the models we tried.

Data Visualization:

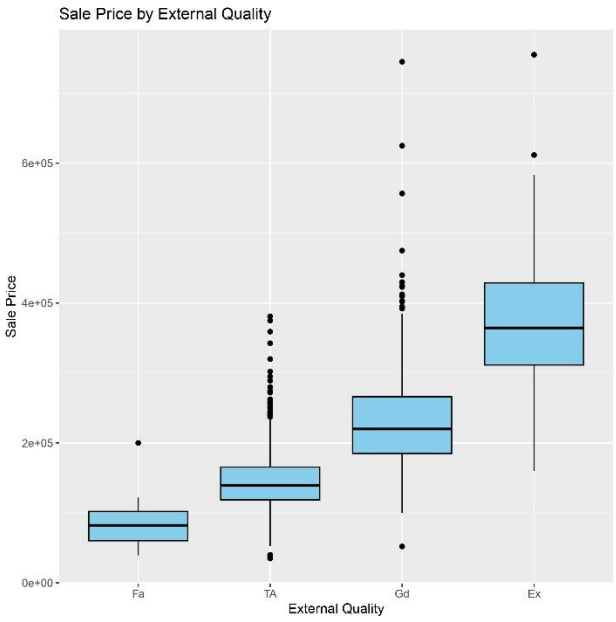




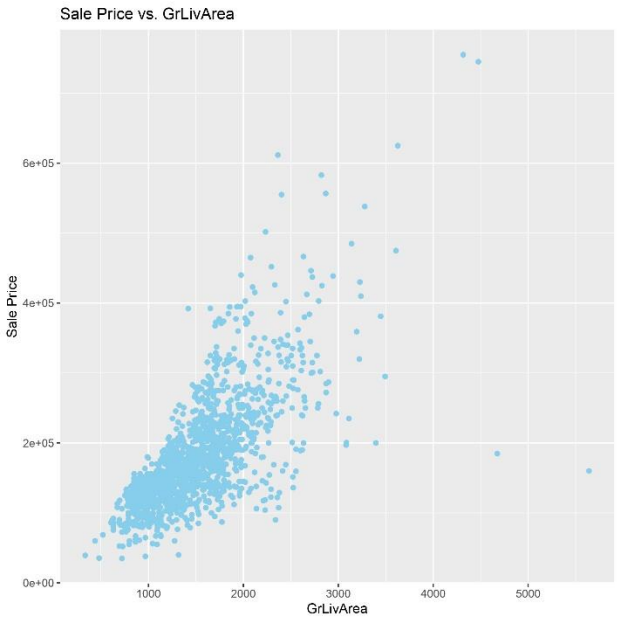
As we can see from the correlation matrix, the following values are the most correlated (Overall Quality, External Quality, GrLivArea) so we plotted them to see how they are distributed.



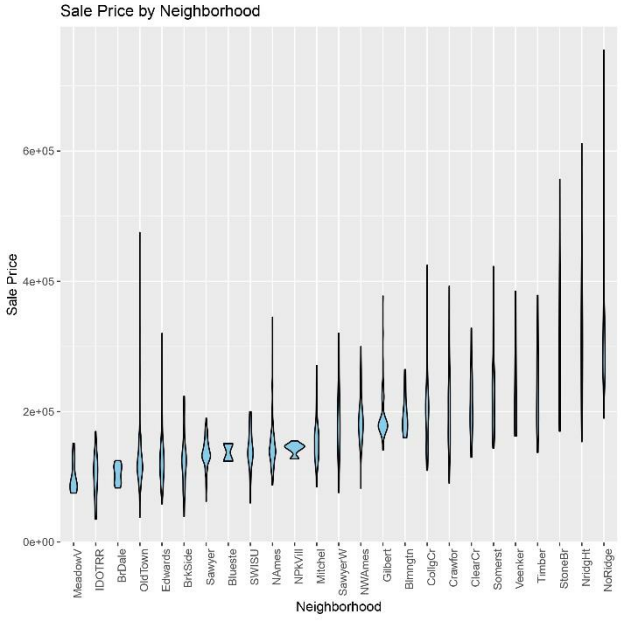
Box Plot for Overall Quality vs Sale Price



Box Plot for External Quality vs Sale Price



Scatter Plot for GrLivArea vs Sale Price



Violin Plot for each Neighborhood's Sale Price Range

Models & Experiments:

We tried 3 main models:

1. Regular Linear Regression
   a. Initially we got a score of 0.33017.
   b. We tried removing uncorrelated variables at different cutoff values and we could achieve best score of 0.16391.

2. XG Boost
   a. After changing many hyper-parameters (learning rate, max_depth, nrounds, etc.). We finally achieved a score of 0.1258.

3. SVR
   a. After trying with different hyper-parameters, we could finally achieve a score of 0.12523.
   b. Which makes this model the best one we could achieve a score with.