# Modeling Assignment 1: Getting to Know Your Data – Exploratory Data Analysis (EDA)

## 1. A Data Survey

The Data

“Data set contains information from the Ames Assessorís Office used in computing assessed values for individual residential properties sold in Ames, IA from 2006 to 2010.” (Ames Housing Data Documentation)

With this data, we will attempt to predict SalePrice of a property/building given certain information about it. We will examine some of the variables first to see which would be appropriate predictors for a simple linear regression.

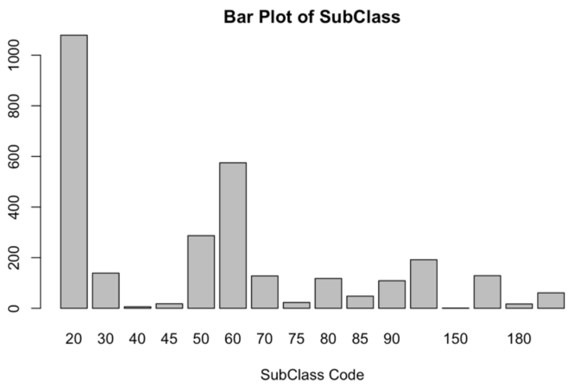
Additionally, the data documentation suggests removing properties with over 4000 sqft of above-ground living area (which produces unusual results).

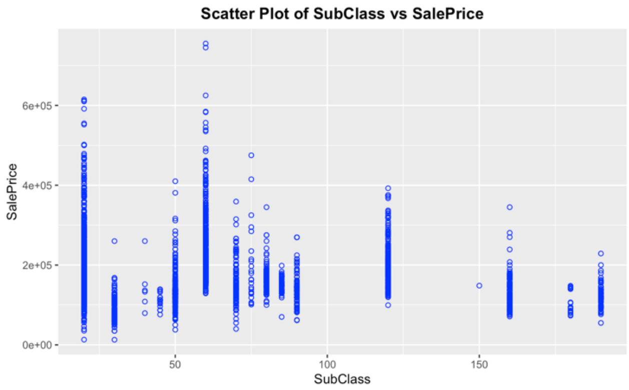
## 2. Define the Sample Population

The “Typical” Property

In order to create a linear regression to predict value of property, we must first use the Waterfall method to determine what a “typical” property is.

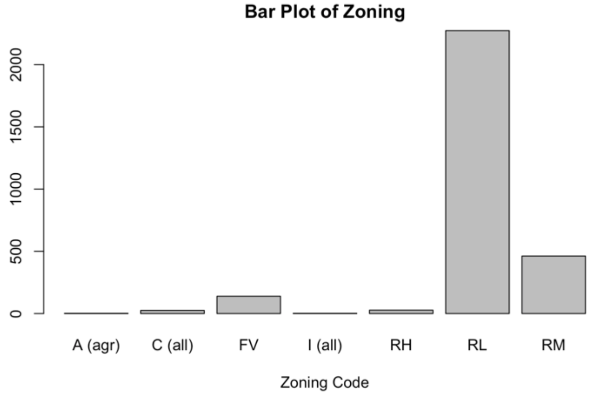
SubClass (Nominal): Identifies the type of dwelling involved in the sale

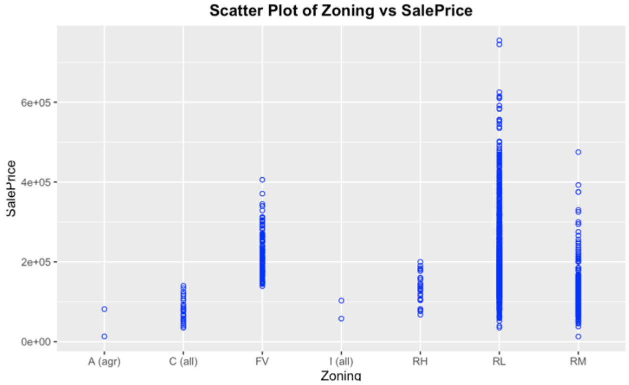




The majority of the data have SubClass codes 20 and 60 which are 1-story and 2-story (respectively) buildings made after 1946. It is evident from the scatterplot that these SubClasses have distributions with higher average SalePrice compared to other SubClasses. We will also include SubClass 50 and SubClass 120 which are 1.5-story buildings and 1-story (Planned Unit Development) buildings made after 1946 due to prevalence in the dataset.

Zoning (Nominal): Identifies the general zoning classification of the sale.





The "typical" property is majority "Residential Low Density" according to Zoning code which has a different distribution of SalePrice compared to other Zoning codes. We will also include the handful of "Residential Medium Density" Zone codes.

Using the waterfall method, we can subset the complete Ames Housing Data by 1) dropping any observations that are not SubClass = 20, 60, 50, 120 and 2) dropping any observations that are not Zone = RL, RM. We also take into account the data documentation note of dropping observations of GrLivArea > 4000. Further drop conditions that may be taken into consideration are discussed in section 4.

### The Sample Population

Note, when using a model with this data in the future, it is important to make sure that the data point used in the model falls within the "typical" categories of SubClass and Zoning as mentioned above:

*A building a 1-story building made after 1946, a 2-story building made after 1946, a 1-story PUD made made after 1946, or a 1.5 story building. AND in a Residential Low or Medium Density Zoning code.*

### The Waterfall

| Step | Number of Obs | Number of Vars |
|---|---|---|
| Ames Housing Data | 2930 | 82 |
| Remove GrLivArea > 4000 | 2925 | 82 |
| Keep SubClass 20, 60, 50, 120 | 2128 | 82 |
| Keep Zone RL, RM | 2008 | 82 |

After subsetting our dataset we are left with 2013 observations which is about 68.5% of the original population.

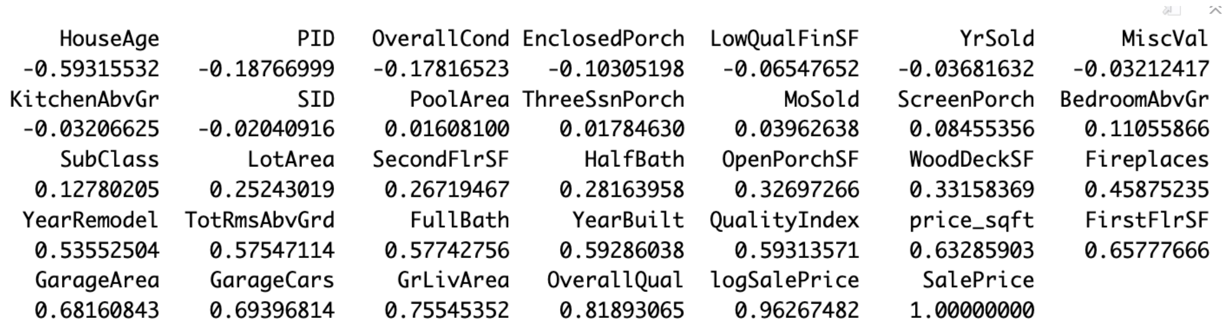
The next section investigates 20 variables as predictors which may also provide further insight on the population of interest.

## 3. A Data Quality Check

Check for impossible values and implausible values (which may or may not include outliers).

NOTE: Data quality check is in the appendix.

For Numeric variables, I chose the variables with the highest magnitude of correlation to SalePrice (so a large negative value like YearBuilt is also important). I chose not to include variables that are too similar like GarageCars and GarageArea because it is bad practice to include highly correlated predictors in a regression model.

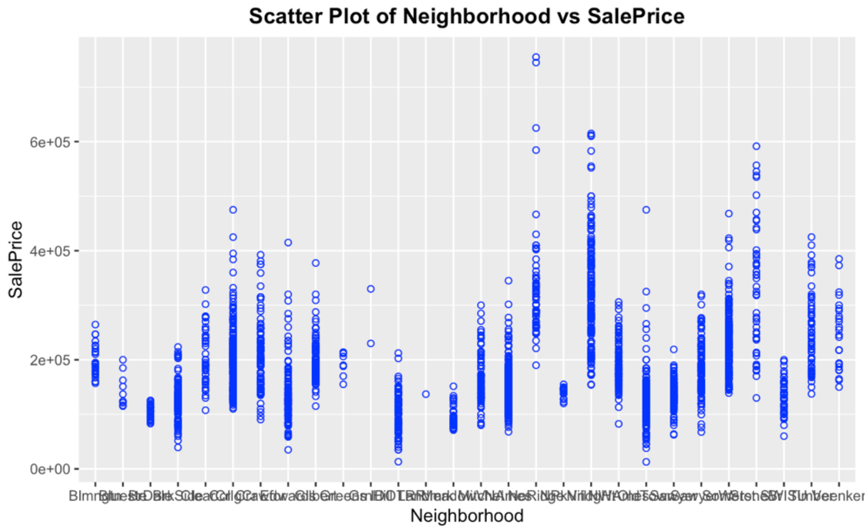


```
     HouseAge           PID   OverallCond EnclosedPorch  LowQualFinSF        YrSold       MiscVal 
  -0.59315532   -0.18766999   -0.17816523   -0.10305198   -0.06547652   -0.03681632   -0.03212417 
 KitchenAbvGr           SID      PoolArea ThreeSsnPorch        MoSold   ScreenPorch  BedroomAbvGr 
  -0.03206625   -0.02040916    0.01608100    0.01784630    0.03962638    0.08455356    0.11055866 
     SubClass       LotArea   SecondFlrSF      HalfBath   OpenPorchSF    WoodDeckSF    Fireplaces 
   0.12780205    0.25243019    0.26719467    0.28163958    0.32697266    0.33158369    0.45875235 
  YearRemodel   TotRmsAbvGrd      FullBath     YearBuilt  QualityIndex    price_sqft   FirstFlrSF 
   0.53552504    0.57547114    0.57742756    0.59286038    0.59313571    0.63285903    0.65777666 
   GarageArea    GarageCars     GrLivArea   OverallQual   logSalePrice     SalePrice 
   0.68160843    0.69396814    0.75545352    0.81893065    0.96267482    1.00000000 
```

Numerical:

1) OverallQual: Rates the overall material and finish of the house
2) GrLivArea: Above grade (ground) living area square feet
3) GarageCars: Size of garage in car capacity
4) YearBuilt: Original construction date
5) FullBath: Basement full bathrooms
6) HouseAge: Year built minus Year sold
7) TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)

For Categorical variables I looked at graphs that suggested different distributions of SalePrice for different categories, but still had an appropriate same size in each category.



For example, you can see here that the distribution of SalePrice for various neighborhoods may be higher or lower depending on the physical location of the building.

Categorical:

8) SubClass: Identifies the type of dwelling involved in the sale.
9) Zoning: Identifies the general zoning classification of the sale.
10) Neighborhood: Physical locations within Ames city limits (map available)

11) ExterQual: Evaluates the quality of the material on the exterior
12) BsmtQual: Evaluates the height of the basement
13) HeatingQC: Heating quality and condition
14) KitchenQual: Kitchen quality

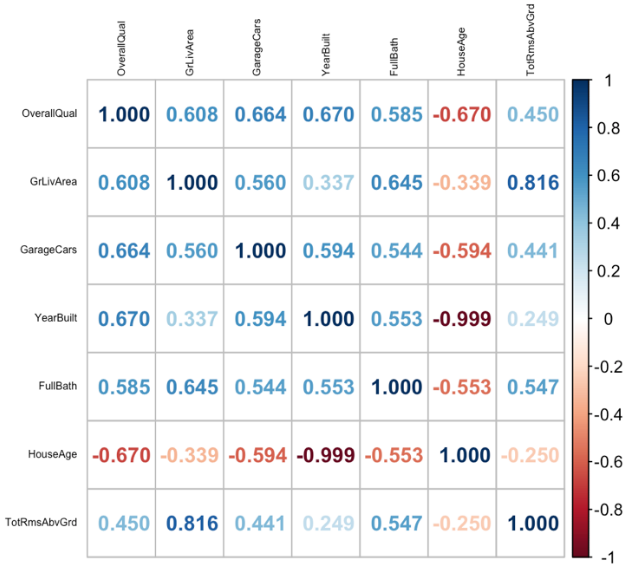
15) HouseStyle: Style of dwelling
16) Foundation: Type of foundation
17) Condition1: Proximity to various conditions
18) BldgType: Type of dwelling
19) RoofStyle: Type of roof
20) LotShape: General shape of property

The Data Quality check for each of the 20 variables is in the appendix. It is found that all data points are possible and plausible. There are a few points in categories that may not have enough data to be represented properly in a regression model. This may be addressed in additional waterfall eliminations as before testing variables in regression models.

## 4. An Initial Exploratory Data Analysis

Of the 20 variables of interest above, here is how I narrowed it down to 10:

I decided to select all of the numeric variables due to their high correlation to SalePrice. I also decided to keep SubClass, Zoning, and Neighborhood of the categorical variables due to their visually different distributions between groups. Variables like LotShape have similar average values and distributions of SalePrice across groups and would not provide much insight in the model. I also decided to exclude the quality categorical variables because the OverallQual numeric variable already provides information on this data and would be overlapping with the categorical quality variables.



| | OverallQual | GrLivArea | GarageCars | YearBuilt | FullBath | HouseAge | TotRmsAbvGrd |
|---|---|---|---|---|---|---|---|
| OverallQual | 1.000 | 0.608 | 0.664 | 0.670 | 0.585 | -0.670 | 0.450 |
| GrLivArea | 0.608 | 1.000 | 0.560 | 0.337 | 0.645 | -0.339 | 0.816 |
| GarageCars | 0.664 | 0.560 | 1.000 | 0.594 | 0.544 | -0.594 | 0.441 |
| YearBuilt | 0.670 | 0.337 | 0.594 | 1.000 | 0.553 | -0.999 | 0.249 |
| FullBath | 0.585 | 0.645 | 0.544 | 0.553 | 1.000 | -0.553 | 0.547 |
| HouseAge | -0.670 | -0.339 | -0.594 | -0.999 | -0.553 | 1.000 | -0.250 |
| TotRmsAbvGrd | 0.450 | 0.816 | 0.441 | 0.249 | 0.547 | -0.250 | 1.000 |

Upon observing the correlation matrix of numeric variables, there seems to be a high correlation between HouseAge and YearBuilt (although negative) which indicates we should drop one or the other from future models.

Additionally, from the 10 variables explored here, an excerpt from the appendix of data quality checks shows some possible outliers in the data distributions when comparing to SalePrice

| | | |
|---|---|---|
| 1) OverallQual | 1-10 (integers) | All values plausible, no outliers |
| 2) GrLivArea | 0-4000 | All values plausible, no outliers |
| 3) GarageCars | integers | All values plausible, possible outliers are 5 pts (4 cars) |
| 4) YearBuilt | year | All values plausible, possible outliers before (1910) |
| 5) FullBath | 0-3 | All values plausible, possible outliers are 4 pts (0 baths) |
| 6) HouseAge | 0-135 | All values plausible, possible outliers ~20 pts (over 100 years old) |
| 7) TotRmsAbvGrd | 2-12 | All values plausible, possible outliers (2, 12 rooms) |
| 8) SubClass | 4 subclasses | All values plausible |
| 9) Zoning | 2 zones | All values plausible |
| 10) Neighborhood | 26 neighborhoods | All values plausible |

The initial EDA shows several things:

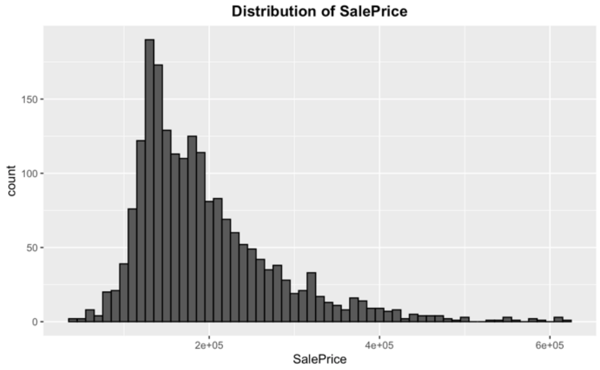
Numeric variables: Most properties were made between 1910-2010 and are under 100 years old. Most have and have 0-3 car garages, 1-3 full baths, and 3-11 rooms above ground level.

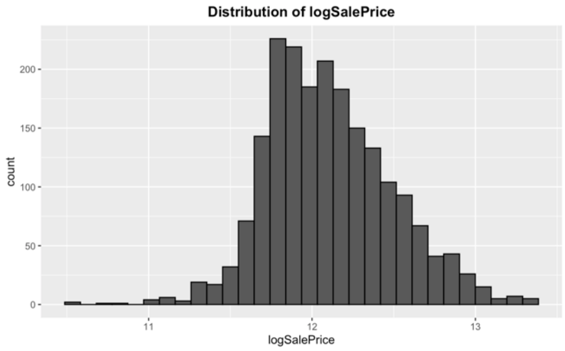
Categorical: all values are plausible and possible, with representative values from each category. SubClass and Zoning observations are already stated in the “Defining the Sample Population” section in Secion 1.

Of the 10 variables explored in the initial EDA, here are the key findings from some of the numeric and categorical variables and their relationships with logSalePrice (see section 5).

## 5. An Initial Exploratory Data Analysis for Modeling

### The Response Variable: SalePrice vs logSalePrice


Distribution of SalePrice

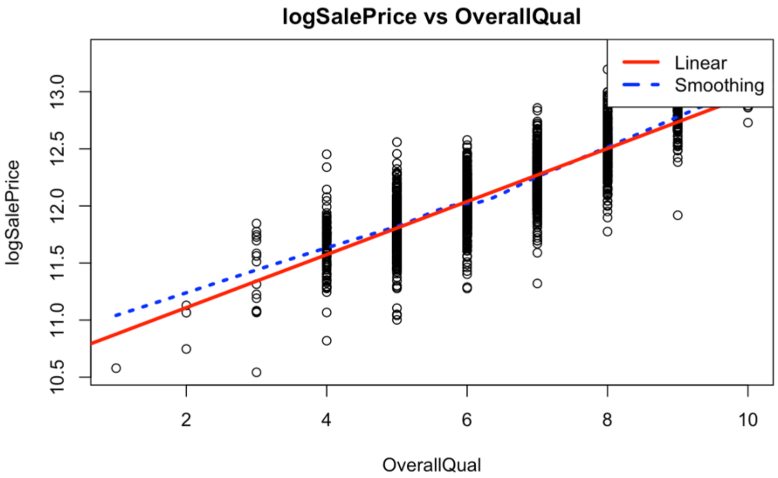

Distribution of logSalePrice

Note that we use logSalePrice in models opposed to SalePrice because the original data is skewed right and is not an approximately normal distribution, whereas the logSalePrice has a more normal distribution.

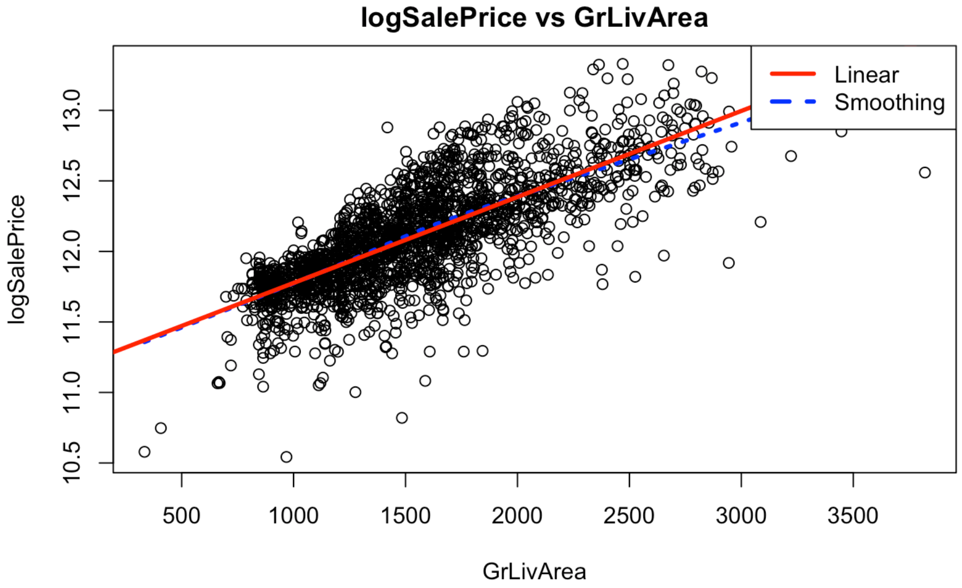
Here is the initial exploratory analyses between a few of the predictor variables and their relationship with logSalePrice,

Numeric Variables:

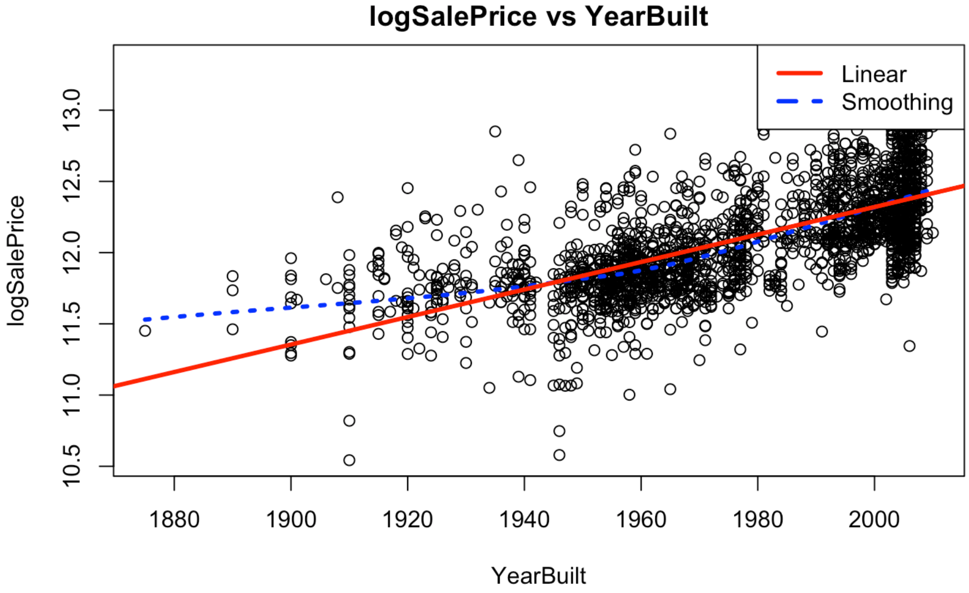
In this EDA, numeric variables are observed using scatterplots with smoothing to show linear and non-linear relationships between the predictors and logSalePrice


logSalePrice vs OverallQual

As assumed, high quality buildings will be sold for higher price. This makes sense, as it may cost more to keep a building in top condition, and buyers will pay more for nicer units.



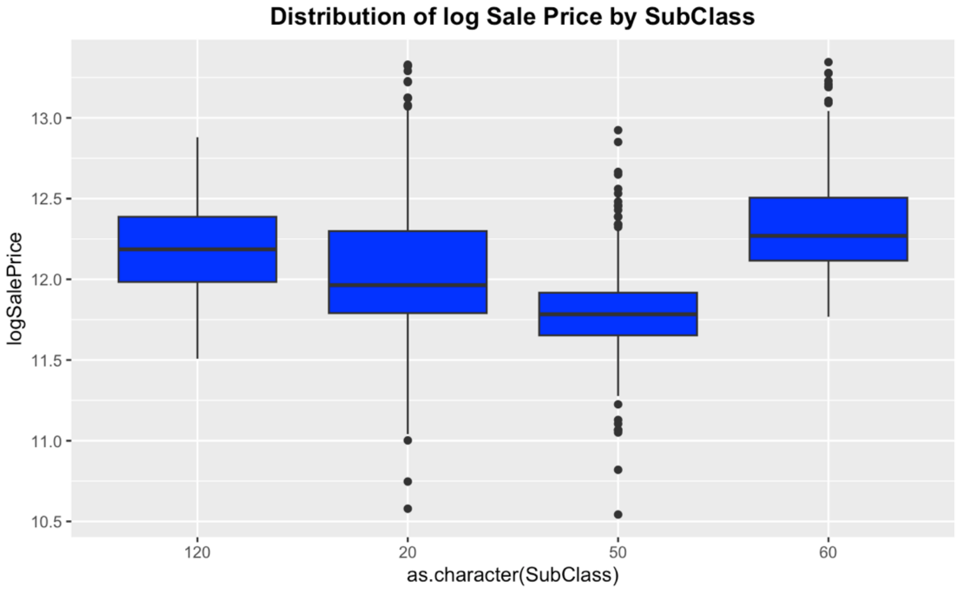
Generally, buildings with more square footage of living area will sell for a higher price. This makes sense, as people will pay more for a larger living space, and sellers need to cover more costs like utilities for larger space.



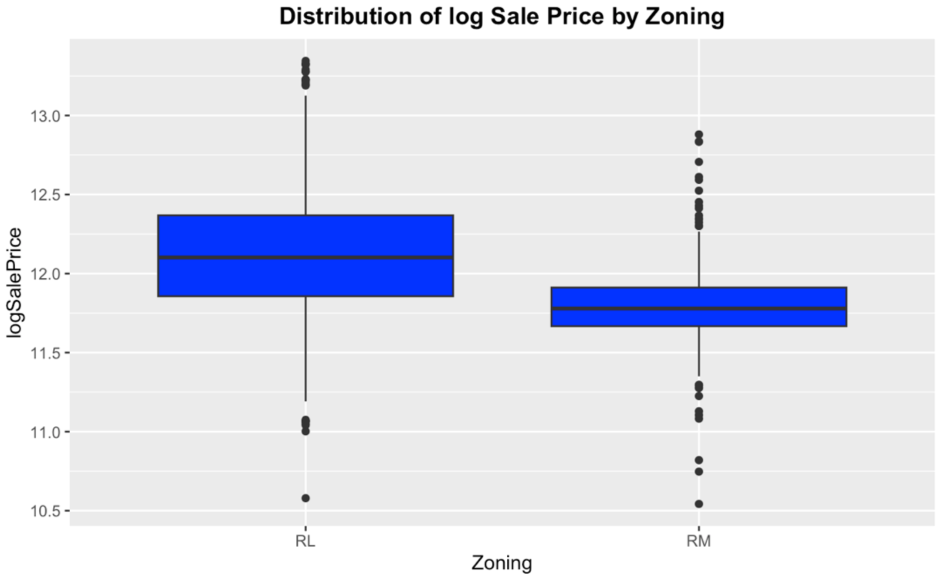
Although slightly less linear of a relationship, buildings built more recently will sell for a higher price. This may possibly due to inflation of currency and/or the depreciation of property value over time.

Categorical Variables:

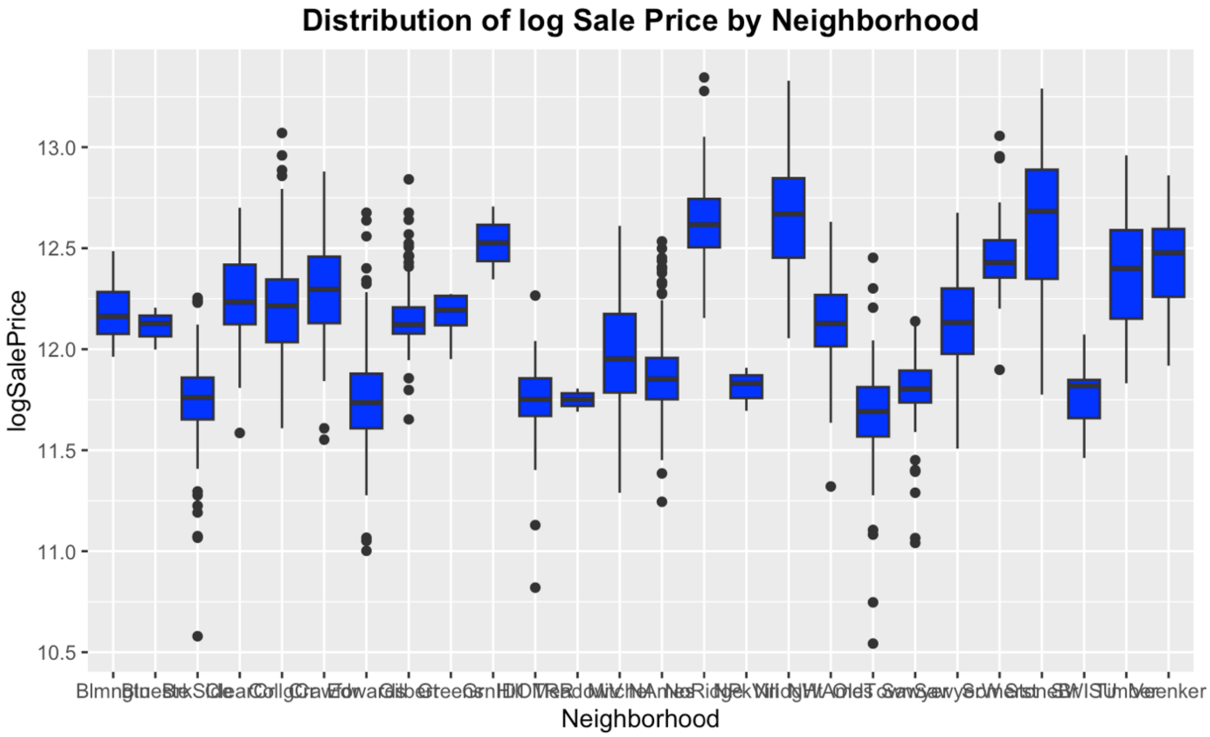
In this EDA, categorical variables are observed using boxplots to show relationships between the different levels in predictors and logSalePrice.



It appears that SubClass 20 and 50 generally have lower SalePrice than SubClass 60 and 120. This indicates that Planned Unit Development (120) and 2-stories (60) will sell higher that non-PUD (20) and 1.5-stories (50).



Low Density residential zones will typically sell higher than Medium Density residential zones. Lower Density residential zones may have a higher cost of living and larger units meant for less people which is why there are less people and a higher SalePrice.



Different neighborhoods have very different SalePrice distributions which can be due to their access to various different resources in the area such as education, grocery stores, transportation, etc.

6. Summary/Conclusions:

From the exploratory data analysis, there are a few key points and concerns to note moving forward.

In order to use this data for a regression model, any data used in the model must fall under the sample population parameters as outlined in section 1. This means that the model will be limited to certain types of buildings in certain populated areas, and cannot be generalized for all types of buildings everywhere. Additional parameters can be added to the model according to Section 3 as seen fit.

Many of the variables in the dataset are highly correlated. Predictors that are highly overlapping and have high correlations with each other must be removed from any regression models that are built. Doing so can result in inaccurate models that overemphasize variables that may have more or less influence together that they truly do individually.

Transformations may be necessary from numeric predictors that are non-normally distributed. Just like logSalePrice, this variable was transformed to have a more normal distribution. Just note that for interpretation purposes, any results must be converted back into non-log SalePrice.

Overall, take caution when proceeding further with models, but overall, this dataset holds promising potential to determine important factors in predicting SalePrice of properties in Ames, IA.

## Appendix

Data Quality Results for 20 Variables

| Variable | Conditions | Results |
|---|---|---|
| 11) OverallQual | 1-10 (integers) | All values plausible, no outliers |
| 12) GrLivArea | 0-4000 | All values plausible, no outliers |
| 13) GarageCars | integers | All values plausible, possible outliers are 5 pts (4 cars) |
| 14) YearBuilt | year | All values plausible, possible outliers before (1910) |
| 15) FullBath | 0-3 | All values plausible, possible outliers are 4 pts (0 baths) |
| 16) HouseAge | 0-135 | All values plausible, possible outliers ~20 pts (over 100 years old) |
| 17) TotRmsAbvGrd | 2-12 | All values plausible, possible outliers (2, 12 rooms) |
| 18) SubClass | 4 subclasses | All values plausible |
| 19) Zoning | 2 zones | All values plausible |
| 20) Neighborhood | 26 neighborhoods | All values plausible |
| 21) ExterQual | 4 qualities | All values plausible |
| 22) BsmtQual | 4 qualities | All values plausible |
| 23) HeatingQC | 5 qualities | All values plausible, possible outliers (only 1 Po) |
| 24) KitchenQual | 5 qualities | All values plausible, possible outliers (only 1 Po) |
| 25) HouseStyle | 6 styles | All values plausible, possible outliers (1 2.5Unf, 1 SFoyer, 2 SLvl) |
| 26) Foundation | 6 types | All values plausible, possible outliers (1 Stone, 4 Wood) |
| 27) Condition1 | 9 conditions | All values plausible |
| 28) BldgType | 4 types | All values plausible, possible outliers (2 fmCon) |
| 29) RoofStyle | 6 styles | All values plausible, possible outliers (3 Shed) |
| 30) LotShape | 4 shapes | All values plausible, possible outliers (13 IR3) |