QUESTIONS for Professor Burstein for Feedback in addition to the write-up:

- Handling of Nulls (on Jupyter Notebook):
    - Would you recommend a different way when it comes to handling null values? (Perhaps specifically dealing with the null values for LotFrontage)?
- How should we encode other categorical variables?
    - Dummy/OneHotEncoder?
    - TfidfVectorizer()?
- EDA (on Jupyter Notebook vs PDF):
    - Can we keep all graphs on the Jupyter Notebook and pick and choose the best visualization for the PDF writeup or should we also be more selective in the visualizations on the Jupyter Notebook?
- Modeling:
    - Other than using Linear Regression, Kfolds, and playing around with features used in the Linear Regression models should we also attempt different models?
        - Others I've heard of are SVM (in this course) and in past courses Decision Tree Regression, KNN Regression Models
- Could you provide us with overall feedback on the write-up so far?

Thank you in advance for your feedback!

Discussion Writeup Instructions

You are expected to include a discussion of **no more than 5 pages** that explains the purpose of your Work. The discussion should include but is not limited to:

- A description of the problem you tried to solve
- The files you included and what they do
- Methods and Techniques you used along with "brief" explanations of how they should work
- What extensions you implemented and how they make your project unique and valuable
- Conclusions of the project, including what you've learnt

MET CS 521 Final Project:
Names: Daniela Demaestri and Alex Mao
GitHub usernames: ddemaest-bu and alexmao0501

Final Project: Predicting Housing Prices (Regression)

Objective

For this final project, we have selected to work with a housing dataset. We started by cleaning the data and handling the missing data, selecting the relevant variables and performing some exploratory data analysis to better understand the data, and finally fitting our data to different regression models in order to predict sale price based on different house features.

Achieving our objectives was split into three phases:

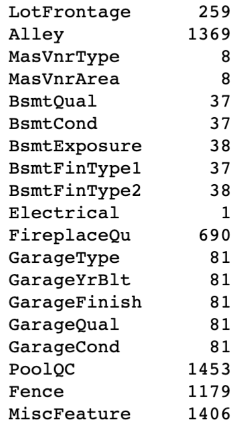
1. Cleaning and Processing the Data: In this phase we handle nulls, encode some categorical variables, and remove some features to reduce the complexity.
2. Exploratory Data Analysis (EDA): our main aim in this phase was to have a better understanding of the features involved in our data. It might be possible that some are left behind but we focused on determining the features that have the highest correlation towards Sale Price in the hopes that it could later help us produce a model to predict sale price.
3. Regression Modeling: We will implement Regression models to predict a possible SalePrice (label) of the house using different features

Cleaning and Processing the Data:

Handling null values: Our initial focus was to handle null values within the dataset. Not only is this an important step in order to get a better understanding of the dataset, but also because many machine learning algorithms and models fail when the dataset contains missing values.

We started looking at the null values within our target variable, SalePrice, because our main objective was to predict SalePrice and identify important housing features in relation to SalePrice. There were 1459 null values in SalePrice (49.98% of our data). Given our objective, we decided to drop these nulls.

After removing the rows with null values in SalePrice, we were left with 1460 rows. Now we were interested in seeing how many nulls there were in the remainder of the dataset, shown below:



```
LotFrontage      259
Alley           1369
MasVnrType         8
MasVnrArea         8
BsmtQual          37
BsmtCond          37
BsmtExposure      38
BsmtFinType1      37
BsmtFinType2      38
Electrical         1
FireplaceQu      690
GarageType        81
GarageYrBlt       81
GarageFinish      81
GarageQual        81
GarageCond        81
PoolQC          1453
Fence           1179
MiscFeature     1406
```

We decided to drop the columns that didn’t seem interesting given the high amount of null values. The features we removed included: PoolQC, MiscFeature, Alley, Fence, FireplaceQu, and MiscVal. We also deduced that the reason behind the high quantity of nulls for these columns is due to the fact these features may not be present in all houses. In other words not all houses have pools, alleys, fences, fireplaces.

Looking at the data description we also noticed that there were nine features related to describing/rating basements. To limit the scope for basement features, we decided to keep BsmtQual, BsmtCond, BsmtExposure, and TotalBsmtSF as they seemed to indicate a more generalized description of the basement, and decided to drop the other basement features.

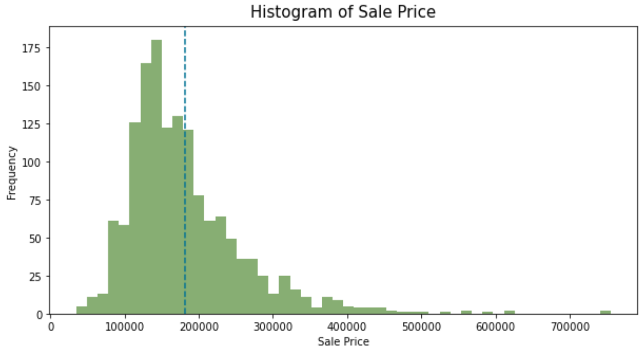
We then handled the remaining null values in the dataset by replacing nulls with the mode for Electrical and MasVnrType, and by replacing the nulls with 0 for MasVnrArea (if MasVnrType was None) and LotFrontage **change LotFrontage once we get feedback from Prof).

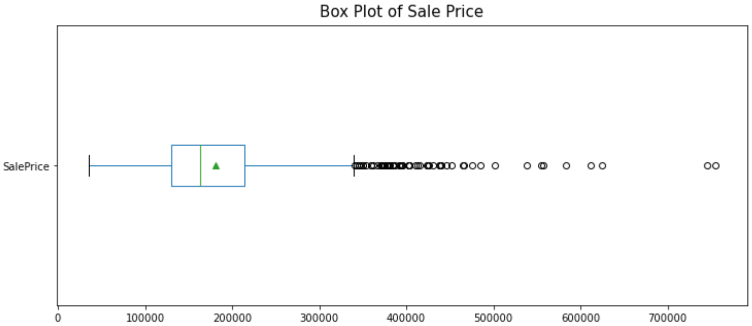
Encoding categorical variables: Looking at the data description a few categorical variables had an implicit order. Hence, we decided to encode these categorical variables as ordered numbers (similar to how OverallCond was already encoded). The variables we manually encoded numerically were: LotShape, Utilities, AllPub, LandSlope, ExterQual, ExterCond, BsmtQual, BsmtCond, HeatingQC, KitchenQual, FireplaceQu, GarageFinish, GarageQual, and GarageCond. Just to showcase one example, we encoded BsmtCond as follows: 5 - Ex, 4 - Gd, 3 - TA, 2 - Fa, 1 - Po, 0 - NA.

***add other encoding we may do categorical variables (select important categorical variables to encode - maybe neighborhood?)*

Exploratory Data Analysis (EDA):

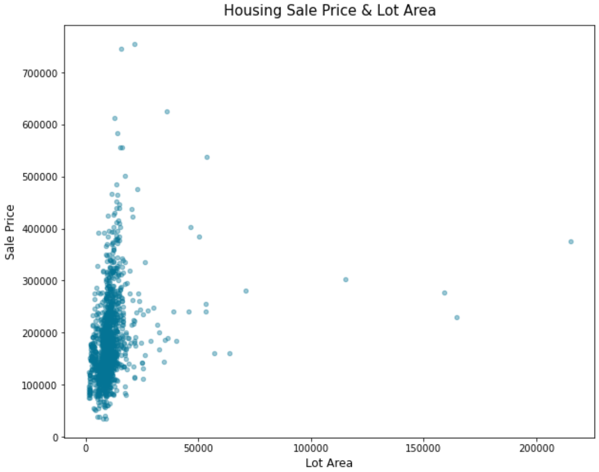
Taking a closer look at the target variable - SalePrice: Looking at the distribution of the target variable, SalePrice, is important especially when it comes to Linear Regression because of the assumptions the model makes: linearity, homoscedasticity, independence, and normal distribution. Below we observe the distribution of SalePrice:

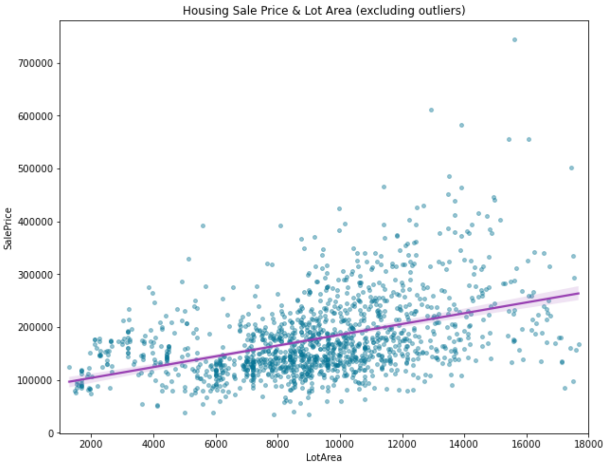




Looking at the distribution of SalePrice (either looking at the histogram or boxplot), we can see the graph is right-skewed. At first we decided to use linear regression ignoring the non-normal distribution; however (**hoping to have time to perhaps use log transformation for one of our models to compare the impact).

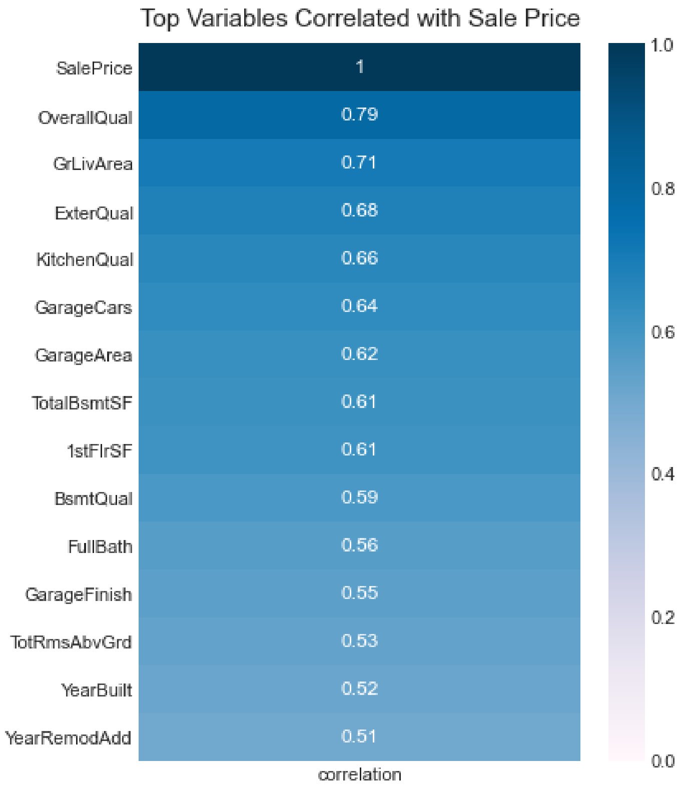
SalePrice vs. LotArea: The next question we wanted to explore was: What is the relationship between SalePrice and LotArea? The reasoning behind our question is due to our assumption that houses on larger lots tend to have a higher price or property value than similar houses on smaller lots. We explore this relationship below:

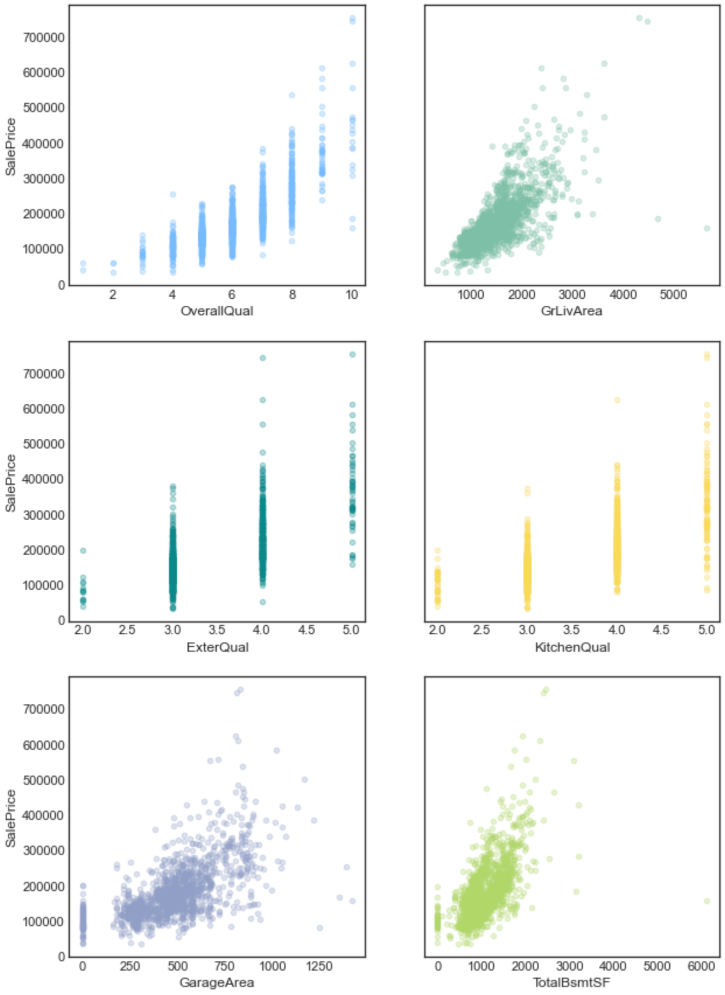




In the left graph, we observe quite a few outliers when it comes to the Lot Area. By removing the outliers (right graph), we can take a better look at the relationship between SalePrice and LotArea. To determine the outliers, we used the interquartile range (IQR) approach where we find the upper and lower bound. Outliers are values that are above and below the dataset's normal range (above the upper bound and below the lower bound). When looking at the graph without the outliers, we can observe a positive correlation between SalePrice and LotArea.

Which are the top (numerical) housing features that are correlated with SalePrice?





Regression Models: