Capstone Project 1

### (Predicting the house prices using Supervised Machine Learning Algorithms)

**Problem statement**:

The aim of the project is “Predicting the house prices using linear regression model”. This project helps people who are looking to buy house with preferences and predicting the price based on these preferences. It can also be used by real estate agents for better price estimation of houses. Generally, people approach brokers for buying houses. If someone is new to that particular area, brokers might cheat the customers by quoting more price for the house compared to the market price. This project is the way to go for them.

**Descriptive Data Analysis**:

The Dataset is taken from the kaggle competition titled “House Prices: Advanced Regression Techniques”. The target of the projects is to predict the house price using multiple regression techniques. The data contains 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa.

MSSubClass: Identifies the type of dwelling involved in the sale.

MSZoning: Identifies the general zoning classification of the sale.

LotFrontage: Linear feet of street connected to property.

LotArea: Lot size in square feet

Street: Type of road access to property

Alley: Type of alley access to property

LotShape: General shape of property

Utilities: Type of utilities available

LotConfig: Lot configuration

LandSlope: Slope of property

Neighborhood: Physical locations within Ames city limits.

Condition1: Proximity to various conditions.

Condition2: Proximity to various conditions (if more than one is present)

BldgType: Type of dwelling.

HouseStyle: Style of dwelling

OverallQual: Rates the overall material and finish of the house

OverallCond: Rates the overall condition of the house

YearBuilt: Original construction date

YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)

RoofStyle: Type of roof

Exterior1st: Exterior covering on house

Exterior2nd: Exterior covering on house (if more than one material)

MasVnrType: Masonry veneer type

MasVnrArea: Masonry veneer area in square feet

ExterQual: Evaluates the quality of the material on the exterior

ExterCond: Evaluates the present condition of the material on the exterior

Foundation: Type of foundation.

BsmtQual: Evaluates the height of the basement

BsmtCond: Evaluates the general condition of the basement

BsmtExposure: Refers to walkout or garden level walls

BsmtFinType1: Rating of basement finished area

BsmtFinSF1: Type 1 finished square feet

BsmtFinType2: Rating of basement finished area (if multiple types)

BsmtFinSF2: Type 2 finished square feet

BsmtUnfSF: Unfinished square feet of basement area

TotalBsmtSF: Total square feet of basement area

Heating: Type of heating

HeatingQC: Heating quality and condition

CentralAir: Central air conditioning

Electrical: Electrical system

1stFlrSF: First Floor square feet.

2ndFlrSF: Second floor square feet

LowQualFinSF: Low quality finished square feet (all floors)

GrLivArea: Above grade (ground) living area square feet

BsmtFullBath: Basement full bathrooms

BsmtHalfBath: Basement half bathrooms

FullBath: Full bathrooms above grade

HalfBath: Half baths above grade

Bedroom: Bedrooms above grade (does NOT include basement bedrooms)

Kitchen: Kitchens above grade

KitchenQual: Kitchen quality

TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)

Functional: Home functionality (Assume typical unless deductions are warranted)

Fireplaces: Number of fireplaces

FireplaceQu: Fireplace quality

GarageYrBlt: Year garage was built

GarageFinish: Interior finish of the garage

GarageCars: Size of garage in car capacity

GarageArea: Size of garage in square feet

GarageQual: Garage quality

GarageCond: Garage condition

PavedDrive: Paved driveway

WoodDeckSF: Wood deck area in square feet

**Data Story:**

I drew scatterplot of price vs most of the features from the dataset.

Price vs bedrooms scatterplot shows there is a positive correlation.

Price vs bathrooms scatterplot also shows there is a positive correlation.

Price vs GrLivArea plot also shows there is a positive correlation.

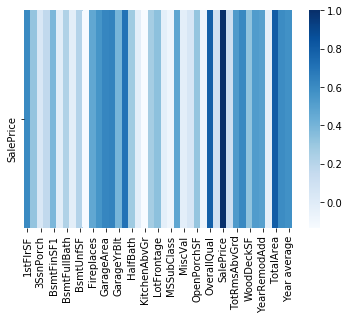
Price vs 3SsnPorch scatter plot shows a weak negative correlation.

Price vs floors scatterplot shows a positive correlation.

Price vs overall Quality scatterplot shows a positive correlation.

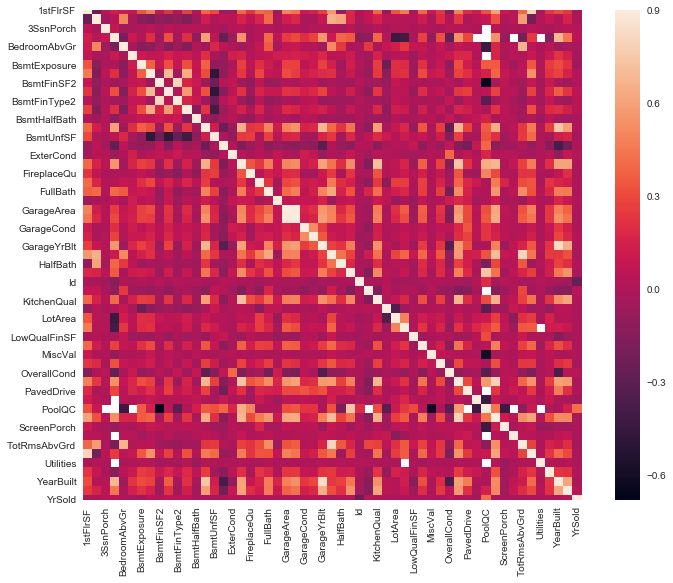
Price vs GarageArea plot also shows there is a positive correlation.

Price vs KitchenAvbrGr plot shows a weak negative correlation

****

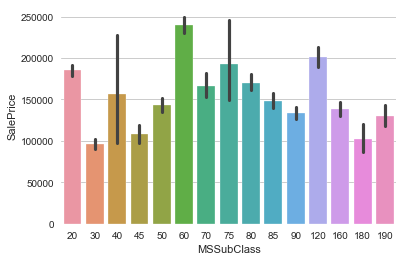
**Heatmap of the features:**

I drew the below heatmap using seaborn library to illustrate the correlation between the different features.

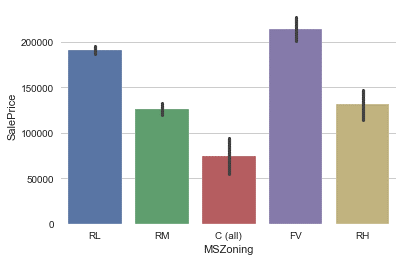
****

**Plotting the relationship between the categorical features and the target:**

Sale price vs MS Subclass**:**



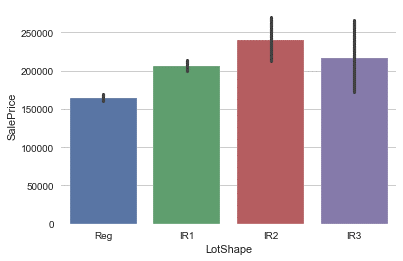
Sale price vs MS Zoning**:**



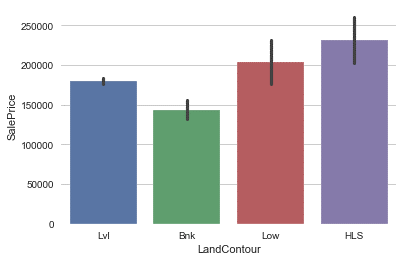
Sale price vs Street**:**



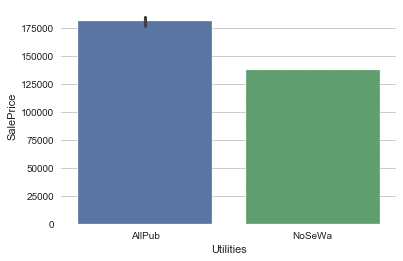
Sale price vs Lot Shape**:**



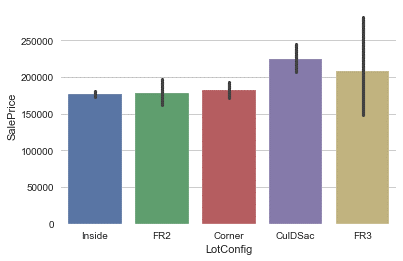
Sale price vs Land Contour**:**



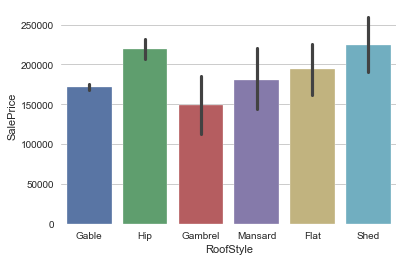
Sale price vs Utilities**:**



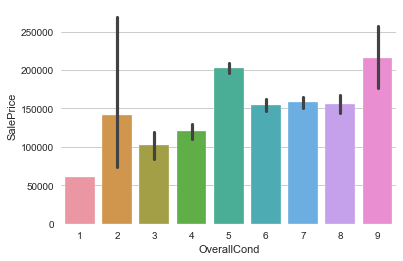
Sale price vs Lot Config**:**



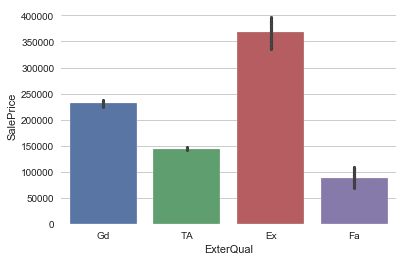
Sale price vs Roof Style**:**



Sale price vs Overall Condition**:**

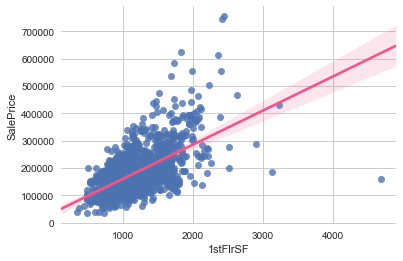


Sale price vs External Quality**:**

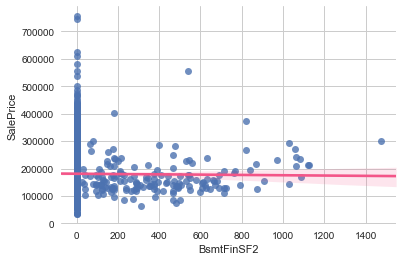


**Plotting the relationship between the numerical features and the target:**

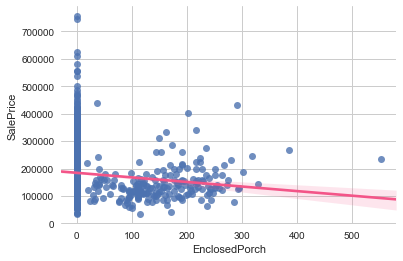
Sale price vs First Floor Area**:**



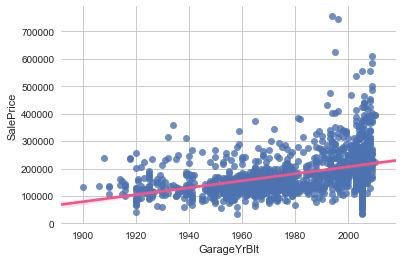
Sale price vs Type 2 finished square feet**:**



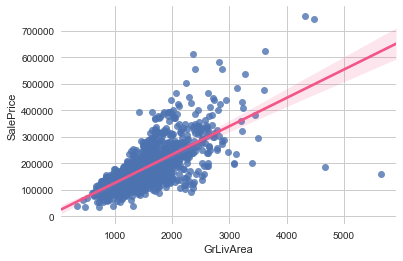
Sale price vs Enclosed porch area**:**



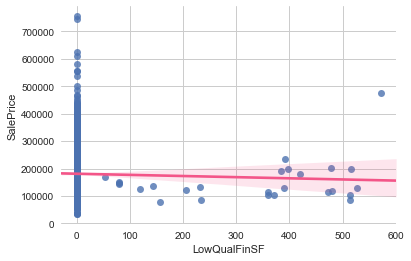
Sale price vs Garage Year Built**:**



Sale price vs Living area**:**



Sale price vs Low quality finished square feet (all floors)**:**



**Data Wrangling:**

.

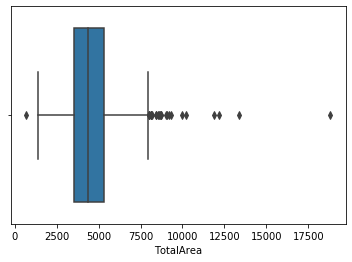
Exploring the data available to check how many non-null variables are present. Checking the correlation of all variables to the final target variable using the correlation method. Using the .corr() the correlated output is plotted on the heatmap to check the correlation. Major columns contributing to the target house price can be identified from the heatmap.

# Identifying the columns with null values and dropping the columns with null values more than 50% using the dropna().Data cleaning to be performed on the remaining data wherever the null values are present. The columns are segregated based on the datatype of the column. The data type of the column can be an object or numerical. If the object is numerical its filled with 0 or relevant number (Mean or Median), if the datatype is object, then its filled with ‘None’ using fillna(). For minor data missing columns the data can be filled using method ffill or bfill.

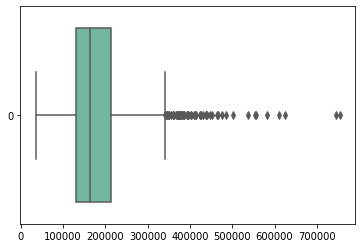
Once the missing values are filled, then plotting scatter plot of the columns with higher correlation to the target price to identify the outliers and removing the same. Visualizing the outliers using the scatter plot and box plot. Function has been written to remove the outliers which are greater than 1.5 IQR (Inter Quartile Range) and less than 1.5 IQR. After removing all the outliers, correlation is again calculated for better understanding.

**Some of the boxplots to identify the outliers:**

Total Area:



Sale Price

:

**Exploratory Data Analysis:**

The below statistical inferences are made on the train dataset.

Performed hypothesis test to check the significant relationship between target variable Price and No. of bathrooms. The p-value for the hypothesis test is less than the level of significance 0.05, so we reject the null hypothesis. So we support that there is a correlation between a number of bathrooms and price.

I also conducted a hypothesis test to check the correlation between a total area and price. The p-value for the hypothesis test is less than the level of significance 0.05, so we reject the null hypothesis and suggest that there is a correlation between a total area and price.

Similarly, I conducted a hypothesis test to check the correlation between overall quality and price. The p-value for the hypothesis test is less than the level of significance 0.05, so we reject the null hypothesis and suggest that there is a correlation between overall quality and price.

Conducted a hypothesis test to check the correlation between a three season porch area in square feet and price. The p-value for the hypothesis test is greater than the level of significance 0.05, so we accept the null hypothesis and suggest that there is no correlation between three season porch area in square feet and price.

I also conducted a hypothesis test to check if there is no statistical importance between mean house price and a number of bedrooms less than 4 and greater than 4. The p-value for the test was greater than the level of significance 0.05, so we fail to reject the null hypothesis. This suggests us that there is no statistical importance between mean house price and a number of bedrooms less than 4 and greater than 4.

**Indepth Analysis using Supervised Machine Learning:**

To predict the house price of the given set of data, I used supervised machine learning techniques. The aim of the model is to obtain the minimum RMSE (Root Mean Square Error) on the test data. I build multiple regression models to predict the house price and calculated RMSE for each model on the Test Data. I tested the model on the test data in Kaggle and received a RMSE 0.1301.

Terminologies used:

Given data with dependent variable is divided as train set and test set and data without dependent variable is called test data.

First, I combined the test data and train data to perform Data Pre-Processing and Exploratory analysis. Once the data is clean and processed, the data is split into train data and test data. Then the train data is divided into independent as x and dependent as y. There are different methods to measure the performance of the regression models such as Mean Squared Error, Root Mean Squared Error, R-Squared score and Mean absolute deviation. I used RMSE (Root Mean Square Error) as the metric for measuring the performance of the model. I normalized the train and test data using robust scaler transformation. I again split the training data into 80:20 as training and testing set. This model learns on the 80% of the train data and tests on the 20% of the unseen data before finally testing on the real test data.

Secondly, I build the Linear Regression as the base model with default parameters, and I fit the train set to the model and predicted the house price on the test set and calculated metrics such as RMSE, R2\_score to check the performance of the model. The RMSE when predicted house price on the test set is 0.1545 and when submitted in the kaggle website on the test data with this model is 0.224.The coefficients of the model can be obtained using” .coef\_ “and intercept using” .intercept\_”. Then I used model Ridge regularization on the dataset with hyperparameter optimization. Using cross validation techniques such as GridSearchCV and RandomSearchCV to tune the hyperparameters to improve the performance of the model. Then the same metrics are used on the Ridge regression and obtained RMSE of 0.102 on the test set and 0.15.

**Results:**

Using the Bayesian Ridge regression model for fitting and predicting the price of the house gives the best RMSE score of 0.130 on the test data when submitted in Kaggle Website.