**PREDICTING THE SALES PRICE BASED ON BOSTON HOUSING DATA**

**JAYACHANDER REDDY KANDAKATLA**

**12th MAY 2019**

**1. INTRODUCTION**

* 1. **Background:**

Meeting daily meal, having clothes to wear and house to live are the three basics needs of life for every human being on the planet. Some buy houses to live and some other to do real estate business. So, Buying a new house has become very important thing in ones life. One would spend nearly handful amount of time in assuming what kind of house is better for his/her own lifestyle and budget.

* 1. **Problem**

If someone is planning to buy a house in coming years and in one reason or the other housing prices are going to be sky high he may lose chance to buy. On the other side investing in Housing has also become a lucky draw sometimes, good investments turn out to be bad investments sometime decisions like Amazon second headquarters announcement in near New York Area but later change in the decision results in plunge down quickly in real estate.

* 1. **Curious**

Being curious on this case study Preparing data from various real estate companies and by applying machine learning tools and techniques predicting the housing prices may solve the problems I have described above to some extent. This study will benefit many household and real estate brokers and businessmen. Here I was interested to study Boston City Housing Data.

**2. DATA ACQUISITION AND ANALYSIS**

**2.1 Data Sources:**

There are number of sources to get housing data One is Kaggle which is very good source of various datasets and stage for various competitions. I have acquired Boston Housing Data from Kaggle and referred to Wikipedia and other others for any missing data.

**2.2 Description of Data:**

I have got this Training and Test data both files were separately given by Kaggle and I have below Features describing the each house corelated with Sales price, I don’t think Each and every features adds an extra buck for the sales price but lets do some visualization on historical taste of each individual investing on their dream house. The 81 features are described clearly below.

* SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.
* MSSubClass: The building class
* MSZoning: The general zoning classification
* LotFrontage: Linear feet of street connected to property
* LotArea: Lot size in square feet
* Street: Type of road access
* Alley: Type of alley access
* LotShape: General shape of property
* LandContour: Flatness of the property
* Utilities: Type of utilities available
* LotConfig: Lot configuration
* LandSlope: Slope of property
* Neighborhood: Physical locations within Ames city limits
* Condition1: Proximity to main road or railroad
* Condition2: Proximity to main road or railroad (if a second is present)
* BldgType: Type of dwelling
* HouseStyle: Style of dwelling
* OverallQual: Overall material and finish quality
* OverallCond: Overall condition rating
* YearBuilt: Original construction date
* YearRemodAdd: Remodel date
* RoofStyle: Type of roof
* RoofMatl: Roof material
* Exterior1st: Exterior covering on house
* Exterior2nd: Exterioubr covering on house (if more than one material)
* MasVnrType: Masonry veneer type
* MasVnrArea: Masonry veneer area in square feet
* ExterQual: Exterior material quality
* ExterCond: Present condition of the material on the exterior
* Foundation: Type of foundation
* BsmtQual: Height of the basement
* BsmtCond: General condition of the basement
* BsmtExposure: Walkout or garden level basement walls
* BsmtFinType1: Quality of basement finished area
* BsmtFinSF1: Type 1 finished square feet
* BsmtFinType2: Quality of second finished area (if present)
* 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: Number of bedrooms above basement level
* Kitchen: Number of kitchens
* KitchenQual: Kitchen quality
* TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
* Functional: Home functionality rating
* Fireplaces: Number of fireplaces
* FireplaceQu: Fireplace quality
* GarageType: Garage location
* 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
* OpenPorchSF: Open porch area in square feet
* EnclosedPorch: Enclosed porch area in square feet
* 3SsnPorch: Three season porch area in square feet
* ScreenPorch: Screen porch area in square feet
* PoolArea: Pool area in square feet
* PoolQC: Pool quality
* Fence: Fence quality
* MiscFeature: Miscellaneous feature not covered in other categories
* MiscVal: $Value of miscellaneous feature
* MoSold: Month Sold
* YrSold: Year Sold
* SaleType: Type of sale
* SaleCondition: Condition of sale

**3. METHODOLOGY**

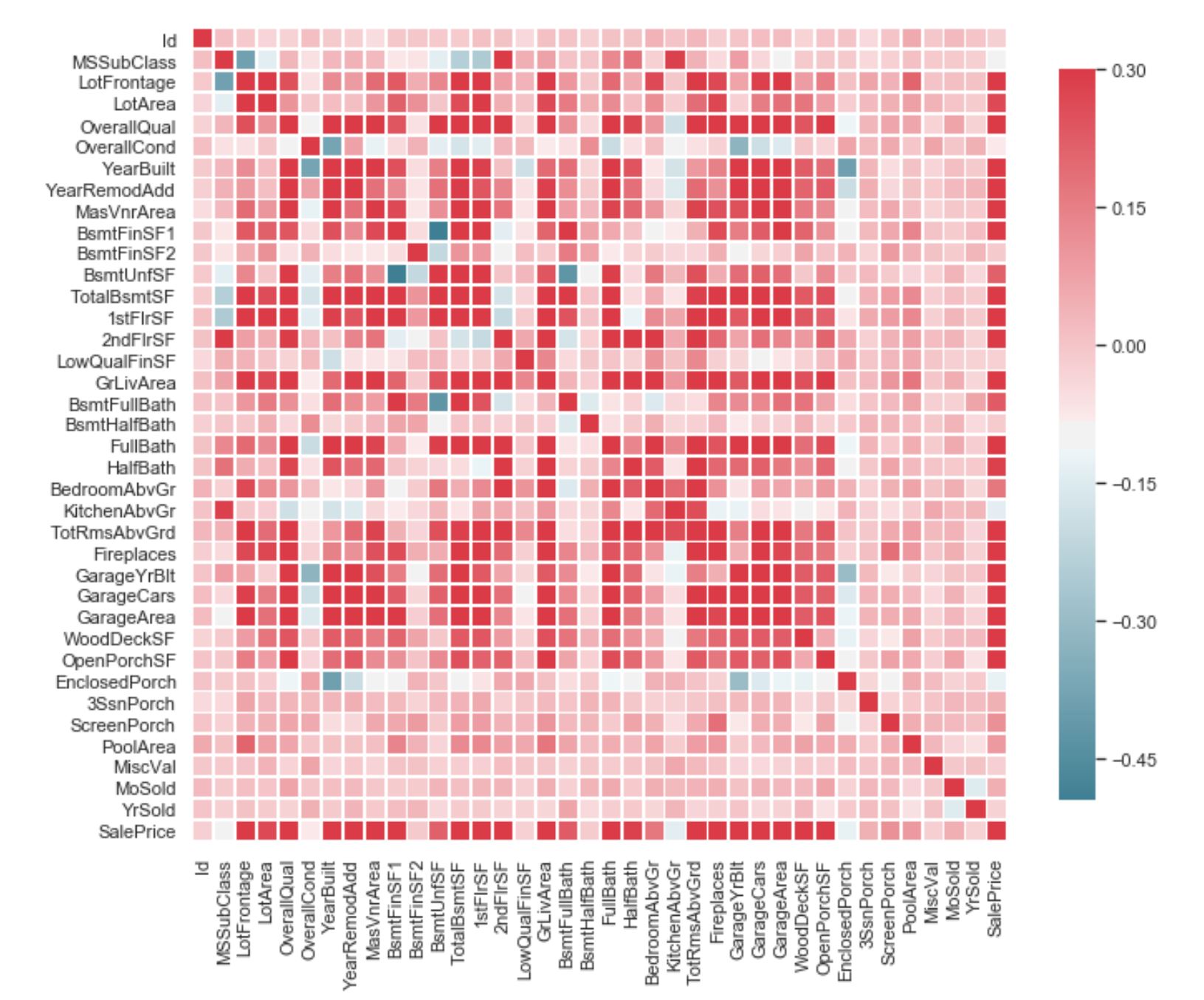
**3.1 Reading of Data**

I have also found missing values for Alley, Lot Frontage, Fireplace Quality and few other features. I have dropped few and used feature scaling techniques to assume.

The Data has 81 features and 1460 observations in train data and 1459 Observation in test data. There are missing values in 19 features and we can drop a few which has more than 50% of the null values in total observations and remaining can be deal with feature scaling and replacing null values with Mean or mode depends on other feature effects on null values.

**3.2 Correlation and Feature Scaling**

The correlation graph shows that “Overall Quality” and “Above grade (ground) living area square feet” has highest impact on sales price. “Total Bsmt Sf” and “1stFlr Sf” has next level impact on Sales price. Although we are having correlation graph on continuous data. We will now analyze on the categorial data as well



**3.3 Data Cleaning (Null Values)**

As I have discussed earlier Lot Area and Lot Frontage many have Null Values, so we are filling null values with mode of the respective Neighborhood feature. I have used similar way to fill null values for “Fireplace Qu”. I have used where function and np.median to fill the null values in Lot Frontage. I have done the similar cleaning with MasVnrType","MasVnrArea" with respect to “Foundation”.

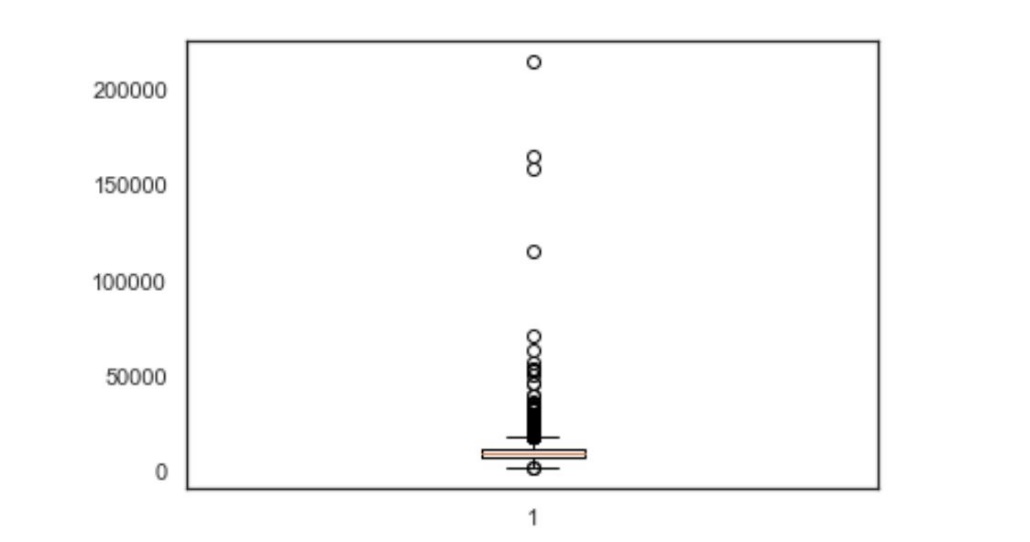
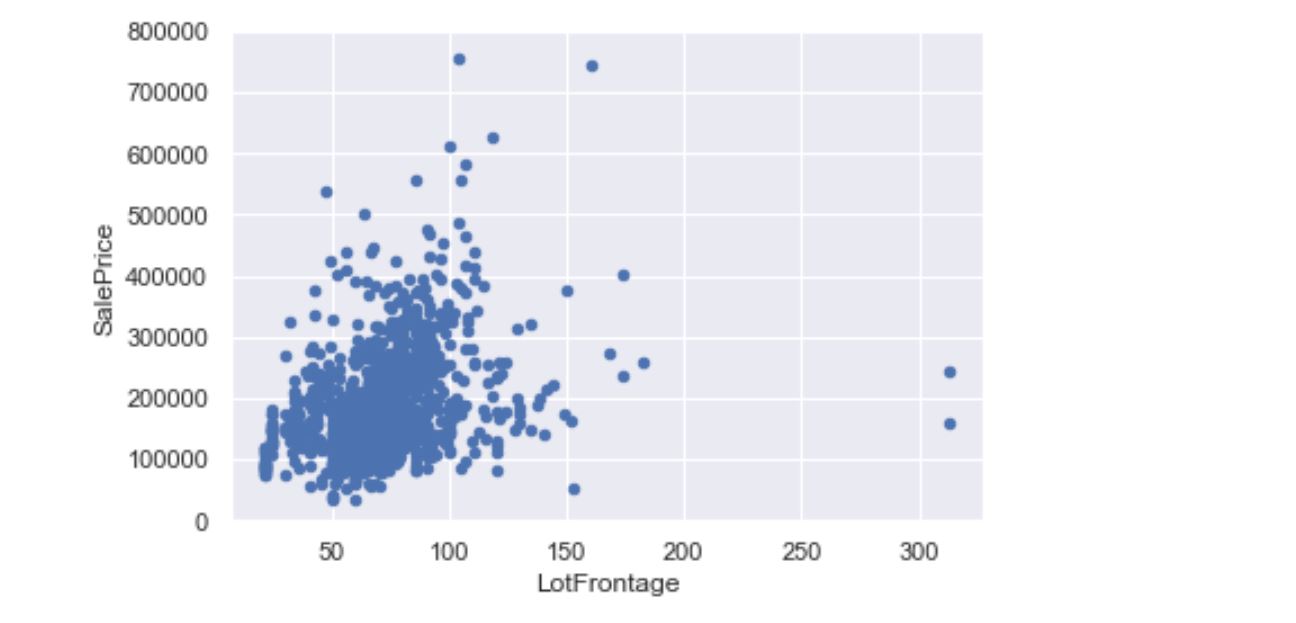
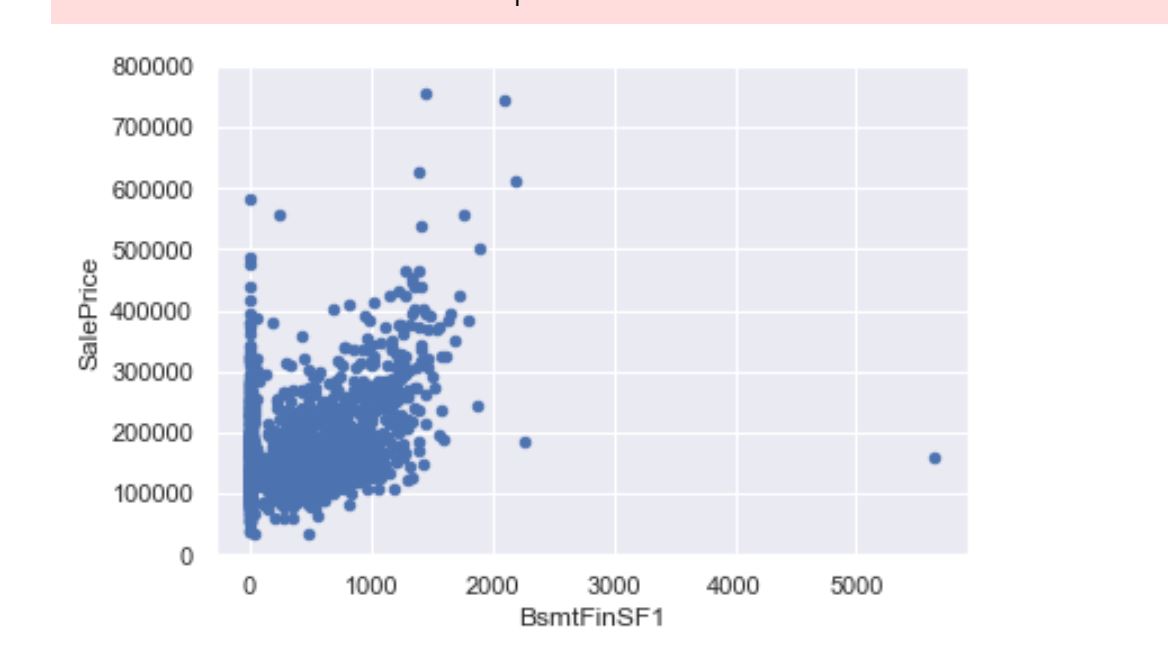
I have also filled with “Bsmt Qual” with Mode but here we have to be extremely careful when dealing with big data or data updating everyday. We cannot simply replace the null value with Mode or median as new data coming in mode and mean values changes so, I have changed value counts to to-list and there I have taken the first values as they are in decreasing order. I have repeated the dame thing with “Bsmt Cond”, “Bsmt Exposure”, “BsmtFinType1”, “BsmtFinType2”, “Electrical”.

Here is a tricky part, “GarageType", "GarageFinish", "GarageQual", "GarageCond", “Garage YrBlt”. All have similar null values. Yes, without Garage Year built, there won’t be any other ways to know details of other Garage features. Best thing is to drop all these but as the observations are less in number I keep them as unknown.

I have dropped the following featues "PoolQC", "Fence", "MiscFeature", "Alley" as they have more than 60% of the observations are null values.

**3.4 Clearing Outliers- Train Data**

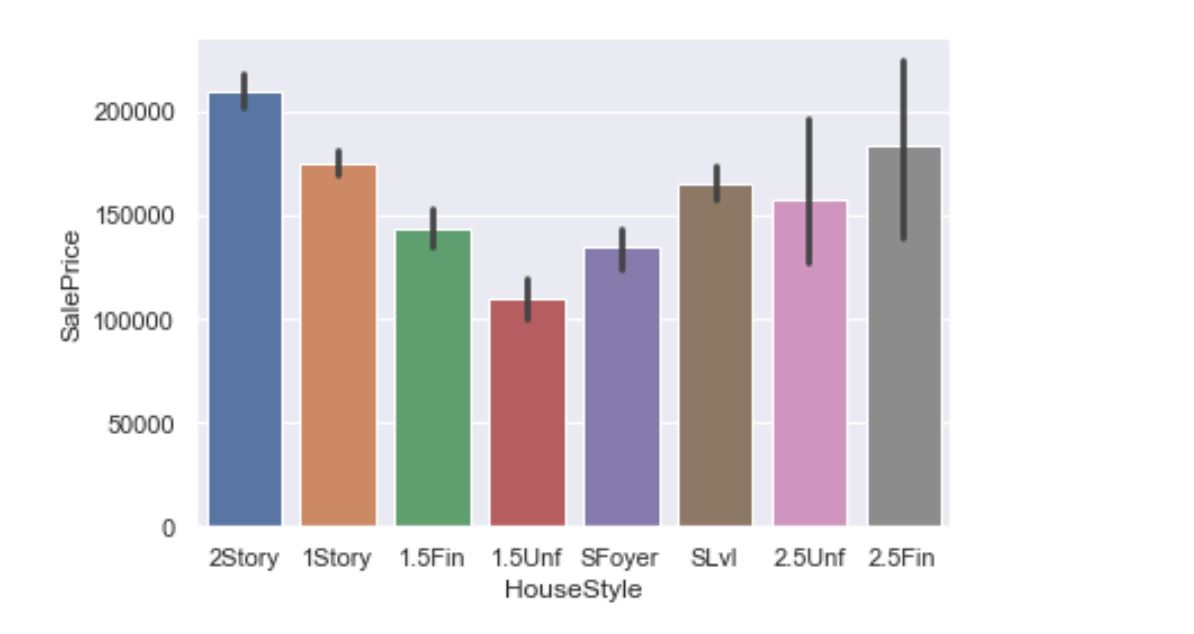
Outliers play an important role in increasing the variance and confuse the model which leads to predicts go wrong on unseen data which is overfitting. We need to clear outliers first. By using Box plot I have observed that “3SsnPorch”, “LotFrontage”, “3SsnPorch”, 'MasVnrArea', 'BsmtFinSF1', 'BsmtFinSF2', 'ScreenPorch', has outliers which can be removed.

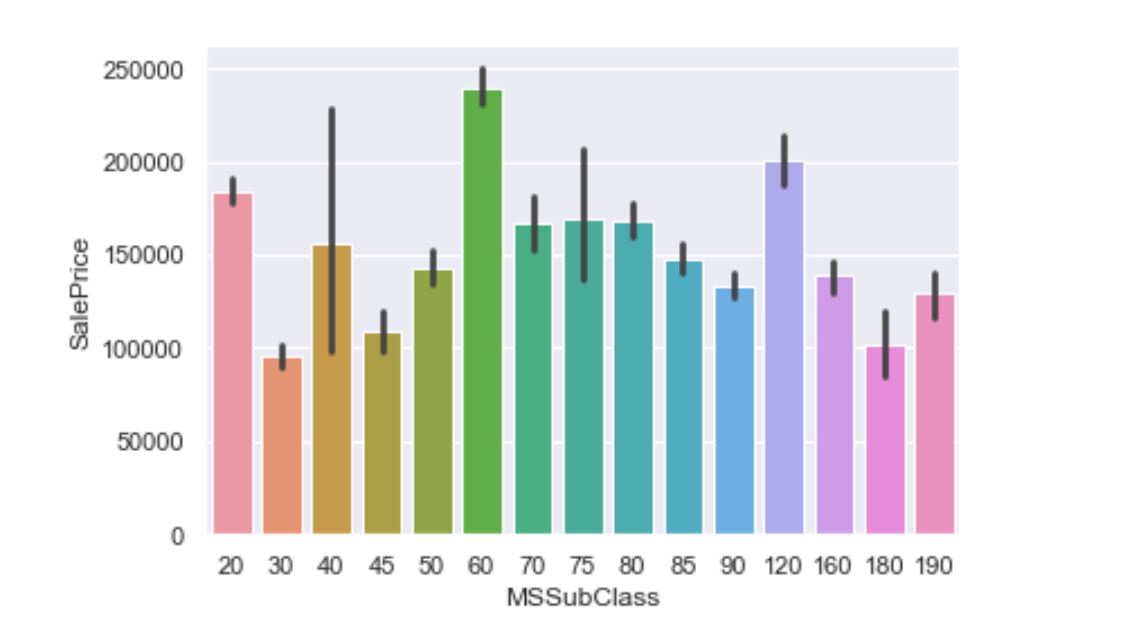
**4. Visualization:**

From Visualization I have following Conclusions:

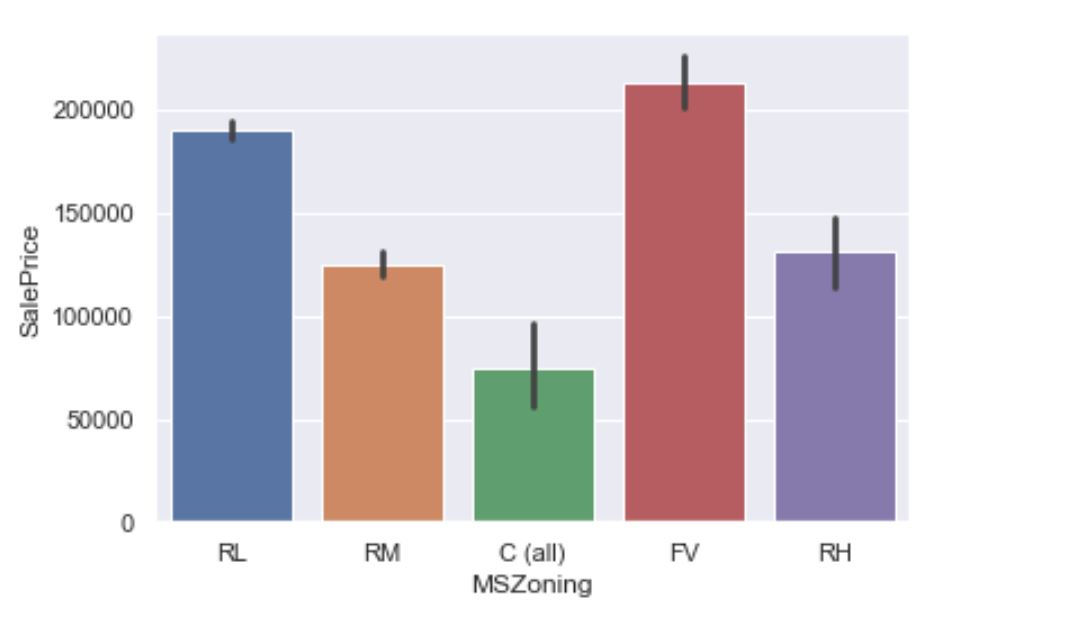
1. People are preferring 2 floor houses and has highest sale price observed ranges to 200000 followed by 2.5 floors with finished which is less than 20000.



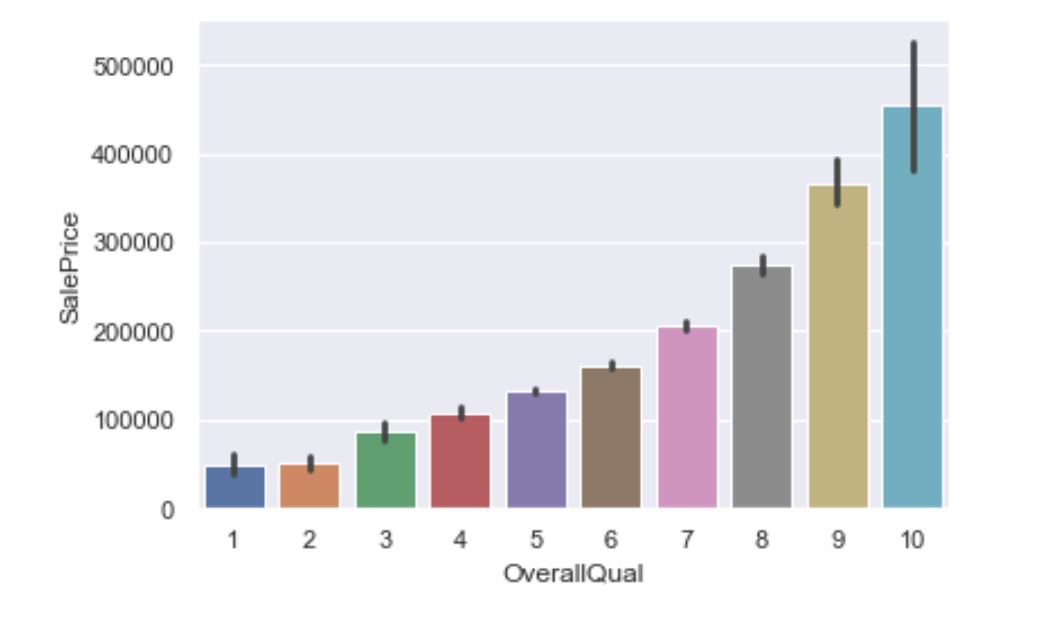
1. Building Class with 60 has highest sale price ranges to almost 250000+ and other classes has not even made it to 200k club except class 120.



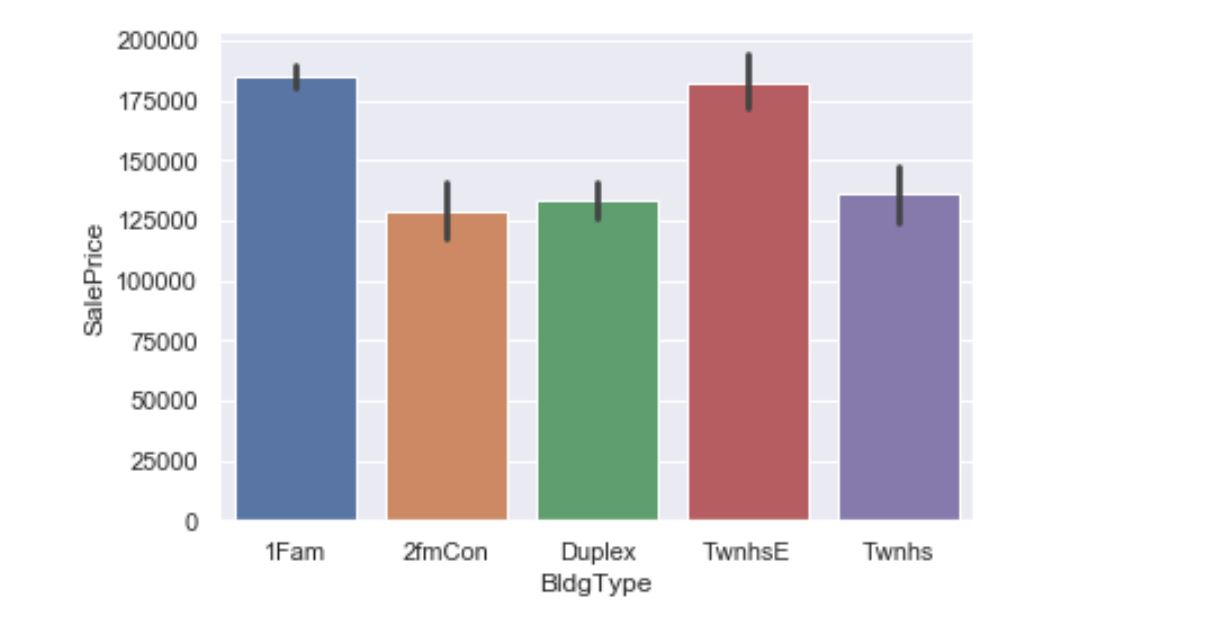
1. People are ready to pay highest price almost 200k+ in Village Residential Zoning.



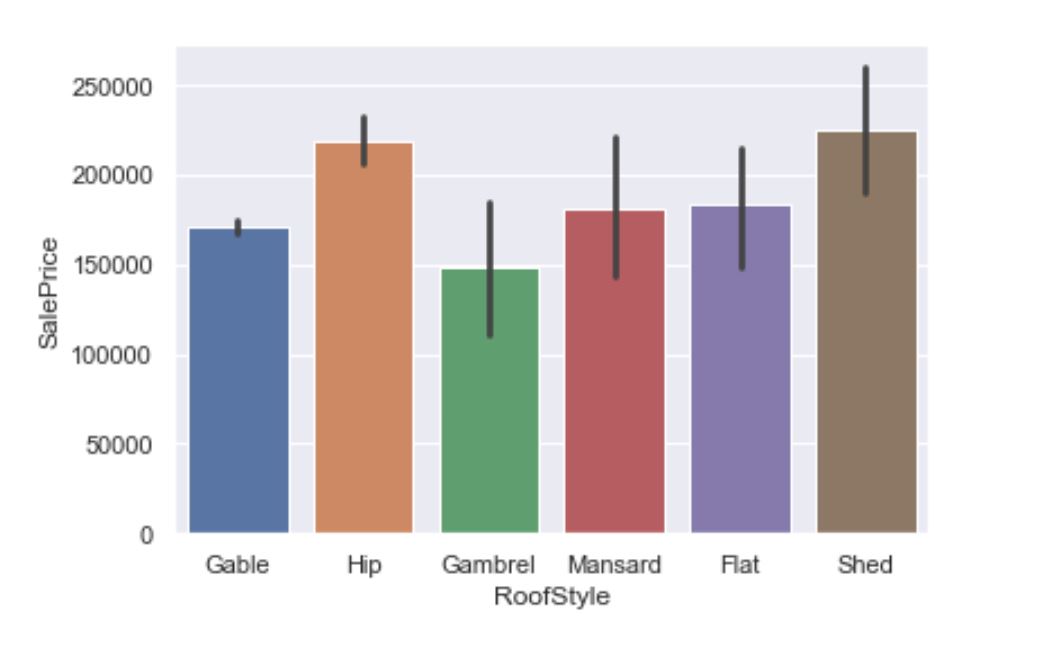
1. As we discussed earlier sale price also increase with Overall Quality in a Ordered way. With 10 being Very Excellent and has highest sale price of 400k+ and others quality not even made it to 400k club.

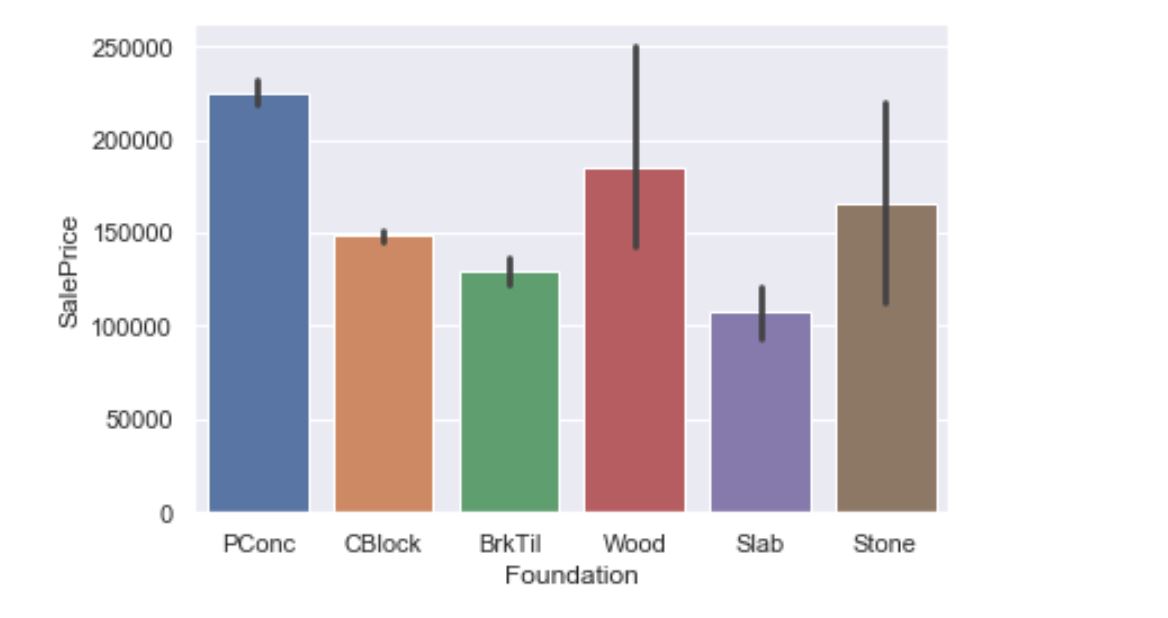


1. Sale price dominates in Town House and Farm land over duplex.

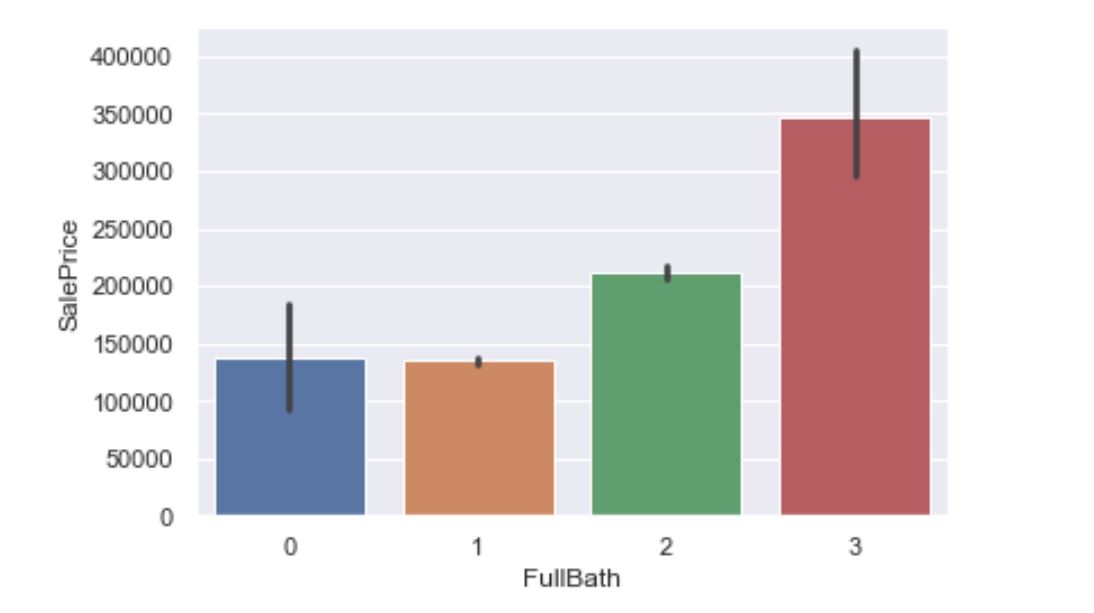


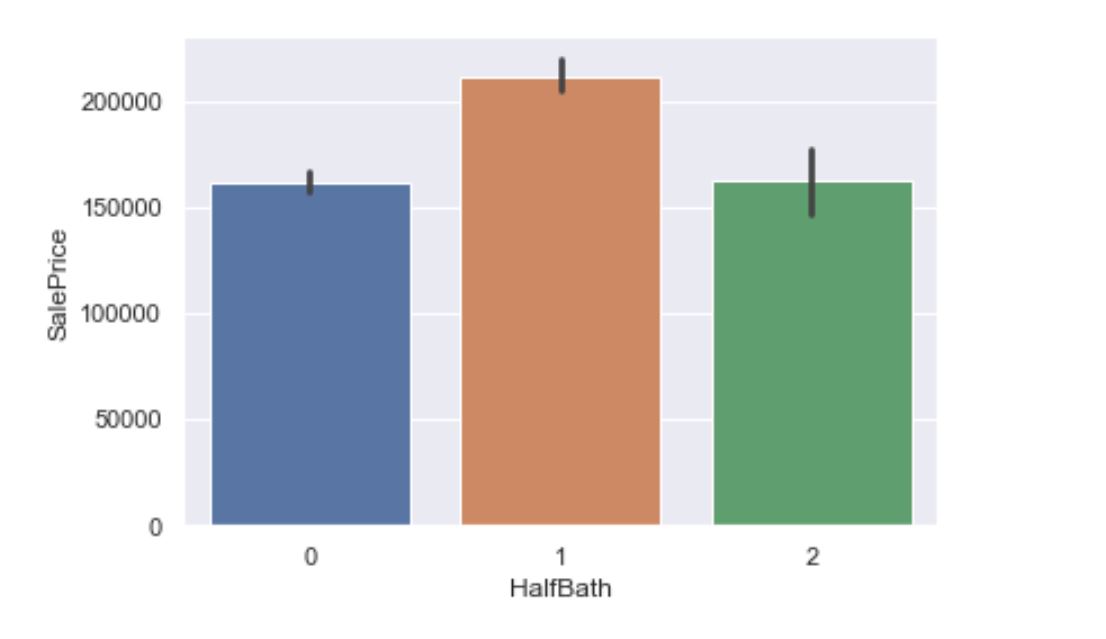
1. Concrete foundation and Roof Style Shed and Hip has highest Sale Price.



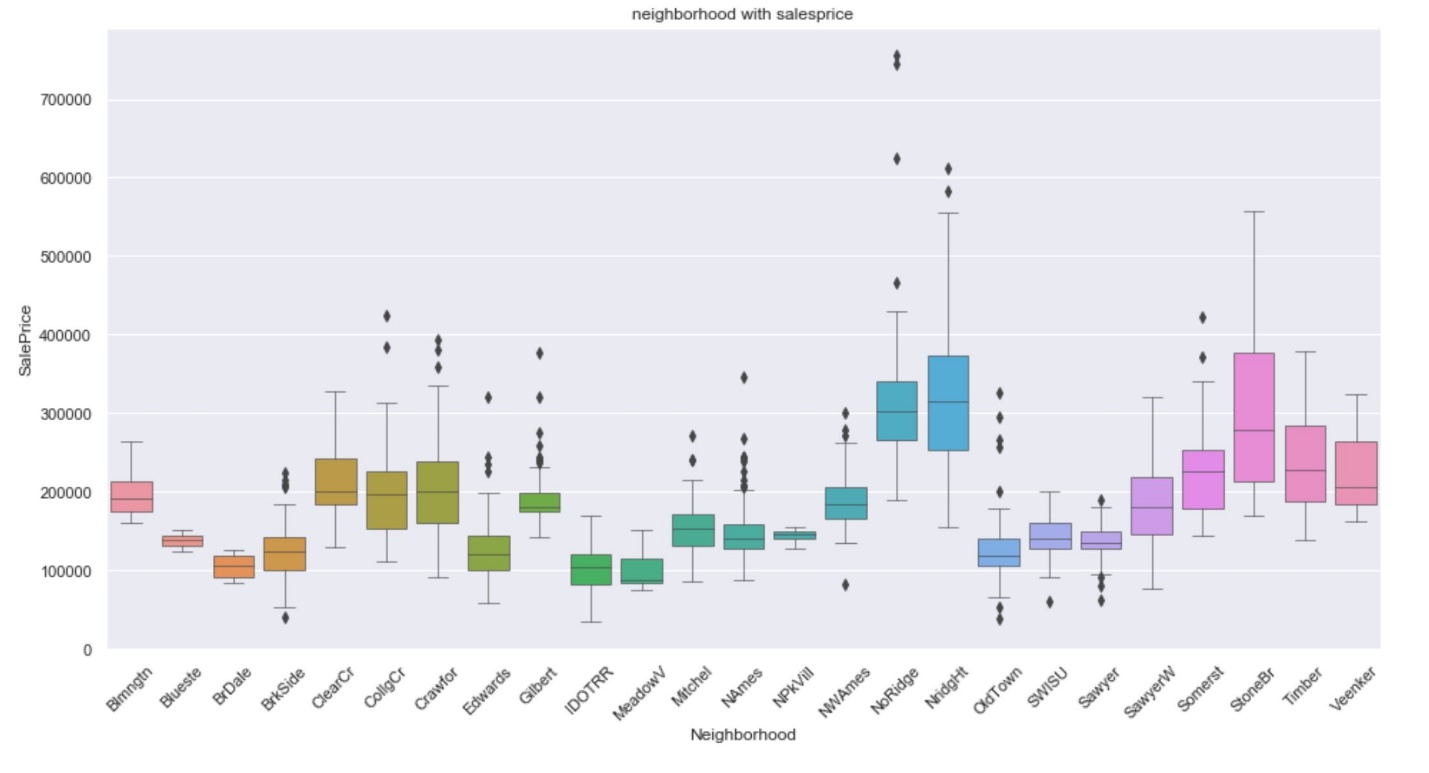
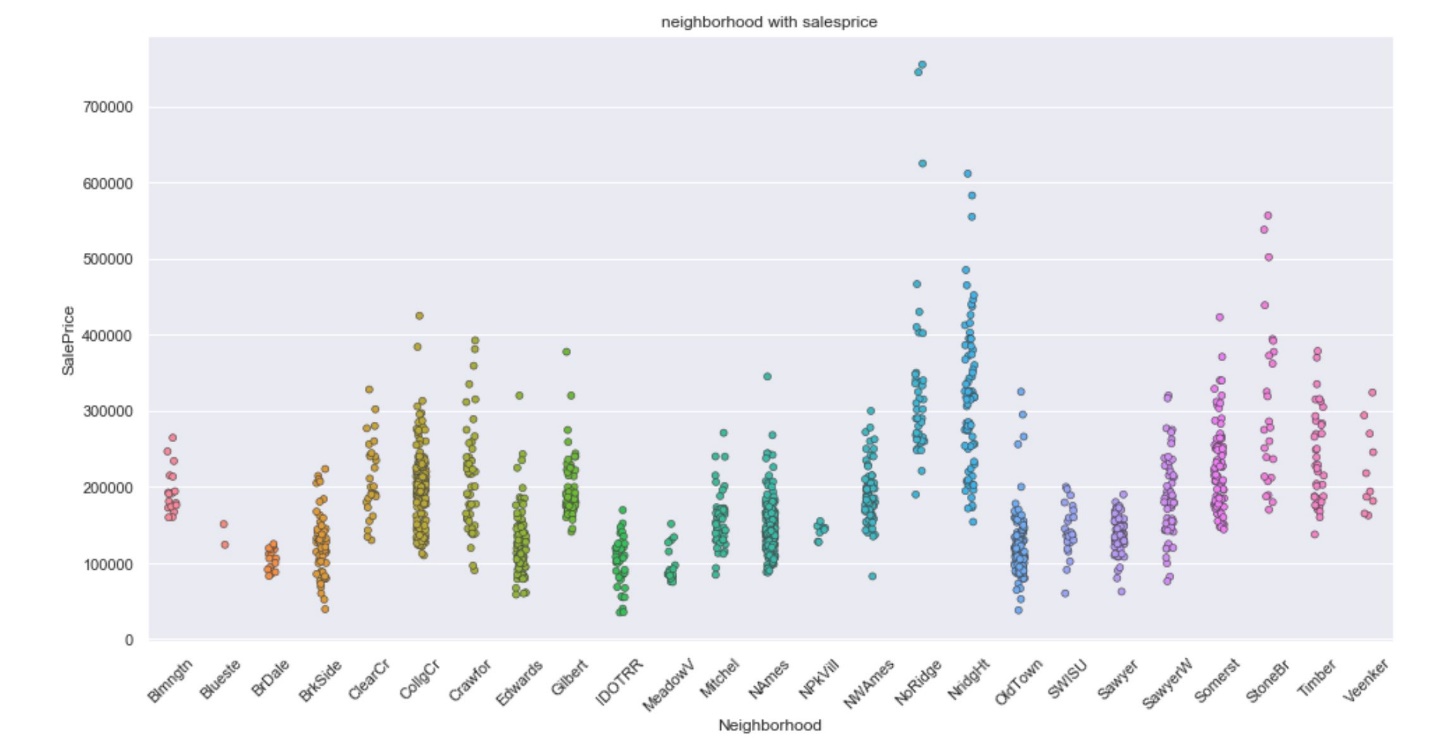
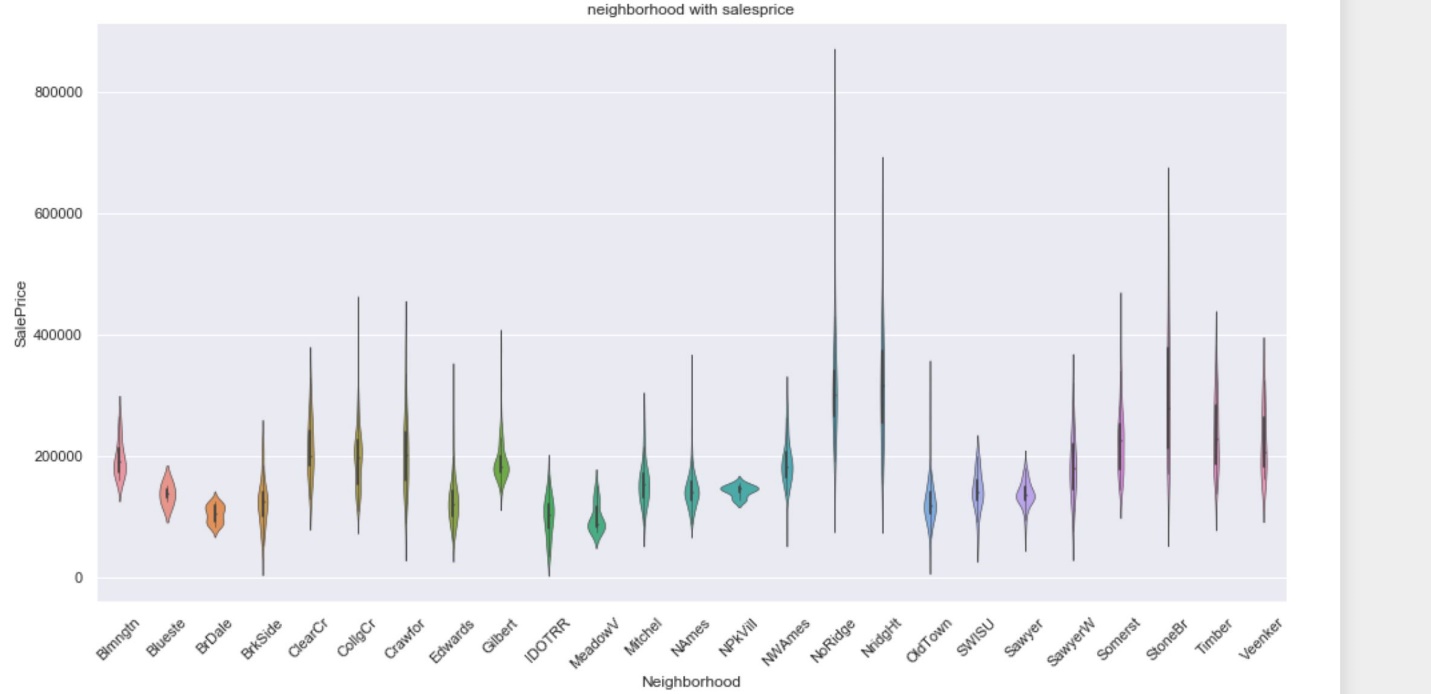


1. With Houses 3 bath and 1 Half bath which is also a factor for highest prices.





1. By Violin Plot, Boxplot and Strip Plot, NwAmes has highest sale price Neighborhood followed by Nridght and starting prices are also starts just below 200k. Brkside, IDotrr, Oldtown Neighborhood being the least priced.



**5. Converting Datatypes**

Datatypes of each observation must be changed to Continuous variables in order to feed to data. It all depends of Nominal or Ordinal Data. In summary, **nominal** variables are used to “name,” or label a series of values. **Ordinal** scales provide good information about the order of choices. We can use map function to map Ordinal Values and Nominal can be converted by creating Dummies. We use these techniques to create dummies for "LotConfig", "Neighborhood", "Condition1" , "Condition2", "RoofMatl", "Exterior1st", "Exterior2nd", "Heating", "Electrical","RoofStyle" , "MasVnrType", "Foundation", "GarageType", "SaleType" , "SaleCondition”, I used map function for other Categorical features.

**6. Data Preprocessing for Test Data**

Lot Frontage many have Null Values, so we are filling null values with mode of the respective Neighborhood feature. I have used similar way to fill null values for “Fireplace Qu”. I have used where function and np.median to fill the null values in Lot Frontage. I have done the similar cleaning with BsmtQual", "BsmtCond", "BsmtExposure", "BsmtFinType1", "BsmtFinType2" with respect to “Foundation”.

I have also filled with “Exterior1st” with Mode As discussed in training Data We cannot simply replace the null value with Mode or median as new data coming in mode and mean values changes so, I have changed value counts to to-list and there I have taken the first values as they are in decreasing order. I have repeated the same thing with “Utilities”, “BsmtFinType1”, “BsmtUnfSf”, “GarageArea” and “GarageCars”

Here is a tricky part, ““GarageType", "GarageFinish", "GarageQual", "GarageCond", “Garage YrBlt”. All have similar null values. Yes, without Garage Year built, there won’t be any other ways to know details of other Garage features. Best thing is to drop all these but as the observations are less in number I keep them as unknown.

I have dropped the following featues "PoolQC", "Fence", "MiscFeature", "Alley" as they have more than 60% of the observations are null values.

I have convert Categorical values to continous based on Nominal and ordinal data similar to train data.

**7. RESULTS:**

After scaling the data I have applied Random Forest Regressor as this is Regression Dataset on train data by splitting the data in to train and validation. Model accuracy is 97% on training data and almost 90% in validation data. This is also a good prediction in terms of model consideration. We can assume that out test results also 90% correctly predicted.

**8. CONCLUSION:**

With enough data feeding to the model the random forest regressor predicted outcomes are very good results and we can predict the house prices with more and more data and more and in other cities. Although the predictions differ with cities and states and same with prices But the process of predicting the model with enough Data is a good thing.

Although We cannot depend 100% on the model prediction because the accuracy is not 100% we can consider the predictions for any future decisions.