

Housing Price Prediction

Submitted by:

Ruchi Tiwari

**ACKNOWLEDGMENT**

Following are the external references which I used:

[www.w3school.com](http://www.w3school.com)

[www.stackoverflow.com](http://www.stackoverflow.com)

[www.google.com](http://www.google.com)

[www.geeksforgeeks.org](http://www.geeksforgeeks.org)

www.kaggle.com

**INTRODUCTION**

* Business Problem Framing

This problem involves predicting the prices of the houses which are continuous and real valued outputs. Thus, this is a **Regression Problem.**

* Conceptual Background of the Domain Problem

Houses are one of the necessary need of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world’s economy. It is a very large market and there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies. Our problem is related to one such housing company.

A US-based housing company named **Surprise Housing** has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia.

* Review of Literature

From the dataset I get to know that it is a Linear Regression problem and the target feature is the sales price and there are so many features which help to find it.

* Motivation for the Problem Undertaken

From this project I get to know of different kind of information like each attribute is play a very import role on purchasing a house. It is really quite interesting to know that each column contributed to make you close to know more about the data and in prediction you can do in many ways

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

The statistical figure I get to know by the data.describe() so many information the min max standard deviation the 25 percentile the 50th percentile the 75 percentile .Then by the help of correlation function I get to know the correlation of each columns with each other. From the heatmap I can visualize to see them clearly that they are positive correlated or the negative correlated the dark side is show the negative correlation among each other the lighter side represent the positive correlation among the each other. **The z-score** function computes the relative **Z-score** of the input data, relative to the sample mean and standard deviation.

Data Sources and their formats

Data I get form the Flip Robo the format was in CSV (Comma Separated Values).The number of columns and row are 1460 and columns are 81.

The data descriptions are as follow:-

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

**20 1-STORY 1946 & NEWER ALL STYLES**

**30 1-STORY 1945 & OLDER**

**40 1-STORY W/FINISHED ATTIC ALL AGES**

**45 1-1/2 STORY - UNFINISHED ALL AGES**

**50 1-1/2 STORY FINISHED ALL AGES**

**60 2-STORY 1946 & NEWER**

**70 2-STORY 1945 & OLDER**

**75 2-1/2 STORY ALL AGES**

**80 SPLIT OR MULTI-LEVEL**

**85 SPLIT FOYER**

**90 DUPLEX - ALL STYLES AND AGES**

**120 1-STORY PUD (Planned Unit Development) - 1946 & NEWER**

**150 1-1/2 STORY PUD - ALL AGES**

**160 2-STORY PUD - 1946 & NEWER**

**180 PUD - MULTILEVEL - INCL SPLIT LEV/FOYER**

**190 2 FAMILY CONVERSION - ALL STYLES AND AGES**

**MSZoning: Identifies the general zoning classification of the sale.**

**A Agriculture**

**C Commercial**

**FV Floating Village Residential**

**I Industrial**

**RH Residential High Density**

**RL Residential Low Density**

**RP Residential Low Density Park**

**RM Residential Medium Density**

**LotFrontage: Linear feet of street connected to property**

**LotArea: Lot size in square feet**

**Street: Type of road access to property**

**Grvl Gravel**

**Pave Paved**

**Alley: Type of alley access to property**

**Grvl Gravel**

**Pave Paved**

**NA No alley access**

**LotShape: General shape of property**

**Reg Regular**

**IR1 Slightly irregular**

**IR2 Moderately Irregular**

**IR3 Irregular**

**LandContour: Flatness of the property**

**Lvl Near Flat/Level**

**Bnk Banked - Quick and significant rise from street grade to building**

**HLS Hillside - Significant slope from side to side**

**Low Depression**

**Utilities: Type of utilities available**

**AllPub All public Utilities (E,G,W,& S)**

**NoSewr Electricity, Gas, and Water (Septic Tank)**

**NoSeWa Electricity and Gas Only**

**ELO Electricity only**

**LotConfig: Lot configuration**

**Inside Inside lot**

**Corner Corner lot**

**CulDSac Cul-de-sac**

**FR2 Frontage on 2 sides of property**

**FR3 Frontage on 3 sides of property**

**LandSlope: Slope of property**

**Gtl Gentle slope**

**Mod Moderate Slope**

**Sev Severe Slope**

**Neighborhood: Physical locations within Ames city limits**

**Blmngtn Bloomington Heights**

**Blueste Bluestem**

**BrDale Briardale**

**BrkSide Brookside**

**ClearCr Clear Creek**

**CollgCr College Creek**

**Crawfor Crawford**

**Edwards Edwards**

**Gilbert Gilbert**

**IDOTRR Iowa DOT and Rail Road**

**MeadowV Meadow Village**

**Mitchel Mitchell**

**Names North Ames**

**NoRidge Northridge**

**NPkVill Northpark Villa**

**NridgHt Northridge Heights**

**NWAmes Northwest Ames**

**OldTown Old Town**

**SWISU South & West of Iowa State University**

**Sawyer Sawyer**

**SawyerW Sawyer West**

**Somerst Somerset**

**StoneBr Stone Brook**

**Timber Timberland**

**Veenker Veenker**

**Condition1: Proximity to various conditions**

**Artery Adjacent to arterial street**

**Feedr Adjacent to feeder street**

**Norm Normal**

**RRNn Within 200' of North-South Railroad**

**RRAn Adjacent to North-South Railroad**

**PosN Near positive off-site feature--park, greenbelt, etc.**

**PosA Adjacent to postive off-site feature**

**RRNe Within 200' of East-West Railroad**

**RRAe Adjacent to East-West Railroad**

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

**Artery Adjacent to arterial street**

**Feedr Adjacent to feeder street**

**Norm Normal**

**RRNn Within 200' of North-South Railroad**

**RRAn Adjacent to North-South Railroad**

**PosN Near positive off-site feature--park, greenbelt, etc.**

**PosA Adjacent to postive off-site feature**

**RRNe Within 200' of East-West Railroad**

**RRAe Adjacent to East-West Railroad**

**BldgType: Type of dwelling**

**1Fam Single-family Detached**

**2FmCon Two-family Conversion; originally built as one-family dwelling**

**Duplx Duplex**

**TwnhsE Townhouse End Unit**

**TwnhsI Townhouse Inside Unit**

**HouseStyle: Style of dwelling**

**1Story One story**

**1.5Fin One and one-half story: 2nd level finished**

**1.5Unf One and one-half story: 2nd level unfinished**

**2Story Two story**

**2.5Fin Two and one-half story: 2nd level finished**

**2.5Unf Two and one-half story: 2nd level unfinished**

**SFoyer Split Foyer**

**SLvl Split Level**

**OverallQual: Rates the overall material and finish of the house**

**10 Very Excellent**

**9 Excellent**

**8 Very Good**

**7 Good**

**6 Above Average**

**5 Average**

**4 Below Average**

**3 Fair**

**2 Poor**

**1 Very Poor**

**OverallCond: Rates the overall condition of the house**

**10 Very Excellent**

**9 Excellent**

**8 Very Good**

**7 Good**

**6 Above Average**

**5 Average**

**4 Below Average**

**3 Fair**

**2 Poor**

**1 Very Poor**

**YearBuilt: Original construction date**

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

**RoofStyle: Type of roof**

**Flat Flat**

**Gable Gable**

**Gambrel Gabrel (Barn)**

**Hip Hip**

**Mansard Mansard**

**Shed Shed**

**RoofMatl: Roof material**

**ClyTile Clay or Tile**

**CompShg Standard (Composite) Shingle**

**Membran Membrane**

**Metal Metal**

**Roll Roll**

**Tar&Grv Gravel & Tar**

**WdShake Wood Shakes**

**WdShngl Wood Shingles**

**Exterior1st: Exterior covering on house**

**AsbShng Asbestos Shingles**

**AsphShn Asphalt Shingles**

**BrkComm Brick Common**

**BrkFace Brick Face**

**CBlock Cinder Block**

**CemntBd Cement Board**

**HdBoard Hard Board**

**ImStucc Imitation Stucco**

**MetalSd Metal Siding**

**Other Other**

**Plywood Plywood**

**PreCast PreCast**

**Stone Stone**

**Stucco Stucco**

**VinylSd Vinyl Siding**

**Wd Sdng Wood Siding**

**WdShing Wood Shingles**

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

**AsbShng Asbestos Shingles**

**AsphShn Asphalt Shingles**

**BrkComm Brick Common**

**BrkFace Brick Face**

**CBlock Cinder Block**

**CemntBd Cement Board**

**HdBoard Hard Board**

**ImStucc Imitation Stucco**

**MetalSd Metal Siding**

**Other Other**

**Plywood Plywood**

**PreCast PreCast**

**Stone Stone**

**Stucco Stucco**

**VinylSd Vinyl Siding**

**Wd Sdng Wood Siding**

**WdShing Wood Shingles**

**MasVnrType: Masonry veneer type**

**BrkCmn Brick Common**

**BrkFace Brick Face**

**CBlock Cinder Block**

**None None**

**Stone Stone**

**MasVnrArea: Masonry veneer area in square feet**

**ExterQual: Evaluates the quality of the material on the exterior**

**Ex Excellent**

**Gd Good**

**TA Average/Typical**

**Fa Fair**

**Po Poor**

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

**Ex Excellent**

**Gd Good**

**TA Average/Typical**

**Fa Fair**

**Po Poor**

**Foundation: Type of foundation**

**BrkTil Brick & Tile**

**CBlock Cinder Block**

**PConc Poured Contrete**

**Slab Slab**

**Stone Stone**

**Wood Wood**

**BsmtQual: Evaluates the height of the basement**

**Ex Excellent (100+ inches)**

**Gd Good (90-99 inches)**

**TA Typical (80-89 inches)**

**Fa Fair (70-79 inches)**

**Po Poor (<70 inches**

**NA No Basement**

**BsmtCond: Evaluates the general condition of the basement**

**Ex Excellent**

**Gd Good**

**TA Typical - slight dampness allowed**

**Fa Fair - dampness or some cracking or settling**

**Po Poor - Severe cracking, settling, or wetness**

**NA No Basement**

**BsmtExposure: Refers to walkout or garden level walls**

**Gd Good Exposure**

**Av Average Exposure (split levels or foyers typically score average or above)**

**Mn Mimimum Exposure**

**No No Exposure**

**NA No Basement**

**BsmtFinType1: Rating of basement finished area**

**GLQ Good Living Quarters**

**ALQ Average Living Quarters**

**BLQ Below Average Living Quarters**

**Rec Average Rec Room**

**LwQ Low Quality**

**Unf Unfinshed**

**NA No Basement**

**BsmtFinSF1: Type 1 finished square feet**

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

**GLQ Good Living Quarters**

**ALQ Average Living Quarters**

**BLQ Below Average Living Quarters**

**Rec Average Rec Room**

**LwQ Low Quality**

**Unf Unfinshed**

**NA No Basement**

**BsmtFinSF2: Type 2 finished square feet**

**BsmtUnfSF: Unfinished square feet of basement area**

**TotalBsmtSF: Total square feet of basement area**

**Heating: Type of heating**

**Floor Floor Furnace**

**GasA Gas forced warm air furnace**

**GasW Gas hot water or steam heat**

**Grav Gravity furnace**

**OthW Hot water or steam heat other than gas**

**Wall Wall furnace**

**HeatingQC: Heating quality and condition**

**Ex Excellent**

**Gd Good**

**TA Average/Typical**

**Fa Fair**

**Po Poor**

**CentralAir: Central air conditioning**

**N No**

**Y Yes**

**Electrical: Electrical system**

**SBrkr Standard Circuit Breakers & Romex**

**FuseA Fuse Box over 60 AMP and all Romex wiring (Average)**

**FuseF 60 AMP Fuse Box and mostly Romex wiring (Fair)**

**FuseP 60 AMP Fuse Box and mostly knob & tube wiring (poor)**

**Mix Mixed**

**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**

**Ex Excellent**

**Gd Good**

**TA Typical/Average**

**Fa Fair**

**Po Poor**

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

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

**Typ Typical Functionality**

**Min1 Minor Deductions 1**

**Min2 Minor Deductions 2**

**Mod Moderate Deductions**

**Maj1 Major Deductions 1**

**Maj2 Major Deductions 2**

**Sev Severely Damaged**

**Sal Salvage only**

**Fireplaces: Number of fireplaces**

**FireplaceQu: Fireplace quality**

**Ex Excellent - Exceptional Masonry Fireplace**

**Gd Good - Masonry Fireplace in main level**

**TA Average - Prefabricated Fireplace in main living area or Masonry Fireplace in basement**

**Fa Fair - Prefabricated Fireplace in basement**

**Po Poor - Ben Franklin Stove**

**NA No Fireplace**

**GarageType: Garage location**

**2Types More than one type of garage**

**Attchd Attached to home**

**Basment Basement Garage**

**BuiltIn Built-In (Garage part of house - typically has room above garage)**

**CarPort Car Port**

**Detchd Detached from home**

**NA No Garage**

**GarageYrBlt: Year garage was built**

**GarageFinish: Interior finish of the garage**

**Fin Finished**

**RFn Rough Finished**

**Unf Unfinished**

**NA No Garage**

**GarageCars: Size of garage in car capacity**

**GarageArea: Size of garage in square feet**

**GarageQual: Garage quality**

**Ex Excellent**

**Gd Good**

**TA Typical/Average**

**Fa Fair**

**Po Poor**

**NA No Garage**

**GarageCond: Garage condition**

**Ex Excellent**

**Gd Good**

**TA Typical/Average**

**Fa Fair**

**Po Poor**

**NA No Garage**

**PavedDrive: Paved driveway**

**Y Paved**

**P Partial Pavement**

**N Dirt/Gravel**

**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**

**Ex Excellent**

**Gd Good**

**TA Average/Typical**

**Fa Fair**

**NA No Pool**

**Fence: Fence quality**

**GdPrv Good Privacy**

**MnPrv Minimum Privacy**

**GdWo Good Wood**

**MnWw Minimum Wood/Wire**

**NA No Fence**

**MiscFeature: Miscellaneous feature not covered in other categories**

**Elev Elevator**

**Gar2 2nd Garage (if not described in garage section)**

**Othr Other**

**Shed Shed (over 100 SF)**

**TenC Tennis Court**

**NA None**

**MiscVal: $Value of miscellaneous feature**

**MoSold: Month Sold (MM)**

**YrSold: Year Sold (YYYY)**

**SaleType: Type of sale**

**WD Warranty Deed - Conventional**

**CWD Warranty Deed - Cash**

**VWD Warranty Deed - VA Loan**

**New Home just constructed and sold**

**COD Court Officer Deed/Estate**

**Con Contract 15% Down payment regular terms**

**ConLw Contract Low Down payment and low interest**

**ConLI Contract Low Interest**

**ConLD Contract Low Down**

**Oth Other**

**SaleCondition: Condition of sale**

**Normal Normal Sale**

**Abnorml Abnormal Sale - trade, foreclosure, short sale**

**AdjLand Adjoining Land Purchase**

**Alloca Allocation - two linked properties with separate deeds, typically condo with a garage unit**

**Family Sale between family members**

**Partial Home was not completed when last assessed (associated with New Homes)**

* Data Preprocessing Done

There were so many null value was present in the dataset but there are some outliers which also get too removed, I Separated the variable into new dataframe from original dataframe which has only numerical values,there is 38 numerical attribute from 81 attributes. Than I fill all the null value None , mode and zero .Then I see the correlation with the sales price and get to know that OverallQual is highly correlated with target feature of saleprice by 82% I check the data that it is normally distributed or not than I get to know that the target varibale is right skewed. Now,I need to tranform this variable and make it normal distributed.

Hardware and Software Requirements and Tools Used

**Hardware** – Laptop

**Software** - anaconda jupyter notebook

**Libraries**- numpy, pandas, seaborn, matplotlib.pyplot, warning

**from sklearn.preprocessing import Label Encoder**

 Label Encoder and One Hot Encoder. These two encoders are parts of the SciKit Learn library in Python, and they are used to convert categorical data, or text data, into numbers, which our predictive models can better understand.

**from sklearn.model\_selection import train\_test\_split,cross\_val\_score**

Train\_test\_split is a function in Sklearn model selection for splitting data arrays into two subsets: for training data and for testing data. With this function, you don't need to divide the dataset manually. By default, Sklearn train\_test\_split will make random partitions for the two subsets.

The algorithm is trained and tested K times, each time a new set is used as testing set while remaining sets are used for training. Finally, the result of the K-Fold Cross-Validation is the average of the results obtained on each set.

**from sklearn.neighbors import linear\_model**

The coefficient estimates for Ordinary Least Squares rely on the independence of the features. When features are correlated and the columns of the design matrix X have an approximate linear dependence, the design matrix becomes close to singular and as a result, the least-squares estimate becomes highly sensitive to random errors in the observed target, producing a large variance. This situation of multicollinearity can arise, for example, when data are collected without an experimental design.

**from sklearn.ensemble import RandomForestRegressor**

Random Forest has multiple decision trees as base learning models. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model.

**from sklearn.ensemble import GradientBoostingRegressor**

Gradient Boosting for regression.

GB builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. In each stage a regression tree is fit on the negative gradient of the given loss function.

**from sklearn.linear\_model import Lasso,Ridge**

In statistics and machine learning, lasso is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces.

* Ridge and Lasso Regression are types of Regularization techniques
* Regularization techniques are used to deal with overfitting and when the dataset is large
* Ridge and Lasso Regression involve adding penalties to the regression function

**from sklearn.linear\_model import ElasticNet**

Elastic net linear regression uses the penalties from both the lasso and ridge techniques to regularize regression models. The technique combines both the [lasso](https://corporatefinanceinstitute.com/resources/knowledge/other/lasso/) and ridge regression methods by learning from their shortcomings to improve on the regularization of statistical models.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

**Descriptive statistics** are used to describe the basic features of the data in a study which are mean count max standard deviations 25% , 75% , 50 % it all help me to understand the data in terms of statistically for the problem solving.

* Testing of Identified Approaches (Algorithms)

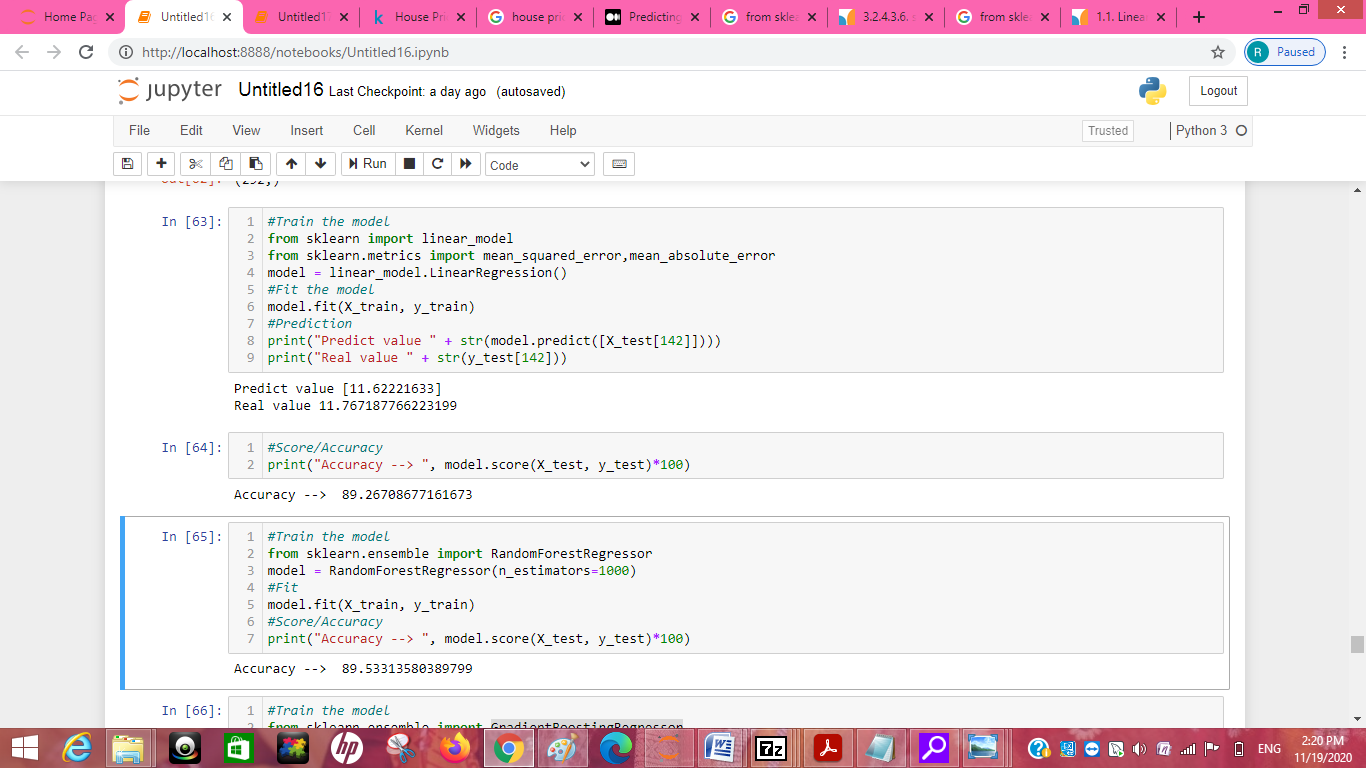
Listing down all the algorithms used for the training and testing.

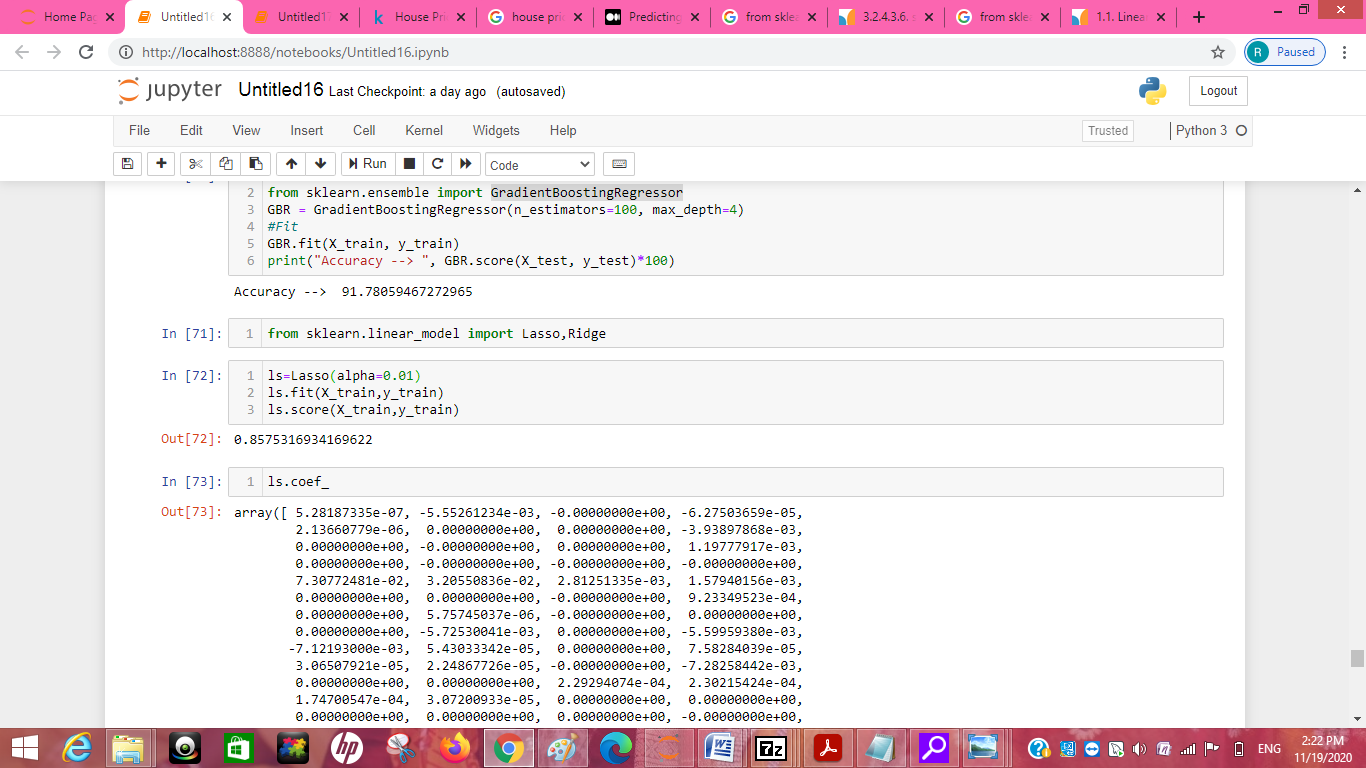
Listing down all the algorithms used for the training and testing.

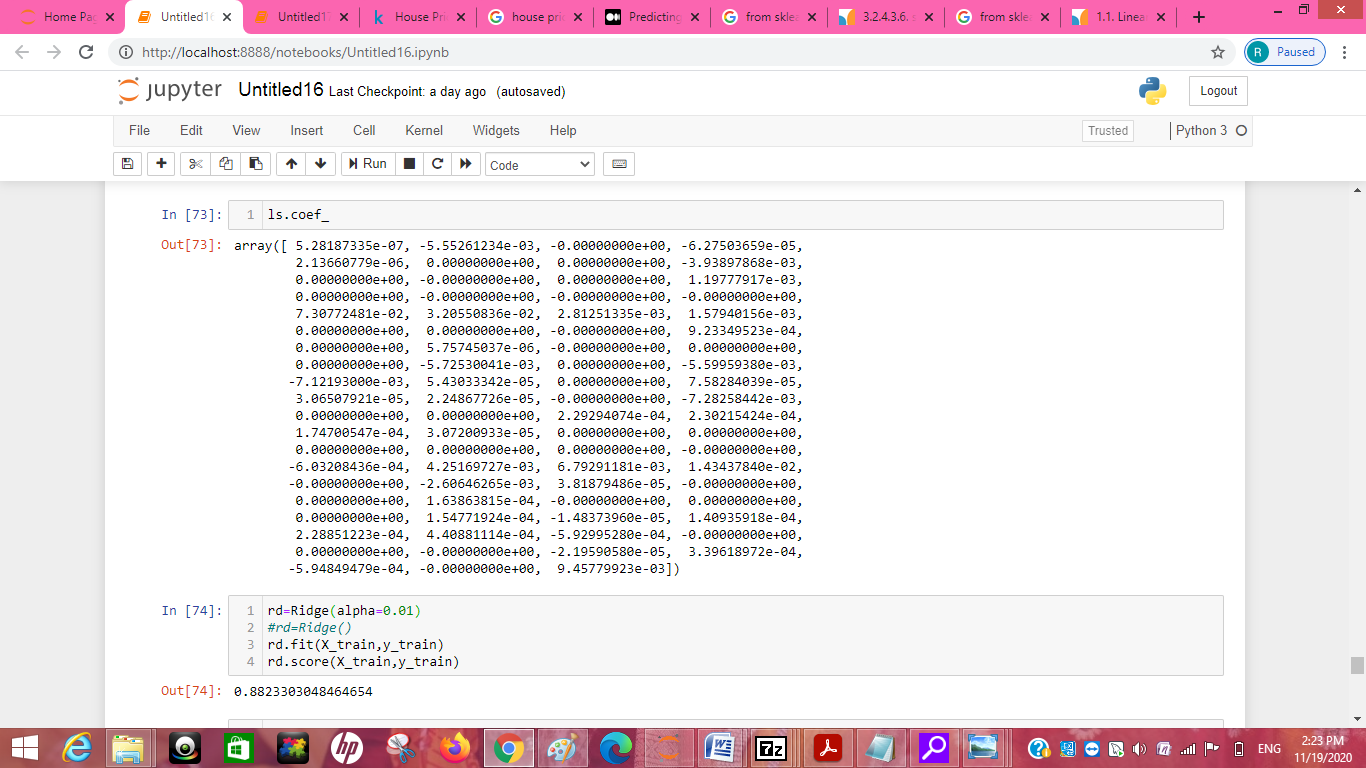
* LR = LinearRegression()
* RFR = RandomForestRegressor(n\_estimators=1000)
* GBR = GradientBoostingRegressor(n\_estimators=100, max\_depth=4)
* enr=ElasticNet
* ls=Lasso
* rd=Ridge

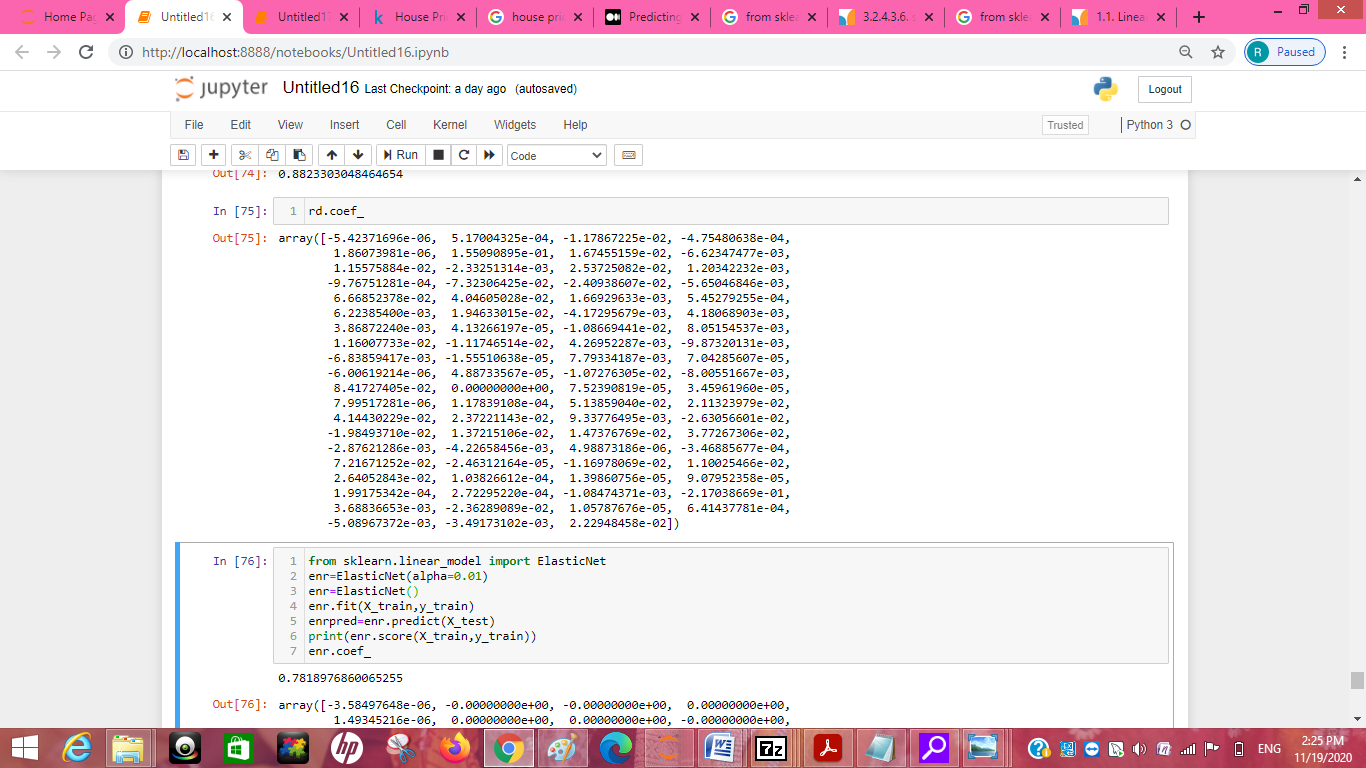
I applied all these algorithms in the dataset.

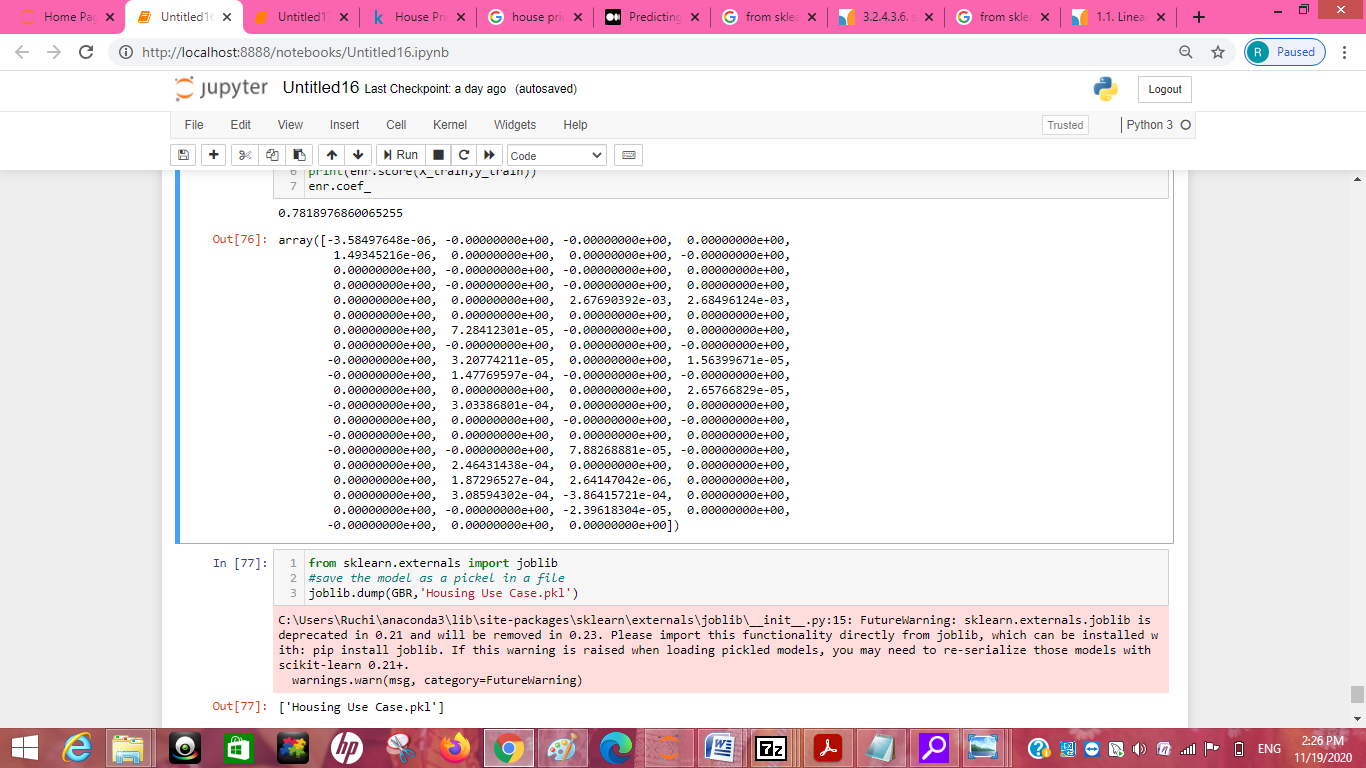
* Run and Evaluate selected models











Visualizations

plt.subplots(figsize=(12,9))

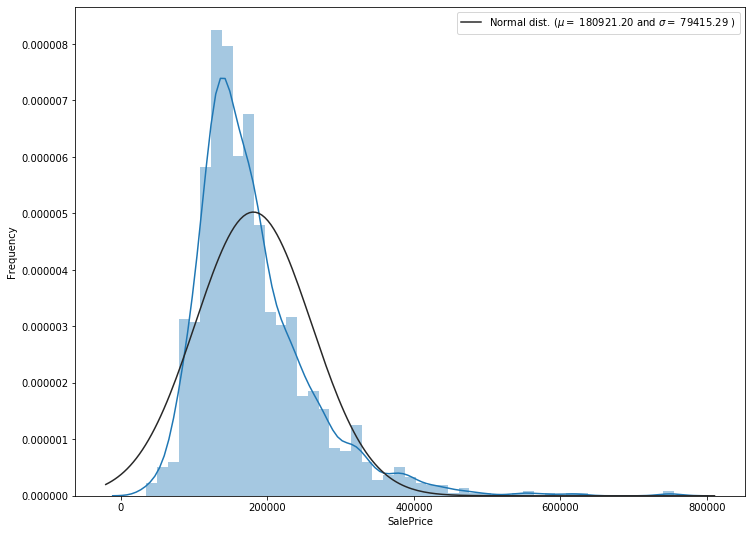
sns.distplot(train['SalePrice'], fit=stats.norm)

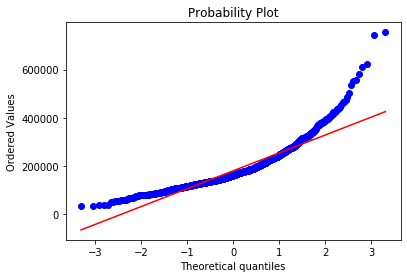
(mu, sigma) = stats.norm.fit(train['SalePrice'])

fig = plt.figure()

stats.probplot(train['SalePrice'], plot=plt)

plt.show()





train['SalePrice'] = np.log1p(train['SalePrice'])

plt.subplots(figsize=(12,9))

sns.distplot(train['SalePrice'], fit=stats.norm)

(mu, sigma) = stats.norm.fit(train['SalePrice'])

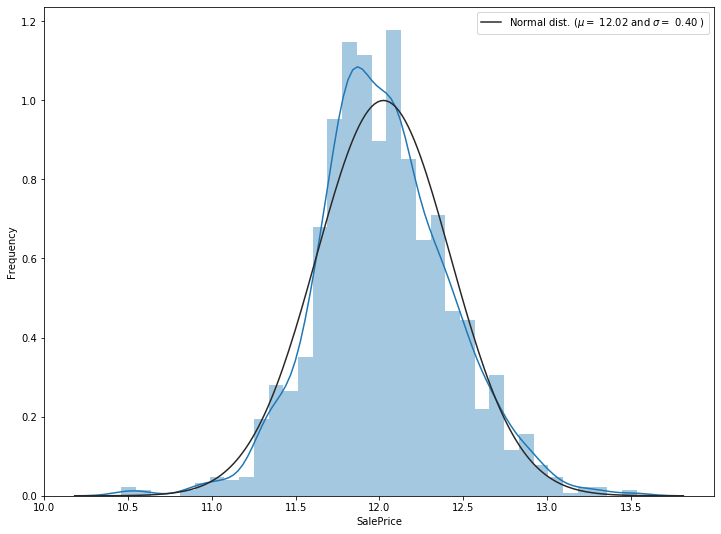
plt.legend(['Normal dist. ($\mu=$ {:.2f} and $\sigma=$ {:.2f} )'.format(mu, sigma)], loc='best')

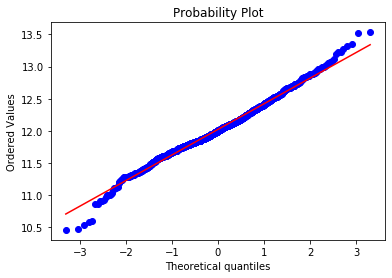
plt.ylabel('Frequency')

fig = plt.figure()

stats.probplot(train['SalePrice'], plot=plt)

plt.show()



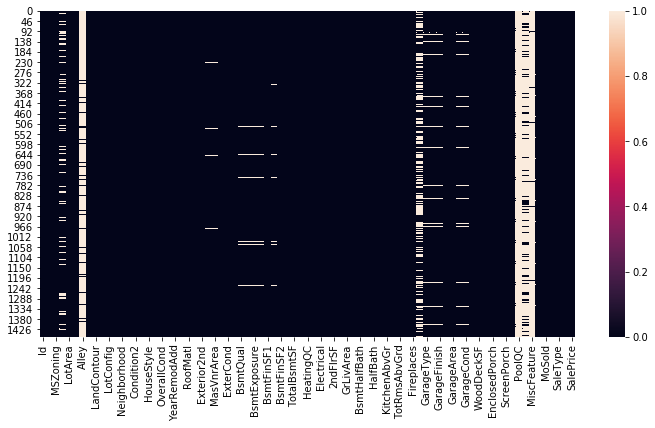


#plot of missing value attributes

plt.figure(figsize=(12, 6))

sns.heatmap(train.isnull())

plt.show()



#plot Missing values

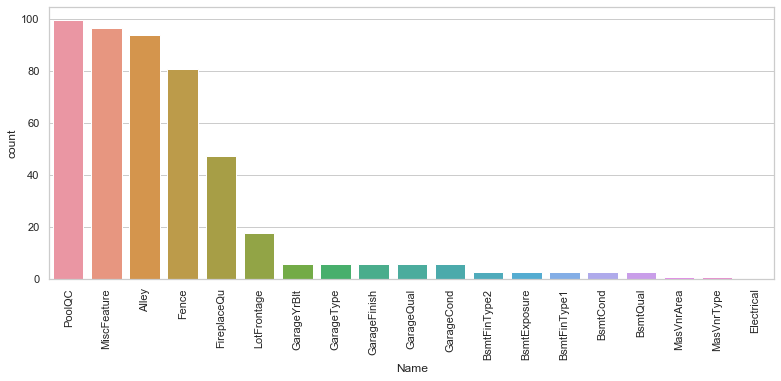
plt.figure(figsize=(13, 5))

sns.set(style='whitegrid')

sns.barplot(x='Name', y='count', data=Isnull)

plt.xticks(rotation = 90)

plt.show()

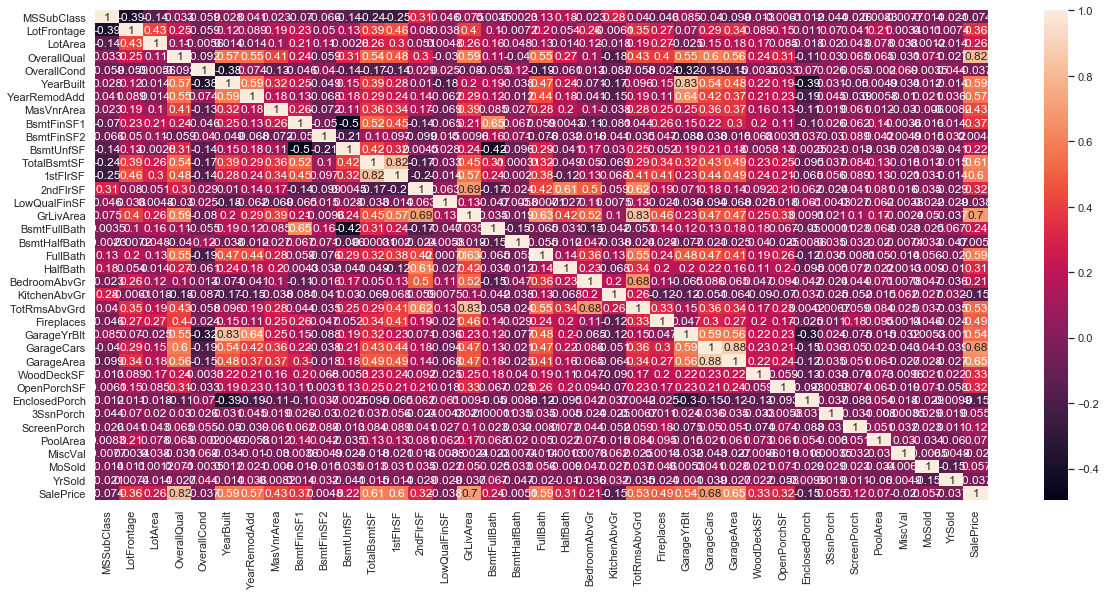


#Coralation plot

corr = train\_corr.corr()

plt.subplots(figsize=(20,9))

sns.heatmap(corr, annot=True)



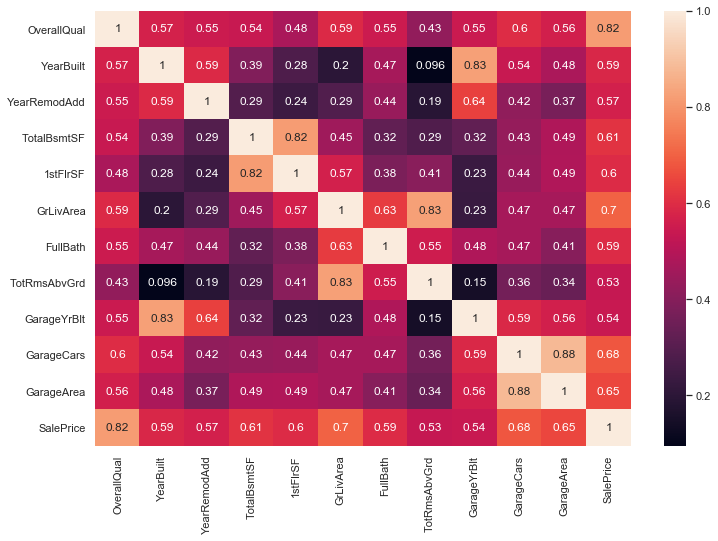
top\_feature = corr.index[abs(corr['SalePrice']>0.5)]

plt.subplots(figsize=(12, 8))

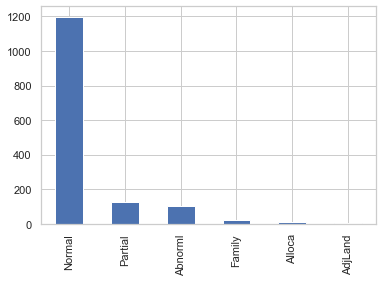
top\_corr = train[top\_feature].corr()

sns.heatmap(top\_corr, annot=True)

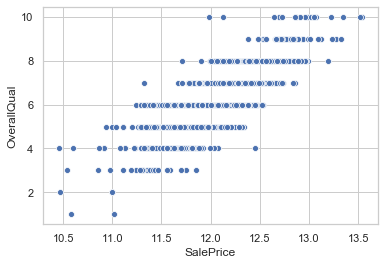
plt.show()



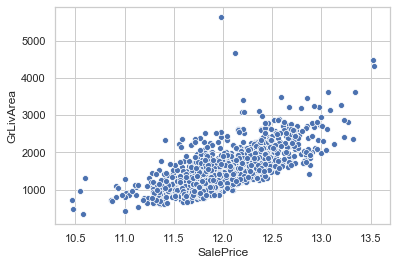
train['SaleCondition'].value\_counts().plot(kind='bar')



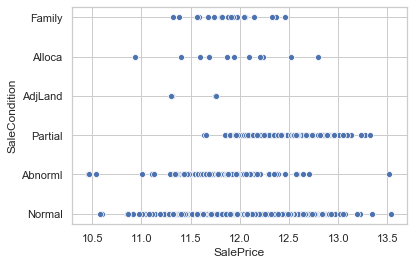
column\_search('OverallQual')



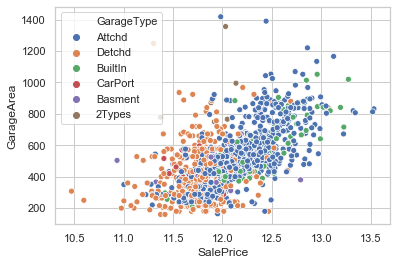
column\_search('GrLivArea')



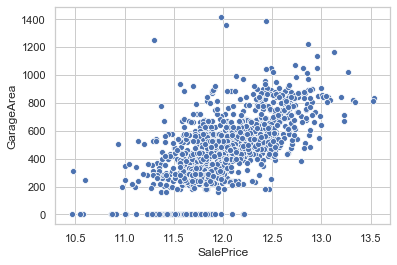
sns.scatterplot(x='SalePrice', y='SaleCondition', data=train)



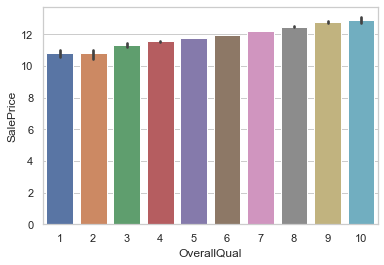
sns.scatterplot(x='SalePrice', y='GarageArea', hue='GarageType', data=train)



column\_search('GarageArea')



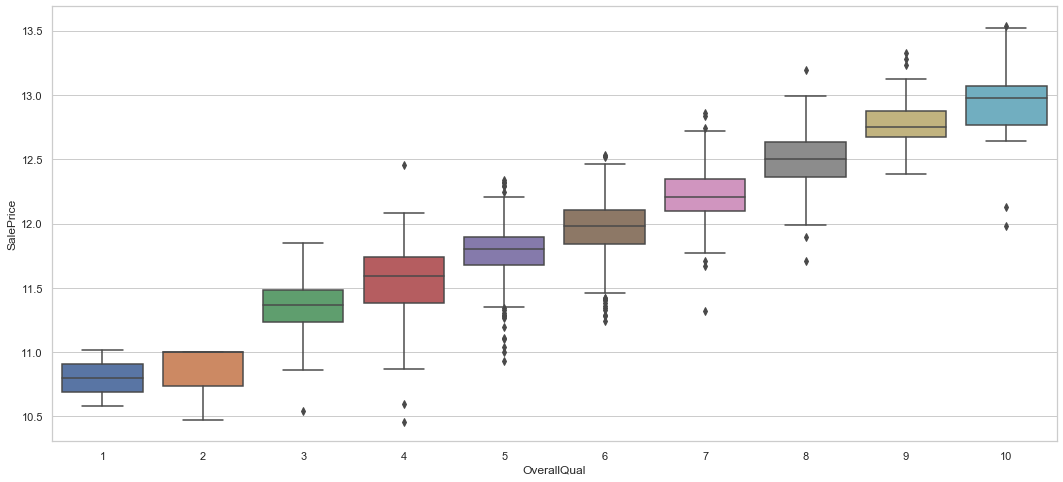
sns.barplot(train.OverallQual, train.SalePrice)



#boxplot

plt.figure(figsize=(18, 8))

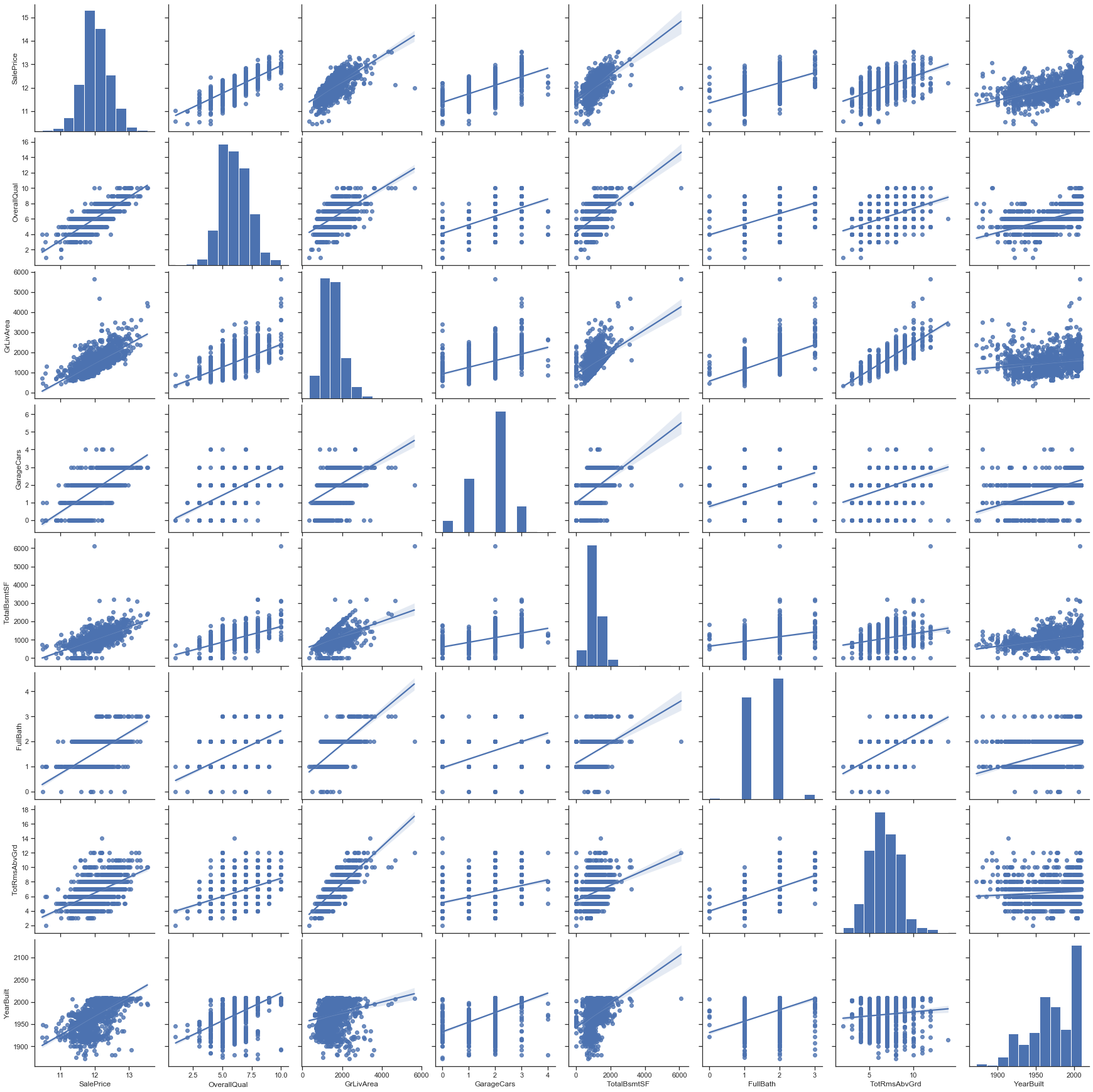
sns.boxplot(x=train.OverallQual, y=train.SalePrice)



col = ['SalePrice', 'OverallQual', 'GrLivArea', 'GarageCars', 'TotalBsmtSF', 'FullBath', 'TotRmsAbvGrd', 'YearBuilt']

sns.set(style='ticks')

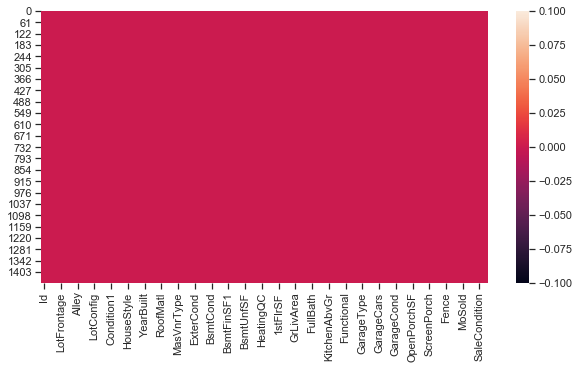
sns.pairplot(train[col], size=3, kind='reg')



#Checking there is any null value or not

plt.figure(figsize=(10, 5))

sns.heatmap(train.isnull())



* Interpretation of the Results

Give a summary of what results were interpreted from the visualizations, preprocessing and modelling.

From the info I get to that their is dtypes: float64(3), int64(35), object(43)

Separate variable into new dataframe from original dataframe which has only numerical values

there is 38 numerical attribute from 81 attributes

Delete Id because that is not need for corralation plot

From the correlation I get to know that OverallQual is highly correlated with target feature of saleprice by 82%

In OverallQual plot I get to know From this we can confirm that the higher the overall quality of the house, most often, the higher the sale price of the house. It is interesting too see that we have a couple outliers though. The first points that caught my eye in this graph were the two points furthest to the left for OverallQual of 10. It is interesting that we have a pretty consistent curve throughout these points except for those two points.

In GrLivArea the numbers show the amount of square feet for each house. We can see that the smallest house size was 334 square feet and the largest was 5642 square feet.

**After looking at this graph there is a very intriguing thing going on. This is, although not confirmed yet, there are two points that have the two highest living area and very low prices given the trend. We will see if these points match with the points from the overall quality points. If they do it will either make our model hard to make or we will need to take these points out because they will be massive outliers.**

From Year sold this we can see that it would be a little uneasy to come to the recession conclusion completely because the data only has houses that was sold from 2006 to 2010. We will go ahead and start looking at features again after taking a little detour

**When looking at the GarageCars variable we can see that the observations lie in line with how many garage cars spaces are available in our dataset homes. The mean is about 2 cars in each home.**

We can see that attached and builtin garages are more expensive than detached cars. We can see that there are two attached garages with the highest garagearea that are between 150,000 and 250,000 dollars which is not constant with the rest of the points in the category. We also have a detached garage that is over 1200 square feet, but less than 100,000. Lets look again at our outlier point and see what the garage situation looks like.

**CONCLUSION**

Key Findings and Conclusions of the Study

**The key findings:**

From this dataset I get to know that each feature play a very import role to understand the data. Data format plays a very important role in the visualization and Appling the models and algorithms.

Learning Outcomes of the Study in respect of Data Science

My learnings :- the power of visualization is helpful for the understanding of data into the graphical representation its help me to understand that what data is trying to say, Data cleaning is one of the most important step to remove missing value or null value fill it by mean median or by mode or by 0.

Various algorithms I used in this dataset and to get out best result and save that model.The best algorithm is GradientBoostingRegressor

The challenges I faced while working on this project basically I was trying to face issue in running the SVC algorithm and during the pair plot also because to huge rows and columns I face the issue to run it since it take more than hour to run I overcome by taking the help of Google I am able to run it. sns.heatmap(dfcor) From this code I get the below picture which represent the correlation among different columns since darker side represents the negative correlation and the higher side represent the positive correlation.