

**HOUSING: PRICE PREDICTION**

**Submitted by:**

**PARAS MALHOTRA**

**ACKNOWLEDGMENT**

I express my sincere gratitude to Flip Robo Technologies for giving me the opportunity to work on this project on Housing Price prediction using machine learning algorithms.

**INTRODUCTION**

**Business Problem Framing**

Houses are one of the necessary needs 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. The company is looking at prospective properties to buy houses to enter the market. You are required to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not. For this company wants to know:

• Which variables are important to predict the price of variable?

• How do these variables describe the price of the house?

**Conceptual Background of the Domain Problem**

Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies.

Hedonic Characteristics of Housing Price: A Hedonic approach is preferred for predicting the sale prices in the housing market because the market displays resilience, flexibility and spatial fixity.

Housing Attributes: Studying the structural, locational, and economic attributes of housing properties is crucial in understanding their mutually inclusive relationships with their pricing.

**Review of Literature**

2 research papers, namely: “House Price Prediction using a Machine Learning Model: A Survey of Literature” and “The impact of housing quality on house prices in eight capital cities, Australia” were reviewed and evaluated to gain insights into all the attributes that influence the price of house.

From studying the papers and analysing the research work it, is learnt that locational attributes and structural attributes are prominent factors in predicting house prices. Studies suggest that there exists a close relationship between House pricing and locational attributes such as distance from the closest shopping center, train station, position offering views of hills or shore, the neighborhood in which the property is situated etc.

Structural attributes of the house like lot size, lot shape, quality and condition of the house, garage capacity, rooms, Lot frontage, number of bedrooms, bathrooms, overall finishing of the house etc play a big role in influencing the house price.

Neighbourhood qualities can be included in deciding house price. Factors like efficiency of public education, community social status, the socio-cultural demographics improve the worth of a property.

The demand side of the housing market is also a necessary component. Although population growth is widely known as a driver in housing demand, the key issue lies in the proportion of people with abundant financial resources.

Variables representing land value such as rents and material costs also demonstrate their influence in explaining house prices, which are positively related to housing prices.

Multiple regression analysis models allow to ascertain price predictions by capturing independent and dependent variable data. In Using multiple regression modelling techniques, we can describe changes brought to a dependent variable with changes in the independent variables.

In this research, various models were built in which the house Sale Price is projected as separate and dependent variable while locational, structural and various other attributes of housing properties were treated as independent variables. Therefore, the house price is set as a target or dependency variable, while other attributes are set as independent variables to determine the main variables by identifying the correlation coefficient of each attribute.

Motivation for the Problem Undertaken

There is a steady rise in house demand with every passing year, and consequently the house prices are rising every year. The problem arises when there are numerous variables such as location and property demand that influence the pricing. Therefore, buyers, sellers, developers and the real estate industry are keen to know the most important factors influencing the house price to help investors make sound decisions and help house builders set the optimal house price. There are many benefits that home buyers, property investors, and house builders can reap from the house-price model. This model aims to serve as a repository of such information and gainful insights to home buyers, property investors and house builders, that will help them determine best house prices. This model can be useful for potential buyers in deciding the characteristics of a house they want that best fits their budget and will be of tremendous benefit, especially to housing developers and researchers, to ascertain the most significant attributes to determine house prices and to acknowledge the best machine learning model to be used to conduct a study in this field.

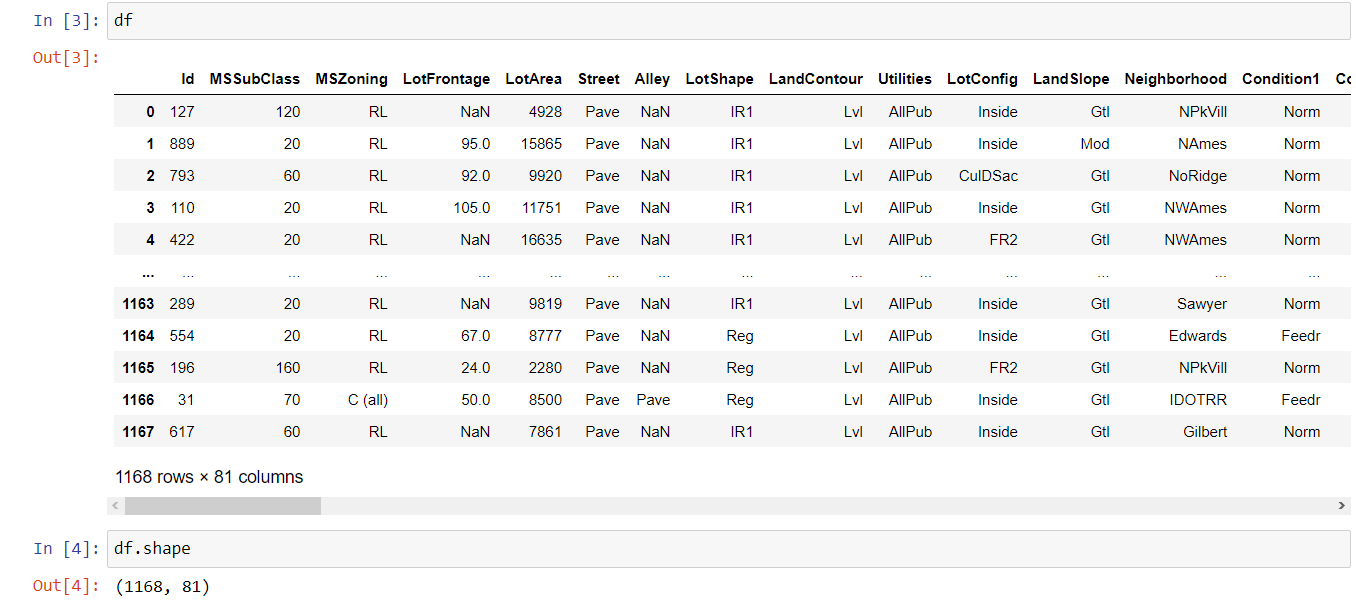
**Analytical Problem Framing**

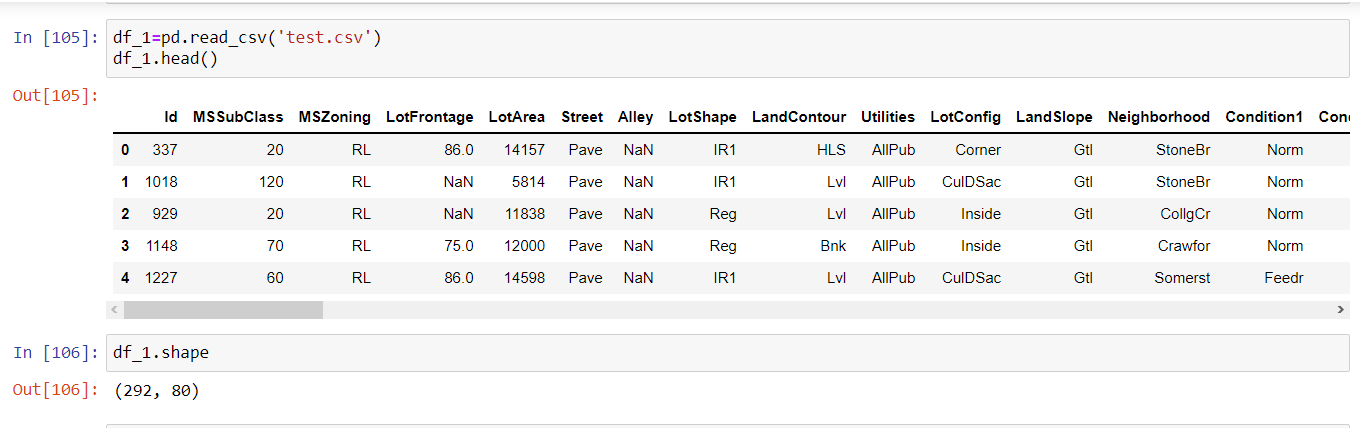
**Mathematical/ Analytical Modelling of the Problem**

In this project we have performed various mathematical and statistical analysis such as we checked description or statistical summary of the data using describe, checked correlation using corr() and also visualized it using heatmap. Then we have used zscore to plot outliers and remove them

Data Sources and their formats

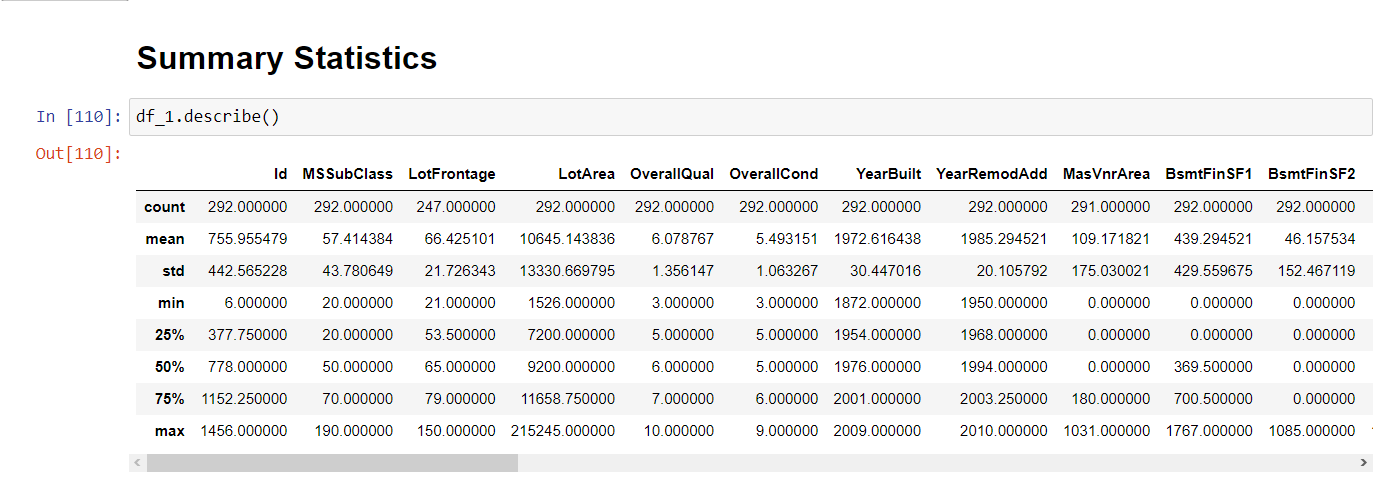
The data was provided to us by our client who is in the Housing Industry. The data was in the form of a CSV file.





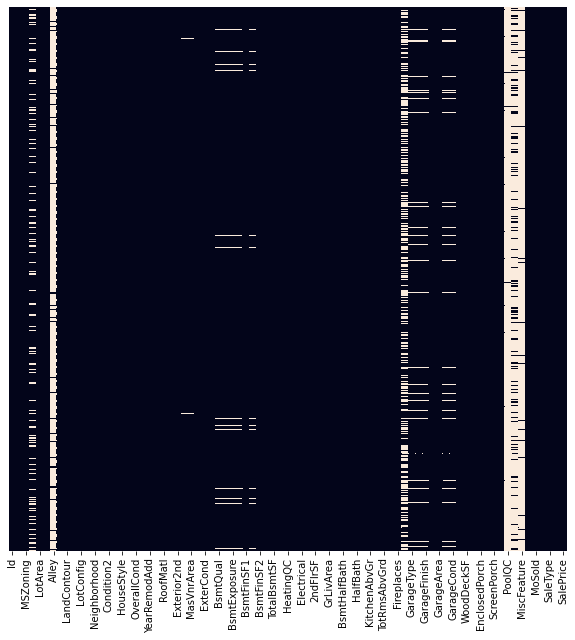
Training Dataset contains 1168 entries and 81 variables,

Test Dataset contains 292 entries and 80 variables.

****

Mean is more than median for SalePrice, MoSold, MiscVal, PoolArea, ScreenPorch, 3SsnPorch, EnclosedPorch, OpenPorchSF, WoodDeckSF, BsmtFinSF1, MasVnrArea, YearRemodAdd, OverallCond, OverallQual, LotArea, LotFrontage, MSSubClass and Id Column.

There is large difference between 75% and maximum for Price column.

****

LotFrontage, Alley, MasVnrType, MasVnrArea, BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinType2, FireplaceQu, GarageType, GarageYrBlt, GarageFinish, GarageQual, GarageCond, PoolQC, Fence and MiscFeature have null values.

**Dataset Description**

**The Independent Feature columns are:**

**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**

**WdSdng 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**

**WdSdng 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**

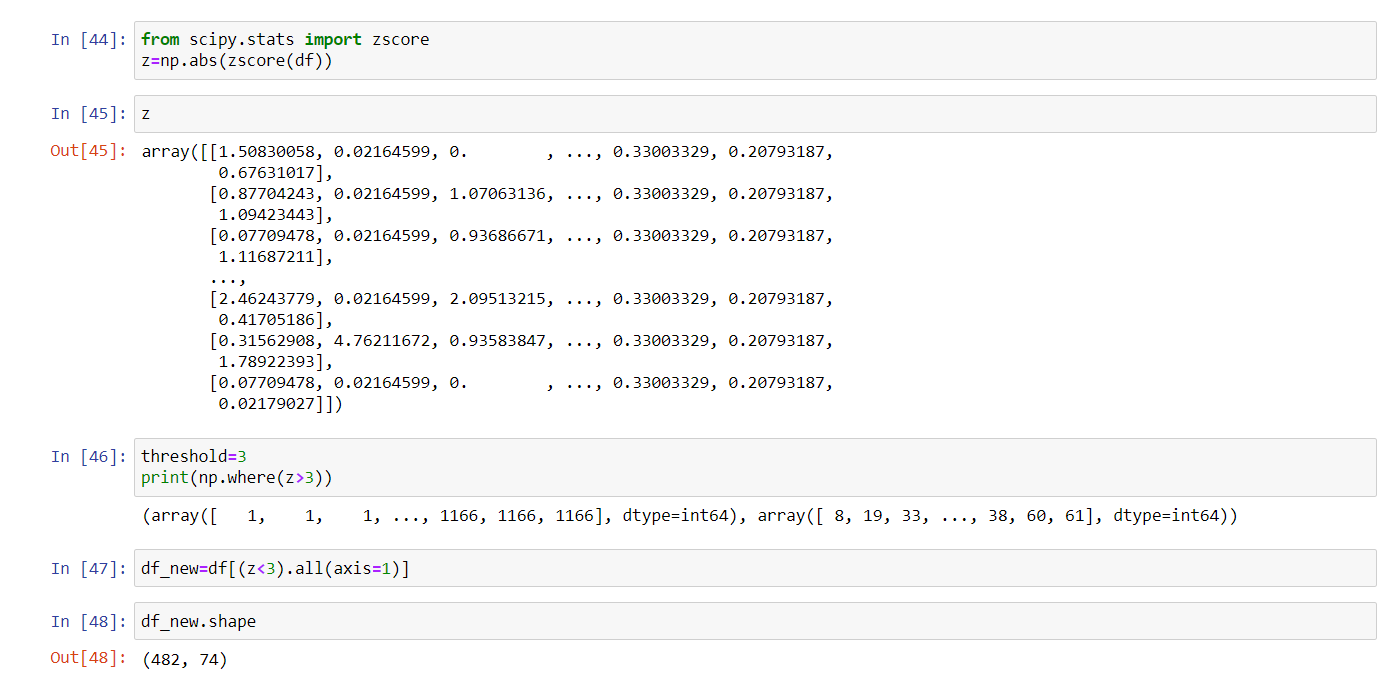
**AdjLand Adjoining Land Purchase**

**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)**

**Removing the Outliers using Z-score**

****

****

**Data Pre-processing Done**

First we will determine whether there are any null values and since there were null values as well as NaN vales present in the dataset we proceeded further by imputing them using Simple Imputer with mean and most frequent as strategies respectively. Next we did Label encoding using label encoder. Then we performed some data visualization in which we observed certain attributes were having skewness and outliers that were plotted using distplot and boxplot. Outliers were removed with the help of Zscore in which 685 rows were removed.

**Data Inputs- Logic- Output Relationships**

The data consists of 80 inputs and one output-“SalePrice”. MSSubClass,OverallCond,KitchenAbvGr,EnclosedPorch and Yr Sold are the least/negatively correlated column with target('SalePrice') variable. OverallQual is highly correlated column with target variable followed by GrLivArea and other attributes.

**Hardware and Software Requirements and Tools Used**

In this project we have used HP Pavilion PC with 64-bit operating system and have Windows 10 pro. We have used python to develop this project in which we have used various libraries such as numpy, pandas, matplotlib, seaborn for handling data or arrays and their visualization. For statistical purpose we have used zscore from scipy.stats to remove outliers. Lastly, to develop the model we have used various libraries and metrics from sklearn such as train\_test\_split, Linear Regression, Lasso, Ridge, Elastic Net, SVR, Decision Tree Regressor, KNeighborsRegressor, Random Forest Regressor, AdaBoost Regressor, Gradient Boosting Regressor, mean\_squared\_error, mean\_absolute\_error and r2\_score.

**Model/s DevelopmenT anD Evaluation**

**Identification of possible problem-solving approaches (methods)**

We have performed various mathematical and statistical analysis such as we checked description or statistical summary of the data using describe, checked correlation using corr and also visualized it using heatmap. Then we have used zscore to plot outliers and remove them. We have used distplot to find the distribution of all attributes.

**Testing of Identified Approaches (Algorithms)**

We have used following algorithms such as: LinearRegression, Lasso, Ridge, ElasticNet, SVR, DecisionTreeRegressor, KNeighborsRegressor, RandomForestRegressor, AdaBoostRegressor and GradientBoostingRegressor.

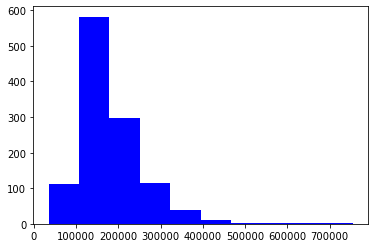
**Run and Evaluate selected models**

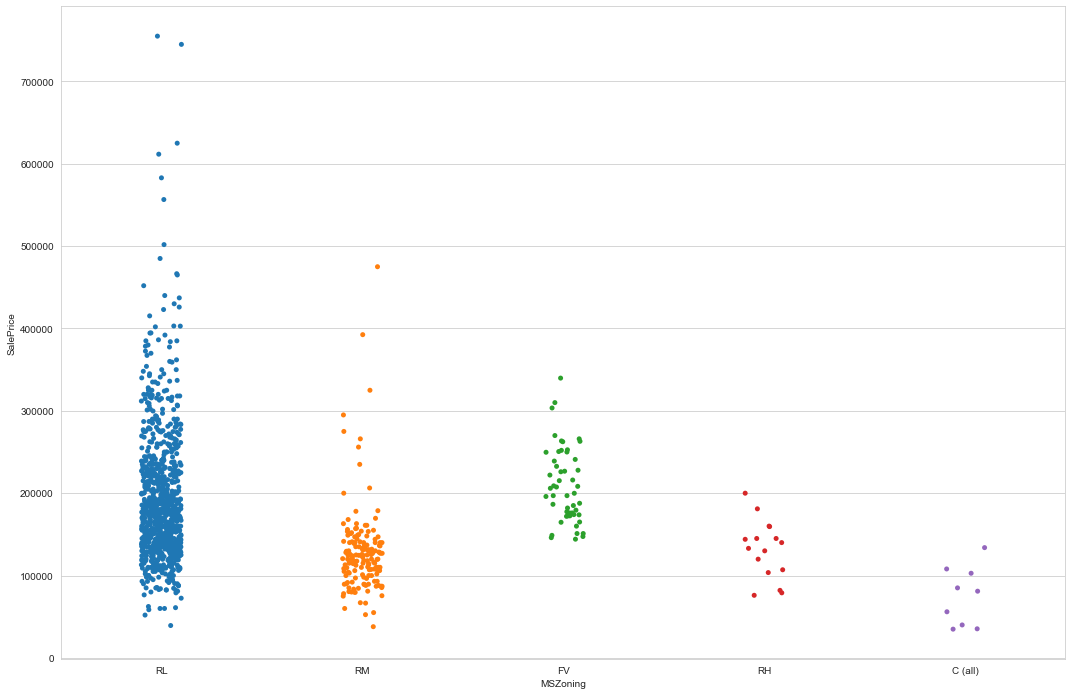
We have formed a loop where all the algorithms will be used one by one and their corresponding Score, Mean Absolute Error, Mean Squared Error, RMSE and r2\_score will be evaluated. • I chose GradientBoostingRegressor as our best model since it's giving us best score and it's performing well. It's r2\_score is also satisfactory and it shows that our model is neither underfitting/overfitting. Then we performed hyperparamter tuning using GridSearchCV on GradientBoostingRegressor from which got 'learning\_rate': 0.1, 'n\_estimators': 500 as best parameters. We got score : 0.999517991577412 after performing hyperparameter tuning and earlier it was 0.9846658425719441. Its r2\_score is also satisfactory.

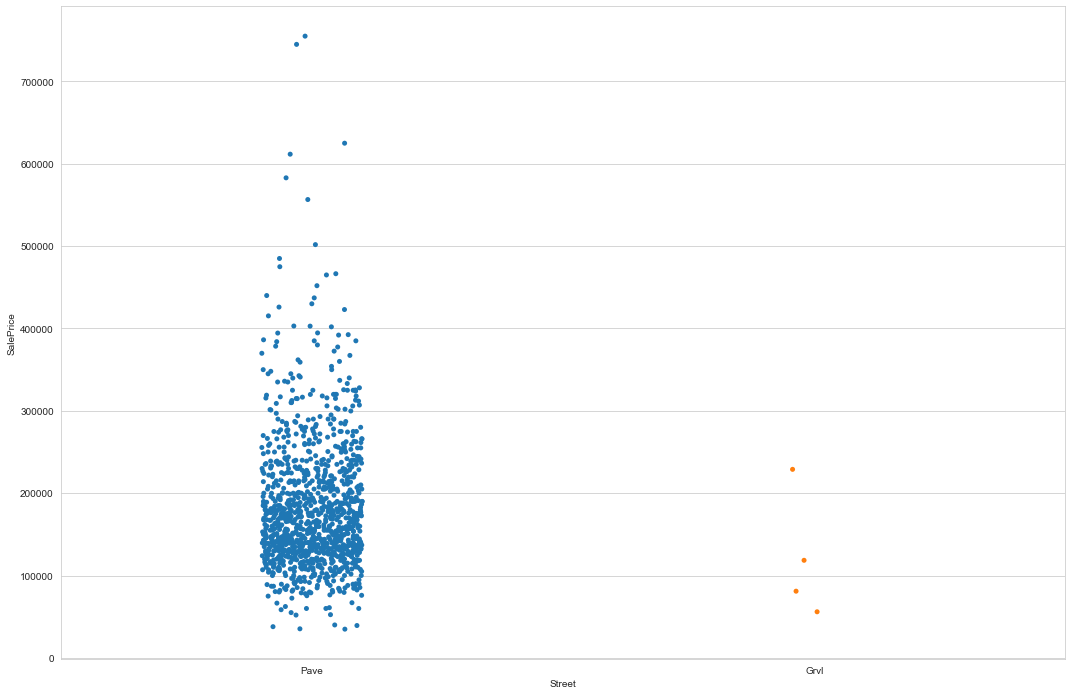
Hence we saved GradientBoostingRegressor as our final model using joblib.

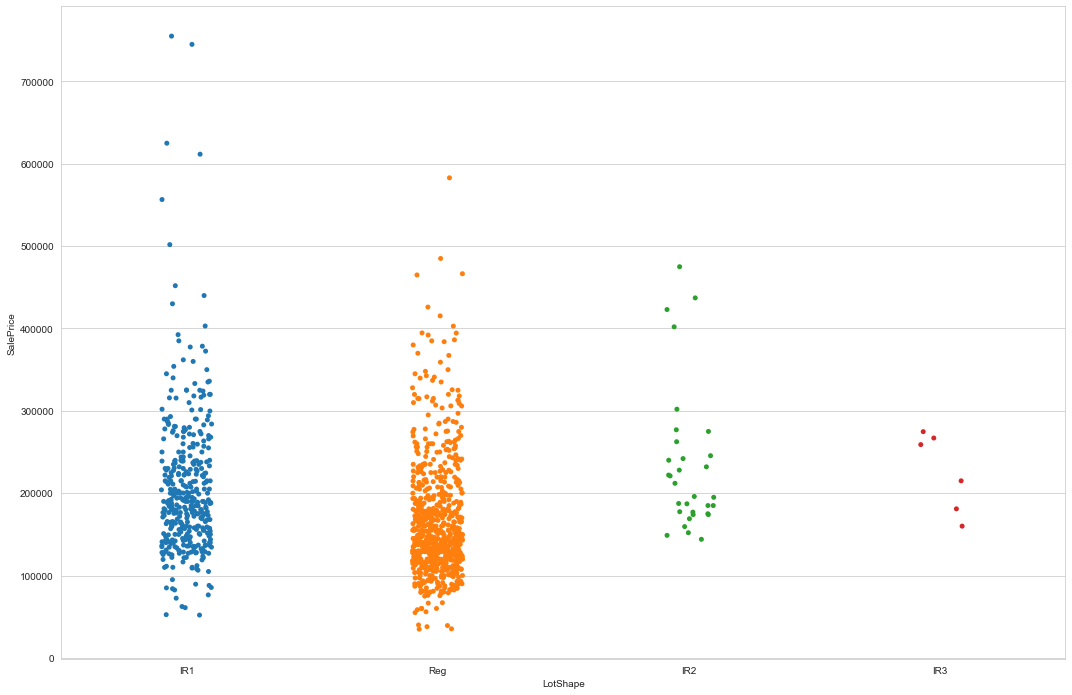
**Key Metrics for success in solving problem under consideration**

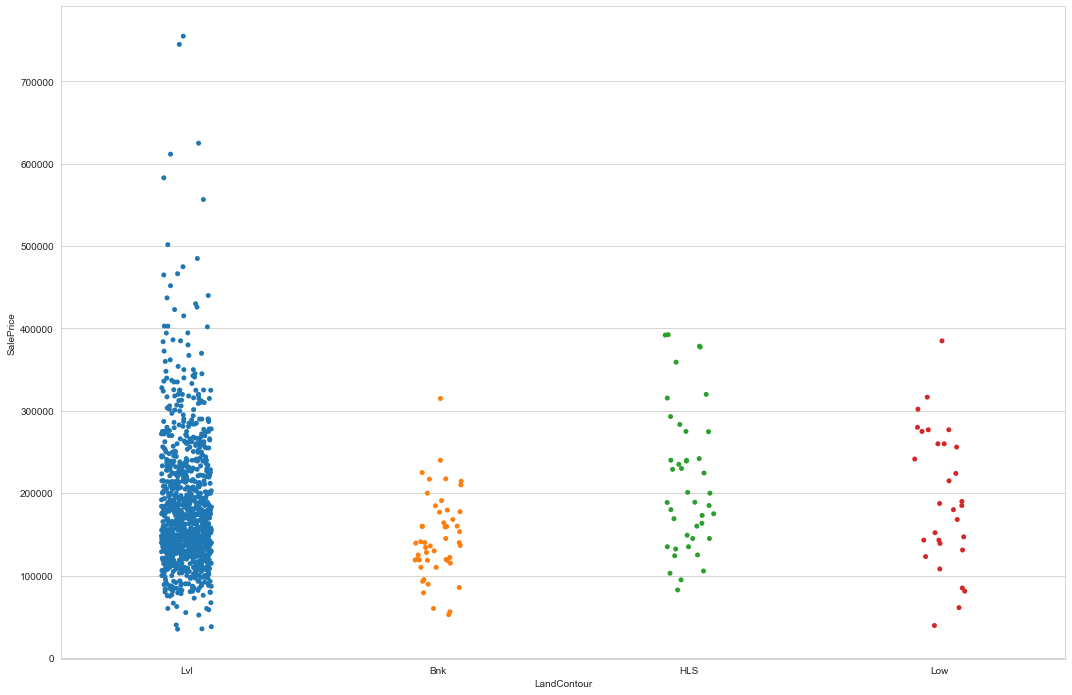
Key metrics used for finalising the model was Score and r2\_score. Since in case of GradientBoostingRegressor it's giving us good score among all other models and it's performing well. It's r2\_score is also satisfactory and it shows that our model is neither underfitting/overfitting .

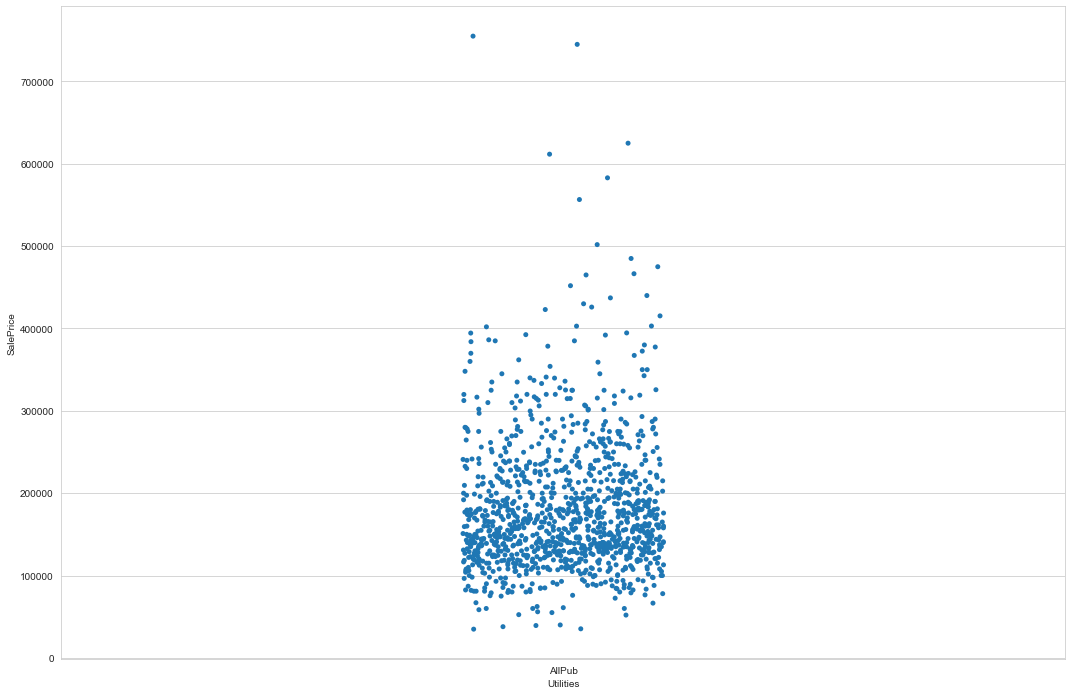




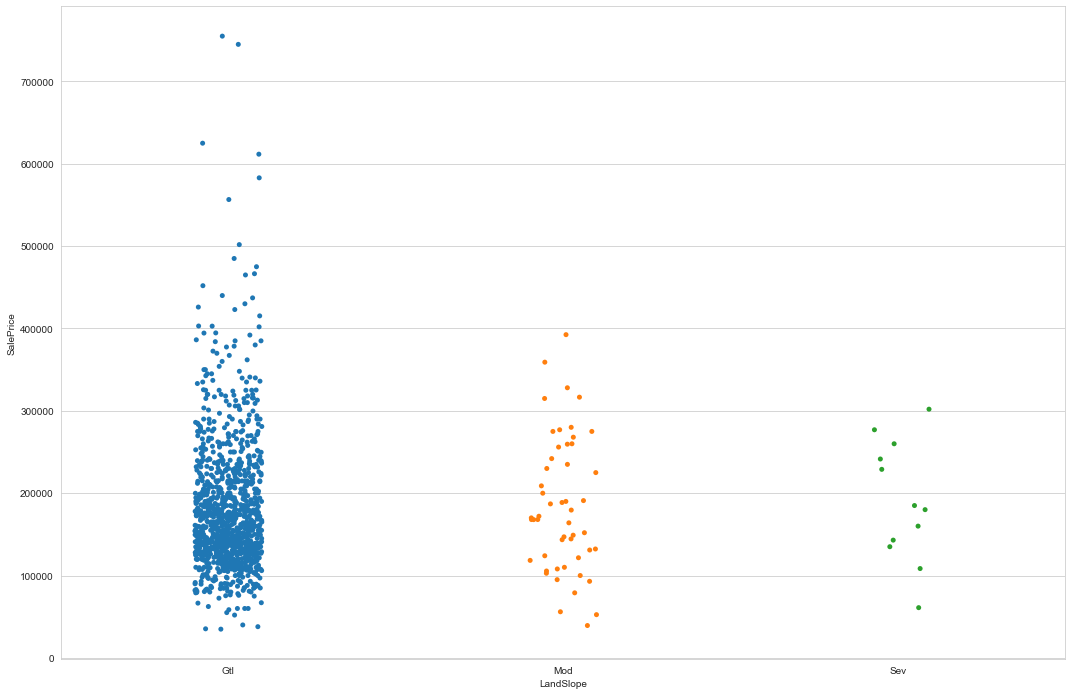


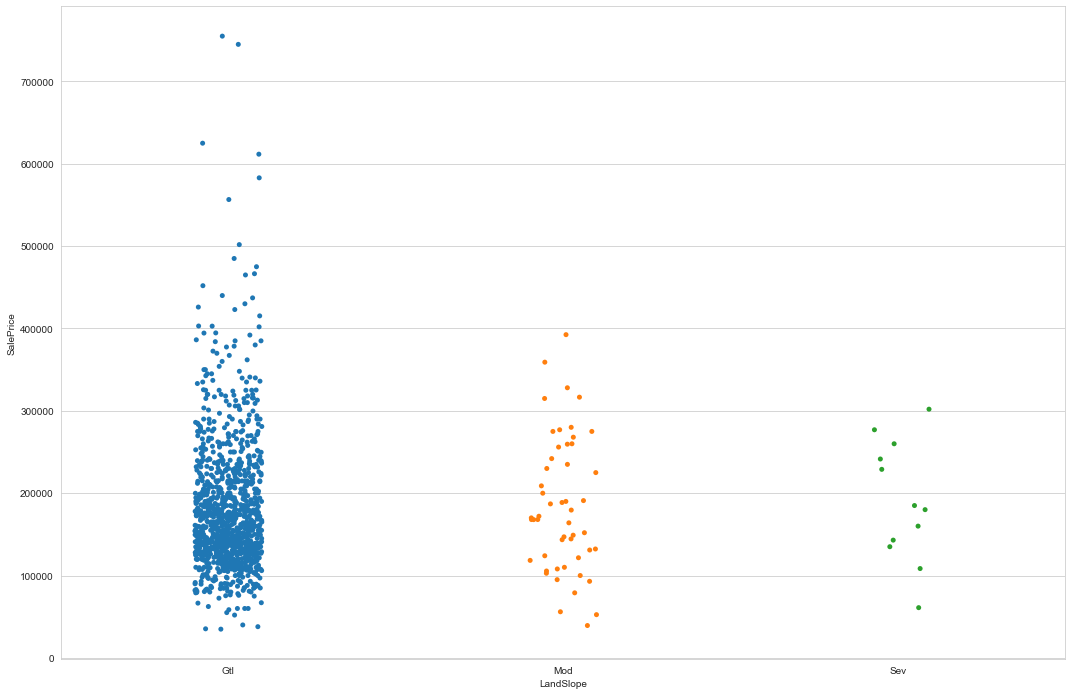


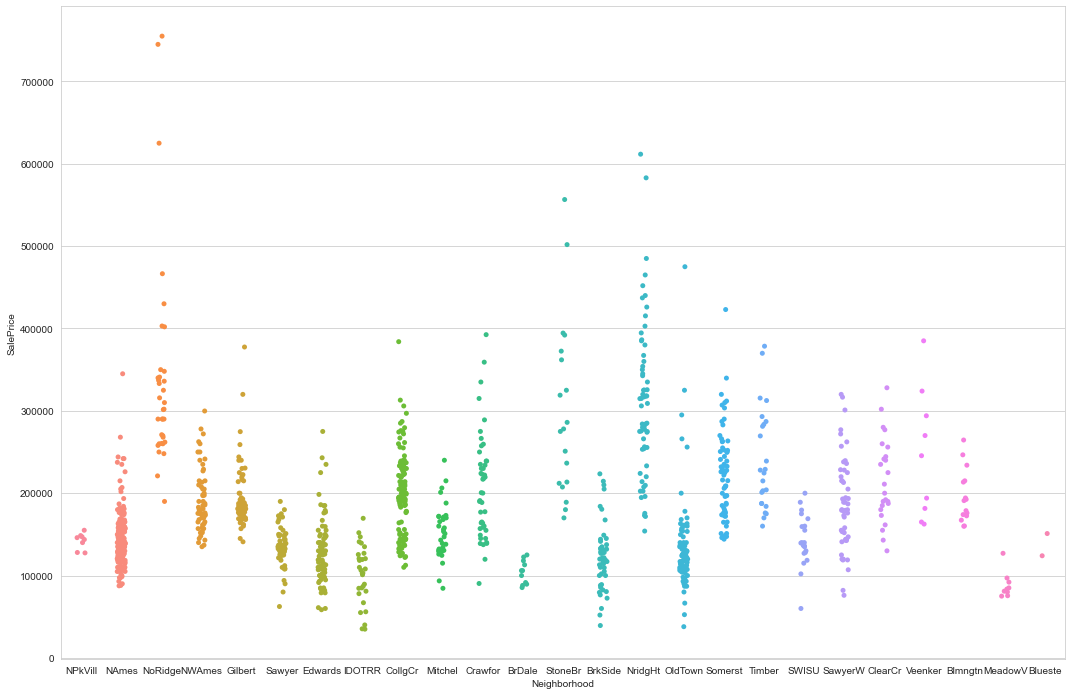
****

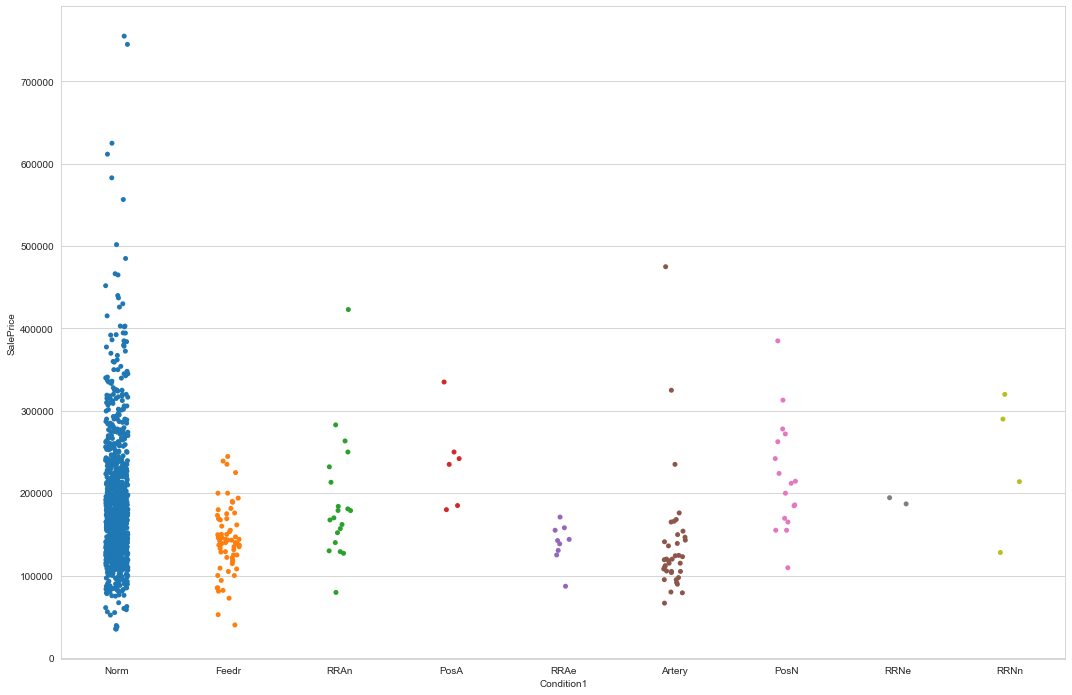
****

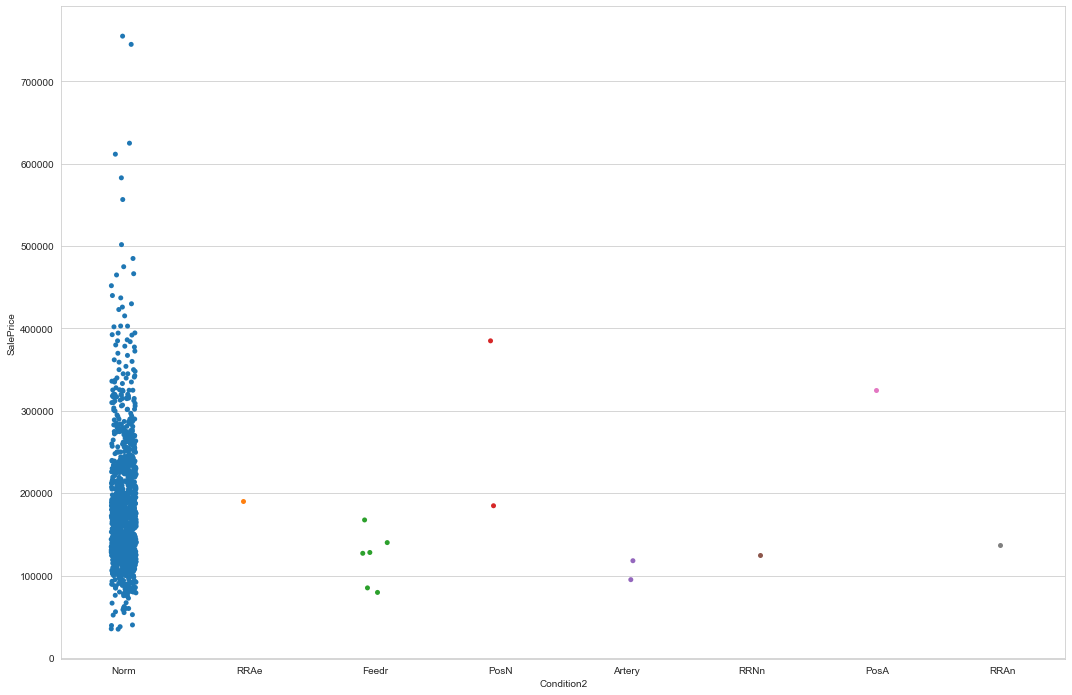
****

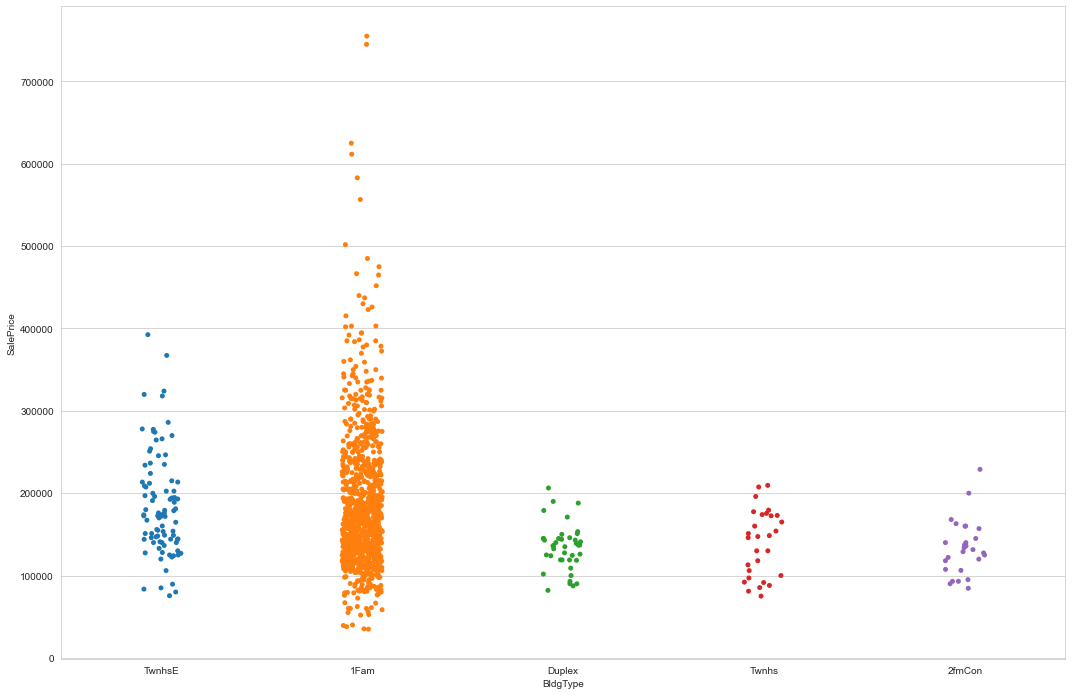
****

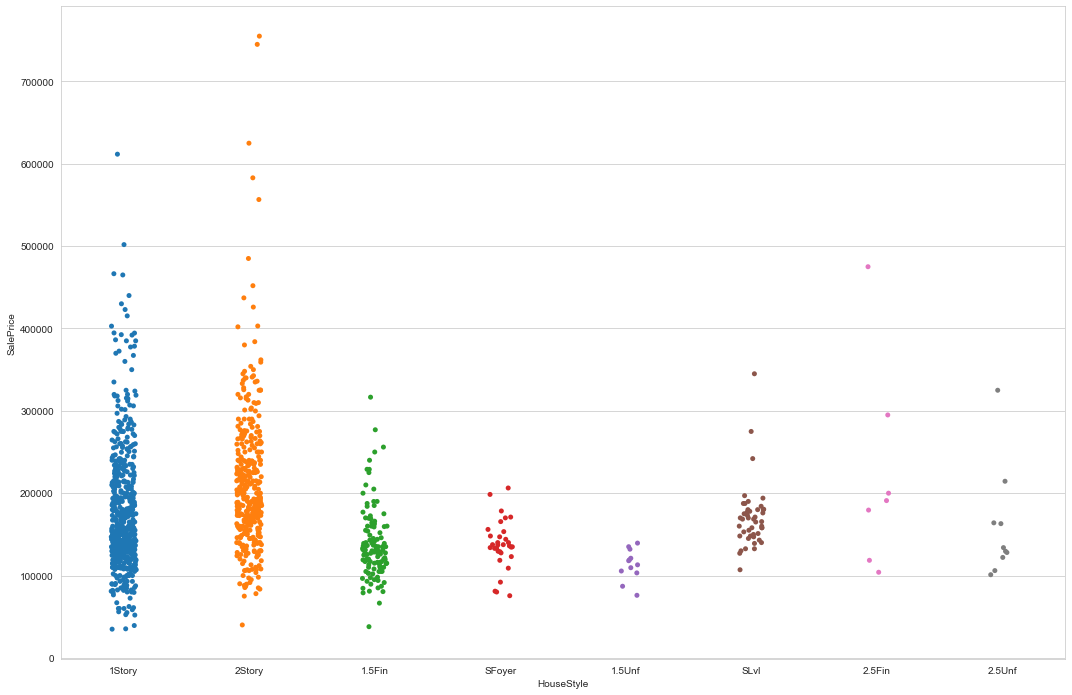
****

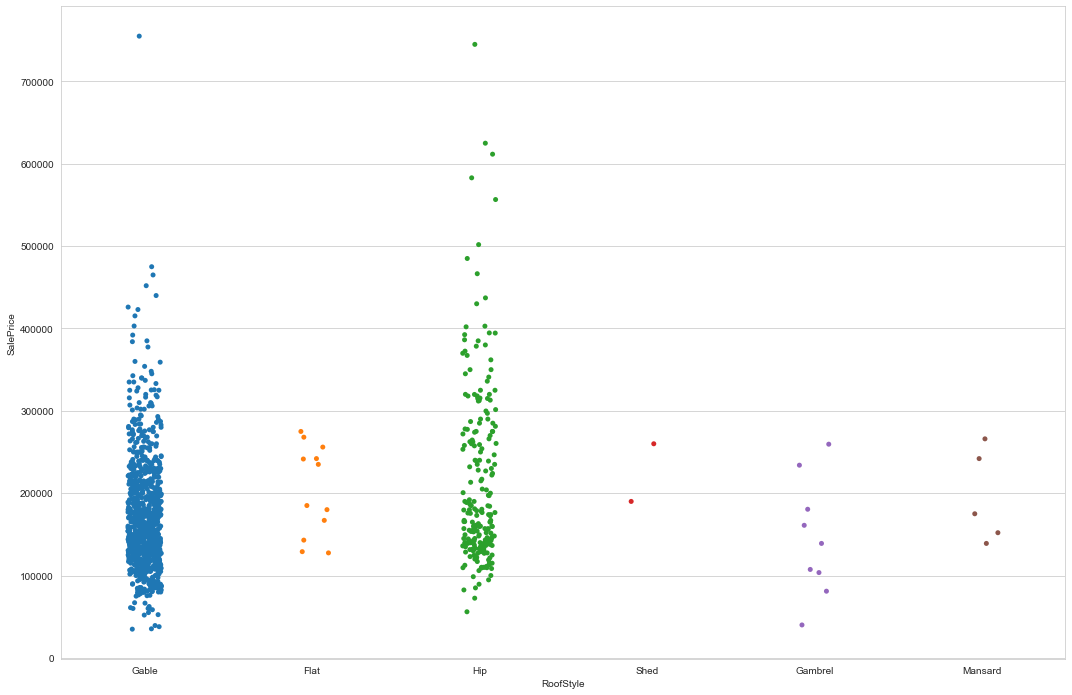
****

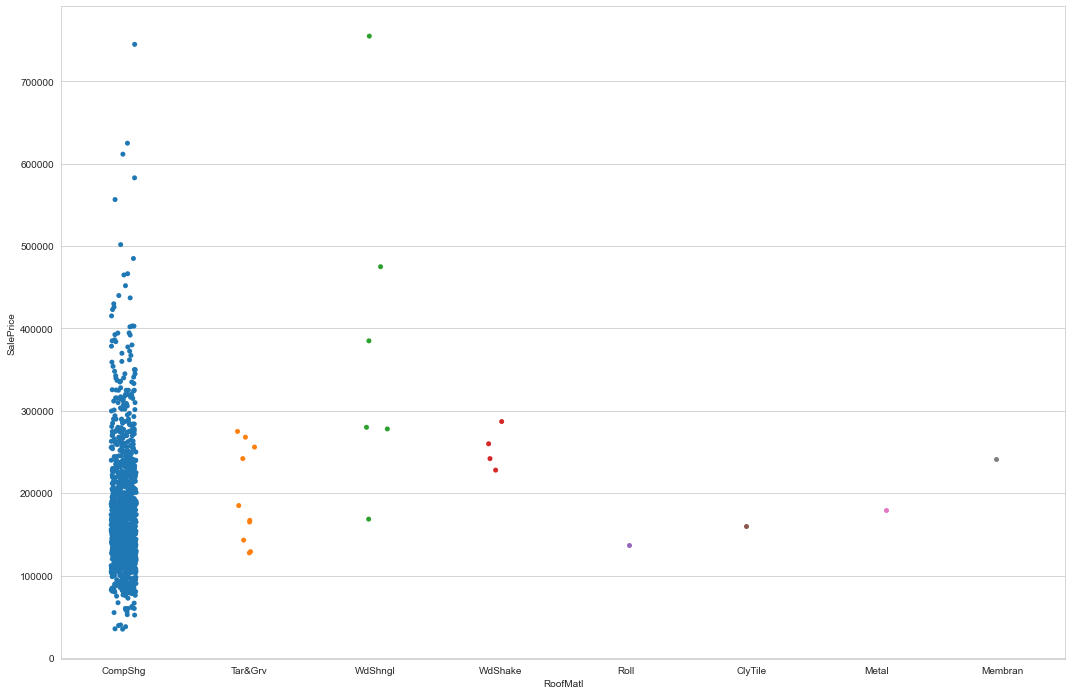
****

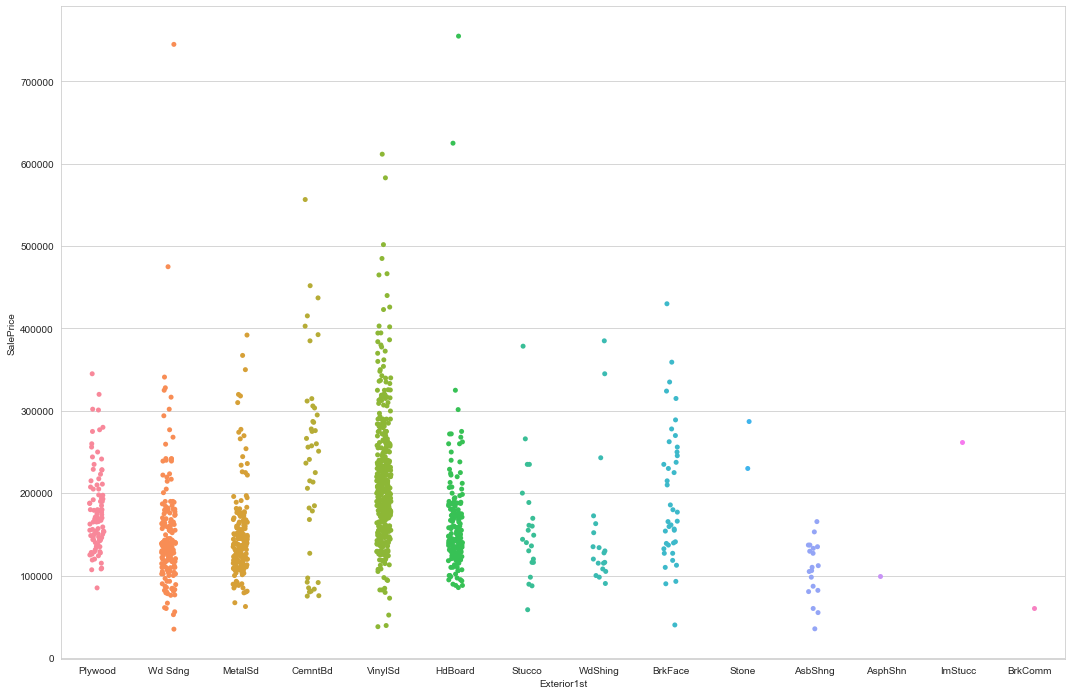
****

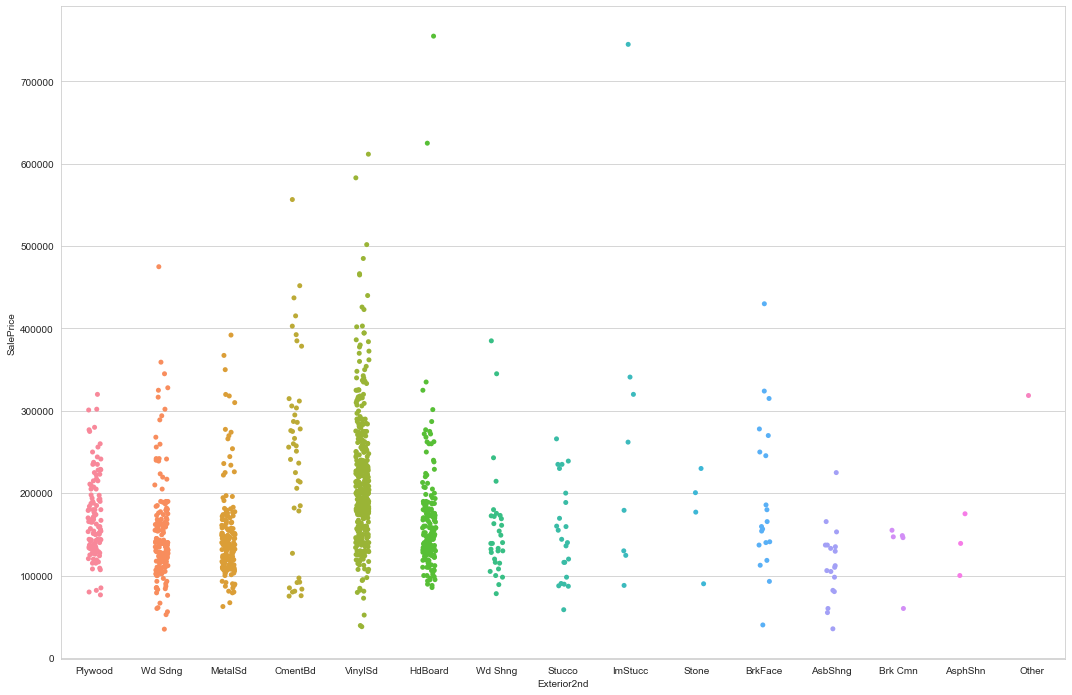
****

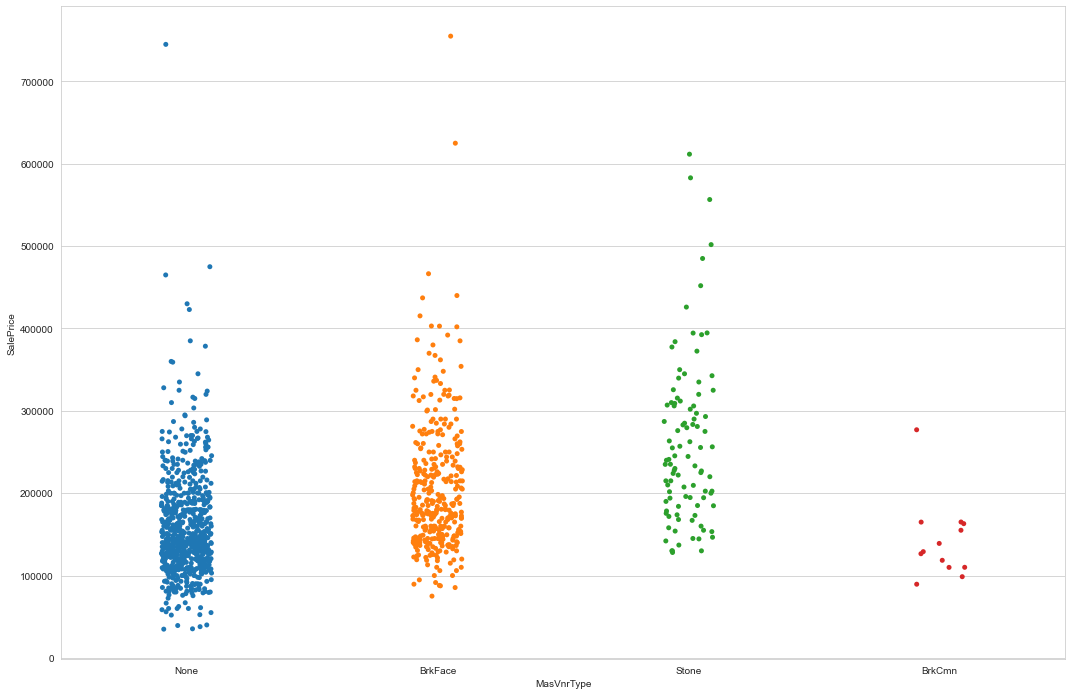
****

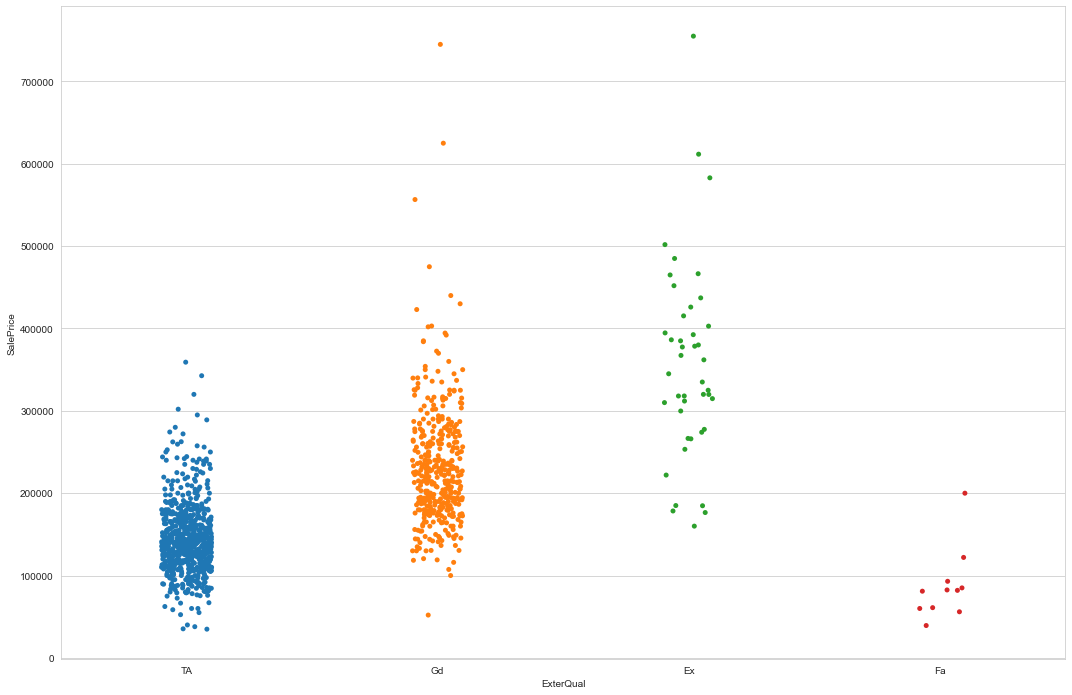
****

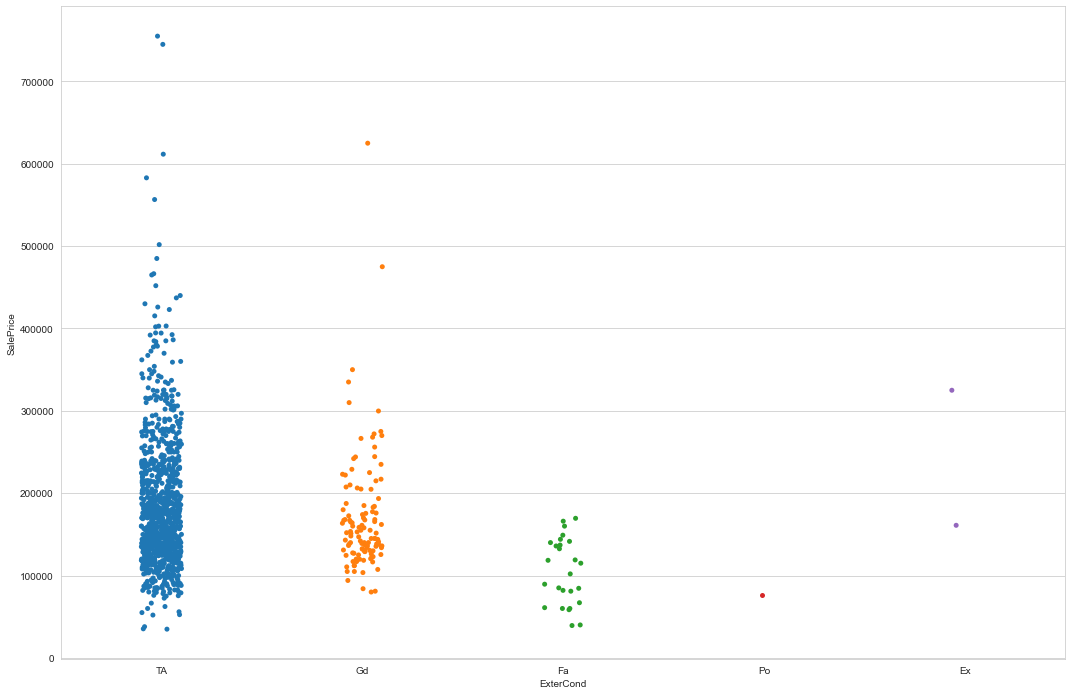
****

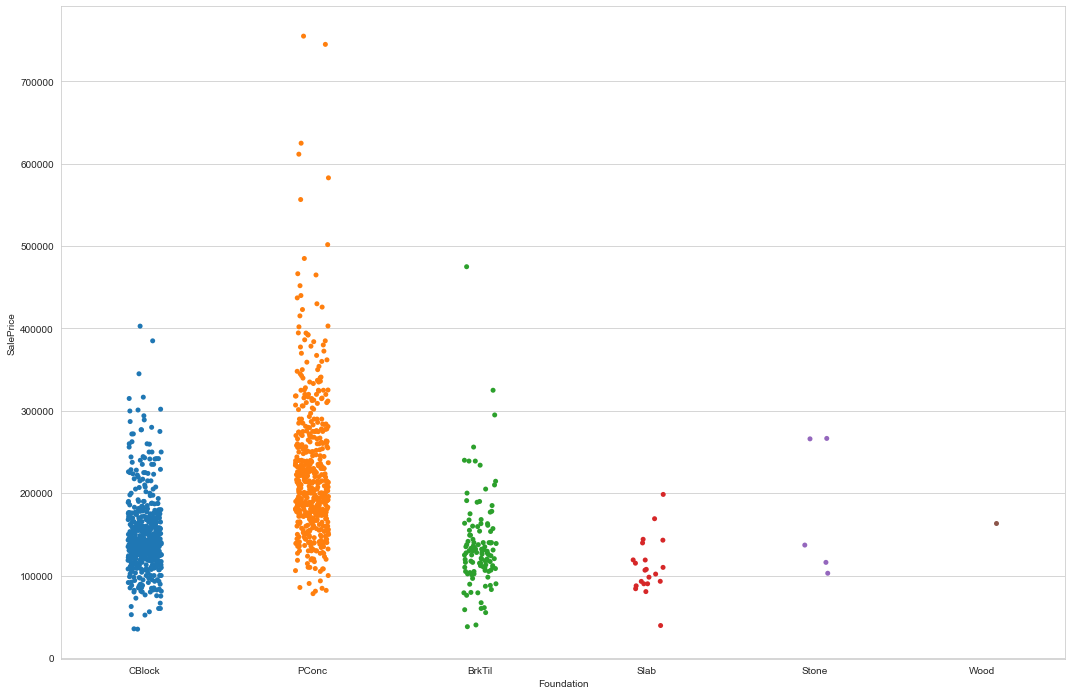
****

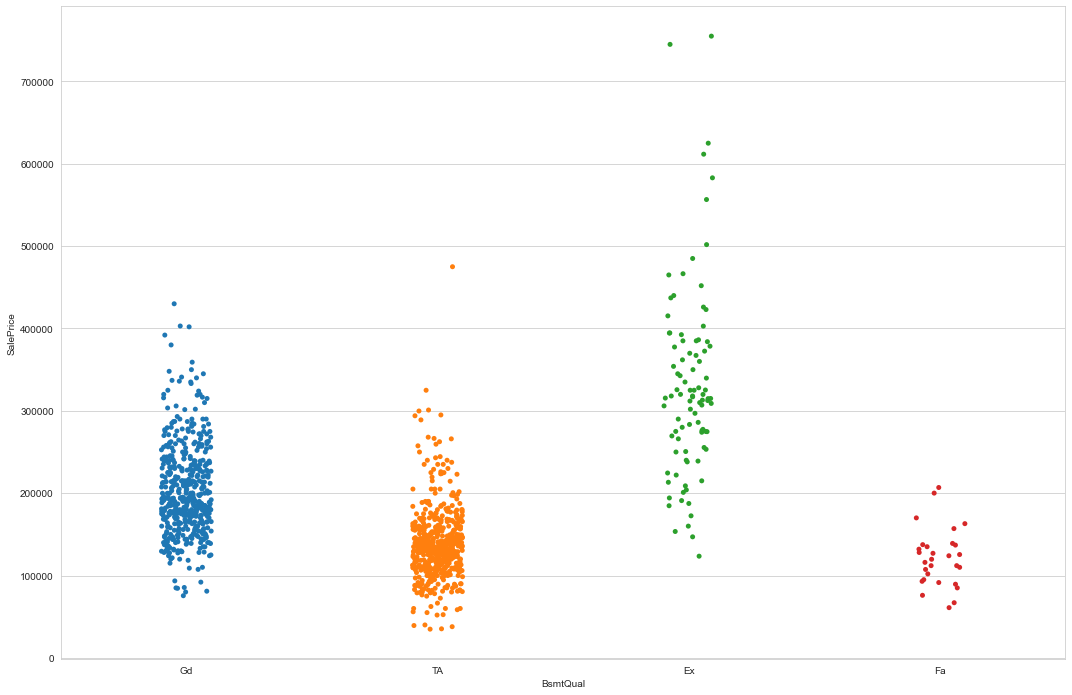
****

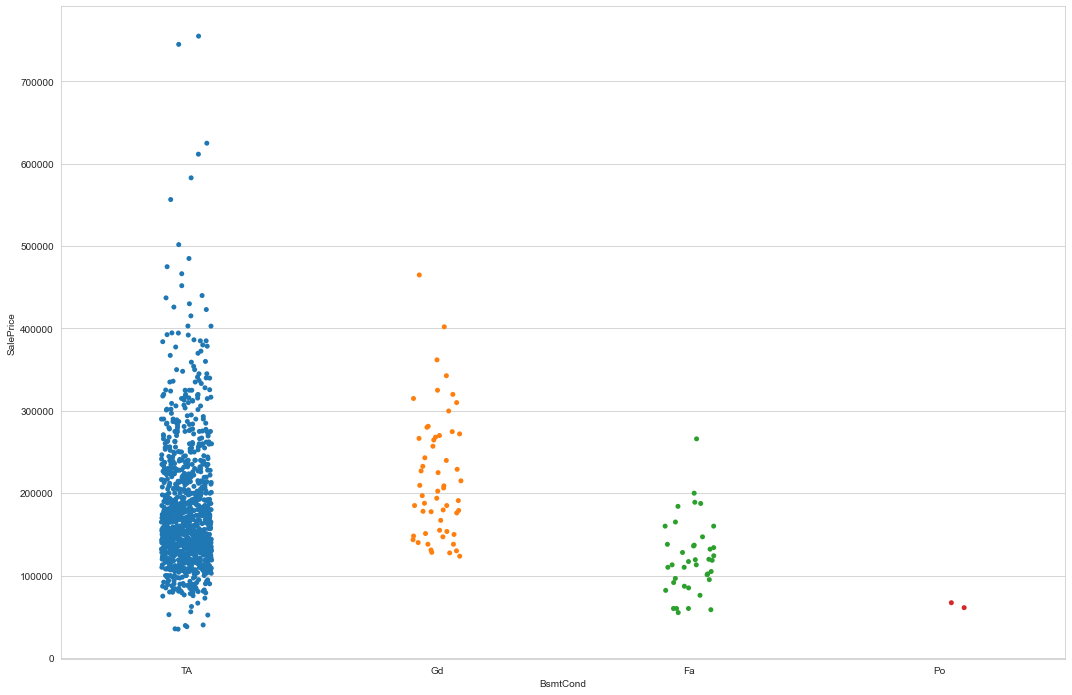
****

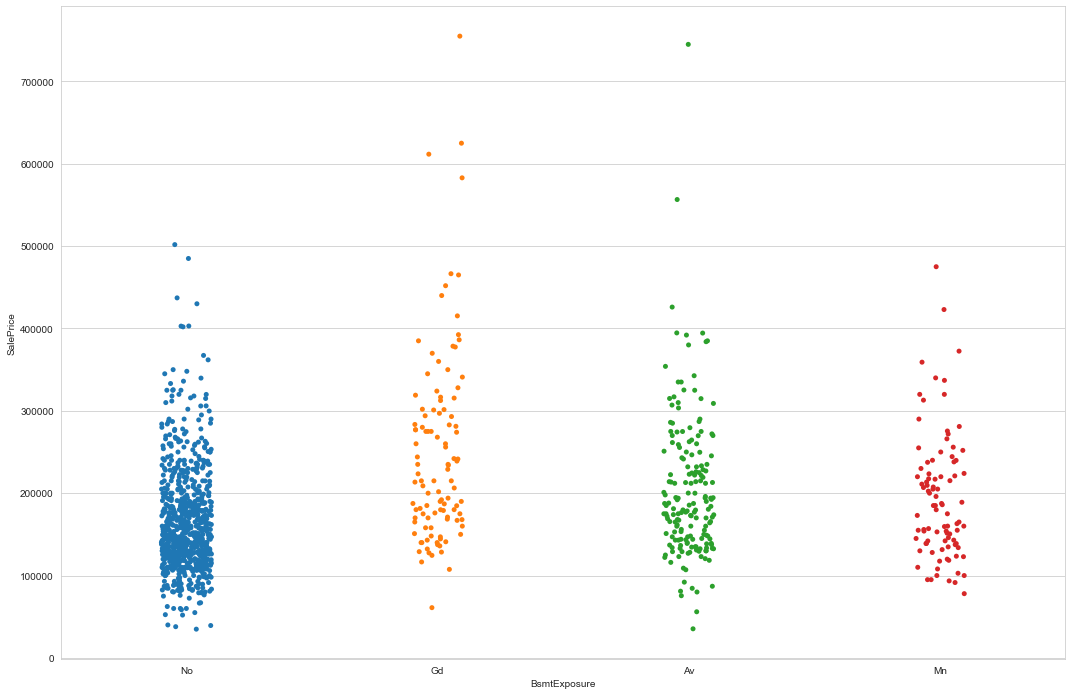
****

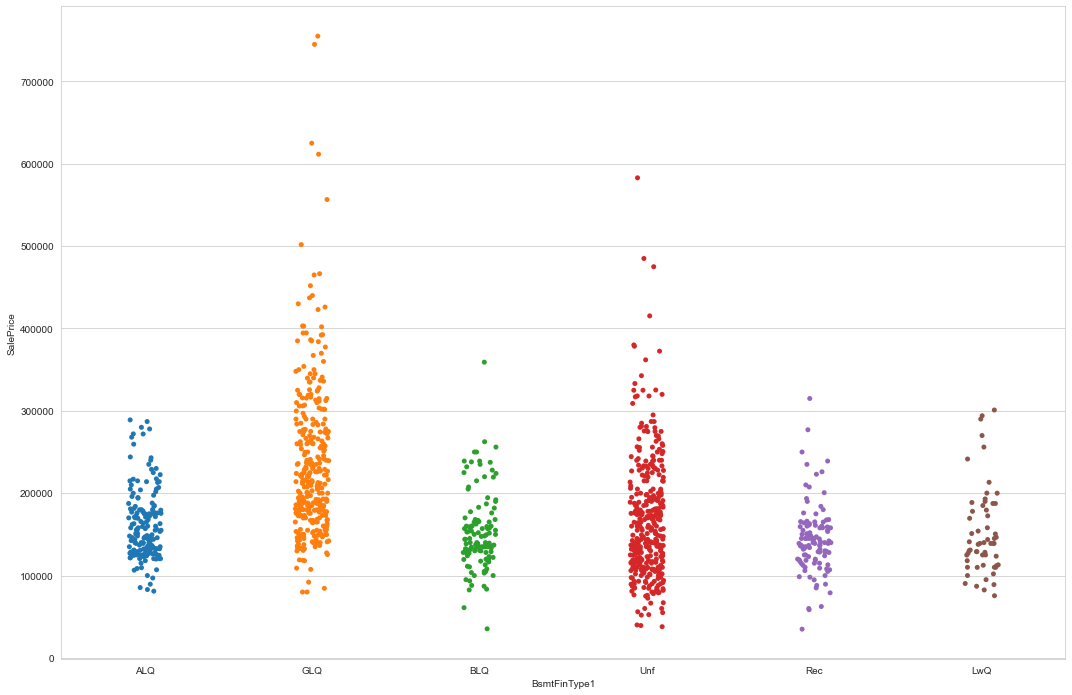
****

****

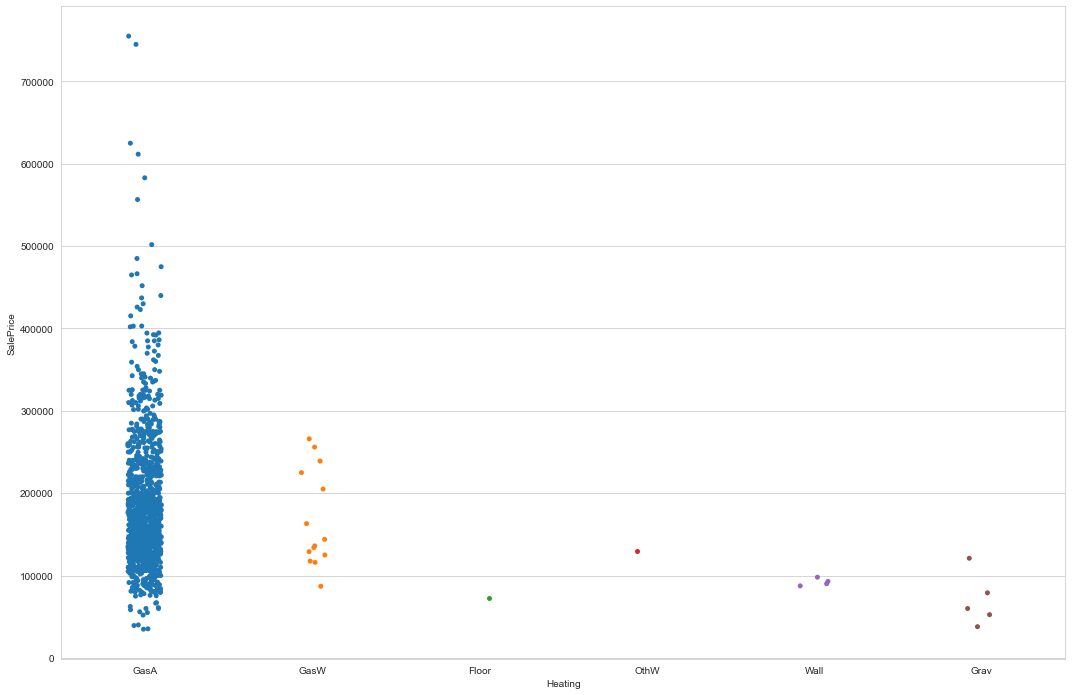
****

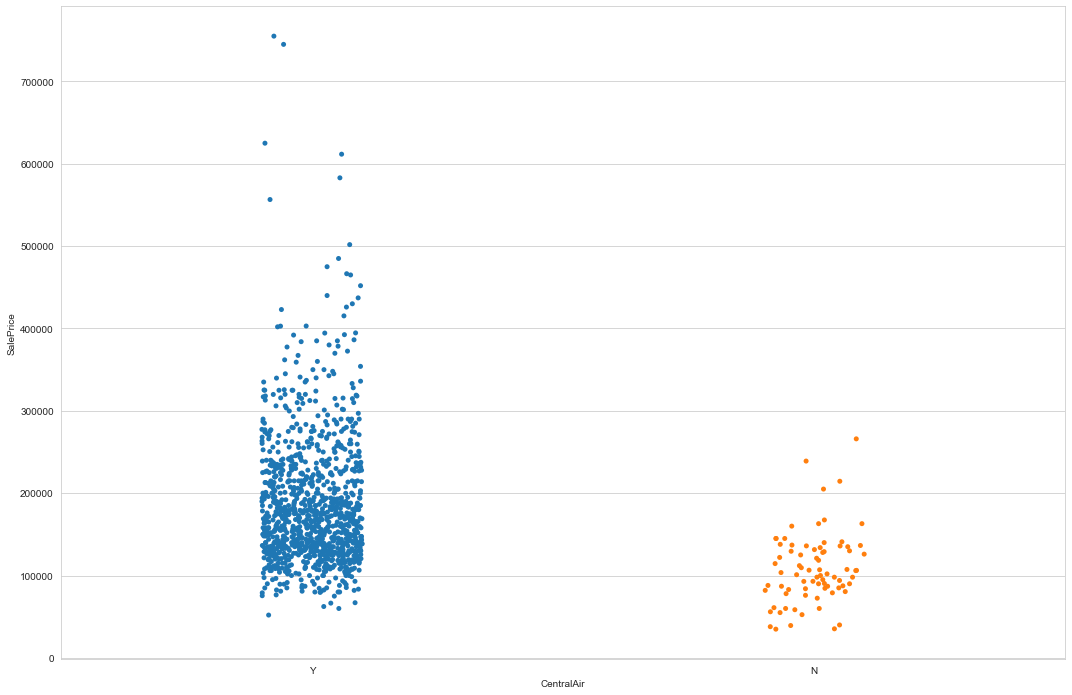
****

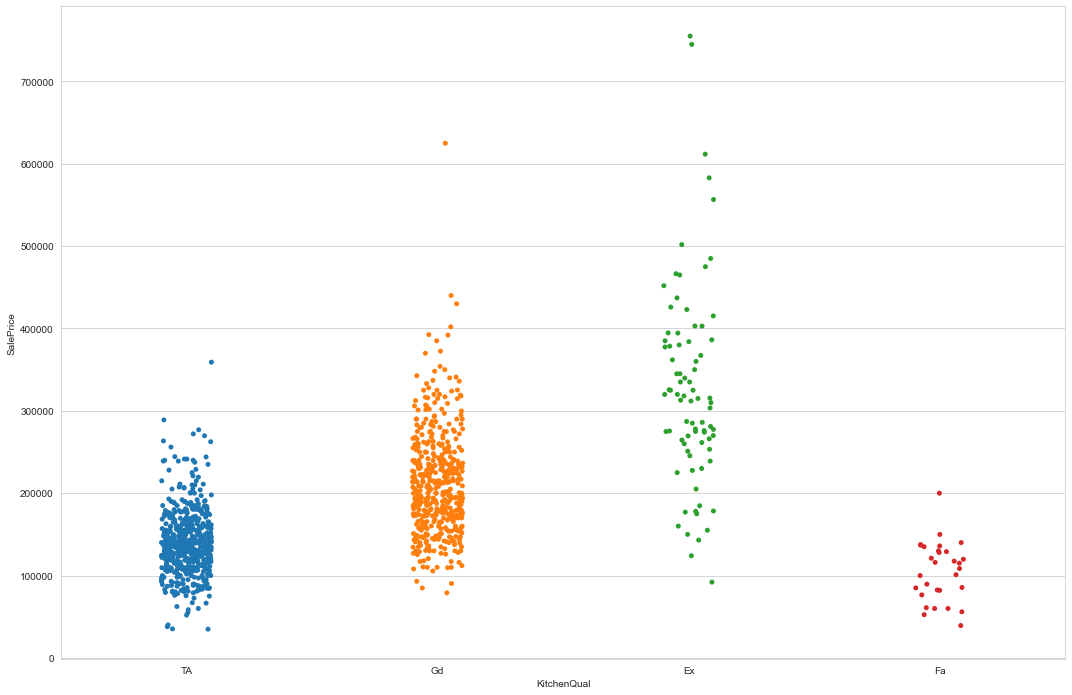
****

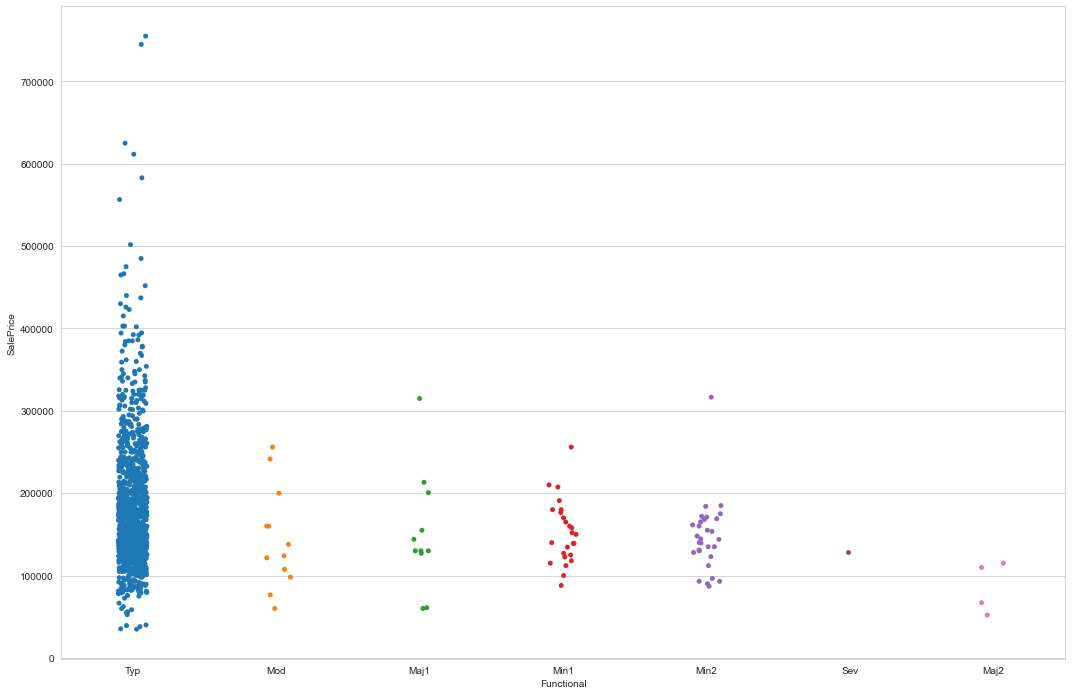
****

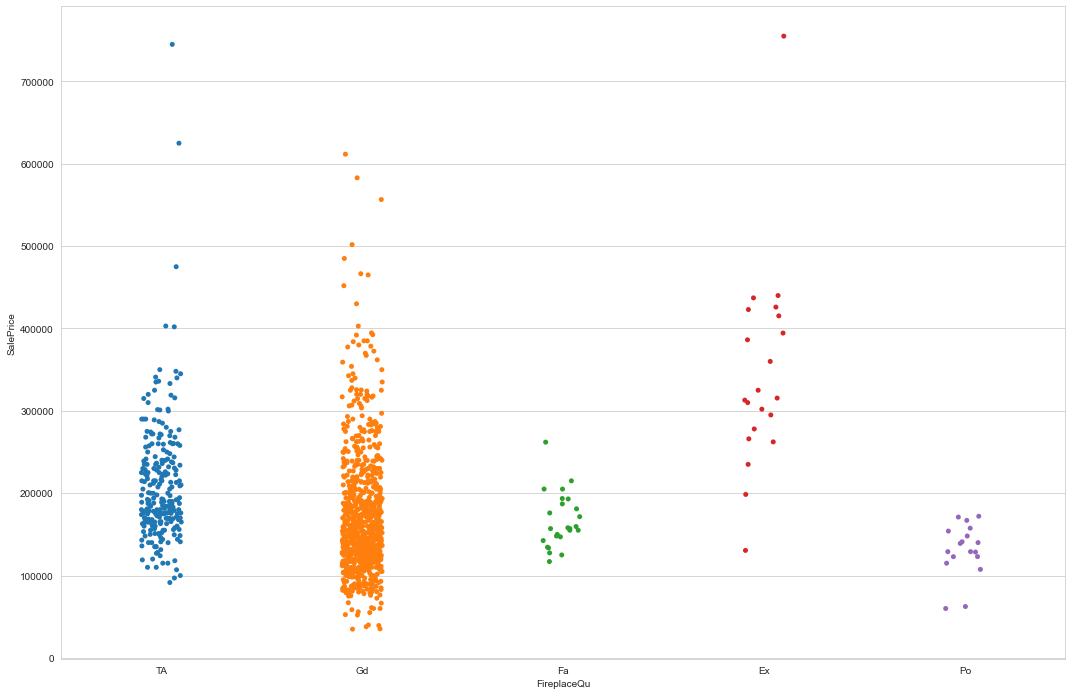
****

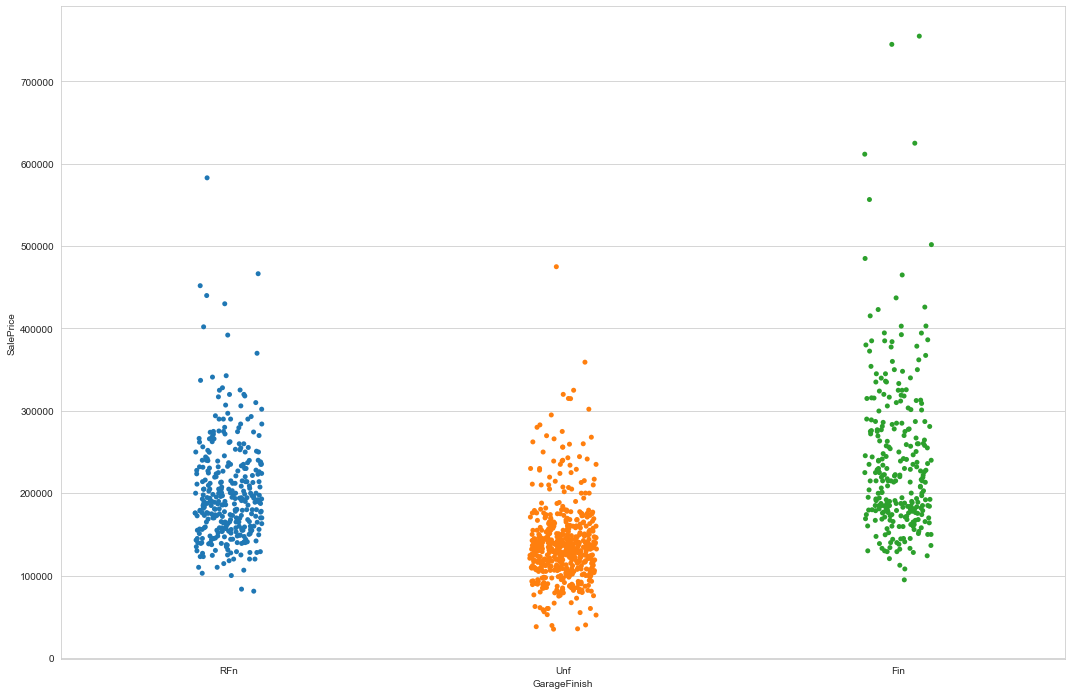
****

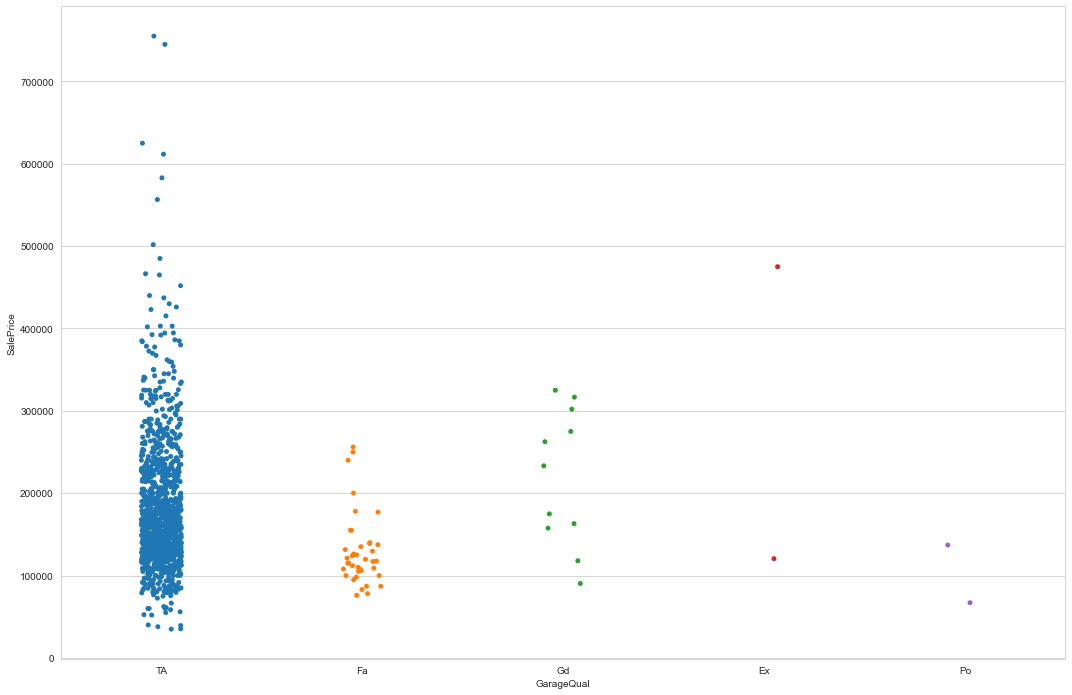
****

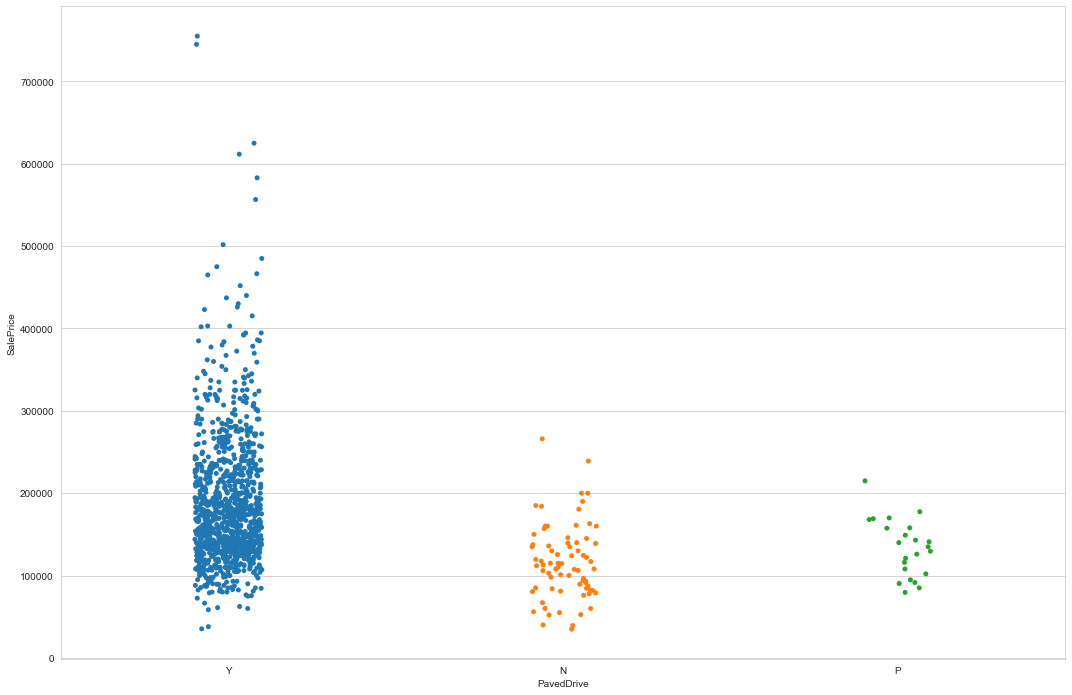
****

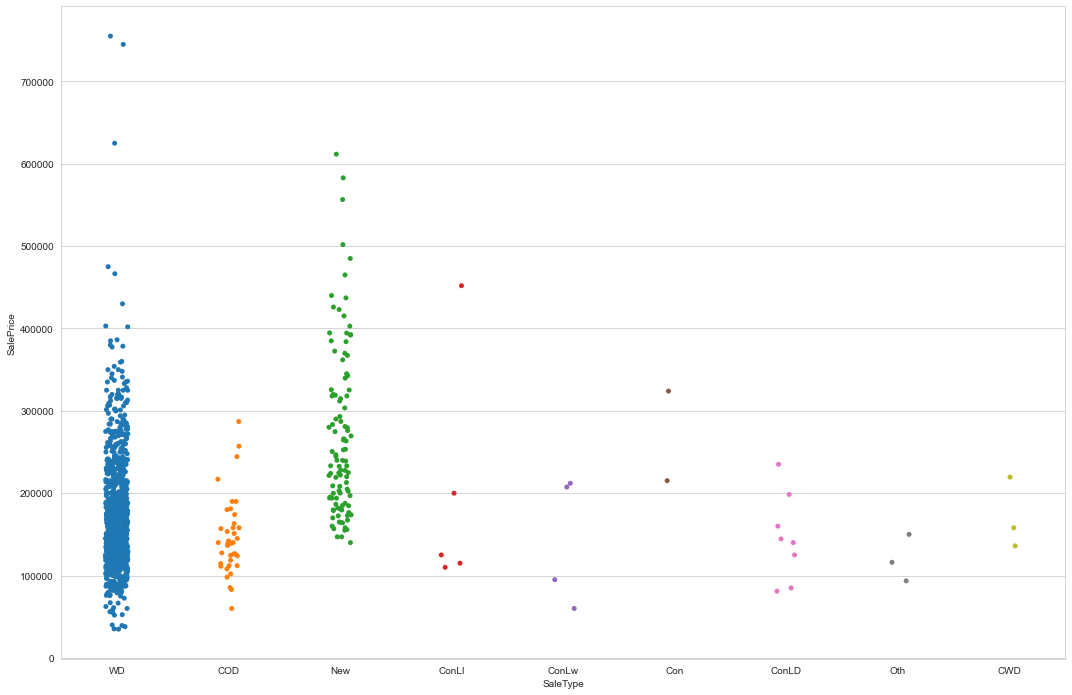
****

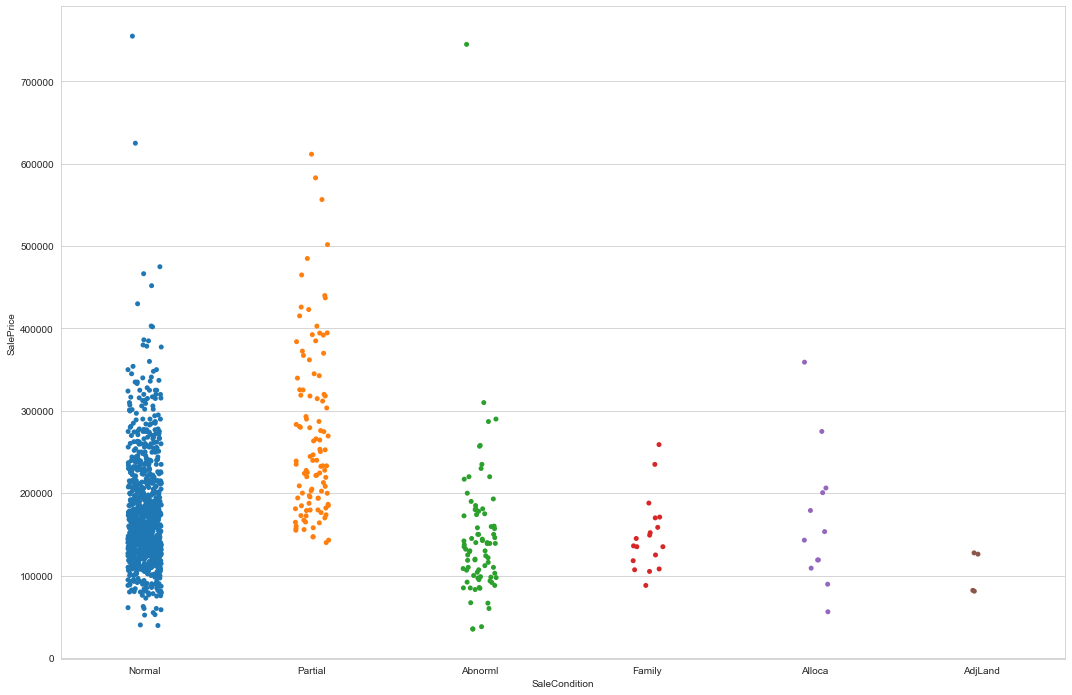
****

****

****

****

****

****

**Interpretation of the Results**

•Least SalePrice is for 30:1-STORY 1945 & OLDER and maximum for 60:2-STORY 1946 & NEWER

• In MSZoing maximum is for category 1 i.e, Floating Village Residential

•Lotshape 1 and 2 have almost similar price and 3 has least.

•Landconotur corresponding to 1 i.e, HLS Hillside - Significant slope from side to side has maximum price.

•Lotconfig corresponding to 1 and 3 have similar price.

•Neighborhoot with (15)NPkVillNorthpark Villa has maximum sales price and (10)IDOTRR Iowa DOT and Rail Road has least.

• Normal condition houses have highest saleprice

• 1Fam Single-family Detached and TwnhsI Townhouse Inside Unit have maximum saleprice.

• In HouseStyle category 3: 2Story Two story has max sale price.

• In OverallQual: SalePrice increase as Ratings increase.

•Similary for OverallCond 5 and 9 have max sale price

• In RoofStyle 5:Shed has maximum.

• In Exterior1st 6:HardBoard and 9:Other have Saleprice

• In Exterior2nd 8:MetalSd Metal Siding

• In MasVnrType, 3:stone has max saleprice and 0:BrkCmn Brick Common has least

• In ExterQual 0:Excellent has maximum price. Similary for ExterCond

• In Foundation 2:PConc Poured Contrete has max price

• In BsmtQual 0: Ex Excellent (100+ inches), In BsmtCond 1: Gd Good, In BsmtExposure 1: Av Average Exposure (split levels or foyers typically score average or above) have max sale prices

• In BsmtFinType1: Rating of basement finished area - 2:GLQ Good Living Quarters has max price

• In HeatingQC: Heating quality and condition 0:Ex Excellent has max price.

• Houses with CentralAir has higher saleprice

• In FireplaceQu: Fireplace quality 0:Ex Excellent - Exceptional Masonry Fireplace has max saleprice.

•GarageType 3:BuiltIn Built-In (Garage part of house - typically has room above garage) has max saleprice

• Finished Garage has more price

• Paved Driveway has more price

• In 2007 maximum houses are sold followed by 2006

• In saletype category 2 and 6 have max sale price

• Normal sale condition has max price.

**CONCLUSION**

* Key Findings and Conclusions of the Study

In this project we have tried to show how the house prices vary and what are the factors related to the changing of house prices.Thebest(minimum) RMSE score was achieved using the best parameters of Ridge Regressor through GridSearchCV though Lasso Regressor model performed well too.

* Learning Outcomes of the Study in respect of Data Science

This project has demonstrated the importance of sampling effectively, modelling and predicting data.

Through different powerful tools of visualization we were able to analyse and interpret different hidden insights about the data.

Through data cleaning we were able to remove unnecessary columns and outliers from our dataset due to which our model would have suffered from overfitting or underfitting.

The few challenges while working on this project where:-

* + Improper scaling
  + Too many features
  + Missing values
  + Skewed data due to outliers

The data was improper scaled so we scaled it to a single scale using sklearns’s package StandardScaler.

There were too many(256) features present in the data so we applied Principal Component Analysis(PCA) and found out the Eigenvalues and on the basis of number of nodes we were able able to reduce our features upto 90 columns.

There were lot of missing values present in different columns which we imputed on the basis of our understanding.

The columns were skewed due to presence of outliers which we handled through winsorization technique.

* Limitations of this work and Scope for Future Work

While we couldn’t reach out goal of minimum RMSE in house price prediction without letting the model to overfit, we did end up creating a system that can with enough time and data get very close to that goal. As with any project there is room for improvement here. The very nature of this project allows for multiple algorithms to be integrated together as modules and their results can be combined to increase the accuracy of the final result. This model can further be improved with the addition of more algorithms into it. However, the output of these algorithms needs to be in the same format as the others. Once that condition is satisfied, the modules are easy to add as done in the code. This provides a great degree of modularity and vesatility to the project.