PFA HOUSING PROJECT

Submitted by:

PRAKHAR PRAKASH

ACKNOWLEDGMENT

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped you and guided you in completion of the project.

In this project(PFA HOUSING) I took the help of a few references and techniques from the github account of KRISH C NAIK(krishnaik06) who is a data scientist in machine learning and deep learning experience in the feature selection part which was used just to depict the features of the highest value in predicting the SalePrice and not in model training or testing.

INTRODUCTION

Technical Requirements: • Data contains 1460 entries each having 81 variables. • Data contains Null values. You need to treat them using the domain knowledge and your own understanding. • Extensive EDA has to be performed to gain relationships of important variable and price. • Data contains numerical as well as categorical variable. You need to handle them accordingly. • You have to build Machine Learning models, apply regularization and determine the optimal values of Hyper Parameters. • You need to find important features which affect the price positively or negatively. • Two datasets are being provided to you (test.csv, train.csv). You will train on train.csv dataset and predict on test.csv file.

- Business Problem Framing

  Describe the business problem and how this problem can be related to the real world.

  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.

  This problem can help the real estate firms in determining the areas of utmost importance,the pricing dynamics of the market and adjust/formulate strategies accordingly in order to yield optimal results.

- Conceptual Background of the Domain Problem

  Describe the domain related concepts that you think will be useful for better understanding of the project.

  The domain knowledge about the features aswell as the techniques/bars used in rating/classifying the features such as – LotConfig: Lot configuration, Neighborhood(Physical locations within Ames city limits), OverallQual,ExterQual etc.

Domain knowledge about the features may help in handling the missing data and the feature selection in our project.

Review of Literature

This is a comprehensive summary of the research done on the topic. The review should enumerate, describe, summarize, evaluate and clarify the research done.

The initial analysis on the dataset revealed that the dataset had a significant number of values missing in a few of the features which were handled accordingly,however replacing the values with a predicted value or mean can never counterbalance the presence of actual values just like the logic,no matter how reasonable is not necessarily a fact.

The missing values of the categorical features were replaced by the category with the highest frequency(mode) whereas the continous(numerical) feature's missing values were replaced with the mean values of the features respectively.

The count of the missing values for the Numerical features looked like:

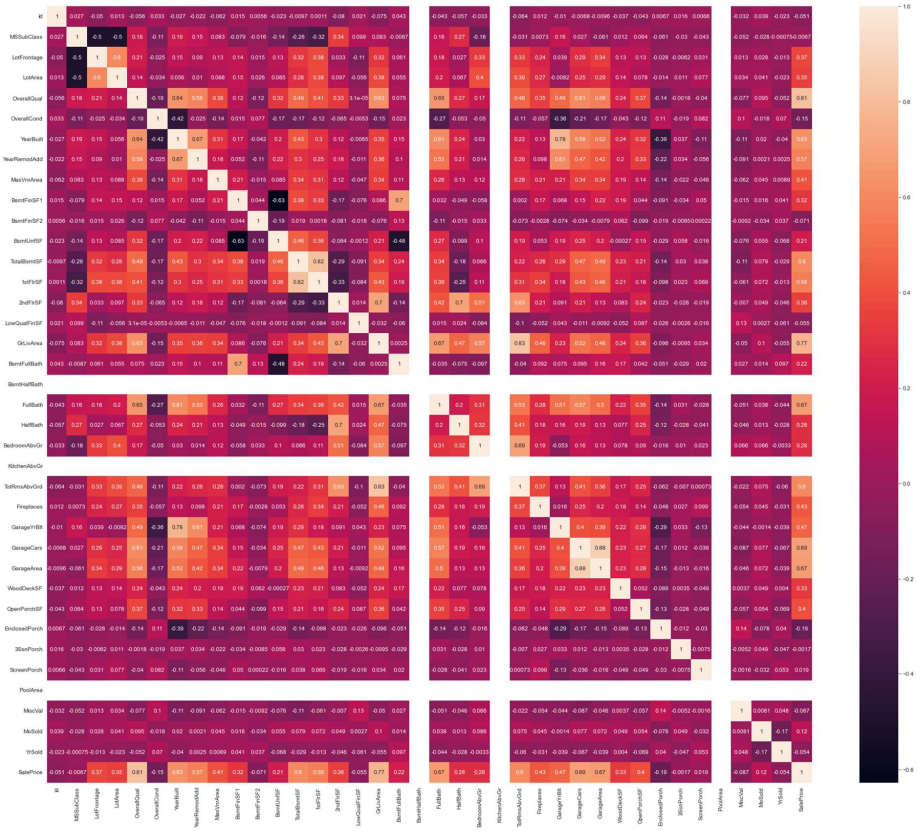
```
Id                0
MSSubClass        0
LotFrontage     214
LotArea           0
OverallQual       0
OverallCond       0
YearBuilt         0
YearRemodAdd      0
MasVnrArea        7
BsmtFinSF1        0
BsmtFinSF2        0
BsmtUnfSF         0
TotalBsmtSF       0
1stFlrSF          0
2ndFlrSF          0
LowQualFinSF      0
GrLivArea         0
BsmtFullBath      0
BsmtHalfBath      0
FullBath          0
HalfBath          0
BedroomAbvGr      0
KitchenAbvGr      0
TotRmsAbvGrd      0
Fireplaces        0
GarageYrBlt      64
GarageCars        0
GarageArea        0
WoodDeckSF        0
OpenPorchSF       0
EnclosedPorch     0
3SsnPorch         0
ScreenPorch       0
PoolArea          0
```

```
MiscVal            0
MoSold             0
YrSold             0
SalePrice          0
dtype: int64
```

The dataset was then checked for outliers and a significant number of outliers was detected in columns such as LotArea,MasVnrArea etc.The outliers had to be removed using the zscore class from sklearn.preprocessing which eliminates all the rows from the dataset which contain numbers that fall beyond 3-standard deviation from the mean.

The numerical and categorical feature names were separated in in lists:num_col and object_col respectively for a more organised and specific approach towards feature engineering.

The data was checked for multicollinearity using a heatmap and the visualization was:



OBSERVATION:

SOME OF THE FEATURES IN THE DATASET SHOW SIGNS OF MULTICOLLINEARITY(FOR EXAMPLE : TotalBsmtSF & 1stFlrSF,GarageCars & GarageArea,TotRmsAbvGrd & GrLivArea).

MERGING THE COLUMNS SHOWING MULTICOLLINEARITY WITH THE TARGET VARIABLES AND AMONGST THEMSELVES INTO A NEW COLUMN CONTAINING THE AVERAGE OFTHE SUM OF THE AFOREMENTIONED FEATURES

'TotalBsmtSF AND 1stFlrSF features were merged into TotalBsmtSF & 1stFlrSF and the former were dropped from the dataset.

GarageCars and GarageArea features were merged into GarageCars & GarageArea and the former were dropped from the dataset.

TotRmsAbvGrd and GrLivArea features were merged into TotRmsAbvGrd & GrLivArea and the former were dropped from the dataset.

THE DATASET WAS THEN CHECKED FOR SKEWNESS USING A HISTPLOT AND features: 'LotFrontage','TotalBsmtSF & 1stFlrSF', 'LotArea','TotRmsAbvGrd & GrLivArea', 'SalePrice **all had skewness present in them.**

The skewness was removed using log transformation technique.

The categorical features were analysed next for missing values and the counts of missing values were:

```
Alley           753
BsmtQual         15
BsmtCond         15
BsmtExposure     16
BsmtFinType1     15
BsmtFinType2     15
FireplaceQu     411
GarageType       37
GarageFinish     37
GarageQual       37
GarageCond       37
PoolQC          803
Fence           644
MiscFeature     781
```

MORE THAN 90% OF DATA was MISSING IN COLUMNS:Alley,PoolQC,MiscFeature.Thus, the features were removed.

The EDA and visualisation done on the dataset brought to light significant patterns and insights such as:

1)for numerical columns

KEY OBSERVATIONS: 1)SalePrice shows that most of the properties with a lot frontage of 40-80 have a price ranging in between 75,000 to 2,50,000. The properties in this sample mostly have a lot frontage ranging between 60 to 80. 2)SalePrice shows that most of the properties with a lot area of 5000-13,000(square feet)have a price ranging from 1,00,000 to 2,50,000.The properties in this sample mostly have a lot area concentrated in 5000-15,000 range. 3)The sale price is directly propotional to the OverallQuality.Most of the properties lie in the quality range of 5 to 8. 4)The sale price is highest for the houses in OverallCond-5(Average condition). 5)The properties built after the year-2000 have a direct positive relationship with the SalePrice. 6)The properties buit between 1940 and 1980 mostly have a SalePrice ranging from 1,00,00 to 1,70,000 (approx). 7)The biggest chunk of the properties were built/remodeled after the year-2000. 8)The SalePrice shows an increase with an increase in BsmtFinType1 area over 200(squarefeet) although a large chunk of properties with a BsmtFinType1 area of 0(square feet) too have prices ranging between 50,000-35,000(approx). 9)Most properties have a BsmtFinType2 area of 0(square feet) and have prices ranging between(50,000 and4,00,000). 10)Most properties have a Unfinished square feet of basement area concentrated between 0 and 1,000. 11)The SalePrice is directly propotional to the Total square feet of basement area over 500 (squarefeet). 12)Most properties have a TotalBsmtSF between 500-1500(square feet). 13)First

Floor square feet has a positive relationship with the SalePrice. 14)The SalePrice increases with an increase in Second floor square feet(2ndFlrSF) above 400(square feet). 15)The SalePrice ranges between 50,000 to 3,50,000 and even upto 4,00,000 in some cases. 16)The SalePrice shows a direct,positive relationship with the GrLivArea(Above grade (ground) living area square feet).Increase in one will correspond to an increase in the other. 17)The sale price is the highest for the FullBath(Full bathrooms above grade 2). 18)Most properties have TotRmsAbvGrd(Total rooms above grade (does not include bathrooms) are mostly between 4 and 9. 19)Properties having GarageYrBlt(Year garage was built) after 2000 have the highest range of salePrice(50,000-4,00,000). 20)Properties built between the year 1940-1980 have prices mostly between 75,000 and 2,00,000. 21)The SalePrice is directly propotion to GarageCars(Size of garage in car capacity). 22)The SalePrice has a direct,positive relation with GarageArea(Size of garage in square feet) above 200(sq.feet).

2)for categorical features with respect to the SalePrice feature

Commented [pr1]:

mszoning: identifies the general zoning classification of the sale. observations: 1)rl-residential low density zones had the highest saleprice amongst the 5 zones classified 2)c-commercial,rm-residential medium density zones depicted the lowest saleprices amongst the 5 zones classified.

street: type of road access to property observation: almost all the properties had access to paved roads. paved(pave) streets had a significantly higher saleprice than gravel(grvl) streets.

LotShape: General shape of property OBSERVATIONS: 1)the properties were mostly Reg-Regular and IR1-Slightly irregular types. 2)the IR1 shaped properties had a higher SalePrice than the others suceeded by Reg shaped properties.

LandContour: Flatness of the property 1)The Lvl type of properties had a significantly higher saleprice than the others. 2)the Bnk type of properties had the lowest saleprice.

LotConfig: Lot configuration 1)The 'Inside' type of (Lot configuration)properties had the highest saleprice. 2)The properties having 'FR2' type of Lot configuration had the lowest saleprice(mean).

LandSlope: Slope of property 1)most properties had a gentle slope 2)properties having a gentle slope had a higher saleprice(mean) than the others.

Neighbourhood 1)NoRidgE-Northridge,StoneBr-Stone Brook AND NridgHt-Northridge Heights had the highest saleprices amongst all the neighborhoods. 2)IDOTRR-Iowa DOT and Rail Road had the lowest saleprice among all.

Condition1: Proximity to various conditions OBSERVATIONS: 1)The properties were mostly Norm-Normal category. 2)The properties under Norm-Normal category had the highest sale price.

Condition2: Proximity to various conditions (if more than one is present) OBSERVATION: Almost 90% of properties were under Norm category and had a the highest saleprice.

BldgType: Type of dwelling OBSERVATION: The 1Fam(Single-family Detached) type of dwelling in properties had the maximum aswell as the minimum

saleprice(highest range). Also they formed the majority of the Type of dwelling in the properties.

HouseStyle: Style of dwelling OBSERVATION: 1)The bulk of the properties had 1Story and 2Story type of dwelling 2)The 2Story style of dwelling had a slightly higher mean of saleprice than the others.

RoofStyle: Type of roof OBSERVATION: 1)The properties were mostly Gable and HIP types. 2)The Hip type of properties had a higher mean sale price than the others.

RoofMatl: Roof material OBSERVATION: Almost all the properties had CompShg Standard (Composite) Shingle type of roof material

Exterior1st: Exterior covering on house OBSERVATION: The VinylSd Vinyl Siding had the highest mean sale price among all the exterior categories followed by MetalSd Metal Siding. 2)The AsbShng Asbestos Shingles had the lowest sale price(mean) among all.

Exterior2nd: Exterior covering on house (if more than one material) OBSERVATIONS: 1)The VinylSd Vinyl Siding had the highest mean sale price among all the exterior2 categories followed by MetalSd Metal Siding. 2)The AsbShng Asbestos Shingles had the lowest sale price(mean) among all.

MasVnrType: Masonry veneer type OBSERVATION: 1)The Stone type of Masonry had the highest SalePrice(mean) 2)The BrkCmn had the lowest mean saleprice.

ExterQual: Evaluates the quality of the material on the exterior OBSERVATION: 1)The Ex-Excellent type of material quality had the highest saleprice(mean) followed by the Gd-good type. 2)Ta had the lowest saleprice and Fa had the lowest mean sale price.

ExterCond: Evaluates the present condition of the material on the exterior OBSERVATION: The majority of the properties had the TA type of ExterCond.The TA also had the highest mean SalePrice.

Foundation: Type of foundation OBSERVATION: 1)The properties with PConc Poured Contrete type of foundation had the highest saleprice(mean). 2)The BrkTil type of foundation had the lowest saleprice.

BsmtQual: Evaluates the height of the basement OBSERVATION: The properties with Ex-(Excellent (100+ inches)) height had the highest sale price(mean).

BsmtCond: Evaluates the general condition of the basement OBSERVATION: TA type of condition had the highest saleprice whereas Gd category had the highest mean saleprice. The majority of the cases were of TA category.

BsmtExposure: Refers to walkout or garden level walls OBSERVATION: The Gd type of properties had the highest SalePrice(mean).

BsmtFinType1: Rating of basement finished area OBSERVATION: The GLQ(Good Living Quarters) had the highest saleprice(mean aswell as maximum).

BsmtFinType2: Rating of basement finished area (if multiple types) OBSERVATION: Majority of properties had Unf Unfinshed type of BsmtFinType2 which also had the highest saleprice.

Heating: Type of heating OBSERVATION: 1)Most properties had a GasA(Gas forced warm air furnace) type of heating which also had the highest SalePrice.

CentralAir: Central air conditioning OBSERVATION: 1)Most properties had central air conditioning. 2)properties with central air conditioning had the highest saleprice.

Electrical: Electrical system OBSERVATION: 1)Majority of the properties had SBrkr Standard Circuit Breakers & Romex type of electrical system. 2)SBrkr Standard Circuit Breakers & Romex type of electrical system had the highest sale price.(mean aswell as maximum)

KitchenQual: Kitchen quality OBSERVATION: 1)The Ex type of kitchen quality had the highest saleprice(mean aswell as maximum). 2)The Fa type of kitchen quality had the lowest SalePrice(minimum).

Functional: Home functionality (Assume typical unless deductions are warranted) OBSERVATIONS: 1)Most properties had Typ Typical Functionality. 2)Typ Typical Functionality had the highest sale price among all.

FireplaceQu: Fireplace quality OBSERVATION: 1)The properties having Ex type of quality had the highest mean saleprice. 2)The properties having Po type of quality had the lowest saleprice(mean).

GarageType: Garage location OBSERVATION: 1)BuiltIn Built-In (Garage part of house - typically has room above garage) had the highest SalePrice(mean) among all the locations. 2)The maximum saleprice was recorded in Attchd Attached to home location.

GarageFinish: Interior finish of the garage OBSERVATION: 1)The properties with Un(Unfinished) GarageFinish had the lowest SalePrice. 2)RFn(Rough Finished) and Fin(Finished) had similar means of SalePrice but Fin category had a slightly higher mean.

MAJORITY OF THE PROPERTIES HAD TA(Typical/Average) CONDITION.

PavedDrive: Paved driveway OBSERVATION: The properties with Y(Paved) driveways had the highest SalePrice(mean aswell as maximum) and also constituted bulk of the sample.

SaleType: Type of sale OBSERVATION: 1)The SaleType New(Home just constructed and sold) had the highest saleprice(mean aswell as maximum). 2)Majority of sales were of WD,New,COD type.(descending order).

SaleCondition: Condition of sale OBSERVATION: The SaleCondition Partial(Home was not completed when last assessed (associated with New Homes)) had the highest SalePrice(mean as well as maximum).

END OF EDA

The model was saved using pickle.dump method from the pickle library as : pfa_housing_price_predictor.sav.

The model was reloaded using pickle.load method and was used to predict the test data.

The features that influcenced the target variable the most were:

```
'MSSubClass', 'FV', 'RL', 'LotFrontage', 'LotArea', 'HLS', 'CulDSac',
       'Mod', 'BrkSide', 'CollgCr', 'Crawfor', 'Edwards', 'IDOTRR', 'NA
mes',
       'NridgHt', 'Somerst', 'Feedr', 'TwnhsE', 'OverallQual', 'Overall
Cond',
```

```
       'YearBuilt', 'YearRemodAdd', 'Gambrel', 'CBlock', 'HdBoard', 'Me
talSd',
       'Plywood', 'VinylSd', 'Fa', 'PConc', 'Slab', 'No', 'BsmtFinSF1',
       'BsmtFinSF2', 'BsmtUnfSF', 'Y', 'SBrkr', '2ndFlrSF', 'LowQualFin
SF',
       'BsmtFullBath', 'FullBath', 'BedroomAbvGr', 'Typ', 'Fireplaces',
       'GarageYrBlt', 'RFn', 'WoodDeckSF', 'OpenPorchSF', 'EnclosedPorc
h',
       '3SsnPorch', 'ScreenPorch', 'MiscVal', 'MoSold', 'YrSold', 'Allo
ca',
       'Normal', 'Partial', 'TotalBsmtSF & 1stFlrSF',
       'GarageCars & GarageArea', 'TotRmsAbvGrd & GrLivArea'],
```

THE ABOVE ARE THE FEATURES THAT CONTRIBUTED THE MOST IN PREDICTING THE SALEPRICE OF THE TEST DATA.

- Motivation for the Problem Undertaken

  Describe your objective behind to make this project, this domain and what is the motivation behind.

DataScience in itself is a field so diverse and interesting that it can be motivating enough for anyone with a curious mind to explore and play with.

This project was given as an assignment during my internship in Flip Robo Technologies with two major objectives:

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

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

Both of these objectives have been completed under the EDA and the feature selection parts respectively.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  Describe the mathematical, statistical and analytics modelling done during this project along with the proper justification.

- Data Sources and their formats

  What are the data sources, their origins, their formats and other details that you find necessary? They can be described here. Provide a proper data description. You can also add a snapshot of the data.

- Data Preprocessing Done

  What were the steps followed for the cleaning of the data? What were the assumptions done and what were the next actions steps over that?

- Data Inputs- Logic- Output Relationships

  Describe the relationship behind the data input, its format, the logic in between and the output. Describe how the input affects the output.

- State the set of assumptions (if any) related to the problem under consideration

  Here, you can describe any presumptions taken by you.

- Hardware and Software Requirements and Tools Used

Listing down the hardware and software requirements along with the tools, libraries and packages used. Describe all the software tools used along with a detailed description of tasks done with those tools.

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  Describe the approaches you followed, both statistical and analytical, for solving of this problem.

- Testing of Identified Approaches (Algorithms)

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

LinearRegression,Lasso,Ridge,DecisionTreeRegressor,RandomForestRegressor,AdaBoostRegressor and GradientBoostingRegressor were used.

- Run and Evaluate selected models

  Describe all the algorithms used along with the snapshot of their code and what were the results observed over different evaluation metrics.

LINEAR REGRESSION

```
score: 0.9488184368722168
M.S.E 0.016254007096186053
MEAN ABSOLUTE ERROR 0.08353082192055095
R2 score: 0.8728042034066253
R.M.S.E: 0.12749120399535827
 Cross validation: mean score: 0.8889058090007358
cross val score: [0.88785884 0.84830088 0.92879418 0.84723265 0.8634732
5 0.84701418
 0.93403808 0.92216169 0.91375455 0.89642978]
```

Lasso

```
score: 0.9200571834654206
M.S.E 0.010617742760808641
```

MEAN ABSOLUTE ERROR 0.07583264146281114
R2 score: 0.9169108121774174
R.M.S.E: 0.10304243184634494
Cross validation: mean score: 0.9045884133057488
cross val score: [0.91245096 0.94149721 0.91150908 0.89023988 0.9273976
5 0.87310974
 0.90859173 0.94511156 0.87486529 0.86111104]

RIDGE

mean score: 0.8889983320614985
cross val score: [0.88791744 0.84884612 0.92878377 0.84732872 0.8636880
7 0.8470689
 0.93406721 0.92216363 0.9137471 0.89637237]

DecisionTreeRegressor

score: 1.0
M.S.E 0.02883628318906452
MEAN ABSOLUTE ERROR 0.12684908115003876
R2 score: 0.7743415522510846
R.M.S.E: 0.16981249420777178
Cross validation: mean score: 0.7207229090345207
cross val score: [0.64935529 0.76009331 0.77676689 0.77106744 0.6907433
1 0.65301005
 0.78193834 0.69094592 0.69270011 0.74060842]

RANDOM FOREST REGRESSOR

score: 0.9808914619953911
M.S.E 0.01691526852245359
MEAN ABSOLUTE ERROR 0.09381069815195504
R2 score: 0.8676294995091288
R.M.S.E: 0.13005871182836462
Cross validation: mean score: 0.8665266419037223
cross val score: [0.8762561 0.89072434 0.89009571 0.87598657 0.9189037
8 0.82798421
 0.85760934 0.88256574 0.85853652 0.7866041]

ADABOOST REGRESSOR

score: 0.8927793312111704
M.S.E 0.025930036260681424
MEAN ABSOLUTE ERROR 0.12100649718646102
R2 score: 0.7970843990435827
R.M.S.E: 0.16102806047605933
Cross validation: mean score: 0.8185870323938571
cross val score: [0.81914835 0.84254528 0.81413889 0.78386369 0.8947737
1 0.80689551
 0.80277597 0.85845146 0.79896241 0.76431506]

GRADIENT BOOST REGRESSOR

score: 0.9724874115663982
M.S.E 0.013133193773183352
MEAN ABSOLUTE ERROR 0.08394238999771032
R2 score: 0.897226140365893
R.M.S.E: 0.11460014735236318
Cross validation: mean score: 0.8895077997205851

```
cross val score: [0.89879235 0.93581665 0.91162561 0.87288682 0.9104984
9 0.8392594
 0.88258582 0.91519058 0.87657636 0.85184593]
```

- Key Metrics for success in solving problem under consideration

  What were the key metrics used along with justification for using it? You may also include statistical metrics used if any.

The key metrics used were mean_squared_error,mean_absolute_error,root_mean_squared_error,r2_score,cross_val_score.

- Visualizations

  Mention all the plots made along with their pictures and what were the inferences and observations obtained from those. Describe them in detail.

  If different platforms were used, mention that as well.

  for visualizing categorical data,catplot was used:

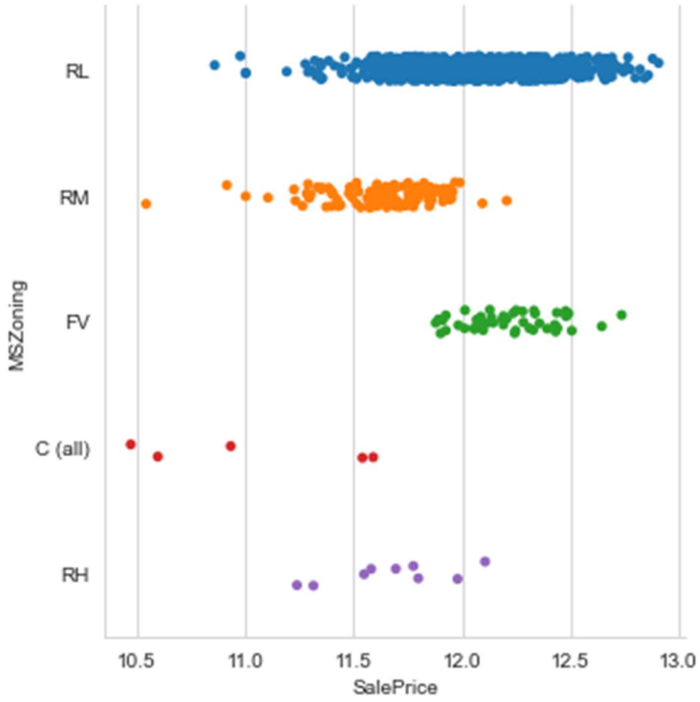
PROCEEDING TO VISUALISING THE CATEGORICAL DATA WITH RESPECT TO THE TARGET FEATURE AND DRAWING PATTERNS

In [2478]:

```
plt.figure(figsize=(7,7))
sns.set_style('whitegrid')
sns.catplot(y='MSZoning',x='SalePrice',data=df)
```

Out[2478]:

```
<seaborn.axisgrid.FacetGrid at 0x1e0f4c4b0d0>
<Figure size 504x504 with 0 Axes>
```



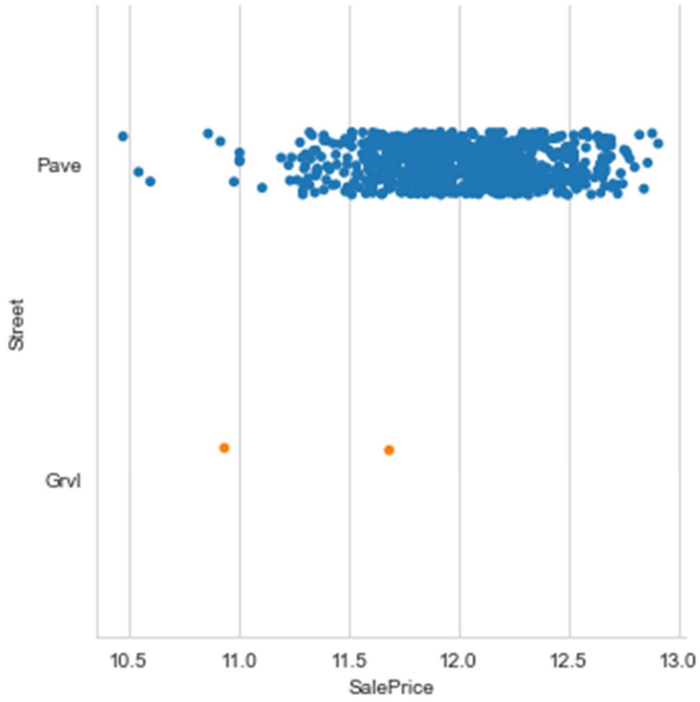
MSZoning: Identifies the general zoning classification of the sale. OBSERVATIONS: 1)RL-Residential Low Density zones had the highest SalePrice amongst the 5 zones classified 2)C-Commercial,RM-Residential Medium Density zones depicted the lowest SalePrices amongst the 5 zones classified.

In [2479]:

```
#plt.figure(figsize=(5,5))
sns.set_style('whitegrid')
sns.catplot(y='Street',x='SalePrice',data=df)
```

Out[2479]:

```
<seaborn.axisgrid.FacetGrid at 0x1e0f5d7c6d0>
```



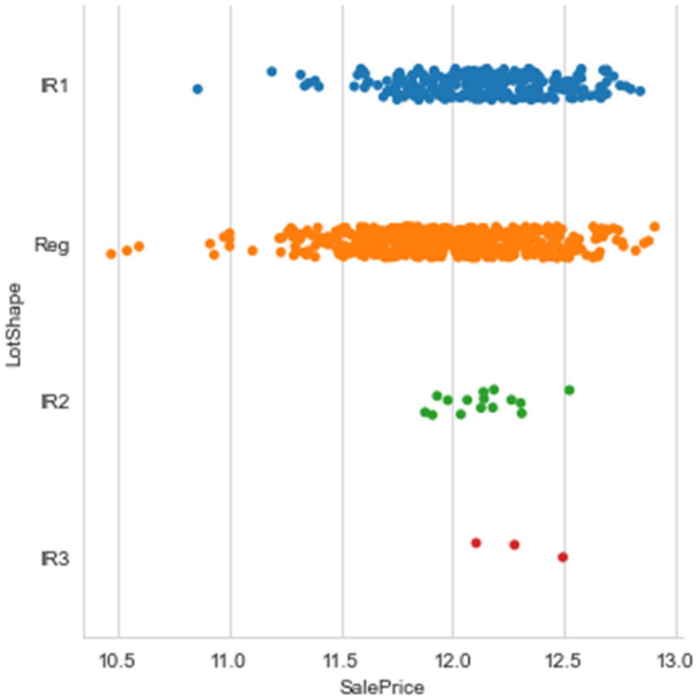
Street: Type of road access to property OBSERVATION: ALMOST ALL THE PROPERTIES HAD ACCESS TO PAVED ROADS. PAVED(PAVE) STREETS HAD A SIGNIFICANTLY HIGHER SALEPRICE THAN GRAVEL(GRVL) STREETS.

In [2480]:

```
plt.figure(figsize=(7,7))
sns.set_style('whitegrid')
sns.catplot(y='LotShape',x='SalePrice',data=df)
```

Out[2480]:

```
<seaborn.axisgrid.FacetGrid at 0x1e0f5e683d0>
<Figure size 504x504 with 0 Axes>
```



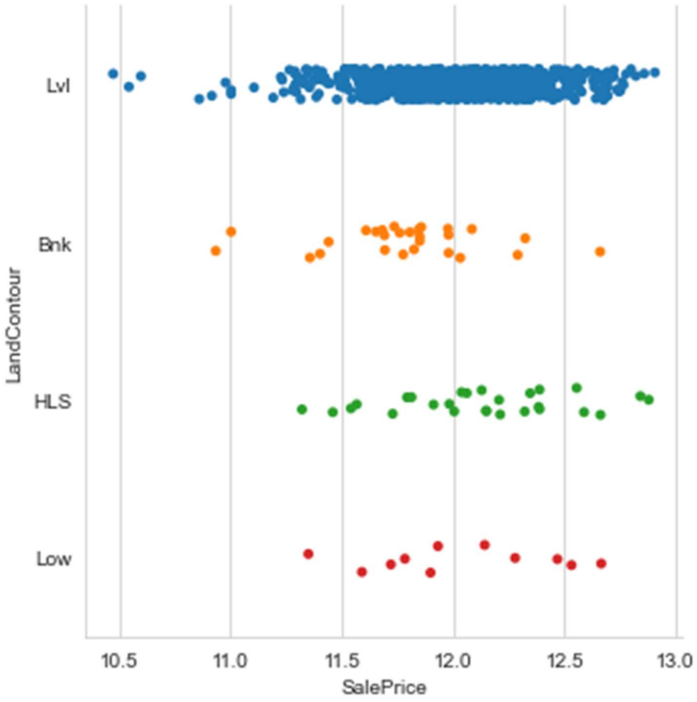
LotShape: General shape of property OBSERVATIONS: 1)the properties were mostly Reg-Regular and IR1-Slightly irregular types. 2)the IR1 shaped properties had a higher SalePrice than the others suceeded by Reg shaped properties.

In [2481]:

```
plt.figure(figsize=(7,7))
sns.set_style('whitegrid')
sns.catplot(y='LandContour',x='SalePrice',data=df)
```

Out[2481]:

```
<seaborn.axisgrid.FacetGrid at 0x1e0f5e82340>
<Figure size 504x504 with 0 Axes>
```



DESCRIPTION: 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
```

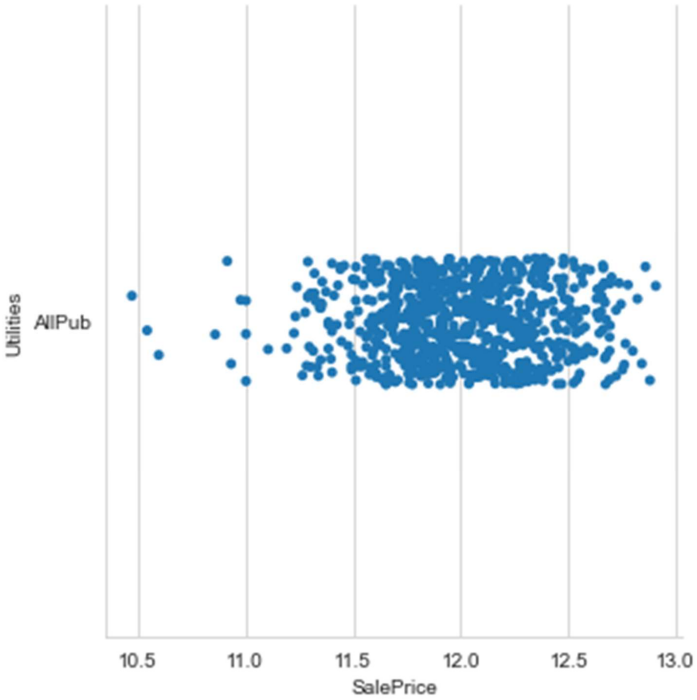
OBSERVATIONs: 1)The Lvl type of properties had a significantly higher saleprice than the others. 2)the Bnk type of properties had the lowest saleprice.

In [2482]:

```
plt.figure(figsize=(7,7))
sns.set_style('whitegrid')
sns.catplot(y='Utilities',x='SalePrice',data=df)
```

Out[2482]:

```
<seaborn.axisgrid.FacetGrid at 0x1e0e3c33d00>
<Figure size 504x504 with 0 Axes>
```



SINCE 'Utilities' feature has only 1 unique value for all the rows,removing it:

In [2483]:

```
df.drop('Utilities',axis=1,inplace=True)
obj_col.pop(4)
```

Out[2483]:

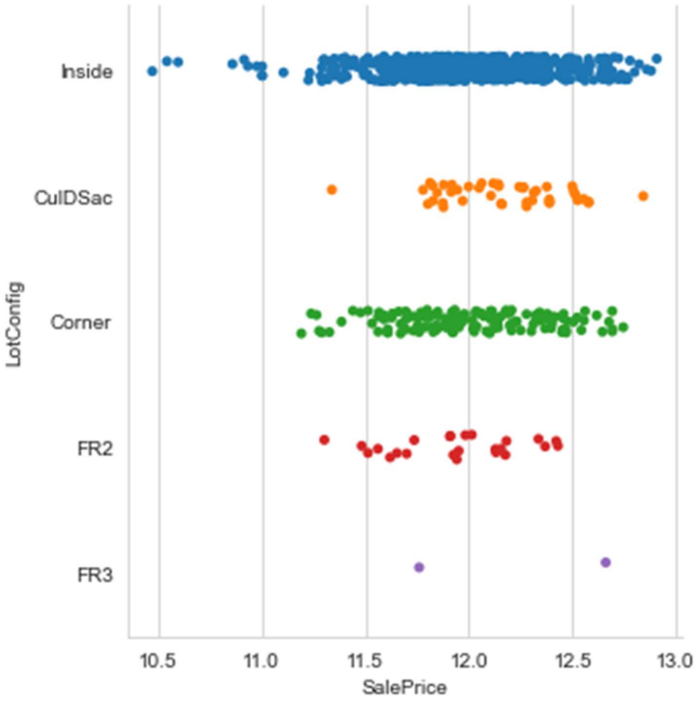
```
'Utilities'
```

In [2484]:

```
plt.figure(figsize=(7,7))
sns.set_style('whitegrid')
sns.catplot(y='LotConfig',x='SalePrice',data=df)
```

Out[2484]:

```
<seaborn.axisgrid.FacetGrid at 0x1e0f5f1ac10>
<Figure size 504x504 with 0 Axes>
```



DESCIPTION: 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
```

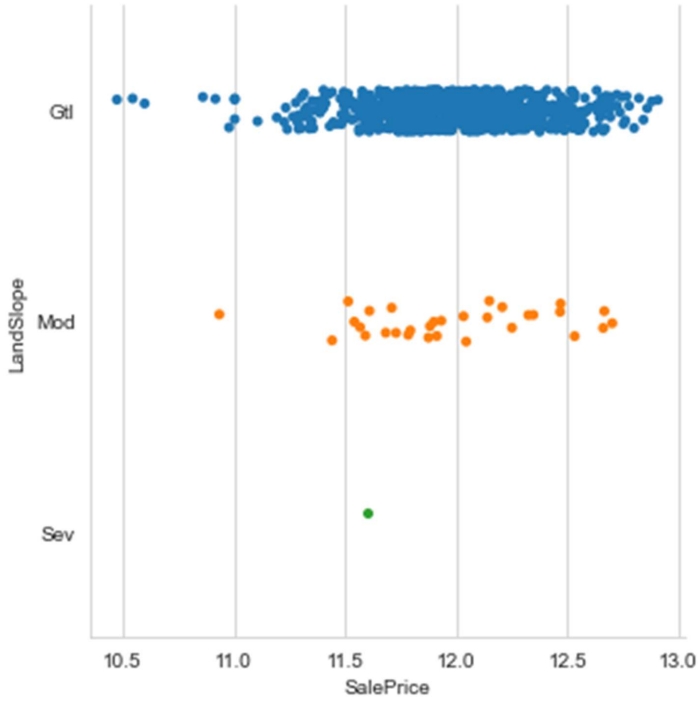
OBSERVATIONS: 1)The 'Inside' type of (Lot configuration)properties had the highest saleprice.
2)The properties having 'FR2' type of Lot configuration had the lowest saleprice(mean).

In [2485]:

```
plt.figure(figsize=(7,7))
sns.set_style('whitegrid')
sns.catplot(y='LandSlope',x='SalePrice',data=df)
```

Out[2485]:

```
<seaborn.axisgrid.FacetGrid at 0x1e0f5e19520>
<Figure size 504x504 with 0 Axes>
```



DESCRIPTION: LandSlope: Slope of property

```
Gtl    Gentle slope

Mod    Moderate Slope

Sev    Severe Slope
```

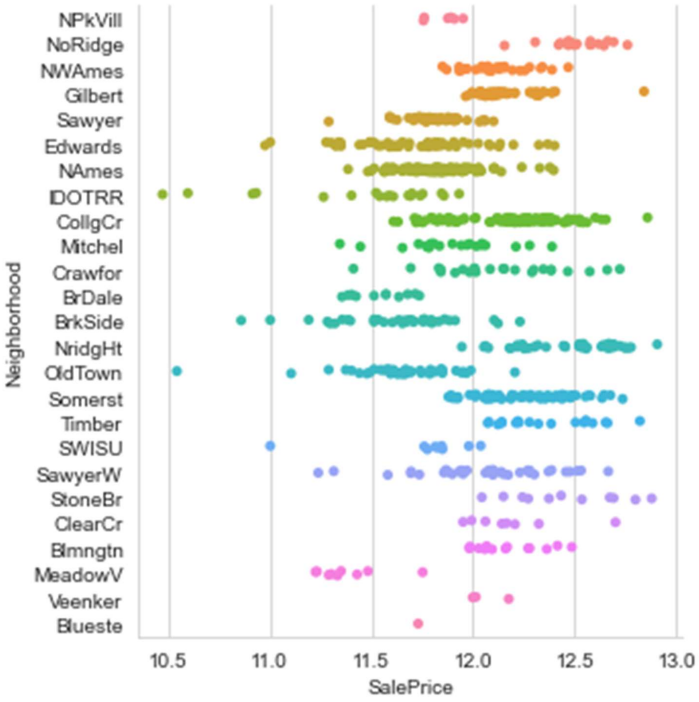
OBSERVATION: 1)most properties had a gentle slope 2)properties having a gentle slope had a higher saleprice(mean) than the others.

In [2486]:

```
plt.figure(figsize=(15,15))
sns.set_style('whitegrid')
sns.catplot(y='Neighborhood',x='SalePrice',data=df)
```

Out[2486]:

```
<seaborn.axisgrid.FacetGrid at 0x1e0f5f11430>
<Figure size 1080x1080 with 0 Axes>
```



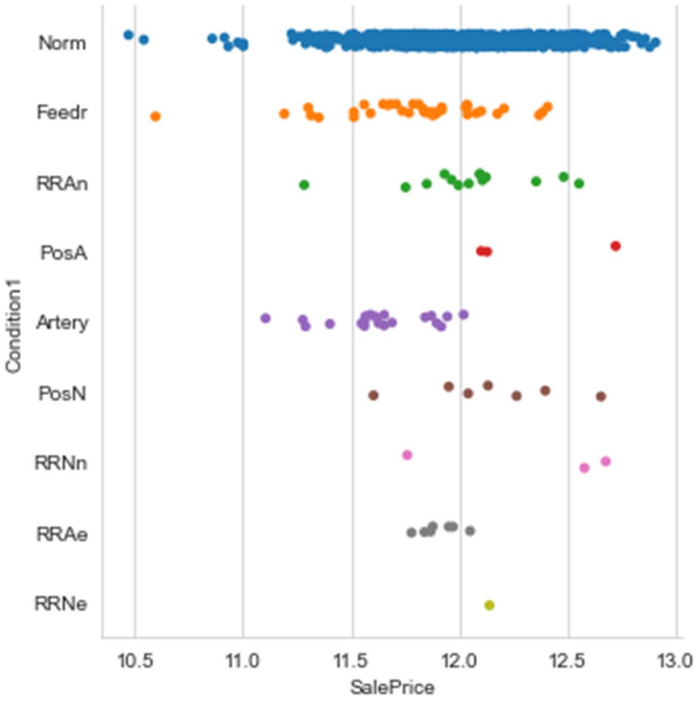
OBSERVATION: 1)NoRidgE-Northridge,StoneBr-Stone Brook AND NridgHt-Northridge Heights had the highest saleprices amongst all the neighborhoods. 2)IDOTRR-Iowa DOT and Rail Road had the lowest saleprice among all.

In [2487]:

```
plt.figure(figsize=(7,7))
sns.set_style('whitegrid')
sns.catplot(y='Condition1',x='SalePrice',data=df)
```

Out[2487]:

```
<seaborn.axisgrid.FacetGrid at 0x1e0f6f9efa0>
<Figure size 504x504 with 0 Axes>
```



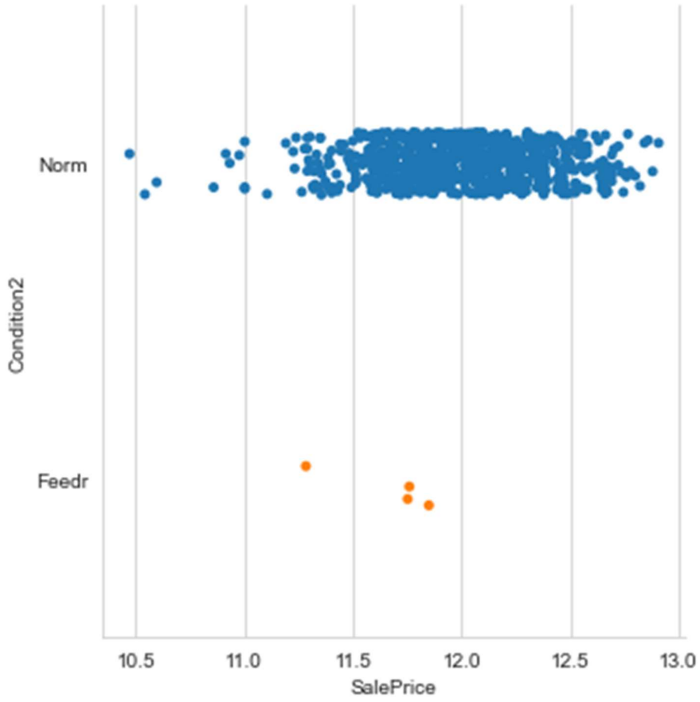
Condition1: Proximity to various conditions OBSERVATIONS: 1)The properties were mostly Norm-Normal category. 2)The properties under Norm-Normal category had the highest sale price.

In [2488]:

```
plt.figure(figsize=(7,7))
sns.set_style('whitegrid')
sns.catplot(y='Condition2',x='SalePrice',data=df)
```

Out[2488]:

```
<seaborn.axisgrid.FacetGrid at 0x1e0f7039a30>
<Figure size 504x504 with 0 Axes>
```



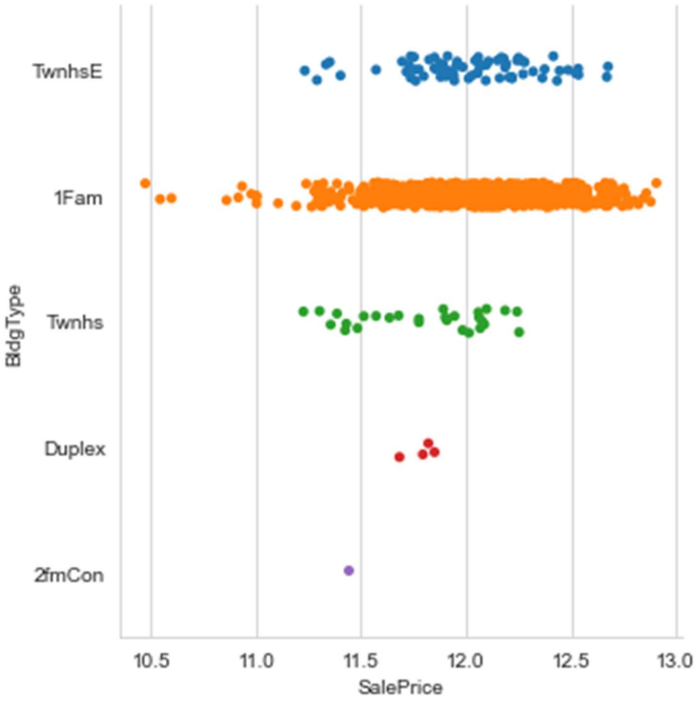
Condition2: Proximity to various conditions (if more than one is present) OBSERVATION: Almost 90% of properties were under Norm category and had a the highest saleprice.

In [2489]:

```
plt.figure(figsize=(7,7))
sns.set_style('whitegrid')
sns.catplot(y='BldgType',x='SalePrice',data=df)
```

Out[2489]:

```
<seaborn.axisgrid.FacetGrid at 0x1e0f5f1ac40>
<Figure size 504x504 with 0 Axes>
```



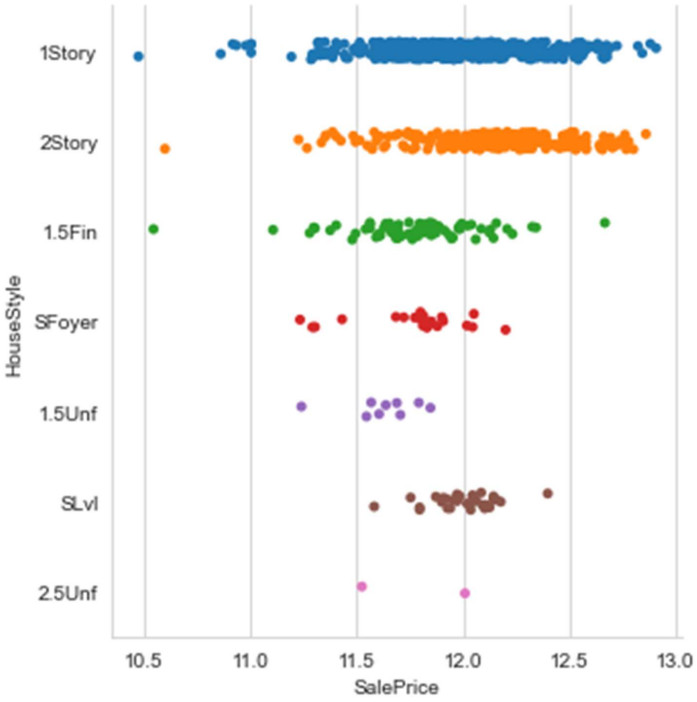
BldgType: Type of dwelling OBSERVATION: The 1Fam(Single-family Detached) type of dwelling in properties had the maximum aswell as the minimum saleprice(highest range). Also they formed the majority of the Type of dwelling in the properties.

In [2490]:

```
plt.figure(figsize=(7,7))
sns.set_style('whitegrid')
sns.catplot(y='HouseStyle',x='SalePrice',data=df)
```

Out[2490]:

```
<seaborn.axisgrid.FacetGrid at 0x1e0f6f1ed60>
<Figure size 504x504 with 0 Axes>
```

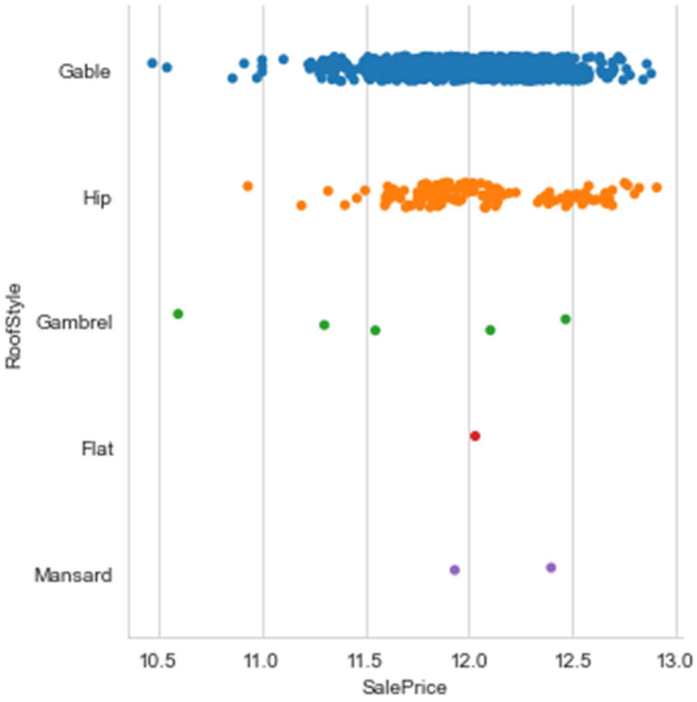


DESCRIPTION: HouseStyle: Style of dwelling OBSERVATION: 1)The bulk of the properties had 1Story and 2Story type of dwelling 2)The 2Story style of dwelling had a slightly higher mean of saleprice than the others.

In [2491]:

```
plt.figure(figsize=(25,7))
sns.set_style('whitegrid')
sns.catplot(y='RoofStyle',x='SalePrice',data=df)
plt.show()
```

```
<Figure size 1800x504 with 0 Axes>
```

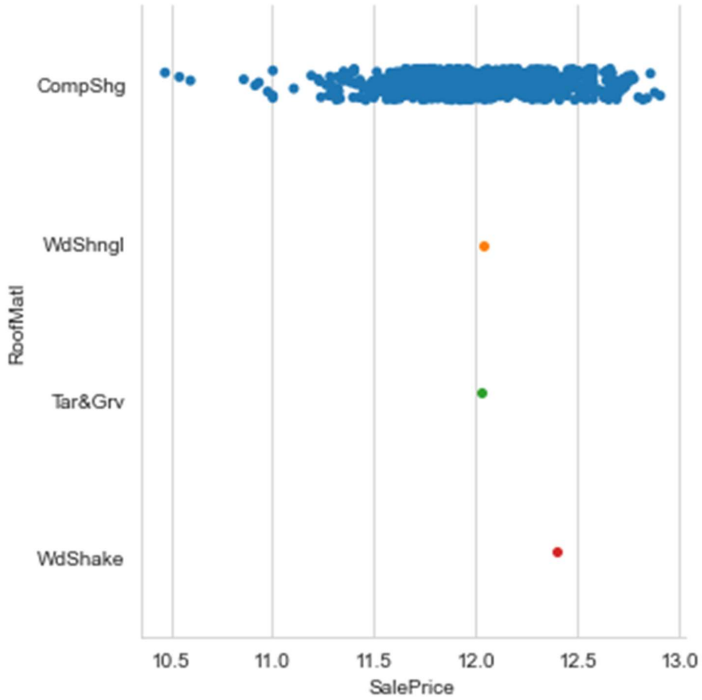


DESCRIPTION: RoofStyle: Type of roof OBSERVATION: 1)The properties were mostly Gable and HIP types. 2)The Hip type of properties had a higher mean sale price than the others.

In [2492]:

```
plt.figure(figsize=(25,7))
sns.set_style('whitegrid')
sns.catplot(y='RoofMatl',x='SalePrice',data=df)
plt.show()
```

```
<Figure size 1800x504 with 0 Axes>
```

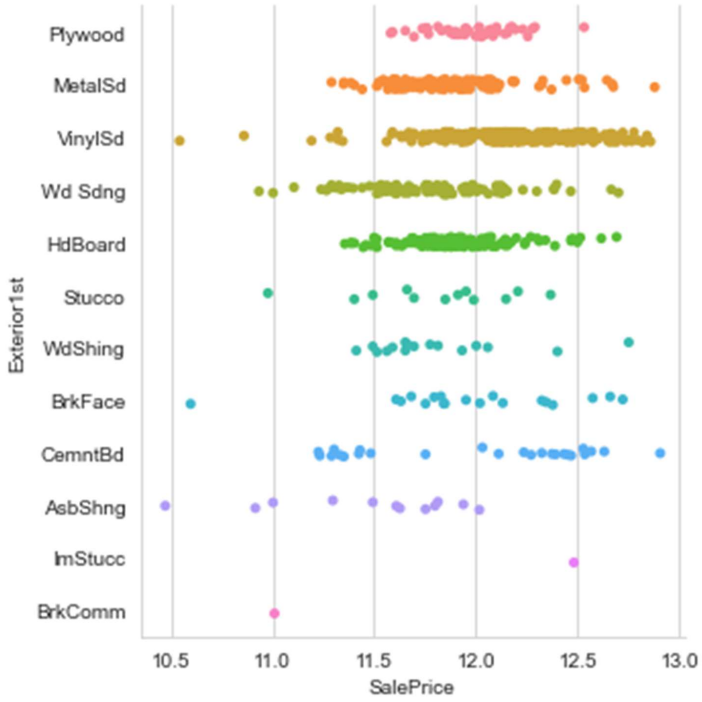


DESCRIPTION: RoofMatl: Roof material OBSERVATION: Almost all the properties had CompShg Standard (Composite) Shingle type of roof material

In [2493]:

```
plt.figure(figsize=(25,7))
sns.set_style('whitegrid')
sns.catplot(y='Exterior1st',x='SalePrice',data=df)
plt.show()
```

```
<Figure size 1800x504 with 0 Axes>
```

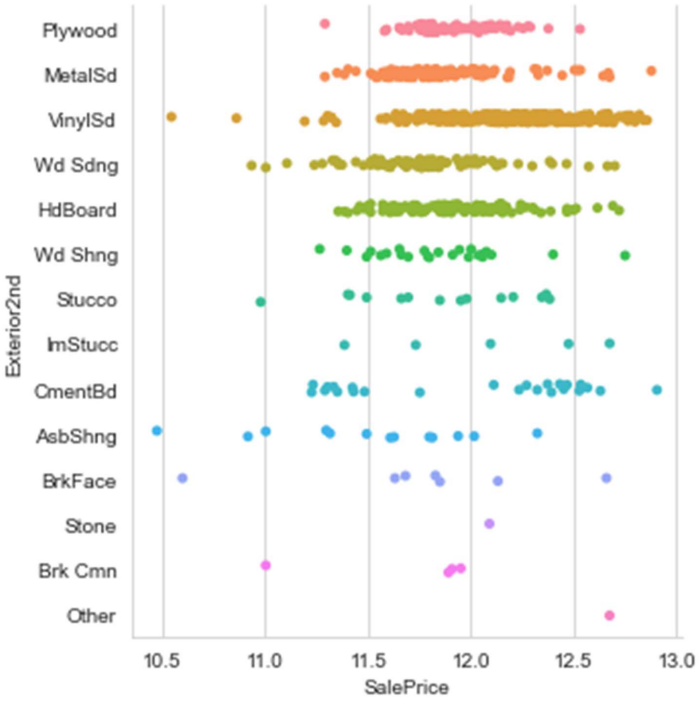


DESCRIPTION: Exterior1st: Exterior covering on house OBSERVATION: The VinylSd Vinyl Siding had the highest mean sale price among all the exterior categories followed by MetalSd Metal Siding. 2)The AsbShng Asbestos Shingles had the lowest sale price(mean) among all.

In [2494]:

```
plt.figure(figsize=(25,7))
sns.set_style('whitegrid')
sns.catplot(y='Exterior2nd',x='SalePrice',data=df)
plt.show()
```

```
<Figure size 1800x504 with 0 Axes>
```

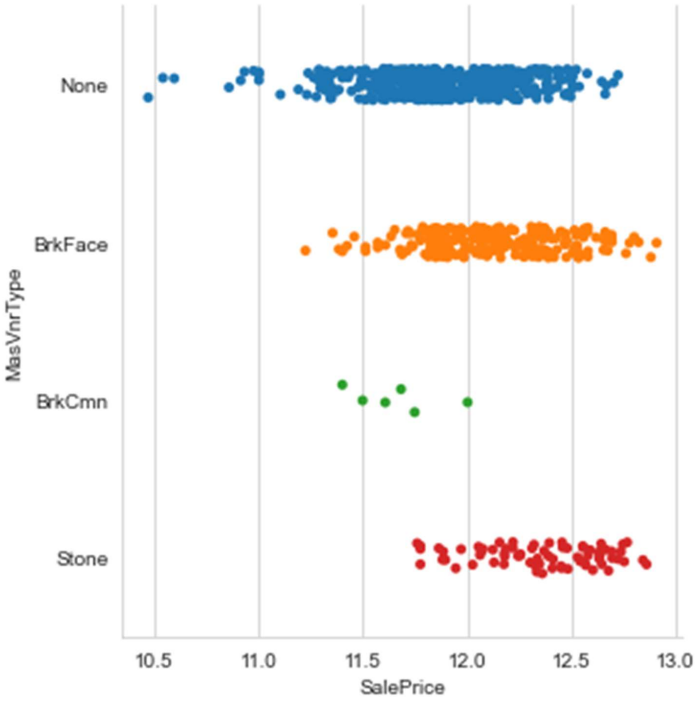


DESCRIPTION: Exterior2nd: Exterior covering on house (if more than one material)
OBSERVATIONS: 1)The VinylSd Vinyl Siding had the highest mean sale price among all the exterior2 categories followed by MetalSd Metal Siding. 2)The AsbShng Asbestos Shingles had the lowest sale price(mean) among all.

In [2495]:

```
plt.figure(figsize=(25,7))
sns.set_style('whitegrid')
sns.catplot(y='MasVnrType',x='SalePrice',data=df)
plt.show()
```

```
<Figure size 1800x504 with 0 Axes>
```

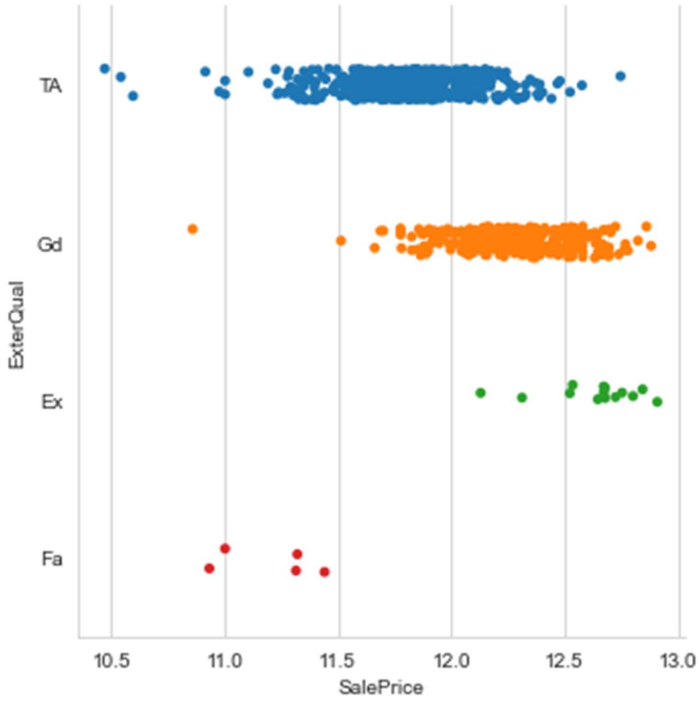


DESCRIPTION: MasVnrType: Masonry veneer type OBSERVATION: 1)The Stone type of Masonry had the highest SalePrice(mean) 2)The BrkCmn had the lowest mean saleprice.

In [2496]:

```
plt.figure(figsize=(25,7))
sns.set_style('whitegrid')
sns.catplot(y='ExterQual',x='SalePrice',data=df)
plt.show()
```

```
<Figure size 1800x504 with 0 Axes>
```

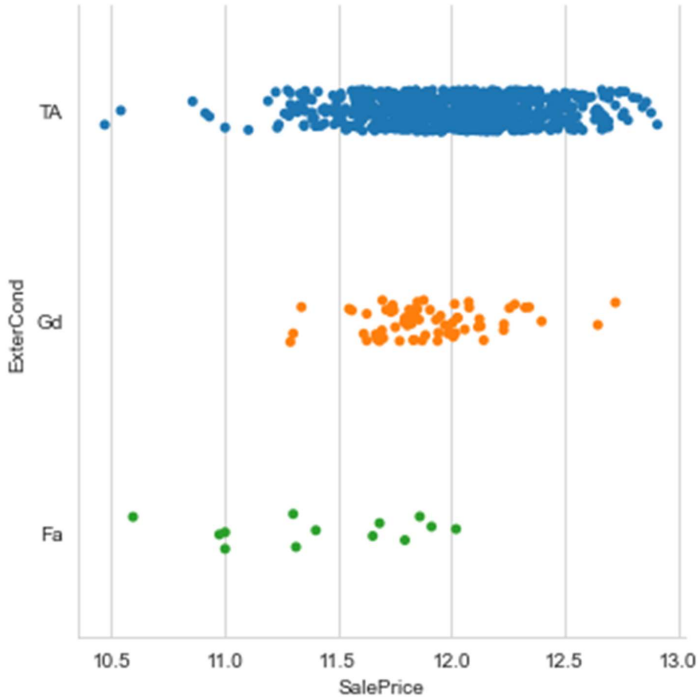


DESCRIPTION: ExterQual: Evaluates the quality of the material on the exterior OBSERVATION: 1)The Ex-Excellent type of material quality had the highest saleprice(mean) followed by the Gd-good type. 2)Ta had the lowest saleprice and Fa had the lowest mean sale price.

In [2497]:

```
plt.figure(figsize=(25,7))
sns.set_style('whitegrid')
sns.catplot(y='ExterCond',x='SalePrice',data=df)
plt.show()
```

```
<Figure size 1800x504 with 0 Axes>
```

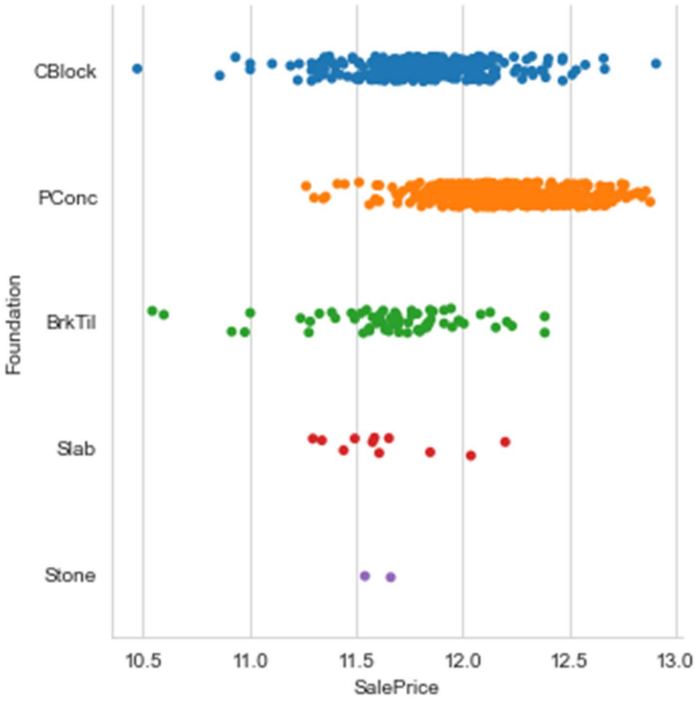


DESCRIPTION: ExterCond: Evaluates the present condition of the material on the exterior
OBSERVATION: The majority of the properties had the TA type of ExterCond.The TA also had the highest mean SalePrice.

In [2498]:

```
plt.figure(figsize=(25,7))
sns.set_style('whitegrid')
sns.catplot(y='Foundation',x='SalePrice',data=df)
plt.show()
```

```
<Figure size 1800x504 with 0 Axes>
```

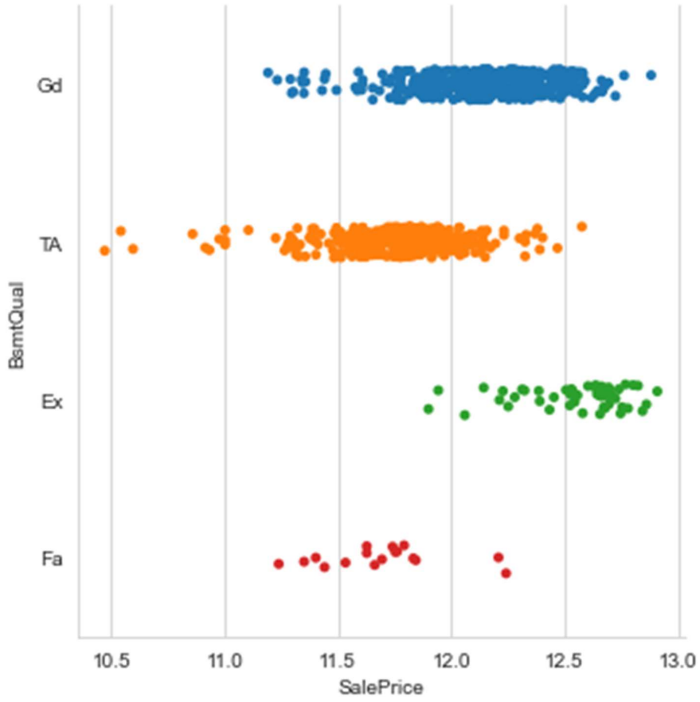


DESCRIPTION: Foundation: Type of foundation OBSERVATION: 1)The properties with PConc Poured Contrete type of foundation had the highest saleprice(mean). 2)The BrkTil type of foundation had the lowest saleprice.

In [2499]:

```
plt.figure(figsize=(25,7))
sns.set_style('whitegrid')
sns.catplot(y='BsmtQual',x='SalePrice',data=df)
plt.show()
```

```
<Figure size 1800x504 with 0 Axes>
```

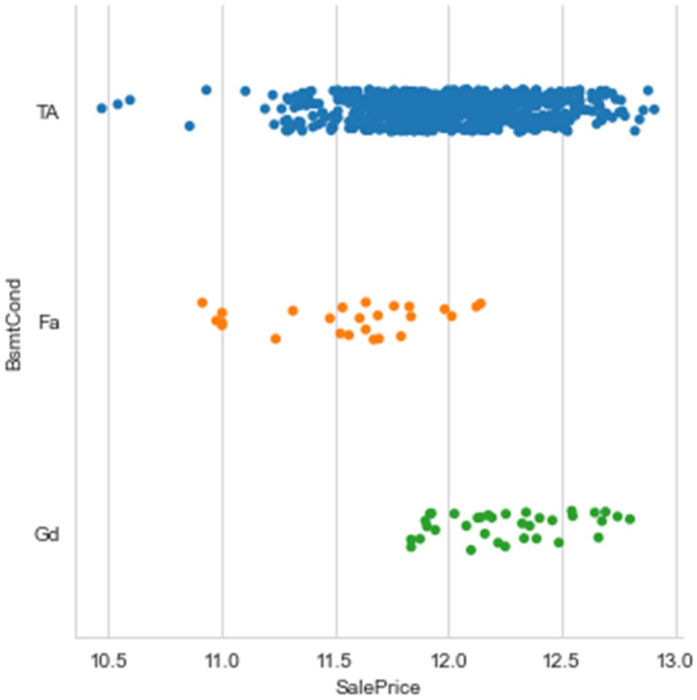


DESCRIPTION: BsmtQual: Evaluates the height of the basement OBSERVATION: The properties with Ex-(Excellent (100+ inches)) height had the highest sale price(mean).

In [2500]:

```
plt.figure(figsize=(25,7))
sns.set_style('whitegrid')
sns.catplot(y='BsmtCond',x='SalePrice',data=df)
plt.show()
```

```
<Figure size 1800x504 with 0 Axes>
```

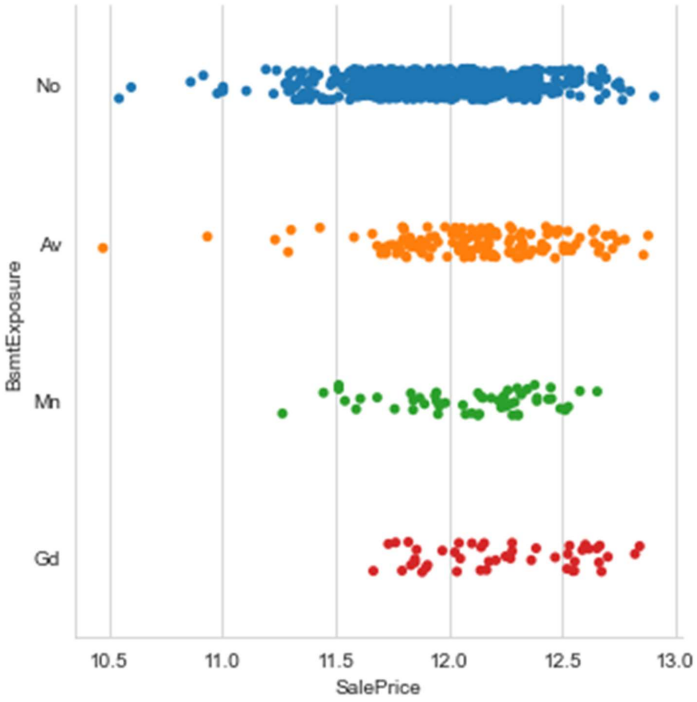


DESCRIPTION: BsmtCond: Evaluates the general condition of the basement OBSERVATION: TA type of condition had the highest saleprice whereas Gd category had the highest mean saleprice. The majority of the cases were of TA category.

In [2501]:

```
plt.figure(figsize=(25,7))
sns.set_style('whitegrid')
sns.catplot(y='BsmtExposure',x='SalePrice',data=df)
plt.show()
```

```
<Figure size 1800x504 with 0 Axes>
```

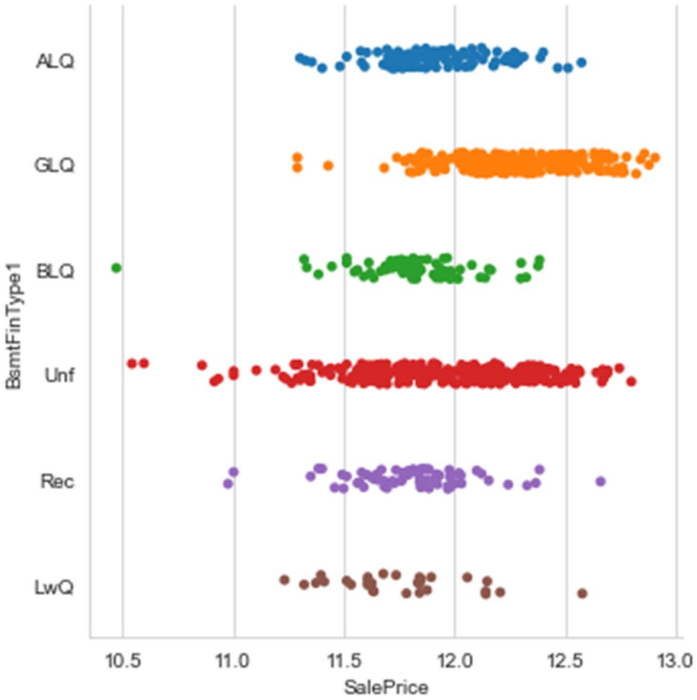


DESCRIPTION: BsmtExposure: Refers to walkout or garden level walls OBSERVATION: The Gd type of properties had the highest SalePrice(mean).

In [2502]:

```
plt.figure(figsize=(25,7))
sns.set_style('whitegrid')
sns.catplot(y='BsmtFinType1',x='SalePrice',data=df)
plt.show()
```

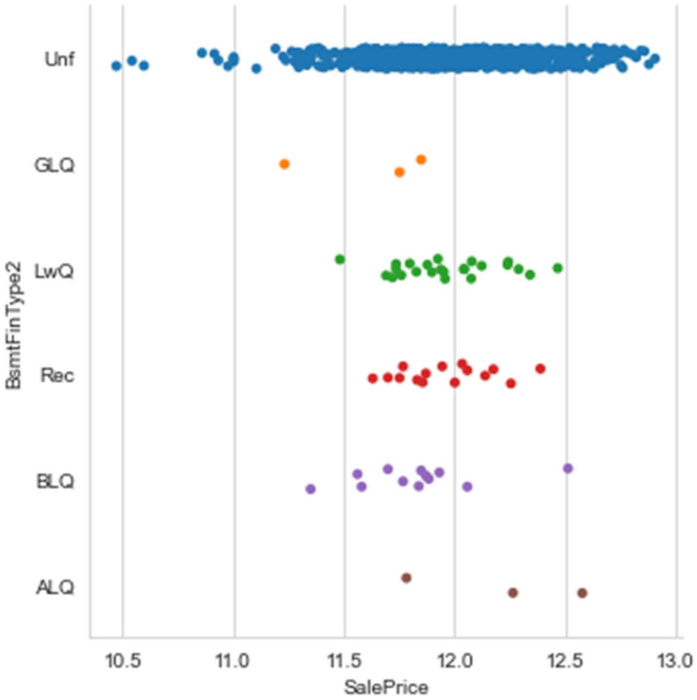
```
<Figure size 1800x504 with 0 Axes>
```



DESCRIPTION: BsmtFinType1: Rating of basement finished area OBSERVATION: The GLQ(Good Living Quarters) had the highest saleprice(mean aswell as maximum)

In [2503]:

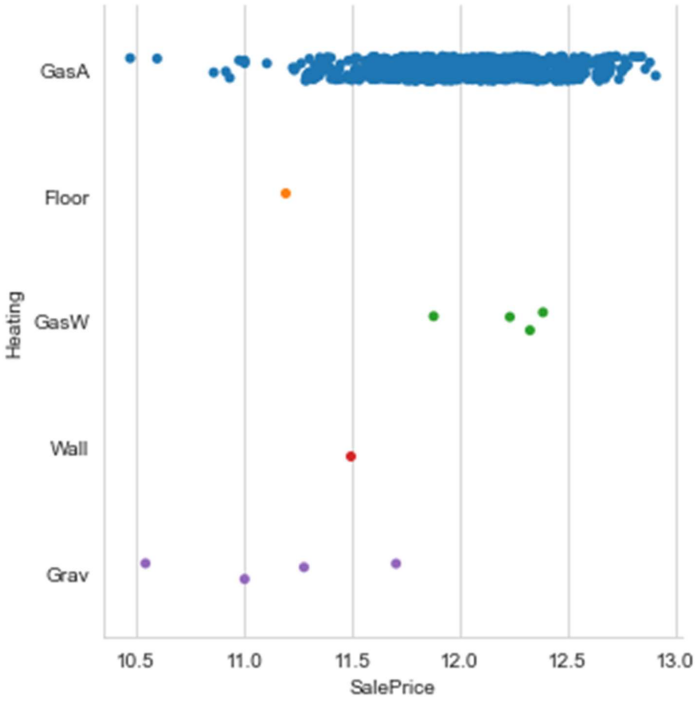
```
sns.set_style('whitegrid')
sns.catplot(y='BsmtFinType2',x='SalePrice',data=df)
plt.show()
```



DESCRIPTION: BsmtFinType2: Rating of basement finished area (if multiple types)
OBSERVATION: Majority of properties had Unf Unfinshed type of BsmtFinType2 which also had the highest saleprice.

In [2504]:

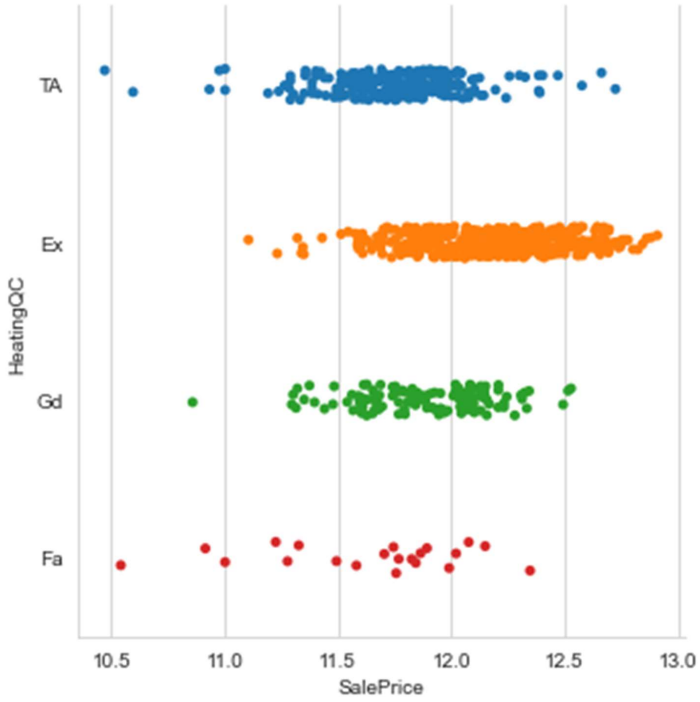
```
sns.set_style('whitegrid')
sns.catplot(y='Heating',x='SalePrice',data=df)
plt.show()
```



DESCRIPTION: Heating: Type of heating OBSERVATION: 1)Most properties had a GasA(Gas forced warm air furnace) type of heating which also had the highest SalePrice.

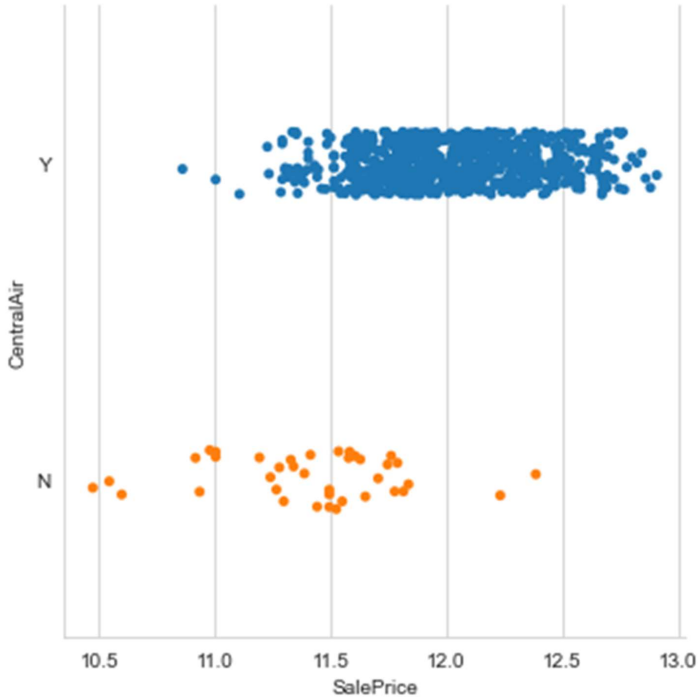
In [2505]:

```
sns.set_style('whitegrid')
sns.catplot(y='HeatingQC',x='SalePrice',data=df)
plt.show()
```



In [2506]:

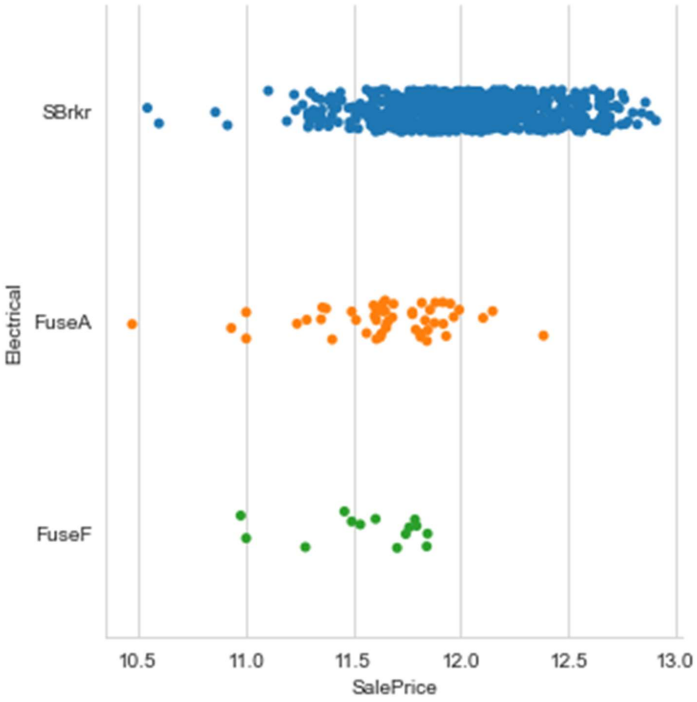
```
sns.set_style('whitegrid')
sns.catplot(y='CentralAir',x='SalePrice',data=df)
plt.show()
```



DESCRIPTION: CentralAir: Central air conditioning OBSERVATION: 1)Most properties had central air conditioning. 2)properties with central air conditioning had the highest saleprice.

In [2507]:

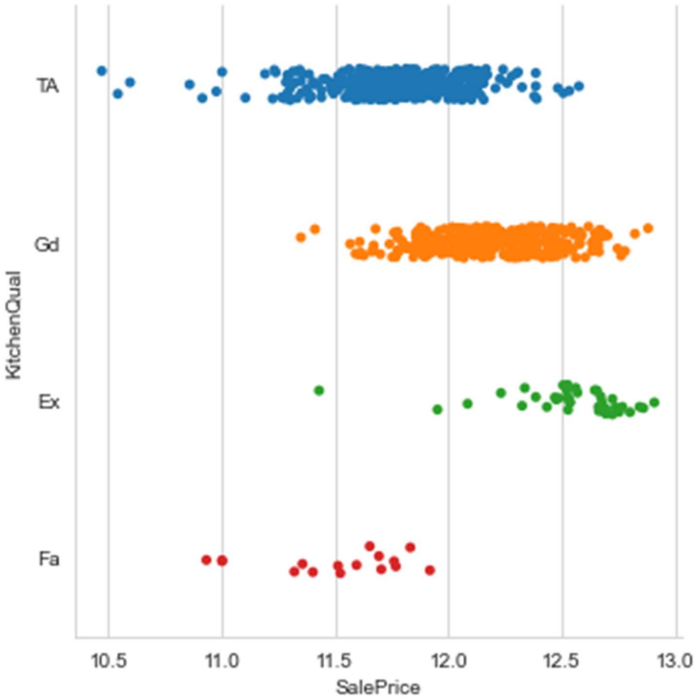
```
sns.set_style('whitegrid')
sns.catplot(y='Electrical',x='SalePrice',data=df)
plt.show()
```



DESCRIPTION: Electrical: Electrical system OBSERVATION: 1)Majority of the properties had SBrkr Standard Circuit Breakers & Romex type of electrical system. 2)SBrkr Standard Circuit Breakers & Romex type of electrical system had the highest sale price.(mean aswell as maximum)

In [2508]:

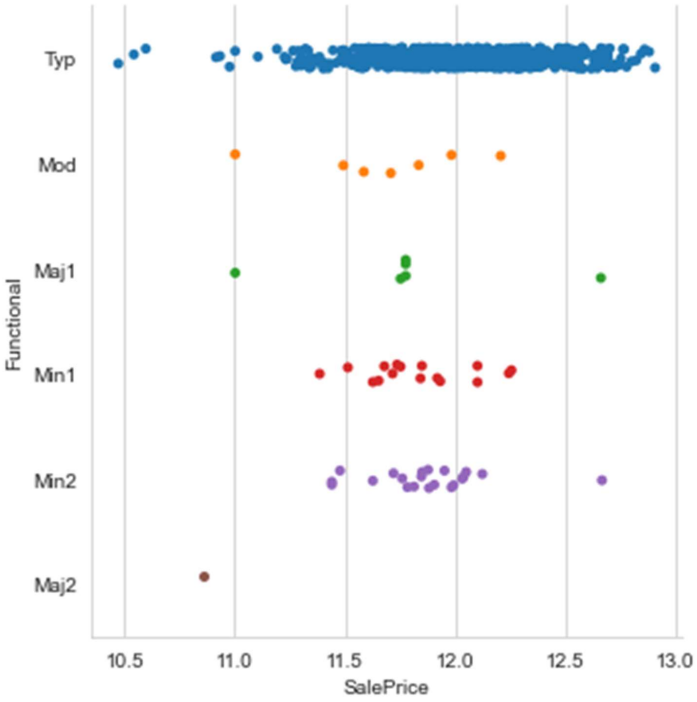
```
sns.set_style('whitegrid')
sns.catplot(y='KitchenQual',x='SalePrice',data=df)
plt.show()
```



DESCRIPTION: KitchenQual: Kitchen quality OBSERVATION: 1)The Ex type of kitchen quality had the highest saleprice(mean aswell as maximum). 2)The Fa type of kitchen quality had the lowest SalePrice(minimum).

In [2509]:

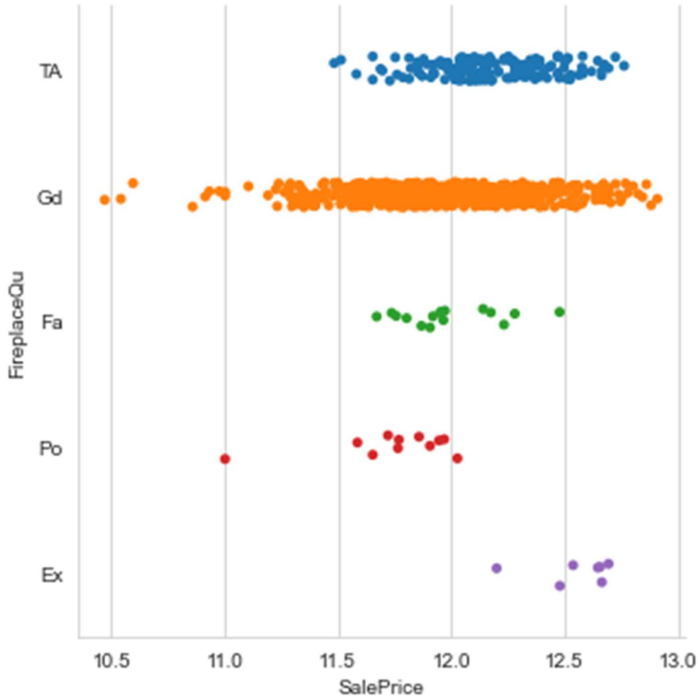
```
sns.set_style('whitegrid')
sns.catplot(y='Functional',x='SalePrice',data=df)
plt.show()
```



DESCRIPTION: Functional: Home functionality (Assume typical unless deductions are warranted) OBSERVATIONS: 1)Most properties had Typ Typical Functionality. 2)Typ Typical Functionality had the highest sale price among all.

In [2510]:

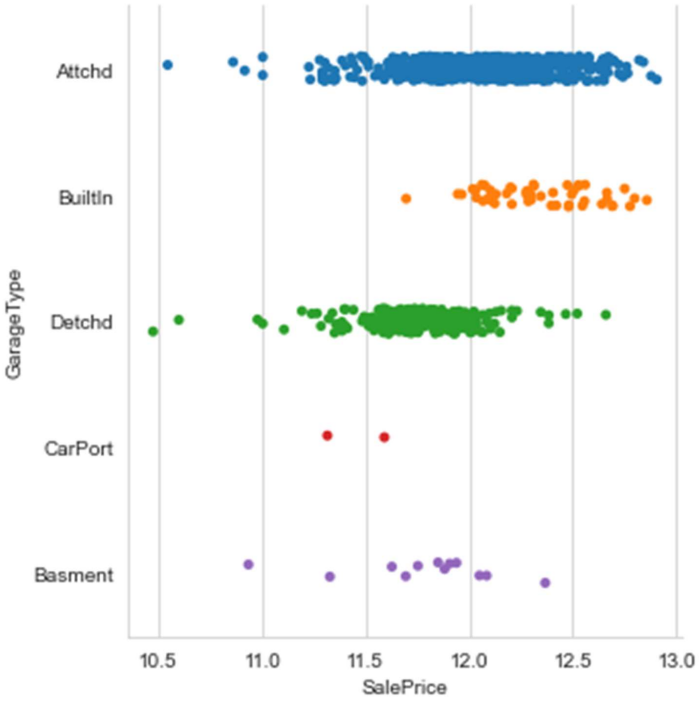
```
sns.set_style('whitegrid')
sns.catplot(y='FireplaceQu',x='SalePrice',data=df)
plt.show()
```



DESCRIPTION: FireplaceQu: Fireplace quality OBSERVATION: 1)The properties having Ex type of quality had the highest mean saleprice. 2)The properties having Po type of quality had the lowest saleprice(mean).

In [2511]:

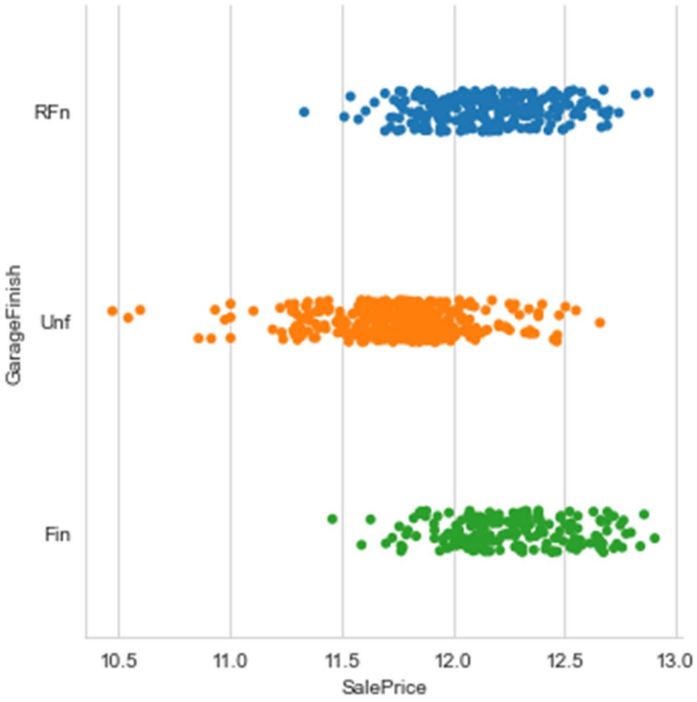
```
sns.set_style('whitegrid')
sns.catplot(y='GarageType',x='SalePrice',data=df)
plt.show()
```



DESCRIPTION: GarageType: Garage location OBSERVATION: 1)BuiltIn Built-In (Garage part of house - typically has room above garage) had the highest SalePrice(mean) among all the locations. 2)The maximum saleprice was recorded in Attchd Attached to home location.

In [2512]:

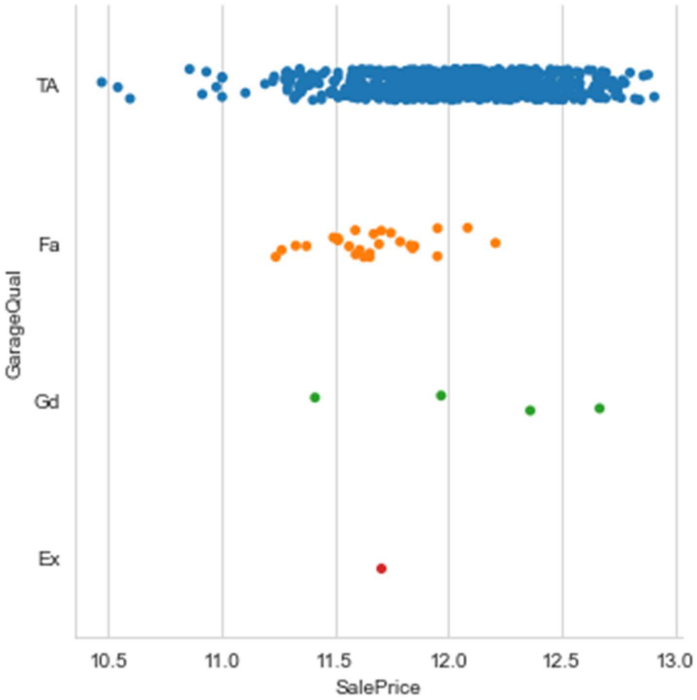
```
sns.set_style('whitegrid')
sns.catplot(y='GarageFinish',x='SalePrice',data=df)
plt.show()
```



DESCRIPTION: GarageFinish: Interior finish of the garage OBSERVATION: 1)The properties with Un(Unfinished) GarageFinish had the lowest SalePrice. 2)RFn(Rough Finished) and Fin(Finished) had similar means of SalePrice but Fin category had a slightly higher mean.

In [2513]:

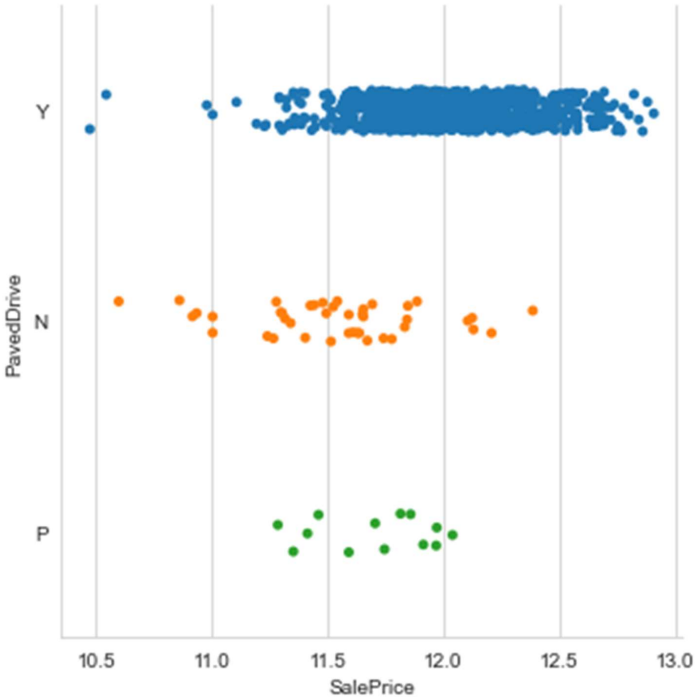
```
sns.set_style('whitegrid')
sns.catplot(y='GarageQual',x='SalePrice',data=df)
plt.show()
```



OBSERVATION: MAJORITY OF THE PROPERTIES HAD TA(Typical/Average) CONDITION.

In [2514]:

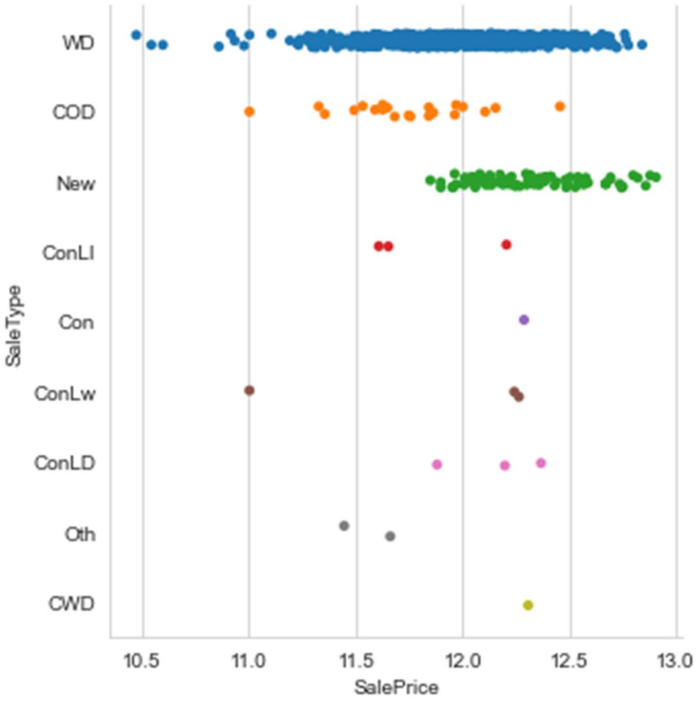
```
sns.set_style('whitegrid')
sns.catplot(y='PavedDrive',x='SalePrice',data=df)
plt.show()
```



DESCRIPTION: PavedDrive: Paved driveway OBSERVATION: The properties with Y(Paved) driveways had the highest SalePrice(mean aswell as maximum) and also constituted bulk of the sample.

In [2515]:

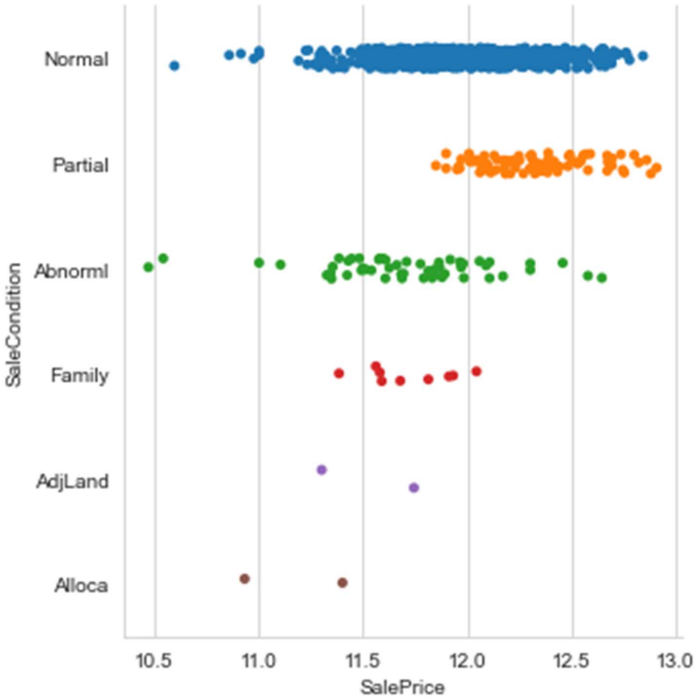
```
sns.set_style('whitegrid')
sns.catplot(y='SaleType',x='SalePrice',data=df)
plt.show()
```



DESCRIPTION: SaleType: Type of sale OBSERVATION: 1)The SaleType New(Home just constructed and sold) had the highest saleprice(mean aswell as maximum). 2)Majority of sales were of WD,New,COD type.(descending order)

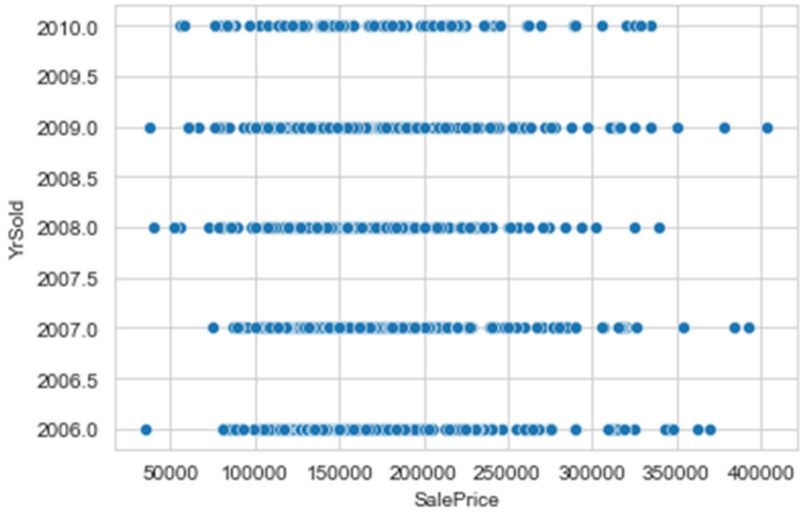
In [2516]:

```
sns.set_style('whitegrid')
sns.catplot(y='SaleCondition',x='SalePrice',data=df)
plt.show()
```



DESCRIPTION: SaleCondition: Condition of sale OBSERVATION: The SaleCondition Partial(Home was not completed when last assessed (associated with New Homes)) had the highest SalePrice(mean as well as maximum)

For contionous data:



- Interpretation of the Results

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

From visualizations, preprocessing and modelling over 5 algorithms it was interpreted that a large number of features were not contributing much to the target prediction and that the regularization techniques worked better in all the evaluation metrics used and the tree algorithms were overfitting the data.The test and the train data had different cardinality for categorical variables and even different datatype for MasVnrArea which made life even more difficult.

The models used for the training and testing of the data with their respective evaluation metrics are:

LINEAR REGRESSION

score: 0.9488184368722168
M.S.E 0.016254007096186053
MEAN ABSOLUTE ERROR 0.08353082192055095
R2 score: 0.8728042034066253
R.M.S.E: 0.12749120399535827
Cross validation: mean score: 0.8889058090007358
cross val score: [0.88785884 0.84830088 0.92879418 0.84723265 0.8634732
5 0.84701418
0.93403808 0.92216169 0.91375455 0.89642978]

Lasso

score: 0.9200571834654206

M.S.E 0.010617742760808641
MEAN ABSOLUTE ERROR 0.07583264146281114
R2 score: 0.9169108121774174
R.M.S.E: 0.10304243184634494
Cross validation: mean score: 0.9045884133057488
cross val score: [0.91245096 0.94149721 0.91150908 0.89023988 0.9273976
5 0.87310974
 0.90859173 0.94511156 0.87486529 0.86111104]

RIDGE

mean score: 0.8889983320614985
cross val score: [0.88791744 0.84884612 0.92878377 0.84732872 0.8636880
7 0.8470689
 0.93406721 0.92216363 0.9137471 0.89637237]

DecisionTreeRegressor

score: 1.0
M.S.E 0.02883628318906452
MEAN ABSOLUTE ERROR 0.12684908115003876
R2 score: 0.7743415522510846
R.M.S.E: 0.16981249420777178
Cross validation: mean score: 0.7207229090345207
cross val score: [0.64935529 0.76009331 0.77676689 0.77106744 0.6907433
1 0.65301005
 0.78193834 0.69094592 0.69270011 0.74060842]

RANDOM FOREST REGRESSOR

score: 0.9808914619953911
M.S.E 0.01691526852245359
MEAN ABSOLUTE ERROR 0.09381069815195504
R2 score: 0.8676294995091288
R.M.S.E: 0.13005871182836462
Cross validation: mean score: 0.8665266419037223
cross val score: [0.8762561 0.89072434 0.89009571 0.87598657 0.9189037
8 0.82798421
 0.85760934 0.88256574 0.85853652 0.7866041]

ADABOOST REGRESSOR

score: 0.8927793312111704
M.S.E 0.025930036260681424
MEAN ABSOLUTE ERROR 0.12100649718646102
R2 score: 0.7970843990435827
R.M.S.E: 0.16102806047605933
Cross validation: mean score: 0.8185870323938571
cross val score: [0.81914835 0.84254528 0.81413889 0.78386369 0.8947737
1 0.80689551
 0.80277597 0.85845146 0.79896241 0.76431506]

GRADIENT BOOST REGRESSOR

score: 0.9724874115663982
M.S.E 0.013133193773183352
MEAN ABSOLUTE ERROR 0.08394238999771032
R2 score: 0.897226140365893
R.M.S.E: 0.11460014735236318
Cross validation: mean score: 0.8895077997205851

```
cross val score: [0.89879235 0.93581665 0.91162561 0.87288682 0.9104984
9 0.8392594
 0.88258582 0.91519058 0.87657636 0.85184593]
```

Remarks: LASSO REGULARIZATION HAD THE BEST CROSS VALIDATION SCORE

Choosing The Model:

SINCE LASSO PERFORMED BEST IN ALMOST ALL THE METRICS:R2_Score,Cross_Val_Score,Mean_Squared_Error,R.M.S.E.,IT HAS BEEN CHOOSEN AS THE BEST MODEL.

CONCLUSION

- Key Findings and Conclusions of the Study

  Describe the key findings, inferences, observations from the whole problem.

  Key finding:

The features that influcenced the target variable the most were:

```
'MSSubClass', 'FV', 'RL', 'LotFrontage', 'LotArea', 'HLS', 'CulDSac',
       'Mod', 'BrkSide', 'CollgCr', 'Crawfor', 'Edwards', 'IDOTRR', 'NA
mes',
       'NridgHt', 'Somerst', 'Feedr', 'TwnhsE', 'OverallQual', 'Overall
Cond',
       'YearBuilt', 'YearRemodAdd', 'Gambrel', 'CBlock', 'HdBoard', 'Me
talSd',
       'Plywood', 'VinylSd', 'Fa', 'PConc', 'Slab', 'No', 'BsmtFinSF1',
       'BsmtFinSF2', 'BsmtUnfSF', 'Y', 'SBrkr', '2ndFlrSF', 'LowQualFin
SF',
       'BsmtFullBath', 'FullBath', 'BedroomAbvGr', 'Typ', 'Fireplaces',
       'GarageYrBlt', 'RFn', 'WoodDeckSF', 'OpenPorchSF', 'EnclosedPorc
h',
       '3SsnPorch', 'ScreenPorch', 'MiscVal', 'MoSold', 'YrSold', 'Allo
ca',
       'Normal', 'Partial', 'TotalBsmtSF & 1stFlrSF',
       'GarageCars & GarageArea', 'TotRmsAbvGrd & GrLivArea'],
```

THE ABOVE ARE THE FEATURES THAT CONTRIBUTED THE MOST IN PREDICTING THE SALEPRICE OF THE TEST DATA.

- Learning Outcomes of the Study in respect of Data Science

List down your learnings obtained about the power of visualization, data cleaning and various algorithms used. You can describe which algorithm works best in which situation and what challenges you faced while working on this project and how did you overcome that.

The biggest challenges faced while working on this project were the different number and datatypes of features in the test and the train data and a large chunk of data missing from a number of features.Also, the dimensionality of the dataset too made it difficult to work on.

The methods used to overcome the aforementioned problems were:

Handling Outliers In the dataset, there are many homes with zero value for Garage Area, indicating that they don't have a garage. We'll transform other features later to reflect this assumption. There are a few outliers as well. Outliers can affect a regression model by pulling our estimated regression line further away from the true population regression line. So, we remove or replace those observations from our data. There are number of values that seem rather strange. In both training and test sets, we have several garages that were built as many as 20 years earlier than their houses and in the training set we have a garage from the future – the record claims that it was built in 2209! Clearly something has gone wrong with these entries and – if we have some means to do so – we would ideally replace them with corrected values. If this is not possible, however, we can proceed to treat them as if they were missing.

The dataset had a significant number of outliers was detected in columns such as LotArea,MasVnrArea etc.The outliers had to be removed using the zscore class from sklearn.preprocessing which eliminates all the rows from the dataset which contain numbers that fall beyond 3-standard deviation from the mean.

Since the majority of available machine learning algorithms can only take numbers (floats or integers) as inputs, we must encode these features numerically if we are to use them in our models. The way we go about this will vary depending on the nature of the feature and the model we decide to use.

For the same,a function named one_hot_encode was created to create dummy variables for the categorical features which had a cardinalty less than 25.The original features were then dropped and the new dummies were concatenated into a new dataframe.

Once the data is clean and we have gained insights about the dataset, we can apply an appropriate machine learning model that fits our dataset. We have selected four algorithms to predict the dependent variable in our dataset. The algorithms that we have selected are basically used as classifiers but we are training them to predict the continuous values. The algorithms we have used here are :

LinearRegression,Lasso,Ridge,DecisionTreeRegressor,RandomForestRegressor,AdaBoostRegressor and GradientBoostingRegressor.

- Limitations of this work and Scope for Future Work

  What are the limitations of this solution provided, the future scope? What all steps/techniques can be followed to further extend this study and improve the results.

  In this research paper, we have used machine learning algorithms to predict the house prices. We have mentioned the step by step procedure to analyze the dataset and finding the correlation between the parameters. Thus we can select the parameters which are not correlated to each other and are independent in nature. These feature set were then given as an input to four algorithms and a csv file was generated consisting of predicted house prices. Hence we calculated the performance of each model using RMSE metric and compared them. We found that Blend_models and Lasso Regression overfits our dataset and give the least RMSE of 0.103 and 0.124 respectively. For future work, we recommend that working on large dataset would yield a better and real picture about the model. We have undertaken only few Machine Learning algorithms that are actually classifiers but we need to train many other classifiers and understand their predicting behavior for continuous values too. By improving the error values this research work can be useful for development of applications for various respective cities.