House Price Prediction

Submitted by:

Er. Deepesh Singh

ACKNOWLEDGMENT

On the very outset of this report, I would like to extend my sincere & heartfelt obligation towards all the personages who have helped me in this endeavour. Without their active guidance, help, cooperation & encouragement, I would not have made headway in the project.

I am ineffably indebted to my SME for their conscientious guidance and encouragement to accomplish this assignment.

I am extremely thankful and pay my gratitude to my SME for his valuable guidance and support on completion of this project.

I extend my gratitude to FLIP ROBO Technologies for giving me this opportunity.

Any omission in this brief acknowledgement does not mean lack of gratitude.

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. We model the price of houses with the available independent variables. This model will then be used by the management to understand how the exactly the prices vary with the variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand the pricing dynamics of a new market.

- ## Conceptual Background of the Domain Problem

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.

- **Motivation for the Problem Undertaken**

  Main objective behind to make this project is that to predicting the actual value of the prospective properties and decide whether to invest in them or not. We build this project for describing the price of the house. 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 and give us to analyse the data and predict the actual price of the houses.

Analytical Problem Framing

- **Mathematical/ Analytical Modelling of the Problem**

  It's a regression problem**,** I use some libraries like Pandas, NumPy, seaborn, matplotlib, for building the project and done EDA with the help of these libraries. And use correlation matrix through heatmap to show the correlation between the each of the columns.

- **Data Sources and their formats**

  Data description: 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.

An 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)

NoSewrElectricity, 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

Neighbourhood: 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

BrkFaceBrick Face

CBlock Cinder Block

CemntBd Cement Board

HdBoard Hard Board

ImStuccImitation Stucco

MetalSd Metal Siding

Other Other

Plywood Plywood

PreCast

Stone Stone

Stucco Stucco

VinylSd Vinyl Siding

Wd Sdng Wood Siding

WdShing Wood Shingles

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

AsbShng Asbestos Shingles

AsphShn Asphalt Shingles

BrkComm Brick Common

BrkFaceBrick Face

CBlock Cinder Block

CemntBd Cement Board

HdBoard Hard Board

ImStucc Imitation Stucco

MetalSd Metal Siding

Other Other

Plywood Plywood

PreCast PreCast

Stone Stone

Stucco Stucco

VinylSd Vinyl Siding

Wd Sdng Wood Siding

WdShing Wood Shingles

MasVnrType: Masonry veneer type

BrkCmn Brick Common

BrkFace Brick Face

CBlock Cinder Block

None None

Stone Stone

MasVnrArea: Masonry veneer area in square feet

ExterQual: Evaluates the quality of the material on the exterior

Ex	Excellent
Gd	Good
TA	Average/Typical
Fa	Fair
Po	Poor

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

Ex	Excellent
Gd	Good
TA	Average/Typical
Fa	Fair
Po	Poor

Foundation: Type of foundation

BrkTil	Brick & Tile
CBlock	Cinder Block
PConc	Poured Contrete
Slab	Slab
Stone	Stone
Wood	Wood

BsmtQual: Evaluates the height of the basement

Ex Excellent (100+ inches)

Gd Good (90-99 inches)

TA Typical (80-89 inches)

Fa Fair (70-79 inches)

Po Poor (<70 inches

NA No Basement

BsmtCond: Evaluates the general condition of the basement

Ex Excellent

Gd Good

TA Typical - slight dampness allowed

Fa Fair - dampness or some cracking or settling

Po Poor - Severe cracking, settling, or wetness

NA No Basement

BsmtExposure: Refers to walkout or garden level walls

Gd Good Exposure

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

Mn Mimimum Exposure

No No Exposure

NA No Basement

BsmtFinType1: Rating of basement finished area

GLQ	Good Living Quarters
ALQ	Average Living Quarters
BLQ	Below Average Living Quarters
Rec	Average Rec Room
LwQ	Low Quality
Unf	Unfinshed
NA	No Basement

BsmtFinSF1: Type 1 finished square feet

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

GLQ	Good Living Quarters
ALQ	Average Living Quarters
BLQ	Below Average Living Quarters
Rec	Average Rec Room
LwQ	Low Quality
Unf	Unfinshed
NA	No Basement

BsmtFinSF2: Type 2 finished square feet

BsmtUnfSF: Unfinished square feet of basement area

TotalBsmtSF: Total square feet of basement area

Heating: Type of heating

Floor	Floor Furnace
GasA	Gas forced warm air furnace
GasW	Gas hot water or steam heat
Grav	Gravity furnace
OthW	Hot water or steam heat other than gas
Wall	Wall furnace

HeatingQC: Heating quality and condition

Ex	Excellent
Gd	Good
TA	Average/Typical
Fa	Fair
Po	Poor

CentralAir: Central air conditioning

N No

Y Yes

Electrical: Electrical system

SBrkr	Standard Circuit Breakers & Romex
FuseA	Fuse Box over 60 AMP and all Romex wiring (Average)
FuseF	60 AMP Fuse Box and mostly Romex wiring (Fair)
FuseP	60 AMP Fuse Box and mostly knob & tube wiring (poor)
Mix	Mixed

1stFlrSF: First Floor square feet

2ndFlrSF: Second floor square feet

LowQualFinSF: Low quality finished square feet (all floors)

GrLivArea: Above grade (ground) living area square feet

BsmtFullBath: Basement full bathrooms

BsmtHalfBath: Basement half bathrooms

FullBath: Full bathrooms above grade

HalfBath: Half baths above grade

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

Kitchen: Kitchens above grade

KitchenQual: Kitchen quality

Ex	Excellent
Gd	Good
TA	Typical/Average
Fa	Fair
Po	Poor

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

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

Typ	Typical Functionality
Min1	Minor Deductions 1
Min2	Minor Deductions 2
Mod	Moderate Deductions
Maj1	Major Deductions 1
Maj2	Major Deductions 2

Sev Severely Damaged

Sal Salvage only

Fireplaces: Number of fireplaces

FireplaceQu: Fireplace quality

Ex Excellent - Exceptional Masonry Fireplace

Gd Good - Masonry Fireplace in main level

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

Fa Fair - Prefabricated Fireplace in basement

Po Poor - Ben Franklin Stove

NA No Fireplace

GarageType: Garage location

2Types More than one type of garage

Attchd Attached to home

Basment Basement Garage

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

CarPort Car Port

Detchd Detached from home

NA No Garage

GarageYrBlt: Year garage was built

GarageFinish: Interior finish of the garage

Fin	Finished
RFn	Rough Finished
Unf	Unfinished
NA	No Garage

GarageCars: Size of garage in car capacity

GarageArea: Size of garage in square feet

GarageQual: Garage quality

Ex	Excellent
Gd	Good
TA	Typical/Average
Fa	Fair
Po	Poor
NA	No Garage

GarageCond: Garage condition

Ex	Excellent

Gd Good

TA Typical/Average

Fa Fair

Po Poor

NA No Garage

PavedDrive: Paved driveway

Y Paved

P Partial Pavement

N Dirt/Gravel

WoodDeckSF: Wood deck area in square feet

OpenPorchSF: Open porch area in square feet

EnclosedPorch: Enclosed porch area in square feet

3SsnPorch: Three season porch area in square feet

ScreenPorch: Screen porch area in square feet

PoolArea: Pool area in square feet

PoolQC: Pool quality

Ex	Excellent
Gd	Good
TA	Average/Typical
Fa	Fair
NA	No Pool

Fence: Fence quality

GdPrv	Good Privacy
MnPrv	Minimum Privacy
GdWo	Good Wood
MnWw	Minimum Wood/Wire
NA	No Fence

MiscFeature: Miscellaneous feature not covered in other categories

Elev	Elevator
Gar2	2nd Garage (if not described in garage section)
Othr	Other
Shed	Shed (over 100 SF)
TenC	Tennis Court
NA	None

MiscVal: $Value of miscellaneous feature

MoSold: Month Sold (MM)

YrSold: Year Sold (YYYY)

SaleType: Type of sale

WD	Warranty Deed - Conventional
CWD	Warranty Deed - Cash
VWD	Warranty Deed - VA Loan
New	Home just constructed and sold
COD	Court Officer Deed/Estate
Con	Contract 15% Down payment regular terms
ConLw	Contract Low Down payment and low interest
ConLI	Contract Low Interest
ConLD	Contract Low Down
Oth	Other

SaleCondition: Condition of sale

Normal Normal Sale

Abnorml Abnormal Sale - trade, foreclosure, short sale

AdjLand Adjoining Land Purchase

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

Family Sale between family members

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

- **Data Preprocessing Done**

  In this first I drop the unusual columns which is not necessary for data building and after that done EDA on the data set w.r.t to Sale price of the houses, after that remove the outliers and skewness of the dataset.

- **Hardware and Software Requirements and Tools Used**

  In Hardware I use the laptop with the i5 processor and of 8 GB of ram. And in software I use an anaconda and in anaconda jupyter notebook software is used. In jupyter notebook I use python for making my project and libraries used for making the project is Pandas, NumPy, seaborn, matplotlib. Through pandas I loaded the dataset into the python.

Model/s Development and Evaluation

- **Testing of Identified Approaches (Algorithms)**

  I use the Train-Test split for training and testing 0f the data.

```
In [41]: #Splitting the data into train and test set
         from sklearn.model_selection import train_test_split

         x_train,x_test,y_train,y_test=train_test_split(x,y,random_state=31)
```

```
In [42]: len(x_train),len(x_test),len(y_train),len(y_test)

Out[42]: (876, 292, 876, 292)
```

- **Run and Evaluate selected models**

I have use the six algorithms for building the model. Snapshot of all those are as follows:

Model Building

```
In [49]: from sklearn.metrics import r2_score,mean_squared_error,mean_absolute_error
         from sklearn.linear_model import Lasso,Ridge
```

```
In [50]: lm=LinearRegression()
         lm.fit(x_train,y_train)
         predlm=lm.predict(x_test)
         print('r2_score:',r2_score(y_test,predlm))
         print('Mean absolute error:',mean_absolute_error(y_test,predlm))
         print('Mean squared error:',mean_squared_error(y_test,predlm))

         r2_score: 0.7526061354329651
         Mean absolute error: 23820.51589034116
         Mean squared error: 1627107369.6812017
```

```
In [67]: rf=RandomForestRegressor(n_estimators=100,random_state=31)
         rf.fit(x_train,y_train)
         predrf=rf.predict(x_test)
         print('r2_score:',r2_score(y_test,predrf))
         print('Mean absolute error:',mean_absolute_error(y_test,predrf))
         print('Mean squared error:',mean_squared_error(y_test,predrf))

         r2_score: 0.8243099933042402
         Mean absolute error: 19629.157054794523
         Mean squared error: 1155511698.6199586
```

```
In [68]: adb=AdaBoostRegressor(n_estimators=100,random_state=31)
         adb.fit(x_train,y_train)
         predadb=adb.predict(x_test)
         print('r2_score:',r2_score(y_test,predadb))
         print('Mean absolute error:',mean_absolute_error(y_test,predadb))
         print('Mean squared error:',mean_squared_error(y_test,predadb))

         r2_score: 0.8018798589291966
         Mean absolute error: 25536.99866068517
         Mean squared error: 1303034504.0397508
```

```
In [69]: gbr=GradientBoostingRegressor(n_estimators=100,random_state=31)
         gbr.fit(x_train,y_train)
         predgbr=gbr.predict(x_test)
         print('r2_score:',r2_score(y_test,predgbr))
         print('Mean absolute error:',mean_absolute_error(y_test,predgbr))
         print('Mean squared error:',mean_squared_error(y_test,predgbr))

         r2_score: 0.8145911309177485
         Mean absolute error: 18630.95923507881
         Mean squared error: 1219432574.9183803
```

```
In [70]: ls=Lasso(alpha=0.0001,random_state=31)
         #ls=lasso(alpha=1.0) #default
         ls.fit(x_train,y_train)
         predls=ls.predict(x_test)
         print('r2_score:',r2_score(y_test,predls))
         print('Mean absolute error:',mean_absolute_error(y_test,predls))
         print('Mean squared error:',mean_squared_error(y_test,predls))

         r2_score: 0.7526061419329668
         Mean absolute error: 23820.496967401396
         Mean squared error: 1627107326.9307446
```

```
In [71]: rd=Ridge(alpha=0.0001)
         #rd=Ridge()
         rd.fit(x_train,y_train)
         predrd=rd.predict(x_test)
         print('r2_score:',r2_score(y_test,predrd))
         print('Mean absolute error:',mean_absolute_error(y_test,predrd))
         print('Mean squared error:',mean_squared_error(y_test,predrd))

         r2_score: 0.7526061539079597
         Mean absolute error: 23820.5156645896
         Mean squared error: 1627107248.1713166
```

- **Visualizations**

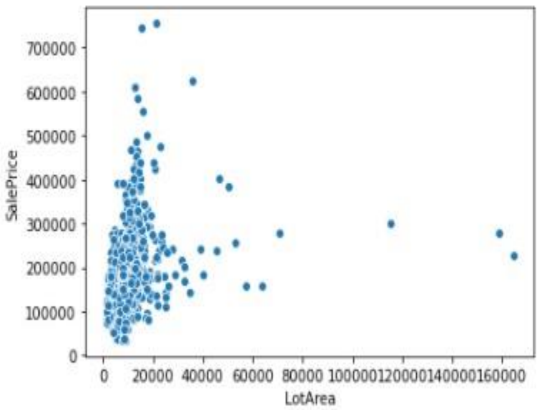
  Firstly I plot the scatterplot which is bivariate analysis. By which we can see that the positive relation is shown between the (lotarea vs saleprice) and (grlivarea vs saleprice).

  Snapshot of that is as below:

Exploratory Data Analysis

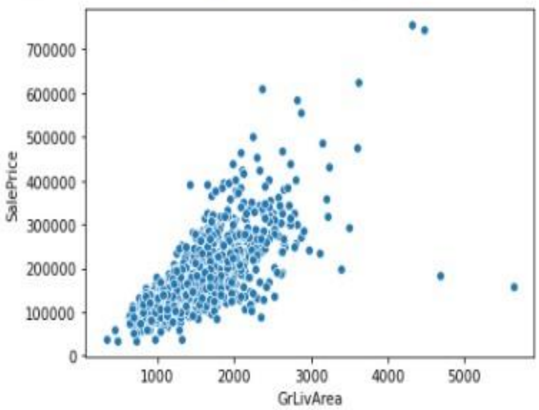
In [20]:
```
#example of bivariate analysis
sns.scatterplot(x='LotArea',y='SalePrice',data=train_df)
```

Out[20]: <matplotlib.axes._subplots.AxesSubplot at 0x1b1c85ec160>



In [21]:
```
sns.scatterplot(x='GrLivArea',y='SalePrice',data=train_df)
```

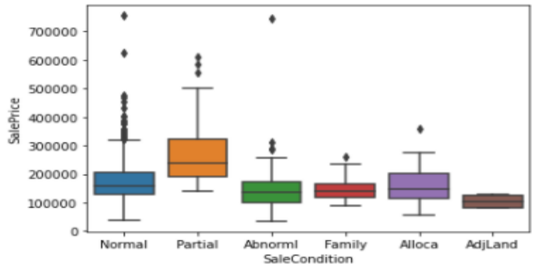
Out[21]: <matplotlib.axes._subplots.AxesSubplot at 0x1b1c83e8eb0>



from above scatterplots we can see outliers are present

After that I plot the box plot comment is on the snapshot itself.

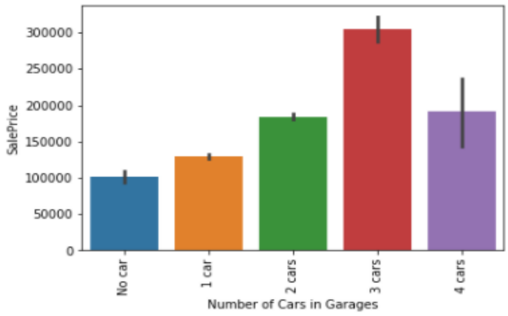
In [23]:
```
#Boxplot
sns.boxplot(y='SalePrice',x='SaleCondition',data=train_df)
```
Out[23]: <matplotlib.axes._subplots.AxesSubplot at 0x1b1c44be0a0>



from above boxplot we can see that there is higher no. of houses are in partial salecondition and these house have more saleprices.

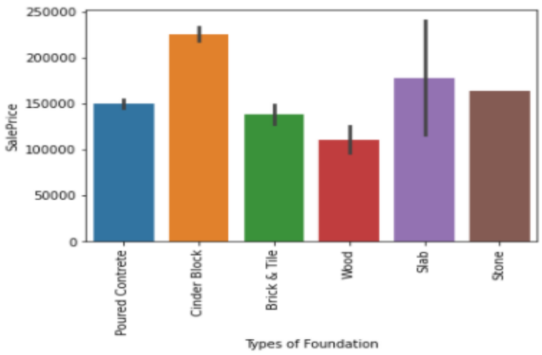
Barplot

In [25]:
```
fig=sns.barplot(x='GarageCars',y='SalePrice',data=train_df)
fig.set_xticklabels(labels=['No car','1 car','2 cars','3 cars','4 cars'], rotation=90)
plt.xlabel("Number of Cars in Garages");
```



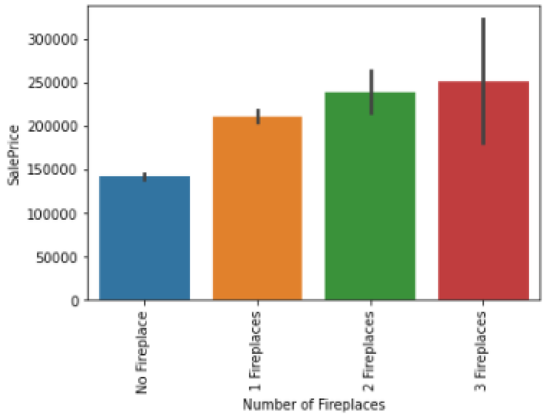
we can see from above barplot that house with garage of 3 cars has the highest sale price

In [26]:
```
fig = sns.barplot(x = 'Foundation',y = 'SalePrice', data = train_df)
fig.set_xticklabels(labels=['Poured Contrete', 'Cinder Block', 'Brick & Tile', 'Wood', 'Slab', 'Stone']
plt.xlabel("Types of Foundation");
```



we can see that from above barplot cinder block types of foundation have the highest saleprice copare to other one.

In [27]:
```
fig = sns.barplot(x = 'Fireplaces',y = 'SalePrice', data = train_df)
fig.set_xticklabels(labels=['No Fireplace', '1 Fireplaces', '2 Fireplaces', '3 Fireplaces'], rotation=
plt.xlabel("Number of Fireplaces");
```

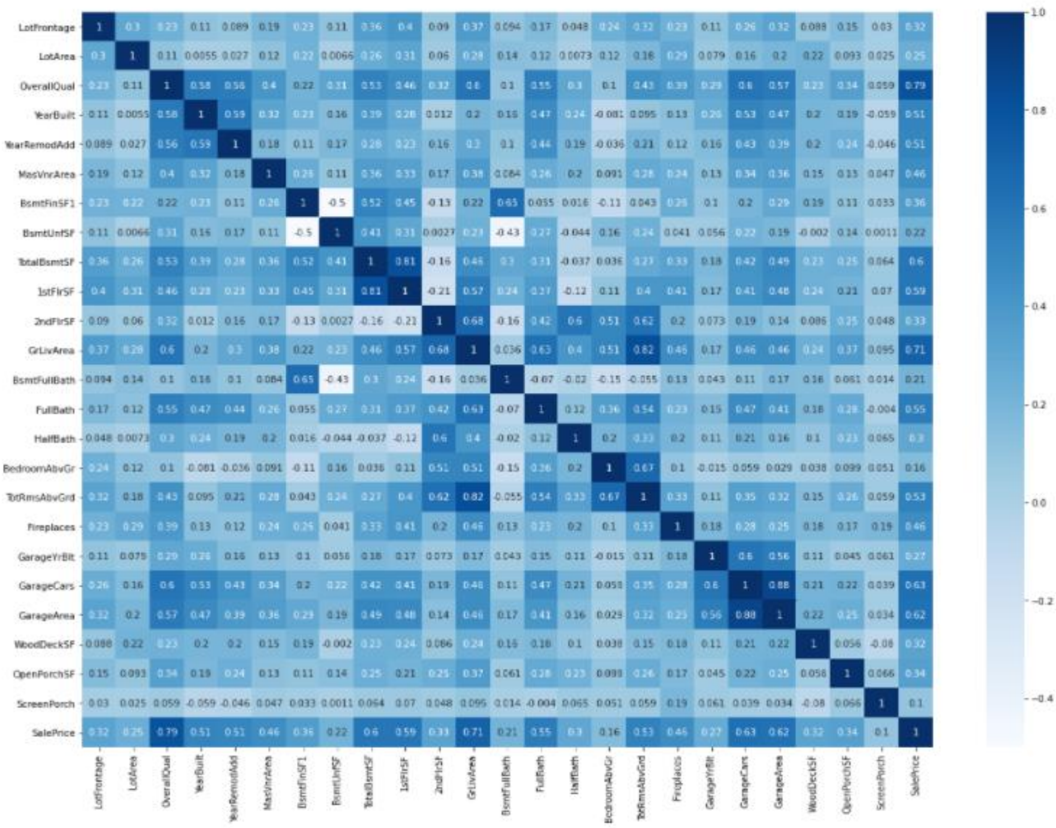


in above barplot we can see that there is most of houses are with 3 fireplaces and highest in saleprice.

After that I plot the correlation matrix of multivariate analysis

In [28]:
```
#correlation matrix type of multivariate analysis
fig=plt.figure(figsize=(20,15))
hc=train_df.corr(method='pearson')
sns.heatmap(hc,annot=True,cmap='Blues')
```

Out[28]: <matplotlib.axes._subplots.AxesSubplot at 0x1b1c98b7190>



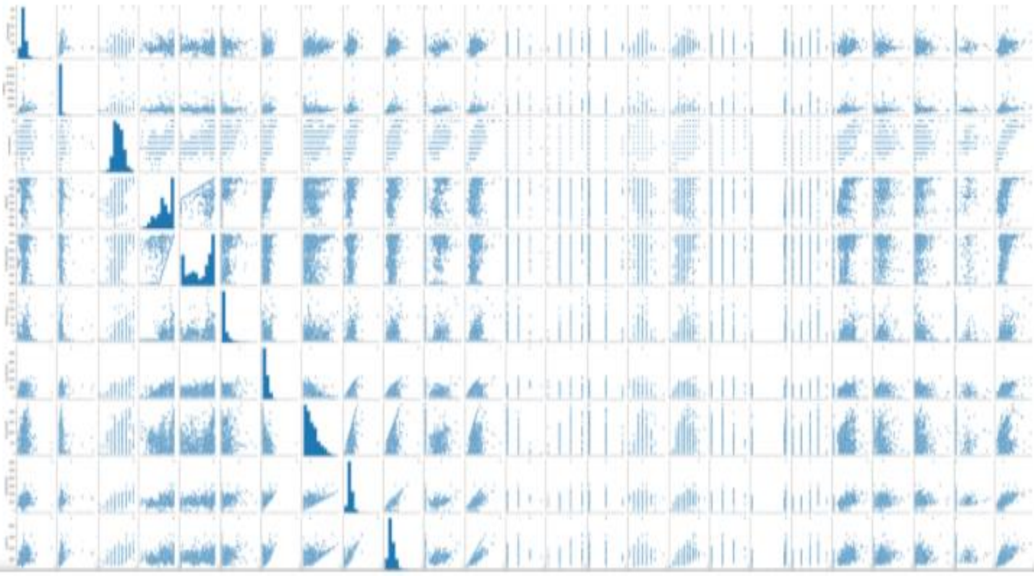
Its shows the correlation matrix of data with respect to saleprice

After that plotting the pair plot which is also an example of multivariate example, in this plot we see all the relation between the columns we scatter plot, density plot.

```
In [29]: #ploting the pairplot which is the type of multivariate analysis

         sns.pairplot(train_df)

Out[29]: <seaborn.axisgrid.PairGrid at 0x1b1ca3cfc70>
```



CONCLUSION

- **Key Findings and Conclusions of the Study**

  I find the highest r2 score of Adaboost regressor with the base estimator of linear regression model after cross validation and hyper parameter tunning my r2 score is 76.30% considering the best model. Then after it I prepare for the test data set and done all the cleaning process and make it for predicting the prices of the houses.

- **Learning Outcomes of the Study in respect of Data Science**

  In the data cleaning I firstly check the null values in the dataset and remove the null values by dropping the columns and using mean of that particular columns. Remove all those columns which is not necessary for the model building. Then I check the outliers and remove it and check the skewness in the data and remove it.

  Done a label encoding on the categorical columns for making the dataset for the model building.

  In model building I use the 6 algorithm and adaboostregressor works best of all in 6.

  In hyperparameter tuning I face the problem for finding the best parameter and running the code after the guidance of my SME I resolve the problem and completed my project.

  This study shows how we can buy the properties. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand the pricing dynamics of a new market.

- **Limitations of this work and Scope for Future Work**

This model helped to identify which characteristics of housing were most strongly associated with price and could explain most of the price variation. Furthermore, we were able to improve our models' prediction accuracy by accounting for the impact of spatial location. We were able to identify most of the residential areas. There may be some more places that have housing complexes or multi-story apartments that are located in commercial areas. Such apartments were not included in this paper and can be counted in the future to give a more accurate result. With more and more demand for housing in metropolitan cities, there is a definite increase in the number of private builders that provide real estate with additional amenities to attract more customers.