PROJECT REPORT

TEAM MEMBERS:

- **Sarah Alalawi (11377856)**
- **Yashaswini Reddy Kalvakol (11542882)**
- **Niharika Ravela (11545991)**

Under the Guidance of Dr. Denise Philpot.

CONTENTS

INTRODUCTION

We as a team analyzed the local housing market activity and have worked on to come up with research questions which we expect to provide guidance to both sellers and buyers respectively. We have followed the step-by-step process to understand the data; prepare the data as per requirements; worked on building research questions to bring the necessary results; also dived into discussions, limitations, and recommendations. We hope this report illustrates the efforts we put in to bring out the best out of the data set provided. In this process we have learnt Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation/Validation, Deployment/ Presentation which is the CRISP-DM Analytics Process.

BUSINESS UNDERSTANDING

The housing data file consists of 25 variables in total, we have taken the SalePrice has a dependent variable, because that is what is the most important to us to give the guidance to sellers and buyers based on other independent variables. We have used Excel as the tool to work on the dataset and in doing the analysis. We have taken the dataset and used tools in Excel to understand the data, perform data preparation, and thereby analyzed the data to perform ANOVA, Correlation and Regression to bring the best model. The variables such as Lot Area, Neighborhood, Bldg Type, House Style, Year Built, BsmtFin SF 1, BsmtFin SF 2, Bsmt Unf SF, Total Bsmt SF, Central Air, Gr Liv Area, Bsmt Full Bath, Bsmt Half Bath, Full Bath, Half Bath, Bedroom AbvGr, Kitchen AbvGr, TotRms AbvGrd, Fireplaces, Garage Cars, Garage Area, Mo Sold, Yr Sold are taken as independent variables in determining the dependent variable SalePrice.

RESEARCH QUESTIONS

The main goal of the project is to find and analyze a pattern in the local housing market by using various methods and techniques like Analysis of Variance, Correlation, Regression. Also, to help as the mediators provide a best bridge between the buyers and sellers. In order to achieve the desired information, we considered following research questions:

1. Considering the level of significance of 0.05, is the Sales Price of house affected with respect to the academic season of the year like Fall, Summer and Spring.(Perform ANOVA)

2. Does the type of building show any change in the sales price of house at 0.05 level of significance? (Perform ANOVA)

3. Is having a central air conditioner in a house leaving an impact on the cost of the house at 0.05 level of significance?(Perform ANOVA)

4. How do the variables in the dataset relate to each other?(Perform correlation for the same)

5. How the independent variables(using the variables in transformed dataset) impact the SalePrice of house thereby build the regression model for the same and what is the regression Equation and R square value?

6. Are the variables in question 5 statistically significant?.Develop the best regression model using the statistically significant variables. Find the best regression model and thereby regression equation for the same.

DATA PREPARATION

The housing data file consists of 25 variables in total, which include:

Variable	Description
Order	id number
Lot Area	size of lot
Neighborhood	name of neighborhood
Bldg Type	single family, duplex, condo, townhouse, townhouse end unit
House Style	1 story, 1.5 story, 2 stories
Year Built	year house was built
BsmtFin SF 1	finished square footage in basement area 1
BsmtFin SF 2	finished square footage in basement area 2
Bsmt Unf SF	unfinished square footage in basement
Total Bsmt SF	total square footage in basement
Central Air	does the house have central air conditioning
Gr Liv Area	gross living area of house
Bsmt Full Bath	number of full baths in basement
Bsmt Half Bath	number of half baths in basement
Full Bath	number of full baths above ground

Half Bath	number of half baths above ground
Bedroom AbvGr	number of bedrooms above ground
Kitchen AbvGr	number of kithchens above ground
TotRms AbvGrd	total rooms above ground
Fireplaces	number of fireplaces
Garage Cars	number of garage spaces
Garage Area	square footage of garage
Mo Sold	month house was sold
Yr Sold	year house was sold
SalePrice	sales price

We have divided the data variables as dependent and independent respectively. The data set consists of numerical values, categorical variables, and also binary variables.

Data Preparation is an important aspect in learning more about the dataset, we combined the variables as per our requirements which would be more efficient which includes combining the data as follows:

TotBsmt Fin= BsmtFin SF 1 + BsmtFin SF 2

Total Bsmt SF= Bsmt Unf SF + BsmtFinTotalArea

TotFullBath = Bsmt Full Bath + Full Bath

TotHalflBath = Bsmt Half Bath + Half Bath

TotBath= TotFullBath + TotHalflBath

Age= Year Built - Yr Sold

We also used the Mo Sold variable in order to convert it into the seasons, to find if the seasons effect the housing prices.

We have taken August to December as FALL, January to May as SPRING and the other months as SUMMER.

Thereby we formed two variables: Season(categorical variables), Season_convt (numerical values)

These provided a clear view of the data and provided us with insights to bring and form the research questions.

Dataset used for Correlation Analysis:

	A	B	C	D	E	F	G	H	I	J
1	Age	Total Bsmt SF	TotBath	TotRms AbvGrd	Kitchen AbvGr	Fireplaces	Garage Area	Gr Liv Area	Lot Area	SalePrice
2	50	1080	2	7	1	2	528	1656	31770	215000
3	49	882	1	5	1	0	730	896	11622	105000
4	52	1329	2	6	1	0	312	1329	14267	172000
5	42	2110	4	8	1	2	522	2110	11160	244000
6	13	928	3	6	1	1	482	1629	13830	189900
7	12	926	3	7	1	1	470	1604	9978	195500
8	9	1338	3	6	1	0	582	1338	4920	213500
9	18	1280	2	5	1	0	506	1280	5005	191500
10	15	1595	3	5	1	1	608	1616	5389	236500
11	11	994	3	7	1	1	442	1804	7500	189000
12	17	763	3	7	1	1	440	1655	10000	175900
13	18	1168	3	6	1	0	420	1187	7980	185000
14	12	789	3	7	1	1	393	1465	8402	180400
15	20	1300	3	5	1	1	506	1341	10176	171500
16	25	1488	3	4	1	0	528	1502	6820	212000
17	7	1650	5	12	1	1	841	3279	53504	538000
18	22	559	2	8	1	0	492	1752	12134	164000
19	0	1856	3	8	1	1	834	1856	11394	394432
20	59	864	1	4	1	0	400	864	19138	141000
21	32	1542	3	7	1	2	500	2073	13175	210000
22	33	1844	2	7	1	1	546	1844	11751	190000
23	36	1053	3	6	1	2	528	1173	10625	170000
24	10	814	4	7	1	0	663	1674	7500	216000
25	40	1004	2	5	1	1	480	1004	11241	149000
26	39	1078	3	6	1	1	500	1078	12537	149900
27	42	1056	2	6	1	1	304	1056	8450	142000

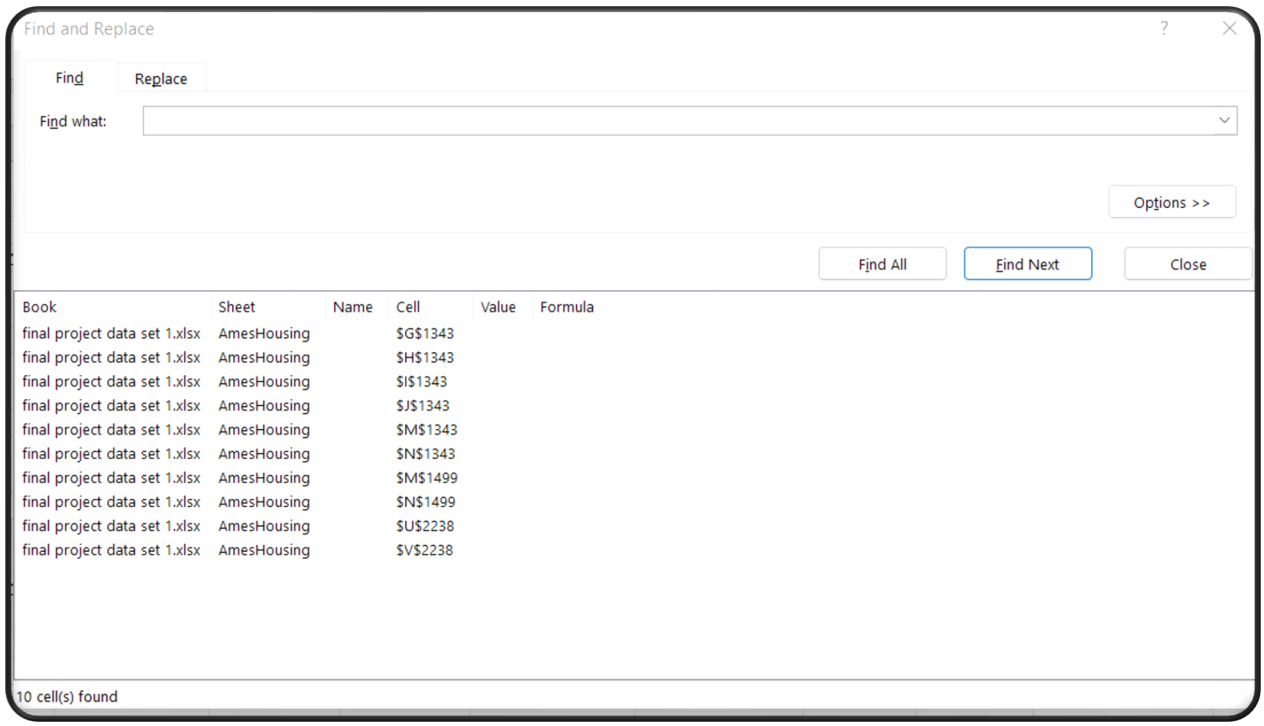
RESULTS

EXPLORATORY DATA ANALYSIS:

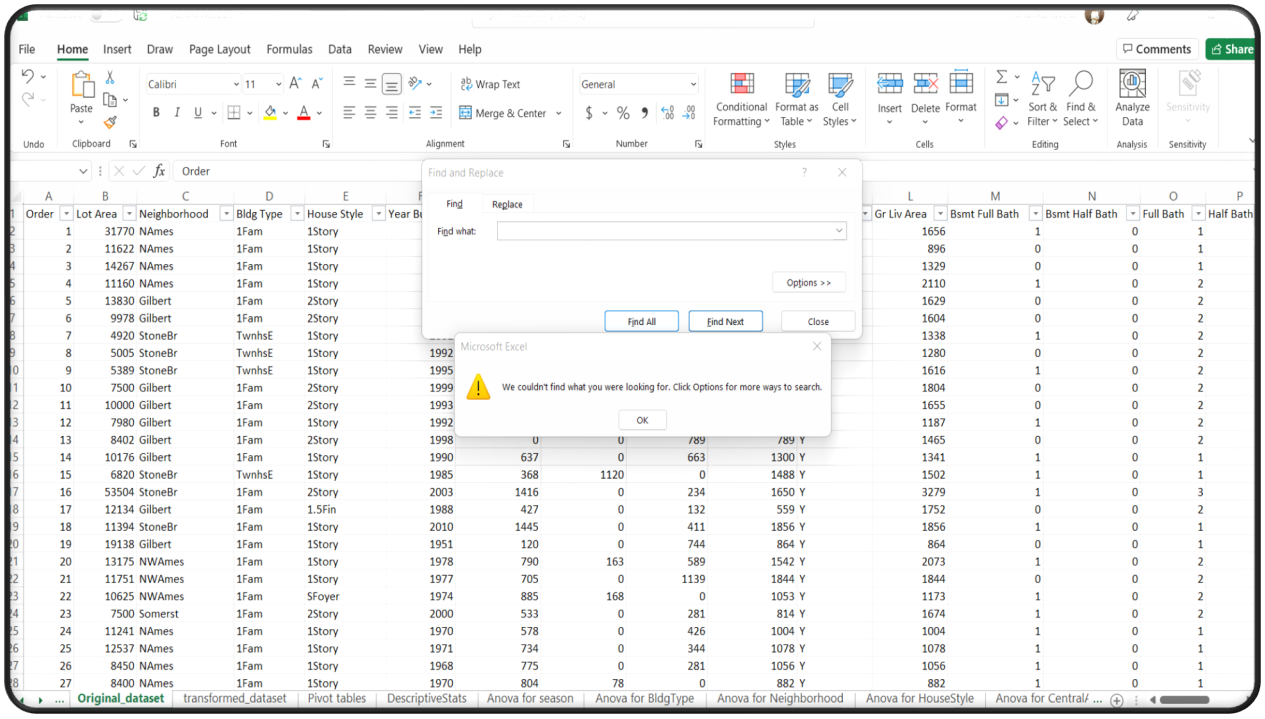
Exploratory Data Analysis is a way to find more about the data and gather insights from it. EDA gives an approach to analyze the data using the statistics, visualization techniques and more. The analysis helps us to find patterns, discover trends, or to check the assumptions by using the summary and graphical representations.

First, we check the missing values in the dataset. In our dataset there are 10 missing values. Usually, handle the missing values by using the data imputation or data interpolation. Data imputation is used when we use the mean, medina, or mode to replace the values. Interpolation is where we replace the value with the help of the values near the observation.

In our case, we used the interpolation and replaced the values based on the observations that lie near the missing values. For Example, the Unfinished Basement where there is no basement in that scenario it will be zero.



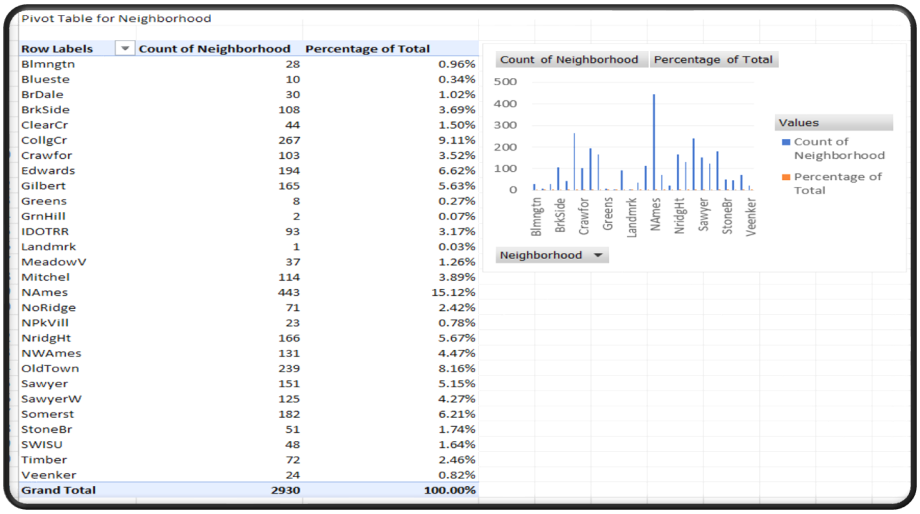
We replaced all the missing values and now that the data is clean, we can use the statistics on it to get meaningful insights that can be useful.



<u>SUMMARY STATISTICS</u>:

The descriptive summary can be done on both numerical and categorical variables. But the thing is for categorical variables we need to use the Pivot table for representation or getting the statistics. For numerical variable we can directly use the Descriptive Statistics. In our dataset there are 5 categorical variables- Neighborhood type, Building type, Season, Central air, and House style. And the rest of them fall under numerical data.

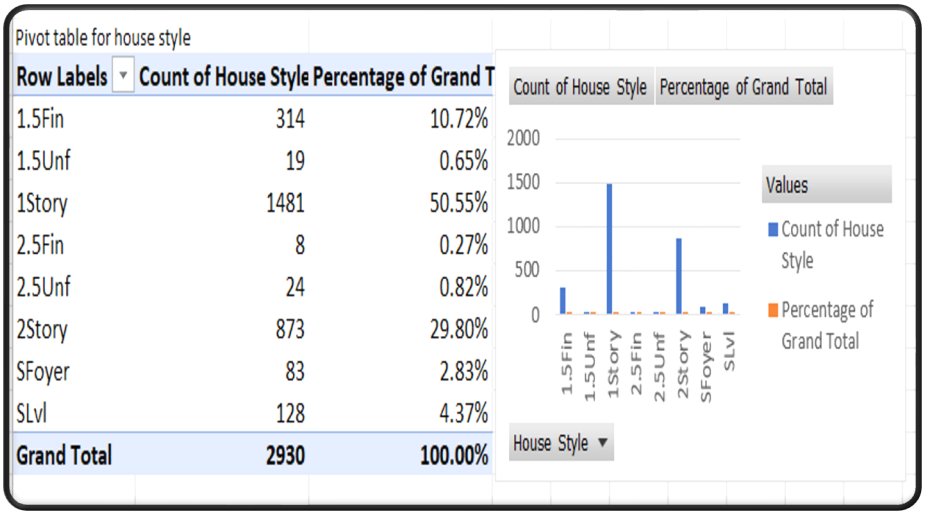
PIVOT TABLE FOR NEIGHBORHOOD TYPE:



Pivot Table for Neighborhood

Row Labels	Count of Neighborhood	Percentage of Total
Blmngtn	28	0.96%
Blueste	10	0.34%
BrDale	30	1.02%
BrkSide	108	3.69%
ClearCr	44	1.50%
CollgCr	267	9.11%
Crawfor	103	3.52%
Edwards	194	6.62%
Gilbert	165	5.63%
Greens	8	0.27%
GrnHill	2	0.07%
IDOTRR	93	3.17%
Landmrk	1	0.03%
MeadowV	37	1.26%
Mitchel	114	3.89%
NAmes	443	15.12%
NoRidge	71	2.42%
NPkVill	23	0.78%
NridgHt	166	5.67%
NWAmes	131	4.47%
OldTown	239	8.16%
Sawyer	151	5.15%
SawyerW	125	4.27%
Somerst	182	6.21%
StoneBr	51	1.74%
SWISU	48	1.64%
Timber	72	2.46%
Veenker	24	0.82%
Grand Total	**2930**	**100.00%**

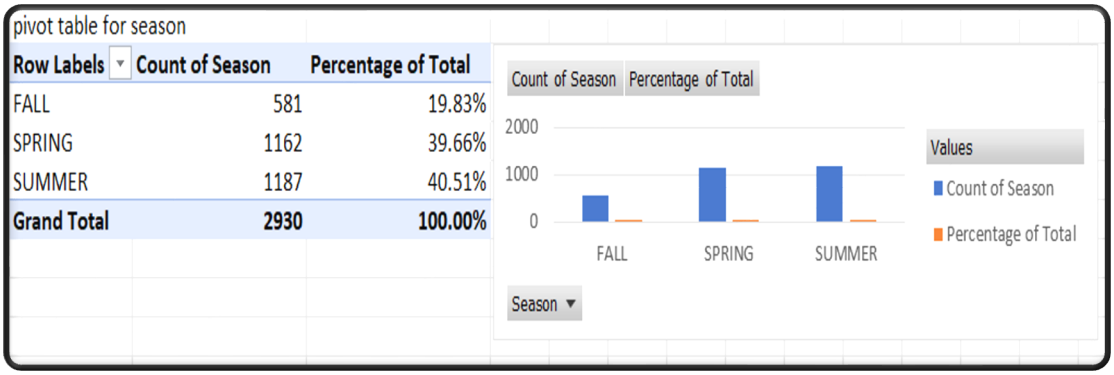
PIVOT TABLE FOR HOUSE STYLE:



Pivot table for house style

Row Labels	Count of House Style	Percentage of Grand T
1.5Fin	314	10.72%
1.5Unf	19	0.65%
1Story	1481	50.55%
2.5Fin	8	0.27%
2.5Unf	24	0.82%
2Story	873	29.80%
SFoyer	83	2.83%
SLvl	128	4.37%
Grand Total	**2930**	**100.00%**

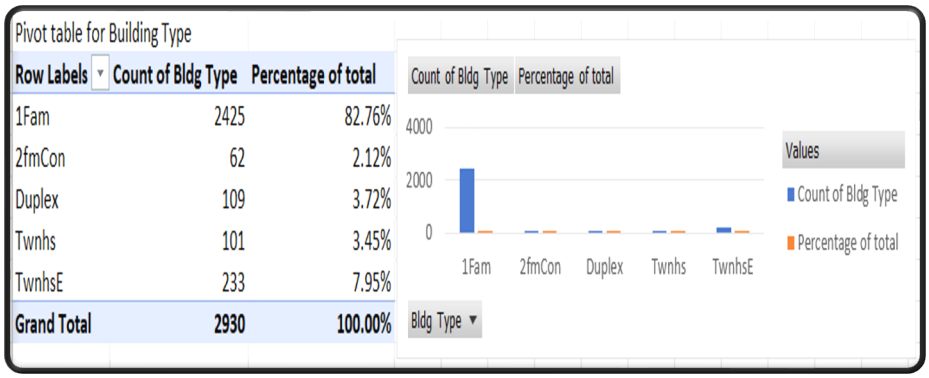
PIVOT TABLE FOR SEASON:



pivot table for season

Row Labels	Count of Season	Percentage of Total
FALL	581	19.83%
SPRING	1162	39.66%
SUMMER	1187	40.51%
Grand Total	**2930**	**100.00%**

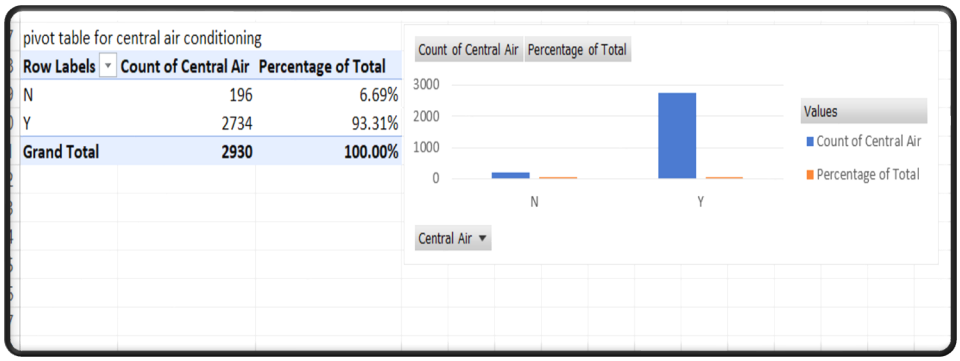
PIVOT TABLE FOR BULDING TYPE:



Pivot table for Building Type

Row Labels	Count of Bldg Type	Percentage of total
1Fam	2425	82.76%
2fmCon	62	2.12%
Duplex	109	3.72%
Twnhs	101	3.45%
TwnhsE	233	7.95%
Grand Total	**2930**	**100.00%**

PIVOT TABLE FOR CENTRAL AIR:



pivot table for central air conditioning

Row Labels	Count of Central Air	Percentage of Total
N	196	6.69%
Y	2734	93.31%
Grand Total	**2930**	**100.00%**

DESCRIPTIVE STATISTICS:

For Age:

Age	
Mean	36.43412969
Standard Error	0.559609302
Median	34
Mode	1
Standard Deviation	30.29135738
Sample Variance	917.5663322
Kurtosis	-0.493459109
Skewness	0.603324863
Range	137
Minimum	-1
Maximum	136
Sum	106752

Count	2930
Confidence Level(95.0%)	1.097267503

For Total Square Footage of Basement Finished:

TotBsmt Fin	
Mean	492.183959
Standard Error	8.818268777
Median	459.5
Mode	0
Standard Deviation	477.3282541
Sample Variance	227842.2622
Kurtosis	5.231418329
Skewness	1.169941527
Range	5644
Minimum	0
Maximum	5644
Sum	1442099
Count	2930
Confidence Level(95.0%)	17.29063425

For Unfinished Square Footage in Basement:

Bsmt Unf SF	
Mean	559.0716724
Standard Error	8.12017068
Median	465.5
Mode	0
Standard Deviation	439.5405711
Sample Variance	193195.9136
Kurtosis	0.409419144
Skewness	0.923044978
Range	2336
Minimum	0
Maximum	2336
Sum	1638080
Count	2930
Confidence Level(95.0%)	15.92182149

For Total Square Footage in Basement:

Total Bsmt SF	
Mean	1051.255631
Standard Error	8.146541648
Median	990
Mode	0
Standard Deviation	440.9680177
Sample Variance	194452.7926
Kurtosis	9.107470636
Skewness	1.150845954
Range	6110
Minimum	0
Maximum	6110
Sum	3080179
Count	2930
Confidence Level(95.0%)	15.973529

For Gross Living Area:

Gr Liv Area	
Mean	1499.690444
Standard Error	9.338884092
Median	1442
Mode	864
Standard Deviation	505.5088875
Sample Variance	255539.2353
Kurtosis	4.137838193
Skewness	1.274109716
Range	5308
Minimum	334
Maximum	5642
Sum	4394093
Count	2930
Confidence Level(95.0%)	18.31144335

For Total number of Full Baths:

TotFullBath	
Mean	1.997610922
Standard Error	0.013920359
Median	2
Mode	2
Standard Deviation	0.753501732
Sample Variance	0.567764861
Kurtosis	0.614136564
Skewness	0.449542073
Range	6
Minimum	0
Maximum	6
Sum	5853
Count	2930
Confidence Level(95.0%)	0.027294682

For Total number of Half Baths:

TotHalfBath	
Mean	0.440614334
Standard Error	0.010092614
Median	0
Mode	0
Standard Deviation	0.546307914
Sample Variance	0.298452337
Kurtosis	0.175311329
Skewness	0.802225175
Range	4
Minimum	0
Maximum	4
Sum	1291
Count	2930
Confidence Level(95.0%)	0.019789338

For Total number of Baths in the House:

TotBath	
Mean	2.438225256
Standard Error	0.017383389
Median	2

Mode	2
Standard Deviation	0.940953725
Sample Variance	0.885393913
Kurtosis	0.869505601
Skewness	0.501140803
Range	7
Minimum	1
Maximum	8
Sum	7144
Count	2930
Confidence Level(95.0%)	0.034084902

For Number of Bedrooms above ground:

Bedroom AbvGr	
Mean	2.854266212
Standard Error	0.01529169
Median	3
Mode	3
Standard Deviation	0.827731142
Sample Variance	0.685138843
Kurtosis	1.891420659
Skewness	0.305694211
Range	8
Minimum	0
Maximum	8
Sum	8363
Count	2930
Confidence Level(95.0%)	0.029983552

For number of Kitchens above ground:

Kitchen AbvGr	
Mean	1.044368601
Standard Error	0.003954892
Median	1
Mode	1
Standard Deviation	0.214076244
Sample Variance	0.045828638

Kurtosis	19.86974309
Skewness	4.313824595
Range	3
Minimum	0
Maximum	3
Sum	3060
Count	2930
Confidence Level(95.0%)	0.007754651

For Total number of rooms above ground:

TotRms AbvGrd	
Mean	6.443003413
Standard Error	0.029059296
Median	6
Mode	6
Standard Deviation	1.572964396
Sample Variance	2.474216992
Kurtosis	1.154588179
Skewness	0.753542562
Range	13
Minimum	2
Maximum	15
Sum	18878
Count	2930
Confidence Level(95.0%)	0.056978718

For Number of Fireplaces in house:

Fireplaces	
Mean	0.599317406
Standard Error	0.011969836
Median	1
Mode	0
Standard Deviation	0.647920917
Sample Variance	0.419801514
Kurtosis	0.101508478
Skewness	0.739215201
Range	4

Minimum	0
Maximum	4
Sum	1756
Count	2930
Confidence Level(95.0%)	0.023470146

For Garage Cars:

Garage Cars	
Mean	1.766211604
Standard Error	0.01406141
Median	2
Mode	2
Standard Deviation	0.761136719
Sample Variance	0.579329105
Kurtosis	0.243365413
Skewness	-0.221162519
Range	5
Minimum	0
Maximum	5
Sum	5175
Count	2930
Confidence Level(95.0%)	0.02757125

For Garage Area:

Garage Area	
Mean	472.6583618
Standard Error	3.975416315
Median	480
Mode	0
Standard Deviation	215.1871957
Sample Variance	46305.5292
Kurtosis	0.94803921
Skewness	0.240064589
Range	1488
Minimum	0
Maximum	1488
Sum	1384889

Count	2930
Confidence Level(95.0%)	7.7948939

For Lot Area:

Lot Area	
Mean	10147.92184
Standard Error	145.5772081
Median	9436.5
Mode	9600
Standard Deviation	7880.017759
Sample Variance	62094679.89
Kurtosis	265.0236706
Skewness	12.82089817
Range	213945
Minimum	1300
Maximum	215245
Sum	29733411
Count	2930
Confidence Level(95.0%)	285.4440394

For Sales Price:

SalePrice	
Mean	180796.0601
Standard Error	1475.844597
Median	160000
Mode	135000
Standard Deviation	79886.69236
Sample Variance	6381883616
Kurtosis	5.118899951
Skewness	1.743500076
Range	742211
Minimum	12789
Maximum	755000

Sum	529732456
Count	2930
Confidence Level(95.0%)	2893.798067

ANOVA:

Analysis of variance (ANOVA) is used to check if the means of two or more groups are significantly different from each other. ANOVA can be done on more than two features in our dataset we perform ANOVA for categorical variables with reference to the most common target variable the sales price of the household.

- ANOVA for Season:

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
FALL	581	108007119	185898.6558	6803974554
SPRING	1162	203875787	175452.4845	5998301807
SUMMER	1187	217849550	183529.5282	6513514060

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	57175795604	2	28587897802	4.490214891	0.011295615	2.998800445
Within Groups	1.86354E+13	2927	6366710391			
Total	1.86925E+13	2929				

From the observations, clearly F calculated > F critical and p value < 0.05(level of significance). As a result we reject the null hypothesis, there is enough evidence to say that there is difference in variability in SalePrice of houses depending on the variable Season i.e SalePrice varies based on the academic seasons such as SPRING,FALL and SUMMER.

- ANOVA for Building Type:

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
SingleFam	2425	448169200	184812.0412	6859450941
Condo	62	7786066	125581.7097	966540833.8
Duplex	109	15239174	139808.9358	1560168910
TownHouse	101	13729340	135934.0594	1758873944
TownHouseEndUnit	233	44808676	192311.9142	4381346182

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	6.45411E+11	4	1.61353E+11	26.15135823	2.47592E-21	2.374970388
Within Groups	1.80471E+13	2925	6169957615			
Total	1.86925E+13	2929				

From the observations, clearly F calculated > F critical and p value < 0.05(level of significance). As a result we reject the null hypothesis, there is enough evidence to say that there is difference in variability in SalePrice of houses depending on the variable Building type, i.e SalePrice varies based on the building types such as SingleFam,Condo,Duplex,TownHouse,TownHouseEndUnit.

- ANOVA for Central Air:

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
N	196	19970534	101890.4796	1413536271
Y	2734	509761922	186452.7879	6260190152

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	1.3078E+12	1	1.3078E+12	220.2639944	4.24747E-48	3.844636702
Within Groups	1.73847E+13	2928	5937410949			
Total	1.86925E+13	2929				

From the observations, clearly F calculated > F critical and p value < 0.05(level of significance). As a result we reject the null hypothesis, there is enough evidence to say that there is difference in variability in SalePrice of houses depending on the variable Central Air, i.e SalePrice varies based on the whether the house has Central Air or not.

CORRELATION:

Dataset used for Correlation Analysis is shown in the Data Preparation, using that:

CORRELATION MATRIX FOR NUMERICAL VARIABLES:

	Age	Total Bsmt SF	TotBath	TotRms AbvGrd	Kitchen AbvGr	Fireplaces	Garage Area	Gr Liv Area	Lot Area	SalePrice
Age	1									
Total Bsmt SF	-0.40760251	1								
TotBath	-0.528437848	0.34773047	1							
TotRms AbvGrd	-0.113074264	0.281626682	0.458847472	1						
Kitchen AbvGr	0.13918201	-0.038612068	0.050899039	0.294444985	1					
Fireplaces	-0.170743774	0.333466654	0.34914869	0.302865275	-0.108085162	1				
Garage Area	-0.480541316	0.485607512	0.431957071	0.3272172	-0.057575506	0.294262112	1			
Gr Liv Area	-0.242510071	0.4451076	0.622862146	0.807772145	0.117835671	0.454924196	0.483970507	1		
Lot Area	-0.024226571	0.253764779	0.170007313	0.216596645	-0.020300505	0.256989114	0.212749106	0.285599214	1	
SalePrice	-0.558906832	0.632528849	0.617377818	0.495474417	-0.11981372	0.474558093	0.640138298	0.706779921	0.26654922	1

The variables Total Bsmt SF and Age, TotBath and Age, TotRms AbvGrd and Age, Kitchen AbvGr and Total Bsmt SF, Fireplaces and Age, Fireplaces and Kitchen AbvGr, Garage Area and Age, Garage Area and Kitchen AbvGr, Gr Liv Area and Age, Lot Area and Age, Lot Area and Kitchen AbvGr, SalePrice and Age, SalePrice and Kitchen AbvGr are negatively correlated.

The variables TotBath and Total Bsmt SF, TotRms AbvGrd and Total Bsmt SF, TotRms AbvGrd and TotBath, Kitchen AbvGr and Age, Kitchen AbvGr and TotBath, Kitchen AbvGr and TotRms AbvGrd, Fireplaces and Total Bsmt SF, Fireplaces and TotBath, Fireplaces and TotRms AbvGrd, Garage Area and Total Bsmt SF, Garage Area and TotBath, Garage Area and TotRms AbvGrd, Garage Area and Fireplaces, Gr Liv Area and Total Bsmt SF, Gr Liv Area and TotBath, Gr Liv Area and TotRms AbvGrd, Gr Liv Area and Kitchen AbvGr, Gr Liv Area and Fireplaces, Gr Liv Area and Garage Area, Lot Area and Total Bsmt SF, Lot Area and TotBath, Lot Area and TotRms AbvGrd, Lot Area and , Lot Area and Fireplaces, Lot Area and Garage Area,

Lot Area and Gr Liv Area, SalePrice and Total Bsmt SF, SalePrice and TotBath, SalePrice and TotRms AbvGrd, SalePrice and Fireplaces, SalePrice and Garage Area, SalePrice and Gr Liv Area, SalePrice and Lot Area are positively correlated.

Gr Liv Area and TotRms AbvGrd being the most highly correlated ones.

REGRESSION MODEL:

The dataset to perform Regression Analysis:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Total Bsmt SF	Bedroom AbvGr	Kitchen AbvGr	season_convt	Age	TotRms AbvGrd	Gr Liv Area	TotBath	Fireplaces	Garage Area	Lot Area	Bsmt Unf SF	SalePrice
2	1080	3	1	0	50	7	1656	2	2	528	31770	441	215000
3	882	2	1	1	49	5	896	1	0	730	11622	270	105000
4	1329	3	1	1	52	6	1329	2	0	312	14267	406	172000
5	2110	3	1	0	42	8	2110	4	2	522	11160	1045	244000
6	928	3	1	0	13	6	1629	3	1	482	13830	137	189900
7	926	3	1	1	12	7	1604	3	1	470	9978	324	195500
8	1338	2	1	0	9	6	1338	3	0	582	4920	722	213500
9	1280	2	1	0	18	5	1280	2	0	506	5005	1017	191500
10	1595	2	1	0	15	5	1616	3	1	608	5389	415	236500
11	994	3	1	1	11	7	1804	3	1	442	7500	994	189000
12	763	3	1	0	17	7	1655	3	1	440	10000	763	175900
13	1168	3	1	0	18	6	1187	3	0	420	7980	233	185000
14	789	3	1	0	12	7	1465	3	1	393	8402	789	180400
15	1300	2	1	0	20	5	1341	3	1	506	10176	663	171500
16	1488	1	1	1	25	4	1502	3	0	528	6820	0	212000
17	1650	4	1	1	7	12	3279	5	1	841	53504	234	538000
18	559	4	1	1	22	8	1752	2	0	492	12134	132	164000
19	1856	1	1	1	0	8	1856	3	1	834	11394	411	394432
20	864	2	1	1	59	4	864	1	0	400	19138	744	141000
21	1542	3	1	0	32	7	2073	3	2	500	13175	589	210000
22	1844	3	1	0	33	7	1844	2	1	546	11751	1139	190000
23	1053	3	1	0	36	6	1173	3	2	528	10625	0	170000
24	814	3	1	0	10	7	1674	4	0	663	7500	281	216000
25	1004	2	1	0	40	5	1004	2	1	480	11241	426	149000
26	1078	3	1	0	39	6	1078	3	1	500	12537	344	149900
27	1056	3	1	1	42	6	1056	2	1	304	8450	281	142000

Regression Model 1:

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.877319028							
5	R Square	0.769688677							
6	Adjusted R Square	0.768741219							
7	Standard Error	38417.00959							
8	Observations	2930							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	12	1.43874E+13	1.19895E+12	812.372084	0			
13	Residual	2917	4.3051E+12	1475866626					
14	Total	2929	1.86925E+13						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	65714.23546	4981.885879	13.19063444	1.2405E-38	55945.86536	75482.6056	55945.8654	75482.6056
18	Total Bsmt SF	41.96285832	2.216791773	18.92954441	1.8174E-75	37.61622272	46.3094939	37.6162227	46.3094939
19	Bedroom AbvGr	-13953.25996	1219.635142	-11.44051978	1.1155E-29	-16344.69319	-11561.8267	-16344.6932	-11561.8267
20	Kitchen AbvGr	-41560.7781	3647.921799	-11.39300138	1.8824E-29	-48713.54136	-34408.0149	-48713.5414	-34408.0149
21	season_convt	1210.265338	955.5883773	1.266513247	0.20543051	-663.4309215	3083.9616	-663.430921	3083.9616
22	Age	-587.1949692	32.54938759	-18.04012341	4.6684E-69	-651.0170784	-523.37286	-651.017078	-523.37286
23	TotRms AbvGrd	5805.608208	929.0085444	6.249251682	4.7257E-10	3984.029089	7627.18733	3984.02909	7627.18733
24	Gr Liv Area	65.73051588	3.111680213	21.12380174	2.7876E-92	59.62920311	71.8318287	59.6292031	71.8318287
25	TotBath	4625.114722	1245.520835	3.713398115	0.00020832	2182.925401	7067.30404	2182.9254	7067.30404
26	Fireplaces	8254.992331	1307.520539	6.313470484	3.1437E-10	5691.235381	10818.7493	5691.23538	10818.7493
27	Garage Area	54.56253621	4.346885495	12.5520988	3.0989E-35	46.03926061	63.0858118	46.0392606	63.0858118
28	Lot Area	0.198547467	0.097297864	2.040614856	0.04137898	0.007767997	0.38932694	0.007768	0.38932694
29	Bsmt Unf SF	-9.635279104	2.046381164	-4.708447904	2.6126E-06	-13.6477774	-5.62278081	-13.6477774	-5.62278081

We first used the independent variables, Total Bsmt SF, Bedroom AbvGr, Kitchen AbvGr, season_convt, Age, TotRms AbvGrd, Gr Liv Area, TotBath, Fireplaces, Garage Area, Lot Area, Bsmt Unf SF where season_convt was not a statistically significant variable as P value > 0.05(level of significance) and t calculated fell in the interval of t critical. The R square value is 76.9% for this model and the F calculated is greater than Fcritical.

The regression equation is as follows:

sales_price = 65714.2354623974 + (B1*Total Bsmt) + (B2*BedroomAbvGr) + (B3*KitchenAbvGr) + (B4*Season_convt) + (B5*Age) + (B6*TotRmsAbvGround) + (B7*GRLivArea) + (B8*TotBath) + (B9*Fireplaces) + (B10*Garage Area) + (B11*LotArea) + (B12*BsmtUnfSF)

So, we did further analysis to get a better regression model.

Regression Model 2:

Here we removed the season_convt variable as it was not statistically significant thereby forming the best regression model with all the variables which are statistically significant and R square of 76.9% as shown below:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.877246846
R Square	0.769562029
Adjusted R Square	0.768693346
Standard Error	38420.98574
Observations	2930

ANOVA

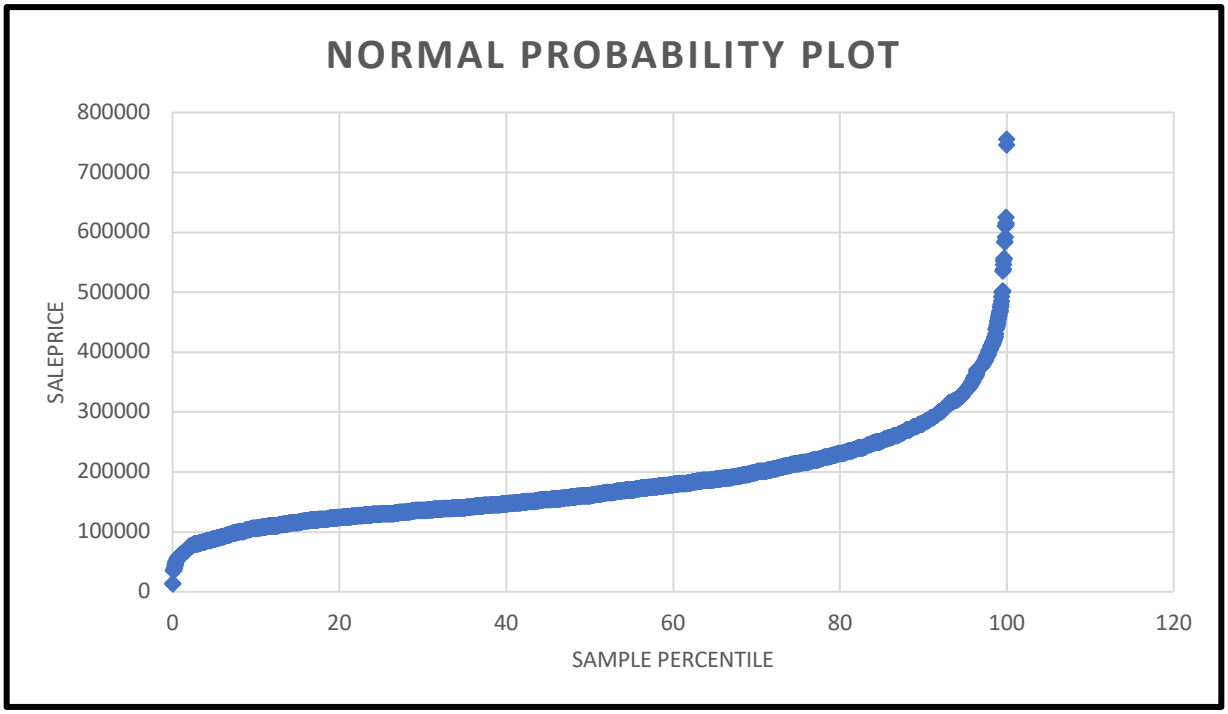
	df	*SS*	*MS*	*F*	*Significance F*
Regression	11	1.43851E+13	1.30773E+12	885.894879	0
Residual	2918	4.30747E+12	1476172145		
Total	2929	1.86925E+13			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	66362.11008	4956.066528	13.39007653	1.004E-39	56644.36736	76079.8528	56644.3674	76079.8528
Age	-587.701553	32.5502985	-18.05518167	3.6433E-69	-651.5254392	-523.877667	-651.525439	-523.877667
Total Bsmt SF	41.94590416	2.216980787	18.92028311	2.1185E-75	37.59889857	46.2929097	37.5988986	46.2929097
Bsmt Unf SF	-9.587688793	2.046247914	-4.685497162	2.9207E-06	-13.59992524	-5.57545234	-13.5999252	-5.57545234
Gr Liv Area	65.78672227	3.111685756	21.14182711	1.9921E-92	59.68539949	71.8880451	59.6853995	71.8880451
TotBath	4622.494287	1245.648028	3.71091527	0.00021037	2180.055917	7064.93266	2180.05592	7064.93266
Bedroom AbvGr	-13898.85159	1219.004485	-11.40180512	1.708E-29	-16289.0479	-11508.6553	-16289.0479	-11508.6553
Kitchen AbvGr	-41388.38354	3645.758687	-11.35247478	2.9353E-29	-48536.90439	-34239.8627	-48536.9044	-34239.8627
TotRms AbvGrd	5779.545924	928.876739	6.222080586	5.6084E-10	3958.225505	7600.86634	3958.2255	7600.86634
Fireplaces	8290.204416	1307.360216	6.34117844	2.6334E-10	5726.762188	10853.6466	5726.76219	10853.6466
Garage Area	54.68600715	4.346241941	12.58236607	2.1538E-35	46.16399463	63.2080197	46.1639946	63.2080197
Lot Area	0.199265922	0.097306281	2.047821793	0.0406668	0.008469976	0.39006187	0.00846998	0.39006187

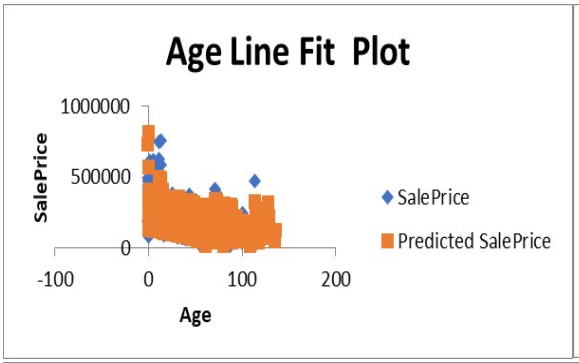
The regression equation is as follows:

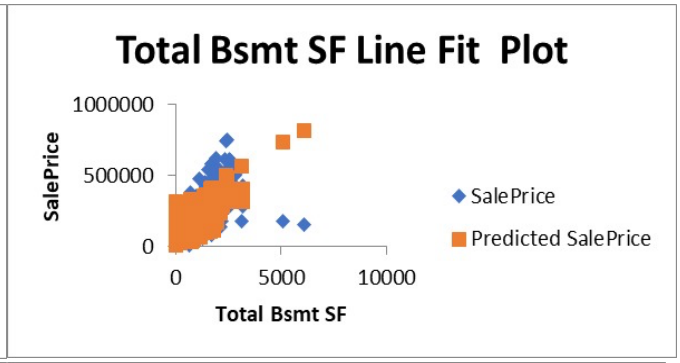
sales_price = 66362.110082619 + (B1*Age) + (B2*Total Bsmt SF) + (B3*BsmtUnfSF) + (B4*GR Liv Area) + (B5*TotBath) + (B6*Bedroom Abvgr) + (B7*Kitchen Abvgr) + (B8*TotRmsAbvgr) + (B9*Fireplaces) + (B10*GarageArea) + (B11*LotArea)

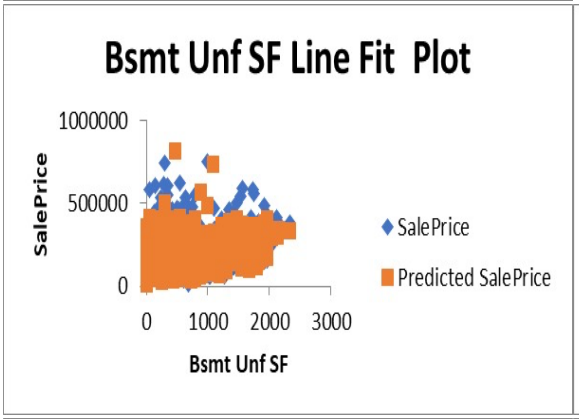
The normal probability plot is as shown:

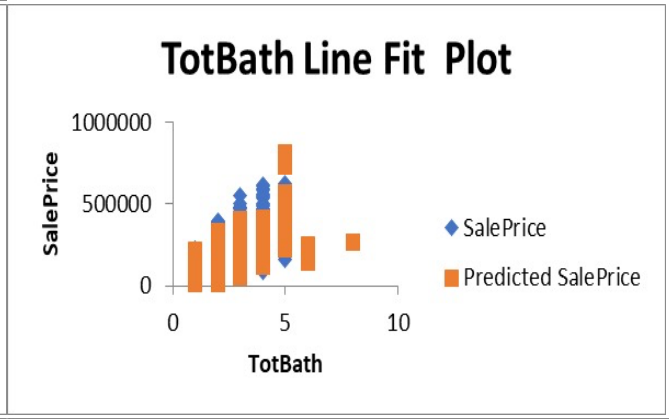


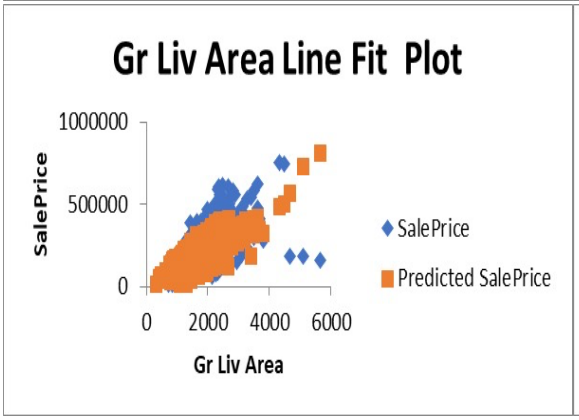
The plots for statistically significant data are as shown:

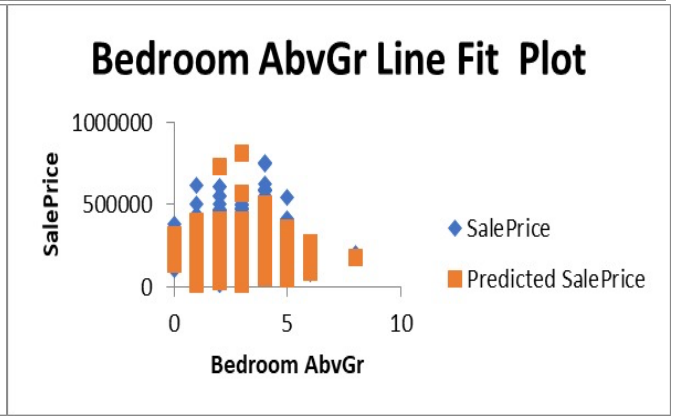


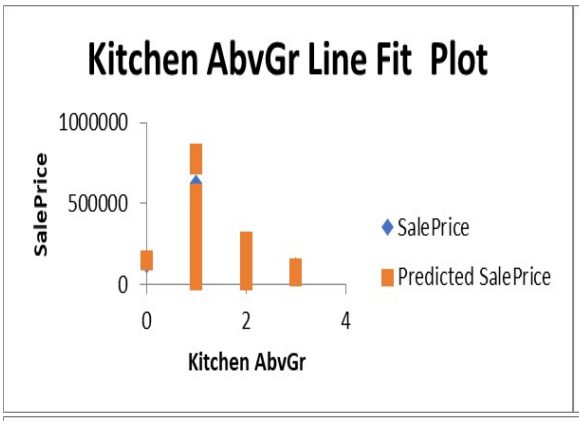


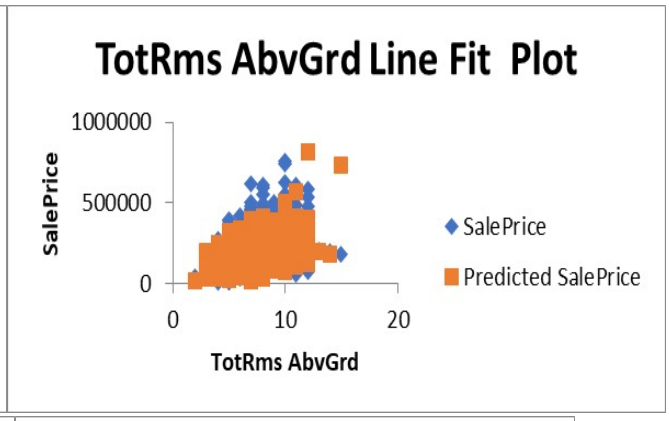


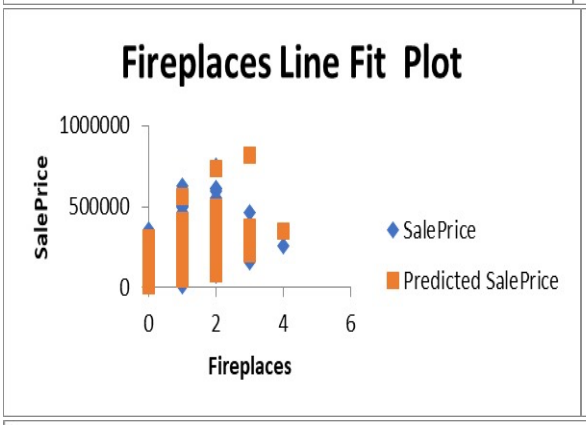


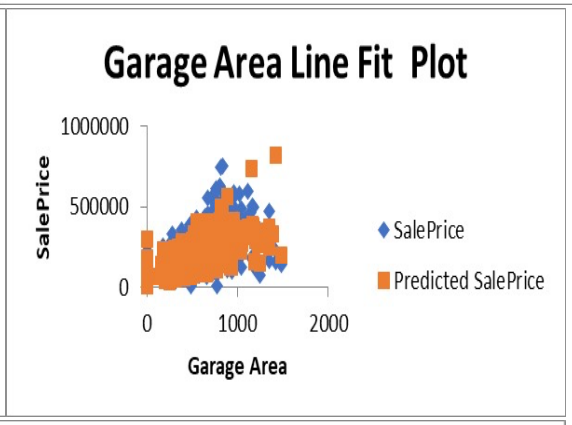


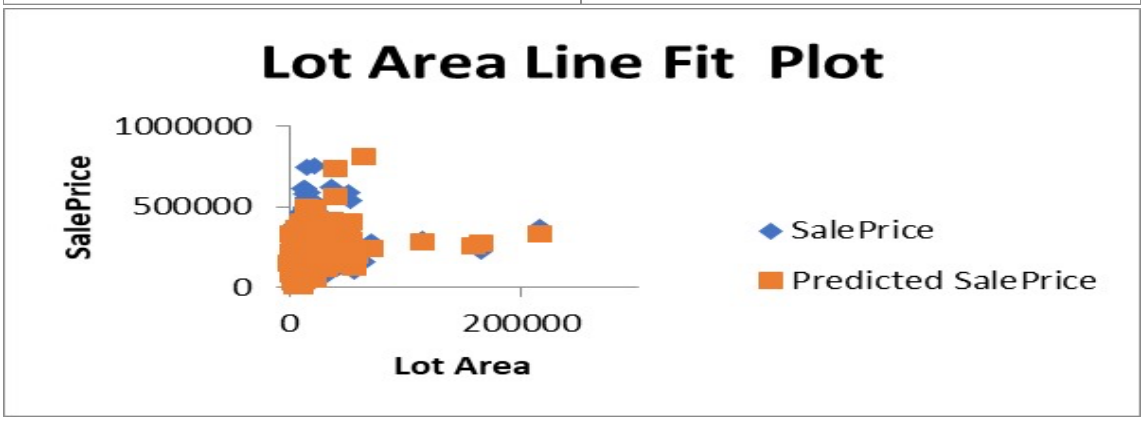



Kitchen AbvGr Line Fit Plot
1000000
500000
0
SalePrice
0
2
4
Kitchen AbvGr
SalePrice
Predicted SalePrice


TotRms AbvGrd Line Fit Plot
1000000
500000
0
SalePrice
0
10
20
TotRms AbvGrd
SalePrice
Predicted SalePrice


Fireplaces Line Fit Plot
1000000
500000
0
SalePrice
0
2
4
6
Fireplaces
SalePrice
Predicted SalePrice


Garage Area Line Fit Plot
1000000
500000
0
SalePrice
0
1000
2000
Garage Area
SalePrice
Predicted SalePrice


Lot Area Line Fit Plot
1000000
500000
0
SalePrice
0
200000
Lot Area
SalePrice
Predicted SalePrice

DISCUSSION

The Season which we classified as SPRING,FALL and SUMMER did have an effect on the SalePrice as we showed in the ANOVA(at 0.05 level of significance), thereby concluding that the academic season does play a role on the decision of Buyers and Sellers. Coming to the type of building, the building type considering the 0.05 level of significance determines the SalePrice and impacts the decision making of the Buyers and Sellers. The Central Air also impacts the price of the house. The Correlation has been performed on the desired variables; some had a positive correlation while others had a negative correlation. The correlation matrix also shows the least correlated, moderately correlated, and highly correlated datasets.

The regression model is obtained in such a way that all the variables are statistically significant with p value < 0.05 and T calculated does not fall in the interval of T critical. Thereby we obtained a model with R square of 76.9% which shows it's a good fit.

LIMITATIONS

We have got the regression model which is what we desired which would indeed provide guidance to Buyers and Sellers. We have seen the independent variables which are not statistically significant and thereby ignored them, this is the case where a sample data set is taken into consideration which may not completely bring out the best of the real correlations between the data. Thereby, accuracy of data with such concentrated variables is the one that needs to be taken care of.

RECOMMENDATIONS

The dataset can be transformed into getting more insights by utilizing more sophisticated tools and methodologies. We have used Excel in our case which helped us to bring out the model with reduced complications, which also helped us to get the R value of 76.9%, but we can't just ignore the rest 23.1% when trying to bring out the results which would help the Buyers and Sellers in reality. Also, several variables can be combined in different ways to form several research questions which may have more impact on the SalePrice and more work needs to be done to play with data which can bring surprising results.