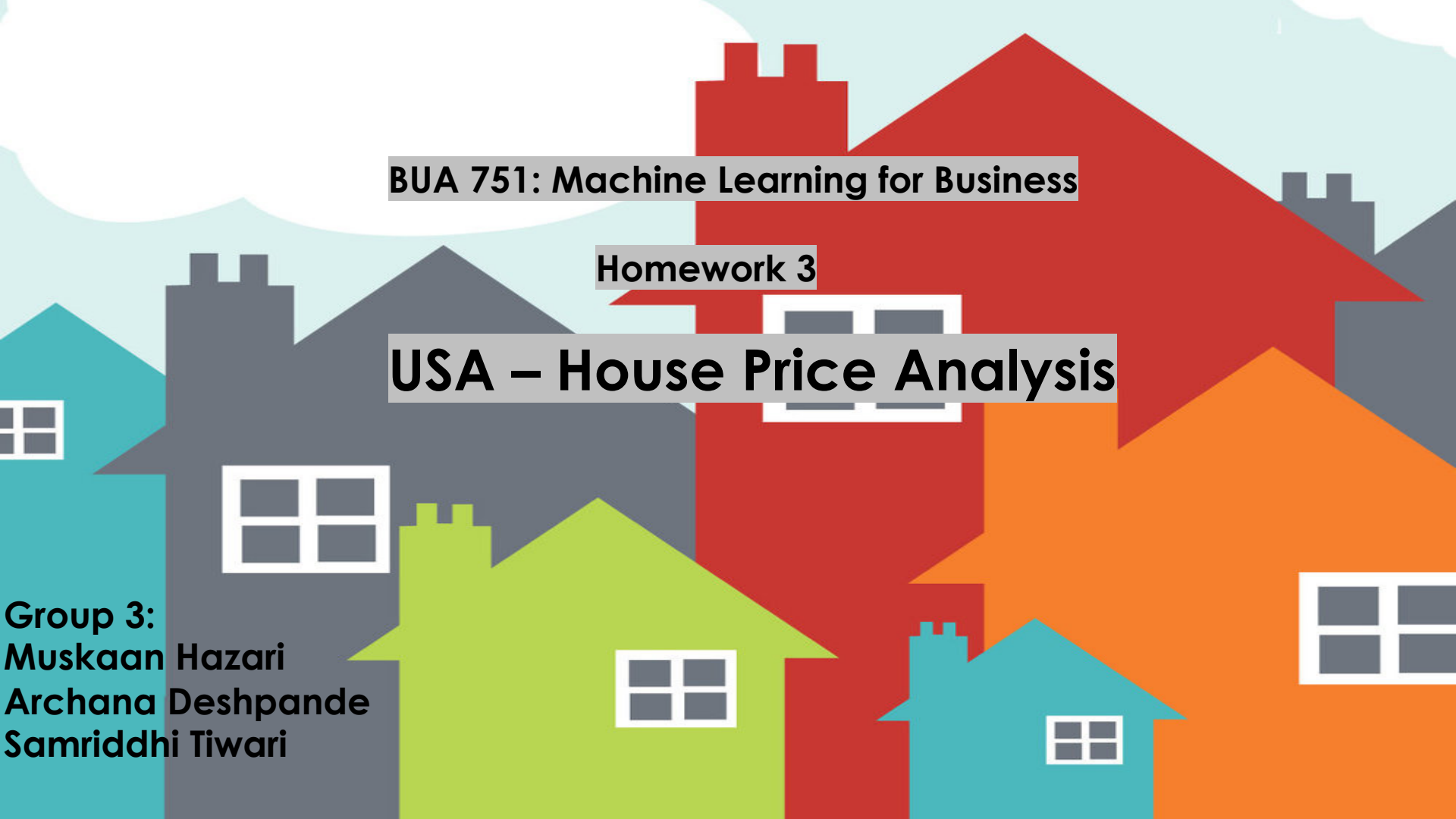


BUA 751: Machine Learning for Business

Homework 3

USA – House Price Analysis

Group 3:
Muskaan Hazari
Archana Deshpande
Samriddhi Tiwari

ABOUT THE DATASET

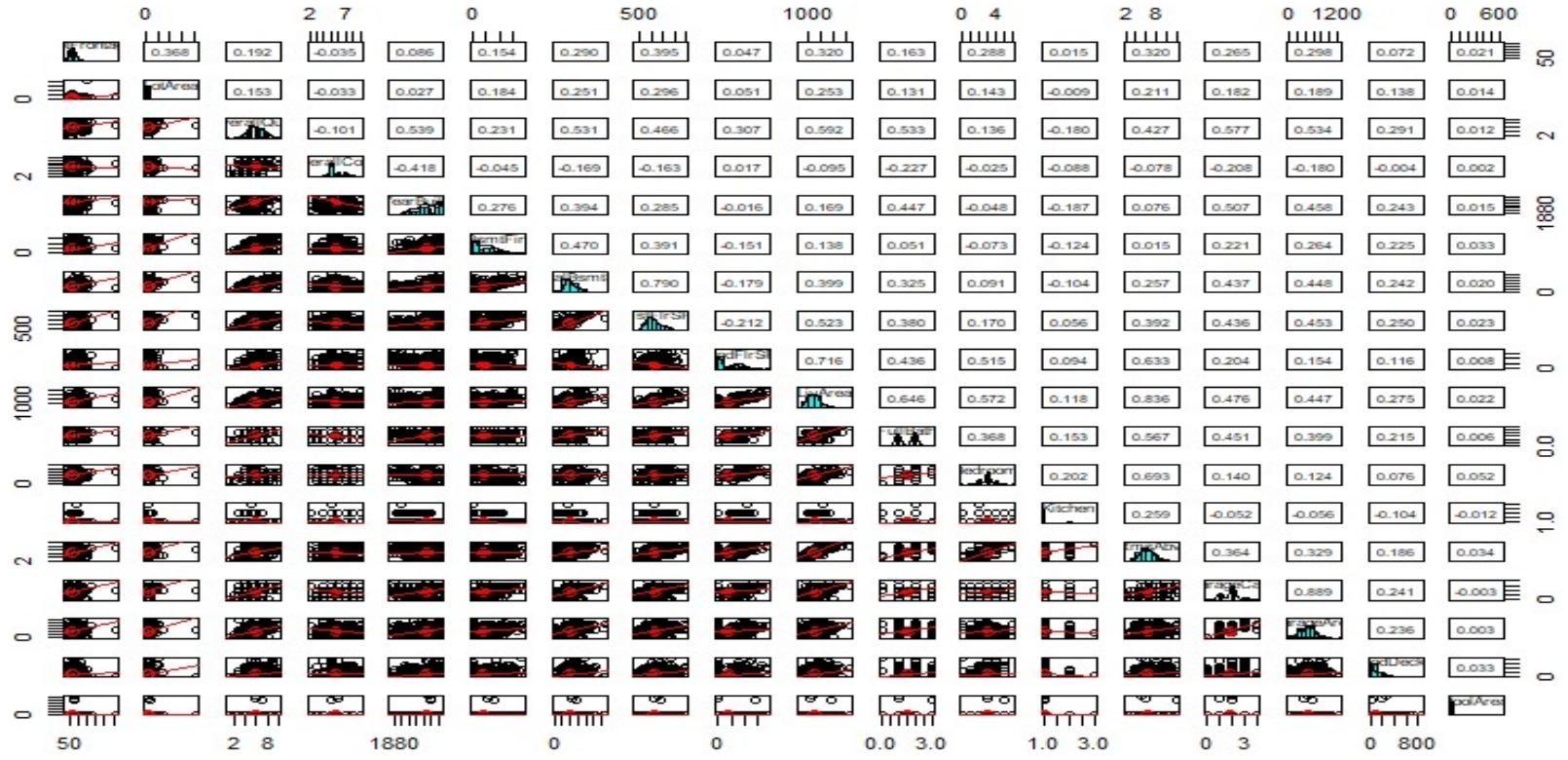
The U.S. housing prices dataset used for analysis contains information on various factors that can impact the price of a house, such as the size of the house, number of bedrooms and bathrooms, location, and more.

Here are the variables included in the dataset:

Housing dataset

SalePrice	final sale price
LotFrontage	width of lot on street
LotArea	square feet of lot
OverallQual	quality of house on scale of 1 to 10
OverallCond	condition of house on scale of 1 to 10
YearBuilt	calendar year of construction
BsmtFin	square feet of finished basement
TotalBsmtSF	total square feet of basement (finished and unfinished)
1stFlrSF	square feet of first floor
2ndFlrSF	square feet of second floor
LivArea	square feet of living area
BsmtFullBath	number of full bathrooms in basement
BsmtHalfBath	number of half bathrooms in basement
FullBath	number of full bathrooms above ground
HalfBath	number of half bathrooms above ground
Bedroom	number of bedrooms
Kitchen	number of kitchens
TotRmsAbvGrd	number of rooms above ground
Fireplaces	number of fireplaces
GarageCars	number of garage spaces for cars
GarageArea	square feet of garage
WoodDeckSF	square feet of wood deck
PoolArea	square feet of pool area

Correlation



Correlation

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1		SalePrice	otFrontage	LotArea	OverallQua	OverallConc	YearBuilt	BsmtFin	TotalBsmtSI	1stFlrSF	2ndFlrSF	LivArea	FullBath	Bedroom	Kitchen	tRmsAbvG	GarageCars	GarageAred	VoodDeckS
2	SalePrice	1																	
3	LotFrontag	0.343363	1																
4	LotArea	0.34521	0.367734	1															
5	OverallQu	0.801069	0.191722	0.152832	1														
6	OverallCor	-0.08136	-0.03498	-0.03325	-0.10117	1													
7	YearBuilt	0.511737	0.086449	0.026779	0.539037	-0.41798	1												
8	BsmtFin	0.429429	0.153806	0.183689	0.231148	-0.04477	0.276224	1											
9	TotalBsmt	0.648682	0.290482	0.2514	0.53112	-0.1689	0.393673	0.47019	1										
10	1stFlrSF	0.624208	0.395408	0.295664	0.46627	-0.16294	0.284757	0.39061	0.789543	1									
11	2ndFlrSF	0.347302	0.047043	0.050818	0.306713	0.017077	-0.01647	-0.1507	-0.17942	-0.21229	1								
12	LivArea	0.738919	0.319991	0.253237	0.591676	-0.09493	0.169474	0.137864	0.399018	0.522782	0.716377	1							
13	FullBath	0.567377	0.162802	0.130751	0.532791	-0.22733	0.446591	0.050577	0.325237	0.379697	0.436268	0.645575	1						
14	Bedroom	0.224133	0.28763	0.143224	0.135581	-0.02451	-0.04777	-0.07297	0.091221	0.170375	0.514563	0.57189	0.368369	1					
15	Kitchen	-0.14775	0.014954	-0.00895	-0.18046	-0.08806	-0.18708	-0.12425	-0.10428	0.056288	0.094029	0.117925	0.153321	0.20248	1				
16	TotRmsAb	0.532797	0.320111	0.211015	0.42739	-0.07817	0.076199	0.015176	0.256784	0.392228	0.632598	0.835742	0.56746	0.693496	0.259405	1			
17	GarageCar	0.635001	0.265366	0.181553	0.577099	-0.20805	0.507223	0.221137	0.436838	0.43599	0.203758	0.476201	0.451453	0.140019	-0.05217	0.364097	1		
18	GarageAre	0.616705	0.298246	0.189024	0.5345	-0.17958	0.458456	0.263659	0.447885	0.453221	0.153807	0.447201	0.399227	0.124458	-0.05631	0.328623	0.888998	1	
19	WoodDecl	0.379098	0.072436	0.137573	0.291494	-0.00414	0.243028	0.225115	0.241763	0.250303	0.11583	0.274614	0.215404	0.076375	-0.10388	0.185946	0.240762	0.235517	1
20	PoolArea	0.019033	0.021363	0.013861	0.011948	0.002264	0.015353	0.032505	0.019806	0.022966	0.007999	0.022386	0.00645	0.051763	-0.01207	0.034065	-0.00312	0.002819	0.033478

	A	B
1		*SalePrice*
2	SalePrice	1
3	LotFrontag	0.343363
4	LotArea	0.34521
5	OverallQua	0.801069
6	OverallCor	-0.08136
7	YearBuilt	0.511737
8	BsmtFin	0.429429
9	TotalBsmt	0.648682
10	1stFlrSF	0.624208
11	2ndFlrSF	0.347302
12	LivArea	0.738919
13	FullBath	0.567377
14	Bedroom	0.224133
15	Kitchen	-0.14775
16	TotRmsAb	0.532797
17	GarageCar	0.635001
18	GarageAre	0.616705
19	WoodDecl	0.379098
20	PoolArea	0.019033

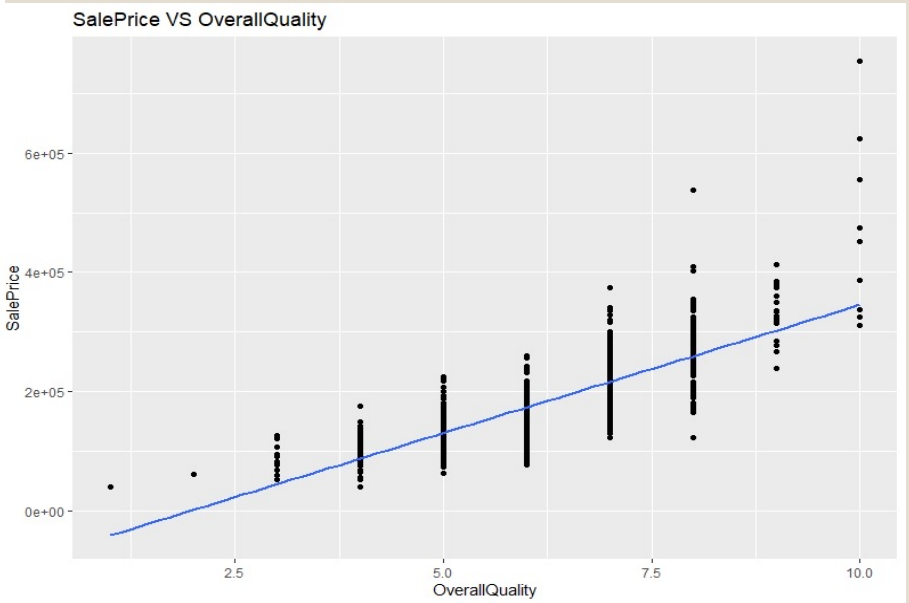
CORRELATION

To understand, the strong relationships between the variables in the dataset and SalePrice, a correlation matrix was formed.

Based on the analysis, we can see the strongest relationship exists between SalePrice and OverallQual followed by LivArea.

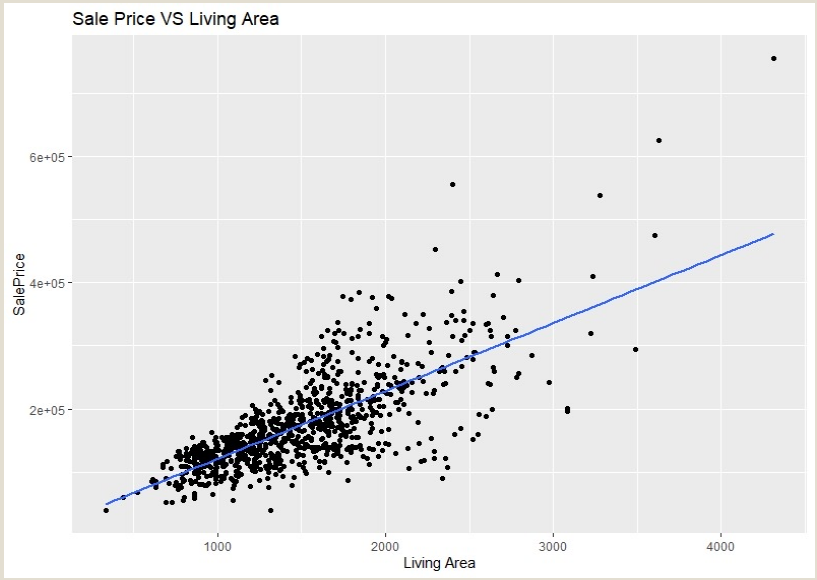
VISUALIZATION

SALEPRICE VS OVERALLQUAL



As the overall quality of a house improves, there is typically a corresponding increase in its sale price. This trend is driven by buyers who are willing to invest more in homes that exhibit superior condition and boast higher quality features.

SALEPRICE VS LIVAREA



As the living area of a house increases, the sale price of the house tends to increase as well. The attractiveness of larger houses generally stems from their ability to provide more space and amenities, features that many buyers find desirable.

Linear Regression with Sale Price as Continuous Dependent Variable

```
R 4.2.2 · ~/
> summary(lmOut)

Call:
lm(formula = SalePrice ~ GarageCars + GarageArea + TotRmsAbvGrd +
    TotalBsmtSF + YearBuilt + OverallQual + FullBath + BsmtFin +
    Bedroom + HalfBath + BsmtFullBath + Fireplaces + LotFrontage +
    Kitchen + OverallCond + LotArea + WoodDeckSF + BsmtHalfBath +
    PoolArea, data = USA_Housing)

Residuals:
     Min       1Q   Median       3Q      Max
-103923  -15703   -2342   13061  323228

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.129e+05  9.294e+04  -4.442 9.95e-06 ***
GarageCars    3.255e+03  2.898e+03   1.123 0.261760
GarageArea    3.253e+01  1.004e+01   3.241 0.001233 **
TotRmsAbvGrd  9.029e+03  1.091e+03   8.279 4.23e-16 ***
TotalBsmtSF   3.162e+01  3.226e+00   9.801  < 2e-16 ***
YearBuilt     1.552e+02  4.680e+01   3.316 0.000949 ***
OverallQual   1.797e+04  1.114e+03  16.124  < 2e-16 ***
FullBath      1.715e+04  2.582e+03   6.641 5.26e-11 ***
BsmtFin       2.521e+01  3.340e+00   7.548 1.04e-13 ***
Bedroom      -6.744e+03  1.727e+03  -3.905 0.000101 ***
HalfBath      1.333e+04  2.222e+03   6.000 2.82e-09 ***
BsmtFullBath  1.341e+03  2.508e+03   0.535 0.593016
Fireplaces    7.447e+03  1.712e+03   4.349 1.52e-05 ***
LotFrontage   1.316e+02  4.643e+01   2.835 0.004682 **
Kitchen      -1.915e+04  4.856e+03  -3.943 8.65e-05 ***
OverallCond   5.955e+03  9.441e+02   6.308 4.36e-10 ***
LotArea       9.522e-01  1.212e-01   7.854 1.09e-14 ***
WoodDeckSF    3.217e+01  7.936e+00   4.054 5.46e-05 ***
BsmtHalfBath -1.308e+03  4.317e+03  -0.303 0.761980
PoolArea     -2.459e+01  2.819e+01  -0.872 0.383238
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28240 on 941 degrees of freedom
Multiple R-squared:  0.8503,    Adjusted R-squared:  0.8473
F-statistic: 281.3 on 19 and 941 DF,  p-value: < 2.2e-16
```

Variance Inflation Factor (VIF)

VIF (Variance Inflation Factor) analysis is a statistical technique used to assess the degree of multicollinearity (correlation) between predictor variables in a multiple regression analysis.

VIF values less than 10 indicate that the variance of a coefficient is inflated due to multicollinearity, with higher values indicating more severe multicollinearity.

LotFrontage	LotArea	OverallQual	OverallCond	YearBuilt	BsmtFin	TotalBsmtSF
1.424288	1.253143	2.924626	1.364292	3.002037	2.237026	3.402420
`1stFlrSF`	`2ndFlrSF`	LivArea	BsmtFullBath	BsmtHalfBath	FullBath	HalfBath
58.777007	83.331675	111.624502	1.933482	1.128702	2.805047	2.177244
Bedroom	Kitchen	TotRmsAbvGrd	Fireplaces	GarageCars	GarageArea	WoodDeckSF
2.228450	1.383041	4.729238	1.499121	5.505599	5.110263	1.198643
PoolArea						
1.015221						

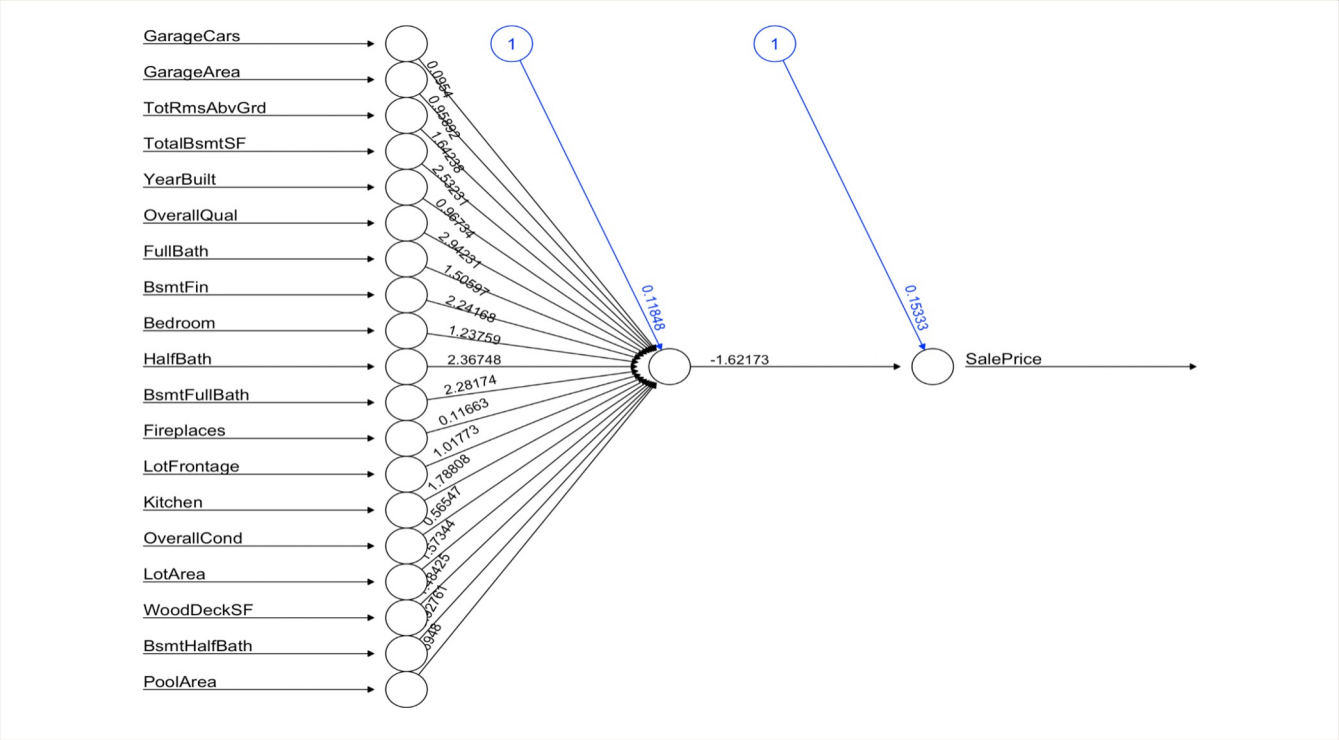
Variables with VIF less than 10

GarageCars	GarageArea	TotRmsAbvGrd	TotalBsmtSF	YearBuilt
5.505599	5.110263	4.729238	3.402420	3.002037
OverallQual	FullBath	BsmtFin	Bedroom	HalfBath
2.924626	2.805047	2.237026	2.228450	2.177244
BsmtFullBath	Fireplaces	LotFrontage	Kitchen	OverallCond
1.933482	1.499121	1.424288	1.383041	1.364292
LotArea	WoodDeckSF	BsmtHalfBath	PoolArea	
1.253143	1.198643	1.128702	1.015221	

The variables with VIF<10 are as follows:

- GarageCars
- GarageArea
- TotRmsAbvGrd
- TotalBsmtSF
- YearBuilt
- OveallQual
- FullBath
- BsmtFin
- Bedroom
- HalfBath
- BsmtFullBath
- Fireplaces
- LotFrontage
- Kitchen
- OverallCond
- LotArea
- WoodDeckSF
- BsmtHalfBath
- PoolArea

Neural network (one hidden layer)



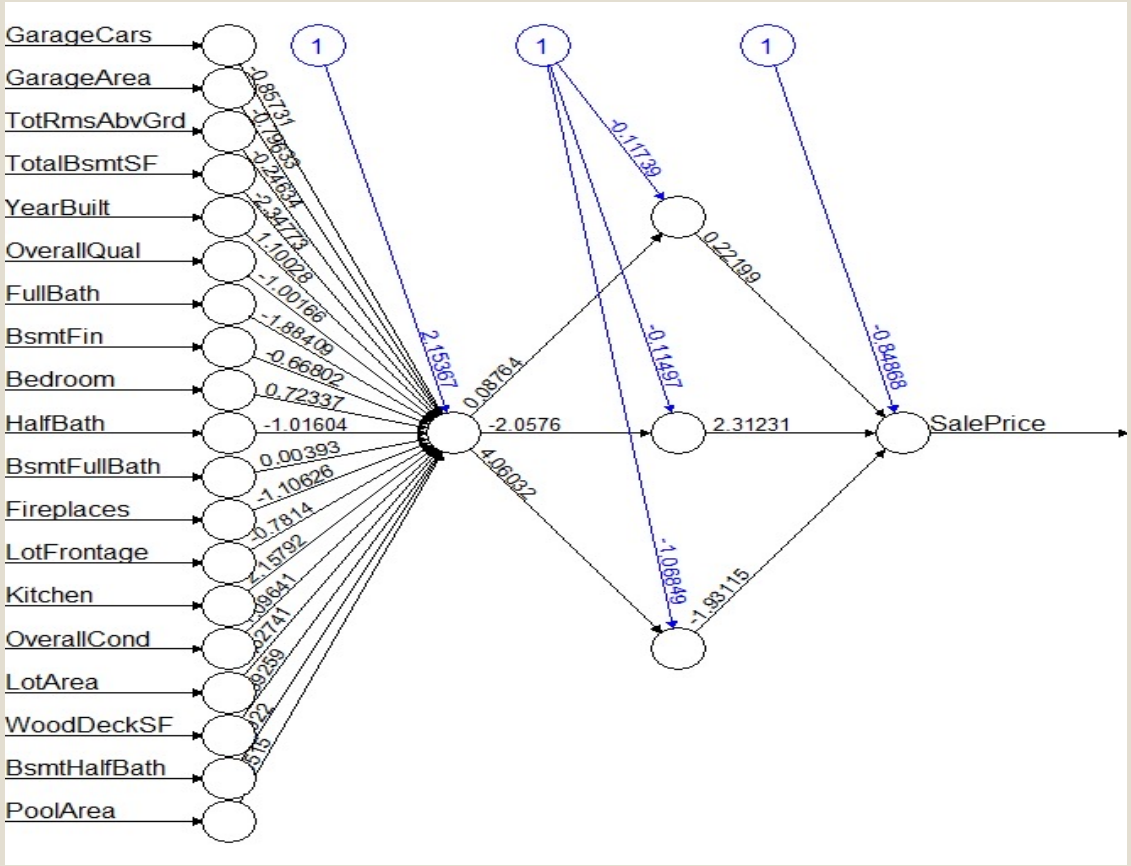
Here we can see the neural network diagram with one hidden layer comprising of all the continuous X variables and a continuous dependent variable Y with a VIF less than 10. First, we split the data into training and testing sets. 70% of the data was used to train the neural network and the remaining 30% of the data was used for testing.

Neural Network (One Layer with loop from 1 to 3 nodes)

```
hidden: 1    thresh: 0.1    rep: 1/1    steps:     248  error: 0.5115   ti
me: 0.08 secs
          [,1]
[1,] 0.9219231
hidden: 2    thresh: 0.1    rep: 1/1    steps:      83  error: 0.70573  ti
me: 0.02 secs
          [,1]
[1,] 0.8856284
hidden: 3    thresh: 0.1    rep: 1/1    steps:     143  error: 0.63948  ti
me: 0.09 secs
          [,1]
[1,] 0.8989678
>
> cur_max_list[which.max(sapply(cur_max_list,max))]
$`1`
[1] 0.9219231
```

Given the large magnitudes of the variables, we normalized the data before supplying it to the neural network. Subsequent to normalization, we established three separate neural networks, each incorporating a single hidden layer. Across iterations, we experimented with different quantities of hidden nodes, ranging from 1 to 3. Upon executing the algorithm, we observed that the highest accuracy, reaching 92.19%, was obtained with a single node, while keeping the hidden layer consistently set at 1.

Neural network (Two hidden layers)



Here we can see the neural network diagram with two hidden layers of all the continuous x variables and a dependent Y variable that is SalePrice.

Second Neural Network (Two layers network with 1-3 nodes in each layer)

```
[1,] 0.6650166
hidden: 3, 2    thresh: 0.1    rep: 1/1    steps:      45      error: 1.1
1102    time: 0.02 secs
         [,1]
[1,] 0.7608515
hidden: 3, 3    thresh: 0.1    rep: 1/1    steps:      50      error: 1.0
5088    time: 0.03 secs
         [,1]
[1,] 0.7916078
>
> cur_max_list[which.max(sapply(cur_max_list,max))]
$`1-3`
[1] 0.8521769
```

The second neural network was designed with two hidden layers, incorporating multiple hidden nodes ranging from 1 to 3 in each iteration and within each hidden layer. Following the execution of the algorithm, it was determined that the maximum accuracy, reaching 85.22%, was accomplished when the neural network featured 1 hidden node in the first layer and 3 hidden nodes in the second layer.

When do visualizations help? And when do they don't?

When visualizations can be helpful:

1. Monitoring performance
2. Presenting research findings
3. Finding insights
4. Highlighting outliers

When visualizations may not be as effective:

1. Poorly designed visualizations
2. Overemphasis on aesthetics
3. Data with little variation
4. Lack of context

Better model?

Neural network 1:

1. 1 Hidden layer and nodes ranging from 1 to 3.
2. Accuracy is 92.12%

Neural network 2:

1. 2 Hidden layers ranging from 1 to 3 in each hidden layer.
2. Accuracy is 85.22%

Thus, from the results, we conclude that Neural Network 1 is better.

Thank You