STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Project: House Price Prediction
Subject: BIA652-B Multivariate Data Analysis (Fall 2017)
Guided By: Prof. Khasha Dehnad

Project Outline

- Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.
- With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, in this project we are predicting the final price of each home using Multiple Linear Regression & PCA.
- So, We choose SalePrice variable as Dependent variable and other variables as Independent Predictor Variables.
- We start with Data transformation and Variable analysis.

Data Source

A closed competition at Kaggle

House Prices: Advanced Regression Techniques

Weblink: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

Acknowledgments:

The Ames Housing dataset was compiled by Dean De Cock for use in data science education. It's an incredible alternative for data scientists looking for a modernized and expanded version of the often cited Boston Housing dataset.

Data Fields

- There are total 79 explanatory variables and 1 dependent variable SalePrice describing (almost) every aspect of residential homes in Ames, Iowa.
- 46 Categorical explanatory variables
- 33 Numerical Continues explanatory variables

- Link: Data Fields

Snapshot of Data

	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Condition2	BldgType	HouseStyle	OverallQual	OverallCond	SalePrice
1	60	RL	65	8450	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	2Story	7	5	208500
2	20	RL	80	9600	Pave	NA	Reg	Lvl	AllPub	FR2	Gtl	Veenker	Feedr	Norm	1Fam	1Story	6	8	181500
3	60	RL	68	11250	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	2Story	7	5	223500
4	70	RL	60	9550	Pave	NA	IR1	Lvl	AllPub	Corner	Gtl	Crawfor	Norm	Norm	1Fam	2Story	7	5	140000
5	60	RL	84	14260	Pave	NA	IR1	Lvl	AllPub	FR2	Gtl	NoRidge	Norm	Norm	1Fam	2Story	8	5	250000
6	50	RL	85	14115	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	Mitchel	Norm	Norm	1Fam	1.5Fin	5	5	143000
7	20	RL	75	10084	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	1Story	8	5	307000
8	60	RL	NA	10382	Pave	NA	IR1	Lvl	AllPub	Corner	Gtl	NWAmes	PosN	Norm	1Fam	2Story	7	6	200000
9	50	RM	51	6120	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	OldTown	Artery	Norm	1Fam	1.5Fin	7	5	129900
10	190	RL	50	7420	Pave	NA	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Artery	Artery	2fmCon	1.5Unf	5	6	118000
11	20	RL	70	11200	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	Sawyer	Norm	Norm	1Fam	1Story	5	5	129500
12	60	RL	85	11924	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	NridgHt	Norm	Norm	1Fam	2Story	9	5	345000
13	20	RL	NA	12968	Pave	NA	IR2	Lvl	AllPub	Inside	Gtl	Sawyer	Norm	Norm	1Fam	1Story	5	6	144000
14	20	RL	91	10652	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	1Story	7	5	279500
15	20	RL	NA	10920	Pave	NA	IR1	Lvl	AllPub	Corner	Gtl	NAmes	Norm	Norm	1Fam	1Story	6	5	157000
16	45	RM	51	6120	Pave	NA	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5Unf	7	8	132000
17	20	RL	NA	11241	Pave	NA	IR1	Lvl	AllPub	CulDSac	Gtl	NAmes	Norm	Norm	1Fam	1Story	6	7	149000
18	90	RL	72	10791	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	Sawyer	Norm	Norm	Duplex	1Story	4	5	90000
19	20	RL	66	13695	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	SawyerW	RRAe	Norm	1Fam	1Story	5	5	159000
20	20	RL	70	7560	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Story	5	6	139000
21	60	RL	10	14215	Pave	NA	IR1	Lvl	AllPub	Corner	Gtl	NridgHt	Norm	Norm	1Fam	2Story	8	5	325300
22	45	RM	57	7449	Pave	Gr	Reg	Bnk	AllPub	Inside	Gtl	IDOTRR	Norm	Norm	1Fam	1.5Unf	7	7	139400
23	20	RL	75	9742	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	1Story	8	5	230000
24	120	RM	44	4224	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	MeadowV	Norm	Norm	TwnhsE	1Story	5	7	129900
25	20	RL	NA	8246	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	Sawyer	Norm	Norm	1Fam	1Story	5	8	154000
26	20	RL	11	14230	Pave	NA	Reg	Lvl	AllPub	Corner	Gtl	NridgHt	Norm	Norm	1Fam	1Story	8	5	256300
27	20	RL	60	7200	Pave	NA	Reg	Lvl	AllPub	Corner	Gtl	NAmes	Norm	Norm	1Fam	1Story	5	7	134800

Data Transformation

- We have transformed Categorical Variables from text to number.
- For Eg., **Pool Quality**

Text(PoolQ)	Number
Ex(Excellent)	6
Gd(Good)	5
TA(Typical)	4
Av(Average)	4
Fa(fair)	3
Po(poor)	2
No	1
NA	0

PoolQC	PoolQC_new
NA	0
Gd	5
NA	0
NA	0
NA	0
NA	0
NA	0
NA	0
NA	0
NA	0
NA	0
NA	0
NA	0
Ex	6

- Here PoolQC stands for Pool Quality. Pool Quality devide into 5 types such as Excellent, Good, Typical, Fair, No pool. Convert this quality into numerical form 6 to 0.
- Transformed some variables into numerical form in this way such as Exterior Quality, Exterior Condition, Kitchen Quality, Basement Quality, Basement Condition Etc…

Data Transformation

- In this type of variables, most of value are in numeric form and some values are defined as “NA”
- So that, we have transformed “NA” to 0

Mixed Value(LotFrontage)	Numeric
Number	Number
NA	0

LotFrontage
65
80
68
60
84
85
75
NA
51
50
70
85
NA

LotFrontage_new
65
80
68
60
84
85
75
0
51
50
70
85
0

Data Transformation

Central Air: Central air Conditioning

Values : N -> No, Y -> Yes

Central Air	Numeric Vlaue
Y	1
N	0

CentralAir
Y
Y
Y
Y
N
N
Y
Y
Y
Y
Y
Y
Y

CentralAir_new
1
1
1
1
0
0
1
1
1
1
1
1
1

Data Transformation

Handling Missing Value

- Find Missing values by running MEANS procedure:
- “MasVnrArea” and “GarageYrBlt” variables had Missing Values.
- Replace these Missing values with 0

SalePrice

The MEANS Procedure

Variable	N Miss	N
Id	0	1460
MSSubClass	0	1460
LotArea	0	1460
OverallQual	0	1460
OverallCond	0	1460
YearBuilt	0	1460
YearRemodAdd	0	1460
MasVnrArea	8	1452
BsmtFinSF1	0	1460
BsmtFinSF2	0	1460
BsmtUnfSF	0	1460
TotalBsmtSF	0	1460
_1stFlrSF	0	1460
_2ndFlrSF	0	1460
LowQualFinSF	0	1460
GrLivArea	0	1460
BsmtFullBath	0	1460
BsmtHalfBath	0	1460
FullBath	0	1460
HalfBath	0	1460
BedroomAbvGr	0	1460
KitchenAbvGr	0	1460
TotRmsAbvGrd	0	1460
Fireplaces	0	1460
GarageYrBlt	81	1379

SalePrice

The MEANS Procedure

Variable	N Miss	N
Id	0	1460
MSSubClass	0	1460
LotArea	0	1460
OverallQual	0	1460
OverallCond	0	1460
YearBuilt	0	1460
YearRemodAdd	0	1460
MasVnrArea	0	1460
BsmtFinSF1	0	1460
BsmtFinSF2	0	1460
BsmtUnfSF	0	1460
TotalBsmtSF	0	1460
_1stFlrSF	0	1460
_2ndFlrSF	0	1460
LowQualFinSF	0	1460
GrLivArea	0	1460
BsmtFullBath	0	1460
BsmtHalfBath	0	1460
FullBath	0	1460
HalfBath	0	1460
BedroomAbvGr	0	1460
KitchenAbvGr	0	1460
TotRmsAbvGrd	0	1460
Fireplaces	0	1460
GarageYrBlt	0	1460

Analysis of Variable: GarageCars

The UNIVARIATE Procedure
Variable: GarageCars

Moments			
N	1460	Sum Weights	1460
Mean	1.76712329	Sum Observations	2580
Std Deviation	0.74731501	Variance	0.55847972
Skewness	-0.3425489	Kurtosis	0.22099776
Uncorrected SS	5374	Corrected SS	814.821918
Coeff Variation	42.2899192	Std Error Mean	0.01955813

Modes	
Mode	Count
2	824

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	90.35237	Pr > |t|	<.0001
Sign	M	689.5	Pr >= |M|	<.0001
Signed Rank	S	475755	Pr >= |S|	<.0001

Basic Statistical Measures			
Location		Variability	
Mean	1.767123	Std Deviation	0.74732
Median	2.000000	Variance	0.55848
Mode	2.000000	Range	4.00000
		Interquartile Range	1.00000

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	1454	4	421
0	1451	4	748
0	1450	4	1191
0	1408	4	1341
0	1350	4	1351

Analysis of Variable: OverallQual

The UNIVARIATE Procedure
Variable: OverallQual

Moments			
N	1460	Sum Weights	1460
Mean	6.09931507	Sum Observations	8905
Std Deviation	1.38299655	Variance	1.91267945
Skewness	0.21694393	Kurtosis	0.09629278
Uncorrected SS	57105	Corrected SS	2790.59932
Coeff Variation	22.6746205	Std Error Mean	0.03619467

Modes	
Mode	Count
5	397

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	168.5142	Pr > |t|	<.0001
Sign	M	730	Pr >= |M|	<.0001
Signed Rank	S	533265	Pr >= |S|	<.0001

Basic Statistical Measures			
Location		Variability	
Mean	6.099315	Std Deviation	1.38300
Median	6.000000	Variance	1.91268
Mode	5.000000	Range	9.00000
		Interquartile Range	2.00000

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
1	534	10	1183
1	376	10	1244
2	1101	10	1299
2	917	10	1374
2	637	10	1443

Analysis of Variable: GrLivArea

The UNIVARIATE Procedure
Variable: GrLivArea

Moments			
N	1460	Sum Weights	1460
Mean	1515.4637	Sum Observations	2212577
Std Deviation	525.480383	Variance	276129.633
Skewness	1.36656036	Kurtosis	4.89512058
Uncorrected SS	3755953259	Corrected SS	402873135
Coeff Variation	34.6745609	Std Error Mean	13.7524502

Modes	
Mode	Count
864	22

Basic Statistical Measures			
Location		Variability	
Mean	1515.464	Std Deviation	525.48038
Median	1464.000	Variance	276130
Mode	864.000	Range	5308
		Interquartile Range	648.50000

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
334	534	3627	1170
438	1101	4316	692
480	917	4476	1183
520	30	4676	524
605	529	5642	1299

Q-Q Plot

The UNIVARIATE Procedure
Probability Plot for GrLivArea
Mean 1515.464
Std Deviation 525.4804
GrLivArea
6000
5000
4000
3000
2000
1000
0
0.01
0.1
1
5
10
25
50
75
90
95
99
99.9
99.99
Normal Percentiles
Normal Line
Mu=1515.5, Sigma=525.48

Histogram

Target Variable: Sale Price

Plots

Box Plot: Overall Quality vs Sale Price

Plots

Plot: Pool Quality vs Saleprice

Plots

Plot: Year Built Vs Overall Quality

Plots

Plot: Garage Cars vs sale Price

Regression Analysis

Multiple Linear Regression Analysis on all predictor variables of transformed dataset

Selection = stepwise
slentry = 0.10
slstay = 0.10

Regression Analysis with stepwise selection

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	1460
Number of Observations Used	1460

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	36	7.783638E12	2.162122E11	216.02	<.0001
Error	1423	1.424273E12	1000894627		
Corrected Total	1459	9.207911E12			

Root MSE	31637	R-Square	0.8453
Dependent Mean	180921	Adj R-Sq	0.8414
Coeff Var	17.48657		

Regression Analysis

Parameter Estimates									
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS	Standardized Estimate	Variance Inflation
Intercept	1	-502513	102720	-4.89	<.0001	4.778942E13	23953626448	0	0
MSSubClass	1	-224.27921	21.11725	-10.62	<.0001	65411302654	1.128995E11	-0.11942	1.16315
LotFrontage_new	1	-52.19481	28.72106	-1.82	0.0694	1225681345	3305536381	-0.02018	1.13386
LotArea	1	0.39675	0.09318	4.26	<.0001	6.130692E11	18145734113	0.04985	1.26091
Alley_new	1	-4077.95268	2320.35948	-1.76	0.0791	1.333279E11	3091449632	-0.02026	1.22315
Utilities_new	1	66147	32312	2.05	0.0408	2480134035	4194413377	0.02179	1.04244
Neighborhood_new	1	-230.39600	132.76128	-1.74	0.0829	50094119964	3014361826	-0.01907	1.11132
Condition1_new	1	-2529.04229	708.77963	-3.57	0.0004	17308009597	12743187909	-0.03865	1.07940
Condition2_new	1	-5707.49101	2054.59286	-2.78	0.0055	4864793712	7723733500	-0.02999	1.07189
OverallQual	1	12948	1163.27312	11.13	<.0001	5.28424E12	1.240118E11	0.22542	3.77288
OverallCond	1	4939.53048	908.65999	5.44	<.0001	145280767	29577188453	0.06919	1.49040
YearBuilt	1	141.04614	50.40733	2.80	0.0052	70526264130	7836523463	0.05362	3.37872
RoofStyle_new	1	2889.20433	969.63323	2.98	0.0029	52337960263	8886482479	0.03295	1.12471
RoofMatl_new	1	-9997.73827	2629.96495	-3.80	0.0001	794461582	14464118199	-0.04132	1.08702
Exterior1st_new	1	-625.76347	251.96761	-2.48	0.0131	3139145056	6173327957	-0.02709	1.09427
MasVnrType_new	1	5467.22156	1165.80003	4.69	<.0001	1109594243	22012713639	0.06619	1.83279
MasVnrArea	1	44.03978	6.51213	6.76	<.0001	2.282753E11	45775380213	0.10019	2.01921
ExterQual_new	1	11651	2505.54393	4.65	<.0001	1.430006E11	21643105814	0.08422	3.01800
BsmtQual_new	1	8018.77192	1860.43331	4.31	<.0001	36087376047	18594130177	0.09941	4.89350
BsmtCond_new	1	-7913.93333	2110.20196	-3.75	0.0002	52571042892	14077485514	-0.06891	3.10568
BsmtExposure_new	1	3599.54005	669.55544	5.38	<.0001	55618008153	28927392868	0.06680	1.42055
BsmtFinType1_new	1	2279.60456	548.50143	4.16	<.0001	34487627955	17288282860	0.06048	1.94837

Regression Analysis

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Regression Analysis

Multiple Linear Regression Analysis on selected variables of last stepwise model

selection = MAXR

regression analysis on variable selected from Stepwise with MAXR selection

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	1460
Number of Observations Used	1460

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	36	7.783638E12	2.162122E11	216.02	<.0001
Error	1423	1.424273E12	1000894627		
Corrected Total	1459	9.207911E12			

Root MSE	31637	R-Square	0.8453
Dependent Mean	180921	Adj R-Sq	0.8414
Coeff Var	17.48657		

Regression Analysis

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS	Standardized Estimate	Variance Inflation
Intercept	1	-502513	102720	-4.89	<.0001	4.778942E13	23953626448	0	0
MSSubClass	1	-224.27921	21.11725	-10.62	<.0001	66411302654	1.128995E11	-0.11942	1.16315
LotFrontage_new	1	-52.19481	28.72106	-1.82	0.0694	1225681345	3305536381	-0.02018	1.13386
LotArea	1	0.39675	0.09318	4.26	<.0001	6.130692E11	18145734113	0.04985	1.26091
Alley_new	1	-4077.95268	2320.35948	-1.76	0.0791	1.333279E11	3091449632	-0.02026	1.22315
Utilities_new	1	66147	32312	2.05	0.0408	2480134035	4194413377	0.02179	1.04244
Neighborhood_new	1	-230.39600	132.76128	-1.74	0.0829	50094119964	3014361826	-0.01907	1.11132
Condition1_new	1	-2529.04229	708.77963	-3.57	0.0004	17308009697	12743187909	-0.03865	1.07940
Condition2_new	1	-5707.49101	2054.59286	-2.78	0.0055	4864793712	7723733500	-0.02999	1.07189
OverallQual	1	12948	1163.27312	11.13	<.0001	5.28424E12	1.240118E11	0.22542	3.77288
OverallCond	1	4939.53048	908.65999	5.44	<.0001	145280767	29577188453	0.06919	1.49040
YearBuilt	1	141.04614	50.40733	2.80	0.0052	70526264130	7836523463	0.05362	3.37872
RoofStyle_new	1	2889.20433	969.63323	2.98	0.0029	52337960263	8886482479	0.03295	1.12471
RoofMatl_new	1	-9997.73827	2629.96495	-3.80	0.0001	794461582	14464118199	-0.04132	1.08702
Exterior1st_new	1	-625.76347	251.96761	-2.48	0.0131	3139145056	6173327957	-0.02709	1.09427
MasVnrType_new	1	5467.22156	1165.80003	4.69	<.0001	1109694243	22012713639	0.06619	1.83279
MasVnrArea	1	44.03978	6.51213	6.76	<.0001	2.282753E11	45775380213	0.10019	2.01921
ExterQual_new	1	11651	2505.54393	4.65	<.0001	1.430006E11	21643105814	0.08422	3.01800
BsmtQual_new	1	8018.77192	1860.43331	4.31	<.0001	36087376047	18594130177	0.09941	4.89350
BsmtCond_new	1	-7913.93333	2110.20196	-3.75	0.0002	52571042892	14077485514	-0.06891	3.10568
BsmtExposure_new	1	3599.54005	669.55544	5.38	<.0001	55618008153	28927392868	0.06680	1.42055
BsmtFinType1_new	1	2279.60456	548.50143	4.16	<.0001	34487627955	17288282860	0.06048	1.94837
Electrical_new	1	3755.35906	2230.75045	1.68	0.0925	618606308	2836542630	0.01961	1.24855
BsmtFullBath	1	7363.36238	2068.22611	3.56	0.0004	17894168192	12688583218	0.04810	1.67899
BedroomAbvGr	1	-5847.77902	1565.39734	-3.74	0.0002	1.436214E11	13967577135	-0.06005	2.37717
KitchenQual_new	1	8040.86673	1993.16181	4.03	<.0001	72982026747	16289545586	0.06718	2.55137
TotRmsAbvGrd	1	2663.28513	1089.30474	2.44	0.0146	2.920075E11	5983081222	0.05449	4.56965
Functional_new	1	4298.40895	1313.83115	3.27	0.0011	663829182	10713335707	0.03613	1.12177
Fireplaces	1	4230.08938	1593.26428	2.65	0.0080	49488457425	7055240226	0.03433	1.53785
GarageYrBlt	1	-25.04776	7.74696	-3.23	0.0013	1146597792	10463172428	-0.14305	18.00786

Regression Analysis

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Regression Analysis

Multiple Linear Regression Analysis after removing variables from previous MAXR

MAXR after removing variables

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	1460
Number of Observations Used	1460

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	32	7.770609E12	2.428315E11	241.09	<.0001
Error	1427	1.437303E12	1007219826		
Corrected Total	1459	9.207911E12			

Root MSE	31737	R-Square	0.8439
Dependent Mean	180921	Adj R-Sq	0.8404
Coeff Var	17.54174		

Regression Analysis

Parameter Estimates									
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS	Standardized Estimate	Variance Inflation
Intercept	1	-532452	97190	-5.48	<.0001	4.778942E13	30230122765	0	0
MSSubClass	1	-223.13047	20.70070	-10.78	<.0001	6541130265 4	1.170233E11	-0.11881	1.11069
LotArea	1	0.41269	0.09278	4.45	<.0001	5.966856E11	19928080795	0.05185	1.24226
Utilities_new	1	62469	32259	1.94	0.0530	2992129904	3776913286	0.02058	1.03249
Condition1_new	1	-2510.40855	708.60467	-3.54	0.0004	20339053243	12841689104	-0.03837	1.07209
Condition2_new	1	-5379.42259	2054.49965	-2.62	0.0089	3901252183	6905314941	-0.02826	1.06506
OverallQual	1	12687	1161.93913	10.92	<.0001	5.453429E12	1.200863E11	0.22087	3.74059
OverallCond	1	4715.96797	898.40230	5.25	<.0001	720530412	27753917299	0.06606	1.44779
YearBuilt	1	158.47717	47.71743	3.32	0.0009	89007667270	11109734293	0.06025	3.00872
RoofStyle_new	1	2720.29528	970.40734	2.80	0.0051	51439061029	7914951544	0.03102	1.11943
RoofMatl_new	1	-10127	2635.82069	-3.84	0.0001	810491247	14867482653	-0.04186	1.08501
Exterior1st_new	1	-620.79999	252.08897	-2.46	0.0139	2732171485	6108295379	-0.02687	1.08845
MasVnrType_new	1	5298.03080	1167.19830	4.54	<.0001	1470106976	20752191875	0.06414	1.82565
MasVnrArea	1	44.31635	6.52069	6.80	<.0001	2.292443E11	46522767176	0.10082	2.01180
ExterQual_new	1	11895	2504.80764	4.75	<.0001	1.435377E11	22716319244	0.08599	2.99728
BsmtQual_new	1	8039.11473	1863.32224	4.31	<.0001	35306233802	18748426576	0.09966	4.87788
BsmtCond_new	1	-8319.46081	2105.68381	-3.95	<.0001	52460566453	15722744368	-0.07244	3.07297
BsmtExposure_new	1	3778.77438	667.26827	5.66	<.0001	58344372562	32301693144	0.07013	1.40200
BsmtFinType1_new	1	2221.80964	548.71736	4.05	<.0001	36318685740	16513555999	0.05895	1.93766
BsmtFullBath	1	7224.02825	2074.04122	3.48	0.0005	18177583507	12219358719	0.04719	1.67784
BedroomAbvGr	1	-6536.88033	1557.67280	-4.20	<.0001	1.382212E11	17738337281	-0.06713	2.33899
KitchenQual_new	1	7898.88919	1996.05207	3.96	<.0001	71599499678	15772938905	0.06600	2.54271
TotRmsAbvGrd	1	2858.82572	1085.52426	2.63	0.0085	2.943547E11	6985871683	0.05849	4.50949
Functional_new	1	4279.91186	1317.58117	3.25	0.0012	562579572	10627697818	0.03597	1.12110
Fireplaces	1	4851.89606	1578.43734	3.07	0.0022	55865482898	9516836216	0.03937	1.49988
GarageYrBlt	1	-24.38006	7.72302	-3.16	0.0016	1041309912	10037359156	-0.13923	17.78434
GarageCars	1	13895	1880.88900	7.39	<.0001	83621170685	54967203542	0.13071	2.86197
GarageQual_new	1	7584.49809	3692.48264	2.05	0.0402	7453215175	4249531743	0.08992	17.51941
WoodDeckSF	1	22.25797	7.28235	3.06	0.0023	12462579123	9409199186	0.03512	1.20683
ScreenPorch	1	51.60474	15.52957	3.32	0.0009	13868416378	11122039542	0.03622	1.08606

Correlation Matrix

Pearson Correlation Coefficients, N = 1460
Prob > |r| under H0: Rho=0

	Utilities_new	OverallQual	OverallCond	YearBuilt	GarageYrBlt	GarageCars	GarageQual_new	WoodDeckSF	SaleType_new	SaleCondition_new	GrLivArea
Utilities_new	1.00000	0.00188 0.9428	-0.00999 0.7028	0.01150 0.6605	-0.00515 0.8441	-0.00816 0.7554	-0.00682 0.7946	0.01969 0.4521	-0.00427 0.8707	0.05127 0.0502	0.00855 0.7442
OverallQual	0.00188 0.9428	1.00000	-0.09193 0.0004	0.57232 <.0001	0.28900 <.0001	0.60067 <.0001	0.28812 <.0001	0.23892 <.0001	0.30567 <.0001	-0.19816 <.0001	0.59301 <.0001
OverallCond	-0.00999 0.7028	-0.09193 0.0004	1.00000	-0.37598 <.0001	-0.00652 0.8036	-0.18576 <.0001	0.01699 0.5167	-0.00333 0.8987	-0.16560 <.0001	0.16483 <.0001	-0.07969 0.0023
YearBuilt	0.01150 0.6605	0.57232 <.0001	-0.37598 <.0001	1.00000	0.27203 <.0001	0.53785 <.0001	0.27651 <.0001	0.22488 <.0001	0.32684 <.0001	-0.21409 <.0001	0.19901 <.0001
GarageYrBlt	-0.00515 0.8441	0.28900 <.0001	-0.00652 0.8036	0.27203 <.0001	1.00000	0.59800 <.0001	0.96937 <.0001	0.11730 <.0001	0.04249 0.1046	0.01847 0.4806	0.16254 <.0001
GarageCars	-0.00816 0.7554	0.60067 <.0001	-0.18576 <.0001	0.53785 <.0001	0.59800 <.0001	1.00000	0.58197 <.0001	0.22634 <.0001	0.25212 <.0001	-0.16244 <.0001	0.46725 <.0001
GarageQual_new	-0.00682 0.7946	0.28812 <.0001	0.01699 0.5167	0.27651 <.0001	0.96937 <.0001	0.58197 <.0001	1.00000	0.11749 <.0001	0.03476 0.1844	0.03060 0.2427	0.16446 <.0001
WoodDeckSF	0.01969 0.4521	0.23892 <.0001	-0.00333 0.8987	0.22488 <.0001	0.11730 <.0001	0.22634 <.0001	0.11749 <.0001	1.00000	0.02008 0.4432	0.01777 0.4976	0.24743 <.0001
SaleType_new	-0.00427 0.8707	0.30567 <.0001	-0.16560 <.0001	0.32684 <.0001	0.04249 0.1046	0.25212 <.0001	0.03476 0.1844	0.02008 0.4432	1.00000	-0.78538 <.0001	0.15871 <.0001
SaleCondition_new	0.05127 0.0502	-0.19816 <.0001	0.16483 <.0001	-0.21409 <.0001	0.01847 0.4806	-0.16244 <.0001	0.03060 0.2427	0.01777 0.4976	-0.78538 <.0001	1.00000	-0.11135 <.0001
GrLivArea	0.00855 0.7442	0.59301 <.0001	-0.07969 0.0023	0.19901 <.0001	0.16254 <.0001	0.46725 <.0001	0.16446 <.0001	0.24743 <.0001	0.15871 <.0001	-0.11135 <.0001	1.00000

Regression Analysis

Regression Analysis

Multiple Linear Regression Analysis after analyzing correlation matrix we remove one more variable

Final Regression Analysis

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	1460
Number of Observations Used	1460

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	29	7.75649E12	2.674652E11	263.52	<.0001
Error	1430	1.451421E12	1014979996		
Corrected Total	1459	9.207911E12			

Root MSE	31859	R-Square	0.8424
Dependent Mean	180921	Adj R-Sq	0.8392
Coeff Var	17.60919		

Regression Analysis

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS	Standardized Estimate	Variance Inflation
Intercept	1	-475261	91613	-5.19	<.0001	4.778942E13	27315714998	0	0
MSSubClass	1	-220.56474	20.76855	-10.62	<.0001	65411302654	1.144768E11	-0.11744	1.10944
LotArea	1	0.40303	0.09305	4.33	<.0001	5.966856E11	19042800291	0.05064	1.23987
Condition1_new	1	-2389.63963	710.46138	-3.36	0.0008	20185115023	11482629631	-0.03652	1.06948
Condition2_new	1	-5455.19742	2062.15072	-2.65	0.0082	3897713658	7102931398	-0.02866	1.06480
OverallQual	1	12687	1166.28016	10.88	<.0001	5.453968E12	1.200987E11	0.22086	3.73978
OverallCond	1	4880.06183	895.96884	5.45	<.0001	745466686	30110801941	0.06836	1.42895
YearBuilt	1	162.54772	47.45899	3.43	0.0006	89300470804	11906449296	0.06180	2.95347
RoofStyle_new	1	3069.24668	969.29911	3.17	0.0016	51782141835	10176664273	0.03500	1.10834
RoofMatl_new	1	-10064	2644.86850	-3.81	0.0001	806704757	14695518134	-0.04160	1.08412
Exterior1st_new	1	-583.40749	252.82564	-2.31	0.0212	2603459783	5404546600	-0.02525	1.08645
MasVnrType_new	1	5162.19191	1171.02225	4.41	<.0001	1320858085	19724022214	0.06250	1.82359
MasVnrArea	1	42.45812	6.52359	6.51	<.0001	2.254905E11	42993770798	0.09659	1.99820
ExterQual_new	1	11910	2508.12318	4.75	<.0001	1.448278E11	22884866819	0.08609	2.98225
BsmtQual_new	1	7824.13798	1866.92502	4.19	<.0001	35464305977	17826937175	0.09699	4.85933
BsmtCond_new	1	-8384.92871	2112.91260	-3.97	<.0001	52892263260	15984283228	-0.07301	3.07045
BsmtExposure_new	1	3796.13525	669.64839	5.67	<.0001	58835278583	32617237493	0.07045	1.40123
BsmtFinType1_new	1	2148.03354	548.55403	3.92	<.0001	35486861110	15563267342	0.05699	1.92170
BsmtFullBath	1	7339.00687	2079.52174	3.53	0.0004	18904090516	12641690200	0.04794	1.67382
BedroomAbvGr	1	-4612.89795	1370.57683	-3.37	0.0008	1.386338E11	11497367279	-0.04737	1.79700
KitchenQual_new	1	7856.40118	2000.70102	3.93	<.0001	70017554456	15650939096	0.06564	2.53504
Functional_new	1	4438.73493	1321.85800	3.36	0.0008	381638443	11444749675	0.03731	1.11977
Fireplaces	1	5127.92564	1578.96931	3.25	0.0012	1.03759E11	10705169360	0.04161	1.48941
GarageYrBlt	1	-9.61591	2.42239	-3.97	<.0001	1709554329	15993772585	-0.05492	1.73627
GarageCars	1	14309	1879.20884	7.61	<.0001	1.431825E11	58847078734	0.13460	2.83501
WoodDeckSF	1	22.07582	7.30961	3.02	0.0026	18559334263	9257690192	0.03483	1.20658
ScreenPorch	1	47.84134	15.49947	3.09	0.0021	10116490627	9670093268	0.03358	1.07359
SaleType_new	1	3875.55694	724.73297	5.35	<.0001	34557797272	29024830481	0.09705	2.98784
SaleCondition_new	1	1972.44704	826.74897	2.39	0.0172	1756474808	5777245992	0.04131	2.71975
GrLivArea	1	57.14464	2.97213	19.23	<.0001	3.752083E11	3.752083E11	0.37799	3.50628

Regression Analysis

Cook's D

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice
Cook's D for SalePrice
Cook's D
Observation
1.5
1.0
0.5
0.0
0
500
1000
1500
1299
524
692
272
1183
1170
30
186
899

Outlier and Leverage

Influence

Influence Diagnostics for SalePrice
DFFITS
Observation
2
0
-2
-4
-6
0
250
500
750
1000
1250
1500
692
1183
272
30
186
314
441
549
899
1170
1244
1047
1004
1182
54
179
279
524
1299

Normalize the Data

We have normalize all predictor variables to run PCA

MSZoning_new	LotFrontage_new	Street_new	Alley_new	LotShape_new	LandContour_new	Utilities_new	LotConfig_new	LandSlope_new
-0.045516344	0.9717583806	-0.064216206	-0.244632991	-0.750473415	0.3145590869	0.0261711961	-0.628100868	-0.225638819
-0.045516344	0.5809762099	-0.064216206	-0.244632991	1.3784602356	0.3145590869	0.0261711961	0.604462663	-0.225638819
-0.045516344	0.3204547629	-0.064216206	-0.244632991	1.3784602356	0.3145590869	0.0261711961	-1.860664398	-0.225638819
-0.045516344	1.1020191041	-0.064216206	-0.244632991	1.3784602356	0.3145590869	0.0261711961	-0.628100868	-0.225638819
-0.045516344	1.134584285	-0.064216206	-0.244632991	1.3784602356	0.3145590869	0.0261711961	0.604462663	-0.225638819
-0.045516344	0.8089324761	-0.064216206	-0.244632991	-0.750473415	0.3145590869	0.0261711961	0.604462663	-0.225638819
-0.045516344	-1.63345609	-0.064216206	-0.244632991	1.3784602356	0.3145590869	0.0261711961	-1.860664398	-0.225638819
1.5367184727	0.0273681349	-0.064216206	-0.244632991	-0.750473415	0.3145590869	0.0261711961	0.604462663	-0.225638819
-0.045516344	-0.005197046	-0.064216206	-0.244632991	-0.750473415	0.3145590869	0.0261711961	-1.860664398	-0.225638819
-0.045516344	0.6461065717	-0.064216206	-0.244632991	-0.750473415	0.3145590869	0.0261711961	0.604462663	-0.225638819
-0.045516344	1.134584285	-0.064216206	-0.244632991	1.3784602356	0.3145590869	0.0261711961	0.604462663	-0.225638819
-0.045516344	-1.63345609	-0.064216206	-0.244632991	0.6688156856	0.3145590869	0.0261711961	0.604462663	-0.225638819
-0.045516344	1.3299753703	-0.064216206	-0.244632991	1.3784602356	0.3145590869	0.0261711961	0.604462663	-0.225638819
-0.045516344	-1.63345609	-0.064216206	-0.244632991	1.3784602356	0.3145590869	0.0261711961	-1.860664398	-0.225638819
1.5367184727	0.0273681349	-0.064216206	-0.244632991	-0.750473415	0.3145590869	0.0261711961	-1.860664398	-0.225638819
-0.045516344	-1.63345609	-0.064216206	-0.244632991	1.3784602356	0.3145590869	0.0261711961	-1.244382633	-0.225638819
-0.045516344	0.7112369335	-0.064216206	-0.244632991	-0.750473415	0.3145590869	0.0261711961	0.604462663	-0.225638819
-0.045516344	0.5158458482	-0.064216206	-0.244632991	-0.750473415	0.3145590869	0.0261711961	0.604462663	-0.225638819
-0.045516344	0.6461065717	-0.064216206	-0.244632991	-0.750473415	0.3145590869	0.0261711961	0.604462663	-0.225638819

Summary: Normalized Data

Mean = 0 ; Standard Deviation = 1

Summary of Normalized Data

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
MSSubClass	1460	5.201319E-17	1.0000000	-0.8722639	3.1465944
MSZoning_new	1460	-8.13961E-16	1.0000000	-4.7922208	1.5367185
LotFrontage_new	1460	-2.92004E-17	1.0000000	-1.6334561	1.5904968
LotArea	1460	-6.32675E-17	1.0000000	-0.9234128	20.5112451
Alley_new	1460	3.698716E-16	1.0000000	-0.2446330	4.8215254
Street_new	1460	-6.08341E-18	1.0000000	-0.0642162	15.5617271
LotShape_new	1460	-2.79837E-16	1.0000000	-0.7504734	1.3784602
LandContour_new	1460	-3.18771E-16	1.0000000	-3.9247295	0.3145591
Utilities_new	1460	5.499406E-16	1.0000000	-38.1837752	0.0261712
LotConfig_new	1460	-5.15265E-16	1.0000000	-1.8606644	0.6044627
LandSlope_new	1460	6.02258E-17	1.0000000	-0.2256388	7.0146398
Neighborhood_new	1460	7.300097E-18	1.0000000	-1.8464690	1.9547785
Condition1_new	1460	-6.20508E-17	1.0000000	-0.3243907	6.2649720
Condition2_new	1460	1.49652E-16	1.0000000	-0.0853365	19.0825545
BldgType_new	1460	2.968706E-16	1.0000000	-0.4115498	2.9265762
HouseStyle_new	1460	-1.56952E-16	1.0000000	-1.5896765	2.0727432
OverallQual	1460	2.238696E-16	1.0000000	-3.6871495	2.8204589
OverallCond	1460	2.311697E-16	1.0000000	-4.1115611	3.0775158
YearBuilt	1460	1.092581E-15	1.0000000	-3.2866975	1.2823996
YearRemodAdd	1460	4.404392E-15	1.0000000	-1.6887898	1.2174256
RoofStyle_new	1460	2.590013E-16	1.0000000	-1.5870136	3.9323762
RoofMatl_new	1460	5.268236E-16	1.0000000	-3.1122978	18.2065248

PCA

PCA Anlysis for all Normalized Variables

The PRINCOMP Procedure

Observations	1460
Variables	79

	MSSubClass	MSZoning_new	LotFrontage_new	LotArea	Alley_new	Street_new	LotShape_new
Mean	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
StD	1.000000000	1.000000000	1.000000000	1.000000000	1.000000000	1.000000000	1.000000000

	Prin1	Prin2	Prin3	Prin4	
MSSubClass	-.012534	0.102314	-.118510	0.015048	0.
MSZoning_new	-.070622	-.018901	0.099871	0.007644	-.
LotFrontage_new	0.001262	0.054605	-.041242	-.012138	-.
LotArea	0.064047	-.016077	0.245544	-.091762	-.
Alley_new	-.061713	0.103691	0.004549	0.015681	-.
Street_new	-.009871	-.036213	0.056489	-.044000	0.
LotShape_new	0.085242	-.023512	0.082367	-.022336	0.
LandContour_new	0.022553	-.040424	-.115241	0.084362	0.
Utilities_new	-.001080	0.008080	-.023461	-.008419	-.
LotConfig_new	-.016763	-.011583	-.096795	-.014855	0.

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	11.4737556	7.2889783	0.1452	0.1452
2	4.1847773	0.5819162	0.0530	0.1982
3	3.6028611	0.4056609	0.0456	0.2438
4	3.1972001	0.3172639	0.0405	0.2843
5	2.8799363	0.4703460	0.0365	0.3207
6	2.4095903	0.3786374	0.0305	0.3512
7	2.0309528	0.1155087	0.0257	0.3770
8	1.9154441	0.0779108	0.0242	0.4012
9	1.8375333	0.1118621	0.0233	0.4245
10	1.7256712	0.0144761	0.0218	0.4463
11	1.7111950	0.0652666	0.0217	0.4680
12	1.6459284	0.0181417	0.0208	0.4888
13	1.6277867	0.1338129	0.0206	0.5094
14	1.4939738	0.0974004	0.0189	0.5283
15	1.3965734	0.0105224	0.0177	0.5460
16	1.3860511	0.0296855	0.0175	0.5635
17	1.3563655	0.0842441	0.0172	0.5807
18	1.2721214	0.0429447	0.0161	0.5968
19	1.2291767	0.0374574	0.0156	0.6124
20	1.1917193	0.0870906	0.0151	0.6275
21	1.1046287	0.0233735	0.0140	0.6414
22	1.0812552	0.0223144	0.0137	0.6551
23	1.0589408	0.0297441	0.0134	0.6685
24	1.0291967	0.0182537	0.0130	0.6816
25	1.0109430	0.0118479	0.0128	0.6943
26	0.9990951	0.0296907	0.0126	0.7070
27	0.9694044	0.0344685	0.0123	0.7193
28	0.9349359	0.0201576	0.0118	0.7311
29	0.9147784	0.0094602	0.0116	0.7427
30	0.9053181	0.0186927	0.0115	0.7541
31	0.8866254	0.0578299	0.0112	0.7654
32	0.8287955	0.0019256	0.0105	0.7759
33	0.8268699	0.0200428	0.0105	0.7863
34	0.8068271	0.0139194	0.0102	0.7965
35	0.7929077	0.0208840	0.0100	0.8066
36	0.7720237	0.0203735	0.0098	0.8163
37	0.7516503	0.0224599	0.0095	0.8259
38	0.7291904	0.0166031	0.0092	0.8351

PCA

Scree Plot & Variance Explained for Principal Components

PCA

PCA with **n = 25**

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10
MSSubClass	-.012534	0.102314	-.118510	0.015048	0.403804	-.115279	0.010640	-.017283	0.241714	-.003300
MSZoning_new	-.070622	-.018901	0.099871	0.007644	-.008851	0.035530	-.031776	-.160667	-.042968	-.082578
LotFrontage_new	0.001262	0.054605	-.041242	-.012138	-.155149	0.007393	0.103975	-.098106	-.058931	0.220096
LotArea	0.064047	-.016077	0.245544	-.091762	-.055839	-.049907	-.164379	0.232787	0.039953	-.079495
Alley_new	-.061713	0.103691	0.004549	0.015681	-.014226	0.075726	0.042545	-.014201	0.141618	0.041147
Street_new	-.009871	-.036213	0.056489	-.044000	0.002621	-.043906	-.109342	0.220849	0.112661	0.001396
LotShape_new	0.085242	-.023512	0.082367	-.022336	0.021374	0.056048	-.136893	0.157903	-.062570	-.148017
LandContour_new	0.022553	-.040424	-.115241	0.084362	0.036122	0.015723	0.144247	-.224522	-.119212	0.143316
Utilities_new	-.001080	0.008080	-.023461	-.008419	-.004746	0.006303	0.024087	0.039141	-.053861	0.005083
LotConfig_new	-.016763	-.011583	-.096795	-.014855	0.009326	-.024210	0.001202	-.071134	0.047479	0.085180
LandSlope_new	0.000239	-.062023	0.158805	-.096159	0.022380	-.038720	-.256079	0.302473	0.132860	-.159215
Neighborhood_new	-.022478	0.032654	0.076390	0.066017	0.005088	-.007298	0.065786	0.013682	0.041133	0.211338
Condition1_new	-.003980	0.037057	0.085547	0.023712	-.020304	0.014318	0.042431	0.051398	-.023069	0.117657
Condition2_new	-.005318	0.017371	0.045384	-.002699	-.022717	-.001715	0.013392	0.015918	0.040790	0.088432
BldgType_new	0.005233	-.045713	-.212966	-.043055	0.269301	-.184024	0.028127	-.082974	0.247521	-.010737
HouseStyle_new	0.078185	0.105804	-.020424	0.118005	0.354303	0.035173	0.013397	0.055358	0.036999	0.002254
OverallQual	0.242947	0.079772	-.031197	-.043913	0.004991	0.067597	0.004346	-.078559	0.012223	-.030120
OverallCond	-.053363	-.048800	0.128721	0.102946	0.019720	0.307006	0.032337	-.109972	-.139655	-.002890
YearBuilt	0.217922	-.074140	-.224418	-.028376	0.077379	-.056655	0.000390	0.070612	-.029101	0.011306
YearRemodAdd	0.182858	0.019414	-.175699	-.052831	0.047974	0.182782	0.033878	0.072560	-.081442	0.072414
RoofStyle_new	0.044283	0.005401	0.097569	-.059200	-.082464	-.098198	0.009254	-.156498	-.106982	0.092881
RoofMatl_new	0.027939	0.005746	0.105982	-.052803	-.003968	-.050274	0.014716	-.014798	0.048194	0.014649
Exterior1st_new	0.057185	0.058036	-.102888	-.005362	-.136348	0.226775	0.002347	0.335948	-.099122	0.096029
Exterior2nd_new	0.034491	0.035671	-.065714	0.021559	-.179500	0.206682	-.014060	0.347475	-.132109	0.091290

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	11.4737556	7.2889783	0.1452	0.1452
2	4.1847773	0.5819162	0.0530	0.1982
3	3.6028611	0.4056609	0.0456	0.2438
4	3.1972001	0.3172639	0.0405	0.2843
5	2.8799363	0.4703460	0.0365	0.3207
6	2.4095903	0.3786374	0.0305	0.3512
7	2.0309528	0.1155087	0.0257	0.3770
8	1.9154441	0.0779108	0.0242	0.4012
9	1.8375333	0.1118621	0.0233	0.4245
10	1.7256712	0.0144761	0.0218	0.4463
11	1.7111950	0.0652666	0.0217	0.4680
12	1.6459284	0.0181417	0.0208	0.4888
13	1.6277867	0.1338129	0.0206	0.5094
14	1.4939738	0.0974004	0.0189	0.5283
15	1.3965734	0.0105224	0.0177	0.5460
16	1.3860511	0.0296855	0.0175	0.5635
17	1.3563655	0.0842441	0.0172	0.5807
18	1.2721214	0.0429447	0.0161	0.5968
19	1.2291767	0.0374574	0.0156	0.6124
20	1.1917193	0.0870906	0.0151	0.6275
21	1.1046287	0.0233735	0.0140	0.6414
22	1.0812552	0.0223144	0.0137	0.6551
23	1.0589408	0.0297441	0.0134	0.6685
24	1.0291967	0.0182537	0.0130	0.6816
25	1.0109430		0.0128	0.6943

PCA

Scree Plot & Variance Explained for n = 25 Principal Components

PCA

Principal Components Prin1 – Prin25 in Dataset

Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10
2.2855067522	0.3067089494	-1.636087346	0.7399614609	1.8999085802	0.3540620706	0.3898895596	0.8475128991	-1.674571775	1.4022920015
0.5264062571	-2.030363623	1.0132972055	-0.069688557	-0.462979081	0.76931645	-0.303279088	-0.580412746	-0.870727656	-0.909637618
2.9931028604	0.2229924844	-1.042064136	0.5399911135	1.5080157471	0.4988791279	-0.571137172	0.7918373965	-1.322231621	-0.23395751
-0.481375152	0.5861686739	1.5387276729	0.6656480147	0.185931096	0.8045993499	0.4613927354	-0.761662866	1.4316523126	-1.189112135
4.9797112373	1.4481944115	0.4572800663	0.3043194569	1.4231244384	0.1193375196	-0.722537445	0.742395725	-1.397276448	0.6311242879
-1.208938277	-2.808270597	0.2240761573	0.8408731974	-0.205500157	1.24054944	0.5701059719	1.7601674674	-2.141127714	0.3623021417
4.0771917256	-1.527340552	-0.02699496	-1.759450279	-0.627773919	-0.129198291	0.1481444863	0.0547065637	-1.371205036	0.9047188628
1.8944207368	0.159843476	3.8816983187	0.8377471551	2.4081059074	0.9294888944	-0.555042426	0.5824081714	-0.022023886	0.8187260642
-2.827232687	2.9458240973	1.7726089459	-0.000112859	-1.603352138	-2.780272202	-0.024320108	-2.449685332	2.2532384119	-0.423701121
-2.179805511	-1.909824135	0.7868536098	0.2604151228	1.4314081259	-2.085237016	0.1141980224	-0.61695102	0.1376309718	-0.088915964
-2.686407134	-2.051733667	-0.032932403	-0.039996409	-0.523007353	-0.753132913	0.749243917	-0.627990903	-1.000822544	0.9804773051
6.6392657371	2.2612484502	0.0356749523	-1.447009014	1.3059212181	-1.443883306	0.3637547549	-2.314119665	0.5221249689	-0.505649986
-2.977451406	-2.998025345	0.5991100995	0.0876195958	-0.026379518	-0.462167375	0.0774979553	-0.353161038	-0.631803847	-0.466020584
4.2518188995	1.0030782075	-2.284440321	-1.11009498	-3.76110981	0.1322476575	0.0220226101	0.1080676973	1.4095249887	-0.962608841
-0.684707691	-1.781767293	1.9132815896	0.0986905823	-0.163204829	-0.559858585	0.1980526706	-1.031026342	-0.957115277	-1.091755985
-2.97237036	-0.662807098	-0.487663506	1.4770320082	-2.324504095	2.2544144372	0.9852234084	-0.04431463	-0.152423406	-0.616684857
-0.191756872	-2.778170446	0.637198344	0.8889804088	-0.637813257	0.2885280739	-0.33923754	1.6934740608	-2.478769247	-0.563791435
-4.424212693	1.6716660423	-1.474387383	2.4587498838	-0.534512703	-6.149344034	1.4229698794	2.7919796134	-1.502679961	1.2598734471
-0.27741129	-1.100350909	-0.977425472	0.2678870403	-0.448325794	0.5764589411	0.8994374568	1.5096078147	-1.301735213	1.9728588175
-2.667648264	-0.962027186	0.2186903445	0.285481996	-1.8364646	-0.809903347	1.0368230195	-1.526357762	0.1730196137	0.3971705095

Regression Analysis on PCA

Run Multiple linear Regression Analysis on Prin1 – Prin25

Selection = MAXR

Multiple Linear Regression Anlysis on PCA

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	1460
Number of Observations Used	1460

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	25	7.38786E12	2.955144E11	232.83	<.0001
Error	1434	1.820051E12	1269212938		
Corrected Total	1459	9.207911E12			

Root MSE	35626	R-Square	0.8023
Dependent Mean	180921	Adj R-Sq	0.7989
Coeff Var	19.69145		

Regression Analysis on PCA

Parameter Estimates										
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS	Standardized Estimate	Squared Partial Corr Type I	Variance Inflation
Intercept	1	180921	932.37542	194.04	<.0001	4.778942E13	4.778942E13	0	.	0
Prin1	1	19771	275.35109	71.80	<.0001	6.543408E12	6.543408E12	0.84299	0.71063	1.00000
Prin2	1	5608.96828	455.93552	12.30	<.0001	1.920851E11	1.920851E11	0.14443	0.07209	1.00000
Prin3	1	5444.51771	491.37815	11.08	<.0001	1.558194E11	1.558194E11	0.13009	0.06302	1.00000
Prin4	1	-7266.93712	521.62054	-13.93	<.0001	2.46336E11	2.46336E11	-0.16356	0.10634	1.00000
Prin5	1	-435.87977	549.60182	-0.79	0.4279	798310043	798310043	-0.00931	0.00038561	1.00000
Prin6	1	1221.27981	600.85267	2.03	0.0423	5243591494	5243591494	0.02386	0.00253	1.00000
Prin7	1	-1795.20625	654.46990	-2.74	0.0062	9549570407	9549570407	-0.03220	0.00463	1.00000
Prin8	1	-1827.89171	673.91458	-2.71	0.0068	9337394516	9337394516	-0.03184	0.00454	1.00000
Prin9	1	-3237.96901	688.05315	-4.71	<.0001	28108384566	28108384566	-0.05525	0.01374	1.00000
Prin10	1	-269.03943	710.00364	-0.38	0.7048	182240632	182240632	-0.00445	0.00009034	1.00000
Prin11	1	2845.35197	713.00052	3.99	<.0001	20212814644	20212814644	0.04685	0.01002	1.00000
Prin12	1	-398.67567	726.99953	-0.55	0.5835	381685534	381685534	-0.00644	0.00019115	1.00000
Prin13	1	1539.78258	731.03952	2.11	0.0354	5630819139	5630819139	0.02473	0.00282	1.00000
Prin14	1	-4829.48920	763.07655	-6.33	<.0001	50839429590	50839429590	-0.07431	0.02554	1.00000
Prin15	1	1109.34881	789.23750	1.41	0.1601	2507582964	2507582964	0.01650	0.00129	1.00000
Prin16	1	5031.67577	792.22763	6.35	<.0001	51198803937	51198803937	0.07457	0.02643	1.00000
Prin17	1	-726.77050	800.85009	-0.91	0.3643	1045265488	1045265488	-0.01065	0.00055414	1.00000
Prin18	1	227.96668	826.94251	0.28	0.7828	96455417	96455417	0.00324	0.00005116	1.00000
Prin19	1	-3438.96831	841.26426	-4.09	<.0001	21209281836	21209281836	-0.04799	0.01125	1.00000
Prin20	1	-3570.73208	854.38304	-4.18	<.0001	22168881668	22168881668	-0.04907	0.01189	1.00000
Prin21	1	99.36373	887.42456	0.11	0.9109	15912096	15912096	0.00131	0.00000864	1.00000
Prin22	1	2772.94850	896.96500	3.09	0.0020	12130176123	12130176123	0.03630	0.00659	1.00000
Prin23	1	-1966.23002	906.36630	-2.17	0.0302	5973042952	5973042952	-0.02547	0.00326	1.00000
Prin24	1	184.03144	919.37015	0.20	0.8414	50855480	50855480	0.00235	0.00002789	1.00000
Prin25	1	-1547.24118	927.63315	-1.67	0.0955	3531002316	3531002316	-0.01958	0.00194	1.00000

Regression Analysis on PCA

Regression Analysis on PCA

Run Multiple linear Regression Analysis on Significant variables of Prin1 – Prin25

Multiple Regression Anlysis on PCA with significant variables

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	1460
Number of Observations Used	1460

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	16	7.379251E12	4.612032E11	363.94	<.0001
Error	1443	1.828661E12	1267263107		
Corrected Total	1459	9.207911E12			

Root MSE	35599	R-Square	0.8014
Dependent Mean	180921	Adj R-Sq	0.7992
Coeff Var	19.67632		

Regression Analysis on PCA

Parameter Estimates										
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS	Standardized Estimate	Squared Partial Corr Type I	Variance Inflation
Intercept	1	180921	931.65897	194.19	<.0001	4.778942E13	4.778942E13	0	.	0
Prin1	1	19771	275.13951	71.86	<.0001	6.543408E12	6.543408E12	0.84299	0.71063	1.00000
Prin2	1	5608.96828	455.58517	12.31	<.0001	1.920851E11	1.920851E11	0.14443	0.07209	1.00000
Prin3	1	5444.51771	491.00057	11.09	<.0001	1.558194E11	1.558194E11	0.13009	0.06302	1.00000
Prin4	1	-7266.93712	521.21972	-13.94	<.0001	2.46336E11	2.46336E11	-0.16356	0.10634	1.00000
Prin6	1	1221.27981	600.39096	2.03	0.0421	5243591494	5243591494	0.02386	0.00253	1.00000
Prin7	1	-1795.20625	653.96699	-2.75	0.0061	9549570407	9549570407	-0.03220	0.00462	1.00000
Prin8	1	-1827.89171	673.39673	-2.71	0.0067	9337394516	9337394516	-0.03184	0.00454	1.00000
Prin9	1	-3237.96901	687.52444	-4.71	<.0001	28108384566	28108384566	-0.05525	0.01374	1.00000
Prin11	1	2845.35197	712.45263	3.99	<.0001	20212814644	20212814644	0.04685	0.01002	1.00000
Prin13	1	1539.78258	730.47778	2.11	0.0352	5630819139	5630819139	0.02473	0.00282	1.00000
Prin14	1	-4829.48920	762.49019	-6.33	<.0001	50839429590	50839429590	-0.07431	0.02552	1.00000
Prin16	1	5031.67577	791.61887	6.36	<.0001	51198803937	51198803937	0.07457	0.02637	1.00000
Prin19	1	-3438.96831	840.61781	-4.09	<.0001	21209281836	21209281836	-0.04799	0.01122	1.00000
Prin20	1	-3570.73208	853.72651	-4.18	<.0001	22168881668	22168881668	-0.04907	0.01186	1.00000
Prin22	1	2772.94850	896.27575	3.09	0.0020	12130176123	12130176123	0.03630	0.00657	1.00000
Prin23	1	-1966.23002	905.66983	-2.17	0.0301	5973042952	5973042952	-0.02547	0.00326	1.00000

Eigenvectors

	Prin1	Prin2	Prin3
MSSubClass	-.012534	0.102314	-.118510
MSZoning_new	-.070622	-.018901	0.099871
LotFrontage_new	0.001262	0.054805	-.041242
LotArea	0.064047	-.016077	0.245544
Alley_new	-.061713	0.103691	0.004549
Street_new	-.009871	-.036213	0.056489
LotShape_new	0.085242	-.023512	0.082367
LandContour_new	0.022553	-.040424	-.115241
Utilities_new	-.001080	0.008080	-.023461
LotConfig_new	-.016763	-.011583	-.096795
LandSlope_new	0.000239	-.062023	0.158805
Neighborhood_new	-.022478	0.032654	0.076390
Condition1_new	-.003980	0.037057	0.085547
Condition2_new	-.005318	0.017371	0.045384
BldgType_new	0.005233	-.045713	-.212966
HouseStyle_new	0.078185	0.105804	-.020424
OverallQual	0.242947	0.079772	-.031197
OverallCond	-.053363	-.048800	0.128721
YearBuilt	0.217922	-.074140	-.224418
YearRemodAdd	0.182858	0.019414	-.175699
RoofStyle_new	0.044283	0.005401	0.097569
RoofMatl_new	0.027939	0.005746	0.105982
Exterior1st_new	0.057185	0.058036	-.102888
Exterior2nd_new	0.034491	0.035671	-.065714
MasVnrType_new	-.124848	0.016146	0.019211
MasVnrArea	0.143792	0.048619	0.032385
ExterQual_new	0.220827	0.051171	-.133819
ExterCond_new	0.006405	-.053317	0.091051
Foundation_new	0.147571	0.037967	-.209762
BsmtQual_new	0.195793	-.056403	-.064454
BsmtCond_new	0.092813	-.108879	0.036556
BsmtExposure_new	0.113641	-.141217	0.038650
BsmtFinType1_new	0.110961	-.250278	0.040764
BsmtFinSF1	0.118959	-.219213	0.156714
BsmtFinType2_new	0.002006	-.135231	0.164692
BsmtFinSF2	-.006204	-.119565	0.170222
BsmtUnfSF	0.074164	0.202536	-.114087
Heating_new	-.031037	0.044258	0.059616
HeatingQC_new	0.153962	0.024885	-.160439
CentralAir_new	0.119134	-.134975	-.017799
Electrical_new	-.103246	0.082953	0.006904
_1stFlrSF	0.177276	-.008956	0.169819
_2ndFlrSF	0.072851	0.329842	0.091914
LowQualFinSF	-.029220	0.096167	0.074664
BsmtFullBath	0.072975	-.234834	0.107989
BsmtHalfBath	-.002612	-.044556	0.065856
FullBath	0.178931	0.219270	-.032102
HalfBath	0.092109	0.155871	0.024995
BedroomAbvGr	0.037447	0.285599	0.187875
KitchenAbvGr	-.055123	0.170597	0.017460
KitchenQual_new	0.211135	0.024634	-.092565
TotRmsAbvGrd	0.136589	0.320478	0.184372
Functional_new	0.037828	-.051775	-.075406
Fireplaces	0.141722	0.027728	0.224428
FireplaceQu_new	0.160636	0.052924	0.155199
GarageType_new	0.184756	-.130030	0.004856
GarageYrBlt	0.146187	-.136632	0.041897
GarageFinish_new	0.217143	-.051763	-.044321
GarageCars	0.230013	0.006497	-.004935
GarageArea	0.221247	-.015510	0.032443
GarageQual_new	0.146684	-.145139	0.053345
GarageCond_new	0.145970	-.150251	0.050988
PavedDrive_new	0.073404	-.116852	0.012339
WoodDeckSF	0.103938	-.040144	0.091186
OpenPorchSF	0.102786	0.094556	0.019248
EnclosedPorch	-.069181	0.067543	0.115867
_3SsnPorch	0.014204	-.016376	0.001956
ScreenPorch	0.022959	-.010490	0.113550
PoolArea	0.028650	0.024080	0.166136
PoolQC_new	0.032840	0.029158	0.170062
MiscVal	-.011183	-.000180	0.061516
MoSold	0.015370	0.031965	-.005066
YrSold	-.011048	-.037183	0.008452
SaleType_new	0.107092	0.067515	-.163220
SaleCondition_new	-.072081	-.066094	0.135832
Fence_new	-.049803	-.051084	0.148156
MiscFeature_new	-.020745	-.017041	0.096383
TotalBsmtSF	0.196092	-.067874	0.110611
GrLivArea	0.188234	0.276317	0.208197

Regression Analysis on PCA

Compare: Regression Analysis

Regression Analysis on PCA

Regression Analysis on PCA

Regression Analysis on PCA

Compare: Regression Analysis

Regression Analysis on PCA

Conclusion

- Data exploration and preparation and takes up to 60% of the effort.
- House price can most predicted by OverallQaul and GrLivArea variables.
- Residual are almost normally distributed and randomly spread in Multiple Linear Regression in both with normal variables and PCA.
- Multiple Linear Regression Model in both with normal variables and with Principal Components are almost same.

Thank You