Assignment #1

Nate Bitting

Introduction:

The purpose of this assignment is to identify 12 variables of the Ames Housing dataset to be used to train a linear regression model to predict the SalePrice for a home in the Ames, Iowa area. Before selecting which 12 variables to use, exploratory data analysis will be performed to understand the relationships that exist in the data and which variables will be most useful in predicting SalePrice.

Results:

In order to first understand each of the variables in the Ames Housing dataset, we must show the basic summary statistics to identify any issues in the data itself. As you can see below, some of the variables appear to be useful, while others might not have enough data available to be useful in a model building exercise. However, further exploratory data analysis must be performed to know for sure.

Summary Statistics

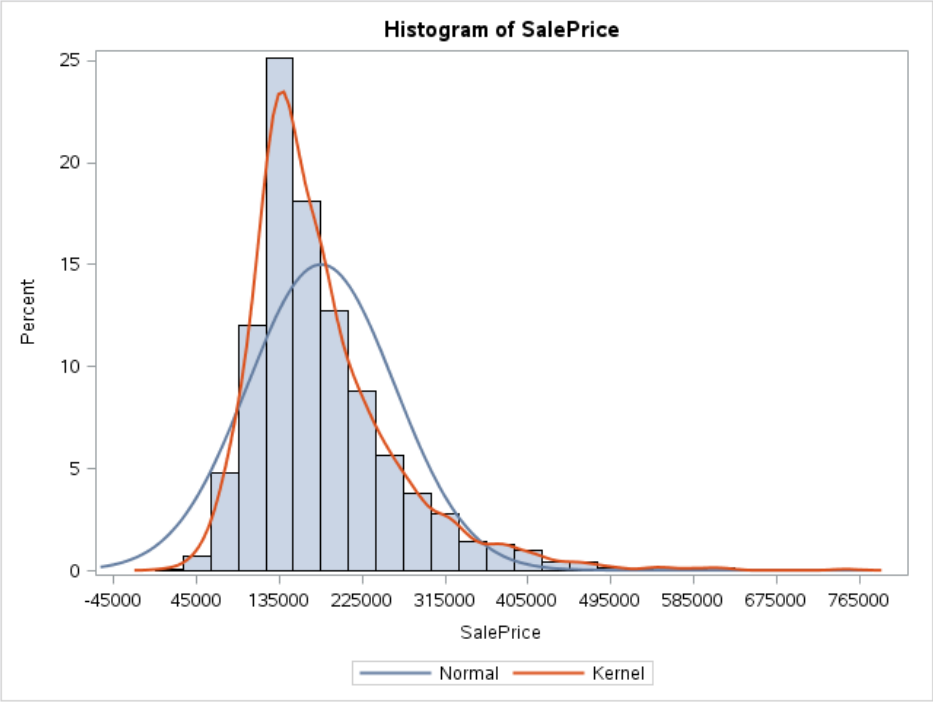
Variable	Minimum	Mean	Median	Mode	Maximum	Std Dev	N	Range
SID	1.0000000	1465.50	1465.50	.	2930.00	845.9624696	2930	2929.00
PID	526301100	714464497	535453620	.	1007100110	188730845	2930	480799010
SubClass	20.0000000	57.3873720	50.0000000	20.0000000	190.0000000	42.6380246	2930	170.0000000
LotFrontage	21.0000000	69.2245902	68.0000000	60.0000000	313.0000000	23.3653350	2440	292.0000000
LotArea	1300.00	10147.92	9436.50	9600.00	215245.00	7880.02	2930	213945.00
OverallQual	1.0000000	6.0948805	6.0000000	5.0000000	10.0000000	1.4110261	2930	9.0000000
OverallCond	1.0000000	5.5631399	5.0000000	5.0000000	9.0000000	1.1115366	2930	8.0000000
YearBuilt	1872.00	1971.36	1973.00	2005.00	2010.00	30.2453606	2930	138.0000000
YearRemodel	1950.00	1984.27	1993.00	1950.00	2010.00	20.8602859	2930	60.0000000
MasVnrArea	0	101.8968008	0	0	1600.00	179.1126106	2907	1600.00
BsmtFinSF1	0	442.6295664	370.0000000	0	5644.00	455.5908391	2929	5644.00
BsmtFinSF2	0	49.7224309	0	0	1526.00	169.1684756	2929	1526.00
BsmtUnfSF	0	559.2625469	466.0000000	0	2336.00	439.4941528	2929	2336.00
TotalBsmtSF	0	1051.61	990.0000000	0	6110.00	440.6150670	2929	6110.00
FirstFlrSF	334.0000000	1159.56	1084.00	864.0000000	5095.00	391.8908853	2930	4761.00
SecondFlrSF	0	335.4559727	0	0	2065.00	428.3957150	2930	2065.00
LowQualFinSF	0	4.6767918	0	0	1064.00	46.3105100	2930	1064.00
GrLivArea	334.0000000	1499.69	1442.00	864.0000000	5642.00	505.5088875	2930	5308.00
BsmtFullBath	0	0.4313525	0	0	3.0000000	0.5248202	2928	3.0000000
BsmtHalfBath	0	0.0611339	0	0	2.0000000	0.2452536	2928	2.0000000
FullBath	0	1.5665529	2.0000000	2.0000000	4.0000000	0.5529406	2930	4.0000000
HalfBath	0	0.3795222	0	0	2.0000000	0.5026293	2930	2.0000000
BedroomAbvGr	0	2.8542662	3.0000000	3.0000000	8.0000000	0.8277311	2930	8.0000000
KitchenAbvGr	0	1.0443686	1.0000000	1.0000000	3.0000000	0.2140762	2930	3.0000000
TotRmsAbvGrd	2.0000000	6.4430034	6.0000000	6.0000000	15.0000000	1.5729644	2930	13.0000000
Fireplaces	0	0.5993174	1.0000000	0	4.0000000	0.6479209	2930	4.0000000
GarageYrBlt	1895.00	1978.13	1979.00	2005.00	2207.00	25.5284113	2771	312.0000000
GarageCars	0	1.7668146	2.0000000	2.0000000	5.0000000	0.7605664	2929	5.0000000
GarageArea	0	472.8197337	480.0000000	0	1488.00	215.0465485	2929	1488.00
WoodDeckSF	0	93.7518771	0	0	1424.00	126.3615619	2930	1424.00
OpenPorchSF	0	47.5334471	27.0000000	0	742.0000000	67.4834001	2930	742.0000000
EnclosedPorch	0	23.0116041	0	0	1012.00	64.1390592	2930	1012.00
ThreeSsnPorch	0	2.5924915	0	0	508.0000000	25.1413310	2930	508.0000000
ScreenPorch	0	16.0020478	0	0	576.0000000	56.0873702	2930	576.0000000
PoolArea	0	2.2433447	0	0	800.0000000	35.5971806	2930	800.0000000
MiscVal	0	50.6351536	0	0	17000.00	566.3442883	2930	17000.00
MoSold	1.0000000	6.2160410	6.0000000	6.0000000	12.0000000	2.7144924	2930	11.0000000
YrSold	2006.00	2007.79	2008.00	2007.00	2010.00	1.3166129	2930	4.0000000
SalePrice	12789.00	180796.06	160000.00	135000.00	755000.00	79886.69	2930	742211.00

When working with familiar data, it is always good to first explore variables you know have an impact on the target variable in question. As with any housing pricing analysis there are some common variables used to determine the sale price:

1. Neighborhood information (Neighborhood)
2. # of bedrooms (BedroomAbvGr)
3. # of full bathrooms (FullBath)

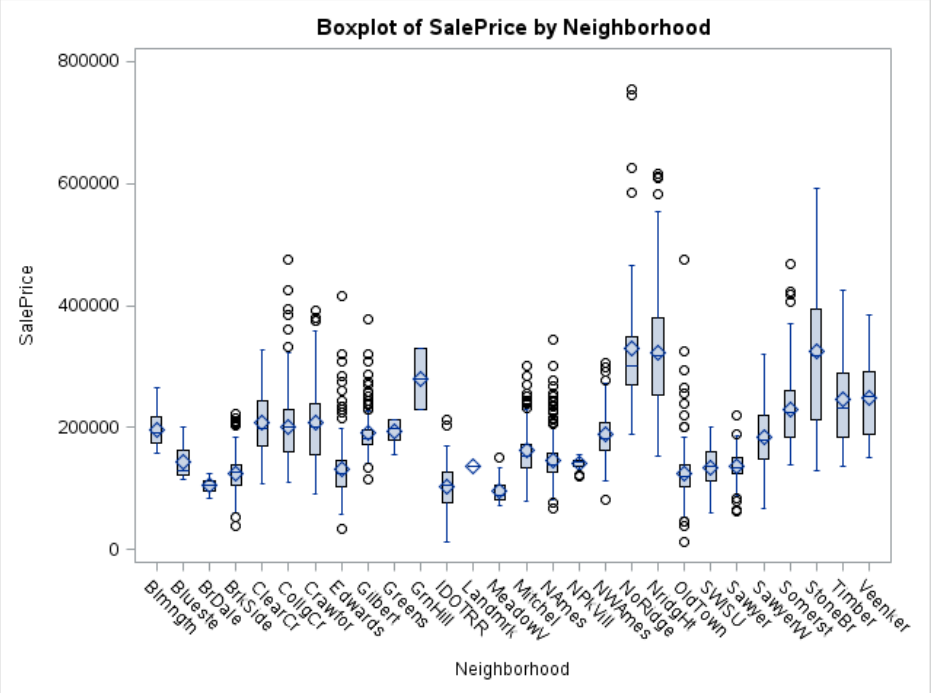
4. # of half bathrooms (HalfBath)
5. Month Sold (MoSold)
6. Garage size (GarageCars)
7. Square Footage (GrLivArea)
8. Lot size (LotArea)
9. Basement size (TotalBsmtSF)
10. Heating information (Heating)
11. Cooling information (CentralAir)
12. Building Type (BldgType)

Before we decide these are the correct variable to use in predicting sale price, let us first explore each of these variables to see if there is any relationship with SalePrice and identify if we have any data quality issues that may exist. Let us first look at the sale price frequency distribution leveraging a histogram and density plot (right).



It appears there is a high number of homes in the $135K bin, which is consistent with the original summary statistics output above. The next step is to begin identifying any relationships that may exist between SalePrice and the explanatory variables we have selected to explore above.

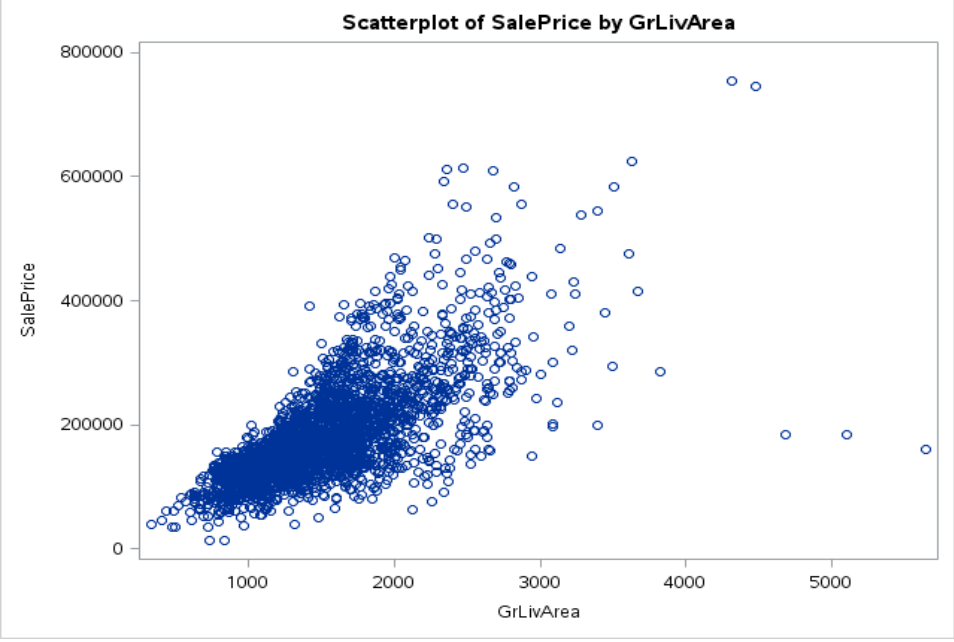
We can first look at the sale price ranges by neighborhood leveraging the boxplot chart to the right. It is apparent that some outliers exist in the dataset, particularly in the Northridge area of Ames. If we were to develop a predictive model we would have to consider removing some of the outliers or performing some sort of factor analysis or principal component analysis to handle some of these extreme observations.



As you can see in the neighborhood frequency table to the left, a majority of the homes in the Ames housing dataset are from the North Ames area. The other areas with a significant number of homes include College Creek, Old Town, Edwards, and Somerset. When looking at the SalePrice in each of these areas, you can see that Northridge, Northridge Heights, and Stone Brook have the highest SalePrice compared to other areas. This tells me that there is a strong relationship between Neighborhood and SalePrice and would be relevant for including in a predictive model.

Neighborhood	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Blmngtn	28	0.96	28	0.96
Blueste	10	0.34	38	1.30
BrDale	30	1.02	68	2.32
BrkSide	108	3.69	176	6.01
ClearCr	44	1.50	220	7.51
CollgCr	267	9.11	487	16.62
Crawfor	103	3.52	590	20.14
Edwards	194	6.62	784	26.76
Gilbert	165	5.63	949	32.39
Greens	8	0.27	957	32.66
GrnHill	2	0.07	959	32.73
IDOTRR	93	3.17	1052	35.90
Landmrk	1	0.03	1053	35.94
MeadowV	37	1.26	1090	37.20
Mitchel	114	3.89	1204	41.09
NAmes	443	15.12	1647	56.21
NPkVill	23	0.78	1670	57.00
NWAmes	131	4.47	1801	61.47
NoRidge	71	2.42	1872	63.89
NridgHt	166	5.67	2038	69.56
OldTown	239	8.16	2277	77.71
SWISU	48	1.64	2325	79.35
Sawyer	151	5.15	2476	84.51
SawyerW	125	4.27	2601	88.77
Somerst	182	6.21	2783	94.98
StoneBr	51	1.74	2834	96.72
Timber	72	2.46	2906	99.18
Veenker	24	0.82	2930	100.00

Now that we have explored the neighborhoods of the Ames area, let us now begin identifying any relationships between the other variables in the dataset. As with most housing datasets, it is common practice to understand the relationship between living area and sale price, which is depicted in the scatterplot to the right. There is a very clear linear relationship between these two variables. By coupling Neighborhood and GrLivingArea, we may have a better prediction of SalePrice; something we can explore when we build the regression model.



Next, we can explore if there is any correlation between some of the continuous variables and SalePrice through the use of the PROC CORR procedure as shown to the right. As you can see, there is a strong correlation between GrLivArea and SalePrice; same goes for TotalBsmtSF and SalePrice as well. However, it doesn't appear to be a strong relationship between LotArea and SalePrice as one would expect.

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations			
	GrLivArea	LotArea	TotalBsmtSF
SalePrice	0.70678 <.0001 2930	0.26655 <.0001 2930	0.63228 <.0001 2929

Now it would be good to explore the frequency tables for each of the categorical variables we are interested in to predict SalePrice to determine if they would be useful for inclusion in a regression model.

Frequency by Number of Bedrooms

Most houses in the Ames area appear to have between 2-4 bedrooms, with the most having 3, representing 55% of the total houses in the dataset.

BedroomAbvGr	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	8	0.27	8	0.27
1	112	3.82	120	4.10
2	743	25.36	863	29.45
3	1597	54.51	2460	83.96
4	400	13.65	2860	97.61
5	48	1.64	2908	99.25
6	21	0.72	2929	99.97
8	1	0.03	2930	100.00

Frequency by Number of Full Bathrooms

Most houses in the Ames area appear to have between 1-2 full bathrooms, with the most having 2, representing 52% of the total houses in the dataset

FullBath	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	12	0.41	12	0.41
1	1318	44.98	1330	45.39
2	1532	52.29	2862	97.68
3	64	2.18	2926	99.86
4	4	0.14	2930	100.00

Frequency by Number of Half Bathrooms

Most houses in the Ames area appear to not have any half baths (63%) while only 36% have at least one half bath

HalfBath	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	1843	62.90	1843	62.90
1	1062	36.25	2905	99.15
2	25	0.85	2930	100.00

Frequency by Garage Size

The majority (55%) of homes in the Ames area have a 2 car garage with 26% of them having a 1 car garage.

GarageCars	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	157	5.36	157	5.36
1	778	26.56	935	31.92
2	1603	54.73	2538	86.65
3	374	12.77	2912	99.42
4	16	0.55	2928	99.97
5	1	0.03	2929	100.00
Frequency Missing = 1				

Frequency by Heating Type

As can be seen in the frequency table to the right, 98% of all homes in the Ames area has a Gas Furnace with Forced Air. This result would indicate it most likely will not be a very reliable explanatory variable in a regression model.

Heating	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Floo	1	0.03	1	0.03
GasA	2885	98.46	2886	98.50
GasW	27	0.92	2913	99.42
Grav	9	0.31	2922	99.73
OthW	2	0.07	2924	99.80
Wall	6	0.20	2930	100.00

Frequency by Central Air (Yes or No)

It appears almost all homes in the area have central air installed (93%). This variable might also not be a very good indicator of SalePrice and we should consider excluding it from a regression model.

CentralAir	Frequency	Percent	Cumulative Frequency	Cumulative Percent
N	196	6.69	196	6.69
Y	2734	93.31	2930	100.00

Frequency by Building Type

It appears almost all homes in the area have central air installed (93%). This variable might also not be a very good indicator of SalePrice and we should consider excluding it from a regression model.

BldgType	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1Fam	2425	82.76	2425	82.76
2fmCon	62	2.12	2487	84.88
Duplex	109	3.72	2596	88.60
Twnhs	101	3.45	2697	92.05
TwnhsE	233	7.95	2930	100.00

Frequency by Month Sold

Based on the frequency distribution in the table to the right, you can clearly see that the bulk of the homes are sold in the summary months, particularly in May, June, and July. Given there is a cluster of the frequencies, month may have an impact on the prediction of SalePrice.

MoSold	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	123	4.20	123	4.20
2	133	4.54	256	8.74
3	232	7.92	488	16.66
4	279	9.52	767	26.18
5	395	13.48	1162	39.66
6	505	17.24	1667	56.89
7	449	15.32	2116	72.22
8	233	7.95	2349	80.17
9	161	5.49	2510	85.67
10	173	5.90	2683	91.57
11	143	4.88	2826	96.45
12	104	3.55	2930	100.00

Conclusions:

The exploratory data analysis conducted above gives us a much better understanding of the Ames housing dataset. Initially, we used variables we *assumed* would be useful in the prediction of SalePrice, but we were able to identify a few variables that have little to no relationship with SalePrice. Given the dataset is relatively small, this process of manual exploration is practical, however as the number of explanatory variables increases, we would need to consider other means of feature selection for a regression model. Some possible options are explanatory factor analysis and principal component analysis.

Code:

```
* Code used to get the data into my library;
ods graphics on;
libname mydata '/courses/d6fc9ae5ba27fe300/c_3505/SAS_Data/' access=readonly;
proc datasets library=mydata; run; quit;
proc print data=mydata.anscombe; run; quit;

* put data into a unique dataset called 'building';
Data building;
SET mydata.ames_housing_data;
run; quit;


* Perform summary statistics on all numerical columns of data;
proc means data=building min mean median mode max stddev n range;
    title "Summary Statistics"
run; quit;

* histogram and density plot of saleprice';
proc sgplot data=building;
    title "Histogram of SalePrice";
    histogram SalePrice / showbins;
    density SalePrice;
    density SalePrice / type=kernel;
    run; quit;

* boxplot of saleprice by neighborhood';
proc sgplot data=building;
    title "Boxplot of SalePrice by Neighborhood";
    vbox SalePrice / category=Neighborhood;
    run; quit;

* Frequency table by Neighborhood;
proc freq data=building;
    tables Neighborhood;
    run; quit;

* scatterplot saleprice by GrLivArea'/*;
Proc sgplot data=building;
    title "Scatterplot of SalePrice by GrLivArea";
    scatter x = GrLivArea y = SalePrice;
    run; quit;

* Correlation analysis for saleprice;
Proc corr data=building;
    var GrLivArea LotArea TotalBsmtSF;
    with SalePrice;
    title "Correlation Analysis for Saleprice";
    run; quit;

* Frequency table by # of Bedrooms;
proc freq data=building;
    tables BedroomAbvGr;
    run; quit;
```

Code continued:

```
* Frequency table by # of Full Bathrooms;
proc freq data=building;
    tables FullBath;
    run; quit;

* Frequency table by # of Half Bathrooms;
proc freq data=building;
    tables HalfBath;
    run; quit;

* Frequency table by Garage Size;
proc freq data=building;
    tables GarageCars;
    run; quit;

* Frequency table by Heating Type;
proc freq data=building;
    tables Heating;
    run; quit;

* Frequency table by CentralAir;
proc freq data=building;
    tables CentralAir;
    run; quit;

* Frequency table by Building Type;
proc freq data=building;
    tables BldgType;
    run; quit;

* Frequency table by Month Sold;
proc freq data=building;
    tables MoSold;
    run; quit;
```