410-57, Assignment #1
Anamitra Bhattacharyya

INTRODUCTION

This is assignment is part of an overall project to build statistical regression models related to modeling sales prices of dwellings in Ames, Iowa. This assignment specifically deals with performing an exploratory data analysis (EDA) of an observational data set from the Ames Assessor's Office used in obtaining values for individual residential properties sold in Ames, Iowa between 2006 and 2010. As described in the data dictionary, the Ames data set comprises 2930 observations (properties sold) with 82 variables, comprising: 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables (with 2 additional observational identifiers). The goal of this EDA assignment is to evaluate and identify potential variables that are predictors of sales prices in the Ames housing market. The overall steps in the assignment comprise:

a) Data Survey
b) Data Quality Check
c) Initial Exploratory Data Analysis

RESULTS

1. Examining Variables in Ames Housing dataset

For the purposes of this assignment, a selection of 20 variables were chosen as potential predictors, from prior knowledge and experience

1. Continuous and discrete variables:
   (a) 5 continuous variables chosen: TotalFlrSF, GrLivArea, BsmtFinSF1, WoodDeckSF, SalePrice.
   (b) 4 Discrete variables chosen: houseage, FullBath, BedroomAbvGr, GarageCars.
2. Categorical variables, 11 were chosen: Zoning KitchenQual OverallQual CentralAir Alley Condition1 BldgType Heating Electrical Utilities RoofMatl.

Included derived variables TotalFlrSF (sum of FirstFlrSF and SecondFlrSF) and houseage (YrSold - YearBuilt).

2. Using PROC SORT procedure

a) PROC SORT by saleprice

2 observations above $700,000 (outliers, see upper table below)
2 observations below $20,000 (outliers, see lower table below)
Looking at all 2930 SalePrice observations indicates they all have a sale price associated with them.

Obs	Zoning	Alley	Utilities	Condition1	BldgType	OverallQual	RoofStyle	BsmtFinSF1	Heating	CentralAir	Electrical	GrLivArea	FullBath	BedroomAbvGr	KitchenQual	GarageCars	WoodDeckSF	SalePrice	TotalFlrSF	houseage
1	RL	NA	AllPub	Norm	1Fam	10	Gable	1455	GasA	Y	SBrkr	4316	3	4	Ex	3	382	755000	4316	13
2	RL	NA	AllPub	Norm	1Fam	10	Hip	2096	GasA	Y	SBrkr	4476	3	4	Ex	3	171	745000	4476	11
3	RL	NA	AllPub	Norm	1Fam	10	Hip	1387	GasA	Y	SBrkr	3627	3	4	Gd	3	361	625000	3627	11
4	RL	NA	AllPub	Norm	1Fam	10	Hip	2257	GasA	Y	SBrkr	2470	1	1	Ex	3	154	615000	2470	5
5	RL	NA	AllPub	Norm	1Fam	9	Hip	2188	GasA	Y	SBrkr	2364	2	2	Ex	3	0	611657	2364	1
6	RL	NA	AllPub	PosA	1Fam	10	Hip	2288	GasA	Y	SBrkr	2674	2	2	Ex	3	360	610000	2674	2
7	RL	NA	AllPub	PosN	1Fam	9	Hip	1101	GasA	Y	SBrkr	2338	2	4	Gd	3	0	591587	2338	1
8	RL	NA	AllPub	Norm	1Fam	9	Hip	292	GasA	Y	SBrkr	3500	3	4	Ex	3	870	584500	3500	13
9	RL	NA	AllPub	Norm	1Fam	9	Hip	0	GasA	Y	SBrkr	2822	3	4	Ex	3	52	582933	2822	1
10	RL	NA	AllPub	Norm	1Fam	9	Hip	240	GasA	Y	SBrkr	2868	3	4	Ex	3	214	556581	2868	1

Obs	Zoning	Alley	Utilities	Condition1	BldgType	OverallQual	RoofStyle	BsmtFinSF1	Heating	CentralAir	Electrical	GrLivArea	FullBath	BedroomAbvGr	KitchenQual	GarageCars	WoodDeckSF	SalePrice	TotalFlrSF	houseage
1	RM	NA	AllPub	Norm	1Fam	2	Gable	0	GasA	N	SBrkr	832	1	2	TA	2	0	12789	832	87
2	A	NA	AllPub	Norm	1Fam	1	Gable	0	Wall	N	FuseA	733	1	2	Fa	2	0	13100	733	56
3	C	NA	AllPub	Norm	1Fam	4	Gable	495	GasA	N	FuseA	720	1	2	TA	0	0	34900	720	89
4	RL	Gr	AllPub	Feedr	1Fam	2	Gable	0	GasA	N	FuseF	498	1	1	TA	1	0	35000	498	84
5	C	NA	AllPub	Norm	1Fam	2	Gable	50	GasA	N	FuseA	480	0	1	TA	1	0	35311	480	57
6	RM	NA	AllPub	Norm	1Fam	3	Gable	0	Grav	N	SBrkr	968	1	2	TA	0	0	37900	968	99
7	RL	NA	AllPub	Norm	1Fam	1	Gable	0	GasA	N	FuseF	334	1	1	Fa	0	0	39300	334	61
8	C	Pa	AllPub	Feedr	1Fam	4	Gambre	0	GasA	N	SBrkr	1317	1	3	TA	1	0	40000	1317	88
9	C	NA	AllPub	Norm	1Fam	3	Mansar	0	GasA	N	SBrkr	797	1	2	TA	0	0	44000	797	109
10	RM	NA	AllPub	Norm	1Fam	2	Gable	0	GasA	N	FuseA	612	1	1	TA	1	0	45000	612	69

b) PROC SORT by houseage

Ascending sort reveals a house with a -1 year's age (which is an erroneous value). While there are numerous houses of zero age, these are likely new construction that were likely built and sold in the same year.

Obs	Zoning	Alley	Utilities	Condition1	BldgType	OverallQual	RoofStyle	BsmtFinSF1	Heating	CentralAir	Electrical	GrLivArea	FullBath	BedroomAbvGr	KitchenQual	GarageCars	WoodDeckSF	SalePrice	TotalFlrSF	houseage
1	RM	Pa	AllPub	Norm	1Fam	8	Gable	259	OthW	N	SBrkr	2358	2	4	TA	0	0	122000	2153	136
2	RL	NA	AllPub	Feedr	1Fam	5	Gable	0	GasA	N	SBrkr	1020	1	2	TA	0	0	94000	1020	135
3	RM	Pa	AllPub	Norm	1Fam	7	Mansar	0	GasW	N	SBrkr	2640	1	4	Gd	4	181	265979	2640	129
4	RL	NA	AllPub	Feedr	1Fam	5	Gable	0	GasA	Y	SBrkr	2016	1	4	TA	2	0	131000	2016	129
5	RM	NA	AllPub	Norm	1Fam	7	Gable	0	GasA	Y	SBrkr	3493	3	3	Gd	3	302	295000	3493	128
6	RM	Gr	AllPub	Norm	1Fam	7	Gable	0	GasA	Y	FuseA	2454	2	3	TA	2	0	185000	2454	128
7	RM	Gr	AllPub	Artery	1Fam	8	Gable	216	GasA	Y	SBrkr	1742	1	4	Gd	2	0	168000	1742	127
8	RM	NA	AllPub	Norm	1Fam	6	Gable	0	GasW	Y	SBrkr	2210	2	5	Fa	1	0	117500	2210	127
9	RM	Gr	AllPub	Artery	1Fam	5	Gable	0	GasA	Y	FuseA	1750	1	3	Ex	1	0	124000	1750	126
10	RM	NA	AllPub	Norm	2fmCon	4	Gable	0	GasA	Y	SBrkr	2290	2	4	TA	2	0	122500	2290	125

Obs	Zoning	Alley	Utilities	Condition1	BldgType	OverallQual	RoofStyle	BsmtFinSF1	Heating	CentralAir	Electrical	GrLivArea	FullBath	BedroomAbvGr	KitchenQual	GarageCars	WoodDeckSF	SalePrice	TotalFlrSF	houseage
1	RL	NA	AllPub	Norm	1Fam	10	Hip	4010	GasA	Y	SBrkr	5095	2	2	Ex	3	546	183850	5095	-1
2	RL	NA	AllPub	Norm	1Fam	9	Hip	1445	GasA	Y	SBrkr	1856	1	1	Ex	3	113	394432	1856	0
3	FV	Pa	AllPub	Norm	1Fam	8	Gable	1032	GasA	Y	SBrkr	1418	1	1	Gd	3	160	267916	1418	0
4	RL	NA	AllPub	Norm	1Fam	7	Gable	24	GasA	Y	SBrkr	1222	2	2	Gd	2	0	187000	1222	0
5	FV	NA	AllPub	RRAn	1Fam	8	Gable	544	GasA	Y	SBrkr	1935	2	3	TA	2	0	263435	1935	0
6	RL	NA	AllPub	Norm	1Fam	9	Gable	986	GasA	Y	SBrkr	2690	2	3	Ex	3	0	398800	2690	0
7	RL	NA	AllPub	Norm	1Fam	10	Hip	1436	GasA	Y	SBrkr	2020	2	3	Ex	3	156	402861	2020	0
8	RL	NA	AllPub	Norm	1Fam	7	Gable	0	GasA	Y	SBrkr	1502	2	3	Gd	2	0	233170	1502	0
9	FV	NA	AllPub	Norm	1Fam	7	Gable	822	GasA	Y	SBrkr	1852	2	3	Gd	2	168	252678	1852	0
10	FV	NA	AllPub	Norm	1Fam	6	Gable	27	GasA	Y	SBrkr	1218	2	2	Gd	2	0	208300	1218	0

c) PROC SORT by BedroomAbvGr

Ascending reveals some houses with zero bedrooms above ground (see lower output table below), these are either spurious and are errors, or alternatively have the bedrooms in the basement (possibly a ranch style). Using descending sort there is also one house with 8 bedrooms from a two-family conversion (see lower output table), but this is an outlier.

Obs	Zoning	Alley	Utilities	Condition1	BldgType	OverallQual	RoofStyle	BsmtFinSF1	Heating	CentralAir	Electrical	GrLivArea	FullBath	BedroomAbvGr	KitchenQual	GarageCars	WoodDeckSF	SalePrice	TotalFlrSF	houseage
1	RH	Pa	AllPub	Feedr	2fmCon	6	Hip	256	GasA	Y	FuseA	3395	2	8	Fa	0	0	200000	2880	93
2	RL	NA	AllPub	Norm	Duplex	5	Hip	1500	GasA	Y	SBrkr	1728	2	6	TA	0	0	84900	1728	48
3	RH	Pa	AllPub	Norm	Duplex	4	Flat	0	GasA	Y	SBrkr	2650	3	6	TA	0	0	160000	2650	43
4	RL	NA	AllPub	Artery	Duplex	6	Gable	0	GasA	Y	SBrkr	2544	2	6	TA	3	0	190000	2544	40
5	RL	NA	AllPub	Artery	Duplex	5	Gable	500	GasA	Y	SBrkr	2634	2	6	TA	4	0	200000	2634	40
6	RM	NA	AllPub	Feedr	2fmCon	7	Gable	169	GasA	N	FuseF	1864	2	6	TA	0	0	97500	1864	107
7	RL	NA	AllPub	Feedr	Duplex	5	Gable	0	GasA	Y	SBrkr	2228	2	6	TA	2	73	147983	2228	30
8	RL	NA	AllPub	Feedr	Duplex	5	Gable	0	GasA	Y	SBrkr	2240	2	6	TA	2	154	142953	2240	30
9	RL	NA	AllPub	Norm	Duplex	7	Hip	820	GasA	Y	SBrkr	2787	4	6	TA	4	312	269500	2787	9
10	RL	NA	AllPub	Feedr	Duplex	7	Hip	820	GasA	Y	SBrkr	2787	4	6	TA	4	312	269500	2787	9

Obs	Zoning	Alley	Utilities	Condition1	BldgType	OverallQual	RoofStyle	BsmtFinSF1	Heating	CentralAir	Electrical	GrLivArea	FullBath	BedroomAbvGr	KitchenQual	GarageCars	WoodDeckSF	SalePrice	TotalFlrSF	houseage
1	RL	NA	AllPub	Norm	Duplex	6	Gable	1056	GasA	Y	SBrkr	1056	0	0	TA	2	264	144000	1056	30
2	RL	NA	AllPub	Norm	1Fam	7	Shed	1258	GasA	Y	SBrkr	1524	0	0	Gd	2	268	260000	1524	31
3	RL	NA	AllPub	Norm	TwnhsE	8	Gable	1153	GasA	Y	SBrkr	1593	1	0	Ex	2	0	286000	1593	7
4	RL	NA	AllPub	Feedr	Duplex	4	Gable	1198	GasA	Y	SBrkr	1258	0	0	TA	2	120	108959	1258	39
5	RL	NA	AllPub	Norm	1Fam	8	Shed	51	GasA	Y	SBrkr	1743	0	0	Gd	2	646	279000	1743	31
6	RM	NA	AllPub	Norm	TwnhsE	6	Gable	16	GasA	Y	SBrkr	936	0	0	TA	2	0	140000	936	11
7	RL	NA	AllPub	Norm	1Fam	9	Gable	1810	GasA	Y	SBrkr	1842	0	0	Gd	3	857	385000	1842	25
8	RL	NA	AllPub	Norm	1Fam	4	Gable	648	GasA	Y	SBrkr	960	0	0	TA	1	88	145000	960	41
9	RL	NA	AllPub	Norm	TwnhsE	8	Gable	368	GasA	Y	SBrkr	1502	1	1	Gd	2	0	212000	1502	25
10	RL	NA	AllPub	Norm	1Fam	9	Hip	1445	GasA	Y	SBrkr	1856	1	1	Ex	3	113	394432	1856	0

Conclusion: We have a reasonable set of variables with which to predict the sales prices including certain variables that were created such at TotalFlSF and houseage. These additional variables were created to provide a better idea of overall square footage of homes and ages of homes. The sales prices in the entire dataset were reviewed and there were none that were zero or negative prices. However, there are outliers in the sales prices of residences in the upper and lower sales range. The houseage data was reviewed and revealed an error (*e.g.* a value of -1 from one residence), while others had a zero age, mostly likely brand new homes. The BedroomAbvGr (number of bedrooms above ground) revealed a house with 8 bedrooms, which is an outlier, arising from a 2-family duplex conversion.

3. Correlation

Numeric variables (10): OverallQual, BsmtFinSF1, GrLivArea, FullBath, BedroomAbvGr, GarageCars, WoodDeckSF, SalePrice, TotalFlrSF, houseage. See output table below.

The CORR Procedure

10 Variables:	OverallQual BsmtFinSF1 GrLivArea FullBath BedroomAbvGr GarageCars WoodDeckSF SalePrice TotalFlrSF houseage

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
OverallQual	2930	6.09488	1.41103	17858	1.00000	10.00000
BsmtFinSF1	2929	442.62957	455.59084	1296462	0	5644
GrLivArea	2930	1500	505.50889	4394093	334.00000	5642
FullBath	2930	1.56655	0.55294	4590	0	4.00000
BedroomAbvGr	2930	2.85427	0.82773	8363	0	8.00000
GarageCars	2929	1.76681	0.76057	5175	0	5.00000
WoodDeckSF	2930	93.75188	126.36156	274693	0	1424
SalePrice	2930	180796	79887	529732456	12789	755000
TotalFlrSF	2930	1495	503.13016	4380390	334.00000	5642
houseage	2930	36.43413	30.29136	106752	-1.00000	136.00000

Comment on which predictor variables have the strongest linear relationships with the response variable,Y?

The predictor variables with the best R^2 coefficient are OverallQual (0.79926), TotalFlrSF (0.71359) and GrLivArea (0.70678); this is followed by GarageCars (0.64788) and FullBath

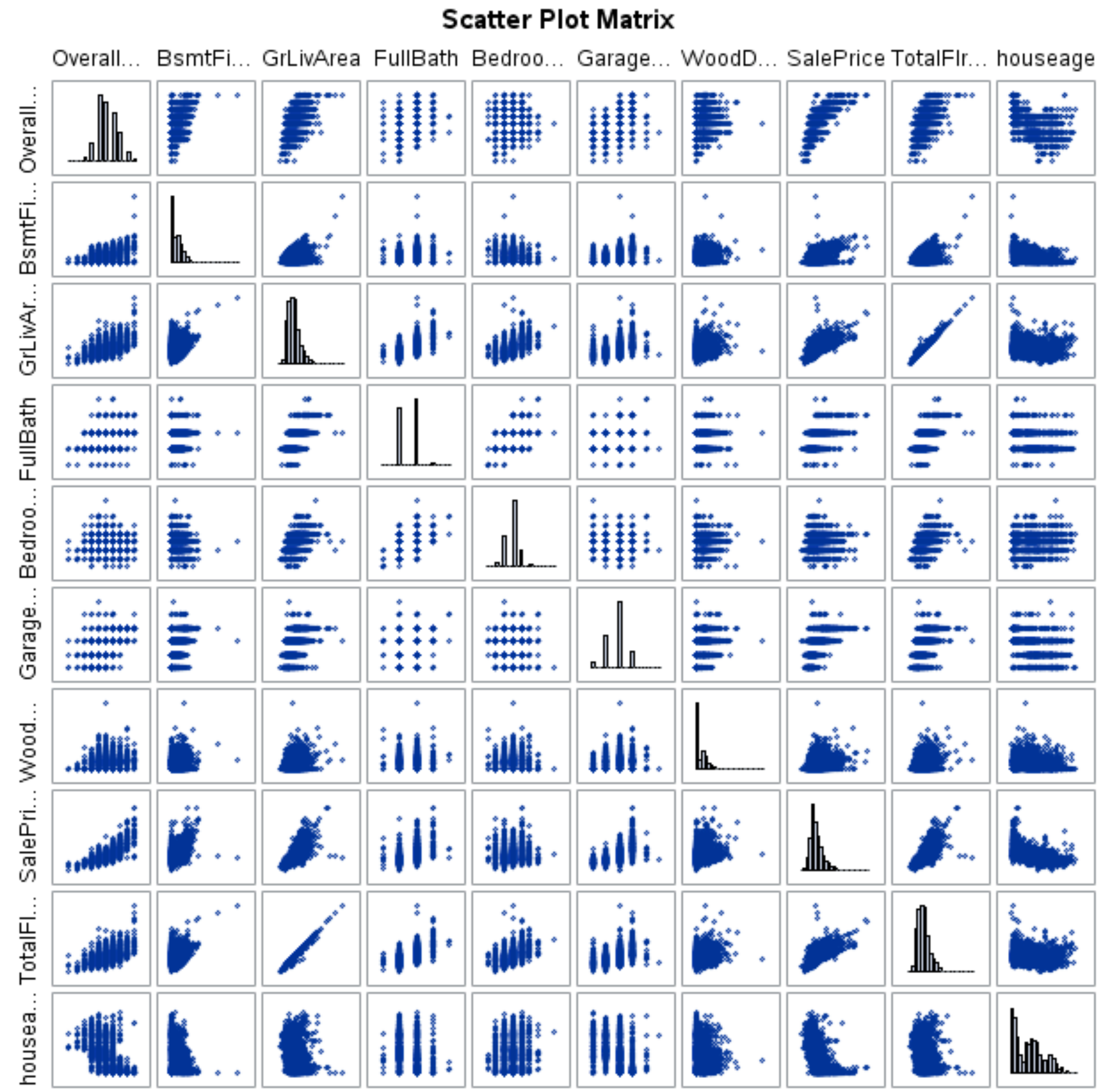
(0.54560), BsmtFinSF1 (0.43291) and WoodDeckSF (0.32714). The relationship between saleprice and BedroomAbvGr (0.14391), seems to have the most surprisingly low correlation. There is also a reasonable negative correlation between saleprice and houseage (-0.55891), indicating newer houses have a higher sale price *versus* older ones. See Pearson correlation coefficients table below.

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	OverallQual	BsmtFinSF1	GrLivArea	FullBath	BedroomAbvGr	GarageCars	WoodDeckSF	SalePrice	TotalFlrSF	houseage
OverallQual	1.00000 2930	0.28412 <.0001 2929	0.57056 <.0001 2930	0.52226 <.0001 2930	0.06329 0.0006 2930	0.59954 <.0001 2929	0.25566 <.0001 2930	0.79926 <.0001 2930	0.57773 <.0001 2930	-0.59702 <.0001 2930
BsmtFinSF1	0.28412 <.0001 2929	1.00000 2929	0.20963 <.0001 2929	0.07777 <.0001 2929	-0.11896 <.0001 2929	0.25548 <.0001 2928	0.22401 <.0001 2929	0.43291 <.0001 2929	0.21672 <.0001 2929	-0.27847 <.0001 2929
GrLivArea	0.57056 <.0001 2930	0.20963 <.0001 2929	1.00000 2930	0.63032 <.0001 2930	0.51681 <.0001 2930	0.48883 <.0001 2929	0.25015 <.0001 2930	0.70678 <.0001 2930	0.99579 <.0001 2930	-0.24251 <.0001 2930
FullBath	0.52226 <.0001 2930	0.07777 <.0001 2929	0.63032 <.0001 2930	1.00000 2930	0.35949 <.0001 2930	0.47818 <.0001 2929	0.17957 <.0001 2930	0.54560 <.0001 2930	0.63354 <.0001 2930	-0.46890 <.0001 2930
BedroomAbvGr	0.06329 0.0006 2930	-0.11896 <.0001 2929	0.51681 <.0001 2930	0.35949 <.0001 2930	1.00000 2930	0.09136 <.0001 2929	0.02971 0.1079 2930	0.14391 <.0001 2930	0.51276 <.0001 2930	0.05423 0.0033 2930
GarageCars	0.59954 <.0001 2929	0.25548 <.0001 2928	0.48883 <.0001 2929	0.47818 <.0001 2929	0.09136 <.0001 2929	1.00000 2929	0.24123 <.0001 2929	0.64788 <.0001 2929	0.49734 <.0001 2929	-0.53760 <.0001 2929
WoodDeckSF	0.25566 <.0001 2930	0.22401 <.0001 2929	0.25015 <.0001 2930	0.17957 <.0001 2930	0.02971 0.1079 2930	0.24123 <.0001 2929	1.00000 2930	0.32714 <.0001 2930	0.25278 <.0001 2930	-0.22858 <.0001 2930
SalePrice	0.79926 <.0001 2930	0.43291 <.0001 2929	0.70678 <.0001 2930	0.54560 <.0001 2930	0.14391 <.0001 2930	0.64788 <.0001 2929	0.32714 <.0001 2930	1.00000 2930	0.71359 <.0001 2930	-0.55891 <.0001 2930
TotalFlrSF	0.57773 <.0001 2930	0.21672 <.0001 2929	0.99579 <.0001 2930	0.63354 <.0001 2930	0.51276 <.0001 2930	0.49734 <.0001 2929	0.25278 <.0001 2930	0.71359 <.0001 2930	1.00000 2930	-0.25691 <.0001 2930
houseage	-0.59702 <.0001 2930	-0.27847 <.0001 2929	-0.24251 <.0001 2930	-0.46890 <.0001 2930	0.05423 0.0033 2930	-0.53760 <.0001 2929	-0.22858 <.0001 2930	-0.55891 <.0001 2930	-0.25691 <.0001 2930	1.00000 2930

What do you notice about the relationship between the numeric correlation measure and the graphical relationship in the scatterplot?

While the tabular view of the correlation produces a discrete value for the R^2 coefficient, in contrast the scatterplot gives a better idea of the distribution, range and scatter of data points. See scatter plot matrix below:



Which predictor variable do you think will be the best single predictor variable. Why? Which will be the worst and why? Are there high correlations within the set of potential predictor variables? This is a primary way to see/identify multicollinearity.

OverallQual (0.79926) is the best single predictor based solely on its highest correlation coefficient, and conversely, the worst is BedroomAbvGr (0.14391). There are other variables displaying high correlations with saleprice in this sample and they include: TotalFlrSF (0.71359) and GrLivArea (0.70678), GarageCars (0.64788) and FullBath (0.54560), in that order. These latter variables are candidates for multilinearity. See the Pearson Correlation Coefficient table shown earlier, and note the row or column for saleprice.

Is the correlation coefficient sufficient information to make a decision regarding a predictor variable and it's usefulness in developing a predictive model? Why?

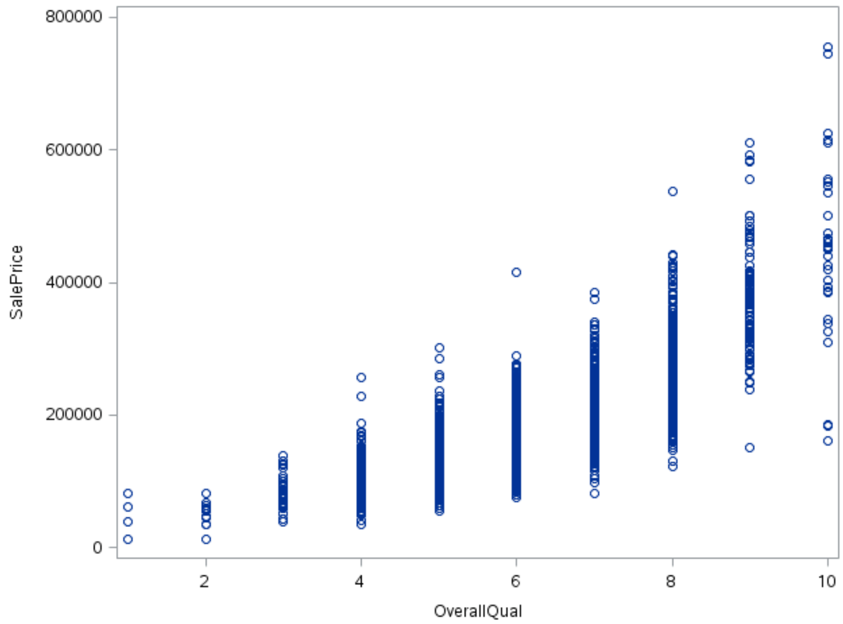
The correlation coefficient alone is not always a good measure of usefulness for a predictor variable. One needs to take into account the distribution and variability within that dataset. For example, in the case of OverallQual and saleprice, there is a lot of variability in saleprice for homes scoring as being excellent quality. This can be seen from the scatterplot but is difficult to observe merely from looking at the correlation coefficient alone. In addition for a variable such as OverallQual, the conferring of a particular score is not immediately transparent; that is, it may be somewhat subjective - what differentiates a score of '10' from one of '8'?

Conclusion: There are candidate variables that from the correlation coefficient matrix indicate may be predictors for sales price of homes, though there is some variability suggested from the scatter plots.

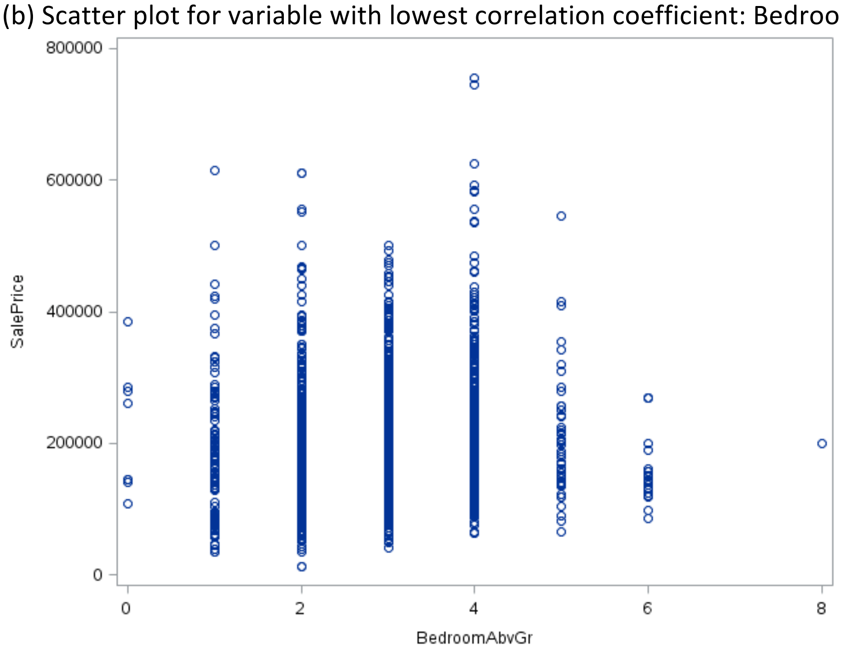
4. Scatter Plots

To investigate the scatter variability further, individual scatter for certain variables was studied.

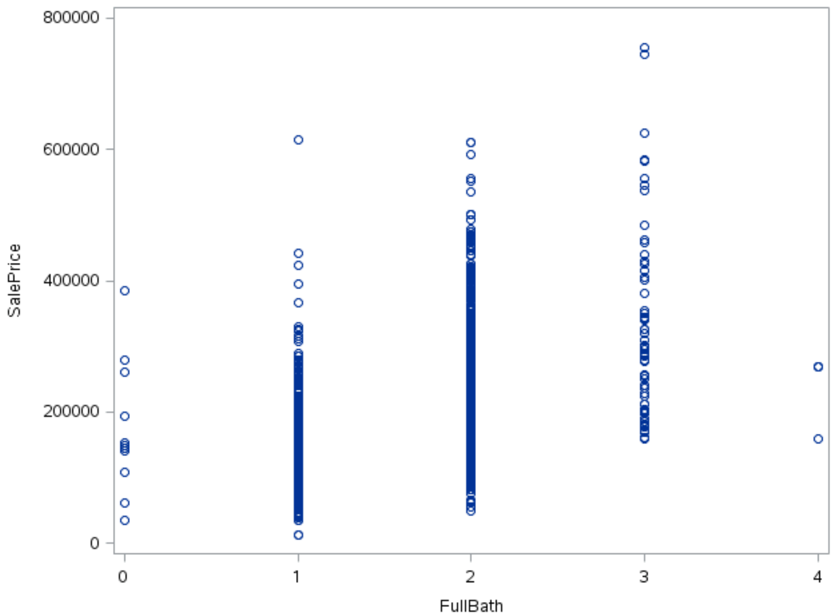
(a) Scatter plot for variable with highest correlation coefficient: OverallQual (0.79926)



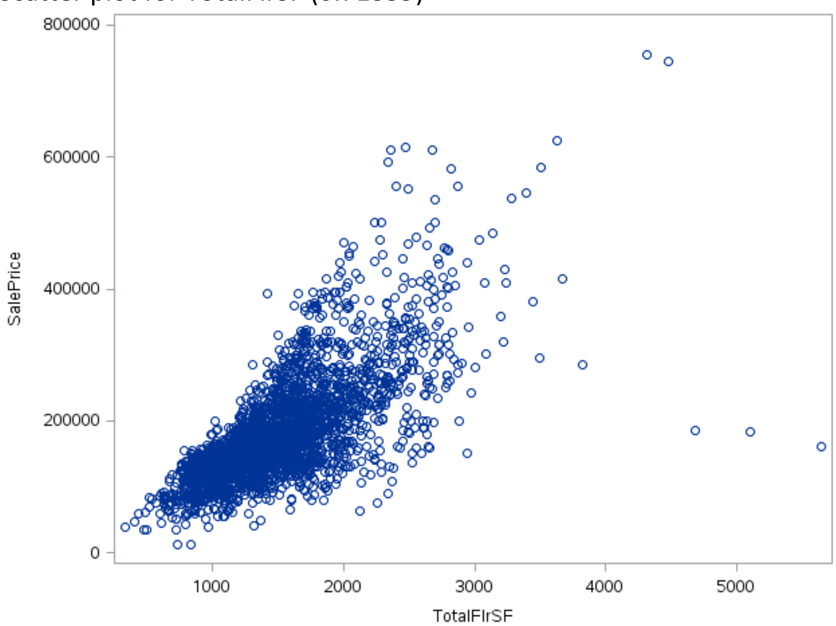
(b) Scatter plot for variable with lowest correlation coefficient: BedroomAbvGr (0.14391)



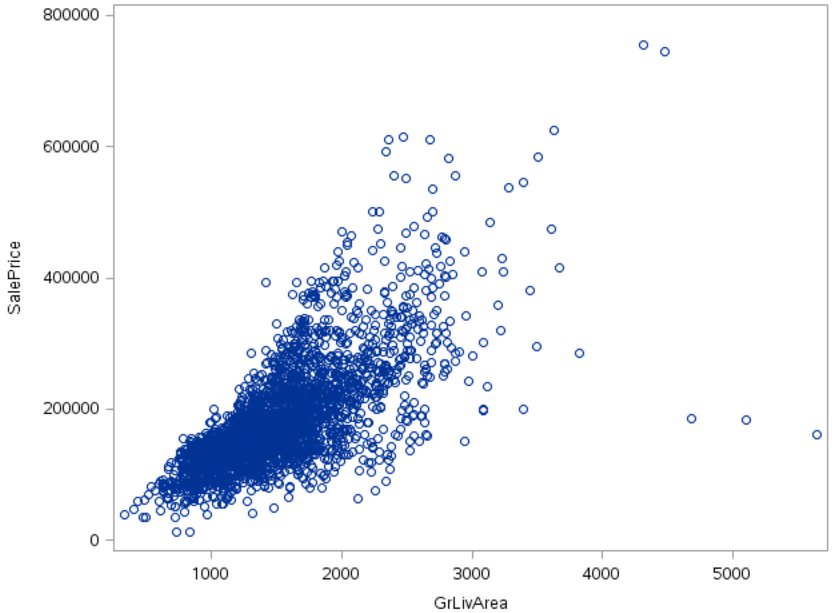
(c) Scatter plot for variable with correlation closet to 0.5: FullBath (0.54560)



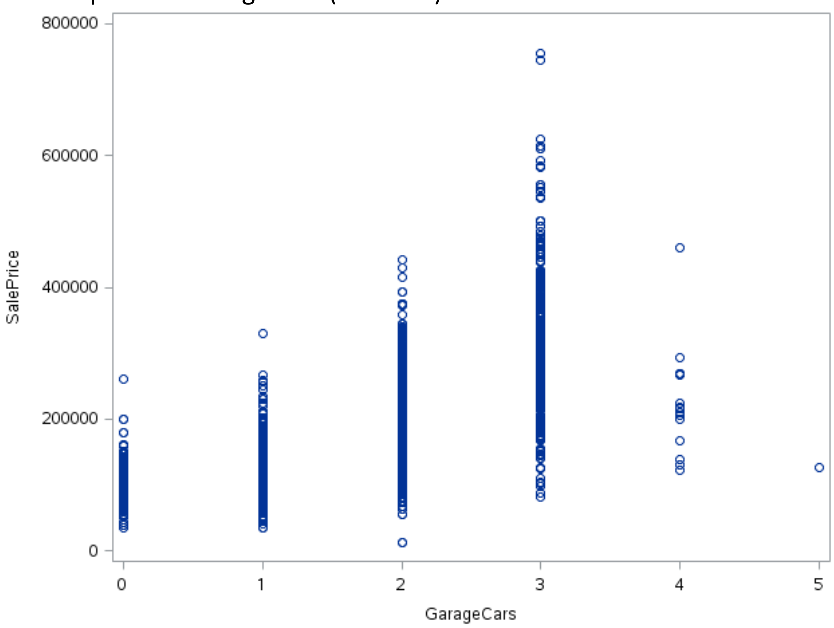
Scatter plot for TotalFlrSF (0.71359)



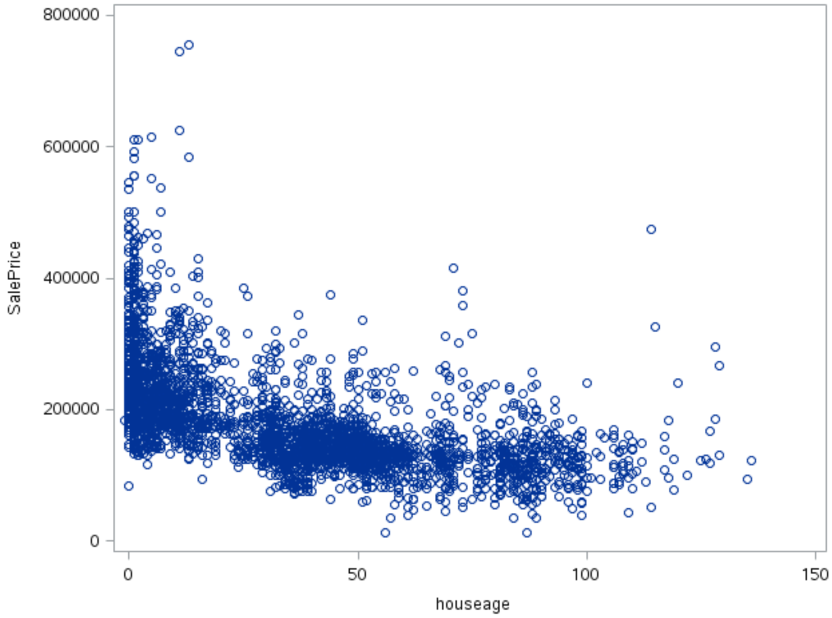
Scatter plot for GrLivArea (0.70678)



Scatter plot for GarageCars (0.64788)



Scatter plot for houseage (-0.55891)



Conclusion: Irrespective of the value of the correlation coefficients the individual scatter plots highlight that although there is a correlation evident, one needs to be wary since there is quite a lot of variation present. This variation is manifest in the form of extensive scatter (*e.g.* fan-tailing in TotalFlrSF, GrLivArea).

5. LOESS Plots

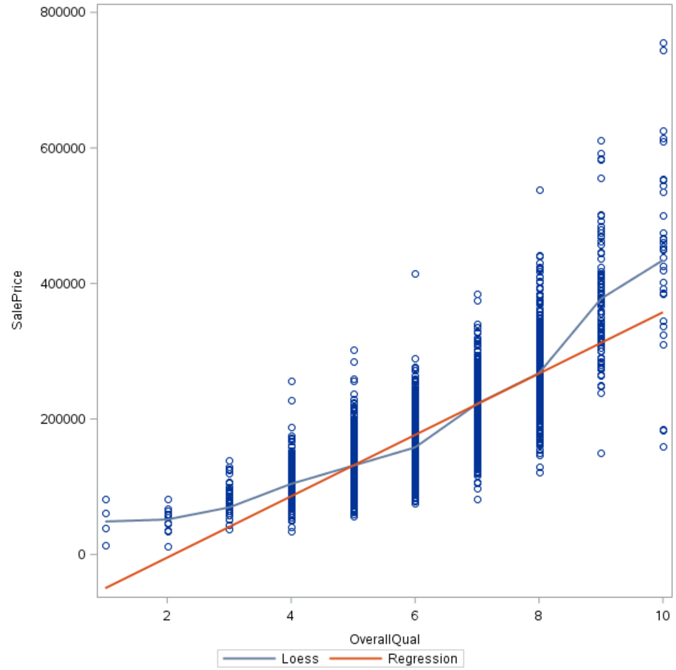
Comment on why we are interested in the LOESS scatterplots and what they are showing us?

LOESS (locally weighted scatterplot smoothing) augments linear regression models, when data sometimes display nonlinear patterns. LOESS sometimes provides a better curve-fit to non-linear patterns. It does this by fitting simple models to localized subsets of the data to build up a function, whereas a linear regression model will specify a global function to the entire dataset.

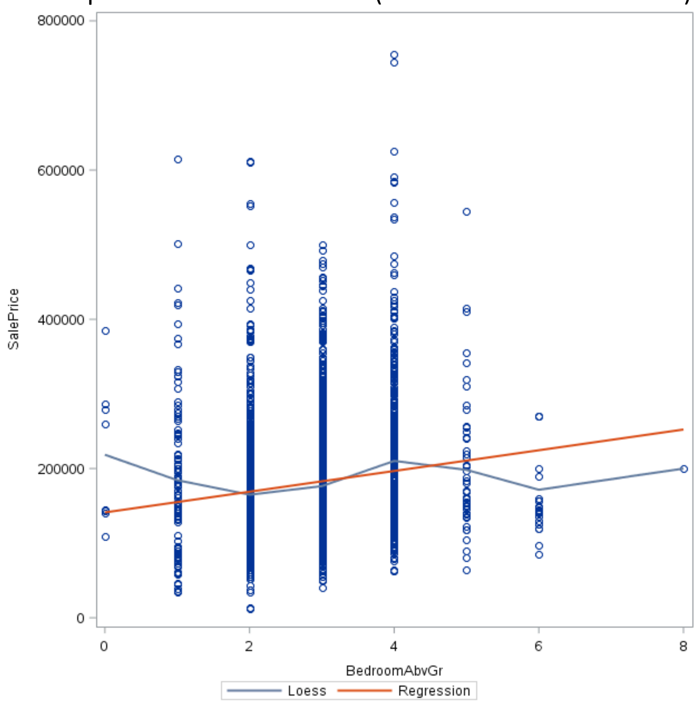
Below are the LOESS plots for:

a) OverallQual (highest correlation earlier)
b) BedroomAbvGr (lowest correlation earlier)
c) FullBath (correlation closest to 0.5 from earlier)

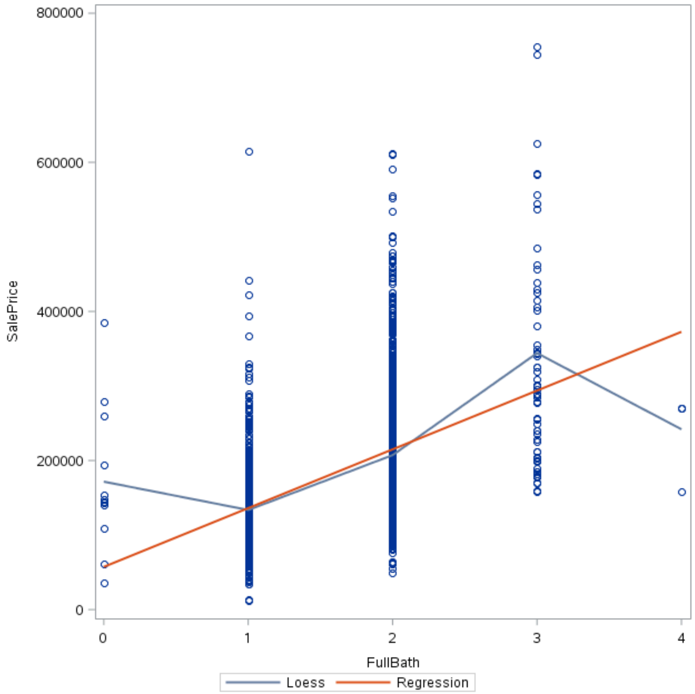
LOESS plot for OverallQual (highest correlation earlier)



LOESS plot for BedroomAbvGr (lowest correlation earlier)



LOESS plot for FullBath (correlation closest to 0.5 from earlier)



Conclusion: Since some of our variables display wide variability, while the linear model tries to fit the entire data for a variable a produces poor fit to the data, as is, the LOESS model due to it localized approach (point-by-point or category-by-category) fits some of the distributions better. Over parts of the datasets shown above, the LOESS and linear regression lines coincide, while at other times the lines diverge.

6. Analysis of categorical variables

Using PROC REQ and histograms to display three specific categorical variables: KitchenQual, BldgType and CentralAir. The frequency distribution tables for these variables are shown below, followed by their respective histograms.

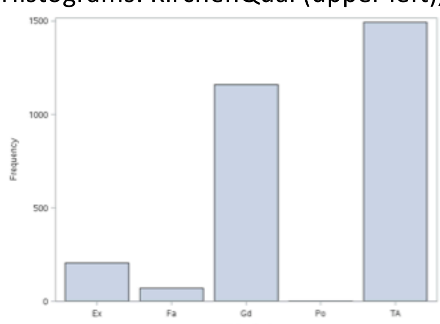
The FREQ Procedure

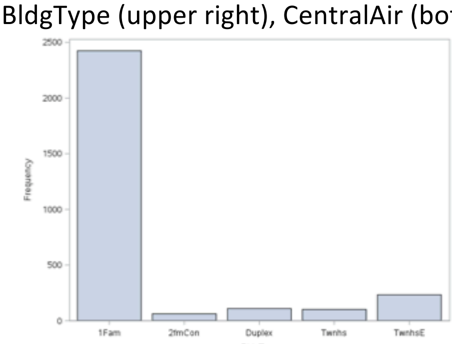
KitchenQual	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Ex	205	7.00	205	7.00
Fa	70	2.39	275	9.39
Gd	1160	39.59	1435	48.98
Po	1	0.03	1436	49.01
TA	1494	50.99	2930	100.00

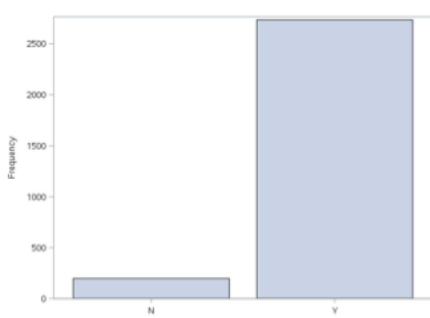
BldgType	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1Fam	2425	82.76	2425	82.76
2fmCon	62	2.12	2487	84.88
Duplex	109	3.72	2596	88.60
Twnhs	101	3.45	2697	92.05
TwnhsE	233	7.95	2930	100.00

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

Histograms: KirchenQual (upper left), BldgType (upper right), CentralAir (bottom).







Conclusion: Three categorical variables were chosen to investigate: KitchenQual, BldgType and CentralAir. From the frequency analysis with KitchenQual, approximately 46% of the observations had kitchen quality that was Good/Excellent while 50% were typical. For BldgType most of the observations, 82%, were single-family homes, and similarly for CentralAir over 90% of homes had central air conditioning.

7. Relating categorical variables with the response

a) Output from PROC MEANS SalePrice by KitchenQual:

The MEANS Procedure

KitchenQual=Ex

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
205	337339.34	114361.31	86000.00	755000.00

KitchenQual=Fa

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
70	105907.04	41974.72	13100.00	260000.00

KitchenQual=Gd

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
1160	210835.58	63585.01	59000.00	625000.00

KitchenQual=Po

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
1	107500.00	.	107500.00	107500.00

KitchenQual=TA

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
1494	139549.95	38447.77	12789.00	375000.00

The average prices of homes with Good or Excellent KitchenQual were $210,835 and $337,339, respectively. Typical kitchen quality homes had an average sale price of $139,549 while poor kitchen quality had a low sale price ($107,500). Thus, KitchenQual is a reasonable indicator of sale price, although it might be a little subjective what makes a kitchen typical but not good.

b) Output from PROC MEANS SalePrice by BldgType:

The MEANS Procedure

BldgType=1Fam

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
2425	184812.04	82821.80	12789.00	755000.00

BldgType=2fmCon

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
62	125581.71	31089.24	55000.00	228950.00

BldgType=Duplex

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
109	139808.94	39498.97	61500.00	269500.00

BldgType=Twnhs

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
101	135934.06	41938.93	73000.00	280750.00

BldgType=TwnhsE

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
233	192311.91	66191.74	71000.00	392500.00

From the analysis in part 6, over 80% of the homes in the Ames data set are single-family homes, whose average sale price is $184,812. Although townhouse inside units has a higher average sale price, $192,311, they only represent approximately 8% of the BldgType in the Ames dataset. The cheaper homes are those, which are two-family conversions, duplexes or townhome end units, as they have lower average sale prices.

c) PROC MEANS saleprice by CentralAir

The MEANS Procedure

CentralAir=N

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
196	101890.48	37597.02	12789.00	265979.00

CentralAir=Y

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
2734	186452.79	79121.36	50000.00	755000.00

As we saw from part 6, most of the Ames homes (>90%) have central air and as seen here these homes have a higher average sale price ($186,452) compared to those without (101,890). Thus, CentralAir is a good indicator/co-indicator of sale price.

Conclusion: From the analysis of categorical variables in this section we know that KitchenQual, BldgType and CentralAir are good indicators of sale price, though they may not be entirely linear with respect to the sales price (see Section 8, EDA). These variables may also be grouped together, that is there may be multi-linearity evident. Thus, a single-family home with central air and a Good/Excellent kitchen quality will likely be at the higher end of the sales prices in Ames.

8. General EDA

Used PROC MEANS for 10 chosen variables to display some general attributes of average homes in the Ames, IA area.

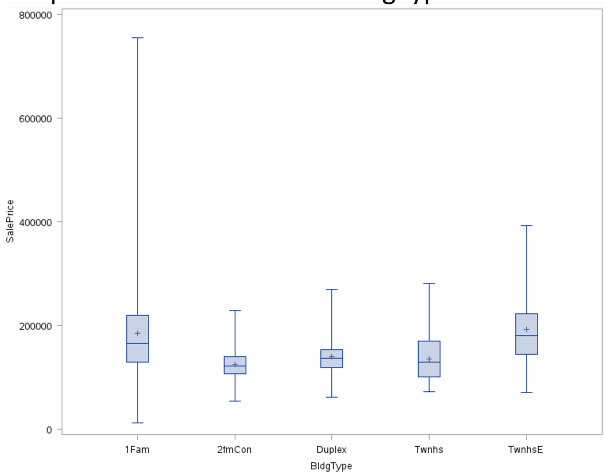
The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
TotalFlrSF	2930	1495.01	503.1301623	334.0000000	5642.00
GrLivArea	2930	1499.69	505.5088875	334.0000000	5642.00
houseage	2930	36.4341297	30.2913574	-1.0000000	136.0000000
FullBath	2930	1.5665529	0.5529406	0	4.0000000
BedroomAbvGr	2930	2.8542662	0.8277311	0	8.0000000
WoodDeckSF	2930	93.7518771	126.3615619	0	1424.00
BsmtFinSF1	2929	442.6295664	455.5908391	0	5644.00
OverallQual	2930	6.0948805	1.4110261	1.0000000	10.0000000
GarageCars	2929	1.7668146	0.7605664	0	5.0000000
SalePrice	2930	180796.06	79886.69	12789.00	755000.00

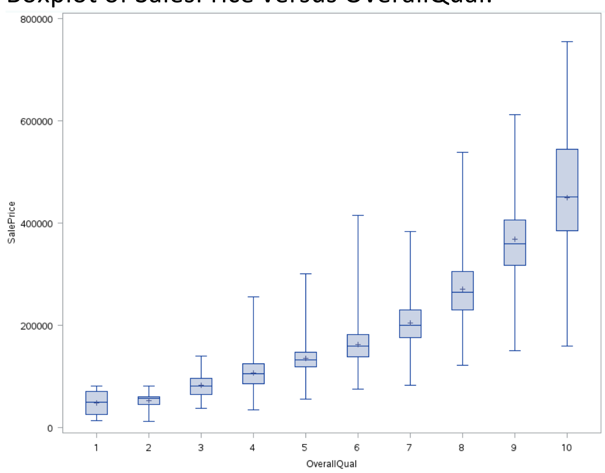
The means of some of these variables shown above provides some general indicators of the Ames house market. So on average an $180,000 home in Ames may comprise an approximately 1500 sq. ft. total living area, with approx. 3 bedrooms and a 2-car garage and 1-2 full bathrooms.

Some boxplots of sales prices by BldgType, OverallQual and GarageCars were generated (see below). Interestingly, with OverallQual boxplot (overall quality of the home) although as the house quality gets better (10=Excellent, 1=Poor) the sale price increases.

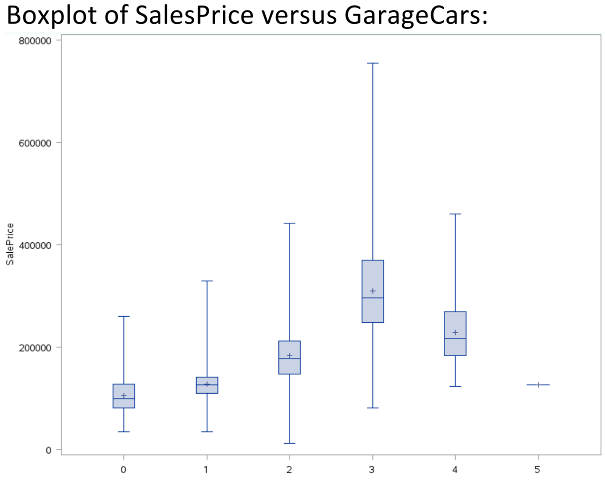
Boxplot of SalesPrice versus BldgType:



Boxplot of SalesPrice versus OverallQual:



Boxplot of SalesPrice versus GarageCars:



However, for most of the scoring categories (4-10), there is a high degree of variability, which increases with the score. This may be an indicator of the subjectivity with which these scores are attributed to each category of home. For OverallQual and GarageCars looking at the means, there is a non-linear relationship that is present.

Conclusion: Some of the EDA of the means of a number of the selected variables provides a snapshot of an average home in Ames, IA. As discussed in earlier sections some of variables when plotted versus sales price show some interesting and non-linear relationships with extensive variability, which needs to be noted for subsequent modeling analyses.

CONCLUSIONS

One of the aims of this initial assignment was to perform:

a) Data Survey
b) Data Quality Check
c) Initial Exploratory Data Analysis (EDA)

As a result of this analysis we can conclude that the data present in this dataset allows us to draw some relationships between some of the numeric and categorical variables *versus* sale price of homes in Ames, IA. From the data survey and quality check we can state that most of the data surveyed in the Ames Housing dataset represents what they are designed to represent. Specifically, sale prices appear valid, although there are some errors in some variables (*e.g.* houseage with a -1 value) and some outliers exist. Twenty variables were chosen from prior knowledge and experience of house sales generally as potential predictors of sales price. Some of the EDA indicated that there are variables chosen that have a reasonable correlation with sales price, specifically: OverallQual, TotalFlrSF, GrLivArea, GarageCars, FullBath and houseage. There are also categorical variables such as KitchenQual, BldgType and GarageCars that can be used as candidate co-predictors of sale price.

As noted above although some predictors of sales price exist, the EDA analysis indicates that even for the best predictor variables (*e.g.* TotalFlrSF, GrLivArea) there is high degree of variability or scatter with increasing higher square footage and sales prices, for example. This is manifest with most, if not all, of the potential continuous and categorical predictors identified. This is a concern when trying to use a single predictor as a measure of sales price, as this can lead to a high error-rate in the prediction. In addition, although some of the relationships between predictor variables are linear (though with high scatter), others have curvi-linear relationships. This might suggest that further analysis with these predictors may require some transformation of the data from these variables. For example, some of the fan tailing (scatter) seen with TotalFlrSF/GrLivArea *versus* SalesPrice may be reduced with a logarithmic transformation to tighten up the distribution of points, leading to a better linear regression model.