Assignment #1
Andrea Bruckner

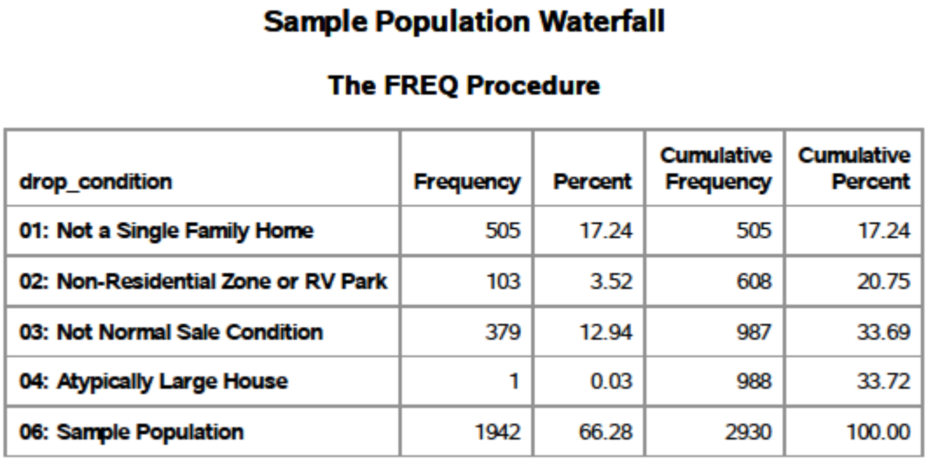
Introduction

In this report we are exploring and familiarizing ourselves with the Ames Housing Data Set so that we can later on build models to predict the value of a property based on various residential variables. Our initial investigation into the data reveals that while certain variables cause predictable increases in sale price, other variables do not reveal a clear pattern, such as why certain Neighborhoods are more expensive than others and why 2-bedroom houses have a larger and higher overall price range than 4-bedroom houses. The price of a property generally increases with property size, but scatter plots later in the report reveal that the price slows in how much it increases after a certain lot size.

Data

The data set we reviewed is from the Ames Assessor's Office and contains 2930 observations with 82 variables concerning residential properties sold between 2006 and 2010 in Ames, Iowa. A broad overview of the data reveals that we have enough data to predict home prices. The data set contains a variety of categorical and numeric variables from which we can draw conclusions, and the data set contains very few missing values.

In order to make the data easier to analyze, we defined a sample from the larger original data set. The waterfall shown below illustrates how many observations were dropped from the original data set with each condition. The fifth drop condition, 05: Pricing Error, does not appear in the table because all of the SalePrice data met the condition that the housing price could not be less than or equal to $0. The waterfall resulted in a sample population of 1942 observations.



Sample Population Waterfall

The FREQ Procedure

drop_condition	Frequency	Percent	Cumulative Frequency	Cumulative Percent
01: Not a Single Family Home	505	17.24	505	17.24
02: Non-Residential Zone or RV Park	103	3.52	608	20.75
03: Not Normal Sale Condition	379	12.94	987	33.69
04: Atypically Large House	1	0.03	988	33.72
06: Sample Population	1942	66.28	2930	100.00

We decided on these drop conditions based on what a typical homebuyer would likely consider when purchasing a home. Most homebuyers are seeking single-family, non-mobile homes; are purchasing the home under normal conditions (i.e., not buying the home from a family member or through a short sale); and are not looking to buy a mansion.

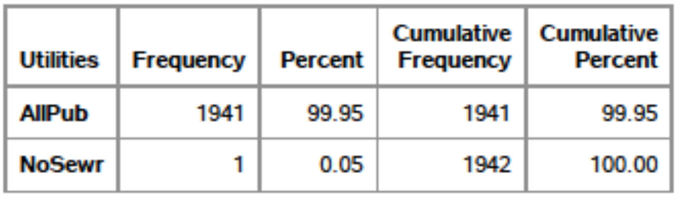
The data documentation available for the Ames Housing Data Set noted that the data set contained five unusual observations, which we were able to easily identify by plotting the sale price against the living area. Creating a scatterplot post waterfall revealed that our drop conditions had effectively removed these observations from the sample data. See Appendix A for the pre- and post-waterfall scatter plots.

After analyzing the results of the sample population, we further reduced our sample population based on what is statistically typical for single-family residences in Ames, Iowa. In the next section we will explain which variables we decided to examine and how we chose the criteria for our next waterfall.

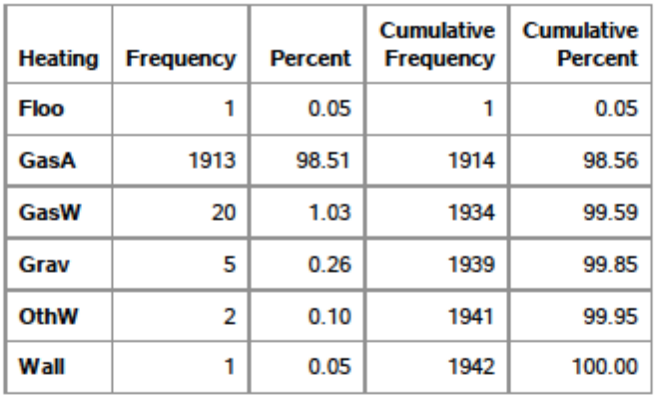
Initial Exploratory Data Analysis

Before we could begin conducting an initial exploratory data analysis (EDA), we had had to select 20 predictor variables from the 82 total variables. A list of these 20 predictor variables appears in Appendix B, along with the data quality checks we performed on each variable. When selecting the 20 variables, we tried to choose a mixture of categorical and numeric variables that would highlight some key features homebuyers likely prioritize when searching for a home. For instance, the size of the house and its property, the quality and condition of the residence, the neighborhood, the month the house was available for purchase, the number of bathrooms and bedrooms, the kinds of utilities available, and the existence and condition of a basement and garage are all features that are important to homebuyers and that affect the sale price of the home.

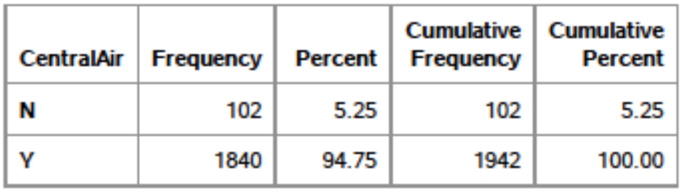
Variables with a Majority Feature



Utilities	Frequency	Percent	Cumulative Frequency	Cumulative Percent
AllPub	1941	99.95	1941	99.95
NoSewr	1	0.05	1942	100.00



Heating	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Floo	1	0.05	1	0.05
GasA	1913	98.51	1914	98.56
GasW	20	1.03	1934	99.59
Grav	5	0.26	1939	99.85
OthW	2	0.10	1941	99.95
Wall	1	0.05	1942	100.00

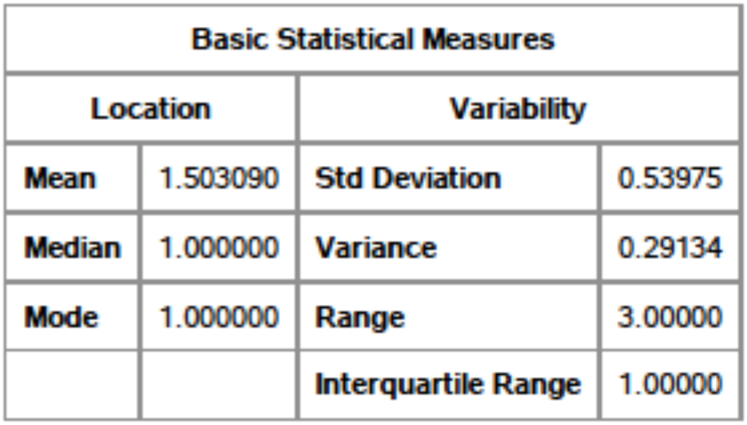


CentralAir	Frequency	Percent	Cumulative Frequency	Cumulative Percent
N	102	5.25	102	5.25
Y	1840	94.75	1942	100.00

From these 20 variables, we selected 10 to conduct an initial EDA. Several variables were easy to eliminate since the data revealed that certain features were found in nearly all the homes. For instance, the Utilities variable revealed that all but one observation in the sample population had all public utilities, and the Heating and CentralAir variables revealed that most homes had gas forced warm air furnaces and central air. Keeping these variables would not reveal any useful information about sale prices and thus could be excluded.

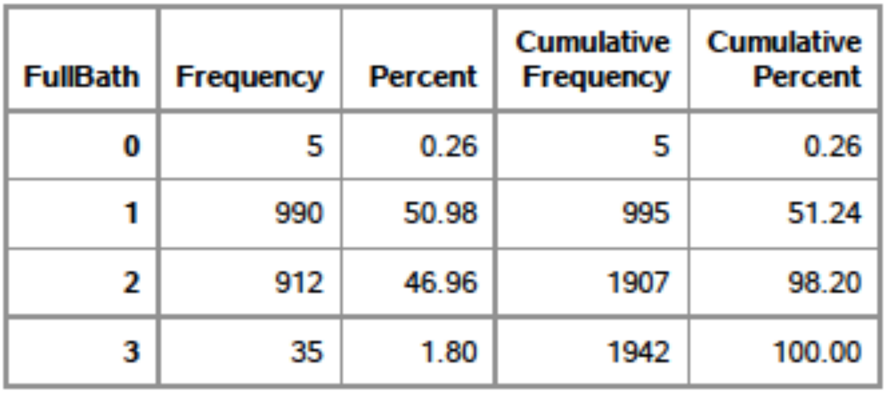
After narrowing down the 20 variables into 10 variables, we began performing an EDA on the selected variables. Because the variables were continuous, discrete, ordinal, or nominal, we had to weigh which methods would be most appropriate for analyzing each variable. In general, we performed PROC FREQ on the ordinal and nominal data and PROC UNIVARIATE and PROC MEANS on the continuous and discrete data. We found this this was not the best method for all of the variables. Although OverallQual is ordinal, its data is discrete numerals and is also readable and somewhat more informative when performing the PROC UNIVARIATE and PROC MEANS functions. Likewise, FullBath, BedroomAbvGr, and MoSold are all discrete data and are also readable and easier to understand when using the PROC FREQ function. For instance, the PROC UNIVARIATE OUTPUT for FullBath can be somewhat misleading. It reveals a mean of 1.5 full bathrooms, which is easy to misinterpret as one full bathroom and one half bathroom. However, HalfBath is a separate variable. The PROC FREQ output makes the data much clearer.

Part of PROC UNIVARIATE Output for FullBath



Basic Statistical Measures			
Location		**Variability**	
Mean	1.503090	**Std Deviation**	0.53975
Median	1.000000	**Variance**	0.29134
Mode	1.000000	**Range**	3.00000
		Interquartile Range	1.00000

PROC FREQ Output for FullBath



FullBath	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	5	0.26	5	0.26
1	990	50.98	995	51.24
2	912	46.96	1907	98.20
3	35	1.80	1942	100.00

After examining the EDA outputs for the 10 variables, we created another waterfall from the sample population that more accurately reflects what is typical of housing transactions in Ames, Iowa. In general, we based the drop conditions on which features were least common in a variable. For the continuous data, LotArea and GrLivArea, we set the condition for the data that fell between the 5th and 95th percentiles. For the rest of the data, we modified the conditions based on trail and error and how many were eliminate during each drop condition. The first waterfall from the sample population contained only 42 observations. We would like to devise a better and more consistent methodology for future reports. Below is the Typical Housing Transaction Waterfall. The specific criteria for each drop condition can be found in the Code section of the report.

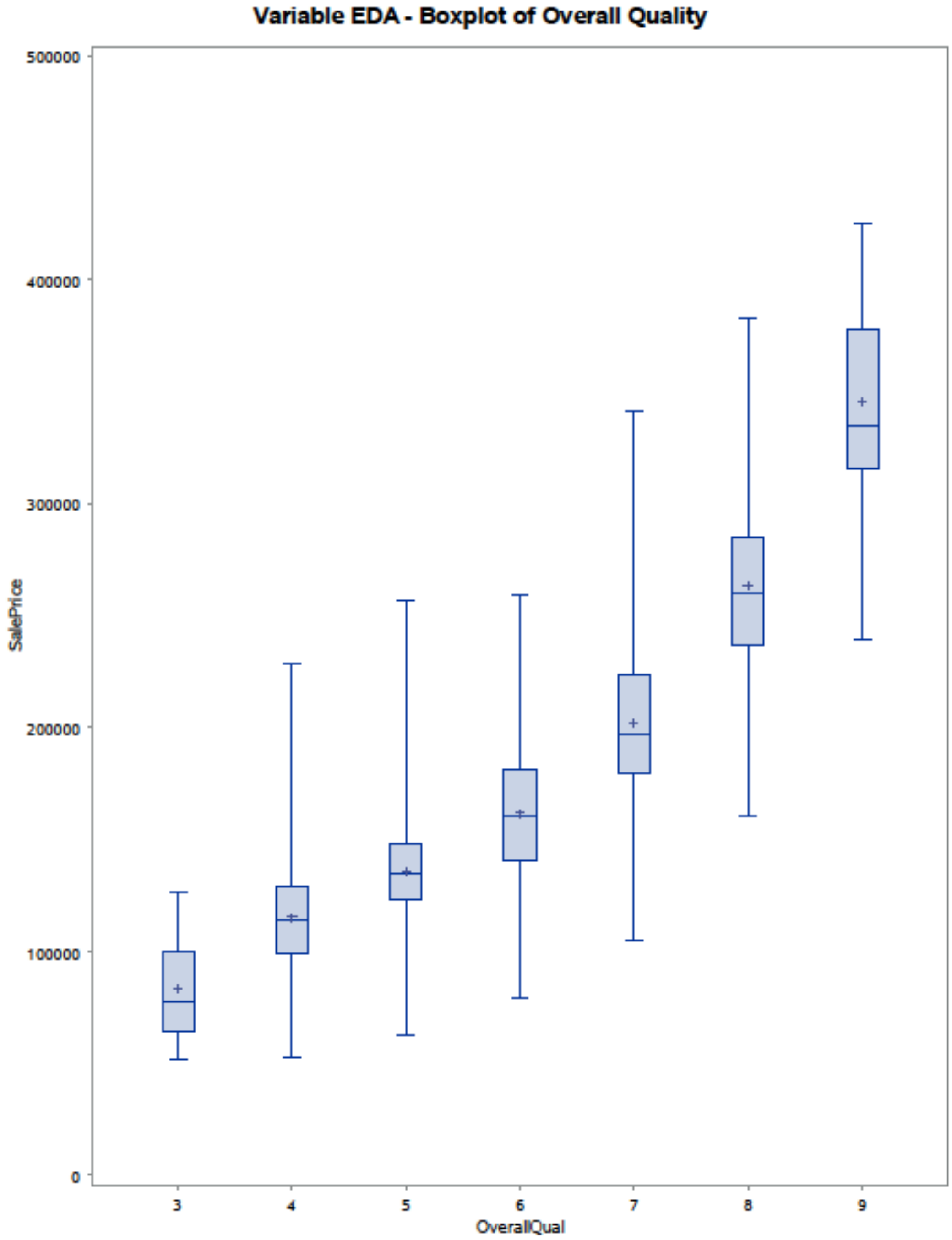
Typical Housing Transaction Waterfall

The FREQ Procedure

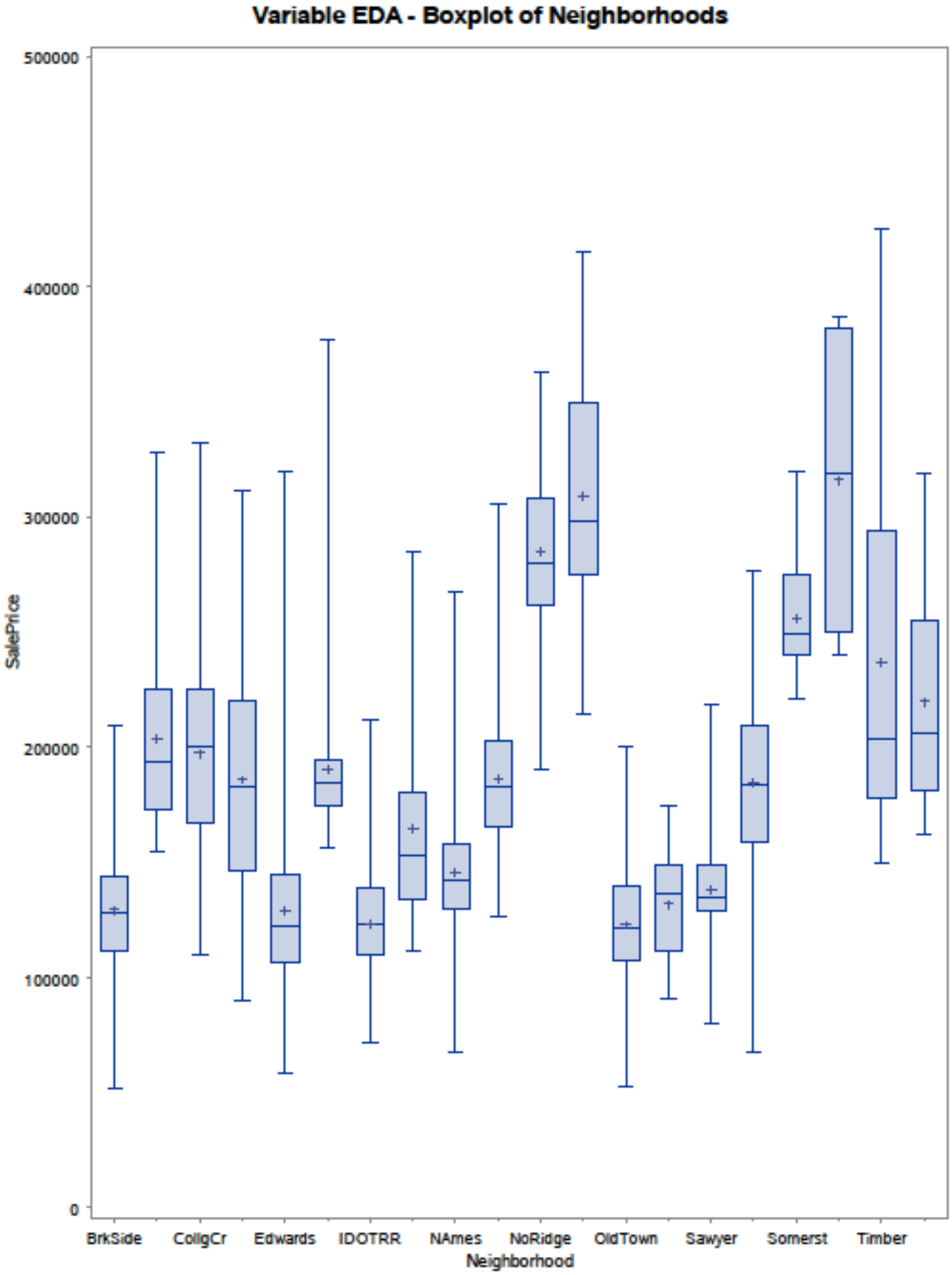
drop_condition	Frequency	Percent	Cumulative Frequency	Cumulative Percent
01: Not Typical Overall Quality	23	1.18	23	1.18
02: Not Typical House Style	37	1.91	60	3.09
03: Not Typical Garage Type	12	0.62	72	3.71
04: Not Typical Basement Type	38	1.96	110	5.66
05: Not Typical Neighboorhood	1	0.05	111	5.72
06: Not Typical Lot Area	87	4.48	198	10.20
07: Not Typical GrLivArea	74	3.81	272	14.01
08: Not Typical Number of Full Bath	11	0.57	283	14.57
09: Not Typical Number of Bedrooms	33	1.70	316	16.27
10: Not Typical Month Sold	105	5.41	421	21.68
11: Typical Housing Transaction	1521	78.32	1942	100.00

Discrete Variables EDA for Modeling

For variables with discrete data, we focused on modeling the relationship between SalePrice and the predictor variables using box plots. The box plots, on a whole, revealed predictable trends. For instance, the box plot depicting the Overall Quality of a house reveals that the sale price increases as quality increases. The lowest quality house has a relatively small range for sale prices, but houses with an Overall Quality rating of 7 or 8 have a larger pricing range.

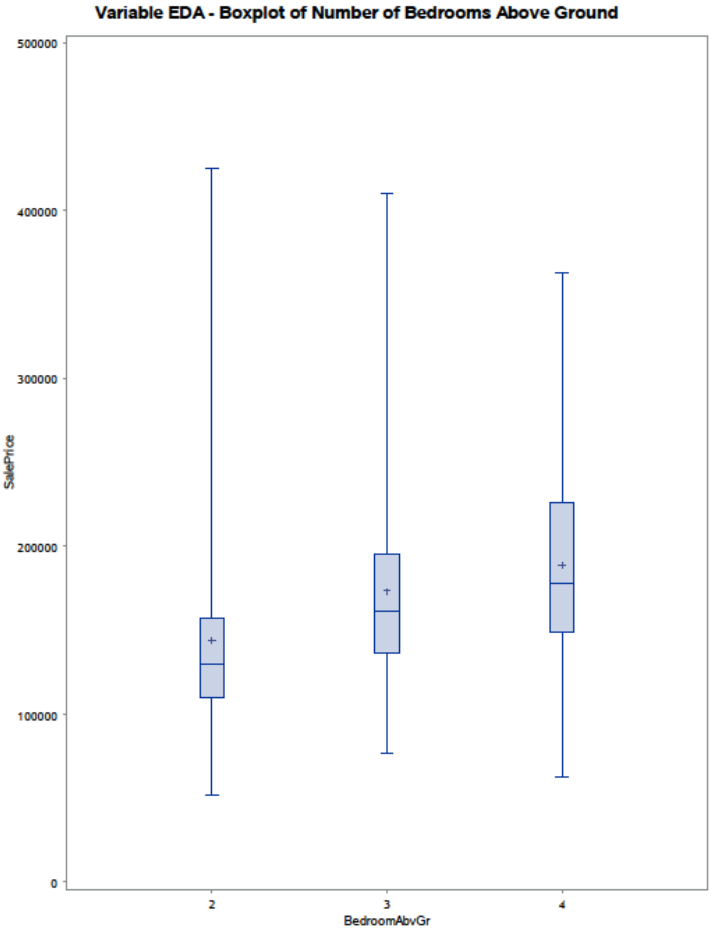


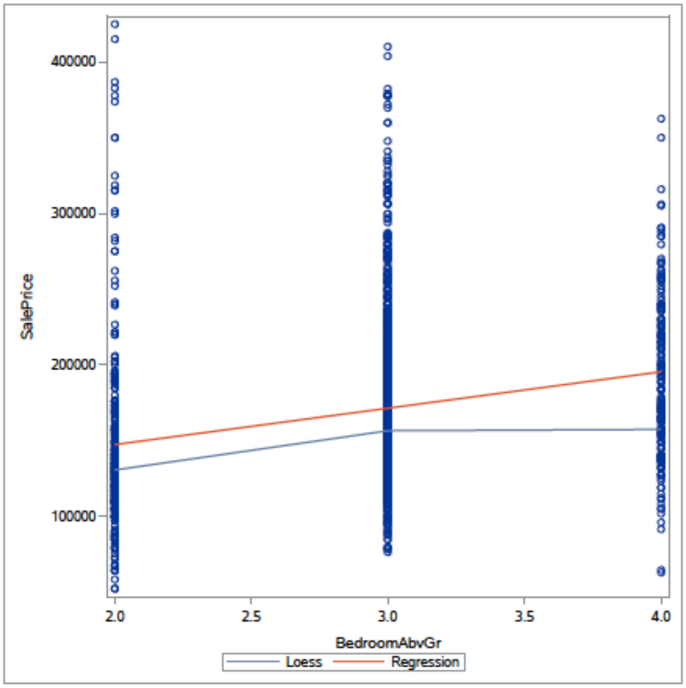
The box plot for sale price according to neighborhood reveals that certain neighborhoods are more expensive to live in. Running further analysis may reveal that the more expensive neighborhoods are made of higher quality materials, are in better condition, have more square footage, or some other features that demand a higher sale price.



Variable EDA - Boxplot of Neighborhoods

While trying to create a graphic of the BedroomAbvGr variable, we learned this data was better suited to a box plot than scatter plot. While houses with 4 bedrooms had higher prices, the price range for 2-bedroom houses is interesting and warrants further exploration. Why are certain 2-bedroom houses more expensive than houses with more bedrooms?

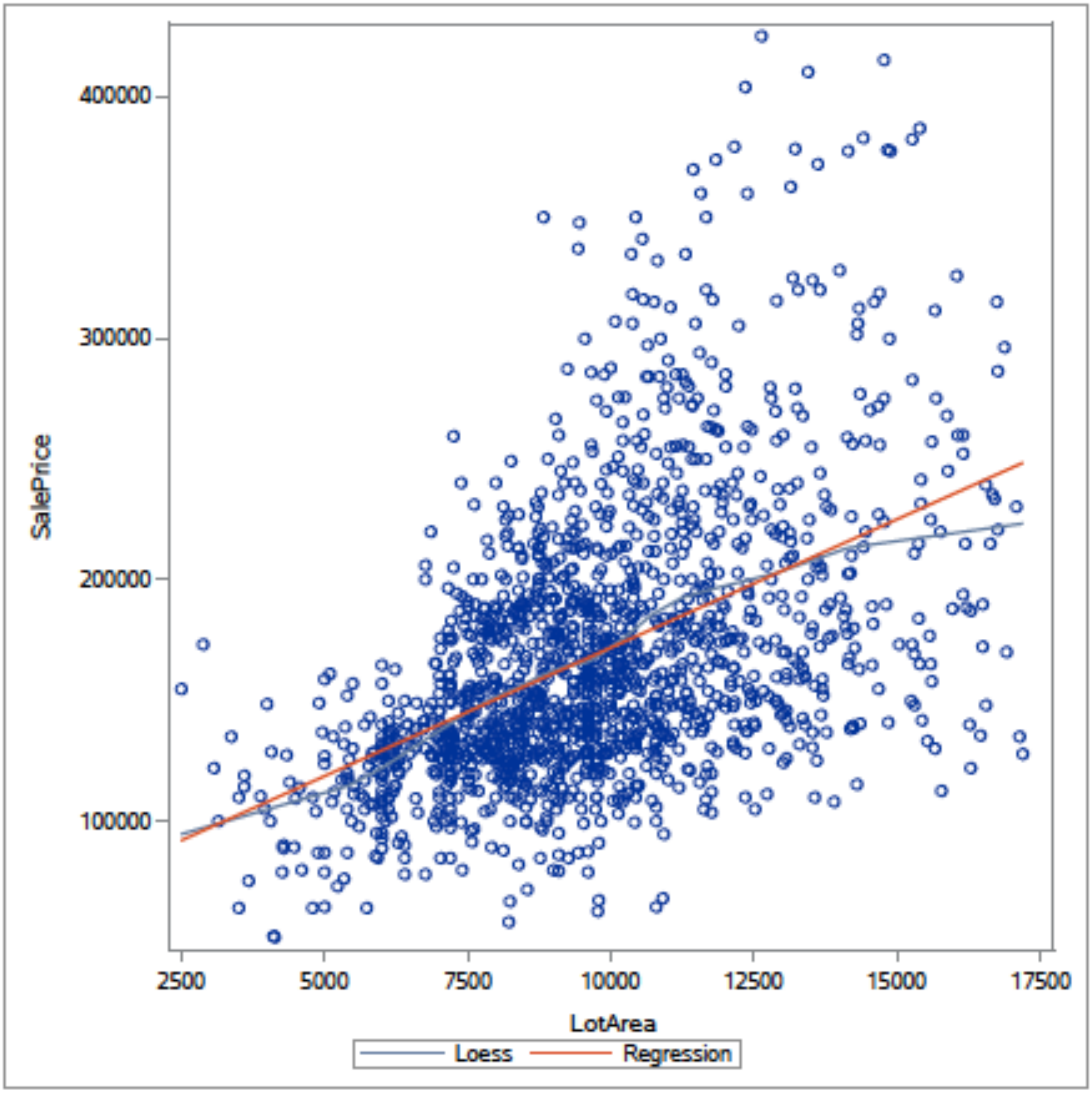




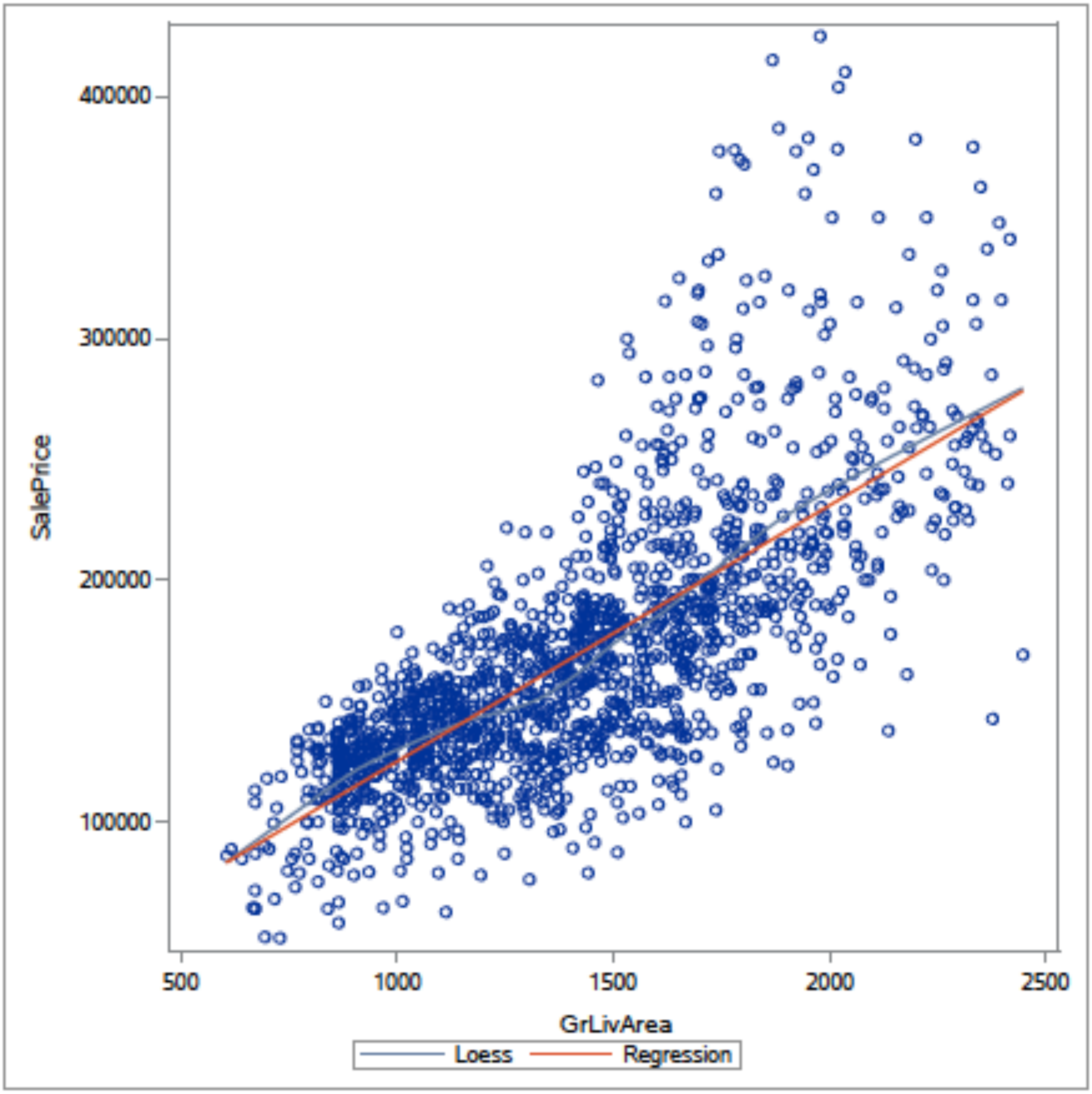
Continuous Variables EDA for Modeling

For both the scatter plots of LotArea v. SalePrice and GrLivArea v. SalePrice, the regression line reveals a logical increase in sale price as area increases. The LOESS curve in the LotArea v. SalePrice scatter plot shows that the increase in sale price begins to taper at the 15,000 sq. ft. mark. For the GrLivArea v. SalePrice scatter plot, the LOESS curve tends to fluctuate more. It is postioned higher, then lower, then higher again in relation to the regression line. Further exploration might reveals a reason for this.

Scatter Plot of LotArea v. SalePrice



Scatter Plot of GrLivArea v. SalePrice



Initial Exploratory Data Analysis Results

While we were unable to successfully use log(SalePrice) in our code, we think that transformations in the predictor variables might be useful at some point when building a model, but we are not yet sure how. The EDA suggests that we will encounter some difficulties in building the model. Though many of the outputs reveal predictable trends, there are certain outputs, such as the Neighborhood v. SalePrice box plot, that indicate further exploration of perhaps more variables is necessary to understand why certain neighborhoods are more expensive than others.

Conclusions

Our initial investigation of the data reveals some predictable patterns, but it also reveals some trends that do not yet have a clear and logical explanation. We would like to further analyze several topics, including how the data for other variables might correlate to the higher sale prices seen in certain neighborhoods. We also would like to see if there is a pattern to the highly priced 2-bedroom houses. Overall we think that we have enough data to eventually create a predictive model, but we still need to explore data relationships a bit more.

Code

```
* Andrea Bruckner
  PREDICT 410, Sec 55
  Winter 2016
  Assignment 1
;

*******************************************************************;
* Preliminary Steps and Data Survey;
*******************************************************************;

* Access library where Ames housing dataset is stored;
libname mydata '/scs/crb519/PREDICT_410/SAS_Data/' access=readonly;
proc datasets library=mydata; run; quit;

* Set Ames housing dataset to short name;
data ames;
set mydata.AMES_HOUSING_DATA;
run;

* Explore Ames housing dataset contents;
proc contents data=ames; run;

* Print first 10 observations as preview of data;
Title "Preview of Dataset";
options obs=10;
proc print data=ames; run;
options obs=max; * reset options to analyze and report on all data;

* Print scatterplot of SalePrice v. GrLivArea to find obvious outliers or unusual data;
ods graphics on;
PROC SGPLOT data=ames;
	scatter X=SalePrice Y=GrLivArea;
	title "Sale Price v. Above Ground Living Area Scatter Plot – No Smoothers";
run;
ods graphics off;
```

```
*****************************************************************;
* Define the Sample Population;
*****************************************************************;

* Create drop conditions;
data temp;
	set ames;
	format drop_condition $35.;
	if (BldgType ne '1Fam') then drop_condition='01: Not a Single Family Home';
	else if (Zoning = 'A' or zoning = 'C' or zoning = 'FV' or zoning = 'I' or zoning = 'RP')
		then drop_condition='02: Non-Residential Zone or RV Park';
	else if (SaleCondition ne 'Normal') then drop_condition='03: Not Normal Sale
Condition';
	else if (GrLivArea >= 4000) then drop_condition='04: Atypically Large House';
	else if (SalePrice <= 0) then drop_condition='05: Pricing Error';
else drop_condition='06: Sample Population';
run;

* View frequency of drop conditions;
proc freq data=temp;
tables drop_condition;
title 'Sample Population Waterfall';
run; quit; * (Reveals that all Sale Prices are positive);

* Subset data for just Sample Population;
data sample;
set temp;
if (drop_condition='06: Sample Population');
run;

* View scatterplot of sample data;
ods graphics on;
PROC SGPLOT data=sample;
	scatter X=SalePrice Y=GrLivArea;
	title "Sample Population Sale Price v. Above Ground Living Area Scatter Plot – No
Smoothers";
run;
ods graphics off;

*****************************************************************;
* Data Quality Checks;
*****************************************************************;

* Create frequency tables for categorical data;
proc freq data=sample;
```

```
tables Utilities Neighborhood Condition1 HouseStyle OverallQual OverallCond
BsmtFinType1 Heating CentralAir KitchenQual GarageType GarageCars SaleCondition;
title;
run; quit;

* Create summary statistics for numeric data;
proc univariate data=sample;
var LotArea GrLivArea LowQualFinSF FullBath HalfBath BedroomAbvGr MoSold;
run; quit;

proc means data=sample min q1 mean median q3 max std;
var LotArea GrLivArea LowQualFinSF FullBath HalfBath BedroomAbvGr MoSold;
run; quit;

*******************************************************************;
* Initial EDA;
*******************************************************************;

* Create frequency tables for Initial EDA categorical data;
proc freq data=sample;
tables Neighborhood HouseStyle OverallQual BsmtFinType1 GarageType;
run; quit;

* Create summary statistics for Initial EDA numeric data;
proc univariate data=sample;
var LotArea GrLivArea FullBath BedroomAbvGr MoSold;
run; quit;

proc means data=sample min q1 mean median q3 max std;
var LotArea GrLivArea FullBath BedroomAbvGr MoSold;
run; quit;

**OverallQual is more readable when using proc univariate and proc means**;
proc univariate data=sample;
var OverallQual;
run; quit;

proc means data=sample min q1 mean median q3 max std;
var OverallQual;
run; quit;

**FullBath BedroomAbvGr MoSold are more readable when using proc freq**;
proc freq data=sample;
tables FullBath BedroomAbvGr MoSold ;
run; quit;
```

```
* Observations in a typical housing transaction;
data temp2;
	set sample;
	format drop_condition $35.;
	if (OverallQual = '1' or OverallQual = '2' or OverallQual = '10') then
drop_condition='01: Not Typical Overall Quality';
	else if (HouseStyle = '1.5Unf' or HouseStyle = '2.5Fin' or HouseStyle = '2.5Unf') then
drop_condition='02: Not Typical House Style';
	else if (GarageType = '2Types' or GarageType = 'Basement' or GarageType =
'Carport') then drop_condition='03: Not Typical Garage Type';
	else if (BsmtFinType1 = 'LWQ' or BsmtFinType1 = 'NA') then drop_condition='04:
Not Typical Basement Type';
	else if (Neighborhood = 'Blmngtn') then drop_condition='05: Not Typical
Neighboorhood'; * I decided to drop neighborhoods with a 3% or less frequency;
	else if (6000 <= LotArea and LotArea >= 17400) then drop_condition='06: Not
Typical Lot Area';
	else if (858 <= GrLivArea and GrLivArea >= 2448) then drop_condition='07: Not
Typical GrLivArea';
	else if (FullBath ne 1 and FullBath ne 2) then drop_condition='08: Not Typical
Number of Full Baths';
	else if (BedroomAbvGr ne 2 and BedroomAbvGr ne 3 and BedroomAbvGr ne 4)then
drop_condition='09: Not Typical Number of Bedrooms';
	else if (MoSold = 1 or MoSold = 12) then drop_condition='10: Not Typical Month
Sold';
else drop_condition='11: Typical Housing Transaction';
run;

* View frequency of drop conditions;
proc freq data=temp2;
tables drop_condition;
title 'Typical Housing Transaction Waterfall';
run; quit;

* Subset data for just Typical Housing Transaction;
data typical;
set temp2;
if (drop_condition='11: Typical Housing Transaction');
run;

******************************************************************;
* Initial EDA for Modeling;
******************************************************************;

* Box plots for Categorical Data;
**(log(SalePrice) did not work)**;
***I'd like to eventually write a macro and perform a t-test to compare group means***;
```

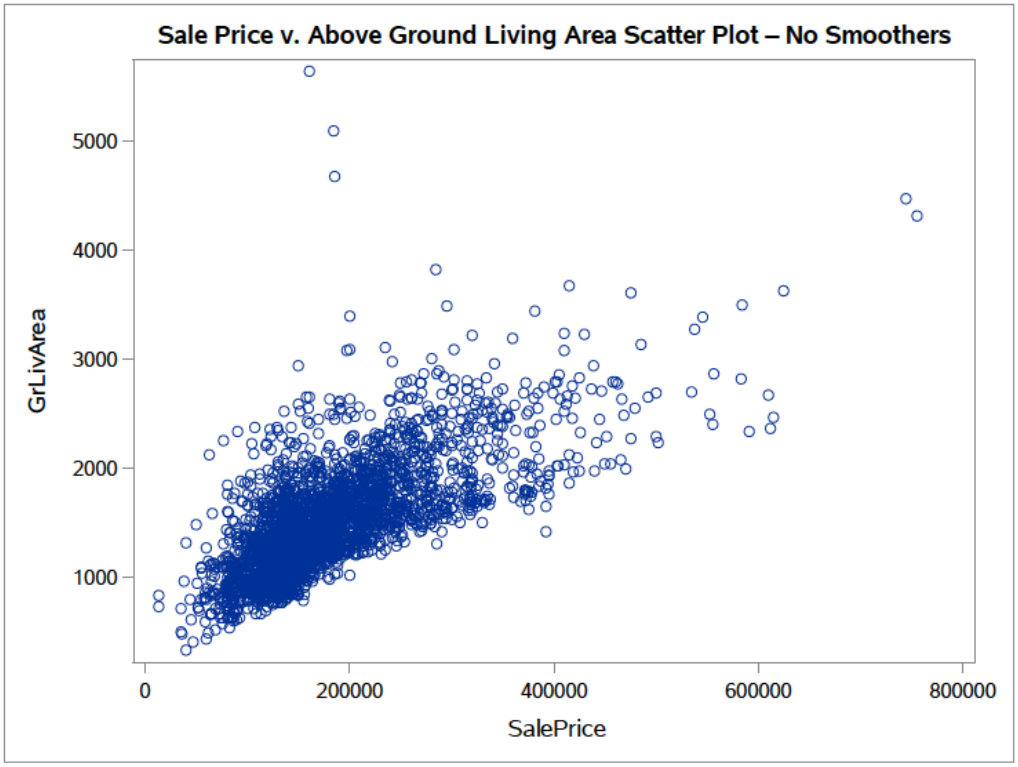
```
proc sort data=typical;
by Neighborhood;
run;
Title "Variable EDA - Boxplot of Neighborhoods";
proc boxplot data=typical;
plot SalePrice * Neighborhood;
run;

proc sort data=typical;
by HouseStyle;
run;
Title "Variable EDA - Boxplot of House Styles";
proc boxplot data=typical;
plot SalePrice * HouseStyle;
run;

proc sort data=typical;
by OverallQual;
run;
Title "Variable EDA - Boxplot of Overall Quality";
proc boxplot data=typical;
plot SalePrice * OverallQual;
run;

proc sort data=typical;
by BsmtFinType1;
run;
Title "Variable EDA - Boxplot of Basement Finished Area";
proc boxplot data=typical;
plot SalePrice * BsmtFinType1;
run;

proc sort data=typical;
by GarageType;
run;
Title "Variable EDA - Boxplot of Garage Type";
proc boxplot data=typical;
plot SalePrice * GarageType;
run;

* Scatter plots for Numeric Data;
Neighborhood HouseStyle OverallQual BsmtFinType1 GarageType
LotArea GrLivArea FullBath BedroomAbvGr MoSold

ods graphics on;
proc sgscatter data=typical;
```

```
compare x=(LotArea)
y=SalePrice / loess reg;
title;
run; quit;
ods graphics off;

ods graphics on;
proc sgscatter data=typical;
compare x=(GrLivArea)
y=SalePrice / loess reg;
run; quit;
ods graphics off;

ods graphics on;
proc sgscatter data=typical;
compare x=(FullBath)
y=SalePrice / loess reg;
run; quit;
ods graphics off;

ods graphics on;
proc sgscatter data=typical;
compare x=(BedroomAbvGr)
y=SalePrice / loess reg;
run; quit;
ods graphics off;

ods graphics on;
proc sgscatter data=typical;
compare x=(MoSold)
y=SalePrice / loess reg;
run; quit;
ods graphics off;

**FullBath BedroomAbvGr MoSold look better as box plots**;
proc sort data=typical;
by FullBath;
run;
Title "Variable EDA - Boxplot of Number of Full Bathrooms";
proc boxplot data=typical;
plot SalePrice * FullBath;
run;

proc sort data=typical;
by BedroomAbvGr;
run;
```

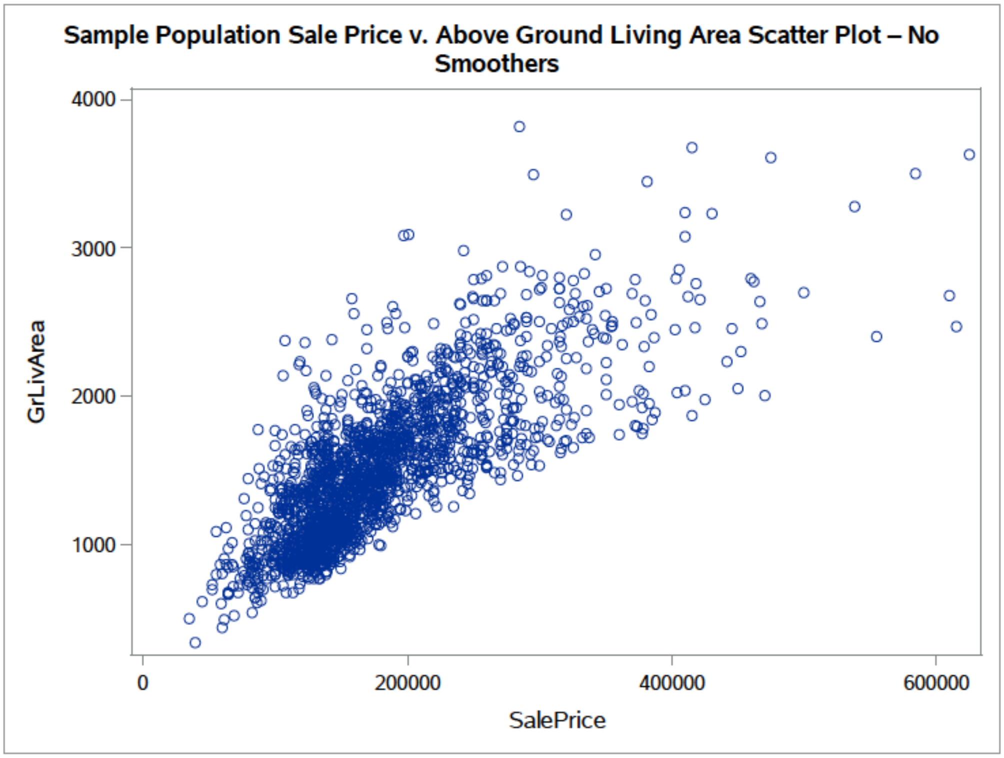
```
Title "Variable EDA - Boxplot of Number of Bedrooms Above Ground";
proc boxplot data=typical;
plot SalePrice * BedroomAbvGr;
run;

proc sort data=typical;
by MoSold;
run;
Title "Variable EDA - Boxplot of Month Sold";
proc boxplot data=typical;
plot SalePrice * MoSold;
run;

********************************************************************;
* END;
********************************************************************;
```

Appendix A

Pre-Sample Population Waterfall Scatter Plot



Post-Sample Population Waterfall Scatter Plot



Appendix B

Variables for Initial EDA:

Discrete and Continuous Variables

- LotArea
- GrLivArea
- LowQualFinSF
- FullBath
- HalfBath
- BedroomAbvGr
- MoSold

Nominal and Ordinal Variables

- Utilities
- Neighborhood
- Condition1
- HouseStyle
- OverallQual
- OverallCond
- BsmtFinType1
- Heating
- CentralAir
- KitchenQual
- GarageType
- GarageCars
- SaleCondition

The FREQ Procedure

Appendix B

Utilities	Frequency	Percent	Cumulative Frequency	Cumulative Percent
AllPub	1941	99.95	1941	99.95
NoSewr	1	0.05	1942	100.00

Neighborhood	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Blmngtn	1	0.05	1	0.05
BrkSide	96	4.94	97	4.99
ClearCr	37	1.91	134	6.90
CollgCr	213	10.97	347	17.87
Crawfor	78	4.02	425	21.88
Edwards	129	6.64	554	28.53
Gilbert	128	6.59	682	35.12
IDOTRR	50	2.57	732	37.69
Mitchel	83	4.27	815	41.97
NAmes	360	18.54	1175	60.50
NWAmes	113	5.82	1288	66.32
NoRidge	66	3.40	1354	69.72
NridgHt	67	3.45	1421	73.17
OldTown	176	9.06	1597	82.23
SWISU	34	1.75	1631	83.99
Sawyer	121	6.23	1752	90.22
SawyerW	89	4.58	1841	94.80
Somerst	22	1.13	1863	95.93
StoneBr	13	0.67	1876	96.60
Timber	49	2.52	1925	99.12
Veenker	17	0.88	1942	100.00

Condition1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Artery	64	3.30	64	3.30
Feedr	111	5.72	175	9.01
Norm	1654	85.17	1829	94.18
PosA	18	0.93	1847	95.11
PosN	34	1.75	1881	96.86
RRAe	20	1.03	1901	97.89
RRAn	32	1.65	1933	99.54

The FREQ Procedure

Appendix B

Condition1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
RRNe	4	0.21	1937	99.74
RRNn	5	0.26	1942	100.00

HouseStyle	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1.5Fin	245	12.62	245	12.62
1.5Unf	17	0.88	262	13.49
1Story	966	49.74	1228	63.23
2.5Fin	6	0.31	1234	63.54
2.5Unf	16	0.82	1250	64.37
2Story	543	27.96	1793	92.33
SFoyer	42	2.16	1835	94.49
SLvl	107	5.51	1942	100.00

OverallQual	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	2	0.10	2	0.10
2	9	0.46	11	0.57
3	24	1.24	35	1.80
4	134	6.90	169	8.70
5	620	31.93	789	40.63
6	511	26.31	1300	66.94
7	393	20.24	1693	87.18
8	187	9.63	1880	96.81
9	50	2.57	1930	99.38
10	12	0.62	1942	100.00

OverallCond	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	2	0.10	2	0.10
2	5	0.26	7	0.36
3	27	1.39	34	1.75
4	54	2.78	88	4.53
5	955	49.18	1043	53.71
6	405	20.85	1448	74.56
7	326	16.79	1774	91.35

The FREQ Procedure

Appendix B

OverallCond	Frequency	Percent	Cumulative Frequency	Cumulative Percent
8	132	6.80	1906	98.15
9	36	1.85	1942	100.00

BsmtFinType1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
ALQ	317	16.32	317	16.32
BLQ	221	11.38	538	27.70
GLQ	494	25.44	1032	53.14
LwQ	116	5.97	1148	59.11
NA	44	2.27	1192	61.38
Rec	225	11.59	1417	72.97
Unf	525	27.03	1942	100.00

Heating	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Floo	1	0.05	1	0.05
GasA	1913	98.51	1914	98.56
GasW	20	1.03	1934	99.59
Grav	5	0.26	1939	99.85
OthW	2	0.10	1941	99.95
Wall	1	0.05	1942	100.00

CentralAir	Frequency	Percent	Cumulative Frequency	Cumulative Percent
N	102	5.25	102	5.25
Y	1840	94.75	1942	100.00

KitchenQual	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Ex	92	4.74	92	4.74
Fa	51	2.63	143	7.36
Gd	742	38.21	885	45.57
Po	1	0.05	886	45.62
TA	1056	54.38	1942	100.00

Appendix B

The FREQ Procedure

GarageType	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2Types	12	0.62	12	0.62
Attchd	1166	60.04	1178	60.66
Basment	19	0.98	1197	61.64
BuiltIn	121	6.23	1318	67.87
CarPort	4	0.21	1322	68.07
Detchd	549	28.27	1871	96.34
NA	71	3.66	1942	100.00

GarageCars	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	71	3.66	71	3.66
1	604	31.10	675	34.76
2	1047	53.91	1722	88.67
3	212	10.92	1934	99.59
4	7	0.36	1941	99.95
5	1	0.05	1942	100.00

SaleCondition	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Normal	1942	100.00	1942	100.00

Appendix B

The UNIVARIATE Procedure
Variable: LotArea

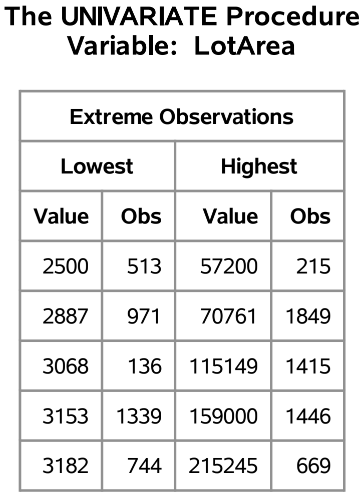
Moments			
N	1942	**Sum Weights**	1942
Mean	10842.7019	**Sum Observations**	21056527
Std Deviation	7853.13265	**Variance**	61671692.5
Skewness	14.7792982	**Kurtosis**	321.141366
Uncorrected SS	3.48014E11	**Corrected SS**	1.19705E11
Coeff Variation	72.4278207	**Std Error Mean**	178.204358

Basic Statistical Measures			
Location		**Variability**	
Mean	10842.70	**Std Deviation**	7853
Median	9758.50	**Variance**	61671692
Mode	9600.00	**Range**	212745
		Interquartile Range	3688

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	60.8442	**Pr > |t|**	<.0001
Sign	**M**	971	**Pr >= |M|**	<.0001
Signed Rank	**S**	943326.5	**Pr >= |S|**	<.0001

Quantiles (Definition 5)	
Level	**Quantile**
100% Max	215245.0
99%	31770.0
95%	17400.0
90%	14670.0
75% Q3	11851.0
50% Median	9758.5
25% Q1	8163.0
10%	6718.0
5%	6000.0
1%	4130.0
0% Min	2500.0

Appendix B



The UNIVARIATE Procedure
Variable: LotArea

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
2500	513	57200	215
2887	971	70761	1849
3068	136	115149	1415
3153	1339	159000	1446
3182	744	215245	669

Appendix B

The UNIVARIATE Procedure
Variable: GrLivArea

Moments			
N	1942	**Sum Weights**	1942
Mean	1489.62925	**Sum Observations**	2892860
Std Deviation	494.100631	**Variance**	244135.434
Skewness	0.86457518	**Kurtosis**	1.07320651
Uncorrected SS	4783155744	**Corrected SS**	473866877
Coeff Variation	33.1693696	**Std Error Mean**	11.212199

Basic Statistical Measures			
Location		**Variability**	
Mean	1489.629	**Std Deviation**	494.10063
Median	1440.000	**Variance**	244135
Mode	864.000	**Range**	3486
		Interquartile Range	645.00000

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	132.8579	**Pr > |t|**	<.0001
Sign	**M**	971	**Pr >= |M|**	<.0001
Signed Rank	**S**	943326.5	**Pr >= |S|**	<.0001

Quantiles (Definition 5)	
Level	**Quantile**
100% Max	3820
99%	2840
95%	2448
90%	2153
75% Q3	1752
50% Median	1440
25% Q1	1107
10%	907
5%	858
1%	672
0% Min	334

Appendix B

The UNIVARIATE Procedure
Variable: GrLivArea

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
334	1276	3500	1650
438	635	3608	1786
492	1775	3627	1646
498	1898	3672	1833
520	935	3820	1052

Appendix B

The UNIVARIATE Procedure
Variable: LowQualFinSF

Moments			
N	1942	**Sum Weights**	1942
Mean	4.98403708	**Sum Observations**	9679
Std Deviation	49.5985559	**Variance**	2460.01675
Skewness	12.1832863	**Kurtosis**	177.372948
Uncorrected SS	4823133	**Corrected SS**	4774892.51
Coeff Variation	995.148213	**Std Error Mean**	1.1254972

Basic Statistical Measures			
Location		**Variability**	
Mean	4.984037	**Std Deviation**	49.59856
Median	0.000000	**Variance**	2460
Mode	0.000000	**Range**	1064
		Interquartile Range	0

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	4.428298	**Pr > |t|**	<.0001
Sign	**M**	12.5	**Pr >= |M|**	<.0001
Signed Rank	**S**	162.5	**Pr >= |S|**	<.0001

Quantiles (Definition 5)	
Level	**Quantile**
100% Max	1064
99%	205
95%	0
90%	0
75% Q3	0
50% Median	0
25% Q1	0
10%	0
5%	0
1%	0
0% Min	0

Appendix B

The UNIVARIATE Procedure
Variable: LowQualFinSF

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	1942	514	1905
0	1941	528	1391
0	1940	572	1786
0	1939	697	945
0	1938	1064	444

Appendix B

The UNIVARIATE Procedure
Variable: FullBath

Moments			
N	1942	**Sum Weights**	1942
Mean	1.5030896	**Sum Observations**	2919
Std Deviation	0.53975468	**Variance**	0.29133512
Skewness	0.28290144	**Kurtosis**	-1.0545442
Uncorrected SS	4953	**Corrected SS**	565.481462
Coeff Variation	35.9096812	**Std Error Mean**	0.01224819

Basic Statistical Measures			
Location		Variability	
Mean	1.503090	**Std Deviation**	0.53975
Median	1.000000	**Variance**	0.29134
Mode	1.000000	**Range**	3.00000
		Interquartile Range	1.00000

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	122.7194	**Pr > |t|**	<.0001
Sign	**M**	968.5	**Pr >= |M|**	<.0001
Signed Rank	**S**	938476.5	**Pr >= |S|**	<.0001

Quantiles (Definition 5)	
Level	**Quantile**
100% Max	3
99%	3
95%	2
90%	2
75% Q3	2
50% Median	1
25% Q1	1
10%	1
5%	1
1%	1
0% Min	0

Appendix B

The UNIVARIATE Procedure
Variable: FullBath

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	1823	3	1650
0	1679	3	1680
0	1449	3	1750
0	520	3	1831
0	153	3	1833

Appendix B

The UNIVARIATE Procedure
Variable: HalfBath

Moments			
N	1942	**Sum Weights**	1942
Mean	0.37487127	**Sum Observations**	728
Std Deviation	0.48950538	**Variance**	0.23961552
Skewness	0.58322842	**Kurtosis**	-1.4871131
Uncorrected SS	738	**Corrected SS**	465.093718
Coeff Variation	130.579594	**Std Error Mean**	0.01110792

Basic Statistical Measures			
Location		**Variability**	
Mean	0.374871	**Std Deviation**	0.48951
Median	0.000000	**Variance**	0.23962
Mode	0.000000	**Range**	2.00000
		Interquartile Range	1.00000

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	33.7481	**Pr > |t|**	<.0001
Sign	**M**	361.5	**Pr >= |M|**	<.0001
Signed Rank	**S**	130863	**Pr >= |S|**	<.0001

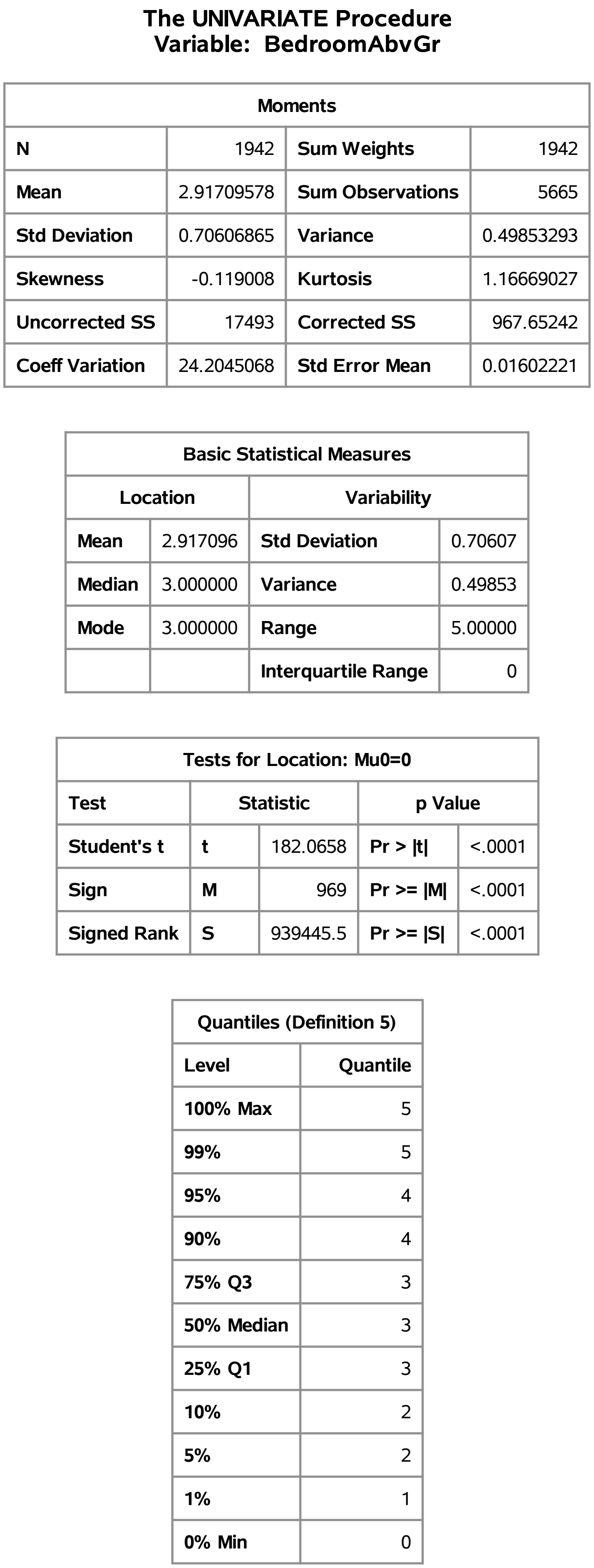
Quantiles (Definition 5)	
Level	**Quantile**
100% Max	2
99%	1
95%	1
90%	1
75% Q3	1
50% Median	0
25% Q1	0
10%	0
5%	0
1%	0
0% Min	0

Appendix B

The UNIVARIATE Procedure
Variable: HalfBath

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	1941	2	65
0	1940	2	381
0	1939	2	835
0	1938	2	1153
0	1937	2	1684

Appendix B



The UNIVARIATE Procedure
Variable: BedroomAbvGr

Moments			
N	1942	**Sum Weights**	1942
Mean	2.91709578	**Sum Observations**	5665
Std Deviation	0.70606865	**Variance**	0.49853293
Skewness	-0.119008	**Kurtosis**	1.16669027
Uncorrected SS	17493	**Corrected SS**	967.65242
Coeff Variation	24.2045068	**Std Error Mean**	0.01602221

Basic Statistical Measures			
Location		Variability	
Mean	2.917096	**Std Deviation**	0.70607
Median	3.000000	**Variance**	0.49853
Mode	3.000000	**Range**	5.00000
		Interquartile Range	0

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	**t**	182.0658	**Pr > |t|**	<.0001
Sign	**M**	969	**Pr >= |M|**	<.0001
Signed Rank	**S**	939445.5	**Pr >= |S|**	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	5
99%	5
95%	4
90%	4
75% Q3	3
50% Median	3
25% Q1	3
10%	2
5%	2
1%	1
0% Min	0

Appendix B

The UNIVARIATE Procedure
Variable: BedroomAbvGr

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	1823	5	1781
0	1679	5	1799
0	1449	5	1833
0	153	5	1838
1	1898	5	1897

Appendix B

The UNIVARIATE Procedure
Variable: MoSold

Moments			
N	1942	**Sum Weights**	1942
Mean	6.10916581	**Sum Observations**	11864
Std Deviation	2.61222672	**Variance**	6.82372841
Skewness	0.24580715	**Kurtosis**	-0.2680176
Uncorrected SS	85724	**Corrected SS**	13244.8568
Coeff Variation	42.7591393	**Std Error Mean**	0.05927701

Basic Statistical Measures			
Location		**Variability**	
Mean	6.109166	**Std Deviation**	2.61223
Median	6.000000	**Variance**	6.82373
Mode	6.000000	**Range**	11.00000
		Interquartile Range	3.00000

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	103.0613	**Pr > |t|**	<.0001
Sign	**M**	971	**Pr >= |M|**	<.0001
Signed Rank	**S**	943326.5	**Pr >= |S|**	<.0001

Quantiles (Definition 5)	
Level	**Quantile**
100% Max	12
99%	12
95%	11
90%	10
75% Q3	7
50% Median	6
25% Q1	4
10%	3
5%	2
1%	1
0% Min	1

The UNIVARIATE Procedure
Variable: MoSold

Appendix B

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
1	1929	12	1832
1	1862	12	1833
1	1844	12	1849
1	1821	12	1864
1	1809	12	1920

The MEANS Procedure

Appendix B

Variable	Minimum	Lower Quartile	Mean	Median	Upper Quartile	Maximum	Std Dev
LotArea	2500.00	8163.00	10842.70	9758.50	11851.00	215245.00	7853.13
GrLivArea	334.0000000	1107.00	1489.63	1440.00	1752.00	3820.00	494.1006313
LowQualFinSF	0	0	4.9840371	0	0	1064.00	49.5985559
FullBath	0	1.0000000	1.5030896	1.0000000	2.0000000	3.0000000	0.5397547
HalfBath	0	0	0.3748713	0	1.0000000	2.0000000	0.4895054
BedroomAbvGr	0	3.0000000	2.9170958	3.0000000	3.0000000	5.0000000	0.7060686
MoSold	1.0000000	4.0000000	6.1091658	6.0000000	7.0000000	12.0000000	2.6122267