Assignment #1

Getting to Know Your Data

Name: Young, Brent

Predict 410 Section #: 57

Quarter: Summer 2017

Introduction

Context

The dataset that we will be working with is called Ames Housing data (includes 2,930 rows) and is observational data collected by Ames Assessor's Office. The data includes houses sold in Ames, Iowa from 2006 to 2010 with SalePrice as the response variable and 81 predictors (includes nominal, ordinal, discrete, and continuous variables). The final goal is to build a Predictive model (e.g., multiple linear regression) to predict SalePrice of a house using other attributes. In order to accomplish this, an iterative regression process focused on statement of the problem, selection of potentially relevant variables, data collection, model specification, parameter estimation, model adequacy checking, model validation and model use will be conducted within the next five weeks.

Objectives/Purpose

The overall purpose/objective of assignment 1 is to understand and obtain a broad overview of the Ames housing data prior to building a predictive model to predict SalePrice of a house using other attributes such as physical characteristics of the house, surrounding areas, and condition of the house. This consists of three components: a data survey, a data quality check, and an initial exploratory data analysis. First, a waterfall of my drop conditions with counts will be provided to define the sample data/population of interest that we will want to use for the modeling purpose and ensure that the sample data is representative of the population that we want to model. Second, a table listing out my twenty variables and data quality results will be created to ensure that the data is clean, examining the data for potential errors, missing values, and outliers. Third, an initial exploratory data analysis (continuous and discrete) will be conducted by providing EDA results for my ten variables using scatterplots and boxplots to help understand important characteristics and properties of the data that may be disguised by numerical summaries (e.g., outliers, distribution, spread, skewness, and relationships between two quantitative variables). Finally, an initial exploratory data analysis for modeling will be conducted by providing EDA results for my three variables to explore the relationship between SalePrice and log(SalePrice). Ultimately, through EDA, potential difficulties/concerns for the model building process will be uncovered and potential transformations in the predictor variables may need to be conducted at some point during the model building process.

Section 1: Sample Definition

Figure 1: Boxplot of Sale Price & Building Style

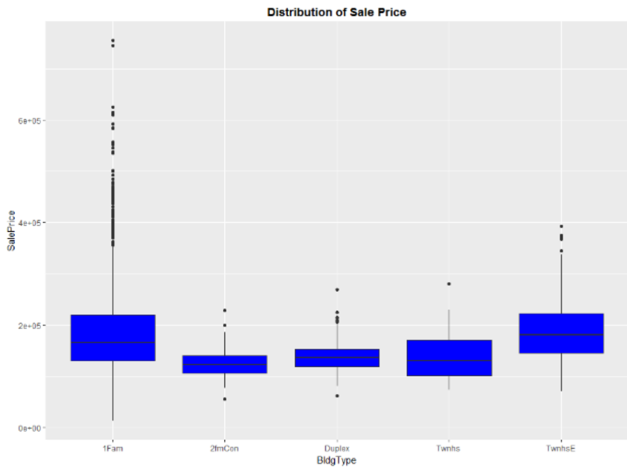

Figure 2: Boxplot of Sale Price & Sale Condition

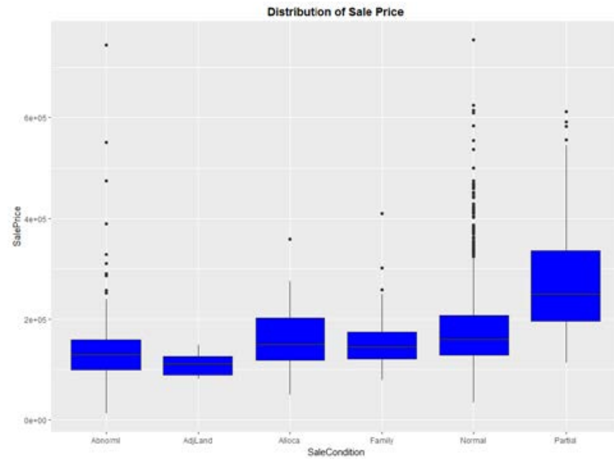

Figure 3: Waterfall of 'Drop Conditions'

```
         1Fam 2fmCon Duplex  Twnhs TwnhsE
 Before  2425     62    109    101    233

         Abnorml AdjLand  Alloca  Family  Normal Partial
Before   190      12      24      46    2413     245

         1Fam 2fmCon Duplex  Twnhs TwnhsE
 After   2002      0      0      0      0

         Abnorml AdjLand  Alloca  Family  Normal Partial
After      0        0        0       0     2002   0
```

Definition of Sample Data & Observations: Figure 1 shows a boxplot of SalePrice & Bldg Type and Figure 2 shows a boxplot of SalePrice & Sale Condition. When comparing figure 1 & 2, 'single-family' homes and 'normal' sale have similar medians as well as the amount and location of the outliers. As a result, based on this, it makes sense for the sample population/data of interest for 'typical' homes in Ames, Iowa to be 'single-family' homes with 'normal' sales in Ames, Iowa. Figure 3 shows the population of interest ('single family' homes and sale condition 'normal' in Ames, Iowa) after the drop conditions were applied, which comes out to 2002 rows and 81 variables.

Section 2: Data Quality Check

Figure 4: Correlation Matrix of Numeric Variables +/- 0.50

	OverallQual	YearBuilt	TotalBsmtSF	FirstFlrSF	GrLivArea	FullBath	TotRmsAbvGrd	GarageCars	GarageArea	TotalFloorSF	HouseAge	SalePrice
OverallQual	1	0.56	0.53	0.46	0.62	0.57	0.51	0.6	0.55	0.63	-0.56	0.8
YearBuilt	0.56	1	0.43	0.32	0.29	0.5	0.23	0.54	0.49	0.3	-1	0.57
TotalBsmtSF	0.53	0.43	1	0.77	0.4	0.34	0.26	0.44	0.46	0.4	-0.42	0.65
FirstFlrSF	0.46	0.32	0.77	1	0.53	0.38	0.36	0.44	0.46	0.53	-0.32	0.64
GrLivArea	0.62	0.29	0.4	0.53	1	0.67	0.83	0.52	0.49	1	-0.29	0.78
FullBath	0.57	0.5	0.34	0.38	0.67	1	0.59	0.52	0.45	0.67	-0.5	0.61
TotRmsAbvGrd	0.51	0.23	0.26	0.36	0.83	0.59	1	0.43	0.37	0.82	-0.23	0.6
GarageCars	0.6	0.54	0.44	0.44	0.52	0.52	0.43	1	0.88	0.53	-0.54	0.66
GarageArea	0.55	0.49	0.46	0.46	0.49	0.45	0.37	0.88	1	0.49	-0.49	0.64
TotalFloorSF	0.63	0.3	0.4	0.53	1	0.67	0.82	0.53	0.49	1	-0.3	0.78
HouseAge	-0.56	-1	-0.42	-0.32	-0.29	-0.5	-0.23	-0.54	-0.49	-0.3	1	-0.56
SalePrice	0.8	0.57	0.65	0.64	0.78	0.61	0.6	0.66	0.64	0.78	-0.56	1

Figure 5: Listing of 20 variables that were chosen

```
[1] "OverallQual"  "YearBuilt"    "TotalBsmtSF"  "FirstFlrSF"   "GrLivArea"
 [6] "FullBath"     "TotRmsAbvGrd" "GarageCars"   "GarageArea"   "TotalFloorSF"
[11] "HouseAge"     "SalePrice"    "LotConfig"    "Neighborhood" "Condition1"
[16] "HouseStyle"   "ExterCond"    "Heating"      "CentralAir"   "GarageType"
```

Observations: Figure 4 shows a Correlation Matrix of numeric variables that had correlations beyond at least +0.5 or –0.5. As a result, I went ahead and included these 12 variables (includes SalePrice) as part of the 20 variables that I chose. The data shows that all the variables were positively correlated between X and Y (Sale Price), except HouseAge. OverallQual, TotalFloorSF, and GrLivArea have the strongest positive correlations with SalePrice. The remaining 8 categorical variables that I chose are denoted in blue (see figure 5). I included these 8 categorical variables based on online research. After doing a data quality check (see Appendix), I did not see any missing values (e.g., SalePrice, except for GarageArea due to No Garage option in GarageType). However, I did notice outliers within the 20 variables. For instance, there were 5 houses that did not have a FullBath, SalePrice for one of the homes was $750000, while the lowest was $35000. I also noticed a small amount of FR2 and FR3 within LotConfig so it might make sense to combine them. I also noticed outliers within Heating & GarageType. For example, majority of the houses had GasA for heating and majority of GarageType was either attached or detached. I also noticed outliers and a “wide range” of values for the following variables: TotalBsmtSF, TotRmsAbvGrd, FirstFlrSF, GrLivArea, TotRmsAbvGrd, GarageCars, GarageArea, TotalFloorSF, YearBuilt, and HouseAge, and SalePrice. As we go on, we will have to investigate these outliers and decide what to do with them. For example, running diagnostic checks or conduct robust regression models to assign differing weights to data points depending on how it’s influencing the regression analysis.

Section 3: Initial Exploratory Data Analysis

Discrete Categorical Observations (Univariate EDA): After conducting a discrete EDA using barplots on GarageType, CentralAir, Heating, and HouseStyle, it appears that majority of the house types are either 1 story or 2 story houses with either attached or detached garages. Additionally, over 1750 houses have Central Air and nearly 2000 of the houses have Gas forced warm air furnaces. The EDA also showed that it may be a good idea to create an "other" category in GarageType, Heating, and HouseStyle in order to collapse the variables that didn't have a large count (see figures 6 & 7 below).

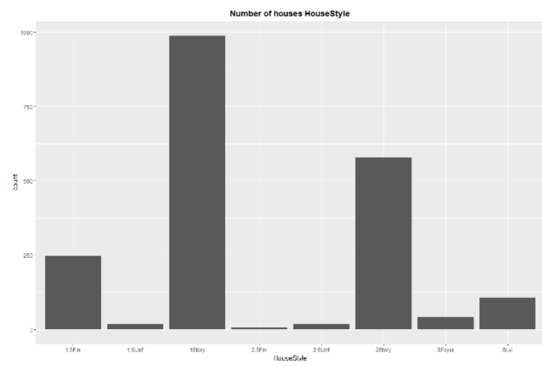

Figure 6: Number of Houses by HouseStyle

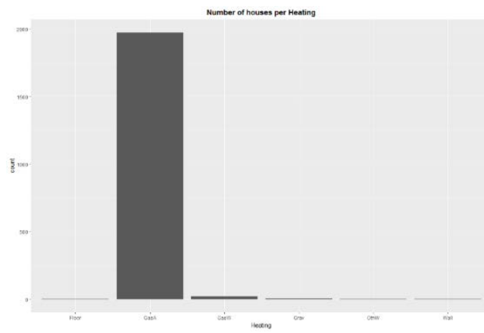

Figure 7: Number of Houses by Heating

Continuous Observations (Univariate EDA): After conducting a continuous EDA using histograms for OverallQual, TotalFloorSF, HouseAge, GarageArea,ToTRmsAbvGrd, and SalePrice, the results showed that OverallQual had a symmetric bell shape with a few outliers on the right & left side and a mean around 6 (figure 8). TotalFloorSF had some outliers on the right hand side (3000+), with a flat peak, slight right skew, with majority of the houses falling in between 1000 to 2000 square feet. HouseAge had noticeable outliers after 100+ years, a right skew, and majority of the houses falling in between the 0 to 50. GarageArea had some outliers on the right hand side around 1000+. Additionally, around 75 garages had GarageArea of 0, most likely due to the "N/A" option for GarageType. GarageArea also had a slight right skew, with majority of the garages falling in between 200 to 600 for GarageArea. TotRmsAbvGrd had symmetric bell shape with a few outliers on the right and left hand side, with majority of the houses falling in between 5 to 7 rooms (figure 9). SalePrice had noticeable outliers on the right tail and a few on the left tail, had a right skew, with majority of the houses falling in the 180k area (figure 10).

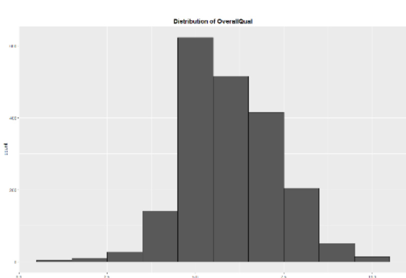

Figure 8: Distribution of OverallQual

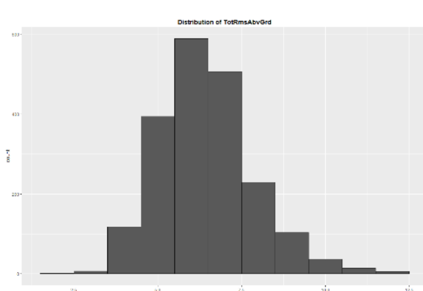

Figure 9: Distribution of TotRmsAbvGrd

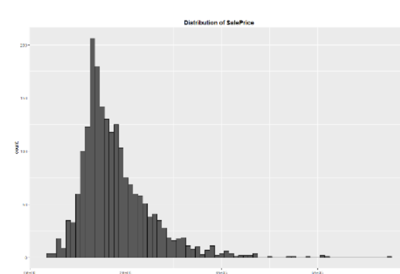

Figure 10: Distribution of SalePrice

Bivariate EDA Observations: Scatterplots of OverallQual vs. HouseAge showed a moderately negative correlation (lower the age, the better the quality; figure 11). Scatterplot of TotalFloorSF and TotRmsAbvGrd showed a positive correlation (more rooms, more square footage; figure 12). A scatterplot of GarageArea and HouseAge showed a moderately negative correlation (lower the age, the larger the garage area). Lastly a scatterplot of TotalFloorSF and HouseAge showed hardly any correlation, but interestingly the populations of old and newer houses were somewhat split (new houses on top and old houses on the bottom, with noticeable outliers on the top and right side of the scatterplot. The results of the scatterplots were confirmed from the correlation matrix (figure 4).

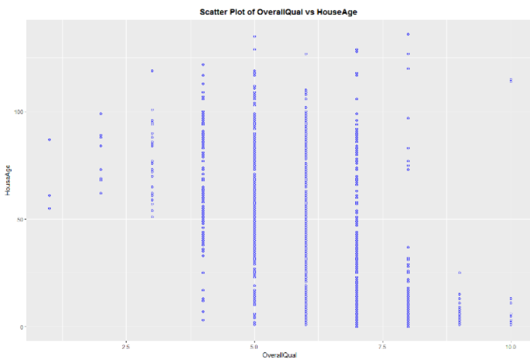

Figure 11: Scatterplot of OverallQual vs. HouseAge

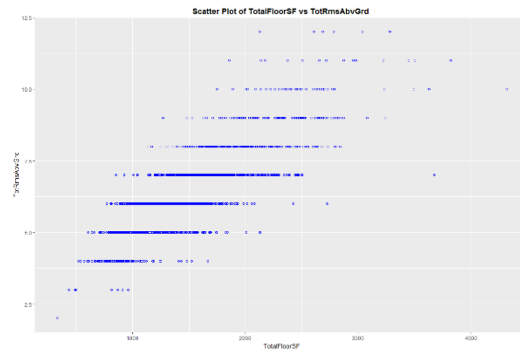

Figure 12: Scatterplot of TotalFloorSF vs. TotRmsAbvGrd

Additionally, I also ran boxplots of HouseAge vs. House Style, Garage Type, Heating, and Central Air. The results showed that majority of the older houses are 1.5Fin, 1.5Unf, 2.5Fin, and 2.5Unf, while majority of the new houses are 1Story, 2Story, SFoyer, and SLvl (figure 13). Additionally, most of the newer houses have either attached or built-in garages, while older houses have carports, detached and no garages at all. Furthermore, the boxplots revealed that newer houses have GasA and CentralAir, while the majority of older houses have GasW, Grav, OthW, and no CentralAir for heating and air conditoning. Furthermore, I also ran boxplots of TotalFloorSF vs. House Style, Garage Type, Heating, and Central Air. There were three insights from the boxplots that were producted. First, majority of the houses with high square feet were 2 story, while houses with lower square feet were 1 story (figure 14). Second, majority of the houses that had high square footage, had built-in garages. Furthermore, the majority of the houses with high square feet had central air and gas for heating.

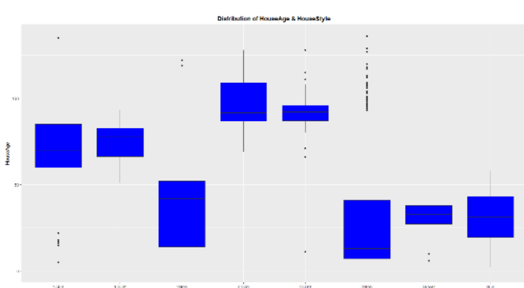

Figure 13: Boxplot of HouseAge & HouseStyle

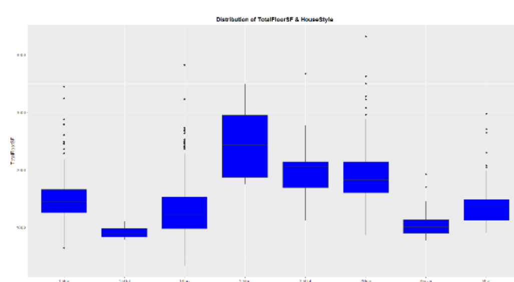

Figure 14: Boxplot of TotalFloorSF & HouseStyle

Section 4: Exploratory Data Analysis for Modeling

Variable: TotalFloorSF

Figure 15: SalePrice vs. TotalFloorSF

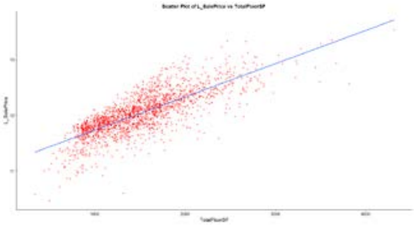

Figure 16: Log SalePrice vs. TotalFloorSF

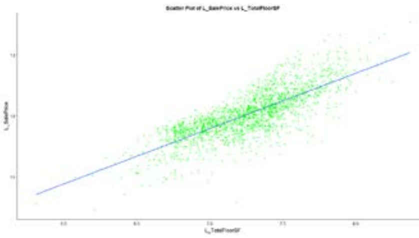

Figure 17: Log SalePrice vs. Log TotalFloorSF

Variable: OverallQual

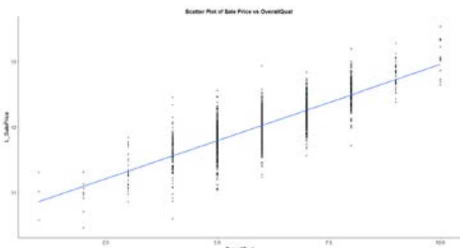

Figure 18: SalePrice vs. OverallQual

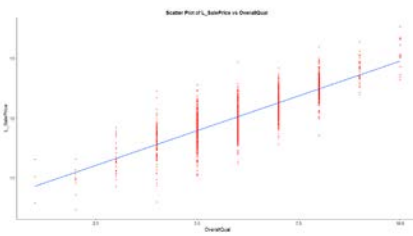

Figure 19: Log SalePrice vs. OverallQual

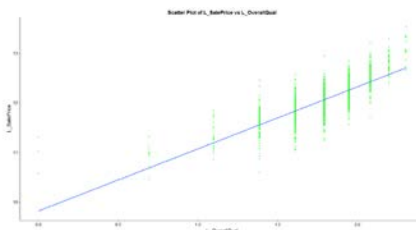

Figure 20: Log SalePrice vs. Log OverallQual

Variable: HouseAge

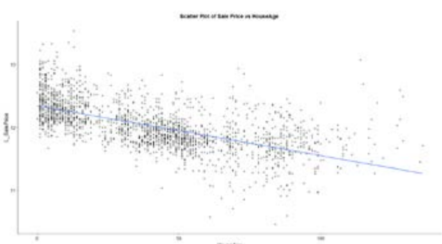

Figure 21: Log SalePrice vs. HouseAge

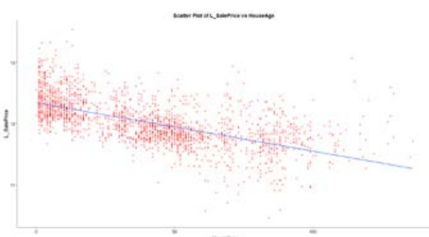

Figure 22: Log SalePrice vs. HouseAge

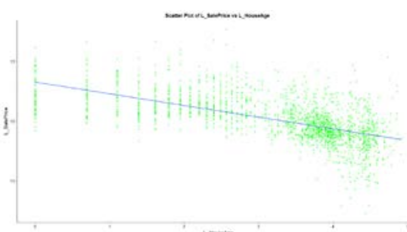

Figure 23: Log SalePrice vs. Log HouseAge

Observations: In regards to the variable TotalFloorSF, the scatterplot (figure 15) shows a “funnel” shape and heteroscedasticity, with a positive correlation between TotalFloorSF and SalePrice (as TotalFloorSF increases, SalePrice increases). In terms of the variable OverallQual, the scatterplot (figure 18) shows a positive correlation between OverallQual and SalePrice (as OverallQual increases, SalePrice increases), but does not show a nice linearly correlated relationship. Additionally, in regards to the variable HouseAge, the scatterplot (figure 21) shows a moderate negative correlation between HouseAge and SalePrice (as HouseAge decreases, SalePrice increases), but the data also does not show a nice linearly correlated relationship. As a result, this shows that one of the concerns for the model building process is the fact that these variables were not linearly correlated and in some cases heteroscedasticity was also evident. Furthermore, it’s also interesting to note that when Log SalePrice is used in conjunction with TotalFloorSF, OverallQual, and HouseAge, the dots in figures: 16, 19, and

22 do not change much, this could also be a potential concern. However, when log is used to transform SalePrice in addition to TotalFloorSF, OverallQual, and HouseAge (figures: 17, 20, and 23) it appears that the variability that we see in Log SalePrice for any choice of Log TotalFloorSF, Log OverallQual, and Log HouseAge decreases and the dots are slightly closer to the least-squares line, but are still not as close as we would like them to be. As a result, this illustrates that it may be beneficial to consider a transformation of SalePrice and also consider transformation in the predictor variables at some point in the model building process. By doing transformation, it will help achieve linearity, homogeneity of variance, and normality/symmetric about the regression equation.

Section 5: Summary/Conclusions

In section 1, we defined the sample population/data of interest for 'typical' homes in Ames, Iowa to be ‘single-family’ homes with ‘normal’ sales in Ames, Iowa using drop conditions and boxplots. In section 2, a Correlation Matrix was used to determine the variables that we would use for this assignment. The data showed that OverallQual, TotalFloorSF, and GrLivArea had the strongest positive correlations with SalePrice. The quality check also showed that there were not any missing values (e.g., SalePrice, except for GarageArea due to No Garage option in GarageType), but that there were outliers among the variables and opportunities to possibly combine categories into one. As we go on, we will have to investigate these outliers and decide what to do with them (e.g., run diagnostic checks or conduct robust regression models to assign differing weights to data). In section 3, an initial EDA (univariate and bivariate) on the discrete and continuous variables were completed. The initial EDA revealed shape, skewness, outliers, correlations between the variables and insightful insights. Lastly, in section 4, we conducted an EDA for modeling and saw that the scatterplots showed a lot of variability, some heteroscedasticity, and non-linear relationships, which are potential concerns for the model building process. Additionally, when Log SalePrice was used in conjunction with TotalFloorSF, OverallQual, and HouseAge, the dots did not get closer to the least-squares line. However, when log was used to transform SalePrice in addition to TotalFloorSF, OverallQual, and HouseAge it appeared that the variability that we saw in Log SalePrice for any choice of Log TotalFloorSF, Log OverallQual, and Log HouseAge decreased and the dots were slightly closer to the least-squares line, but are still not as close as we would like them to be. However, this improvement showed that there may be a need to consider transformations in the predictor variables at some point in the building process so that the model can achieve linearity, homogeneity of variance, and normality.

Appendix for Section (Data Quality Check)

```
> summary(subdat)
  OverallQual       YearBuilt     TotalBsmtSF       FirstFlrSF       GrLivArea   
 Min.   : 1.000   Min.   :1872   Min.   :   0.0   Min.   : 334.0   Min.   : 334  
 1st Qu.: 5.000   1st Qu.:1950   1st Qu.: 801.2   1st Qu.: 882.2   1st Qu.:1111  
 Median : 6.000   Median :1968   Median : 974.0   Median :1062.5   Median :1445  
 Mean   : 5.996   Mean   :1968   Mean   :1031.3   Mean   :1145.0   Mean   :1494  
 3rd Qu.: 7.000   3rd Qu.:1996   3rd Qu.:1228.0   3rd Qu.:1344.0   3rd Qu.:1762  
 Max.   :10.000   Max.   :2010   Max.   :3206.0   Max.   :3820.0   Max.   :4316  

    FullBath      TotRmsAbvGrd      GarageCars     GarageArea     TotalFloorSF 
 Min.   :0.000   Min.   : 2.000   Min.   :0.00   Min.   :   0   Min.   : 334  
 1st Qu.:1.000   1st Qu.: 5.000   1st Qu.:1.00   1st Qu.: 312   1st Qu.:1107  
 Median :1.000   Median : 6.000   Median :2.00   Median : 472   Median :1442  
 Mean   :1.512   Mean   : 6.437   Mean   :1.74   Mean   : 468   Mean   :1489  
 3rd Qu.:2.000   3rd Qu.: 7.000   3rd Qu.:2.00   3rd Qu.: 576   3rd Qu.:1755  
 Max.   :3.000   Max.   :12.000   Max.   :5.00   Max.   :1488   Max.   :4316  

    HouseAge        SalePrice         LotConfig     Neighborhood   Condition1  
 Min.   :  0.00   Min.   : 35000   Corner : 373   NAmes  :360   Norm   :1709  
 1st Qu.: 11.25   1st Qu.:130063   CulDSac: 139   CollgCr:213   Feedr  : 114  
 Median : 40.00   Median :161875   FR2    :  47   OldTown:177   Artery :  65  
 Mean   : 40.36   Mean   :179185   FR3    :   8   Edwards:129   PosN   :  34  
 3rd Qu.: 58.00   3rd Qu.:212450   Inside :1435   Gilbert:128   RRAn   :  32  
 Max.   :136.00   Max.   :755000                  Sawyer :121   RRAe   :  20  
                                                  (Other):874   (Other):  28  
   HouseStyle   ExterCond   Heating      CentralAir   GarageType  
 1Story :987   Ex:  10   Floor:   1   N: 109       2Types :  12  
 2Story :577   Fa:  40   GasA :1973   Y:1893       Attchd :1207  
 1.5Fin :248   Gd: 244   GasW :  20                Basment:  19  
 SLvl   :107   Po:   1   Grav :   5                BuiltIn: 124  
 SFoyer : 42   TA:1707   OthW :   2                CarPort:   5  
 1.5Unf : 18             Wall :   1                Detchd : 561  
 (Other): 23                                       NA's   :  74  
```

```
> library(Hmisc)
> describe(subdat)
subdat

 20  Variables      2002  Observations
---------------------------------------------------------------------------------
----
OverallQual
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25
    2002        0       10    0.943    5.996     1.46        4        5        5
     .50      .75      .90      .95
       6        7        8        8

Value          1     2     3     4     5     6     7     8     9    10
Frequency      3    10    27   141   622   515   415   204    51    14
Proportion 0.001 0.005 0.013 0.070 0.311 0.257 0.207 0.102 0.025 0.007
---------------------------------------------------------------------------------
----
YearBuilt
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25
    2002        0      113        1     1968    33.73     1915     1923     1950
     .50      .75      .90      .95
    1968     1996     2004     2006

lowest : 1872 1875 1879 1880 1882, highest: 2006 2007 2008 2009 2010
---------------------------------------------------------------------------------
----
TotalBsmtSF
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25
    2002        0      860        1     1031    430.4    456.6    644.4    801.2
     .50      .75      .90      .95
   974.0   1228.0   1568.0   1720.0

lowest :    0  105  160  173  190, highest: 2633 2846 3094 3200 3206
---------------------------------------------------------------------------------
----
FirstFlrSF
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25
    2002        0      899        1     1145    389.2    704.0    773.0    882.2
     .50      .75      .90      .95
  1062.5   1344.0   1651.9   1800.0

lowest :  334  432  438  442  448, highest: 2674 2696 2726 3228 3820
---------------------------------------------------------------------------------
----
GrLivArea
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25
    2002        0     1085        1     1494    545.5    858.1    907.0   1111.0
     .50      .75      .90      .95
  1445.0   1761.5   2141.8   2446.8

lowest :  334  438  492  498  520, highest: 3608 3627 3672 3820 4316
---------------------------------------------------------------------------------
----
FullBath
       n  missing distinct     Info     Mean      Gmd
```

```
    2002        0        4    0.765    1.512   0.5415

Value          0     1     2     3
Frequency      5  1003   957    37
Proportion 0.002 0.501 0.478 0.018
---------------------------------------------------------------------------------
----
TotRmsAbvGrd
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25
    2002        0       11    0.949    6.437    1.547        4        5        5
     .50      .75      .90      .95
       6        7        8        9

Value          2     3     4     5     6     7     8     9    10    11    12
Frequency      1     7   117   394   588   506   229   103    37    14     6
Proportion 0.000 0.003 0.058 0.197 0.294 0.253 0.114 0.051 0.018 0.007 0.003
---------------------------------------------------------------------------------
----
GarageCars
       n  missing distinct     Info     Mean      Gmd
    2002        0        6    0.813     1.74   0.7376

Value          0     1     2     3     4     5
Frequency     74   615  1080   225     7     1
Proportion 0.037 0.307 0.539 0.112 0.003 0.000
---------------------------------------------------------------------------------
----
GarageArea
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25
    2002        0      538        1      468    223.6    186.2    240.0    312.0
     .50      .75      .90      .95
   472.0    576.0    730.0    839.0

lowest :    0  100  160  162  164, highest: 1184 1231 1248 1314 1488
---------------------------------------------------------------------------------
----
TotalFloorSF
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25
    2002        0     1084        1     1489    543.3    856.1    904.0   1107.0
     .50      .75      .90      .95
  1442.0   1755.0   2133.9   2442.8

lowest :  334  438  492  498  520, highest: 3500 3627 3672 3820 4316
---------------------------------------------------------------------------------
----
HouseAge
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25
    2002        0      125        1    40.36    33.69     2.00     4.00    11.25
     .50      .75      .90      .95
   40.00    58.00    85.00    92.95

lowest :   0   1   2   3   4, highest: 127 128 129 135 136
---------------------------------------------------------------------------------
----
SalePrice
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25
    2002        0      722        1   179185    75562    95000   110000   130063
```

```
     .50      .75      .90      .95
  161875   212450   271450   316475

lowest :  35000  39300  40000  45000  52000, highest: 584500 610000 615000 625000
755000
---------------------------------------------------------------------------------
----
LotConfig
       n  missing distinct
    2002        0        5

Value       Corner CulDSac     FR2     FR3  Inside
Frequency      373     139      47       8    1435
Proportion   0.186   0.069   0.023   0.004   0.717
---------------------------------------------------------------------------------
----
Neighborhood
       n  missing distinct
    2002        0       21

lowest : Blmngtn BrkSide ClearCr CollgCr Crawfor, highest: Somerst StoneBr SWISU
Timber  Veenker
---------------------------------------------------------------------------------
----
Condition1
       n  missing distinct
    2002        0        9

Value      Artery  Feedr   Norm   PosA   PosN   RRAe   RRAn   RRNe   RRNn
Frequency      65    114   1709     18     34     20     32      4      6
Proportion  0.032  0.057  0.854  0.009  0.017  0.010  0.016  0.002  0.003
---------------------------------------------------------------------------------
----
HouseStyle
       n  missing distinct
    2002        0        8

Value      1.5Fin 1.5Unf 1Story 2.5Fin 2.5Unf 2Story SFoyer   SLvl
Frequency     248     18    987      6     17    577     42    107
Proportion  0.124  0.009  0.493  0.003  0.008  0.288  0.021  0.053
---------------------------------------------------------------------------------
----
ExterCond
       n  missing distinct
    2002        0        5

Value         Ex    Fa    Gd    Po    TA
Frequency     10    40   244     1  1707
Proportion 0.005 0.020 0.122 0.000 0.853
---------------------------------------------------------------------------------
----
Heating
       n  missing distinct
    2002        0        6

Value      Floor  GasA  GasW  Grav  OthW  Wall
Frequency      1  1973    20     5     2     1
```

```
Proportion 0.000 0.986 0.010 0.002 0.001 0.000
--------------------------------------------------------------------------------------
----
CentralAir
       n  missing distinct
    2002        0        2

Value          N     Y
Frequency    109  1893
Proportion 0.054 0.946
--------------------------------------------------------------------------------------
----
GarageType
       n  missing distinct
    1928       74        6

Value       2Types  Attchd Basment BuiltIn CarPort  Detchd
Frequency       12    1207      19     124       5     561
Proportion   0.006   0.626   0.010   0.064   0.003   0.291
--------------------------------------------------------------------------------------
----
```