/\*\* Import your XLSX file \*\*/

PROC IMPORT DATAFILE="/folders/myfolders/sasuser.v94/house\_price\_filtered.xlsx"

OUT=WORK.HOUSE\_PRICE\_FILTER\_OUT

DBMS=XLSX

REPLACE;

RUN;

/\* Run Means Procedure to check for Missing Values \*/

proc means data=House\_price\_filter\_out NMISS N;

RUN;

/\* identify data sets \*/

data House\_price\_filter\_out\_train;

set House\_price\_filter\_out;

/\* Remove observations 524 and 1299 \*/

if id = 524 then delete;

if id = 1299 then delete;

RUN;

/\*log transformation due to outliers \*/

data log\_House\_price\_filter\_out\_train;

set House\_price\_filter\_out\_train;

logGrLivArea = log(GrLivArea);

logSalePrice = log(SalePrice);

RUN;

/\*exploratory data analysis\*/

proc univariate data=log\_House\_price\_filter\_out\_train;

var SalePrice logSalePrice;

histogram SalePrice logSalePrice;

RUN;

proc univariate data=log\_House\_price\_filter\_out\_train;

var GrLivArea logGrLivArea;

histogram GrLivArea logGrLivArea;

RUN;

/\* frequency table to show size of neighborhoods in dataset \*/

proc freq data= House\_price\_filter\_out\_train;

tables Neighborhood

/plots=freqplot;

RUN;

/\*boxplots of saleprice by neighborhood\*/

proc sgplot data=House\_price\_filter\_out\_train;

hbox SalePrice / category=Neighborhood;

RUN;

/\*boxplots of GrLivArea by neighborhood \*/

proc sgplot data= House\_price\_filter\_out\_train;

hbox GrLivArea / category= Neighborhood;

RUN;

ods graphics on / attrpriority=none;

Title 'SalePrice by Neighborhood';

proc sgplot data=House\_price\_filter\_out\_train;

scatter y=SalePrice x=GrLivArea /group= Neighborhood markerattrs=(size=10px);

styleattrs datasymbols=(circle square star);

RUN;

proc glm data=House\_price\_filter\_out\_train plots =all;

class Neighborhood(ref = 'BrkSide');

model SalePrice = GrLivArea | Neighborhood /solution;

RUN;

/\* lm for log transformed data \*/

proc glm data=log\_House\_price\_filter\_out\_train plots =all;

class Neighborhood(ref = 'BrkSide');

model logSalePrice = logGrLivArea | Neighborhood /solution;

RUN;

**The MEANS Procedure**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Label** | **N Miss** | **N** |
| Id  MSSubClass  LotFrontage  LotArea  OverallQual  OverallCond  YearBuilt  YearRemodAdd  MasVnrArea  BsmtFinSF1  BsmtFinSF2  BsmtUnfSF  TotalBsmtSF  \_1stFlrSF  \_2ndFlrSF  LowQualFinSF  GrLivArea  BsmtFullBath  BsmtHalfBath  FullBath  HalfBath  BedroomAbvGr  KitchenAbvGr  TotRmsAbvGrd  Fireplaces  GarageYrBlt  GarageCars  GarageArea  WoodDeckSF  OpenPorchSF  EnclosedPorch  \_3SsnPorch  ScreenPorch  PoolArea  MiscVal  MoSold  YrSold  SalePrice | Id  MSSubClass  LotFrontage  LotArea  OverallQual  OverallCond  YearBuilt  YearRemodAdd  MasVnrArea  BsmtFinSF1  BsmtFinSF2  BsmtUnfSF  TotalBsmtSF  \_1stFlrSF  \_2ndFlrSF  LowQualFinSF  GrLivArea  BsmtFullBath  BsmtHalfBath  FullBath  HalfBath  BedroomAbvGr  KitchenAbvGr  TotRmsAbvGrd  Fireplaces  GarageYrBlt  GarageCars  GarageArea  WoodDeckSF  OpenPorchSF  EnclosedPorch  \_3SsnPorch  ScreenPorch  PoolArea  MiscVal  MoSold  YrSold  SalePrice | 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 | 383  383  383  383  383  383  383  383  383  383  383  383  383  383  383  383  383  383  383  383  383  383  383  383  383  383  383  383  383  383  383  383  383  383  383  383  383  383 |

**The UNIVARIATE Procedure**

**Variable: SalePrice (SalePrice)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Moments** |  |  |  | | | |
| **N** | | | | 381 | **Sum Weights** | 381 |
| **Mean** | | | | 137882.383 | **Sum Observations** | 52533188 |
| **Std Deviation** | | | | 38164.4224 | **Variance** | 1456523136 |
| **Skewness** | | | | 1.22174907 | **Kurtosis** | 4.10391607 |
| **Uncorrected SS** | | | | 7.79688E12 | **Corrected SS** | 5.53479E11 |
| **Coeff Variation** | | | | 27.6789692 | **Std Error Mean** | 1955.22259 |

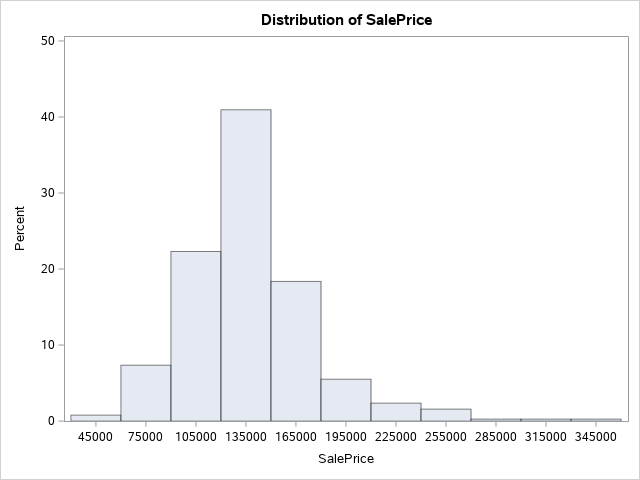
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Basic Statistical Measures** |  |  |  | | | | |
| **Location** |  | | | | **Variability** |  | |
| **Mean** | | | | 137882.4 | **Std Deviation** | | 38164 |
| **Median** | | | | 135000.0 | **Variance** | | 1456523136 |
| **Mode** | | | | 140000.0 | **Range** | | 305700 |
|  | | | |  | **Interquartile Range** | | 39000 |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Tests for Location: Mu0=0** |  |  |  |  | | | | | | |
| **Test** | | | | | **Statistic** |  | | **p Value** |  | |
| **Student's t** | | | | | **t** | | 70.52004 | **Pr > |t|** | | <.0001 |
| **Sign** | | | | | **M** | | 190.5 | **Pr >= |M|** | | <.0001 |
| **Signed Rank** | | | | | **S** | | 36385.5 | **Pr >= |S|** | | <.0001 |

|  |  |  |
| --- | --- | --- |
| **Quantiles (Definition 5)** |  | |
| **Level** | | **Quantile** |
| **100% Max** | | 345000 |
| **99%** | | 268000 |
| **95%** | | 207000 |
| **90%** | | 180000 |
| **75% Q3** | | 155000 |
| **50% Median** | | 135000 |
| **25% Q1** | | 116000 |
| **10%** | | 97000 |
| **5%** | | 82500 |
| **1%** | | 60000 |
| **0% Min** | | 39300 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Extreme Observations** |  |  |  | | | | |
| **Lowest** |  | | | | **Highest** |  | |
| **Value** | | | | **Obs** | **Value** | | **Obs** |
| 39300 | | | | 22 | 244000 | | 349 |
| 52000 | | | | 31 | 268000 | | 290 |
| 58500 | | | | 151 | 274970 | | 153 |
| 60000 | | | | 84 | 320000 | | 107 |
| 60000 | | | | 28 | 345000 | | 255 |

**The UNIVARIATE Procedure**



**The UNIVARIATE Procedure**

**Variable: logSalePrice**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Moments** |  |  |  | | | |
| **N** | | | | 381 | **Sum Weights** | 381 |
| **Mean** | | | | 11.7975216 | **Sum Observations** | 4494.85573 |
| **Std Deviation** | | | | 0.27349706 | **Variance** | 0.07480064 |
| **Skewness** | | | | -0.3226264 | **Kurtosis** | 1.97549745 |
| **Uncorrected SS** | | | | 53056.5818 | **Corrected SS** | 28.4242444 |
| **Coeff Variation** | | | | 2.31825863 | **Std Error Mean** | 0.01401168 |

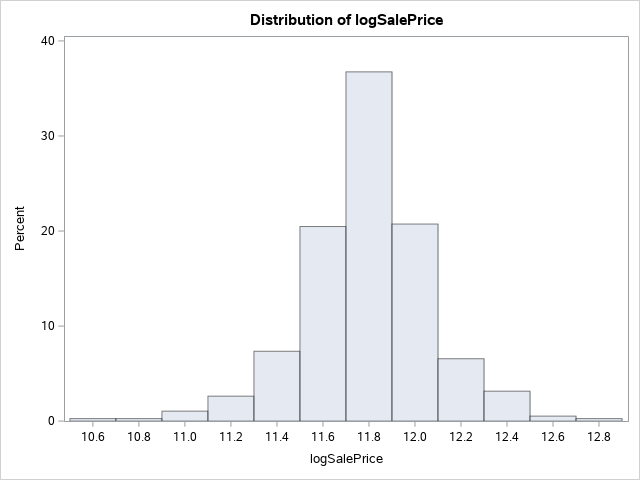
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Basic Statistical Measures** |  |  |  | | | | |
| **Location** |  | | | | **Variability** |  | |
| **Mean** | | | | 11.79752 | **Std Deviation** | | 0.27350 |
| **Median** | | | | 11.81303 | **Variance** | | 0.07480 |
| **Mode** | | | | 11.84940 | **Range** | | 2.17232 |
|  | | | |  | **Interquartile Range** | | 0.28983 |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Tests for Location: Mu0=0** |  |  |  |  | | | | | | |
| **Test** | | | | | **Statistic** |  | | **p Value** |  | |
| **Student's t** | | | | | **t** | | 841.9777 | **Pr > |t|** | | <.0001 |
| **Sign** | | | | | **M** | | 190.5 | **Pr >= |M|** | | <.0001 |
| **Signed Rank** | | | | | **S** | | 36385.5 | **Pr >= |S|** | | <.0001 |

|  |  |  |
| --- | --- | --- |
| **Quantiles (Definition 5)** |  | |
| **Level** | | **Quantile** |
| **100% Max** | | 12.7513 |
| **99%** | | 12.4987 |
| **95%** | | 12.2405 |
| **90%** | | 12.1007 |
| **75% Q3** | | 11.9512 |
| **50% Median** | | 11.8130 |
| **25% Q1** | | 11.6613 |
| **10%** | | 11.4825 |
| **5%** | | 11.3206 |
| **1%** | | 11.0021 |
| **0% Min** | | 10.5790 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Extreme Observations** |  |  |  | | | | |
| **Lowest** |  | | | | **Highest** |  | |
| **Value** | | | | **Obs** | **Value** | | **Obs** |
| 10.5790 | | | | 22 | 12.4049 | | 349 |
| 10.8590 | | | | 31 | 12.4987 | | 290 |
| 10.9768 | | | | 151 | 12.5244 | | 153 |
| 11.0021 | | | | 84 | 12.6761 | | 107 |
| 11.0021 | | | | 28 | 12.7513 | | 255 |

**The UNIVARIATE Procedure**



**The UNIVARIATE Procedure**

**Variable: GrLivArea (GrLivArea)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Moments** |  |  |  | | | |
| **N** | | | | 381 | **Sum Weights** | 381 |
| **Mean** | | | | 1281.5853 | **Sum Observations** | 488284 |
| **Std Deviation** | | | | 396.265051 | **Variance** | 157025.991 |
| **Skewness** | | | | 1.15280014 | **Kurtosis** | 2.05121043 |
| **Uncorrected SS** | | | | 685447474 | **Corrected SS** | 59669876.5 |
| **Coeff Variation** | | | | 30.9199123 | **Std Error Mean** | 20.3012736 |

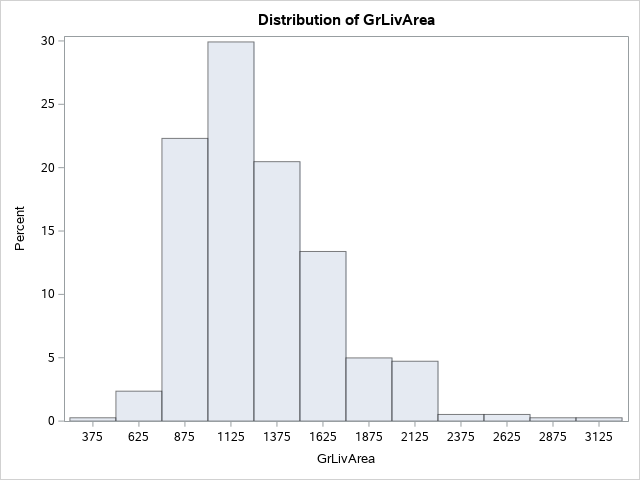
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Basic Statistical Measures** |  |  |  | | | | |
| **Location** |  | | | | **Variability** |  | |
| **Mean** | | | | 1281.585 | **Std Deviation** | | 396.26505 |
| **Median** | | | | 1200.000 | **Variance** | | 157026 |
| **Mode** | | | | 864.000 | **Range** | | 2778 |
|  | | | |  | **Interquartile Range** | | 484.00000 |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Tests for Location: Mu0=0** |  |  |  |  | | | | | | |
| **Test** | | | | | **Statistic** |  | | **p Value** |  | |
| **Student's t** | | | | | **t** | | 63.12832 | **Pr > |t|** | | <.0001 |
| **Sign** | | | | | **M** | | 190.5 | **Pr >= |M|** | | <.0001 |
| **Signed Rank** | | | | | **S** | | 36385.5 | **Pr >= |S|** | | <.0001 |

|  |  |  |
| --- | --- | --- |
| **Quantiles (Definition 5)** |  | |
| **Level** | | **Quantile** |
| **100% Max** | | 3112 |
| **99%** | | 2634 |
| **95%** | | 2108 |
| **90%** | | 1812 |
| **75% Q3** | | 1486 |
| **50% Median** | | 1200 |
| **25% Q1** | | 1002 |
| **10%** | | 864 |
| **5%** | | 816 |
| **1%** | | 616 |
| **0% Min** | | 334 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Extreme Observations** |  |  |  | | | | |
| **Lowest** |  | | | | **Highest** |  | |
| **Value** | | | | **Obs** | **Value** | | **Obs** |
| 334 | | | | 22 | 2380 | | 262 |
| 520 | | | | 3 | 2634 | | 362 |
| 605 | | | | 91 | 2704 | | 255 |
| 616 | | | | 30 | 2978 | | 205 |
| 672 | | | | 139 | 3112 | | 190 |

**The UNIVARIATE Procedure**



**The UNIVARIATE Procedure**

**Variable: logGrLivArea**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Moments** |  |  |  | | | |
| **N** | | | | 381 | **Sum Weights** | 381 |
| **Mean** | | | | 7.11183407 | **Sum Observations** | 2709.60878 |
| **Std Deviation** | | | | 0.29509401 | **Variance** | 0.08708048 |
| **Skewness** | | | | 0.09937127 | **Kurtosis** | 0.73908727 |
| **Uncorrected SS** | | | | 19303.3786 | **Corrected SS** | 33.0905808 |
| **Coeff Variation** | | | | 4.14933769 | **Std Error Mean** | 0.01511812 |

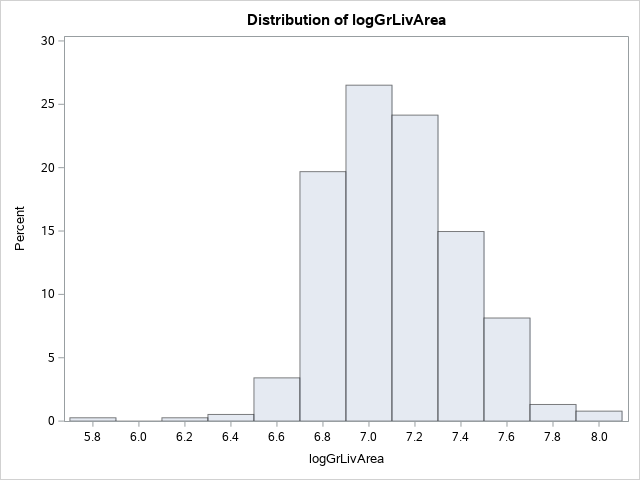
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Basic Statistical Measures** |  |  |  | | | | |
| **Location** |  | | | | **Variability** |  | |
| **Mean** | | | | 7.111834 | **Std Deviation** | | 0.29509 |
| **Median** | | | | 7.090077 | **Variance** | | 0.08708 |
| **Mode** | | | | 6.761573 | **Range** | | 2.23188 |
|  | | | |  | **Interquartile Range** | | 0.39409 |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Tests for Location: Mu0=0** |  |  |  |  | | | | | | |
| **Test** | | | | | **Statistic** |  | | **p Value** |  | |
| **Student's t** | | | | | **t** | | 470.4178 | **Pr > |t|** | | <.0001 |
| **Sign** | | | | | **M** | | 190.5 | **Pr >= |M|** | | <.0001 |
| **Signed Rank** | | | | | **S** | | 36385.5 | **Pr >= |S|** | | <.0001 |

|  |  |  |
| --- | --- | --- |
| **Quantiles (Definition 5)** |  | |
| **Level** | | **Quantile** |
| **100% Max** | | 8.04302 |
| **99%** | | 7.87626 |
| **95%** | | 7.65349 |
| **90%** | | 7.50219 |
| **75% Q3** | | 7.30384 |
| **50% Median** | | 7.09008 |
| **25% Q1** | | 6.90975 |
| **10%** | | 6.76157 |
| **5%** | | 6.70441 |
| **1%** | | 6.42325 |
| **0% Min** | | 5.81114 |

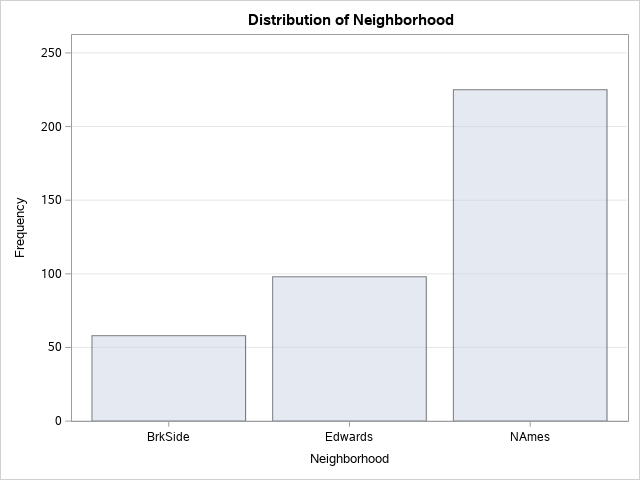
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Extreme Observations** |  |  |  | | | | |
| **Lowest** |  | | | | **Highest** |  | |
| **Value** | | | | **Obs** | **Value** | | **Obs** |
| 5.81114 | | | | 22 | 7.77486 | | 262 |
| 6.25383 | | | | 3 | 7.87626 | | 362 |
| 6.40523 | | | | 91 | 7.90249 | | 255 |
| 6.42325 | | | | 30 | 7.99901 | | 205 |
| 6.51026 | | | | 139 | 8.04302 | | 190 |

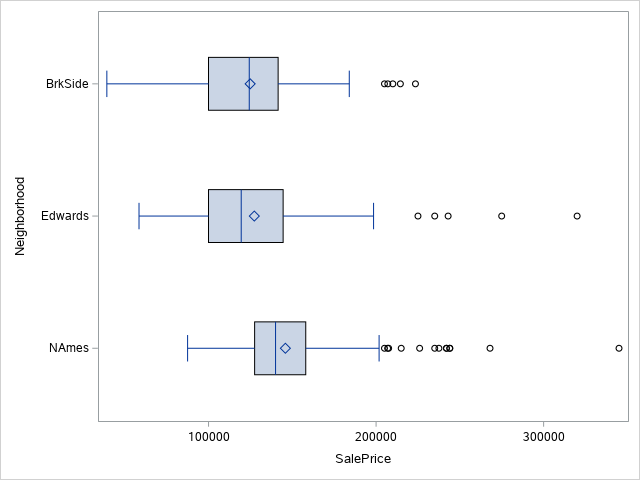
**The UNIVARIATE Procedure**

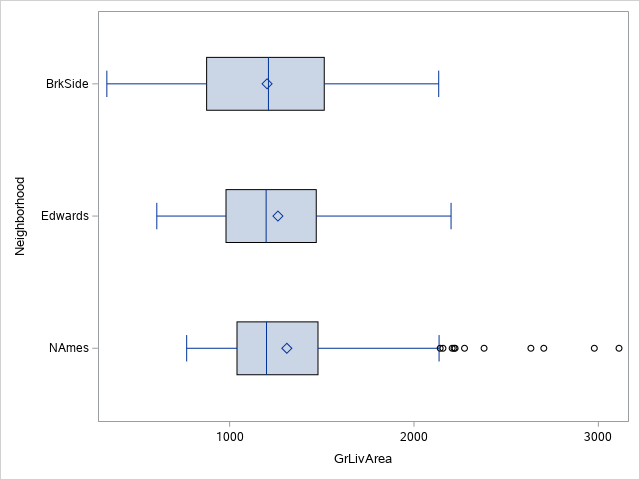


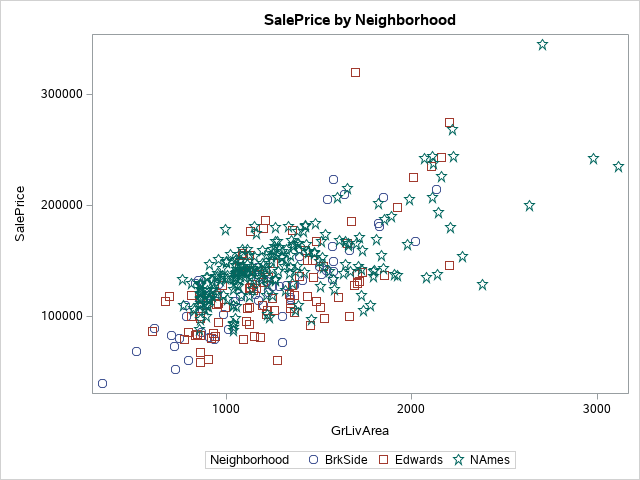
**The FREQ Procedure**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Neighborhood** |  |  |  |  | | | | |
| **Neighborhood** | | | | | **Frequency** | **Percent** | **Cumulative**  **Frequency** | **Cumulative**  **Percent** |
| **BrkSide** | | | | | 58 | 15.22 | 58 | 15.22 |
| **Edwards** | | | | | 98 | 25.72 | 156 | 40.94 |
| **NAmes** | | | | | 225 | 59.06 | 381 | 100.00 |









**SalePrice by Neighborhood**

**The GLM Procedure**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class Level Information** |  |  | | |
| **Class** | | | **Levels** | **Values** |
| **Neighborhood** | | | 3 | Edwards NAmes BrkSide |

|  |  |
| --- | --- |
| **Number of Observations Read** | 381 |
| **Number of Observations Used** | 381 |

**SalePrice by Neighborhood**

**The GLM Procedure**

**Dependent Variable: SalePrice SalePrice**

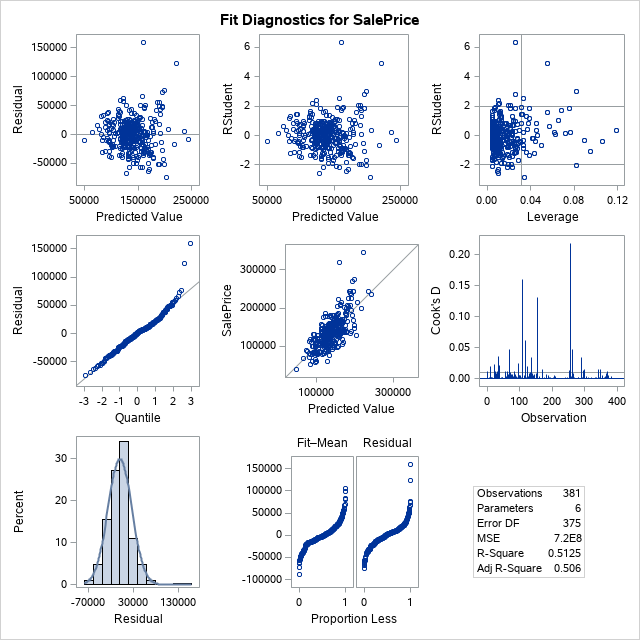
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 5 | 283631588727 | 56726317745 | 78.83 | <.0001 |
| **Error** | 375 | 269847202943 | 719592541.18 |  |  |
| **Corrected Total** | 380 | 553478791670 |  |  |  |

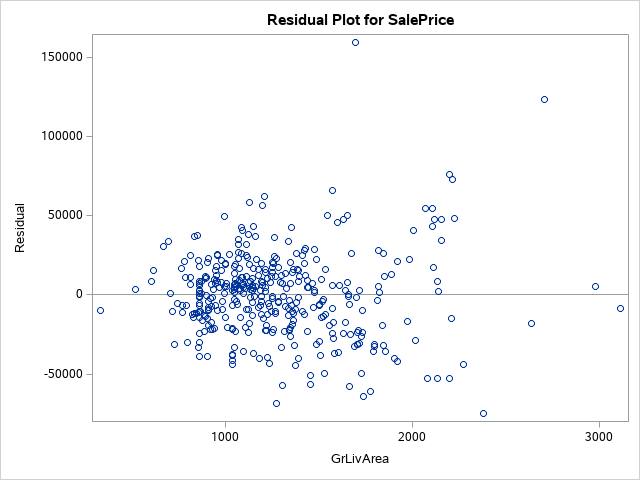
|  |  |  |  |
| --- | --- | --- | --- |
| **R-Square** | **Coeff Var** | **Root MSE** | **SalePrice Mean** |
| 0.512452 | 19.45515 | 26825.22 | 137882.4 |

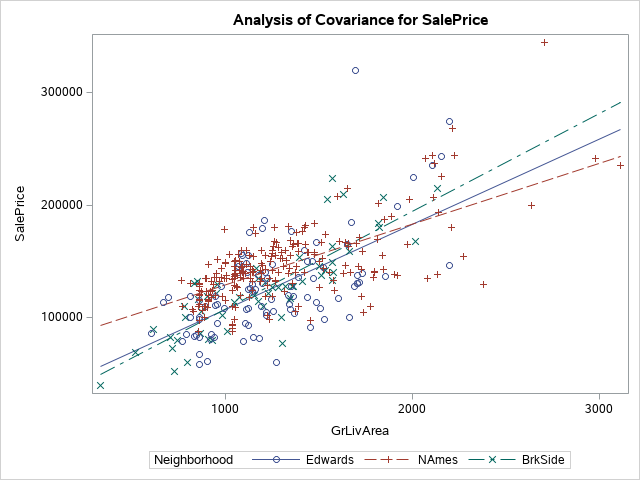
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Type I SS** | **Mean Square** | **F Value** | **Pr > F** |
| **GrLivArea** | 1 | 253099925462 | 253099925462 | 351.73 | <.0001 |
| **Neighborhood** | 2 | 20620955996 | 10310477998 | 14.33 | <.0001 |
| **GrLivArea\*Neighborho** | 2 | 9910707269.1 | 4955353634.6 | 6.89 | 0.0012 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| **GrLivArea** | 1 | 210178319563 | 210178319563 | 292.08 | <.0001 |
| **Neighborhood** | 2 | 18323983921 | 9161991960.5 | 12.73 | <.0001 |
| **GrLivArea\*Neighborho** | 2 | 9910707269.1 | 4955353634.6 | 6.89 | 0.0012 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **Estimate** |  | **Standard**  **Error** | **t Value** | **Pr > |t|** |
| **Intercept** | 19971.51379 | B | 11604.02869 | 1.72 | 0.0861 |
| **GrLivArea** | 87.16253 | B | 9.19027 | 9.48 | <.0001 |
| **Neighborhood Edwards** | 11457.04080 | B | 15343.23590 | 0.75 | 0.4557 |
| **Neighborhood NAmes** | 54704.88774 | B | 13042.61747 | 4.19 | <.0001 |
| **Neighborhood BrkSide** | 0.00000 | B | . | . | . |
| **GrLivArea\*Neighborho Edwards** | -11.18610 | B | 11.96277 | -0.94 | 0.3504 |
| **GrLivArea\*Neighborho NAmes** | -32.84667 | B | 10.16117 | -3.23 | 0.0013 |
| **GrLivArea\*Neighborho BrkSide** | 0.00000 | B | . | . | . |
| **Note:** | | | The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable. | | |







**SalePrice by Neighborhood**

**The GLM Procedure**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class Level Information** |  |  | | |
| **Class** | | | **Levels** | **Values** |
| **Neighborhood** | | | 3 | Edwards NAmes BrkSide |

|  |  |
| --- | --- |
| **Number of Observations Read** | 381 |
| **Number of Observations Used** | 381 |

**SalePrice by Neighborhood**

**The GLM Procedure**

**Dependent Variable: logSalePrice**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 5 | 15.00591625 | 3.00118325 | 83.87 | <.0001 |
| **Error** | 375 | 13.41832813 | 0.03578221 |  |  |
| **Corrected Total** | 380 | 28.42424438 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **R-Square** | **Coeff Var** | **Root MSE** | **logSalePrice Mean** |
| 0.527927 | 1.603403 | 0.189162 | 11.79752 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Type I SS** | **Mean Square** | **F Value** | **Pr > F** |
| **logGrLivArea** | 1 | 12.59227060 | 12.59227060 | 351.91 | <.0001 |
| **Neighborhood** | 2 | 1.73746625 | 0.86873313 | 24.28 | <.0001 |
| **logGrLivA\*Neighborho** | 2 | 0.67617940 | 0.33808970 | 9.45 | <.0001 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| **logGrLivArea** | 1 | 11.78567132 | 11.78567132 | 329.37 | <.0001 |
| **Neighborhood** | 2 | 0.75967130 | 0.37983565 | 10.62 | <.0001 |
| **logGrLivA\*Neighborho** | 2 | 0.67617940 | 0.33808970 | 9.45 | <.0001 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **Estimate** |  | **Standard**  **Error** | **t Value** | **Pr > |t|** |
| **Intercept** | 5.912920736 | B | 0.49641696 | 11.91 | <.0001 |
| **logGrLivArea** | 0.819648056 | B | 0.07046839 | 11.63 | <.0001 |
| **Neighborhood Edwards** | 1.010172153 | B | 0.69820714 | 1.45 | 0.1488 |
| **Neighborhood NAmes** | 2.579806905 | B | 0.59016472 | 4.37 | <.0001 |
| **Neighborhood BrkSide** | 0.000000000 | B | . | . | . |
| **logGrLivA\*Neighborho Edwards** | -0.146310441 | B | 0.09867902 | -1.48 | 0.1390 |
| **logGrLivA\*Neighborho NAmes** | -0.346624454 | B | 0.08344620 | -4.15 | <.0001 |
| **logGrLivA\*Neighborho BrkSide** | 0.000000000 | B | . | . | . |
| **Note:** | | | The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable. | | |

