Project 1

Wentian Chen, Yufei Zhang(Amber), Peiyu Wang

2024-10-28

#Introduction

Our group used the Ames Housing dataset, which provides comprehensive property details from Ames, Iowa and we used this dataset to analyze housing price determinants. Our group project focuses on estimating a linear regression model, where SalePrice as the dependent variable is predicted based on characteristics like LotFrontage (street connectivity in feet), WoodDeckSF (wood deck square footage), GarageArea (garage square footage), and GrLivArea (above-ground living area). This model focuses on quantifying each feature's influence on housing prices, assisting us with identifying significant predictors of property value and better understanding the characteristics contributing to price variations in Ames. Our group is estimating a linear regression model in order to understand the relationship between SalePrice and each predictor. Our main purpose is to explain the impact of each factor on sales price, assess statistical significance, and validate model assumptions. Techniques which are involved include summary statistics, normality transformation, confidence intervals, and bootstrap estimates to ensure robustness.

The Ames Housing dataset was created by Dean De Cock, a professor at Iowa State University for educational purposes and is now widely available for predictive modeling in real estate, including through Kaggle. It serves as a credible benchmark dataset in housing price analysis and machine learning applications (De Cock, 2011). Reference: De Cock, Dean. "Ames, Iowa: Alternative to the Boston Housing Data as an End-of-Semester Regression Project." Journal of Statistics Education, vol. 19, no. 3, 2011.

```
getwd()
```

```
## [1] "C:/Users/WANGPEIYU/Downloads"
```

```
## [1] "C:/Program Files/R"
setwd("C:/Program Files/R")

library(readxl)
data <- read.csv("C:/Program Files/R/train.csv")
head(data)
```

```
##   Id MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape LandContour
## 1  1         60       RL          65    8450   Pave  <NA>      Reg         Lvl
## 2  2         20       RL          80    9600   Pave  <NA>      Reg         Lvl
## 3  3         60       RL          68   11250   Pave  <NA>      IR1         Lvl
## 4  4         70       RL          60    9550   Pave  <NA>      IR1         Lvl
## 5  5         60       RL          84   14260   Pave  <NA>      IR1         Lvl
## 6  6         50       RL          85   14115   Pave  <NA>      IR1         Lvl
##   Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType
## 1    AllPub    Inside       Gtl      CollgCr       Norm       Norm     1Fam
## 2    AllPub       FR2       Gtl      Veenker      Feedr       Norm     1Fam
## 3    AllPub    Inside       Gtl      CollgCr       Norm       Norm     1Fam
## 4    AllPub    Corner       Gtl      Crawfor       Norm       Norm     1Fam
## 5    AllPub       FR2       Gtl      NoRidge       Norm       Norm     1Fam
## 6    AllPub    Inside       Gtl      Mitchel       Norm       Norm     1Fam
##   HouseStyle OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle RoofMatl
## 1     2Story           7           5      2003         2003     Gable  CompShg
## 2     1Story           6           8      1976         1976     Gable  CompShg
## 3     2Story           7           5      2001         2002     Gable  CompShg
## 4     2Story           7           5      1915         1970     Gable  CompShg
## 5     2Story           8           5      2000         2000     Gable  CompShg
## 6     1.5Fin           5           5      1993         1995     Gable  CompShg
##   Exterior1st Exterior2nd MasVnrType MasVnrArea ExterQual ExterCond Foundation
## 1     VinylSd     VinylSd    BrkFace        196        Gd        TA      PConc
## 2     MetalSd     MetalSd       None          0        TA        TA     CBlock
## 3     VinylSd     VinylSd    BrkFace        162        Gd        TA      PConc
## 4     Wd Sdng     Wd Shng       None          0        TA        TA     BrkTil
## 5     VinylSd     VinylSd    BrkFace        350        Gd        TA      PConc
## 6     VinylSd     VinylSd       None          0        TA        TA       Wood
##   BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1 BsmtFinType2
## 1       Gd       TA           No          GLQ        706          Unf
## 2       Gd       TA           Gd          ALQ        978          Unf
## 3       Gd       TA           Mn          GLQ        486          Unf
## 4       TA       Gd           No          ALQ        216          Unf
## 5       Gd       TA           Av          GLQ        655          Unf
## 6       Gd       TA           No          GLQ        732          Unf
##   BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir Electrical
## 1          0       150         856    GasA        Ex          Y      SBrkr
## 2          0       284        1262    GasA        Ex          Y      SBrkr
## 3          0       434         920    GasA        Ex          Y      SBrkr
## 4          0       540         756    GasA        Gd          Y      SBrkr
## 5          0       490        1145    GasA        Ex          Y      SBrkr
## 6          0        64         796    GasA        Ex          Y      SBrkr
##   X1stFlrSF X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath
## 1       856       854            0      1710            1            0        2
## 2      1262         0            0      1262            0            1        2
## 3       920       866            0      1786            1            0        2
## 4       961       756            0      1717            1            0        1
## 5      1145      1053            0      2198            1            0        2
## 6       796       566            0      1362            1            0        1
##   HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Functional
## 1        1            3            1          Gd            8        Typ
## 2        0            3            1          TA            6        Typ
## 3        1            3            1          Gd            6        Typ
## 4        0            3            1          Gd            7        Typ
## 5        1            4            1          Gd            9        Typ
## 6        1            1            1          TA            5        Typ
##   Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars
## 1          0        <NA>     Attchd        2003          RFn          2
## 2          1          TA     Attchd        1976          RFn          2
## 3          1          TA     Attchd        2001          RFn          2
## 4          1          Gd     Detchd        1998          Unf          3
## 5          1          TA     Attchd        2000          RFn          3
## 6          0        <NA>     Attchd        1993          Unf          2
##   GarageArea GarageQual GarageCond PavedDrive WoodDeckSF OpenPorchSF
## 1        548         TA         TA          Y          0          61
## 2        460         TA         TA          Y        298           0
## 3        608         TA         TA          Y          0          42
## 4        642         TA         TA          Y          0          35
## 5        836         TA         TA          Y        192          84
## 6        480         TA         TA          Y         40          30
##   EnclosedPorch X3SsnPorch ScreenPorch PoolArea PoolQC Fence MiscFeature
## 1             0          0           0        0   <NA>  <NA>        <NA>
## 2             0          0           0        0   <NA>  <NA>        <NA>
## 3             0          0           0        0   <NA>  <NA>        <NA>
## 4           272          0           0        0   <NA>  <NA>        <NA>
## 5             0          0           0        0   <NA>  <NA>        <NA>
## 6             0        320           0        0   <NA> MnPrv        Shed
##   MiscVal MoSold YrSold SaleType SaleCondition SalePrice
## 1       0      2   2008       WD        Normal    208500
## 2       0      5   2007       WD        Normal    181500
## 3       0      9   2008       WD        Normal    223500
## 4       0      2   2006       WD       Abnorml    140000
## 5       0     12   2008       WD        Normal    250000
## 6     700     10   2009       WD        Normal    143000
```

```
#Response variable: SalePrice
#Predictors: LotFrontage, WoodDeckSF, GarageArea, GrLivArea

#Summary statistics of selected variables.
#SalePrice: Target variable representing the sale price of the house.

#LotFrontage: Linear feet of street connected to property.
#Properties with higher frontage might be more valuable.

#WoodDeckSF: Square footage of wood deck area, potentially affecting
#property value due to added amenity.

#GarageArea: Square footage of the garage, which adds functional value
#and often affects sale price.

#GrLivArea: Above-ground living area, typically a strong predictor of
# property price.
```

```
# Use psych package to perform descriptive statistics and display variable distribution
# Transformation may be necessary for some variables to meet normality assumptions
library(psych)
describe(data[c("SalePrice", "LotFrontage", "WoodDeckSF", "GarageArea", "GrLivArea")])
```

```
##             vars    n      mean       sd median   trimmed      mad   min    max
## SalePrice      1 1460 180921.20 79442.50 163000 170783.29 56338.80 34900 755000
## LotFrontage    2 1201     70.05    24.28     69     68.94    16.31    21    313
## WoodDeckSF     3 1460     94.24   125.34      0     71.76     0.00     0    857
## GarageArea     4 1460    472.98   213.80    480    469.81   177.91     0   1418
## GrLivArea      5 1460   1515.46   525.48   1464   1467.67   483.33   334   5642
##              range skew kurtosis      se
## SalePrice   720100 1.88     6.50 2079.11
## LotFrontage    292 2.16    17.34    0.70
## WoodDeckSF     857 1.54     2.97    3.28
## GarageArea    1418 0.18     0.90    5.60
## GrLivArea     5308 1.36     4.86   13.75
```

#Histogram and Fitted Distribution

```
library(ggplot2)
```

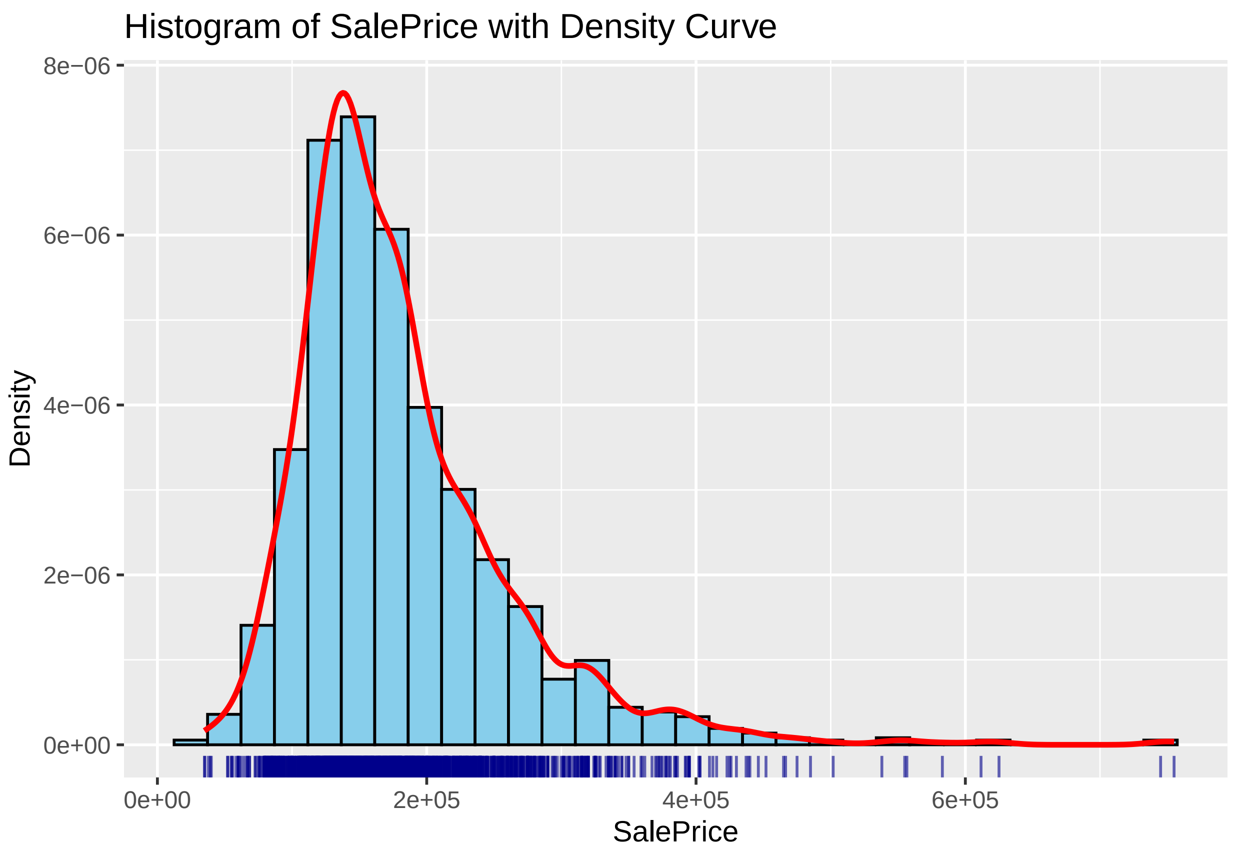
```
##
##    'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```
#Histogram of SalePrice, including density line
# Plot histogram of SalePrice with density curve to observe distribution shape
# SalePrice shows skewness, indicating the need for transformation to achieve normality
ggplot(data, aes(x = SalePrice)) +
geom_histogram(aes(y = ..density..), bins = 30, fill = "skyblue", color = "black") +
geom_density(color = "red", size = 1) + geom_rug(sides = "b", color = "darkblue", alpha = 0.6)+
labs(title = "Histogram of SalePrice with Density Curve",
x = "SalePrice", y = "Density")
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



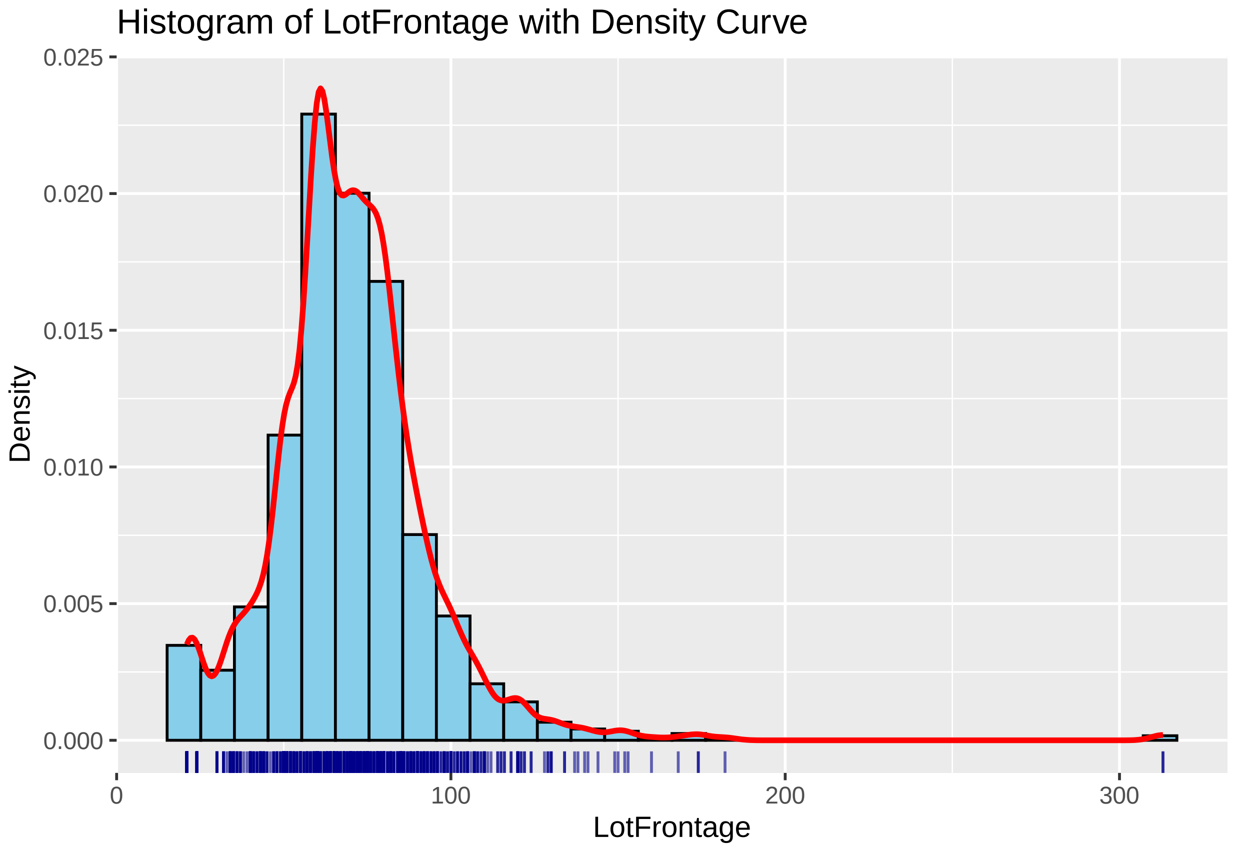
Comment:

This histogram displays the distribution of the SalePrice variable with an overlayed density curve. Observing the right-skewed shape, it suggests that SalePrice may need a log or Box-Cox transformation to meet normality assumptions. The red density curve shows that higher sale prices are less frequent. The right-skewness could impact model accuracy, as normality is a key assumption. Higher-priced homes are outliers, possibly representing luxury properties.

```
#Histogram of LotFrontage, including density line
# Histograms and density curves of other variables to observe distribution shape
# Transformation might be needed if skewness is evident
ggplot(data, aes(x = LotFrontage)) +
geom_histogram(aes(y = ..density..), bins = 30, fill = "skyblue", color = "black") +
geom_density(color = "red", size = 1) + geom_rug(sides = "b", color = "darkblue", alpha = 0.6)+
labs(title = "Histogram of LotFrontage with Density Curve",
x = "LotFrontage", y = "Density")
```

```
## Warning: Removed 259 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

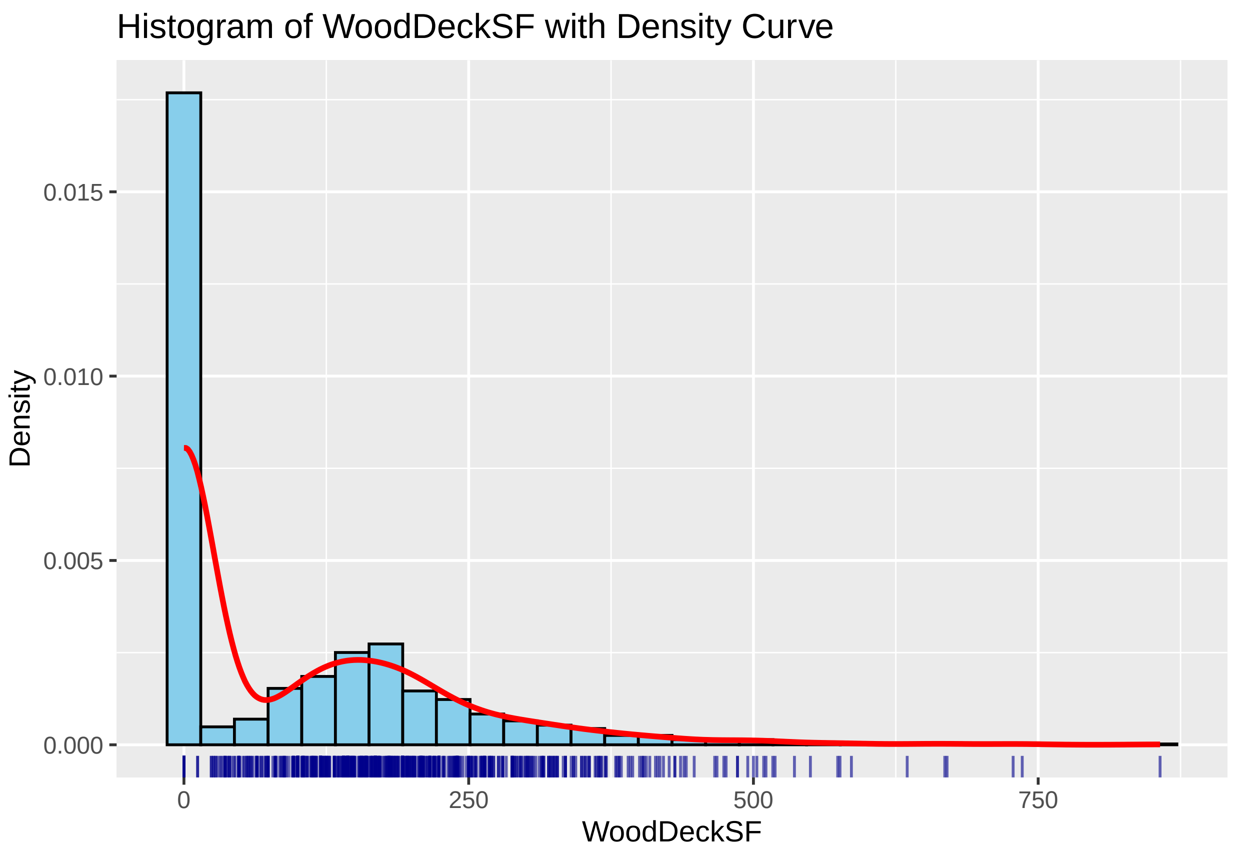
```
## Warning: Removed 259 rows containing non-finite outside the scale range
## (`stat_density()`).
```



Comment:

Most values clustered around the lower range. The right-skewed suggests outliers exists or larger lots that are less common. Log transformation might help in reducing skewness. Skewness may violate regression assumptions, suggesting the need for transformation. Larger frontages could indicate premium lots, impacting housing prices.

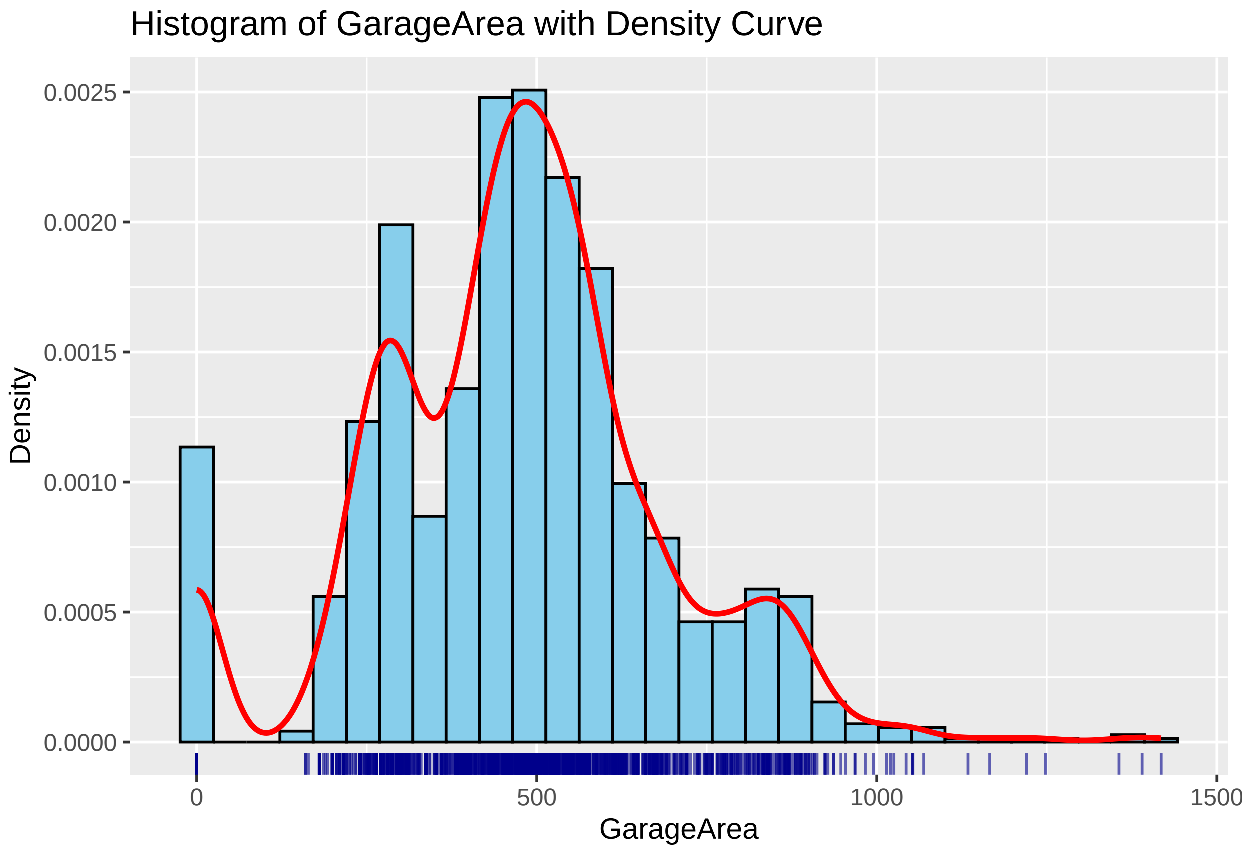
```
#Histogram of WoodDeckSF, including density line
# Similar plots for other variables to determine if transformation is necessary
# Consider transformation if skewness is substantial
ggplot(data, aes(x = WoodDeckSF)) +
geom_histogram(aes(y = ..density..), bins = 30, fill = "skyblue", color = "black") +
geom_density(color = "red", size = 1) + geom_rug(sides = "b", color = "darkblue", alpha = 0.6)+
labs(title = "Histogram of WoodDeckSF with Density Curve",
x = "WoodDeckSF", y = "Density")
```



Comment:

High frequency of properties having zero deck area exist. The density curve indicates a right-skewed distribution, with many properties lacking a wood deck. A log transformation might normalize the data by reducing the concentration to zero. High skewness may affect regression assumptions. Properties with decks may have higher values, representing added amenities.

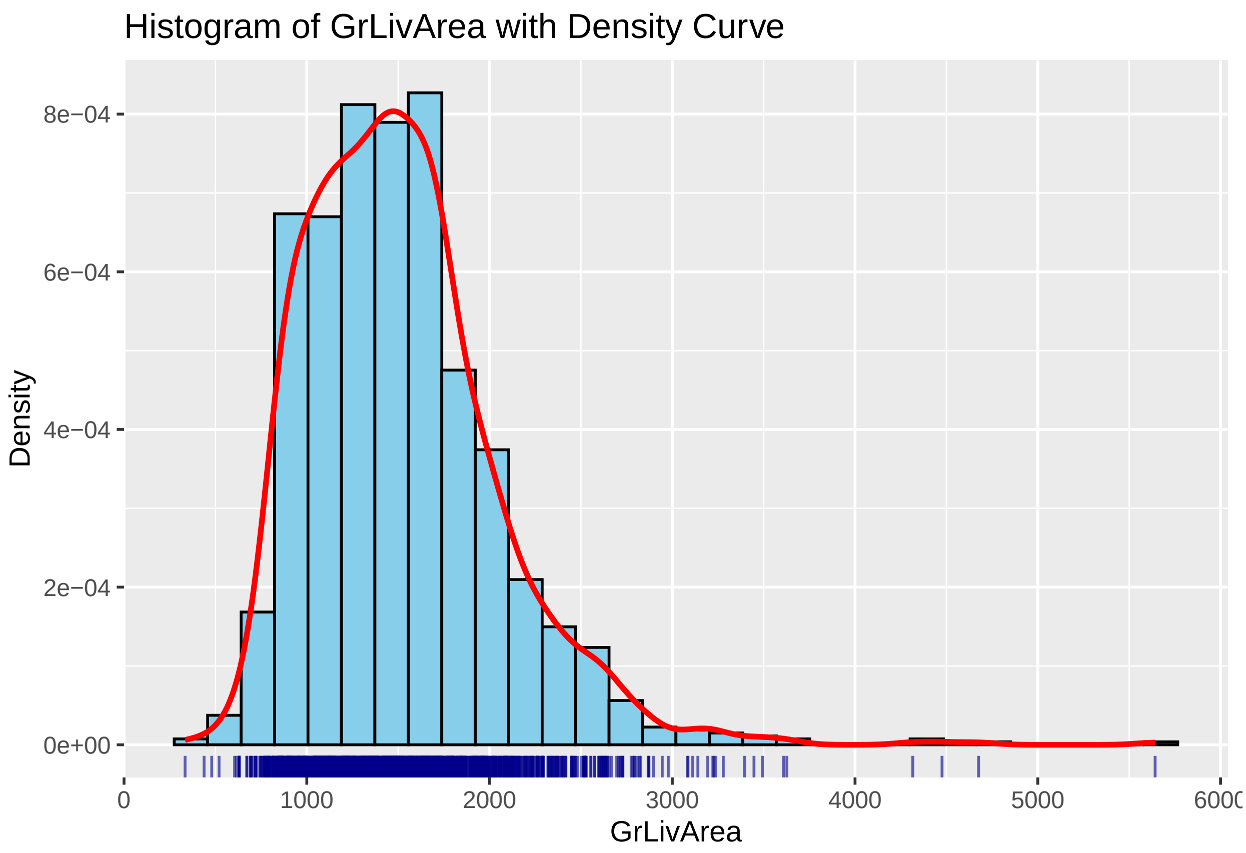
```
#Histogram of GarageArea, including density line
ggplot(data, aes(x = GarageArea)) +
geom_histogram(aes(y = ..density..), bins = 30, fill = "skyblue", color = "black") +
geom_density(color = "red", size = 1) + geom_rug(sides = "b", color = "darkblue", alpha = 0.6)+
labs(title = "Histogram of GarageArea with Density Curve",
x = "GarageArea", y = "Density")
```



Comment:

This histogram displays GarageArea distribution, with a noticeable right skew and several properties without a garage. The density curve indicates skewness, which could affect statistical assumptions in modeling. Skewed distribution may impact model accuracy if untransformed. Larger garage areas could help increase home values, representing an important amenity for buyers.

```r
#Histogram of GrLivArea, including density line
ggplot(data, aes(x = GrLivArea)) +
geom_histogram(aes(y = ..density..), bins = 30, fill = "skyblue", color = "black") +
geom_density(color = "red", size = 1) + geom_rug(sides = "b", color = "darkblue", alpha = 0.6)+
labs(title = "Histogram of GrLivArea with Density Curve",
x = "GrLivArea", y = "Density")
```



Comment:

A density curve shows a mild right skew, implying the potential need for transformation for normality. A log transformation may improve distribution symmetry. A more symmetrical distribution would enhance model accuracy.

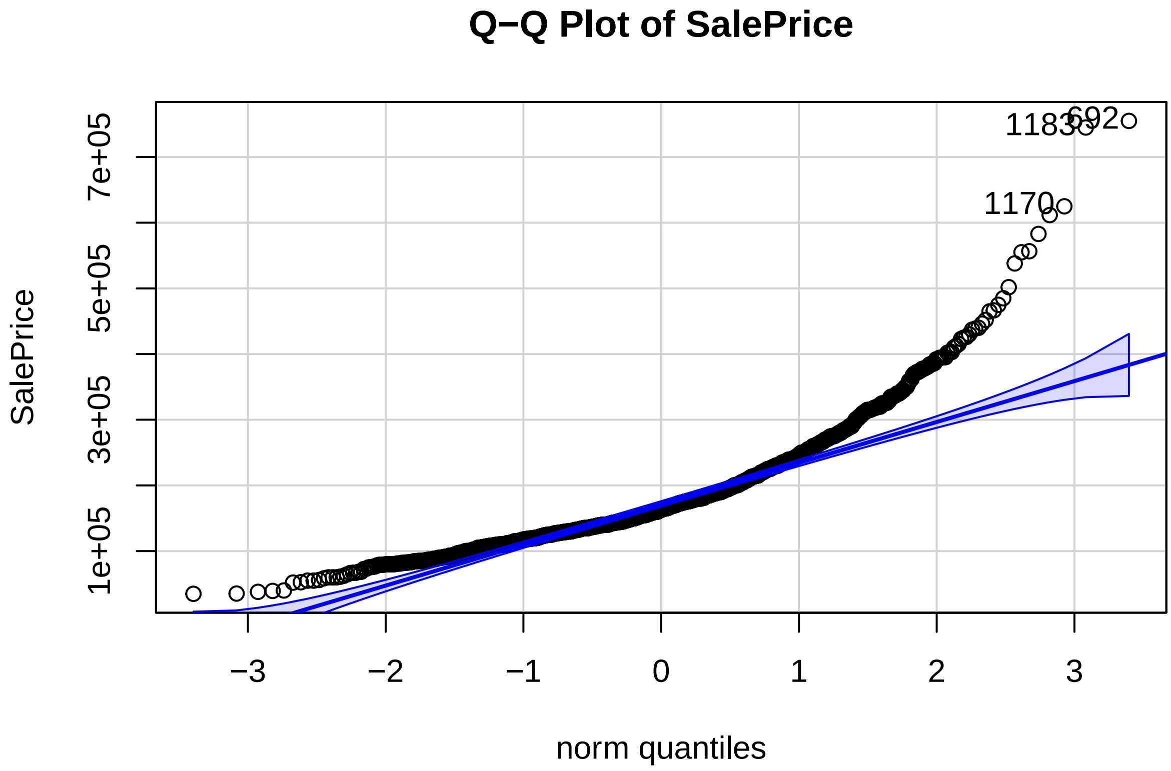
```
# Q-Q plots for each variable
# Q-Q plots to check normality; if points deviate significantly from the normal line,
# transformation is recommended.

# List of variables to create Q-Q plots for
library(car)
```

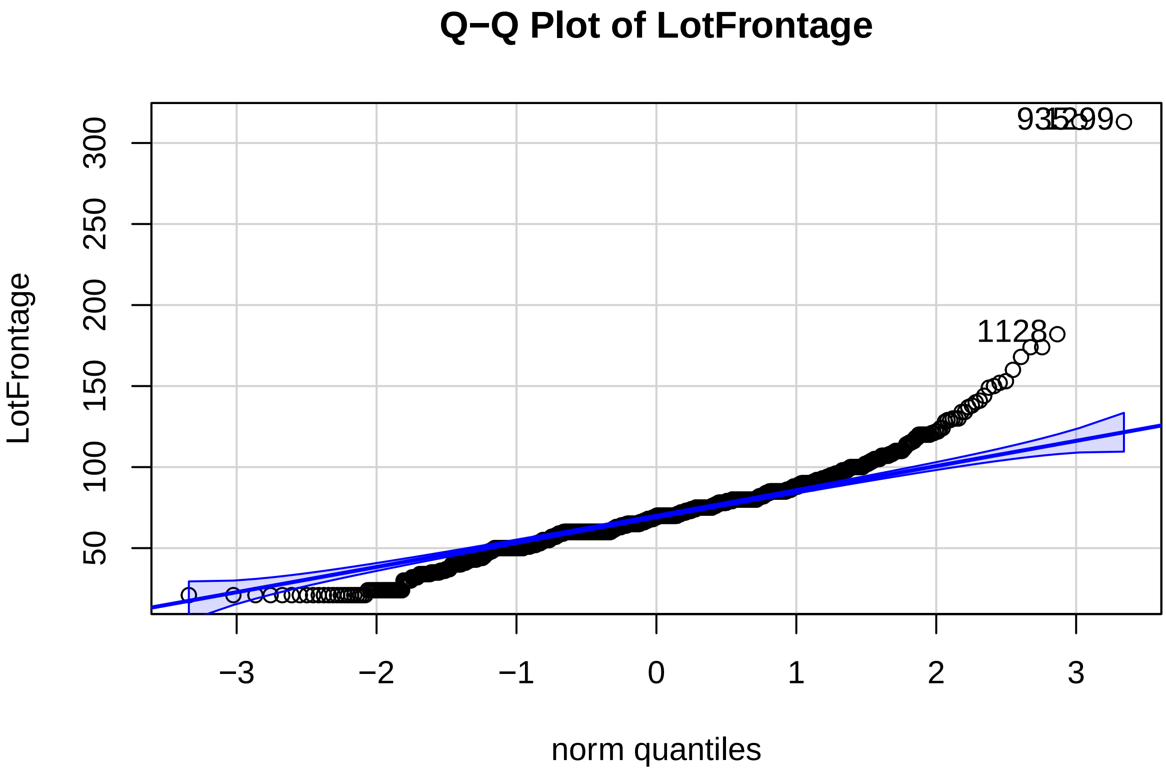
```
##      carData
```

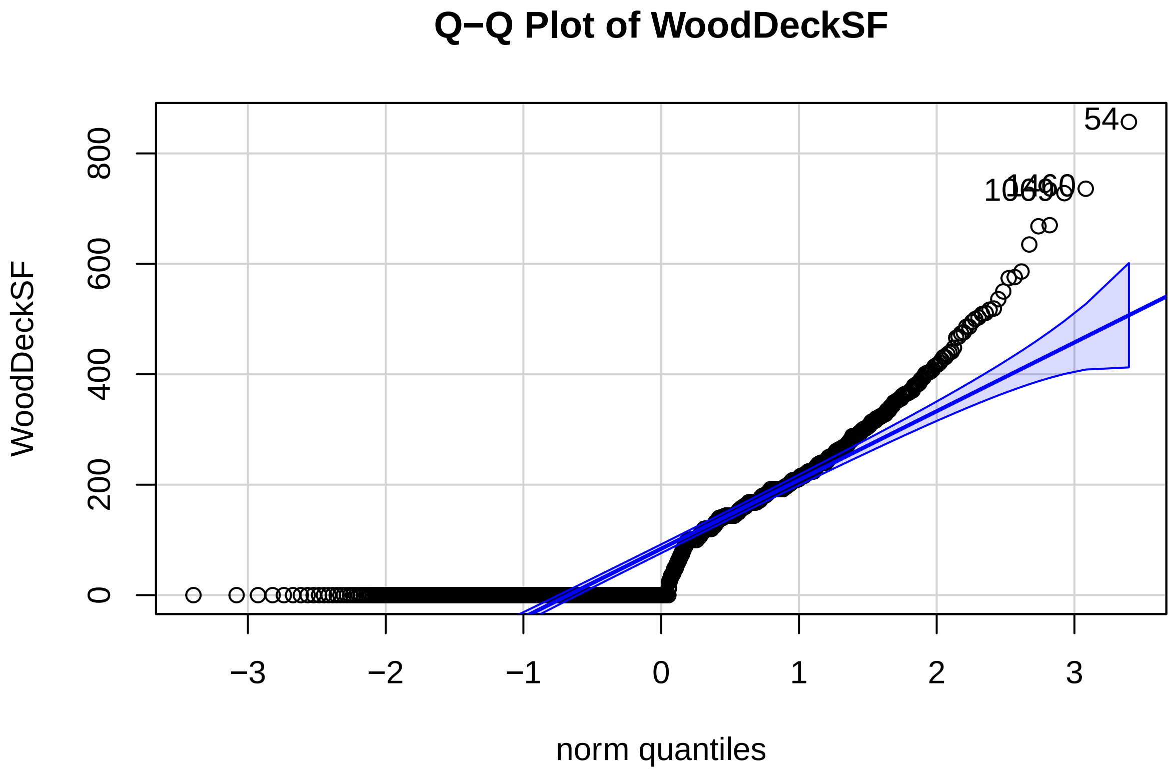
```
##
##    'car'
```

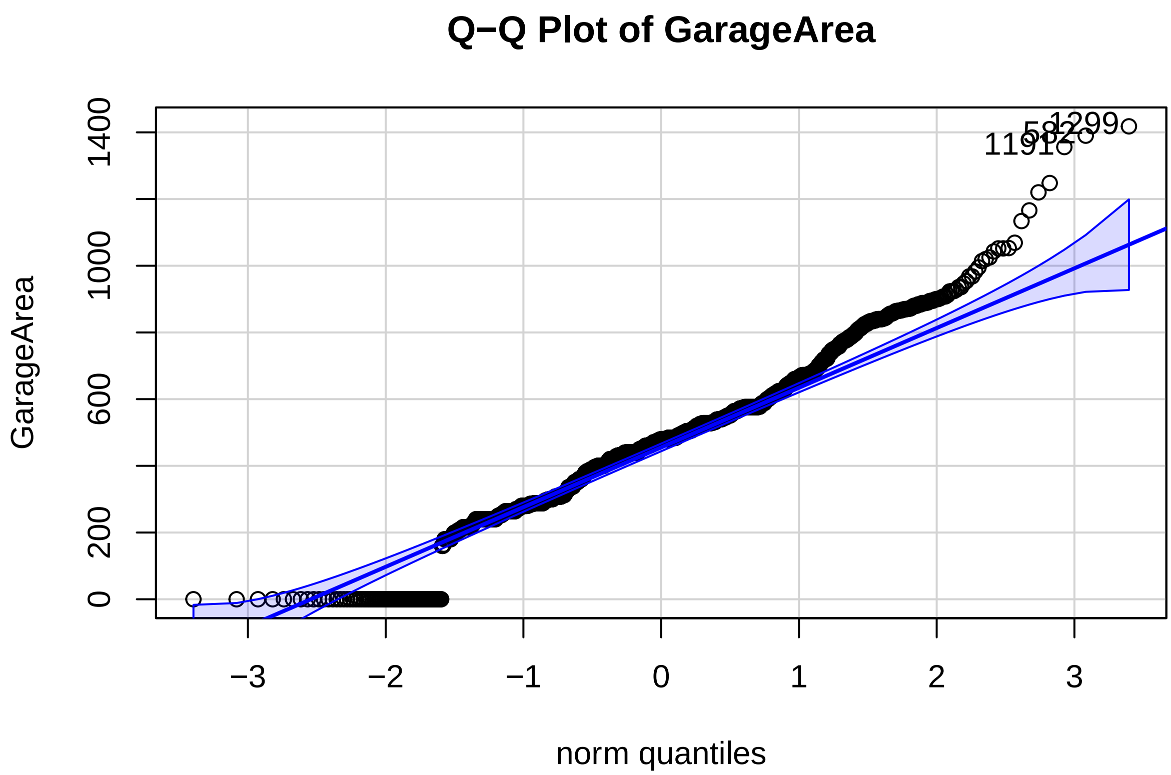
```
## The following object is masked from 'package:psych':
##
##     logit
```

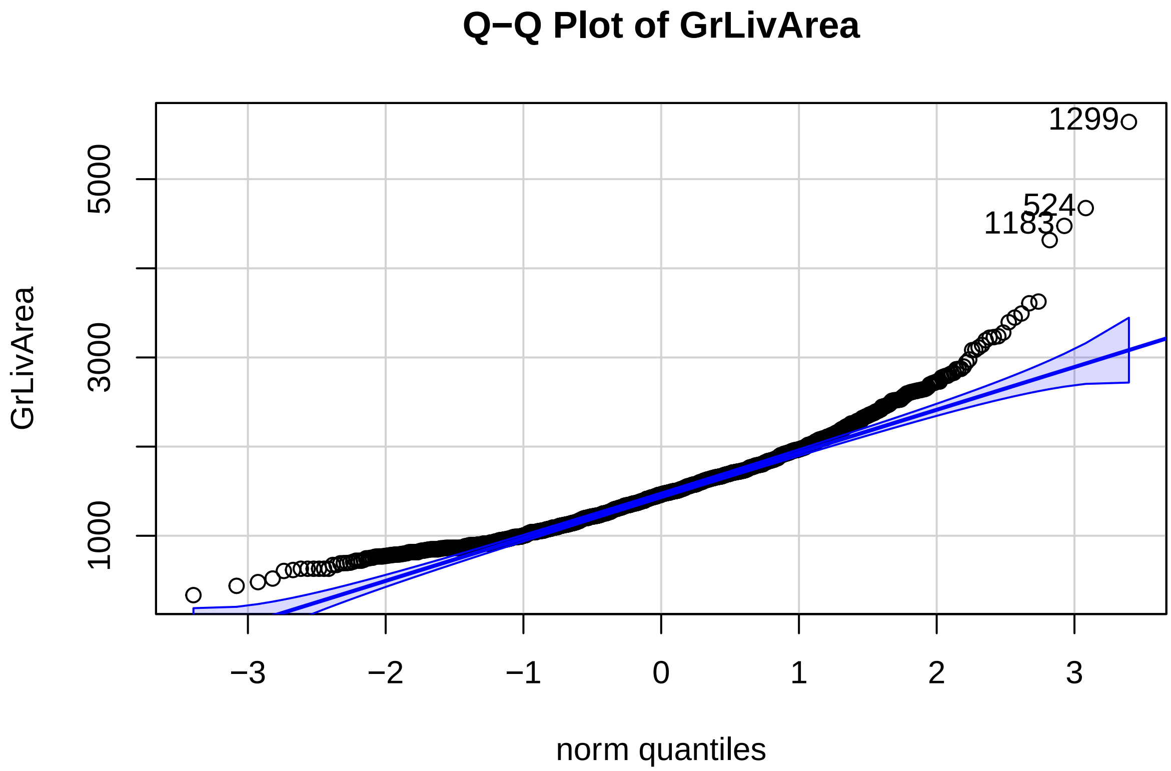
```
variables <- c("SalePrice", "LotFrontage", "WoodDeckSF", "GarageArea", "GrLivArea")
# Loop to generate Q-Q plots for each variable
for (var in variables) {
# Generate Q-Q plot with car::qqPlot
qqPlot(data[[var]],
main = paste("Q-Q Plot of", var),
ylab = var,
id = list(n = 3)) # Label the top 3 extreme points
}
```




Q−Q Plot of LotFrontage
LotFrontage
300
250
200
150
100
50
935
1299
1128
−3
−2
−1
0
1
2
3
norm quantiles


Q−Q Plot of WoodDeckSF
WoodDeckSF
0
200
400
600
800
−3
−2
−1
0
1
2
3
norm quantiles
54
1460
1069


Q−Q Plot of GarageArea
GarageArea
0
200
600
1000
1400
−3
−2
−1
0
1
2
3
norm quantiles
1299
582
1191



Comments:

1. The Q-Q plot for SalePrice shows an obvious deviation from the normal line, particularly at the upper tail. This implies that SalePrice does not follow the normal distribution, showing right skewness and outliers in the higher range. The Q-Q plot reveals deviations from normality, especially at the upper quantiles. The tail on the right side curves upwards, indicating positive skewness. This skew suggests that SalePrice may require transformation for better linear modeling.

2. The Q-Q plot for LotFrontage shows deviation from the normal line, especially in the upper tail. This implies that LotFrontage is right-skewed and contains extreme values (outliers). The plot shows right-tail deviations from the normal line, indicating positive skewness. # Extreme values in the upper quantiles may affect linearity and normality assumptions. Transforming LotFrontage could help normalize the distribution for analysis.

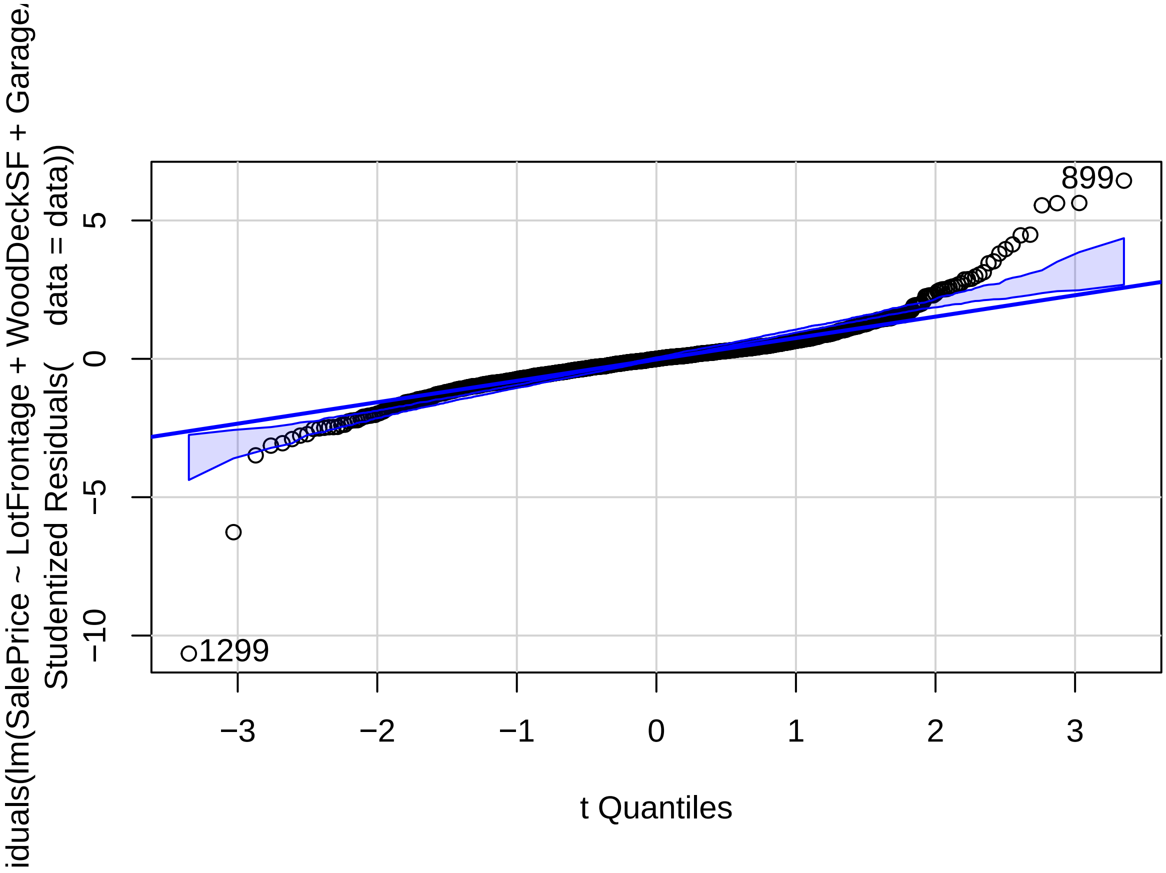
3. The Q-Q plot for WoodDeckSF shows a significant deviation from the normal line, particularly at the lower and upper ends. This implies a high degree of skewness and the presence of extreme values. Deviations in the upper quantiles indicate a right-skewed distribution. This plot suggests that WoodDeckSF has a few large values not following normal distribution. A transformation may reduce skewness for model building.

4. The Q-Q plot for GarageArea shows deviations from the normal line, especially in the upper tail, indicating skewness and extreme values. Many properties have nearly no value for GarageArea, and this contributes to the skewness.The distribution shows slight right skew with outliers in the upper quantiles. Points above the line on the right side indicate larger garages than typical. Addressing this skew could improve the data's fit for linear models.

5. The Q-Q plot for GrLivArea shows mild deviations from the normal line, particularly in the upper tail where some extreme values are present. This indicates slight right skewness and might affect

statistical analysis assuming normality. The plot displays right-tail deviation, indicating a positive skew in the data. Larger values deviate from the normal distribution, suggesting a transformation may be beneficial. This skew could impact regression models if not addressed.

```
qqPlot(lm(SalePrice ~ LotFrontage+WoodDeckSF+GarageArea+GrLivArea, data=data),
envelope=.99)
```

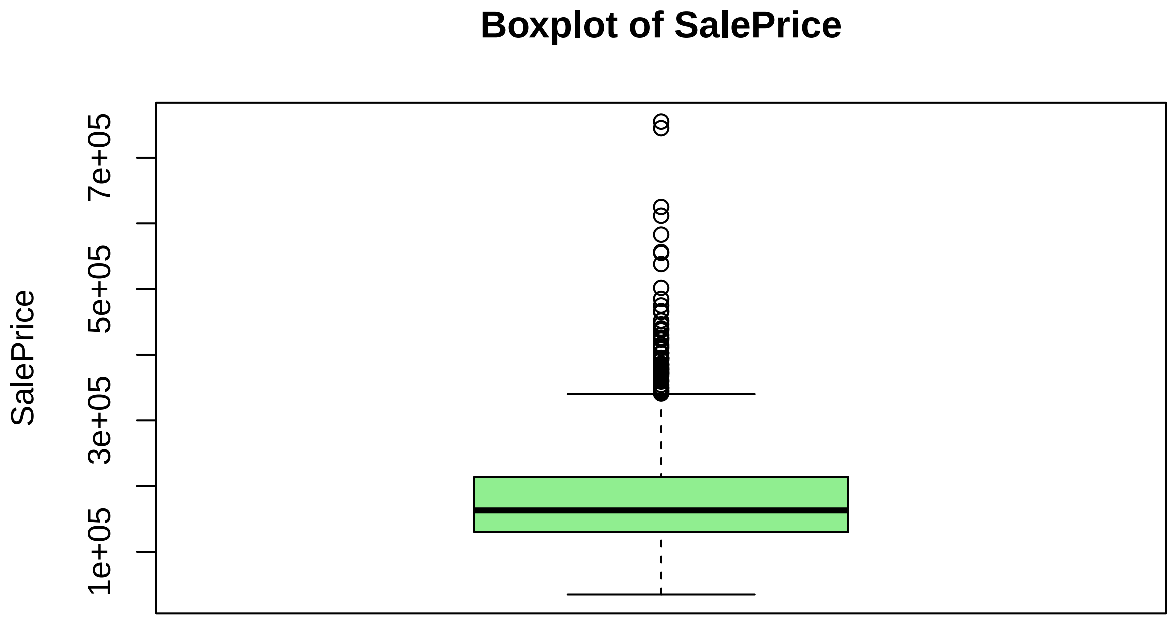


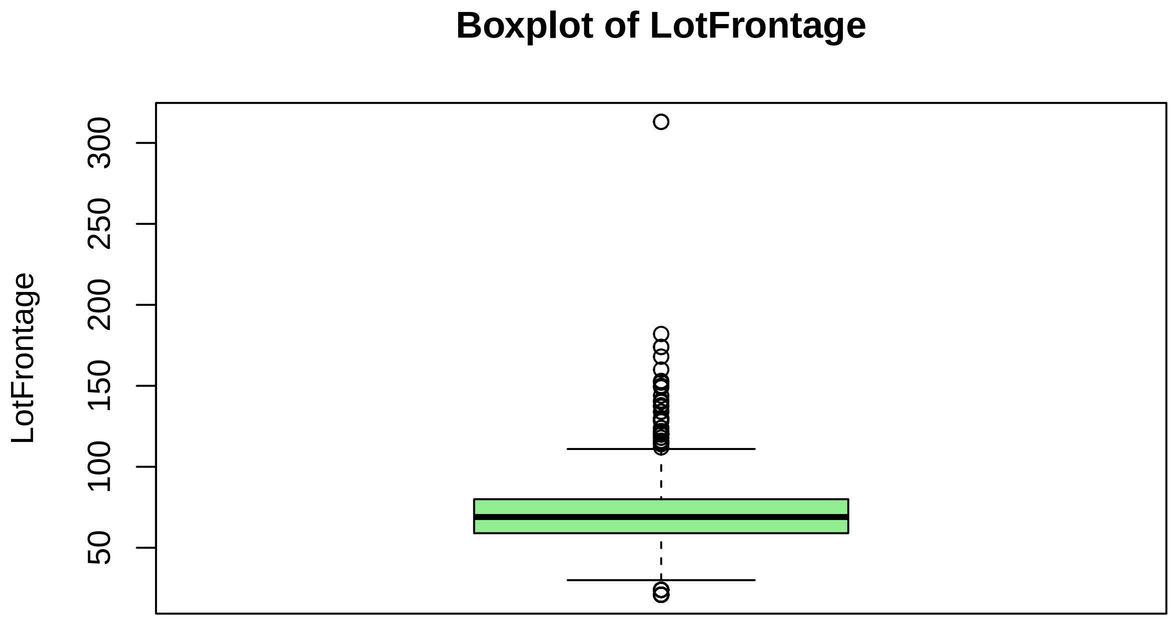
```
## [1]  899 1299
```

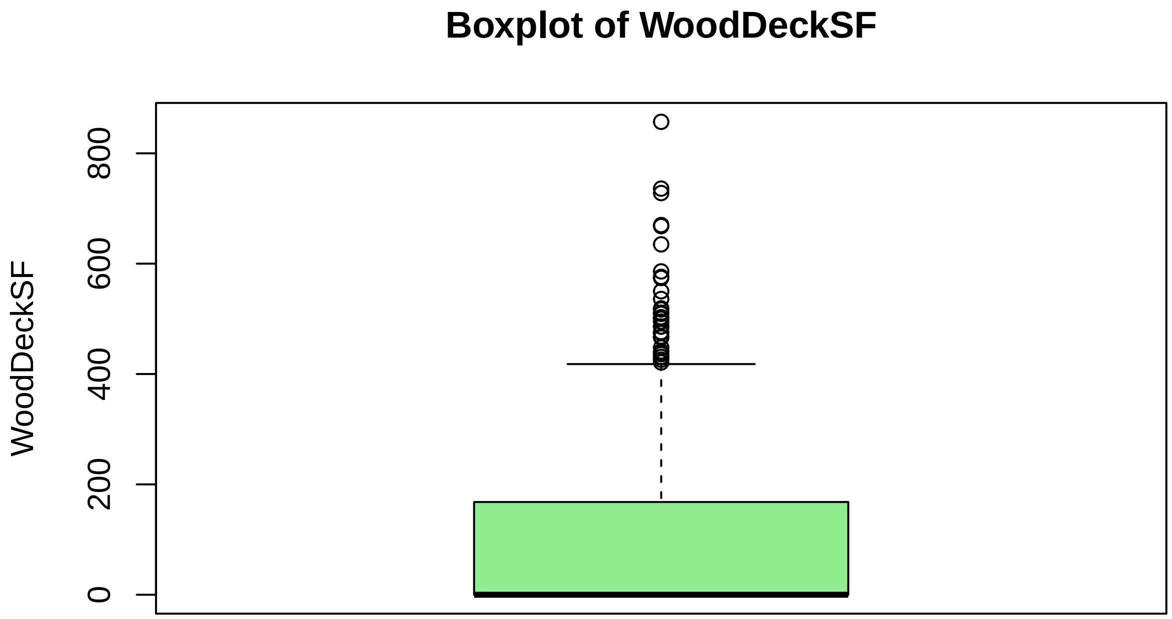
Comment:

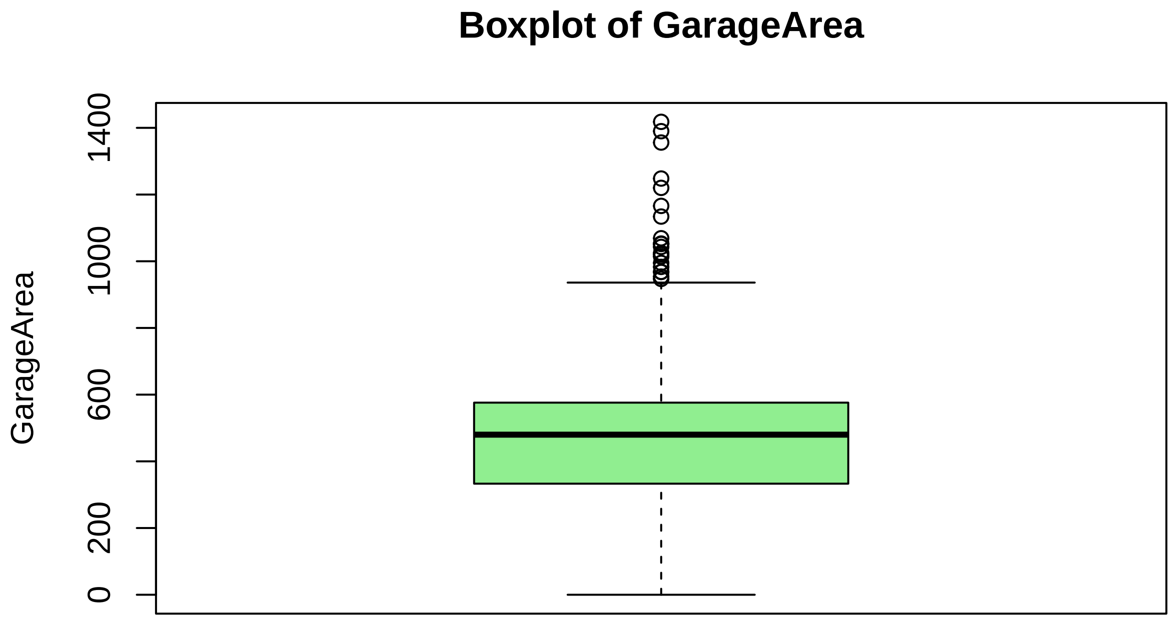
This Q-Q plot shows the distribution of residuals for the model (SalePrice ~ LotFrontage + WoodDeckSF + GarageArea + GrLivArea). The plot compares the residual quantiles to a theoretical normal distribution. Ideally, if residuals are normally distributed, they would align closely along the 45-degree reference line. Here, deviations from the line, particularly in the upper tail, indicate that some residuals deviate from normality, suggesting the presence of outliers. This may affect model assumptions and inference accuracy. The plot for residuals aligns closely with the normal line, indicating a near-normal distribution. This suggests that the chosen model adequately fits the data with minimal skew in residuals. Minimal transformation may be needed if this pattern holds across other diagnostics.

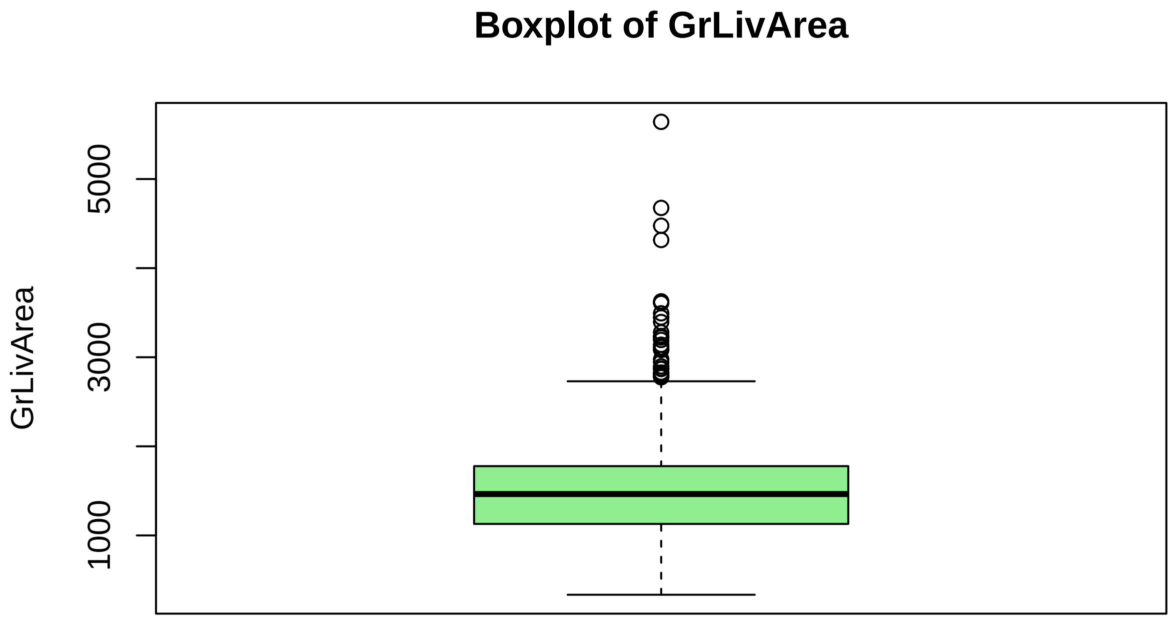
```
#Boxplots for all variables
for (var in variables) {
boxplot(data[[var]], main = paste("Boxplot of", var),
ylab = var, col = "lightgreen")
}
```


Boxplot of SalePrice
SalePrice
1e+05
3e+05
5e+05
7e+05


Boxplot of LotFrontage
LotFrontage
50
100
150
200
250
300


Boxplot of WoodDeckSF
WoodDeckSF
0
200
400
600
800

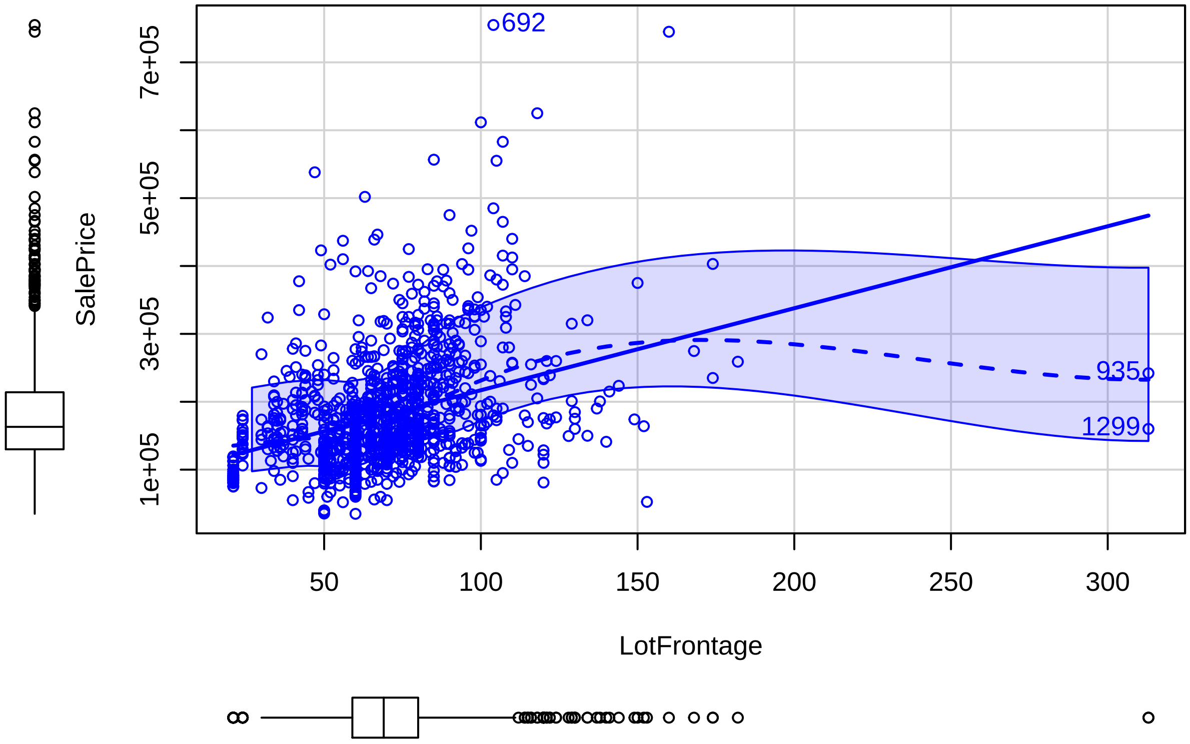

Boxplot of GarageArea
GarageArea
0
200
600
1000
1400



Comments:

1. The boxplot of SalePrice shows that most values are concentrated within the interquartile range with a few outliers above the upper whisker that represent properties with unusually high sale prices, which causes the distribution of SalePrice to be right-skewed. The median is towards the lower end of the interquartile range, suggesting a right-skewed distribution.
2. The boxplot of LotFrontage shows that most values are concentrated within the interquartile range, with several outliers above the upper whisker that represent properties with unusually large lot frontages, imply a right-skewed distribution for LotFrontage might because of specific geographic or zoning factors. The distribution's central tendency is around the median, with a relatively compact IQR, demonstrating moderate variability.
3. The boxplot of WoodDeckSF shows that most values are low, with a large number of properties having zero or very small deck areas that represent properties with larger wood deck areas, causing the distribution to be right-skewed, reveals a positively skewed distribution for Wood Deck square footage. Data concentrated near the lower end, with a few outliers representing properties with much larger decks. The IQR is relatively narrow, highlighting limited variability in deck sizes for most properties.
4. The boxplot of GarageArea shows most values concentrated within the interquartile range, with several outliers above the upper whisker which causes the distribution of GarageArea to be slightly right-skewed. The distribution of Garage Area is fairly concentrated within the IQR, with outliers above the upper whisker. The median is situated slightly below the center of the IQR, suggesting a slight skew.
5. The boxplot shows that most GrLivArea values are concentrated within the interquartile range with a few high outliers above the upper whisker. This indicates that while most properties have similar above-ground living areas, there are a few with significantly larger spaces, which stand out as outliers in the data. Median centrally located within the IQR. This suggests a more symmetrical distribution within the core range, although extreme values exist.

```
#scatterplots for all variables
scatterplot(SalePrice ~ LotFrontage, data = data, lwd = 3, id = list(n = 3))
```

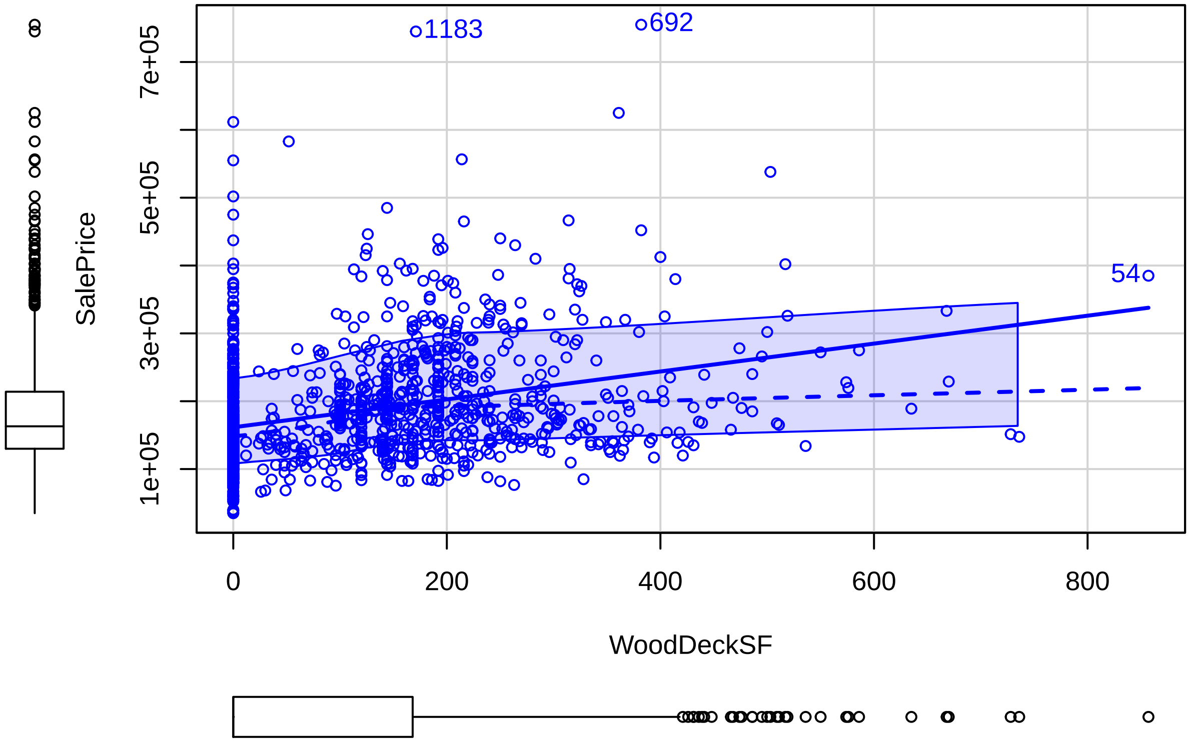


```
##  692  935 1299
##  574  774 1070
```

Comment:

The scatterplot shows a weak positive relationship between LotFrontage and SalePrice, with a large spread and several outliers, particularly at higher values of LotFrontage. The shaded area represents confidence bands, and the dashed line suggests a non-linear trend. Outliers are visible in both variables, especially in LotFrontage. A transformation on LotFrontage clarified its relationship with SalePrice due to the widespread.

```
scatterplot(SalePrice ~ WoodDeckSF, data = data, lwd = 3, id = list(n = 3))
```

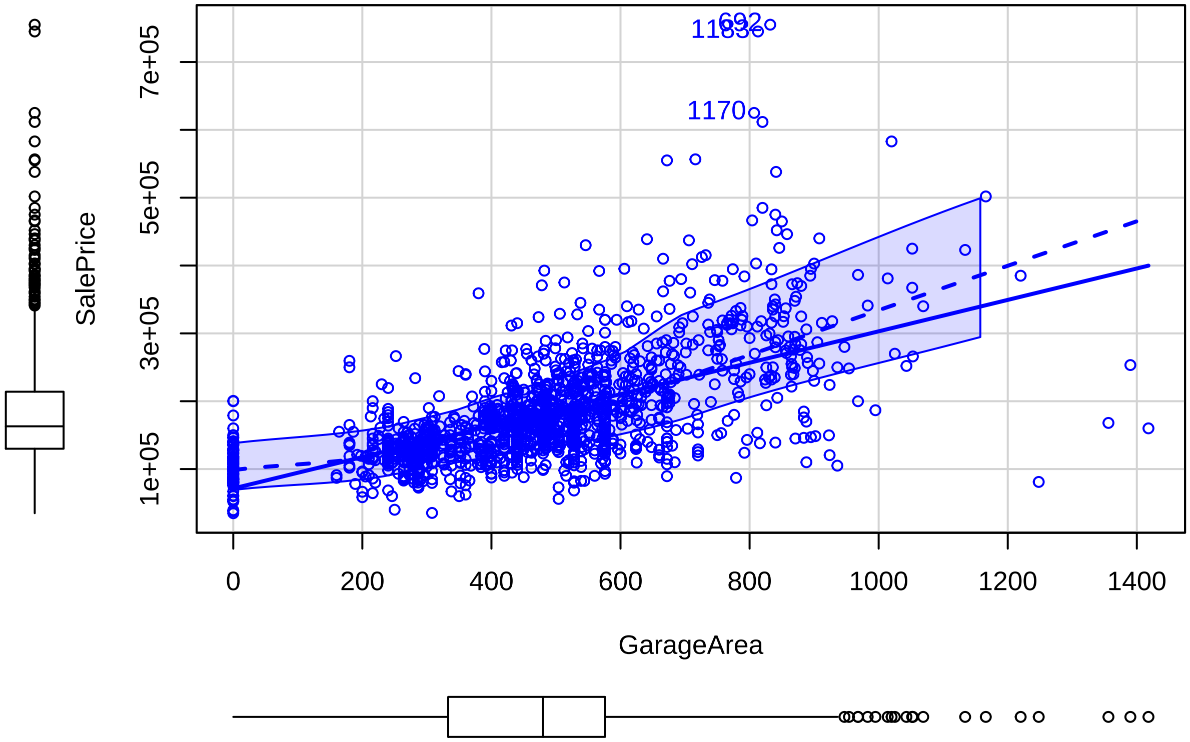


```
## [1]   54  692 1183
```

Comment:

The scatterplot indicates a weak positive association between WoodDeckSF and SalePrice, with a large spread of points and a few outliers. The dashed line shows a possible non-linear pattern. Points are clustered around lower values of WoodDeckSF. Transforming WoodDeckSF could possibly reduce skewness and improve interpretation of the relationship.

```
scatterplot(SalePrice ~ GarageArea, data = data, lwd = 3, id = list(n = 3))
```

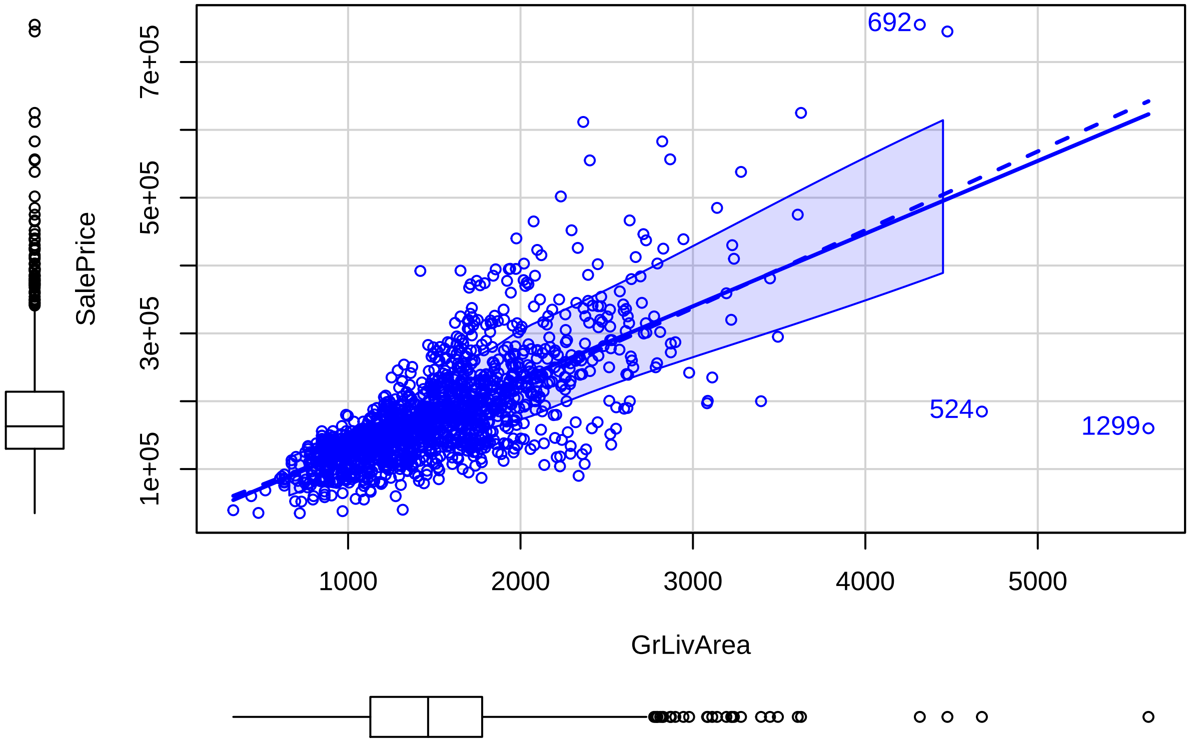


```
## [1]  692 1170 1183
```

Comment:

The scatterplot shows a moderate positive relationship between GarageArea and SalePrice, with SalePrice generally increasing as GarageArea increases. A few high outliers are visible, mostly for larger GarageArea values. The shaded area represents confidence intervals, and the dashed line suggests a potential non-linear trend. Higher values of GarageArea correspond to higher SalePrice.

```
scatterplot(SalePrice ~ GrLivArea, data = data, lwd = 3, id = list(n = 3))
```

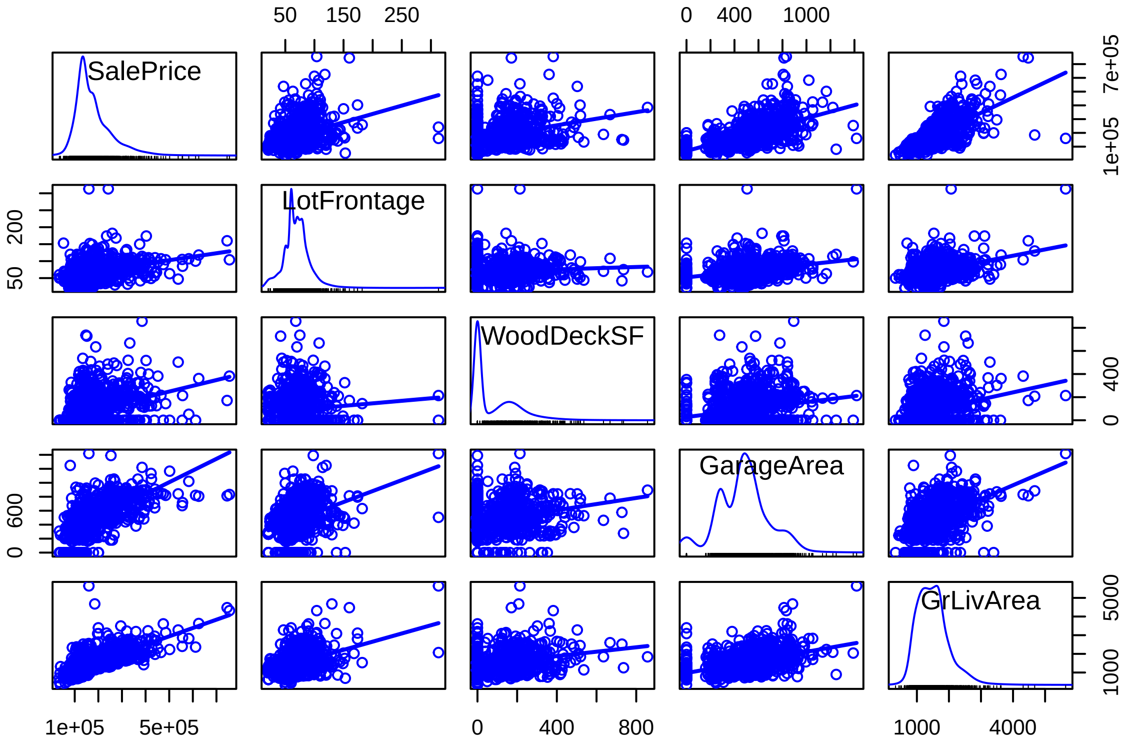


```
## [1]  524  692 1299
```

Comment:

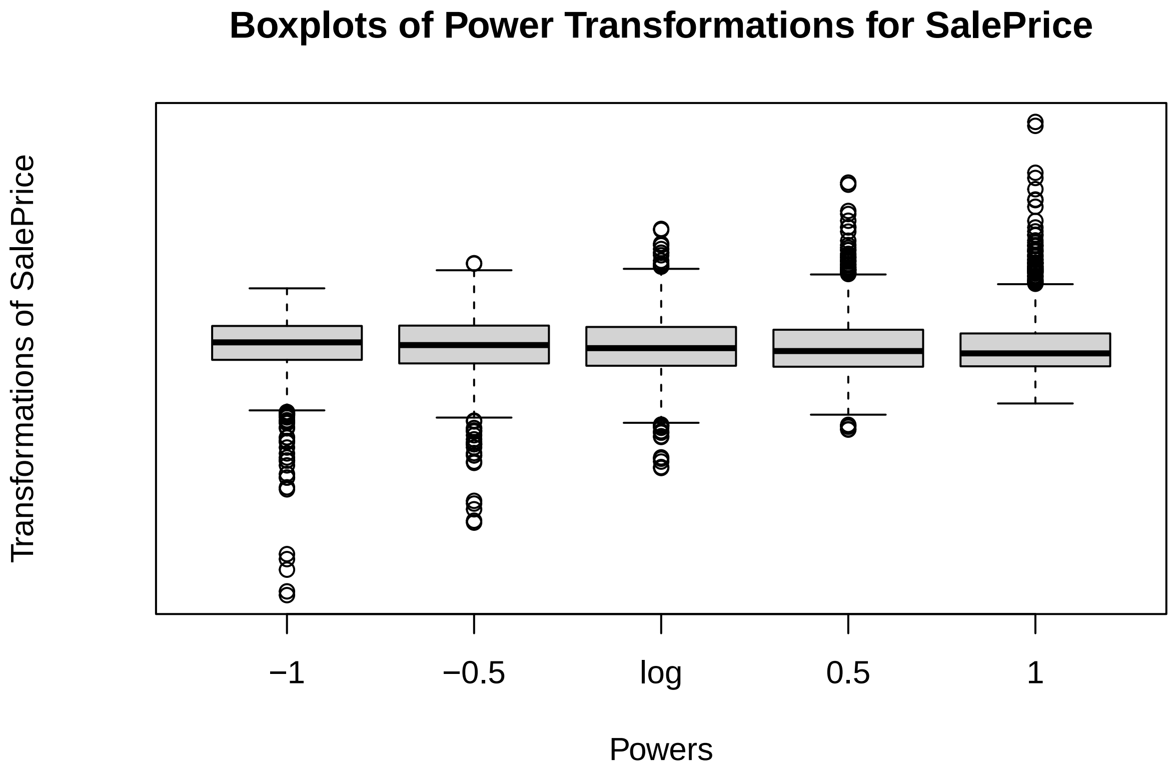
The scatterplot shows a strong positive relationship between GrLivArea and SalePrice, with SalePrice increasing as GrLivArea increases. A few high outliers are present, particularly at larger GrLivArea values. The shaded area represents confidence intervals, and the dashed line indicates a linear trend. There is a strong positive linear trend between GrLivArea and SalePrice. Outliers in GrLivArea are apparent, possibly impact the model fit.

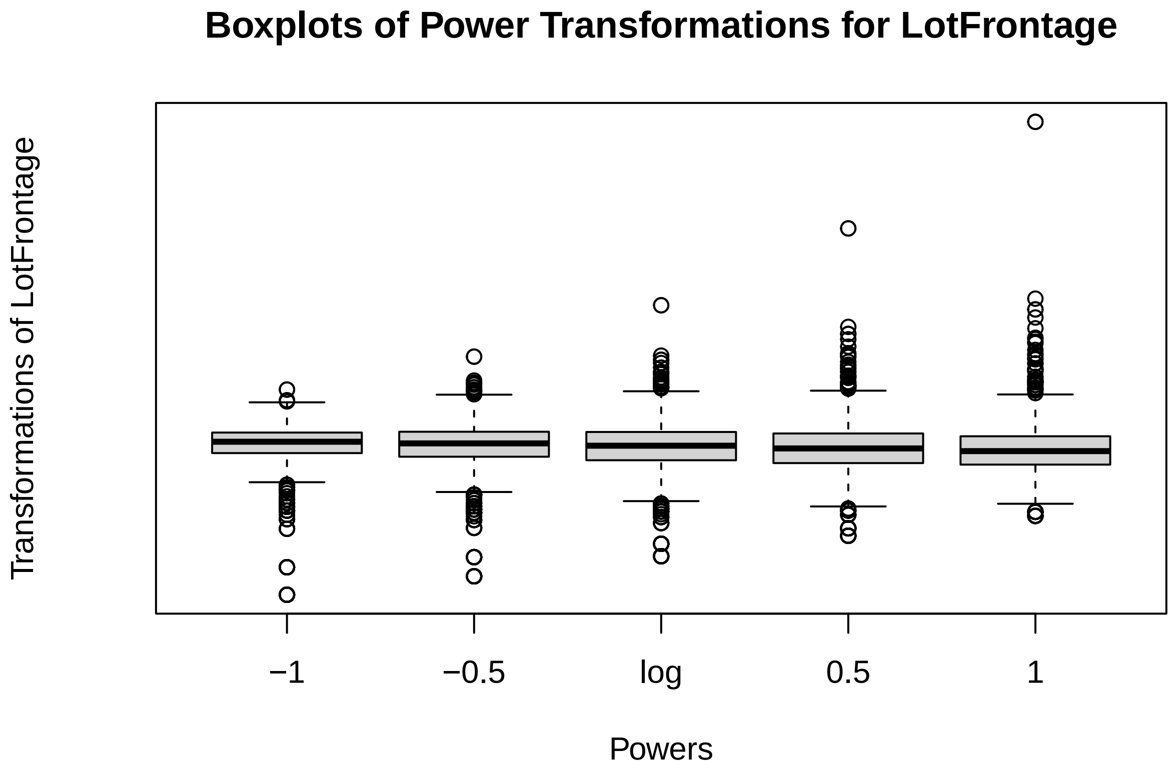
```r
# Scatterplot matrix with specified variables and options
scatterplotMatrix(~ SalePrice + LotFrontage + WoodDeckSF + GarageArea + GrLivArea,
data = data,
smooth = FALSE,
ellipse = list(levels = 0.5))
```



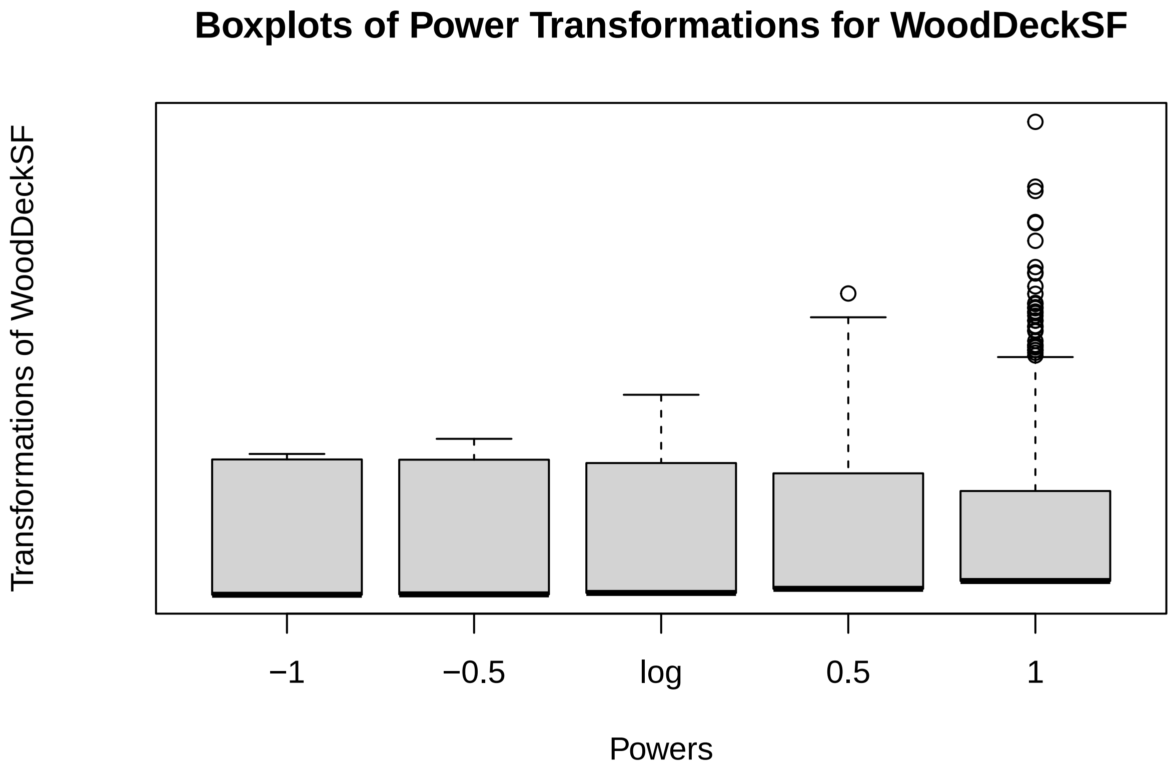
Comment: The scatterplot matrix shows pairwise relationships between key variables. SalePrice has a positive association with most variables, particularly GrLivArea and GarageArea. Density plots on the diagonal indicate right-skewness in SalePrice, LotFrontage, and WoodDeckSF. Ellipses show the spread and concentration of points for each variable pair. Displays pairwise scatterplots between SalePrice, LotFrontage, WoodDeckSF, GarageArea, and GrLivArea. Relationships vary, with stronger trends observed in GrLivArea and GarageArea vs. SalePrice. Consider transformations for variables with weak or skewed relationships to improve linearity.

```r
#linearity test
for (var in variables) {
formula <- as.formula(paste("~", var))
symbox(formula, data = data, main = paste("Boxplots of Power Transformations for", var))
}
```

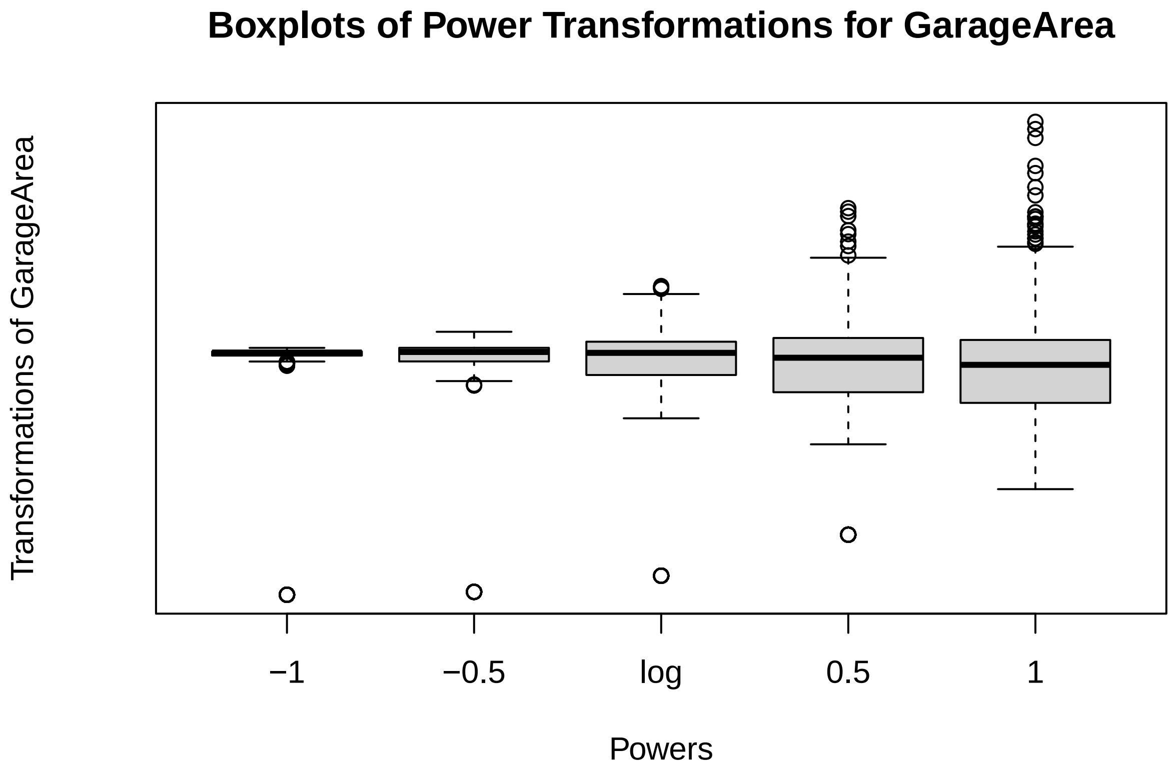

Boxplots of Power Transformations for SalePrice
Transformations of SalePrice
−1
−0.5
log
0.5
1
Powers

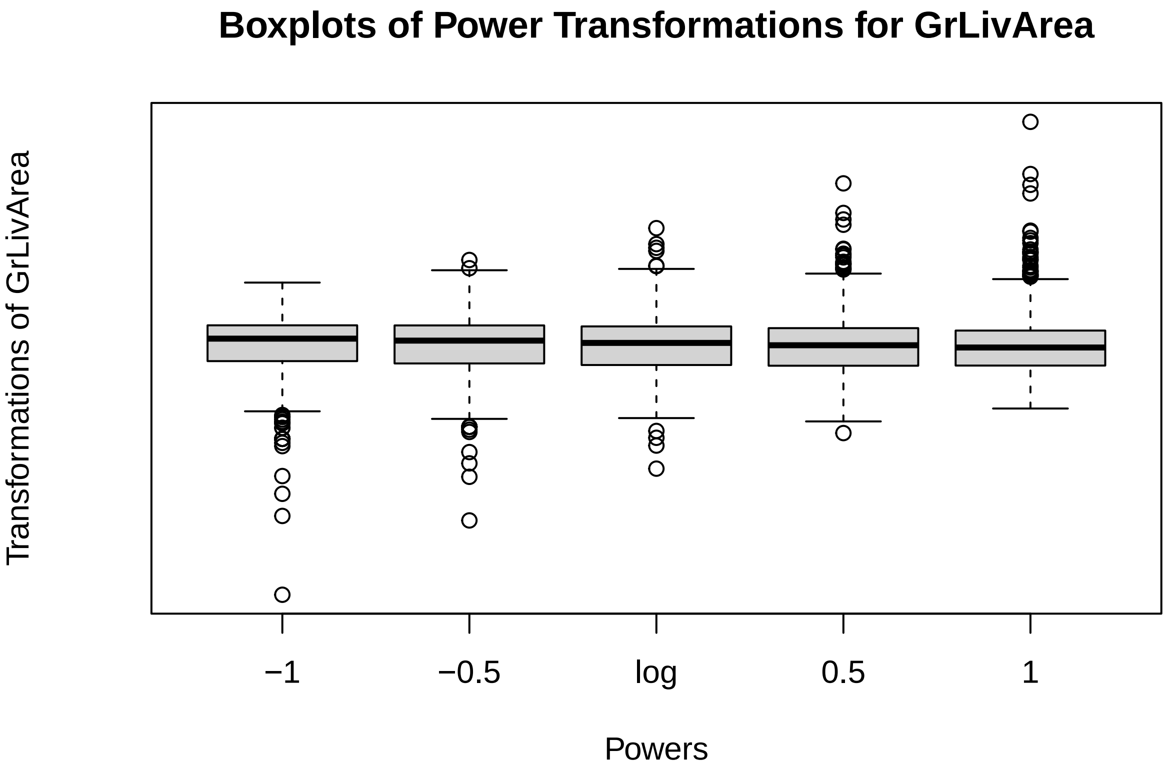


```
## Warning in symbox.default(as.vector(mf[[1]]), ylab = ylab, ...): start set to
## 8.57
```



```
## Warning in symbox.default(as.vector(mf[[1]]), ylab = ylab, ...): start set to
## 14.18
```


Boxplots of Power Transformations for GarageArea
Transformations of GarageArea
−1
−0.5
log
0.5
1
Powers



Comments:

1. This plot shows boxplots of different power transformations for SalePrice, including -1, -0.5, log, 0.5, and 1. Each transformation adjusts the distribution of SalePrice differently. This helps identify the transformation that makes the data more symmetric, potentially improving linearity in regression models.

2. This plot displays boxplots of different power transformations for LotFrontage, such as -1, -0.5, log, 0.5, and 1. These transformations are used to identify a more symmetric distribution, which may help improve the variable's behavior in regression models.

3. This plot displays boxplots of different power transformations for WoodDeckSF, including -1, -0.5, log, 0.5, and 1. The transformations illustrate the distribution changes, allowing us to identify a more symmetric or balanced transformation if needed for regression.

4. This plot displays boxplots of different power transformations for GarageArea, including -1, -0.5, log, 0.5, and 1. The transformations show variations in the distribution symmetry, assisting in identifying a suitable transformation for linear modeling if needed.

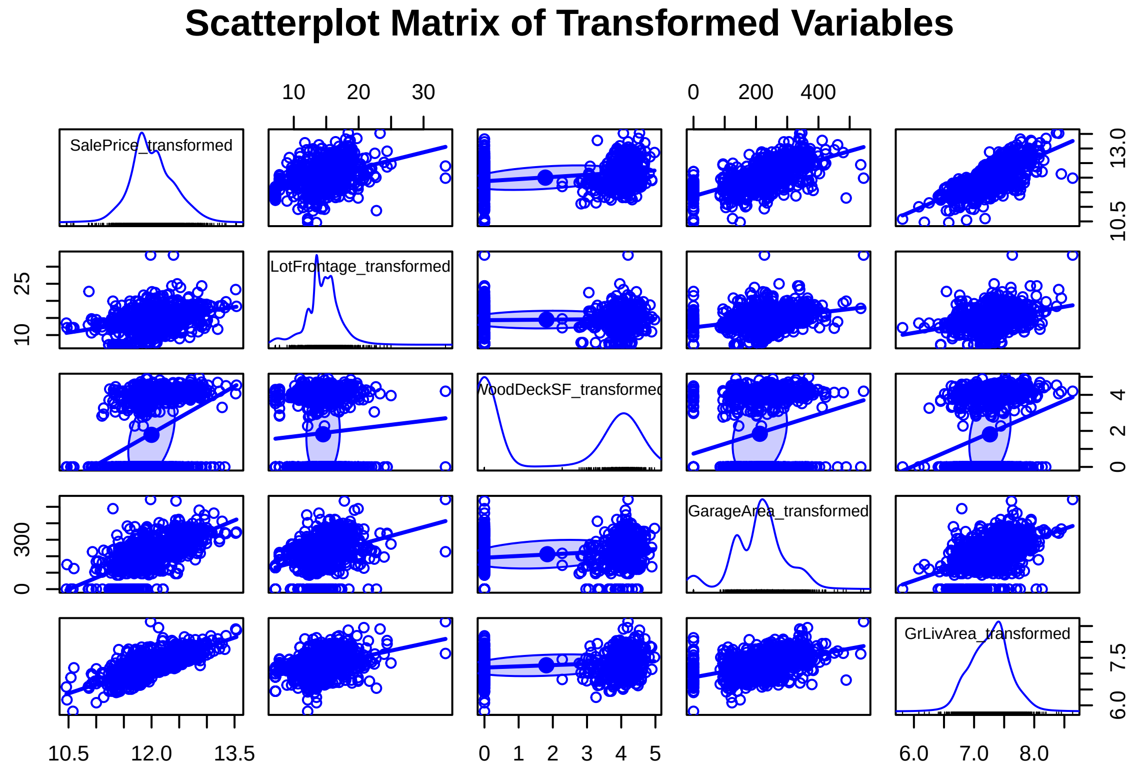
5. This plot shows boxplots for different power transformations of GrLivArea (powers: -1, -0.5, log, 0.5, 1), helping to assess distribution symmetry. It aids in selecting a transformation that may improve model assumptions.

```
# Apply powerTransform to multiple variables
# Use Box-Cox transformation to assess whether variables need normalization
# GarageArea and WoodDeckSF contain negative values, requiring shifting before transformation
# Variable wooddecksf and GarageArea contains negative values so we need to transform it by
```

```r
# adding 1 to each
WoodDeckSF1 = data$WoodDeckSF + 1
GarageArea1 = data$GarageArea + 1

# Box-Cox transformation helps reduce skewness, improving normality and variance stability
# Apply powerTransform to multiple variables
a3 <- powerTransform(cbind(SalePrice, LotFrontage, WoodDeckSF1, GarageArea1, GrLivArea) ~ 1,
                     data = data)

# View the summary of the transformations
summary(a3)
```

```
## bcPower Transformations to Multinormality
##             Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## SalePrice     -0.0303        0.00      -0.1042       0.0436
## LotFrontage    0.4226        0.50       0.3323       0.5130
## WoodDeckSF1   -0.0957       -0.10      -0.1320      -0.0594
## GarageArea1    0.8452        0.85       0.8001       0.8902
## GrLivArea      0.0065        0.00      -0.1012       0.1143
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                  LRT df       pval
## LR test, lambda = (0 0 0 0 0) 2494.766  5 < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
##                                  LRT df       pval
## LR test, lambda = (1 1 1 1 1) 4587.745  5 < 2.22e-16
```

```r
# Transform selected variables using Box-Cox with coefficients from a3
# Generate transformed data for further analysis
transformeddata <- as.data.frame(bcPower(
with(data, cbind(SalePrice, LotFrontage, WoodDeckSF1, GarageArea1, GrLivArea)),
coef(a3, round = TRUE)
))

# Rename columns in transformeddata for clarity (optional)
colnames(transformeddata) <- c("SalePrice_transformed", "LotFrontage_transformed",
"WoodDeckSF_transformed", "GarageArea_transformed",
"GrLivArea_transformed")

# Plot scatterplot matrix of transformed variables to observe distribution shape after
# transformation
scatterplotMatrix(~ SalePrice_transformed + LotFrontage_transformed + WoodDeckSF_transformed +
GarageArea_transformed + GrLivArea_transformed,
data = transformeddata,
smooth = FALSE, ellipse = list(levels = 0.5),
main = "Scatterplot Matrix of Transformed Variables")
```



Comment:

This scatterplot matrix visualizes relationships between the transformed variables. The transformations are intended to improve linearity and reduce skewness, helping to meet model assumptions. Ellipses show data concentration at the 0.5 level.

```r
#fit data
# Model fitting and evaluation for each variable
# Transformed variables are expected to better meet normality and linearity assumptions

library(broom)
# Fit linear models
# Model fitting and evaluation for each variable
# Transformed variables are expected to better meet normality and linearity assumptions

model1 <- lm(SalePrice_transformed ~ LotFrontage_transformed, data = transformeddata)
# Model 1: SalePrice ~ LotFrontage

model2 <- lm(SalePrice_transformed ~ WoodDeckSF_transformed, data = transformeddata)
# Model 2: SalePrice ~ WoodDeckSF

model3 <- lm(SalePrice_transformed ~ GarageArea_transformed, data = transformeddata)
# Model 3: SalePrice ~ GarageArea

model4 <- lm(SalePrice_transformed ~ GrLivArea_transformed, data = transformeddata)
# Model 4: SalePrice ~ GrLivArea
# Summarize each model's output
# Model summaries provide coefficients, standard errors, p-values, and R-squared values.
# Coefficients reveal the direction and magnitude of relationships, with p-values indicating
# statistical significance.

summary_model1 <- summary(model1)
summary_model2 <- summary(model2)
summary_model3 <- summary(model3)
summary_model4 <- summary(model4)
# Store summaries in a list for easy viewing
model_summaries <- list(
 model1 = summary_model1,
 model2 = summary_model2,
 model3 = summary_model3,
 model4 = summary_model4
)
# Display the summaries
model_summaries
```

```
## $model1
##
## Call:
## lm(formula = SalePrice_transformed ~ LotFrontage_transformed,
##     data = transformeddata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.60057 -0.24263 -0.03524  0.25241  1.33355
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             11.217331   0.058977  190.20   <2e-16 ***
## LotFrontage_transformed  0.055052   0.003993   13.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3867 on 1199 degrees of freedom
##   (    259     )
## Multiple R-squared:  0.1369, Adjusted R-squared:  0.1361
## F-statistic: 190.1 on 1 and 1199 DF,  p-value: < 2.2e-16
##
##
## $model2
##
## Call:
## lm(formula = SalePrice_transformed ~ WoodDeckSF_transformed,
##     data = transformeddata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.43494 -0.24201 -0.02119  0.21855  1.42874
##
## Coefficients:
```

```
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            11.89519    0.01358  876.24   <2e-16 ***
## WoodDeckSF_transformed  0.06654    0.00483   13.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3759 on 1458 degrees of freedom
## Multiple R-squared:  0.1152, Adjusted R-squared:  0.1145
## F-statistic: 189.7 on 1 and 1458 DF,  p-value: < 2.2e-16
##
##
## $model3
##
## Call:
## lm(formula = SalePrice_transformed ~ GarageArea_transformed,
##     data = transformeddata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.56536 -0.15902  0.01365  0.18022  1.10551
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            1.138e+01  2.125e-02  535.69   <2e-16 ***
## GarageArea_transformed 3.038e-03  9.332e-05   32.56   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3041 on 1458 degrees of freedom
## Multiple R-squared:  0.4209, Adjusted R-squared:  0.4206
## F-statistic:  1060 on 1 and 1458 DF,  p-value: < 2.2e-16
##
##
## $model4
##
## Call:
## lm(formula = SalePrice_transformed ~ GrLivArea_transformed, data = transformeddata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.35338 -0.14260  0.02864  0.16585  0.86377
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            5.66812    0.15589   36.36   <2e-16 ***
## GrLivArea_transformed  0.87454    0.02143   40.81   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.273 on 1458 degrees of freedom
## Multiple R-squared:  0.5333, Adjusted R-squared:  0.533
## F-statistic:  1666 on 1 and 1458 DF,  p-value: < 2.2e-16
```

comments:

Model 1: SalePrice ~ LotFrontage Statistically significant p-value for LotFrontage coefficient would indicate it as a meaningful predictor of SalePrice. Economically, a positive coefficient implies higher LotFrontage correlates with increased property value. Interpretation: For each unit increase in LotFrontage, SalePrice is expected to increase by the coefficient value, all else equal.

Model 2: SalePrice ~ WoodDeckSF Statistically, a significant p-value would mean WoodDeckSF is a meaningful predictor of SalePrice. Economically, a positive coefficient suggests that larger deck space adds value to properties. Interpretation: Each additional square foot in WoodDeckSF is expected to increase SalePrice by the coefficient amount, holding other factors constant.

Model 3: SalePrice ~ GarageArea A statistically significant coefficient for GarageArea would indicate it as a key factor influencing SalePrice. Economically, larger garage space is often valued by buyers, potentially adding to property worth. Interpretation: The coefficient here implies that each additional square foot of garage area increases SalePrice, assuming other variables remain unchanged.

Model 4: SalePrice ~ GrLivArea Statistically, low p-values and high R-squared suggest GrLivArea as a strong predictor of SalePrice. Economically, this reflects buyers' high valuation of living space, making GrLivArea a key factor in pricing. Interpretation: The coefficient indicates how much SalePrice increases for each additional square foot of GrLivArea, highlighting its impact on property value.

```
#R-squared and Adjusted R-squared:
#Model 4 has the highest R-squared (0.5333), indicating that approximately 53.33% of the
#variance in SalePrice_#Model 3 is next with an R-squared of 0.4209, indicating that
#GarageArea_transformed explains about 42.09% 32
#Statistical Significance:
#All models show highly significant p-values (< 2e-16), indicating that the predictors have a
#statistically #Economic Significance:
#The coefficient for GrLivArea_transformed (0.8745) suggests that for each unit increase in
#GrLivArea_transformed, #The coefficient for LotFrontage_transformed (0.0551) is smaller,
#indicating a weaker economic impact compared
```

```
#model4 is the best model because of low p values and high R^2
# Calculate confidence intervals for Model 4
# Confidence intervals provide a range within which the true parameter value is likely to lie,
#with a specified confidence level (e.g., 95%).
# Narrow intervals indicate greater precision in parameter estimates; if the interval does not
#include zero, it suggests statistical significance.

confint_model4 <- confint(model4, level = 0.95)
print(confint_model4)
```

```
##                           2.5 %   97.5 %
## (Intercept)           5.3623350 5.973914
## GrLivArea_transformed 0.8325049 0.916566
```

```
#confidence intervals are within a small range
```

```
#Performing bootstrapping for parameters with 1000 samples
```

```
# Bootstrapping for Model 4 estimates
# Bootstrapping generates resampled estimates to assess the stability and variability of
#parameters without assuming a normal distribution.
# This approach provides robust estimates, especially if the original data distribution is
#unknown or violates normality.
```

```
# The mean of bootstrap estimates close to the original values indicates reliable estimates,
#with distributions visualized to understand parameter variability.

library(boot)
```

```
##
##    'boot'
```

```
## The following object is masked from 'package:car':
##
##     logit
```

```
## The following object is masked from 'package:psych':
##
##     logit
```

```
library(car)
set.seed(3435)
betahat.boot = Boot(model4, R=1000)# Perform bootstrap sampling with 1000 samples
usualEsts = summary(model4)$coef[, 1:2]
summary(betahat.boot)# Summarize bootstrap results
```

```
##
## Number of bootstrap replications R = 1000
##                        original   bootBias   bootSE bootMed
## (Intercept)             5.66812  0.0081272 0.187748 5.66987
## GrLivArea_transformed   0.87454 -0.0011562 0.026107 0.87462
```

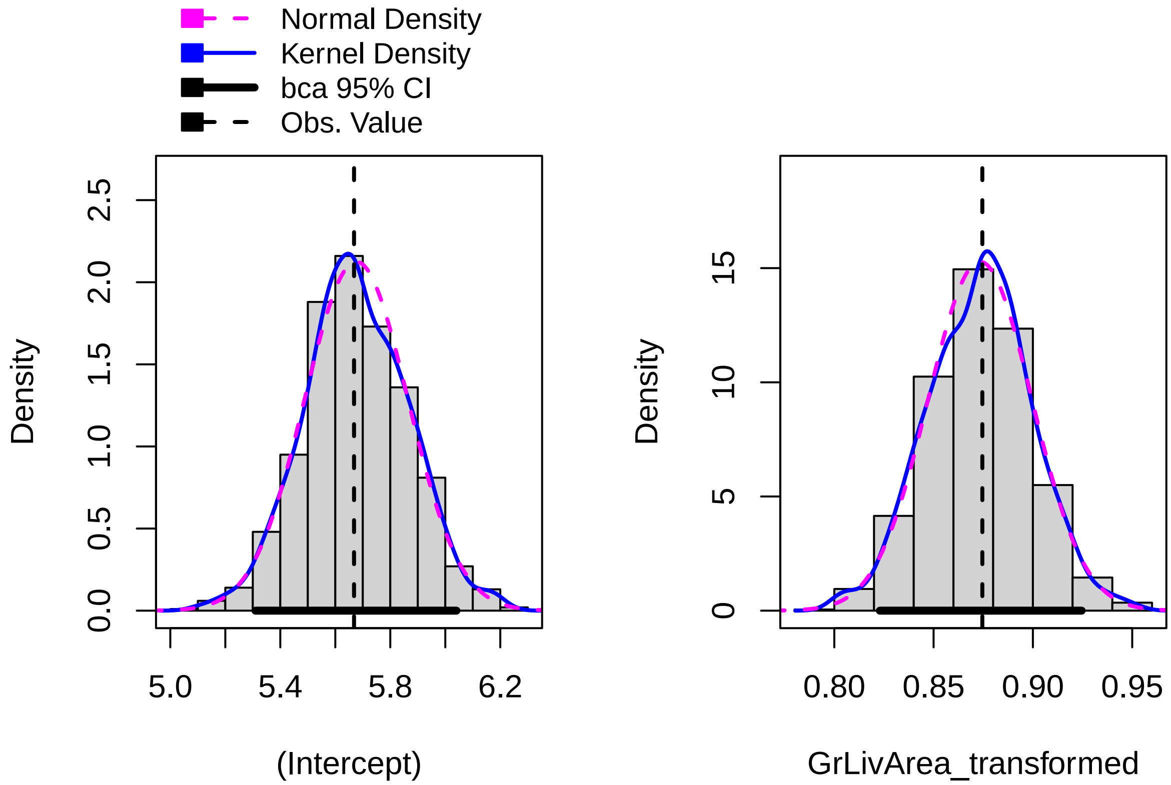
```
# Confidence intervals from bootstrapped estimates
# These intervals help validate the original confidence intervals, providing additional robustness.

confint(betahat.boot)
```

```
## Warning in confint.boot(betahat.boot): BCa method fails for this problem.
## Using 'perc' instead
```

```
## Bootstrap percent confidence intervals
##
##                            2.5 %    97.5 %
## (Intercept)            5.3097093 6.0405171
## GrLivArea_transformed 0.8229458 0.9245575
```

```
hist(betahat.boot)
```

```
## Warning in confint.boot(x, type = ci, level = level): BCa method fails for this
## problem.  Using 'perc' instead
```



Comment:

This plot shows the bootstrap density estimates for the Intercept and the GrLivArea_transformed variable. The solid line represents the Kernel density, and the dashed line represents the normal density, allowing for shape comparison. The thick black line represents the 95% bootstrap confidence interval.

```
#Means of parameters after bootstrapping are close to the original values
```

```
# Bootstrapping for R-squared

# Define the bootstrapping function
boot_fn <- function(data, indices) {
  # Resample the data using the indices
  d <- data[indices, ]

  # Fit the linear model
  model <- lm(SalePrice_transformed ~ GrLivArea_transformed, data = d)

  # Extract coefficients
  coefs <- coef(model)

  # Extract R-squared
  r_squared <- summary(model)$r.squared

  # Return both coefficients and R-squared
  return(c(coefs, r_squared))
}
```

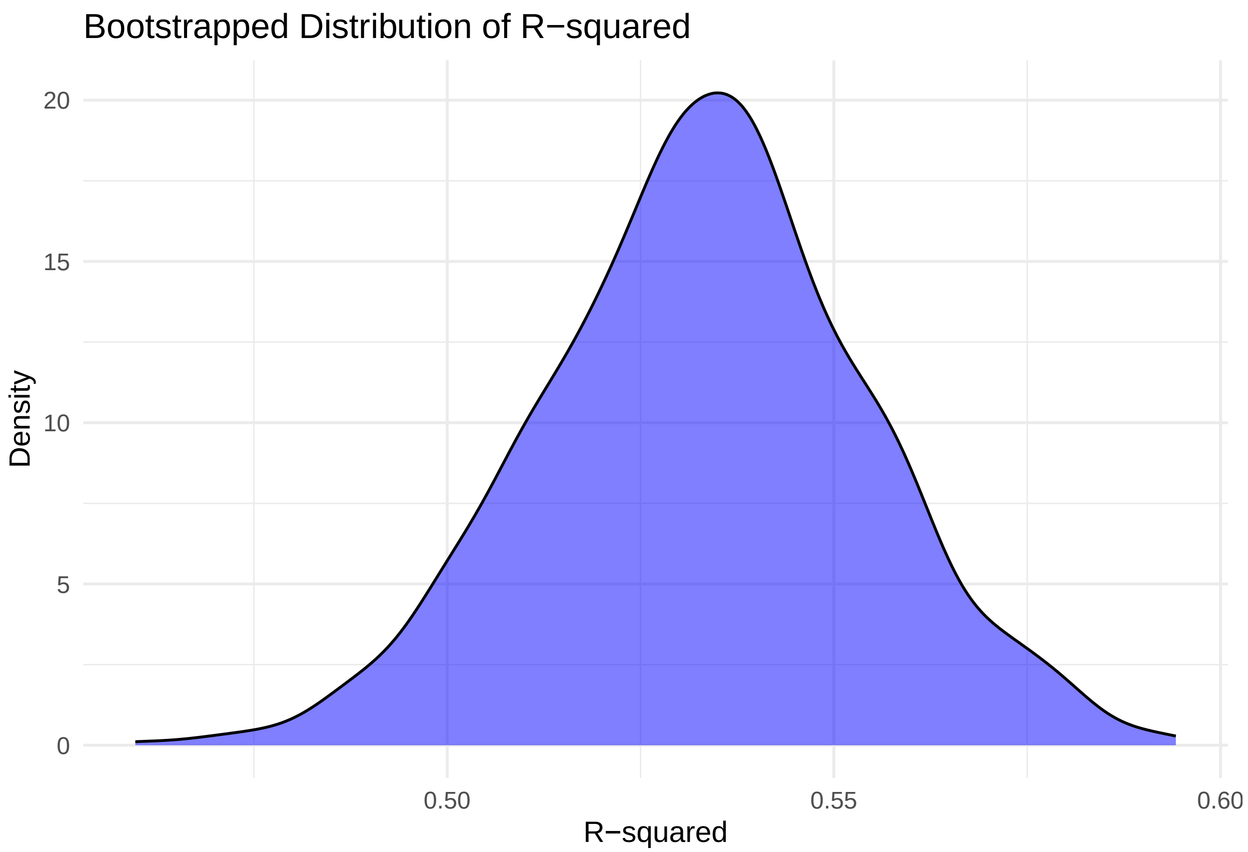
```r
#Set parameters for bootstrapping
set.seed(3435)

R <- 1000 # Number of bootstrap samples
dd <- transformeddata # Your dataset

# Perform bootstrapping
betahat.boot <- boot(data = dd, statistic = boot_fn, R = R)

# Check the structure of the bootstrapped results
str(betahat.boot)
```

```
## List of 11
##  $ t0       : Named num [1:3] 5.668 0.875 0.533
##   ..- attr(*, "names")= chr [1:3] "(Intercept)" "GrLivArea_transformed" ""
##  $ t        : num [1:1000, 1:3] 5.66 5.82 5.71 5.53 5.39 ...
##  $ R        : num 1000
##  $ data     :'data.frame':   1460 obs. of  5 variables:
##   ..$ SalePrice_transformed  : num [1:1460] 12.2 12.1 12.3 11.8 12.4 ...
##   ..$ LotFrontage_transformed: num [1:1460] 14.1 15.9 14.5 13.5 16.3 ...
##   ..$ WoodDeckSF_transformed : num [1:1460] 0 4.39 0 0 4.13 ...
##   ..$ GarageArea_transformed : num [1:1460] 243 210 266 278 348 ...
##   ..$ GrLivArea_transformed  : num [1:1460] 7.44 7.14 7.49 7.45 7.7 ...
##  $ seed     : int [1:626] 10403 624 -1829144649 -728749940 -1307864035 -1108863942 -266787725 -13516
##  $ statistic:function (data, indices)
##  $ sim      : chr "ordinary"
##  $ call     : language boot(data = dd, statistic = boot_fn, R = R)
##  $ stype    : chr "i"
##  $ strata   : num [1:1460] 1 1 1 1 1 1 1 1 1 1 ...
##  $ weights  : num [1:1460] 0.000685 0.000685 0.000685 0.000685 0.000685 ...
##  - attr(*, "class")= chr "boot"
##  - attr(*, "boot_type")= chr "boot"
```

```r
# Check the number of columns in betahat.boot$t
ncol(betahat.boot$t) # Ensure it returns 3
```

```
## [1] 3
```

```r
# Extract coefficients and R-squared
boot_coefs <- betahat.boot$t[, 1:2] # Coefficients
boot_r2 <- betahat.boot$t[, 3] # R-squared values

# Plotting the distribution of bootstrapped R-squared values
# A tight distribution around the original R-squared value shows consistency,
#reinforcing model fit.
ggplot(data.frame(R2 = boot_r2), aes(x = R2)) +
geom_density(fill = "blue", alpha = 0.5) +
labs(title = "Bootstrapped Distribution of R-squared",
x = "R-squared",
y = "Density") +
theme_minimal()
```

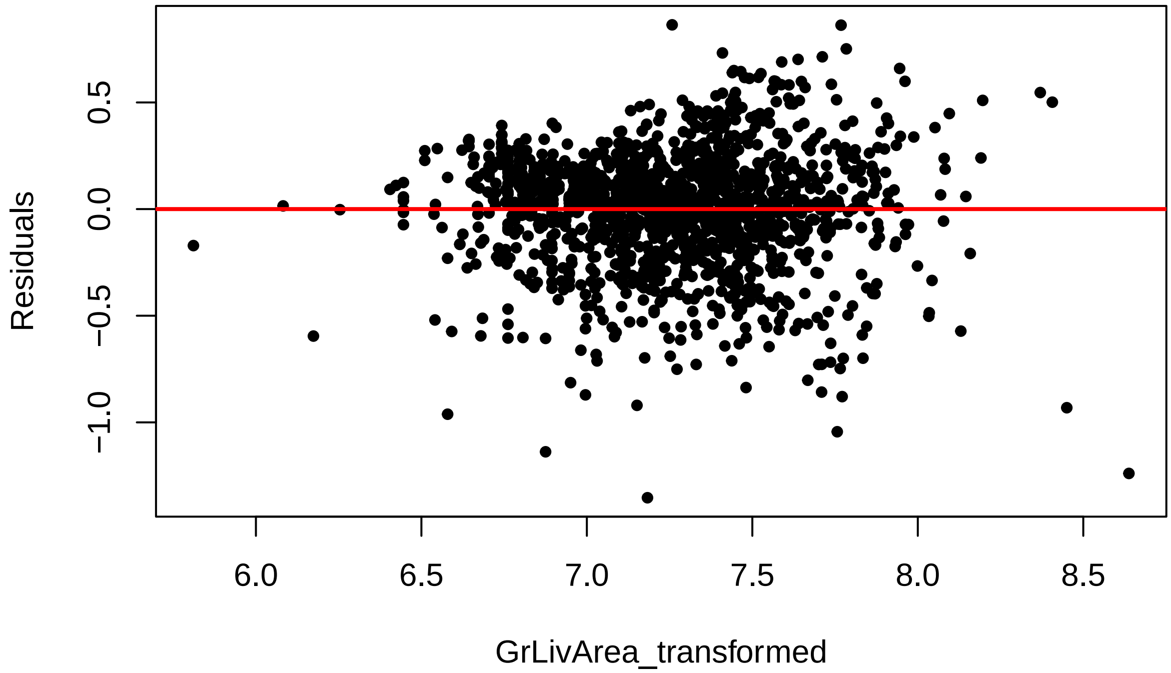


Comments:

This density plot visualizes the distribution of R-squared values generated from 1,000 bootstrap samples of the model. The shape and concentration of the distribution provide insight into the consistency of R-squared as a measure of the model's explanatory power. The peak around 0.52 reflects the most frequent R-squared value across these samples, reinforcing that the model captures approximately 52% of the variance in the response variable with moderate explanatory strength. The relatively tight spread around this central peak indicates that R-squared values remain stable across resamples, suggesting the model's fit is robust.

```
#the mean of boostrapping adjusted R^2 is close to the original value
```

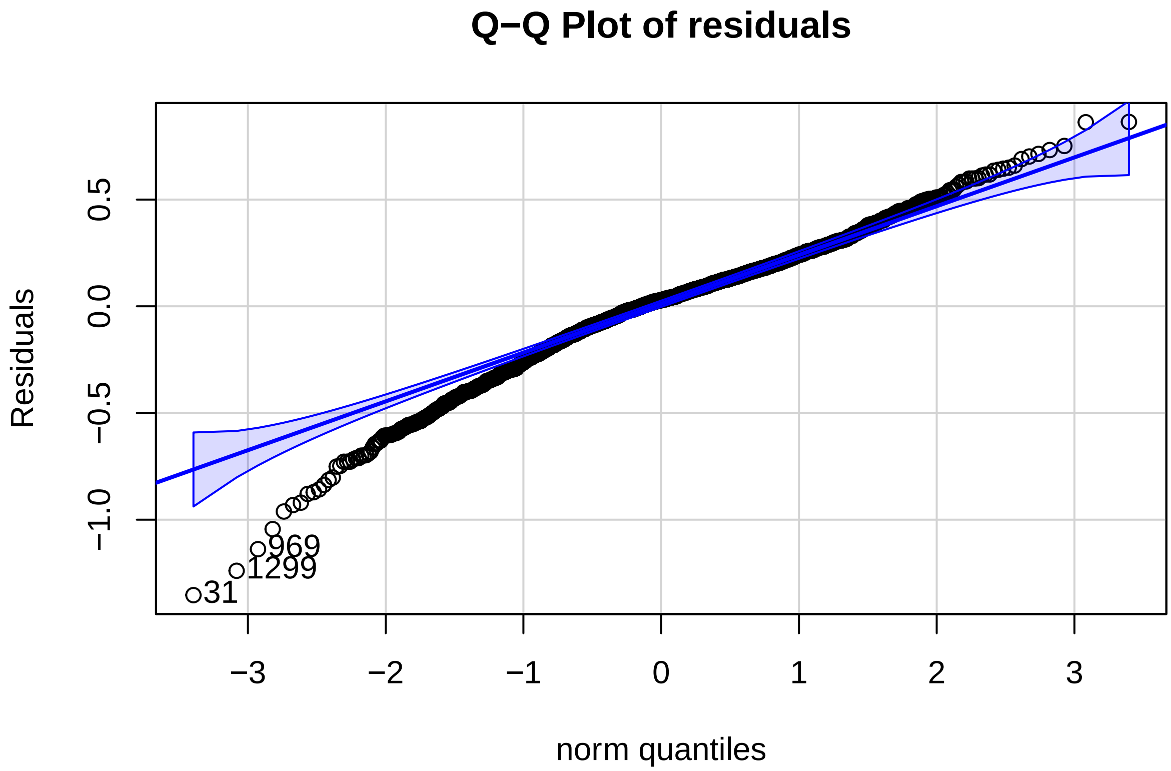
```
#Plot residuals of model4 and plot its qqplot
plot(transformeddata$GrLivArea_transformed,model4$residuals,pch=20, ylab="Residuals",
     xlab="GrLivArea_transformed")
abline(h=0, lwd=2, col="red")
```



Comments:

This residual plot displays the residuals against the transformed GrLivArea predictor, with a red horizontal line at zero representing an ideal mean for residuals. The plot helps assess homoscedasticity, an assumption that residuals have constant variance across levels of the predictor variable. Here, the residuals are randomly scattered around zero without forming any visible pattern, suggesting that the model does not exhibit heteroscedasticity, or unequal spread. This absence of structure indicates that the linear model appropriately fits the transformed predictor, supporting the assumptions of linearity and homoscedasticity for accurate model performance.

```r
residuals <- resid(model4)
qqPlot(residuals,
main = "Q-Q Plot of residuals",
ylab = "Residuals",
id = list(n = 3)) # Label the top 3 extreme points
```

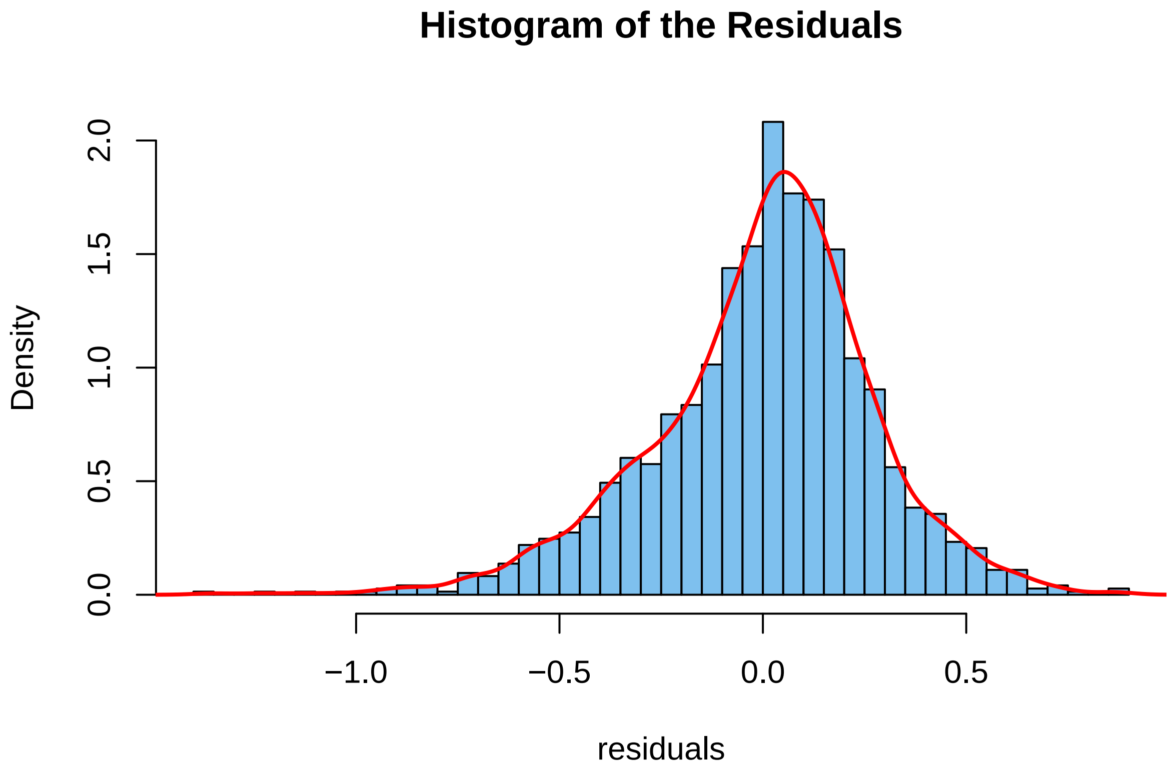


```
## [1]   31 1299  969
```

Comments:

This Q-Q plot compares the residuals to a theoretical normal distribution, with points that should ideally fall along the 45-degree line if the residuals are normally distributed. The close alignment of most points along this line confirms the assumption of approximate normality, indicating that the residuals largely fit the expected distribution shape. Minor deviations at the tails suggest slight non-normality in extreme values, but overall, the residuals appear approximately normal. This plot supports the assumption of normally distributed residuals in the model.

```
hist(residuals,breaks ="FD",col="skyblue2", freq = FALSE, ylab = "Density",
main = "Histogram of the Residuals")
lines(density(residuals),lwd = 2, col ="red")
```



Comments:

This histogram of the residuals is centered around zero, which indicates that our model does not consistently over- or under-predict the response variable, suggesting a good fit in terms of bias. The approximate symmetry around zero and the high peak near the center support the assumption of normally distributed residuals. The density curve (in red) overlaid on the histogram closely aligns with the shape of a normal distribution, although there is slight evidence of skewness in the tails, which may indicate minor deviations from normality. This skewness might affect predictions at the extremes but does not appear severe enough to invalidate the overall model. The residuals' distribution thus supports the validity of the normality assumption underlying the model.

#Summary

In this project, we analyze the determinants of house prices using the Ames Housing dataset, focusing on the key variables SalePrice, LotFrontage, WoodDeckSF, GarageArea, and GrLivArea. The goal is to estimate linear regression models to understand the relationship between SalePrice and each predictor.

We first performed a descriptive analysis of these variables and visualized their distributions through histograms, box plots, and QQ plots. Initially, the skewness of the SalePrice distribution was revealed, preparing us to consider transformations to achieve normality. We plotted scatter plots to explore the relationship between SalePrice and each predictor, identifying trends and nonlinear patterns that may require further transformation. To ensure that our model meets linear regression assumptions, we applied appropriate transformations where necessary. We then estimated separate linear regression models for each predictor for SalePrice. The results of each model were interpreted according to statistical significance, providing insight into how changes in each predictor affect house prices.

Throughout the analysis, we also and confidence intervals and p-values to assess statistical significance, and residual analyses to ensure that model assumptions were met. After a thorough evaluation, the model using GrLivArea emerged as the best choice.