STA9797_Group_Project_Code

Brandon Kokin, Ayrat Aymetov, Mohammed Saadman Chowdhury

2025-12-14

```r
# Load the dataset
AmesHousing <- read.csv("data/AmesHousing.csv")

# Print the first few rows
head(AmesHousing)
```

```
##   Order       PID MS.SubClass MS.Zoning Lot.Frontage Lot.Area Street Alley
## 1     1 526301100          20        RL          141    31770   Pave  <NA>
## 2     2 526350040          20        RH           80    11622   Pave  <NA>
## 3     3 526351010          20        RL           81    14267   Pave  <NA>
## 4     4 526353030          20        RL           93    11160   Pave  <NA>
## 5     5 527105010          60        RL           74    13830   Pave  <NA>
## 6     6 527105030          60        RL           78     9978   Pave  <NA>
##   Lot.Shape Land.Contour Utilities Lot.Config Land.Slope Neighborhood
## 1       IR1          Lvl    AllPub     Corner        Gtl        NAmes
## 2       Reg          Lvl    AllPub     Inside        Gtl        NAmes
## 3       IR1          Lvl    AllPub     Corner        Gtl        NAmes
## 4       Reg          Lvl    AllPub     Corner        Gtl        NAmes
## 5       IR1          Lvl    AllPub     Inside        Gtl      Gilbert
## 6       IR1          Lvl    AllPub     Inside        Gtl      Gilbert
##   Condition.1 Condition.2 Bldg.Type House.Style Overall.Qual Overall.Cond
## 1        Norm        Norm      1Fam      1Story            6            5
## 2       Feedr        Norm      1Fam      1Story            5            6
## 3        Norm        Norm      1Fam      1Story            6            6
## 4        Norm        Norm      1Fam      1Story            7            5
## 5        Norm        Norm      1Fam      2Story            5            5
## 6        Norm        Norm      1Fam      2Story            6            6
##   Year.Built Year.Remod.Add Roof.Style Roof.Matl Exterior.1st Exterior.2nd
## 1       1960           1960        Hip   CompShg      BrkFace      Plywood
## 2       1961           1961      Gable   CompShg      VinylSd      VinylSd
## 3       1958           1958        Hip   CompShg      Wd Sdng      Wd Sdng
## 4       1968           1968        Hip   CompShg      BrkFace      BrkFace
## 5       1997           1998      Gable   CompShg      VinylSd      VinylSd
## 6       1998           1998      Gable   CompShg      VinylSd      VinylSd
##   Mas.Vnr.Type Mas.Vnr.Area Exter.Qual Exter.Cond Foundation Bsmt.Qual
## 1        Stone          112         TA         TA     CBlock        TA
## 2         None            0         TA         TA     CBlock        TA
## 3      BrkFace          108         TA         TA     CBlock        TA
## 4         None            0         Gd         TA     CBlock        TA
## 5         None            0         TA         TA      PConc        Gd
## 6      BrkFace           20         TA         TA      PConc        TA
##   Bsmt.Cond Bsmt.Exposure BsmtFin.Type.1 BsmtFin.SF.1 BsmtFin.Type.2
```

```
## 1        Gd            Gd            BLQ          639           Unf
## 2        TA            No            Rec          468           LwQ
## 3        TA            No            ALQ          923           Unf
## 4        TA            No            ALQ         1065           Unf
## 5        TA            No            GLQ          791           Unf
## 6        TA            No            GLQ          602           Unf
##   BsmtFin.SF.2 Bsmt.Unf.SF Total.Bsmt.SF Heating Heating.QC Central.Air
## 1            0         441          1080    GasA         Fa           Y
## 2          144         270           882    GasA         TA           Y
## 3            0         406          1329    GasA         TA           Y
## 4            0        1045          2110    GasA         Ex           Y
## 5            0         137           928    GasA         Gd           Y
## 6            0         324           926    GasA         Ex           Y
##   Electrical X1st.Flr.SF X2nd.Flr.SF Low.Qual.Fin.SF Gr.Liv.Area Bsmt.Full.Bath
## 1      SBrkr        1656           0               0        1656              1
## 2      SBrkr         896           0               0         896              0
## 3      SBrkr        1329           0               0        1329              0
## 4      SBrkr        2110           0               0        2110              1
## 5      SBrkr         928         701               0        1629              0
## 6      SBrkr         926         678               0        1604              0
##   Bsmt.Half.Bath Full.Bath Half.Bath Bedroom.AbvGr Kitchen.AbvGr Kitchen.Qual
## 1              0         1         0            3             1           TA
## 2              0         1         0            2             1           TA
## 3              0         1         1            3             1           Gd
## 4              0         2         1            3             1           Ex
## 5              0         2         1            3             1           TA
## 6              0         2         1            3             1           Gd
##   TotRms.AbvGrd Functional Fireplaces Fireplace.Qu Garage.Type Garage.Yr.Blt
## 1             7        Typ          2          Gd      Attchd          1960
## 2             5        Typ          0        <NA>      Attchd          1961
## 3             6        Typ          0        <NA>      Attchd          1958
## 4             8        Typ          2          TA      Attchd          1968
## 5             6        Typ          1          TA      Attchd          1997
## 6             7        Typ          1          Gd      Attchd          1998
##   Garage.Finish Garage.Cars Garage.Area Garage.Qual Garage.Cond Paved.Drive
## 1           Fin           2         528          TA          TA           P
## 2           Unf           1         730          TA          TA           Y
## 3           Unf           1         312          TA          TA           Y
## 4           Fin           2         522          TA          TA           Y
## 5           Fin           2         482          TA          TA           Y
## 6           Fin           2         470          TA          TA           Y
##   Wood.Deck.SF Open.Porch.SF Enclosed.Porch X3Ssn.Porch Screen.Porch Pool.Area
## 1          210            62              0           0           0         0
## 2          140             0              0           0         120         0
## 3          393            36              0           0           0         0
## 4            0             0              0           0           0         0
## 5          212            34              0           0           0         0
## 6          360            36              0           0           0         0
##   Pool.QC Fence Misc.Feature Misc.Val Mo.Sold Yr.Sold Sale.Type Sale.Condition
## 1    <NA>  <NA>         <NA>        0       5    2010       WD          Normal
## 2    <NA> MnPrv         <NA>        0       6    2010       WD          Normal
## 3    <NA>  <NA>         Gar2    12500       6    2010       WD          Normal
## 4    <NA>  <NA>         <NA>        0       4    2010       WD          Normal
## 5    <NA> MnPrv         <NA>        0       3    2010       WD          Normal
```

```
## 6    <NA>  <NA>         <NA>        0       6    2010       WD          Normal
##   SalePrice
## 1    215000
## 2    105000
## 3    172000
## 4    244000
## 5    189900
## 6    195500
```

Setup and basic preparation

```
set.seed(123)

str(AmesHousing)
```

```
## 'data.frame':    2930 obs. of  82 variables:
##  $ Order          : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ PID            : int  526301100 526350040 526351010 526353030 527105010 527105030 527127150 527145
##  $ MS.SubClass    : int  20 20 20 20 60 60 120 120 120 60 ...
##  $ MS.Zoning      : chr  "RL" "RH" "RL" "RL" ...
##  $ Lot.Frontage   : int  141 80 81 93 74 78 41 43 39 60 ...
##  $ Lot.Area       : int  31770 11622 14267 11160 13830 9978 4920 5005 5389 7500 ...
##  $ Street         : chr  "Pave" "Pave" "Pave" "Pave" ...
##  $ Alley          : chr  NA NA NA NA ...
##  $ Lot.Shape      : chr  "IR1" "Reg" "IR1" "Reg" ...
##  $ Land.Contour   : chr  "Lvl" "Lvl" "Lvl" "Lvl" ...
##  $ Utilities      : chr  "AllPub" "AllPub" "AllPub" "AllPub" ...
##  $ Lot.Config     : chr  "Corner" "Inside" "Corner" "Corner" ...
##  $ Land.Slope     : chr  "Gtl" "Gtl" "Gtl" "Gtl" ...
##  $ Neighborhood   : chr  "NAmes" "NAmes" "NAmes" "NAmes" ...
##  $ Condition.1    : chr  "Norm" "Feedr" "Norm" "Norm" ...
##  $ Condition.2    : chr  "Norm" "Norm" "Norm" "Norm" ...
##  $ Bldg.Type      : chr  "1Fam" "1Fam" "1Fam" "1Fam" ...
##  $ House.Style    : chr  "1Story" "1Story" "1Story" "1Story" ...
##  $ Overall.Qual   : int  6 5 6 7 5 6 8 8 8 7 ...
##  $ Overall.Cond   : int  5 6 6 5 5 6 5 5 5 5 ...
##  $ Year.Built     : int  1960 1961 1958 1968 1997 1998 2001 1992 1995 1999 ...
##  $ Year.Remod.Add : int  1960 1961 1958 1968 1998 1998 2001 1992 1996 1999 ...
##  $ Roof.Style     : chr  "Hip" "Gable" "Hip" "Hip" ...
##  $ Roof.Matl      : chr  "CompShg" "CompShg" "CompShg" "CompShg" ...
##  $ Exterior.1st   : chr  "BrkFace" "VinylSd" "Wd Sdng" "BrkFace" ...
##  $ Exterior.2nd   : chr  "Plywood" "VinylSd" "Wd Sdng" "BrkFace" ...
##  $ Mas.Vnr.Type   : chr  "Stone" "None" "BrkFace" "None" ...
##  $ Mas.Vnr.Area   : int  112 0 108 0 0 20 0 0 0 0 ...
##  $ Exter.Qual     : chr  "TA" "TA" "TA" "Gd" ...
##  $ Exter.Cond     : chr  "TA" "TA" "TA" "TA" ...
##  $ Foundation     : chr  "CBlock" "CBlock" "CBlock" "CBlock" ...
##  $ Bsmt.Qual      : chr  "TA" "TA" "TA" "TA" ...
##  $ Bsmt.Cond      : chr  "Gd" "TA" "TA" "TA" ...
##  $ Bsmt.Exposure  : chr  "Gd" "No" "No" "No" ...
##  $ BsmtFin.Type.1 : chr  "BLQ" "Rec" "ALQ" "ALQ" ...
##  $ BsmtFin.SF.1   : int  639 468 923 1065 791 602 616 263 1180 0 ...
##  $ BsmtFin.Type.2 : chr  "Unf" "LwQ" "Unf" "Unf" ...
```

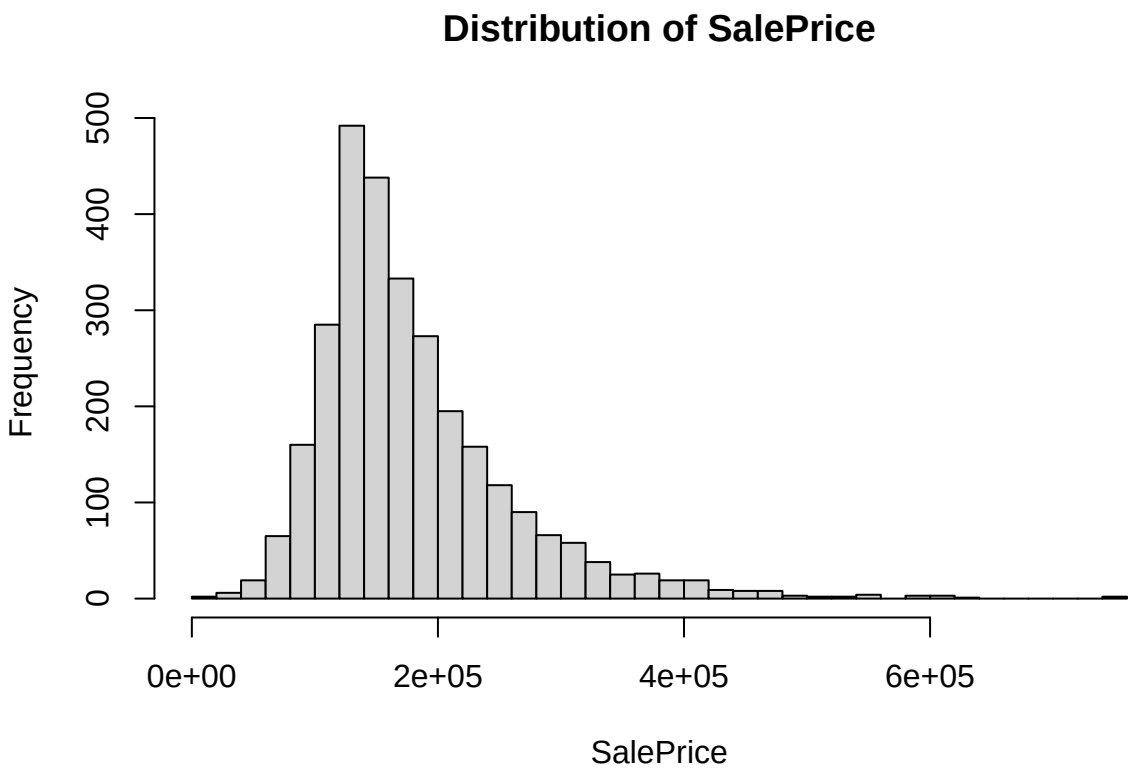
```
##  $ BsmtFin.SF.2   : int  0 144 0 0 0 0 0 0 0 0 ...
##  $ Bsmt.Unf.SF    : int  441 270 406 1045 137 324 722 1017 415 994 ...
##  $ Total.Bsmt.SF  : int  1080 882 1329 2110 928 926 1338 1280 1595 994 ...
##  $ Heating        : chr  "GasA" "GasA" "GasA" "GasA" ...
##  $ Heating.QC     : chr  "Fa" "TA" "TA" "Ex" ...
##  $ Central.Air    : chr  "Y" "Y" "Y" "Y" ...
##  $ Electrical     : chr  "SBrkr" "SBrkr" "SBrkr" "SBrkr" ...
##  $ X1st.Flr.SF    : int  1656 896 1329 2110 928 926 1338 1280 1616 1028 ...
##  $ X2nd.Flr.SF    : int  0 0 0 0 701 678 0 0 0 776 ...
##  $ Low.Qual.Fin.SF: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Gr.Liv.Area    : int  1656 896 1329 2110 1629 1604 1338 1280 1616 1804 ...
##  $ Bsmt.Full.Bath : int  1 0 0 1 0 0 1 0 1 0 ...
##  $ Bsmt.Half.Bath : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Full.Bath      : int  1 1 1 2 2 2 2 2 2 2 ...
##  $ Half.Bath      : int  0 0 1 1 1 1 0 0 0 1 ...
##  $ Bedroom.AbvGr  : int  3 2 3 3 3 3 2 2 2 3 ...
##  $ Kitchen.AbvGr  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Kitchen.Qual   : chr  "TA" "TA" "Gd" "Ex" ...
##  $ TotRms.AbvGrd  : int  7 5 6 8 6 7 6 5 5 7 ...
##  $ Functional     : chr  "Typ" "Typ" "Typ" "Typ" ...
##  $ Fireplaces     : int  2 0 0 2 1 1 0 0 1 1 ...
##  $ Fireplace.Qu   : chr  "Gd" NA NA "TA" ...
##  $ Garage.Type    : chr  "Attchd" "Attchd" "Attchd" "Attchd" ...
##  $ Garage.Yr.Blt  : int  1960 1961 1958 1968 1997 1998 2001 1992 1995 1999 ...
##  $ Garage.Finish  : chr  "Fin" "Unf" "Unf" "Fin" ...
##  $ Garage.Cars    : int  2 1 1 2 2 2 2 2 2 2 ...
##  $ Garage.Area    : int  528 730 312 522 482 470 582 506 608 442 ...
##  $ Garage.Qual    : chr  "TA" "TA" "TA" "TA" ...
##  $ Garage.Cond    : chr  "TA" "TA" "TA" "TA" ...
##  $ Paved.Drive    : chr  "P" "Y" "Y" "Y" ...
##  $ Wood.Deck.SF   : int  210 140 393 0 212 360 0 0 237 140 ...
##  $ Open.Porch.SF  : int  62 0 36 0 34 36 0 82 152 60 ...
##  $ Enclosed.Porch : int  0 0 0 0 0 0 170 0 0 0 ...
##  $ X3Ssn.Porch    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Screen.Porch   : int  0 120 0 0 0 0 0 144 0 0 ...
##  $ Pool.Area      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Pool.QC        : chr  NA NA NA NA ...
##  $ Fence          : chr  NA "MnPrv" NA NA ...
##  $ Misc.Feature   : chr  NA NA "Gar2" NA ...
##  $ Misc.Val       : int  0 0 12500 0 0 0 0 0 0 0 ...
##  $ Mo.Sold        : int  5 6 6 4 3 6 4 1 3 6 ...
##  $ Yr.Sold        : int  2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 ...
##  $ Sale.Type      : chr  "WD " "WD " "WD " "WD " ...
##  $ Sale.Condition : chr  "Normal" "Normal" "Normal" "Normal" ...
##  $ SalePrice      : int  215000 105000 172000 244000 189900 195500 213500 191500 236500 189000 ...
```

```
summary(AmesHousing$SalePrice)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12789  129500  160000  180796  213500  755000
```

Distribution of SalePrice (raw scale)

```r
hist(
AmesHousing$SalePrice,
breaks = 50,
main = "Distribution of SalePrice",
xlab = "SalePrice"
)
```



Initial linear model using raw SalePrice

```r
ols_initial <- lm(
SalePrice ~ Gr.Liv.Area + Total.Bsmt.SF + Garage.Area +
Lot.Area + Overall.Qual + Overall.Cond +
Year.Built + Year.Remod.Add +
Neighborhood,
data = AmesHousing
)

summary(ols_initial)
```

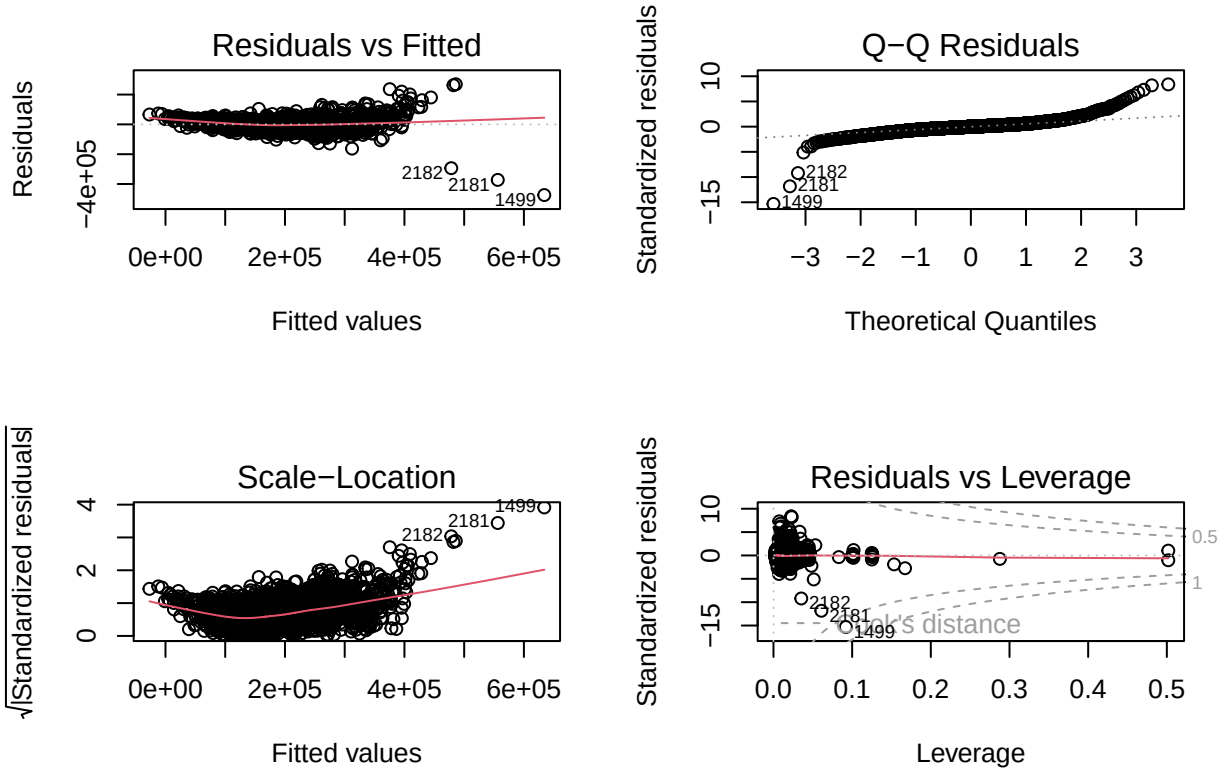
```
##
## Call:
## lm(formula = SalePrice ~ Gr.Liv.Area + Total.Bsmt.SF + Garage.Area +
##     Lot.Area + Overall.Qual + Overall.Cond + Year.Built + Year.Remod.Add +
##     Neighborhood, data = AmesHousing)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -474106  -13237    -461   11585  269018
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -1.132e+06  1.078e+05 -10.493  < 2e-16 ***
## Gr.Liv.Area          4.702e+01  1.711e+00  27.486  < 2e-16 ***
## Total.Bsmt.SF        2.439e+01  1.847e+00  13.204  < 2e-16 ***
## Garage.Area          3.857e+01  3.842e+00  10.038  < 2e-16 ***
## Lot.Area             7.054e-01  8.804e-02   8.012 1.62e-15 ***
## Overall.Qual         1.458e+04  7.837e+02  18.600  < 2e-16 ***
## Overall.Cond         6.253e+03  6.707e+02   9.323  < 2e-16 ***
## Year.Built           3.973e+02  4.871e+01   8.156 5.09e-16 ***
## Year.Remod.Add       1.381e+02  4.448e+01   3.104  0.00193 **
## NeighborhoodBlueste -1.118e+04  1.205e+04  -0.927  0.35384
## NeighborhoodBrDale  -1.493e+04  8.750e+03  -1.706  0.08806 .
## NeighborhoodBrkSide  9.804e+03  7.607e+03   1.289  0.19759
## NeighborhoodClearCr  1.625e+04  8.222e+03   1.977  0.04817 *
## NeighborhoodCollgCr  5.864e+03  6.500e+03   0.902  0.36707
## NeighborhoodCrawfor  3.111e+04  7.414e+03   4.197 2.79e-05 ***
## NeighborhoodEdwards -4.671e+02  6.974e+03  -0.067  0.94660
## NeighborhoodGilbert  2.419e+03  6.745e+03   0.359  0.71992
## NeighborhoodGreens   1.206e+04  1.314e+04   0.918  0.35873
## NeighborhoodGrnHill  1.099e+05  2.380e+04   4.616 4.08e-06 ***
## NeighborhoodIDOTRR   1.878e+03  7.798e+03   0.241  0.80971
## NeighborhoodLandmrk -1.593e+04  3.306e+04  -0.482  0.62986
## NeighborhoodMeadowV -3.326e+03  8.441e+03  -0.394  0.69363
## NeighborhoodMitchel  7.685e+02  7.026e+03   0.109  0.91291
## NeighborhoodNAmes    2.510e+03  6.733e+03   0.373  0.70936
## NeighborhoodNoRidge  5.676e+04  7.455e+03   7.614 3.57e-14 ***
## NeighborhoodNPkVill -1.133e+04  9.274e+03  -1.222  0.22177
## NeighborhoodNridgHt  6.019e+04  6.730e+03   8.944  < 2e-16 ***
## NeighborhoodNWAmes  -4.498e+03  6.989e+03  -0.644  0.51992
## NeighborhoodOldTown -2.816e+03  7.405e+03  -0.380  0.70375
## NeighborhoodSawyer   3.915e+03  7.030e+03   0.557  0.57764
## NeighborhoodSawyerW -1.107e+03  6.885e+03  -0.161  0.87232
## NeighborhoodSomerst  1.517e+04  6.624e+03   2.290  0.02207 *
## NeighborhoodStoneBr  7.012e+04  7.723e+03   9.080  < 2e-16 ***
## NeighborhoodSWISU    9.174e+02  8.416e+03   0.109  0.91321
## NeighborhoodTimber   2.202e+04  7.351e+03   2.995  0.00277 **
## NeighborhoodVeenker  2.221e+04  9.179e+03   2.420  0.01559 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32460 on 2892 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.8369, Adjusted R-squared:  0.8349
## F-statistic:   424 on 35 and 2892 DF,  p-value: < 2.2e-16
```

Diagnostics for initial SalePrice model

```
par(mfrow = c(2, 2))
plot(ols_initial)
```

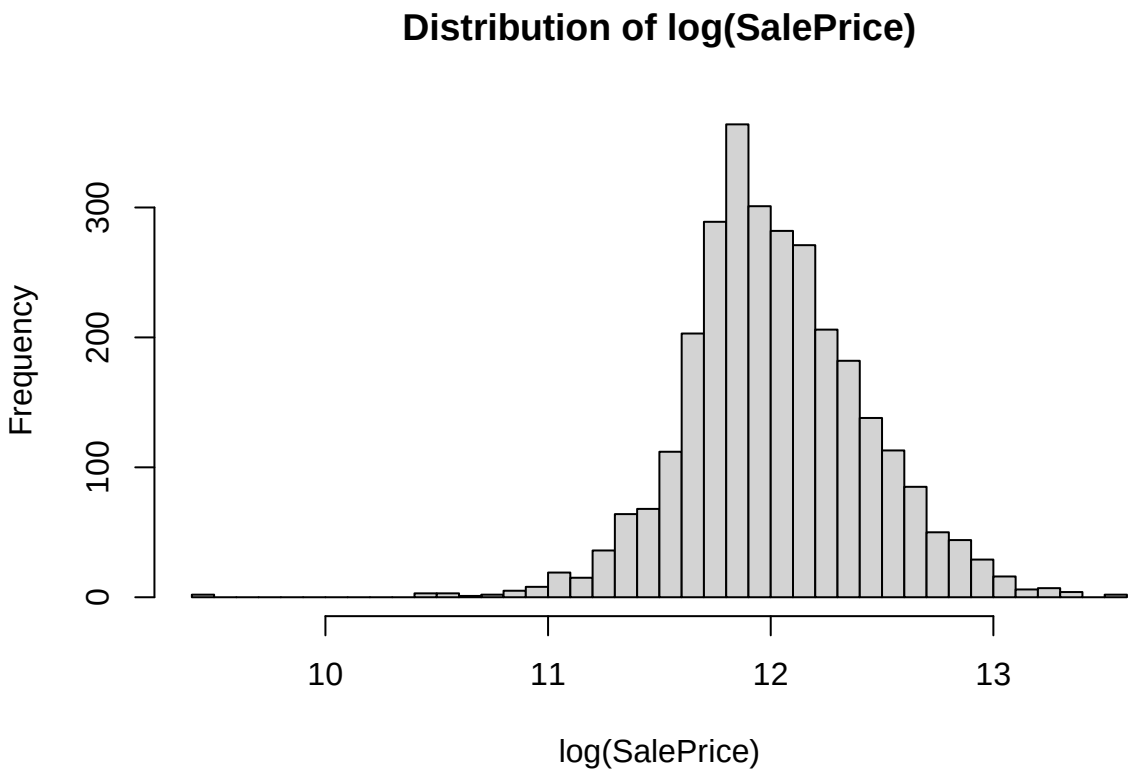
```
## Warning: not plotting observations with leverage one:
##   2787
```




```
par(mfrow = c(1, 1))
```

Log transformation of SalePrice

```
AmesHousing$logSalePrice <- log(AmesHousing$SalePrice)
hist(
AmesHousing$logSalePrice,
breaks = 50,
main = "Distribution of log(SalePrice)",
xlab = "log(SalePrice)"
)
```



Linear model using log(SalePrice)

```
ols_log <- lm(
logSalePrice ~ Gr.Liv.Area + Total.Bsmt.SF + Garage.Area +
Lot.Area + Overall.Qual + Overall.Cond +
Year.Built + Year.Remod.Add +
Neighborhood,
data = AmesHousing
)

summary(ols_log)
```

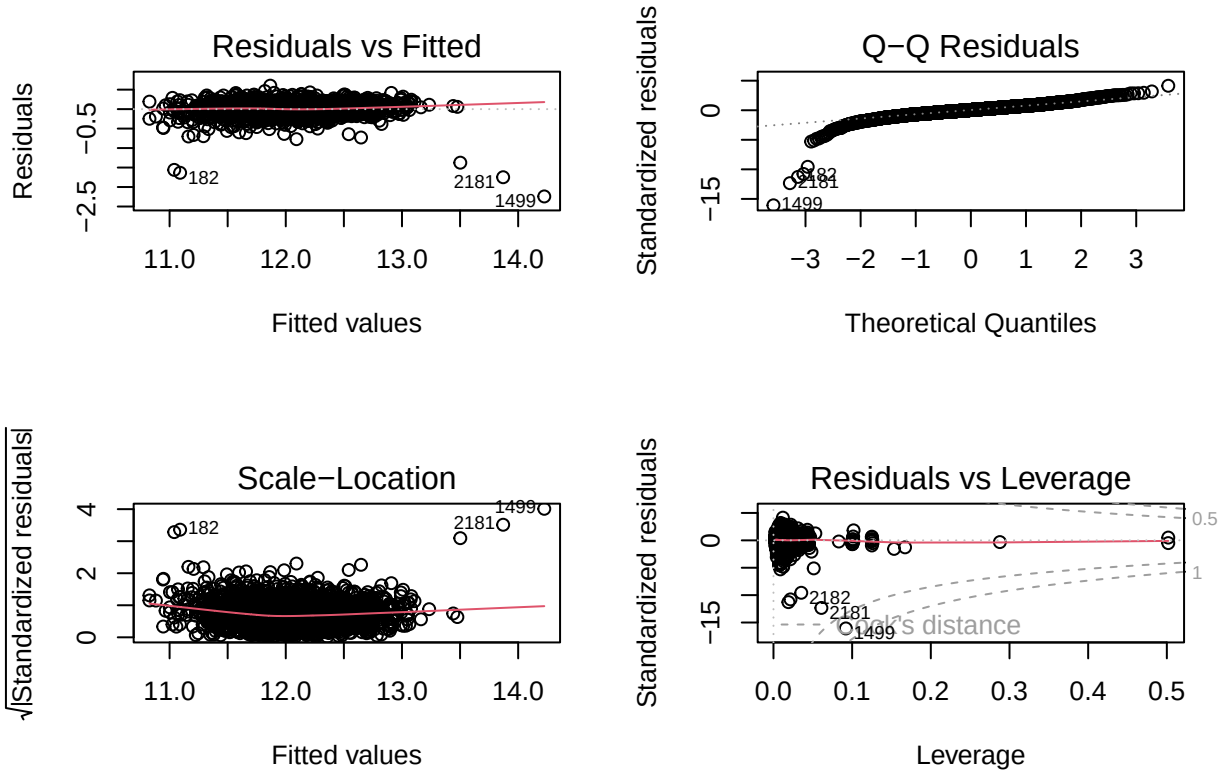
```
##
## Call:
## lm(formula = logSalePrice ~ Gr.Liv.Area + Total.Bsmt.SF + Garage.Area +
##     Lot.Area + Overall.Qual + Overall.Cond + Year.Built + Year.Remod.Add +
##     Neighborhood, data = AmesHousing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.24173 -0.06421  0.00636  0.07729  0.60177
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.364e+00  4.860e-01   6.922 5.45e-12 ***
## Gr.Liv.Area          2.388e-04  7.710e-06  30.973  < 2e-16 ***
```

```
## Total.Bsmt.SF        1.095e-04  8.325e-06  13.155  < 2e-16 ***
## Garage.Area          1.951e-04  1.732e-05  11.267  < 2e-16 ***
## Lot.Area             2.505e-06  3.968e-07   6.313 3.15e-10 ***
## Overall.Qual         8.284e-02  3.532e-03  23.455  < 2e-16 ***
## Overall.Cond         5.627e-02  3.023e-03  18.616  < 2e-16 ***
## Year.Built           2.747e-03  2.195e-04  12.513  < 2e-16 ***
## Year.Remod.Add       9.225e-04  2.005e-04   4.602 4.37e-06 ***
## NeighborhoodBlueste -1.132e-01  5.432e-02  -2.085 0.037178 *
## NeighborhoodBrDale  -2.107e-01  3.943e-02  -5.344 9.81e-08 ***
## NeighborhoodBrkSide -2.606e-02  3.428e-02  -0.760 0.447219
## NeighborhoodClearCr  1.126e-01  3.705e-02   3.040 0.002387 **
## NeighborhoodCollgCr  1.618e-02  2.929e-02   0.552 0.580834
## NeighborhoodCrawfor  1.490e-01  3.341e-02   4.461 8.47e-06 ***
## NeighborhoodEdwards -7.120e-02  3.143e-02  -2.266 0.023550 *
## NeighborhoodGilbert  2.142e-02  3.040e-02   0.705 0.481055
## NeighborhoodGreens   7.351e-02  5.920e-02   1.242 0.214392
## NeighborhoodGrnHill  4.937e-01  1.073e-01   4.604 4.33e-06 ***
## NeighborhoodIDOTRR  -1.583e-01  3.514e-02  -4.505 6.91e-06 ***
## NeighborhoodLandmrk -1.172e-01  1.490e-01  -0.787 0.431483
## NeighborhoodMeadowV -2.054e-01  3.804e-02  -5.399 7.23e-08 ***
## NeighborhoodMitchel -3.369e-03  3.166e-02  -0.106 0.915274
## NeighborhoodNAmes   -7.839e-04  3.034e-02  -0.026 0.979391
## NeighborhoodNoRidge  1.135e-01  3.360e-02   3.379 0.000737 ***
## NeighborhoodNPkVill -8.605e-02  4.179e-02  -2.059 0.039596 *
## NeighborhoodNridgHt  1.317e-01  3.033e-02   4.341 1.47e-05 ***
## NeighborhoodNWAmes  -1.316e-02  3.150e-02  -0.418 0.676005
## NeighborhoodOldTown -1.013e-01  3.337e-02  -3.036 0.002420 **
## NeighborhoodSawyer  -4.680e-03  3.168e-02  -0.148 0.882567
## NeighborhoodSawyerW -1.572e-02  3.103e-02  -0.507 0.612353
## NeighborhoodSomerst  4.627e-02  2.985e-02   1.550 0.121262
## NeighborhoodStoneBr  1.579e-01  3.481e-02   4.537 5.94e-06 ***
## NeighborhoodSWISU   -1.493e-02  3.793e-02  -0.394 0.693915
## NeighborhoodTimber   7.810e-02  3.313e-02   2.358 0.018462 *
## NeighborhoodVeenker  7.216e-02  4.136e-02   1.744 0.081191 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1463 on 2892 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.8727, Adjusted R-squared:  0.8711
## F-statistic: 566.4 on 35 and 2892 DF,  p-value: < 2.2e-16
```

Diagnostics for log-transformed model

```
par(mfrow = c(2, 2))
plot(ols_log)
```

```
## Warning: not plotting observations with leverage one:
##   2787
```




```
par(mfrow = c(1, 1))
```

ANOVA for categorical predictors (log scale)

```
anova_neighborhood <- aov(
logSalePrice ~ Neighborhood,
data = AmesHousing
)
summary(anova_neighborhood)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## Neighborhood   27  283.4   10.50   149.9 <2e-16 ***
## Residuals    2902  203.2    0.07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova_quality <- aov(
logSalePrice ~ factor(Overall.Qual),
data = AmesHousing
)
summary(anova_quality)
```

```
##                         Df Sum Sq Mean Sq F value Pr(>F)
## factor(Overall.Qual)     9  335.3   37.26   719.3 <2e-16 ***
```

```
## Residuals            2920  151.2    0.05
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Final OLS model (log SalePrice)

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.4.2
```

```
## Loading required package: carData
```

```
final_ols <- lm(
logSalePrice ~ Gr.Liv.Area + Total.Bsmt.SF + Garage.Area +
Lot.Area + TotRms.AbvGrd + Overall.Qual + Overall.Cond +
Year.Built + Year.Remod.Add +
Neighborhood + House.Style + Bldg.Type + Sale.Condition,
data = AmesHousing
)

summary(final_ols)
```

```
##
## Call:
## lm(formula = logSalePrice ~ Gr.Liv.Area + Total.Bsmt.SF + Garage.Area +
##     Lot.Area + TotRms.AbvGrd + Overall.Qual + Overall.Cond +
##     Year.Built + Year.Remod.Add + Neighborhood + House.Style +
##     Bldg.Type + Sale.Condition, data = AmesHousing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.21093 -0.06534  0.00287  0.07290  0.59587
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             3.073e+00  4.924e-01   6.240 5.01e-10 ***
## Gr.Liv.Area             2.566e-04  1.284e-05  19.993  < 2e-16 ***
## Total.Bsmt.SF           1.003e-04  1.010e-05   9.929  < 2e-16 ***
## Garage.Area             1.694e-04  1.705e-05   9.932  < 2e-16 ***
## Lot.Area                1.904e-06  3.907e-07   4.873 1.16e-06 ***
## TotRms.AbvGrd           2.313e-03  3.245e-03   0.713 0.476091
## Overall.Qual            7.820e-02  3.511e-03  22.270  < 2e-16 ***
## Overall.Cond            5.226e-02  2.972e-03  17.582  < 2e-16 ***
## Year.Built              2.957e-03  2.220e-04  13.320  < 2e-16 ***
## Year.Remod.Add          8.516e-04  1.967e-04   4.329 1.55e-05 ***
## NeighborhoodBlueste    -4.686e-02  5.362e-02  -0.874 0.382250
## NeighborhoodBrDale     -1.132e-01  4.143e-02  -2.733 0.006310 **
## NeighborhoodBrkSide    -5.945e-02  3.496e-02  -1.700 0.089192 .
## NeighborhoodClearCr     8.812e-02  3.723e-02   2.367 0.017990 *
## NeighborhoodCollgCr    -1.922e-02  3.042e-02  -0.632 0.527565
## NeighborhoodCrawfor     1.251e-01  3.350e-02   3.735 0.000192 ***
## NeighborhoodEdwards    -1.027e-01  3.204e-02  -3.206 0.001359 **
```

```
## NeighborhoodGilbert   -2.089e-02  3.169e-02  -0.659 0.509806
## NeighborhoodGreens     1.385e-01  5.836e-02   2.374 0.017674 *
## NeighborhoodGrnHill    4.967e-01  1.039e-01   4.782 1.83e-06 ***
## NeighborhoodIDOTRR    -1.804e-01  3.559e-02  -5.069 4.26e-07 ***
## NeighborhoodLandmrk   -1.333e-02  1.454e-01  -0.092 0.926975
## NeighborhoodMeadowV   -1.635e-01  3.844e-02  -4.252 2.18e-05 ***
## NeighborhoodMitchel   -3.355e-02  3.255e-02  -1.031 0.302768
## NeighborhoodNAmes     -2.862e-02  3.112e-02  -0.919 0.357922
## NeighborhoodNoRidge    8.012e-02  3.450e-02   2.322 0.020278 *
## NeighborhoodNPkVill   -2.426e-02  4.158e-02  -0.583 0.559666
## NeighborhoodNridgHt    1.185e-01  3.031e-02   3.910 9.44e-05 ***
## NeighborhoodNWAmes    -4.566e-02  3.225e-02  -1.416 0.156855
## NeighborhoodOldTown   -1.257e-01  3.382e-02  -3.715 0.000207 ***
## NeighborhoodSawyer    -3.933e-02  3.249e-02  -1.211 0.226182
## NeighborhoodSawyerW   -4.166e-02  3.187e-02  -1.307 0.191252
## NeighborhoodSomerst    3.989e-02  3.015e-02   1.323 0.185864
## NeighborhoodStoneBr    1.464e-01  3.408e-02   4.296 1.79e-05 ***
## NeighborhoodSWISU     -3.964e-02  3.841e-02  -1.032 0.302172
## NeighborhoodTimber     4.636e-02  3.359e-02   1.380 0.167656
## NeighborhoodVeenker    5.069e-02  4.111e-02   1.233 0.217687
## House.Style1.5Unf     -4.134e-02  3.425e-02  -1.207 0.227540
## House.Style1Story     -1.430e-03  1.124e-02  -0.127 0.898750
## House.Style2.5Fin     -4.948e-02  5.329e-02  -0.929 0.353193
## House.Style2.5Unf      1.140e-02  3.090e-02   0.369 0.712139
## House.Style2Story     -2.077e-02  1.123e-02  -1.850 0.064468 .
## House.StyleSFoyer      6.836e-02  2.002e-02   3.415 0.000647 ***
## House.StyleSLvl        1.398e-02  1.619e-02   0.863 0.388181
## Bldg.Type2fmCon       -1.093e-02  1.903e-02  -0.574 0.565840
## Bldg.TypeDuplex       -1.146e-01  1.601e-02  -7.158 1.03e-12 ***
## Bldg.TypeTwnhs        -1.398e-01  2.063e-02  -6.774 1.51e-11 ***
## Bldg.TypeTwnhsE       -5.524e-02  1.317e-02  -4.195 2.81e-05 ***
## Sale.ConditionAdjLand  1.650e-01  4.328e-02   3.813 0.000140 ***
## Sale.ConditionAlloca   1.118e-01  3.274e-02   3.414 0.000650 ***
## Sale.ConditionFamily   6.529e-02  2.341e-02   2.789 0.005321 **
## Sale.ConditionNormal   1.042e-01  1.091e-02   9.547  < 2e-16 ***
## Sale.ConditionPartial  1.030e-01  1.521e-02   6.774 1.51e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1414 on 2875 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.8817, Adjusted R-squared:  0.8795
## F-statistic:   412 on 52 and 2875 DF,  p-value: < 2.2e-16
```

```
vif(final_ols)
```

```
##                     GVIF Df GVIF^(1/(2*Df))
## Gr.Liv.Area     6.160741  1        2.482084
## Total.Bsmt.SF   2.899440  1        1.702774
## Garage.Area     1.968482  1        1.403026
## Lot.Area        1.387720  1        1.178015
## TotRms.AbvGrd   3.811926  1        1.952415
## Overall.Qual    3.591457  1        1.895114
## Overall.Cond    1.597326  1        1.263854
```

```
## Year.Built      6.595293  1        2.568130
## Year.Remod.Add  2.463572  1        1.569577
## Neighborhood   60.162881 27        1.078824
## House.Style     6.066848  7        1.137433
## Bldg.Type       4.685837  4        1.212965
## Sale.Condition  1.678885  5        1.053179
```

Examine statistical significance of predictors

```
#Extract coefficient table

ols_coef_table <- summary(final_ols)$coefficients

#View predictors with p-values

ols_coef_table[, "Pr(>|t|)"]
```

```
##           (Intercept)           Gr.Liv.Area         Total.Bsmt.SF
##          5.013441e-10          2.276614e-83          7.207317e-23
##           Garage.Area              Lot.Area         TotRms.AbvGrd
##          7.060855e-23          1.159281e-06          4.760910e-01
##          Overall.Qual          Overall.Cond            Year.Built
##         1.710593e-101          8.300485e-66          2.540449e-39
##        Year.Remod.Add   NeighborhoodBlueste    NeighborhoodBrDale
##          1.552468e-05          3.822500e-01          6.309840e-03
##   NeighborhoodBrkSide   NeighborhoodClearCr   NeighborhoodCollgCr
##          8.919172e-02          1.798969e-02          5.275652e-01
##   NeighborhoodCrawfor   NeighborhoodEdwards   NeighborhoodGilbert
##          1.915565e-04          1.359246e-03          5.098062e-01
##    NeighborhoodGreens   NeighborhoodGrnHill    NeighborhoodIDOTRR
##          1.767362e-02          1.826621e-06          4.256070e-07
##   NeighborhoodLandmrk   NeighborhoodMeadowV   NeighborhoodMitchel
##          9.269748e-01          2.184453e-05          3.027676e-01
##     NeighborhoodNAmes   NeighborhoodNoRidge   NeighborhoodNPkVill
##          3.579223e-01          2.027828e-02          5.596664e-01
##   NeighborhoodNridgHt    NeighborhoodNWAmes   NeighborhoodOldTown
##          9.443388e-05          1.568549e-01          2.068130e-04
##    NeighborhoodSawyer   NeighborhoodSawyerW   NeighborhoodSomerst
##          2.261817e-01          1.912521e-01          1.858638e-01
##   NeighborhoodStoneBr     NeighborhoodSWISU    NeighborhoodTimber
##          1.794773e-05          3.021721e-01          1.676556e-01
##   NeighborhoodVeenker     House.Style1.5Unf     House.Style1Story
##          2.176874e-01          2.275398e-01          8.987503e-01
##     House.Style2.5Fin     House.Style2.5Unf     House.Style2Story
##          3.531929e-01          7.121387e-01          6.446823e-02
##     House.StyleSFoyer       House.StyleSLvl       Bldg.Type2fmCon
##          6.468706e-04          3.881809e-01          5.658402e-01
##       Bldg.TypeDuplex        Bldg.TypeTwnhs       Bldg.TypeTwnhsE
##          1.033526e-12          1.512436e-11          2.807236e-05
## Sale.ConditionAdjLand  Sale.ConditionAlloca  Sale.ConditionFamily
##          1.404060e-04          6.496750e-04          5.320815e-03
##  Sale.ConditionNormal Sale.ConditionPartial
##          2.751729e-21          1.508283e-11
```

```r
#Identify predictors significant at 5% level

significant_predictors <- rownames(ols_coef_table)[
ols_coef_table[, "Pr(>|t|)"] < 0.05
]

significant_predictors
```

```
##  [1] "(Intercept)"           "Gr.Liv.Area"           "Total.Bsmt.SF"
##  [4] "Garage.Area"           "Lot.Area"              "Overall.Qual"
##  [7] "Overall.Cond"          "Year.Built"            "Year.Remod.Add"
## [10] "NeighborhoodBrDale"    "NeighborhoodClearCr"   "NeighborhoodCrawfor"
## [13] "NeighborhoodEdwards"   "NeighborhoodGreens"    "NeighborhoodGrnHill"
## [16] "NeighborhoodIDOTRR"    "NeighborhoodMeadowV"   "NeighborhoodNoRidge"
## [19] "NeighborhoodNridgHt"   "NeighborhoodOldTown"   "NeighborhoodStoneBr"
## [22] "House.StyleSFoyer"     "Bldg.TypeDuplex"       "Bldg.TypeTwnhs"
## [25] "Bldg.TypeTwnhsE"       "Sale.ConditionAdjLand" "Sale.ConditionAlloca"
## [28] "Sale.ConditionFamily"  "Sale.ConditionNormal"  "Sale.ConditionPartial"
```

Reduced OLS model using statistically significant predictors

```r
final_ols_reduced <- lm(
logSalePrice ~ Gr.Liv.Area + Total.Bsmt.SF + Garage.Area +
Lot.Area + Overall.Qual + Overall.Cond +
Year.Built + Year.Remod.Add +
Neighborhood + Bldg.Type + Sale.Condition,
data = AmesHousing
)

summary(final_ols_reduced)
```

```
##
## Call:
## lm(formula = logSalePrice ~ Gr.Liv.Area + Total.Bsmt.SF + Garage.Area +
##     Lot.Area + Overall.Qual + Overall.Cond + Year.Built + Year.Remod.Add +
##     Neighborhood + Bldg.Type + Sale.Condition, data = AmesHousing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.22955 -0.06514  0.00271  0.07317  0.59013
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.976e+00  4.840e-01   6.149 8.87e-10 ***
## Gr.Liv.Area             2.464e-04  7.715e-06  31.941  < 2e-16 ***
## Total.Bsmt.SF           1.080e-04  8.120e-06  13.303  < 2e-16 ***
## Garage.Area             1.753e-04  1.698e-05  10.319  < 2e-16 ***
## Lot.Area                1.957e-06  3.901e-07   5.016 5.61e-07 ***
## Overall.Qual            7.885e-02  3.474e-03  22.696  < 2e-16 ***
## Overall.Cond            5.257e-02  2.971e-03  17.697  < 2e-16 ***
## Year.Built              3.011e-03  2.173e-04  13.858  < 2e-16 ***
## Year.Remod.Add          8.519e-04  1.963e-04   4.339 1.48e-05 ***
```

```
## NeighborhoodBlueste   -5.962e-02  5.353e-02  -1.114 0.265450
## NeighborhoodBrDale    -1.269e-01  4.103e-02  -3.092 0.002009 **
## NeighborhoodBrkSide   -5.852e-02  3.455e-02  -1.694 0.090411 .
## NeighborhoodClearCr    9.009e-02  3.707e-02   2.430 0.015148 *
## NeighborhoodCollgCr   -2.471e-02  3.022e-02  -0.817 0.413724
## NeighborhoodCrawfor    1.263e-01  3.334e-02   3.789 0.000154 ***
## NeighborhoodEdwards   -9.672e-02  3.183e-02  -3.038 0.002400 **
## NeighborhoodGilbert   -2.751e-02  3.141e-02  -0.876 0.381084
## NeighborhoodGreens     1.331e-01  5.828e-02   2.283 0.022476 *
## NeighborhoodGrnHill    4.998e-01  1.041e-01   4.802 1.65e-06 ***
## NeighborhoodIDOTRR    -1.783e-01  3.534e-02  -5.046 4.80e-07 ***
## NeighborhoodLandmrk   -2.917e-02  1.458e-01  -0.200 0.841403
## NeighborhoodMeadowV   -1.543e-01  3.803e-02  -4.058 5.07e-05 ***
## NeighborhoodMitchel   -2.545e-02  3.232e-02  -0.787 0.431122
## NeighborhoodNAmes     -2.684e-02  3.099e-02  -0.866 0.386578
## NeighborhoodNoRidge    7.498e-02  3.416e-02   2.195 0.028258 *
## NeighborhoodNPkVill   -3.209e-02  4.153e-02  -0.773 0.439784
## NeighborhoodNridgHt    1.140e-01  3.031e-02   3.762 0.000172 ***
## NeighborhoodNWAmes    -4.468e-02  3.213e-02  -1.391 0.164466
## NeighborhoodOldTown   -1.236e-01  3.357e-02  -3.682 0.000236 ***
## NeighborhoodSawyer    -3.306e-02  3.236e-02  -1.022 0.306988
## NeighborhoodSawyerW   -4.553e-02  3.166e-02  -1.438 0.150577
## NeighborhoodSomerst    3.088e-02  2.988e-02   1.034 0.301423
## NeighborhoodStoneBr    1.437e-01  3.407e-02   4.217 2.55e-05 ***
## NeighborhoodSWISU     -3.835e-02  3.794e-02  -1.011 0.312137
## NeighborhoodTimber     4.466e-02  3.356e-02   1.331 0.183385
## NeighborhoodVeenker    5.201e-02  4.083e-02   1.274 0.202822
## Bldg.Type2fmCon       -6.967e-03  1.894e-02  -0.368 0.713051
## Bldg.TypeDuplex       -1.015e-01  1.524e-02  -6.657 3.34e-11 ***
## Bldg.TypeTwnhs        -1.425e-01  2.035e-02  -7.005 3.06e-12 ***
## Bldg.TypeTwnhsE       -5.760e-02  1.273e-02  -4.525 6.28e-06 ***
## Sale.ConditionAdjLand  1.737e-01  4.331e-02   4.010 6.22e-05 ***
## Sale.ConditionAlloca   1.229e-01  3.267e-02   3.762 0.000172 ***
## Sale.ConditionFamily   6.758e-02  2.345e-02   2.882 0.003981 **
## Sale.ConditionNormal   1.030e-01  1.093e-02   9.422  < 2e-16 ***
## Sale.ConditionPartial  1.016e-01  1.522e-02   6.676 2.94e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1419 on 2883 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.8806, Adjusted R-squared:  0.8787
## F-statistic:   483 on 44 and 2883 DF,  p-value: < 2.2e-16
```

```
vif(final_ols_reduced)
```

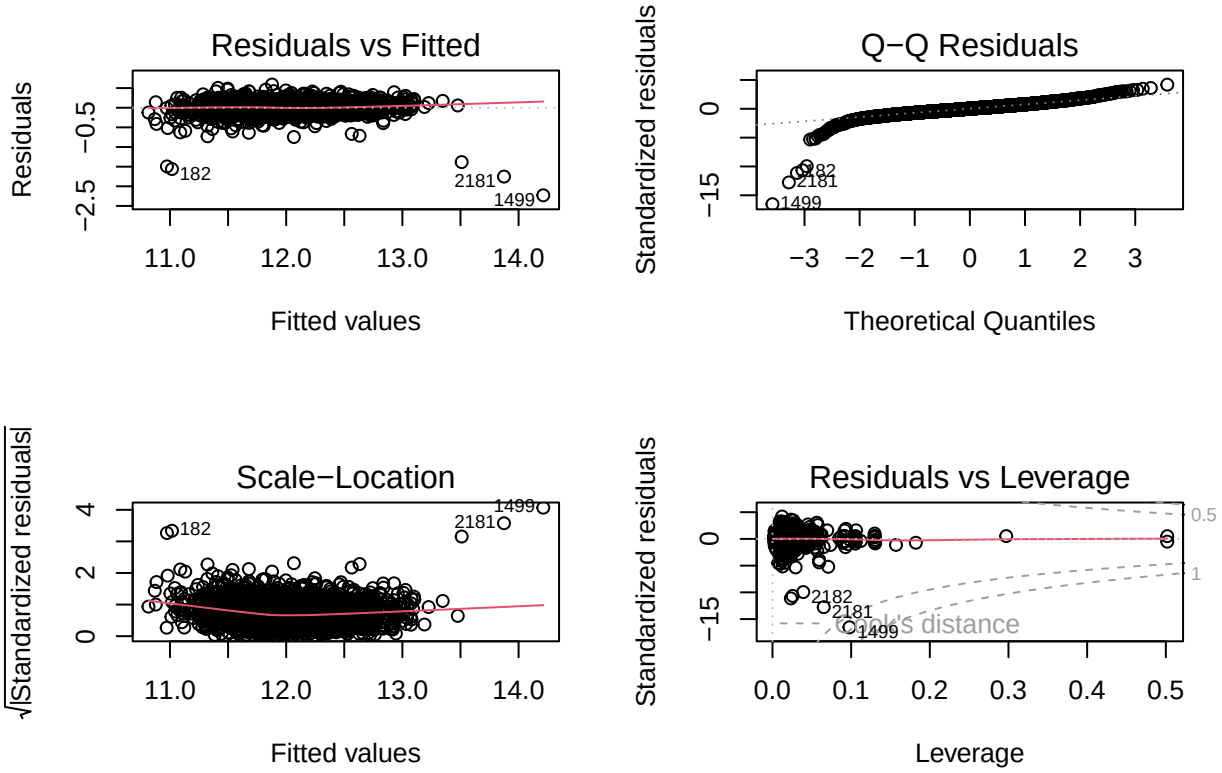
```
##                     GVIF Df GVIF^(1/(2*Df))
## Gr.Liv.Area     2.211126  1        1.486986
## Total.Bsmt.SF   1.861159  1        1.364243
## Garage.Area     1.939077  1        1.392507
## Lot.Area        1.374317  1        1.172313
## Overall.Qual    3.491716  1        1.868613
## Overall.Cond    1.584870  1        1.258916
## Year.Built      6.273775  1        2.504750
```

```
## Year.Remod.Add  2.436807  1        1.561028
## Neighborhood   39.777240 27        1.070589
## Bldg.Type       3.836062  4        1.183003
## Sale.Condition  1.630585  5        1.050109
```

OLS diagnostics (final_ols_reduced)

```
par(mfrow = c(2, 2))
plot(final_ols_reduced)
```

```
## Warning: not plotting observations with leverage one:
##   2787
```




```
par(mfrow = c(1, 1))
```

Construct binary response: AboveExpected

```
model_data <- model.frame(final_ols_reduced)

model_data$SalePrice_obs <- exp(model_data$logSalePrice)
model_data$FittedPrice <- exp(fitted(final_ols_reduced))

model_data$AboveExpected <- ifelse(
model_data$SalePrice_obs > model_data$FittedPrice, 1, 0
```

```
)

table(model_data$AboveExpected)
```

```
##
##    0    1
## 1428 1500
```

```
prop.table(table(model_data$AboveExpected))
```

```
##
##         0         1
## 0.4877049 0.5122951
```

Logistic regression model

```
final_logit <- glm(
AboveExpected ~ Gr.Liv.Area + Total.Bsmt.SF + Garage.Area +
Lot.Area + Overall.Qual + Overall.Cond +
Year.Built + Year.Remod.Add +
Neighborhood + Bldg.Type + Sale.Condition,
family = binomial,
data = model_data
)

summary(final_logit)
```

```
##
## Call:
## glm(formula = AboveExpected ~ Gr.Liv.Area + Total.Bsmt.SF + Garage.Area +
##     Lot.Area + Overall.Qual + Overall.Cond + Year.Built + Year.Remod.Add +
##     Neighborhood + Bldg.Type + Sale.Condition, family = binomial,
##     data = model_data)
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -1.026e+01  7.057e+00  -1.454 0.145933
## Gr.Liv.Area            4.618e-04  1.160e-04   3.980 6.89e-05 ***
## Total.Bsmt.SF          5.340e-04  1.232e-04   4.333 1.47e-05 ***
## Garage.Area            1.360e-04  2.470e-04   0.551 0.581924
## Lot.Area               1.786e-05  8.588e-06   2.080 0.037565 *
## Overall.Qual          -1.422e-01  5.056e-02  -2.812 0.004929 **
## Overall.Cond          -7.730e-02  4.332e-02  -1.784 0.074366 .
## Year.Built             7.715e-03  3.188e-03   2.420 0.015536 *
## Year.Remod.Add        -2.944e-03  2.835e-03  -1.039 0.298996
## NeighborhoodBlueste    3.942e-01  7.746e-01   0.509 0.610794
## NeighborhoodBrDale     8.634e-01  5.892e-01   1.465 0.142831
## NeighborhoodBrkSide    1.175e+00  5.028e-01   2.337 0.019457 *
## NeighborhoodClearCr    7.612e-01  5.418e-01   1.405 0.160049
## NeighborhoodCollgCr    4.150e-01  4.380e-01   0.948 0.343300
## NeighborhoodCrawfor    9.011e-01  4.857e-01   1.855 0.063573 .
```

```
## NeighborhoodEdwards    1.330e+00  4.661e-01   2.854 0.004319 **
## NeighborhoodGilbert    4.433e-01  4.556e-01   0.973 0.330499
## NeighborhoodGreens     7.321e-01  8.322e-01   0.880 0.379008
## NeighborhoodGrnHill    6.877e-01  1.512e+00   0.455 0.649148
## NeighborhoodIDOTRR     1.800e+00  5.179e-01   3.476 0.000509 ***
## NeighborhoodLandmrk   -1.118e+01  1.970e+02  -0.057 0.954752
## NeighborhoodMeadowV    9.267e-01  5.484e-01   1.690 0.091053 .
## NeighborhoodMitchel    7.332e-01  4.691e-01   1.563 0.118049
## NeighborhoodNAmes      9.409e-01  4.506e-01   2.088 0.036792 *
## NeighborhoodNoRidge    3.747e-01  4.940e-01   0.759 0.448108
## NeighborhoodNPkVill    5.285e-01  5.951e-01   0.888 0.374446
## NeighborhoodNridgHt    3.143e-01  4.386e-01   0.717 0.473579
## NeighborhoodNWAmes     9.282e-01  4.658e-01   1.993 0.046267 *
## NeighborhoodOldTown    1.515e+00  4.901e-01   3.091 0.001995 **
## NeighborhoodSawyer     1.213e+00  4.705e-01   2.578 0.009942 **
## NeighborhoodSawyerW    8.109e-01  4.589e-01   1.767 0.077212 .
## NeighborhoodSomerst    4.700e-01  4.320e-01   1.088 0.276574
## NeighborhoodStoneBr    3.722e-01  4.928e-01   0.755 0.450087
## NeighborhoodSWISU      9.055e-01  5.514e-01   1.642 0.100572
## NeighborhoodTimber     4.527e-01  4.848e-01   0.934 0.350401
## NeighborhoodVeenker    7.659e-01  5.945e-01   1.288 0.197635
## Bldg.Type2fmCon       -1.817e-01  2.746e-01  -0.662 0.508103
## Bldg.TypeDuplex       -4.827e-01  2.217e-01  -2.177 0.029510 *
## Bldg.TypeTwnhs         6.685e-01  2.966e-01   2.254 0.024204 *
## Bldg.TypeTwnhsE        3.428e-01  1.860e-01   1.843 0.065292 .
## Sale.ConditionAdjLand -4.963e-01  6.281e-01  -0.790 0.429375
## Sale.ConditionAlloca  -3.544e-01  4.816e-01  -0.736 0.461757
## Sale.ConditionFamily  -2.148e-01  3.364e-01  -0.639 0.523126
## Sale.ConditionNormal  -1.388e-01  1.585e-01  -0.876 0.381232
## Sale.ConditionPartial -8.123e-02  2.196e-01  -0.370 0.711414
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4057.3  on 2927  degrees of freedom
## Residual deviance: 3939.8  on 2883  degrees of freedom
## AIC: 4029.8
##
## Number of Fisher Scoring iterations: 10
```

Odds ratios

```
or_table <- exp(
cbind(
OR = coef(final_logit),
confint.default(final_logit)
)
)

or_table
```

```
##                                 OR         2.5 %        97.5 %
```

```
## (Intercept)           3.498453e-05  3.445952e-11  3.551754e+01
## Gr.Liv.Area           1.000462e+00  1.000234e+00  1.000689e+00
## Total.Bsmt.SF         1.000534e+00  1.000293e+00  1.000776e+00
## Garage.Area           1.000136e+00  9.996520e-01  1.000620e+00
## Lot.Area              1.000018e+00  1.000001e+00  1.000035e+00
## Overall.Qual          8.674850e-01  7.856422e-01  9.578536e-01
## Overall.Cond          9.256095e-01  8.502596e-01  1.007637e+00
## Year.Built            1.007744e+00  1.001467e+00  1.014061e+00
## Year.Remod.Add        9.970599e-01  9.915351e-01  1.002616e+00
## NeighborhoodBlueste   1.483213e+00  3.250029e-01  6.768923e+00
## NeighborhoodBrDale    2.371182e+00  7.471833e-01  7.524932e+00
## NeighborhoodBrkSide   3.237780e+00  1.208527e+00  8.674378e+00
## NeighborhoodClearCr   2.140851e+00  7.402702e-01  6.191313e+00
## NeighborhoodCollgCr   1.514446e+00  6.418780e-01  3.573180e+00
## NeighborhoodCrawfor   2.462331e+00  9.503691e-01  6.379705e+00
## NeighborhoodEdwards   3.782156e+00  1.516932e+00  9.430026e+00
## NeighborhoodGilbert   1.557899e+00  6.378829e-01  3.804853e+00
## NeighborhoodGreens    2.079490e+00  4.069790e-01  1.062531e+01
## NeighborhoodGrnHill   1.989190e+00  1.027876e-01  3.849565e+01
## NeighborhoodIDOTRR    6.050596e+00  2.192594e+00  1.669698e+01
## NeighborhoodLandmrk   1.400603e-05 3.066354e-173 6.397466e+162
## NeighborhoodMeadowV   2.526273e+00  8.623259e-01  7.400981e+00
## NeighborhoodMitchel   2.081720e+00  8.301035e-01  5.220502e+00
## NeighborhoodNAmes     2.562323e+00  1.059424e+00  6.197235e+00
## NeighborhoodNoRidge   1.454607e+00  5.523967e-01  3.830364e+00
## NeighborhoodNPkVill   1.696421e+00  5.284634e-01  5.445683e+00
## NeighborhoodNridgHt   1.369350e+00  5.796574e-01  3.234874e+00
## NeighborhoodNWAmes    2.530045e+00  1.015483e+00  6.303532e+00
## NeighborhoodOldTown   4.548600e+00  1.740689e+00  1.188596e+01
## NeighborhoodSawyer    3.363284e+00  1.337395e+00  8.457991e+00
## NeighborhoodSawyerW   2.249964e+00  9.152970e-01  5.530815e+00
## NeighborhoodSomerst   1.599967e+00  6.861734e-01  3.730682e+00
## NeighborhoodStoneBr   1.450896e+00  5.523094e-01  3.811447e+00
## NeighborhoodSWISU     2.473125e+00  8.392193e-01  7.288140e+00
## NeighborhoodTimber    1.572504e+00  6.080901e-01  4.066452e+00
## NeighborhoodVeenker   2.151014e+00  6.707823e-01  6.897710e+00
## Bldg.Type2fmCon       8.338131e-01  4.867505e-01  1.428338e+00
## Bldg.TypeDuplex       6.171383e-01  3.996034e-01  9.530942e-01
## Bldg.TypeTwnhs        1.951239e+00  1.091083e+00  3.489498e+00
## Bldg.TypeTwnhsE       1.408865e+00  9.785291e-01  2.028452e+00
## Sale.ConditionAdjLand 6.087512e-01  1.777538e-01  2.084782e+00
## Sale.ConditionAlloca  7.015831e-01  2.730014e-01  1.802991e+00
## Sale.ConditionFamily  8.066858e-01  4.171919e-01  1.559815e+00
## Sale.ConditionNormal  8.704044e-01  6.379632e-01  1.187535e+00
## Sale.ConditionPartial 9.219771e-01  5.995345e-01  1.417836e+00
```

Odds ratio visualization (forest plot)

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.3
```

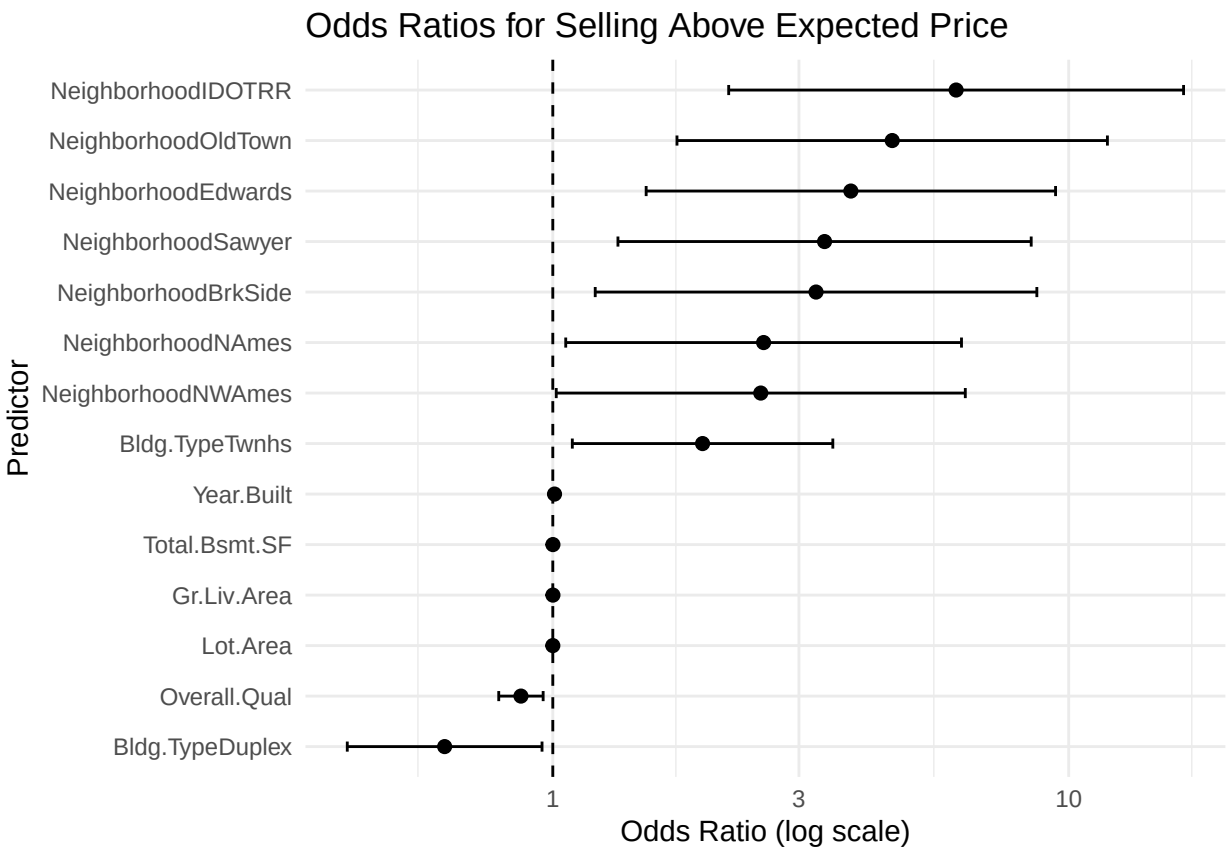
```r
# Create odds ratio data frame
or_df <- data.frame(
  term  = rownames(or_table),
  OR    = or_table[, 1],
  lower = or_table[, 2],
  upper = or_table[, 3]
)

# Remove intercept
or_df <- or_df[or_df$term != "(Intercept)", ]

# Remove non-finite or extreme confidence intervals
or_df <- or_df[
  is.finite(or_df$OR) &
  is.finite(or_df$lower) &
  is.finite(or_df$upper) &
  or_df$lower > 0 &
  or_df$upper < 100,
]

# OPTIONAL but recommended: keep only statistically significant predictors
or_df <- or_df[!(or_df$lower <= 1 & or_df$upper >= 1), ]

# Forest plot of odds ratios
ggplot(or_df, aes(x = OR, y = reorder(term, OR))) +
  geom_point(size = 2) +
  geom_errorbarh(aes(xmin = lower, xmax = upper), height = 0.2) +
  geom_vline(xintercept = 1, linetype = "dashed") +
  scale_x_log10() +
  labs(
    title = "Odds Ratios for Selling Above Expected Price",
    x = "Odds Ratio (log scale)",
    y = "Predictor"
  ) +
  theme_minimal()
```



Bootstrap: OLS coefficient (Gr.Liv.Area)

```
library(boot)
```

```
##
## Attaching package: 'boot'

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

```
set.seed(123)
B <- 1000

ols_data <- model.frame(final_ols)
X <- model.matrix(final_ols, data = ols_data)
y <- ols_data$logSalePrice

boot_ols <- function(data, indices) {
X_b <- X[indices, , drop = FALSE]
y_b <- y[indices]
fit <- lm.fit(x = X_b, y = y_b)
fit$coefficients
}
```

```r
boot_ols_results <- boot(
data = ols_data,
statistic = boot_ols,
R = B
)

idx_grliv <- which(names(coef(final_ols)) == "Gr.Liv.Area")

boot.ci(
boot.out = boot_ols_results,
type = "perc",
index = idx_grliv
)
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_ols_results, type = "perc", index = idx_grliv)
##
## Intervals :
## Level     Percentile
## 95%   ( 0.0002,  0.0003 )
## Calculations and Intervals on Original Scale
```

Bootstrap: Logistic regression

```r
B <- 1000

# Fit final logistic regression model
final_logit <- glm(
  AboveExpected ~ Gr.Liv.Area + Total.Bsmt.SF + Garage.Area +
    Lot.Area + Overall.Qual + Overall.Cond +
    Year.Built + Year.Remod.Add +
    Neighborhood + Bldg.Type + Sale.Condition,
  family = binomial,
  data = model_data
)

# Store full coefficient names to enforce fixed length
logit_coef_names <- names(coef(final_logit))

# Define bootstrap statistic function
boot_logit <- function(data, indices) {

  # Resample rows
  d <- data[indices, ]

  # Refit logistic model
  fit <- glm(
    AboveExpected ~ Gr.Liv.Area + Total.Bsmt.SF + Garage.Area +
      Lot.Area + Overall.Qual + Overall.Cond +
      Year.Built + Year.Remod.Add +
```

```
      Neighborhood + Bldg.Type + Sale.Condition,
    family = binomial,
    data = d
  )

  # Initialize fixed-length coefficient vector
  beta_full <- rep(NA, length(logit_coef_names))
  names(beta_full) <- logit_coef_names

  # Extract coefficients
  beta_hat <- coef(fit)

  # Align by name
  beta_full[names(beta_hat)] <- beta_hat

  # Return numeric vector
  return(beta_full)
}

# Run bootstrap
boot_logit_results <- boot(
  data = model_data,
  statistic = boot_logit,
  R = B
)

# Sanity checks
dim(boot_logit_results$t)
```

```
## [1] 1000   45
```

```
length(boot_logit_results$t0)
```

```
## [1] 45
```

Percentile Bootstrap CI for Overall.Qual

```
#Locate coefficient index

idx_qual <- which(logit_coef_names == "Overall.Qual")

#Percentile confidence interval (log-odds scale)

boot_ci_qual <- boot.ci(
boot.out = boot_logit_results,
type = "perc",
index = idx_qual
)

boot_ci_qual
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
```

```
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_logit_results, type = "perc", index = idx_qual)
##
## Intervals :
## Level     Percentile
## 95%   (-0.2469, -0.0415 )
## Calculations and Intervals on Original Scale
```

```
#Convert to odds ratio scale

#Point estimate

or_hat <- exp(coef(final_logit)["Overall.Qual"])

#Percentile CI on odds ratio scale

ci_log <- boot_ci_qual$percent[4:5]
exp(c(or_hat, ci_log))
```

```
## Overall.Qual
##    2.3809153    0.7812099    0.9593826
```

Predicting whether a home sells at a premium or a discount

```
model_data <- model.frame(final_ols_reduced)

model_data$SalePrice_obs <- exp(model_data$logSalePrice)
model_data$ExpectedPrice <- exp(fitted(final_ols_reduced))

model_data$AboveExpected <- ifelse(
model_data$SalePrice_obs > model_data$ExpectedPrice, 1, 0
)

#Fit final logistic regression classifier
final_logit <- glm(
AboveExpected ~ Gr.Liv.Area + Total.Bsmt.SF + Garage.Area +
Lot.Area + Overall.Qual + Overall.Cond +
Year.Built + Year.Remod.Add +
Neighborhood + Bldg.Type + Sale.Condition,
family = binomial,
data = model_data
)

summary(final_logit)
```

```
##
## Call:
## glm(formula = AboveExpected ~ Gr.Liv.Area + Total.Bsmt.SF + Garage.Area +
##     Lot.Area + Overall.Qual + Overall.Cond + Year.Built + Year.Remod.Add +
##     Neighborhood + Bldg.Type + Sale.Condition, family = binomial,
##     data = model_data)
```

```
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -1.026e+01  7.057e+00  -1.454 0.145933
## Gr.Liv.Area            4.618e-04  1.160e-04   3.980 6.89e-05 ***
## Total.Bsmt.SF          5.340e-04  1.232e-04   4.333 1.47e-05 ***
## Garage.Area            1.360e-04  2.470e-04   0.551 0.581924
## Lot.Area               1.786e-05  8.588e-06   2.080 0.037565 *
## Overall.Qual          -1.422e-01  5.056e-02  -2.812 0.004929 **
## Overall.Cond          -7.730e-02  4.332e-02  -1.784 0.074366 .
## Year.Built             7.715e-03  3.188e-03   2.420 0.015536 *
## Year.Remod.Add        -2.944e-03  2.835e-03  -1.039 0.298996
## NeighborhoodBlueste    3.942e-01  7.746e-01   0.509 0.610794
## NeighborhoodBrDale     8.634e-01  5.892e-01   1.465 0.142831
## NeighborhoodBrkSide    1.175e+00  5.028e-01   2.337 0.019457 *
## NeighborhoodClearCr    7.612e-01  5.418e-01   1.405 0.160049
## NeighborhoodCollgCr    4.150e-01  4.380e-01   0.948 0.343300
## NeighborhoodCrawfor    9.011e-01  4.857e-01   1.855 0.063573 .
## NeighborhoodEdwards    1.330e+00  4.661e-01   2.854 0.004319 **
## NeighborhoodGilbert    4.433e-01  4.556e-01   0.973 0.330499
## NeighborhoodGreens     7.321e-01  8.322e-01   0.880 0.379008
## NeighborhoodGrnHill    6.877e-01  1.512e+00   0.455 0.649148
## NeighborhoodIDOTRR     1.800e+00  5.179e-01   3.476 0.000509 ***
## NeighborhoodLandmrk   -1.118e+01  1.970e+02  -0.057 0.954752
## NeighborhoodMeadowV    9.267e-01  5.484e-01   1.690 0.091053 .
## NeighborhoodMitchel    7.332e-01  4.691e-01   1.563 0.118049
## NeighborhoodNAmes      9.409e-01  4.506e-01   2.088 0.036792 *
## NeighborhoodNoRidge    3.747e-01  4.940e-01   0.759 0.448108
## NeighborhoodNPkVill    5.285e-01  5.951e-01   0.888 0.374446
## NeighborhoodNridgHt    3.143e-01  4.386e-01   0.717 0.473579
## NeighborhoodNWAmes     9.282e-01  4.658e-01   1.993 0.046267 *
## NeighborhoodOldTown    1.515e+00  4.901e-01   3.091 0.001995 **
## NeighborhoodSawyer     1.213e+00  4.705e-01   2.578 0.009942 **
## NeighborhoodSawyerW    8.109e-01  4.589e-01   1.767 0.077212 .
## NeighborhoodSomerst    4.700e-01  4.320e-01   1.088 0.276574
## NeighborhoodStoneBr    3.722e-01  4.928e-01   0.755 0.450087
## NeighborhoodSWISU      9.055e-01  5.514e-01   1.642 0.100572
## NeighborhoodTimber     4.527e-01  4.848e-01   0.934 0.350401
## NeighborhoodVeenker    7.659e-01  5.945e-01   1.288 0.197635
## Bldg.Type2fmCon       -1.817e-01  2.746e-01  -0.662 0.508103
## Bldg.TypeDuplex       -4.827e-01  2.217e-01  -2.177 0.029510 *
## Bldg.TypeTwnhs         6.685e-01  2.966e-01   2.254 0.024204 *
## Bldg.TypeTwnhsE        3.428e-01  1.860e-01   1.843 0.065292 .
## Sale.ConditionAdjLand -4.963e-01  6.281e-01  -0.790 0.429375
## Sale.ConditionAlloca  -3.544e-01  4.816e-01  -0.736 0.461757
## Sale.ConditionFamily  -2.148e-01  3.364e-01  -0.639 0.523126
## Sale.ConditionNormal  -1.388e-01  1.585e-01  -0.876 0.381232
## Sale.ConditionPartial -8.123e-02  2.196e-01  -0.370 0.711414
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4057.3  on 2927  degrees of freedom
```

```
## Residual deviance: 3939.8  on 2883  degrees of freedom
## AIC: 4029.8
##
## Number of Fisher Scoring iterations: 10
```

```
#Generate predicted probabilities
model_data$prob_premium <- predict(
final_logit,
type = "response"
)

#Classify using a 0.50 probability cutoff
model_data$predicted_class <- ifelse(
model_data$prob_premium >= 0.5, 1, 0
)
```

Evaluate classification performance

```
#Confusion matrix

confusion_matrix <- table(
Actual = model_data$AboveExpected,
Predicted = model_data$predicted_class
)
confusion_matrix
```

```
##       Predicted
## Actual   0   1
##      0 811 617
##      1 557 943
```

```
#Accuracy

accuracy <- mean(model_data$AboveExpected == model_data$predicted_class)
accuracy
```

```
## [1] 0.5990437
```

```
#Sensitivity (true positive rate: premium correctly identified)

sensitivity <- confusion_matrix["1","1"] / sum(confusion_matrix["1",])
sensitivity
```

```
## [1] 0.6286667
```

```
#Specificity (true negative rate: discount correctly identified)

specificity <- confusion_matrix["0","0"] / sum(confusion_matrix["0",])
specificity
```

```
## [1] 0.5679272
```

ROC curve and AUC (model reliability)

```
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 4.4.3
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
roc_obj <- roc(
response = model_data$AboveExpected,
predictor = model_data$prob_premium
)
```

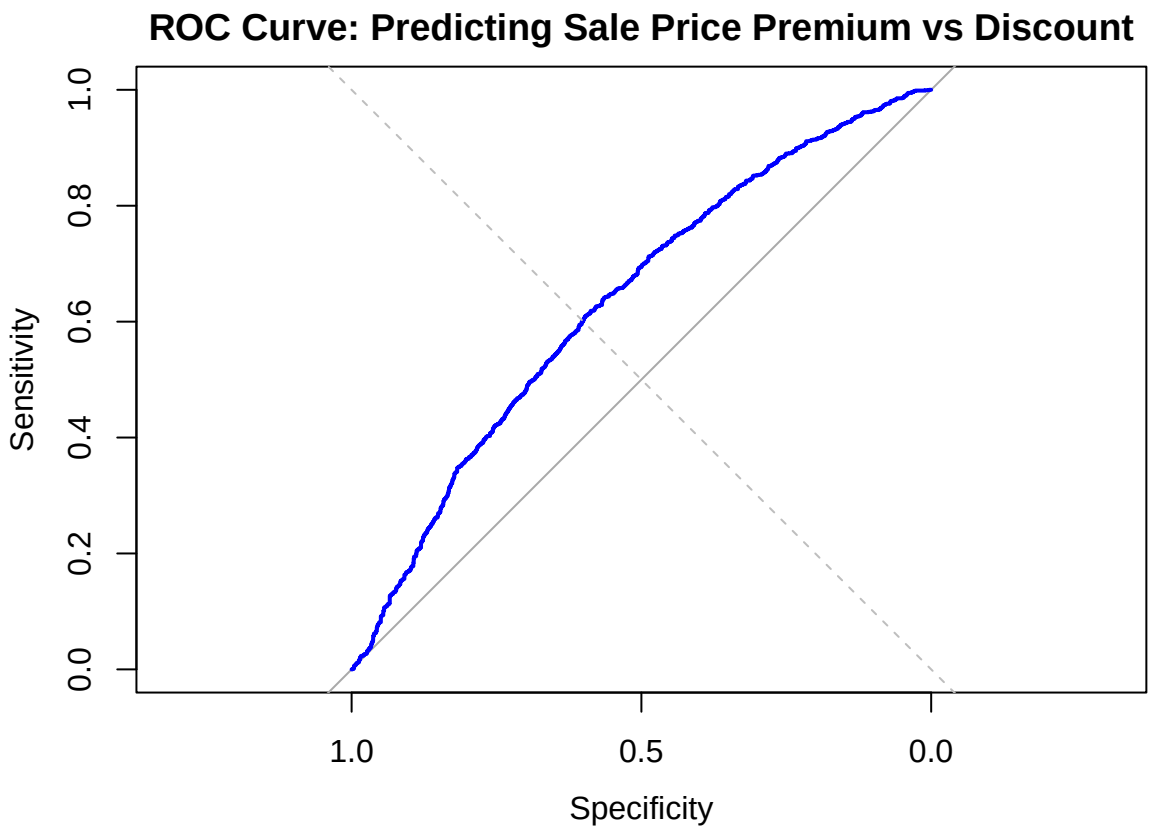
```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
auc_value <- auc(roc_obj)
auc_value
```

```
## Area under the curve: 0.6336
```

```
plot(
roc_obj,
main = "ROC Curve: Predicting Sale Price Premium vs Discount",
col = "blue",
lwd = 2
)
abline(a = 0, b = 1, lty = 2, col = "gray")
```



Example: Interpretable prediction for a single home

```
example_home <- model_data[1, ]

example_prob <- predict(
final_logit,
newdata = example_home,
type = "response"
)

example_prob
```

```
##         1
## 0.6530183
```

```
if (example_prob >= 0.5) {
"Predicted to sell at a premium"
} else {
"Predicted to sell at a discount"
}
```

```
## [1] "Predicted to sell at a premium"
```