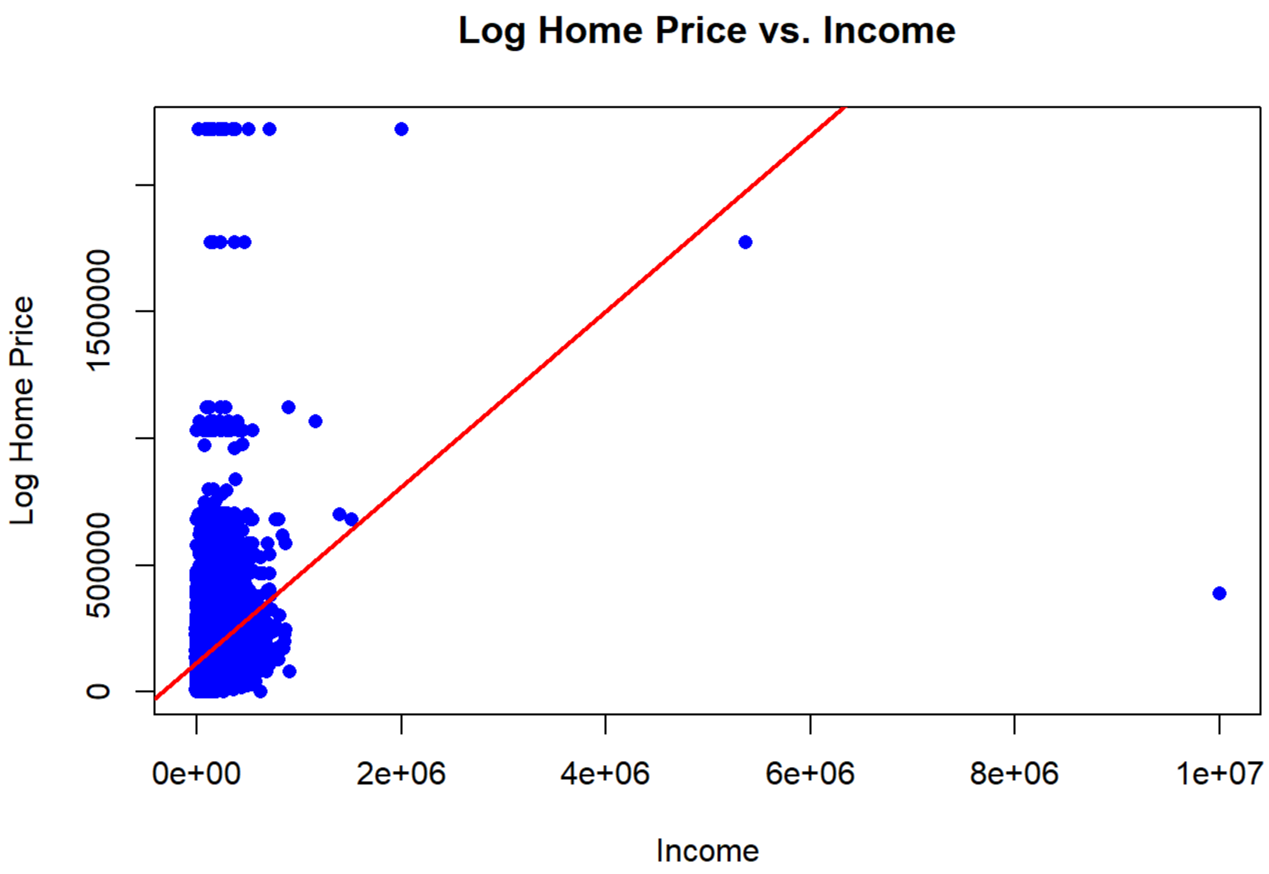
Econ 573: Problem Set 2 - Part 3

```
Homes = read.csv("C:/Users/mateo/OneDrive - University of North Carolina at Chapel Hill/Courses/
Spring 2025/Econ 573/homes2004.csv")
```

Exercise 1: Plot some relationships and tell a story.

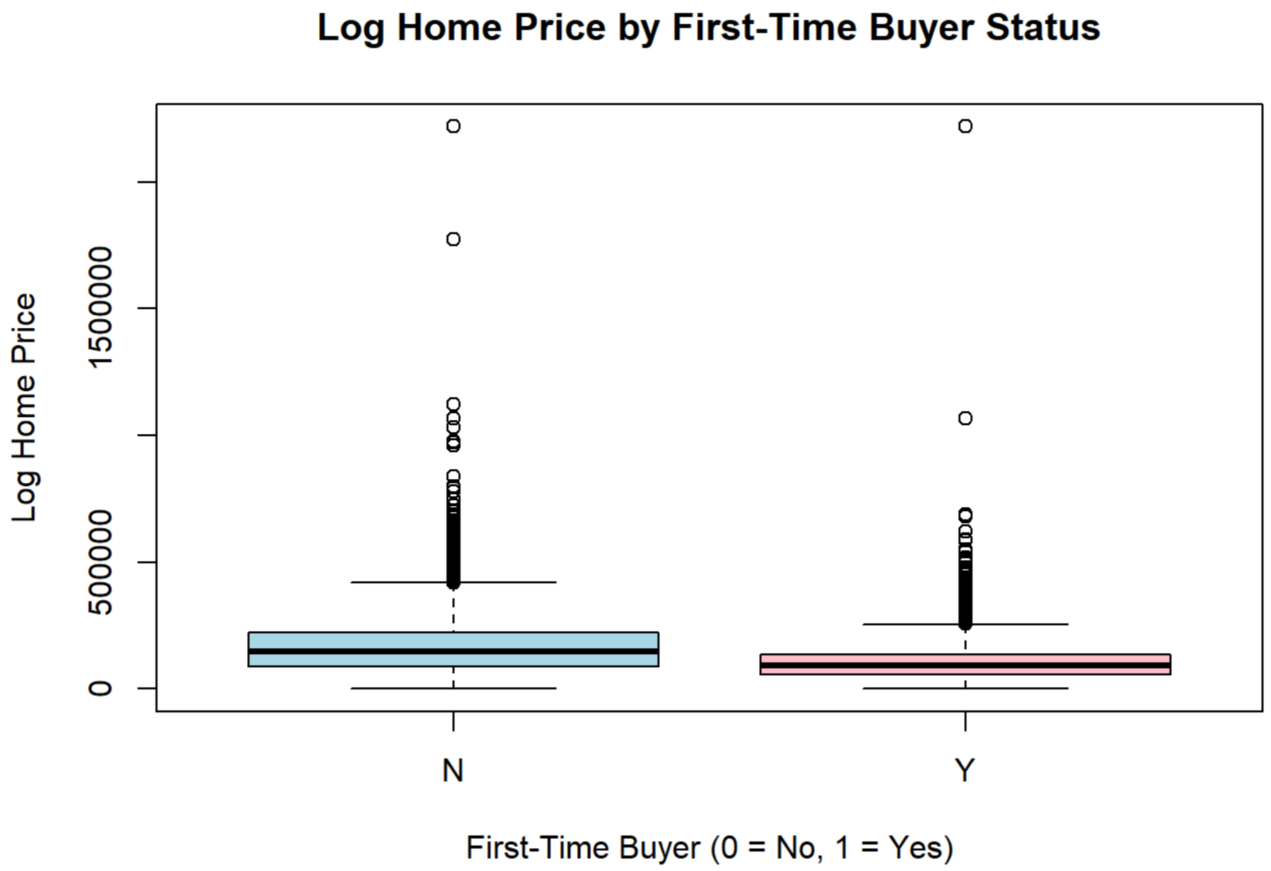
```
plot(Homes$ZINC2, Homes$LPRICE,
     main = "Log Home Price vs. Income",
     xlab = "Income",
     ylab = "Log Home Price",
     pch = 16, col = "blue")

abline(lm(LPRICE ~ ZINC2, data = Homes), col = "red", lwd = 2)
```



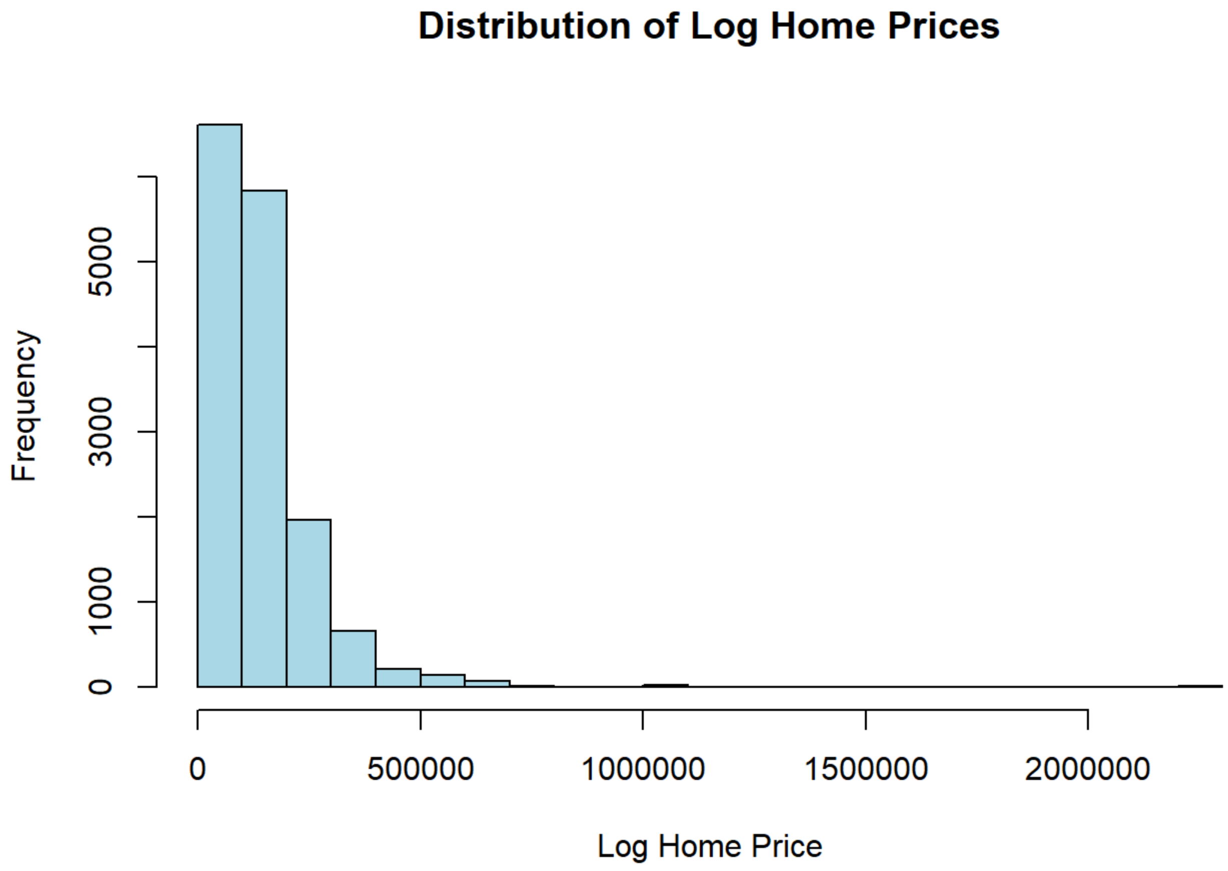
The scatterplot of log home price (LPRICE) vs. income (ZINC2) shows a positive relationship, indicating that higher-income buyers tend to purchase more expensive homes. However, we can see tha the true relationship is far from linear. This alludes to the fact that income alone doesn’t fully determine home value.

```
boxplot(LPRICE ~ FRSTHO, data = Homes,
        main = "Log Home Price by First-Time Buyer Status",
        xlab = "First-Time Buyer (0 = No, 1 = Yes)",
        ylab = "Log Home Price",
        col = c("lightblue", "pink"))
```



The boxplot comparing LPRICE across first-time (FRSTHO) vs. repeat buyers reveals that first-time buyers generally purchase lower-priced homes. This aligns with expectations, as first-time buyers often have smaller down payments, lower incomes, or less access to high-value properties compared to repeat buyers who have built more financial resources.

```
hist(Homes$LPRICE,
     main = "Distribution of Log Home Prices",
     xlab = "Log Home Price",
     ylab = "Frequency",
     col = "lightblue",
     breaks = 30)
```



The histogram shows that most home prices are clustered at lower values, meaning cheaper homes are more common. The distribution is slightly skewed right, meaning a smaller number of very expensive homes exist.

Exercise 2: Regress log value onto all but mortgage and purchases.How many coefficients are jointly significant at 10%? Re-run regression with only the significant covariates, and compare R2 to the full model.

```
full_model <- lm(LPRICE ~ . - AMMORT - VALUE, data = Homes)
summary(full_model)
```

```
##
## Call:
## lm(formula = LPRICE ~ . - AMMORT - VALUE, data = Homes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1777798   -44069    -7067    31276  2059345
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.442e+04  7.821e+03   4.401 1.08e-05 ***
## EAPTBLY         -3.825e+03  2.947e+03  -1.298 0.194339
## ECOM1Y          -6.394e+03  2.417e+03  -2.645 0.008172 **
## ECOM2Y          -3.456e+03  6.036e+03  -0.573 0.566965
## EGREENY          6.753e+03  1.759e+03   3.838 0.000124 ***
## EJUNKY          -8.041e+03  6.412e+03  -1.254 0.209882
## ELOW1Y          -3.250e+03  2.904e+03  -1.119 0.263186
## ESFDY            9.334e+03  3.713e+03   2.514 0.011946 *
## ETRANSY          1.101e+03  3.181e+03   0.346 0.729236
## EABANY          -5.111e+03  4.520e+03  -1.131 0.258163
## HOWHgood         1.960e+03  3.306e+03   0.593 0.553264
## HOWNgood         1.622e+04  2.751e+03   5.894 3.84e-09 ***
## ODORAY          -2.468e+02  4.161e+03  -0.059 0.952711
## STRNAY          -5.848e+03  2.019e+03  -2.897 0.003777 **
## ZINC2            1.994e-01  6.957e-03  28.664  < 2e-16 ***
## PER              7.065e+03  7.855e+02   8.993  < 2e-16 ***
## ZADULT          -1.494e+04  1.366e+03 -10.933  < 2e-16 ***
## HHGRADBach       2.268e+04  2.879e+03   7.878 3.54e-15 ***
## HHGRADGrad       3.949e+04  3.238e+03  12.196  < 2e-16 ***
## HHGRADHS Grad   -4.531e+03  2.727e+03  -1.662 0.096589 .
## HHGRADNo HS     -1.003e+04  3.998e+03  -2.509 0.012110 *
## NUNITS           5.824e+01  6.536e+01   0.891 0.372853
## INTW            -8.248e+03  5.538e+02 -14.894  < 2e-16 ***
## METROurban       1.301e+04  2.270e+03   5.728 1.03e-08 ***
## STATECO         -1.374e+04  3.669e+03  -3.744 0.000182 ***
## STATECT         -1.105e+04  3.925e+03  -2.816 0.004869 **
## STATEGA         -4.266e+04  3.904e+03 -10.926  < 2e-16 ***
## STATEIL         -6.779e+04  7.246e+03  -9.356  < 2e-16 ***
## STATEIN         -5.792e+04  3.857e+03 -15.016  < 2e-16 ***
## STATELA         -6.481e+04  4.633e+03 -13.990  < 2e-16 ***
## STATEMO         -4.871e+04  4.200e+03 -11.599  < 2e-16 ***
## STATEOH         -3.993e+04  4.106e+03  -9.725  < 2e-16 ***
## STATEOK         -8.138e+04  4.121e+03 -19.747  < 2e-16 ***
## STATEPA         -6.161e+04  4.257e+03 -14.473  < 2e-16 ***
## STATETX         -8.525e+04  4.309e+03 -19.783  < 2e-16 ***
## STATEWA          2.185e+04  3.886e+03   5.622 1.92e-08 ***
## BATHS            5.313e+04  1.455e+03  36.508  < 2e-16 ***
## BEDRMS           8.714e+03  1.263e+03   6.897 5.53e-12 ***
## MATBUYY          4.138e+04  1.718e+03  24.083  < 2e-16 ***
## DWNPAYprev home  2.412e+04  2.242e+03  10.760  < 2e-16 ***
## FRSTHOY         -1.529e+04  2.166e+03  -7.058 1.76e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 102600 on 15524 degrees of freedom
## Multiple R-squared:  0.407,  Adjusted R-squared:  0.4055
## F-statistic: 266.4 on 40 and 15524 DF,  p-value: < 2.2e-16
```

The regression results indicate that several factors significantly influence log home price (LPRICE). Key predictors with strong statistical significance ($p < 0.001$) include income (ZINC2), bathrooms (BATHS), bedrooms (BEDRMS), graduate education (HHGRADGrad), down payment type (DWNPAYprev home), and first-time buyer status (FRSTHOY). Higher income, more bathrooms, and being a repeat homebuyer are associated with higher home values, while first-time buyers tend to purchase lower-priced homes.

Location also plays a significant role, as indicated by state-level variables (STATEGA, STATEIL, STATETX, etc.), many of which have large coefficients and p-values near zero, suggesting regional differences in home prices. Additionally, urban areas (METROurban) are associated with higher home values.

Overall, tjuse are 13 predictors that are jointly significant at the 10% level ($p<0.10$), meaning they have a statistically significant impact on log home price (LPRICE). These include income (ZINC2), household size (ZADULT), education level (HHGRAD), first-time buyer status (FRSTHO), bathrooms (BATHS), bedrooms (BEDRMS), and down payment type (DWNPAY), which directly influence home values. Additionally, location factors (METRO, STATE) show significant regional variations in pricing. Other notable factors include material of purchase (MATBUY), percent change in price (PER), and transportation accessibility (INTW), suggesting that both personal financial characteristics and external market factors play crucial roles in determining home prices.

```
reduced_model <- lm(LPRICE ~ ZINC2 + ZADULT + HHGRAD + FRSTHO + BATHS + BEDRMS + DWNPAY + METRO
+ STATE + MATBUY + PER + INTW, data = Homes)

# Display summary of the reduced model
summary(reduced_model)
```

```
##
## Call:
## lm(formula = LPRICE ~ ZINC2 + ZADULT + HHGRAD + FRSTHO + BATHS +
##     BEDRMS + DWNPAY + METRO + STATE + MATBUY + PER + INTW, data = Homes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1796669   -44317    -6922    31450  2072356
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.642e+04  6.334e+03   8.908  < 2e-16 ***
## ZINC2            2.016e-01  6.973e-03  28.910  < 2e-16 ***
## ZADULT          -1.497e+04  1.370e+03 -10.926  < 2e-16 ***
## HHGRADBach       2.322e+04  2.885e+03   8.050 8.91e-16 ***
## HHGRADGrad       4.013e+04  3.244e+03  12.371  < 2e-16 ***
## HHGRADHS Grad   -5.088e+03  2.734e+03  -1.861  0.06280 .
## HHGRADNo HS     -1.081e+04  4.007e+03  -2.698  0.00698 **
## FRSTHOY         -1.657e+04  2.168e+03  -7.645 2.22e-14 ***
## BATHS            5.440e+04  1.448e+03  37.560  < 2e-16 ***
## BEDRMS           9.568e+03  1.240e+03   7.717 1.26e-14 ***
## DWNPAYprev home  2.525e+04  2.246e+03  11.243  < 2e-16 ***
## METROurban       9.526e+03  2.222e+03   4.287 1.82e-05 ***
## STATECO         -1.477e+04  3.657e+03  -4.038 5.42e-05 ***
## STATECT         -1.063e+04  3.929e+03  -2.706  0.00682 **
## STATEGA         -4.323e+04  3.896e+03 -11.096  < 2e-16 ***
## STATEIL         -6.845e+04  7.259e+03  -9.430  < 2e-16 ***
## STATEIN         -5.749e+04  3.861e+03 -14.891  < 2e-16 ***
## STATELA         -6.630e+04  4.629e+03 -14.322  < 2e-16 ***
## STATEMO         -4.909e+04  4.210e+03 -11.660  < 2e-16 ***
## STATEOH         -4.075e+04  4.096e+03  -9.947  < 2e-16 ***
## STATEOK         -7.997e+04  4.126e+03 -19.382  < 2e-16 ***
## STATEPA         -6.216e+04  4.255e+03 -14.609  < 2e-16 ***
## STATETX         -8.410e+04  4.314e+03 -19.493  < 2e-16 ***
## STATEWA          2.073e+04  3.891e+03   5.327 1.01e-07 ***
## MATBUYY          4.149e+04  1.720e+03  24.115  < 2e-16 ***
## PER              7.081e+03  7.865e+02   9.003  < 2e-16 ***
## INTW            -8.698e+03  5.528e+02 -15.734  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 103000 on 15538 degrees of freedom
## Multiple R-squared:  0.4025, Adjusted R-squared:  0.4015
## F-statistic: 402.5 on 26 and 15538 DF,  p-value: < 2.2e-16
```

The adjusted R-Squared dropped slightly from 0.4055 to 0.4015, meaning the reduced model explains almost the same variation in log home prices than the full model. This confirms that the removed predictors did not contribute much explanatory power, making the reduced model more efficient without losing significant predictive ability.

Exercise 3: Fit a regression for whether the buyer had ≥ 20% down (again, onto everything but AMMORT and LPRICE). Interpret effects for 1st home buyers and # of bathrooms. Add + describe interaction for 1st home-buyers and #baths

```
Homes$DOWN20 <- ifelse(Homes$DWNPAY == "prev home", 1, 0)
logit_model <- glm(DOWN20 ~ . - AMMORT - LPRICE, data = Homes, family = binomial)
```

```
## Warning: glm.fit: algorithm did not converge
```

```
summary(logit_model)
```

```
## 
## Call:
## glm(formula = DOWN20 ~ . - AMMORT - LPRICE, family = binomial, 
##     data = Homes)
## 
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -2.657e+01  2.723e+04  -0.001    0.999
## EAPTBLY         -6.794e-12  1.023e+04   0.000    1.000
## ECOM1Y           8.286e-12  8.388e+03   0.000    1.000
## ECOM2Y          -5.354e-11  2.095e+04   0.000    1.000
## EGREENY          1.811e-12  6.106e+03   0.000    1.000
## EJUNKY           7.202e-12  2.225e+04   0.000    1.000
## ELOW1Y           4.503e-12  1.008e+04   0.000    1.000
## ESFDY            3.590e-12  1.289e+04   0.000    1.000
## ETRANSY          4.133e-12  1.104e+04   0.000    1.000
## EABANY          -1.503e-11  1.568e+04   0.000    1.000
## HOWHgood         1.296e-11  1.147e+04   0.000    1.000
## HOWNgood        -1.041e-11  9.560e+03   0.000    1.000
## ODORAY          -2.464e-12  1.444e+04   0.000    1.000
## STRNAY           6.608e-12  7.004e+03   0.000    1.000
## ZINC2            3.023e-18  2.475e-02   0.000    1.000
## PER             -5.010e-13  2.726e+03   0.000    1.000
## ZADULT           6.816e-13  4.742e+03   0.000    1.000
## HHGRADBach      -6.662e-12  1.001e+04   0.000    1.000
## HHGRADGrad      -3.073e-12  1.130e+04   0.000    1.000
## HHGRADHS Grad    3.100e-13  9.461e+03   0.000    1.000
## HHGRADNo HS      4.467e-12  1.387e+04   0.000    1.000
## NUNITS          -1.029e-13  2.268e+02   0.000    1.000
## INTW             3.521e-14  1.927e+03   0.000    1.000
## METROurban       5.993e-13  7.906e+03   0.000    1.000
## STATECO         -1.258e-11  1.288e+04   0.000    1.000
## STATECT          1.005e-11  1.376e+04   0.000    1.000
## STATEGA          2.519e-11  1.402e+04   0.000    1.000
## STATEIL          6.484e-12  2.548e+04   0.000    1.000
## STATEIN          5.842e-12  1.398e+04   0.000    1.000
## STATELA          5.796e-12  1.651e+04   0.000    1.000
## STATEMO         -1.546e-12  1.499e+04   0.000    1.000
## STATEOH          1.044e-11  1.465e+04   0.000    1.000
## STATEOK          1.714e-12  1.506e+04   0.000    1.000
## STATEPA          4.149e-12  1.528e+04   0.000    1.000
## STATETX          1.371e-12  1.575e+04   0.000    1.000
## STATEWA          9.564e-12  1.350e+04   0.000    1.000
## BATHS           -2.550e-12  5.254e+03   0.000    1.000
## BEDRMS          -1.102e-12  4.409e+03   0.000    1.000
## MATBUYY         -1.059e-12  5.962e+03   0.000    1.000
## DWNPAYprev home  5.313e+01  7.795e+03   0.007    0.995
## VALUE           -3.794e-17  2.134e-02   0.000    1.000
## FRSTHOY         -4.176e-12  7.527e+03   0.000    1.000
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
```

```
##     Null deviance: 1.9957e+04  on 15564  degrees of freedom
## Residual deviance: 9.0302e-08  on 15523  degrees of freedom
## AIC: 84
##
## Number of Fisher Scoring iterations: 25
```

The coefficient for FRSTHOY is negative (−4.176e−12) but has a p-value of 1.000, meaning it is not statistically significant. This suggests that being a first-time buyer does not have a meaningful effect on the likelihood of putting down at least 20%.

The coefficient for BATHS is very small (−2.550e−12) with a p-value of 1.000, meaning it is not statistically significant. This suggests that the number of bathrooms in a home does not influence the likelihood of a 20% down payment.

```
interaction_model <- glm(DOWN20 ~ FRSTHO * BATHS, data = Homes, family = binomial)
summary(interaction_model)
```

```
##
## Call:
## glm(formula = DOWN20 ~ FRSTHO * BATHS, family = binomial, data = Homes)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.52581    0.06461  -8.139 3.99e-16 ***
## FRSTHOY       -19.04026  376.28693  -0.051    0.960
## BATHS           0.44052    0.03063  14.384  < 2e-16 ***
## FRSTHOY:BATHS  -0.44052  218.38525  -0.002    0.998
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 19957  on 15564  degrees of freedom
## Residual deviance: 11937  on 15561  degrees of freedom
## AIC: 11945
##
## Number of Fisher Scoring iterations: 18
```

The results show that first-time buyer status (FRSTHOY) has no significant effect on the likelihood of putting down ≥ 20% (p = 0.960). However, more bathrooms (BATHS) significantly increase this likelihood (p < 2e-16), suggesting that buyers of larger, higher-value homes tend to make larger down payments. The interaction term (FRSTHOY * BATHS) is not significant (p = 0.998), meaning the effect of bathrooms on down payment likelihood is the same for both first-time and repeat buyers. This indicates that home size, rather than buyer type, is a key predictor of down payment behavior.

Exercise 4: Re-fit your model from Q3 for only homes worth > 100k. Compare in-sample fit to R2 for predicting homes worth < 100 k.

```
Homes_over_100K <- subset(Homes, VALUE > 100000)
Homes_under_100K <- subset(Homes, VALUE <= 100000)
logit_over_100K <- glm(DOWN20 ~ ZINC2 + ZADULT + HHGRAD + FRSTHO + BATHS + BEDRMS + DWNPAY + MET
RO + STATE + MATBUY + PER + INTW,
                        data = Homes_over_100K, family = binomial)
```

```
## Warning: glm.fit: algorithm did not converge
```

```
summary(logit_over_100K)
```

```
##
## Call:
## glm(formula = DOWN20 ~ ZINC2 + ZADULT + HHGRAD + FRSTHO + BATHS +
##     BEDRMS + DWNPAY + METRO + STATE + MATBUY + PER + INTW, family = binomial,
##     data = Homes_over_100K)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -2.657e+01  2.549e+04  -0.001    0.999
## ZINC2           -4.529e-18  2.465e-02   0.000    1.000
## ZADULT          -9.448e-13  5.410e+03   0.000    1.000
## HHGRADBach      -6.135e-12  1.115e+04   0.000    1.000
## HHGRADGrad      -4.211e-12  1.234e+04   0.000    1.000
## HHGRADHS Grad    1.075e-12  1.093e+04   0.000    1.000
## HHGRADNo HS      1.974e-11  1.756e+04   0.000    1.000
## FRSTHOY         -3.495e-12  8.687e+03   0.000    1.000
## BATHS           -1.726e-12  5.533e+03   0.000    1.000
## BEDRMS          -3.349e-12  4.846e+03   0.000    1.000
## DWNPAYprev home  5.313e+01  8.490e+03   0.006    0.995
## METROurban       1.304e-12  9.203e+03   0.000    1.000
## STATECO         -1.057e-11  1.284e+04   0.000    1.000
## STATECT          8.941e-13  1.407e+04   0.000    1.000
## STATEGA          3.345e-11  1.400e+04   0.000    1.000
## STATEIL          1.015e-11  3.147e+04   0.000    1.000
## STATEIN         -2.561e-12  1.469e+04   0.000    1.000
## STATELA         -2.913e-13  1.843e+04   0.000    1.000
## STATEMO         -1.499e-11  1.594e+04   0.000    1.000
## STATEOH          4.590e-12  1.541e+04   0.000    1.000
## STATEOK         -2.703e-12  1.782e+04   0.000    1.000
## STATEPA          4.454e-12  1.780e+04   0.000    1.000
## STATETX          2.320e-12  1.919e+04   0.000    1.000
## STATEWA          5.629e-12  1.371e+04   0.000    1.000
## MATBUYY          1.176e-12  6.688e+03   0.000    1.000
## PER             -3.739e-12  3.067e+03   0.000    1.000
## INTW             3.921e-12  2.540e+03   0.000    1.000
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1.6272e+04  on 12143  degrees of freedom
## Residual deviance: 7.0454e-08  on 12117  degrees of freedom
## AIC: 54
##
## Number of Fisher Scoring iterations: 25
```

All predictors except DWNPAYprev home have p-values of 1.000, meaning they are statistically insignificant in explaining the likelihood of putting down ≥ 20%. This suggests that factors like income (ZINC2), home size (BATHS, BEDRMS), education (HHGRAD), and location (STATE) do not meaningfully contribute to predicting down payment size for higher-value homes.

```
r2_over_100K <- 1 - (logit_over_100K$deviance / logit_over_100K$null.deviance)
cat("McFadden's R^2 for Homes > $100K:", r2_over_100K, "\n")
```

```
## McFadden's R^2 for Homes > $100K: 1
```

```
predicted_probs <- predict(logit_over_100K, newdata = Homes_under_100K, type = "response")
predicted_classes <- ifelse(predicted_probs >= 0.5, 1, 0)
accuracy <- mean(predicted_classes == Homes_under_100K$DOWN20)

cat("Prediction Accuracy for Homes < $100K:", accuracy, "\n")
```

```
## Prediction Accuracy for Homes < $100K: 1
```

While the model appears perfect, this is almost certainly due to data separation, overfitting, or an overpowering variable rather than truly perfect predictability.