Home Values Analysis

(Student: Vinh Luong - 442069)

1.DATA IMPORT & PRE-PROCESSING

First of all, we read in the data and change the columns to make them less of an eyesore. We remove the *FirstMortgageAmount* and *PurchasePrice* variables as they are very directly linked to the *CurrentValue* variable of interest. We also get rid of several rows that have negative *Household Income*.

```
file_name <- "homes2004.csv"
homes <- read.csv(file_name, stringsAsFactors = TRUE)
homes <- as.data.table(homes)
setnames(homes, c("FirstMortgageAmount", "NearApartments", "NearBusiness", "NearIndustry",
                  "NearGreen", "NearTrash", "NearSingleFamilyTownhouses", "NearSingleFamilyHomes",
                  "NearMajorTransportLink", "NearAbandonedBuildings", "HomeRating",
                  "NeighborhoodRating", "NeighborhoodBadSmells", "NeighborhoodNoisy",
                  "HouseholdIncome", "NumPersons", "NumAdults", "EducationLevel",
                  "NumUnitsInBuilding", "MortgateInterestRate", "RuralOrUrban", "State",
                  "PurchasePrice", "NumBathrooms", "NumBedrooms", "MortgateBoughtSameYear",
                  "DownpaymentSource", "CurrentValue", "FirstHome"))
homes <- homes[HouseholdIncome >= 0, ]
excluded_vars = homes[, .(FirstMortgageAmount, PurchasePrice)]
homes[, c("FirstMortgageAmount", "PurchasePrice") := NULL]
```

2. LINEAR REGRESSION OF log(VALUE)

Below, we run a linear regression of the *log* of *CurrentValue* on all other variables except *MortgageAmount* and *PurchasePrice*:

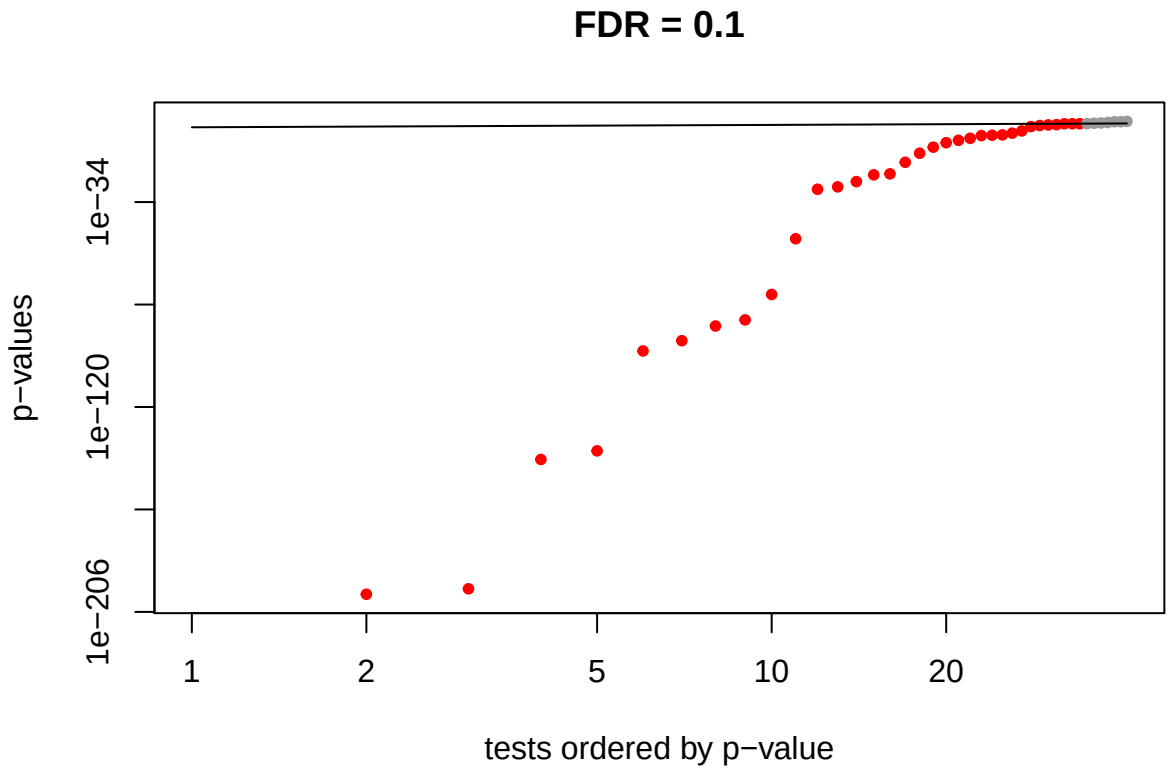
```
linear_model <- train(log(CurrentValue) ~ ., data = copy(homes), method = 'glm')
results <- coef(summary(linear_model$finalModel))[, c("Estimate", "Pr(>|t|)")]
results
```

```
##                                 Estimate      Pr(>|t|)
## (Intercept)                 1.160051e+01  0.000000e+00
## NearApartmentsY            -4.212260e-02  7.275616e-02
## NearBusinessY              -2.708860e-02  1.593572e-01
## NearIndustryY              -8.581602e-02  7.414738e-02
## NearGreenY                  9.184912e-03  5.120445e-01
## NearTrashY                 -1.265735e-01  1.316575e-02
## NearSingleFamilyTownhousesY  2.653726e-02  2.512538e-01
## NearSingleFamilyHomesY       2.916473e-01  7.098407e-23
## NearMajorTransportLinkY    -1.587580e-02  5.308383e-01
## NearAbandonedBuildingsY    -1.605475e-01  8.479411e-06
## HomeRatinggood              1.299966e-01  8.105695e-07
## NeighborhoodRatinggood      1.189541e-01  5.763041e-08
## NeighborhoodBadSmellsY      1.029798e-02  7.558753e-01
```

```
## NeighborhoodNoisyY           -3.623373e-02  2.423051e-02
## HouseholdIncome               6.243801e-07  2.394195e-29
## NumPersons                    9.626206e-03  1.237876e-01
## NumAdults                    -1.869489e-02  8.566313e-02
## EducationLevelBach            1.322092e-01  8.182772e-09
## EducationLevelGrad            1.959812e-01  3.130868e-14
## `EducationLevelHS Grad`      -6.127260e-02  4.778140e-03
## `EducationLevelNo HS`        -1.956959e-01  8.233226e-10
## NumUnitsInBuilding           -9.251729e-04  7.538443e-02
## MortgateInterestRate         -4.671977e-02  3.920188e-26
## RuralOrUrbanurban             8.680240e-02  1.594981e-06
## StateCO                      -2.911438e-01  2.726322e-23
## StateCT                      -3.459025e-01  2.394833e-28
## StateGA                      -6.551847e-01  3.365937e-97
## StateIL                      -8.616949e-01  4.111913e-50
## StateIN                      -7.792566e-01 3.827217e-139
## StateLA                      -7.216665e-01  3.617263e-84
## StateMO                      -6.643065e-01  9.701868e-87
## StateOH                      -6.731513e-01  6.527271e-93
## StateOK                      -9.973124e-01 4.989152e-197
## StatePA                      -8.713551e-01 9.258785e-143
## StateTX                      -1.048815e+00 2.589841e-199
## StateWA                      -1.228880e-01  7.221676e-05
## NumBathrooms                  2.113446e-01  1.638043e-73
## NumBedrooms                   8.723059e-02  4.914555e-18
## MortgateBoughtSameYearY      -2.937492e-02  3.184927e-02
## `DownpaymentSourceprev home`  1.213431e-01  1.115125e-11
## FirstHomeY                   -8.403527e-02  1.120203e-06
```

```
r_square <- 1 - summary(linear_model$finalModel)$deviance /
  summary(linear_model$finalModel)$null.deviance
```

This model has an R^2 statistic of **0.3053277**.



In order to control the expected False Discovery Rate at 10%, we should only consider variables with coefficients that have p-values smaller than **0.0753844**. The list of such variables is below:

```
results[pvals < alpha, ]
```

```
##                                 Estimate      Pr(>|t|)
## (Intercept)                 1.160051e+01  0.000000e+00
## NearApartmentsY            -4.212260e-02  7.275616e-02
## NearIndustryY              -8.581602e-02  7.414738e-02
## NearTrashY                 -1.265735e-01  1.316575e-02
## NearSingleFamilyHomesY      2.916473e-01  7.098407e-23
## NearAbandonedBuildingsY    -1.605475e-01  8.479411e-06
## HomeRatinggood              1.299966e-01  8.105695e-07
## NeighborhoodRatinggood      1.189541e-01  5.763041e-08
## NeighborhoodNoisyY         -3.623373e-02  2.423051e-02
## HouseholdIncome             6.243801e-07  2.394195e-29
## EducationLevelBach          1.322092e-01  8.182772e-09
## EducationLevelGrad          1.959812e-01  3.130868e-14
## `EducationLevelHS Grad`    -6.127260e-02  4.778140e-03
## `EducationLevelNo HS`      -1.956959e-01  8.233226e-10
## MortgateInterestRate       -4.671977e-02  3.920188e-26
## RuralOrUrbanurban           8.680240e-02  1.594981e-06
## StateCO                    -2.911438e-01  2.726322e-23
## StateCT                    -3.459025e-01  2.394833e-28
## StateGA                    -6.551847e-01  3.365937e-97
## StateIL                    -8.616949e-01  4.111913e-50
```

```
## StateIN                       -7.792566e-01 3.827217e-139
## StateLA                       -7.216665e-01  3.617263e-84
## StateMO                       -6.643065e-01  9.701868e-87
## StateOH                       -6.731513e-01  6.527271e-93
## StateOK                       -9.973124e-01 4.989152e-197
## StatePA                       -8.713551e-01 9.258785e-143
## StateTX                       -1.048815e+00 2.589841e-199
## StateWA                       -1.228880e-01  7.221676e-05
## NumBathrooms                   2.113446e-01  1.638043e-73
## NumBedrooms                    8.723059e-02  4.914555e-18
## MortgateBoughtSameYearY       -2.937492e-02  3.184927e-02
## `DownpaymentSourceprev home`   1.213431e-01  1.115125e-11
## FirstHomeY                    -8.403527e-02  1.120203e-06
```

We run a second regression on these variables only:

```
linear_model_fewer_vars <- train(log(CurrentValue) ~ NearApartments + NearIndustry + NearTrash +
                                   NearSingleFamilyHomes + NearAbandonedBuildings + HomeRating +
                                   NeighborhoodRating + NeighborhoodNoisy + HouseholdIncome +
                                   EducationLevel + MortgateInterestRate + RuralOrUrban + State +
                                   NumBathrooms + NumBedrooms + MortgateBoughtSameYear +
                                   DownpaymentSource + FirstHome,
                                 data = copy(homes), method = "glm")
results_2 <- coef(summary(linear_model_fewer_vars$finalModel))[, c("Estimate", "Pr(>|t|)")]
results_2
```

```
##                                 Estimate      Pr(>|t|)
## (Intercept)                   1.158923e+01  0.000000e+00
## NearApartmentsY              -4.897865e-02  2.574900e-02
## NearIndustryY                -9.845035e-02  3.609951e-02
## NearTrashY                   -1.255075e-01  1.365005e-02
## NearSingleFamilyHomesY        2.876846e-01  8.960303e-23
## NearAbandonedBuildingsY      -1.608349e-01  7.910933e-06
## HomeRatinggood                1.294143e-01  8.871379e-07
## NeighborhoodRatinggood        1.188917e-01  5.496940e-08
## NeighborhoodNoisyY           -3.963469e-02  1.232701e-02
## HouseholdIncome               6.191985e-07  3.916349e-29
## EducationLevelBach            1.324746e-01  7.361642e-09
## EducationLevelGrad            1.957428e-01  3.022193e-14
## `EducationLevelHS Grad`      -6.204207e-02  4.259637e-03
## `EducationLevelNo HS`        -1.976716e-01  5.134861e-10
## MortgateInterestRate         -4.713071e-02  1.175789e-26
## RuralOrUrbanurban             8.385259e-02  2.946534e-06
## StateCO                      -2.874840e-01  5.173652e-23
## StateCT                      -3.446987e-01  3.080224e-28
## StateGA                      -6.560752e-01  2.184925e-98
## StateIL                      -8.620022e-01  3.175000e-50
## StateIN                      -7.777076e-01 6.681303e-139
## StateLA                      -7.245829e-01  2.803478e-85
## StateMO                      -6.647062e-01  7.094244e-87
## StateOH                      -6.788921e-01  4.550427e-95
## StateOK                      -9.967420e-01 5.126824e-197
## StatePA                      -8.679300e-01 3.538992e-142
```

```
## StateTX                       -1.049796e+00 3.393676e-200
## StateWA                       -1.212725e-01  8.811018e-05
## NumBathrooms                   2.127695e-01  7.679510e-75
## NumBedrooms                    8.908465e-02  1.908951e-21
## MortgateBoughtSameYearY       -2.777407e-02  4.174313e-02
## `DownpaymentSourceprev home`   1.215572e-01  9.906898e-12
## FirstHomeY                    -8.288506e-02  1.448077e-06
```

```
r_square_2 <- 1 - summary(linear_model_fewer_vars$finalModel)$deviance /
  summary(linear_model_fewer_vars$finalModel)$null.deviance
```

The model with fewer variables has an R^2 statistic of **0.3048328**, which is almost identical to the previous model' R^2 of 0.3053277, suggesting that the removed variables do not indeed matter much.

3. LOGISTIC REGRESSION

We create a new variable indicating whether the buyer had over 20% downpayment in the purchase, and regress it on the same independent variables as above:

```
homes$Over20PercentDownpayment <-
  factor(excluded_vars$FirstMortgageAmount < 0.8 * excluded_vars$PurchasePrice)
logistic_model <- train(Over20PercentDownpayment ~ ., data = copy(homes),
                        method = "glm", family = "binomial")
r_square <- 1 - summary(logistic_model$finalModel)$deviance /
  summary(logistic_model$finalModel)$null.deviance
summary(logistic_model)
```

```
##
## Call:
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4450  -0.8081  -0.5985   1.0693   2.4750
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -1.294e+00  1.833e-01  -7.063 1.63e-12 ***
## NearApartmentsY                1.797e-02  7.028e-02   0.256 0.798181
## NearBusinessY                 -1.621e-01  5.812e-02  -2.789 0.005287 **
## NearIndustryY                 -3.133e-01  1.600e-01  -1.958 0.050198 .
## NearGreenY                    -4.971e-04  3.985e-02  -0.012 0.990047
## NearTrashY                    -8.353e-03  1.608e-01  -0.052 0.958571
## NearSingleFamilyTownhousesY    4.505e-02  6.632e-02   0.679 0.496945
## NearSingleFamilyHomesY        -2.632e-01  8.285e-02  -3.177 0.001486 **
## NearMajorTransportLinkY       -6.076e-02  7.617e-02  -0.798 0.425096
## NearAbandonedBuildingsY       -8.489e-02  1.160e-01  -0.732 0.464445
## HomeRatinggood                -1.416e-01  7.952e-02  -1.780 0.075064 .
## NeighborhoodRatinggood         1.616e-01  6.736e-02   2.398 0.016464 *
## NeighborhoodBadSmellsY         1.055e-01  9.810e-02   1.075 0.282291
## NeighborhoodNoisyY            -9.893e-02  4.741e-02  -2.086 0.036936 *
```

```
## HouseholdIncome              -1.155e-07  1.851e-07  -0.624 0.532683
## NumPersons                   -1.251e-01  1.856e-02  -6.744 1.54e-11 ***
## NumAdults                     2.050e-02  3.187e-02   0.643 0.520053
## EducationLevelBach            1.786e-01  6.597e-02   2.707 0.006786 **
## EducationLevelGrad            2.688e-01  7.291e-02   3.687 0.000227 ***
## `EducationLevelHS Grad`      -2.152e-02  6.377e-02  -0.338 0.735716
## `EducationLevelNo HS`        -7.051e-02  9.847e-02  -0.716 0.473918
## NumUnitsInBuilding            2.362e-03  1.427e-03   1.655 0.098010 .
## MortgateInterestRate         -6.359e-02  1.372e-02  -4.633 3.60e-06 ***
## RuralOrUrbanurban            -8.097e-02  5.392e-02  -1.502 0.133213
## StateCO                      -2.698e-02  8.500e-02  -0.317 0.750920
## StateCT                       7.855e-01  8.830e-02   8.897  < 2e-16 ***
## StateGA                      -2.233e-01  9.462e-02  -2.360 0.018282 *
## StateIL                       5.863e-01  1.635e-01   3.586 0.000336 ***
## StateIN                       2.408e-01  9.358e-02   2.573 0.010078 *
## StateLA                       5.884e-01  1.079e-01   5.455 4.89e-08 ***
## StateMO                       5.309e-01  9.732e-02   5.455 4.88e-08 ***
## StateOH                       7.649e-01  9.483e-02   8.066 7.27e-16 ***
## StateOK                       1.299e-01  1.028e-01   1.264 0.206329
## StatePA                       6.009e-01  1.007e-01   5.965 2.45e-09 ***
## StateTX                       2.932e-01  1.073e-01   2.732 0.006288 **
## StateWA                       1.543e-01  8.823e-02   1.748 0.080381 .
## NumBathrooms                  2.447e-01  3.421e-02   7.155 8.39e-13 ***
## NumBedrooms                  -2.096e-02  2.911e-02  -0.720 0.471579
## MortgateBoughtSameYearY       2.609e-01  3.929e-02   6.640 3.14e-11 ***
## `DownpaymentSourceprev home`  7.397e-01  4.859e-02  15.222  < 2e-16 ***
## CurrentValue                  1.483e-06  1.451e-07  10.219  < 2e-16 ***
## FirstHomeY                   -3.701e-01  5.172e-02  -7.156 8.32e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 18851  on 15550  degrees of freedom
## Residual deviance: 16953  on 15509  degrees of freedom
## AIC: 17037
##
## Number of Fisher Scoring iterations: 4
```

The coefficient on *FirstHome* is negative (-0.37), implying that young persons who buy homes for the first time are likely to have less equity financing available and hence are more reliant on debt. The odds of people buying their first homes putting down over 20% downpayment is exp(-0.37) = 0.69 times that of people buying subsequent homes doing so.

The coefficent on *NumBathrooms* is positive (0.24), implying that people buying larger homes also tend to have a greater % downpayment.

This logistic regression has an R^2 statistic of **0.1006724**.

We fit a second logistic regression with the interactions of the above two variables:

```
logistic_model_with_interaction <- train(Over20PercentDownpayment ~ . + FirstHome * NumBathrooms,
                                         data = copy(homes), method = "glm", family = "binomial")
summary(logistic_model_with_interaction)
```

```
##
```

```
## Call:
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4351  -0.8053  -0.5980   1.0661   2.4438
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -1.378e+00  1.853e-01  -7.437 1.03e-13 ***
## NearApartmentsY                1.518e-02  7.022e-02   0.216 0.828826
## NearBusinessY                 -1.610e-01  5.809e-02  -2.772 0.005572 **
## NearIndustryY                 -3.183e-01  1.598e-01  -1.992 0.046382 *
## NearGreenY                    -1.211e-03  3.988e-02  -0.030 0.975777
## NearTrashY                    -3.995e-03  1.606e-01  -0.025 0.980156
## NearSingleFamilyTownhousesY    4.824e-02  6.632e-02   0.727 0.467022
## NearSingleFamilyHomesY        -2.675e-01  8.284e-02  -3.229 0.001241 **
## NearMajorTransportLinkY       -5.957e-02  7.613e-02  -0.782 0.433934
## NearAbandonedBuildingsY       -9.513e-02  1.159e-01  -0.821 0.411683
## HomeRatinggood                -1.368e-01  7.943e-02  -1.722 0.085026 .
## NeighborhoodRatinggood         1.648e-01  6.734e-02   2.448 0.014380 *
## NeighborhoodBadSmellsY         1.036e-01  9.804e-02   1.057 0.290588
## NeighborhoodNoisyY            -9.917e-02  4.740e-02  -2.092 0.036435 *
## HouseholdIncome               -1.353e-07  1.875e-07  -0.722 0.470565
## NumPersons                    -1.265e-01  1.859e-02  -6.803 1.03e-11 ***
## NumAdults                      2.299e-02  3.192e-02   0.720 0.471375
## EducationLevelBach             1.805e-01  6.598e-02   2.736 0.006217 **
## EducationLevelGrad             2.728e-01  7.297e-02   3.739 0.000185 ***
## `EducationLevelHS Grad`       -2.063e-02  6.375e-02  -0.324 0.746234
## `EducationLevelNo HS`         -7.579e-02  9.839e-02  -0.770 0.441117
## NumUnitsInBuilding             2.269e-03  1.415e-03   1.604 0.108629
## MortgateInterestRate          -6.451e-02  1.372e-02  -4.703 2.57e-06 ***
## RuralOrUrbanurban             -8.499e-02  5.394e-02  -1.576 0.115114
## StateCO                       -3.702e-02  8.525e-02  -0.434 0.664118
## StateCT                        7.726e-01  8.842e-02   8.739  < 2e-16 ***
## StateGA                       -2.323e-01  9.495e-02  -2.447 0.014410 *
## StateIL                        5.733e-01  1.635e-01   3.506 0.000455 ***
## StateIN                        2.345e-01  9.375e-02   2.501 0.012377 *
## StateLA                        5.848e-01  1.080e-01   5.415 6.12e-08 ***
## StateMO                        5.196e-01  9.751e-02   5.329 9.87e-08 ***
## StateOH                        7.514e-01  9.496e-02   7.913 2.52e-15 ***
## StateOK                        1.184e-01  1.030e-01   1.150 0.249986
## StatePA                        5.816e-01  1.010e-01   5.761 8.37e-09 ***
## StateTX                        2.873e-01  1.075e-01   2.673 0.007527 **
## StateWA                        1.551e-01  8.832e-02   1.756 0.079015 .
## NumBathrooms                   2.989e-01  3.826e-02   7.813 5.59e-15 ***
## NumBedrooms                   -2.167e-02  2.915e-02  -0.743 0.457269
## MortgateBoughtSameYearY        2.613e-01  3.932e-02   6.645 3.03e-11 ***
## `DownpaymentSourceprev home`   7.318e-01  4.871e-02  15.024  < 2e-16 ***
## CurrentValue                   1.442e-06  1.458e-07   9.895  < 2e-16 ***
## FirstHomeY                    -2.590e-02  1.185e-01  -0.219 0.826974
## `NumBathrooms:FirstHomeY`     -1.995e-01  6.211e-02  -3.212 0.001319 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 18851  on 15550  degrees of freedom
## Residual deviance: 16942  on 15508  degrees of freedom
## AIC: 17028
##
## Number of Fisher Scoring iterations: 4
```

The coefficients on *FirstHome* and *NumBathrooms* are now both more moderate in absolute value, and there is a statistically significant coefficient of -0.2 on the interaction term.

4. IN-SAMPLE AND OUT-OF-SAMPLE TESTS

We re-fit the first logistic regression on a the data set with homes with value over $100,000 only:

```
logistic_model_over_100k <- train(Over20PercentDownpayment ~ .,
                                  data = copy(homes[CurrentValue > 1e5, ]),
                                  method = "glm", family = "binomial")
r_square <- 1 - summary(logistic_model_over_100k$finalModel)$deviance /
  summary(logistic_model_over_100k$finalModel)$null.deviance
```

The R^2 statistic for this model is **0.1043649**.

We then use this model to fit the data set with home values below $100,000, and measure the goodness of fit R^2 from the deviance statistics:

```
pred <- predict(logistic_model_over_100k$finalModel,
                as.data.frame(model.matrix(~ ., copy(homes[CurrentValue < 1e5, ]))),
                type = 'response')
source('deviance.R')
r_square_2 <- R2(copy(homes[CurrentValue < 1e5, ])$Over20PercentDownpayment, pred,
                 family = "binomial")
```

This model has an R^2 statistic of **0.0321295**, which is much worse than the previous model's R^2 statistic of 0.1043649. This suggests that the patterns are very different in the two data sets above and below the $100,000 valuation threshold.