```
library(readr)
```

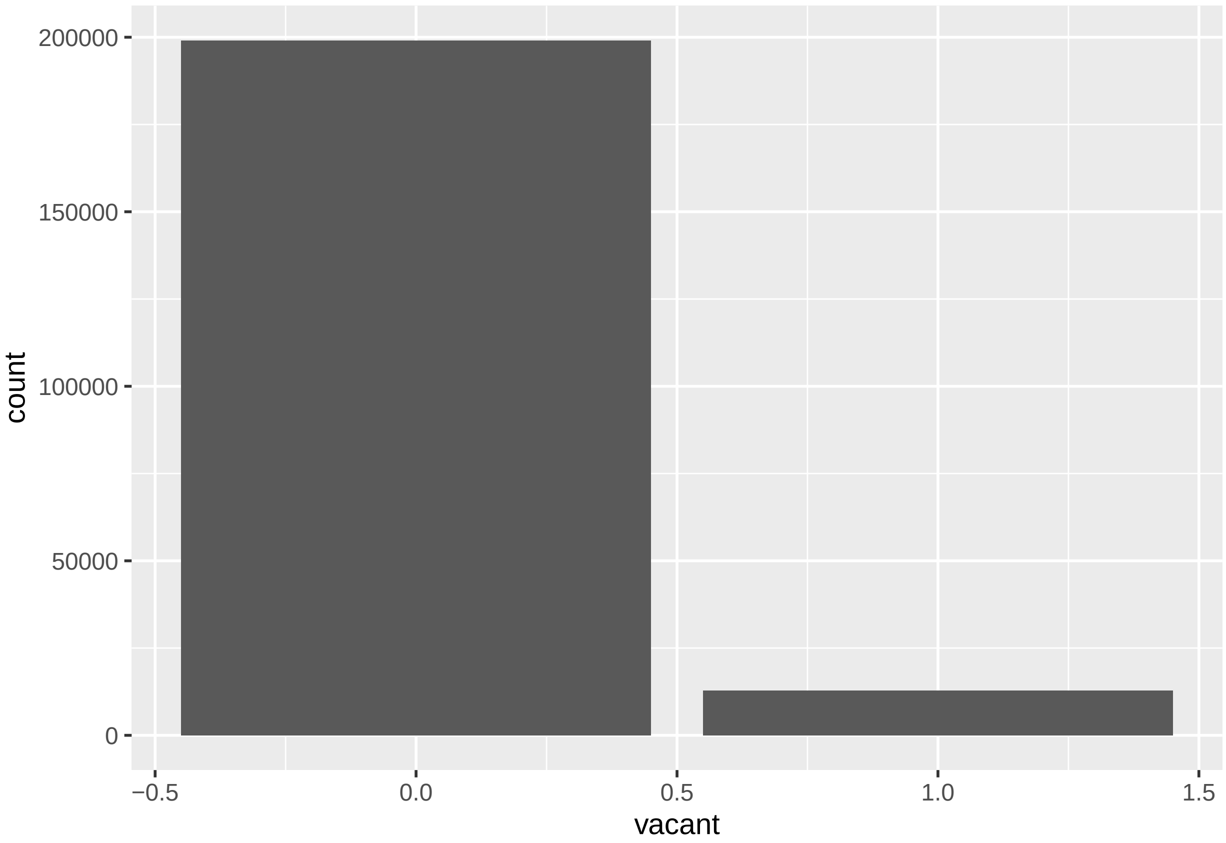
```
## Warning: package 'readr' was built under R version 4.3.3
```

```
data = read.csv("final_cleaned_detroit.csv")
head(data)
```

```
##   Parcel.ID                     Taxpayer.1 Property.Class Total.Floor.Area
## 1 21062470.                853 ASHLAND LLC            401             2468
## 2 21063979.               CASTLE, ROBERT M            401             1389
## 3 21068429.               CITIMORTGAGE INC            401             1293
## 4 21069819.               ROBINSON, DENISE            401             1540
## 5 21070146.              BRADFORD, WILLIAM            401              920
## 6 21069974. EQUITY TRUST COMPANY CUSTODIAN            401             2040
##   Total.Acreage Frontage Depth Building.Count Year.Built Sale.Price
## 1         0.079       30   115              1       1916      65000
## 2         0.000        0     0              1       2002      13600
## 3         0.115       40   126              1       1938      79992
## 4         0.105       40   114              1       1929       8912
## 5         0.107       40   116              1       1938      20000
## 6         0.105       40   114              1       1929      27900
##   Assessed.Value Previous.Assessed.Value Taxable.Value Previous.Taxable.Value
## 1          61300                   46200         61300                  16361
## 2          40100                   31700         13286                  12654
## 3          40400                   33400         40400                  15835
## 4          39900                   31300         14391                  13706
## 5          20300                   16200          9396                   8949
## 6          43700                   34400         23095                  21996
##         Neighborhood fine_amount yearly_average vacant Binary.Tax.Status
## 1 Jefferson Chalmers         250         1602.2      0                 1
## 2        Morningside           0         3541.0      0                 1
## 3      Moross-Morang           0         3200.6      0                 1
## 4        Morningside           0         3541.0      0                 1
## 5  Outer Drive-Hayes           0         4247.8      0                 1
## 6        Morningside           0         3541.0      1                 1
##   Binary.Blight.Violation Binary.Building.Permit.Status Sale.Date.Year
## 1                       1                             0           2021
## 2                       1                             0           2014
## 3                       1                             0           2023
## 4                       1                             0           2010
## 5                       1                             0           1987
## 6                       1                             0           2019
##   Taxpayer.City.is.Detroit neighborhood_population normcrime
## 1                        1                    1695 0.9452507
## 2                        0                    3606 0.9819745
## 3                        0                    2610 1.2262835
## 4                        1                    3606 0.9819745
## 5                        1                    3234 1.3134818
## 6                        0                    3606 0.9819745
##   num_vacant_neighborhood
## 1                      49
## 2                     152
## 3                      85
## 4                        152
## 5                        304
## 6                        152
```

Vacancy Distribution

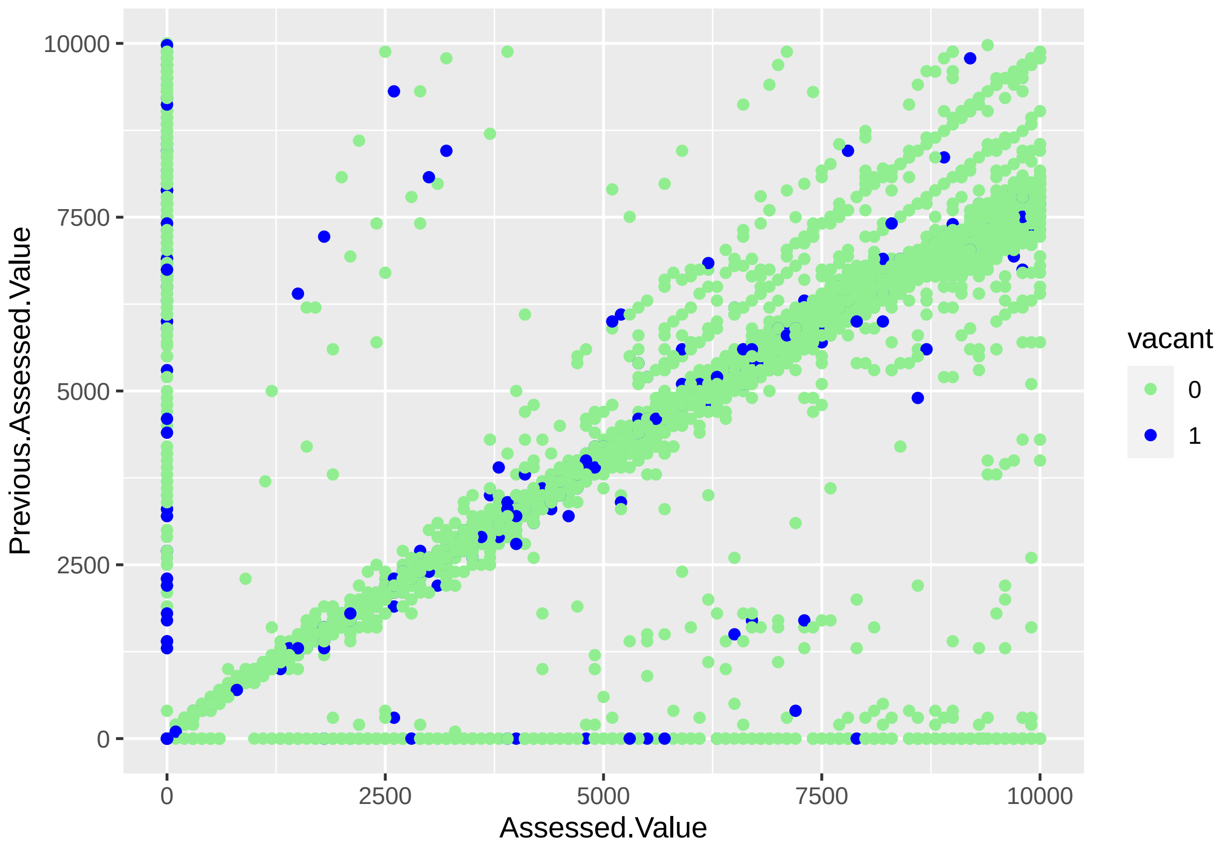
```
library(ggplot2)
ggplot(data) +
  geom_bar(aes(x = vacant))
```



Distribution of assessed value and previous assessed value to show why it can be used as a predictor

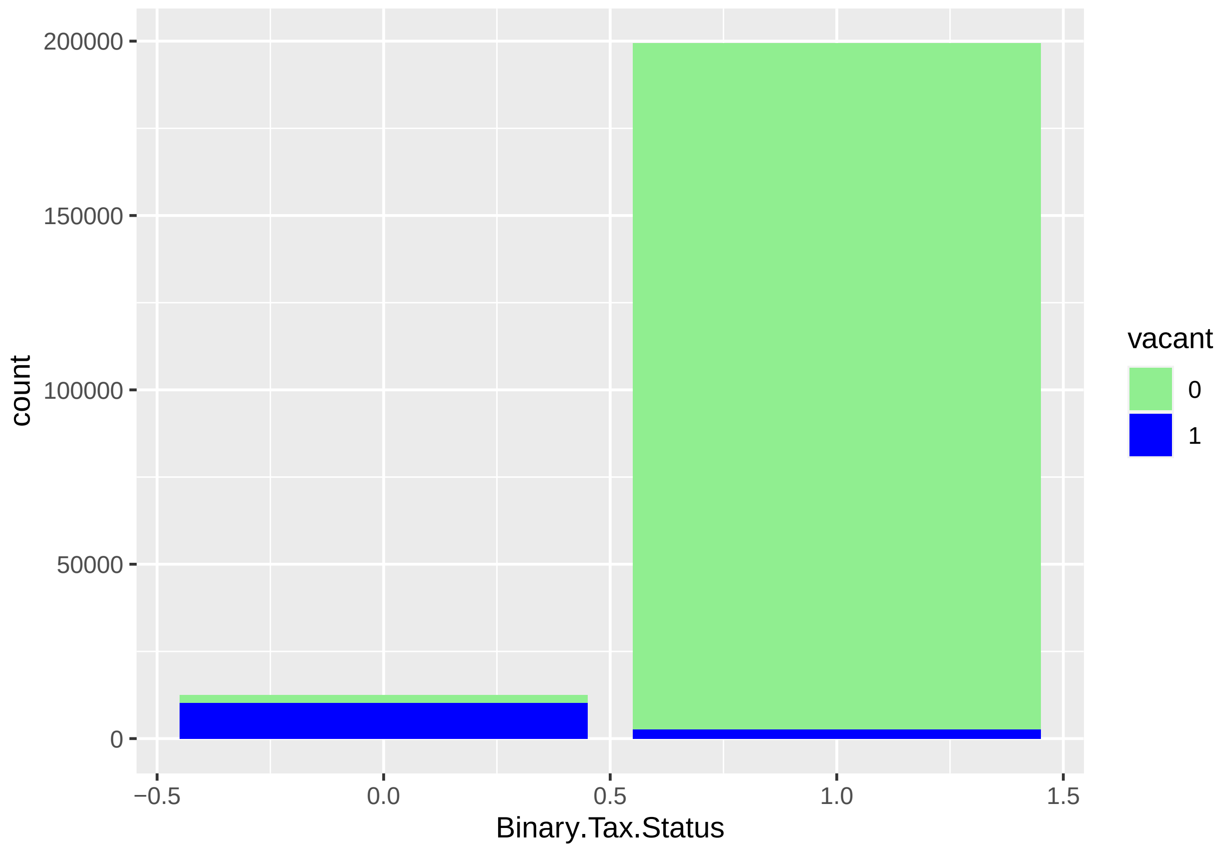
```
ggplot(data)+
  geom_point(aes(x = Assessed.Value, y = Previous.Assessed.Value, col = factor(vacant)))+
  labs(col = "vacant")+
  xlim(0,10000)+
  ylim(0,10000)+
  scale_color_manual(values = c("lightgreen", "blue"))
```

```
## Warning: Removed 185722 rows containing missing values (‘geom_point()‘).
```



Why binary tax status should not be used as a predictor

```
ggplot(data)+
  geom_bar(aes(x = Binary.Tax.Status, fill = factor(vacant)))+
  scale_fill_manual(values = c("lightgreen", "blue"))+
  labs(fill = "vacant")
```



```
data2 <- data[c("Total.Floor.Area",
                "Total.Acreage",
                "Frontage",
                "Depth",
                "Building.Count",
                "Year.Built",
                "Sale.Price",
                "Assessed.Value",
                "Previous.Assessed.Value",
                "Taxable.Value",
                "Previous.Taxable.Value",
                "fine_amount",
                "yearly_average",
                "vacant",
                #"Binary.Tax.Status",
                "Binary.Building.Permit.Status",
                "Sale.Date.Year",
                "Taxpayer.City.is.Detroit",
                "neighborhood_population",
                "normcrime",
                "num_vacant_neighborhood"
                )]
data2[is.na(data2)] <- 0
head(data2)
```

```
##   Total.Floor.Area Total.Acreage Frontage Depth Building.Count Year.Built
## 1             2468         0.079       30   115              1       1916
## 2             1389         0.000        0     0              1       2002
## 3             1293         0.115       40   126              1       1938
## 4             1540         0.105       40   114              1       1929
## 5              920         0.107       40   116              1       1938
## 6             2040         0.105       40   114              1       1929
##   Sale.Price Assessed.Value Previous.Assessed.Value Taxable.Value
## 1      65000          61300                   46200         61300
## 2      13600          40100                   31700         13286
## 3      79992          40400                   33400         40400
## 4       8912          39900                   31300         14391
## 5      20000          20300                   16200          9396
## 6      27900          43700                   34400         23095
##   Previous.Taxable.Value fine_amount yearly_average vacant
## 1                 16361         250         1602.2      0
## 2                 12654           0         3541.0      0
## 3                 15835           0         3200.6      0
## 4                 13706           0         3541.0      0
## 5                  8949           0         4247.8      0
## 6                 21996           0         3541.0      1
##   Binary.Building.Permit.Status Sale.Date.Year Taxpayer.City.is.Detroit
## 1                            0           2021                        1
## 2                            0           2014                        0
## 3                            0           2023                        0
## 4                            0           2010                        1
## 5                            0           1987                        1
## 6                            0           2019                        0
##   neighborhood_population normcrime num_vacant_neighborhood
## 1                    1695 0.9452507                      49
## 2                    3606 0.9819745                     152
## 3                    2610 1.2262835                      85
## 4                    3606 0.9819745                     152
## 5                    3234 1.3134818                     304
## 6                    3606 0.9819745                     152
```

Logistic Regression Model

```
#splitting the data
set.seed(34)
trainsample = sample(1:211865, size = 150000)
train = data2[trainsample,]
test = data2[-trainsample,]

train_model = glm(vacant~., data = train, family = binomial)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

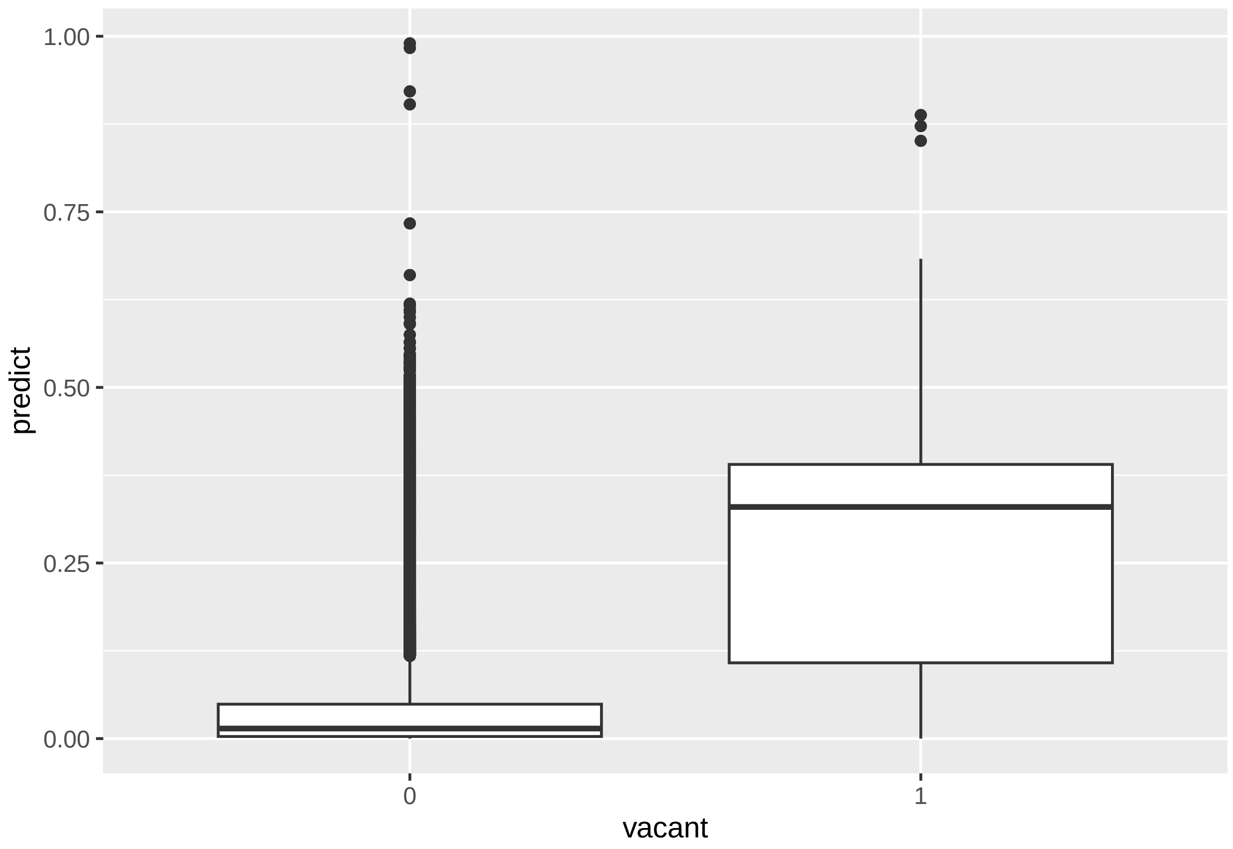
```
summary(train_model)
```

```
##
## Call:
## glm(formula = vacant ~ ., family = binomial, data = train)
##
```

```
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    7.438e-02  5.435e-01   0.137  0.89114
## Total.Floor.Area               2.813e-04  2.321e-05  12.123  < 2e-16 ***
## Total.Acreage                 -8.444e-01  6.771e-01  -1.247  0.21235
## Frontage                      -6.197e-03  2.190e-03  -2.830  0.00466 **
## Depth                          9.568e-04  5.175e-04   1.849  0.06447 .
## Building.Count                -1.682e+00  5.116e-01  -3.287  0.00101 **
## Year.Built                     1.228e-04  8.680e-05   1.415  0.15710
## Sale.Price                     7.082e-07  1.411e-07   5.020 5.18e-07 ***
## Assessed.Value                -9.839e-05  3.445e-06 -28.557  < 2e-16 ***
## Previous.Assessed.Value       -4.277e-05  4.327e-06  -9.884  < 2e-16 ***
## Taxable.Value                  6.070e-05  3.949e-06  15.372  < 2e-16 ***
## Previous.Taxable.Value        -4.210e-05  5.706e-06  -7.378 1.60e-13 ***
## fine_amount                    5.422e-04  6.362e-05   8.523  < 2e-16 ***
## yearly_average                 1.728e-04  1.973e-05   8.757  < 2e-16 ***
## Binary.Building.Permit.Status -1.143e+00  5.930e-02 -19.267  < 2e-16 ***
## Sale.Date.Year                 1.154e-04  1.550e-05   7.446 9.60e-14 ***
## Taxpayer.City.is.Detroit       3.663e-01  4.135e-02   8.858  < 2e-16 ***
## neighborhood_population       -2.802e-04  2.381e-05 -11.768  < 2e-16 ***
## normcrime                      5.406e-03  9.863e-03   0.548  0.58358
## num_vacant_neighborhood        1.124e-03  8.011e-05  14.032  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 68261  on 149999  degrees of freedom
## Residual deviance: 48489  on 149980  degrees of freedom
## AIC: 48529
##
## Number of Fisher Scoring iterations: 9
```

adding the predicted probability to the testing data and why we should use .25 threshold

```r
test.predict <- predict(train_model, newdata = test, type = "response")
#test.predict
test$predict <- predict(train_model, newdata = test, type = "response")
#summary(test$predict)
ggplot(test)+
  geom_boxplot(aes(x = factor(vacant), y = predict))+
  labs(x = "vacant")
```



```
# for 0.5 threshold
table(test$vacant, test.predict > 0.5)
```

```
##
##     FALSE  TRUE
##   0 58030    61
##   1  3706    68
```

```
# overall accuracy
(58030+68)/(211865-150000)
```

```
## [1] 0.9391094
```

```
# predicted inhabited
(58030+3706)/(211865-150000)
```

```
## [1] 0.9979148
```

```
# proportion inhabited
(58030+61)/(211865-150000)
```

```
## [1] 0.9389962
```

```r
# accuracy among vacant houses
68/(3706 + 68)
```

```
## [1] 0.01801802
```

```r
# accuracy among inhabited houses
58030/(58030+61)
```

```
## [1] 0.9989499
```

```r
# predicted vacant
(68+61)/(211865-150000)
```

```
## [1] 0.002085185
```

```r
# for 0.25 threshold
table(test$vacant, test.predict > 0.25)
```

```
##
##     FALSE  TRUE
##   0 55531  2560
##   1  1312  2462
```

```r
# overall accuracy
(55531+2462)/(211865-150000)
```

```
## [1] 0.9374121
```

```r
# predicted inhabited
(55531+1312)/(211865-150000)
```

```
## [1] 0.9188232
```

```r
# proportion inhabited
(55531+2460)/(211865-150000)
```

```
## [1] 0.9373798
```

```r
# accuracy among vacant houses
2462/(1312+2462)
```

```
## [1] 0.6523582
```

```r
# accuracy among inhabited houses
55531/(55531+2560)
```

```
## [1] 0.9559312
```

```
# predicted vacant
(2560+2462)/(211865-150000)
```

```
## [1] 0.08117676
```

```
# proportion vacant
1-((55531+2460)/(211865-150000))
```

```
## [1] 0.06262022
```

```
(0.5)^211865
```

```
## [1] 0
```

Vacancy proportion in entire dataset

```
table(data$vacant)
```

```
##
##      0      1
## 199060  12805
```

```
12805/(199060+12805)
```

```
## [1] 0.06043943
```