Final Project | STAT 419

Group 4

Abby Drongpa, Alex Sullivan, Chris Liu, Jett Palmer, Sumanth Thokala

2025-06-08

Introduction

Our data set is a subset of homes in Cincinnati, Ohio, from the year 2002. It includes 552 observations and seven variables, each representing an individual house. These data are meant to identify factors that influence and may predict the prices of homes.

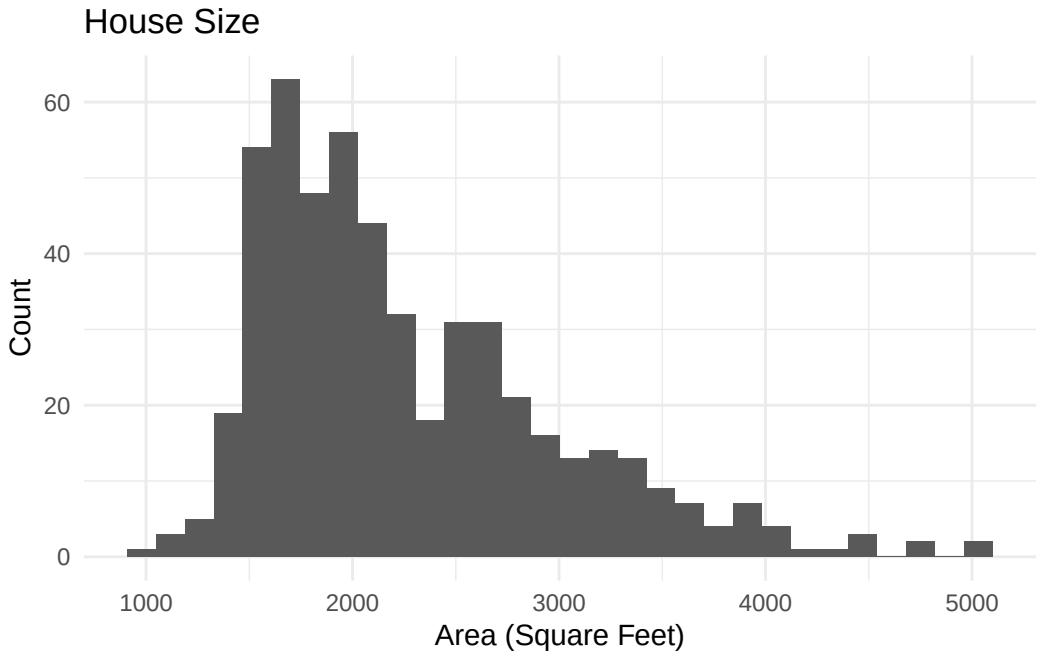
Our ordinal grouping variable, "Pricerank", categorizes homes based on their selling prices. Houses sold for less than $190,000 are coded as 1, those sold between $190,000 and $285,000 are coded as 2, and houses sold for more than $285,000 are coded as 3.

The remaining six variables are explanatory:

- Area: The size of the house in square feet.
- BR: The number of bedrooms.
- BA: The number of bathrooms.
- Garage: The number of cars the garage(s) can accommodate.
- Quality: An index of construction quality, where 1 indicates high quality, 2 indicates medium quality, and 3 indicates low quality.
- Age: The age of the house as of 2002.

Graphs and Summary Statistics

House Area Histogram

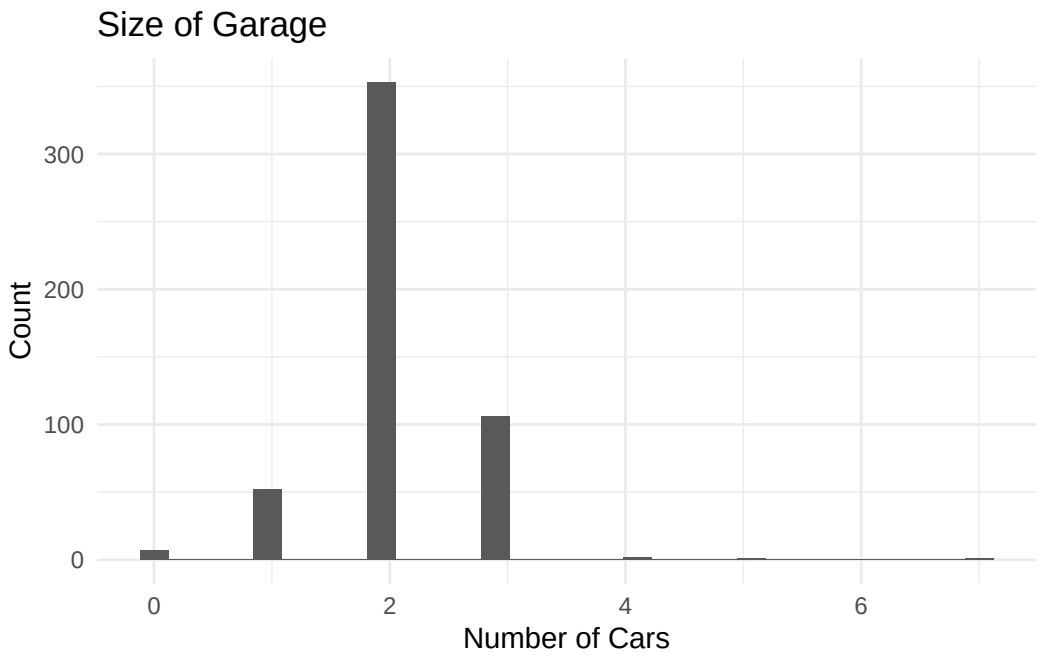



```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 980    1701    2061    2261    2636    5032
```

St Dev
711.066

The distribution of house sizes is right skewed with a mean area of 2261 ft^2 and a median of 2061 ft^2. The standard deviation is 711.06 ft^2.

Car Storage Histogram

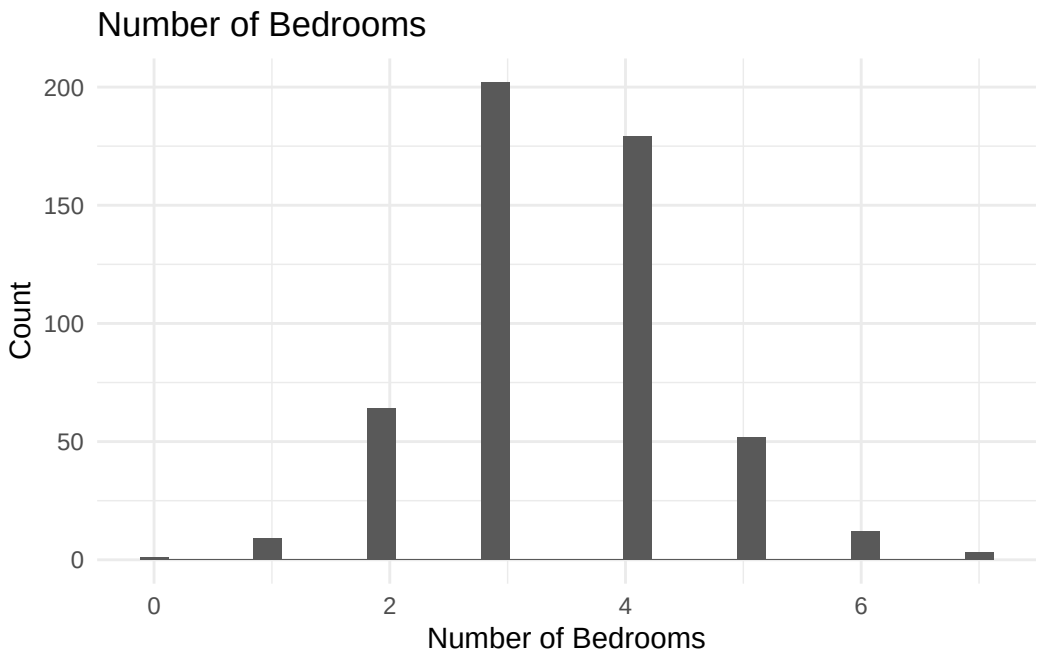


On average, garages accommodate a mean of 2.1 cars and a median of 2 cars, with a standard deviation of 0.65. The distribution reveals several high and low outliers at 0, 4 , 5, 6, and 7 cars. In other words, the distribution is highly concentrated at 2 car garages.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0     2.0     2.0     2.1     2.0     7.0
```

St Dev
0.654

Bedroom Histogram

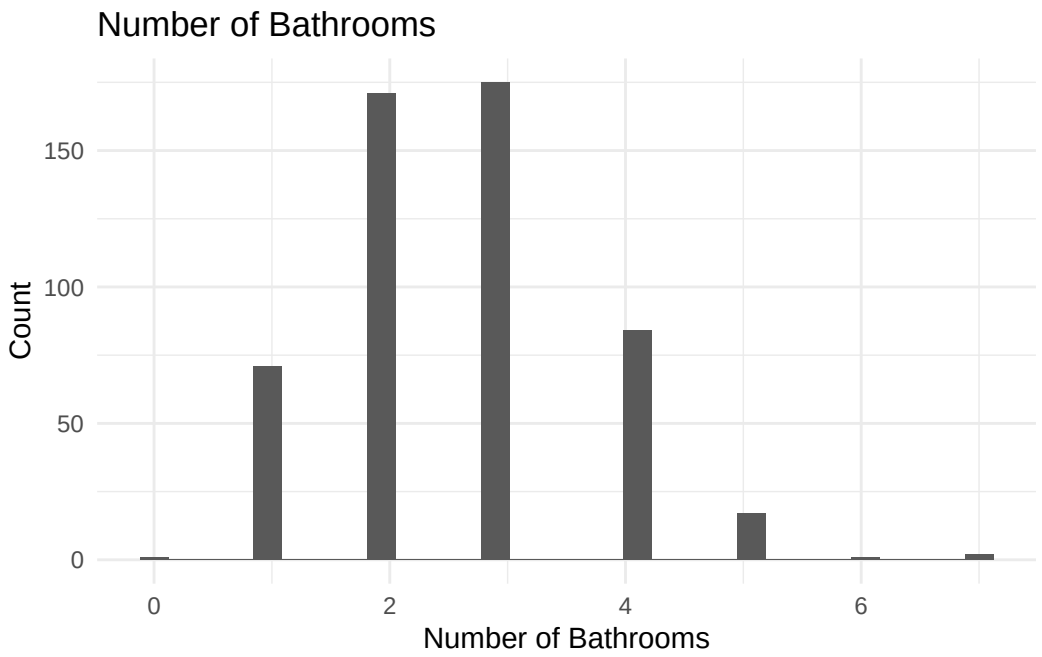


The distribution of number of bedrooms in each house is roughly normal, with a mean of 3.47 and median of 3 bedrooms. The standard deviation is 1.01 bedrooms.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   3.000   3.000   3.471   4.000   7.000
```

St Dev
1.014

Bathrooms Histogram

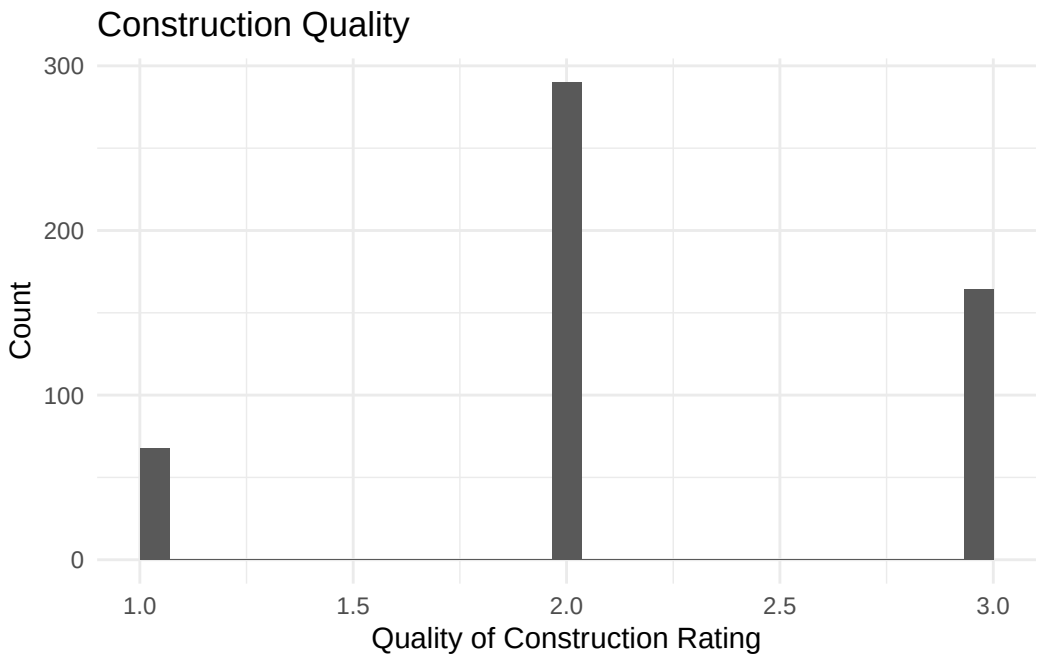


The distribution of number of bathrooms is right skewed with a mean of 2.64 and median of 3 bedrooms. The standard deviation is 1.06.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   2.000   3.000   2.642   3.000   7.000
```

St Dev
1.064

Construction Quality Bar Chart




```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.000   2.000   2.184   3.000   3.000
```

St Dev
0.641

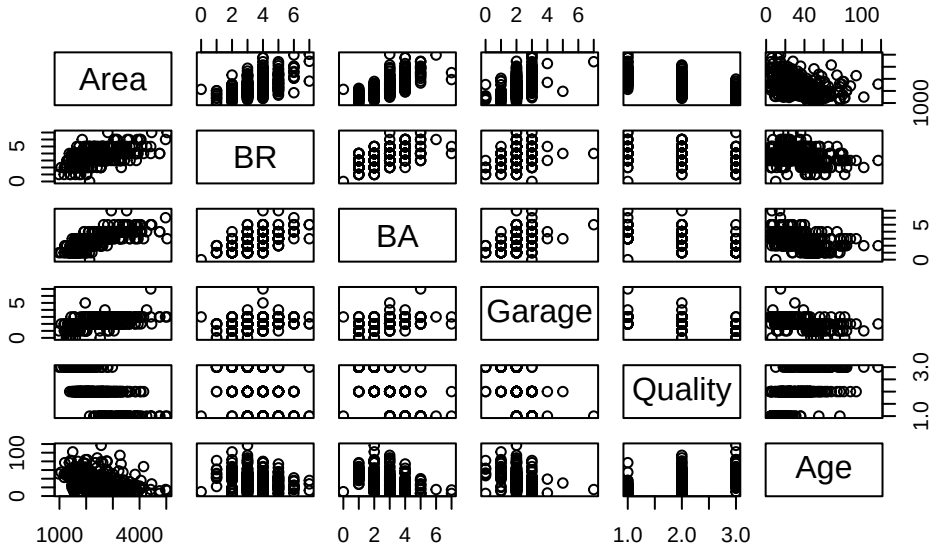
Most houses were given a quality rating of 2 (medium), followed by 3 (low quality), and 1 (high quality).

Discriminant and Classification Analysis

Correlated Quantitative Variables

	Area	BR	BA	Garage	Quality	Age
Area	1.0000000	0.5578378	0.7552729	0.5337665	-0.6955529	-0.4411967
BR	0.5578378	1.0000000	0.5834469	0.3168137	-0.3783218	-0.2686924
BA	0.7552729	0.5834469	1.0000000	0.4898981	-0.6822149	-0.5128410
Garage	0.5337665	0.3168137	0.4898981	1.0000000	-0.5470967	-0.4617604
Quality	-0.6955529	-0.3783218	-0.6822149	-0.5470967	1.0000000	0.6175260

	Area	BR	BA	Garage	Quality	Age
Age	-0.4411967	-0.2686924	-0.5128410	-0.4617604	0.6175260	1.0000000



We will consider correlations above |0.80| to be strong. Provided that no correlations exceed this threshold, we will proceed with all six variables.

Discriminant Analysis

```
            [,1]        [,2]
[1,]  0.70478207 -0.54597681
[2,]  0.04815262  0.43649046
[3,]  0.33024101  0.08605737
[4,]  0.11313689 -0.01035619
[5,] -0.38983060 -0.38524224
[6,] -0.22310228  0.07040947
```

$$Z_1 = 0.705(y_1) + 0.048(y_2) + 0.330(y_3) + 0.113(y_4) - 0.390(y_5) - 0.223(y_6)$$

$$Z_2 = -0.546(y_1) + 0.436(y_2) + 0.086(y_3) - 0.010(y_4) - 0.385(y_5) + 0.070(y_6)$$

Variable	Z1
y1	0.705
y5	-0.390
y3	0.330
y6	-0.223
y4	0.113
y2	0.048

Variable	Z2
y1	-0.546
y2	0.436
y5	-0.385
y3	0.081
y6	0.070
y4	-0.010

The standardized coefficients are ranked above according to their variable contributions to group separation in the presence of all other variables.

H0: $\boldsymbol{\alpha_1} = \boldsymbol{\alpha_2} = \mathbf{0}$

Ha: At least one $\boldsymbol{\alpha_i} \neq \mathbf{0}$ for i = 1, 2

	Lambda	V	P-Values
LD1	0.279	658.852	0
LD2	0.874	69.690	0

Since both p-values are less than 0.025, we have sufficient evidence to reject the null hypothesis and conclude that both discriminant functions are significant.

H0: Variable y_i contributes to group separation after adjusting for the remaining variables.

Ha: Variable y_i does not contribute to group separation after adjusting for the remaining variables.

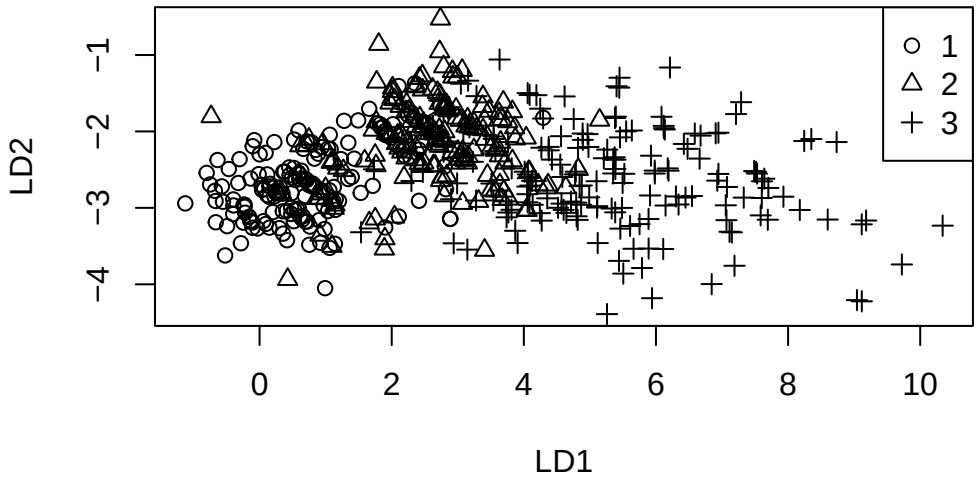
For i = Area, BR, BA, Garage, Quality, Age

	Lambda	F.stat	p.value
Area	0.7833485	71.078759	0.0000000
Quality	0.9157522	23.643594	0.0000000

	Lambda	F.stat	p.value
BR	0.9412469	16.042072	0.0000002
BA	0.9670125	8.766977	0.0001803
Age	0.9816153	4.813347	0.0084903
Garage	0.9952284	1.232173	0.2925182

Since the p-values for Area, Quality, BR, and BA are < 0.0083, at the 5% significance level, we reject the null hypothesis and conclude that these variables significantly contribute to group separation, adjusting for the remaining variables.

Age and Garage, however, do not significantly contribute to group separation, adjusting for the other variables ($p = 0.0085$, $p = 0.2925$).



The plot of the first two linear discriminant functions shows that LD1 creates meaningful separation between the Pricerank groups. Group 1 (circles) and Group 3 (pluses) appear on opposite ends of the LD1 axis, indicating strong discrimination between lower and higher priced homes. Group 2 (triangles), however, overlaps with both groups, particularly along LD2. This suggests that while the first discriminant function captures most of the between group variation, the second function contributes less to distinguishing the groups. Overall, the functions separate the groups moderately well, but the separation isn't perfect, especially between groups 2 and 3.

Classification Analysis

The four most important variables from our discriminant analysis were y1, y2, y5, and y3. We will use these four variables to construct linear classification functions and evaluate classification performance.

Table 6: Linear Classification Function Coefficients (Top 4 Variables)

V1	V2	V3	V4	c0
0.012	0.454	20.273	3.615	-42.143
0.012	1.261	17.783	4.343	-40.120
0.017	0.696	16.957	5.039	-49.057

Each group has its own classification function of the form

$$\text{Score}_k = \boldsymbol{b}_k^\top \boldsymbol{x} + c_{0k}$$

where ($\mathbf{x}$) is the vector of the four selected predictors for an observation, and ($\mathbf{b}k$ *) and (* $c\{0\text{k}\}$) are the corresponding coefficients for group (k).

Classification for Observation #1

```
Predicted Group: 1
```

```
Actual Group: 1
```

```
Correct Prediction: TRUE
```

The classification was accurate for observation #1.

Performance Analysis:

Table 7: Confusion Matrix

1	2	3
135	37	0
28	125	21
3	31	142

```
ACCR: 0.77
```

```
AER: 0.23
```

The linear classification functions based on Area, BR, Quality, and BA achieve an apparent correct classification rate of 77%, which indicates a reasonably strong ability to separate the three Pricerank groups. The confusion matrix shows that most misclassifications occur between similar ranking categories, which is expected.

Summary

As stated in the introduction, this project analysed factors that affect the selling price of homes in 2002 in Cincinnati, Ohio, with each observation in our data set representing a single house. The homes are grouped into three categories of the the Pricerank variable, with homes sold for less than \$190,000 = "1", homes sold for between \$190,000 and \$285,000 = 2, and homes sold for more than \$285,000 = 3, with six explanatory variables to help with classifying a home by Pricerank.

Before performing discriminant analysis, we checked for any strong correlations (i.e., $r \geq \pm 0.8$) between the explanatory variables. No pairs of explanatory variables yielded strong correlations, leading us to an analysis with two discriminant functions. The standardized discriminant function suggests that area, number of bedrooms, number of bathrooms, and overall quality rating contribute the most to separating the groups in the second linear discriminant function.

Performing a classification for Pricerank based on the four previously mentioned variables yields an apparent correct classification rate of 0.77 and an apparent error rate of 0.23. These values indicate a reasonably strong ability to classify house Pricerank by area, number of bedrooms, number of bathrooms, and overall quality rating. Most misclassifications occurred between similar categories and were narrowly mistaken.

Generally speaking, our team did not encounter highly surprising results when grouping houses on Pricerank. However, one variable our team felt would be useful for separation would be location score because homes vary by neighborhood. This location score would be an integer that takes into consideration school district quality, crime rate, proximity to downtown, etc. Homes with similar features can vary significantly in price when taking their neighborhood into consideration.

R Code

Section A

Load dependencies.

```
library(tidyverse)
library(here)
library(janitor)
library(knitr)
source(here("all_custom_functions.R"))
```

Read in data.

```
cincy <- read_csv(here("data", "cincy.csv")) |>
  rename(
    Pricerank = PRICERANK,
    Age = AGE
  )
```

Section B

House area histogram.

```
cincy |> ggplot(aes(x=Area)) + geom_histogram() +
  labs(x="Area (Square Feet)", y = "Count", title="House Size")
```

House area summary statistics.

```
summary(cincy$Area)
round(sd(cincy$Area), 3) |>
  kable(col.names = "St Dev", format = "latex", booktabs = FALSE) |>
  kable_styling(latex_options = c(), position = "left")
```

Car storage histogram.

```
cincy |> ggplot(aes(x=Garage)) + geom_histogram() +
  labs(x="Number of Cars", y = "Count", title="Size of Garage")
```

Car storage summary statistics.

```
summary(cincy$Garage)
round(sd(cincy$Garage), 3) |>
  kable(col.names = "St Dev", format = "latex", booktabs = FALSE) |>
  kable_styling(latex_options = c(), position = "left")
```

Number of bedrooms histogram.

```
cincy |> ggplot(aes(x=BR)) + geom_histogram() +
  labs(x="Number of Bedrooms", y = "Count", title="Number of Bedrooms")
```

Number of bedrooms summary statistics.

```
summary(cincy$BR)
round(sd(cincy$BR), 3) |>
  kable(col.names = "St Dev", format = "latex", booktabs = FALSE) |>
  kable_styling(latex_options = c(), position = "left")
```

Number of bathrooms histogram.

```
cincy |> ggplot(aes(x=BA)) + geom_histogram() +
  labs(x="Number of Bathrooms", y = "Count", title="Number of Bathrooms")
```

Number of bathrooms summary statistics.

```
summary(cincy$BA)
round(sd(cincy$BA), 3) |>
  kable(col.names = "St Dev", format = "latex", booktabs = FALSE) |>
  kable_styling(latex_options = c(), position = "left")
```

Construction quality bar chart.

```
cincy |> ggplot(aes(x=Quality)) + geom_histogram() +
  labs(x="Quality of Construction Rating", y = "Count", title="Construction Quality")
```

Construction quality summary statistics.

```
summary(cincy$Quality)
round(sd(cincy$Quality), 3) |>
  kable(col.names = "St Dev", format = "latex", booktabs = FALSE) |>
  kable_styling(latex_options = c(), position = "left")
```

Section C

Correlation coefficients between all quantitative explanatory variables.

```
kable(cor(cincy))
```

Plot of correlation between all quantitative explanatory variables.

```
plot(cincy[,2:7])
```

Discriminant analysis, standardized functions.

```
discrim(cincy[,-1], cincy$Pricerank)$a.stand
```

Ordered standardized coefficients for both functions, by absolute value from zero, descending.

```
kable(data.frame(Variable = c("y1", "y5", "y3", "y6", "y4", "y2"),
                 Z1 = c(0.70478207,
                        -0.38983060,
                        0.33024101,
                        -0.22310228,
                        0.11313689,
                        0.04815262)),
      digits = 3,
      col.names = c("Variable", "Z1"))

kable(data.frame(Variable = c("y1", "y2", "y5", "y3", "y6", "y4"),
                 Z2 = c(-0.54597681,
                        0.43649046,
                        -0.38524224,
                        0.08065737,
                        0.07040947,
                        -0.01035619)),
      digits = 3,
      col.names = c("Variable", "Z2"))
```

Hypothesis test on discriminant functions.

```
test <- discr.sig(cincy[,-1], cincy$Pricerank)
kable(test, digits = 3, col.names = c("Lambda",
                                      "V",
                                      "P-Values"))
```

Plot of discriminant functions.

```
cincy$Pricerank <- as.factor(cincy$Pricerank)
discr.plot(cincy[, -1], cincy$Pricerank)
```

Classification analysis with top four variables.

```r
top_vars <- c("Area", "BR", "Quality", "BA")

lin_out <- lin.class(cincy[, top_vars], cincy$Pricerank)

coefs <- as.data.frame(lin_out$coefs)
coefs$c0 <- lin_out$c.0
rownames(coefs) <- levels(as.factor(cincy$Pricerank))

kable(coefs, digits = 3, caption = "Linear Classification Function Coefficients (Top 4 Var
```

Classification for the first observation (home) in the data set.

```r
obs1 <- as.numeric(cincy[1, top_vars])
group_names <- rownames(coefs)

scores <- apply(coefs[, -ncol(coefs)], 1, function(b) sum(b * obs1)) + coefs$c0
predicted <- group_names[which.max(scores)]
actual <- as.character(cincy$Pricerank[1])

cat("Predicted Group:", predicted, "\n")
cat("Actual Group:", actual, "\n")
cat("Correct Prediction:", predicted == actual, "\n")
```

Performance analysis of classification.

```r
class_results <- rates(cincy[, top_vars], cincy$Pricerank, method = "l")

kable(class_results$`Confusion Matrix`, caption = "Confusion Matrix")

cat("ACCR:", round(class_results$`Correct Class Rate`, 3), "\n")
cat("AER:", round(class_results$`Error Rate`, 3), "\n")
```