STA 135 Project 1 - Exploratory Analysis

Grant Gambetta

1/25/2022

Introduction

For this project, the goal was to perform exploratory analysis on a dataset consisting of financial information among bankrupt firms and non-bankrupt firms. Exploratory analysis of the data was conducted using a varierty of graphical and numerical techniques such as sample mean vectors, correlation matrices, scatterplots, boxplots, and histograms. The main questions that this exploratory analysis aims to answer are whether any variables were highly correlated, whether any of the variables had a discriminating effect, whether the variables followed a normal distribution, and whether there were any variables that contained outliers. Additionally, the data visualizations were useful in providing high level insights about the data.

This report begins with the introduction and then proceeds to the materials and results section, where I discuss the variables in the dataset, the visualization, numerical, and statistical methods that were used, and the exploratory analysis results. Within the methods and results section, the three subsections are titled "Dataset Overview," "Methods Used," and "Results of Exploratory Analysis." The report concludes with the appendix where the full R code can be found.

Materials and Results

Dataset Overview

To begin with, it is important to understand the dataset behind this project. The dataset consisted of five variables and 46 rows, where four of the variables were numeric/continuous and the fifth was a binary categorical variable. The four continuous variables were denoted as X_1, X_2, X_3, and X_4 and the factor variable represented the population, specifically whether each firm was bankrupt or non-bankrupt (0 = bankrupt, 1 = non-bankrupt). There were 21 bankrupt firms and 25 non-bankrupt firms in the dataset. Table 1 provides the calculations for each continuous variable.

Table 1: Calculations for continuous variables.

	Calculation
X_1	$\frac{\text{cash flow}}{\text{total debt}}$
X_2	$\frac{\text{net income}}{\text{total assets}}$
X_3	$\frac{\text{current assets}}{\text{current liability}}$
X_4	$\frac{\text{current assets}}{\text{net sales}}$

Methods Used

To conduct the exploratory analysis for this project, several visualization, statistical, and numerical techniques were used. First, the sample mean vector and sample correlation matrices for each population were obtained with the `colMeans()` and `cor()` functions, respectively. For the visualizations, I began by creating pairwise scatterplots of all the continuous variables using the `pairs()` function and color coded the data points by population. Also, I created side-by-side histograms and normal Q-Q plots of the continuous variables for both populations to determine if they followed a normal distribution. To statistically verify whether the variables were normally distributed, I carryed out a Shapiro-Wilk test for each variable using the `shapiro.test()` function. Lastly, one important part of this exploratory analysis was determining if any of the variables had a discriminating effect based on each population and if the variables had outliers, so I created side by side boxplots to answer these questions.

Results of Exploratory Analysis

Summary Statistics and Correlation Analysis

To begin with, summary statistics such as the sample mean vector and sample correlation matrix provide high level insights about the variables in each population. The sample mean vectors allow us to understand differences in the means of all continuous variables and the correlation matrices show us the direction and strength of linear relationship between the continuous variables. The sample mean vectors and sample correlation matrices for bankrupt firms and non-bankrupt firms can be found in tables 2-5.

Table 2: Sample mean vector for bankrupt firms.

X_1	X_2	X_3	X_4
-0.069	-0.081	1.367	0.438

Table 3: Sample mean vector for non-bankrupt firms.

X_1	X_2	X_3	X_4
0.235	0.056	2.594	0.427

Table 4: Sample correlation matrix for bankrupt firms.

	X_1	X_2	X_3	X_4
X_1	1	0.935	0.405	0.094
X_2	0.935	1	0.443	0.112
X_3	0.405	0.443	1	0.383
X_4	0.094	0.112	0.383	1

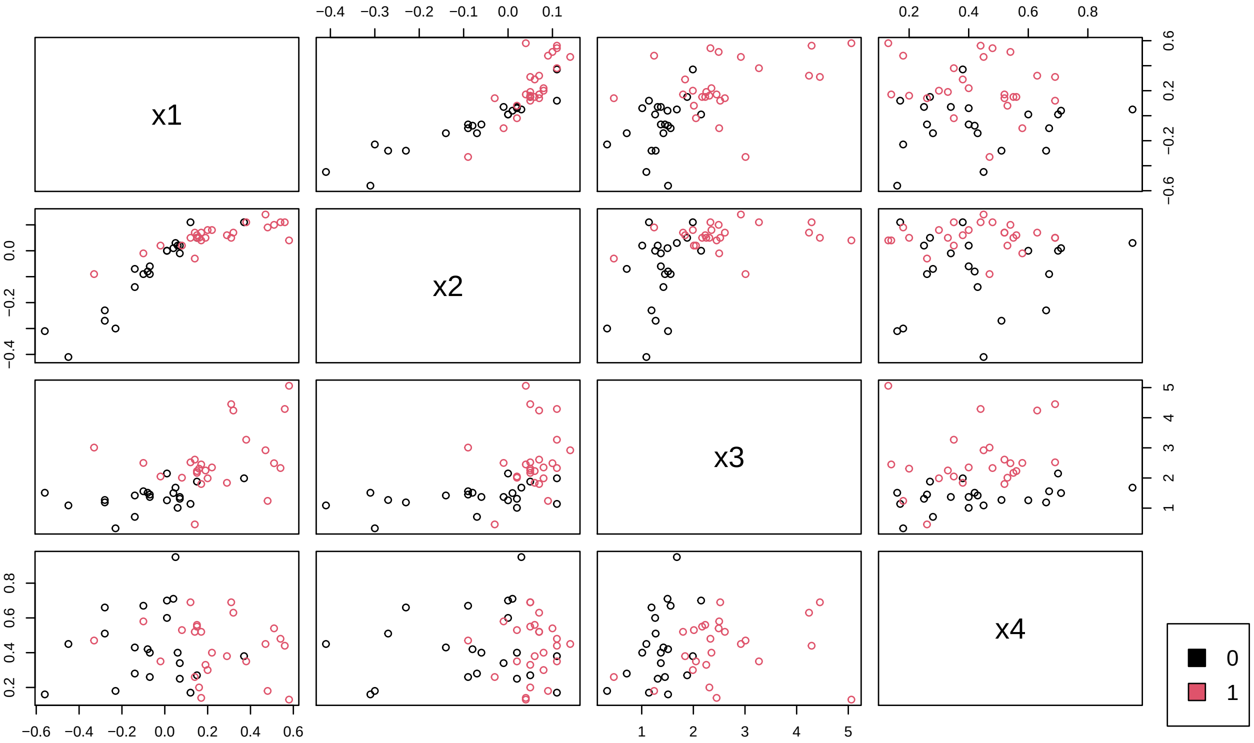
Table 5: Sample correlation matrix for non-bankrupt firms.

	X_1	X_2	X_3	X_4
X_1	1	0.805	0.338	-0.190
X_2	0.805	1	0.172	0.023
X_3	0.338	0.172	1	0.196
X_4	-0.190	0.023	0.196	1

Comparing the sample mean vectors for both populations, we notice that X_1, X_2, and X_3 for non-bankrupt firms have a larger mean than those variables do for bankrupt firms, and the mean for X_4 is very similar for both populations. Observing the sample correlation matrix for bankrupt firms, we see that X_1 and X_2 have a high correlation of 0.935 while the rest of the pairs have medium to low correlation. Observing the sample correlation matrix for non-bankrupt firms, we also notice that the correlation between X_1 and X_2 is rather high at 0.805, which means that in general, X_1 and X_2 are highly correlated.

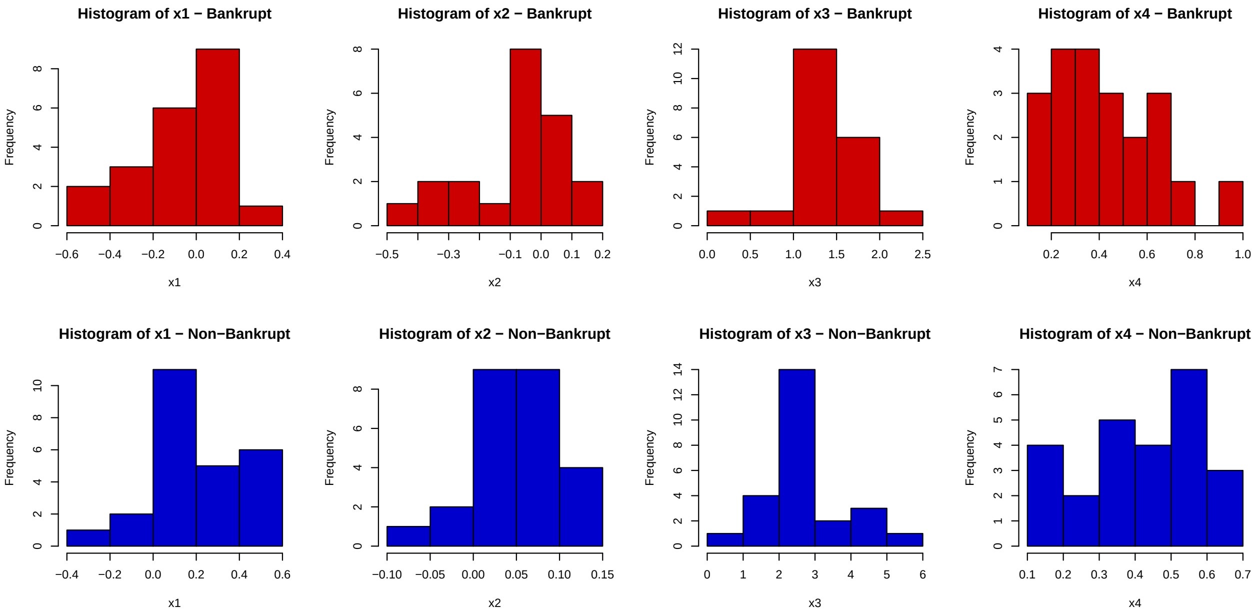
Pairwise Scatterplot and Side-by-Side Histograms

Next, a pairwise scatterplot and side-by-side histograms were created to observe trends in the data. First, the pairwise scatterplot provides insights about the relationships that are present between the continuous variables. In the pairwise scatterplot below, the black data points represent bankrupt firms (population = 0) and the red data points represent non-bankrupt firms (population = 1).



From looking at the pairwise scatterplot, one of the main trends I notice is that primarily for plots involving X_1, X_2, and X_3, the red data points (non-bankrupt firms) tend to be larger than the black data points (bankrupt firms). Also, looking at the scatterplot for X_1 and X_2, we see that the data has a clear positive relationship which supports the high correlation coefficients for these two variables. Additionally, many of the scatterplots are simply a "cluster" of data points which means that there is little to no relationship between certain variables.

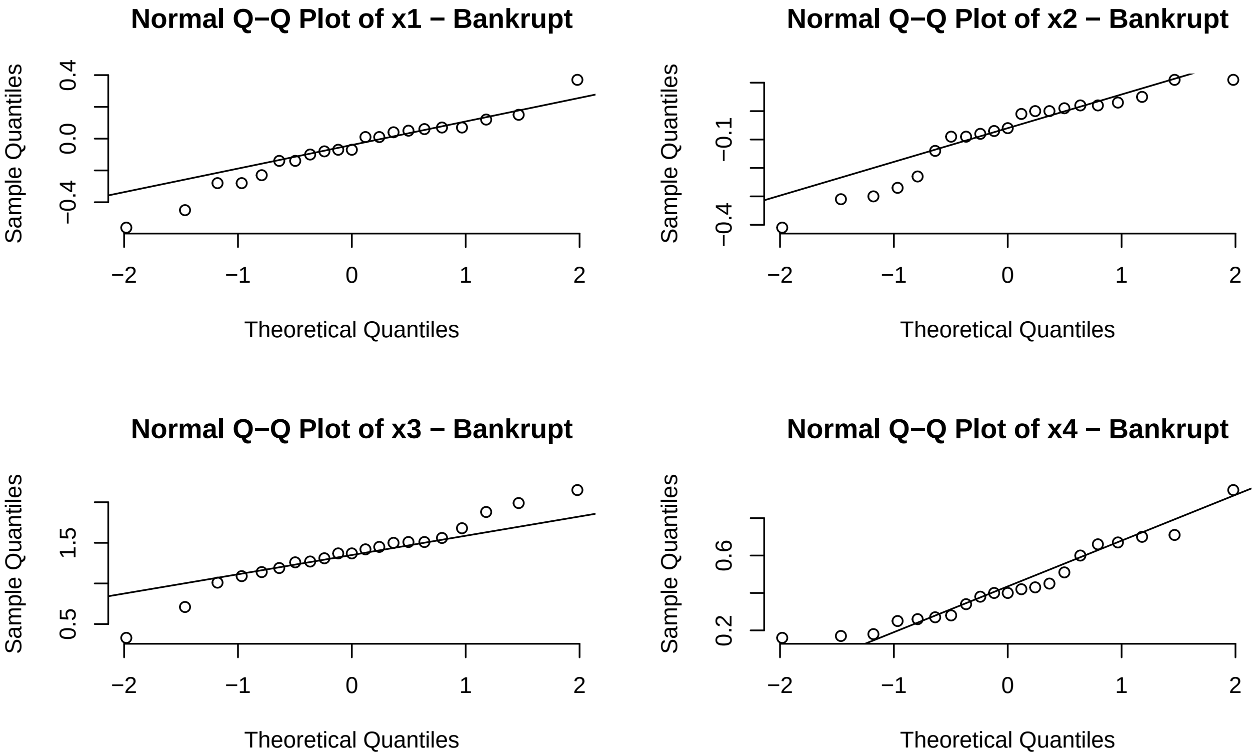
The side-by-side histograms show us the distribution of the data and are useful in gaining an initial insight into whether the variables are normally distributed. The histograms I created are below and for simplicity, the red histograms represent bankrupt firms and the blue histograms represent non-bankrupt firms.



After observing the histograms, it is difficult to determine whether the variables in both populations follow a normal distribution because all of the histograms do not have a clear bell shape. Further analysis of normality is done using normal Q-Q plots and the Shapiro-Wilk test in the next section.

Marginal Normality of the Variables

To check the marginal normality of the variables in each population, I created normal Q-Q plots and utilized the Shapiro-Wilk normality test. For the normal Q-Q plots, the data is considered normal if the points fall within a straight line.

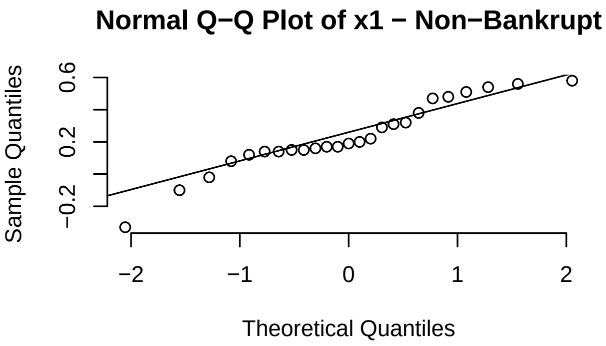


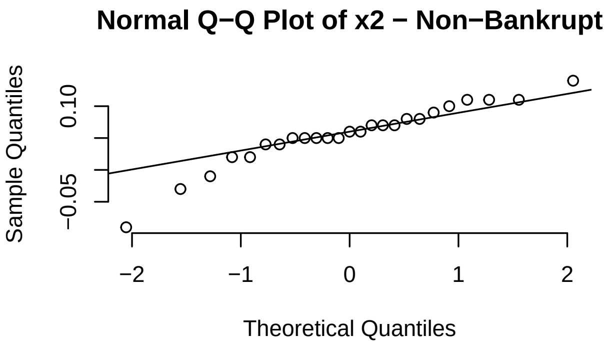
Observing the normal Q-Q plots for bankrupt firms (population = 0), we see that X_1 and X_2 look a bit left skewed, X_3 looks a bit light tailed, and X_4 looks a bit right skewed but not far from normal. To statistically support the normality conclusions from the normal Q-Q plots for bankrupt firms, I conducted Shapiro-Wilk tests for each variable among bankrupt firms. The null hypothesis (H_0) for the Shapiro-Wilk test is that the variable is normally distributed and if the p-value is less than the significance level (α), we reject the null hypothesis. Below is the table of p-values from the Shapiro-Wilk tests for each variable among bankrupt firms.

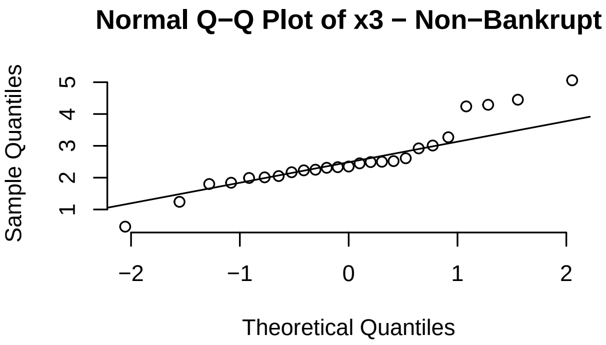
Table 6: P-values from the Shapiro-Wilk test for bankrupt firms.

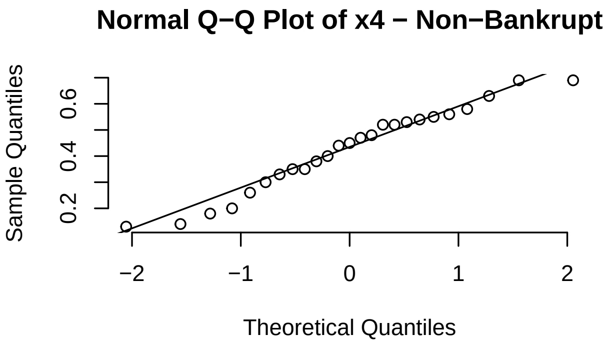
	x1	x2	x3	x4
p.value	0.480	0.057	0.506	0.192

Testing H_0 at $\alpha = 0.1$, we can statistically conclude that X_1, X_3, and X_4 for bankrupt firms are normally distributed since their p-values are greater than 0.1. Observing the normal Q-Q plots for non-bankrupt firms (population = 1), X_1 and X_2 look a bit left skewed, while X_3 looks a bit light tailed, and X_4 looks close to normal. It is important once again to run Shapiro-Wilk tests to statistically verify whether the variables are normally distributed.









Below is the table of p-values from the Shapiro-Wilk tests for each variable among non-bankrupt firms. Similar to the tests earlier, the null hypothesis (H_0) is that the variable is normally distributed.

Table 7: P-values from the Shapiro-Wilk test for non-bankrupt firms.

	x1	x2	x3	x4
p.value	0.162	0.063	0.027	0.443

Testing H_0 at $\alpha = 0.1$, we can statistically conclude that X_1 and X_4 for non-bankrupt firms are normally distributed since their p-values are greater than 0.1.

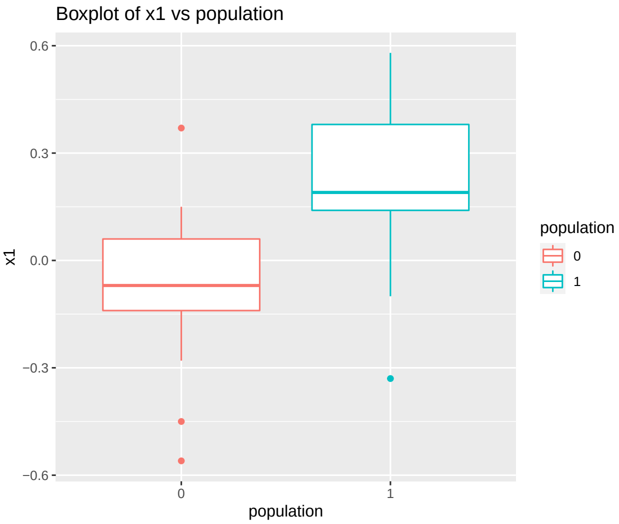
Discriminating Effects and Outlier Detection

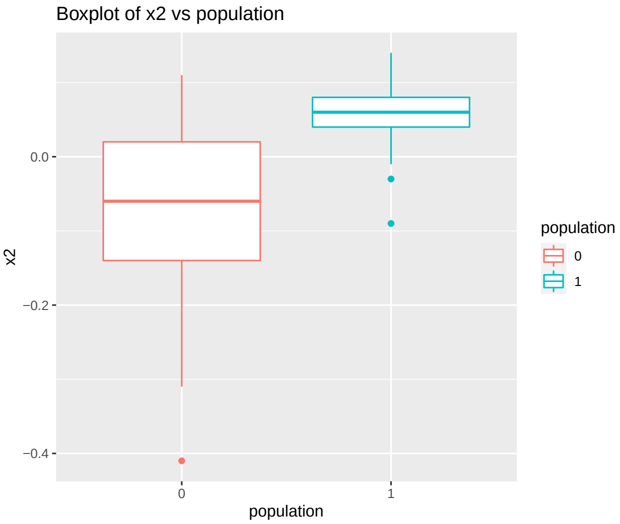
Lastly, to analyze the data for discriminating effects and outliers, I created side-by-side boxplots for all the continuous variables. Looking at the boxplots for X_1, it appears that the values for population 1 (non-bankrupt firms) are larger than the values for population 0 (bankrupt firms), which implies that there is a discriminating effect for X_1. Additionally, there appear to be three outliers for population 0 and one outlier for population 1.

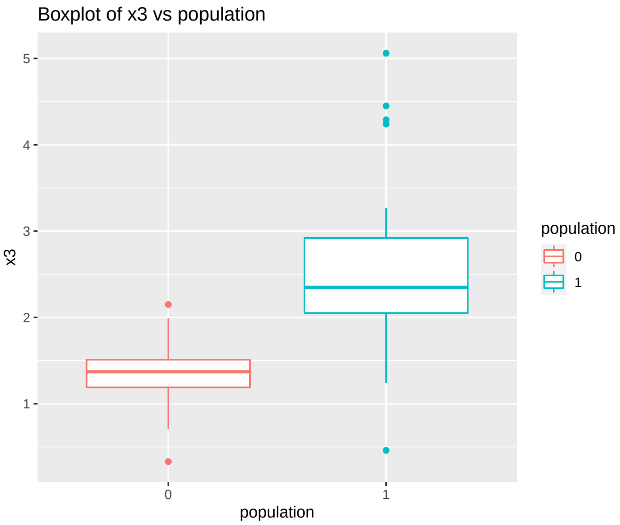
Looking at the boxplots for X_2, it appears that the values for population 1 (non-bankrupt firms) are larger than the values for population 0 (bankrupt firms), which implies that a discriminating effect is present for X_2 since the values of population 1 are larger than the values of population 0. There appears to be one outlier for population 0 and two outliers for population 1.

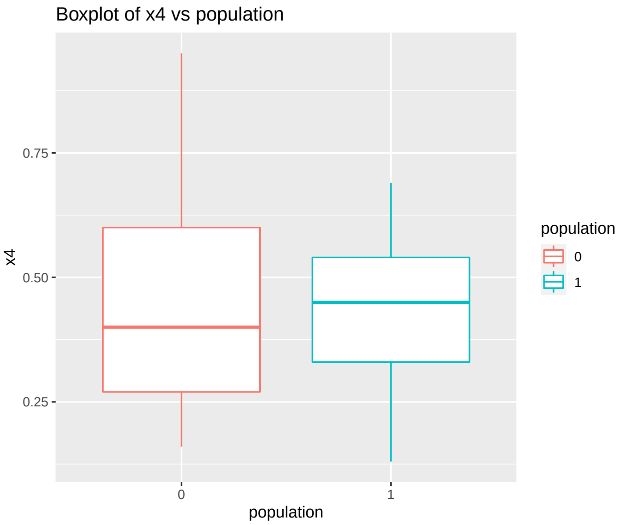
Looking at the boxplots for X_3, it also appears that the values for population 1 (non-bankrupt firms) are larger than the values for population 0 (bankrupt firms), which means there is a dicriminating effect for this variable. Additionally, there appear to be two outliers for population 0 and five outliers for population 1.

Looking at the boxplots for X_4, we are unable to conclude if there is a discriminating effect because the IQR for population 0 (bankrupt firms) is larger than the IQR for population 1 (non-bankrupt firms) and the values of both populations significantly overlap. This implies that the values for both populations could be roughly equal and no major difference exists between the values. Additionally, there appear to no outliers for both populations in X_4.


Boxplot of x1 vs population


Boxplot of x2 vs population


Boxplot of x3 vs population


Boxplot of x4 vs population

Appendix

```
library(tidyr)
library(dplyr)
library(ggplot2)
library(ggpubr)
library(stargazer)

# read in the data and change column data types
data <- read.table('Project_1_Data.txt', header = T)
data[c('x1', 'x2', 'x3', 'x4')] <- sapply(data %>% select('x1', 'x2', 'x3', 'x4'), as.numeric)
data$population <- as.factor(data$population)
head(data)

# create seperate dataframes for the two populations
bankrupt <- data %>% filter(population == 0)
non_bankrupt <- data %>% filter(population == 1)

# sample mean vectors
stargazer(colMeans(bankrupt %>% select(-population)))
stargazer(colMeans(non_bankrupt %>% select(-population)))

# sample correlation matrices
stargazer(cor(bankrupt %>% select(-population)))
stargazer(cor(non_bankrupt %>% select(-population)))

# pairwise scatterplots color coded by population
pairs(data[,1:4], col = data$population, oma = c(3,3,3,8))
par(xpd = TRUE)
legend('bottomright', fill = unique(data$population), legend = c(levels(data$population)))
```

```
# side by side histograms for bankrupt
par(mfrow=c(2,4))
for (col in colnames(bankrupt)[1:4]) {
  hist(bankrupt[[col]], xlab = col, main = sprintf('Histogram of %s - Bankrupt', col))
}

# side by side histograms for non-bankrupt
for (col in colnames(non_bankrupt)[1:4]) {
  hist(non_bankrupt[[col]], xlab = col, main = sprintf("Histogram of %s - Non-Bankrupt", col))
}
```

```
# normal Q-Q plots for bankrupt
par(mfrow=c(2,2))
for(col in colnames(bankrupt[,1:4])) {
  qqnorm(bankrupt[,col], frame = F, main = sprintf('Normal Q-Q Plot of %s - Bankrupt', col))
  qqline(bankrupt[,col])
}
```

```
# shapiro-wilk tests for normality - bankrupt
SW_pvalue_bankrupt <- matrix(rep(NA, 4), nrow = 1, dimnames = list(c('p.value'), colnames(bankrupt)[1:4]
for(i in 1:4) {
  SW_pvalue_bankrupt[i] <- shapiro.test(bankrupt[,i])$p.value
}
stargazer(SW_pvalue_bankrupt)
```

```r
# normal Q-Q plots for non-bankrupt
par(mfrow=c(2,2))
for(col in colnames(non_bankrupt[,1:4])) {
  qqnorm(non_bankrupt[,col], frame = F, main = sprintf('Normal Q-Q Plot of %s - Non-Bankrupt', col))
  qqline(non_bankrupt[,col])
}
```

```r
# shapiro-wilk tests for normality - non bankrupt
SW_pvalue_nonBankrupt <- matrix(rep(NA, 4), nrow = 1, dimnames = list(c('p.value'), colnames(non_bankrup
for(i in 1:4) {
  SW_pvalue_nonBankrupt[i] <- shapiro.test(non_bankrupt[,i])$p.value
}
stargazer(SW_pvalue_nonBankrupt)

# boxplots for outlier detection
p1 <- ggplot(data = data, aes(population, x1)) +
  geom_boxplot(col = 'blue') +
  ggtitle('Boxplot of x1 vs population')

p2 <- ggplot(data = data, aes(population, x2)) +
  geom_boxplot(col = 'red') +
  ggtitle('Boxplot of x2 vs population')

p3 <- ggplot(data = data, aes(population, x3)) +
  geom_boxplot(col = 'darkgreen') +
  ggtitle('Boxplot of x3 vs population')

p4 <- ggplot(data = data, aes(population, x4)) +
  geom_boxplot(col = 'gold4') +
  ggtitle('Boxplot of x4 vs population')

ggarrange(p1, p2, p3, p4)
```