Project 2: Biodiesel

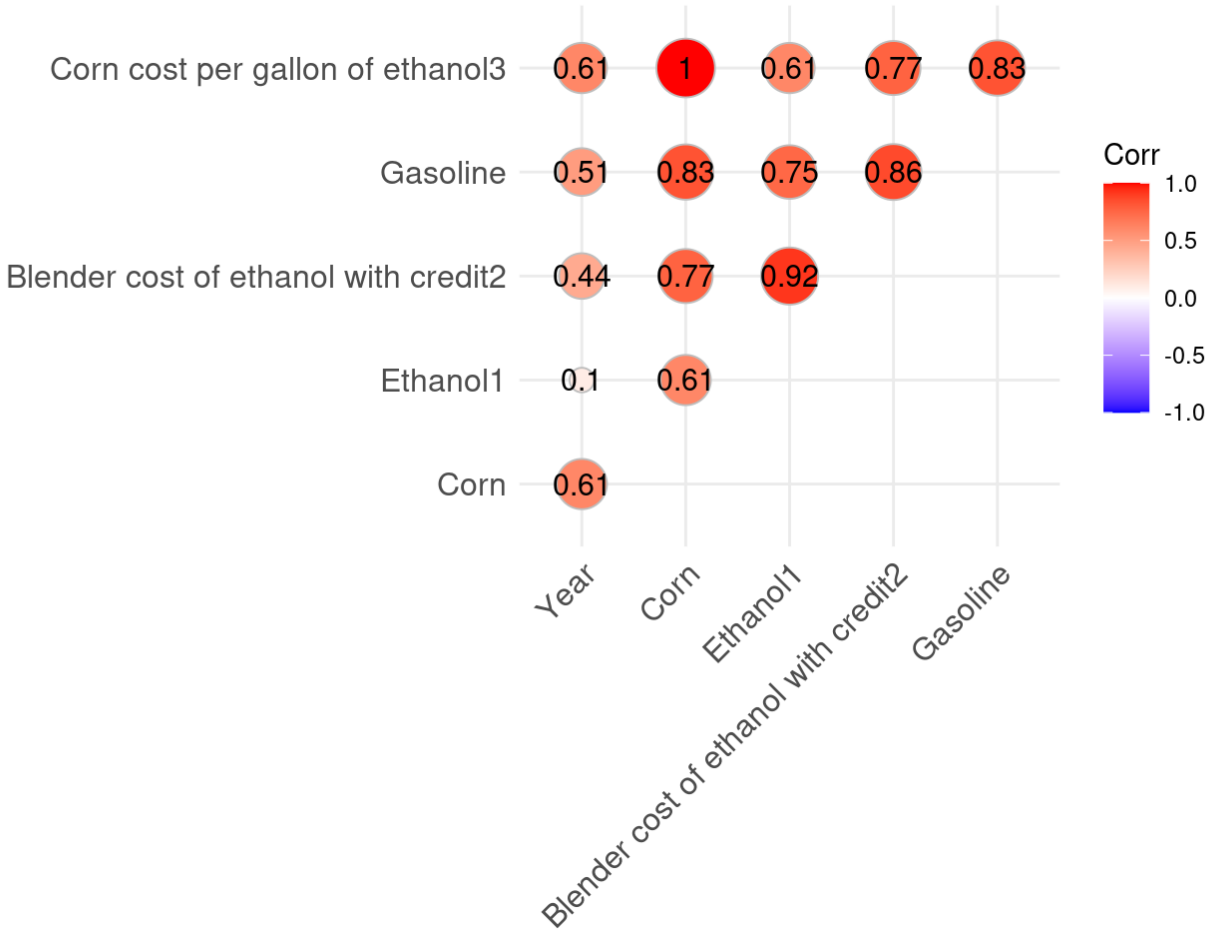
Introduction

The `biodiesel` dataset is U.S. Bioenergy Statistics found on the USDA's Economic Research Service (ERS) website. To develop this dataset, the ERS combined data from the U.S. Department of Energy and USDA. I chose to examine this dataset because I am interested in the pricing of biofuels and how the cost of corn, ethanol, and gasoline affect each other if at all. Each row represents data from each month during the years 2000 to 2023. The variables include the year, the price of corn (dollars per bushel), the price of ethanol (dollars per gallon), the blender cost of ethanol, the price of gasoline (dollars per gallon), and the corn cost per gallon of ethanol. Since 10% of gasoline is made with ethanol and ethanol is typically made from corn biomass, I expect a positive relationship between the cost of corn, ethanol, and gasoline. Each variable has its own column and each observation has its own row, so the dataset is considered to be already tidy. In this project, I will explore the following research question: How does the various pricing variables of corn and ethanol affect the cost of gasoline?

```
# load dataset
biodiesel <- read_csv("~/sds 322e/biodiesel.csv", show_col_types = FALSE)

# filter for numeric variables
biodiesel <- biodiesel %>%
  select(where(is.numeric)) %>%
  select(-ends_with('geg4'))
```

Exploratory Data Analysis

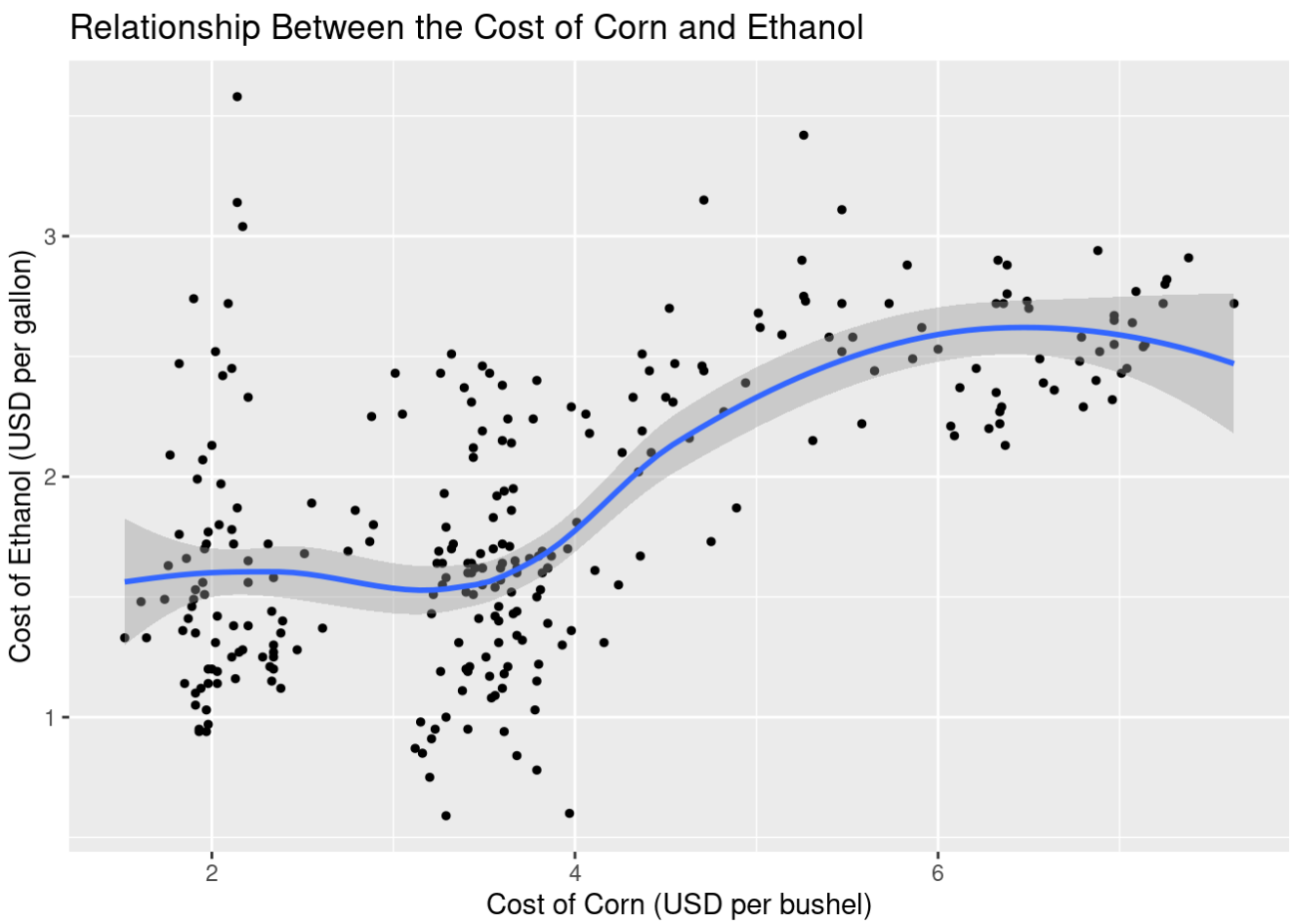
```
# create a correlation matrix
ggcorrplot(cor(biodiesel),
           type = "upper", # upper diagonal
           lab = TRUE, # print values
           method = "circle")
```



The correlation matrix values for this dataset ranges from as low as 0.1 to as high as 1.0. The variables with the greatest correlation are the cost of corn and the cost of corn per gallon of ethanol. The variables with the lowest correlation are the year and the corresponding cost of ethanol (dollars per gallon).

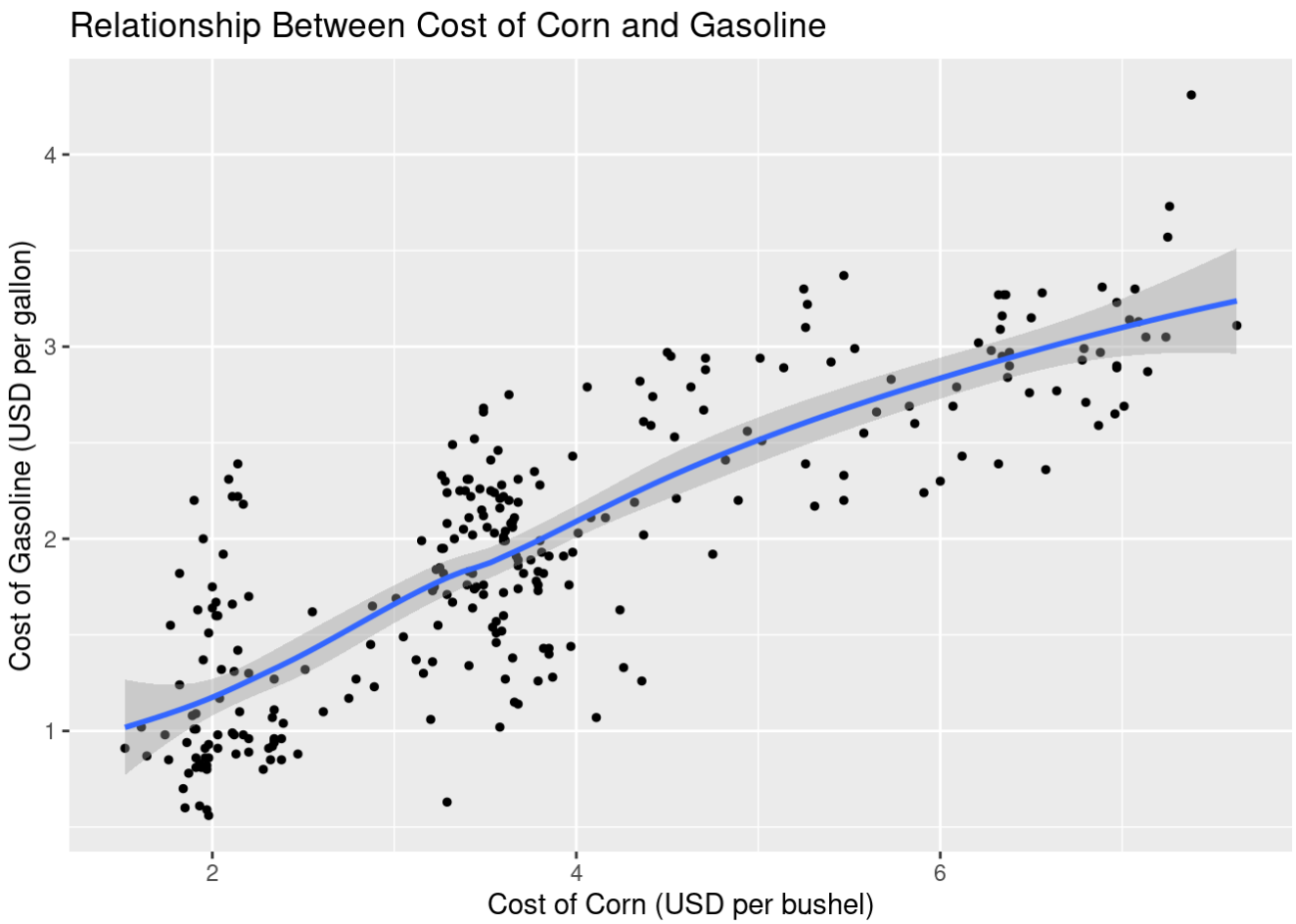
Investigate Relationships between Variables

```
# visualize corn and ethanol cost
biodiesel %>%
  ggplot(aes(x=Corn, y=Ethanol1)) +
  geom_point(size=1) +
  geom_smooth() +
  labs(x="Cost of Corn (USD per bushel)", y="Cost of Ethanol (USD per gallon)",title="Rela
tionship Between the Cost of Corn and Ethanol")
```



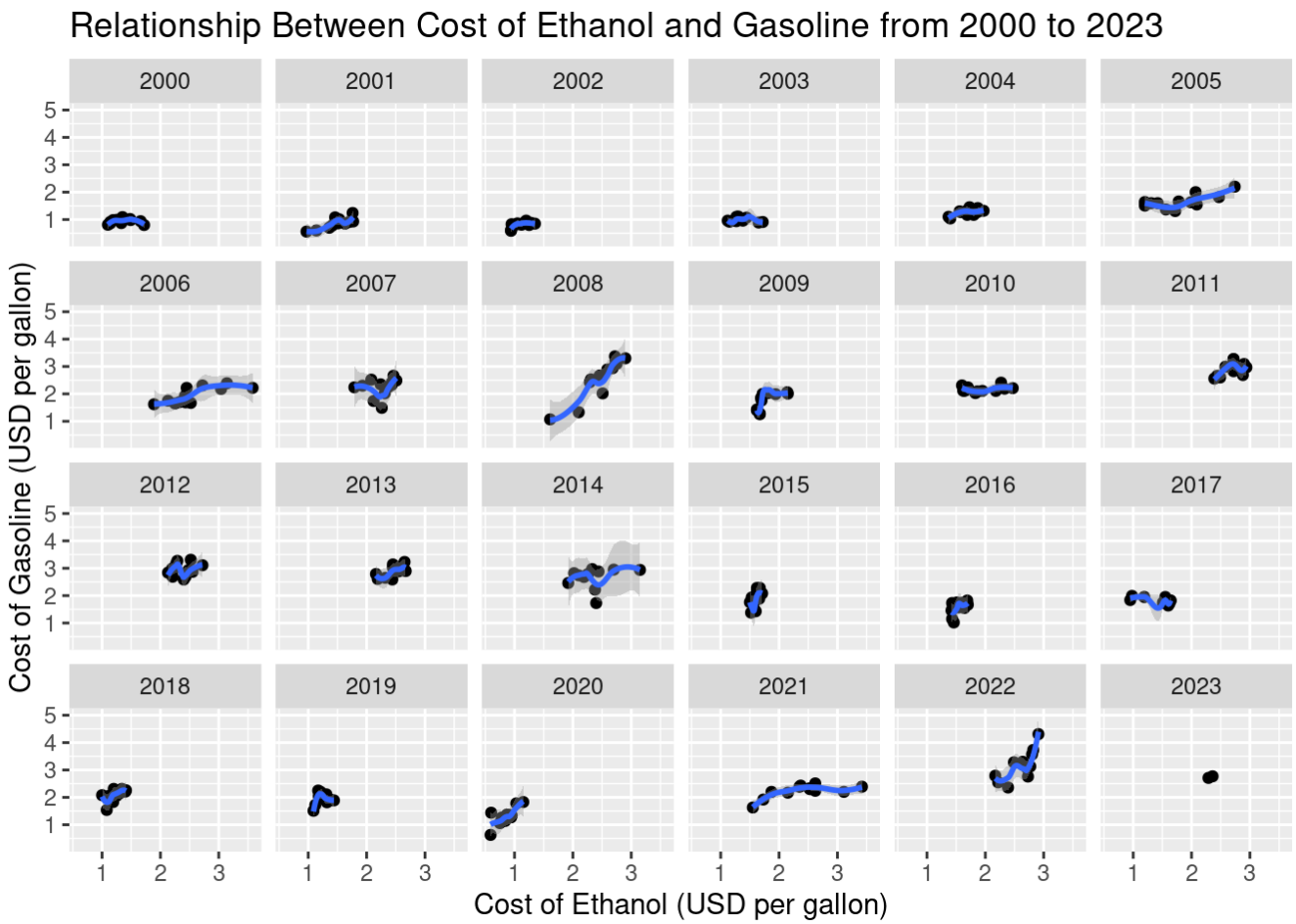
The scatterplot depicts the relationship between the cost of corn (U.S. dollars per bushel) and the cost of Ethanol (U.S. dollars per gallon). The cost of corn and ethanol largely have a positive relationship in that the greater the cost of corn, the greater the cost of Ethanol is. There is a plateau in the cost of Ethanol when the cost of corn is $2.50/bushel or less. There is a brief negative relationship when the cost of corn is around $2.50 to $3.30/bushel. From $3.30/bushel, there is a positive relationship between the cost of corn and Ethanol until it turns into a negative relationship around $7/bushel of corn.

```
# visualize corn and gasoline cost
biodiesel %>%
  ggplot(aes(x=Corn, y=Gasoline)) +
  geom_point(size=1) +
  geom_smooth() +
  labs(x="Cost of Corn (USD per bushel)", y="Cost of Gasoline (USD per gallon)", title="Re
lationship Between Cost of Corn and Gasoline")
```



The scatterplot depicts the relationship between the cost of corn (U.S. dollars per bushel) and the cost of gasoline (U.S. dollars per gallon). The cost of corn and gasoline largely have a positive relationship in that the greater the cost of corn, the greater the cost of gasoline is.

```
# visualize ethanol and gasoline cost
biodiesel %>%
  ggplot(aes(x=Ethanol1, y=Gasoline)) +
  geom_point() +
  facet_wrap(~Year, ncol=6) +
  geom_smooth() +
  labs(x="Cost of Ethanol (USD per gallon)", y="Cost of Gasoline (USD per gallon)",title
="Relationship Between Cost of Ethanol and Gasoline from 2000 to 2023")
```



The scatterplots depict the relationship between the cost of Ethanol (U.S. dollars per gallon) and the cost of gasoline (U.S. dollars per gallon). Each scatterplot represents a different year from the years 2000 to 2023. The cost of ethanol and gasoline largely have a positive relationship in that the greater the cost of ethanol, the greater the cost of gasoline is.

Prediction and Cross-Validation

Linear Regression

```
# linear regression
bio_lin <- lm(Gasoline ~ ., data=biodiesel)
summary(bio_lin)
```

```
##
## Call:
## lm(formula = Gasoline ~ ., data = biodiesel)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.00604 -0.21668 -0.02743  0.24627  0.75876
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        -37.416163  14.311631  -2.614  0.00944
## Year                                 0.018667   0.007106   2.627  0.00911
##
## (Intercept)                        **
## Year                               **
##  [ reached getOption("max.print") -- omitted 4 rows ]
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3313 on 272 degrees of freedom
## Multiple R-squared:  0.808,  Adjusted R-squared:  0.8045
## F-statistic:   229 on 5 and 272 DF,  p-value: < 2.2e-16
```

```
# make predictions and calculate residuals
biodiesel %>%
  mutate(predictions = predict(bio_lin)) %>%
  mutate(residuals = Gasoline - predictions) %>%
  select(predictions,residuals,`Blender cost of ethanol with credit2`, Corn, `Corn cost pe
r gallon of ethanol3`, `Ethanol1`, Gasoline, Year)
```

```
## # A tibble: 278 × 8
##    predictions residuals Blender cost of e…¹  Corn Corn …² Ethan…³ Gasol…⁴  Year
##          <dbl>     <dbl>               <dbl> <dbl>   <dbl>   <dbl>   <dbl> <dbl>
##  1       0.841   -0.0309                0.56  1.91    0.71    1.1     0.81  2000
##  2       0.938   -0.0783                0.6   1.98    0.73    1.14    0.86  2000
##  3       0.933   -0.0231                0.6   2.03    0.75    1.14    0.91  2000
##  4       0.963    0.0168                0.65  2.03    0.75    1.19    0.98  2000
##  5       1.01    -0.0212                0.71  2.11    0.78    1.25    0.99  2000
##  6       0.991    0.0987                0.81  1.91    0.71    1.35    1.09  2000
##  7       0.927   -0.0567                0.79  1.64    0.61    1.33    0.87  2000
##  8       0.959   -0.0493                0.79  1.52    0.56    1.33    0.91  2000
##  9       1.00     0.0201                0.94  1.61    0.6     1.48    1.02  2000
## 10       1.11    -0.133                 0.95  1.74    0.64    1.49    0.98  2000
## # … with 268 more rows, and abbreviated variable names
## #   ¹`Blender cost of ethanol with credit2`,
## #   ²`Corn cost per gallon of ethanol3`, ³Ethanol1, ⁴Gasoline
```

```
# check performance of linear regression model (RMSE)
sqrt(mean(resid(bio_lin)^2))
```

```
## [1] 0.3276777
```

I developed a linear regression model to understand how the different pricing variables of corn and Ethanol impact the cost of gasoline. Then, I used this regression model to make the predicted values. I then calculated residuals to represent the difference between the actual values and the predicted values. Lastly, I calculated the root mean square error (RMSE) to reflect the performance of the linear regression model. The RMSE is 0.3277, which is somewhat low and reflects that the model fits the `biodiesel` dataset fairly well.

Cross-Validation

```
# cross validation of 5 folds
lin_cv <- train(Gasoline ~ .,
      data = biodiesel ,
      method = "lm",
      trControl = trainControl(method = "cv", number = 5))
lin_cv
```

```
## Linear Regression
##
## 278 samples
##   5 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 223, 222, 223, 222, 222
## Resampling results:
##
##   RMSE       Rsquared   MAE
##   0.3360259  0.8119067  0.2675571
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

I perfomed a 5-fold cross-validation for the linear regression model. The average performance of the model across 5 folds is reflected in the RMSE value given: 0.335. Since the RMSE value for the cross-validation model is similar to the RMSE value for the linear regression model, the cross validation model is fairly accurate in predicting new observations.

Dimensionality Reduction

Principal Component Analysis

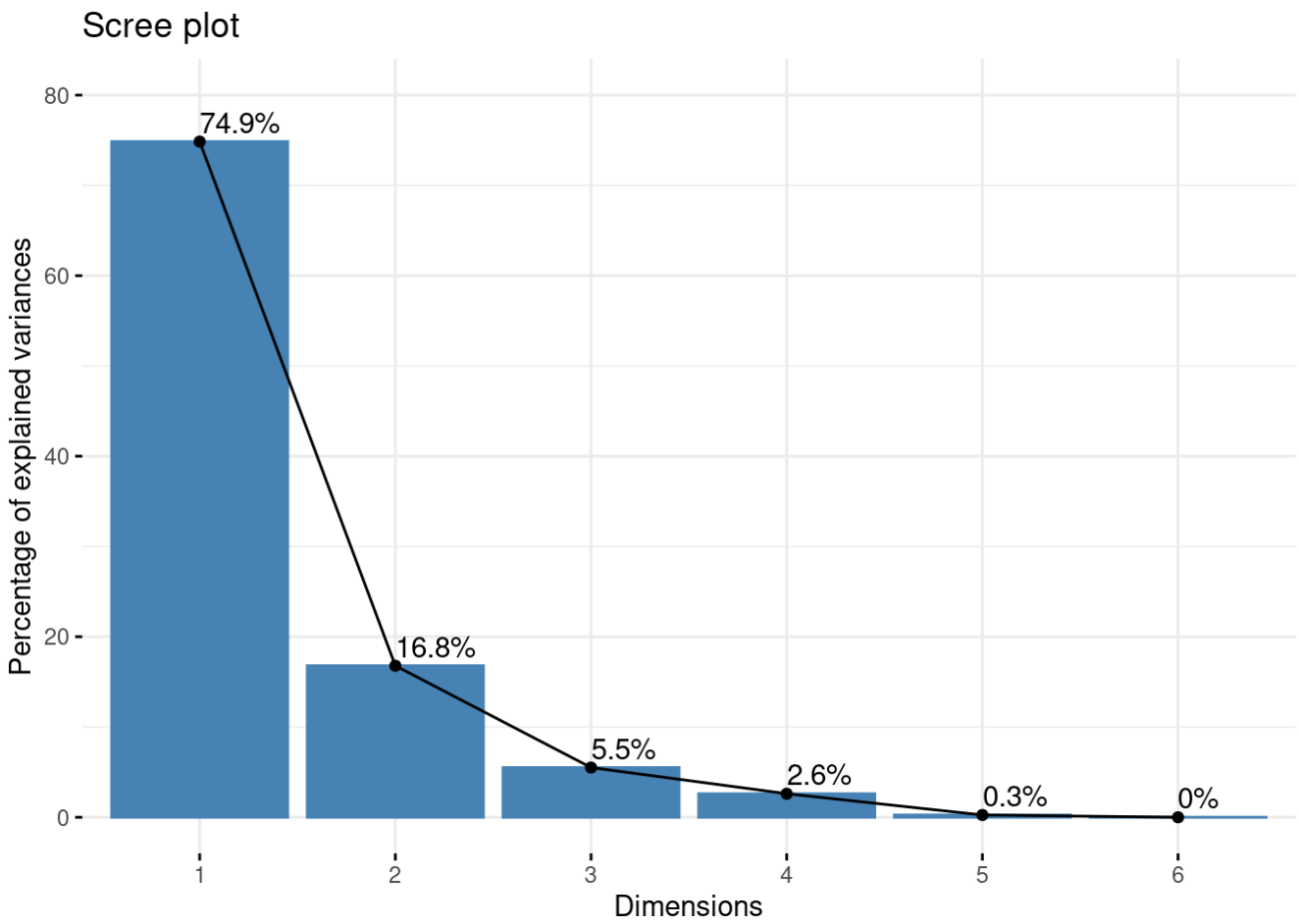
```
# scale the dataset
bio_scaled <- biodiesel %>%
  scale %>%
  as.data.frame

# perform PCA
pca <- bio_scaled %>%
  prcomp

# view objects
names(pca)
```

```
## [1] "sdev"     "rotation" "center"   "scale"    "x"
```

```
# view element x
pca$x %>% as.data.frame
```

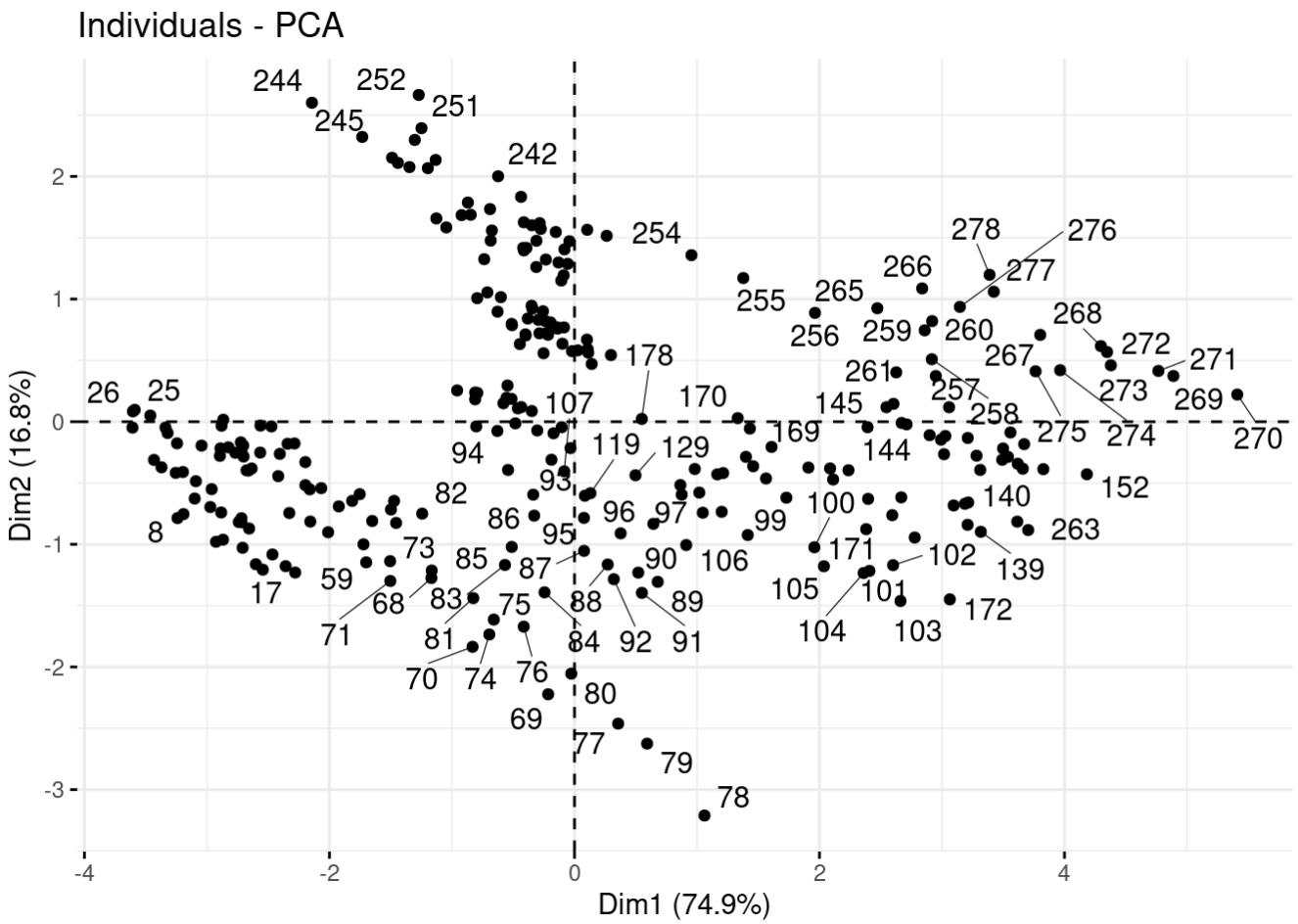
```
##         PC1        PC2       PC3        PC4         PC5          PC6
## 1 -3.370972 -0.3719753 0.7173878 0.01019626 -0.09445391 -0.003737291
##  [ reached 'max' / getOption("max.print") -- omitted 277 rows ]
```

```
# create scree plot
fviz_eig(pca, addlabels = TRUE, ylim = c(0, 80))
```



To conduct principal component analysis (PCA), the dataset was scaled. Once PCA was performed, I viewed the five different objects and decided to view element x. The dataframe for element x indicates the new values for the six principal components made. The scree plot depicts how much variance each PC explains. The first dimension explains about 75% of the variance.

```
# visualize observations for PC 1 and PC 2
fviz_pca_ind(pca, repel = TRUE)
```



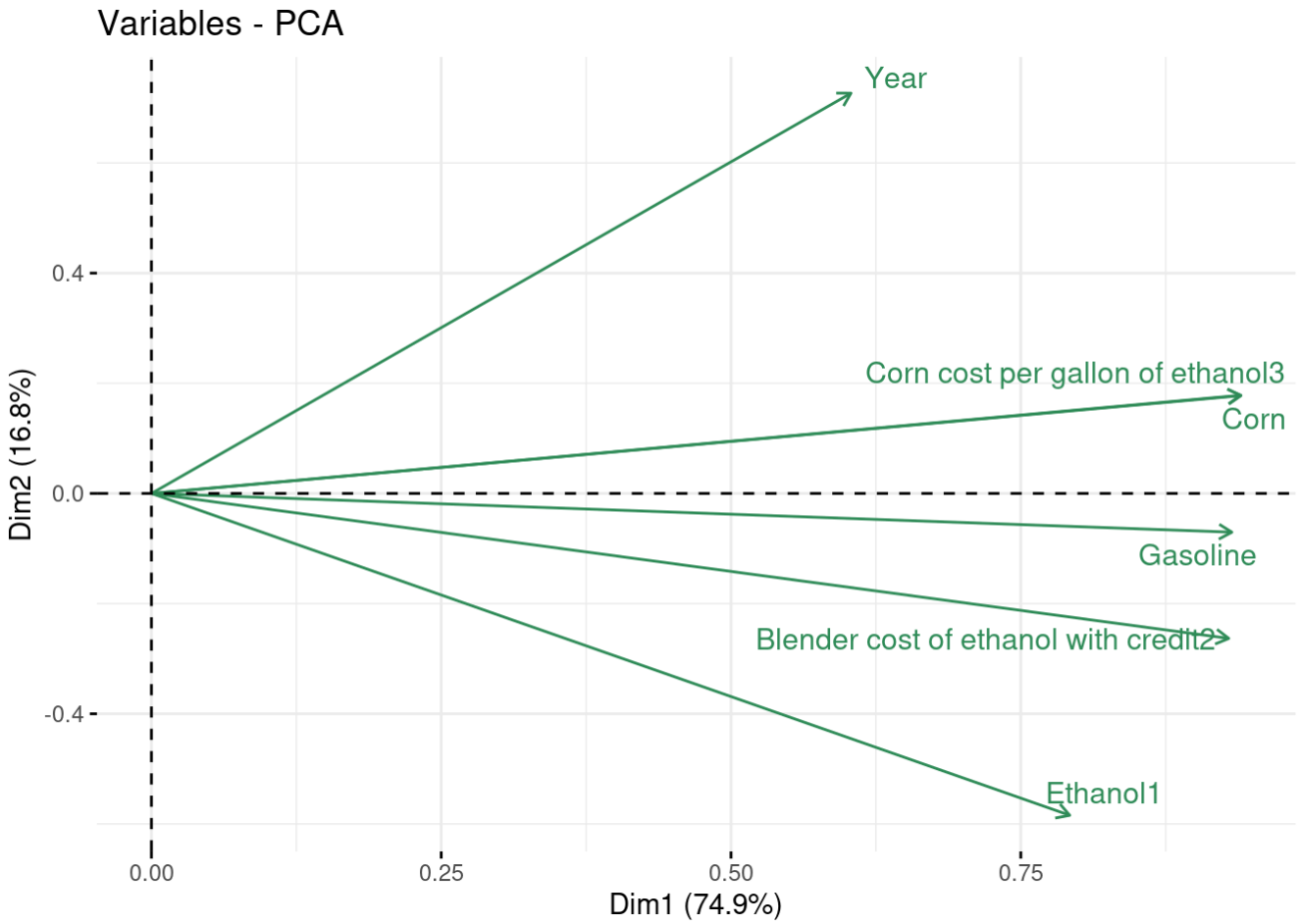

```
# contribution of each variable to each component
get_pca_var(pca)$coord %>% as.data.frame
```

```
##          Dim.1    Dim.2      Dim.3      Dim.4      Dim.5        Dim.6
## Year 0.6035264 0.726706 -0.3229422 0.04910012 0.03085187 2.823004e-06
##  [ reached 'max' / getOption("max.print") -- omitted 5 rows ]
```

```
# select for the variables that contributes the most and least to the first PC
get_pca_var(pca)$coord %>% as.data.frame %>%
  filter(Dim.1 == max(Dim.1) | Dim.1 == min(Dim.1)) %>%
  select(Dim.1)
```

```
##          Dim.1
## Year 0.6035264
## Corn 0.9399761
```

```
# contributions of variables to dimension 1 and 2
fviz_pca_var(pca, col.var = "seagreen", repel = TRUE)
```

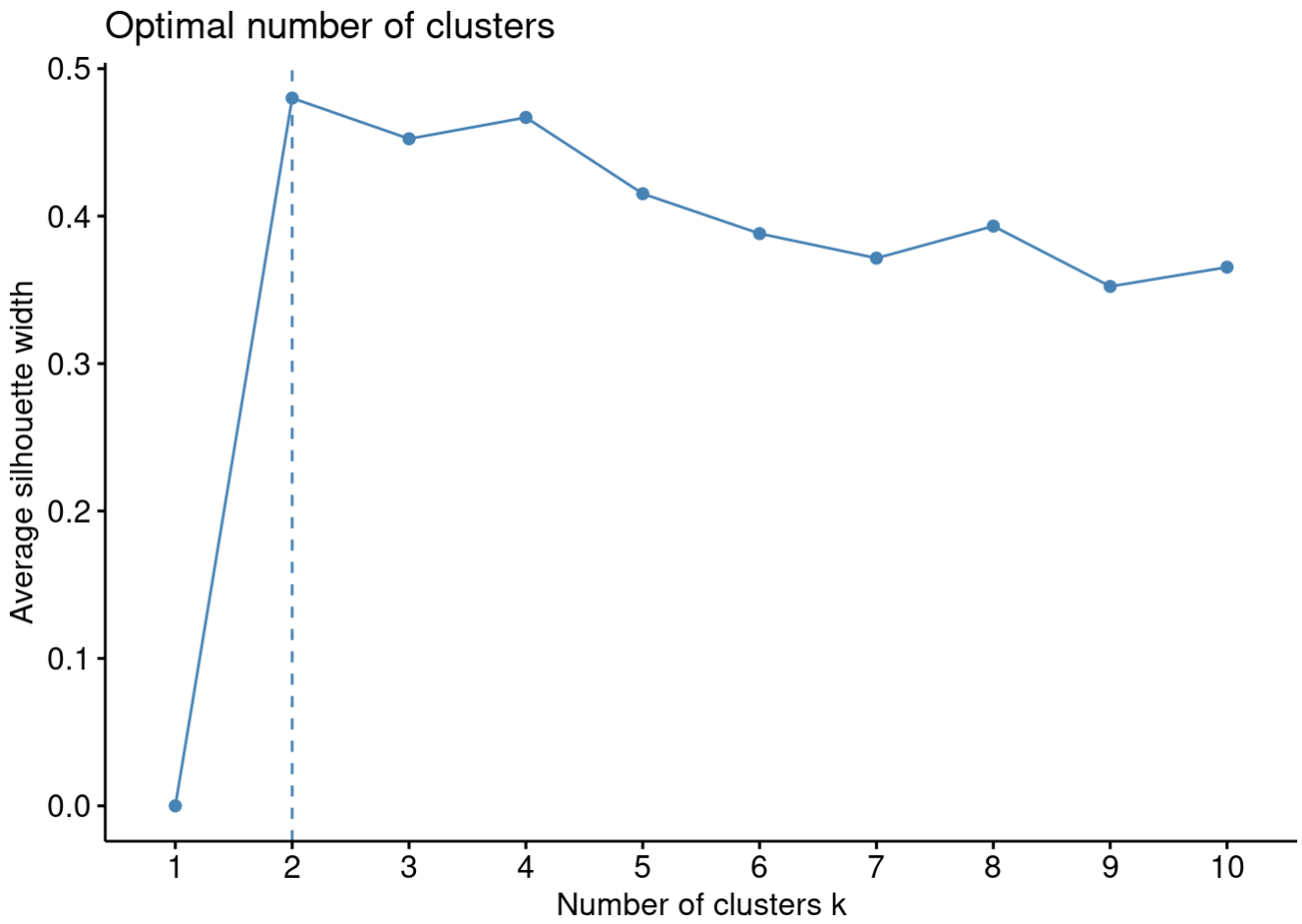


I first visualized the observations for the first and second PC in the plot. The dataframe shows how each variable contributes to each PC. Since PC 1 explains 75% of the variance, I examined which variable contributes the most and least to it. I found that the cost of corn contributes the most, while the year variable contributes the least to PC 1. Then, I visualized how the different variables contribute to PC 1 (75% of the variance) and PC 2 (17% of the variance). I found that for PC 1, all the variables except for year contributes. All of the variables also exhibited correlation with each other. Ethanol, the blender cost of Ethanol, and the cost of gasoline have opposing correlation to the rest of the variables for PC 2.

Clustering

PAM Clustering

```
# scale dataset
bio_pam_scaled <- biodiesel %>%
  scale()

# determine number of clusters
fviz_nbclust(bio_pam_scaled, pam, method = "silhouette")
```

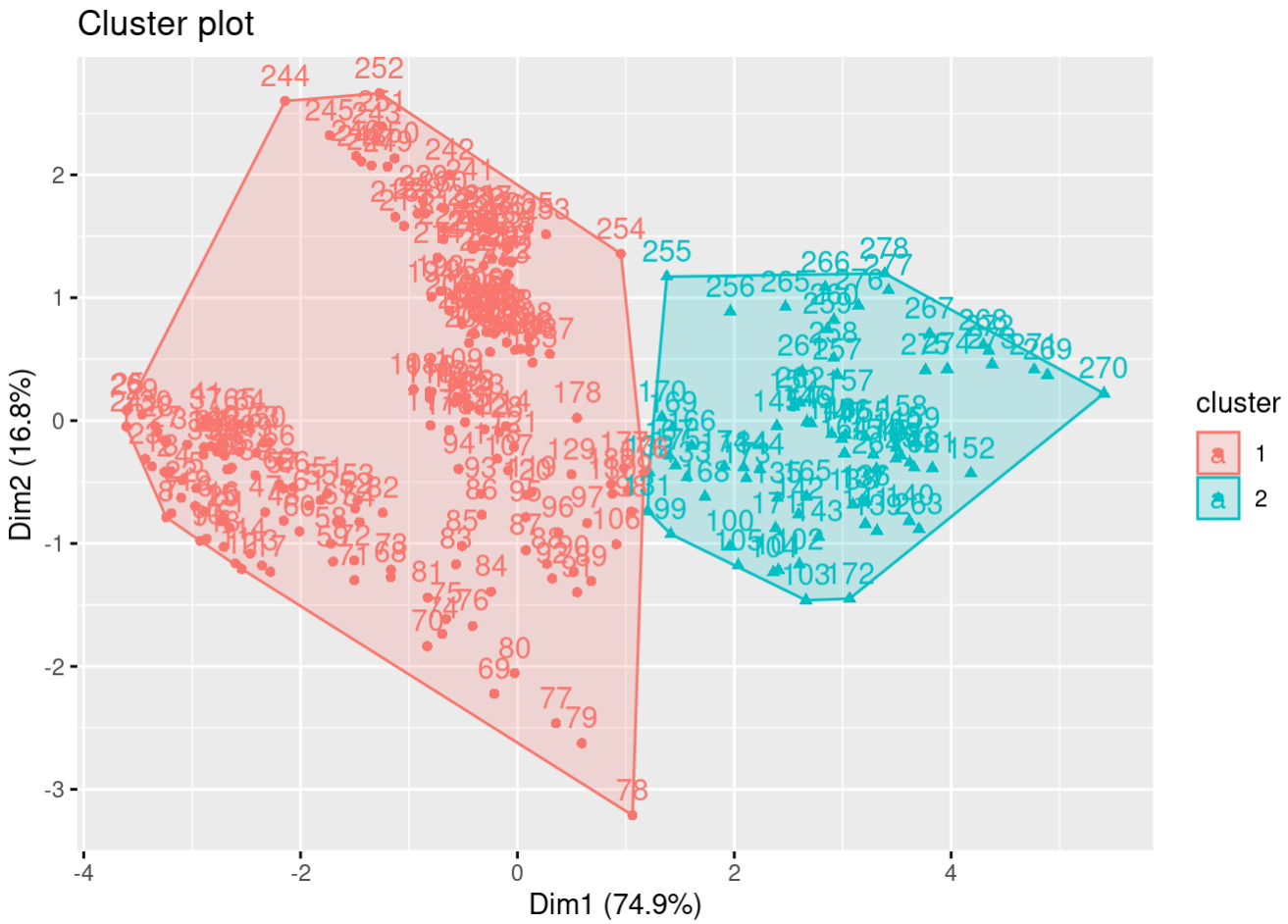


The plot indicates that 2 clusters should be utilized when conducting PAM clustering for the dataset.

```
# find clusters
pam_results <- bio_pam_scaled %>%
  pam(k = 2)

# view resulting object
pam_results
```

```
## Medoids:
##       ID       Year       Corn   Ethanol1 Blender cost of ethanol with credit2
## [1,] 117 -0.3114503 -0.3529929 -0.2783779                            -0.5418001
##        Gasoline Corn cost per gallon of ethanol3
## [1,] -0.1267504                         -0.358711
##  [ reached getOption("max.print") -- omitted 1 row ]
## Clustering vector:
##  [1] 1 1 1 1 1 1 1 1 1 1
##  [ reached getOption("max.print") -- omitted 268 entries ]
## Objective function:
##    build     swap
## 1.676816 1.614893
##
## Available components:
##  [1] "medoids"    "id.med"     "clustering" "objective"  "isolation"
##  [6] "clusinfo"   "silinfo"    "diss"       "call"       "data"
```

```
# plot clusters after dimension reduction
fviz_cluster(pam_results, data = bio_pam_scaled)
```



The cluster plot visualizes the two clusters for PC 1 and PC 2.

```
# stats for clusters
pam_results$medoids
```

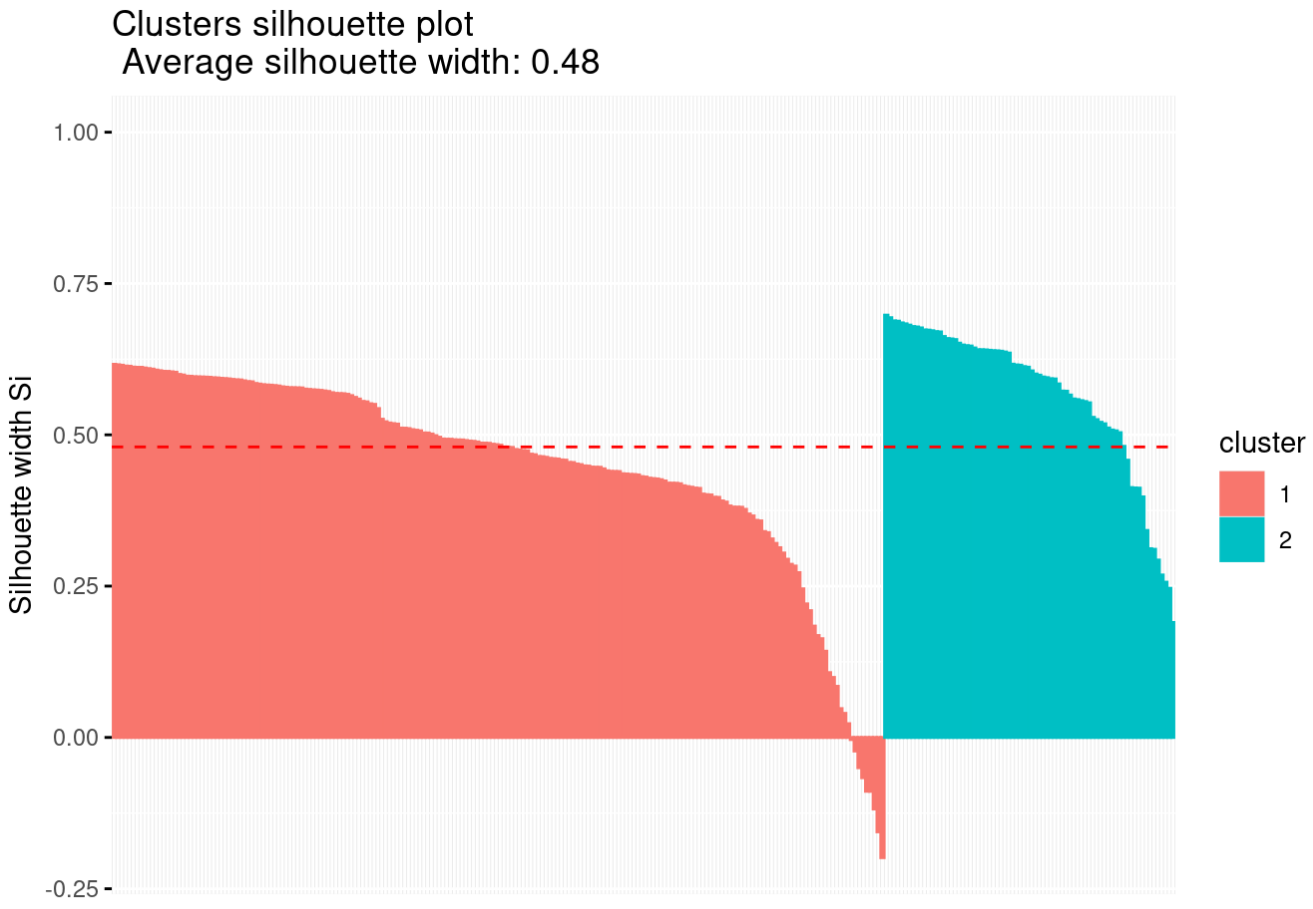
```
##            Year       Corn   Ethanol1 Blender cost of ethanol with credit2
## [1,] -0.3114503 -0.3529929 -0.2783779                            -0.5418001
##        Gasoline Corn cost per gallon of ethanol3
## [1,] -0.1267504                         -0.358711
##  [ reached getOption("max.print") -- omitted 1 row ]
```

```
pam_results$clusinfo
```

```
##      size max_diss  av_diss diameter separation
## [1,]  202 4.374667 1.722137 7.002103  0.3824148
## [2,]   76 2.768041 1.329850 4.463508  0.3824148
```

```
sile <- silhouette(pam_results$cluster, dist(bio_pam_scaled))
fviz_silhouette(sile)
```

```
##   cluster size ave.sil.width
## 1       1  202          0.45
## 2       2   76          0.57
```



The medoids matrix returns rows that are the medoids and columns that are the objects representing each cluster at the center. The maximum dissimilarity between observations in the cluster and the cluster's medoid is 4.37 for PC 1 and 2.77 for PC 2. The average dissimilarity between observations in the cluster and the cluster's medoid is 1.72 for PC 1 and 1.33 for PC 2. The diameter, or the maximum dissimilarity

between two observations in the cluster, is 7.002 for PC 1 and 4.46 for PC 2. The minimal dissimilarity between observations in each cluster is 0.382. Cluster 1 has 202 observations, while cluster 2 has 76 observations. The average silhouette width is 0.45 for cluster 1 and 0.57 for cluster 2. This indicates that cluster 1 has a weak structure and cluster 2 has a reasonable structure.

Discussion

To determine how the various pricing variables of corn and ethanol affect the cost of gasoline, I conducted exploratory visualizations, correlation matrix, multiple linear regression, predictions, cross validation, principal component analysis, and PAM clustering. The exploratory visualizations consisted of multiple scatterplots depicting the relationship between different variables, which generally exhibited positive correlations. The multiple linear regression indicates none of the variables are significant in linearly contributing to the cost of gasoline. I calculated the predicted values as well as the residuals in order to calculate the average distance between the predicted and actual values in the dataset. This RMSE value for the multiple linear regression model is 0.33, which reflects that the model fits the dataset fairly well. Next, I conducted 5-fold cross-validation for the linear regression model. The RMSE value for this model is 0.34, which is quite similar to the RMSE value of the linear regression model. This indicates that the cross validation model is fairly accurate in predicting new observations. Principal component analysis (PCA) developed six principal components (PCs) for the dataset. The scree plot identified that PC 1 explains 75% of the variance and PC 2 explains 17% of the variance. Among all the variables, the year variable contributed the least and the cost of corn variable contributed the most to PC 1. I also visualized the contributions of the variables to PC 1 and PC 2 with a loading plot. All of the variables except for the year variable contributes to PC 1. None of the variables formed a 90 degree or 180 degree vector. In fact, all the vectors formed angles that are less than 90 degrees, which indicates that all of the variables have a positive relationship with each other. For PC 2, the variables ethanol, blender cost of ethanol, and cost of gasoline have opposing correlation to the other variables such as year, corn cost per gallon of ethanol, and cost of corn per bushel. Lastly, I conducted PAM clustering on the scaled dataset. The plot indicates that 2 clusters should be utilized. The average silhouette width is 0.48, which indicates that the clusters have a weak structure. More specifically, cluster 1 has an average silhouette width of 0.45, which is weak. Cluster 2 has an average silhouette width of 0.57, which indicates a reasonable structure. This project was challenging because there was a lot of different ways to go about analyzing the dataset to answer the question. The hardest part of the project was figuring out how to interpret the various plots for the research question, which greatly contributed to my learning process.

Acknowledgement: Dr Guyot.