Project 2: Drunk Driving Deaths and Seat Belt Safety Part 2, CJ Yun

1. Introduction

Description of the datasets

The first dataset is called “US Seat Belts” that explores the relationship of traffic fatalities depending on the seat belt usage in different states over the years. This dataset interested me because sometimes, I honestly forget to wear my seat belt when driving until my car alerts me. I wondered what the effects of not wearing seat belts were in a greater scale. Each row represents each states from years 1983-1997. They also contain 12 different variables that include 6 numeric and 6 categorical variables. It is obtained from R Studio datasets: https://vincentarelbundock.github.io/Rdatasets/doc/AER/USSeatBelts.html (https://vincentarelbundock.github.io/Rdatasets/doc/AER/USSeatBelts.html).

The second dataset is called “Fatality” that explores how different drunk driving laws across the different states over the years has an effect of traffic fatality rate. It is obtained from R Studio datasets: https://vincentarelbundock.github.io/Rdatasets/doc/Ecdat/Fatality.html (https://vincentarelbundock.github.io/Rdatasets/doc/Ecdat/Fatality.html). This dataset particularly interested me since I wondered whether or not having a strict or loose alcohol laws had an effect on people dying in traffic due to drunk driving, especially since drinking is frequent in universities like UT. Each row also represents each states but from the years 1982 to 1988. It contains 10 variables that include 8 numeric and 2 categorical variables. This dataset was supported by another dataset to identify state ID code, which was obtained from a dataset called “stateID” : https://gist.github.com/dantonnoriega/bf1acd2290e15b91e6710b6fd3be0a53 (https://gist.github.com/dantonnoriega/bf1acd2290e15b91e6710b6fd3be0a53).

Based on these two datasets, I can join them with state and year, since they are the two variables that are common and present in both datasets. Moreover, by joining the two datasets with states and year, we can potentially explore the relationship between seatbelt usage from “US Seat Belts” and jail sentences or death due to drunk driving from “Fatality”. We can also explore possible relationship of drunk driving death in correlation to different income and ages.

I predict that my predictor variables will correctly predict my outcome variable fatalities. Moreover, I predict that seat belt usage and age will have the highest correlation and impact on the outcome. The more seat belt usage and higher the age, I predict that the fatality will decrease.

Always a good idea to take a look at the dataset:

```r
# Take a Look
head(joint_dataset)
```

```
## # A tibble: 6 × 23
##   ...1.x state  year miles fatalities seatbelt speed65 speed70 drinkage alcohol
##    <dbl> <chr> <dbl> <dbl>      <dbl>    <dbl> <chr>   <chr>   <chr>    <chr>
## 1     17 AL     1984 32961       932.    0.130 no      no      no       no
## 2     18 AL     1985 35091       882.    0.170 no      no      yes      no
## 3     19 AL     1986 34003      1081.    0.290 no      no      yes      no
## 4     20 AL     1987 37426      1111.    0.210 yes     no      yes      no
## 5     21 AL     1988 39684      1024.    0.290 yes     no      yes      no
## 6     34 AR     1986 17555       603.    0.198 no      no      yes      no
## # … with 13 more variables: income <dbl>, age <dbl>, enforce <chr>,
## #   ...1.y <dbl>, mrall <dbl>, beertax <dbl>, mlda <dbl>, jaild <chr>,
## #   comserd <chr>, vmiles <dbl>, unrate <dbl>, perinc <dbl>, fatal <fct>
```

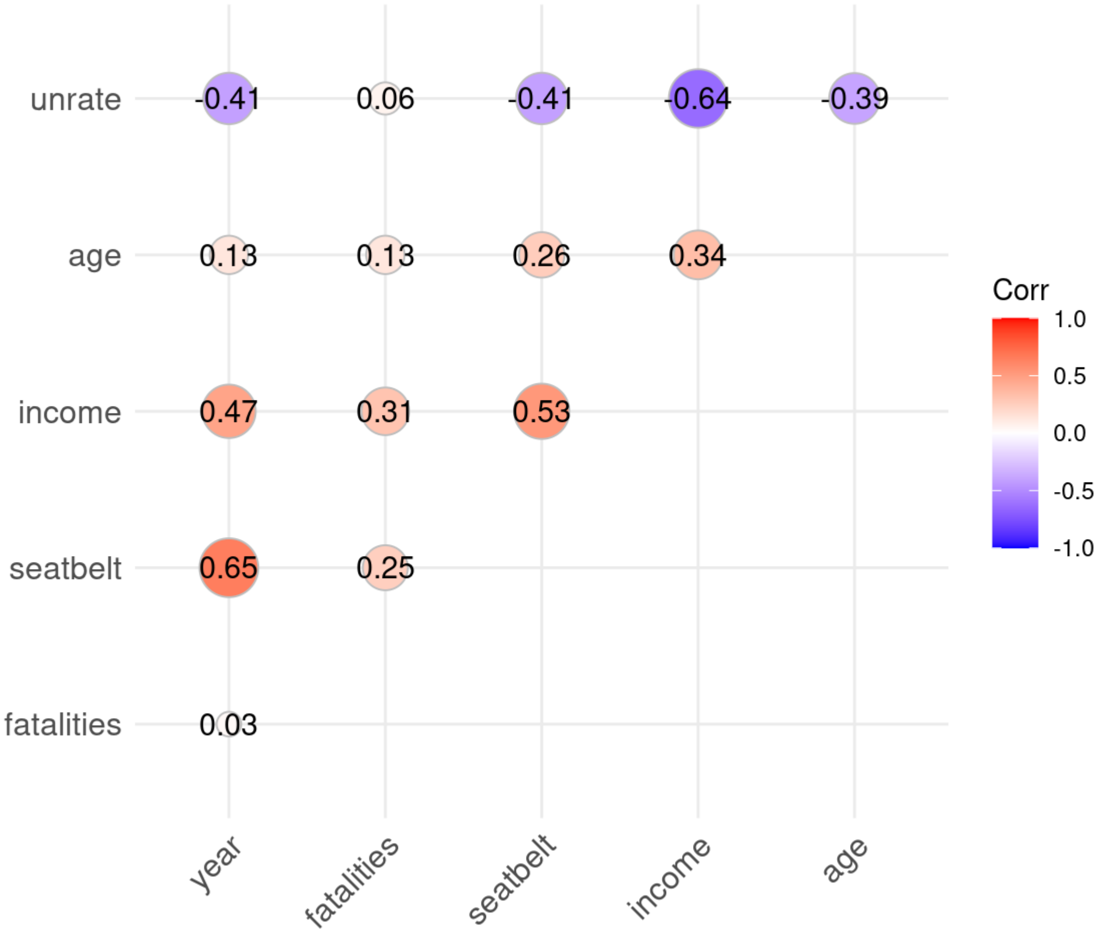
Define a research question

Can year, seat belt usage, age, income, unemployment rate, and mandatory jail sentence predict the status of fatality?

2. Exploratory Data Analysis

Create a correlation matrix:

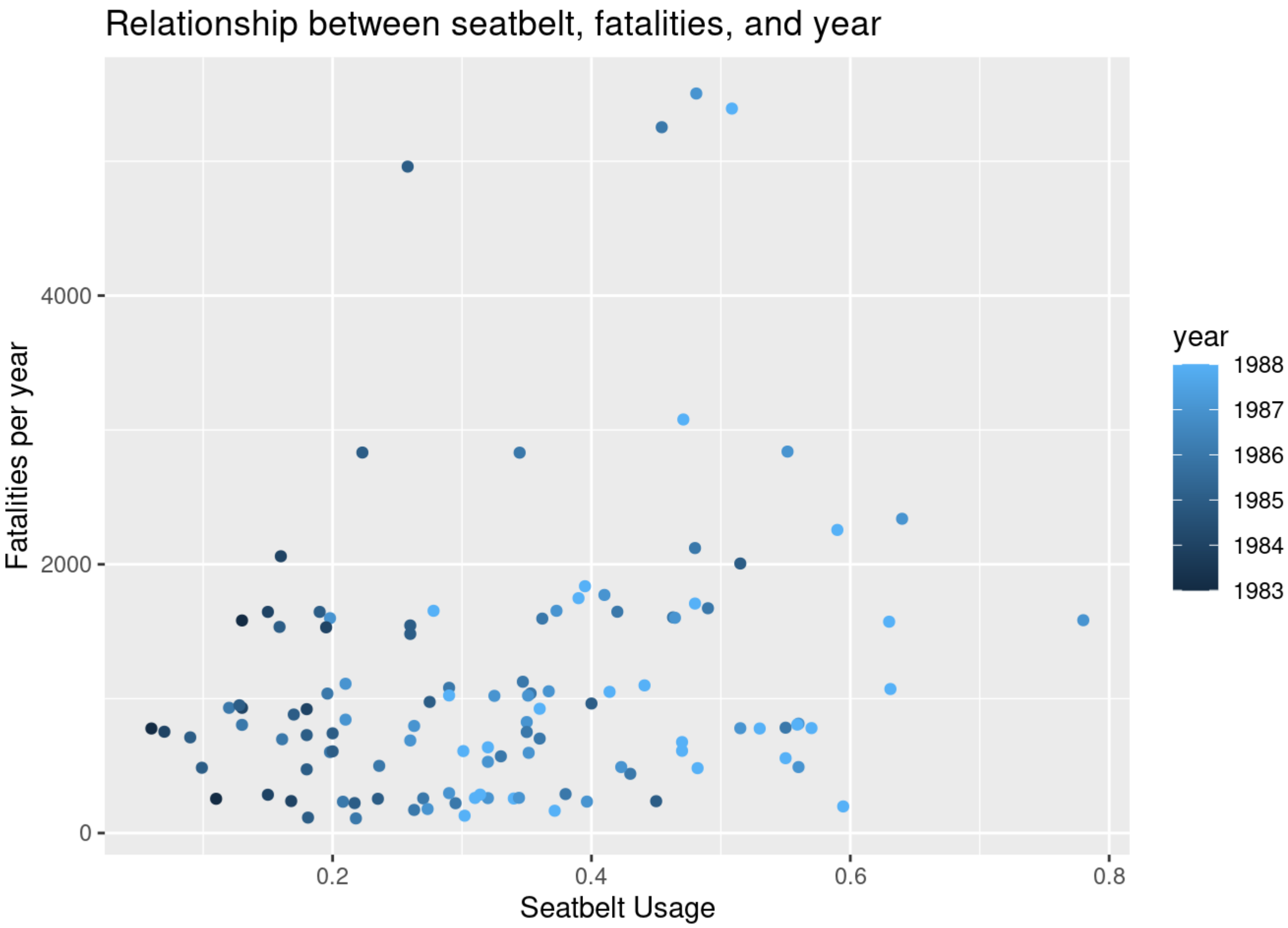
```
# Creating a data with only numeric variables to use them in ggcorrplot
numeric_data <- joint_dataset %>%
  select(year, fatalities, seatbelt, income, age, unrate)

## Use the ggcorrplot to visualize the correlation matrix
ggcorrplot(cor(numeric_data),
           type = "upper", # upper diagonal
           lab = TRUE, # print values
           method = "circle") # use circles with different sizes
```



In the correlation matrix above, it can be seen that seatbelt and year have the highest correlation while year and fatality have the lowest correlation. It makes sense to me that seatbelt and year have the highest correlation since seatbelt usage were enforced more frequently as the years progressed. However, seeing how low the correlation is between fatalities and year surprised me since I thought that as the years went by, they would issue a safer driving leading to a decreased traffic fatalities.

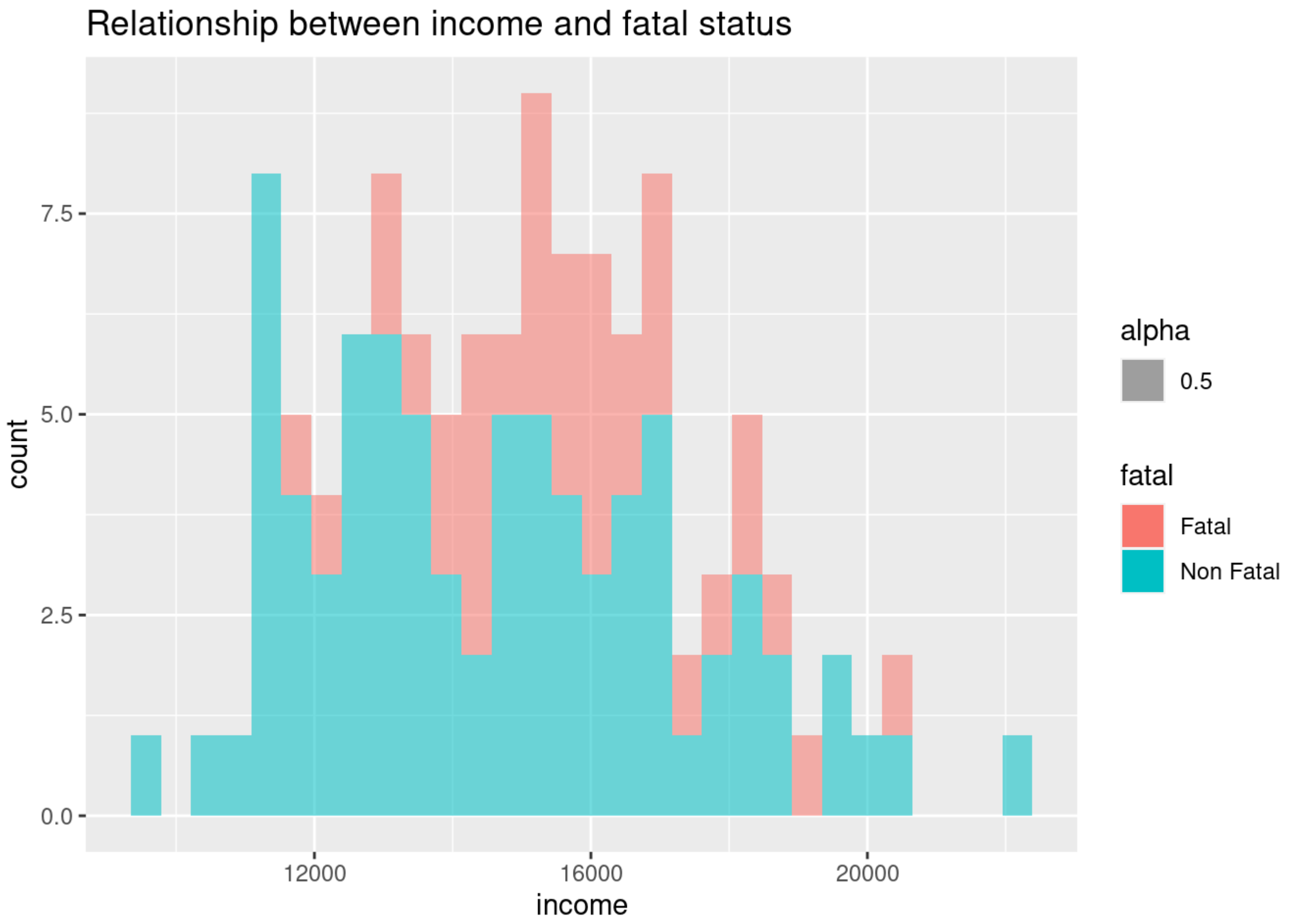
Create visualizations to investigate relationships:

```
# Visualizing between seatbelt, year and fatalities
ggplot(joint_dataset, aes(x=seatbelt, y=fatalities, col = year)) +
  geom_point() +
  labs(title = 'Relationship between seatbelt, fatalities, and year',
       x = 'Seatbelt Usage',
       y = 'Fatalities per year')
```



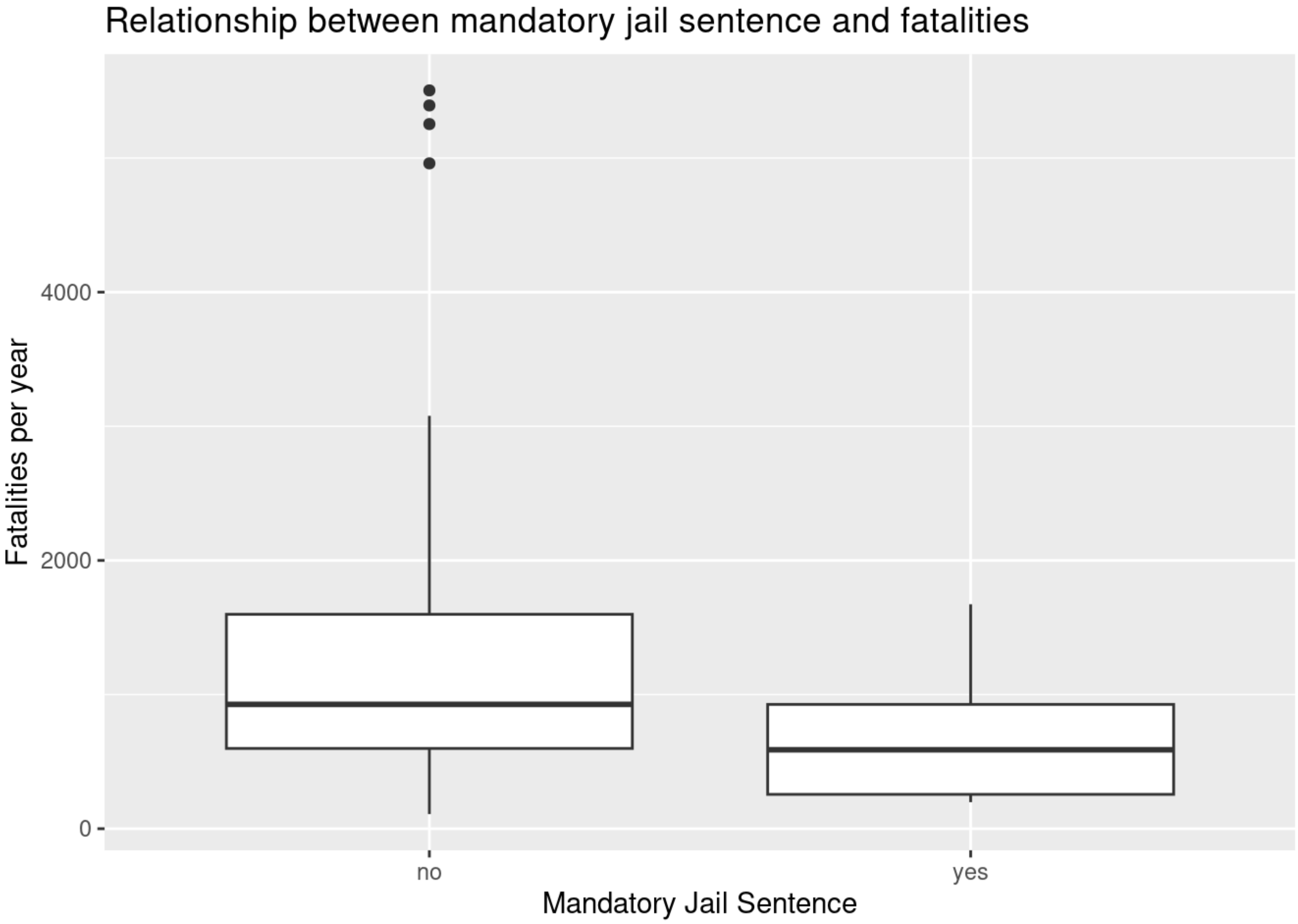
In the visualization above, it can be seen that as the years progressed, the seatbelt usage went up, depicted by lighter colored dots. However, fatalities seemed to remain the same despite seatbelt usage and year based on this visualization.

```
# Visualizing between income and fatalities
ggplot(joint_dataset, aes(x=income, fill=fatal, alpha=0.5)) +
  geom_histogram() +
  labs(title = 'Relationship between income and fatal status')
```



In this visualization above, it can be seen that as income increases, fatality seems to be more fatal, especially around 16000 range. However, on the lower income around 12000, it seems that fatality seems to be more non fatal.

```
# Visualization between mandatory jail and fatalities
ggplot(joint_dataset, aes(x=jaild, y=fatalities)) +
  geom_boxplot() +
  labs(title = 'Relationship between mandatory jail sentence and fatalities',
       x = 'Mandatory Jail Sentence',
       y = 'Fatalities per year')
```



In the third visualization, it can be seen that when there is a mandatory jail sentence when caught drunk driving, fatalities on traffic deaths are less compared to when there is no mandatory jail sentence, which makes sense since people would take more caution in order to avoid jail time.

3. Prediction and Cross-Validation

Train the model

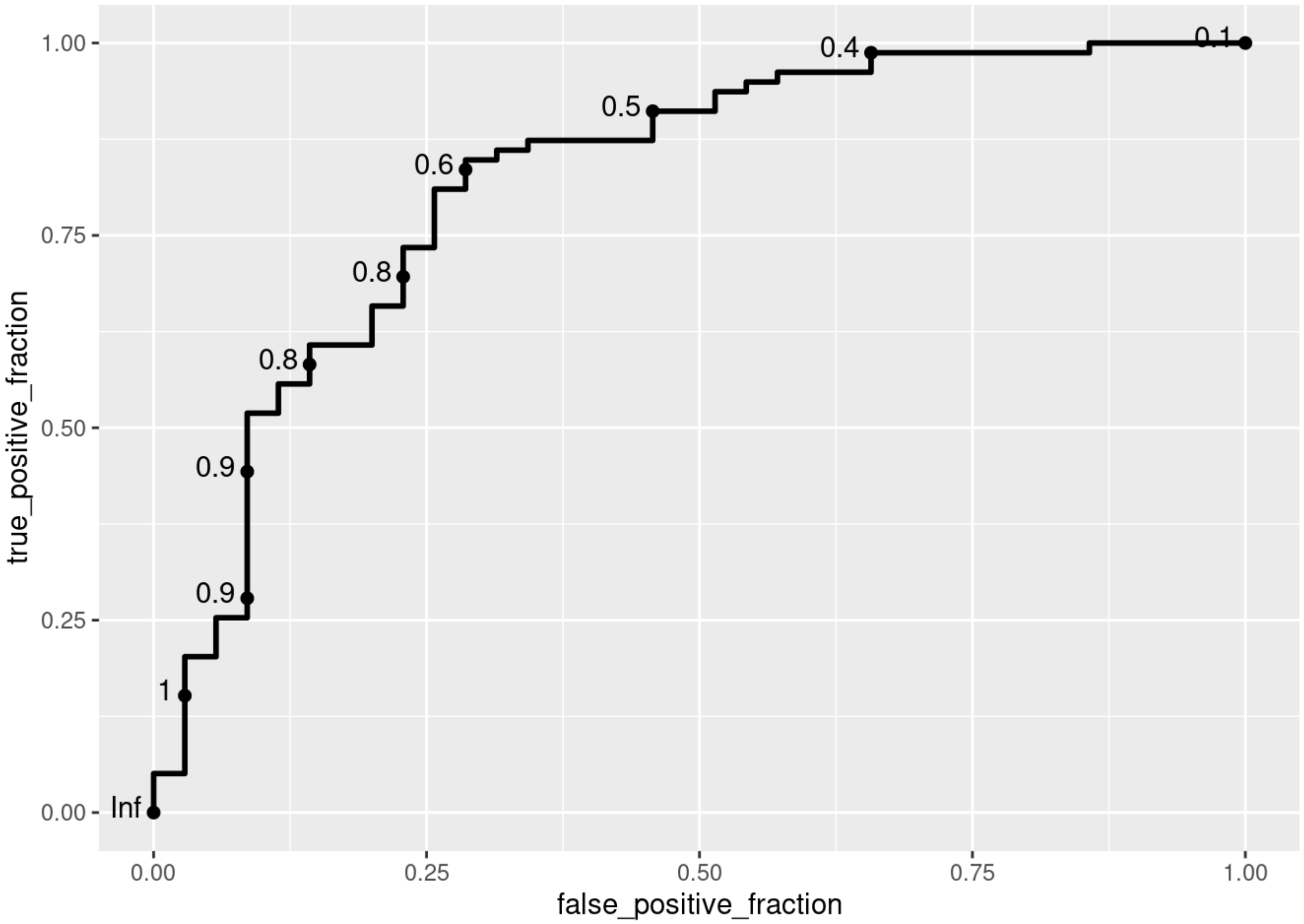
Fitting the model with logistic regression

```
# Logistic Regression
logistic_fatality <- glm(fatal ~ jaild+age+seatbelt+year+income+unrate, data = joint_dataset, fa
mily = 'binomial')

summary(logistic_fatality)
```

```
##
## Call:
## glm(formula = fatal ~ jaild + age + seatbelt + year + income +
##     unrate, family = "binomial", data = joint_dataset)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6515  -0.7404   0.4200   0.6665   1.7356
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.038e+03  5.076e+02  -2.045 0.040861 *
## jaildyes     1.475e+00  7.458e-01   1.977 0.048035 *
##  [ reached getOption("max.print") -- omitted 5 rows ]
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 140.61  on 113  degrees of freedom
## Residual deviance: 104.94  on 107  degrees of freedom
## AIC: 118.94
##
## Number of Fisher Scoring iterations: 5
```

Finding AUC by building ROC:

```
# Build a ROC curve for logistic model:
fatal_ROC <- joint_dataset %>%
  mutate(predictions = predict(logistic_fatality, type = 'response')) %>%
  ggplot() +
  geom_roc(aes(d=fatal, m = predictions), n.cuts = 10)

fatal_ROC
```



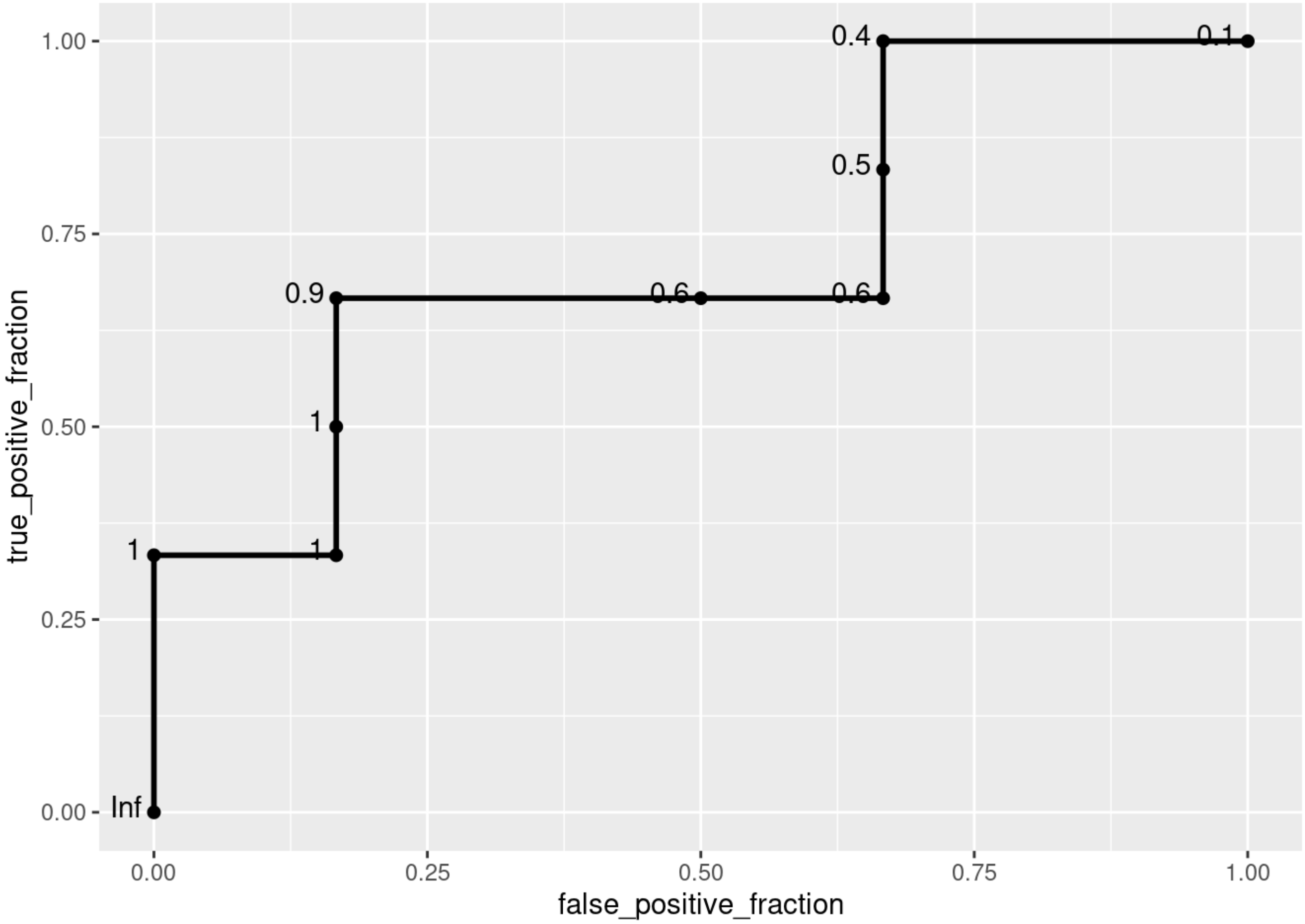
```
# Value of ROC
calc_auc(fatal_ROC)
```

```
##   PANEL group       AUC
## 1     1    -1 0.8202532
```

In this model, the AUC value comes out to 0.8202532, which is not too bad of a prediction for this model.

Perform cross-validation

```
# Choose number of folds
k = 10

# Randomly order rows in the dataset
data <- joint_dataset[sample(nrow(joint_dataset)), ]

# Create k folds from the dataset
folds <- cut(seq(1:nrow(data)), breaks = k, labels = FALSE)

# Initialize a vector to keep track of the performance
perf_k <- NULL

# Use a for loop to get diagnostics for each test set
for(i in 1:k){
  # Create train and test sets
  train_not_i <- data[folds != i, ] # all observations except in fold i
  test_i <- data[folds == i, ]  # observations in fold i

  # Train model on train set (all but fold i)
  fatal_log <- glm(fatal ~ jaild+age+seatbelt+year+income+unrate, data = train_not_i %>%
                    mutate(fatal = ifelse(fatal == "Fatal", 0, 1)),
                  family = "binomial")

  # Test model on test set (fold i)
  predict_i <- data.frame(
    predictions = predict(fatal_log, newdata = test_i, type = "response"),
    fatal = test_i$fatal)

  # Consider the ROC curve for the test dataset
  ROC <- ggplot(predict_i) +
    geom_roc(aes(d = fatal, m = predictions))

  # Get diagnostics for fold i (AUC)
  perf_k[i] <- calc_auc(ROC)$AUC
}

ROC
```



```
# Average performance
mean(perf_k)
```

```
## [1] 0.7651212
```

When using the 10-fold cross-validation with the same type of logistic regression model, it can be seen that the AUC value of this model comes out to 0.7686442, which is quite a bit lower than the previous observation. This displays potential signs of overfitting which made the average performance of the model lower.

Dimensionality reduction

```
# Perform PCA with prcomp()

# Prepare the dataset
scaled_dataset <- joint_dataset %>%
  select(year, fatalities, seatbelt, income, age, unrate) %>%
  scale %>%
  as.data.frame()

# Take a look at the scaled data
head(scaled_dataset)
```

```
##       year fatalities  seatbelt    income        age   unrate
## 1 -1.71753 -0.1785812 -1.365009 -1.780312 -0.4807597 2.099003
##  [ reached 'max' / getOption("max.print") -- omitted 5 rows ]
```

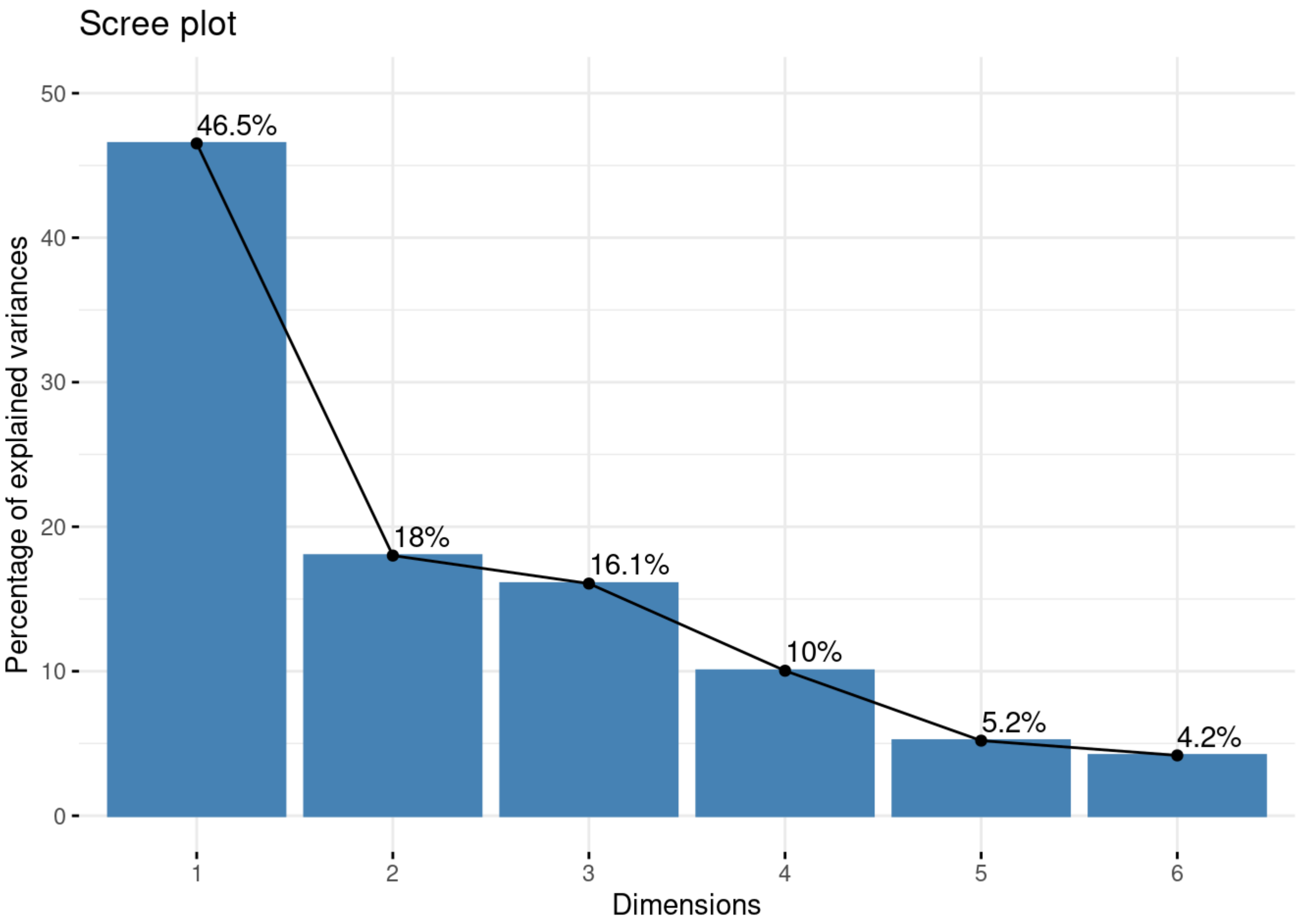
```
# PCA performed with the function prcomp()
p2_pca <- scaled_dataset %>%
  prcomp

# The output creates 5 different objects
names(p2_pca)
```

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

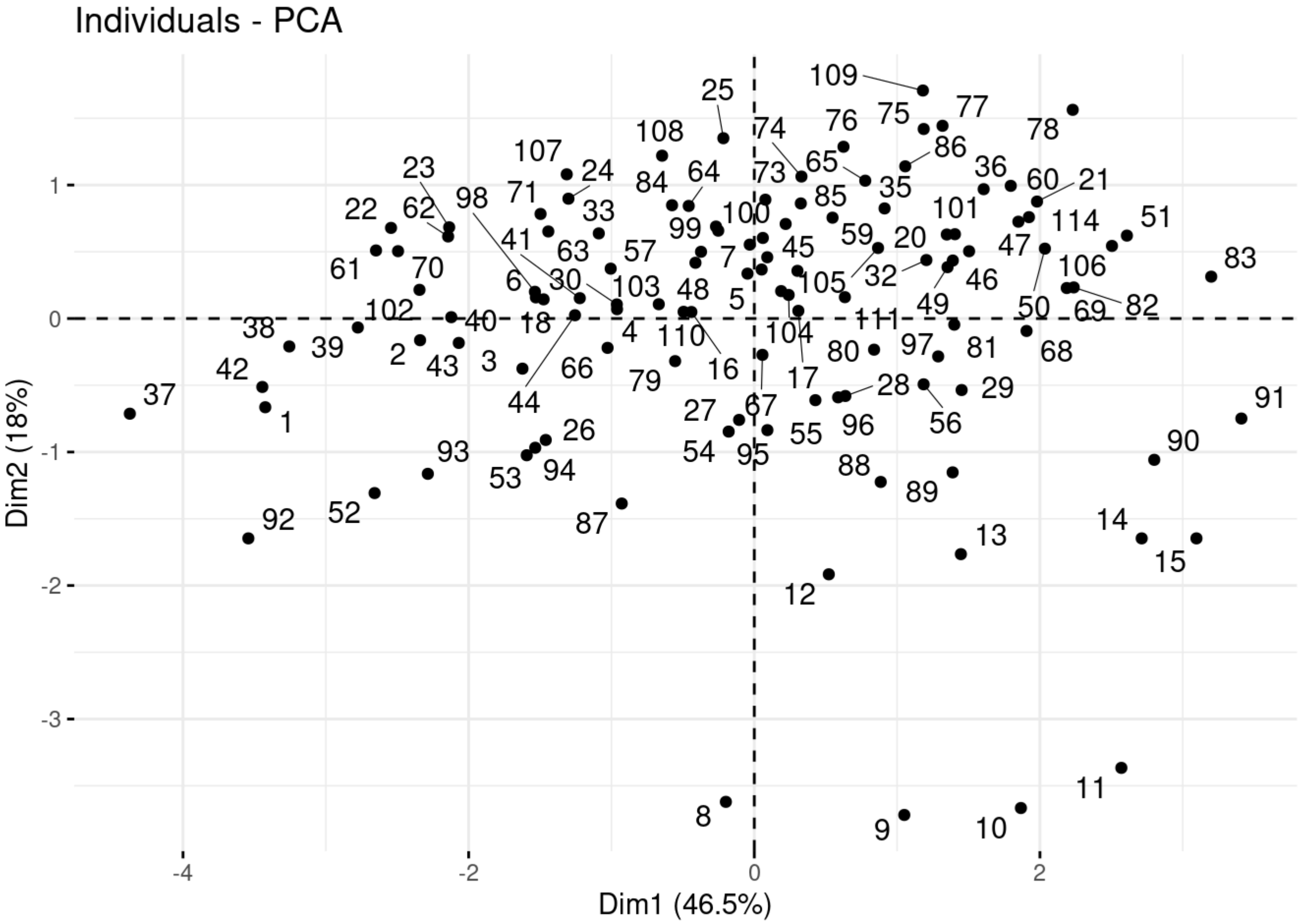
```
# New perspective on our data
p2_pca$x %>% as.data.frame
```

```
##         PC1        PC2       PC3       PC4       PC5        PC6
## 1 -3.424021 -0.6645107 0.2368261 0.5897149 0.1089973 -0.2650955
##  [ reached 'max' / getOption("max.print") -- omitted 113 rows ]
```

```
# Visualize percentage of variance explained for each PC in a scree plot
fviz_eig(p2_pca, addlabels = TRUE, ylim = c(0, 50))
```



```
  # Keep first 3

# Visualize the individuals according to PC1 and PC2
fviz_pca_ind(p2_pca,
             repel = TRUE) # Avoid text overlapping for the row number
```



```
# Visualize the contributions of the variables to the PCs in a table
get_pca_var(p2_pca)$coord %>% as.data.frame
```

```
##         Dim.1     Dim.2      Dim.3     Dim.4      Dim.5     Dim.6
## year 0.723328 0.2516738 -0.4831554 0.2198379 -0.3464769 0.1079019
##  [ reached 'max' / getOption("max.print") -- omitted 5 rows ]
```

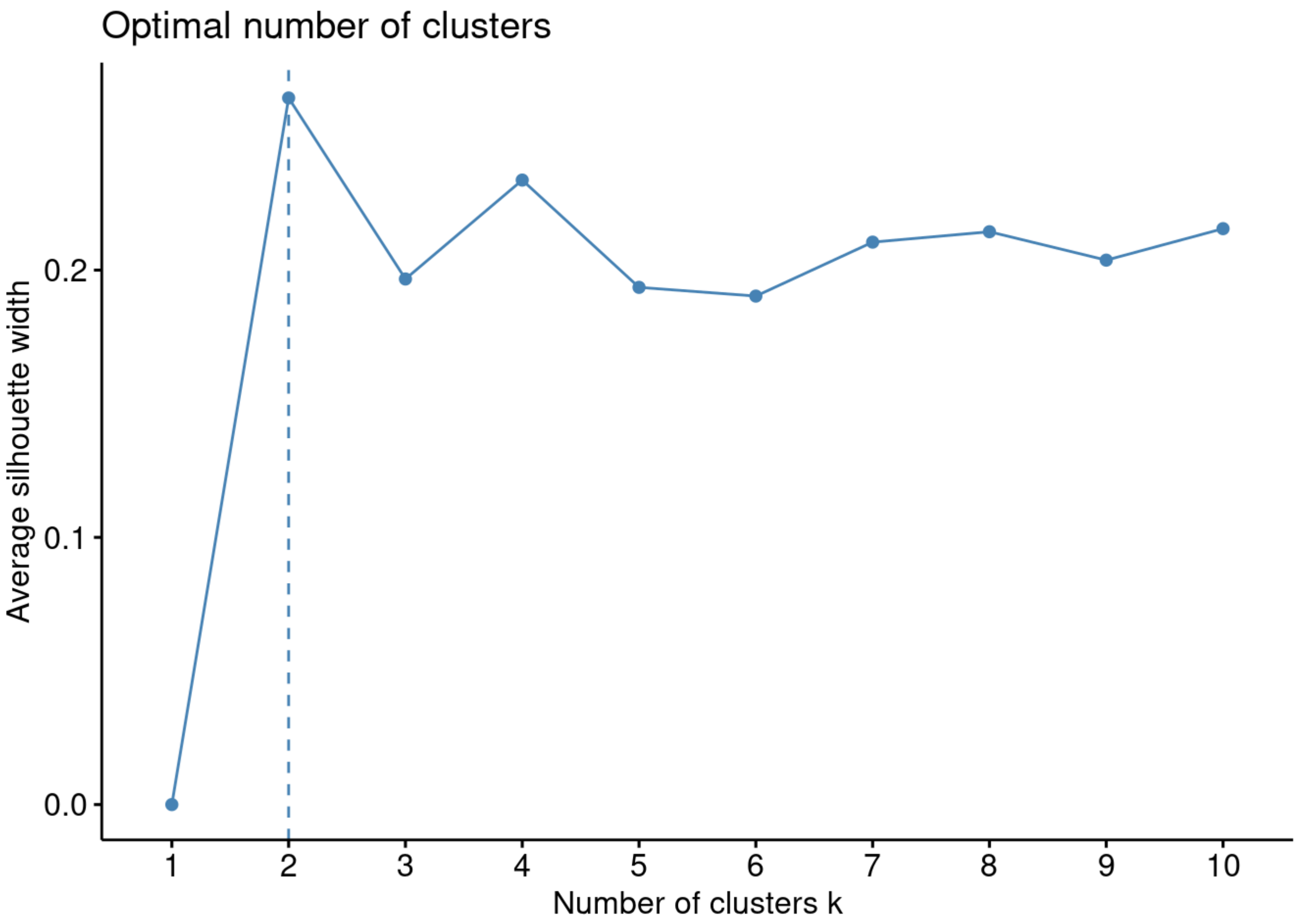
Based on the percentages of variation explained in the scree plot, 3 PCs should be retained since they combine to around 80%.

As can be seen in the scree plot, it can be said that this data was able to score high on the first 2 PCs, which means that the data points are located in a region of the dataset that is strongly influenced by these two PCs. In other words, it indicates that the varaibles that contribute most to these PCs have a strong effect on the data points.

Moreover, it can be seen that income and seatbelts were high contributers to the PCs.

Clustering

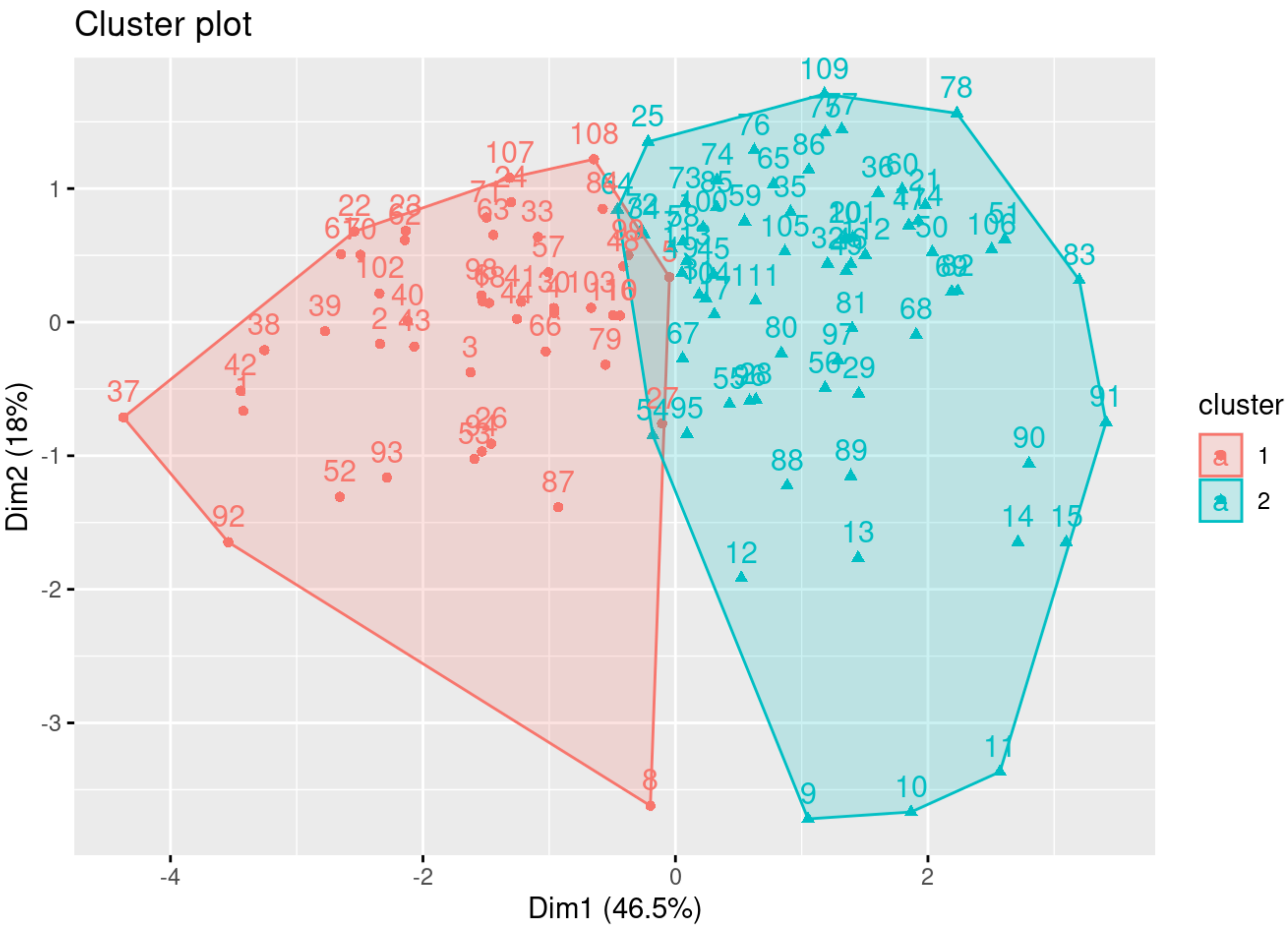
```
# Determine the number of clusters based on the silhouette width
fviz_nbclust(scaled_dataset, pam, method = "silhouette")
```



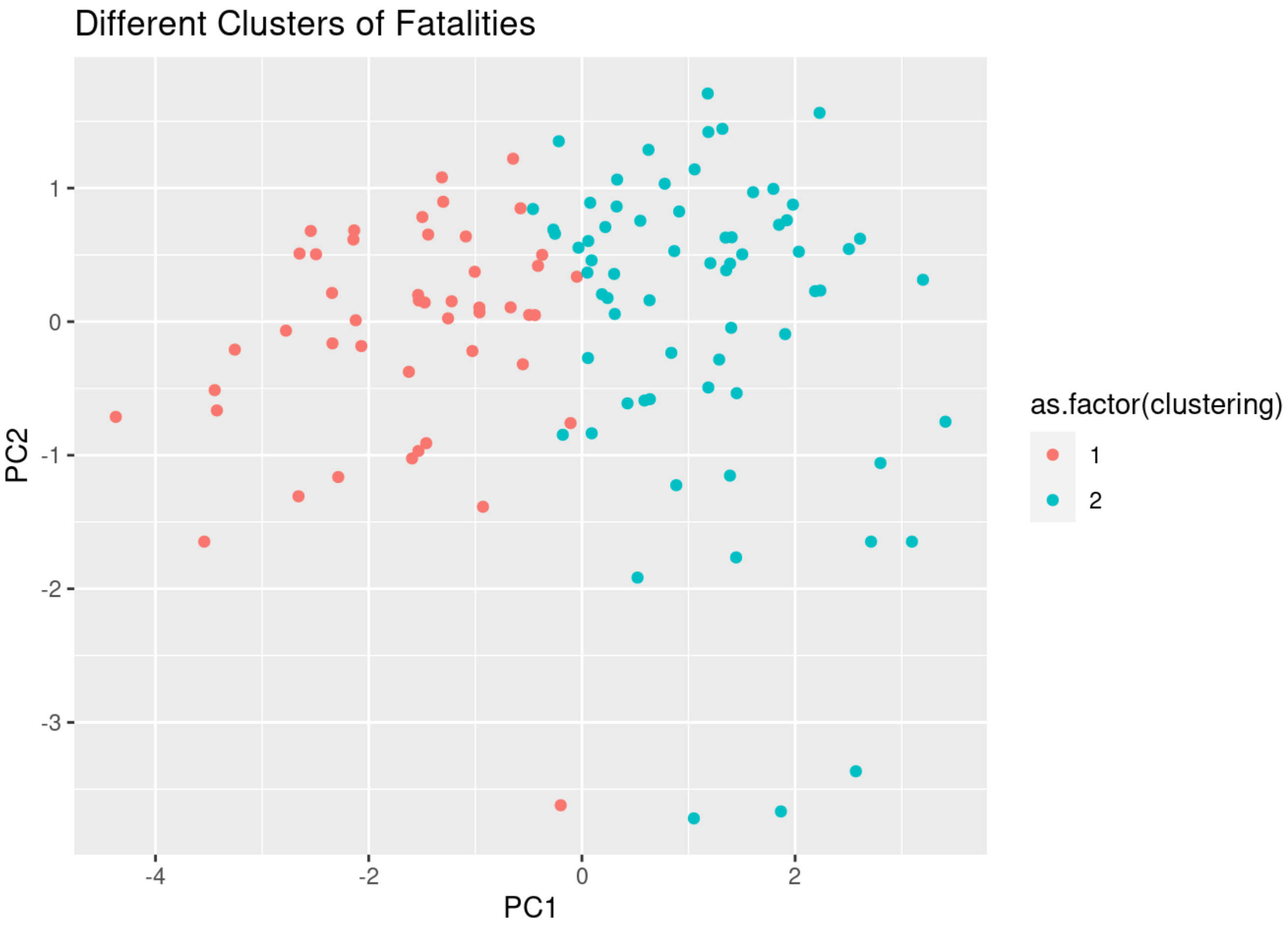
```
# Use the function pam() to find clusters
p2_pam_results <- scaled_dataset %>%
  pam(k = 2) # k is the number of clusters

# Take a look at the resulting object
p2_pam_results
```

```
## Medoids:
##      ID       year fatalities   seatbelt     income        age     unrate
## [1,] 98 -0.2398060 -0.4051284 -1.1570968 -0.7990901 -0.1707015  0.7817468
##  [ reached getOption("max.print") -- omitted 1 row ]
## Clustering vector:
##  [1] 1 1 1 1 1 1 2 1 2 2
##  [ reached getOption("max.print") -- omitted 104 entries ]
## Objective function:
##    build     swap
## 1.979968 1.968582
##
## Available components:
##  [1] "medoids"    "id.med"     "clustering" "objective"  "isolation"
##  [6] "clusinfo"   "silinfo"    "diss"       "call"       "data"
```

```
# Let's visualize the clusters after dimension reduction
fviz_cluster(p2_pam_results, data = scaled_dataset)
```



```
# Visualizing the clusters in a 2-dimensional plot
p2_pca$x %>%
  as.data.frame %>% # change to data frame
  select(PC1, PC2) %>% # select the first two principal components
  # add species and clustering
  mutate(clustering = p2_pam_results$clustering) %>%
  # add ggplot
  ggplot(aes(x=PC1, y=PC2, color = as.factor(clustering))) +
  geom_point() +
  labs(title = 'Different Clusters of Fatalities',
       xlab = 'PC1',
       ylab = 'PC2')
```



Based on the silhouette width, there should be 2 clusters since that is the optimal peak.

Therefore, in the visualizations, it can be seen that there are two main clusters that forms from the variables fatalities, age, year, seatbelt, income, and unrate. This also represents how there are two groups with similar datapoints with the variables stated above.

One of the cluster displays the combination of variables that lead to a non fatal status from the data while the other cluster displays the combination of variables that lead to a fatal status from the data.

6. Discussion

From these following data and visualization, it can be concluded that the following variables, ‘age, year, seatbelt, income, jaild, and unrate,’ predicts the outcome of fatality by a signficant amount. From the logistic regression model, it can be seen that according to the higher end of AUC value, it can be concluded that the predictor variables were able to predict the outcome majority of the times. Although 10-fold cross-validation had a lower prediction due to its overfitting and the clusters weren’t defined too clearly, based on some correlations that did exist in the data allowed the predictor variables to the correct outcome.

However, some challenges did exist as the variables in the data did not have significant correlation. The highest correlation that existed were .65, which isn’t too high. It was difficult and challenging to try to mesh different data and try to find a significant correlation. But, through this process, I was able to learn how to deal with different and uncorrelated variables and still find a cluster by having many variables that can offer contribution.

7. Formatting

Remember to knit your file and produce a pdf with multiple pages to upload on Gradescope. It is ok to not follow the structure of the project in order but make sure to identify the pages accordingly.