419 Final Project

Justin Koida, Brandon Kim, Nick Soares, Shrav Sheth

Table of Contents

Set Up

Getting stuff running

Libraries:

```
library(tidyverse)
library(tidymodels)
library(MASS)
```

The Data:

```
pollution <- read_csv(here::here("pollution_419.csv"))

pollution <- pollution %>%
  mutate(MORTRANK = as.factor(MORTRANK))

knitr::kable(head(pollution), caption = "Glimpse of Dataset")
```

Table 1: Glimpse of Dataset

MORTRANK	PRECIP	EDUC	NONWHITE	NOX	SO2
1	13	12.2	3.0	32	3
1	28	12.1	7.5	2	1
1	10	12.1	5.9	66	20
1	43	9.5	2.9	7	32
1	25	12.1	2.0	11	26
1	35	11.8	14.8	1	1

Section A

Overview of the Dataset

The dataset used in this project contains environmental and demographic data collected from 60 metropolitan areas in the United States. The primary focus of the dataset is to analyze factors associated with mortality rates in these areas. The total age-adjusted mortality rate from all causes (MORT) represents the number of deaths per 100,000 people. Although MORT itself is not included as a variable in the dataset, a grouping variable, **MORTRANK**, is derived from it. This grouping allows for classification into three categories:

- **MORTRANK = 1**: Areas where mortality rates are below 912 deaths per 100,000 people.

- **MORTRANK = 2**: Areas where mortality rates range between 912 and 968 deaths per 100,000 people.
- **MORTRANK = 3:** Areas where mortality rates are 968 or higher per 100,000 people.

The purpose of this project is to analyze whether environmental and socioeconomic factors contribute to higher or lower mortality rates across different metropolitan regions.

Description of Variables

The dataset includes the following variables:

- **MORTRANK** *(Grouping Variable)*: Represents the classification of metropolitan areas based on their mortality rates.
- **PRECIP** *(Mean Annual Precipitation in inches)*: This variable measures the average annual precipitation in each metropolitan area. Higher or lower precipitation levels might influence environmental factors related to health.
- **EDUC** *(Median Number of School Years Completed for Individuals Aged 25 and Older)*: This represents the median education level in each area, which could be a key indicator of socioeconomic status and access to health-related information.
- **NONWHITE** *(Percentage of Population That is Non-White)*: This variable represents the racial composition of each metropolitan area. Demographic factors may be correlated with health outcomes due to social determinants of health.
- **NOX** *(Relative Pollution Potential of All Nitrogen Oxides)*: This metric represents the estimated pollution impact of nitrogen oxides in each metropolitan area. Higher NOX levels might be associated with respiratory and cardiovascular diseases.
- **SO2** *(Relative Pollution Potential of Sulfur Dioxide)*: This measures the estimated pollution potential of sulfur dioxide. Like NOX, higher SO2 levels could indicate poorer air quality, potentially leading to adverse health outcomes.

Purpose of the Analysis

The goal of this project is to investigate the relationships between these environmental and socioeconomic variables and mortality rates. Specifically, we aim to determine which factors might contribute to higher or lower mortality rankings (MORTRANK). By analyzing these variables, we can assess whether pollution levels, education, racial composition, and climate factors are significant determinants of health outcomes in urban settings.

Section B

To make our analysis easier for predictor-wise visualizations, we decided to develop an interactive RShiny applet to display all of the appropriate histograms as needed:

Histograms Interpretations

(Please keep in note of the varying x axis per distribution)

Precipitation: Whilst the centers of the precipitations distributions across all groups are relatively similar (considering their standard deviations), it's interesting to note that the group 1 has considerably higher variance than the other groups.

Education Level: All the groups distributions for education level are really similar. This is reflected in the overall distribution's shape looking very similar to each groups.

Nonwhite %: Group 1 has significantly smaller nonwhite % in comparison to the other two groups. It might seem like group 2 and group 3 are also very different, but when considering their standard deviations, it's more marginal than it seems. All 3 distributions seem to all show some right skew as well.

Nitrogen Oxides: Group 1 has massive outliers, skewing the mean and standard deviation to being way larger than they seem. Despite this, all 3 distributions have relatively high variance in consideration to their mean and median.

Sulfur Dioxides: Despite all distributions having roughly similar shape, group 3 has significantly larger spread compared to the other two groups. This high variance seems to really affect its mean and median.

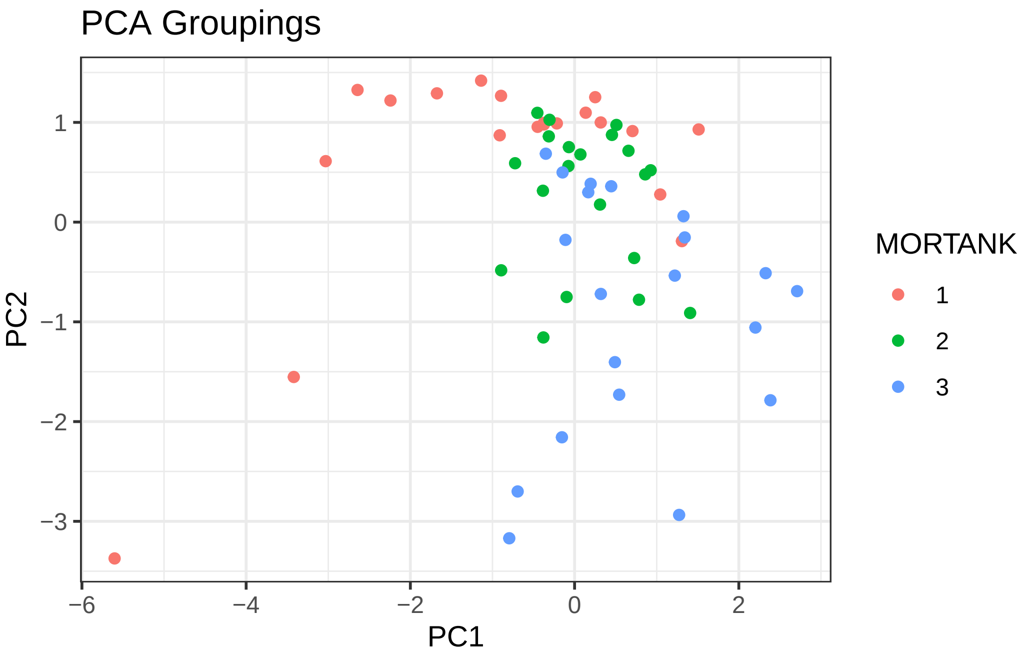
Overall, it seems like group 1 has massively different characteristics than the other two groups. We can see if this is the case visually through Principle Component Analysis, mapping the points on a 2 dimensional plane with PC1 and PC2 as its axis.

PCA Clusters

```r
prepped_recipe_pca <- recipe(~., data=dplyr::select(pollution,-MORTRANK)) %>%
  step_normalize(all_numeric_predictors()) %>%
  step_pca(all_numeric_predictors(), num_comp = 2) %>%
  prep()
```

```
pca_data <- prepped_recipe_pca %>%
  bake(new_data = dplyr::select(pollution, -MORTRANK)) %>%
  mutate(MORTANK = pollution$MORTRANK)

pca_data %>%
  ggplot(aes(x=PC1, y=PC2, color=MORTANK)) +
    geom_point() +
    theme_bw() +
    labs(title = 'PCA Groupings')
```



As hypothesized, group 1 seems to have better separation from the other 2 groups. This might assist LDA into developing a function that can separate group 1 away from the other two. We can do a more formal examination with the k-means algorithm. To ensure reproducibility, we will set the starting centroids as the original scaled means of the groups.

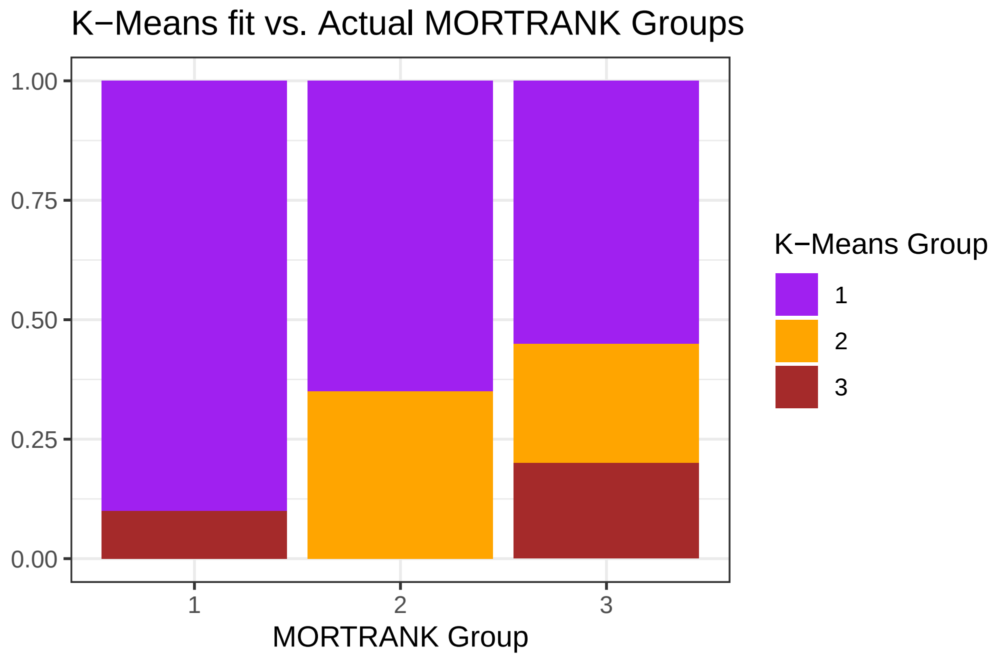
Scaled K-means

```
scaled_predictors <- pollution %>%
  mutate(across(PRECIP:SO2, ~ (. - mean(.)/sd(.))))
```

```
starting_points <- scaled_predictors %>%
  group_by(MORTRANK) %>%
  summarize(across(PRECIP:SO2, list(mean = mean))) %>%
  dplyr::select(-MORTRANK) %>%
  as.matrix()

kmeans_result <- kmeans(dplyr::select(scaled_predictors, -MORTRANK),
                        centers = starting_points,
                        iter.max = 100)

scaled_predictors %>%
  mutate(kmeans_groups = kmeans_result$cluster) %>%
  ggplot(aes(x=as.factor(MORTRANK), fill=as.factor(kmeans_groups))) +
    geom_bar(position = 'fill') +
    scale_fill_manual(values = c('purple', 'orange', 'brown')) +
    labs(x = 'MORTRANK Group', y = '',
         title = 'K-Means fit vs. Actual MORTRANK Groups',
         fill = 'K-Means Group') +
    theme_bw()
```



Unfortunately, the k-means algorithm does not fit the data well. Ideally, we would have 3 k-means groups of equal sizes that align with the MORTRANK groups well. However, k-means does not have a restriction on group sizes, so it seems as if 1 group is dominating. The k-means group that is dominating originally started with the centroid positioned at the 1st MORTRANK group's mean values, which indicates that the other two initial centroids were not able to diverege and capture their points properly. This is probably because those two initial centroids were far closer to each other in terms of proximity to points than the dominant intial centroid, allowing for the dominant group to capture more of the outlying points whilst the two other centroids "fought for control" in their limited shared area. This aligns with the findings from our principle component analysis, as group 1 has a far clearer separation than group 2 and 3.

Section C

C.1

```r
cor_matrix <- round(cor(pollution[,-1]), 6)
knitr::kable(cor_matrix, caption = "Correlation Matrix of Pollution Variables")
```

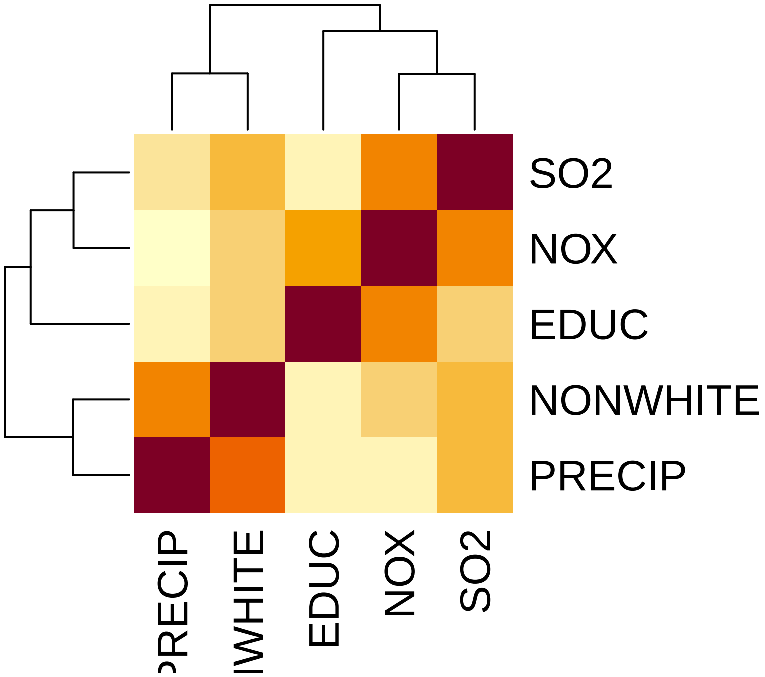
Table 2: Correlation Matrix of Pollution Variables

	PRECIP	EDUC	NONWHITE	NOX	SO2
PRECIP	1.000000	-0.490425	0.413204	-0.487321	-0.106924
EDUC	-0.490425	1.000000	-0.208774	0.224402	-0.234346
NONWHITE	0.413204	-0.208774	1.000000	0.018385	0.159293
NOX	-0.487321	0.224402	0.018385	1.000000	0.409394
SO2	-0.106924	-0.234346	0.159293	0.409394	1.000000

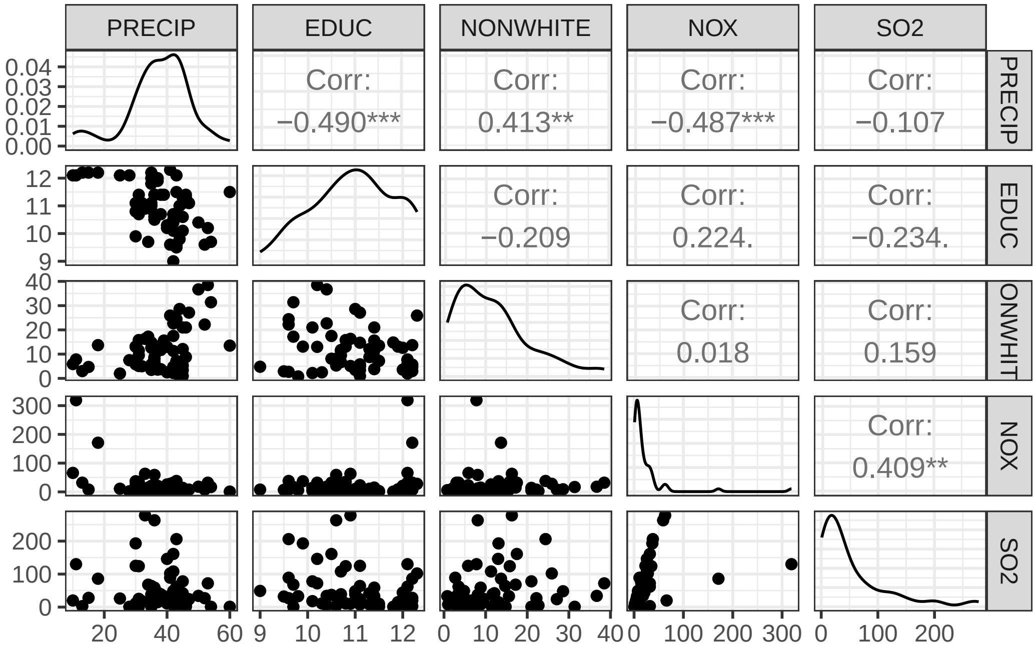
The most correlation we see between variables is EDUC and PRECIP with a correlation of -.4904. This is not correlated significantly, so we will keep all the variables in this data set.

Additionally, below is a heatmap as well as a correlation scatterplot of all of the independent variables:

```r
heatmap(cor_matrix)
```



```
GGally::ggpairs(dplyr::select(pollution,-MORTRANK)) +
  theme_bw()
```



C.2

C.2 Part 1

Write the complete form of all discriminant functions. Be sure to use standardized coefficients. Based on these coefficients, produce a ranking of variable importance (in the presence of other variables).

To start, we have $k = 3$ groups and $p = 6$ variables, so we have min$(p, k - 1) = 2$ discriminant functions.

```
discrim(pollution[,-1], pollution[[1]])
```

```
$a
              [,1]        [,2]
[1,] -0.01840701  0.54309519
[2,]  0.97027032 -0.40472511
[3,] -0.23875696 -0.70885075
[4,]  0.01988270 -0.19533883
[5,] -0.02892210  0.02487075

$a.stand
             [,1]       [,2]
[1,] -0.1722439  5.0820234
[2,]  0.7507216 -0.3131456
[3,] -1.7675389 -5.2476849
[4,]  0.9193970 -9.0326716
[5,] -1.6834803  1.4476618
```

Here is the complete form of both discriminant functions:

$$LD1 = -0.1722439y_1 + 0.7507216y_2 - 1.7675389y_3 + 0.9193970y_4 - 1.6834803y_5$$

$$LD2 = -5.0820234y_1 + 0.3131456y_2 + 5.2476849y_3 + 9.0326716y_4 - 1.4476618y_5$$

For the importance of variables based on the standardized coefficients, we have

LD1: 3, 5, 4, 2, 1 –> NONWHITE, SO2, NOX, EDUC, PRECIP

LD2: 4, 3, 1, 5, 2 –> NOX, NONWHITE, PRECIP, SO2, EDUC

C.2 Part 2

Carry out tests of significance for the discriminant functions. Be sure to specify corresponding null and alternative hypotheses, test statistic values, and p-values. Be sure to provide a conclusion for each test.

```
knitr::kable(discr.sig(pollution[,-1], pollution[[1]]),row.names = T)
```

	Lambda	V	p.values
LD1	0.4375897	45.456049	0.0000018
LD2	0.9616608	2.150142	0.7081668

let $\alpha_1, \alpha_2, ..., \alpha_s$ be the population discriminant functions

For LD1,

$H_0 : \alpha_1 = \alpha_2 = 0$ <—— where 0 is the 0 vector

Ha : at least one of $\alpha_1, \alpha_2 \neq 0$

$\alpha^* = .05/2 = .025$

With a p-value of 1.799089e-06, we reject the Null Hypothesis at $\alpha^* = .025$, so we conclude that we have significant evidence that at least one of $\alpha_1, \alpha_2 \neq 0$ at the $\alpha^* = .025$ level.

For LD2,

$H_0 : \alpha_2 = 0$ <—— where 0 is the 0 vector

$H_a : \alpha_2 \neq 0$

$\alpha^* = .05/2 = .025$

With a p-value of 7.081668e-01, we do not reject the Null Hypothesis at $\alpha^* = .025$, so we do not have significant evidence to suggest that α_2, the second linear discriminant function, has a significant impact on group separability.

Conclude:

We have sufficient evidence to conclude α1 significantly contributes to group separability at $\alpha^* = .025$. We have insufficient evidence to conclude α2 significantly contributes to group separability at $\alpha^* = .025$.

C.2 Part 3

Carry out tests of significance of each non-grouping variable, after adjusting the presence of other non-grouping variables. Be sure to specify corresponding null and alternative hypotheses, test statistic values, and p-values. Be sure to provide a conclusion for each test.

```
knitr::kable(partial.F(pollution[,-1], pollution[[1]]), row.names=T)
```

	Lambda	F.stat	p.value
NONWHITE	0.7351297	9.5480596	0.0002875
SO2	0.8028967	6.5054908	0.0029750
NOX	0.9394342	1.7084679	0.1909673
EDUC	0.9559307	1.2216740	0.3029006
PRECIP	0.9948165	0.1380797	0.8713416

Using $\alpha^* = .05/5 = .01$

We will check variable importance for each non-grouping variable after adjusting for the presence of the other non-grouping variables.

NONWHITE

H_0 : The variable NONWHITE has no significant effect on group separation

H_a : The variable NONWHITE has a significant effect on group separation

With a Lambda stat of 0.7351297 and a p-value of 0.0002875051, we reject the null hypothesis at $\alpha^* = .01$, so there is strong evidence that NONWHITE individually contributes significantly to group separability, adjusting for other variables.

SO2

H_0 : The variable SO2 has no significant effect on group separation

H_a : The variable SO2 has a significant effect on group separation

With a Lambda stat of 0.8028967 and a p-value of 0.0029749625, we reject the null hypothesis at $\alpha^* = .01$, so there is strong evidence that SO2 individually contributes significantly to group separability, adjusting for other variables.

NOX

H_0 : The variable NOX has no significant effect on group separation

H_a : The variable NOX has a significant effect on group separation

With a Lambda stat of 0.9394342 and a p-value of 0.1909672716, we do not reject the null hypothesis at $\alpha^* = .01$, and so there is insufficient evidence to conclude that NOX, individually, contributes significantly to group separation adjusting for other variables.

EDUC

H_0 : The variable EDUC has no significant effect on group separation

H_a : The variable EDUC has a significant effect on group separation

With a Lambda stat of 0.9559307 and a p-value of 0.3029006436, we do not reject the null hypothesis at $\alpha^* = .01$, and so there is insufficient evidence to conclude that EDUC, individually, contributes significantly to group separation adjusting for other variables.

PRECIP

H_0 : The variable PRECIP has no significant effect on group separation

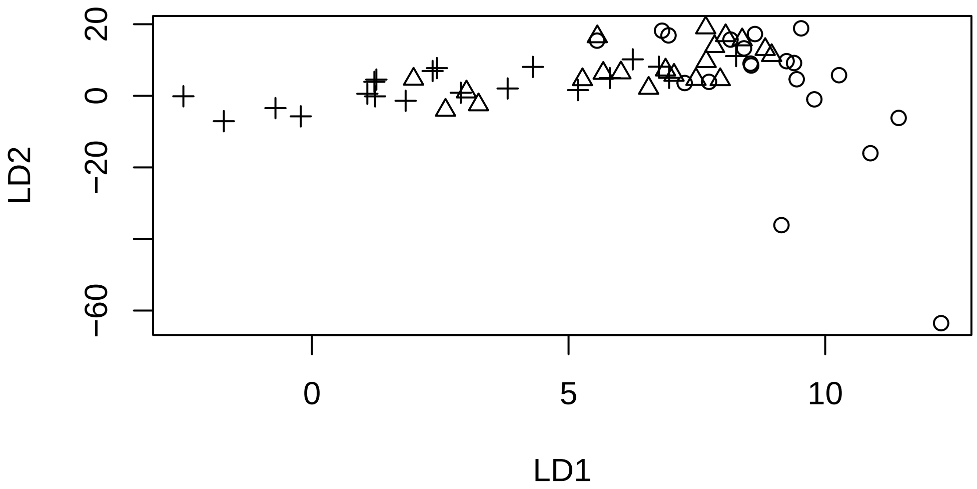
H_a : The variable PRECIP has a significant effect on group separation

With a Lambda stat of 0.9948165 and a p-value of 0.8713416049, we do not reject the null hypothesis at $\alpha^* = .01$, and so there is insufficient evidence to conclude that PRECIP, individually, contributes significantly to group separation adjusting for other variables.

C.2 Part 4

Produce a plot of the first two linear discriminant functions. Be sure to include this plot in the report. Comment on the plot with regard to how well the discriminant functions separate the groups in the data. NOTE: When using the corresponding function, the system will wait for you to click on the graph to place a legend for the symbols used in the graph. Be sure to click on a location that is empty for the legend placement. If the legend placement causes your system to lock up, modify the original function and remove the code that inserts the legend. In that event, you can manually insert your own legend by annotating the graph

```
discr.plot(pollution[,-1], pollution[[1]])
```



LD1 seems to separate a decent amount of the pluses and circles when looking at the projection onto LD1. However, there still seems to be some overlap between a few of the plues, a few of the circles, and many of the triangles.

LD2 does not separate the groups well at all.

C.3

C.3 Part 1

Specify the linear classification functions for each of the group levels in the data.

For classification analysis, we will use the four most important variables, as determined by the discriminant analysis. These variables are: NONWHITE, SO2, NOX, and EDUC.

```
pollution <- as.data.frame(pollution)
linclass <- lin.class(pollution[,-c(1,2)],pollution[,1])
linclass
```

```
$coefs
         [,1]      [,2]        [,3]       [,4]
[1,] 20.47375 0.1161389 -0.09341959 0.06391368
```

```
[2,] 19.89371 0.2146811 -0.11154897 0.07789059
[3,] 19.30087 0.3883845 -0.11797003 0.09583682

$c.0
[1] -116.6708 -111.0938 -108.3684
```

The linear classification functions are specified below:

$$L1(y) = 20.47375y_1 + 0.1161389y_2 - 0.09341959y_3 + 0.06391368y_4 - 116.6708$$

$$L2(y) = 19.89371y_1 + 0.2146811y_2 - 0.11154897y_3 + 0.07789059y_4 - 111.0938$$

$$L3(y) = 19.30087y_1 + 0.3883845y_2 - 0.11797003y_3 + 0.09583682y_4 - 108.3684$$

C.3 Part 2

Using Observation #1 from the data, apply the classification functions from the previous step to predict which group that observation should be classified as. Then compare this to the actual group classification from the data and state whether or not the prediction was correct.

```
obs1<-pollution[1,-c(1,2)]
lincomb <- linclass$coefs %*% t(obs1) + linclass$c.0
rownames(lincomb) <- c('L1', 'L2', 'L3')
colnames(lincomb) <- c('Lin. Combo')
knitr::kable(lincomb, row.names = T)
```

	Lin. Combo
L1	130.6597
L2	128.9176
L3	124.7798

The predicted classification of Observation #1 is MORTRANK Group 1, because this corresponds to the linear classification function that resulted in the largest value $L1(y) = 130.66$. Observation #1 belongs in Group 1, so the prediction is correct.

C.3 Part 3

Based on linear classification functions you generated, provide the corresponding "Confusion Matrix", Apparent Error Rate, and Apparent Correct Classification Rate. Comment on how well the linear classification functions are classifying observations.

```
pol_rates<-rates(pollution[,-c(1,2)],pollution[,1])
knitr::kable(pol_rates[[4]], caption = "Confusion Matrix", row.names = T)
```

Table 6: Confusion Matrix

	1	2	3
1	15	5	0
2	5	11	4
3	1	5	14

Correct Classification Rate:

```
pol_rates$`Correct Class Rate`
```

```
[1] 0.6666667
```

Error Rate:

```
pol_rates$`Error Rate`
```

```
[1] 0.3333333
```

Apparent Error Rate (AER): 0.333 Apparent Correct Classification Rate (ACCR): 0.667

When taking into consideration racial demographics (NONWHITE), pollution (SO2 and NOX), and education level (EDU), the linear classification functions assign observations to the correct age-adjusted mortality rank (MORTRANK) 66.7% of the time and assign them to an incorrect group 33.3%. This means that on average 1 out of every 3 classifications will be incorrect.

Section D

For our multivariate analysis, we investigated the “pollution” dataset, which contained age-adjusted mortality rate (inclusive of all causes) for 60 United States cities. The grouping variable, designated MORTRANK, separated these cities into 3 groups of 20, with a value of “1” corresponding to the group with the lowest mortality rate and a value of “3” representing the highest mortality rate.

Preliminary analysis, using the distributions across each variable in the dataset, revealed that the MORTRANK Group 1 was notably different from the other groups. Group 1 exhibited the highest variance in precipitation, smallest percentage of nonwhite mortality, and very large nitrogen oxide pollution potential outliers. For this reason, Group 1 displayed the highest degree of separation from the other groups using PCA.

Upon conducting the discriminant and classification analyses, we did not find any significant correlations between the variables, so we were able to proceed with them all included. We found one significant discriminant function that contributed to group separation, as well as two variables (Nonwhite % and sulfur dioxide pollution potential) that exhibited strong evidence of contributing to group separability. Using the 4 most important variables, we obtained 3 classification functions to predict the MORTRANK of future observations. While the classification functions had an Apparent Correct Classification Rate (ACCR) of 66.7%, this essentially means that a third of observations will be incorrectly classified. According to the confusion matrix, Group 1 received the most correct classifications, which is consistent with the group separation we observed during our preliminary analysis.

We found it surprising that sulfur dioxide pollution and nitrogen oxide pollution were not correlated with each other, and that only sulfur dioxide contributed to group separation, considering they are both criteria pollutants that are byproducts of electricity generation, industrial processes, and automobile emissions that result in similar side effects (breathing difficulties).

If we were to change one thing about our analysis, we would opt for a quadratic discriminant analysis, which would offer a more flexible, non-linear decision boundary for class assignments. We believe this would be a more appropriate fit for the data as we suspect a significant, non-linear classification rule to exist between the variables that would split the data in a more concise manner. Lastly, we believe that including variables such as the per capita crime rate variable would be a useful addition to this dataset. We would expect that cities with higher MORTRANK (Group 3) would be more easily separated from the lower groups by a crime rate variable. To ensure that our analysis is ethical, we would also like to include protected features in our dataset, such as race and sex demographics. We wouldn’t use this type of data in fitting our model, but if we had that type of data, in our analysis, we could take into consideration if our model was discriminating towards cities that had these protected characeristics.