Edx Capstone 2 Project

Engineers and Hairstylists

October 21, 2022

M. A. White

Table of Contents

Introduction

A recent report by the Society of Women Engineers and the National Society of Black Engineers (Rincon and Yates, 2017).noted that there has been an ongoing call, for decades, to diversify the workers within engineering professions. Efforts to increase gender and racial diversity in engineering has occurred in education (K-12 and postsecondary) as well as in the workplace. However, the report noted that even with the increased amount of resources dedicated to the effort along with research and attention, women and women of color are still underrepresented in nearly all engineering occupations.

Underrepresentation is a term that defines the expected percentages, say in a given occupation, relative to their actual numbers, as in employment. Underrepresentation means that the group's percentages (by gender, race, ability status, etc.) in an occupation do not equal their percentages of the general workplace population. The report referenced above and entitled, *Women of Color in the Engineering Workplace*, notes that in education, "only 20% of all engineering bachelor's degree holders are women, and …less than 4% of engineering bachelor's degrees are awarded to African American, Hispanic, and Native American women combined" (Rincon and Yates, 2017).

The low numbers of women in engineering education courses transfers over into the workplace. The latest estimate is that women of color make up less than 2% of all engineering professionals. Again, the *Women of Color in the Engineering Workplace* report (2017) found that, "in addition, approximately one in four women leaves the engineering profession within the first five years, a rate much higher than their male counterparts."

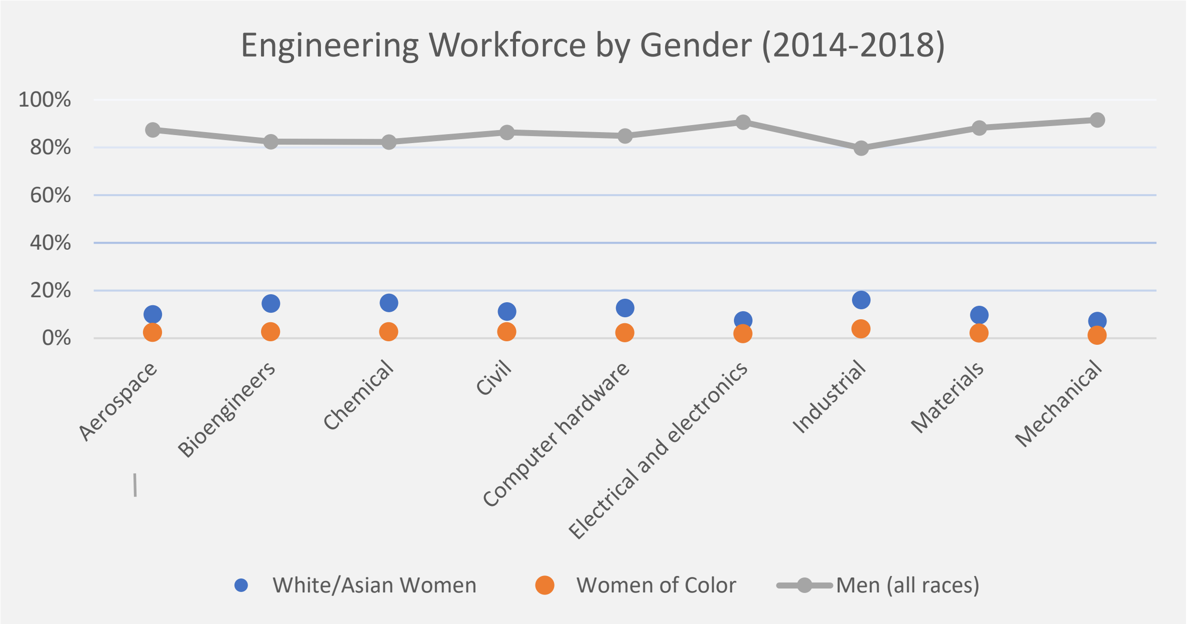
The table below provides a quick overview of the representation of women in the U.S. by race/ethnicity and engineering occupation.

Race/Ethnicity (alpha order), % of Population, and % in Engineering Occupations

Race/Ethnicity	Female Population Estimates (2021)	% Female in Science and Engineering jobs
American Indian	1.3%	<1%
Asian	3%	7%
Black	7%	2%
Hispanic	8%	2%
Native Hawaiian	0.3%	<1%
Two or more races	2.9%	N/A
White	31%	18%

Source: US Census QuickFacts (https://www.census.gov/quickfacts/fact/table/US/PST045221)

The chart below, developed in Excel, gives a visual of the rate of all women and all men working in the U.S. field of engineering based on the EEO tables from 2014-2018. Women of Color, defined here as American Indian, Black, Hispanic, Native American, and Native Hawaiian/Pacific Islander, are compiled into one category and labeled as Women of Color.



Source: U.S. Census Bureau EEO Tables (2014-2018)
https://www.census.gov/acs/www/data/eeo-data/eeo-tables-2018/

A number of researchers have examined and continue to explore reasons for the underrepresentation of women in some science, technology, engineering, and math (STEM) education fields as well as the high attrition rates of women in STEM jobs.

Recently, a collaborative group of engineering professionals, professional associations, and industry leaders decided to provide resources to support and extend opportunities for women of color in the engineering workplace. Interestingly, this group defined women of color as a woman who identifies as being one of the following racial/ethnic groups: American Indian, Asian, Black, Hispanic, or Pacific Islander/Native Hawaiian.

This study's researcher was interested in the inclusion of female, Asian American engineers as Women of Color because the stereotype of Asians is that they are overrepresented in STEM fields. In fact, the National Science Foundation does not include Asians in their definition of underrepresented racial and ethnic groups in STEM. They, instead, list Black, Hispanic, American Indians, or Alaska Natives (by race and ethnicity). A recent NSF publications states that Asians are overrepresented among science and engineering degrees as well as in science and engineering professions (https://www.nsf.gov/statistics/2017/nsf17310/digest/introduction/). Also, as can be seen in some demographic data, it appears that Asian women are represented in proportion to their segment of the general population.

However, it is obvious that all women are underrepresented in engineering professions compared to men. Is it important, therefore, to classify females in engineering by race and ethnicity? How dramatic are differences in representation by subgroups? Are some racial/ethnic groups' representation more alike or have more affinity with one group than another? These are some of the questions that led to this study and analysis. The question of affinity does not answer the overall question of underrepresentation, but as one researcher has suggested, treating women as a homogeneous group can potentially risk obscuring important racial and ethnic difference among women in engineering (Johnson, 2011). These are differences that may be an important lever or variable to explore while conducting continued research on underrepresentation.

The purpose of this research paper is to identify a machine learning technique that might help classify the racial/ethnic groupings of women in engineering and to examine the probability of whether one racial/ethnic group has some or more affinity to another group. The report uses a machine learning approach that includes first training an algorithm then applying it to a future (test set) to make a prediction of the categorial outcomes (Irizarry, 2019).

Project Overview

This research project is the second product of the EdX Capstone course where students are allowed to choose their own project. The impetus for this project is based on the question of whether, for females in engineering, one racial/ethnic groups' representation is more alike or have more affinity with another group. The goal is to see if there is a way to identify racial/ethnic differences and similarities that might serve future research on the topic of underrepresentation. This report is one of three project files: 1.) this report in PDF form, 2.) an RMD file, and 3.) a file with the R script that performs the report's machine learning tasks. Access to these three files along with their related datasets are included in the following GitHub repository, linked here: https://github.com/mwhiteaz/Capstone2.git

Dataset

The dataset used in the report is from the Equal Employment Opportunity tables of the U.S. Census Bureau. According to its website, the Equal Employment Opportunity Tabulations has provided "the primary external benchmark for comparing the race, ethnicity, and sex composition of an organization's internal workforce, and the analogous external labor market, within a specified geography and job category" (US Census Bureau, 2022). The goal of the tables is to assist organizations in reviewing workplace data related to Equal Employment Opportunity. This tabulation began in 1970 and continues today though the Census Bureau. The data used in this report were developed from 5-year American Community Survey data. In this case the five-year span included 2006-2010 and 2014-2018. The complete data file contains over 450 Census occupation categories based on the 2010 Standard Occupational Classification (SOC). This tabulation included estimates and percentages of the labor force for race and ethnicity by sex for all counties and for places of 50,000 or more, covering nearly 6,500 geographic entities.

For this study, the dataset was reduced in this analysis to simply include race/ethnicity and occupation only. The occupations selected were 16 engineering fields (e.g., aerospace, mechanical, electrical), and other occupations that are more often considered to be more stereotypically "female" careers such as being a nurse and a K-12th grade teacher. In fact, according to a recent article by the International Labour Organization, the percent of employment by gender and occupation from 121 countries show that hairstylists, health care professionals, and teachers make up four of the top six occupations dominated by women workers.

The dataset for this report used census data so was a little different than a data set of a sample or subset file of the population. The differences associated with census data, which is practically the whole population, is not due as much to random chance as in sampling. According to this article, the same statistical techniques can be used with census data although it is harder to attribute differences to random chance as in sampling.

Methods

In this section, the methods and analysis of the study will be covered in order to explain the process and techniques used. This includes data cleaning, data exploration and visualization, and the modeling approaches used.

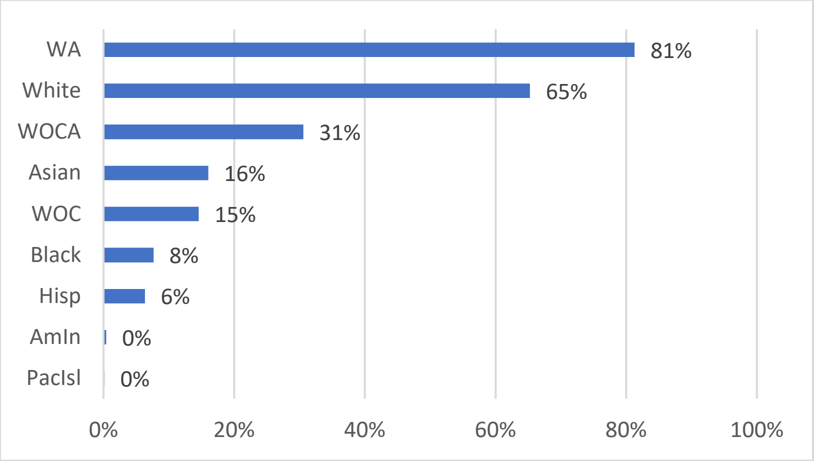
The goal of the project is to figure out which method of classification can best identify affinities between the selected racial/ethnic groups. Namely, classification algorithms such as K Means Cluster and linear discriminant analysis (LDA) were used to determine which is best able to identify affinities and differences between the racial/ethnic groups of females in engineering and in other occupations.

Data cleaning

As mentioned earlier, the dataset used was from the Equal Employment Opportunity Online Tables. The website has a data tools page which allows for an immediate download of an Excel (.xls) file. Selected fields included EEO Occupation Codes by Race/Ethnicity. The geography selected can be by nation, state, or county. Nation was selected. The Occupation Codes selected were a number of Engineering Codes ranging from aerospace engineers (1305) to petroleum and mining engineers (code 1520). Also included as engineering was the code for Drafters, Engineering Technicians, and mapping technicians. To help compare race and gender by occupation, more stereotypically “female” occupations were selected such as all Nursing, Teachers (PreK-12), and Hairstylists or Personal Care Workers.

Once the data file was downloaded into Excel, the researcher manually created two additional Race Categories which included women of color or “WOC” (Black, American Indian, Hispanic, and PI) and WOCA (same as WOC but also included Asian). One other combination included was “WAS” which was the total of White and Asian counts. The Excel chart below displays the percentages of female within the dataset, by racial group:

Females by Race/Ethnicity in dataset

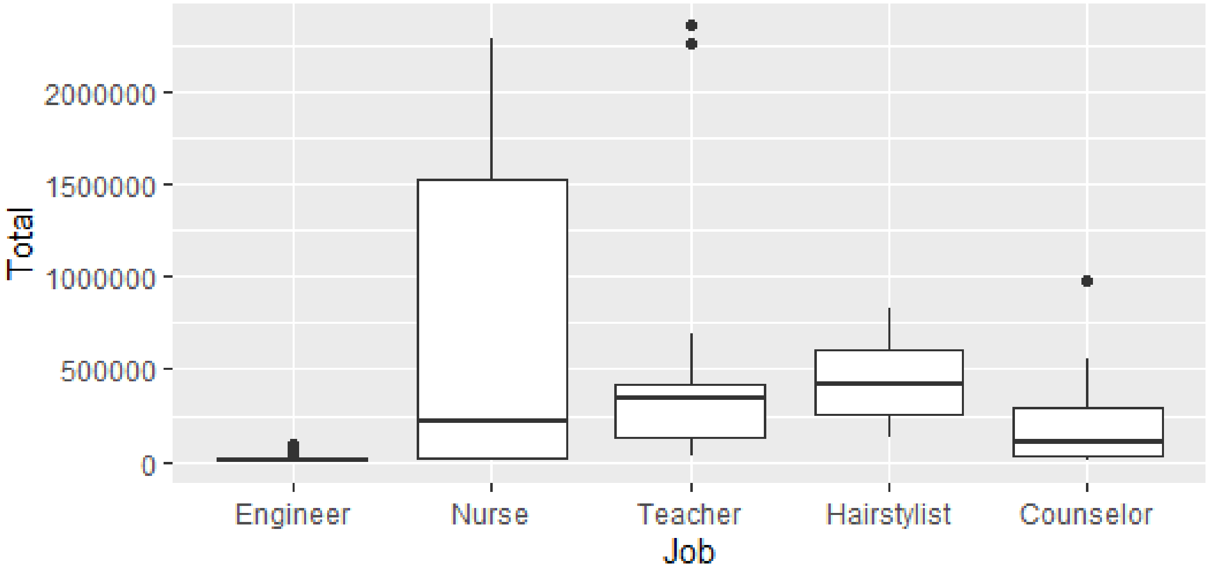


Male data was removed from the table for both years in the analysis presented below. Those with two or more races were assigned to the WOC or WOCA category except for the White/Asian group which was only assigned to the WOCA variable. Race was recoded as either a factor or numeric variable and the various sub categories for all occupation codes were simplified to be one of the final five: Engineer, Nurse, Teacher, Counselor, and Hairstylist.

Data exploration and visualization

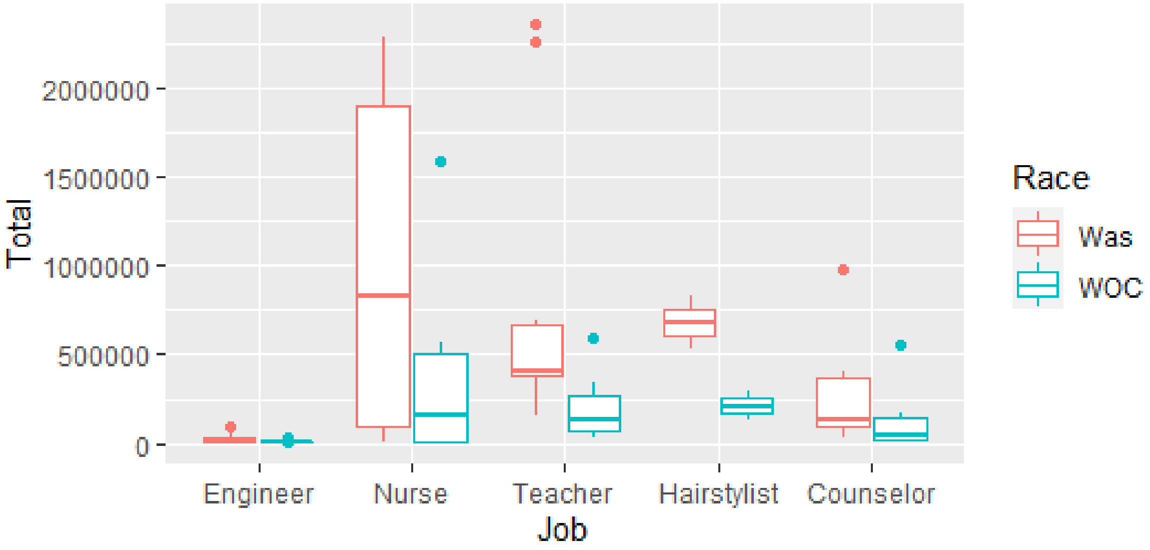
There were a few variations of the final dataset used in the analysis, but the base data set included 112 rows and 23 columns. The variables included the EEO table Year, the occupation code, the occupation recoded into one of the 5 careers (Engineering, Counselor, Nurse, Teacher, Hairstylist), Total count, Gender, and each Racial/Ethnic category total then percentage of the total. The races in the base data set included American Indian, Asian, Black, Hispanic, Native Hawaiian/Pacific Islander, and White.

As shown in the chart below, women have a very low representation in engineering jobs, but are very well represented in other fields such as nursing. The chart isn't comparative to men, it just compares female employment by selected occupation in the dataset.

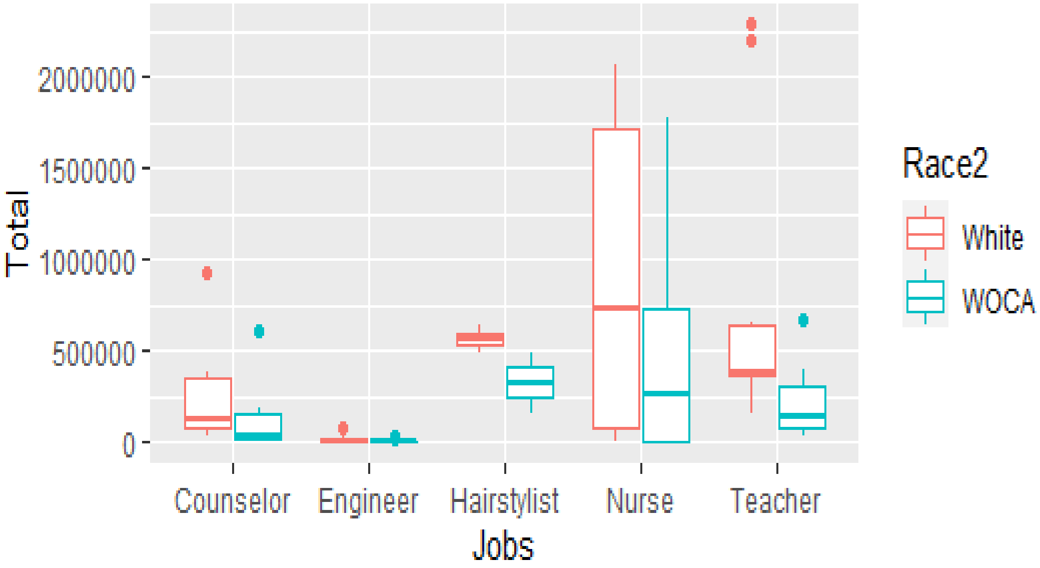


For some researchers, women of color in engineering includes all racial/ethnic groups other than White. There has been some debate on whether Asian American women should be classified alongside their White peers or whether their representation is more like women of color. As shown in the two charts below, regardless of where Asian women were placed (with White women (WAS)) or with women of color (WOCA)), women were least represented in engineering compared to other fields. For teachers, the very high outliers are the individuals who comprise the teacher-aide type positions in the K-12 school setting.

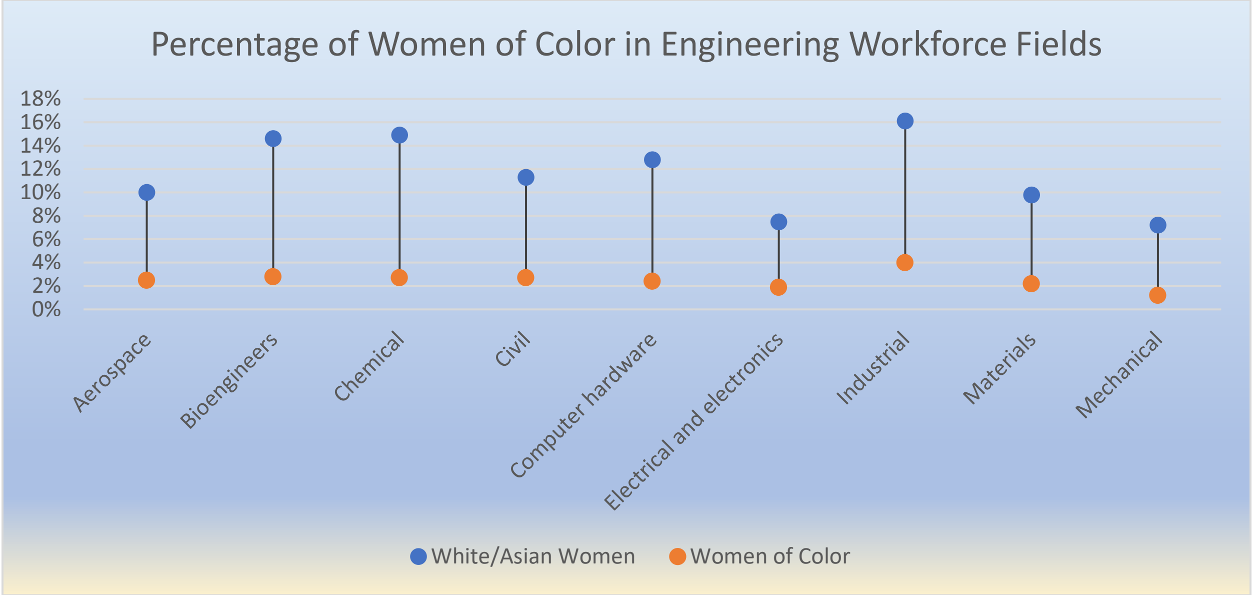
Asian and White vs WOC in occupations



Asian and WOC vs White-only in occupations



Finally, this dumbbell chart displays, for 2014-2018 data only, the distribution of two groups within the different fields of engineering. This split is the WAS (White and Asian) versus WOC (Women of Color). Females are least represented in electrical and mechanical engineering and most represented in bio, chemical, and industrial engineering.



Source: U.S. Census Bureau EEO Tables (2014-2018)
https://www.census.gov/acs/www/data/eeo-data/eeo-tables-2018/

Modeling Approach

A number of approaches were used to explore the dataset and are described here.

1.) Naïve Bayes classification. At first glance, this approach seemed best suited for the research project. This approach is used to help solve classification problems using probability. The goal is to predict "membership probabilities" for each group. This approach assumes independence among the variables and was supported by various online research and articles.

2.) The second approach was the Wilcoxon test in R. This test allows for a comparison between two groups as a non-parametric test. This test was selected based on the data type for this study. The Wilcoxon test in R performs both the Mann-Whitney-Wilcoxon test and the Wilcoxon signed-rank test which is used for independent and dependent samples respectively.

3.) K Means Clustering is a technique that uses clustering to find subgroups within a dataset. This is one way to classify observations or groups into the same group or against different groups. Citation

4.) Linear Discriminant Analysis (LDA). This approach is used to help solve classification problems as well. Linear boundaries are produced and the model predicts which cases belong within a It is basically a dimensionality reduction technique. LDA tries to predict the class of the given observations.

These approaches were all selected in order to examine affinity within racial/ethnic groupings and occupations. The goal was to uncover similarity between racial/ethnic groupings by occupation. This was also expected to help answer whether, in engineering, are Asian females more aligned to their White counterparts or to women of color.

Results

This section presents the modeling results for the four selected algorithms in the order presented above. The modeling analysis and results are included in this section. The following section, the conclusion, provides a discussion of the best model performance.

The results section is listed in the order of the machine learning techniques listed above.

1.) Naïve Bayes classification,
2.) Wilcoxon test,
3.) K Means Clustering, and
4.) Linear Discriminant Analysis.

1.) Naïve Bayes Classification

For the first analysis, the probability of White and Asian women being in certain careers was compared to that of Women of Color (WOC). The code for the analysis is provided below along with a table that displays the conditional probabilities, by racial/ethnic group and occupation. Because the data is based on summary, census data rather than a sample, the A-priori probabilities reflect the higher number of careers in the set for engineering. The four career choices included in this analysis were:

1. Engineer
2. Hairstylist
3. Nurse
4. Teacher

The code for the analysis included the following:

```
# Installing Packages
install.packages("e1071")
install.packages("caTools")
install.packages("caret")

# Loading package
library(e1071)
library(caTools)
library(caret)
df3 <-NBtest3
Job3 <- as.factor(df3$Job)
df4=as.data.frame(df3)
# Splitting data into train and test data
split <- sample.split(df4, SplitRatio = 0.7)
train_cl <- subset(df4, split == "TRUE")
test_cl <- subset(df4, split == "FALSE")

# Feature Scaling
train_scale <- scale(train_cl)
test_scale <- scale(test_cl)
```

```
# Fitting Naive Bayes Model to training dataset
set.seed(1)  # Setting Seed
classifier_cl <- naiveBayes(Job ~ ., data = train_cl)
classifier_cl

# Predicting on test data'
y_pred <- predict(classifier_cl, newdata = test_cl)

# Confusion Matrix
cm <- table(test_cl$Job, y_pred)
cm

# Model Evaluation
confusionMatrix(cm)
```

```
Results
Naive Bayes Classifier for Discrete Predictors
Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
              1            2            3            4
     0.61290323   0.03225806   0.12903226   0.22580645

Conditional probabilities:

  Hisp
Y       [,1]      [,2]
  1  1279.211  1707.293
  2 38380.000        NA
  3 57403.750 87749.597
  4 71357.857 93292.296

  White
Y         [,1]       [,2]
  1   14447.89   14333.77
  2  494440.00         NA
  3 1011631.25 1088043.39
  4  916047.14  914616.25

  Black
Y         [,1]      [,2]
  1   1462.632   1453.11
  2  79830.000        NA
  3 142733.750 161192.03
  4 111602.857 101008.17

  AmIn
Y         [,1]       [,2]
```

```
  1   80.21053   86.33757
  2 2705.00000         NA
  3 5132.50000 5770.38055
  4 5502.85714 4402.98276

   Asian
Y         [,1]       [,2]
  1   3942.895   4658.083
  2  36670.000         NA
  3 108966.250 121872.201
  4  34075.000  28649.606

   PacIsl
Y         [,1]       [,2]
  1   37.84211   55.34364
  2  340.00000         NA
  3 1145.00000 1338.16417
  4 1005.71429  748.34500

   WOC
Y         [,1]       [,2]
  1   2859.895   3127.133
  2 121255.000         NA
  3 206415.000 246291.122
  4 189469.286 187527.340

   WOCA
Y         [,1]       [,2]
  1   6802.789   7609.813
  2 157925.000         NA
  3 315381.250 365838.794
  4 223544.286 214738.212

   WA
Y         [,1]       [,2]
  1   18390.79   18516.85
  2  531110.00         NA
  3 1120597.50 1209570.57
  4  950122.14  942277.97

>
> # Predicting on test data'
> y_pred <- predict(classifier_cl, newdata = test_cl)
>
> # Confusion Matrix
> cm <- table(test_cl$Job, y_pred)
> cm
>
> # Model Evaluation
> confusionMatrix(cm)

Confusion Matrix and Statistics
```

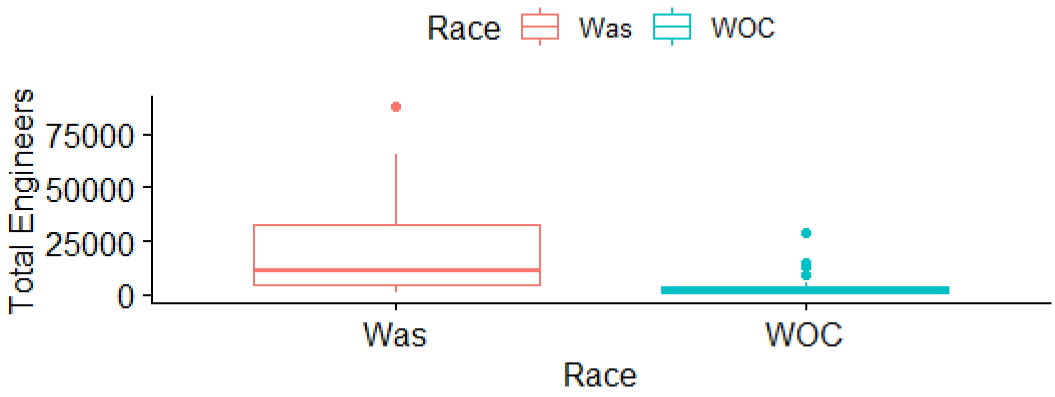
```
   y_pred
    1 2 3 4
  1 9 0 0 2
  2 0 0 1 0
  3 1 0 1 0
  4 0 0 0 4

Overall Statistics

               Accuracy : 0.7778
                 95% CI : (0.5236, 0.9359)
    No Information Rate : 0.5556
    P-Value [Acc > NIR] : 0.04535

                  Kappa : 0.6129

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity            0.9000       NA  0.50000   0.6667
Specificity            0.7500  0.94444  0.93750   1.0000
Pos Pred Value         0.8182       NA  0.50000   1.0000
Neg Pred Value         0.8571       NA  0.93750   0.8571
Prevalence             0.5556  0.00000  0.11111   0.3333
Detection Rate         0.5000  0.00000  0.05556   0.2222
Detection Prevalence   0.6111  0.05556  0.11111   0.2222
Balanced Accuracy      0.8250       NA  0.71875   0.8333
```

The model only achieved 78% accuracy with a p-value of less than .05. The second career option, Hairstylist returned N/A values for some reason. But for the other occupations, the Sensitivity, Specificity, and Balanced accuracy were good. But it is skewed by the number of engineering occupations included in the data set compared to the others. The analysis attempted to examine the probability of “belonging” to Engineering or other fields, based on the probability of being from one racial group or another.

The probabilities, however, may be of interest to future predictions related to job employment by gender, but the analysis does not shed much light on the differences, by race, within each occupation. For each race, Hispanic, Black, White, Asian, and American Indian, the probabilities for being a nurse or a teacher were the highest.

2.) Wilcoxon test in R. In the second analysis, the Wilcoxon test was employed to examine mean differences between the racial/ethnic groupings. The charts and results below compare the two racial groupings for women in engineering only.

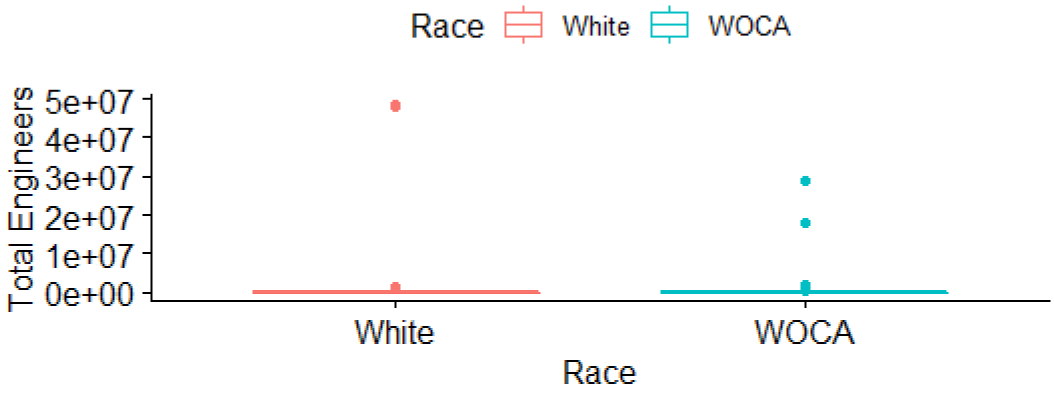
White and Asian women vs WOC



The results of the Wilcoxon rank sum exact test are shown below.
W = 749, p-value = 3.463e-06
alternative hypothesis: true location shift is not equal to 0

The p-value of the test is 3.463, which is much higher than the significance level alpha = 0.05. We can conclude that there is no significantly different from the two groups.

Chart X, White vs WOC along with Asian



The results of the Wilcoxon rank sum exact test are shown below.

W = 650, p-value = 0.06459

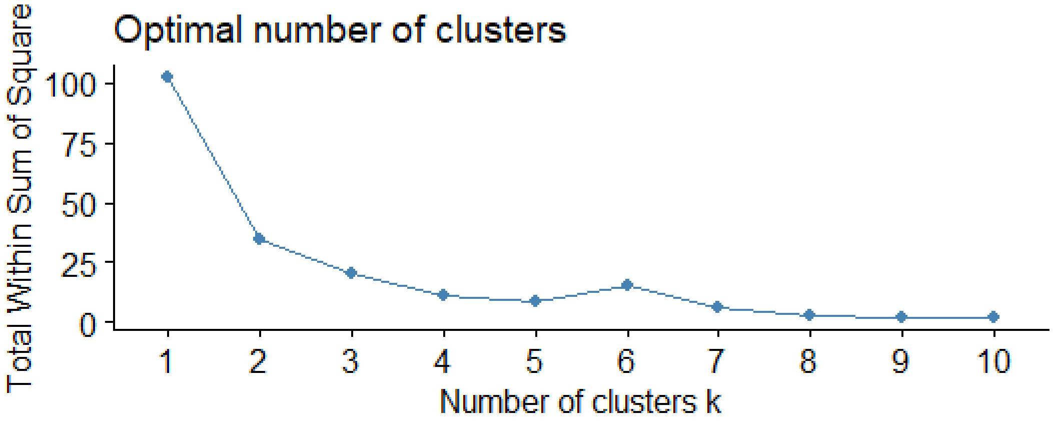
alternative hypothesis: true location shift is not equal to 0

The p-value of the test is 0.06459, which is kind of close to the significance level alpha = 0.05. We can conclude that there may be a slight difference between the two groups.

3.) K Means Clustering. The third analysis used K Means clustering to see if the groups would form natural clusters by racial/ethnic groups and occupations. The code used for the analysis is below.

```
library (tidyverse)
library (cluster)
library (factoextra)
df30<-LDA
df30<- na.omit(df30)
df30$Race <- as.numeric(as.factor(df30$Race))
df30 <- scale(df30)
head(df30)
distance <- get_dist(df30)
fviz_dist(distance, gradient = list(low = "#00AFBB", mid = "white", high = "#FC4E07"))
k2 <- kmeans(df30, centers = 2, nstart = 25)
str(k2)
k2
```

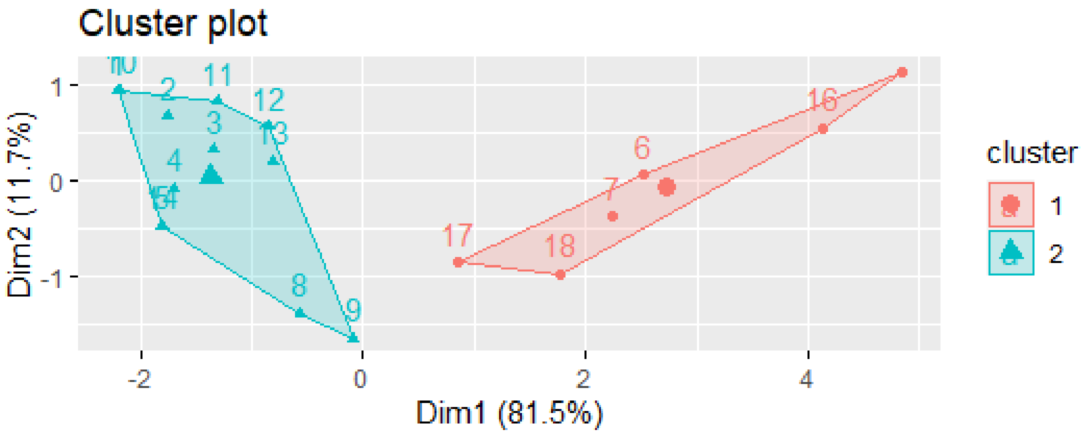
The “Elbow Method” was used and, as shown in the chart below, suggested that 2 clusters be used in the analysis.



The charts below show the clustering of racial groups by varied professions. The racial groups are defined as:

Race =
1 AI
2 Asian
3 Black
4 Hispanic
5 PI
6 WAS
7 White
8 WOC
9 WOCA
10 AI
11 Asian
12 Black
13 Hispanic
14 PI
15 WAS
16 White
17 WOC
18 WOCA

Applying the center at 2 with the code above below provided the following results. As can be seen, the groups are separated so that White and WAS are different than the other racial/ethnic groups except for the WOC group (17) and WOCA group (18).



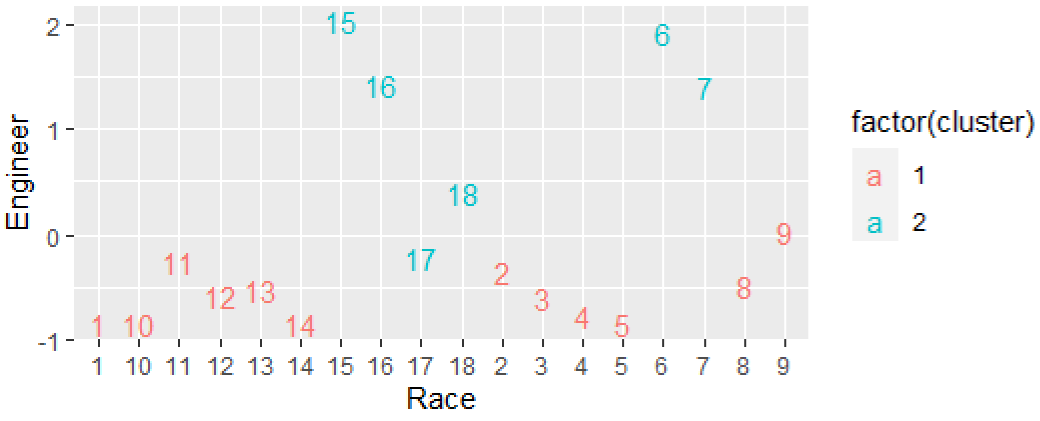
The results are presented below and show the cluster to which the different racial groups are allocated to. The majority are to Nurse and Teacher in both groups.

```
Cluster means:

Within cluster sum of squares by cluster:
[1] 20.41796 14.42998
 (between_SS / total_SS =  65.8 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"
[5] "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"


# A tibble: 2 × 7
  Cluster  Race Engineer Counselor    Nurse  Teacher Hairstylist
    <int> <dbl>    <dbl>     <dbl>    <dbl>    <dbl>       <dbl>
1       1    NA   216411    862380 8700681. 4446058.      541445
2       2    NA   30215.   140499.  767404.  458394.      75742.
```

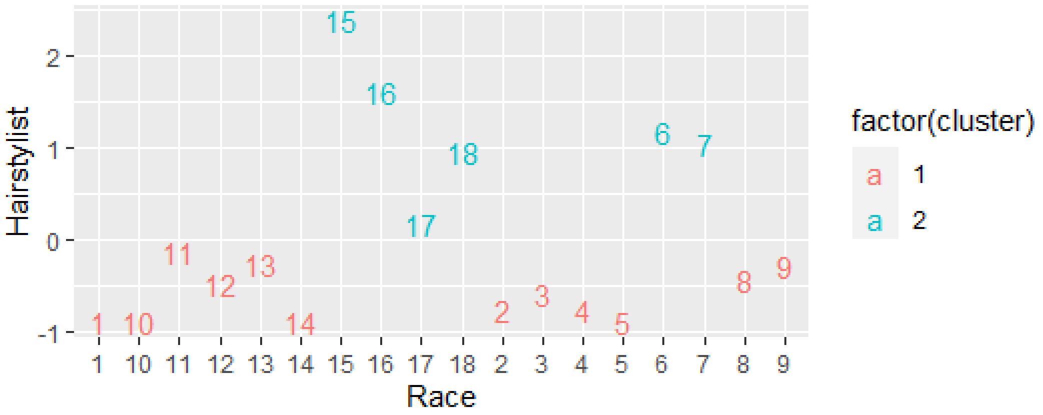
To further explore the data, individual occupations were selected for review. The groupings, by occupation, do not differ widely from the overall chart on the page above. In addition to discrete occupations, the code includes a request to display the groupings based on different cluster values.

```
df30 %>%
  as_tibble() %>%
  mutate(cluster = k2$cluster,
         Race = row.names(LDA)) %>%
  ggplot(aes(Race, Engineer, color = factor(cluster), label = Race)) +
  geom_text()
df30 %>%
  as_tibble() %>%
  mutate(cluster = k2$cluster,
         Race = row.names(LDA)) %>%
  ggplot(aes(Race, Hairstylist, color = factor(cluster), label = Race)) +
  geom_text()
df30 %>%
  as_tibble() %>%
  mutate(cluster = k2$cluster,
         Race = row.names(LDA)) %>%
  ggplot(aes(Race, Nurse, color = factor(cluster), label = Race)) +
  geom_text()
df30 %>%
  as_tibble() %>%
  mutate(cluster = k2$cluster,
         Race = row.names(LDA)) %>%
  ggplot(aes(Race, Teacher, color = factor(cluster), label = Race)) +
  geom_text()

k3 <- kmeans(df30, centers = 3, nstart = 25)
k4 <- kmeans(df30, centers = 4, nstart = 25)
k5 <- kmeans(df30, centers = 5, nstart = 25)

# plots to compare
p1 <- fviz_cluster(k2, geom = "point", data = df30) + ggtitle("k = 2")
p2 <- fviz_cluster(k3, geom = "point",  data = df30) + ggtitle("k = 3")
p3 <- fviz_cluster(k4, geom = "point",  data = df30) + ggtitle("k = 4")
p4 <- fviz_cluster(k5, geom = "point",  data = df30) + ggtitle("k = 5")

library(gridExtra)
grid.arrange(p1, p2, p3, p4, nrow = 2)

df30 %>%
  as_tibble() %>%
  mutate(cluster = k5$cluster,
         Race = row.names(LDA)) %>%
  ggplot(aes(Race, Engineer, color = factor(cluster), label = Race)) +
  geom_text()

set.seed(123)

fviz_nbclust(df30, kmeans, method = "wss")

set.seed(123)
final <- kmeans(df30, 2, nstart = 25)
print(final)
fviz_cluster(final, data = df30)
```

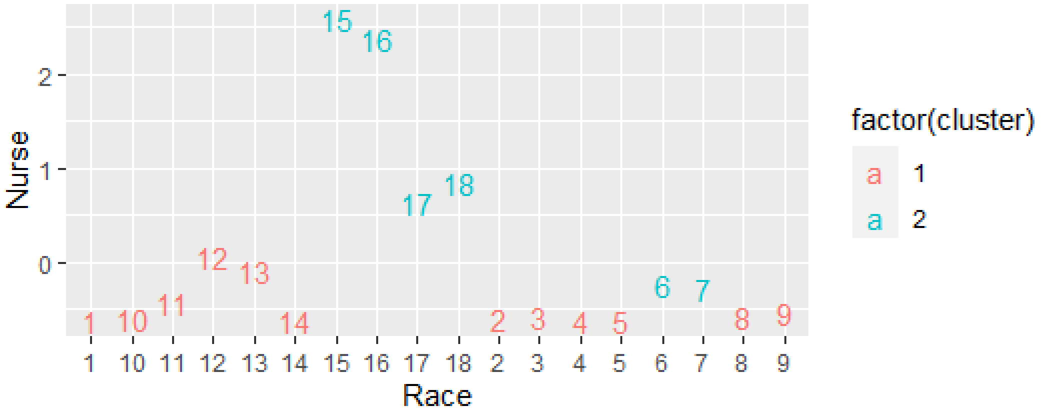
Engineer Cluster



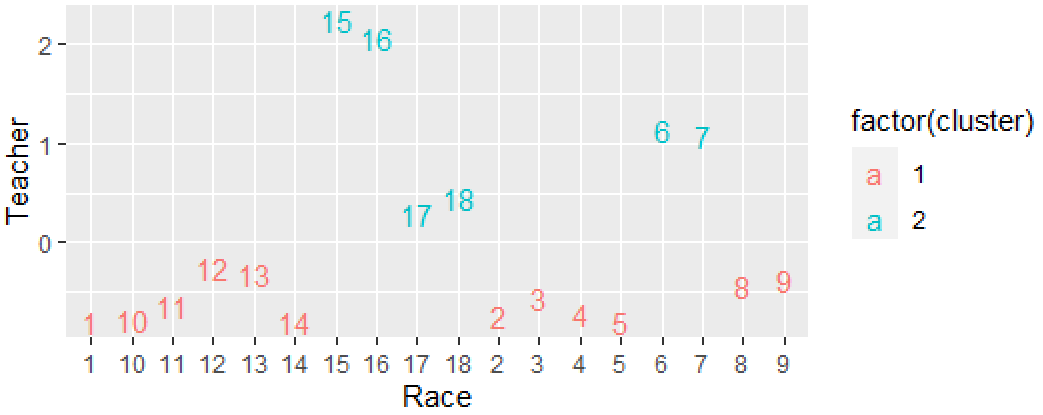
Hairstylist Cluster



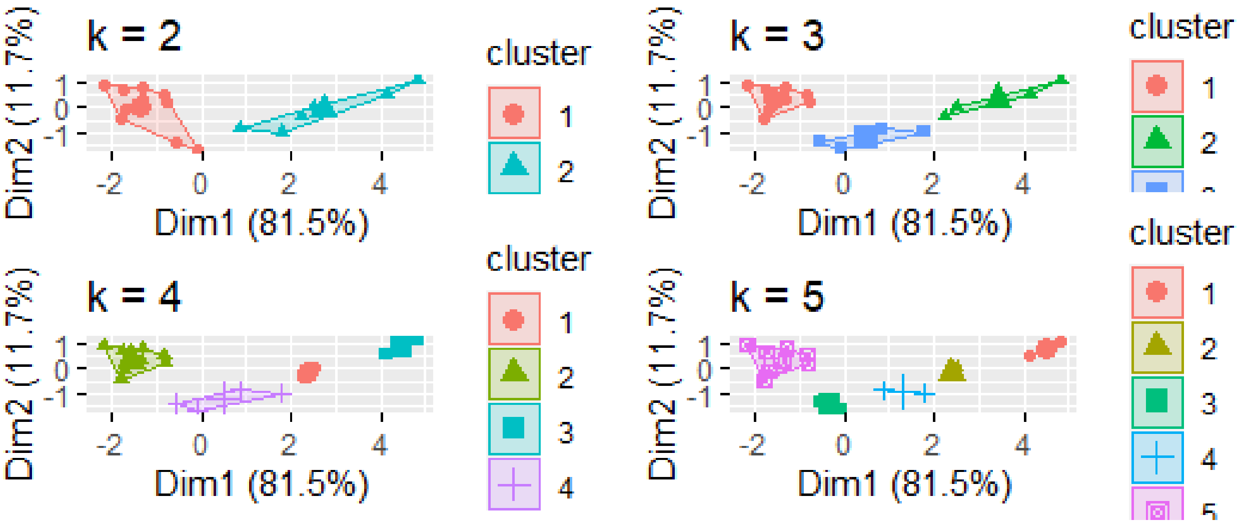
Nurse Cluster



Teacher Cluster



Although two clusters were used in the analysis, the following charts show the use of other, different values of K. Namely, 3, 4, and 5 clusters were used to examine differences. Upon closer review of individual charts, the racial groupings from the two cluster analysis was very similar to other others.



4.) Linear Discriminant Analysis. The fourth analysis attempted to distinguish affinities between the racial/ethnic groups in different professions. The code below was used to generate prior probabilities for each group, group means, and coefficients of linear discriminants. Those statistics were then used to help predict a model of groupings.

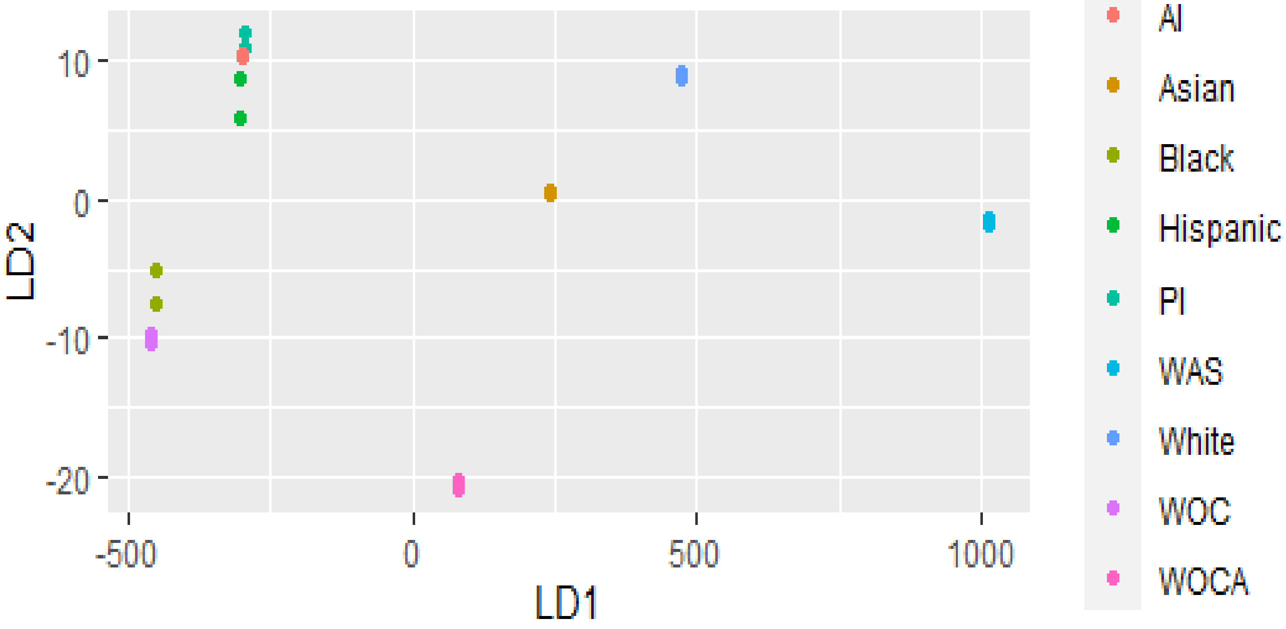
```
> library(MASS)
Attaching package: 'MASS'
The following object is masked from 'package:dplyr':
    select
Code: df7 <-LDA
model <- lda(Race ~., data = df7)
model
predictions <- model %>% predict(df7)
names(predictions)
head(predictions$class)
head(predictions$posterior)
head(predictions$x)
lda.data <- cbind(df7, predict(model)$x)
ggplot(lda.data, aes(LD1, LD2))+
  geom_point(aes(color=Race7))
mean(predictions$class ==df7$Race)


Results
df7 <-LDA
> model <- lda(Race ~., data = df7)
> model
Call:
lda(Race ~ ., data = df7)

Prior probabilities of groups:
      AI     Asian     Black  Hispanic        PI       WAS     White
```

```
0.1111111 0.1111111 0.1111111 0.1111111 0.1111111 0.1111111 0.1111111 
      WOC      WOCA 
0.1111111 0.1111111 

Group means:
         Engineer Counselor     Nurse   Teacher Hairstylist
AI         1561.5   12427.5   79014.0   42449.5      3495.0
Asian     59847.5   44742.5  664027.5  263032.5    116980.0
Black     28602.5  288695.0 1975467.5  899582.5     94202.5
Hispanic  23655.0  139575.0 1492275.0  647230.0     99465.0
PI          561.5    1852.5    6134.5   18644.0       720.0
WAS      303790.0 1028870.0 9691882.5 5500892.5    684080.0
White    243897.5  984127.5 9027855.0 5237860.0    567100.0
WOC       54380.5  442550.0 3552891.0 1607906.0    197882.5
WOCA     114228.0  487292.5 4216918.5 1870938.5    314862.5

Coefficients of linear discriminants:
                      LD1           LD2           LD3           LD4
Engineer     1.141755e-02 -2.058631e-04 -1.396712e-04 -3.598075e-05
Counselor   -1.298154e-03 -1.438310e-04  1.459382e-05 -4.794144e-06
Nurse        4.055999e-05 -5.707217e-06 -1.422485e-06 -4.472342e-07
Teacher     -1.400879e-04  4.614954e-05  8.382575e-06  1.179452e-06
Hairstylist -6.568046e-04 -7.815167e-07 -2.991371e-06  2.013411e-05

Proportion of trace:
   LD1    LD2    LD3    LD4 
0.9994 0.0005 0.0001 0.0000 
> predictions <- model %>% predict(df7)
> names(predictions)
[1] "class"     "posterior" "x"        
> head(predictions$class)
[1] AI       Asian    Black    Hispanic PI       WAS     
Levels: AI Asian Black Hispanic PI WAS White WOC WOCA
> head(predictions$posterior)
            AI Asian Black     Hispanic           PI WAS White
1 9.992958e-01     0     0 1.148157e-06 7.030927e-04   0     0
2 0.000000e+00     1     0 0.000000e+00 0.000000e+00   0     0
3 0.000000e+00     0     1 0.000000e+00 0.000000e+00   0     0
4 1.293062e-03     0     0 9.987069e-01 2.769241e-16   0     0
5 1.160697e-07     0     0 1.585722e-21 9.999999e-01   0     0
6 0.000000e+00     0     0 0.000000e+00 0.000000e+00   1     0
           WOC WOCA
1 0.000000e+00    0
2 0.000000e+00    0
3 9.151618e-15    0
4 0.000000e+00    0
5 0.000000e+00    0
6 0.000000e+00    0
> head(predictions$x)
        LD1        LD2       LD3         LD4
1 -297.6673 10.2967939 -1.939121 -0.11210055
2  243.2838  0.4021063 -8.278092 -1.40123757
3 -452.3705 -7.5172518  2.207269 -0.15097032
```

```
4 -302.1674  8.8103596 -1.297904  0.23332958
5 -292.6563 10.8708837 -2.115867 -0.09833442
6 1010.7672 -1.3673325  1.530111 -0.55832512
> lda.data <- cbind(df7, predict(model)$x)
> ggplot(lda.data, aes(LD1, LD2))+
+   geom_point(aes(color=Race7))
> mean(predictions$class ==df7$Race)
[1] 1
```

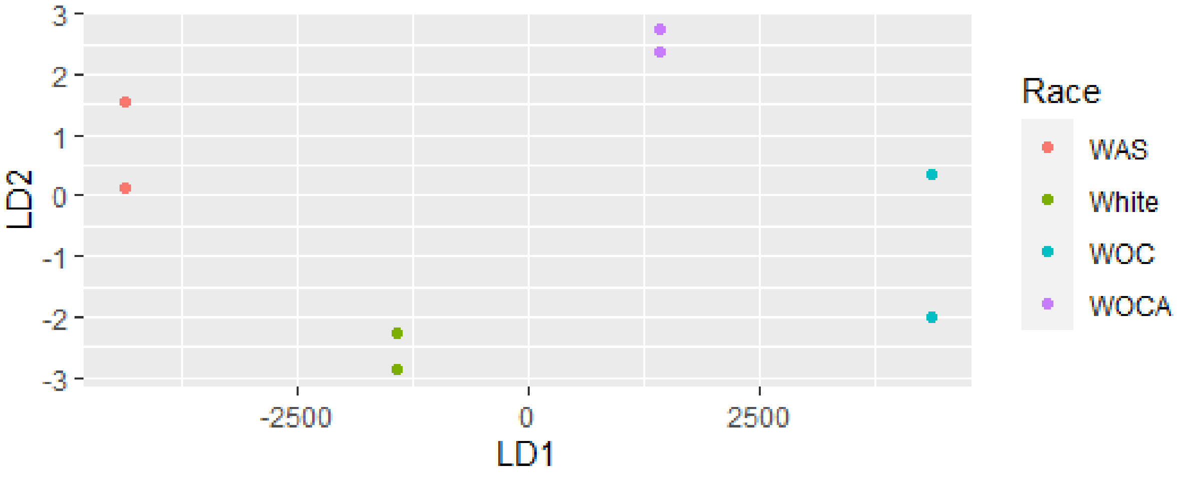
In the first analysis, all races were included and resultant graphic, shown below, displays racial ethnic group affinities within the data set.

LDA plot for females in engineering by race



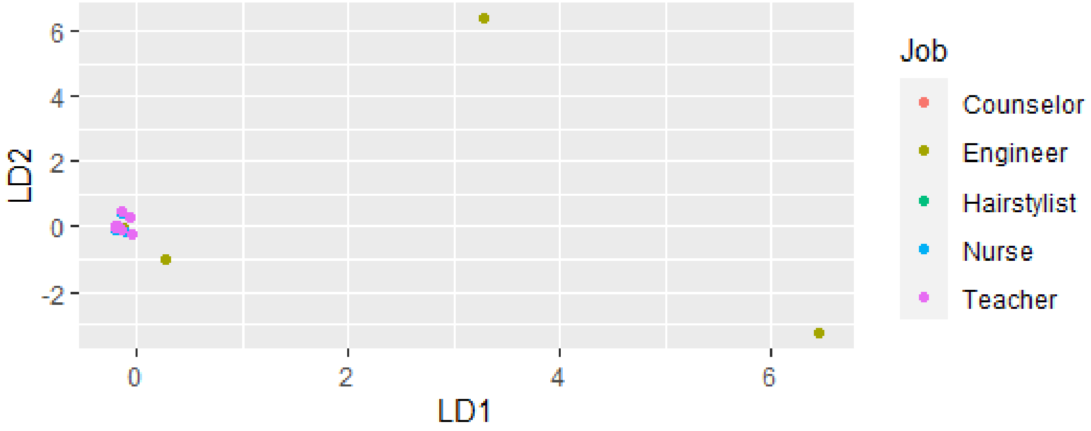
The racial/ethnic groups were recast as combinations in the second analysis, As was done earlier, the comparison was of aligning Asian with White or aligning Asian with women of color. The resultant graphic, shown below, displays the affinities by racial groups.

LDA plot for females in engineering by race



A final analysis was completed using LDA to simply review groupings by job categories as shown in the chart below. As can be seen, all stereotypical female jobs are closely aligned whereas Engineering jobs are separate, like outliers.

LDA plot for females in selected occupations



In the final step, the first model was used to predict a test model using the code below.

```
> p2 <- predict(model, test.transform)$class
> tab1 <- table (Predicted = p2, Actual = test.transform$Race)
> tab1
```

The accuracy of the model is 1 for the test set and 1 for the training set.

Conclusions

All women are underrepresented in engineering professions and for women of color, their representation is even smaller. Representation in engineering and other STEM fields is defined as the extent to a group's representation relative to their representation in the U.S. population.

A number of approaches were used to explore whether there were ways to classify or group the racial/ethnic representation within engineering for women. The algorithms from Naïve Bayes, LDA, and K-Means clustering were used along with single measure group mean differences via the Wilcoxon mean test. Of all these approaches, the LDA approach was the most promising in that it provided the clearest partitioning of the racial/ethnic groups into clusters. The model accuracy was high (1) for both the test and the training set, but that may well be a function of the underlying data frame (census data).

As shown in the chart on page 24, there were distinct differences between racial/ethnic groups and it does appear that Asian women are more aligned to the representation of White women than of woman of color. Since the outcome is categorical, the machine learning task was about classification. This output can then be, in some ways, the start of a decision rule to help predict affinity between underrepresented groups (Irizarry, 2019).

There are some new, additional tools that might serve to better identify differences in future work. Walker-Data.com identifies the dissimilarity index as a way to assess neighborhood segregation between two groups within an area, based on census data. This analysis was limited due to not having state or regional level data. Also, as stated earlier, the question of affinity between groups may not help to answer the overall question of underrepresentation, but it may shed some light on differentiating solutions. As one researcher has suggested, disaggregated data analysis can help the development of policies and programs that address the diversity of each group's experiences in the field.

References

Fonseca, L. (2019). Clustering Analysis in R using K-means. Retrieved online from: towardsdatascience.com/clustering-analysis-in-r-using-k-menas-73eca4fb7967

International Labour Organization (2020). These occupations are dominated by women. Retrieved online from: https://ilostat.ilo.org/these-occupations-are-dominated-by-women/#:~:text=Cleaning%20roles%2C%20teaching%2C%20clerical%20support,occupations%20overwhelmingly%20held%20by%20men.).

Irizarry, R. A. (2019). Introduction to Data Science: Data Analysis and Prediction Algorithms with R. Retrieved online from: http://rafalab.dfci.harvard.edu/dsbook/index.html

Johnson, D., (2011). Women of color in science, technology, engineering, and mathematics (STEM). New Directions for Institutional Research, 152, pages 75 – 85. Retrieved online from http://research.pomona.edu/janice-hudgings/files/2017/08/Johnson_2011.pdf

Naïve Bayes Classifier in R Programming (2021). Retrieved online from: https://www.geeksforgeeks.org/naive-bayes-classifier-in-r-programming/#:~:text=Naive%20Bayes%20is%20a%20Supervised,between%20the%20features%20or%20variables.

National Science Foundation (2022). Women, Minorities, and Persons with Disabilities in Science and Engineering. Retrieved online from https://www.nsf.gov/statistics/2017/nsf17310/digest/introduction/

ndthl (https://stats.stackexchange.com/users/10905/ndthl), Census Statistical Techniques, URL (version: 2012-04-30): https://stats.stackexchange.com/q/27211

R-Bloggers (2021. Linear Discriminant Analysis in R. Retrieved online from: https://www.r-bloggers.com/2021/05/linear-discriminant-analysis-in-r/

Rincon, R. M. & Yates, N. (2017). Women of Color in the Engineering Workplace. Retrieved online: https://alltogether.swe.org/wp-content/uploads/2018/02/Women-of-Color-in-the-Engineering-Workplace.pdf

Soetewey, A. (2020. Wilcoxon test in R: how to compare 2 groups under the non-normality assumption. Retrieved online from: https://statsandr.com/blog/wilcoxon-test-in-r-how-to-compare-2-groups-under-the-non-normality-assumption/

U.C. Business Analytics R Programming Guide (2022). K-means Cluster Analysis. Retrieved online from: https://uc-r.github.io/kmeans_clustering.

U.S. Census Bureau (2022). Equal Employment Opportunity Tabulation, Retrieved online from: https://www.census.gov/topics/employment/equal-employment-opportunity-tabulation.html

US Census QuickFacts (2022). Retrieved online from https://www.census.gov/quickfacts/fact/table/US/PST045221

Walker, K. (2022). *Analyzing US Census Data: Methods, Maps, and Models in R*. Retrieved online from: https://walker-data.com/census-r/index.html

Zach (2020). Linear Discriminant Analysis in R. Retrieved online from: https://www.statology.org/linear-discriminant-analysis-in-r/