College

Jason Flores UT EID: *jf36995*

This is the dataset I will be using:

```
salary_potential <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/maste
```

More information about the data set can be found at https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-03-10/readme.md

Data Prep:

```
#The data set used are already tidy
college <- salary_potential %>%
  filter(state_name %in% c("Texas","California"))
```

Introduction:

The purpose is to analyze the colleges between California and Texas (`state_name`) to see a correlation between `early_career_pay` and `mid_career_pay`.

The data set used in this project is taken from the GitHub thread `rfordatascience/tidytuesday`. The focus will be on the salary potential from the colleges in California and Texas. These two are picked because recently Californians have been moving to Texas, specifically to Harris county. The salary potential data is taken from the website `payscale.com`. It shows how much salary potential a recent bachelor's graduate can earn. The data was filtered by California and Texas, where the variables picked are the `rank`, `early_career_pay`, `mid_career_pay`, `make_world_better_percent` and `stem_percent`. The data set `college` has 50 observations with 7 variables.

Clustering

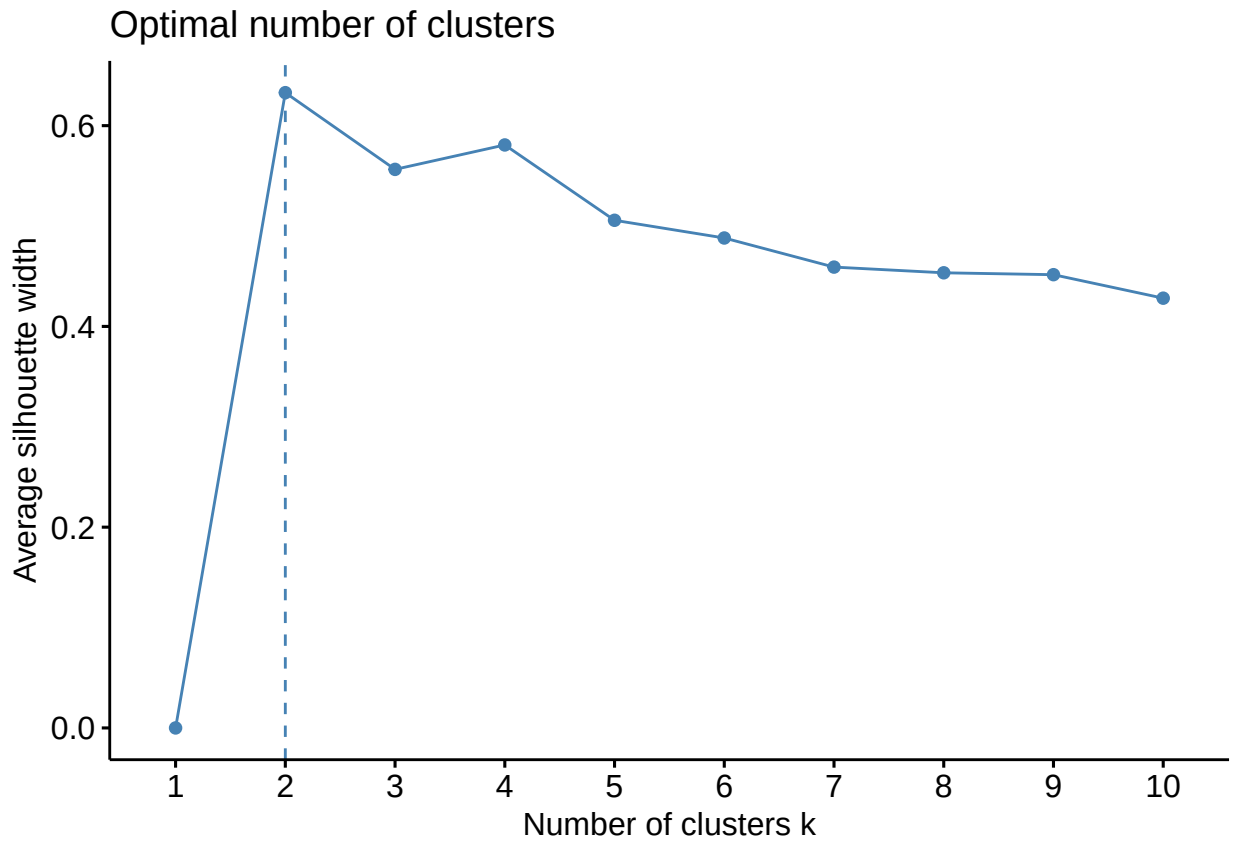
Approach:

The question I will be answering for this portion is: Is there a correlation between California and Texas colleges salary potential?

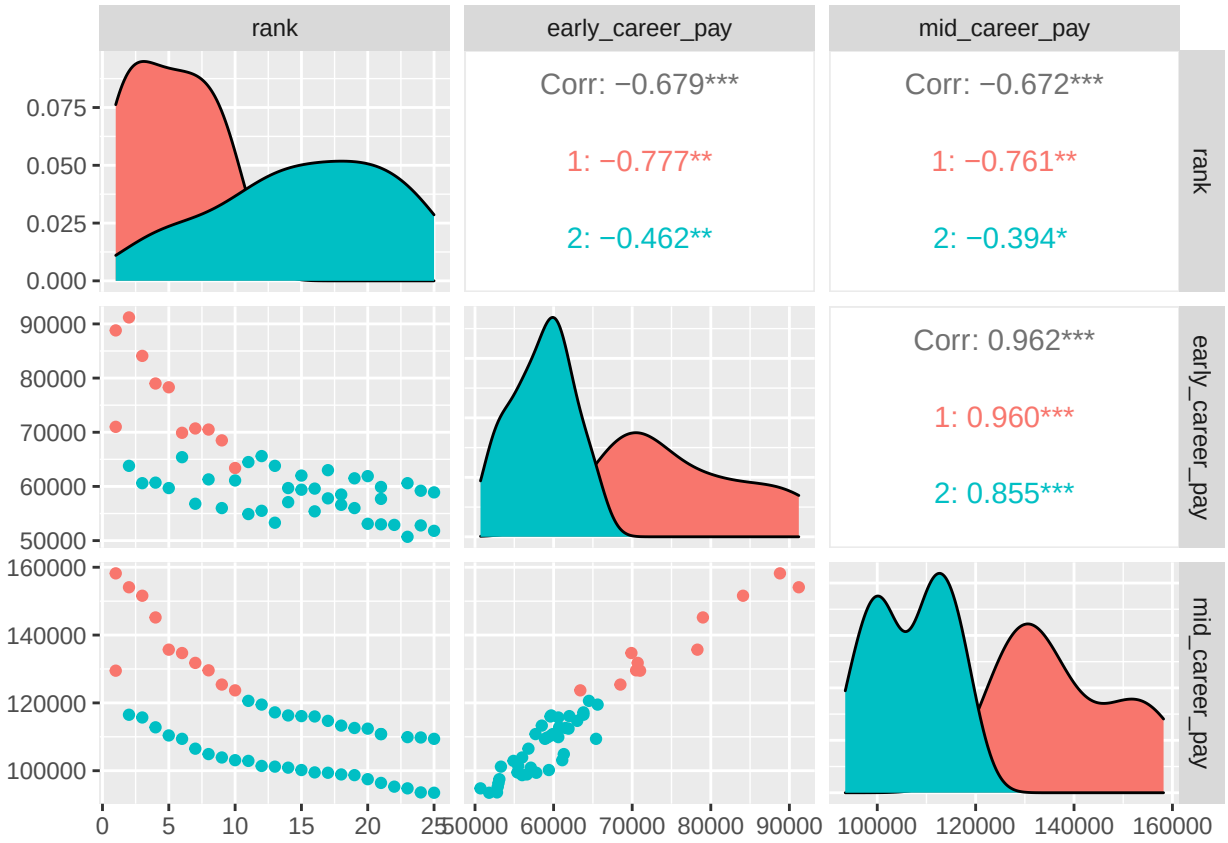
This question will be answered by creating a correlation matrix and using PAM clustering to view any trends. The PAM clustered will be done using the `silhouette` method where it will be visualized using `ggpairs` and `fviz_cluster`.

Analysis:

```
#Finding optimal clusters
opclust <- college %>%
  select(-name,-state_name,-make_world_better_percent,-stem_percent)
#fviz_cluster
fviz_nbclust(opclust, pam, method = "silhouette")
```




```
#k=2
pam_results <- opclust %>%
  pam(k=2)
#correlation matrix
college %>%
  mutate(
    cluster = as.factor(pam_results$clustering)
    ) %>%
  ggpairs(columns = c("rank","early_career_pay","mid_career_pay"), aes(color = cluster))
```




```
#finding avg of each variable
mean <- college %>%
 mutate(
   cluster = as.factor(pam_results$clustering)
      )%>%
#excluding unwanted variables
  select(-name,-state_name)%>%
  group_by(cluster) %>%
#finding average
  summarize_if(is.numeric, mean, na.rm = T)
```

```
#cluster plot
fviz_cluster(pam_results, data = college) +
  geom_point(aes(shape = college$state_name))+
  guides(shape = guide_legend(title = "shape"))
```

Discussion

There seems to be a strong positive correlation between early and mid career pay. This observations aligns well being that the salary increases the longer a person is in their respective field. Additionally, rank with the early and mid career pay reveals that the higher the rank a college is the higher the early and mid career pay is. Furthermore, the cluster plot reveals that California's colleges have a higher starting and mid career pay than Texas. This observation is correct as Californian's pay more for housing cost compared to Texas.

Dimensionality Reduction

Approach

The question I will be answering in this portion: Between California and Texas colleges, which one has a higher starting and mid career pay?

Although this question can be answered easily by simply taking the average, a PCA analysis will be performed where an arrow format is used to reveal the distribution and a plot by variance is used to reveal how much of the variable contributes to the overall plot. Essentially, principal components will be analyzed.

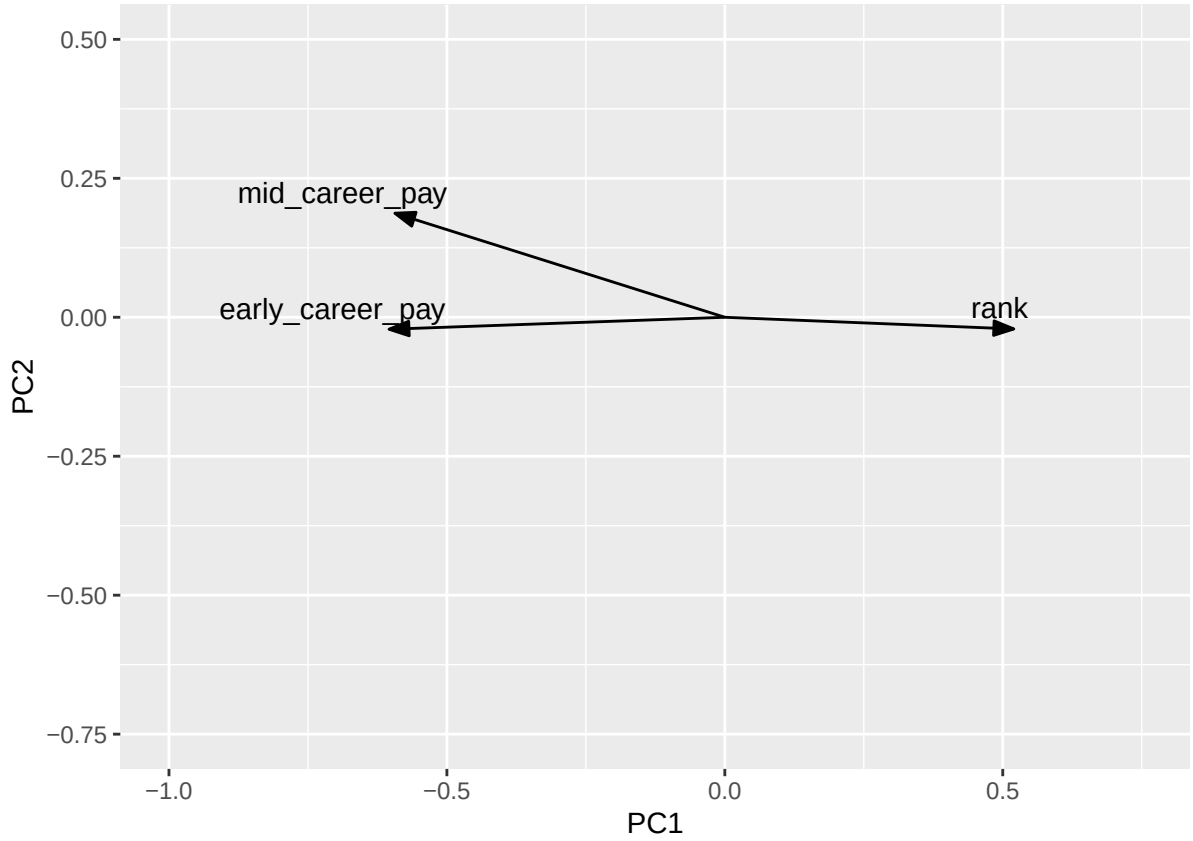
Analysis

```
# additional tidying
sum_college <- college %>%
  select(state_name,rank,early_career_pay,mid_career_pay,make_world_better_percent)%>%
  na.omit()

#pca calc
pca_fit_college <- sum_college %>%
```

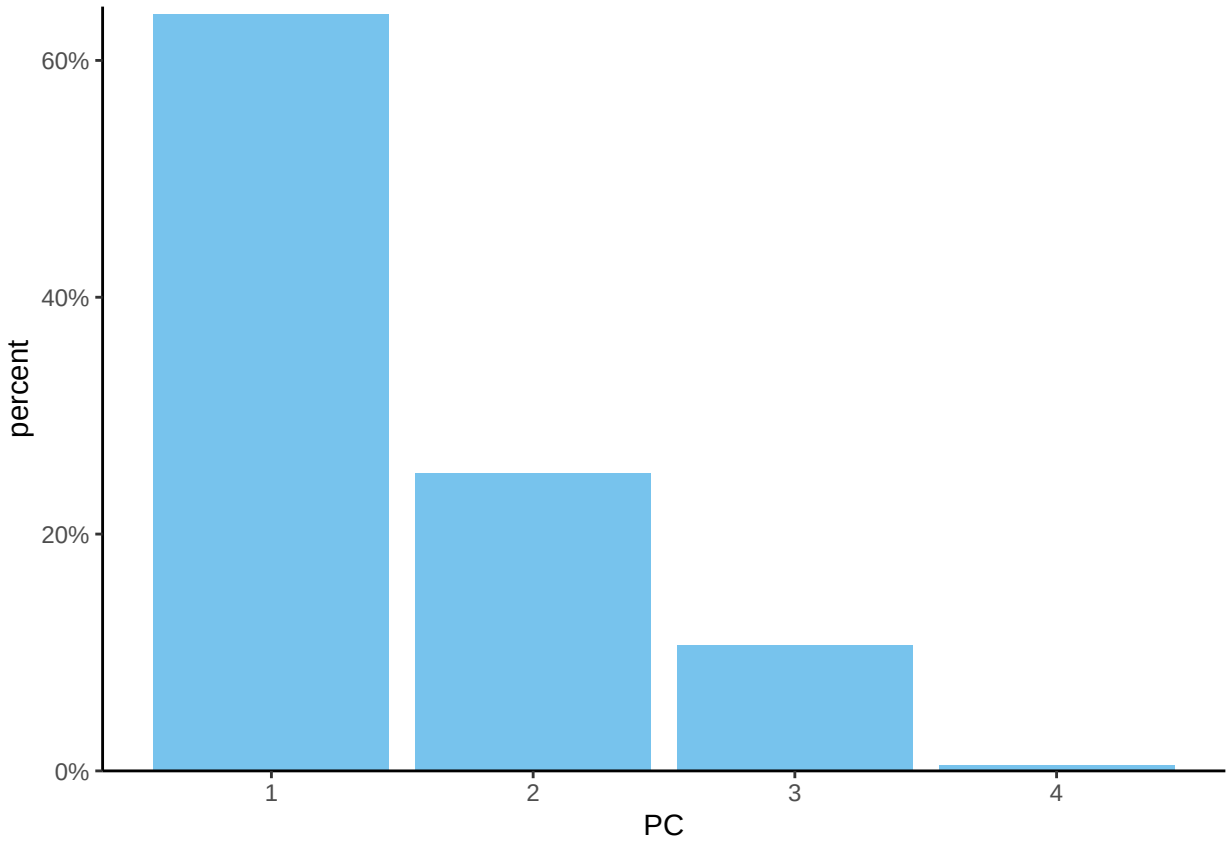
```
  select(where(is.numeric)) %>%
  scale()%>%
  prcomp()

#arrow
  arrow_style <- arrow(
  angle = 20, length = grid::unit(8, "pt"),
  ends = "first", type = "closed"
)
#to wide format
pca_fit_college %>%
  #extract a rotation matrix
  tidy(matrix = "rotation") %>%
  pivot_wider(
    names_from = "PC", values_from = "value",
    names_prefix = "PC"
  ) %>%

#rotation-plot
ggplot(aes(PC1, PC2)) +
geom_segment(
    xend = 0, yend = 0,
    arrow = arrow_style) +
geom_text(aes(label = column),
            hjust = 0.75, vjust = -0.5) +
xlim(-1.0, 0.75) + ylim(-0.75, 0.5) +
coord_fixed()
```

```
## Warning: Removed 1 rows containing missing values (geom_segment).
```

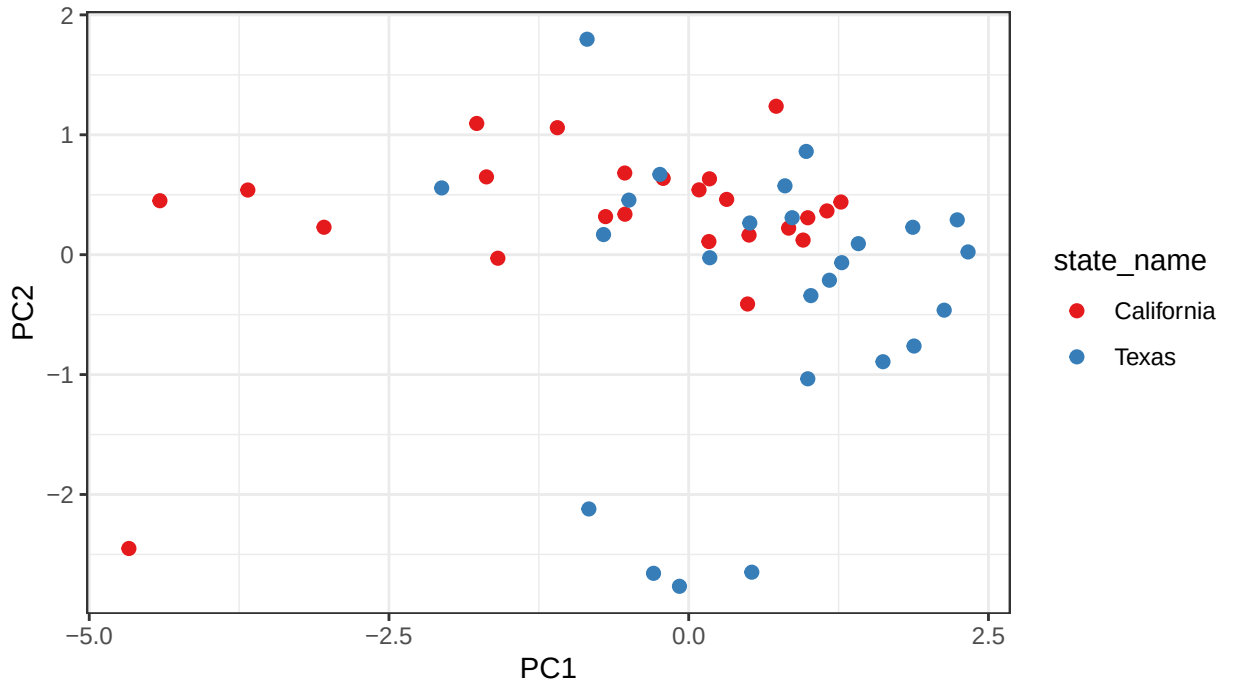
```
## Warning: Removed 1 rows containing missing values (geom_text).
```




```
# plot variance explained plot
pca_fit_college %>%
  tidy(matrix = "eigenvalues") %>%
  ggplot(aes(PC, percent)) +
  geom_col(fill = "#56B4E9", alpha = 0.8) +
  scale_x_continuous(breaks = 1:9) +
  scale_y_continuous(
    labels = scales::percent_format(),
    expand = expansion(mult = c(0, 0.01))
  ) +
  theme_classic()
```




```
pca_fit_college %>%
  augment(sum_college) %>%
  ggplot(aes(.fittedPC1, .fittedPC2)) +
  geom_point(aes(color = state_name),
             size = 2) +
  coord_fixed()+
  labs(x= "PC1", y= "PC2")+
  theme_bw()+
  scale_color_brewer(palette = "Set1")
```



Discussion

After performing PCA and plotting the arrow and scatter plot. It can be concluded values that have a positive PC1 score have a high college rank and positive PC2 value have a higher starting and mid career pay. Additionally, the first component explains about 65 % of the graph, the second explains about 25 %, the third explains about 10% and the fourth explains about 5%. Furthermore, it can be concluded that Californian colleges have a higher starting and mid career pay which supports the argument that it is expensive to live in California due to housing cost. Because it is expensive to live in California the salary would be higher.

Classification and Cross-validation

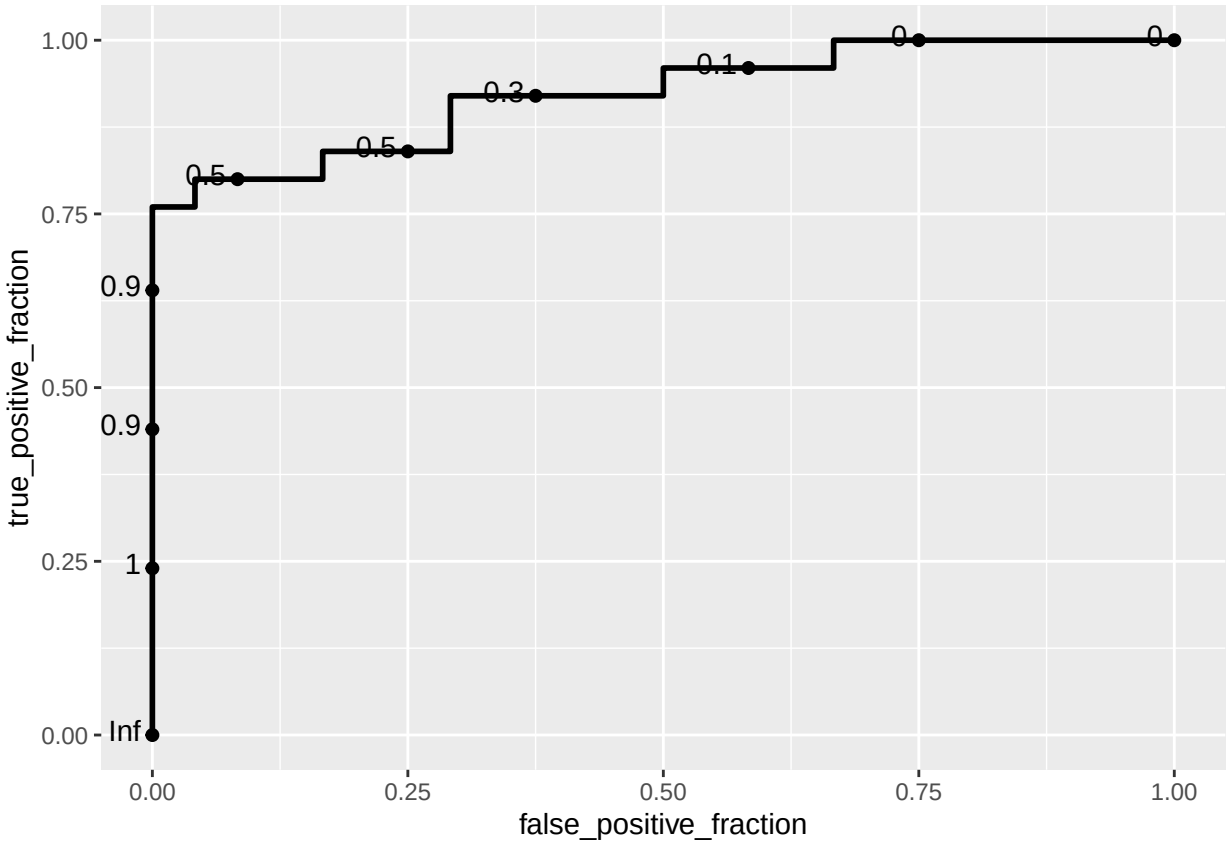
Approach

In this section the purpose is to do a logistic regression model to predict `state_name` status from `early_career_pay` and `mid_career_pay`. We find the probabilities of being what `state_name` and add them to the data frame called `log_college`. We would then use the `log_college` to build an ROC curve to compute the AUC. Furthermore, a 10-fold cross-validation will be performed on the model where the average perfomance is measured and to determine any signs of overfitting.

Analysis

```
sum_college <- sum_college %>%
#overwriting rank as a factor
  mutate(
    rank = as.factor(rank)
  ) %>%
#changing cali and texas to number values
```

```
  mutate(
    state_name = case_when(
      state_name == "California" ~ 0,
      state_name == "Texas"      ~ 1
    )
  ) %>%
#changing state_name to numeric
  mutate(
    state_name = as.numeric(state_name)
  )

#logistic
logistic_fit <- glm(state_name ~ early_career_pay + mid_career_pay,
                    data = sum_college, family = "binomial")

# Calculate a predicted probability
log_college <- sum_college %>%
  mutate(probability = predict(logistic_fit, type = "response"))

#roc with log_pokemon
ROC <- ggplot(log_college,
              aes(d = state_name, m = probability))+
  geom_roc()
ROC
```



```r
calc_auc(ROC)
```

```
##   PANEL group       AUC
## 1     1    -1 0.9216667
```

```r
# Set a seed to get reproducible results
set.seed(322)
# Choosing number of folds
k = 10

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

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

# Using a for loop to get diagnostics for each test set
diags_k <- NULL

for(i in 1:k){
  # Create training and test sets
  train <- data[folds != i, ] # all observations except in fold i
  test <- data[folds == i, ]  # observations in fold i

  # Train model on training set (all but fold i)
  fit <- glm(state_name ~ early_career_pay + mid_career_pay,
                    data = train, family = "binomial")

  # Test model on test set (fold i)
  df <- data.frame(
    prob = predict(fit, newdata = test, type = "response"),
    y = test$state_name)

  # Consider the ROC curve for the test dataset
  ROCplot <- ggplot(df) +
    geom_roc(aes(d = y, m = prob), n.cuts = 0)


  # Get diagnostics for fold i (AUC)
  diags_k[i] <- calc_auc(ROCplot)$AUC
}
```

```
## Warning in verify_d(data$d): D not labeled 0/1, assuming 1 = 0 and NA = 1!
```

```r
# Average performance
mean(diags_k, na.rm = TRUE)
```

```
## [1] 0.9722222
```

```r
#calculation ROCplot
calc_auc(ROCplot)
```

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

Discussion

After performing a logistic regression the AUC score was good, a value of 0.92, but after performing a k-fold cross-validation the AUC score improved to 1.0, the best possible value to obtain. This means the model is good. No signs of overfitting are seen since the ROC has little to no gaps. This means that the classifier predicts new observations 100 % correct beacuse it has an AUC value of 1.