Admission Rates of Colleges and Universities - Prediction Models

Md Abrar Nasir, Sumaita Imam Anika

March 20, 2022

```
library("tidyverse")
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library("patchwork")
library("car")
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
df <- read.csv("VPdataset.csv")
```

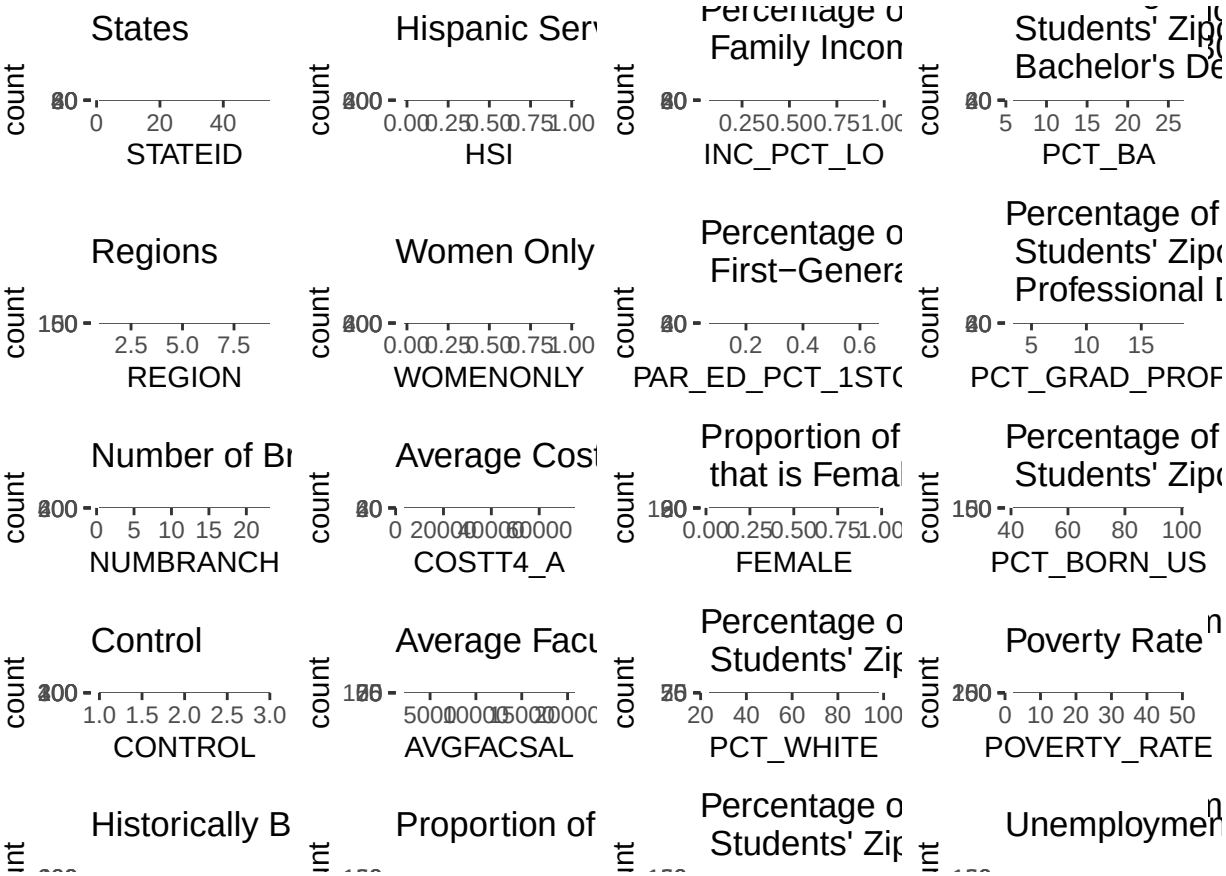
```
df <- transform(df, STATEID = as.numeric(factor(STABBR)))
df <- df[, !(names(df) %in% c("UNITID", "INSTNM", "STABBR"))]
df <- df %>% relocate(STATEID)
df <- df %>% relocate(ADM_RATE)
```

```
hist1 <- ggplot(df, aes(x = ADM_RATE)) + geom_histogram() + ggtitle("Admission Rate")
hist2 <- ggplot(df, aes(x = STATEID)) + geom_histogram() + ggtitle("States")
hist3 <- ggplot(df, aes(x = REGION)) + geom_histogram() + ggtitle("Regions")
hist4 <- ggplot(df, aes(x = NUMBRANCH)) + geom_histogram() + ggtitle("Number of Branches")
hist5 <- ggplot(df, aes(x = CONTROL)) + geom_histogram() + ggtitle("Control")
hist6 <- ggplot(df, aes(x = HBCU)) + geom_histogram() + ggtitle("Historically Black")
hist7 <- ggplot(df, aes(x = PBI)) + geom_histogram() + ggtitle("Predominantly Black")
hist8 <- ggplot(df, aes(x = TRIBAL)) + geom_histogram() + ggtitle("Tribal")
hist9 <- ggplot(df, aes(x = HSI)) + geom_histogram() + ggtitle("Hispanic Serving")
hist10 <- ggplot(df, aes(x = WOMENONLY)) + geom_histogram() + ggtitle("Women Only")
hist11 <- ggplot(df, aes(x = COSTT4_A)) + geom_histogram() + ggtitle("Average Cost of Attendance")
```

```r
hist12 <- ggplot(df, aes(x = AVGFACSAL)) + geom_histogram() + ggtitle("Average Faculty Salary")
hist13 <- ggplot(df, aes(x = PFTFAC)) + geom_histogram() + ggtitle("Proportion of Full-Time Faculty Meml
hist14 <- ggplot(df, aes(x = PCTPELL)) + geom_histogram() + ggtitle("Percentage of Undergraduates \n Rec
hist15 <- ggplot(df, aes(x = UG25ABV)) + geom_histogram() + ggtitle("Percentage of Undergraduates \n Age
hist16 <- ggplot(df, aes(x = INC_PCT_LO)) + geom_histogram() + ggtitle("Percentage of Aided Students who
hist17 <- ggplot(df, aes(x = PAR_ED_PCT_1STGEN)) + geom_histogram() + ggtitle("Percentage of \n First-Ge
hist18 <- ggplot(df, aes(x = FEMALE)) + geom_histogram() + ggtitle("Proportion of Student Body \n that i
hist19 <- ggplot(df, aes(x = PCT_WHITE)) + geom_histogram() + ggtitle("Percentage of the Population from
hist20 <- ggplot(df, aes(x = PCT_BLACK)) + geom_histogram() + ggtitle("Percentage of the Population from
hist21 <- ggplot(df, aes(x = PCT_ASIAN)) + geom_histogram() + ggtitle("Percentage of the Population from
hist22 <- ggplot(df, aes(x = PCT_HISPANIC)) + geom_histogram() + ggtitle("Percentage of the Population f
hist23 <- ggplot(df, aes(x = PCT_BA)) + geom_histogram() + ggtitle("Percentage of the Population from \n
hist24 <- ggplot(df, aes(x = PCT_GRAD_PROF)) + geom_histogram() + ggtitle("Percentage of the Population
hist25 <- ggplot(df, aes(x = PCT_BORN_US)) + geom_histogram() + ggtitle("Percentage of the Population fr
hist26 <- ggplot(df, aes(x = POVERTY_RATE)) + geom_histogram() + ggtitle("Poverty Rate")
hist27 <- ggplot(df, aes(x = UNEMP_RATE)) + geom_histogram() + ggtitle("Unemployment Rate")
hist28 <- ggplot(df, aes(x = MD_FAMINC)) + geom_histogram() + ggtitle("Median Family Income of Students

hist1 + hist2 + hist3 + hist4 + hist5 + hist6 + hist7 + hist8 + hist9 + hist10 + hist11 + hist12 + hist1
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
full <- lm(ADM_RATE ~., data = df)
summary(full)
```

```
##
## Call:
## lm(formula = ADM_RATE ~ ., data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.56521 -0.11176  0.00225  0.12449  0.41536
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.763e-01  3.272e-01   1.150 0.250475
## STATEID          7.233e-04  4.520e-04   1.600 0.109931
## NUMBRANCH        8.252e-03  2.140e-03   3.855 0.000126 ***
## CONTROL         -2.582e-02  2.107e-02  -1.225 0.220918
## REGION           1.128e-03  4.503e-03   0.250 0.802281
## HBCU            -3.156e-02  4.801e-02  -0.657 0.511118
## PBI              6.739e-03  5.456e-02   0.124 0.901724
## TRIBAL          -1.303e-02  1.922e-01  -0.068 0.945941
## HSI              6.038e-02  3.021e-02   1.998 0.046043 *
## WOMENONLY        6.517e-02  7.899e-02   0.825 0.409630
## COSTT4_A        -2.253e-06  8.620e-07  -2.613 0.009156 **
## AVGFACSAL       -2.958e-05  4.309e-06  -6.865 1.43e-11 ***
## PFTFAC          -1.956e-02  2.798e-02  -0.699 0.484570
```

```
## PCTPELL           -1.031e-01  8.163e-02  -1.264 0.206761
## UG25ABV            1.620e-02  6.325e-02   0.256 0.797876
## INC_PCT_LO        -5.263e-02  1.610e-01  -0.327 0.743903
## PAR_ED_PCT_1STGEN  2.731e-01  1.362e-01   2.005 0.045345 *
## FEMALE             1.241e-01  5.632e-02   2.203 0.027900 *
## MD_FAMINC          1.227e-06  8.503e-07   1.443 0.149529
## PCT_WHITE          3.272e-03  2.456e-03   1.332 0.183183
## PCT_BLACK          9.672e-05  2.584e-03   0.037 0.970154
## PCT_ASIAN          4.828e-03  4.642e-03   1.040 0.298625
## PCT_HISPANIC      -2.122e-03  1.485e-03  -1.429 0.153439
## PCT_BA             1.034e-02  4.538e-03   2.278 0.022993 *
## PCT_GRAD_PROF     -1.186e-02  6.061e-03  -1.957 0.050677 .
## PCT_BORN_US        5.229e-05  1.773e-03   0.029 0.976484
## POVERTY_RATE      -6.009e-03  4.050e-03  -1.484 0.138332
## UNEMP_RATE         5.674e-02  1.708e-02   3.322 0.000940 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1673 on 722 degrees of freedom
## Multiple R-squared:  0.2835, Adjusted R-squared:  0.2567
## F-statistic: 10.58 on 27 and 722 DF,  p-value: < 2.2e-16
```