Data_607_Project_2_Seung_Min_Song_Untidy_Data_Admit

Enid Roman

2022-10-09

ABOUT THE DATASET:

This dataset was created by Seung Min Song which information was taken from the following website:

https://www.randomservices.org/random/data/Berkeley.html

The dataset represents admissions data at the University of California, Berkeley in 1973 according to the variables department (A, B, C, D, E), gender (male, female), and outcome admitted or denied.

```
# Upload the libraries

library(tidyr)
library(tidyverse)
```

Were there gender bias during the application process?

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v dplyr   1.0.9
## v tibble  3.1.8     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(ggplot2)
```

```
# Import the data from github.
# Link is provided to the csv file below:
# https://github.com/enidroman/data_607_data_aquisition_and_management_project/blob/main/Seung%20Min%20

urlfile <- "https://raw.githubusercontent.com/enidroman/data_607_data_aquisition_and_management_project/

admit_reject <- read.csv(urlfile)
admit_reject
```

```
##     Gender Dept Admitted Rejected
## 1     Male     A      512      313
## 2   Female     A       89       19
## 3     Male     B      353      207
## 4   Female     B       17        8
## 5     Male     C      120      205
## 6   Female     C      202      391
## 7     Male     D      138      279
## 8   Female     D      131      244
## 9     Male     E       53      138
## 10  Female     E       94      299
## 11    Male     F       22      351
## 12  Female     F       24      317
```

DATA CLEANING AND TRANSFORMATION

In observing the dataset I see that:

1. Admited and Rejected should not have their seperate columns. There should be only one column for both Admitted and Rejected. A new column should be created for Admit and Reject and be named Outcome. That column should be the first column.

2, A new column must be created for Numbers of Applicants for the numbers of Admitted and Rejected, which should go at the end after Department. The numbers of applicants should be aligned with the Admiitted and Rejected and Male and Female and the Deptartment.

```
# Transform the dataframe into a long format to have Admitted and Rejected in one column and have Numbe

#outcome <- admit_reject %>%
 # pivot_longer (c(`Admitted`, `Rejected`), names_to = "Outcome", values_to = "Numbers_of_Applicants")
#outcome

outcome <- admit_reject %>%
  gather(key = "Outcome", value = "Number_of_Applicants", Admitted:Rejected) %>%
   arrange(desc(Gender)) %>%
   arrange(Dept)

outcome
```

3. Columns should be rearranged, Outcome, Gender, Dept, and Number_of_Applicants.

```
##     Gender Dept  Outcome Number_of_Applicants
## 1     Male     A Admitted                  512
## 2     Male     A Rejected                  313
## 3   Female     A Admitted                   89
## 4   Female     A Rejected                   19
## 5     Male     B Admitted                  353
## 6     Male     B Rejected                  207
## 7   Female     B Admitted                   17
```

```
## 8  Female     B Rejected                  8
## 9    Male     C Admitted                120
## 10   Male     C Rejected                205
## 11 Female     C Admitted                202
## 12 Female     C Rejected                391
## 13   Male     D Admitted                138
## 14   Male     D Rejected                279
## 15 Female     D Admitted                131
## 16 Female     D Rejected                244
## 17   Male     E Admitted                 53
## 18   Male     E Rejected                138
## 19 Female     E Admitted                 94
## 20 Female     E Rejected                299
## 21   Male     F Admitted                 22
## 22   Male     F Rejected                351
## 23 Female     F Admitted                 24
## 24 Female     F Rejected                317
```

```
# Rearranged the columns.

outcome <- outcome[, c("Outcome", "Gender", "Dept", "Number_of_Applicants")]
outcome
```

```
##     Outcome  Gender Dept Number_of_Applicants
## 1  Admitted   Male     A                  512
## 2  Rejected   Male     A                  313
## 3  Admitted Female     A                   89
## 4  Rejected Female     A                   19
## 5  Admitted   Male     B                  353
## 6  Rejected   Male     B                  207
## 7  Admitted Female     B                   17
## 8  Rejected Female     B                    8
## 9  Admitted   Male     C                  120
## 10 Rejected   Male     C                  205
## 11 Admitted Female     C                  202
## 12 Rejected Female     C                  391
## 13 Admitted   Male     D                  138
## 14 Rejected   Male     D                  279
## 15 Admitted Female     D                  131
## 16 Rejected Female     D                  244
## 17 Admitted   Male     E                   53
## 18 Rejected   Male     E                  138
## 19 Admitted Female     E                   94
## 20 Rejected Female     E                  299
## 21 Admitted   Male     F                   22
## 22 Rejected   Male     F                  351
## 23 Admitted Female     F                   24
## 24 Rejected Female     F                  317
```

ANALYSIS

No analysis was requested on the discussion but I created my own analysis.

```
# Summary of each column.

summary(outcome)
```

I see the dataframe is 24 rows in length. Outcome, Gender, Dept is class as characters. The Number of Applicants Min is 8.0, 1st Quarter is 80., Median 170.0, Mean is 188.6, 3rd Quarter is 302.5, the Max is 512.0.

```
##    Outcome              Gender              Dept           Number_of_Applicants
##  Length:24          Length:24          Length:24          Min.   :  8.0
##  Class :character   Class :character   Class :character   1st Qu.: 80.0
##  Mode  :character   Mode  :character   Mode  :character   Median :170.0
##                                                           Mean   :188.6
##                                                           3rd Qu.:302.5
##                                                           Max.   :512.0
```

```
# Aggregate Function to compute summary statistic for subsets of the data, Average of Number of Applica

group_mean <- aggregate(x = outcome$Number_of_Applicants,
                    by = list(outcome$Outcome),
                    FUN = mean)

colnames(group_mean) <- c("Outcome", "Mean")

group_mean
```

There were more applicants that were rejected then Admitted.

```
##    Outcome     Mean
## 1 Admitted 146.2500
## 2 Rejected 230.9167
```

There were more applicants that were male then female.

```
# Aggregate Function to compute summary statistic for subsets of the data, Average of Number of Applica

group_mean <- aggregate(x = outcome$Number_of_Applicants,
                    by = list(outcome$Gender),
                    FUN = mean)

colnames(group_mean) <- c("Gender", "Mean")

group_mean
```

Please note: I don't know why Female came out twice in this dataframe. For some reason the female count comes out to 10. I did checked everything it seem fine.

```
##    Gender   Mean
## 1  Female 181.00
## 2 Female   12.50
## 3   Male  224.25
```

```
sum(outcome$Gender=='Female')
```

Please note: I don't know why the # number of count is wrong below.

```
## [1] 10
```

```
sum(outcome$Gender=='Male')
```

```
## [1] 0
```

```
# Aggregate Function to compute summary statistic for subsets of the data, Average of Number of Applica

group_mean <- aggregate(x = outcome$Number_of_Applicants,
                    by = list(outcome$Dept),
                    FUN = mean)
colnames(group_mean) <- c("Dept", "Mean")

group_mean
```

There were more applicants in Dept C and less in Dept E.

```
##   Dept   Mean
## 1    A 233.25
## 2    B 146.25
## 3    C 229.50
## 4    D 198.00
## 5    E 146.00
## 6    F 178.50
```

Please note again I have the extra set of Female Admit and Reject and I don't know why. '

There were 557 Female that were admitted and 1278 Female that were rejected.

```
# Aggregate function to aggregate the sum to summarize the data frame based on the two variables, Outco

list_aggregate <- aggregate(outcome$Number_of_Applicants, by = list(outcome$Outcome, outcome$Gender), FU

colnames(list_aggregate) <- c("Outcome", "Gender", "Number_of_Applicants")

list_aggregate
```

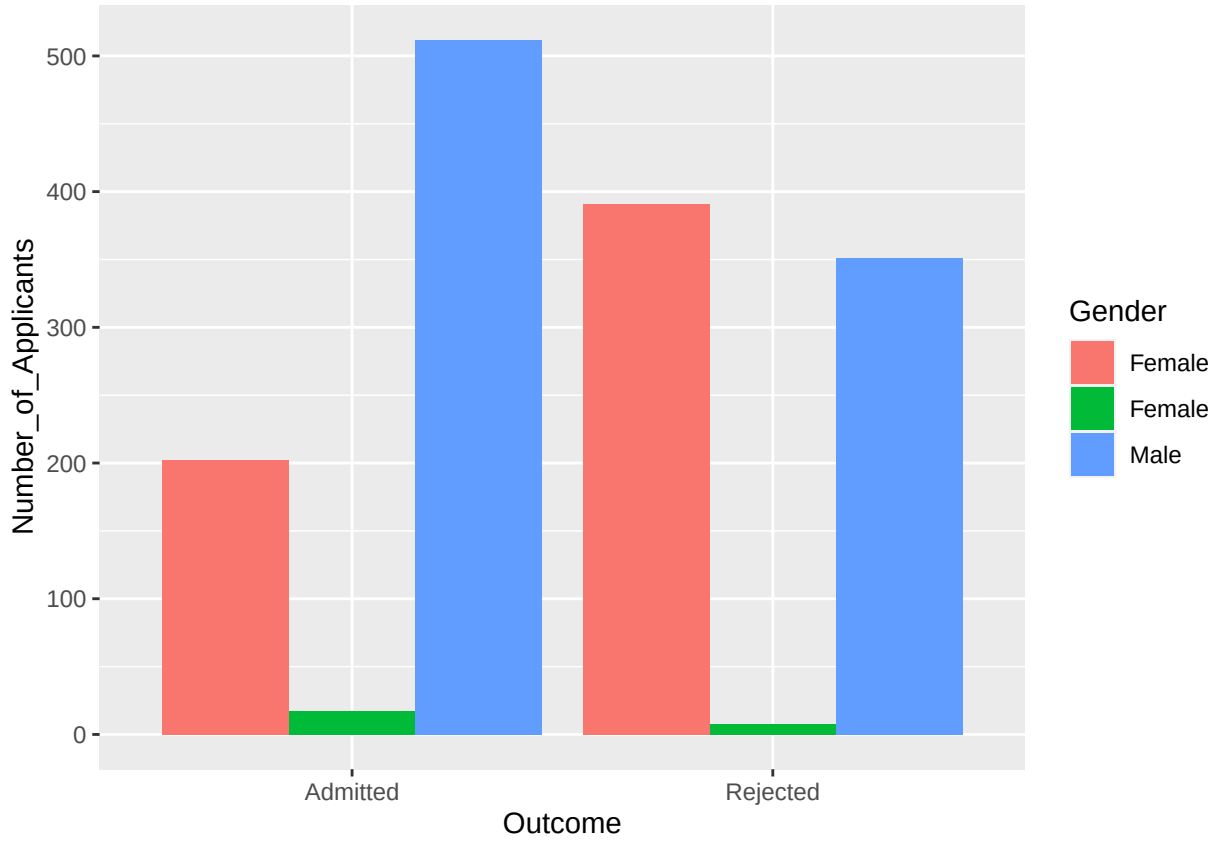
There were 1198 Male that were admitted and 1493 Male that were Rejected.

```
##    Outcome  Gender Number_of_Applicants
## 1 Admitted  Female                  540
## 2 Rejected  Female                 1270
## 3 Admitted Female                   17
## 4 Rejected Female                    8
## 5 Admitted   Male                 1198
## 6 Rejected   Male                 1493
```

```
# Bar gaph showing Net Value per Boro Block Lot by Neighborhood.

graph <- ggplot(outcome, aes(x = Outcome, y = Number_of_Applicants, fill = Gender)) +
  geom_col(position = "dodge")
graph
```

As per the graph below more male applicants were admitted vs female applicants.



CONCLUSION

In my analysis I observed that there were gender bias during the application process since more male were admitted then female. But as the University of California, Berkley states there were more male applicants then female that had applied. In regards to the women were

applying for admission in harder departments I have yet to see since there is no data the Departments that the applicants applied to. Only that they are listed as A, B, C, D, E, and F. Further investication has to be conducted to see if this application process was actually a gender bias.

As per the University of California Berkley An analysis of just the variables gender and admissions shows a correlation that suggests gender bias: the proportion of women admitted was significantly lower than the proportion of men admitted. However, when the department variable is taken into account, the gender bias disappears. Generally, the women were applying for admission in the harder departments, those with low admission rates.

A data set in which a correlation between two variables disapears, or even reverses, when a third variable is taken into account is known as Simpson's paradox.