Group 3: Exercise 3: Data Exploration

2023-01-31

Data Processing:

```
df1 = read.csv("D:/MMA Material/Term 3/Talent Analytics/preprocessed_file.csv")
df1$status <- ifelse(df1$latest_date > "2017-01-01", 1, 0)
df2 <- subset(df1, select = c(2,7,8,9,17,18,19,22,23))
df3 <- aggregate(application_number ~ ., data = df2, FUN = length)

date_counts <- table(df3$latest_date)
top_dates <- names(sort(date_counts, decreasing=TRUE))[1:5]
top_counts <- sort(date_counts, decreasing=TRUE)[1:20]
top_dates
```

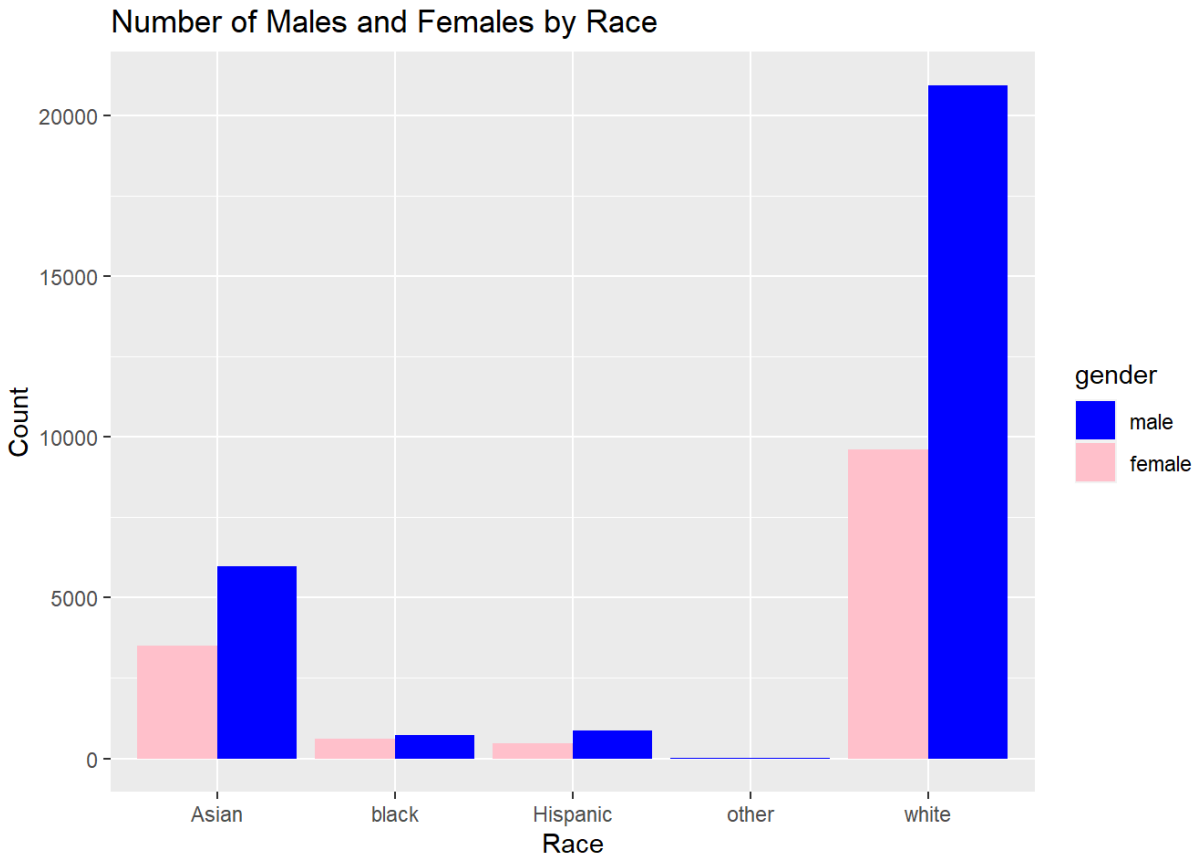
```
## NULL
```

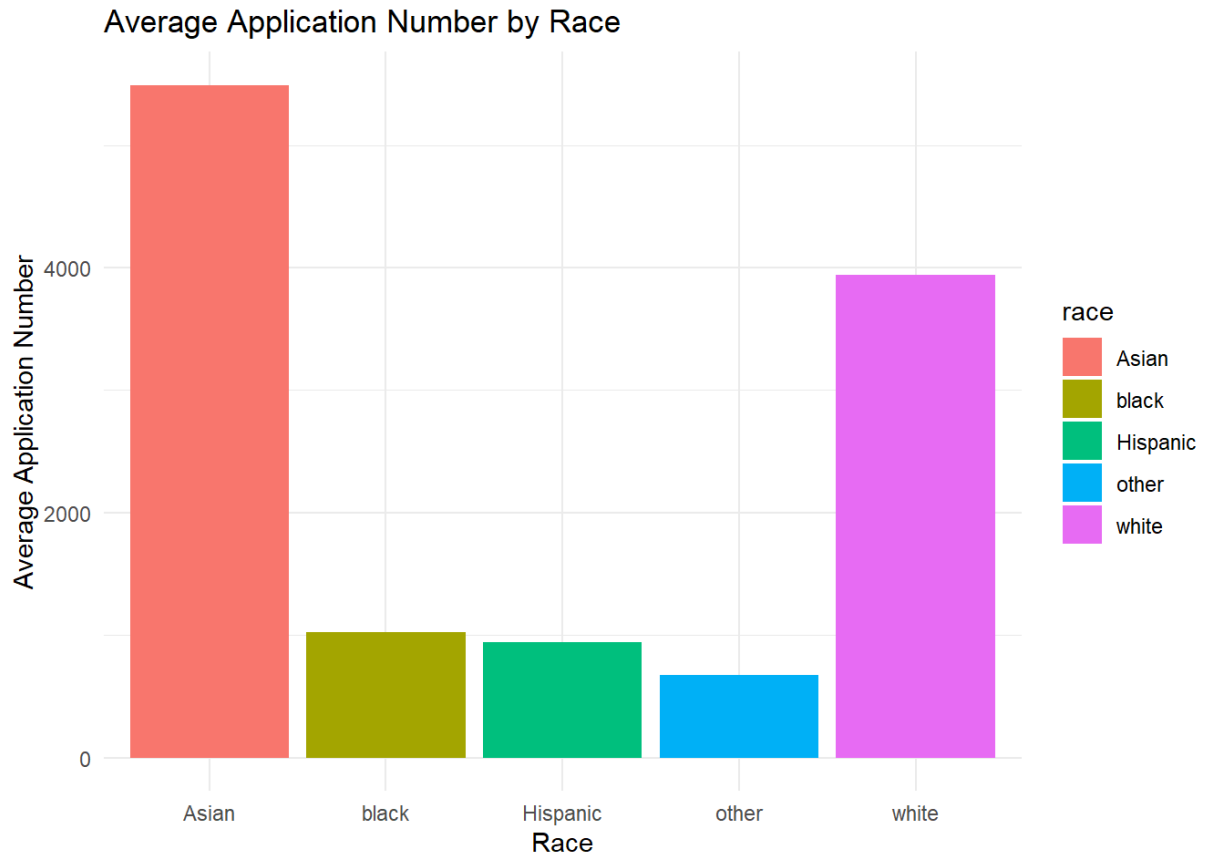
```
top_counts
```

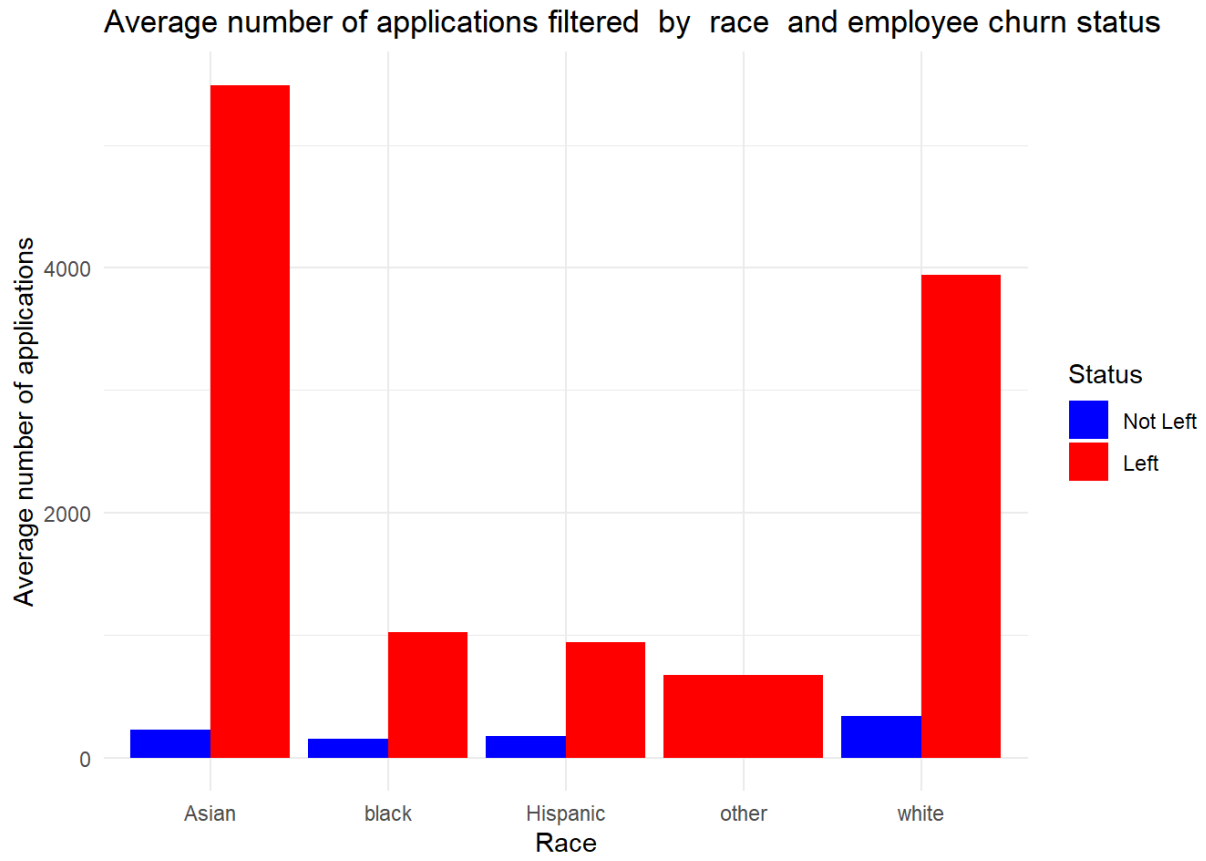
```
##  [1] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
```

```
df3$uspc_class = as.factor(df3$uspc_class)
df3$tc = as.factor(df3$tc)
df3$gender = as.factor(df3$gender)
df3$race = as.factor(df3$race)
df3$examiner_art_unit = as.factor(df3$examiner_art_unit)
df <- df3[ -c(1,2,3) ]
library(ggplot2)
```

Data Exploration:

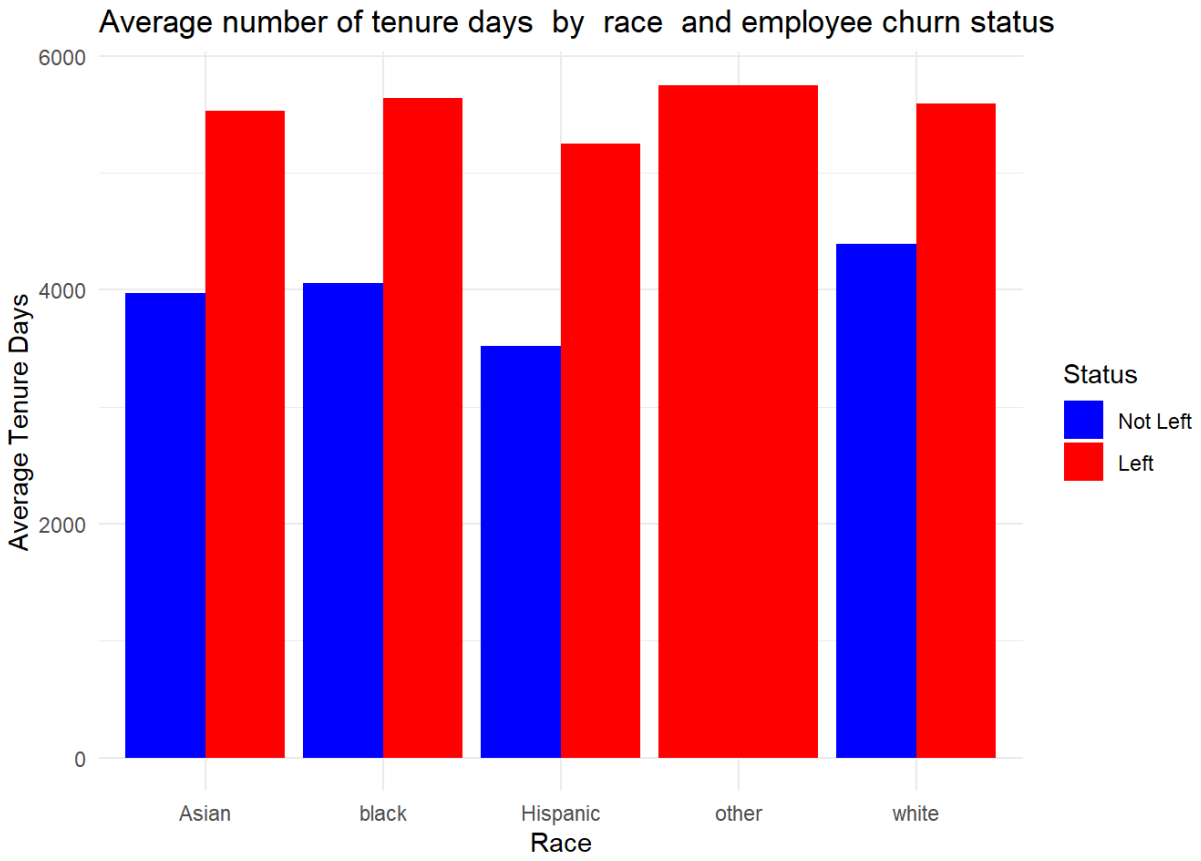



Average Application Number by Race
Average Application Number
4000
2000
0
Asian
black
Hispanic
other
white
Race
race
Asian
black
Hispanic
other
white



Average number of tenure days by race and employee churn status:

```
df_agg <- aggregate(tenure_days ~ race + status, data=df, mean)
```


Average number of tenure days by race and employee churn status
6000
4000
2000
0
Average Tenure Days
Asian
black
Hispanic
other
white
Race
Status
Not Left
Left