02_EDA

2023-09-16

```
tuesdata <- tidytuesdayR::tt_load('2022-11-01')
```

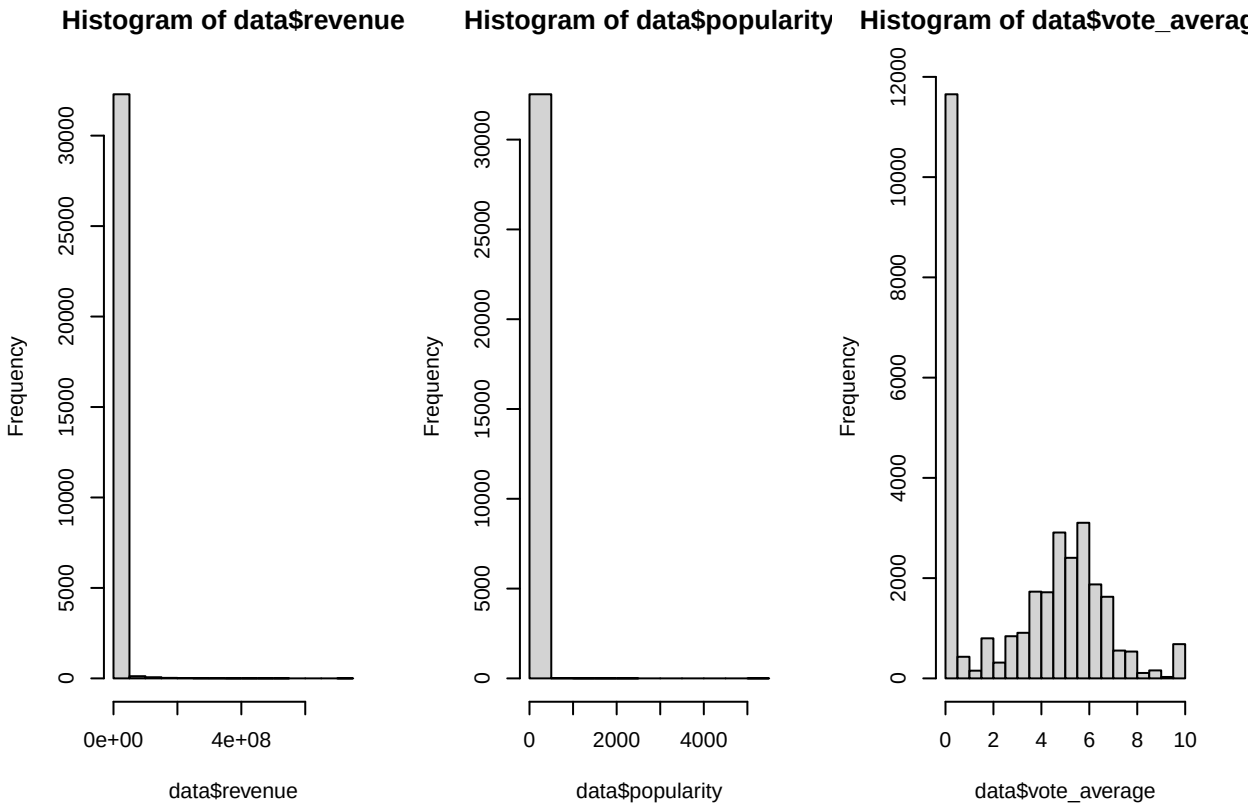
```
## --- Compiling #TidyTuesday Information for 2022-11-01 ----

## --- There is 1 file available ---

## --- Starting Download ---

##
##  Downloading file 1 of 1: ‘horror_movies.csv‘

## --- Download complete ---
```

```
data <- tuesdata$horror_movies
```

Section 1

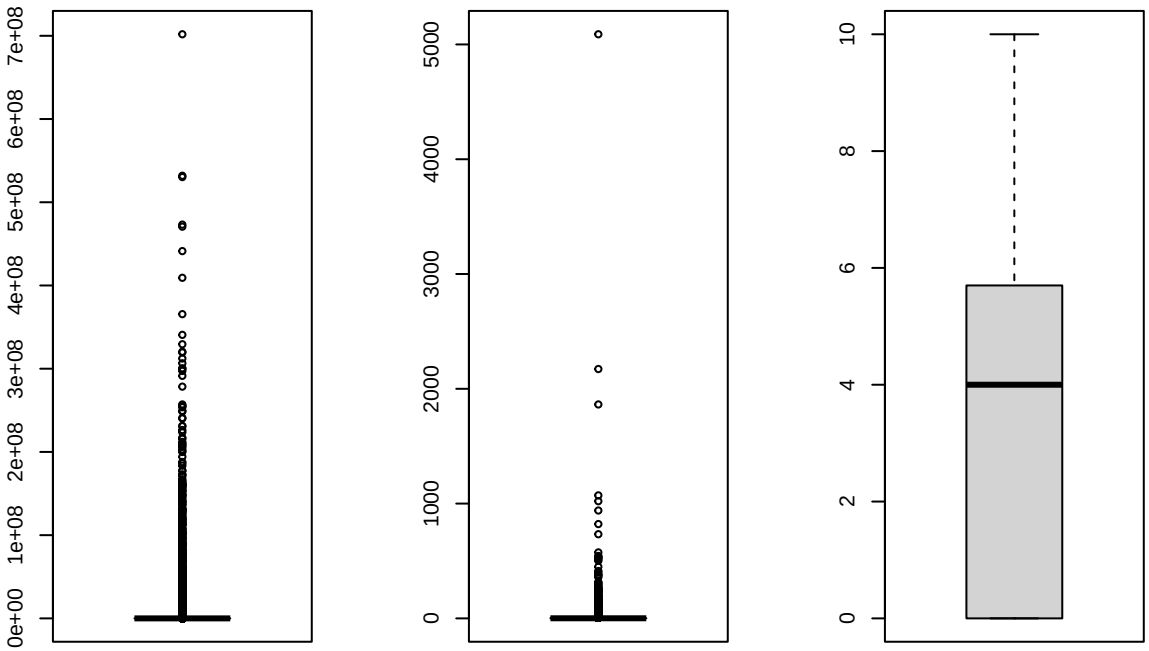
The response variables I'm interested in are revenue, popularity, and rating.

```
par(mfrow=c(1,3))
revhist <- hist(data$revenue)
pophist <- hist(data$popularity)
rathist <- hist(data$vote_average)
```




```
par(mfrow=c(1,1))
```

```
par(mfrow=c(1,3))
revboxplot <- boxplot(data$revenue)
popboxplot <- boxplot(data$popularity)
ratboxplot <- boxplot(data$vote_average)
```




```
par(mfrow=c(1,1))
```

Section 2

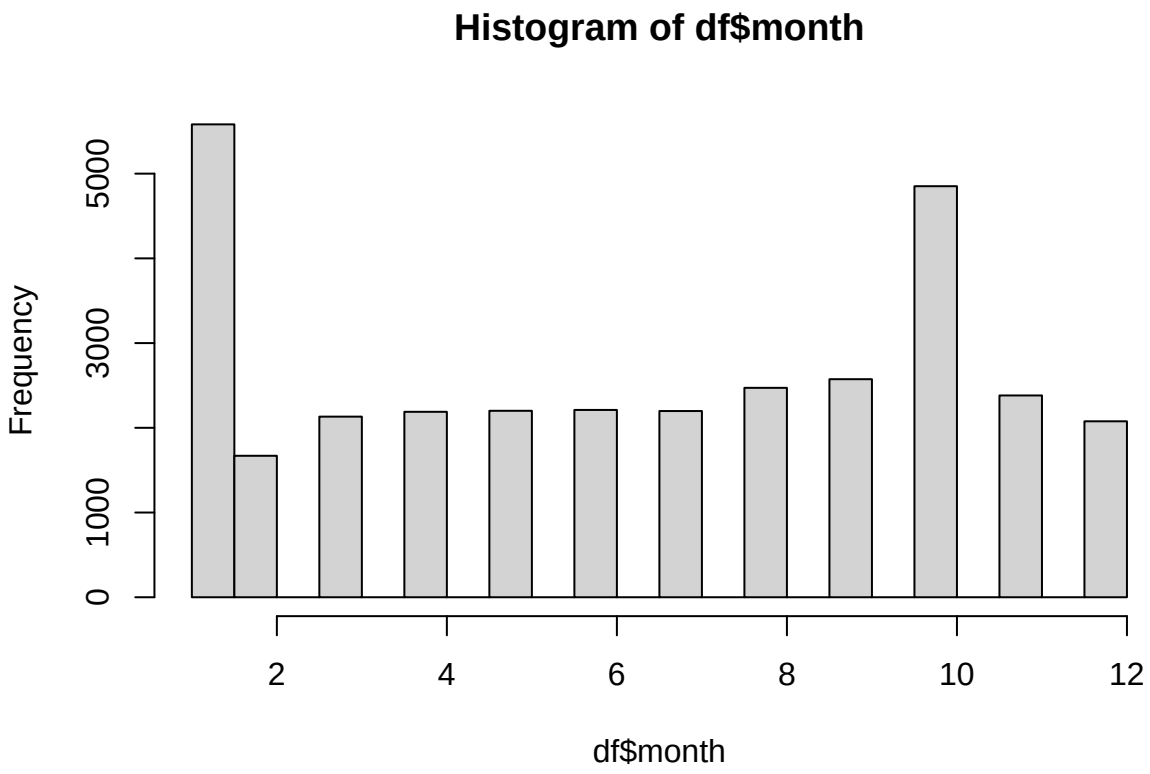
The explanatory variable I'm interested in is month of release.

```
#install.packages("lubridate")
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
df <- data.frame(release_date = as.Date(data$release_date))
df$monthname <- month(df$release_date, label = TRUE)
df$month <- month(df$release_date)
hist(df$month)
```



Section 3 After considering the variables further and previous documentation, I'm not sure what the popularity variable measures. Certainly, it should say that a higher number means the movie was more popular, but in what sense? Therefore, I will not be including it as on of my response variables in my killer graph.

```
#install.packages("dplyr")
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

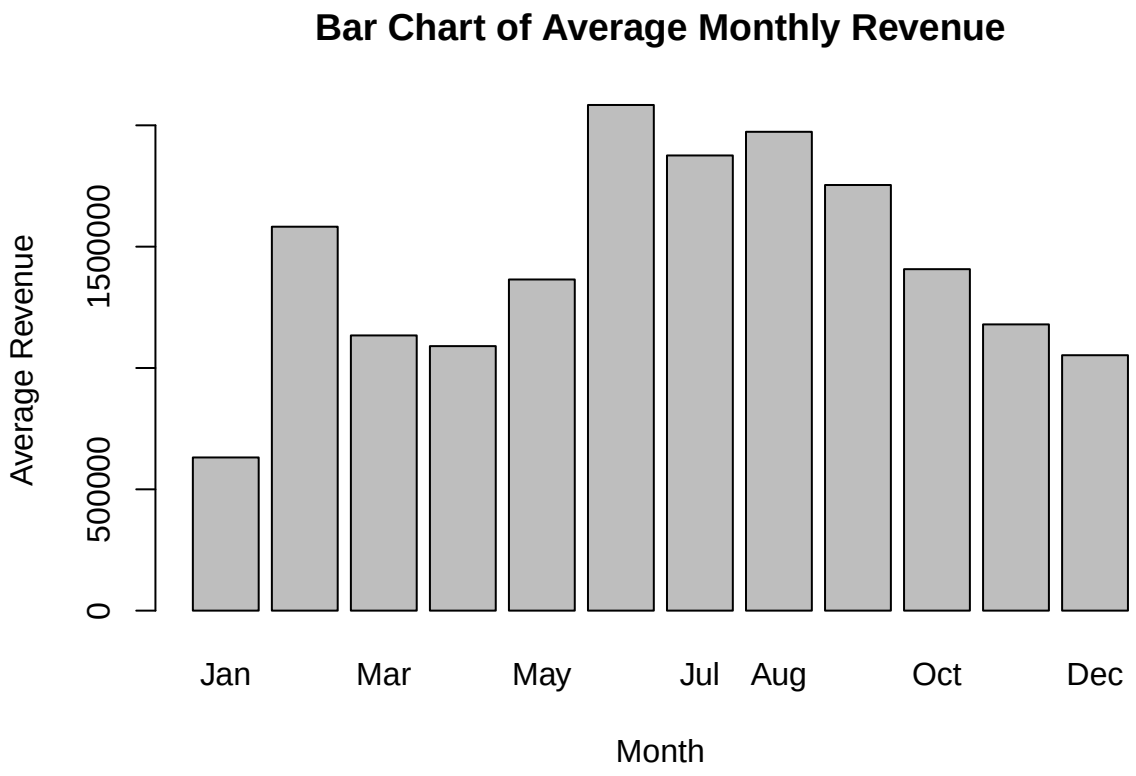
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

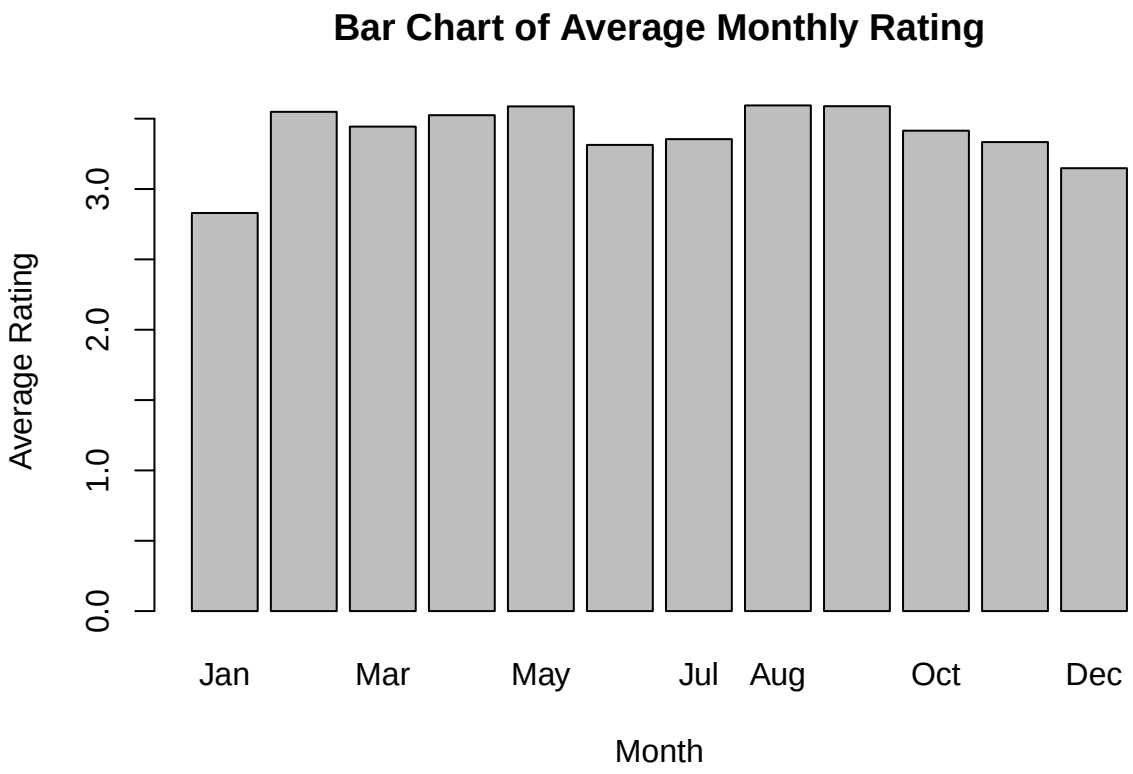
```
data$month <- df$monthname
result <- data %>%
  group_by(month) %>%
  summarise(mean_revenue = mean(revenue, na.rm = TRUE), mean_rating = mean(vote_average, na.rm = TRUE))
print(result)
```

```
## # A tibble: 12 x 3
##    month mean_revenue mean_rating
##    <ord>        <dbl>       <dbl>
##  1 Jan         631209.        2.83
##  2 Feb        1582155.        3.55
##  3 Mar        1134126.        3.44
##  4 Apr        1090063.        3.52
##  5 May        1364678.        3.59
##  6 Jun        2083872.        3.31
##  7 Jul        1875757.        3.35
##  8 Aug        1973259.        3.59
##  9 Sep        1754071.        3.59
## 10 Oct        1407257.        3.41
## 11 Nov        1179617.        3.33
## 12 Dec        1052624.        3.15
```

```
barplot(as.numeric(as.matrix(result)[,2]),names.arg = as.matrix(result)[,1],main = "Bar Chart of Averag
```




```
barplot(as.numeric(as.matrix(result)[,3]),names.arg = as.matrix(result)[,1],main = "Bar Chart of Averag
```



Section 4

```
save.image()
```