EDA

Juwon Park

2024-09-20

```r
# read in data
data<- read.csv("amazon_reviews.csv")
# remove unnecessary columns
data<-data[, -c(1, 10:12)]
data <- na.omit(data)
View(data)
```

```r
# summary of overall ratings
data$overall<- as.numeric(data$overall)
summary(data$overall)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   1.000   5.000   5.000   4.588   5.000   5.000
```

```r
# average of each rating
rating_counts <- data %>%
  group_by(overall) %>%
  summarize(count = n(), .groups = 'drop')
rating_avg_counts <- rating_counts %>%
  arrange(desc(overall))
rating_avg_counts <- rating_avg_counts %>%
  rename(Rating = overall, Average_Count = count)
rating_avg_counts
```

```
## # A tibble: 5 x 2
##   Rating Average_Count
##    <dbl>         <int>
## 1      5          3922
## 2      4           527
## 3      3           142
## 4      2            80
## 5      1           244
```

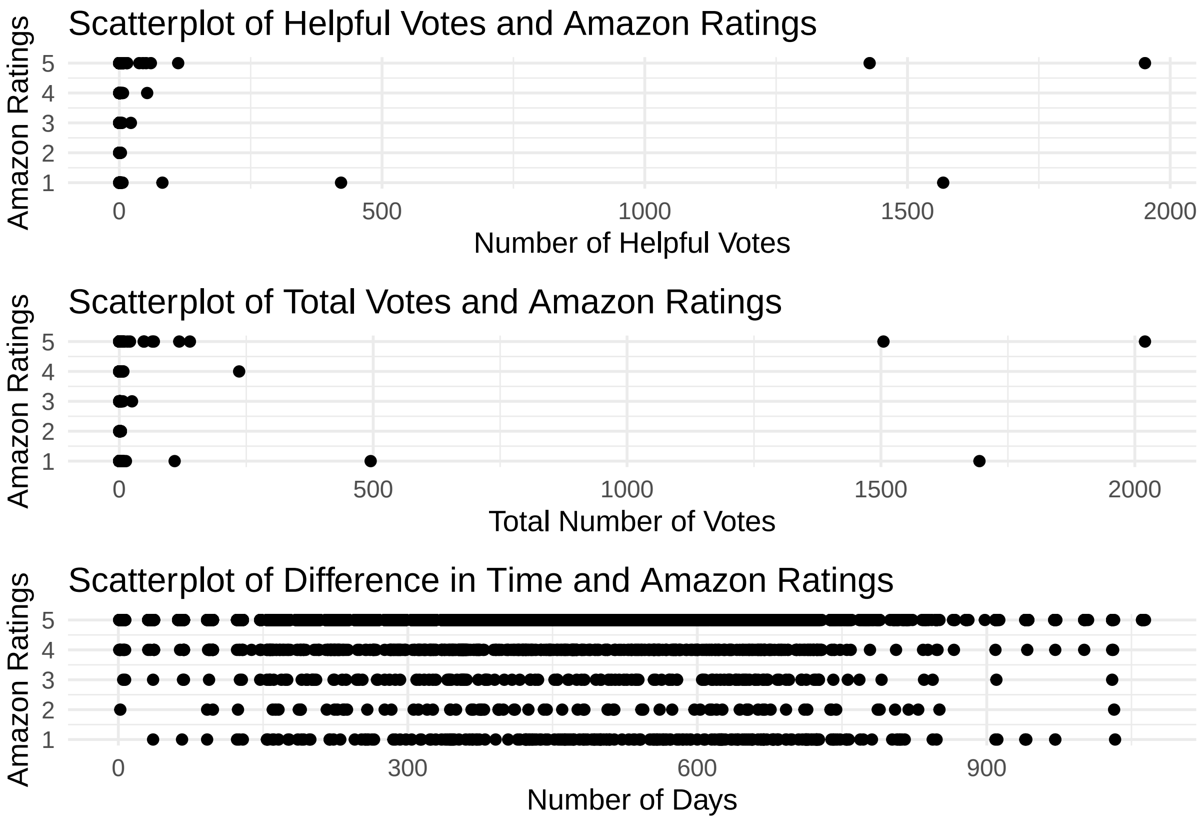
```r
# scatterplots of helpful ratings, total votes, day diff
data$day_diff <- as.numeric(data$day_diff)
data$helpful_yes<- as.numeric(data$helpful_yes)
data$helpful_no<- as.numeric(data$helpful_no)
data$total_vote<- as.numeric(data$total_vote)

library(patchwork)
```

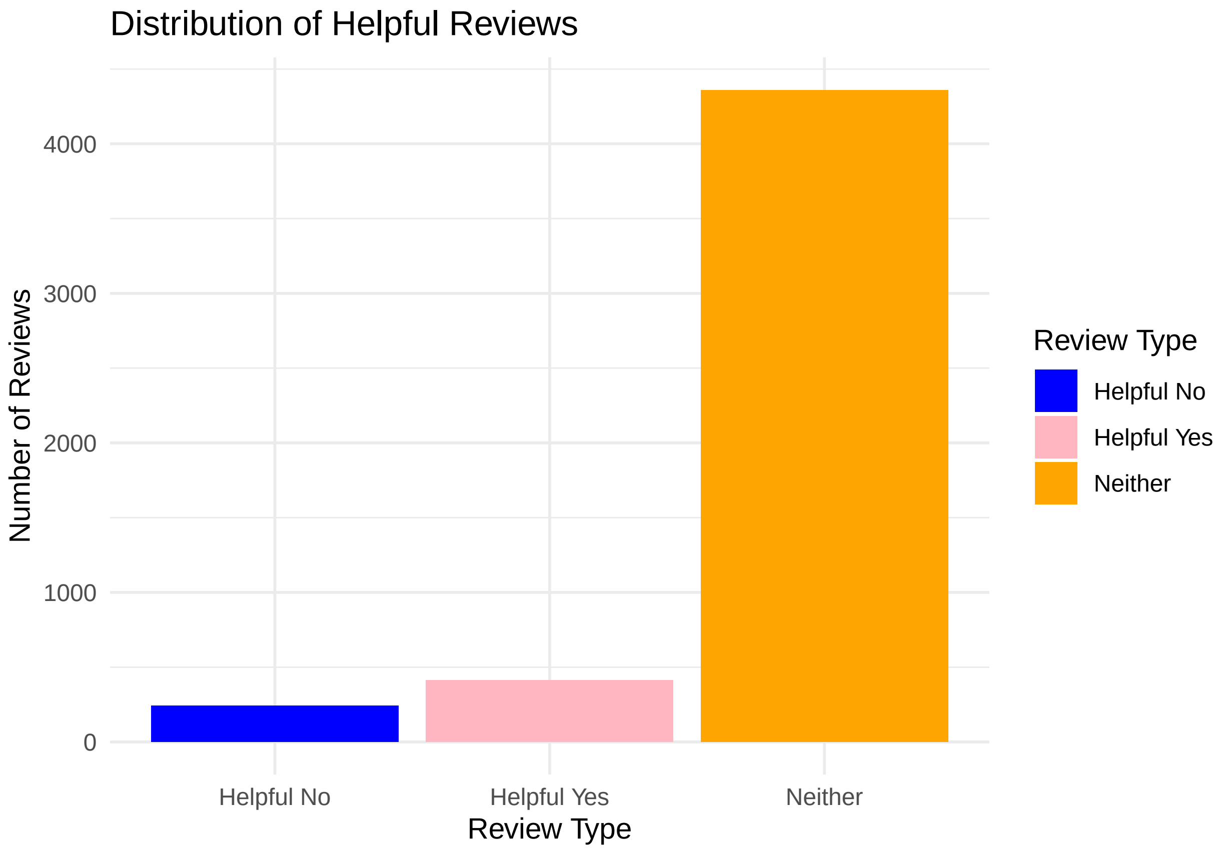
```r
plot1<- ggplot(data)+geom_point(
           aes(x = helpful_yes, y = overall))+
          labs(title = "Scatterplot of Helpful Votes and Amazon Ratings", x = "Number of Helpful Votes"
  theme_minimal()

plot2<- ggplot(data)+geom_point(
  aes(x = total_vote, y = overall))+
  labs(title = "Scatterplot of Total Votes and Amazon Ratings", x = "Total Number of Votes", y = "Amazon
  theme_minimal()

plot3<- ggplot(data)+geom_point(
  aes(x = day_diff, y = overall))+
  labs(title = "Scatterplot of Difference in Time and Amazon Ratings", x = "Number of Days", y = "Amazon
  theme_minimal()
combined_plot <- plot1 + plot2 + plot3 + plot_layout(ncol = 1)
combined_plot
```



```r
# distribution of helpful ratings
data <- data.frame(
    Review_Type = c("Helpful Yes", "Helpful No", "Neither"),
    Count = c(413, 241,4360)  # Example numbers
  )
ggplot(data, aes(x = Review_Type, y = Count, fill = Review_Type)) +
    geom_bar(stat = "identity") +
    labs(title = "Distribution of Helpful Reviews", x = "Review Type", y = "Number of Reviews", fill =
    theme_minimal() +
    scale_fill_manual(values = c("Helpful Yes" = "lightpink", "Helpful No" = "blue", "Neither"= "orange
```


Distribution of Helpful Reviews
Number of Reviews
Helpful No
Helpful Yes
Neither
Review Type
Review Type
Helpful No
Helpful Yes
Neither