Portfolio Assignment 3 - Chi square test

Nadia Hajighassem

2024-01-12

```
#install packages
pacman::p_load('car',
               'ggplot2',
               'tidyverse',
               'dplyr',
               'rcompanion',
               'hrbrthemes')
```

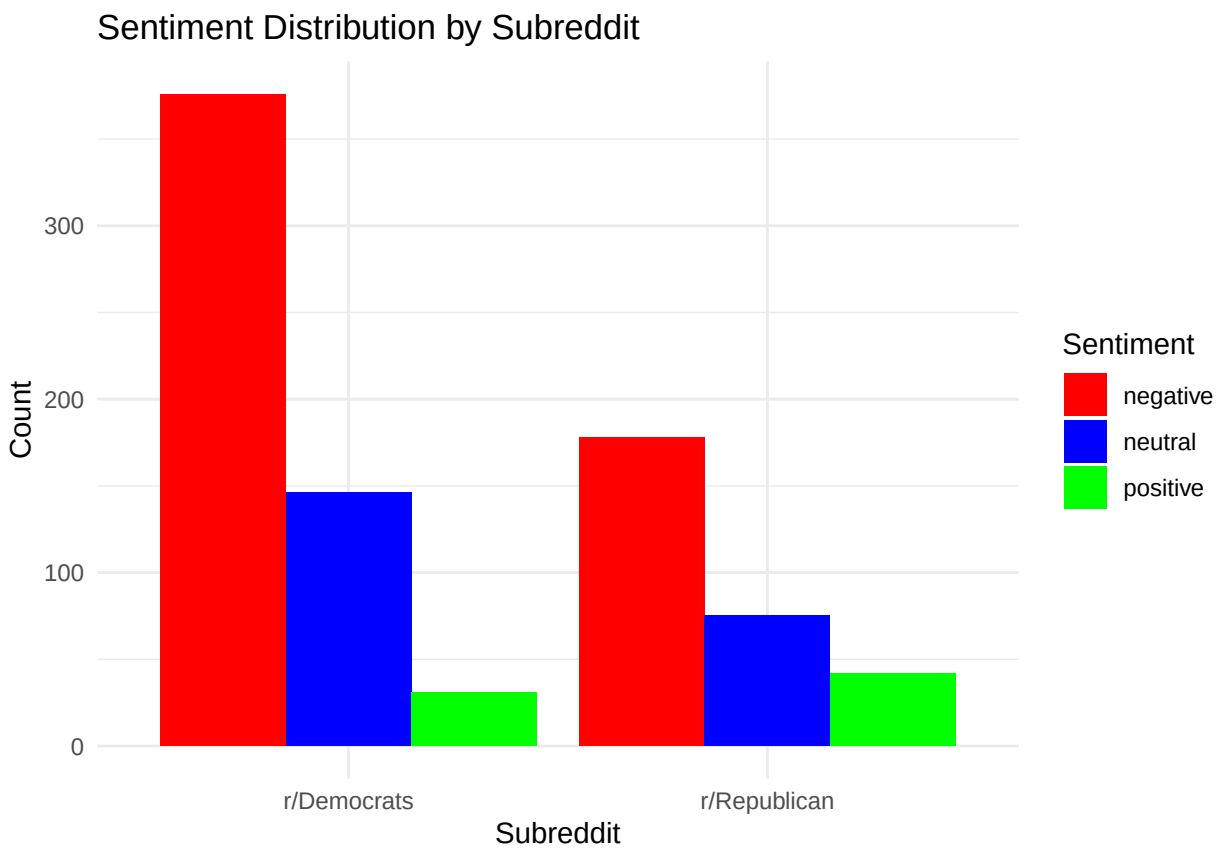
```
# load in data
df <- read.csv("combined_df.csv")
```

Make Contingency table

```
# prepare data for chi-square test
sentiment_table <- table(df$Subreddit, df$Sentiment)

# View the table
print(sentiment_table)
```

```
##
##                negative neutral positive
##   r/Democrats       376     146       31
##   r/Republican      178      75       42
```

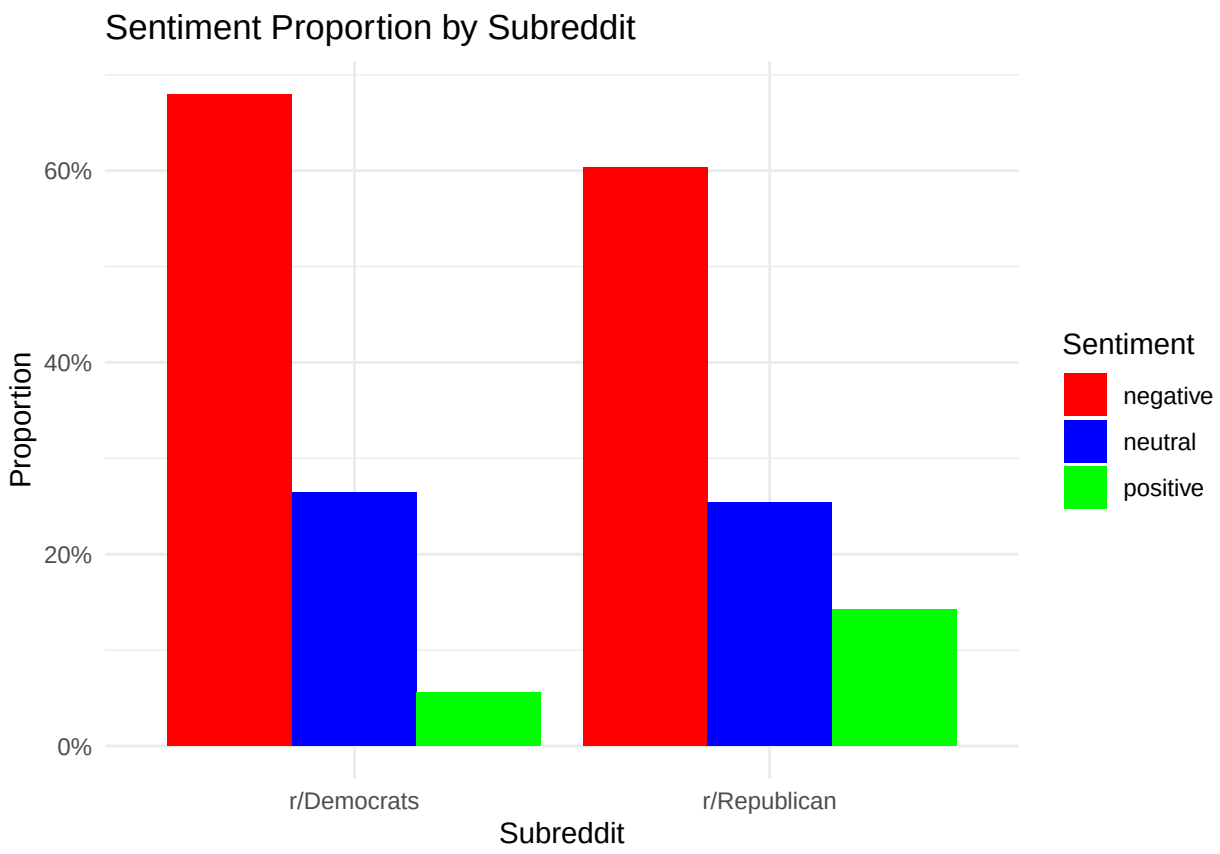
```
color <- c("red", "blue", "green")


ggplot(df, aes(x = Subreddit, fill = Sentiment)) +
  geom_bar(position = "dodge") +
  scale_fill_manual(values = color) +
  labs(title = "Sentiment Distribution by Subreddit",
       x = "Subreddit",
       y = "Count") +
    theme_minimal()
```




```
# Calculate the proportion of each sentiment within each Subreddit
df_prop <- df %>%
  group_by(Subreddit, Sentiment) %>%
  summarise(Count = n()) %>%
  mutate(Proportion = Count / sum(Count))
```

```
## `summarise()` has grouped output by 'Subreddit'. You can override using the
## `.groups` argument.
```

```
# Plot the data as proportions
ggplot(df_prop, aes(x = Subreddit, y = Proportion, fill = Sentiment)) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_manual(values = color) +
  labs(
    title = "Sentiment Proportion by Subreddit",
    x = "Subreddit",
    y = "Proportion",
    fill = "Sentiment"
  ) +
  theme_minimal() +
  scale_y_continuous(labels = scales::percent)
```




```
# lets check the numbers
df_prop
```

```
## # A tibble: 6 x 4
## # Groups:   Subreddit [2]
##   Subreddit    Sentiment Count Proportion
##   <chr>        <chr>     <int>      <dbl>
## 1 r/Democrats  negative    376     0.680
## 2 r/Democrats  neutral     146     0.264
## 3 r/Democrats  positive     31     0.0561
## 4 r/Republican negative    178     0.603
## 5 r/Republican neutral      75     0.254
## 6 r/Republican positive     42     0.142
```

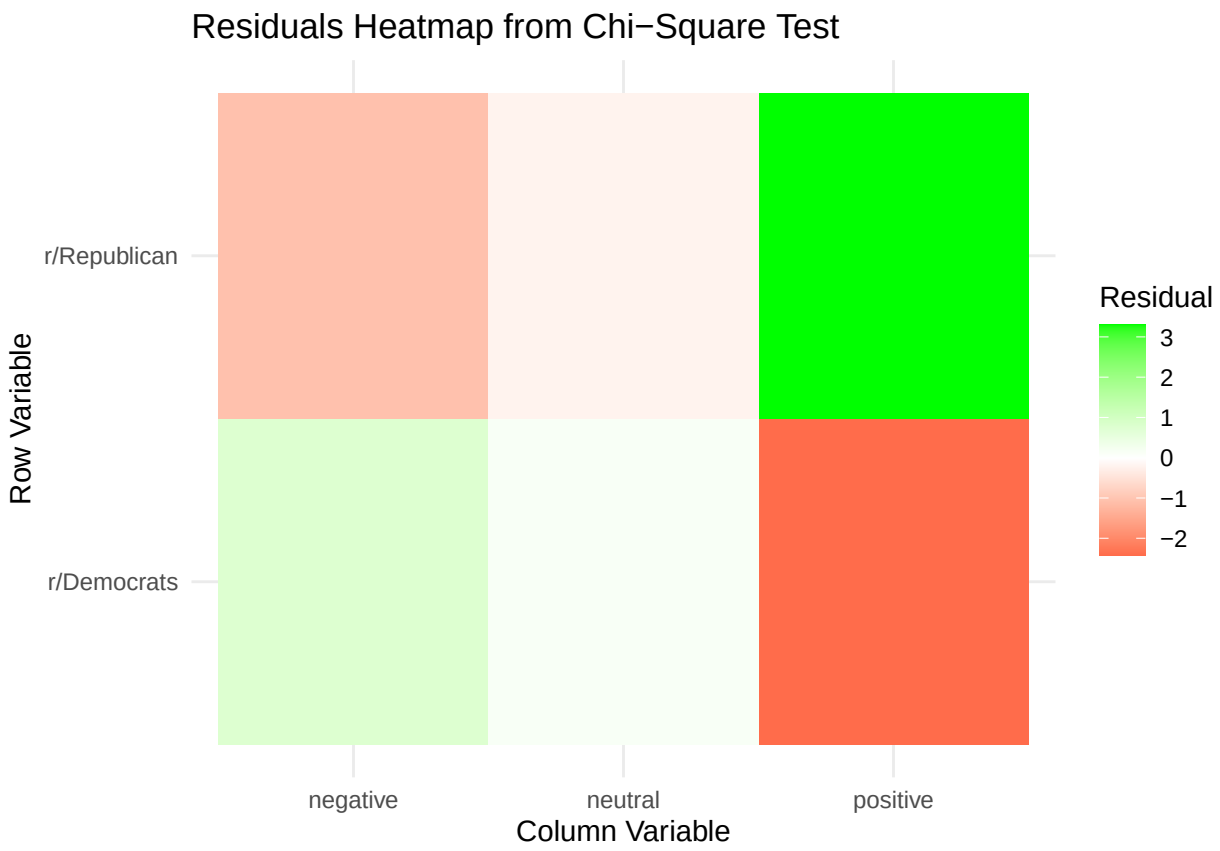
```
# chi-square test of independence
chi_sq_test <- chisq.test(sentiment_table)
print(chi_sq_test)
```

```
##
##  Pearson's Chi-squared test
##
## data:  sentiment_table
## X-squared = 18.445, df = 2, p-value = 9.88e-05
```

```r
# lets check the residuals
residuals <- chi_sq_test$residuals
residuals
```

```
##
##                  negative    neutral   positive
##   r/Democrats   0.7746543  0.1566766 -2.4066425
##   r/Republican -1.0606193 -0.2145140  3.2950585
```

```r
library(ggplot2)

# convert to df
residuals_df <- as.data.frame(as.table(residuals))
colnames(residuals_df) <- c("Row", "Column", "Residual")

# heatmap
ggplot(residuals_df, aes(x = Column, y = Row, fill = Residual)) +
  geom_tile() +
  scale_fill_gradient2(low = "red", mid = "white", high = "green") + # green is positive, red is negati
  labs(title = "Residuals Heatmap from Chi-Square Test",
       x = "Column Variable",
       y = "Row Variable",
       fill = "Residual") +
  theme_minimal()
```



```r
# Calculate Cramer's V
cramers_v <- cramerV(sentiment_table)

# Print the result
print(cramers_v)
```

```
## Cramer V
##   0.1475
```