Project Proposal

The BEST Fit - Olivia Encarcion, Leo Yang, Philip Lin, Allison Yang

```
library(tidyverse)
library(tidymodels)
library(ggplot2)
library(dplyr)

newspopularity <- read_csv("data/OnlineNewsPopularity.csv")
```

Introduction

...

Data description

...

Data Processing

To prepare the OnlineNewsPopularity dataset for analysis, we need to perform several processing steps. The first step is go through the dataset and handle and missing data. This includes checking for missing values and remove them from the dataset. If a feature has too many missing values, we may decide to drop it completely. We also will chose to merge different data channels into one categorical variable. The dataset includes multiple variables representing different data channels. Instead of having separate columns for each channel, we can create a single categorical variable with different levels. Additionally, since the dataset has multiple variables for whether the article was uploaded on a weekday or weekend we are going to merge these variables in a meaningful way. This transformation simplifies modeling and visualization. Lastly, we can convert highly skewed variabls using log transformations to improve model performances.

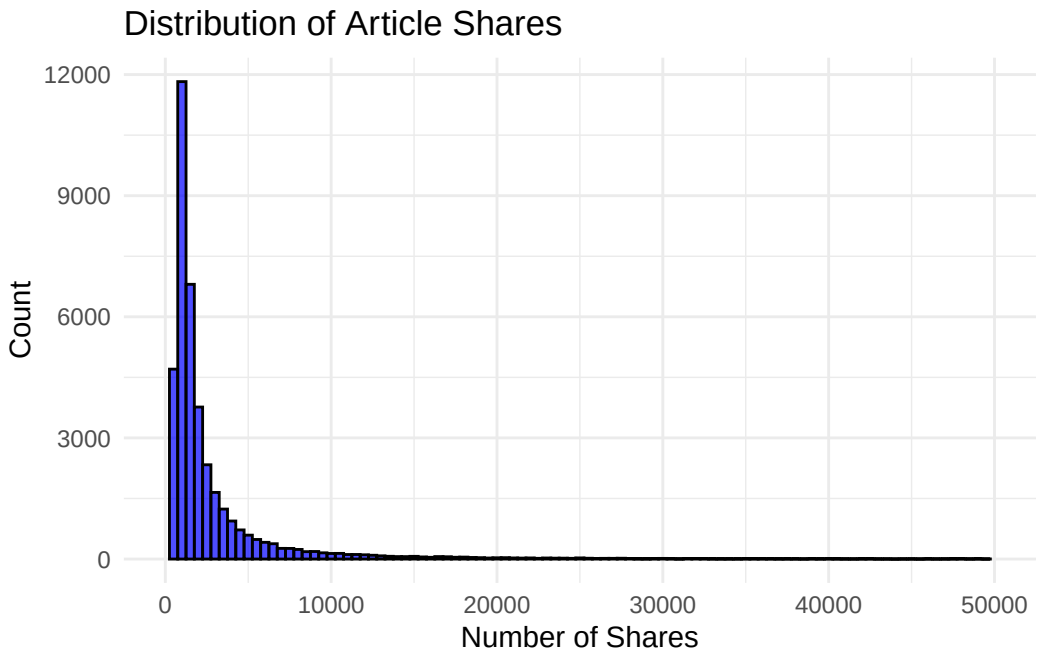
```r
summary(newspopularity$shares)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      1     946    1400    3395    2800  843300 
```

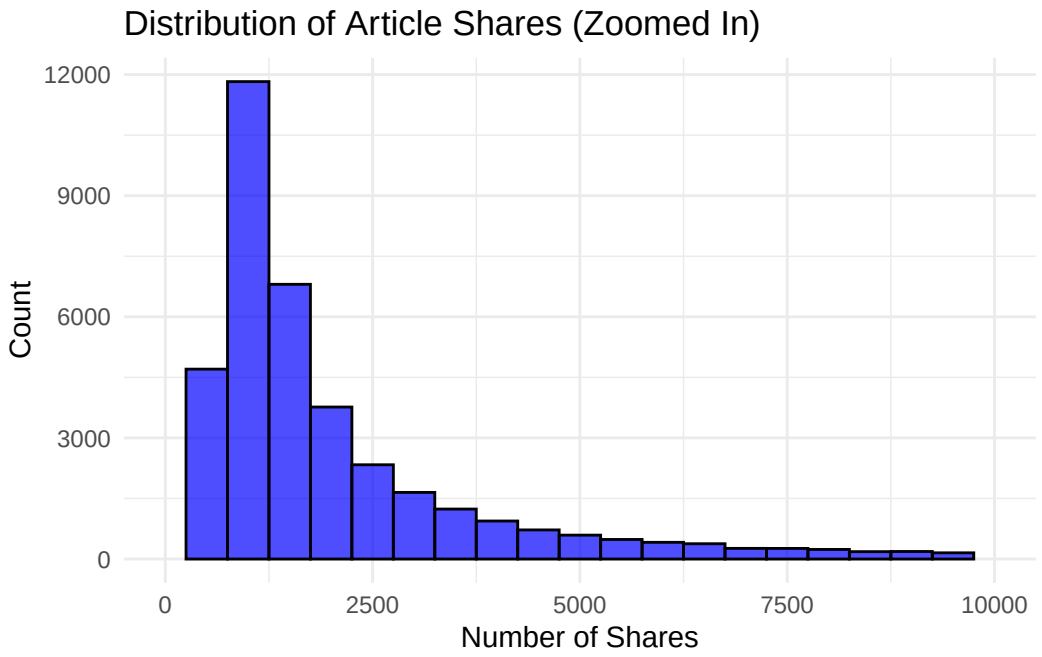
```r
newspopularity |>
  summarize(
    mean_shares = mean(shares),
    median_shares = median(shares),
    sd_shares = sd(shares),
    min_shares = min(shares),
    max_shares = max(shares),
    q1 = quantile(shares, 0.25),
    q3 = quantile(shares, 0.75)
  )
```

```
# A tibble: 1 x 7
  mean_shares median_shares sd_shares min_shares max_shares    q1    q3
        <dbl>         <dbl>     <dbl>      <dbl>      <dbl> <dbl> <dbl>
1       3395.          1400    11627.          1     843300   946  2800
```

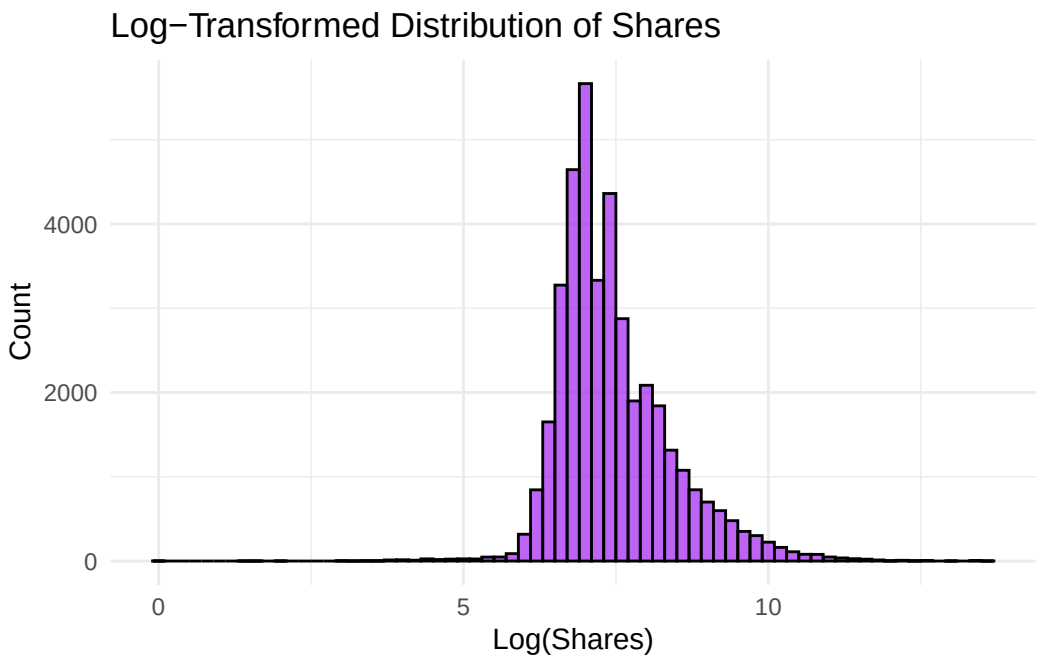
```r
ggplot(newspopularity, aes(x = shares)) +
  geom_histogram(binwidth = 500, fill = "blue", color = "black", alpha = 0.7) +
  scale_x_continuous(limits = c(0, 50000)) +  # Limit for better visualization
  labs(title = "Distribution of Article Shares",
       x = "Number of Shares",
       y = "Count") +
  theme_minimal()
```




```
ggplot(newspopularity, aes(x = shares)) +
  geom_histogram(binwidth = 500, fill = "blue", color = "black", alpha = 0.7) +
  scale_x_continuous(limits = c(0, 10000)) +  # Zoom in on the first 10,000 shares
  labs(title = "Distribution of Article Shares (Zoomed In)",
       x = "Number of Shares",
       y = "Count") +
  theme_minimal()
```




```
ggplot(newspopularity, aes(x = log(shares))) +
  geom_histogram(binwidth = 0.2, fill = "purple", color = "black", alpha = 0.7) +
  labs(title = "Log-Transformed Distribution of Shares",
       x = "Log(Shares)",
       y = "Count") +
  theme_minimal()
```



...

Analysis approach

...

Data dictionary

The data dictionary can be found here [Update the link and remove this note!]