Value Types

Numeric:

timedelta, n_tokens_title, n_tokens_content, n_unique_tokens, n_non_stop_words, n_non_stop_unique_tokens, num_hrefs, num_self_hrefs, num_imgs, num_videos, average_token_length, num_keywords, global_subjectivity, global_sentiment_polarity, global_rate_positive_words, global_rate_negative_words, shares, etc.

Categorical:

url, data_channel_is_lifestyle, data_channel_is_entertainment, data_channel_is_bus, data_channel_is_socmed, data_channel_is_tech, data_channel_is_world, weekday_is_monday, weekday_is_tuesday, etc. Boolean: is_weekend

Coding Schemes

data_channel_is_lifestyle, data_channel_is_entertainment, etc. use 1 for "Yes" and 0 for "No". weekday_is_monday, weekday_is_tuesday, etc. are using 1 to represent "True" and 0 to represent "False".

Data Quantity

Format: The data is in CSV format. The dataset of Mashable articles written before 2015, obtained from the UCI Machine Learning Repository

Database Size

Dataset contains 61 columns and has 39,644 unique values.

Basic Statistics

```
mashable_data <- read.csv("/Users/keiichiro_watanabe/Desktop/CSC324/IndividualProject/data/OnlineNewsPo

# Calculate the mean for shares
mean_shares <- mean(mashable_data$shares, na.rm = TRUE)
cat("Mean Shares:", mean_shares, "\n")
```

```
## Mean Shares: 3395.38
```

```
# Correlation between shares and global_rate_positive_words
cor_shares_positive_words <- cor(mashable_data$shares, mashable_data$global_rate_positive_words, use =
cat("Correlation between shares and global_rate_positive_words:", cor_shares_positive_words, "\n")
```

```
## Correlation between shares and global_rate_positive_words: 0.0005432341
```

```r
# Correlation between shares and n_tokens_content
cor_shares_tokens <- cor(mashable_data$shares, mashable_data$n_tokens_content, use = "complete.obs")
cat("Correlation between shares and n_tokens_content:", cor_shares_tokens, "\n")
```

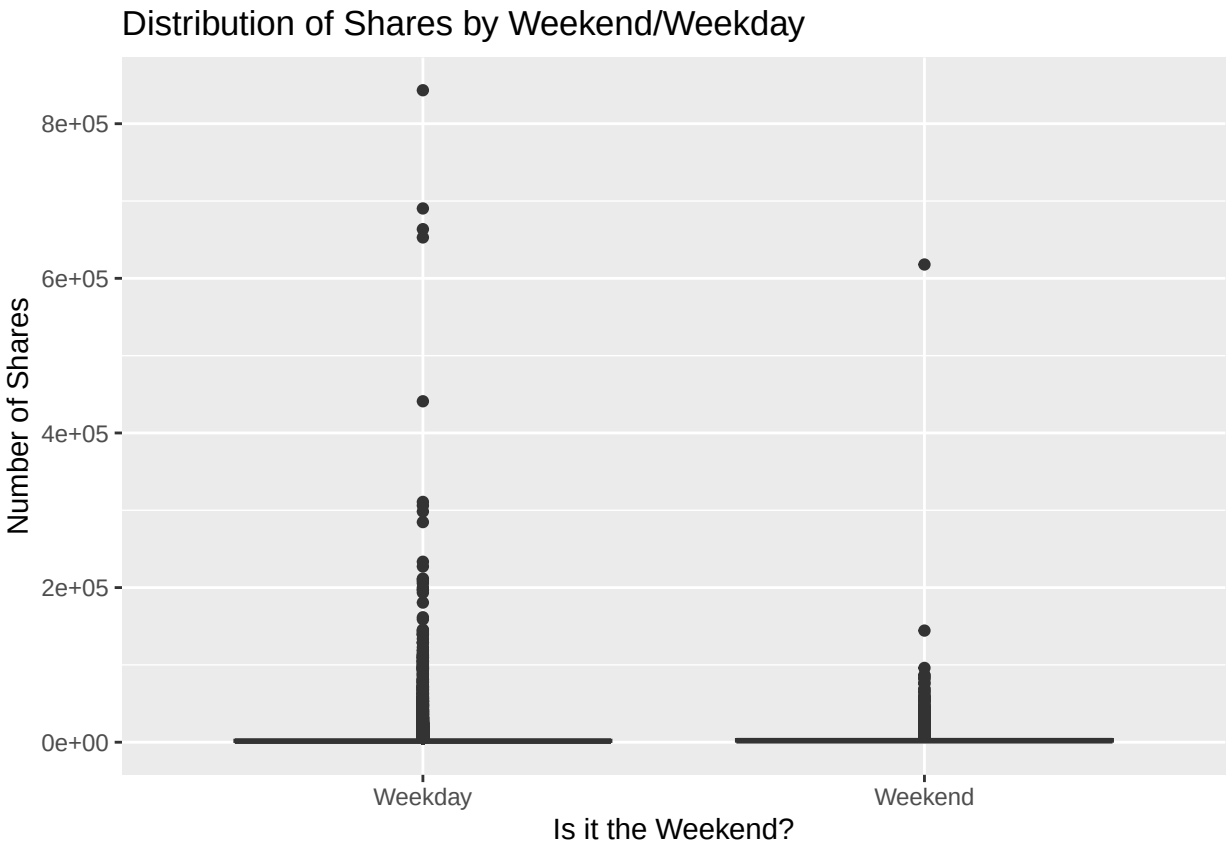
```
## Correlation between shares and n_tokens_content: 0.002458984
```

```r
# Correlation between shares and global_rate_negative_words
cor_shares_negative_words <- cor(mashable_data$shares, mashable_data$global_rate_negative_words, use =
cat("Correlation between shares and global_rate_negative_words:", cor_shares_negative_words, "\n")
```

```
## Correlation between shares and global_rate_negative_words: 0.006615173
```

```r
library(ggplot2)
# Convert the is_weekend column to a factor for better labeling
mashable_data$is_weekend <- factor(mashable_data$is_weekend, levels = c(0, 1), labels = c("Weekday", "W

# Create the box plot
ggplot(mashable_data, aes(x=is_weekend, y=shares)) +
  geom_boxplot() +
  ggtitle("Distribution of Shares by Weekend/Weekday") +
  xlab("Is it the Weekend?") +
  ylab("Number of Shares")
```



Data Quality

The data includes multiple characteristics that could be relevant to the business question of what contributes to an article's popularity. Are certain topics more popular than others? What days of the week should we publish articles to get the most shares?

variables below might be useful: n_tokens_content, global_rate_negative_words, weekday_is_monday, weekday_is_tuesday, etc. is_weekend, num_shares, num_videos