Disney Reviews

Analysis of Disneyland Reviews

NOVEMBER 21, 2023

The enchanting world of Disney, a global cultural phenomenon, has captivated audiences for decades, transcending generations and geographical boundaries. Founded by Walt Disney in 1923, the Walt Disney Company has become synonymous with magic, imagination, and storytelling. Disney, with its iconic characters, timeless animations, and immersive theme parks, has created a world of joy and wonder that resonates in the hearts of millions. Disneyland parks embody enchantment, inviting visitors to enter fantastical realms where dreams unfold.

In this project, our aim is to thoroughly explore and analyze Disneyland reviews, uncovering valuable insights into visitor sentiments, preferences, and key themes across Disneyland California, Disneyland Hong Kong, and Disneyland Paris. By delving into a vast corpus of reviews, we seek to understand the unique experiences and perceptions of visitors. Through a journey of data exploration and refinement, we aim to extract actionable insights that provide a holistic view of the enchanting world of Disneyland.

Questions of interest

Our primary goal is to address the following four key questions of interest through the analysis of Disneyland reviews:

- What are the most frequently mentioned aspects of the Disney experience in general?
- Are there differences in the types of feedback or topics discussed between the branches in California, Paris and Hong Kong?
- Are there any trends or seasonal patterns in the reviews?
- Do certain times of the year lead to different types of feedback?

These questions are compelling as they target key aspects of the Disneyland experience, promising valuable insights for both visitor satisfaction and strategic park management. Identifying frequently mentioned aspects provides a blueprint for prioritizing and enhancing universally appreciated elements. Analyzing differences in feedback between branches acknowledges diverse preferences, enabling tailored experiences for varied visitor bases. Investigating trends and seasonal patterns equips Disneyland with the foresight to optimize operations and address recurring issues during specific periods. Understanding if feedback varies across seasons guides targeted improvements, ensuring a dynamic and responsive approach to visitor expectations. Overall, addressing these questions enhances the Disneyland experience, making it more personalized, culturally attuned, and enjoyable for visitors worldwide.

Loading packages

```
library(tidytext)
library(readr)
library(dplyr)
library(tidyr)
library(ggplot2)
library(tm)
library(topicmodels)
library(keyATM)
library(quanteda)
library(forcats)
library(rmarkdown)
```

Loading the dataset

```
disney <- read_csv("Disneyland_Reviews.csv")
```

The dataset used for this analysis was obtained from Kaggle, comprising 42,655 observations, each representing a distinct review. The dataset contains six columns:

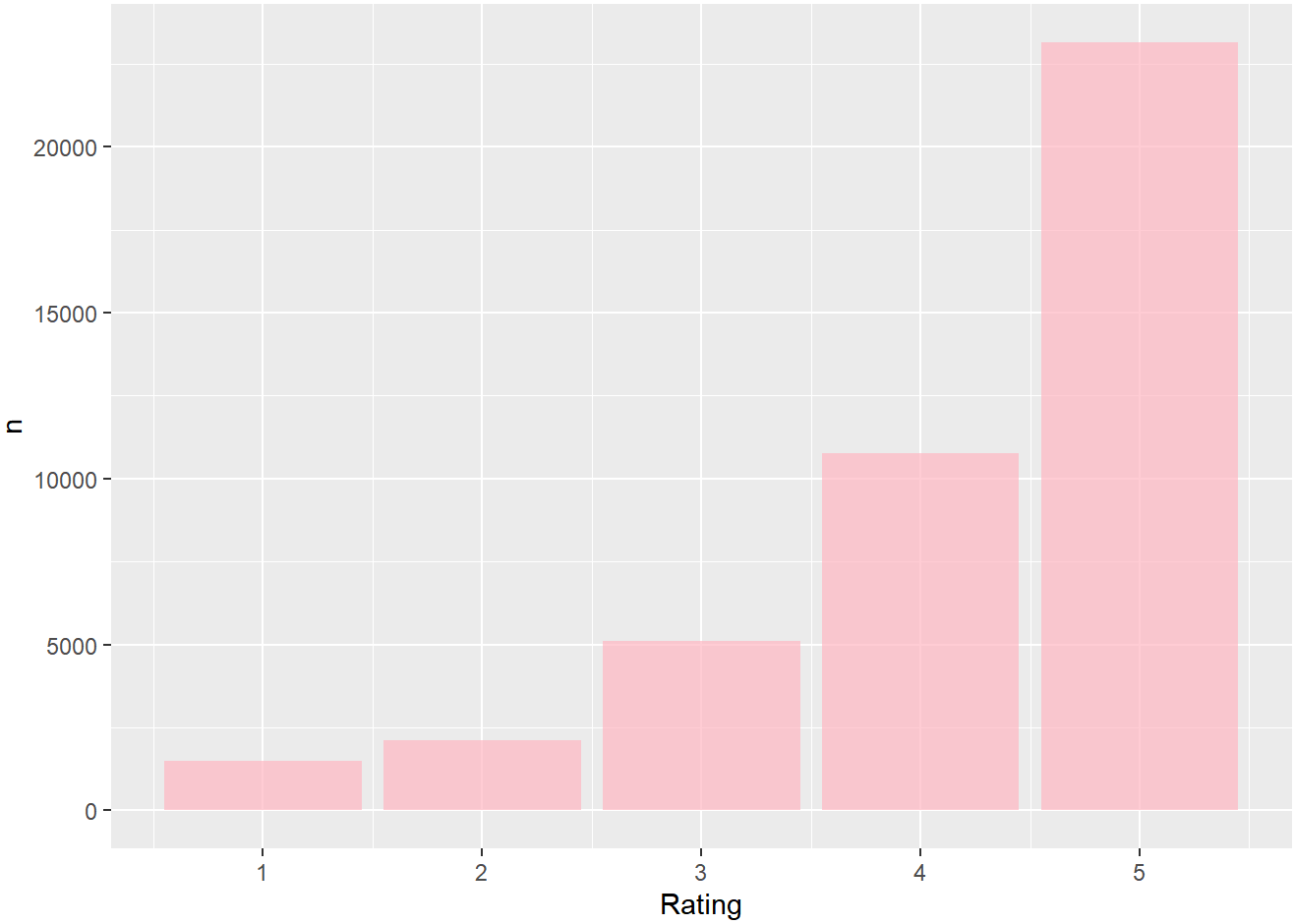
1. **Review ID:** The unique identifier for each review.
2. **Rating:** The numerical rating (from 1 to 5) assigned by the reviewer.
3. **Year_Month:** The time stamp indicating when the review was submitted.
4. **Reviewer Location:** The location of the reviewer.
5. **Review Text:** The actual content of the review.
6. **Branch:** The specific Disneyland branch associated with the review.

Exploring the Data

First, we are going to explore our data. This involves taking a closer look to understand its patterns, check out how key elements are distributed, and spot any interesting trends. This exploration sets the stage for more detailed analyses, helping us ask informed questions and dive deeper into the dataset.

Count of Rating

```
disney %>%
  count(Rating) %>%
  ggplot(aes(Rating, n)) +
  geom_col(fill = "lightpink", alpha = 0.7)
```



Looking at the graph, we observe a left-skewed distribution. This means that most people who provided reviews tended to give very positive ratings. In fact, the majority of reviewers gave a rating of 5, which is the highest possible score in the range of 1 to 5.

This skewness indicates a strong positive sentiment among the reviewers. It suggests that Disneyland is highly appreciated by a significant portion of the audience. On the other hand, fewer reviewers gave lower ratings, indicating that dissatisfaction or negative experiences are less common among the reviewers.

Distribution of Words in the Reviews

We are going to prepare the 'tidydisney' dataset for textual analysis. This includes tokenizing the 'Review_Text' column, removing common stop words, digits, and punctuation, ensuring that the resulting 'tidydisney' dataset only contains meaningful words for subsequent analysis. Additionally, we filtered out any entries with missing or empty words, finalizing the dataset for further exploration.

```
tidydisney <- disney %>%
  unnest_tokens(input=Review_Text,
                output = word)%>%
   anti_join(stop_words)
```

```
Joining with `by = join_by(word)`
```

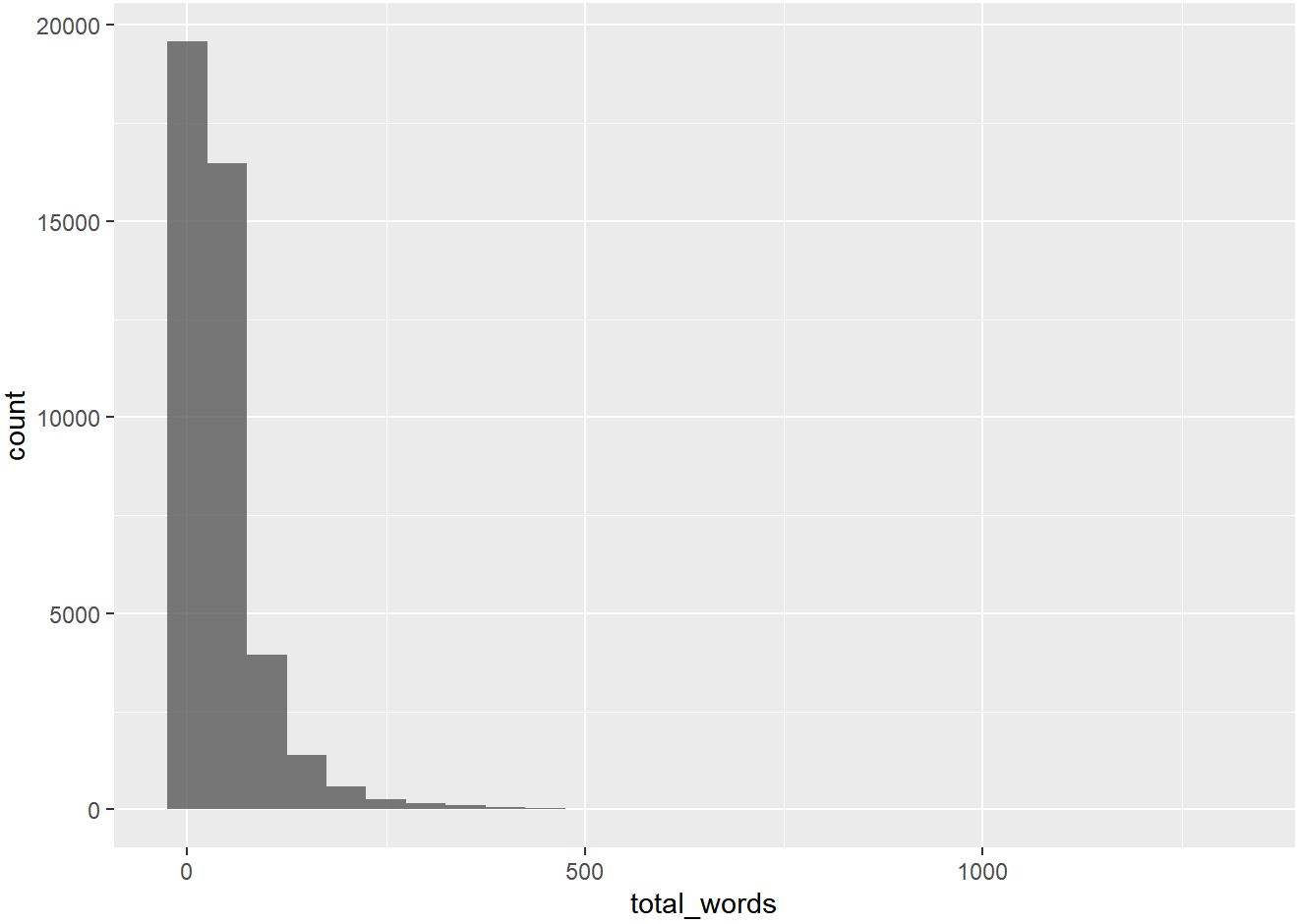
```
tidydisney$word <- gsub('[0-9]+', '', tidydisney$word)
tidydisney$word <- gsub('[[:punct:] ]+', '', tidydisney$word)

any(is.na(tidydisney$word) | nchar(tidydisney$word) == 0)
```

```
[1] TRUE
```

```
tidydisney <- subset(tidydisney, nchar(word) > 0)
```

With the preprocessing steps completed, the 'tidydisney' dataset is now prepared for further analysis.

```
wordsperreview <- tidydisney %>%
   count(Review_ID, name = "total_words")

wordsperreview %>%
  ggplot(aes(total_words)) +
  geom_histogram(binwidth = 50, alpha = 0.8)
```



This histogram visualizes the distribution of total word counts in the reviews. The x-axis represents the range of word counts, divided into intervals of 100 words each. The y-axis displays the frequency of reviews falling within each word count range.

From the histogram, it's evident that the largest number of reviews have a word count between 0 and 250. This indicates that the majority of reviewers tend to provide feedback with relatively concise descriptions or comments, using fewer than 250 words. Reviews with word counts exceeding 250 are less common in this dataset.

```
unique_locations <- unique(disney$Reviewer_Location)
length(unique_locations)
```

```
[1] 162
```

With contributions from 162 distinct locations, this dataset reflects a broad geographic representation of Disney reviews.

Years of the Reviews

```
unique_year_month <- unique(disney$Year_Month)

unique(substr(unique_year_month, 1, 4))
```

```
 [1] "2019" "2018" "miss" "2017" "2016" "2015" "2014" "2013" "2012" "2011"
[11] "2010"
```

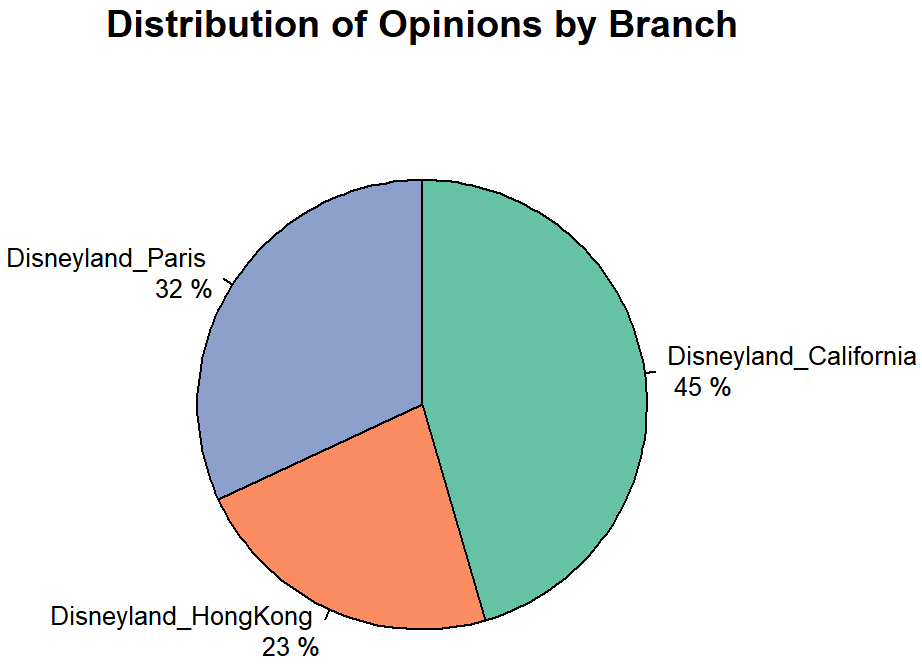
```
sum(grepl("missing", disney$Year_Month, ignore.case = TRUE))
```

```
[1] 2613
```

We have 2613 missing values in the Year_Month Column. This dataset includes reviews spanning from 2010 to 2019.

Distribution of Opinions by Branch

```
branch_counts <- table(disney$Branch)

custom_colors <- c("#66c2a5", "#fc8d62", "#8da0cb")


pie(branch_counts,
    labels = paste(names(branch_counts), "\n", round(prop.table(branch_counts)*100), "%"),
    main = "Distribution of Opinions by Branch",
    col = custom_colors,
    cex = 0.8,
    clockwise = TRUE
)
```



I found that the distribution of reviews across the three branches of Disneyland is as follows: approximately 23% of the reviews pertain to Disneyland Hong Kong, 32% are specific to Disneyland Paris, and the majority, constituting 45% of the reviews, focus on Disneyland California.

Sentiment Analysis

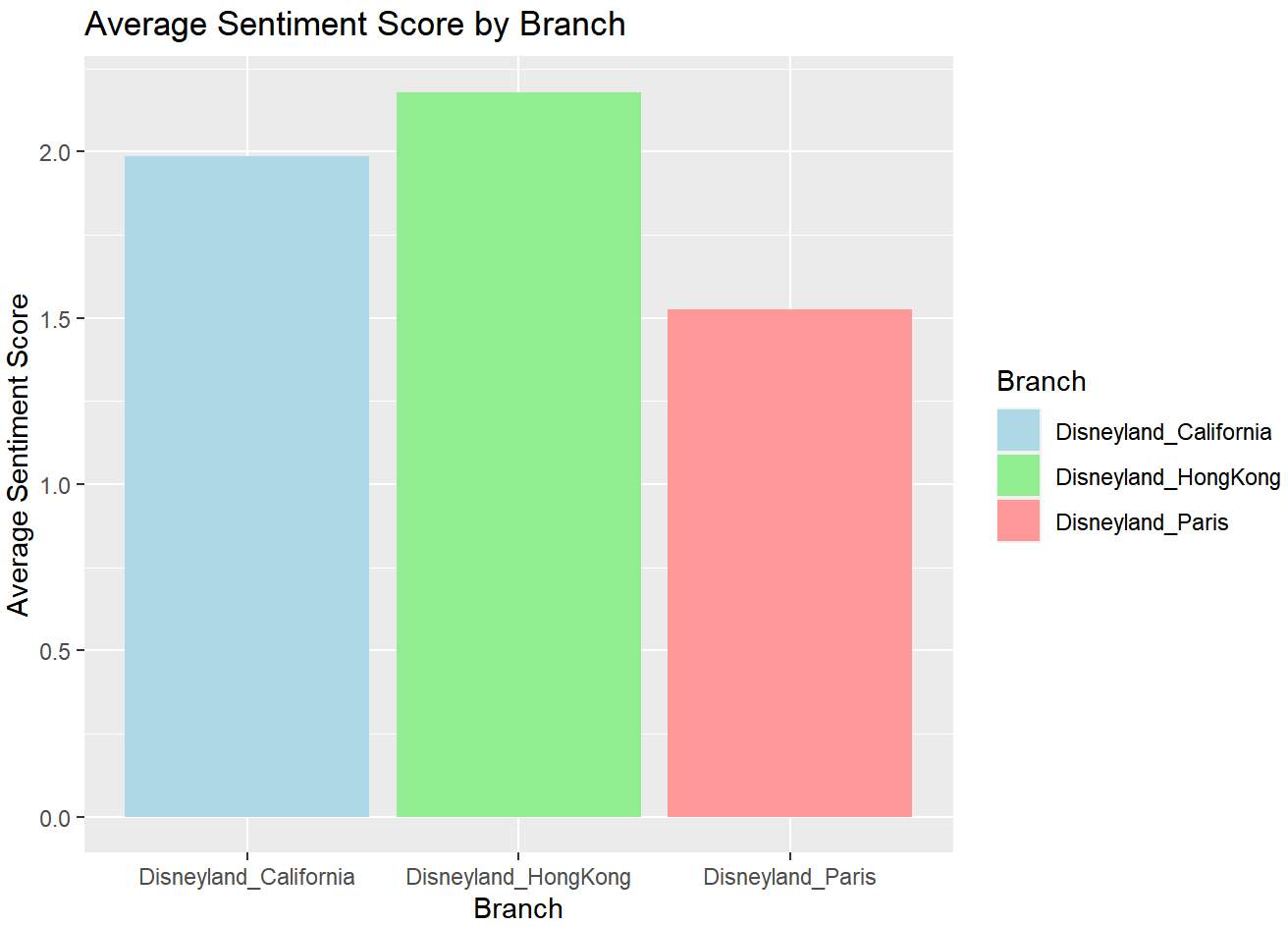
Moving forward, our focus is going to shift to conducting sentiment analysis specific to each branch. This entails evaluating and understanding the emotional tone expressed in the reviews associated with different Disneyland branches. By examining sentiment scores, we aim to uncover the varied perceptions and experiences reported by visitors at Disneyland California, Disneyland Hong Kong, and Disneyland Paris.

```
sentiment_scores <- tidydisney %>%
  inner_join(get_sentiments(), relationship = "many-to-many") %>%
  count(Branch, Review_ID, sentiment) %>%
  pivot_wider(names_from = sentiment, values_from = n) %>%
  mutate(sentiment_score = positive - negative)
```

```
Joining with `by = join_by(word)`
```

```
average_sentiment_by_branch <- sentiment_scores %>%
  group_by(Branch) %>%
  summarise(mean_sentiment_score = mean(sentiment_score, na.rm = TRUE))
```

```
ggplot(average_sentiment_by_branch, aes(x = Branch, y = mean_sentiment_score, fill = Branch)) +
  geom_col() +
  labs(title = "Average Sentiment Score by Branch",
       x = "Branch",
       y = "Average Sentiment Score") +
  scale_fill_manual(values = c("lightblue", "lightgreen", "#FF9999"))
```

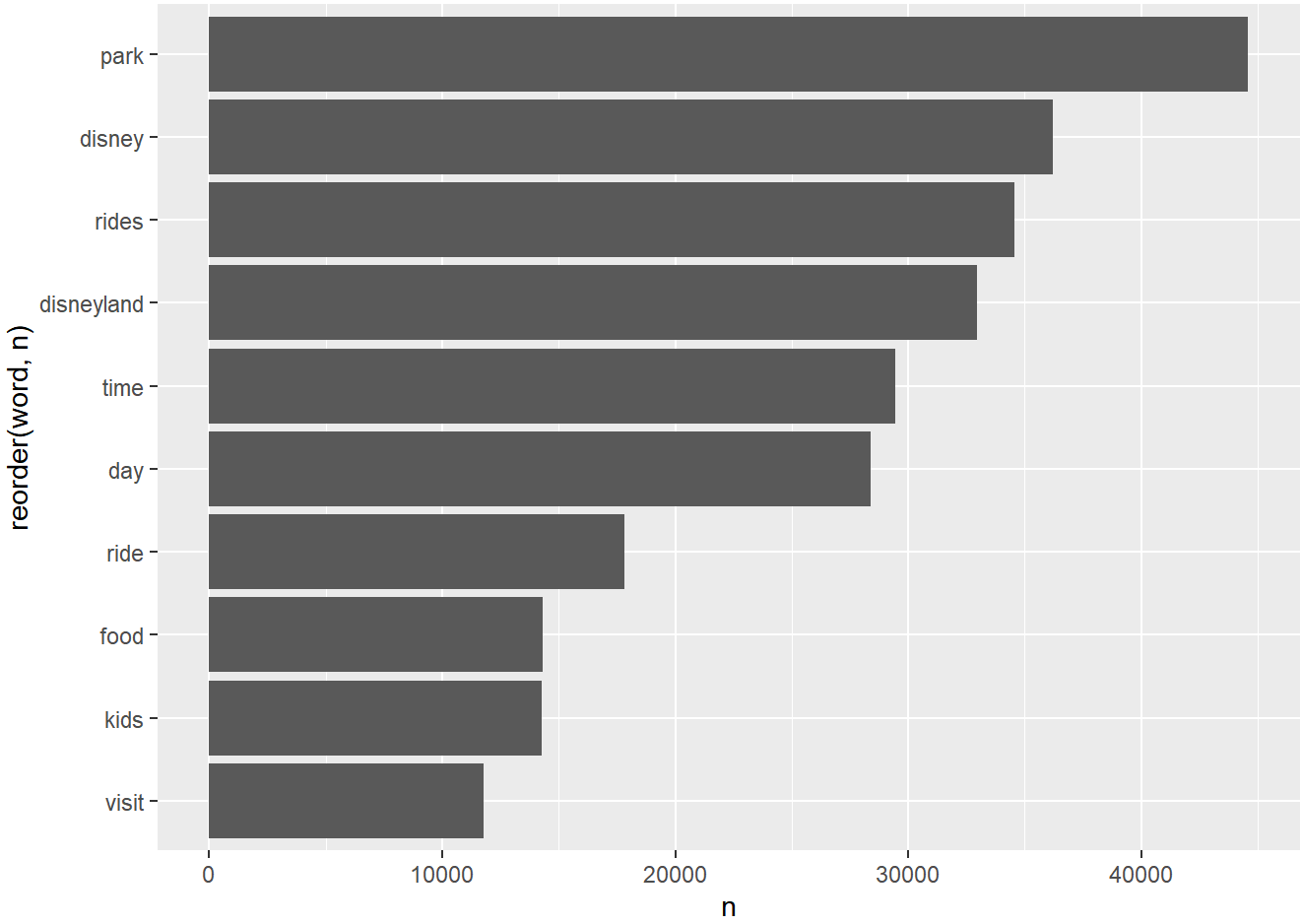


The sentiment scores for various Disneyland parks reveal intriguing insights. Disneyland California boasts a positive sentiment score of 1.987593, reflecting a favorable reception. Disneyland Hong Kong surpasses expectations with a sentiment score of 2.179445, an impressive feat considering it has the lowest number of reviews among the mentioned parks. Meanwhile, Disneyland Paris registers a respectable sentiment score of 1.524333.

Next, we are going to delve into addressing the specific questions of interest. This phase involves applying analytical methods and techniques to extract meaningful insights and provide informed responses.

- **What are the most frequently mentioned aspects of the Disney experience in general?**

```
tidydisney %>%
  count(word) %>%
  top_n(10, n) %>%
  ggplot(aes(y = reorder(word, n), x = n)) +
  geom_col()
```



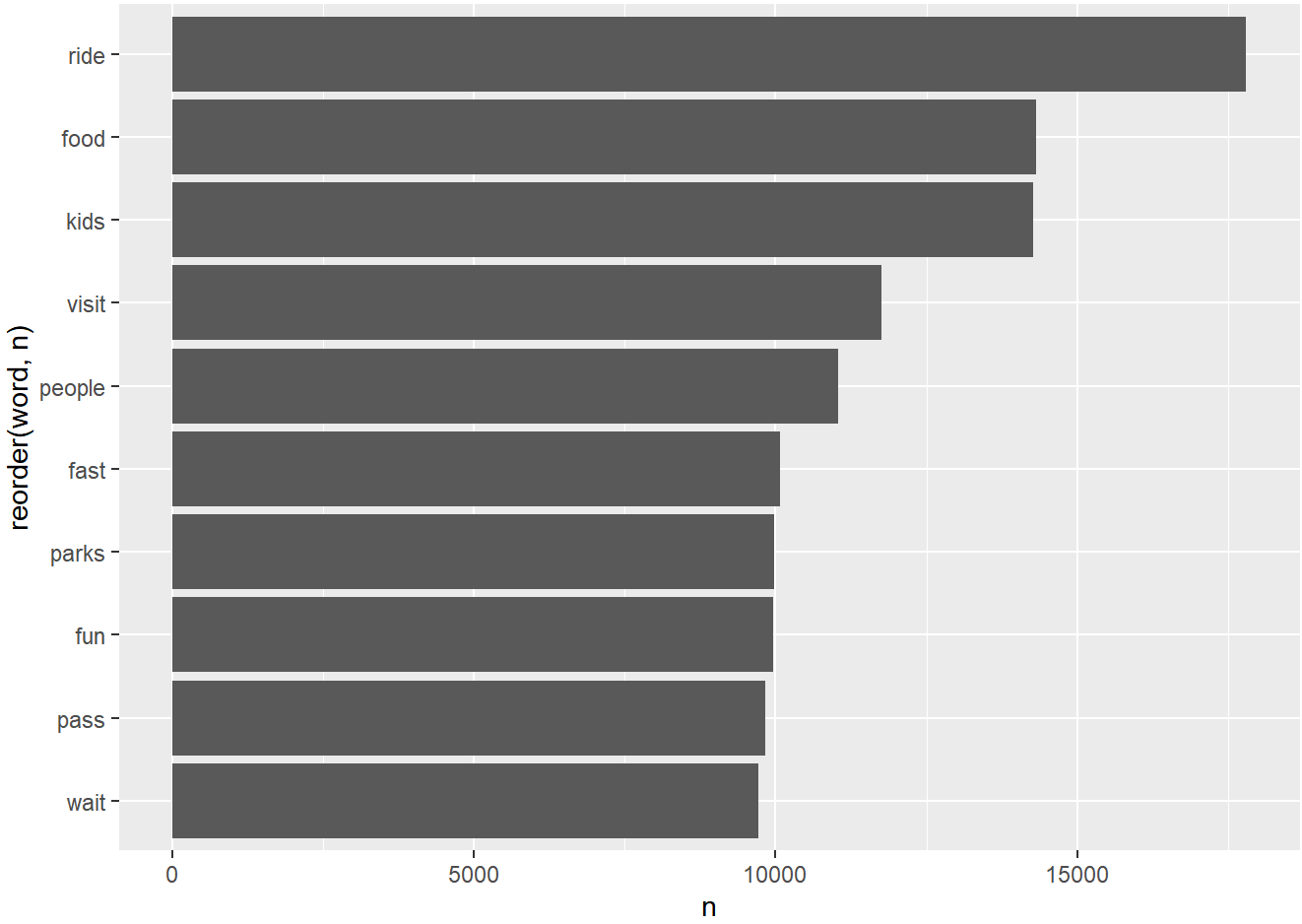
The top 10 most frequently occurring words in the dataset are “park,” “Disney,” “rides,” “Disneyland,” “time,” “day,” “ride,” “food,” “kids,” and “visit.” These words appear most frequently in the reviews and are likely central to the topics being discussed.

To refine our analysis, I have identified specific words (“Disney”, “Disneyland”, “park”, “time”,“rides” and “day”) that are highly specific to the context and are likely to occur frequently. I am designating them as custom stop words. By excluding these words from our analysis, I aim to focus on more insightful and distinctive content in the reviews, allowing me to extract more detailed information about visitors’ experiences.

```
new_stopsdisney1 <- tibble(word = c("disney","disneyland","park","time","day","rides"))
```

```
tidydisney %>%
  anti_join(new_stopsdisney1)%>%
  count(word) %>%
  top_n(10, n) %>%
  ggplot(aes(y = reorder(word, n), x = n)) +
  geom_col()
```

```
Joining with `by = join_by(word)`
```

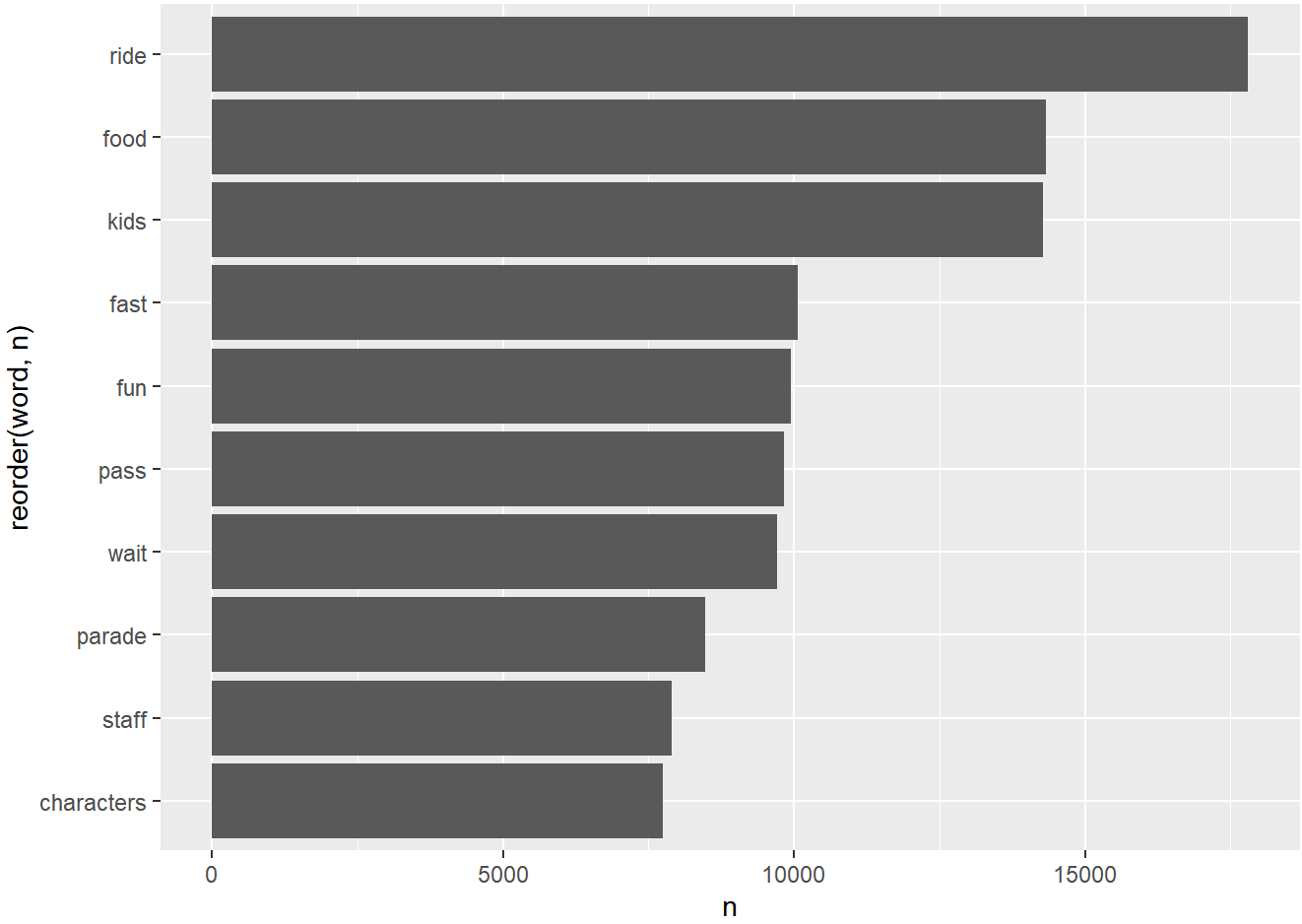


As we refine our list of stop words, we enhance our ability to uncover deeper patterns and extract more valuable information from the text data, ultimately leading to more accurate and insightful results.

```
new_stopsdisney2 <- tibble(word = c("disney","disneyland","park","time","day","rides","parks","2"
```

```
tidydisney %>%
  anti_join(new_stopsdisney2)%>%
  count(word) %>%
  top_n(10, n) %>%
  ggplot(aes(y = reorder(word, n), x = n)) +
  geom_col()
```

```
Joining with `by = join_by(word)`
```



The most common words in the reviews, which include “ride,” “food,” “kids,” “fast,” “fun,” “pass,” “wait,” “parade,” “staff,” and “characters,” provide significant insights into the key aspects of the visitors’ experiences at the park.

- “Ride” and “food” are central elements, indicating that visitors frequently discuss their experiences with attractions and dining options. These are critical components of the overall park experience.
- “Kids” suggests that the family-friendly nature of the park is a prominent theme. Many visitors likely discuss their experiences with children and family-oriented activities.
- “Fast” and “pass” could refer to features like FastPass systems or strategies to minimize wait times for attractions, highlighting the importance of efficient park navigation.
- “Fun” is a fundamental sentiment, indicating that visitors generally have an enjoyable time at the park.
- “Wait” indicates that ride queue times are a significant consideration for visitors, influencing their overall experience.
- “Parade” suggests that special events and entertainment offerings, such as parades, play a notable role in the visitor experience.
- “Staff” is crucial, as it signifies the importance of positive interactions with park employees in shaping visitor satisfaction.

- "Characters" indicate that encounters with iconic Disney characters are memorable experiences for visitors.

- **Are there differences in the types of feedback or topics discussed between the branches in California, Paris and Hong Kong?**

We are now set to examine the top words used in each branch after refining our stop words. This process aims to enhance the precision of our analysis by excluding irrelevant terms and providing a more focused exploration of the language patterns within each branch.

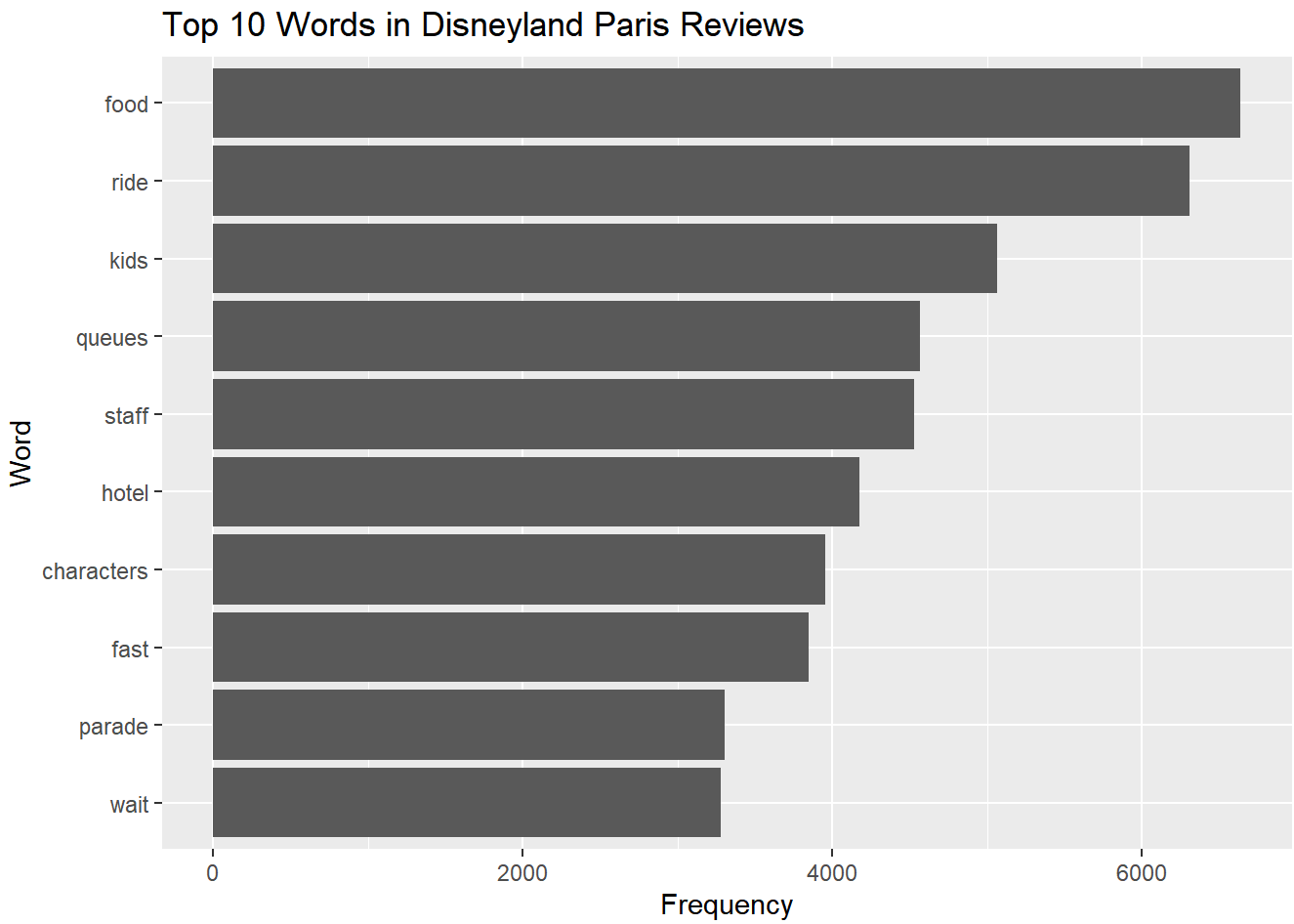
Paris

```
stopsparis <- tibble(word = c("disney","disneyland","park","time","day","rides","parks","2","visi
```

```
paris_reviews <- tidydisney %>%
  anti_join(stopsparis)%>%
  filter(Branch == "Disneyland_Paris") %>%
  count(word, sort = TRUE) %>%
  top_n(10, n)
```

```
Joining with `by = join_by(word)`
```

```
ggplot(data = paris_reviews, aes(y = reorder(word, n), x = n)) +
  geom_col() +
  labs(title = "Top 10 Words in Disneyland Paris Reviews") +
  xlab("Frequency") +
  ylab("Word")
```



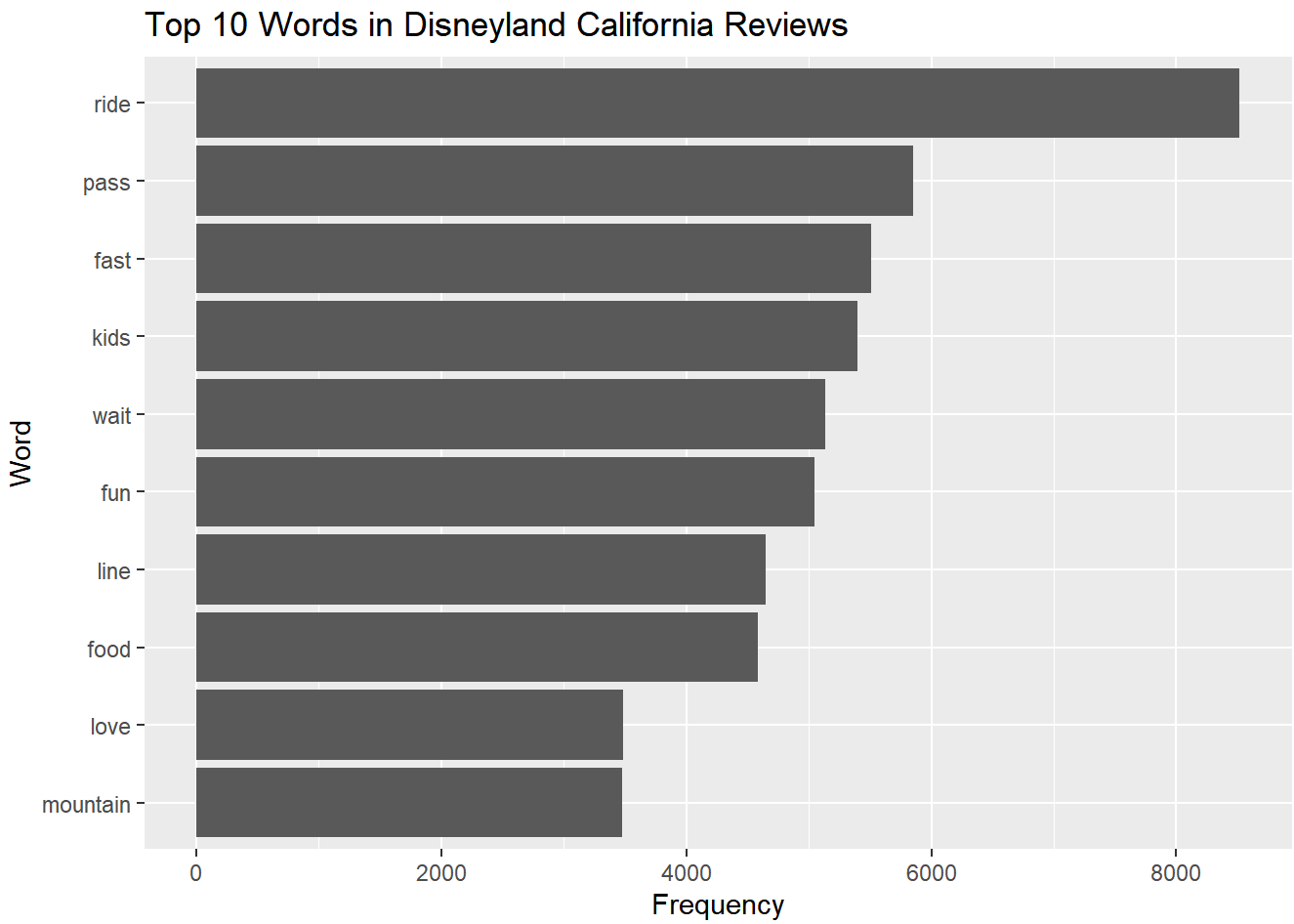
California

```
stopscali <- tibble(word = c("disney","disneyland","park","time","day","rides","parks","2","visit
```

```
california_reviews <- tidydisney %>%
  anti_join(stopscali)%>%
  filter(Branch == "Disneyland_California") %>%
  count(word, sort = TRUE) %>%
  top_n(10, n)
```

```
Joining with `by = join_by(word)`
```

```
ggplot(data = california_reviews, aes(y = reorder(word, n), x = n)) +
  geom_col() +
  labs(title = "Top 10 Words in Disneyland California Reviews") +
  xlab("Frequency") +
  ylab("Word")
```



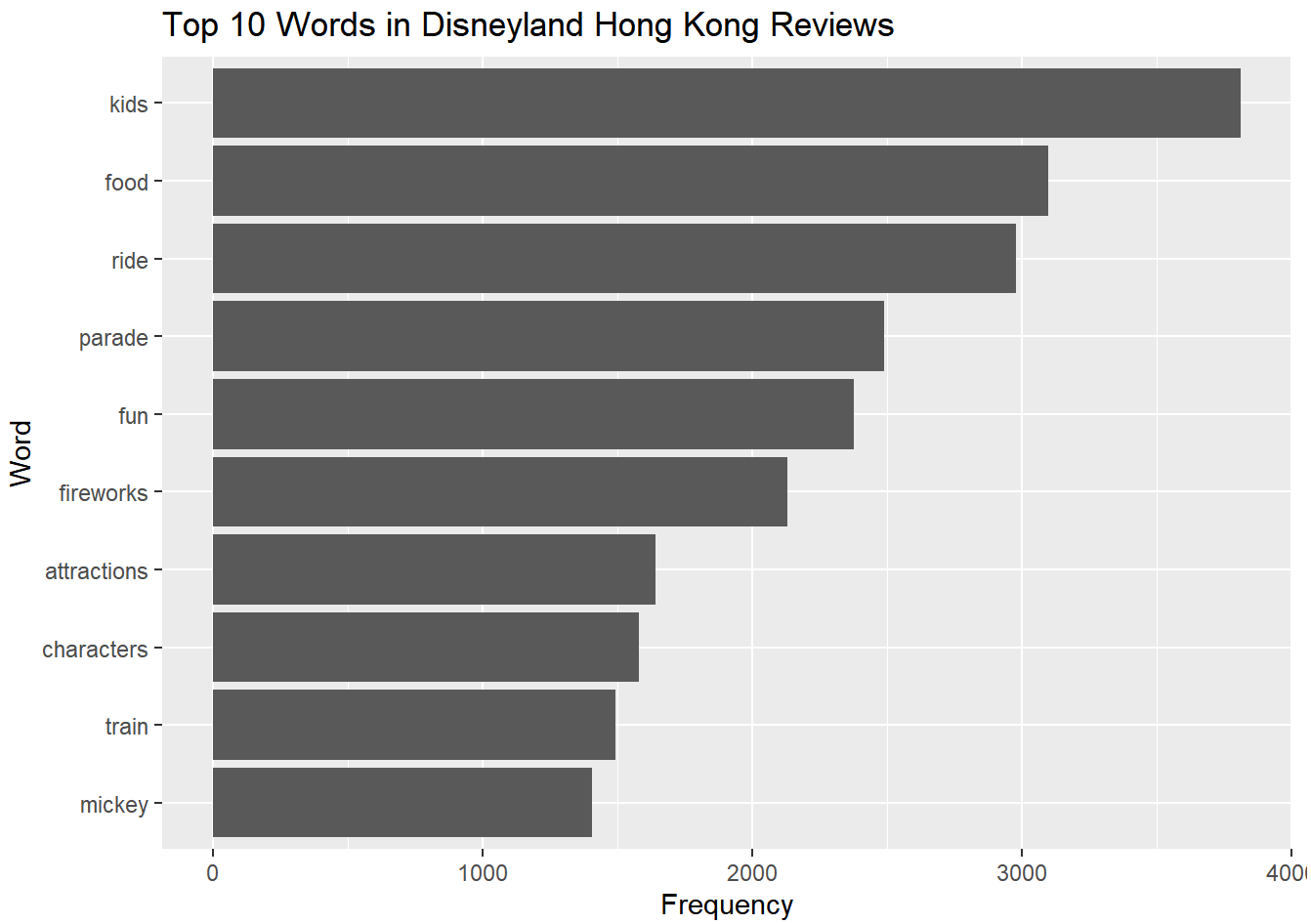
Hong Kong

```
stopshk <- tibble(word = c("disney","disneyland","park","time","day","rides","parks","2","visit",
```

```
hk_reviews <- tidydisney %>%
  anti_join(stopshk)%>%
  filter(Branch == "Disneyland_HongKong") %>%
  count(word, sort = TRUE) %>%
  top_n(10, n)
```

```
Joining with `by = join_by(word)`
```

```
ggplot(data = hk_reviews, aes(y = reorder(word, n), x = n)) +
  geom_col() +
  labs(title = "Top 10 Words in Disneyland Hong Kong Reviews") +
  xlab("Frequency") +
  ylab("Word")
```



While there are some commonalities across all branches, such as 'food' and 'ride' being frequently mentioned, there are also notable differences. For instance, California reviews highlight aspects related to passes, fast access, wait times, and specific attractions. Paris reviews emphasize character interactions and accommodations, suggesting a focus on immersive experiences. Hong Kong reviews showcase a strong emphasis on family-oriented experiences and unique attractions, indicating a focus on entertainment and cultural elements.

Recognizing that the top words by branch might not provide sufficiently insightful results, our approach is going to now shift to implementing TF-IDF (Term Frequency-Inverse Document Frequency) analysis specific to each branch. By tailoring TF-IDF to individual branches, we aim to extract more meaningful and relevant terms, taking into account the distinct characteristics and nuances associated with Disneyland California, Disneyland Hong Kong, and Disneyland Paris.

```
branch_words <- disney %>%
  unnest_tokens(word, Review_Text) %>%
  count(Branch, word, sort = TRUE)
```

```
total_words <- branch_words %>%
  group_by(Branch) %>%
  summarize(total = sum(n))
```

```
branch_words <- left_join(branch_words, total_words)
```

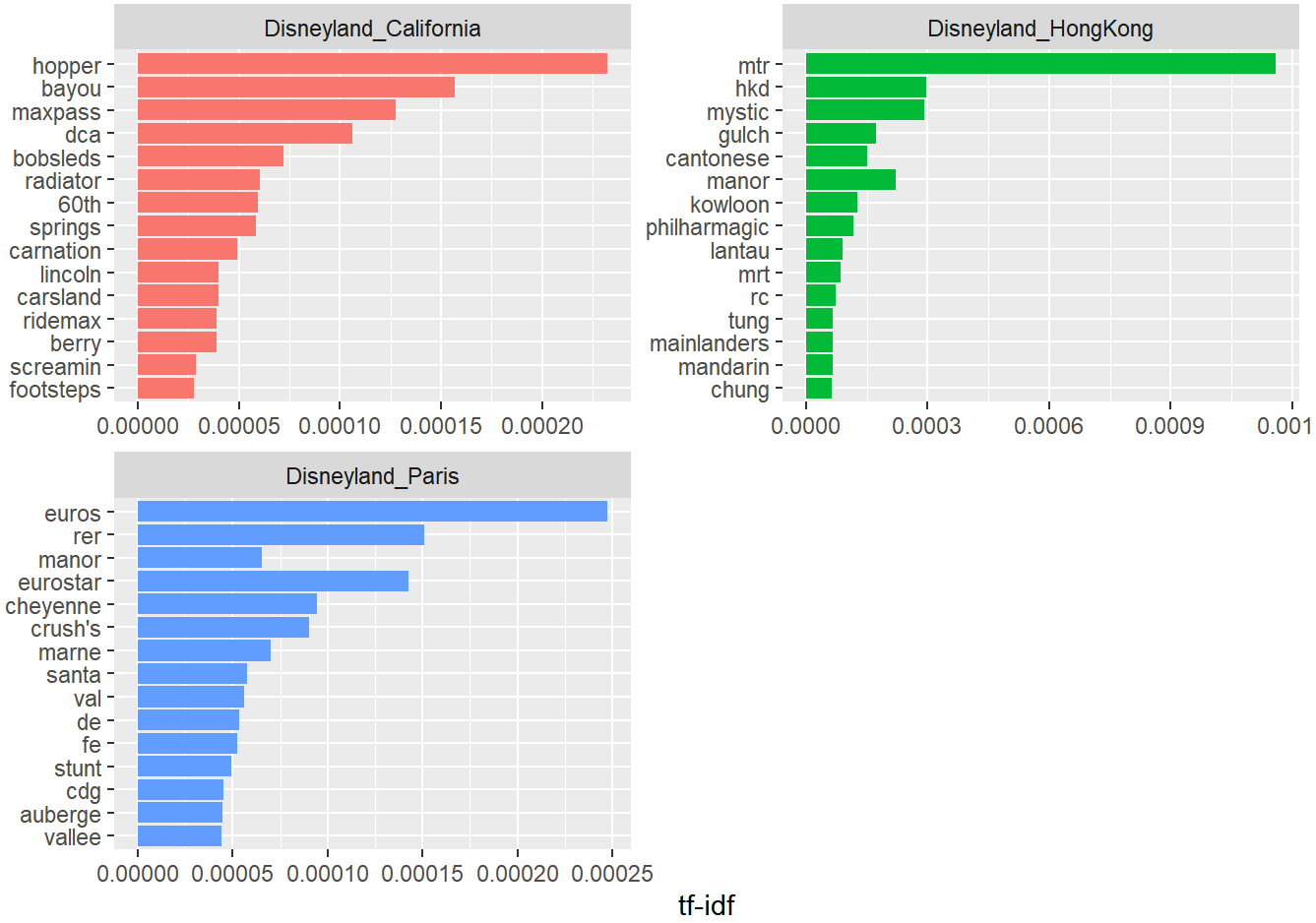
```
Joining with `by = join_by(Branch)`
```

```
disney_tf_idf <- branch_words %>%
  bind_tf_idf(word, Branch, n)
```

```
disney_tf_idf %>%
  select(-total) %>%
  arrange(desc(tf_idf))
```

```
# A tibble: 94,508 × 6
   Branch               word          n       tf   idf   tf_idf
   <chr>                <chr>     <int>    <dbl> <dbl>    <dbl>
 1 Disneyland_HongKong  mtr        1039 0.00106  1.10  0.00116
 2 Disneyland_HongKong  hkd         266 0.000270 1.10  0.000297
 3 Disneyland_HongKong  mystic      711 0.000722 0.405 0.000293
 4 Disneyland_Paris     euros      1411 0.000611 0.405 0.000248
 5 Disneyland_California hopper    1286 0.000573 0.405 0.000233
 6 Disneyland_HongKong  manor       536 0.000544 0.405 0.000221
 7 Disneyland_HongKong  gulch       418 0.000425 0.405 0.000172
 8 Disneyland_California bayou      320 0.000143 1.10  0.000157
 9 Disneyland_Paris     rer         317 0.000137 1.10  0.000151
10 Disneyland_HongKong  cantonese   135 0.000137 1.10  0.000151
# ℹ 94,498 more rows
```

```
disney_tf_idf %>%
  group_by(Branch) %>%
  slice_max(tf_idf, n = 15) %>%
  ungroup() %>%
  ggplot(aes(tf_idf, fct_reorder(word, tf_idf), fill = Branch)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~Branch, ncol = 2, scales = "free") +
  labs(x = "tf-idf", y = NULL)
```

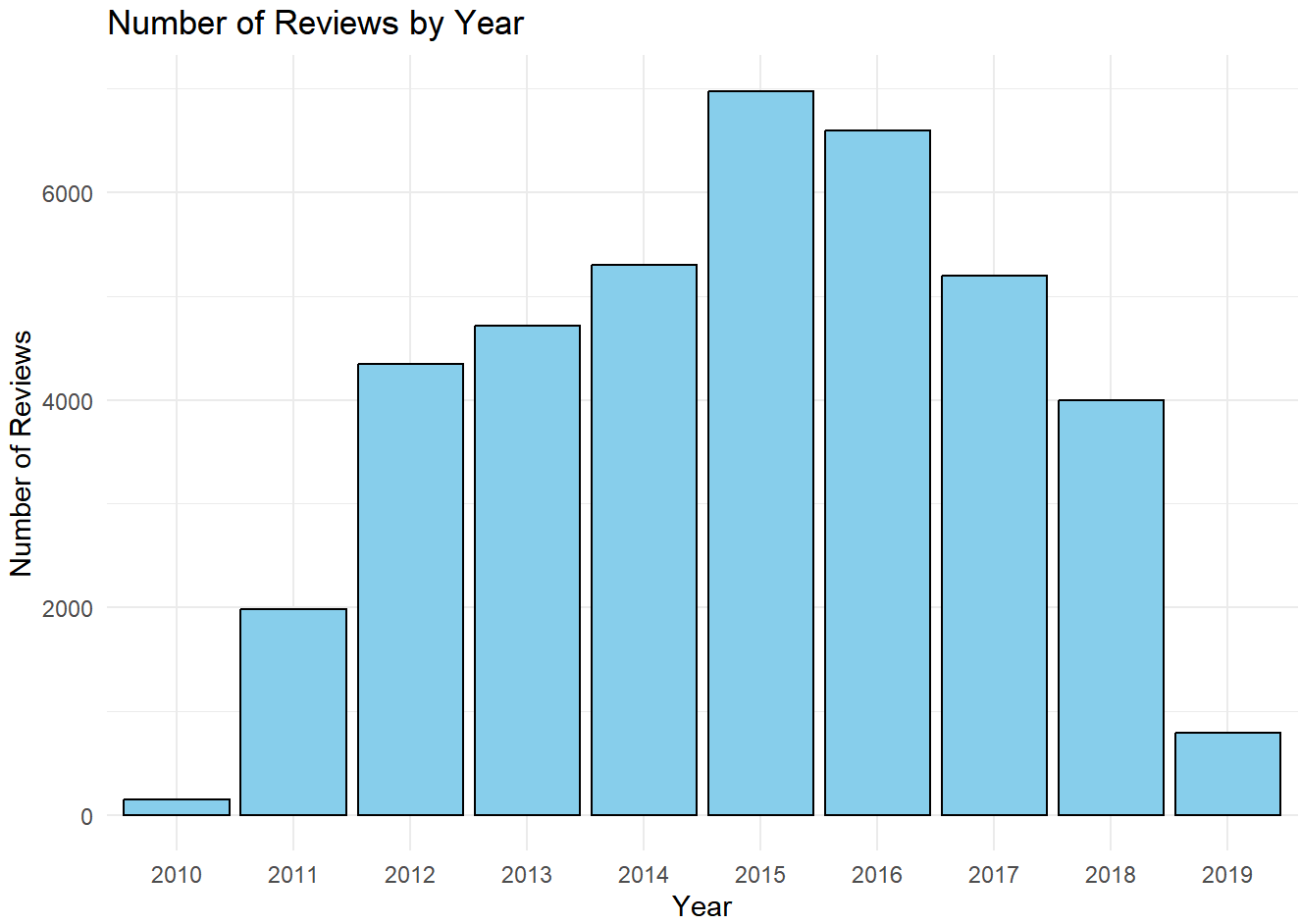


By looking at the words with high tf-idf, things change a little bit. The words I got are significant and specific to each branch, providing insights into the experiences associated with each location. For example, In California, the term “Hopper” is a specific park, while “Bayou” is a restaurant within the park. In Hong Kong, “MTR” is the train to access the park, while “Mystic” and “Gulch” indicate attractions and park areas. For Paris, “RER” and “Eurostar” are trains to access the park, while “Euros” reflects the currency used.

- **Are there any trends or seasonal patterns in the rates?**

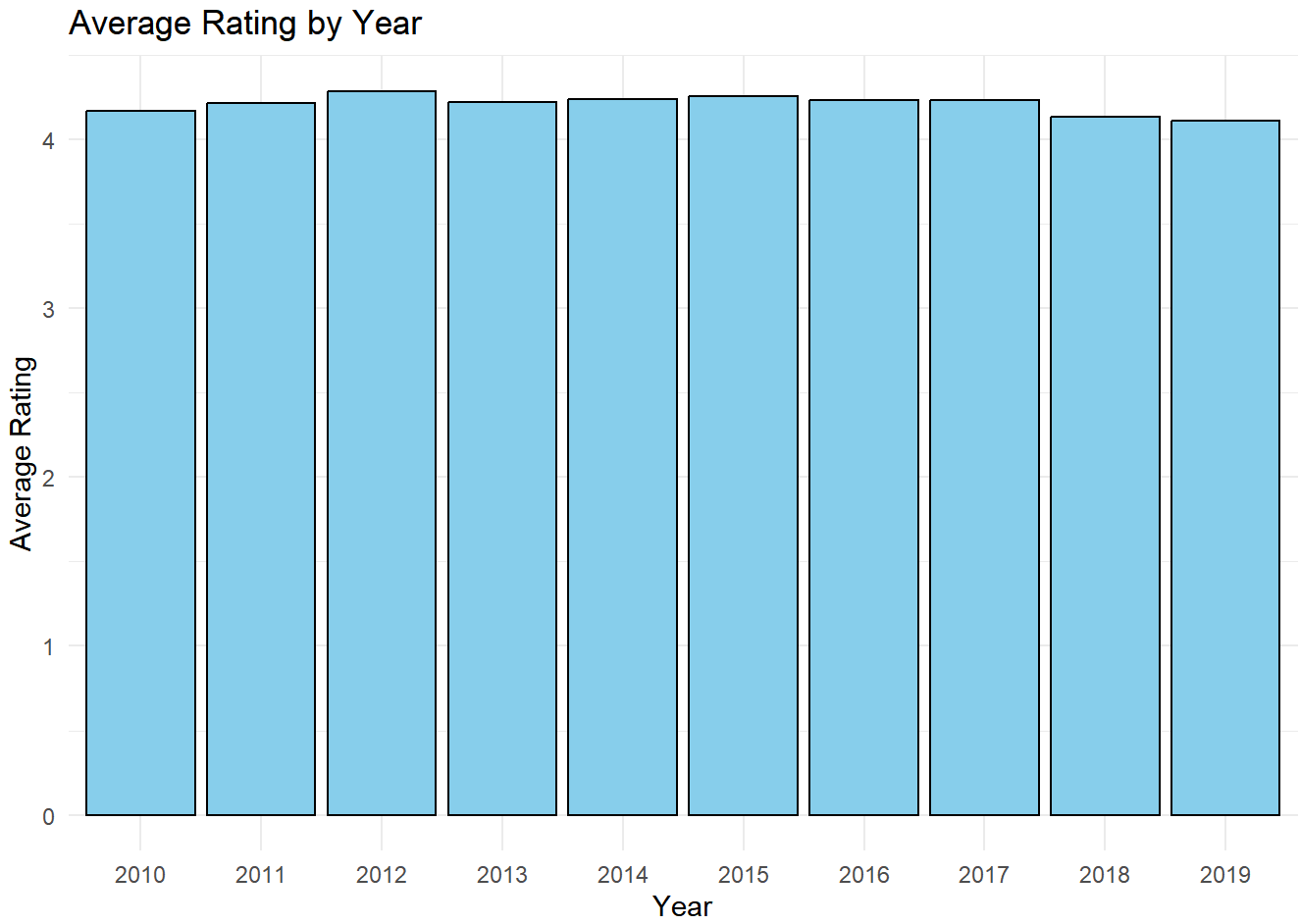
```
tidydata <- disney %>%
  separate(Year_Month, into = c("Year", "Month"), sep = "-")
```

```
tidydata %>%
  filter(Year != "missing") %>%
  group_by(Year) %>%
  summarize(Review_Count = n()) %>%
  ggplot(aes(x = Year, y = Review_Count)) +
    geom_col(fill = "skyblue", color = "black") +
    labs(title = "Number of Reviews by Year",
         x = "Year",
         y = "Number of Reviews") +
    theme_minimal()
```



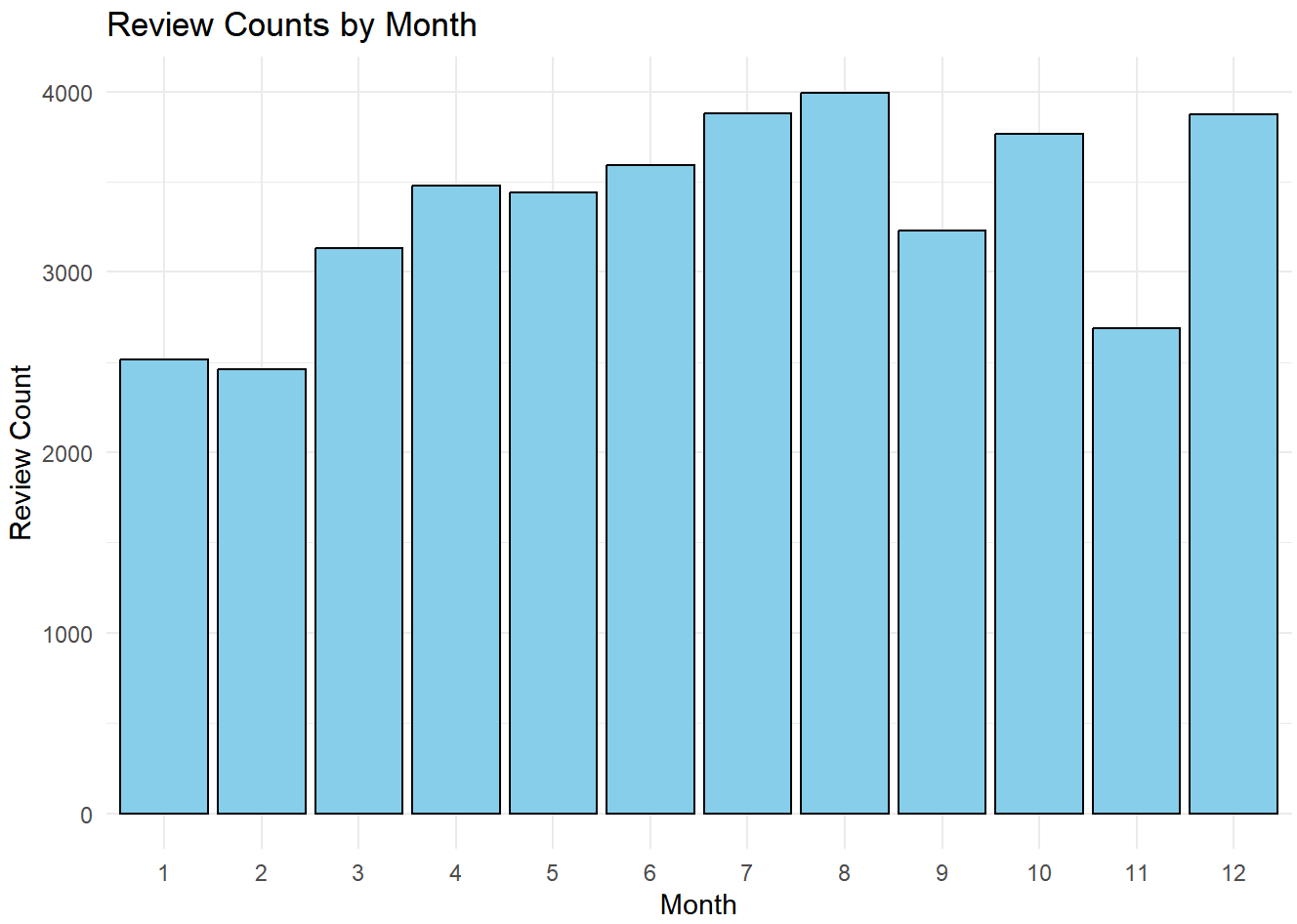
It is evident that the volume of reviews has shown a steady increase over the years, with a notable peak in 2015. This surge in reviews may indicate a heightened level of interest or engagement with Disney experiences during that period.

```
tidydata %>%
  filter(Year != "missing") %>%
  group_by(Year) %>%
  summarize(Average_Rating = mean(Rating)) %>%
  ggplot(aes(x = Year, y = Average_Rating)) +
    geom_col(fill = "skyblue", color = "black") +
    labs(title = "Average Rating by Year",
         x = "Year",
         y = "Average Rating") +
    theme_minimal()
```



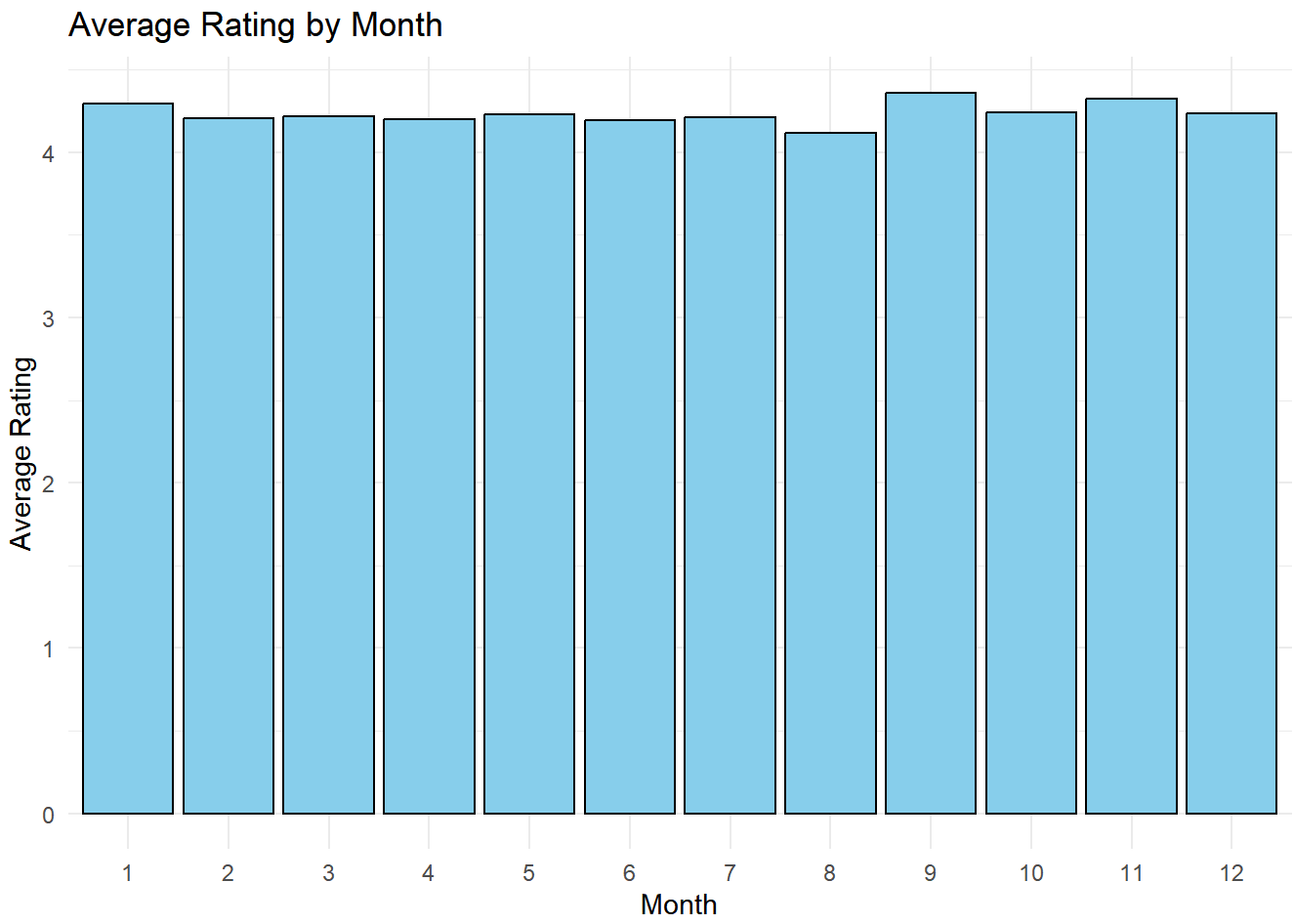
It suggests that there is little variation in customer sentiment over time. This indicates a stable and consistent level of satisfaction with Disney experiences throughout the years covered in the dataset.

```
tidydata %>%
  filter(Year != "missing") %>%
  group_by(Month) %>%
  summarize(Review_Count = n()) %>%
  ggplot(aes(x = reorder(Month, as.numeric(Month)), y = Review_Count)) +
    geom_col(fill = "skyblue", color = "black") +
    labs(title = "Review Counts by Month",
         x = "Month",
         y = "Review Count") +
    theme_minimal()
```



The analysis indicates that October experiences a surge in reviews, likely driven by Halloween festivities. Similarly, December sees a peak in reviews, which aligns with the holiday season, particularly Christmas. Furthermore, the months of June, July, and August stand out, reflecting heightened customer engagement during the summer months. This pattern suggests that seasonal events and holidays play a significant role in influencing customer feedback and experiences at Disney.

```
tidydata %>%
  filter(Year != "missing") %>%
  group_by(Month) %>%
  summarize(Average_Rating = mean(Rating)) %>%
  ggplot(aes(x = reorder(Month, as.numeric(Month)), y = Average_Rating)) +
    geom_col(fill = "skyblue", color = "black") +
    labs (title = "Average Rating by Month",
          x = "Month",
          y = "Average Rating") +
    theme_minimal()
```



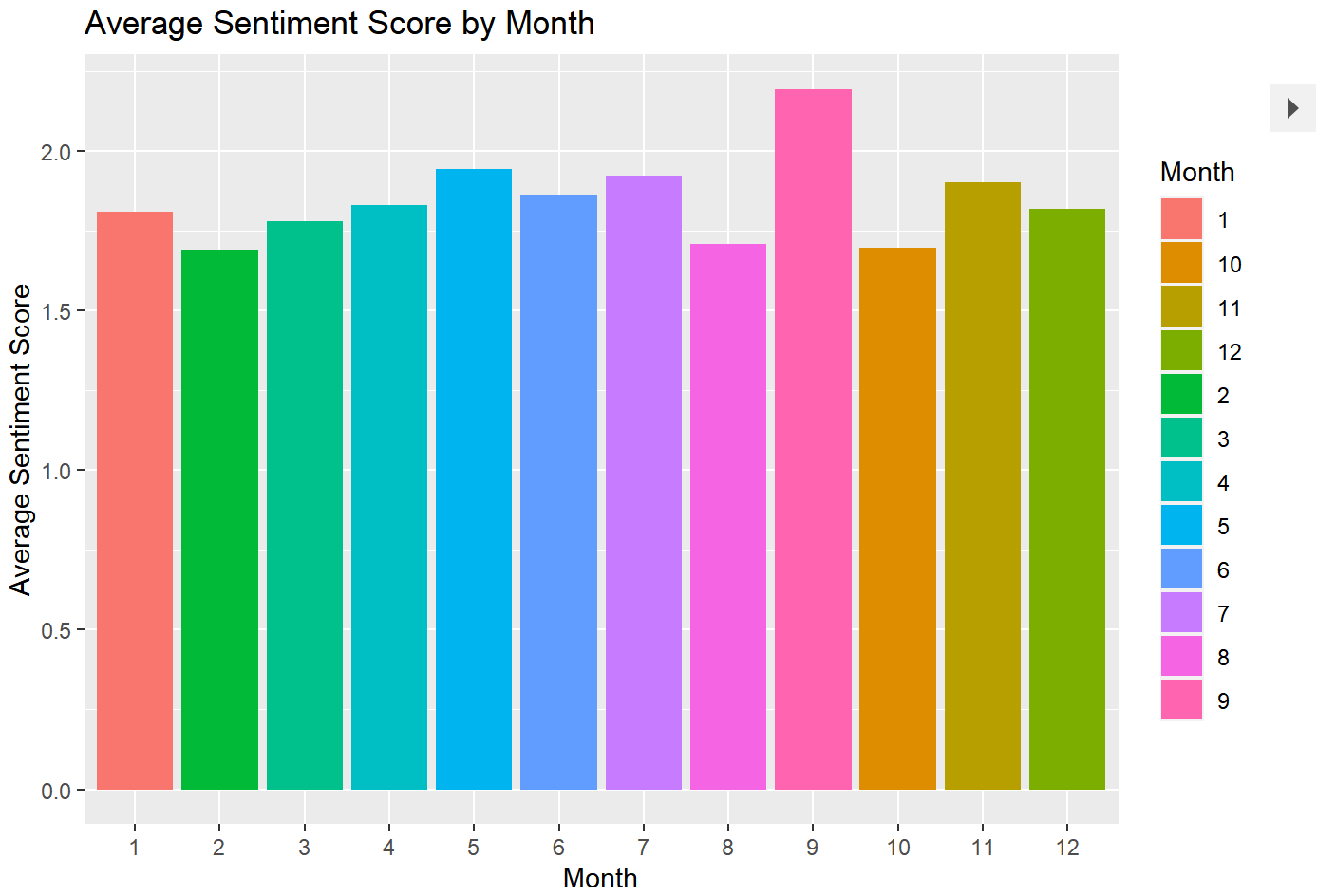
The consistent average ratings across all months indicate a remarkable stability in customer satisfaction levels throughout the year. This suggests that customers' experiences at Disney remain consistently positive, regardless of the season.

- **Do certain times of the year lead to different types of feedback?**

```
tidydisney <- tidydisney %>%
  separate(Year_Month, into = c("Year", "Month"), sep = "-")

sentiment_scores_by_month <- tidydisney %>%
  inner_join(get_sentiments(), relationship = "many-to-many") %>%
  count(Month,Review_ID, sentiment) %>%
  pivot_wider(names_from = sentiment, values_from = n) %>%
  mutate(sentiment_score = positive - negative)

average_sentiment_by_month <- sentiment_scores_by_month %>%
  group_by(Month) %>%
  summarise(mean_sentiment_score1 = mean(sentiment_score, na.rm = TRUE)) %>%
  filter(!is.na(Month))
```

```
ggplot(average_sentiment_by_month, aes(x = reorder(Month, as.numeric(Month)), y = mean_sentiment_
  geom_col() +
  labs(title = "Average Sentiment Score by Month",
```

```
        x = "Month",
        y = "Average Sentiment Score")
```



The observation that January, September, and November have the highest average sentiment scores suggests an interesting pattern. It's possible that this trend is influenced by the months immediately preceding them. Summer months (June, July, August), October, and December are often associated with high attendance at the park due to summer vacations, Halloween, and Christmas, respectively. As a result, visitors may be more inclined to share positive feedback in the subsequent months, as they reflect on their enjoyable experiences.

Another plausible explanation could be that, given that these months (January, September, and November) are typically associated with lower park attendance, visitors who do choose to visit during these times may have a more positive and less crowded experience. The lower crowd levels might contribute to a more enjoyable visit, influencing visitors to share positive feedback. In this scenario, the contrast with the higher attendance months might lead to a higher likelihood of positive sentiment, as visitors appreciate the quieter and more relaxed atmosphere.

LDA

After addressing the questions of interest, the next step involves exploring whether any discernible topics can be identified through Latent Dirichlet Allocation (LDA). This technique allows us to uncover latent themes or patterns within the dataset that might not be immediately apparent. By applying LDA, we aim to

extract meaningful topics from the reviews, providing a more structured and interpretable framework for understanding the content and sentiments expressed by visitors across various aspects of the Disneyland experience.

As a preliminary step, we will implement customized stop words.

```
custom_stopwords1 <- tibble(word = c("disney","disneyland","park","time","day","rides","parks","2

all_stopwords1 <- c(stop_words$word, custom_stopwords1$word)
```

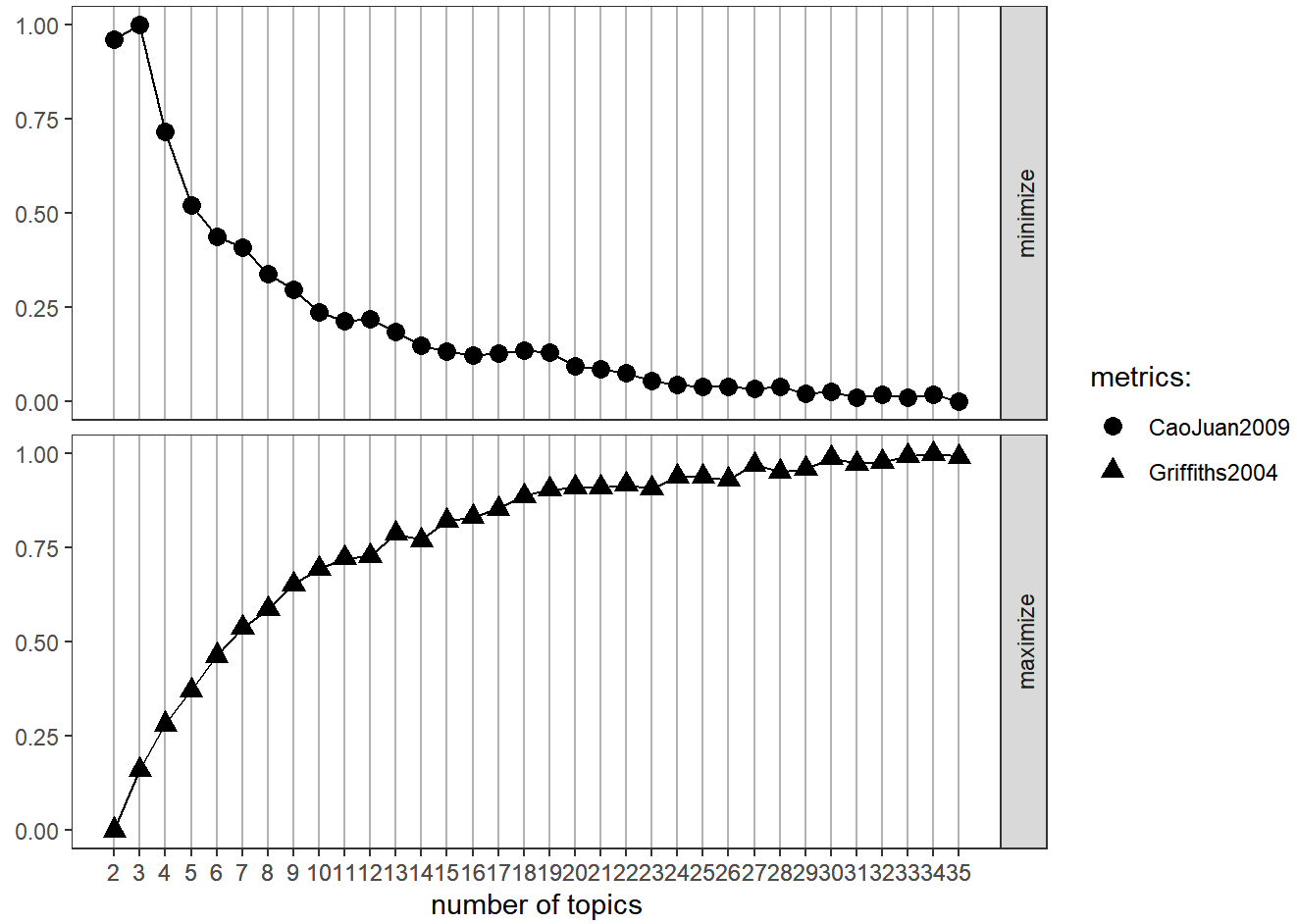
```
branch_reviewid1 <- tidydisney %>%
  filter(!word %in% all_stopwords1) %>%
  mutate(document = paste(Branch, Review_ID, sep = "-")) %>%
  count(document, Review_ID, Branch, word, sort = TRUE)
```

```
word_counts2 <- branch_reviewid1 %>%
  anti_join(stop_words) %>%
  count(document,word,sort=T)
```

```
Joining with `by = join_by(word)`
```

```
disney_dtm1 <- word_counts2 %>%
  cast_dtm(document = document,
           term = word,
           value = n)
```

```
fit models... done.
calculate metrics:
  CaoJuan2009... done.
  Griffiths2004... done.
```



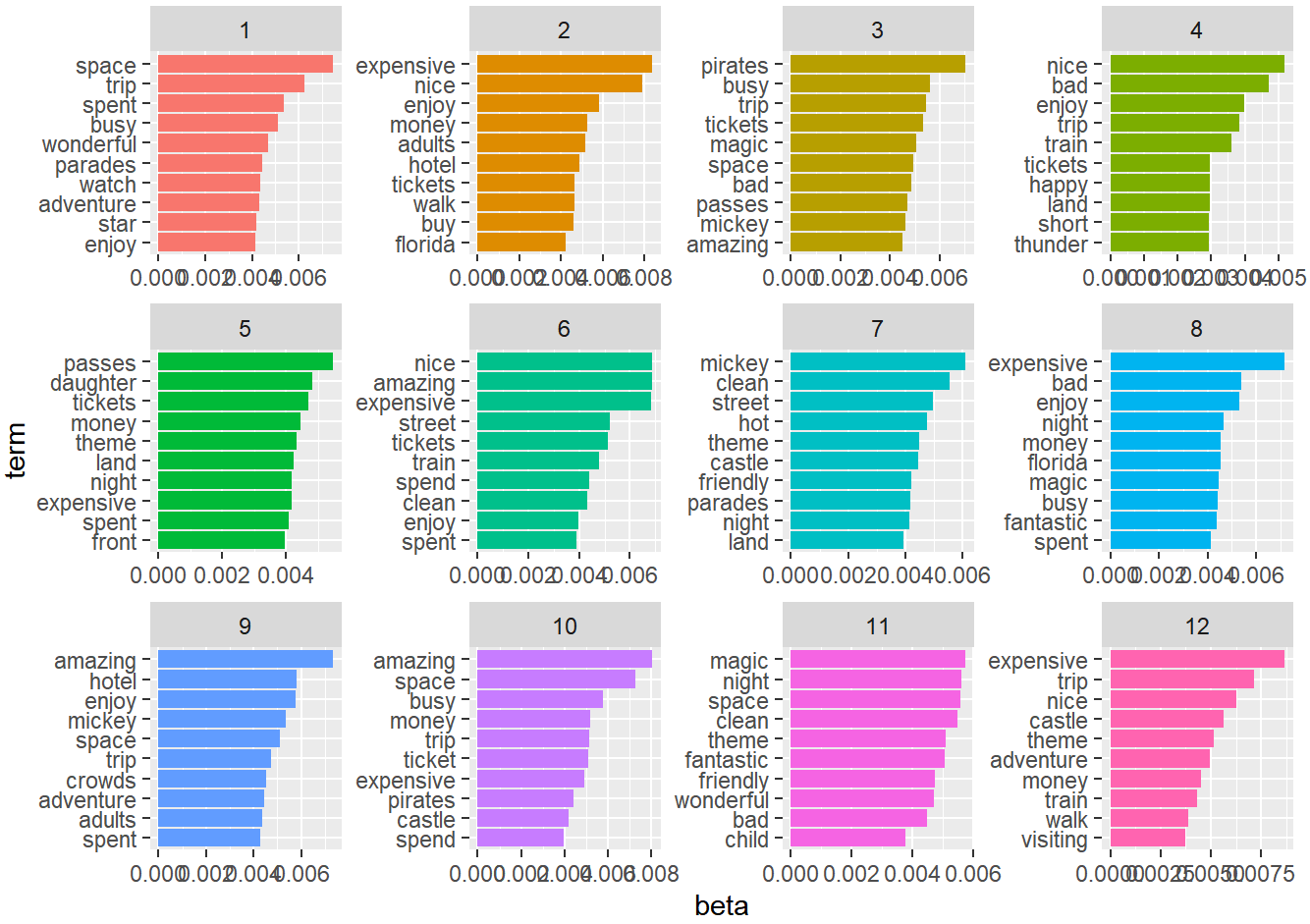
After employing the FindTopicsNumber method, I've chosen to proceed with 24 topics. This decision is informed by the observation that, beyond this point, the line on the graph remains mostly straight, suggesting that significant changes in topic distribution are less apparent.

```
disney_lda1 <- LDA(disney_dtm1, k = 24,
                   control = list(seed = 1234))
```

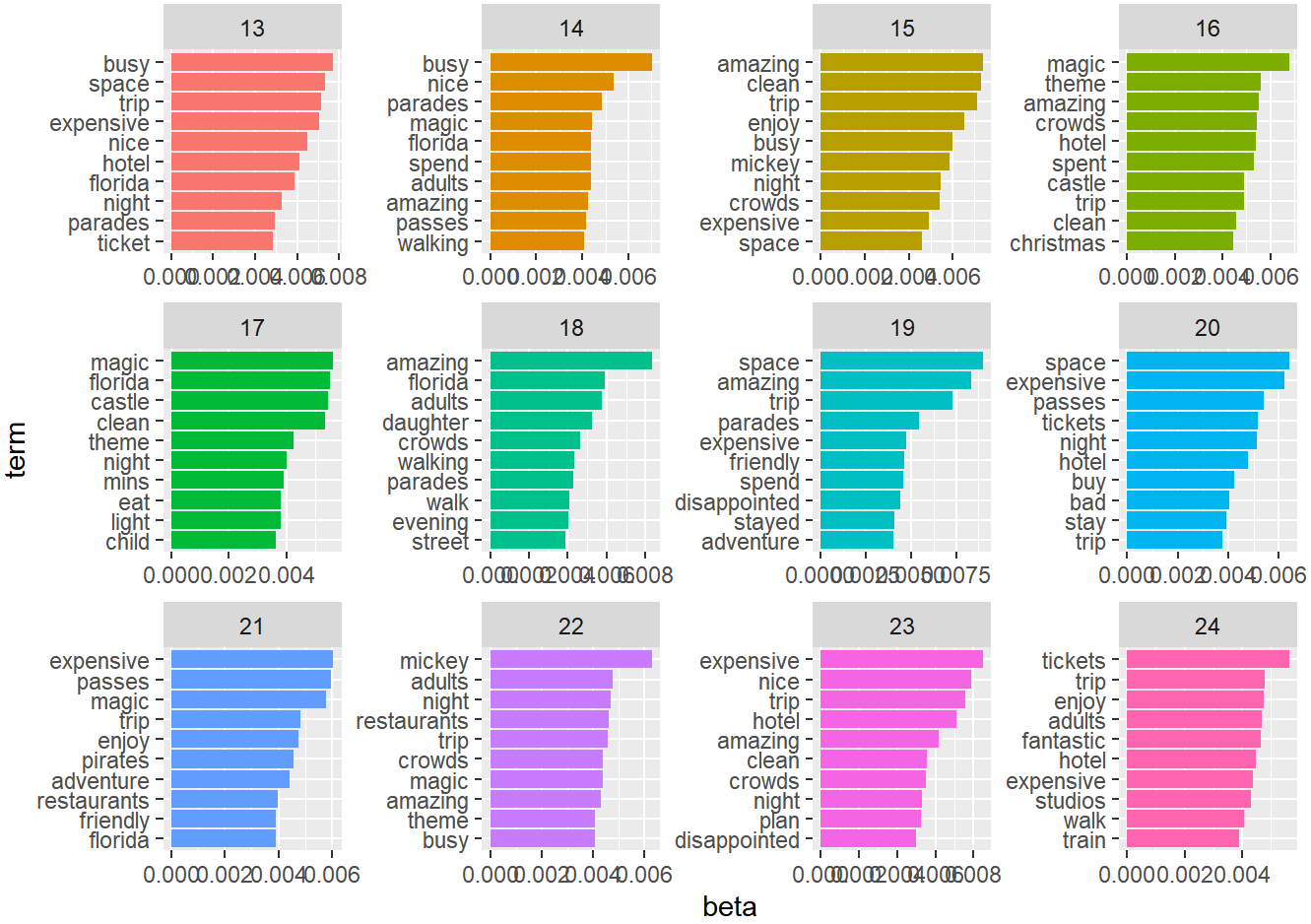
Now, we are going to see which are the 10 most common words in each topic. For enhanced readability, we'll divide our results into two separate charts.

```
disney_topics1 <- tidy(disney_lda1,
                       matrix = "beta")

top_disney_terms1 <- disney_topics1 %>%
  group_by(topic) %>%
  slice_max(beta, n = 10) %>%
  ungroup() %>%
  arrange(topic, -beta)

top_disney_terms1 %>%
  filter(topic < 13) %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(x = beta, y = term, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
```

```
  facet_wrap(~topic  , scales = "free" ) +
  scale_y_reordered()
```




```
top_disney_terms1 %>%
  filter(topic > 12) %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(x = beta, y = term, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~topic  , scales = "free" ) +
  scale_y_reordered()
```



Although the exploration of 24 topics in the LDA modeling process highlighted difficulties in generating coherent and meaningful topics, further experimentation with various topic counts—ranging from smaller to larger numbers—continued to reveal challenges in achieving coherence and meaningful thematic patterns. This consistent struggle across different topic counts suggests a potential underlying similarity among reviews, emphasizing the need to explore alternative modeling approaches to better capture the diverse nuances present in the dataset.

In light of this observation, we are transitioning to another technique to explore whether different modeling approaches can yield more insightful results and better capture the diverse nuances present in the dataset.

KeyATM

In our ongoing quest for more insightful findings, we are going to be employing KeyATM to explore and identify potentially more useful topics within the dataset.

Our initial approach involves seeding three topics: “Hotel,” “Food,” and “Park” as key themes.

```
disney_keywords <- list("Hotel" = c("hotel","accomodation","night"),
                        "Food" = c("restaurant","food","drink","restaurants"),
                        "Park" = c("park","parks","attractions","queues","queue","line","lines"))
```

```
tidydisney1 <- tidydisney %>%
  rename(text=word)%>%
```

```
  filter(text != "")
```

Here we are preparing our Disney-related text data for analysis with the keyATM package:

```
KeyATM_docs2 <- keyATM_read(tidydisney1)
```

```
ℹ Using tibble.
```

Subsequently, we are going to delegate the selection of the remaining two topics to KeyATM, allowing the algorithm to autonomously identify and assign relevant themes based on the intrinsic patterns within the dataset. This combination of predefined and algorithmically determined topics aims to strike a balance between specific areas of interest and the discovery of latent themes that might be less apparent in the initial selection.

To save time, the following KeyATM model named `out_disney` was previously created in another document with the specified parameters and saved as an RDS file.

```
{r}
out_disney <- invisible({keyATM(docs = KeyATM_docs2,
              model = "base",
              no_keyword_topics = 2, #number of additional
                                     #topics to model
              keywords = disney_keywords)
})
```

Now, I am loading this precomputed model from the RDS file using the `readRDS` function and storing the result in the `out_disney` variable.

```
out_disney <- readRDS("C:/Users/maria/OneDrive - University of Denver/Desktop/COMPLEX DATA ANALYT
```

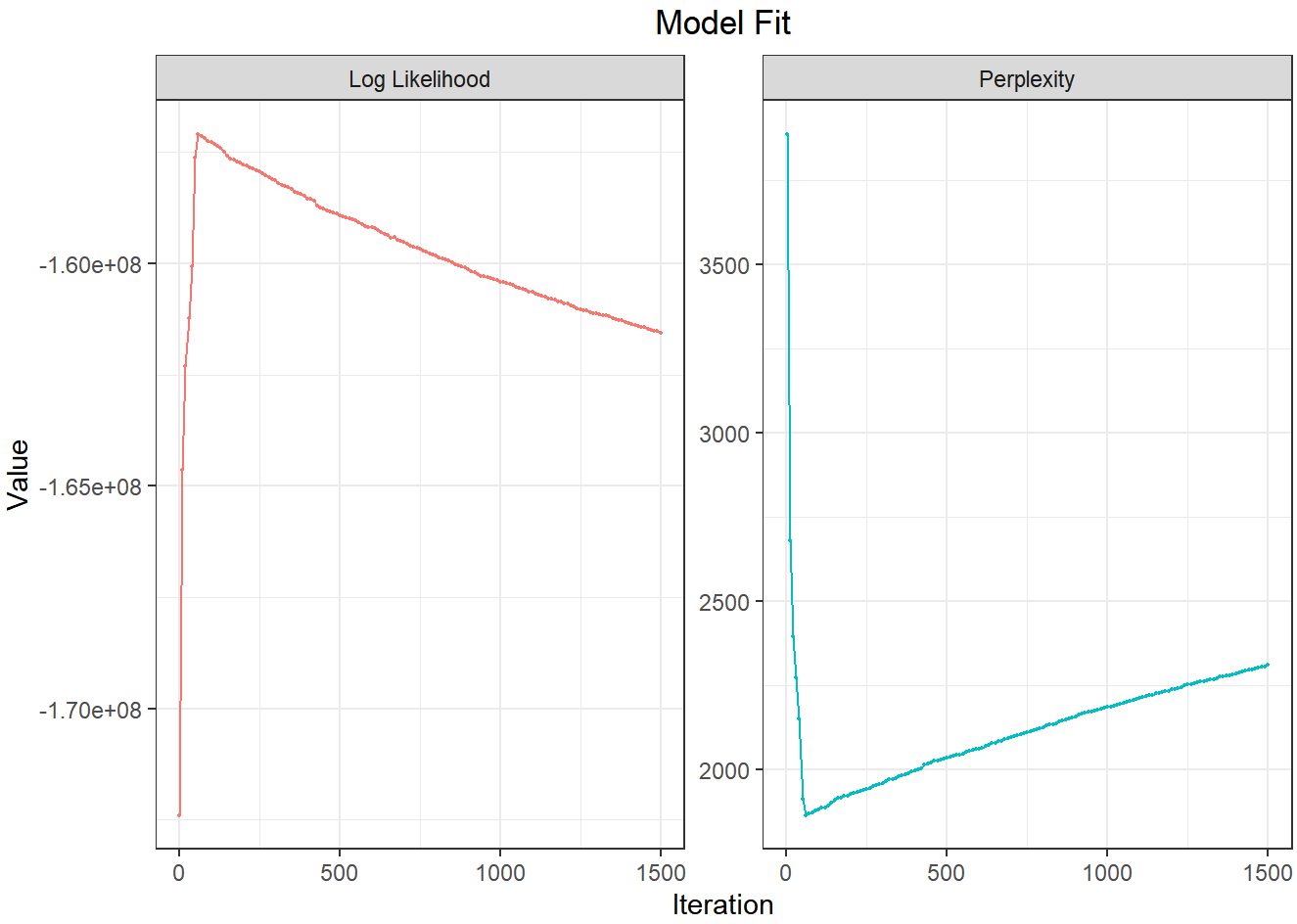
```
saveRDS(out_disney, file = "out_disney.rds")
```

```
top_words(out_disney)
```

	1_Hotel	2_Food	3_Park	Other_1	Other_2
1	fighting	food [✓]	park [✓]	edge	sweeping
2	ops	restaurant [✓]	disney	spanish	greatthe
3	caves	drink [✓]	rides	perfection	recommending
4	emergency	tomorrowland	disneyland	cheesy	showthe
5	effectively	total	time	showers	meticulous
6	pathetic	location	day	soaking	mature
7	spooky	boy	ride	cared	harbour
8	youngsters	polite	kids	kudos	prevented
9	ashamed	sat	visit	toystory	slipping
10	seaworld	entertaining	people	naps	dodge

Despite the initial effort to seed three topics ("Hotel," "Food," and "Park") and allow KeyATM to determine the remaining two, it appears challenging to interpret or distinguish these topics based on the output or analysis performed. Further investigation and exploration may be necessary to refine the topic modeling parameters or adjust the approach to improve the interpretability and relevance of the identified topics.

```
plot_modelfit(out_disney)
```



The initial increase in log likelihood and decrease in perplexity suggest that the model is learning and improving its representation of the training data.

The decrease in log likelihood and increase in perplexity may indicate overfitting, where the model becomes too specialized to the training data and loses generalization ability.

Conclusion

In conclusion, our exploration and analysis of Disneyland reviews across its branches in California, Hong Kong, and Paris have unveiled valuable insights into the multifaceted world of Disney experiences.

Branch-specific analyses highlighted unique characteristics. Despite Hong Kong having the fewest reviews, it garnered the highest sentiment score, emphasizing the quality of experiences shared by its visitors. Our exploration into stop words and sentiment scores shed light on visitors' priorities and sentiments across various aspects of the parks.

The temporal dimension revealed intriguing patterns. Peaks in sentiment scores during January, September, and November do not necessarily align with high attendance periods but indicate interesting patterns in visitor sentiments. While Summer months, October, and December are typically associated with higher park attendance due to summer vacations, Halloween, and Christmas, respectively, the sentiment scores of the succeed months suggest positive reflections on experiences without a direct correlation to attendance figures. These findings open avenues for exploring the unique factors contributing to positive sentiments during these specific months.

The application of LDA and KeyATM, unfortunately, did not yield insightful or cohesive results. One plausible explanation for this outcome is the possibility that overall reviews across the Disneyland branches share a substantial degree of similarity in their themes and content. It appears that the diversity in topics might be limited, contributing to challenges in the convergence and effectiveness of these topic modeling techniques. This observation suggests that the overarching narratives in Disneyland reviews may not exhibit distinct topical patterns that these models could effectively capture.

This project exemplifies the power of data analytics in unraveling the complexities of user-generated content, offering actionable insights for park management and contributing to a nuanced understanding of the dynamic interplay between visitor sentiments, thematic preferences, and the enchanting world of Disneyland.