TOP-SELLLING BOOKS

SUMMARY

In this study, the change in the price distribution of the top 50 best-selling books on Amazon between 2009 and 2019, the change in the distribution of reader ratings of the best-selling books over the years, and the price distribution of books by genre were investigated.

DATA INTRODUCTION

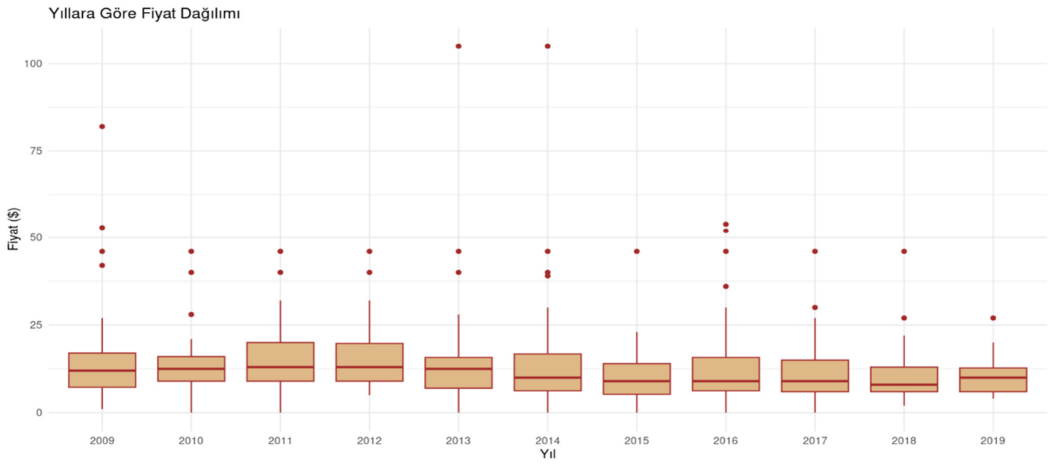
The dataset used in the analysis is the Amazon Top 50 Bestselling Books 2009 - 2019 dataset obtained from Kaggle. The dataset contains 550 books, and the data is categorized as fiction and non-fiction using Goodreads. User rating, reviews, price, year and genre variables were used in the study.

RESEARCH QUESTIONS

RESEARCH QUESTION 1: EXPLORE THE VARITAION IN PRICE DISTRIBUTION OF TOP-SELLING BOOKS OVER THE YEARS.

```
data <- bestsellers_with_categories

install.packages("ggplot2")
install.packages("tidyr")
install.packages("diplyr")
library(tidyr)
library(diplyr)
library(ggplot2)
```

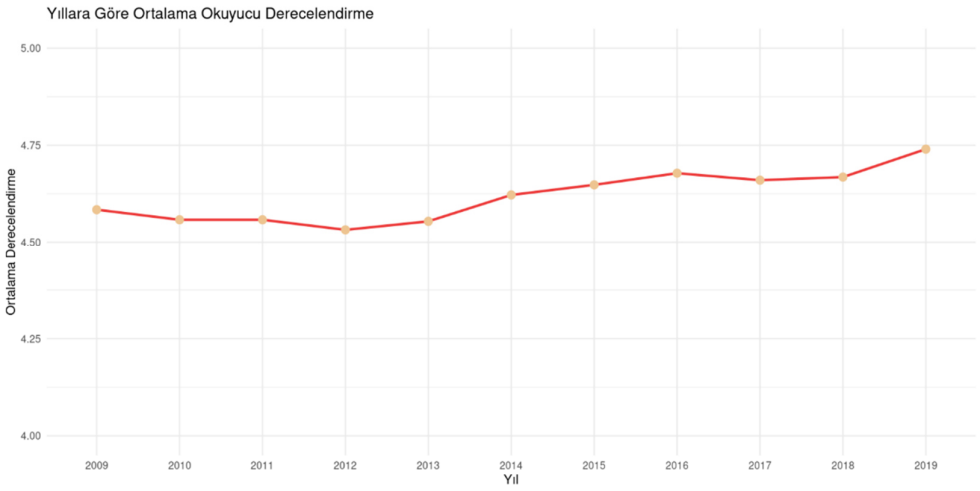
```
#• Explore the variation in price distribution of top-selling books over the years.
ggplot(data, aes(x = as.factor(Year), y = Price)) +
  geom_boxplot(fill = "burlywood", color = "brown") +
  labs(
    title = "Yıllara Göre Fiyat Dağılımı",
    x = "Yıl",
    y = "Fiyat ($)"
  ) +
  theme_minimal()
```



Using the above codes, we obtained the price distribution of books by year, and we displayed it with a box plot. According to this graph, prices generally vary between $10 and $15 over the years. There are outliers in each year, but outliers are more frequent in 2010, 2011, 2016 and 2018. The average price distribution seems to remain similar in each year. The graph shows that the price range is wider in 2014, 2015 and 2016.

Research Question 2: Investigate the change in reader rating distribution of top-selling books over the years.

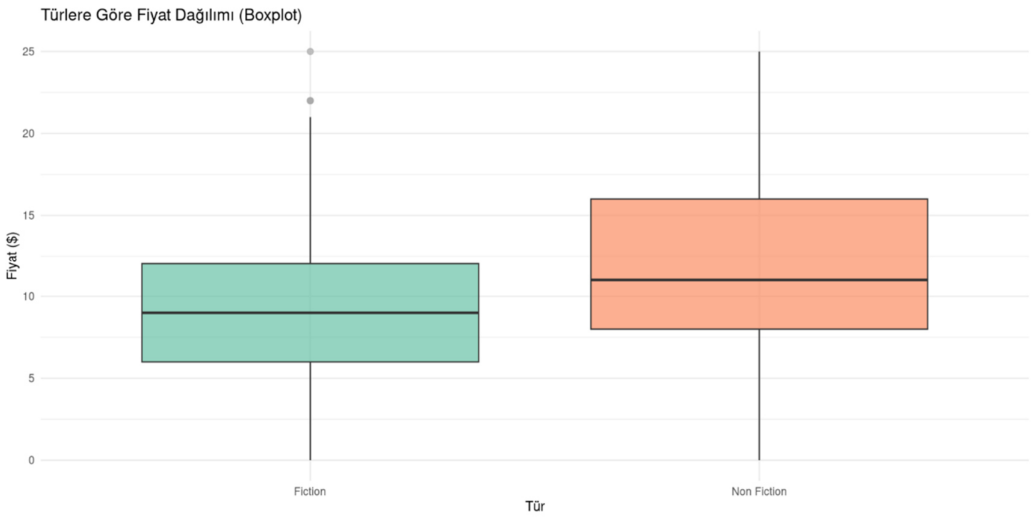
```
# Ortalama derecelendirme hesapla
avg_rating <- data %>%
  group_by(Year) %>%
  summarize(Average_Rating = mean(`User Rating`, na.rm = TRUE))

# Çizgi Grafiği
ggplot(avg_rating, aes(x = as.factor(Year), y = Average_Rating)) +
  geom_line(group = 1, color = "brown2", size = 1) +
  geom_point(color = "burlywood2", size = 3) +
  scale_y_continuous(expand = c(0, 0), limits = c(0, 5)) +
  labs(
    title = "Yıllara Göre Ortalama Okuyucu Derecelendirme",
    x = "Yıl",
    y = "Ortalama Derecelendirme"
  ) +
  theme_minimal() +
  ylim(4,5)
```



Using the codes above, we obtained the distribution of average ratings of readers for books by year, and we showed it with a line graph. According to this graph, the average rating in the period 2009-2013 is flat between 4.50 and 4.55, then an increase is seen starting from 2013. The average rating increases significantly until 2015. A slowdown in the rate of increase is seen in the period 2016-2018, but a slight increase still continues. The highest average rating is observed in 2019.

RESEARCH QUESION 3: ANALYZE THE PRICE DISTRIBUTION OF BOOKS BY GENRE.

```
#• Analyze the price distribution of books by genre.

ggplot(data, aes(x = Genre, y = Price, fill = Genre)) +
  geom_boxplot(alpha = 0.7, outlier.color = "darkgrey", outlier.size = 2) +
  labs(
    title = "Türlere Göre Fiyat Dağılımı (Boxplot)",
    x = "Tür",
    y = "Fiyat ($)"
  ) +
  theme_minimal() +
  theme(legend.position = "none") +
  scale_fill_brewer(palette = "Set2") +
  ylim(0,25)
```



Using the codes above, we obtained a box plot of the price distribution by book type. According to this graph, the price of fiction books is between $5 and $12.5, while the price of non-fiction books is between $7.5 and $17.5. The average price of fiction books is approximately $8.5, while the average price of non-fiction books is approximately $11. According to this graph, we can say that the prices of non-fiction books are slightly higher than fiction books.

FRIENDS TV SERIES

SUMMARY

This study aims to analyze the distribution of IMDb ratings and votes of the TV series "Friends" by season. The research focuses on determining which seasons have the most skewed rating distribution and the most variability.

DATA INTRODUCTION

The dataset used in the analysis is the Friends Episode Data dataset obtained from Kaggle. The dataset includes IMDb ratings, vote counts, season and episode information for each episode of the series. Season, rating, votes variables were used in this study.

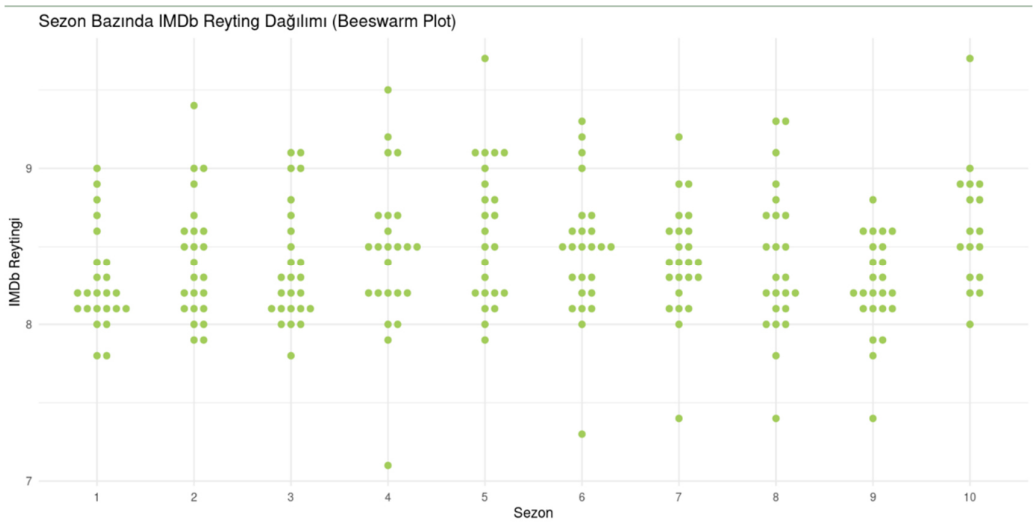
RESEARCH QUESTIONS

RESEARCH QUESTION 1: INVESTIGATE THE DISTRIBUTION OF IMDB RATINGS BY SEASON. IDENTIFY THE SEASON WITH THE MOST SKEWED RATING DISTRIBUTION.

```
if (!requireNamespace("ggbeeswarm")) install.packages("ggbeeswarm")
library(ggbeeswarm)

# Beeswarm plot
ggplot(ratings_by_season, aes(x = as.factor(season), y = imdb_rating)) +
  geom_beeswarm(color = "darkolivegreen3", size = 2) +
  theme_minimal() +
  labs(title = "Sezon Bazında IMDb Reyting Dağılımı (Beeswarm Plot)",
       x = "Sezon",
       y = "IMDb Reytingi")
```

```
install.packages("e1071")
library(e1071)

# Çarpıklık hesaplama
skewness_values <- ratings_by_season %>%
  group_by(season) %>%
  summarize(Skewness = skewness(imdb_rating, na.rm = TRUE))

# En çarpık sezonu belirleme
most_skewed_season <- skewness_values %>%
  filter(Skewness == max(Skewness))
print(most_skewed_season)
```

```
# A tibble: 1 × 2
  season Skewness
   <dbl>    <dbl>
1     10    0.776
>
```



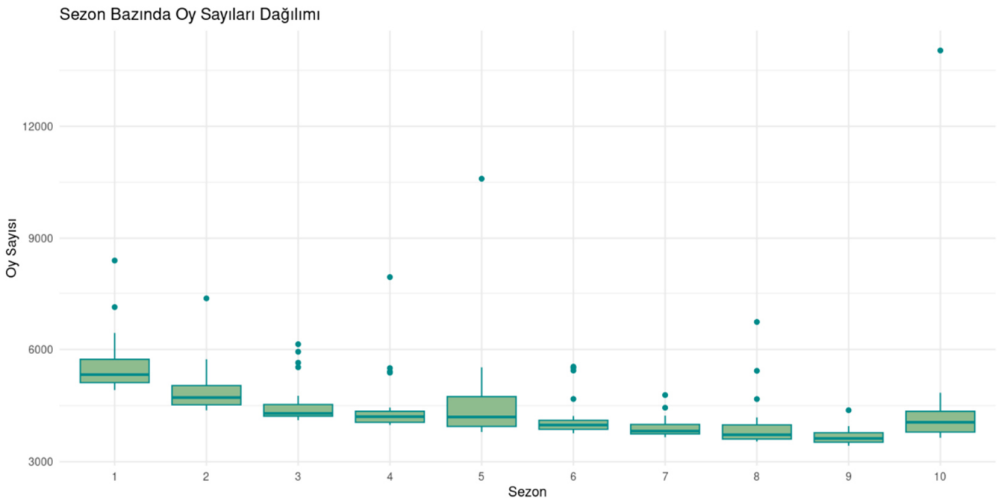
Using the codes above, we obtained the beeswarm plot of the Friends series' IMDB rating distribution by season and which season had a more skewed distribution. According to this graph, the ratings vary between 7 and 9, which shows that the series is generally liked. In the 1st, 2nd and 3rd seasons, the ratings are generally concentrated between 8 and 9, and the episodes have almost the same ratings. In the 4th and 5th seasons, some episodes have received ratings above 9, while in the 6th and 7th seasons, there is a more scattered rating, and it is observed that some episodes have fallen below 8. In the 8th, 9th and 10th seasons, as in the first three seasons, the IMDB ratings are between 8 and 9.

Research Question 2: Examine the distribution of vote counts by season. Determine the season with the highest variation in vote counts.

```
votes_by_season <- friends %>%
  select(season, total_votes)

# Sezon bazında standart sapma hesaplama
variation_by_season <- votes_by_season %>%
  group_by(season) %>%
  summarize(StandardDeviation = sd(total_votes, na.rm = TRUE))

# En değişken sezonu bulma
most_variable_season <- variation_by_season %>%
  filter(StandardDeviation == max(StandardDeviation))
print(most_variable_season)
```

```
ggplot(votes_by_season, aes(x = as.factor(season), y = total_votes)) +
  geom_boxplot(fill = "darkseagreen", color = "cyan4") +
  theme_minimal() +
  labs(title = "Sezon Bazında Oy Sayıları Dağılımı",
       x = "Sezon",
       y = "Oy Sayısı")
```

```
# A tibble: 1 × 2
  season StandardDeviation
   <dbl>             <dbl>
1     10             2443.
```



Using the codes above, we created a box plot of the distribution of votes for the Friends series by season and determined the season with the highest change in vote counts. According to this graph, it is observed that the vote changes in seasons 1, 5 and 10 are higher than the others, yet the highest average vote is received by season 1. While there are outliers in every season, it is observed that the outlier value of season 10 is high at 12000.