T.R

ESKISEHIR TECHNICAL UNIVERSITY

FACULTY OF SCIENCE

DEPARTMENT OF STATSTICS

HOMEWORK #4

Name and Surname: Mustafa Kemal Karaman

ID Number: 28747425986

big_bang_theory_compiled.csv

Summary of the Dataset

This dataset consists of episode-level information across 12 seasons of The Big Bang Theory. Below are the key features and their relevance:

Episode Metadata:

- **Season and Episode Numbers:** Essential for chronological analyses and observing trends across seasons.
- **Titles:** Useful for identifying thematic patterns or notable episodes.

Production Details:

- **Directors and Writers:** Enable frequency and qualitative analysis of creative contributors to the show.
- **Production Codes**: Provide insights into the filming and scheduling process.

Broadcast Information:

- **Original Air Date:** Allows temporal analysis of viewership trends.
- **US Viewers (in millions):** A quantitative metric reflecting the show's popularity and reception.

This dataset provides ample opportunities for data summaries, such as:

- **Frequency Analysis:** Commonly used to determine the most frequent directors or prolific writers.
- **Quantitative Summaries:** Metrics like mean, median, and quantiles of viewership to understand audience trends.
- **Time-Series Analysis:** How viewership fluctuated across the show's lifespan.

Leveraging these insights, one can explore relationships between creative decisions and audience responses, investigate patterns of success, or highlight milestone episodes in the series.

Data Introduction

The dataset contains information about all episodes of The Big Bang Theory, one of the most popular sitcoms in television history. It includes both quantitative and qualitative variables, providing a detailed perspective on the show's production, broadcasting, and viewership trends.

This dataset is particularly suited for summarization and exploratory data analysis due to its well-structured content. Key aspects include categorical variables such as episode titles, production details, and directors, alongside numerical variables like viewership statistics. By summarizing and visualizing these components, we can gain insights into patterns like audience engagement, thematic evolution, and contributions from key creators over time.

The dataset serves as a foundation for various statistical summaries, such as frequency counts of directors, distributions of viewership data, and time-series trends across seasons. These analyses provide a gateway to understanding the show's impact and its relationship with the audience.

Research Questions of big_bang_theory_compiled.csv

1) Analyze the distribution of IMDB ratings by season. Identify the season with the most skewed rating distribution.

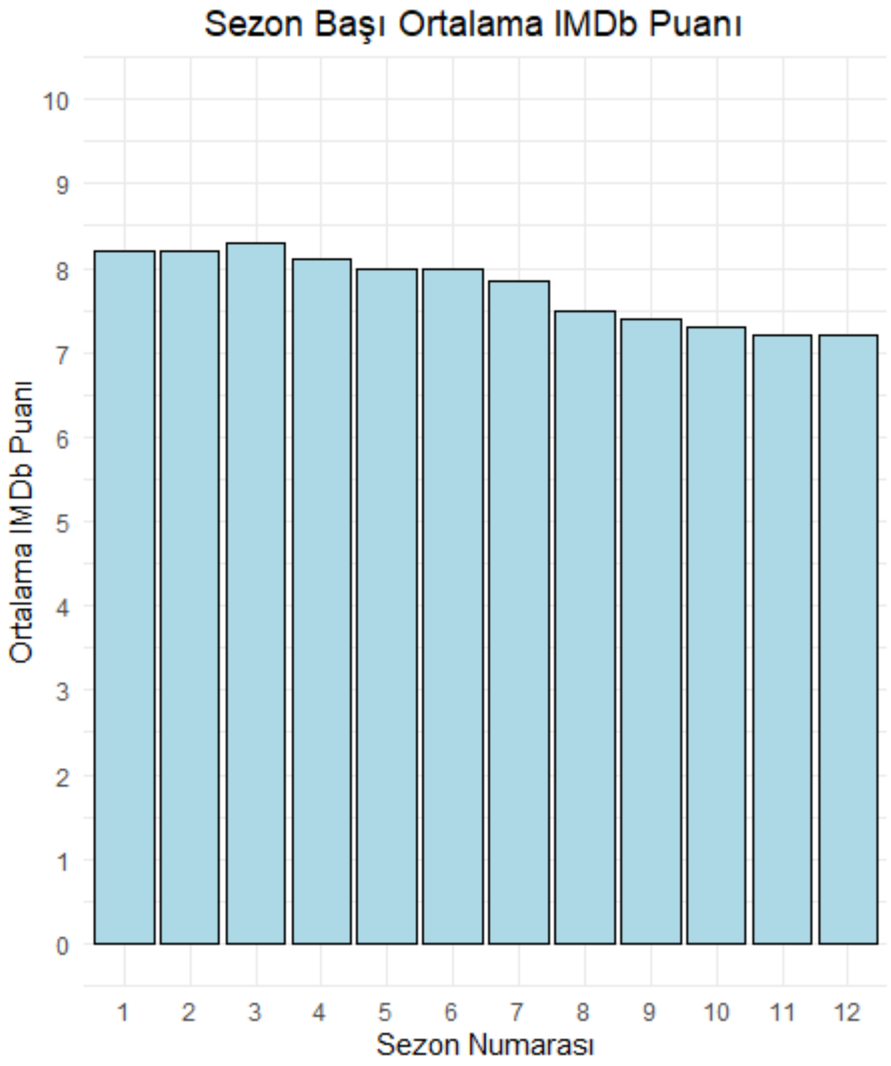
Analyzing the distribution of IMDB ratings by season is necessary to understand how audience reception varies over time. By identifying the season with the most skewed rating distribution, we can uncover trends such as declining quality, standout episodes, or polarizing content that divides viewers. This analysis helps pinpoint specific seasons that significantly impact the overall success or failure of a series, offering valuable insights for producers, critics, and viewers.

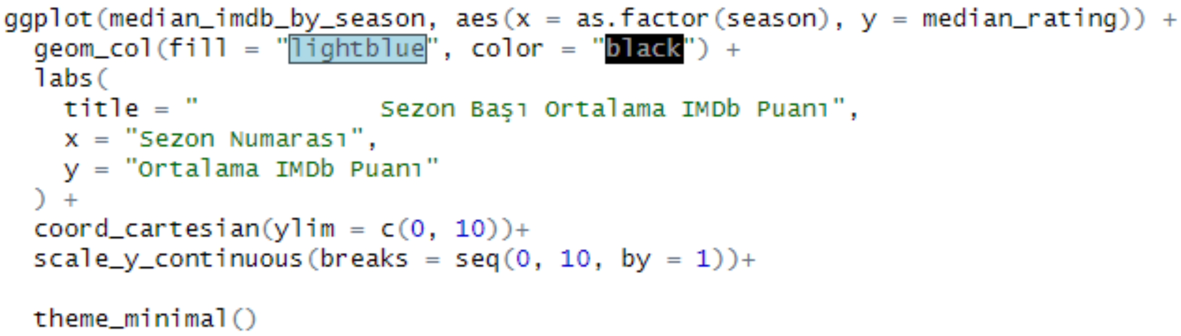
2) Investigate the distribution of vote counts by season. Determine the season with the highest differences in vote counts

Investigating the distribution of vote counts by season is essential to understand audience engagement and interest over time. Determining the season with the highest differences in vote counts can reveal key patterns, such as episodes that attracted unusually high attention or seasons where interest fluctuated dramatically. This analysis helps identify factors driving viewer participation, such as standout episodes, marketing efforts, or shifts in the show's popularity.

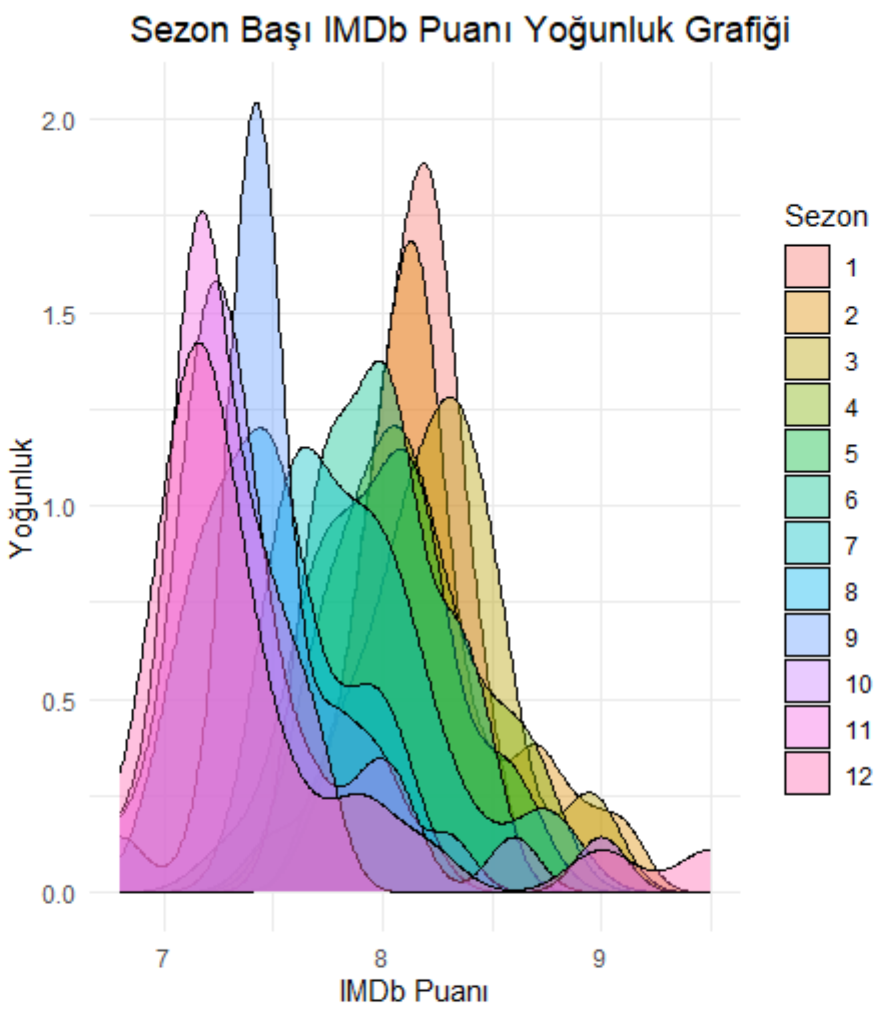
Analysis

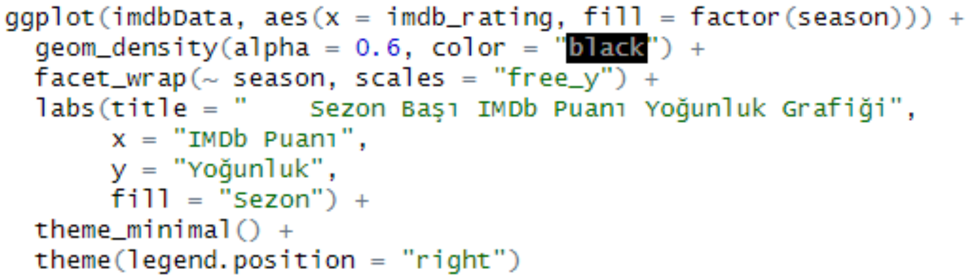
Analyze the distribution of IMDB ratings by season. Identify the season with the most skewed rating distribution.



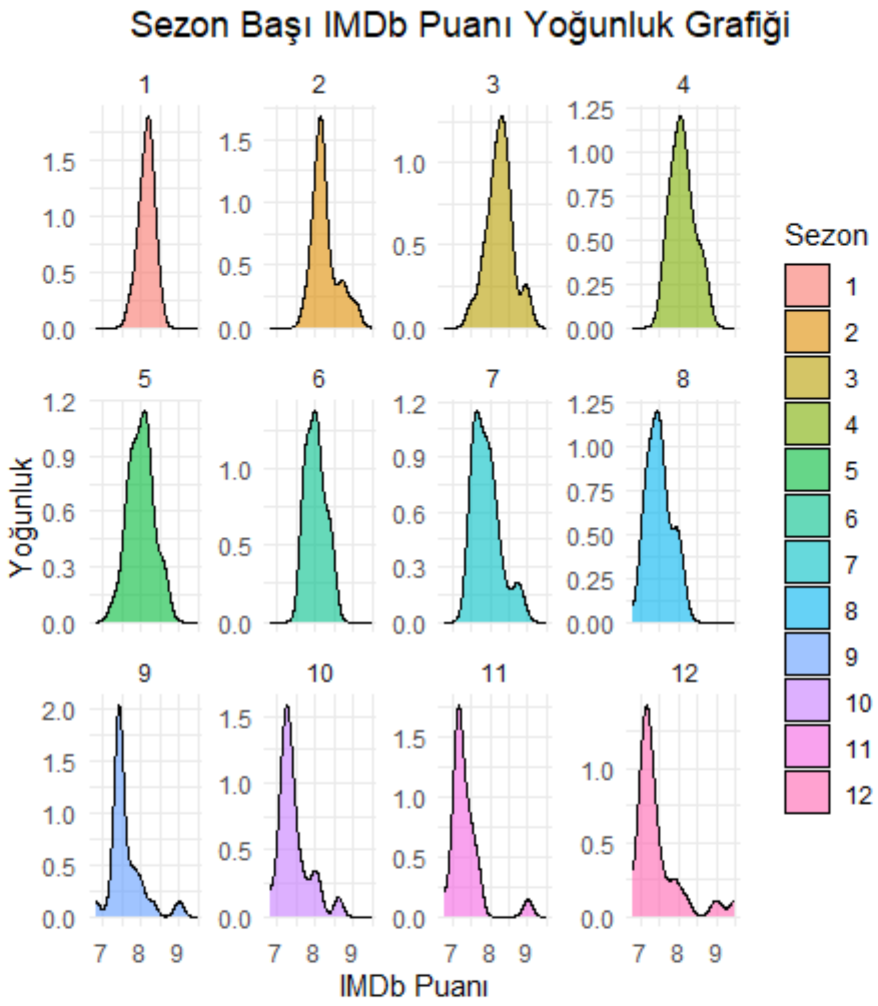


```
ggplot(median_imdb_by_season, aes(x = as.factor(season), y = median_rating)) +
  geom_col(fill = "lightblue", color = "black") +
  labs(
    title = "               Sezon Başı Ortalama IMDb Puanı",
    x = "Sezon Numarası",
    y = "Ortalama IMDb Puanı"
  ) +
  coord_cartesian(ylim = c(0, 10))+
  scale_y_continuous(breaks = seq(0, 10, by = 1))+

  theme_minimal()
```

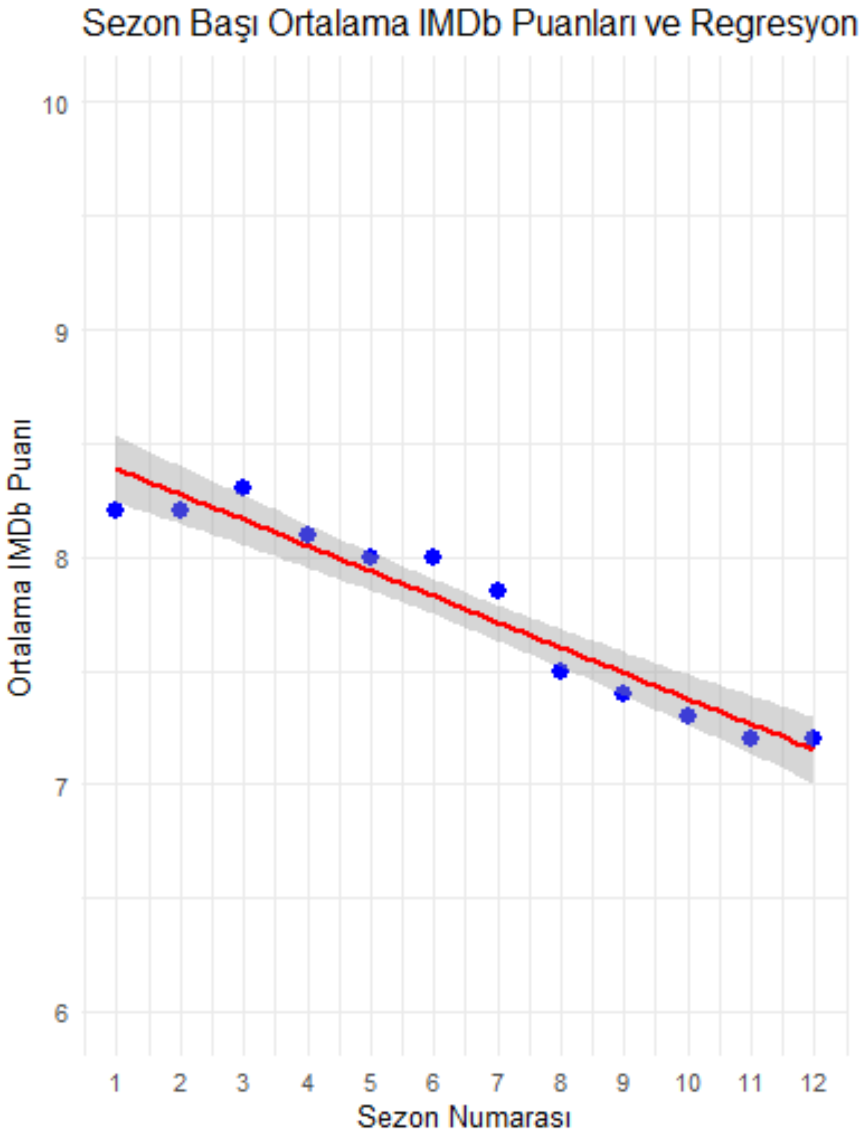


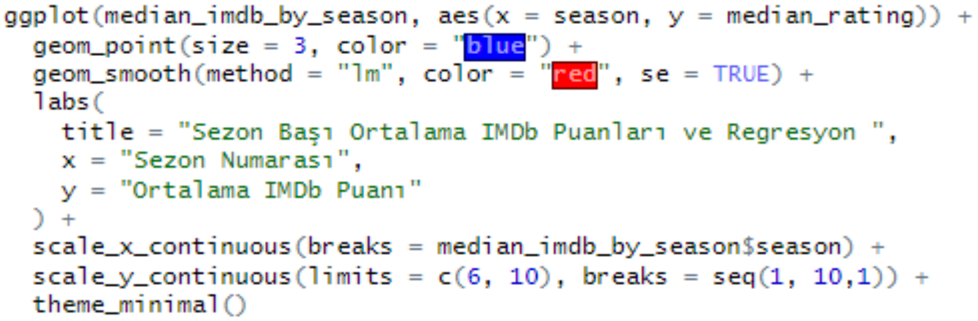


```
ggplot(imdbData, aes(x = imdb_rating, fill = factor(season))) +
  geom_density(alpha = 0.6, color = "black") +
  facet_wrap(~ season, scales = "free_y") +
  labs(title = "     Sezon Başı IMDb Puanı Yoğunluk Grafiği",
       x = "IMDb Puanı",
       y = "Yoğunluk",
       fill = "Sezon") +
  theme_minimal() +
  theme(legend.position = "right")
```



```
ggplot(imdbData, aes(x = imdb_rating, fill = factor(season))) +
  geom_density(alpha = 0.4) +
  labs(title = "     Sezon Başı IMDb Puanı Yoğunluk Grafiği",
       x = "IMDb Puanı",
       y = "Yoğunluk",
       fill = "Sezon") +
  theme_minimal() +
  theme(legend.position = "right")
```





```
ggplot(median_imdb_by_season, aes(x = season, y = median_rating)) +
  geom_point(size = 3, color = "blue") +
  geom_smooth(method = "lm", color = "red", se = TRUE) +
  labs(
    title = "Sezon Başı Ortalama IMDb Puanları ve Regresyon ",
    x = "Sezon Numarası",
    y = "Ortalama IMDb Puanı"
  ) +
  scale_x_continuous(breaks = median_imdb_by_season$season) +
  scale_y_continuous(limits = c(6, 10), breaks = seq(1, 10,1)) +
  theme_minimal()
```

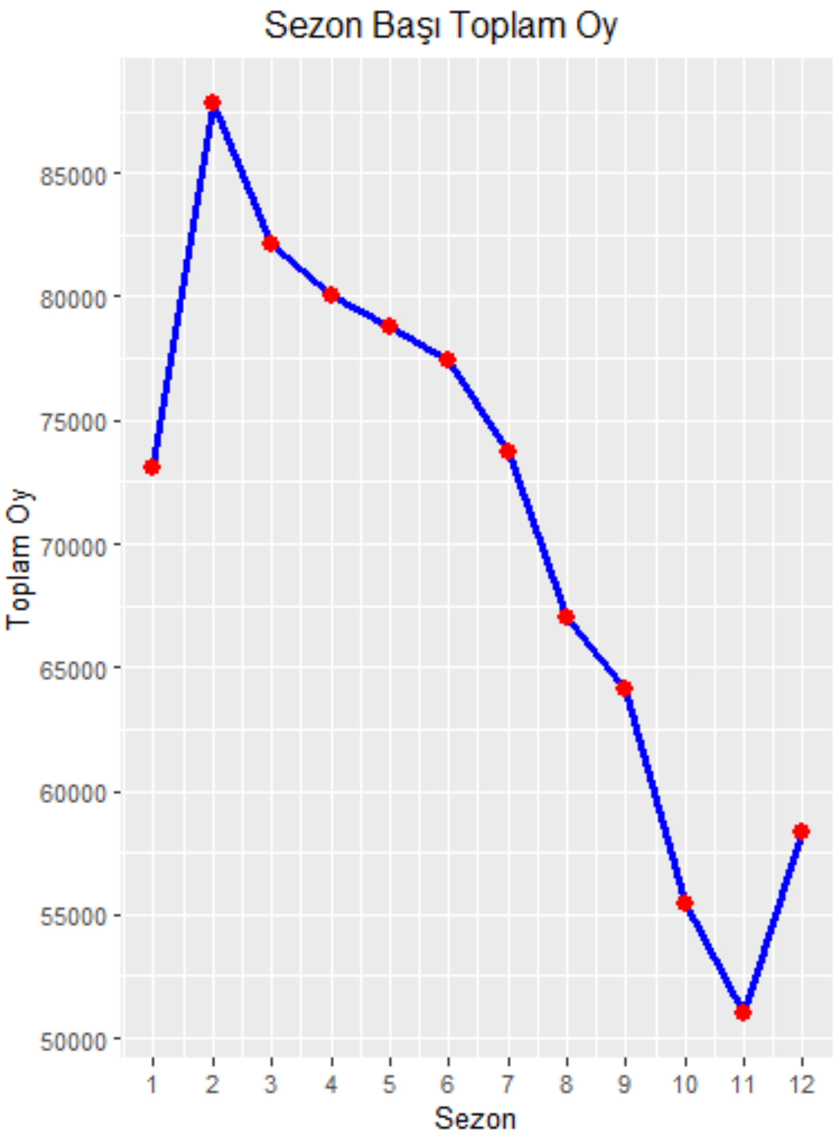
Conclusions of Research Question

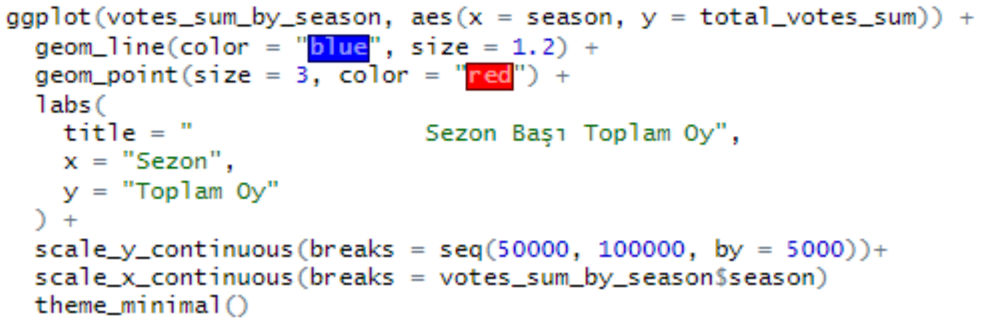
The distribution of IMDB ratings varies significantly across the seasons. Most seasons show a relatively symmetric distribution of ratings, centering around higher values (7.5–9.0), indicating overall audience appreciation.

However, the season with the most skewed rating distribution is Season 1, where the ratings exhibit a broader spread, including episodes with lower scores (as low as 6.7). This skew suggests a mixed audience reception, possibly due to the early stages of the show as it established its characters and storyline.

Subsequent seasons demonstrate a more consistent, less skewed distribution, reflecting the series' increasing popularity and stable episode quality over time.

Investigate the distribution of vote counts by season. Determine the season with the highest differences in vote counts.





```
ggplot(votes_sum_by_season, aes(x = season, y = total_votes_sum)) +
  geom_line(color = "blue", size = 1.2) +
  geom_point(size = 3, color = "red") +
  labs(
    title = "                    Sezon Başı Toplam Oy",
    x = "Sezon",
    y = "Toplam Oy"
  ) +
  scale_y_continuous(breaks = seq(50000, 100000, by = 5000))+
  scale_x_continuous(breaks = votes_sum_by_season$season)
  theme_minimal()
```

Conclusions of Research Question

Conclusion:

The analysis of vote count distributions across seasons reveals notable differences:

- Season 1 exhibits the highest variability in vote counts, with significant differences between episodes. This suggests that early episodes had inconsistent audience engagement, likely due to the show's gradual build-up in popularity.
- Later seasons show more compact vote count distributions, reflecting stable audience engagement and consistent popularity as the series matured.

In summary, Season 1 stands out with the largest discrepancies in vote counts, indicating varied viewer interest during the show's initial phase.

mdc.csv

Summary of the Dataset

This dataset consists of film-level information spanning multiple decades of Marvel and DC cinematic productions. It provides a comprehensive view that enables comparisons between two of the most prominent superhero franchises in the entertainment industry. Below are the key features and their relevance:

Film Metadata:

- Universe Affiliation (Marvel or DC): Fundamental for categorizing each film and enabling comparative analyses between the two leading superhero domains.
- Title and Release Year: Important for identifying individual films and examining trends and evolutions over time.

Production Details:

- Directors and Writers: Facilitate frequency and qualitative analysis of creative contributors, highlighting patterns in storytelling styles and talent involvement.
- Production Attributes (if available): Provide insights into production companies, budgets, and filming logistics.

Critical Reception:

- IMDb Ratings: A numerical measure reflecting audience perception and engagement.
- Rotten Tomatoes Scores: A critic-centered perspective that helps assess critical acclaim and broader industry reception.

This dataset supports a wide range of analytical approaches, such as:

- **Frequency Analysis:** Determine the most prolific directors or identify which universe has produced the most films over certain periods.
- **Descriptive Statistics:** Compute metrics like mean or median IMDb and Rotten Tomatoes scores to understand general reception patterns.
- **Chronological and Comparative Studies:** Explore how reception, creative contributors, and production values vary across different time frames, and contrast Marvel vs. DC trends.

Data Introduction

The dataset contains a collection of Marvel and DC films, two of the most influential and enduring brands in the superhero genre. It includes both quantitative and qualitative variables, offering a well-rounded perspective on factors such as critical reception, production attributes, and creative talent. This dataset is particularly suited for summarization and exploratory data analysis due to its structured format.

Key aspects include categorical variables such as film titles, associated universes (Marvel or DC), and directors, alongside numerical variables like IMDb ratings and Rotten Tomatoes scores. By summarizing and visualizing these components, we can gain insights into patterns such as audience engagement, critical acclaim, and the evolution of storytelling across different eras and production teams.

The dataset serves as a foundation for various statistical analyses, including frequency counts of directors, distributions of film ratings, and release trends over time. These analyses can help illuminate how superhero cinema has grown, adapted, and resonated with audiences, while also highlighting the roles of key creators and the shifting landscapes of both Marvel and DC properties.

Research Questions of mdc.csv

1) Investigate the IMDB rating distribution of Marvel movies before and after the year 2000. Determine which period had more variation in ratings.

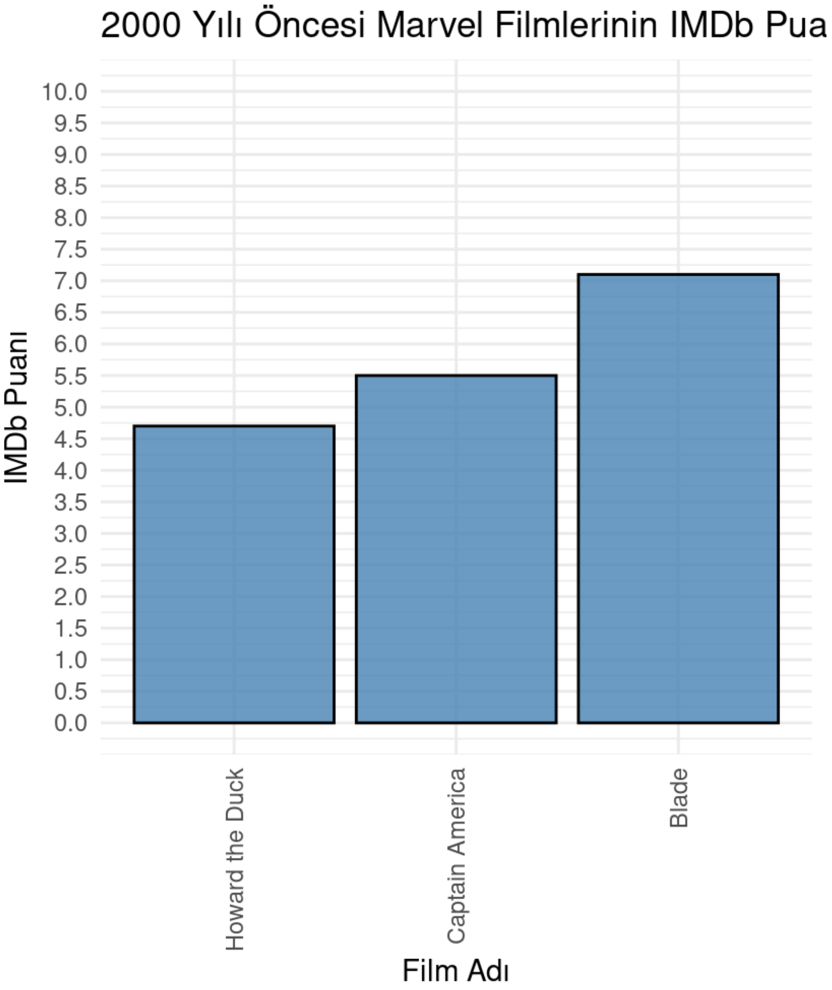
Investigating the IMDB rating distribution of Marvel movies before and after the year 2000 is crucial for understanding how audience perception of Marvel films has evolved over time. Determining which period had more variation in ratings can highlight differences in quality, consistency, and audience reception. This analysis can provide insights into the impact of changes in storytelling, production quality, and the rise of the Marvel Cinematic Universe on movie ratings.

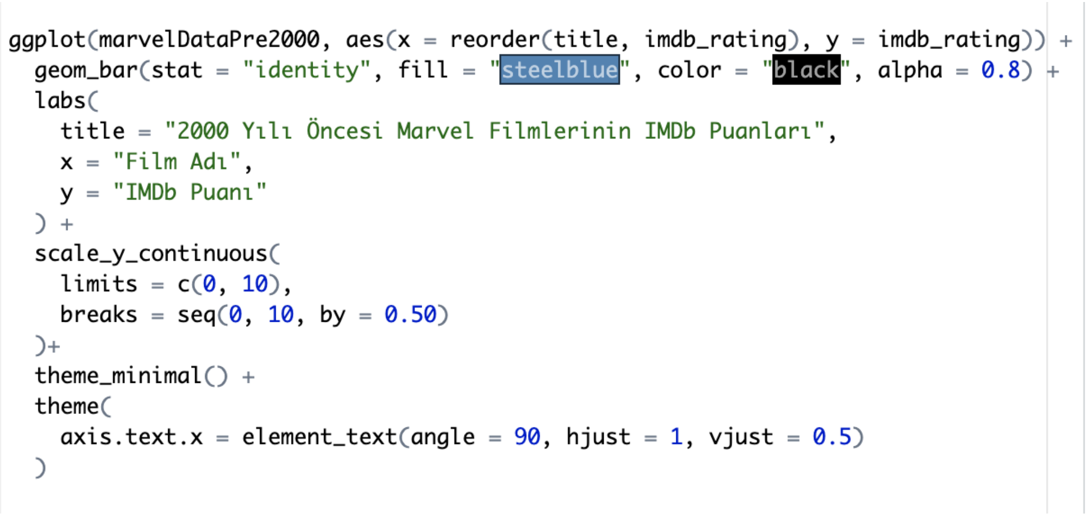
2) Examine the duration distribution of Marvel movies before and after the year 2000. Determine which period had more variation in durations.

Examining the duration distribution of Marvel movies before and after the year 2000 is essential to identify trends in movie lengths over time. Determining which period had more variation in durations can reveal changes in storytelling approaches, production styles, and audience expectations. This analysis helps to understand whether modern Marvel movies have become more standardized or if there was greater flexibility in runtimes during earlier periods.

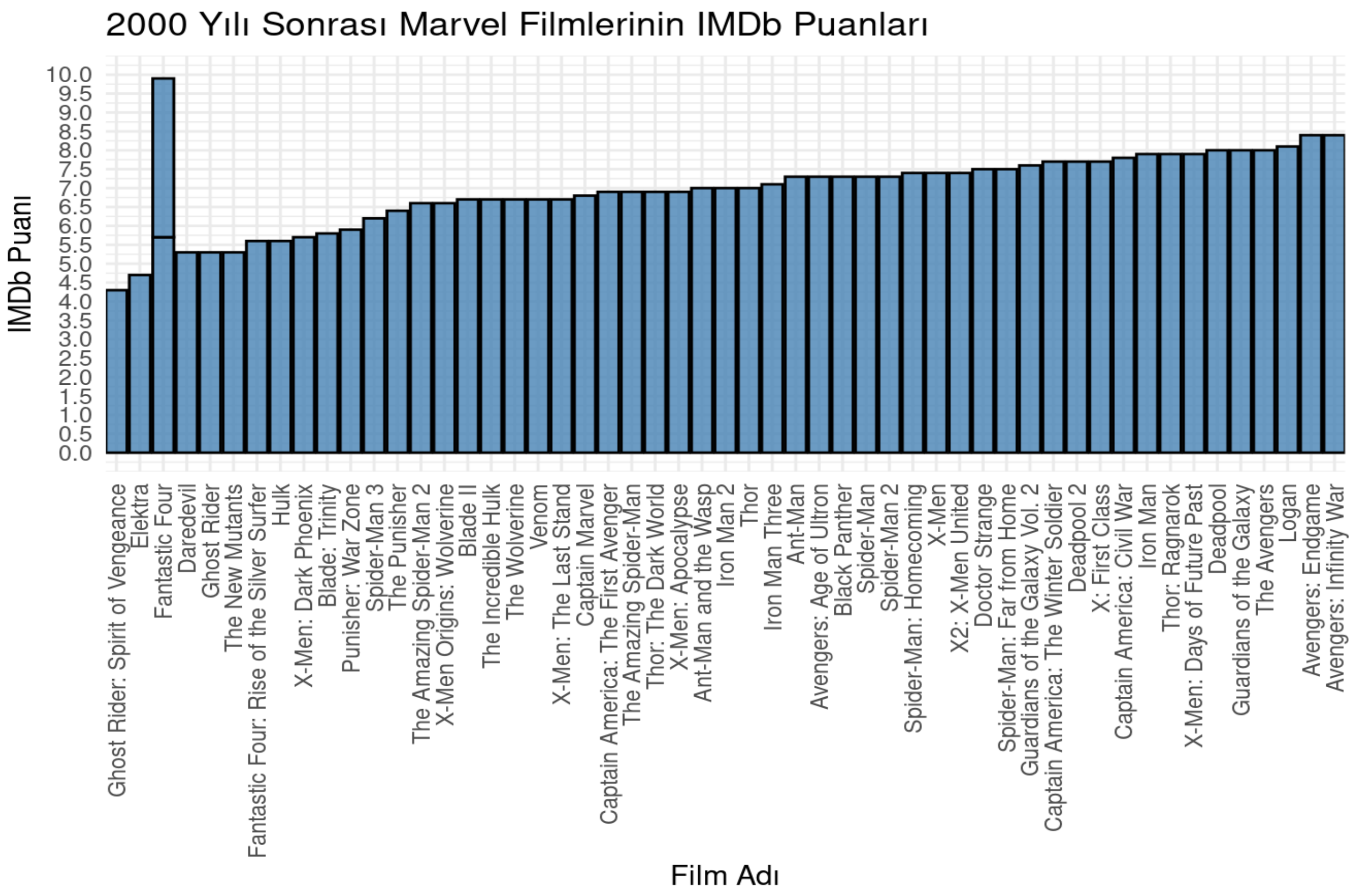
Analysis

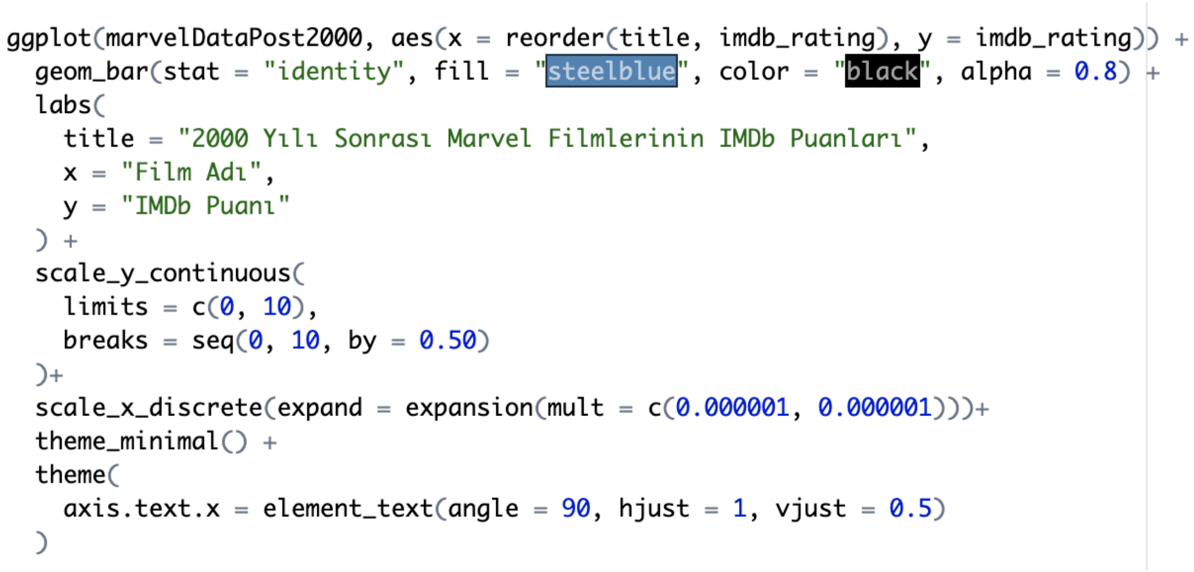
Investigate the IMDB rating distribution of Marvel movies before and after the year 2000. Determine which period had more variation in ratings.



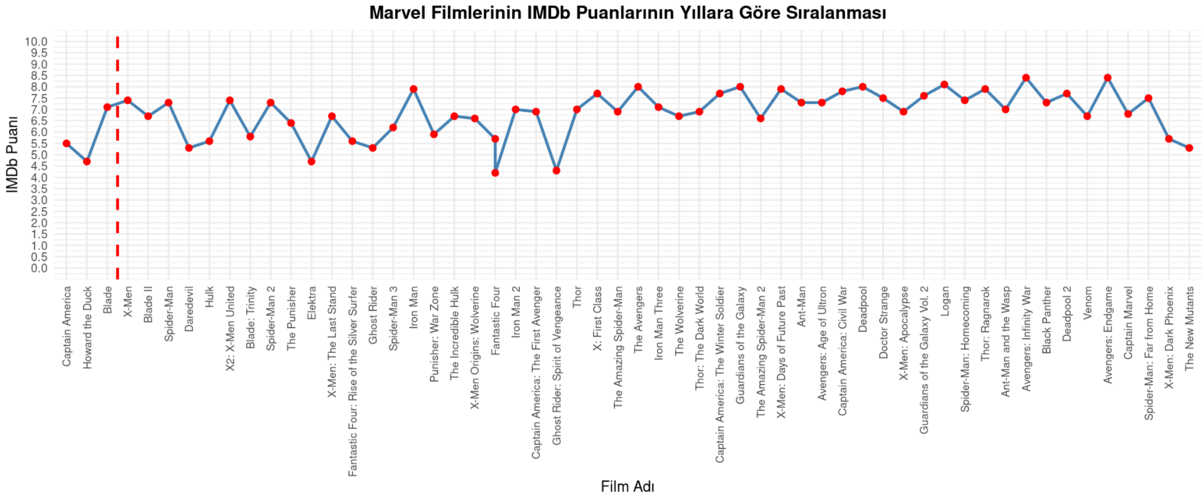


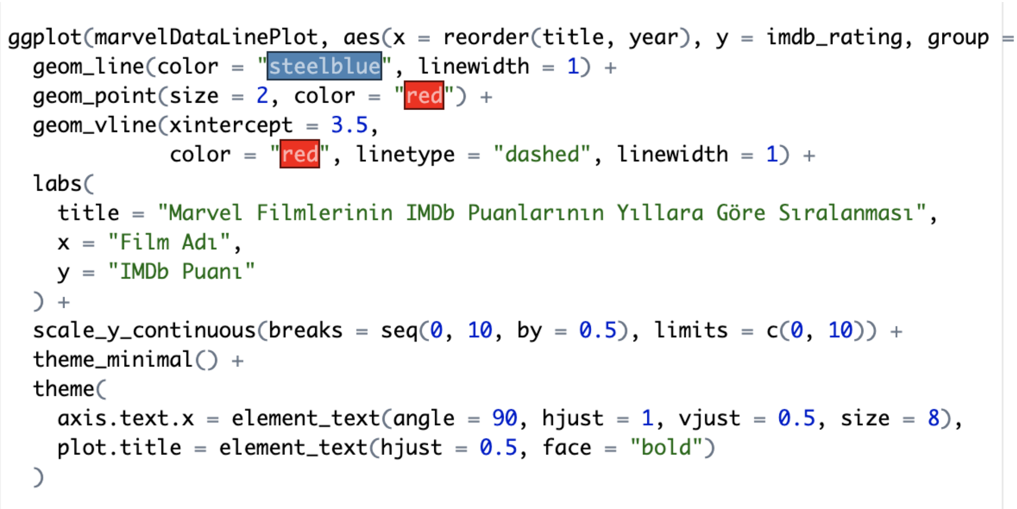
```r
ggplot(marvelDataPre2000, aes(x = reorder(title, imdb_rating), y = imdb_rating)) +
  geom_bar(stat = "identity", fill = "steelblue", color = "black", alpha = 0.8) +
  labs(
    title = "2000 Yılı Öncesi Marvel Filmlerinin IMDb Puanları",
    x = "Film Adı",
    y = "IMDb Puanı"
  ) +
  scale_y_continuous(
    limits = c(0, 10),
    breaks = seq(0, 10, by = 0.50)
  )+
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5)
  )
```





```
ggplot(marvelDataPost2000, aes(x = reorder(title, imdb_rating), y = imdb_rating)) +
  geom_bar(stat = "identity", fill = "steelblue", color = "black", alpha = 0.8) +
  labs(
    title = "2000 Yılı Sonrası Marvel Filmlerinin IMDb Puanları",
    x = "Film Adı",
    y = "IMDb Puanı"
  ) +
  scale_y_continuous(
    limits = c(0, 10),
    breaks = seq(0, 10, by = 0.50)
  )+
  scale_x_discrete(expand = expansion(mult = c(0.000001, 0.000001)))+
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5)
  )
```



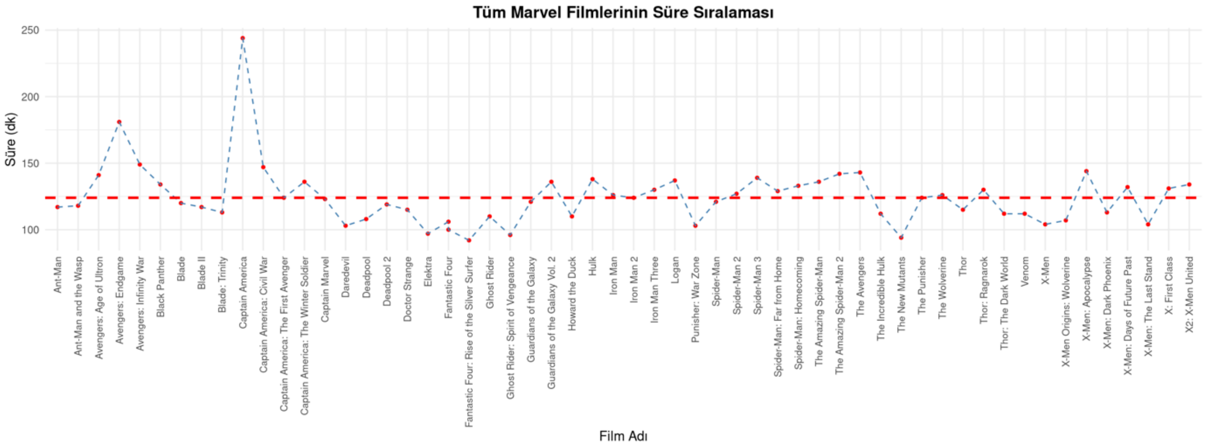


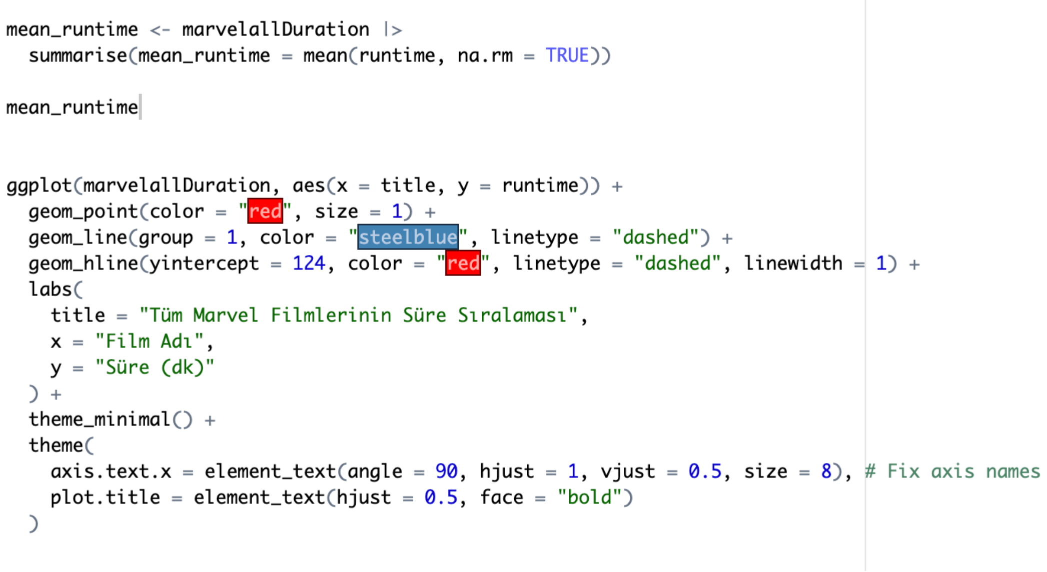
```r
ggplot(marvelDataLinePlot, aes(x = reorder(title, year), y = imdb_rating, group = 1)) +
  geom_line(color = "steelblue", linewidth = 1) +
  geom_point(size = 2, color = "red") +
  geom_vline(xintercept = 3.5,
             color = "red", linetype = "dashed", linewidth = 1) +
  labs(
    title = "Marvel Filmlerinin IMDb Puanlarının Yıllara Göre Sıralanması",
    x = "Film Adı",
    y = "IMDb Puanı"
  ) +
  scale_y_continuous(breaks = seq(0, 10, by = 0.5), limits = c(0, 10)) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5, size = 8),
    plot.title = element_text(hjust = 0.5, face = "bold")
  )
```

Conclusions of Research Question

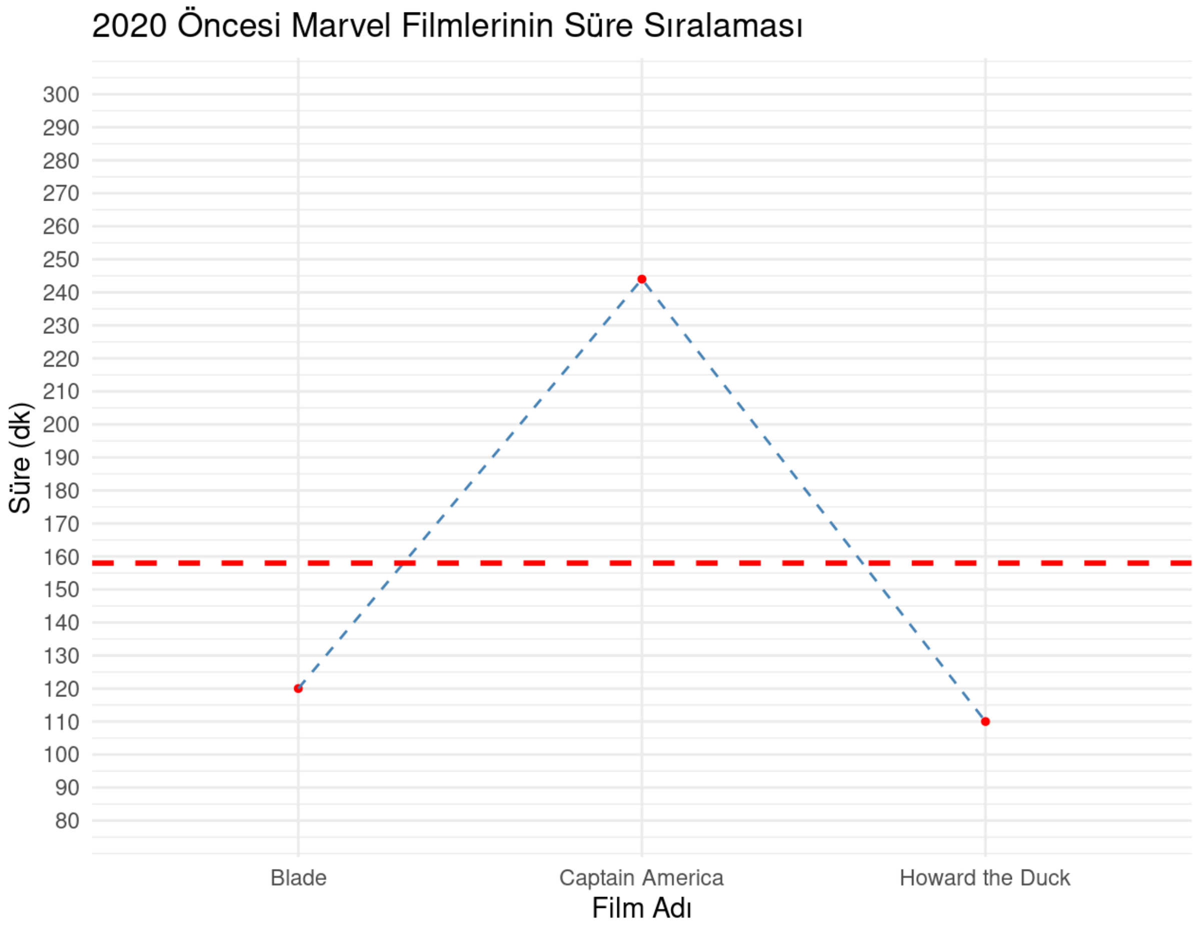
The variation in IMDb ratings of Marvel movies before the year 2000 was higher compared to those after the year 2000. This indicates that the ratings were more spread out and inconsistent in the earlier period, while they became relatively more consistent in the later period.

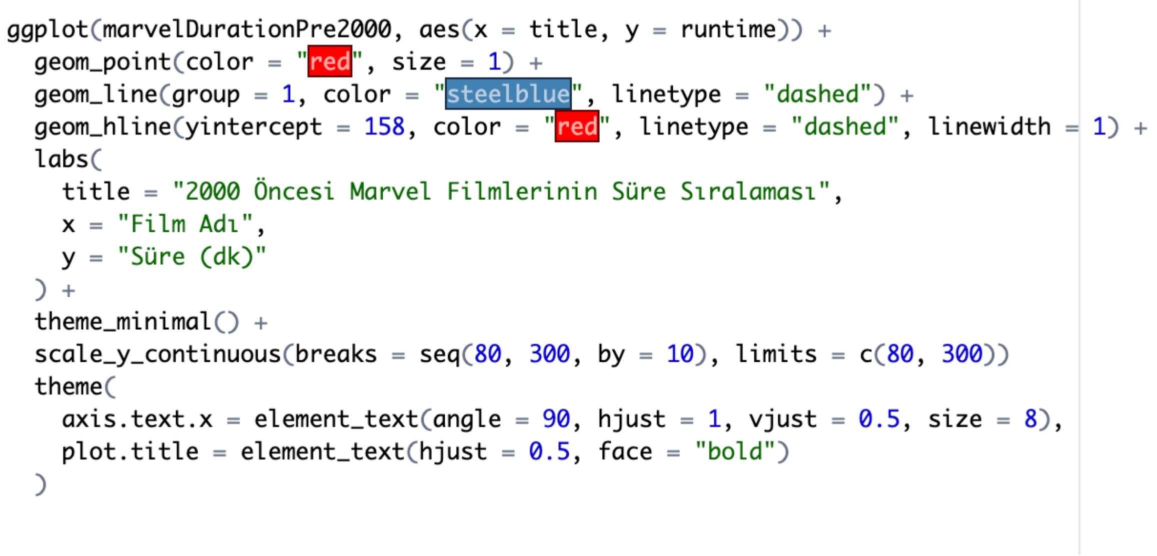
Examine the duration distribution of Marvel movies before and after the year 2000. Determine which period had more variation in durations.

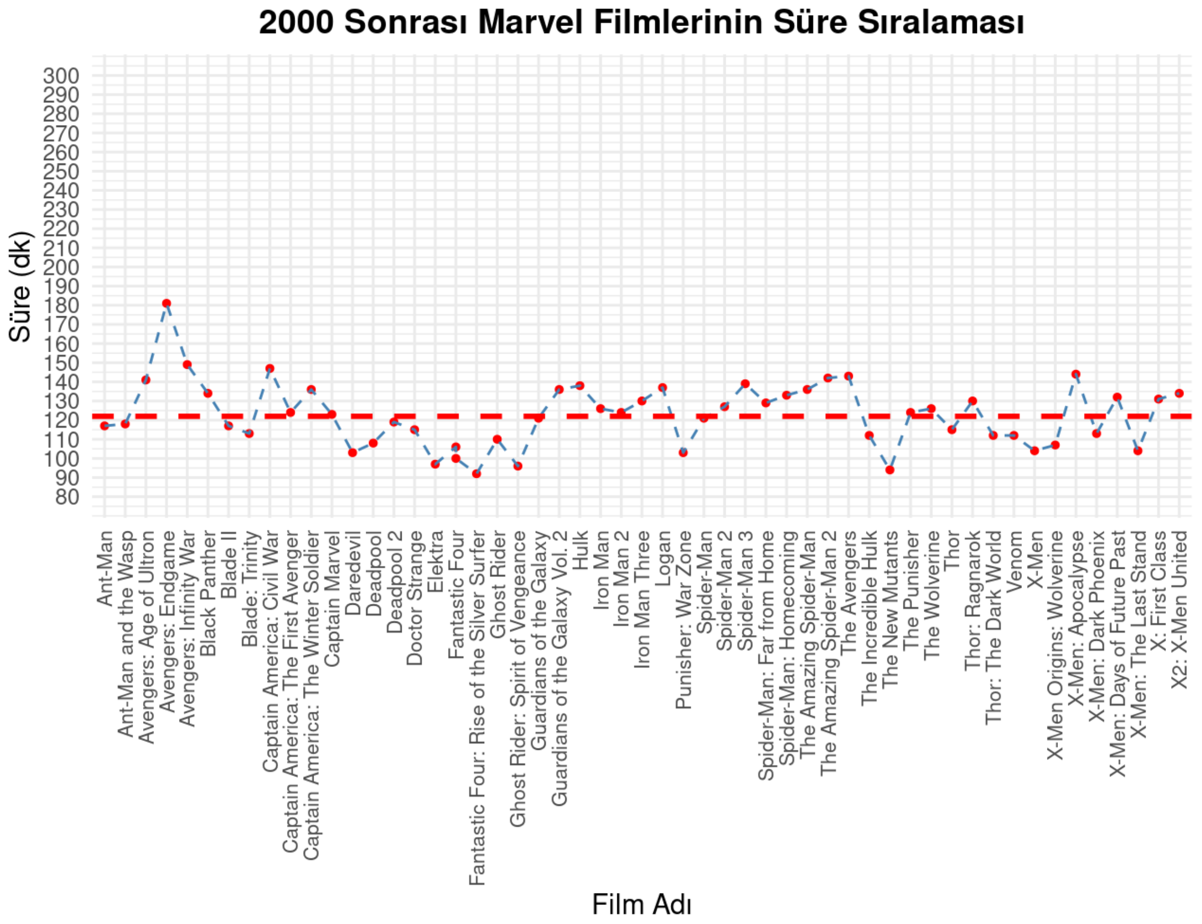


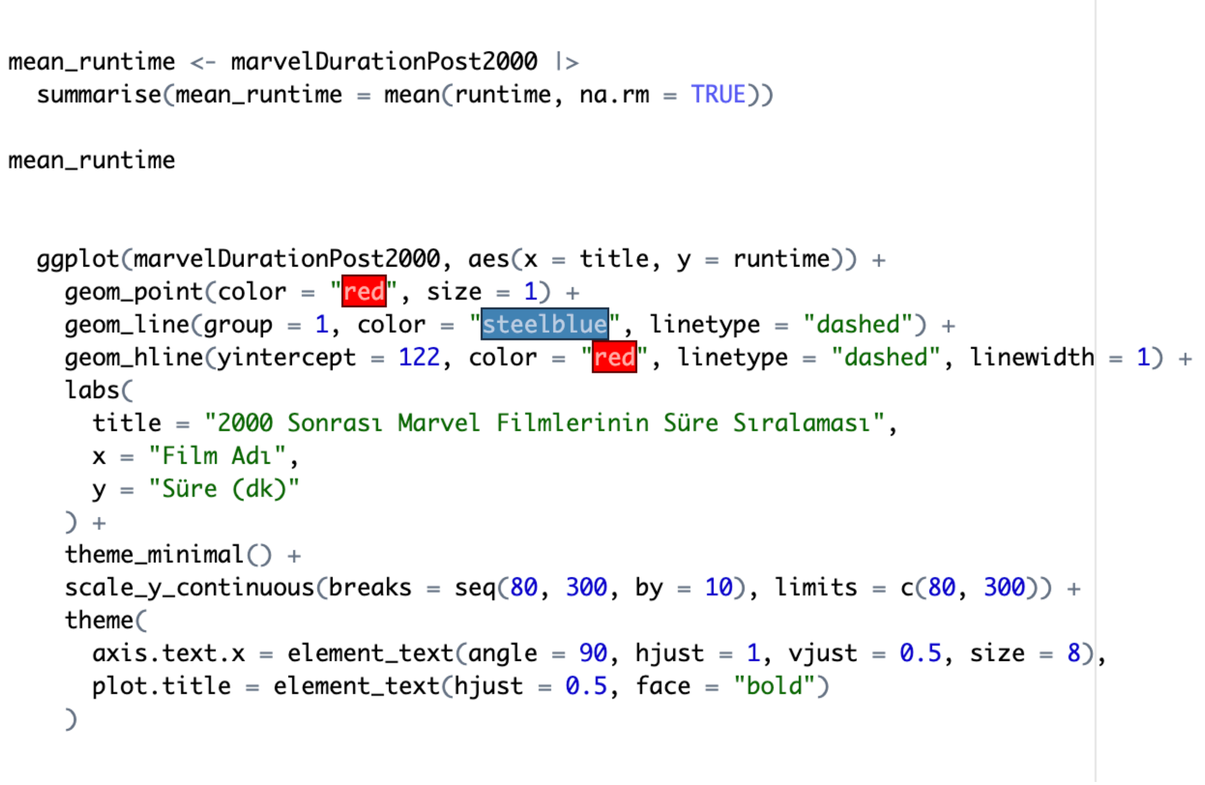


```r
mean_runtime <- marvelallDuration |>
  summarise(mean_runtime = mean(runtime, na.rm = TRUE))

mean_runtime


ggplot(marvelallDuration, aes(x = title, y = runtime)) +
  geom_point(color = "red", size = 1) +
  geom_line(group = 1, color = "steelblue", linetype = "dashed") +
  geom_hline(yintercept = 124, color = "red", linetype = "dashed", linewidth = 1) +
  labs(
    title = "Tüm Marvel Filmlerinin Süre Sıralaması",
    x = "Film Adı",
    y = "Süre (dk)"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5, size = 8), # Fix axis names
    plot.title = element_text(hjust = 0.5, face = "bold")
  )
```




```r
ggplot(marvelDurationPre2000, aes(x = title, y = runtime)) +
  geom_point(color = "red", size = 1) +
  geom_line(group = 1, color = "steelblue", linetype = "dashed") +
  geom_hline(yintercept = 158, color = "red", linetype = "dashed", linewidth = 1) +
  labs(
    title = "2000 Öncesi Marvel Filmlerinin Süre Sıralaması",
    x = "Film Adı",
    y = "Süre (dk)"
  ) +
  theme_minimal() +
  scale_y_continuous(breaks = seq(80, 300, by = 10), limits = c(80, 300))
  theme(
    axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5, size = 8),
    plot.title = element_text(hjust = 0.5, face = "bold")
  )
```





```
mean_runtime <- marvelDurationPost2000 |>
  summarise(mean_runtime = mean(runtime, na.rm = TRUE))

mean_runtime


  ggplot(marvelDurationPost2000, aes(x = title, y = runtime)) +
    geom_point(color = "red", size = 1) +
    geom_line(group = 1, color = "steelblue", linetype = "dashed") +
    geom_hline(yintercept = 122, color = "red", linetype = "dashed", linewidth = 1) +
    labs(
      title = "2000 Sonrası Marvel Filmlerinin Süre Sıralaması",
      x = "Film Adı",
      y = "Süre (dk)"
    ) +
    theme_minimal() +
    scale_y_continuous(breaks = seq(80, 300, by = 10), limits = c(80, 300)) +
    theme(
      axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5, size = 8),
      plot.title = element_text(hjust = 0.5, face = "bold")
    )
```

Conclusions of Research Question

The variation in movie durations of Marvel movies before the year 2000 was significantly higher compared to those after the year 2000. This indicates that movie durations were far more inconsistent in the earlier period, while they became more standardized and consistent in the later period