# Step 2: Exploratory Data Analysis Report

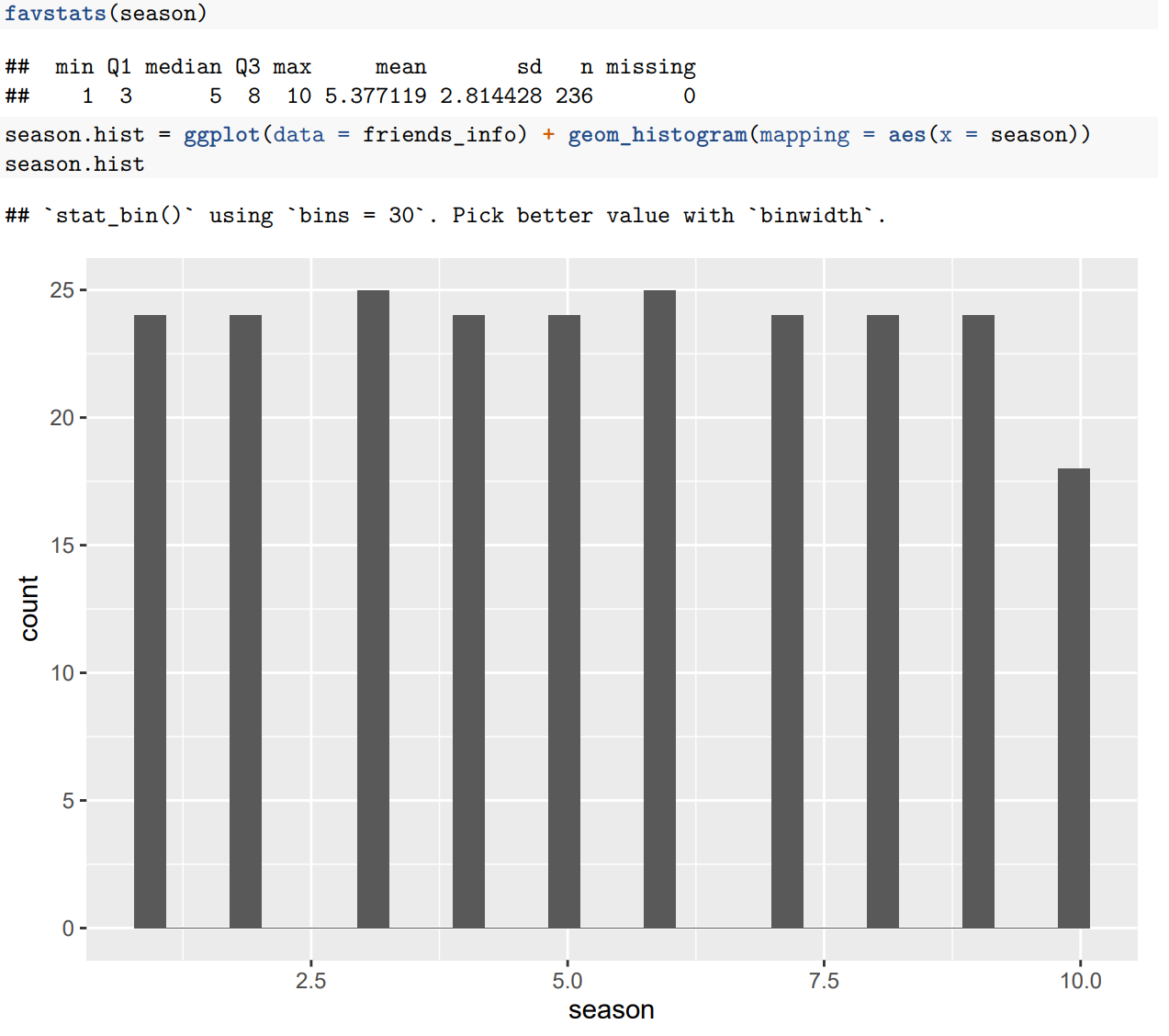
Khadeeja Naseer

Ruoqi Sun

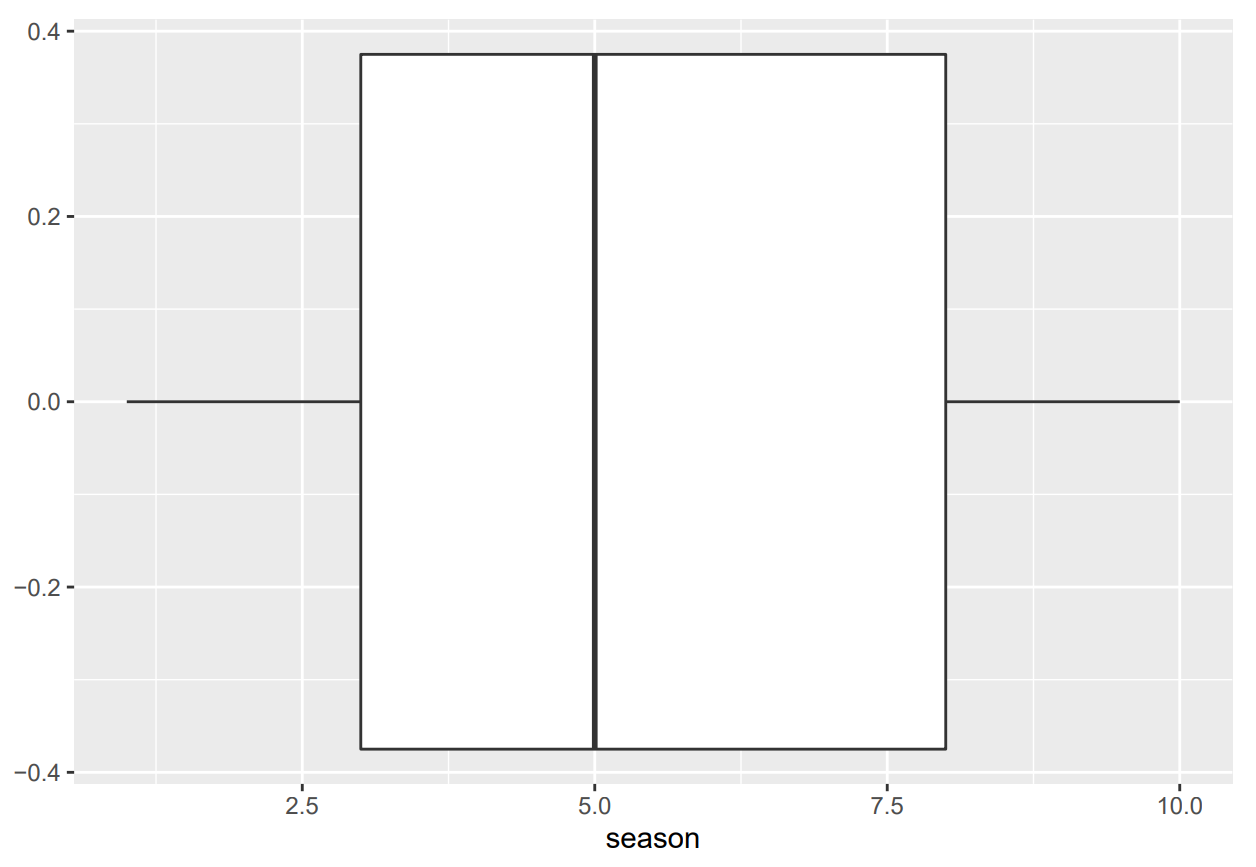
Michael Skotar

**1) Exploring Variables**

1. **Season (season number)**

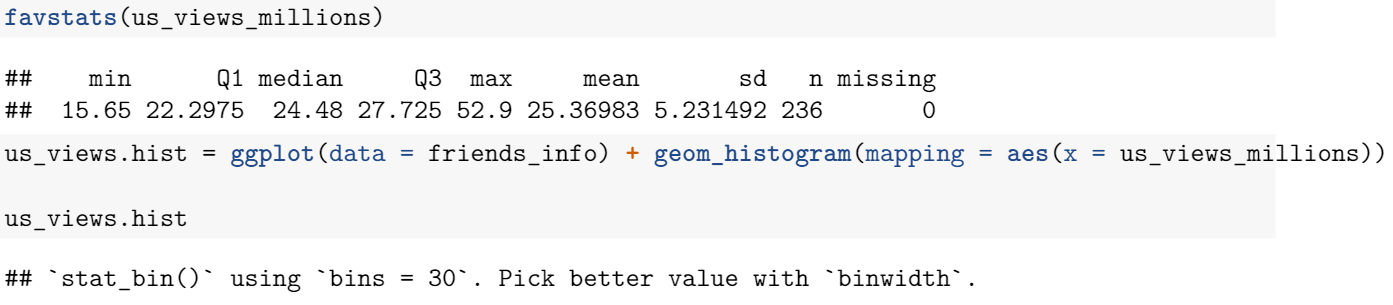


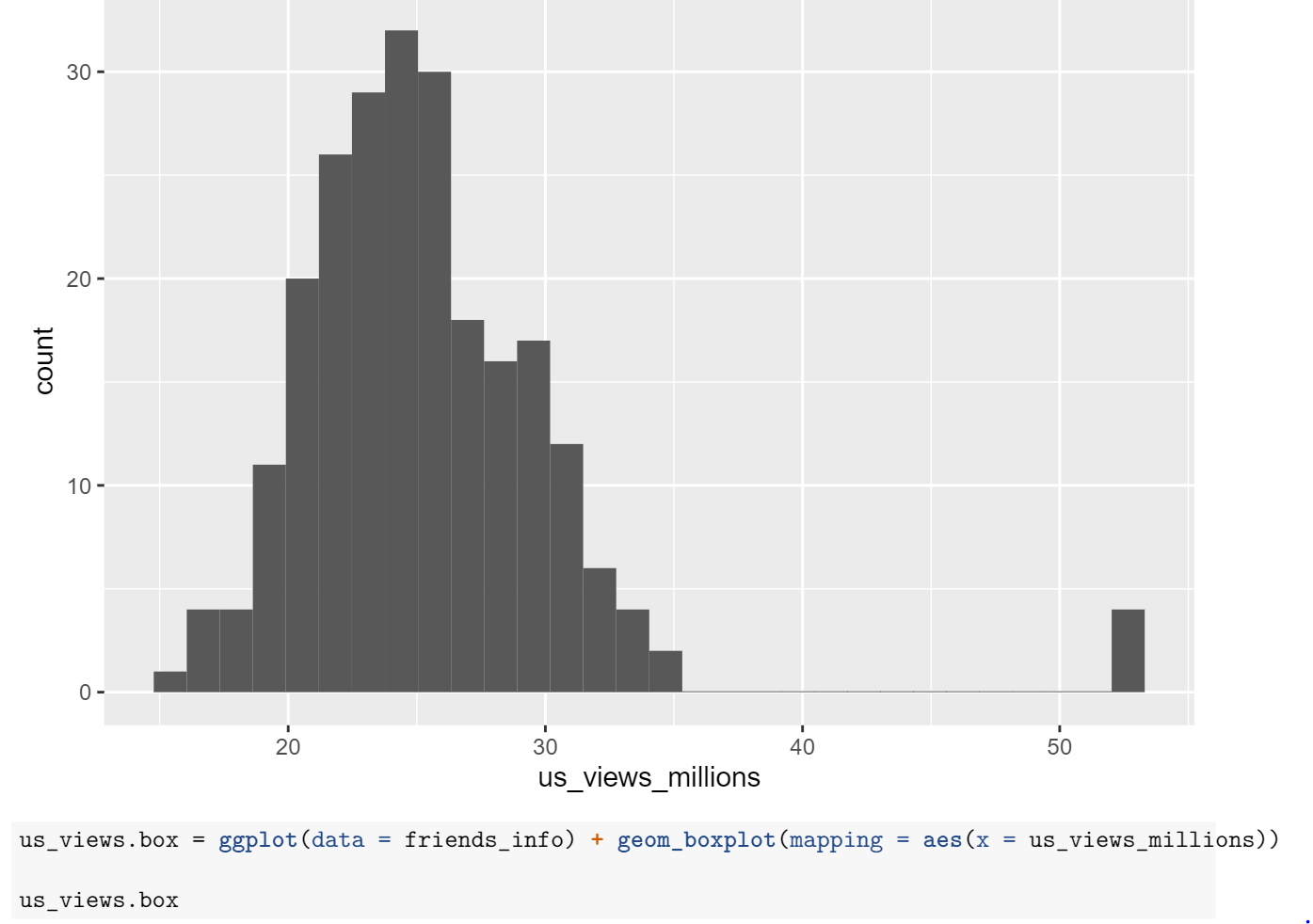


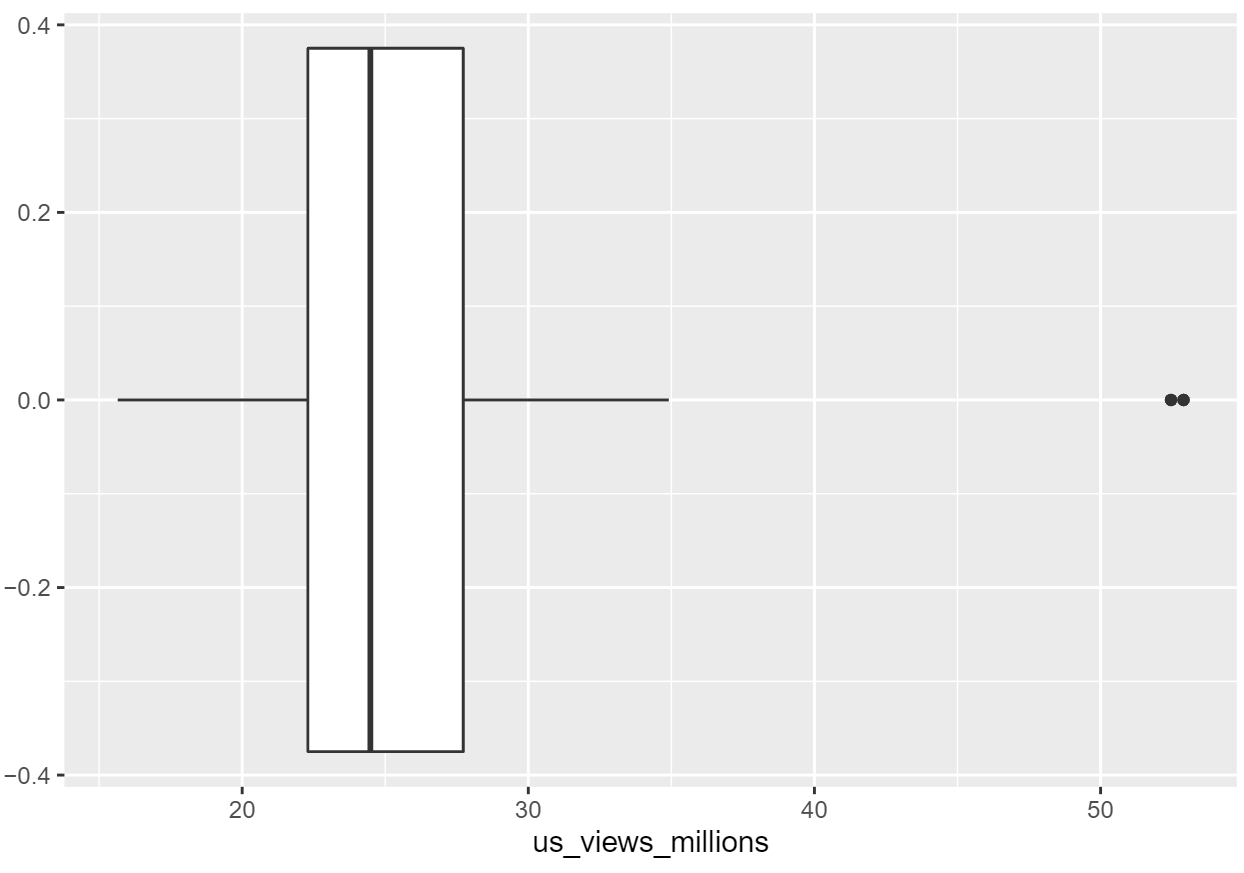


The seasons of Friends ranged from 1 to 10, with 5.377119 seasons on average. Median seasons were 5, indicating that 50% of the seasons had a season number of at most 5, and the other 50% had a season number that was at least 5. The middle 50% of the distribution of seasons ranged from 3 seasons to 8 seasons. The standard deviation of the seasons was 2.814428. Thus, we can expect that the season number differs from the mean by 2.814428, on average. Based on the histogram of the data, the distribution of the seasons appears to be uniform as almost all the bars have the same height, with the exception of season 10. The shape of the box plot appears to be slightly skewed to the right. There are also no extreme values.

1. **us\_views\_millions**

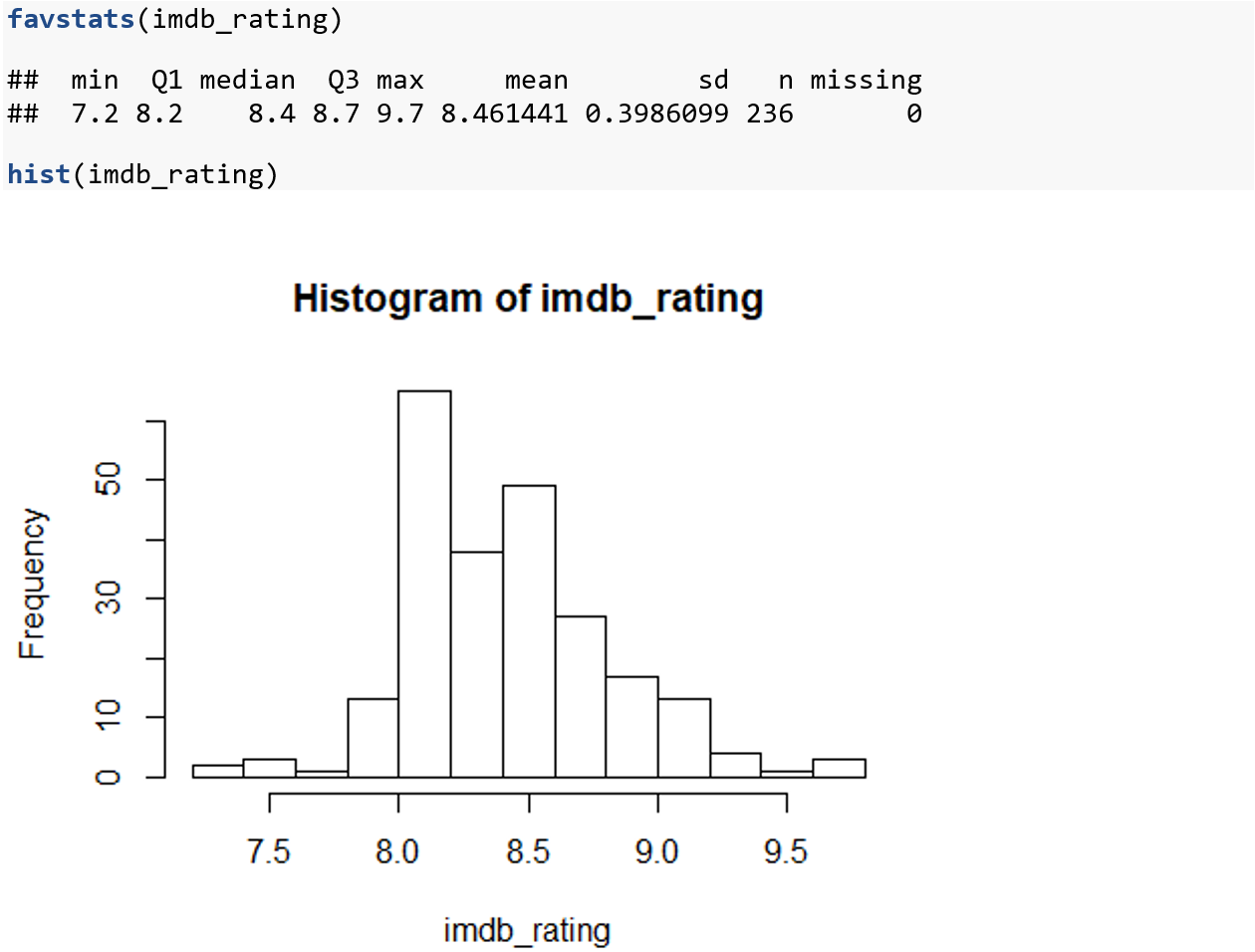




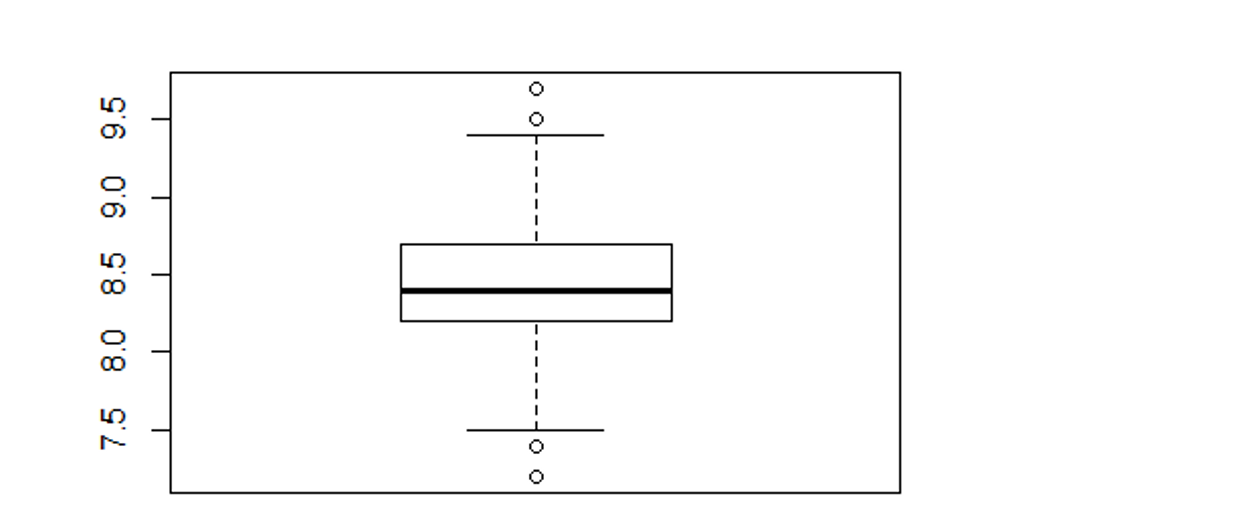


By analyzing the us\_views\_millions dataset in the Friends R package, there are two main plots of data and a basic summary that can be obtained. The number of American viewers from the first episode in season 1 to the last episode in season 10 has an average of 25.36983 millions. Median number of viewers are 24.48 millions, which means 50% of the data have at most 24.48 millions US viewers, and the other 50% have at least 24.48 millions viewers. The middle 50% of the distribution of US viewers range from 22.2975 millions to 27.725 millions. The standard deviation illustrates the number of American viewers differs from the mean by 5.231482, on average. Based on the histogram of the data, the distribution of the number of viewers appears to be unimodal since only one peak occurs in the histogram. The shape of the box plot appears to be skewed to the right. Few extreme values show on the right side of the plot.

**(c) imdb\_rating**

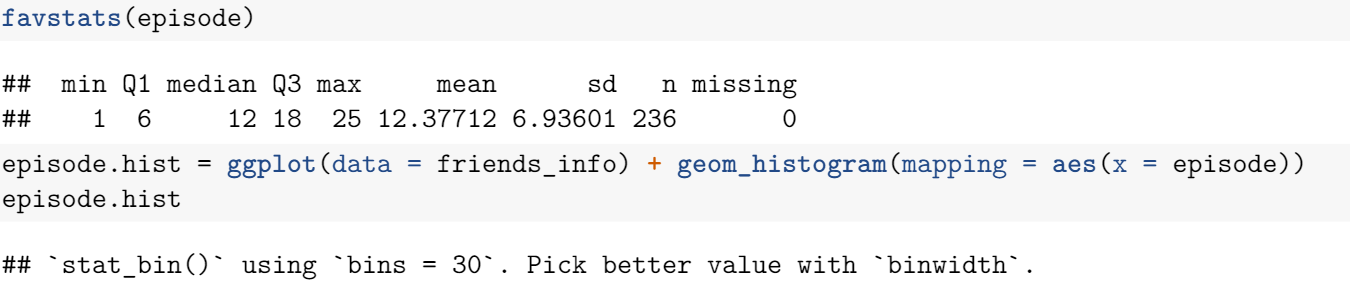


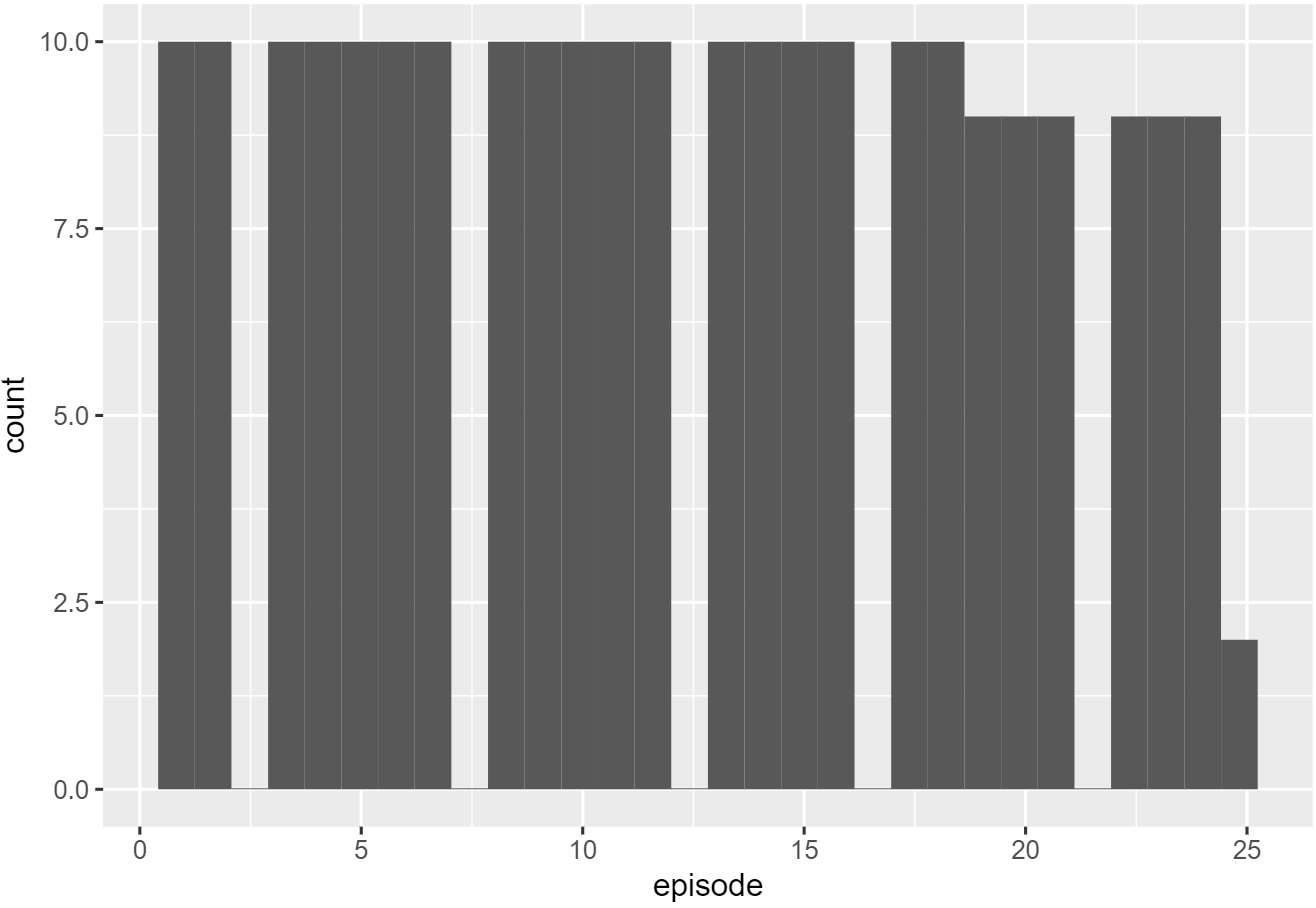




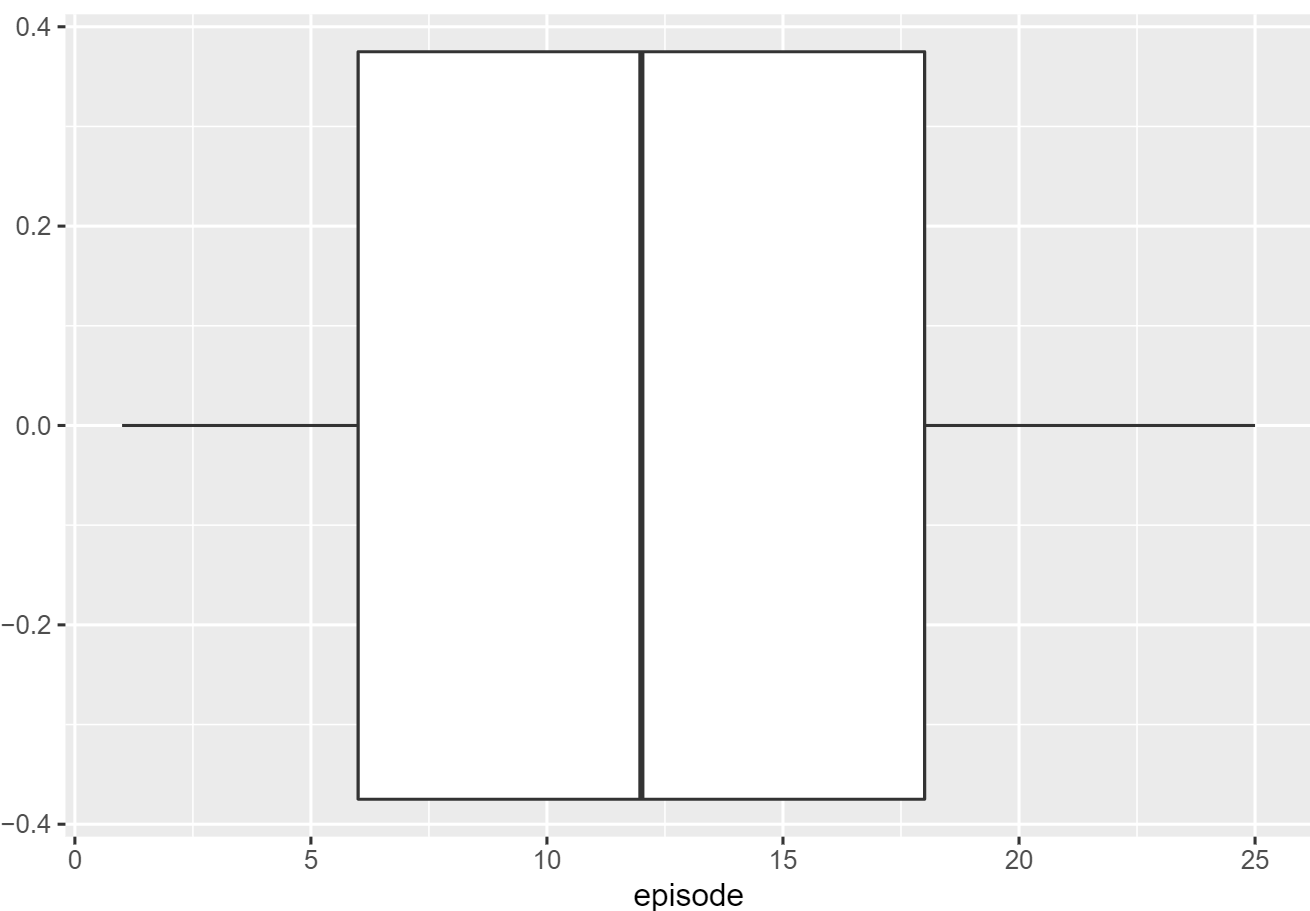
The imdb ratings of the tv show friends range from 7.2 to 9.7. The imdb rating system is on a 10 point scale where 1 is the lowest rating and 10 is the highest. Looking at the summary statistics for the imdb\_rating we see that the median is 8.4 which indicates that half of the ratings are above and half are below. With a standard deviation of 0.3986099 we see that the variation of the values is relatively low. There are 236 episodes over the lifespan of friends, and with a large population we see that there are a few outliers. There are two that are above 9.5 and two below 7.5. With regards to the boxplot these are outside of the range and therefore are part of the outliers in the data.

**(d) episode**





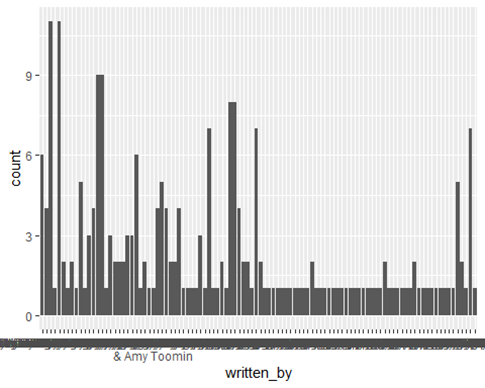


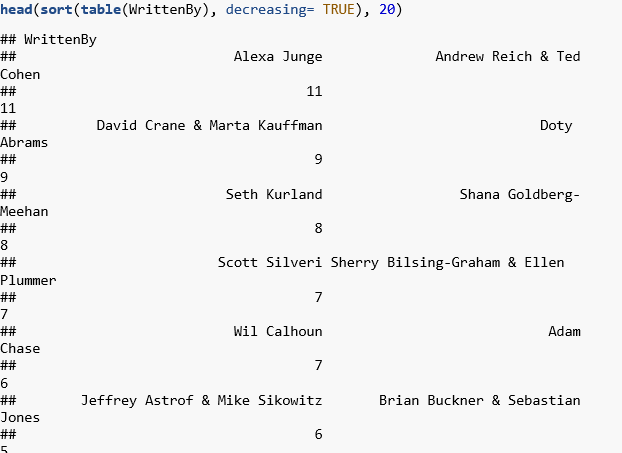


The episodes of Friends ranged from 1 to 25, with 12.37712 episodes on average. Median episodes were 12, indicating that 50% of the episodes of Friends had an episode number of at most 12, and the other 50% had an episode number that was at least 12. The middle 50% of the distribution of seasons ranged from 6 episodes to 18 episodes. The standard deviation of the episodes was 6.93601. Thus, we can expect that the episode number differs from the mean by 6.93601, on average. Based on the histogram of the data, the distribution of the episodes appears to be uniform as almost all the bars have the same height, with the exception of episode 25. The shape of the boxplot appears to be symmetric. There are also no extreme values.

**(e) written\_by**

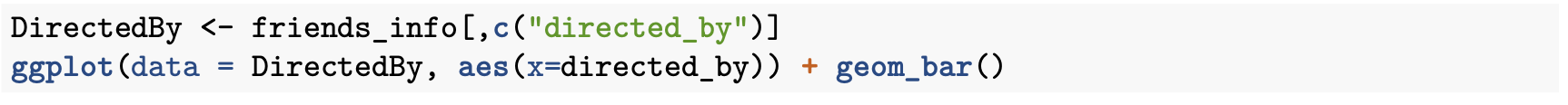


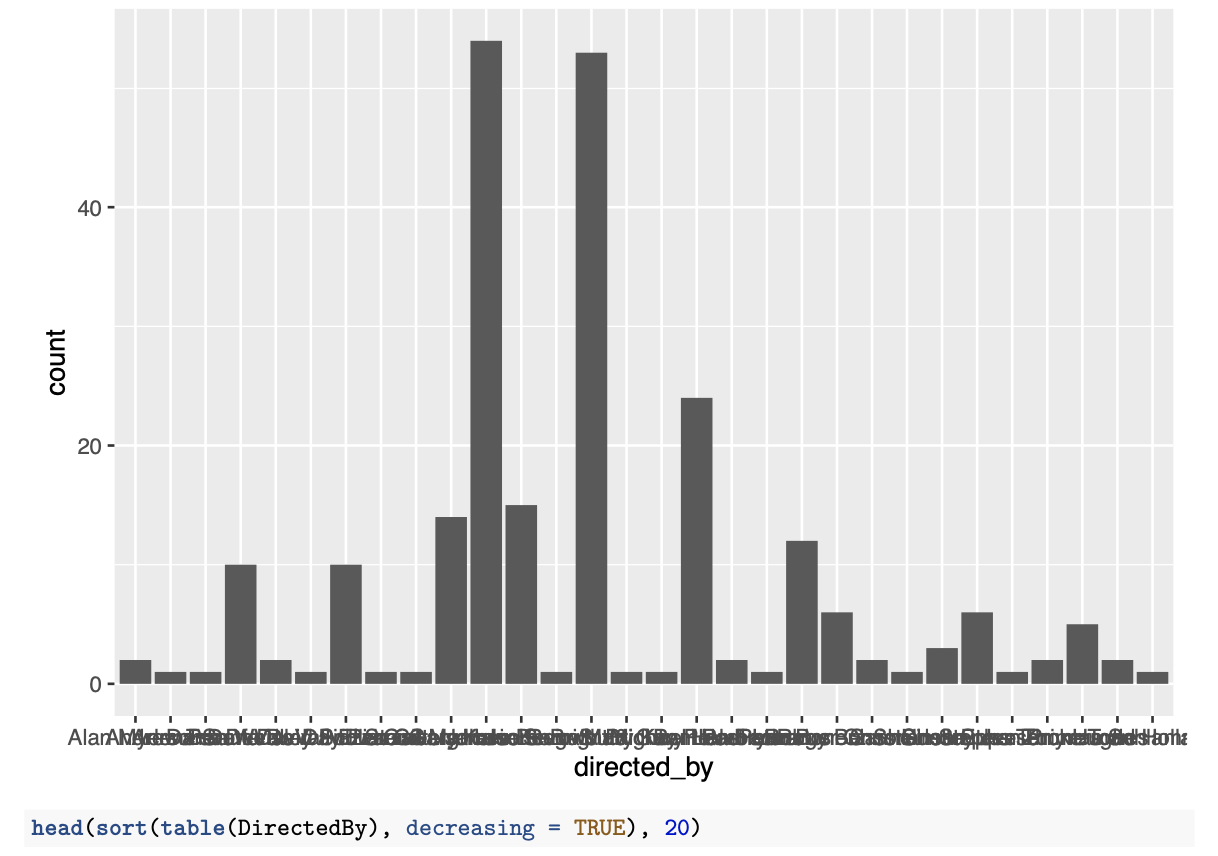


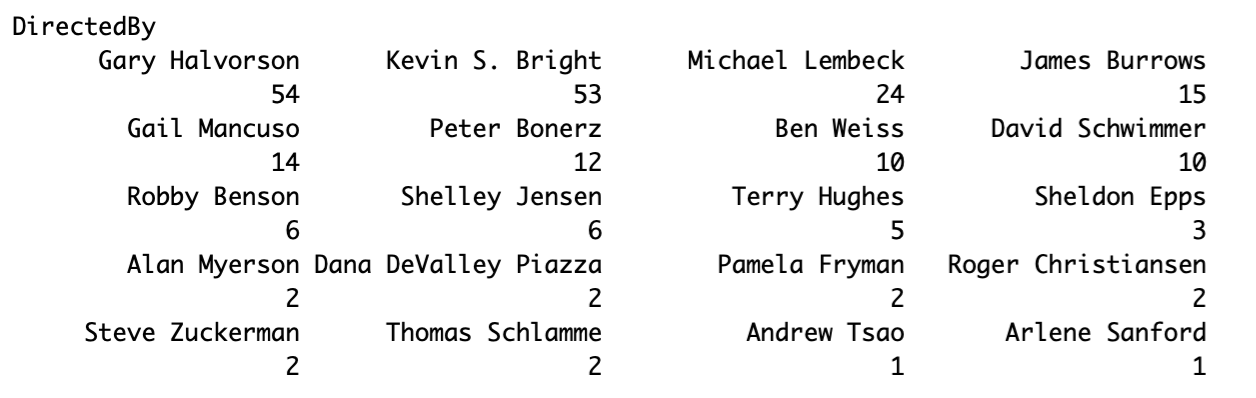


The variable written\_by is a categorical variable that shows who wrote that specific episode. In this case since it was shown to be names we cannot take the mean or the interquartile range of this statistic. In order to get a summary statistic of this variable we look at it with regards to frequency. If we look at how often the names of the writers appear we can see which writers were most popular. By taking the writer with the most frequency we have that Alexa Junge and Andrew Reich and Ted Cohen are the writers with the largest credit of 11. After this we have writers that come in with a frequency of 9, 8, all the way to 1. Most of the writers have a frequency of 1. This is due to them only writing one episode.

**(f) directed\_by**

****

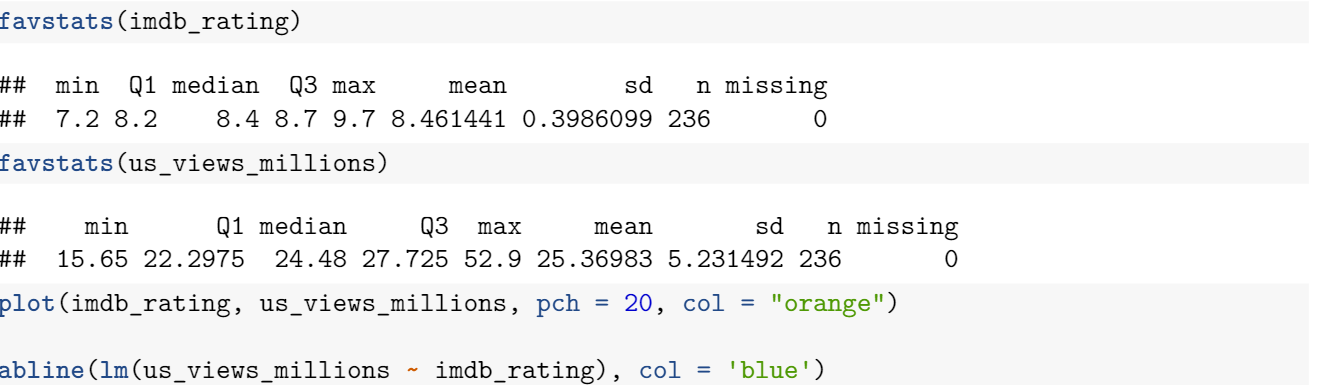
****

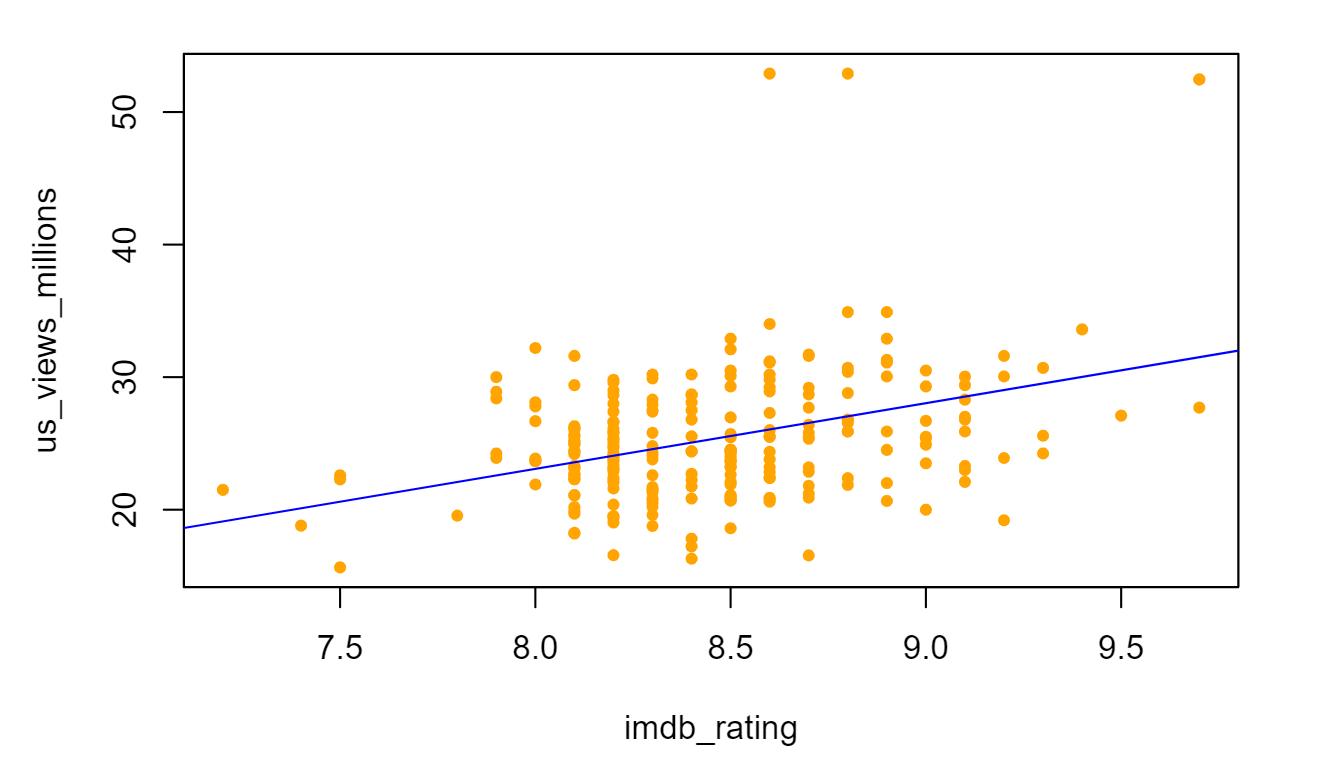
****

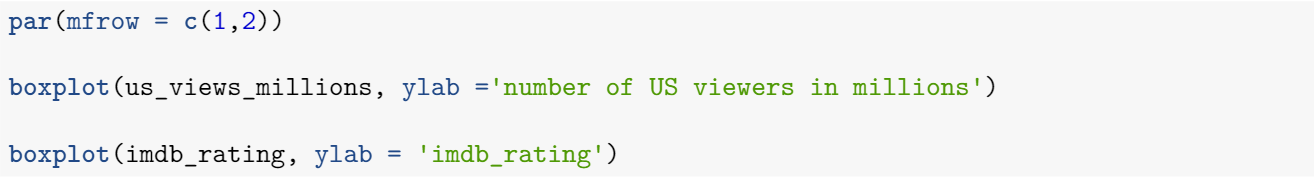
The variable "directed\_by" is a categorical variable that shows who directed that specific episode. The data was shown to be names, so we cannot generate a summary or boxplots to analyze the distribution of directors' names. In order to obtain a summary statistics of this variable, we can look at the frequency of each director's occurrence. By plotting the histogram with regard to each director, Gary Halvorson and Kevin S. Bright directed the most episodes, with 54 and 53 episodes respectively. After that Michael Lembeck has 24 episodes in this Friends series. Only two directors participated in one episode's direction, which were Andrew Tsao and Arlene Sanford. Overall, there were a total of 20 directors that joined in the filming of Friends.

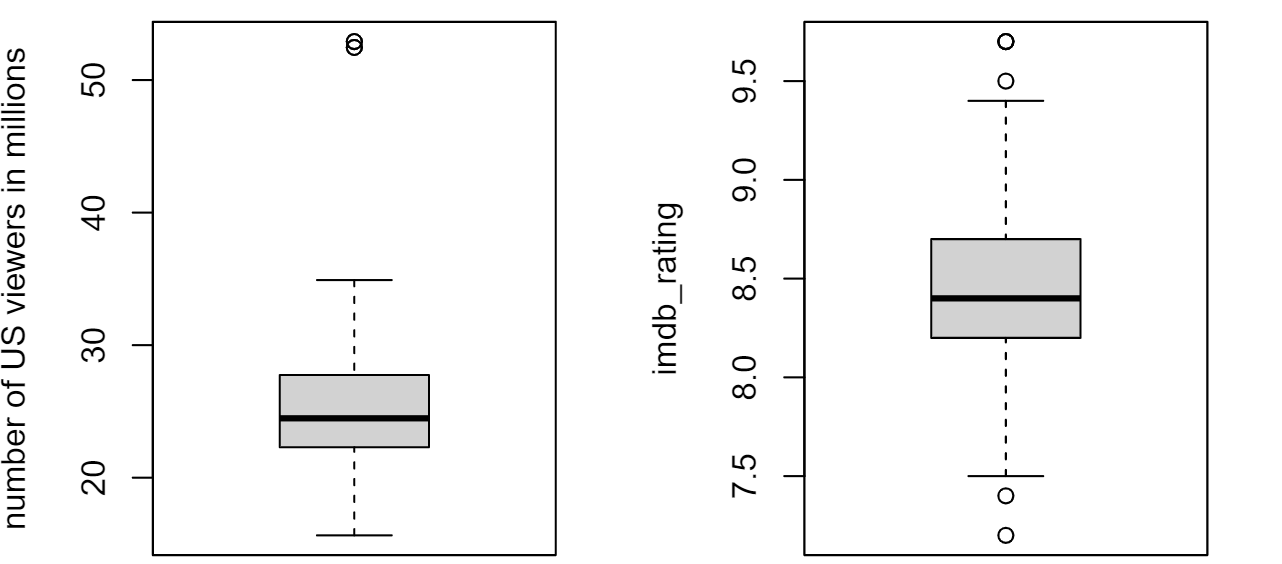
**2) Exploring Relationships Among Variables**

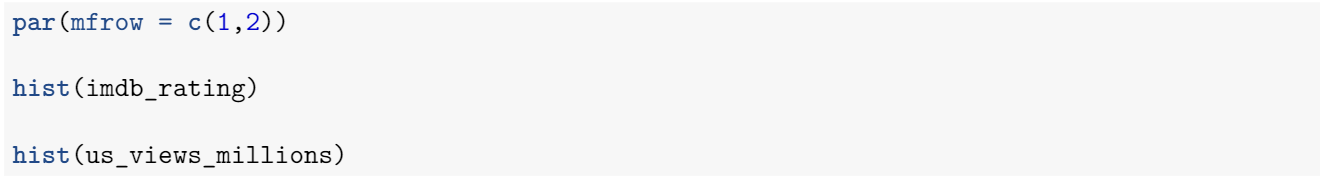
**us\_views\_millions vs imdb\_rating**

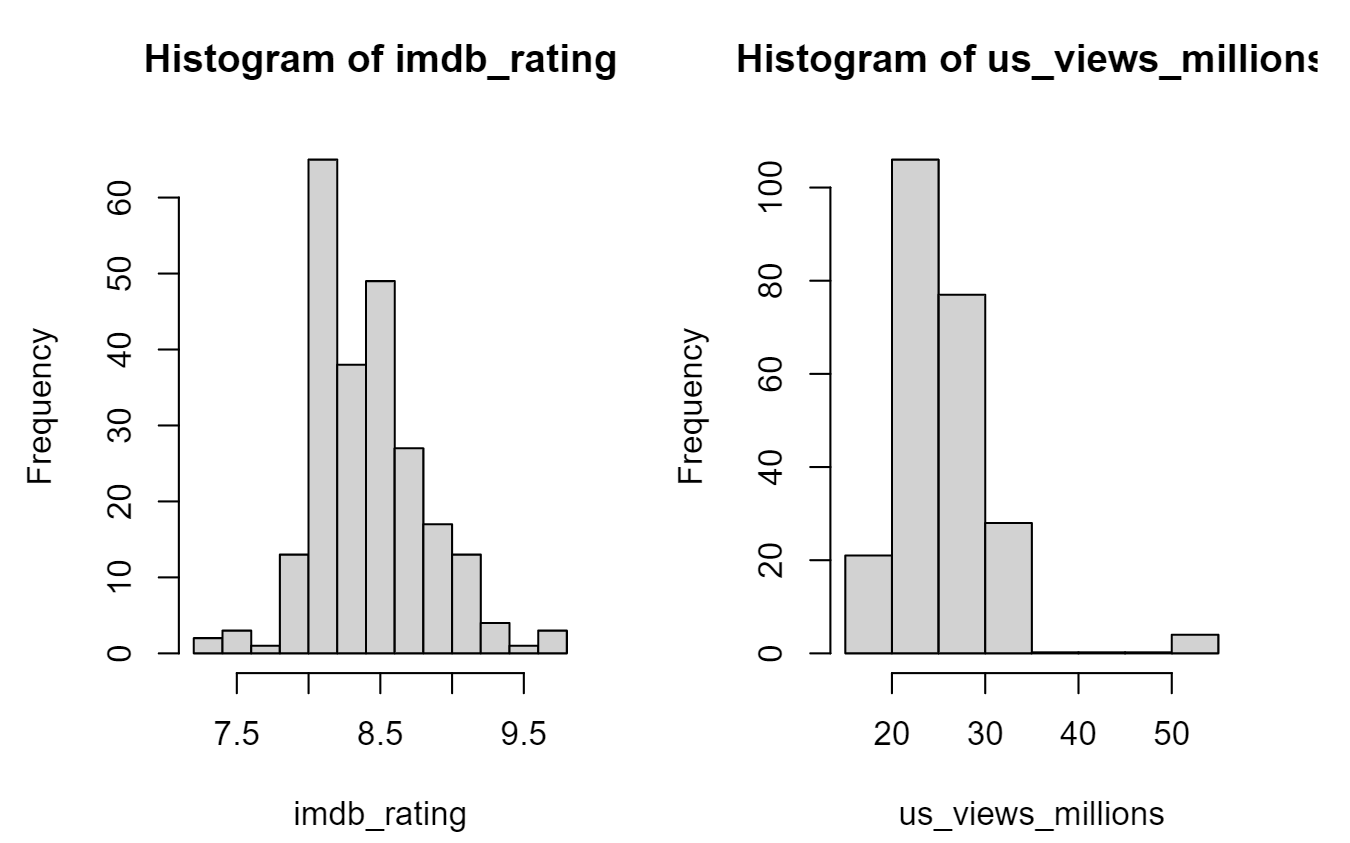
****





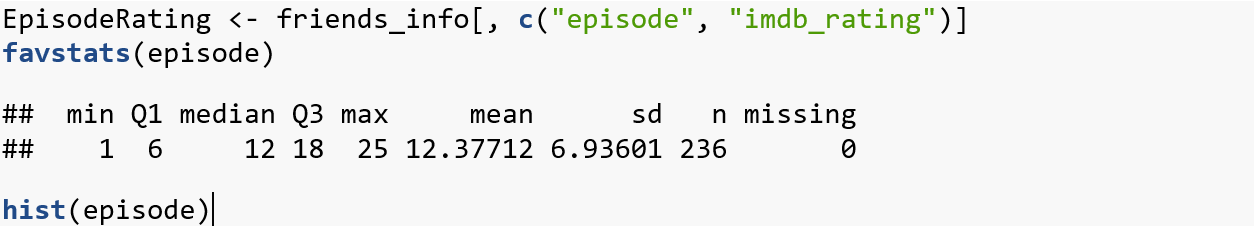


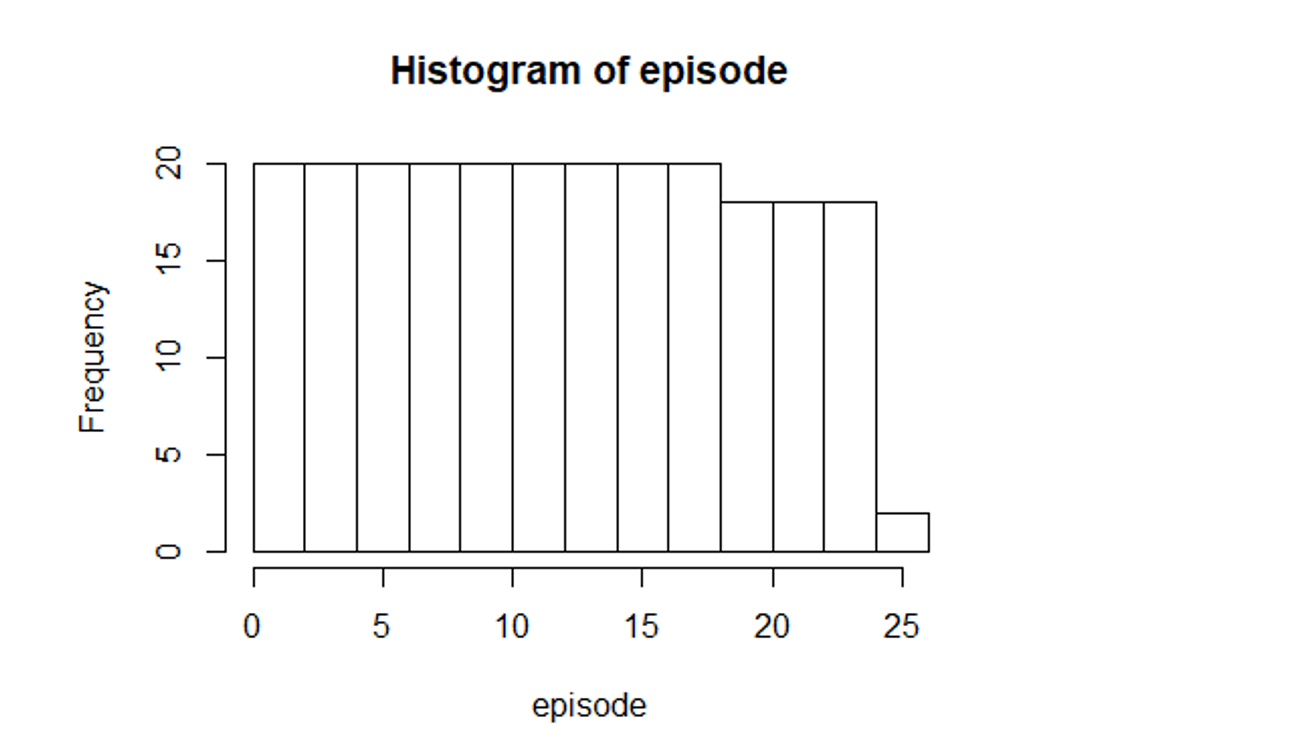




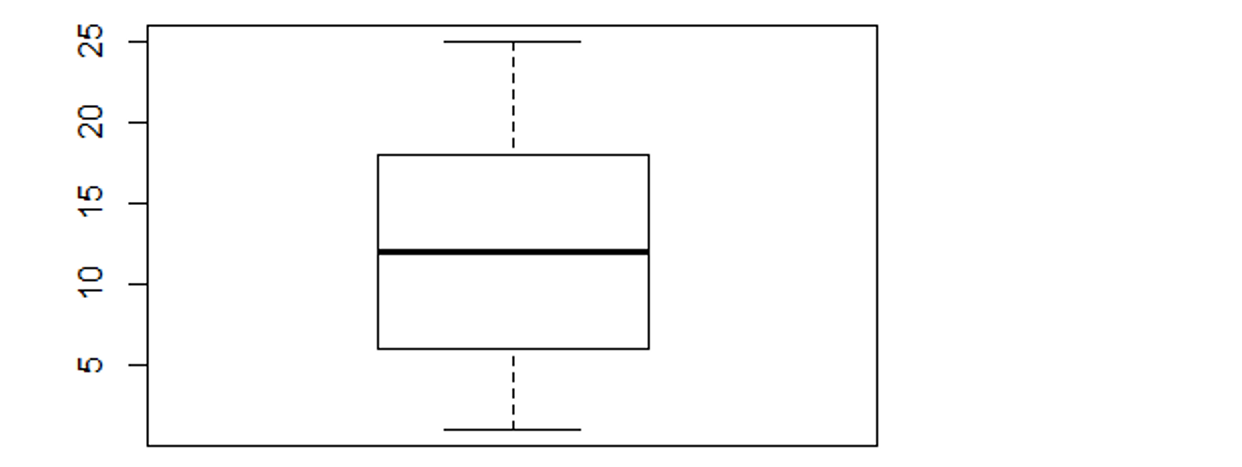
The scatterplot examines the relationship between us\_views\_millions and imdb\_rating. Although the points are a little bit scattered and the relationship is not obvious, the plot still shows a low positive correlation that can be verified by the blue line on the graph. We can conclude that as the imdb\_rating increases, the number of US audiences slightly goes up. Comparing the boxplots of two quantitative variables, us\_views\_millions appears to be skewed to the right, and imdb\_rating shows a relatively symmetric distribution. Both histograms are unimodal; however, imdb\_rating shows a bell-shaped distribution compared with us\_views\_millions. Both datasets have some outliers. The standard deviation of imdb\_rating(0.3986099) is quite smaller than us\_views\_millions(5.231492) since the imdb\_rating follows normality and its average distance to the mean is comparatively small.

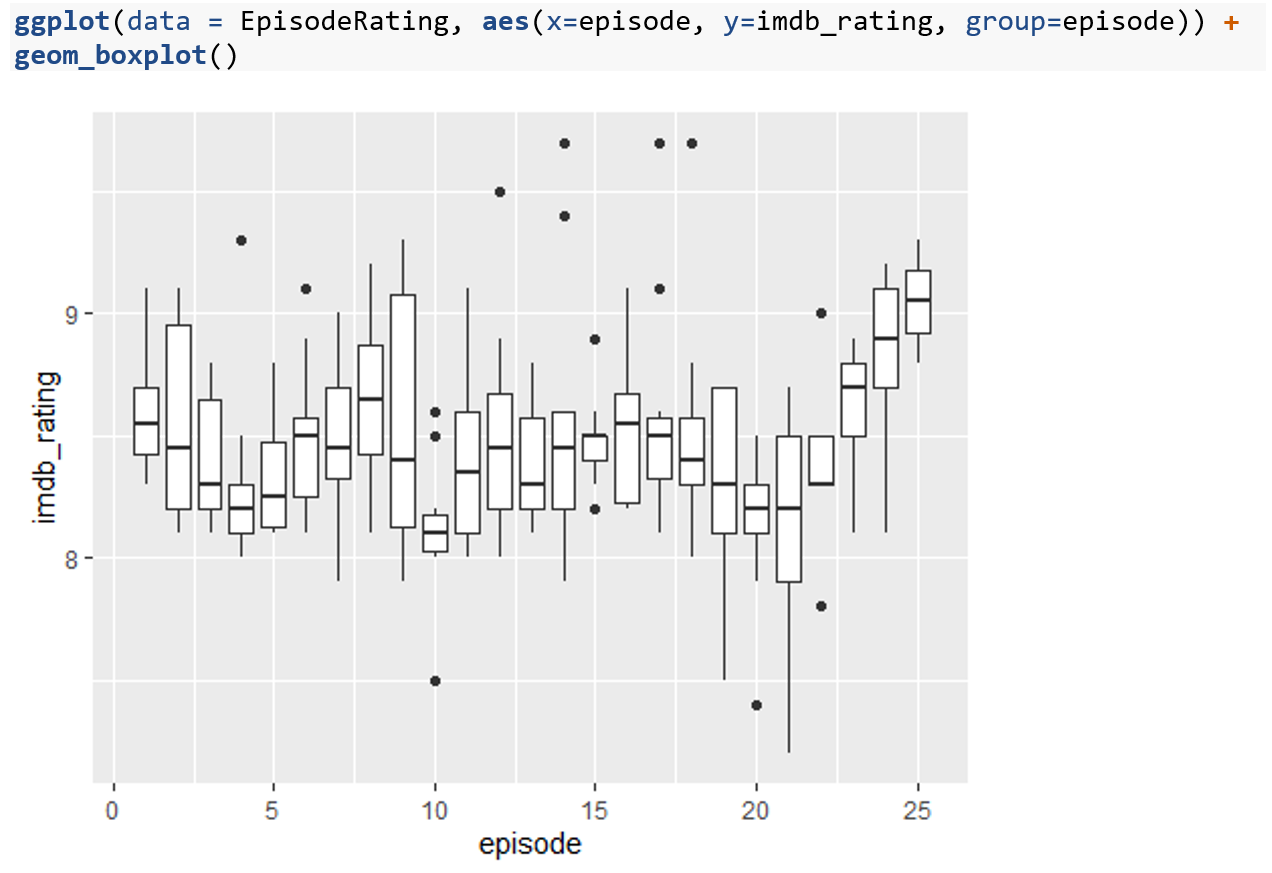
**episode vs imdb\_rating**











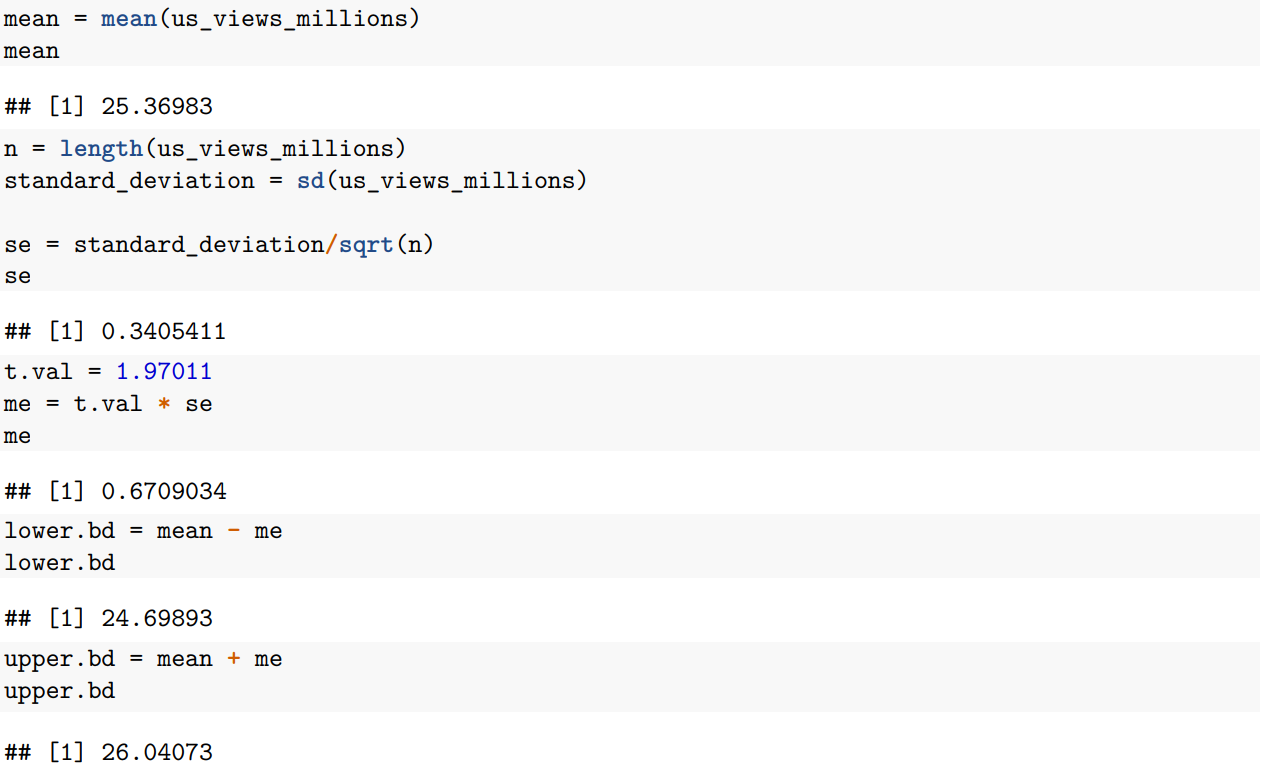
In this case here we have a comparison between two different types of variables. The first variable is episode, which is a categorical variable, and the second is imdb\_rating which is a quantitative variable. Here we see a relationship between these when we separate the imdb\_rating to be for each episode number. As we can see for the episodes they start at 1 and go to 25. By creating a boxplot one for each episode, we can see what the spread is for the rating based on the episode number. By treating the episodes as a categorical variable we are able to see this type of relationship. Some of the distributions are very small as it is with episode 4, while others are larger as in episode 9. Here we can see that there is randomness here, meaning that the episode number does not affect the imdb\_rating.

**3) Describe Interval Estimations**

1. **95% Confidence Interval for Population Mean of us\_views\_millions**

Assumptions:

* Independence: We can assume the number of views when the episode airs are independent from each other.
* Randomization: The data is random as every US resident has an equal chance of being represented in the data.
* Normal Population: The histogram and boxplot display the distribution to be unimodal and approximately symmetric.

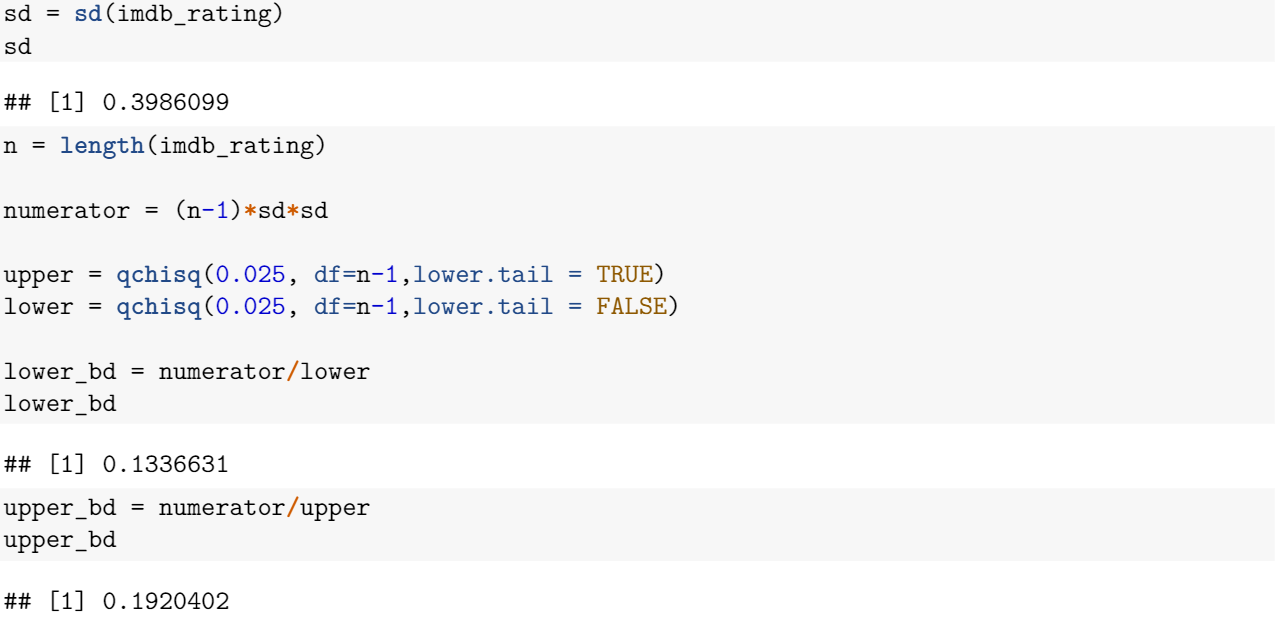


The standard error of 0.3405411 is our estimate of the amount of variation in the sample mean we expect to see when we sample the amount of US viewers, in millions, who watched Friends when the episode aired. The margin of error of 0.6709034 indicates that we are 95% confident that the amount of US viewers, in million, who watched Friends when the episode aired is within 0.6709034 +/- 25.36983 million. Thus, we are 95% confident that the true mean of US residents who watched Friends when it aired is between approximately 24.69893 million and 26.04073 million.

1. **95% Confidence Interval for Population Variance of imdb\_rating**

Assumptions:

* Normality: The histogram of the data is unimodal and approximately bell-shaped.
* Independent Observations/Randomization: The data is random, and thus independent, as every imdb rating has an equal chance of being represented in the data.



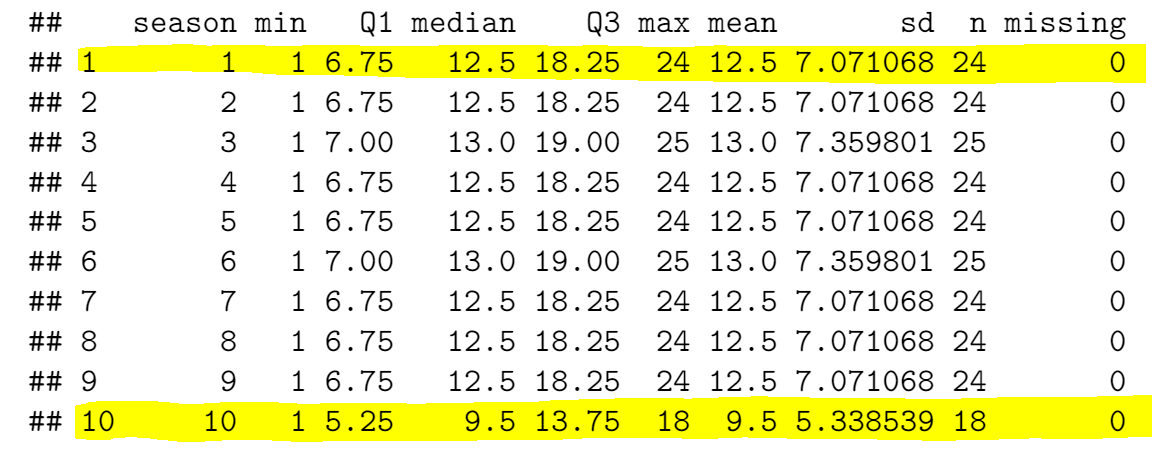
We are 95% confident that the population variance of the imdb\_rating for Friends is between 0.1336631 to 0.1920402 points.

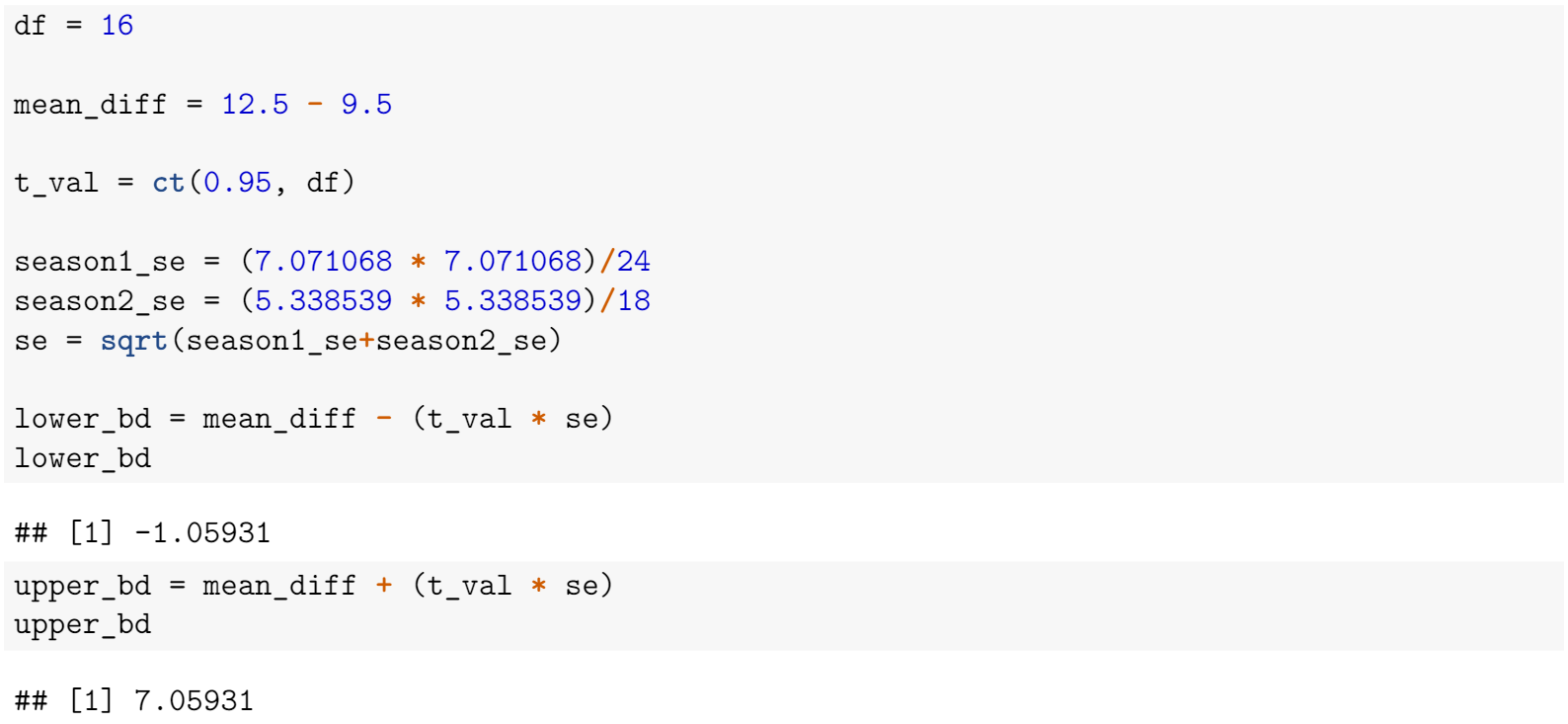
1. **95% Confidence Interval for the Difference Between two Population Means**

Assumptions:

* Independence/Randomization: The data in each group is drawn independently and at random.
* Independent Groups: The number of episodes in season one and in season 10 are independent of each other.
* Normality: Both samples have n ≥ 10.







We are 95% confident that the mean number of episodes for season 1 could be between 1.05931 episodes less than season 10 to 7.05931 episodes more than season 10.