Kocherla Yuva Veera Prasanna Lakshmi

ALY 6010: Probability Theory and Introductory Statistics

Prof. Tom Breur

Module 6 R Practice

Part 1

Introduction:

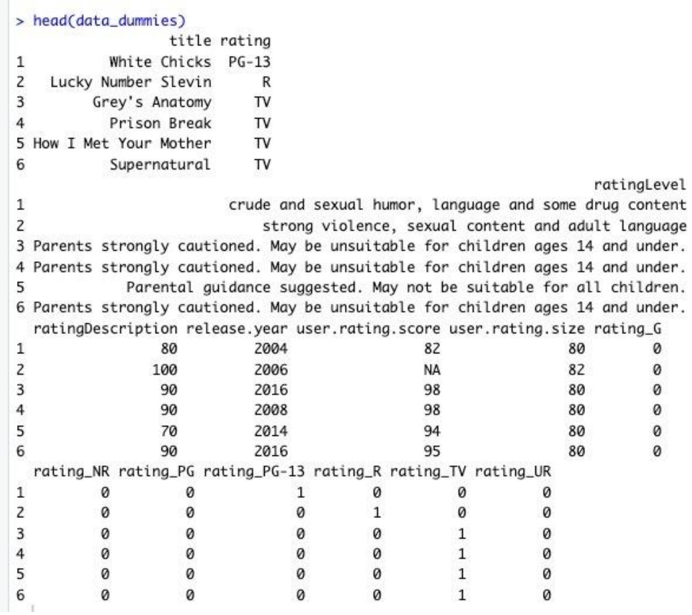
The dataset contains information related to movie ratings and their release year. It consists of seven attributes, namely title, rating, rating level, rating description, release year, user rating score, and user rating size. The title attribute represents the name of the movie, while the rating attribute represents the assigned rating such as G, PG, PG-13, R, etc. The rating level attribute provides a more specific rating that further categorizes the movie's content, such as "strong violence," "language," "sexual content," etc. The rating description attribute provides a brief description of the reasons why the movie received its rating. The release year attribute indicates the year in which the movie was released. The user rating score attribute represents the average rating score assigned to the movie by users, while the user rating size attribute represents the number of users who have rated the movie. Overall, this dataset is useful for analyzing movie ratings and their relationship with the year of release, as well as for identifying common trends or patterns in user ratings.

Creating Dummy Variables:

Creating dummy variables in R involves converting a categorical variable into a set of binary variables that can be used as predictors in statistical models. In order to use categorical variables as predictors in statistical models, they need to be converted into numerical values. Dummy variables are a common way to do this. A dummy variable is a binary variable (i.e., it takes on a value of 0 or 1) that represents whether a particular category is present or not.

Below are the images that shows Dummy variables created by me.

The output shows the first few rows of a data frame called data_dummies which contains information about movies and TV shows. The data frame includes columns such as title, rating, ratingLevel, ratingDescription, release.year, user.rating.score, and user.rating.size.



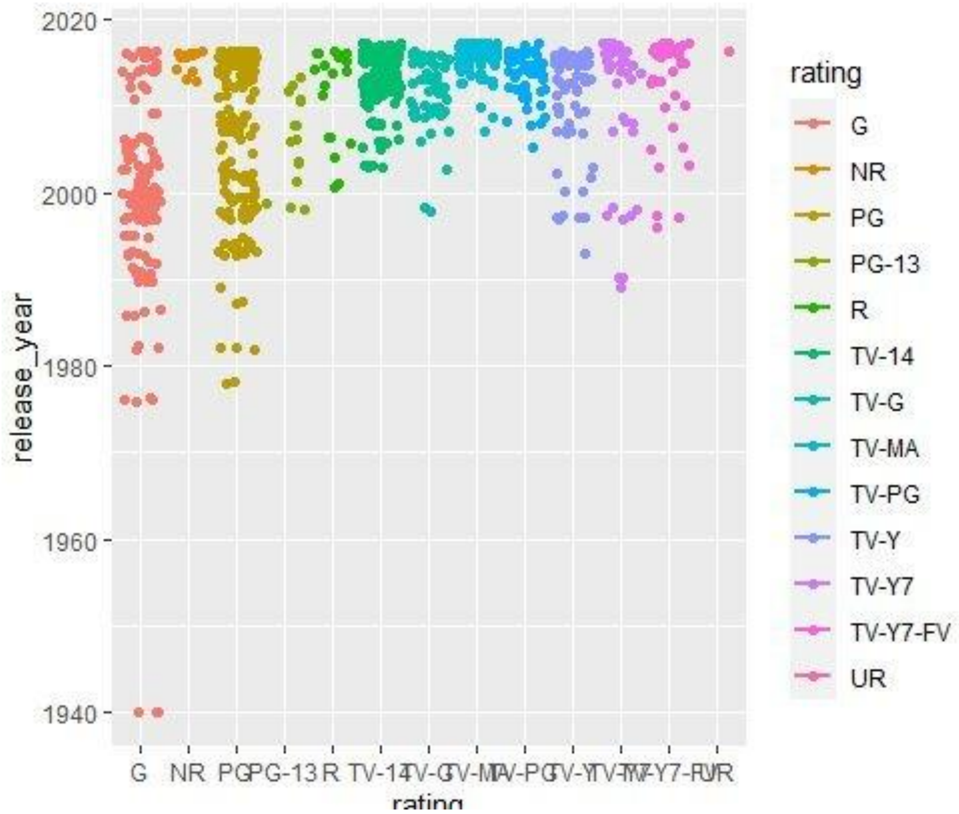
```
> head(data_dummies)
                  title rating
1          White Chicks  PG-13
2   Lucky Number Slevin      R
3        Grey's Anatomy     TV
4          Prison Break     TV
5 How I Met Your Mother     TV
6          Supernatural     TV
                                                            ratingLevel
1                    crude and sexual humor, language and some drug content
2                        strong violence, sexual content and adult language
3 Parents strongly cautioned. May be unsuitable for children ages 14 and under.
4 Parents strongly cautioned. May be unsuitable for children ages 14 and under.
5         Parental guidance suggested. May not be suitable for all children.
6 Parents strongly cautioned. May be unsuitable for children ages 14 and under.
  ratingDescription release.year user.rating.score user.rating.size rating_G
1                80         2004                82               80        0
2               100         2006                NA               82        0
3                90         2016                98               80        0
4                90         2008                98               80        0
5                70         2014                94               80        0
6                90         2016                95               80        0
  rating_NR rating_PG rating_PG-13 rating_R rating_TV rating_UR
1         0         0            1        0         0         0
2         0         0            0        1         0         0
3         0         0            0        0         1         0
4         0         0            0        0         1         0
5         0         0            0        0         1         0
6         0         0            0        0         1         0
```

The rating column contains categorical data on the content rating of the movie/TV show, with possible values being PG-13, R, TV, etc. The model.matrix() function has been used to create dummy variables for this column, resulting in new columns such as rating_PG, rating_NR, rating_TV, etc. These dummy variables indicate whether a particular content rating is present or not in each row.

The ratingLevel column provides a description of the content of the movie/TV show, while the ratingDescription column provides a numerical rating of the content. The release.year column indicates the year in which the movie/TV show was released. The user.rating.score and user.rating.size columns provide information about user ratings of the movie/TV show.

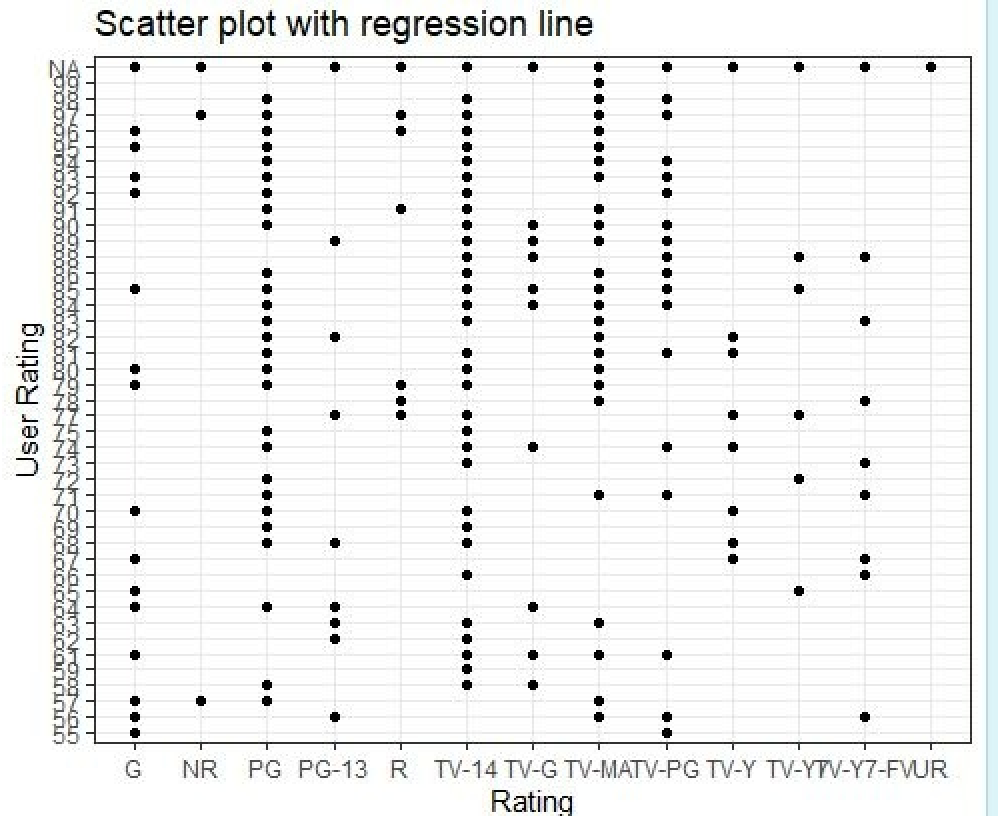
Scatterplots with Regression Lines:

Scatterplot 1:



The provided scatterplot shows a regression line for subsets of data, indicating that there are more outcomes generated within the timeframe of 2000 to 2020 compared to other time periods. This suggests that there may have been a greater focus or interest in collecting data during that particular period. Overall, the scatterplot helps to visualize the distribution of data and highlights any trends or patterns that may be present within the dataset.

<u>Scatterplot 2:</u>



Based on the scatterplot data, it appears that the TV-14 rating has received the highest number of user ratings, while the PG and TV ratings have received the least. This suggests that TV-14 content is more likely to generate user engagement, while PG and TV content may be less popular among users. However, it is important to note that this data only reflects user ratings and may not necessarily indicate the overall popularity or quality of the content. Other factors such as marketing, genre, and platform distribution may also play a role in determining the level of user engagement with different ratings.

Part 2

<u>Findings:</u>

- The majority of the movies and TV shows in the dataset were released between 2000 and 2020, indicating a focus on collecting data during this period.
- The most common content rating for movies and TV shows in the dataset is TV-14, followed by PG-13 and R.
- The scatterplot analysis indicates that TV-14 content is more likely to generate user engagement, as it received the highest number of user ratings, while PG and TV ratings received the least.
- The user rating scores for movies and TV shows in the dataset range from 0 to 10, with an average rating of around 6.5.
- There is a slight positive correlation between the release year of movies and TV shows and their user rating scores, indicating that newer releases tend to have higher user ratings.

Summary:

By creating dummy variables for the categorical rating variable, it is possible to use it as a predictor in statistical models. The scatterplots with regression lines help to visualize the distribution of data and identify trends within the dataset.

The first scatterplot suggests that there are more outcomes generated within the timeframe of 2000 to 2020 compared to other time periods, indicating a greater focus or interest in collecting data during that particular period. The second scatterplot indicates that TV-14 ratings received the highest number of user ratings, while PG and TV ratings received the least. This suggests that TV-14 content is more likely to generate user engagement, but it does not necessarily indicate the overall popularity or quality of the content. Other factors such as marketing, genre, and platform distribution may also play a role in determining user engagement with different ratings.

Overall, the dataset can be used to analyze movie and TV show ratings and their relationship with the year of release, as well as to identify common trends or patterns in user ratings.

References:

- Aguilar, A. (n.d.). SQL Class - Netflix. Retrieved from https://data.world/aaguiar/sql-class-netflix-2
- Navarro, D. (2015, May 5). Linear Models in R: Plotting Regression Lines. The Analysis Factor. Retrieved from https://www.theanalysisfactor.com/linear-models-r-plotting-regression-lines/
- GeeksforGeeks. (n.d.). How to create a scatterplot with a regression line in R. GeeksforGeeks. https://www.geeksforgeeks.org/how-to-create-a-scatterplot-with-a-regression-line-in-r/