

Mock Quiz ADA

4th Trimester

António Fidalgo

Miguel Salema

April 22, 2024

Instructions

- You have 70 minutes to complete your quiz.
- This is an individual exam.
- You may not use the internet. You must have your Wi-Fi disconnected.
- All relevant R code and written answers must be answered in a .qmd file and then rendered to pdf. In case you cannot render to pdf, you may render it to html, but you will have a 5 points discount.
- Upload the full directory of your assignment to Moodle by the indicated deadline as a .zip file. The .zip file should include the .qmd, the .pdf/.html and the data.
- Your code must run on the grader's computer.
- The detection of any form of plagiarism in your work will result in a “fail” grade.
- You must separate all your questions and sub-questions using headers in the .qmd document.
- You can access all class materials but internet access is forbidden.

IMDb TV Shows

To answer this quiz, you must download the file `imdb_top_shows.csv` from Moodle. The dataset contains information about TV shows from the website *IMDb*. The dataset contains the best rated TV shows that have at least 10000 votes. Each observation is a TV show. Table 1 contains the description of the variables. **Assume that you know nothing else about TV shows and draw no conclusions that cannot be drawn from this data alone.**

Table 1: Description of the Variables

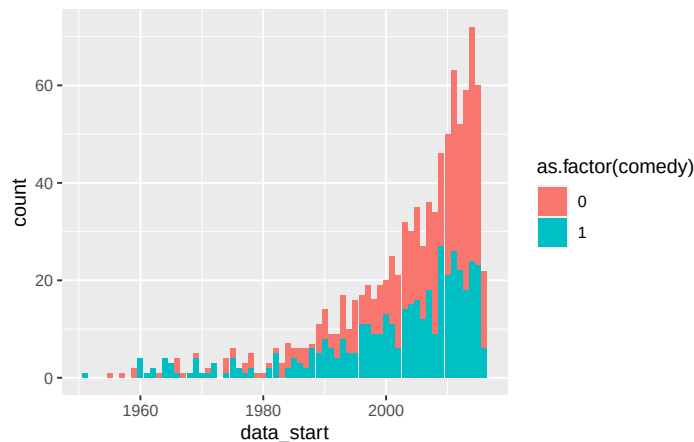
Variable	Description
show_name	Name of the show.
show_rating	Average user rating, from 1 to 10.
show_votes	Number of users that rated the show.
data_start	Year in which the show started.
data_end	Year in which the show ended.
genre	Several dummies indicating the genre.

Questions

1. (20 points) Read the data into R and name the object `df`.
2. (30 points) Copy the following sentence into your Quarto file and, using the dataset `df`, fill the “xxx” using inline R code: “The average rating of a show is xxx, from 0 to 10.”
3. (30 points) Print a dataset containing only the 10 shows with the highest rating. We did it and it looks like this:

```
# A tibble: 10 x 11
  show_name      show_rating show_votes data_start data_end comedy drama crime
  <chr>          <dbl>      <dbl>    <dbl>    <dbl> <dbl> <dbl> <dbl>
1 Breaking Bad      9.5      2130170    2008     2013     0     1     1
2 The Wire          9.3      376456     2002     2008     0     1     1
3 Avatar: The La~    9.3      371137     2005     2008     0     0     0
4 Game of Thrones    9.2     2279488    2011     2019     0     1     0
5 The Sopranos       9.2     469309     1999     2007     0     1     1
6 Sherlock           9.1     996114    2010     2017     0     1     1
7 Rick and Morty      9.1     600020    2013      NA     1     0     0
8 Leyla ile Mecn~     9.1      96370    2011     2023     1     1     0
9 Shingeki no Ky~     9.1     508133    2013     2023     0     0     0
10 Hagane no renk~    9.1     196903    2009     2010     0     0     0
# i 3 more variables: romance <dbl>, action <dbl>, horror <dbl>
```

4. (30 points) How many shows are still ongoing? (Hint: those shows don’t have an end date.)
5. (50 points) Build a bar plot with the number of shows by starting date. Moreover, divide them into comedy and non-comedy shows. We did it and it looks like this:



6. (40 points) What is the effect of being a comedy on the shows’ rating? To answer the question, estimate a linear model while controlling for the log of the number of votes.