Case Study 2 - Analyzing data from MovieLens

Data Science with R

Introduction

Desired outcome of the case study. In this case study we will look at the movies data set from MovieLens. It contains data about users and how they rate movies. The idea is to analyze the data set, make conjectures, support or refute those conjectures with data, and tell a story about the data!

Problem 1: Importing the MovieLens data set and merging it into a single data frame

https://raw.githubusercontent.com/dnchari/DS501_MovieLens/master/Results/unifiedMLDataMulti.csv

```
movielens = 'https://raw.githubusercontent.com/dnchari/DS501_MovieLens/master/Results/unifiedMLDataMulti
mlData = read.csv(movielens)
```

Report some basic details of the data you collected. For example:

- How many movies have an average rating over 4.5 overall?
- How many movies have an average rating over 4.5 among men? How about women?
- How many movies have an median rating over 4.5 among men over age 30? How about women over age 30?
- What are the ten most popular movies?
  - Choose what you consider to be a reasonable defintion of "popular".
  - Be perpared to defend this choice.
- Make some conjectures about how easy various groups are to please? Support your answers with data!
  - For example, one might conjecture that people between the ages of 1 and 10 are the easiest to please since they are all young children. This conjecture may or may not be true, but how would you support or disprove either conclusion with with data?
  - Be sure to come up with your own conjectures and support them with data!

Problem 2: Expand our investigation to histograms

An obvious issue with any inferences drawn from Problem 1 is that we did not consider how many times a movie was rated.

- Plot a histogram of the ratings of all movies.
- Plot a histogram of the number of ratings each movie recieved.
- Plot a histogram of the average rating for each movie.
- Plot a histogram of the average rating for movies which are rated more than 100 times.
  - What do you observe about the tails of the histogram where you use all the movies versus the one where you only use movies rated more than 100 times?
  - Which highly rated movies would you trust are actually good? Those rated more than 100 times or those rated less than 100 times?
- Make some conjectures about the distribution of ratings? Support your answers with data!
  - For example, what age range do you think has more extreme ratings? Do you think children are more or less likely to rate a movie 1 or 5?
  - Be sure to come up with your own conjectures and support them with data!

Problem 3: Correlation: Men versus women

Let us look more closely at the relationship between the pieces of data we have.

- Make a scatter plot of men versus women and their mean rating for every movie.
- Make a scatter plot of men versus women and their mean rating for movies rated more than 200 times.
- Compute the correlation coefficent between the ratings of men and women.
  - What do you observe?
  - Are the ratings similiar or not? Support your answer with data!
- Conjecture under what circumstances the rating given by one gender can be used to predict the rating given by the other gender.
  - For example, are men and women more similar when they are younger or older?
  - Be sure to come up with your own conjectures and support them with data!

Problem 4: Open Ended Question: Business Intelligence

- Do any of your conjectures in Problems 1, 2, and 3 provide insights that a movie company might be interested in?
- Propose a business question that you think this data can answer.
- Suppose you are a Data Sciencetist at a movie company. Convince your boss that your conjecture is correct!

Done

All set!

What do you need to submit?

1. Report: please prepare a report based on what you found in the data.

- What data you collected?
- Why this topic is interesting or important to you? (Motivations)
- How did you analyze the data?
- What did you find in the data? (please include figures or tables in the report)

2. R Code with RMarkdown, compile it to PDF

How to submit: - Submit PDF file on Course Webpage on Canvas only. Do not email it to me.