UTAustinX: UT.7.01x Foundations of Data Analysis

KarenWest ▾

Courseware Course Info Discussion Syllabus Download R and RStudio R Tutorials Readings Contact Us
Progress Office Hours Community

Reflect on the Question Analyze the Data Draw Conclusions

Lab 11: Top Grossing Films

Like most Americans, people in Austin are fascinated with cinema. The American film industry has captured the attention of audiences around the world, making film a multibillion-dollar-a-year industry. Most of the top-grossing films of all times have been produced by the same five major studios: 20th-Century Fox, Paramount, Sony Pictures, Universal Pictures and Warner Bros. This data set focuses on the 151 films made by these studios that made the list of the 245 top-grossing films of all times, as determined by authoritative source Box Office Mojo. For each of the films, data includes film genre, MPAA rating, measures of film critic and user rankings, and production outcomes such as budget, time in theaters and amount grossed.

Primary Research Questions

1. Does a film’s rating (PG, PG-13, or R) impact its cost to produce?
2. Does a film’s rating (PG, PG-13, or R) influence its IMDB score?

(3 points possible)

Check the Data

Let's begin by examining our data in R.

1. Open RStudio. Make sure you've installed the SDSFoundations package.
2. Type **`library(SDSFoundations)`** This will automatically load the data for the labs.

3.Type **`film <- FilmData`** This will assign the data to your Workspace.

4. Look at the spreadsheet view of the data to answer the following questions.

Alternatively, you can use follow the steps in the "Importing a Data Frame" R tutorial video, and use the FilmData.csv file. (Right-click and "Save As.") Make sure to **name** the dataframe "film" when importing.

1. Open RStudio.
2. Click on "Import Dataset" button at the top of the workspace window. Choose *"from text file."*

3. Click on the location of the FilmData.csv file you just downloaded.
4. Click on the FilmData.csv file. Then, click Upload.
5. Look at the spreadsheet view of the data to answer the following questions.

1a. What rank does *Titanic* hold among highest-grossing films?

Answer: 2

1b. What is the name of the highest ranked film made by Universal Studios?

Answer: Jurassic Park

1c. What was the lowest IMDB rating given to a film that ranked in the top 10 grossing films of all time?

Answer: 7.6

Hide Answer *You have used 0 of 2 submissions*

(6 points possible)

Check the Variables of Interest

Help

2a. Which variable tells us whether a film is **rated** PG, PG-13 or R?

The variable name in the dataset is □ Rating , which is a □ categorical variable.

2b. Which variable tells us the average score the film received on **IMDB**?

The variable name in the dataset is □ IMDB , which is a □ quantitative variable.

2c. Which variable tells us how much it **cost** to produce the film?

The variable name in the dataset is □ Budget , which is a □ quantitative variable.

Hide Answer *You have used 0 of 2 submissions*

(2 points possible)

Reflect on the Method

Which method should we be using for this analysis and why?

3a. We will use **ANOVA** to help us answer each of these questions. Why?

○ We want to determine if the category to which a film belongs has an impact on some other quantitative measure. ✓

○ We believe the difference in scores given by film critics and IMDB users is significant.

○ We cannot compare the means of these groups using any other method because the groups are related.

Help

CORRECT. AN ANOVA IS USEFUL TO DETERMINE WHETHER THE MEANS FOR 3 OR MORE GROUPS ARE EQUAL IN THE POPULATION. HERE, THE OBSERVATIONS ARE FILMS GROUPED BY RATING, AND WE WANT TO KNOW WHETHER PRODUCTION COST OR IMDB SCORE ARE RELATED TO THESE GROUPING. IF THEY ARE, THE MEANS FOR THE GROUPS WOULD BE STATISTICALLY DIFFERENT.

3b. We will conduct **post-hoc tests**, specifically Tukey’s HSD, if the result of either ANOVA is significant. Why?

○ We want to confirm that the assumptions of ANOVA have been met.

○ We want to locate which group means are different from each other. ✓

○ We want to determine the average score for each category.

CORRECT. THE ANOVA TEST WILL SIMPLY TELL US WHETHER OR NOT THE MEANS ARE EQUAL. IF THEY DO NOT SEEM TO BE EQUAL BASED ON OUR ANOVA RESULTS, THEN WE NEED TO CONDUCT POST-HOC TESTS IN ORDER TO DETERMINE WHICH AND HOW THE MEANS DIFFER FROM ONE ANOTHER.

Hide Answer *You have used 0 of 2 submissions*

Help

EdX offers interactive online classes and MOOCs from the world's best universities. Online courses from MITx, HarvardX, BerkeleyX, UTx and many other universities. Topics include biology, business, chemistry, computer science, economics, finance, electronics, engineering, food and nutrition, history, humanities, law, literature, math, medicine, music, philosophy, physics, science, statistics and more. EdX is a non-profit online initiative created by founding partners Harvard and MIT.

© 2015 edX Inc.

EdX, Open edX, and the edX and Open edX logos are registered trademarks or trademarks of edX Inc.

Terms of Service and Honor Code

Privacy Policy (Revised 10/22/2014)

About edX

About

News

Contact

FAQ

edX Blog

Donate to edX

Jobs at edX

Follow Us

Twitter

Facebook

Meetup

LinkedIn

Google+