Data analysis project 1:

Hypothesis testing of movie ratings data

Mission command preamble: As in general, we won't tell you how to do something. That is up to you and your creative problem solving skills. However, we will tell you what we would like you to do. One exception: We do expect you to do this work yourself, so it reflects your intellectual contribution.

Purpose: In this project, you will demonstrate the essential skills involved in hypothesis testing. To do so, we will use a real dataset that stems from a replication attempt of published research (Wallisch & Whritner, 2017). Please write a report (1-2 pages, as needed) that answers all the questions below. You can pretend you are working for a major movie studio that needs to answer these questions in order to optimize their operations. You can use figures as needed to buttress/illustrate your argument.
Note that you will need to do a lot of tests (of your choice, as appropriate) to answer these questions, so to cut down on false positives, set the per-test significance level α to 0.005 (as per Benjamin et al., 2018).

Dataset description: This dataset features ratings data of 400 movies from 1097 research participants.
1st row: Headers (Movie titles/questions) – note that the indexing in this list is from 1
Row 2-1098: Responses from individual participants
Columns 1-400: These columns contain the ratings for the 400 movies (0 to 4, and missing)
Columns 401-421: These columns contain self-assessments on sensation seeking behaviors (1-5)
Columns 422-464: These columns contain responses to personality questions (1-5)
Columns 465-474: These columns contain self-reported movie experience ratings (1-5)
Column 475: Gender identity (1 = female, 2 = male, 3 = self-described)
Column 476: Only child (1 = yes, 0 = no, -1 = no response)
Column 477: Movies are best enjoyed alone (1 = yes, 0 = no, -1 = no response)
Note that we did most of the data munging for you already (e.g. Python interprets commas in a csv file as separators, so we removed all commas from movie titles), but you still need to handle missing data.

Questions corporate would like you to answer in the report (each is worth 10% of the grade score):
1) Are movies that are more popular (operationalized as having more ratings) rated higher than movies that are less popular? [*Hint: You can do a median-split of popularity to determine high vs. low popularity movies*]
2) Are movies that are newer rated differently than movies that are older? [*Hint: Do a median split of year of release to contrast movies in terms of whether they are old or new*]
3) Is enjoyment of 'Shrek (2001)' gendered, i.e. do male and female viewers rate it differently?
4) What proportion of movies are rated differently by male and female viewers?
5) Do people who are only children enjoy 'The Lion King (1994)' more than people with siblings?
6) What proportion of movies exhibit an "only child effect", i.e. are rated different by viewers with siblings vs. those without?
7) Do people who like to watch movies socially enjoy 'The Wolf of Wall Street (2013)' more than those who prefer to watch them alone?
8) What proportion of movies exhibit such a "social watching" effect?
9) Is the ratings distribution of 'Home Alone (1990)' different than that of 'Finding Nemo (2003)'?
10) There are ratings on movies from several franchises (['Star Wars', 'Harry Potter', 'The Matrix', 'Indiana Jones', 'Jurassic Park', 'Pirates of the Caribbean', 'Toy Story', 'Batman']) in this dataset. How many of these are of inconsistent quality, as experienced by viewers? [Hint: You can use the keywords in quotation marks featured in this question to identify the movies that are part of each franchise]
Extra Credit: Tell us something interesting and true (supported by a significance test of some kind) about the movies in this dataset that is not already covered by the questions above [for 5% of the grade score].