data analysis project

November 7, 2022

Data Analysis Project 1

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]

[ ]:
```
import pandas as pd
import numpy as np
from scipy import stats
df = pd.read_csv('movieReplicationSet.csv')
```

[ ]:
```
df.head()
```

[ ]:
```
   The Life of David Gale (2003)  Wing Commander (1999)  \
0                            NaN                    NaN
1                            NaN                    NaN
2                            NaN                    NaN
3                            NaN                    NaN
4                            NaN                    NaN

   Django Unchained (2012)  Alien (1979)  \
0                      4.0           NaN
1                      1.5           NaN
2                      NaN           NaN
3                      2.0           NaN
4                      3.5           NaN

   Indiana Jones and the Last Crusade (1989)  Snatch (2000)  \
0                                        3.0            NaN
1                                        NaN            NaN
2                                        NaN            NaN
3                                        3.0            NaN
4                                        0.5            NaN

   Rambo: First Blood Part II (1985)  Fargo (1996)  \
0                                NaN           NaN
1                                NaN           NaN
2                                NaN           NaN
3                                NaN           NaN
```

```
4                                0.5           1.0

   Let the Right One In (2008)  Black Swan (2010)  …  \
0                          NaN               NaN  …
1                          NaN               NaN  …
2                          NaN               NaN  …
3                          NaN               4.0  …
4                          NaN               0.0  …

   When watching a movie I cheer or shout or talk or curse at the screen  \
0                                                  1.0
1                                                  3.0
2                                                  5.0
3                                                  3.0
4                                                  2.0

   When watching a movie I feel like the things on the screen are happening to
me  \
0                                                  6.0
1                                                  1.0
2                                                  4.0
3                                                  1.0
4                                                  3.0

   As a movie unfolds I start to have problems keeping track of events that
happened earlier  \
0                                                  2.0
1                                                  1.0
2                                                  3.0
3                                                  1.0
4                                                  2.0

   The emotions on the screen "rub off" on me - for instance if something sad is
happening I get sad or if something frightening is happening I get scared  \
0                                                  5.0
1                                                  6.0
2                                                  5.0
3                                                  4.0
4                                                  5.0

   When watching a movie I get completely immersed in the alternative reality of
the film  \
0                                                  5.0
1                                                  5.0
2                                                  5.0
3                                                  5.0
4                                                  6.0
```

```
   Movies change my position on social economic or political issues  \
0                                                5.0
1                                                3.0
2                                                4.0
3                                                3.0
4                                                4.0

   When watching movies things get so intense that I have to stop watching  \
0                                                1.0
1                                                2.0
2                                                4.0
3                                                1.0
4                                                4.0

   Gender identity (1 = female; 2 = male; 3 = self-described)  \
0                                                1.0
1                                                1.0
2                                                1.0
3                                                1.0
4                                                1.0

   Are you an only child? (1: Yes; 0: No; -1: Did not respond)  \
0                                                  0
1                                                  0
2                                                  1
3                                                  0
4                                                  1

   Movies are best enjoyed alone (1: Yes; 0: No; -1: Did not respond)
0                                                  1
1                                                  0
2                                                  0
3                                                  1
4                                                  1

[5 rows x 477 columns]
```

[ ]:
```
df1 = df.iloc[:,:400]
pop = df1.loc[:,df1.isna().sum() < df1.isna().sum().median()]
not_pop = df1.loc[:,df1.isna().sum() >= df1.isna().sum().median()]
```

[ ]:
```
size = pop.shape[1]
array_pop = np.empty(0)
for i in range(size):
    a = pop.iloc[:,i].dropna()
    array_pop = np.concatenate((array_pop, a), axis=None)
```

```
array_not_pop = np.empty(0)
for i in range(400-size):
    b = not_pop.iloc[:,i].dropna()
    array_not_pop = np.concatenate((array_not_pop, b), axis=None)
# simple summary
# use mannwhitney u test
u,p = stats.mannwhitneyu(array_pop, array_not_pop, alternative="greater")
print(u, p)
```

```
1242808144.5 0.0
```

I first do a median split of number of ratings to determine high vs low popularity movies. Since the movie ratings are meaningless if reduced to mean and we are testing two sample, I used the Mann Whiteney U test instead of t-test to test whether there is a difference in frequency of ratings between high popularity movies and low popularity moviews. For question 1 to 8, I used the Mann Whitney U test for the same reason, but according to different test circumstance I used different hypothesis. For question 1, the null hypothesis is that high popularity movies would are not higher than low popularity movies. Under $\alpha = 0.05$ We get a p-value that is much less than 0.05, so there is evidence that popular movies have higher ratings than movies that are not that popular.

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]

[ ]:
```
# select movie names
df1.columns[0]
list = []
# get the release year from the name
for i in range(400):
    list.append(int(df1.columns[i][-5:-1]))
median_year = np.median(list)
```

[ ]:
```
# split by release year
newer_movies = df1.iloc[:,list >= median_year]
older_movies = df1.iloc[:,list < median_year]
size = newer_movies.shape[1]
array_newer = np.empty(0)
for i in range(size):
    a = newer_movies.iloc[:,i].dropna()
    array_newer = np.concatenate((array_newer, a), axis=None)
array_older = np.empty(0)
for i in range(400-size):
    b = older_movies.iloc[:,i].dropna()
    array_older = np.concatenate((array_older, b), axis=None)
```

[ ]:
```
u,p = stats.mannwhitneyu(array_older, array_newer, alternative="two-sided")
p
```

[ ]: 1.2849216001533932e-06

I also do a median split to determine older, newer movies. The median of release year of movies in this dataset is 1999, so I splited them, into movie released before 1999 as the group of older movies and movies released after (including) 1999 as the group of newer movies. Again, for movie ratings, it is more reasonable to use a Mannu Whitney U test to detect difference in frequency than difference in mean. The null hypothesis of this test is that ratings of older movies and newer movies are the same. The p value is really small, at 0.05 significance level, we rejected the null hypothesis that movies released in recent years are rated the same as movies released in earlier.

3) Is enjoyment of 'Shrek (2001)' gendered, i.e. do male and female viewers rate it differently?

[ ]:
```
# find the gender column
df.iloc[:,-3].head()
```

[ ]:
```
0    1.0
1    1.0
2    1.0
3    1.0
4    1.0
Name: Gender identity (1 = female; 2 = male; 3 = self-described), dtype: float64
```

[ ]:
```
female = df.loc[df.iloc[:,-3] == 1, :].loc[:,"Shrek (2001)"].dropna()
male = df.loc[df.iloc[:,-3] == 2, :].loc[:,"Shrek (2001)"].dropna()
```

[ ]:
```
u,p = stats.mannwhitneyu(female, male, alternative="two-sided")
p
```

[ ]:
```
0.050536625925559006
```

I compared the Shrek ratings by male and female using the Mannu Whitney U test. Since we want to test whether the two group of rating has difference, the null hypothesis is that males rating and females rating are the same (no difference). At 0.05 significance level, we failed to rejecte that Shrek (2001) is rated differently by male and female. But it is hard to say, if we pick larger α (i.e. $\alpha = 0.1$), we would not reject null hypothesis. So we should not believe enjoyment of Shrek is gendered.

4) What proportion of movies are rated differently by male and female viewers?

[ ]:
```
sum = 0
for i in range(400):
    female = df.loc[df.iloc[:,-3] == 1, :].iloc[:,i].dropna()
    male = df.loc[df.iloc[:,-3] == 2, :].iloc[:,i].dropna()
    u,p = stats.mannwhitneyu(female, male, alternative="two-sided")
    if p < 0.05:
        sum += 1
str(sum/400*100) + '%'
```

[ ]:
```
'30.5%'
```

Using the Mann Whitney U test for all the movies, under the null that movies are rated the same by male and female with 0.05 significance level, there are 30.5% of the movies have a p-value that

is less than 0.05, so approximately 30.5% of the movies are rated differently by male and female.

5) Do people who are only children enjoy 'The Lion King (1994)' more than people with siblings?

```
[ ]: only_child = df.loc[df.iloc[:,-2] == 1, :].loc[:,"The Lion King (1994)"].
      ↪dropna()
     not_only_child = df.loc[df.iloc[:,-2] == 0, :].loc[:,"The Lion King (1994)"].
      ↪dropna()
```

```
[ ]: u,p = stats.mannwhitneyu(only_child, not_only_child, alternative="greater")
     p
```

```
[ ]: 0.978419092554931
```

I compared The Lion King 1994 ratings by who are single child and those who with siblings using the Mannu Whitney U test. The null hypothesis is that rating by single child are not greater (i.e. less or equal) than rating by people with siblings and the alternative hypothsis is that single child enjoy the movie more. At 0.05 significance level, we fail reject the null hypothesis (p = 0.978419092554931). As a result, we have no evidence that rating of only child is greater.

6) What proportion of movies exhibit an "only child effect", i.e. are rated different by viewers with siblings vs. those without?

```
[ ]: sum = 0
     for i in range(400):
         only_child  = df.loc[df.iloc[:,-2] == 1, :].iloc[:,i].dropna()
         not_only_child = df.loc[df.iloc[:,-2] == 0, :].iloc[:,i].dropna()
         u,p = stats.mannwhitneyu(only_child, not_only_child,alternative="two-sided")
         if p < 0.05:
             sum += 1
     sum/400
```

```
[ ]: 0.1
```

Using the same test and hypothesis for question 6, and under 0.05 significance level, we find that 10% of the cases we generated a p-value that is less than 0.05. So under $\alpha = 0.05$, there are 10% of movies have "only child effect".

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?

```
[ ]: alone = df.loc[df.iloc[:,-1] == 1, :].loc[:,"The Wolf of Wall Street (2013)"].
      ↪dropna()
     socially = df.loc[df.iloc[:,-1] == 0, :].loc[:,"The Wolf of Wall Street␣
      ↪(2013)"].dropna()
     u,p = stats.mannwhitneyu(socially, alone, alternative="greater")
     p
```

```
[ ]: 0.9436657996253056
```

Since it is comparing I filtered out who enjoy watching movie socially and who enjoy watching alone and used the Mann Whitney U test to test whether the ratings of who like to watch movies socially is higher than those who like to watch movies alone. The null hypothesis is that the rating of who like to watch socially is less or equal to the rating of who like to watch moview alone. The p-value of this test is 0.9436657996253056, which fail to reject the null under 0.05 significance level. We have no evidence to believe 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?

[ ]:
```
sum = 0
for i in range(400):
    alone  = df.loc[df.iloc[:,-1] == 1, :].iloc[:,i].dropna()
    socially = df.loc[df.iloc[:,-1] == 0, :].iloc[:,i].dropna()
    u,p = stats.mannwhitneyu(alone, socially,alternative="two-sided")
    if p < 0.05:
        sum += 1
sum/400
```

[ ]:
```
0.0825
```

By performing the same test in question 7 on every movie in the dataset, 8.25% of the movies generated a p-value less than 0.05. So under the 0.05 significance level, 8.25% of the movies exhibit such a "social watching" effect.

9) Is the ratings distribution of 'Home Alone (1990)' different than that of 'Finding Nemo (2003)'?

[ ]:
```
home_alone = df.loc[:,"Home Alone (1990)"].dropna()
finding_nemo = df.loc[:,"Finding Nemo (2003)"].dropna()
d, p = stats.ks_2samp(home_alone, finding_nemo)
p
```

[ ]:
```
6.379397182836346e-10
```

This is still two sample test, since we want to compare whether the two distributions where these two samples came from are the same, we could use the KS(Kolmogorov-Smirnov) test, it is a kind of goodness-of-fit test. I filtered out the movie ratings of Home Alone and Finding Nemo as the two samples, the p-value is printed above, which failed to reject the null hypothesis that the two distribution are the same, under 0.05 significance level. So we have evidence to say 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]

[ ]:
```
def func(s,x):
    return s in x
```

```
franchises = ["Star Wars", "Harry Potter", "The Matrix", "Indiana Jones",␣
 ↪"Jurassic Park", "Pirates of the Caribbean", "Toy Story", "Batman"]
list = [[False]*len(df1.columns) for i in range(len(franchises))]
for j in range(len(franchises)):
    for i in range(len(df1.columns)):
        list[j][i] = (func(franchises[j], df1.columns[i]))
```

[ ]:
```
# Star Wars
star_wars = df1.loc[:,list[0]].dropna()
h,p = stats.kruskal(star_wars.iloc[:,0],
star_wars.iloc[:,1],
star_wars.iloc[:,2],
star_wars.iloc[:,3],
star_wars.iloc[:,4],
star_wars.iloc[:,5])
p
```

[ ]:
```
6.940162236984522e-40
```

[ ]:
```
# Harry Potter
harry_potter = df1.loc[:,list[1]].dropna()
h,p = stats.kruskal(harry_potter.iloc[:,0],
harry_potter.iloc[:,1],
harry_potter.iloc[:,2],
harry_potter.iloc[:,3])
p
```

[ ]:
```
0.11790622831256074
```

[ ]:
```
# The Matrix
the_matrix = df1.loc[:,list[2]].dropna()
h,p = stats.kruskal(the_matrix.iloc[:,0],
the_matrix.iloc[:,1],
the_matrix.iloc[:,2])
p
```

[ ]:
```
1.7537323830838066e-09
```

[ ]:
```
# Indiana Jones
indiana = df1.loc[:,list[3]].dropna()
h,p = stats.kruskal(indiana.iloc[:,0],
indiana.iloc[:,1],
indiana.iloc[:,2],
indiana.iloc[:,3])
p
```

[ ]:
```
1.020118354785894e-11
```

[ ]:
```
# Jurassic Park
jurassic = df1.loc[:,list[4]].dropna()
h,p = stats.kruskal(jurassic.iloc[:,0],
jurassic.iloc[:,1],
jurassic.iloc[:,2])
p
```

[ ]:
```
1.8492328391686058e-11
```

[ ]:
```
# Toy Story
pirates = df1.loc[:,list[5]].dropna()
h,p = stats.kruskal(pirates.iloc[:,0],
pirates.iloc[:,1],
pirates.iloc[:,2])
p
```

[ ]:
```
0.035792727694248905
```

[ ]:
```
# Pirates of the Caribbean
toy = df1.loc[:,list[6]].dropna()
h,p = stats.kruskal(toy.iloc[:,0],
toy.iloc[:,1],
toy.iloc[:,2])
p
```

[ ]:
```
7.902234665149812e-06
```

[ ]:
```
# Batman
batman = df1.loc[:,list[7]].dropna()
h,p = stats.kruskal(batman.iloc[:,0],
batman.iloc[:,1],
batman.iloc[:,2])
p
```

[ ]:
```
4.1380499020034183e-19
```

I first filtered out all the franchise movies, and there are more than two movies for each franchise, so we cannot use Mann-Whitney again. And we cannot use normal one-way anova, since the data is movie rating. I found that Kruskal-Wallis H-test is a non-parametric version of ANOVA, which is to test the median for 2 or more samples, which may have different sample sizes. I also find the Friedman test is a appriopriate in this circumstance but it does not support different sample size for each sample, so I finally choose Kruskal-Wallis H test. The null hypothesis of the test is that all movies in the franchise have consistent quality (the median of the ratings have no difference). Using the KW H-test on the 8 franchise movies, I found that only the Harry Potter Series generated a p-value higher than 0.05, that is saying, under 0.05 significance level, We fail to reject that Harry Potter Series has consistent quality, while for the other 7 franchise movies, we rejeceted that their quality is consistent. Inconclusion, 7 out of 8 franchise movies have inconsistent quanlity.

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].

I tried to find out whether newer released moview are more popular than older movies (have more ratings). I would use the median split as in question 2. For this two sample circumstance, I would use the two-sample t-test, since I only care about count of rating, so I only need to test whether there is a difference in the mean of counts. And I assume the rating counts are independent, so I would use the independent t-test

[ ]:
```
rows = newer_movies.shape[0]
newer_movie_rating_counts = []
older_movie_rating_counts = []
for i in range(newer_movies.shape[1]):
    newer_movie_rating_counts.append(rows - newer_movies.iloc[:,i].isna().sum())
for i in range(older_movies.shape[1]):
    older_movie_rating_counts.append(rows - older_movies.iloc[:,i].isna().sum())
t1,p1 = stats.ttest_ind(newer_movie_rating_counts, older_movie_rating_counts,␣
 ↪alternative="greater")
p1
```

[ ]:
```
4.2443663905911793e-05
```

By spliting the movie ratings into those before 1999 and after 1999, I used the independent t-test and rejected the null that the count of newer movie ratings are less or equal to the count of older movie ratings under 0.05 significance level. That it is to say, we have evidence to believe that newer movies are more popular, or younger generation movie viewers are more likely to give ratings of movie.