Problem 3 – hypothesis testing (30%)

Next, you are to generate various hypotheses about the data, and perform statistical tests of those hypotheses. For example, you may hypothesize that there is a positive correlation between the budget of a movie and the revenue it generated. We expect you will generate 4-6 hypotheses, though the actual number is not as important as what those hypotheses are testing (i.e., are they meaningful vs. non-sensical).

You can use any of the hypothesis testing techniques presented in the lectures (and reading lists.)

Hypothesis 1: Movies released in recent years (i.e. since 2000) have lower average Metascore ratings.

1) Make observations on the data.
    a. In Part 1 and Part 2 of the homework, I noticed that in the scatter plot between Year of Release and Metascore, there appeared to be an inverse relationship between the two variables. Though it does look like movies always tended to do well, as this is the top 1000 movies in IMDB, I noted that the "minimum" lowest Metascore per year appeared to be on a trending decrease, with lower Metascore ratings becoming more common as time progressed. Because of this, I began to think: Could it be that movies released in recent years have lower average Metascore ratings?
2) Conduct background research
    a. When doing readings to see what currently existed on the topic, I came across this article on Medium that essentially argued that older movies get better ratings due to the principle of Quantity vs. Quality. The article makes the claim that "more and more movies are being made each year on average, and most of them are bad". I wanted to see if this statement done by Medium also applied to that of the dataset of the top 1000 movies on IMDB, and if more recent movies have lower average Metascore ratings.
3) Pose a question.
    a. Do movies released in recent years have lower average Metascore ratings than that of older movies. For the sake of this question, we will be comparing movies released in recent years (since 2000) to those released prior to 2000.
4) Run Experiment.
    a. For the sake of this experiment, we do a t test for two independent samples of scores (aka scipy.stats.ttest_ind).

    H_0: Null hypothesis: There is no significant difference between the Metascores of movies released in recent years and that of Metascores of movies released in prior years.

***H_a*: Alternative Hypothesis**: The mean of movie Metascores released in recent years is greater than the movie Metascores of movies released in prior years.

We first took a sample size of 30 from the set of movies released past 2000 and the set of movies released prior to 2000. Then, we ran a right tailed unpaired two sample t-test to see if the population mean of movie Metascores for movies released past 2000 is greater than that of Metascores for movies prior to 2000.

```
In [74]: #sort into recent and older movies
recent_movies = df[df['Year of Release'] >= 2000]['Metascore of movie']
older_movies = df[df['Year of Release'] < 2000]['Metascore of movie']

sample_size = 30
recent_sample = np.random.choice(recent_movies, sample_size)
older_sample = np.random.choice(older_movies, sample_size)

#we put greater because we want to be able to do a right tailed test and see if the population mean 1 is greater
#than population mean 2
t_statistic, p_value = stats.ttest_ind(recent_sample, older_sample, equal_var=False, alternative = "greater")
print("p value: ",p_value)

p value:  0.624792328358243
```

We end up with the p-value above, which is greater than $\alpha = 0.05$. Thus, we do not reject the null hypothesis.

Hypothesis 2: Recently released movies have significantly longer watch times than older movies.

1) Make observations on the data.
    a. When looking at the pair plot in p2, I noticed that for the scatter plot given for Year of Release and Watch time showed a general change in trend towards Watch Time length as the years progressed. While the average length of movies appeared to be mainly constant throughout the years, the minimum length appeared to be slowly shifting to become longer and longer, and thus, I wanted to know if there was a growing trend for watch times in terms of recently released movies.
2) Conduct background research
    a. When conducting background research on the topic, I came across Chartr's article, which claimed that there was a trend for movies to be growing longer in terms of watch time. WhattoWatch makes a similar claim by doing a detailed breakdown of runtimes of movies in the top 10 at box office for particularly chosen years, which does indeed show a growing trend towards movies growing longer. This made me wonder if the 1000 top movies of IMDB follow the same trend, and if the newer movies in the dataset had longer watch times.
3) Pose a question.
    a. Thus, I pose the question: Do recently released movies have significantly longer watch times than older movies?

4) Run Experiment.
    a. For the sake of this experiment, we do a t test for two independent samples of scores (aka scipy.stats.ttest_ind) that is unpaired.

***H_0*: Null hypothesis**: There is no difference in watch time in terms of recently released movies and older movies.

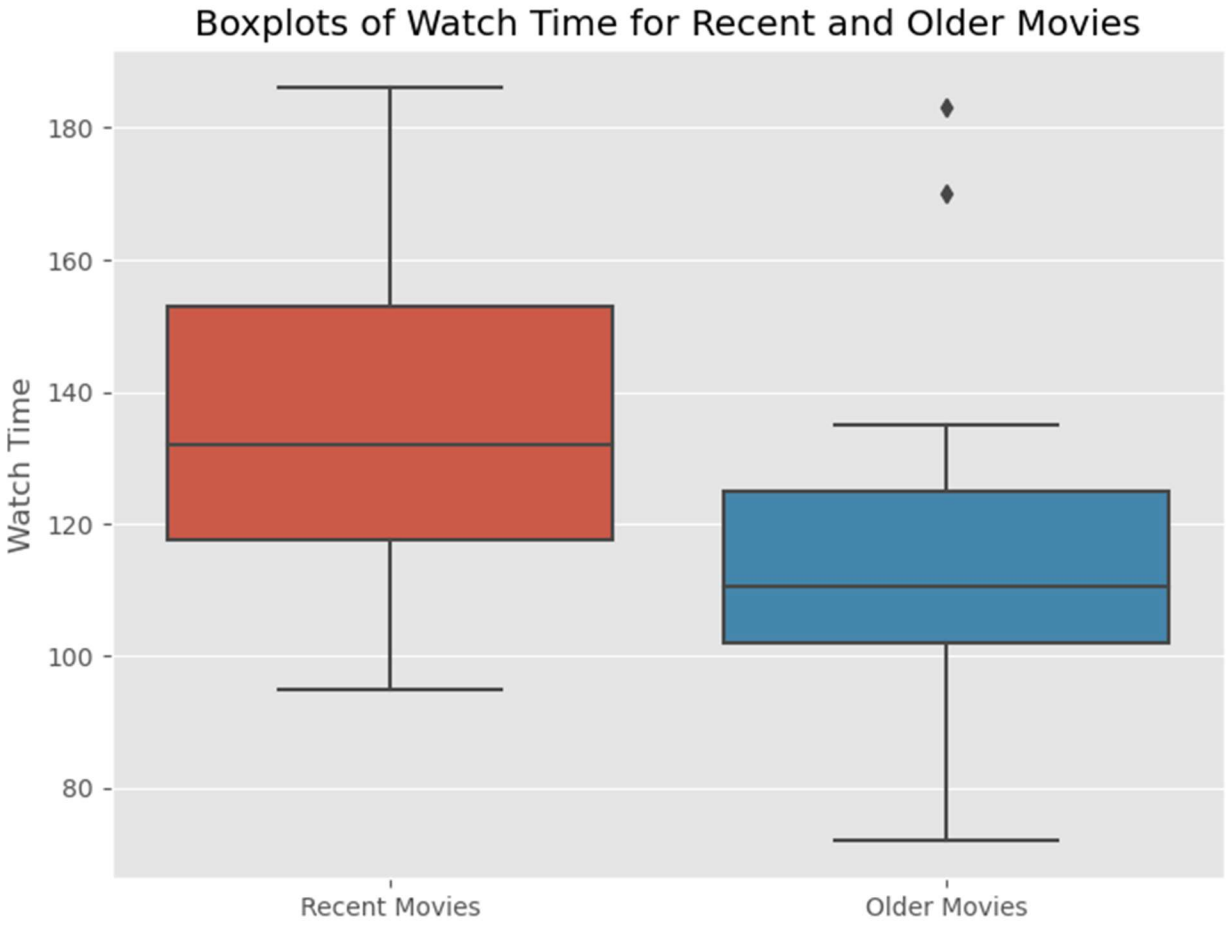
***H_a*: Alternative Hypothesis**: The watch time of recently released movies are significantly longer than that of older movies. (Aka, recent means past 2000, and older means prior to 2000)

First we took a sample size of 30 from the set of movies whose year of release was past 2000 and a sample size of 30 from the set of movies whose year of release was prior to 2000. We then ran a two sample right-tailed t-test on these randomly chosen samples and determined the p value to be as follows.

In [79]:
```
# Filter the DataFrame to get watch times for movies released since 2000
recent_watch = df[df['Year of Release'] >= 2000]['Watch Time']

# Filter the DataFrame to get watch times for movies released before 2000
older_watch = df[df['Year of Release'] < 2000]['Watch Time']

sample_size = 30
recent_sample = np.random.choice(recent_watch, sample_size)
older_sample = np.random.choice(older_watch, sample_size)

# Perform a t-test to compare the means of the two groups,
t_statistic, p_value = stats.ttest_ind(recent_sample, older_sample, equal_var=False, alternative = 'greater')
print("p value: " ,p_value)
```

```
p value:  0.00032531041245484755
```

Providing a boxplot of the data gives us the following, which helps me to better visualize what data we randomly selected.



We end up with a p-value of above, which is less than $\alpha = 0.05$. This tells us to reject the null hypothesis and say that the data is statistically significant at $\alpha = 0.05$.

<u>Hypothesis 3: Movies with a longer watch time tend to receive higher ratings than those with lower watch times.</u>

1) Make observations on the data.
    a. When looking at the scatter plot between Watch Time and movie ratings in P2, I noticed that while most of the data was clustered in a way that showed that movies of a variable length could receive the same movie rating score, between the watch times of 150 and 200 minutes, some of the movies scored exceptionally high. Whether this is because these values are outliers, or whether these values imply that movies with longer watch time receive higher ratings, I got to thinking: could there possibly be a connection between movies with a longer watch time with higher ratings?

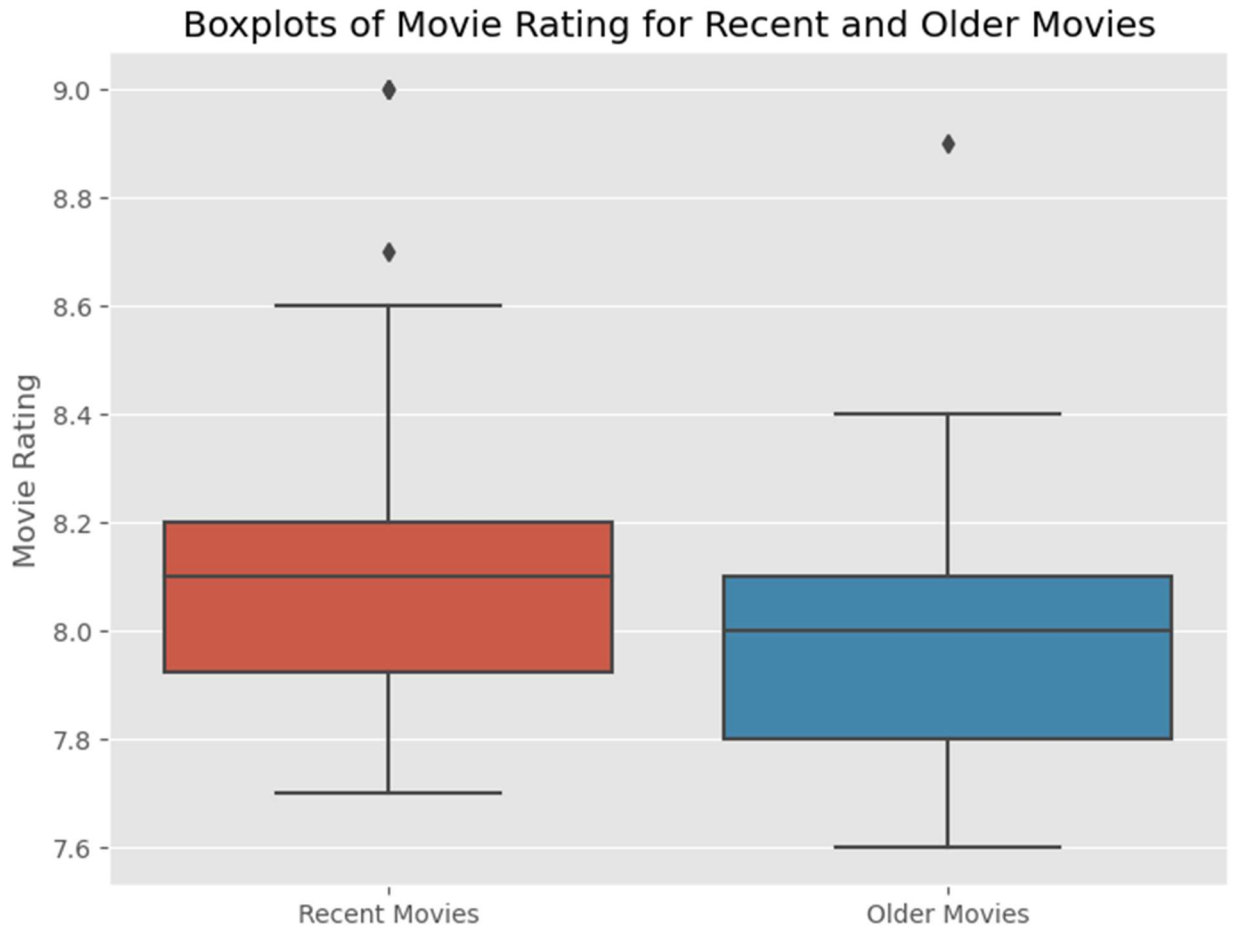
2) Conduct background research
    a. When looking into this question, I came across ScreenRant's article on how longer movies are better and more likely to be rated "Fresh" on *Rotten Tomatoes*. Though this is a different site than that of IMDB, I wanted to see if the statistic for *Rotten Tomatoes* also applied to that of the sample of IMDB's top 1000 movies. Similarly to ScreenRant, in 2019, Rotten Tomatoes posted an article that also implied that longer movies typically have higher Tomatometer averages. Again, this made me wonder if there was truly a difference in rating between longer movies and shorter movies.
3) Pose a question.
    a. Thus, I pose the question: Do movies with longer watch times tend to receive higher ratings than that of lower watch times? For the sake of this problem, since the mean watch time of movies is 124 minutes, let's say that a long watch time is 3 hours, or 180 minutes. In this experiment, we test to see if it the ratings are significantly different.
4) Run Experiment.
    a. For the sake of this experiment, we do a t test for two independent samples of scores (aka scipy.stats.ttest_ind) that is unpaired.

$\boldsymbol{H_0}$**: Null hypothesis**: There is no difference between the ratings of movies with higher watch times than those with shorter watch times. (aka movies with watch times of above 180 minutes and below 180 minutes)

$\boldsymbol{H_a}$**: Alternative Hypothesis**: There is a significant difference between the ratings of movies with higher watch times than those with shorter watch times.

First, we selected a sample size of 30 from the movie ratings of movies who had watch time over 180 and another sample size from the movies who had a watch time below 180 minutes. With this random selection, we thus did a two sample right-tailed t test on the two sets to determine the p value.

We also visualize the random sample sizes for both sets below in the form of boxplots.




```
In [89]: #first, we need to split the data into two groups.
         #Those with watch times over 180 and those below so we can
         #get the averages of the movie ratings

         long_watch = df[df['Watch Time'] >= 180]['Movie Rating']
         short_watch = df[df['Watch Time'] < 180]['Movie Rating']
         # print(long_watch.mean())
         # print(short_watch.mean())

         sample_size = 30
         long_sample = np.random.choice(long_watch, sample_size)
         short_sample = np.random.choice(short_watch, sample_size)

         t_test, p_val = stats.ttest_ind(long_sample, short_sample, equal_var=False,alternative = "greater")

         print("p value: " ,p_value)

         p value:  0.00032531041245484755
```

We end up with the above p-value, which is below $\alpha = 0.05$. Thus, we reject the null hypothesis and say that the data is statistically significant at $\alpha = 0.05$.

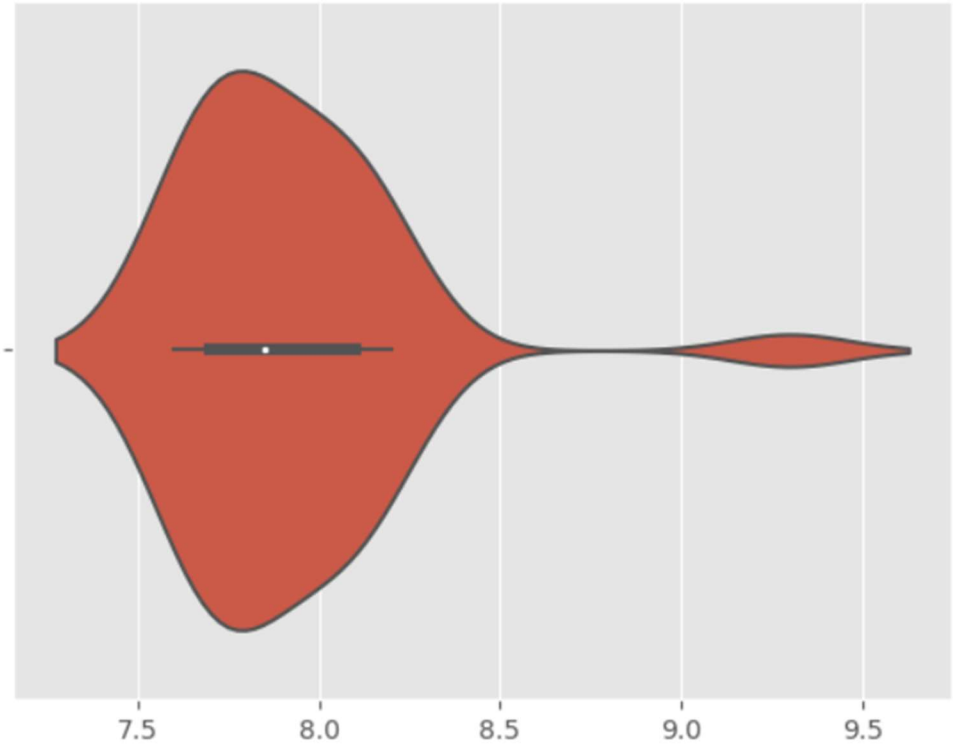
Hypothesis 4: The movie rating average of the sample, or the 1000 top movies on IMDB is significantly different than that of the general population.

1) Make observations on the data.
    a. When looking at the results of P1, I noticed that the average movie rating in the dataset is 7.97. While this indicates general positive reception from audiences, because this dataset is of the top 1000 movies on IMDB, I was surprised that the average rating was around 7.97 and not higher. This got me thinking: Is the average movie rating of the top 1000 movies on IMDB significantly different than that of all movies? I felt like average movie ratings were probably in a range from 6-7.5, but before I assume this, I need to perform some background research beforehand and see what is truly the case.
2) Conduct background research
    a. According to IMDB, the average movie ratings of movies on the site is 7.0, which brings the question to mind: are the top 1000 movies on IMDB significantly different than that of the website's in terms of movie ratings?
3) Pose a question.
    a. I want to see if the average movie ratings of IMDB's top 1000 movies are different than the mean of all movies on IMDB to see if there is a significant difference between the two. In short, I want to see if there is a significant difference between the sample mean (IMDB's top 1000) and the population mean (IMDB as a whole).
4) Run Experiment.
    a. For the sake of this experiment, we do a one sample t test (aka ttest_1samp).

H_0: **Null hypothesis**: There is no difference between the ratings of movies with higher watch times than those with shorter watch times.
H_a: **Alternative Hypothesis**: There is a significant difference between the ratings of movies with higher watch times than those with shorter watch times.

First, we took a sample size of 30 from the data frame's Movie Rating column, and created a violin plot to view the randomly selected datapoints. We end up with the

following.



Then, we run the 1 sample t test to see our p value.

```
In [10]: #testing to see if different from the general average 7.0
         ttest, p_value = ttest_1samp(rating_sample, 7)
         p_value

Out[10]: 6.125891693192224e-18
```

Since the p-value is less than 0.05, we reject the null hypothesis and say that the data is statistically significant at $\alpha = 0.05$.