Project: Wikidata, Movies, and Success

What actually makes movie successful? Do various criteria for success correlate with each other?
As we were only interested in the ratings, we took the average rating from audiences and critics from *rotten tomatoes* data. In this test, we would like to know whether those two ratings are correlated to each other. The null hypothesis is that those variables has no correlation at all. In addition, make sure we don't have empty or invalid data, we dropped the rows that has NaN entries.

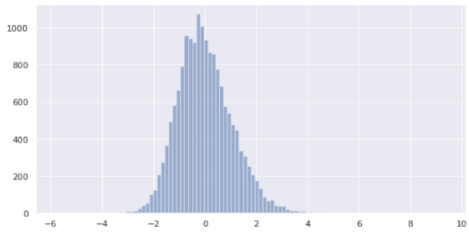
Next, as we think the *critics average ratings* are linearly related to the *audience average rating*, we used *linregress* to obtain the fitted line regression with on *critics average ratings* on y-axis and *audience average rating* on x-axis. we managed to get the following result:

```
Fitted value of the average ratings:
        slope ->  2.2143834539977956
        intercept ->  -0.29287443971619376
        rvalue ->  0.6991003609011144
        r2value ->  0.4887413146120684
        pvalue ->  0.0
        stderr ->  0.017509981224292495
```

Using the slope and intercept we could get a new prediction for *critics average ratings* and plotted it against *audience average rating.*

Based on the plot and the linregress result we could conclude a few things. First, the p-value of 0.00 is smaller than 0.005, thus we can reject the null hypotheses, and deduced that *critics average ratings* and au*dience average rating* has correlation. From, the graph, slope value, and r-value (0.699), we could tell that the *critics average ratings* and *audience average rating* have positive moderate correlation. Unfortunately, with r-squared of 0.488, it is not possible not explain those variation in *critics average ratings* using the *audience average rating*.

Lineregression has a few conditions that has to be fulfilled for it to be valid, one of it is the residual have to be pretty much normally distributed. By taking the predicted *critics average ratings* subtracted with the original *critics average ratings*, we could get the residual. From that, we could plot the residual histogram, and see that it's pretty much normally distributed.



Can we predict whether a movie made profit just using their average ratings?

There are quite a number of data that has no information about the profit, but perhaps we could try to fill the score in which the values are missing.

```
Number entries without empty average: 16732
Number entries without empty made_profit: 791
```

First, we started by taking several columns from *wikidata* data (rotten_tomatoes_id, made_profit) and *rotten tomatoes* data (rotten_tomatoes_id, audience_average, critic_average). Then we had to convert *audience average ratings* and *audience average ratings* into percentage so it's standardized. Then we join the wikidata and rotten tomatoes data on 'rotten_tomatoes_id'. Then we cleaned the data by dropping the one with empty data. We set 'audience_average' and 'critic_average' as the X, the factor, and 'made_profit' as y, what we want to predict. Then we split the data into training and testing data.

Next, we set up the model for the prediction, Bayes, KNN, and SVC, and try to train the model. The accuracy score is as follows:

```
Bayes accuracy:  0.7989130434782609
KNN accuracy:  0.8097826086956522
SVC accuracy:  0.8641304347826086
```

After a few trials, we realized that the accuracy score is consistently "good" at 0.80 range. As we suspected that there might be sample imbalance in the movies that have made profit and movies that made loss, we decided to check that.

```
Percentage of movies that made profit: 84.91847826086956 %
Percentage of movies that made loss: 15.081521739130435 %

Difference in number of sample 514
```

From the result above, we could tell that there are much more positive sample (made profit), which is about 80% which made a lot of sense when we compare with the accuracy. We could conclude that there is sample imbalance, so we decided to fix it by dropping the number of sample of movies that made profit. This was done by shuffling the positive sample. Grab as many data as the negative data, and then combine the positive and the negative data. We followed the same step as above but added MinMaxScaler() on the defining the model part which makes the accuracy fluctuate less and have slightly higher result.

```
Bayes accuracy:  0.6428571428571429
KNN accuracy:  0.7321428571428571
SVC accuracy:  0.6428571428571429
```

We could get around 0.6 up to 0.7 accuracy score, but it fluctuates too much. Unfortunately, not much that we can conclude from here, except we could not really predict whether a movie would make a profit based to the average ratings score. Perhaps we could get better score when we have more data, or if there is more technique that we could introduce to this study.

Do the movies that made profit has higher rating?

We would like to know if movies that made profit has higher average ratings. The null hypothesis here is "Having profit means higher rating". We used the same data like before, and separate the *wikidata* that has been joined with *rotten tomato* data based on whether they make profit or loss. Then we tried Mann-Whiteney-U test on both *audience average rating* and *critics average ratings*. The first input is the positive ratings (made profit), and the second input is the negative result (made loss).

```
The Audience Average Ratings p-value: 1.5012176450003027e-11
The Critics Average Ratings p-value: 3.1205039528034646e-09
```

Both p-value are lesser than 0.05, thus we can reject the null hypothesis that the movies that made profit has higher rating.