**Mini Project**

**Introduction**

The topic of this project is about analyzing Hollywood’s most profitable stories from 2007 to 2011. There are 7 variables in the dataset, which are genre, studio, audience score, profitability, rotten tomatoes percentage (an American review-aggregation website for film and television), worldwide gross, and year. Hence, there are many options for the analysis, e.g., the relationship between audience scores and worldwide gross (does the high worldwide gross must have a high audience score), comparing the worldwide gross of a different genre to see the preference of the audience, etc. I found the dataset on tableau public (<https://public.tableau.com/en-us/s/resources>), a software for creating and sharing data visualization online, and the origin of the dataset is <https://www.informationisbeautiful.net/data/>. The strength of this dataset is that it has many variables, so we can analyze a lot of information from the sample. Also, the dataset is organized and cleaned, so I don’t need to spend time cleaning and transforming data. The weakness of the dataset is that it’s not large enough to have an accurate statistical result, and there are some null values in the dataset. My research question is to analyze the relationship between audience score and worldwide gross (does a film with a high worldwide gross must have a high audience score), and compare the difference between the profitability of drama and romance.

**Descriptive Statistics**

I have plotted histograms for three variables respectively that I will use in this project by R, which are audience score, profitability, and worldwide grass. I have included the density line and normal curve for better checking their normality and distribution.

Chart, line chart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Besides graphs, I have calculated the mean, median, and standard deviation of three variables by R, and I have created a table to better display them.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mean | Median | Standard deviation |
| Audience score | 64.45 | 64.5 | 13.51 |
| Profitability | 4.78 | 2.64 | 8.34 |
| Worldwide gross | 141.93 | 85.89 | 159.69 |

**Inferential Statistics**

I have conducted four hypotheses tests and one confidence interval of the dataset by using R, and the confidence interval is reflected in the hypothesis test.

Graphical user interface, text, application

Description automatically generated

The Shapiro Wilk test is used to check the normality of the data audience score.

Text

Description automatically generated

This one is the chi-square test between audience score and worldwide gross, the purpose of it is to see if there is a relation between these two variables.

Text, letter

Description automatically generated

This one is the student’s t-test between the profitability of drama films and the profitability of romance films. The purpose of this test is to check if there is a difference between the mean of these two categories of film. And the test returns a 95% confidence interval.

Text

Description automatically generated

The last hypothesis test I did is the Wilcoxon rank-sum test (Mann-Whitney U test) between the profitability of drama films and romance films. The reason I did this test is that the distribution of the dataset is not normal, and the sample size is too small.

**Discussion**

From the histogram of the audience score, it seems to have normal distribution since the density line and normal curve are similar. The reason for checking normality is that we can conduct hypothesis testing for a sample drawn from a normal distribution population by referencing the table in the textbook (Leemis, 406), which provided more statistical power than a nonnormally distributed dataset. I have conducted the Shapiro-Wilk test to check the normality of audience scores. The null hypothesis of the test is that the data of audience score is normally distributed, and the alternative hypothesis is that the data is not normally distributed. From the result, p-value=0.1405 > significance level =0.05, so we don’t have enough statistical evidence to reject the null hypothesis, which means that the data of audience score is normally distributed. I did the normality hypothesis test for the other two variables (profitability and worldwide gross), and the result is both profitability and worldwide gross are non-normally distributed.

After doing the normality testing, by observing the graphs and running the Shapiro-Wilk test, I have better learning of the dataset, which helps do hypothesis testing. To test the relationship between audience scores and worldwide gross, I ran the chi-square test in R. The null hypothesis of the test is that two variables are independent, and the alternative hypothesis is that two variables have a relation. Comparing the p-value with the significance level, 0.2976>0.05, we retain the null hypothesis and reject the alternative hypothesis, so we can conclude that audience scores and worldwide gross are two independent variables, which means the increase in audience scores of a movie doesn’t necessarily mean that it will have higher worldwide gross.

Then I did a student’s t-test for the profitability of drama films and romance films, and I assumed they are normally distributed before I run the hypothesis testing. Following the 4 steps of the hypothesis test in the class, the null hypothesis will be the mean of profitability of drama film equal to the mean of profitability of romance film. On the other hand, the alternative hypothesis is that they are different. So, it’s a two-sided hypothesis since we are supposed to pay attention to both tails of the critical area. The test statistic is the t score, which is calculated by (Leemis, 406), following student’s t distribution (variance is unknown). We compare the test statistics with the critical value to conclude. Another way we usually use is to compare the p-value with the significance level ɑ=0.05. Therefore, in this case, we have p-value=0.4394, which is greater than ɑ=0.05. We can conclude from the t-test that we have sufficient evidence to retain the null hypothesis that the mean profitability of drama films is equal to the mean profitability of romance films, which means there is no difference between the profitability of drama films and romance films. Moreover, we can use a confidence interval to conclude. In this case, the 95% confidence interval is (-6.90, 14.97), which means we are 95% confident that the mean difference between the profitability of drama films and romance films falls into the range (-6.90, 14.97). The way we use the confidence interval to conclude is that we reject the null hypothesis if the parameter value we want to check in the null hypothesis does not fall in the confidence interval. For example, in this case, the null hypothesis is (X=profitability of drama films, Y=profitability of romance films), so we reject the null hypothesis if 0 does not fall in the confidence interval. In this case, since 0 is in the range (-6.90, 14.97), we have insufficient evidence to reject the null hypothesis and we get the same conclusion as the significance testing method we applied above.

In addition, I also conduct a Mann-Whitney U test on the profitability of drama films and romance films. The reason for that is the dataset is not normally distributed, so a non-parametric hypothesis test will provide a more accurate and reliable result than a parametric hypothesis. With the same purpose in the t-test, p-value = 0.793 in the Wilcoxon Rank sum test is significantly greater than the significance level of 0.05, so we draw the same conclusion as in the t-test that there is no difference between the mean profitability of drama films and romance films at significance level = 0.05.

**Reflection**

From the process of data analysis, I have followed the basic steps of data analysis, that is collecting data, visualization, calculating descriptive statistics of the dataset, and running hypothesis testing. I have a deeper understanding of how to run data analysis and all concepts of hypothesis testing, such as p-value, confidence intervals, etc. I enjoy doing this data analysis since watching films is one of my hobbies. The passion for analyzing data depends on the content of the dataset and the research questions. If the dataset and research direction is interesting to me, the process of data analysis is interesting. I have learned a general pattern of analyzing data, and most importantly, I have learned what effort should be put to solve the research question, and how different process of data analysis connects to solve the problem. I appreciate both data collectors and data analysts. After doing this project, I appreciate data collectors more because collecting, organizing, and cleaning data is the start and one of the most crucial parts of data analysis. We cannot run an efficient data analysis without a clean dataset from a reliable source. Also, I have developed my personal perspective on statistical conclusions. Statistical conclusions should be considered as a powerful reference instead of a decisive tool since there are so many uncertain elements in data analysis, and none of the datasets is perfect in terms of data source, distribution, etc.