**7COM1079 – Team Research and Development Project**

**Group 171 - Coursework**

Contents

[Introduction: 2](#_Toc61037050)

[Visualisation 2](#_Toc61037051)

[Analysis: 3](#_Toc61037052)

[Conclusion 5](#_Toc61037053)

[Bibliography 5](#_Toc61037054)

## Introduction:

We have chosen “Netflix\_Titles” dataset from Kaggle.com[[1]](#footnote-1). The data set includes Show ID, type, title, director, release\_ year, date\_added, cast, country, movie rating, duration, listed in, description and genre. From these columns we have chosen “type” and “release\_year” performed our analysis to show that the TV show proportion is higher than that of Movie since last decade.

The objective of this study is to find out that there is a difference in proportion between “Movie” and “TV Show” in Netflix over the last decade.

**Research Question:**

* Is there a difference in the proportion of Tv Show and Movie available in Netflix over the last decade?

**Null Hypothesis:**

* There is no difference in the proportions of Tv Show, and Movie available in Netflix over the last decade.

**Alternative Hypothesis:**

* There is a difference in the proportions of Tv Show, and Movie available in Netflix over the last decade.

## 

## Visualisation[[2]](#footnote-2)

Data visualisation is the graphical representation of information and data. By using visual elements like charts, graphs and maps, data visualisation tools provide an accessible way to see and understand trends, outliers, and patterns in data.

Chart, bar chart

Description automatically generated

Figure 1: Proportion of Netflix TV Show and Movie

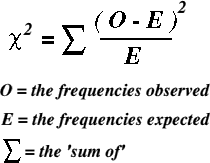
To plot the above graph, we have used “release\_year” column values from 2009 to 2020 and “type” column values of “Movie” and “Tv Show” from the dataset that we have used in our research. We plotted “year” on x-axis and "Fraction" on y-axis. We have plotted the graph in R-Script by using below libraries.

**library(ggplot2)[[3]](#footnote-3)**

**library(scales)[[4]](#footnote-4)**

To plot the bar chart in the R-Script, we have used This library **(ggplot2)** and to plot the percentages on the y-axis we have used **library(scales).** From this graph, we can clearly see that how the Tv show proportion has been increased over the years.

## Analysis:

To perform analysis on our data we have used a chi-squared test, also written as the χ2 test, is a true-to-be mathematical hypothesis test in which the numbers of the chi-squared test are distributed under the null hypothesis, particularly the Pearson chi-squared test and its variants. Pearson's chi-squared test is used to assess if there is a statistically important difference in one or more classes of a contingency table between the expected frequencies and the observed frequencies.  
  
The chi-squared test is used to test the goodness of fit. In this test of the parameters, it is usually assumed that the random variable follows a particular distribution like Binomial, Poisson, normal distribution, etc. However, often the need is felt to confirm whether our assumption is true or not. So, based on outcomes of a trial or observational data. The chi-squared test is performed, which measures the discrepancy between the observed frequencies and theoretically determined frequencies from the assumed distribution for the same event. If the discrepancy is not large, it is considered that our assumption about the distribution of the variable is correct, otherwise not.  
  
**Chi-squared test χ2 can be derived as**:  
[[5]](#footnote-5)

In the R script, we have calculated X2 =Chi-Square, df (Degrees of freedom) and p value. The values which we have obtained from the analysis are as follows:

* **Chi Square** **X2 = 212.47,**
* **Degrees of Freedom (df) = 11,**
* **P-Value < 2.2e-16**(nearly equivalent to zero)
* **Observed values:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 |
| Movie | 87 | 111 | 100 | 125 | 177 | 213 | 363 | 593 | 682 | 646 | 400 | 6 |
| TvShow | 34 | 38 | 36 | 58 | 60 | 75 | 154 | 237 | 277 | 417 | 443 | 19 |

* **Expected values:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 |
| Movie | 79.21192 | 97.54195 | 89.03158 | 119.79985 | 155.15063 | 188.53747 | 338.4509 | 543.3545 | 627.8036 | 6958866 | 551.8649 | 16.3661 |
| TvShow | 41.78808 | 51.45805 | 46.96842 | 63.2001 | 81.84937 | 99.46253 | 178.5491 | 286.6455 | 331.1964 | 367.1134 | 291.1351 | 8.6339 |

As per the above analysis, the p-value which we have obtained is **2.2e-16(~0)** which is less than **0.05**. Thus, we are rejecting the null hypothesis.

## Conclusion

After performing the visualization and analysis on the “Netflix\_titles” dataset we concluded that “there is a difference in the proportions of Tv Show and Movie available in Netflix over the last decade”. As per the results obtained from the R-Script we can see that p-value is 2.2e-16(~0), which is less than 0.05. Hence, we reject null hypothesis.

## Bibliography

1.‘Https://En.Wikipedia.Org/Wiki/Chi-Squared\_test’, n.d.

2.‘Https://Support.Rstudio.Com/Hc/En-Us/Articles/201057987-Quick-List-of-Useful-R-Packages’, n.d.

3.‘Https://Www.Kaggle.Com/Shivamb/Netflix-Shows’, n.d.

4.‘Https://Www.Tableau.Com/En-Gb/Learn/Articles/Data-Visualization’, n.d.

1. ‘Https://Www.Kaggle.Com/Shivamb/Netflix-Shows’. [↑](#footnote-ref-1)
2. ‘Https://Www.Tableau.Com/En-Gb/Learn/Articles/Data-Visualization’. [↑](#footnote-ref-2)
3. ‘Https://Support.Rstudio.Com/Hc/En-Us/Articles/201057987-Quick-List-of-Useful-R-Packages’. [↑](#footnote-ref-3)
4. ‘Https://Support.Rstudio.Com/Hc/En-Us/Articles/201057987-Quick-List-of-Useful-R-Packages’. [↑](#footnote-ref-4)
5. ‘Https://En.Wikipedia.Org/Wiki/Chi-Squared\_test’. [↑](#footnote-ref-5)