Executive Summary

Canopy is a boutique streaming service seeking to provide quality streaming services for viewers of French-Language movies. This report explores a dataset of over 14000 movies to assess if Canopy can achieve its business goals and also to assess the general market surrounding movies released in France against movies not released in France. It is this report's suggestion that Canopy can achieve its business goals. It can accomplish this by focusing its initial customer base on the older 18+ market and the Drama, Comedy, Thriller and Romance genres where the market is most abundant. Canopy should also look to advise French Filmmakers of the areas of less focus, such as younger and family based audiences, as well as the Sci-Fi, Musical/Music, Sport and Western genres.

Table of Contents

1. Introduction

Canopy is a boutique streaming service seeking to provide quality streaming services for viewers of French-Language movies. They have requested analytics be performed on movies to provide guidance on business decisions. This report examines a database of over 14000 movie titles using the Python (Van Rossum, G. & Drake, F.L., 2009) programming language to produce a dashboard of visualisations to assist in business decision making. The report will first explore the reasoning behind the use of Python (Van Rossum, G. & Drake, F.L., 2009) and the data visualisation libraries used. The report will then explore the actual visualisations created and the dashboard provided for Canopy to use. Finally the report will summarise the observations and conclusions made from the visualisations and how that impacts Canopy's business goals.

2. Python and Data Visualisation

The Python (Van Rossum, G. & Drake, F.L., 2009) programming language is a high-level widely adopted language that is extremely well suited for data analytics, primarily due to the robust numerical and tabulated data storing libraries built for it such as Numpy (Harris, C.R. et al., 2020) and Pandas (McKinney, W. & others, 2010.). It is easy to learn and use due to the general flow and structure closely resembling the English language while being versatile and powerful for complex programming.

Data visualisation involves turning data into a format that conveys information visually, typically in the form of graphs. There are several data visualisation libraries for Python (Van

Rossum, G. & Drake, F.L., 2009) and this report focuses on Seaborn (Waskom, M. et al., 2017) for most of the visualisations and Matplotlib (Hunter, J.D., 2007) for the dashboarding. Seaborn is the most simple of the common visualisation libraries to generate graphs for and its default settings are often more than suitable for use. Matplotlib (Hunter, J.D., 2007) is more complex than Seaborn (Waskom, M. et al., 2017), requiring more lines of code to accomplish the same level of graphing but in turn has more flexibility in its application. Matplotlib (Hunter, J.D., 2007) also has functionality that allows multiple visualisations to be graphed to the same code cell and can also take on completed Seaborn (Waskom, M. et al., 2017 visualisations in its dashboard.

3. Visualisations and Dashboard

Four primary visualisations were created to assist Canopy with assessing their business goals. The first is a Seaborn (Waskom, M. et al., 2017) box-and-whisker plot of the score distribution of movies that were released in France, against movies that were not released in France. The box-and-whisker provides a quick, yet robust understanding of the distribution of data by showing the five-number summary. The visualisation then provides a simple observation demonstrating that the scores for films released in France score higher on average than films not released in France.

The second visualisation is a Seaborn (Waskom, M. et al., 2017) bar graph of the number of films by age category for films released in the French Language, designed to highlight areas of focus within the film industry that may be suitable for French filmmakers to pursue, as well as likely audience targets for the current market of movies. Currently the adult market of 18+ dominates the count, at over double the next highest age category of 13+. The lowest categories are those for youths (all) or niche age brackets (16+).

The third visualisation is another Seaborn (Waskom, M. et al., 2017) bar graph, this time of the number of films by genre released in the French Language. Again this is to highlight areas of focus within the film industry that may be suitable avenues for film creation by French filmmakers as well as the likely audience of the current market. Drama dominates the market, at nearly triple that of the next highest genre of Comedy, which is closely followed by Thriller and Romance. The weakest genre representations are that of some niche genres, such as News and Shorts, as well as some more popular genres such as Sci-Fi, Musical/Music, Sport and Western.

The fourth and final visualisation is another Seaborn (Waskom, M. et al., 2017) box-and-whisker plot, this time highlighting the runtime of movies released in France against movies not released in France. It shows similarly to the first box-and-whisker plot and shows that movies in France tend to run longer than movies not released in France.

These four visualisations are then produced in a single code cell, using a Matplotlib (Hunter, J.D., 2007) subplot grid, arranged to a two by two square as seen in Figure 1 below.

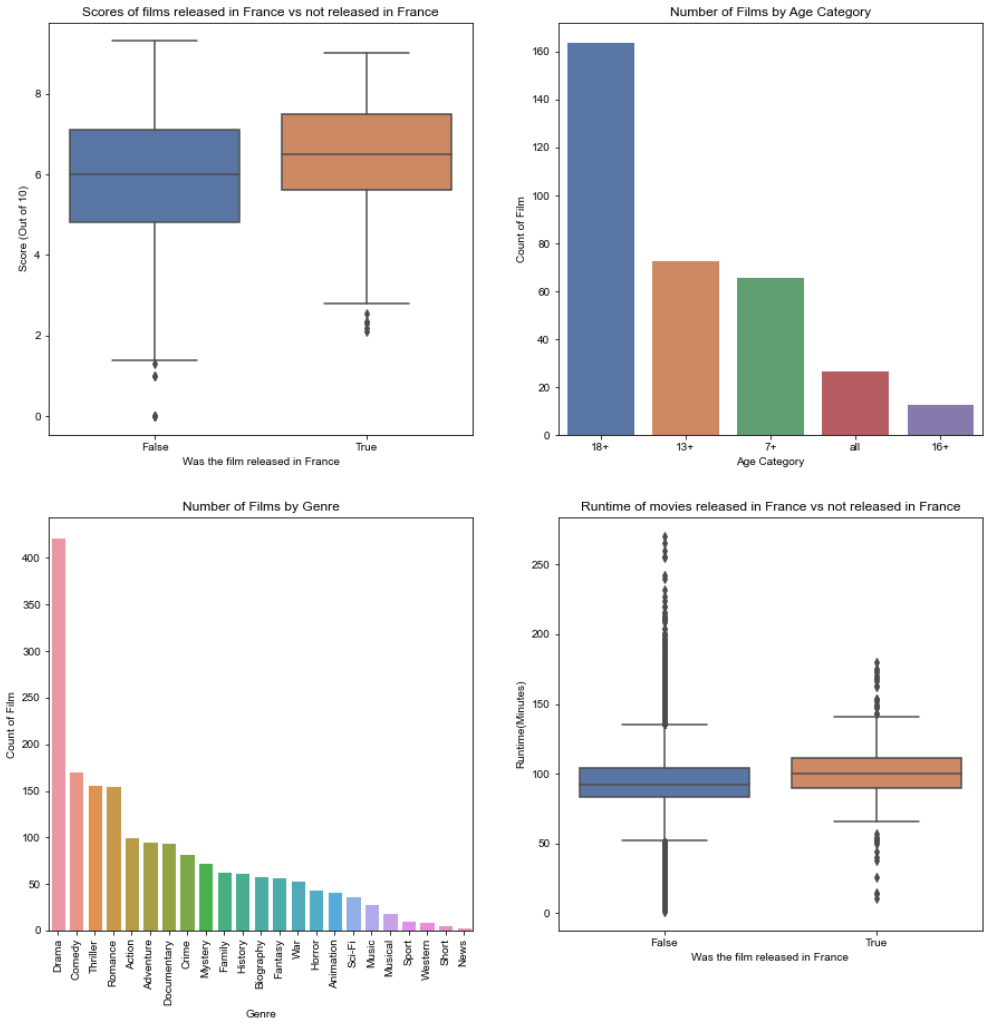

Figure 1: The dashboard output of four visualisations to assist Canopy in business goal assessment. The 4 graphs are all plotted via Seaborn (Waskom, M. et al., 2017) and the dashboard grid was created using Matplotlib (Hunter, J.D., 2007). The top left graph is a box-and-whisker of scores for movies released in France against movies not released in France. The top right is a bar chart of the counts of films released in French for each age category. The bottom left is a bar chart of the counts of films released in French for each genre. The bottom right is a box-and-whisker plot of runtimes for movies released in France against movies not released in France.

The dashboard is laid out such that the box-and-whisker plots and the bar plots are diagonally opposite each other so as to not cause any confusion against the axes used for same style graphs.

4. Observations and Conclusions

It has been observed that films released in France score better and run longer than films not released in France on average. It has also been determined that the market for French-Language viewers caters primarily to older customers and the genre of Drama. The initial focus and design of Canopy should be aiming to cater towards that demographic. Canopy should also consider providing feedback to French filmmakers regarding the lack of content provided for genres such as Sci-Fi, Music/Musical, Sport and Western. They should also advise of the lack of content provided towards younger and more family orientated audiences. It is this report's conclusion that Canopy should be capable of accomplishing their business goals of catering toward French-Language viewers.

References

Harris, C.R. et al., 2020. Array programming with NumPy. *Nature*, 585, pp.357–362.

Hunter, J.D., 2007. Matplotlib: A 2D graphics environment. Computing in science & engineering, 9(3), pp.90–95.

McKinney, W. & others, 2010. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*. pp. 51–56.

Van Rossum, G. & Drake, F.L., 2009. *Python 3 Reference Manual*, Scotts Valley, CA: CreateSpace.

Waskom, M. et al., 2017. *mwaskom/seaborn: v0.8.1 (September 2017)*, Zenodo. Available at: https://doi.org/10.5281/zenodo.883859.