IMDB Top 100 Popular Movies Income Visual Analysis and Storytelling

1. Introduction

1.1 Introduction

Over the past two decades, the film industry has undergone significant evolution due to technological advancements and the global proliferation of innovative ideas. This transformation has enriched both the production and viewing experiences of films. It is crucial for filmmakers, film critics, and audiences to understand the myriad factors that influence financial success.

In light of this, we have chosen to analyze the top 100 movies from the years 2003 to 2022, sourced from the IMDb dataset available on Kaggle. We have placed a strong emphasis on visualisations as the primary method of analysis, utilizing them to narrate the data's story and offer audiences an insightful and intuitive understanding. In doing so, we also furnish dependable recommendations for achieving financial success for movie producers.

To facilitate a robust visual analysis and storytelling, we've structured a series of progressive subtasks, guiding us from basic to intricate analyses, each aimed at crafting a data narrative through appealing and user-friendly visualizations. This data is categorized into three types: high-dimensional, graphical, and dynamic data, enabling us to simultaneously explore multiple factors, visualize entity relationships, and observe temporal trends. Our visual analytics systems are purposefully designed to deliver these insights in an intuitive and interactive format, facilitating efficient data exploration and the extraction of valuable insights.

1.2. Data set

The original full data set we choose to use is Top 100 popular movies from 2003 to 2022 (iMDB). Specifically in this report, we focus on analysing high income movies. So our data set is the sub data set that the income of those movies are all higher than 256.70.

After data cleaning described in the implementation part below, some samples are further dropped, and we add 2 new columns which are called Income_million and Budget_million, for visualising 2 original attributes in numerical types and consistent units.

The new dataset (extracted subsets) contains 13 attributes as follows and our analysis only relate 9 attributes and 396 samples

- Title: The movie name
- Rating: The rating of the movie according to IMDB users

- Year: The release year of the movie
- Runtime: The length of the movie in minutes
- Director/s: The person/people who directed the movie
- Stars: Actors playing in the movie
- Genre/s: The genre/s of the movie
- Budget_million: The money spent on the movie
- Income_million: The money earned by the movie

For each subtasks we defined below, we further extract 4 subsets, and this will be detailly introduced in the data processing section.

2. Design

2.1. Tasks

Our objective is to analyse high-income movies and provide insights for global movie makers including movies directors, producers, etc. regarding financial success. To achieve this, we have defined a series of subtasks that progress from simple to complex, allowing for a step-by-step and in-depth analysis.

Our first task aims to gain an overview of high-income movies by exploring the relationship between movie directors and their incomes. This involves analysing and visualising graph data via tree map. As part of this task, we also identify the top five movie directors with the highest income and present their information in an intuitive manner in the next task.

Moving on to our second task, we dive into high-dimensional categorical data analysis. Building upon the findings from the previous task, we primarily focus on the top five income directors. Our goal is to explore the potential correlation between income and these high-income directors.

Lastly, we embark on dynamic data visualisation, aiming to uncover insights into the trends of high-income movies over the years. In this task, we aim to gain a deeper understanding of the evolving and dynamic landscape of the data.

Moreover, a visual analysis system is implemented to enhance the user engagement, visual intuition and show the data storyline.

2.2. Data processing

Graph Data

In terms of the graph data analysis, we aim to visualise the overview income of the data set with focus on displaying the network and relationship between directors and income. Consequently, we choose the sub data set consisting of attributes: ‘Directors’, ‘Title’, ‘Income’ to analyse, and ‘Genre’ is also displayed to provide further information.

High-Dimensional Data

For high-dimensional data, we primarily focus on high-dimensional categorical data analysis with respect to the top 5 income directors ('Anthony Russo, Joe Russo', 'James Cameron', 'Christopher Nolan', 'Peter Jackson' and 'David Yates') identified from the above task. In order to display detailed information of each director/s, we select further attributes: Title, Directors, Stars, Genre and Income. Moreover, to support our analysis results in a statistical manner, we employ the hypothesis test between Directors and Income of all directors in our data set.

Dynamic Data

In order to observe the pattern and evaluate the trends over years, we employ ‘Year', ‘Director’, ‘Title’, ‘Genre’, and ‘Income’ as our primary attributes for analysis. Additionally, in order to engage the audience and include more information for the storyline, ‘Stars’ is also selected as part of the sub data set.

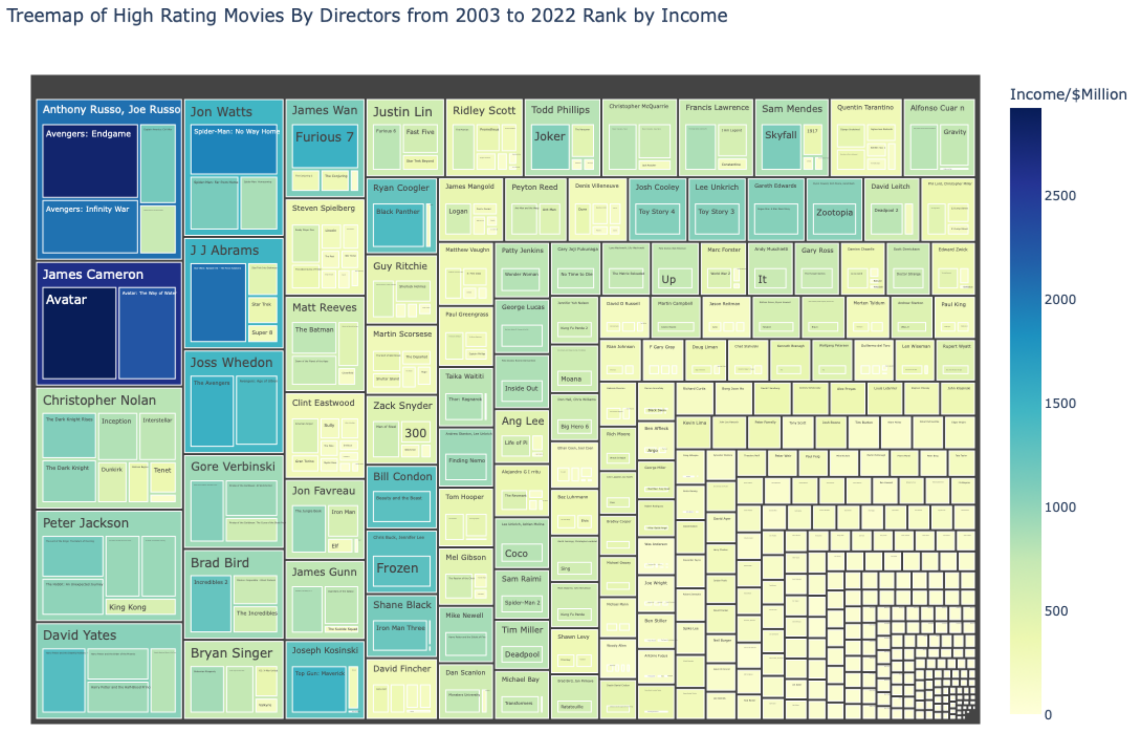
Visual Analysis system

As the visualisation needs to provide comprehensive visualisations of data, except the attributes mentioned above, we also include Runtime and Ratings to provide more information of each movie.

3. Visual Anlysis with Storytelling

3.1 Analysis

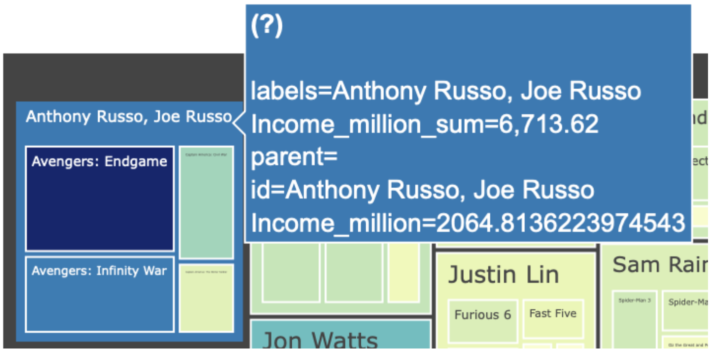
Visualisation 1



Treemap of High Rating Movies By Directors from 2003 to 2022 Rank by Income

Storytelling 1

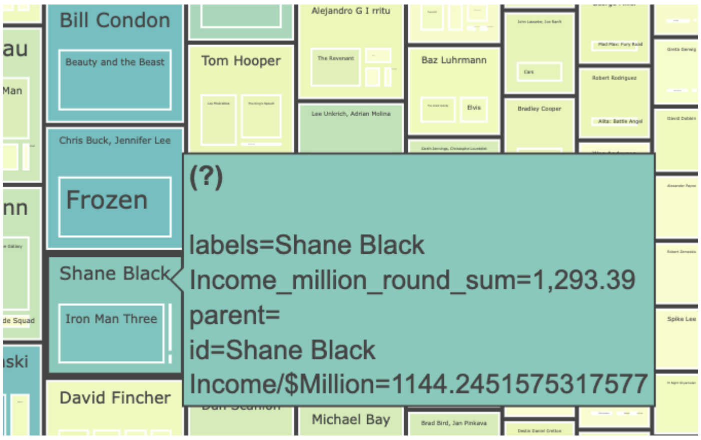
The tree map above displays the overall income situation of all movies and directors, and reveals the potential pattern about directors and income. The structural layout consists of square nodes from small to big indicating the income of each movie and director. From the graph, we can see that the 5 biggest block are 'Anthony Russo, Joe Russo', 'James Cameron', 'Christopher Nolan', 'Peter Jackson' and 'David Yates' who are also our top 5 income directors. The income of Anthony Russo and Joe Russo gained the total income of 6713.62, and their superhero series of movies contribute to the high income, which matches the real movie market that shows the high popularity of superhero movies around the world. So as for the rest of the directors who are also famous for other classical IP (intellectual properties) movies such as the Harry Potter series and the Dark Knight series.



(?)

labels=Anthony Russo, Joe Russo
Income_million_sum=6,713.62
parent=
id=Anthony Russo, Joe Russo
Income_million=2064.8136223974543

Except for those familiar to the ears, well-known works. Most movies have lower income with the range from around 250 to 500 million. Some of the movie directors have directed more than 3 movies, but some of others only directed 1 or 2 movies. However, fewer movies directed doesn't mean the director is financially unsuccessful, there are indeed some directors such as Bill Conton and Shane Black also gained more than 1000 million from their works.



(?)

labels=Shane Black
Income_million_round_sum=1,293.39
parent=
id=Shane Black
Income/$Million=1144.2451575317577

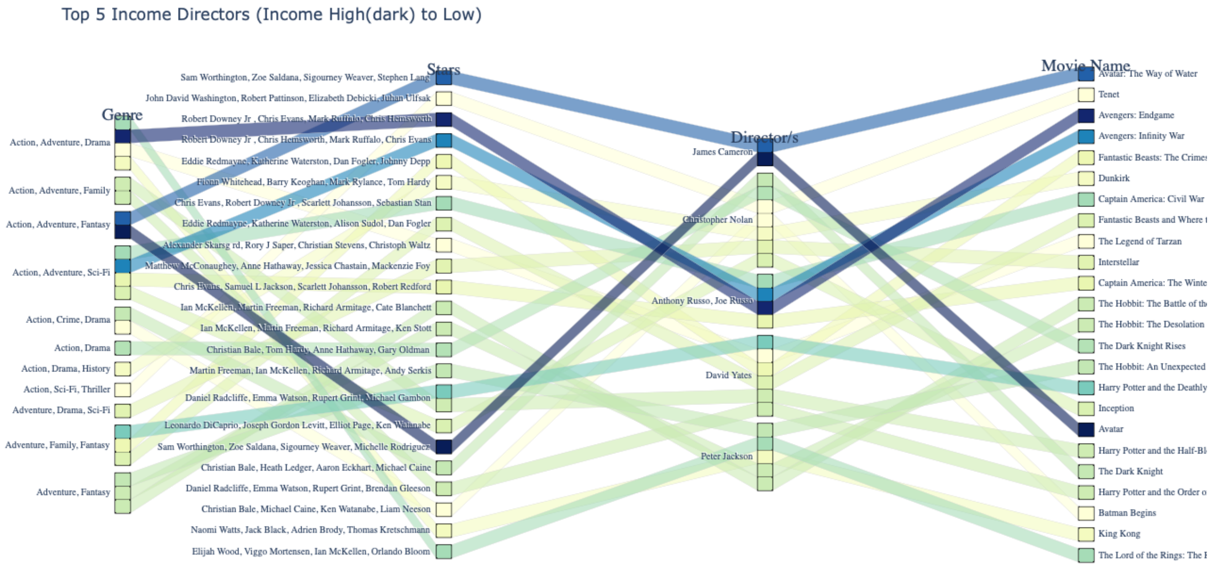
Based on findings above, we have made conclusions as follows:

1. Regarding head directors, the well-known series superhero movies have gained huge financial success.
2. Most movies have a lower income range from 250 to 500 million.

Furthermore we provide suggestions as follows for the movie producers and directors.

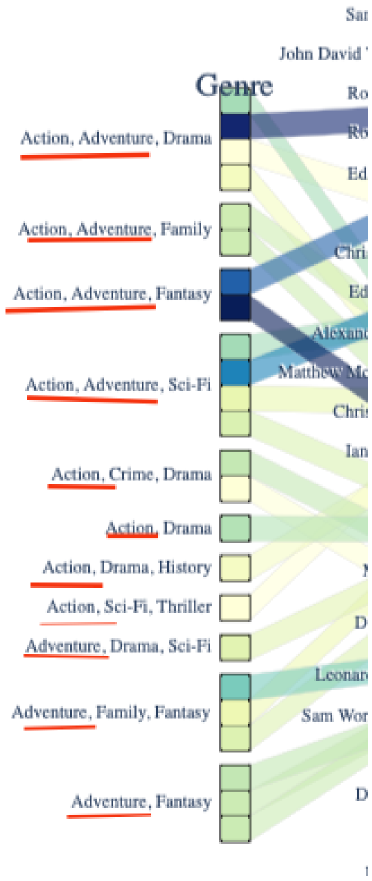
1. In order to gain financial success, it is important to shape an IP (It can be a work that is characterised by the virtual world or a character inside) and promote it to the market with a moving story line.
2. Continued sequels to already well-known works can also generate predictable high income

Visualisation 2

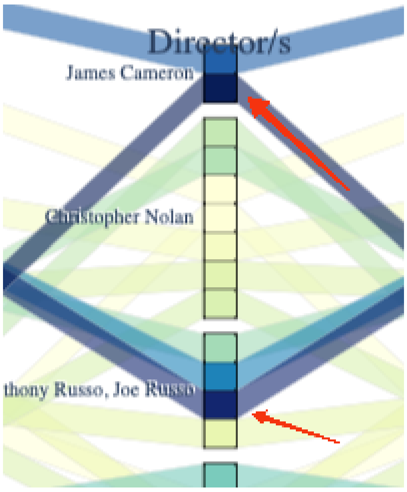


Storytelling 2

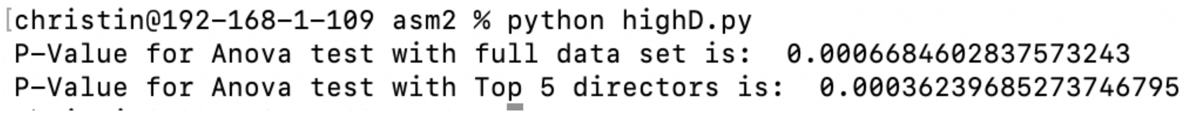
The parallel category graph above provides additional insights and supports our findings from the first visualisation. It confirms that high-income movies are primarily associated with series movies or those adapted from well-known intellectual properties (IPs), such as the Harry Potter series. This observation strengthens the notion that established IPs contribute significantly to generating high income in the movie industry.



Furthermore, an interesting observation is that all the listed genres in the graph include either action or adventure, or both. This suggests that the action and adventure genres have a broad market appeal, with action movies being particularly prominent.



Additionally, we notice that some directors consistently produce movies with high income. This observation implies a potential correlation between the income generated and the director's involvement. This aligns with the notion that renowned directors attract larger audiences, as people are more inclined to watch movies directed by well-known and highly-regarded individuals.



```
[christin@192-168-1-109 asm2 % python highD.py
 P-Value for Anova test with full data set is:  0.0006684602837573243
 P-Value for Anova test with Top 5 directors is:  0.00036239685273746795
```

To further support this finding statistically, we conducted a one-way ANOVA test to analyse the relationship between directors (independent variable) and income (dependent variable). Our null hypothesis (H0) assumes that the income values among all director groups are the same, with an alpha value of 0.1. However, both p-values obtained from the analysis of the top 5 income directors and all directors are less than the alpha level, leading us to reject the null hypothesis. This indicates a statistically significant correlation between the income generated and the directors involved.

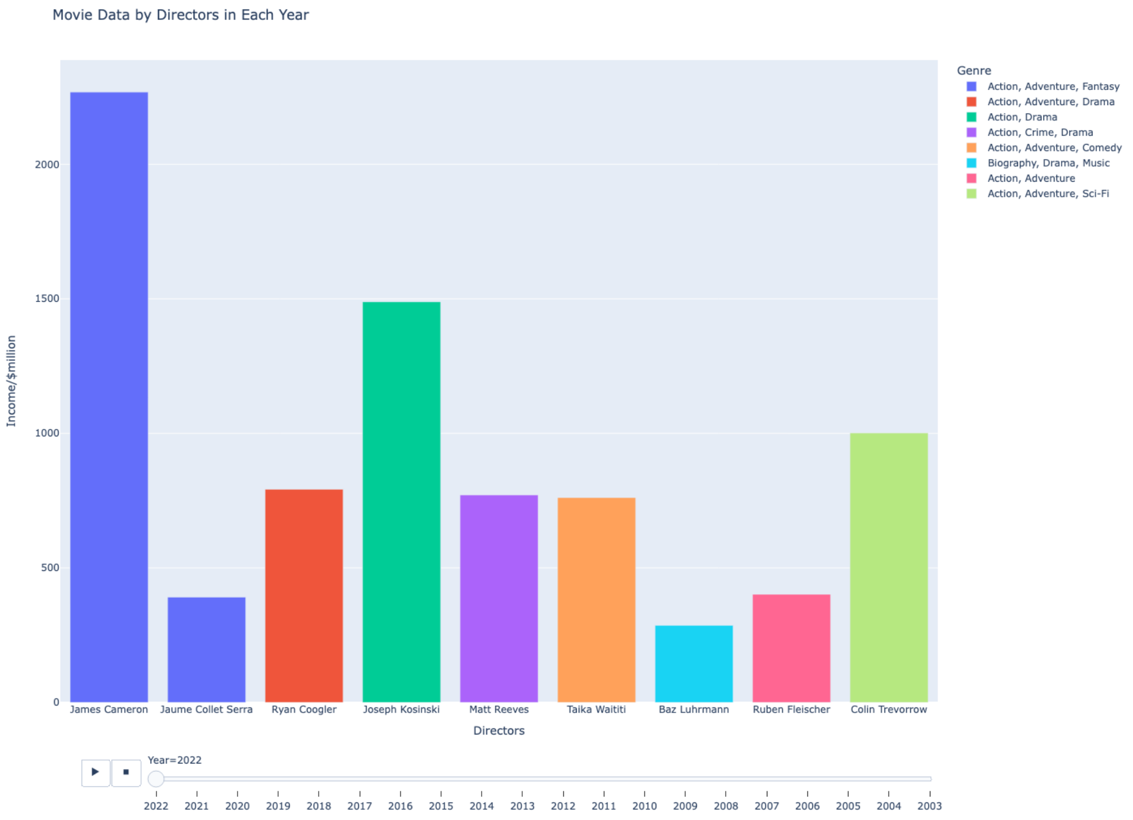
In summary, firstly, established IPs contribute significantly to generating high income in the movie industry. Secondly, action elements are prevalent in high-income movies. Thirdly, famous directors appeal more to the market and gain higher income.

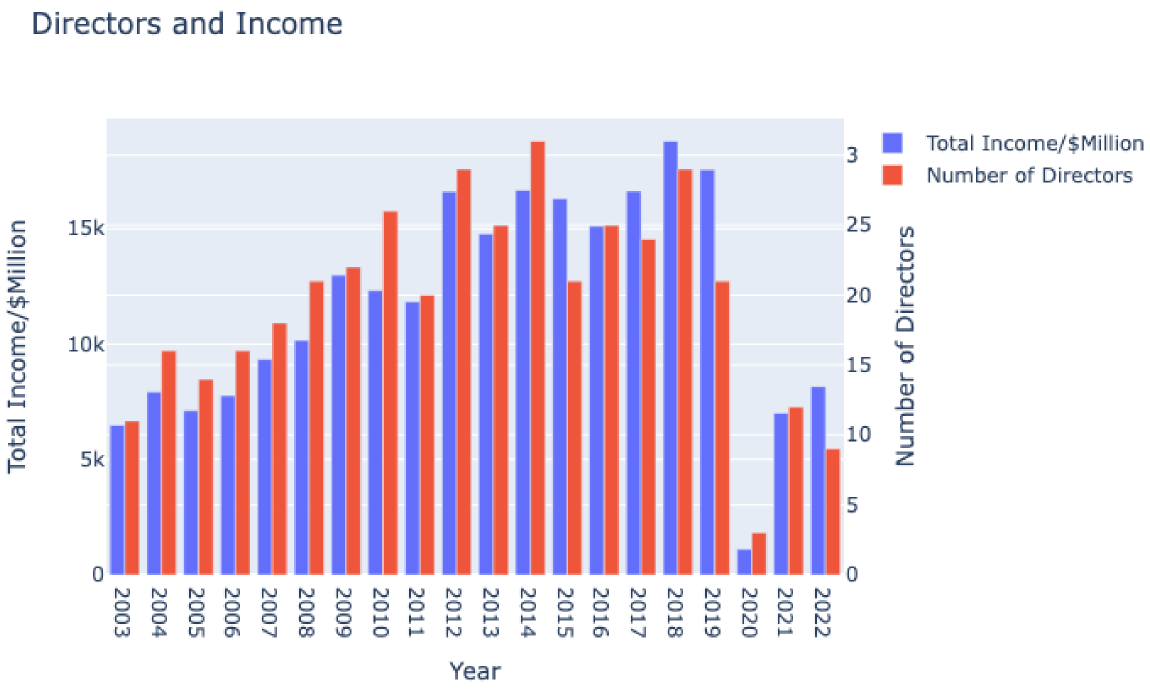
Based on above findings, we provide suggestions as follows:

1. Movie producers can invest on famous directors' new movies, especially investing on the sequels of the famous IPs.

2. One of the helpful promotion strategies is to highlight the director as an attraction point, if the movie is directed by directors that have a high level of reputation.
3. Movie producers can consider making action movies for higher financial income.

Visualisation 3





Storytelling 3

We analyse the visualisations above to find insights about how high income movies evolved over years. From the first dynamic graph, while we hover through each year, we can find the general trends are as follows: though the number of directors fluctuate, there is an increasing trend from 03 to 2012 overall. The number of directors peaked in 2018, in which year there are more than 25 directors producing new movies. Then it begins to generally decrease, failing rapidly to only 3 directors in 2020.

The income keeps the similar trend as the number of directors, with some minor differences, such as the income ranking at the peak in 2014. After 2020, the movie industry will recover, as both income and number of directors pick up, but both of them are no longer as high as previous years.

When looking at Genre data, we can find that in most years, there are various different types of movies released. Though the tags of each genre consist of usually more than 3 types, Action types appear in a very high frequency.

In conclusion, the movie industry is developing from 03 to 19 but rapidly decreases in 2020. This rapidly decreasing is possibly because of the COVID-19 epidemic from 2020, and the whole movie industry is recovering slowly in the following years. Additionally, the movies are in different types, while the action is always the common element.

Most of the above findings are responded with above 2 other visualisations. When the Time dimension is added to data analysis, we find that the COVID-19 epidemic period has been a shark point and hard time for all movies and directors. So we suggest that the movie industry employers and employees stay adaptable and agile in the face of challenges. Continuously evaluating and adjusting their strategies to navigate through uncertain times and seize opportunities can help both the movie makers and the industry recovery and growth.

Visual Analysis System

The visual analysis system is constructed based on contexts of diverse visualisation and aims to display more information, allow to conduct a higher level and more intuitive visualisation analysis. The details can be viewed in our system.

3.2. Discussion

Limitation

Firstly, as our report mainly focuses on income, and our data spans a time range of nearly 20 years, an unavoidable problem is inflation. Though the income is presented in American dollars, it is essential to acknowledge that inflation can impact the true value of these earnings. The fluctuating purchasing power of the currency over time may lead to distortions in the comparative analysis of income levels.

Secondly, it is important to acknowledge that our data set is extracted from the top 100 movies on IMDB, which introduces a limitation in terms of data set bias. There are numerous other popular and high-income movies that may not be included in the IMDB website. This limitation restricts our ability to make predictions or draw conclusions about the entire global movie market.

Thirdly, the size of our data set is relatively small, consisting of only 396 values. This can make the conclusions drawn from the analysis less reliable. There can also be sampling error and limited representativeness. To ensure more accurate monitoring of data and support informed decision-making for movie makers in producing new movies, it is crucial to acquire a larger and more diverse dataset.

Summary

A common theme observed in all above visualisations is the prominence of Intellectual Properties (IPs). The high-income movies mentioned in the top rankings are predominantly well-known series movies that have gained international recognition. These movies have transcended the realm of mere entertainment and have become iconic properties that captivate audiences with their unique storylines and immersive worlds. For example, the magical and enchanting universe of the Harry Potter movies has resonated with people worldwide, appealing to their emotions and captivating their hearts.

Those famous directors who enjoy high income either created those famous movies and made them as an IP, or made the movies on existing IP and enhanced their influence. The high income movies feed back to the directors and improve their reputation. So the audience are paying for both the IP and the directors.

Another recurring word that appears frequently is ‘action’. Specifically, a significant number of movies within our dataset incorporate elements of action. This observation suggests that there is a strong demand among audiences for action-oriented films. The popularity of action movies underscores the audience's inclination towards high-octane and visually engaging storytelling, which further emphasises the commercial viability of this genre.

Overall the income of movies is evolving quickly, except the low peak of 2020, but it proves the importance of data analysis that can monitor trends, identify patterns and furthermore analyse audience preference.

Therefore, we suggest the movie producers and the directors or the whole team to make their movie as an IP, this includes keeping an eye on emerging genres, popular themes, and new storytelling techniques that resonate with audiences. If it is hard to create an attractive new IP initially, it is also a good option to start from cooperating with those famous directors, and adapting movies from popular novels or mangas.

While considering income potential, prioritising quality and unique storytelling is also very important. A compelling and well-executed story can attract viewers and contribute to the success of a movie both in finance and reputation of the team.

The impact of the COVID-19 pandemic on the global market has been significant. Data analysis plays a crucial role in helping movie directors and producers navigate these uncertain times by providing insights into the latest market demands and trends. The ability to analyse data empowers the movie industry to anticipate and respond to market shifts, enabling a more efficient recovery process, and we expect the blooming of excellent movies around the world in the future years.

4. References

1, “Plotly express in Python.” *Plotly*, https://plotly.com/python/plotly-express/. Accessed 28 May 2023.
2. “Parallel categories diagram in Python.” *Plotly*, https://plotly.com/python/parallel-categories-diagram/. Accessed 28 May 2023.
3. “Anova test in python” http://reneshbedre.com/blog/anova.html. Accessed 28 May 2023.

5. Appendix

A: Data cleaning

```
import pandas as pd
import numpy as np

df = pd.read_csv("movies2.csv")

df = df.dropna(how='all')
df = df[df['Income'] != "Unknown"]
df = df[df['Runtime'] != "Unknown"]
df = df[df['Budget'] != "Unknown"]
df = df[~np.isnan(df['Rating'])]

df["Income_dollar"] = ["".join(e.split("$")[1].split(",")) for e in
df['Income']]
df["Income_dollar"] = df['Income_dollar'].astype(float)
df['Income_million'] = [round(x/1000000,2) for x in df['Income_dollar']]

df = df[df['Budget'].str.startswith('$')]
df['Budget_dollar'] = ["".join(e.split("$")[1].split(",")) for e in
df['Budget']]
df["Budget_dollar"] = df['Budget_dollar'].astype(int)
df['Budget_million'] = [x/1000000 for x in df['Budget_dollar']]

df['Runtime'] = df['Runtime'].astype(int)

df.Income_million.quantile([0.25,0.5,0.75])
df_high_income = df[df['Income_million'] >= 256.70]
df_high_income.to_csv("movies_high_income.csv")
```

B: Graph Data

```
import pandas as pd
import plotly.express as px

df = pd.read_csv("movies_high_income.csv")

fig = px.treemap(df, color = 'Income_million', values = 'Income_million',
             path = ['Directors', 'Title'], hover_name = 'Genre',
             title="Treemap of High Income Movies By Directors from 2003 to 2022",
             height=800, width=1150, color_continuous_scale='ylgnbu')

fig.update_traces(root_color="lightgrey")

fig.show()
```

C: High Dimensional Data

```
import pandas as pd
import plotly.express as px
import plotly.graph_objects as go

df = pd.read_csv("movies_high_income.csv")
df_sub = df[df['Directors'].isin(['James Cameron', 'David Yates', 'Anthony Russo, Joe
Russo', 'Christopher Nolan',
                                'Peter Jackson'])]

from scipy.stats import f_oneway
CategoryGroupLists=df.groupby('Directors')['Income_million'].apply(list)
AnovaResults = f_oneway(*CategoryGroupLists)
print('P-Value for Anova test with full data set is: ', AnovaResults[1])

CategoryGroupLists=df_sub.groupby('Directors')['Income_million'].apply(list)
AnovaResults = f_oneway(*CategoryGroupLists)
print('P-Value for Anova test with Top 5 directors is: ', AnovaResults[1])

# Create dimensions
genre_dim = go.parcats.Dimension(
   values=df_sub.Genre,
   categoryorder='category ascending', label="Genre"
)

director_dim = go.parcats.Dimension(values=df_sub.Directors, label="Director/s")

stars_dim = go.parcats.Dimension(
   values=df_sub.Stars, label="Country"
)

title_dim = go.parcats.Dimension(
   values=df_sub.Title, label="Movie Name"
)

# Create parcats trace
color = df_sub.Income_million

fig = go.Figure(data = [go.Parcats(dimensions=[genre_dim, stars_dim,
director_dim,title_dim],
       line={'color': color, 'colorscale': 'ylgnbu'},
       hoveron='color', hoverinfo='count+probability',
       labelfont={'size': 18, 'family': 'Times'},
       tickfont={'size': 10, 'family': 'Times'},
       arrangement='freeform'),
      ],
       layout=go.Layout(title = "Top 5 Income Directors (Income High(dark) to Low)"
                        ))


fig.show()
```

D: Dynamic Data - Animation

```
import pandas as pd
import plotly.express as px

df = pd.read_csv("movies_high_income.csv")

fig = px.bar(df, x="Directors", y="Income_million",
            color="Genre",
                animation_frame="Year",
                animation_group="Directors",
          hover_data=['Directors','Title', 'Stars'],
          title="Movie Data by Directors in Each Year")

fig.update_layout(yaxis_title = "Income/$million",
               legend_title = "Genre")
fig.show()
```

E: VA system - Website within all visualisations: please look through all files in the VA folder for full codes.

F: VA system - dynamic web page (codes in VA/pages/dynamic.py)

```
import dash
from dash import html, dcc, callback, Input, Output
import pandas as pd
import plotly.express as px

import plotly.graph_objects as go
from plotly.subplots import make_subplots

dash.register_page(
   __name__,
   path='/analytics_story_line',
   title='Dynamic Visulization',
   name='Dynamic Visulization'
)

df = pd.read_csv("../movies_high_income.csv")
all_directors = df['Directors'].unique()
```

```
years = df['Year'].unique()
years_num = years.astype(int)
df['Year'] = df['Year'].astype(int)


def render_income_year():
    fig = make_subplots(specs=[[{"secondary_y": True}]])
    avg = []
    num_dit = []
    sum = []
    #print(years)
    for i in years:
        df_sub2 = df[df['Year'] == int(i)]
        sum2 = 0
        num_dit.append(len(df_sub2['Directors'].unique()))
        for e in df_sub2['Income_million']:
            sum2 += e
        avg.append(round(sum2/df_sub2.shape[0],2))
        sum.append(sum2)
        #print(i, sum2, df_sub2['Directors'].unique())
    # Add traces
    #fig = make_subplots(specs=[[{"secondary_y": True}]])
    fig = go.Figure(
    data=[
        go.Bar(name='Total Income/$Million', x=years, y=sum, yaxis='y',
offsetgroup=1),
        go.Bar(name='Number of Directors', x=years, y=num_dit,
yaxis='y2', offsetgroup=2)
    ],
    layout={
        'yaxis': {'title': 'Total Income/$Million'},
        'yaxis2': {'title': 'Number of Directors', 'overlaying': 'y',
'side': 'right'}
    }
)

    # Add figure title
    fig.update_layout(
        title_text="Directors and Income",
        barmode='group',
        xaxis =dict(tickmode = 'array',
                    tickvals = years)
    )

    # Set x-axis title
    fig.update_xaxes(title_text="Year")
```

```
   return html.Div(dcc.Graph(figure = fig))


layout = html.Div(children=[
   html.H2("Story Line of Movie Data from "+str(years_num.min())+" to
"+str(years_num.max())),
   html.H3("Overview"),
   render_income_year(),
   html.H3("Dynamic Visulisations of High Income Movies by Years"),
   html.Div([
        "Select to visulise graph in each year: ",
       dcc.RadioItems(
                      years,
       value="2022",
       id='year-select', inline=True),
       html.Div(id='year-output'),
       dcc.Graph(id = 'year-graph', figure={}),
       html.Br(),
       html.Div("Click the play button in the below graph to view the
animation version",
                className="Note"),
       html.Iframe(
               src="./assets/dynamic2.html",
               style={"height": "1067px", "width": "100%"},
           )
   ]),
   html.Br(),
   html.H3(children='Dynamic Visulisations of High Income Movies by
Directors'),
   html.Div([
       "Select Directors: ",
       dcc.Dropdown(all_directors,
       value="Steven Spielberg",
       id='director-select')
   ]),
   html.Div(id='director-output'),
   dcc.Graph(id = 'director_graph', figure={})
])


@callback(
   Output(component_id='director-output',
component_property='children'),
   Input(component_id='director-select', component_property='value')
)
```

```
def update_director_selected(input_value):
   return f'You selected: {input_value}'

@callback(
   Output(component_id='year-output', component_property='children'),
   Input(component_id='year-select', component_property='value')
)
def update_year_selected(input_value):
   df_sub2 = df[df['Year'] == int(input_value)]
   sum = 0
   num_dit = len(df_sub2['Directors'].unique())
   for e in df_sub2['Income_million']:
       sum += e
   avg = round(sum/df_sub2.shape[0],2)
   return f'The total income of {input_value} is {sum}, average is {avg}
and {num_dit} directors produce new movies'

@callback(
   Output(component_id="year-graph", component_property="figure"),
   Input(component_id='year-select', component_property="value"))

def update_year_graph(input):
  # print(input)
   dfsub = df[df['Year'] == int(input)]
   #print(dfsub)
   fig = px.bar(dfsub, x="Directors", y="Income_million",
            color="Genre",
          hover_data=['Directors','Title', 'Stars'],
          title="Movie Data by Directors in "+str(input))
   fig.update_traces(width = 0.5)
   return fig

@callback(
       Output(component_id="director_graph",
component_property="figure"),
       Input(component_id='director-select',
component_property="value"))

def update_bar_chart(input_value):
   df_director = df.query("Directors == @input_value")
   min_year = df_director['Year'].min()
   max_year = df_director['Year'].max()

   fig = px.bar(df_director, x = "Year", y="Income_million",
                color="Title", hover_data=["Stars", "Genre", "Rating",
"Runtime"],
                title="High Income Movies Directored by 
"+str(input_value))
   fig.update_traces(width=0.5)
   if min_year != max_year:
       fig.update_layout(yaxis_title = "Income/$Million",
                       legend_title = "Movie names",
                       xaxis = dict(
                               tickmode = 'linear',
                               tick0 = min_year,
                               dtick = 1
                           ))
   else:
       fig.update_layout(yaxis_title = "Income/$Million",
                       legend_title = "Movie names",
                       xaxis = dict(
                               tickmode = 'array',
                               tickvals = [min_year],
                           ))
   return fig
```