Project 1

March 27, 2024

1 Project: Investigate a Dataset (TMD 5000 Movie Dataset)

1.1 Table of Contents

Introduction

[1]:
```
import pandas as pd
import numpy as np
import matplotlib
import matplotlib.pyplot as plt
#import seaborn as sns
pd.options.mode.chained_assignment = None  # default='warn'

%matplotlib inline
```

Data Wrangling

1.1.1 General Properties

[2]:
```
# loading data ...
df = pd.read_csv('tmdb-movies.csv')
```

[3]:
```
# checking info for df ...
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10866 entries, 0 to 10865
Data columns (total 21 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   id                    10866 non-null  int64
 1   imdb_id               10856 non-null  object
 2   popularity            10866 non-null  float64
```

```
 3   budget                10866 non-null  int64
 4   revenue               10866 non-null  int64
 5   original_title        10866 non-null  object
 6   cast                  10790 non-null  object
 7   homepage              2936 non-null   object
 8   director              10822 non-null  object
 9   tagline               8042 non-null   object
 10  keywords              9373 non-null   object
 11  overview              10862 non-null  object
 12  runtime               10866 non-null  int64
 13  genres                10843 non-null  object
 14  production_companies  9836 non-null   object
 15  release_date          10866 non-null  object
 16  vote_count            10866 non-null  int64
 17  vote_average          10866 non-null  float64
 18  release_year          10866 non-null  int64
 19  budget_adj            10866 non-null  float64
 20  revenue_adj           10866 non-null  float64
dtypes: float64(4), int64(6), object(11)
memory usage: 1.7+ MB
```

[4]:
```
# viewing the data ...
df
```

[4]:
```
           id    imdb_id  popularity     budget     revenue  \
0      135397  tt0369610   32.985763  150000000  1513528810
1       76341  tt1392190   28.419936  150000000   378436354
2      262500  tt2908446   13.112507  110000000   295238201
3      140607  tt2488496   11.173104  200000000  2068178225
4      168259  tt2820852    9.335014  190000000  1506249360
…         …          …           …          …           …
10861      21  tt0060371    0.080598          0           0
10862   20379  tt0060472    0.065543          0           0
10863   39768  tt0060161    0.065141          0           0
10864   21449  tt0061177    0.064317          0           0
10865   22293  tt0060666    0.035919      19000           0

                     original_title  \
0                    Jurassic World
1                Mad Max: Fury Road
2                         Insurgent
3      Star Wars: The Force Awakens
4                         Furious 7
…                               …
10861            The Endless Summer
10862                    Grand Prix
10863           Beregis Avtomobilya
```

```
10864       What's Up, Tiger Lily?
10865      Manos: The Hands of Fate

                                                    cast  \
0      Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi…
1      Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic…
2      Shailene Woodley|Theo James|Kate Winslet|Ansel…
3      Harrison Ford|Mark Hamill|Carrie Fisher|Adam D…
4      Vin Diesel|Paul Walker|Jason Statham|Michelle …
…                                                  …
10861  Michael Hynson|Robert August|Lord 'Tally Ho' B…
10862  James Garner|Eva Marie Saint|Yves Montand|Tosh…
10863  Innokentiy Smoktunovskiy|Oleg Efremov|Georgi Z…
10864  Tatsuya Mihashi|Akiko Wakabayashi|Mie Hama|Joh…
10865  Harold P. Warren|Tom Neyman|John Reynolds|Dian…

                                                homepage           director  \
0                        http://www.jurassicworld.com/     Colin Trevorrow
1                          http://www.madmaxmovie.com/       George Miller
2         http://www.thedivergentseries.movie/#insurgent   Robert Schwentke
3      http://www.starwars.com/films/star-wars-episod…       J.J. Abrams
4                             http://www.furious7.com/           James Wan
…                                                  …                 …
10861                                              NaN         Bruce Brown
10862                                              NaN  John Frankenheimer
10863                                              NaN     Eldar Ryazanov
10864                                              NaN         Woody Allen
10865                                              NaN    Harold P. Warren

                                                 tagline  …  \
0                                     The park is open.  …
1                                    What a Lovely Day.  …
2                                One Choice Can Destroy You  …
3                           Every generation has a story.  …
4                                      Vengeance Hits Home  …
…                                                  …  …
10861                                              NaN  …
10862  Cinerama sweeps YOU into a drama of speed and …  …
10863                                              NaN  …
10864                            WOODY ALLEN STRIKES BACK!  …
10865        It's Shocking! It's Beyond Your Imagination!  …

                                                overview runtime  \
0      Twenty-two years after the events of Jurassic …     124
1      An apocalyptic story set in the furthest reach…     120
2      Beatrice Prior must confront her inner demons …     119
3      Thirty years after defeating the Galactic Empi…     136
```

```
4      Deckard Shaw seeks revenge against Dominic Tor…     137
…                                                …     …
10861  The Endless Summer, by Bruce Brown, is one of …      95
10862  Grand Prix driver Pete Aron is fired by his te…     176
10863  An insurance agent who moonlights as a carthie…      94
10864  In comic Woody Allen's film debut, he took the…      80
10865  A family gets lost on the road and stumbles up…      74

                                          genres  \
0      Action|Adventure|Science Fiction|Thriller
1      Action|Adventure|Science Fiction|Thriller
2             Adventure|Science Fiction|Thriller
3       Action|Adventure|Science Fiction|Fantasy
4                          Action|Crime|Thriller
…                                            …
10861                                Documentary
10862                     Action|Adventure|Drama
10863                             Mystery|Comedy
10864                              Action|Comedy
10865                                     Horror

                                    production_companies release_date  \
0      Universal Studios|Amblin Entertainment|Legenda…       6/9/15
1      Village Roadshow Pictures|Kennedy Miller Produ…      5/13/15
2      Summit Entertainment|Mandeville Films|Red Wago…      3/18/15
3              Lucasfilm|Truenorth Productions|Bad Robot     12/15/15
4      Universal Pictures|Original Film|Media Rights …       4/1/15
…                                                …          …
10861                                  Bruce Brown Films      6/15/66
10862  Cherokee Productions|Joel Productions|Douglas …     12/21/66
10863                                            Mosfilm       1/1/66
10864                             Benedict Pictures Corp.      11/2/66
10865                                          Norm-Iris     11/15/66

      vote_count  vote_average  release_year   budget_adj   revenue_adj
0           5562           6.5          2015 1.379999e+08  1.392446e+09
1           6185           7.1          2015 1.379999e+08  3.481613e+08
2           2480           6.3          2015 1.012000e+08  2.716190e+08
3           5292           7.5          2015 1.839999e+08  1.902723e+09
4           2947           7.3          2015 1.747999e+08  1.385749e+09
…            …            …            …            …            …
10861         11           7.4          1966 0.000000e+00  0.000000e+00
10862         20           5.7          1966 0.000000e+00  0.000000e+00
10863         11           6.5          1966 0.000000e+00  0.000000e+00
10864         22           5.4          1966 0.000000e+00  0.000000e+00
10865         15           1.5          1966 1.276423e+05  0.000000e+00
```

```
[10866 rows x 21 columns]
```

Now that we have our data, we want to check for NaN values and clean our data …

[5]:
```python
# using .isnull() then .sum() to figure out how many NaN values each column
↪contains ...
df.isnull().sum()
```

[5]:
```
id                         0
imdb_id                   10
popularity                 0
budget                     0
revenue                    0
original_title             0
cast                      76
homepage                7930
director                  44
tagline                 2824
keywords                1493
overview                   4
runtime                    0
genres                    23
production_companies    1030
release_date               0
vote_count                 0
vote_average               0
release_year               0
budget_adj                 0
revenue_adj                0
dtype: int64
```

We now see:

- imdb_id has 10.
- cast has 76.
- homepage has 7930.
- director has 44.
- tagline has 2824.
- keywords has 1493.
- overview has 4.
- genres has 23.
- production_companies has 1030.

1.1.2 We now ask questions about the data and from there we determine which NaN values to get rid of:

- Which genre was the most common from year to year?
- Does runtime affect the movie's revenues?

1.1.3 Now we choose which NaN values to drop in accordance with our questions:

- Looking at imdb_id, we notice that it is just the ids for the movies, which doesn't affect our analysis. Hence, we leave its NaN values.
- Looking at cast, it doesn't have an effect either. Like imdb_id, we leave its NaN values.
- Looking at homepage, it lists the homepages of the movies. We leave its NaN values.
- Looking at director, it lists the directors (obviously). We leave its NaN values.
- Looking at tagline, it lists the taglines (catchphrases), which aren't important to our analysis. We leave its NaN values.
- Looking at keywords, it lists key search words, again, not important to our analysis. We leave its NaN values.
- Looking at overview, it lists the summary and main theme of the movie. We leave its NaN values.
- Looking at genres, it lists the genres corresponding to the movies, we need those for our questions, so we drop its NaN values.
- Looking at production_companies, it lists the companies involved in the production of the movie. We leave its NaN values.

[6]:
```
# dropping the rows for 'genre'
df.dropna(subset=['genres'], inplace = True)
# resetting the indicies ...
df.reset_index(inplace = True, drop = True)
df
```

[6]:
```
           id    imdb_id  popularity     budget     revenue  \
0      135397  tt0369610   32.985763  150000000  1513528810
1       76341  tt1392190   28.419936  150000000   378436354
2      262500  tt2908446   13.112507  110000000   295238201
3      140607  tt2488496   11.173104  200000000  2068178225
4      168259  tt2820852    9.335014  190000000  1506249360
…         …         …          …          …          …
10838      21  tt0060371    0.080598          0           0
10839   20379  tt0060472    0.065543          0           0
10840   39768  tt0060161    0.065141          0           0
10841   21449  tt0061177    0.064317          0           0
10842   22293  tt0060666    0.035919      19000           0

                     original_title  \
0                    Jurassic World
1                Mad Max: Fury Road
2                         Insurgent
3      Star Wars: The Force Awakens
4                         Furious 7
…                                 …
10838            The Endless Summer
10839                    Grand Prix
10840           Beregis Avtomobilya
10841          What's Up, Tiger Lily?
```

```
10842     Manos: The Hands of Fate

                                                    cast  \
0      Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi…
1      Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic…
2      Shailene Woodley|Theo James|Kate Winslet|Ansel…
3      Harrison Ford|Mark Hamill|Carrie Fisher|Adam D…
4      Vin Diesel|Paul Walker|Jason Statham|Michelle …
…                                                   …
10838  Michael Hynson|Robert August|Lord 'Tally Ho' B…
10839  James Garner|Eva Marie Saint|Yves Montand|Tosh…
10840  Innokentiy Smoktunovskiy|Oleg Efremov|Georgi Z…
10841  Tatsuya Mihashi|Akiko Wakabayashi|Mie Hama|Joh…
10842  Harold P. Warren|Tom Neyman|John Reynolds|Dian…

                                                homepage           director  \
0                          http://www.jurassicworld.com/    Colin Trevorrow
1                            http://www.madmaxmovie.com/      George Miller
2         http://www.thedivergentseries.movie/#insurgent   Robert Schwentke
3      http://www.starwars.com/films/star-wars-episod…        J.J. Abrams
4                               http://www.furious7.com/          James Wan
…                                                   …                  …
10838                                                NaN        Bruce Brown
10839                                                NaN  John Frankenheimer
10840                                                NaN     Eldar Ryazanov
10841                                                NaN        Woody Allen
10842                                                NaN   Harold P. Warren

                                                 tagline  …  \
0                                     The park is open.  …
1                                    What a Lovely Day.  …
2                                One Choice Can Destroy You  …
3                          Every generation has a story.  …
4                                   Vengeance Hits Home  …
…                                                   …  …
10838                                                NaN  …
10839  Cinerama sweeps YOU into a drama of speed and …  …
10840                                                NaN  …
10841                              WOODY ALLEN STRIKES BACK!  …
10842       It's Shocking! It's Beyond Your Imagination!  …

                                                overview runtime  \
0      Twenty-two years after the events of Jurassic …     124
1      An apocalyptic story set in the furthest reach…     120
2      Beatrice Prior must confront her inner demons …     119
3      Thirty years after defeating the Galactic Empi…     136
4      Deckard Shaw seeks revenge against Dominic Tor…     137
```

```
…                                                    …     …
10838  The Endless Summer, by Bruce Brown, is one of …      95
10839  Grand Prix driver Pete Aron is fired by his te…     176
10840  An insurance agent who moonlights as a carthie…      94
10841  In comic Woody Allen's film debut, he took the…      80
10842  A family gets lost on the road and stumbles up…      74

                                          genres  \
0      Action|Adventure|Science Fiction|Thriller
1      Action|Adventure|Science Fiction|Thriller
2             Adventure|Science Fiction|Thriller
3       Action|Adventure|Science Fiction|Fantasy
4                          Action|Crime|Thriller
…                                            …
10838                                Documentary
10839                     Action|Adventure|Drama
10840                             Mystery|Comedy
10841                              Action|Comedy
10842                                     Horror

                                    production_companies release_date  \
0      Universal Studios|Amblin Entertainment|Legenda…       6/9/15
1      Village Roadshow Pictures|Kennedy Miller Produ…      5/13/15
2      Summit Entertainment|Mandeville Films|Red Wago…      3/18/15
3              Lucasfilm|Truenorth Productions|Bad Robot     12/15/15
4      Universal Pictures|Original Film|Media Rights …       4/1/15
…                                                    …          …
10838                                  Bruce Brown Films      6/15/66
10839  Cherokee Productions|Joel Productions|Douglas …     12/21/66
10840                                            Mosfilm       1/1/66
10841                             Benedict Pictures Corp.      11/2/66
10842                                          Norm-Iris     11/15/66

      vote_count  vote_average  release_year    budget_adj   revenue_adj
0           5562           6.5          2015  1.379999e+08  1.392446e+09
1           6185           7.1          2015  1.379999e+08  3.481613e+08
2           2480           6.3          2015  1.012000e+08  2.716190e+08
3           5292           7.5          2015  1.839999e+08  1.902723e+09
4           2947           7.3          2015  1.747999e+08  1.385749e+09
…            …              …             …            …            …
10838         11           7.4          1966  0.000000e+00  0.000000e+00
10839         20           5.7          1966  0.000000e+00  0.000000e+00
10840         11           6.5          1966  0.000000e+00  0.000000e+00
10841         22           5.4          1966  0.000000e+00  0.000000e+00
10842         15           1.5          1966  1.276423e+05  0.000000e+00

[10843 rows x 21 columns]
```

[7]:
```
# running .isnull().sum() again to check ...
df.isnull().sum()
```

[7]:
```
id                         0
imdb_id                    8
popularity                 0
budget                     0
revenue                    0
original_title             0
cast                      75
homepage                7912
director                  42
tagline                 2806
keywords                1475
overview                   3
runtime                    0
genres                     0
production_companies    1016
release_date               0
vote_count                 0
vote_average               0
release_year               0
budget_adj                 0
revenue_adj                0
dtype: int64
```

1.1.4 We check for duplicates in our dataset ...

[8]:
```
df.duplicated().sum()
```

[8]:
```
1
```

[9]:
```
# dropping duplicated rows ...
df.drop_duplicates(inplace=True)
df.reset_index(inplace = True, drop = True)
df
```

[9]:
```
           id    imdb_id  popularity     budget     revenue  \
0      135397  tt0369610   32.985763  150000000  1513528810
1       76341  tt1392190   28.419936  150000000   378436354
2      262500  tt2908446   13.112507  110000000   295238201
3      140607  tt2488496   11.173104  200000000  2068178225
4      168259  tt2820852    9.335014  190000000  1506249360
…         …          …           …          …           …
10837      21  tt0060371    0.080598          0           0
10838   20379  tt0060472    0.065543          0           0
10839   39768  tt0060161    0.065141          0           0
```

```
10840   21449  tt0061177    0.064317          0           0
10841   22293  tt0060666    0.035919      19000           0

                     original_title  \
0                    Jurassic World
1                Mad Max: Fury Road
2                         Insurgent
3      Star Wars: The Force Awakens
4                         Furious 7
…                               …
10837             The Endless Summer
10838                    Grand Prix
10839           Beregis Avtomobilya
10840         What's Up, Tiger Lily?
10841       Manos: The Hands of Fate

                                                    cast  \
0      Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi…
1      Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic…
2      Shailene Woodley|Theo James|Kate Winslet|Ansel…
3      Harrison Ford|Mark Hamill|Carrie Fisher|Adam D…
4      Vin Diesel|Paul Walker|Jason Statham|Michelle …
…                                                   …
10837  Michael Hynson|Robert August|Lord 'Tally Ho' B…
10838  James Garner|Eva Marie Saint|Yves Montand|Tosh…
10839  Innokentiy Smoktunovskiy|Oleg Efremov|Georgi Z…
10840  Tatsuya Mihashi|Akiko Wakabayashi|Mie Hama|Joh…
10841  Harold P. Warren|Tom Neyman|John Reynolds|Dian…

                                                homepage            director  \
0                          http://www.jurassicworld.com/     Colin Trevorrow
1                            http://www.madmaxmovie.com/       George Miller
2         http://www.thedivergentseries.movie/#insurgent    Robert Schwentke
3      http://www.starwars.com/films/star-wars-episod…         J.J. Abrams
4                               http://www.furious7.com/           James Wan
…                                                   …                   …
10837                                                NaN         Bruce Brown
10838                                                NaN  John Frankenheimer
10839                                                NaN      Eldar Ryazanov
10840                                                NaN         Woody Allen
10841                                                NaN    Harold P. Warren

                                                tagline  …  \
0                                   The park is open.  …
1                                  What a Lovely Day.  …
2                              One Choice Can Destroy You  …
3                          Every generation has a story.  …
```

```
4                                    Vengeance Hits Home  …
…                                                    …  …
10837                                                NaN  …
10838  Cinerama sweeps YOU into a drama of speed and …  …
10839                                                NaN  …
10840                          WOODY ALLEN STRIKES BACK!  …
10841       It's Shocking! It's Beyond Your Imagination!  …

                                                overview runtime  \
0      Twenty-two years after the events of Jurassic …     124
1      An apocalyptic story set in the furthest reach…     120
2      Beatrice Prior must confront her inner demons …     119
3      Thirty years after defeating the Galactic Empi…     136
4      Deckard Shaw seeks revenge against Dominic Tor…     137
…                                                    …     …
10837  The Endless Summer, by Bruce Brown, is one of …      95
10838  Grand Prix driver Pete Aron is fired by his te…     176
10839  An insurance agent who moonlights as a carthie…      94
10840  In comic Woody Allen's film debut, he took the…      80
10841  A family gets lost on the road and stumbles up…      74

                                         genres  \
0      Action|Adventure|Science Fiction|Thriller
1      Action|Adventure|Science Fiction|Thriller
2             Adventure|Science Fiction|Thriller
3       Action|Adventure|Science Fiction|Fantasy
4                          Action|Crime|Thriller
…                                            …
10837                                Documentary
10838                     Action|Adventure|Drama
10839                             Mystery|Comedy
10840                              Action|Comedy
10841                                     Horror

                                  production_companies release_date  \
0      Universal Studios|Amblin Entertainment|Legenda…       6/9/15
1      Village Roadshow Pictures|Kennedy Miller Produ…      5/13/15
2      Summit Entertainment|Mandeville Films|Red Wago…      3/18/15
3              Lucasfilm|Truenorth Productions|Bad Robot     12/15/15
4      Universal Pictures|Original Film|Media Rights …       4/1/15
…                                                    …        …
10837                                Bruce Brown Films      6/15/66
10838  Cherokee Productions|Joel Productions|Douglas …     12/21/66
10839                                          Mosfilm       1/1/66
10840                            Benedict Pictures Corp.      11/2/66
10841                                        Norm-Iris     11/15/66
```

```
       vote_count  vote_average  release_year    budget_adj   revenue_adj
0            5562           6.5          2015  1.379999e+08  1.392446e+09
1            6185           7.1          2015  1.379999e+08  3.481613e+08
2            2480           6.3          2015  1.012000e+08  2.716190e+08
3            5292           7.5          2015  1.839999e+08  1.902723e+09
4            2947           7.3          2015  1.747999e+08  1.385749e+09
…               …             …             …             …             …
10837          11           7.4          1966  0.000000e+00  0.000000e+00
10838          20           5.7          1966  0.000000e+00  0.000000e+00
10839          11           6.5          1966  0.000000e+00  0.000000e+00
10840          22           5.4          1966  0.000000e+00  0.000000e+00
10841          15           1.5          1966  1.276423e+05  0.000000e+00

[10842 rows x 21 columns]
```

Exploratory Data Analysis

1.2 Now we answer our first question:

Which genre was the most common from year to year?

First, we notice that genres has pipe-seperated values (ie, genre_0|genre_1|genre_2| … |genre_n), so, we would like to seperate these values to be able to do operations and answer our question.

One way to do this is to make a series with all the seperated genres contained in a list. Once that's over, we want to create a new data frame with columns with the name of each genre, and then place 1 wherever the movie has a genre which is equal to the column's name, and leave a NaN wherever it isn't equal. Then we concatenate our original data frame and this new one.

We may find it helpful to define two functions for this process.

[10]:
```python
# defining a function to seperate all pipe-seperated values into a list ...
# col -> column to seperate values for
# char -> separating character

def seperate_into_series(col, char):
    return df[col].str.split(char)
```

[11]:
```python
# defining a function to make a new data frame with our series
# rows -> number of rows to iterate on
# col_list -> list containing desired column names (that are also the names we␣
↪filter on)
# d_series -> the series that has the data

def df_new(rows, col_list, d_series):
    df_temp = pd.DataFrame(index = range(0, rows), columns = col_list)
```

```
    for i in range(0, rows):
        list_temp = d_series[i]
        for j in col_list:
            for k in range(0, len(list_temp)):
                if list_temp[k] == j:
                    df_temp.at[i, j] = 1
    return df_temp
```

Now we desire to find what the list of all genres (col_list) is. An approach to this is we could form a dictionary that adds a genre as the key. Then we do list(dict.keys()) to get a list of all the keys (genres in this case).

We may find it helpful to define this as a function

[12]:
```
# defining a function to get the list of columns ...
# rows -> number of rows to iterate on
# d_series -> the series that has the data

def col_list(rows, d_series):
    dict_temp = {}
    for i in range(0, rows):
        list_temp = d_series[i]
        for j in range(0, len(list_temp)):
            if not list_temp[j] in dict_temp:
                dict_temp[list_temp[j]] = 1
    return list(dict_temp.keys())
```

Now we are ready to answer our first question:

[13]:
```
# first we seperate the data into a series ...
series_1a = seperate_into_series('genres', '|')
```

[14]:
```
# second we run col_list() to obtain a list of all genres.
list_1a = col_list(10842, series_1a)
```

[15]:
```
# third we create our dataframe ...
df_1a = df_new(10842, list_1a, series_1a)
```

[16]:
```
# we now concatenate df and df_1 ...
df = pd.concat([df, df_1a], axis = 1)
df
```

[16]:
```
           id    imdb_id  popularity     budget     revenue  \
0      135397  tt0369610   32.985763  150000000  1513528810
1       76341  tt1392190   28.419936  150000000   378436354
2      262500  tt2908446   13.112507  110000000   295238201
3      140607  tt2488496   11.173104  200000000  2068178225
4      168259  tt2820852    9.335014  190000000  1506249360
…           …          …           …          …           …
```

```
10837      21  tt0060371    0.080598         0           0
10838   20379  tt0060472    0.065543         0           0
10839   39768  tt0060161    0.065141         0           0
10840   21449  tt0061177    0.064317         0           0
10841   22293  tt0060666    0.035919     19000           0

                original_title  \
0               Jurassic World
1           Mad Max: Fury Road
2                    Insurgent
3      Star Wars: The Force Awakens
4                    Furious 7
…                         …
10837           The Endless Summer
10838                  Grand Prix
10839          Beregis Avtomobilya
10840       What's Up, Tiger Lily?
10841     Manos: The Hands of Fate

                                                 cast  \
0      Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi…
1      Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic…
2      Shailene Woodley|Theo James|Kate Winslet|Ansel…
3      Harrison Ford|Mark Hamill|Carrie Fisher|Adam D…
4      Vin Diesel|Paul Walker|Jason Statham|Michelle …
…                                                   …
10837  Michael Hynson|Robert August|Lord 'Tally Ho' B…
10838  James Garner|Eva Marie Saint|Yves Montand|Tosh…
10839  Innokentiy Smoktunovskiy|Oleg Efremov|Georgi Z…
10840  Tatsuya Mihashi|Akiko Wakabayashi|Mie Hama|Joh…
10841  Harold P. Warren|Tom Neyman|John Reynolds|Dian…

                                             homepage            director  \
0                       http://www.jurassicworld.com/     Colin Trevorrow
1                         http://www.madmaxmovie.com/       George Miller
2        http://www.thedivergentseries.movie/#insurgent   Robert Schwentke
3      http://www.starwars.com/films/star-wars-episod…        J.J. Abrams
4                            http://www.furious7.com/           James Wan
…                                                   …                 …
10837                                             NaN         Bruce Brown
10838                                             NaN  John Frankenheimer
10839                                             NaN      Eldar Ryazanov
10840                                             NaN         Woody Allen
10841                                             NaN    Harold P. Warren

                                              tagline  … Comedy Mystery  \
0                                  The park is open.  …    NaN     NaN
```

```
1                                          What a Lovely Day.  …    NaN     NaN
2                                 One Choice Can Destroy You  …    NaN     NaN
3                              Every generation has a story.  …    NaN     NaN
4                                        Vengeance Hits Home  …    NaN     NaN
…                                                          …   …    …      …
10837                                                    NaN  …    NaN     NaN
10838  Cinerama sweeps YOU into a drama of speed and …  …    NaN     NaN
10839                                                    NaN  …      1       1
10840                                  WOODY ALLEN STRIKES BACK!  …      1     NaN
10841       It's Shocking! It's Beyond Your Imagination!  …    NaN     NaN

       Romance  War History Music  Horror  Documentary  TV Movie  Foreign
0          NaN  NaN     NaN   NaN     NaN          NaN       NaN      NaN
1          NaN  NaN     NaN   NaN     NaN          NaN       NaN      NaN
2          NaN  NaN     NaN   NaN     NaN          NaN       NaN      NaN
3          NaN  NaN     NaN   NaN     NaN          NaN       NaN      NaN
4          NaN  NaN     NaN   NaN     NaN          NaN       NaN      NaN
…        …  …     …   …     …          …       …      …
10837      NaN  NaN     NaN   NaN     NaN            1       NaN      NaN
10838      NaN  NaN     NaN   NaN     NaN          NaN       NaN      NaN
10839      NaN  NaN     NaN   NaN     NaN          NaN       NaN      NaN
10840      NaN  NaN     NaN   NaN     NaN          NaN       NaN      NaN
10841      NaN  NaN     NaN   NaN       1          NaN       NaN      NaN

[10842 rows x 41 columns]
```

[17]:
```python
# we now want to know the range of years, which is simple ...
print(f'Maximum year is {df["release_year"].max()}')
print(f'Minimum year is {df["release_year"].min()}')
# now we create a list with all years in between (including the max and min) ...
list_years = []
for i in range(1960, 2015 + 1):
    list_years.append(i)
```

```
Maximum year is 2015
Minimum year is 1960
```

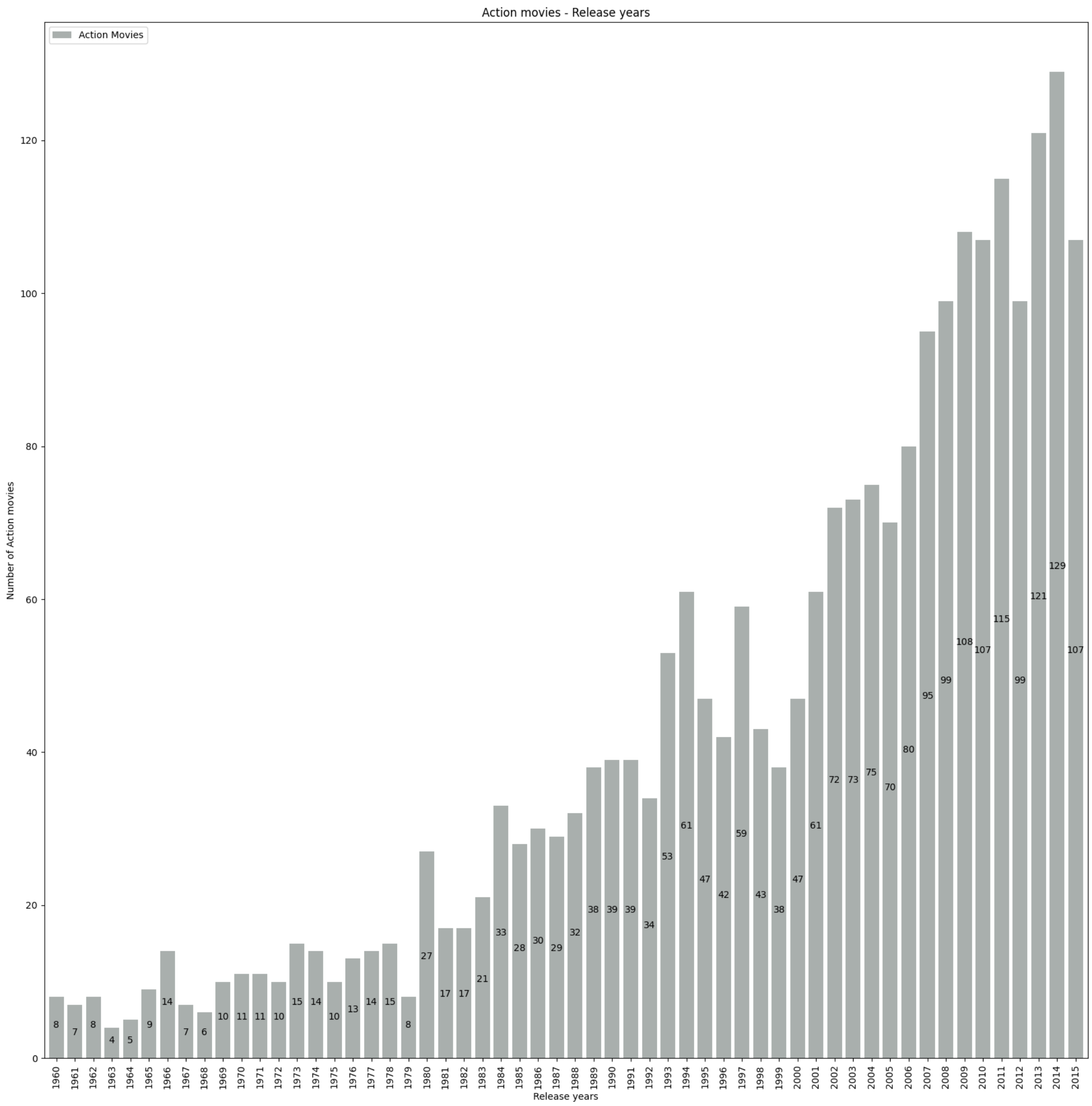
[18]:
```python
# defining a function ...
# col1 -> column to group by
# col2 -> column to get values for
# list_val -> list of values which match the values of **unique** col1

def df_part(col1, col2, list_val):
    df_temp = pd.DataFrame(df.groupby(col1)[col2].value_counts().
 ↪sort_index(ascending = True))
    list_temp = list(df_temp.index.get_level_values(0))
    dict_temp = {}
```

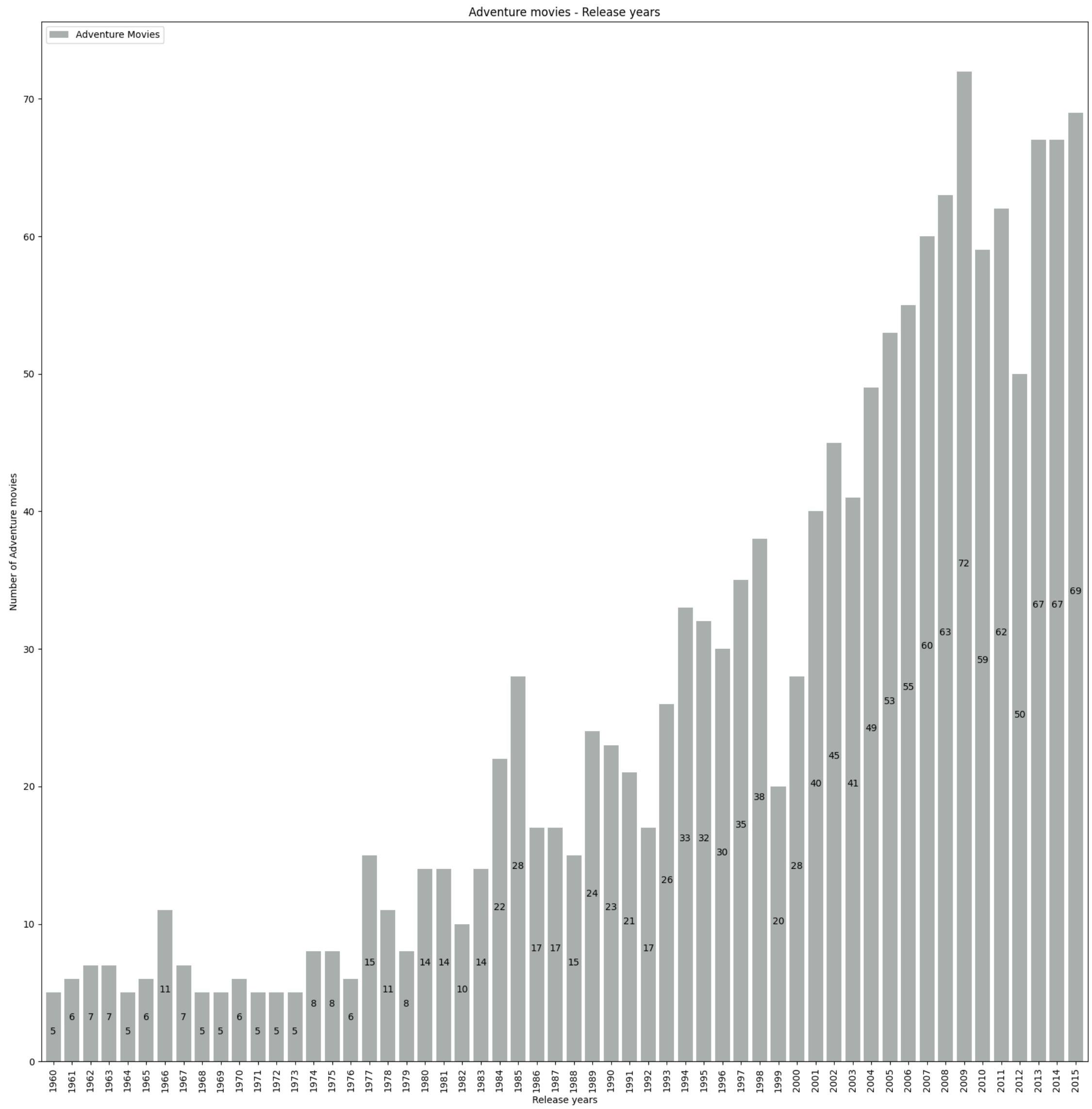
```
    for i in list_val:
        # safety check.
        if i in list_temp:
            dict_temp[i] = df_temp[col2][i][1]
    return pd.DataFrame(dict_temp, index = [col2]).T
```

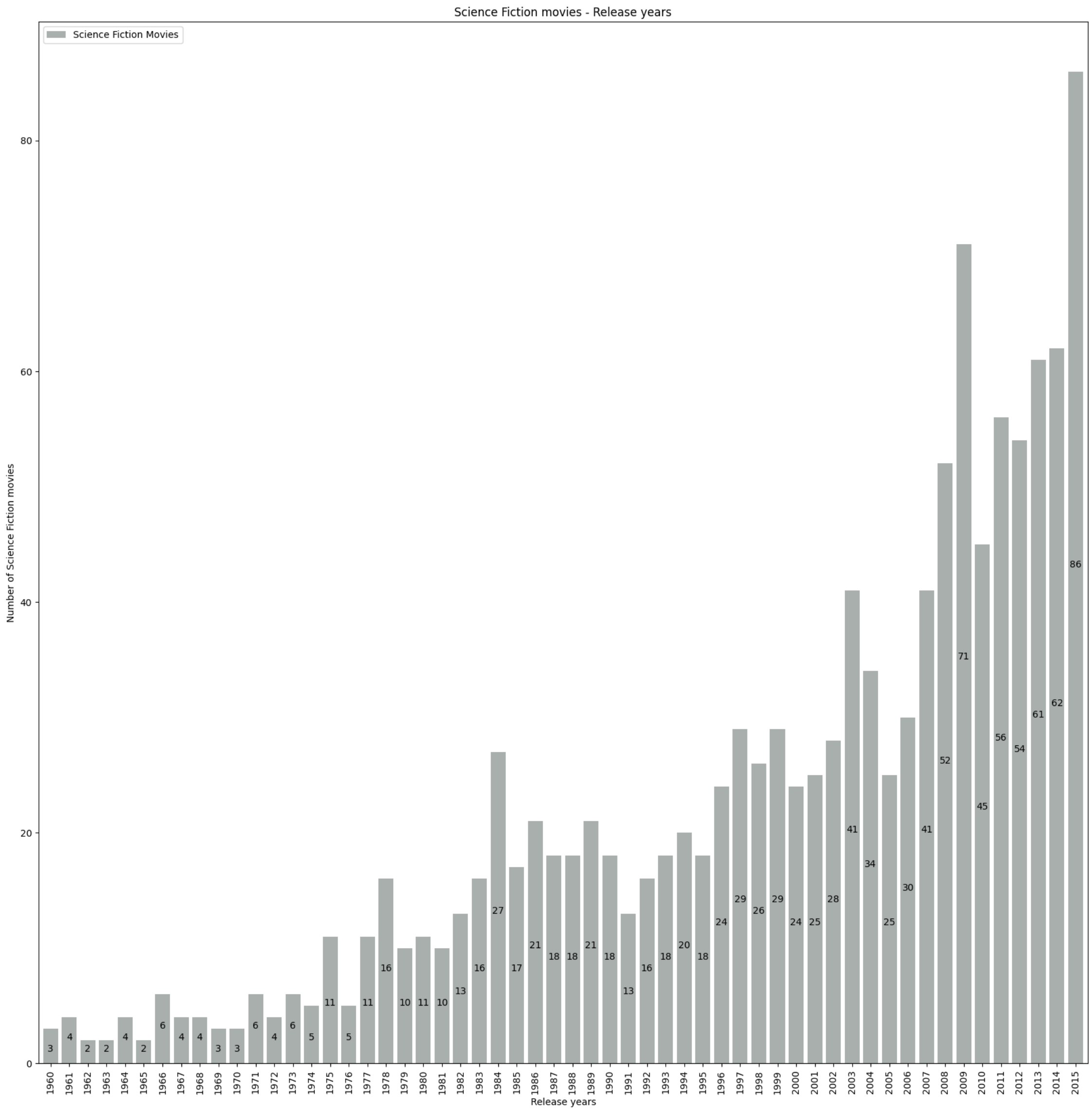
In the following 2 lines of code, we plot all the plots into one cell window. The reader is encouraged to make the window larger for ease of viewing. These plots show the trend of each genre with each passing year. We see that most of the genres follow an increasing trend as we tend to the right side of the graph, only going down a bit at 2015 sometimes.

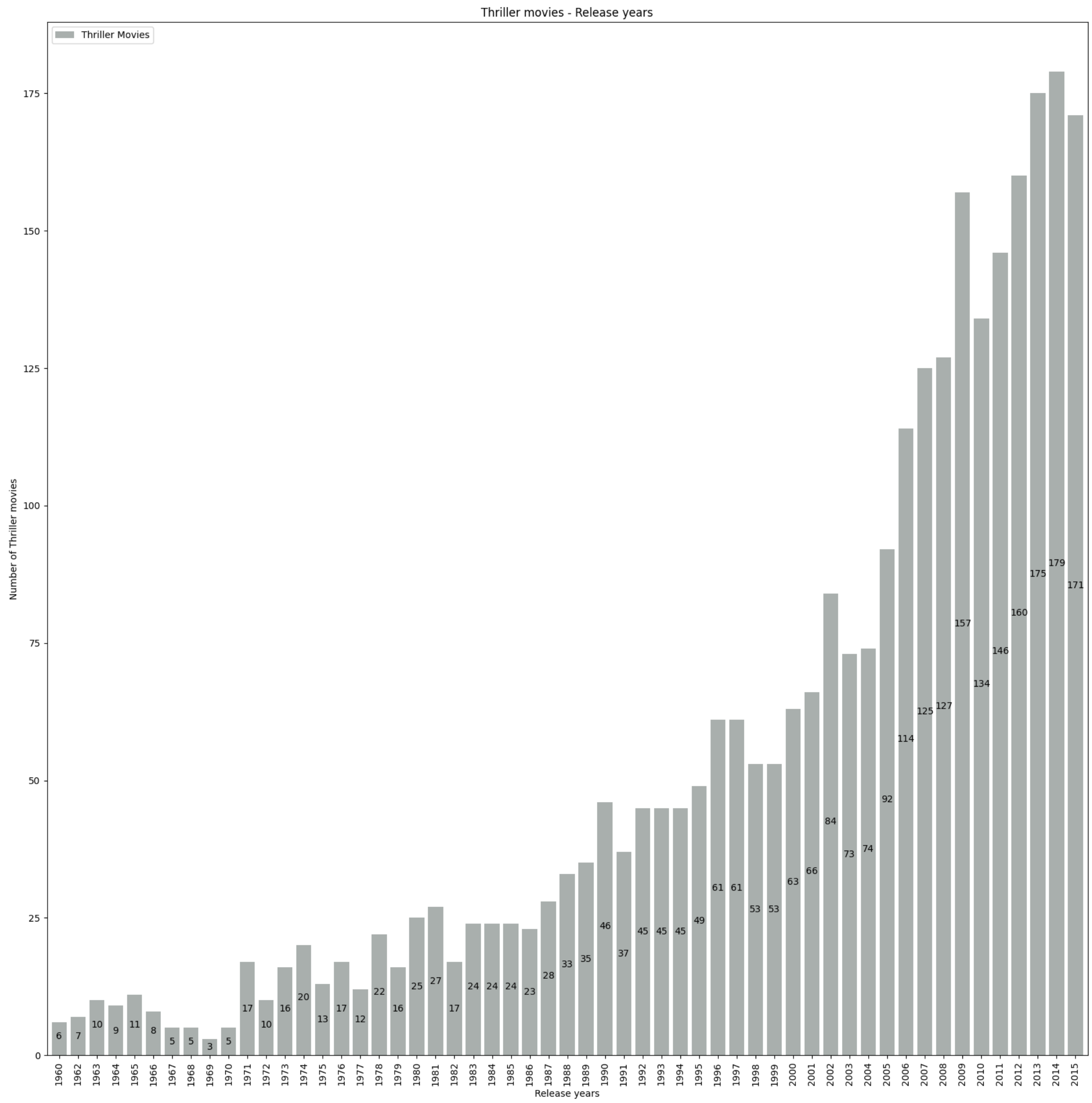
We present our answer for this question in the cell after the following one.

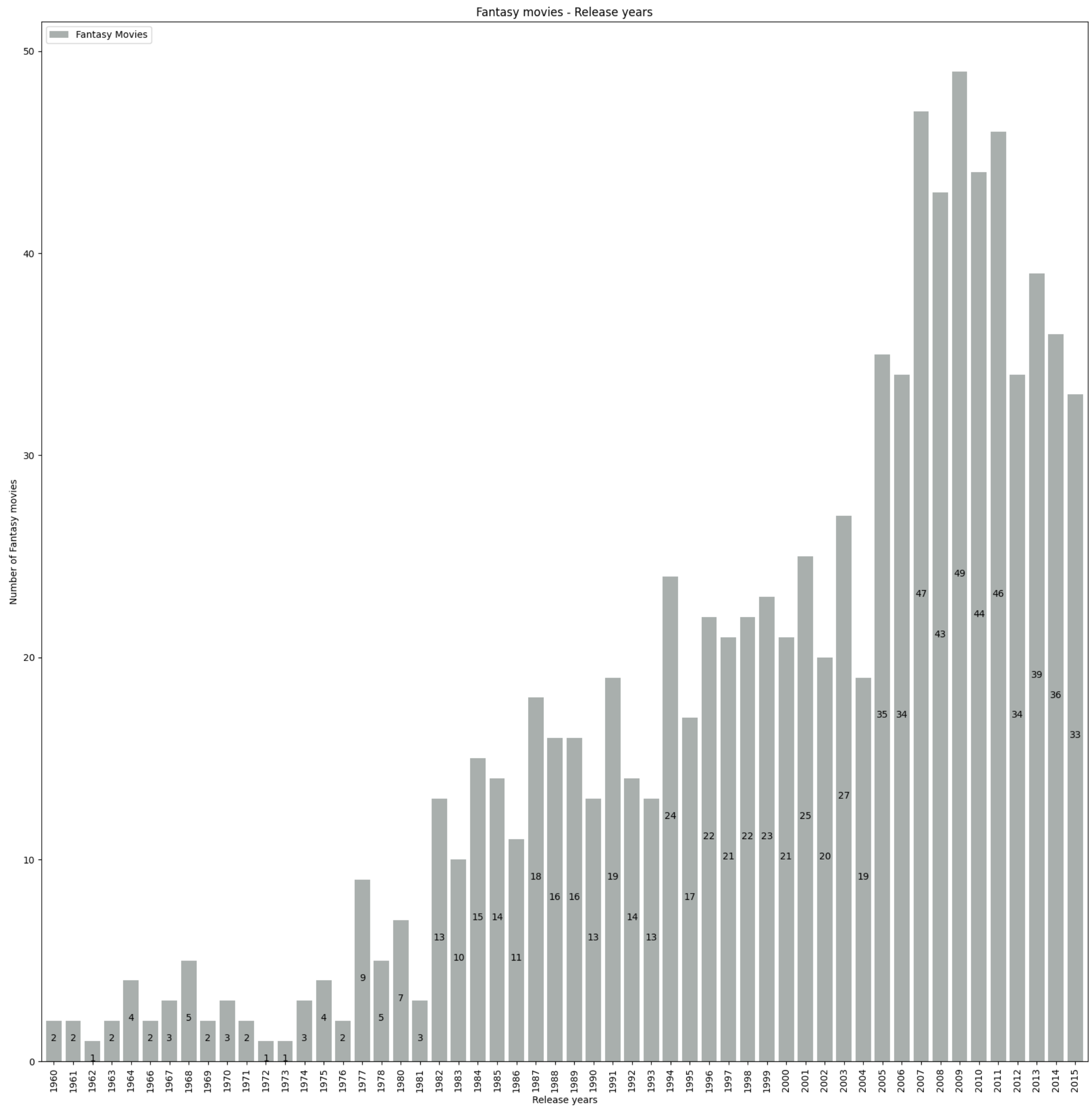
[19]:
```
# this definition of a function is available on: https://www.geeksforgeeks.org/
↪adding-value-labels-on-a-matplotlib-bar-chart/ , I reused it here.
def addlabels(x,y):
    for i in range(len(x)):
        plt.text(i, y[i]//2, y[i], ha = 'center')

for i in list_1a:
    vals = []
    df_2a = df_part('release_year', i, list_years)
    df_2a.plot(kind = 'bar', figsize = (20, 20), width=0.8, color = (0.166, 0.
↪224, 0.204, 0.4), xlabel = 'Release years', ylabel = f'Number of {i}␣
↪movies', title = f'{i} movies - Release years')
    plt.legend([f'{i} Movies'], loc = 'upper left')
    # producing a list of values to render in the box plot ...
    for j in list(df_2a[i].index.get_level_values(0)):
        vals.append(df_2a[i][j])
    # adding the values as labels ...
    addlabels(list(df_2a[i].index.get_level_values(0)), vals)
```

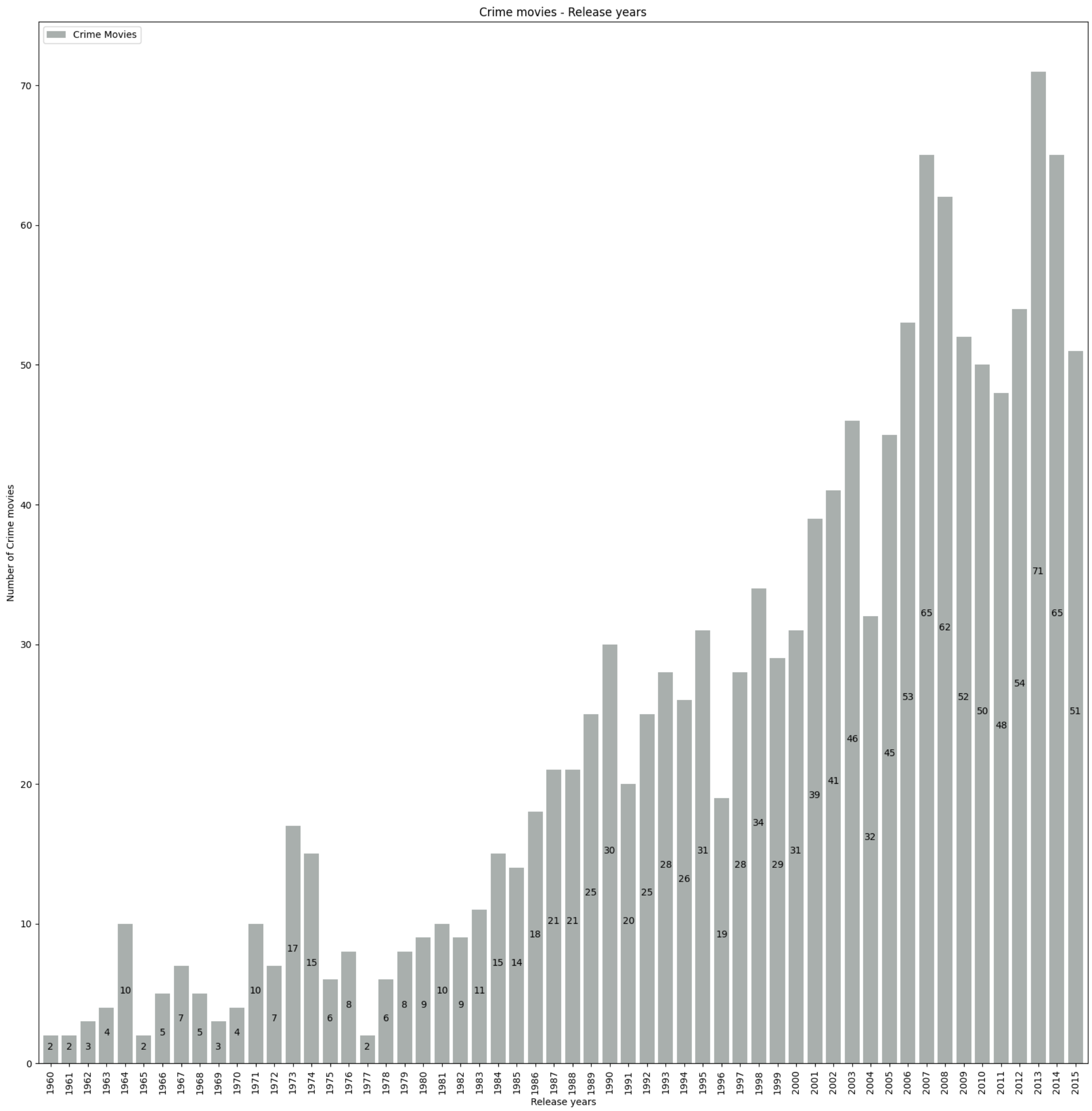

Action movies - Release years
Action Movies
Number of Action movies
0 20 40 60 80 100 120
Release years
1960 1961 1962 1963 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
8 7 8 4 5 9 14 7 6 10 11 11 10 15 14 10 13 14 15 8 27 17 17 21 33 28 30 29 32 38 39 39 34 53 61 47 42 59 43 38 47 61 72 73 75 70 80 95 99 108 107 115 99 121 129 107

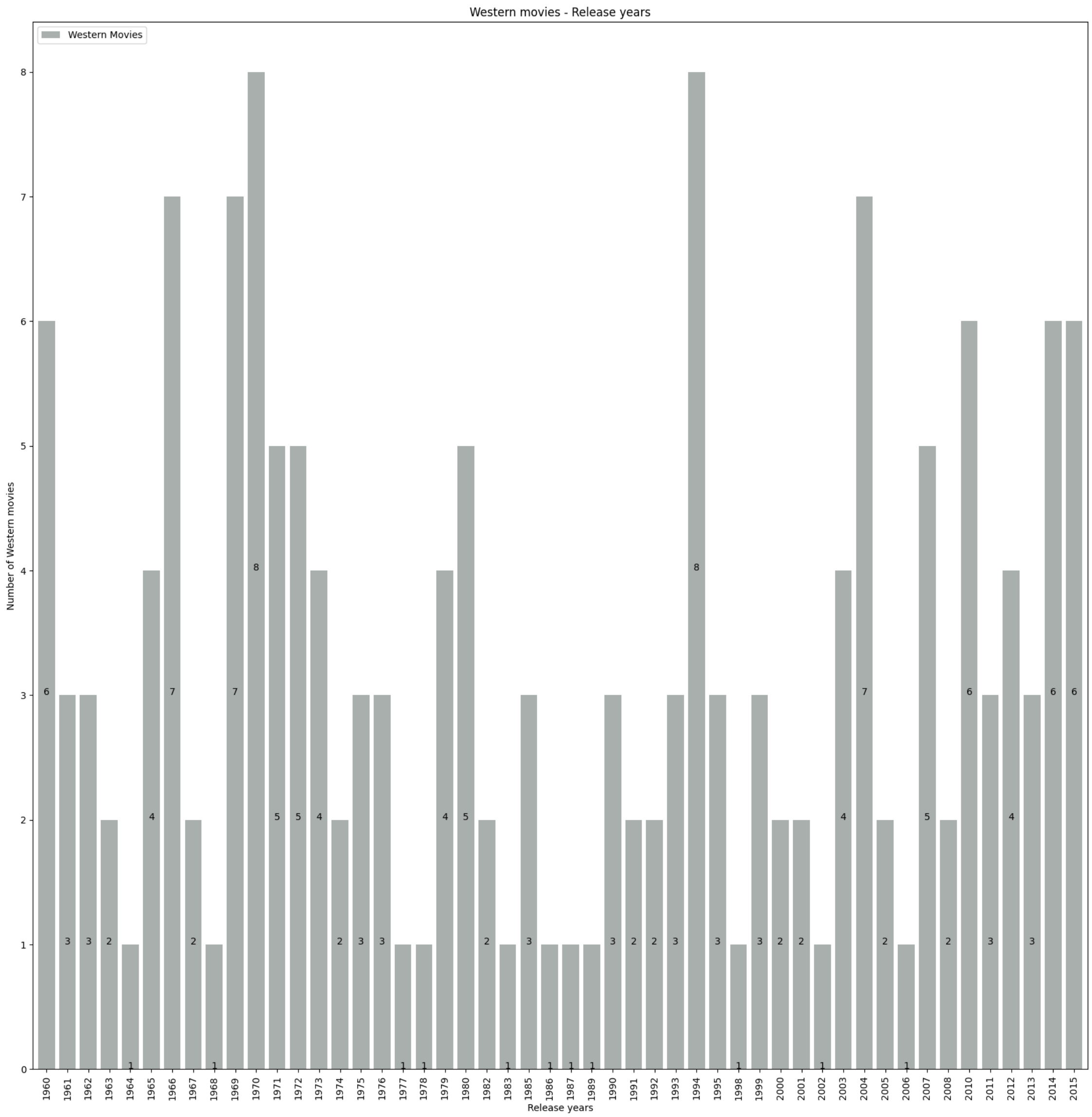

Adventure movies - Release years
Adventure Movies
Number of Adventure movies
Release years

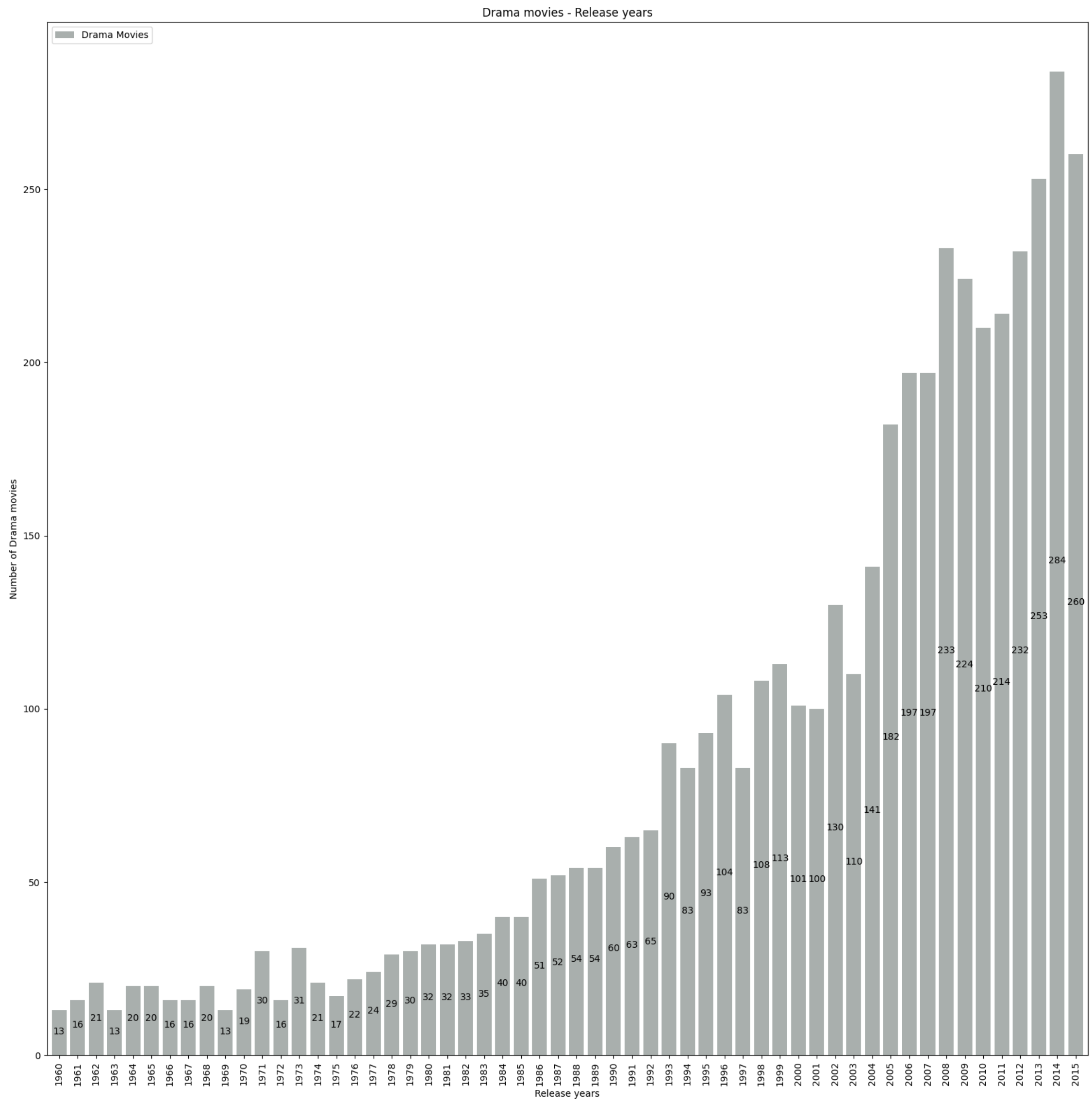

Science Fiction movies - Release years
Science Fiction Movies
Number of Science Fiction movies
Release years
0
20
40
60
80
1960 1961 1962 1963 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
3 4 2 2 4 2 6 4 4 3 3 6 4 6 5 11 5 11 16 10 11 10 13 16 27 17 21 18 18 21 18 13 16 18 20 18 24 29 26 29 24 25 28 41 34 25 30 41 52 71 45 56 54 61 62 86

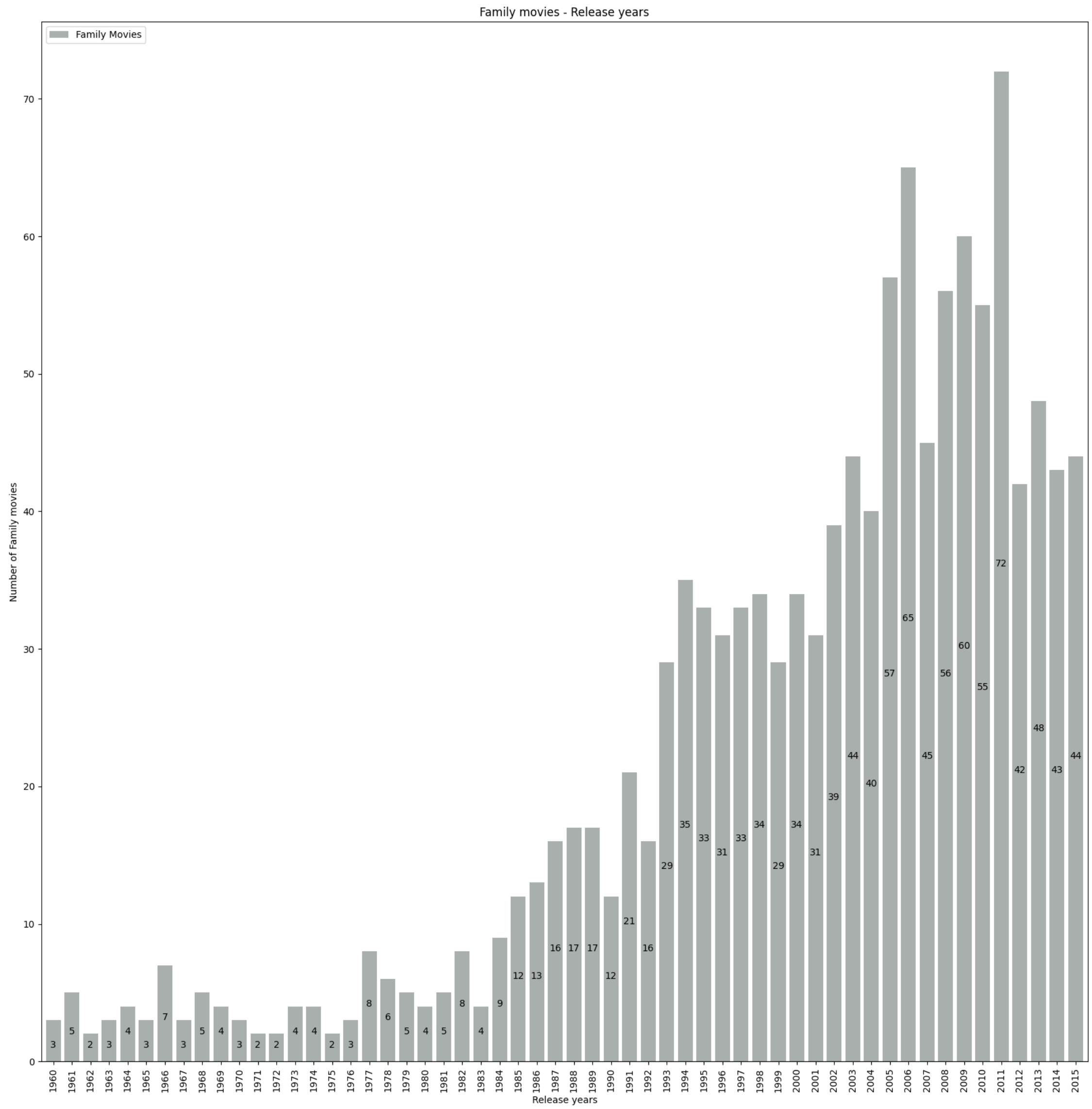

Thriller movies - Release years
Thriller Movies
Number of Thriller movies
Release years
0
25
50
75
100
125
150
175
1960 1962 1963 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
6 7 10 9 11 8 5 5 3 5 17 10 16 20 13 17 12 22 16 25 27 17 24 24 24 23 28 33 35 46 37 45 45 45 49 61 61 53 53 63 66 84 73 74 92 114 125 127 157 134 146 160 175 179 171

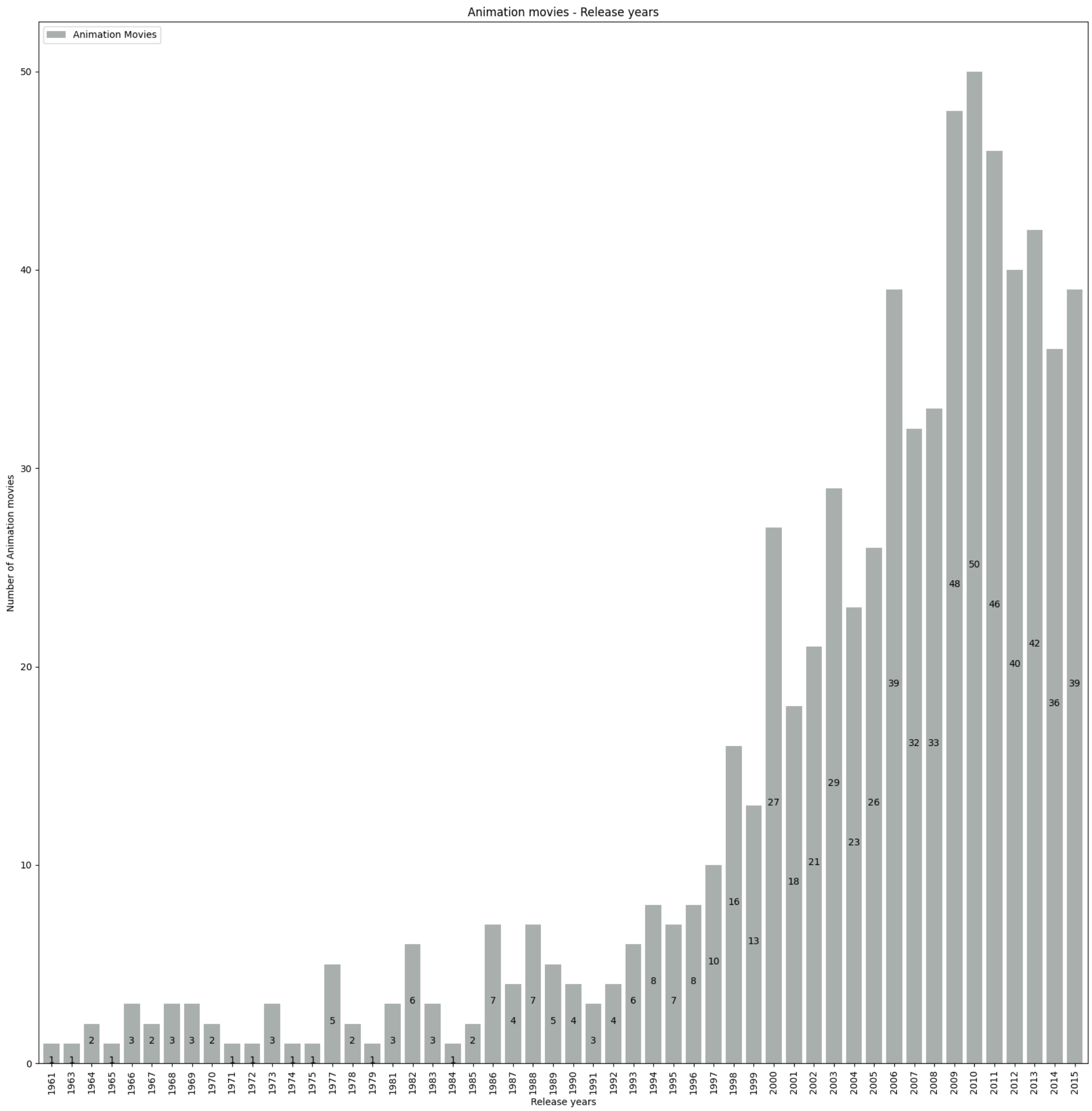

Fantasy movies - Release years
Fantasy Movies
Number of Fantasy movies
Release years

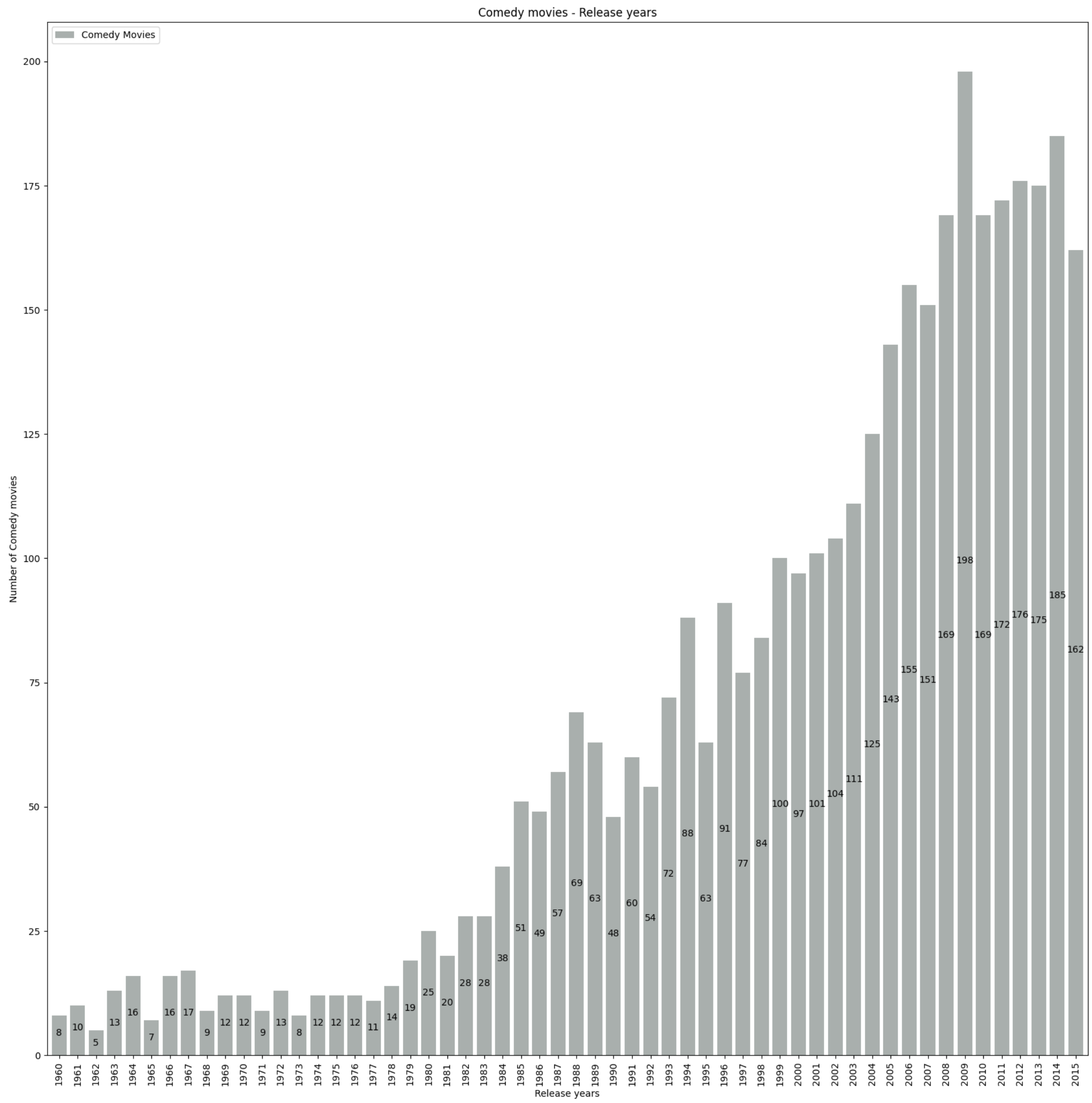

Crime movies - Release years
Crime Movies
Number of Crime movies
Release years
0
10
20
30
40
50
60
70
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2 2 3 4 10 2 5 7 5 3 4 10 7 17 15 6 8 2 6 8 9 10 9 11 15 14 18 21 21 25 30 20 25 28 26 31 19 28 34 29 31 39 41 46 32 45 53 65 62 52 50 48 54 71 65 51


Western movies - Release years
Western Movies
Number of Western movies
Release years

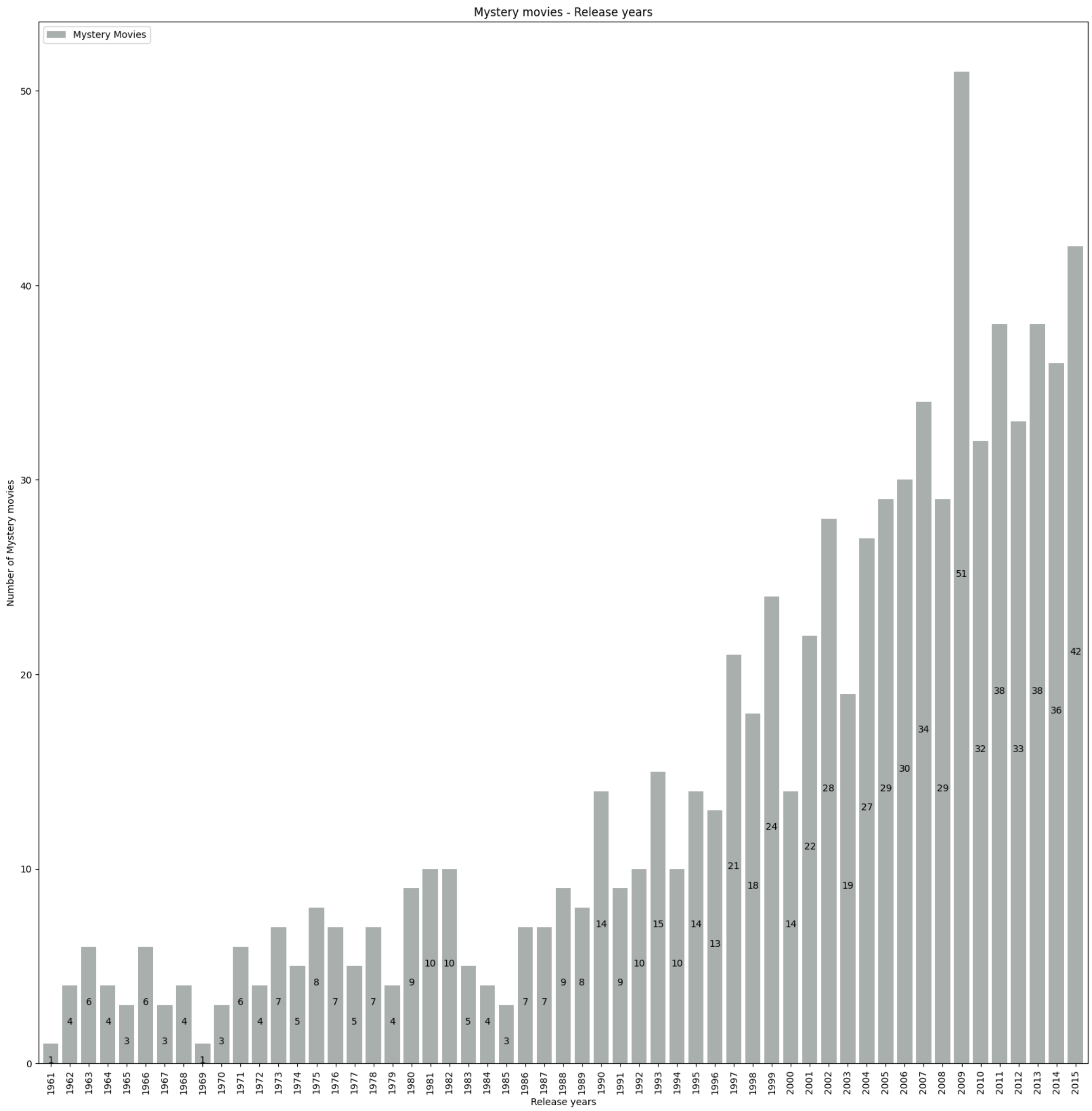

Drama movies - Release years
Drama Movies
Number of Drama movies
Release years

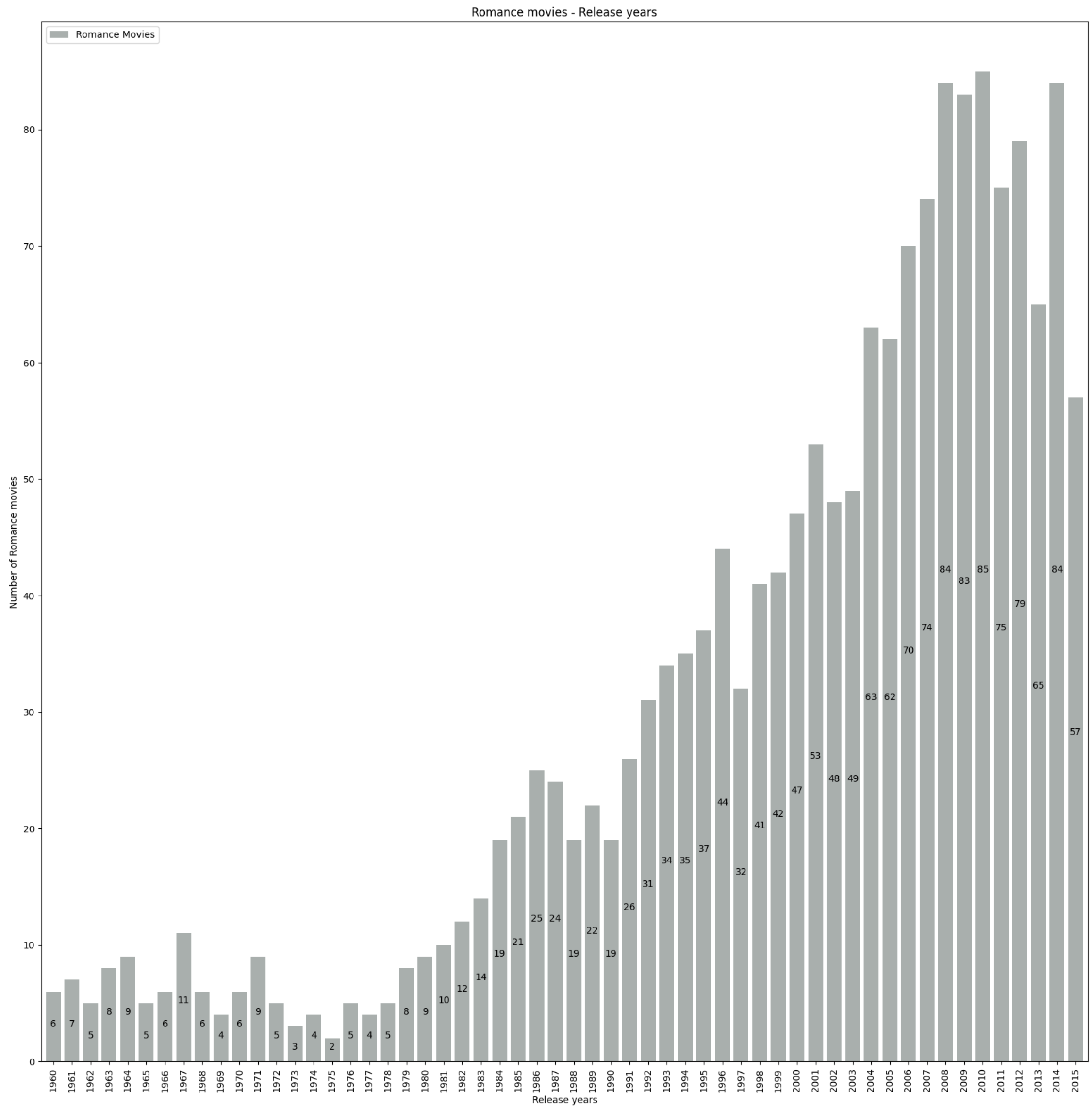

Family movies - Release years
Family Movies
Number of Family movies
Release years


Animation movies - Release years
Animation Movies
Number of Animation movies
Release years
0
10
20
30
40
50
1961
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1977
1978
1979
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
1
1
2
1
3
2
3
3
2
1
1
3
1
1
5
2
1
3
6
3
1
2
7
4
7
5
4
3
4
6
8
7
8
10
16
13
27
18
21
29
23
26
39
32
33
48
50
46
40
42
36
39

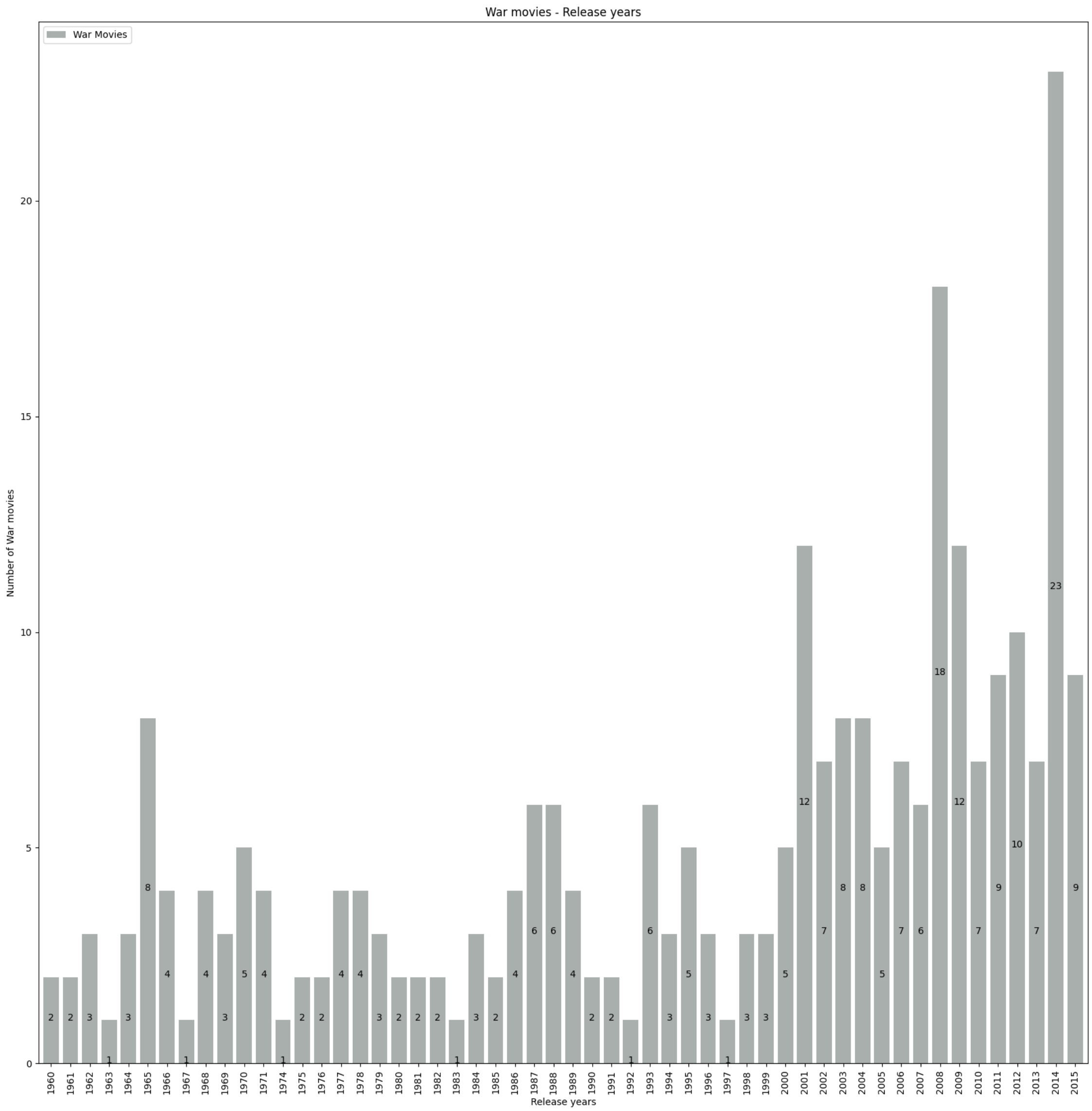


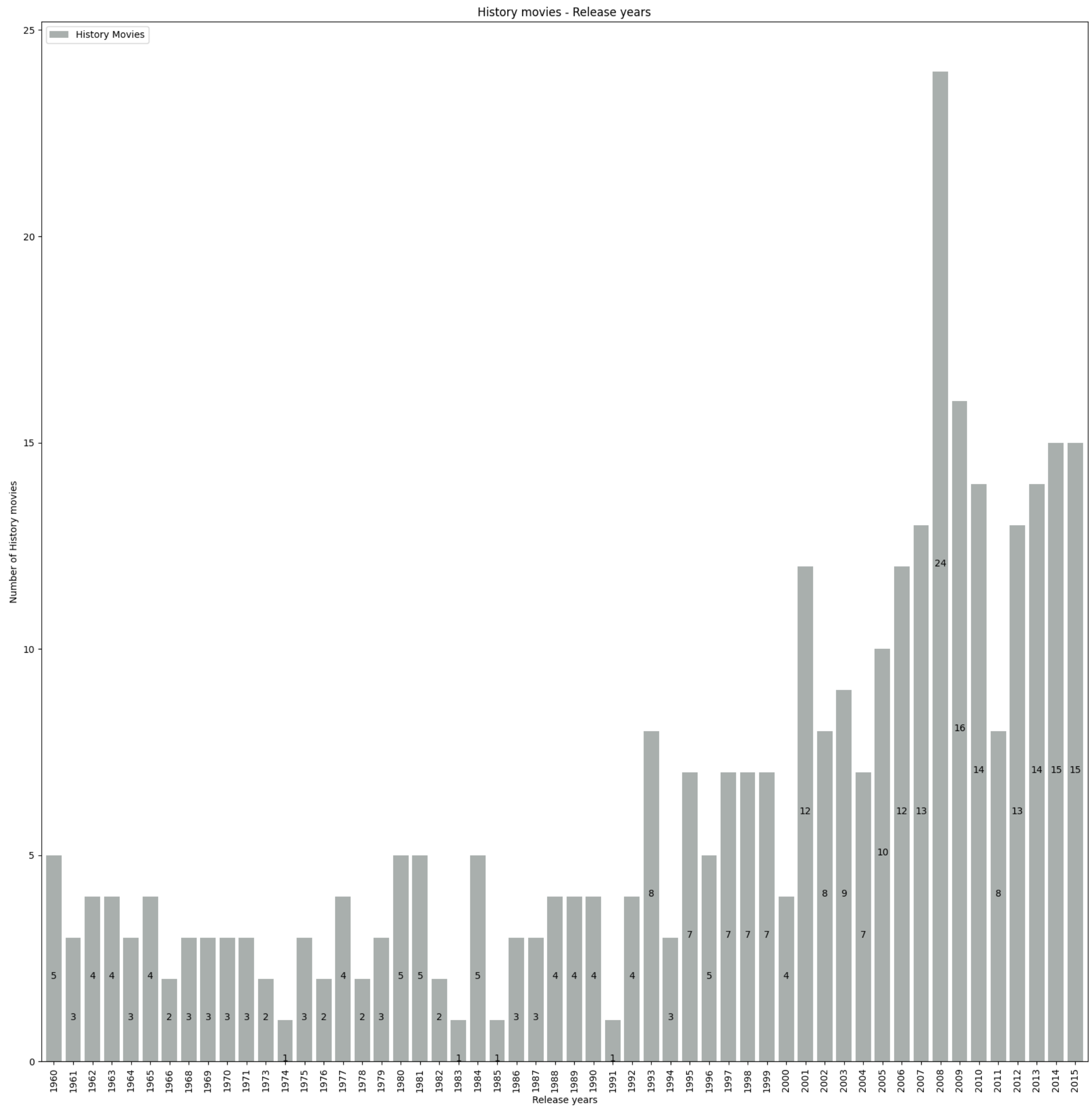
Comedy movies - Release years
Comedy Movies
Number of Comedy movies
0
25
50
75
100
125
150
175
200
8
10
5
13
16
7
16
17
9
12
12
9
13
8
12
12
12
11
14
19
25
20
28
28
38
51
49
57
69
63
48
60
54
72
88
63
91
77
84
100
97
101
104
111
125
143
155
151
169
198
169
172
176
175
185
162
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
Release years

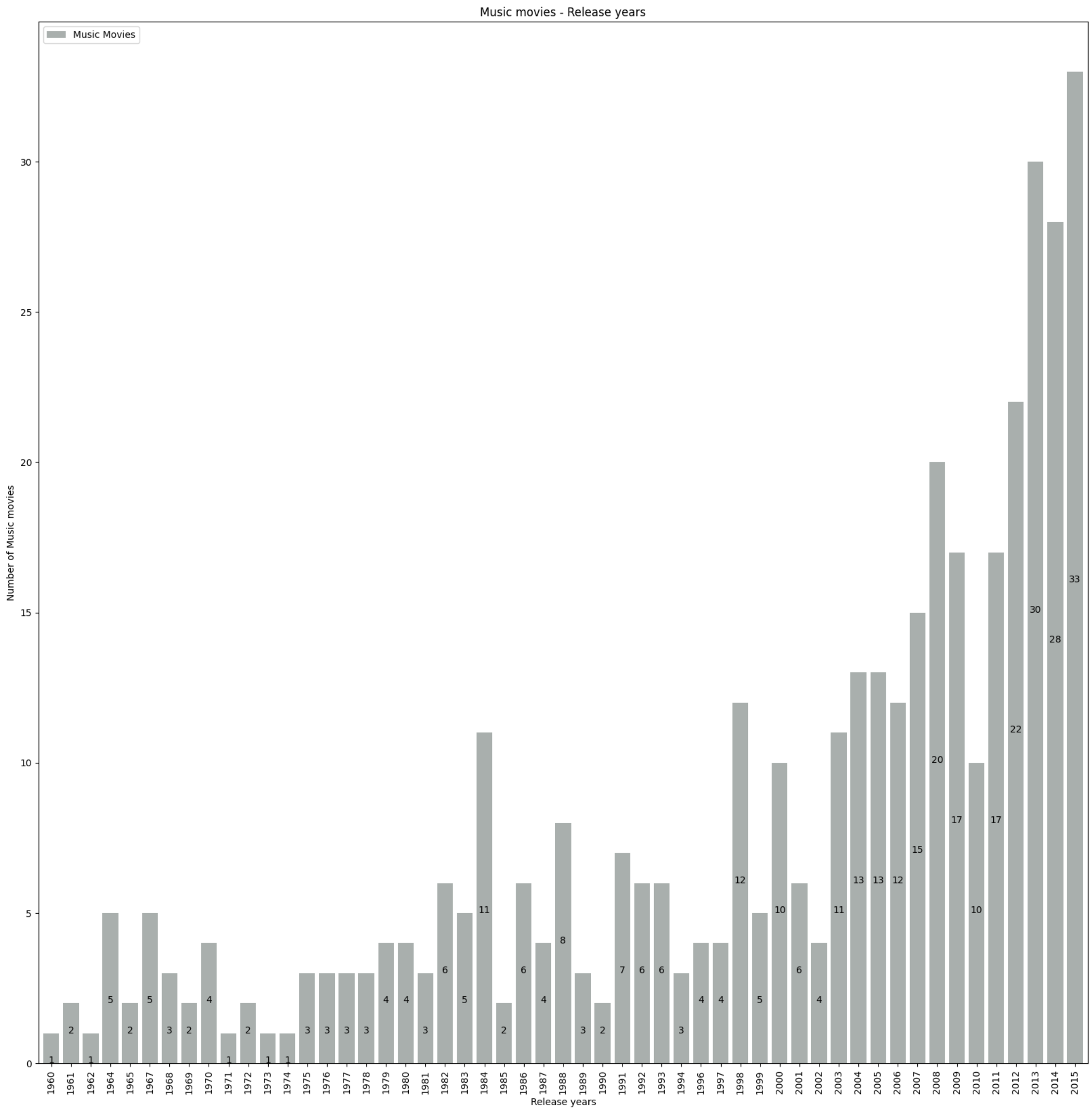

Mystery movies - Release years
Mystery Movies
Number of Mystery movies
Release years
0
10
20
30
40
50
1961 1962 1963 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
1 4 6 4 3 6 3 4 1 3 6 4 7 5 8 7 5 7 4 9 10 10 5 4 3 7 7 9 8 14 9 10 15 10 14 13 21 18 24 14 22 28 19 27 29 30 34 29 51 32 38 33 38 36 42

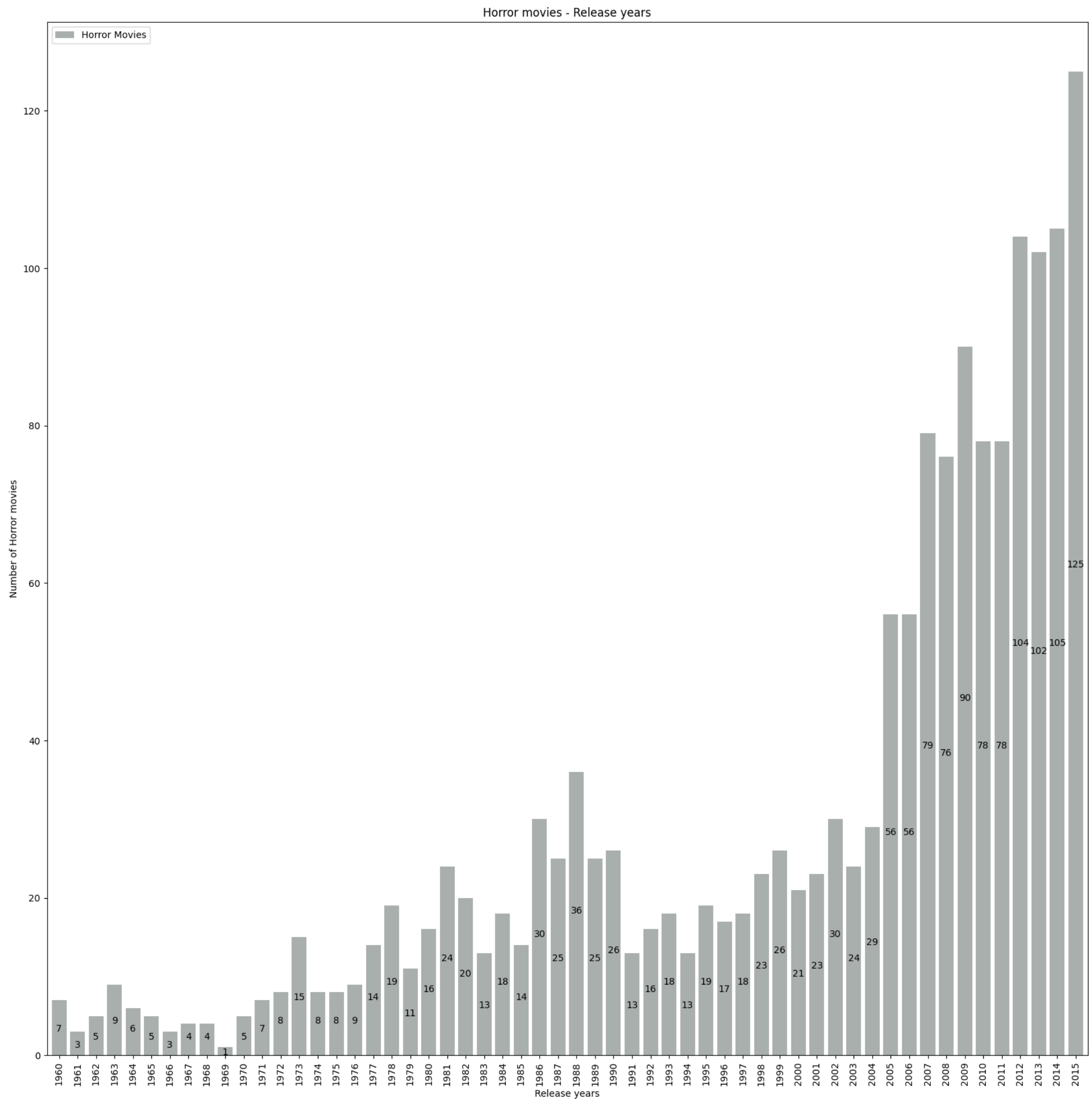


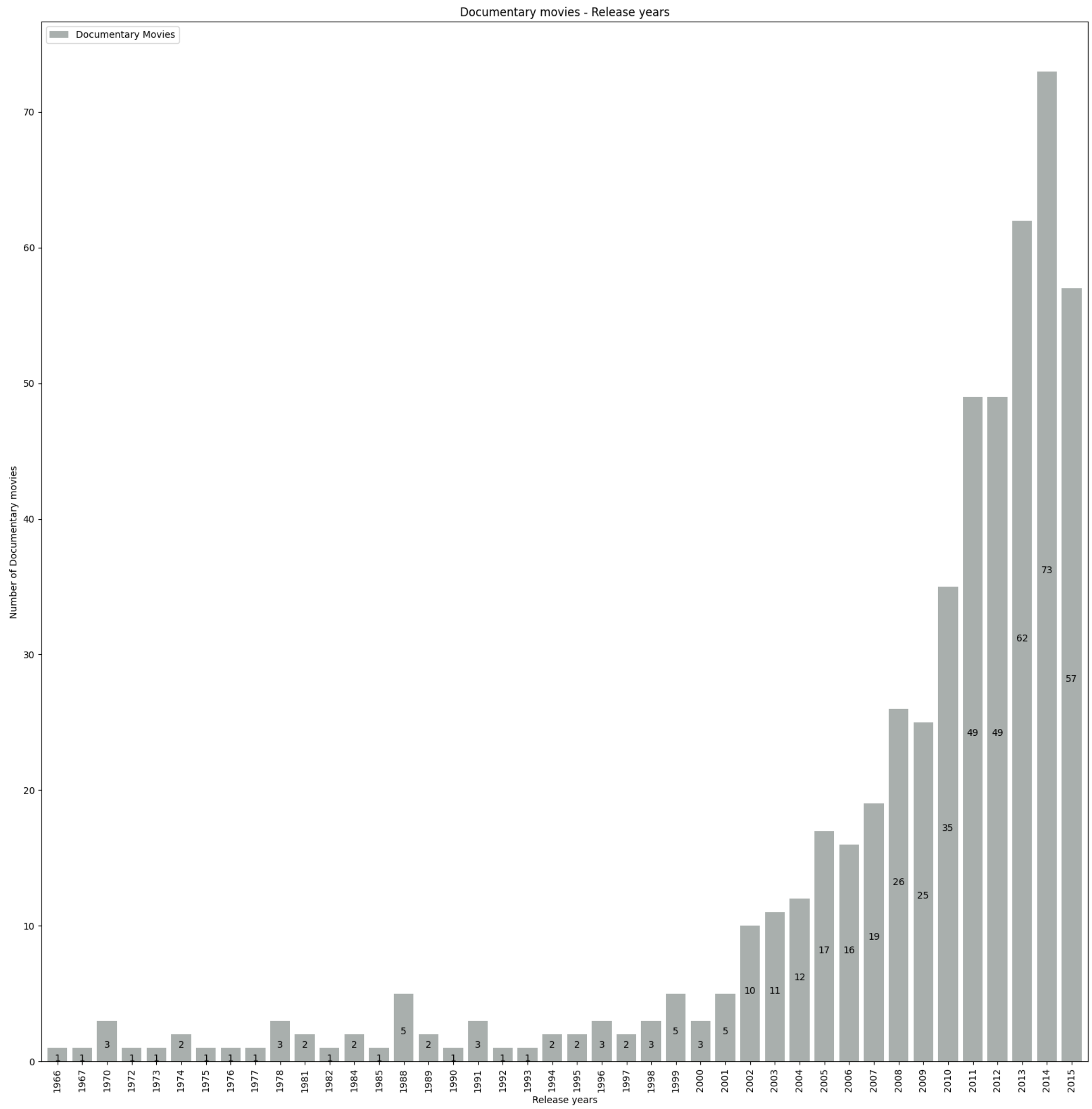
Romance movies - Release years
Romance Movies
Number of Romance movies
Release years

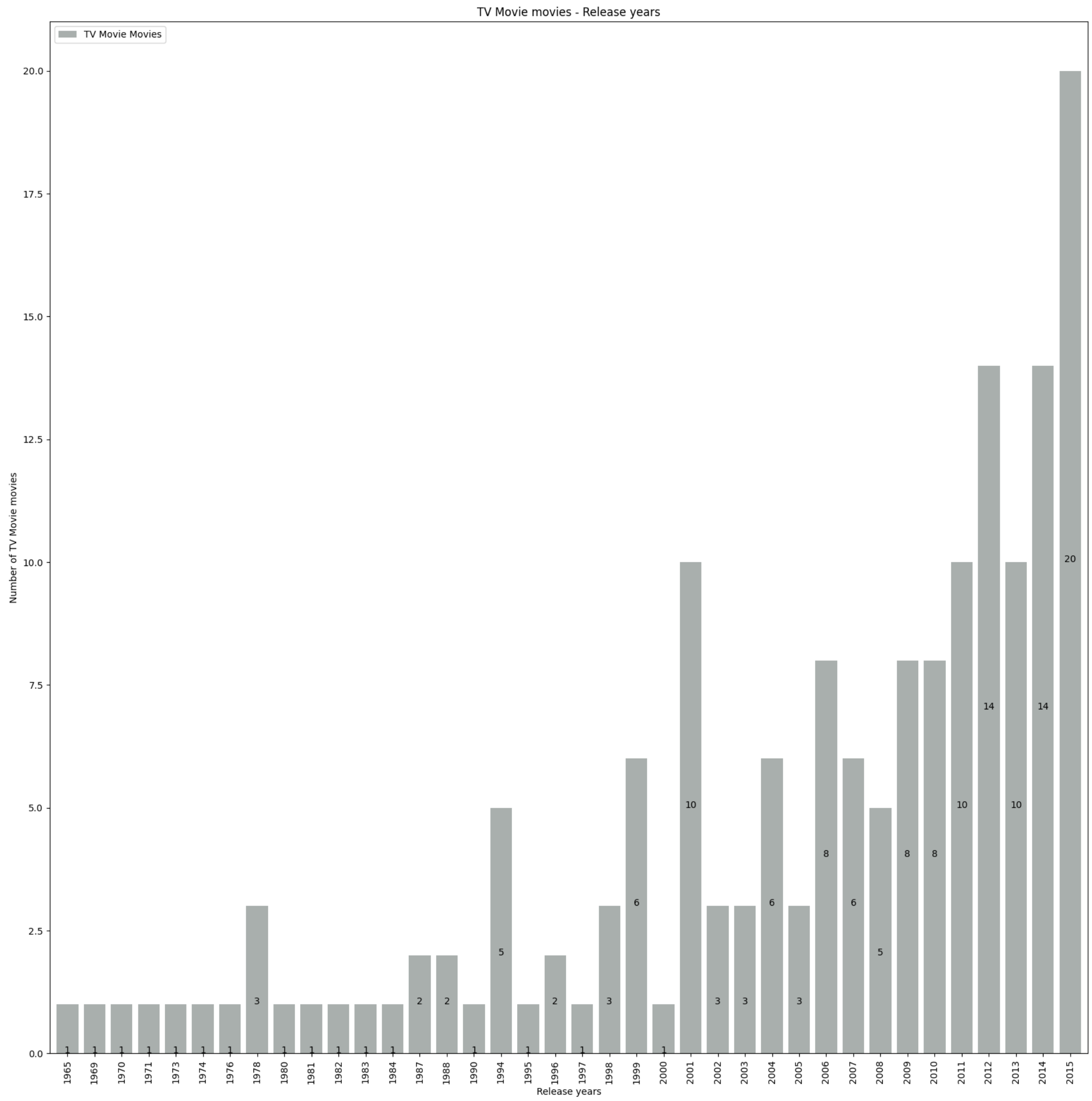

War movies - Release years
War Movies
Number of War movies
Release years
0
5
10
15
20
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015


History movies - Release years
History Movies
Number of History movies
Release years

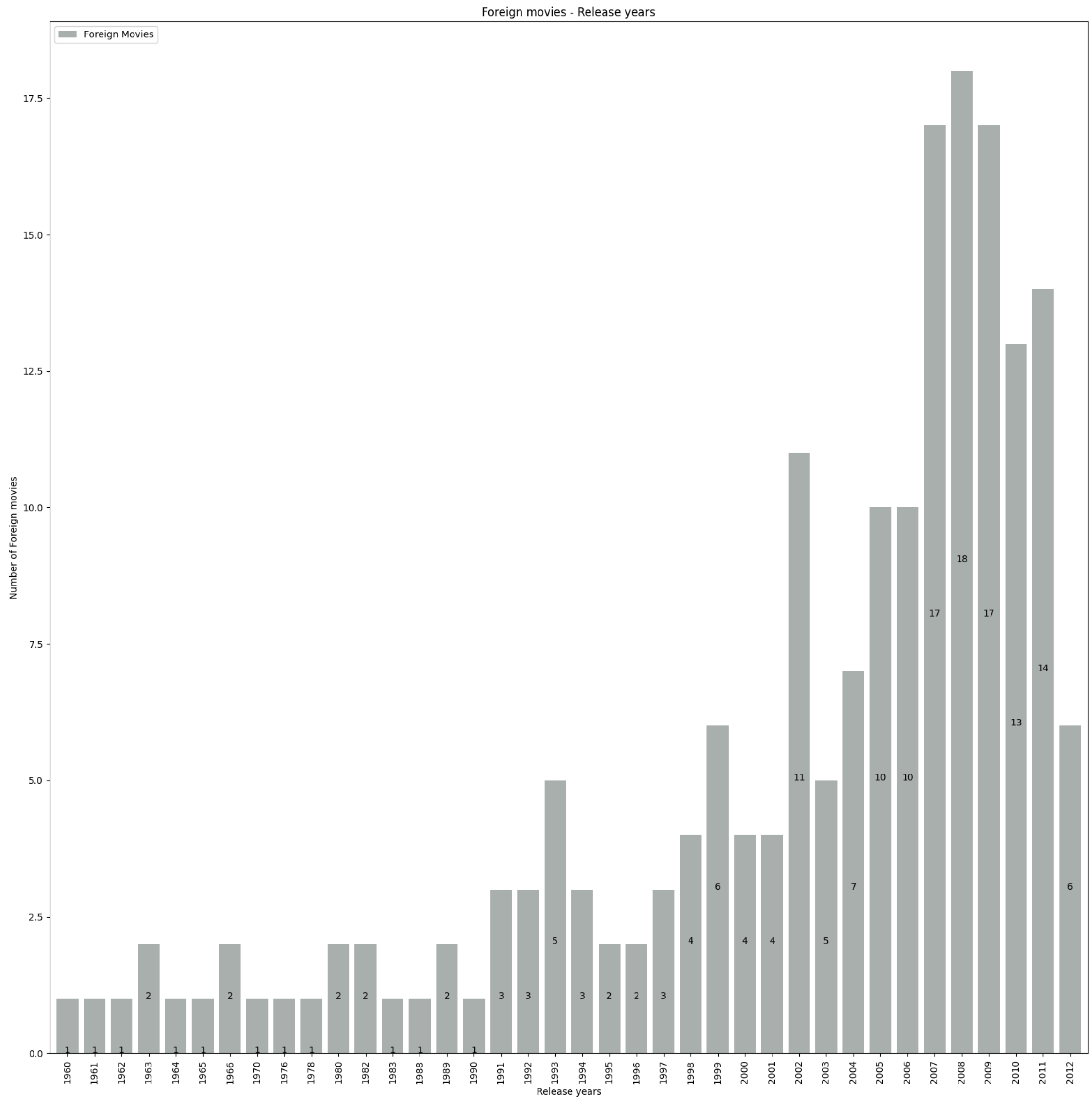

Music movies - Release years
Music Movies
Number of Music movies
Release years


Horror movies - Release years
Horror Movies
Number of Horror movies
Release years


Documentary movies - Release years
Documentary Movies
Number of Documentary movies
Release years



TV Movie movies - Release years
TV Movie Movies
Number of TV Movie movies
Release years



Answer: from the plots we have plotted, we can see that for:

[20]:
```
# here, we print the most popular genre for each year, along with its number.

for i in list_years:
    list_compare = []
    dict_compare = {}
    for j in list_1a:
        df_part('release_year', j, list_years)
        if i in list(df_part('release_year', j, list_years)[j].index.
 ↪get_level_values(0)):
            list_compare.append(df_part('release_year', j, list_years)[j][i])
            dict_compare[j] = df_part('release_year', j, list_years)[j][i]
```

```python
    # finally, print the following formatted string.
    print(f'* For {i}, the most common genre is {str(max(dict_compare,␣
↪key=dict_compare.get))} of value {str(max(list_compare))}.')
```

```
* For 1960, the most common genre is Drama of value 13.
* For 1961, the most common genre is Drama of value 16.
* For 1962, the most common genre is Drama of value 21.
* For 1963, the most common genre is Drama of value 13.
* For 1964, the most common genre is Drama of value 20.
* For 1965, the most common genre is Drama of value 20.
* For 1966, the most common genre is Drama of value 16.
* For 1967, the most common genre is Comedy of value 17.
* For 1968, the most common genre is Drama of value 20.
* For 1969, the most common genre is Drama of value 13.
* For 1970, the most common genre is Drama of value 19.
* For 1971, the most common genre is Drama of value 30.
* For 1972, the most common genre is Drama of value 16.
* For 1973, the most common genre is Drama of value 31.
* For 1974, the most common genre is Drama of value 21.
* For 1975, the most common genre is Drama of value 17.
* For 1976, the most common genre is Drama of value 22.
* For 1977, the most common genre is Drama of value 24.
* For 1978, the most common genre is Drama of value 29.
* For 1979, the most common genre is Drama of value 30.
* For 1980, the most common genre is Drama of value 32.
* For 1981, the most common genre is Drama of value 32.
* For 1982, the most common genre is Drama of value 33.
* For 1983, the most common genre is Drama of value 35.
* For 1984, the most common genre is Drama of value 40.
* For 1985, the most common genre is Comedy of value 51.
* For 1986, the most common genre is Drama of value 51.
* For 1987, the most common genre is Comedy of value 57.
* For 1988, the most common genre is Comedy of value 69.
* For 1989, the most common genre is Comedy of value 63.
* For 1990, the most common genre is Drama of value 60.
* For 1991, the most common genre is Drama of value 63.
* For 1992, the most common genre is Drama of value 65.
* For 1993, the most common genre is Drama of value 90.
* For 1994, the most common genre is Comedy of value 88.
* For 1995, the most common genre is Drama of value 93.
* For 1996, the most common genre is Drama of value 104.
* For 1997, the most common genre is Drama of value 83.
* For 1998, the most common genre is Drama of value 108.
* For 1999, the most common genre is Drama of value 113.
* For 2000, the most common genre is Drama of value 101.
* For 2001, the most common genre is Comedy of value 101.
* For 2002, the most common genre is Drama of value 130.
* For 2003, the most common genre is Comedy of value 111.
```

```
* For 2004, the most common genre is Drama of value 141.
* For 2005, the most common genre is Drama of value 182.
* For 2006, the most common genre is Drama of value 197.
* For 2007, the most common genre is Drama of value 197.
* For 2008, the most common genre is Drama of value 233.
* For 2009, the most common genre is Drama of value 224.
* For 2010, the most common genre is Drama of value 210.
* For 2011, the most common genre is Drama of value 214.
* For 2012, the most common genre is Drama of value 232.
* For 2013, the most common genre is Drama of value 253.
* For 2014, the most common genre is Drama of value 284.
* For 2015, the most common genre is Drama of value 260.
```

1.2.1 The code above prints the following:

- For 1960, the most common genre is Drama of value 13.
- For 1961, the most common genre is Drama of value 16.
- For 1962, the most common genre is Drama of value 21.
- For 1963, the most common genre is Drama of value 13.
- For 1964, the most common genre is Drama of value 20.
- For 1965, the most common genre is Drama of value 20.
- For 1966, the most common genre is Drama of value 16.
- For 1967, the most common genre is Comedy of value 17.
- For 1968, the most common genre is Drama of value 20.
- For 1969, the most common genre is Drama of value 13.
- For 1970, the most common genre is Drama of value 19.
- For 1971, the most common genre is Drama of value 30.
- For 1972, the most common genre is Drama of value 16.
- For 1973, the most common genre is Drama of value 31.
- For 1974, the most common genre is Drama of value 21.
- For 1975, the most common genre is Drama of value 17.
- For 1976, the most common genre is Drama of value 22.
- For 1977, the most common genre is Drama of value 24.
- For 1978, the most common genre is Drama of value 29.
- For 1979, the most common genre is Drama of value 30.
- For 1980, the most common genre is Drama of value 32.
- For 1981, the most common genre is Drama of value 32.
- For 1982, the most common genre is Drama of value 33.
- For 1983, the most common genre is Drama of value 35.
- For 1984, the most common genre is Drama of value 40.
- For 1985, the most common genre is Comedy of value 51.
- For 1986, the most common genre is Drama of value 51.
- For 1987, the most common genre is Comedy of value 57.
- For 1988, the most common genre is Comedy of value 69.
- For 1989, the most common genre is Comedy of value 63.
- For 1990, the most common genre is Drama of value 60.
- For 1991, the most common genre is Drama of value 63.
- For 1992, the most common genre is Drama of value 65.

- For 1993, the most common genre is Drama of value 90.
- For 1994, the most common genre is Comedy of value 88.
- For 1995, the most common genre is Drama of value 93.
- For 1996, the most common genre is Drama of value 104.
- For 1997, the most common genre is Drama of value 83.
- For 1998, the most common genre is Drama of value 108.
- For 1999, the most common genre is Drama of value 113.
- For 2000, the most common genre is Drama of value 101.
- For 2001, the most common genre is Comedy of value 101.
- For 2002, the most common genre is Drama of value 130.
- For 2003, the most common genre is Comedy of value 111.
- For 2004, the most common genre is Drama of value 141.
- For 2005, the most common genre is Drama of value 182.
- For 2006, the most common genre is Drama of value 197.
- For 2007, the most common genre is Drama of value 197.
- For 2008, the most common genre is Drama of value 233.
- For 2009, the most common genre is Drama of value 224.
- For 2010, the most common genre is Drama of value 210.
- For 2011, the most common genre is Drama of value 214.
- For 2012, the most common genre is Drama of value 232.
- For 2013, the most common genre is Drama of value 253.
- For 2014, the most common genre is Drama of value 284.
- For 2015, the most common genre is Drama of value 260.

We can conclude that Drama is the most common movie genre overall, and has dominated early TV.

Comedy was also a very good competitor from 1985 to around 2003.

We can also conclude that with the rising number of Drama movies over the time interval, we can say that TV has gotten extremely popular and might've reached its peak at 2014 only dropping a bit during 2015.

1.3 Second question:

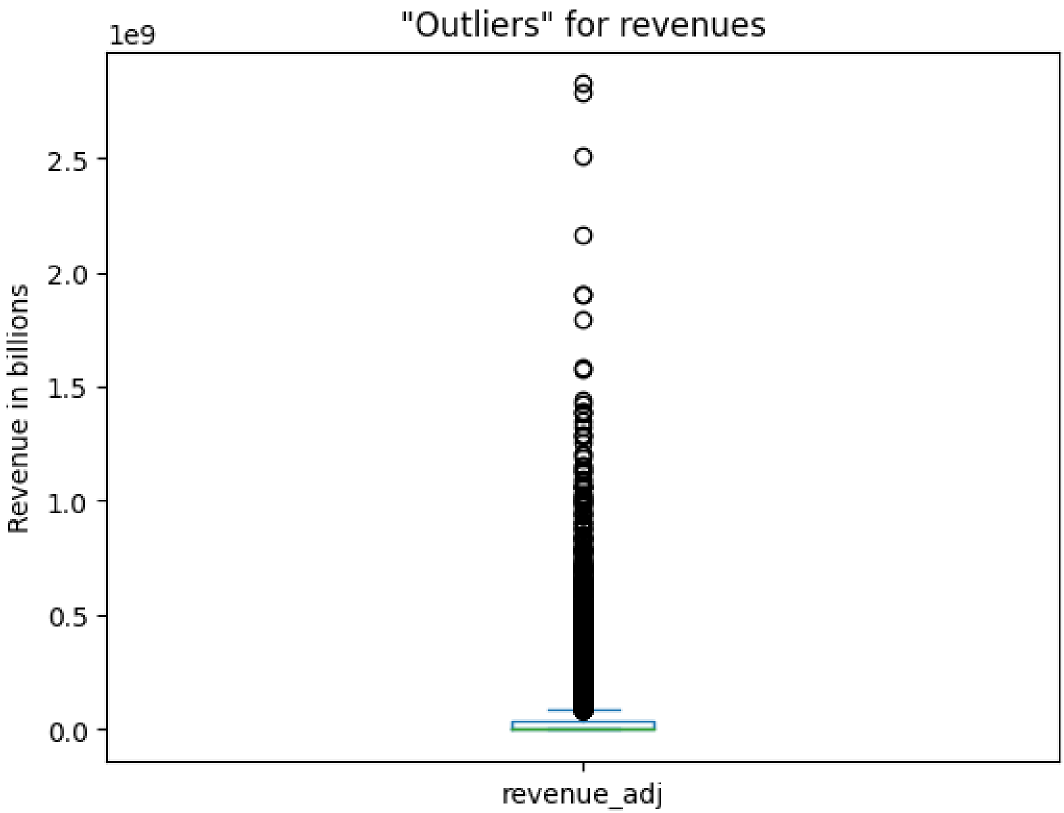
Does runtime affect the movie's revenues?

For this one, we need to isolate our two columns 'runtime' and 'revenue_adj'. Then get the mean for each group of runtime, which will give us a rough idea for our question.

(runtime is measured in minutes and revenue_adj accounts for inflation.)

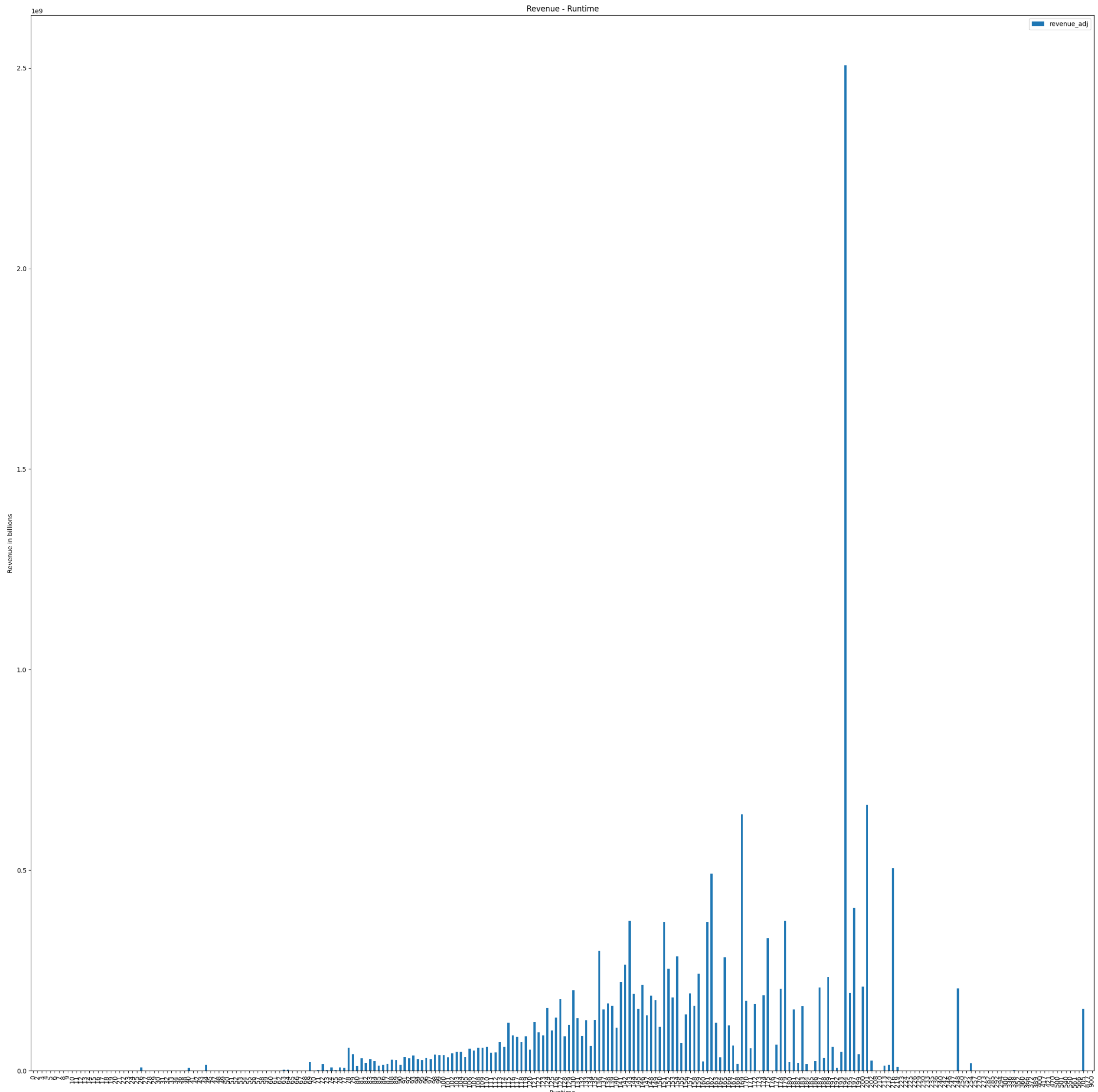
However, let's check for outliers.

[21]:
```
df['revenue_adj'].plot(kind = 'box', ylabel = 'Revenue in billions', title =␣
↪'"Outliers" for revenues');
```



Apparently most movies flop which gives us a ton of outliers.

[22]:
```
# the y-values are measured in billions.
pd.DataFrame(df.groupby('runtime')['revenue_adj'].mean()).plot(kind = 'bar',␣
↪figsize = (30, 30), xlabel = 'Runtime', ylabel = 'Revenue in billions',␣
↪title = 'Revenue - Runtime');
```



[23]:
```
#identifying what value that large spike (apparent outlier) is ...
df.groupby('runtime')['revenue_adj'].mean().max()
```

[23]: 2506405735.41947

Now, this is a very rough plot, many values are plotted on the x-axis and so you don't get a good look at the x-values. However, the plot is very clear, and we'll be making use of this in our answer. First, we see that for values of runtime under a certain value, these movies got almost no revenue, and for above that certain value, we see a very positive correlation between runtime and revenue.

This trend doesn't continue on for long though. Eventually, the revenue approaches zero steadily once again except for a few outliers which I think are probably related

to documentaries and large movie productions that are very hyped up and has a good marketing team.

Now, let's call the values that are at the bounds of the region where there is positive correlation t_a (lower bound) and t_b (upper bound). We see that for values:

- $t < t_a$ we use our observations to quantify this and we get that the revenue $r \to 0$ (approaches zero).
- $t_a < t < t_b$, the revenue r has a positive correlation and is expected to rise (almost) regularly until t_b.
- $t > t_b$, the revenue r has a negative correlation (which is true for most of the movies in that interval, except a few outliers) and swiftly approaches zero ($r \to 0$).

We can tell what t_a and t_b are by inspecting the plot, which is definitely hard, but, by inspection:

- t_a is around the 80 mark.
- t_b is around the 152 mark.

1.3.1 So, for optimal revenue, assuming the movie's theme, marketing strategies and whatnot are optimised, your movie's runtime should fall in the $(80, 152)$ minute interval.

Conclusions

1.3.2 Finally, at the end of our two questions, we can conclude that for a movie to be successful it might have two criteria to fill:

- It might have to be a Drama film/movie. Reasoning behind this is most movies and films are of genre 'Drama', we can conclude that most production companies (including very very big ones) produce mostly Drama films.
- And it has to fall in the 80 to 152 minute interval for its runtime. Reasoning behind this is we notice a slight positive correlation between runtime and revenue in our graph/plot, which leads us to think that on average, people are more likely to watch a movie or film if its runtime doesn't exceed 2.53 hours (152 minutes) and doesn't fall beneath 1.3 hours (80 minutes) and ideally is inside that region, not teetering on the edge of it.

1.3.3 That doesn't always mean success for our movie or film, but it does have a slight effect.

1.3.4 We also have to recognise the limitations of this analysis. If we wanted to analyse production companies, you might want to seperate them (they're in a pipe-separated format), for which case you would get many many many more values than when we did that to genres, which would take much more time to analyse. We also state that point for analysing directors.

1.3.5 With revenue vs runtime, the plot we produced is very very full of values, which makes it hard to analyse it.