TMDb dataset investigation

Table of Contents

Introduction

overview

> This data set contains information about 10,000 movies collected from The Movie Database (TMDb), including user ratings and revenue.
> • Certain columns, like ‘cast’ and ‘genres’, contain multiple values separated by pipe (|) characters.
> • There are some odd characters in the ‘cast’ column. Don’t worry about cleaning them. You can leave them as is.
> • The final two columns ending with “_adj” show the budget and revenue of the associated movie in terms of 2010 dollars, accounting for inflation over time.
>
> ### *strategy of analysis
>
> Giving overview of the key points about all movies by proposing the frist 9 questions.
> Next choosing four categories of revenue to get some insights to see if there is a relationship between the profit and the increasing in the budget of the movie.
>
> **Analytic questions for the first part of the strategy:-**
>
> 1.which movie has the biggest and lowest profit?
> 2.which movie has the biggest and lowest budget?
> 3.which movies has the biggest and lowest revenue?
> 4.which movies has the longest and shortest runtime?
> 5.what is the average runtime of all movies?.
> 6.what are the most successful genres of movies?
> 7.the most repeated cast?
> 8.what is the average budget ?

```
9. what is the  average Revenue ?<br>
```

For the secound part of the starategy>>> based on the comparsion between four revenue categories of movies more than (25,50,100,150)M according to the following questions it can let us to some conclusions

> 1.What is the average budget of the movie?
> 2.What is the average revenue of the movie?
> 3.What is the average runtime of the movie?
> 4.Which are the successfull genres?
> 5.Which are the most frequent cast involved?
>
> **the insides we get from answers of those questions can lead us to a conclusion about the next questions**
>
> 1.what are the best genres of movies constantly?

```
2. what the best cast for the different categories ?<br>
3.is runtime varies according to the cat.?<br>
4. is there a relation between the budget and revenue (BR) and the categories of movies ?
<br>
5.is there a relation between avg. of the profit (SR) and the categories ?
```

In [168]:

```
#loading necessary libraries

import pandas as pd
import numpy as np
import operator
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

Data collecting

In [169]:

```
tmdb_data = pd.read_csv('tmdb-movies2.csv')
#printing first five rows
tmdb_data.head()
```

Out[169]:

	id	imdb_id	popularity	budget	revenue	original_title	cast	homepage	directo
0	135397	tt0369610	32.985763	150000000	1513528810	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	http://www.jurassicworld.com/	Coli Trevorrov
1	76341	tt1392190	28.419936	150000000	378436354	Mad Max: Fury Road	Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...	http://www.madmaxmovie.com/	Georg Mille
2	262500	tt2908446	13.112507	110000000	295238201	Insurgent	Shailene Woodley|Theo James|Kate Winslet|Ansel...	http://www.thedivergentseries.movie/#insurgent	Rober Schwentk
3	140607	tt2488496	11.173104	200000000	2068178225	Star Wars: The Force Awakens	Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...	http://www.starwars.com/films/star-wars-episod...	J.. Abram
4	168259	tt2820852	9.335014	190000000	1506249360	Furious 7	Vin Diesel|Paul Walker|Jason Statham|Michelle ...	http://www.furious7.com/	Jame Wa

5 rows × 22 columns

Data set observation

1.)unifiy the currency to dollar.
2.) we cannot conculed the popularity of the movies based on the average vote account becouse vote_count is different for all the movies

Data Cleaning

1. Removing Unused columns

Columns that we need to delete are

- id, imdb_id, popularity, budget_adj, revenue_adj, homepage, keywords, overview, production_companies, vote_count and vote_average.

In [170]:

```
# list of columb to be deleted
del_col=[ 'id', 'imdb_id', 'popularity', 'budget_adj', 'revenue_adj', 'homepage', 'keywords', 'overview',

#deleting the columns
tmdb_data= tmdb_data.drop(del_col,1)

#previewing the new dataset
tmdb_data.head(4)
```

Out[170]:

	budget	revenue	original_title	cast	director	tagline	runtime	genres	release_date	release_year	
0	150000000	1513528810	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	Colin Trevorrow	The park is open.	124	Action|Adventure|Science Fiction|Thriller	06/09/2015	2015	13
1	150000000	378436354	Mad Max: Fury Road	Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...	George Miller	What a Lovely Day.	120	Action|Adventure|Science Fiction|Thriller	5/13/15	2015	2
2	110000000	295238201	Insurgent	Shailene Woodley|Theo James|Kate Winslet|Ansel...	Robert Schwentke	One Choice Can Destroy You	119	Adventure|Science Fiction|Thriller	3/18/15	2015	1
3	200000000	2068178225	Star Wars: The Force Awakens	Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...	J.J. Abrams	Every generation has a story.	136	Action|Adventure|Science Fiction|Fantasy	12/15/15	2015	18

2. remove the duplicated rows if excist

figure out how many entries we have in the database

In [171]:

```
rows, col = tmdb_data.shape
#We need to reduce the count of row by one as contain header row also.
print('There are {} total entries of movies and {} no.of columns in it.'.format(rows-1, col))
```

```
There are 10865 total entries of movies and 11 no.of columns in it.
```

removing the duplicated rows if any excist

In [261]:

```
tmdb_data.drop_duplicates(keep ='first', inplace=True)
rows, col = tmdb_data.shape

print('There are {}  entries of movies and {} number of columns.'.format(rows-1, col))
```

```
There are 3853  entries of movies and 12 number of columns.
```

So there was a duplicate row and it has been removed now.

3. deleting the zero values from budget and the revenue columns

In [262]:

```
# making a seperate list of revenue and budget columns
temp_list=['budget', 'revenue']

# relpacing all zeros with NAN
tmdb_data[temp_list] = tmdb_data[temp_list].replace(0, np.NAN)

#Removing all the row which has NaN value in temp_list
tmdb_data.dropna(subset = temp_list, inplace = True)

rows, col = tmdb_data.shape
print(' we now have only {} no.of movies.'.format(rows-1))
```

```
 we now have only 3853 no.of movies.
```

4. separating the release date to reformate it

In [263]:

```
tmdb_data.release_date = pd.to_datetime(tmdb_data['release_date'])
```

In [264]:

```
# printing the new dataset
tmdb_data.head(5)
```

Out[264]:

	budget	revenue	profit_earned	original_title	cast	director	tagline	runtime	genres	release_date
0	150000000	1513528810	1363528810	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	Colin Trevorrow	The park is open.	124	Action|Adventure|Science Fiction|Thriller	2015-06-09
1	150000000	378436354	228436354	Mad Max: Fury Road	Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...	George Miller	What a Lovely Day.	120	Action|Adventure|Science Fiction|Thriller	2015-05-13
2	110000000	295238201	185238201	Insurgent	Shailene Woodley|Theo James|Kate Winslet|Ansel...	Robert Schwentke	One Choice Can Destroy You	119	Adventure|Science Fiction|Thriller	2015-03-18
3	200000000	2068178225	1868178225	Star Wars: The Force Awakens	Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...	J.J. Abrams	Every generation has a story.	136	Action|Adventure|Science Fiction|Fantasy	2015-12-15
4	190000000	1506249360	1316249360	Furious 7	Vin Diesel|Paul Walker|Jason Statham|Michelle ...	James Wan	Vengeance Hits Home	137	Action|Crime|Thriller	2015-04-01

5. replacing zeros with NAN in runtime column

In [176]:

```
#replacing zeros with NaN of runtime column
tmdb_data['runtime'] =tmdb_data['runtime'].replace(0, np.NAN)
```

6. reformattin of budget and revenue column.

Let's check the current format of columns in the dataset

In [265]:

```
#printing the data type
tmdb_data.dtypes
```

Out[265]:

```
budget                     int64
revenue                    int64
profit_earned              int64
original_title            object
cast                      object
director                  object
tagline                   object
runtime                    int64
genres                    object
release_date      datetime64[ns]
release_year               int64
profit                     int64
dtype: object
```

In [266]:

```
change_type=['budget', 'revenue']
#changing data type
tmdb_data[change_type]=tmdb_data[change_type].applymap(np.int64)
#printing the new information
tmdb_data.dtypes
```

Out[266]:

```
budget                     int64
revenue                    int64
profit_earned              int64
original_title            object
cast                      object
director                  object
tagline                   object
runtime                    int64
genres                    object
release_date      datetime64[ns]
release_year               int64
profit                     int64
dtype: object
```

Data exploration

1. Calculating the profit of the each movie

In [269]:

```
#insert function with three parameters(index of the column in the dataset, name of the column, value to
tmdb_data.insert(2,'profit_earned0',tmdb_data['revenue']-tmdb_data['budget'])

#previewing the new value of profit_earned in the dataset
tmdb_data.head(3)
```

Out[269]:

	budget	revenue	profit_earned0	profit_earned2	profit_earned	original_title	cast	director	tagline	runtime	
0	150000000	1513528810	1363528810	1363528810	1363528810	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	Colin Trevorrow	The park is open.	124	Action|Adve F
1	150000000	378436354	228436354	228436354	228436354	Mad Max: Fury Road	Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...	George Miller	What a Lovely Day.	120	Action|Adve F
2	110000000	295238201	185238201	185238201	185238201	Insurgent	Shailene Woodley|Theo James|Kate Winslet|Ansel...	Robert Schwentke	One Choice Can Destroy You	119	Adve F

Answers to the proposed questions for the first step in the strategy >>

1.which movie has the biggest and lowest profit?

In [271]:

```
import pprint
def calculate(column):
    #the highest profit
    high= tmdb_data[column].idxmax()
    high_details=pd.DataFrame(tmdb_data.loc[high])

    #the lowest profit
    low= tmdb_data[column].idxmin()
    low_details=pd.DataFrame(tmdb_data.loc[low])

    #getting data in one place
    info=pd.concat([high_details, low_details], axis=1)

    return info

calculate('profit_earned0')
```

Out[271]:

	1386	2244
budget	237000000	425000000
revenue	2781505847	11087569
profit_earned0	2544505847	-413912431
profit_earned2	2544505847	-413912431
profit_earned	2544505847	-413912431
original_title	Avatar	The Warrior's Way
cast	Sam Worthington|Zoe Saldana|Sigourney Weaver|S...	Kate Bosworth|Jang Dong-gun|Geoffrey Rush|Dann...
director	James Cameron	Sngmoo Lee
tagline	Enter the World of Pandora.	Assassin. Hero. Legend.
runtime	162	100
genres	Action|Adventure|Fantasy|Science Fiction	Adventure|Fantasy|Action|Western|Thriller
release_date	2009-12-10 00:00:00	2010-12-02 00:00:00
release_year	2009	2010
profit	2544505847	-413912431

Avatar movie has the hieghest profit value = 2544505847 .
Whereas The warrior's way has the lowest profit value = -413912431 it seems that it lost alot of money

2.which movie has the biggest and lowest budget?

In [181]:

```
# we will call the same function **calculate(column)** again to calculate the highest and lowest budget
calculate('budget')
```

Out[181]:

	2244	2618
budget	425000000	1
revenue	11087569	100
profit_earned	-413912431	99
original_title	The Warrior's Way	Lost & Found
cast	Kate Bosworth|Jang Dong-gun|Geoffrey Rush|Dann...	David Spade|Sophie Marceau|Ever Carradine|Step...
director	Sngmoo Lee	Jeff Pollack
tagline	Assassin. Hero. Legend.	A comedy about a guy who would do anything to ...
runtime	100	95
genres	Adventure|Fantasy|Action|Western|Thriller	Comedy|Romance
release_date	2010-12-02 00:00:00	1999-04-23 00:00:00
release_year	2010	1999
profit	-413912431	99

.The Warrior's Way has the biggest budget = 425000000
Whereas Lost & Found has the lowest budget = 1 dollar

3.which movies has the giggest and lowest revenue?

In [272]:

```
# we will call the same function **calculate(column)** again for calculating the hieghest and lowest val
calculate('revenue')
```

Out[272]:

	1386	5067
budget	237000000	6000000
revenue	2781505847	2
profit_earned0	2544505847	-5999998
profit_earned2	2544505847	-5999998
profit_earned	2544505847	-5999998
original_title	Avatar	Shattered Glass
cast	Sam Worthington|Zoe Saldana|Sigourney Weaver|S...	Hayden Christensen|Peter Sarsgaard|Chloë Sevig...
director	James Cameron	Billy Ray
tagline	Enter the World of Pandora.	NaN
runtime	162	94
genres	Action|Adventure|Fantasy|Science Fiction	Drama|History
release_date	2009-12-10 00:00:00	2003-11-14 00:00:00
release_year	2009	2003
profit	2544505847	-5999998

Avatar has the biggest revenue = 2781505847 dollar.
Whereas Shattered Glass has the lowest revenue = 2 dollar

4.which movies has the longest and shortest runtime?

In [183]:

```
# we will call the same function **calculate(column)** again to calculate the longest and shortest runti
calculate('runtime')
```

Out[183]:

	2107	5162
budget	18000000	10
revenue	871279	5
profit_earned	-17128721	-5
original_title	Carlos	Kid's Story
cast	Edgar Ramírez|Alexander Scheer|Fadi Abi Samra|...	Clayton Watson|Keanu Reeves|Carrie-Anne Moss|K...
director	Olivier Assayas	Shinichiro Watanabe
tagline	The man who hijacked the world	NaN
runtime	338	15
genres	Crime|Drama|Thriller|History	Science Fiction|Animation
release_date	2010-05-19 00:00:00	2003-06-02 00:00:00
release_year	2010	2003
profit	-17128721	-5

Carlos has the longest runtime = 338 minutes.
Whereas Kid's Story has the shortest runtime =15 minutes

5.what is the average runtime of all movies?

In [273]:

```
# making a function to find average of a column
def avg_fun(column):
    return tmdb_data[column].mean()
```

In [274]:

```
#calling above function
avg_fun('runtime')
```

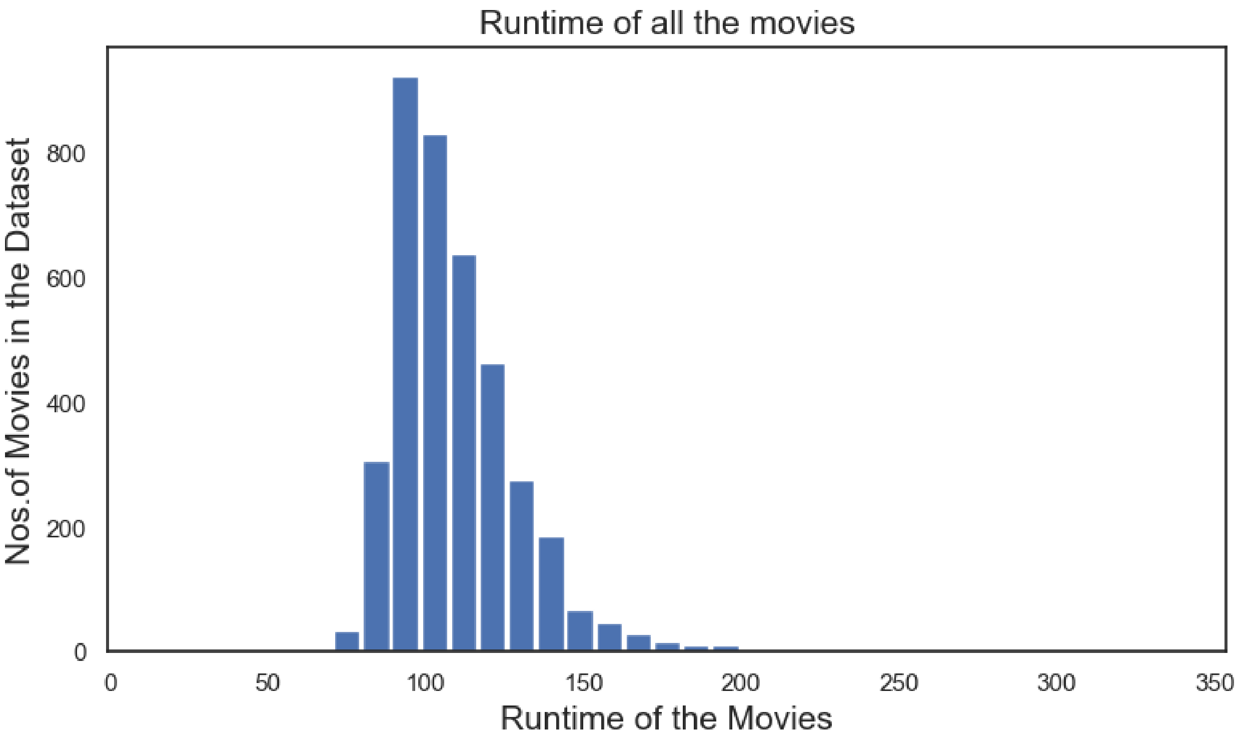
Out[274]:

```
109.22029060716139
```

So the average runtime a movie is 109 minutes. Lets analyse it in a visual form i.e. by graphical approach.

In [186]:

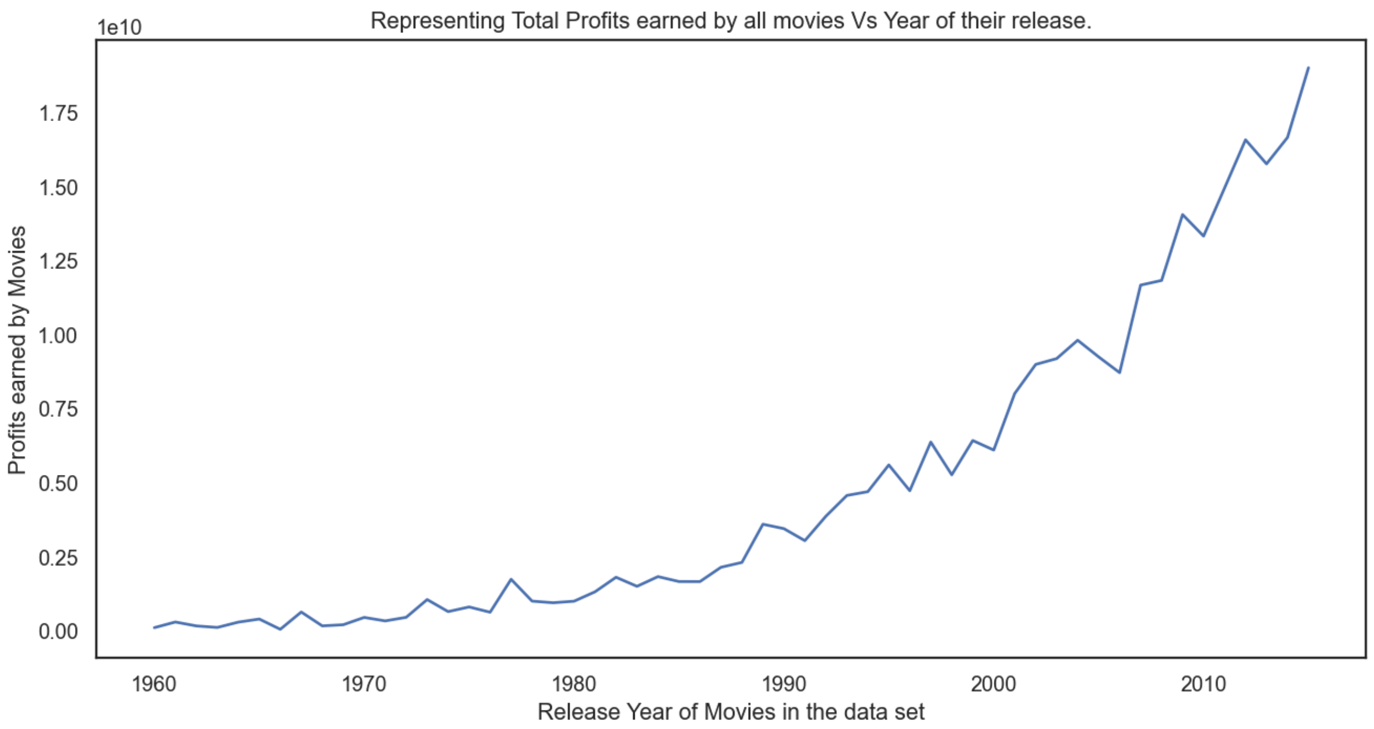
```
#plotting a histogram of runtime of movies
#giving the figure size(width, height)
plt.figure(figsize=(9,5), dpi = 100)
#On x-axis
plt.xlabel('Runtime of the Movies', fontsize = 15)
#On y-axis
plt.ylabel('Nos.of Movies in the Dataset', fontsize=15)
#Name of the graph
plt.title('Runtime of all the movies', fontsize=15)
#giving a histogram plot
plt.hist(tmdb_data['runtime'], rwidth = 0.9, bins =35)
#displays the plot
plt.show()
```



The distribution of the above formed graph is positively skewed or right skewed! Most of the movies are timed between 80 to 115 minutes. Almost 1000 and more no.of movies fall in this criteria.

In [187]:

```
#plotting the relationship between years and its profit
profits_year = tmdb_data.groupby('release_year')['profit_earned'].sum()
#figure size(width, height)
plt.figure(figsize=(12,6), dpi = 130)
#on x-axis
plt.xlabel('Release Year of Movies in the data set', fontsize = 12)
#on y-axis
plt.ylabel('Profits earned by Movies', fontsize = 12)
#title of the line plot
plt.title('Representing Total Profits earned by all movies Vs Year of their release.')
#plotting the graph
plt.plot(profits_year)
#displaying the line plot
plt.show()
```



In [188]:

```
#To find that which year made the highest profit?
profits_year.idxmax()
```

Out[188]:

```
2015
```

So we can conclude both graphically as well as by calculations that year 2015 was the year where movies made the highest profit
We will now find characteristics of profitable movies

6. what are the most successful genres of movies?

In [275]:

```
profit_data = tmdb_data[tmdb_data['profit_earned'] != 0 ]
#reindexing new data
profit_data.index = range(len(profit_data))
#we will start from 1 instead of 0
profit_data.index = profit_data.index + 1
#printing the changed dataset
profit_data.head(3)
```

Out[275]:

	budget	revenue	profit_earned0	profit_earned2	profit_earned	original_title	cast	director	tagline	runtime	
1	150000000	1513528810	1363528810	1363528810	1363528810	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	Colin Trevorrow	The park is open.	124	Action|Adve F
2	150000000	378436354	228436354	228436354	228436354	Mad Max: Fury Road	Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...	George Miller	What a Lovely Day.	120	Action|Adve F
3	110000000	295238201	185238201	185238201	185238201	Insurgent	Shailene Woodley|Theo James|Kate Winslet|Ansel...	Robert Schwentke	One Choice Can Destroy You	119	Adve F

In [276]:

```
#function which will take any column as argument from and keep its track
def data(column):
    #will take a column, and separate the string by '|'
    data = profit_data[column].str.cat(sep = '|')
    #giving pandas series and storing the values separately
    data = pd.Series(data.split('|'))
    #arranging in descending order
    count = data.value_counts(ascending = False)
    return count
```

In [277]:

```
#variable to store the retured value
count = data('genres')
#printing top 10 values
count.head(10)
```

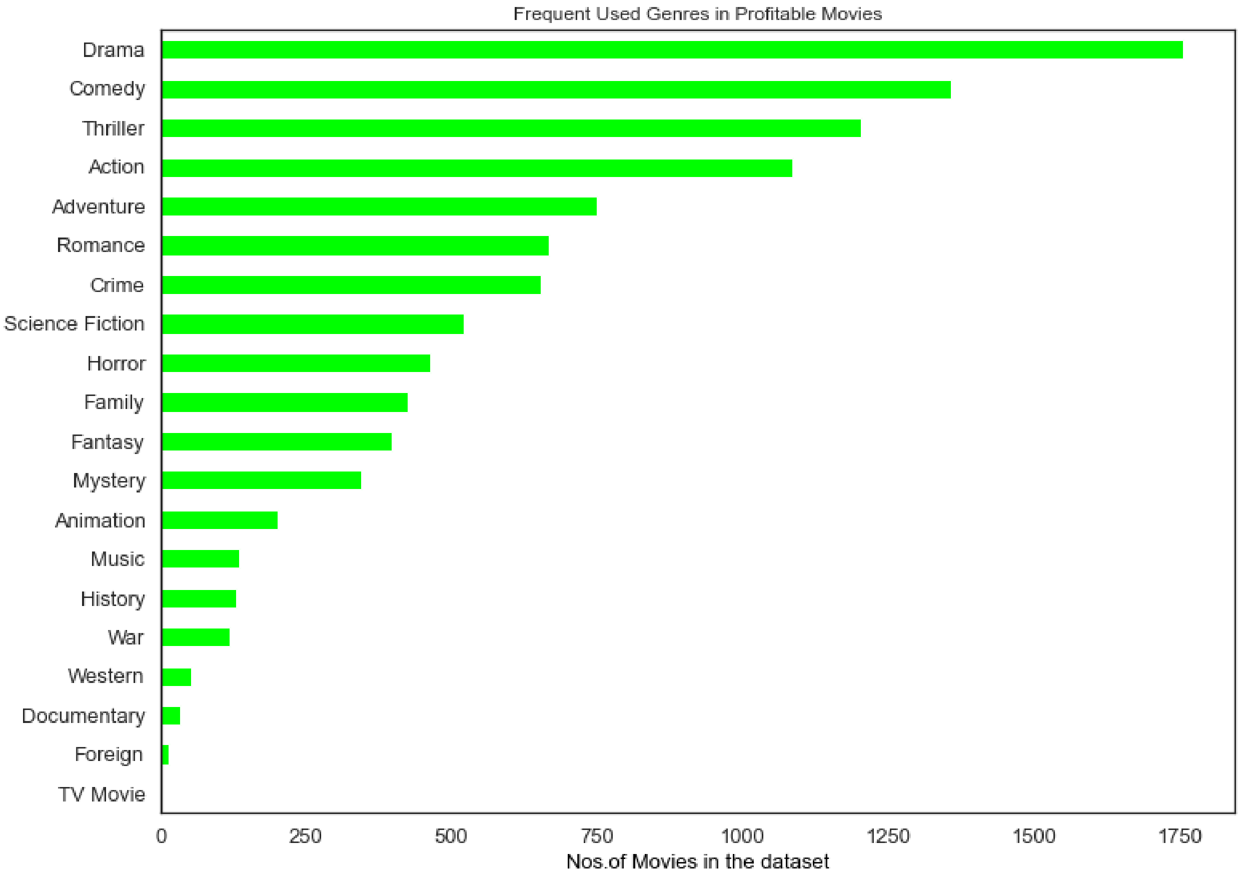
Out[277]:

```
Drama              1755
Comedy             1356
Thriller           1204
Action             1084
Adventure           749
Romance             667
Crime               651
Science Fiction     519
Horror              463
Family              425
dtype: int64
```

Lets to a graphical analysis of the above collected data

In [278]:

```
#lets plot the points in descending order top to bottom as we have data in same format.
count.sort_values(ascending = True, inplace = True)
#ploting
lt = count.plot.barh(color = '#00FF00', fontsize = 13)
#title
lt.set(title = 'Frequent Used Genres in Profitable Movies')
# on x axis
lt.set_xlabel('Nos.of Movies in the dataset', color = 'black', fontsize = '13')
#figure size(width, height)
lt.figure.set_size_inches(12, 9)
#ploting the graph
plt.show()
```



the top 10 genre are Drama,Comedy,Thriller,Action,Adventure,Romance,Crime,Science Fiction,Horror,Family

7.the most repeated cast?

We will call the same function data(column) again for this analysis

In [279]:

```
#variable to store the retured value
count = data('cast')
#printing top 10 values
count.head(10)
```

Out[279]:

```
Robert De Niro       52
Bruce Willis         46
Samuel L. Jackson    44
Nicolas Cage         43
Matt Damon           36
Johnny Depp          35
Tom Hanks            34
Morgan Freeman       34
Harrison Ford        34
Brad Pitt            34
dtype: int64
```

the top 3 cast are Robert De Niro with 52 cast , Bruce Willis with 46 cast , Samuel L. Jackson with 44 cast

8.what is the average budget ?

In [280]:

```
#New function to find average
def profit_avg(column):
    return profit_data[column].mean()
```

In [281]:

```
# calling the above function for budget
profit_avg('budget')
```

Out[281]:

```
37241986.903376624
```

So the average budget of all movies are equal to 37 millon dollars

9. what is the average Revenue earned ?

In [196]:

```
# calling the above function for revenue
profit_avg('revenue')
```

Out[196]:

```
107798135.0535065
```

so the average revenue = 107 millon dollars

In [198]:

```
# calculating ratio between profit mean of the category and No of movies in it
go=(profit_data['profit'].mean()/len(profit_data))
print(go)
```

```
18326.27224678698
```

- #### secound part of the strategy

  Before moving further we need to clean our data again. We will be considering only those movies who have earned a significant amount of profit

  peaking up 4 catergory of movies according to its profit

  choosing catergory of revenue more than (25,50,100,150)M$

1)category of more than 25M and less than 50M profit movies

What is the average budget of the movie w.r.t Profit of movies making more than 25M and less than 50M Dollars?

In [199]:

```
# Dataframe which has data of movies which made profit of more the 25M Dollars.
tmdb_data['profit'] = tmdb_data['revenue'] - tmdb_data['budget']
tmdb_profit_data = tmdb_data[(tmdb_data['profit'] >= 25000000) + (tmdb_data['profit'] < 50000000) ^ (tmdb_
# Reindexing the dataframe
tmdb_profit_data.index = range(len(tmdb_profit_data))
#showing the dataset
tmdb_profit_data.head()
```

```
C:\Users\Mustafa\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.9_qbz5n2kfra8p0\LocalCache\loca
packages\Python39\site-packages\pandas\core\computation\expressions.py:204: UserWarning: evaluating in Py
thon space because the '+' operator is not supported by numexpr for the bool dtype, use '|' instead
  warnings.warn(
```

Out[199]:

	budget	revenue	profit_earned	original_title	cast	director	tagline	runtime	genres	release_date
0	6000000	35401758	29401758	Room	Brie Larson|Jacob Tremblay|Joan Allen|Sean Bri...	Lenny Abrahamson	Love knows no boundaries	117	Drama|Thriller	2015-10-16
1	75000000	108145109	33145109	The Man from U.N.C.L.E.	Henry Cavill|Armie Hammer|Alicia Vikander|Eliz...	Guy Ritchie	Saving the world never goes out of style.	116	Comedy|Action|Adventure	2015-08-13
2	11800000	40272135	28472135	Carol	Cate Blanchett|Rooney Mara|Kyle Chandler|Sarah...	Todd Haynes	Some people change your life forever.	118	Romance|Drama	2015-11-20
3	60000000	101134059	41134059	Joy	Jennifer Lawrence|Bradley Cooper|Robert De Nir...	David O. Russell	NaN	124	Comedy|Drama	2015-12-24
4	105000000	133718711	28718711	Point Break	Edgar Ramírez|Luke Bracey|Teresa Palmer|Delroy...	Ericson Core	The only law that matters is gravity	114	Action|Crime|Thriller	2015-12-03

In [200]:

```
# Printing the info of the new dataframe
tmdb_profit_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 453 entries, 0 to 452
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   budget          453 non-null    int64
 1   revenue         453 non-null    int64
 2   profit_earned   453 non-null    int64
 3   original_title  453 non-null    object
 4   cast            452 non-null    object
 5   director        453 non-null    object
 6   tagline         433 non-null    object
 7   runtime         453 non-null    int64
 8   genres          453 non-null    object
 9   release_date    453 non-null    datetime64[ns]
 10  release_year    453 non-null    int64
 11  profit          453 non-null    int64
dtypes: datetime64[ns](1), int64(6), object(5)
memory usage: 42.6+ KB
```

We can see that we have 453 movies which has profit more than 25M and less than 50M Dollars

In [282]:

```
# Finfd the average budget of movies which made profit more then 25M Dollars
tmdb_profit_data['budget'].mean()
```

Out[282]:

```
26543858.077262692
```

In [283]:

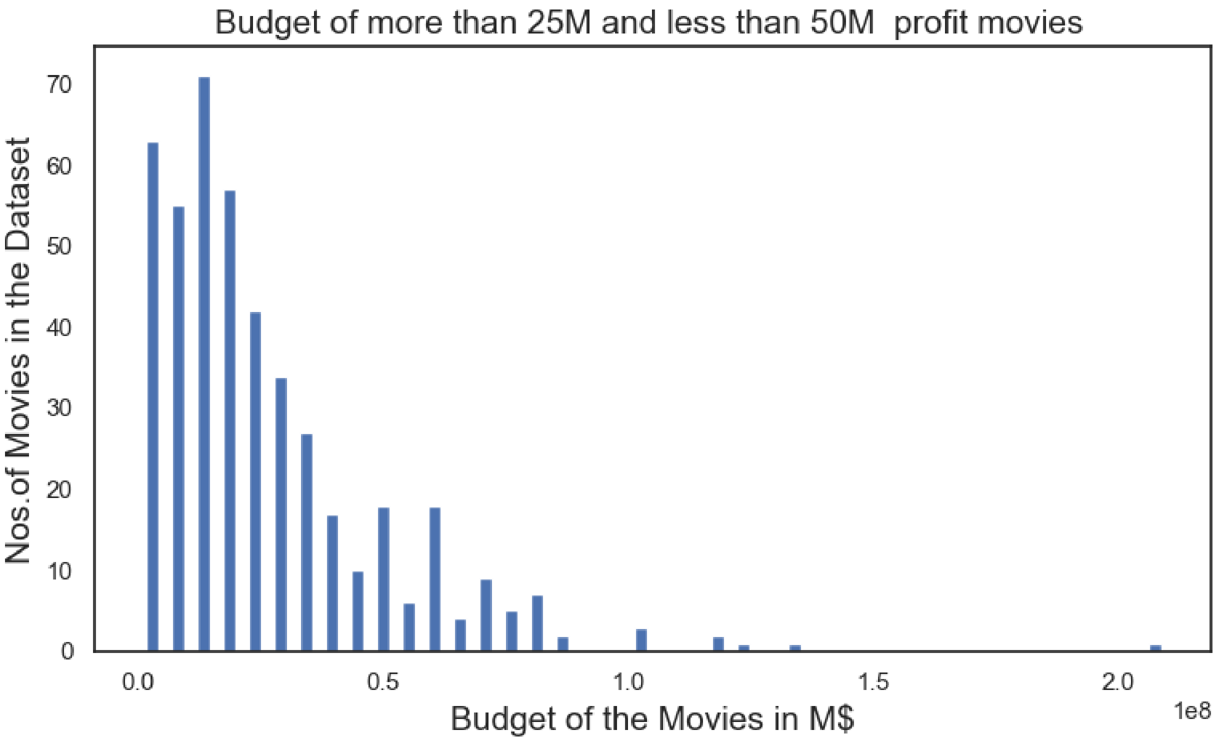
```
#plotting a histogram of budget of movies

#giving the figure size(width, height)
```

```
plt.figure(figsize=(9,5), dpi = 100)

#On x-axis
plt.xlabel('Budget of the Movies in M$', fontsize = 15)
#On y-axis
plt.ylabel('Nos.of Movies in the Dataset', fontsize=15)
#Name of the graph
plt.title('Budget of more than 25M and less than 50M  profit movies ', fontsize=15)

#giving a histogram plot
plt.hist(tmdb_profit_data['budget'], rwidth = 0.5, bins =40)
#displays the plot
plt.show()
```



So the average budget of the movies is 26543858.08 Dollars
more than 90% of the movies in this category has budget less than 30M dollar

What is the average revenue of the movie w.r.t Profit of movies making more then 25M and less than 50M Dollars?

In [284]:

```
# Finfd the average revenue of movies which made profit more then 25M Dollars
tmdb_profit_data['revenue'].mean()
```

Out[284]:

```
62817673.401766
```

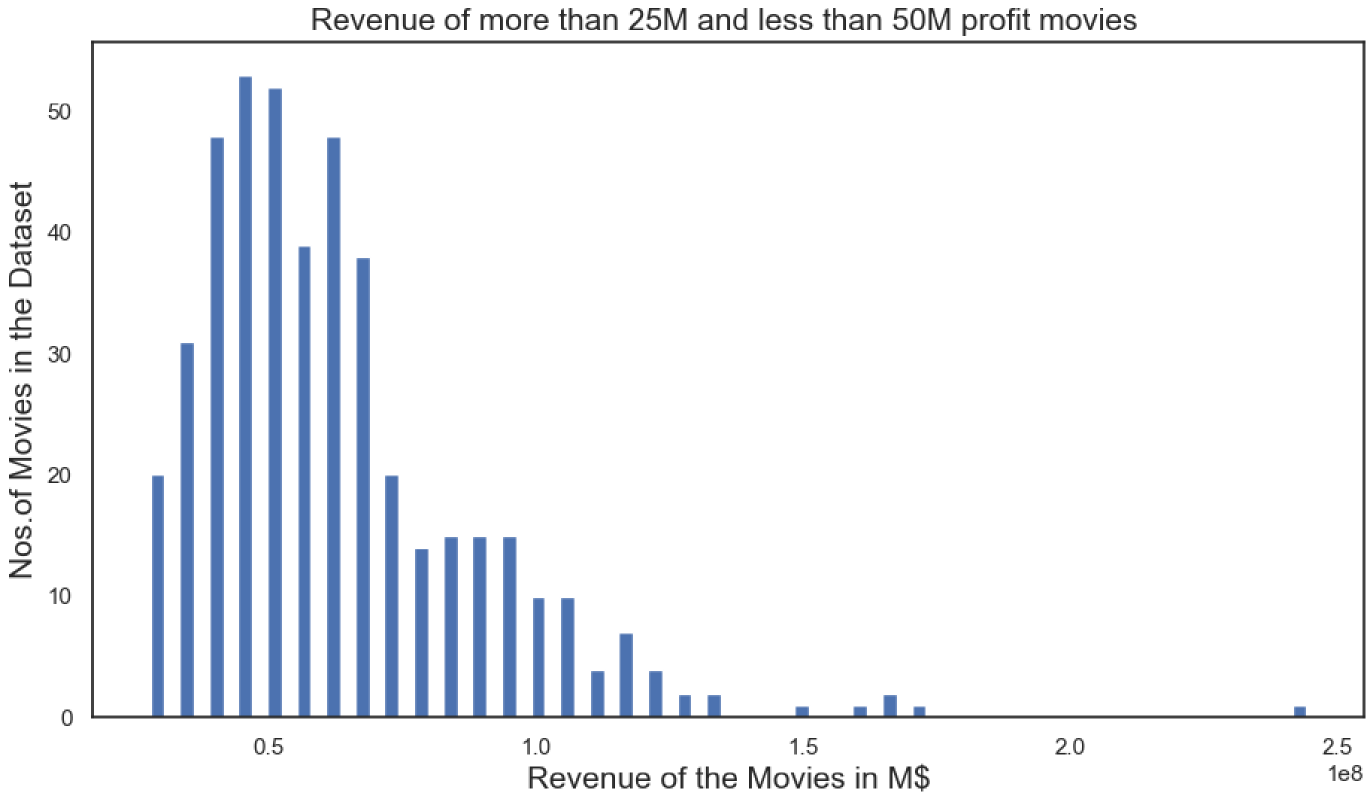
In [296]:

```
#plotting a histogram of revenue of movies

#giving the figure size(width, height)
plt.figure(figsize=(11,6), dpi = 100)

#On x-axis
plt.xlabel('Revenue of the Movies in M$', fontsize = 15)
#On y-axis
plt.ylabel('Nos.of Movies in the Dataset', fontsize=15)
#Name of the graph
plt.title('Revenue of more than 25M and less than 50M profit movies ', fontsize=15)

#giving a histogram plot
plt.hist(tmdb_profit_data['revenue'], rwidth = 0.5, bins =40)
#displays the plot
plt.show()
```

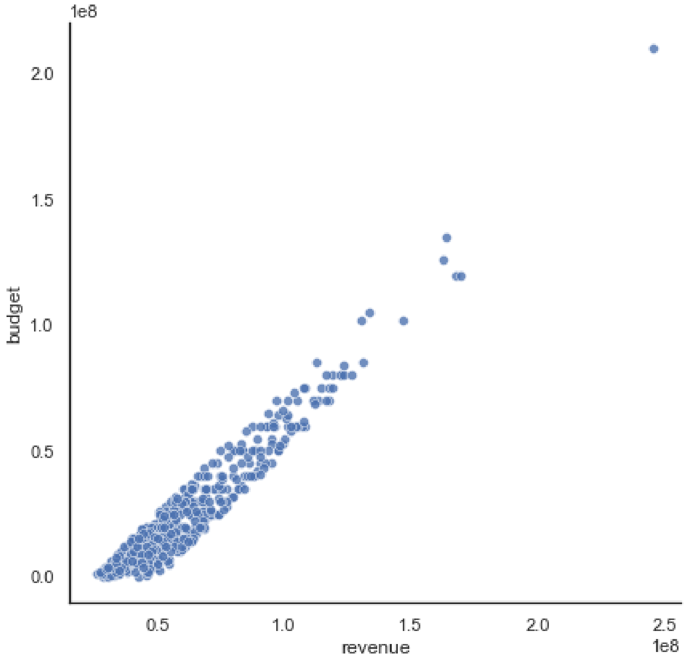

Revenue of more than 25M and less than 50M profit movies

So the average revenue of the movies is 62817673.4 Dollars more than 70% of movies in this category has revenue more than 30M dollars

let's plot a relationship between budget and revenue

In [205]:

```
#let's polt a relationship between budget and revenue
sns.set_theme(style="white")
# Plot miles per gallon against horsepower with other semantics
sns.relplot(x="revenue", y="budget",
            sizes=(40, 400), alpha=.8, palette="muted",
            height=6,data=tmdb_profit_data,facet_kws=dict(sharex=False))
plt.show()
```



more than 90% of the movies in this category have budget less than 30M and revenue >= 30M
it seems that there is a consistancy between the two values

next we calculate value SR which will be a stander to compare between the four category

In [206]:

```
#calculating thr ratio between profit and No of movies in the category
SR=(tmdb_profit_data['profit'].mean()/len(tmdb_profit_data))
print(SR)
```

```
80074.64751545984
```

it seems that each movie in this category hase an average of profit = 80074.65 dollars

What is the average runtime of the movie w.r.t Profit of movies making more then 25M and less than 50M Dollars?

In [286]:

```
# Finfd the average runtime of movies which made profit more then 25M Dollars
tmdb_profit_data['runtime'].mean()
```

Out[286]:

```
109.31567328918322
```

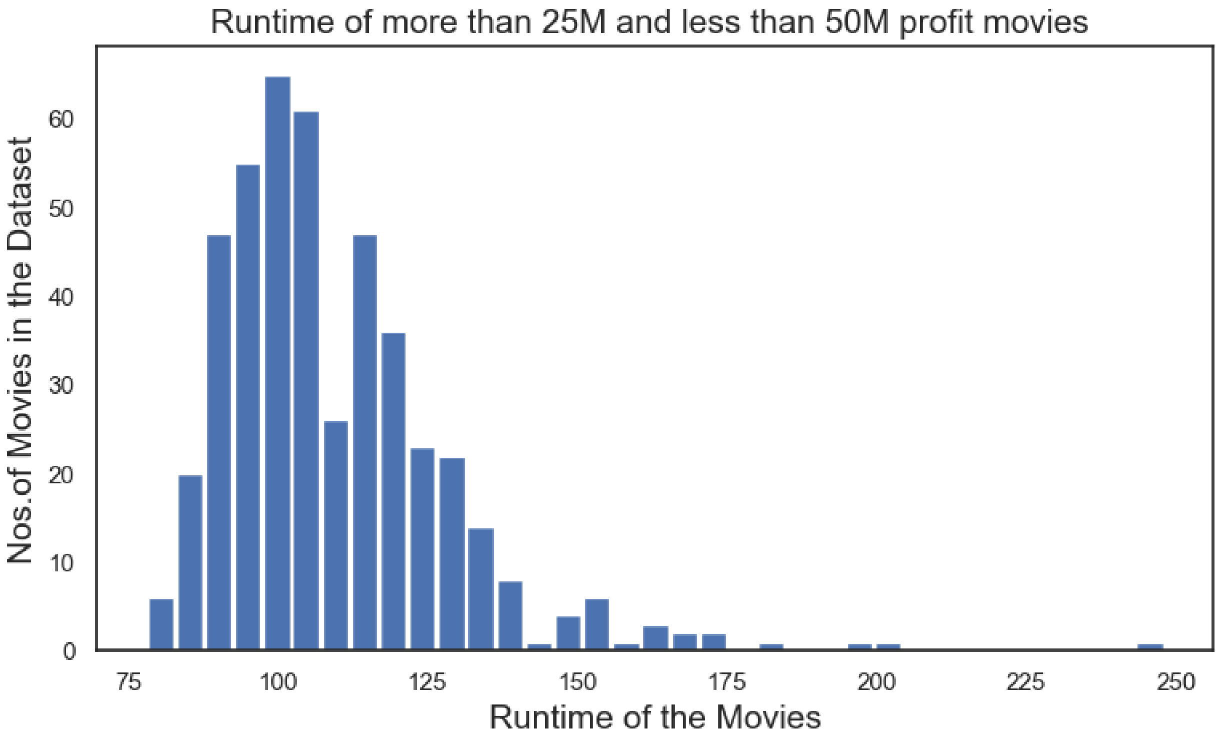
In [287]:

```
#plotting a histogram of runtime of movies

#giving the figure size(width, height)
plt.figure(figsize=(9,5), dpi = 100)

#On x-axis
plt.xlabel('Runtime of the Movies', fontsize = 15)
#On y-axis
plt.ylabel('Nos.of Movies in the Dataset', fontsize=15)
#Name of the graph
plt.title('Runtime of more than 25M and less than 50M profit movies ', fontsize=15)

#giving a histogram plot
plt.hist(tmdb_profit_data['runtime'], rwidth = 0.9, bins =35)
#displays the plot
plt.show()
```

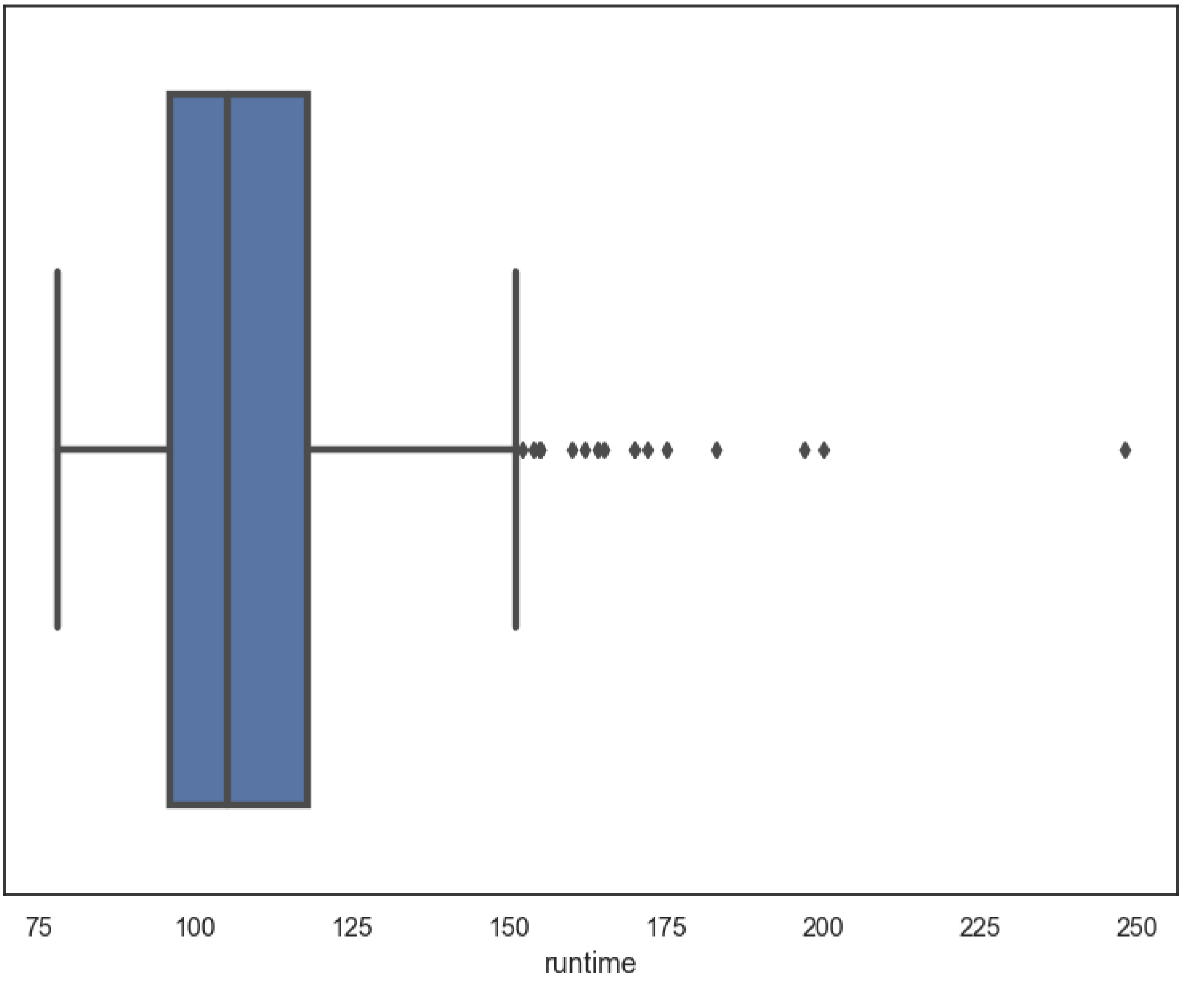


In [209]:

```
#The First plot is box plot of the runtime of the movies
plt.figure(figsize=(9,7), dpi = 105)

#using seaborn to generate the boxplot
sns.boxplot(tmdb_profit_data['runtime'], linewidth = 3)
#diplaying the plot
plt.show()
```

```
C:\Users\Mustafa\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.9_qbz5n2kfra8p0\LocalCache\loca
packages\Python39\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variable as
a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other
arguments without an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```



So the average runtime of the movies is 112.56 Minutes and concentrated between 90-124 minutes

Which are the successfull genres w.r.t Profit of movies making more then 25M and less than 50M Dollars?

In [288]:

```
# This will first concat all the data with | from the whole column and then split it using | and count t
genres_count = pd.Series(tmdb_profit_data['genres'].str.cat(sep = '|').split('|')).value_counts(ascending
genres_count
```

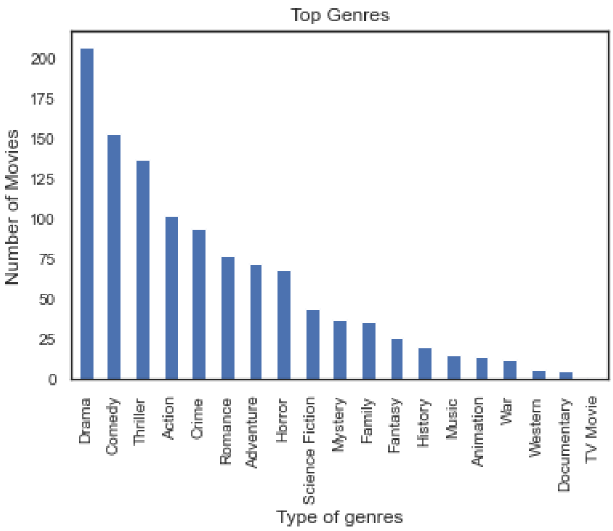
Out[288]:

```
Drama              207
Comedy             153
Thriller           137
Action             102
Crime               94
Romance             77
Adventure           72
Horror              68
Science Fiction     44
Mystery             37
Family              36
Fantasy             26
History             20
Music               15
Animation           14
War                 12
Western              6
Documentary          5
TV Movie             1
dtype: int64
```

So the Top 10 Genres are Drama, Comedy, Thriller, Action, Crime, Romance,Adventure,Horror ,Scince Fiction,mystery Lets visualize this with a plot

In [289]:

```
# Initialize the plot
diagram = genres_count.plot.bar(fontsize = 10)
# Set a title
diagram.set(title = 'Top Genres')
# x-label and y-label
diagram.set_xlabel('Type of genres')
diagram.set_ylabel('Number of Movies')
# Show the plot
plt.show()
```



We can clearly see in the visualization that most movies has Drama as a genre which tends to higher profit

Which are the most frequent cast involved w.r.t Profit of movies making more then 25M Dollars?

In [291]:

```
# This will first concat all the data with | from the whole column and then split it using | and count ti
cast_count = pd.Series(tmdb_profit_data['cast'].str.cat(sep = '|').split('|')).value_counts(ascending = Fa
cast_count.head(10)
```

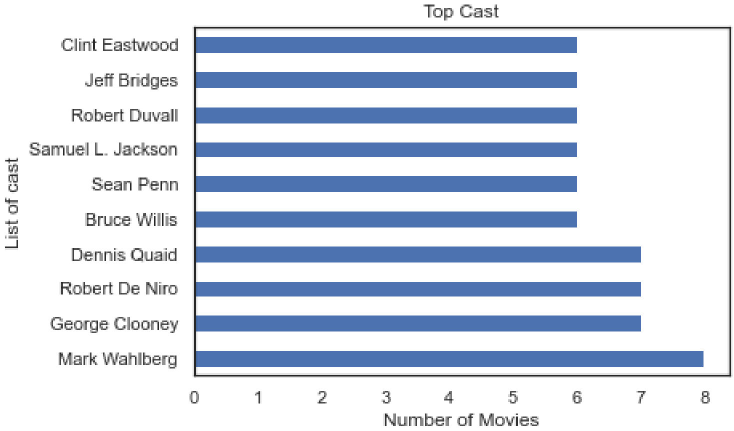
Out[291]:

```
Mark Wahlberg         8
George Clooney        7
Robert De Niro        7
Dennis Quaid          7
Bruce Willis          6
Sean Penn             6
Samuel L. Jackson     6
Robert Duvall         6
Jeff Bridges          6
Clint Eastwood        6
dtype: int64
```

So the Top 5 Mark Wahlberg,George Clooney,Robert De Niro,Dennis Quaid,Bruce Willis Lets visualize this with a plot

In [294]:

```
# Initialize the plot
diagram = cast_count.head(10).plot.barh(fontsize = 11)
# Set a title
diagram.set(title = 'Top Cast')
# x-label and y-label
diagram.set_xlabel('Number of Movies')
diagram.set_ylabel('List of cast')
# Show the plot
plt.show()
```



We can clearly see in the visualization that most movies have Mark Wahlberg as a cast which tends to higher profit.

2)category of more than 50M and less than 100M revenue movies

What is the average budget of the movie w.r.t Profit of movies making more then 50M and less than 100M Dollars?

In [297]:

```
# Dataframe which has data of movies which made profit of more the 50M Dollars.
tmdb_data['profit'] = tmdb_data['revenue'] - tmdb_data['budget']
tmdb_profit_data0 = tmdb_data[(tmdb_data['profit'] >= 50000000) + (tmdb_data['profit'] < 100000000) ^(tmd
# Reindexing the dataframe
tmdb_profit_data0.index = range(len(tmdb_profit_data0))
#showing the dataset
tmdb_profit_data0.head()
```

```
C:\Users\Mustafa\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.9_qbz5n2kfra8p0\LocalCache\loca
packages\Python39\site-packages\pandas\core\computation\expressions.py:204: UserWarning: evaluating in Py
thon space because the '+' operator is not supported by numexpr for the bool dtype, use '|' instead
  warnings.warn(
```

Out[297]:

	budget	revenue	profit_earned0	profit_earned2	profit_earned	original_title	cast	director	tagline	runtime	
0	30000000	91709827	61709827	61709827	61709827	Southpaw	Jake Gyllenhaal|Rachel McAdams|Forest Whitaker...	Antoine Fuqua	Believe in Hope.	123	
1	20000000	88346473	68346473	68346473	68346473	Spotlight	Mark Ruffalo|Michael Keaton|Rachel McAdams|Lie...	Tom McCarthy	Break the story. Break the silence.	128	I
2	49000000	102069268	53069268	53069268	53069268	Chappie	Sharlto Copley|Dev Patel|Ninja|Yolandi Visser|...	Neill Blomkamp	I am consciousness. I am alive. I am Chappie.	120	
3	58000000	150170815	92170815	92170815	92170815	Goosebumps	Jack Black|Dylan Minnette|Odeya Rush|Amy Ryan|...	Rob Letterman	The stories are alive.	103	Adve
4	11000000	62076141	51076141	51076141	51076141	Brooklyn	Saoirse Ronan|Domhnall Gleeson|Emory Cohen|Emi...	John Crowley	Two countries, two loves, one heart	111	

In [215]:

```
# Printing the info of the new dataframe
tmdb_profit_data0.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 512 entries, 0 to 511
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   budget          512 non-null    int64
 1   revenue         512 non-null    int64
 2   profit_earned   512 non-null    int64
 3   original_title  512 non-null    object
 4   cast            512 non-null    object
 5   director        512 non-null    object
 6   tagline         500 non-null    object
 7   runtime         512 non-null    int64
 8   genres          512 non-null    object
 9   release_date    512 non-null    datetime64[ns]
 10  release_year    512 non-null    int64
 11  profit          512 non-null    int64
dtypes: datetime64[ns](1), int64(6), object(5)
memory usage: 48.1+ KB
```

We can see that we have 512 movies which has profit more then 50M and less than 100M Dollars

In [298]:

```
# Finfd the average budget of movies which made profit more then 50M Dollars
tmdb_profit_data0['budget'].mean()
```

Out[298]:

```
37819309.318359375
```

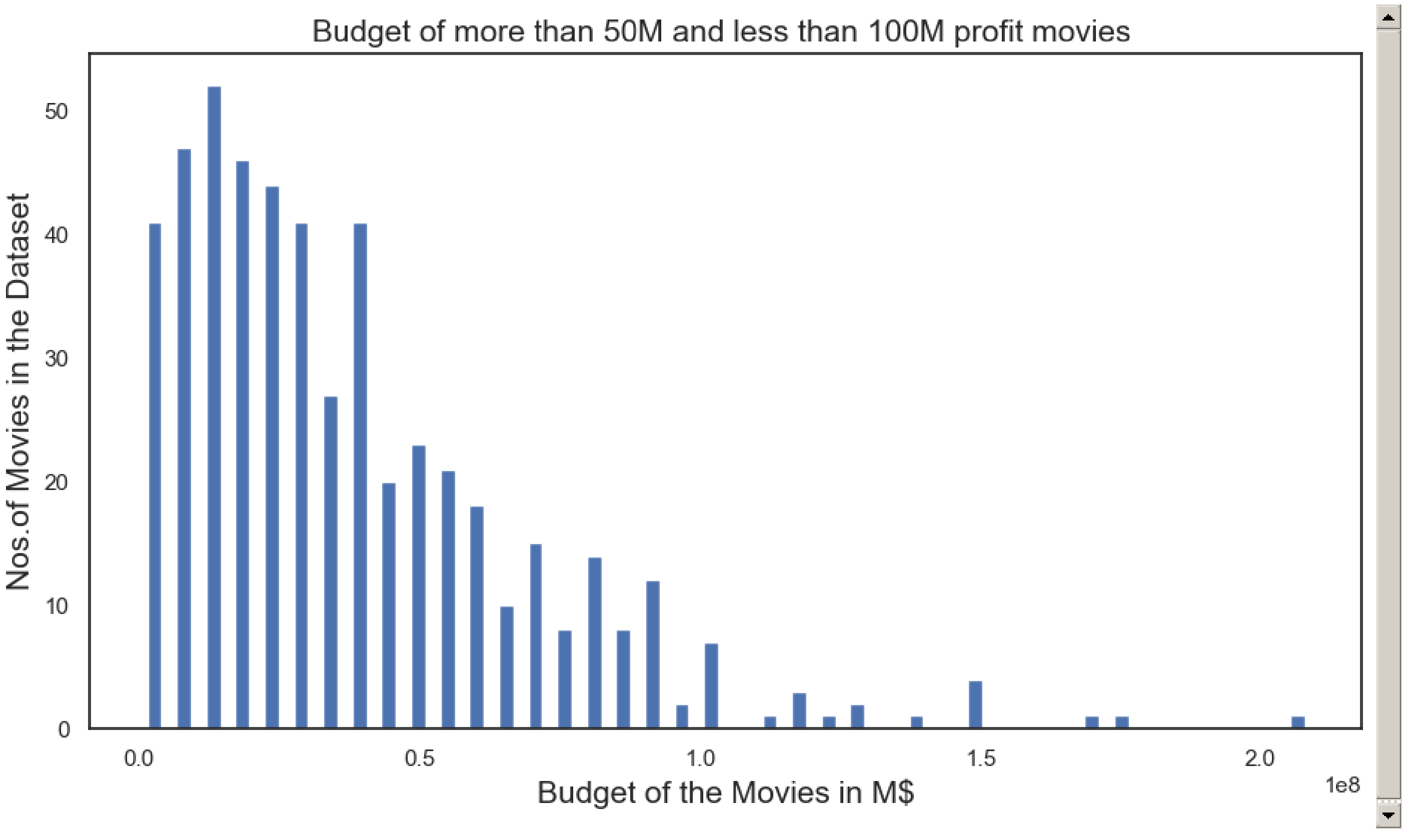
In [299]:

```
#plotting a histogram of budget of movies

#giving the figure size(width, height)
plt.figure(figsize=(11,6), dpi = 100)

#On x-axis
plt.xlabel('Budget of the Movies in M$', fontsize = 15)
#On y-axis
plt.ylabel('Nos.of Movies in the Dataset', fontsize=15)
#Name of the graph
plt.title('Budget of more than 50M and less than 100M profit movies ', fontsize=15)

#giving a histogram plot
plt.hist(tmdb_profit_data0['budget'], rwidth = 0.5, bins =40)
#displays the plot
plt.show()
```



So the average budget of the movies is 37819309 Dollars
more than 80% of movies has budget less than 40M dollar

What is the average revenue of the movie w.r.t Profit of movies making more then 50M and less than 100M Dollars?

In [300]:

```
# Finfd the average revenue of movies which made profit more then 50M Dollars
tmdb_profit_data0['revenue'].mean()
```

Out[300]:

```
109164816.16992188
```

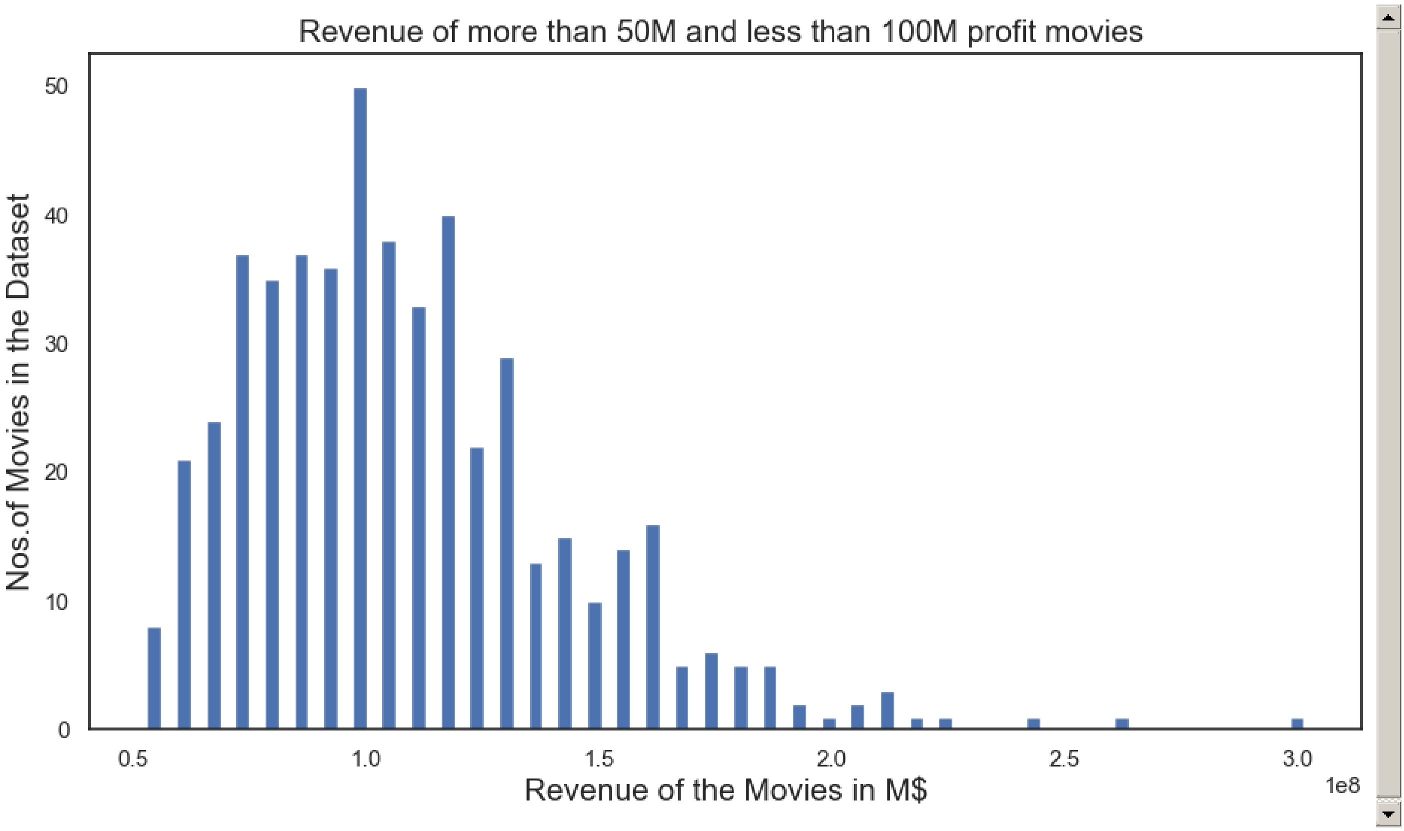
In [301]:

```
#plotting a histogram of revenue of movies

#giving the figure size(width, height)
plt.figure(figsize=(11,6), dpi = 100)

#On x-axis
plt.xlabel('Revenue of the Movies in M$', fontsize = 15)
#On y-axis
plt.ylabel('Nos.of Movies in the Dataset', fontsize=15)
#Name of the graph
plt.title('Revenue of more than 50M and less than 100M profit movies ', fontsize=15)

#giving a histogram plot
plt.hist(tmdb_profit_data0['revenue'], rwidth = 0.5, bins =40)
#displays the plot
plt.show()
```

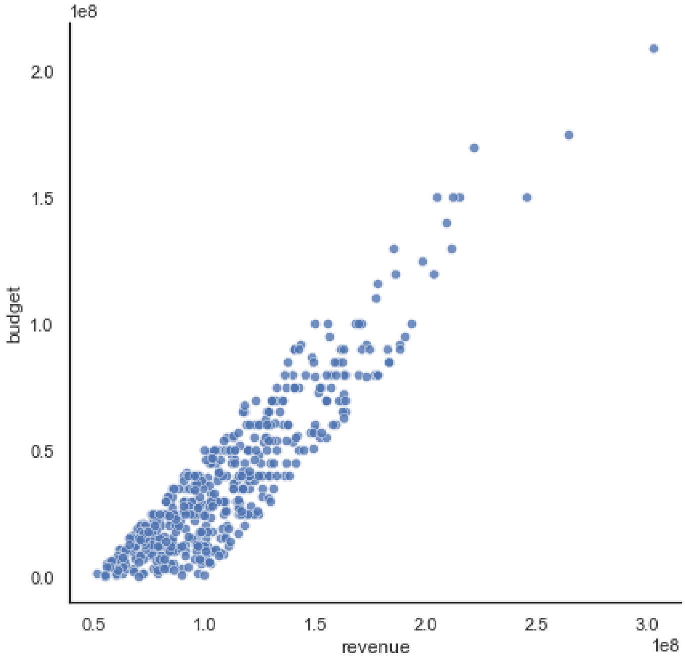


So the average revenue of the movies is 109164816.2 Dollars and more than 50% of movies in this category have revenue greater than 80M dollar

let's plot a relationship between budget and revenue

In [220]:

```
#let's polt a relationship between budget and revenue
sns.set_theme(style="white")
# Plot miles per gallon against horsepower with other semantics
sns.relplot(x="revenue", y="budget",
            sizes=(40, 400), alpha=.8, palette="muted",
            height=6 ,data=tmdb_profit_data0,facet_kws=dict(sharex=False))
plt.show()
```



this much better than the previous category it shows that most of the movies in this cat. have revenue more than 60M dollar with budget less than 30M dollars

In [302]:

```
#calculating thr ratio between profit and No of movies in the category
SR2=(tmdb_profit_data0['profit'].mean()/len(tmdb_profit_data))
print(SR2)
```

```
157495.6001138245
```

the average profit of each movie in this cat. = 157495.6 dollar much better than the previous cat.

What is the average runtime of the movie w.r.t Profit of movies making more then 50M and less than 100M Dollars?

In [303]:

```
# Finfd the average runtime of movies which made profit more then 50M Dollars
tmdb_profit_data0['runtime'].mean()
```

Out[303]:

```
110.50390625
```

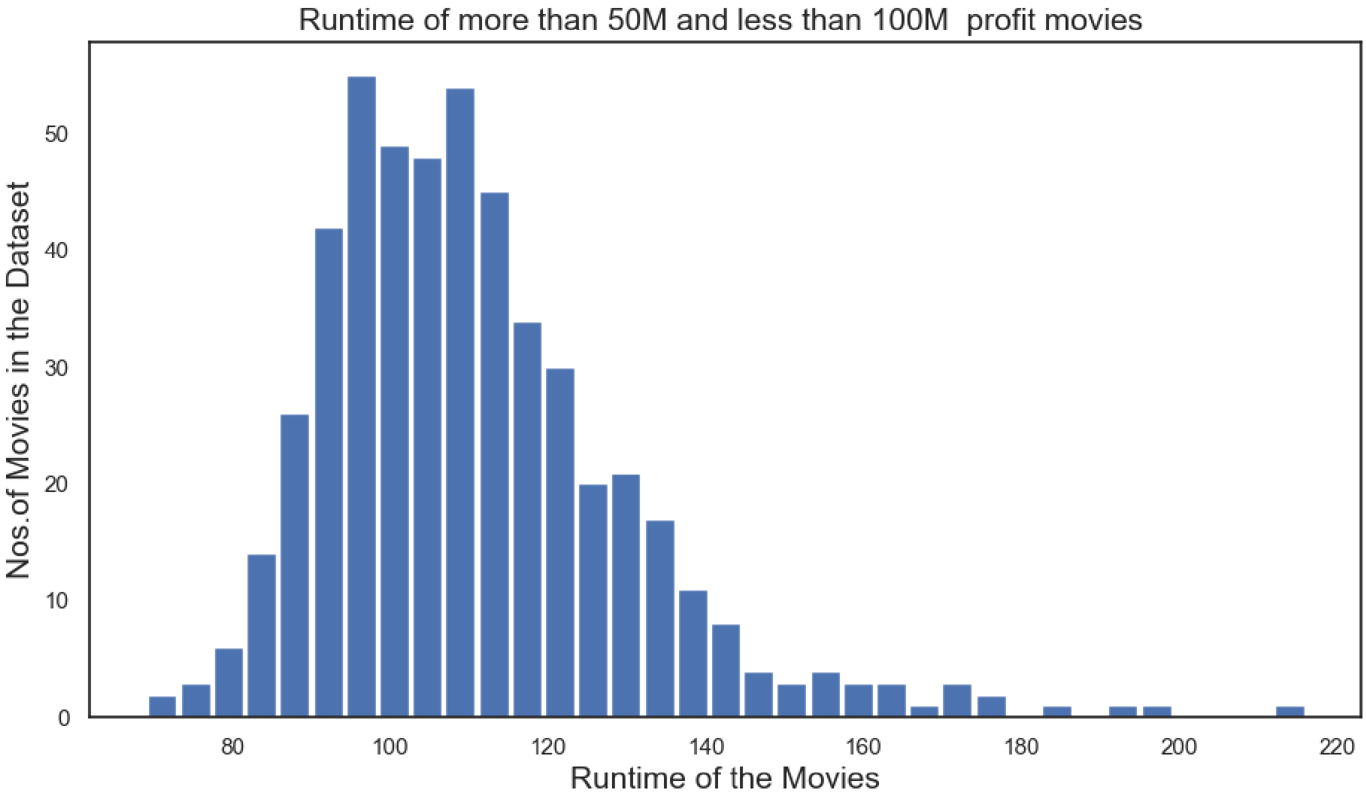
In [304]:

```
#plotting a histogram of runtime of movies

#giving the figure size(width, height)
plt.figure(figsize=(11,6), dpi = 100)

#On x-axis
plt.xlabel('Runtime of the Movies', fontsize = 15)
#On y-axis
plt.ylabel('Nos.of Movies in the Dataset', fontsize=15)
#Name of the graph
plt.title('Runtime of more than 50M and less than 100M  profit movies ', fontsize=15)

#giving a histogram plot
plt.hist(tmdb_profit_data0['runtime'], rwidth = 0.9, bins =35)
#displays the plot
plt.show()
```

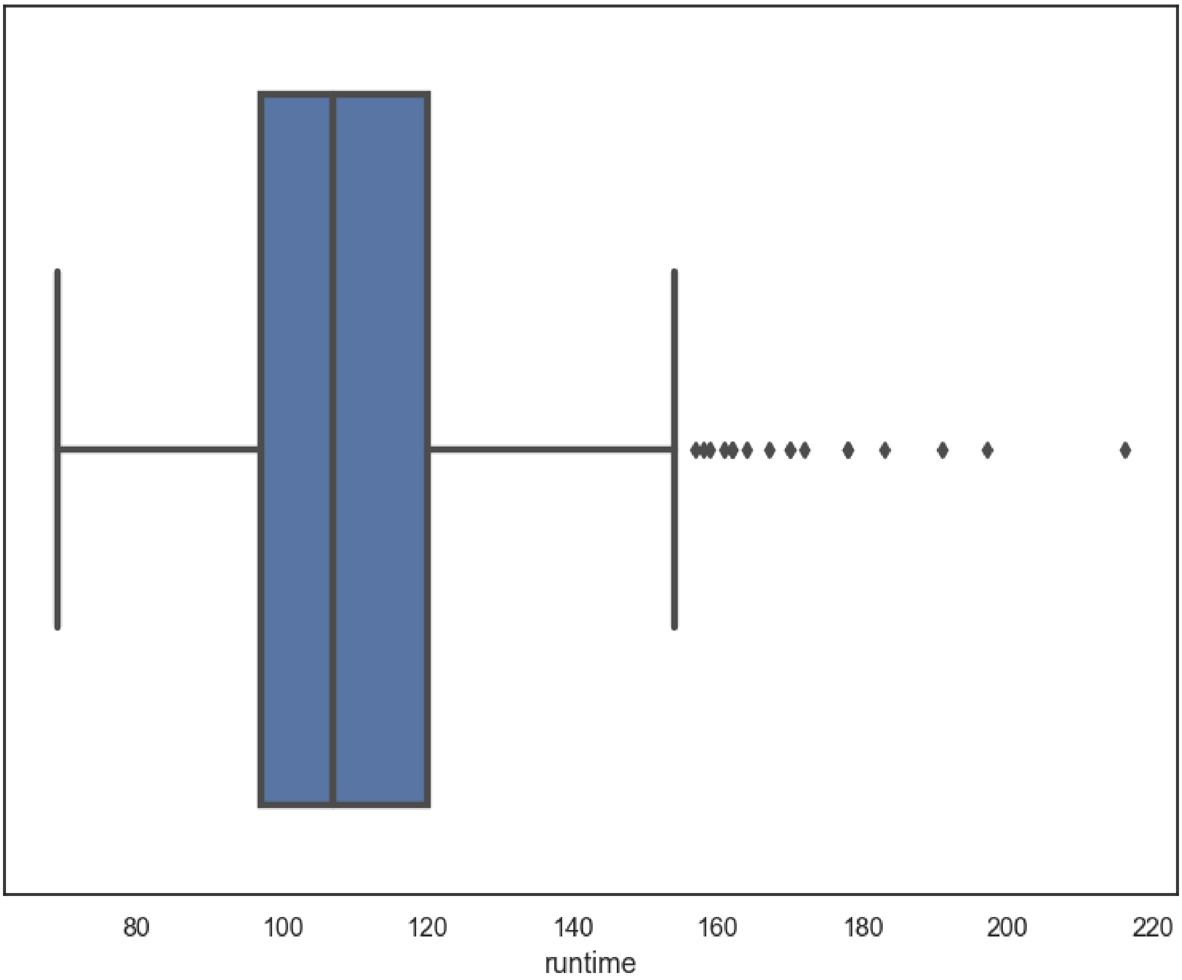


In [224]:

```
#The First plot is box plot of the runtime of the movies
plt.figure(figsize=(9,7), dpi = 105)

#using seaborn to generate the boxplot
sns.boxplot(tmdb_profit_data0['runtime'], linewidth = 3)
#diplaying the plot
plt.show()
```

```
C:\Users\Mustafa\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.9_qbz5n2kfra8p0\LocalCache\loca
packages\Python39\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variable as
a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other
arguments without an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```



So the average runtime of the movies is 109 Minutes focused in 95-120 minutes

Which are the successfull genres w.r.t Profit of movies making more then 50M and less than 100M Dollars?

In [225]:

```
# This will first concat all the data with | from the whole column and then split it using | and count t
genres_count = pd.Series(tmdb_profit_data0['genres'].str.cat(sep = '|').split('|')).value_counts(ascendin
genres_count
```

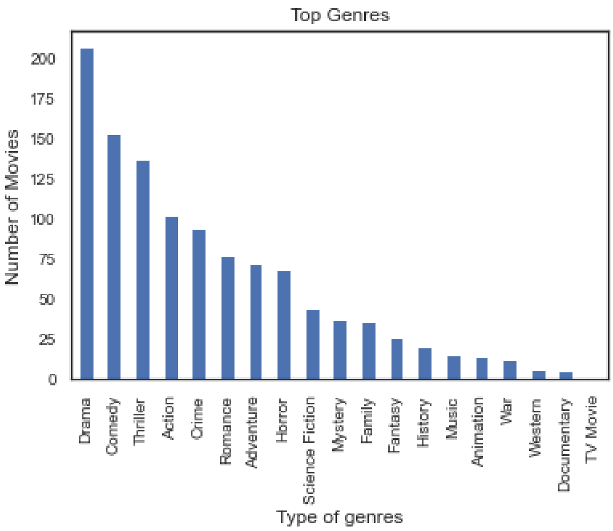
Out[225]:

```
Drama              217
Comedy             193
Thriller           172
Action             147
Adventure          101
Romance             83
Crime               81
Horror              68
Family              63
Science Fiction     58
Fantasy             57
Mystery             47
History             25
Animation           20
War                 17
Music               16
Western              4
Documentary          3
Foreign              1
dtype: int64
```

So the Top 10 Genres are Drama,Comedy,Thriller,Action,Adventure,Romance,Crime,Horror,Family,Science Fiction Lets visualize this with a plot

In [305]:

```
# Initialize the plot
diagram = genres_count.plot.bar(fontsize = 10)
# Set a title
diagram.set(title = 'Top Genres')
# x-label and y-label
diagram.set_xlabel('Type of genres')
diagram.set_ylabel('Number of Movies')
# Show the plot
plt.show()
```



We can clearly see in the visualization that most movies has Drama as a genre which tends to higher profit

Which are the most frequent cast involved w.r.t Profit of movies making more then 50M and less than 100M Dollars?

In [307]:

```
# This will first concat all the data with | from the whole column and then split it using | and count t
cast_count = pd.Series(tmdb_profit_data0['cast'].str.cat(sep = '|').split('|')).value_counts(ascending = 
cast_count.head(10)
```

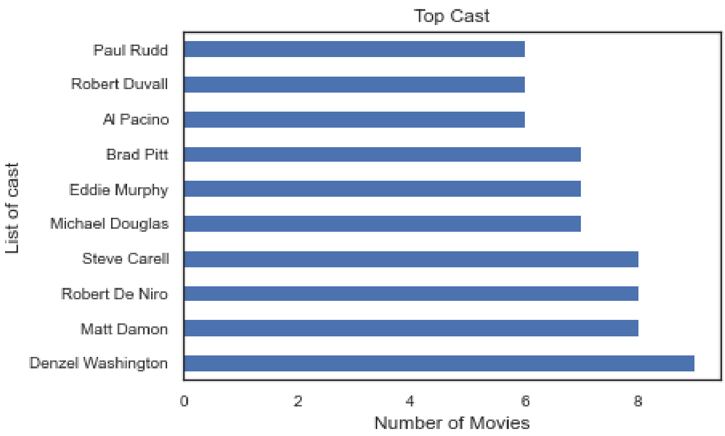
Out[307]:

```
Denzel Washington    9
Matt Damon           8
Robert De Niro       8
Steve Carell         8
Michael Douglas      7
Eddie Murphy         7
Brad Pitt            7
Al Pacino            6
Robert Duvall        6
Paul Rudd            6
dtype: int64
```

So the Top 5 cast are Denzel Washington,Matt Damon,Robert De Niro,Steve Carell,Michael Douglas
Lets visualize this with a plot

In [308]:

```
# Initialize the plot
diagram = cast_count.head(10).plot.barh(fontsize = 10)
# Set a title
diagram.set(title = 'Top Cast')
# x-label and y-label
diagram.set_xlabel('Number of Movies')
diagram.set_ylabel('List of cast')
# Show the plot
plt.show()
```



We can clearly see in the visualization that most movies have Denzel Washington as a cast which tends to higher profit.

3)category of more than 100M and less than 150M revenue movies

What is the average budget of the movie w.r.t Profit of movies making more than 100M and less than 150M Dollars?

In [309]:

```
# Dataframe which has data of movies which made profit of more the 50M Dollars.
tmdb_data['profit'] = tmdb_data['revenue'] - tmdb_data['budget']
tmdb_profit_data2 = tmdb_data[(tmdb_data['profit'] >= 100000000) + (tmdb_data['profit'] < 150000000) ^ (t
# Reindexing the dataframe
tmdb_profit_data2.index = range(len(tmdb_profit_data2))
#showing the dataset
tmdb_profit_data2.head()
```

```
C:\Users\Mustafa\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.9_qbz5n2kfra8p0\LocalCache\loca
packages\Python39\site-packages\pandas\core\computation\expressions.py:204: UserWarning: evaluating in Py
thon space because the '+' operator is not supported by numexpr for the bool dtype, use '|' instead
  warnings.warn(
```

Out[309]:

	budget	revenue	profit_earned0	profit_earned2	profit_earned	original_title	cast	director	tagline	runtime	
0	44000000	155760117	111760117	111760117	111760117	The Hateful Eight	Samuel L. Jackson|Kurt Russell|Jennifer Jason ...	Quentin Tarantino	No one comes up here without a damn good reason.	167	Crime|Dr
1	28000000	133346506	105346506	105346506	105346506	The Big Short	Christian Bale|Steve Carell|Ryan Gosling|Brad ...	Adam McKay	This is a true story.	130	
2	68000000	215863606	147863606	147863606	147863606	Ted 2	Mark Wahlberg|Seth MacFarlane|Amanda Seyfried|...	Seth MacFarlane	Ted is Coming, Again.	115	
3	40000000	162610473	122610473	122610473	122610473	Bridge of Spies	Tom Hanks|Mark Rylance|Amy Ryan|Alan Alda|Seba...	Steven Spielberg	In the shadow of war, one man showed the world...	141	
4	55000000	203427584	148427584	148427584	148427584	Everest	Jason Clarke|Jake Gyllenhaal|Josh Brolin|John ...	Baltasar Kormákur	The Storm Awaits.	121	

In [230]:

```
# Printing the info of the new dataframe
tmdb_profit_data2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 278 entries, 0 to 277
Data columns (total 12 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   budget           278 non-null    int64
 1   revenue          278 non-null    int64
 2   profit_earned    278 non-null    int64
 3   original_title   278 non-null    object
 4   cast             278 non-null    object
 5   director         278 non-null    object
 6   tagline          270 non-null    object
 7   runtime          278 non-null    int64
 8   genres           278 non-null    object
 9   release_date     278 non-null    datetime64[ns]
 10  release_year     278 non-null    int64
 11  profit           278 non-null    int64
dtypes: datetime64[ns](1), int64(6), object(5)
memory usage: 26.2+ KB
```

We can see that we have 278 movies in this cat.

In [310]:

```
# Finfd the average budget of movies which made profit more then 50M Dollars
tmdb_profit_data2['budget'].mean()
```

Out[310]:

```
51263946.44964029
```

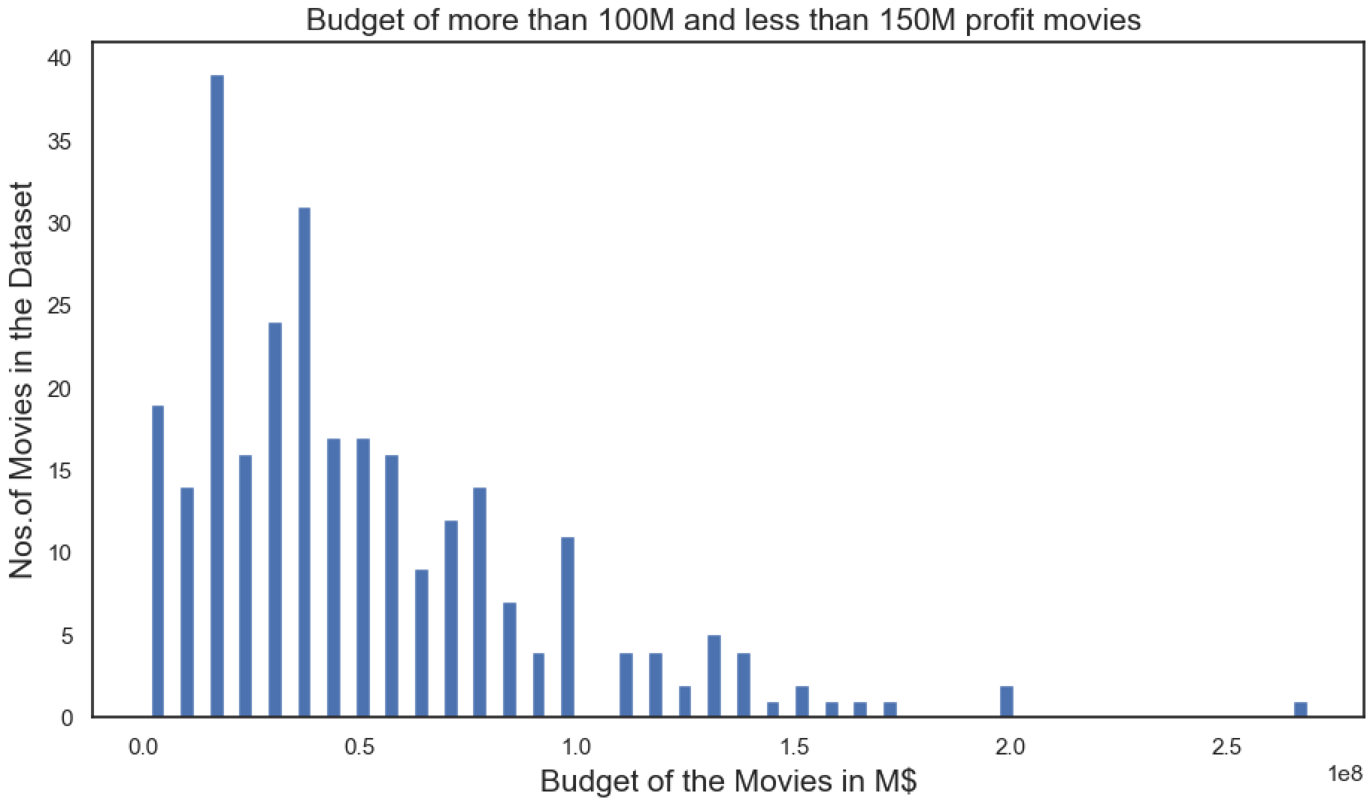
In [311]:

```
#plotting a histogram of budget of movies

#giving the figure size(width, height)
plt.figure(figsize=(11,6), dpi = 100)

#On x-axis
plt.xlabel('Budget of the Movies in M$', fontsize = 15)
#On y-axis
plt.ylabel('Nos.of Movies in the Dataset', fontsize=15)
#Name of the graph
plt.title('Budget of more than 100M and less than 150M profit movies ', fontsize=15)

#giving a histogram plot
plt.hist(tmdb_profit_data2['budget'], rwidth = 0.5, bins =40)
#displays the plot
plt.show()
```



So the average budget of the movies is 51263946.4 Dollars and around 60% are more than 30M

What is the average revenue of the movie w.r.t Profit of movies making more then 100M and less than 150M Dollars?

In [312]:

```
# Finfd the average revenue of movies which made profit more then 50M Dollars
tmdb_profit_data2['revenue'].mean()
```

Out[312]:

```
173312035.4028777
```

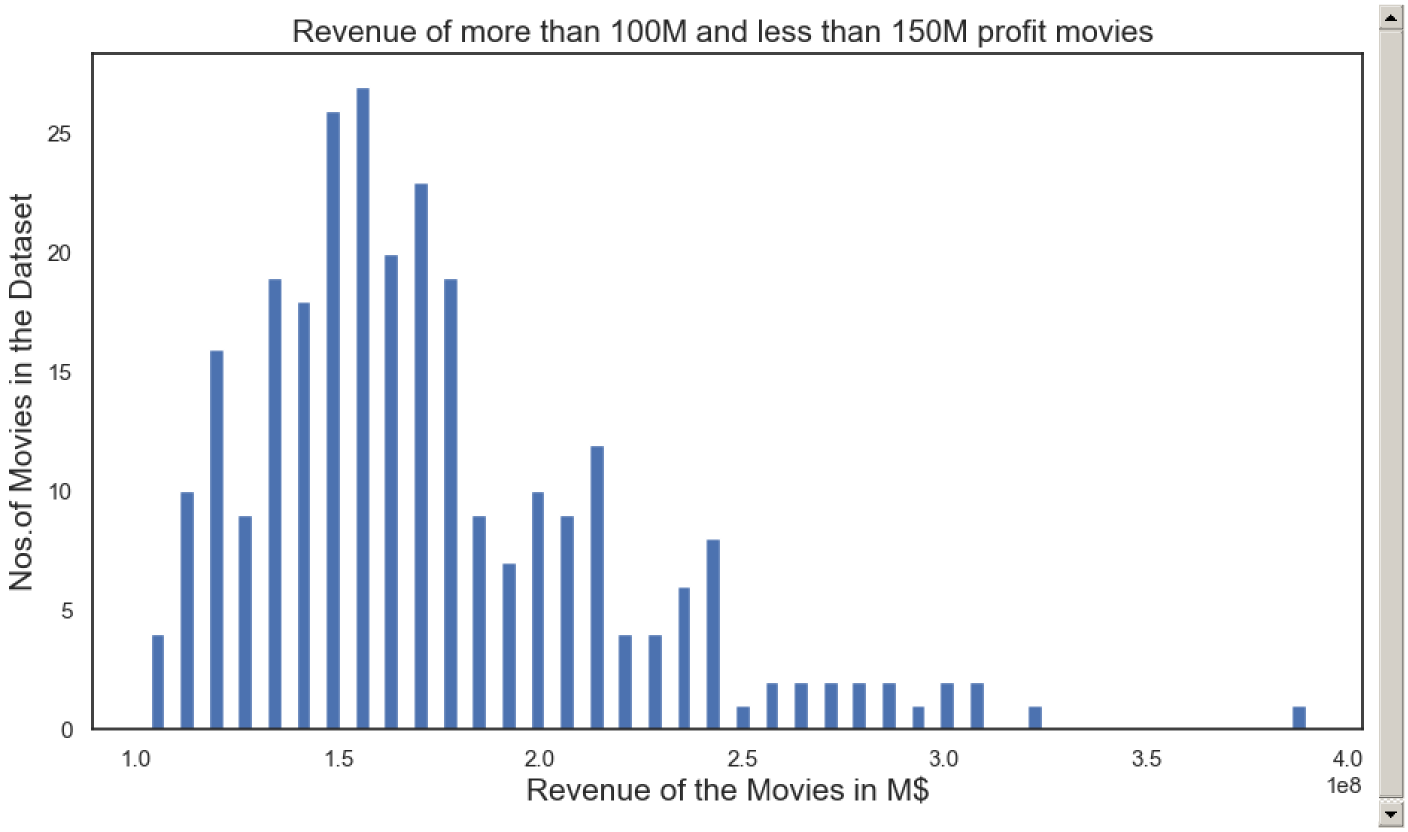
In [313]:

```
#plotting a histogram of revenue of movies

#giving the figure size(width, height)
plt.figure(figsize=(11,6), dpi = 100)

#On x-axis
plt.xlabel('Revenue of the Movies in M$', fontsize = 15)
#On y-axis
plt.ylabel('Nos.of Movies in the Dataset', fontsize=15)
#Name of the graph
plt.title('Revenue of more than 100M and less than 150M profit movies ', fontsize=15)

#giving a histogram plot
plt.hist(tmdb_profit_data2['revenue'], rwidth = 0.5, bins =40)
#displays the plot
plt.show()
```

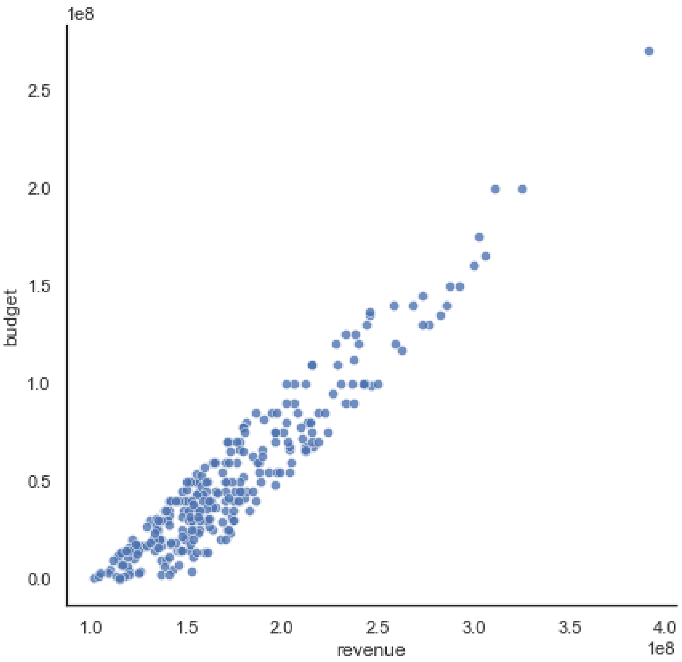


So the average revenue of the movies is 173312035.4 Dollars and more than 55% of the movies have revenue more than 130M

let's plot a relationship between budget and revenue

In [235]:

```
#let's polt a relationship between budget and revenue
sns.set_theme(style="white")
# Plot miles per gallon against horsepower with other semantics
sns.relplot(x="revenue", y="budget",
            sizes=(40, 400), alpha=.8, palette="muted",
            height=6 ,data=tmdb_profit_data2,facet_kws=dict(sharex=False))
plt.show()
```



more than 75% of the movies have budget less than 30M with revenue more than 130M and it is better than the previous one

In [236]:

```
#calculating thr ratio between profit and No of movies in the category
SR1=(tmdb_profit_data2['profit'].mean()/len(tmdb_profit_data))
print(SR1)
```

```
269421.8299188464
```

each movie in this cat. has mean profit = 269421.8 dollars

What is the average runtime of the movie w.r.t Profit of movies making more then 100M and less than 150M Dollars?

In [314]:

```
# Finfd the average runtime of movies which made profit more then 50M Dollars
tmdb_profit_data2['runtime'].mean()
```

Out[314]:

```
112.07194244604317
```

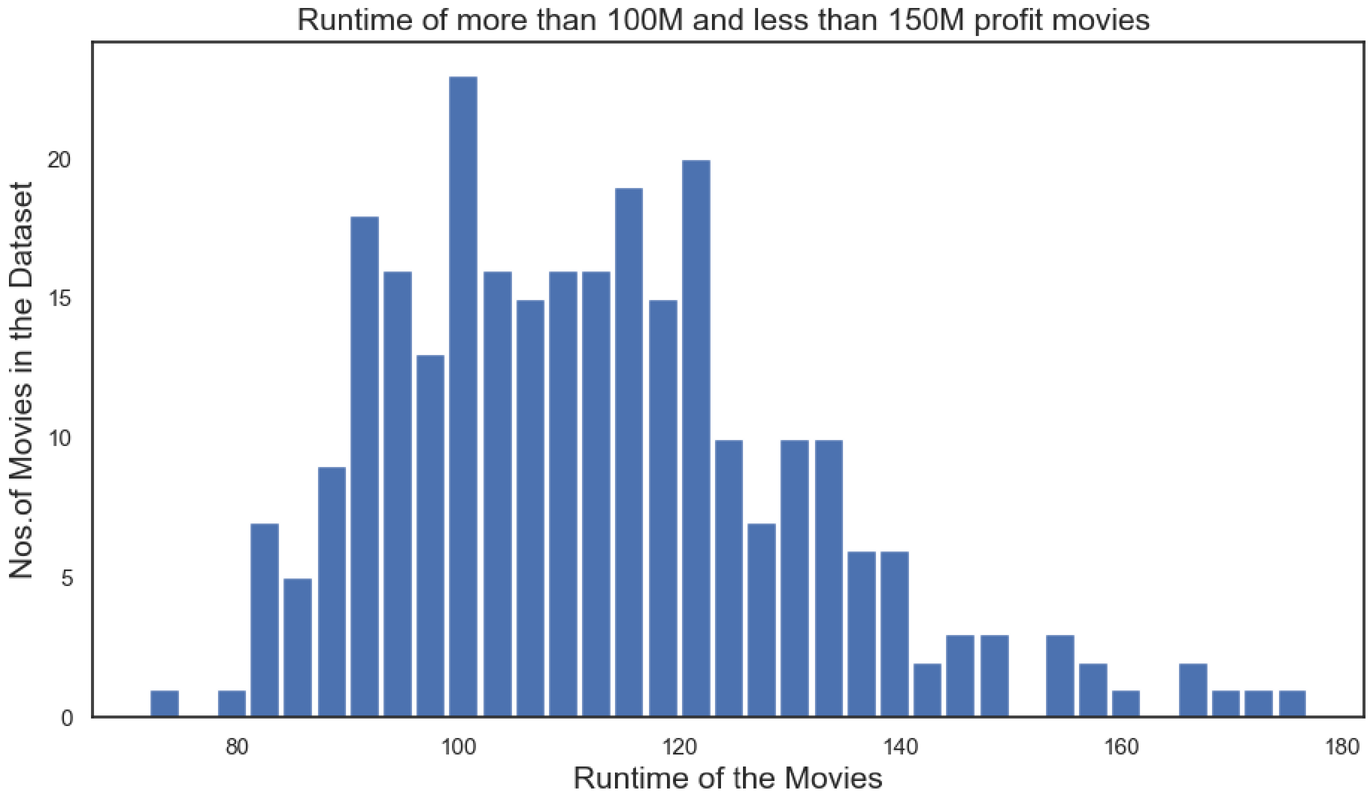
In [315]:

```
#plotting a histogram of runtime of movies

#giving the figure size(width, height)
plt.figure(figsize=(11,6), dpi = 100)

#On x-axis
plt.xlabel('Runtime of the Movies', fontsize = 15)
#On y-axis
plt.ylabel('Nos.of Movies in the Dataset', fontsize=15)
#Name of the graph
plt.title('Runtime of more than 100M and less than 150M profit movies ', fontsize=15)

#giving a histogram plot
plt.hist(tmdb_profit_data2['runtime'], rwidth = 0.9, bins =35)
#displays the plot
plt.show()
```

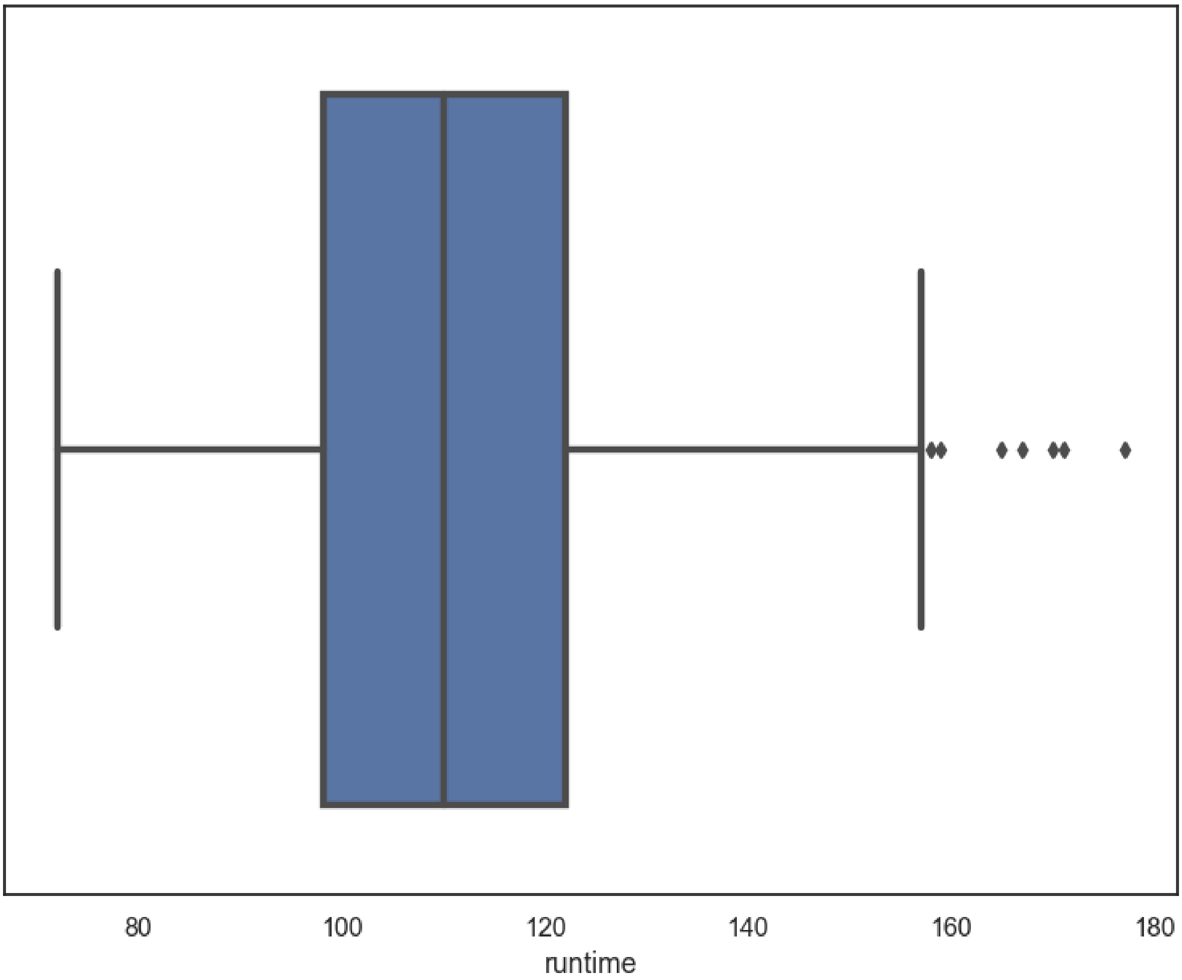


In [239]:

```
#The First plot is box plot of the runtime of the movies
plt.figure(figsize=(9,7), dpi = 105)

#using seaborn to generate the boxplot
sns.boxplot(tmdb_profit_data2['runtime'], linewidth = 3)
#diplaying the plot
plt.show()
```

```
C:\Users\Mustafa\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.9_qbz5n2kfra8p0\LocalCache\loca
packages\Python39\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variable as
a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other
arguments without an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```



So the average runtime of the movies is 112 Minutes and lying between 100-120 minutes

Which are the successfull genres w.r.t Profit of movies making more then 100M and less than 150M Dollars?

In [240]:

```
# This will first concat all the data with | from the whole column and then split it using | and count t
genres_count = pd.Series(tmdb_profit_data2['genres'].str.cat(sep = '|').split('|')).value_counts(ascendin
genres_count
```

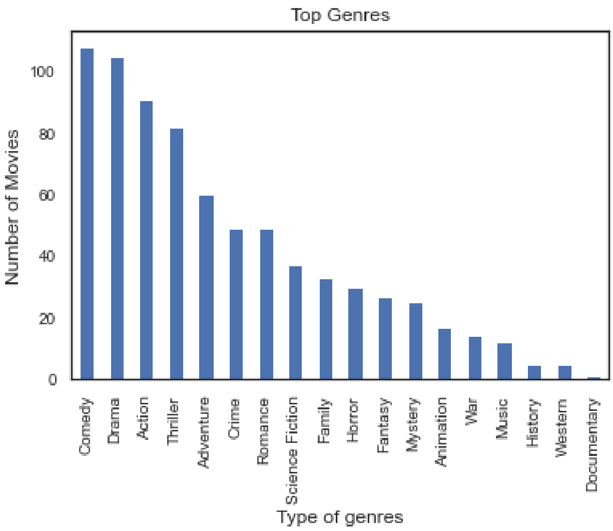
Out[240]:

```
Comedy             108
Drama              105
Action              91
Thriller            82
Adventure           60
Crime               49
Romance             49
Science Fiction     37
Family              33
Horror              30
Fantasy             27
Mystery             25
Animation           17
War                 14
Music               12
History              5
Western              5
Documentary          1
dtype: int64
```

So the Top 10 Genres are Comedy,Drama,Action,Thriller,Adventure,Crime,Romance,Science Fiction,Family,Horror Lets visualize this with a plot

In [241]:

```
# Initialize the plot
diagram = genres_count.plot.bar(fontsize = 10)
# Set a title
diagram.set(title = 'Top Genres')
# x-label and y-label
diagram.set_xlabel('Type of genres')
diagram.set_ylabel('Number of Movies')
# Show the plot
plt.show()
```



We can clearly see in the visualization that most movies has Comedy as a genre which tends to higher profit

Which are the most frequent cast involved w.r.t Profit of movies making more then 100M and less than 150M Dollars?

In [319]:

```
# This will first concat all the data with | from the whole column and then split it using | and count ti
cast_count = pd.Series(tmdb_profit_data2['cast'].str.cat(sep = '|').split('|')).value_counts(ascending = 1
cast_count.head(10)
```

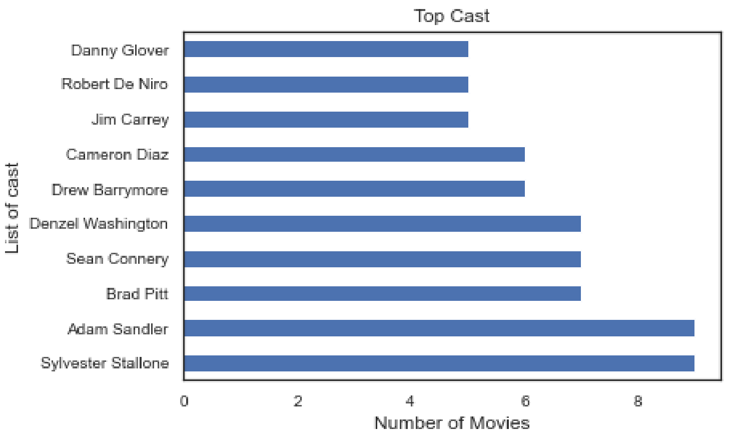
Out[319]:

```
Sylvester Stallone    9
Adam Sandler          9
Brad Pitt             7
Sean Connery          7
Denzel Washington     7
Drew Barrymore        6
Cameron Diaz          6
Jim Carrey            5
Robert De Niro        5
Danny Glover          5
dtype: int64
```

So the Top 5 cast are Sylvester Stallone,Adam Sandler,Brad Pitt,Sean Connery,Denzel Washington Lets visualize this with a plot

In [318]:

```
# Initialize the plot
diagram = cast_count.head(10).plot.barh(fontsize = 10)
# Set a title
diagram.set(title = 'Top Cast')
# x-label and y-label
diagram.set_xlabel('Number of Movies')
diagram.set_ylabel('List of cast')
# Show the plot
plt.show()
```



We can clearly see in the visualization that most movies have Sylvester Stallone as a cast which tends to higher profit.

category of more than 150M revenue movies

What is the average budget of the movie w.r.t Profit of movies making more then 150M Dollars?

In [320]:

```
# Dataframe which has data of movies which made profit of more the 150M Dollars.
tmdb_data['profit'] = tmdb_data['revenue'] - tmdb_data['budget']
tmdb_profit_data1 = tmdb_data[(tmdb_data['profit'] >= 150000000) ]
# Reindexing the dataframe
tmdb_profit_data1.index = range(len(tmdb_profit_data1))
#showing the dataset
tmdb_profit_data1.head()
```

Out[320]:

	budget	revenue	profit_earned0	profit_earned2	profit_earned	original_title	cast	director	tagline	runtime	
0	150000000	1513528810	1363528810	1363528810	1363528810	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	Colin Trevorrow	The park is open.	124	Action|
1	150000000	378436354	228436354	228436354	228436354	Mad Max: Fury Road	Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...	George Miller	What a Lovely Day.	120	Action|
2	110000000	295238201	185238201	185238201	185238201	Insurgent	Shailene Woodley|Theo James|Kate Winslet|Ansel...	Robert Schwentke	One Choice Can Destroy You	119	
3	200000000	2068178225	1868178225	1868178225	1868178225	Star Wars: The Force Awakens	Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...	J.J. Abrams	Every generation has a story.	136	Action|
4	190000000	1506249360	1316249360	1316249360	1316249360	Furious 7	Vin Diesel|Paul Walker|Jason Statham|Michelle ...	James Wan	Vengeance Hits Home	137	Ac

In [321]:

```
# Printing the info of the new dataframe
tmdb_profit_data1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 548 entries, 0 to 547
Data columns (total 14 columns):
 #   Column          Non-Null Count  Dtype         
---  ------          --------------  -----         
 0   budget          548 non-null    int64         
 1   revenue         548 non-null    int64         
 2   profit_earned0  548 non-null    int64         
 3   profit_earned2  548 non-null    int64         
 4   profit_earned   548 non-null    int64         
 5   original_title  548 non-null    object        
 6   cast            548 non-null    object        
 7   director        548 non-null    object        
 8   tagline         543 non-null    object        
 9   runtime         548 non-null    int64         
 10  genres          548 non-null    object        
 11  release_date    548 non-null    datetime64[ns]
 12  release_year    548 non-null    int64         
 13  profit          548 non-null    int64         
dtypes: datetime64[ns](1), int64(8), object(5)
memory usage: 60.1+ KB
```

There are 548 movies with revenue more than 150M

In [246]:

```
# Finfd the average budget of movies which made profit more then 50M Dollars
tmdb_profit_data1['budget'].mean()
```

Out[246]:

```
86241770.0729927
```

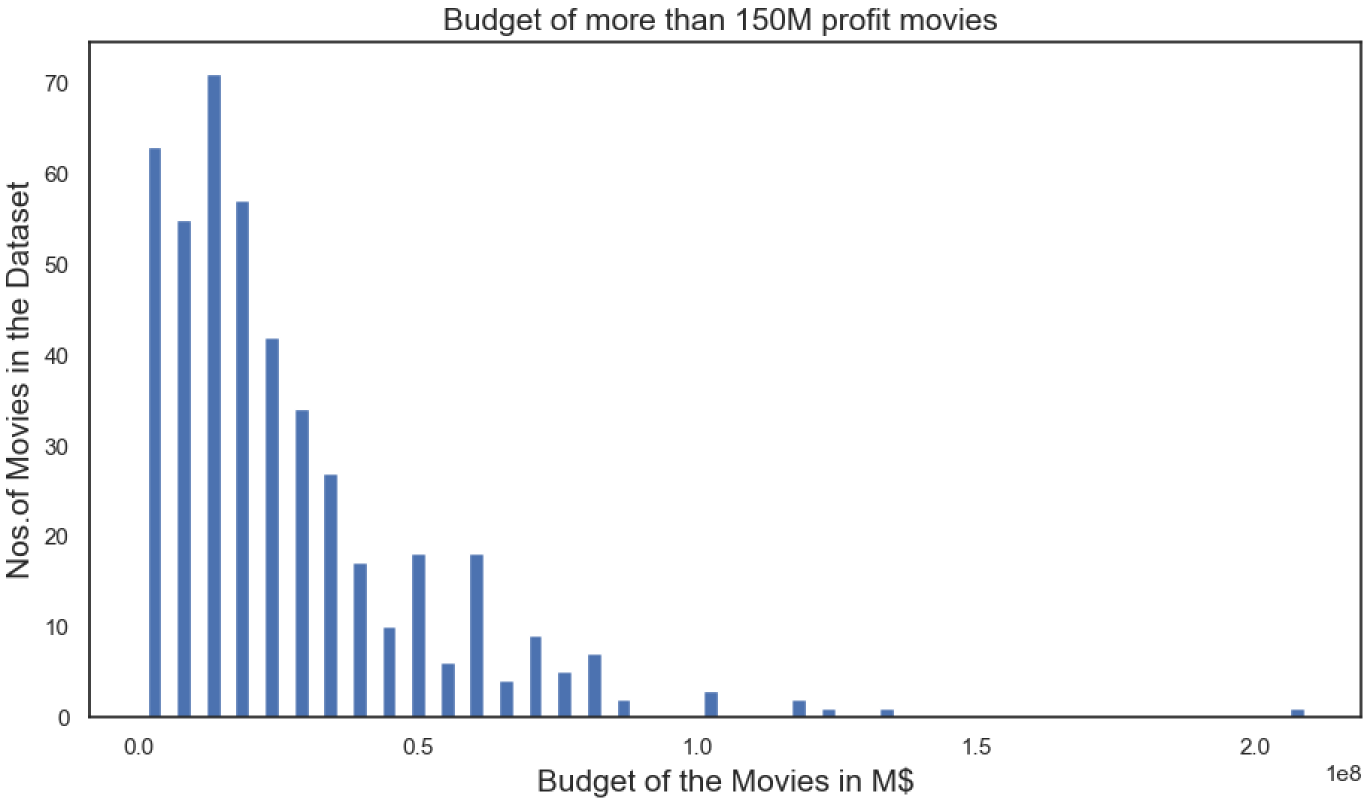
In [322]:

```
#plotting a histogram of budget of movies

#giving the figure size(width, height)
plt.figure(figsize=(11,6), dpi = 100)

#On x-axis
plt.xlabel('Budget of the Movies in M$', fontsize = 15)
#On y-axis
plt.ylabel('Nos.of Movies in the Dataset', fontsize=15)
#Name of the graph
plt.title('Budget of more than 150M profit movies ', fontsize=15)

#giving a histogram plot
plt.hist(tmdb_profit_data['budget'], rwidth = 0.5, bins =40)
#displays the plot
plt.show()
```



So the average budget of the movies is 86241770.1 Dollars and more than 80% with budget less than 100M

What is the average revenue of the movie w.r.t Profit of movies making more then 150M Dollars?

In [323]:

```
# Finfd the average revenue of movies which made profit more then 150M Dollars
tmdb_profit_data1['revenue'].mean()
```

Out[323]:

```
432591643.8521898
```

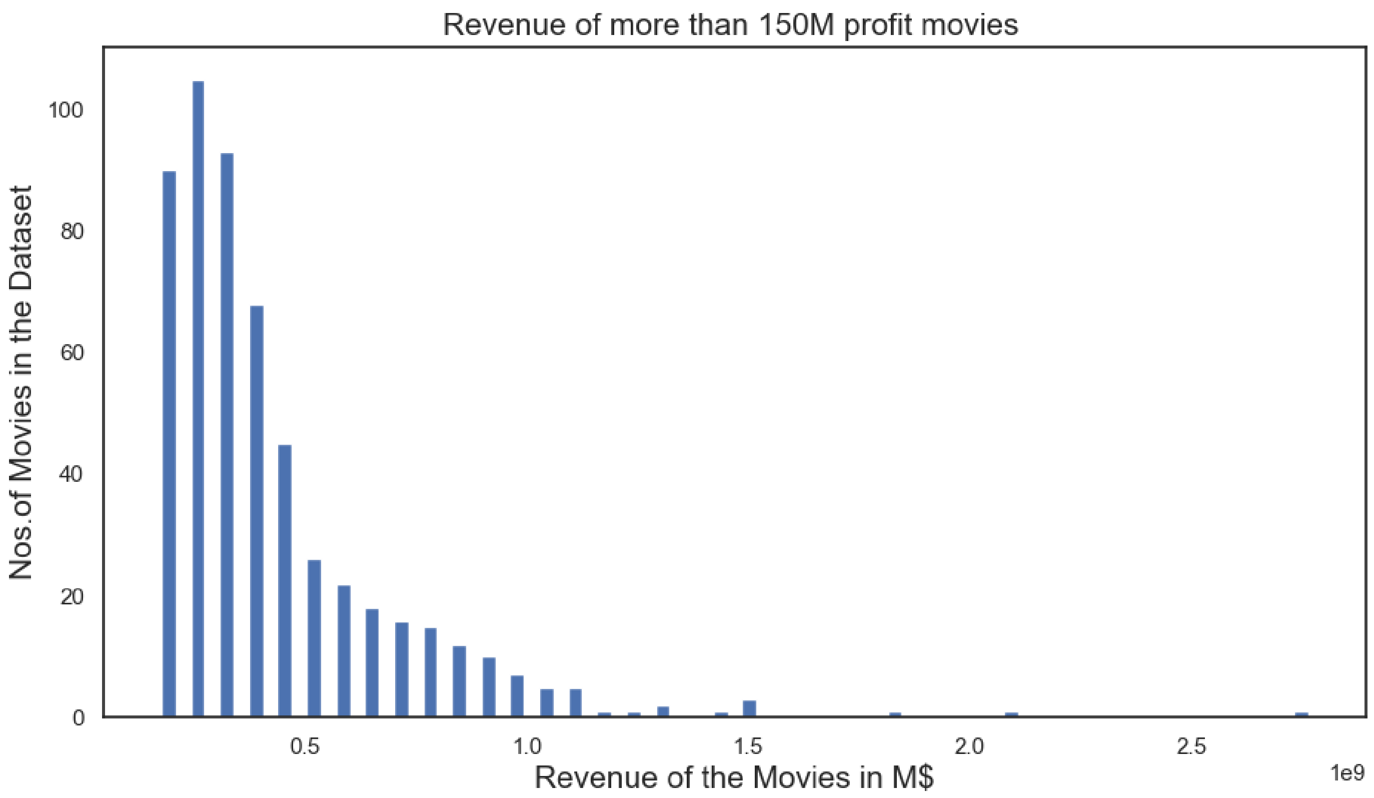
In [324]:

```
#plotting a histogram of revenue of movies

#giving the figure size(width, height)
plt.figure(figsize=(11,6), dpi = 100)

#On x-axis
plt.xlabel('Revenue of the Movies in M$', fontsize = 15)
#On y-axis
plt.ylabel('Nos.of Movies in the Dataset', fontsize=15)
#Name of the graph
plt.title('Revenue of more than 150M profit movies ', fontsize=15)

#giving a histogram plot
plt.hist(tmdb_profit_data1['revenue'], rwidth = 0.5, bins =40)
#displays the plot
plt.show()
```

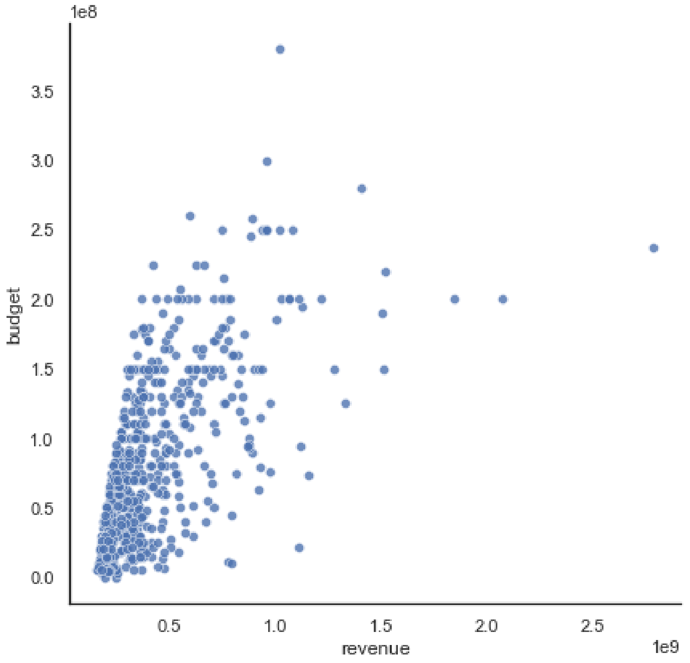


So the average revenue of the movies is 432591643.9 Dollars more than 90% of movies have revenue more than 250M

let's plot a relationship between budget and revenue

In [250]:

```
#let's polt a relationship between budget and revenue
sns.set_theme(style="white")
# Plot miles per gallon against horsepower with other semantics
sns.relplot(x="revenue", y="budget",
            sizes=(40, 400), alpha=.8, palette="muted",
            height=6,data=tmdb_profit_data1,facet_kws=dict(sharex=False))
plt.show()
```



90% of the movies has budget less than 100M and revenue more than 250M

In [325]:

```
#calculating thr ratio between profit and No of movies in the category
SR3=(tmdb_profit_data1['profit'].mean()/len(tmdb_profit_data))
print(SR3)
```

```
764569.2577907221
```

each movie in this cat. has mean profit = 764569.3 dollar much better than the previous cat.

What is the average runtime of the movie w.r.t Profit of movies making more then 150M Dollars?

In [326]:

```
# Finfd the average runtime of movies which made profit more then 50M Dollars
tmdb_profit_data1['runtime'].mean()
```

Out[326]:

```
117.43248175182482
```

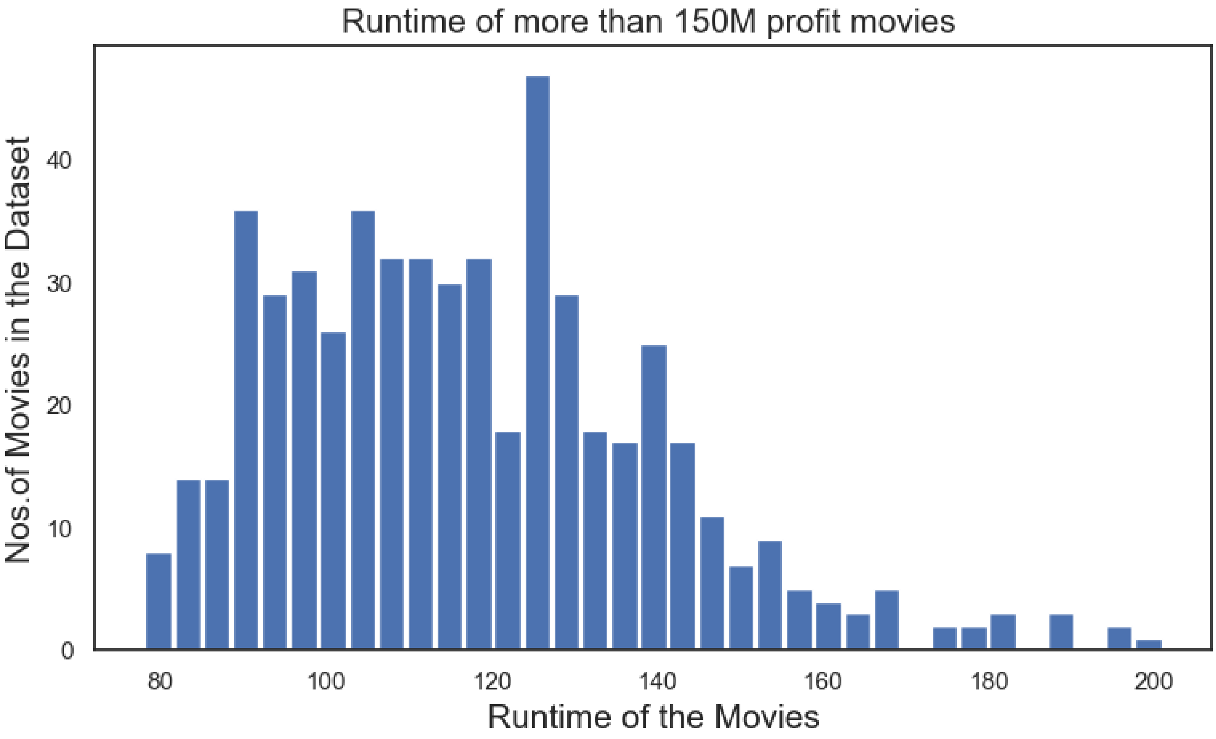
In [327]:

```
#plotting a histogram of runtime of movies

#giving the figure size(width, height)
plt.figure(figsize=(9,5), dpi = 100)

#On x-axis
plt.xlabel('Runtime of the Movies', fontsize = 15)
#On y-axis
plt.ylabel('Nos.of Movies in the Dataset', fontsize=15)
#Name of the graph
plt.title('Runtime of more than 150M profit movies ', fontsize=15)

#giving a histogram plot
plt.hist(tmdb_profit_data1['runtime'], rwidth = 0.9, bins =35)
#displays the plot
plt.show()
```

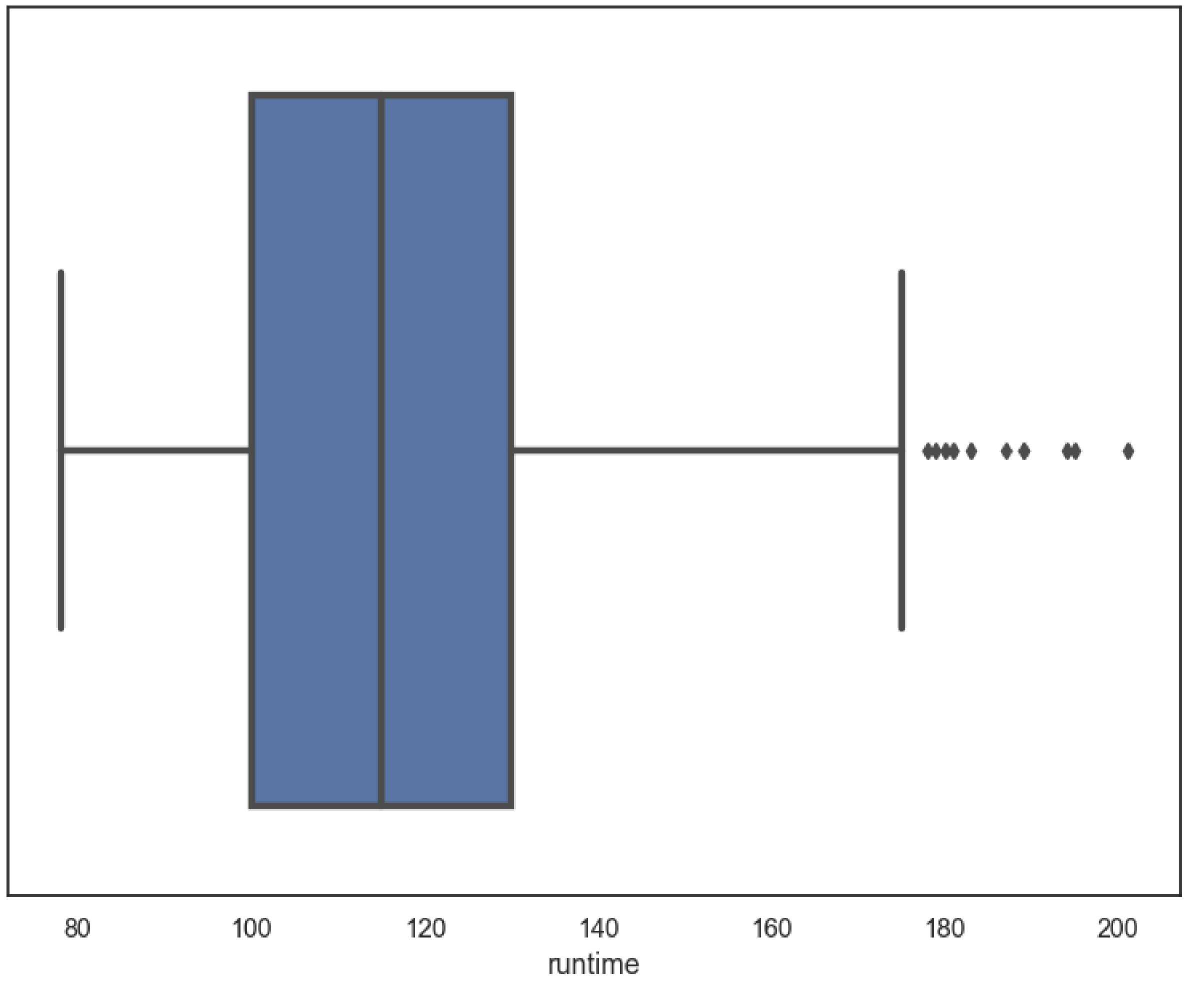


In [254]:

```
#The First plot is box plot of the runtime of the movies
plt.figure(figsize=(9,7), dpi = 105)

#using seaborn to generate the boxplot
sns.boxplot(tmdb_profit_data1['runtime'], linewidth = 3)
#diplaying the plot
plt.show()
```

```
C:\Users\Mustafa\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.9_qbz5n2kfra8p0\LocalCache\loca
packages\Python39\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variable as
a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other
arguments without an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```



So the average runtime of the movies is 117.4 Minutes and it is lying between 100-125 minutes

Which are the successfull genres w.r.t Profit of movies making more then 150M Dollars?

In [328]:

```
# This will first concat all the data with | from the whole column and then split it using | and count t
genres_count = pd.Series(tmdb_profit_data1['genres'].str.cat(sep = '|').split('|')).value_counts(ascendin
genres_count
```

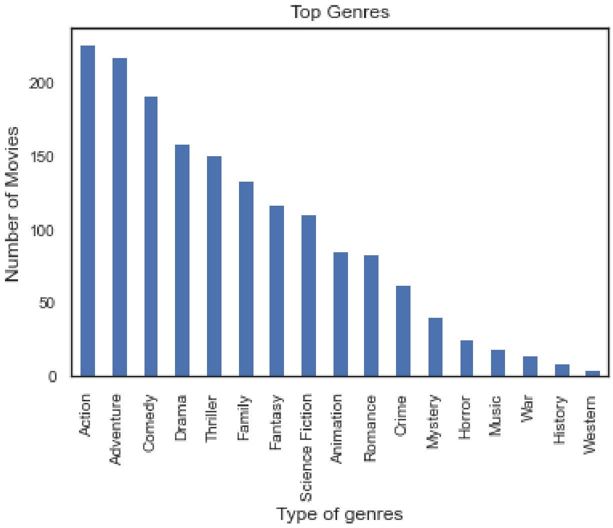
Out[328]:

```
Action             226
Adventure          218
Comedy             191
Drama              159
Thriller           151
Family             133
Fantasy            117
Science Fiction    111
Animation           85
Romance             83
Crime               63
Mystery             41
Horror              25
Music               19
War                 15
History              9
Western              5
dtype: int64
```

So the Top 10 Genres are Action,Adventure,Comedy,Drama,Thriller,Family,Fantasy,Science Fiction,Animation,Romance Lets visualize this with a plot

In [329]:

```
# Initialize the plot
diagram = genres_count.plot.bar(fontsize = 10)
# Set a title
diagram.set(title = 'Top Genres')
# x-label and y-label
diagram.set_xlabel('Type of genres')
diagram.set_ylabel('Number of Movies')
# Show the plot
plt.show()
```



We can clearly see in the visualization that most movies has Action as a genre which tends to higher profit

Which are the most frequent cast involved w.r.t Profit of movies making more then 150M Dollars?

In [330]:

```
# This will first concat all the data with | from the whole column and then split it using | and count t
cast_count = pd.Series(tmdb_profit_data1['cast'].str.cat(sep = '|').split('|')).value_counts(ascending =
cast_count.head(20)
```

Out[330]:

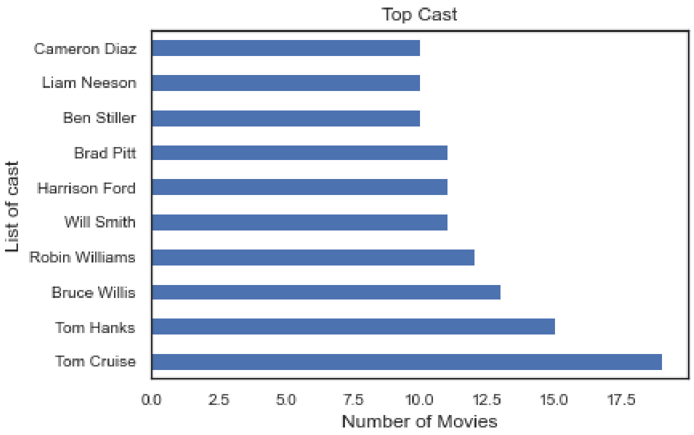
```
Tom Cruise          19
Tom Hanks           15
Bruce Willis        13
Robin Williams      12
Will Smith          11
Harrison Ford       11
Brad Pitt           11
Ben Stiller         10
Liam Neeson         10
Cameron Diaz        10
Julia Roberts       10
Gary Oldman         10
Anne Hathaway       10
Matt Damon          10
Samuel L. Jackson   10
Emma Watson          9
Jim Carrey           9
Leonardo DiCaprio    9
Ralph Fiennes        9
Angelina Jolie       9
dtype: int64
```

So the Top 5 cast are Tom Cruise,Tom Hanks,Bruce Willis,Robin Williams,Brad Pitt
Lets visualize this with a plot

In [331]:

```
# Initialize the plot
diagram = cast_count.head(10).plot.barh(fontsize = 10)
# Set a title
diagram.set(title = 'Top Cast')
# x-label and y-label
diagram.set_xlabel('Number of Movies')
diagram.set_ylabel('List of cast')
# Show the plot
```

```
plt.show()
```



We can clearly see in the visualization that most movies have Tom Cruise as a cast which tends to higher profit.

In [332]:

```
#the data about the ratio between profit and number of movies in each category is collected in the book
SR= pd.read_csv('book.csv')
SR.head()
```

Out[332]:

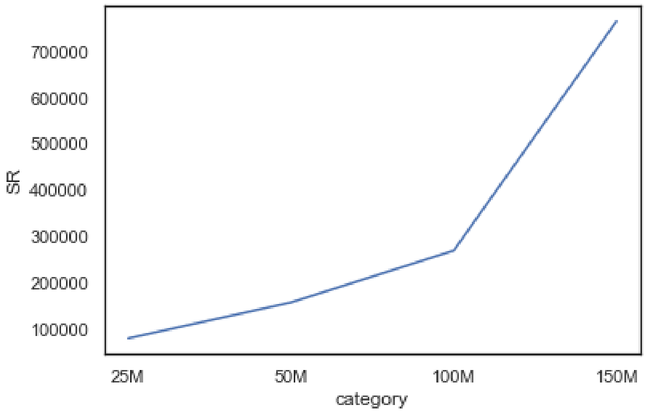
	category	SR
0	25M	80074.64752
1	50M	157495.60010
2	100M	269421.82990
3	150M	764569.25780

here we make a relationship which can be used to make assumption about relationship between the budget and revenue it as the follow .
SR=profit(revenue-budget)/number of movies in each category

next we ploting the relationship between the categories and SR

In [333]:

```
# plotting a relatinship to figure out some conclusion
sns.lineplot(data=SR,x="category",y="SR")
plt.show()
```



there is an expontional relationship between two variant

In [334]:

```
BR= pd.read_csv('Book1.csv')
BR.head()
```

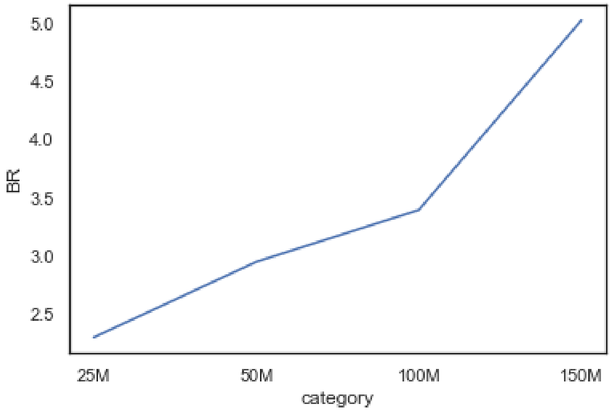
Out[334]:

```
   category        BR
0       25M  2.296296
1       50M  2.945946
2      100M  3.392157
3      150M  5.023256
```

Next is a relation between the ratio of budget and revenue (BR) and each responding cat.

In [335]:

```
# plotting a relatinship to figure out some conclusion
sns.lineplot(data=BR,x="category",y="BR")
plt.show()
```



ther is a direct relationship between both variants

conclusion

1.) 1.what are the best genres of movies constantly?
A1) Drama , Comedy , Action , Thriller and Adventure
2.) what the best cast for the different categories ? A2)for the general categories Robert De Nero ,Bruce Willis and Matt Demain
for the high profittable movies Tom cruis
3.)is runtime varies according to the cat.?
A3) no , it is seemly constant (109-117) minutes
4.) is there a relation between the budget and revenue (BR) and the categories of movies ?
A4)there is an direct relationship , so for better revenue movie should has a good budget
5.)is there a relation between avg. of the profit (SR) and the categories ?
A5) there is an expontial relationship , so with increasing in revenue ther is much increasing in the profit **tip for the movie makers
there is a increamental relationship between the budget and the revenue rate , so the make a profitable movie the budget should not be less than 60M

limitations

1) I relied on movies which revenue in the range more than(25,50,100,150)M dollars and it is not inclusive
2) Data may not be up to date and it affects
3) there are different currency in the revenue and budget columns it differs according to the production country
4) there were a drop in rows which contains missing values