DATA ANALYST NANODEGREE

Project # 2

INVESTIGATE A DATASET

(TMDb Movies)

Himanshu Saini
Delhi-NCR, India

Overview :

The TMDb movie dataset provide many information on all movies. The data contains information that are provided from The Movie Database (TMDb). It collects 5000+ movies basic move information and movie matrices, including user ratings, popularity and revenue data. These metrics can be seen as how successful these movies are. The movie basic information contained like cast, director, keywords, runtime, genres, etc.

The primary goal of the project is to go through the general data analysis process — using basic data analysis technique with NumPy, pandas, and Matplotlib. It contains three parts:

1. Data Wrangling
2. Exploratory Data Analysis
3. Conclusion

In this report I am going to explore following questions :

Question 1: Which movies are the most profitable to the market?

Question 2: Which movie has the Least and maximum profit, budget, runtime?

Question 3: Top 10 movies by profit, budget, and runtime?

Question 4: Which years do movies made the most profits?

Question 5: Movie Release year's vs Total budget made by movies?

Question 6: No. of movies release in every month of a year?

Question 7: Find the top casts, directors and genres?

Question 8: What is the Average Budget of the movies?

Question 9: What is the Average Revenue earned by the movies?

Question 10: What is the Average duration of the movies?

1) **Data Wrangling**

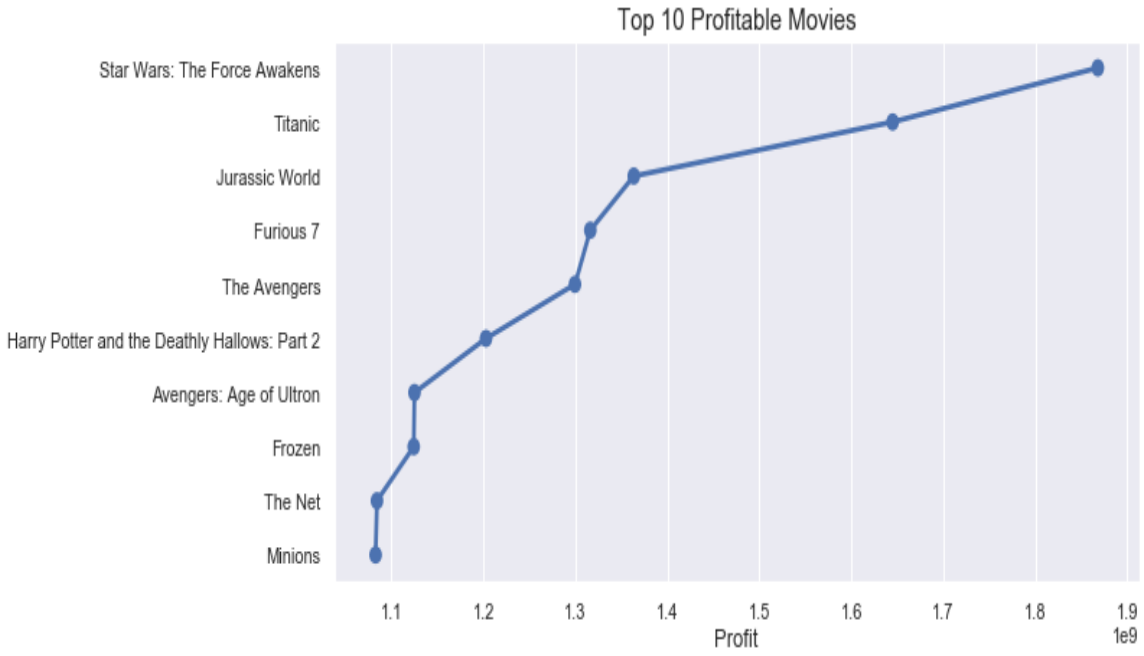
In this process the main idea is to take a quick glance on the data set, find the potential unreasonable data value, unnecessary variables for my research question, null data or duplicates, and then make data clearing decisions.

- Basic Exploration
- Delete the columns that is not required
- Null values and zero values
    - Zero Values in Budget and Revenue Columns
    - Zero Values in Runtime Columns
- Data Cleaning
    - Drop Duplicates
    - Replace zero values with null values

2) Exploratory data analysis:

Question 1. Which movies are the most profitable to the market?

⇨ The most profitable movies in the market is **"Star Wars : The Force Awakens"**



Question 2: Which movie has the Least and maximum profit, budget, runtime?

⇨ Least and maximum Budget:

	2244	2618
popularity	0.25054	0.090186
budget	4.25e+08	1
revenue	11087569	100
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
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	-4.13912e+08	99

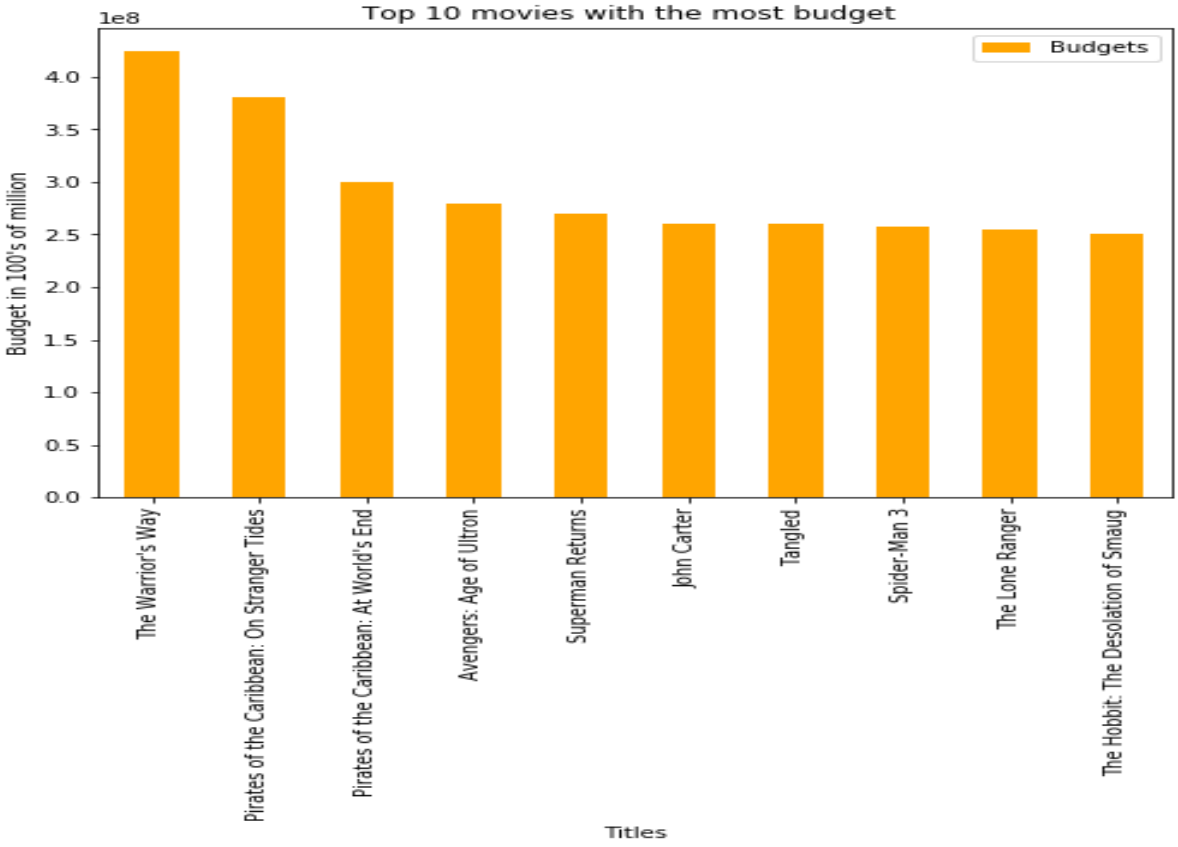
⇨ Least and maximum profit :

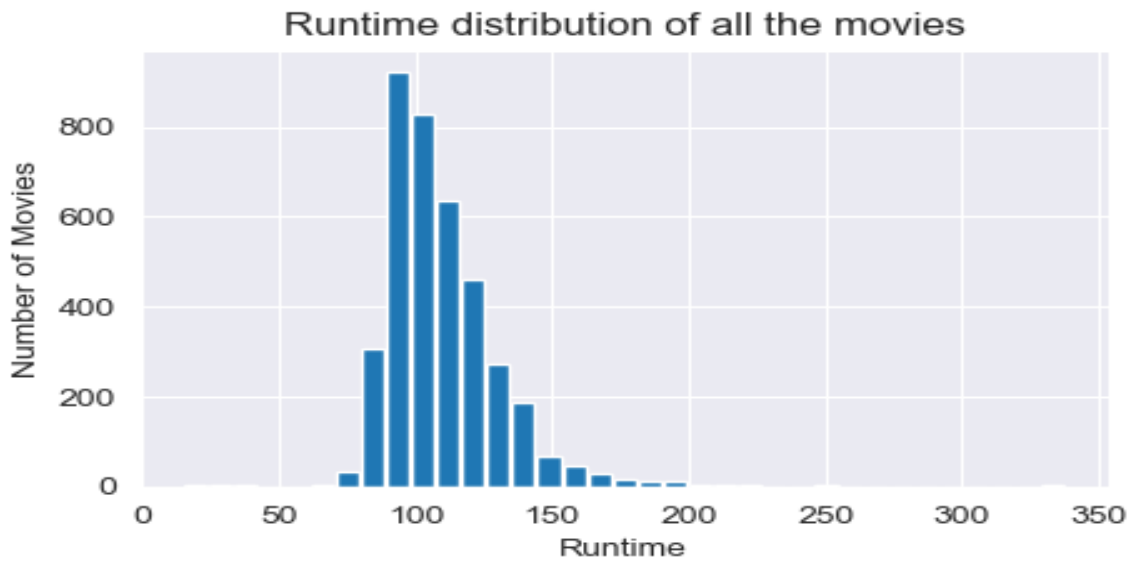
	3	1386
popularity	11.1731	9.43277
budget	2e+08	2.37e+08
revenue	2068178225	-1513461449
original_title	Star Wars: The Force Awakens	Avatar
cast	Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...	Sam Worthington|Zoe Saldana|Sigourney Weaver|S...
director	J.J. Abrams	James Cameron
runtime	136	162
genres	Action|Adventure|Science Fiction|Fantasy	Action|Adventure|Fantasy|Science Fiction
release_date	2015-12-15 00:00:00	2009-12-10 00:00:00
release_year	2015	2009
profit	1.86818e+09	-1.75046e+09

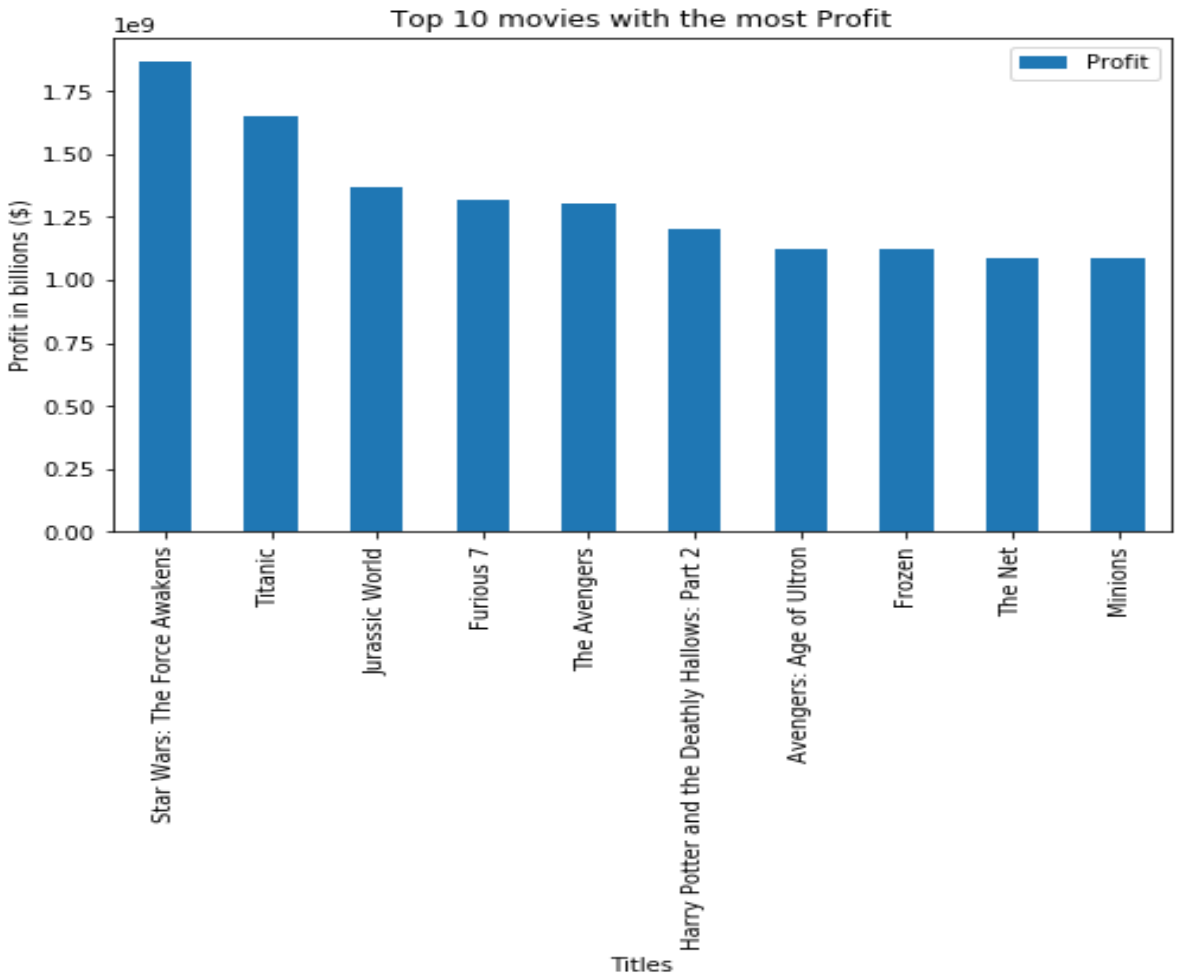
⇨ Least and maximum runtime:

	2107	5162
popularity	0.534192	0.208637
budget	1.8e+07	10
revenue	871279	5
original_title	Carlos	Kid's Story
cast	Edgar RamÃrez|Alexander Scheer|Fadi Abi Samra...	Clayton Watson|Keanu Reeves|Carrie-Anne Moss|K...
director	Olivier Assayas	Shinichiro Watanabe
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	-1.71287e+07	-5

Question 3: Top 10 movies by profit , budget, runtime?

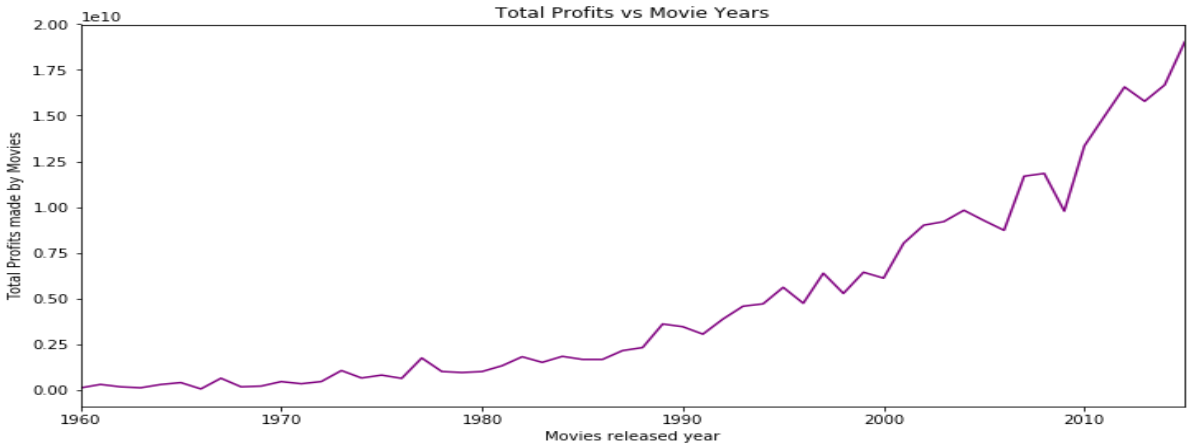



Runtime distribution of all the movies
Number of Movies
800
600
400
200
0
0
50
100
150
200
250
300
350
Runtime

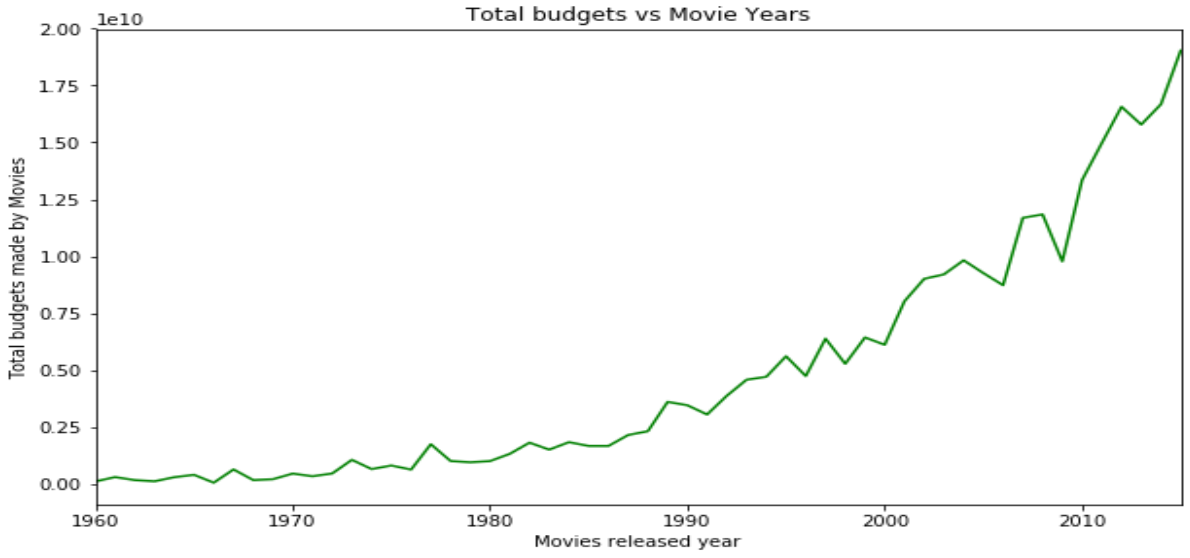

Top 10 movies with the most Profit
1e9
Profit
Profit in billions ($)
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Star Wars: The Force Awakens
Titanic
Jurassic World
Furious 7
The Avengers
Harry Potter and the Deathly Hallows: Part 2
Avengers: Age of Ultron
Frozen
The Net
Minions
Titles

Question 4: Which years do movies made the most profits ?

Release year	Profit
1960	108198052.0
1961	299083188.0
1962	166879846.0
.....	
2013	1.578274e+10
2014	1.667620e+10
2015	1.903215e+10



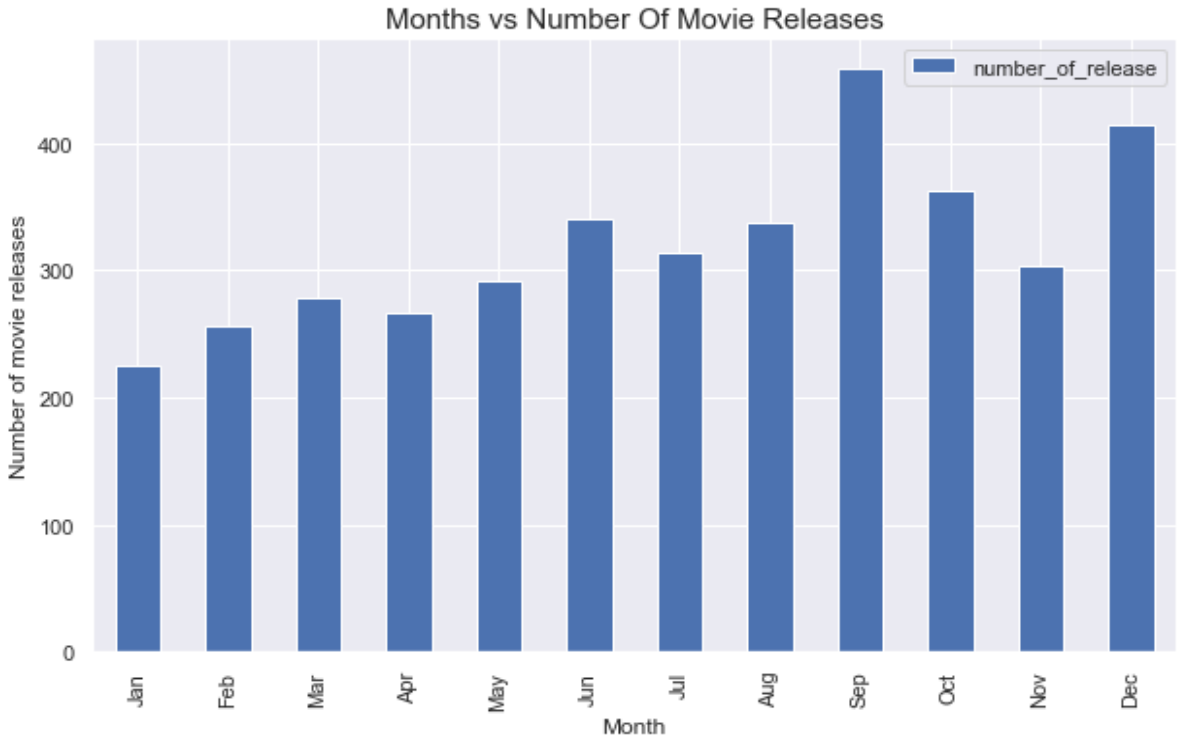
Question 5: Movie Release years vs Total budget made by movies ?



Question 6. No. of movies release in every month of a year?

⇨ Number of release Movies By month:

Jan	225		**Jul**	314
Feb	257		**Aug**	337
Mar	279		**Sept**	459
Apr	266		**Oct**	363
May	291		**Nov**	303
Jun	340		**Dec**	415



Question 7. Find the top casts, directors and genres?

Most Frequent Cast		Most successful Directors		Most Successful Genres	
Robert De Niro	52	Steven Spielberg	28	Drama	1753
Bruce Willis	46	Clint Eastwood	24	Comedy	1357
Samuel L. Jackson	44	Ridley Scott	21	Thriller	1203
Nicolas Cage	43	Woody Allen	18	Action	1085
Matt Damon	36	Martin Scorsese	17	Adventure	749

Question 8: What is the Average Budget of the movies

- ⇨ Average budget of the movies is 60312904.6252 (60.62 Million $)

Question 9: What is the Average Revenue earned by the movies

- ⇨ Average Revenue earned by the movies is 253067948.17 (253.17 Million $)

Question 10: What is the Average duration of the movies

- ⇨ Average duration of the movies is 113.63 (1 hour 53 minutes)

3) Conclusions:

This was a very interesting data analysis. We came out with some very interesting facts about movies. After this analysis we can conclude following:

For a Movie to be in successful criteria

- Average Budget must be around 60 million dollar
- Average duration of the movie must be 113 minutes (1 hr 53min)
- Profits: profits has positive relationship with budget and popularity
- Any one of these should be in the cast: Robert De Niro , Bruce Willis, Sylvester Stallone, Samuel L. Jackson
- Genre must be: Comedy, Drama. Action, Thriller.
- Director must be: Steven Spielberg, Clint Eastwood, Ridley Scott, and Woody Allen.

Limitation:

Although we successfully predicted the above properties on TMDb movie dataset, there are many infomation removed such as rows contained 0 values and null values. The dataset was cut by few thousand rows of movies, which would definitly affect the result.