

trainity

PROJECT 5

BY ASHISH KUMAR SAMANTARAY





Ashish Kumar Samantaray

B.Tech, Computer Science and Engineering



ashish.kumar.samantaray2003@gmail.com



7205691104

HYPERLINK

OF EXCEL SHEET

[Click here to get the working file](#)

PROJECT DESCRIPTION:

The dataset provided is related to **IMDB** Movies. A potential problem to investigate could be: "What factors influence the success of a movie on IMDB?" Here, success can be defined by high **IMDB ratings**. The impact of this problem is significant for movie producers, directors, and investors who want to understand what makes a movie successful to make informed decisions in their future projects.

Our task is to create a report that tells a story with your data. This should include your initial problem, your findings, and the **insights** you've gained. Use visualizations to help tell your story and make your findings more understandable.

TECH STACK USED:

MICROSOFT EXCEL

CANVA FOR CREATING PPT

I chose **Microsoft Excel** because it is thw most convenient spreadhseet and can be used efficiently to view statistics and analyse the data set given very quickly.

I chose **Canva** so as to make my PPT look more visually appealing.

MADE IN
Canva



insights AHEAD

WITH DETAILED APPROACH AND OUTPUT AND FORMULA BOX
(GRAPH IF ASKED)

TASK A

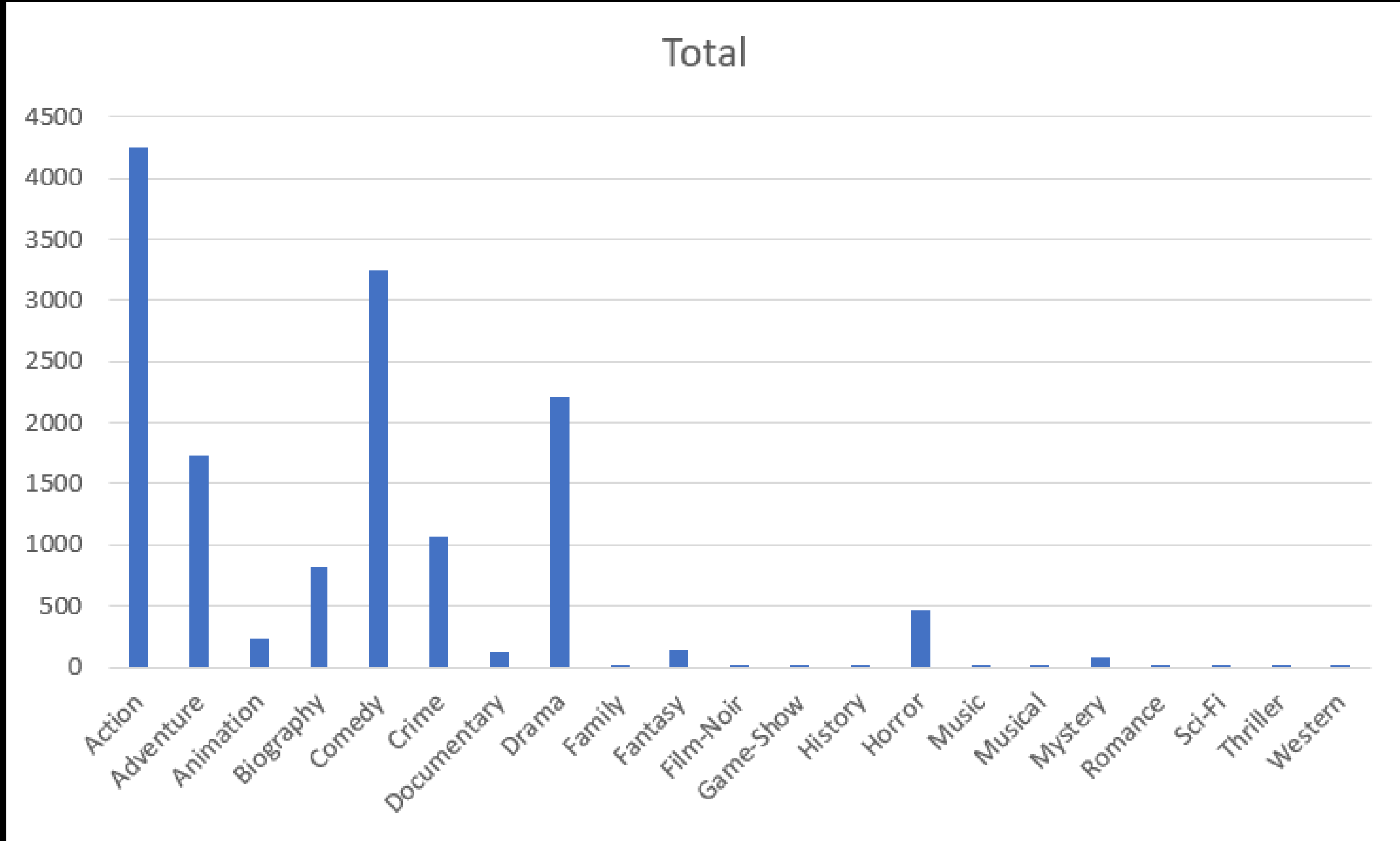
Movie Genre Analysis



[illegible]

TASK A

Graph:



TASK A

FORMULA FOR MEDIAN:(ALL OTHERS DONE THE SAME WAY)

Action =MEDIAN(B5:B74)

Descriptive Analysis:

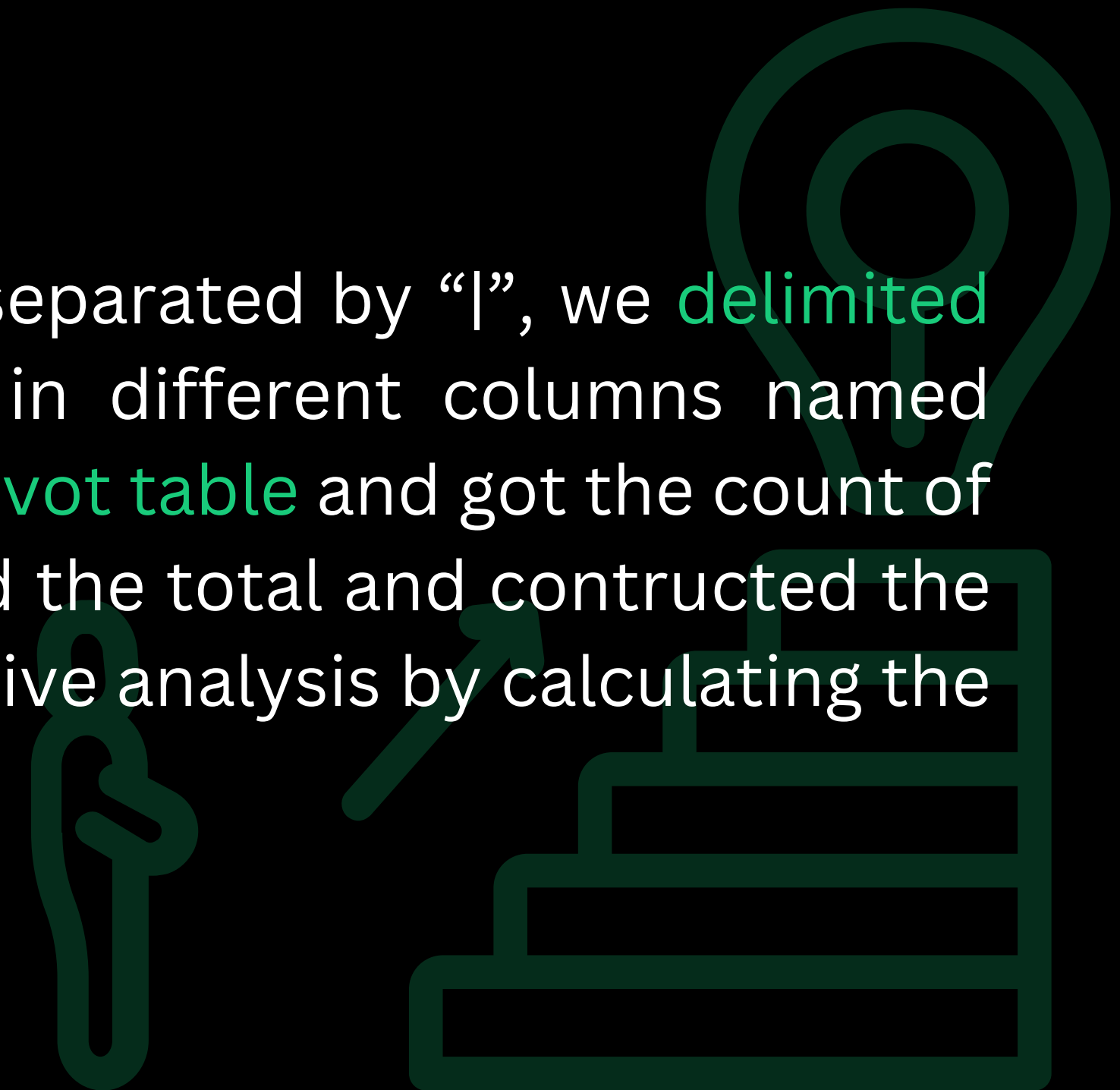
genres	Average of imdb_score	StdDev of imdb_score	Var of imdb_score	Max of imdb_score	Min of imdb_score	Sum of imdb_score	Range
Action	6.239895924	1.118349975	1.250706667	9.1	1.7	7194.6	7.4
Adventure	6.525165563	1.117879095	1.24965367	8.6	2.3	2955.9	6.3
Animation	6.631147541	1.177875412	1.387390486	8.4	3.7	404.5	4.7
Biography	7.159126984	0.691986466	0.47884527	8.9	4.5	1804.1	4.4
Comedy	6.198194131	1.08706592	1.181712314	9.5	1.9	8237.4	7.6
Crime	6.906876791	0.975257656	0.951127495	9.3	3.1	2410.5	6.2
Documentary	7.167857143	1.180310766	1.393133503	8.7	1.6	602.1	7.1
Drama	6.76718107	0.996518154	0.993048432	9.1	2	6577.7	7.1
Family	5.709090909	1.875901987	3.519008264	8.6	2.8	62.8	5.8
Fantasy	6.445283019	0.890741908	0.793421146	7.9	4.3	341.6	3.6
Film-Noir	7.6	0	0	7.6	7.6	7.6	0
Game-Show	2.9	0	0	2.9	2.9	2.9	0
History	7.5	0	0	7.5	7.5	7.5	0
Horror	5.671551724	1.127080913	1.270311385	8.5	2.2	1315.8	6.3
Music	7.2	0	0	7.2	7.2	7.2	0
Musical	6	1.541103501	2.375	7.2	3.4	24	3.8
Mystery	6.551515152	1.114945633	1.243103765	8.5	3.2	216.2	5.3
Romance	5.883333333	0.794599829	0.631388889	7.1	5.1	35.3	2
Sci-Fi	6	1.428824153	2.041538462	8.2	2.8	78	5.4
Thriller	5.590909091	1.270615131	1.61446281	8.1	3.4	123	4.7
Western	6.583333333	1.531792704	2.346388889	8.9	3.8	79	5.1

Count of genres		
genres	Total	Median
Action	1153	5.65
Adventure	453	6
Animation	61	6.75
Biography	252	6.95
Comedy	1329	5.45
Crime	349	6.65
Documentary	84	7.15
Drama	972	5.95
Family	11	5.8
Fantasy	53	6.35
Film-Noir	1	7.6
Game-Show	1	2.9
History	1	7.5
Horror	232	5.6
Music	1	7.2
Musical	4	6.7
Mystery	33	6.7
Romance	6	6.2
Sci-Fi	13	6.2
Thriller	22	5.6
Western	12	6.95
Grand Total	5043	

TASK A

APPROACH:

Since, many movies had more than **one genre** separated by “|”, we **delimited** the cells and separated the multiple genres in different columns named **genre, genre2 and so on**. Then, we created the **pivot table** and got the count of all the genre fields till genre8. Then we counted the total and constructed the **bar graph** for the same. We also did the descriptive analysis by calculating the mean, max, min, variance and so on.



TASK B

Movie Duration Analysis



TASK B

Output(With Descriptive Analysis):

genres	Sum of duration	StdDev of duration	Var of duration	Average of duration	Average of imdb_score
Biography	31596	30.29638177	917.8707483	125.3809524	7.159126984
Drama	109678	26.5093026	702.7431244	112.8374486	6.76718107
Crime	38891	36.31175895	1318.543838	111.4355301	6.906876791
Action	127634	23.85823557	569.2154047	110.6973114	6.239895924
Mystery	3549	18.67763354	348.8539945	107.5454545	6.551515152
Adventure	48344	27.92165779	779.6189738	106.7196468	6.525165563
Western	1272	20.61148547	424.8333333	106	6.583333333
Thriller	2285	19.37499667	375.3904959	103.8636364	5.590909091
Musical	415	4.145780988	17.1875	103.75	6
Sci-Fi	1301	14.10652625	198.9940828	100.0769231	6
Fantasy	5257	22.05315757	486.3417586	99.18867925	6.445283019
Comedy	131820	18.26908534	333.759479	99.18735892	6.198194131
Horror	22472	14.65075998	214.6447681	96.86206897	5.671551724
Family	1061	11.41262169	130.2479339	96.45454545	5.709090909
Film-Noir	95	0	0	95	7.6
Music	93	0	0	93	7.2
Documentary	7736	26.42041916	698.0385488	92.0952381	7.167857143
History	90	0	0	90	7.5
Animation	5016	21.5379143	463.8817522	82.2295082	6.631147541
Game-Show	60	0	0	60	2.9
Romance	342	56.88585061	3236	57	5.883333333

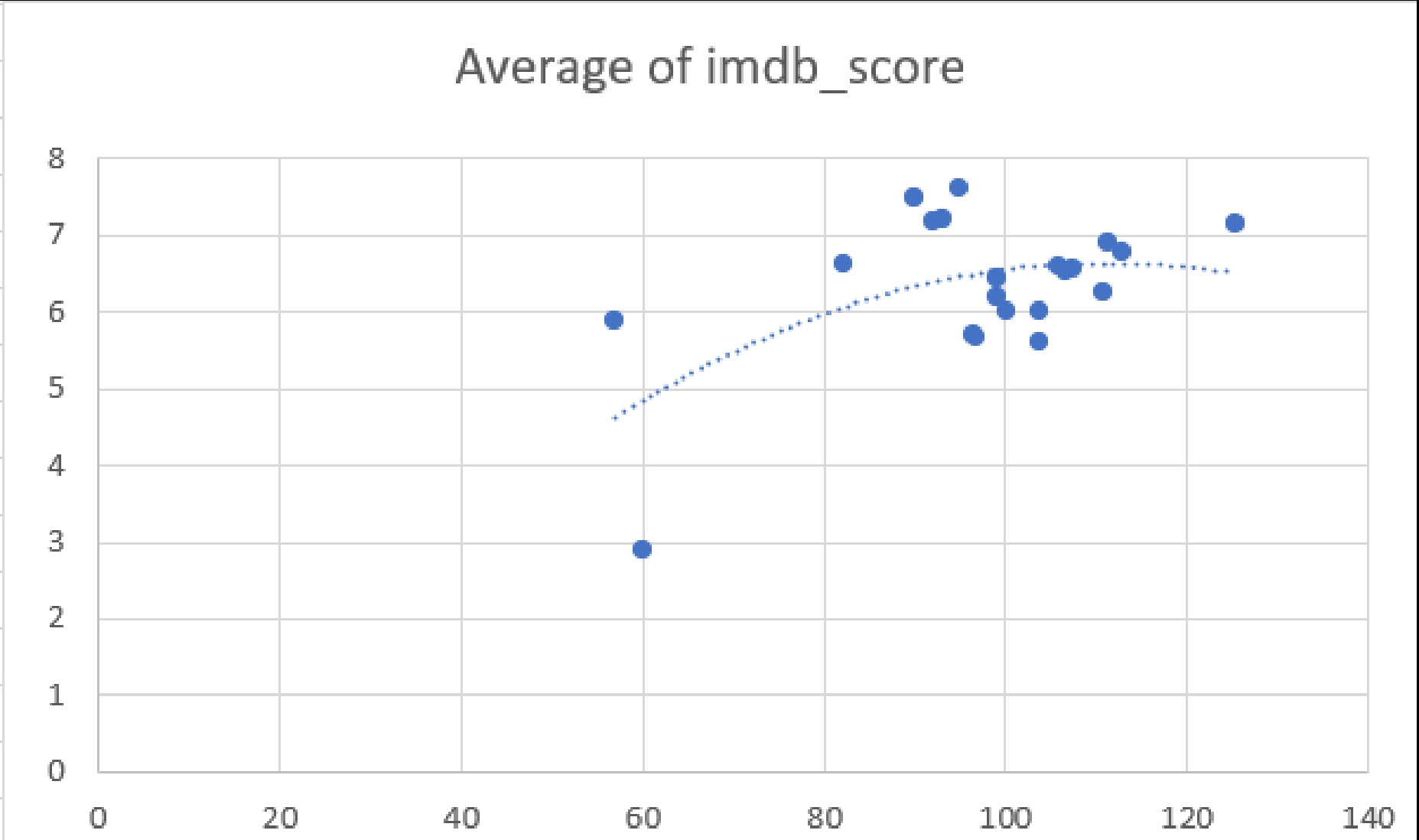
TASK B

Output:

Total Duration	
genres	Total
Comedy	131820
Action	127634
Drama	109678
Adventure	48344
Crime	38891
Biography	31596
Horror	22472
Documentary	7736
Fantasy	5257
Animation	5016
Mystery	3549
Thriller	2285
Sci-Fi	1301
Western	1272
Family	1061
Musical	415
Romance	342
Film-Noir	95
Music	93
History	90
Game-Show	60
Grand Total	539007

Average of duration	Average of imdb_score
125.3809524	7.159126984
112.8374486	6.76718107
111.4355301	6.906876791
110.6973114	6.239895924
107.5454545	6.551515152
106.7196468	6.525165563
106	6.583333333
103.8636364	5.590909091
103.75	6
100.0769231	6
99.18867925	6.445283019
99.18735892	6.198194131
96.86206897	5.671551724
96.45454545	5.709090909
95	7.6
93	7.2
92.0952381	7.167857143
90	7.5
82.2295082	6.631147541
60	2.9
57	5.883333333

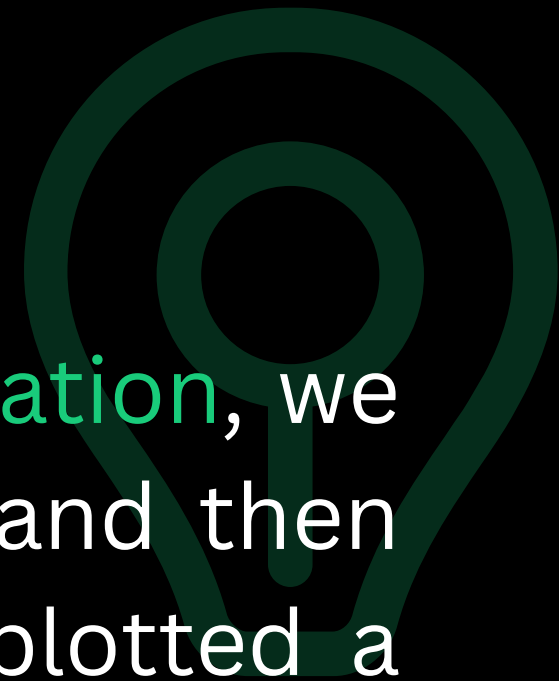
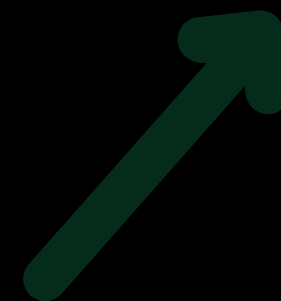
Graph:



TASK B

APPROACH:

Since we had to analyze the relation between **IMDB score** and **duration**, we first derived some values based on the imdb scores of the films and then related the average **IMDB score** with the average **duration** and plotted a **scattered graph along with the trend line** inserted.



TASK C

Language Analysis



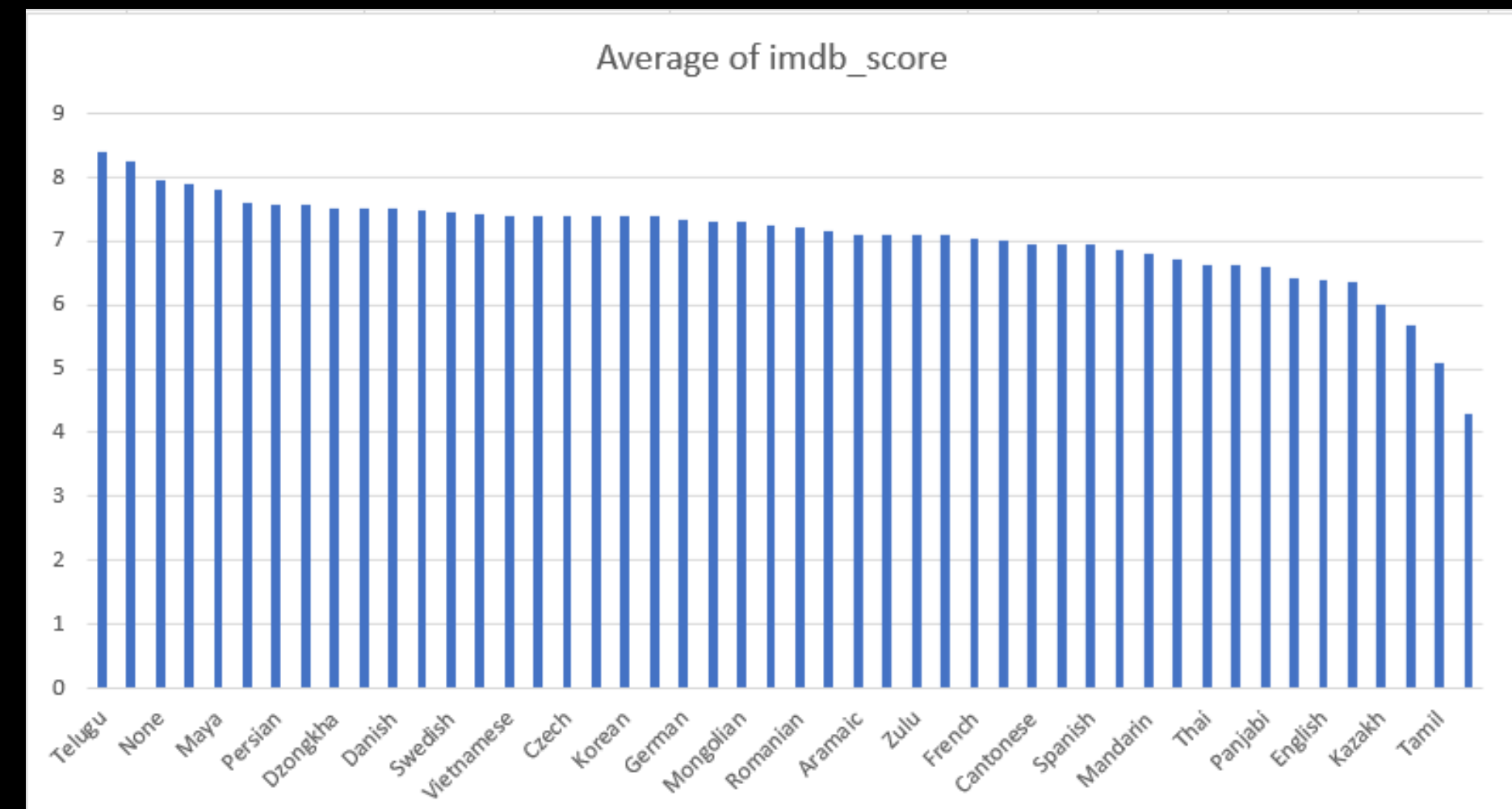
TASK C

Output:

language	Total
Telugu	8.4
Polish	8.25
None	7.95
Indonesian	7.9
Maya	7.8
Hebrew	7.58
Persian	7.575
Icelandic	7.55
Dzongkha	7.5
Dari	7.5
Danish	7.5
Portuguese	7.4875
Swedish	7.44
Dutch	7.425
Vietnamese	7.4
Swahili	7.4
Czech	7.4
Japanese	7.394444444
Korean	7.3875
Arabic	7.38
German	7.342105263
Greek	7.3
Mongolian	7.3
Italian	7.227272727
Romanian	7.2
Norwegian	7.15
Aramaic	7.1
Hungarian	7.1
Zulu	7.1

language	Total
English	4704
French	73
Spanish	40
Hindi	28
Mandarin	26
German	19
Japanese	18
Not Specified	12
Russian	11
Cantonese	11
Italian	11
Portuguese	8
Korean	8
Swedish	5
Danish	5
Hebrew	5
Arabic	5
Polish	4
Persian	4
Dutch	4
Norwegian	4
Thai	3
Chinese	3
Zulu	2
Romanian	2
Icelandic	2
Indonesian	2
Dari	2

Graph:



TASK C

Output:

language	Count of language	Average of duration
Aboriginal	2	111
Arabic	5	102.2
Aramaic	1	120
Bosnian	1	127
Cantonese	11	97.45454545
Chinese	3	117
Czech	1	113
Danish	5	96.2
Dari	2	105.5
Dutch	4	116.25
Dzongkha	1	108
English	4704	106.7841466
Filipino	1	132
French	73	103.9452055
German	19	128.8947368
Greek	1	94
Hebrew	5	95
Hindi	28	142.037037
Hungarian	1	134
Icelandic	2	296
Indonesian	2	99
Italian	11	109.1818182
Japanese	18	116.7777778
Kannada	1	#DIV/0!
Kazakh	1	112
Korean	8	127.375
Mandarin	26	112.0384615

Maya	1	139
Mongolian	1	126
None	2	101
Norwegian	4	95.75
Not Specified	12	90.63636364
Panjabi	1	141
Persian	4	102.25
Polish	4	58.25
Portuguese	8	113.625
Romanian	2	108.5
Russian	11	108.7272727
Slovenian	1	83
Spanish	40	106.575
Swahili	1	60
Swedish	5	141.2
Tamil	1	155
Telugu	1	159
Thai	3	173.6666667
Urdu	1	#DIV/0!
Vietnamese	1	135
Zulu	2	105

TASK C

Descriptive Analysis:

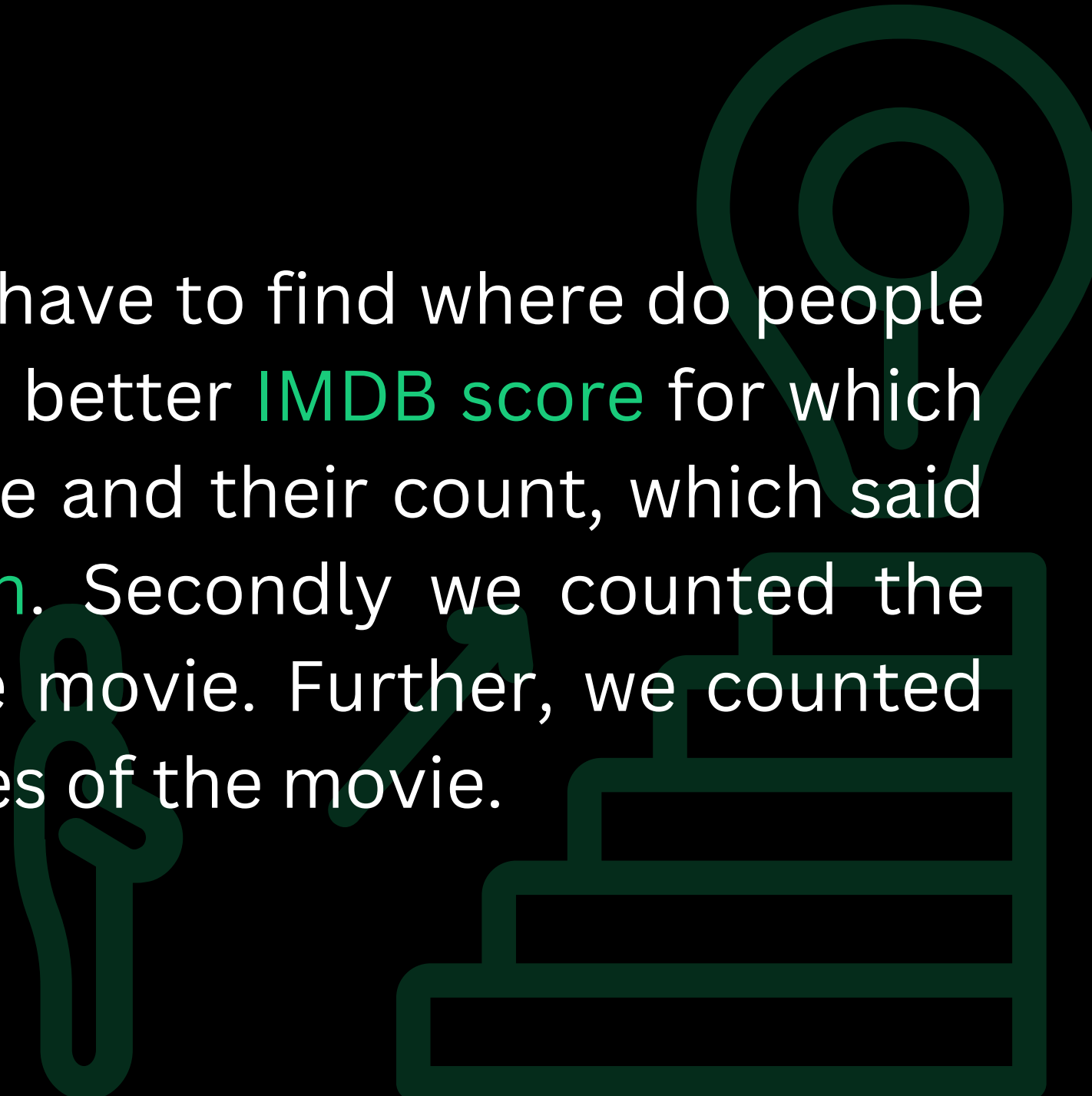
**For more languages refer to workbook and the Task C sheet

language	▼	Average of imdb_score	Average of duration	Min of imdb_score	Max of imdb_score	StdDev of imdb_score	Var of imdb_score
Telugu		8.4	159	8.4	8.4	0	0
Polish		8.25	58.25	7.4	9.1	0.85	0.7225
None		7.95	101	7.4	8.5	0.55	0.3025
Indonesian		7.9	99	7.6	8.2	0.3	0.09
Maya		7.8	139	7.8	7.8	0	0
Hebrew		7.58	95	7.2	8	0.299332591	0.0896
Persian		7.575	102.25	5.9	8.5	1.042532973	1.086875
Icelandic		7.55	296	6.9	8.2	0.65	0.4225
Dzongkha		7.5	108	7.5	7.5	0	0
Dari		7.5	105.5	7.4	7.6	0.1	0.01
Danish		7.5	96.2	5.7	8.3	0.963327566	0.928
Portuguese		7.4875	113.625	6.1	8.7	0.826797285	0.68359375
Swedish		7.44	141.2	6.6	8.2	0.677052435	0.4584
Dutch		7.425	116.25	7	7.8	0.376662979	0.141875
Vietnamese		7.4	135	7.4	7.4	0	0
Swahili		7.4	60	7.4	7.4	0	0
Czech		7.4	113	7.4	7.4	0	0
Japanese		7.394444444	116.7777778	5.6	8.7	0.962907762	0.927191358
Korean		7.3875	127.375	5.7	8.4	0.772071078	0.59609375
Arabic		7.38	102.2	6	8.2	0.790948797	0.6256
German		7.342105263	128.8947368	4.9	8.5	0.928675225	0.862437673
Greek		7.3	94	7.3	7.3	0	0
Mongolian		7.3	126	7.3	7.3	0	0
Italian		7.227272727	109.1818182	5.1	8.9	1.186354929	1.407438017
Romanian		7.2	108.5	6.5	7.9	0.7	0.49
Norwegian		7.15	95.75	6.4	7.6	0.497493719	0.2475
Aramaic		7.1	120	7.1	7.1	0	0
Hungarian		7.1	134	7.1	7.1	0	0
Zulu		7.1	105	6.9	7.3	0.2	0.04

TASK C

APPROACH:

Since we have to perform **language analysis**, we have to find where do people spend more time and which **language** movie has better **IMDB score** for which we created **pivot tables** and viewed the language and their count, which said that most of the movies were made in **english**. Secondly we counted the average duration spent in each language of the movie. Further, we counted the average **IMDB** score of the different languages of the movie.

A decorative graphic in the bottom right corner of the slide. It features a stylized green outline of a person standing on a set of stairs and climbing upwards. At the top of the stairs is a large green lightbulb, symbolizing an idea or a goal.

TASK D

Director Analysis



TASK D

Output:

Full Table is in The excel sheet***

director_name	Average of imdb_score	Sum of director_facebook_likes	Average of duration
Chatrichalerm Yukol	6.6	6	300
Ron Maxwell	7	66	275.5
Peter Flinth	6.6	5	270
Michael Cimino	7.5	1034	254
Joseph L. Mankiewicz	7	311	251
George Stevens	6.6	126	225
Michael Wadleigh	8.1	14	215
Stanley Kramer	7.95	352	191.5
David Lean	8	3068	188
Yash Chopra	7.4	294	184
Kevin Costner	7.166666667	0	184
Billy Bob Thornton	6.9	0	184
Edward Hall	7.2	0	180
Ken Annakin	7.8	19	178
Aleksey German	6.7	23	177
William Wyler	8.1	355	172
Paolo Sorrentino	7.7	667	172
John Sturges	8.3	120	172
Anthony Mann	6.7	75	172
Mervyn LeRoy	7.2	54	171
Peter Jackson	7.675	0	170.6666667
George Cukor	7.9	165	170
Christopher Spencer	5.6	25	170
Richard Attenborough	7.6	0	169.25

TASK D

Percentile Function Implementation:

Full Table is in The excel sheet***

director_name	Average of imdb_score	Sum of director_facebook_likes	Average of duration
John Blanchard	9.5	0	65
John Stockwell	9.1	134	90
Frank Darabont	8.9	0	165.5
Francis Ford Coppola	8.9	0	228
Sidney Lumet	8.9	0	96
Peter Jackson	8.8	0	178.3333333
Irvin Kershner	8.8	883	127
Lana Wachowski	8.7	0	136
Cary Bell	8.7	0	78
Mitchell Altieri	8.7	9	87
Sadyk Sher-Niyaz	8.7	135	135
David Fincher	8.7	42000	139
Fernando Meirelles	8.7	353	135
George Lucas	8.7	0	125
Akira Kurosawa	8.7	0	202
Robert Zemeckis	8.65	0	129
Sergio Leone	8.633333333	0	179.3333333
Charles Chaplin	8.6	0	87
Bryan Singer	8.6	0	106
Frank Capra	8.6	964	118
Christopher Nolan	8.6	154000	143.4285714
Michael Curtiz	8.6	345	82
Jonathan Demme	8.6	438	138
Mike Mayhall	8.6	14	88
Steven Spielberg	8.575	56000	149

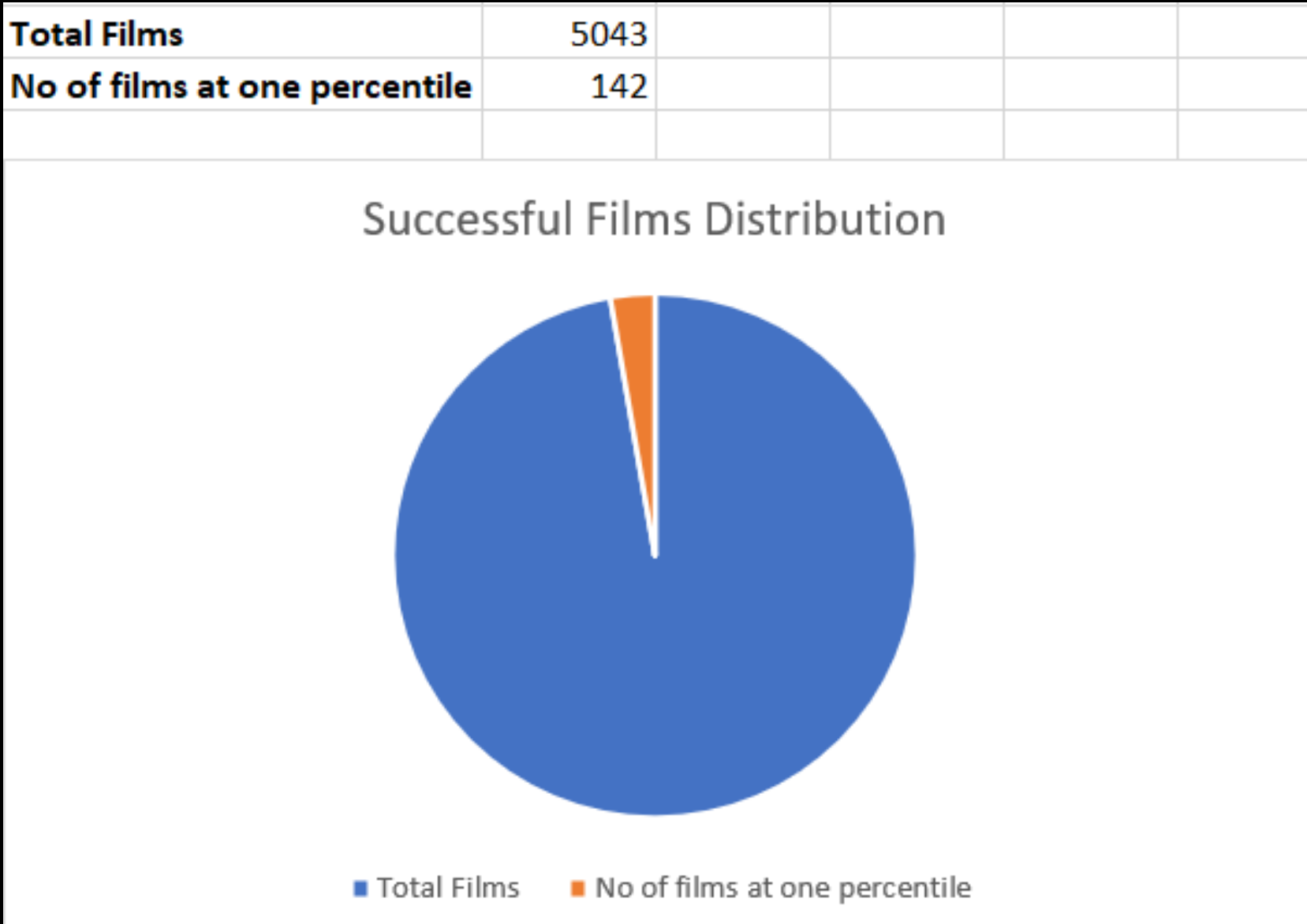
Top 1 Percentile Score for main table	=PERCENTILE(B5:B2403,0.99)		
Directors having Imdb averagerating >8.3 are top directors			

TASK D

Successful film proportion among all the films:

Total Films	=COUNTA(R2:R5044)	
No of films at one percentile	142	

Total Films	5043		
No of films at one percentile	=COUNTIF(Z2:Z5044,">=8.3")		



TASK D

APPROACH:

Since we have to analyse the **directors** and choose the **top rated directors**, we first have to arrange them according to their **average duration** watched by the audience. We can also compare them on the basis of their average **IMDB scores** and further filter them and sort them from largest to smallest or else we can see the popularity of the directors according to their total count of **facebook likes**. Further, **percentile** function was implemented so as to know the top 1 percentile directors among all the directors of the other film according to their average imdb score.

TASK E

Budget Analysis



TASK E

Correlation

0.142390456

**Data set is arranged according to profit from largest to smallest.

Output:

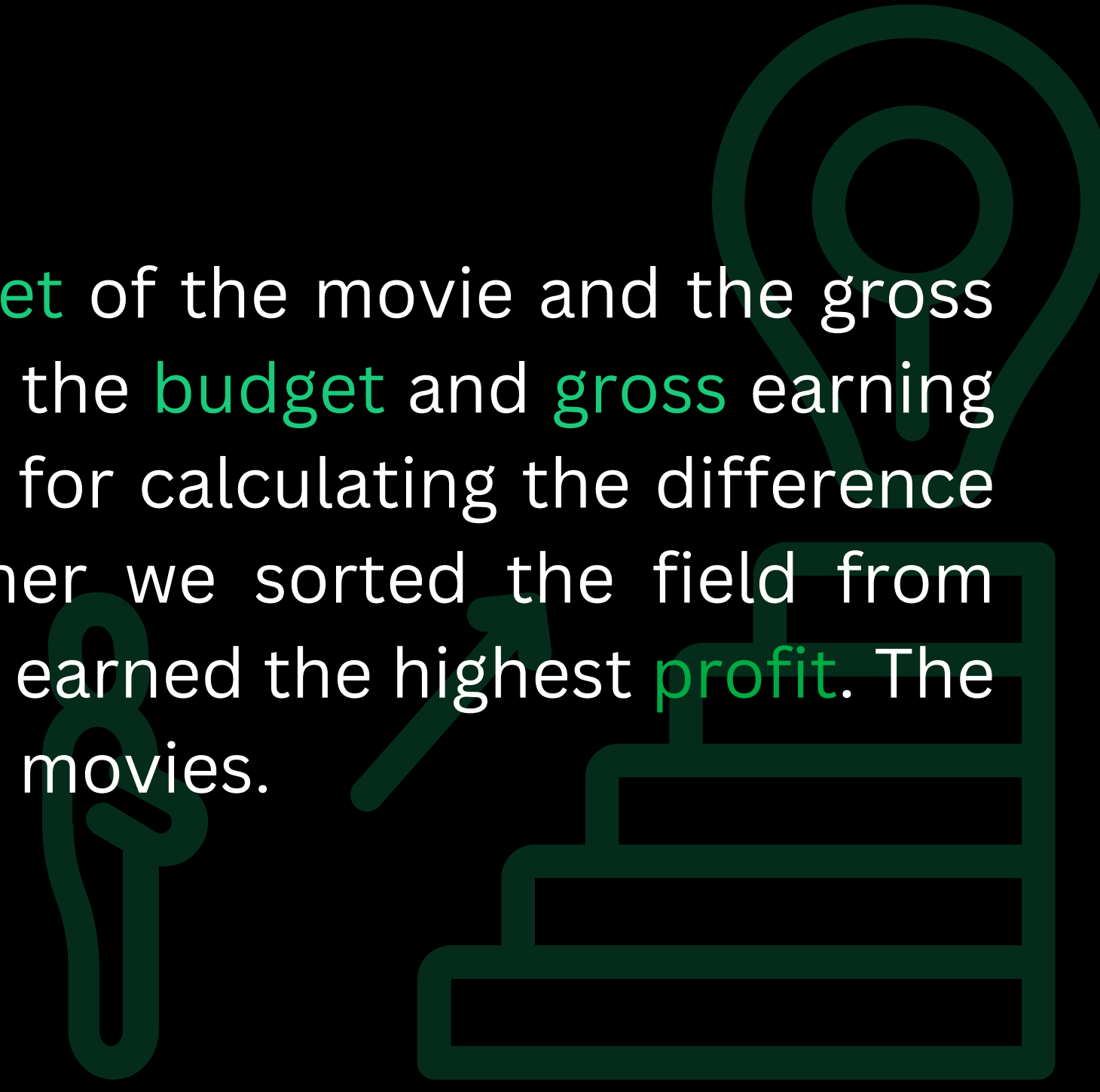
movie_title	Sum of gross	Sum of budget	Sum of Profit Margin
The Avengers	1246559094	440000000	806559094
Avatar	760505847	237000000	523505847
Jurassic World	652177271	150000000	502177271
Titanic	658672302	200000000	458672302
Star Wars: Episode IV - A New Hope	460935665	11000000	449935665
E.T. the Extra-Terrestrial	434949459	10500000	424449459
The Lion King	422783777	45000000	377783777
The Jungle Book	725290282	350000000	375290282
Star Wars: Episode I - The Phantom Menace	474544677	115000000	359544677
The Dark Knight	533316061	185000000	348316061
The Twilight Saga: Breaking Dawn - Part 2	584597846	240000000	344597846
The Hunger Games	407999255	78000000	329999255
The Fast and the Furious	433536930	114000000	319536930
Twilight	382898950	74000000	308898950
Deadpool	363024263	58000000	305024263
The Hunger Games: Catching Fire	424645577	130000000	294645577
Jurassic Park	356784000	63000000	293784000
Despicable Me 2	368049635	76000000	292049635
American Sniper	350123553	58800000	291323553
Finding Nemo	380838870	94000000	286838870
Shrek 2	436471036	150000000	286471036
The Lord of the Rings: The Return of the King	377019252	94000000	283019252
Star Wars: Episode VI - Return of the Jedi	309125409	32500000	276625409
Forrest Gump	329691196	55000000	274691196

The Avengers is the highest profitable movie along all.

TASK E

APPROACH:

In order to derive a relation between the **budget** of the movie and the gross earning we used the **correl** function to correlate the **budget** and **gross** earning of the movie. Also, we created a separate field for calculating the difference between the budget and gross earning. Further we sorted the field from largest to smallest so as to get which movie has earned the highest **profit**. The negative ones in this field denote the **loss** of the movies.

A decorative graphic in the bottom right corner. It features a stylized green outline of a person climbing a set of stairs. At the top of the stairs is a target symbol with concentric circles. An arrow points upwards from the person towards the target.



IMDb

RESULT

This project made me experience a real time data analysis of a thousand of movie and filter them on the basis of their properties which will be further used by the marketing team to push them to the audience. Following says my experience in an **IMDb** theme.

PROJECT QUALITY



“Fan-trainistic



OVERALL EXPERIENCE



“Loved It





Ashish Kumar Samantaray

B.Tech, Computer Science and Engineering



ashish.kumar.samantaray2003@gmail.com



7205691104