

NAME-AKSHAY ANAND

PROJECT DESCRIPTION: Dataset having various columns of different IMDB Movies.

APPROACH: WE USED MS EXCEL FOR DATA CLEANING AND FUNCTIONS TO CALCULATE REQUIRED PARAMETERS FROM THE DATASET PROVIDED.

TECH-STACK USED: MS-EXCEL

INSIGHTS: EXPLAINED WITH EACH PLOT.

Cleaning the data:: This is one of the most important step to perform before moving forward with the analysis. Use your knowledge learned till now to do this. (Dropping columns, removing null values, etc.)

Your task: Clean the data

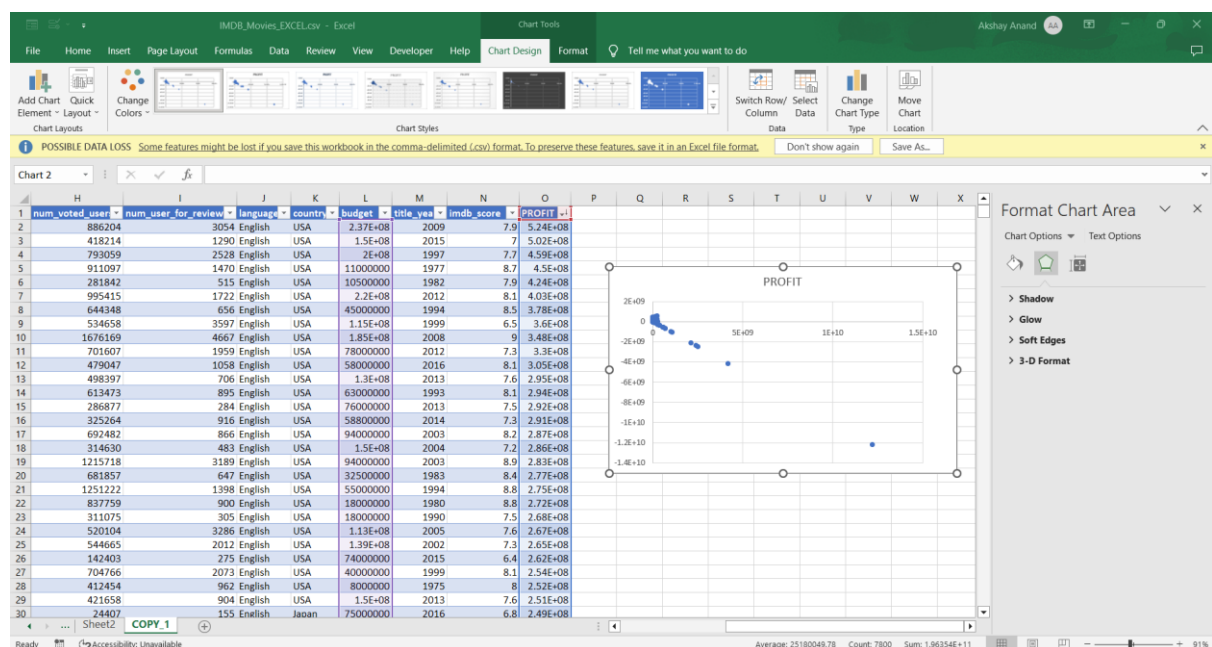
Columns like 'Color', 'director_facebook_likes', 'actor_3_facebook_likes', 'actor_2_name', 'actor_1_facebook_likes', 'cast_total_facebook_likes', 'actor_3_name', 'facenumber_in_posts', 'plot_keywords', 'movie_imdb_link', 'content_rating', 'actor_2_facebook_likes', 'aspect_ratio', 'movie_facebook_likes' are the columns which have been dropped. Also dropped rows having null values. At the end we got rid of duplicate values using Remove Duplicates in data tab.

Movies with highest profit: Create a new column called profit which contains the difference of the two columns: gross and budget. Sort the column using the profit column as reference.

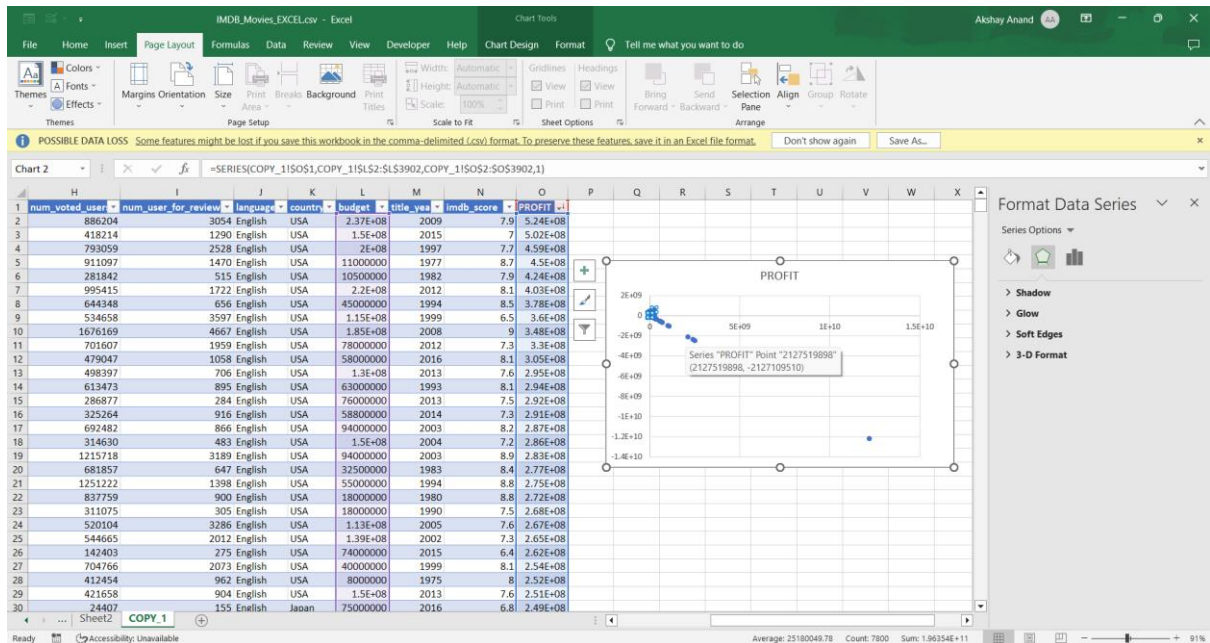
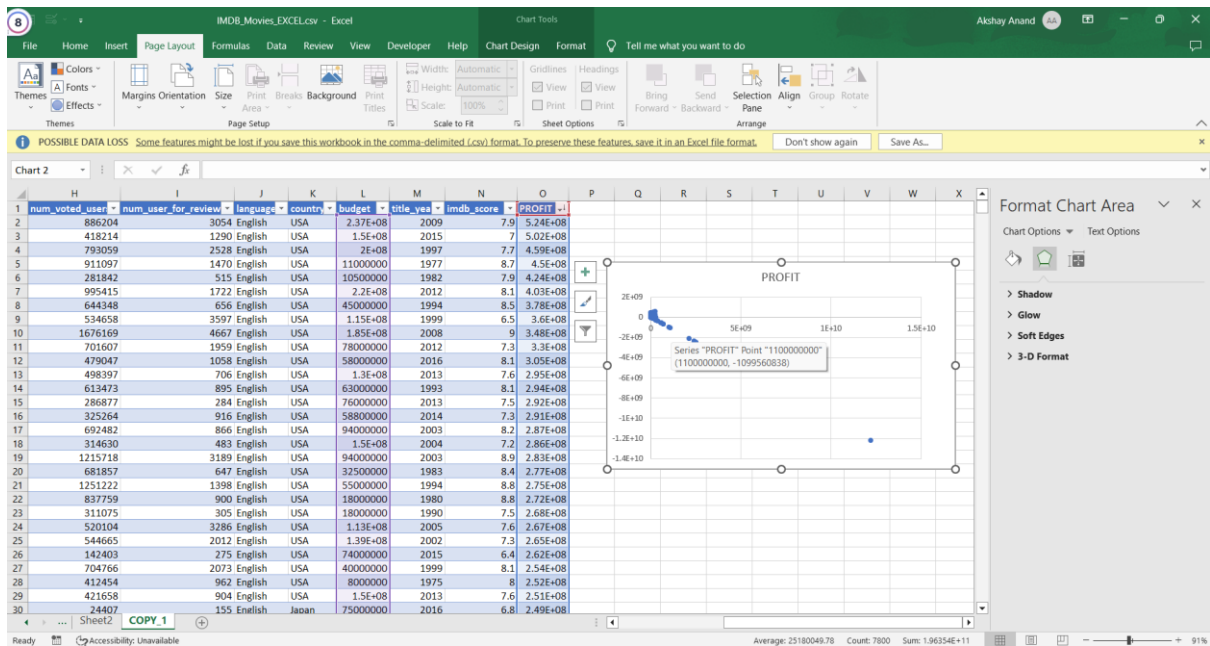
Plot profit (y-axis) vs budget (x- axis) and observe the outliers using the appropriate chart type.

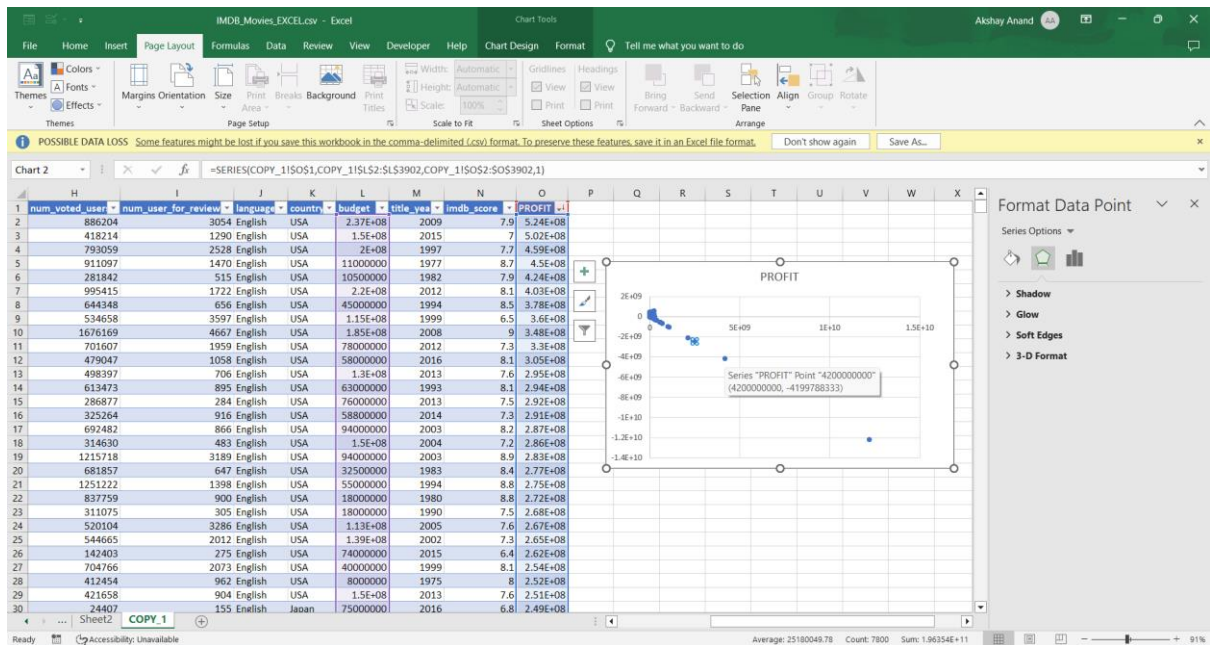
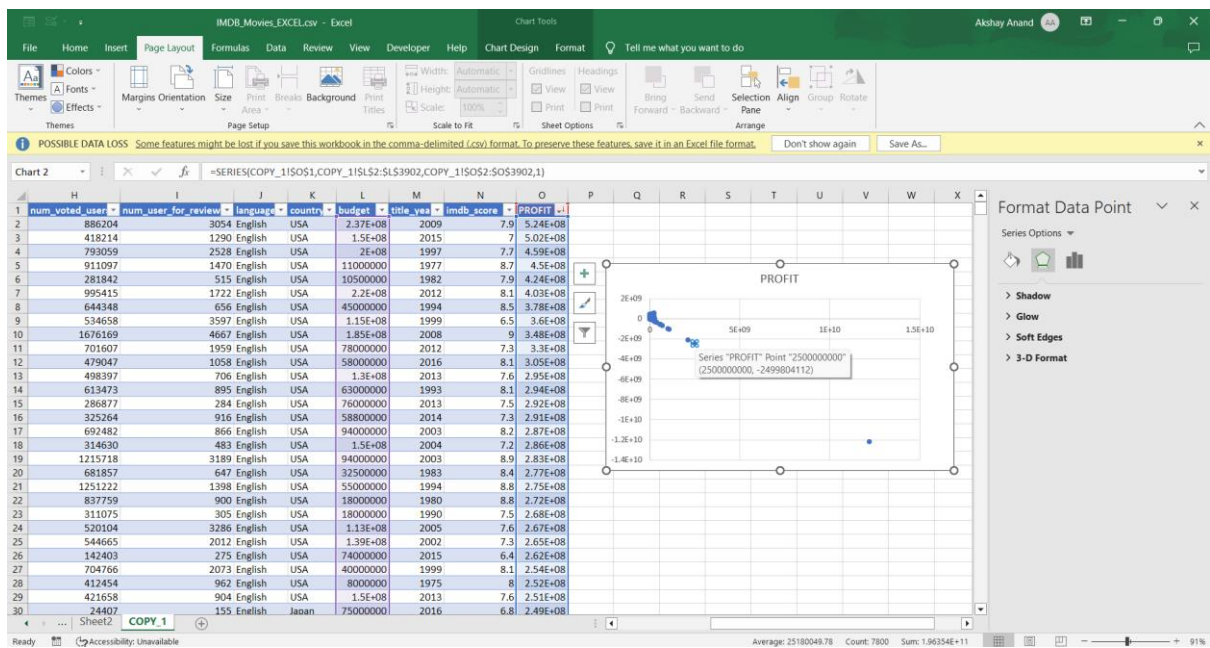
OUTLIERS->

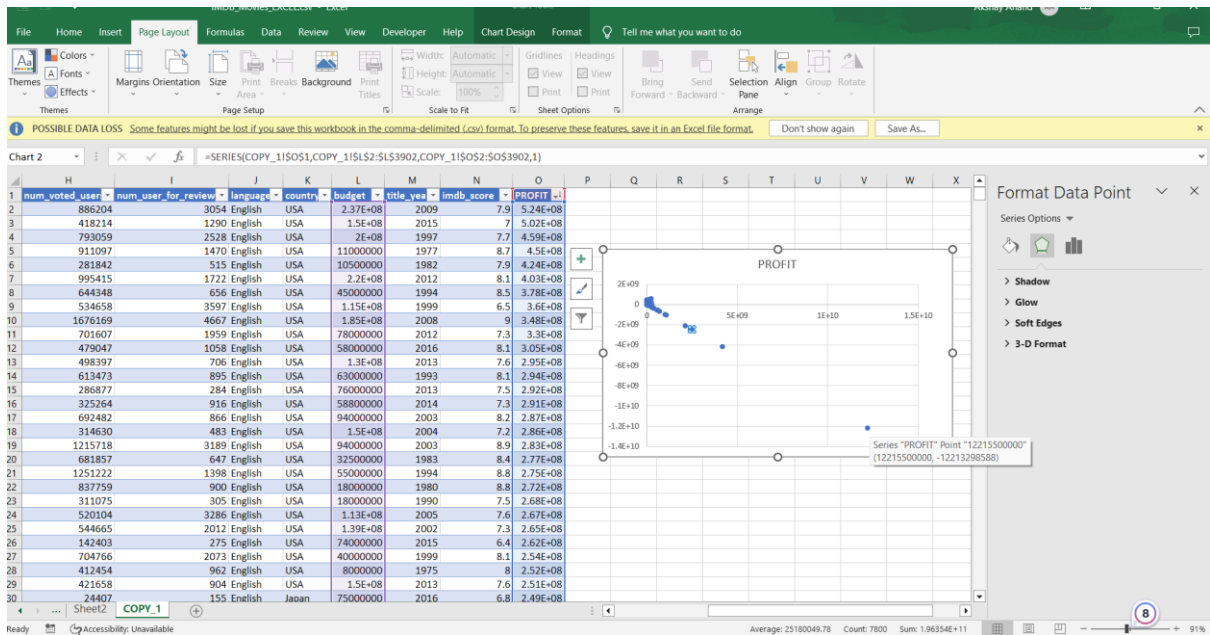
-1099560838, -2127109510, -2499804112, -4199788333, -12213298588



POINTS THAT ARE SIGNIFICANTLY AWAY FROM THE EXPECTED RANGE ARE CALLED OUTLIERS. FROM THE GRAPH WE CAN CLEARLY SEE OUTLIERS.







num_voted_user	num_user_for_review	language	country	budget	title_year	imdb_score	PROFIT
886204	3054	English	USA	2.37E+08	2009	7.9	5.24E+08
418214	1290	English	USA	1.5E+08	2015	7	5.02E+08
793059	2528	English	USA	2E+08	1997	7.7	4.59E+08
911097	1470	English	USA	11000000	1977	8.7	4.5E+08
281842	515	English	USA	10500000	1982	7.9	4.24E+08
995415	1722	English	USA	2.2E+08	2012	8.1	4.03E+08
644348	656	English	USA	45000000	1994	8.5	3.78E+08
534658	3597	English	USA	1.15E+08	1999	6.5	3.6E+08
1676169	4667	English	USA	1.85E+08	2008	9	3.48E+08
701607	1959	English	USA	78000000	2012	7.3	3.3E+08
479047	1058	English	USA	58000000	2016	8.1	3.05E+08
498397	706	English	USA	1.3E+08	2013	7.6	2.95E+08
613473	895	English	USA	63000000	1993	8.1	2.94E+08
286877	284	English	USA	76000000	2013	7.5	2.92E+08
325264	916	English	USA	58800000	2014	7.3	2.91E+08
692482	866	English	USA	94000000	2003	8.2	2.87E+08
314630	483	English	USA	1.5E+08	2004	7.2	2.86E+08
1215718	3189	English	USA	94000000	2003	8.9	2.83E+08
681857	647	English	USA	32500000	1983	8.4	2.77E+08
1251222	1398	English	USA	55000000	1994	8.8	2.75E+08
837759	900	English	USA	18000000	1980	8.8	2.72E+08
311075	305	English	USA	18000000	1990	7.5	2.68E+08
520104	3286	English	USA	1.13E+08	2005	7.6	2.67E+08
544665	2012	English	USA	1.39E+08	2002	7.3	2.65E+08
142403	275	English	USA	74000000	2015	6.4	2.62E+08
704766	2073	English	USA	40000000	1999	8.1	2.54E+08
412464	962	English	USA	8000000	1975	8	2.52E+08
421658	904	English	USA	1.5E+08	2013	7.6	2.51E+08
24407	155	English	Japan	75000000	2016	6.8	2.49E+08

THE MOVIE WITH THE HIGHEST PROFIT IS Avatar.

Top 250: Create a new column IMDb_Top_250 and store the top 250 movies with the highest IMDb Rating (corresponding to the column: imdb_score). Also make sure that for all of these movies, the num_voted_users is greater than 25,000. Also add a Rank column containing the values 1 to 250 indicating the ranks of the corresponding films.

Extract all the movies in the MDb_Top_250 column which are not in the English language and store them in a new column named Top_Foreign_Lang_Film. You can use your own imagination also!

Your task: Find IMDB Top 250

TOP 250

IMDb_Movies_EXCEL.csv - Excel

File Home Insert Page Layout Formulas Data Review View Developer Help Tell me what you want to do

Queries & Connections Properties Refresh All

Sort Filter Clear Reapply Advanced Text to Columns Data Validation Manage Data Model What-If Analysis Forecast Sheet Group Ungroup Subtotal Solver

Get & Transform Data

Get or Transform Data

Queries & Connections

Sort & Filter

Advanced

Data Tools

What-If Analysis

Forecast Sheet

Group Ungroup Subtotal

Solver

Outline

Analyze

POSSIBLE DATA LOSS Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.

Don't show again Save As...

Q8

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
	director_name	num_crits_for_review	duration	gross	gross_us	gross_intl	actor_1_name	imdb_score	num_voted_user	num_user_for_review	language	country	budget	title_year	imdb_score	RT_RANK			
1	Frank Darabont	398	142	28341469	Crim	Dr Morgan Freeman The Shawshank		1689764		4144	English	USA	25000000	1994	9.5	1			
2	Francis Ford Copp	208	175	134821952	Crim	Dr Al Pacino The Godfather		1155770		2238	English	USA	6000000	1972	9.2	2			
3	Christopher Nolan	645	152	535318061	Action	Dr Christian Bale The Dark Knight		1676149		4667	English	USA	185000000	2008	9	3			
4	Francis Ford Copp	149	220	175000000	Crim	Dr Robert De Niro The Godfather		795025		650	English	USA	13000000	1974	9	4			
5	Peter Jackson	328	192	377019252	Action	Dr Orlando Bloom The Lord of the		1215718		3189	English	USA	94000000	2003	8.9	5			
6	Steven Spielberg	174	185	96087179	Biography	Liam Neeson Schindler's List		865020		1273	English	USA	22000000	1993	8.9	6			
7	Quentin Tarantino	215	178	107892000	Crim	Dr Bruce Willis Pulp Fiction		1246480		2195	English	USA	8000000	1994	8.9	7			
8	Sergio Leone	181	142	6100000	Western	Clint Eastwood The Good, the		501509		780	Italian	Italy	1200000	1966	8.9	8			
9	Christopher Nolan	642	148	209168861	Action	Dr Leonardo DiCaprio Inception		1448200		2803	English	USA	160000000	2010	8.8	9			
10	Peter Jackson	197	171	11181757	Action	Dr Christopher Lee The Lord of the		1378746		5040	English	New Zeala	91000000	2001	8.8	10			
11	David Fincher	315	151	37023595	Drama	Brad Pitt Fight Club		1347461		2968	English	USA	63000000	1999	8.8	11			
12	Robert Zemeckis	149	147	329691194	Comedy	Tom Hanks Forrest Gump		125122		1358	English	USA	55000000	1994	8.8	12			
13	Joan Kempner	225	127	200181951	Action	Dr Harrison Ford Star Wars: Epi		837759		900	English	USA	70000000	1980	8.6	13			
14	Peter Jackson	294	172	34047889	Action	Dr Christopher Lee The Lord of the		1100446		2417	English	USA	94000000	2002	8.7	14			
15	Lana Wachowski	313	136	171383253	Action	Dr Keanu Reeves The Matrix		1217752		3646	English	USA	63000000	1999	8.7	15			
16	Martin Scorsese	392	146	66830394	Biography	Robert De Niro Goodfellas		728465		989	English	USA	25000000	1990	8.6	16			
17	George Lucas	282	125	460935685	Action	Dr Harrison Ford Star Wars: Epi		911097		1470	English	USA	11000000	1977	8.7	17			
18	Milos Forman	149	133	11200000	Drama	Scatman Crother One Flew Over		680241		760	English	USA	4400000	1975	8.7	18			
19	Fernando Meirelles	134	135	7863397	Crim	Dr Alice Braga City of God		533120		749	Portuguese	Brazil	3100000	2002	8.6	19			
20	Akira Kurosawa	153	202	260061	Action	Dr Takashi Shimura Seven Samurai		229012		598	Japanese	Japan	2000000	1954	8.7	20			
21	Christopher Nolan	712	169	189991439	Adventure	Matthew McConaughey Interstellar		928227		2725	English	USA	165000000	2014	8.6	21			
22	Steven Spielberg	219	169	216111491	Action	Dr Tom Hanks Saving Private		881236		2277	English	USA	70000000	1998	8.6	22			
23	David Fincher	216	127	100125540	Crim	Dr Morgan Freeman Se7en		1023511		1080	English	USA	35000000	1995	8.6	23			
24	Jonathan Demme	185	135	110727000	Crim	Dr Anthony Hopkins The Silence of		887467		916	English	USA	19000000	1991	8.6	24			
25	Hayao Miyazaki	146	125	10049886	Adventure	Bunta Sugawara Spirited Away		417971		902	Japanese	Japan	19000000	2001	8.6	25			
26	Tony Kaye	162	101	6712241	Crim	Dr Ethan Suplee American Hist		782437		1420	English	USA	7500000	1998	8.6	26			
27	Bryan Singer	162	106	227207208	Crim	Dr Kevin Spacey The Usual Sus		740918		1242	English	USA	6000000	1995	8.6	27			
28	Charles Chaplin	120	87	165145	Comedy	Chaplinette Goddard Modern Times		145086		211	English	USA	1500000	1936	8.6	28			
29	Christopher Nolan	813	164	448130642	Action	Dr Tom Hardy The Dark Knight		1144337		2701	English	USA	250000000	2012	8.5	29			
30	Kubrick Scott	265	177	147670868	Action	Dr Djimon Hounsou Gladiator		962637		2368	English	USA	100000000	2000	8.5	30			
31	James Cameron	235	193	208484350	Action	Dr Joe Morton Terminator 2: The		744893		983	English	USA	100000000	1991	8.5	31			
32	Quentin Tarantino	765	165	162804648	Drama	Dr Leonardo DiCaprio Django Unch		955174		1193	English	USA	100000000	2012	8.5	32			
33	Martin Scorsese	352	151	152373442	Crim	Dr Leonardo DiCaprio The Departed		879469		2054	English	USA	30000000	2006	8.5	33			
34	Roger Ailes	186	73	42278777	Adventure	Matthew Broderick The Lion King		644348		656	English	USA	45000000	1994	8.5	34			
35	Frank Darabont	186	189	158601974	Crim	Dr Tom Hanks The Green Mile		782610		1377	English	USA	60000000	1999	8.5	35			
36	Christopher Nolan	541	150	15082743	Drama	Dr Christian Bale The Prestige		844052		1120	English	USA	40000000	2006	8.5	36			
37	Roman Polanski	199	150	15232922	Biography	Emilia Fox The Pianist		487946		761	English	France	35000000	2002	8.5	37			
38	Francis Ford Copp	261	289	78800000	Drama	Dr Harrison Ford Apocalypse Ni		450676		993	English	USA	31500000	1979	8.5	38			

IMDb_Movies_EXCEL COPY 2 COPY 1

Ready Accessibility: Unavailable

TOP 250 NOT IN ENGLISH

IMDb_Movies_EXCEL - Excel

FileHomeInsertPage LayoutFormulasDataReviewViewDeveloperHelpTell me what you want to do

Get DataFrom Text/CSVFrom WebFrom Table/Range

Existing SourcesRecent Connections

Queries & ConnectionsPropertiesRefreshAllEdit Links

SortFilterClearReapplyAdvancedText to ColumnsData ValidationManage Data Model

Flash FillConsolidateRemove DuplicatesRelationshipsWhat-If AnalysisForecast SheetGroup Ungroup SubtotalSolver

Get & Transform Data

Queries & ConnectionsSort & FilterData ToolsData AnalysisOutlineAnalyze

POSSIBLE DATA LOSSSome features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.

Don't show againSave As...

Q8

director_name

num_crits_for_review

duration

gross

gross_us

gross_intl

actor_1_name

imdb_score

num_voted_user

num_user_for_review

language

country

budget

title_year

imdb_score

RT_RANK

1

Sergio Leone

181

142

6100000

Western

Clint Eastwood

The Good, the

501509

780

Italian

Italy

1200000

1966

8.9

8

2

Fernando Meirelles

214

135

7563397

Crim

Dr Alice Braga

City of God

533120

749

Portuguese

Brazil

3100000

2002

8.7

19

3

Akira Kurosawa

153

202

15501940

Biography

Thomas Kretschmann

Downfall

228012

598

Japanese

Japan

2000000

1954

8.7

20

4

Hayao Miyazaki

146

125

10049886

Adventure

Bunta Sugawara

Spirited Away

417971

902

Japanese

Japan

19000000

2001

8.6

25

5

Florian Henckel von Donners

215

137

11284637

Drama

Th Sebastian Koch

The Lives of O

259379

407

German

Germany

2000000

2006

8.5

45

6

Majid Majidi

46

89

625402

Drama

Pa Behrouz Soudji

Children of he

27887

130

Persian

Iran

180000

1997

8.5

46

7

Jean-Pierre Jeunet

242

122

33201861

Comedy

R Mathieu Kassovitz

Amélie

534262

1314

French

France

7700000

2001

8.4

49

8

S.S. Rajamouli

44

159

6488000

Action

Ad Tamannaah Bhat

Baahubali: Th

627556

410

Telugu

India

18026148

2015

8.4

50

9

Hayao Miyazaki

174

134

2228193

Adventure

Minnie Driver

Princess Mon

211352

570

Japanese

Japan

240000

1987

8.4

53

10

Wolfgang Petersen

96

293

11433134

Adventure

Jürgen Prochnow

Das Boot

168203

426

German

West Ger

14000000

1981

8.4

57

11

 Chan-wook Park | 305 | 120 | 2181290 | Drama | M Min-sik Choi | Oldboy | | 356181 | | 808 | Korean | South Kore | 3000000 | 2003 | 8.4 | 59 |

12

 Angharad | 354 | 123 | 708492 | Drama | M Shahab Hosseini | A Separation | | 515812 | | 264 | Persian | Iran | 500000 | 2011 | 8.4 | 61 |

13

 Fritz Lang | 260 | 145 | 26435 | Drama | Sc Brigitte Helm | Metropolis | | 111841 | | 413 | German | Germany | 6000000 | 1927 | 8.3 | 74 |

14

 Oliver Hirschbiegel | 192 | 178 | 5501940 | Biography | Thomas Kretschmann | Downfall | | 248354 | | 564 | German | Germany | 13500000 | 2004 | 8.3 | 76 |

15

 Thomas Vinterberg | 186 | 115 | 102668 | Drama | Thomas Lejacen | The Hunt | | 370235 | | 248 | Danish | Denmark | 3800000 | 2012 | 8.3 | 82 |

16

 Hayao Miyazaki | 212 | 119 | 4710455 | Adventure | Christian Bale | How's Movin | | 214091 | | 350 | Japanese | Japan | 2400000 | 2004 | 8.2 | 97 |

17

 Guillermo del Toro | 406 | 122 | 37623143 | Drama | Pa Inma Baquerro | Pan's Labrynth | | 467234 | | 1083 | Spanish | Spain | 13500000 | 2006 | 8.2 | 99 |

18

 Denis Villeneuve | 226 | 139 | 6837099 | Drama | M Lubna Aual | Incendies | | 80429 | | 150 | French | Canada | 6800000 | 2010 | 8.2 | 101 |

19

 Juan José Campanella | 262 | 129 | 20167424 | Drama | M Ricardo Darín | The Secret in T | | 131831 | | 231 | Spanish | Argentina | 2000000 | 2009 | 8.2 | 103 |

20

 Aléjandro Amenábar | 157 | 125 | 206345 | Biography | Belén Rueda | The Sea inside | | 44556 | | 140 | Spanish | Spain | 10000000 | 2004 | 8.1 | 137 |

21

 Je-kyu Kang | 86 | 148 | 1101018 | Action | Ad Min-sik Choi | The God of Th | | 319483 | | 224 | Korean | South Kore | 31800000 | 2004 | 8.1 | 138 |

22

 Katsuhiro A'tomo | 150 | 124 | 439162 | Action | An Mitsuo Iwata | Akira | | 106160 | | 450 | Japanese | Japan | 1.1E+09 | 1988 | 8.1 | 140 |

23

 José Padilha | 142 | 115 | 8060 | Action | Cr Wagner Moura | Elite Squad | | 81644 | | 107 | Portuguese | Brazil | 4000000 | 2007 | 8.1 | 141 |

24

 Alejandro G. S. Jodorowsky | 157 | 115 | 5363814 | Drama | Dr Adriana Barraza | Amores Perros | | 173551 | | 561 | Spanish | Mexico | 2000000 | 2000 | 8.1 | 150 |

25

 Thomas Vinterberg | 198 | 105 | 1647780 | Drama | Ulrich Thomsen | The Celebratin | | 65951 | | 258 | Danish | Denmark | 1300000 | 1998 | 8.1 | 151 |

26

 Karen Johar | 210 | 128 | 4018895 | Adventure | Shah Rukh Khan | My Name is K | | 69759 | | 235 | Hindi | India | 12000000 | 2010 | 8 | 193 |

27

 Vincent Paronnaud | 89 | 243 | 89 | 4614603 | Animation | Catherine Deneuve | Reservoir | | 70284 | | 158 | French | France | 7500000 | 2007 | 8 | 195 |

28

 Walter Salles | 71 | 113 | 5595428 | Drama | Fernanda Montez | Central Statu | | 28951 | | 257 | Portuguese | Brazil | 2800000 | 1998 | 8 | 201 |

29

 Art Follman | 231 | 90 | 2283376 | Animation | Art Follman | Wait with Ba | | 48107 | | 156 | Hebrew | Israel | 1500000 | 2008 | 8 | 204 |

30

 Sergio Leone | 122 | 99 | 8200002 | Action | Dr Clint Eastwood | A Fistful of Do | | 147566 | | 235 | Italian | Italy | 200000 | 1964 | 8 | 205 |

31

 Yimou Zhang | 283 | 80 | 84861 | Action | Ad Jet Li | Hero | | 149414 | | 841 | Mandarin | China | 31000000 | 2002 | 7.9 | 221 |

32

 Ang Lee | 287 | 120 | 126067808 | Action | Dr Chen Chang | Crouching Tig | | 217740 | | 1641 | Mandarin | Taiwan | 15000000 | 2000 | 7.9 | 237 |

33

 Clint Eastwood | 251 | 141 | 13753931 | Drama | Dr Yuki Matsuyama | Letters from I | | 332449 | | 316 | Japanese | USA | 19000000 | 2006 | 7.9 | 240 |

34

 Michael Haneke | 247 | 127 | 215377 | Drama | Dr Isabelle Huppert | Amour | | 703882 | | 1 | French | France | 8900000 | 2012 | 7.9 | 244 |

35

 Chah-Chah Chah | 39 | 193 | 2921788 | Drama | Dr Khatun Bahar Khan | Ven-Ze-Nak | | 244149 | | 54 | Hindi | India | 5000000 | 2004 | 7.9 | 249 |

36

 Christophe Bruneau | 112 | 97 | 8629758 | Drama | M Jean-Baptiste Ma | The Chorus | | 44415 | | 110 | French | France | 55000000 | 2004 | 7.9 | 249 |

37

 Fabrizio Bressi | 84 | 114 | 1221261 | Crim | Dr Riccardo Dar | Nine Queens | | 382215 | | 125 | Spanish | Argentina | 15000000 | 2000 | 7.9 | 253 |

38

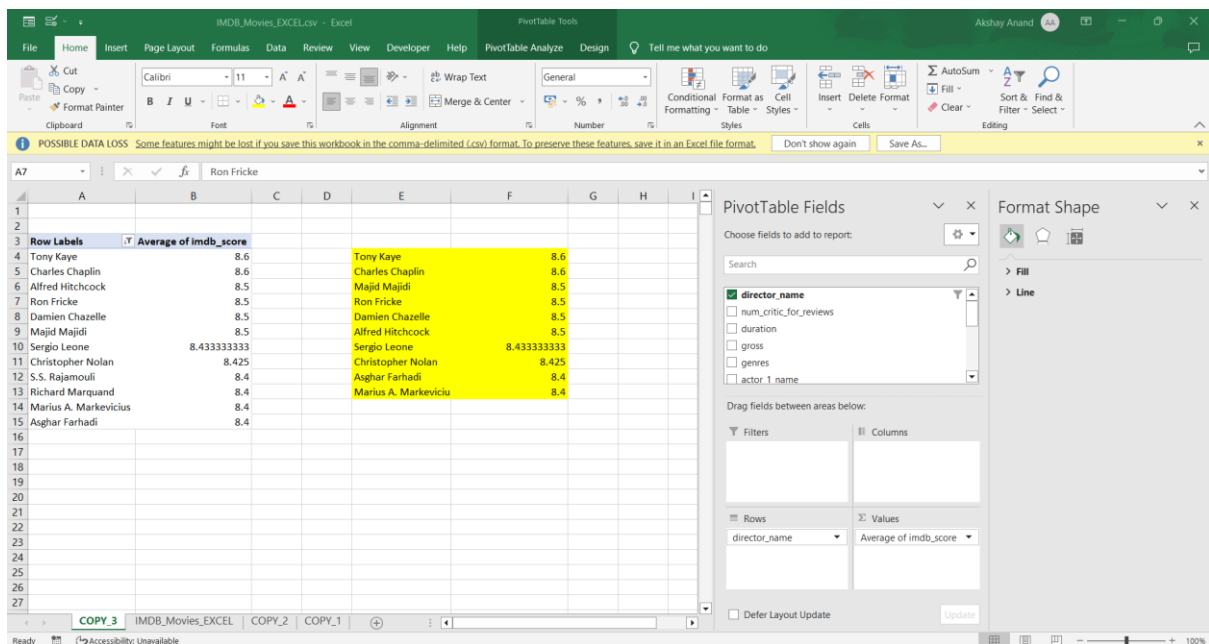
 Gracia Nolasco | 233 | 133 | 1195783 | Drama | Jeanne-Marie | 4 Months, 3 W | | 44763 | | 127 | Romanian | Romania | 900000 | 2007 | 7.9 | 254 |

IMDb_Movies_EXCEL

COPY_2COPY_1

Ready38 of 3900 records found

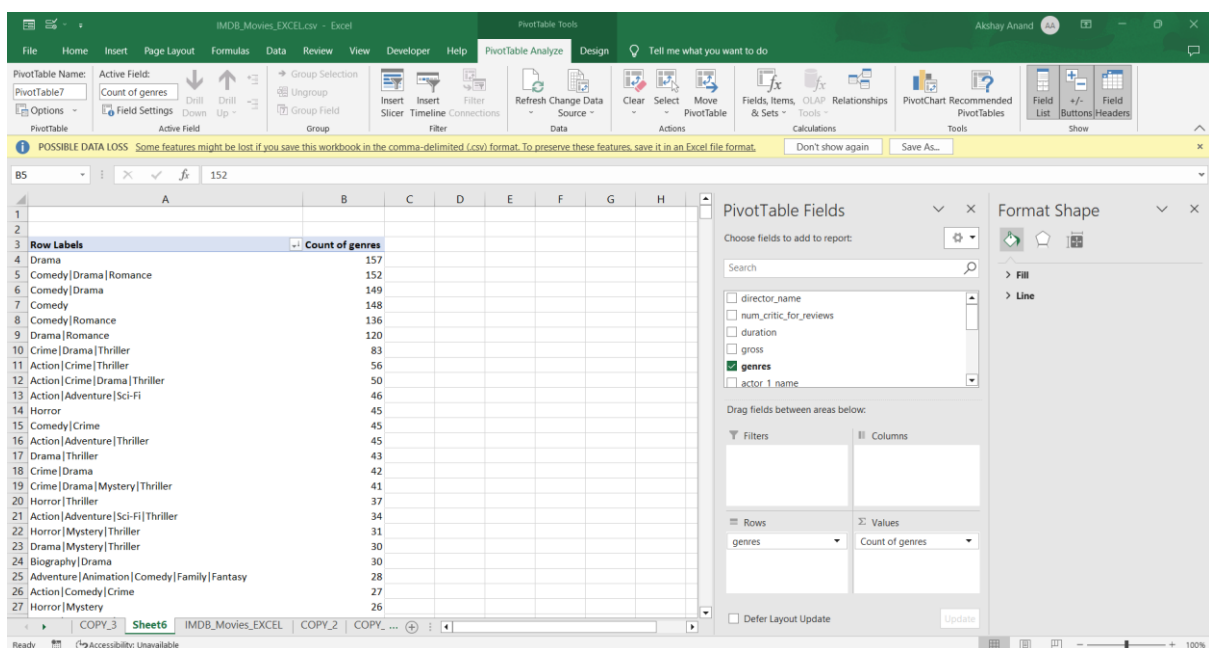
Copy toClipboardUnavailable



WE USED PIVOT TABLES TO SORT DATA BASED ON IMDB SCORE.

Popular Genres: Perform this step using the knowledge gained while performing previous steps.

Your task: Find popular genres



WE USED PIVOT TABLES TO SORT DATA BASED ON THE COUNT OF GENRES.

Charts: Create three new columns namely, Meryl_Streep, Leo_Caprio, and Brad_Pitt which contain the movies in which the actors: 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' are the lead actors. Use only the actor_1_name column for extraction. Also, make sure that you use the names 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' for the said extraction.

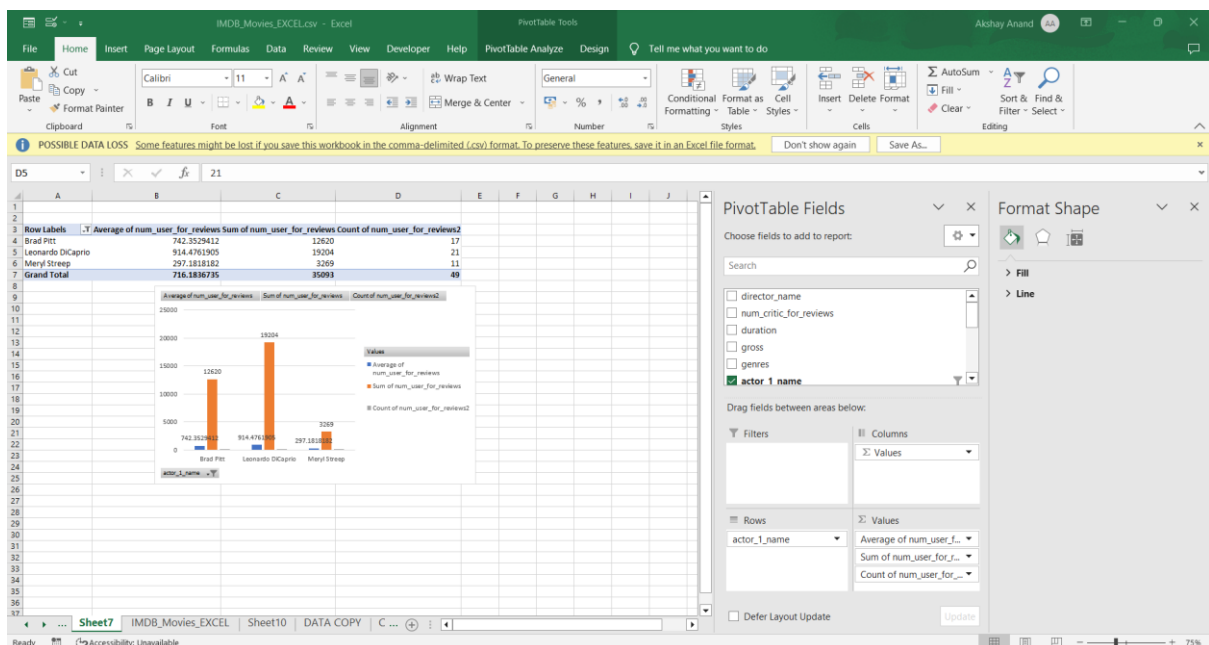
Append the rows of all these columns and store them in a new column named Combined.

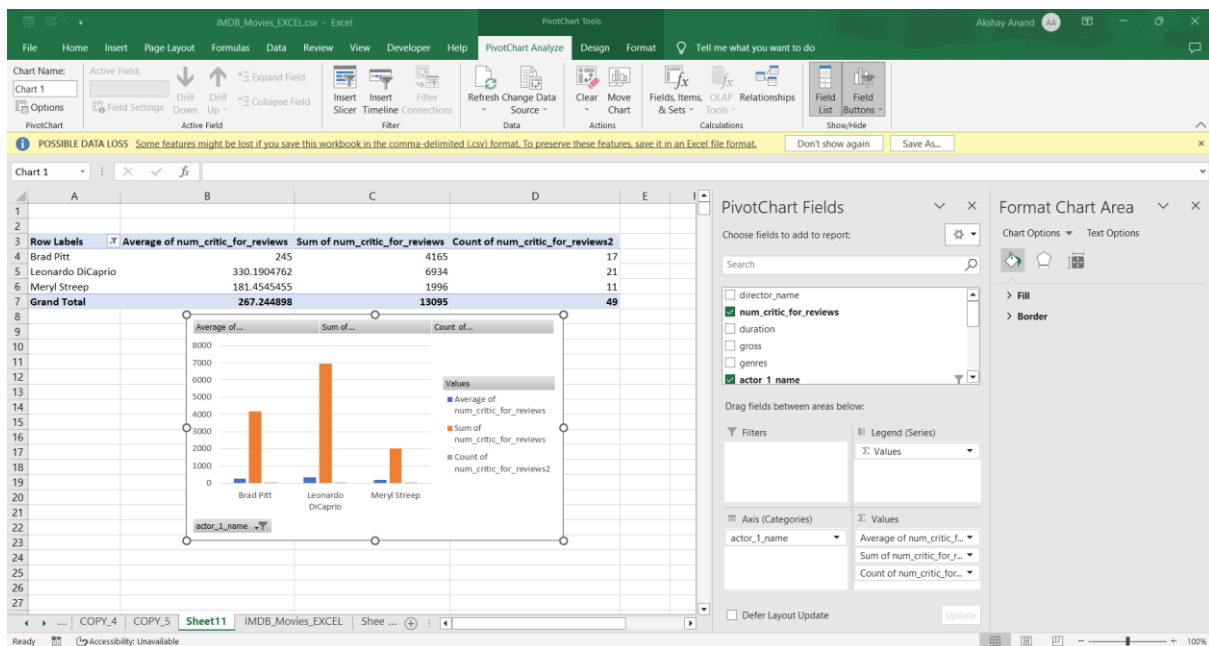
Group the combined column using the actor_1_name column.

Find the mean of the num_critic_for_reviews and num_users_for_review and identify the actors which have the highest mean.

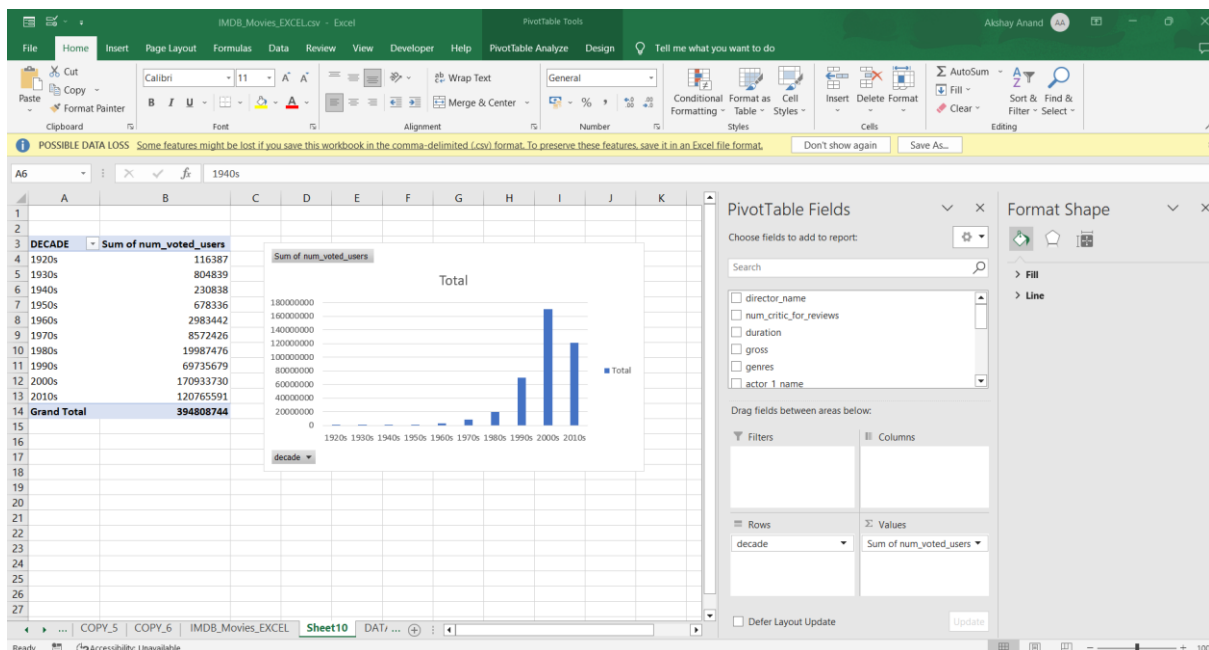
Observe the change in number of voted users over decades using a bar chart. Create a column called decade which represents the decade to which every movie belongs to. For example, the title_year year 1923, 1925 should be stored as 1920s. Sort the column based on the column decade, group it by decade and find the sum of users voted in each decade. Store this in a new data frame called df_by_decade.

Your task: Find the critic-favorite and audience-favorite actors





FROM ABOVE IT IS CLEAR THAT LEONARDO DICAPRIO IS BOTH THE CRITIC FAVOURITE AND THE AUDIENCE FAVOURITE ACTOR.



DECADES HAVE BEEN MADE BY: =FLOOR(M2,10) & "s"

WE USED PIVOT TABLES TO DRAW/PLOT GRAPHS.

RESULT: WE USED MS EXCEL FUNCTIONS SUCH AS FORMULAE, SORT, FILTER, PIVOT TABLES ETC TO PERFORM AN ANALYSIS OF THE DATASET AND GET INSIGHTS.