IMDb

Name: **Paritosh Gandre**

Email-Id: **paritoshkrcg@gmail.com**

PROJECT DESCRIPTION

This project is about finding out all the outliers, user votes for a particular film, an actor or a director from IMDb dataset provided. Each and every question is attempted and completed with desired results achieved.

In this project I have performed queries on following tasks:

A. **Cleaning Data**: Here we are supposed to clean the data which is provided by the Trainity team. When we gather a data from multiple sources there is a high probability that the data may contain null values or the data can be inconsistent, to avoid problems while querying or to get proper results we have to clean the data first then start querying the data.
   My task: Clean the Data.

B. **Movies with highest profit:** Here we are told to create a column called profit containing the difference of the two columns that are gross and budget column. Then we have to sort the column by profit and plot them using scatter plot or any plot for better understanding.
   My task: Finding movies with highest profits.

C. **TOP 250:** Here we are supposed to find out the top 250 movies with a condition of number of voted users should be greater than 25,000.
   My task: Find top 250 IMDb Movies

D. **Best Directors:** Here I am supposed to find top 10 directors with mean IMDb score being amongst the highest.
   My Task: Find best directors

E. **Popular Genres:** By using our previous knowledge we have to calculate the most popular genre from the dataset.
   My Task: find out popular genres from dataset

F. **Charts:** Here, we have to create 3 new columns named after actors mentioned in the question, segregating their rows and calculating the mean of critic reviews and user reviews per actor, then visualizing the results for better understanding.
   My Task: find audience and critic favourite actors.

APPROACH

My approach to this project is by understanding the dataset first, then diving into the problems given by the Trainity team.

After understanding the dataset, I have also understood all the problems given and made a gist of what I need to do to solve these queries.

Name: **Paritosh Gandre**

Email-Id: **paritoshkrcg@gmail.com**

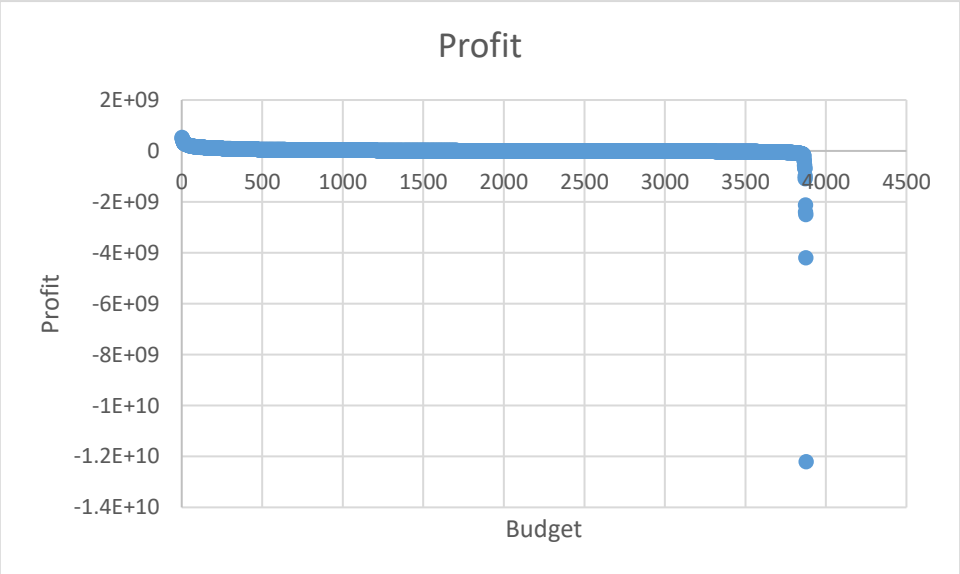
RESULT

A. Cleaning Data: I have cleaned the data by eradicating rows with empty cells, and making the dataset suitable for further analysis and querying.

First what I did is while going through dataset I filled the column name with yellow colour to indicate that those columns need to be dropped as they are irrelevant to my project or analysis.
There were 5044 rows before and 3877 after getting rid of rows with empty cells.
And I got rid of columns like "color","director_facebook_likes","actor_1_facebook_likes", "actor_2_facebook_likes","
actor_3_facebook_likes","cast_total_facebook_likes","actor_3_name","facenumber_in_posts","plot_keywords","movie_imdb_link","content_rating","aspect_ratio","movie_facebook_likes".

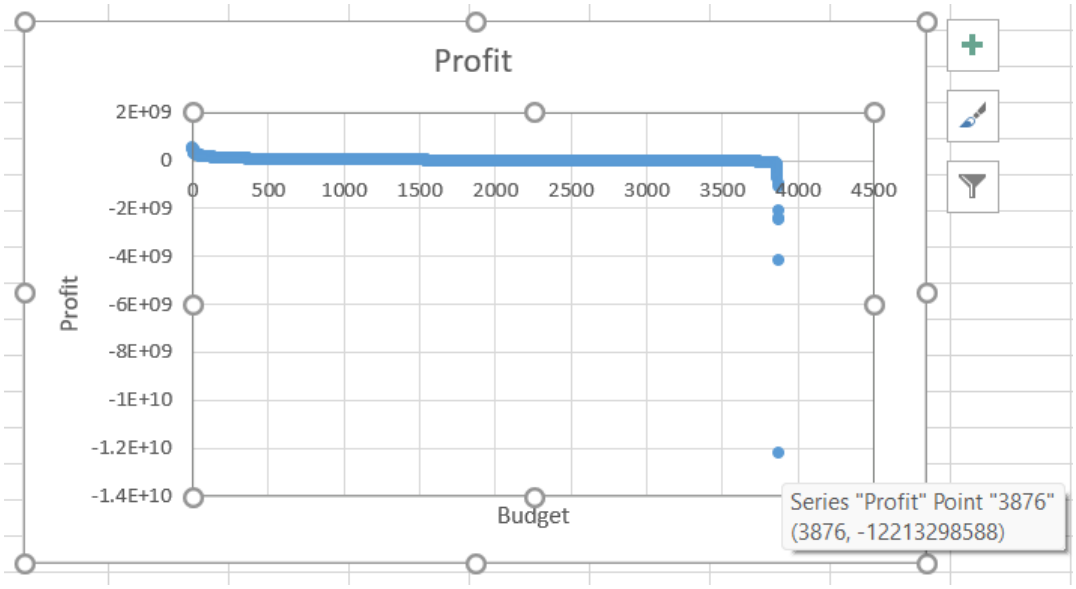
B. Movies with Highest Profit: To find the profit I have computed the difference of two columns named gross and budget to find the profit.(Gross - Budget)

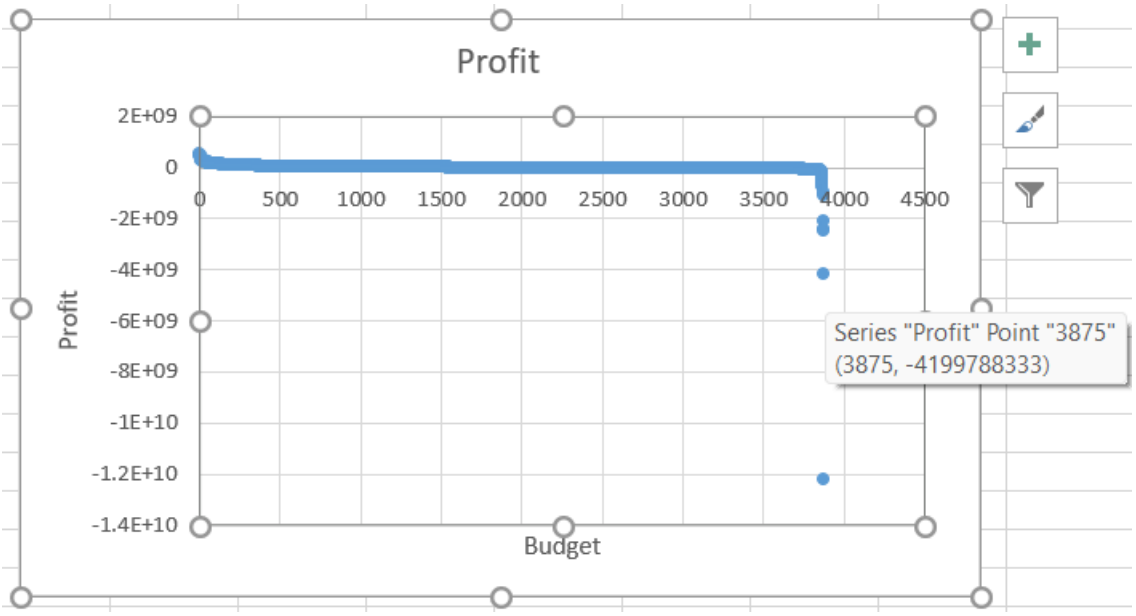
I have found Profit for each row and then plotted them into a scatter plot to make it clear how dispersed the profit is within the dataset.



Name: **Paritosh Gandre**

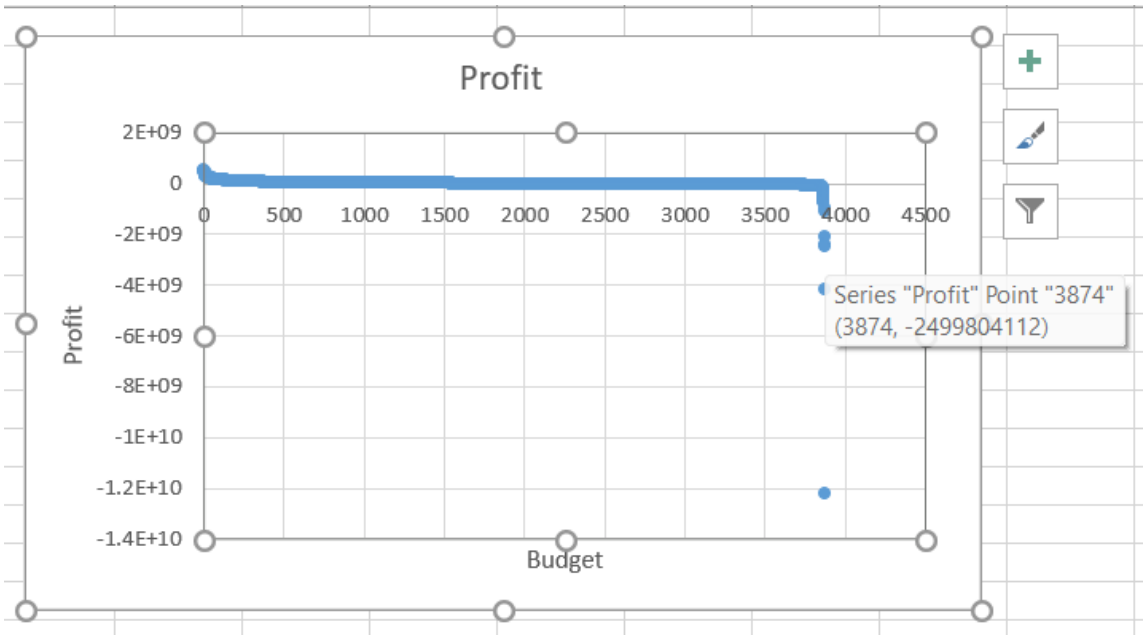
Email-Id: **paritoshkrcg@gmail.com**





Name: **Paritosh Gandre**

Email-Id: **paritoshkrcg@gmail.com**



- Outliers are :
  1. -1099560838
  2. -2127109510
  3. -2397701809
  4. -2499804112
  5. -4199788333
  6. 12213298588

- **Dataset of Movies with Highest Profit:**

Sr no	director_name	num_critic_for_reviews	duration	actor_2_name	gross	genres	actor_1_name	movie_title
1	James Cameron	723	178	Joel David Moore	760505847	Action|Adventure|Fantasy|Sci-Fi	CCH Pounder	AvatarÂ
2	Colin Trevorrow	644	124	Judy Greer	652177271	Action|Adventure|Sci-Fi|Thriller	Bryce Dallas Howard	Jurassic WorldÂ
3	James Cameron	315	194	Kate Winslet	658672302	Drama|Romance	Leonardo DiCaprio	TitanicÂ
4	George Lucas	282	125	Peter Cushing	460935665	Action|Adventure|Fantasy|Sci-Fi	Harrison Ford	Star Wars: Episode IV - A New HopeÂ
5	Steven Spielberg	215	120	Dee Wallace	434949459	Family|Sci-Fi	Henry Thomas	E.T. the Extra-TerrestrialÂ
6	Joss Whedon	703	173	Robert Downey Jr.	623279547	Action|Adventure|Sci-Fi	Chris Hemsworth	The AvengersÂ

Name: **Paritosh Gandre**

Email-Id: **paritoshkrcg@gmail.com**

Sr No.	num_voted_users	actor_3_name	num_user_for_reviews	language	country	title_year	imdb_score	budget	Profit
1	886204	Wes Studi	3054	English	USA	2009	7.9	237000000	523505847
2	418214	Omar Sy	1290	English	USA	2015	7	150000000	502177271
3	793059	Gloria Stuart	2528	English	USA	1997	7.7	200000000	458672302
4	911097	Kenny Baker	1470	English	USA	1977	8.7	11000000	449935665
5	281842	Peter Coyote	515	English	USA	1982	7.9	10500000	424449459
6	995415	Scarlett Johansson	1722	English	USA	2012	8.1	220000000	403279547

C. **TOP 250**: TOP 250 is something that everyone keeps an eye on. So we were told to find out top 250 movies with a criteria that the number of voted users should be greater than 25000.

To calculate the top 250 with the condition mentioned above, I have first used the FILTER option in excel and gave a condition for the column "num_voted_users" > 25,000 and sorted it in descending order.

Because I also have to make two different extractions for English and Non English language movies: I simply applied a filter on the language column and selected "Select all" option for all language and selected everything except "English" for Non English movies.

And later I used the RANK():RANK(P2,P$2:$P$251,0)+COUNTIFS($P$2:P2,P2)-1

Name: **Paritosh Gandre**

Email-Id: **paritoshkrcg@gmail.com**

- **FOR ENGLISH LANGUAGE:**

Sr no.	director_name	num_critic_for_reviews	duration	actor_2_name	gross	genres	actor_1_name	movie_title	actor_3_name
1	Frank Darabont	199	142	Jeffrey DeMunn	28341469	Crime|Drama	Morgan Freeman	The Shawshank RedemptionÂ	Bob Gunton
2	Francis Ford Coppola	208	175	Marlon Brando	134821952	Crime|Drama	Al Pacino	The GodfatherÂ	Robert Duvall
3	Christopher Nolan	645	152	Heath Ledger	533316061	Action|Crime|Drama|Thriller	Christian Bale	The Dark KnightÂ	Morgan Freeman
4	Francis Ford Coppola	149	220	Al Pacino	57300000	Crime|Drama	Robert De Niro	The Godfather: Part IIÂ	Robert Duvall
5	Quentin Tarantino	215	178	Eric Stoltz	107930000	Crime|Drama	Bruce Willis	Pulp FictionÂ	Phil LaMarr
6	Peter Jackson	328	192	Billy Boyd	377019252	Action|Adventure|Drama|Fantasy	Orlando Bloom	The Lord of the Rings: The Return of the KingÂ	Bernard Hill

Sr no.	country	title_year	budget	Profit	Rank	imdb_score	num_voted_users	language	country
1	USA	1994	25000000	3341469	1	9.3	1689764	English	USA
2	USA	1972	6000000	128821952	2	9.2	1155770	English	USA
3	USA	2008	185000000	348316061	4	9	1676169	English	USA
4	USA	1974	13000000	44300000	4	9	790926	English	USA
5	USA	1994	8000000	99930000	6	8.9	1324680	English	USA
6	USA	2003	94000000	283019252	7	8.9	1215718	English	USA

Name: **Paritosh Gandre**

Email-Id: **paritoshkrcg@gmail.com**

- **FOR NON English LANGUAGES:**

Sr no.	director_name	num_critic_for_reviews	duration	actor_2_name	gross	genres	actor_1_name	movie_title	num_voted_users
1	Sergio Leone	181	142	Luigi Pistilli	6100000	Western	Clint Eastwood	The Good, the Bad and the UglyÂ	503509
2	Fernando Meirelles	214	135	Seu Jorge	7563397	Crime|Drama	Alice Braga	City of GodÂ	533200
3	Akira Kurosawa	153	202	Minoru Chiaki	269061	Action|Adventure|Drama	Takashi Shimura	Seven SamuraiÂ	229012
4	Hayao Miyazaki	246	125	RyÃ»nosuke Kamiki	10049886	Adventure|Animation|Family|Fantasy	Bunta Sugawara	Spirited AwayÂ	417971
5	Florian Henckel von Donnersmarck	215	137	Ulrich MÃ¼he	11284657	Drama|Thriller	Sebastian Koch	The Lives of OthersÂ	259379
6	Majid Majidi	46	89	Amir Farrokh Hashemian	925402	Drama|Family	Bahare Seddiqi	Children of HeavenÂ	27882

Sr no	actor_3_name	num_user_for_reviews	language	country	title_year	imdb_score	budget	Profit	Rank
1	Enzo Petito	780	Italian	Italy	1966	8.9	1200000	4900000	8
2	Alexandre Rodrigues	749	Portuguese	Brazil	2002	8.7	3300000	4263397	19
3	Kamatari Fujiwara	596	Japanese	Japan	1954	8.7	2000000	-1730939	20
4	Miyu Irino	902	Japanese	Japan	2001	8.6	19000000	-8950114	27
5	Martina Gedeck	407	German	Germany	2006	8.5	2000000	9284657	45
6	Mohammad Amir Naji	130	Persian	Iran	1997	8.5	180000	745402	46

Name: **Paritosh Gandre**

Email-Id: **paritoshkrcg@gmail.com**

D. **Best Directors:** Here I have to find out the mean for each director first and then sort them in descending order to know which director is the best according to the IMDb score.

I used pivot table to get the desired outcome with Rows as "director_name" and values as "AVG(IMDb_score)".

- **Top 10 directors using pivot table:**

Row Labels	Average of imdb_score
Charles Chaplin	8.6
Tony Kaye	8.6
Alfred Hitchcock	8.5
Damien Chazelle	8.5
Majid Majidi	8.5
Ron Fricke	8.5
Sergio Leone	8.433333333
Christopher Nolan	8.425
Asghar Farhadi	8.4
Marius A. Markevicius	8.4

E. **Popular Genres:** Here I have used my previous knowledge to get the best results and appropriate one to the query to find popular genre from the dataset.

To find out the most popular genre, I again used the pivot table. For the rows I have used "Genre" and for the values "Count(genre)".

Name: **Paritosh Gandre**

Email-Id: **paritoshkrcg@gmail.com**

- **Top 10 genre :**

Genre Labels	Count of genres
Comedy|Drama|Romance	147
Drama	144
Comedy	144
Comedy|Drama	139
Comedy|Romance	132
Drama|Romance	117
Crime|Drama|Thriller	82
Action|Crime|Thriller	57
Action|Crime|Drama|Thriller	50
Action|Adventure|Sci-Fi	48
Grand Total	**1060**

F. **Charts :**

a. I have created 3 independent sheets for “Meryl Streep”, “Leonardo DiCaprio”, and “Brad Pitt”. I have segregated the data where “actor_1_name” = the names mentioned above respectively. And with the help of pivot table I calculated the average, sum and count for Critic reviews and User Reviews, and then plotted it into charts.

b. For the latter problem, I used pivot table to get the count of votes by users for each year from 1920 to 2020. Then, I created a class interval for years and summed up the count for each interval to calculate the growth per decade.

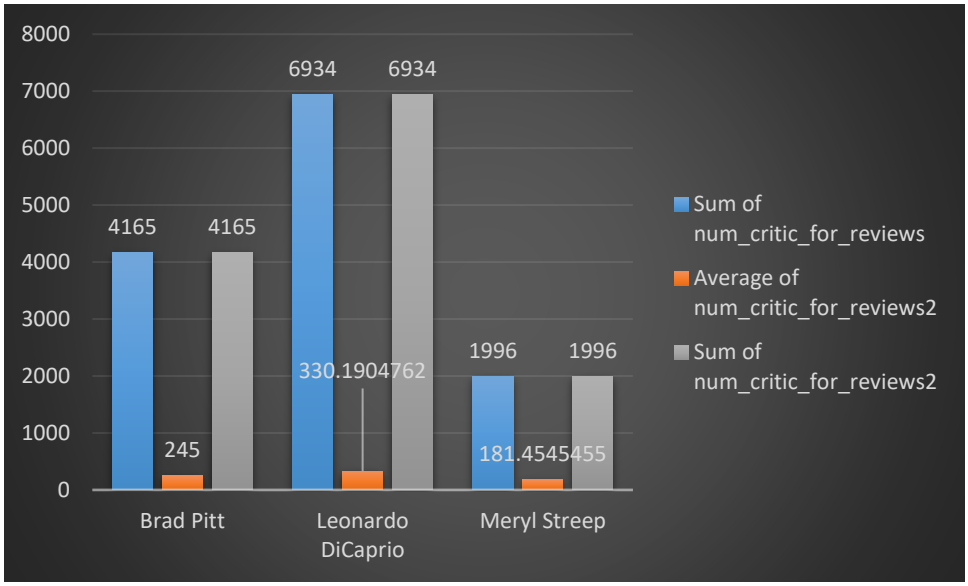
- **Result of ‘a’ question:**

Critic mean

Row Labels	Sum of num_critic_for_rev	Average of num_critic_for_revie	Sum of num_critic_for_reviews2
Brad Pitt	4165	245	4165
Leonardo DiCaprio	6934	330.1904762	6934
Meryl Streep	1996	181.4545455	1996
Grand Total	**13095**	**267.244898**	**13095**

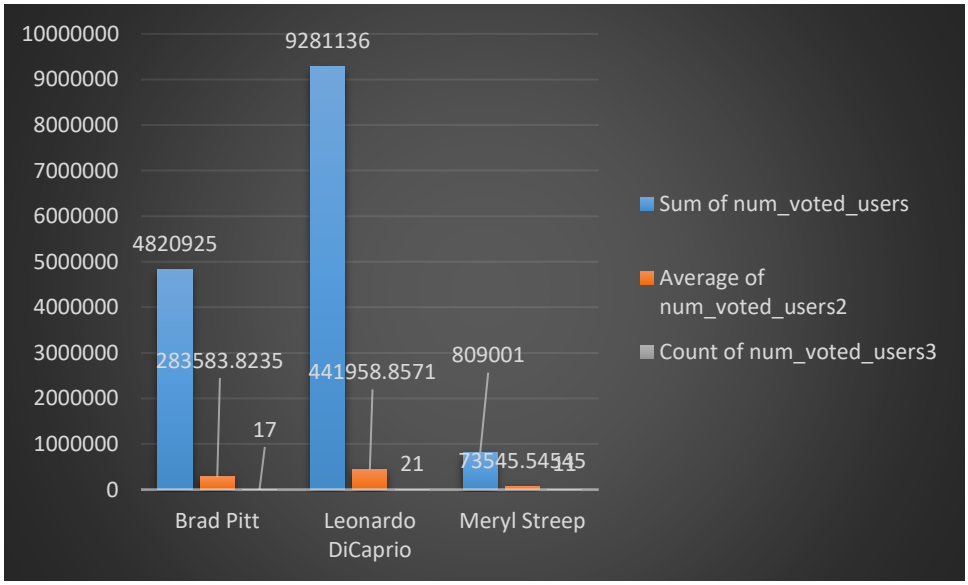
Name: **Paritosh Gandre**

Email-Id: **paritoshkrcg@gmail.com**



User mean :

Row Labels	Sum of num_voted_users	Average of num_voted_users2	Count of num_voted_users3
Brad Pitt	4820925	283583.8235	17
Leonardo DiCaprio	9281136	441958.8571	21
Meryl Streep	809001	73545.54545	11
Grand Total	**14911062**	**304307.3878**	**49**

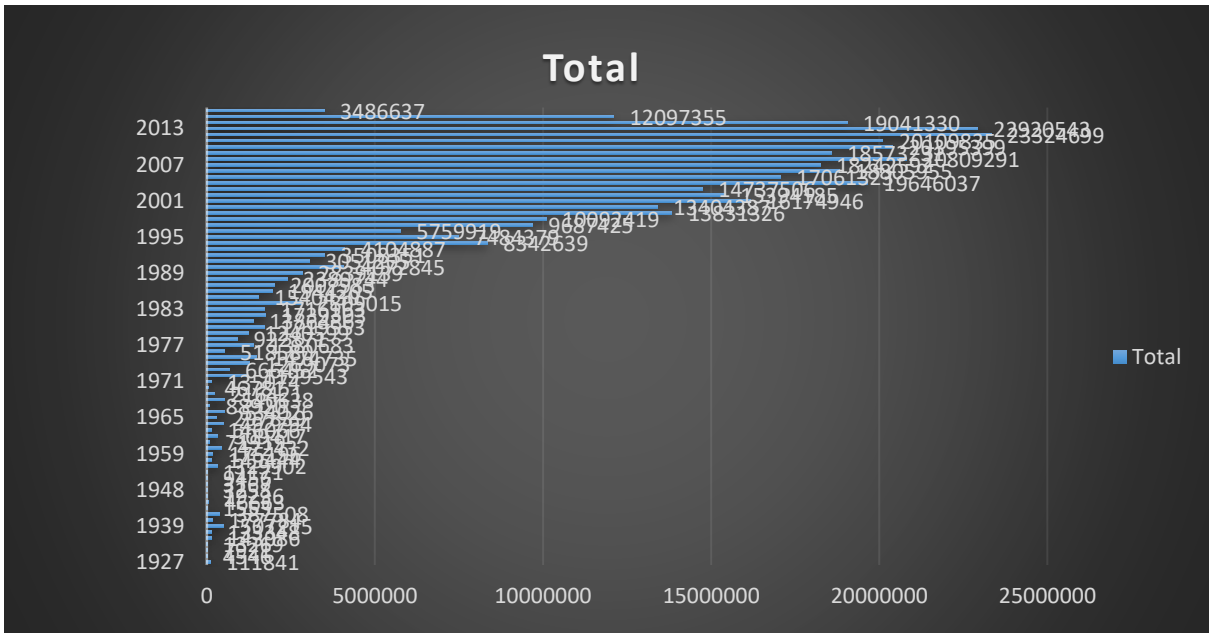


Name: **Paritosh Gandre**

Email-Id: **paritoshkrcg@gmail.com**

- **Result of 'b' question:**

  Growth per decade

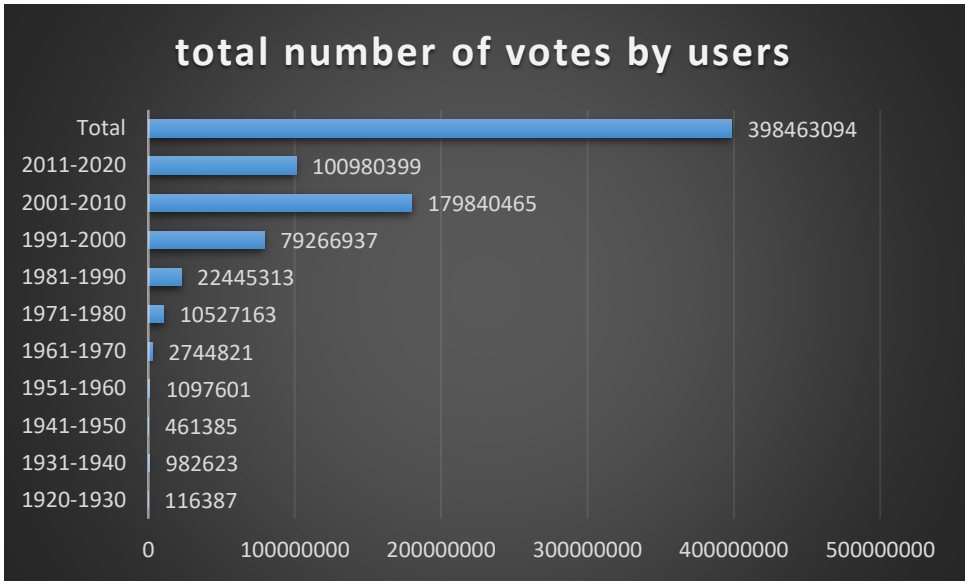


Class interval of years and sum of number of votes by users:

Range of years	total number of votes by users
1920-1930	116387
1931-1940	982623
1941-1950	461385
1951-1960	1097601
1961-1970	2744821
1971-1980	10527163
1981-1990	22445313
1991-2000	79266937
2001-2010	179840465
2011-2020	100980399
Total	398463094

Name: **Paritosh Gandre**

Email-Id: **paritoshkrcg@gmail.com**



In this way, I have answered each question properly with visualization as well.

THANK YOU