# Project: Visualizing Movie Data

Complete each section. When you are ready, save your file as a PDF document and submit it [here](https://classroom.udacity.com/nanodegrees/nd008/parts/c7cbb25d-deae-4be1-bc9c-71f465b849f8/project).

## Step 1: Data Cleanup and Attribute Selection

* I have cleaned the dataset using python’s libraries in jupyter notebook.
* There were total 8873 rows with missing values.
* **Duplicate values were removed.**
* **Null values with respect to primary key ('imdb\_id) were removed.**
* **Extraneous Columns were removed.**

I decided to drop the following following columns as that will not provide any useful info :-

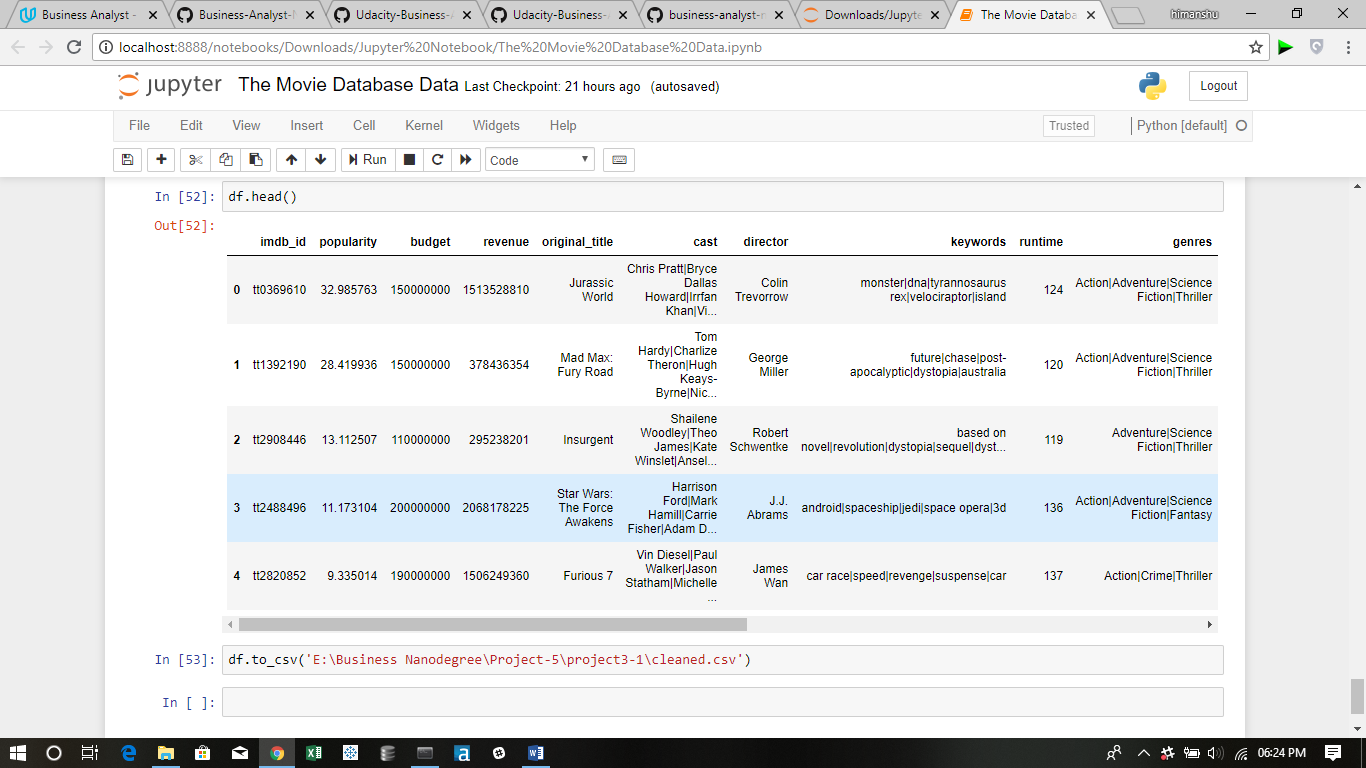
> 'id'

> 'homepage'

> 'tagline'

> 'overview'

* Cleaning dataframe :-



* I decided to explore the following attributes to dive further in my visualizations.

->Popularity

->Budget

->Revenue

->Runtime

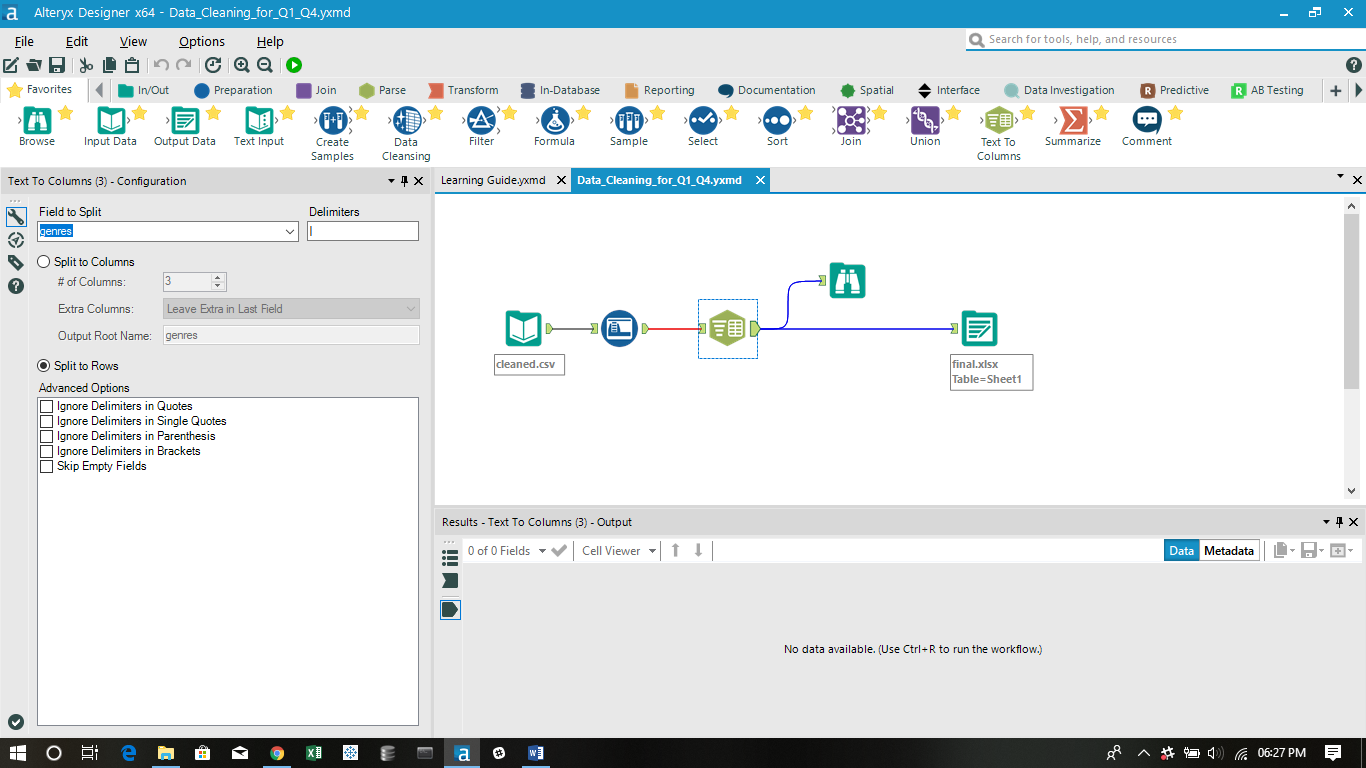
->Genres

->Release\_date

->Vote\_count

* For question 1 and question 4 :-

To split the hybrid values in the genres column , I have used Alteryx and used text to rows feature of Alteryx and using delimeter as ‘|’.



* For question 2 and 3 I have continued with cleaned file with python as splitting the ‘genres’ column in rows will draw duplicated values in other questions.

## Step 2: Tableau Visualizations

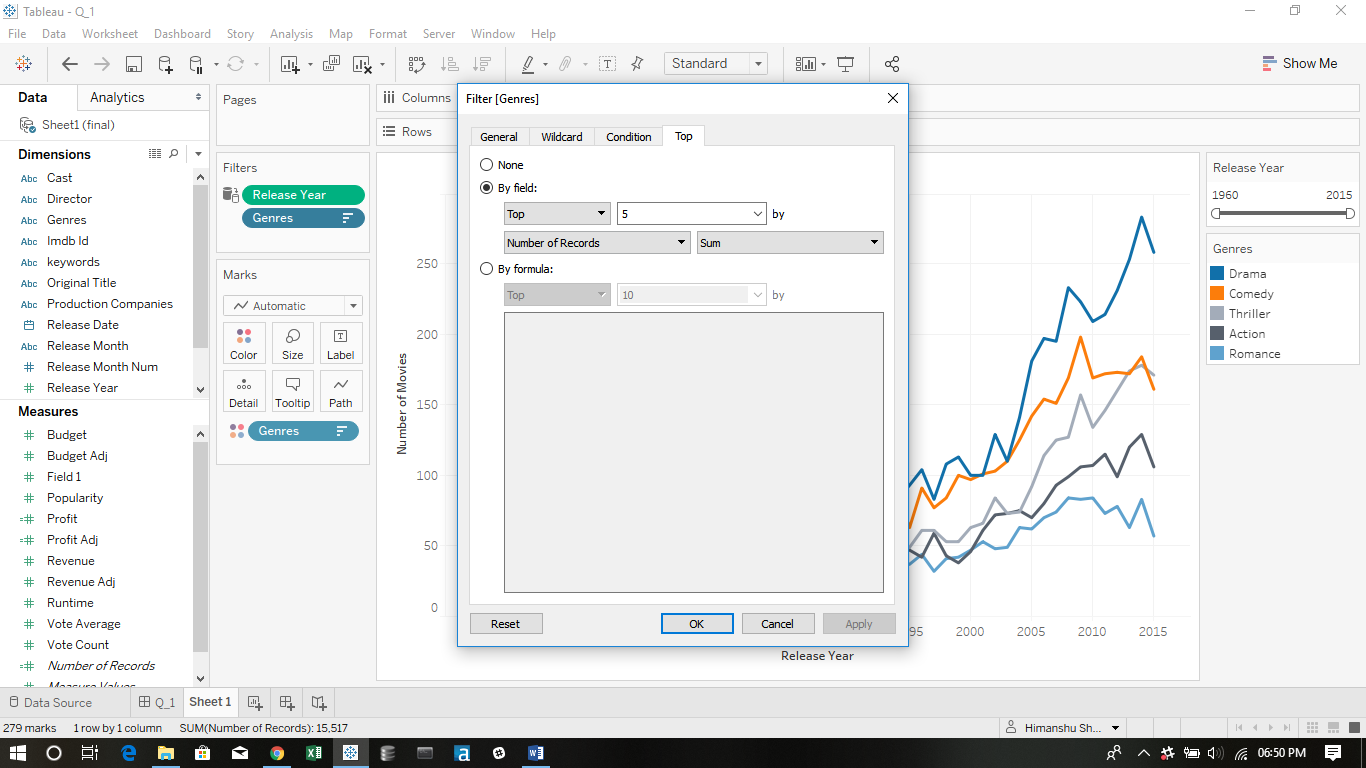
* I have followed the rubric instructions and attached all the files in the zip format and the links of my public dashboards is as follows:-

**Question -1**

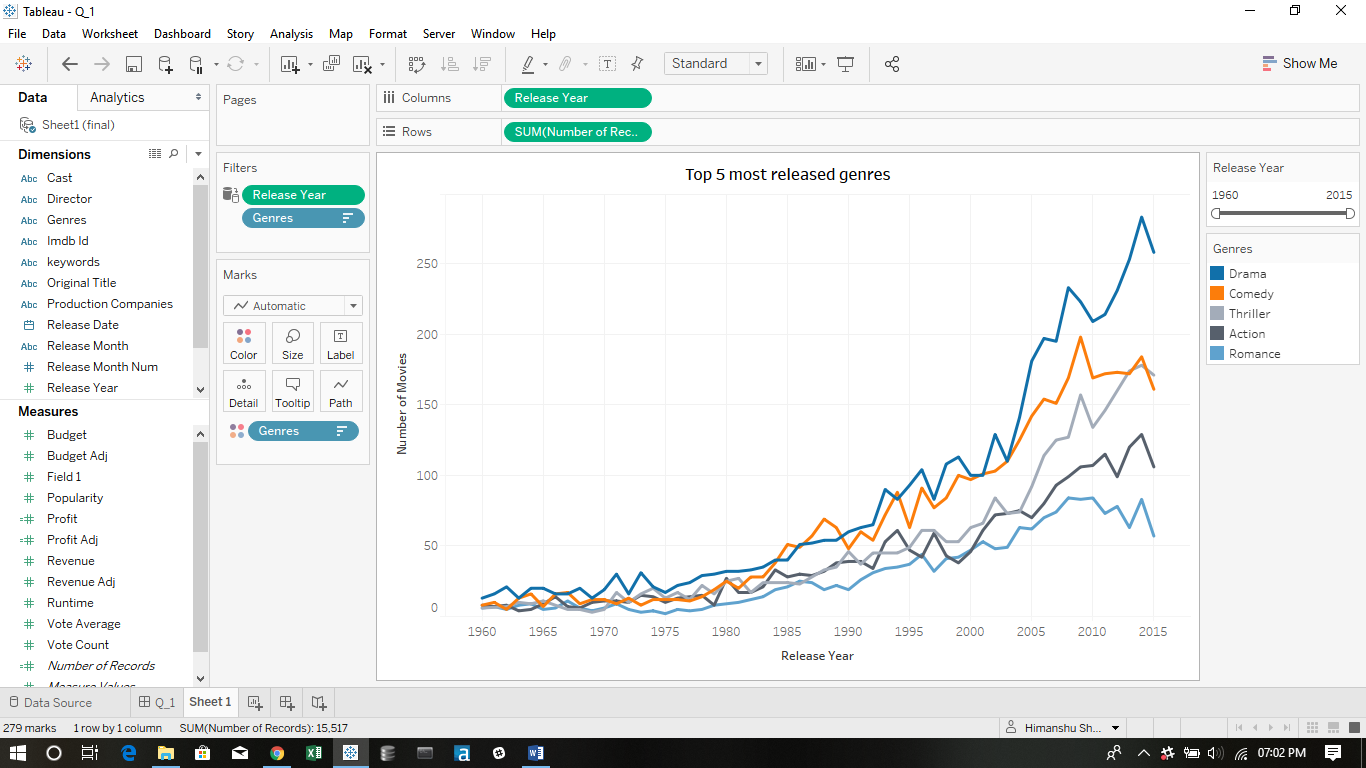
* <https://public.tableau.com/profile/himanshu.sharma3138#!/vizhome/Q_1_4/Q_1?publish=yes>

Concepts Used:-

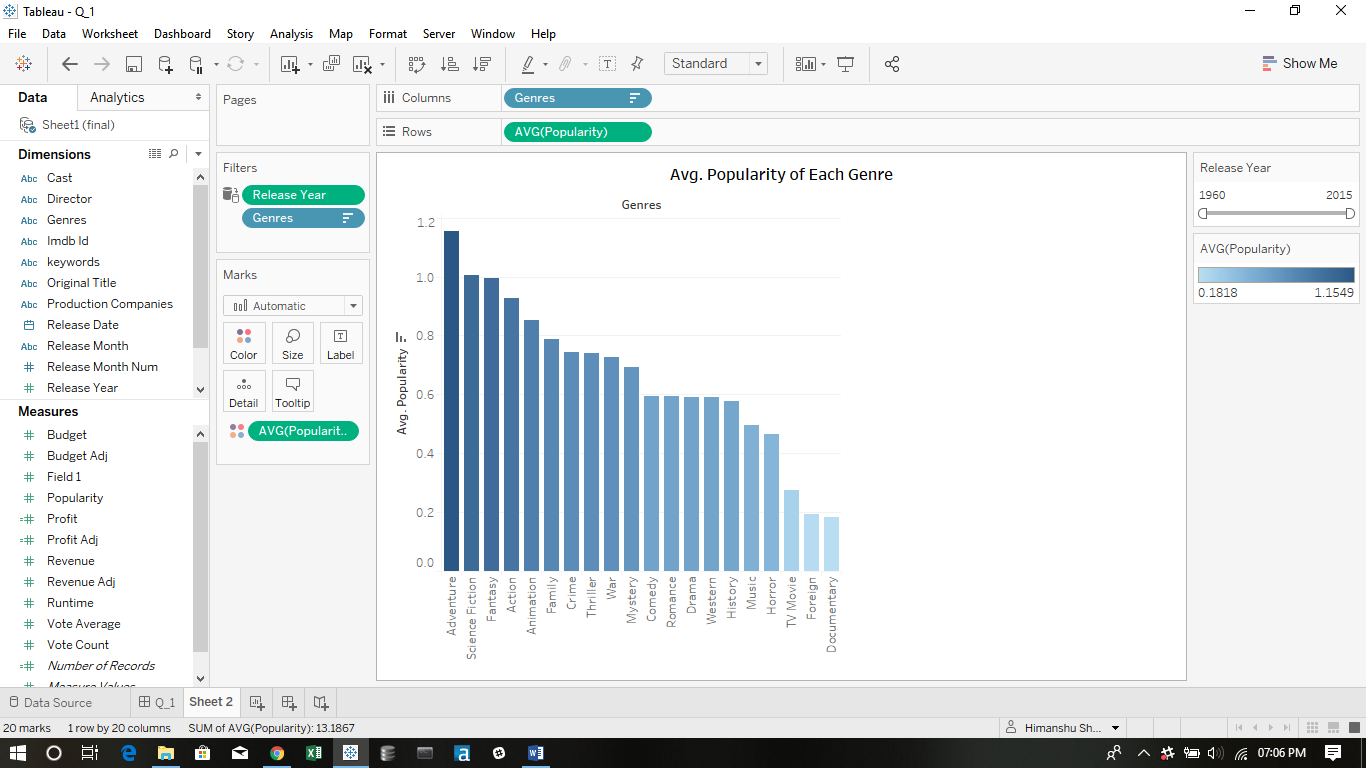
* Visualisation 1
* I have filtered the genre dimension by top 5 values according to the sum of total number of records.
* Then I have plotted the line graphs on a single visualization between the number of records and release\_year.
* I have applied the colour formatting in the marks table on the basis of genres.
* I have provided a filter for the release\_year .



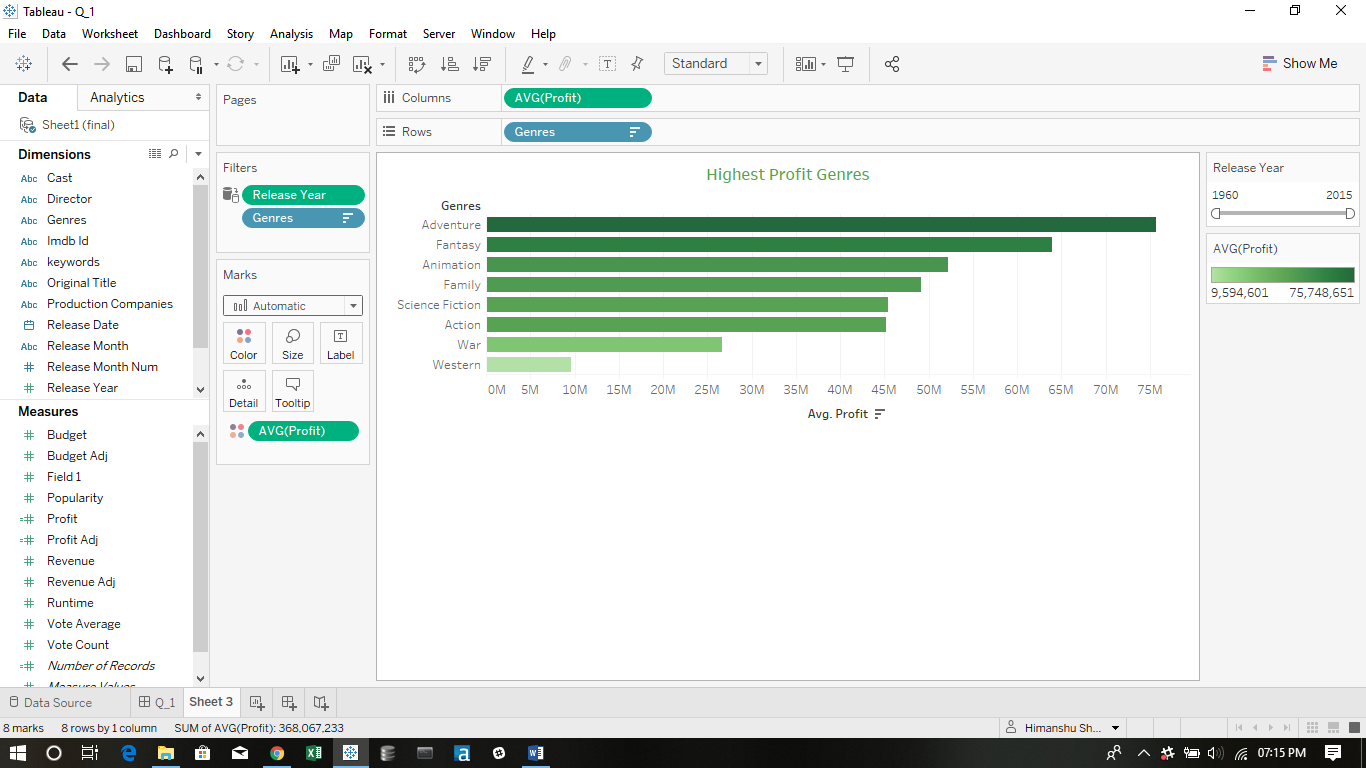
* Final visualisation



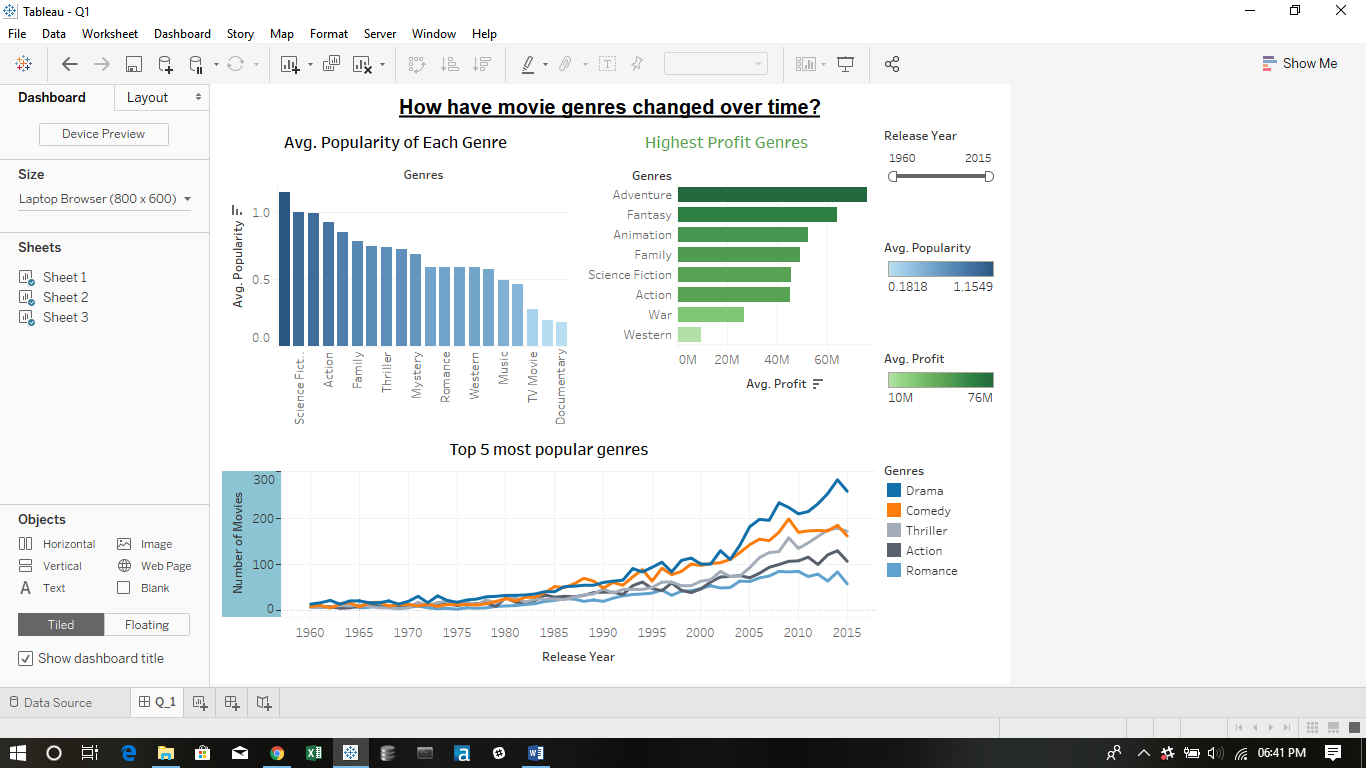
* Visualisation 2
* I have plotted a bar graph between the average(Popularity) and genre.
* I have given a filter for release\_year.
* I have also done colour foramatting on the basis of Average popularity,i.e, more the average popularity, darker the colour.
* Final Visaulization :-



* Visualizatin 3:-
* I have made a new calculated field ‘Profit’ by applying the formula ‘revenue’-‘budget’.
* I have drawn horizontal bars between average profit and genres.
* I have given a filter for release\_year.
* I have also done colour foramatting on the basis of profit ,i.e, more the profit, darker the colour.
* I have shown only the top most profiting genres using the filter by condition.
* Final visualization :-



* Final Dashboard for Q\_1



**Question -2**

* <https://public.tableau.com/profile/himanshu.sharma3138#!/vizhome/Q2_234/Q_2?publish=yes>

Concepts Used-

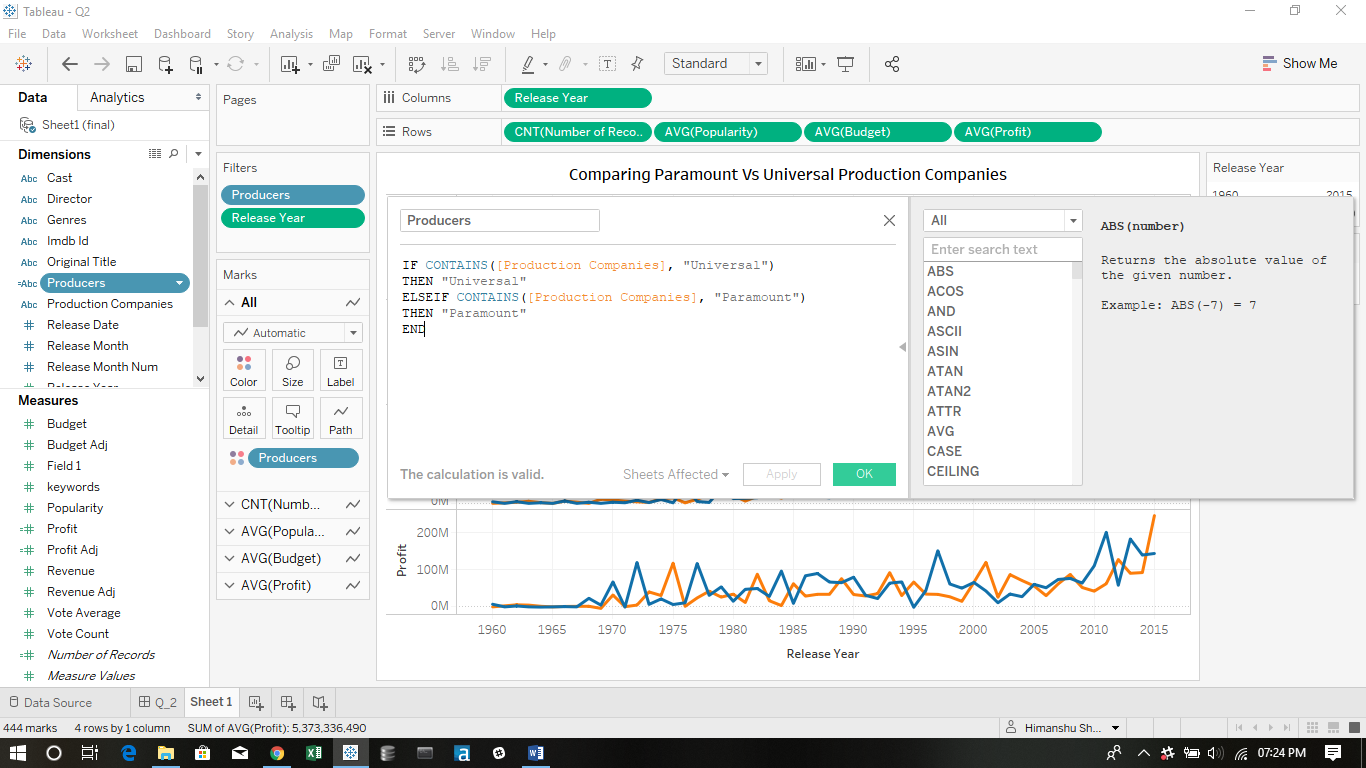
* Visualization 1-
* I have made a new calculated field ‘producers’ which on the following formula .

IF CONTAINS([Production Companies], "Universal")

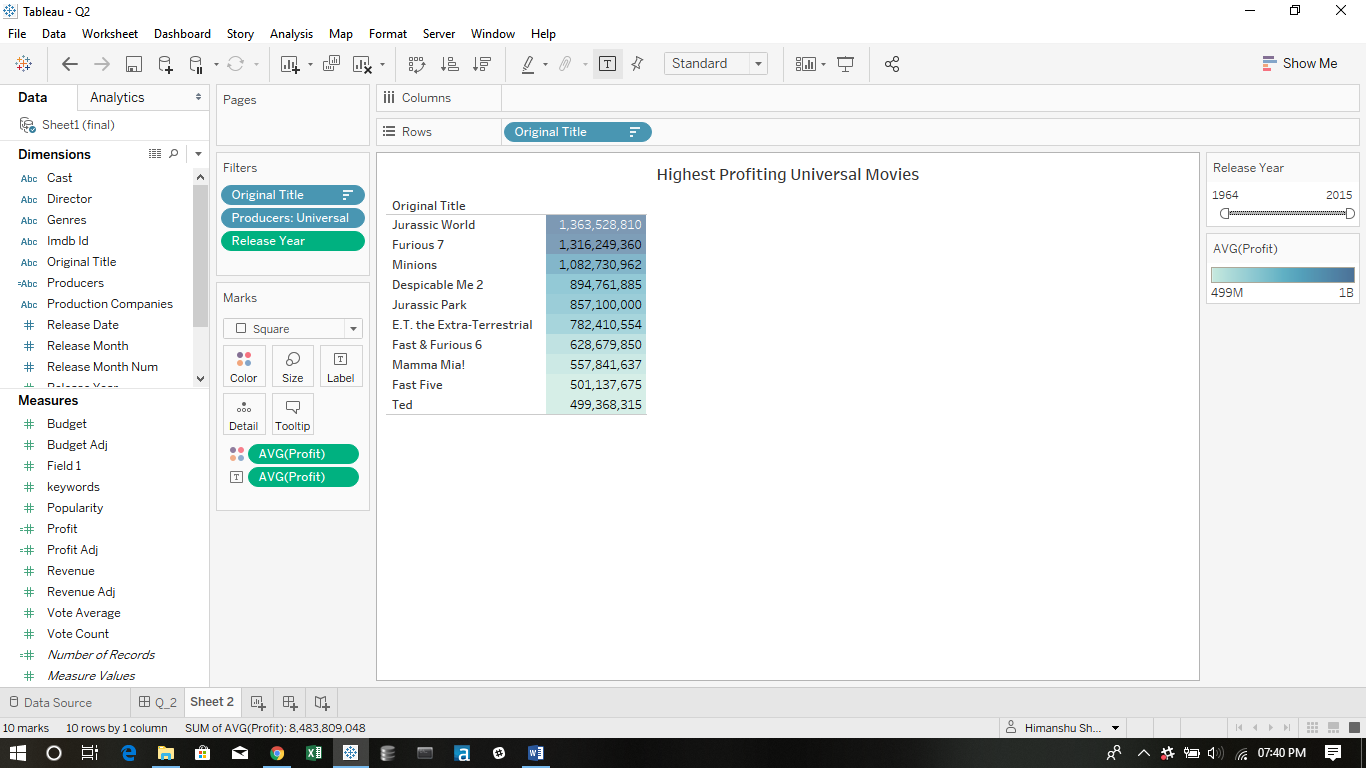
THEN "Universal" ELSEIF CONTAINS([Production Companies], "Paramount")

THEN "Paramount"

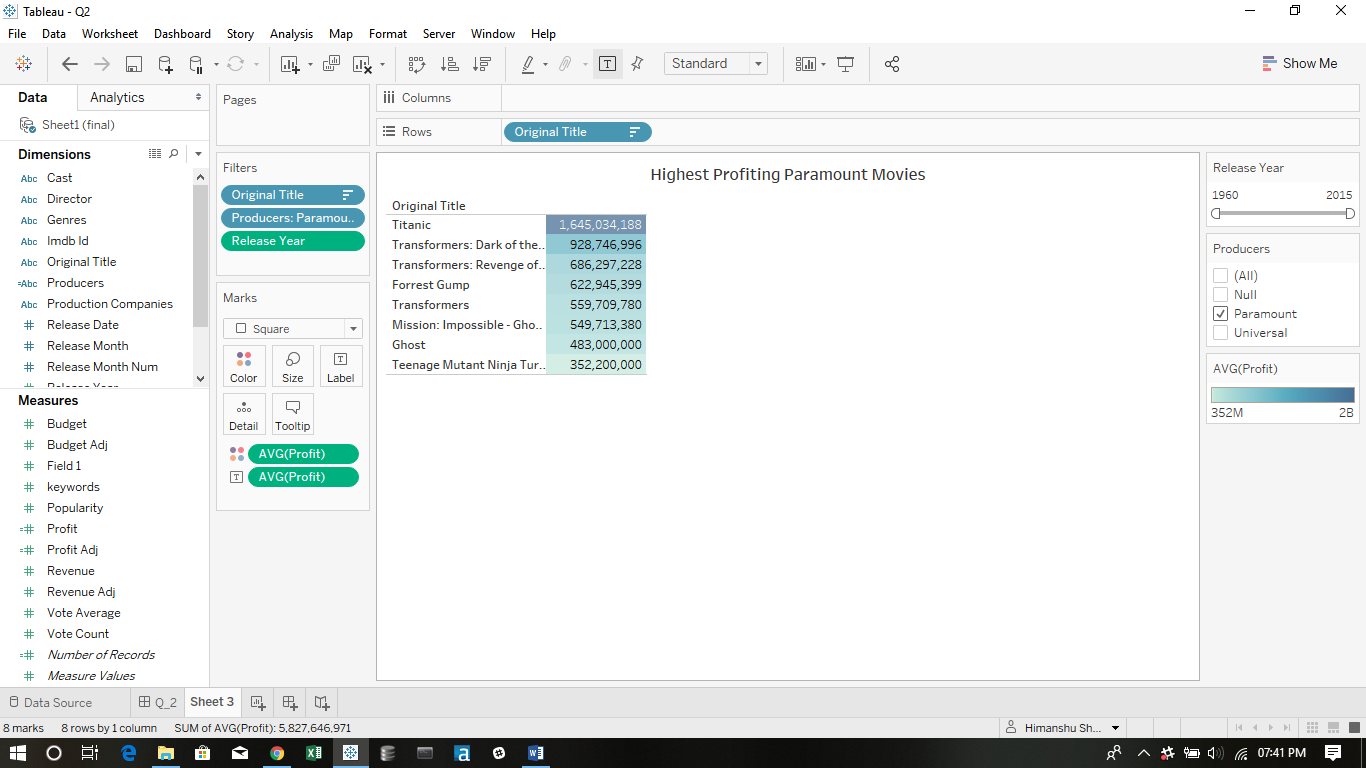
END



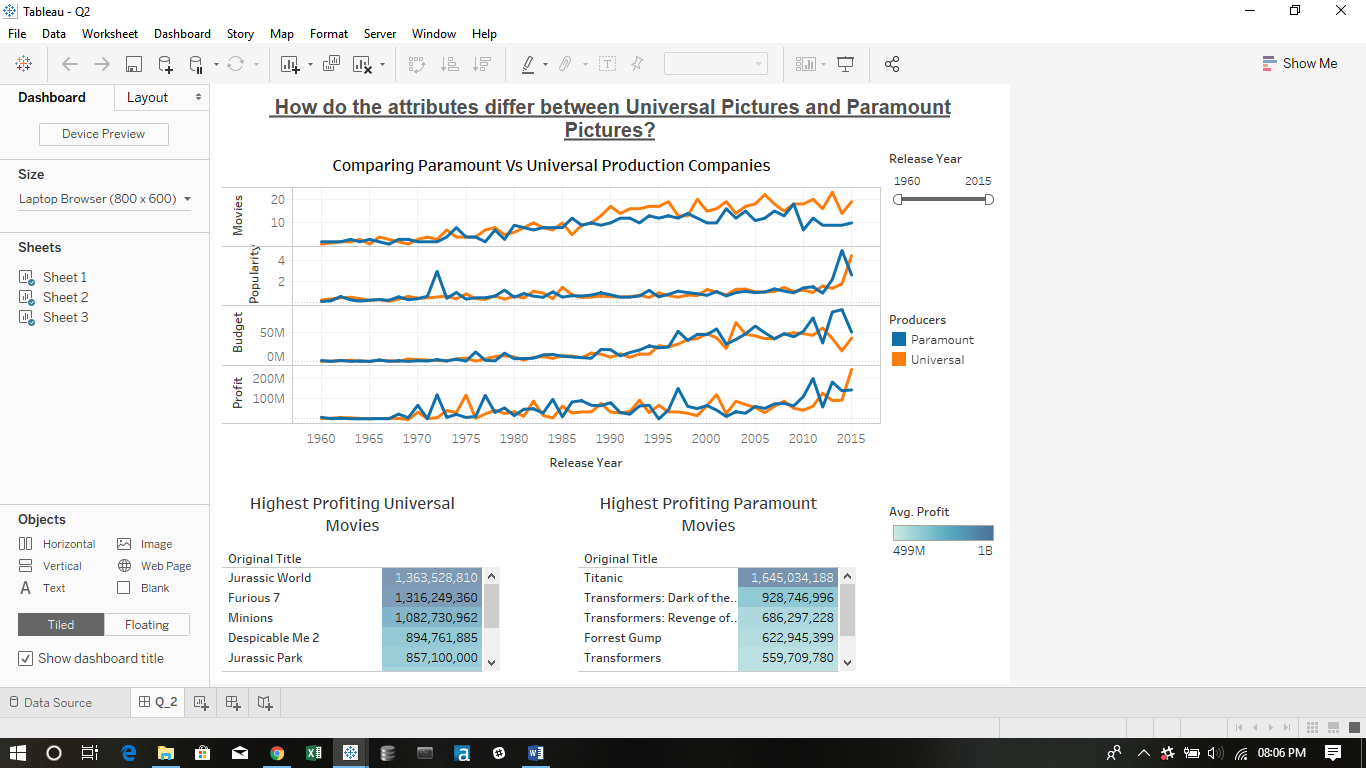
* Drawn lines graphs with dual axis method for release year against number of records, average\_popularity , average\_budget and average\_profit.
* I have provided a filter for release year
* I have edited the colour formatting in the marks table and assigned the ‘color blind’ pallete .
* Data for paramout production is shown in ‘blue’ colour and data for universal productions is shown in ‘orange colour’.
* Visualization 2
* For this I have drawn a text table for Original title and avg(profit) .
* I have filtered the movies by dragging producers column to the fiters table and selecting ‘Universal’ from the values shown.
* Now,I have applied an condition to show the list of Universal movies that are in list of Top 100 highest profitable movies of all time .
* I have provided a filter for the release\_date.
* I have done the colour formatting on profit column on the basis of avg(profit).



* Visualization 3
* For this I have drawn a text table for Original title and avg(profit) .
* I have filtered the movies by dragging producers column to the fiters table and selecting ‘Paramount’ from the values shown.
* Now,I have applied an condition to show the list of Paramount movies that are in list of Top 100 highest profitable movies of all time .
* I have provided a filter for the release\_date.
* I have done the colour formatting on profit column on the basis of avg(profit).



* Final Dashboard –

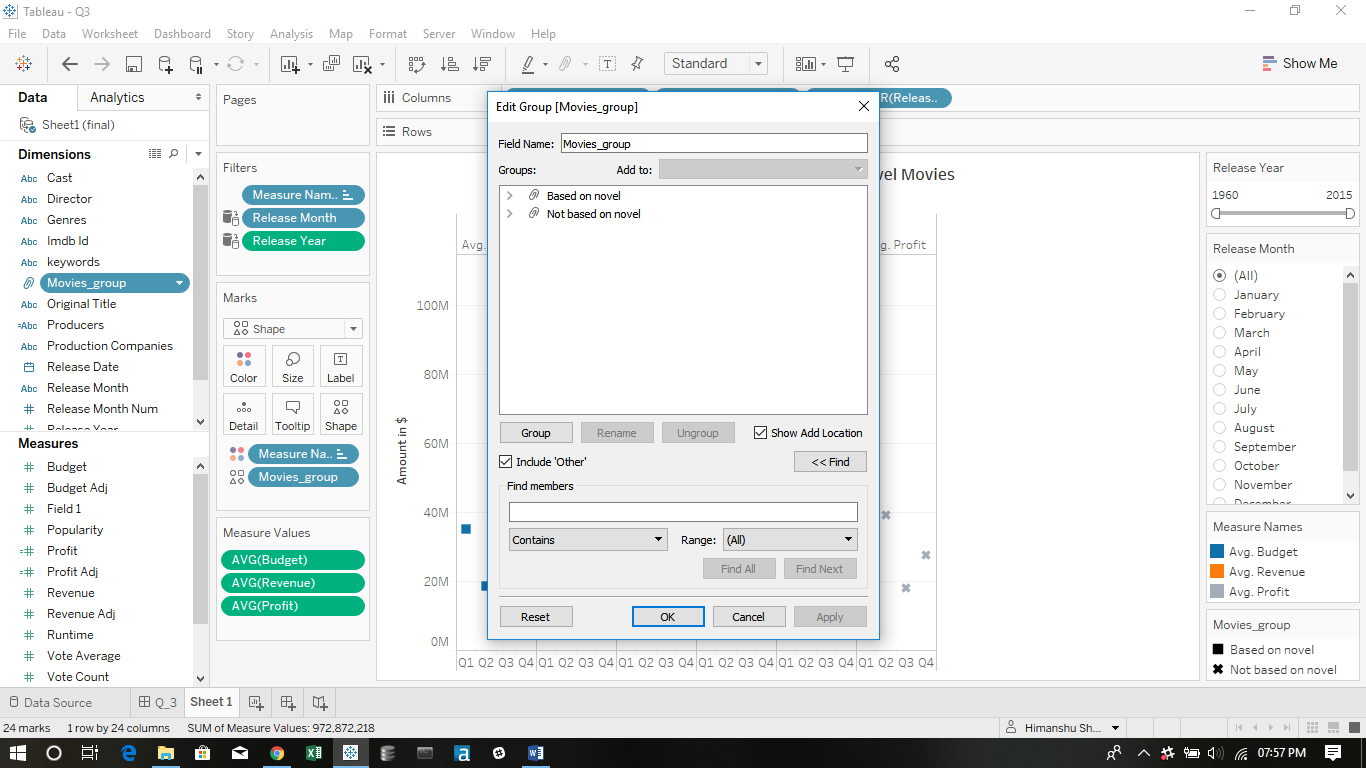


**Question -3**

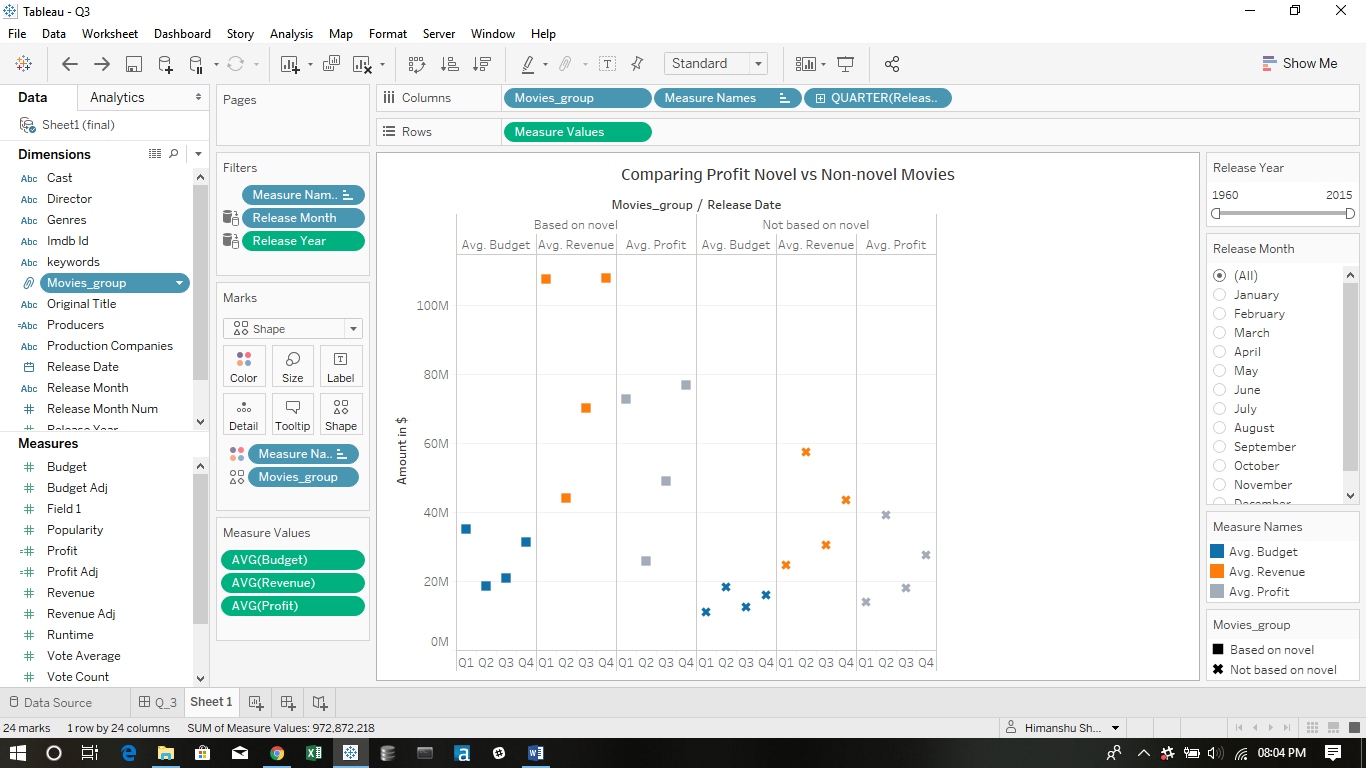
* <https://public.tableau.com/profile/himanshu.sharma3138#!/vizhome/Q3_224/Q_3?publish=yes>

Concepts Used-

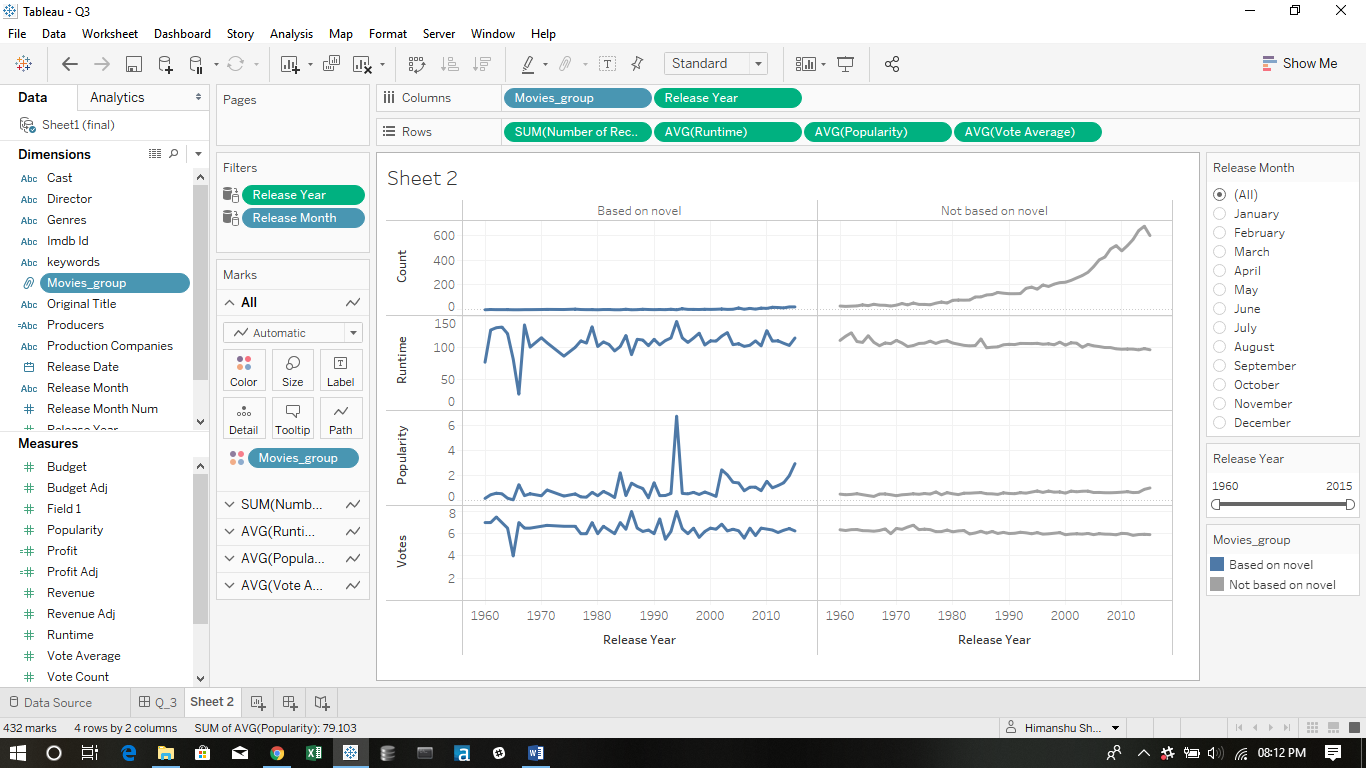
* Visualization 1-
* I have grouped the keywords column into two groups and named the field ‘Movies\_group’ using the condition string contains and named the groups as based on novel and not based on novel.



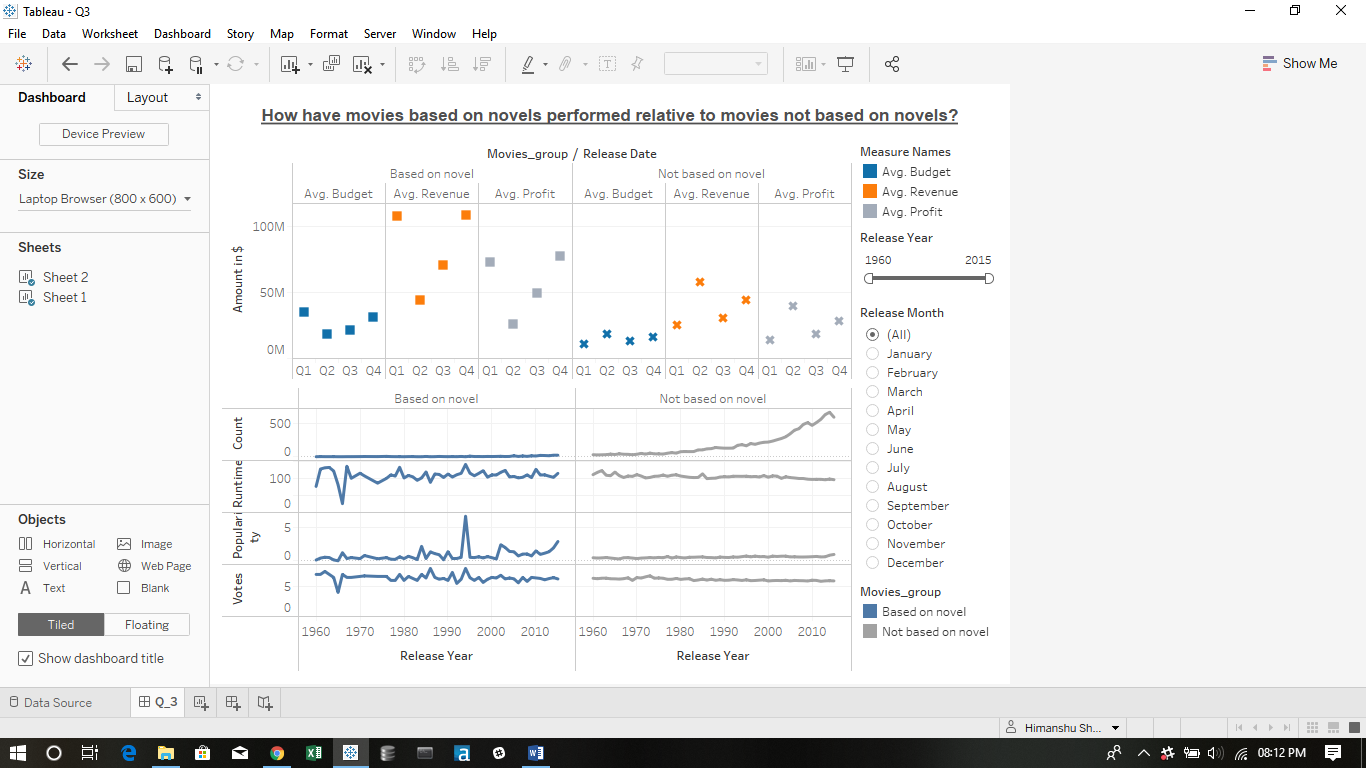
* I have plotted the lines(discrete) plot for the movies\_group column againt the avg(budget) , avg(revenue) and avg(profit).
* I have changed the release date to continuos and dragged it to columns and drilled down to year>>quarter and then removed the year measure.
* I have given colour formatting on the basis of measure names.
* I have given shape formatting on the basis of movies\_group.
* I have provided a filter for release date and release month.
* Small multiple plots were derived after that.



* Visualization 2
* I have plotted the lines(continuous) plot for the total number of records , average runtime , average popularity and average vote\_count against the release\_year for each movie\_group.
* I have provided a filter for release date and release month.
* I have coloured the line plots on the basis of movies\_group.



* Final dashboard for question 3

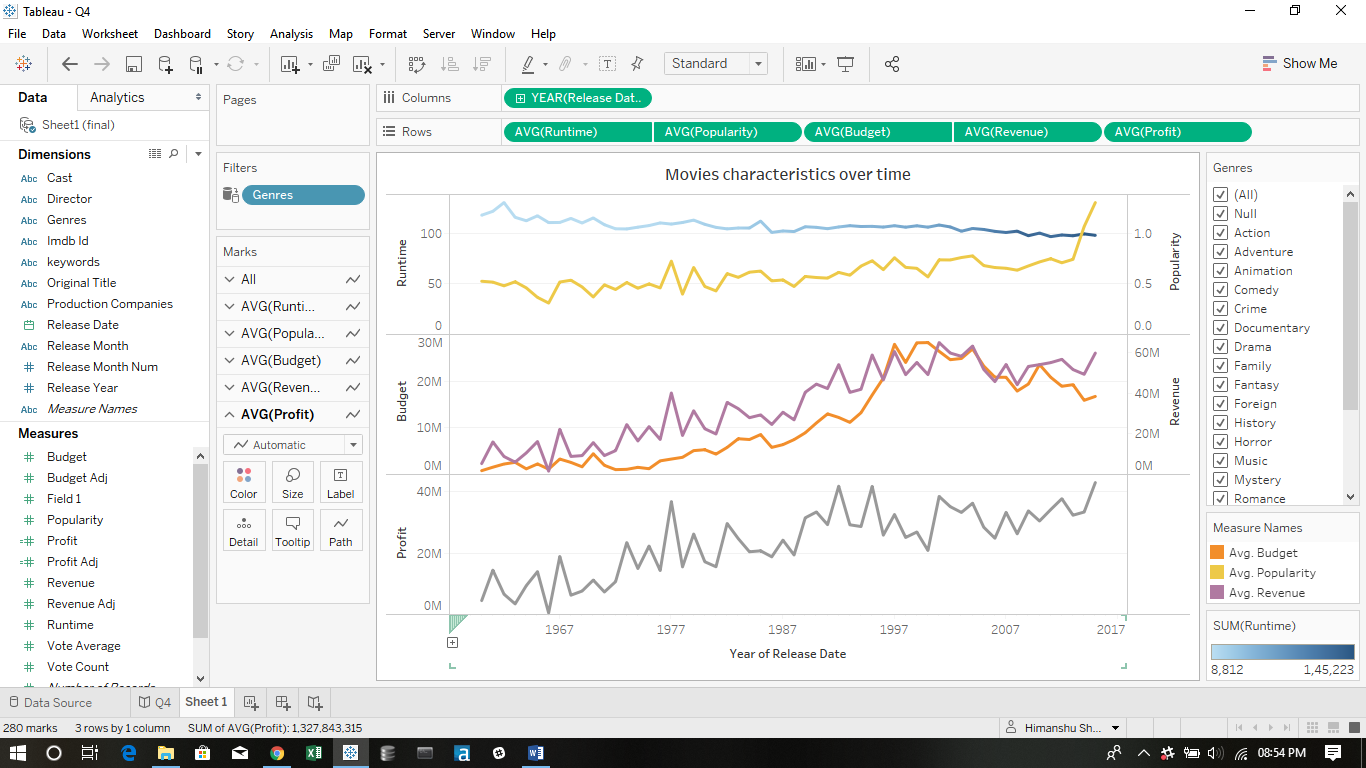


**Question -4 (Have the movie trends changed over time ?)**

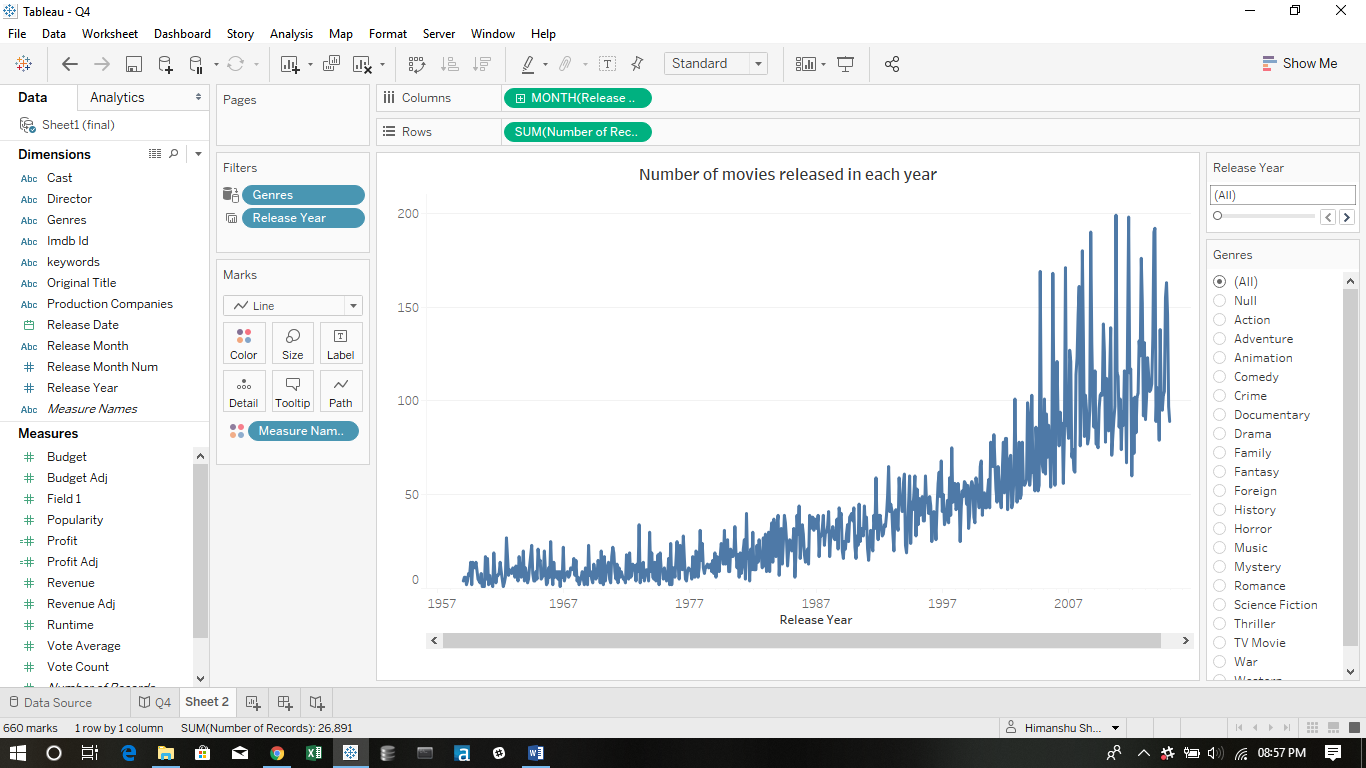
* <https://public.tableau.com/profile/himanshu.sharma3138#!/vizhome/Q4_197/Q4?publish=yes>

Concepts Used-

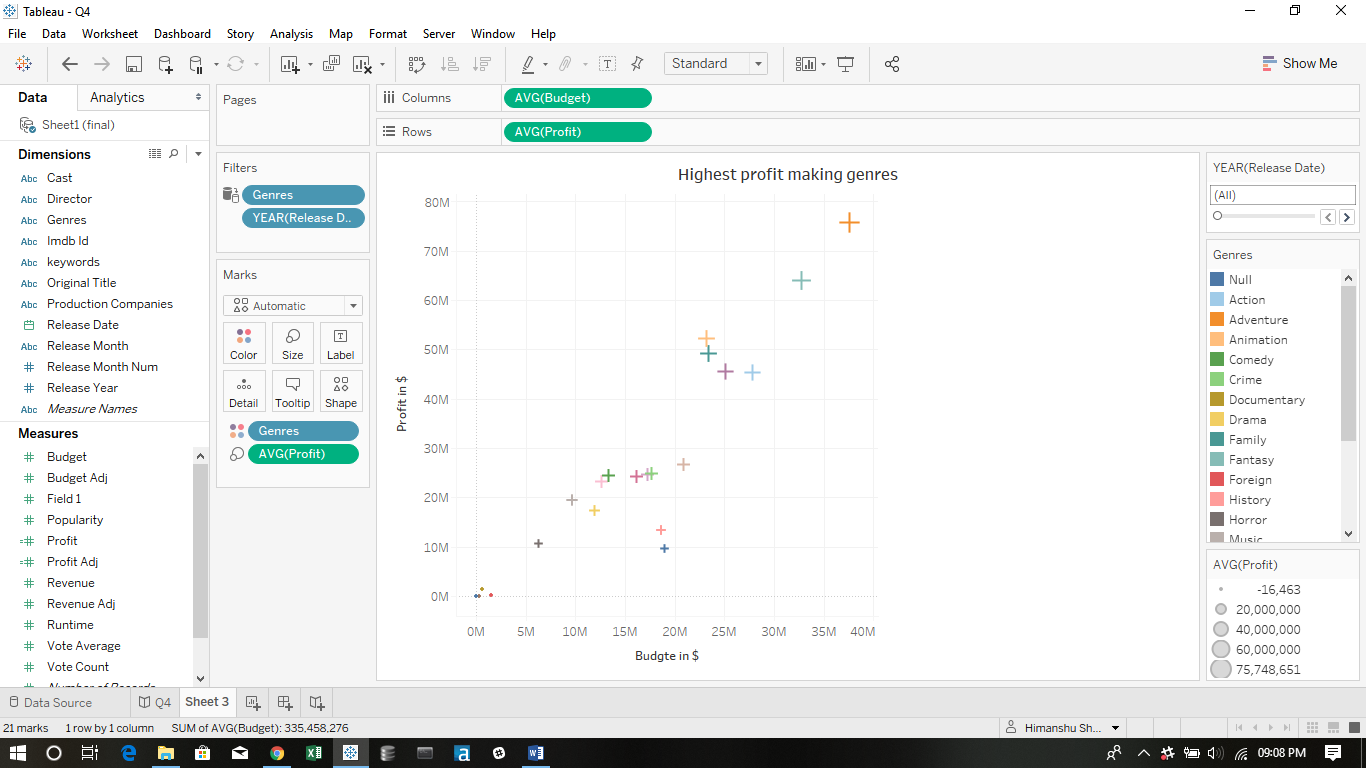
* Visualization 1 –
* I have plotted a dual axis line graph for average budget and average budget over time.
* I have also plotted a dual axis graph for runtime and popularity over time.
* I have plotted a line graph for the profit against release years.
* I have given filters for genre\_type.



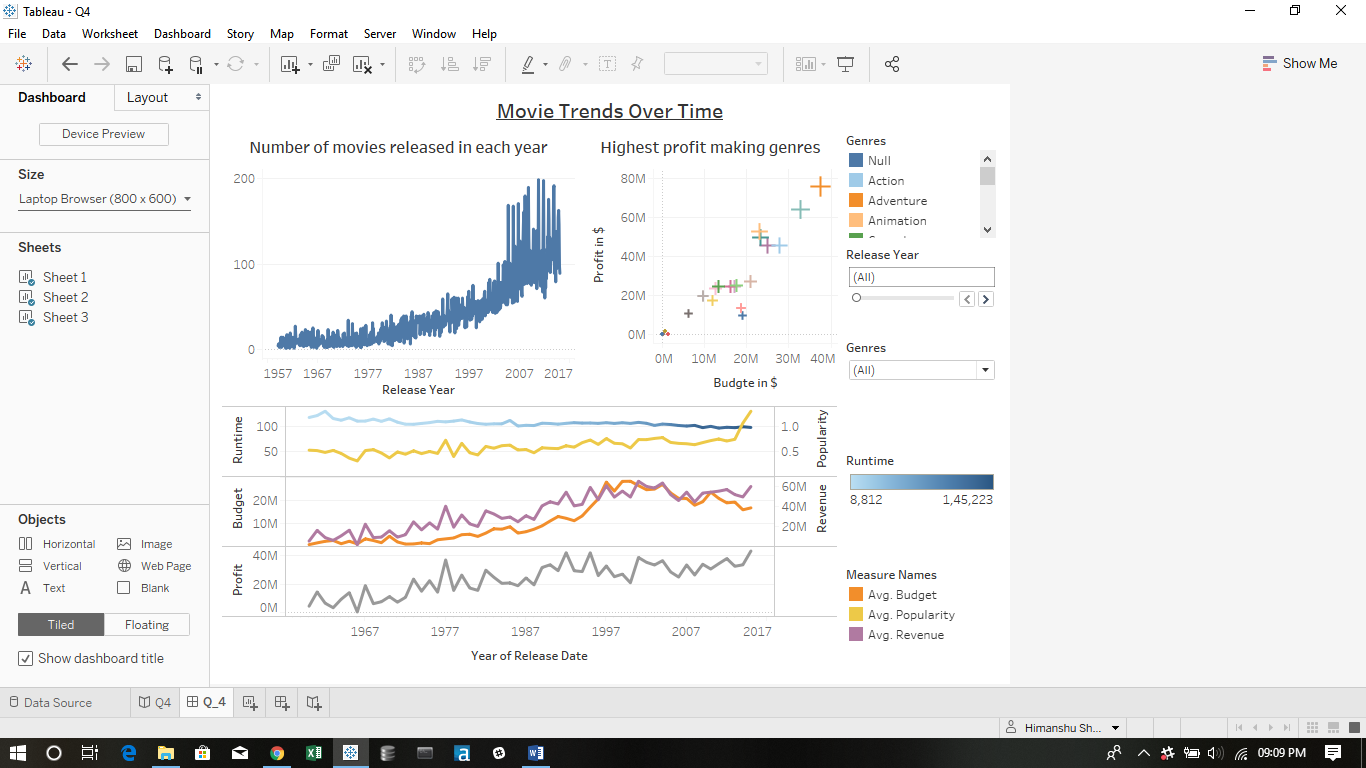
* Visualization 2
* I have changed the released date to continuous and added it to columns.
* I have drilled two levels to months.
* I have plotted the number of movies released over by years by plotting against number of records.
* I have provided a filter for genre\_type.
* I have provided a filter for release\_year.



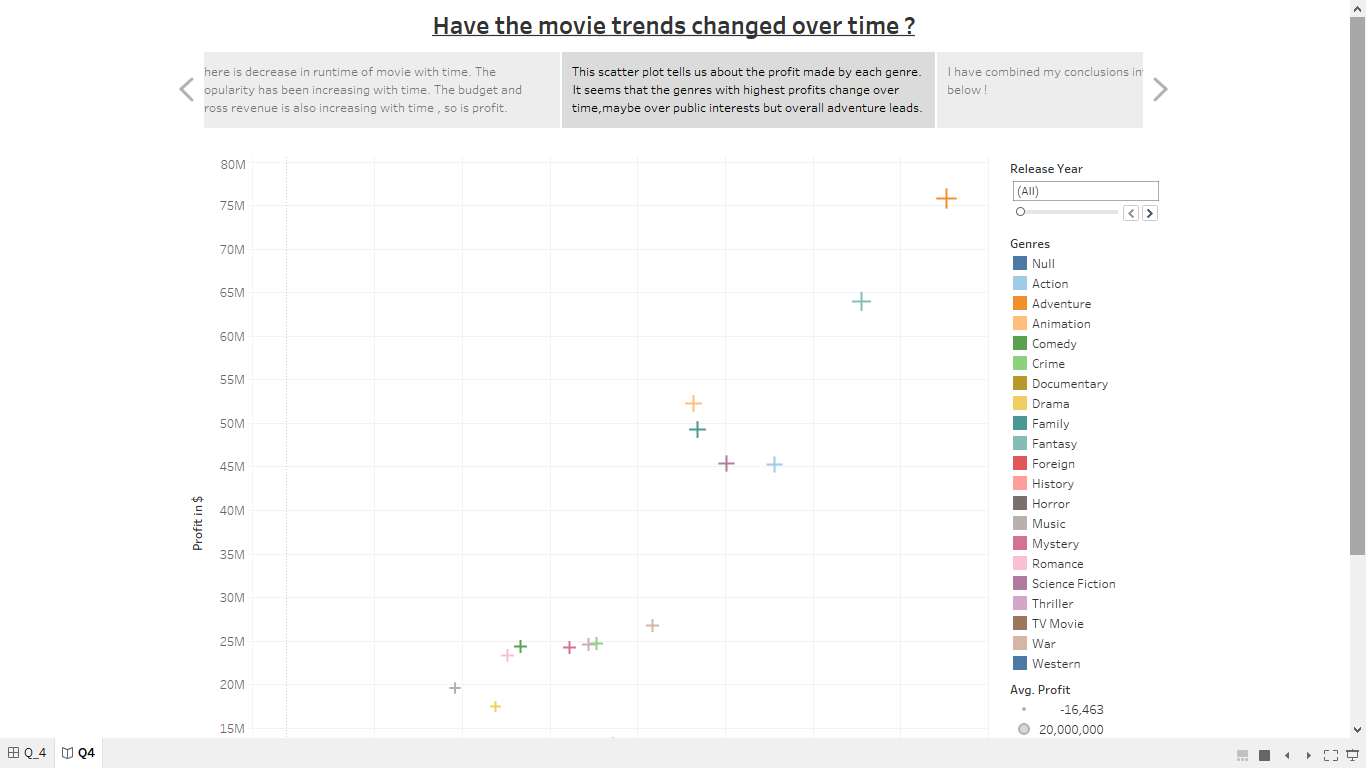
* Visualization 3
* I have drawn a scatter plot between average\_profit and average\_budget.
* I have filtered the data by genres.
* I have provided a filter for release\_year and genre\_type.
* I have given colours with respect to genre\_type.
* I have given size with respect to profit amount.



* Final dashboard-



* I have made a story for this question



## Step 3: Questions

* + **Question 1:** How have movie genres changed over time?
* Yes, the movie genres have changed over time .
* Earlier most popular genres were family, TV movie and action but nowadays adventure, sci fi , action and western genres are more popular.
* In the 1960’s most profitable genres were animation, adventure and family. These years almost the same genres are the most profitable but also including action, sci fi and fantasy.
* Comparing with the rest of genres ,movies of drama, comedy , thriller , action and romance are produced the most from the 1980’s.
  + **Question 2:** How do the attributes differ between Universal Pictures and Paramount Pictures?
* From the 1980’s Univeral production company is producing more movies than paramount.
* There is a tough competition for popularity and profit between universal and paramount production companies.
* In the last 3 to 4 years paramount has produced much high budgeted movies as compared to universal.
  + **Question 3:** How have movies based on novels performed relative to movies not based on novels?
* The popularity of novel based movies has rapidly increased from the year 2000.
* The number of movies based on novels are very few these years as compared to movies not based on movies.
* These years novel based movies have average runtime of 115 minutes and not novel based have 95 minutes.
* The average votes to both the categories are almost same.
* The movies based on novels have always grossed greater revenue and so have made greater profits then movies not based on novels.
* What is your additional question that you proposed? What is the answer? How did you come up with this question?
* My question is ‘**Have the movie trends changed over time ?**‘

According to the visualizations-

* The number of movies released in a year are increasing year by year .
* The average runtime of movies has decreased from 117 minutes to 97 minutes. It seem that people don’t like too lengthy movies.
* The popularity of moves have increased from the year 2000.
* The revenue of the movie is already increasing year by year.
* High budgeted movies are being made from the 1980’s.
* The profit has been increasing year by year because of the revenue increase.
* In the 1960’s top profit making genres were western, history , war and adventure. But now the trend has changed. Top grossing genres these years are adventure, action, fantasy, family , western and animation.
* In the 1960’s movies average profit was 10-15 million $ but movies have started gaining profit more than 100 million $.