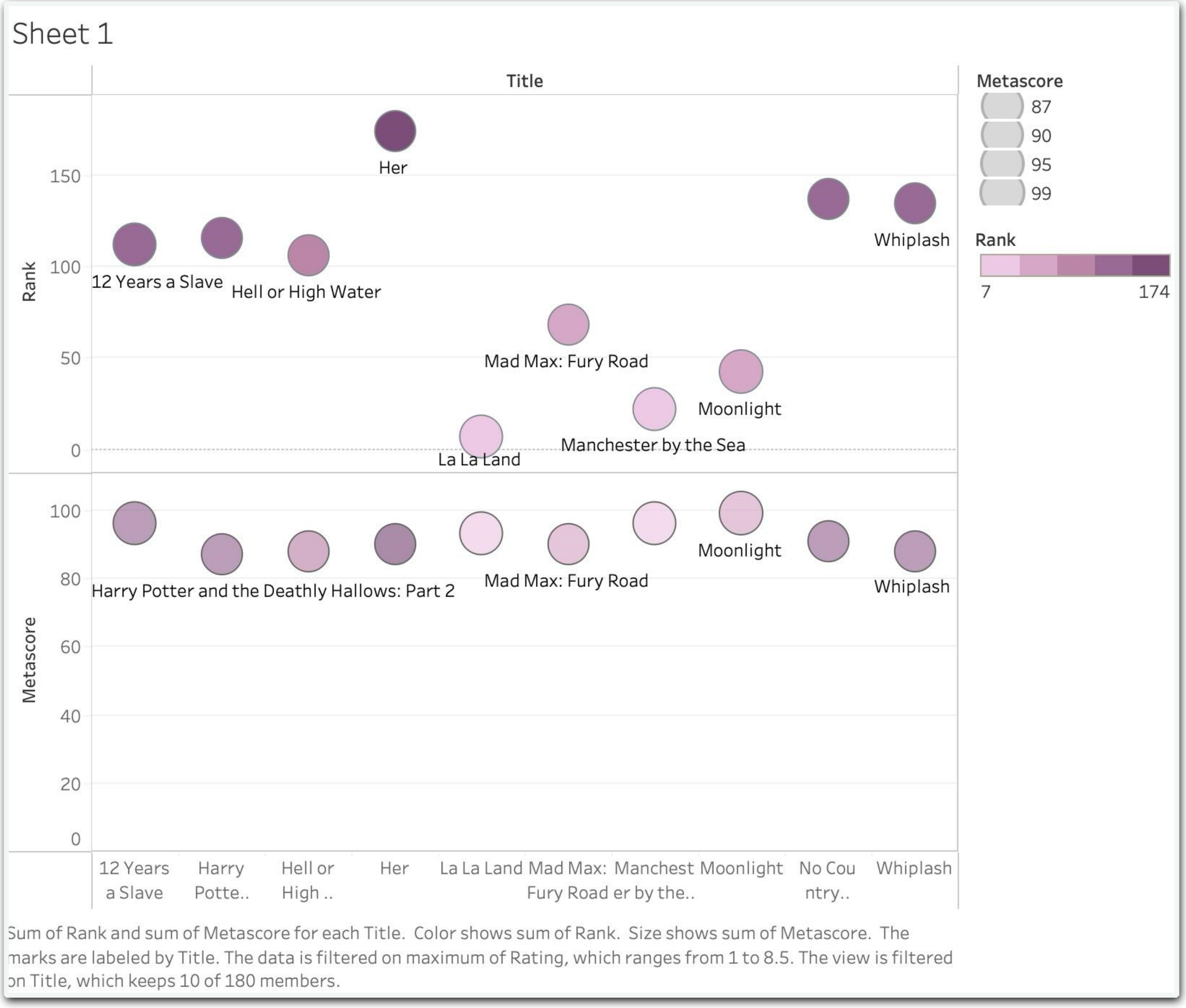
<u>PROJECT TASK</u>: Extract a dataset from GitHub and analyse top 3 questions with reference to tableau tool.

<u>Source</u> : This movie mashup dataset was compiled by me from GitHub. The data was gathered from about 1000 movie rows and then it was imported into the tableau.

Source URL : **<u>https://github.com/laxmimerit/All-CSV-ML-Data-FilesDownload/blob/master/IMDB-Movie-Data.csv\</u>**

- Title, Rank, Genre, Actors, Year, Runtime, Votes and Meta score are the variables and rows of the dataset respectively.

Question 1 : Which movies uphold the top 10 position based on the Meta score element trailing by the ranking criteria ?



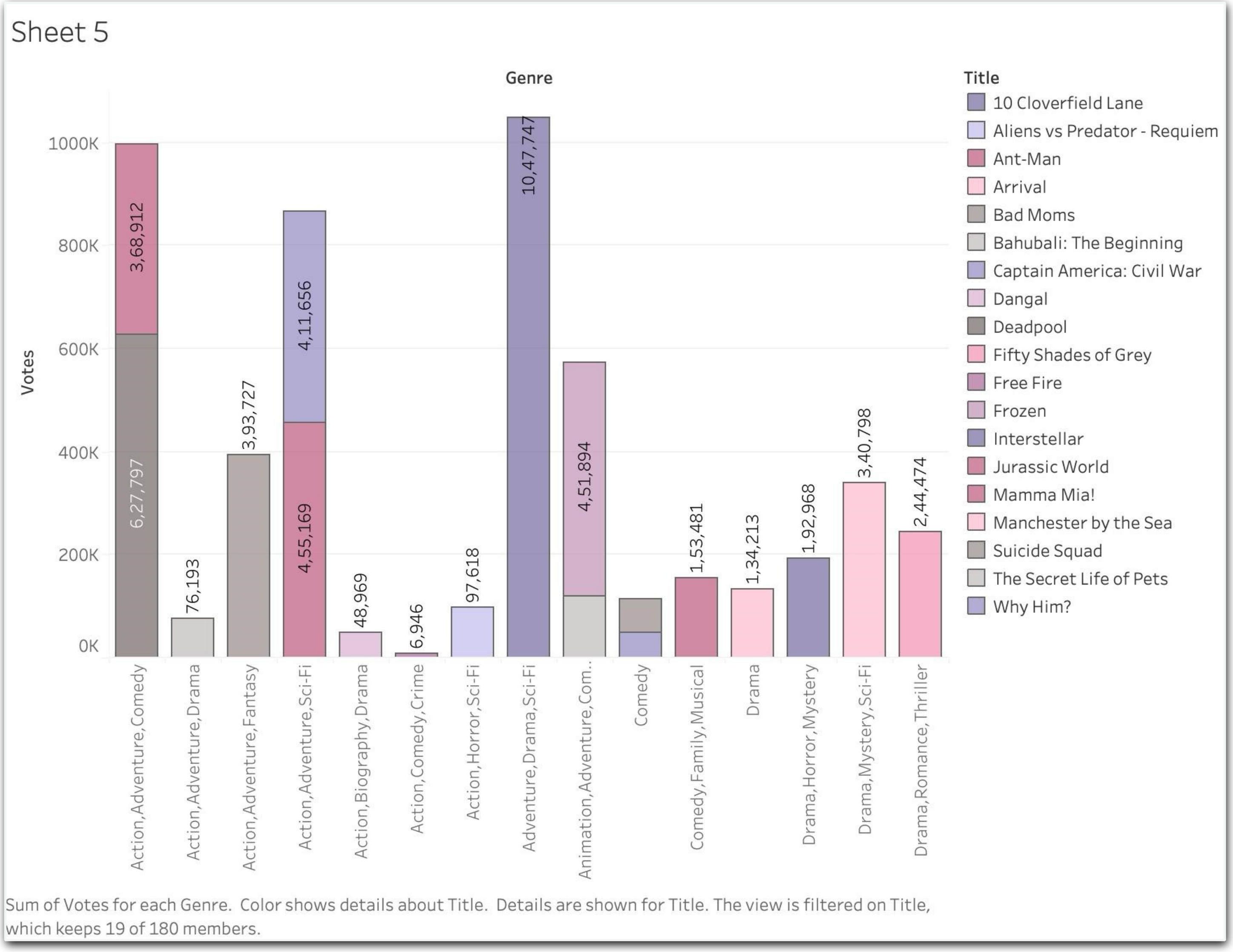
Sum of Rank and sum of Metascore for each Title. Color shows sum of Rank. Size shows sum of Metascore. The marks are labeled by Title. The data is filtered on maximum of Rating, which ranges from 1 to 8.5. The view is filtered on Title, which keeps 10 of 180 members.

Answer : The above scatter plot visualization depicts the top 10 movie meta score ranging on a scale of 1 to 8.5 (It represents as 0 to 100 on the graph for better understanding) along with the same 10 movies ranking plot placed on top of it. These 10 movies dataset is filtered from a 180 movies mash up dataset and surprisingly, few movies that last with top meta score didn’t abate with the ranking criteria. For instance, the movie Moonlight shows the highest meta score level with a

range value of 99. But, when we look into its ranking , the movie is levelled off on a lowest scale with the rank of 42.

Likewise , the few other movies which comes under this criteria (high meta score and modest rank) includes : La La Land , Mad Max , Manchester by the sea, moonlight and it is specified by light color dots. Lastly, the maximum value in terms of rank is backed up by “Her” movie along with inflated value of meta score of 174 and 90 respectively. Such measures are indicated by dark color dots. All the remaining consecutive films rely on the similar edge with reference to ranking and meta score and this divergence can be clearly assessed with such scatter plot visualization.

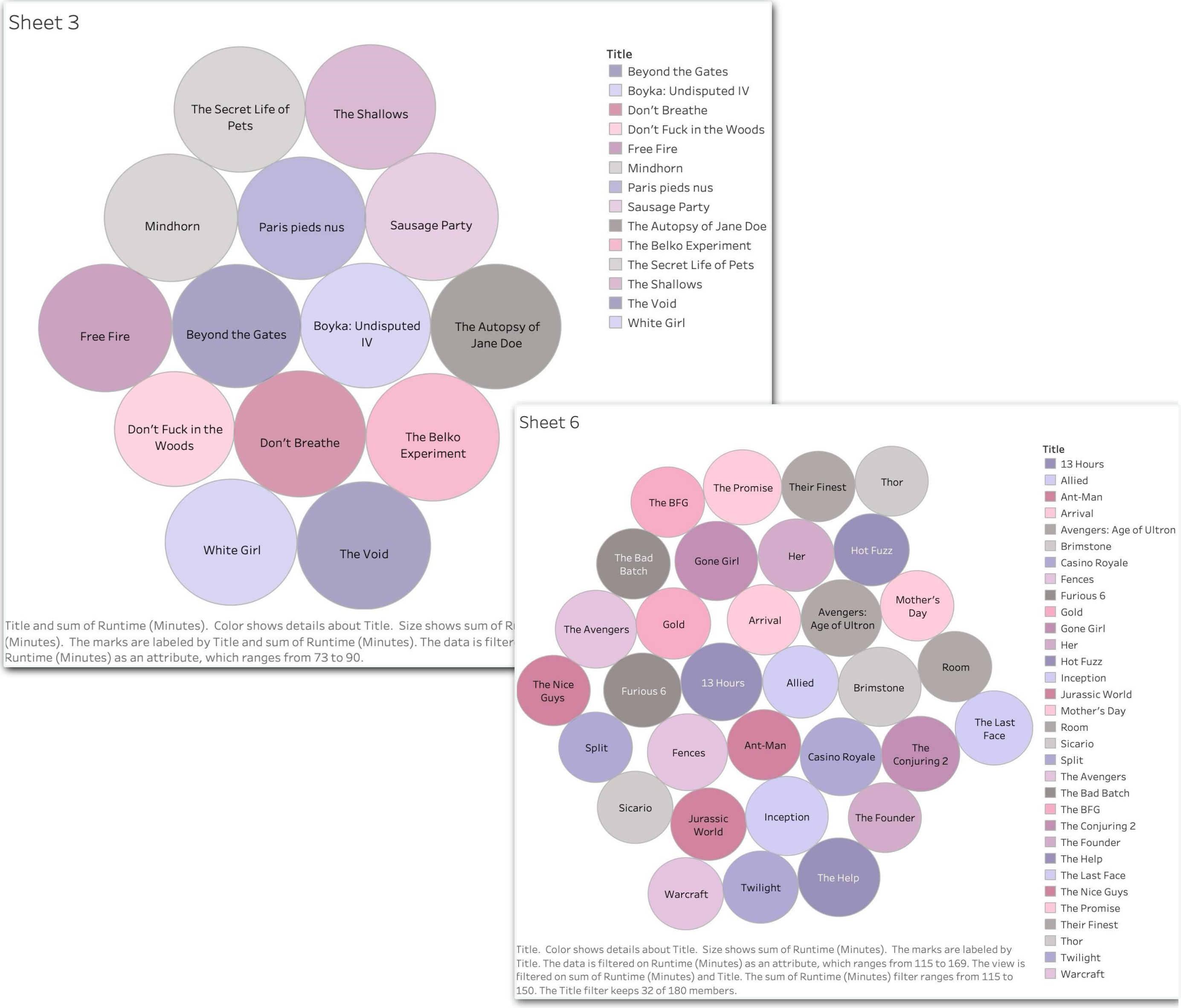
Question 2 : What is the distribution of votes across different genres based on their respective titles?



Sum of Votes for each Genre. Color shows details about Title. Details are shown for Title. The view is filtered on Title, which keeps 19 of 180 members.

Answer : This chart portrays a horizontal bar graph visual representation stating movie genres in accord with voting category. The above graph demonstrates more than 1 movie title if they fall under the same genre hence, the differentiation is outlined by total number of votes. This visualization gives a brief sense about most of the population\ people's interest in terms of genre. Yet, the accurate level of rendition is not elevated because of merging the name of the movies within the bar chart. Nonetheless, the color part of this graph stands as an highlighting point , as they are allotted in contrasting colours of the same shade so as to recognise which movie holds up with how many number of votes. On observing the grid , one can clearly predict that most of the people would love to binge watch movies that contains "Adventure , Drama , Sci-Fi" elements as this genre retains with highest number of votes (10,47,747). Under this genre it includes only one movie name according to the movie mash up dataset I casted, and it is "Interstellar". Moving to the med line level of the votes it distributes the votes among the ratio in between 200K to 650K with genres - Action , comedy , adventure and mystery and in this level it encompasses more than 1 movie title as mentioned above. Some films that comes underneath this level are: "Bad Moms, Arrival ,Deadpool , Captain America , Frozen , Manchester by the sea". In the end the movies that are inclined towards low number level are : "Bahubali , Dangal ,Mamma Mia , Why him and few others". Genres of such movies are Drama , Crime and musical in which people are barely interested. Therefore, the sum of voting percentage of such genre films are also on the downturn values from 6K to 1,53,481.

Question 3 : What is the gradient of movie runtimes in a dataset illustrated by two bubble packed charts ?



Title and sum of Runtime (Minutes). Color shows details about Title. Size shows sum of R
(Minutes). The marks are labeled by Title and sum of Runtime (Minutes). The data is filter
Runtime (Minutes) as an attribute, which ranges from 73 to 90.

Title. Color shows details about Title. Size shows sum of Runtime (Minutes). The marks are labeled by Title. The data is filtered on Runtime (Minutes) as an attribute, which ranges from 115 to 169. The view is filtered on sum of Runtime (Minutes) and Title. The sum of Runtime (Minutes) filter ranges from 115 to 150. The Title filter keeps 32 of 180 members.

Answer : The two packed - bubble schematic representations above unveils the segregation of movies with run times of less than 100 minutes and greater than 100 minutes to 170 minutes. This plot has been created using an 180 rows of movie dataset and transformed into these nested circles with different colours , so as to get a general picture of movies length time along with its title. Out of 180 data columns the tool filtered 14 movie sets that account for the duration in between 73 to 90 minutes (presented in sheet 3). On the other hand, the bubbles representing in sheet 6 gives a view of crowded bubbles since most of the movies last more than an hour and the filter applied for this runtime ranges between 115 minutes to 170 minutes. Thought it was assimilated for more than 100 minutes , the data I opted doesn't include any movie with duration attribute between 100-114 minutes hence, this sheet inbuilt starts with the minimum duration of 1 hour 15 mins. There are 3 distinct colours to stipulate these runtimes and I added slight opacity for the colours as it contains additional movie sets in sheet 6. There are 6 bubbles of Gray ash color that indicate in between 115 to 139 minutes. Purple shaded bubbles prudent to the next time length range that is 141 to 150 minutes. Finally the pink color shaded bubble implicates the maximum run times (151 to 169 minutes) approaching the movies like "Arrival , Warcraft , Avengers , Conjuring 2". The same phenomenon is followed for sheet 3 that is Gray ash color bubbles indicate the initial or base period for instance in this chart it shows this color for movies : "The Autopsy of Jane Doe , The secret life of pets , Mindhorn" quoting 73 , 79 and 81 minutes. On the whole, this bubble chart relate a translucent contrasting feature for duration attribute.

CONCLUSION or SUMMARY:

- Tableau was my choice for constructing all my visualization tools for my project report because, it makes the task easier as it shows any vast amounts of data in a way that is both clear and intelligible and produces visually impressive dashboards and reports. In order to meet our demands, this software also provides various levels of customisation while designing graphics. It may not inherently handle all data sources, which might render difficulties in aggregating data from different sources, this limitation affects some particular tableau use cases. Additionally, several users remarked that it might require considerable effort and trial-and-error to use various capabilities, including developing customised computations.

- Addressing the very first visualization tool, the **Scatter Plot diagram**, one of this graph benefits is that, for a particular group of data it effectively displays the connection between two variables (in this example, movie ranking and meta score). To emphasise those specific movies, it can be useful to use various coloured dots to symbolise the movies that suit the great meta score/ modest rank criterion and those that fit the opposite requirement. In respect to user interface, the scatter plot has a clean and straightforward structure with both variables depicted on their corresponding axes and the movie titles shown when the pointer is hovered over the dots.

- The **Horizontal Bar Graph** addresses the second visualization tool. It efficiently displays the voting distribution across various movie genres with a simple and intuitive layout. (genres points towards y-axis and vote percentages towards x-axis). Users can easily determine which movie has received the most votes within each genre owing to the usage of several tints of the same color for each title within a genre and the inclusion of a legend. On the other hand, the merging of movie titles inside the bars, however, makes it challenging to compare the amount of votes for several films belonging to the same genre, which is one of this tool's biggest flaws.

- Closing with the final visualization, the **Packed-Bubble Plot**, Tableau's power here rely's in its potential to provide an interesting and dynamic depiction of data points. The option to sort the data depending on length is very helpful for analysing the data in further depth. Employing various hues and opacities to distinguish amongst data sets helps to make the chart better understandable. The plot's nested circles make it possible to clearly segregate movies with different run spans. Detailing about the UI characteristic, this plot would be especially suitable for those who are interested in examining the spectrum of movie run times and finding trends or patterns in the data. But few drawbacks in this interface design includes more versatility in the size and location of the bubbles as some users could find this to be perplexing , as well as making it simpler to recognise the movie titles connected to each data point.