VISUALIZATION OF IMDB MOVIES DATA SET

BDA-504 FINAL

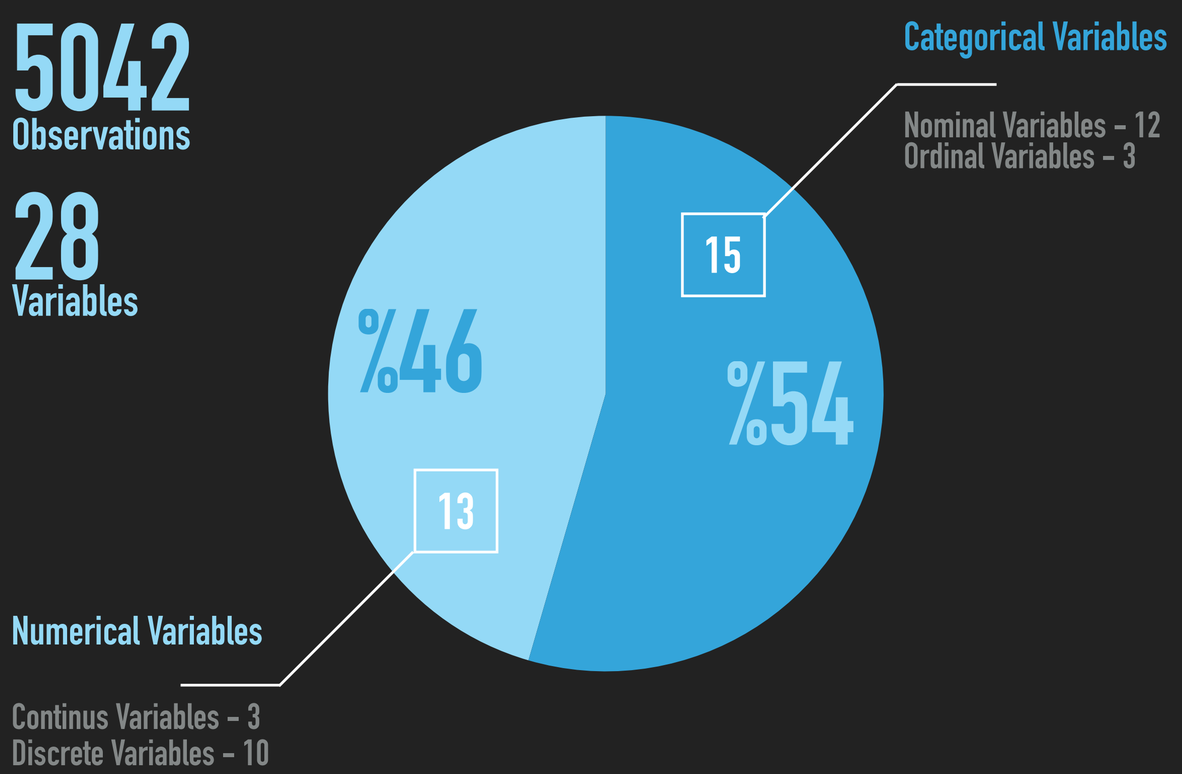
BY CEM KILIÇLI - 3.4.2017

VARIABLES – CATEGORICAL & NUMERICAL

- In IMDB Movies data set there are 5042 observations. The observations does not have all the listed variables filled. Some data about the observations is either missing or not available.
- In data set there are 28 variables in total. In addition to these variables that are available I split up one specific variable “Genre” into 8 different categorical variables.
- Each newly created categorical variable is formed by the availability of a specific genre for the defined title. (Ex. Avengers - Action | Adventure | Sci-Fi)
- Created genres are used as a hierarchical field. For the simplicity and making the graph more understandable first of 8 categories (genre) in the created hierarchy used in the project. Also an option to dig deeper is available.

5042
Observations

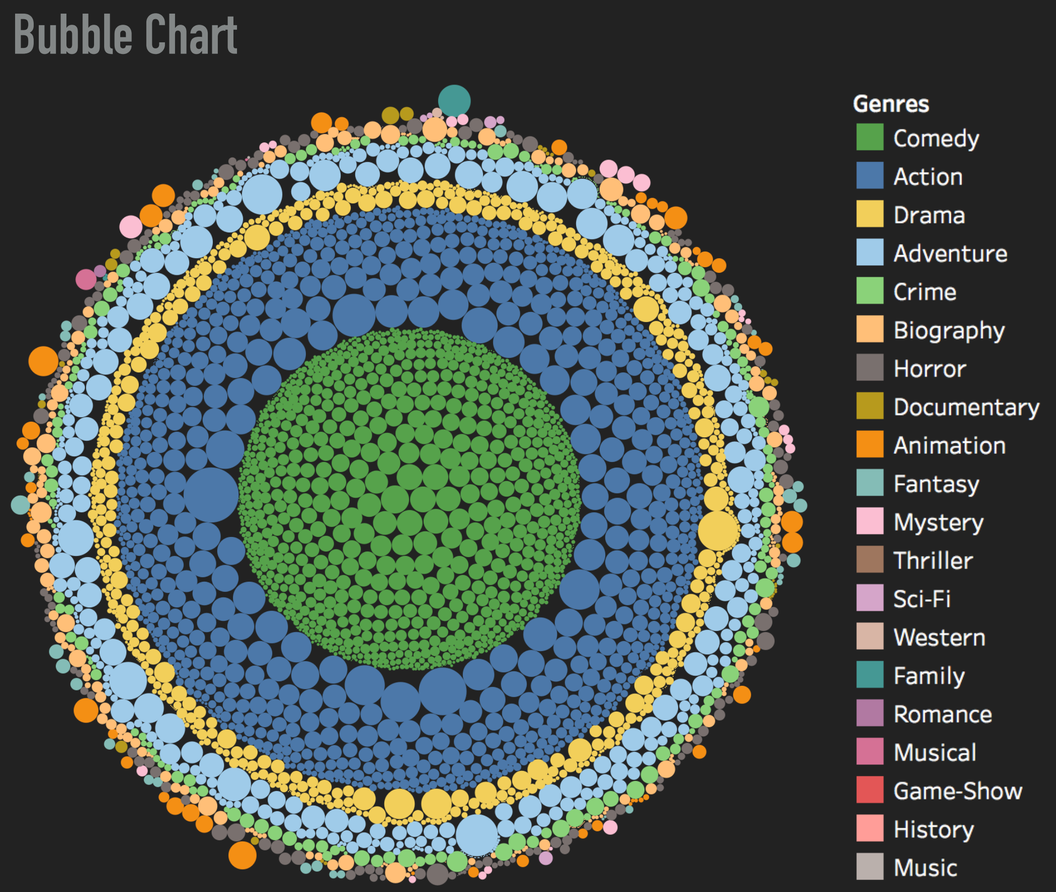
28
Variables



The data set can be accessible by; https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset

GENRES & GROSS

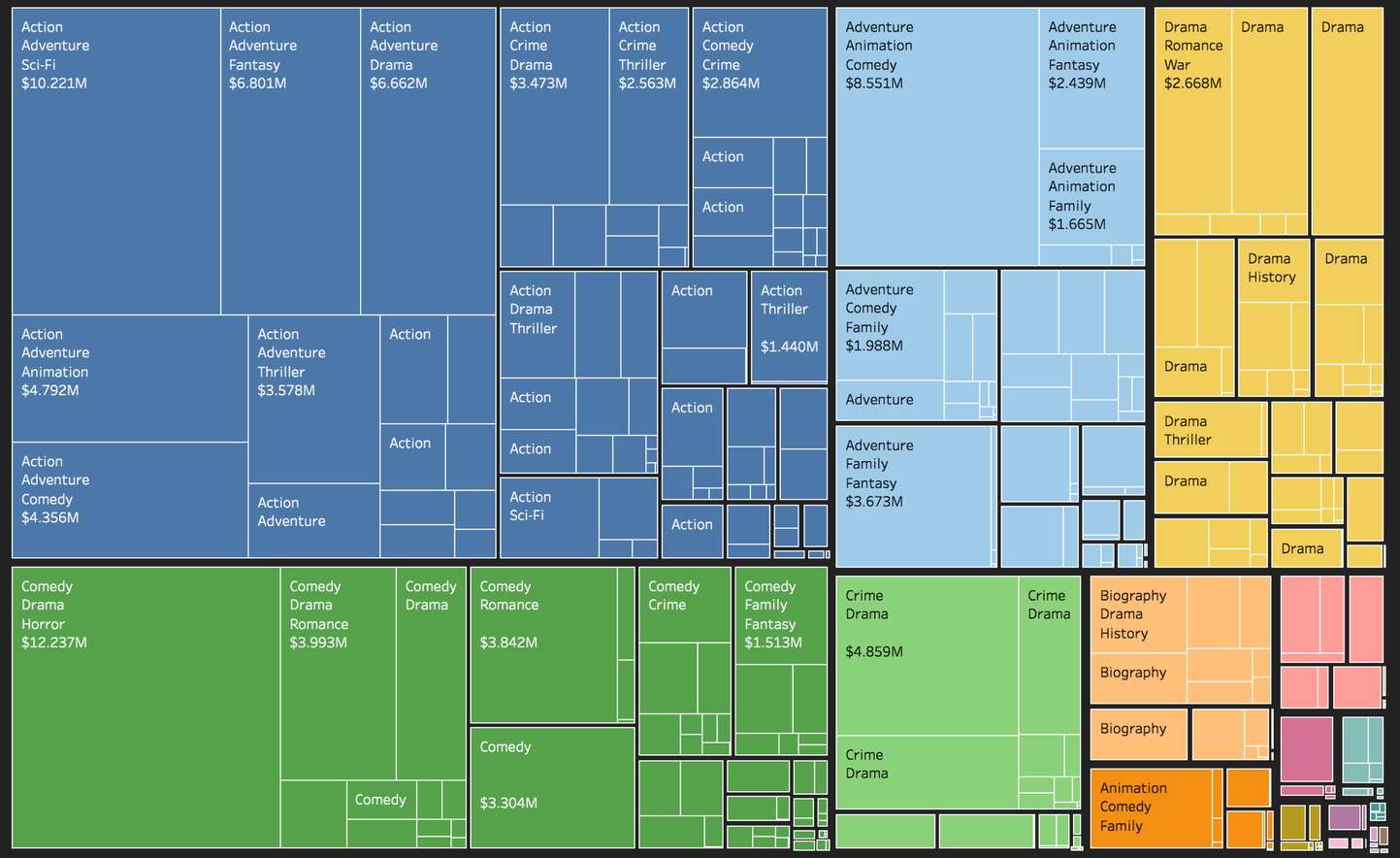
- In this graph I use “Gross” as a indicate of scale. Bubble sizes create a comparison between movies “Gross”. Scale of area is used to compare the difference between “Gross” numbers.
- Colouring done due to “Genres” so that the same coloured bubbles are in the same “Genre” . Genres also used as attribute to display the selected genre in the tooltip.
- This data covers all the movies that are represented in the data set.
- I choose to use bubble graph to represent this data because; it is a multi-variable graph. These charts typically used to compare and show the relationship between labels/ categories.
- This chart can be filtered to create a more readable version on an interaction basis. This is the only avoidance that can be used in bubble graph since the data size is large.
- This feature scan be used as a filter on the interactive dashboard that is created.



BUDGET BY GENRES

- Tree-maps charts are one of the best way to visualise hierarchical structure. Each genre is assigned to a rectangle area with their sub-genre rectangles nested inside of it.
- The size of the rectangles represent the “Budget” spent to film the movie. Scale of area is used to compare the difference between “Budgets”.
- As a quantity “budget”, assigned to genre categories, its area size is displayed in proportion to that quantity and to the other quantities within the same parent category in a part-to-whole relationship.
- Since the amount of data and the genre considered this graph is selected for; a more compact and space-efficient option for displaying hierarchies, that gives a quick overview of the structure.
- This feature scan be used as a filter on the interactive dashboard that is created.

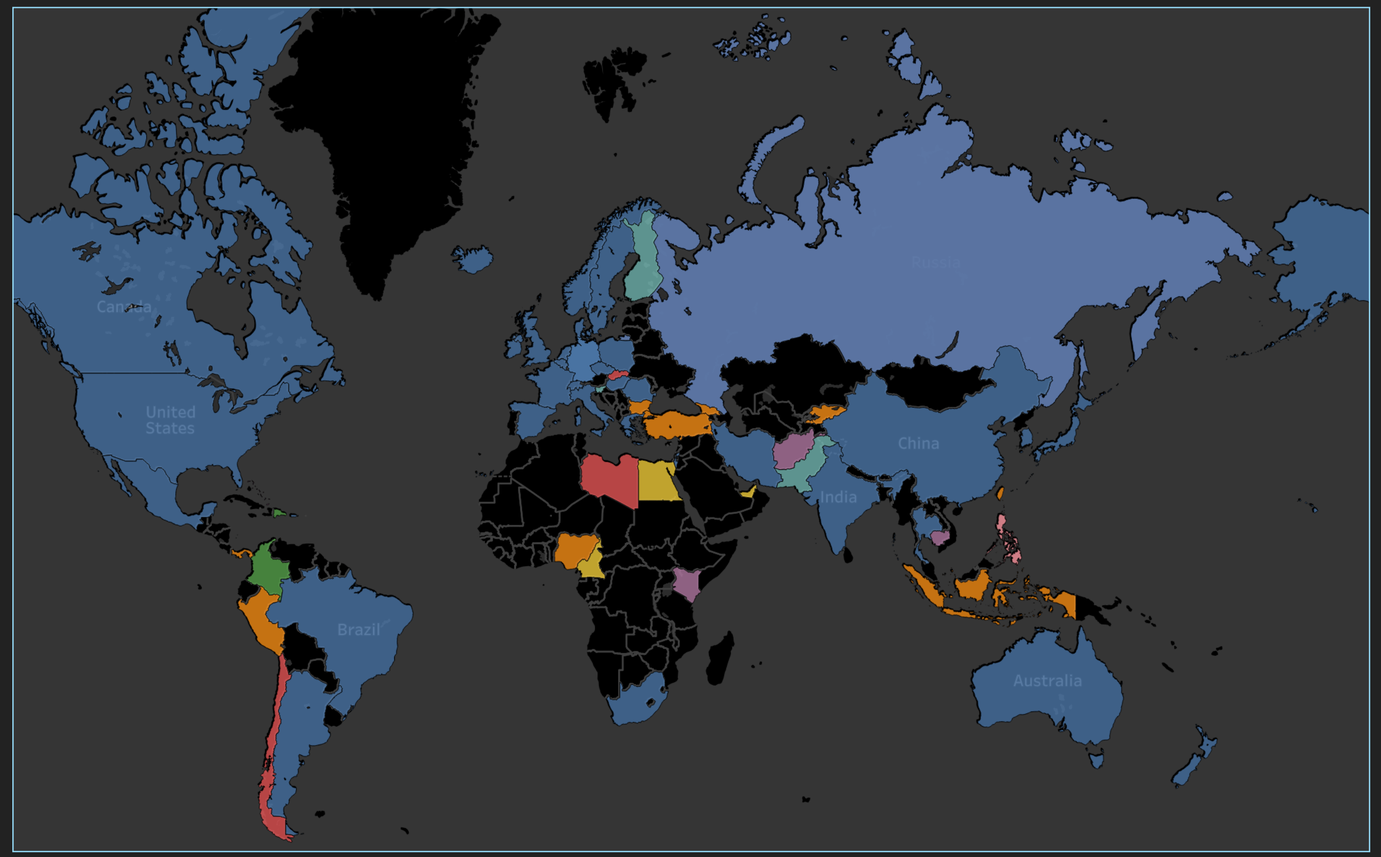
Tree Map Chart



MOVIES BY COUNTRY

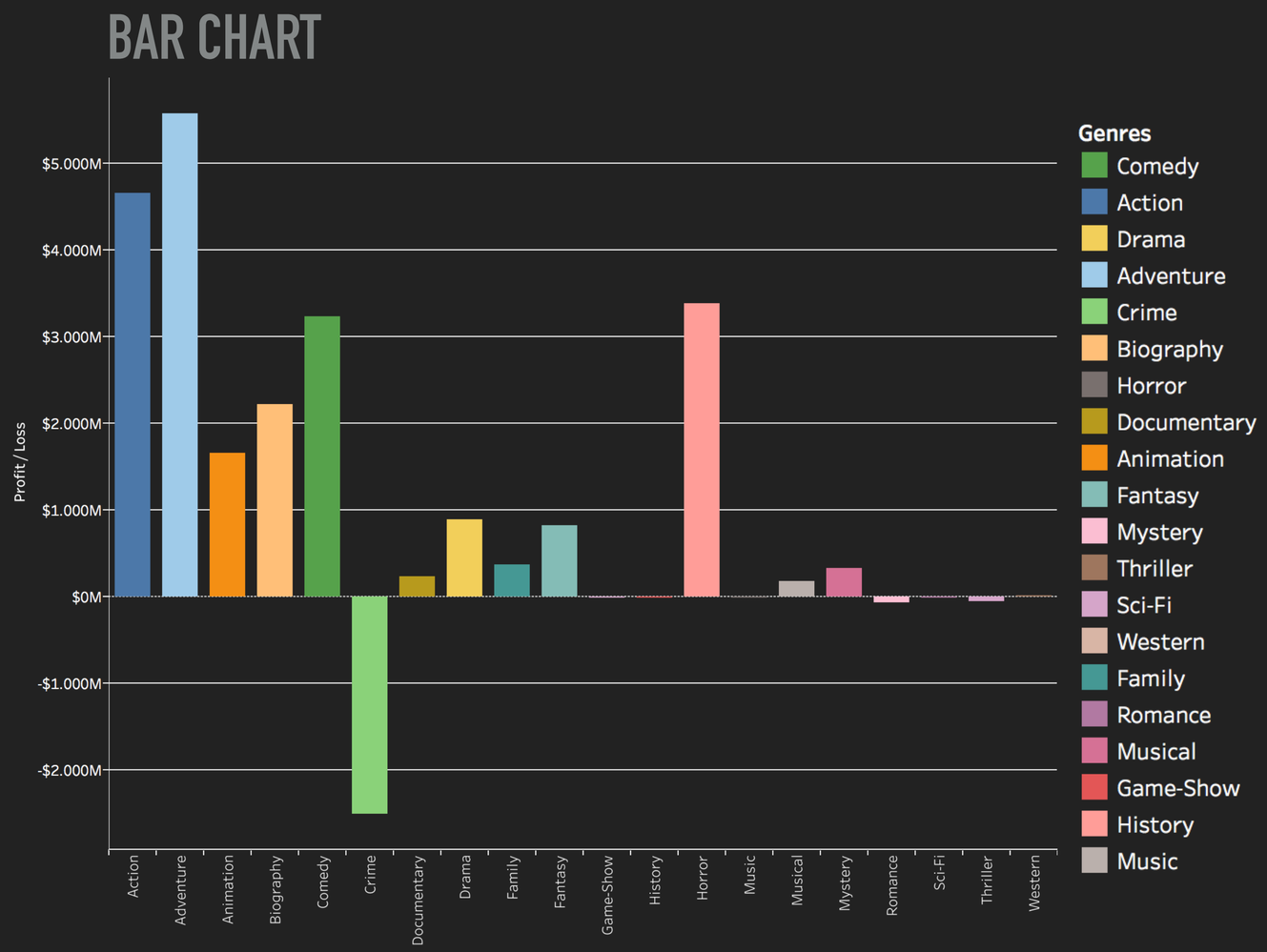
- Since I have countries listed as one of the categorical variable it makes sense, to use a map representation of the data meaningful.
- Using the longitude, latitude conversion feature of software, I distribute the movie origins to map and visualise it.
- Genres used in this graph as attribute that is colorised. Due to the restrictions in software we only can assign “Genres” via color which has the most observations.
- This feature scan be used as a filter on the interactive dashboard that is created.

World Map Chart



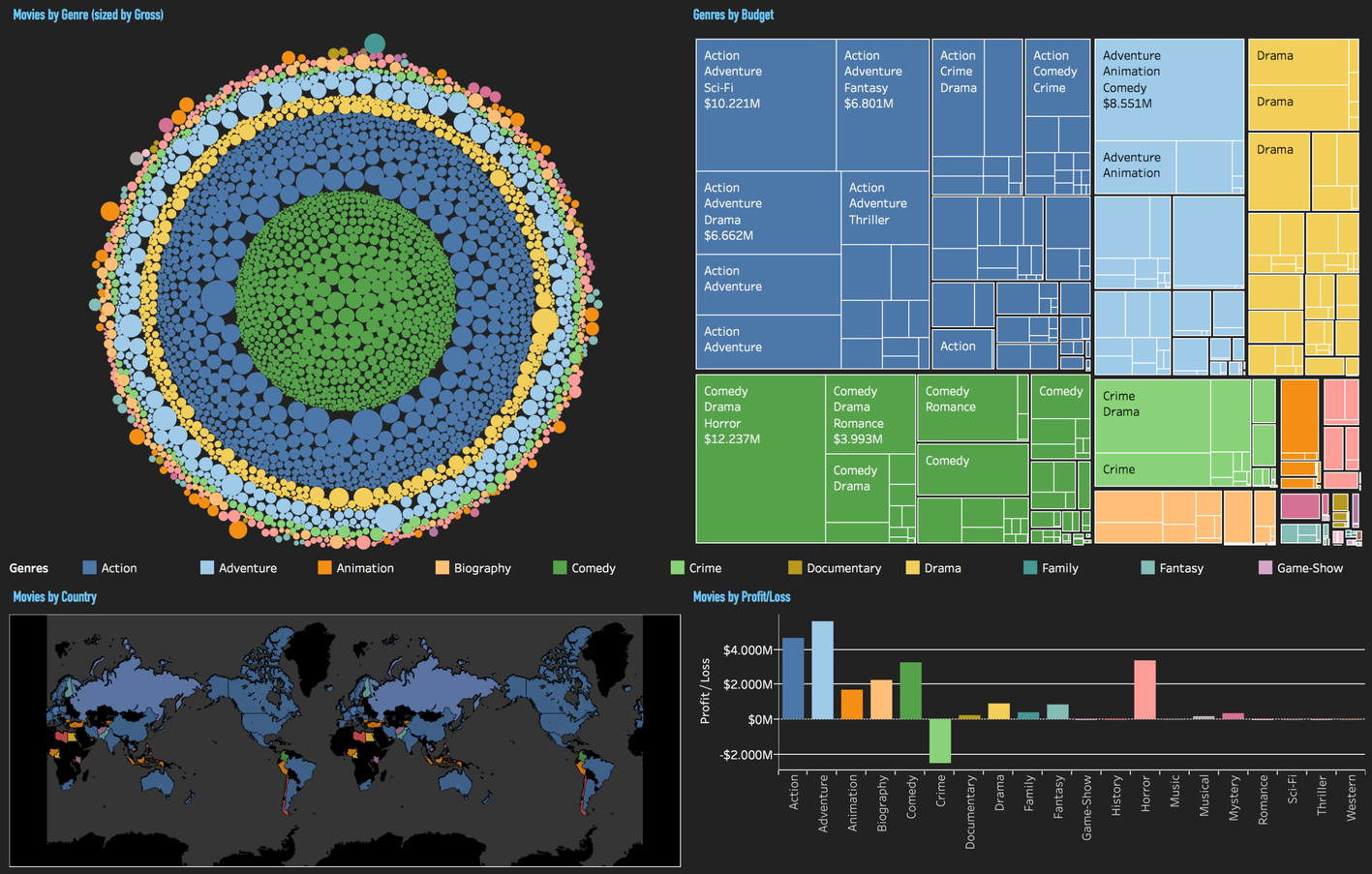
MOVIES BY TITLE YEARS

- To display the profit/loss I have choose Bar chart. Since this variable is a discrete variable, this one of the best way to comparing numerical data across categories (Genre).
- Bar Chart's discrete data is numerical data and therefore answers the question of "how much?" In each category.
- Since the use of labelling on bar chart can be problematic. The label representation is avoided.
- This feature scan be used as a filter on the interactive dashboard that is created.



INTERACTIVE DASHBOARD

- Bubble chart can be used to filter and dig in to data by “Movie Title” basis. If a specific movie is selected the related data based on selected movie will be filtered and displaced across the whole dashboard.
- Tree-Map chart can be used to filter out certain genre for the rest of the charts. If a specific genre is selected the related data based on selected genre will be filtered and displaced across the whole dashboard.
- Map chart can be used to filter the data in a bases of countries. If a specific country is selected the related data based on selected country will be filtered and displaced across the whole dashboard.
- Bar chart can be used to filter the data in a bases of genres. The genre that is selected, will filter out the data related with top level hierarchy in genres and displaced across the whole dashboard.



Interactive Dashboard link:
https://public.tableau.com/views/IMDBMOVIESVIZUALIZATION/Dashboard3?:embed=y&:display_count=yes

CONCULUTION

- Biggest budget is spend on Action movies. Second biggest budget is spend to Comedy movies. This may be caused due to missing points within the data.
- Biggest profit is also gained by Adventure movies. Second one is the Action movies. The weird part of the analysis is that Crime movies come out with huge losses. Which might be the cause of lack of data availability.
- Also the movie count per country is very limited. Most movies come out from USA.
- This data can be used only for studies that the resulting analysis is does not affect in any scientific research.

LESSONS LEARNED

- There are limits of what out of the package software can do. Considering the software I use it lacks of many of the chart types available in data visualisation.
- It is hard to assign distinct colours if you have too many categories under one categorical variable.
- Most of the times visualising big data is tricky and the software have limits in the max number of inputs that it can get for a specific field. In example you can not assign more than 5000 rows in the software that I use.
- Creating a dashboard to display data from various angles need time and expertise on the topic. It might be tricky to spot out the important information.

THANK YOU

VISUALIZATION OF IMDB MOVIES DATA SET

BY CEM KILIÇLI