KOLEJ PROFESIONAL MARA BERANANG

DIPLOMA IN COMPUTER SCEINCE

COURSE NAME	:	DATA ANALYSIS & VISUALIZATION
COURSE CODE	:	CSC 2823
SESSION	:	3 2024/2025
TYPE OF ASSESSMENT	:	FINAL PROJECT
DURATION	:	03/02/2025 – 21/02/2025

CLO 3 Prepare data visualization for effective presentation using computer software tool.

INSTRUCTION TO CANDIDATES:

1. Students are required to answer ALL questions.
2. Students need to submit the assignment report in hardcopy.

Personal Details	
Name	
I/D Number	
Class	☐ DCS 5A ☐ DCS 5B ☐ DCS 5C ☐ DCS 5D
Lecturer	☐ **DR.ZALINA AYOB** ☐ **MAWARWIDURI BT AB HALIK**

Task	**Marks**
1	
2	
3	
4	
5	
Total	**/ 50**

PROJECT SCENARIO

Assume you are a data analyst working for a company, government agency, or nonprofit organization. Your team has been tasked with providing actionable insights using data to address specific challenges or make better decisions. An example of dataset relevant to a specific domain as shown in Figure 1 below. You are required to identify and analyze the selected dataset, create visualizations, and present your findings clearly, informative, and useful for stakeholders.

Domain	Example of dataset
Healthcare	· Patient health records (age, gender, diagnosis, treatments, recovery rates) · Disease outbreaks (geographical spread, case counts, recovery rates) · Hospital performance (admissions, bed availability, patient satisfaction)
Retail & E-commerce	· Product sales (date, category, region, revenue, units sold) · Customer behaviours (age group, purchase frequency, cart abandonment rate) · Inventory data (stock levels, restocking frequency).
Education	· Student performance (grades, attendance, participation) · Enrolments trends (age, gender, region, program type) · Online learning metrics (time spent on platform, quiz scores, course completion).
Transportation	· Public transport usage (daily ridership, routes, timings) · Traffic data (vehicle count, congestion levels, accident locations) · Flight or train delays (departure time, delay length, reason for delay).
Sports	· Player statistics (points, assists, performance over seasons) · Match results (team, score, location, audience size) · Fan engagement (ticket sales, merchandise, social media activity).
Environmental science	· Climate data (temperature, precipitation, CO2 levels, time) · Air quality (pollutants, AQI, region, time) · Wildlife populations (species, count, region, time).
Energy and utilities	· Energy consumption (time, location, source type, usage) · Renewable energy output (solar, wind, hydro generation by region) · Utility service complaints (frequency, type, resolution time).
Public safety and law enforcement	· Crime statistics (type, location, time, resolution status). · Emergency response times (type of emergency, location). · Traffic violations (type, location, time of day).
Gaming industry	· Player activity (time spent, levels completed, purchases). · Game performance metrics (bugs, crashes, frame rate). · Monetization (in-app purchases, subscriptions, ad revenue).
Media and entertainment	· Streaming platform data (viewership, time spent, genres watched). · Box office sales (ticket sales, revenue, regions). · Social media activity (hashtags, trends, engagement).

Urban Planning and smart city	▪ City infrastructure usage (traffic, utilities, public transport). ▪ Population density (by region, time of day). ▪ Pollution levels (air, water, noise by region).

Figure 1

The following steps outline the processes and the tasks that you need to complete for the project:

Task 1: Identification of dataset and analysis background

- Identify domain/key areas based on Figure 1. You may use suggested dataset or other relevant examples.

Task 2: Gather Relevant Data

- Gather data to support the generation of effective visualization.

Task 3: Clean and Preprocess the Data

- Raw data is cleaned and pre-processed to ensure it can be used effectively.

Task 4: Analyze the data to extract meaningful insights and trends.

- Analyze the data to extract meaningful insights and trends using:
  - ☐ Univariate Data Analysis
  - ☐ Bivariate Data analysis

Task 5: Create visually appealing and informative data visualizations to convey the identified insights

- Utilize charts, graphs, maps, and other visualization techniques to make the information easily understandable. The visualization should use different types of variables and suitable presentation approaches.

Assessment Rubrics:

No	Attribute	Task	(1 mark)	(2 Marks)	(3 Marks)	(4 Marks)	Marks Weighted	Marks Obtained
1	Gather	**Identification of dataset and analysis background**						
		Identify domain area and the analysis background	Able to provide at least **ONE (1)** of the following: ☐ Provide objective of the analysis correctly ☐ Able to identify target audience appropriately	Able to provide **All** of the following: ☐ Provide objectives of the analysis correctly. ☐ Able to identify target audience appropriately	Able to provide **All** of the following: ☐ Precisely explain the objective of the analysis correctly. ☐ Able to identify target audiences appropriately	Able to provide **All** of the following: ☐ Precisely explain the objective of the analysis correctly. ☐ establish the scope of the analysis, outlining what will be included and excluded ☐ Able to identify target audience appropriately and explain the benefit gained from the analysis	1	
2		**Gather Relevant Data**						
		Gather data to support the generation of effective visualization.	Able to provide **ALL** the following: ☐ Able to justify appropriately the reason for your data selection. ☐ Able to select suitable dataset to fulfill the objective(s)	Able to provide **ALL** the following: ☐ Able to justify appropriately the reason for your data selection. ☐ Able to select suitable dataset to fulfill the objective(s)			1	

			☐ Provide evidence for the dataset from various sources (website/pre-installed RStudio etc)	☐ Provide evidence for the dataset from various sources (website/pre-installed RStudio etc) ☐ Able to specify suitable variable(s) from the dataset for the analysis				
3		**Clean and Preprocess the Data**						
		Clean and Preprocess the Data	☐ Provide a screen shot of with insufficient steps to treat any missing values or quantities of zero in the dataset	☐ Provide a screen shot of the correct steps to treat any missing values or quantities of zero in the dataset	☐ Provide a screen shot of the correct steps to treat any missing values or quantities of zero in the dataset. ☐ Provide a minimal explanation of the steps.	☐ Provide a shot of the correct steps to treat any missing values or quantities of zero in the dataset. ☐ Provide a precise and clear explanation of each step in the screen shots	1	
4	Reproduce and Process Information	**Analyze the data to extract meaningful insights and trends.**						
		Univariate Data analysis	☐ Able to provide **one** category of univariate analysis ☐ Able to display one visual on category with a unsuitable selection of a plot or graph for any category	☐ Able to provide **one** category of univariate analysis ○ Able to display one visual on category with a suitable selection of a plot or graph for any category	○ Able to provide **more than one** category of univariate analysis ○ Able to display visual on categories with a suitable selection of a plot or graph for each category	○ Able to provide **more than one** category of univariate analysis ○ Able to display visual on categories with a suitable selection of a plot or graph for each category ○ Provide a clear and concise	2	

						finding explanation based on the analysis conducted for each category.		
		Bivariate Data Analysis	○ Able to provide **one** category of Bivariate analysis ○ Able to display one visual on category with a unsuitable selection of a plot or graph for any category	○ Able to provide **one** category of Bivariate analysis ○ Able to display one visual on category with a suitable selection of a plot or graph for any category	○ Able to provide **more than one** category of Bivariate analysis ○ Able to display visual on categories with a suitable selection of a plot or graph for each category	○ Able to provide **more than one** category of univariate analysis ○ Able to display visual on categories with a suitable selection of a plot or graph for each category ○ Provide a clear and concise finding explanation based on the analysis conducted for each category	2	
5		**Create visually appealing and informative data visualizations to convey the identified insights**						
		Use suitable chart /graph type	☐ Provide suitable types of presentations for the targeted audience. ☐ Use the ggplot2 library to create the graphic. ☐ The created visualisation is simple and only caters to **ONE (1)** of	☐ Provide suitable types of presentations for the targeted audience. ☐ Use the ggplot2 library to create the graphic. ☐ The created visualisation is simple and only caters to	○ Provide suitable types of presentations for the targeted audience. ○ Use the ggplot2 library to create the graphic. ○ The created visualisation is simple and only caters to **THREE (3)** of the following: ○ Colors and contrast	○ Provide suitable types of presentations for the targeted audience. ○ Use the ggplot2 library to create the graphic. ○ The created visualisation is simple and only cater to **ALL** of the following: ○ Colors and contrast	2	

	Convey		the following: ○ Colors and contrast ○ size ○ theme ○ scale	**TWO (2)** of the following: ○ Colors and contrast ○ size ○ theme ○ scale	○ size ○ theme ○ scale	○ size ○ theme ○ scale		
		Customize graph in creative ways	☐ The visualisation produced covers only **ONE(1)** of the aspects of below: ○ Good quality with clear and accurate ○ Have aesthetic (colors,label & formatting) appeal with visually pleasing ☐ easy for the audience to understand ☐ relevance and effective to communicate the intended message	☐ The visualisation produced covers only **TWO(1)** of the aspects below: ○ Good quality with clear and accurate ○ Have aesthetic (colors,label & formatting) appeal with visually pleasing ○ easy for the audience to understand ○ relevance and effective to communicate the intended message	☐ The visualisation produced covers only **THREE (3)** of the aspects of below: ○ Good quality with clear and accurate ○ Have aesthetic (colors, label & formatting) appeal with visually pleasing ○ easy for the audience to understand ○ relevance and effective to communicate the intended message	☐ The visualisation produced covers **ALL** of the aspects of below: ○ Good quality with clear and accurate ○ Have aesthetic (colors, label & formatting) appeal with visually pleasing ○ easy for the audience ○ relevance and effective to communicate the intended message	2	

		Summarize the findings	☐ Able to highlight some of the important findings/result. ☐ No interpretation of the result provided.	☐ Able to highlight some of the important findings. ☐ Poor interpretation of results provided.	☐ Able to highlight all the important findings. ☐ The interpretation of results is briefly explained but inappropriate, based on the analysis of the data.	☐ Able to highlight most of the important findings. ☐ The interpretation of results is briefly explained and appropriate, based on the analysis of the data.	2	
Total Marks Earned								/50
Total Percentage (40%)								

Table of Contents

1.0 Introduction

In today's digital era, movies remain one of the most popular forms of entertainment across all generations. The film industry has evolved with changing audience preferences, shifting from traditional cinemas to streaming platforms. By analyzing a dataset from Kaggle.com, I explored key aspects of the movie industry, such as average ratings, popular genres over time, and the financial impact of production budgets on revenue. This analysis helps determine trends in audience preferences, the success factors behind high-grossing films, and whether investing in movies is a profitable venture

2.0 Identification of dataset

a. Objective

The objective of this analysis is to extract meaningful insights from a dataset containing movie information from 1930 to 2016. This analysis explores the average movie ratings over time and identifies the most popular genres during different periods. Additionally, it examines the revenue generated based on genre and investigates whether a high budget is necessary to achieve high revenue.

Scope

Included (state all variables that use to achieve objective)

Genre - Category of the film

Budget - Money spent on making the movie and its publicity

Box Office(revenue) - Money from ticket sales

IMDb score - Score out of 10 that is calculated from the votes of registered IMDb users on the movie

Excluded

Movie - Name of the film

Director - Film director that controls the making of the movie and supervises the actors and technical crew

Running time - The length of time in minutes of the movie

Actor 1, Actor 2 and Actor 3 - Three different actors that participate in the movie

Actors Box Office % - Percentage that reflects how many times the actors managed to at least double the budget in their other films movies

Director Box Office % - Percentage that reflects how many times the director managed to at least double the budget in their other films movies
Oscars and Golden Globes nominations - Amount of nominations that the movie had in the Oscars and the Golden Globes
Oscars and Golden Globes awards - Amount of awards that the movie had in the Oscars and the Golden Globes
Release year - Year when the movie was first released

b. Target audience and benefit of the analysis

- Movie productions and producers: better decisions making leads to avoid financial loss and what drives the movie to success.
- Investor: to get high return on investment. By understanding the revenue and budget pattern this can lead to better decision making.
- Movie critics and influences: play a crucial role to shape public by analyzing the movie, is it worth the hype like overrated or underrated or not?

3.0 Dataset

a. Justification using dataset

The reason this dataset is chosen is because it meets the requirements to do univariate and bivariate analysis, which includes 2 types of variables which is categorical and numerical data types. Another thing is, the datasets provide budget and box office(revenue) perfect to analyze movies profitability. Also, it provides IMDb scores for the public to criticize the movie. Lasty, genre to know what movie production should focus on.

b. Evidence of the data

https://www.kaggle.com/datasets/delfinaoliva/movies

Appendix1

c. Suitable dataset used (from included)

Genre - Category of the film
Budget - Money spent on making the movie and its publicity

Box Office(revenue) - Money from ticket sales

IMDb score - Score out of 10 that is calculated from the votes of registered IMDb users on the movie

4.0 Data cleaning

```
> df_status(data)
                              variable q_zeros p_zeros q_na p_na q_inf p_inf      type unique
1                                Movie       0    0.00    0 0.00     0     0 character   3907
2                             Director       0    0.00    0 0.00     0     0 character   1760
3                         Running.time       0    0.00    0 0.00     0     0   integer    155
4                              Actor.1       0    0.00    0 0.00     0     0 character   1591
5                              Actor.2       0    0.00    0 0.00     0     0 character   2367
6                              Actor.3       0    0.00    0 0.00     0     0 character   2782
7                                Genre       0    0.00    0 0.00     0     0 character     14
8                               Budget       0    0.00    0 0.00     0     0   integer    374
9                           Box.Office       0    0.00    0 0.00     0     0   numeric    994
10                  Actors.Box.Office..     321    8.08    0 0.00     0     0   numeric    234
11                Director.Box.Office..     854   21.49    0 0.00     0     0   numeric     39
12                             Earnings      68    1.71    0 0.00     0     0   numeric   1244
13 Oscar.and.Golden.Globes.nominations    2839   71.44    0 0.00     0     0   integer     22
14      Oscar.and.Golden.Globes.awards    3504   88.17    3 0.08     0     0   integer     14
15                        Release.year       0    0.00    0 0.00     0     0   integer     86
16                          IMDb.score       0    0.00    0 0.00     0     0   numeric     77
```

The image above shows that this data set has two variables that have percentage zero more than 60 percent. Variable 60 percent can lead to bias decisions.

```
12  # 2. Data cleaning remove percentage zero value
13  my_data = df_status(data)
14  arrange(my_data, -p_zeros) %>% select(variable, q_zeros,p_zeros)
15  vars_to_remove <- filter(my_data, p_zeros > 60) %>% .$variable
16  vars_to_remove
17  movies = select(data, -one_of(vars_to_remove))
18  df_status(movies)
```

- In line 13, assign a new variable to be use in line 14
- In line 14, arrange data in descending order by percentage zero. The output should be displayed like the image below:

```
> arrange(my_data, -p_zeros) %>% select(variable, q_zeros,p_zeros)
                            variable q_zeros p_zeros
1       Oscar.and.Golden.Globes.awards    3504   88.17
2  Oscar.and.Golden.Globes.nominations    2839   71.44
3                 Director.Box.Office..     854   21.49
4                   Actors.Box.Office..     321    8.08
5                             Earnings      68    1.71
6                                Movie       0    0.00
7                             Director       0    0.00
8                         Running.time       0    0.00
9                              Actor.1       0    0.00
10                             Actor.2       0    0.00
11                             Actor.3       0    0.00
12                               Genre       0    0.00
13                              Budget       0    0.00
14                          Box.Office       0    0.00
15                        Release.year       0    0.00
16                          IMDb.score       0    0.00
```

- In line 15, is the code to remove variables that have percentage zero more than 60%
- Line 16 is to execute the code in line 15. The output should be displayed like the image below

```
> vars_to_remove <- filter(my_data, p_zeros > 60) %>% .$variable
> vars_to_remove
[1] "Oscar.and.Golden.Globes.nominations" "Oscar.and.Golden.Globes.awards"
```

- In line 17, assign new variable to keeping all columns except the ones present in “vars_to_remove”.
- Line 18 is to see variable. The output should be displayed like the image below:

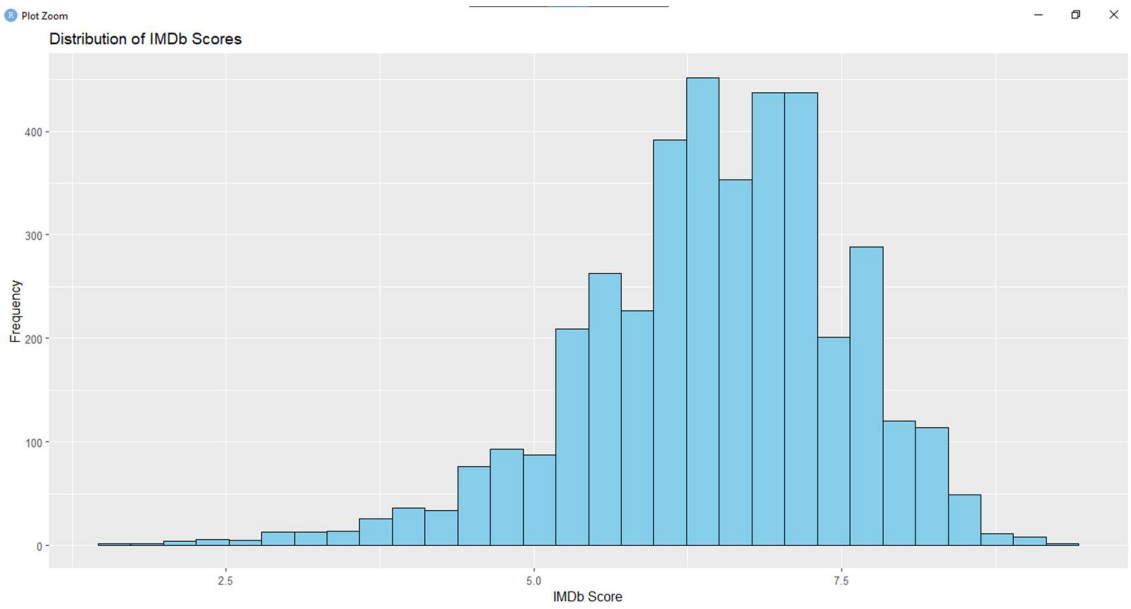
```
> movies = select(data, -one_of(vars_to_remove))
> df_status(movies)
                variable q_zeros p_zeros q_na p_na q_inf p_inf      type unique
1                  Movie       0    0.00    0    0     0     0 character   3907
2               Director       0    0.00    0    0     0     0 character   1760
3           Running.time       0    0.00    0    0     0     0   integer    155
4                Actor.1       0    0.00    0    0     0     0 character   1591
5                Actor.2       0    0.00    0    0     0     0 character   2367
6                Actor.3       0    0.00    0    0     0     0 character   2782
7                  Genre       0    0.00    0    0     0     0 character     14
8                 Budget       0    0.00    0    0     0     0   integer    374
9             Box.Office       0    0.00    0    0     0     0   numeric    994
10   Actors.Box.Office..     321    8.08    0    0     0     0   numeric    234
11 Director.Box.Office..     854   21.49    0    0     0     0   numeric     39
12             Earnings      68    1.71    0    0     0     0   numeric   1244
13          Release.year       0    0.00    0    0     0     0   integer     86
14            IMDb.score       0    0.00    0    0     0     0   numeric     77
```

5.0 EDA Data analysis

Univariate analysis

Numerical analysis: average IMDb score

```
25  ggplot(movies, aes(x = IMDb.score)) +
26    geom_histogram(fill = "skyblue", color = "black") +
27    labs(title="Distribution of IMDb Scores",x="IMDb Score",y="Frequency" )
28
29  summary(movies)
```



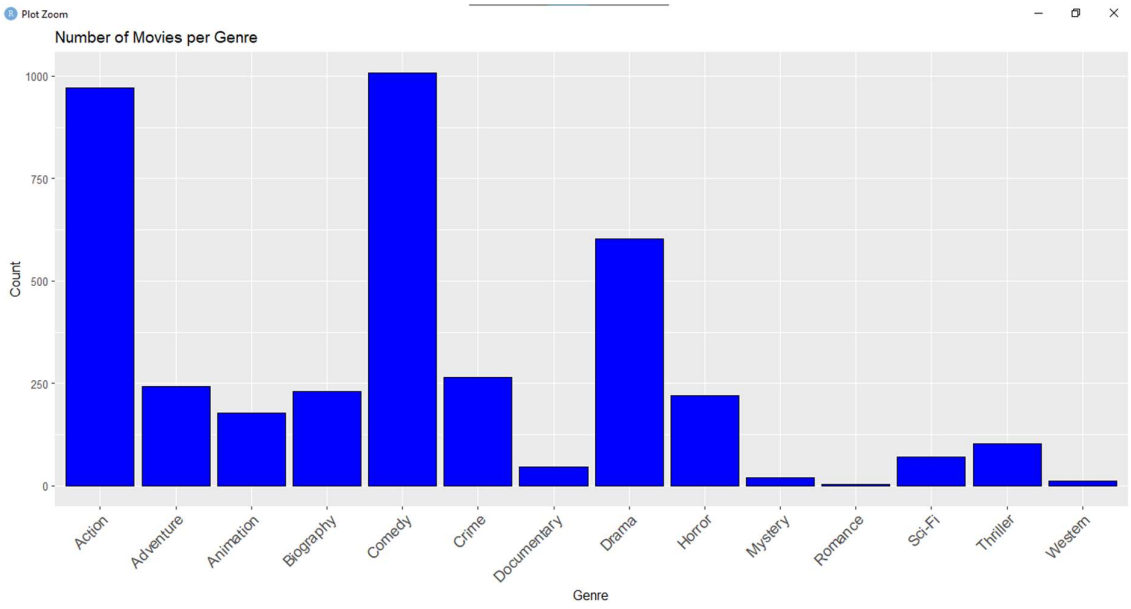
by using summary () function:

- The lowest rating is 1.6/10
- Highest rating is 9.3/10
- And the average rating is 6.48/10
- In conclusion, most of the public audience give moderate rating more than high and low.

```
   IMDb.score
Min.   :1.600
1st Qu.:5.900
Median :6.600
Mean   :6.468
3rd Qu.:7.200
Max.   :9.300
```

Categorical analysis: Number of movies by genre

```
31  # categorical analysis: Number of movies by genre
32  ggplot(movies, aes(x = Genre)) +
33    geom_bar(fill = "blue", color = "black") +
34    theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 12)) +
35    labs(title = "Number of Movies per Genre",x="Genre",y="Count")
```



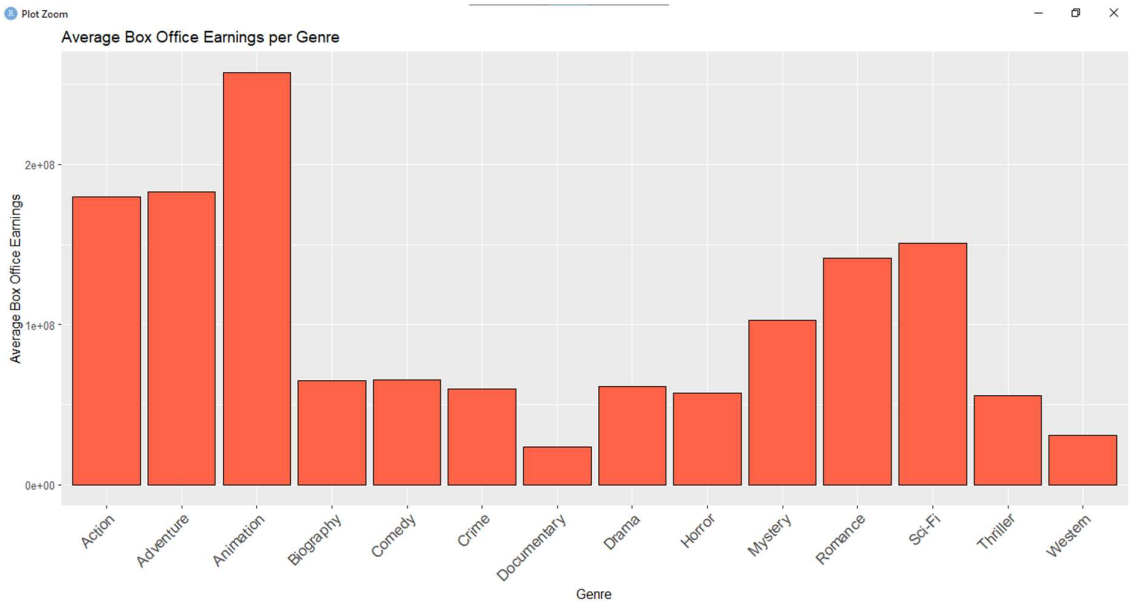
Based on the bar image:

- Comedy genre is the most produced movies
- The second goes to action
- The third is Drama
- And the least produce goes to romance
- In conclusion, movies production tend to produce more comedy and action genre.

Bivariate analysis

categorical analysis vs numerical analysis: average box office earnings per genre

```
38  # categorical vs numerical: average box offfice per genre
39  movies %>%
40    group_by(Genre) %>%
41    summarise(Average_Box_Office = mean(Box.Office)) %>%
42    ggplot(aes(x = Genre, y = Average_Box_Office)) +
43    geom_bar(stat = "identity", fill = "tomato", color = "black") +
44    theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 12)) +
45    labs(title="Average Box Office Earnings per Genre", x="Genre",
46         y="Average Box Office Earnings")
```



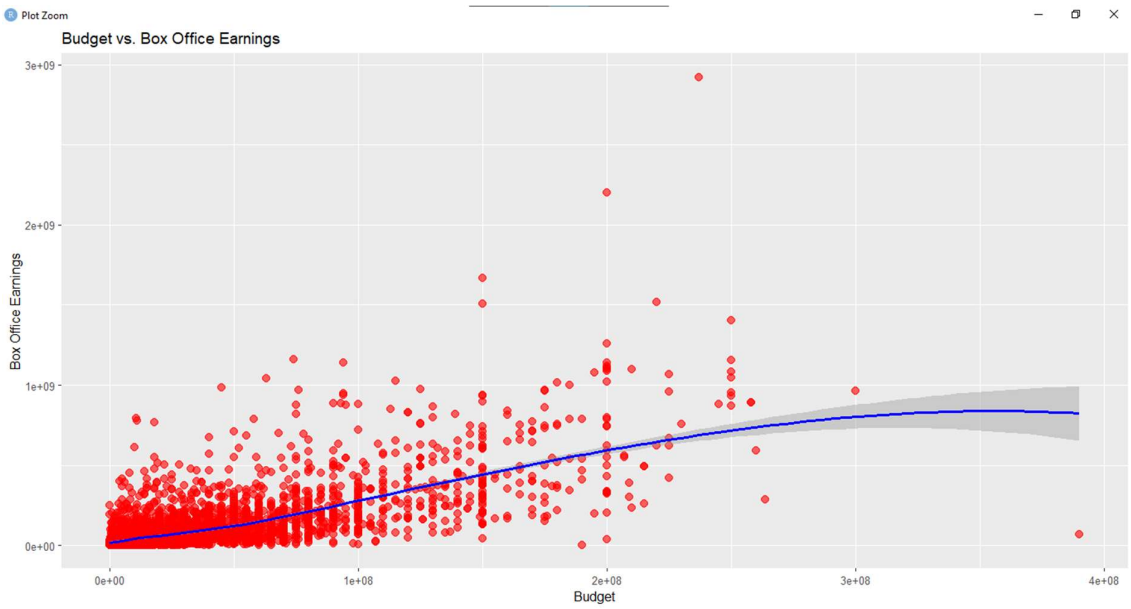
Based on the above image:

- Animation performs better than any other genre
- Meanwhile average box office earnings go to adventure and action
- Least, box office earnings(revenue) is documentary
- From summary() function:

```
   Box.Office
Min.   :5.000e+04
1st Qu.:1.200e+07
Median :4.300e+07
Mean   :1.087e+08
3rd Qu.:1.250e+08
Max.   :2.923e+09
```

numerical analysis vs numerical analysis

```r
# numerical vs numerical:
## Numerical vs. Numerical: Relationship between Budget and Box Office Earnings
ggplot(movies, aes(x = Budget, y = Box.Office)) +
  geom_point(color = "red", alpha = 0.6, size = 3) +
  geom_smooth(method = "loess", color = "blue",size = 1.2) +
  ggtitle("Budget vs. Box Office Earnings") +
  xlab("Budget") +
  ylab("Box Office Earnings")
```



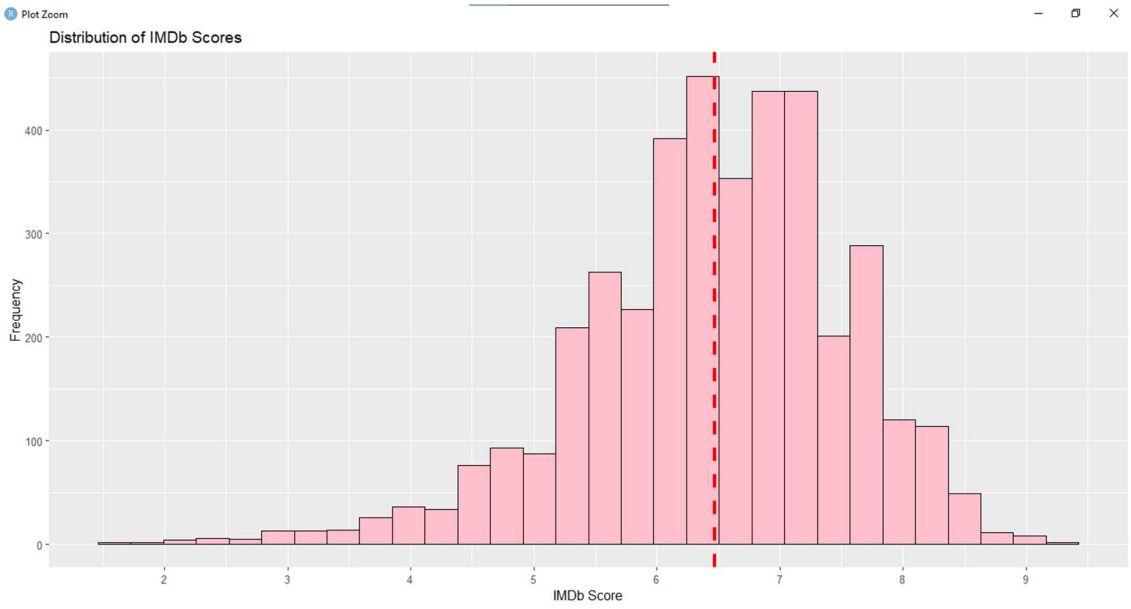
Based on the image above:

- The scatter plot shows a positive correlation, higher budget led to higher box office earnings(revenue)
- But it is not guaranteed that higher budgets lead to higher box office earnings(revenue) because some movies with higher budgets perform poorly
- From summary function ()

```
     Budget          Box.Office
Min.   :     1100  Min.   :5.000e+04
1st Qu.:  9000000  1st Qu.:1.200e+07
Median : 22000000  Median :4.300e+07
Mean   : 36906392  Mean   :1.087e+08
3rd Qu.: 50000000  3rd Qu.:1.250e+08
Max.   :390000000  Max.   :2.923e+09
```

6.0 Data visualization

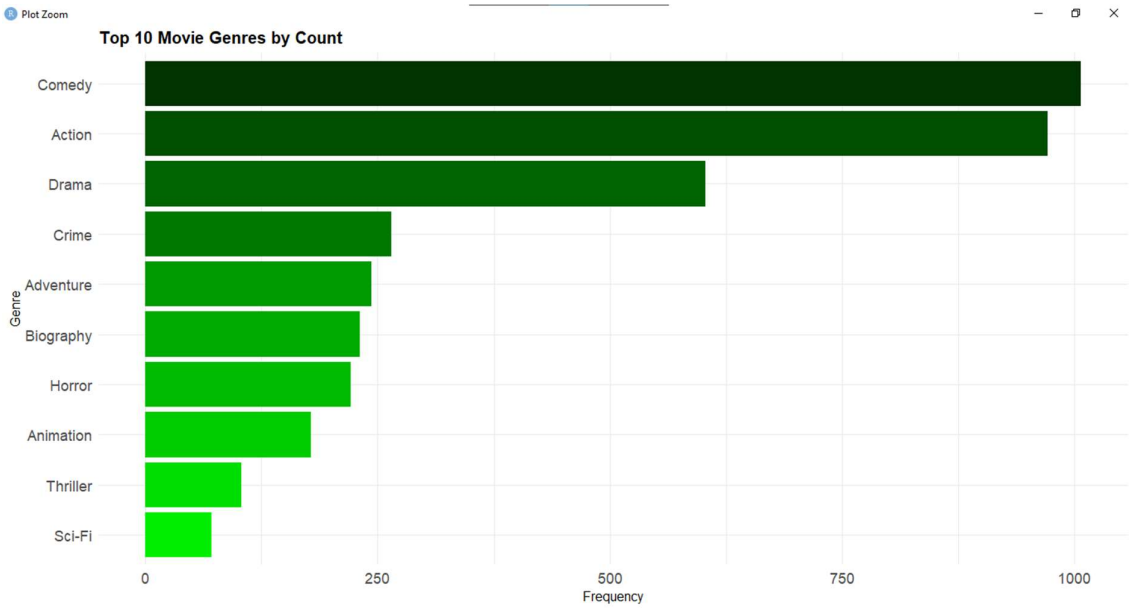
Average IMDb score



Most movies receive mid-range scores rather than being considered masterpieces or complete failures.

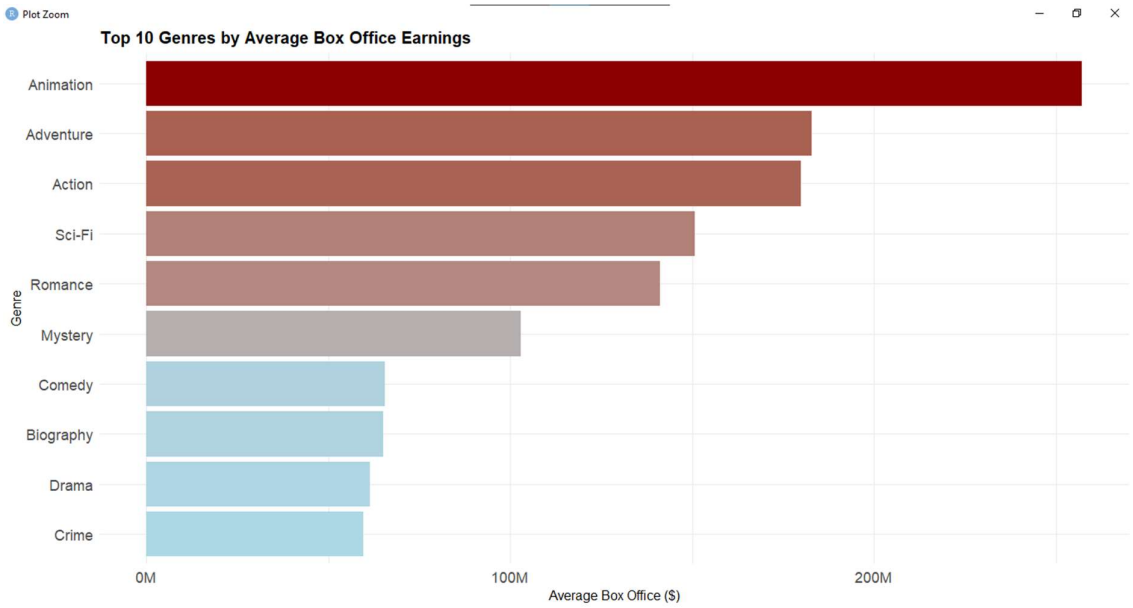
The data suggests that getting an IMDb score above 8 is rare, but there are also fewer movies rated below 3.

Number of movies by genre



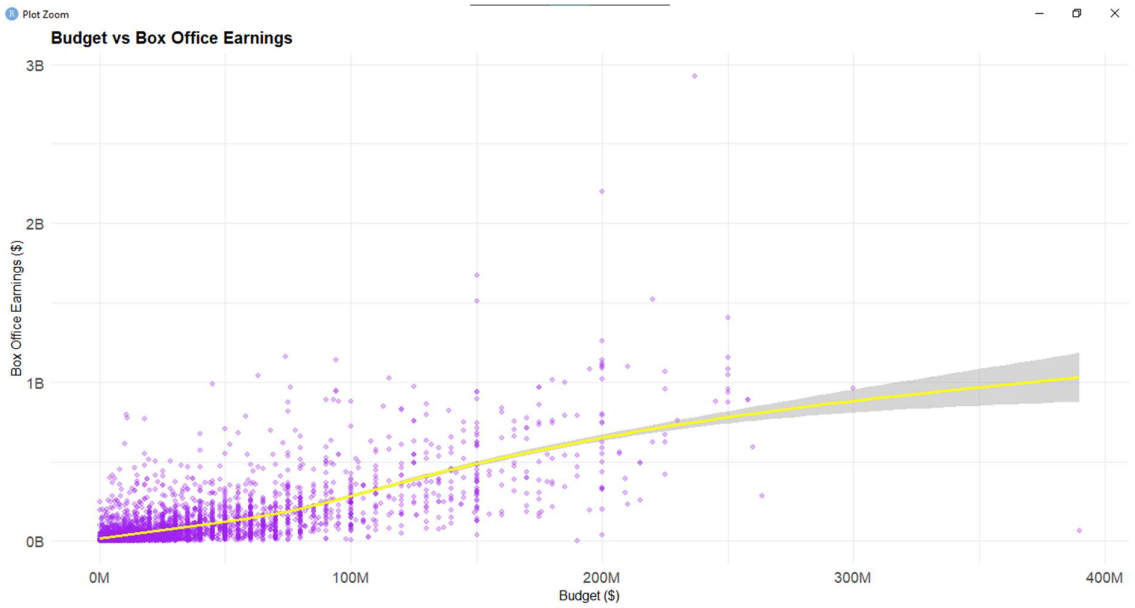
The top three genres (Comedy, Action, and Drama) are significantly more common than other genres, indicating a strong preference for these types of films in movie production.

Average box office earnings per genre



Animation has the highest average box office earnings, suggesting that animated movies are highly profitable. This could be due to their broad appeal to both children and adults, strong franchise potential, and high merchandising revenue.

Budget vs box office earnings



The yellow trend line shows a general upward trend, indicating that movies with higher budgets tend to generate higher box office earnings.

However, the relationship is not perfectly linear, meaning a higher budget does not always guarantee proportionally higher earnings.

6.5 Summarize findings

Most movies receive mid-range scores, with very few rated above 8 or below 3.

Comedy, Action, and Drama are the most common genres, indicating a strong production preference. Animation has the highest average earnings, likely due to its broad appeal, franchise potential, and merchandising opportunities.Higher budgets generally lead to higher box office earnings, but the relationship is not strictly linear, meaning a large budget does not always guarantee success.

7.0 References

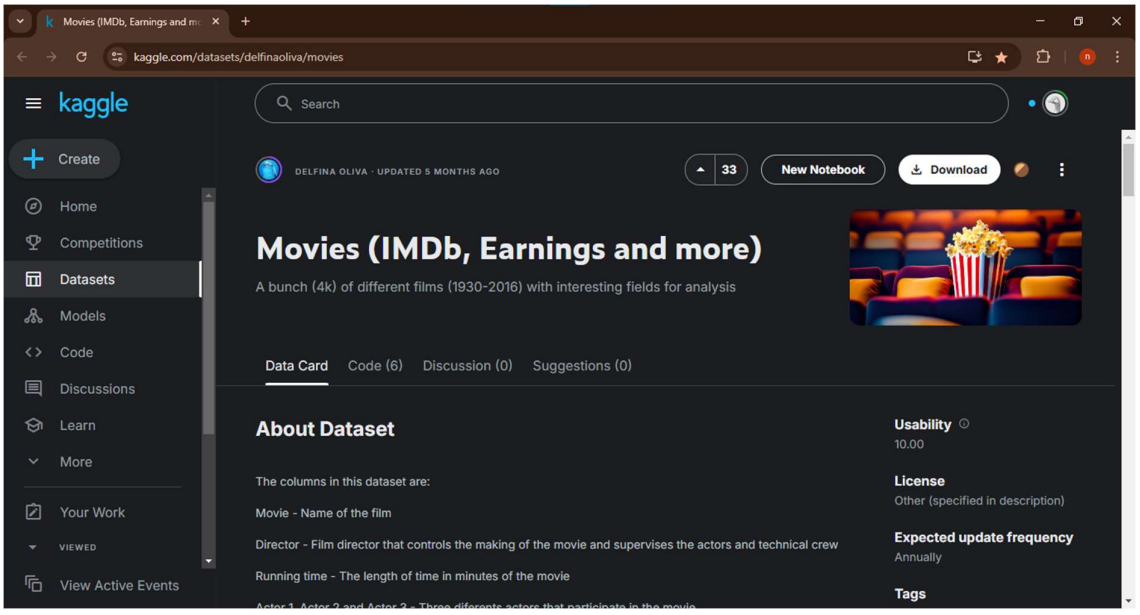
1. GeeksforGeeks. (2024, October 3). *Display an axis value in millions in ggplot using R*. GeeksforGeeks. https://www.geeksforgeeks.org/display-an-axis-value-in-millions-in-ggplot-using-r/?ref=header_outind

2. GeeksforGeeks. (2023, December 20). *Histogram in R using ggplot2*. GeeksforGeeks. https://www.geeksforgeeks.org/histogram-in-r-using-ggplot2/

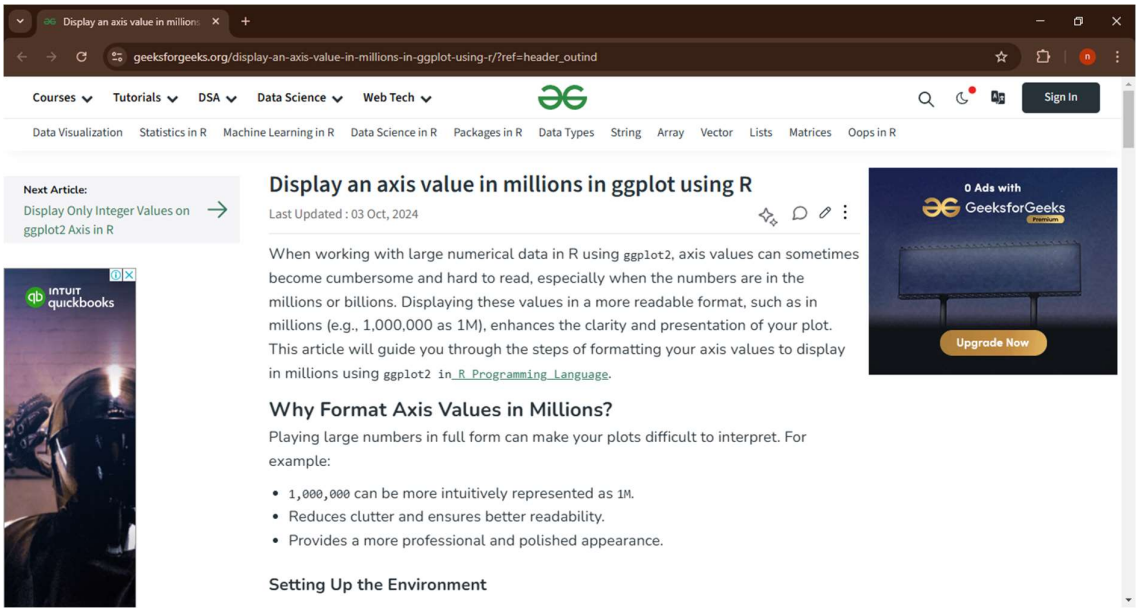
3. Pedersen, H. W. D. N. a. T. L. (n.d.). *ggplot2: Elegant Graphics for Data Analysis (3e) - 11 Colour scales and legends*. https://ggplot2-book.org/scales-colour

4. Mph, S. P. S. I., & Mph, S. P. S. I. (2024, April 24). *A practical guide to selecting top N values by group in R | R-Bloggers*. R-bloggers. https://www.r-bloggers.com/2024/04/a-practical-guide-to-selecting-top-n-values-by-group-in-r/#google_vignette

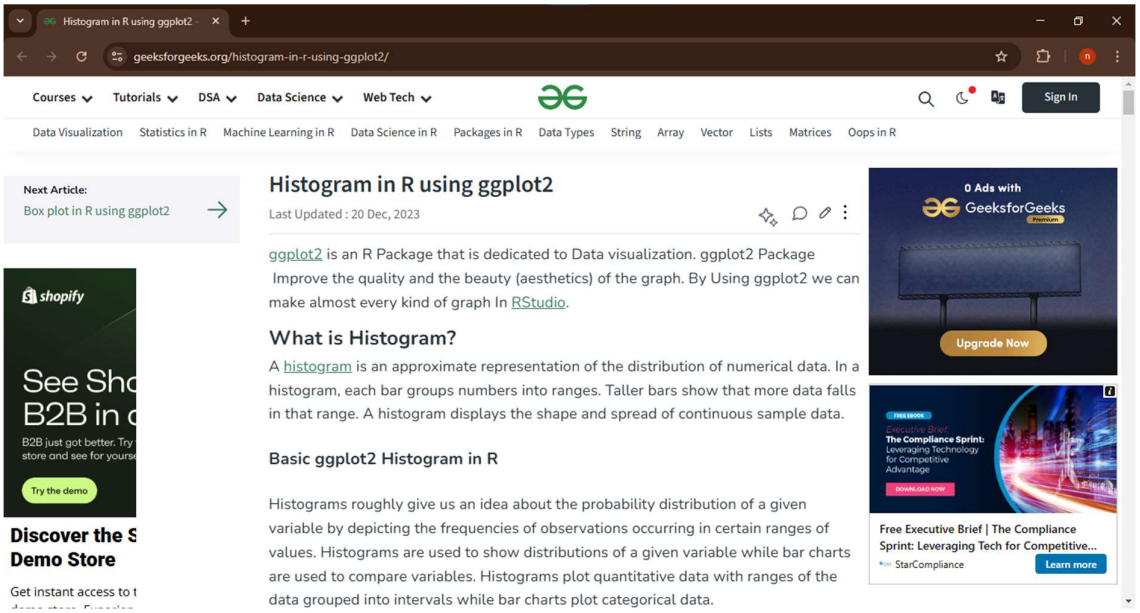
8.0 Appendix



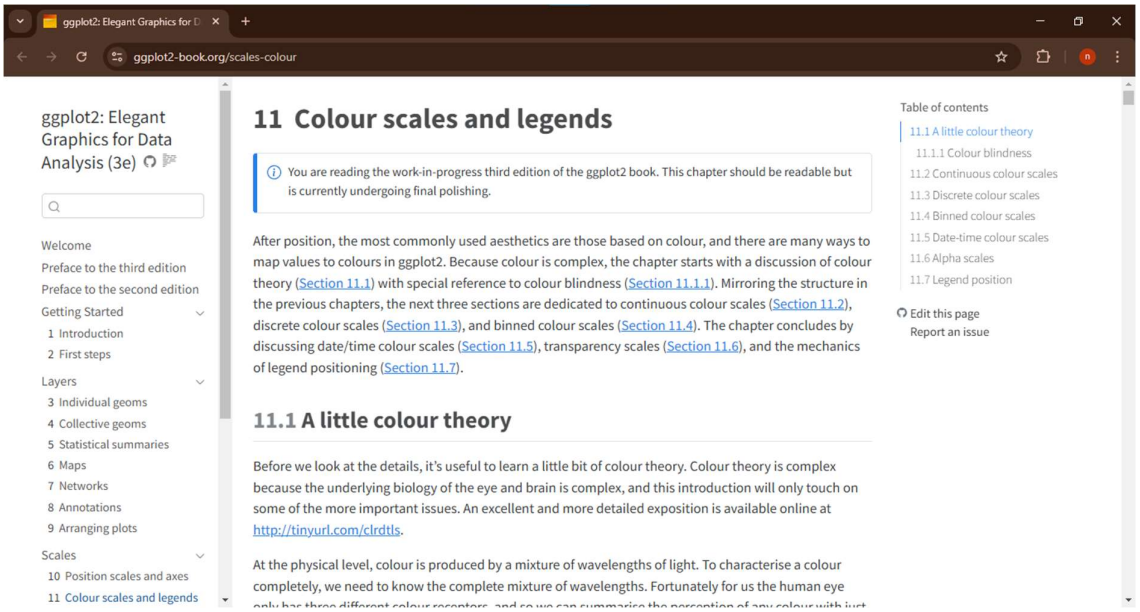
Appendix 1



Appendix 2



Appendix 3


ggplot2: Elegant Graphics for D
ggplot2-book.org/scales-colour
ggplot2: Elegant Graphics for Data Analysis (3e)
Welcome
Preface to the third edition
Preface to the second edition
Getting Started
1 Introduction
2 First steps
Layers
3 Individual geoms
4 Collective geoms
5 Statistical summaries
6 Maps
7 Networks
8 Annotations
9 Arranging plots
Scales
10 Position scales and axes
11 Colour scales and legends
11 Colour scales and legends
You are reading the work-in-progress third edition of the ggplot2 book. This chapter should be readable but is currently undergoing final polishing.
After position, the most commonly used aesthetics are those based on colour, and there are many ways to map values to colours in ggplot2. Because colour is complex, the chapter starts with a discussion of colour theory (Section 11.1) with special reference to colour blindness (Section 11.1.1). Mirroring the structure in the previous chapters, the next three sections are dedicated to continuous colour scales (Section 11.2), discrete colour scales (Section 11.3), and binned colour scales (Section 11.4). The chapter concludes by discussing date/time colour scales (Section 11.5), transparency scales (Section 11.6), and the mechanics of legend positioning (Section 11.7).
11.1 A little colour theory
Before we look at the details, it's useful to learn a little bit of colour theory. Colour theory is complex because the underlying biology of the eye and brain is complex, and this introduction will only touch on some of the more important issues. An excellent and more detailed exposition is available online at http://tinyurl.com/clrdtls.
At the physical level, colour is produced by a mixture of wavelengths of light. To characterise a colour completely, we need to know the complete mixture of wavelengths. Fortunately for us the human eye
Table of contents
11.1 A little colour theory
11.1.1 Colour blindness
11.2 Continuous colour scales
11.3 Discrete colour scales
11.4 Binned colour scales
11.5 Date-time colour scales
11.6 Alpha scales
11.7 Legend position
Edit this page
Report an issue

Appendix 4