A REPORT
BY SUBBULAKSHMI NATARAJAN

Manga & Anime

Table of Contents

1. Introduction

Project Title: Anime and Manga insights and analysis

Introduction: Recently, both children and adults have started to watch and read more anime and manga series. It's said that the characters and storylines of manga and anime differ significantly from those of Marvel, DC, and other novels. Many young people read manga and then watch the corresponding anime to visualise the story they've read. The popularity of anime saw a significant increase in 2019 [1,2], and after the COVID-19 pandemic, it became even more popular. Additionally, this project will explore the distinction between the genre preference of the user's watching anime and manga, popularity as well as score. This is why I've chosen to analyse the popular anime and manga adaptations that are currently captivating young teens and adults.

Motivation: As a huge fan of the One-Piece anime and manga, I was inspired to choose this topic. Since my childhood, I've read and watched a lot of anime, so I thought it would be fitting to explore this topic for my visualisation project, given my familiarity with the popular culture of manga and anime.

Questions: (I have changed the question according to the data source variables for better visualisation for the report)

1) How has the manga impacted anime series adaptations, considering factors such as type of media and demographic in manga compared to anime?
2) How do the anime and manga genres differ from each other in terms of score and popularity?

2. Data Wrangling

Data wrangling focuses on manipulating and transforming data from one format or structure to another to make it ready for analysis.

Description of the data and data sources

In detail I have added full data variables explanation for each data source in the appendix section 2.1 in detail for the variables that are there in the dataset.

- **Manga & Anime Dataset 2024**

Tabular data in CSV format, consisting of two files named anime.csv and manga.csv. Anime.csv contains 10,000 rows and 10 columns, while manga.csv contains 10,000 rows and 16 columns. Both datasets contain textual, numerical, and categorical data types.
Link : https://www.kaggle.com/datasets/duongtruongbinh/manga-and-anime-dataset

- **MyAnimeList Anime and Manga Dataset 2023**

Tabular data in CSV format, comprising two files named anime.csv and manga.csv. Anime.csv contains 24,000 rows and 39 columns, while manga.csv contains 64,000 rows and 30 columns. Both datasets contain textual, numerical, and categorical data types.
Link : https://www.kaggle.com/datasets/andreuvallhernndez/myanimelist

- **Anime and Manga Dataset 2023**

Tabular data in CSV format, consisting of two files named MAL_anime.csv and MAL_manga.csv. MAL_anime.csv contains 12,000 rows and 10 columns, while MAL_manga.csv contains 17,000 rows and 10 columns. Both datasets contain textual data types.
Link : https://www.kaggle.com/datasets/nikhil1e9/myanimelist-anime-and-manga

There are few steps that I took when I was doing data wrangling for each data sources:

A. Merging Multiple Datasets

We have three data sources for visualisation, each containing two CSV files for manga and anime. To merge the datasets for visualisation of each question as given in the report, I employed excel to join both datasets by unique columns.

B. Formatting the Data

Each data source had to be formatted so I splitted the datasets where there is an error and how it will be formatted using Python and excel. For further information about how I did in python go to the section appendix 2.3 for knowing how I formatted the data.

➢ **Anime and Manga (2023)**

I used a simple tool like Excel for cleaning the dataset. For error checking, I used Python. In the Anime and Manga Dataset 2023, specifically in the MAL Manga and MAL Anime CSV files, I observed question marks "?" in MAL_manga and "?" in MAL_anime. I replaced these question marks with “Not known" using Python to make the data consistent. The screenshot of the error checking is given in the appendix.

The following datasets in the data source had few formatting errors:

- **Manga Dataset**

I replaced the question marks with "unknown" for better clarity. I also adjusted the dataset in Excel to ensure consistency in data when using it for Tableau or R for visualisation purposes.

- **Anime Dataset**

Similarly, I detected and cleared question marks using Python and then Excel for the episode column, analogous to the manga dataset's volume column.

C. **Dropping Unnecessary Columns**

➢ **Anime and Manga (2023)**

Using Excel, I filtered out unwanted columns like image URL and page URL, which are unnecessary for visualisation purposes. After filtering, we focused on columns like "ID," "Title," "Rank," "Type," "Volume" (for manga dataset), and "Episode" (for anime dataset), "Published," "Member," and "Score."

➢ **Manga and Anime (2024)**

I removed unwanted columns in these datasets to streamline them for the data visualisation section. There were not many errors found in the dataset since the data was already processed.

➢ **MyAnimeList Anime and Manga**

In this data source, I had to remove unnecessary columns from the Manga and Anime dataset.

In both datasets, we had to remove the columns such as members, favourites, episode duration, total duration, created at, updated at, start year, start season, real start date, real end date, broadcast day, broadcast time, studios, producers, licensors, synopsis, background, main picture, URL, trailer URL, title English, title Japanese.

From removing these columns from each respective dataset, I could take a good visualisation graph as per the question for this project.

D. Changing Column Names

To see the code that I used for changing the column name. Check the screenshot appendix 2.1.

➢ **Anime and Manga (2023)**

I changed "Unnamed_0" to "ID" for data consistency, as "Unnamed_0" didn't make sense in both the anime and manga dataset I have change it.

➢ **My AnimeList Anime and Manga**

I corrected "on_hitatus" to the "ongoing" category in the status column in both manga and anime datasets using Excel.

3. Data Checking

Data checking, on the other hand, focuses on ensuring the quality and accuracy of the data. This involves identifying and correcting errors and inconsistencies in the data.
Here are some common data checking methods:

A. Identifying missing values

In the following data sources, which I have cited in the report, I did data checking according to the following steps.

➢ **My AnimeList Anime and Manga**

In the manga dataset, I replaced blank columns in the end date and start date with "Present" for clarity and consistency. Similarly, blank values in the volume and chapter columns were replaced with "unknown" for consistency. For columns like genre, themes, and demographic, where "[]" indicated no assigned value, I replaced it with "unknown." Similarly for the anime dataset also it is the same case as the manga dataset where I replaced blank columns with Unknown and for the end date it should be “Present”.

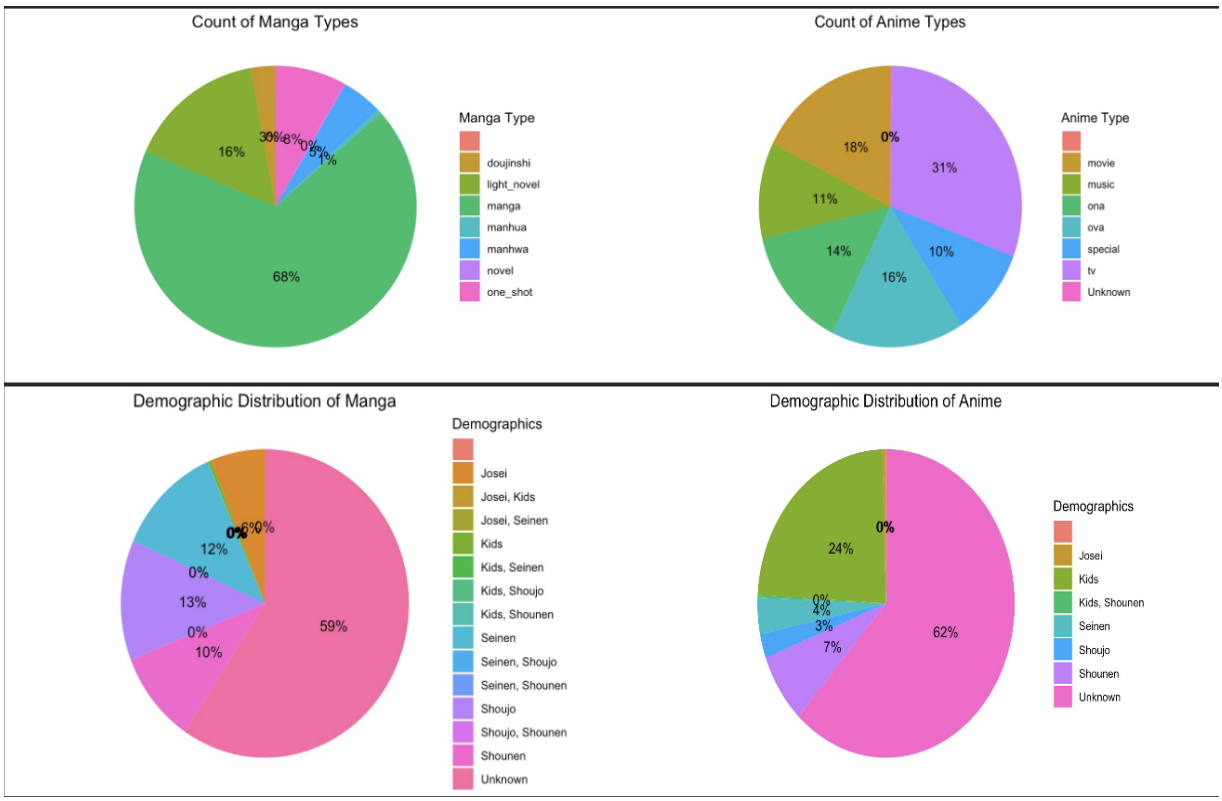
B. Identifying and correcting inconsistencies in data format

For identifying and correcting the data format in the data sources, I have already performed this process in the data wrangling steps so I think for now there is no errors that are found in this step.

4. Data Exploration

I will be using R to get the results of the questions that we are focusing on these questions.

1) How has the popularity of manga impacted anime series adaptations, considering factors such as type of media and genre in manga compared to anime?



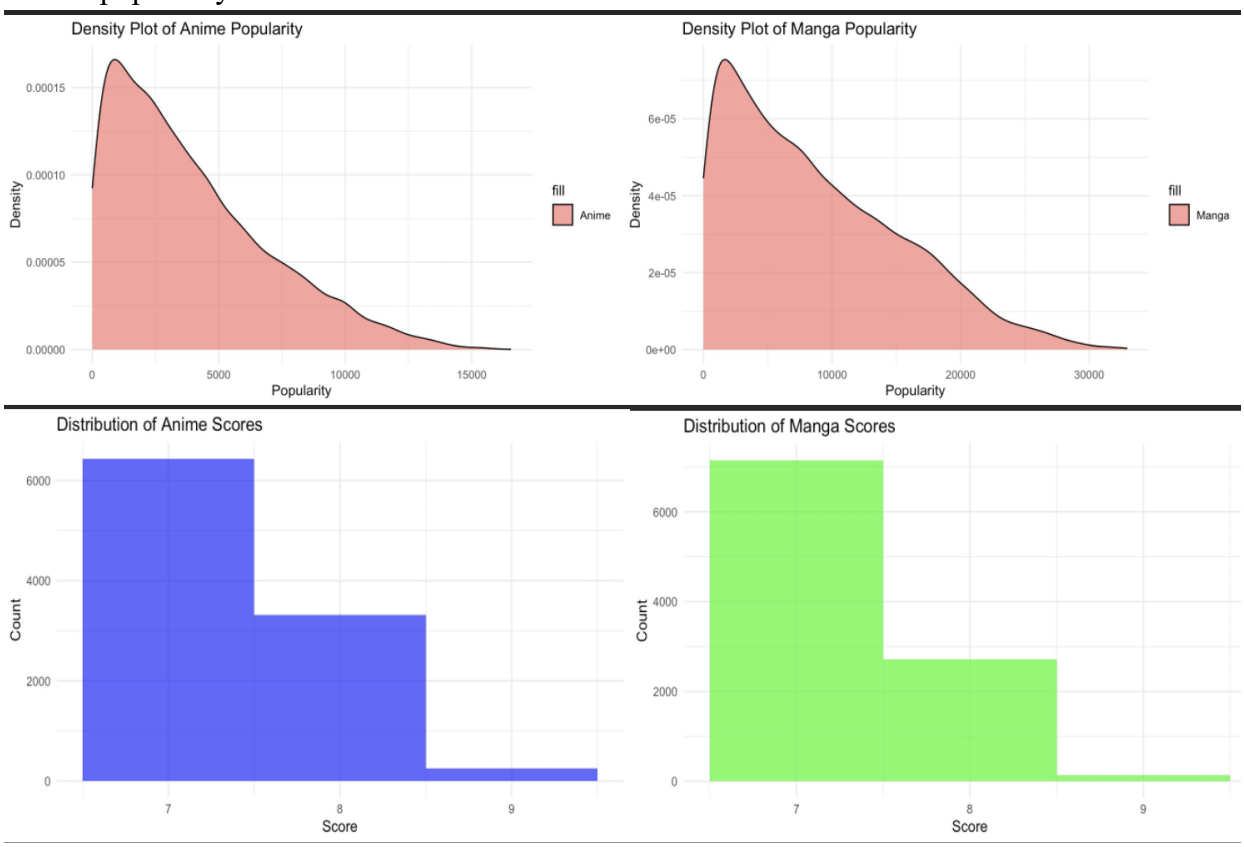
In this question, we must analyse the demographic and type of media in manga and anime using the data source given in the report. I will be using my anime and manga dataset to visualise the graph for the question I will be analysing. To see the type of media most people prefer is Manga and second is light novel. This is mainly because in the light novel and manga they visually are engaging to the readers but whereas in the anime it is the opposite since the anime is like a movie adaptation of a manga novel. So, from the graph we can see that most of the people watch this anime through TV channels where they broadcast the anime and second most is movies which are rated too higher. Since anime movies are being released in many theatres these days many anime binges who love watching anime go for these. For screenshot of the code please go to appendix (4.2) to see how I created the graph.

Whereas, when we investigate demographics for both manga and anime, we can see that Shounen is the most popular in terms of the manga and unknown in anime. The Shounen demographic is very popular because it is about a young male centric storyline which majority of the people read this book. Shounen manga has had a profound

cultural impact in Japan and beyond. Many iconic and influential manga series, such as "Dragon Ball," "Naruto," and "One Piece," fall under the shounen category. On the other hand, when it comes to anime, most of them voted for the unknown category since they don't know whether the category will include shounen or either Josei. But it can be said that kids watch most of the anime since anime is meant for all ages including kids.

So, comparing these two graphs we can see that anime has been chosen because most of the people prefer watching anime through the allocated time that channel broadcasts or releases in the theatre. Also, the survey has shown that because of its compelling storyline and diverse array of recognizable and endearing characters, anime is now more widely known and engaged with than many other aspects of Japanese culture.

2) How do the anime and manga genres differ from each other in terms of score and popularity?



In this question we will be considering the two factors one is popularity and other is score. So, for visualising these two graphs we will be using anime and manga 2024 dataset. So first we will focus on popularity for anime and manga then we will focus on score. For screenshot of the code please go to appendix (4.1) to see how I created the graph.

As we can see that for popularity for anime and manga we have used density plots. It can be said that the popularity distribution indicates that most manga fall within the range of 0 to

30000, encompassing over 600 to 800 manga. But on the other hand, it can be said that anime is said to be the most popular among young adults. It is the same as manga. So, we cannot originally prove the question that we have indicated in the report since the data given in the Kaggle is inappropriate when visualising the manga and anime's popularity.

Next going to score, it can be said that we have taken in consideration of scale of 7 to 9 where it can be said that the highest point is 7 where the least was 9 but it can be said that histogram shown in the report is not in appropriate to show the highest score between in the anime and manga so we cannot distinguish through histogram or any graph for telling the highest score.

5. Conclusion

Through this report, what I have done is data wrangling, checking, and exploring the data. When I delved into the data that is the world of anime and manga, I have sought into many insights such as their impact, popularity, and genre preference. The project mainly focus on how the manga and anime is linked together, explore demographic and media preference, and compare genre difference in terms of their score and popularity.

When we first merge multiple datasets from different data source and format the data to ensure proper data consistency and clarity. Then after that we will be doing data checking where we will be addressing the missing values and data inconsistency. After we finish all these steps, we will be going to data exploration where we will be answering to the question that mentioned in the report using R. The two questions that we will be mentioning are as follows:

1. **The influence of manga on the adaptation of anime series:**

Manga continues to be the favoured medium, especially among young adults and readers of the shounen genre, according to our analysis of demographic and media preferences. Despite their popularity, anime adaptations frequently stick true to the manga's plot, demonstrating the power and allure of manga stories. The preference for watching anime on television over going to the movies emphasises how accessible and popular animated adaptations have remained over time.

2. **Comparison of Anime and Manga Genres in Terms of Score and Popularity:**

Nevertheless, the data created difficulties when attempting to compare preferences for different genres in terms of popularity and score. Density plots showed popularity distributions, but they were unable to definitively differentiate between manga and anime fans. In a similar vein, the score comparison histograms were unclear, which made it challenging to determine with certainty which genres were preferred based only on rating.

In conclusion, when I was this project, I found many valuable insights and analysis into the world of anime and manga where there was a lot of limitation during the data analysis process.

Furthermore, when I was doing research and data processing such as data wrangling and data checking had given good robust conclusion about the genre preference and their impact on manga and anime adaptions. Nonetheless, this project gave me good understanding of many dynamic relationship between manga and anime and highlighted the popularity of both medium among the audience.

6. Reflection

When I was doing this project, I had many valuable insights into the world of anime and manga. One of lesson that I learned during this project is the importance of data wrangling and checking process to ensure data quality and consistency for the dataset that can be used for data analysis.
Second is that data wrangling stage was proved to be more challenging since it was dealing with formatting errors and inconsistency across multiple data sources. While other tools like Excel and Python were key elements to identify the errors and clean the data as well as formatting the dataset.

The data wrangling stage proved to be more challenging than anticipated, particularly in dealing with formatting errors and inconsistencies across multiple datasets. While tools like Excel and Python were instrumental in cleaning and formatting the data, it became evident that having a standardised approach to data collection and storage could streamline this process in future projects.

Interesting trends and patterns were also found throughout the data investigation phase, including the influence of manga popularity on anime adaptations and the distinctions in genre preferences between anime and manga consumers. These results offered insightful information about trends and audience preferences in the anime and manga community.

To get deeper insights from the data, future studies could consider incorporating more sophisticated statistical analysis approaches. Even while this project's descriptive analysis yielded insightful results, more sophisticated techniques like regression analysis or clustering could improve our comprehension of the variables impacting the popularity of anime and manga.

Overall, this project was a valuable learning experience that deepened my understanding of data analysis techniques and the intricacies of the anime and manga industry. Moving forward, I will apply the lessons learned from this project to future endeavours, ensuring a more robust and comprehensive approach to data analysis and interpretation.

7. Bibliography

GeeksforGeeks. (n.d.). Data Wrangling in Python. Retrieved from https://www.geeksforgeeks.org/data-wrangling-in-python/

ProjectPro. (n.d.). Data Wrangling. Retrieved from https://www.projectpro.io/article/data-wrangling/806#:~:text=Python%20comes%20with%20a%20rich,merging%20datasets%2C%20and%20reshaping%20data.

Simplilearn. (2019, November 12). Merging DataFrames in Pandas | Concatenation [Video]. YouTube. https://www.youtube.com/watch?v=x0Sy6Kl0Mzw

Kaggle. (n.d.). Anime and Manga Dataset. Retrieved from https://www.kaggle.com/code/hbsaakashyadav/anime-and-manga-dataset

Data Carpentry. (n.d.). Merging DataFrames. Retrieved from https://datacarpentry.org/python-ecology-lesson/05-merging-data.html

Kiran, A. (2019, July 25). Merging Two Dataset Together For Visualization [Video]. YouTube. https://www.youtube.com/watch?v=uPDcx0wUBIY

DataCamp. (n.d.). Merging DataFrames in R. Retrieved from https://www.datacamp.com/tutorial/merging-datasets-r

Exploring the Anime and Manga Global Takeover. (n.d.). Brandwatch. https://www.brandwatch.com/blog/anime-manga-global-interest/#:~:text=For%20the%20January%20to%20July,1%25%20being%20%E2%80%9Cangry%E2%80%9D

Brzeski, P. (2022, October 24). *How Japanese Anime Became the World’s Most Bankable Genre*. The Hollywood Reporter. https://www.hollywoodreporter.com/business/business-news/japanese-anime-worlds-most-bankable-genre-1235146810/

r-charts.com. (n.d.). Pie Chart Percentages with ggplot2. Retrieved from https://r-charts.com/part-whole/pie-chart-percentages-ggplot2/#pie

Data wrangling: What it is & why it's important. (2021, January 19). Business Insights Blog. https://online.hbs.edu/blog/post/data-wrangling

8. Appendix

2.1 Data source variables description

1) Manga and Anime dataset 2024

Data field description

- **Title**: Title of the anime (in both Japanese and English).
- **Score**: Score given by users (out of 10).
- **Votes**: Number of user votes for the anime
- **Ranked**: Rank of the anime based on score.
- **Popularity**: Popularity rank.
- **Episodes**: Number of episodes.
- **Status**: Current airing status (e.g., Finished Airing).
- **Aired**: Airing period.
- **Premiered**: Premiere season and year.
- **Producers**: Production companies.
- **Licensors**: Licensing companies.
- **Studios**: Animation studios.
- **Source**: Source material (e.g., Manga).
- **Duration**: Duration per episode.

2) Anime and Manga Dataset 2023

- Title: Name of the anime/manga
- Rank: Ranking of the anime/manga
- Type: Category of anime/manga e.g. One-shot, Light novel, etc
- Episodes(*anime*): Number of episodes of the anime
- Volumes(*manga*): Number of volumes of the manga
- Aired(*anime*): Date of airing of an anime.
- Published(*manga*): Date of publishing of a manga.
- Members: Number of members who have watched/read the anime/manga
- Page URL: The URL link to the page of the anime/manga.
- image URL: The URL link to the cover image of the anime/manga.
- Score: Average user rating/score of the anime/manga

2.2 Changing column names (screenshot)

```
[50]
anime_dataset.rename(columns = {'Unnamed:_0':'id'}, inplace = True)
manga_dataset.rename(columns = {'Unnamed:_0':'id'}, inplace = True)
Python
```

2.3 Formatting data (Screenshot)

Anime dataset replacing "?" to Not known.

```
[57]
anime_dataset['Episodes']=anime_dataset['Episodes'].replace(['?'], 'Not Known')
Python

[56]
anime_dataset[anime_dataset['Episodes']=='?']
Python
```

Manga dataset replacing "?" to Not known.

```
[36]
manga_dataset[manga_dataset['Volumes']=='?']
Python

+ Code + Markdown

[37]
manga_dataset['Volumes'] = manga_dataset['Volumes'].replace(['?'], 'Not Known')
Python
```

4.1 Visualizing the pie chart of demographic and type of media (R codes)

```
library(ggplot2)

# Load manga dataset
manga <- read.csv("manga_cleaned data.csv")

# Load anime dataset
anime <- read.csv("anime_cleaneddata.csv")

# Count the occurrences of each manga type
manga_type_counts <- manga %>%
  group_by(type) %>%
  summarise(count = n())

# Count the occurrences of each anime type
anime_type_counts <- anime %>%
  group_by(type) %>%
  summarise(count = n())

# Pie chart for count of manga types with percentage labels
print(ggplot(manga_type_counts, aes(x = "", y = count, fill = type)) +
        geom_bar(stat = "identity", width = 1) +
        coord_polar("y") +
        labs(title = "Count of Manga Types",
             fill = "Manga Type",
             y = "Count") +
        theme_void() +
        theme(legend.position = "right",
              plot.title = element_text(hjust = 0.5)) +
        geom_text(aes(label = paste0(round(count/sum(count) * 100), "%")),
                  position = position_stack(vjust = 0.5)))

# Pie chart for count of anime types with percentage labels
print(ggplot(anime_type_counts, aes(x = "", y = count, fill = type)) +
        geom_bar(stat = "identity", width = 1) +
        coord_polar("y") +
        labs(title = "Count of Anime Types",
             fill = "Anime Type",
             y = "Count") +
        theme_void() +
        theme(legend.position = "right",
              plot.title = element_text(hjust = 0.5)) +
        geom_text(aes(label = paste0(round(count/sum(count) * 100), "%")),
                  position = position_stack(vjust = 0.5)))
```

```
# Create pie chart for manga demographic distribution
manga_demographic_counts <- manga %>%
  group_by(demographics) %>%
  summarise(count = n())

print(ggplot(manga_demographic_counts, aes(x = "", y = count, fill = demographics)) +
        geom_bar(stat = "identity", width = 1) +
        coord_polar("y") +
        labs(title = "Demographic Distribution of Manga",
             fill = "Demographics",
             y = "Count") +
        theme_void() +
        theme(legend.position = "right",
              plot.title = element_text(hjust = 0.5)) +
        geom_text(aes(label = paste0(round(count/sum(count) * 100), "%")),
                  position = position_stack(vjust = 0.5)))

# Create pie chart for anime demographic distribution
anime_demographic_counts <- anime %>%
  group_by(demographics) %>%
  summarise(count = n())

print(ggplot(anime_demographic_counts, aes(x = "", y = count, fill = demographics)) +
        geom_bar(stat = "identity", width = 1) +
        coord_polar("y") +
        labs(title = "Demographic Distribution of Anime",
             fill = "Demographics",
             y = "Count") +
        theme_void() +
        theme(legend.position = "right",
              plot.title = element_text(hjust = 0.5)) +
        geom_text(aes(label = paste0(round(count/sum(count) * 100), "%")),
                  position = position_stack(vjust = 0.5)))
```

4.2 Histogram for distribution of anime and manga score and density plot for anime and manga popularity

```
# Load necessary libraries
library(ggplot2)
library(tidyverse)

# Load manga dataset
manga <- read.csv("manga cleaned.csv")

# Load anime dataset
anime <- read.csv("anime_cleaned.csv")

# Plot histogram for anime scores
ggplot(anime, aes(x = Score)) +
  geom_histogram(binwidth = 1, fill = "blue", alpha = 0.7) +
  labs(title = "Distribution of Anime Scores",
       x = "Score",
       y = "Count") +
  theme_minimal()

# Plot histogram for manga scores
ggplot(manga, aes(x = Score)) +
  geom_histogram(binwidth = 1, fill = "green", alpha = 0.7) +
  labs(title = "Distribution of Manga Scores",
       x = "Score",
       y = "Count") +
  theme_minimal()

# Plot density plot for anime popularity
ggplot(anime, aes(x = Popularity, fill = "Anime")) +
  geom_density(alpha = 0.7) +
  labs(title = "Density Plot of Anime Popularity",
       x = "Popularity",
       y = "Density") +
  theme_minimal()

# Plot density plot for manga popularity
ggplot(manga, aes(x = Popularity, fill = "Manga")) +
  geom_density(alpha = 0.7) +
  labs(title = "Density Plot of Manga Popularity",
       x = "Popularity",
       y = "Density") +
  theme_minimal()
```