

NANYANG TECHNOLOGICAL UNIVERSITY

SINGAPORE

MH3511 Data Analysis with Computer - Group Project

Name	Matriculation Number
Cao Shuwen	U1922953B
Joelle Thng	U1920277K
Luo Wenyu	U1922009G
Tan Ching Fhen	U1920787D
Wang Anyi	U1922992C
Wei Yao	U1921669G

Abstract

Anime (Japanese: Anime) is hand-drawn, and computer animation originated in Japan. Animations are delivered directly to home media in a dramatic way through television broadcasts and over the Internet. In addition to the original work, anime is usually an adaptation of Japanese manga, light novels, or video games. The animation industry is divided into 430 production companies, including large studios such as Studio Ghibli, Sunrise and Toei Animation. Since the 1980s, the media has also achieved international success with the rise of foreign dubbing and subtitled programming. As of 2016, Japanese anime accounts for 60% of the world's anime TV shows.

Table of Content

Abstract

1. Introduction	1
2. Data Description	2
3. Description and Cleaning of Dataset	2
3.1 Summary statistics for the Main Variable of Interest: Score	3
3.2 Summary statistics for other variables	3
3.2.1 Popularity	4
3.2.2 Members	4
3.2.3 Favorites	4
3.2.4 Duration_min	4
3.2.5 Number of episodes of anime	5
3.2.6 Year of airing	5
3.2.7 Season of airing	5
3.2.8 Type	6
4. Statistical Analysis and Tests	6
4.1 Score VS Genre	6
4.2 Score VS Type	7
4.3 Score VS Year	9
4.4 Score VS Season	11
4.5 Score VS Continuous Data	13
4.6 Multiple Linear Regression	15
5. Conclusion and Discussion	16
6. Appendix	17

1. Introduction

In recent years, anime has seen a surge in popularity across the world. With major streaming services like netflix incorporating anime into their platforms, anime popularity continues to amplify. In 2020, anime audiences on netflix have more than doubled. Therefore, we seek to understand more about anime through statistical analysis. Particularly, our objective is to investigate what makes an anime well-liked?

This report will detail the data description and statistical analysis using R. The dataset we utilise consists of information on different animes, including their scores, duration, genre and many other variables. Among these columns, we picked the 'score' attribute as our target variable, which is indicative of how well-liked the anime is in general. In particular, we would like to answer the following questions:

1. Is the score of an anime dependent on its genre? If so, which genre(s) have statistically significant impact among others?
2. Is the score of an anime dependent on its type of release (eg. TV, movie, OVA etc)? If so, which type has the highest score?
3. Is the score of an anime dependent on its year of release? If so, in which year did anime have higher scores?
4. Is the score of an anime dependent on its season of air? If so, in which seasons did anime have higher scores?
5. How are the continuous variables episodes, popularity, members, favorites and duration related to score? Which numeric variable is the strongest determinant of score?

2. Data Description

The dataset was crawled from MyAnimeList (<https://myanimelist.net>), which is a representative sample of the internet anime community. The dataset contains a list of anime, with title, title synonyms, genre, studio, licensor, producer, duration, rating, score, airing date, episodes, source (manga, light novel etc.) and many other important data about individual anime. The dataset as a whole contains 14,478 unique anime.

3. Description and Cleaning of Dataset

The following data cleaning steps were performed in Python to further process the original dataset:

1. Columns which had extremely high cardinality were removed e.g production studio.
2. Animes with no assigned genre (i.e missing values) were dropped .
3. Animes that were still being aired were dropped - we will narrow our focus to animes that have finished airing.
4. Genres were one-hot-encoded i.e converted into boolean columns.
5. The 'aired' column which describes the period during which an anime was aired was converted to a categorical variable, 'season'.
6. Severely unbalanced categorical variables such as some genres that were 95% True and 5% False, were dropped.
7. Remaining columns were dropped to keep to a manageable number of variables.

The cleaned dataset has 4656 observations and 56 variables - 6 are numeric and the remaining are categorical. The variables are described below:

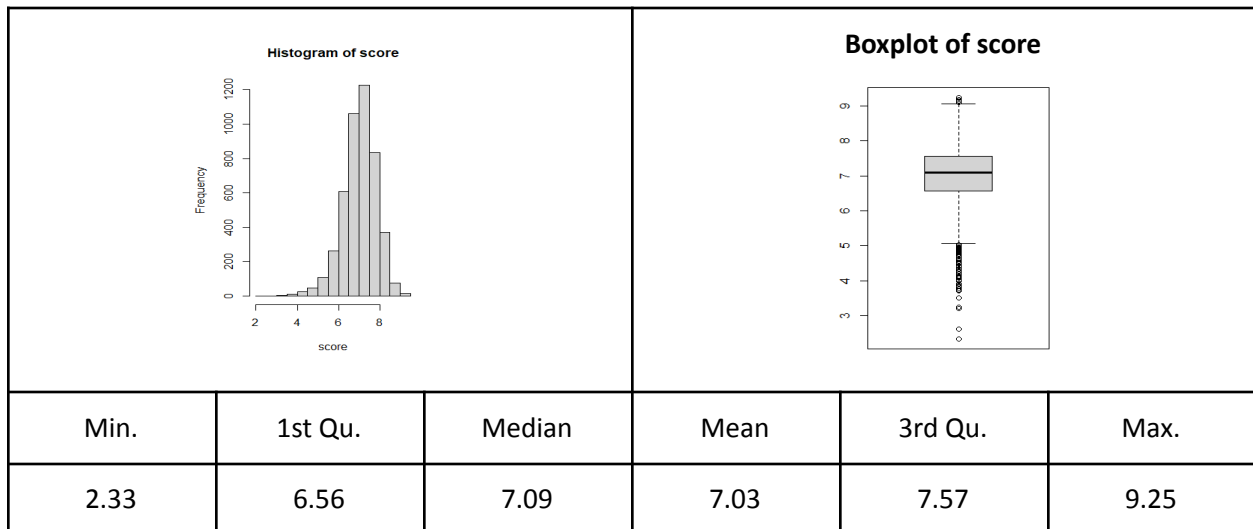
1. Score (target variable) - how well-liked the anime is
2. Episodes - number of episodes
3. Popularity - popularity ranking on MyAnimeList.net
4. Members - number of users who subscribed on MyAnimeList.net

5. Favorites - number of users who favorited on MyAnimeList.net
6. Duration - length of a single episode in minutes
7. Type - type of release e.g TV or movie
8. Year - year of airing
9. Season - season of airing e.g winter
10. Genre - genre (accounts for the remaining columns)

Additionally, a separate csv file was created containing score and a single genre column. The genre columns (in 10.) were converted into a single column using python in order to construct the boxplots and analysis for genres in section 4.1.

3.1 Summary statistics for the Main Variable of Interest: Score

The following plots show the overall distribution of the variable Score.

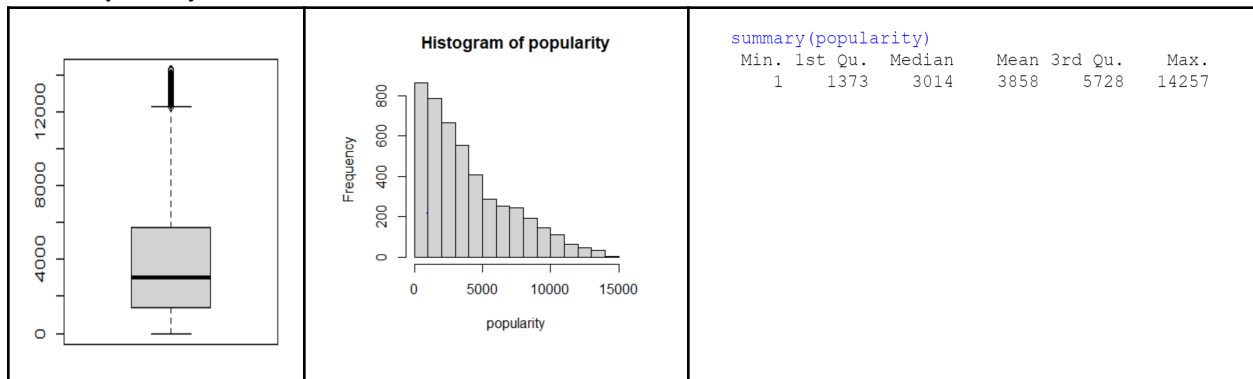


The variable score has a slight negative skew. However, we will not perform any transformation as the tests as the statistical tests we perform are robust to slight deviations from normality. In subsequent statistical tests, we will assume that that population normality holds.

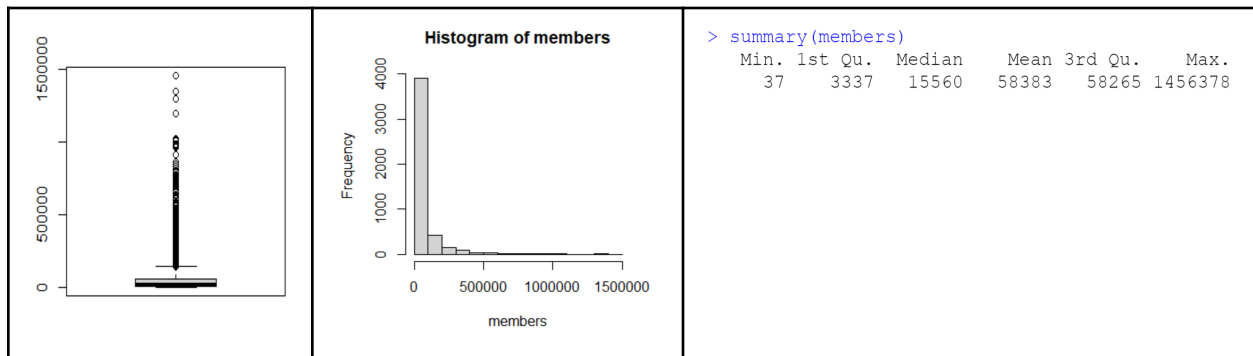
3.2 Summary statistics for other variables

The following sections display the summary statistics and plots for each of the dependent variables.

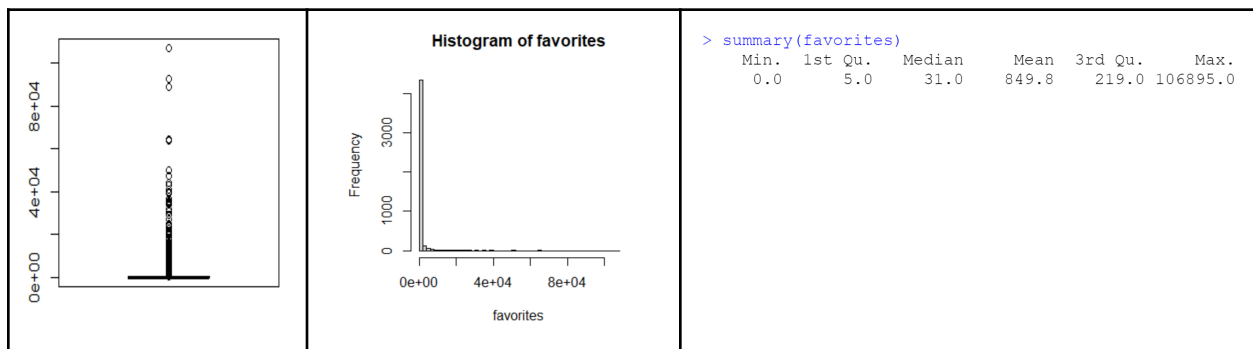
3.2.1 Popularity



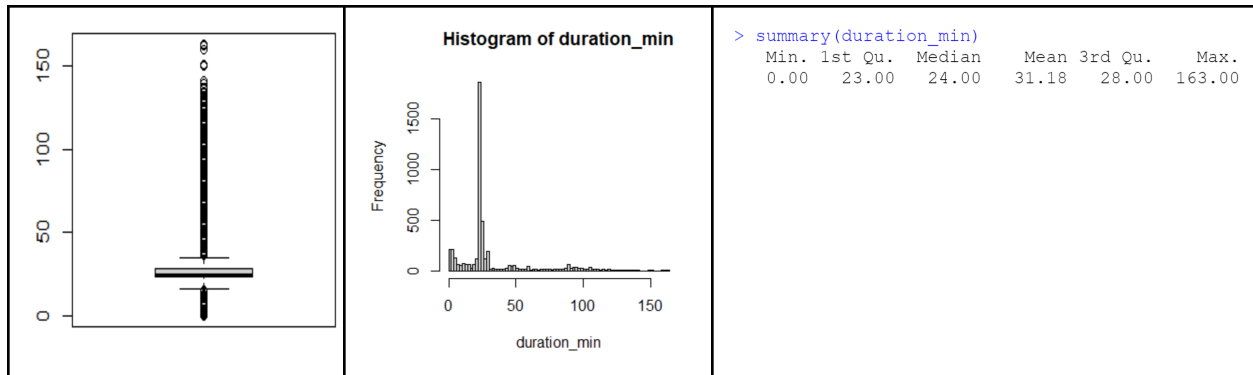
3.2.2 Members



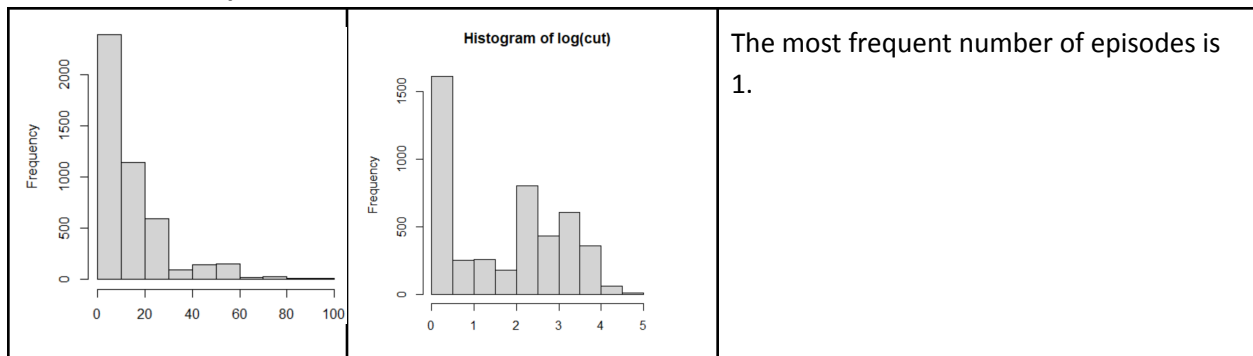
3.2.3 Favorites



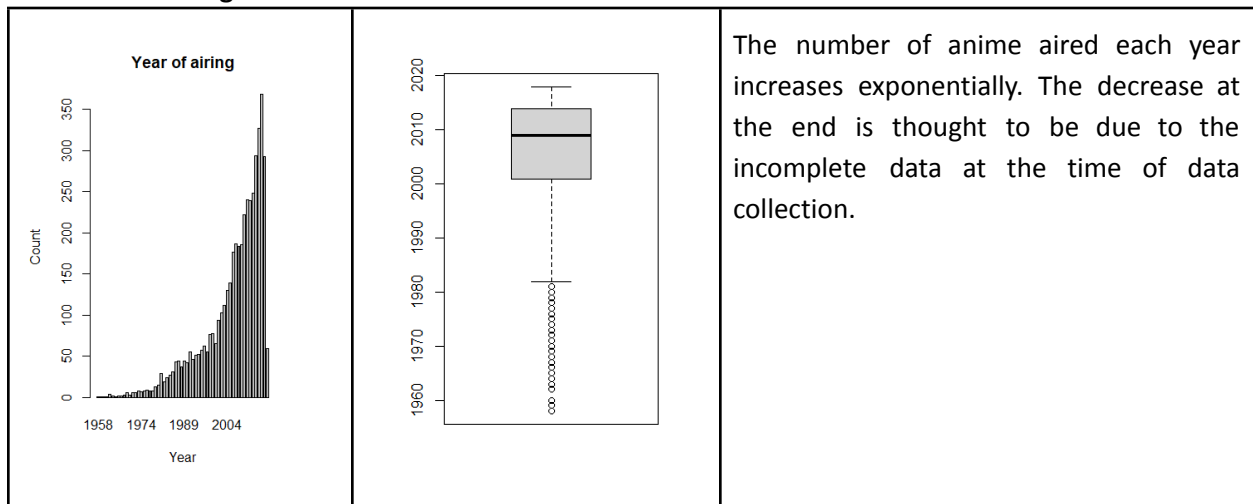
3.2.4 Duration_min



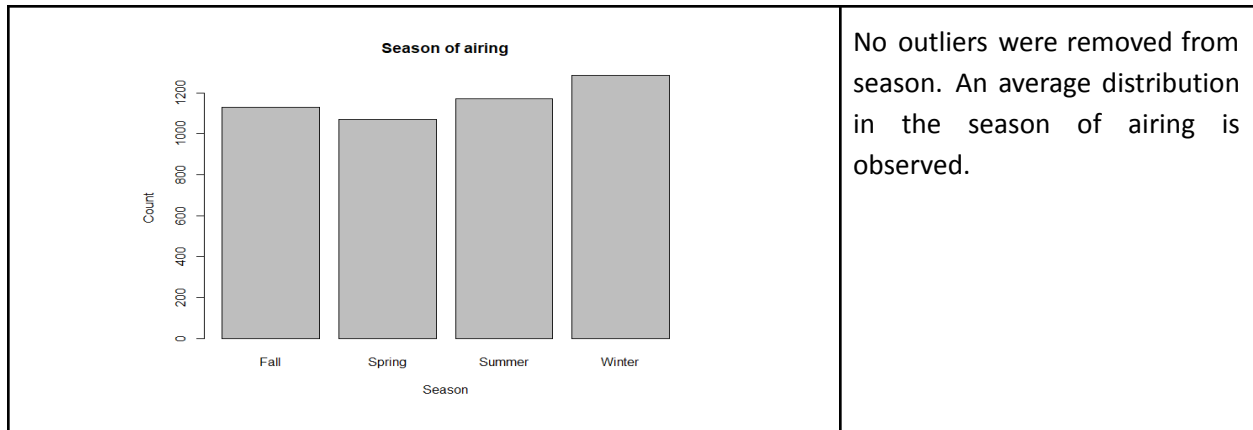
3.2.5 Number of episodes of anime



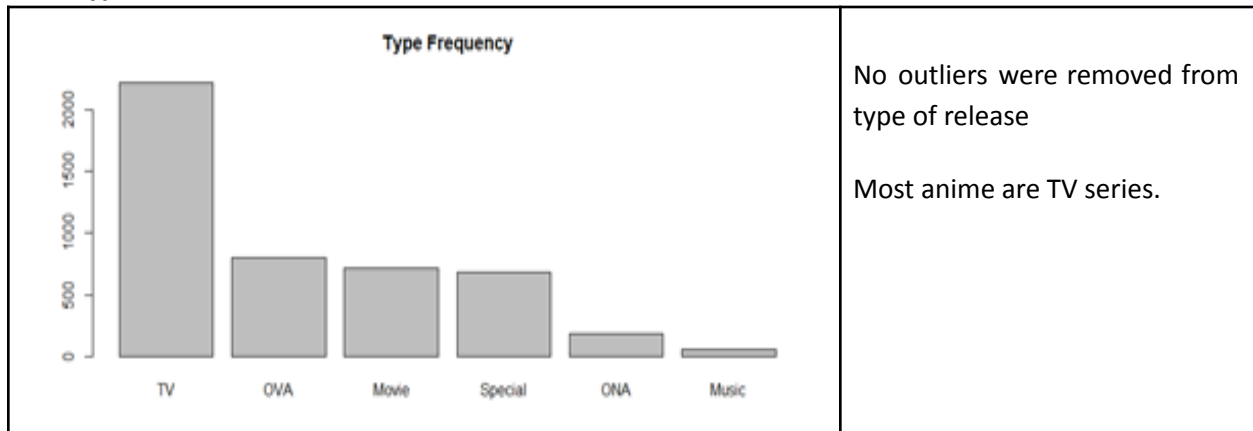
3.2.6 Year of airing



3.2.7 Season of airing



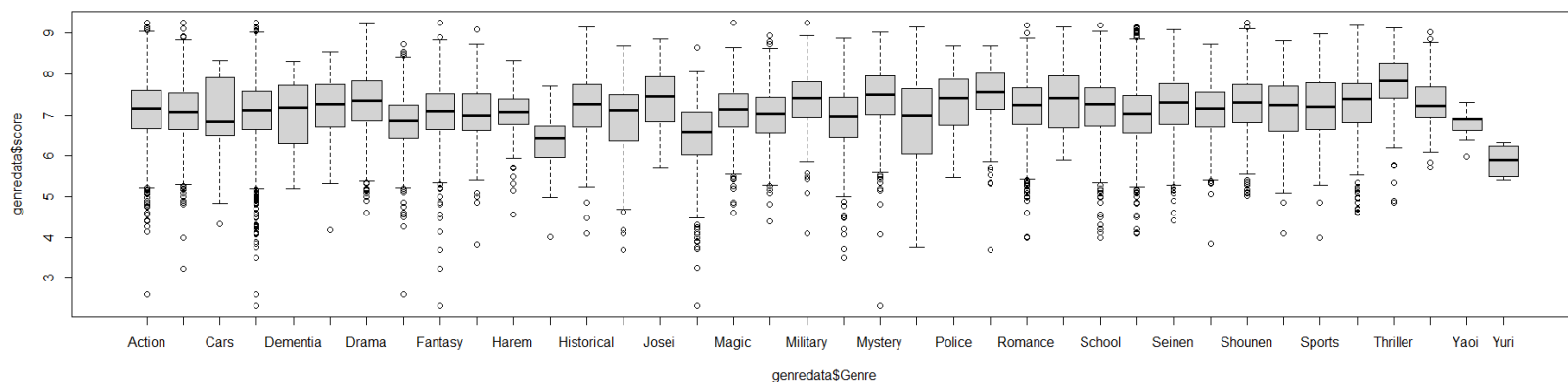
3.2.8 Type



4. Statistical Analysis and Tests

4.1 Score VS Genre

This section will address the research questions, “Is the score of an anime dependent on its genre? If so, which genre(s) have statistically significant impact among others?”



Based on the boxplot, we observe that different genres have different medians, so we hypothesize that scores are dependent on genres. Also, based on the boxplot, we will assume that the variances are unequal between different genres. Since, Anova test assumes the homogeneity of variances, we decide that Anova test is not suitable for this particular variable. Instead, we will use Kruskal Wallis Test to test the equality of the median score between different genres.

$H_0: \mu_{Action} = \mu_{Comedy} = \dots = \mu_{Yuri}$ for all genres, where μ refers to the median score

H_1 : median score of at least one genre is not equal to the rest

```
> kruskal.test(genredata$score~genredata$Genre)
```

```
kruskal-wallis rank sum test
```

```
data: genredata$score by genredata$Genre
```

```
Kruskal-wallis chi-squared = 771.62, df = 37, p-value < 2.2e-16
```

From Kruskal Wallis Test, p-value (2.2e-16) is below 0.05. Thus, there is significant evidence against H_0 and we reject H_0 in favor of the alternative hypothesis - median score of at least one genre differs from the rest. Therefore, score does depend on genre.

Next, we built a multilinear regression model to understand how each anime genre can positively or negatively affect the score. We used backward elimination to select the most fitted model.

```
> modelgenre <- lm(score ~ Comedy+ Supernatural+ Romance+ Shounen+ Parody+ School+ Magic+ Shoujo+ Drama+ Fantasy+ Kids+ Action+ Music+ Josei+ Harem+ Adventure+ Sci.Fi+ Ecchi+ Seinen+ Game+ Sports+ Demons+ Historical+ Horror+ Mystery+ Psychological+ Vampire+ Mecha+ Military+ Space+ Samurai+ Thriller+Hentai+ Yaoi+ Police+ Cars+ Dementia+ Yuri)
> step(modelgenre, direction="backward")
```

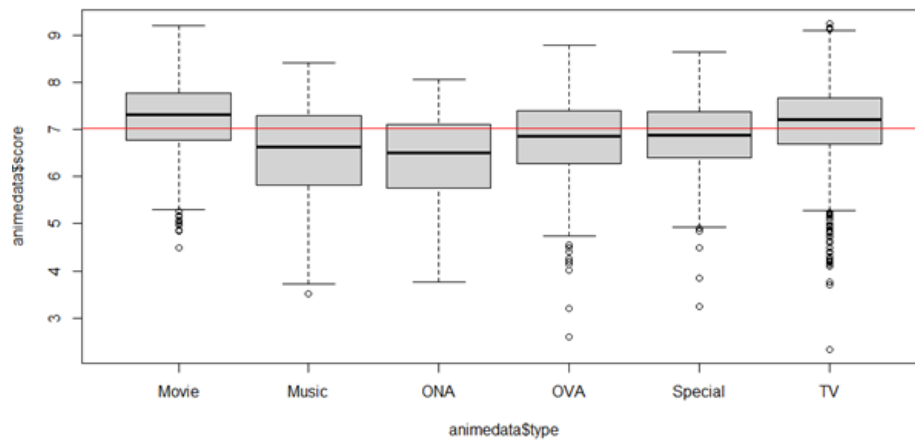
```
Call:
lm(formula = score ~ Comedy + Supernatural + Romance + Shounen + School + Magic + Shoujo + Drama + Fantasy + Kids + Action + Music + Josei + Adventure + Ecchi + Seinen + Sports + Historical + Horror + Mystery + Psychological + Vampire + Military + Space + Samurai + Thriller + Hentai + Police + Yuri)
```

```
Coefficients:
(Intercept)      ComedyTrue  SupernaturalTrue    RomanceTrue    ShounenTrue
      6.48145         0.16972         0.29127         0.15157         0.31739
SchoolTrue      MagicTrue    ShoujoTrue      DramaTrue      FantasyTrue
      0.26046         0.05524         0.15655         0.33601         0.14158
KidsTrue        ActionTrue    MusicTrue      JoseiTrue      AdventureTrue
      -0.42697        0.13398         0.08943         0.30913         0.11194
EcchiTrue       SeinenTrue    SportsTrue    HistoricalTrue    HorrorTrue
      -0.25085        0.29832         0.15183         0.15299        -0.34808
MysteryTrue     PsychologicalTrue  VampireTrue    MilitaryTrue    SpaceTrue
      0.30518         0.33180         0.14704         0.34823         0.11395
SamuraiTrue     ThrillerTrue    HentaiTrue    PoliceTrue      YuriTrue
      0.28175         0.37821        -0.53706         0.21843        -0.69287
```

We observe that genres like Thriller, Drama, Military and Psychological have relatively larger positive coefficients of above 0.3, thus they positively affect the score by greater extents. On the other hand, genres like Yuri, Hentai and Kids have coefficients below -0.4, so they affect the score negatively to a greater extent.

4.2 Score VS Type

This section addresses the research question, “Is the score of an anime dependent on its type of release (eg. TV, movie, OVA etc)? If so, which type has the highest score?”



```
> bartlett.test(data$score~data$type)
```

Bartlett test of homogeneity of variances

data: data\$score by data\$type

Bartlett's K-squared = 59.979, df = 5, p-value = 1.227e-11

Looking at boxplot above, there appears to be statistically different distributions based on type of release. Furthermore, the variances of each category appear unequal. Based on the Bartlett test above, a test for homogeneity of variances in k samples, variances between categories are unequal because p-value (1.227e-11) is below 0.05. Thus, we will employ Kruskal Wallis Test to test for equality of medians between categories instead of ANOVA (which assumes equality of variances).

$$H_0 : \mu_{Movie} = \mu_{Music} = \mu_{ONA} = \mu_{OVA} = \mu_{Special} = \mu_{TV}$$

, where μ refers to median score

H_1 : median score of at least one type does not equal the rest

```
> kruskal.test(data$score~data$type)
```

Kruskal-Wallis rank sum test

data: data\$score by data\$type

Kruskal-Wallis chi-squared = 321.81, df = 5, p-value < 2.2e-16

From Kruskal Wallis Test above, p-value (2.2e-16) is below 0.05. Thus, there is significant evidence against H_0 and we reject H_0 in favor of the alternative hypothesis - median score of at least one type of release differs from the rest. To answer the research question, we conclude that score does depend on the type of release. From the previous boxplot, movie releases appear to have higher scores. We will confirm this using a pairwise t-test, while still assuming variances are unequal based on the previous Bartlett test.

$$H_0 : \mu_{Movie} = \mu_i, i = Music, ONA, OVA, Special, Tv$$

$$H_1 : \mu_{Movie} > \mu_i, \text{where } \mu \text{ refers to mean score}$$

```
> pairwise.t.test(data$score,data$type, p.adjust.method = "none",var.equal = FALSE,alternative = "less")
```

Pairwise comparisons using t tests with pooled SD

data: data\$score and data\$type

	Movie	Music	ONA	OVA	Special
Music	1.4e-12	-	-	-	-
ONA	< 2e-16	0.1029	-	-	-
OVA	< 2e-16	0.9978	1.0000	-	-
Special	< 2e-16	0.9998	1.0000	0.9598	-
TV	0.0077	1.0000	1.0000	1.0000	1.0000

Looking at the above table, all p-values corresponding to 'movie' are below 0.05, thus there is sufficient evidence against the null hypothesis. Thus, to answer the second part of the research question, we conclude that movie releases have statistically significant higher scores than other types of releases.

Similarly, we built a multilinear regression to find out the best fitted model for type of release.

```
> modeltype <- lm(score~type)
> step(modeltype, direction="backward")
Start: AIC=-2228.65
score ~ type
```

	Df	Sum of Sq	RSS	AIC
<none>			2877.5	-2228.7
- type	5	245.44	3122.9	-1857.5

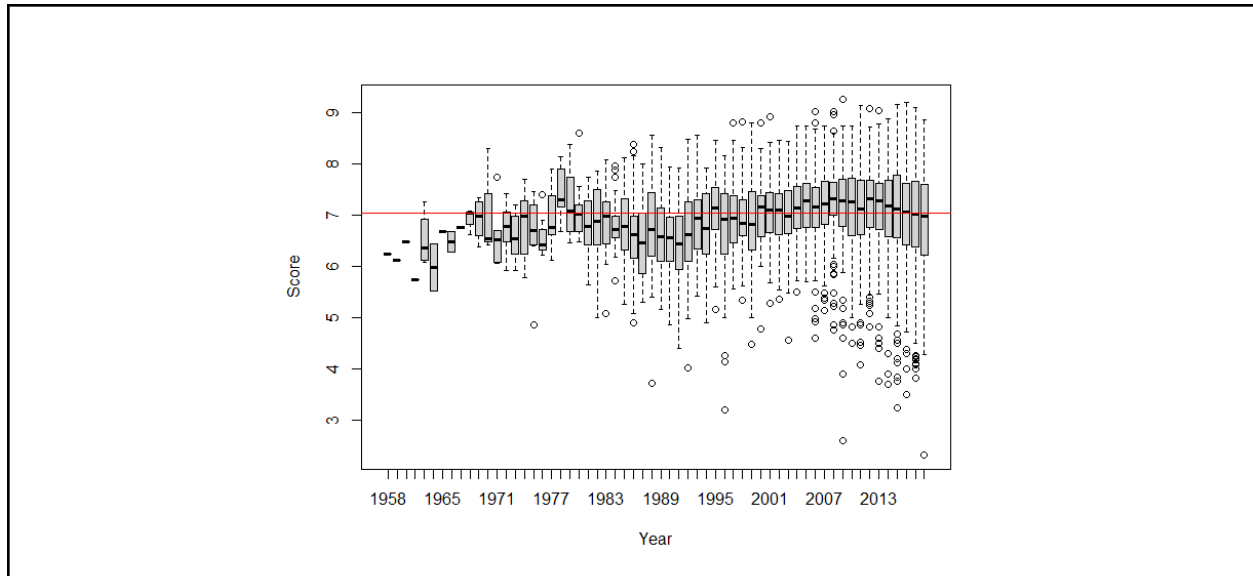
Call:
lm(formula = score ~ type)

Coefficients:
(Intercept) typeMusic typeONA typeOVA typeSpecial typeTV
 7.24492 -0.76474 -0.91666 -0.45517 -0.38337 -0.08173

From the model, we can see that the ONA and Music type of release have the largest negative coefficients of -0.9166 and -0.7647, thus they pull down the score of an anime to a greater extent. On the other hand, TV type of release, as the most mainstream type, has a coefficient that is close to zero, thus it has the least impact on the score.

4.3 Score VS Year

This section deals with the research question - Is the score of an anime dependent on its year of release? If so, in which year did anime have higher scores?



Based on the boxplot, the spread of scores in each category appear to be different, so we will assume that the variances are unequal. This renders the Anova test (assumes equal variances between k samples) unsuitable, so we will employ Kruskal Wallis test - tests the homogeneity of median scores across years.

$H_0 : \mu_{1958} = \mu_{1959} = \dots = \mu_{2018}$, where μ refers to median score
 $H_1 : \text{median score of at least one year is different from other years}$

```
> kruskal.test(data$score~data$aired_from_year)
```

Kruskal-wallis rank sum test

data: data\$score by data\$aired_from_year
Kruskal-wallis chi-squared = 258.22, df = 59, p-value < 2.2e-16

Since p-value (2.2e-16) is below 0.05, there is sufficient evidence against the null hypothesis that median scores are equal between years. Thus, we can conclude that the score depends on the year aired.

Observing the boxplot, 2012 appears to have the higher median score in recent years. We will test this hypothesis using pairwise t-test, with the assumption that the variances between each year are unequal.

$$H_0 : \mu_{2012} = \mu_i$$

, where $i = 2010, 2011, 2013, \dots, 2018$

$$H_1 : \mu_{2012} > \mu_j$$

```
> pairwise.t.test(scores2010onwards,after2010, p.adjust.method = "none",var.equal = FALSE,alternative = 'less')
```

Pairwise comparisons using t tests with pooled SD

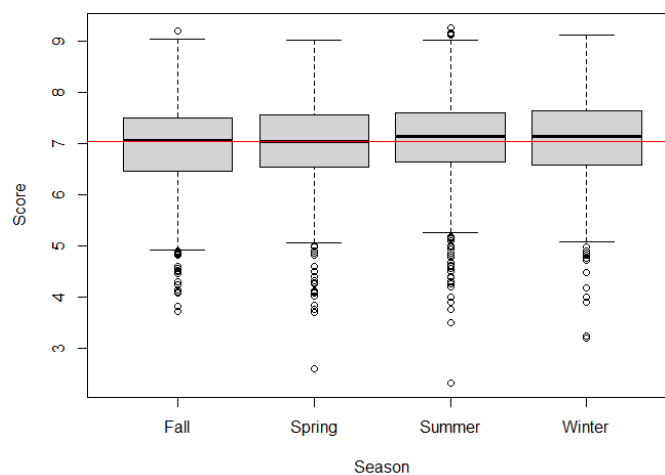
data: scores2010onwards and after2010

	2010	2011	2012	2013	2014	2015	2016	2017
2011	0.29789	-	-	-	-	-	-	-
2012	0.76413	0.89859	-	-	-	-	-	-
2013	0.60054	0.78961	0.31549	-	-	-	-	-
2014	0.25032	0.45227	0.07252	0.16669	-	-	-	-
2015	0.07844	0.19320	0.01277	0.04086	0.21562	-	-	-
2016	0.01681	0.05719	0.00145	0.00651	0.06143	0.22519	-	-
2017	0.00439	0.01747	0.00029	0.00147	0.01787	0.08574	0.25054	-
2018	0.00803	0.01865	0.00191	0.00467	0.02023	0.05330	0.11241	0.20647

For the purpose of presentation, we will narrow down our analysis to the years 2010 onwards. Comparing the year 2012 against 2010, 2011, 2013 and 2014, p-values are 0.60, 0.79, 0.31 and 0.07 respectively, which are all above 0.05. Therefore, there is insufficient evidence against the null hypothesis for these years. Comparing the year 2012 and the years 2015 and onwards, p-values were 0.0408, 0.0065, 0.0014 and 0.0046 respectively - all below 0.05. Thus, there is sufficient evidence against the null hypothesis for these years. In conclusion, to answer our 5th research question, there is statistically significant evidence to show that the mean score of anime was higher in 2012 than the recent years (2015 onwards).

4.4 Score VS Season

This section will continue to answer the research question - Is the score of an anime dependent on its season of air? If so, in which seasons did anime have higher scores?



```
> bartlett.test(data$score~data$prem_season)
```

Bartlett test of homogeneity of variances

```
data: data$score by data$prem_season
Bartlett's K-squared = 1.3429, df = 3, p-value = 0.719
```

From the above boxplot, the median scores in different seasons appear somewhat equal. Based on Bartlett test, a test for homogeneity of variances between k groups, variances can be assumed equal as the p-value (0.719) is above 0.05. Therefore, we will utilize the Anova test to test if different seasons have equal mean scores.

$$H_0: \mu_{Fall} = \mu_{Spring} = \mu_{Summer} = \mu_{Winter}$$

, where μ refers to mean score

H_1 : mean score of at least one season does not equal the rest

```
> anova = aov(data$score~data$prem_season)
> summary(anova)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
data\$prem_season	3	11.2	3.739	5.59	0.000798 ***
Residuals	4652	3111.7	0.669		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Contrary to our previous observations from the boxplot, the Anova test suggests the opposite. The derived p-value (0.000789) is smaller than 0.05, thus there is sufficient evidence against the null hypothesis. To answer our research question, scores are dependent on season aired because the mean score of at least one season differs from the other seasons.

Upon closer inspection, summer and winter appears to have a higher median score based on the boxplot. We will test this theory using pairwise t-test, with the assumption that the variances are equal.

$$H_0: \mu_i = \mu_j$$

, where $i = \text{summer, winter, } j = \text{fall, spring}$

$$H_1: \mu_i > \mu_j$$

```
> pairwise.t.test(data$score,data$prem_season, p.adjust.method = "none",var.equal = T,alternative = "greater")
```

Pairwise comparisons using t tests with pooled SD

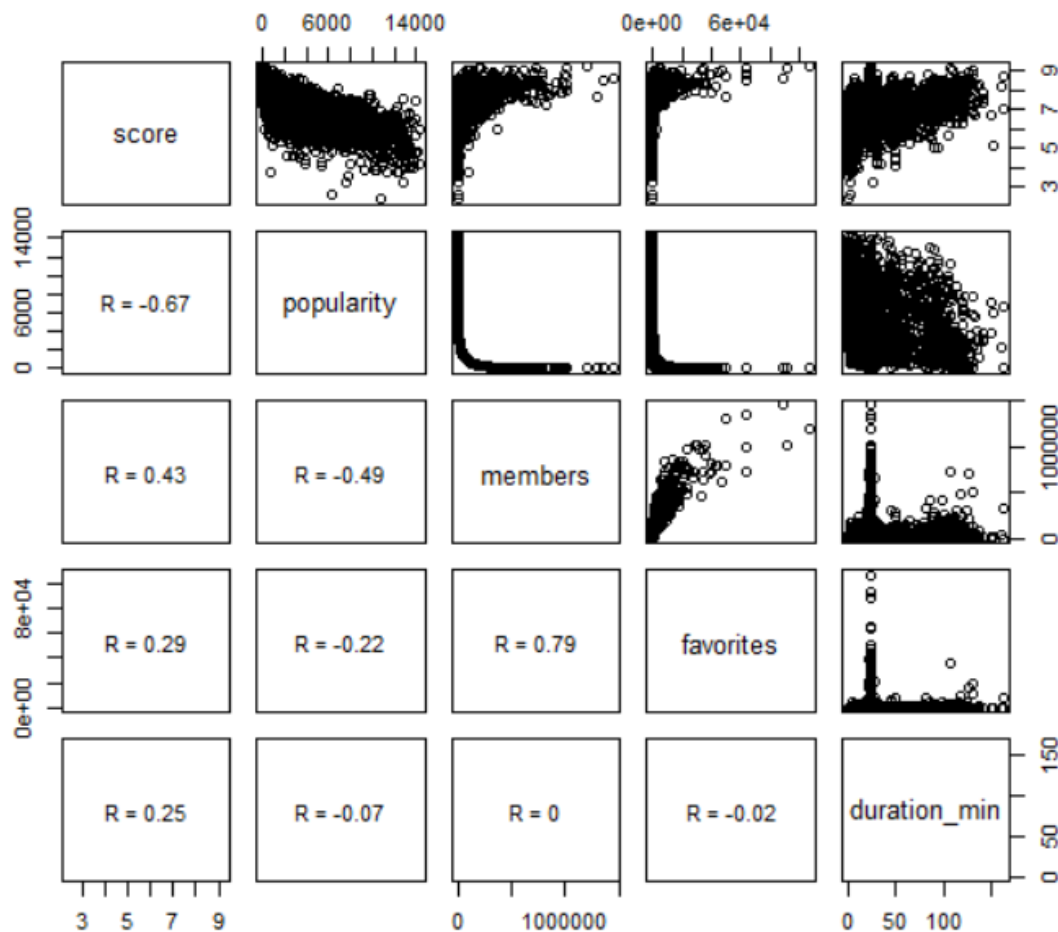
data: data\$score and data\$prem_season

	Fall	Spring	Summer
Spring	0.30114	-	-
Summer	0.00087	0.00518	-
Winter	0.00076	0.00479	0.51133

Comparing p-values of summer against fall and spring, p-values were 0.00087 and 0.00076 ,which are below 0.05. Comparing p-values of winter against fall and spring, p-values were 0.00518 and 0.00479, which are below 0.05. Therefore, there is sufficient evidence against the null hypothesis. In response to the research question, we conclude that summer and winter have statistically significant higher mean scores than fall and spring.

4.5 Score VS Continuous Data

In this section, we will investigate the correlations between anime score and the numerical variables.



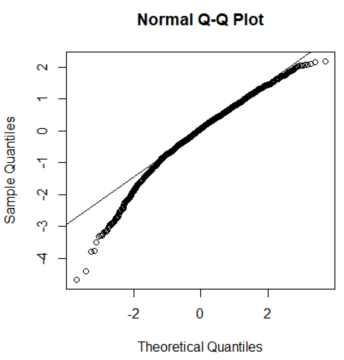
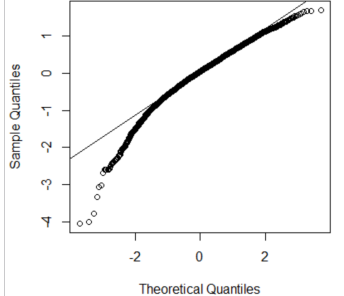
From the plot above, observe that score is most correlated to popularity and members. Between score and popularity there is a Pearson's correlation coefficient of -0.67, implying a relatively strong negative linear correlation. Between score and members the Pearson's correlation coefficient is 0.43, implying a relatively strong positive linear correlation. On the other hand, against favorites and duration, there is a lower Pearson's correlation value and relatively weaker linear correlation. Therefore, we conclude that more popular animes have higher scores.

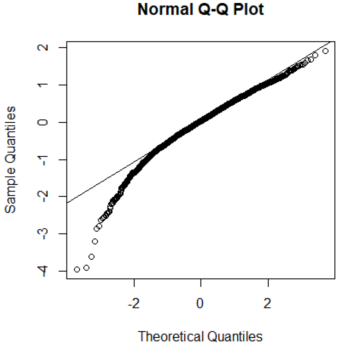
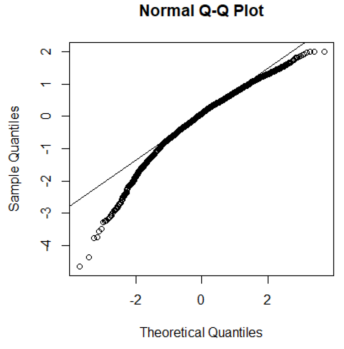
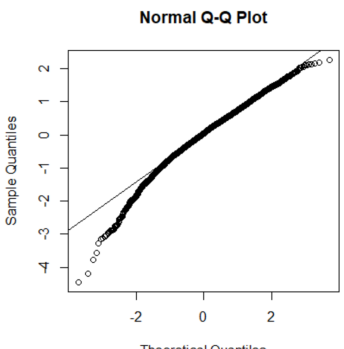
On another note, an interesting observation is that members and favorites have the strongest linear correlation among all pairs. This suggests that anime with more members on MyAnimeList tend to have more favorites.

Next, perform a simple linear regression analysis which of the 4 numerical variables is the most important to model score linearly.

$$Score = \beta_0 + \beta_1 * X + \epsilon$$

Comparing the R-squared values, 'members' variable has the highest value of 0.4587. Therefore we conclude that the number of members an anime has on MyAnimeList is the single most important predictor of how well-liked an anime is.

Variable	Fitted Model	P-value	R^2	QQ plot for residuals
Episodes	$Y = 7.002 + 0.00159X$	$< 2e-16$	0.005569	 <p>A Normal Q-Q Plot for the residuals of the 'Episodes' variable. The x-axis is labeled 'Theoretical Quantiles' and ranges from -2 to 2. The y-axis is labeled 'Sample Quantiles' and ranges from -4 to 2. The data points follow a straight line, indicating that the residuals are approximately normally distributed.</p>
Popularity	$Y = 10.387 - 0.431X$	$< 2e-16$	0.3837	 <p>A Normal Q-Q Plot for the residuals of the 'Popularity' variable. The x-axis is labeled 'Theoretical Quantiles' and ranges from -2 to 2. The y-axis is labeled 'Sample Quantiles' and ranges from -4 to 1. The data points follow a straight line, indicating that the residuals are approximately normally distributed.</p>

Members	$Y = 4.385 + 0.279X$	$< 2e-16$	0.4587	 <p>Normal Q-Q Plot</p> <p>The plot shows Sample Quantiles on the y-axis (ranging from -4 to 2) and Theoretical Quantiles on the x-axis (ranging from -2 to 2). The data points closely follow the diagonal line, indicating a normal distribution.</p>
Favorites	$Y = 6.980 + 5.44e-05X$	$< 2e-16$	0.08281	 <p>Normal Q-Q Plot</p> <p>The plot shows Sample Quantiles on the y-axis (ranging from -4 to 2) and Theoretical Quantiles on the x-axis (ranging from -2 to 2). The data points closely follow the diagonal line, indicating a normal distribution.</p>
Duration in Minutes	$Y = 6.789 + 0.00763X$	$< 2e-16$	0.06061	 <p>Normal Q-Q Plot</p> <p>The plot shows Sample Quantiles on the y-axis (ranging from -4 to 2) and Theoretical Quantiles on the x-axis (ranging from -2 to 2). The data points closely follow the diagonal line, indicating a normal distribution.</p>

With high R-squared value and low p-value, the popularity and members of an anime is shown to be significant in causing the variation in the score of an anime, and they can be explained by the model. While the rest of the features are shown to be significant, its effect cannot be well explained by the simple linear model.

4.6 Multiple Linear Regression

In this section, we attempted to build a multiple linear model for score based on the features that shows significant relationship with the score of an anime, namely popularity, members, and duration in minutes.

The most fitted model is found below, where we see the number of members has contributed most significantly to the score of an anime on the website.


```

> model1 <- lm(score ~ popularity + members + duration_min)
> summary(model1)

Call:
lm(formula = score ~ popularity + members + duration_min)

Residuals:
    Min       1Q   Median       3Q      Max
-3.8054 -0.3257  0.0326  0.3724  2.1629

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.350e+00  2.048e-02  358.85  <2e-16 ***
popularity   -1.513e-04  3.110e-06  -48.66  <2e-16 ***
members       1.062e-06  8.201e-08   12.95  <2e-16 ***
duration_min  6.358e-03  3.213e-04   19.79  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5777 on 4652 degrees of freedom
Multiple R-squared:  0.5029,    Adjusted R-squared:  0.5026
F-statistic: 1569 on 3 and 4652 DF,  p-value: < 2.2e-16

```

With the R-squared value being 0.5777, the model explains about 58% of the variation within the data. The low p-value for all three variables shows that the data favors the hypothesis that there is a non-zero correlation between the variables and the target variable score.

5. Conclusion and Discussion

The anime community has been growing rapidly in recent years. Large entertainment corporations like Netflix are cognizant of this rising trend and are interested to find the best anime for profits. In this report, we hope to answer a number of questions to find out what makes an anime well-liked (measured using 'score' attribute)

Based on the analysis conducted, we conclude that:

- The score of an anime is largely dependent on genres. Particularly, animes with the genre 'thriller' have higher scores. On the contrary, 'hentai' and 'yuri' genres negatively impact scores.
- Type of release does affect score. Particularly, movie releases have statistically significant higher scores while ONA or music releases affect score negatively.
- Score of an anime is also dependent on the year of release. Animes from the year 2012 had higher scores compared to recent years.
- Season of release does affect the anime score. To elaborate, animes released during fall and spring have lower scores.
- For continuous variables, popularity and number of members appear to be the most significant determinant of score. In the multilinear model we built, members contribute the most to the score of the anime.

While these conclusions are meaningful, we acknowledge that there may be more sophisticated underlying relationships in the variables which were not taken into consideration. For instance, a combination of multiple genres, or a particular genre only in a particular year, may result in high scores. However, determining these relationships may be beyond the scope of this course and was simplified in our analysis.

6. Appendix

```
anime = read.csv("AnimeCleanedData.csv",row.names=1)
attach(anime)
```

3.1

```
hist(score)
boxplot(score)
summary(score)
```

3.2.1

```
summary(popularity)
boxplot(popularity)
hist(popularity)
```

3.2.2

```
summary(members)
boxplot(members)
hist(members)
```

3.2.3

```
favorites = as.numeric(favorites)
boxplot(favorites)
hist(favorites, breaks=sqrt(length(favorites)))
summary(favorites)
```

3.2.4

```
boxplot(duration_min)
hist(duration_min, breaks=sqrt(length(duration_min)))
summary(duration_min)
```

3.2.5

```
hist(epsodes)
hist(log(epsodes))
```

3.2.6

```
barplot(table(aired_from_year),
  main = 'Year of airing',
  xlab = 'Year',
  ylab = 'Count')
boxplot(aired_from_year)
```

3.2.7

```
barplot(table(prem_season),
  main="Season of airing",
  xlab="Season",
```

```
ylab="Count")
```

3.2.8

```
Typefreq <- as.data.frame(type)
barplot(Typefreq$Freq,names.arg=Typefreq$Var1, main="Type Frequency")
```

4.1

```
genredata <- read.csv("genreNew.csv", header = TRUE) # genreNew.csv contains score and single column of genre
                                                    from preprocessing
```

```
boxplot(genredata$score~genredata$Genre)
kruskal.test(genredata$score~genredata$Genre)
modelgenre <- lm(score ~ Comedy+ Supernatural+ Romance+ Shounen+ Parody+ School+ Magic+ Shoujo+ Drama+
Fantasy+ Kids+ Action+ Music+ Josei+ Harem+ Adventure+ Sci.Fi+ Ecchi+ Seinen+ Game+ Sports+ Demons+
Historical+ Horror+ Mystery+ Psychological+ Vampire+ Mecha+ Military+ Space+ Samurai+ Thriller+Hentai+ Yaoi+
Police+ Cars+ Dementia+ Yuri)
step(modelgenre, direction="backward")
```

4.2

```
bartlett.test(score~type)
kruskal.test(score~type)
pairwise.t.test(score, type, p.adjust.method = "none", var.equal = FALSE, alternative = "less")
modeltype <- lm(score~type)
step(modeltype, direction="backward")
```

4.3

```
boxplot(score~factor(aired_from_year),
        xlab = 'Year',
        ylab = 'Score')
abline(h=mean(score),col='red')
kruskal.test(score~aired_from_year)
pairwise.t.test(score2010onwards, after2010, p.adjust.method = "none", var.equal = FALSE, alternative = "less")
modelyear <- lm(score~aired_from_year)
summary(modelyear)
```

4.4

```
boxplot(score~factor(prem_season),
        xlab = 'Season',
        ylab = 'Score')
abline(h=mean(score), col='red')
bartlett.test(score~prem_season)
anova = aov(score~prem_season)
summary(anova)
pairwise.t.test(score, prem_season, p.adjust.method = "none", var.equal = FALSE, alternative = "greater")
```

4.5

```
num_var = anime[,c("episodes","score","popularity","members","favorites","duration_min")]
```

```
library("PerformanceAnalytics")
chart.Correlation(animatedata, histogram=TRUE, pch=19)
modelepisodes <- lm(score~episodes)
summary(modelepisodes)
modelpop <- lm(score~popularity)
summary(modelpop)
modelmembers <- lm(score~members)
summary(modelmembers)
modelfav <- lm(score~favorites)
summary(modelfav)
modelmin <- lm(score~duration_min)
summary(modelmin)
model1 <- lm(score ~ log(popularity) + log(members) + duration_min)
summary(model1)
```

4.6

```
model1 <- lm(score ~ popularity + members + duration_min)
step(model1, direction="backward")
summary(model1)
```