Data Analysis Project (Descriptive Analysis)

MCSD 1113-01 Statistic for Data Science
Group members:
Chong Xian Jun (MCS211047)
Li Yexin (MCS211046)

1) Introduction / background

The dataset used is an open dataset obtained from Kaggle (Data source: Steam Store Games (Clean dataset) | Kaggle) based on games on a video games digital distribution market, Steam. As a gaming company, our business spans across video game development, investment, publishing, platforms development and other game-related aspects. It is thus, of our utmost interest to understand the market of video games to assemble better informed business strategies. Many questions are asked to explore the market from various perspectives:

Learning the big market in gaming industry

1. Trend of game releases (quantitative)
2. The revenue involved in the video game market between 2014 and 2018

Understanding video game products

3. The distribution of the number of owners for each video games
4. Median number of owners of each genres
5. Find out the distribution of the number of owners for the video game genre with the highest median number of owners.

2) Data Collection

The dataset is based on games on a video games digital distribution market, Steam. Deploying systematic sampling, the developer scrapped most games released on the store from June 1997 to May 2019 using data gathered from the Steam Store and SteamSpy APIs. The dimension of the dataset is (27075, 18). Table 1 below shows the data schema.

Table 1 dataset schema

No.	Column	dtype	Description	Example of data
1	appid	int	Unique identifier for each title	10
2	name	chr	Title of app (game)	Counter-Strike
3	release_date	date	Release date in format YYYY-MM-DD	2018/11/1
4	english	IDate	Language support: 1 if is in English	1
5	developer	int	Name (or names) of developer(s). Semicolon delimited if multiple	Valve
6	publisher	chr	Name (or names) of publisher(s). Semicolon delimited if multiple	Valve
7	platforms	chr	Semicolon delimited list of supported platforms. At most includes: windows;mac;linux	windows;mac;linux
8	required_age	int	Minimum required age according to PEGI UK standards. Many with 0 are unrated or unsupplied.	0
9	categories	chr	Semicolon delimited list of game categories, e.g. single-player;multi-player	Multi-player;Online Multi-Player;Local Multi-Player;Valve Anti-Cheat enabled

10	genres	chr	Semicolon delimited list of game genres, e.g. action;adventure	Action
11	steamspy_tags	chr	Semicolon delimited list of top steamspy game tags, similar to genres but community voted, e.g. action;adventure	Action;FPS;Multiplayer
12	achievements	int	Number of in-games achievements, if any	0
13	positive_ratings	int	Number of positive ratings, from SteamSpy	124534
14	negative_ratings	int	Number of negative ratings, from SteamSpy	3339
15	average_playtime	int	Average user playtime, from SteamSpy	17612
16	median_playtime	int	Median user playtime, from SteamSpy	317
17	owners	chr	Estimated number of owners. Contains lower and upper bound (like 20000-50000). May wish to take mid-point or lower	10000000-20000000
18	price	num	Current full price of title in GBP, (pounds sterling)	7.19

3) Data Analysis

Learning the big market in gaming industry

1. Trend of game releases (quantitative)

   Over the course of the last few years, we have intuitively felt the immense growth of the video game industry. Nevertheless, it is important to verify the intuition with empirical data.

   a) Plot graph of the number of game released from 2000-2018 by year

```
steam_trend <- steam[(year(steam$release_date) >= 2000) &
(year(steam$release_date) <= 2018)]
vis_stream_trend <-steam_trend %>%
                    group_by(year(release_date)) %>%
                    summarize(n_game_release =n())

# plot line graph
ggplot(data=vis_steam_trend, aes(x=release_year, y=n_game_release)) +
  geom_line()+
  geom_point()
```

Output:

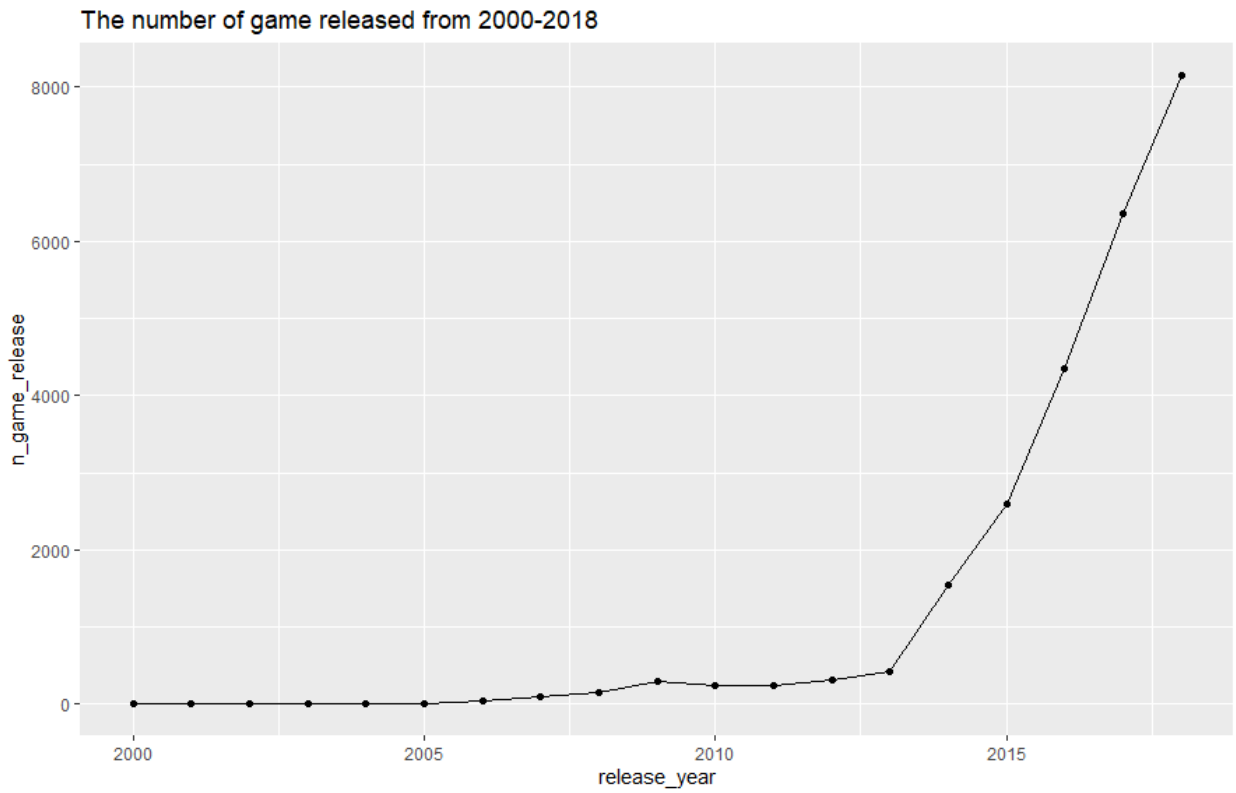


Figure 1 the trend of the number of games released from 2000-2018

Since 2014, the number of games released has come to a significant increase. The growth trend remains steady since and does not show any sign of slowing down even until 2018.

2. <u>The revenue involved in the vido game market between 2014 and 2018</u>

```
library(tidyr)

#separate owners column and change its type to numeric
st <- separate(steam, col=owners, into = c('owners_lower',
'owners_upper'),sep="-")
st$owners_lower <- as.numeric(as.character(st$owners_lower))
st$owners_upper <- as.numeric(as.character(st$owners_upper))

# calculate the midpoints for owners
st$owners_midpt <- ((st$owners_lower + st$owners_upper)/2)


# calculate the number of total owners for games released between 14 to 18
```

```
and the average game price
total_owners_14_18 <- sum(st[(year(st$release_date) >= 2014) &
(year(st$release_date) <= 2018)]$owners_midpt)
avg_game_price <- mean(st[(year(st$release_date) >= 2014) &
(year(st$release_date) <= 2018)]$price)

#calculate total revenue
revenue_14_18 <- total_owners_14_18*avg_game_price
revenue_14_18
```

Output:

[1] 15093750877

Since the explosion of the video game industry from 2014, it had accumulated an estimated total revenue of 15,093,750,877 GBP (~15 Billion GBP) at the end of 2018. This shows that the video game industry is extremely large and lucrative.

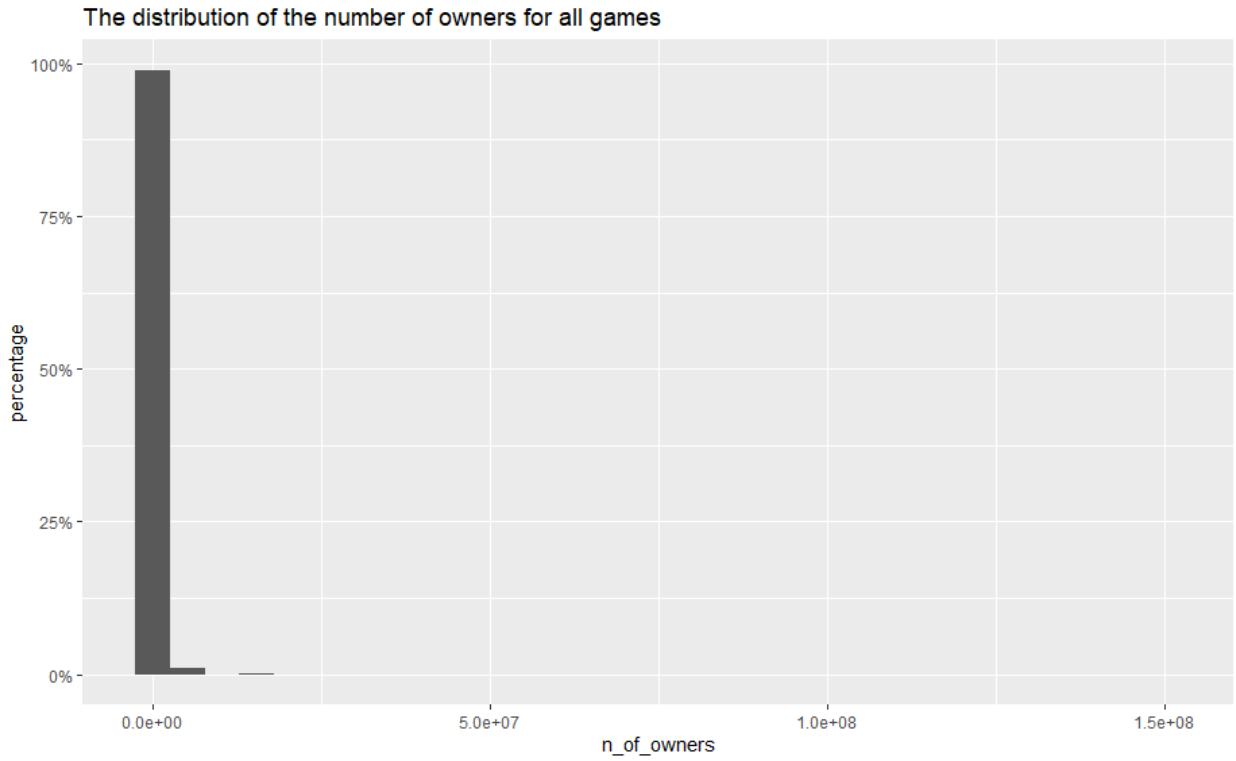
Understanding video game products

3. The distribution of the number of owners for each video games

```
library(scales)

# plot percentage histogram
ggplot(data=st, aes(x=owners_midpt)) +
  geom_histogram(aes(y = (..count..)/sum(..count..))) +
  scale_y_continuous(labels=percent) +
  labs(title = 'The distribution of the number of owners for all games',
x='n_of_owners', y='percentage')
```

Output:



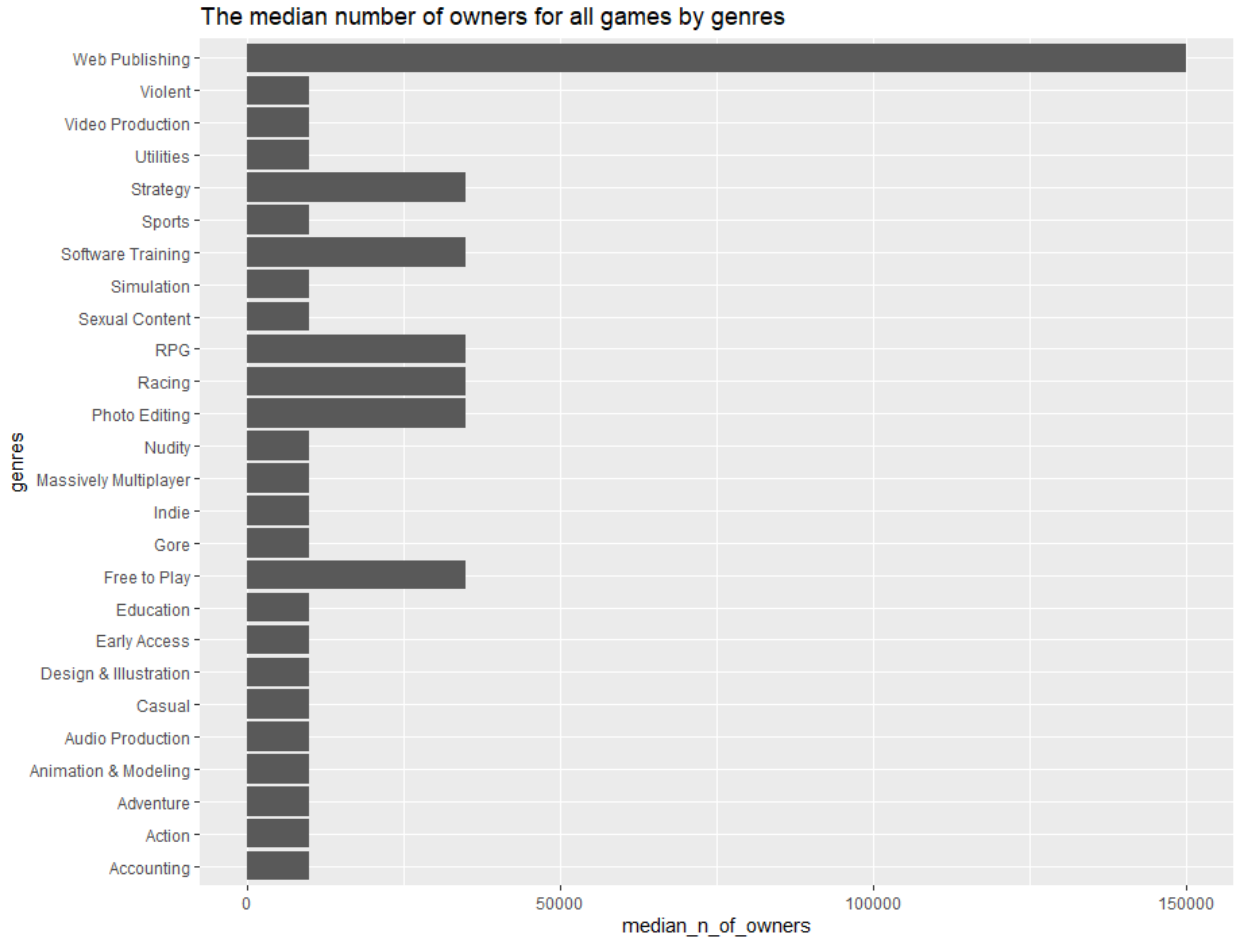
From the histogram, we can see that nearly 100% of the games have a very low number of owners. Despite the large number of games on the platform, only a very small number of them become very popular. While it means that most of the games end up not very profitable, it also means that the ones that make it eat up a tremendous part of the market.

4. Median number of owners of each genres

As the original genres column consist of too many labels, only the main genre (the first genre) of each game is extracted to be analyzed.

```
# visualize the median by genres
ggplot(data=genre, aes(y= genre1, x = median_of_owners)) +
  geom_bar(stat = 'identity') +
  labs(title = 'The median number of owners for all games by genres',
x='median_n_of_owners', y='genres')
```

Output:



We can see that in general, most of the genres have very similar median number of owners other than a few, i.e. web publishing, strategy, software training, RPG, racing, photo editing, free to play. Out of all these genres, only strategy, racing and RPG are purely gaming genres.

a. Is there any evidence that the differences in genres are significant? (inference statistics)

```
#Chi-square goodness of fit test to proof the differences are significant
expected_prob <-  rep( 1/(nrow(genre)), times = nrow(genre))
expected_prob

res <- chisq.test(genre$median_of_owners, p = expected_prob)
res
```

Output:

```
	Chi-squared test for given probabilities

data:  genre$median_of_owners
X-squared = 950909, df = 25, p-value < 2.2e-16
```

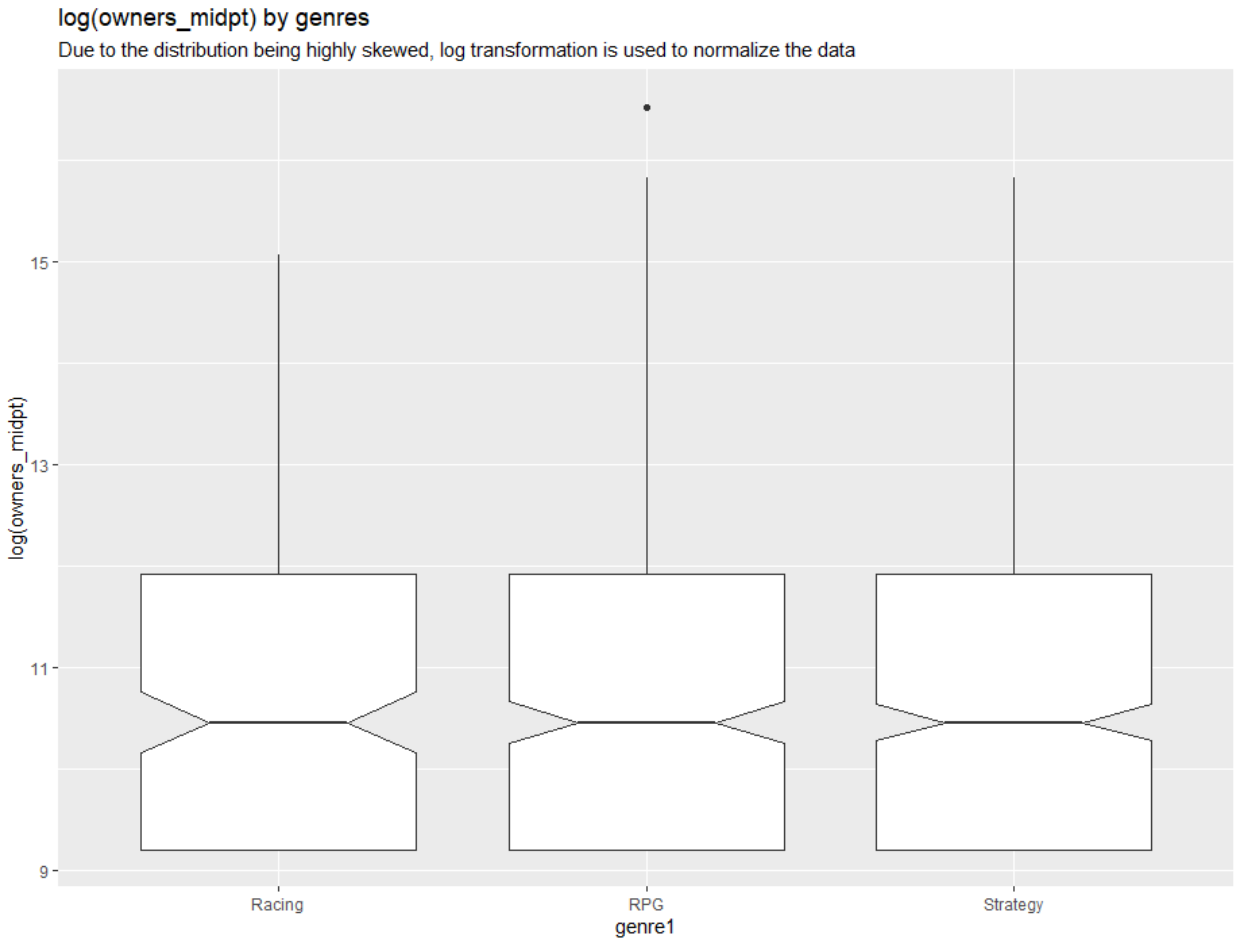
Null hypothesis is that all genres have the same median_n_of_users. As the p-value is 2.2e-16, which is smaller than the critical value of 0.05, the null hypothesis is rejected. The difference between the median number of users in different genres are significant. But it is also important not to base investment decisions on game genres based solely on this significance test as it is affected by many other factors such as the amount of supply in each genre.

5. Find out the distribution of the number of owners for racing, strategy and RPG with boxplot

```
# boxplot for racing, RPG and strategy
genre_top3 <- subset(st, genre1 %in% c('Racing', 'RPG', 'Strategy'))

ggplot(genre_top3, aes(x=genre1, y=log(owners_midpt))) +
  geom_boxplot(notch=TRUE) +
  labs(title='log(owners_midpt) by genres', subtitle = 'Due to the
distribution being highly skewed, log transformation is used to normalize
the data')
```

Output:



a. <u>From the boxplot, we assume that there is no significant difference between the number of owners for the three genres. ANOVA test is carried out to prove the statement.</u>

```
summary(aov(log(owners_midpt) ~ genre1, data = genre_top3))
```

Output:

```
             Df Sum Sq Mean Sq F value Pr(>F)
genre1        2      5   2.525   0.873  0.418
Residuals  1133   3276   2.891
```

F(2, 1133) = 1.375, P(>F) = 0.418, which falls outside of the critical value of $P < 0.05$. Null hypothesis is accepted, there is no significant difference between the three most popular gaming genres.

4) Conclusion

Overall, we know that the gaming industry is a huge market with continually growing trends, especially since 2014. However, it is also a very challenging one as we see that most of the video games on the Steam platform have only 0-20000 users, which is not a really profitable number. However, this also implies that the market is mostly taken up by few very successful games, which make them incredibly lucrative.

In order to find out what makes a game popular, we look to the most obvious variable- genres. As the original genres column consist of too many labels, only the main genre (the first genre) of each game is extracted to be analyzed. While most genres are having a similar median number of owners, there are certain genres that are significantly more popular than the others, i.e. web publishing, strategy, software training, RPG, racing, photo editing, free to play. A Chi-Squared test is carried out to prove that the difference is significant.

As there are also a few non-games software available on Steam, some of the apps provided are not video games. Among the genres mentioned above, only ‘Racing’, ‘RPG’ and ‘Strategy’ are actual gaming genres. Boxplots and ANOVA tests were carried out, and the results showed that there is no significant difference between these three genres. This shows that our company may invest more resources on games that consist of elements in these three genres.