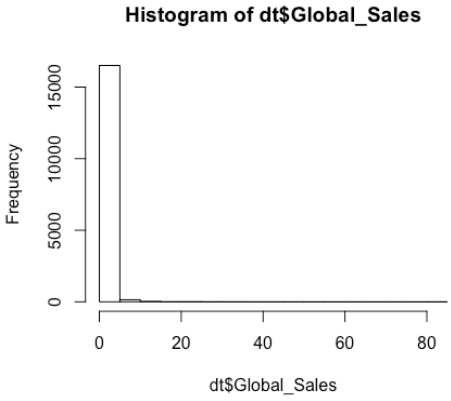
Data Analysis Report

Team Aboleths
Wenyu Ma
Xianru Liu
Qi Di
Yue Ru
Dawei Wang

We've noticed there are a few outliers that could influence the readability of our data charts and figures in our dataset. Meanwhile, more than 90% of our global sales data are under 2 million dollars while the rest of them disperse from 2 to 82.53 million dollars. Therefore, we draw all the figures (Expect for Figure. 1) that under 2 million dollars below in this report.



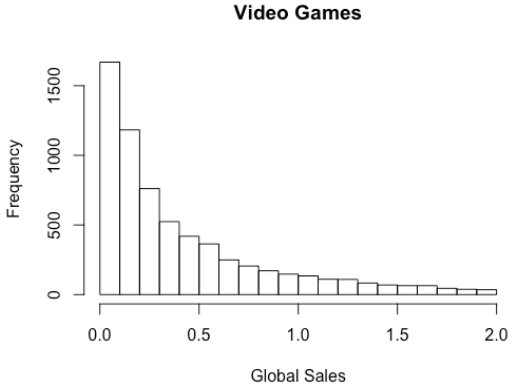


Figure.1 Histogram of original global sales **Figure.2** Histogram of global sales without outliers

In addition, figure.2 shows that it is not symmetric, it is right skewed which means the mean and median will be different. the reason why it is right-skewed is because of the lower boundary of the dataset.

```
> summary(sale$Global_Sales)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0100  0.1000  0.2500  0.4157  0.5800  1.9900
```

```
> summary(sale$Critic_Score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  13.00   61.00   71.00   69.15   79.00   98.00
```

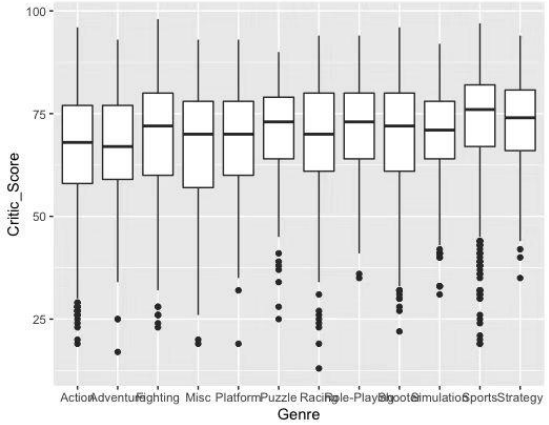
```
> summary(sale$User_Score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.500   6.400   7.500   7.143   8.200   9.600
```

According to the five numbers of Global sales which is our dependent variable, we can see the minimum is 0.01, the maximum is 1.99, the median is 0.25. and the quantile for 25 and, 75% is 0.1 and 0.58 respectively. we also put the summary for the Global sales, Critic score, and user score to compare what is the difference between them.

We can see that the critic score and user score is higher than Global sales. this also means they have a strong relationship with our dependent variable which is Global sales.

Moreover, we looked through the distributions of Critic Scores, User Scores, and Global Sales in different game types. There are 12 different game types in this dataset. In order of showing in this report, from left to right, they are action, adventure, fighting, misc, platform, puzzle, racing, role-playing, shooter, simulation, sports, and strategy. In these box plots, we can see the quartiles, median, outliners of all game types.

The first box plot (Figure.3) shows the distribution of the score from Metacritic staff among all genres. The mean of critic score for all these types is around 70 and the distributions are similar to each other. Several outliers show in the lower part which accommodates the real situation. In the real world, the score from critics will be in a range and there might be a small percentage of poor-quality games that deserve a low score.



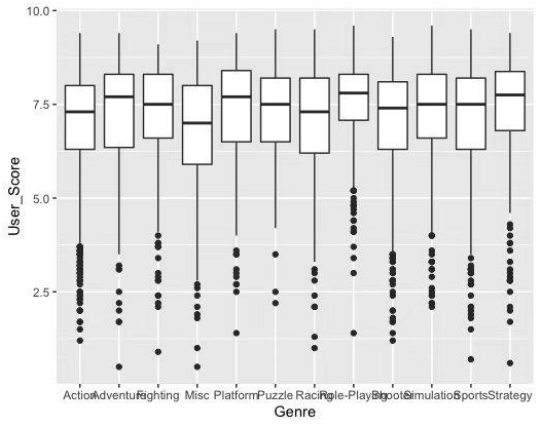


Figure.3 Boxplot of critic score for all genres

Figure.4 Boxplot of user score for all genres

The users' score (Figure.4) is in the scare of 10 which could be matched with critic score by multiple 10. The user shows more lenient than Metacritic staff. Thus, the parameters in the box plot are a bit higher than the first plot. The data are more condensed to the median than the critic score as the boxes from the first plot are generally longer than this plot.

Figure.5 shows the distribution of global sales of different genres. The adventure, puzzle, and strategy games do not have good sales credit among all the game types. The sales of other types are generally closer to each other while the sports game may have the best performance in the market. To make the best representation results, though we display the data under 2 million dollars in

this report, the major sales are under 0.7 billion dollars for most games.

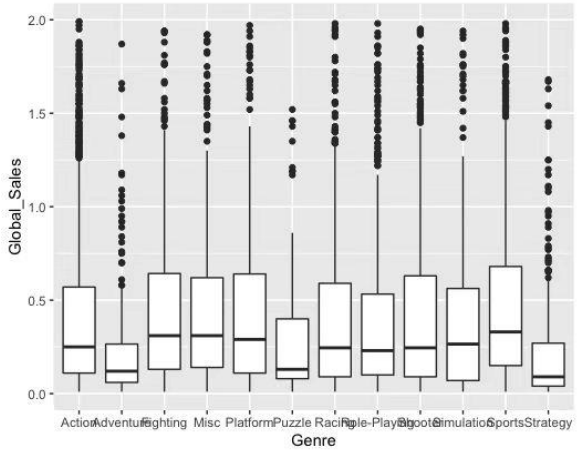


Figure.5 Boxplot of global sales for all genres

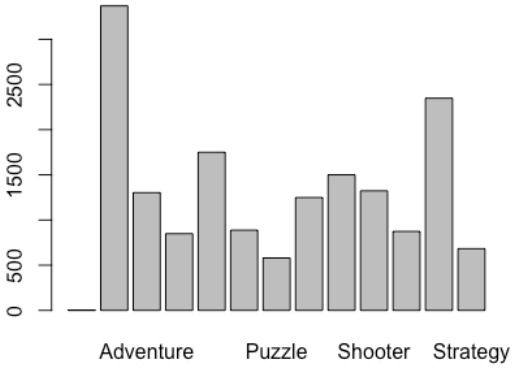


Figure.6 Bar chart of quantity of all genres

```
> table(dt$Genre)

                Action    Adventure     Fighting         Misc     Platform       Puzzle       Racing
             2    3370         1303          849         1750          888          580         1249
Role-Playing     Shooter   Simulation       Sports     Strategy
          1500       1323          874         2348          683
```

There is a total of 12 games genre. In this bar chart, start from the left bar to the right bar which represents Action, Adventure, Fighting, Misc, Platform, Puzzle, Racing, role-Playing, Shooter, Simulation, Sports and Strategy. But there are only showing 4 games genres in the x-axis. We can see different proportions of different games genre in this bar chart. The action game genre is the largest among all the game genres. However, Puzzle is the smallest proportion of those game genres. Thus, based on the observation, it means the action is the most popular game type for people. We only can say that people like this type of game a lot, but we cannot make a conclusion to say that this Action game has the highest critic score or highest global sales. we still need to make more correlation between the game genre and its sales.

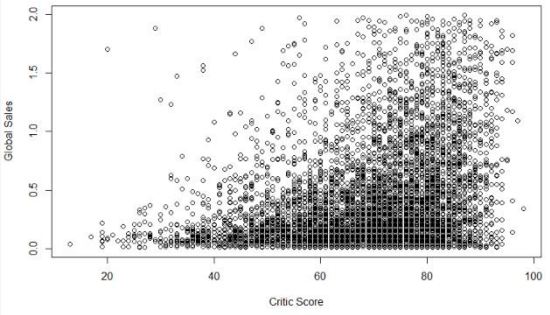


Figure.7 Critic score vs. Global sales

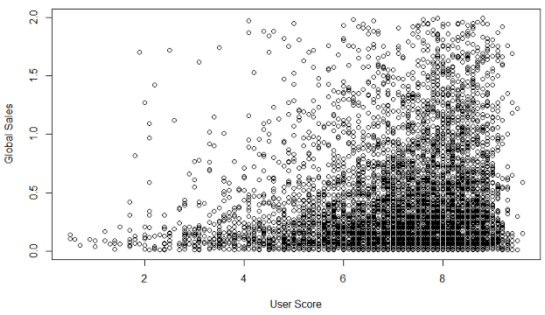


Figure.8 User score vs. Global sales

Scatter plots show how much one variable is influenced by another. According to the Figure.7, the correlation between the Global Sales and Critic Score is positive. We can tell that the higher critic score leads to more global sales. These dots become more and greater along with the critic score larger and larger. In

the Figure.8, the horizontal variable is the user score, which is the score by subscribers. The correlation between User Score and Global Sales is also positive. A higher user score on a video game leads to more global sales. From these two scatter plots, both of the critic score and user score on a video game are the factors that may affect the global sales, if a game with a relatively high critic or user score, it would be easier to become popular all around the world.

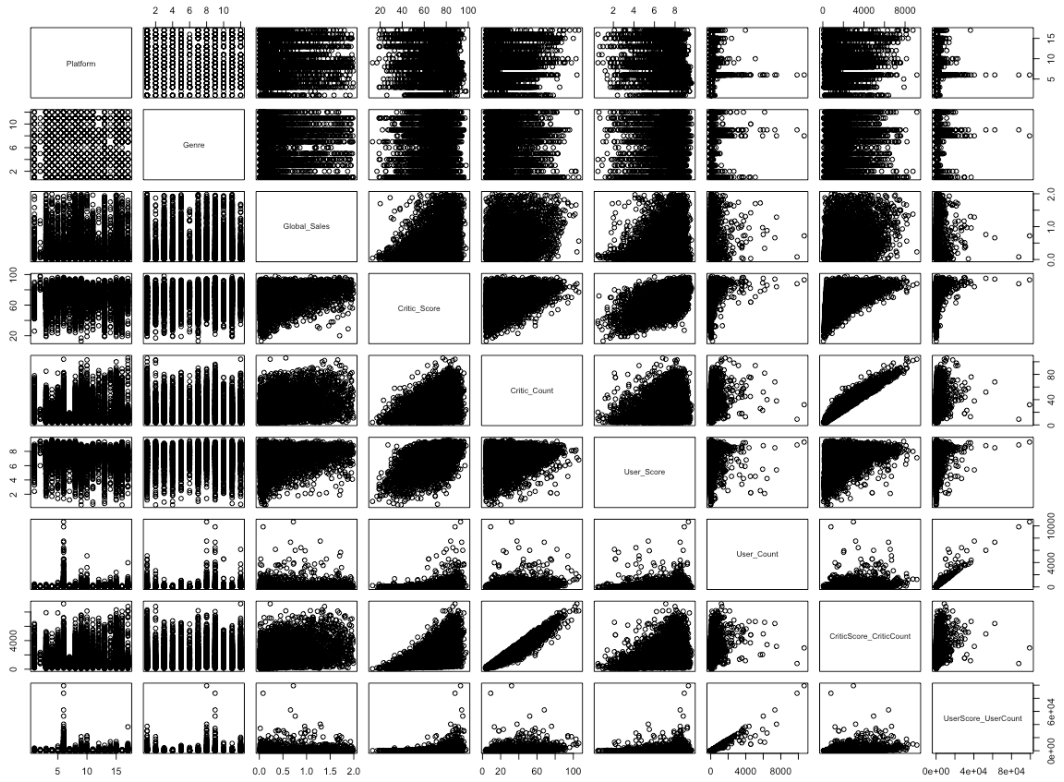


Figure.9 Correlation Matrixes of all variables

In this correlation matrixes, we can see there is a strong positive relationship between global sales and critic score with few outliers. The relationship between global sales and user score seems to be a positive correlation with some outliers. We also see an almost perfect positive correlation between Critic Score/Critic Count and CriticScore_CriticCount, and User Score/User Count and UserScore_UserCount. That's because our two new variables CriticScore_CriticCount and UserScore_UserCount are calculated by Critic Score times Critic Count and User Score times User Count. Based on these observations, we believe the global sales on each video game will be impacted by its critic score and user score. When the critic score and user score are high, the global sales will be high accordingly.