Video Game Sales Analysis

Batch no: 18

Shashwath Kumar - PES2201800623
Keshav Shivkumar - PES2201800168

Objectives

- Analyzing Genre trends by year
- Analyzing Platform trends by year
- Trend of PS platforms over the years
- Trend of XBox platforms over the years
- Competition between Microsoft and SONY
- Sales in North America
- Sales in Eurasia
- Sales in Japan
- Sales in Other countries
- Deducing the factors that affect number of game sales
- Predicting game sales using the different factors in every region

Dataset

- Title : Video Game Sales
- Source : https://www.kaggle.com/ashaheedq/video-games-sales-2019
- Description : Analysis and prediction of Video game sales worldwide
- No. of rows : 55,792
- Variables involved : Name, Genre, ESRB Rating, Platform, Publisher, Developer, Critic Score, Global Sales, NA Sales, PAL Sales, JP Sales, Other Sales, Year

Variable Description

- Name - Name of the game
- Platform - Platform of the game (i.e. PC, PS4, XOne, etc.)
- Genre - Genre of the game
- ESRB Rating - ESRB Rating of the game
- Publisher - Publisher of the game
- Developer - Developer of the game
- Critic Score - Critic score of the game from 10
- Global Sales - Total worldwide sales (in millions)
- NA Sales - Sales in North America (in millions)
- PAL Sales - Sales in Europe (in millions)
- JP Sales - Sales in Japan (in millions)
- Other Sales - Sales in the rest of the world (in millions)
- Year - Year of release of the game

No. of columns: 14

Approach

- Data Preparation : Scraped from the website vgchartz.com. Removed columns which are unnecessary such as img_url, User score, vg chartz score. The year 2020 is also removed from the dataset.
- Data Analysis : Analyzing Trend of Genre by year, Trends of platforms per Year, Trends of Playstation, Xbox and the competition between PlayStation and Microsoft. Pie Chart representation of Sales in different regions of the world.
- Data Redundancy : None present

Literature Survey

Paper by Amar Aziz , Shuhaida Ismail , Muhammad Fakri Othman , Aida Mustapha

Empirical Analysis on Sales of Video Games - 2018

- Source : https://www.researchgate.net/publication/326510277_Empirical_Analysis_on_Sales_of_Video_Games_A_Data_Mining_Approach

Methodology

- Decision Tree is used to predict and to find the correlation between features and as before for pre-processing the dataset.
- Data cleaning is performed in order to remove noise and to correct the inconsistencies existing in the data.
- The next step is merging two different data sets by using data integration technique.
- Classification is performed to analyse the statistical value between pair or two items.

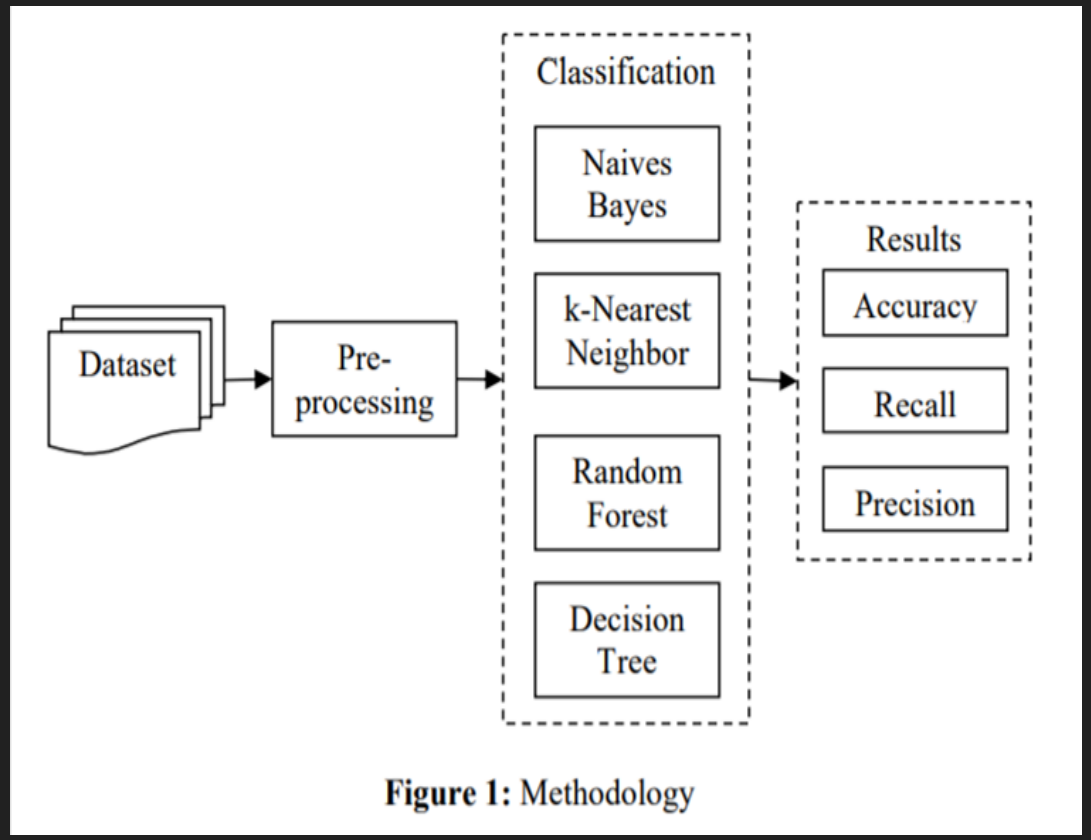


Figure 1: Methodology

Experiments

A. Data Pre-Processing : Data pre-processing technique that has been used in this experiment is normalization. Normalization technique is generally used to rescale the attribute values in the dataset.
B. Data Transformation : cross-validation technique is used to estimate the statistical performance of the learning operator and to estimate the accurate of model performance in training and testing phase.
C. Operators Parameters : suitable operators are chosen in order to produce the result and to fix error occurs during the experiment.
D. Assessment Criteria : Assessments criteria accuracy is calculated by taking the percentage of correct predictions over the total number of examples.

Operators and Parameters

- Naïve Bayes: This technique is used to create a Bayesian model that predicts the value of a target attribute (often called class or label) based on several input attributes of the dataset.
- k-Nearest Neighbor: By comparing a given test example with training examples that are similar
- Random Forest: The Random Forest operator is used to generate a set of random trees.
- Decision Tree: The technique used to create a classification model that predicts the value of a target attribute (often called class or label) based on several input attributes of the dataset.

Table 1: Operators and Parameter

No	Type	Parameter
1	Frequency-based Discretization	Filter type: regular expression Regular expression: Usedprice Number of bin: 2 Range name type: long
2	Decision Tree	Criterion: gain ratio Maximal depth: 20 Confidence: 0.25
3	k-Nearest Neighbor (k-NN)	k: 1 Measure types: Mixed Measures Mixed Measure: Mixed Euclidean Distance
4	Random Forest	Number of Tree: 10 Criterion: gain ratio Maximal depth: 20 Confidence: 0.25
5	Naïve Bayes	Laplace Correction: yes
6	Performance Classification	Main criterion: First Accuracy: Yes Weighted mean recall: Yes Weighted mean precision: Yes

Results

The accuracy, recall and precision of the data were gathered by based on four machine learning algorithms :

- Naïve Bayes
- Decision Tree
- k-NN
- Random Forest.

Techniques	Accuracy	Recall	Precision
Naïve Bayes	81.58%	43.00%	44.12%
Decision Tree	99.55%	86.61%	86.20%
K-NN	24.86%	15.07%	13.97%
Random Forest	26.89%	3.45%	0.93%

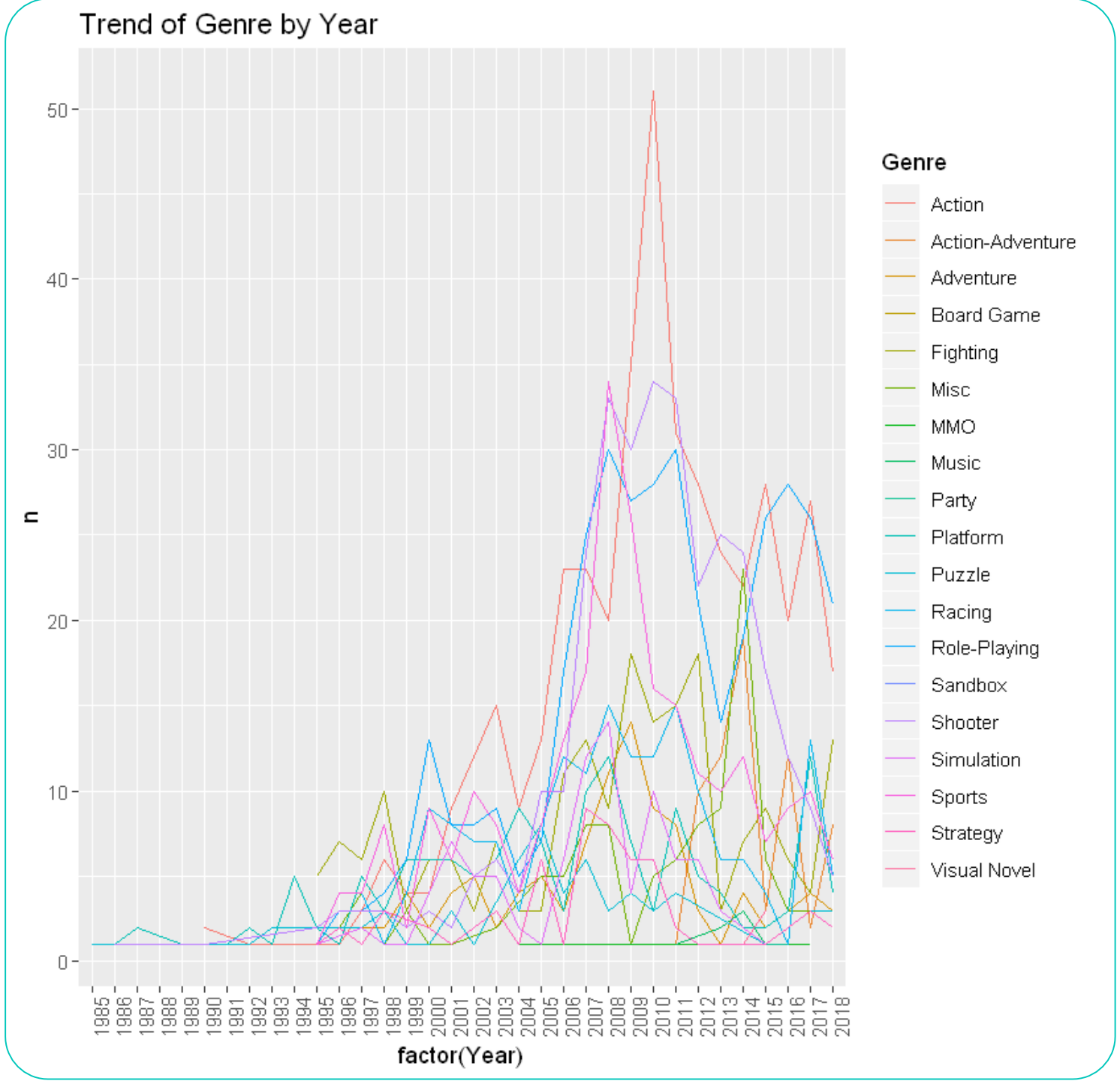
Data Pre-processing

- Removed rank, img url, vgchartz score, last update, basename from columns, user score.
- If Sales have N/A remove the row.
- Reduced to close to 5000 rows.

Data exploration

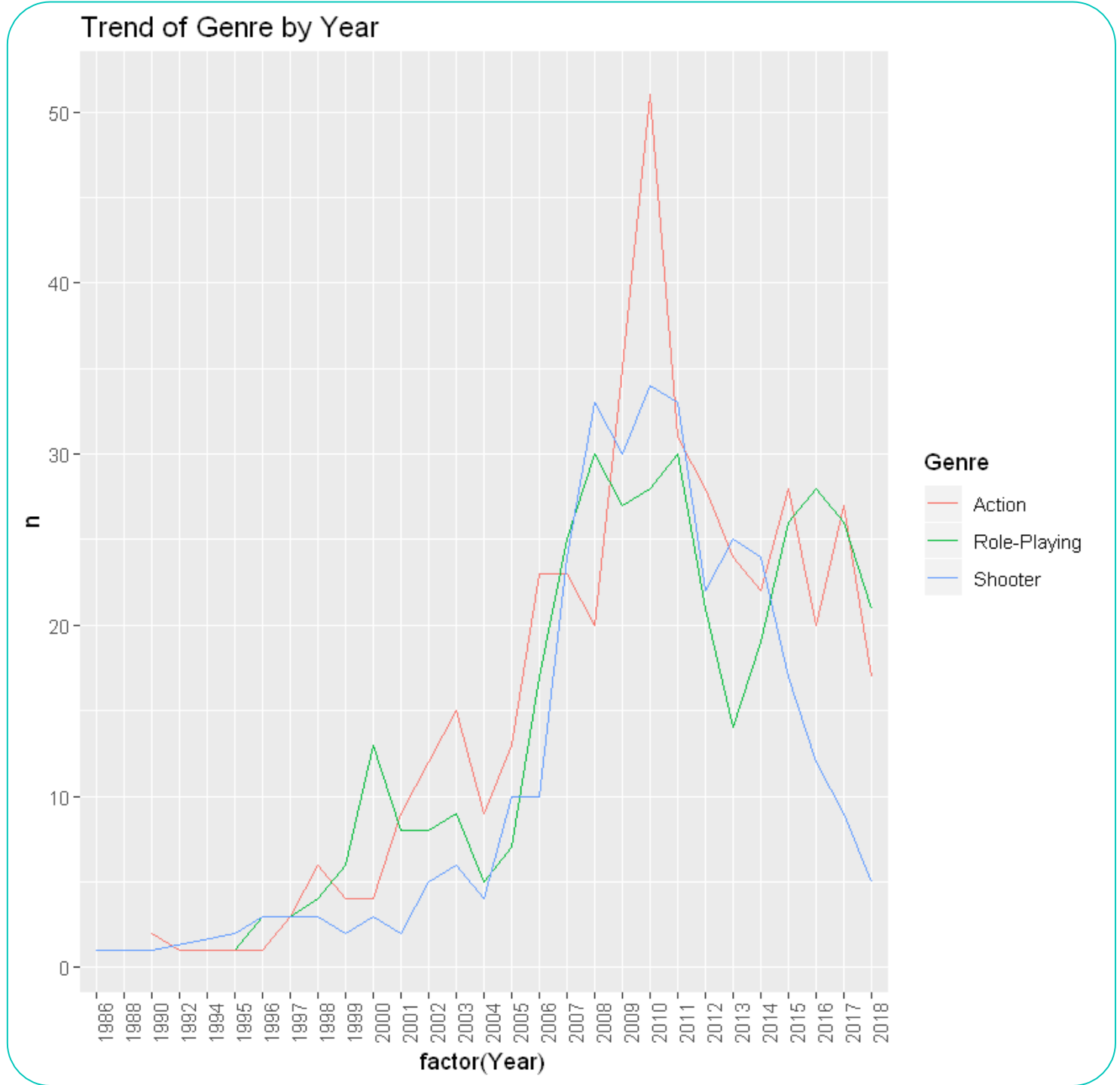
Trends of all Genres by Year

```
ggplot(q1, aes(x = factor(Year), y = n, colour=Genre, group=Genre)) + geom_line()+ggtitle("Trend of Genre by Year")+theme(axis.text.x = element_text(angle = 90,hjust = 1))
```



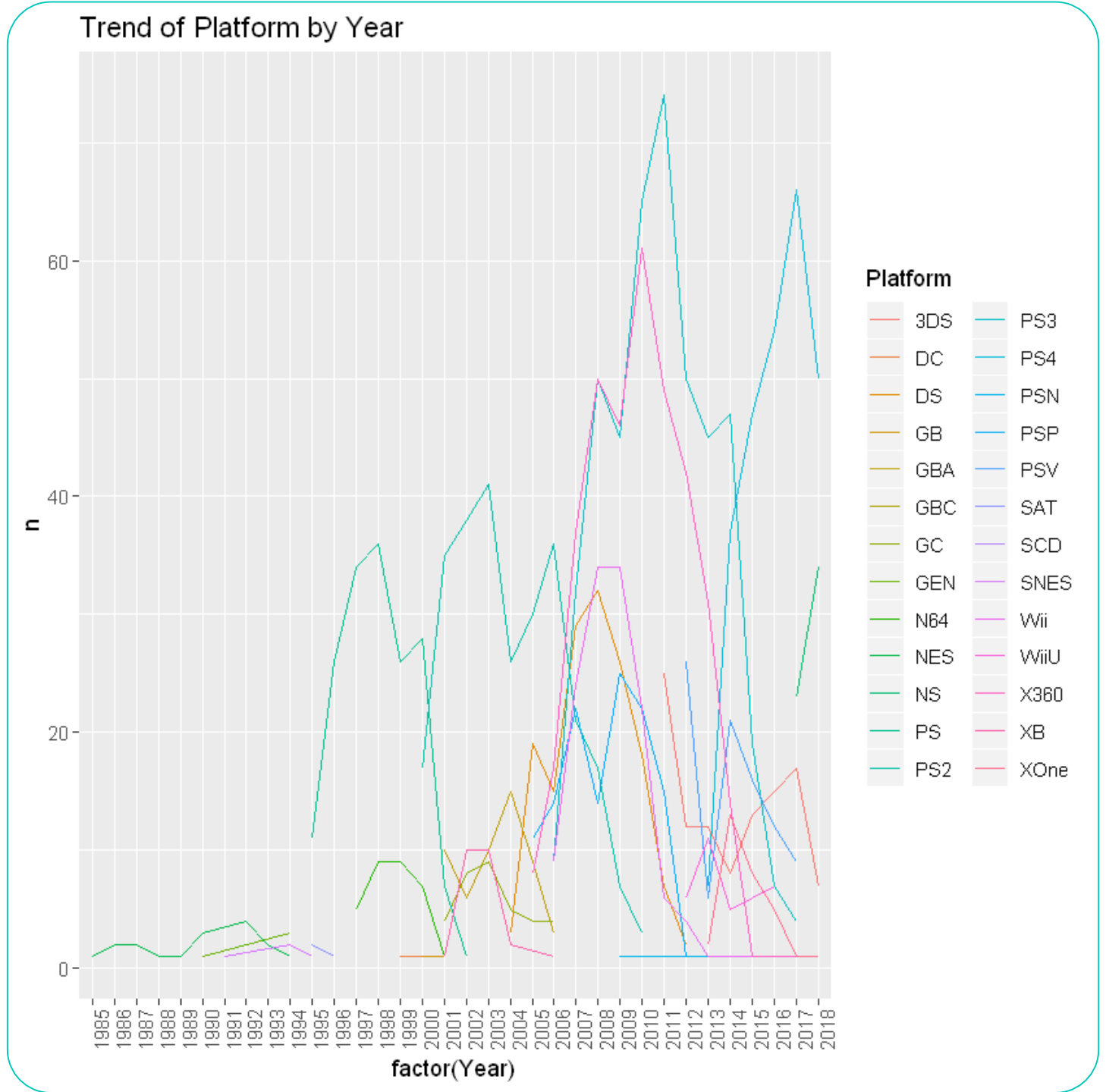
Trends of Top 3 Genres

```
ggplot(q1main, aes(x = factor(Year), y = n, colour=Genre, group=Genre)) + geom_line()+ggtitle("Trend of Genre by Year")+theme(axis.text.x = element_text(angle = 90,hjust = 1))
```



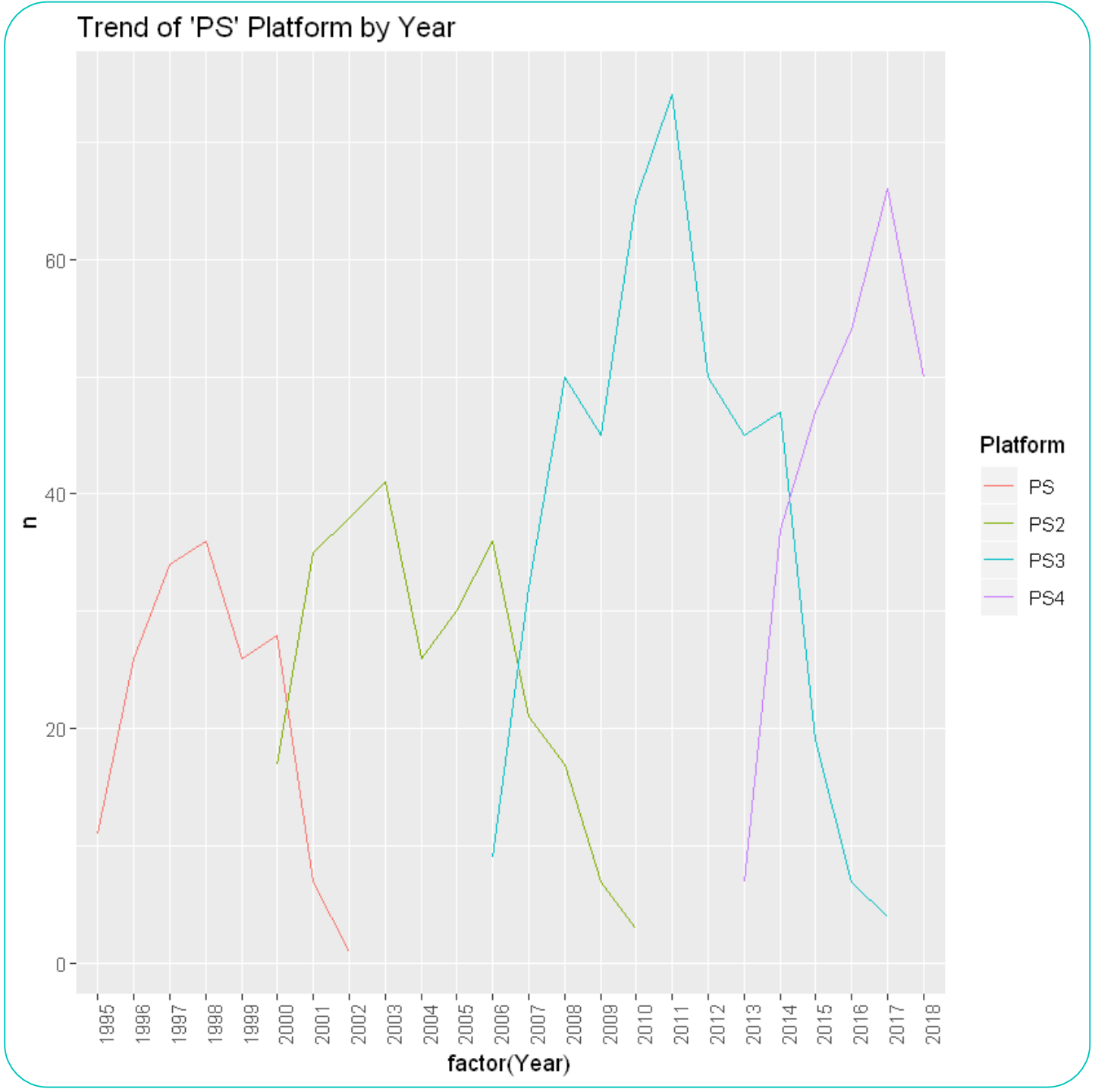
Trend of Platform by Year

```
ggplot(q2, aes(x = factor(Year), y = n, colour=Platform, group=Platform)) + geom_line()+ggtitle("Trend of Platform by Year")+theme(axis.text.x = element_text(angle = 90,hjust = 1))
```



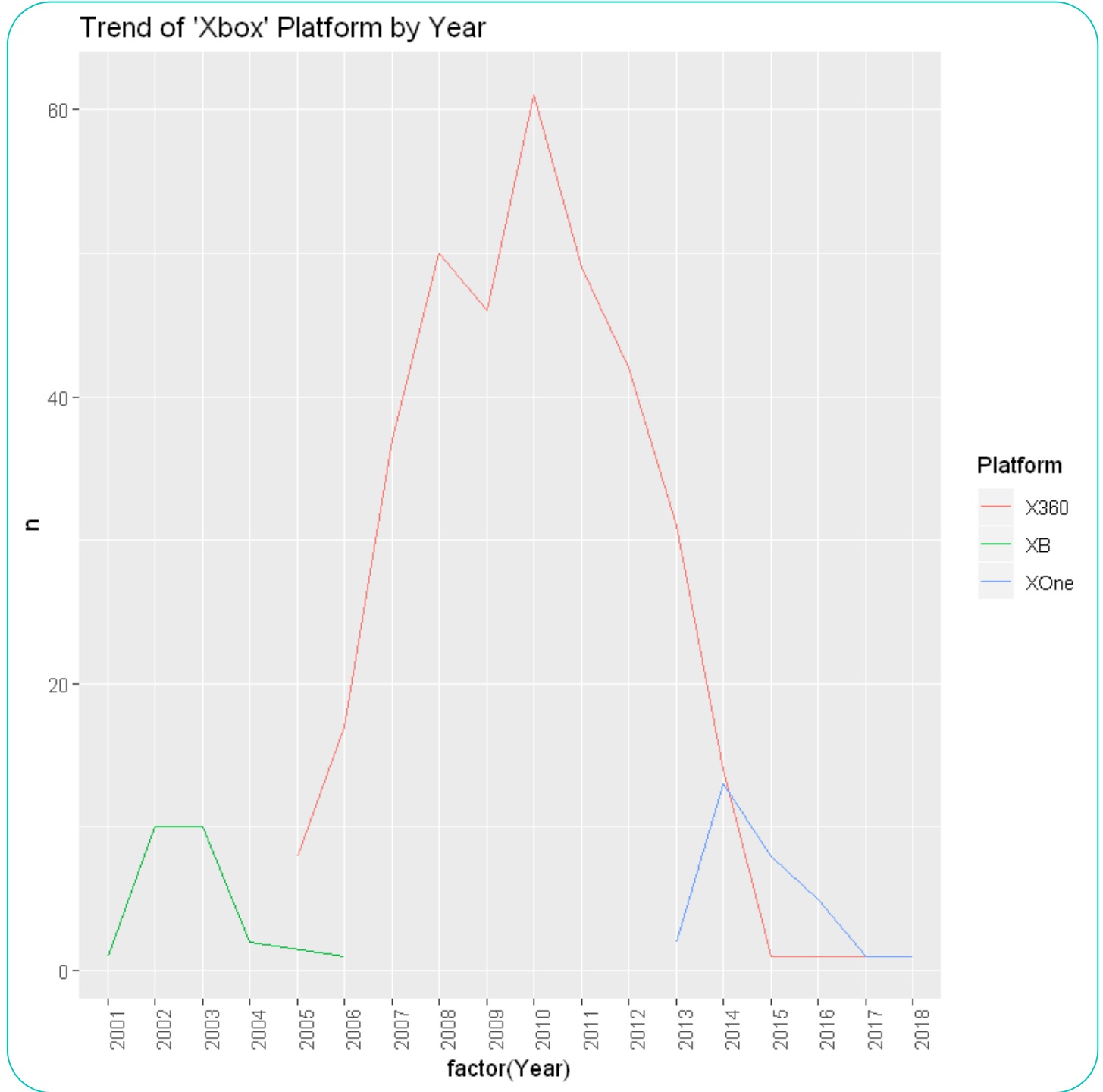
PlayStation platforms

```
q3_ps<-q2[q2$Platform=="PS" |
q2$Platform=="PS2" |
q2$Platform=="PS3"|
q2$Platform=="PS4",]

ggplot(q3_ps, aes(x =
factor(Year), y = n,
colour=Platform, group=Platform))
+ geom_line()+ggtitle("Trend of 'PS'
Platform by
Year")+theme(axis.text.x =
element_text(angle = 90,hjust = 1))
```



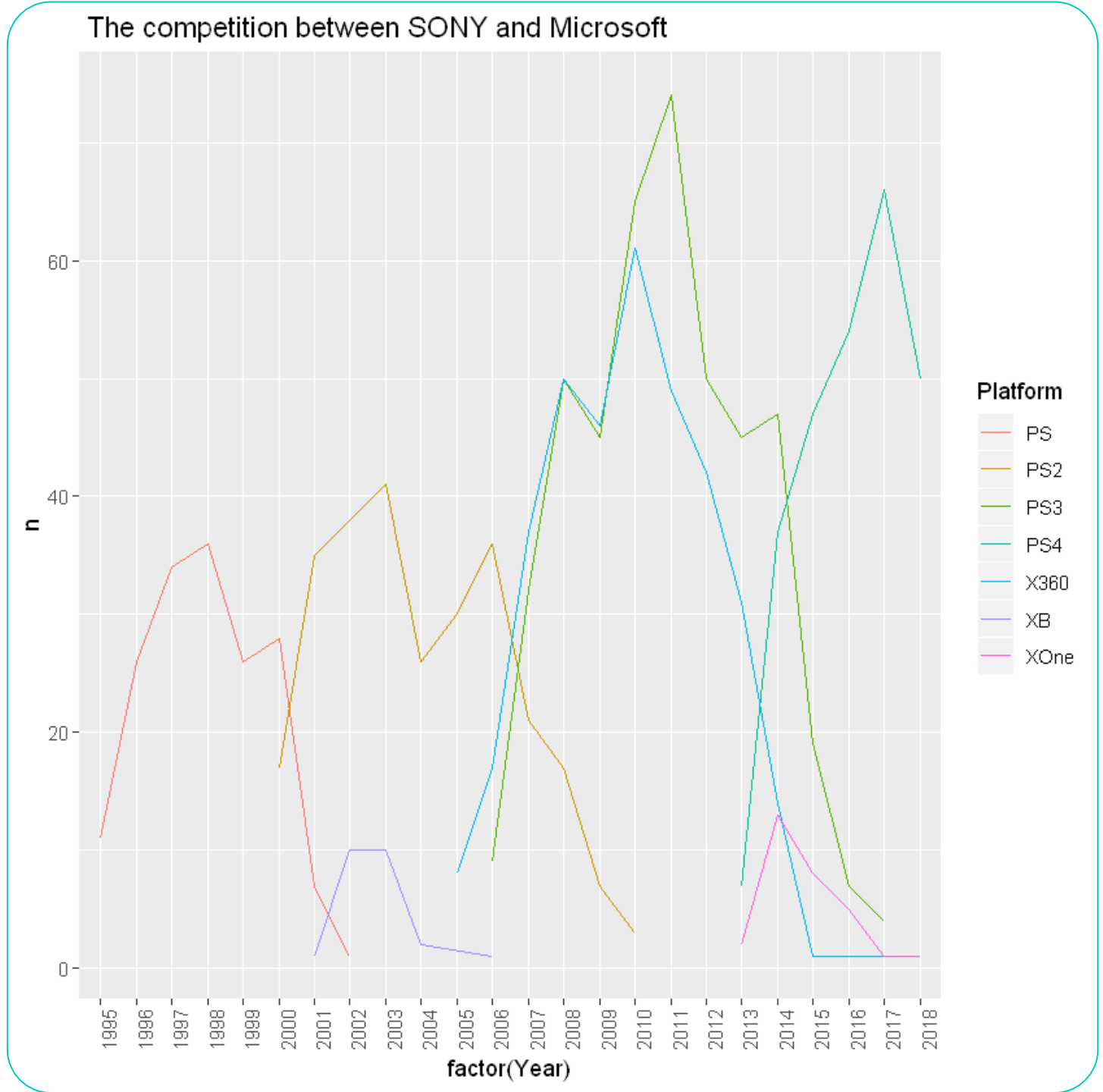
XBOX Platform

```
q4_xb<-q2[q2$Platform=="XB" |
q2$Platform=="X360" |
q2$Platform=="XOne",]

ggplot(q4_xb, aes(x =
factor(Year), y = n,
colour=Platform, group=Platform))
+ geom_line()+ggtitle("Trend of
'Xbox' Platform by
Year")+theme(axis.text.x =
element_text(angle = 90,hjust = 1))
```



PlayStation vs XBox

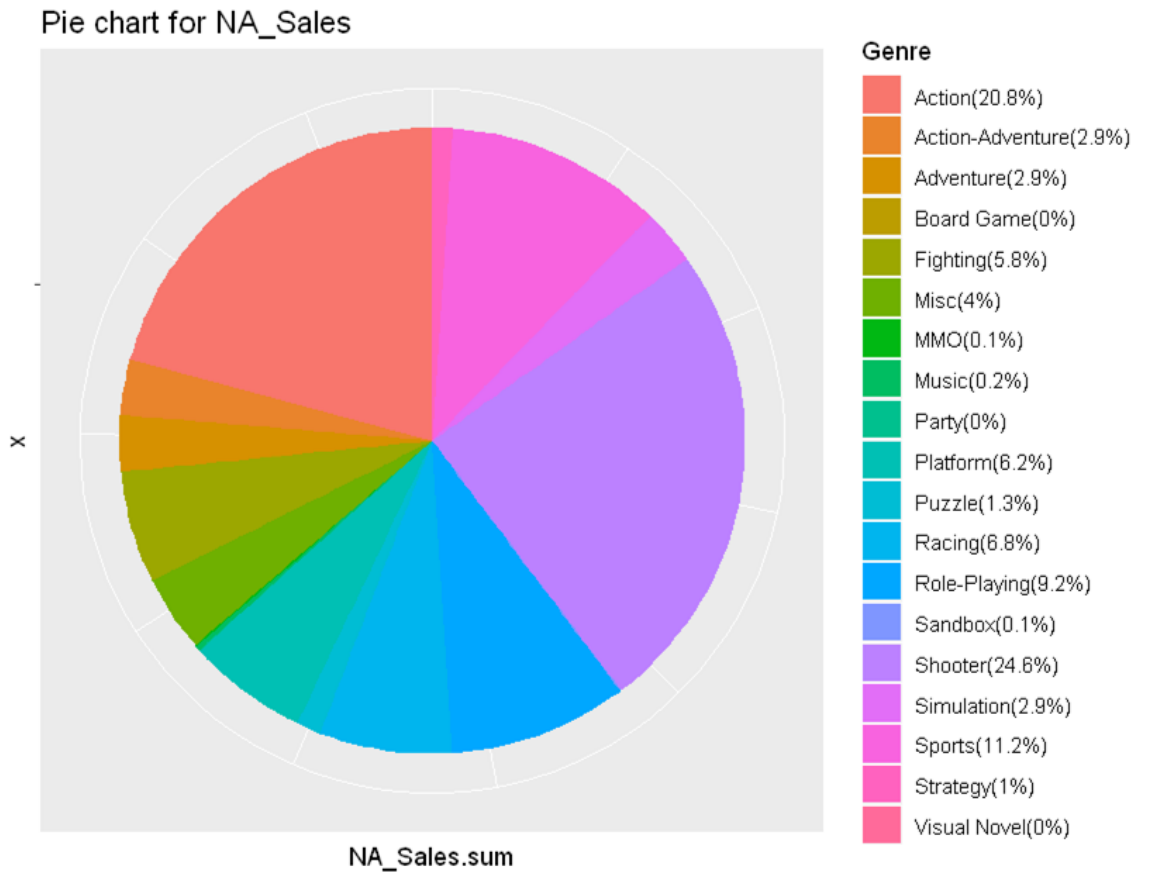
```
q5_mix<-q2[q2$Platform=="XB" |
q2$Platform=="X360"|
q2$Platform=="XOne" |
q2$Platform=="PS"|q2$Platform=="
PS2" | q2$Platform=="PS3" |
q2$Platform=="PS4" ,]

ggplot(q5_mix, aes(x =
factor(Year), y = n,
colour=Platform, group=Platform))
+ geom_line()+ggtitle(" The
competition between SONY and
Microsoft")+theme(axis.text.x =
element_text(angle = 90,hjust = 1))
```



NA Sales by Genre

Action and Shooter are the most popular genres in NA

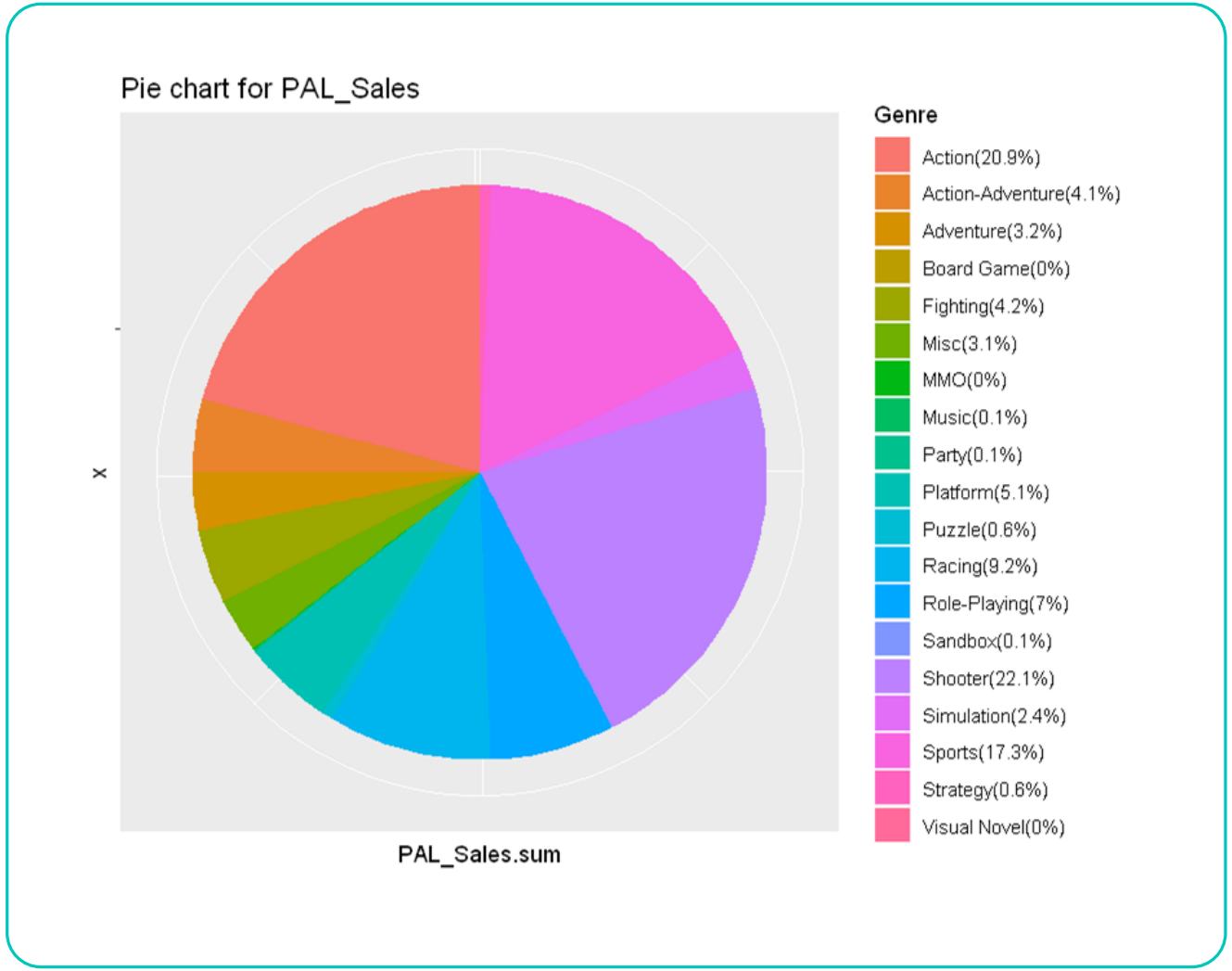
```
ggplot(data = sum_sale, mapping =
aes(x = 'Content', y = NA_Sales.sum, fill =
Genre )) + geom_bar(stat = 'identity',
position = 'stack', width = 1)+
coord_polar(theta ="y") + ggtitle("Pie
chart for NA_Sales")+ theme(axis.text =
element_blank())+
scale_fill_discrete(labels =labelNA)
```



PAL Sales by Genre

Action and Shooters are the most popular genres in Eurasia

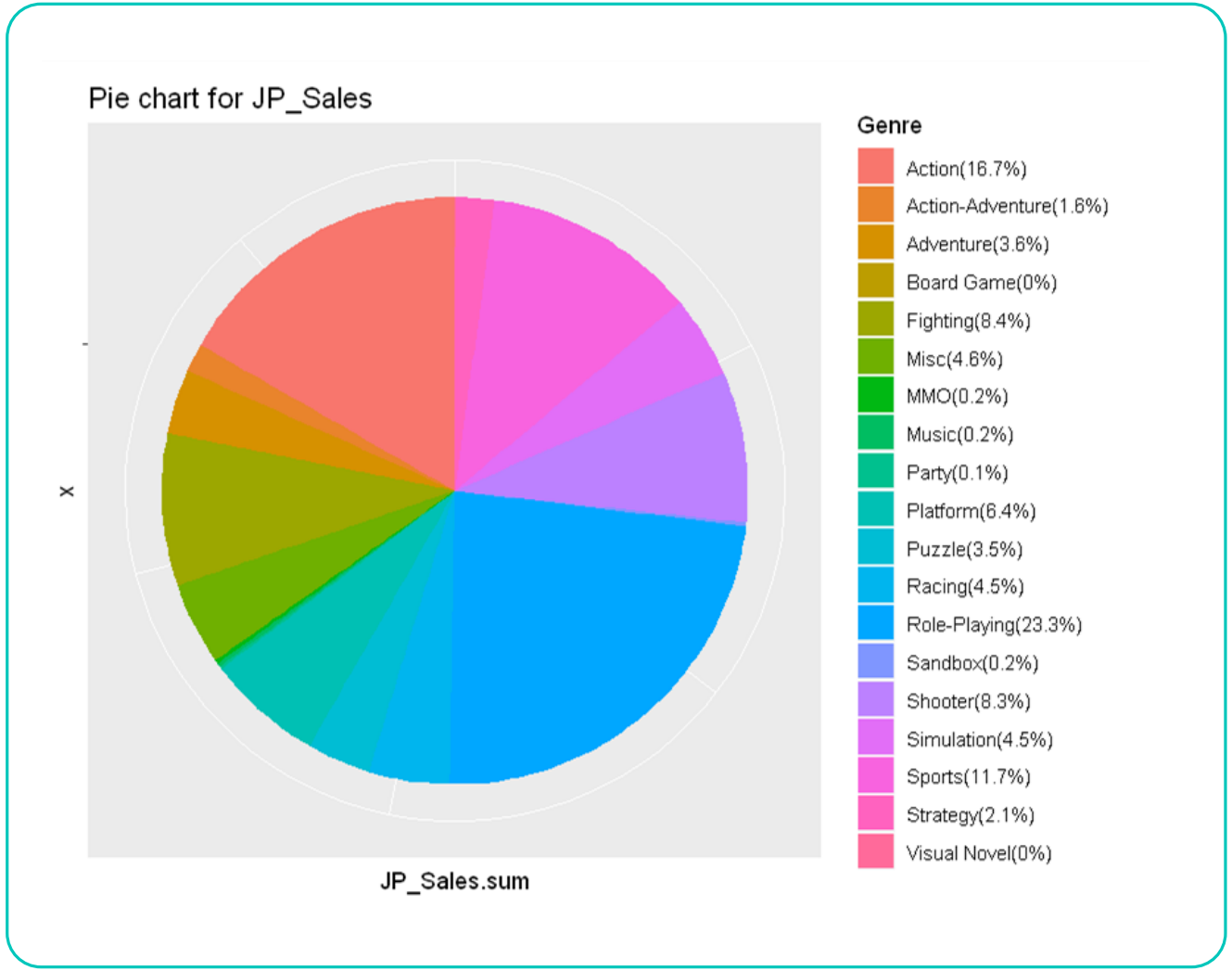
```
ggplot(data = sum_sale, mapping = aes(x = 'Content', y = PAL_Sales.sum, fill = Genre )) +
geom_bar(stat = 'identity', position = 'stack', width = 1)+ coord_polar(theta ="y") +
ggtitle("Pie chart for PAL_Sales")+
theme(axis.text = element_blank())+
scale_fill_discrete(labels =labelPAL)
```



Japan Sales by Genre

Action and Role-Playing games are the most popular

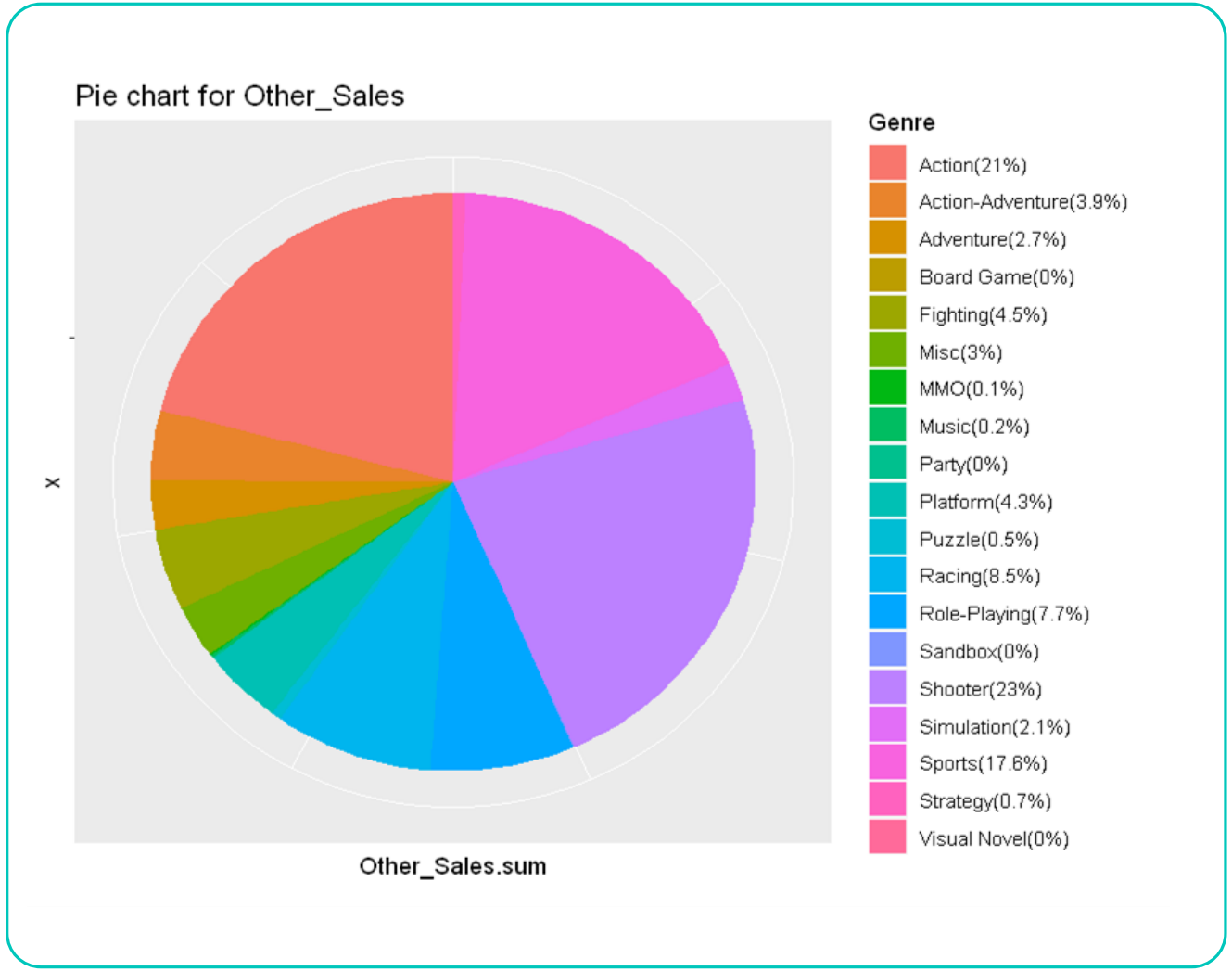
```
ggplot(data = sum_sale, mapping = aes(x = 'Content', y = JP_Sales.sum, fill = Genre )) +
geom_bar(stat = 'identity', position = 'stack', width = 1)+ coord_polar(theta ="y") +
ggtitle("Pie chart for JP_Sales")+
theme(axis.text = element_blank())+
scale_fill_discrete(labels =labelJP)
```



Other Sales

Action, Shooters and Sports games are the most popular

```
ggplot(data = sum_sale, mapping = aes(x = 'Content', y = Other_Sales.sum, fill = Genre )) +
geom_bar(stat = 'identity', position = 'stack', width = 1)+ coord_polar(theta ="y") +
ggtitle("Pie chart for Other_Sales")+
theme(axis.text = element_blank())+
scale_fill_discrete(labels =labelOTHER)
```



Exploratory Analysis

- Action, Role-Playing and Shooter games are the most popular games Worldwide
- PS3 had the most selling games for SONY
- Xbox 360 had the most selling games for Microsoft
- Xbox 360 was the best competitor to the same gen PS platforms
- Action and Shooters are popular among the world. But, in Japan role-playing games are much more popular than action and shooter games.

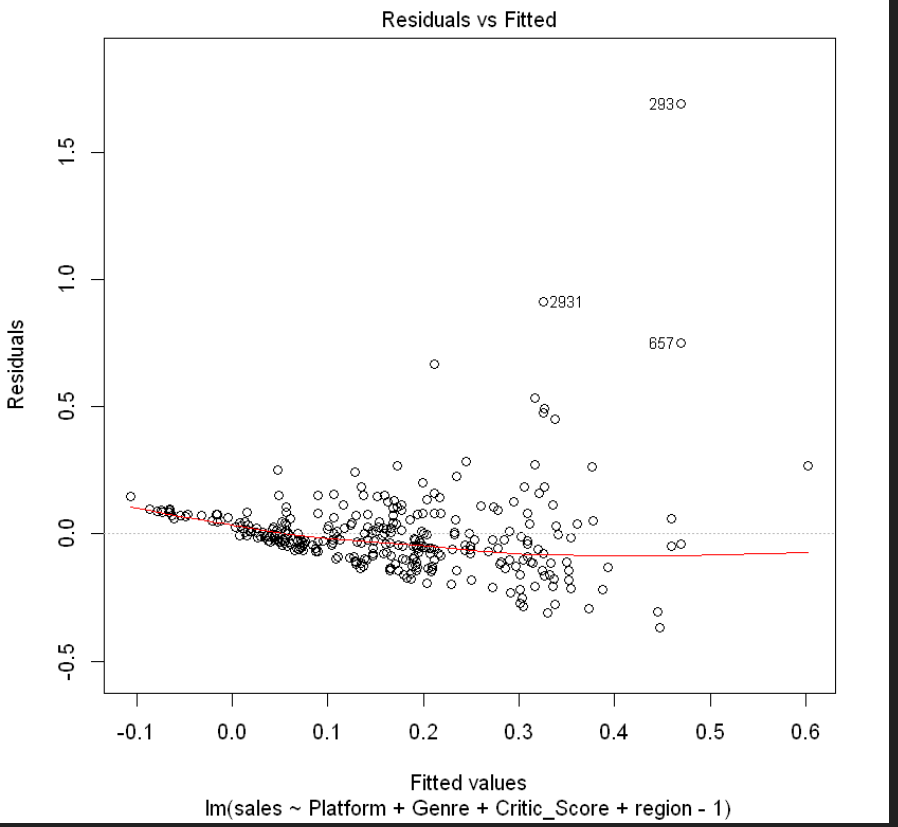
Data Prediction and Regression

Linear Regression model

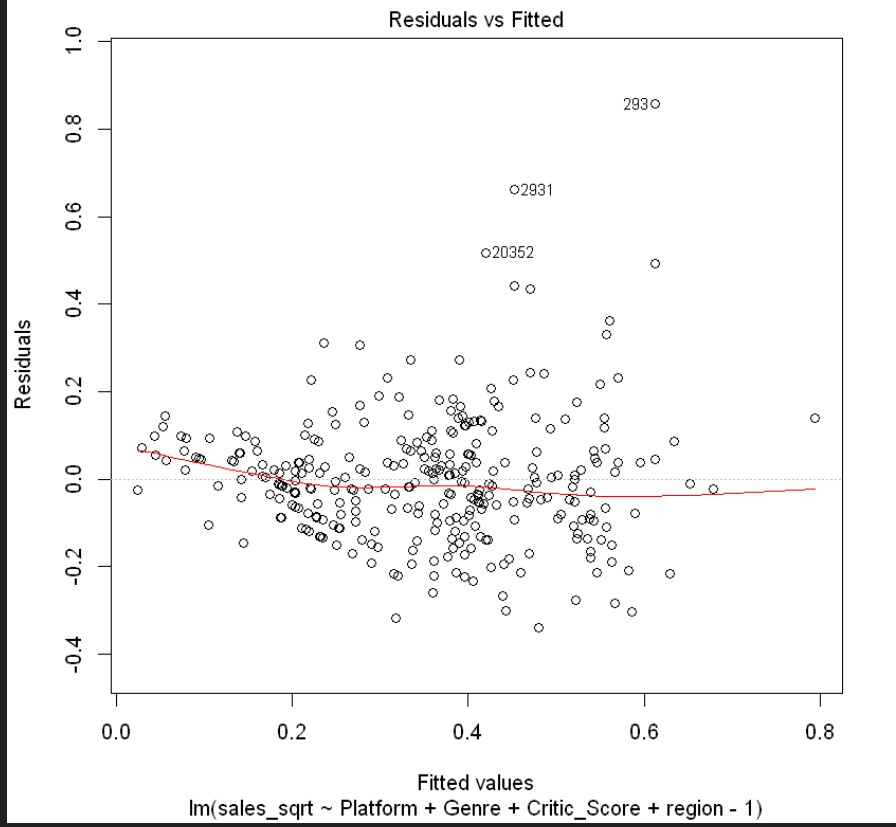
- The majority of video games were not selling well, and their data fits the linear regression model very well.
- The data of best-selling games(Super Mario, Call of Duty, etc.) doesn't fit the model well.
- When sales is big enough, sales variable itself is also an influential factor to regression model.
  Finally, we used square rooted sales as a transform of response Y thus eliminate the error due to butterfly effect.
- The higher R-square presents a better fitting model.

Residuals vs FITTED

A residual is the difference between the observed value of the dependent variable (y) and the predicted value (ŷ).




```
new_video_nintendo_training$sales_sqrt
<-
sqrt(new_video_nintendo_training$sales)
new_video_nintendo_test$sales_sqrt <-
sqrt(new_video_nintendo_test$sales)
```

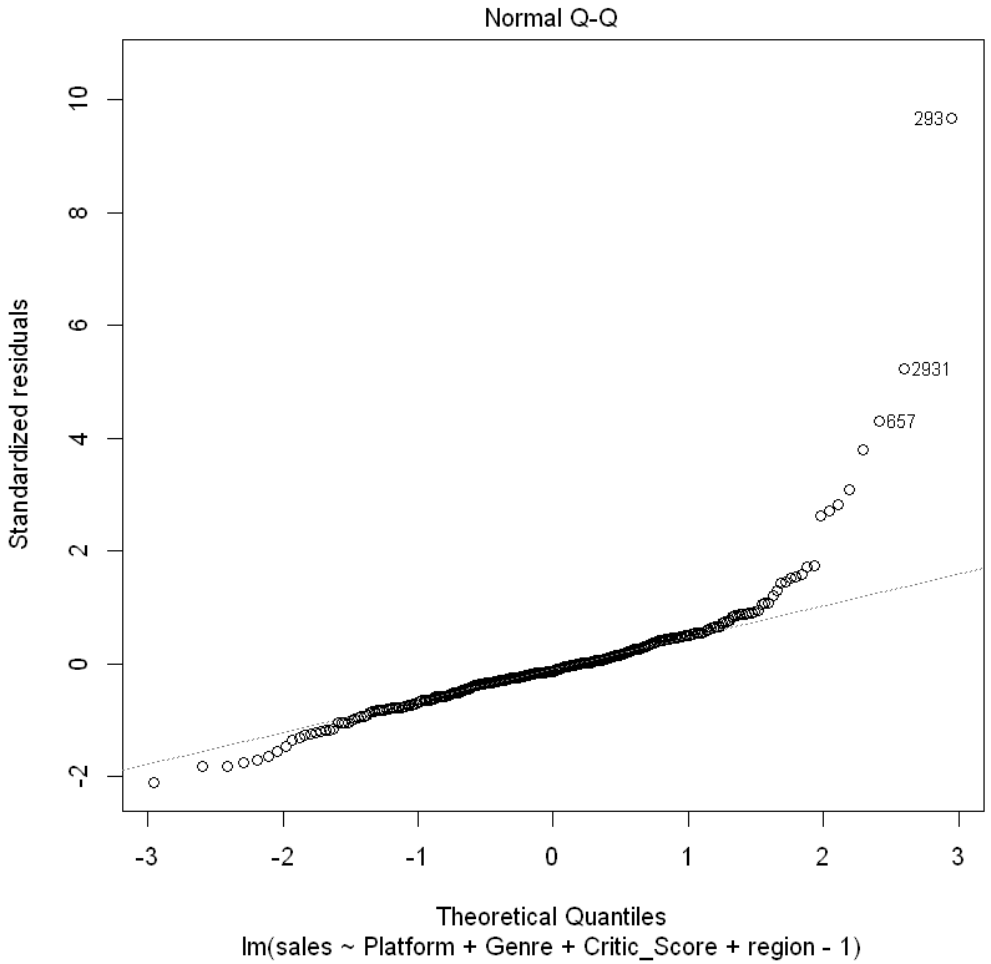


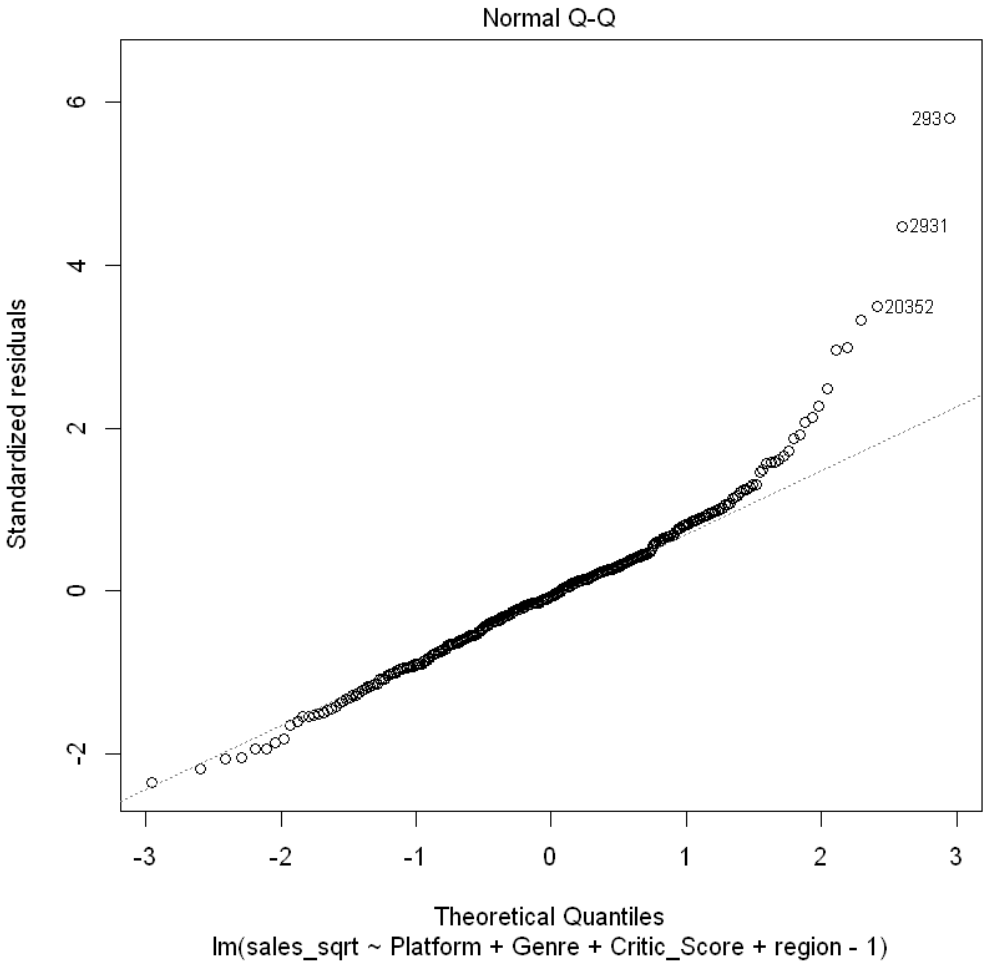

```
lm1 <- lm(sales~Platform+Genre+Critic_Score+region-
1,data=new_video_nintendo_training)
```

```
lm2 <-
lm(sales_sqrt~Platform+Genre+Critic_Score+region-
1,data=new_video_nintendo_training)
```

Q-Q Plot

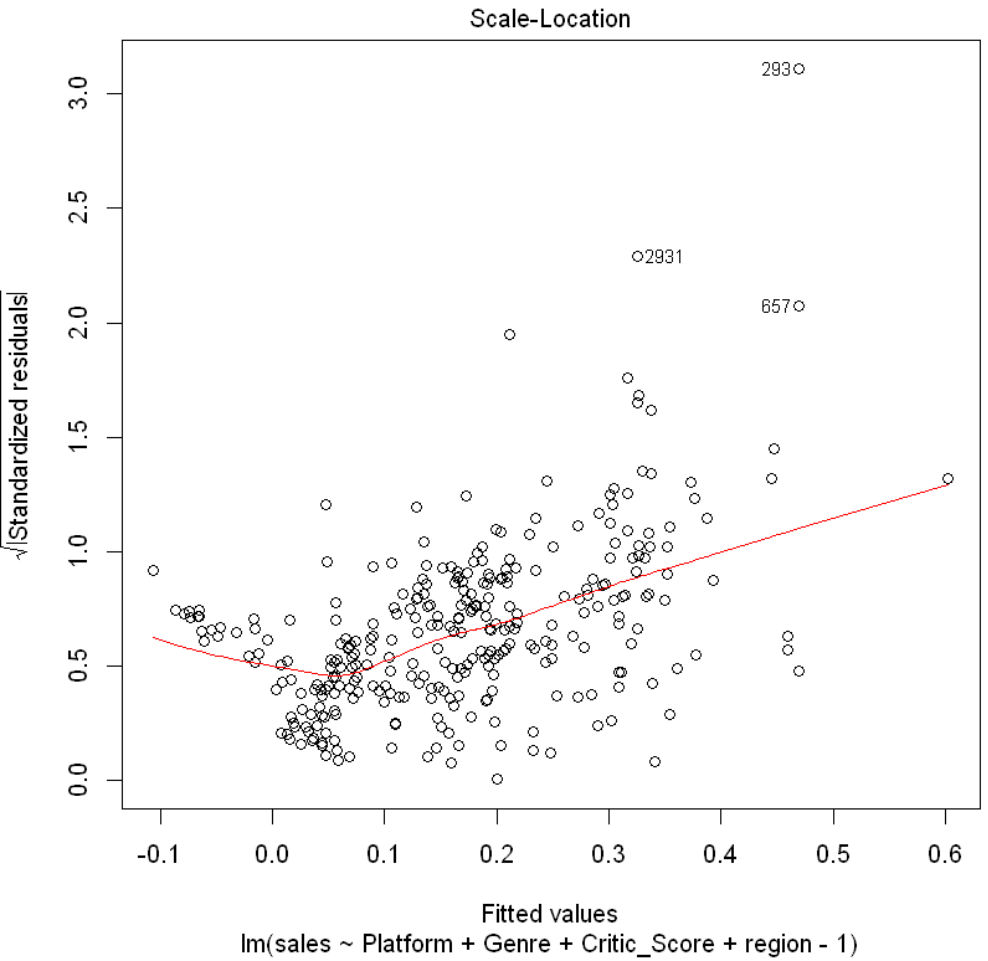
A **Q-Q plot** is a scatterplot created by **plotting** two sets of quantiles against one another.

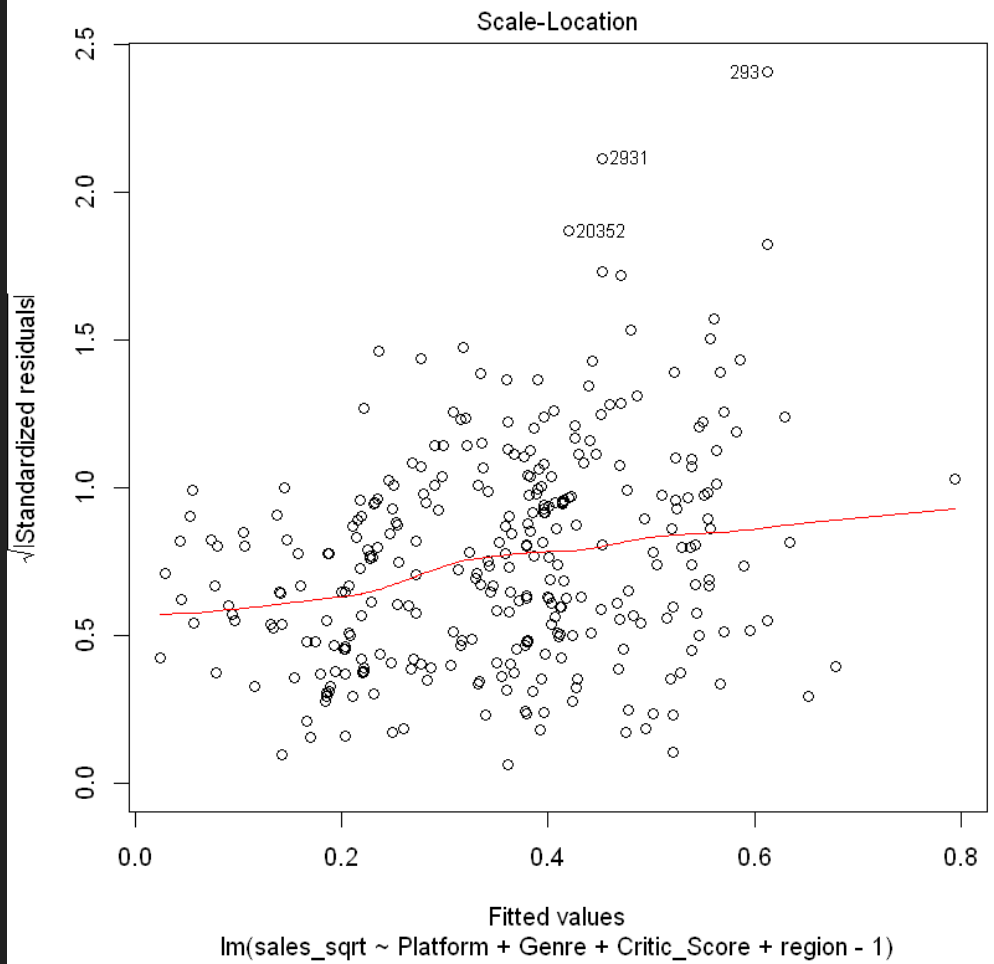




Scale Location

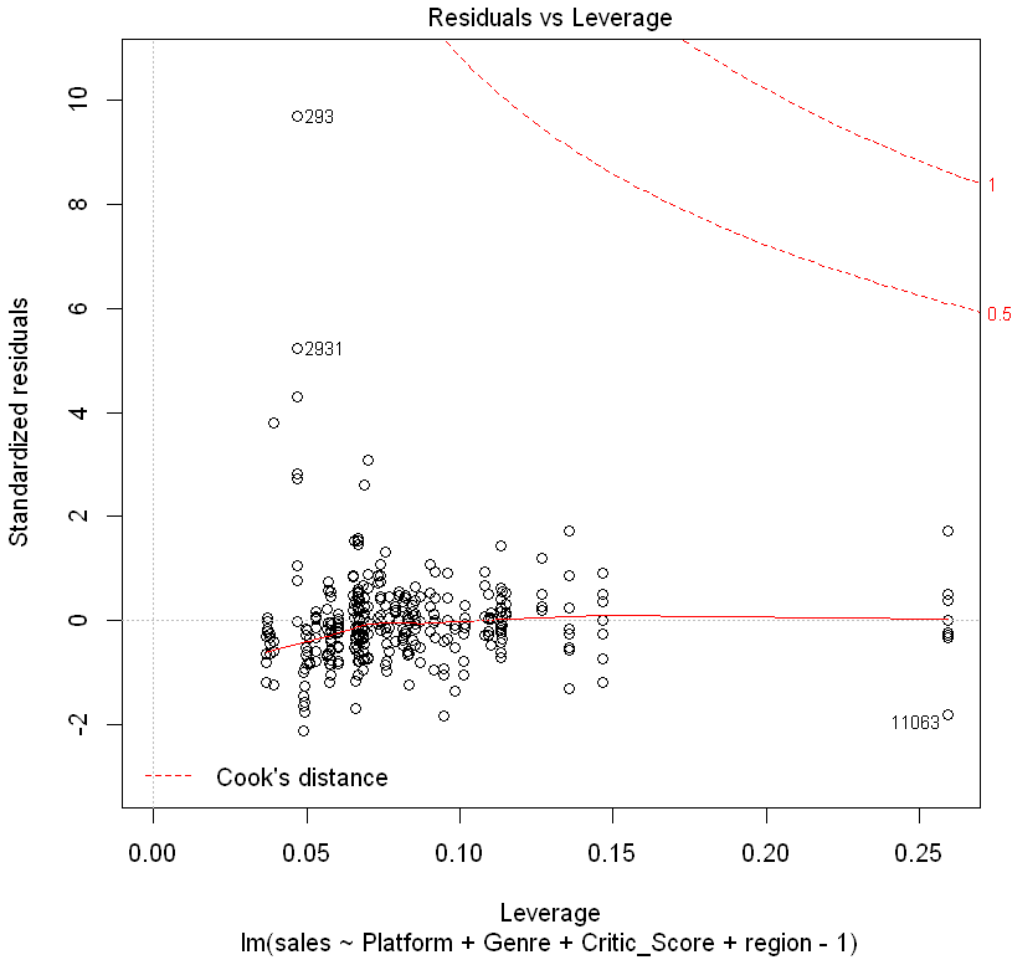
This plot shows if residuals are spread equally along the ranges of predictors

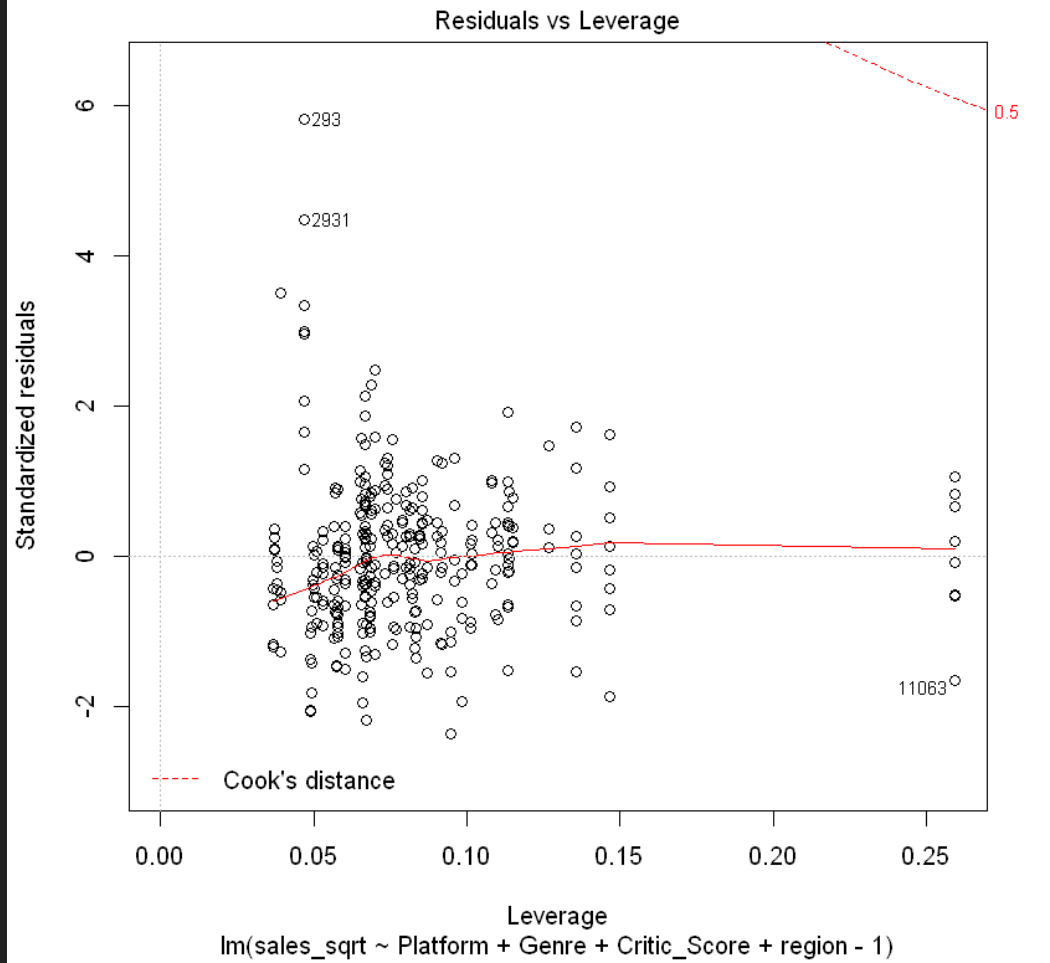




Residuals vs Leverage

The Residuals vs. Leverage plots helps you identify influential data points on your model





Prediction

We use lm2 to predict values and find the RMSE values, which predicts how far off the predicted values are to the observed values.

In [116]:
```
pred<- predict(lm2,new_video_nintendo_test1)
Eval <- data.frame(Game= new_video_nintendo_test1$Name, Actual = new_video_nintendo_test1$sales_sqrt)
pred <- round(pred,2)
Eval <- Eval[1:length(pred),]
Eval$Predicted <- abs(pred)
Eval$diff <- abs(Eval$Predicted - Eval$Actual)
row <- !is.na(Eval$Predicted)
eval1 <- Eval[row,]
head(eval1)
```

	Game	Actual	Predicted	diff
2	The Legend of Zelda: Oracle of Ages	0.9591663	0.80	0.15916630
3	Donkey Kong Jungle Beat	0.9165151	0.42	0.49651514
4	Star Fox: Assault	0.8246211	0.43	0.39462113
6	Excitebike 64	0.8062258	0.54	0.26622577
7	Paper Mario: Color Splash	0.5916080	0.50	0.09160798
8	Endless Ocean: Blue World	0.6782330	0.72	0.04176700

In [110]:
```
RMSE <- sqrt(mean(eval1$diff^2))
RMSE
```

0.142725576093331

We get RMSE value to be 0.14

Hypothesis Testing

H0 : NA_sales does not depend on Genre

```
fit <- lm( NA_Sales ~ Genre , data = vgsales)
summary(fit)
```

```
Residual standard error: 0.887 on 2396 degrees of freedom
Multiple R-squared:  0.05802,   Adjusted R-squared:  0.05095
F-statistic: 8.199 on 18 and 2396 DF,  p-value: < 2.2e-16
```

We find that the p-value < 0.05, So we can reject the hypothesis.

North American sales depends on the genre of the game.

H0 : PAL_sales does not depend on Genre

```
fit <- lm( PAL_Sales ~ Genre , data = vgsales)
summary(fit)
```

```
Residual standard error: 0.7609 on 2396 degrees of freedom
Multiple R-squared:  0.0659,    Adjusted R-squared:  0.05888
F-statistic: 9.391 on 18 and 2396 DF,  p-value: < 2.2e-16
```

We find that the p-value < 0.05, So we can reject the hypothesis.

European sales depends on the genre of the game.

H0 : JP_sales does not depend on Genre

```
fit <- lm(JP_Sales ~ Genre , data = vgsales)
summary(fit)
```

```
Residual standard error: 0.1895 on 2396 degrees of freedom
Multiple R-squared:  0.03288,   Adjusted R-squared:  0.02561
F-statistic: 4.525 on 18 and 2396 DF,  p-value: 7.398e-10
```

We find that the p-value < 0.05, So we can reject the hypothesis.

Japanian sales depends on the genre of the game.

Conclusion

- The majority of video games were not selling well, and their data fits the linear regression model very well.
- The data of best-selling games(Super Mario, Call of Duty, etc.) doesn't fit the model well.
- When sales is big enough, sales variable itself is also an influential factor to regression model.
- Finally, we used square rooted sales as a transform of response Y thus eliminate the error due to butterfly effect. The higher R-square presents a better fitting model.
- Lm2 has a R-Squared value of 0.8621 compared to lm1 with 0.5519

THANK YOU