Data 115 Final Project: Nintendo Game Analysis

Libby Stephan

The Big Question

What would have been the ideal game for Nintendo to release during the 2010s? How do North America's and Japan's tastes differ in this hypothetical game?

This analysis attempts to find characteristics (genre, platform, franchise) of a Nintendo game that would have sold the most number of copies during the 2010s. This curiosity extends to determine if Japan, the home country of Nintendo, would have differing preferences in games than North America.

The Data

The data used for this analysis was taken from Kaggle. The original poster, Gregory Smith, claims to have scraped the data from vgdata using a script based on BeautifulSoup using Python. The record consists of more than 16,500 games with sales (defined by copies sold) greater than 100,000. However, two games were removed due to incompleteness, and little else is mentioned about this dataset's processing or its original purpose.

The variables in this dataset are the game's name, platform, year released, genre, and publisher. It also includes the number of copies sold in millions from North America, Europe, Japan, other (rest of the world), and total globally. Lastly, a rank is included based on global sales. Unfortunately, the original author does not specify the period of time the sales were based on, but judging on the most recent games included, it can assume it is not current as of December 2022.

Nonetheless, I selected this data for this project because I have a personal interest in video games and a soft spot for Nintendo. This data will also be sufficient to answer my big question. Sadly, the most recent Nintendo game in this dataset is from 2016 and does not include the newest console release, the Nintendo Switch, which saw massive success and brought an onslaught of new games. Therefore, modern predictions are not appropriate, which is why I am limiting my research question to the 2010s.

Initial Cleaning and Processing

```
vgsales.df = read.csv("~/Documents/School/Data 115/FinalProject/vgsales.csv")
```

Managing Missing Years:

I first noticed some missing values in the Year column. To handle this, I converted Year into a numeric value, then replaced each NA with the mean of all the years (2006).

```
#replacing n/a with mean
vgsales.df$Year = as.numeric(vgsales.df$Year)
summary(vgsales.df$Year)
#mean of years = 2006
vgsales.df$Year[is.na(vgsales.df$Year)] = 2006
```

Splitting Up Nintendo:

It was at this point I decided to narrow my research question to Nintendo publishing, so a new data frame was created to only hold Nintendo games.

```
nintendo.df = vgsales.df[vgsales.df$Publisher == "Nintendo",]
summary(nintendo.df)
```

```
Rank              Name               Year              Genre                 NA_Sales
Min.   :    1     Length:703         Min.   :1983      Length:703            Min.   : 0.000
1st Qu.:  719     Class :character   1st Qu.:2000      Class :character      1st Qu.: 0.005
Median : 2335     Mode  :character   Median :2005      Mode  :character      Median : 0.370
Mean   : 3861                        Mean   :2004                            Mean   : 1.162
3rd Qu.: 6010     Platform           3rd Qu.:2009      Publisher             3rd Qu.: 0.960
Max.   :16545     Length:703         Max.   :2016      Length:703            Max.   :41.490
                  Class :character                     Class :character
                  Mode  :character                     Mode  :character

EU_Sales          JP Sales           Other_Sales       Global_Sales
Min.   : 0.0000   1st Qu.: 0.0900    Min.   :0.0000    Min.   : 0.010
1st Qu.: 0.0000   Min.   : 0.0000    1st Qu.:0.0000    1st Qu.: 0.290
Median : 0.1200   Median : 0.2800    Median :0.0300    Median : 0.890
Mean   : 0.5956   Mean   : 0.6478    Mean   :0.1356    Mean   : 2.541
3rd Qu.: 0.4800   3rd Qu.: 0.7300    3rd Qu.:0.0900    3rd Qu.: 2.250
Max.   :29.0200   Max.   :10.2200    Max.   :8.4600    Max.   :82.740
```

The resulting Nintendo data frame holds 703 games, which is quite the decrease from the original data set. But this size decreased further by only including games released in the 2010 decade.

```
nintendoRecent.df = nintendo.df[nintendo.df$Year >= 2010,]
```

This left a sample size of 170. Not ideal compared to the original data set, but not terrible.

Creating "Franchise" Column:

Franchises play a major role in sales especially for a brand like Nintendo. Because Nintendo houses a large number of dominating gaming franchises, a new column was created to reflect the associated franchise of the game. To accomplish this, specific words were searched in the game's name to determine if a game was marketed within a certain franchise.

```
library(stringr)

#formalize the name strings to lowercase
nintendo.df$Name = str_to_lower(nintendo.df$Name)

nintendo.df$Franchise = "Other"
nintendo.df$Franchise[str_count(nintendo.df$Name, "pokemon") != 0] = "Pokemon"
nintendo.df$Franchise[str_count(nintendo.df$Name, "pokémon") != 0] = "Pokemon"
nintendo.df$Franchise[str_count(nintendo.df$Name, "mario") != 0] = "Mario"
nintendo.df$Franchise[str_count(nintendo.df$Name, "yoshi") != 0] = "Yoshi"
nintendo.df$Franchise[str_count(nintendo.df$Name, "zelda") != 0] = "Zelda"
nintendo.df$Franchise[str_count(nintendo.df$Name, "animal crossing") != 0] = "Animal
Crossing"
nintendo.df$Franchise[str_count(nintendo.df$Name, "donkey kong") != 0] = "Donkey Kong"
nintendo.df$Franchise[str_count(nintendo.df$Name, "kirby") != 0] = "Kirby"
nintendo.df$Franchise[str_count(nintendo.df$Name, "wii") != 0] = "Wii Exclusive"
nintendo.df$Franchise[str_count(nintendo.df$Name, "smash bro") != 0] = "Smash Bros"
```

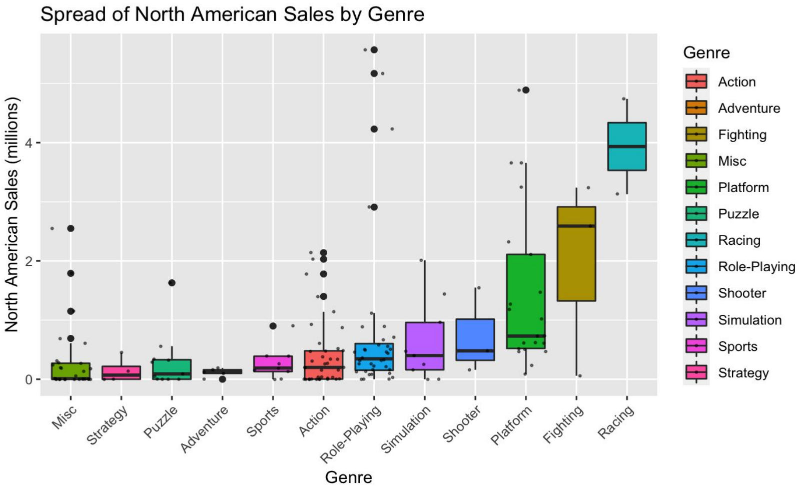
It is important to note that duplicates in names were ignored as these duplicates may have different platforms, and therefore technically a different entry.

Exploratory Analysis

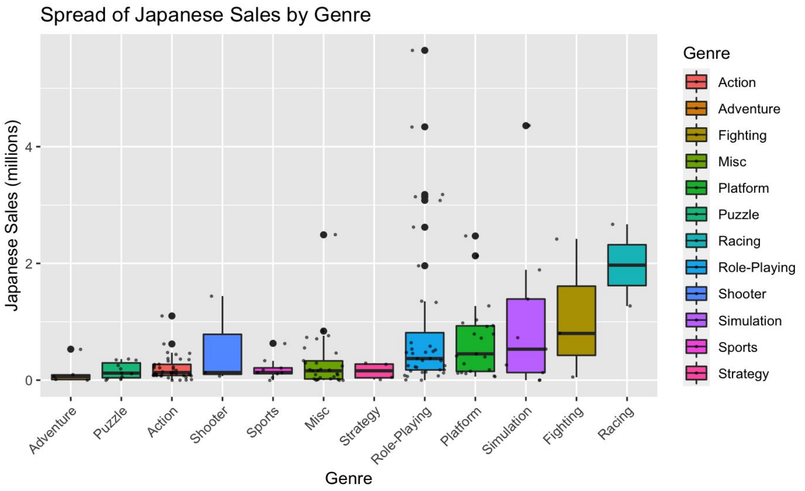
This section examines how each variable (genre, platform, franchise) influenced the sales from North America, Japan, and globally.

Genre:

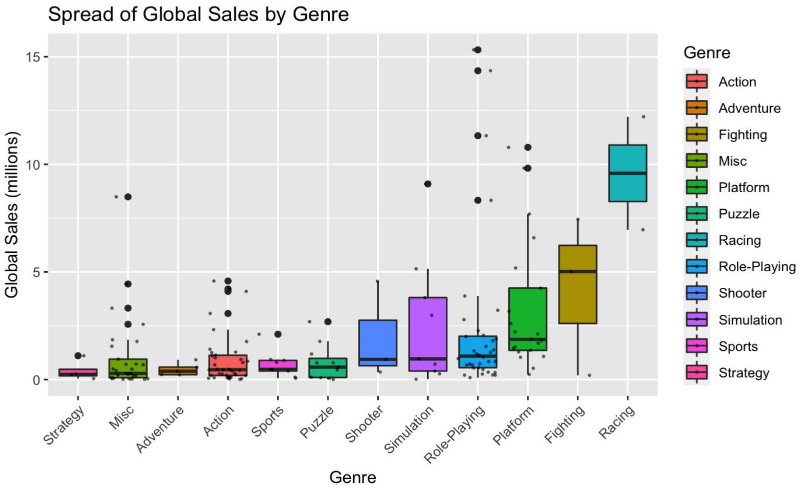
```
ggplot(nintendoRecent.df, aes(x =
reorder(Genre, NA_Sales, FUN =
median), y = NA_Sales, fill= Genre)) +
  geom_boxplot() +
  scale_x_discrete(guide =
guide_axis(angle = 45)) +
  labs(x = "Genre", y = "North
American Sales (millions)", title =
"Spread of North American Sales by
Genre") +
  geom_jitter(color="black", size=0.4,
alpha=0.5)
```



```
ggplot(nintendoRecent.df, aes(x =
reorder(Genre, JP_Sales, FUN =
median), y = JP_Sales, fill= Genre)) +
  geom_boxplot() +
  scale_x_discrete(guide =
guide_axis(angle = 45)) +
  labs(x = "Genre", y = "Japanese
Sales (millions)", title = "Spread of
Japanese Sales by Genre")+
  geom_jitter(color="black", size=0.4,
alpha=0.5)
```



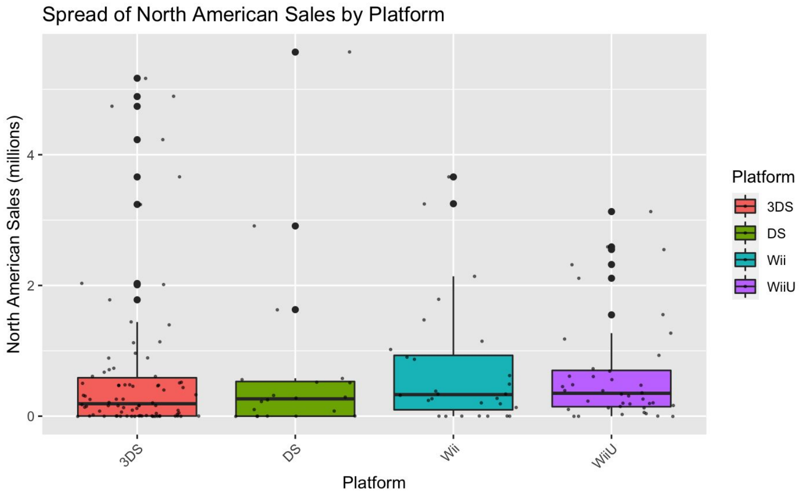
```
ggplot(nintendoRecent.df, aes(x =
reorder(Genre, Global_Sales, FUN =
median), y = Global_Sales, fill=
Genre)) +
  geom_boxplot() +
  scale_x_discrete(guide =
guide_axis(angle = 45)) +
  labs(x = "Genre", y = "Global Sales
(millions)", title = "Spread of Global
Sales by Genre")+
  geom_jitter(color="black", size=0.4,
alpha=0.5)
```



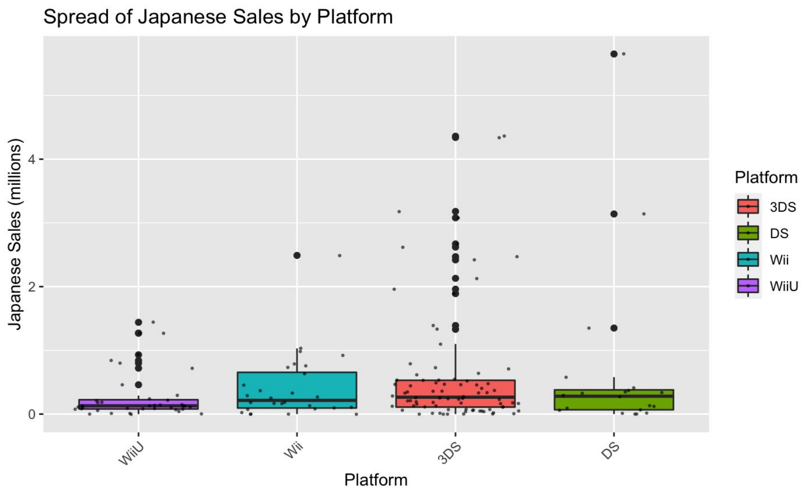
From these plots, it becomes evident that racing is the genre with the highest median sales across all three regions with fighting in second.

Platform:

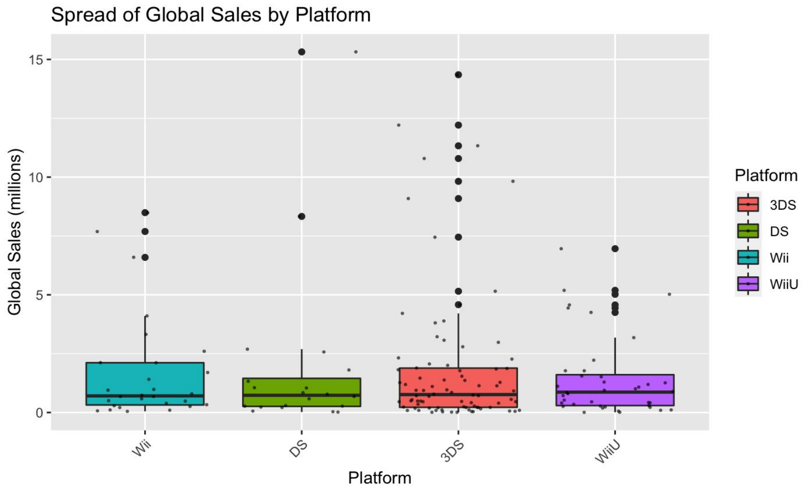
```
ggplot(nintendoRecent.df, aes(x =
reorder(Platform, NA_Sales, FUN =
median), y = NA_Sales, fill =
Platform)) +
  geom_boxplot() +
  scale_x_discrete(guide =
guide_axis(angle = 45)) +
  labs(x = "Platform", y = "North
American Sales (millions)", title =
"Spread of North American Sales by
Platform") +
  geom_jitter(color="black", size=0.4,
alpha=0.5)
```


Spread of North American Sales by Platform

```
ggplot(nintendoRecent.df, aes(x =
reorder(Platform, JP_Sales, FUN =
median), y = JP_Sales, fill =
Platform)) +
  geom_boxplot() +
  scale_x_discrete(guide =
guide_axis(angle = 45)) +
  labs(x = "Platform", y = "Japanese
Sales (millions)", title = "Spread of
Japanese Sales by Platform") +
  geom_jitter(color="black", size=0.4,
alpha=0.5)
```


Spread of Japanese Sales by Platform

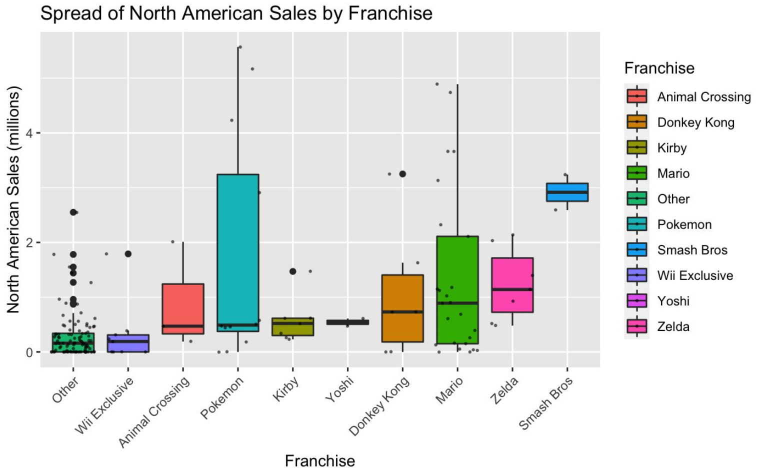
```
ggplot(nintendoRecent.df, aes(x =
reorder(Platform, Global_Sales, FUN =
median), y = Global_Sales, fill =
Platform)) +
  geom_boxplot() +
  scale_x_discrete(guide =
guide_axis(angle = 45)) +
  labs(x = "Platform", y = "NGlobal
Sales (millions)", title = "Spread of
Global Sales by Platform") +
  geom_jitter(color="black", size=0.4,
alpha=0.5)
```


Spread of Global Sales by Platform

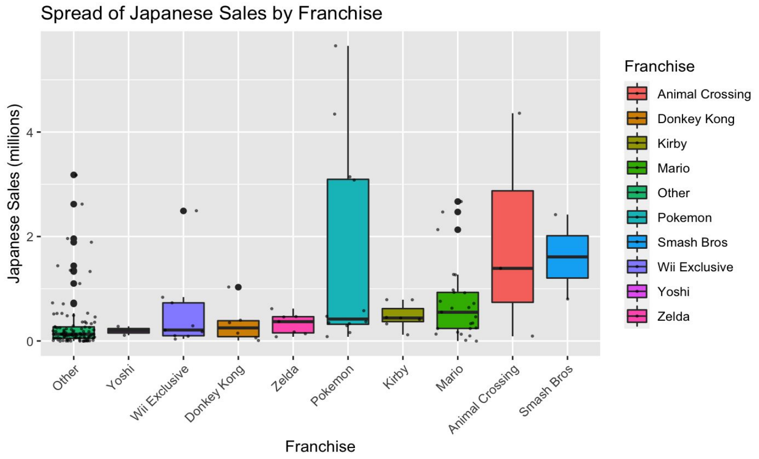
Based on these plots, we can conclude that the platform has very little to do with sales success as all medians have very little difference between each other, and the three region's leading platforms are inconsistent.

Franchise:

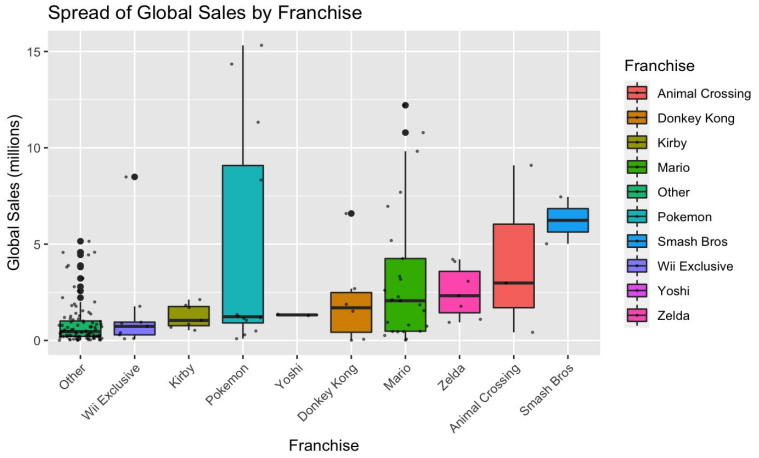
```
ggplot(nintendoRecent.df, aes(x =
reorder(Franchise, NA_Sales, FUN = median),
y = NA_Sales, fill = Franchise)) +
  geom_boxplot() +
  scale_x_discrete(guide = guide_axis(angle
= 45)) +
  labs(x = "Franchise", y = "North American
Sales (millions)", title = "Spread of North
American Sales by Franchise") +
  geom_jitter(color="black", size=0.4,
alpha=0.5)
```


Spread of North American Sales by Franchise

```
ggplot(nintendoRecent.df, aes(x =
reorder(Franchise, JP_Sales, FUN =
median), y = JP_Sales, fill = Franchise))
+
  geom_boxplot() +
  scale_x_discrete(guide =
guide_axis(angle = 45)) +
  labs(x = "Franchise", y = "Japanese
Sales (millions)", title = "Spread of
Japanese Sales by Franchise") +
  geom_jitter(color="black", size=0.4,
alpha=0.5)
```


Spread of Japanese Sales by Franchise

```
ggplot(nintendoRecent.df, aes(x =
reorder(Franchise, Global_Sales, FUN =
median), y = Global_Sales, fill =
Franchise)) +
  geom_boxplot() +
  scale_x_discrete(guide =
guide_axis(angle = 45)) +
  labs(x = "Franchise", y = "Global Sales
(millions)", title = "Spread of Global
Sales by Franchise") +
  geom_jitter(color="black", size=0.4,
alpha=0.5)
```


Spread of Global Sales by Franchise

As seen in these plots, Smash Bros appears to lead all three regions in sales based on franchise. Animal Crossing comes in second for Japan and global, whereas North America's second is held by Zelda. Interestingly, Pokemon in all three regions has an extremely wide interquartile range.

From these nine plots, we can estimate that platform will have little effect on sales. On the other hand, genre and franchise may have a much higher effect. While all three regions (North America, Japan, global) appear to appreciate the Smash Bros. franchise and racing genre the most, North America seems to appreciate the Animal Crossing franchise far less than Japan, but is more receptive to Zelda. Universally, Pokemon games appear to receive varying degrees of success as each has a large spread. Finally, franchises made exclusively for the Wii and those unaffiliated with larger franchises (other) consistently rank last across all regions.

Modeling

To model this data, various multiple regressions were used where the independent variables were platform, genre, and franchise, and the dependent variables were North American sales, Japanese sales, and global sales.

This method of analysis was chosen because of its ability to predict a numerical output (sales) based on multiple categorical input (platform, genre, franchise) for an ideal hypothetical game.

The pro to this method are the predictions the model will allow given various input variables. However, the con is that the input variables have a high chance of collinearity. This results in a model that is not optimally fitted.

Alternative methods were scarce given the current statistical knowledge I possess.

Checking for Collinearity:

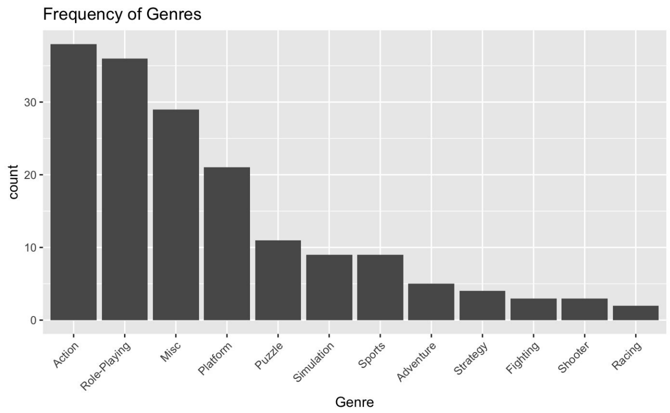
There was great suspicion of collinearity between genre and franchise because franchises typically stick to one or two genres of gameplay. To determine this suspicion's accuracy, a frequency table between genre and franchise was created.

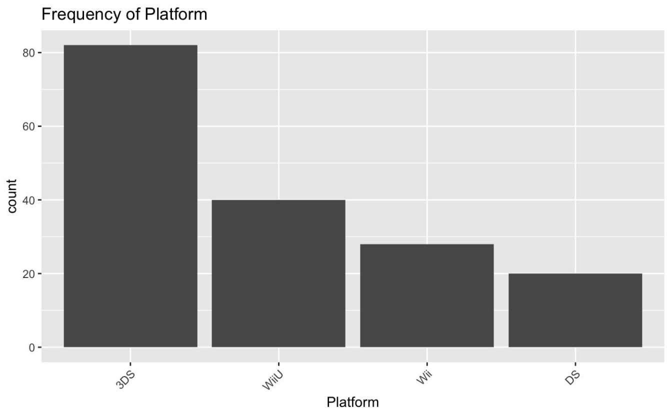
```
table(nintendoRecent.df$Genre, nintendoRecent.df$Franchise)
```

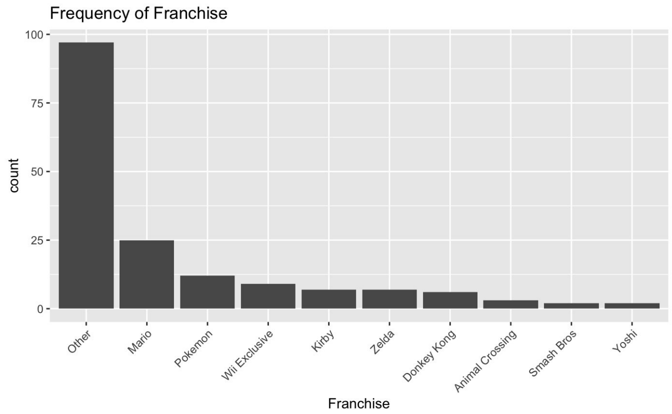
```
                Animal Crossing Donkey Kong Kirby Mario Other Pokemon Smash Bros Wii Exclusive Yoshi Zelda
  Action                      0           0     2     2    24       2          0             1     0     7
  Adventure                   0           0     0     0     5       0          0             0     0     0
  Fighting                    0           0     0     0     1       0          2             0     0     0
  Misc                        1           0     0     4    16       2          0             6     0     0
  Platform                     0           3     5     7     4       0          0             0     2     0
  Puzzle                      0           3     0     1     7       0          0             0     0     0
  Racing                      0           0     0     2     0       0          0             0     0     0
  Role-Playing                0           0     0     3    25       8          0             0     0     0
  Shooter                     0           0     0     0     3       0          0             0     0     0
  Simulation                  2           0     0     0     7       0          0             0     0     0
  Sports                      0           0     0     6     1       0          0             2     0     0
  Strategy                    0           0     0     0     4       0          0             0     0     0
```

As expected, this table is extremely sparse and uneven. In fact, when creating a bar chart of each input variable, it is evident that each group's sizes are wildly uneven.

```
library(forcats)
ggplot(nintendoRecent.df, aes(x = fct_infreq(Genre))) + geom_bar()
+
  labs(x = "Genre", title = "Frequency of Genres") +
  scale_x_discrete(guide = guide_axis(angle = 45))
gzgplot(nintendoRecent.df, aes(x = fct_infreq(Platform))) +
geom_bar() +
  labs(x = "Platform", title = "Frequency of Platform") +
  scale_x_discrete(guide = guide_axis(angle = 45))
ggplot(nintendoRecent.df, aes(x = fct_infreq(Franchise))) +
geom_bar() +
  labs(x = "Franchise", title = "Frequency of Franchise") +
  scale_x_discrete(guide = guide_axis(angle = 45))
```







While this uneven disbursement does not bode well for the model's fit, a more formal check for collinearity was performed using Chi-squared tests.

```
chisq.test(nintendoRecent.df$Genre, nintendoRecent.df$Franchise)
chisq.test(nintendoRecent.df$Genre, nintendoRecent.df$Platform)
chisq.test(nintendoRecent.df$Platform, nintendoRecent.df$Franchise)
```

```
	Pearson's Chi-squared test

data:  nintendoRecent.df$Genre and nintendoRecent.df$Franchise
X-squared = 322.15, df = 99, p-value < 2.2e-16

	Pearson's Chi-squared test

data:  nintendoRecent.df$Genre and nintendoRecent.df$Platform
X-squared = 46.066, df = 33, p-value = 0.06496


	Pearson's Chi-squared test

data:  nintendoRecent.df$Platform and nintendoRecent.df$Franchise
X-squared = 43.986, df = 27, p-value = 0.02075
```

With p-values of less than 0.5 for the first and third Chi-squared tests, it is concluded that genre and franchise, and platform and franchise are not independent of each other and thus, have collinearity.

Given the wild unevenness of cell sizes and the results of the Chi-squared tests, the model made will not accurately reflect the data as well as one would hope. Despite this, interesting insights can still be made however good the fit.

Making the Models:

For each region, a simple model was made for genre, platform, and franchise before combining all three into a multiple regression model.

North America:

```
summary(lm(NA_Sales ~ Genre, nintendoRecent.df))
summary(lm(NA_Sales ~ Platform, nintendoRecent.df)
summary(lm(NA_Sales ~ Franchise, nintendoRecent.df))
nintendoRecentNA.lm = lm(NA_Sales ~ Genre + Franchise + Platform, data = nintendoRecent.df)
summary(nintendoRecentNA.lm)
```

```
Call:
lm(formula = NA_Sales ~ Genre, data = nintendoRecent.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.9033 -0.4158 -0.2491  0.0617  4.7558 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)         0.4158     0.1529   2.720  0.00726 ** 
GenreAdventure     -0.3018     0.4483  -0.673  0.50181    
GenreFighting       1.5475     0.5651   2.738  0.00688 ** 
GenreMisc          -0.1172     0.2324  -0.504  0.61478    
GenrePlatform       1.0328     0.2562   4.031 8.63e-05 ***
GenrePuzzle        -0.1176     0.3226  -0.365  0.71596    
GenreRacing         3.5192     0.6836   5.148 7.74e-07 ***
GenreRole-Playing   0.3984     0.2192   1.818  0.07101 .  
GenreShooter        0.3142     0.5651   0.556  0.57900    
GenreSimulation     0.2164     0.3493   0.620  0.53644    
GenreSports        -0.1480     0.3493  -0.424  0.67237    
GenreStrategy      -0.2683     0.4953  -0.542  0.58885    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9423 on 158 degrees of freedom
Multiple R-squared:   0.27,	Adjusted R-squared:  0.2192 
F-statistic: 5.313 on 11 and 158 DF,  p-value: 3.966e-07

Call:
lm(formula = NA_Sales ~ Franchise, data = nintendoRecent.df)

Residuals:
   Min     1Q Median     3Q    Max 
-1.710 -0.282 -0.152  0.093  3.860 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)  
(Intercept)              0.8900     0.5385   1.653   0.1003  
FranchiseDonkey Kong     0.1667     0.6595   0.253   0.8008  
FranchiseKirby          -0.3114     0.6436  -0.484   0.6291  
FranchiseMario           0.4412     0.5699   0.774   0.4400  
FranchiseOther          -0.6080     0.5467  -1.112   0.2678  
FranchisePokemon         0.8200     0.6020   1.362   0.1751  
FranchiseSmash Bros      2.0250     0.8514   2.378   0.0186 *
FranchiseWii Exclusive  -0.5656     0.6218  -0.910   0.3644  
FranchiseYoshi          -0.3500     0.8514  -0.411   0.6816  
FranchiseZelda           0.3443     0.6436   0.535   0.5934  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9327 on 160 degrees of freedom
Multiple R-squared:  0.2759,	Adjusted R-squared:  0.2351 
F-statistic: 6.773 on 9 and 160 DF,  p-value: 3.276e-08
```

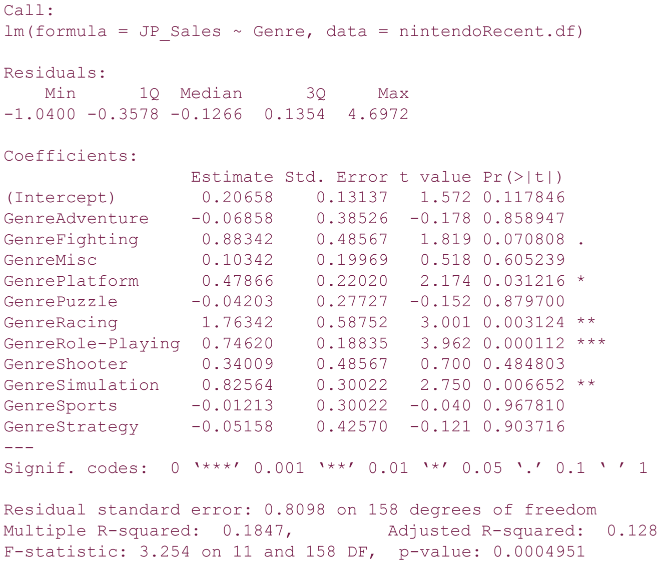
```
Call:
lm(formula = NA_Sales ~ Platform, data = nintendoRecent.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.7157 -0.6128 -0.3984 -0.0330  4.8790 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.643049   0.118789   5.413 2.13e-07 ***
PlatformDS   0.047951   0.268263   0.179    0.858    
PlatformWii  0.072666   0.235447   0.309    0.758    
PlatformWiiU 0.009201   0.207456   0.044    0.965    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.076 on 166 degrees of freedom
Multiple R-squared:  0.0006802,	Adjusted R-squared:  -0.01738 
F-statistic: 0.03766 on 3 and 166 DF,  p-value: 0.9902

Call:
lm(formula = NA Sales ~ Genre + Franchise + Platform, data = 
nintendoRecent.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.4810 -0.3437 -0.0717  0.1831  3.6801 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)             0.67709    0.55078   1.229   0.2209    
GenreAdventure         -0.05727    0.41035  -0.140   0.8892    
GenreFighting          -0.02334    0.84046  -0.028   0.9779    
GenreMisc              -0.12710    0.22820  -0.557   0.5784    
GenrePlatform           1.24738    0.27970   4.460 1.63e-05 ***
GenrePuzzle             0.04470    0.31180   0.143   0.8862    
GenreRacing             3.20247    0.63727   5.025 1.45e-06 ***
GenreRole-Playing       0.31929    0.21228   1.504   0.1347    
GenreShooter            0.67875    0.50354   1.348   0.1798    
GenreSimulation         0.40699    0.34309   1.186   0.2375    
GenreSports            -0.36387    0.35076  -1.037   0.3013    
GenreStrategy           0.05380    0.44365   0.121   0.9036    
FranchiseDonkey Kong   -0.30876    0.65216  -0.473   0.6366    
FranchiseKirby         -1.10706    0.64235  -1.723   0.0869 .  
FranchiseMario          0.07952    0.56197   0.141   0.8877    
FranchiseOther         -0.59375    0.52285  -1.136   0.2580    
FranchisePokemon        0.80391    0.58959   1.364   0.1748    
FranchiseSmash Bros     2.28532    1.13648   2.011   0.0462 *  
FranchiseWii Exclusive -0.31037    0.59838  -0.519   0.6048    
FranchiseYoshi         -1.36040    0.81947  -1.660   0.0990 .  
FranchiseZelda          0.53374    0.62615   0.852   0.3954    
PlatformDS              0.08957    0.22176   0.404   0.6869    
PlatformWii             0.26049    0.19706   1.322   0.1883    
PlatformWiiU           -0.04814    0.17340  -0.278   0.7817    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8236 on 146 degrees of freedom
Multiple R-squared:  0.4847,	Adjusted R-squared:  0.4036 
F-statistic: 5.972 on 23 and 146 DF,  p-value: 3.774e-12
```

From the North American simple models and their p-values, it is observed that genres of fighting, platforming, and racing are statistically significant, meaning that their influence on North American sales is not due to chance. Platform has no significance, but the Smash Bros. franchise does.

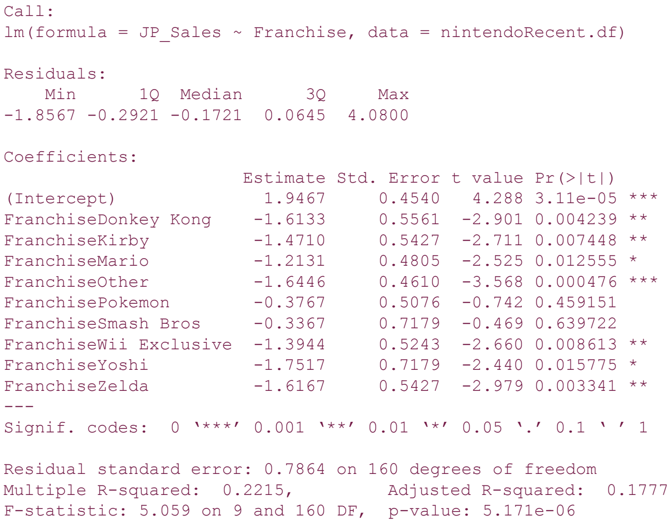
From the multiple regression model, a similar result is achieved. However, fighting as a genre is not statistically significant.

Japan:

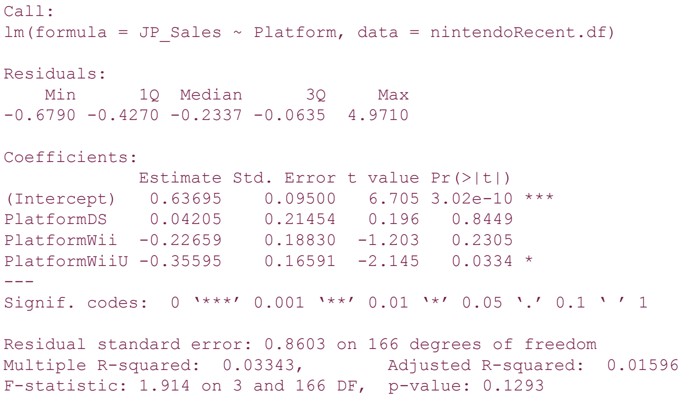
```
summary(lm(JP_Sales ~ Genre, nintendoRecent.df))
summary(lm(JP_Sales ~ Platform, nintendoRecent.df))
summary(lm(JP_Sales ~ Franchise, nintendoRecent.df))
nintendoRecentJP.lm = lm(JP_Sales ~ Genre + Franchise + Platform, data = nintendoRecent.df)
summary(nintendoRecentJP.lm)
```



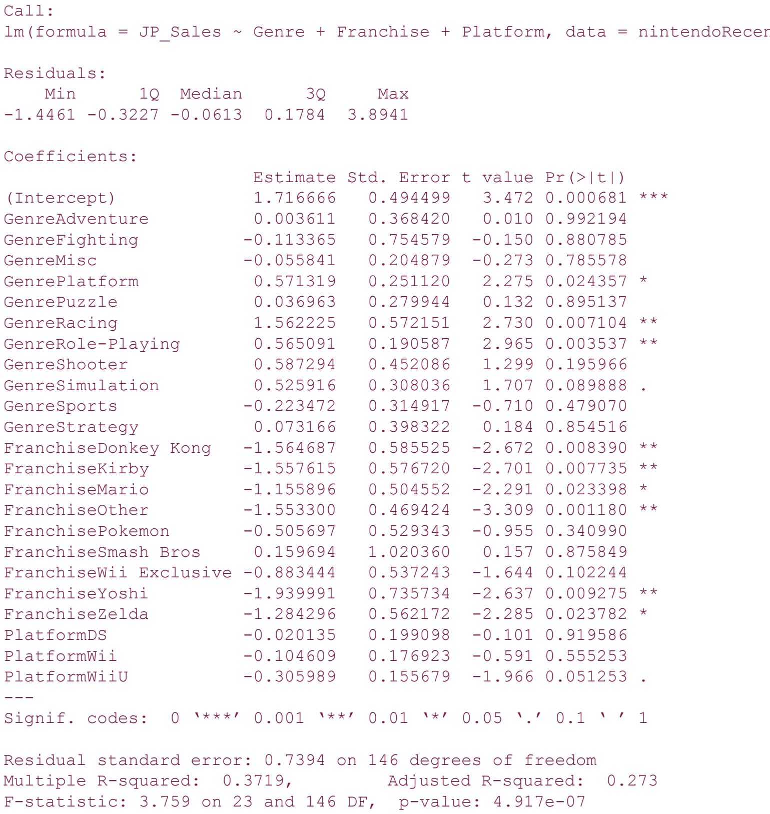
```
Call:
lm(formula = JP_Sales ~ Genre, data = nintendoRecent.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.0400 -0.3578 -0.1266  0.1354  4.6972 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)        0.20658    0.13137   1.572 0.117846    
GenreAdventure    -0.06858    0.38526  -0.178 0.858947    
GenreFighting      0.88342    0.48567   1.819 0.070808 .  
GenreMisc          0.10342    0.19969   0.518 0.605239    
GenrePlatform      0.47866    0.22020   2.174 0.031216 *  
GenrePuzzle       -0.04203    0.27727  -0.152 0.879700    
GenreRacing        1.76342    0.58752   3.001 0.003124 ** 
GenreRole-Playing  0.74620    0.18835   3.962 0.000112 ***
GenreShooter       0.34009    0.48567   0.700 0.484803    
GenreSimulation    0.82564    0.30022   2.750 0.006652 ** 
GenreSports       -0.01213    0.30022  -0.040 0.967810    
GenreStrategy     -0.05158    0.42570  -0.121 0.903716    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8098 on 158 degrees of freedom
Multiple R-squared:  0.1847,	Adjusted R-squared:  0.128 
F-statistic: 3.254 on 11 and 158 DF,  p-value: 0.0004951
```



```
Call:
lm(formula = JP_Sales ~ Franchise, data = nintendoRecent.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.8567 -0.2921 -0.1721  0.0645  4.0800 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)              1.9467     0.4540   4.288 3.11e-05 ***
FranchiseDonkey Kong    -1.6133     0.5561  -2.901 0.004239 ** 
FranchiseKirby          -1.4710     0.5427  -2.711 0.007448 ** 
FranchiseMario          -1.2131     0.4805  -2.525 0.012555 *  
FranchiseOther          -1.6446     0.4610  -3.568 0.000476 ***
FranchisePokemon        -0.3767     0.5076  -0.742 0.459151    
FranchiseSmash Bros     -0.3367     0.7179  -0.469 0.639722    
FranchiseWii Exclusive  -1.3944     0.5243  -2.660 0.008613 ** 
FranchiseYoshi          -1.7517     0.7179  -2.440 0.015775 *  
FranchiseZelda          -1.6167     0.5427  -2.979 0.003341 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7864 on 160 degrees of freedom
Multiple R-squared:  0.2215,	Adjusted R-squared:  0.1777 
F-statistic: 5.059 on 9 and 160 DF,  p-value: 5.171e-06
```



```
Call:
lm(formula = JP_Sales ~ Platform, data = nintendoRecent.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.6790 -0.4270 -0.2337 -0.0635  4.9710 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.63695    0.09500   6.705 3.02e-10 ***
PlatformDS    0.04205    0.21454   0.196   0.8449    
PlatformWii  -0.22659    0.18830  -1.203   0.2305    
PlatformWiiU -0.35595    0.16591  -2.145   0.0334 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8603 on 166 degrees of freedom
Multiple R-squared:  0.03343,	Adjusted R-squared:  0.01596 
F-statistic: 1.914 on 3 and 166 DF,  p-value: 0.1293
```

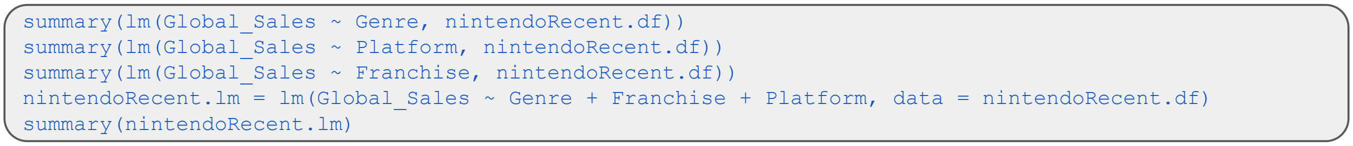


```
Call:
lm(formula = JP_Sales ~ Genre + Franchise + Platform, data = nintendoRecer

Residuals:
    Min      1Q  Median      3Q     Max 
-1.4461 -0.3227 -0.0613  0.1784  3.8941 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)             1.716666   0.494499   3.472 0.000681 ***
GenreAdventure          0.003611   0.368420   0.010 0.992194    
GenreFighting          -0.113365   0.754579  -0.150 0.880785    
GenreMisc              -0.055841   0.204879  -0.273 0.785578    
GenrePlatform           0.571319   0.251120   2.275 0.024357 *  
GenrePuzzle             0.036963   0.279944   0.132 0.895137    
GenreRacing             1.562225   0.572151   2.730 0.007104 ** 
GenreRole-Playing       0.565091   0.190587   2.965 0.003537 ** 
GenreShooter            0.587294   0.452086   1.299 0.195966    
GenreSimulation         0.525916   0.308036   1.707 0.089888 .  
GenreSports            -0.223472   0.314917  -0.710 0.479070    
GenreStrategy           0.073166   0.398322   0.184 0.854516    
FranchiseDonkey Kong   -1.564687   0.585525  -2.672 0.008390 ** 
FranchiseKirby         -1.557615   0.576720  -2.701 0.007735 ** 
FranchiseMario         -1.155896   0.504552  -2.291 0.023398 *  
FranchiseOther         -1.553300   0.469424  -3.309 0.001180 ** 
FranchisePokemon       -0.505697   0.529343  -0.955 0.340990    
FranchiseSmash Bros     0.159694   1.020360   0.157 0.875849    
FranchiseWii Exclusive -0.883444   0.537243  -1.644 0.102244    
FranchiseYoshi         -1.939991   0.735734  -2.637 0.009275 ** 
FranchiseZelda         -1.284296   0.562172  -2.285 0.023782 *  
PlatformDS             -0.020135   0.199098  -0.101 0.919586    
PlatformWii            -0.104609   0.176923  -0.591 0.555253    
PlatformWiiU           -0.305989   0.155679  -1.966 0.051253 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7394 on 146 degrees of freedom
Multiple R-squared:  0.3719,	Adjusted R-squared:  0.273 
F-statistic: 3.759 on 23 and 146 DF,  p-value: 4.917e-07
```

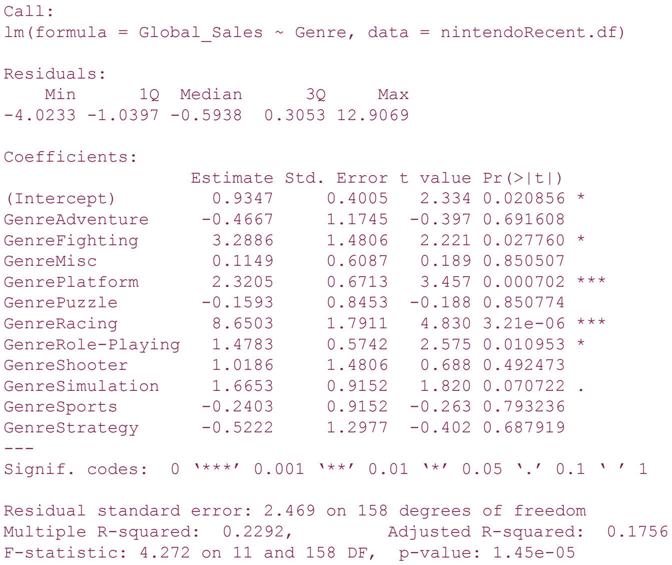
From the Japanese simple models and their p-values, it is observed that genres of platform, racing, role-playing, and simulation are statistically significant. Interestingly, fighting did not reach significance. The only significant platform is the WiiU. All franchises were significant except for Pokemon and Smash Bros.

From the multiple regression model, platform, racing, and role-playing are statistically significant in genres. All franchises are significant except Pokemon, Smash Bros., and Wii exclusives, and no platforms are significant.

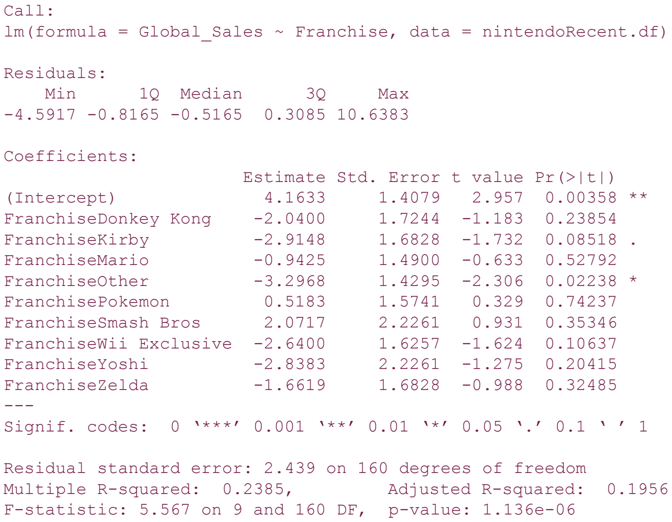
<u>Global:</u>



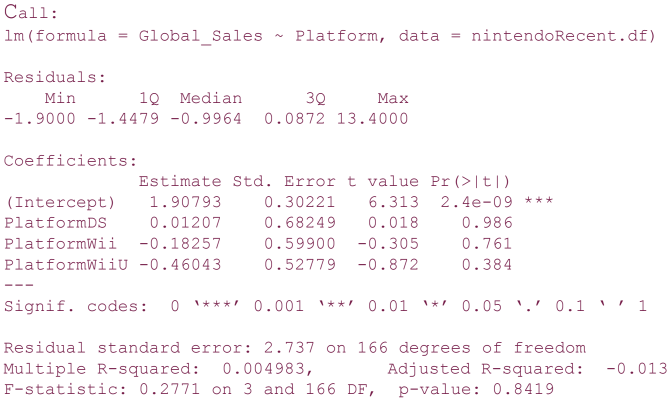
```
summary(lm(Global_Sales ~ Genre, nintendoRecent.df))
summary(lm(Global_Sales ~ Platform, nintendoRecent.df))
summary(lm(Global_Sales ~ Franchise, nintendoRecent.df))
nintendoRecent.lm = lm(Global_Sales ~ Genre + Franchise + Platform, data = nintendoRecent.df)
summary(nintendoRecent.lm)
```



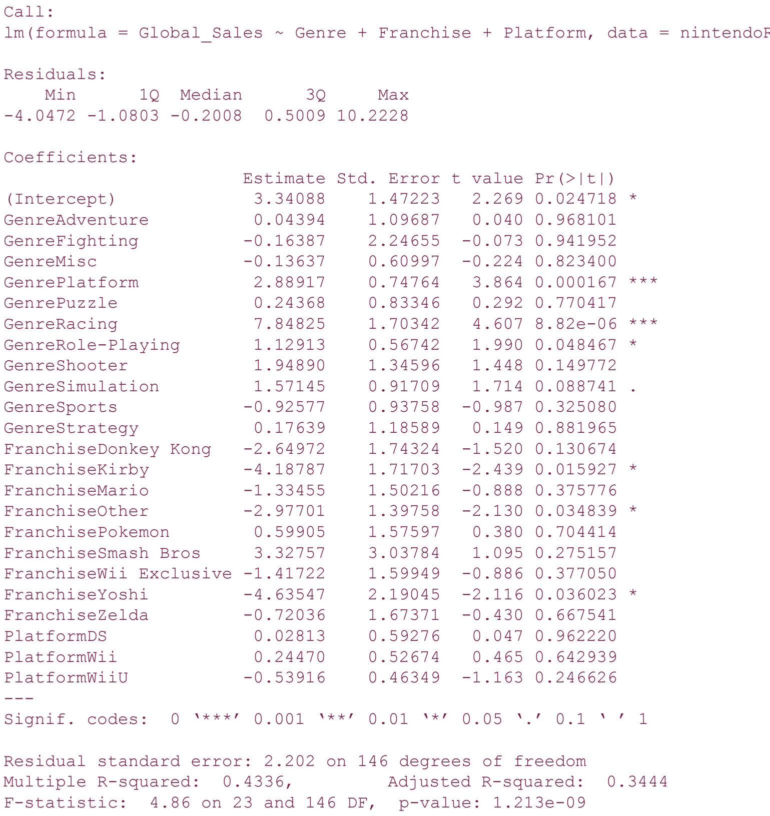
```
Call:
lm(formula = Global_Sales ~ Genre, data = nintendoRecent.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.0233 -1.0397 -0.5938  0.3053 12.9069 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)         0.9347     0.4005   2.334 0.020856 *  
GenreAdventure     -0.4667     1.1745  -0.397 0.691608    
GenreFighting       3.2886     1.4806   2.221 0.027760 *  
GenreMisc           0.1149     0.6087   0.189 0.850507    
GenrePlatform       2.3205     0.6713   3.457 0.000702 ***
GenrePuzzle        -0.1593     0.8453  -0.188 0.850774    
GenreRacing         8.6503     1.7911   4.830 3.21e-06 ***
GenreRole-Playing   1.4783     0.5742   2.575 0.010953 *  
GenreShooter        1.0186     1.4806   0.688 0.492473    
GenreSimulation     1.6653     0.9152   1.820 0.070722 .  
GenreSports        -0.2403     0.9152  -0.263 0.793236    
GenreStrategy      -0.5222     1.2977  -0.402 0.687919    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.469 on 158 degrees of freedom
Multiple R-squared:  0.2292,	Adjusted R-squared:  0.1756 
F-statistic: 4.272 on 11 and 158 DF,  p-value: 1.45e-05
```



```
Call:
lm(formula = Global_Sales ~ Franchise, data = nintendoRecent.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.5917 -0.8165 -0.5165  0.3085 10.6383 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)   
(Intercept)             4.1633     1.4079   2.957  0.00358 **
FranchiseDonkey Kong   -2.0400     1.7244  -1.183  0.23854   
FranchiseKirby         -2.9148     1.6828  -1.732  0.08518 . 
FranchiseMario         -0.9425     1.4900  -0.633  0.52792   
FranchiseOther         -3.2968     1.4295  -2.306  0.02238 * 
FranchisePokemon        0.5183     1.5741   0.329  0.74237   
FranchiseSmash Bros     2.0717     2.2261   0.931  0.35346   
FranchiseWii Exclusive -2.6400     1.6257  -1.624  0.10637   
FranchiseYoshi         -2.8383     2.2261  -1.275  0.20415   
FranchiseZelda         -1.6619     1.6828  -0.988  0.32485   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.439 on 160 degrees of freedom
Multiple R-squared:  0.2385,	Adjusted R-squared:  0.1956 
F-statistic: 5.567 on 9 and 160 DF,  p-value: 1.136e-06
```



```
Call:
lm(formula = Global_Sales ~ Platform, data = nintendoRecent.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.9000 -1.4479 -0.9964  0.0872 13.4000 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   1.90793    0.30221   6.313  2.4e-09 ***
PlatformDS    0.01207    0.68249   0.018    0.986    
PlatformWii  -0.18257    0.59900  -0.305    0.761    
PlatformWiiU -0.46043    0.52779  -0.872    0.384    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.737 on 166 degrees of freedom
Multiple R-squared:  0.004983,	Adjusted R-squared:  -0.013 
F-statistic: 0.2771 on 3 and 166 DF,  p-value: 0.8419
```



```
Call:
lm(formula = Global_Sales ~ Genre + Franchise + Platform, data = nintendoR

Residuals:
    Min      1Q  Median      3Q     Max 
-4.0472 -1.0803 -0.2008  0.5009 10.2228 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)             3.34088    1.47223   2.269 0.024718 *  
GenreAdventure          0.04394    1.09687   0.040 0.968101    
GenreFighting          -0.16387    2.24655  -0.073 0.941952    
GenreMisc              -0.13637    0.60997  -0.224 0.823400    
GenrePlatform           2.88917    0.74764   3.864 0.000167 ***
GenrePuzzle             0.24368    0.83346   0.292 0.770417    
GenreRacing             7.84825    1.70342   4.607 8.82e-06 ***
GenreRole-Playing       1.12913    0.56742   1.990 0.048467 *  
GenreShooter            1.94890    1.34596   1.448 0.149772    
GenreSimulation         1.57145    0.91709   1.714 0.088741 .  
GenreSports            -0.92577    0.93758  -0.987 0.325080    
GenreStrategy           0.17639    1.18589   0.149 0.881965    
FranchiseDonkey Kong   -2.64972    1.74324  -1.520 0.130674    
FranchiseKirby         -4.18787    1.71703  -2.439 0.015927 *  
FranchiseMario         -1.33455    1.50216  -0.888 0.375776    
FranchiseOther         -2.97701    1.39758  -2.130 0.034839 *  
FranchisePokemon        0.59905    1.57597   0.380 0.704414    
FranchiseSmash Bros     3.32757    3.03784   1.095 0.275157    
FranchiseWii Exclusive -1.41722    1.59949  -0.886 0.377050    
FranchiseYoshi         -4.63547    2.19045  -2.116 0.036023 *  
FranchiseZelda         -0.72036    1.67371  -0.430 0.667541    
PlatformDS              0.02813    0.59276   0.047 0.962220    
PlatformWii             0.24470    0.52674   0.465 0.642939    
PlatformWiiU           -0.53916    0.46349  -1.163 0.246626    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.202 on 146 degrees of freedom
Multiple R-squared:  0.4336,	Adjusted R-squared:  0.3444 
F-statistic:  4.86 on 23 and 146 DF,  p-value: 1.213e-09
```

From the global simple models and their p-values, it is observed that genres of fighting, platforming, racing, and role-playing are statistically significant. Platform has no significance, but the “other” franchise does.

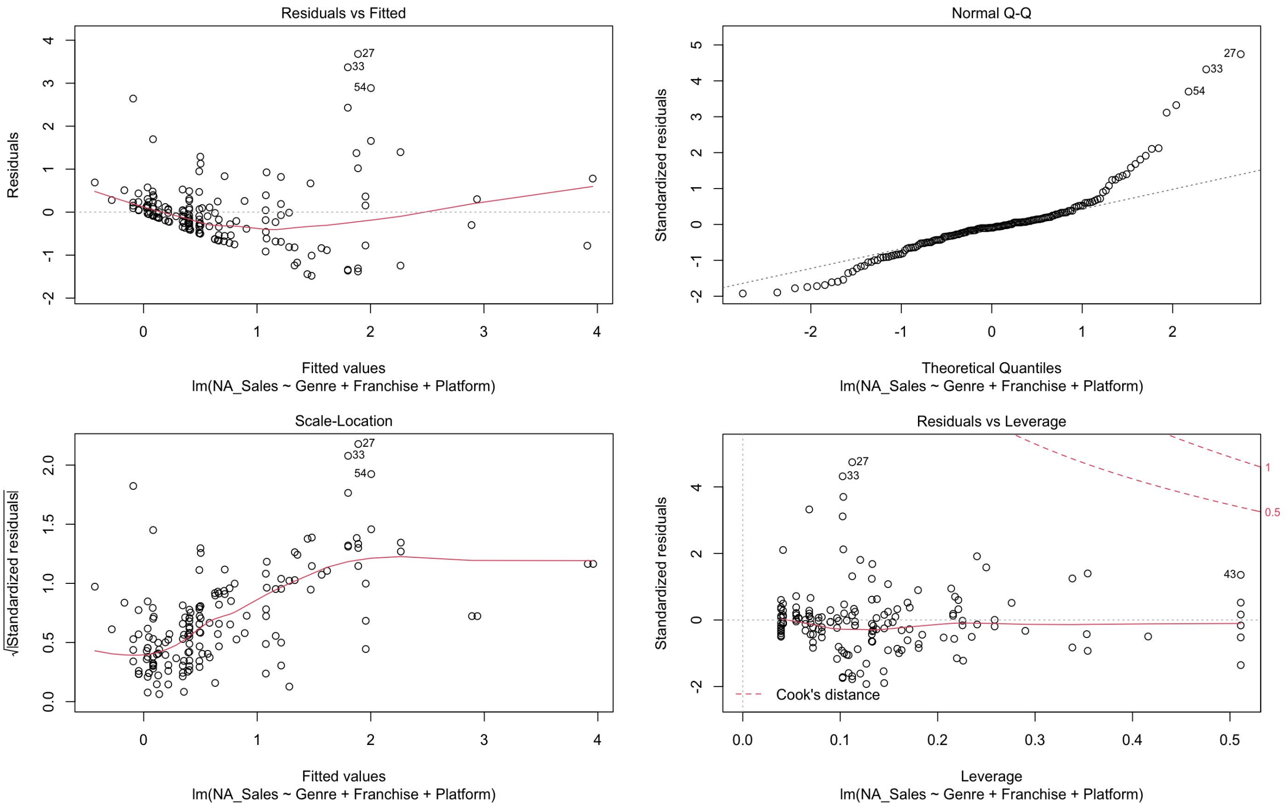
From the multiple regression model, platform, racing, and role-playing are statistically significant. Franchises Kirby, “other,” and Wii exclusives are also significant, and no platforms are significant.

Evaluating the Models:

To evaluate the fit of the multiple regressions model for each region, diagnostic plots were made.

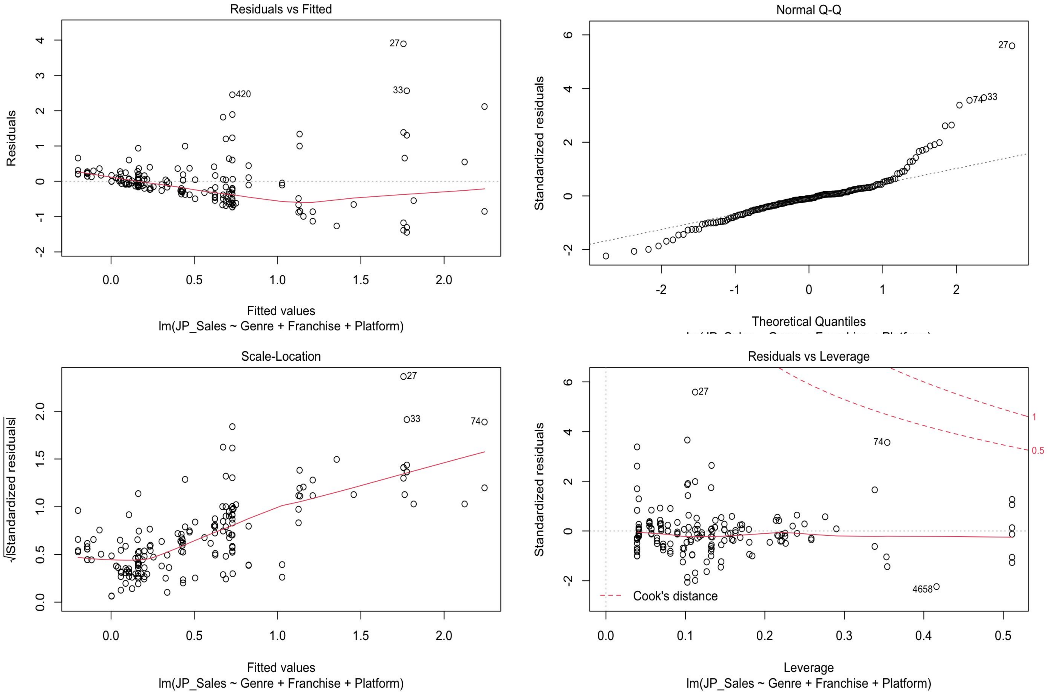
North America:

```
par(mcfrow = c(2,2))
plot(nintendoRecentNA.lm)
```



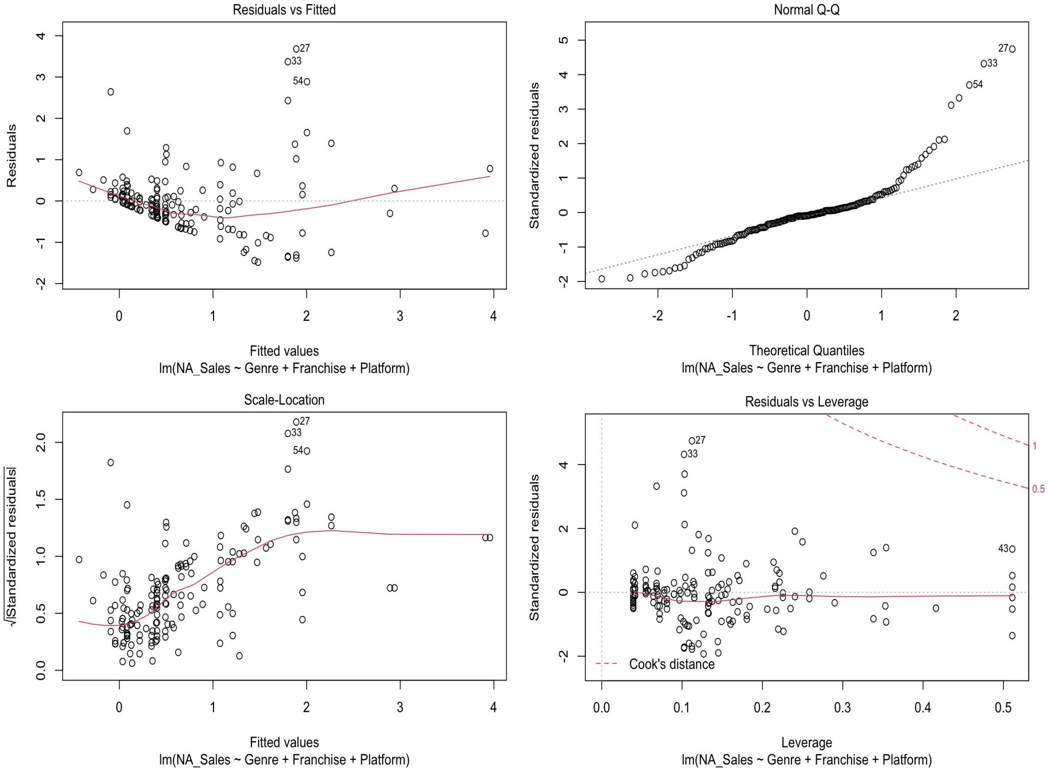
Japan:

```
par(mcfrow = c(2,2))
plot(nintendoRecentJP.lm))
```



Global:

```
par(mcfrow = c(2,2))
plot(nintendoRecent.lm)
```



According to these diagnostic plots, all three regression models are not of the highest fit. However, the underlying cause is known and therefore, can be taken into account when further accessing predictions.

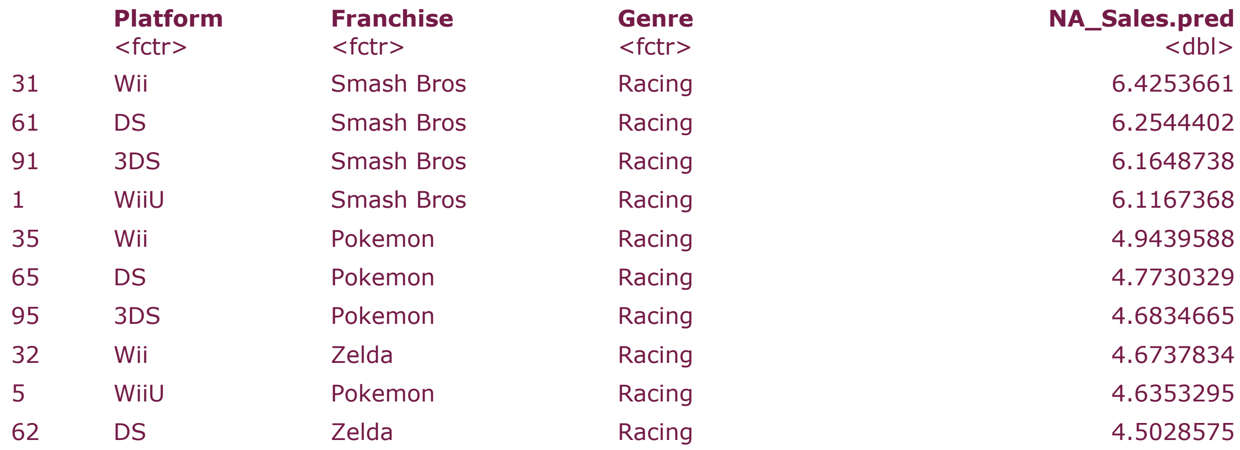
Predicting Sales from Models

To predict what combination of attributes an ideal Nintendo game during the 2010s would be, multiple data frames containing different combinations of the attributes were fed into the predictions. The chosen attributes were based on the highest and interesting spread of sales from each region as seen in the box plots in the exploratory analysis.

North America:

Genre: racing, fighting, platform, shooter, simulation, role-playing
Platform: WiiU, Wii, DS, 3 DS
Franchise: Smash Bros., Zelda, Mario, Donkey Kong, Pokemon

```
NAplatform = c(rep(as.factor(c("WiiU")), 30), rep(as.factor(c("Wii")), 30),
rep(as.factor(c("DS")), 30),rep(as.factor(c("3DS")), 30))

NAgenre = as.factor(rep(c(rep("Racing", 5), rep("Fighting", 5), rep("Platform", 5),
rep("Shooter", 5),rep("Simulation", 5), rep("Role-Playing", 5)), 4))

NAfranchise = as.factor(rep(c("Smash Bros", "Zelda", "Mario", "Donkey Kong", "Pokemon"), 24))

newNAdf = data.frame(Platform = NAplatform, Franchise = NAfranchise, Genre = NAgenre)
newNAdf = cbind(newNAdf, NA_Sales.pred = predict(nintendoRecentNA.lm, newNAdf))
arrange(newNAdf, -NA_Sales.pred)
```



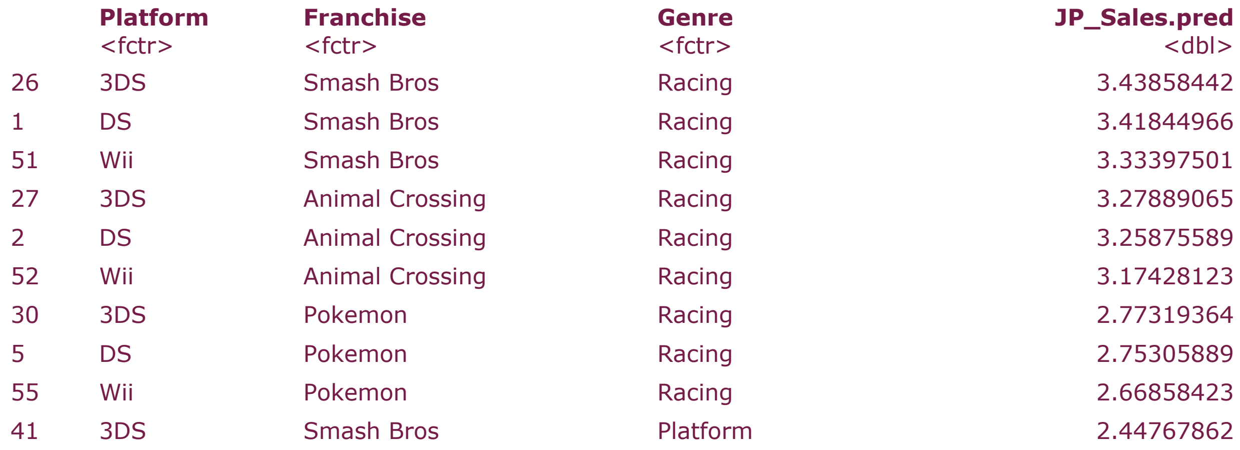
```
     Platform    Franchise    Genre        NA_Sales.pred
     <fctr>      <fctr>       <fctr>               <dbl>
31   Wii         Smash Bros   Racing           6.4253661
61   DS          Smash Bros   Racing           6.2544402
91   3DS         Smash Bros   Racing           6.1648738
1    WiiU        Smash Bros   Racing           6.1167368
35   Wii         Pokemon      Racing           4.9439588
65   DS          Pokemon      Racing           4.7730329
95   3DS         Pokemon      Racing           4.6834665
32   Wii         Zelda        Racing           4.6737834
5    WiiU        Pokemon      Racing           4.6353295
62   DS          Zelda        Racing           4.5028575
```

Based on this model, the ideal Nintendo game during the 2010s for North America is…
Wii racing game in Smash Bros. franchise.

Japan:

Genre: racing, fighting, simulation, platform, role-playing
Platform: DS, 3 DS, Wii
Franchise: Smash Bros, Animal Crossing, Mario, Kirby, Pokemon

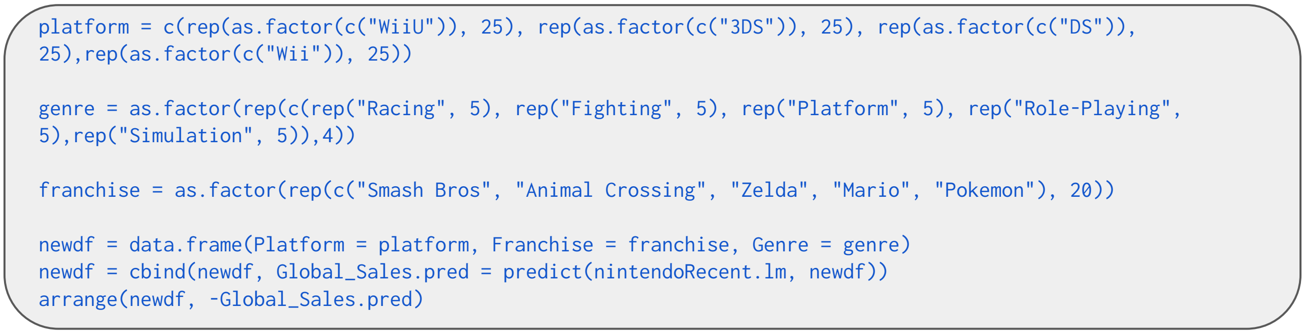
```
JPplatform = c(rep(as.factor(c("DS")), 25), rep(as.factor(c("3DS")), 25),
rep(as.factor(c("Wii")), 25))

JPgenre = as.factor(rep(c(rep("Racing", 5), rep("Fighting", 5), rep("Simulation", 5),
rep("Platform", 5),rep("Role-Playing", 5)), 3))

JPfranchise = as.factor(rep(c("Smash Bros", "Animal Crossing", "Mario", "Kirby", "Pokemon"),
15))

newJPdf = data.frame(Platform = JPplatform, Franchise = JPfranchise, Genre = JPgenre)
newJPdf = cbind(newJPdf, JP_Sales.pred = predict(nintendoRecentJP.lm, newJPdf))
arrange(newJPdf, -JP_Sales.pred)
```



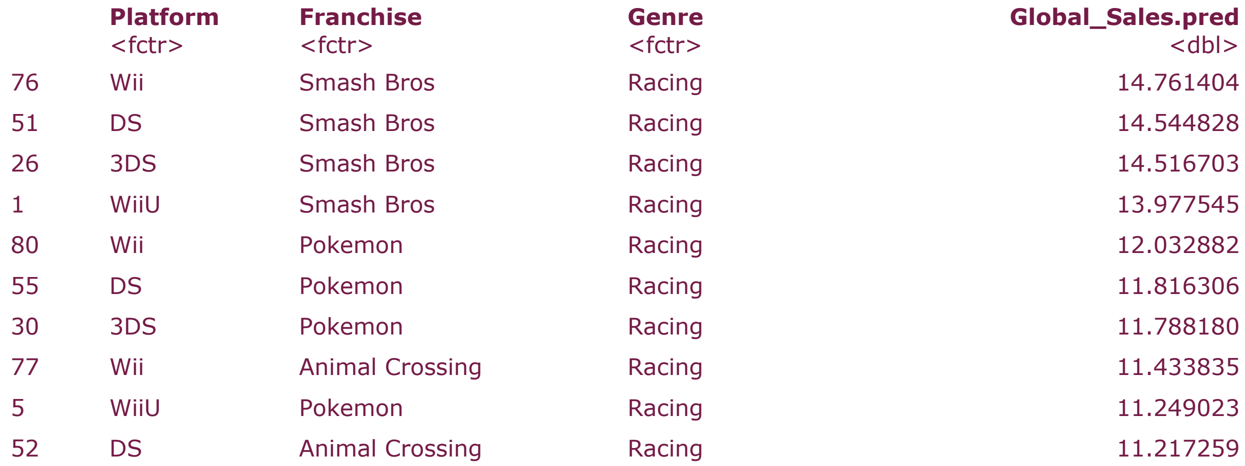
```
     Platform   Franchise         Genre          JP_Sales.pred
     <fctr>     <fctr>            <fctr>                 <dbl>
26   3DS        Smash Bros        Racing            3.43858442
1    DS         Smash Bros        Racing            3.41844966
51   Wii        Smash Bros        Racing            3.33397501
27   3DS        Animal Crossing   Racing            3.27889065
2    DS         Animal Crossing   Racing            3.25875589
52   Wii        Animal Crossing   Racing            3.17428123
30   3DS        Pokemon           Racing            2.77319364
5    DS         Pokemon           Racing            2.75305889
55   Wii        Pokemon           Racing            2.66858423
41   3DS        Smash Bros        Platform          2.44767862
```

Based on this model, the ideal Nintendo game during the 2010s for Japan is…
3 DS racing game in Smash Bros. franchise.

Global:
Genre: racing, fighting, platform, role-playing, simulation
Platform: WiiU, 3DS, DS, Wii
Franchise: Smash Bros., Animal Crossing, Zelda, Mario, Pokemon



```
platform = c(rep(as.factor(c("WiiU")), 25), rep(as.factor(c("3DS")), 25), rep(as.factor(c("DS")),
25),rep(as.factor(c("Wii")), 25))

genre = as.factor(rep(c(rep("Racing", 5), rep("Fighting", 5), rep("Platform", 5), rep("Role-Playing",
5),rep("Simulation", 5)),4))

franchise = as.factor(rep(c("Smash Bros", "Animal Crossing", "Zelda", "Mario", "Pokemon"), 20))

newdf = data.frame(Platform = platform, Franchise = franchise, Genre = genre)
newdf = cbind(newdf, Global_Sales.pred = predict(nintendoRecent.lm, newdf))
arrange(newdf, -Global_Sales.pred)
```



```
     Platform   Franchise         Genre          Global_Sales.pred
     <fctr>     <fctr>            <fctr>                     <dbl>
76   Wii        Smash Bros        Racing                 14.761404
51   DS         Smash Bros        Racing                 14.544828
26   3DS        Smash Bros        Racing                 14.516703
1    WiiU       Smash Bros        Racing                 13.977545
80   Wii        Pokemon           Racing                 12.032882
55   DS         Pokemon           Racing                 11.816306
30   3DS        Pokemon           Racing                 11.788180
77   Wii        Animal Crossing   Racing                 11.433835
5    WiiU       Pokemon           Racing                 11.249023
52   DS         Animal Crossing   Racing                 11.217259
```

Based on this model, the ideal Nintendo game during the 2010s for the globe is…
Wii racing game in Smash Bros. franchise.

Results

Across all three regions (North America, Japan, Global), racing games in the Smash Bros. franchise appear to contribute to the highest sales in their respective regions. Platform had a minimal impact on sales as one can infer that games are typically only released on the latest console. Differences in gaming taste between North America and Japan are small, but Japan appears to purchase games in the Animal Crossing franchise more than North America.

However, the models used to determine these results are flawed due to uneven cells and collinearity. Nonetheless, the results are still supported by the box plots of each input variable as seen in the exploratory analysis section.

Future Analysis

The original question in mind for this analysis asked what the ideal Nintendo game would be. However, with the outdated data set given, the question was restricted to ask what the ideal Nintendo game would be in the 2010s. Therefore, I would like to continue to answer this question with more accuracy given a fully updated data set that includes games released for the Nintendo Switch.

Additionally, I believe more aspects of a game would accurately reflect what factors are more attractive to sales as genre and franchise can only convey so much. If columns included whether a game has multiplayer, solo, or both, whether it is linear gameplay or is open world, and what its intended age range is, then I believe that I could gather more useful results.

Conclusion

This analysis found that across all three regions (North America, Japan, Global), racing games in the Smash Bros. franchise seem to contribute to the highest sales while the platform had little impact. However, the models used to find these results did not have as sufficient of a fit as one would hope due to uneven cell sizes and collinearity.

Although the results of this analysis are not as strong as initially hoped, the success of recent games released by Nintendo offer alternative support. According to Nintendo's corporate website, the top selling Nintendo game as of September 2022 is Mario Kart 8 Deluxe with 48.41 million copies sold. The second most selling game is Animal Crossing: New Horizons with 40.17 million copies sold. Finally, the third most selling game is Super Smash Bros. Ultimate with 29.53 million copies sold.

Mario Kart is a racing game with a large array of Nintendo characters across franchises. This character variety is a large commonality with the Super Smash Bros. franchise and may be the factor that entices players. Couple these aspects with a multiplayer and party atmosphere, and this may be another variable that contributes to the success of Nintendo games. While Animal Crossing: New Horizons has little similarities between these two games, I predict it is the release date that drove the sales as this game was released March 20, 2020 at the early onset of the Pandemic. With isolation just beginning, a relaxing solo-player simulator game seemed to appeal the most to the public.

Upon reviewing the success of the newer games, future analysis can continue to answer the original research question of what the ideal Nintendo game would be. Additions to data could include updated games and consoles, and more attributes such as the number of players, the linearity of gameplay, and intended audience. These new factors and new games could help to reduce some of the problems this analysis faced by adding more to the sample size and reducing collinearity.