Stat 4620 Final Project

Austin Usher

2024-11-20

Load Data

```
path = "C:/Users/austi/OneDrive/Stat 4620/Final Project/train.csv"

data = read.csv(path)

head(data)
```

```
##   X                     Title Features.Handheld. Features.Max.Players
## 1 1          Super Mario 64 DS               True                   1
## 2 2     Lumines: Puzzle Fusion               True                   1
## 3 3         WarioWare Touched!               True                   2
## 4 4   Hot Shots Golf: Open Tee               True                   1
## 5 5               Spider-Man 2               True                   1
## 6 6 The Urbz: Sims in the City               True                   1
##   Features.Multiplatform. Features.Online.                Metadata.Genres
## 1                    True             True                         Action
## 2                    True             True                       Strategy
## 3                    True             True Action,Racing / Driving,Sports
## 4                    True             True                         Sports
## 5                    True             True                         Action
## 6                    True             True                     Simulation
##   Metadata.Licensed. Metadata.Publishers Metadata.Sequel. Metrics.Review.Score
## 1               True            Nintendo             True                   85
## 2               True             Ubisoft             True                   89
## 3               True            Nintendo             True                   81
## 4               True                Sony             True                   81
## 5               True          Activision             True                   61
## 6               True                  EA             True                   67
##   Metrics.Sales Metrics.Used.Price Release.Console Release.Rating
## 1          4.69              24.95     Nintendo DS              E
## 2          0.56              14.95        Sony PSP              E
## 3          0.54              22.95     Nintendo DS              E
## 4          0.49              12.95        Sony PSP              E
## 5          0.45              14.95     Nintendo DS              E
## 6          0.41              12.95     Nintendo DS              M
##   Release.Re.release. Release.Year Length.All.PlayStyles.Average
## 1               True         2004                     22.716667
## 2               True         2004                     10.100000
## 3               True         2004                      4.566667
## 4               True         2004                      0.000000
```

```
## 5                True         2004                     13.250000
## 6                True         2004                     21.933333
##   Length.All.PlayStyles.Leisure Length.All.PlayStyles.Median
## 1                      31.90000                      24.48333
## 2                      11.01667                      10.00000
## 3                      11.56667                       2.50000
## 4                       0.00000                       0.00000
## 5                      48.38333                      10.00000
## 6                      25.50000                      20.00000
##   Length.All.PlayStyles.Polled Length.All.PlayStyles.Rushed
## 1                           57                    14.300000
## 2                            5                     9.516667
## 3                           57                     2.266667
## 4                            0                     0.000000
## 5                           37                     7.066667
## 6                            7                    16.733333
##   Length.Completionists.Average Length.Completionists.Leisure
## 1                      29.76667                       35.03333
## 2                       0.00000                        0.00000
## 3                      10.00000                       14.10000
## 4                       0.00000                        0.00000
## 5                      72.56667                       78.86667
## 6                      30.03333                       30.03333
##   Length.Completionists.Median Length.Completionists.Polled
## 1                     30.00000                           20
## 2                      0.00000                            0
## 3                      7.25000                           16
## 4                      0.00000                            0
## 5                     72.56667                            2
## 6                     30.03333                            2
##   Length.Completionists.Rushed Length.Main...Extras.Average
## 1                     22.01667                      24.91667
## 2                      0.00000                       9.75000
## 3                      6.80000                       3.85000
## 4                      0.00000                       0.00000
## 5                     66.28333                      12.76667
## 6                     30.03333                      20.83333
##   Length.Main...Extras.Leisure Length.Main...Extras.Median
## 1                    29.966667                    25.000000
## 2                     9.866667                     9.750000
## 3                     5.666667                     3.333333
## 4                     0.000000                     0.000000
## 5                    17.316667                    12.500000
## 6                    25.200000                    20.000000
##   Length.Main...Extras.Polled Length.Main...Extras.Rushed
## 1                          16                   18.333333
## 2                           2                    9.616667
## 3                          11                    2.783333
## 4                           0                    0.000000
## 5                          12                   10.483333
## 6                           3                   16.450000
##   Length.Main.Story.Average Length.Main.Story.Leisure Length.Main.Story.Median
## 1                 14.333333                 18.316667                14.500000
## 2                 10.333333                 11.083333                10.000000
```

```
## 3                  1.916667                   2.933333                   1.833333
## 4                  0.000000                   0.000000                   0.000000
## 5                  8.350000                  11.083333                   8.000000
## 6                 15.500000                  15.750000                  15.500000
##   Length.Main.Story.Polled Length.Main.Story.Rushed
## 1                       21                 9.700000
## 2                        3                 9.583333
## 3                       30                 1.433333
## 4                        0                 0.000000
## 5                       23                 5.333333
## 6                        2                15.250000
```

Does EDA suggest modeling approached to try

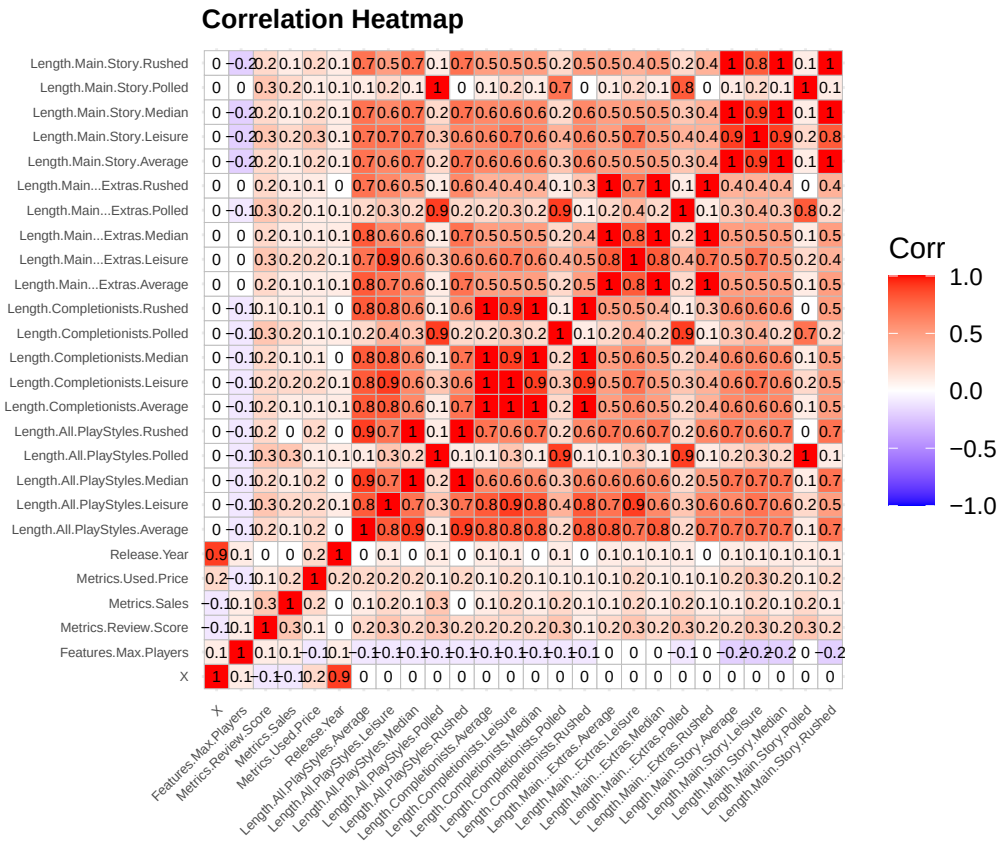
```
# Correlation heatmap for numeric variables
library(ggcorrplot)
```

```
## Loading required package: ggplot2
```

```
numeric_data <- data[sapply(data, is.numeric)]
cor_matrix <- cor(numeric_data, use = "complete.obs")

cor_matrix = round(cor_matrix,1)

# Plot heatmap with larger size and improved readability
ggcorrplot(cor_matrix,
           lab = TRUE,
           lab_size = 2,
           tl.cex = 5,
           title = "Correlation Heatmap",
           ggtheme = ggplot2::theme_minimal()) +
  theme(
    plot.title = element_text(size = 10, face = "bold"),
    axis.text = element_text(size = 4)
  )
```




```
# Extract correlations of 'Metrics.Sales' with all other variables
sales_correlations <- cor_matrix["Metrics.Sales", ]

# Print the correlation values
print(sales_correlations)
```

```
##                                X          Features.Max.Players
##                             -0.1                          0.1
##           Metrics.Review.Score                Metrics.Sales
##                             0.3                          1.0
##             Metrics.Used.Price                 Release.Year
##                             0.2                          0.0
## Length.All.PlayStyles.Average Length.All.PlayStyles.Leisure
##                             0.1                          0.2
##  Length.All.PlayStyles.Median  Length.All.PlayStyles.Polled
##                             0.1                          0.3
##  Length.All.PlayStyles.Rushed Length.Completionists.Average
##                             0.0                          0.1
## Length.Completionists.Leisure  Length.Completionists.Median
##                             0.2                          0.1
##  Length.Completionists.Polled  Length.Completionists.Rushed
##                             0.2                          0.1
##  Length.Main...Extras.Average  Length.Main...Extras.Leisure
##                             0.1                          0.2
##   Length.Main...Extras.Median   Length.Main...Extras.Polled
```

```
##                            0.1                            0.2
##   Length.Main...Extras.Rushed    Length.Main.Story.Average
##                            0.1                            0.1
##     Length.Main.Story.Leisure     Length.Main.Story.Median
##                            0.2                            0.1
##      Length.Main.Story.Polled     Length.Main.Story.Rushed
##                            0.2                            0.1
```

Summary:

Because many variables have very small correlation with the Sales variable both Lasso and Ridge would be useful to diminish the impact of non important variables. Lasso would work best to selected the needed features and to simplify the model as some coefficients are set to 0. Ridge does not to feature selection so would not be as ideal. Also there does not seem to be large multicollinearity between variables so PLS would not be suitable.

Kind of variables in data set

```
# Check the structure of the data
str(data)
```

```
## 'data.frame':    728 obs. of  37 variables:
##  $ X                            : int  1 2 3 4 5 6 7 9 11 12 ...
##  $ Title                        : chr  "Super Mario 64 DS" "Lumines: Puzzle Fusion" "WarioWare Touch
##  $ Features.Handheld.           : chr  "True" "True" "True" "True" ...
##  $ Features.Max.Players         : int  1 1 2 1 1 1 1 1 1 1 ...
##  $ Features.Multiplatform.      : chr  "True" "True" "True" "True" ...
##  $ Features.Online.             : chr  "True" "True" "True" "True" ...
##  $ Metadata.Genres              : chr  "Action" "Strategy" "Action,Racing / Driving,Sports" "Sports"
##  $ Metadata.Licensed.           : chr  "True" "True" "True" "True" ...
##  $ Metadata.Publishers          : chr  "Nintendo" "Ubisoft" "Nintendo" "Sony" ...
##  $ Metadata.Sequel.             : chr  "True" "True" "True" "True" ...
##  $ Metrics.Review.Score         : int  85 89 81 81 61 67 88 68 62 75 ...
##  $ Metrics.Sales                : num  4.69 0.56 0.54 0.49 0.45 0.41 0.36 0.25 0.2 0.16 ...
##  $ Metrics.Used.Price           : num  24.9 14.9 22.9 12.9 14.9 ...
##  $ Release.Console              : chr  "Nintendo DS" "Sony PSP" "Nintendo DS" "Sony PSP" ...
##  $ Release.Rating               : chr  "E" "E" "E" "E" ...
##  $ Release.Re.release.          : chr  "True" "True" "True" "True" ...
##  $ Release.Year                 : int  2004 2004 2004 2004 2004 2004 2004 2004 2004 2004 ...
##  $ Length.All.PlayStyles.Average: num  22.72 10.1 4.57 0 13.25 ...
##  $ Length.All.PlayStyles.Leisure: num  31.9 11 11.6 0 48.4 ...
##  $ Length.All.PlayStyles.Median : num  24.5 10 2.5 0 10 ...
##  $ Length.All.PlayStyles.Polled : int  57 5 57 0 37 7 6 0 4 9 ...
##  $ Length.All.PlayStyles.Rushed : num  14.3 9.52 2.27 0 7.07 ...
##  $ Length.Completionists.Average: num  29.8 0 10 0 72.6 ...
##  $ Length.Completionists.Leisure: num  35 0 14.1 0 78.9 ...
##  $ Length.Completionists.Median : num  30 0 7.25 0 72.57 ...
##  $ Length.Completionists.Polled : int  20 0 16 0 2 2 1 0 0 0 ...
##  $ Length.Completionists.Rushed : num  22 0 6.8 0 66.3 ...
##  $ Length.Main...Extras.Average : num  24.92 9.75 3.85 0 12.77 ...
##  $ Length.Main...Extras.Leisure : num  29.97 9.87 5.67 0 17.32 ...
##  $ Length.Main...Extras.Median  : num  25 9.75 3.33 0 12.5 ...
##  $ Length.Main...Extras.Polled  : int  16 2 11 0 12 3 2 0 0 2 ...
##  $ Length.Main...Extras.Rushed  : num  18.33 9.62 2.78 0 10.48 ...
##  $ Length.Main.Story.Average    : num  14.33 10.33 1.92 0 8.35 ...
```

```
##  $ Length.Main.Story.Leisure    : num  18.32 11.08 2.93 0 11.08 ...
##  $ Length.Main.Story.Median     : num  14.5 10 1.83 0 8 ...
##  $ Length.Main.Story.Polled     : int  21 3 30 0 23 2 3 0 4 7 ...
##  $ Length.Main.Story.Rushed     : num  9.7 9.58 1.43 0 5.33 ...
```

```r
# Numeric, Factor, and Character Variable Count Without %>%
numeric_vars <- sum(sapply(data, is.numeric))
factor_vars <- sum(sapply(data, is.factor))
character_vars <- sum(sapply(data, is.character))

# Create a summary table for variable types
var_summary <- data.frame(
  numeric_vars = numeric_vars,
  factor_vars = factor_vars,
  character_vars = character_vars
)

# Print the summary table
print(var_summary)
```

```
##   numeric_vars factor_vars character_vars
## 1           26           0             11
```

```r
# Summary for each variable in the dataset
summary(data)
```

```
##        X             Title           Features.Handheld. Features.Max.Players
##  Min.   :   1.0   Length:728         Length:728         Min.   :1.00
##  1st Qu.: 283.8   Class :character   Class :character   1st Qu.:1.00
##  Median : 571.0   Mode  :character   Mode  :character   Median :1.00
##  Mean   : 587.1                                        Mean   :1.69
##  3rd Qu.: 887.2                                        3rd Qu.:2.00
##  Max.   :1211.0                                        Max.   :8.00
##  Features.Multiplatform. Features.Online.   Metadata.Genres
##  Length:728              Length:728         Length:728
##  Class :character        Class :character   Class :character
##  Mode  :character        Mode  :character   Mode  :character
##
##
##
##  Metadata.Licensed. Metadata.Publishers Metadata.Sequel.   Metrics.Review.Score
##  Length:728         Length:728          Length:728         Min.   :19.00
##  Class :character   Class :character    Class :character   1st Qu.:60.00
##  Mode  :character   Mode  :character    Mode  :character   Median :70.00
##                                                           Mean   :68.99
##                                                           3rd Qu.:79.00
##                                                           Max.   :98.00
##  Metrics.Sales     Metrics.Used.Price Release.Console    Release.Rating
##  Min.   : 0.0100   Min.   : 4.95      Length:728         Length:728
##  1st Qu.: 0.0900   1st Qu.:14.95      Class :character   Class :character
##  Median : 0.2200   Median :17.95      Mode  :character   Mode  :character
##  Mean   : 0.5294   Mean   :17.27
##  3rd Qu.: 0.4900   3rd Qu.:17.95
```

```
##  Max.   :14.6600   Max.   :49.95
##  Release.Re.release.  Release.Year  Length.All.PlayStyles.Average
##  Length:728          Min.   :2004   Min.   :  0.000
##  Class :character    1st Qu.:2006   1st Qu.:  2.975
##  Mode  :character    Median :2007   Median :  8.667
##                      Mean   :2007   Mean   : 13.960
##                      3rd Qu.:2008   3rd Qu.: 15.908
##                      Max.   :2008   Max.   :279.733
##  Length.All.PlayStyles.Leisure Length.All.PlayStyles.Median
##  Min.   :  0.000               Min.   :  0.000
##  1st Qu.:  3.425               1st Qu.:  2.458
##  Median : 12.000               Median :  8.000
##  Mean   : 27.094               Mean   : 11.315
##  3rd Qu.: 28.617               3rd Qu.: 13.912
##  Max.   :476.267               Max.   :126.000
##  Length.All.PlayStyles.Polled Length.All.PlayStyles.Rushed
##  Min.   :   0.00              Min.   :  0.000
##  1st Qu.:   1.00              1st Qu.:  2.117
##  Median :   6.00              Median :  6.583
##  Mean   :  44.03              Mean   :  9.467
##  3rd Qu.:  25.00              3rd Qu.: 11.367
##  Max.   :1900.00              Max.   :120.200
##  Length.Completionists.Average Length.Completionists.Leisure
##  Min.   :  0.00                Min.   :  0.00
##  1st Qu.:  0.00                1st Qu.:  0.00
##  Median :  5.20                Median :  5.50
##  Mean   : 20.34                Mean   : 26.47
##  3rd Qu.: 20.71                3rd Qu.: 26.07
##  Max.   :683.13                Max.   :691.57
##  Length.Completionists.Median Length.Completionists.Polled
##  Min.   :  0.00               Min.   :  0.00
##  1st Qu.:  0.00               1st Qu.:  0.00
##  Median :  5.00               Median :  1.00
##  Mean   : 19.33               Mean   :  5.37
##  3rd Qu.: 20.00               3rd Qu.:  3.00
##  Max.   :683.13               Max.   :214.00
##  Length.Completionists.Rushed Length.Main...Extras.Average
##  Min.   :  0.000              Min.   :  0.000
##  1st Qu.:  0.000              1st Qu.:  0.000
##  Median :  4.433              Median :  7.183
##  Mean   : 16.916              Mean   : 13.100
##  3rd Qu.: 18.071              3rd Qu.: 16.233
##  Max.   :674.700              Max.   :291.000
##  Length.Main...Extras.Leisure Length.Main...Extras.Median
##  Min.   :  0.000              Min.   :  0.0
##  1st Qu.:  0.000              1st Qu.:  0.0
##  Median :  7.875              Median :  7.0
##  Mean   : 19.408              Mean   : 12.5
##  3rd Qu.: 21.617              3rd Qu.: 15.0
##  Max.   :478.933              Max.   :291.0
##  Length.Main...Extras.Polled Length.Main...Extras.Rushed
##  Min.   :  0.00              Min.   :  0.000
##  1st Qu.:  0.00              1st Qu.:  0.000
##  Median :  1.00              Median :  6.283
```

```
##  Mean   : 13.55              Mean   : 10.712
##  3rd Qu.:  5.25              3rd Qu.: 12.854
##  Max.   :630.00              Max.   :291.000
##  Length.Main.Story.Average Length.Main.Story.Leisure Length.Main.Story.Median
##  Min.   : 0.000            Min.   :  0.00            Min.   : 0.000
##  1st Qu.: 0.000            1st Qu.:  0.00            1st Qu.: 0.000
##  Median : 6.325            Median :  8.00            Median : 6.000
##  Mean   : 8.219            Mean   : 10.76            Mean   : 8.024
##  3rd Qu.:11.088            3rd Qu.: 14.56            3rd Qu.:10.550
##  Max.   :69.833            Max.   :107.95            Max.   :62.000
##  Length.Main.Story.Polled Length.Main.Story.Rushed
##  Min.   :   0.00          Min.   : 0.000
##  1st Qu.:   0.00          1st Qu.: 0.000
##  Median :   3.00          Median : 5.058
##  Mean   :  25.35          Mean   : 6.762
##  3rd Qu.:  14.00          3rd Qu.: 9.213
##  Max.   :1100.00          Max.   :60.000
```

Sales shows a right-skewed distribution indicating few games achieve very high sales while the majority have fewer sales. Review scores also display a broad variation which could be useful in predicting sales. Features such as number of players, multiplatform, sequels, and online capabilities could also be helpful in predicting sales. Some game play styles have very large outliers which could affect prediction.