THE UNIVERSITY OF SYDNEY

Explore, Clean, Define

COMP5310 Project Stage 2

Prepared by
Chenlin Du & Xin Xu

Student ID:
470053517 & 510060497
UniKey:
chdu9178 & xixu7480

Setup

Streaming platforms are expanded rapidly and receive millions of broadcasters and audiences every month (Foster, 2016). Video game streams, as one of the most popular and effective mediums, can be seen as an important marketing tool to attract viewers and sponsors. In order to continue expanding the market share, it is common for streaming platforms to buy out exclusive broadcast rights of some most welcomed video games. Therefore, it becomes crucial to successfully predict their popularity, and then to make further business decisions based on that prediction. The goal of this report is to evaluate the potential commercial value of video games by testing whether the stream performances at peak time have a significant effect on the total game popularity.

The amount of viewers and the number of channels at peak time, which are regarded as peak indices, along with the streamed hours on platform, the number of streamers are the factors to be examined to see whether they will affect the influence of a specific video game. The popularity or influence of a game can be represented by the total hours watched during a certain period of time. The null and alternative hypotheses are set as following:

- H_0: two peak indices (the amount of viewers and the number of channels at peak time) will have no significant effect on total hours watched on game streaming platforms.
- H_1: there is a significant relationship between the two peak indices and the game's total viewing time.

The multiple regression approach is implemented to support the hypothesis testing. It can use explanatory variables to predict the outcome of a response variable. Prediction can be seen as the output of an algorithm after being trained on a historical dataset when a new data is applied (Datarobot, 2021). With the aim of finding the optimal regression line, a large amount of relevant data has been fitted. R-squared value is used to measure how well the regression model can explain the observed data. After generating the r-squared value of each model in null hypothesis and alternative hypothesis, a t-test will be applied to determine the significant difference between two. The α is set as 0.05.

The chosen dataset, Top games on Twitch 2016 - 2021, is available publicly in Kaggle. 12 attributes of top 200 games on the platform are recorded monthly (Krish, 2021). Since the focus of this report is to predict the outcome of total hours watched by using explanatory variables including hours streamed, peak viewers, peak channels, streamers and average viewers. We extracted and cleaned these data in python and exported them as a csv file to R for further analysis.

Approach

The first step of conducting a multiple linear regression is to analyse the correlation and directionality of the data. The correlation between two variables are tested (Table 1) and visually represented in figure 2. The correlation between hours watched and average viewers is almost one and much higher than others, meaning the average viewers can dominantly determine the response variable. This can also be illustrated in the scatter plot between ave_viewers and hours_watched in figure 3, where points are perfectly fitted on the regression line. In order to explore other useful information, we assumed the data of average viewers will not be given. Thus, this report will only consider the amount of viewers and the number of channels at peak time, streamed hours on platform and the number of streamers as independent variables.

To examine whether two peak indices will affect the influence of a specific video game, two regression models have been built in the following, one of which is complete model and the other one is reduced model(benchmark model).

- Complete model: Hours_watched = $\beta_0 + \beta_1$*hours_streamed + β_2*peak_viewers + β_3*peak_channels + β_4*streamers
- Reduced model: Hours_watched = $\beta_0 + \beta_1$*hours_streamed + β_4*streamers

Two multiple regression models will be effectively evaluated by means of numerous criterias. A component-plus-residual plot is built to attempt showing the relationship between the response variable and the four explanatory variables respectively given that other independent variables are in the model. As shown in figure 4, although there are some outliers, the scatter plot of hours_stremed and streamers show a linear relationship to the total hours of viewing. However, linear relationships to the dependent variable are still observed in the peak_viewers and peak_channels scatter plot, but dots are concentrated in a small range of values and do not spread along the line. This points out the weak linear relations. The gap between error line and prediction line in both plots are a lot wider than the other two, indicating a higher fitting deviation.

With the aim of providing specific details about our predicted model, a model summary table can be implemented to report the strength of the relationship between the independent and dependent variables. The first step in interpreting the multiple regression analysis is to examine the F-statistic and its associated p-value. From our summary table 1, there is a highly significant p-value of the F-statistic (less than 2.2e-16). This means at least one of the predictor variables is significantly related to the outcome variable. For a given predictor, the t-statistic evaluates whether or not there is a strong relationship between the predictor and the outcome dependent variable -- total hours watched. From table 1, it can be found that the p-value of the T-statistic is < 2.2e-16 for each predictor, except the p-value of 0.00083 Streamers which is still smaller than 0.001. Therefore, all predictors are significantly associated with total hours of watched.

The estimated coefficients can be interpreted as the average effect on y of a one unit increase in one predictor, holding all other predictors fixed. For example, increasing an additional 100 viewers at the peak time leads to an increase in total hours of watch by approximately 61*100 = 6100 hours on average. The R-squared represents the correlation coefficient between the observed values of the outcome variable (y) and the fitted (i.e., predicted) values of y. R-squared value of closing to 1 indicates that the model explains a large portion of the variance in the outcome variable. The R-squared is 0.7442, meaning that 74% of the variance in the measure of hours_watched can be predicted by hours streamed, peak viewers, peak channels, streamers and average viewers. Overall, the evaluation result for our predicted model is acceptable as each predictor shows a closed relationship to the dependent variable.

```
Call:
lm(formula = Hours_watched ~ Hours_Streamed + Peak_viewers +
    Peak_channels + Streamers, data = data)

Residuals:
       Min         1Q     Median         3Q        Max
-170942415    -792046     181250     671247  174590856

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -8.008e+05  7.375e+04 -10.859  < 2e-16 ***
Hours_Streamed  1.934e+01  5.144e-01  37.592  < 2e-16 ***
Peak_viewers    6.061e+01  7.231e-01  83.811  < 2e-16 ***
Peak_channels  -1.769e+03  4.502e+01 -39.290  < 2e-16 ***
Streamers       1.572e+01  4.701e+00   3.343  0.00083 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7621000 on 12745 degrees of freedom
Multiple R-squared:  0.7442,    Adjusted R-squared:  0.7442
F-statistic:  9271 on 4 and 12745 DF,  p-value: < 2.2e-16
```

Table 1.

```
Call:
lm(formula = Hours_watched ~ Hours_Streamed + Streamers, data = data)

Residuals:
      Min        1Q    Median        3Q       Max
-99900648  -1347832   -904526   -493539 252874106

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    9.857e+05  8.907e+04   11.07   <2e-16 ***
Hours_Streamed 1.429e+01  5.974e-01   23.93   <2e-16 ***
Streamers      7.851e+01  5.856e+00   13.41   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9612000 on 12747 degrees of freedom
Multiple R-squared:  0.5931,    Adjusted R-squared:  0.5931
F-statistic:  9291 on 2 and 12747 DF,  p-value: < 2.2e-16
```

Table 2.

As there are still some fitting errors for two peak indices (peak_channels, pel_viewers), the following research is to conduct the accuracy comparison between our predicted model and reduced model (benchmark model). Summary tables for both models will be firstly analysed. The table 2 demonstrates the regression analysis for reduced model, in which the p-value of the T-statistic is < 2.2e-16 for each predictor, showing the number of Streamed hours and Streamers significantly influence the total hours of watch. Though both p-value of F-statistic of the predicted model and reduced model close to 0, the predicted model has a larger R-squared value (0.5931) than the predicted model (0.7442). It means that the predicted model has a better goodness of fit than the reduced model.

The interaction term is a subset of our predictor variables, of which we applied a partial F-test to understand the significance. This is essentially the same as a T-test because our subsets are the two predictors in the benchmark model. An Analysis of Variance (ANOVA) table has been applied, as shown in table 3, to test the difference between the means of two groups, where model 1 is the complete model and model 2 removes two peak indices. The p-value of partial F-statistic is < 2.2e-16, indicating the two peak indices can not be ignored as these two are significantly associated with total hours of watched.

The linear hypothesis test is also useful for determining the significance of two peak indices. From table 4, the hypothesis set is the same as our null hypothesis that two peak indices (the amount of viewers and the number of channels at peak time) will have no significant effect on total hours watched. The null hypothesis is rejected because the result of p-value is < 2.2e-16, which is the same as the partial F-test from the previous table.

Result

From the above analyses, we can conclude that our null hypothesis is rejected, and the alternative hypothesis is true as there is a significant relationship between the two peak indices and the game's total viewing time.

There are many types of criteria that we can use to determine whether or not one model of a set of data is better than another model of that same set of data. In order to find an optimal linear regression model to give future analysis about the relationship between total hours watch and several predictors, we used stepAIC() function to choose the best multiple linear regression model between the complete model and the reduced model. The scope argument in stepAIC allows you to specify the range of variables that will be included in the model. The model with the fewest variables is the lower model, and the model with the most variables is the upper model. Both the upper and lower scope components can be explicitly specified. The best model produced from stepAIC is the multiple linear regression model where

Hours_Streamed + Streamers + Peak_viewers + Peak_channels are independent variables to the Hours_watched. This model is just as the same model as we predicted before. Therefore, this result once again proves the importance of these peak indices for linear regression model fitting optimization. It also shows that we should not ignore these two factors, because they are significantly correlated with the dependent variable.

As one of the most populated games on this platform, the statistical data of Fortnite also demonstrates this trend. A clear correlation between total hours watched and peak viewers for Fortnite can be seen in figure 5. Two variables follow a similar fluctuation, with some sharp peaks, showing a high correlation between hours viewing and peak viewers.

Nevertheless, limitations still exist in this model. The hypothesis testing and regression models are conducted under the assumption that the data of the average viewers is not given. The fact that the acquisition of this data is straightforward for streaming platform companies -- real-time monitoring the active members. Since the average viewers can dominantly determine the response variable as previously mentioned, the practical value of the complete model has been reduced hugely. Despite this fact, two peak data are still important indices when two games are similar in the amount of average viewers.

The quantile-quantile plot, a good graphical tool assisting us to assess if a set of data come from the same theoretical distribution, is applied in this study. Points from two sets of quantiles with a common distribution should form a straight line along the reference line. As shown in figure 6, points in lower and upper quantiles of the plot do not fall on the line. It illustrates that our residuals do not come from a normal distribution. To conclude, our data distribution is relatively unconcentrated, some outliers are considered to take negative effects for fitting our linear regression model.

Conclusion

It was the first data analytical study we had come across and this study really enhanced the understanding of both of us regarding data analysis on a specific topic. Collecting useful data from numerous information resources, cleaning raw data and applying statistical approaches are the skills we have learned through this study. We also understood the collected data might be unordered and really hard to generate any valuable information. Outcomes of derived data may not perform as expected.

In terms of recommending the complete model as a solution to whether the stream performances at peak time have a significant effect on the total game popularity, it depends on the information we have. If the data regarding average viewers is given, we can just simply use this data as a simple explanatory variable to predict the hours watched instead of using the complete model. The statistics indicate they are highly correlated and points are perfectly concentrated along the regression line in the scatter plot. Under the assumption where the average viewers is not given or two games are similar in the amount of average viewers, the complete model can be implemented as the solution. The above statistical analyses have proved the importance of the data regarding two peak indices for linear regression model fitting optimization. It is sufficient to show that the two peak variables are significantly correlated to the hours watched when the average viewers is not given.

Reference

Datarobot. (2021). *What does Prediction mean in Machine Learning?* https://www.datarobot.com/wiki/prediction/#:~:text=What%20does%20Prediction%20mean%20in,will%20churn%20in%2030%20days.

Foster, L. B. (2016). Effects of Video Game Streaming on Consumer Attitudes and Behaviors. *East Tennessee State University*. https://dc.etsu.edu/cgi/viewcontent.cgi?article=4451&context=etd

Krish.R. (2021). *Top games on Twitch 2016-2021* [data set]. Kaggle. https://www.kaggle.com/rankirsh/evolution-of-top-games-on-twitch

Appendix

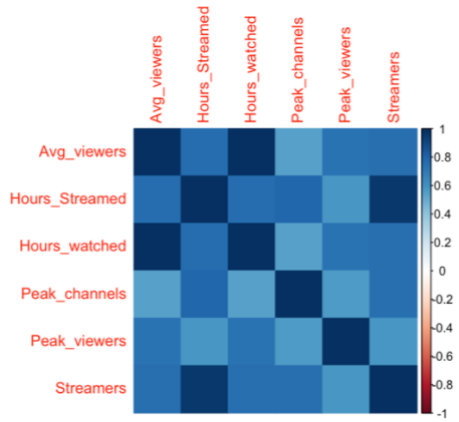

Figure 2: correlation matrix

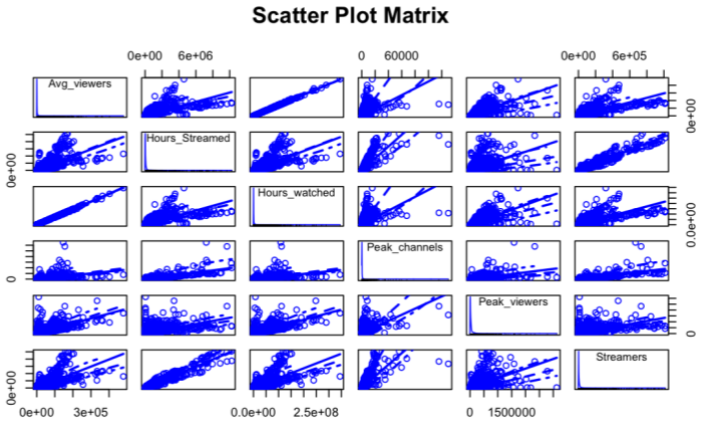

Figure 3: scatter plot matrix

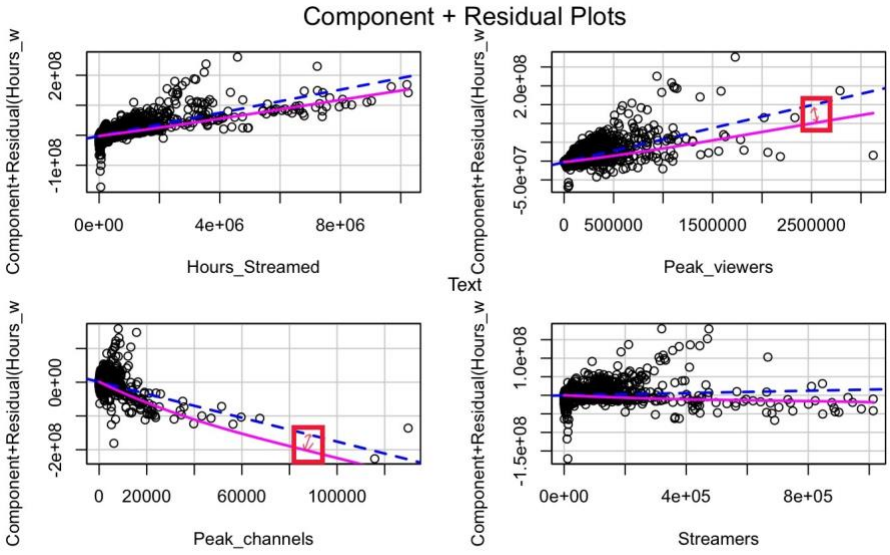

Figure 4. Model1 Component+Residual plots.

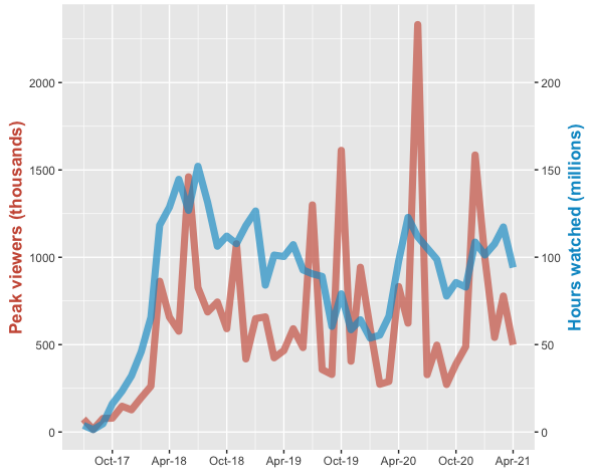

Figure 5

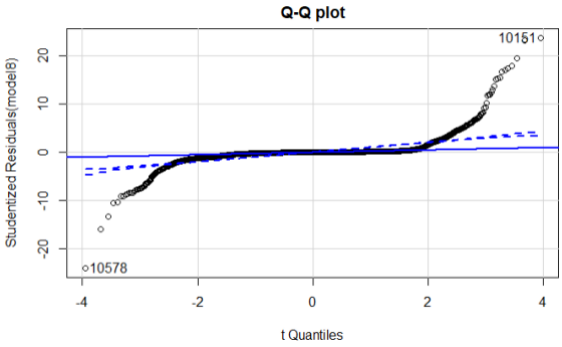

Figure 6

```
> cor(data)
```

	Avg_viewers	Hours_Streamed	Hours_watched	Peak_channels	Peak_viewers	Streamers
Avg_viewers	1.0000000	0.7654585	0.9995457	0.5471015	0.7323499	0.7574036
Hours_Streamed	0.7654585	1.0000000	0.7664120	0.7971457	0.5823754	0.9624282
Hours_watched	0.9995457	0.7664120	1.0000000	0.5472144	0.7335951	0.7581867
Peak_channels	0.5471015	0.7971457	0.5472144	1.0000000	0.5603652	0.7562565
Peak_viewers	0.7323499	0.5823754	0.7335951	0.5603652	1.0000000	0.5882104
Streamers	0.7574036	0.9624282	0.7581867	0.7562565	0.5882104	1.0000000

Table 1: correlation table

```
Analysis of Variance Table

Model 1: Hours_watched ~ Hours_Streamed + Peak_viewers + Peak_channels +
    Streamers
Model 2: Hours_watched ~ Hours_Streamed + Streamers
  Res.Df        RSS Df   Sum of Sq      F    Pr(>F)
1  12745 7.4032e+17
2  12747 1.1777e+18 -2 -4.3738e+17 3764.9 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 2.

```
Linear hypothesis test

Hypothesis:
Peak_viewers = 0
Peak_channels = 0

Model 1: restricted model
Model 2: Hours_watched ~ Hours_Streamed + Peak_viewers + Peak_channels +
    Streamers

  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
1  12747 1.1777e+18
2  12745 7.4032e+17  2 4.3738e+17 3764.9 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 3.

```
Start:  AIC=410009.5
Hours_watched ~ Hours_Streamed + Streamers

                Df  Sum of Sq        RSS    AIC
+ Peak_viewers   1 3.4771e+17 8.2999e+17 405552
+ Peak_channels  1 2.9361e+16 1.1483e+18 409691
<none>                       1.1777e+18 410010

Step:  AIC=405551.6
Hours_watched ~ Hours_Streamed + Streamers + Peak_viewers

                Df  Sum of Sq        RSS    AIC
+ Peak_channels  1 8.9670e+16 7.4032e+17 404097
<none>                       8.2999e+17 405552
- Peak_viewers   1 3.4771e+17 1.1777e+18 410010

Step:  AIC=404097.3
Hours_watched ~ Hours_Streamed + Streamers + Peak_viewers + Peak_channels

                Df  Sum of Sq        RSS    AIC
<none>                       7.4032e+17 404097
- Peak_channels  1 8.9670e+16 8.2999e+17 405552
- Peak_viewers   1 4.0802e+17 1.1483e+18 409691

Call:
lm(formula = Hours_watched ~ Hours_Streamed + Streamers + Peak_viewers +
    Peak_channels, data = data)

Residuals:
       Min         1Q     Median         3Q        Max
-170942415    -792046     181250     671247  174590856

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -8.008e+05  7.375e+04 -10.859  < 2e-16 ***
Hours_Streamed  1.934e+01  5.144e-01  37.592  < 2e-16 ***
Streamers       1.572e+01  4.701e+00   3.343  0.00083 ***
Peak_viewers    6.061e+01  7.231e-01  83.811  < 2e-16 ***
Peak_channels  -1.769e+03  4.502e+01 -39.290  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7621000 on 12745 degrees of freedom
Multiple R-squared:  0.7442,    Adjusted R-squared:  0.7442
F-statistic:  9271 on 4 and 12745 DF,  p-value: < 2.2e-16
```

Table 4.