"A Time Series Analysis of Video Game Player Bases"

Abstract:

In this project, our group analyzed and modeled time series of average monthly player base of four different video games using ARIMA modeling techniques, in the hopes of seeing if useful models could be built for video games, and to see if games of different genres and sizes would model similarly. Overall, we found that useful models could be made, and while there is not one universal model for games, certain games of different genres and sizes still might model similarly.

1. Introduction

1.1 Background and Significance

Video games have been a popular pastime for decades and have only become increasingly popular in recent years. By 2025, the global gaming market is expected to be worth over 250 billion dollars. The video game industry's growth has been propelled by technological advancements that are constantly improving the quality of the games as well as the overall experience of the users (Mordor Intelligence). Today, there are many different ways for people to play video games without having to own every expensive gaming console or to own expensive physical copies of games. Digital game distribution platforms allow users to purchase and download games online for either a specific console or a PC. Steam is one of if not the most popular digital game distribution platform with over 120 million monthly active users in 2020 (Irwin 2020). Steam offers many of the available games on the market for PC users with over 50,000 distinct games as of 2020. With so many available games and plenty of players, it would be of importance to find useful statistics and metrics of users play patterns for each game. The results of such could influence decisions of Steam and Game Creators, potentially impacting on the future the economics, popularity, and risk involved. Knowing trends in player bases is particularly important for studios running servers for multiplayer games. By being able to have an understanding of what their player base might be going into the future, they can increase or decrease server capacity to best handle their fluctuating playerbases.

1.2 Data Description

The dataset we used contains player information for 38164 Steam games from June 30th 2012 to October 23rd 2021. The data includes the average monthly players and peak players, along with the change in players, on a month by month basis. The range of average monthly players varied from 0 to a max of 1,584,886. The average monthly players for all games in the set was 329, with a median of only 2.5. This indicates that many games in the dataset had few players, with some games that had a substantially higher playerbase. In this project we looked at four games in particular. The first game selected is titled Counter Strike: Global Offensive. This game is an online multiplayer tactical shooter, with multiple game modes for players to have fun with their teams. The game has large competitive communities with professional teams and players that compete for millions of dollars. The second game that was selected for our analysis is called Stardew Valley, which is a farming simulation role-playing game. Players create their own characters and build up their farms and the surrounding community.

Also, players can invite other players online to join their farms and work together. Another one of the games selected was Sid Meier's Civilization V. In this strategy game, players choose a famous historical leader and build up an empire over time, playing through the various eras of human history. While most commonly played single player, the game does have a multiplayer option. Finally, Universe Sandbox is a physics-based simulator. It allows players to explore unrealistic and often impossible scenarios on a galactic scale by modifying or adding physical systems or rules to space, including situations such as what would happen if our Sun was replaced by a black hole or if a planet in our Solar System was thrown off its orbit. Players can change values like the gravitational constant on a planet, its mass, and other characteristics that make objects behave a certain way according to the laws of physics.

1.3 Hypotheses

With this data, we wanted to see two things. First, is it possible to create useful models from the data? Secondly, do these games with different genres and consumer sizes model the same way? That is to say, we want to see if it's possible to create a universal model for monthly players over time, across all video games, or if different video games require different parameterizations. It may even be the case that games that are similar in genre are modeled similarly and games that are dissimilar have different best models. We will test this by modeling four different video games of different genres and sizes, and see if we can create useful models for each and independently reach similar models or not.

2. Methods

2.1 Data Collection

The data we used was downloaded from the website Kaggle and was uploaded by user Connor Wynkoop. According to his description, he had scraped the data off the website steamcharts.com, which contains the average players, peak players, and player change for all games on the platform ranging back to June 30th 2012.

2.2 Variable Creation

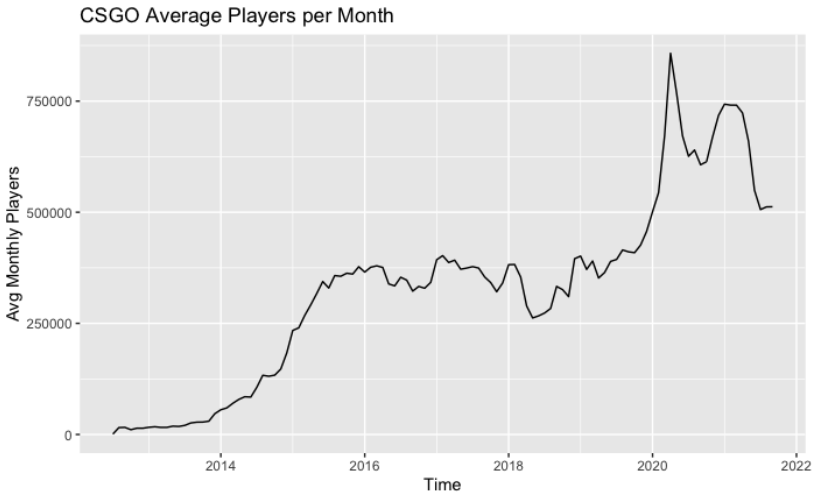
From the entire data set, we separated the four games we planned to analyze, and transformed these four new data sets into separate time series of monthly average players ranging from June 2012 to October 2021.

2.3 Analytic Methods

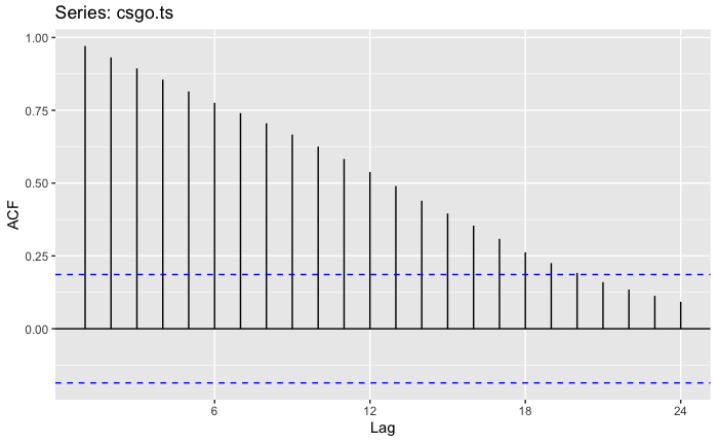
Our main method to analyze the data was to fit it to autoregressive integrated moving average, or ARIMA, models. An ARIMA model consists of three parts; AR, I, and MA. In the ARIMA(p, d, q) model, p denotes the number of autoregressive terms, d denotes the difference applied to the time series data, and q denotes the number of moving average terms. Stationarity analysis was performed in order to determine the optimal d for each game's final model. To analyze data stationarity, the Dickey-Fuller test and associated methods were used. The Dickey-Fuller method tests the null hypothesis of stationarity being not present in the data, with the alternative hypothesis that stationary is present. The independence of data in the time series was determined through the Ljung-Box test. The Ljung-Box method tests for the null hypothesis of data independence, and an alternative of serial autocorrelation.

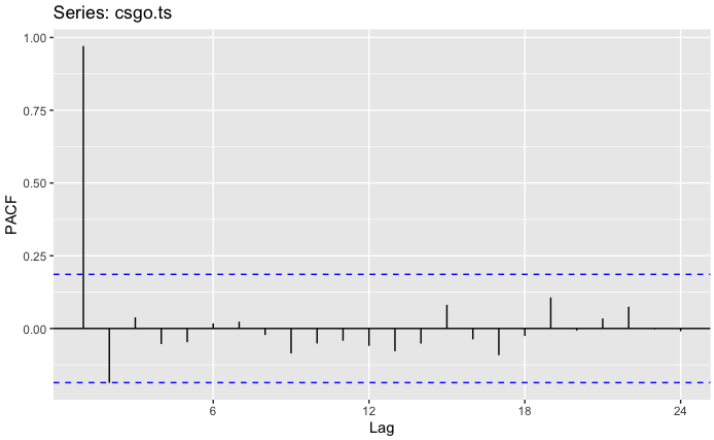
2.3.1 Exploratory Data Analysis (EDA)

Counter Strike: Global Offensive



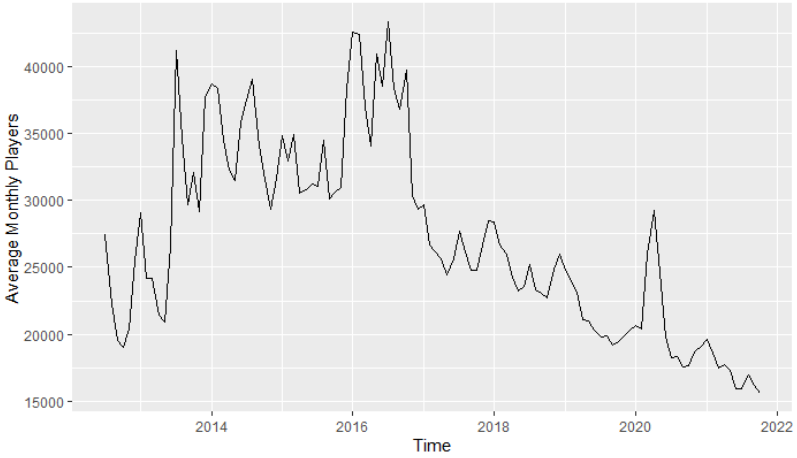
Counter Strike: Global Offensive (CSGO for short) was released in August of 2012, and remains one of the highest played games on steam, and within this data timeframe is in the top 5 most popular games. The Initial plots of the ACF, PACF, and overall Average of Players vs Time plot are shown above. The average monthly players over time plot indicates the game has an upward trend overall, although a dip in players after 2020. The ACF plot shows a steadily decreasing trend that is significant for many time lags, and the PACF shows a large spike at 1 and then drops off. Based on the time series plot, we can assume that this is not stationary as there is no evidence of a constant mean over time, and the variance does not seem to be constant as well observing the peaks of the graph. The steadily decreasing ACF for many lags also provides supporting evidence to this assumption.

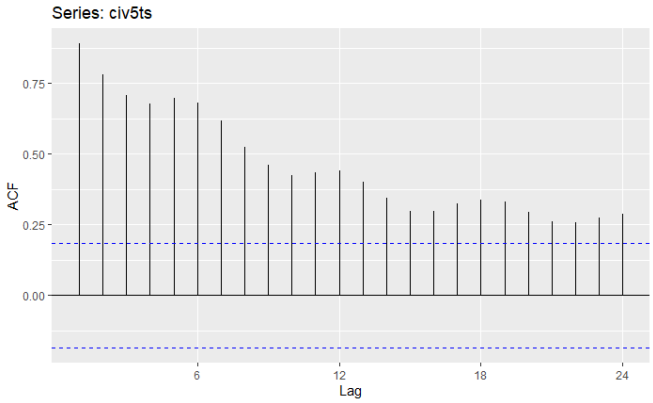




Sid Meier's Civilization V

On the right is the time series plot for average players for Civ 5. Much of the spikes in average players is probably due to the way the game is marketed. Civ 5 would often release major updates or downloaded content that would refresh the game and bring new players to the game and bring old players back, increasing the player base for periods of time. Additionally, the game would often sell for a heavily discounted price, bringing new players to the game when it entered a price range they could afford. Eventually though, with the release of the sequel in 2016 and its inevitable sales and content, people migrated from the game in 2017. The beginning of the Covid-19 pandemic did help to see a spike in player base, as with more time on people's hands they returned to one of their old favorites. This revival was short lived, and the playerbase for the game continues to trend downwards.

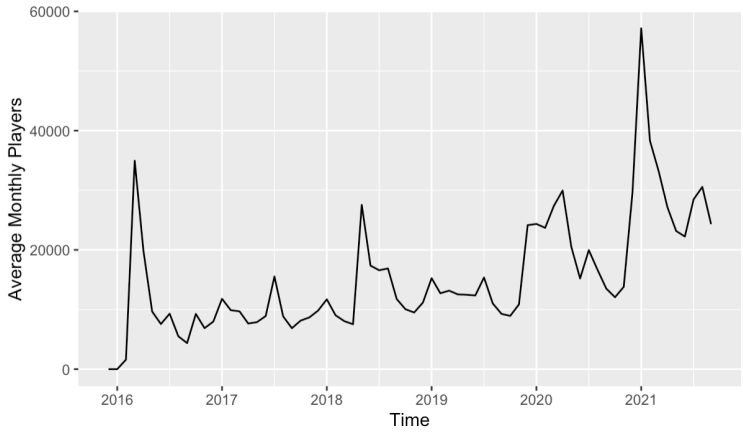




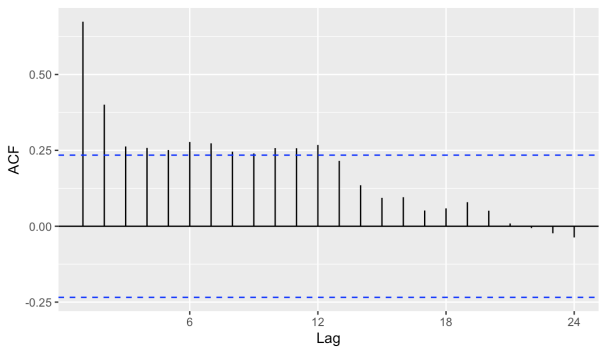
From the ACF plot of the time series, we see a trend that lends itself towards either an AR model or a differenced model, and upon some short analysis we found that differencing the time series helps to greatly decrease the residual lag in the model, which will be later explored in the modeling procedures.

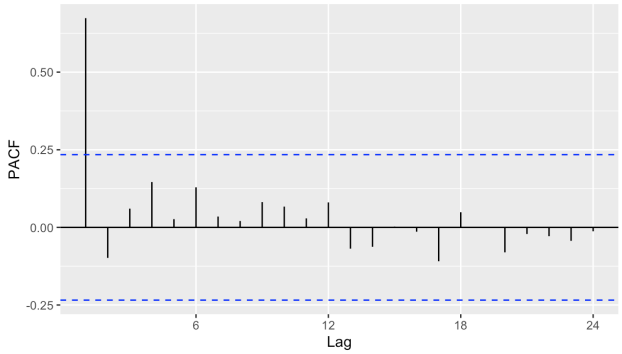
Stardew Valley

Before plotting the average monthly player data from Stardew Valley, the data needed to be sorted so that the older values appeared before more recent ones. Below is a plot of Stardew Valley's average monthly players on Steam over time.



The game was officially released on Steam in February of 2016, which can be seen in the plot as the first significant spike. The second larger spike that appears after 2018 correlates to the release of a highly anticipated multiplayer feature of the game that would allow up to four players to collaborate on a farm together. It is clear from the plot above that the average monthly player data for Stardew Valley is not stationary. The mean consistently increases over time and the many large spikes in the graph show that the variance is not constant. Next, we will look at the ACF and PACF plots for the data.

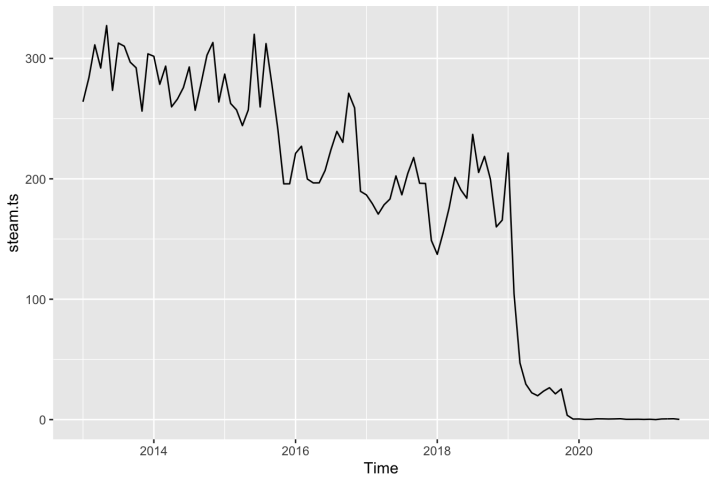




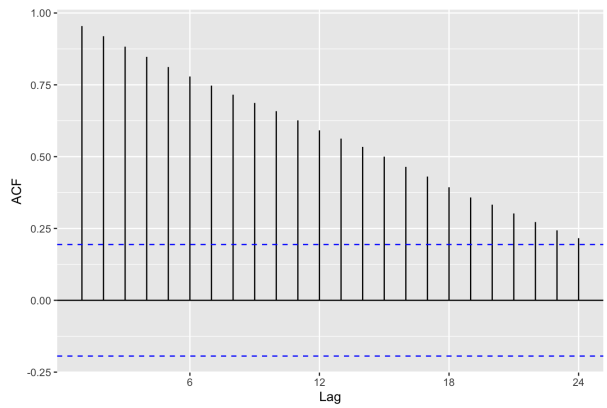
The ACF plot on the left shows one large spike at time lag 1, a slightly smaller spike at lag 2, and then the autocorrelation decreases slightly for the following lag values. This suggests that there may be an auto-regressive component to the data. Because there is no significant drop off of ACF values after the first or second large spike, the ACF plot does not look similar to that of an MA(1) or MA(2) process. In the PACF plot, we see a large spike at lag 1 as well, but then the autocorrelation values drop off after that. This suggests that an AR(1) component may be present.

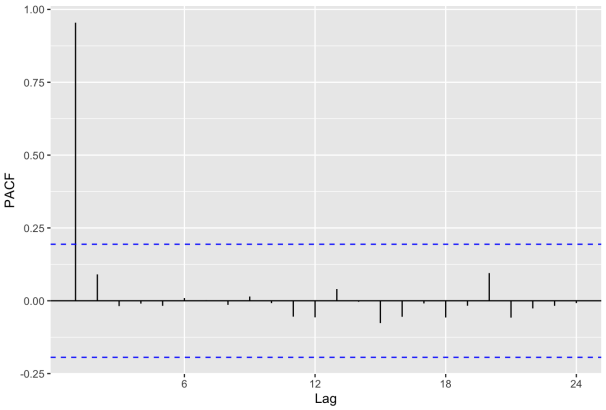
Universe Sandbox

The plot below is the time series plot of the monthly players of Universe Sandbox from 2012 to 2021. We see that in 2019, there was a massive decline in monthly players, eventually dropping the user base to 0. This can be attributed to the release of the sequel, Universe Sandbox 2, that showcased newer features and more advanced simulation.



The ACF plot is extremely similar to the ACF plot of the AR(1) process. It features slowly and constantly decreasing autocorrelation as lag increases, with all ACF values being above the blue, dotted confidence lines indicating significance. The PACF plot further verifies this by showing us that the autocorrelation at lag 1 when removing the effect of all other lags is greatest - this is a function of the one autoregressive component presumably powering this distribution. This is helpful intuition for building models later on, as we now guess that an ARIMA(1, 0, 0) model might be appropriate in modeling this time series.

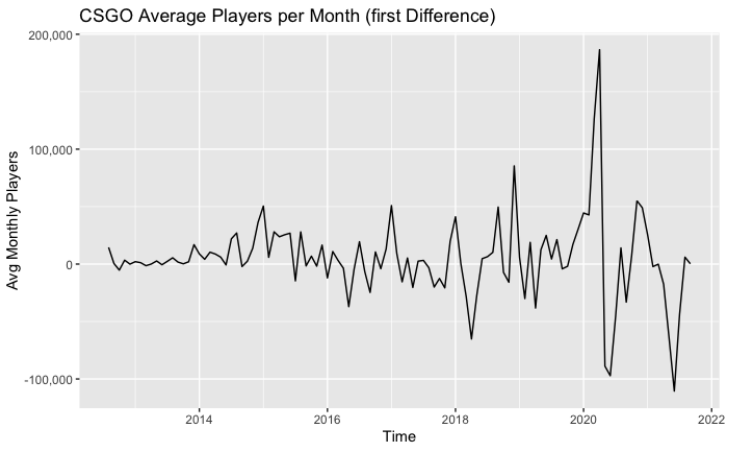


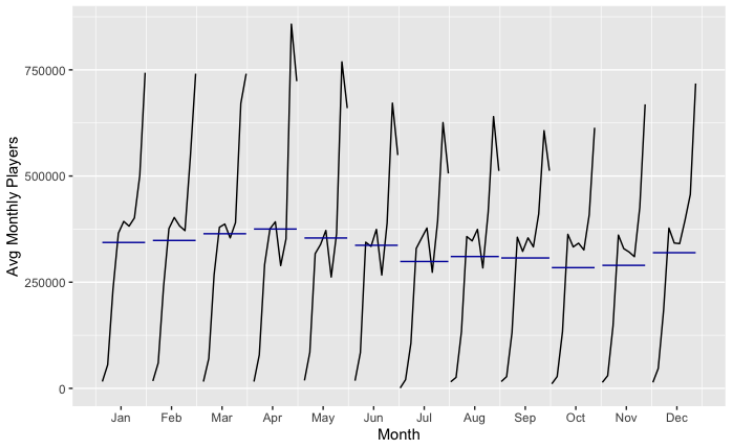


2.3.2 Modeling Procedures

Counter Strike: Global Offensive

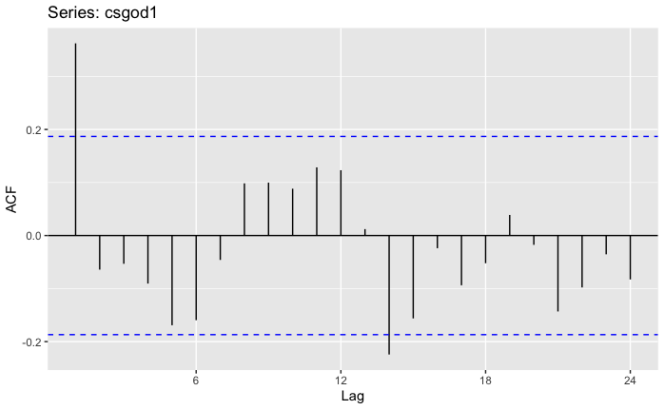
To help correct stationarity issues in the original time series of the CSGO data, the first difference was taken. This visually produced a time series that met more of the assumptions of stationarity, however do recognize that the variance seems to increase over time. This is mostly due to the increased player base over time, as the later values appear more volatile than earlier time periods. To add supporting evidence that this first difference is able to meet assumptions, the Dickey-Fuller test was applied and found the p-value to be less than 0.01. This means that the model can be assumed stationary by the Dickey-Fuller hypothesis. Looking at seasonality of the time series on a monthly basis plot, there does not appear to be any suspicion of seasonality as most months are relatively the same spread.

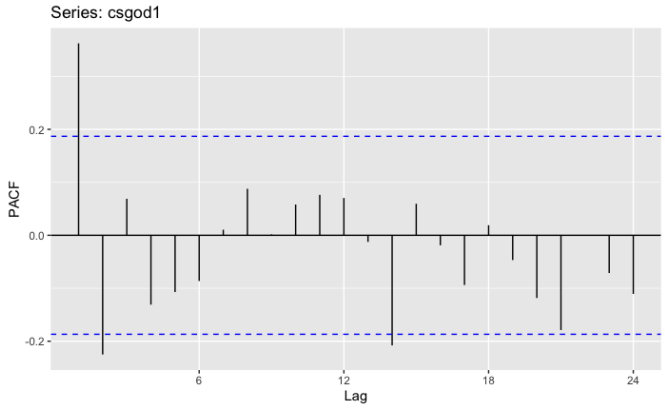




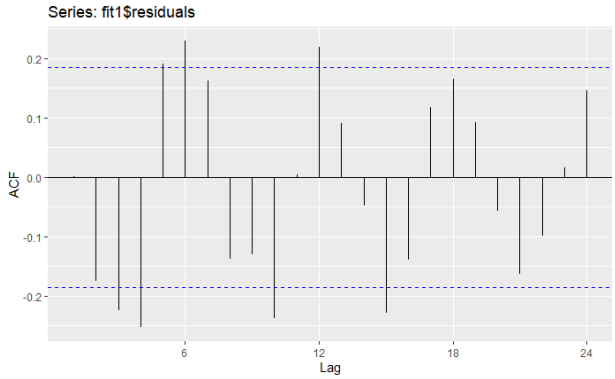
The ACF plot of the first difference of the CSGO data shows a spike at lag 1, that then drops off under the threshold for the remaining time periods.

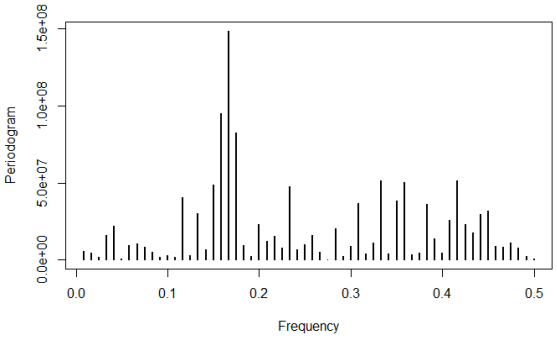
There is a single spike 14 time lag, but this is most likely due to random chance that this is significant due to the structure of the other times. This suggests an MA model and looking at the PACF, there is a significant spike at lag 1 and a smaller but significant spike at lag 2. This suggests that an IMA(1,1) or IMA(1,2) may well fit this data.



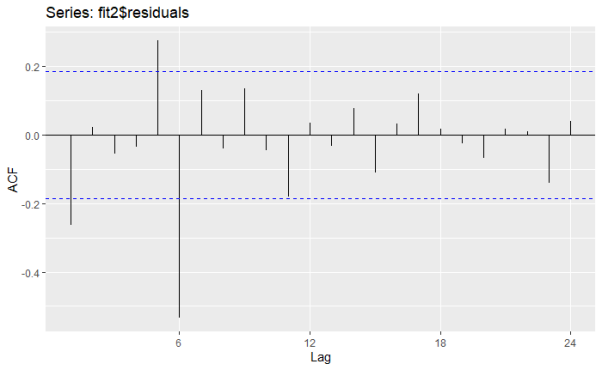


Sid Meier's Civilization V



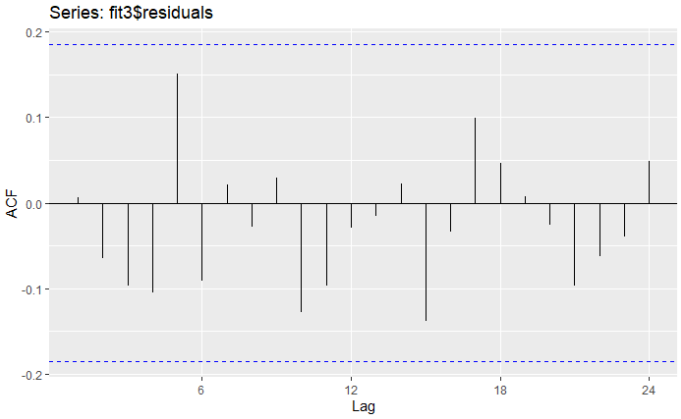


After taking the first difference of the data, we see some sort of seasonal pattern in the residual lags. Through a periodogram of the residuals, we see a distinct spike at 1/6. This means our time series has a 6 month seasonal trend, so we should take the 6 month seasonal difference.



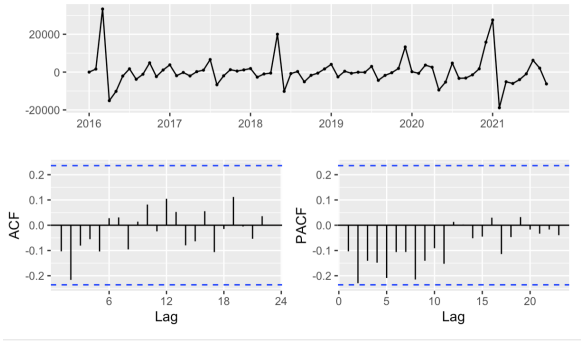
After adding the MA components to the model, we see no significant spikes left in the residuals, and we do not see any sort of seasonal pattern.

Looking at the residuals after taking the 6 month seasonal difference, we see significantly better residuals, but we still have significant spikes at lag 1, 5, and 6. By adding a MA(1) and seasonal MA(1) component, we should be able to get rid of these significant spikes.

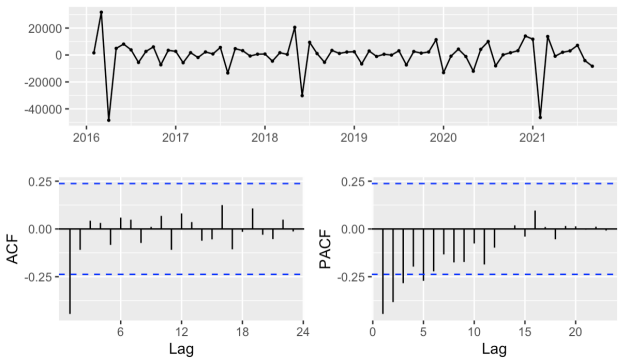


Stardew Valley

To begin modeling the average monthly players of Stardew Valley, the first and second differences were taken to determine if differencing would improve the stationarity of the data. The plots of the first difference appear on the right and show that taking the first difference does give the data a more constant mean. Also, there are no significant autocorrelation values for any of the lags in the ACF or PACF plots. The results of the Dickey-Fuller test produce a p-value less than 0.05, meaning that the null hypothesis is rejected and the data is stationary.




```
        Augmented Dickey-Fuller Test

data:  first_diff
Dickey-Fuller = -5.5241, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary
```



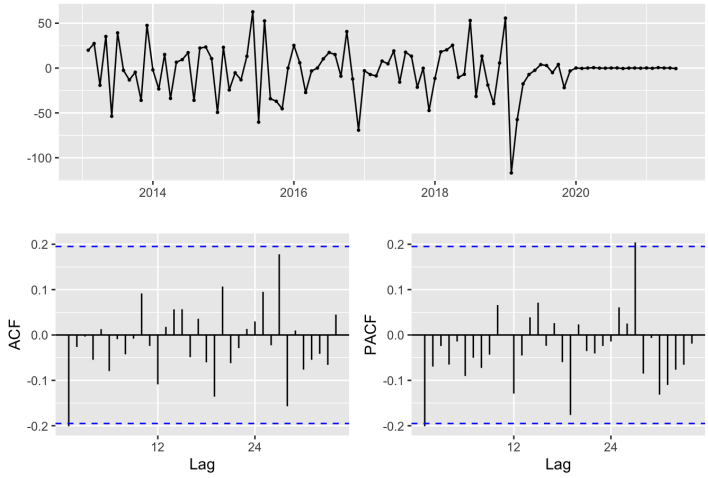
To the left are the plots for the second difference of the Stardew Valley data. Again, we can see that differencing has made the mean constant. However, there are still some significant autocorrelation values in the ACF and PACF plots for the second \

```
        Augmented Dickey-Fuller Test

data:  second_diff
Dickey-Fuller = -6.7705, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary
```

difference. Even though there does appear to be some non constant variance in the plot, the Dickey-Fuller Test reports that the data after taking the second difference is stationary.≈

Universe Sandbox

Before modeling, the data was differenced to see if there was any effect on stationarity and whether it was induced. The unchanged data is not stationary, as can be seen by the systematic change in the mean over time. Differencing did bring about stationarity - this is indicated by the autocorrelation values in the ACF plot of the differenced data always being between the blue dotted confidence lines.



The large dip caused by the abandonment of this game for its sequel may be cause for concern in terms of heteroskedasticity, or non-constant variance, but it is out of the ordinary and caused due to a special, known, directly causal circumstance. Taking the logarithm of the differenced data reduces the scale of the variance, but it is not necessary to use the log data to model the original here.

Since differencing induced stationarity, the model search was two-pronged: verify that ARIMA(1, 0, 0) models this data well and search for ARIMA parameterizations with $d = 1$.

3. Results

3.1 Model Comparison

Counter Strike: Global Offensive

For all fit measures AIC, AICc, and BIC the IMA(1,1) Model performed better than the IMA(1,2).

Model	AIC	AICc	BIC
IMA(1,1)	2600.540	2600.653	2605.941
IMA(1,2)	2602.461	2602.687	2610.562

Thus, the IMA(1,1) model will be considered as the best model.

Sid Meier's Civilization V

Comparing the model chosen earlier to the simpler models, and more complex models with added AR components, we get the following results.

	AICc	Box Test p-value
I(1)	2108.239	0.0000594
I(1), sI(1)[6]	2040.726	0.0000000
IMA(1,1),sIMA(1,1)[6]	1986.674	0.5607752
ARIMA(1,1,1),sIMA(1,1)[6]	1987.516	0.6340424
IMA(1,1),sARIMA(1,1,1)[6]	1987.787	0.7160990
ARIMA(1,1,1),sARIMA(1,1,1)[6]	1988.832	0.8247488

Based on having the best AICc score, along with a significant Box Test p-value, we stick with the ARIMA(0,1,1), sARIMA(0,1,1)$_6$ model.

Stardew Valley

Several ARIMA models were fit to the Stardew Valley data, starting with ARIMA(1,0,0) because we noted a potential AR(1) component in the data while looking at the ACF and PACF plots. The p, d, and q components were incremented one by one in order to compare as many combinations of values as possible. Below is a table that reports the AICc value for the top 8 models that were found as well as the model that was produced by the auto.arima function in the "forecast" package in R.

Fit	AICc
ARIMA(1,0,0)	1447.809
ARIMA(1,1,0)	1434.699
ARIMA(1,2,0)	1454.094
ARIMA(2,1,0)	1433.342
ARIMA(2,1,1)	1425.860
ARIMA(2,1,2)	1424.196
ARIMA(2,2,2)	1410.809
ARIMA(1,1,1)	1426.023
Auto Arima	1433.273

The auto.arima function reported that ARIMA(2,1,0) was the best model for the average monthly players of Stardew Valley, However, several other models produced lower AICc values than ARIMA(2,1,0), which indicates better performance. The ARIMA(2,2,2) model produced the lowest AICc as well as the lowest RMSE, so it was selected as the final model for the average monthly players of Stardew Valley.

Universe Sandbox

	Model	AIC	S2
1	ARIMA(0,1,1)	953.5974	723.0344
2	ARIMA(1,0,0)	970.7345	743.3099
3	ARIMA(2,0,0)	969.5433	719.6601

Among all models being compared, the ARIMA(0, 1, 1) had the lowest AIC, making it the best model to model monthly player data for Universe Sandbox.

3.2 Selected Models and Equations

Counter Strike: Global Offensive

The CSGO model using an IMA(1,1) can be written as:

$$Y_t = Y_{t-1} + \varepsilon_t + 0.4896\varepsilon_{t-1}$$

Sid Meier's Civilization V

Based on everything above, we get a final model of

$$y_t = y_{t-1} + y_{t-6} - y_{t-7} + \varepsilon_t - 0.2138\varepsilon_{t-1} - 0.732\varepsilon_{t-6} + 0.1565\varepsilon_{t-7}$$

Stardew Valley

The summary of the final fit for the Stardew Model is shown below.

```
ARIMA(2,2,2)

Coefficients:
        ar1      ar2      ma1     ma2
      0.613  -0.2054  -1.9891  1.0000
s.e.  0.120   0.1209   0.0831  0.0833

sigma^2 = 46858185:  log likelihood = -699.92
AIC=1409.84   AICc=1410.81   BIC=1420.94

Training set error measures:
                    ME     RMSE      MAE       MPE     MAPE      MASE
ACF1
Training set -35.39736 6545.362 4196.94 -12263.77 14118.34 0.6333341
-0.01273487

        Box-Ljung test

data:  residuals(fit8)
X-squared = 1.7956, df = 5, p-value = 0.8766
```

The Ljung-Box test produced a p-value that is much larger than 0.05, meaning that there is insufficient evidence to reject the null hypothesis that states the data is independently distributed. The final model equation for the Stardew Valley data is surprisingly the most complicated compared to the other final models that were selected. The final model equation is

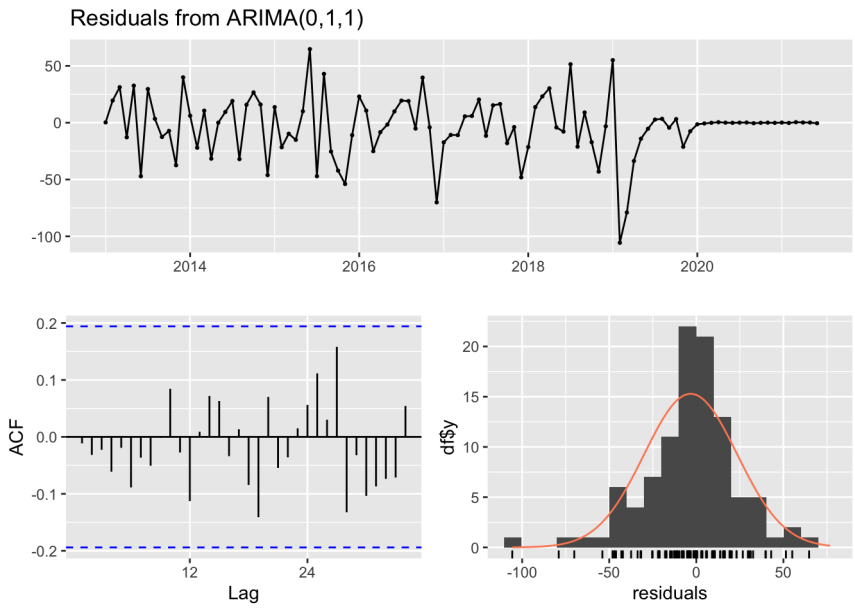
$$\Delta^2 Y_t = 0.62\Delta^2 Y_{t-1} - 0.21\Delta^2 Y_{t-2} + e_t - 1.99e_{t-1} + e_{t-2}$$

where $\Delta^2 Y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2})$.

Universe Sandbox

```
Call:
arima(x = steam.ts, order = c(0, 1, 1))

Coefficients:
          ma1
      -0.2043
s.e.   0.1000

sigma^2 estimated as 723:  log likelihood = -475.8,  aic = 953.6
```

```
 Ljung-Box test

data:  Residuals from ARIMA(0,1,1)
Q* = 9.6912, df = 19, p-value = 0.9602

Model df: 1.   Total lags used: 20
```



With a p-value of 0.96 in the Ljung-Box test, the ARIMA(0, 1, 1) was the best model considered amongst all the other models. The MA(1) theta coefficient was 0.2. The model equation can then be written as

$$(1 - B)Y_t = \varepsilon_t - 0.2\varepsilon_{t-1}$$

where B denotes the backshift operator.

4. Discussion

4.1 Conclusions

The models the different games arrived at indicates that there may be some potential for a conglomerate model, depending on the type of game and its popularity. For the games Counter Strike: Global Offensive and Universe Sandbox, both had a considered IMA(1,1) as the best modeling choice albeit with different values of theta. Similarly, Sid Meier's Civilization V arrived at a best model choice of an ARIMA(0,1,1) SARIMA(0,1,1)[6]. These three previous models all involve a difference, and a measure of the current and past moving averages. This could suggest there is a potential for having a model that can evaluate multiple games from different genres and popularities. However, the game Stardew Valley had arrived at a much different model of ARIMA(2,2,2). This inclusion suggests against our initial hypothesis of a possibility of a universal model structure, but opens up another question: can different games be categorized into their respective conglomerate model based on available metrics?

4.2 Limitations

There are several limitations to our research from working with this dataset, one being that it ignores players that don't use Steam to access the video games. Another limitation of the dataset is that it only contains the average monthly player data until October of 2021. Having the most recent data would allow us to more accurately model the time series for each game. A third limitation in our research methods was only choosing four games to analyze. While all four games do fall into different genres, there are plenty of different video game genres and subgenres. Because of this, it is difficult to draw any meaningful conclusions on how the genre of a video game influences how its monthly average players will be modeled as a time series. A much larger subset of video games would need to be analyzed in order to make more definitive conclusions on the relationship between game genres and average monthly players.

4.3 Future Direction

From our current analysis, it is clear that genre does have an impact on how the monthly average players of a video game will be modeled. A next step would be to further investigate genre as a factor by selecting a subset of video games from a single genre and performing time series analysis on the average monthly players of each game to identify any similarities or differences in the models. Another way to continue this research would be to analyze the impact of other factors, such as game popularity or whether the game is a part of a series, on models of the average monthly players.

4.4 Summary

Overall, our analyses suggest that certain video games might model the same, or at least contain certain shared elements, but there is no one shared model that can be used for all video games. This makes sense, there are so many styles of video games; different genres, varying popularity, single player or multiplayer. Going into this project we did not expect to reach a singular model for all games. Despite this, we did gain interesting knowledge about how to model games, and how certain games might model similarly.

Acknowledgements

We would like to thank Dr. Tatjiana Miljkovic for our learning experience in the classroom and supporting us in this project.

References

"Gaming Market Forecast, Revenue, Trends: 2022 - 27: Industry Growth." Gaming Market Forecast, Revenue, Trends | 2022 - 27 | Industry Growth, Mordor Intelligence, https://mordorintelligence.com/industry-reports/global-gaming-market.

Irwin, Kate. "Valve's Steam Store Breaks All-Time Highs with 28 Million Users." Input, Input, 4 Jan. 2022, https://www.inputmag.com/gaming/valves-steam-store-breaks-all-time-highs-with-28-million-users.

Witkowski, Wallace. "Videogames Are a Bigger Industry than Movies and North American Sports Combined, Thanks to the Pandemic." MarketWatch, MarketWatch, 22 Dec. 2020, https://www.marketwatch.com/story/videogames-are-a-bigger-industry-than-sports-and-movies-combined-thanks-to-the-pandemic-11608654990/.