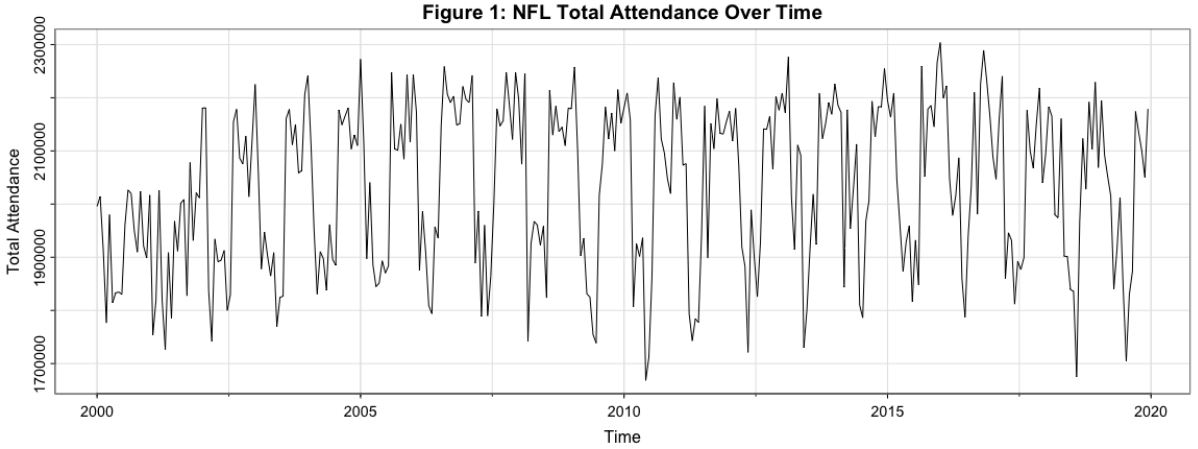
Ryan Schaefer and AnnaMarie Fernandes

# Time Series Project

## I. Introduction

The data set that we decided to use was the attendance at the games for NFL teams during the 17-game season for 20 years (2000-2019). We retrieved this data from the R for Data Science GitHub repository (https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-02-04/readme.md). Figure One is the plot of the time series and there are a lot of interesting aspects about this time series. First, the time series exhibits a seasonal pattern where it oscillates between the values of 1750000 and 2250000. A time series is stationary when there is a constant mean, variance, and autocorrelation over time. This time series looks to be stationary because the oscillations over time look to be consistent, which correlates with the time series having a constant mean, variance, and autocorrelation over time. The aim of our project is to forecast the cumulative NFL attendance for the next season (next year). Some questions that come to mind when doing this project are: will the attendance go up next season, is it reasonable to predict attendance for the upcoming season based on the attendance for the last 19 seasons, and which part of the season has lower attendance compared to the part of the season with higher attendance and why?

**Figure 1: NFL Total Attendance Over Time**



## II. Statistical Analysis

### II.I. ACF and PACF

**Figure 2: NFL Total Attendance ACF**

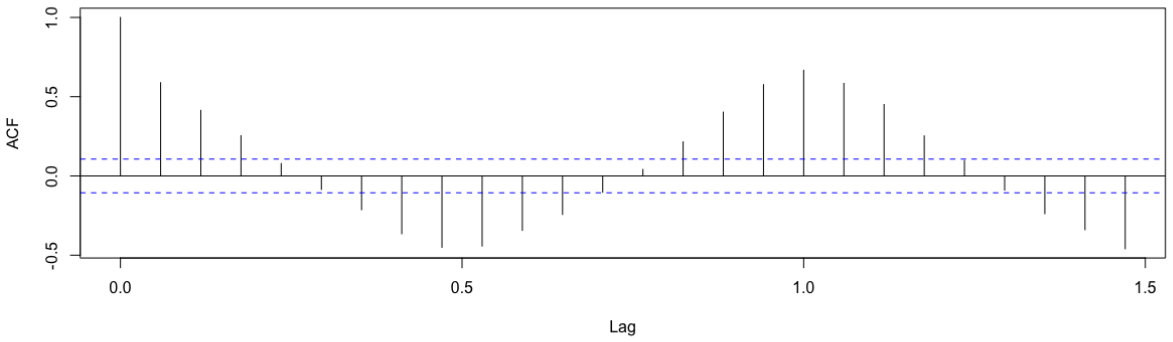
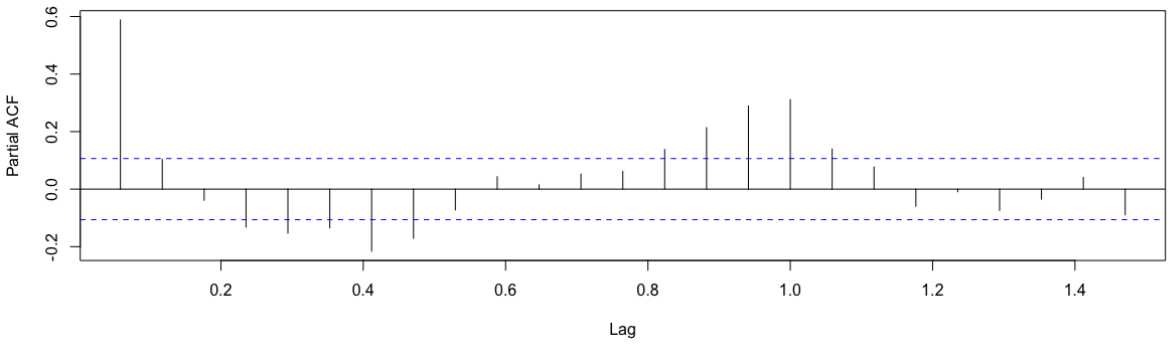



**Figure 3: NFL Total Attendance PACF**




We first did data transformation where we grouped by the season and the game week, and then took the sum of the attendance of all teams to get a cumulative sum. We also split the data into training which included the data from 2000 to 2018, and testing which included just the 2019 data. We used training to fit our models and testing to check to see how accurate our forecasts were. We then created the ACF and PACF plots with the transformed data which are shown in Figure 2 and Figure 3, respectively. Figure 2 (ACF) illustrates a clear seasonal pattern with the oscillation from positive to negative values over time, which is consistent with the time series plot. The ACF confirms that there is a period of about one season based on the sinusoidal behavior of the plot. The PACF plot also validates that the time series is seasonal because of the positive and negative patterns. Furthermore, we conducted a Dickey-Fuller test and found that the p-value was less than 0.01, which means that the time series is stationary.

## II.II. Holt Winters Model

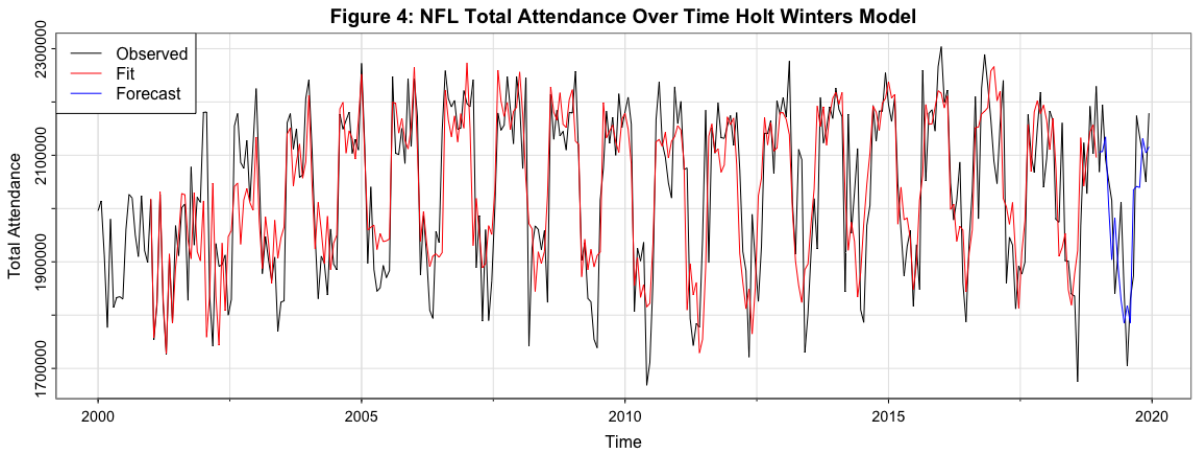



Figure 4 shows a plot of the NFL total attendance time series (black line), the Holt-Winters model fit on the training data (red line), and the Holt-Winters model forecast of the 2019 season's attendance (blue line). We chose to fit a Holt Winters model for this data because it considers trend and seasonal effects. We fit an additive Holt-Winters model because the within-season variation appears to be approximately constant throughout most of the time series. A Ljung-Box test on the residuals of this model gives a p-value of 0.8011, meaning that the residuals behave like a white noise time series and that the model is a good fit for the data.

## II.III. SARIMA

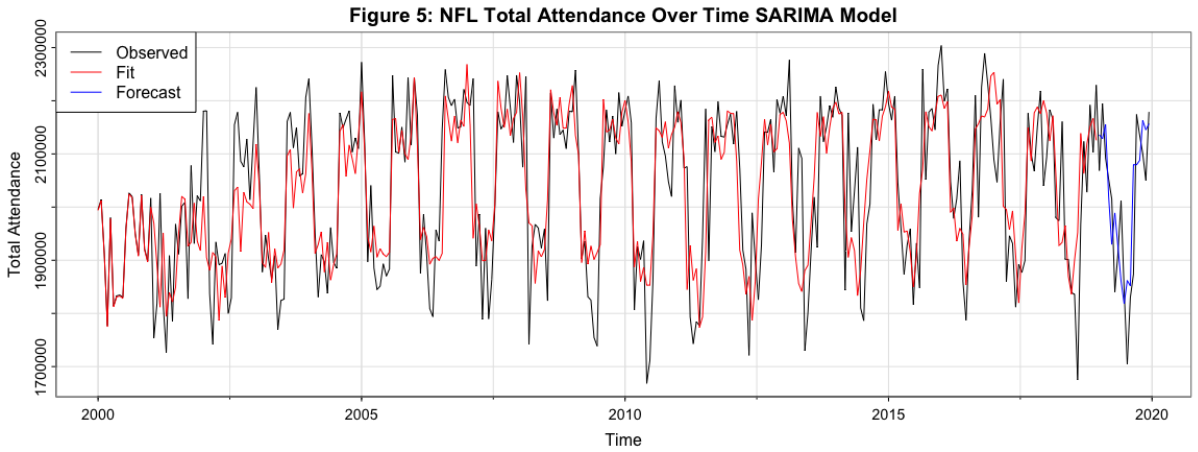



Figure 5 shows a plot of the NFL total attendance time series (black line), the SARIMA model fit on the training data (red line), and the SARIMA model forecast of the 2019 season's attendance (blue line). We determined the model and seasonal orders using the auto arima function on the training data and found that the order was 1, 0, 1 and the seasonal order was 0, 1, 1. We used the training data, the model order, the seasonal order, and the period of 17 (due to the number of games in the season) to fit the model. A Ljung-Box test on the residuals of this model gives a p-value of 0.3108, meaning that the residuals behave like a white noise time series and that the model is a good fit for the data.

## II.IV. Linear Regression

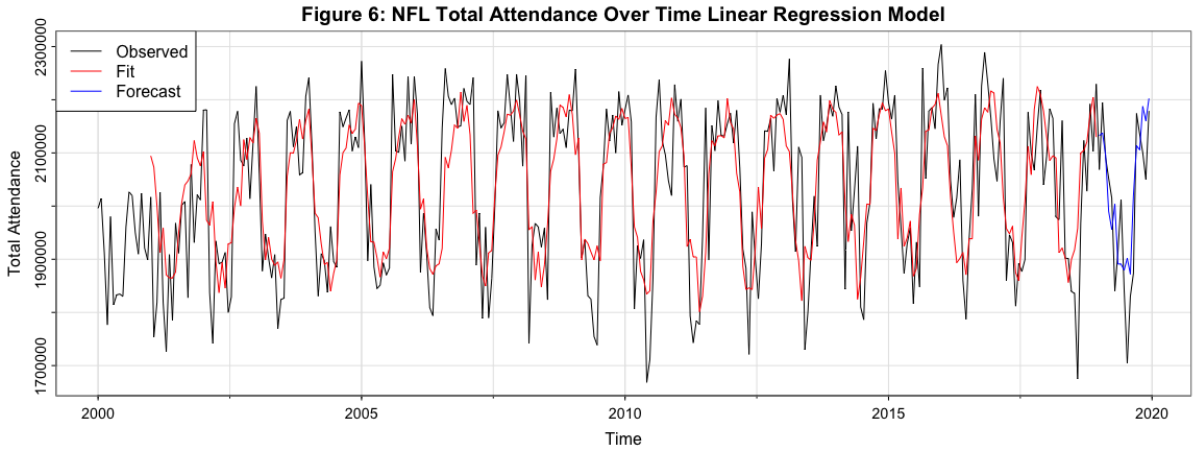



Figure 6 shows a plot of the NFL total attendance time series (black line), the linear regression model fit on the training data (red line), and

the linear regression model forecast of the 2019 season's attendance (blue line). To detect the reflect the seasonality of the time series, sin, and cos terms were used. A lag term of 17 was also used, because of the approximately 1 season period in the ACF plot (Figure 2). No trend term was used because the mean of the time series is approximately constant. A Ljung-Box test on the residuals of this model gives a p-value of 0.9216, meaning that the residuals behave like a white noise time series and that the model is a good fit for the data.

### II.V. Results

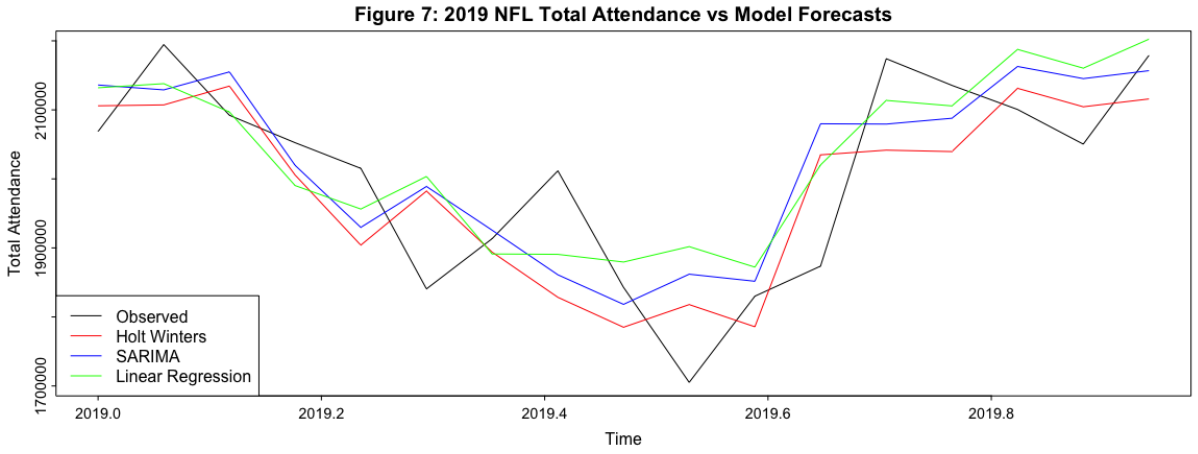



Table 1: RMSE of NFL Attendance Forecasts

| Model Type | RMSE |
|---|---|
| Holt Winters | 96414.18 |
| SARIMA | 96660.92 |
| Linear Regression | 92131.05 |

Figure 7 shows the total NFL attendance for each game week in the 2019 season (black line), along with the forecasted attendance from the Holt-Winters model (red line), SARIMA model (blue line), and linear regression model (green line). Table 1 gives the RMSE for each model's

forecasts of the total NFL attendance in the 2019 season. Figure 7 shows that, although there are a few observations that do not follow the forecasted patterns, all of the models reflect the seasonality of the time series well. Table 1 shows that the linear regression model produced the smallest RMSE, suggesting that it provides the best fit for this time series.

## III. Conclusion

Overall, we found that with this NFL Attendance data set that ranged from 2000 to 2019, the linear regression model looked to be the model with the best fit. Based on the RMSE scores for all of the models, it confirmed that the linear regression model was the best because it had the lowest RMSE score. All three models had Ljung-Box test p-values that exhibited that the residuals behaved like a white noise time series, which verified that the models were a good fit for the data. Going back to the questions that were asked in the introduction, due to the model being stationary, it looks like the average attendance is not increasing. We also think that it is reasonable to predict attendance for the upcoming season based on the attendance for the last 19 seasons because of the consistent seasonality that we see in the time series plot. Looking back at Figure 7, it seems like the part of the season that has lower attendance occurs at around the halfway point of the season, and that the beginning and end of the season have the highest attendance. Looking at the full-time series in Figure 1, this pattern appears to be consistent across prior seasons.