Big Data Project

Influence of football matches on the crime rate in Philadelphia

August, Critty, Jack, Michael

February 18, 2020

Introduction

In an article written by NBC, the Philadelphia Eagles fan base is regarded as one of the most disliked fan bases in the NFL (Campitelli 2016). The Eagles fan base reputation may be due to public perception that Eagles fans are thought to commit more immoral acts than other sports fan bases. As an example, Eagles fans have been involved in acts such as rushing onto the field to assault the mascot of the opposing team, cheering when an opposing player received a career ending injury, and throwing snowballs at a man dressed as Santa Claus. We take this theory that Eagles fans are more likely to commit immoral acts and test if they tend to commit more crimes on game days. If this were the case, we would expect for the frequency of crimes in Philadelphia to be higher on days when the Eagles play than on days when they do not. In order to test this hypothesis, we compared crime frequencies in Philadelphia using the presence or absence of games as our explanatory variable and the total crime frequency as our response variable. Our null and alternative hypotheses are outlined below:

- H_0: The mean frequency of crimes in Philadelphia are similar between game days and non-game days.
- H_A: The mean frequency of crimes in Philadelphia are different between game days and non-game days.

We tested this hypothesis using Welch's two-sample t-test to understand if crime frequencies in Philadelphia are influenced by the presence or absence of a football game.

Methods

The basis of our analysis involved converting our two raw data sets into usable forms and then combining them into a single data set to analyze. Our Philadelphia crime data set consisted of reports for crimes committed everyday in years 2006 to 2019. Our Eagles data consisted of several seasonal data sets from the years 2005 to 2019, each of which had data regarding the games played in each season. The bulk of our analysis included a series of data cleaning methods which aimed to make the data compatible for t-testing. Our study aimed to test whether or not the presence or absence of a game day influences the observed frequency of crimes. We tested this using a two-sample t-test. A two-sample t-test requires two groups with the same dependent variable. In the case of our data, the explanatory variable was the presence or absence of a game day, which delineated our two sample groups. The dependent variable for each sample group was the continuous variable of total crime frequency for any given day of the year. Together these were used to compare the mean crime frequencies for game days and non-game days within the range of dates that composed the entirety of a given Eagles season. The following subsections outline the various methodologies we used to prepare our data for a series of t-tests.

Workflow

We began our workflow by creating a repository on GitHub, accessible at the following link: GitHub Repository. We then downloaded our data sets: "Eagles(2006-2019)".csv

files and the "Philly_Raw_Crime_Data.csv" file. We set our GitHub repository as our working directory. We created the folders, "Eagles" and "Figures." We then set up our pathways for our different folders in our working directory. We visualized both of the Eagles and crime data files as data frames on Rstudio. We manipulated the data to remove the variables we did not want in either of the csv files. After this, we further manipulated the data by merging the two data frames into one encompassing dataset. We then added a new column with binary variables, indicating either the presence or absence of a game. We visualized the data once more and then tested our data based on this new dataset using two-sample t-tests. We tested both for overall mean frequency and for individual frequencies, giving us the results of our experiment.

Data Cleaning - Eagles Games Data

We had fourteen data sets that indicated the total amount of games the Philadelphia Eagles played in a given season, from the years 2006 to 2019. To make our data compatible for analysis, we had to refine it. Our first challenge was to combine all fourteen seasons of the Philadelphia Eagles into one data frame. After accomplishing this, we then cleaned unnecessary data, leaving only the date of each match and a binary vector indicating that a match took place on this date. To interpret the dates, we had to use the as.Date() function to instruct R how to read the date format. We wrote a looped function that read each .csv file, bound all fourteen seasons together, removed erroneous entries and, output a binary value associated with dates in which a match took place.

Data Cleaning - Philadelphia Crime Rates

Our Philadelphia crime rate document was gigantic, with 2.7 million rows and 17 variables. To clean our data, we removed unnecessary columns by programming a for loop that systematically removed columns that were not applicable to our investigation. We also needed to standardize our date format to match with our Eagles data. The next task was to consolidate crimes by each day; the original format of our data set had a row for each individual crime committed, leading to us having 2.7 million rows. We wanted to find the frequency of crime on each individual day, so we wrote a function that first created an empty vector for each unique day and then filled the vector with consecutive dates. This reduced our number of rows from 2.7 million to about five thousand. Our function then created a second column and filled each date with the corresponding frequency of crimes. Alongside the Eagles data code, this code is fully scalable and reproducible.

Merging Data

We wanted to merge our cleaned Eagles and crime data. To accomplish this, we used the merge() function. This function merged our two data frames by matching dates. When we merged, only our Eagles data came with a binary column indicating whether or not a game was played on this date. If a game was not played on the date in the crime data, an NA value was listed in the binary column. To remove these NA values and replace them with zeros, we applied a function that converted all of the missing (NA) values in the binary column to zeros, giving a value of 0 to any date in which a

match was not played by the Eagles. We then had a completed data set with binary columns that differentiated the days in which a game was played, which was automatically exported to our Data Output folder as a csv file named FinalCleanData.csv. The large size of our datasets made them incompatible with our GitHub repository. Thus, the Git repository only has our FinalCleanData.csv file and all git users should load the libraries at the beginning and then skip to line 200 of our script to reproduce our results. For use of the original raw data, we can be contacted for file transfer at august.ramberg-gomez@questu.ca.

Analysis

We conducted a series of Welch two-sample t-tests for each season from 2006 to 2019, as well as a final t-test across all seasons. The two-samples used in the seasonal t-test were delineated by the presence or absence of a game. There were roughly 17 game days a season and 120 non-game days. The length of the season was dependent on the success of the Eagles in that season. Thus, some seasons were longer with more game days and non-game days if the Eagles were doing well. Respectively, the two-samples used for each season's t-test had different sample sizes and different variance, giving us reason to use the Welch t-test.

Results

Each of the t-tests tested for the mean difference in crime frequencies and was dependent on the presence or absence of a game day. For each season, the t-test performed highlighted the direction of the difference in crime frequency on game days and non-game days. Generally, game days were not only significantly different in their mean crime frequency from non-game days, but there were also significantly less frequencies of crime than on non-game days (Table 1). Due to our low P-Values, we can say that the results indicate a rejection of our null hypothesis. We also examined each of the fourteen seasons individually to compare the means between the two categories. For 13 of the 14 seasons we analyzed, there are significantly fewer crimes committed on average on game days compared to non-game days (Table 1).

Year	Mean of Non Game days	Mean of Game Days	Mean Diff.	P-Value
2006	645.97	607.17	38.8	0.16
2007	608.76	548.56	60.2	0.00028
2008	571.85	503.63	68.22	0.00051
2009	531.9	466.76	65.14	0.0075
2010	524.93	459.18	65.75	0.0012
2011	523.69	453.75	69.94	0.0002
2012	519.22	449	70.22	2.23e-05
2013	493.83	428.41	65.42	0.004
2014	532.05	449.06	82.99	0.00099
2015	477.54	405.31	72.23	0.00089
2016	462.44	412.75	49.69	0.0065
2017	445.52	370.26	75.26	0.00052
2018	442.19	385.39	56.8	0.0011
2019	465.19	391.47	73.72	0.0003
All Years	516.6169	452.285	64.3319	2.2e-16

Table 1: The means of game days and non game days for each year. This also includes the difference in the means for each year.

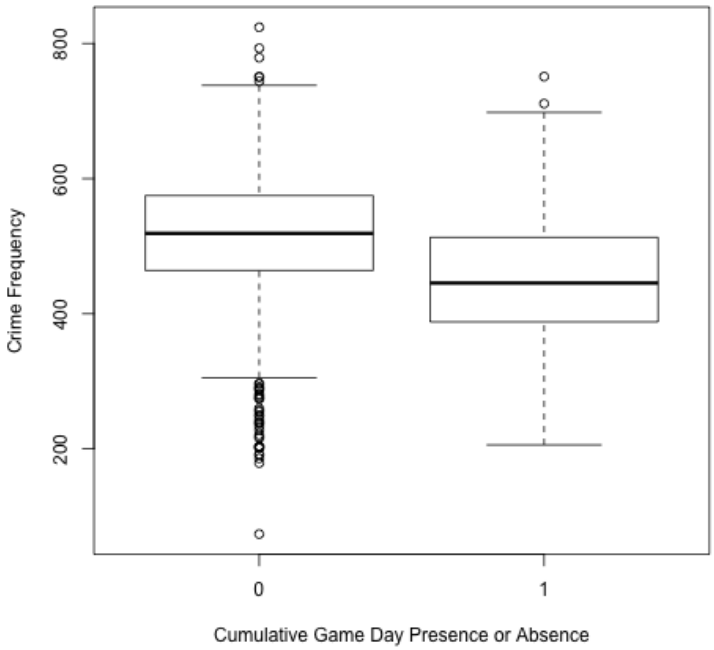

Figure 1: Box Plot for the all Eagles seasons from 2006 - 2019. 0 = Non-game days, 1 = game days.

Assumptions

Our analysis is contingent upon various assumptions. Firstly, we ignore the presence of any variables that would potentially confound our data. Daily crime frequency is certainly influenced by an immense range of variables. These could include variables such as game days of other sports teams, the implementation of new laws, increased policing, or various illicit activities. We also make the assumption that the overall frequency of crimes is affected by the presence of a game.

Discussion

The scope of our study was confined by the analysis we performed. Our analysis fails to explain the inter variation of crime frequency within game days themselves. A common trend we noticed in the data which was not quantified was that game days that had Eagles losses had higher crime frequencies than game days with Eagles wins. In 2013, PhillyMag reported statistics on the crime rates in response to the incidence of losing or winning a game. The article found that crime rates nearly tripled when the Eagles lost games compared to when they win in a sample of 64 game days (McManus 2013). Our analysis only looked at the presence or absence of a game as a predictor for the mean difference crime frequencies. Finally, effect sizes seemed to become more prominent with the increase sample sizes. We received smaller P-Values the larger the samples were, with the overall t-test (years 2006 - 2019) having the largest sample size and accounting for the biggest effect picked up by a t-test (p>2.2e-16).

2006 is the only year with a higher P-Value than our confidence interval. This may be due to several factors outside the scope of our study. The most prominent reason is due to a higher mean crime rate in 2006. As seen in our table, the mean frequency of crimes in 2006 is higher compared to other years. According to Reuters, homicides in Philadelphia rose by 7 percent in 2006 (Hurdle 2007). This indicates that violent crimes were greatly inflated in our 2006 season, which affected our 2006 season data as a whole. Therefore, we can assume that the presence of games had no significant effect on crime rates during the 2006 season due to overall inflation of crime during that period of time in Philadelphia.

References

Campitelli, E., & Dunne, E. (2016, May 17). Eagles fans named the most hated in the NFL, obviously. Retrieved from https://www.nbcsports.com/philadelphia/the700level/eagles-fans-named-most-hated-nfl-obviously

McManus, T. (2013, December 6). Wake-Up Call: How the Eagles Impact the Crime Rate. Retrieved from https://www.phillymag.com/birds247/2013/12/06/wake-call-8/

Hurdle, J. (2007, January 31). On streets of Philadelphia, crime is back. Retrieved from https://www.reuters.com/article/us-usa-crime-philadelphia/on-streets-of-philadelphia-crime-is-back-idUSN2545126720070131

Appendix

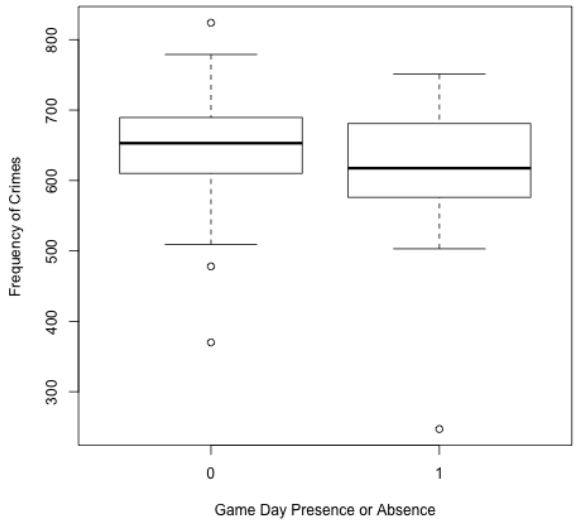

Figure 2: Box Plot for the 2006 Season. 0 = Non-game days, 1 = game days.

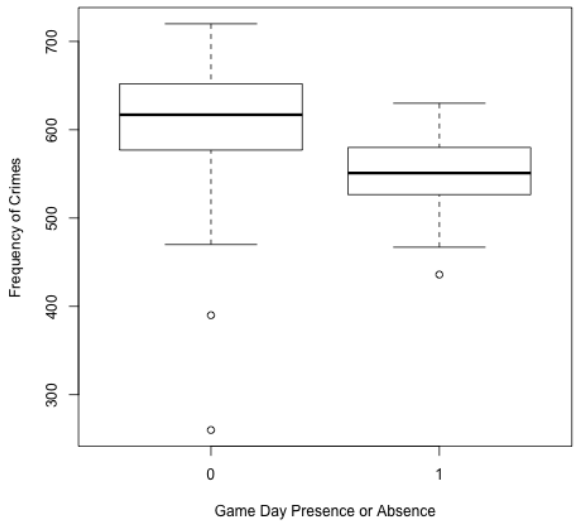

Figure 3: Box Plot for the 2007 Season. 0 = Non-game days, 1 = game days.

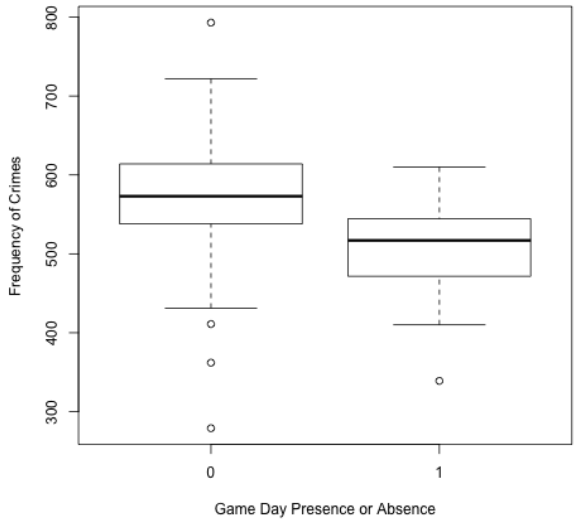

Figure 4: Box Plot for the 2008 Season. 0 = Non-game days, 1 = game days.

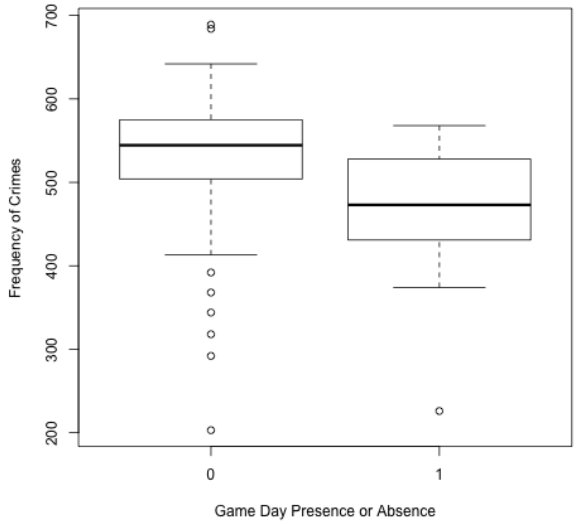

Figure 5: Box Plot for the 2009 Season. 0 = Non-game days, 1 = game days.

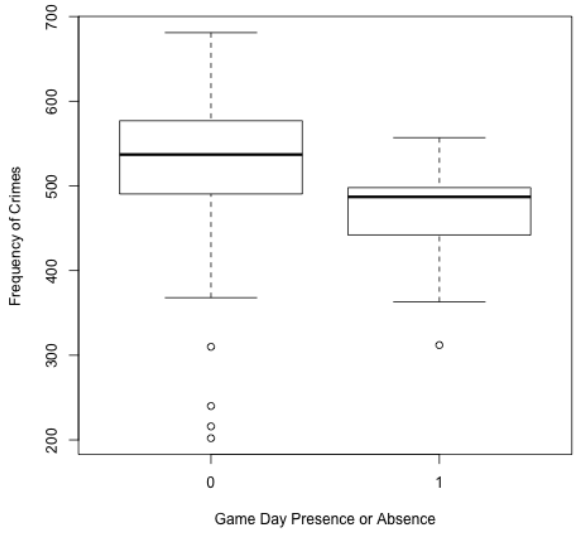

Figure 6: Box Plot for the 2010 Season. 0 = Non-game days, 1 = game days.

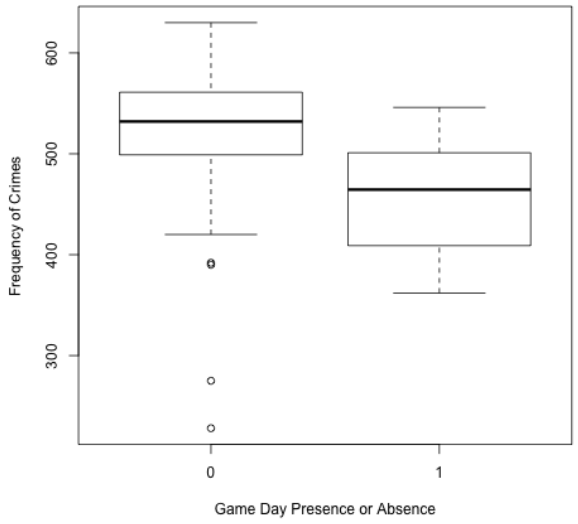

Figure 7: Box Plot for the 2011 Season. 0 = Non-game days, 1 = game days.

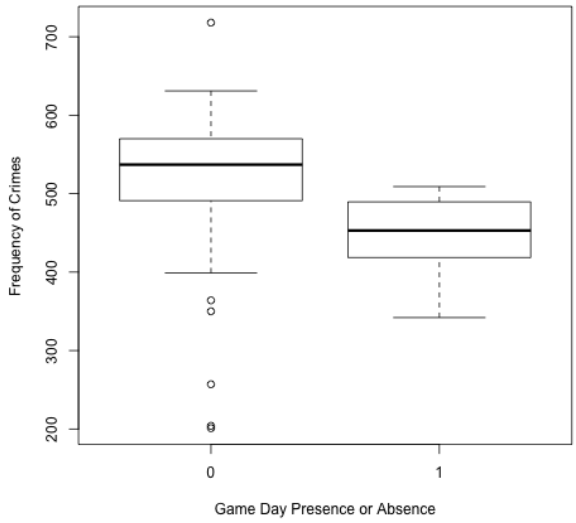

Figure 8: Box Plot for the 2012 Season. 0 = Non-game days, 1 = game days.

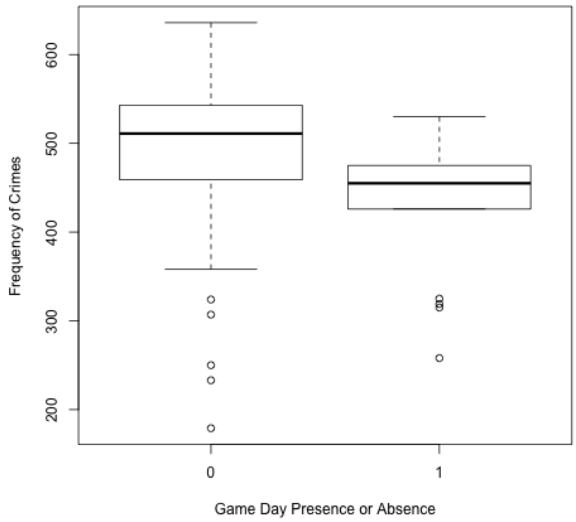

Figure 9: Box Plot for the 2013 Season. 0 = Non-game days, 1 = game days.

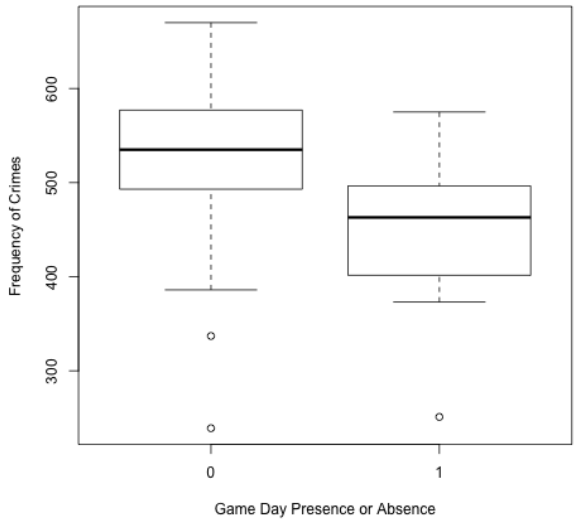

Figure 10: Box Plot for the 2014 Season. 0 = Non-game days, 1 = game days.

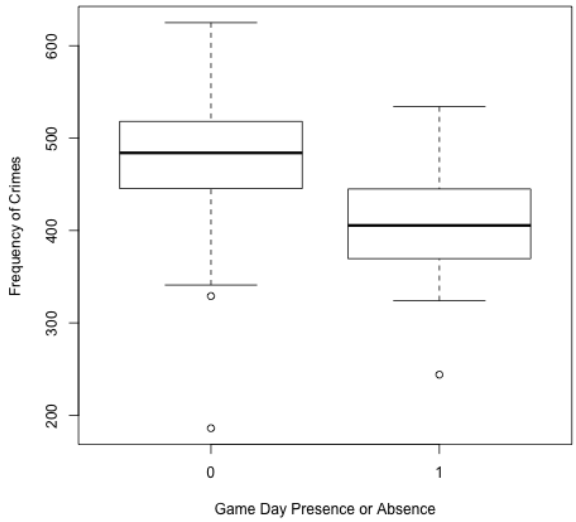

Figure 11: Box Plot for the 2015 Season. 0 = Non-game days, 1 = game days.

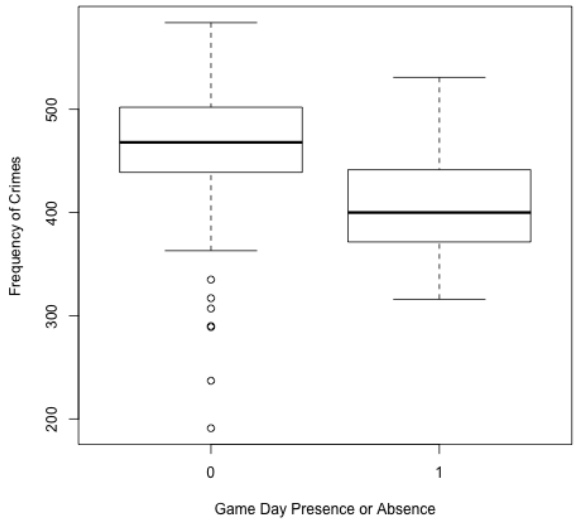

Figure 12: Box Plot for the 2016 Season. 0 = Non-game days, 1 = game days.

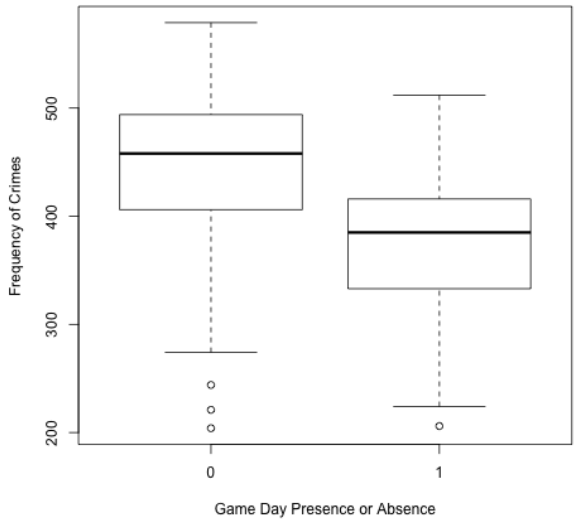

Figure 13: Box Plot for the 2017 Season. 0 = Non-game days, 1 = game days.

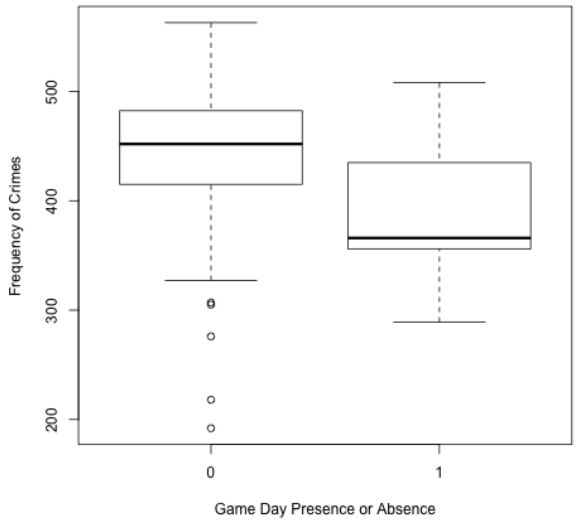

Figure 14: Box Plot for the 2018 Season. 0 = Non-game days, 1 = game days.

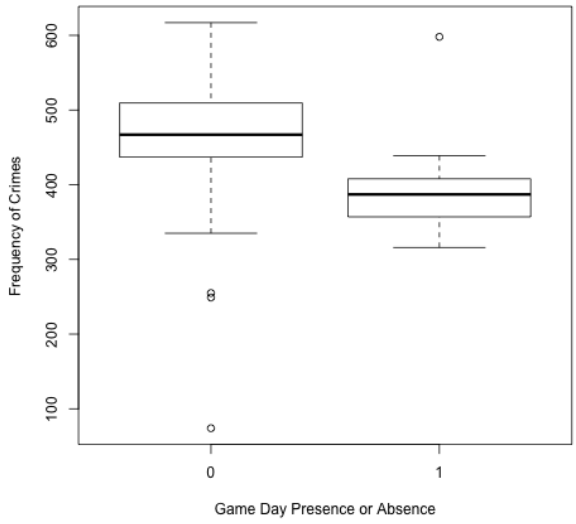

Figure 15: Box Plot for the 2019 Season. 0 = Non-game days, 1 = game days.