Dodgers Marketing Promotion

Gabriel Valenzuela

6/20/2020

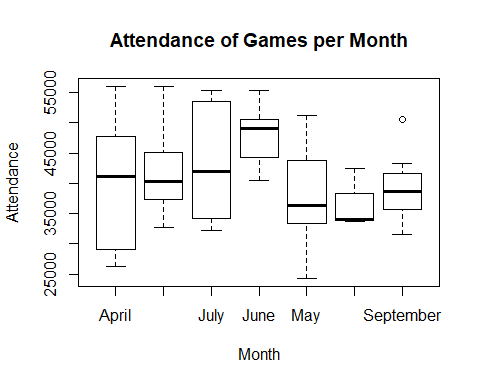
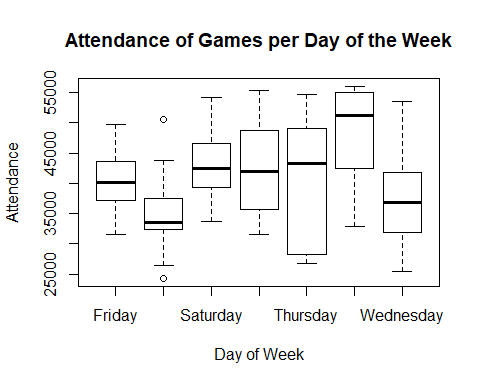
# 2012 Dodgers Game Attendance and Promotions Data and Data Structure

## Warning: package 'ggplot2' was built under R version 3.6.2

## month day attend day\_of\_week opponent temp skies day\_night cap shirt  
## 1 April 10 56000 Tuesday Pirates 67 Clear Day 0 0  
## 2 April 11 29729 Wednesday Pirates 58 Cloudy Night 0 0  
## 3 April 12 28328 Thursday Pirates 57 Cloudy Night 0 0  
## 4 April 13 31601 Friday Padres 54 Cloudy Night 0 0  
## 5 April 14 46549 Saturday Padres 57 Cloudy Night 0 0  
## 6 April 15 38359 Sunday Padres 65 Clear Day 0 0  
## 7 April 23 26376 Monday Braves 60 Cloudy Night 0 0  
## 8 April 24 44014 Tuesday Braves 63 Cloudy Night 0 0  
## 9 April 25 26345 Wednesday Braves 64 Cloudy Night 0 0  
## 10 April 27 44807 Friday Nationals 66 Clear Night 0 0  
## fireworks bobblehead  
## 1 0 0  
## 2 0 0  
## 3 0 0  
## 4 1 0  
## 5 0 0  
## 6 0 0  
## 7 0 0  
## 8 0 0  
## 9 0 0  
## 10 1 0

## 'data.frame': 81 obs. of 12 variables:  
## $ month : Factor w/ 7 levels "April","August",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ day : int 10 11 12 13 14 15 23 24 25 27 ...  
## $ attend : int 56000 29729 28328 31601 46549 38359 26376 44014 26345 44807 ...  
## $ day\_of\_week: Factor w/ 7 levels "Friday","Monday",..: 6 7 5 1 3 4 2 6 7 1 ...  
## $ opponent : Factor w/ 17 levels "Angels","Astros",..: 13 13 13 11 11 11 3 3 3 10 ...  
## $ temp : int 67 58 57 54 57 65 60 63 64 66 ...  
## $ skies : Factor w/ 2 levels "Clear ","Cloudy": 1 2 2 2 2 1 2 2 2 1 ...  
## $ day\_night : Factor w/ 2 levels "Day","Night": 1 2 2 2 2 1 2 2 2 2 ...  
## $ cap : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ shirt : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ fireworks : int 0 0 0 1 0 0 0 0 0 1 ...  
## $ bobblehead : int 0 0 0 0 0 0 0 0 0 0 ...

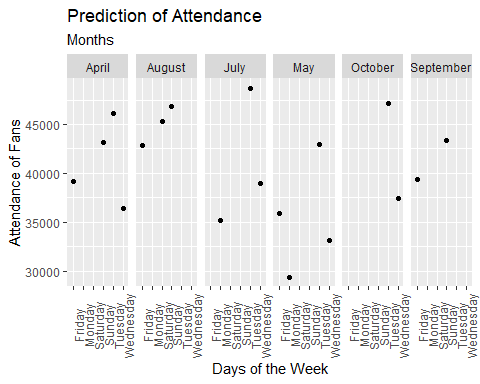
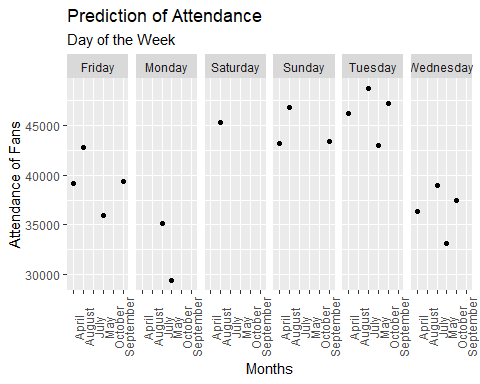
# Box Plots of Attendance of Games Per Day of the Week & Month



# Splitting the data for training and testing of model to be fitted for prediction of attendance

##   
## Call:  
## lm(formula = attendance\_model, data = trainSet)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14839.5 -5038.5 121.9 4108.5 13944.9   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 39122.0 3601.5 10.863 4.9e-14 \*\*\*  
## monthAugust 3673.9 3574.9 1.028 0.3097   
## monthJuly 2569.2 3985.4 0.645 0.5225   
## monthJune 8113.0 3712.9 2.185 0.0343 \*   
## monthMay -3201.9 3601.7 -0.889 0.3788   
## monthOctober 1032.3 7845.5 0.132 0.8959   
## monthSeptember 237.8 3705.5 0.064 0.9491   
## day\_of\_weekMonday -6530.3 3303.9 -1.977 0.0544 .   
## day\_of\_weekSaturday 2526.6 3109.0 0.813 0.4208   
## day\_of\_weekSunday 4017.6 3449.4 1.165 0.2504   
## day\_of\_weekThursday -479.2 3909.2 -0.123 0.9030   
## day\_of\_weekTuesday 7032.3 3564.1 1.973 0.0548 .   
## day\_of\_weekWednesday -2744.7 3766.7 -0.729 0.4701   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6960 on 44 degrees of freedom  
## Multiple R-squared: 0.4467, Adjusted R-squared: 0.2958   
## F-statistic: 2.96 on 12 and 44 DF, p-value: 0.004239

# Facet Grid Scatter Plots of Predicted Attendance for Days of the week and Month



# Process and Recommendation

In applying a regression model on the 2012 Dodgers Game Attendance and Promotion data set, I was seeking to predict the day of the week and month with the lowest attendance that would call for more promotions for those set date games. Before beginning my analysis, I performed data preparation of the set within a Jupyter Notebook using Python. I removed initials of months, changed YES/NO to 1/0, and converted variables to be used in a better scope. By using the box plots initially on the given data, it allowed me to see the distribution of attendance between the months and the days of the week. The main challenge was to find the combination of the both since I was able to see solely months and solely the days from those visuals. Once I had an idea of which months and days, I randomly split the data for training and testing and created the model with the 70% of the allocated training data followed by using the remaining testing data to make predictions on attendance of games. By creating a multi-scatter plot for days of the week and months combined, I was able to see that in the month of May on Mondays have the lowest attendance for games. Therefore, I recommend that it would be best to run a marketing promotion for Mondays in May to increase attendance for those days and attendance overall for the team.