**Name:** Andrea Fox

**Date:** September 19, 2020

**Class:** DSC630 - T301 – Predictive Analytics

**Assignment:** 3.3 – Using Data to Improve a Marketing Promotion

**Introduction**

For this assignment we were supposed to use the dodgers.csv dataset to predict the best day/night to run a marketing promotion to increase attendance. It was up to us to decide if we wanted to recommend a specific date or simply to recommend a day of the week.

**Requirements**

* Use a combination of Python and R
* Display a boxplot
* Display a scatter plot
* Have a regression model

**Process**

Python:

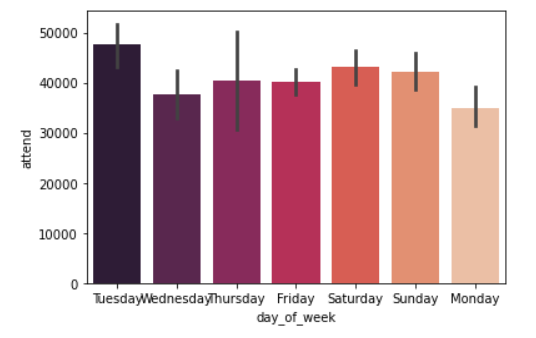
1. Read in the dodgers.csv file and used head() to verify it read in correctly
2. Used describe() to look at count, mean, standard deviation, and min/max
3. Checked for null values
4. Checked for missing values
5. Created bar plot using seaborn to look at attendance by day of week
6. Created bar plot to look at attendance by month
7. Created bar plot to look at attendance by day or night
8. Created bar plot to look at attendance by skies
9. Changed promotion values from Yes/No to 1/0
10. Looked at value counts for promotions
11. Created boxplots of promotions by attendance
12. Created scatter plot to look at attendance by day of week
13. Grouped promotions by day of week and attendance using mean values

R

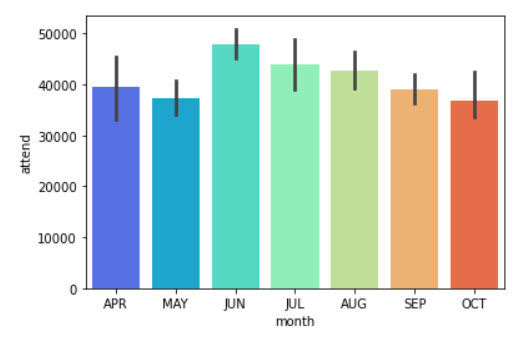
1. Loaded in the file dodgers.csv
2. Used glimpse() to get a picture of what was in the data
3. Split by dataset into train and test
4. Performed linear regression for attendance by month and day of the week
5. Performed linear regression for attendance by promotion type
6. Performed linear regression for attendance for bobblehead based on day of the week

**Analysis**

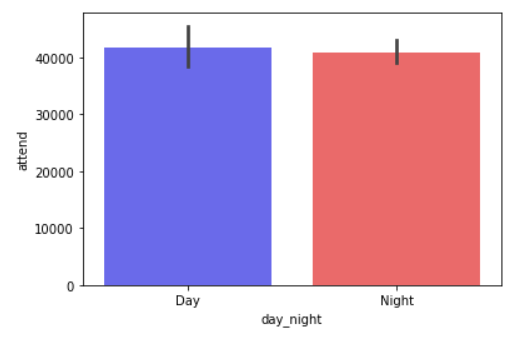
The first thing I really wanted to do when analyzing my data was to use some simple bar plots to get a picture of some features that I thought might play into my prediction.

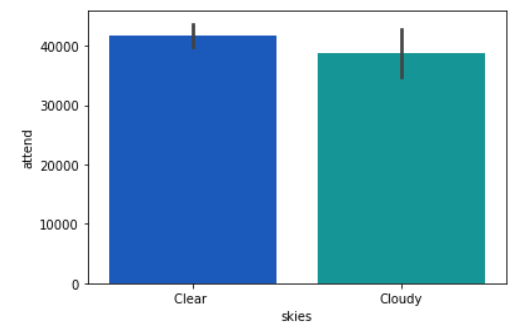


The first plot I looked at attendance by day of the week. Here I could visibly see that Tuesdays looked to be the most popular day of the week to attend a dodgers game. Monday does not look like it has a high attendance rating, and my initial thought was weekends would have the highest attendance rating but the plot proved me wrong. I then created a second bar plot to look at attendance by month.

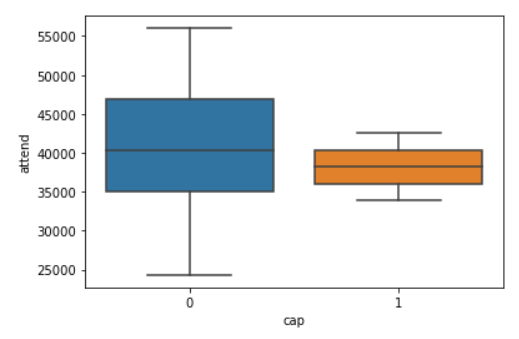


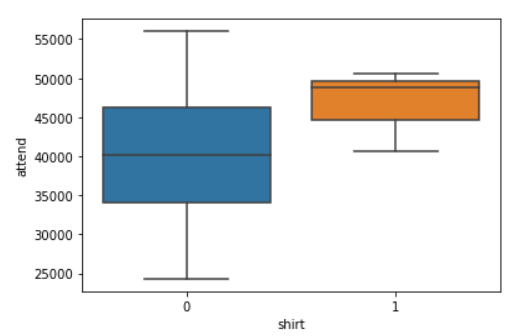
On this plot I could see June looked to be the most popular month to attend a game. My theory is June is usually a mild month compared to some of the others where you can really experience some extreme high and low temperatures. The lowest attendance ratings occurred in October and May. I wager October would be due to play-offs and not all teams playing in them but have no data to prove this theory. The last two bar plots I did was look at attendance by day/night and by skies.

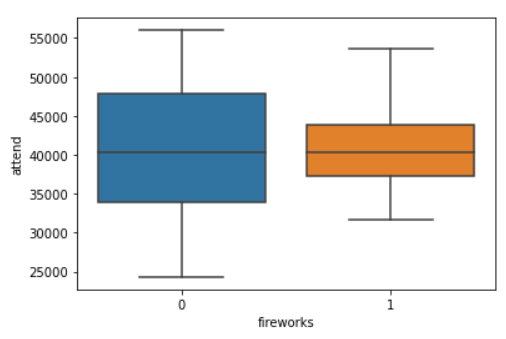


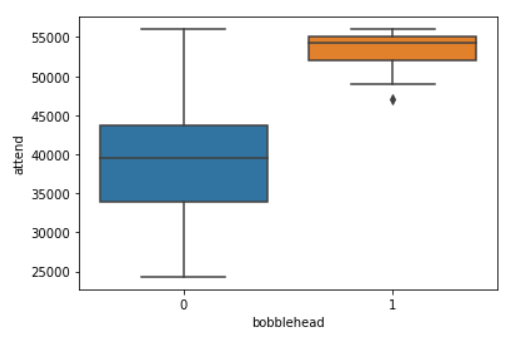


Looking at both of these plots there really isn’t a significant difference between the two, so I concluded day\_night and skies would not play a role in my prediction. This is where I decided to change the Yes/No into numeric values so I could get counts on each. Looking at just value counts I found that cap had 79 no’s and 2 yes’s, shirt had 78 no’s and 3 yes’s, fireworks had 67 no’s and 14 yes’s, and bobblehead had 70 no’s and 11 yes’s. Looking at these values it does not look like the shirt and cap promotions are used very often, while fireworks and bobblehead promotions happen more frequently. I then wanted to look at attendance against each promotion to see if attendance was higher for any specific promotion.

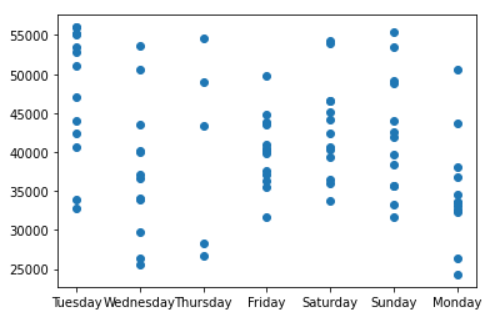








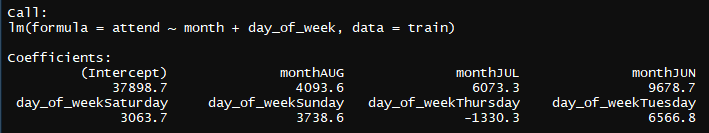
Based on the box plots, bobblehead had much higher attendance when the promotion was offered. Shirts came in second with the most attendance, while cap and fireworks did not have a noticeable attendance increase when promotions were offered. The last thing I wanted to do was create a scatter plot of attendance by day of week just to see some of the individual points.

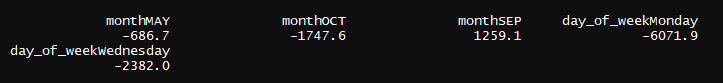


Here I could see a high concentration of higher attendance on Tuesdays followed closely by Sundays. Mondays and Fridays looked to have a lower concentration of attendance. I then moved over to Rstudio to do my dataset split and perform linear regression.

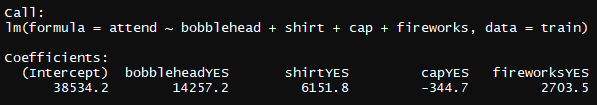
#Splitting the data into train and test  
dt = sort(sample(nrow(dodgers), nrow(dodgers)\*.7))  
train <- dodgers[dt,]  
test <- dodgers[-dt,]  
  
#Verify that it split randomly  
train  
test  
  
#Linear regression for attendance by month and day of the week  
lm(attend~month+day\_of\_week, data = train)  
  
#Linear regression for attendance by promotion type  
lm(attend~bobblehead+shirt+cap+fireworks, data = train)  
  
#Linear regression for attendance for bobblehead based on day of the week  
lm(attend~bobblehead+day\_of\_week, data = train)

I added the data for your convenience but will also be submitting along with his document. I split the dataset into train and test then performed a few linear regressions. First, I looked at attendance by month and day of the week. Looking at the output to me it looked like June and Tuesdays had the highest attendance.

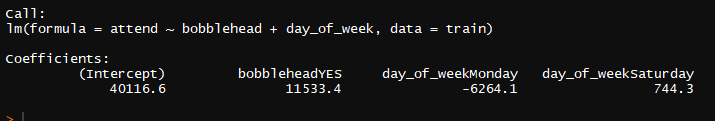


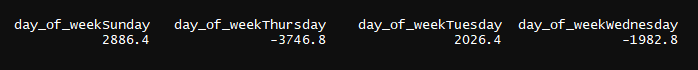


I then tried looking at attendance by promotion type. On this bobblehead had the most attendance compared to the other promotion types.



Last I ran attendance by bobblehead and day of the week. On this Sunday performed the best with Tuesday following.





After reviewing all my data, I would run my promotion marketing on Tuesdays based off attendance alone. If I want to be more specific, I would run it on Tuesdays in June.