**Using Data to Improve a Marketing Promotion**

**Week 3**

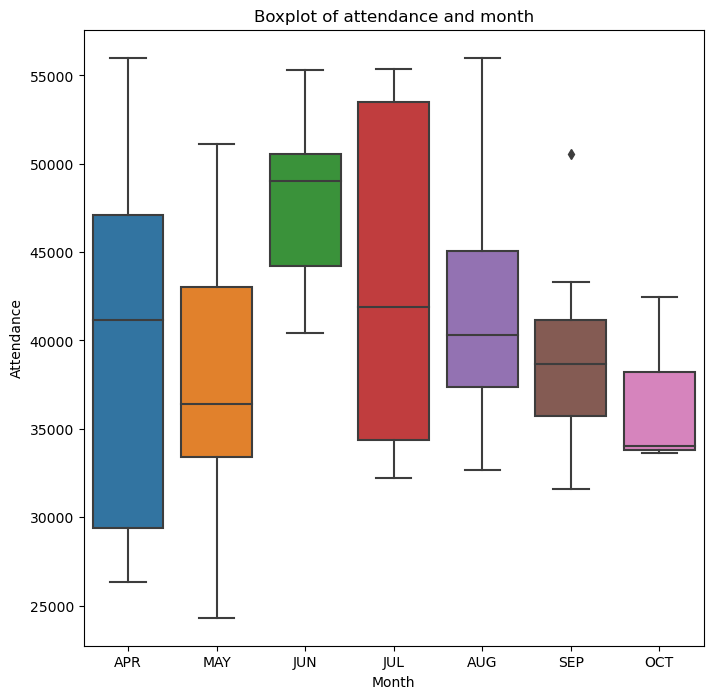
**DSC 630**

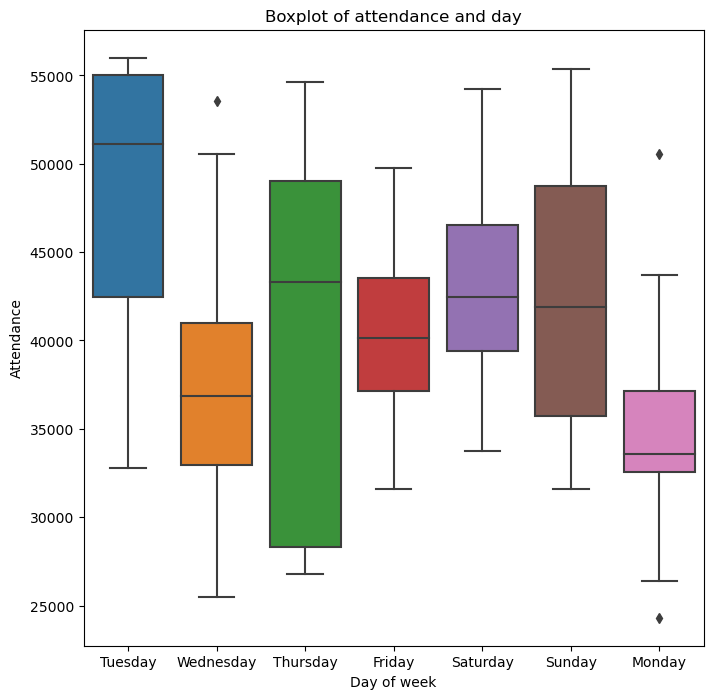
For this week’s assignment we're going to use Dodgers Major League Baseball data from 2012. The data file you will be using is contained in the [dodgers.csv](http://content.bellevue.edu/cst/dsc/630/dodgers.csv) file. I would like you to determine what night would be the best to run a marketing promotion to increase attendance. It is up to you if you decide to recommend a specific date or if you recommend a day of the week (e.g., Tuesdays) or month and day of the week (e.g., July Tuesdays). Use R and/or Python to accomplish this assignment. It is important to remember, there will be lots of ways to solve this problem. Explain your thought process and how you used various techniques to come up with your recommendation. From this data, at a minimum, you should be able to demonstrate the following:

Box plots

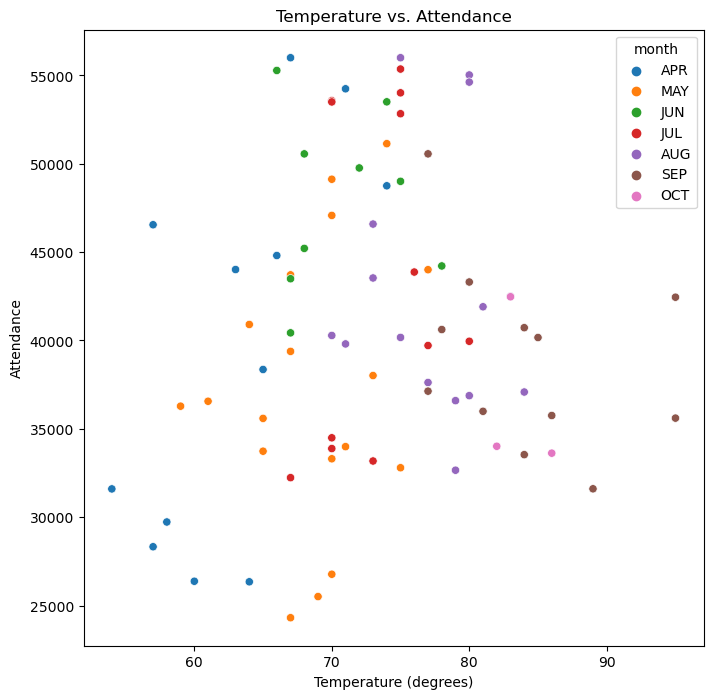
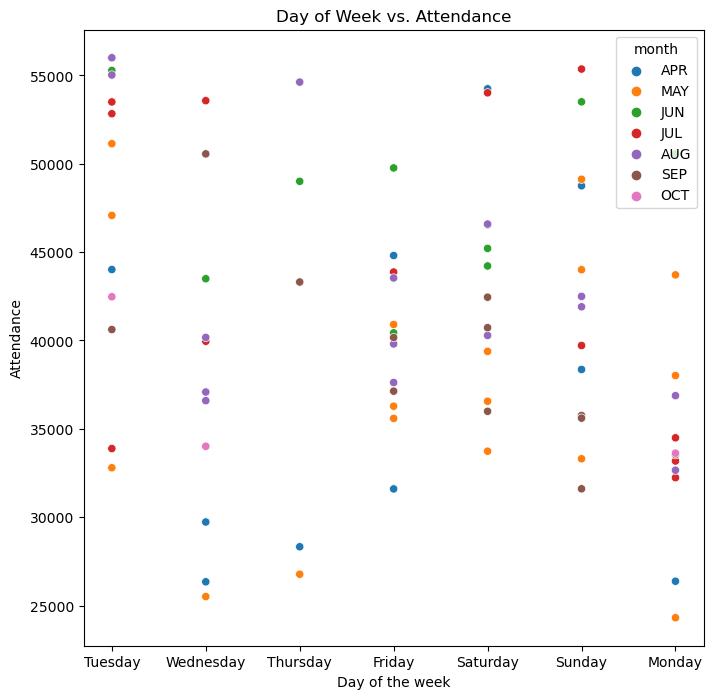
Scatter plots

Regression Model

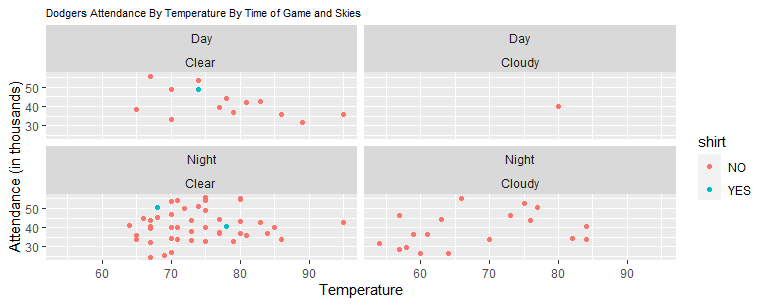
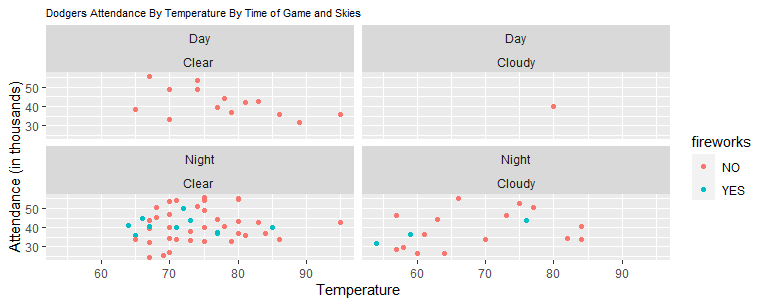
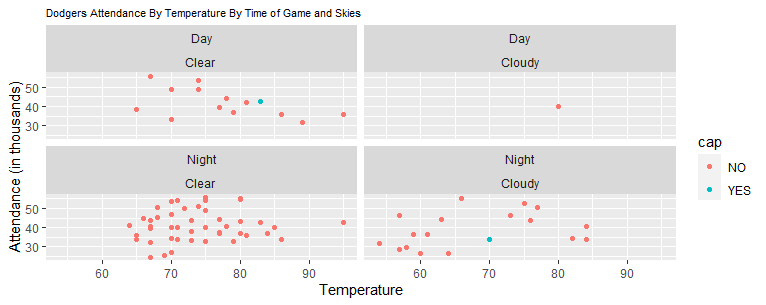
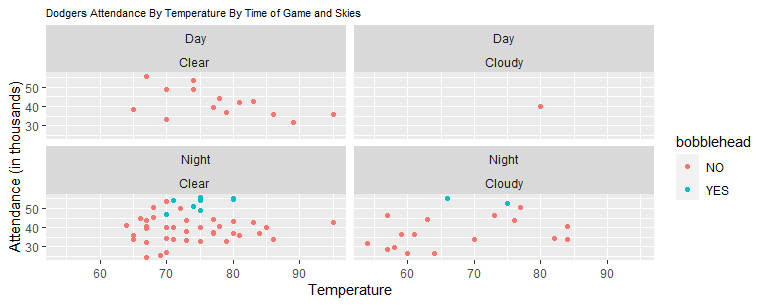




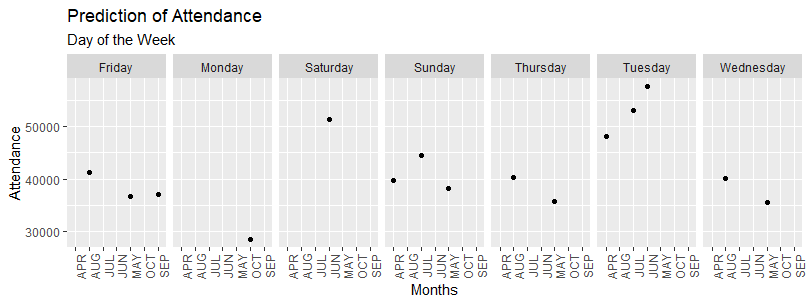
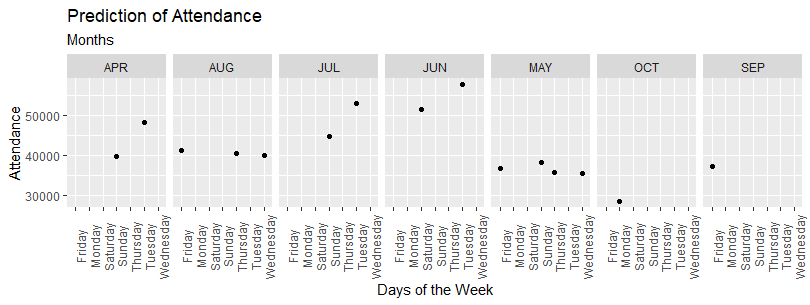
Based on above data with both boxplot it seems that June has the highest attendance with mean being 49,940 attendees. June has the high range of attendees ranging from 40,432 to maximum of 55,279. In another boxplot we can see that Tuesday has the highest attendance lowest being around 33,000 and highest being 57,000. There are outliers in both the boxplots. In the first one it is on September whereas in the second one we have 3 outliers; one occurring on Wednesday and other two on Monday.



Above we have two scatterplots; first being day of the week and second being temperature and y axis variable being attendance. As we can see that on day of the week Thursday had the lowest attendance and Sunday had the higher attendance whereas with temperature; when it came between 70-80 degree Fahrenheit, there was higher occupancy with the lowest ranging from 30k plus. On the other hand, October as the month has the lowest occupancy compared to other months. This can be due to beginning of winter season compared to April till September.



From above we have 4 scatterplots, which is taken from R and used geom\_point() to create the scatterplot. I needed to know how the promotional type worked based on day/night with variable such as temperature. We had higher attendance when there was bobblehead promotion and the sky was clear, cap doesn’t seem to have much effect on attendance since there was only one cap promotion and it does seem to alter the attendance.



**PROCESS:**

I used both Python and R for this assignment which can be viewed from the attested GitHub link. Before beginning my analysis, I performed data preparation of the set within Jupyter Notebook using Python. I converted promotional type variable from YES/NO to 1/0. I applied regression model at the end of the coding process in R. I used month and day of the week as model parameters. I randomly split the data for training and testing and created the model with the 80% of the allocated training data followed by using the remaining testing data to make predictions on attendance of games. From above boxplots and scatter plots we can see that Monday has the lowest attendance among all other days of the week and October as the month having the lowest attendance. We can also see from prediction graph (day of the week) that on October month, Monday is likely to have the lowest attendance based on our model. Even on the graph (day of the week) Monday on October has the likely lower attendance of all other days and months. Using all above graphs/charts and prediction I came to conclusion that it would be best to run a marketing promotion for Mondays in May to increase attendance for those days and attendance overall.