**Summary Report**

**Professor: Fadi Alsaleem**

Student: Pritam Shrestha

Date:06/24/2020

Language: Python and R

File1: assignment\_3.3\_using\_python( used for python)

File2: data\_modeling\_and\_plotting\_using\_r( used notebook for r command for better visualization)

For this assignment, I have used two programming languages Python and R. First of all, I have loaded the data using the pandas data frame and displayed the structure of the data. Data is a very important part of any project so I have checked, missing data, duplicates, and types of data for further analysis. After seeing its structure and shape, I did little iteration on data variables to count its values and data types. And plotted them into a histogram.

There is no problem with data quality. It means there are no missing values and unusual inputs. Besides that, I have found a couple of categorical data. I planned to perform some data conversion while creating a model and fitting model. So I decided to do that part using R and saved the data.

In R, I have performed the following operations

1. Loaded data : **DodgersData <- read.csv("DodgersData.csv")**Checked structure of data: **str(DodgersData)**
2. Evaluate the factor level of month, day\_of \_week, etc.
3. Boxplot: completed boxplot of each of them based on attending.

Based on the box plot I have found the following facts about it.

1. Promotion ranges from April to August
2. There are four types of promotions, e.g. cap, shirt, fireworks, and bubblehead.
3. Data is divided into seven days of the week.
4. The highest attendance is 56000.

Box plot is one of the best ways to find statistics of the data. Hence based on the attendance Tuesday was the best day and Monday was a worse day. Likewise, June was best and October was worse based on the attendance.

**Design Predictive Model:**

To measure the impact of a promotion on attendance, I have built a linear model. The intension was to predict the attendance using month, day of the week, and promotion (bubblehead) indicator. First of all, I have created divided main data into two datasets such as testing and training. To evaluate the effect of the month, day of the week, and bubblehead on attendance, I have created a model and fitted those data into two sets of newly created datasets. Besides that, I have predicted the values on the training set as well as calculated accuracy based on the test set.

A summary is the best way to find the statistics of the datasets so I have gone through each part of the summary report fitted model of main data and found the following facts.

1) T-value: The coefficient t-value is a measure of how many standard deviations our coefficient estimate is far away from 0. In this case, overall t-values are not that far away from 0, meaning, a relationship exists between the variables.

2) p-value: The p-value is significantly low it means true relationship exists between the variables.

3) R-Squared: The R-squared value is which means 54.4% of the variance found in the response variable can be explained by the predictor variable.

Looking at all the matrix, I used another way to look at the highest attendance day using the following formula.

which.max(DodgersData$attend)=[1] which means,

It displays array [1], it means the highest value is in the first index

Max value=56000

Based on the available value we can conclude that

Day =10, month=1(January) and day\_of\_week is 6

Hence, the actual data of maximum attendance to watch the game is Friday, January 10.

If we do marketing promotion on that day we might get more attendance in the future.

Thanks for letting us resubmit!!!!