Business Analyst Intern Technical Interview

```
## Warning: package 'caret' was built under R version 3.6.2
## Loading required package: lattice
## Loading required package: ggplot2
## Warning: package 'ggplot2' was built under R version 3.6.2
## Warning: package 'dplyr' was built under R version 3.6.2
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Q1. Create a dataset of game-by-game home attendance for FC Cincinnati's 2019 season using at least day of week, time of game, win percentage, and opponent.

```
df <- read.csv('match-by-match-report.csv')

print(names(df))
```

```
##  [1] "Round"        "Day"          "Date"         "Time"         "Venue"
##  [6] "Result"       "GF"           "GA"           "Opponent"     "xG"
## [11] "xGA"          "Attendance"   "Captain"      "Formation"    "Referee"
## [16] "Match.Report" "Notes"
```

We definitely dont need the first (Round) and the last two columns (Match Report and Notes) since they have no relation to the fan attendence Thus, we can remove them from our dataset.

Additionally, variables like Captain, Formation and Refree shouldn't really affect the attendence in anyway (Though formation and results might be correlated, and the attendence is usually related to the results, so we could say formation is correlated as well, but for this particular case we will not consider that complexity)

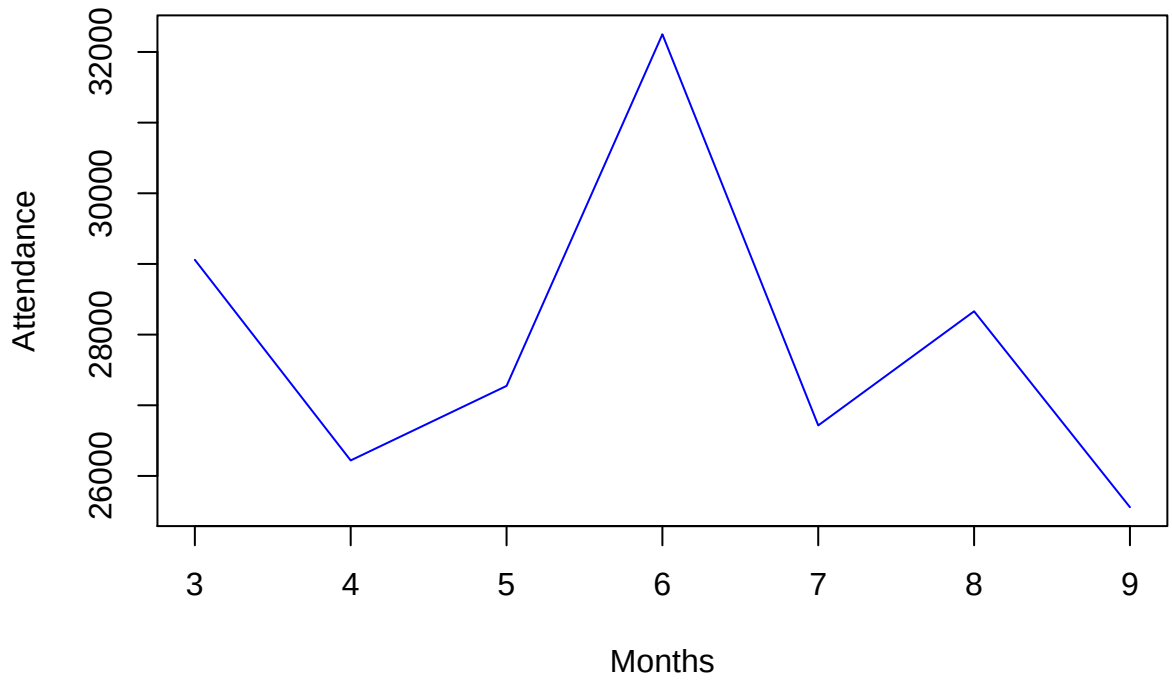
Finally, we are only conisdering home games here. Thus, we can get rid of the away games from our dataset.

```
#Removing columns - round, match report and notes
df = df[,2:15]

#Removing columns - formation and referee
df = df[,1:11]

#Removing rows for home games
df = df[df$Venue!='Away',]

head(df)
```

```
##    Day       Date          Time Venue Result GF GA       Opponent  xG xGA
## 3  Sun 2019-03-17 17:00 (14:00)  Home      W  3  0       Portland 1.5 1.0
## 5  Sat 2019-03-30 19:30 (16:30)  Home      L  0  2   Philadelphia 0.4 2.0
## 6  Sun 2019-04-07 15:00 (12:00)  Home      D  1  1    Sporting KC 1.8 2.4
## 8  Fri 2019-04-19 19:30 (16:30)  Home      L  0  3 Real Salt Lake 0.9 2.1
## 12 Sat 2019-05-11 13:00 (10:00)  Home      W  2  1       Montreal 0.7 1.2
## 14 Sat 2019-05-25 19:30 (16:30)  Home      L  0  2   NY Red Bulls 1.8 1.1
##    Attendance
## 3       32250
## 5       25867
## 6       26023
## 8       26416
## 12      26258
## 14      28290
```

Let us now check how the attendance for the game varied over the course of the year.

```
df$ModifiedDate <- sapply(df$Date, function(x) (as.character(x)))
df$Month <- sapply(df$ModifiedDate, function(x) substr(x,6,7))

groupedbymth <- aggregate(df[,"Attendance"], list(df$Month), mean)
plot(groupedbymth$Group.1, groupedbymth$x,
     type = "l", xlab = "Months"
     , ylab = "Attendance", col="blue")
```



The attendance for home games rose up to 32000 by June, which could be due to the summer break.

Q2. Determine whether or not weekend games have a statistically significant impact on attendance and

construct a game-by-game attendance forecast model for the 2020 season.

Note: In my analysis, I am considering the term weekend as 'Saturday' and 'Sunday'. I will next encode my Day variable into 1s for weekend and 0s for weekday.

Also, since our data points for home games played in 2019 are very low, I have decided to append the data for home attendance for FC Cincinnati's home games in the USL as well (since we are only focusing on the Day Of The Week and Attendance)

```
#Reading in data
homeattendance <- read.csv('FCCincinnatiHomeAttendance.csv')

#Getting rid of the away games
homeattendance <- homeattendance[df$Venue!='Away',c('Day','Attendance')]
homeattendance <- rbind(homeattendance, df[,c('Day','Attendance')])

homeattendance <- homeattendance %>% mutate(Day = ifelse(Day == "Sat" | Day == "Sun",1,0))
homeattendance$Day <- as.factor(homeattendance$Day)

#Dummy encoding the day variable
dmy <- dummyVars(" ~ .", data = homeattendance)
trsf <- data.frame(predict(dmy, newdata = homeattendance))
colnames(trsf)[1] <- "NotWeekend"
colnames(trsf)[2] <- "Weekend"
head(trsf)
```

```
##   NotWeekend Weekend Attendance
## 1          0       1       4598
## 2          0       1      17535
## 3          0       1      25667
## 4          0       1       1418
## 5          0       1      24505
## 6          0       1       2785
```

```
weekendimpact <- lm(Attendance~., data=trsf)
summary(weekendimpact)
```

```
##
## Call:
## lm(formula = Attendance ~ ., data = trsf)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -17129 -11601   3373   8525  15832
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    17742       1282  13.840   <2e-16 ***
## NotWeekend     -4801       3015  -1.592    0.115
## Weekend           NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10570 on 81 degrees of freedom
## Multiple R-squared:  0.03034,    Adjusted R-squared:  0.01837
## F-statistic: 2.534 on 1 and 81 DF,  p-value: 0.1153
```

Thus, according to our model, the impact of a weekdays on the attendance is not statistically significant.

Before we build a forecasting model, let us add a few more variables to our data that logically would affect the attendance for a particular game.

The attendence for a game usually depends a lot on the opponent quality. To check this, we can include opponent quality in our analysis. To calculate the opponent quality, I will consider two metrics for the last three seasons - Win Percentage and Goals Scored per game in the last three seasons.

The reason I chose these two metrics is:

Win Percentage - This metric usually tells us about how well has the team played over the duration of time. Since I am considering the past three season data, I am also taking into consideration the consistency of their performances.

Goals Per Game - The other reason why crowds love watching game is the team the home team is playing against. If the opponent is a team that scores a lot of goals, it usually means that the team is very exciting to watch.

```
#Import previous 3 season tables
table2017 <- read.csv('MLS-2017-Table.csv')
table2018 <- read.csv('MLS-2018-Table.csv')
table2019 <- read.csv('MLS-2019-Table.csv')

#Combining all the 3 season data into one dataframe
prev3season <- rbind(table2017[,c('Squad','MP','W','D','GF')],
      table2018[,c('Squad','MP','W','D','GF')],
      table2019[,c('Squad','MP','W','D','GF')])

#Grouping data by the team
aggTeamData <- aggregate(as.matrix(prev3season[,2:5]),
                         by=list(Team=prev3season$Squad),
                         FUN = sum)

#Win Percentage for each team
aggTeamData$WinPct <- (aggTeamData$W + (0.5 * aggTeamData$D))/aggTeamData$MP

#Goals scored by the team per game
aggTeamData$GoalPerGame <- (aggTeamData$GF)/aggTeamData$MP

#Getting a list of the top performing teams according to our metrics
aggTeamData[order(-aggTeamData$WinPct),]
```

```
##             Team  MP  W  D  GF    WinPct GoalPerGame
## 1        Atlanta 102 54 20 198 0.6274510   1.9411765
## 8          NYCFC 102 50 27 178 0.6225490   1.7450980
## 7   NY Red Bulls 102 50 19 168 0.5833333   1.6470588
## 11    Toronto FC 102 43 26 190 0.5490196   1.8627451
## 10  Philadelphia 102 42 21 157 0.5147059   1.5392157
## 3       Columbus 102 40 23 135 0.5049020   1.3235294
## 4    D.C. United 102 36 25 133 0.4754902   1.3039216
## 6    New England 102 34 29 152 0.4754902   1.4901961
## 2        Chicago 102 34 27 164 0.4656863   1.6078431
```

```
## 5       Montreal 102 37 15 146 0.4362745   1.4313725
## 9   Orlando City 102 27 23 126 0.3774510   1.2352941
## 12 FC Cincinnati  34  6  6  31 0.2647059   0.9117647
```

```
#Merging this dataframe into our main dataframe
df_merged <- merge(df[,c("Opponent","Day","Attendance")], aggTeamData[, c("Team","WinPct","GoalPerGame"
```

Build a forecasting model:

```
forcastmodel <- lm(Attendance~WinPct+GoalPerGame , data=df_merged)
summary(forcastmodel)
```

```
##
## Call:
## lm(formula = Attendance ~ WinPct + GoalPerGame, data = df_merged)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1406.6  -943.3   180.3   755.2  1533.5
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    29951       2631  11.383  3.2e-06 ***
## WinPct         27391       8298   3.301  0.01084 *
## GoalPerGame   -11055       2809  -3.935  0.00432 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1177 on 8 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.6597, Adjusted R-squared:  0.5746
## F-statistic: 7.754 on 2 and 8 DF,  p-value: 0.01341
```

We can see that for our model, the Goals Per Game variable and Win Percentage are statitically significant. Again, whether the day was a weekday or a weekend was surprisingly not making that big a difference, so I got rid of it. But then again, it could easily be because of the lack of data. The above model has an adjusted R-squared of 0.57, which is way better than the model we created using on the day of the week.

Now as a final step, we can predict the attendence for the upcoming season.

Q3.

Determine whether or not win percentage at the time of the game has a significant impact on attendance.

Assumption: My interpretation is that a win percentage at the time of a game is the number of games the team has won plus the 0.5 times number of games the team drew. This whole thing is then divided by the number of games played till the time of the game. These calculations are done in a cumulative way.

```
df <- df %>%
  mutate(EncodedResult = ifelse(Result == "W" ,1,ifelse(Result == "D" ,0.5,0)))
df_modified <- df[,c("EncodedResult","Attendance")]
df_modified$winratio <- cumsum(df_modified[, 1])/seq(nrow(df_modified))
head(df_modified)
```

```
##   EncodedResult Attendance  winratio
## 1           1.0      32250 1.0000000
## 2           0.0      25867 0.5000000
## 3           0.5      26023 0.5000000
## 4           0.0      26416 0.3750000
```

```
## 5           1.0      26258 0.5000000
## 6           0.0      28290 0.4166667
```

```
print(cor(df_modified$Attendance, df_modified$winratio))
```

```
## [1] 0.4607355
```

```
winpctmodel <- lm(Attendance~winratio, data = df_modified)
summary(winpctmodel)
```

```
##
## Call:
## lm(formula = Attendance ~ winratio, data = df_modified)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2102.4 -1599.3  -805.2   819.4  5135.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    24977       1285  19.443 4.75e-12 ***
## winratio        5986       2977   2.011   0.0627 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2153 on 15 degrees of freedom
## Multiple R-squared:  0.2123, Adjusted R-squared:  0.1598
## F-statistic: 4.042 on 1 and 15 DF,  p-value: 0.06271
```

The win percentage ratio has a decent correlation of 0.46 with the attendance. However, we clearly see that the winratio vairable is not statitically significant in determining the attendance values.

Q4. What other data, variables, or means of analysis would you consider for looking at this problem further?

1. Ideally, I would have loved to include the performance of the Orange and Blue against a particular team during the previous seasons to predict the amount of people attending a game against that team. However, since the 2019 season was the first season when FC Cincinnati played in the MLS, we cannot calculate that statistic.
2. Historically, games against rivals are the ones that have the maximum attendence (especially if it is a derby, since the people from both teams can attend the games). However, this belief does not hold true since the attendance for the home game against Columbus Crew (Cincinnati's rivals) was only 3rd in the ranking.

```
head(df[order(-df$Attendance),c('Day','Opponent','Attendance')])
```

```
##    Day     Opponent Attendance
## 1  Sun     Portland      32250
## 7  Sat    LA Galaxy      32250
## 13 Sun     Columbus      30611
## 9  Thu  D.C. United      28774
## 6  Sat NY Red Bulls      28290
## 12 Sat        NYCFC      27273
```

However, once possible justification for this that I can think of is that the game against Portland was the first home game of the season for FC Cincinnati. And given that this was their first home game of their MLS campaign as well, the attendance was expected to be the highest.

The possible explanation for the second highest attendance to be for the LA Galaxy game is that LA Galaxy

had ZLATAN last season!! And come on, who doesn't want to watch Zlatan play?

3. Another variable which I believe might affect the attendence would be the weather. If the weather was very bad (heavy rains, snow or winds), the attendence is expected to be negatively affected.

SIDE NOTE: I thoroughly enjoyed this take home assignment. Even more since I am a huge soccer fan myself. I'm new to MLS though, and the upcoming season would be the first season that I will be watching from start to end. I'm really looking forward to FC Cincinnati's performance this season. Good luck for your second season in the top flight!!