cuongAnalysis

###Cuong's data analysis using a two-way ANOVA with interactions model

###Introduction

The questions that I want to explore and address are:

1. Do we want to use promos at all?
2. If we decide to use promos, which game tiers would benefit from it the most?

In particular, I want to find out if there is a difference between the attendance number for when there is a promo and when there is not. Furthermore, I want to see if this difference is dependent on the game tier. In essence, I want to know if these two independent variables - game tier and promo status - affect each other to influence the attendance number.

With these 2 main questions in mind, let's proceed to the analysis.

###Data

Importing the necessary libraries

```
library(mosaic)
```

```
## Registered S3 method overwritten by 'mosaic':
##   method                           from
##   fortify.SpatialPolygonsDataFrame ggplot2
```

```
##
## The 'mosaic' package masks several functions from core packages in order to add
## additional features.  The original behavior of these functions should not be affected by this.
```

```
##
## Attaching package: 'mosaic'
```

```
## The following objects are masked from 'package:dplyr':
##
##     count, do, tally
```

```
## The following object is masked from 'package:Matrix':
##
##     mean
```

```
## The following object is masked from 'package:ggplot2':
##
##     stat
```

```
## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,
##     quantile, sd, t.test, var

## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v tibble  3.1.6     v purrr   0.3.4
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x mosaic::count()            masks dplyr::count()
## x purrr::cross()             masks mosaic::cross()
## x mosaic::do()               masks dplyr::do()
## x tidyr::expand()            masks Matrix::expand()
## x dplyr::filter()            masks stats::filter()
## x ggstance::geom_errorbarh() masks ggplot2::geom_errorbarh()
## x dplyr::lag()               masks stats::lag()
## x tidyr::pack()              masks Matrix::pack()
## x mosaic::stat()             masks ggplot2::stat()
## x mosaic::tally()            masks dplyr::tally()
## x tidyr::unpack()            masks Matrix::unpack()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(knitr)
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'

## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
library(readxl)
library(ggplot2)
library(DescTools)
```

```
## Warning: package 'DescTools' was built under R version 4.1.3
```

```
##
## Attaching package: 'DescTools'
```

```
## The following object is masked from 'package:mosaic':
##
##     MAD
```

Importing and cleaning the data set to use.

```
h1819 <- read_csv("hackathon-business-track/data/hackathon_2018_19_attendance.csv",
                  show_col_types = FALSE)
h2122 <- read_csv("hackathon-business-track/data/hackathon_2021_22_attendance.csv",
                  show_col_types = FALSE)
```

```
h1819 <- h1819 %>%
  janitor::clean_names() %>%
  mutate(game_date = mdy(game_date)) %>%
  #Code not functioning as desired. If there is something in promo,
  #then promo_status = 1, else = 0 but I got NA instead of 0
  mutate(promo_status = if_else(promo != "NA", 1, 0, 0))

h2122 <- h2122 %>%
  janitor::clean_names() %>%
  mutate(game_date = mdy(game_date)) %>%
  #Code not functioning as desired. If there is something in promo,
  #then promo_status = 1, else = 0 but I got NA instead of 0
  mutate(promo_status = if_else(promo != "NA", 1, 0, 0))

df <- rbind(h1819, h2122)
```

I decided to bind the two separate data sets into one data frame since this will give me more observations to work with. Furthermore, I don't think there would be a clear trend in attendance when they use promos in 2018-2019 vs in 2021-2022.

###Two-way ANOVA with interactions

###Hypotheses

Since this two-way ANOVA consider the effect of two categorical variables and the effect of the categorical variables on each other, there are three pairs of null or alternative hypotheses to consider.

H0: The means of promo status are equal (no difference between using a promo and not using)

HA: The means of at least one promo status is different

H0: The means of game tier are equal

HA: The means of at least one game tier is different

H0: There is no interaction between promo status and game tier

HA: There is interaction between promo status and game tier

Components of our ANOVA model:

#We can change this if needed

- Response variable (y): Attendance. Attendance and attendance at tip are relatively close to each other, so I decided to use attendance as the response variable
- Explanatory varialbe (x1): Game tier (from A to D, with A meaning the highest-ranked matches)
- Explanatory variable (x2): Promo status (1 for promo and 0 for none)

With these components, I craft my model as follows:

```
mod <- lm(total_attendance ~ game_tier * promo_status, data = df)
```

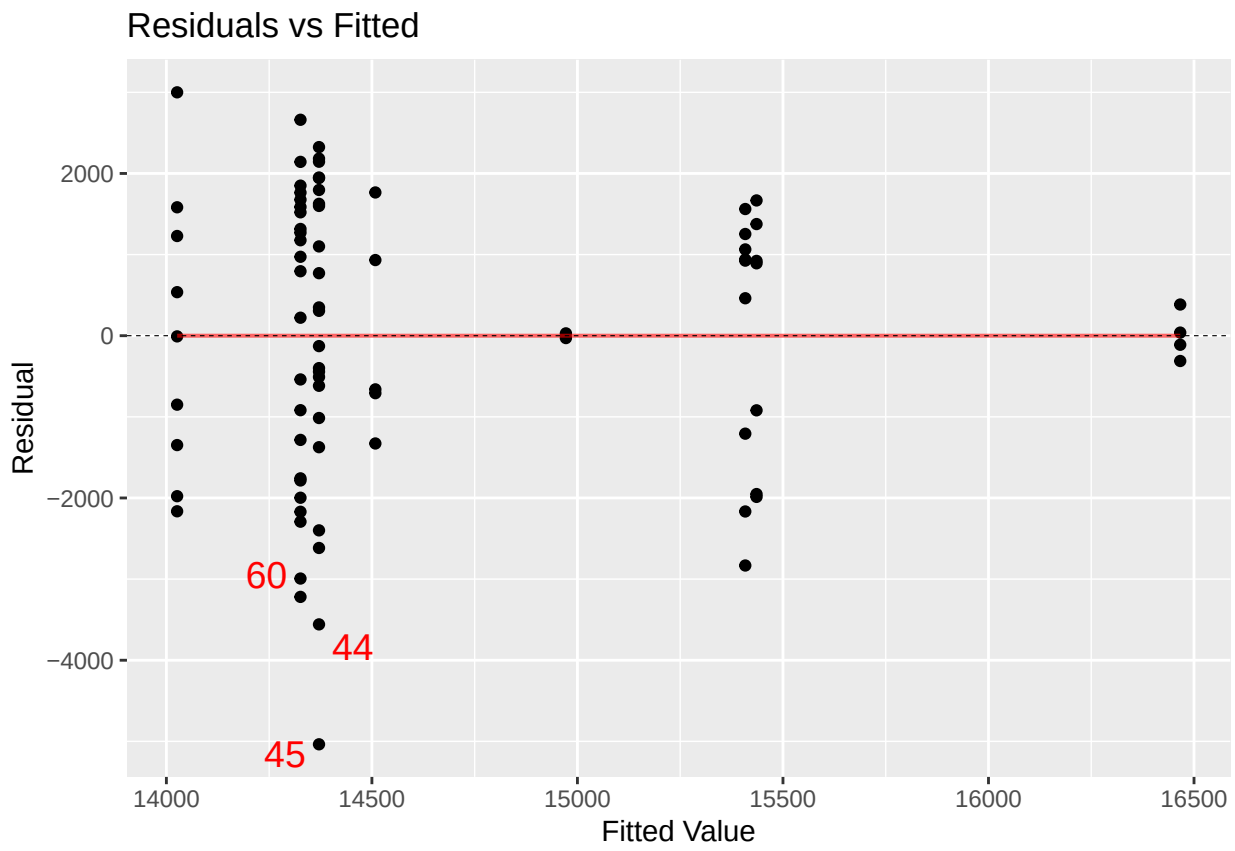
###Check conditions

There are 4 conditions to check if we want to proceed with a two-way ANOVA:

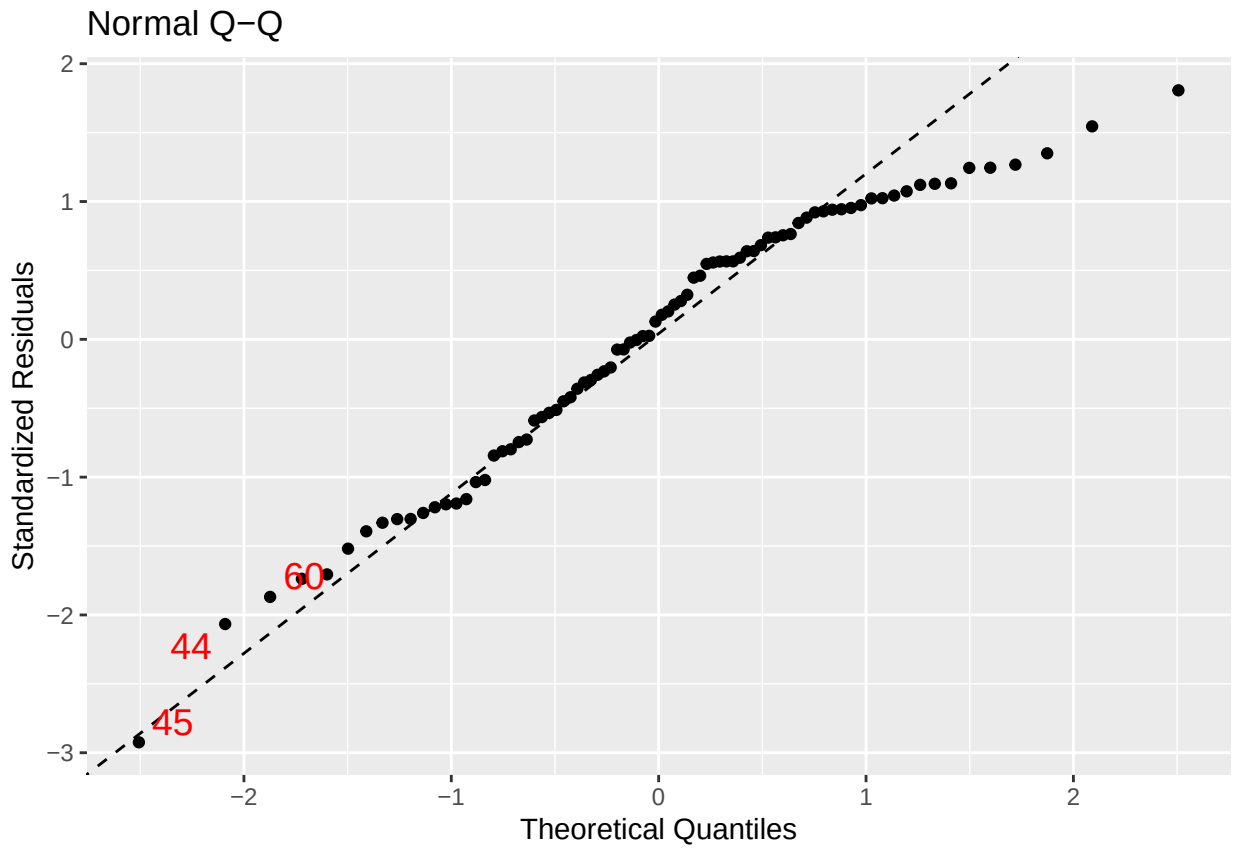
1. Mean 0: This is true automatically
2. Independence: It is safe to assume that the attendance of one game doesn't affect the attendance of another
3. Equal variance per group: There does seem to be a bit of a problem with constant variance, as the strokes are not of equal length. Ideally, we would want a length of a stroke to be no more than twice the length of another stroke. Indeed, the condition violates rule of 2, as the first stroke/fifth stroke > 2.
4. Normality: Based on the second plot below, we can see that the majority of the points stay closely to the dotted line. However, about 1/4 of the residuals trail off from the line (at the head and tail). This implies that normality condition might not be met. With that being said, I would argue that having 3/4 of the residuals following this rule is sufficient, so we would press on.

```
#Equal variance
mplot(mod, which = 1)
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
#Normality
mplot(mod, which = 2)
```



Since we want to use the interaction hypothesis (the third one), we must check for interaction before interpreting. Hence, there is another condition to check.

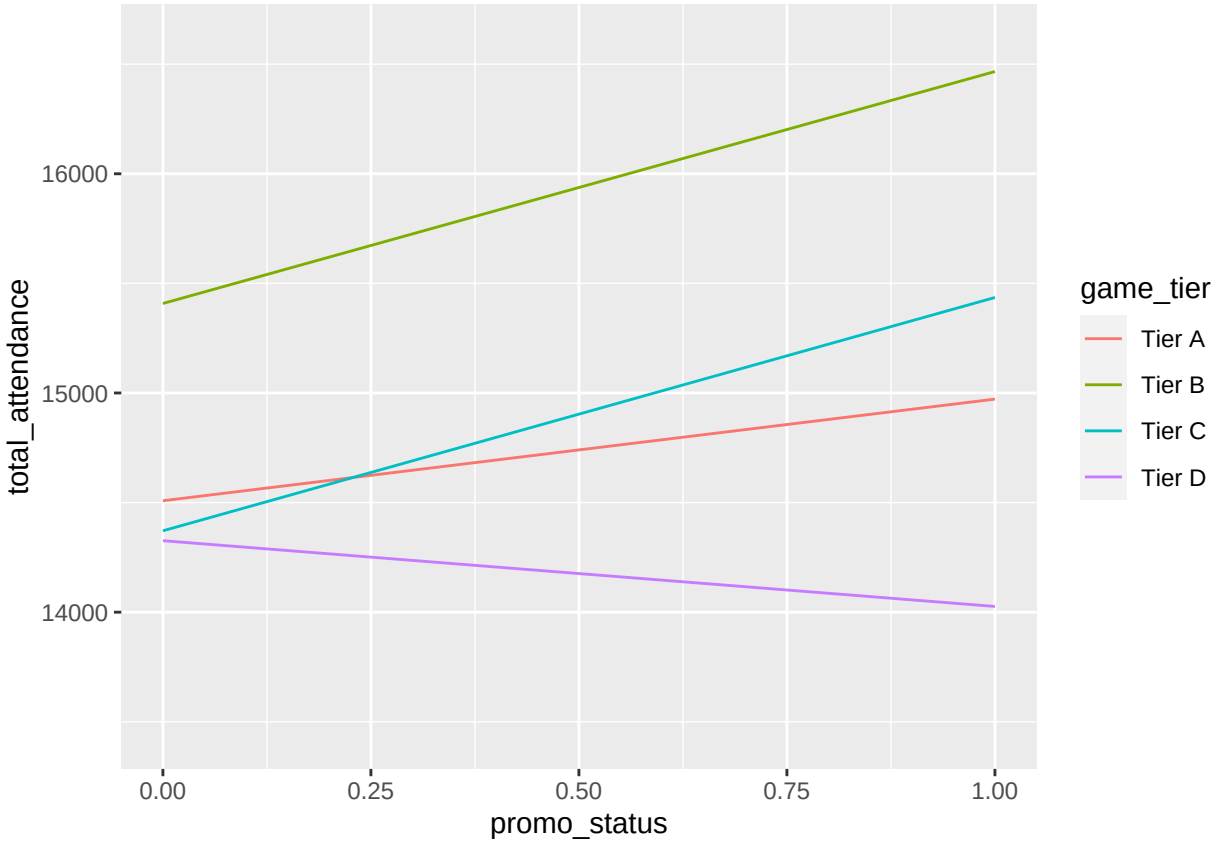
```
#Could probably polish this visualization morex

#mod2 <- ggplot(data = df,
 #              mapping = aes(y = "total_attendance", x = "promo_status"),
  #             color = game_tier)

#mod2 + geom_line()

gf_line(total_attendance ~ promo_status, color = ~ game_tier, data = df, group = ~ game_tier, stat = "su
```

```
## No summary function supplied, defaulting to `mean_se()`
```



We can see that there is a difference in differences. Meaning that the changes in using promos differ among game tiers. Therefore, using a two-way ANOVA with interactions is justified.

To summarize, I would proceed to do inference on this model. However, it is worth noticing that the conditions are nor entirely met, so we should take what we find here with a grain of salt.

###Inference

```
#Two-way ANOVA using game_tier and promo_status
anova(mod)
```

```
## Analysis of Variance Table
##
## Response: total_attendance
##                          Df    Sum Sq Mean Sq F value  Pr(>F)
## game_tier                 3  20634337 6878112  2.2180 0.09311 .
## promo_status              1   3382377 3382377  1.0907 0.29971
## game_tier:promo_status    3   6688605 2229535  0.7190 0.54381
## Residuals                74 229480440 3101087
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we can see from the ANOVA table, there isn't significant evidence to conclude that there is a difference on average in the number of attendance among game tiers. We say this based on an F-statistic of 2.218 and a p-value of 0.09311. Similarly, we can conclude that the difference on average in the attendance between two promo status and among the interactions (game tier and promo status) are both insignificant (with p-values of 0.2997 and 0.5438 accordingly). To summarize, non of our initial factors are significant.

###Conclusion: Where to go from here

There are many possible conclusions we can make. Here I will list the 3 most relevant ones:

1. The problems in conditions might be bigger than we anticipated, making the model less useful and credible. This means that whatever we found out in this analysis might not be accurate and applicable to a bigger data set.
2. Another conclusion might simply be: We don't have enough data/variables to create any sensible output. Without going to deep into the details, having more variables might have helped me pick a better model, which will create a more desired output.
3. Under the assumption that the conditions are met and the data is big enough: It is simply true that promos are insignificant in this model. This translates to "there isn't sufficient evidence to conclude that using promos would ,on average, make a difference in the total attendance".

Final words: Since the conditions are loosely met, we should take this findings with a grain of salt. Thank you!