EDA - Game of Thrones battle data

Aakash Mayur Kumar Shah - S3636808 & Simarpreet Luthra - S3706588

Analysis of Categorical Data Course Project - Phase I

Table of Contents

1.Introduction
2.Data Set Source and Description
3. Initial Setup and Input
4.Data Preprocessing
5.Summary
6.Plot for Attacks by major houses
7.Plot for Attacks defended by major houses
8.Plot for battle type
9.Plot for season of attack
10.Plot for major captures and deaths
11.Plot for attacking allied houses
12.Results and Discussion
13.Conclusion

Introduction

The Game of Thrones is a fatasy drama series that is based on the popular book series written by R.R.Martin. The book was made into a series by HBO. It has been a sucessful book and series having major fan following. The series premiered on April 17, 2011 and has had seven successful seasons so far, with the eighth and final season set to premiere in 2018.The series has won 38 academy awards, making it the most academy awarded series. In additon to academy awards, the series has also won 3 Hugo awards and a Peabody award, not to mention 5 time nominations for the Golden Globe award for the Best Series. In terms of the cast 9 of the cast members have won Primetime Emmy Awards and one of them has also one the Golden Globe Award.

The aim of this analysis is to identify if there is an effect on the outcome of the battles based on the various variables such as battle type, the attacker or the defender etc. It is to see if the Game of Thrones could have been won or played out differently or not given the change in any of these parameters and how drastic a chnage would it be. # Data set source and description The data in this study was sourced from Kaggle and stored on (https://www.kaggle.com/mylesoneill/game-of-thrones/home), which has a compilation of data from a wide range of sources. The data set contains the collection of all the battles waged trhoughout the series.

- name: The name of the Battle.
- year: The year in which the battle occured.
- battle_number: The order number in which the battles took place.
- attacker_king: The name of the king of the kingdom/clan that was attacking in the course of the battle.
- defender_king: The name of the king of the kingdom/clan that was defending in the course of the battle.
- attacker_1: The name of the house that was attacking in the battle. The main attacker.
- attacker_2: The name of the second house that was attacking as a part of a collaboration if any.
- attacker_3: The name of the third house that was attacking as a part of a collaboration if any.
- attacker_4: The name of the fourth house that was attacking as a part of a collaboration if any.
- defender_1: The name of the house that was defending in the battle. The main defending.
- defender_2: The name of the second house that was defending as a part of a collaboration if any.
- defender_3: The name of the third house that was defending as a part of a collaboration if any.
- defender_4: The name of the fourth house that was defending as a part of a collaboration if any.
- attacker_outcome: The outcome of the battle. It is a binomail variable with "win" or "loss" as possible values, where "win" implies that the attcker won the battle, and "loss" implies that the defender won the battle.
- battle_type: The type of the battle waged. It takes 4 values "Pitched Battle","Ambush","Siege","Razing".
- major_death: A binomial variable taking "1" or "0" and tells if there was a death of a major character in the battle or not. If there was a death it is denoted by "1", if not "0".
- major_capture: A binomial variable taking "1" or "0" and tells if there was a capture of a major character in the battle or not. If there was a capture it is denoted by "1", if not "0".
- attacker_size: The size of the attacker's army.
- defender_size: The size of the defender's army.
- attacker_commander: The names of the commanders of the attacker's army.
- defender_commander: The names of the commanders of the defender's army.
- summer: A binomial variable which denotes if it is winter or not, as given by "1" and "0".
- region: The region of the fanatasy map of Westoros the battle was waged in.

## Initial setup and input

```
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
battle <- read_csv("battles.csv")
```

-> In the above cod chunk: *The necessary libraries have been loaded.* The dataset has been imported and store in the variable called battle. ## Data preprocessing

```
battle[!battle$attacker_1 %in%
c("Stark","Greyjoy","Lannister","Baratheon","Bolton","Frey"),"attacker_1"] <-
"Others"
battle$attacker_1 <- as.factor(battle$attacker_1)
battle[!battle$defender_1 %in%
c("Stark","Greyjoy","Lannister","Baratheon","Tully"),"defender_1"] <-
"Others"
battle$defender_1 <- as.factor(battle$defender_1)
battle$attacker_outcome <- as.factor(battle$attacker_outcome)
battle$battle_type <- as.factor(battle$battle_type)
battle$attacker_collab = 3 - is.na(battle$attacker_4) -
is.na(battle$attacker_3) - is.na(battle$attacker_2)
battle$defender_collab = 3 - is.na(battle$defender_4) -
is.na(battle$defender_3) - is.na(battle$defender_2)
colSums(is.na(battle))
```

```
##               name               year      battle_number 
##                  0                  0                  0 
##      attacker_king      defender_king         attacker_1 
##                  2                  3                  0 
##         attacker_2         attacker_3         attacker_4 
##                 28                 35                 36 
##         defender_1         defender_2         defender_3 
##                  0                 36                 38 
##         defender_4   attacker_outcome        battle_type 
##                 38                  1                  1 
##        major_death      major_capture      attacker_size 
##                  1                  1                 14 
##      defender_size attacker_commander defender_commander 
##                 19                  1                 10 
##             summer           location             region 
##                  1                  1                  0 
##               note    attacker_collab    defender_collab 
##                 33                  0                  0
```

```
battle <- battle %>% filter( ! is.na(attacker_outcome),! is.na(defender_1),!
is.na(summer))
battle <- battle %>%
select(attacker_1,defender_1,attacker_outcome,battle_type,major_capture,major
_death,summer,attacker_collab,defender_collab)
colSums(is.na(battle))
```

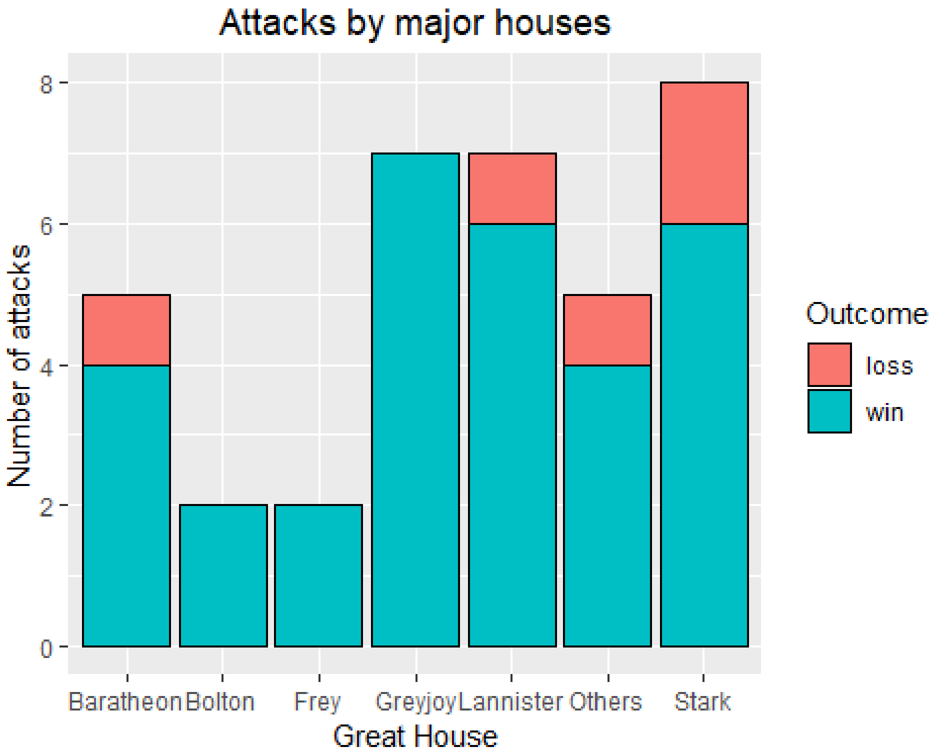
```
##       attacker_1       defender_1 attacker_outcome      battle_type 
##                0                0                0                0
##    major_capture      major_death          summer  attacker_collab
##                0                0               0                0
##  defender_collab
##                0
```

In the above code chunk: *We have recoded the data to have all the minor and nonsignificant to be stored or labelled as “Others” in both the vraiables attacker_1, and defender_1.* The variable attacker_1 has been converetd to a factor to ease further analysis. *Similary the variables defender_1, attacker_outcome and defender_collab have been converted to a factor to ease further analysis.* The attacker_collab and defender_collab have been checked and Nas have been handled. *Finally all the Nas in the remianing variables have been handled.
## Summary

```
head(battle)

## # A tibble: 6 x 9
##   attacker_1 defender_1 attacker_outcome battle_type major_capture
##   <fct>      <fct>      <fct>            <fct>               <int>
## 1 Lannister  Tully      win              pitched ba~             0
## 2 Lannister  Baratheon  win              ambush                  0
## 3 Lannister  Tully      win              pitched ba~             1
## 4 Stark      Lannister  loss             pitched ba~             1
## 5 Stark      Lannister  win              ambush                  1
## 6 Stark      Lannister  win              ambush                  0
## # ... with 4 more variables: major_death <int>, summer <int>,
## #   attacker_collab <dbl>, defender_collab <dbl>

summary(battle)

##      attacker_1     defender_1 attacker_outcome         battle_type
##  Baratheon:5     Baratheon:4     loss: 5           ambush        :10
##  Bolton   :2     Greyjoy  :3     win :31           pitched battle:13
##  Frey     :2     Lannister:9                       razing        : 2
##  Greyjoy  :7     Others   :9                       siege         :11
##  Lannister:7     Stark    :7
##  Others   :5     Tully    :4
##  Stark    :8
##  major_capture     major_death        summer       attacker_collab
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :1.0000   Median :0.0000
##  Mean   :0.3056   Mean   :0.3611   Mean   :0.7222   Mean   :0.3333
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.2500
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :3.0000
##
##  defender_collab
##  Min.   :0.00000
##  1st Qu.:0.00000
##  Median :0.00000
##  Mean   :0.02778
##  3rd Qu.:0.00000  
##  Max.   :1.00000  
## 
```

In the above code chunk: *The data has been displayed.* The summary of the data has been displayed. # Plot for Attacks by major houses

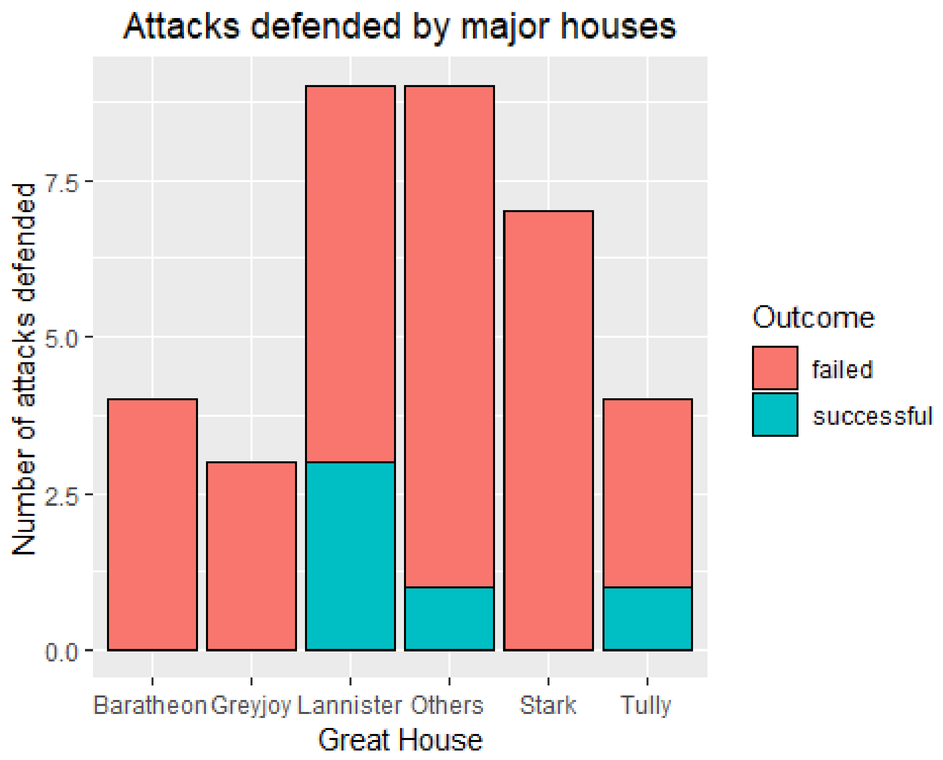
```
a <- ggplot(data = battle,mapping = aes(x = attacker_1,fill = 
attacker_outcome))
a <- a + geom_bar(stat = "count", colour = "black",position = "stack")
a + xlab("Great House") + ylab("Number of attacks") + ggtitle("Attacks by 
major houses") + theme(plot.title = element_text(hjust = 0.5)) + 
scale_fill_discrete(name = "Outcome")
```



In the above code chunk: *A ggplot variable named “a” has been created having the attacker_1 set to X axis and the attacker_outcome has been set to Y axis.* The colour of the abr chart has been set to the battle_outcome. * Finally the labels have been set appropriately.

Thus, we obtain a graph that gives us the number of victories the houses have managed to get and how they have faired. # Plot for Attacks defended by major houses

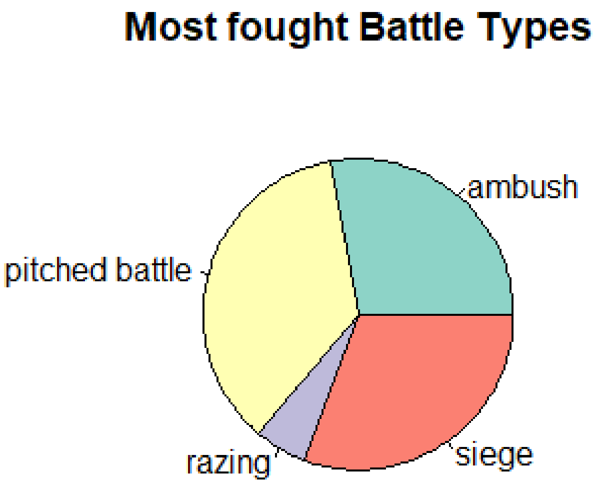
```
b <- ggplot(data = battle,mapping = aes(x = defender_1,fill = 
forcats::fct_rev(attacker_outcome)))
b <- b + geom_bar(stat = "count", colour = "black",position = "stack")
b + xlab("Great House") + ylab("Number of attacks defended") + 
ggtitle("Attacks defended by major houses") + theme(plot.title = 
element_text(hjust = 0.5)) + scale_fill_discrete(name = "Outcome", labels = 
c("failed", "successful"))
```



In the above code chunk: *A ggplot variable named “b” has been created having the defender_1 set to X axis and the rverese attacker_outcome has been set to fill of the chart.* The labels have been set and the positioning as well.

Thus, we get an graph which explains how the houses have faired when it comes to denfending in battles. # Plot for battle type

```
battle_type_summary <- battle %>% group_by(battle_type) %>% summarise(n =
n())
battle_type_summary

## # A tibble: 4 x 2
##   battle_type        n
##   <fct>          <int>
## 1 ambush            10
## 2 pitched battle    13
## 3 razing             2
## 4 siege             11

pie(battle_type_summary$n, labels = battle_type_summary$battle_type, main =
"Most fought Battle Types",col = c("#8dd3c7","#ffffb3","#bebada","#fb8072"))
```



In the above code chunk: *A summary table for the battle_type has been constucted and stored in the variable "battle_type_summary".* Finally a pie chart has been plotted for the newly created summary table.

The pie chart thus obtained tells which battle type has been prefered.

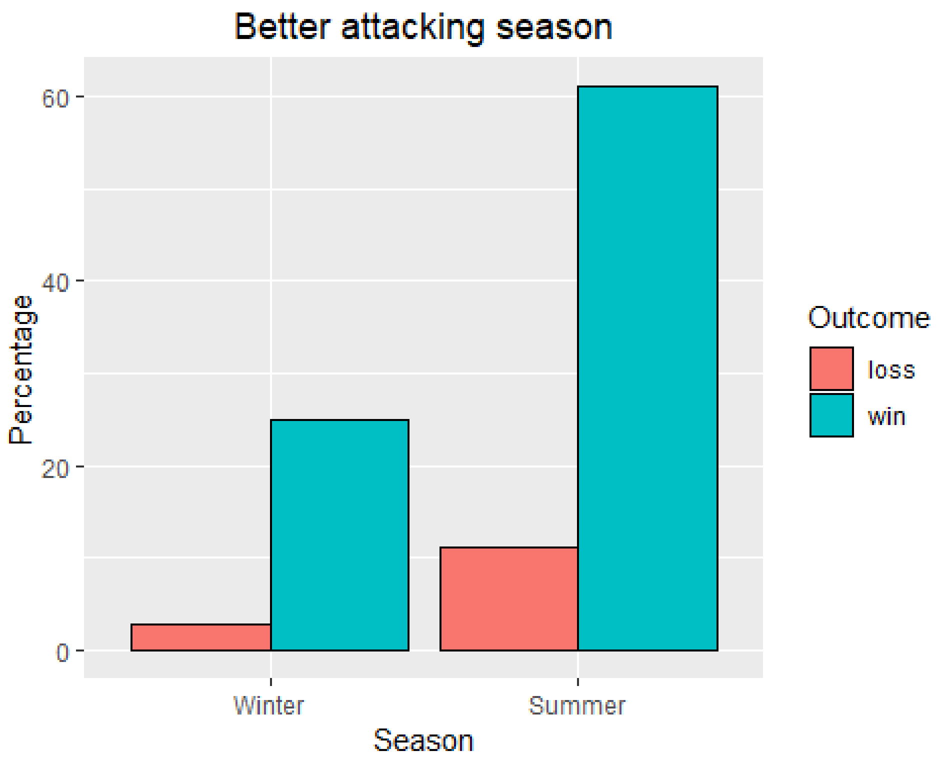
Plot for season of attack

```
summer_summary <- battle %>% group_by(summer,attacker_outcome) %>%
summarise(n = n())
summer_summary$summer <- factor(summer_summary$summer,levels = c(0,1),labels
= c("Winter","Summer"))
summer_summary$percent <- summer_summary$n/nrow(battle)*100
summer_summary
```

```
## # A tibble: 4 x 4
## # Groups:   summer [?]
##   summer attacker_outcome     n percent
##   <fct>  <fct>            <int>   <dbl>
## 1 Winter loss                 1    2.78
## 2 Winter win                  9   25
## 3 Summer loss                 4   11.1
## 4 Summer win                 22   61.1
```

```
c <- ggplot(data = summer_summary,mapping = aes(x = summer,y = percent,fill =
attacker_outcome))
c <- c + geom_bar(stat = "identity", colour = "black",position = "dodge")
```

```
c + xlab("Season") + ylab("Percentage") + ggtitle("Better attacking season")
+ theme(plot.title = element_text(hjust = 0.5)) + scale_fill_discrete(name =
"Outcome")
```



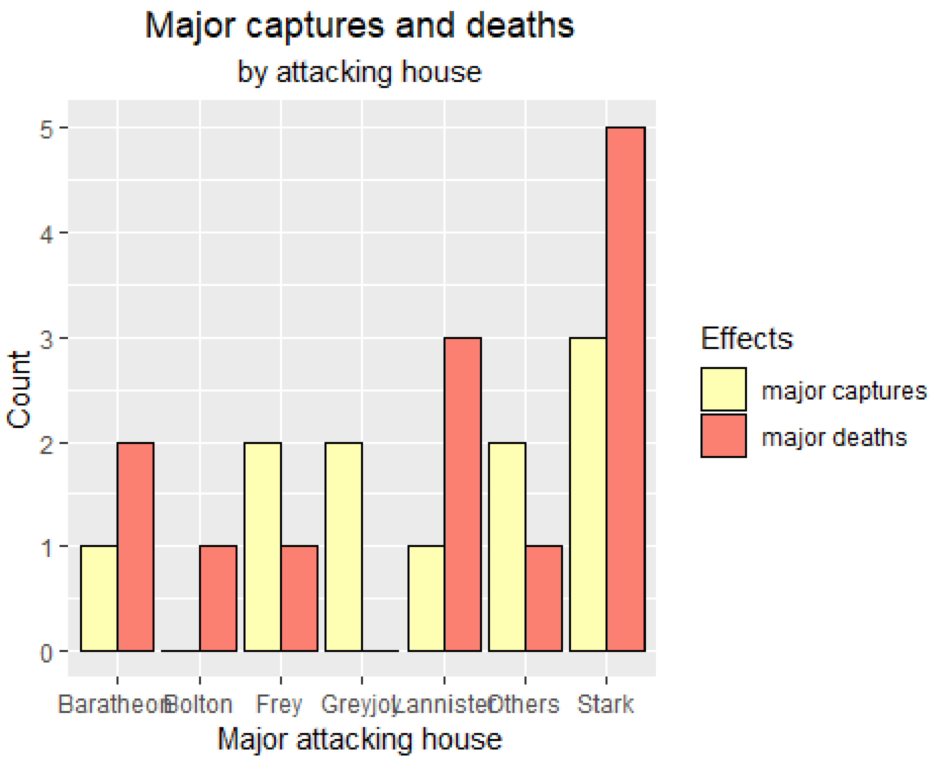
In the above code chunk: *A summary table has been made on the summer variable and has been stored in the variable summer_summary.* The newly created variable has been made into a factor. *A new ggplot varible “c” has been created with the summer variable set to X axis and the percent set to Y axis.* The attacker_outcome has been used to set the fill of the chart. *Finally the labels of the chart has been set.

We thus get a graph that explains if the summer has any effect on the battle outcome.

Plot for major captures and deaths

```
war_effects <- battle %>% group_by(attacker_1) %>% summarise(major_captures =
sum(major_capture),major_deaths = sum(major_death))
war_effects <- war_effects %>% gather(c('major_captures','major_deaths'),key
= "effect",value = "count")
war_effects

## # A tibble: 14 x 3
##    attacker_1 effect         count
##    <fct>      <chr>          <int>
##  1 Baratheon  major_captures     1
##  2 Bolton     major_captures     0
##  3 Frey       major_captures     2
##  4 Greyjoy    major_captures     2
##   5 Lannister  major_captures     1
##   6 Others     major_captures     2
##   7 Stark      major_captures     3
##   8 Baratheon  major_deaths       2
##   9 Bolton     major_deaths       1
##  10 Frey       major_deaths       1
##  11 Greyjoy    major_deaths       0
##  12 Lannister  major_deaths       3
##  13 Others     major_deaths       1
##  14 Stark      major_deaths       5
```

```
d <- ggplot(data = war_effects,mapping = aes(x = attacker_1, y = count,fill =
effect))
d <- d + geom_bar(stat = "identity",colour = "black",position = "dodge")
d + xlab("Major attacking house") + ylab("Count") + ggtitle("Major captures
and deaths",subtitle = "by attacking house") + theme(plot.title =
element_text(hjust = 0.5),plot.subtitle = element_text(hjust = 0.5)) +
scale_fill_manual(values = c("#ffffb3","#fb8072"),name = "Effects",labels =
c("major captures","major deaths"))
```



In the above code chunk: *The sum of all the major deaths and captures in the battle have been calculated.* A ggplot avriable having the attacker_1 mapped to X axis, and count to Y axis has been created. * The labels and the positioning has been set.

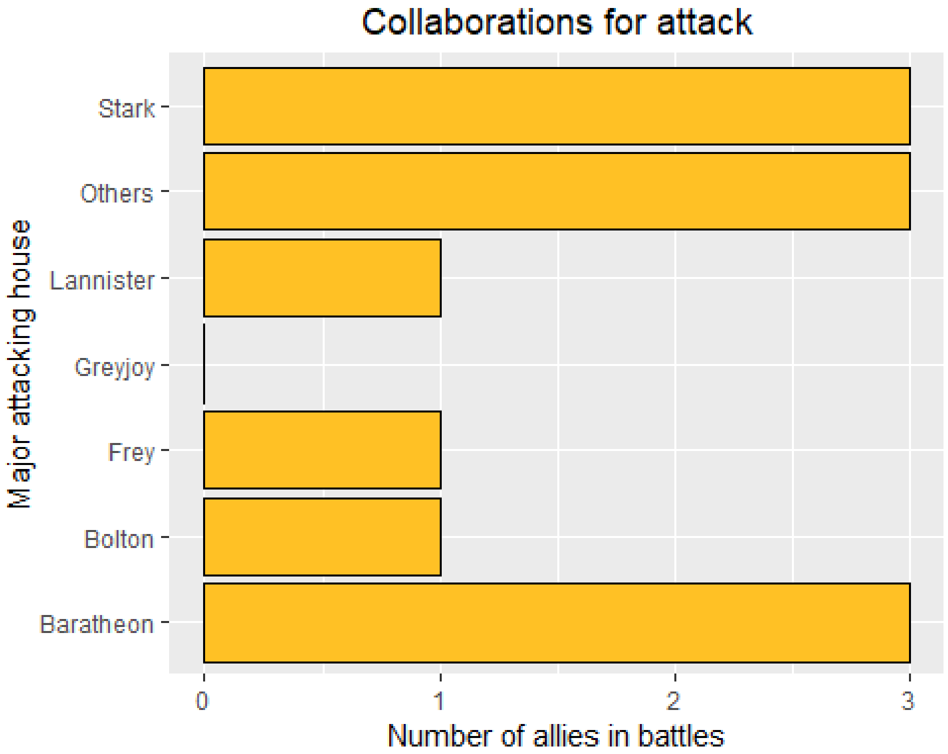
We thus obtain a plot that explains which house had or was responsible for major deaths and captures.

Plot for attacking allied houses

```
allies <- battle %>% group_by(attacker_1) %>% summarise(allies =
sum(attacker_collab))
allies
```

```
## # A tibble: 7 x 2
##   attacker_1 allies
##   <fct>       <dbl>
## 1 Baratheon       3
## 2 Bolton          1
## 3 Frey            1
## 4 Greyjoy         0
## 5 Lannister       1
## 6 Others          3
## 7 Stark           3
```

```
e <- ggplot(data = allies,mapping = aes(x = attacker_1,y = allies))
e <- e + geom_bar(stat = "identity",fill = "#FFC125",colour =
"black",position = "dodge")
e + xlab("Major attacking house") + ylab("Number of allies in battles") +
ggtitle("Collaborations for attack") + theme(plot.title = element_text(hjust
= 0.5)) + coord_flip()
```



In the above code chunk: *A ggplot variable “e” has been cerated with the attacker_1 and the allies mapped to X and Y axis repsectively.* Finally the positioning and labels have been set.

We thus have a graph explaining the houses and how they allied.

Results and Discussion

Analysis of the data has revealed that there were 36 battles fought during the course of the enitre saga of Game of Thrones. The most success has been enjoyed by the Greyjoys who won all of the 7 battles they waged followed by the lannisters who won 6 out of the 7 battles waged. The starks were the once who waged the most battles being 8 in total but won only 6. The boltons and the freys were the ones who waged the least battles being just 2 but both houses were undefeated. The baratheons faired well as well as they won 4 out of 5 battles.

In the case of defending it was observed that almost none of the houses that were defending faired well. It was seen that although none of the houses were good at defending in battles, the lannistes relatively were better as they successfully defended in 3 out of 8 battles. The tullys and the others were successful in defending 1 out of the 4 and 8 battles they waged respectively. The starks, greyjoys and the baratheons had not successfully defended in any battle waged.

The data reveals that when it came to battle type, ptiched battle was the most popular type of battle fought followed by ambush, and seige, and finally razing. It could also be seen that the number of battles won in summer is far more than the number of battles won in winter. The data also shows that death or the capture of major characters when the starks were involved in the battle, followed by the lannisters in terms of deaths and the greyjoys and freys in terms of captures. Finally it could be seen that except the greyjoys all the other houses have been involved in collaboration. The starks, baratheons and others had the maximum collaborations while the lannisters, freys and boltons relatively had less collaboration.

Conclusion

Preliminary analysis of the data suggests that there are several factors that can affect the battle outcomes. It is to be analysed that if the season does indeed affect the battle outcome or a particular house in terms of winning or loosing. It would also be interesting to see if the seasons affect a particular type of battle.