Final Project

Question 1

```
setwd("/Users/jessicasaini/Desktop/UoW/Stat 847/Final Project")

library(GGally)
```

```
## Loading required package: ggplot2
```

```
## Warning in register(): Can't find generic `scale_type` in package ggplot2 to
## register S3 method.
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(mice)
```

```
##
## Attaching package: 'mice'
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
library(sentimentr)
library(plyr)
library(tidyverse)
```

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

```
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1
## v purrr   0.3.4
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::arrange()   masks plyr::arrange()
## x purrr::compact()   masks plyr::compact()
## x dplyr::count()     masks plyr::count()
## x dplyr::failwith()  masks plyr::failwith()
## x dplyr::filter()    masks mice::filter(), stats::filter()
## x dplyr::id()        masks plyr::id()
## x dplyr::lag()       masks stats::lag()
## x dplyr::mutate()    masks plyr::mutate()
## x dplyr::rename()    masks plyr::rename()
## x dplyr::summarise() masks plyr::summarise()
## x dplyr::summarize() masks plyr::summarize()
```

Reading The data

```
df = read.csv("Gamelog T20I Stat 847.csv")
head(df)
```

```
##   Format MatchNo TeamBowling TeamBatting Inning Over Ball Bowler BowlerID
## 1   T20I      33         AUS          BD      1    0    1  B Lee       17
## 2   T20I      33         AUS          BD      1    0    2  B Lee       17
## 3   T20I      33         AUS          BD      1    0    3  B Lee       17
## 4   T20I      33         AUS          BD      1    0    4  B Lee       17
## 5   T20I      33         AUS          BD      1    0    4  B Lee       17
## 6   T20I      33         AUS          BD      1    0    5  B Lee       17
##       Batsman BatsmanID Fielder FielderID Outcome NumOutcome BallType
## 1 Tamim Iqbal      1041                NA      no          0      run
## 2 Tamim Iqbal      1041                NA      no          0      run
## 3 Tamim Iqbal      1041                NA      no          0      run
## 4 Tamim Iqbal      1041                NA       1          1     wide
## 5 Tamim Iqbal      1041                NA      no          0      run
## 6 Tamim Iqbal      1041                NA      no          0      run
##   NumBallType  Notes
## 1           0   good
## 2           0  short
## 3           0  short
## 4           2  Tamim
## 5           0 fuller
## 6           0   good
##
## 1                                            good start by Lee   dug in short of a length outside of
## 2                                         short of a length outside off again   this time Tamim gets
## 3                                            short again and aimed at the body   Tamim gets on the ba
## 4 Tamim backs away to whack that over the off side   Lee senses it and thuds it in short, the ball sa
## 5                                              fuller in length and inviting the drive   Tamim flashe
## 6                                                                               good length aimed at
##   IDflag Wickets
## 1      0       0
## 2      0       0
## 3      0       0
## 4      0       0
## 5      0       0
## 6      0       0
```

##Question 1 Make summary statistics (25 of 100 points)

Answer: Variables used: Over, Wickets and NumOutcome. The data is grouped by MatchNo, NumOutcome,Wickets and Over and frequency of the variable is calculated for T20 dataset.

```
t20_df = subset(df, df$Format=="T20I")
head(t20_df)
```

```
##   Format MatchNo TeamBowling TeamBatting Inning Over Ball Bowler BowlerID
## 1   T20I      33         AUS          BD      1    0    1  B Lee       17
## 2   T20I      33         AUS          BD      1    0    2  B Lee       17
## 3   T20I      33         AUS          BD      1    0    3  B Lee       17
## 4   T20I      33         AUS          BD      1    0    4  B Lee       17
## 5   T20I      33         AUS          BD       1    0    4  B Lee       17
## 6   T20I      33         AUS          BD       1    0    5  B Lee       17
##       Batsman BatsmanID Fielder FielderID Outcome NumOutcome BallType
## 1 Tamim Iqbal      1041                NA      no          0      run
## 2 Tamim Iqbal      1041                NA      no          0      run
## 3 Tamim Iqbal      1041                NA      no          0      run
## 4 Tamim Iqbal      1041                NA       1          1     wide
## 5 Tamim Iqbal      1041                NA      no          0      run
## 6 Tamim Iqbal      1041                NA      no          0      run
##   NumBallType  Notes
## 1           0   good
## 2           0  short
## 3           0  short
## 4           2  Tamim
## 5           0 fuller
## 6           0   good
##
## 1                                            good start by Lee   dug in short of a length outside of
## 2                                         short of a length outside off again   this time Tamim gets
## 3                                            short again and aimed at the body   Tamim gets on the ba
## 4 Tamim backs away to whack that over the off side   Lee senses it and thuds it in short, the ball sa
## 5                                              fuller in length and inviting the drive   Tamim flashe
## 6                                                                               good length aimed at
##   IDflag Wickets
## 1      0       0
## 2      0       0
## 3      0       0
## 4      0       0
## 5      0       0
## 6      0       0
```

```
#Change -1 in NumOutcome to 0
t20_df$NumOutcome[t20_df$NumOutcome == -1]<-0
```

HeatMap of Outcome of a Ball over Wickets and Over

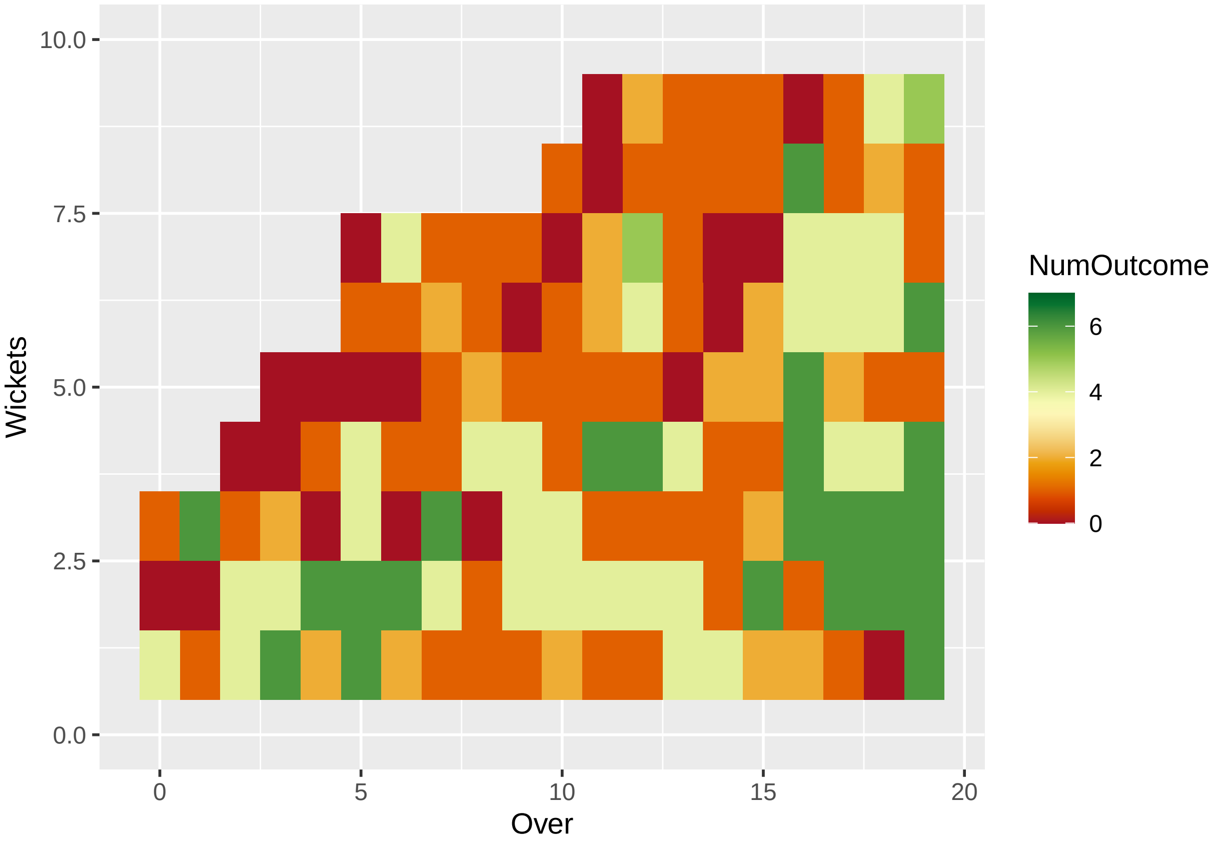
Variables used: Over, Wickets and NumOutcome

Approach: The data is grouped by MatchNo, NumOutcome,Wickets and Over and frequency of the variable is calculated for T20 dataset. The dataset is grouped by MatchNo, NumOutcome, Wickets and Over and frequency of the variables is calculated. The heatmap showcases the different trends observed when visualising and grouping the data by above mentioned variables.

```
a<- t20_df
a = a[, c("MatchNo", "NumOutcome", "Wickets", "Over")]
a<-a %>%
  group_by(MatchNo, NumOutcome,Wickets,Over) %>%
  dplyr::summarise(n = n())%>%
mutate(freq = n / sum(n))
```

```
## `summarise()` has grouped output by 'MatchNo', 'NumOutcome', 'Wickets'. You can
## override using the `.groups` argument.
```

```
#View(a)
ggplot(a, mapping = aes(x = Over, y = Wickets, fill = NumOutcome )) + geom_tile()+ylim(0,10)+ scale_fil
```



Inference: Inference: The above graph depicts the outcome of the ball i.e. (0,1,2,3,4,5,6,7) as a function of Overs and Wickets. All games start in the upper-left corner, with 20 overs and 10 wickets remaining. The following observations can be made from the heatmap:

- Most of the runs scored are between 0 and 2.
- Wickets 1 to 4 have higher chances of hitting a six and a boundary. This could be because the initial players are primarily batsmen and have a more chance of scoring higher runs than the later wickets.
- In the above chart, most of the sixes happen after the 15th over. As the game approaches the end, it is common for the batsman to hit a six in order to score more runs for his team.
- The light green square box after the 15th over and Wicket 5 onwards, indicates that the players try to score more boundaries (just like the sixes), as the game comes to a conclusion.

Bubble Chart

Variables Used: Wickets, Over, Innings and Runs scored on the ball(NumOutcome)

Approach: The runs 3, 5 and 7 are rare in cricket. The dataset is grouped by MatchNo,Inning, NumOutcome, Wickets and Over. The frequency of the variables is calculated. The bubble chart showcases the different trends observed when visualising and grouping the data by above mentioned variables. The following graph visualizes the chances of a ball fetching 3, 5 and 7.

```
#Bubble Chart Final for Balls with 3, 5 and 7 runs
a<- df
a$NumOutcome[a$NumOutcome == -1]<-0
a <- subset(a,NumOutcome %in% c(3,5,7))
a <- subset(a,Wickets<11)

print(unique(a$NumOutcome))
```

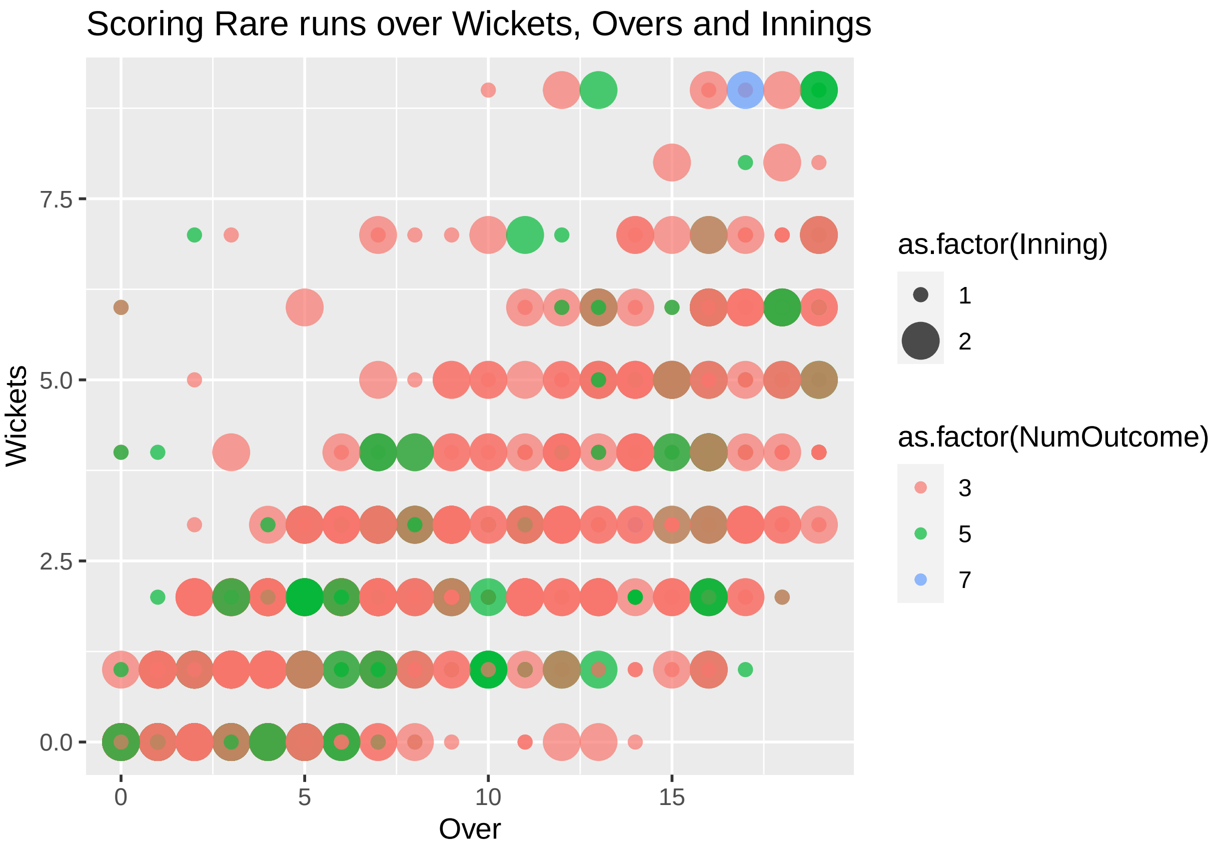
```
## [1] 5 3 7
```

```
a<-a %>%
     group_by(MatchNo,NumOutcome,Wickets,Over,Inning) %>%
     dplyr::summarise(n = n())%>%
     mutate(freq = n / sum(n))
```

```
## `summarise()` has grouped output by 'MatchNo', 'NumOutcome', 'Wickets', 'Over'.
## You can override using the `.groups` argument.
```

```
ggplot(a, aes(x=Over, y=Wickets, size = as.factor(Inning),color=as.factor(NumOutcome))) +
    geom_point(alpha=0.7)+ ggtitle("Scoring Rare runs over Wickets, Overs and Innings")
```

```
## Warning: Using size for a discrete variable is not advised.
```



Inference:

- The larger bubble represent the rare runs scored during second inning whereas the smaller bubbles represent the rare runs scored during first inning.
- Rare Runs(3,5 and 7) are common both during the first and the second inning
- 7 are the rarest whereas 3 runs are common.
- 5 runs are more prominent in the first 10 overs.

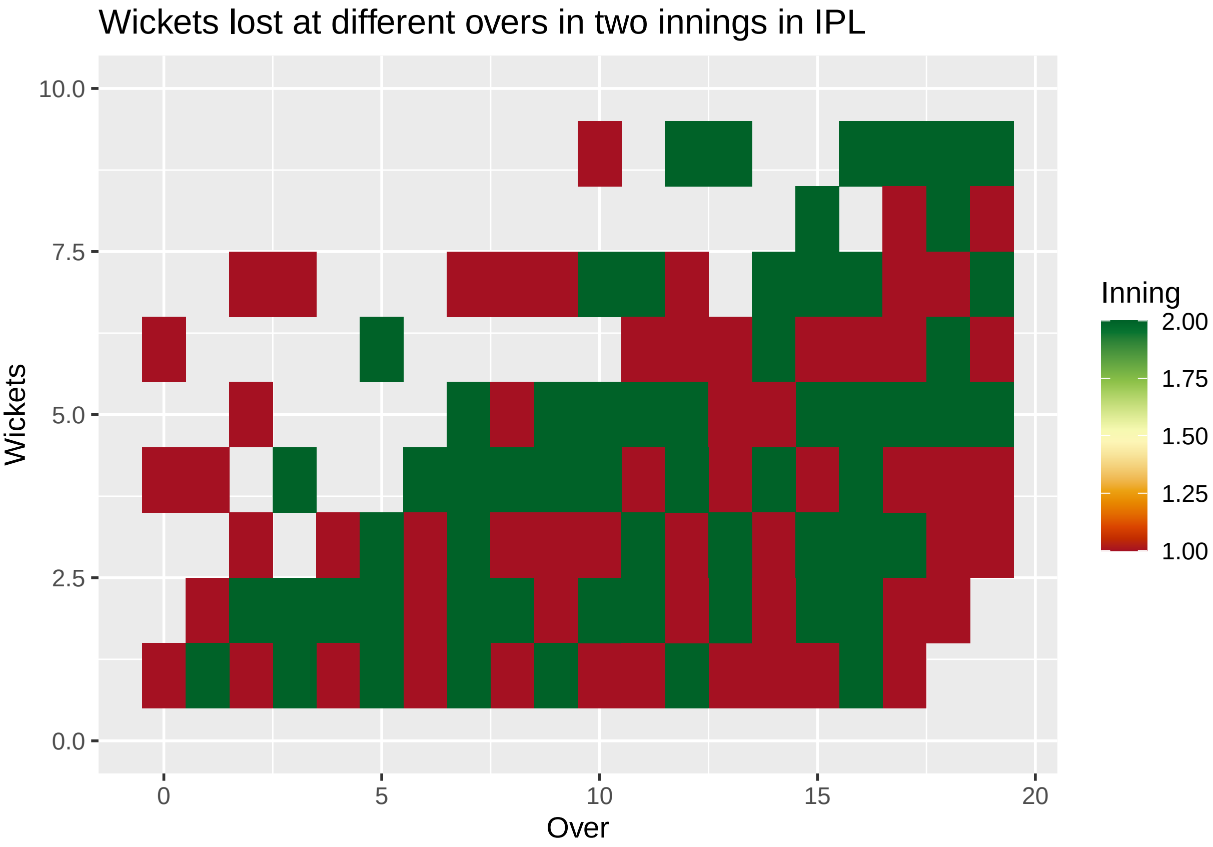
Additional trends found during analysis:

```
#HeatMap 2
a<- df
a$NumOutcome[a$NumOutcome == -1]<-0
a <- subset(a,NumOutcome %in% c(3,5,7))
a<-a %>%
     group_by(MatchNo,NumOutcome,Wickets,Over,Inning) %>%
     dplyr::summarise(n = n())%>%
     mutate(freq = n / sum(n))
```

```
## `summarise()` has grouped output by 'MatchNo', 'NumOutcome', 'Wickets', 'Over'.
## You can override using the `.groups` argument.
```

```
ggplot(a, mapping = aes(x = Over, y = Wickets, fill = Inning )) + geom_tile()+ylim(0,10)+ scale_fill_gra
```

```
## Warning: Removed 3 rows containing missing values (geom_tile).
```



Approach: The dataset is grouped by MatchNo,Inning, NumOutcome, Wickets and Over. The data is visualized for Over, Wickets and Inning.

Inference: - During the second innings, the teams lose around 3 wickets in the first 5 overs. - In the first innings, the first 3 wickets are majorly lost after 5 overs

Question 2 Identify second inning 'turning points'

Filter the data for second innings only and use the sentiment package to calculate the sentiment score.

```
#Question 2 :  Identify second inning 'turning points'

#Filter inning 2 data
setwd("/Users/jessicasaini/Desktop/UoW/Stat 847/Final Project")
df = read.csv("Gamelog T20I Stat 847.csv")
library(dplyr)
dat2 = filter(df , Inning == 2) #filter for second inning only
#View(dat2)
```

```
dat2 <- subset(dat2, !is.na(dat2$FullNotes))

dat2$FullNotes <- as.character(dat2$FullNotes)
dat2 <- dat2[-c(21815) , ]

sentiment = sentiment_by(dat2$FullNotes) # sentiment Df
```

```
## Warning: Each time `sentiment_by` is run it has to do sentence boundary disambiguation when a
## raw `character` vector is passed to `text.var`. This may be costly of time and
## memory.  It is highly recommended that the user first runs the raw `character`
## vector through the `get_sentences` function.
```

```
dat2$sentiment_score = sentiment$ave_sentiment # ave_sentiment is sentiment score
```

```
#View first few sentiment scores and check range of sentiment scores
head(dat2$sentiment_score)
```

```
## [1]  0.00000000  0.26726124  0.08354082  0.00000000  0.15000000 -0.31622777
```

```
min(dat2$sentiment_score)
```

```
## [1] -1.576641
```

```
max(dat2$sentiment_score)
```

```
## [1] 1.63087
```

After calculating the sentiment score, the next step is to filter the data based on extreme sentiments. The range of sentiment scores is from approx. -1.5 to 1.6. Higher the absolute value of the score, extreme is the sentiment.
Filtering the balls in terms of extreme positive and negative sentiment based on a threshold.

Positive Sentiment Ball by Ball Analysis

```
positive_sentiment_df = filter(dat2, dat2$Over > 13 & dat2$sentiment_score > 0.7)
print("The number of rows")
```

```
## [1] "The number of rows"
```

```
print(nrow(positive_sentiment_df))
```

```
## [1] 50
```

```
head(positive_sentiment_df)
```

```
##   Format MatchNo TeamBowling TeamBatting Inning Over Ball        Bowler
## 1   T20I      13         AUS         ENG      2   15    5     A Symonds
## 2   T20I     198         AUS         ENG      2   16    5       SW Tait
## 3   T20I     184         PAK         AUS      2   14    1 Mohammad Amir
## 4   T20I     184         PAK         AUS      2   15    5   Saeed Ajmal
## 5   T20I       4         AUS          SA      2   16    2      JR Hopes
## 6   T20I      97         AUS          SL      2   16    2         B Lee
##   BowlerID    Batsman BatsmanID Fielder FielderID Outcome NumOutcome BallType
## 1       25   PA Nixon      2061                NA       1          1      run
## 2       39 TT Bresnan      2050                NA       1          1      run
## 3     5066 MEK Hussey        54                NA       1          1  leg bye
## 4     5046  SPD Smith        46                NA      no          0      run
## 5       13    AJ Hall      6011                NA     OUT         -1      out
## 6       17  J Mubarak      7021                NA      no          0      run
##   NumBallType Notes
## 1           0 quick
## 2           0  very
## 3           4 Aamer
## 4           0  full
## 5           0 Hopes
## 6           0     a
##
## 1
## 2                               very sharp and a good line from Tait   Bresnan stays leg side of the
## 3 Aamer strikes Hussey on the pad first ball   but that's sliding well down the leg side and the appe
## 4                                                                         full toss   but hammered s
## 5        Hopes to Hall   OUT (Caught), struck well in the air to deep midwicket, Hussey takes it very
## 6                                                                           a very fast and full deli
##   IDflag Wickets sentiment_score
## 1      0       5       0.7071068
## 2      0       5       0.7226596
## 3      0       4       0.7248824
## 4      0       5       0.7954951
## 5      0       6       0.7110696
## 6      0       4       0.9600000
```

Negative Sentiment Ball by Ball Analysis

```r
negative_sentiment_df = filter(dat2, dat2$Over > 13 & sentiment_score < - 0.9)
print("The number of rows")
```

```
## [1] "The number of rows"
```

```r
nrow(negative_sentiment_df)
```

```
## [1] 39
```

```r
head(negative_sentiment_df)
```

```
##   Format MatchNo TeamBowling TeamBatting Inning Over Ball             Bowler
## 1   T20I     331         AUS         IND      2   16    2        JP Faulkner
## 2   T20I     210          SA         AUS      2   17    2           M Morkel
## 3   T20I     333          NZ          BD      2   14    5           KD Mills
## 4   T20I     117          BD          WI      2   14    2 Mehrab Hossain jnr
## 5   T20I     316          BD         ZIM      2   17    1      Shafiul Islam
## 6   T20I     317          NZ         ENG      2   16    1        CJ Anderson
##   BowlerID      Batsman BatsmanID Fielder FielderID Outcome NumOutcome BallType
## 1       75 Yuvraj Singh      3033                NA       1          1  leg bye
## 2     6044    DJ Hussey        14                NA      no          0      run
## 3     4021  Mahmudullah        NA                NA      no          0      run
## 4     1053    TM Dowlin      8065                NA      no          0      run
## 5     1050     P Utseya      9037                NA      no          0      run
## 6     4074    RS Bopara      2060                NA      no          0      run
##   NumBallType         Notes
## 1           4       pitched
## 2           0        Hussey
## 3           0 well-executed
## 4           0        tossed
## 5           0         lands
## 6           0        length
##
## 1                                          pitched up and heading for the pads   Yuvraj looking to whi
## 2                  Hussey charges out aggressively at Morkel   but swings across the line too much and
## 3                                                                             well-executed slower bal
## 4 tossed up   landed on the off stump and spun away just a touch, Dowlin went for a massive heave ove
## 5
## 6
##   IDflag Wickets sentiment_score
## 1      0       4      -1.3205134
## 2      0       4      -0.9167659
## 3      1       6      -0.9143593
## 4      0       2      -0.9275568
## 5      0       6      -0.9013878
## 6      0       4      -1.1759495
```

The total number of combined observations includes positive and negative. We need to construct a diverse ### Constructing the Highlight Reel

For Positive Sentiment:

```
positive_boundaries = filter(positive_sentiment_df, positive_sentiment_df$NumOutcome == 4)
positive_sixes = filter(positive_sentiment_df, positive_sentiment_df$NumOutcome == 6)
positive_exceptions = filter(positive_sentiment_df, positive_sentiment_df$NumOutcome %in% c(5,7))
positive_wickets = filter(positive_sentiment_df, positive_sentiment_df$BallType == "out")
positive_wides = filter(positive_sentiment_df, positive_sentiment_df$BallType == "wide")
positive_sentiment = bind_rows(positive_boundaries,positive_sixes,positive_exceptions,positive_wickets,p
head(positive_sentiment)
```

```
##   Format MatchNo TeamBowling TeamBatting Inning Over Ball          Bowler
## 1   T20I     151          SL          NZ      2   19    3      SL Malinga
## 2   T20I      72          SL         PAK      2   19    4 KMDN Kulasekara
## 3   T20I     189          SA         ZIM      2   14    2     RJ Peterson
## 4    IPL  201172         KKR          MI      2   18    1           B Lee
## 5    IPL  201360         RCB         KKR      2   16    1        A Mithun
## 6    IPL  201340         RCB          RR      2   16    4   R Vinay Kumar
##   BowlerID         Batsman BatsmanID Fielder FielderID Outcome NumOutcome
## 1     7020     NL McCullum      4045                NA    FOUR          4
## 2       NA    Shoaib Malik      5026                NA    FOUR          4
## 3     6023      BRM Taylor      9036                NA    FOUR          4
## 4       17 Harbhajan Singh     16007                NA    FOUR          4
## 5    18021       JH Kallis      6013                NA    FOUR          4
## 6     3056       SR Watson        27                NA     SIX          6
##   BallType NumBallType     Notes
## 1      run           0     found
## 2      run           0 beautiful
## 3      run           0 brilliant
## 4      run           0     Lucky
## 5      run           0        he
## 6      run           0       low
##
## 1            found the gap   full delivery and he moved across and swing that between fine leg and d
## 2
## 3 brilliant stuff from Taylor   really classy stuff, skipping down the ground, getting inside the lin
## 4              Lucky top edge! It was a lovely bouncer at the body   a cramped-up Harbhajan is late
## 5                                                     he bowls a quick short delivery outside off
## 6                                                  low full toss and Watson clears it comfortably
##   IDflag Wickets sentiment_score
## 1      0       6       0.8325383
## 2      1       7       0.7794229
## 3      0       2       0.7394255
## 4      0       6       0.7227786
## 5      0       3       0.7550471
## 6      0       3       0.9231591
```

```
#View(positive_sentiment)
```

For Negative examples

```
negative_boundaries = filter(negative_sentiment_df, negative_sentiment_df$NumOutcome == 4)
negative_sixes = filter(negative_sentiment_df, negative_sentiment_df$NumOutcome == 6)
negative_exceptions = filter(negative_sentiment_df, negative_sentiment_df$NumOutcome %in% c(5,7))
negative_wickets = filter(negative_sentiment_df, negative_sentiment_df$BallType == "out")
negative_wides = filter(negative_sentiment_df, negative_sentiment_df$BallType == "wide")
negative_sentiment = bind_rows(negative_boundaries,negative_sixes,negative_exceptions,negative_wickets,n
print(negative_sentiment)
```

```
##   Format MatchNo TeamBowling TeamBatting Inning Over Ball      Bowler BowlerID
## 1    IPL  201458         KXP         CSK      2   18    6    AR Patel    15055
## 2    IPL  201542         KKR          DD      2   17    5      Mishra       NA
## 3   T20I      78          NZ          WI      2   19    4  TG Southee     4050
## 4   T20I     247          WI          NZ      2   15    2  FH Edwards       NA
## 5    IPL  201419          RR         KKR      2   18    5 JP Faulkner       75
## 6    IPL  200915         KXP          RR      2   16    2   IK Pathan    15003
## 7    IPL  200915         KXP          RR      2   16    2   IK Pathan    15003
## 8    IPL  200940          DC          RR      2   14    6    RP Singh    11009
## 9    IPL  200950          DD          RR      2   17    1   DP Nannes       66
##         Batsman BatsmanID  Fielder FielderID Outcome NumOutcome BallType
## 1      MS Dhoni      3008                 NA       4          4     byes
## 2     IK Pathan     15003                 NA    FOUR          4      run
## 3     JE Taylor      8040 MCCULLUM      4045     OUT         -1      out
## 4      JDP Oram      4024    SAMMY      8048     OUT         -1      out
## 5 R Vinay Kumar      3056                 NA     OUT         -1      out
## 6      SK Warne     19008                 NA       1          1     wide
## 7      SK Warne     19008                 NA       1          1     wide
## 8      M Morkel      6044                 NA       1          1     wide
## 9    SK Trivedi     19015                 NA       1          1     wide
##   NumBallType    Notes
## 1           3   swings
## 2           0   offers
## 3           0    swing
## 4           0     Oram
## 5           0 Faulkner
## 6           2    Warme
## 7           2    Warme
## 8           2     goes
## 9           2 bouncer,
##
## 1
## 2
## 3
## 4 Oram drops a short ball outside off into the covers   thinks of a single before pausing, but Latham
## 5
## 6
## 7
## 8
## 9
##   IDflag Wickets sentiment_score
## 1      0       6      -1.0000000
## 2      1       5      -0.9053616
## 3      0       7      -0.9356092
## 4      1       3      -0.9610181
## 5      0       6      -0.9074537
## 6      0       6      -1.0914103
## 7      0       6      -1.0914103
## 8      0       5      -1.0914103
## 9      0       9      -0.9799872
```

```
#head(negative_sentiment)
```

##Highlight Reel

```
reel = bind_rows(positive_sentiment,negative_sentiment)
reel2 = data.frame(reel)

print(reel2[1:20,])
```

```
##    Format MatchNo TeamBowling TeamBatting Inning Over Ball          Bowler
## 1    T20I     151          SL          NZ      2   19    3      SL Malinga
## 2    T20I      72          SL         PAK      2   19    4 KMDN Kulasekara
## 3    T20I     189          SA         ZIM      2   14    2     RJ Peterson
## 4     IPL  201172         KKR          MI      2   18    1           B Lee
## 5     IPL  201360         RCB         KKR      2   16    1        A Mithun
## 6     IPL  201340         RCB          RR      2   16    4   R Vinay Kumar
## 7    T20I       4         AUS          SA      2   16    2        JR Hopes
## 8    T20I     301         ENG          NZ      2   15    5       LJ Wright
## 9    T20I     125          SA         ENG      2   15    1       JA Morkel
## 10    IPL  201345         CSK         KXP      2   19    2        DJ Bravo
## 11    IPL  201424          SH         RCB      2   19    1       IK Pathan
## 12   T20I      40         IND         ENG      2   19    3       IK Pathan
## 13    IPL  201458         KXP         CSK      2   18    6        AR Patel
## 14    IPL  201542         KKR          DD      2   17    5          Mishra
## 15   T20I      78          NZ          WI      2   19    4      TG Southee
## 16   T20I     247          WI          NZ      2   15    2      FH Edwards
## 17    IPL  201419          RR         KKR      2   18    5     JP Faulkner
## 18    IPL  200915         KXP          RR      2   16    2       IK Pathan
## 19    IPL  200915         KXP          RR      2   16    2       IK Pathan
## 20    IPL  200940          DC          RR      2   14    6        RP Singh
##    BowlerID         Batsman BatsmanID  Fielder FielderID Outcome NumOutcome
## 1      7020     NL McCullum      4045                 NA    FOUR          4
## 2        NA    Shoaib Malik      5026                 NA    FOUR          4
## 3      6023      BRM Taylor      9036                 NA    FOUR          4
## 4        17 Harbhajan Singh     16007                 NA    FOUR          4
## 5     18021       JH Kallis      6013                 NA    FOUR          4
## 6      3056       SR Watson        27                 NA     SIX          6
## 7        13          AJ Hall      6011                NA     OUT          -1
## 8      2066     NL McCullum      4045  BUTTLER      2092     OUT          -1
## 9      6041        IJL Trott      2070     KUHN      6070     OUT          -1
## 10     8059 Gurkeerat Singh     15046   HUSSEY        54     OUT          -1
## 11    15003         MA Starc        50                NA     OUT          -1
## 12     3023        LJ Wright      2066    DHONI      3008       1           1
## 13    15055         MS Dhoni      3008                NA       4           4
## 14       NA        IK Pathan     15003                NA    FOUR           4
## 15     4050        JE Taylor      8040 MCCULLUM      4045     OUT          -1
## 16       NA         JDP Oram      4024    SAMMY      8048     OUT          -1
## 17       75   R Vinay Kumar      3056                NA     OUT          -1
## 18    15003         SK Warne     19008                NA       1           1
## 19    15003         SK Warne     19008                NA       1           1
## 20    11009        M Morkel      6044                NA       1           1
##    BallType NumBallType      Notes
## 1       run           0      found
## 2       run           0  beautiful
## 3       run           0  brilliant
## 4       run           0      Lucky
## 5       run           0         he
## 6       run           0        low
## 7       out           0      Hopes
## 8       out           0    massive
## 9       out           0 impressive
## 10      out           0     slower
## 11      out           0      Starc
## 12     wide           2        way
## 13     byes           3     swings
## 14      run           0     offers
## 15      out           0      swing
## 16      out           0       Oram
## 17      out           0   Faulkner
## 18     wide           2      Warme
## 19     wide           2      Warme
## 20     wide           2       goes
##
## 1
## 2
## 3                                                                                                   l
## 4
## 5
## 6
## 7
## 8
## 9
## 10
## 11
## 12
## 13
## 14
## 15
## 16 Oram drops a short ball outside off into the covers   thinks of a single before pausing, but Latha
## 17
## 18
## 19
## 20
##    IDflag Wickets sentiment_score
## 1       0       6       0.8325383
## 2       1       7       0.7794229
## 3       0       2       0.7394255
## 4       0       6       0.7227786
## 5       0       3       0.7550471
## 6       0       3       0.9231591
## 7       0       6       0.7110696
## 8       0       6       0.7176968
## 9       0       2       0.8308505
## 10      0       4       0.8646402
## 11      0       5       0.7033533
## 12      0       5       0.8112490
## 13      0       6      -1.0000000
## 14      1       5      -0.9053616
## 15      0       7      -0.9356092
## 16      1       3      -0.9610181
## 17      0       6      -0.9074537
## 18      0       6      -1.0914103
## 19      0       6      -1.0914103
## 20      0       5      -1.0914103
```

Highlight Reel: After calculating the sentiment score, the next step is to filter the data based on extreme sentiments. The range of sentiment scores is from approx. -1.5 to 1.6. Higher the absolute value of the score, extreme is the sentiment. Filter the data in terms of extreme positive and negative sentiment based on a threshold. After the data is filtered, we constructed a highlight right for 20 balls. Since most of the turning points in the second innings during the last over, the data is filtered 6 overs. The highlight reel as mentioned above contains sixes, wickets, boundaries, rare runs and wides for both positive and negative sentiment.

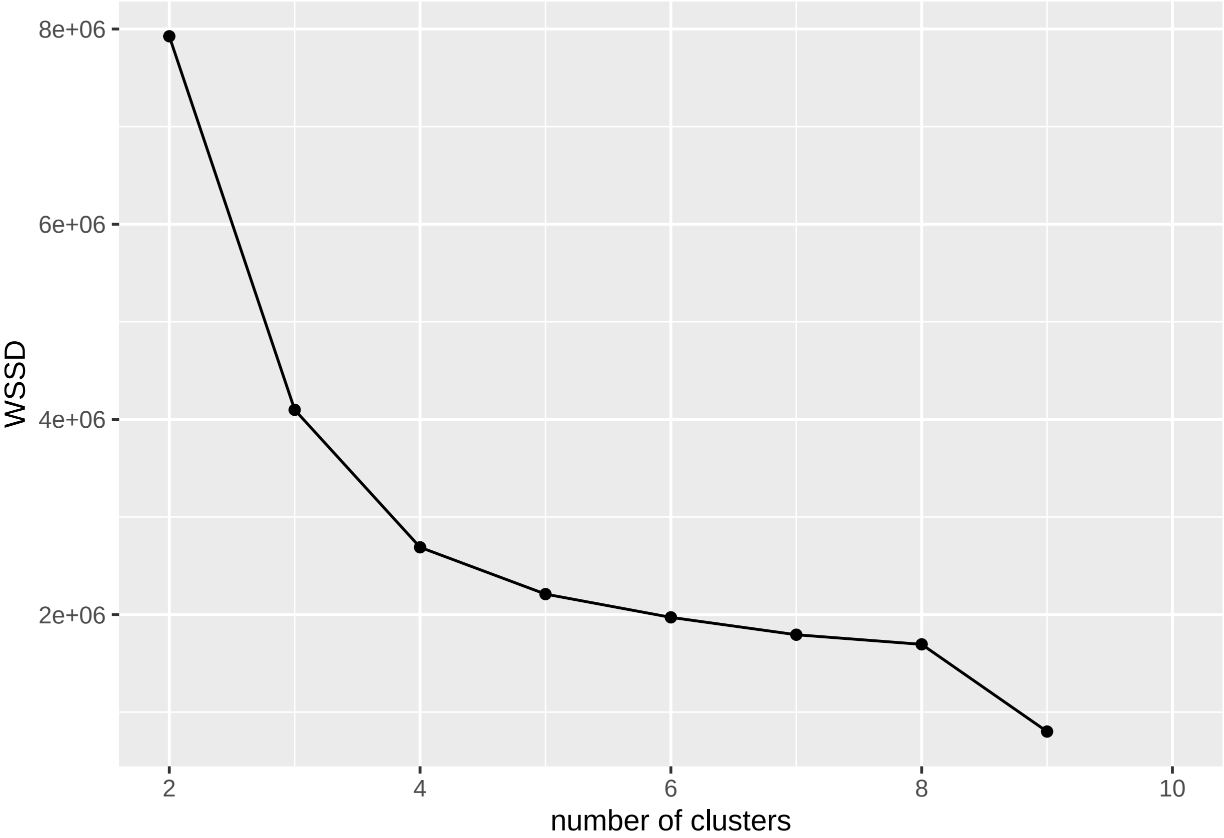
Question 3 Find a meaningful clustering of matches (25 of 100 points)

```
df =  read.csv("Gamelog T20I Stat 847.csv")
df6 = df
df_matches = ddply( df6, "MatchNo" , summarize ,
total_runs = sum( pmax(NumOutcome, 0)) ,
 fielder_mentions = length(which(Fielder != "")),
balls_until_1st_wicket = length(which(Wickets == 0)) ,
average_wickets_in_during_match = mean(Wickets, na.rm=TRUE),
format=Format[1],
teamBowling = TeamBowling[1],
teamBatting = TeamBowling[1]
)
```

#Games with Early Wickets, High scoring, Average Wickets and Fielder Mentions

```
df_kmeans = subset(df_matches, select = c(total_runs, average_wickets_in_during_match,balls_until_1st_wi
                                      ))

df_kmeans = na.omit(df_kmeans)
wssd <- rep(NA,9)

for(k in 2:9)
 {
     emo_clust <- kmeans(df_kmeans, centers = k)
     wssd[k-1] <- emo_clust$tot.withinss
}

 centers <- 2:10
 dat <- data.frame(centers, wssd)
 gr3 <- ggplot(dat, aes(x=centers, y=wssd)) +
     geom_line() +
     geom_point() +
     xlab("number of clusters") +
     ylab("WSSD")
 plot(gr3)
```

```
## Warning: Removed 1 row(s) containing missing values (geom_path).

## Warning: Removed 1 rows containing missing values (geom_point).
```



```
#Choosing number of clusters as 4

k_mean_cluster <- kmeans(df_kmeans, centers = 4)
k_mean_cluster$centers
```

```
##   total_runs average_wickets_in_during_match balls_until_1st_wicket
## 1   697.6667                        2.862072               98.29167
## 2   243.0105                        2.726626               41.16084
## 3  4319.0000                        2.471311               50.00000
## 4   335.7919                        2.353010               53.57297
##   fielder_mentions
## 1         48.41667
## 2         20.45105
## 3         27.00000
## 4         22.81892
```

```
msd <- sqrt(k_mean_cluster$withinss / k_mean_cluster$size)
```

Approach: In this question, k-means clustering is used for the Games with Early Wickets, High scoring, Average Wickets and Fielder Mentions. First of all, "MatchNo" is used for identifying unique values. Then we summarize the data for those rows into new variables. NA values are handled and a graph for choosing the right number of clusters is used.

Justification on Choice of Cluster: Within Cluster Sum of Squares (WCSS) measures the squared average distance of all the points within a cluster to the cluster centroid. K-means consists of two major steps that attempt to minimize the sum of WSSDs over all the clusters. If we plot the total WSSD versus the number of clusters, we see that the decrease in total WSSD levels off (or forms an "elbow shape") when we reach roughly the right number of clusters. In our graph, cluster size=4 is optimal based on the elbow method.

Features of Each Cluster - Games with Early Wickets, High scoring, Average Wickets and Fielder Mentions

Interpretation: - In Cluster 1: Highest scoring games with around 50 balls until the first wicket and around 27 fielder mentions and 2.4 average wickets per match balls are present in cluster 1.

- In Cluster 2: Games with scores of around 335 with 53 balls until the first wicket and around 22 fielder mentions and 2.3 average wickets per match balls are present in cluster 2.
- In Cluster 3: Games with scores around 700 but with highest balls until frst wicket and highest fielder mentions with highest average wickets during match are present in Cluster 3.
- In Cluster 4: Least scoring games with least balls until the first wicket and least field mentions with 2.7 average wickets during match are present in Cluster 4.

Question 4 Optimize Duckworth-Lewis

```
df10 = read.csv("Gamelog T20I Stat 847.csv")
df10 = subset(df10, !is.na(MatchNo))
df10 = subset(df10, Inning %in% c(1))
df10$over2 <- df10$Over + df10$Ball/6
matches <- unique(df10$MatchNo)

df <- data.frame(Match=numeric(),
                 Innings1=numeric())
count <- 0
Runs1 <- 0

#Calculate the table for total runs scored in first innings of each match
for(i in matches)
  {
  dataset <- subset(df10, MatchNo %in% i)
  dataset <- subset(dataset, !is.na(NumOutcome))
  Runs1 <- 0
  for(j in 1:nrow(dataset))
    {
    row <- dataset[j,]
    Runs1 <- Runs1 + row$NumOutcome
  }
  new <- c(i, Runs1)
  df[nrow(df) + 1, ] <- new
}


count <- 0
Runs1 <- 0
prop <- c()
counter <- 1


#Calculate the proportion


for(i in matches)
  {
  dataset <- subset(df10, MatchNo %in% i)
  dataset <- subset(dataset, !is.na(NumOutcome))

  Runs1 <- 0
  for(j in 1:nrow(dataset))
    {
    row <- dataset[j,]
    Runs1 <- Runs1 + row$NumOutcome
    if(df10$MatchNo[counter]==i)
    {

    prop <- c(prop, Runs1/df$Innings1[counter])
    }
  }
  counter <- counter + 1
}
```

#Define Loss Function

```
loss_function = function(x, prop)
{

A = x[1]
B = x[2]
C = x[3]
D = x[4] #Interaction variable

prop_smooth = A*(df10$over2) + B*((df10$Wickets)) + C*(df10$over2)^2 + D*(df10$Wickets*df10$over2)


error = sum( (prop - prop_smooth)^2)
return(error)

}
```

```
#Reference Additional Guidance File
options(warn = -1)
best_params = optim(par=c(0,0,0,0), loss_function, prop=prop)$par

A = best_params[1]
```

```r
B = best_params[2]
C = best_params[3]
D = best_params[4]


newDLT = matrix(NA, nrow=20, ncol=10)

for(overcount in 1:20)
{

newDLT[overcount,] = 1 - A*overcount + B*(9:0) + C*overcount^2 + D*(9:0)
}

range = max(newDLT) - min(newDLT)
newDLT2 = round((newDLT - min(newDLT)) / range, 3)
newDLT2
```

```
##        [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10]
##  [1,] 1.000 0.999 0.997 0.996 0.995 0.993 0.992 0.991 0.989 0.988
##  [2,] 0.969 0.968 0.966 0.965 0.964 0.962 0.961 0.960 0.958 0.957
##  [3,] 0.936 0.934 0.933 0.932 0.930 0.929 0.928 0.926 0.925 0.924
##  [4,] 0.900 0.898 0.897 0.896 0.895 0.893 0.892 0.891 0.889 0.888
##  [5,] 0.862 0.860 0.859 0.858 0.856 0.855 0.854 0.852 0.851 0.850
##  [6,] 0.821 0.820 0.819 0.817 0.816 0.815 0.813 0.812 0.811 0.809
##  [7,] 0.779 0.777 0.776 0.775 0.773 0.772 0.771 0.769 0.768 0.767
##  [8,] 0.734 0.732 0.731 0.730 0.728 0.727 0.726 0.724 0.723 0.722
##  [9,] 0.686 0.685 0.684 0.682 0.681 0.680 0.678 0.677 0.676 0.674
## [10,] 0.637 0.635 0.634 0.633 0.631 0.630 0.629 0.627 0.626 0.625
## [11,] 0.585 0.583 0.582 0.581 0.579 0.578 0.577 0.575 0.574 0.573
## [12,] 0.530 0.529 0.528 0.526 0.525 0.524 0.522 0.521 0.520 0.518
## [13,] 0.474 0.472 0.471 0.470 0.468 0.467 0.466 0.464 0.463 0.462
## [14,] 0.415 0.413 0.412 0.411 0.409 0.408 0.407 0.405 0.404 0.403
## [15,] 0.353 0.352 0.351 0.349 0.348 0.347 0.345 0.344 0.343 0.341
## [16,] 0.290 0.288 0.287 0.286 0.284 0.283 0.282 0.280 0.279 0.278
## [17,] 0.224 0.222 0.221 0.220 0.218 0.217 0.216 0.214 0.213 0.212
## [18,] 0.156 0.154 0.153 0.152 0.150 0.149 0.148 0.146 0.145 0.144
## [19,] 0.085 0.084 0.082 0.081 0.080 0.078 0.077 0.076 0.074 0.073
## [20,] 0.012 0.011 0.009 0.008 0.007 0.005 0.004 0.003 0.001 0.000
```

Inference:

Resource numbers represent the proportion of runs that a team is expected to still score in a match, given the current overs and wickets lost. For example, teams at the beginning of their 7th over with 3 wickets lost have 0.776 resource.

The 'optim' function was used to optimize the Duckworth Lewis Table (DLT). First of all, the total runs were calculated along with the proportion of runs scored at a particular ball for a team. The smoothing function involves 4 variables: prop_smooth = A$(df10over2) + B * ((df10Wickets)) + C(\mathrm{df10}over2)^2 + D *$ $(df$10Wickets*df10$over2).

Comparison:

Comparing the obtained DLT to the given DLT, it can be observed somewhat similar results have been obtained considering the 12th over, 50% of the resources have been utilized while scoring 50% of the runs. At 12th over 4th wicket, 54.6% resources are utilised as compared 11th over and 4th wicket. This proves that the Duckworth Lewis Table has been optimized.