Data Taming Assignment 1

Dongju Ma

Date you finished your assignment

Setup

```
#Load the required packages
library(tidyverse)
library(inspectdf)
library(lubridate)
library(caret)
library(moments)
library(tidymodels)
library(ISLR)
library(car)
```

Q1. Loading the data

```
# Your student number goes here
ysn = 1942340
# Calculate your student number modulo 3
filenum <- ysn %% 3
filenum
```

```
## [1] 2
```

```
filename <- paste0("./data/afl_",filenum,".csv")
filename
```

```
## [1] "./data/afl_2.csv"
```

```
# Read in the data
afl<-read_csv("./data/afl_2.csv")
# Display the first 10 lines of the data
head(afl,10)
```

```
## # A tibble: 10 x 24
##    Team    State Round01 Round02 Round03 Round04 Round05 Round06 Round07 Round08
##    <chr>   <chr> <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>
##  1 Collin~ VIC   away g~ home g~ away g~ home g~ home g~ away g~ home g~ away g~
##  2 St Kil~ VIC   away g~ home g~ home g~ home g~ away g~ away g~ home g~ home g~
##  3 Carlton VIC   away g~ away g~ home g~ away g~ home g~ home g~ away g~ away g~
##  4 North ~ VIC   away g~ away g~ home g~ home g~ away g~ home g~ away g~ home g~
##  5 Essend~ VIC   away g~ home g~ away g~ away g~ away g~ home g~ home g~ away g~
##  6 Melbou~ VIC   home g~ away g~ home g~ away g~ home g~ away g~ home g~ home g~
##  7 Hawtho~ bict~ away g~ home g~ away g~ away g~ home g~ away g~ away g~ away g~
##  8 Wester~ VIC   home g~ away g~ home g~ away g~ home g~ home g~ away g~ home g~
##  9 testX1  test~ testX1  testX1  testX1  testX1  testX1  testX1  testX1  testX1
## 10 Geelong VIC   home g~ away g~ away g~ home g~ away g~ home g~ home g~ away g~
## # i 14 more variables: Round09 <chr>, Round10 <chr>, Round11 <chr>,
## #   Round12 <chr>, Round13 <chr>, Round14 <chr>, Round15 <chr>, Round16 <chr>,
## #   Round17 <chr>, Round18 <chr>, Round19 <chr>, Round20 <chr>, Round21 <chr>,
## #   Round22 <chr>
```

Q2. The dimensions of the data set

```r
#Use dim to show the numbers of rows and columns
dim(afl)
```

```
## [1] 18 24
```

The data set has 18 rows and 24 columns.

Q3. Random permutation of the rows

```r
# Set the random seed
set.seed(1942340)
# Use sample_n to get the random permutation of the rows
afl1<-sample_n(afl,18,replace = FALSE)
afl1
```

```
## # A tibble: 18 x 24
##    Team    State Round01 Round02 Round03 Round04 Round05 Round06 Round07 Round08
##    <chr>   <chr> <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>
##  1 Carlton VIC   away g~ away g~ home g~ away g~ home g~ home g~ away g~ away g~
##  2 Port A~ SA    home g~ away g~ home g~ away g~ home g~ away g~ away g~ home g~
##  3 Geelong VIC   home g~ away g~ away g~ home g~ away g~ home g~ home g~ away g~
##  4 Brisba~ Quee~ home g~ home g~ away g~ home g~ away g~ away g~ home g~ home g~
##  5 Freman~ WA    home g~ away g~ home g~ away g~ home g~ away g~ away g~ home g~
##  6 testX1  test~ testX1  testX1  testX1  testX1  testX1  testX1  testX1  testX1
##  7 Collin~ VIC   away g~ home g~ away g~ home g~ home g~ away g~ home g~ away g~
##  8 West C~ WA    away g~ home g~ away g~ home g~ away g~ home g~ home g~ away g~
##  9 St Kil~ VIC   away g~ home g~ home g~ home g~ away g~ away g~ home g~ home g~
## 10 Adelai~ New ~ away g~ home g~ away g~ home g~ away g~ home g~ home g~ away g~
## 11 Carlton VIC   away g~ away g~ home g~ away g~ home g~ home g~ away g~ away g~
## 12 Richmo~ VIC   home g~ home g~ away g~ home g~ away g~ away g~ away g~ home g~
## 13 Sydney  NSW   home g~ away g~ home g~ away g~ home g~ home g~ away g~ away g~
## 14 North ~ VIC   away g~ away g~ home g~ home g~ away g~ home g~ away g~ home g~
## 15 Melbou~ VIC   home g~ away g~ home g~ away g~ home g~ away g~ home g~ home g~
## 16 Hawtho~ bict~ away g~ home g~ away g~ away g~ home g~ away g~ away g~ away g~
## 17 Wester~ VIC   home g~ away g~ home g~ away g~ home g~ home g~ away g~ home g~
## 18 Essend~ VIC   away g~ home g~ away g~ away g~ away g~ home g~ home g~ away g~
## # i 14 more variables: Round09 <chr>, Round10 <chr>, Round11 <chr>,
## #   Round12 <chr>, Round13 <chr>, Round14 <chr>, Round15 <chr>, Round16 <chr>,
## #   Round17 <chr>, Round18 <chr>, Round19 <chr>, Round20 <chr>, Round21 <chr>,
## #   Round22 <chr>
```

Q4. Adding an extra column of row numbers

```
# Use mutate to add a column at the far right of the data set
afl1<-mutate(afl1,RowNum=c(1:18))
# Then use relocate to move the new column to the far left
afl1<-relocate(afl1,"RowNum", .before = Team)
afl1
```

```
## # A tibble: 18 x 25
##    RowNum Team     State Round01 Round02 Round03 Round04 Round05 Round06 Round07
##     <int> <chr>    <chr> <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>
##  1      1 Carlton  VIC   away g~ away g~ home g~ away g~ home g~ home g~ away g~
##  2      2 Port Ad~ SA    home g~ away g~ home g~ away g~ home g~ away g~ away g~
##  3      3 Geelong  VIC   home g~ away g~ away g~ home g~ away g~ home g~ home g~
##  4      4 Brisban~ Quee~ home g~ home g~ away g~ home g~ away g~ away g~ home g~
##  5      5 Fremant~ WA    home g~ away g~ home g~ away g~ home g~ away g~ away g~
##  6      6 testX1   test~ testX1  testX1  testX1  testX1  testX1  testX1  testX1
##  7      7 Colling~ VIC   away g~ home g~ away g~ home g~ home g~ away g~ home g~
##  8      8 West Co~ WA    away g~ home g~ away g~ home g~ away g~ home g~ home g~
##  9      9 St Kilda VIC   away g~ home g~ home g~ home g~ away g~ away g~ home g~
## 10     10 Adelaide New ~ away g~ home g~ away g~ home g~ away g~ home g~ home g~
## 11     11 Carlton  VIC   away g~ away g~ home g~ away g~ home g~ home g~ away g~
## 12     12 Richmond VIC   home g~ home g~ away g~ home g~ away g~ away g~ away g~
## 13     13 Sydney   NSW   home g~ away g~ home g~ away g~ home g~ home g~ away g~
## 14     14 North M~ VIC   away g~ away g~ home g~ home g~ away g~ home g~ away g~
## 15     15 Melbour~ VIC   home g~ away g~ home g~ away g~ home g~ away g~ home g~
## 16     16 Hawthorn bict~ away g~ home g~ away g~ away g~ home g~ away g~ away g~
## 17     17 Western~ VIC   home g~ away g~ home g~ away g~ home g~ home g~ away g~
## 18     18 Essendon VIC   away g~ home g~ away g~ away g~ away g~ home g~ home g~
## # i 15 more variables: Round08 <chr>, Round09 <chr>, Round10 <chr>,
## #   Round11 <chr>, Round12 <chr>, Round13 <chr>, Round14 <chr>, Round15 <chr>,
## #   Round16 <chr>, Round17 <chr>, Round18 <chr>, Round19 <chr>, Round20 <chr>,
## #   Round21 <chr>, Round22 <chr>
```

Q5 Data cleaning

Q5(a)

```
# Use filter to extract the rows without text data.
afl1<-filter(afl1,Team!="testX1")
# Make sure the row numbers are updated
afl1<-mutate(afl1,Rownumber=c(1:17))
afl1
```

```
## # A tibble: 17 x 26
##    RowNum Team     State Round01 Round02 Round03 Round04 Round05 Round06 Round07
##     <int> <chr>    <chr> <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>
##  1      1 Carlton  VIC   away g~ away g~ home g~ away g~ home g~ home g~ away g~
##  2      2 Port Ad~ SA    home g~ away g~ home g~ away g~ home g~ away g~ away g~
##  3      3 Geelong  VIC   home g~ away g~ away g~ home g~ away g~ home g~ home g~
##  4      4 Brisban~ Quee~ home g~ home g~ away g~ home g~ away g~ away g~ home g~
##  5      5 Fremant~ WA    home g~ away g~ home g~ away g~ home g~ away g~ away g~
##  6      7 Colling~ VIC   away g~ home g~ away g~ home g~ home g~ away g~ home g~
##  7      8 West Co~ WA    away g~ home g~ away g~ home g~ away g~ home g~ home g~
##  8      9 St Kilda VIC   away g~ home g~ home g~ home g~ away g~ away g~ home g~
##  9     10 Adelaide New ~ away g~ home g~ away g~ home g~ away g~ home g~ home g~
## 10     11 Carlton  VIC   away g~ away g~ home g~ away g~ home g~ home g~ away g~
## 11     12 Richmond VIC   home g~ home g~ away g~ home g~ away g~ away g~ away g~
## 12     13 Sydney   NSW   home g~ away g~ home g~ away g~ home g~ home g~ away g~
## 13     14 North M~ VIC   away g~ away g~ home g~ home g~ away g~ home g~ away g~
## 14     15 Melbour~ VIC   home g~ away g~ home g~ away g~ home g~ away g~ home g~
## 15     16 Hawthorn bict~ away g~ home g~ away g~ away g~ home g~ away g~ away g~
## 16     17 Western~ VIC   home g~ away g~ home g~ away g~ home g~ home g~ away g~
## 17     18 Essendon VIC   away g~ home g~ away g~ away g~ away g~ home g~ home g~
## # i 16 more variables: Round08 <chr>, Round09 <chr>, Round10 <chr>,
## #   Round11 <chr>, Round12 <chr>, Round13 <chr>, Round14 <chr>, Round15 <chr>,
## #   Round16 <chr>, Round17 <chr>, Round18 <chr>, Round19 <chr>, Round20 <chr>,
## #   Round21 <chr>, Round22 <chr>, Rownumber <int>
```

Q5(b)

```
# Change Team name "Adelaide" to "Port Adelaide"
afl1[9,]$Team<-str_replace(afl1[9,]$Team,"Adelaide","Port Adelaide")
# Change Team name "Melbourne" to "North Melbourne"
afl1[14,]$Team<-str_replace(afl1[14,]$Team,"Melbourne","North Melbourne")
# Change State "Queensld" to "QLD"
afl1[4,]$State<-str_replace(afl1[4,]$State,"Queensld","QLD")
# Change State "New South Wales" to "SA"
afl1[9,]$State<-str_replace(afl1[9,]$State,"New South Wales","SA")
# Change State "bictoria" to "VIC"
afl1[15,]$State<-str_replace(afl1[15,]$State,"bictoria","VIC")
afl1
```

```
## # A tibble: 17 x 26
##    RowNum Team     State Round01 Round02 Round03 Round04 Round05 Round06 Round07
##     <int> <chr>    <chr> <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>
##  1      1 Carlton  VIC   away g~ away g~ home g~ away g~ home g~ home g~ away g~
##  2      2 Port Ad~ SA    home g~ away g~ home g~ away g~ home g~ away g~ away g~
##  3      3 Geelong  VIC   home g~ away g~ away g~ home g~ away g~ home g~ home g~
##  4      4 Brisban~ QLD   home g~ home g~ away g~ home g~ away g~ away g~ home g~
##  5      5 Fremant~ WA    home g~ away g~ home g~ away g~ home g~ away g~ away g~
##  6      7 Colling~ VIC   away g~ home g~ away g~ home g~ home g~ away g~ home g~
##  7      8 West Co~ WA    away g~ home g~ away g~ home g~ away g~ home g~ home g~
##  8      9 St Kilda VIC   away g~ home g~ home g~ home g~ away g~ away g~ home g~
##  9     10 Port Ad~ SA    away g~ home g~ away g~ home g~ away g~ home g~ home g~
## 10     11 Carlton  VIC   away g~ away g~ home g~ away g~ home g~ home g~ away g~
## 11     12 Richmond VIC   home g~ home g~ away g~ home g~ away g~ away g~ away g~
## 12     13 Sydney   NSW   home g~ away g~ home g~ away g~ home g~ home g~ away g~
## 13     14 North M~ VIC   away g~ away g~ home g~ home g~ away g~ home g~ away g~
## 14     15 North M~ VIC   home g~ away g~ home g~ away g~ home g~ away g~ home g~
## 15     16 Hawthorn VIC   away g~ home g~ away g~ away g~ home g~ away g~ away g~
## 16     17 Western~ VIC   home g~ away g~ home g~ away g~ home g~ home g~ away g~
## 17     18 Essendon VIC   away g~ home g~ away g~ away g~ away g~ home g~ home g~
## # i 16 more variables: Round08 <chr>, Round09 <chr>, Round10 <chr>,
## #   Round11 <chr>, Round12 <chr>, Round13 <chr>, Round14 <chr>, Round15 <chr>,
## #   Round16 <chr>, Round17 <chr>, Round18 <chr>, Round19 <chr>, Round20 <chr>,
## #   Round21 <chr>, Round22 <chr>, Rownumber <int>
```

Q5(c)

```
# Use arrange to sort the tibble by team name
afl1<-arrange(afl1,Team)
afl1
```

```
## # A tibble: 17 x 26
##    RowNum Team     State Round01 Round02 Round03 Round04 Round05 Round06 Round07
##     <int> <chr>    <chr> <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>
##  1      4 Brisban~ QLD   home g~ home g~ away g~ home g~ away g~ away g~ home g~
##  2      1 Carlton  VIC   away g~ away g~ home g~ away g~ home g~ home g~ away g~
##  3     11 Carlton  VIC   away g~ away g~ home g~ away g~ home g~ home g~ away g~
##  4      7 Colling~ VIC   away g~ home g~ away g~ home g~ home g~ away g~ home g~
##  5     18 Essendon VIC   away g~ home g~ away g~ away g~ away g~ home g~ home g~
##  6      5 Fremant~ WA    home g~ away g~ home g~ away g~ home g~ away g~ away g~
##  7      3 Geelong  VIC   home g~ away g~ away g~ home g~ away g~ home g~ home g~
##  8     16 Hawthorn VIC   away g~ home g~ away g~ away g~ home g~ away g~ away g~
##  9     14 North M~ VIC   away g~ away g~ home g~ home g~ away g~ home g~ away g~
## 10     15 North M~ VIC   home g~ away g~ home g~ away g~ home g~ away g~ home g~
## 11      2 Port Ad~ SA    home g~ away g~ home g~ away g~ home g~ away g~ away g~
## 12     10 Port Ad~ SA    away g~ home g~ away g~ home g~ away g~ home g~ home g~
## 13     12 Richmond VIC   home g~ home g~ away g~ home g~ away g~ away g~ away g~
## 14      9 St Kilda VIC   away g~ home g~ home g~ home g~ away g~ away g~ home g~
## 15     13 Sydney   NSW   home g~ away g~ home g~ away g~ home g~ home g~ away g~
## 16      8 West Co~ WA    away g~ home g~ away g~ home g~ away g~ home g~ home g~
## 17     17 Western~ VIC   home g~ away g~ home g~ away g~ home g~ home g~ away g~
## # i 16 more variables: Round08 <chr>, Round09 <chr>, Round10 <chr>,
## #   Round11 <chr>, Round12 <chr>, Round13 <chr>, Round14 <chr>, Round15 <chr>,
## #   Round16 <chr>, Round17 <chr>, Round18 <chr>, Round19 <chr>, Round20 <chr>,
## #   Round21 <chr>, Round22 <chr>, Rownumber <int>
```

Q6

Q6(a)

```
# Use gather to convert the data set to long form
afl1<- gather(afl1,key = "round",value = "details",'Round01':'Round22')
afl1
```

```
## # A tibble: 374 x 6
##    RowNum Team            State Rownumber round   details
##     <int> <chr>           <chr>     <int> <chr>   <chr>
##  1      4 Brisbane Lions  QLD           4 Round01 home game, scored 16 goals an~
##  2      1 Carlton         VIC           1 Round01 away game, scored 18 goals an~
##  3     11 Carlton         VIC          10 Round01 away game, scored 18 goals an~
##  4      7 Collingwood     VIC           6 Round01 away game, scored 19 goals an~
##  5     18 Essendon        VIC          17 Round01 away game, scored 13 goals an~
##  6      5 Fremantle       WA            5 Round01 home game, scored 17 goals an~
##  7      3 Geelong         VIC           3 Round01 home game, scored 19 goals an~
##  8     16 Hawthorn        VIC          15 Round01 away game, scored 17 goals an~
##  9     14 North Melbourne VIC          13 Round01 away game, scored 12 goals an~
## 10     15 North Melbourne VIC          14 Round01 home game, scored 8 goals and~
## # i 364 more rows
```

Q6(b)

```
# Use sting replace to remove all the "Round" string in column round
afl1$round<-str_replace(afl1$round,"Round","")
afl1
```

```
## # A tibble: 374 x 6
##    RowNum Team            State Rownumber round details
##     <int> <chr>           <chr>     <int> <chr> <chr>
##  1      4 Brisbane Lions  QLD           4 01    home game, scored 16 goals and ~
##  2      1 Carlton         VIC           1 01    away game, scored 18 goals and ~
##  3     11 Carlton         VIC          10 01    away game, scored 18 goals and ~
##  4      7 Collingwood     VIC           6 01    away game, scored 19 goals and ~
##  5     18 Essendon        VIC          17 01    away game, scored 13 goals and ~
##  6      5 Fremantle       WA            5 01    home game, scored 17 goals and ~
##  7      3 Geelong         VIC           3 01    home game, scored 19 goals and ~
##  8     16 Hawthorn        VIC          15 01    away game, scored 17 goals and ~
##  9     14 North Melbourne VIC          13 01    away game, scored 12 goals and ~
## 10     15 North Melbourne VIC          14 01    home game, scored 8 goals and 1~
## # i 364 more rows
```

Q6(c)

```
afl1<-afl1 %>%
  mutate("home"=is.na(str_match(afl1$details,"away"))[,1])
afl1
```

```
## # A tibble: 374 x 7
##    RowNum Team            State Rownumber round details                    home
##     <int> <chr>           <chr>     <int> <chr> <chr>                      <lgl>
##  1      4 Brisbane Lions  QLD           4 01    home game, scored 16 goal~ TRUE
##  2      1 Carlton         VIC           1 01    away game, scored 18 goal~ FALSE
##  3     11 Carlton         VIC          10 01    away game, scored 18 goal~ FALSE
##  4      7 Collingwood     VIC           6 01    away game, scored 19 goal~ FALSE
##  5     18 Essendon        VIC          17 01    away game, scored 13 goal~ FALSE
##  6      5 Fremantle       WA            5 01    home game, scored 17 goal~ TRUE
##  7      3 Geelong         VIC           3 01    home game, scored 19 goal~ TRUE
##  8     16 Hawthorn        VIC          15 01    away game, scored 17 goal~ FALSE
##  9     14 North Melbourne VIC          13 01    away game, scored 12 goal~ FALSE
## 10     15 North Melbourne VIC          14 01    home game, scored 8 goals~ TRUE
## # i 364 more rows
```

Q6(d)

```
afl1<-mutate(afl1,goals=str_match(afl1$details,"(\\d+) goals and (\\d+)")[,2])
afl1<-mutate(afl1,behinds=str_match(afl1$details,"(\\d+) goals and (\\d+)")[,3])
afl1
```

```
## # A tibble: 374 x 9
##    RowNum Team            State Rownumber round details      home  goals behinds
##     <int> <chr>           <chr>     <int> <chr> <chr>        <lgl> <chr> <chr>
##  1      4 Brisbane Lions  QLD           4 01    home game, ~ TRUE  16    18
##  2      1 Carlton         VIC           1 01    away game, ~ FALSE 18    12
##  3     11 Carlton         VIC          10 01    away game, ~ FALSE 18    12
##  4      7 Collingwood     VIC           6 01    away game, ~ FALSE 19    15
##  5     18 Essendon        VIC          17 01    away game, ~ FALSE 13    16
##  6      5 Fremantle       WA            5 01    home game, ~ TRUE  17    16
##  7      3 Geelong         VIC           3 01    home game, ~ TRUE  19    11
##  8     16 Hawthorn        VIC          15 01    away game, ~ FALSE 17    15
##  9     14 North Melbourne VIC          13 01    away game, ~ FALSE 12    10
## 10     15 North Melbourne VIC          14 01    home game, ~ TRUE  8     13
## # i 364 more rows
```

Q6(e)

```
afl1<-mutate(afl1,details=NULL)
afl1
```

```
## # A tibble: 374 x 8
##    RowNum Team            State Rownumber round home  goals behinds
##     <int> <chr>           <chr>     <int> <chr> <lgl> <chr> <chr>
##  1      4 Brisbane Lions  QLD           4 01    TRUE  16    18
##  2      1 Carlton         VIC           1 01    FALSE 18    12
##  3     11 Carlton         VIC          10 01    FALSE 18    12
##  4      7 Collingwood     VIC           6 01    FALSE 19    15
##  5     18 Essendon        VIC          17 01    FALSE 13    16
##  6      5 Fremantle       WA            5 01    TRUE  17    16
##  7      3 Geelong         VIC           3 01    TRUE  19    11
##  8     16 Hawthorn        VIC          15 01    FALSE 17    15
##  9     14 North Melbourne VIC          13 01    FALSE 12    10
## 10     15 North Melbourne VIC          14 01    TRUE  8     13
## # i 364 more rows
```

Q6(f)

```r
afl1<-mutate(afl1,TidyRowNum=(1:374), .after=RowNum)
afl1
```

```
## # A tibble: 374 x 9
##    RowNum TidyRowNum Team            State Rownumber round home  goals behinds
##     <int>      <int> <chr>           <chr>     <int> <chr> <lgl> <chr> <chr>
##  1      4          1 Brisbane Lions  QLD           4 01    TRUE  16    18
##  2      1          2 Carlton         VIC           1 01    FALSE 18    12
##  3     11          3 Carlton         VIC          10 01    FALSE 18    12
##  4      7          4 Collingwood     VIC           6 01    FALSE 19    15
##  5     18          5 Essendon        VIC          17 01    FALSE 13    16
##  6      5          6 Fremantle       WA            5 01    TRUE  17    16
##  7      3          7 Geelong         VIC           3 01    TRUE  19    11
##  8     16          8 Hawthorn        VIC          15 01    FALSE 17    15
##  9     14          9 North Melbourne VIC          13 01    FALSE 12    10
## 10     15         10 North Melbourne VIC          14 01    TRUE  8     13
## # i 364 more rows
```

Q7. Identifying data types

- variable1: type and justification
- variable2: type and justification
- etc

Q8. Taming the data

etc.

etc.

etc.

etc.