# 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

```
#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

```
# 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,Rownumber=c(1:18))
# Then use relocate to move the new column to the far left
afl1<-relocate(afl1,"Rownumber", .before = Team)
afl1
```

```
## # A tibble: 18 x 25
##    Rownumber Team  State Round01 Round02 Round03 Round04 Round05 Round06 Round07
##        <int> <chr> <chr> <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>
##  1         1 Carl~ VIC   away g~ away g~ home g~ away g~ home g~ home g~ away g~
##  2         2 Port~ SA    home g~ away g~ home g~ away g~ home g~ away g~ away g~
##  3         3 Geel~ VIC   home g~ away g~ away g~ home g~ away g~ home g~ home g~
##  4         4 Bris~ Quee~ home g~ home g~ away g~ home g~ away g~ away g~ home g~
##  5         5 Frem~ WA    home g~ away g~ home g~ away g~ home g~ away g~ away g~
##  6         6 test~ test~ testX1  testX1  testX1  testX1  testX1  testX1  testX1
##  7         7 Coll~ VIC   away g~ home g~ away g~ home g~ home g~ away g~ home g~
##  8         8 West~ WA    away g~ home g~ away g~ home g~ away g~ home g~ home g~
##  9         9 St K~ VIC   away g~ home g~ home g~ home g~ away g~ away g~ home g~
## 10        10 Adel~ New ~ away g~ home g~ away g~ home g~ away g~ home g~ home g~
## 11        11 Carl~ VIC   away g~ away g~ home g~ away g~ home g~ home g~ away g~
## 12        12 Rich~ VIC   home g~ home g~ away g~ home g~ away g~ away g~ away g~
## 13        13 Sydn~ NSW   home g~ away g~ home g~ away g~ home g~ home g~ away g~
## 14        14 Nort~ VIC   away g~ away g~ home g~ home g~ away g~ home g~ away g~
## 15        15 Melb~ VIC   home g~ away g~ home g~ away g~ home g~ away g~ home g~
## 16        16 Hawt~ bict~ away g~ home g~ away g~ away g~ home g~ away g~ away g~
## 17        17 West~ VIC   home g~ away g~ home g~ away g~ home g~ home g~ away g~
## 18        18 Esse~ 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 25
##    Rownumber Team  State Round01 Round02 Round03 Round04 Round05 Round06 Round07
##        <int> <chr> <chr> <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>
##  1         1 Carl~ VIC   away g~ away g~ home g~ away g~ home g~ home g~ away g~
##  2         2 Port~ SA    home g~ away g~ home g~ away g~ home g~ away g~ away g~
##  3         3 Geel~ VIC   home g~ away g~ away g~ home g~ away g~ home g~ home g~
##  4         4 Bris~ Quee~ home g~ home g~ away g~ home g~ away g~ away g~ home g~
##  5         5 Frem~ WA    home g~ away g~ home g~ away g~ home g~ away g~ away g~
##  6         6 Coll~ VIC   away g~ home g~ away g~ home g~ home g~ away g~ home g~
##  7         7 West~ WA    away g~ home g~ away g~ home g~ away g~ home g~ home g~
##  8         8 St K~ VIC   away g~ home g~ home g~ home g~ away g~ away g~ home g~
##  9         9 Adel~ New ~ away g~ home g~ away g~ home g~ away g~ home g~ home g~
## 10        10 Carl~ VIC   away g~ away g~ home g~ away g~ home g~ home g~ away g~
## 11        11 Rich~ VIC   home g~ home g~ away g~ home g~ away g~ away g~ away g~
## 12        12 Sydn~ NSW   home g~ away g~ home g~ away g~ home g~ home g~ away g~
## 13        13 Nort~ VIC   away g~ away g~ home g~ home g~ away g~ home g~ away g~
## 14        14 Melb~ VIC   home g~ away g~ home g~ away g~ home g~ away g~ home g~
## 15        15 Hawt~ bict~ away g~ home g~ away g~ away g~ home g~ away g~ away g~
## 16        16 West~ VIC   home g~ away g~ home g~ away g~ home g~ home g~ away g~
## 17        17 Esse~ 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(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 25
##    Rownumber Team  State Round01 Round02 Round03 Round04 Round05 Round06 Round07
##        <int> <chr> <chr> <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>
##  1         1 Carl~ VIC   away g~ away g~ home g~ away g~ home g~ home g~ away g~
##  2         2 Port~ SA    home g~ away g~ home g~ away g~ home g~ away g~ away g~
##  3         3 Geel~ VIC   home g~ away g~ away g~ home g~ away g~ home g~ home g~
```

```
##  4         4 Bris~ QLD   home g~ home g~ away g~ home g~ away g~ away g~ home g~
##  5         5 Frem~ WA    home g~ away g~ home g~ away g~ home g~ away g~ away g~
##  6         6 Coll~ VIC   away g~ home g~ away g~ home g~ home g~ away g~ home g~
##  7         7 West~ WA    away g~ home g~ away g~ home g~ away g~ home g~ home g~
##  8         8 St K~ VIC   away g~ home g~ home g~ home g~ away g~ away g~ home g~
##  9         9 Port~ SA    away g~ home g~ away g~ home g~ away g~ home g~ home g~
## 10        10 Carl~ VIC   away g~ away g~ home g~ away g~ home g~ home g~ away g~
## 11        11 Rich~ VIC   home g~ home g~ away g~ home g~ away g~ away g~ away g~
## 12        12 Sydn~ NSW   home g~ away g~ home g~ away g~ home g~ home g~ away g~
## 13        13 Nort~ VIC   away g~ away g~ home g~ home g~ away g~ home g~ away g~
## 14        14 Nort~ VIC   home g~ away g~ home g~ away g~ home g~ away g~ home g~
## 15        15 Hawt~ VIC   away g~ home g~ away g~ away g~ home g~ away g~ away g~
## 16        16 West~ VIC   home g~ away g~ home g~ away g~ home g~ home g~ away g~
## 17        17 Esse~ 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(c)

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

```
## # A tibble: 17 x 25
##    Rownumber Team  State Round01 Round02 Round03 Round04 Round05 Round06 Round07
##        <int> <chr> <chr> <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>
##  1         4 Bris~ QLD   home g~ home g~ away g~ home g~ away g~ away g~ home g~
##  2         1 Carl~ VIC   away g~ away g~ home g~ away g~ home g~ home g~ away g~
##  3        10 Carl~ VIC   away g~ away g~ home g~ away g~ home g~ home g~ away g~
##  4         6 Coll~ VIC   away g~ home g~ away g~ home g~ home g~ away g~ home g~
##  5        17 Esse~ VIC   away g~ home g~ away g~ away g~ away g~ home g~ home g~
##  6         5 Frem~ WA    home g~ away g~ home g~ away g~ home g~ away g~ away g~
##  7         3 Geel~ VIC   home g~ away g~ away g~ home g~ away g~ home g~ home g~
##  8        15 Hawt~ VIC   away g~ home g~ away g~ away g~ home g~ away g~ away g~
##  9        13 Nort~ VIC   away g~ away g~ home g~ home g~ away g~ home g~ away g~
## 10        14 Nort~ VIC   home g~ away g~ home g~ away g~ home g~ away g~ home g~
## 11         2 Port~ SA    home g~ away g~ home g~ away g~ home g~ away g~ away g~
## 12         9 Port~ SA    away g~ home g~ away g~ home g~ away g~ home g~ home g~
## 13        11 Rich~ VIC   home g~ home g~ away g~ home g~ away g~ away g~ away g~
## 14         8 St K~ VIC   away g~ home g~ home g~ home g~ away g~ away g~ home g~
## 15        12 Sydn~ NSW   home g~ away g~ home g~ away g~ home g~ home g~ away g~
## 16         7 West~ WA    away g~ home g~ away g~ home g~ away g~ home g~ home g~
## 17        16 West~ VIC   home g~ away g~ home g~ away g~ home g~ home g~ away 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>
```

# 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 5
##    Rownumber Team            State round   details
##        <int> <chr>           <chr> <chr>   <chr>
##  1         4 Brisbane Lions  QLD   Round01 home game, scored 16 goals and 18 be~
##  2         1 Carlton         VIC   Round01 away game, scored 18 goals and 12 be~
##  3        10 Carlton         VIC   Round01 away game, scored 18 goals and 12 be~
##  4         6 Collingwood     VIC   Round01 away game, scored 19 goals and 15 be~
##  5        17 Essendon        VIC   Round01 away game, scored 13 goals and 16 be~
##  6         5 Fremantle       WA    Round01 home game, scored 17 goals and 16 be~
##  7         3 Geelong         VIC   Round01 home game, scored 19 goals and 11 be~
##  8        15 Hawthorn        VIC   Round01 away game, scored 17 goals and 15 be~
##  9        13 North Melbourne VIC   Round01 away game, scored 12 goals and 10 be~
## 10        14 North Melbourne VIC   Round01 home game, scored 8 goals and 13 beh~
## # 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 5
##    Rownumber Team            State round details
##        <int> <chr>           <chr> <chr> <chr>
##  1         4 Brisbane Lions  QLD   01    home game, scored 16 goals and 18 behi~
##  2         1 Carlton         VIC   01    away game, scored 18 goals and 12 behi~
##  3        10 Carlton         VIC   01    away game, scored 18 goals and 12 behi~
##  4         6 Collingwood     VIC   01    away game, scored 19 goals and 15 behi~
##  5        17 Essendon        VIC   01    away game, scored 13 goals and 16 behi~
##  6         5 Fremantle       WA    01    home game, scored 17 goals and 16 behi~
##  7         3 Geelong         VIC   01    home game, scored 19 goals and 11 behi~
##  8        15 Hawthorn        VIC   01    away game, scored 17 goals and 15 behi~
##  9        13 North Melbourne VIC   01    away game, scored 12 goals and 10 behi~
## 10        14 North Melbourne VIC   01    home game, scored 8 goals and 13 behin~
## # i 364 more rows
```

## Q6(C)

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

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