Data Taming Assignment 1

Dongju Ma

12/06/2024

Setup

```
#Load the required packages
library(tidyverse)
library(inspectdf)
```

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,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,RowNum=c(1:17))
afl1
```

```
## # A tibble: 17 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 Colling~ VIC   away g~ home g~ away g~ home g~ home g~ away g~ home g~
##  7      7 West Co~ WA    away g~ home g~ away g~ home g~ away g~ home g~ home g~
##  8      8 St Kilda VIC   away g~ home g~ home g~ home g~ away g~ away g~ home g~
##  9      9 Adelaide New ~ away g~ home g~ away g~ home g~ away g~ home g~ home g~
## 10     10 Carlton  VIC   away g~ away g~ home g~ away g~ home g~ home g~ away g~
## 11     11 Richmond VIC   home g~ home g~ away g~ home g~ away g~ away g~ away g~
## 12     12 Sydney   NSW   home g~ away g~ home g~ away g~ home g~ home g~ away g~
## 13     13 North M~ VIC   away g~ away g~ home g~ home g~ away g~ home g~ away g~
## 14     14 Melbour~ VIC   home g~ away g~ home g~ away g~ home g~ away g~ home g~
## 15     15 Hawthorn bict~ away g~ home g~ away g~ away g~ home g~ away g~ away g~
## 16     16 Western~ VIC   home g~ away g~ home g~ away g~ home g~ home g~ away g~
## 17     17 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(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
##    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      6 Colling~ VIC   away g~ home g~ away g~ home g~ home g~ away g~ home g~
##  7      7 West Co~ WA    away g~ home g~ away g~ home g~ away g~ home g~ home g~
##  8      8 St Kilda VIC   away g~ home g~ home g~ home g~ away g~ away g~ home g~
##  9      9 Port Ad~ SA    away g~ home g~ away g~ home g~ away g~ home g~ home g~
## 10     10 Carlton  VIC   away g~ away g~ home g~ away g~ home g~ home g~ away g~
```

```
## 11     11 Richmond VIC   home g~ home g~ away g~ home g~ away g~ away g~ away g~
## 12     12 Sydney   NSW   home g~ away g~ home g~ away g~ home g~ home g~ away g~
## 13     13 North M~ VIC   away g~ away g~ home g~ home g~ away g~ home g~ away g~
## 14     14 North M~ VIC   home g~ away g~ home g~ away g~ home g~ away g~ home g~
## 15     15 Hawthorn VIC   away g~ home g~ away g~ away g~ home g~ away g~ away g~
## 16     16 Western~ VIC   home g~ away g~ home g~ away g~ home g~ home g~ away g~
## 17     17 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(c)

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

```
## # A tibble: 17 x 25
##    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     10 Carlton  VIC   away g~ away g~ home g~ away g~ home g~ home g~ away g~
##  4      6 Colling~ VIC   away g~ home g~ away g~ home g~ home g~ away g~ home g~
##  5     17 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     15 Hawthorn VIC   away g~ home g~ away g~ away g~ home g~ away g~ away g~
##  9     13 North M~ VIC   away g~ away g~ home g~ home g~ away g~ home g~ away g~
## 10     14 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      9 Port Ad~ SA    away g~ home g~ away g~ home g~ away g~ home g~ home g~
## 13     11 Richmond VIC   home g~ home g~ away g~ home g~ away g~ away g~ away g~
## 14      8 St Kilda VIC   away g~ home g~ home g~ home g~ away g~ away g~ home g~
## 15     12 Sydney   NSW   home g~ away g~ home g~ away g~ home g~ home g~ away g~
## 16      7 West Co~ WA    away g~ home g~ away g~ home g~ away g~ home g~ home g~
## 17     16 Western~ 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
##    RowNum Team            State round   details
##     <int> <chr>           <chr> <chr>   <chr>
##  1      4 Brisbane Lions  QLD   Round01 home game, scored 16 goals and 18 behin~
##  2      1 Carlton         VIC   Round01 away game, scored 18 goals and 12 behin~
##  3     10 Carlton         VIC   Round01 away game, scored 18 goals and 12 behin~
##  4      6 Collingwood     VIC   Round01 away game, scored 19 goals and 15 behin~
##  5     17 Essendon        VIC   Round01 away game, scored 13 goals and 16 behin~
##  6      5 Fremantle       WA    Round01 home game, scored 17 goals and 16 behin~
##  7      3 Geelong         VIC   Round01 home game, scored 19 goals and 11 behin~
##  8     15 Hawthorn        VIC   Round01 away game, scored 17 goals and 15 behin~
##  9     13 North Melbourne VIC   Round01 away game, scored 12 goals and 10 behin~
## 10     14 North Melbourne VIC   Round01 home game, scored 8 goals and 13 behinds
## # 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
##    RowNum Team            State round details
##     <int> <chr>           <chr> <chr> <chr>
##  1      4 Brisbane Lions  QLD   01    home game, scored 16 goals and 18 behinds
##  2      1 Carlton         VIC   01    away game, scored 18 goals and 12 behinds
##  3     10 Carlton         VIC   01    away game, scored 18 goals and 12 behinds
##  4      6 Collingwood     VIC   01    away game, scored 19 goals and 15 behinds
##  5     17 Essendon        VIC   01    away game, scored 13 goals and 16 behinds
##  6      5 Fremantle       WA    01    home game, scored 17 goals and 16 behinds
##  7      3 Geelong         VIC   01    home game, scored 19 goals and 11 behinds
##  8     15 Hawthorn        VIC   01    away game, scored 17 goals and 15 behinds
##  9     13 North Melbourne VIC   01    away game, scored 12 goals and 10 behinds
## 10     14 North Melbourne VIC   01    home game, scored 8 goals and 13 behinds
## # i 364 more rows
```

Q6(c)

```
# Judge is away in details column, and rename the result column 1 into home
afl1<-afl1 %>%
  mutate("home"=is.na(str_match(afl1$details,"away"))[,1])
afl1
```

```
## # A tibble: 374 x 6
##    RowNum Team            State round details                              home
##     <int> <chr>           <chr> <chr> <chr>                                <lgl>
##  1      4 Brisbane Lions  QLD   01    home game, scored 16 goals and 18 b~ TRUE
##  2      1 Carlton         VIC   01    away game, scored 18 goals and 12 b~ FALSE
##  3     10 Carlton         VIC   01    away game, scored 18 goals and 12 b~ FALSE
##  4      6 Collingwood     VIC   01    away game, scored 19 goals and 15 b~ FALSE
```

```
##  5     17 Essendon        VIC   01    away game, scored 13 goals and 16 b~ FALSE
##  6      5 Fremantle       WA    01    home game, scored 17 goals and 16 b~ TRUE
##  7      3 Geelong         VIC   01    home game, scored 19 goals and 11 b~ TRUE
##  8     15 Hawthorn        VIC   01    away game, scored 17 goals and 15 b~ FALSE
##  9     13 North Melbourne VIC   01    away game, scored 12 goals and 10 b~ FALSE
## 10     14 North Melbourne VIC   01    home game, scored 8 goals and 13 be~ TRUE
## # i 364 more rows
```

Q6(d)

```
# Dig the numbers by str_match and put the result into column goals and column behinds
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 8
##    RowNum Team            State round details                home  goals behinds
##     <int> <chr>           <chr> <chr> <chr>                  <lgl> <chr> <chr>
##  1      4 Brisbane Lions  QLD   01    home game, scored 16 ~ TRUE  16    18
##  2      1 Carlton         VIC   01    away game, scored 18 ~ FALSE 18    12
##  3     10 Carlton         VIC   01    away game, scored 18 ~ FALSE 18    12
##  4      6 Collingwood     VIC   01    away game, scored 19 ~ FALSE 19    15
##  5     17 Essendon        VIC   01    away game, scored 13 ~ FALSE 13    16
##  6      5 Fremantle       WA    01    home game, scored 17 ~ TRUE  17    16
##  7      3 Geelong         VIC   01    home game, scored 19 ~ TRUE  19    11
##  8     15 Hawthorn        VIC   01    away game, scored 17 ~ FALSE 17    15
##  9     13 North Melbourne VIC   01    away game, scored 12 ~ FALSE 12    10
## 10     14 North Melbourne VIC   01    home game, scored 8 g~ TRUE  8     13
## # i 364 more rows
```

Q6(e)

```
# Delete the details column
afl1<-mutate(afl1,details=NULL)
afl1
```

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

Q6(f)

```
# Add the TidyRowNum column right next to the origin RowNum
afl1<-mutate(afl1,TidyRowNum=(1:374), .after=RowNum)
afl1
```

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

Q7. Identifying data types

- Row Num: Categorical Ordinal. The numbers represent the teams and round status is home or away. For example number 1 indicates team Carlton's away games.
- Tidy Row Num: Categorical Ordinal. The tidy row numbers are integers indicate the order of this data set.
- Team: Categorical Nominal. They are the names of teams in AFL.
- State: Categorical Nominal.. They are the names of the states.
- Round: Categorical Nominal. The characters represents the rounds in the match season, which is in the range of 01 to 22.
- home: Categorical Nominal. There are only two categories in this variables, TRUE means the game is home and FALSE means away.
- goals: Quantitative Discrete. The numbers are integers represent the goals' points in each game and they can be really huge theoretically.
- behinds: Quantitative Discrete. The numbers are integers represent the points in behinds and they can be really huge theoretically.

Q8. Taming the data

```
# Change the blank spaces in Team into "_"
afl1$Team<-str_replace(afl1$Team," ","_")
afl1
```

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

```
# Change the number characters into integers
afl1$round<-as.integer(afl1$round)
afl1$goals<-as.integer(afl1$goals)
afl1$behinds<-as.integer(afl1$behinds)
# Check if there is any NA
inspect_na(afl1)
```

```
## # A tibble: 8 x 3
##   col_name     cnt  pcnt
##   <chr>      <int> <dbl>
## 1 RowNum         0     0
## 2 TidyRowNum     0     0
## 3 Team           0     0
## 4 State          0     0
## 5 round          0     0
## 6 home           0     0
## 7 goals          0     0
## 8 behinds        0     0
```

Q9 Set the new data set

```
set.seed(1942340)
afl2<-sample_n(afl1,200)
afl2
```

```
## # A tibble: 200 x 8
##    RowNum TidyRowNum Team            State round home  goals behinds
##     <int>      <int> <chr>           <chr> <int> <lgl> <int>   <int>
##  1     12         15 Sydney          NSW       1 TRUE     13      10
##  2     14        299 North_Melbourne VIC      18 FALSE    11       8
##  3     16        170 Western_Bulldogs VIC     10 FALSE    14       6
##  4      9        301 Port_Adelaide   SA       18 FALSE    11      14
##  5      1        172 Carlton         VIC      11 TRUE     15      11
##  6      6        174 Collingwood     VIC      11 TRUE     17      11
##  7     12        338 Sydney          NSW      20 FALSE    14      12
```

```
##  8     13        281 North_Melbourne  VIC      17 TRUE     18      11
##  9      3         75 Geelong          VIC       5 FALSE     9      14
## 10      4        120 Brisbane_Lions   QLD       8 TRUE     10      14
## # i 190 more rows
```

Q10(a) Insert two new columns

```
# Calculate the score and accuracy and insert the new columns
afl2<-mutate(afl2,score=goals*6+behinds)
afl2<-mutate(afl2,accuracy=goals/(goals+behinds))
afl2
```

```
## # A tibble: 200 x 10
##    RowNum TidyRowNum Team         State round home  goals behinds score accuracy
##     <int>      <int> <chr>        <chr> <int> <lgl> <int>   <int> <dbl>    <dbl>
##  1     12         15 Sydney       NSW       1 TRUE     13      10    88    0.565
##  2     14        299 North_Melbo~ VIC      18 FALSE    11       8    74    0.579
##  3     16        170 Western_Bul~ VIC      10 FALSE    14       6    90    0.7
##  4      9        301 Port_Adelai~ SA       18 FALSE    11      14    80    0.44
##  5      1        172 Carlton      VIC      11 TRUE     15      11   101    0.577
##  6      6        174 Collingwood  VIC      11 TRUE     17      11   113    0.607
##  7     12        338 Sydney       NSW      20 FALSE    14      12    96    0.538
##  8     13        281 North_Melbo~ VIC      17 TRUE     18      11   119    0.621
##  9      3         75 Geelong      VIC       5 FALSE     9      14    68    0.391
## 10      4        120 Brisbane_Li~ QLD       8 TRUE     10      14    74    0.417
## # i 190 more rows
```

The score variable is Quantitative Discrete while the accuracy variable is Quantitative Continuous. The score's type is incorrect, it should be integers and the accuracy's is correct.

```
afl2$score<-as.integer(afl2$score)
afl2
```

```
## # A tibble: 200 x 10
##    RowNum TidyRowNum Team         State round home  goals behinds score accuracy
##     <int>      <int> <chr>        <chr> <int> <lgl> <int>   <int> <int>    <dbl>
##  1     12         15 Sydney       NSW       1 TRUE     13      10    88    0.565
##  2     14        299 North_Melbo~ VIC      18 FALSE    11       8    74    0.579
##  3     16        170 Western_Bul~ VIC      10 FALSE    14       6    90    0.7
##  4      9        301 Port_Adelai~ SA       18 FALSE    11      14    80    0.44
##  5      1        172 Carlton      VIC      11 TRUE     15      11   101    0.577
##  6      6        174 Collingwood  VIC      11 TRUE     17      11   113    0.607
##  7     12        338 Sydney       NSW      20 FALSE    14      12    96    0.538
##  8     13        281 North_Melbo~ VIC      17 TRUE     18      11   119    0.621
##  9      3         75 Geelong      VIC       5 FALSE     9      14    68    0.391
## 10      4        120 Brisbane_Li~ QLD       8 TRUE     10      14    74    0.417
## # i 190 more rows
```

Q10(b)

```
summarise(group_by(afl2,Team),mean_score=mean(score))
```

```
## # A tibble: 14 x 2
##    Team             mean_score
##    <chr>                 <dbl>
##  1 Brisbane_Lions         81.4
##  2 Carlton                92.6
##  3 Collingwood           107.
##  4 Essendon               90.8
##  5 Fremantle             104.
##  6 Geelong               114.
##  7 Hawthorn               98.4
##  8 North_Melbourne        82.3
##  9 Port_Adelaide          82.9
## 10 Richmond               75.3
## 11 St_Kilda               87.7
## 12 Sydney                 89.3
## 13 West_Coast             82.6
## 14 Western_Bulldogs       88.4
```

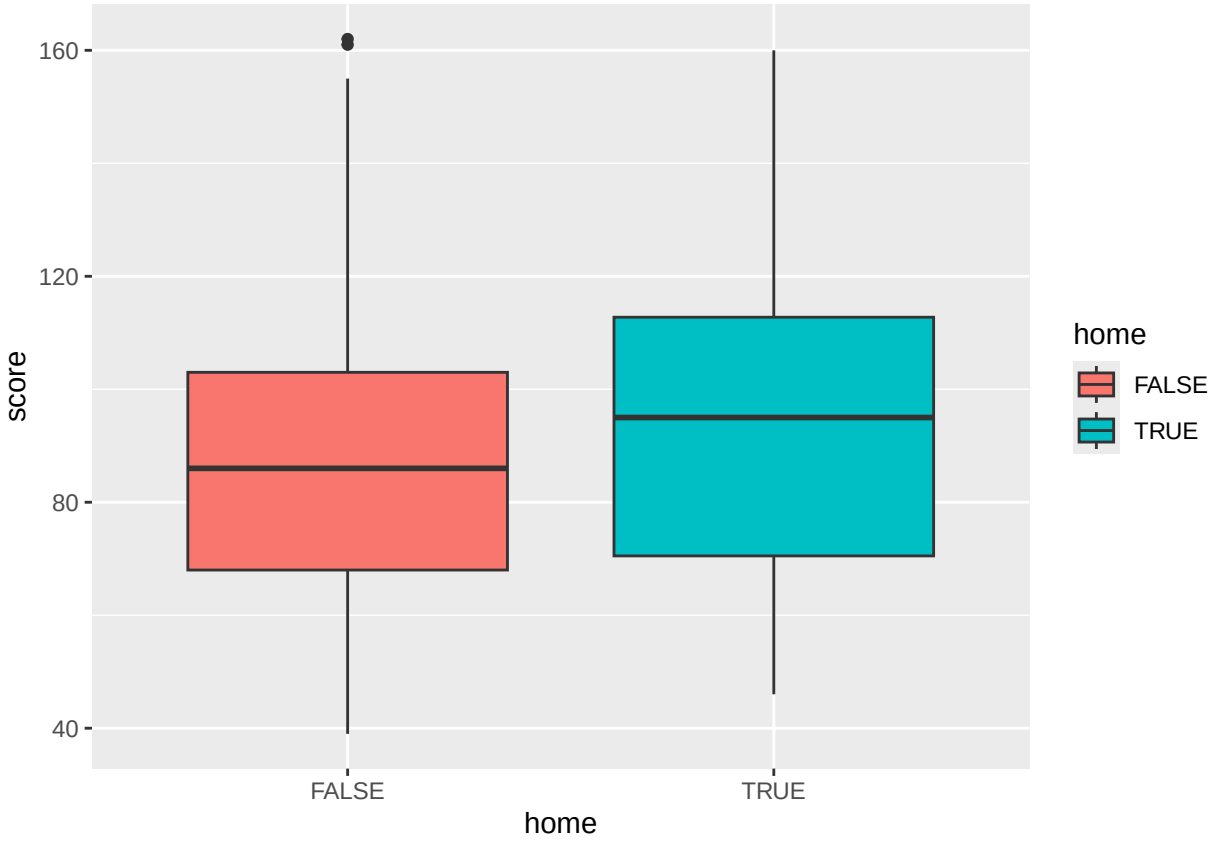
```
summarise(group_by(afl2,Team),mean_accuracy=mean(accuracy))
```

```
## # A tibble: 14 x 2
##    Team             mean_accuracy
##    <chr>                    <dbl>
##  1 Brisbane_Lions           0.487
##  2 Carlton                  0.564
##  3 Collingwood              0.477
##  4 Essendon                 0.535
##  5 Fremantle                0.567
##  6 Geelong                  0.565
##  7 Hawthorn                 0.566
##  8 North_Melbourne          0.532
##  9 Port_Adelaide            0.498
## 10 Richmond                 0.522
## 11 St_Kilda                 0.529
## 12 Sydney                   0.515
## 13 West_Coast               0.491
## 14 Western_Bulldogs         0.538
```

- i. Fremantle 104.50000
- ii. Richmond 75.33333
- iii. Fremantle 0.5674431
- iv. Collingwood 0.4771722

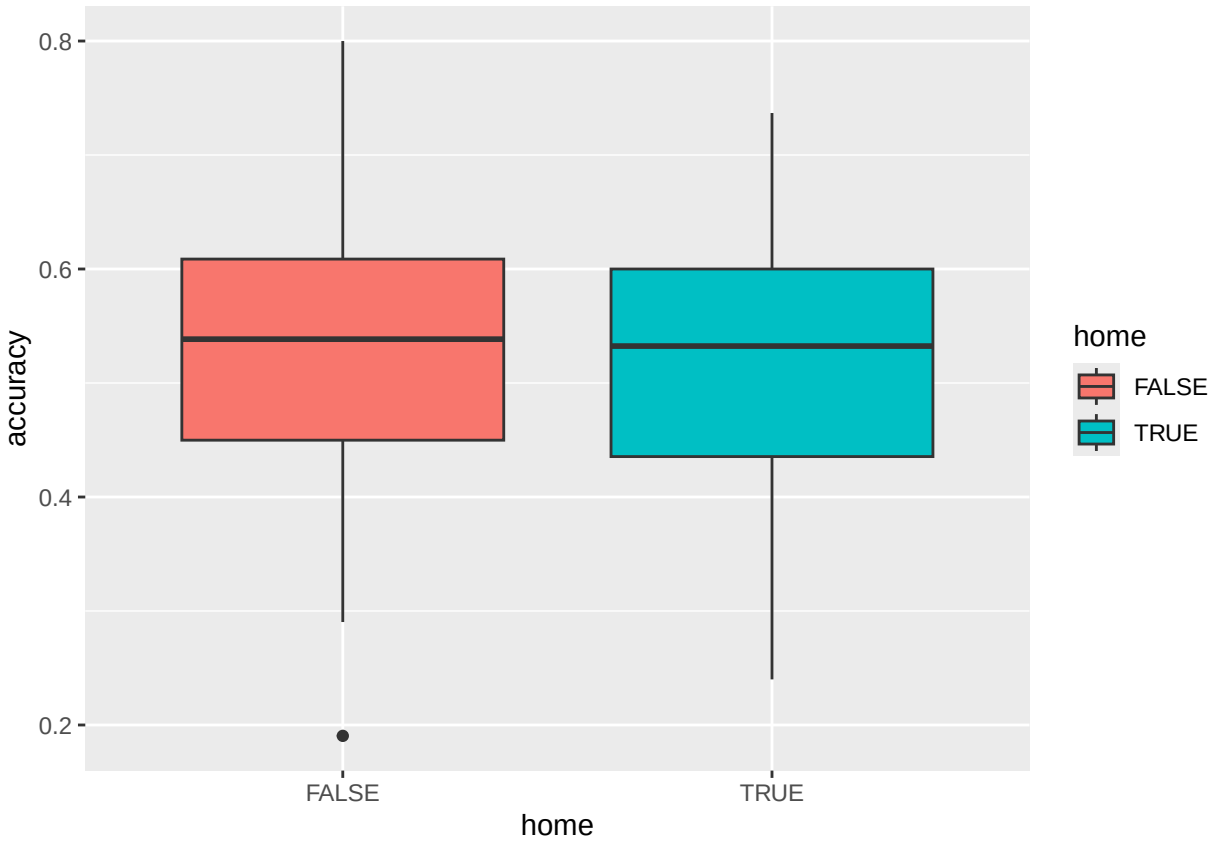
Q11(a)

```
ggplot(afl2,aes(home,score,fill=home))+
  geom_boxplot()
```



Q11(b)

```
ggplot(afl2,aes(home,accuracy,fill=home))+
  geom_boxplot()
```



The home games have better probabilities to win more scores but the accuracy between home and away is very close. With the graph we can see the average line of home score is higher and the top and bottom is higher than away's. But when it comes to the accuracy graph their posistion is much closer. # Q12

```
afl_home<-filter(afl2,home==TRUE)
afl_away<-filter(afl2,home==FALSE)
afl_home
```

```
## # A tibble: 98 x 10
##    RowNum TidyRowNum Team         State round home  goals behinds score accuracy
##     <int>      <int> <chr>        <chr> <int> <lgl> <int>   <int> <int>    <dbl>
##  1     12         15 Sydney       NSW       1 TRUE     13      10    88    0.565
##  2      1        172 Carlton      VIC      11 TRUE     15      11   101    0.577
##  3      6        174 Collingwood  VIC      11 TRUE     17      11   113    0.607
##  4     13        281 North_Melbo~ VIC      17 TRUE     18      11   119    0.621
##  5      4        120 Brisbane_Li~ QLD       8 TRUE     10      14    74    0.417
##  6      7        186 West_Coast   WA       11 TRUE     14      14    98    0.5
##  7     17        345 Essendon     VIC      21 TRUE     10       8    68    0.556
##  8     15        212 Hawthorn     VIC      13 TRUE     14      18   102    0.438
##  9     12        253 Sydney       NSW      15 TRUE     12      13    85    0.48
## 10      4        290 Brisbane_Li~ QLD      18 TRUE      9      10    64    0.474
## # i 88 more rows
```

```
afl_away
```

```
## # A tibble: 102 x 10
##    RowNum TidyRowNum Team         State round home  goals behinds score accuracy
```

```
##     <int>      <int> <chr>        <chr> <int> <lgl> <int>   <int> <int>    <dbl>
##  1     14        299 North_Melbo~ VIC      18 FALSE    11       8    74    0.579
##  2     16        170 Western_Bul~ VIC      10 FALSE    14       6    90    0.7
##  3      9        301 Port_Adelai~ SA       18 FALSE    11      14    80    0.44
##  4     12        338 Sydney       NSW      20 FALSE    14      12    96    0.538
##  5      3         75 Geelong      VIC       5 FALSE     9      14    68    0.391
##  6     13        230 North_Melbo~ VIC      14 FALSE     9       9    63    0.5
##  7      4         86 Brisbane_Li~ QLD       6 FALSE    13       9    87    0.591
##  8      6        310 Collingwood  VIC      19 FALSE    14      23   107    0.378
##  9      2         62 Port_Adelai~ SA        4 FALSE    10       4    64    0.714
## 10     13        366 North_Melbo~ VIC      22 FALSE    17      11   113    0.607
## # i 92 more rows
```

Q13

```
inspect_num(afl_home)
```

```
## # A tibble: 7 x 10
##   col_name     min     q1  median    mean    q3     max      sd pcnt_na hist
##   <chr>      <dbl>  <dbl>   <dbl>   <dbl> <dbl>   <dbl>   <dbl>   <dbl> <named >
## 1 RowNum       1     6      10       9.58   13    17       4.77        0 <tibble>
## 2 TidyRowNum   1    89.5   172.    179.    271   370     109.          0 <tibble>
## 3 round        1     6      11      11.0    16    22       6.38        0 <tibble>
## 4 goals        6    10      14      13.6    16    24       4.44        0 <tibble>
## 5 behinds      4    10      12      12.4    15    23       3.84        0 <tibble>
## 6 score       46    70.5    95      94.1   113.  160      27.6         0 <tibble>
## 7 accuracy     0.24  0.435   0.532   0.522   0.6   0.737   0.105       0 <tibble>
```

```
inspect_num(afl_away)
```

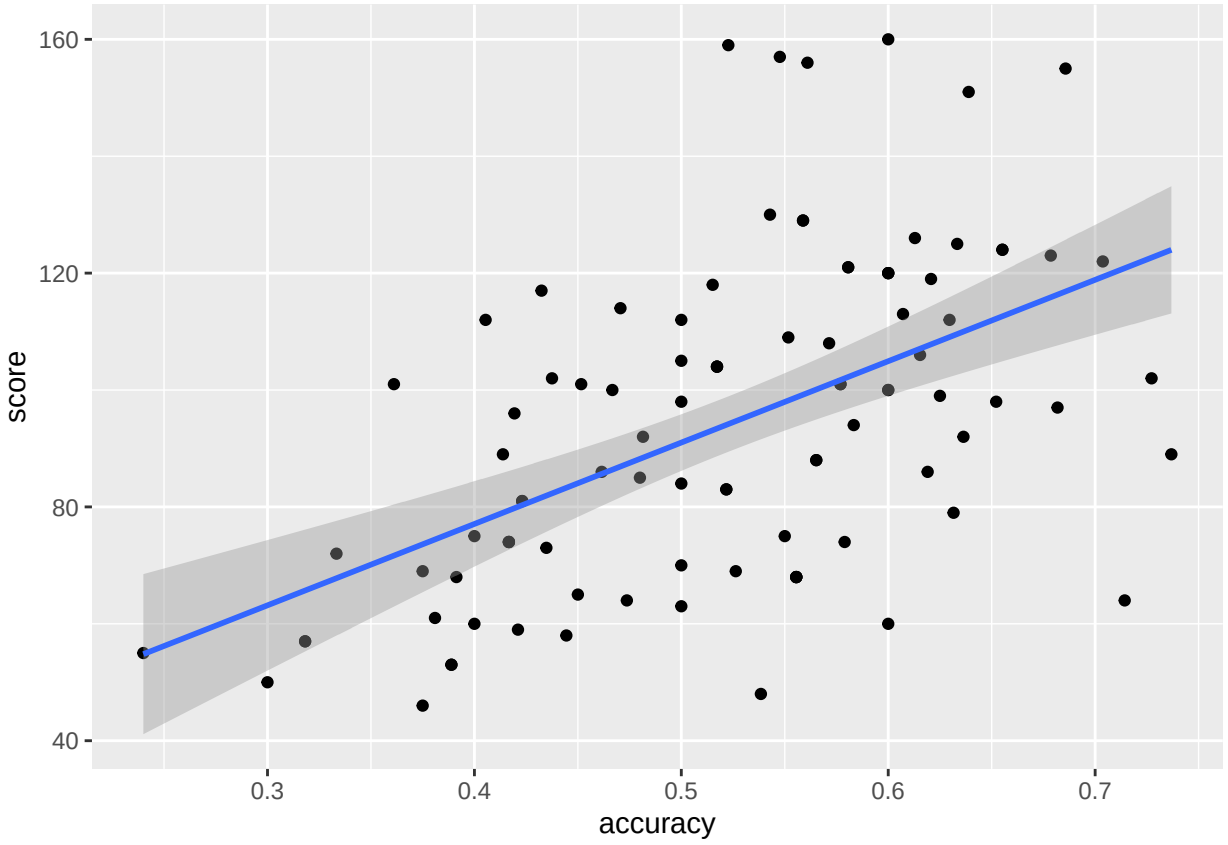
```
## # A tibble: 7 x 10
##   col_name     min     q1  median    mean      q3   max      sd pcnt_na hist
##   <chr>      <dbl>  <dbl>   <dbl>   <dbl>   <dbl> <dbl>   <dbl>   <dbl> <named >
## 1 RowNum     1      4.25    9.5     8.98   13      17     5.13        0 <tibble>
## 2 TidyRowN~  4     83     176.    181.    284.    373   112.          0 <tibble>
## 3 round      1      5.25   11      11.2    17      22     6.59        0 <tibble>
## 4 goals      4     10      12      12.6    15      25     4.23        0 <tibble>
## 5 behinds    3      8      10      11.1    14      23     4.39        0 <tibble>
## 6 score     39     68      86      87.0   103     162    26.7         0 <tibble>
## 7 accuracy   0.190  0.450   0.538   0.537   0.609   0.8   0.114       0 <tibble>
```

The average score of home games is 94.0510204 while the average accuracy is 0.5218593. Also the average score of away games 86.9803922 while the average accuracy is 86.9803922.

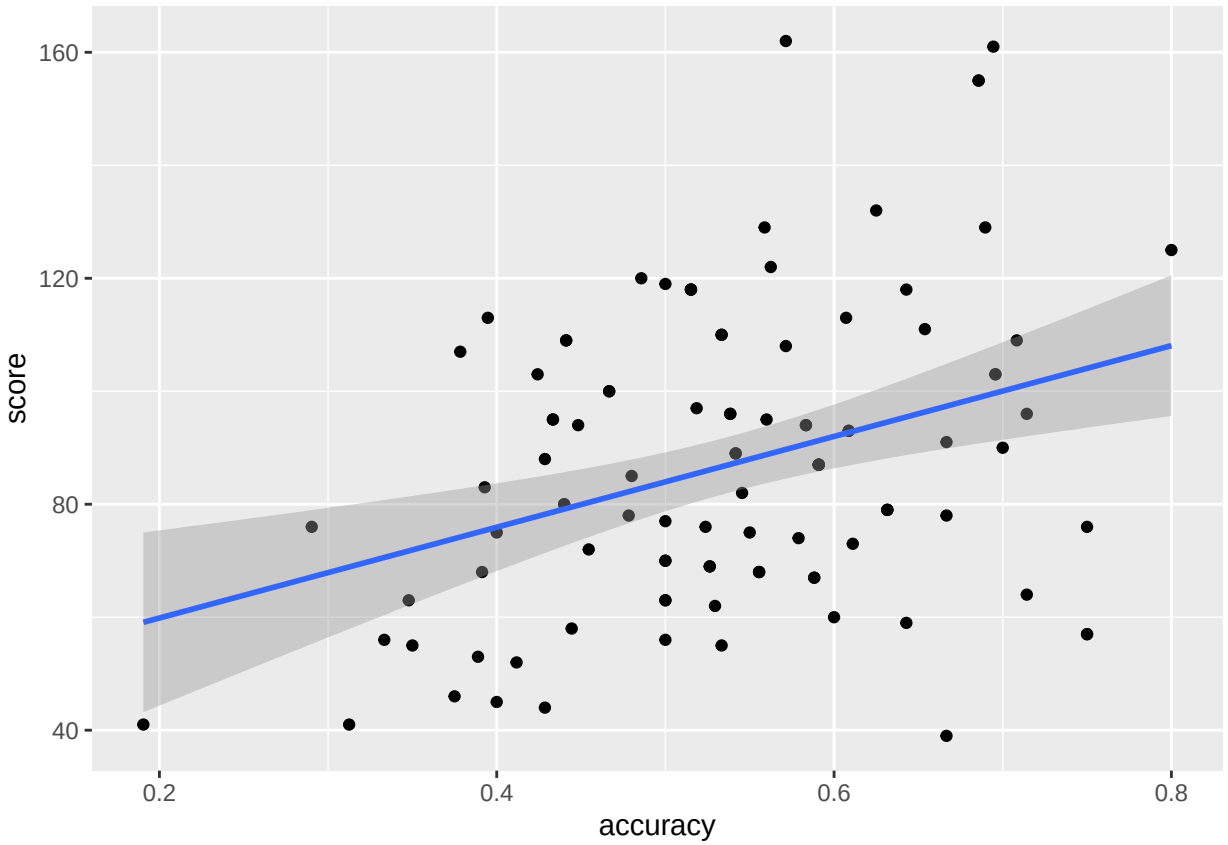
The data does support the claim.

Q14

```
ggplot(afl_home, aes(x = accuracy, y = score)) +
  geom_point() +
  geom_smooth(method="lm")
```




```
ggplot(afl_away, aes(x = accuracy, y = score)) +
  geom_point() +
  geom_smooth(method="lm")
```



The calculation of score is to multiple goals numbers with 6 and behinds just 1 time and the accuracy represent the proportion of goals, which infers that with higher accuracy come to higher goals. And the higher goals change into higher scores. So I choose the accuracy to be the independent variable and the score to be predictor.

Q15

As the scatter plots shown, the higher accuracy come to higher scores. And it is similar for both home and away teams.