DATA TAMING ASSIGNMENT 1

VINEETH MARIKUNTEMATHA RAVISHARADHYA-a1896845

23/02/2024

Setup

```
#Load the required packages
library(here)
library(tidyverse)
library(tidyr)
library(stringr)
library(forcats)
library(lubridate)
library(inspectdf)
library(ggplot2)
```

Q1. Loading the data

```
# Your student number goes here
ysn = 1896845
# Calculate your student number modulo 3
filenum <- ysn %% 3
filenum
```

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

```
filename <- paste0("D:/r studio files/Assignment_1/archery_2.csv")
filename
```

```
## [1] "D:/r studio files/Assignment_1/archery_2.csv"
```

```
# Read in the data
archery_2 <- read_csv("archery_2.csv")
# Display the first 10 lines of the data

archery_2
```

```
## # A tibble: 200 x 9
##    name    experienced started session_1 session_2 session_3 session_4 session_5
```

```
##    <chr>   <chr>       <chr>   <chr>     <chr>     <chr>     <chr>     <chr>
##  1 Eli     No          01-Feb~ Target h~ Target h~ Target h~ Target h~ Target h~
##  2 Aria    No          01-Feb~ Target h~ Target h~ Target h~ Target h~ Target h~
##  3 Isabel~ Yes         15-Dec~ Target h~ Target h~ Target h~ Target h~ Target h~
##  4 Sofia   No          01-Feb~ Target h~ Target h~ Target h~ Target h~ Target h~
##  5 Macken~ Yes         17-Apr~ Target h~ Target h~ Target h~ Target h~ Target h~
##  6 Audrey  Yes         14-Apr~ Target h~ Target h~ Target h~ Target h~ Target h~
##  7 Abigail No          01-Feb~ Target h~ Target h~ Target h~ Target h~ Target h~
##  8 Zoe     Yes         15-Nov~ Target h~ Target h~ Target h~ Target h~ Target h~
##  9 Claire  No          01-Feb~ Target h~ Target h~ Target h~ Target h~ Target h~
## 10 Floren~ No          01-Feb~ Target h~ Target h~ Target h~ Target h~ Target h~
## # i 190 more rows
## # i 1 more variable: session_6 <chr>
```

```
print(n=10,archery_2)
```

```
## # A tibble: 200 x 9
##    name    experienced started session_1 session_2 session_3 session_4 session_5
##    <chr>   <chr>       <chr>   <chr>     <chr>     <chr>     <chr>     <chr>
##  1 Eli     No          01-Feb~ Target h~ Target h~ Target h~ Target h~ Target h~
##  2 Aria    No          01-Feb~ Target h~ Target h~ Target h~ Target h~ Target h~
##  3 Isabel~ Yes         15-Dec~ Target h~ Target h~ Target h~ Target h~ Target h~
##  4 Sofia   No          01-Feb~ Target h~ Target h~ Target h~ Target h~ Target h~
##  5 Macken~ Yes         17-Apr~ Target h~ Target h~ Target h~ Target h~ Target h~
##  6 Audrey  Yes         14-Apr~ Target h~ Target h~ Target h~ Target h~ Target h~
##  7 Abigail No          01-Feb~ Target h~ Target h~ Target h~ Target h~ Target h~
##  8 Zoe     Yes         15-Nov~ Target h~ Target h~ Target h~ Target h~ Target h~
##  9 Claire  No          01-Feb~ Target h~ Target h~ Target h~ Target h~ Target h~
## 10 Floren~ No          01-Feb~ Target h~ Target h~ Target h~ Target h~ Target h~
## # i 190 more rows
## # i 1 more variable: session_6 <chr>
```

Q2. Taking a random sample of archers

```
set.seed(1896845)
sample_archery_2 <- sample_n(archery_2, 100)

sample_archery_2
```

```
## # A tibble: 100 x 9
##    name    experienced started session_1 session_2 session_3 session_4 session_5
##    <chr>   <chr>       <chr>   <chr>     <chr>     <chr>     <chr>     <chr>
##  1 River   No          01-Feb~ Target h~ Target h~ Target h~ Target h~ Target h~
##  2 Benjam~ No          01-Feb~ Target h~ Target h~ Target h~ Target h~ Target h~
##  3 Chloe   No          01-Feb~ Target h~ Target h~ Target h~ Target h~ Target h~
##  4 Lincoln Yes         15-Jul~ Target h~ Target h~ Target h~ Target h~ Target h~
##  5 Charlo~ Yes         07-Dec~ Target h~ Target h~ Target h~ Target h~ Target h~
##  6 Scarle~ No          01-Feb~ Target h~ Target h~ Target h~ Target h~ Target h~
##  7 Poppy   No          01-Feb~ Target h~ Target h~ Target h~ Target h~ Target h~
##  8 Gabrie~ No          01-Feb~ Target h~ Target h~ Target h~ Target h~ Target h~
```

```
##  9 Zayn    No          01-Feb~ Target h~ Target h~ Target h~ Target h~ Target h~
## 10 Logan   Yes         18-Sep~ Target h~ Target h~ Target h~ Target h~ Target h~
## # i 90 more rows
## # i 1 more variable: session_6 <chr>
```

Q3. Tidying the data

Q3(a). Converting from wide form to long form

```
library(tidyr)

archery_wide_long <- gather(sample_archery_2, key = "session", value = "result", session_1:session_6)
sample_archery_2
```

```
## # A tibble: 100 x 9
##    name    experienced started session_1 session_2 session_3 session_4 session_5
##    <chr>   <chr>       <chr>   <chr>     <chr>     <chr>     <chr>     <chr>
##  1 River   No          01-Feb~ Target h~ Target h~ Target h~ Target h~ Target h~
##  2 Benjam~ No          01-Feb~ Target h~ Target h~ Target h~ Target h~ Target h~
##  3 Chloe   No          01-Feb~ Target h~ Target h~ Target h~ Target h~ Target h~
##  4 Lincoln Yes         15-Jul~ Target h~ Target h~ Target h~ Target h~ Target h~
##  5 Charlo~ Yes         07-Dec~ Target h~ Target h~ Target h~ Target h~ Target h~
##  6 Scarle~ No          01-Feb~ Target h~ Target h~ Target h~ Target h~ Target h~
##  7 Poppy   No          01-Feb~ Target h~ Target h~ Target h~ Target h~ Target h~
##  8 Gabrie~ No          01-Feb~ Target h~ Target h~ Target h~ Target h~ Target h~
##  9 Zayn    No          01-Feb~ Target h~ Target h~ Target h~ Target h~ Target h~
## 10 Logan   Yes         18-Sep~ Target h~ Target h~ Target h~ Target h~ Target h~
## # i 90 more rows
## # i 1 more variable: session_6 <chr>
```

Q3(b). Replacing `result` with 2 new columns

```
sample <- str_match(archery_wide_long$result,"Target hit (\\d+) times from (\\d+) shots")
archery_wide_long <- mutate(archery_wide_long,result = NULL)
archery_wide_long <-mutate(archery_wide_long,
shots = sample[,3],
hits = sample[,2]
)
archery_wide_long
```

```
## # A tibble: 600 x 6
##    name      experienced started     session   shots hits
##    <chr>     <chr>       <chr>       <chr>     <chr> <chr>
##  1 River     No          01-Feb-2024 session_1 37    19
##  2 Benjamin  No          01-Feb-2024 session_1 54    23
##  3 Chloe     No          01-Feb-2024 session_1 41    24
##  4 Lincoln   Yes         15-Jul-2023 session_1 62    53
##  5 Charlotte Yes         07-Dec-2022 session_1 66    59
```

```
##  6 Scarlett  No          01-Feb-2024 session_1 55    28
##  7 Poppy     No          01-Feb-2024 session_1 44    23
##  8 Gabriella No          01-Feb-2024 session_1 57    19
##  9 Zayn      No          01-Feb-2024 session_1 38    19
## 10 Logan     Yes         18-Sep-2022 session_1 61    53
## # i 590 more rows
```

```
archery_wide_long
```

```
## # A tibble: 600 x 6
##    name      experienced started     session   shots hits
##    <chr>     <chr>       <chr>       <chr>     <chr> <chr>
##  1 River     No          01-Feb-2024 session_1 37    19
##  2 Benjamin  No          01-Feb-2024 session_1 54    23
##  3 Chloe     No          01-Feb-2024 session_1 41    24
##  4 Lincoln   Yes         15-Jul-2023 session_1 62    53
##  5 Charlotte Yes         07-Dec-2022 session_1 66    59
##  6 Scarlett  No          01-Feb-2024 session_1 55    28
##  7 Poppy     No          01-Feb-2024 session_1 44    23
##  8 Gabriella No          01-Feb-2024 session_1 57    19
##  9 Zayn      No          01-Feb-2024 session_1 38    19
## 10 Logan     Yes         18-Sep-2022 session_1 61    53
## # i 590 more rows
```

Q3(c). Replace data in `session` column with session number

```
trial <- str_match(archery_wide_long$session, "session_(\\d+)")


archery_wide_long <- mutate(archery_wide_long,
                            session = as.integer(trial[, 2]))


archery_wide_long
```

```
## # A tibble: 600 x 6
##    name      experienced started     session shots hits
##    <chr>     <chr>       <chr>         <int> <chr> <chr>
##  1 River     No          01-Feb-2024       1 37    19
##  2 Benjamin  No          01-Feb-2024       1 54    23
##  3 Chloe     No          01-Feb-2024       1 41    24
##  4 Lincoln   Yes         15-Jul-2023       1 62    53
##  5 Charlotte Yes         07-Dec-2022       1 66    59
##  6 Scarlett  No          01-Feb-2024       1 55    28
##  7 Poppy     No          01-Feb-2024       1 44    23
##  8 Gabriella No          01-Feb-2024       1 57    19
##  9 Zayn      No          01-Feb-2024       1 38    19
## 10 Logan     Yes         18-Sep-2022       1 61    53
## # i 590 more rows
```

Q3(d). Replace "Yes/No" with "TRUE/FALSE"

```
archery_wide_long$experienced[archery_wide_long$experienced == "Yes"] <- TRUE
archery_wide_long$experienced[archery_wide_long$experienced == "No"] <- FALSE

archery_wide_long
```

```
## # A tibble: 600 x 6
##    name      experienced started     session shots hits
##    <chr>     <chr>       <chr>         <int> <chr> <chr>
##  1 River     FALSE       01-Feb-2024       1 37    19
##  2 Benjamin  FALSE       01-Feb-2024       1 54    23
##  3 Chloe     FALSE       01-Feb-2024       1 41    24
##  4 Lincoln   TRUE        15-Jul-2023       1 62    53
##  5 Charlotte TRUE        07-Dec-2022       1 66    59
##  6 Scarlett  FALSE       01-Feb-2024       1 55    28
##  7 Poppy     FALSE       01-Feb-2024       1 44    23
##  8 Gabriella FALSE       01-Feb-2024       1 57    19
##  9 Zayn      FALSE       01-Feb-2024       1 38    19
## 10 Logan     TRUE        18-Sep-2022       1 61    53
## # i 590 more rows
```

```
#print first 10 outputs
print(n=10,archery_wide_long)
```

```
## # A tibble: 600 x 6
##    name      experienced started     session shots hits
##    <chr>     <chr>       <chr>         <int> <chr> <chr>
##  1 River     FALSE       01-Feb-2024       1 37    19
##  2 Benjamin  FALSE       01-Feb-2024       1 54    23
##  3 Chloe     FALSE       01-Feb-2024       1 41    24
##  4 Lincoln   TRUE        15-Jul-2023       1 62    53
##  5 Charlotte TRUE        07-Dec-2022       1 66    59
##  6 Scarlett  FALSE       01-Feb-2024       1 55    28
##  7 Poppy     FALSE       01-Feb-2024       1 44    23
##  8 Gabriella FALSE       01-Feb-2024       1 57    19
##  9 Zayn      FALSE       01-Feb-2024       1 38    19
## 10 Logan     TRUE        18-Sep-2022       1 61    53
## # i 590 more rows
```

Q4. Identifying data types

name = It is a "Categorical Nominal",as they are not categorized on the basis of hierarchy , making it as a nominal variable.

"experienced" = "Categorical Nominal" , as it appears to be the status of the archers ,with true and false and also its a nominal variable as it has no inherent order from all the categories.

"started" = it is a"Quantitative Continuous" as it represents continuous range of values and also the this variable holds the date of each athlete or archer started

"session" = it is a "Quantitative Discrete" as this variable is countable and distinct value as it doesnt have fractions or decimals

"shots" = here this variable has the number of shots taken by every archer and which tends to have countable and distinct values, so it is "Quantitative Discrete" .

"hits" = it is similar to shots as it is countable and has distinct value so it is "Quantitative Discrete"

"days_experience" = it is a "Quantitative Continuous" as this variable can take any non variable non negative real value within a continuous range.

"accuracy" = it is a "Quantitative Continuous, as it represents the ratio of hits to shots along with the accuracy of every archer.

Q5(a). Taming the data

```
archery_wide_long$hits <- as.integer(archery_wide_long$hits)
archery_wide_long$shots <- as.integer(archery_wide_long$shots)


archery_wide_long
```

```
## # A tibble: 600 x 6
##    name      experienced started     session shots  hits
##    <chr>     <chr>       <chr>         <int> <int> <int>
##  1 River     FALSE       01-Feb-2024       1    37    19
##  2 Benjamin  FALSE       01-Feb-2024       1    54    23
##  3 Chloe     FALSE       01-Feb-2024       1    41    24
##  4 Lincoln   TRUE        15-Jul-2023       1    62    53
##  5 Charlotte TRUE        07-Dec-2022       1    66    59
##  6 Scarlett  FALSE       01-Feb-2024       1    55    28
##  7 Poppy     FALSE       01-Feb-2024       1    44    23
##  8 Gabriella FALSE       01-Feb-2024       1    57    19
##  9 Zayn      FALSE       01-Feb-2024       1    38    19
## 10 Logan     TRUE        18-Sep-2022       1    61    53
## # i 590 more rows
```

Q5(b). Taming the data

‘

```
archery_wide_long$started = dmy(archery_wide_long$started)
archery_wide_long$experienced = as.logical(archery_wide_long$experienced)
archery_wide_long
```

```
## # A tibble: 600 x 6
##    name      experienced started    session shots  hits
##    <chr>     <lgl>       <date>       <int> <int> <int>
##  1 River     FALSE       2024-02-01       1    37    19
##  2 Benjamin  FALSE       2024-02-01       1    54    23
##  3 Chloe     FALSE       2024-02-01       1    41    24
```

```
##  4 Lincoln   TRUE        2023-07-15       1    62    53
##  5 Charlotte TRUE        2022-12-07       1    66    59
##  6 Scarlett  FALSE       2024-02-01       1    55    28
##  7 Poppy     FALSE       2024-02-01       1    44    23
##  8 Gabriella FALSE       2024-02-01       1    57    19
##  9 Zayn      FALSE       2024-02-01       1    38    19
## 10 Logan     TRUE        2022-09-18       1    61    53
## # i 590 more rows
```

```
#print first 10 outputs
print(n=10,archery_wide_long)
```

```
## # A tibble: 600 x 6
##    name      experienced started    session shots  hits
##    <chr>     <lgl>       <date>       <int> <int> <int>
##  1 River     FALSE       2024-02-01       1    37    19
##  2 Benjamin  FALSE       2024-02-01       1    54    23
##  3 Chloe     FALSE       2024-02-01       1    41    24
##  4 Lincoln   TRUE        2023-07-15       1    62    53
##  5 Charlotte TRUE        2022-12-07       1    66    59
##  6 Scarlett  FALSE       2024-02-01       1    55    28
##  7 Poppy     FALSE       2024-02-01       1    44    23
##  8 Gabriella FALSE       2024-02-01       1    57    19
##  9 Zayn      FALSE       2024-02-01       1    38    19
## 10 Logan     TRUE        2022-09-18       1    61    53
## # i 590 more rows
```

Q6. Adding two new colums to dataset

adding days_experience column

```
library(lubridate)
library(dplyr)

archery_wide_long <- mutate(archery_wide_long,
                            days_experience = as.integer(difftime(as.Date("2024-02-01"), started, units


archery_wide_long
```

```
## # A tibble: 600 x 7
##    name      experienced started    session shots  hits days_experience
##    <chr>     <lgl>       <date>       <int> <int> <int>           <int>
##  1 River     FALSE       2024-02-01       1    37    19               0
##  2 Benjamin  FALSE       2024-02-01       1    54    23               0
##  3 Chloe     FALSE       2024-02-01       1    41    24               0
##  4 Lincoln   TRUE        2023-07-15       1    62    53             201
##  5 Charlotte TRUE        2022-12-07       1    66    59             421
##  6 Scarlett  FALSE       2024-02-01       1    55    28               0
```

```
##  7 Poppy     FALSE       2024-02-01       1    44    23               0
##  8 Gabriella FALSE       2024-02-01       1    57    19               0
##  9 Zayn      FALSE       2024-02-01       1    38    19               0
## 10 Logan     TRUE        2022-09-18       1    61    53             501
## # i 590 more rows
```

adding accuracy column

```
archery_wide_long <- mutate(archery_wide_long,
                            accuracy = hits / shots)
archery_wide_long$accuracy <- archery_wide_long$hits / archery_wide_long$shots

archery_wide_long
```

```
## # A tibble: 600 x 8
##    name      experienced started    session shots  hits days_experience accuracy
##    <chr>     <lgl>       <date>       <int> <int> <int>           <int>    <dbl>
##  1 River     FALSE       2024-02-01       1    37    19               0    0.514
##  2 Benjamin  FALSE       2024-02-01       1    54    23               0    0.426
##  3 Chloe     FALSE       2024-02-01       1    41    24               0    0.585
##  4 Lincoln   TRUE        2023-07-15       1    62    53             201    0.855
##  5 Charlotte TRUE        2022-12-07       1    66    59             421    0.894
##  6 Scarlett  FALSE       2024-02-01       1    55    28               0    0.509
##  7 Poppy     FALSE       2024-02-01       1    44    23               0    0.523
##  8 Gabriella FALSE       2024-02-01       1    57    19               0    0.333
##  9 Zayn      FALSE       2024-02-01       1    38    19               0    0.5
## 10 Logan     TRUE        2022-09-18       1    61    53             501    0.869
## # i 590 more rows
```

```
#print first 10 outputs
print(n=10,archery_wide_long)
```

```
## # A tibble: 600 x 8
##    name      experienced started    session shots  hits days_experience accuracy
##    <chr>     <lgl>       <date>       <int> <int> <int>           <int>    <dbl>
##  1 River     FALSE       2024-02-01       1    37    19               0    0.514
##  2 Benjamin  FALSE       2024-02-01       1    54    23               0    0.426
##  3 Chloe     FALSE       2024-02-01       1    41    24               0    0.585
##  4 Lincoln   TRUE        2023-07-15       1    62    53             201    0.855
##  5 Charlotte TRUE        2022-12-07       1    66    59             421    0.894
##  6 Scarlett  FALSE       2024-02-01       1    55    28               0    0.509
##  7 Poppy     FALSE       2024-02-01       1    44    23               0    0.523
##  8 Gabriella FALSE       2024-02-01       1    57    19               0    0.333
##  9 Zayn      FALSE       2024-02-01       1    38    19               0    0.5
## 10 Logan     TRUE        2022-09-18       1    61    53             501    0.869
## # i 590 more rows
```

Q7. Display the sample statistics for the numerical values in your dataset

```
inspect_num(archery_wide_long)
```

```
## # A tibble: 5 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 session    1      2       3.5     3.5     5         6   1.71        0 <tibble>
## 2 shots     22     43      51      52.6    62        96  12.2         0 <tibble>
## 3 hits      10     24      39.5    39.6    54        83  16.6         0 <tibble>
## 4 days_exp~  0      0     206.    269.    497.      910 297.          0 <tibble>
## 5 accuracy   0.333  0.560   0.762   0.726   0.898     1   0.181       0 <tibble>
```

#q8. Subset of dataset to contain experienced ones

```
archery_wide_long <- filter(archery_wide_long, experienced == TRUE)
archery_wide_long
```

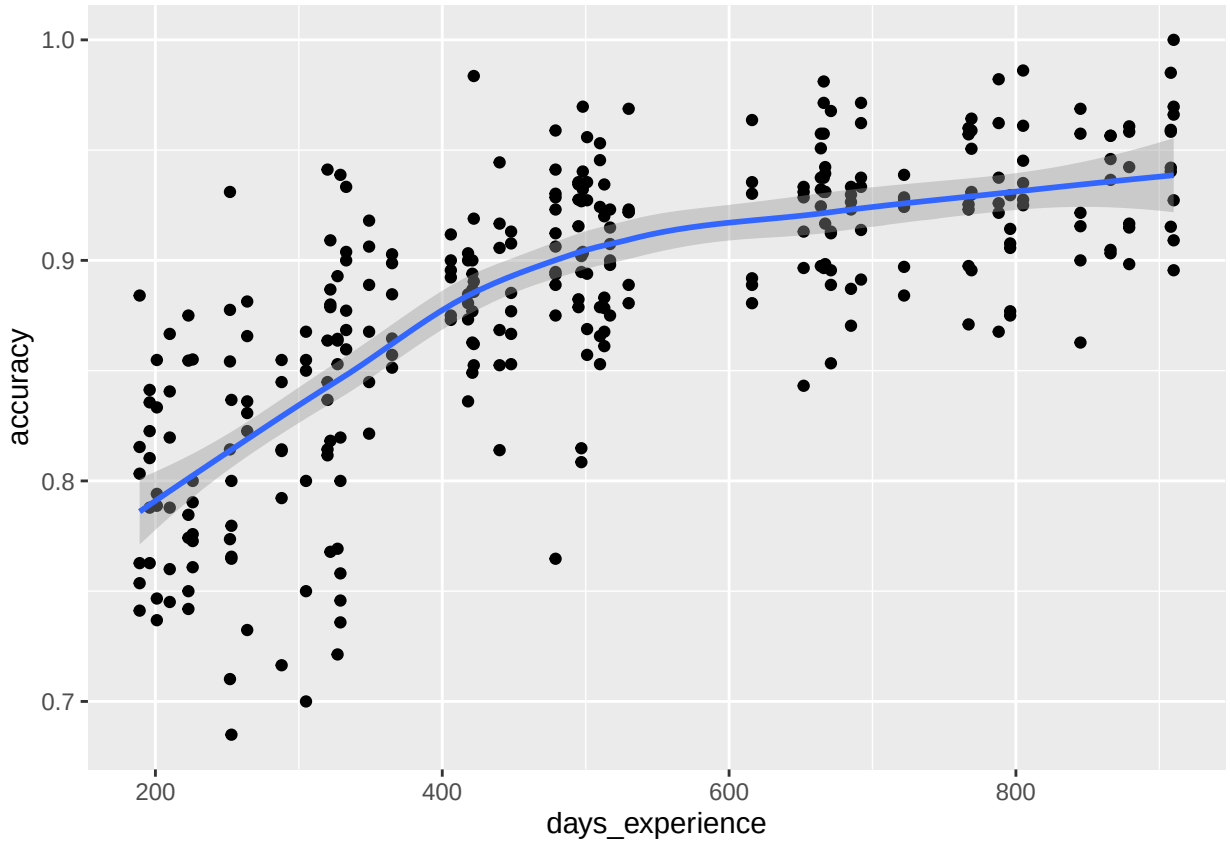
```
## # A tibble: 318 x 8
##    name      experienced started    session shots  hits days_experience accuracy
##    <chr>     <lgl>       <date>       <int> <int> <int>           <int>    <dbl>
##  1 Lincoln   TRUE        2023-07-15       1    62    53             201    0.855
##  2 Charlotte TRUE        2022-12-07       1    66    59             421    0.894
##  3 Logan     TRUE        2022-09-18       1    61    53             501    0.869
##  4 Sophia    TRUE        2022-04-20       1    51    43             652    0.843
##  5 Bella     TRUE        2021-10-09       1    64    62             845    0.969
##  6 Layla     TRUE        2022-09-09       1    66    58             510    0.879
##  7 Charlie   TRUE        2023-04-19       1    67    48             288    0.716
##  8 Millie    TRUE        2023-02-17       1    84    69             349    0.821
##  9 Zara      TRUE        2022-05-26       1    43    40             616    0.930
## 10 Emily     TRUE        2022-09-02       1    54    49             517    0.907
## # i 308 more rows
```

```
#print first 10 outputs
print(n=10,archery_wide_long)
```

```
## # A tibble: 318 x 8
##    name      experienced started    session shots  hits days_experience accuracy
##    <chr>     <lgl>       <date>       <int> <int> <int>           <int>    <dbl>
##  1 Lincoln   TRUE        2023-07-15       1    62    53             201    0.855
##  2 Charlotte TRUE        2022-12-07       1    66    59             421    0.894
##  3 Logan     TRUE        2022-09-18       1    61    53             501    0.869
##  4 Sophia    TRUE        2022-04-20       1    51    43             652    0.843
##  5 Bella     TRUE        2021-10-09       1    64    62             845    0.969
##  6 Layla     TRUE        2022-09-09       1    66    58             510    0.879
##  7 Charlie   TRUE        2023-04-19       1    67    48             288    0.716
##  8 Millie    TRUE        2023-02-17       1    84    69             349    0.821
##  9 Zara      TRUE        2022-05-26       1    43    40             616    0.930
## 10 Emily     TRUE        2022-09-02       1    54    49             517    0.907
## # i 308 more rows
```

#Q9. Scatter plot of accuracy and days_experience

```
ggplot(archery_wide_long, aes(x = days_experience, y = accuracy)) +
  geom_point() +
  geom_smooth()
```



q10. Since there is a rise in the days_experience , its a sign of positive trend, where it implies that the archers are getting more experience and their accuracy level tends to improve. However at the curve over 750 days becaomes more stable, so to be the better archer it takes anywhere around 750 to 800 days by taking plot as the reference.