EDA olympics

Usman Siddiqui

11/29/2020

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## Warning: package 'ggplot2' was built under R version 3.6.2

## Warning: package 'tibble' was built under R version 3.6.2

## Warning: package 'tidyr' was built under R version 3.6.2

## Warning: package 'purrr' was built under R version 3.6.2

## Warning: package 'dplyr' was built under R version 3.6.2

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(stringr)
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 3.6.2

##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(ggrepel)
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 3.6.2
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(cluster)
library(ggforce)
```

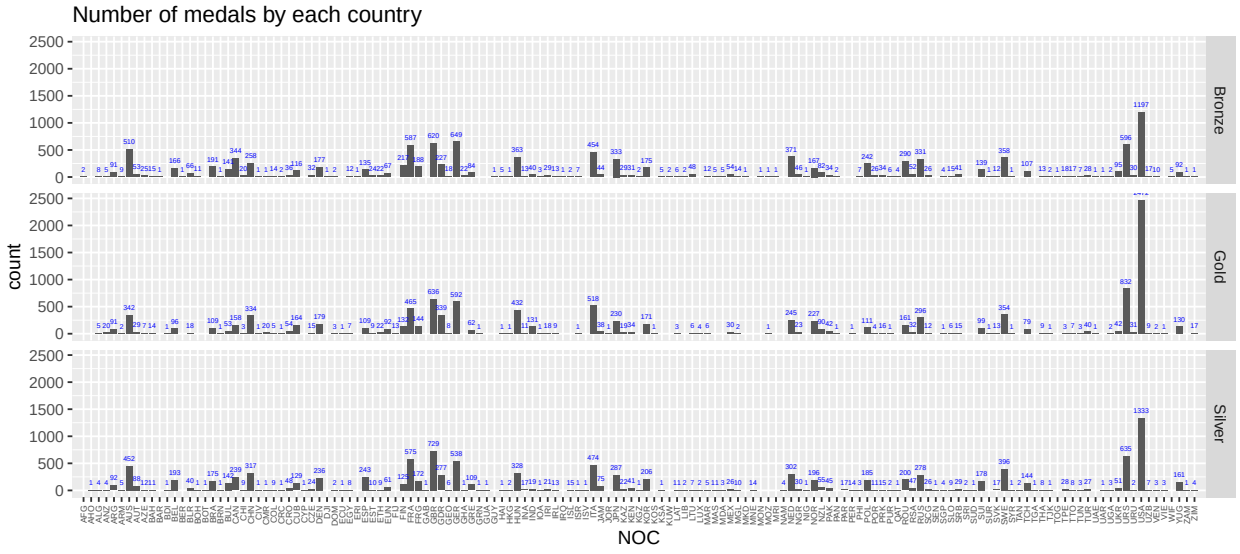
```
## Warning: package 'ggforce' was built under R version 3.6.2
```

```
athletes <- read.csv("../data/athlete_events.csv")
noc <- read.csv("../data/noc_regions.csv")
vertical_x <- theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1,size = 5))
```

Exploring data

Find the medal tally of each country

```
medals <- athletes %>%
  filter(!(is.na(Medal))) %>%
  filter(Season == "Summer")

ggplot(data = medals , aes(x = NOC))+
  geom_bar() + vertical_x + labs(title = "Number of medals by each country" , c) +
  scale_x_discrete(expand = c(0.01,0.01))+
  geom_text(stat='count', aes(label=..count..), vjust=-1 , size = 1.5 , color = 'blue') +
  facet_grid(rows  = vars(Medal))
```



Find the country with the highest medal tally and see how they fare over the years

```
best_perf <- medals %>% summarise(tally = n())
top_country <- best_perf %>% filter(tally == 5002)
top_country
```

```
## [1] tally
## <0 rows> (or 0-length row.names)
```

This concludes that USA has been the best country in the summer olympics over the years

USA EDA

```
usa <- medals %>% filter(NOC == "USA")

total_diff_medals <- usa %>% group_by(Year , Medal) %>%
  summarise("medals per year" = n())
```

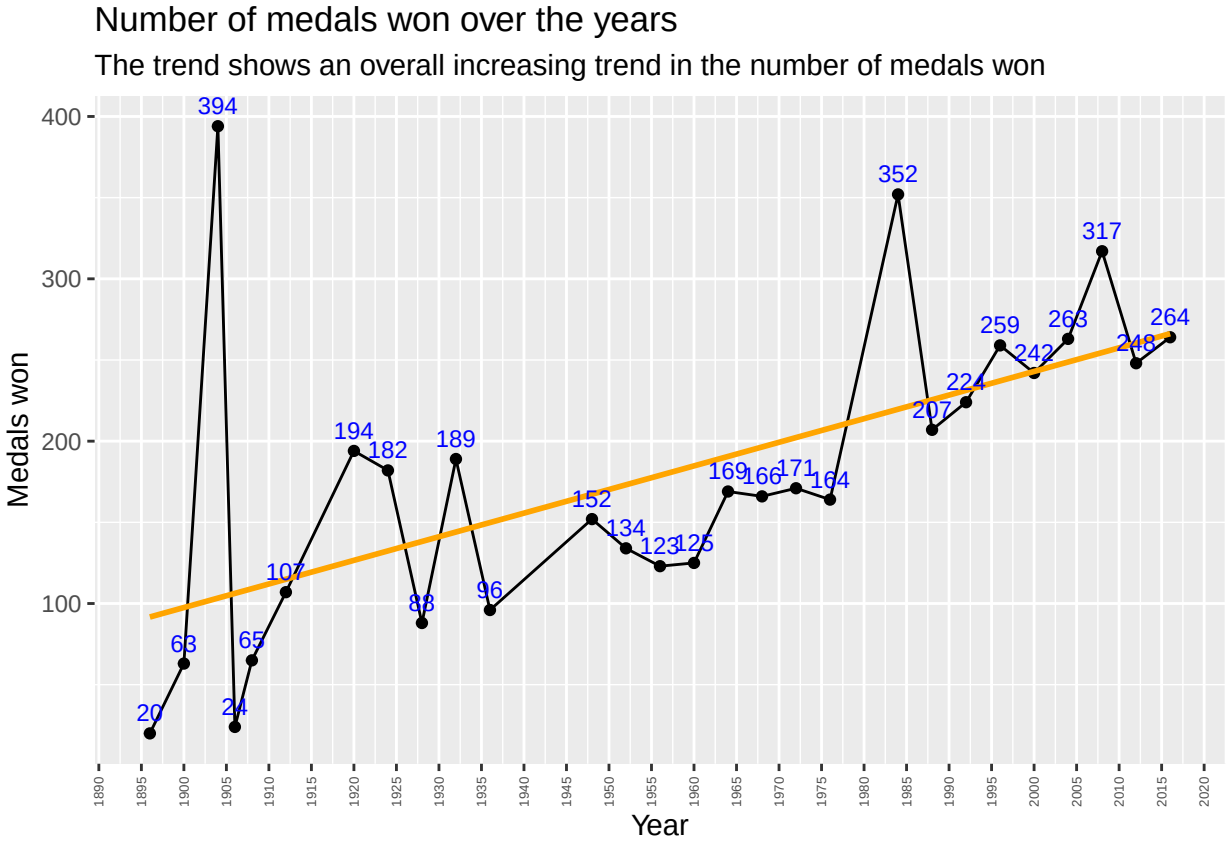
```
## `summarise()` regrouping output by 'Year' (override with `.groups` argument)
```

```
total_medals <- total_diff_medals %>% group_by(Year) %>%
  summarise(tally = sum(`medals per year`))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
ggplot(total_medals , aes(x =Year , y = tally)) +
  geom_line()+
  geom_point() +
  geom_smooth(se=FALSE , method = "lm" , color = 'orange')+
  geom_text(aes(x = Year , y = tally, label = tally),color = 'blue' , vjust = -0.7 , size = 3) +
  scale_x_continuous(expand = c(0.05, 0.5), n.breaks = 30) + vertical_x+
  labs(title = "Number of medals won over the years" ,
       subtitle = "The trend shows an overall increasing trend in the number of medals won" ,
       y = "Medals won")
```

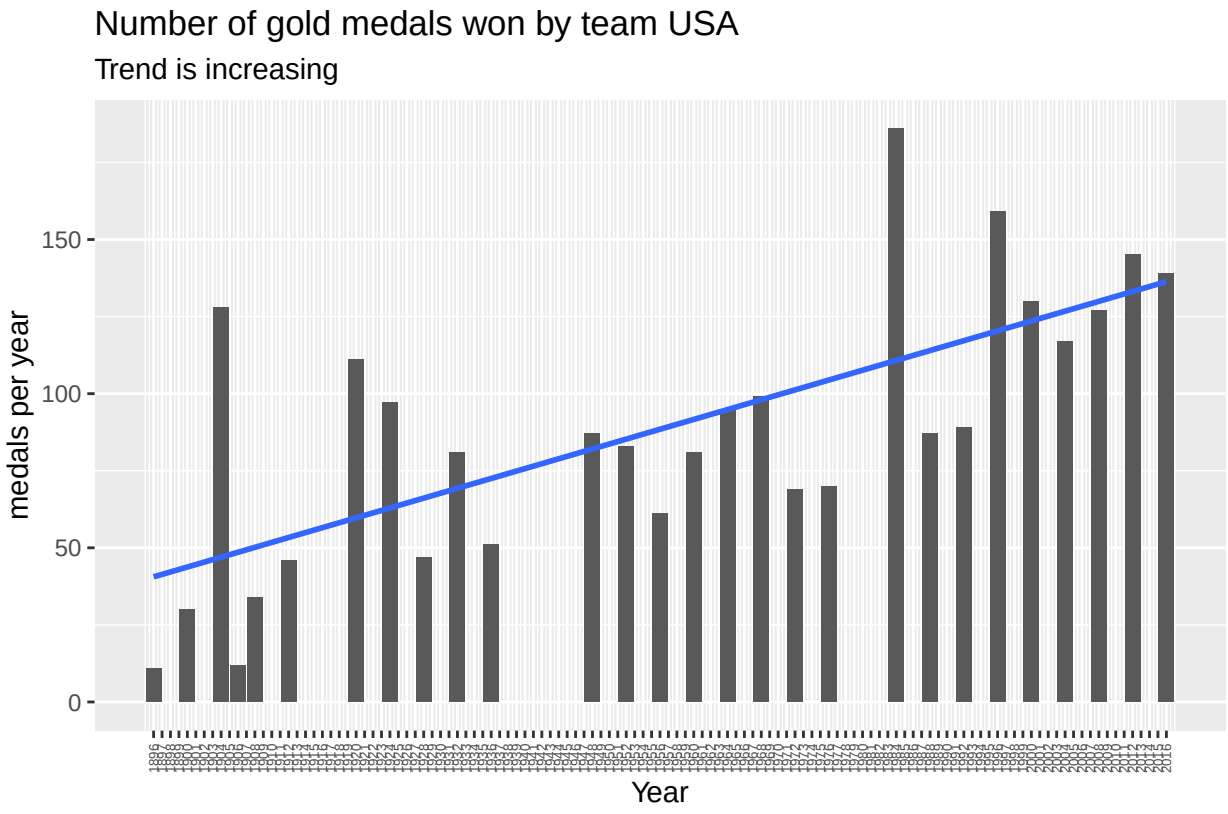
```
## `geom_smooth()` using formula 'y ~ x'
```



Next, we will explore the number of gold medals won over the years for USA

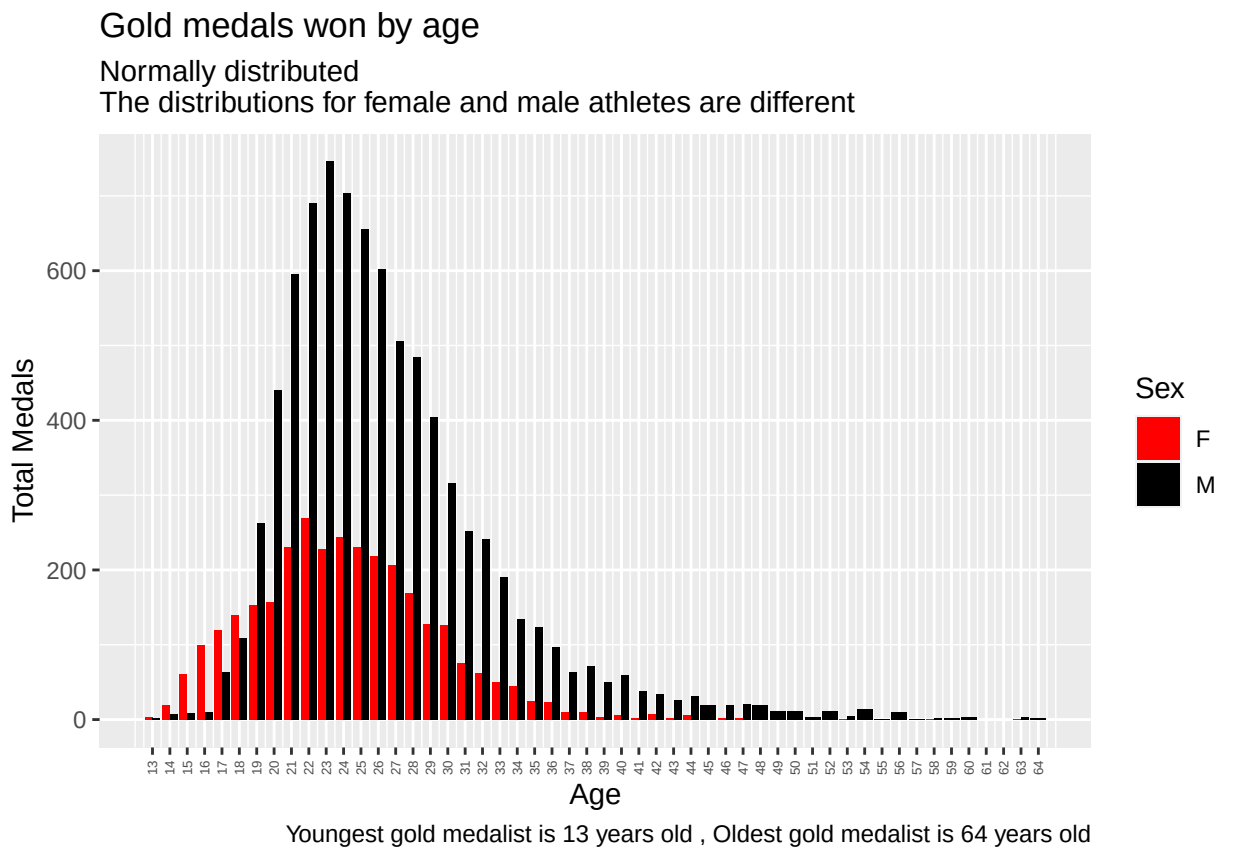
```
only_gold_usa <- total_diff_medals %>% filter(Medal == "Gold")

x_axis_labels <- min(only_gold_usa[,"Year"]):max(only_gold_usa[,"Year"])

all_x_values <- scale_x_continuous(labels = x_axis_labels, breaks = x_axis_labels)

ggplot(only_gold_usa , aes(x =Year , y = `medals per year`))+
  geom_col() + vertical_x+
  geom_smooth(se = FALSE , method = "lm") + all_x_values +
  labs(title = "Number of gold medals won by team USA" ,
       subtitle = "Trend is increasing" ,
       caption = "Olympics did not always take place in 4 year intervals , but why?")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Distribution of gold medals by age

```
gold_overall <- medals %>%filter(!is.na(Age)) %>%
  filter(Medal == "Gold") %>% group_by(Age , Medal)

x_axis_labels_2 <- min(gold_overall[,"Age"]):max(gold_overall[,"Age"])

all_x_values_2 <- scale_x_continuous(labels = x_axis_labels_2, breaks = x_axis_labels_2)
ggplot(data = gold_overall) +
  geom_bar(aes(x = Age , fill = Sex) , position = position_dodge()) +
  all_x_values_2 +
  vertical_x +
  scale_fill_manual("Sex" , values = c("M" = "black" , "F" = "red")) +
  labs(title = "Gold medals won by age" ,
       subtitle = "Normally distributed \nThe distributions for female and male athletes are different"
       caption = "Youngest gold medalist is 13 years old , Oldest gold medalist is 64 years old" ,
       y = "Total Medals")
```



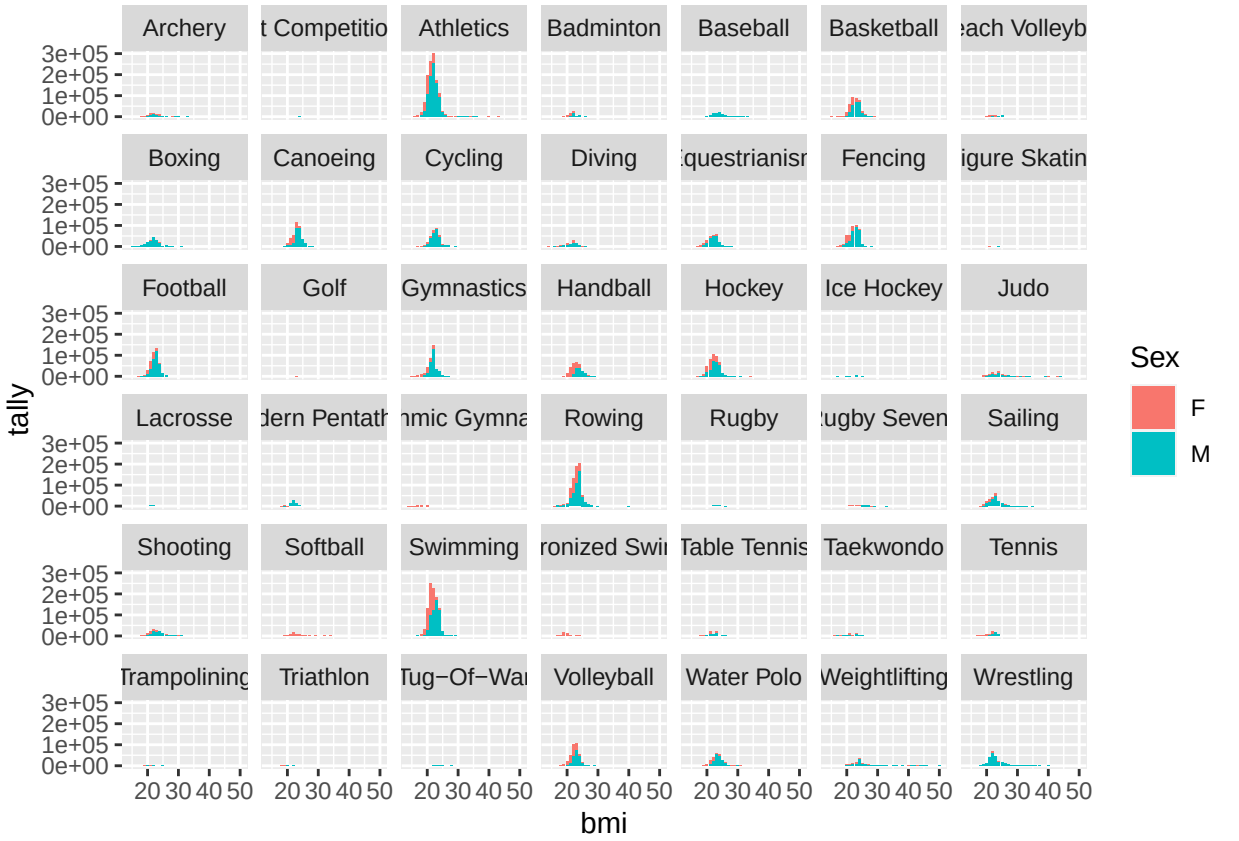
Relationship between BMI and Medals

Firstly, I would like to see if the number of medals won is related to the BMI of the athlete

```
bmi <- medals %>% filter(!is.na(Height)) %>%
  filter(!is.na(Weight)) %>%
  mutate(bmi = Weight/(Height/100)^2) %>%
  mutate(bmi = as.integer(bmi))

to_plot<- bmi %>% filter(Medal == "Gold") %>%
  group_by(Medal , bmi) %>%
  summarise(Sport , Sex , tally = n())
```

```
## `summarise()` regrouping output by 'Medal', 'bmi' (override with `.groups` argument)
```

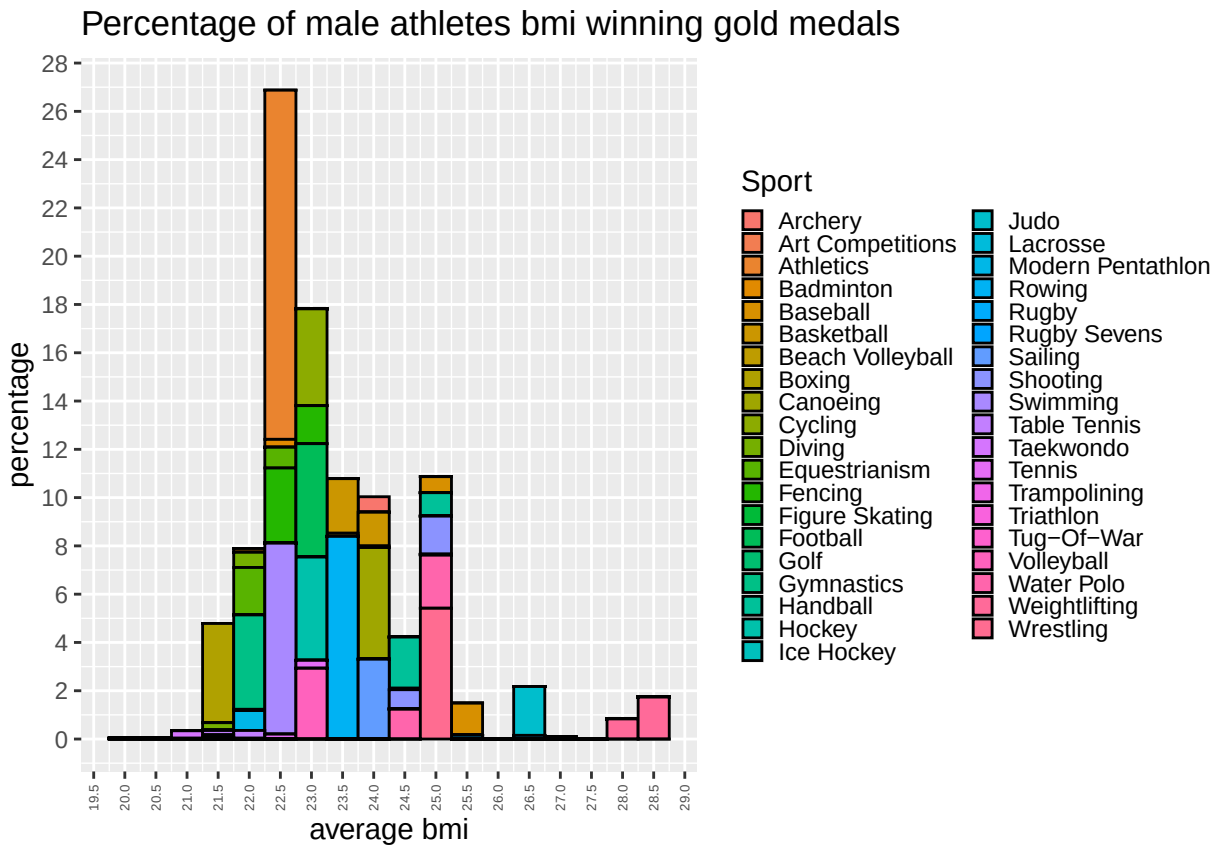
```
ggplot(to_plot , aes(x = bmi , y = tally , fill = Sex))+
  geom_col()+
  facet_wrap(vars(Sport))
```




```
bmi_avg <- bmi %>% filter(Sex == "M") %>% group_by(Medal , Sport) %>%
  summarise(avg_bmi =mean(bmi) , NOC)
```

```
## `summarise()` regrouping output by 'Medal', 'Sport' (override with `.groups` argument)
```

```
ggplot(data = bmi_avg , aes(x = avg_bmi , fill = Sport))+
  geom_histogram(aes(y = (..count..)/sum(..count..)*100) , color = "black" , binwidth = 0.5)+
  scale_x_continuous(n.breaks = 20) +
  scale_y_continuous(n.breaks = 20)+
  labs(y = "percentage" , x = "average bmi",
                                        title = "Percentage of male athletes bmi winning gold medals
  theme(legend.key.size = unit(3 , "mm")) +vertical_x
```



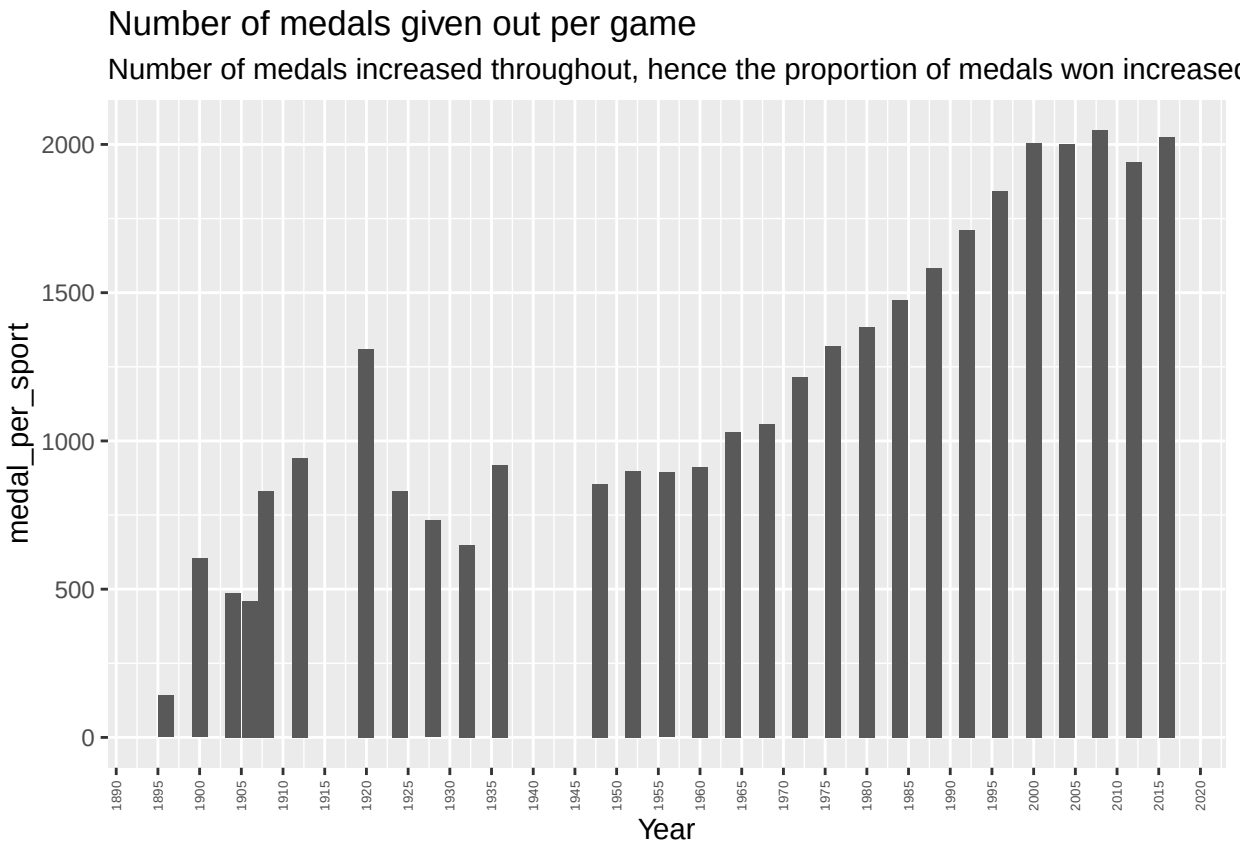
Questions to be answered 1) Find the density of age and how it changes overtime using geom density and make nearest graph more transparent. 2) Jaccard index is the number of medals won in the same events for diff countries / union of all events they have participated in 5) From the first graph, explore why was their a spike and a dip. Gain insights from dataset

```
num_medal <- medals %>% group_by(Sport, Medal) %>%
  summarise(total =n() , Year) %>% group_by(Sport , Year) %>% summarise(medal_per_sport = n())
```

```
## `summarise()` regrouping output by 'Sport', 'Medal' (override with `.groups` argument)
```

```
## `summarise()` regrouping output by 'Sport' (override with `.groups` argument)
```

```
medal_plot <-ggplot(data = num_medal , aes(x = Year , y = medal_per_sport))+
  geom_col()+
  scale_x_continuous(n.breaks = 30) + vertical_x + labs(title = "Number of medals given out per game",
                                                        subtitle = "Number of medals increased throughou
medal_plot
```

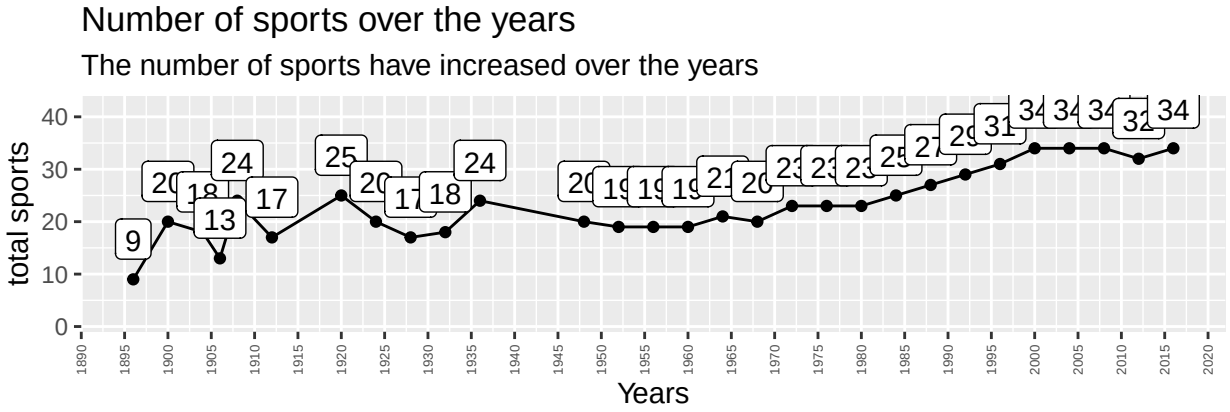


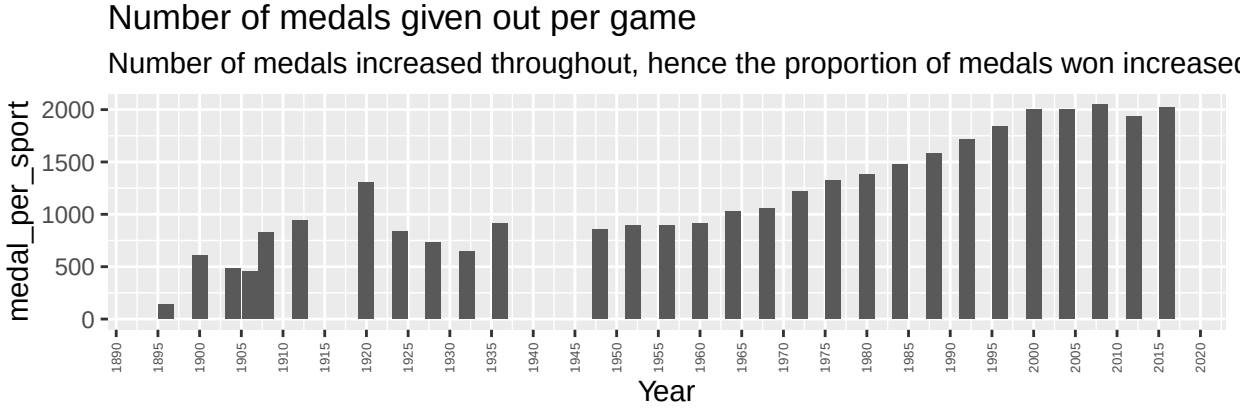
3) Number of sports that changed throughout the years and group the sports into different categories

```
num_sports <- medals %>% group_by(Year,Sport) %>%
  summarise(total_sports = n()) %>%
  summarise(total_sports = n())
```

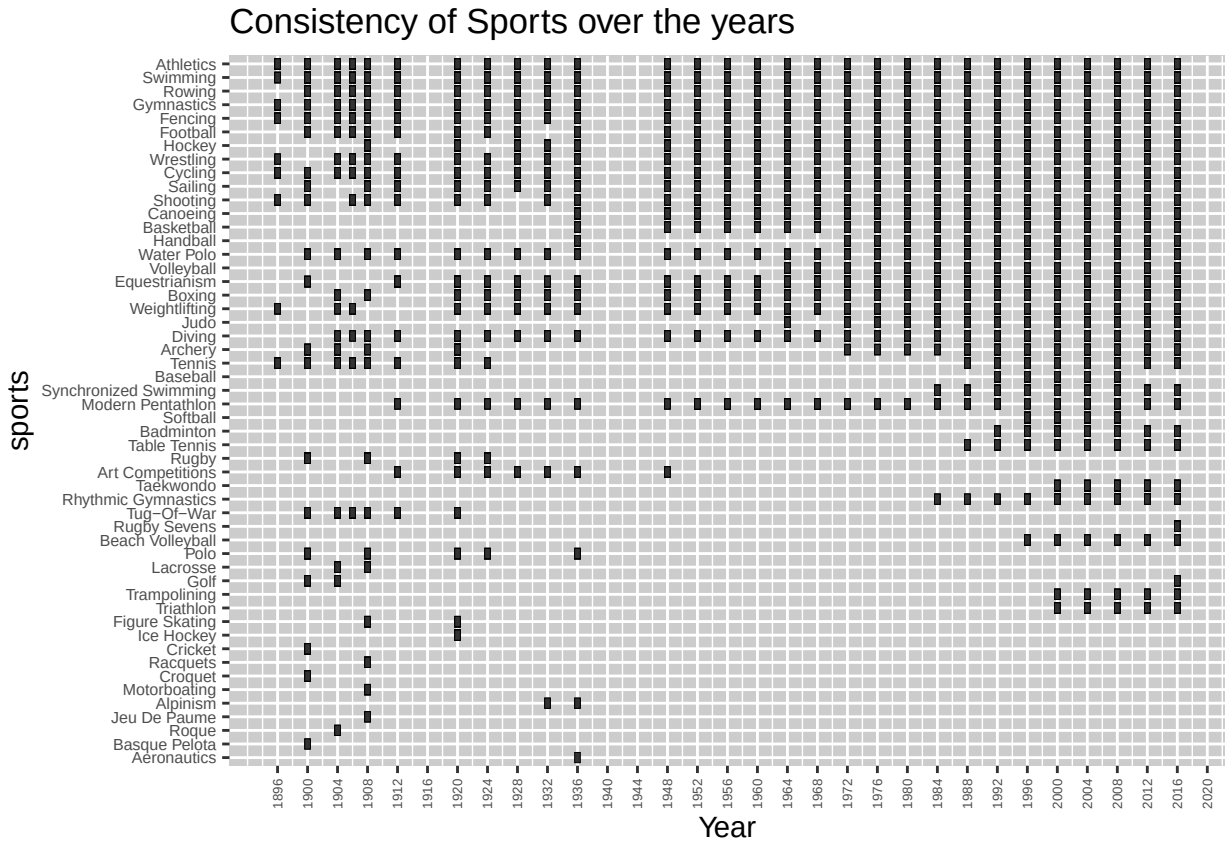
```
## `summarise()` regrouping output by 'Year' (override with `.groups` argument)

## `summarise()` ungrouping output (override with `.groups` argument)
```

```
sport_plot<- ggplot(data = num_sports , aes(x = Year , y = total_sports))+
  geom_line()+
  geom_point()+
  scale_x_continuous(n.breaks = 30) + vertical_x+
  scale_y_continuous(expand = c(0 ,10))+
  geom_label(aes(label = total_sports) ,vjust = -0.5) +
  labs(title = "Number of sports over the years" ,
       subtitle = "The number of sports have increased over the years" ,
       y = "total sports", x = "Years" )
gridExtra::grid.arrange(sport_plot,medal_plot , ncol = 1)
```






```
sports_overyrs <- medals %>% dplyr::select(Year,Sport)
  ggplot(sports_overyrs)+
    geom_tile(aes(Year , fct_rev(fct_infreq(Sport))),height=.75, width=.75 , color = "black") +
    scale_x_continuous(breaks = seq(1896,2020 ,4)) +vertical_x + labs(title = "Consistency of Sports ov
                                                                      y = "sports")+
    theme(axis.text.y = element_text(size = 6) , panel.background = element_rect(fill = "grey80",
                                                                                  color = "black",
                                                                                  size = 0.5 , linetype =
```

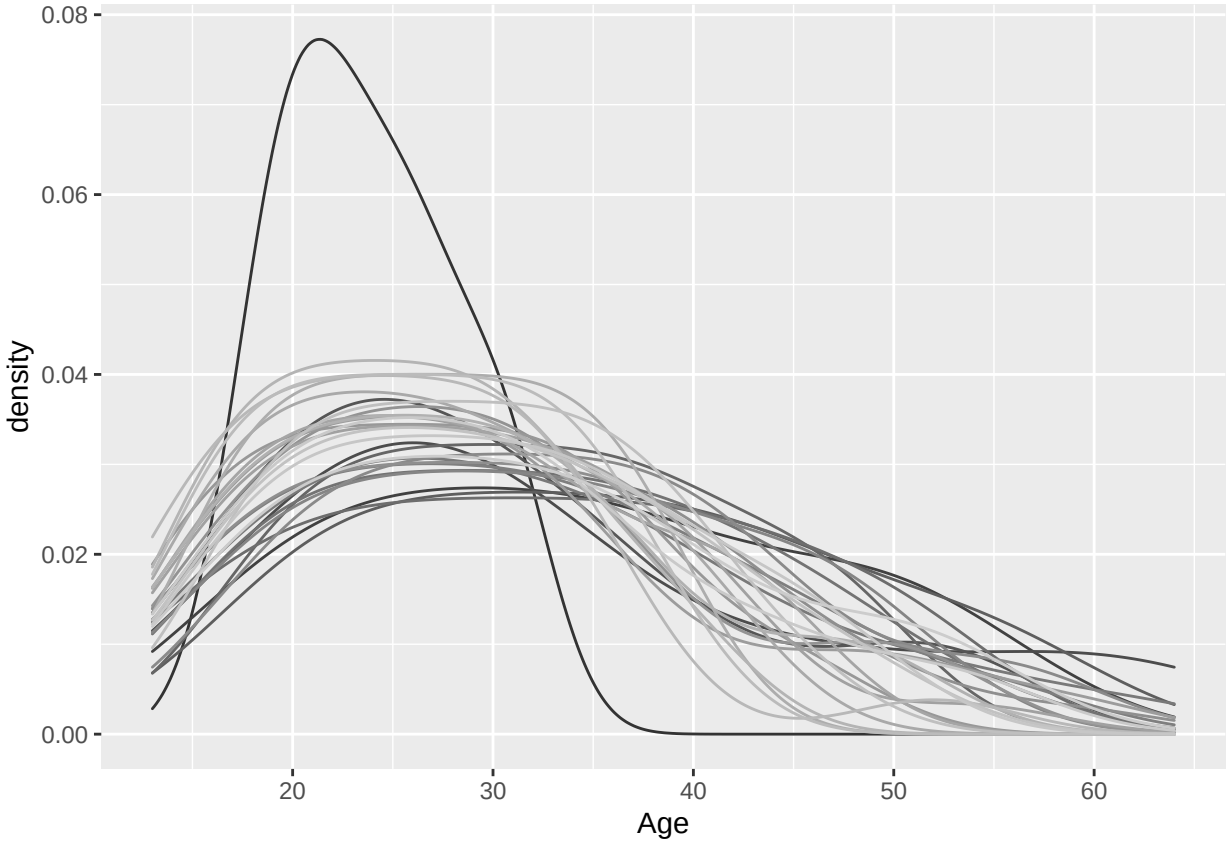



```
data_medal <- medals %>% filter(Medal == "Gold") %>%
  arrange(Year) %>%
  group_by(ID,Name) %>%
  slice_min(Year,n=1) %>%
  group_by(Age, Year) %>%
  summarise(tally = n()) %>%
  arrange(Year) %>% ungroup()
```

```
## `summarise()` regrouping output by 'Age' (override with `.groups` argument)
```

```
to_plot <- data_medal %>%
  mutate(alpha = ((Year-min(Year))/(max(Year) - min(Year))))

ggplot(data = to_plot , aes(x = Age)) +
  geom_density(aes(color = factor(alpha)) , show.legend = FALSE)+
  scale_color_grey()
```

```
## Warning: Removed 14 rows containing non-finite values (stat_density).
```


0.08
0.06
0.04
0.02
0.00
density
20
30
40
50
60
Age