Assignment\_DataVisualization\_CA3\_StoryTelling

Pranav\_Deogaonkar

09/01/2020

Table of Contents

[1. Introduction - 2](#_Toc29423378)

[2. Rationale : 2](#_Toc29423379)

[3. Dataset & variables 2](#_Toc29423380)

[4. Research Question/s: 2](#_Toc29423381)

[5. Data Analysis 3](#_Toc29423382)

[5.1 Importing data from “athlete\_event” and “noc\_regions” csv files and checking the dataframe structure and variables 3](#_Toc29423383)

[5.2 Performing Data transformation : 5](#_Toc29423384)

[5.3 Participation of athletes over time period from 1st Athens Olympics to 2016 Rio olympics 5](#_Toc29423385)

[5.4 Effects of 1972 turmoil at Munich Olympics 6](#_Toc29423386)

[5.5 Participation of athletes across time period as per gender: 11](#_Toc29423387)

[5.6 Participation of athletes as per season 12](#_Toc29423388)

[5.7 Relationship of height of athletes and their success rate at Olympic events for Male participants 13](#_Toc29423389)

[5.8 Relationship of height of athletes and their success rate at Olympic events for Female participants 16](#_Toc29423390)

[5.9 Is age a barrier to win in the Olympics events. 19](#_Toc29423391)

[5.10 Country wise performance at London 2012 Summer Olympic Games 21](#_Toc29423392)

[5.11 Cities that have hosted the Olympics: 23](#_Toc29423393)

[Cities around the world that have bid and won to host olympics more than once 24](#_Toc29423394)

[5.12 5 Trending Summer and Winter Olympic games before 1995: 25](#_Toc29423395)

[5.13 Trending Summer and Winter Olympic games from 1995 to recent Rio Olympics 28](#_Toc29423396)

[6. Insights: 31](#_Toc29423397)

[7. References: 31](#_Toc29423398)

# 1. Introduction -

The modern Olympic games are considered to be an elite global sports event which comprises of summer and winter sports contests. Thousands of athletes participate in the Olympics across the globe in vivid competitions. The Olympic games are conducted every four years, alternating of summer and winter Olympics with two years apart. The International Olympic Committee(IOC) was founded by “Baron Pierre de Coubertin” in 1894. Thus, the foundation for modern Olympic games was laid in 1896.(En.wikipedia.org, 2020)

Ancient Olympics:

Modern Olympics are influenced by the ancient Olympic Games, held in Olympia, Greece during the period from 8th century BC to the 4th century AD. These games were religious and athletic fiesta for the Greeks. Competitions were among different delegates of city-states and kingdoms of ancient Greece.(En.wikipedia.org, 2020)

Modern Olympic Games:

Under the patronage of IOC first Games were held in Athens in 1896. Evangelos Zappas was a pioneer leader in establishing the Modern Olympic Games. (En.wikipedia.org, 2020)

This brings us to a thought that from 1896 till now how have the Olympics changed and in what manner.

# 2. Rationale :

We will investigate into evolution of Olympic Games from the 19th century till the 21st century.

# 3. Dataset & variables

Dataset we have obtained from kaggle, which is “120 years of Olympic history: athletes and results”.It consists of two csv files named as “athlete\_events.csv” and “noc\_regions.csv”. The athelete\_events file contains data of all participants from 1896 to 2016 olympic games such as : Name Personal Traits : Sex,Age,Height and Weight Participant Country Details: Team,NOC,Region Games Details: Year,Games(held at Summer/Winter),Season Competetions: Sport,Event Host City: City Outcome of Competetion: Medal

(Kaggle.com, 2020)

# 4. Research Question/s:

To address the research question of how has Olympics evolved from the initial Industrial Revolution era to Information era. We will formulate sub research questions as below:

1. How has participation of athletes changed over the years.
2. Is there difference in participation as per Gender at Olympic events
3. Differences in Summer and Winter olympics across the period
4. Relationship of Age and Height versus Achievement at olympics.
5. Is there any change in popularity of Olympic games across the era.

# 5. Data Analysis

library(ggplot2)  
library(gridExtra)

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

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following object is masked from 'package:gridExtra':  
##   
## combine

## The following objects are masked from 'package:stats':  
##   
## filter, lag

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

## 5.1 Importing data from “athlete\_event” and “noc\_regions” csv files and checking the dataframe structure and variables

olympics\_dataframe <- read.csv("D://Submissions//RLab//Assignment3//120-years-of-olympic-history-athletes-and-results//athlete\_events.csv",header = TRUE,sep = ",",stringsAsFactors = FALSE)  
  
regions <- read.csv("D://Submissions//RLab//Assignment3//120-years-of-olympic-history-athletes-and-results//noc\_regions.csv",header = TRUE,sep = ",",stringsAsFactors = FALSE)  
  
View(olympics\_dataframe)  
str(olympics\_dataframe)

## 'data.frame': 271116 obs. of 15 variables:  
## $ ID : int 1 2 3 4 5 5 5 5 5 5 ...  
## $ Name : chr "A Dijiang" "A Lamusi" "Gunnar Nielsen Aaby" "Edgar Lindenau Aabye" ...  
## $ Sex : chr "M" "M" "M" "M" ...  
## $ Age : int 24 23 24 34 21 21 25 25 27 27 ...  
## $ Height: int 180 170 NA NA 185 185 185 185 185 185 ...  
## $ Weight: num 80 60 NA NA 82 82 82 82 82 82 ...  
## $ Team : chr "China" "China" "Denmark" "Denmark/Sweden" ...  
## $ NOC : chr "CHN" "CHN" "DEN" "DEN" ...  
## $ Games : chr "1992 Summer" "2012 Summer" "1920 Summer" "1900 Summer" ...  
## $ Year : int 1992 2012 1920 1900 1988 1988 1992 1992 1994 1994 ...  
## $ Season: chr "Summer" "Summer" "Summer" "Summer" ...  
## $ City : chr "Barcelona" "London" "Antwerpen" "Paris" ...  
## $ Sport : chr "Basketball" "Judo" "Football" "Tug-Of-War" ...  
## $ Event : chr "Basketball Men's Basketball" "Judo Men's Extra-Lightweight" "Football Men's Football" "Tug-Of-War Men's Tug-Of-War" ...  
## $ Medal : chr NA NA NA "Gold" ...

summary(olympics\_dataframe)

## ID Name Sex Age   
## Min. : 1 Length:271116 Length:271116 Min. :10.00   
## 1st Qu.: 34643 Class :character Class :character 1st Qu.:21.00   
## Median : 68205 Mode :character Mode :character Median :24.00   
## Mean : 68249 Mean :25.56   
## 3rd Qu.:102097 3rd Qu.:28.00   
## Max. :135571 Max. :97.00   
## NA's :9474   
## Height Weight Team NOC   
## Min. :127.0 Min. : 25.0 Length:271116 Length:271116   
## 1st Qu.:168.0 1st Qu.: 60.0 Class :character Class :character   
## Median :175.0 Median : 70.0 Mode :character Mode :character   
## Mean :175.3 Mean : 70.7   
## 3rd Qu.:183.0 3rd Qu.: 79.0   
## Max. :226.0 Max. :214.0   
## NA's :60171 NA's :62875   
## Games Year Season City   
## Length:271116 Min. :1896 Length:271116 Length:271116   
## Class :character 1st Qu.:1960 Class :character Class :character   
## Mode :character Median :1988 Mode :character Mode :character   
## Mean :1978   
## 3rd Qu.:2002   
## Max. :2016   
##   
## Sport Event Medal   
## Length:271116 Length:271116 Length:271116   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##

## 5.2 Performing Data transformation :

# Converting variables "Sex" and "Season" to factor variables  
  
olympics\_dataframe <- olympics\_dataframe%>%mutate\_at(vars("Sex","Season"),funs(factor))

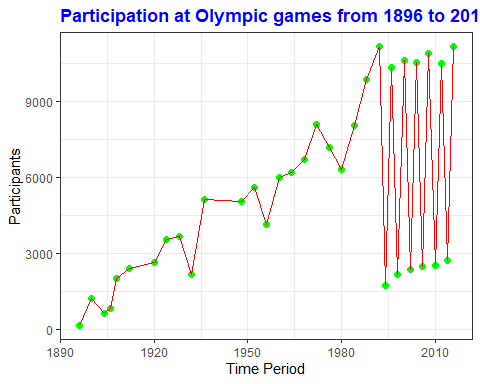
## Warning: funs() is soft deprecated as of dplyr 0.8.0  
## Please use a list of either functions or lambdas:   
##   
## # Simple named list:   
## list(mean = mean, median = median)  
##   
## # Auto named with `tibble::lst()`:   
## tibble::lst(mean, median)  
##   
## # Using lambdas  
## list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))  
## This warning is displayed once per session.

str(olympics\_dataframe)

## 'data.frame': 271116 obs. of 15 variables:  
## $ ID : int 1 2 3 4 5 5 5 5 5 5 ...  
## $ Name : chr "A Dijiang" "A Lamusi" "Gunnar Nielsen Aaby" "Edgar Lindenau Aabye" ...  
## $ Sex : Factor w/ 2 levels "F","M": 2 2 2 2 1 1 1 1 1 1 ...  
## $ Age : int 24 23 24 34 21 21 25 25 27 27 ...  
## $ Height: int 180 170 NA NA 185 185 185 185 185 185 ...  
## $ Weight: num 80 60 NA NA 82 82 82 82 82 82 ...  
## $ Team : chr "China" "China" "Denmark" "Denmark/Sweden" ...  
## $ NOC : chr "CHN" "CHN" "DEN" "DEN" ...  
## $ Games : chr "1992 Summer" "2012 Summer" "1920 Summer" "1900 Summer" ...  
## $ Year : int 1992 2012 1920 1900 1988 1988 1992 1992 1994 1994 ...  
## $ Season: Factor w/ 2 levels "Summer","Winter": 1 1 1 1 2 2 2 2 2 2 ...  
## $ City : chr "Barcelona" "London" "Antwerpen" "Paris" ...  
## $ Sport : chr "Basketball" "Judo" "Football" "Tug-Of-War" ...  
## $ Event : chr "Basketball Men's Basketball" "Judo Men's Extra-Lightweight" "Football Men's Football" "Tug-Of-War Men's Tug-Of-War" ...  
## $ Medal : chr NA NA NA "Gold" ...

## 5.3 Participation of athletes over time period from 1st Athens Olympics to 2016 Rio olympics

# Participation of athletes over time period from 1st Athens Olympics to 2016 Rio olympics  
  
Participants\_Year <- olympics\_dataframe%>%group\_by(Year)%>%summarise(Participants\_Year = n\_distinct(ID))  
plot\_year <- ggplot(data=Participants\_Year,aes(x=Year,y=Participants\_Year))+geom\_point(size=2,colour="green")+geom\_line(colour="red")+theme\_bw()+  
 theme(plot.title = element\_text(face = "bold",colour = "blue"))+  
 labs(title = "Participation at Olympic games from 1896 to 2016",x="Time Period",y="Participants")  
plot\_year



## 5.4 Effects of 1972 turmoil at Munich Olympics

olympics\_dataframe <- olympics\_dataframe%>%left\_join(regions, by = "NOC")  
  
  
munich\_participation <- olympics\_dataframe%>%filter(Games == "1972 Summer")%>%group\_by(region)%>%summarise(Munich\_part = n\_distinct(ID))%>%mutate(Country\_max\_part = min\_rank(desc(Munich\_part)))%>%filter(Country\_max\_part %in% c("1":"3"))%>%mutate(No\_Participants\_Munich = sum(Munich\_part))  
  
munich\_participation

## # A tibble: 3 x 4  
## region Munich\_part Country\_max\_part No\_Participants\_Munich  
## <chr> <int> <int> <int>  
## 1 Germany 721 1 1492  
## 2 Russia 371 3 1492  
## 3 USA 400 2 1492

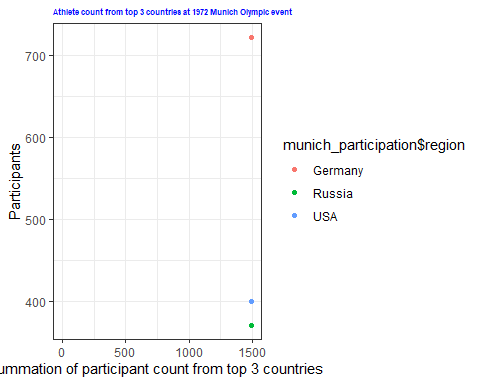
canada\_participation <- olympics\_dataframe%>%filter(Games == "1976 Summer")%>%group\_by(region)%>%summarise(canada\_part = n\_distinct(ID))%>%mutate(Country\_max\_part = min\_rank(desc(canada\_part)))%>%filter(Country\_max\_part %in% c("1":"3"))%>%mutate(No\_Participants\_canada = sum(canada\_part))  
canada\_participation

## # A tibble: 3 x 4  
## region canada\_part Country\_max\_part No\_Participants\_canada  
## <chr> <int> <int> <int>  
## 1 Germany 557 1 1363  
## 2 Russia 410 2 1363  
## 3 USA 396 3 1363

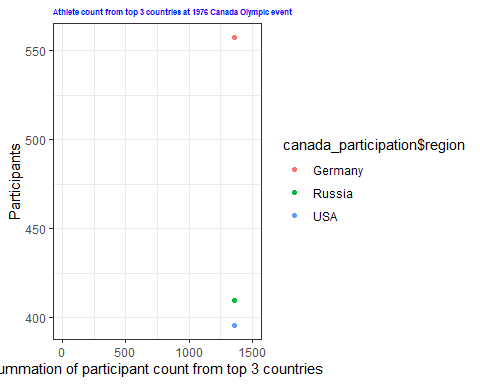
russia\_participation <- olympics\_dataframe%>%filter(Games == "1980 Summer")%>%group\_by(region)%>%summarise(russia\_part = n\_distinct(ID))%>%mutate(Country\_max\_part = min\_rank(desc(russia\_part)))%>%filter(Country\_max\_part %in% c("1":"3"))%>%mutate(No\_Participants\_russia = sum(russia\_part))  
  
russia\_participation

## # A tibble: 3 x 4  
## region russia\_part Country\_max\_part No\_Participants\_russia  
## <chr> <int> <int> <int>  
## 1 Germany 346 2 1141  
## 2 Poland 306 3 1141  
## 3 Russia 489 1 1141

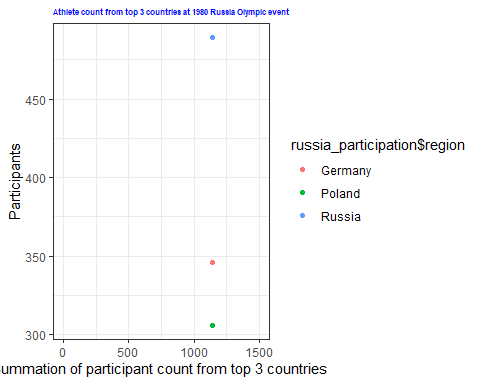
plot\_1 <- ggplot(data=munich\_participation,aes(x=munich\_participation$No\_Participants\_Munich,y=Munich\_part,fill=munich\_participation$region)) + geom\_point(aes(colour=munich\_participation$region))+scale\_x\_continuous(limits = c(0,1500))+theme\_bw()+  
 theme(plot.title = element\_text(face = "bold",colour = "blue",size = 6))+  
 labs(title = "Athlete count from top 3 countries at 1972 Munich Olympic event ",x="Summation of participant count from top 3 countries ",y="Participants")  
plot\_1



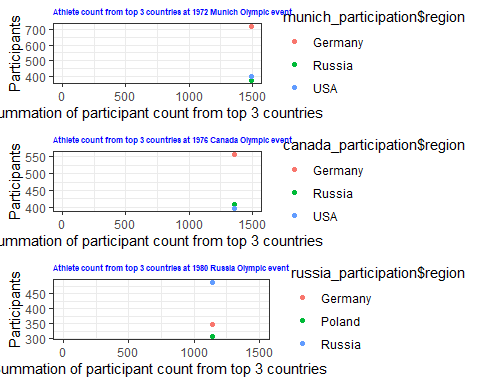
plot\_2 <- ggplot(data=canada\_participation,aes(x=canada\_participation$No\_Participants\_canada,y=canada\_part,fill=canada\_participation$region)) + geom\_point(aes(colour=canada\_participation$region)) +scale\_x\_continuous(limits = c(0,1500))+theme\_bw()+  
 theme(plot.title = element\_text(face = "bold",colour = "blue",size = 6))+  
 labs(title = "Athlete count from top 3 countries at 1976 Canada Olympic event ",x="Summation of participant count from top 3 countries ",y="Participants")  
plot\_2



plot\_3 <- ggplot(data=russia\_participation,aes(x=russia\_participation$No\_Participants\_russia,y=russia\_part,fill=russia\_participation$region)) + geom\_point(aes(colour=russia\_participation$region)) +scale\_x\_continuous(limits = c(0,1500))+theme\_bw()+  
 theme(plot.title = element\_text(face = "bold",colour = "blue",size = 6))+  
 labs(title = "Athlete count from top 3 countries at 1980 Russia Olympic event ",x="Summation of participant count from top 3 countries ",y="Participants")  
plot\_3



grid.arrange(plot\_1,plot\_2,plot\_3,nrow=3)

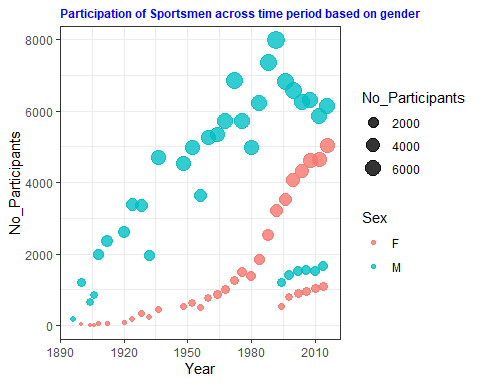


## 5.5 Participation of athletes across time period as per gender:

participation\_olympics <- olympics\_dataframe%>%group\_by(Year,Sex)%>%summarise(No\_Participants =n\_distinct(ID))  
(participation\_olympics)

## # A tibble: 69 x 3  
## # Groups: Year [35]  
## Year Sex No\_Participants  
## <int> <fct> <int>  
## 1 1896 M 176  
## 2 1900 F 23  
## 3 1900 M 1201  
## 4 1904 F 6  
## 5 1904 M 644  
## 6 1906 F 6  
## 7 1906 M 835  
## 8 1908 F 44  
## 9 1908 M 1980  
## 10 1912 F 53  
## # ... with 59 more rows

plot <- ggplot(participation\_olympics,aes(x=Year,y=No\_Participants,group = Sex,fill = Sex,size = No\_Participants,color= Sex))+ geom\_point(alpha=0.8)+theme\_bw()+  
 theme(plot.title = element\_text(face = "bold",colour = "blue",size = 9))+  
 labs(title = "Participation of Sportsmen across time period based on gender")  
plot

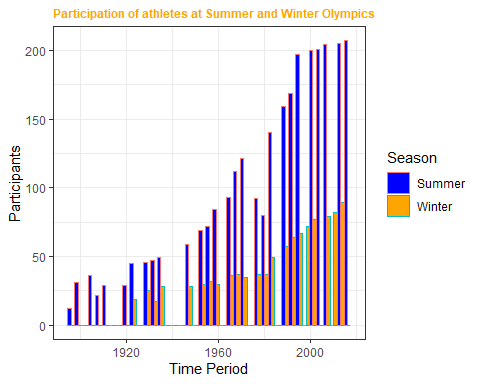


## 5.6 Participation of athletes as per season

#Participation of athletes as per season  
Athlete\_Participation\_Season <- olympics\_dataframe%>%group\_by(Year,Season,NOC)%>%summarise(Athletes\_As\_Per\_Season= n\_distinct(ID))  
Athlete\_Participation\_Season

## # A tibble: 3,837 x 4  
## # Groups: Year, Season [51]  
## Year Season NOC Athletes\_As\_Per\_Season  
## <int> <fct> <chr> <int>  
## 1 1896 Summer AUS 1  
## 2 1896 Summer AUT 3  
## 3 1896 Summer DEN 3  
## 4 1896 Summer FRA 12  
## 5 1896 Summer GBR 10  
## 6 1896 Summer GER 19  
## 7 1896 Summer GRE 102  
## 8 1896 Summer HUN 7  
## 9 1896 Summer ITA 1  
## 10 1896 Summer SUI 3  
## # ... with 3,827 more rows

Athlete\_plot <- ggplot(Athlete\_Participation\_Season,aes(x=Year,color=Season,group=Season),alpha=0.5)+geom\_histogram(aes(fill=Season),position = "dodge",binwidth = 3)+scale\_fill\_manual(values=c("blue","orange"))+theme\_bw()+  
 theme(plot.title = element\_text(face = "bold",colour = "orange",size = 9))+  
 labs(title = "Participation of athletes at Summer and Winter Olympics",x="Time Period",y="Participants")  
Athlete\_plot



## 5.7 Relationship of height of athletes and their success rate at Olympic events for Male participants

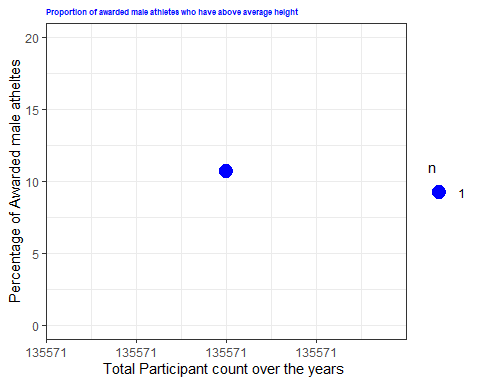
# Does Height affect your chances to win a medal at Olympics ?  
  
  
# Male participation:  
 Male\_Participation\_Height <- olympics\_dataframe%>%filter(Sex=="M",Height >mean(!is.na(olympics\_dataframe$Height)),Medal%in%c("Gold","Silver","Bronze"))%>%summarise(Successful\_Atheltes\_M = n\_distinct(ID))%>%mutate(Proportion\_SuccessfulMales = Successful\_Atheltes\_M/n\_distinct(olympics\_dataframe$ID)\*100)  
  
Male\_Participation\_Height

## Successful\_Atheltes\_M Proportion\_SuccessfulMales  
## 1 14461 10.66674

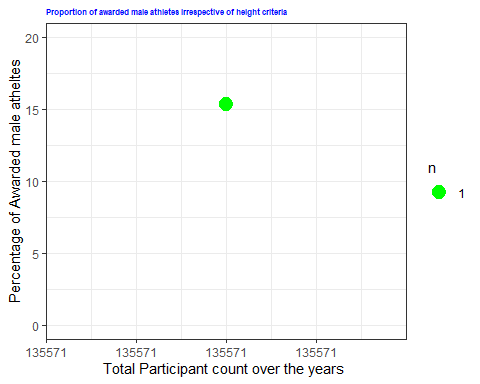
Male\_Participation\_Won <- olympics\_dataframe%>%filter(Sex=="M",Medal%in%c("Gold","Silver","Bronze"))%>%summarise(Successful\_Atheltes\_M = n\_distinct(ID))%>%mutate(Proportion\_SuccessfulMales = Successful\_Atheltes\_M/n\_distinct(olympics\_dataframe$ID)\*100)  
  
Male\_Participation\_Won

## Successful\_Atheltes\_M Proportion\_SuccessfulMales  
## 1 20797 15.3403

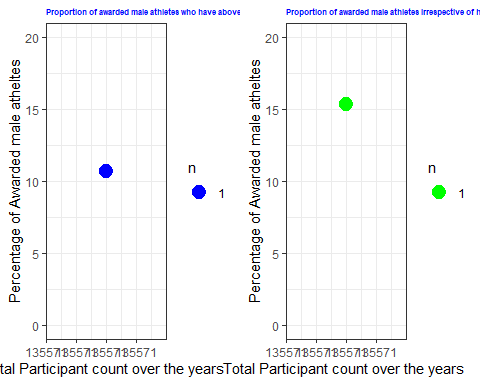
plot\_Male\_Successful\_Height <- ggplot(Male\_Participation\_Height,aes(x=n\_distinct(olympics\_dataframe$ID),y=Proportion\_SuccessfulMales)) + geom\_count(colour="blue") +scale\_y\_continuous(limits = c(0,20))+ theme\_bw()+  
 theme(plot.title = element\_text(face = "bold",colour = "blue",size=6))+  
 labs(title = "Proportion of awarded male athletes who have above average height ",x="Total Participant count over the years ",y="Percentage of Awarded male atheltes")  
plot\_Male\_Successful\_Height



plot\_Male\_Successful <- ggplot(Male\_Participation\_Won,aes(x=n\_distinct(olympics\_dataframe$ID),y=Proportion\_SuccessfulMales)) + geom\_count(colour="green") +scale\_y\_continuous(limits = c(0,20))+ theme\_bw() +  
 theme(plot.title = element\_text(face = "bold",colour = "blue",size=6))+  
 labs(title = "Proportion of awarded male athletes irrespective of height criteria ",x="Total Participant count over the years ",y="Percentage of Awarded male atheltes")  
plot\_Male\_Successful



grid.arrange(plot\_Male\_Successful\_Height,plot\_Male\_Successful,ncol=2)



## 5.8 Relationship of height of athletes and their success rate at Olympic events for Female participants

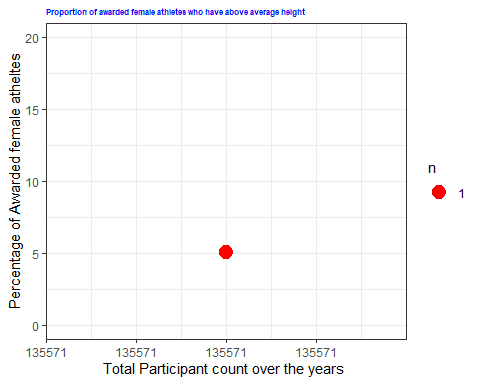
#Female participation:  
 Female\_Participation\_Height <- olympics\_dataframe%>%filter(Sex=="F",Height >mean(!is.na(olympics\_dataframe$Height)),Medal%in%c("Gold","Silver","Bronze"))%>%summarise(Successful\_Atheltes\_F = n\_distinct(ID))%>%mutate(Proportion\_Sucessful\_Females = Successful\_Atheltes\_F/n\_distinct(olympics\_dataframe$ID)\*100)  
  
Female\_Participation\_Height

## Successful\_Atheltes\_F Proportion\_Sucessful\_Females  
## 1 6866 5.064505

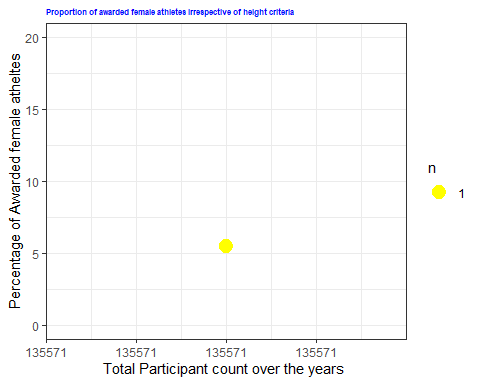
Female\_Participation\_Won <- olympics\_dataframe%>%filter(Sex=="F",Medal%in%c("Gold","Silver","Bronze"))%>%summarise(Successful\_Atheltes\_F = n\_distinct(ID))%>%mutate(Proportion\_Sucessful\_Females = Successful\_Atheltes\_F/n\_distinct(olympics\_dataframe$ID)\*100)  
  
Female\_Participation\_Won

## Successful\_Atheltes\_F Proportion\_Sucessful\_Females  
## 1 7454 5.498226

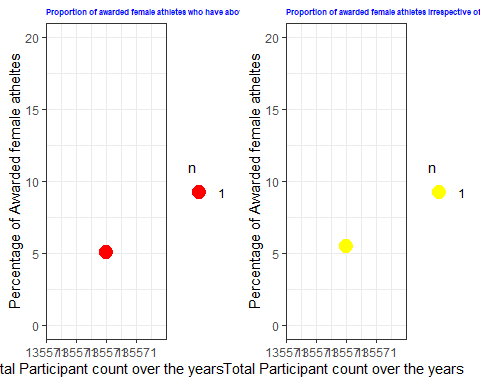
plot\_Female\_Successful\_Height <- ggplot(Female\_Participation\_Height,aes(x=n\_distinct(olympics\_dataframe$ID),y=Proportion\_Sucessful\_Females)) + geom\_count(colour="red") +scale\_y\_continuous(limits = c(0,20)) + theme\_bw()+  
 theme(plot.title = element\_text(face = "bold",colour = "blue",size=6))+  
 labs(title = "Proportion of awarded female athletes who have above average height ",x="Total Participant count over the years ",y="Percentage of Awarded female atheltes")  
plot\_Female\_Successful\_Height



plot\_Female\_Successful <- ggplot(Female\_Participation\_Won,aes(x=n\_distinct(olympics\_dataframe$ID),y=Proportion\_Sucessful\_Females)) + geom\_count(colour="yellow") +scale\_y\_continuous(limits = c(0,20)) + theme\_bw()+  
 theme(plot.title = element\_text(face = "bold",colour = "blue",size=6))+  
 labs(title = "Proportion of awarded female athletes irrespective of height criteria ",x="Total Participant count over the years ",y="Percentage of Awarded female atheltes")  
plot\_Female\_Successful



grid.arrange(plot\_Female\_Successful\_Height,plot\_Female\_Successful,ncol=2)



## 5.9 Is age a barrier to win in the Olympics events.

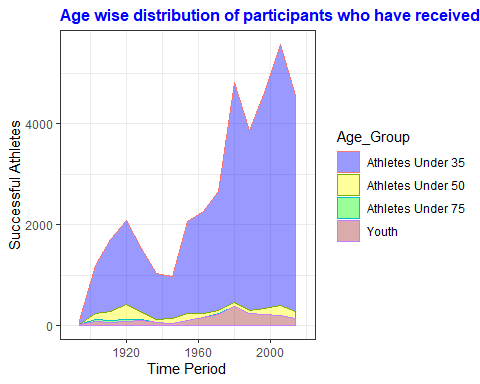
# Does Age have an effect on participation and wining formula of athletes in Olympics  
  
  
olympics\_dataframe <- olympics\_dataframe%>%mutate(Age\_Group = case\_when(Age >= 10 & Age <= 18 ~ 'Youth',  
 Age >= 19 & Age <= 36 ~ 'Athletes Under 35',  
 Age >= 37 & Age <= 50 ~ 'Athletes Under 50',   
 Age >= 51 & Age <= 75 ~'Athletes Under 75',  
 Age >= 76 & Age <100 ~ 'Senior Atheletes' ))  
  
str(olympics\_dataframe)

## 'data.frame': 271116 obs. of 18 variables:  
## $ ID : int 1 2 3 4 5 5 5 5 5 5 ...  
## $ Name : chr "A Dijiang" "A Lamusi" "Gunnar Nielsen Aaby" "Edgar Lindenau Aabye" ...  
## $ Sex : Factor w/ 2 levels "F","M": 2 2 2 2 1 1 1 1 1 1 ...  
## $ Age : int 24 23 24 34 21 21 25 25 27 27 ...  
## $ Height : int 180 170 NA NA 185 185 185 185 185 185 ...  
## $ Weight : num 80 60 NA NA 82 82 82 82 82 82 ...  
## $ Team : chr "China" "China" "Denmark" "Denmark/Sweden" ...  
## $ NOC : chr "CHN" "CHN" "DEN" "DEN" ...  
## $ Games : chr "1992 Summer" "2012 Summer" "1920 Summer" "1900 Summer" ...  
## $ Year : int 1992 2012 1920 1900 1988 1988 1992 1992 1994 1994 ...  
## $ Season : Factor w/ 2 levels "Summer","Winter": 1 1 1 1 2 2 2 2 2 2 ...  
## $ City : chr "Barcelona" "London" "Antwerpen" "Paris" ...  
## $ Sport : chr "Basketball" "Judo" "Football" "Tug-Of-War" ...  
## $ Event : chr "Basketball Men's Basketball" "Judo Men's Extra-Lightweight" "Football Men's Football" "Tug-Of-War Men's Tug-Of-War" ...  
## $ Medal : chr NA NA NA "Gold" ...  
## $ region : chr "China" "China" "Denmark" "Denmark" ...  
## $ notes : chr "" "" "" "" ...  
## $ Age\_Group: chr "Athletes Under 35" "Athletes Under 35" "Athletes Under 35" "Athletes Under 35" ...

olympics\_dataframe$Age\_Group <- as.factor(olympics\_dataframe$Age\_Group)  
str(olympics\_dataframe)

## 'data.frame': 271116 obs. of 18 variables:  
## $ ID : int 1 2 3 4 5 5 5 5 5 5 ...  
## $ Name : chr "A Dijiang" "A Lamusi" "Gunnar Nielsen Aaby" "Edgar Lindenau Aabye" ...  
## $ Sex : Factor w/ 2 levels "F","M": 2 2 2 2 1 1 1 1 1 1 ...  
## $ Age : int 24 23 24 34 21 21 25 25 27 27 ...  
## $ Height : int 180 170 NA NA 185 185 185 185 185 185 ...  
## $ Weight : num 80 60 NA NA 82 82 82 82 82 82 ...  
## $ Team : chr "China" "China" "Denmark" "Denmark/Sweden" ...  
## $ NOC : chr "CHN" "CHN" "DEN" "DEN" ...  
## $ Games : chr "1992 Summer" "2012 Summer" "1920 Summer" "1900 Summer" ...  
## $ Year : int 1992 2012 1920 1900 1988 1988 1992 1992 1994 1994 ...  
## $ Season : Factor w/ 2 levels "Summer","Winter": 1 1 1 1 2 2 2 2 2 2 ...  
## $ City : chr "Barcelona" "London" "Antwerpen" "Paris" ...  
## $ Sport : chr "Basketball" "Judo" "Football" "Tug-Of-War" ...  
## $ Event : chr "Basketball Men's Basketball" "Judo Men's Extra-Lightweight" "Football Men's Football" "Tug-Of-War Men's Tug-Of-War" ...  
## $ Medal : chr NA NA NA "Gold" ...  
## $ region : chr "China" "China" "Denmark" "Denmark" ...  
## $ notes : chr "" "" "" "" ...  
## $ Age\_Group: Factor w/ 5 levels "Athletes Under 35",..: 1 1 1 1 1 1 1 1 1 1 ...

Age\_Plot <- olympics\_dataframe%>%filter(!is.na(Age\_Group),Medal %in% c("Gold","Silver","Bronze"))%>%ggplot(aes(Year,fill=Age\_Group))+geom\_area(alpha=0.4,aes(color=Age\_Group),stat= "bin",bins = 15)+scale\_fill\_manual(values=c("blue","yellow","green","brown","red"))+theme\_bw()+  
 theme(plot.title = element\_text(face = "bold",colour = "blue",size = 12))+  
 labs(title = "Age wise distribution of participants who have received medal at olympic events",x="Time Period",y="Successful Athletes")  
Age\_Plot



Age\_Distribution <- olympics\_dataframe%>%filter(!is.na(Age\_Group),Medal %in% c("Gold","Silver","Bronze"))%>%group\_by(Age\_Group)%>%summarise(No\_Participants\_Agewise = n())  
Age\_Distribution

## # A tibble: 4 x 2  
## Age\_Group No\_Participants\_Agewise  
## <fct> <int>  
## 1 Athletes Under 35 35074  
## 2 Athletes Under 50 1735  
## 3 Athletes Under 75 203  
## 4 Youth 2039

## 5.10 Country wise performance at London 2012 Summer Olympic Games

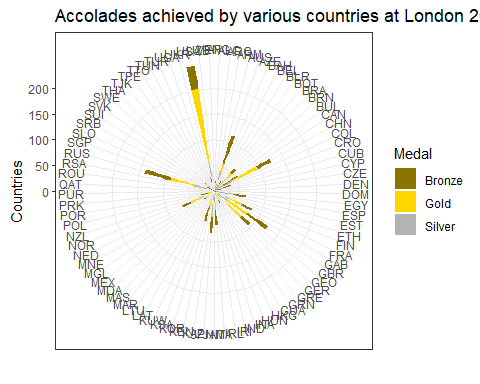
# Medals received by various countries in 2012  
  
Medals\_2012 <- olympics\_dataframe%>%filter(Year=="2012",!is.na(Medal))%>%group\_by(NOC,Medal)%>%summarise(No\_medals =length(Medal))%>%group\_by(NOC)%>%mutate(top\_25 = min\_rank(desc(sum(No\_medals))))%>%filter(between(top\_25,1,25))  
  
Medals\_2012

## # A tibble: 191 x 4  
## # Groups: NOC [85]  
## NOC Medal No\_medals top\_25  
## <chr> <chr> <int> <int>  
## 1 AFG Bronze 1 1  
## 2 ALG Gold 1 1  
## 3 ARG Bronze 3 1  
## 4 ARG Gold 1 1  
## 5 ARG Silver 15 1  
## 6 ARM Bronze 2 1  
## 7 ARM Silver 1 1  
## 8 AUS Bronze 59 1  
## 9 AUS Gold 18 1  
## 10 AUS Silver 37 1  
## # ... with 181 more rows

Medals\_2012$NOC <- as.factor(Medals\_2012$NOC)  
  
Medal\_plot <- ggplot(Medals\_2012, aes(x=Medals\_2012$NOC, y=No\_medals,fill=Medal)) +  
 geom\_col() +  
 coord\_flip() +  
 scale\_fill\_manual(values=c("gold4","gold","gray70")) +  
 theme(  
   
 axis.text.y = element\_text(size=3,angle=3),plot.title = element\_text(face = "bold",colour = "blue",size = 12)) +theme\_bw()+  
   
 labs(title = "Accolades achieved by various countries at London 2012 Olympics ",x=" ",y="Countries")+coord\_polar(theta = "x")

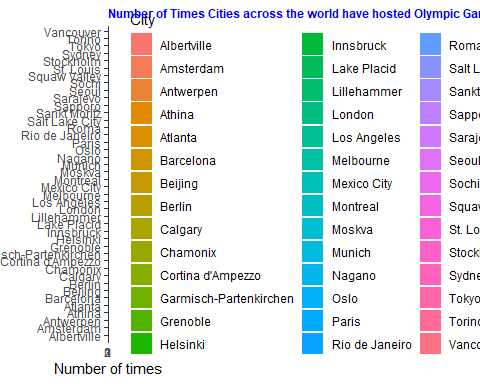
## Coordinate system already present. Adding new coordinate system, which will replace the existing one.

Medal\_plot



## 5.11 Cities that have hosted the Olympics:

Host\_City\_Times <- olympics\_dataframe%>%group\_by(City,Games)%>%summarise(Host\_City\_Count = n\_distinct(City))%>%group\_by(City,Host\_City\_Count)%>%summarise(Number\_Olympics\_Held =sum(Host\_City\_Count))  
  
Host\_Allcities <- ggplot(Host\_City\_Times,aes(x=City,y=Number\_Olympics\_Held,fill=City))+geom\_col()+coord\_flip()+ theme\_bw()+  
 theme(plot.title = element\_text(face = "bold",colour = "blue",size = 9))+  
 labs(title = "Number of Times Cities across the world have hosted Olympic Games ",x="Cities which have hosted olympics",y="Number of times")  
Host\_Allcities

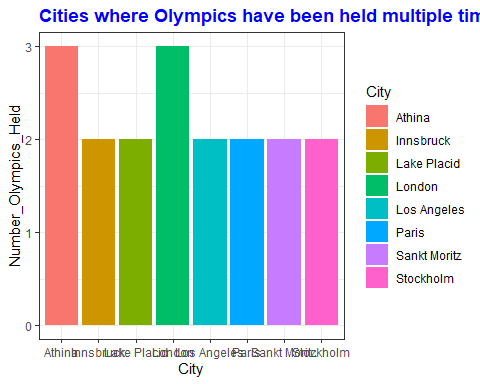


## Cities around the world that have bid and won to host olympics more than once

Host\_City <- olympics\_dataframe%>%group\_by(City,Games)%>%summarise(Host\_City\_Count = n\_distinct(City))%>%group\_by(City,Host\_City\_Count)%>%summarise(Number\_Olympics\_Held =sum(Host\_City\_Count))%>%filter(Number\_Olympics\_Held>1)  
Host\_City

## # A tibble: 8 x 3  
## # Groups: City [8]  
## City Host\_City\_Count Number\_Olympics\_Held  
## <chr> <int> <int>  
## 1 Athina 1 3  
## 2 Innsbruck 1 2  
## 3 Lake Placid 1 2  
## 4 London 1 3  
## 5 Los Angeles 1 2  
## 6 Paris 1 2  
## 7 Sankt Moritz 1 2  
## 8 Stockholm 1 2

Host\_City\_plot <- ggplot(Host\_City,aes(x=City,y=Number\_Olympics\_Held,fill=City))+geom\_col()+ theme\_bw()+  
 theme(plot.title = element\_text(face = "bold",colour = "blue"))+  
 labs(title = "Cities where Olympics have been held multiple times ")  
Host\_City\_plot



## 5.12 5 Trending Summer and Winter Olympic games before 1995:

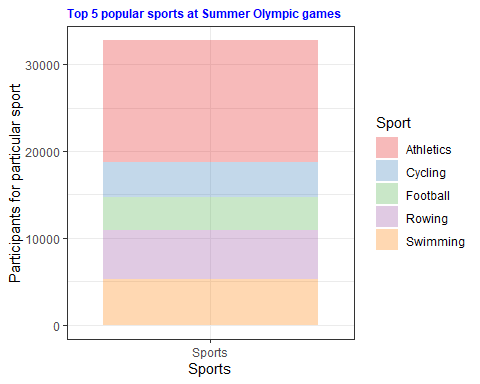
#Which are the top 5 popular sports in Summer and Winter Olympics in terms of participation for the sport from time period 1896 to 1994:  
  
olympics\_sport\_atheletes <- olympics\_dataframe%>%filter(Season=="Summer", between(Year,1896,1994))%>%group\_by(Sport)%>%summarise(Athletes\_Per\_Sport = n\_distinct(Name))%>%mutate(pop\_sport = min\_rank(desc(Athletes\_Per\_Sport)))%>%filter(pop\_sport %in% c("1":"5"))%>%arrange(pop\_sport)  
  
olympics\_sport\_atheletes

## # A tibble: 5 x 3  
## Sport Athletes\_Per\_Sport pop\_sport  
## <chr> <int> <int>  
## 1 Athletics 14032 1  
## 2 Rowing 5640 2  
## 3 Swimming 5220 3  
## 4 Cycling 3990 4  
## 5 Football 3869 5

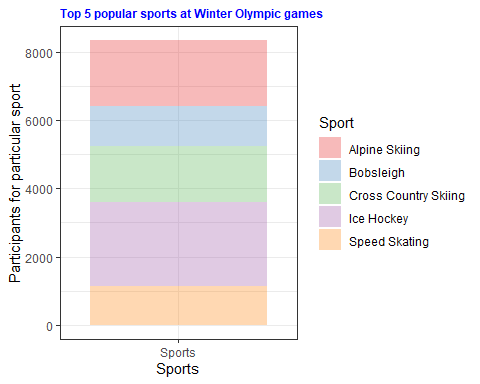
olympics\_sport\_atheletes\_Winter <- olympics\_dataframe%>%filter(Season=="Winter", between(Year,1896,1994))%>%group\_by(Sport)%>%summarise(Athletes\_Per\_Sport = n\_distinct(Name))%>%mutate(pop\_sport = min\_rank(desc(Athletes\_Per\_Sport)))%>%filter(pop\_sport %in% c("1":"5"))%>%arrange(pop\_sport)  
  
olympics\_sport\_atheletes\_Winter

## # A tibble: 5 x 3  
## Sport Athletes\_Per\_Sport pop\_sport  
## <chr> <int> <int>  
## 1 Ice Hockey 2469 1  
## 2 Alpine Skiing 1925 2  
## 3 Cross Country Skiing 1650 3  
## 4 Bobsleigh 1167 4  
## 5 Speed Skating 1132 5

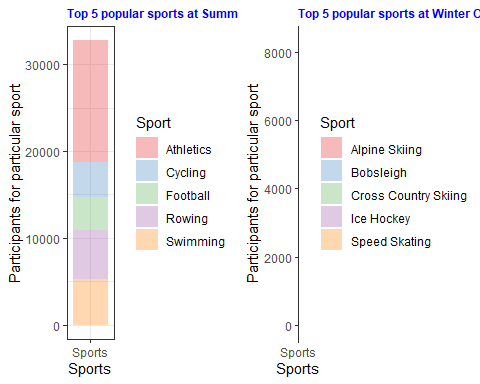
Sport\_plot <- olympics\_sport\_atheletes%>%ggplot(aes(x="Sports",y=Athletes\_Per\_Sport,fill=Sport))+geom\_col(alpha=0.3,aes(fill= Sport))+scale\_fill\_brewer(palette = "Set1")+theme\_bw()+  
 theme(plot.title = element\_text(face = "bold",colour = "blue",size=9))+  
 labs(title = "Top 5 popular sports at Summer Olympic games",x="Sports",y="Participants for particular sport")  
Sport\_plot



Sport\_plot\_Winter <- olympics\_sport\_atheletes\_Winter%>%ggplot(aes(x="Sports",y=Athletes\_Per\_Sport,fill=Sport))+geom\_col(alpha=0.3,aes(fill= Sport))+scale\_fill\_brewer(palette = "Set1")+theme\_bw()+  
 theme(plot.title = element\_text(face = "bold",colour = "blue",size=9))+  
 labs(title = "Top 5 popular sports at Winter Olympic games",x="Sports",y="Participants for particular sport")  
Sport\_plot\_Winter



grid.arrange(Sport\_plot,Sport\_plot\_Winter,ncol=2)



## 5.13 Trending Summer and Winter Olympic games from 1995 to recent Rio Olympics

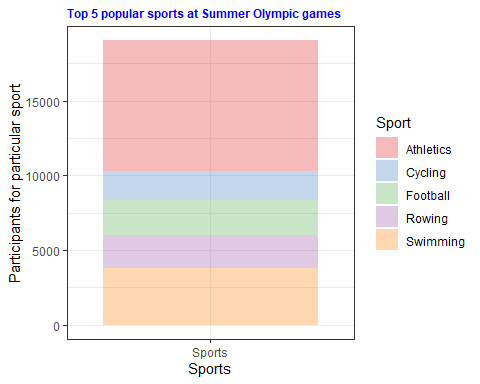
#Which are the top 5 popular sports in Summer and Winter Olympics in terms of participation for the sport from time period 1995 to 2016  
  
olympics\_sport\_atheletes\_recent <- olympics\_dataframe%>%filter(Season=="Summer", between(Year,1995,2016))%>%group\_by(Sport)%>%summarise(Athletes\_Per\_Sport = n\_distinct(Name))%>%mutate(pop\_sport = min\_rank(desc(Athletes\_Per\_Sport)))%>%filter(pop\_sport %in% c("1":"5"))%>%arrange(pop\_sport)  
  
olympics\_sport\_atheletes

## # A tibble: 5 x 3  
## Sport Athletes\_Per\_Sport pop\_sport  
## <chr> <int> <int>  
## 1 Athletics 14032 1  
## 2 Rowing 5640 2  
## 3 Swimming 5220 3  
## 4 Cycling 3990 4  
## 5 Football 3869 5

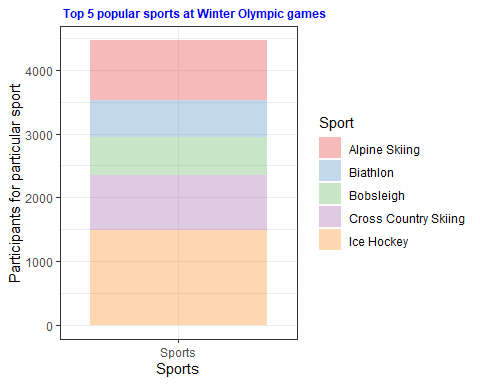
olympics\_sport\_atheletes\_Winter\_recent <- olympics\_dataframe%>%filter(Season=="Winter", between(Year,1995,2016))%>%group\_by(Sport)%>%summarise(Athletes\_Per\_Sport = n\_distinct(Name))%>%mutate(pop\_sport = min\_rank(desc(Athletes\_Per\_Sport)))%>%filter(pop\_sport %in% c("1":"5"))%>%arrange(pop\_sport)  
  
olympics\_sport\_atheletes\_Winter\_recent

## # A tibble: 5 x 3  
## Sport Athletes\_Per\_Sport pop\_sport  
## <chr> <int> <int>  
## 1 Ice Hockey 1479 1  
## 2 Alpine Skiing 935 2  
## 3 Cross Country Skiing 866 3  
## 4 Bobsleigh 603 4  
## 5 Biathlon 580 5

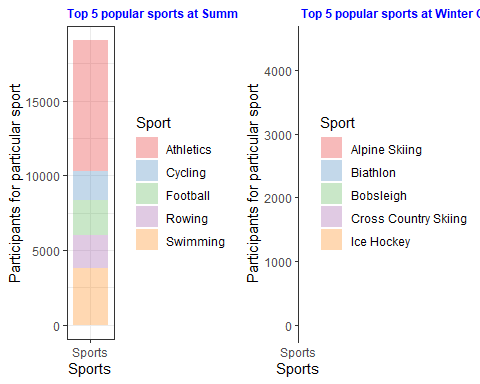
Sport\_plot\_recent <- olympics\_sport\_atheletes\_recent%>%ggplot(aes(x="Sports",y=Athletes\_Per\_Sport,fill=Sport))+geom\_col(alpha=0.3,aes(fill= Sport))+scale\_fill\_brewer(palette = "Set1")+theme\_bw()+  
 theme(plot.title = element\_text(face = "bold",colour = "blue",size=9))+  
 labs(title = "Top 5 popular sports at Summer Olympic games",x="Sports",y="Participants for particular sport")  
Sport\_plot\_recent



Sport\_plot\_Winter\_recent <- olympics\_sport\_atheletes\_Winter\_recent%>%ggplot(aes(x="Sports",y=Athletes\_Per\_Sport,fill=Sport))+geom\_col(alpha=0.3,aes(fill= Sport))+scale\_fill\_brewer(palette = "Set1")+theme\_bw()+  
 theme(plot.title = element\_text(face = "bold",colour = "blue",size=9))+  
 labs(title = " Top 5 popular sports at Winter Olympic games",x="Sports",y="Participants for particular sport")  
Sport\_plot\_Winter\_recent



grid.arrange(Sport\_plot\_recent,Sport\_plot\_Winter\_recent,ncol=2)



# 6. Insights:

1. Over the years, participants at Olympics are increasing except for few years in history.
2. Country representatives for Summer Olympics are more than twice as compared to Winter Olympics due to lesser Olympic events held during Winter Olympics.
3. Participation of Women now and also in coming years will be on similar level as Men, which was not seen earlier.
4. Age plays a vital role in becoming an Olympic medalist.
5. The USA and European Nations dominate on accolades achieved.
6. Change in popular sports at Olympic events over time.
7. Height of athltes seems to be a factor contributing to winning formula at Olympics especially among women participants.

# 7. References:

En.wikipedia.org. (2020). Olympic Games. [online] Available at: <https://en.wikipedia.org/wiki/Olympic_Games>

Kaggle.com. (2020). Datasets | Kaggle. [online] Available at: <https://www.kaggle.com/datasets>

International Olympic Committee. (2020). Olympics | Olympic Games, Medals, Results, News | IOC. [online] Available at: <https://www.olympic.org/>