CHAITANYA BHARTHI INSTITUTE OF TECHNOLOGY(AUTONOMOUS),HYDERABAD

Introduction to Inference and Interpretation – Final Report

TOPIC:

DATA ANALYSIS AND VISUALIZATION OF OLYMPICS

Department of Computer Science & Engineering

BE (AI&ML) III Semester

Introduction to Inference and Interpretation – Final Report

Team No: 04

S.no.	Roll No.	Name	Role	Remark
1	160121729040	K Manoj Kumar	Team Leader	
2	160121729041	K. SAITEJA	Member	
3	160121729042	K Nithin	Member	
4	160121729047	M Naresh	Member	
5	160121729049	P Guru Prasad	Member	

AIM: To analyse and visualize the Olympics data and check any gender discrimination is present in Olympics.

ABSTRACT:

This is all about analysis and data visualization of the Olympics data. Analysis through R language is very easy and user-friendly when compared to other platforms. Here we mainly focused upon medals won by different countries at different venues. Here we also focused on participants at different venues, and also compared number of participants by gender. Based on the difference in the number of participants in at Olympics we made some conclusions. Visualisation of the data over various factors will provide us with the statistical view of the various factors which lead to the evolution of the Olympic Games and Improvement in the performance of various Countries/Players over time. We also compared India performance over the year and visualized the data. We also compared India with other top comparative countries.

PROBLEM DESINATION:

- **Visualizations of number of participants in Olympics over the years.**
- **Comparing different countries based on medals won at different venues.**
- **Comparing total number of participants by gender over the years.**
- **Analysing growth in total number of participants.**
- **Comparing India performance with other countries.**
- **Finding highest medal won by individual at different venue.**

Introduction:

The modern Olympic Games or Olympics are leading international sporting events featuring summer and winter sports competitions in which thousands of athletes from around the world participate in a variety of competitions. The Olympic Games are considered the world's foremost sports competition with more than 200 nations participating. The Olympic Games are normally held every four years, alternating between the Summer and Winter Olympics every two years in the four years.

Various scenarios come to our mind when we look into the Evolution of the Olympic Games over the years. These scenarios are: Increase in the number of participating nations, increase in the number of participating Athletes, Increase/Decrease in the number of events, increase in the expenditure cost of the event, improvement in the performance of the particular country, improvement in the performance of a particular player, Increase in women participation, Participation Ratio of Men to Women. This analysis would help in future prediction.

Information:

- **Lubridate**: provides tools that make it easier to parse and manipulate dates. These tools are grouped below by common purpose. More information about each function can be found in its help documentation.
- **Country code:** Converts long country names into one of many different coding schemes. Translates from one scheme to another. Converts country name or coding scheme to the official short English country name. Creates a new variable with the name of the continent or region to which each country belongs.
- **Highchart**: is a mature javascript charting library. Highcharts provide a various type of charts, from scatters to heatmaps or treemaps.
- **Direct labels:** Add direct labels to a plot, and hide the color legend. Modern plotting packages like lattice and ggplot2 show automatic legends based on the variable specified for color, but these legends can be confusing if there are too many colors. Direct labels are a useful and clear alternative to a confusing legend in many common plots.
- **geom_path()** connects the observations in the order in which they appear in the data. **geom_line()** connects them in order of the variable on the x axis. **geom_step()** creates a stairstep plot, highlighting exactly when changes occur. The group aesthetic determines which cases are connected together.
- **facet_wrap()** wraps a 1d sequence of panels into 2d. This is generally a better use of screen space than **facet_grid()** because most displays are roughly rectangular.
- scale_x_discrete() and scale_y_discrete() are used to set the values for discrete x and y scale aesthetics. For simple manipulation of scale labels and limits, you may wish to use labs() and lims() instead.
- mutate() creates new columns that are functions of existing variables. It can also modify (if the name is the same as an existing column) and delete columns (by setting their value to NULL).
- Hcboxplot(): Shortcut for Box plot.

Analysis and Visualization:

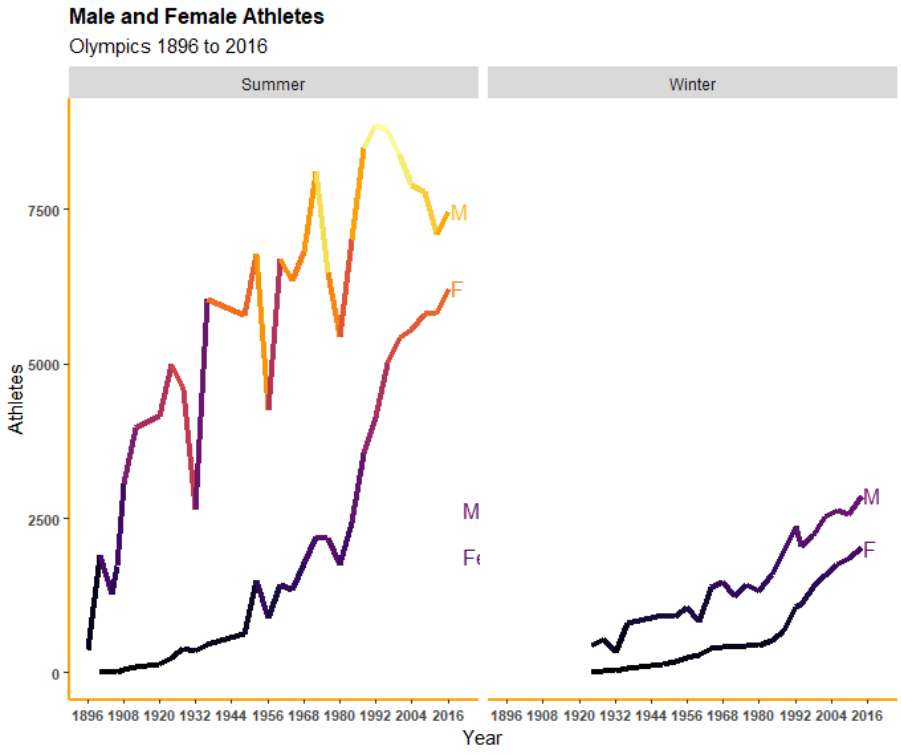
1. Number of athletics over at each venue:

In this we plot the graph between number of athletics participated in the Olympics at a particular venue.

Code:

```
ggplot(season, aes(x=Year, y=total,colour=total,group=Sex))+
        geom_line(size=1.5)+
        geom_dl(aes(label = Sex), method = list(dl.trans(x = x + .05), "last.points")) +
        facet_wrap(~Season)+
        scale_color_viridis(option = "B")+
        scale_x_continuous(breaks = seq(1896, 2016, by = 12))+
        xlab("Year")+ylab("Athletes")+
        theme(axis.line =element_line(color=
"orange",size=1))+theme(panel.background=element_blank())+
        theme(legend.position = "none",
            axis.text = element_text(size = 8,face="bold"),
            plot.title = element_text(size=12,face = "bold")) +
        ggtitle("Male and Female Athletes",subtitle = "Olympics 1896 to 2016 ")
```

Graph:

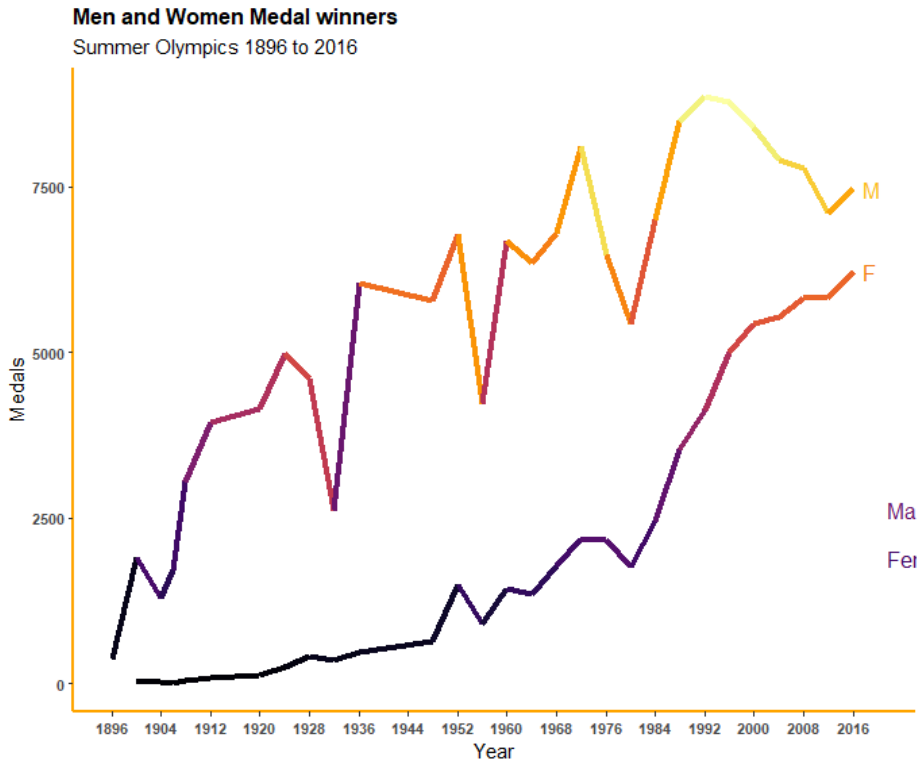


2. Number of medals won by men and women at different venue

In this we plot the graph between medals won by men and women at different venues of the Olympics.

Code:

```
gender = data_events%>%filter(!is.na(Medal),Season=="Summer")
        gender =gender%>%group_by(Year,Sex)%>%summarize(total=n())
       ggplot(gender, aes(x=Year, y=total,colour=total,group=Sex))+
        geom_line(size=1.5)+
        geom_dl(aes(label = Sex), method = list(dl.trans(x = x + .2),
"last.points")) +
        scale_color_viridis(option = "B")+
        scale_x_continuous(breaks = seq(1896, 2016, by = 8))+
        xlab("Year")+ylab("Medals")+
        theme(axis.line = element_line(color =
"orange",size=1))+theme(panel.background=element_blank())+
        theme(legend.position = "none",
            axis.text = element_text(size = 8,face="bold"),
            plot.title = element_text(size=12,face = "bold")) +
        ggtitle("Men and Women Medal winners ",subtitle = "Summer
Olympics 1896 to 2016 ")
```

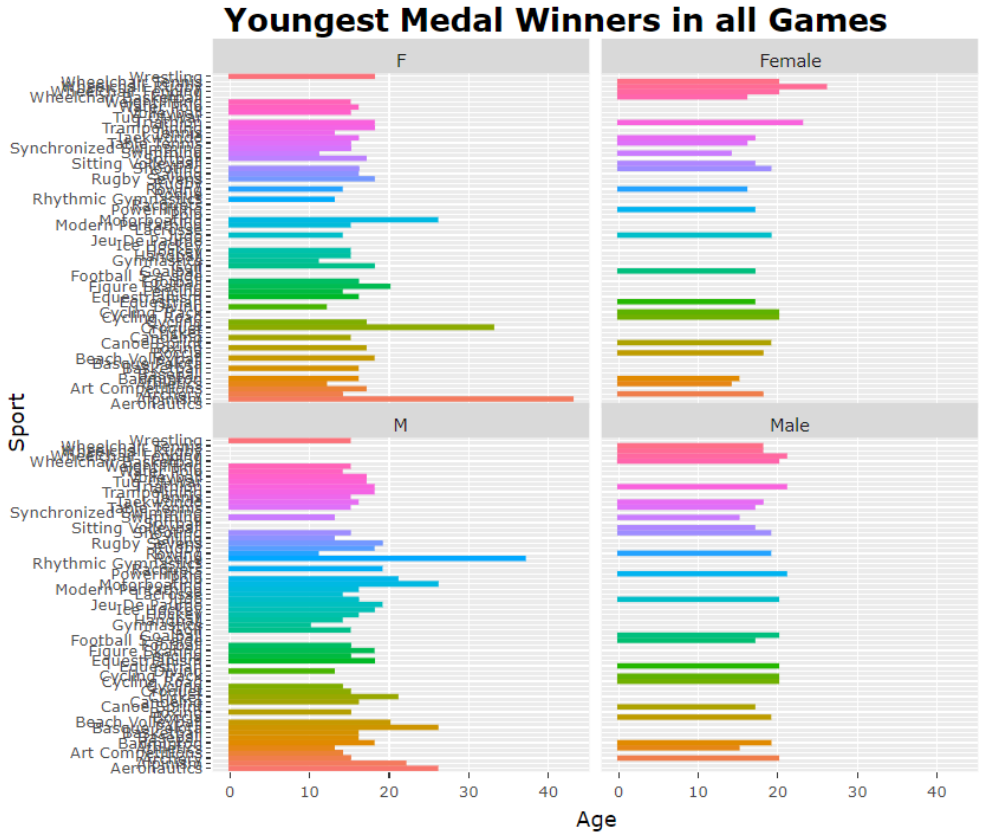


3. **Finding youngest medallist at every sport:**

This analysis is used to find who is the youngest athletic to won a gold medal at different sport.

Code:

```
age_mi <- data_events%>%filter(!is.na(Medal),Season=='Summer')
      age_mi <-
age_mi%>%group_by(Sex,Sport)%>%summarize(Age=min(Age,na.rm =
TRUE))
age_min <-
data_events%>%filter(!is.na(Medal),Season=="Summer")%>%right_join(age
_mi,by=c("Sex","Sport","Age"))
  c <-ggplot(age_min,aes(Sport,Age, color=Sport,fill=Name)) +
       geom_bar(position = "dodge",  width =.5,stat="identity") +
       coord_flip()+
       facet_wrap(~Sex)+
       theme_grey() +
       scale_x_discrete() +
       xlab("Sport")+ylab("Age")+
       theme(legend.position = "none",
            axis.text = element_text(size = 8,face="bold"),
            plot.title = element_text(size=16,face = "bold")) +
       ggtitle("Youngest Medal Winners in all Games")
ggplotly(c)
```



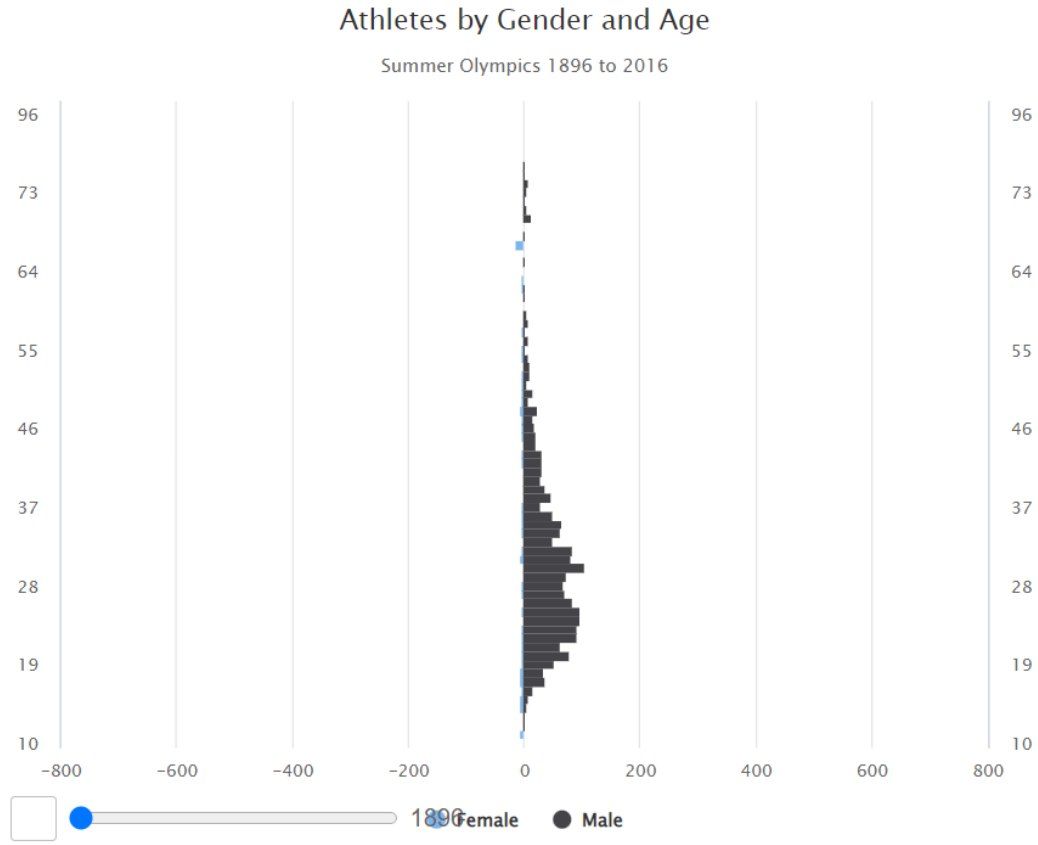
4. Analysis of gender and age:

Here we plot an animation containing frequency of age of an athletic at an venue.

Code:

```
age_mi <- data_events%>%filter(!is.na(Medal),Season=='Summer')
      age_mi <-age_mi%>%group_by(Sex,Sport)%>%summarize(Age=min(Age,na.rm =
TRUE))
      age_min <-
data_events%>%filter(!is.na(Medal),Season=="Summer")%>%right_join(age_mi,by=c("Sex
","Sport","Age"))
       c <-ggplot(age_min,aes(Sport,Age, color=Sport,fill=Name)) +
        geom_bar(position = "dodge",  width =.5,stat="identity") +
        coord_flip()+
        facet_wrap(~Sex)+
        theme_grey() +
        scale_x_discrete() +
        xlab("Sport")+ylab("Age")+
        theme(legend.position = "none",
              axis.text = element_text(size = 8,face="bold"),
              plot.title = element_text(size=16,face = "bold")) +
        ggtitle("Youngest Medal Winners in all Games")

      ggplotly(c)
      df <-
data_events%>%filter(!is.na(Age),Season=='Summer')%>%group_by(Sex,Age,Year)%>%su
mmarize(pop=n())
      df$Sex <- ifelse(df$Sex=="F","Female","Male")

      df <- df%>%
        mutate(athletes = pop*ifelse(Sex == "Female", -1, 1))

      series <- df %>%
        group_by(Sex, Age)%>%
        do(data = list(sequence = .$athletes)) %>%
        ungroup() %>%
        group_by(Sex) %>%
        do(data = .$data) %>%
        mutate(name = Sex)%>%
        list_parse()

      maxpop <- max(abs(df$athletes))

      xaxis <- list(categories = sort(unique(df$Age)),
                    reversed = FALSE, tickInterval = 3,
                    labels = list(step= 3))

      yrs <-  sort(unique(df$Year))

      highchart() %>%
        hc_chart(type = "bar") %>%
        hc_motion(enabled = TRUE, labels =yrs, series = c(0,1), autoplay = TRUE,
updateInterval = 4) %>%
        hc_add_series_list(series) %>%
        hc_plotOptions(
```

```
  series = list(stacking = "normal"),
  bar = list(groupPadding = 0, pointPadding =  0, borderWidth = 0)
 ) %>%
 hc_tooltip(shared = TRUE) %>%
 hc_yAxis(
  min=-800,max=800)%>%
 hc_xAxis(
  xaxis,
  rlist::list.merge(xaxis, list(opposite = TRUE, linkedTo = 0))
 ) %>%
 hc_tooltip(shared = FALSE,
        formatter = JS("function () { return '<b>' + this.series.name + ', Age ' +
this.point.category + '</b><br/>' + 'athletes: ' +
Highcharts.numberFormat(Math.abs(this.point.y), 0);}")
 ) %>%
 hc_title(text = " Athletes by Gender and Age") %>%
 hc_subtitle(text = "Summer Olympics 1896 to 2016")
```



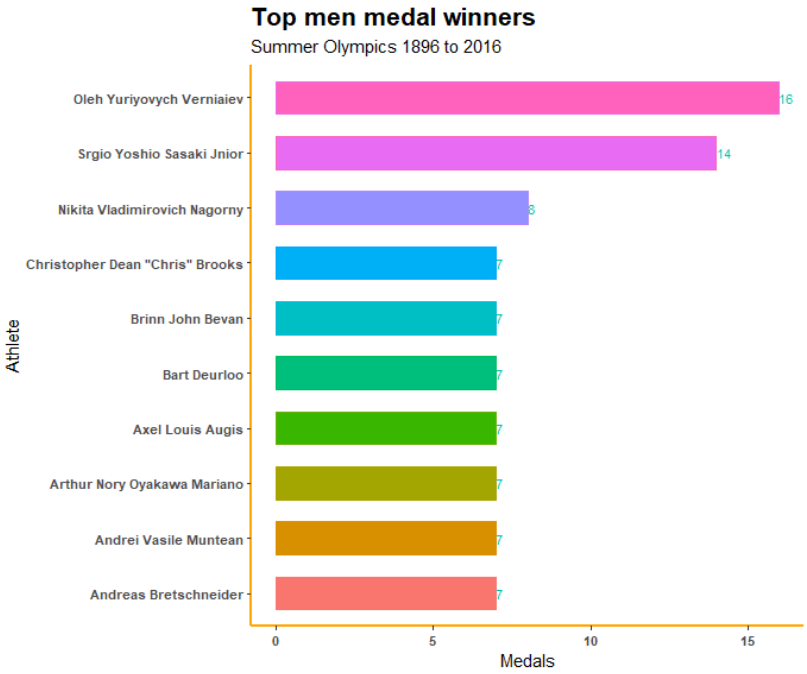
5. **Graph of highest medal won by athletics at each venue:**

Here we will find which athletic has won the most medals at a venue.

Code:

```
athlete = athlete%>%left_join(data_team,by=NULL)
      athlete
=athlete%>%group_by(Sex,Name)%>%summarize(total=n())%>%arrange(desc(total))
      women =athlete%>%filter(Sex=="F")%>%head(n=30)
      men =athlete%>%filter(Sex=='M')%>%head(n=30)
       men$Name <- factor(men$Name,levels = men$Name[order(men$total)])
     ggplot(men,aes(Name,total,color=Name,fill=Name)) +
```

```
geom_bar(position = "stack",  width =.6,stat="identity") +
    coord_flip()+
    geom_text(aes(label=total,hjust=-.03,  colour="black"),size=3)+
     theme(axis.line = element_line(color = "orange",size=1))+
    theme(panel.background=element_blank())+
    scale_x_discrete() +
    xlab("Athlete")+ylab("Medals")+
    theme(legend.position = "none",
          axis.text = element_text(size = 8,face="bold"),
          plot.title = element_text(size=16,face = "bold")) +
    ggtitle("Top men medal winners " ,subtitle = "Summer Olympics 1896 to 2016")
```

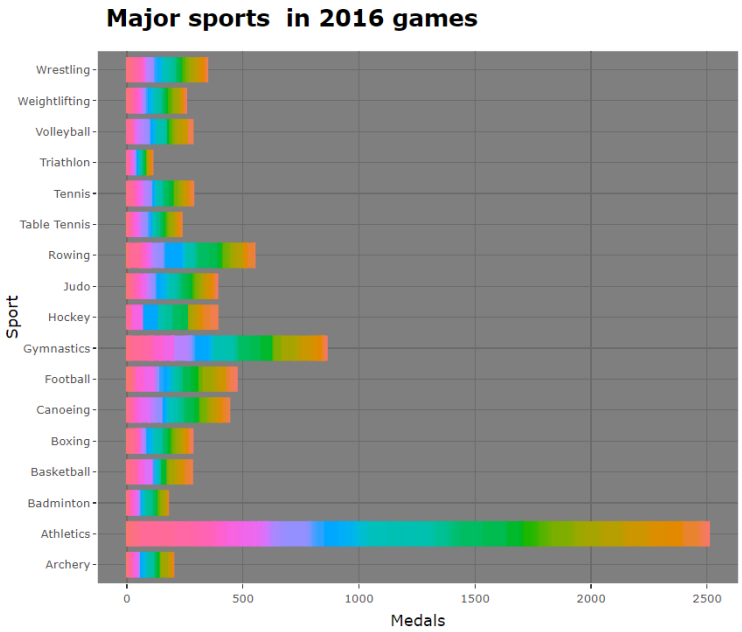


6. Graph of major sport at 2016 venue:

This graph represents the major sport at the 2016 Rio Olympics venue.

Code:

```
slist
=c('Archery','Athletics','Badminton','Baseball','Basketball','Boxing','Canoeing','Football','Gym
nastics','Hockey','Judo','Rowing',' Shooting','Swimming ','Table
Tennis','Tennis','Triathlon','Weightlifting','Wrestling','Volleyball')
    sport <- filter(sport,Sport%in%slist,total>=1)
    sport$Team <- as.factor(sport$Team)

    p<-ggplot(sport,aes(Sport,total,color=Team,fill=Team)) +
     geom_bar(position = "stack",  width =.75,stat="identity") +
     coord_flip()+
     theme_dark() +
     scale_x_discrete() +
     xlab("Sport")+ylab("Medals")+
     theme(legend.position = "none",
          axis.text = element_text(size = 8,face="bold"),
          plot.title = element_text(size=16,face = "bold")) +
     ggtitle("Major sports  in 2016 games")

    ggplotly(p)
```

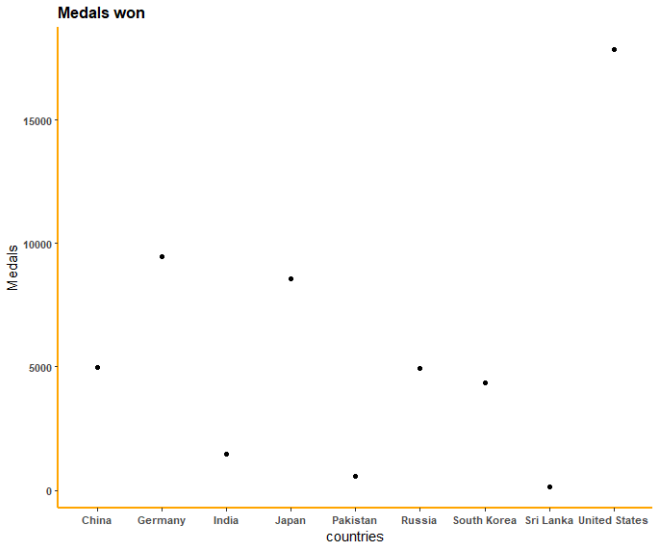


7. Comparing India with other countries:

In this we compare medals won by India and Other countries and compare them.

Code:

```
k<-count_medal
k<-filter(k,(Team=='India'|Team=='United
States'|Team=='China'|Team=='Russia'|Team=='Germany'|Team=='Japan'|Team=='Pakistan'|
Team=='Sri Lanka'|Team=='South Korea'))
ggplot(data=k,aes(x=Team,y=Medals))+
  geom_point() +
  scale_color_viridis(option = "B")+
  xlab("contries")+ylab("Medals")+
  theme(axis.line = element_line(color =
"orange",size=1))+theme(panel.background=element_blank())+
  theme(legend.position = "none",
        axis.text = element_text(size = 8,face="bold"),
        plot.title = element_text(size=12,face = "bold")) +
 ggtitle("Medals won")
```



Conclusion and Inference:

- R has its own significant for its vast and advanced packages and functions.
- When compared to other platforms working with R is much more user friendly and easy to learn.
- Visualization and Analyze through R is very easy and effective.
- We were able to learn new functions and packages of R very easily.

- We have concluded that there is a discrimination towards women in Olympics.
- But over the years growth in the women athletics is much more then growth in men athletics.
- From that we can conclude that female athletics are also encouraged to participate in Olympics in modern times.
- We have also concluded that only a few countries are continuously winning major medals in the Olympics.
- And we have also concluded that despite of its vast population India is far more behind in Olympics medals and participants , when compared to some top and neighboring countries.