ST. Clair College

PROJECT 1

Group No. 1

Yash P Mecwan (0734905)
Ronak R Prajapati (0734905)

Basic Stats & EDA (DAB501)

Undertaking

We understand St. Clair College's Academic Integrity policies and consequences of plagiarism. It is our assurance that nothing presented in this project is plagiarised. And any resource and material used for assistance will mentioned in references. We understand the fact that copying, and cheating might lead harsh measures, and also fathom that allowing someone else copy our work is also a serious offence and thus we state that we haven't allowed anyone use our work and neither have we copied someone else's work or project.

Contents

Resources & Documentation

Software & Packages

- r version 3.5.2
- r studio version desktop open source edition
- ggplot2
- plotly

Data Set Documentation

- Creators: Farhan Ahmed, SPScientist, www.sports-reference.com
- Data frame “a”: data about athlete’s events

  https://www.kaggle.com/itsfarhan/athletes-events-datasets

- Data frame “b”: data about student performance

  https://www.kaggle.com/spscientist/students-performance-in-exams

- Summary StudentPerformance:It contains marks of students in various subjects in a school. The data set has classifying column and also continuous column.
- Summary athlets-events : This dataset has data on modern Olympics games starting from Athens 1896 leading up to Rio 2016.

Documentation on changes done on Dataset

- In student performance dataset for getting efficient output two new columns added which are total score which shows total score of all three scores given in dataset and percentage which are calculated based on total score.

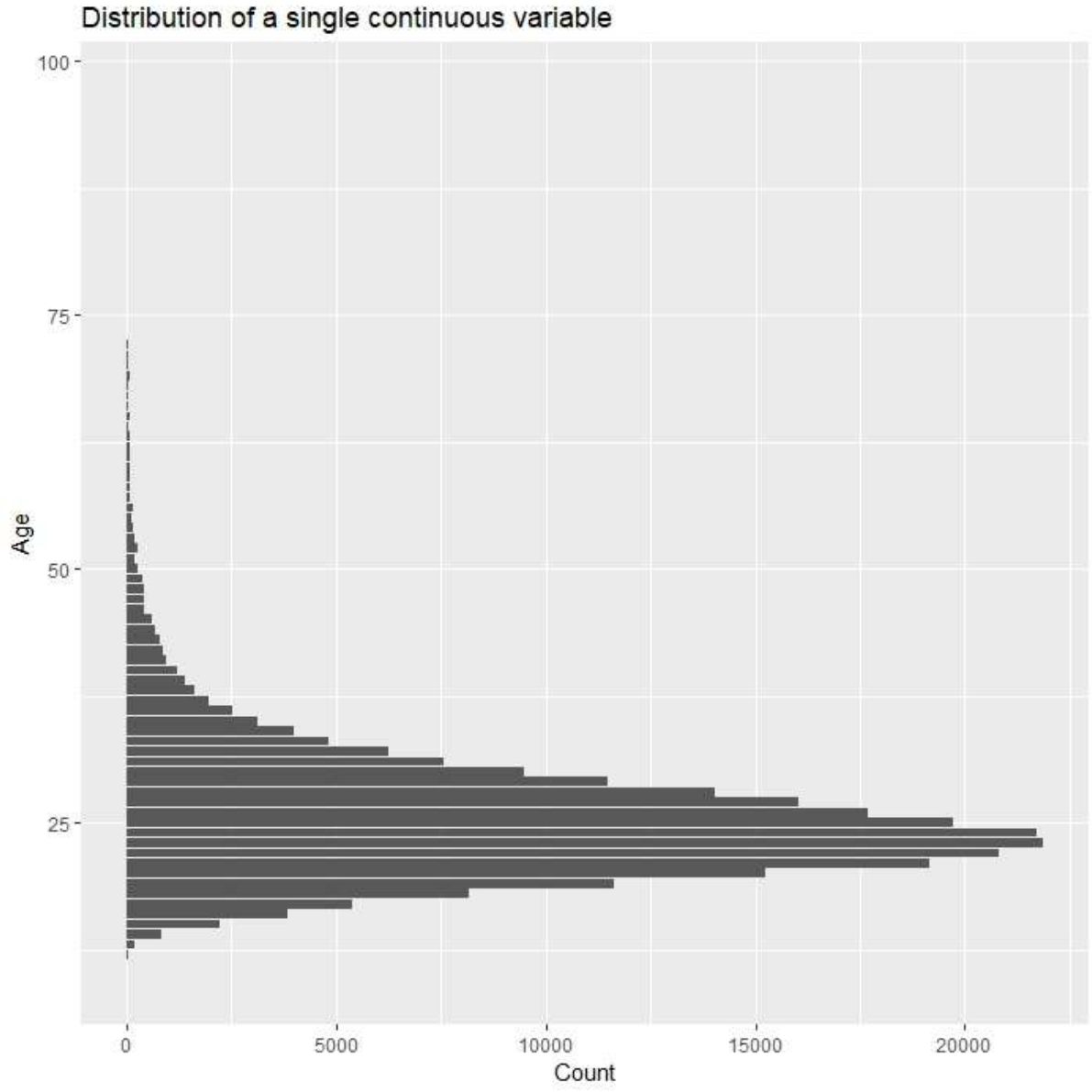
Plots

Plot 1a

Code:

```
ggplot(data = athlete_events) +
geom_bar(mapping = aes(x = Age)) +
coord_flip()+labs(title='Distribution of a single continuous
variable',x='Age',y= 'Count')
```

Output:

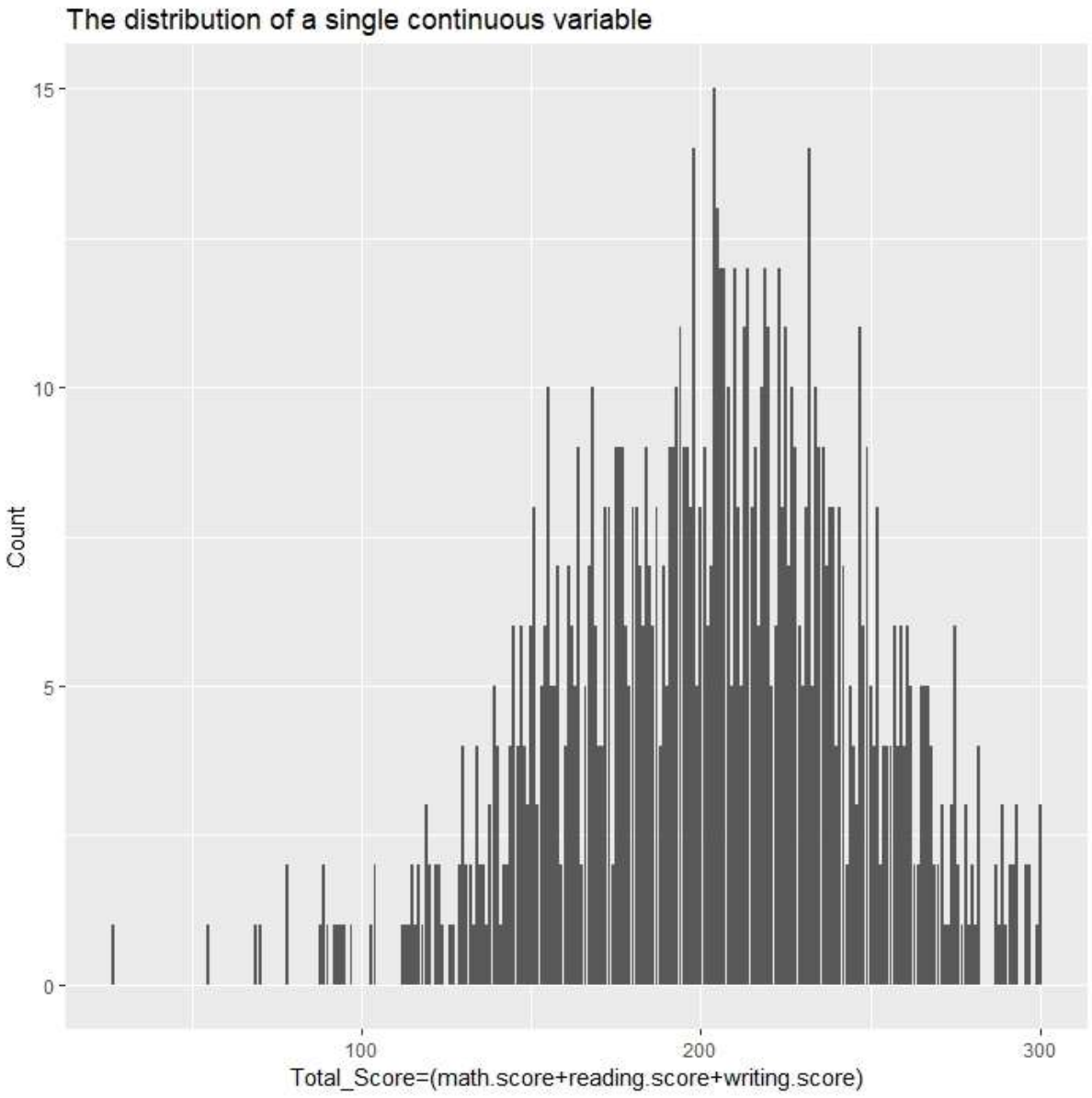


Plot 1b

Code:

- StudentsPerformance$total.score <-(StudentsPerformance$math.score + StudentsPerformance$reading.score + StudentsPerformance$writing.score)
- ggplot(StudentsPerformance,aes(total.score))+ geom_bar()+labs(title='The distribution of a single continuous variable',x='Total_Score=(math.score+reading.score+writing.score)', y='Count')

Output:

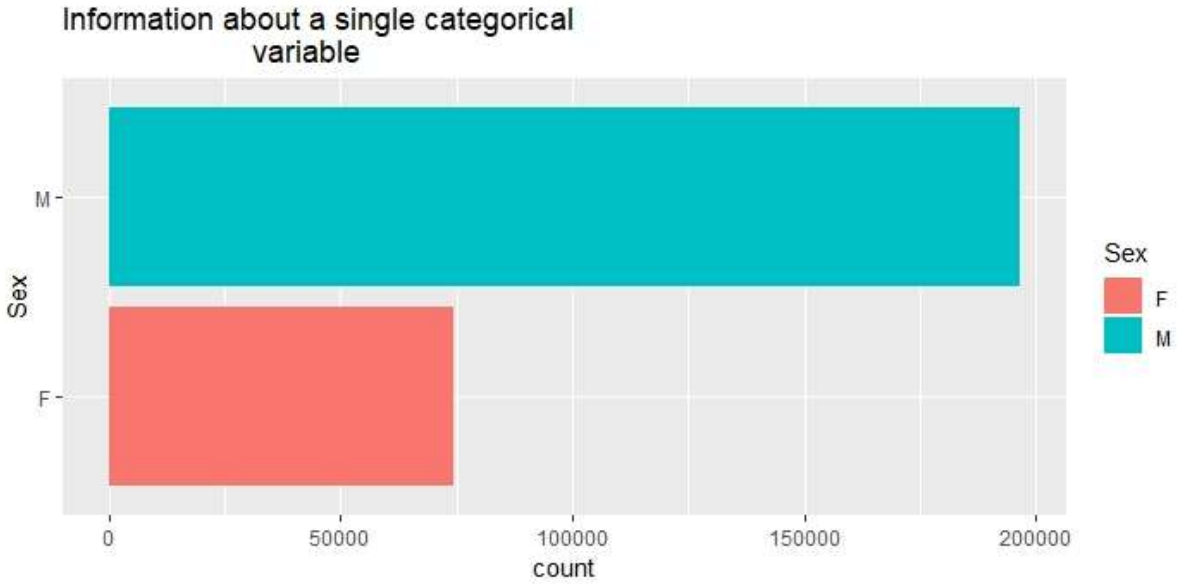


Plot 2a

Code:

```
ggplot(data = athlete_events) +
geom_bar(mapping = aes(x = Sex, fill = Sex))+
coord_flip()+labs(title='Information about a single categorical
variable',x='Sex',y='count')
```

Output:

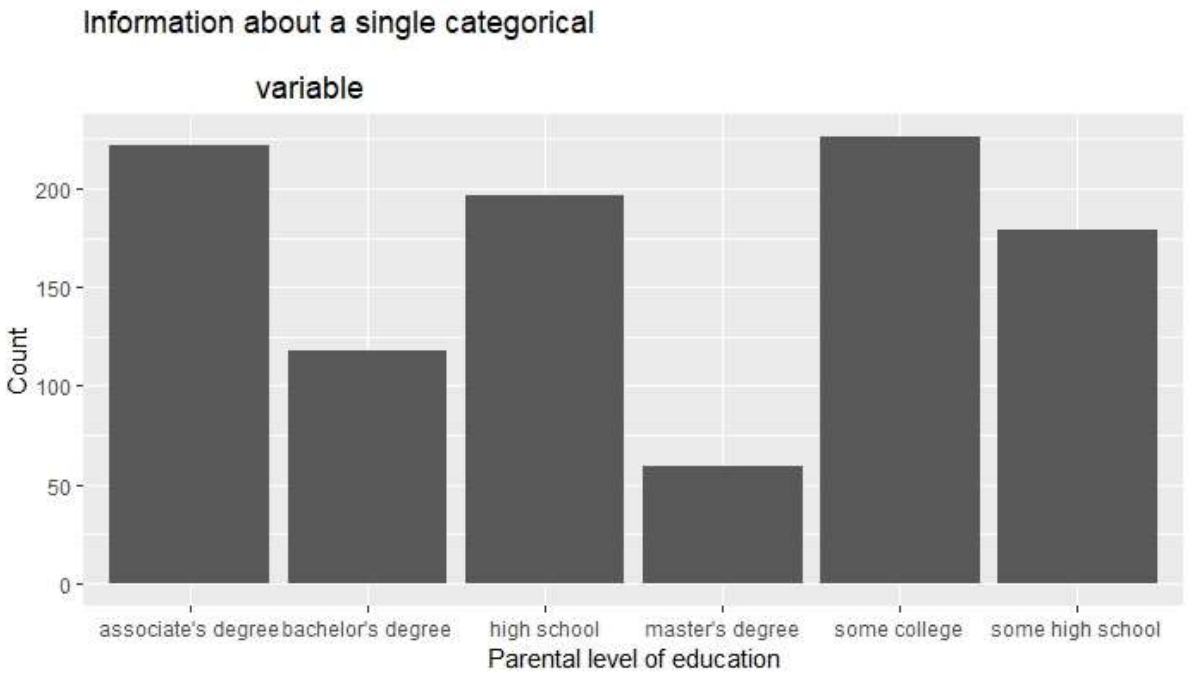


Plot 2b

Code:

```
ggplot(StudentsPerformance,aes(parental.level.of.education))+
geom_bar()+labs(title = 'Information about a single categorical
variable',x='Parental level of education',y='Count')
```

Output:

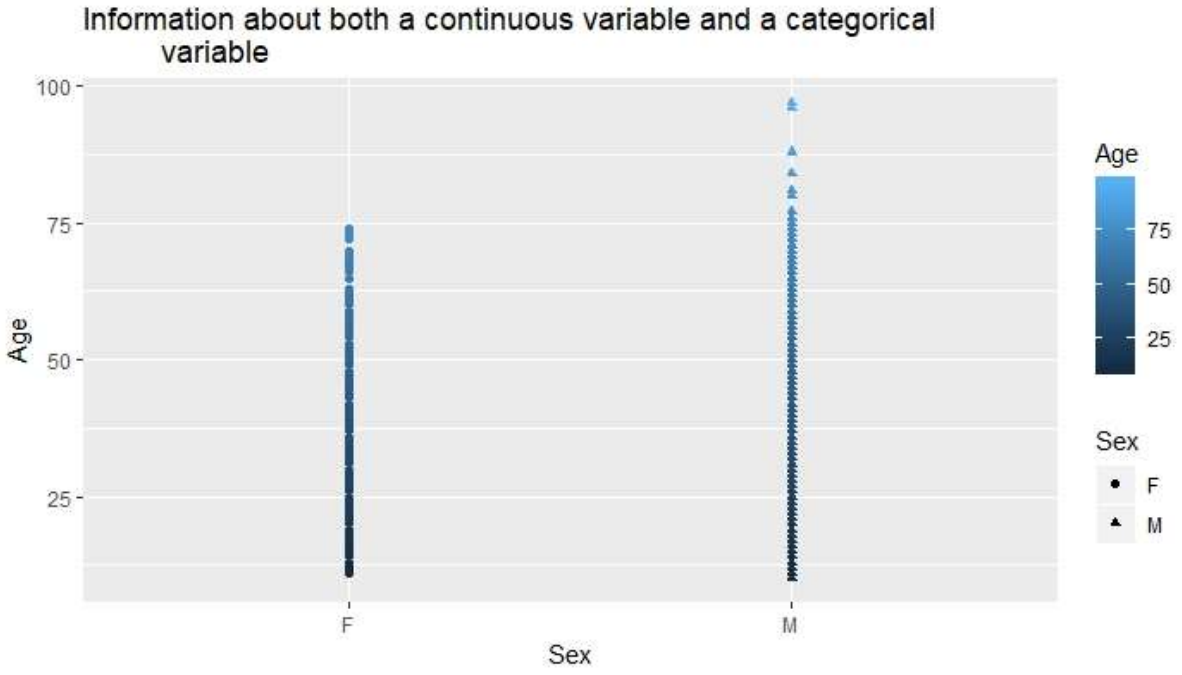


Plot 3a

Code:

```
ggplot(data = athlete_events) +
geom_point(mapping = aes(x = Sex, y = Age, color = Age,
shape = Sex))+
labs(title = 'Information about both a continuous variable and a categorical
variable',x= 'Sex',y='Age')
```

Output:

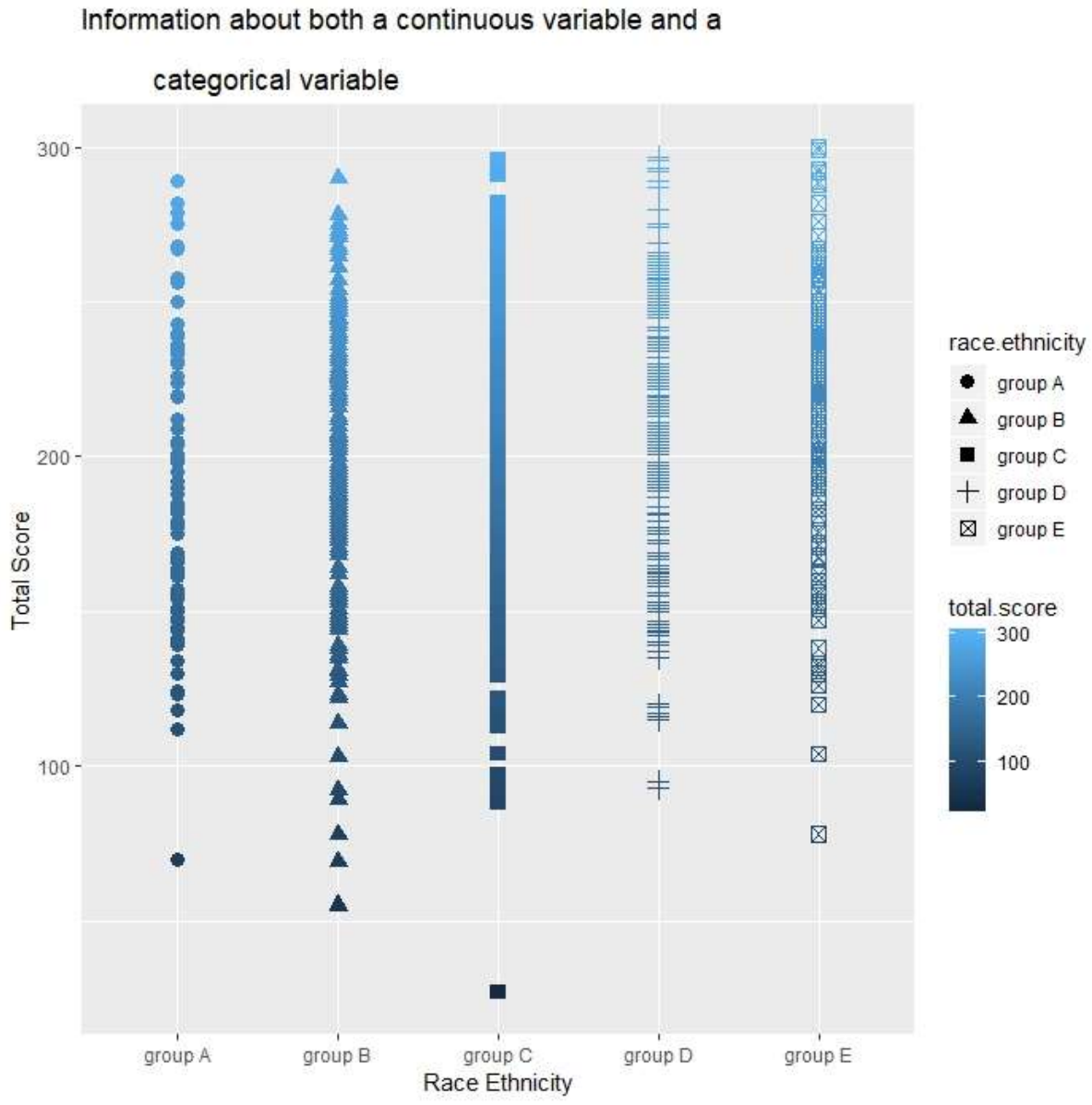


Plot 3b

Code:

```
ggplot(StudentsPerformance,aes(x=race.ethnicity,y=total.score,
color = total.score,shape=race.ethnicity))+
geom_point(size= 3)+
labs(title='Information about both a continuous variable and a
categorical variable',x='Race Ethnicity',y='Total Score')
```

Output:

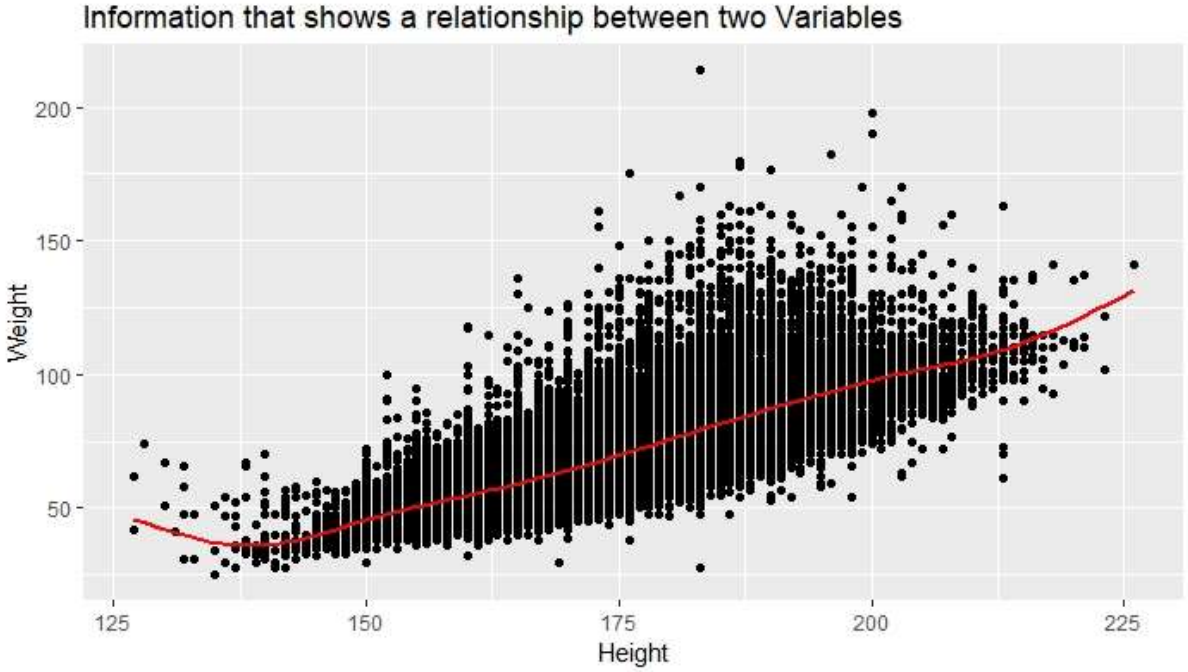


Plot 4a

Code:

```
ggplot(data = athlete_events, aes(x = Height, y = Weight)) +
geom_point() + geom_smooth(color = "red", se = FALSE)+
labs(title = 'Information that shows a relationship between two
Variables',x='Height',y='Weight')
```

Output:

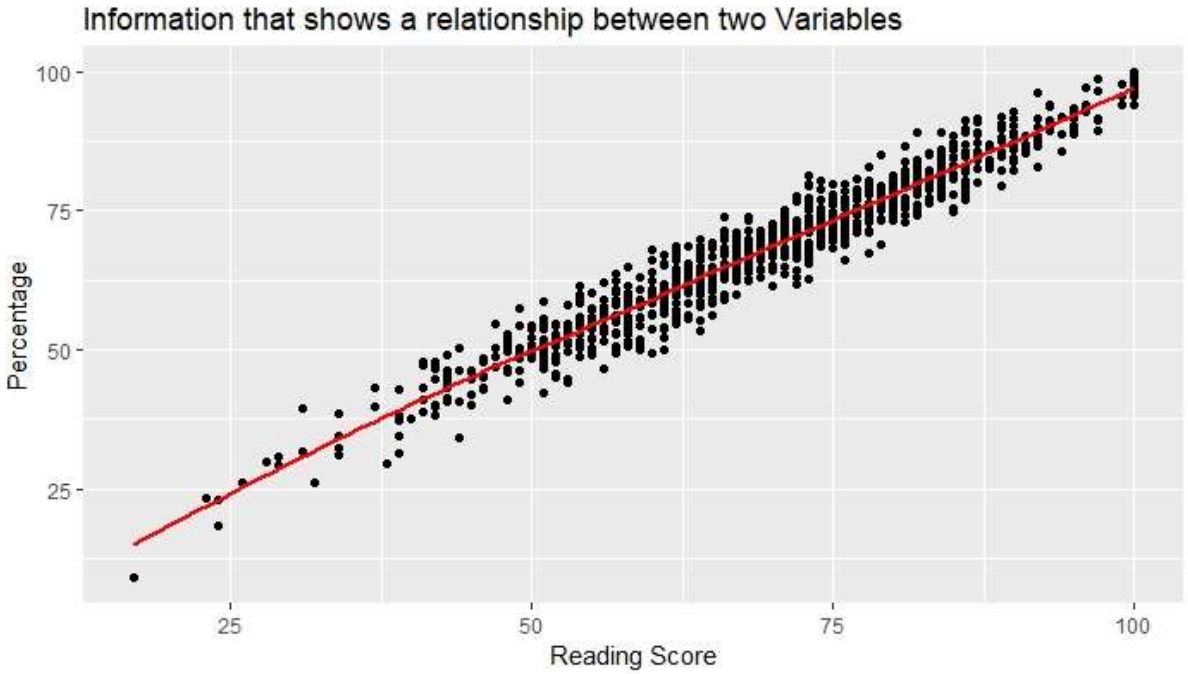


Plot 4b

Code:

- StudentsPerformance$percentage <-((StudentsPerformance$total.score * 100)/300)
- ggplot(StudentsPerformance,aes(x=reading.score,y=percentage))+ geom_point()+geom_smooth(se=FALSE,color="red")+ labs(title='Information that shows a relationship between two Variables',x='Reading Score',y='Percentage')

Output:

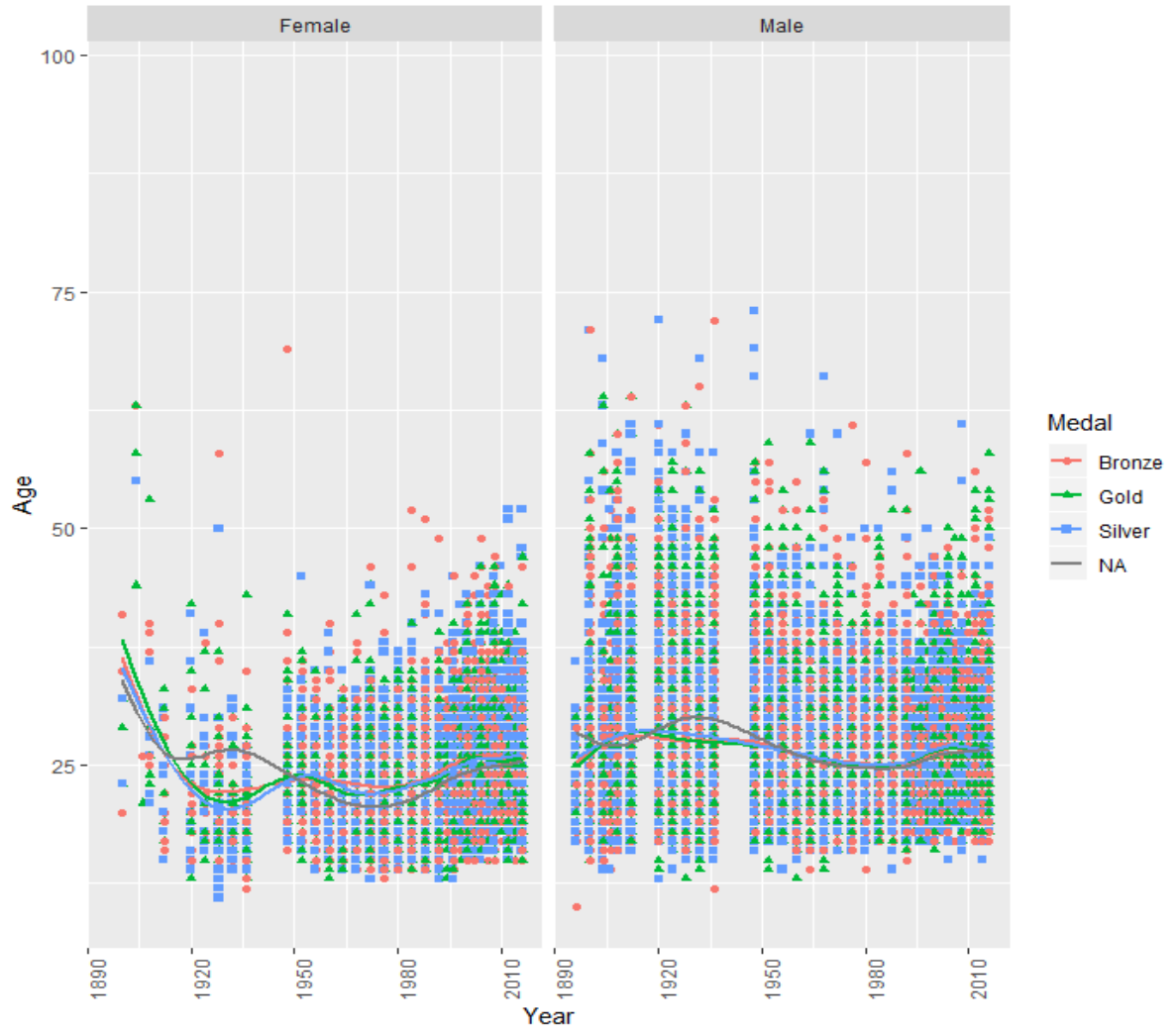


Plot 5a

Code:

- labels <- c(F = "Female" , M= "Male")
- ggplot(data = athlete_events, aes(x = Year, y = Age,color=Medal)) +
  geom_point(mapping = aes(shape=Medal)) +
  geom_smooth(aes(color=Medal), se = FALSE)+
  facet_grid(.~Sex,labeller = labeller(Sex=labels))+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  labs(title = 'Use faceting and display information about 4 variables',x='Year',y='Age')

Output:

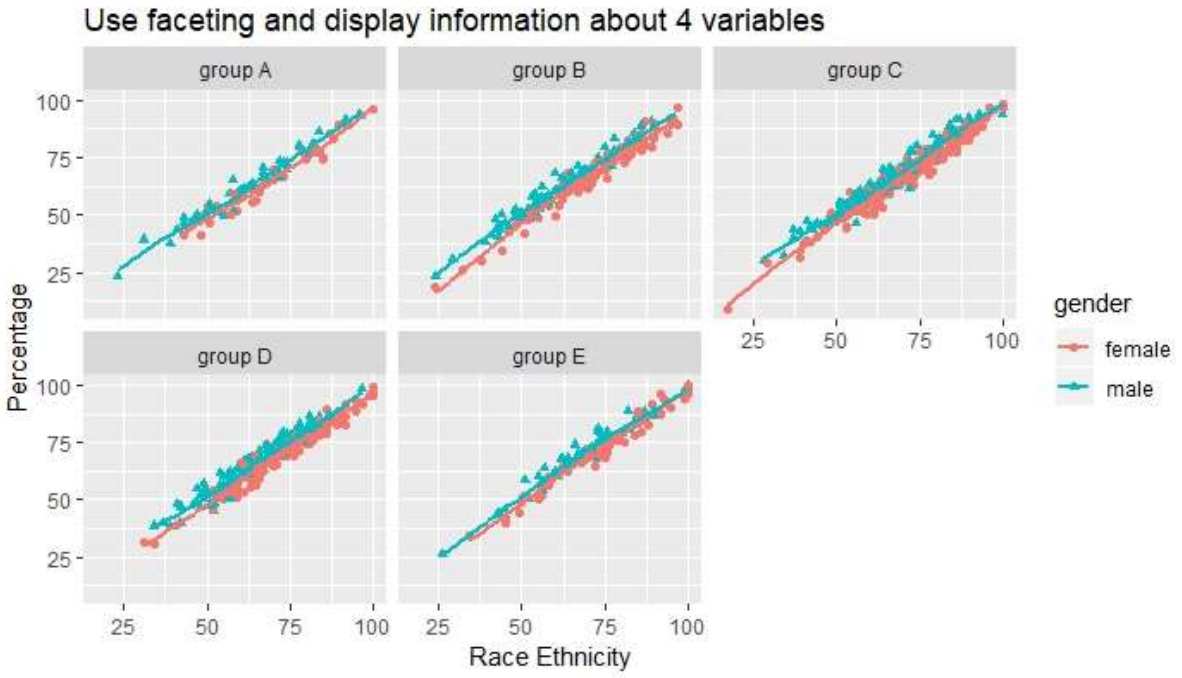


Plot 5b

Code:

```
ggplot(StudentsPerformance,aes(x=reading.score,y=percentage,
shape=gender,color=gender))+geom_point()+
geom_smooth(aes(color=gender),se=FALSE)+
facet_wrap(vars(race.ethnicity))+
labs(title='Use faceting and display information about 4
variables',x='Race Ethnicity',y='Percentage')
```

Output:



Plot 6a

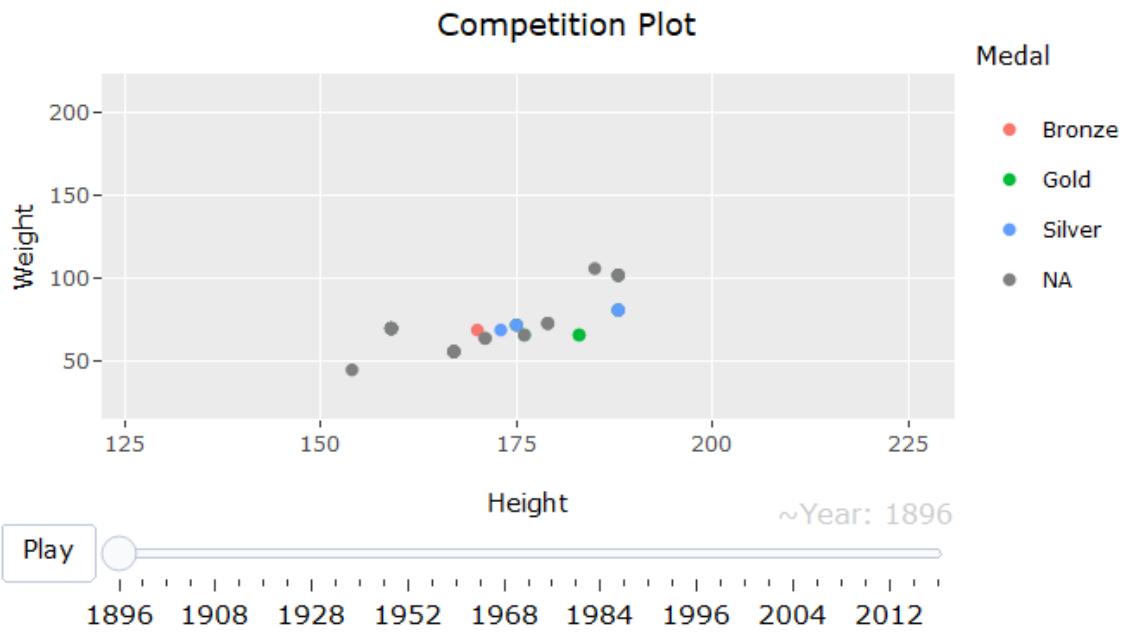
Code:

- p <- ggplot(athlete_events,

  aes(x=Height, y=Weight, frame = Year, color = Medal)) +

  geom_point()+labs(title = 'Competition Plot',x='Height',y='Weight')
- p <- ggplotly(p)
- p

Output:

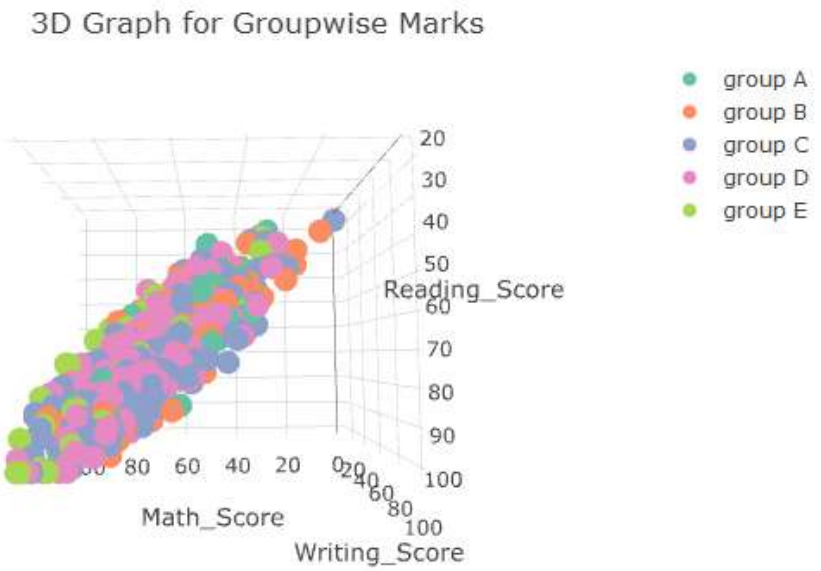


Plot 6b

Code:

```
plot_ly(StudentsPerformance,x=~math.score,
y=~reading.score,z=~writing.score)
%>% add_markers(color=~race.ethnicity) %>% layout (title = '3D Graph
for Groupwise Marks',scene = list(xaxis = list(title = 'Math_Score'),yaxis =
list(title = 'Reading_Score'),zaxis = list(title = 'Writing_Score')))
```

Output:



Questions

Question 1:

In what ways do you think data visualization is important to understanding a data set?

Answer 1:

Data visualization makes data understandable by using interactive tools like charts, bars, etc. Observing and understanding data set is very laborious and frivolous because it would consume loads of time and the ultimate output would be nothing. Where as in case of data visualization the output is very comprehensible and helps user understanding the data set by providing necessary information.

Question 2:

In what ways do you think data visualization is important to communicating important aspects of a data set?

Answer 2:

Data visualization makes it possible to integrate huge and complex data into graphical format. This makes it easy for business owners to understand and analyse current situation by means of patterns generated from data.

Question 3:

What role does integrity play when creating a data visualization for communicating results to others?

Answer 3:

Integrity is the key aspect of data visualization, because it makes sure that the data the is interpreted in correct manner.

Question 4

How many variables do you think you can successfully represent in a visualization? What happens when you exceed this number?

Answer 4

5 variables can be successfully represented, but exceeding this limit might make visualization complicated.

References

- https://www.kaggle.com/itsfarhan/athletes-events-datasets
- https://www.kaggle.com/spscientist/students-performance-in-exams
- https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf
- https://plotly-book.cpsievert.me/

Appendix:

- Plot1a: We are using Athlete_events data frame and plotted one continuous variable “Age”
- Plot1b: We are using StudentsPerformance data frame and plotted one continuous variable “total_ score” which is derive from addition of math.score, reading.score, writing.score.
- Plot2a: We are using Athlete_events data frame and plotted one categorical variable “Sex”.
  Plot2b: We are using StudentsPerformance data frame and plotted one categorical variable “parental.level.of.education”.
- Plot3a: We are using Athlete_events data frame and plotted one categorical variable “Sex” and one continuous variable ”Age”.
- Plot3b: we are using StudentsPerformance data frame and plotted one categorical variable “race.ethnicity” and one continuous variable ”total.score”.
- Plot4a: we are using Athlete_events data frame and plotted two variables “Height” and ” Weight”.
- Plot4b: we are using StudentsPerformance dataframe and plotted two variables “percentage” and ”reading.score”. Where percentage can be calculate from (total.score*100)/300.
- Plot5a:In this plot first we create vector to give label to facets and then use “Year” for x-axis and “Age” for y-axis and “medal” used for color. Variable “Sex” is used in facet.
- Plot5b:we plotted “reading.score” and “percentage” variables and “gender” used for color as well as shape. “race.ethnicity” used for creating facet.
- Plot6a:In this plot we used Athlete_events data frame and mapped “Height” to x-axis ,“Weight” to y-axis,”Year” with frame and “medal” with color. Here we used “ggplotly” for animation.
- Plot6b: Here we have done 3D plotting on StudentsPerformance data frame by using plotly.X=”math.score”, Y=”reading.score”, Z=”writing.score” and color=”race.ethnicity”.