

Data visualisation in R

Mohsin Mohammed

```
library(dplyr)
```

```
##  
## Attaching package: 'dplyr'
```

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

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

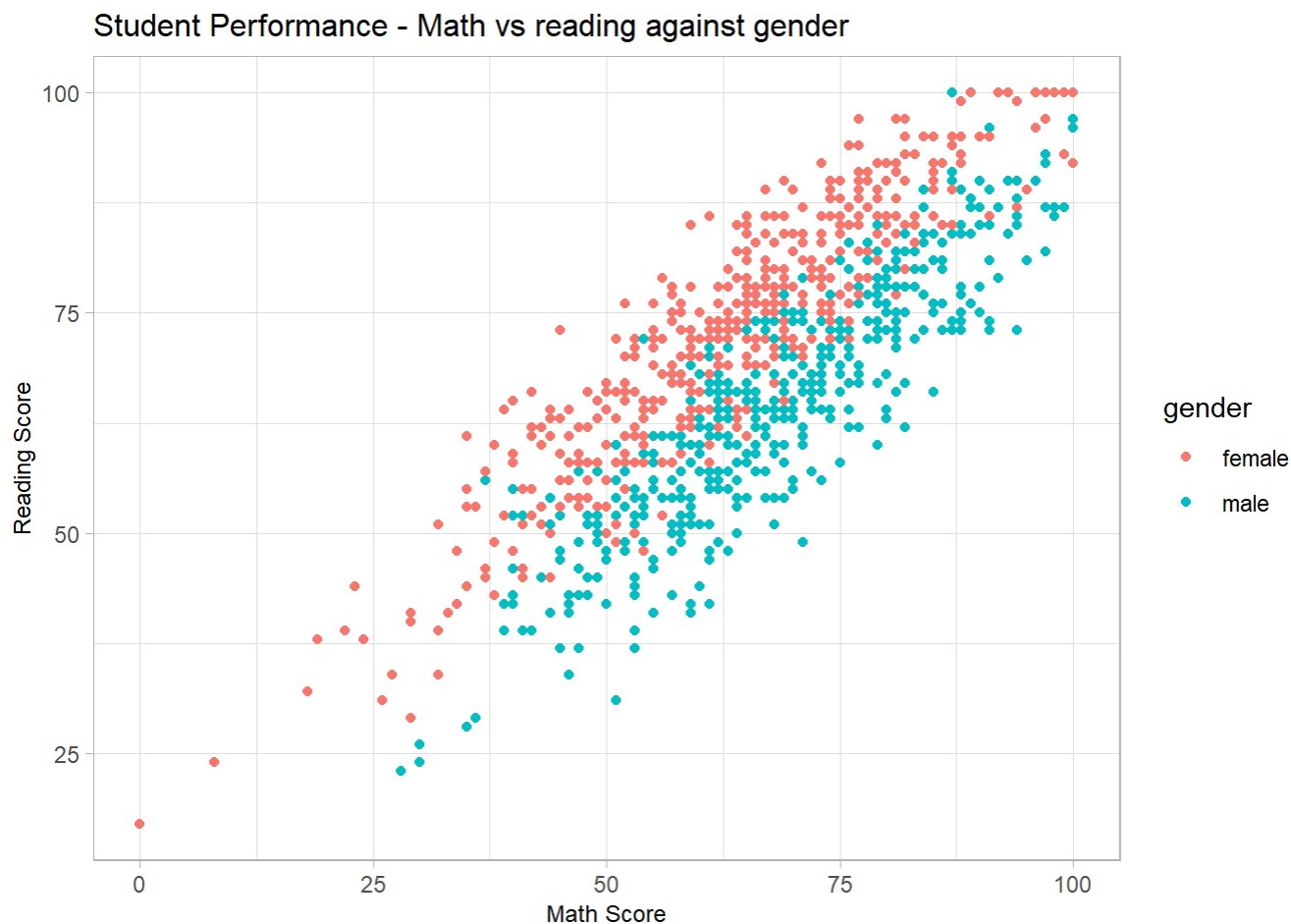
```
library(ggplot2)
```

Lets load the Student Performance data set and do exploratory data analysis.

```
data <- read.csv('./StudentsPerformance.csv')
```

Data visualization - 1 (Scatter plot)

```
ggplot(data, aes(x=math.score, y=reading.score)) + geom_point(aes(color=gender)) +  
labs(y="Reading Score", x="Math Score", title="Student Performance - Math vs reading against  
gender") + theme_light() + theme(plot.title = element_text(size=12), axis.title=element_text  
(size=9))
```

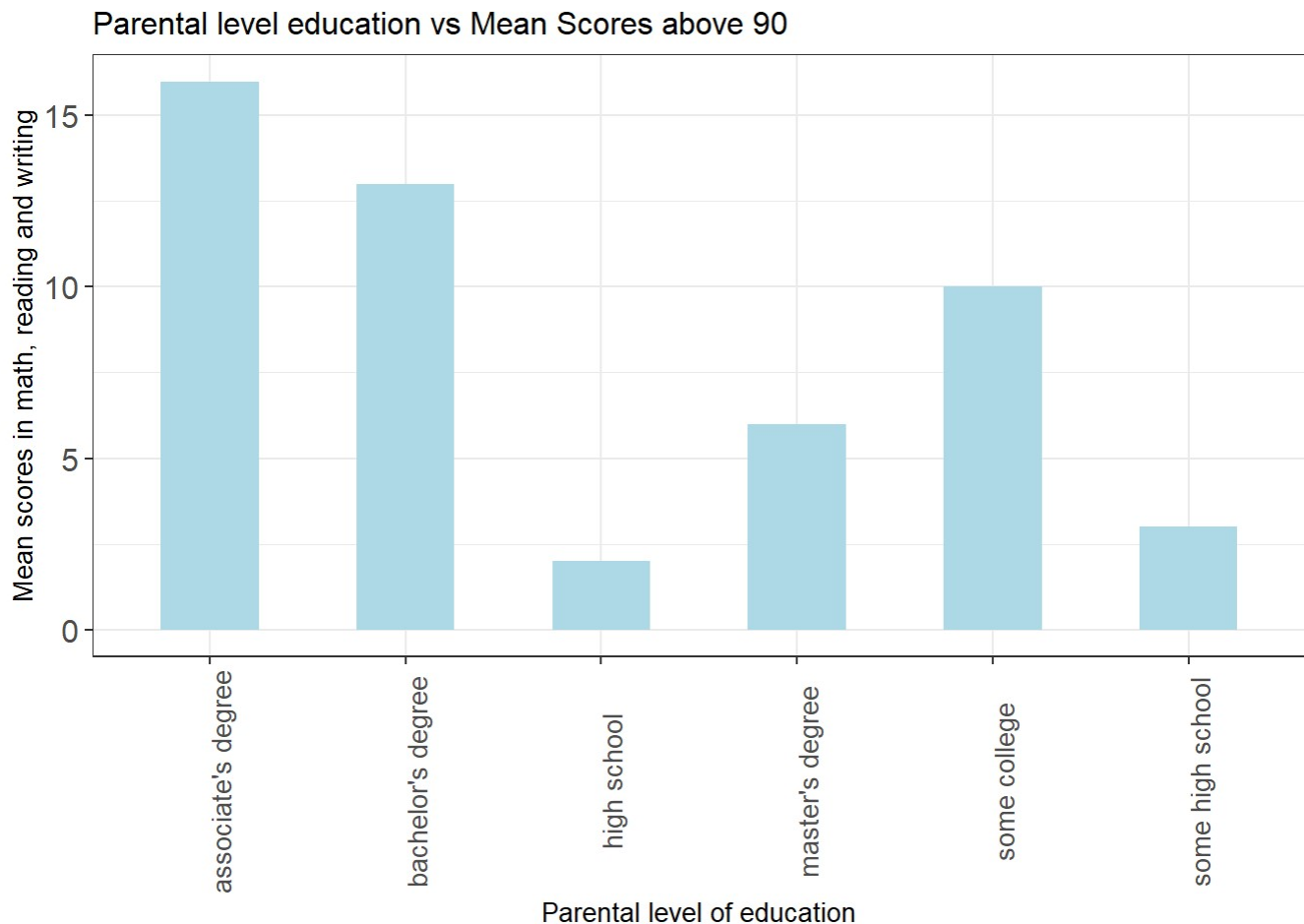


We can see that females scored better on reading compared to their counterpart. While males scored better on math. We can test this by switching the x and the y labels in the aes function parameter.

Data visualization - 2 (Bar Chart)

```
## calculating the mean of math, reading and writing scores:
ndata <- mutate(data, mean=(math.score + reading.score + writing.score)/3)

## plotting the mean scores > 90 against the parental level of education to understand the relation between them.
ggplot(ndata[ndata$mean>90,], aes(parental.level.of.education))+
  geom_bar(stat="count", width = 0.5, fill="lightblue")+
  labs(x="Parental level of education",
       y="Mean scores in math, reading and writing",
       title="Parental level education vs Mean Scores above 90")+
  theme_bw()+
  theme(plot.title = element_text(size=12),axis.text.x= element_text(size=10, angle=90),
        axis.text.y= element_text(size=12), axis.title=element_text(size=10))
```



We started by finding the average scores of all students for math, reading and writing and plotted the means that were greater than 90 and plotted them against their parental level of education. We found that students whose parents have an associates degree scored the highest followed by parents who held a bachelors degree. This shows that there is a correlation between student performance and their parental level of education.

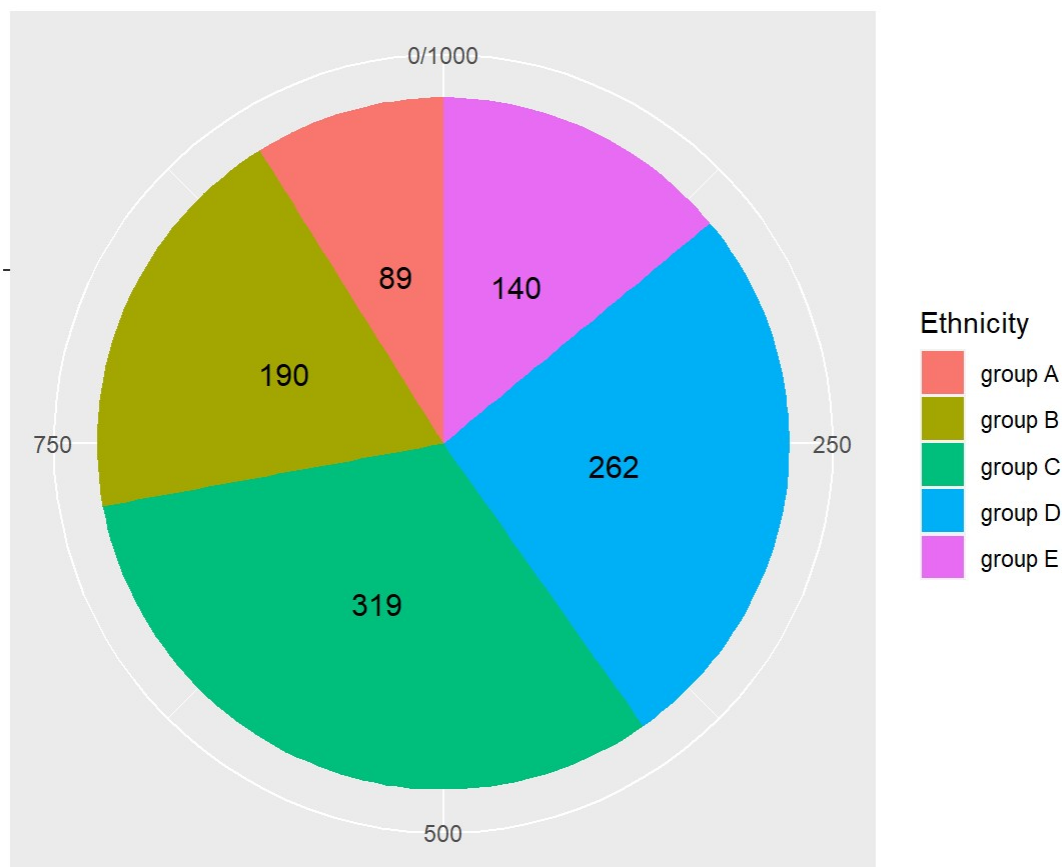
Data visualization - 3 (Pie Chart)

```
# Creating a table of counts of all the different categories in the ethnicity column
ethnicity_counts <- table(data$race.ethnicity)

# converting the counts table into a data frame:
ethnicity_df <- data.frame(ethnicity = names(ethnicity_counts),
                           count = ethnicity_counts)

# creating a pie chart of the ethnicity and their counts
ggplot(ethnicity_df, aes(x = "", y = count.Freq, fill = ethnicity)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar(theta = "y") +
  labs(fill = "Ethnicity", x = NULL, y = NULL,
       title = "Pie Chart of Ethnicity in Student Performance Dataset")+
  geom_text(aes(label=count.Freq), size=4, position=position_stack(vjust=0.5))
```

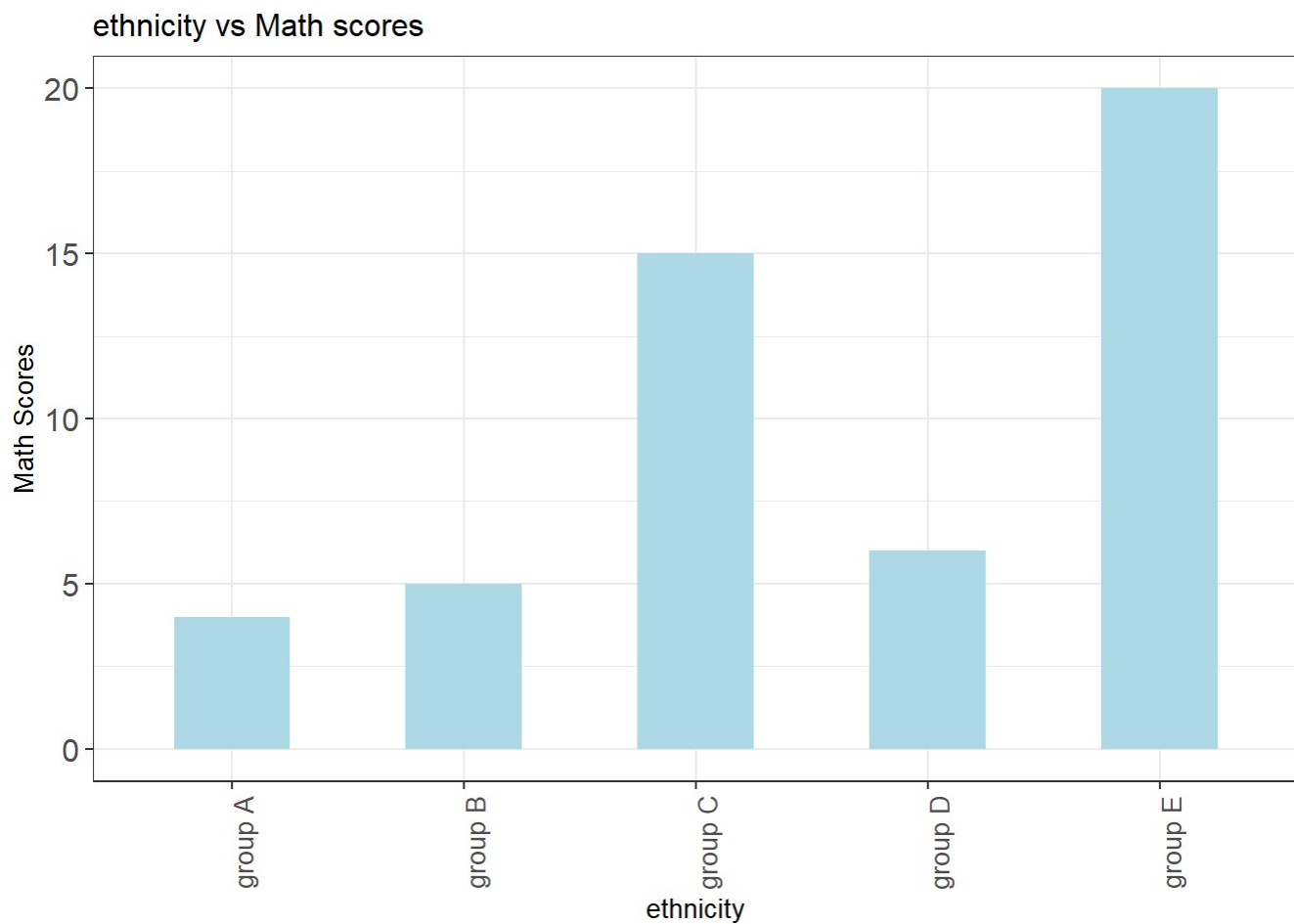
Pie Chart of Ethnicity in Student Performance Dataset



We can see that the majority of the students belong to the group C ethnicity followed by group D and Group A has the lowest representation.

Data Visualization - 4 (Bar Chart)

```
# Lets see which ethnicity scores the highest in math
## plotting the mean scores > 90 against the parental level of education to understand the relation between them.
ggplot(ndata[data$math.score>90,], aes(race.ethnicity))+
  geom_bar(stat="count", width = 0.5, fill="lightblue")+
  labs(x="ethnicity",
       y="Math Scores",
       title="ethnicity vs Math scores")+
  theme_bw()+
  theme(plot.title = element_text(size=12),axis.text.x= element_text(size=10, angle=90),
        axis.text.y= element_text(size=12), axis.title=element_text(size=10))
```



We can see that students belonging to ethnicity group E scored the highest in Math followed by group C.