Day 2

Agenda

- Descriptive Analysis
  - Min, Max, Range, Sum, Median, Var, Sd etc.
- Graphical Analysis
  - Histogram, Box Plot, Bar Chart, Correlation Matrix, Scatter Plots etc.

How many observations?
How many variables?

```
> dim(grades)
[1] 105   22
```

How many observations are there in a variable?

```
> length(grades$ethnicity)
[1] 105
```

Know More about Data

- min(grades$ethnicity)
- max(grades$ethnicity)
- range(grades$ethnicity)
- sum(grades$ethnicity)
- median(grades$ethnicity)
- var(grades$ethnicity)
- sd(grades$ethnicity)

Know structure of data

```
> str(grades)
'data.frame':   105 obs. of  22 variables:
 $ Sr_No    : int  1 2 3 4 5 6 7 8 9 10 ...
 $ id       : int  106484 108642 127285 132931 140219 142630 153964 15444
1 157147 164605 ...
 $ lastname : Factor w/ 99 levels "AHGHEL","ANDERSON",..: 92 89 28 59 32
65 86 51 5 46 ...
 $ firstname: Factor w/ 98 levels "AARON","ALFRED",..: 2 80 35 3 92 89 17
 40 49 19 ...
 $ gender   : int  2 2 1 1 1 1 2 1 2 1 ...
 $ ethnicity: int  2 4 4 3 2 4 2 5 4 3 ...
 $ year     : int  2 3 4 2 4 3 3 2 3 3 ...
 $ lowup    : int  1 2 2 1 2 2 2 1 2 2 ...
 $ section  : int  2 2 2 2 1 3 3 1 1 2 ...
 $ gpa      : num  1.18 2.19 2.46 3.98 1.84 3.9 2.84 3.57 3.95 3.49 ...
 $ extrc    : int  1 2 2 1 1 1 2 1 2 2 ...
 $ review   : int  2 1 2 1 1 2 1 2 2 1 ...
 $ quiz1    : int  6 10 10 7 7 10 10 10 10 10 ...
 $ quiz2    : int  5 10 7 8 8 10 9 9 10 10 ...
 $ quiz3    : int  7 7 8 7 9 10 10 10 10 9 ...
 $ quiz4    : int  6 6 9 7 8 9 10 10 10 10 ...
 $ quiz5    : int  3 9 7 6 10 9 10 10 9 10 ...
 $ final    : int  53 54 57 68 66 74 63 71 74 75 ...
 $ total    : int  80 96 98 103 108 122 112 120 123 124 ...
 $ percent  : int  64 77 78 82 86 98 90 96 98 99 ...
 $ grade    : Factor w/ 5 levels "A","B","C","D",..: 4 3 3 2 2 1 1 1 1 1
...
 $ passfail : Factor w/ 3 levels "F","O","P": 3 3 3 3 3 3 3 3 3 3 ...
>
```

Top and Bottom 6 Data Points

- head(grades) for top 6 data points
- tail (grades) for bottom 6 data points

Summary of Data

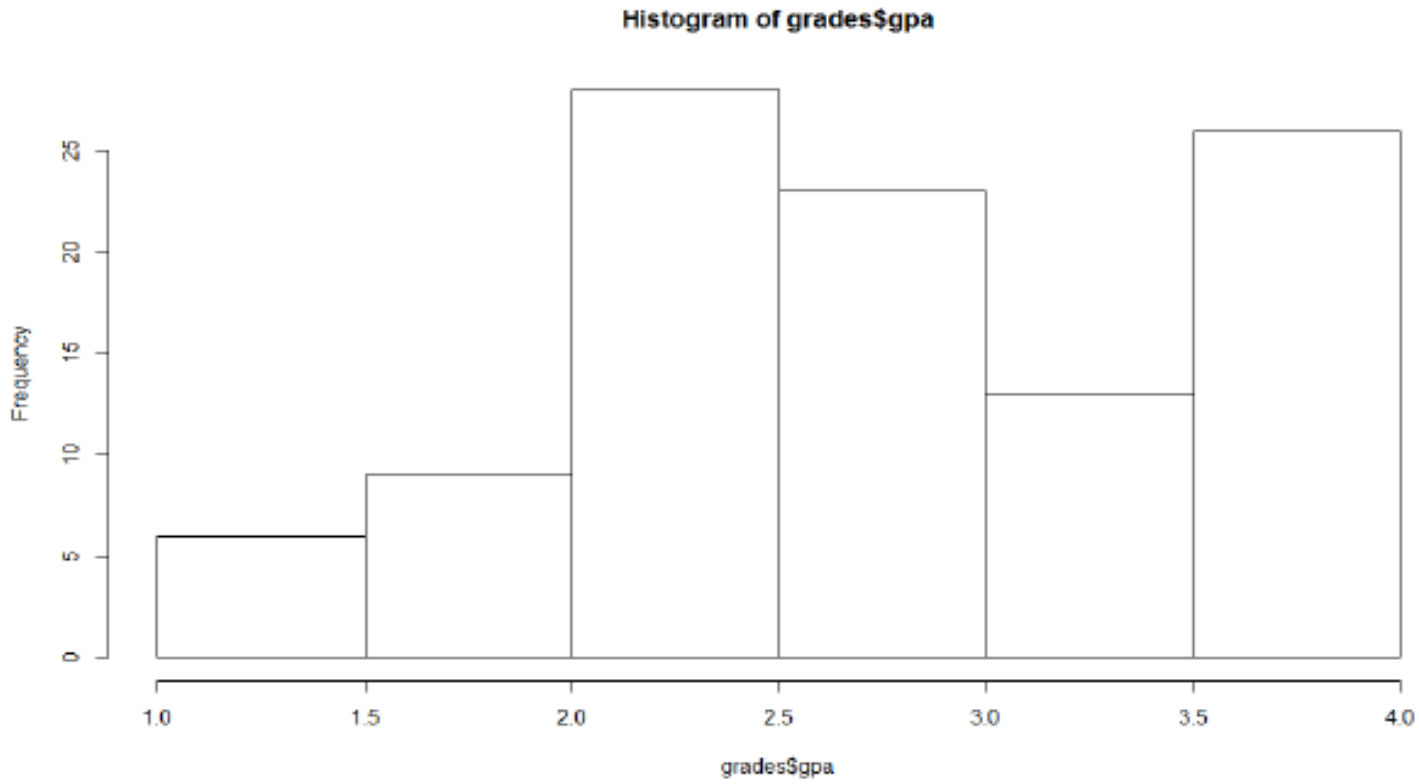
- summary(grades)

- For further better understanding
  - Install package (psych)
  - Load package (psych)
  - describe(grades)

Graphical Analysis

Histogram of gpa

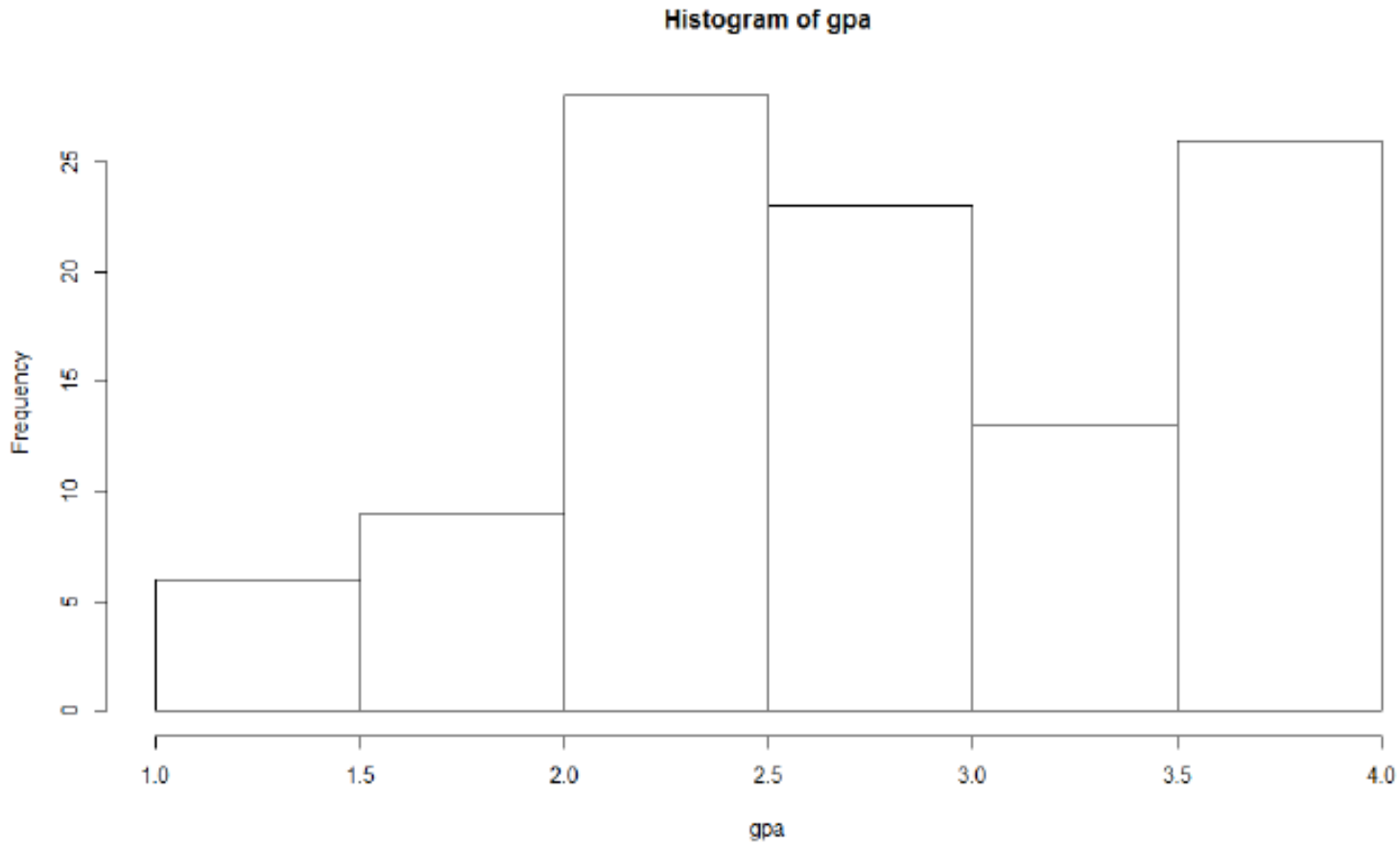
```
> hist(grades$gpa)
```



Histogram of gpa

Do proper labelling and Heading

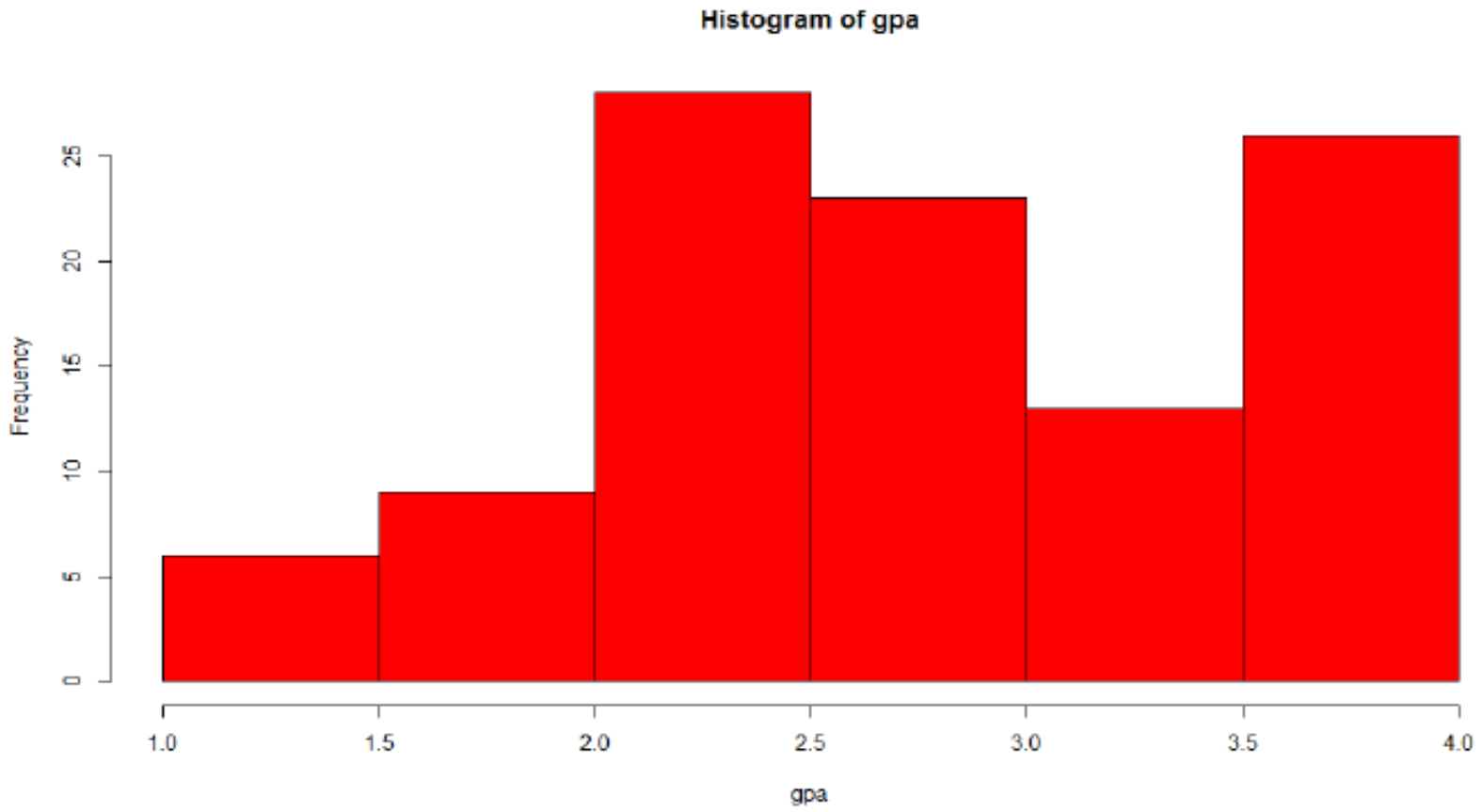
```
hist(grades$gpa, xlab="gpa", ylab="Frequency", main = "Histogram of gpa" )
```



Histogram of gpa

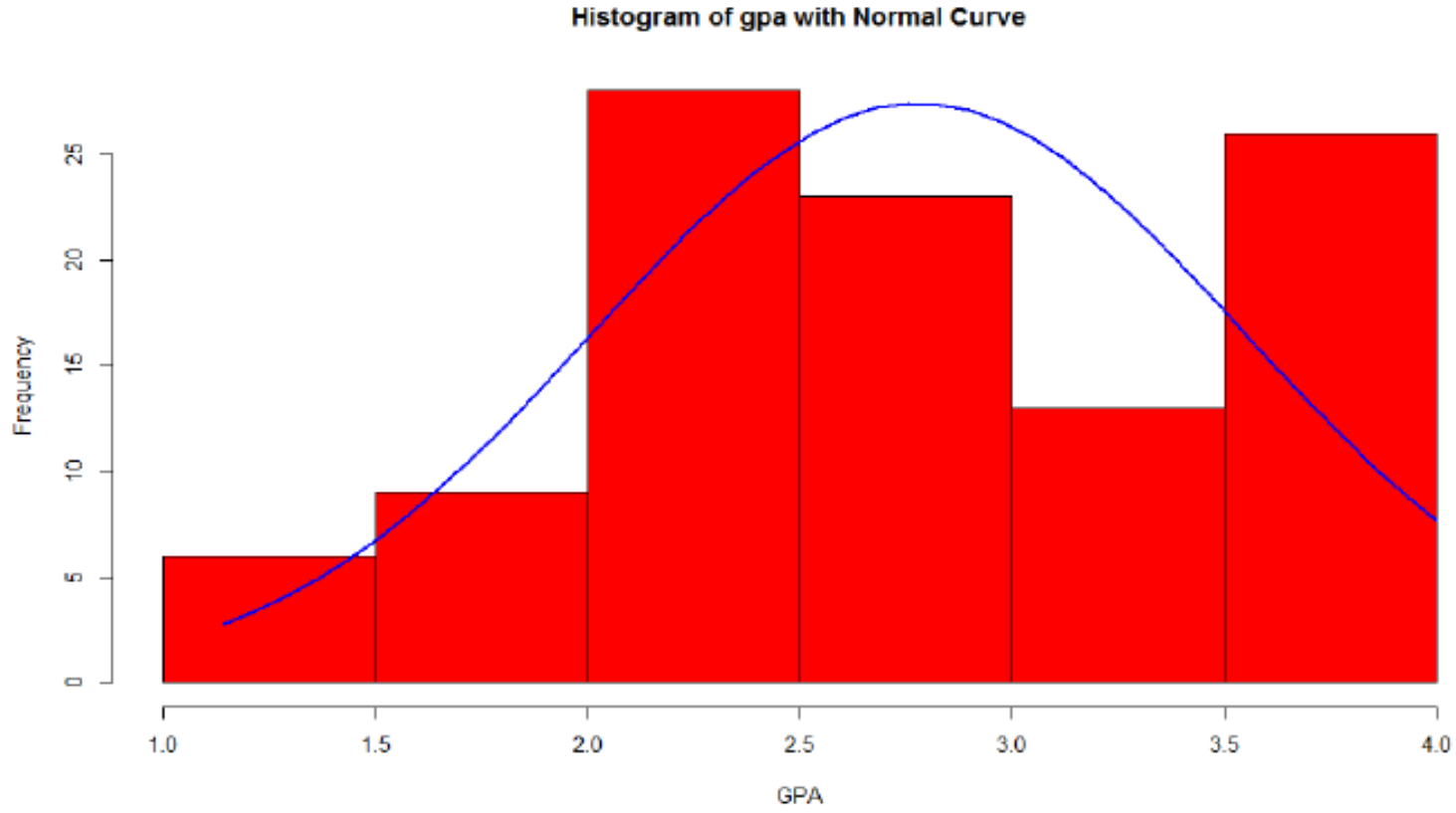
Fill red colour

```
 > hist(grades$gpa, xlab="gpa", ylab = "Frequency", main = "Histogram of
gpa", col = "red" )
```



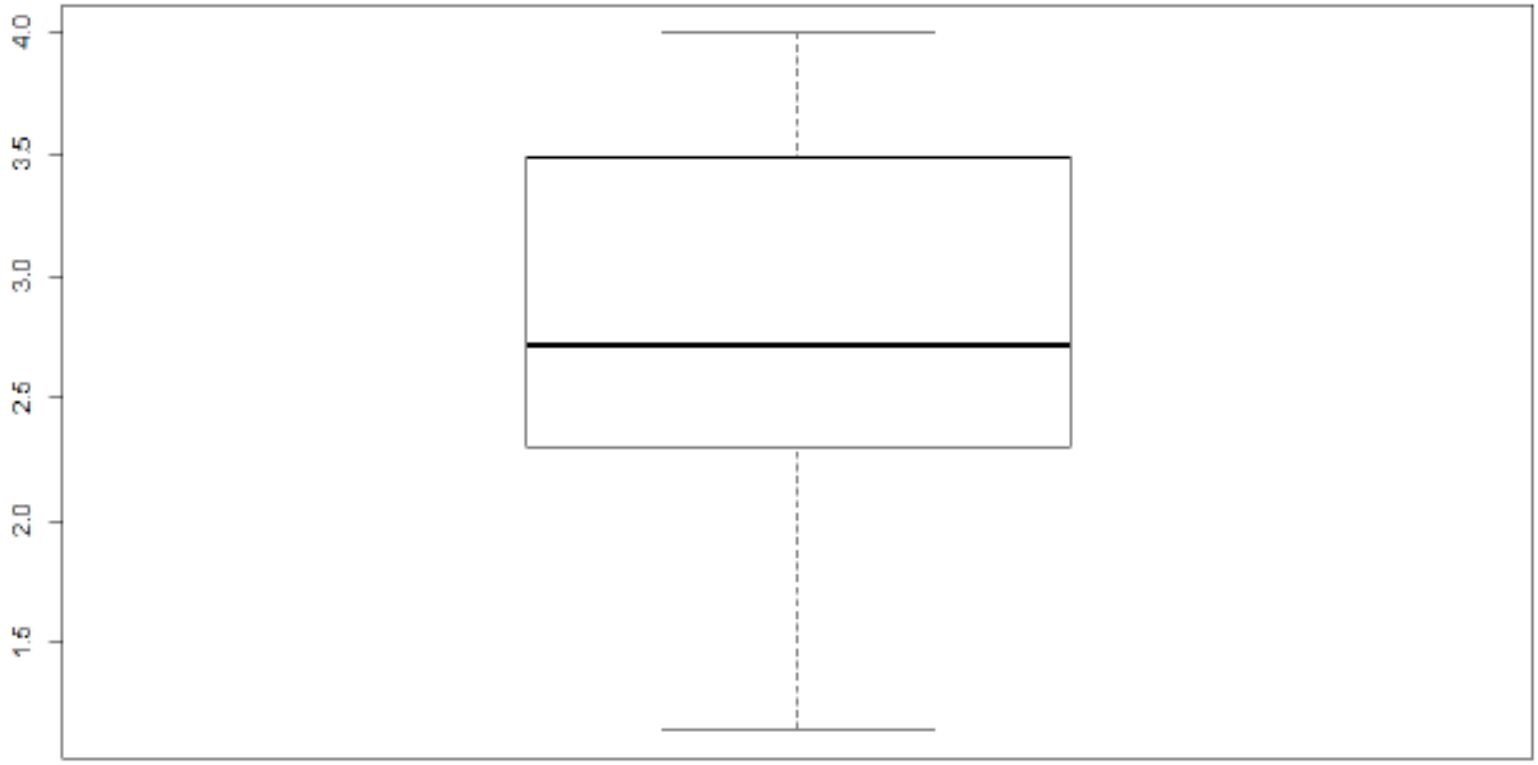
Histogram of gpa with Normal Curve

```
> x<-grades$gpa
> h<-hist(x, breaks=10, col="red", xlab ="GPA", main="Histogram of gpa
with Normal Curve")
> xfit<-seq(min(x), max(x), length =40)
> yfit<-dnorm(xfit, mean=mean(x), sd=sd(x))
> yfit<-yfit*diff(h$mids[1:2]*length(x))
> lines(xfit, yfit, col="blue", lwd=2)
```



Box Plot of gpa

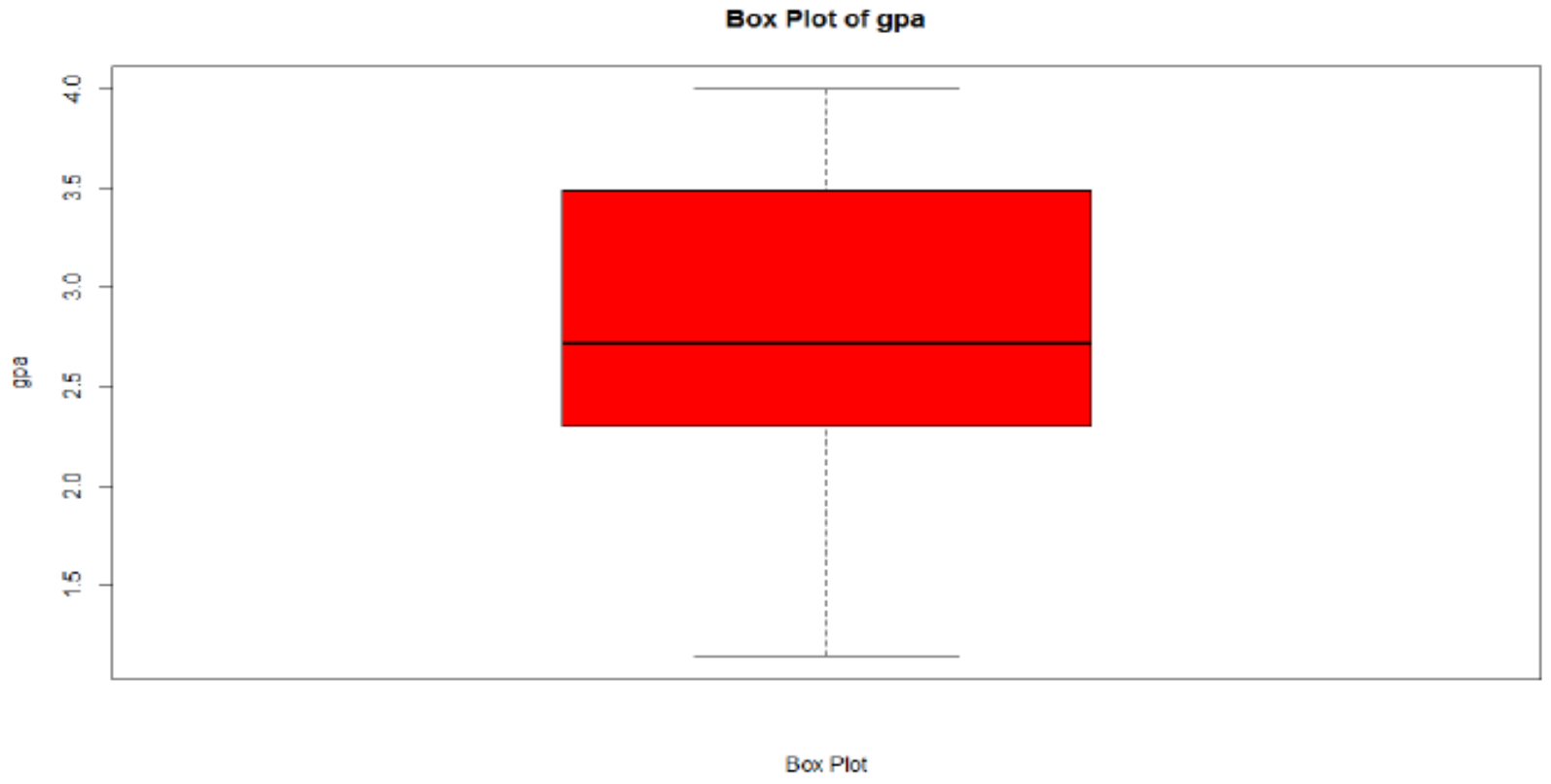
```
> boxplot(grades$gpa)
```



Box Plot of gpa

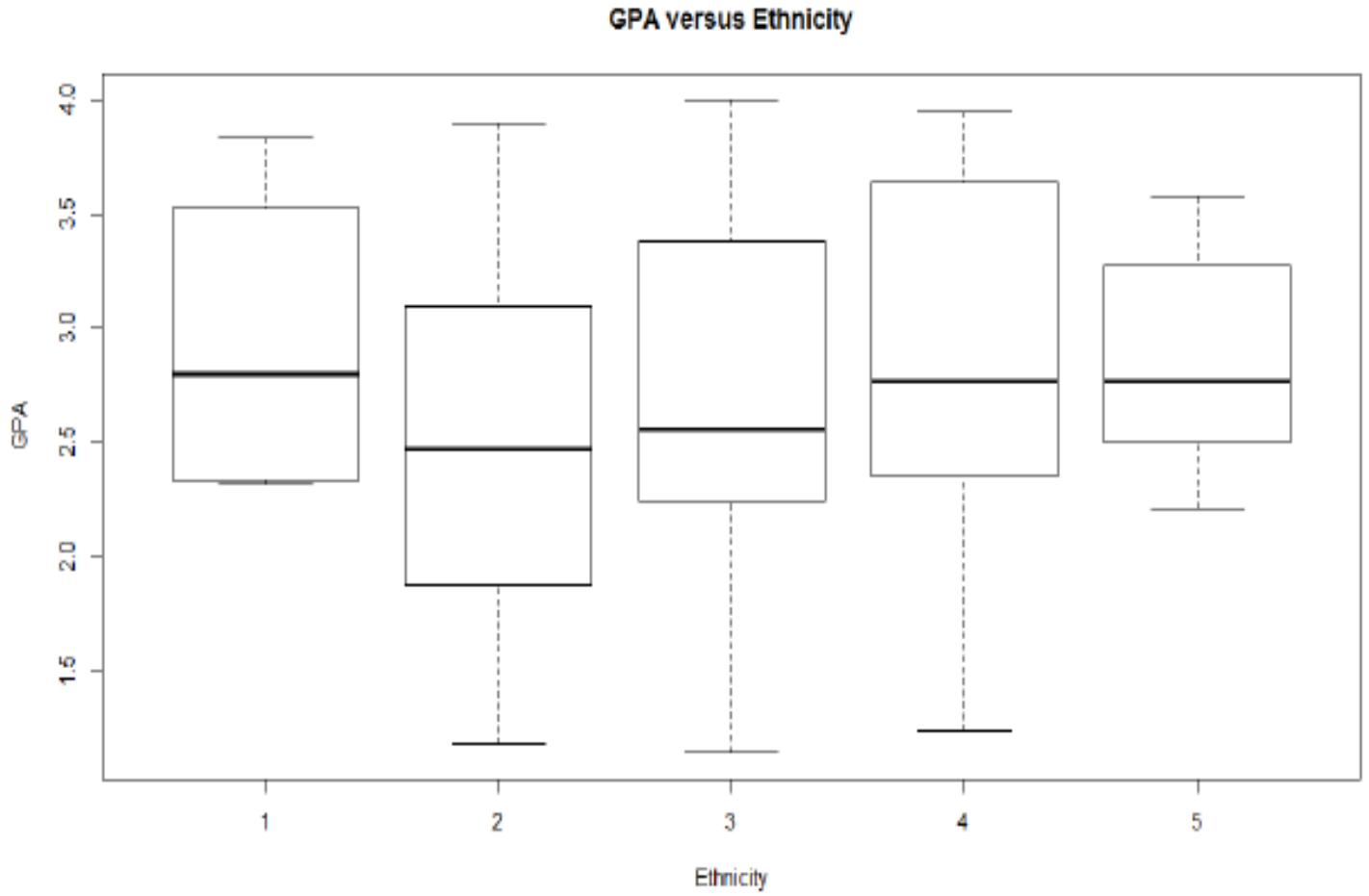
Label and colour

```
> boxplot(grades$gpa, xlab="Box Plot", ylab="gpa", main = "Box Plot of
gpa", col="red")
```



Box Plot of *gpa* versus *ethnicity*

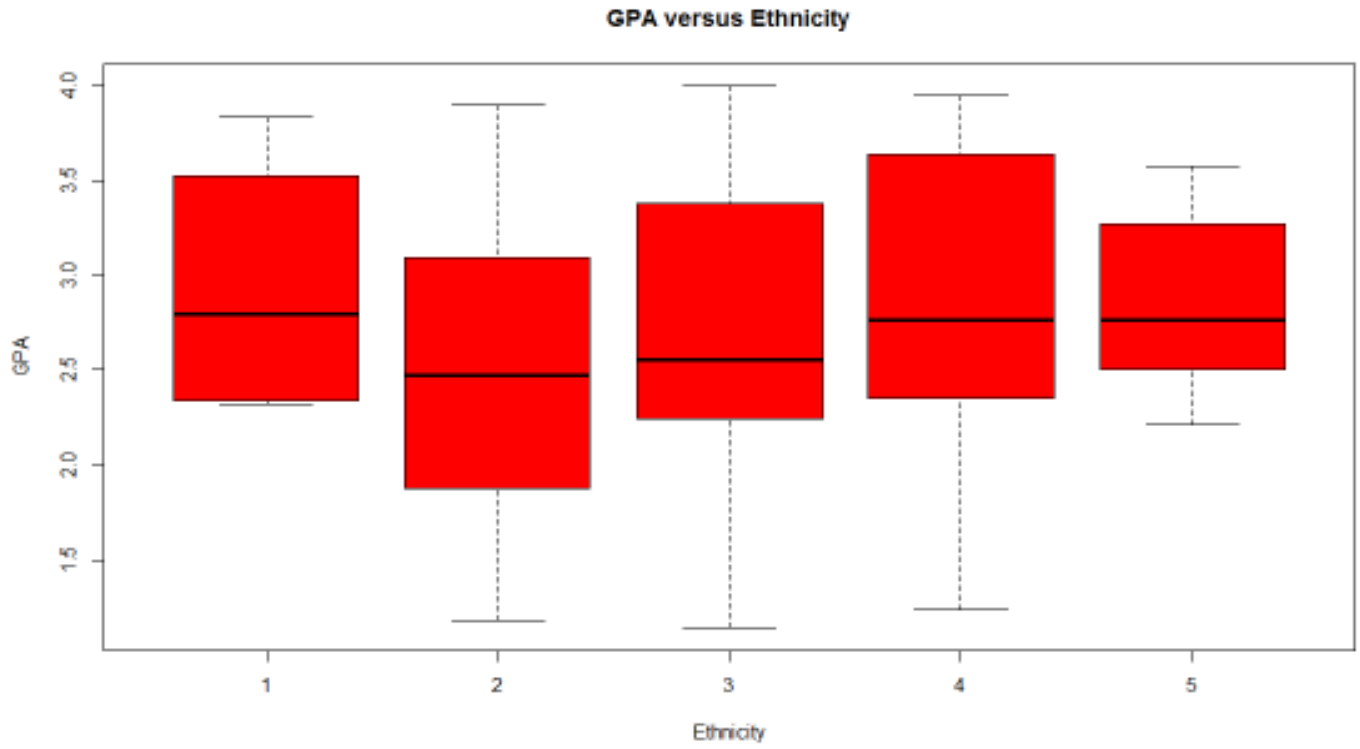
```
> boxplot(gpa~ethnicity, data = grades, main = "GPA versus Ethnicity",
xlab="Ethnicity", ylab="GPA")
```



Box Plot of *gpa* versus *ethnicity*

One colour 'red'

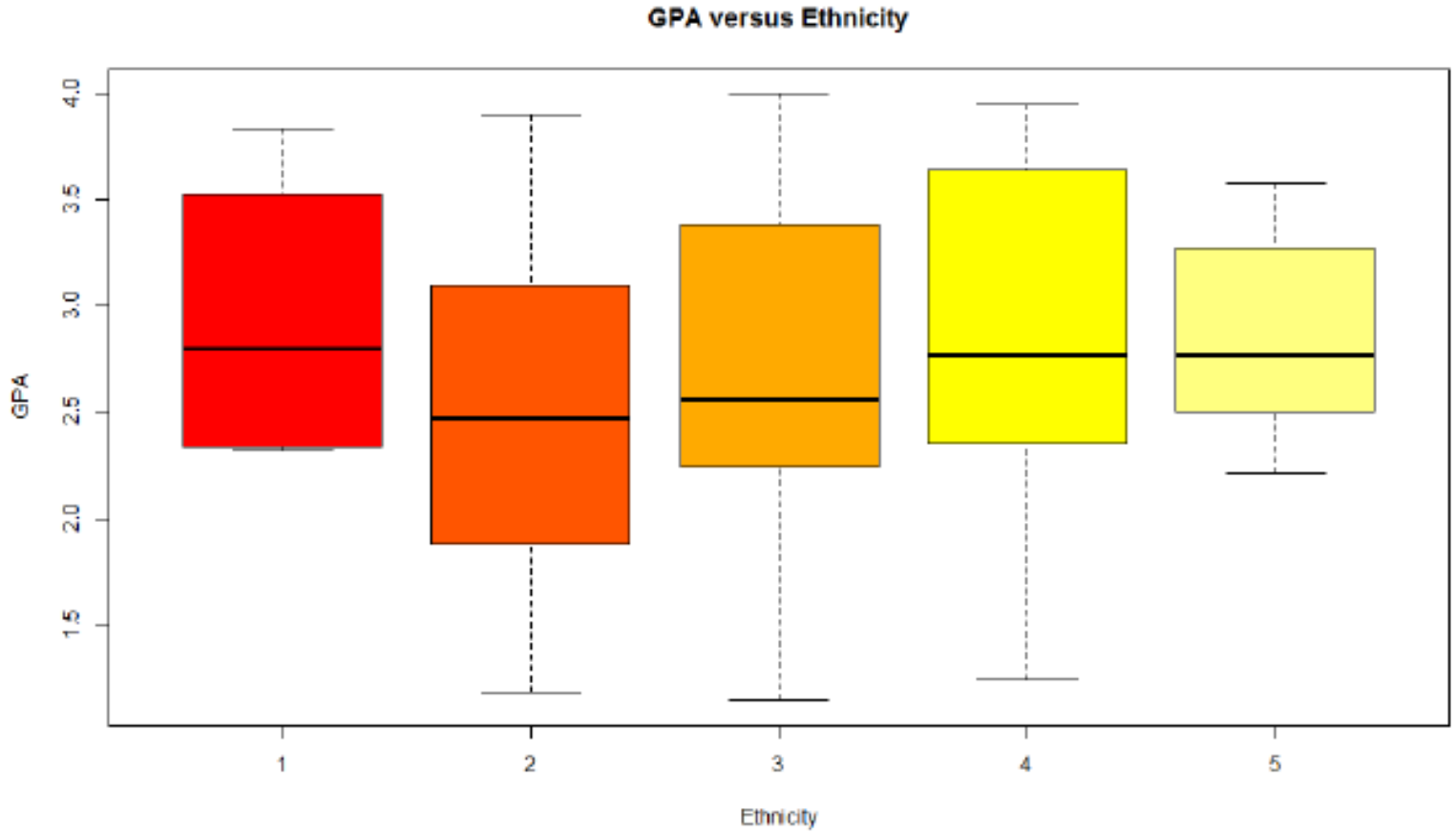
```
> boxplot(gpa~ethnicity, data = grades, main = "GPA versus Ethnicity",
xlab="Ethnicity", ylab="GPA", col= "red")
```



Box Plot of *gpa* versus *ethnicity*

Different colours 'heat.colors(5)'

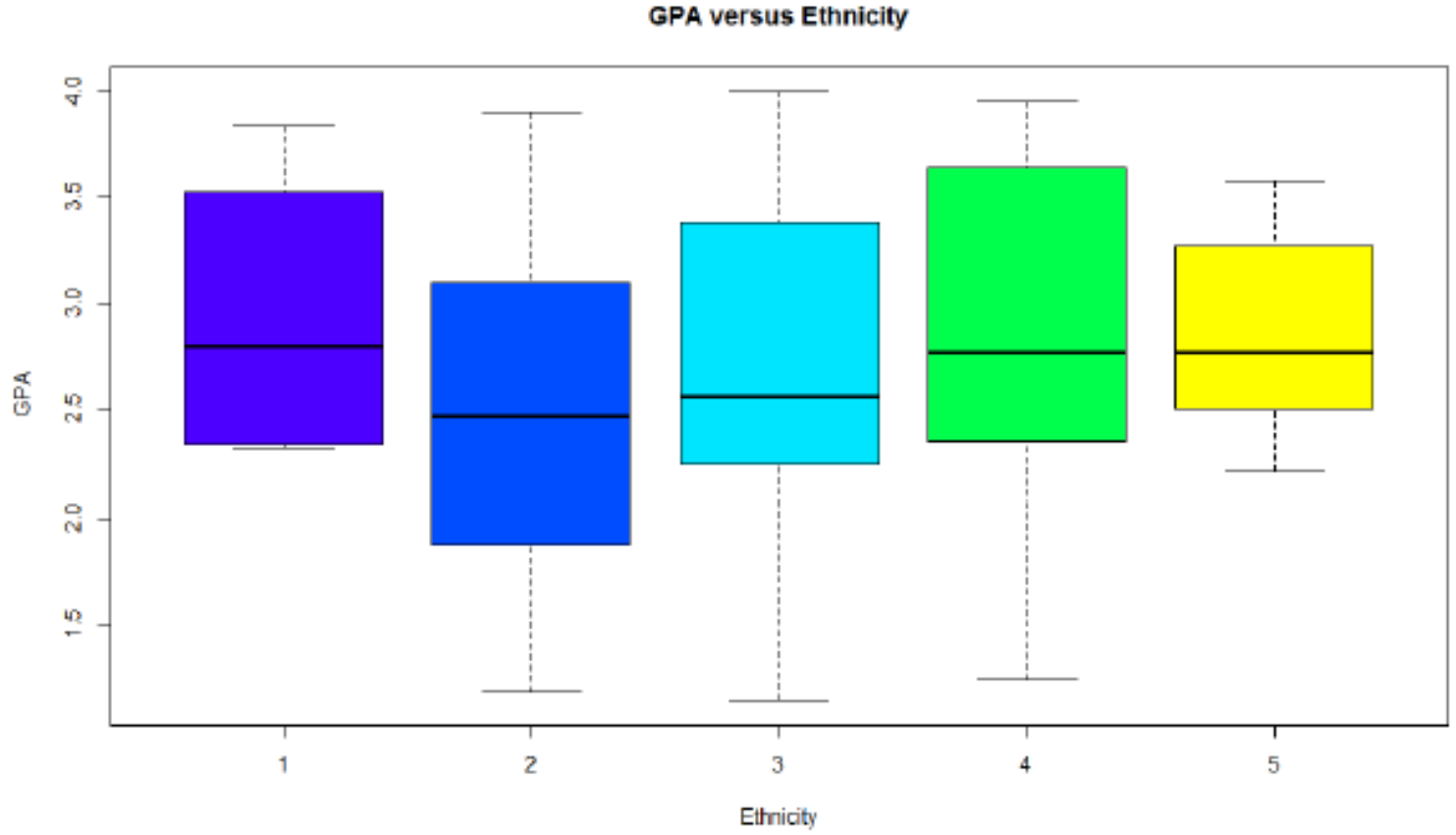
```
> boxplot(gpa~ethnicity, data = grades, main = "GPA versus Ethnicity",
xlab="Ethnicity", ylab="GPA", col= heat.colors(5))
```



Box Plot of *gpa* versus *ethnicity*

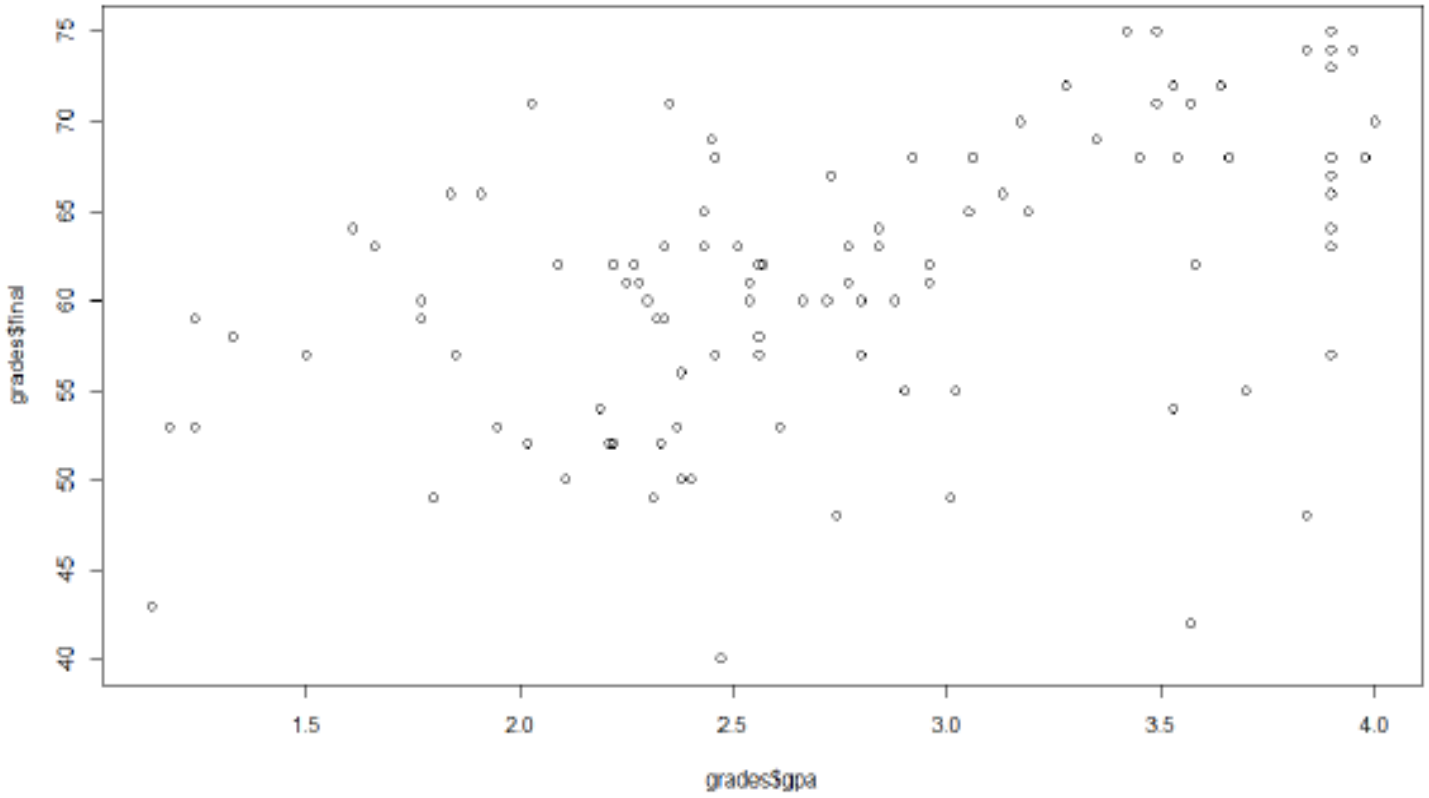
Different colours 'topo.colors(5)'

```
> boxplot(gpa~ethnicity, data = grades, main = "GPA versus Ethnicity",
xlab="Ethnicity", ylab="GPA", col= topo.colors(5))
```



Scatter Plot of Final versus gpa

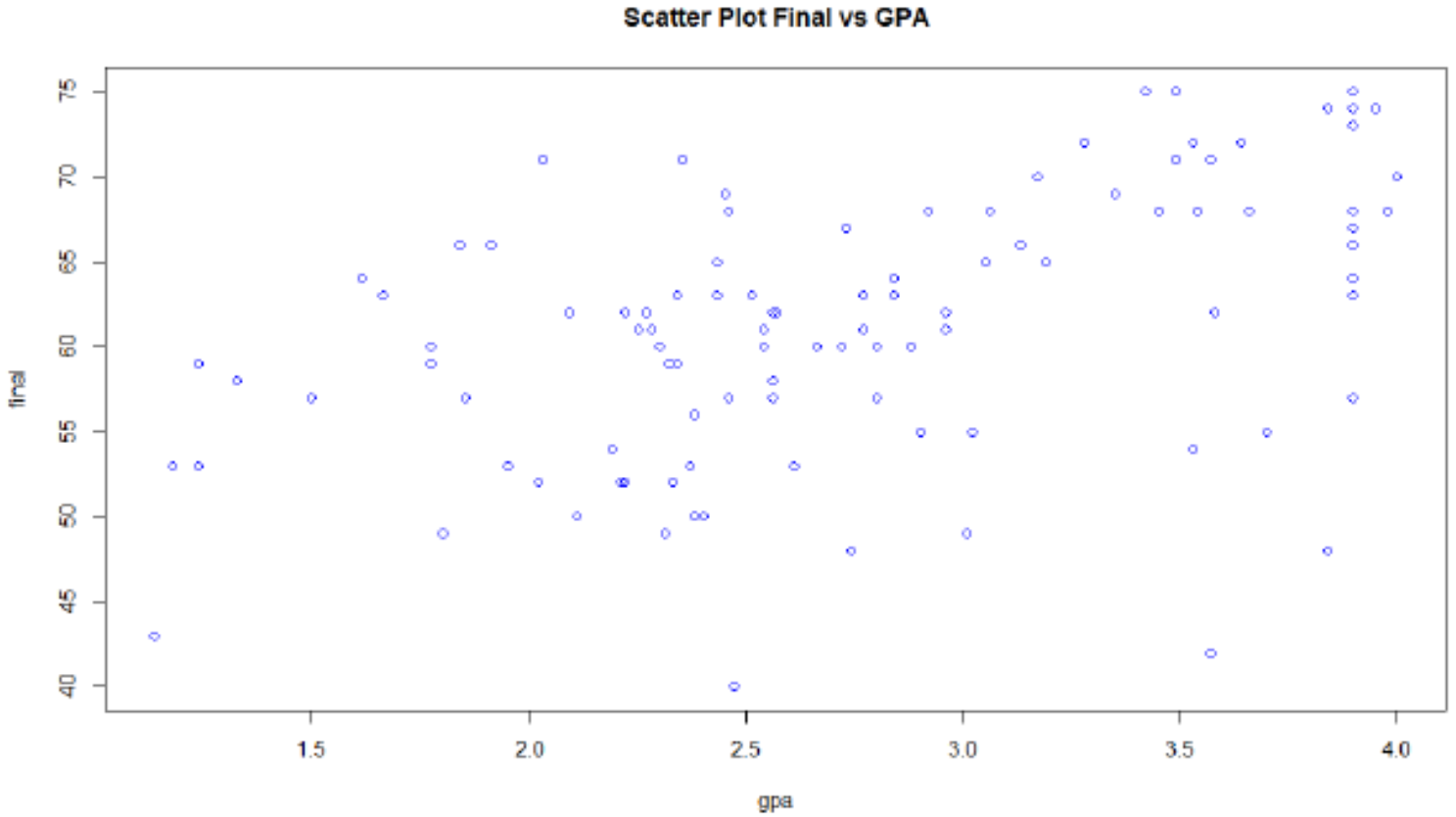
```
> plot(grades$gpa, grades$final)
```



Scatter Plot of Final versus gpa

Label, Heading and colour dots as blue color

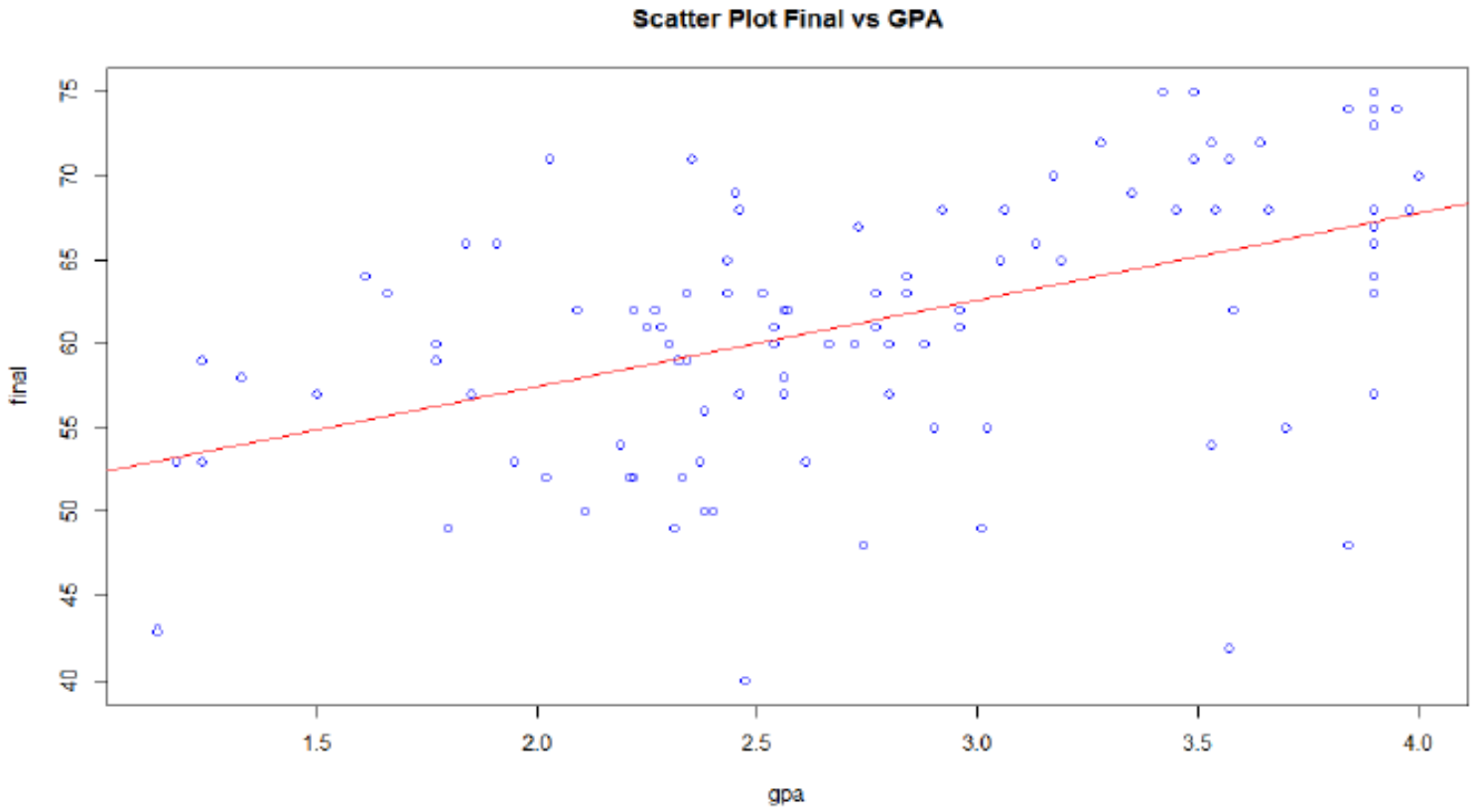
```
> plot(grades$gpa, grades$final, xlab = "gpa", ylab = "final", main
="Scatter Plot Final vs GPA", col="blue")
```



Scatter Plot of Final versus gpa

Label, Heading, colour dots as blue colour and add regression line in colour red

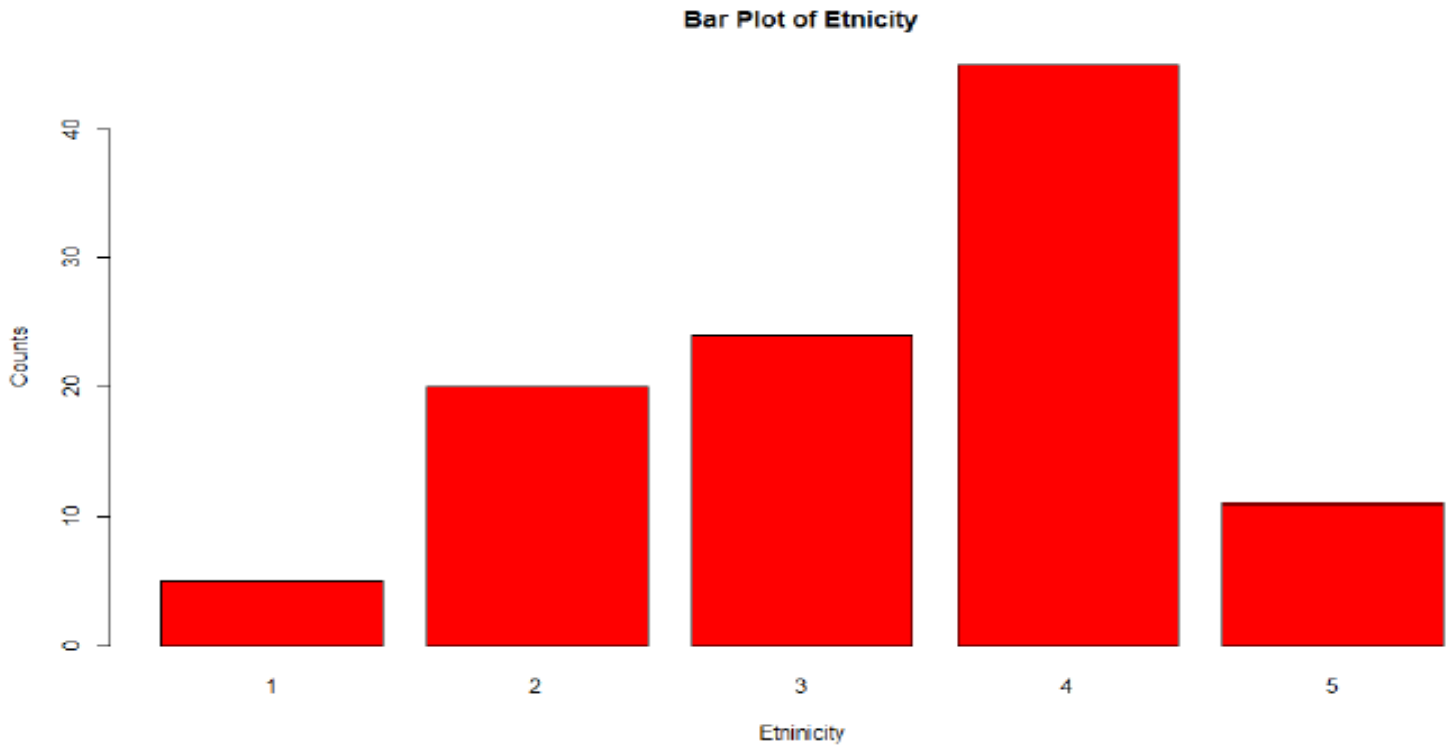
```
> plot(grades$gpa, grades$final, xlab = "gpa", ylab = "final", main
="Scatter Plot Final vs GPA", col="blue")

> abline(lm(grades$final~grades$gpa), col="red")
```



Bar Plot of ethnicity

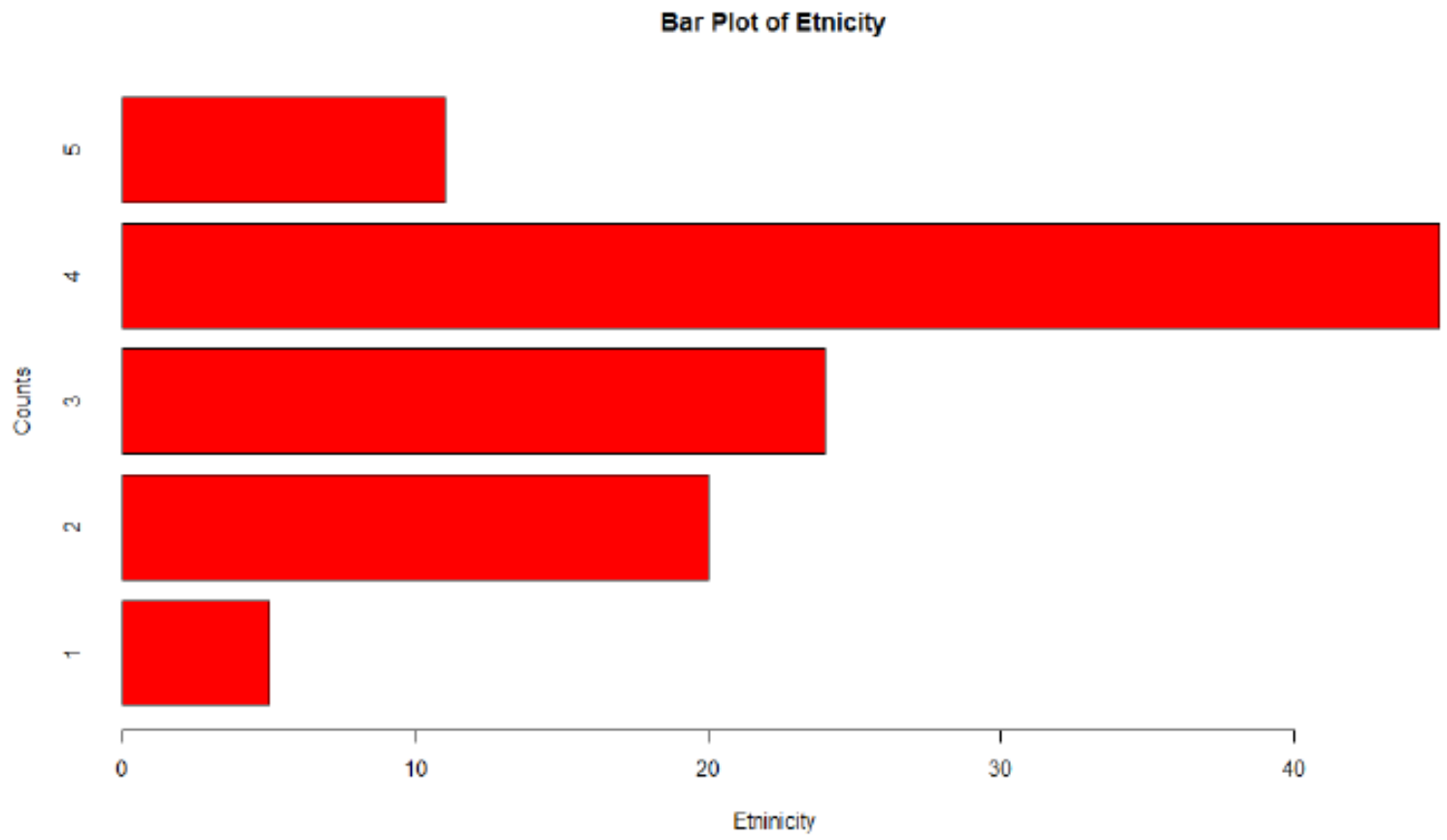
Package barplot in available in base packages

```
> grades <- read.csv("C:/Users/inurture1/Desktop/Datafiles/grades.csv")
>   View(grades)
> counts<-table(grades$ethnicity)
> barplot(counts, main = "Bar Plot of Etnicity", xlab="Etninicity", ylab =
"Counts", col = "red")
```



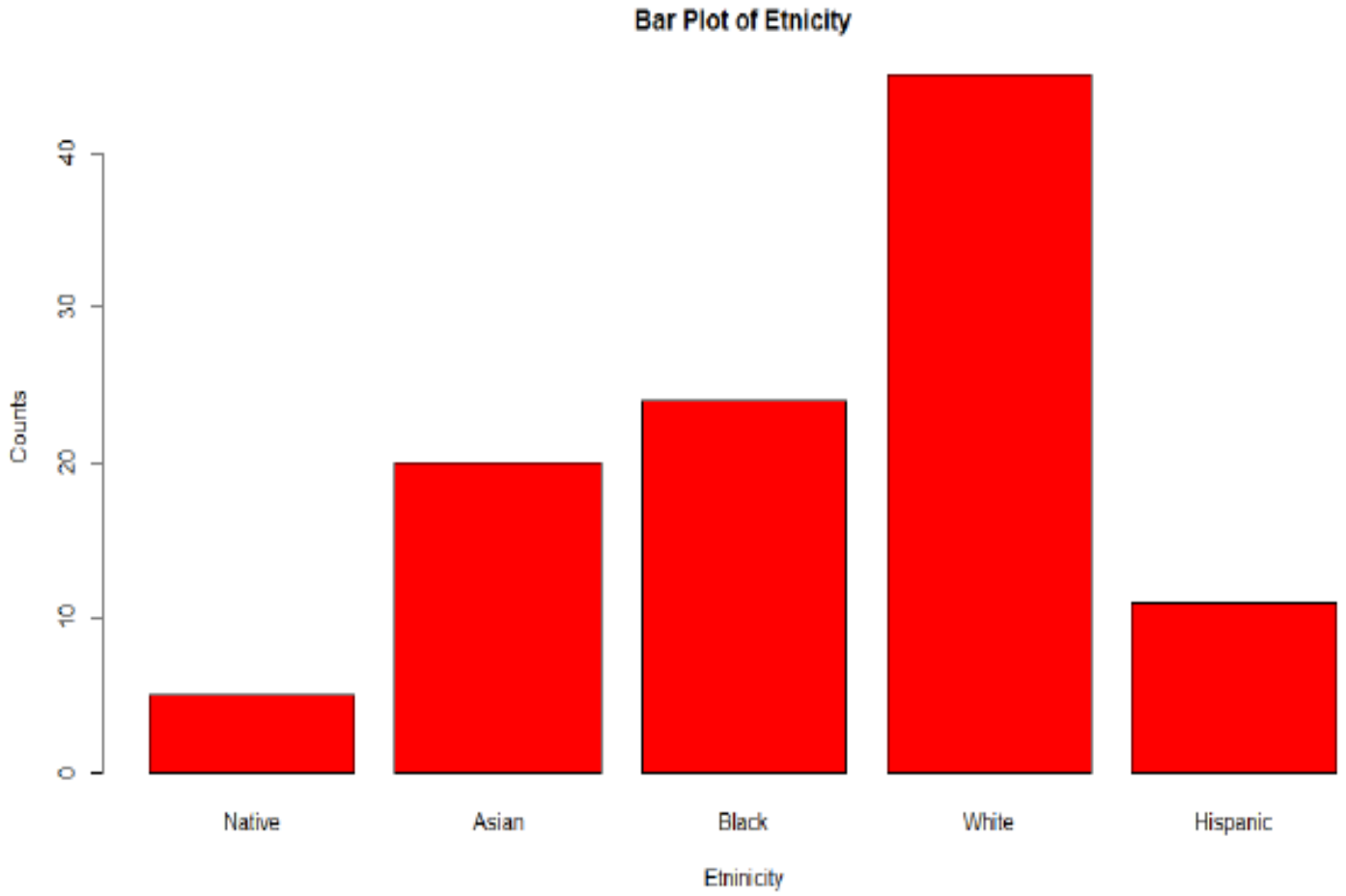
For making plot horizontal

```
> barplot(counts, main = "Bar Plot of Etnicity", xlab="Etninicity", horiz
= T, ylab = "Counts", col = "red")
```



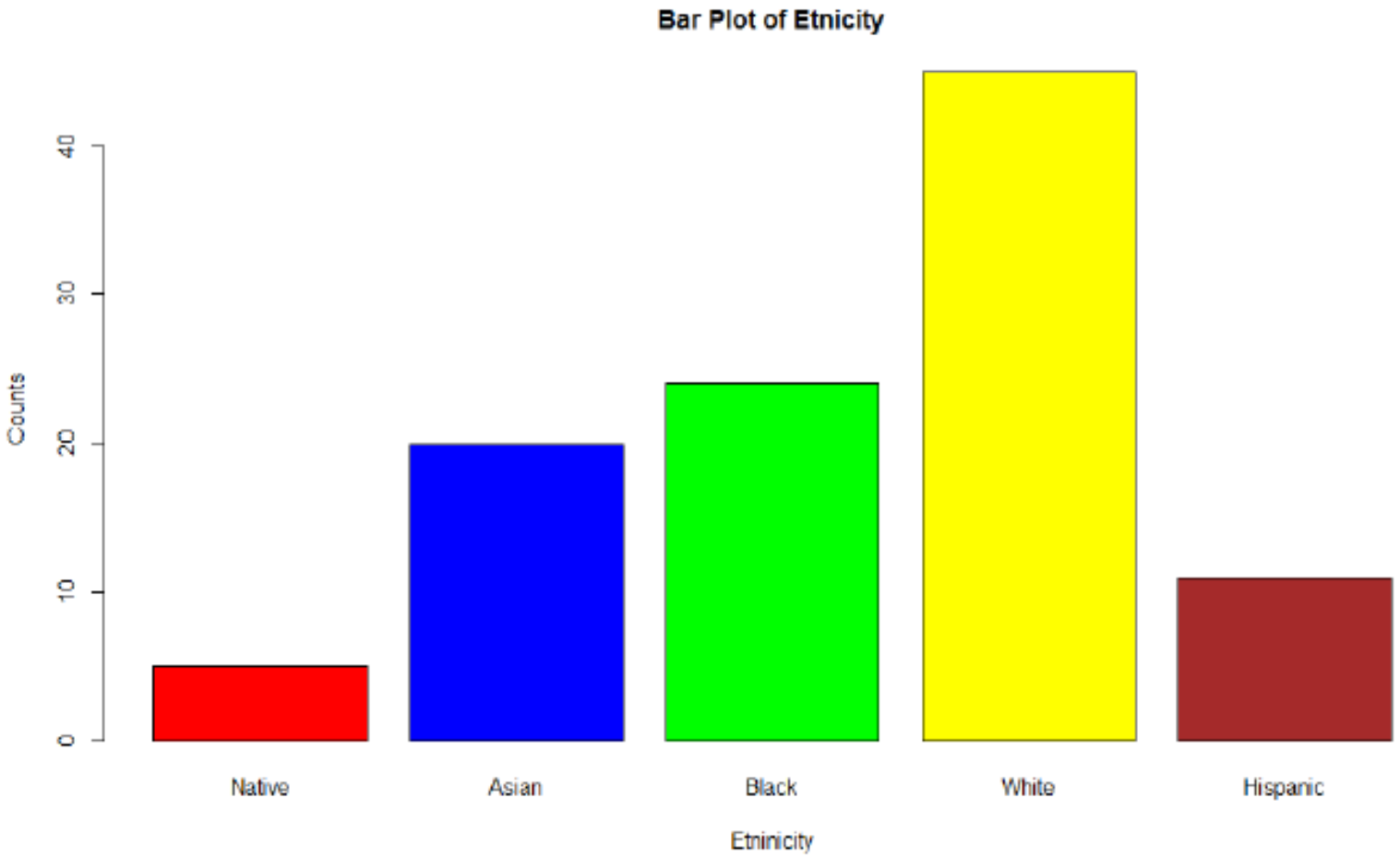
Label 1, 2, 3, 4 and 5 as Native, Asian, Black, White & Hispanic respectively

```
> barplot(counts, main = "Bar Plot of Etnicity", xlab="Etninicity", ylab =
"Counts", col = "red", names.arg = c("Native", "Asian", "Black", "White",
"Hispanic"))
```



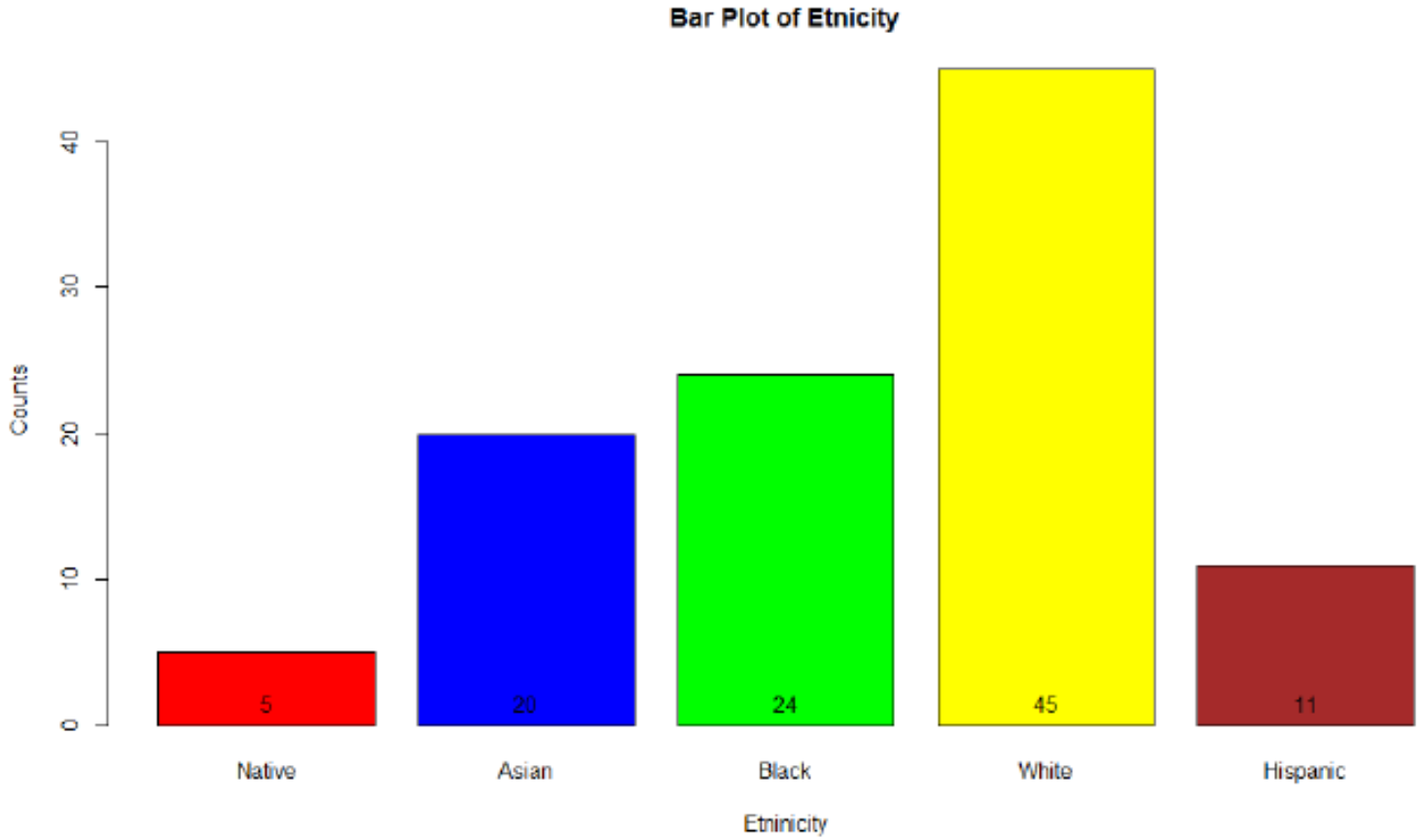
Different colours for each bar

```
> barplot(counts, main = "Bar Plot of Etnicity", xlab="Etninicity", ylab =
"Counts", col = c("red", "blue", "green", "yellow", "brown"), names.arg =
c("Native", "Asian", "Black", "White", "Hispanic"))
```



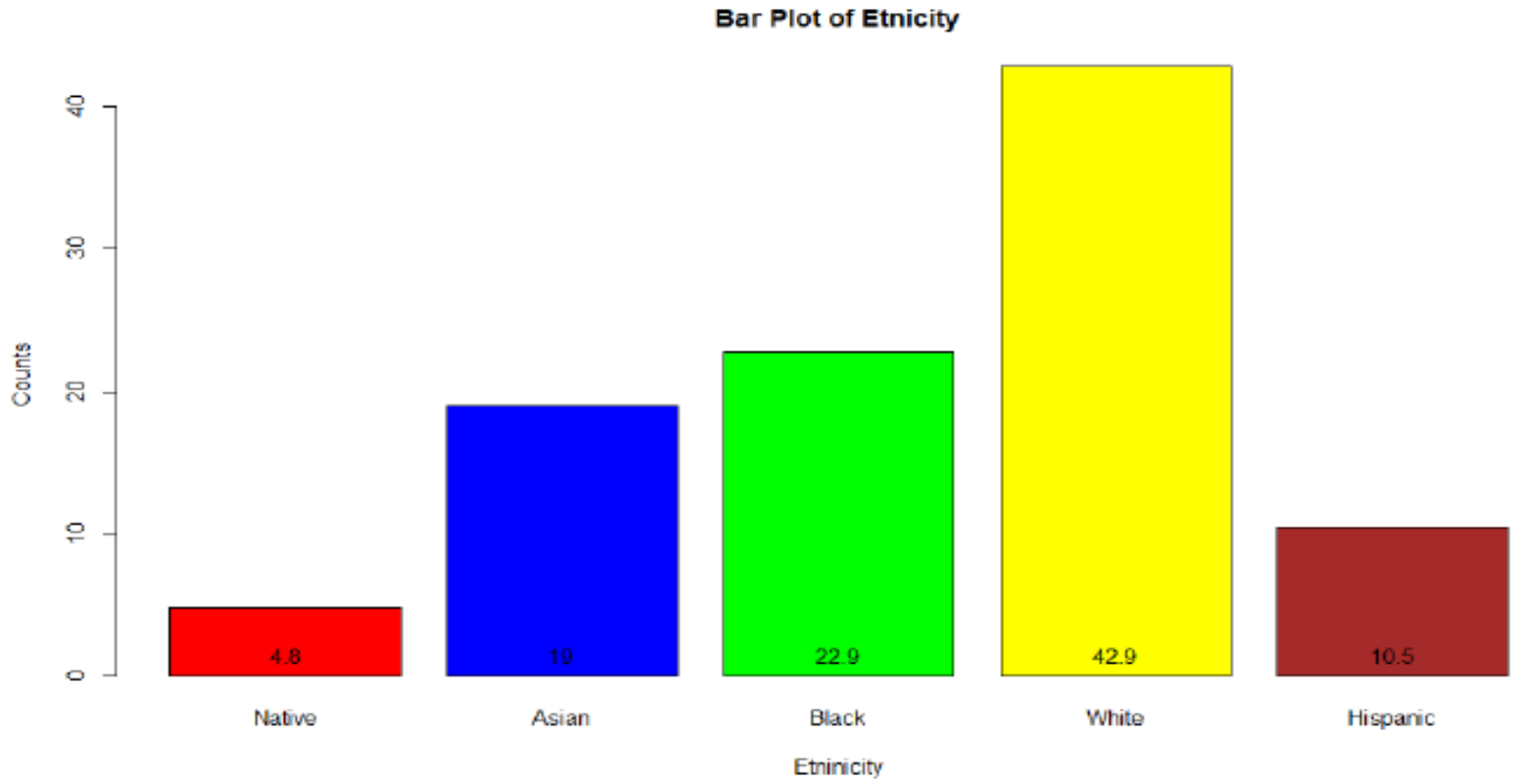
Insert Counts within bars

```
> bp<-barplot(counts, main = "Bar Plot of Etnicity", xlab="Etninicity",
ylab = "Counts", col = c("red", "blue", "green", "yellow", "brown"),
names.arg = c("Native", "Asian", "Black", "White", "Hispanic"))

> text(bp,0,counts, cex=1,pos=3)
```



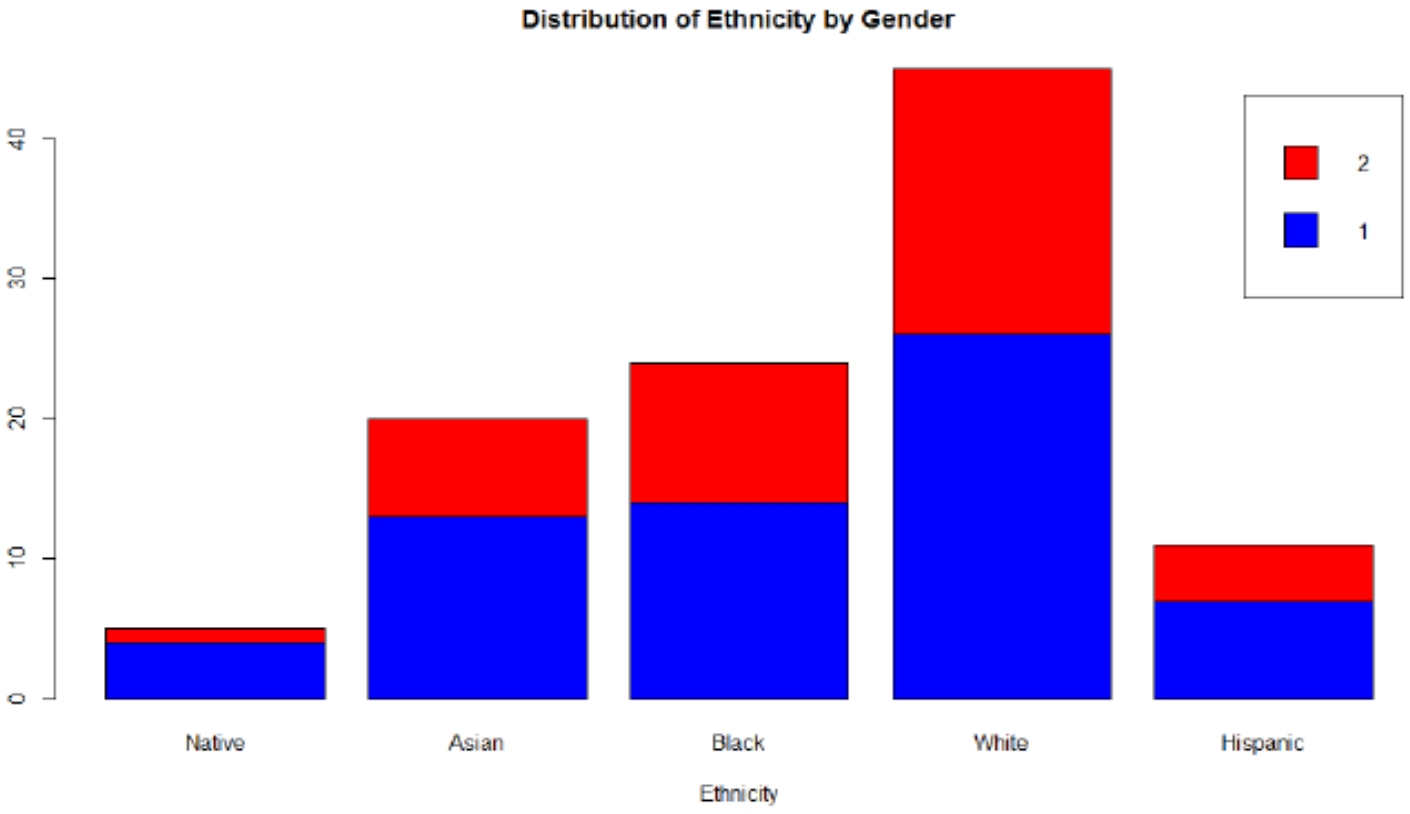
Insert Percentages within bars

```
> prop.counts<-counts/(sum(counts))*100
> bp<-barplot(prop.counts, main = "Bar Plot of Etnicity",
xlab="Etninicity", ylab = "Counts", col = c("red", "blue", "green",
"yellow", "brown"), names.arg = c("Native", "Asian", "Black", "White",
"Hispanic"))
> text(bp,0,round(prop.counts,1), cex =1, pos = 3)
```



Stacked Bar Chart ethnicity versus gender

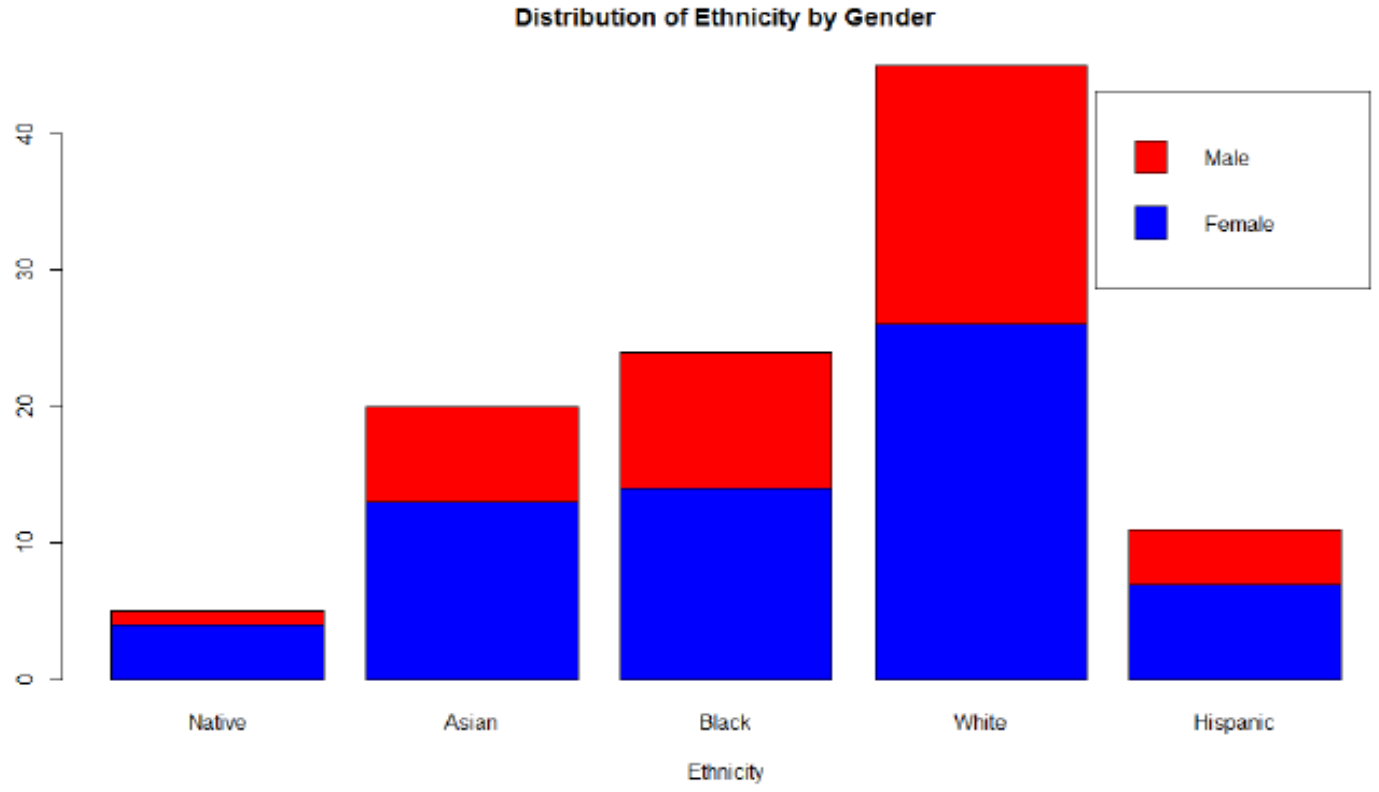
```
> counts<-table(grades$gender, grades$ethnicity)
> barplot(counts, main = "Distribution of Ethnicity by Gender",
xlab="Ethnicity", col=c("blue", "red"), legend = rownames(counts),
names.arg = c("Native", "Asian", "Black", "White", "Hispanic"))
```



Stacked Bar Chart ethnicity versus gender

(Give names Female to 1 and Male to 2)

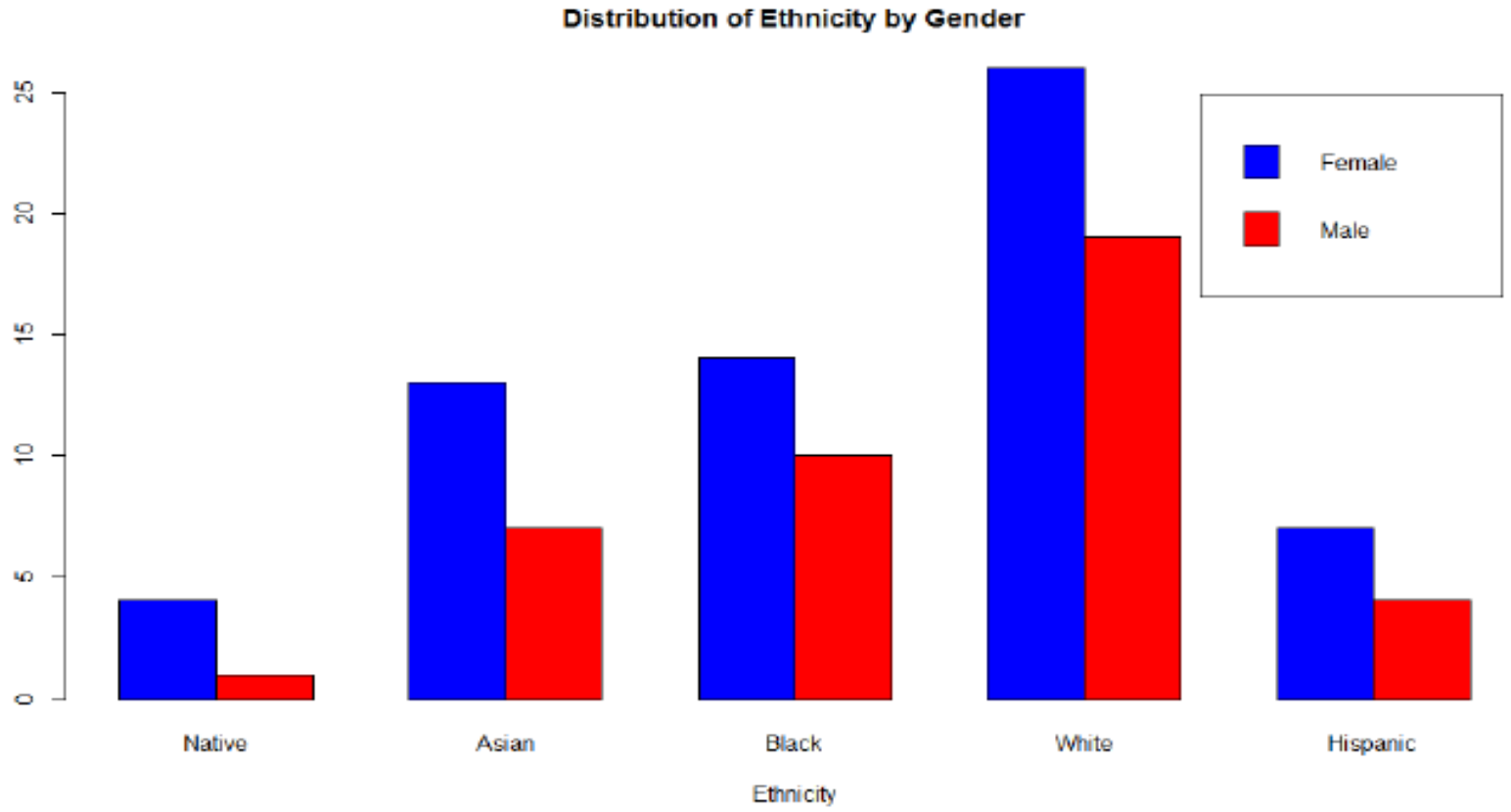
```
> barplot(counts, main = "Distribution of Ethnicity by Gender",
xlab="Ethnicity", col=c("blue", "red"), legend = c("Female", "Male"),
names.arg = c("Native", "Asian", "Black", "White", "Hispanic"))
```



Grouped Bar Chart ethnicity versus gender

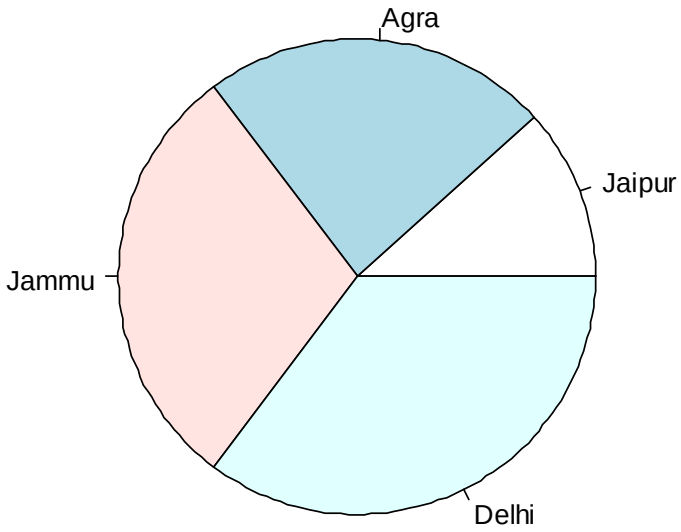
(Give names Female to 1 and Male to 2)

```
> counts<-table(grades$gender, grades$ethnicity)
> barplot(counts, main = "Distribution of Ethnicity by Gender",
xlab="Ethnicity", col=c("blue", "red"), legend = c("Female", "Male"),
names.arg = c("Native", "Asian", "Black", "White", "Hispanic"), beside=T)
```



Pie Chart

```
prashant<- c(10, 20, 25, 30)
names(prashant)<-c("Jaipur", "Agra", "Jammu", "Delhi")
pie(prashant)
```



Practice Assignment

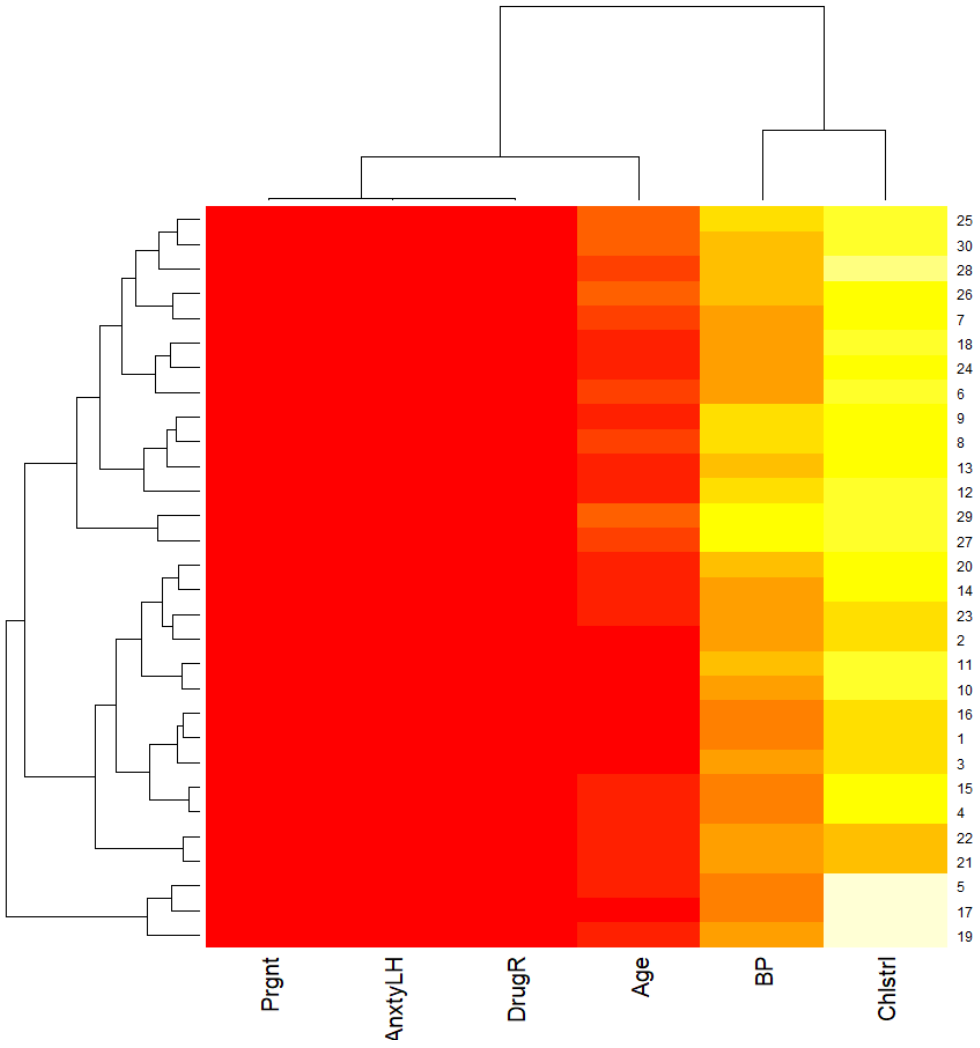
- Data File (grades.xls, CS2M.xls)
  - Convert this in csv
  - Read in the R-system
  - Do descriptive analysis (both tabular and graphical)

Covered

- Descriptive Analysis
  - Basic statistics, summary and describe (install a package in R)
  - Correlations
  - Pairs.panels

  Graphical

  Hist, scatter, Bar, Pie, Box, Heatmaps



Simple Heat Map

Red = 0, White is High

Data should be in matrix form

```
#simple heat map
cs2m<- as.matrix(cs2m)
heatmap(cs2m, scale = 'none')
```

	BP	Chlstrl	Age	Prgnt	AnxtyLH	DrugR
1	100	150	20	0	0	0
2	120	160	16	0	0	0
3	110	150	18	0	0	0
4	100	175	25	0	0	0
5	95	250	36	0	0	0