DDC ONLINE TRAINING SESSION:

DATA REPRESENTATION / EXPLORING DATA

DATA VISUALIZATION

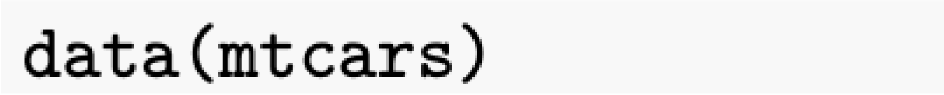
16/02/2023

CBS COPENHAGEN BUSINESS SCHOOL

HANDELSHØJSKOLEN

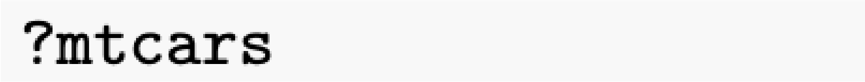
Loading the datasets

1. Load the dataset:



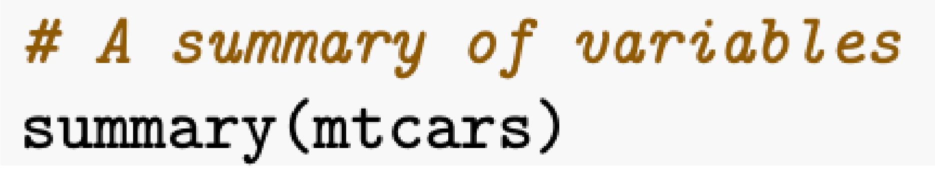
```
data(mtcars)
```

2. Query the documentation for the datasets included in packages:



```
?mtcars
```

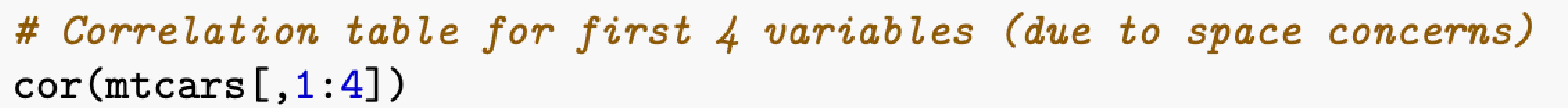
3. Get a sense of the data by visualizing a summary of the variables:



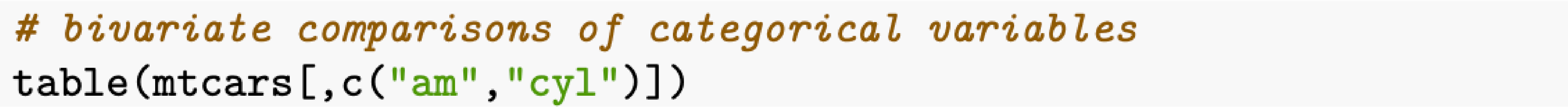
```
# A summary of variables
summary(mtcars)
```

Datasets' statistics

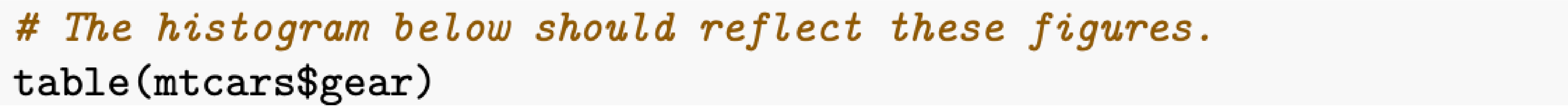
1. Display the correlation table (limited to the first 4 variables for space concern):
2. Display the bivariate comparisons of categorical data:



```
# Correlation table for first 4 variables (due to space concerns)
cor(mtcars[,1:4])
```



```
# bivariate comparisons of categorical variables
table(mtcars[,c("am","cyl")])
```

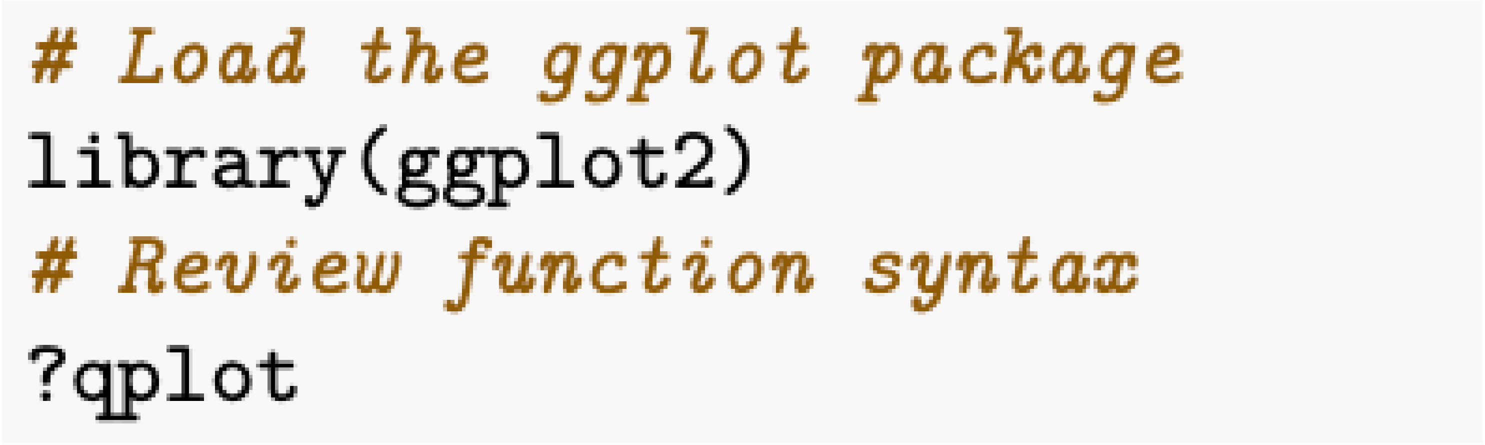


```
# The histogram below should reflect these figures.
table(mtcars$gear)
```

Plotting with qplot()

1. Load the ggplot package:

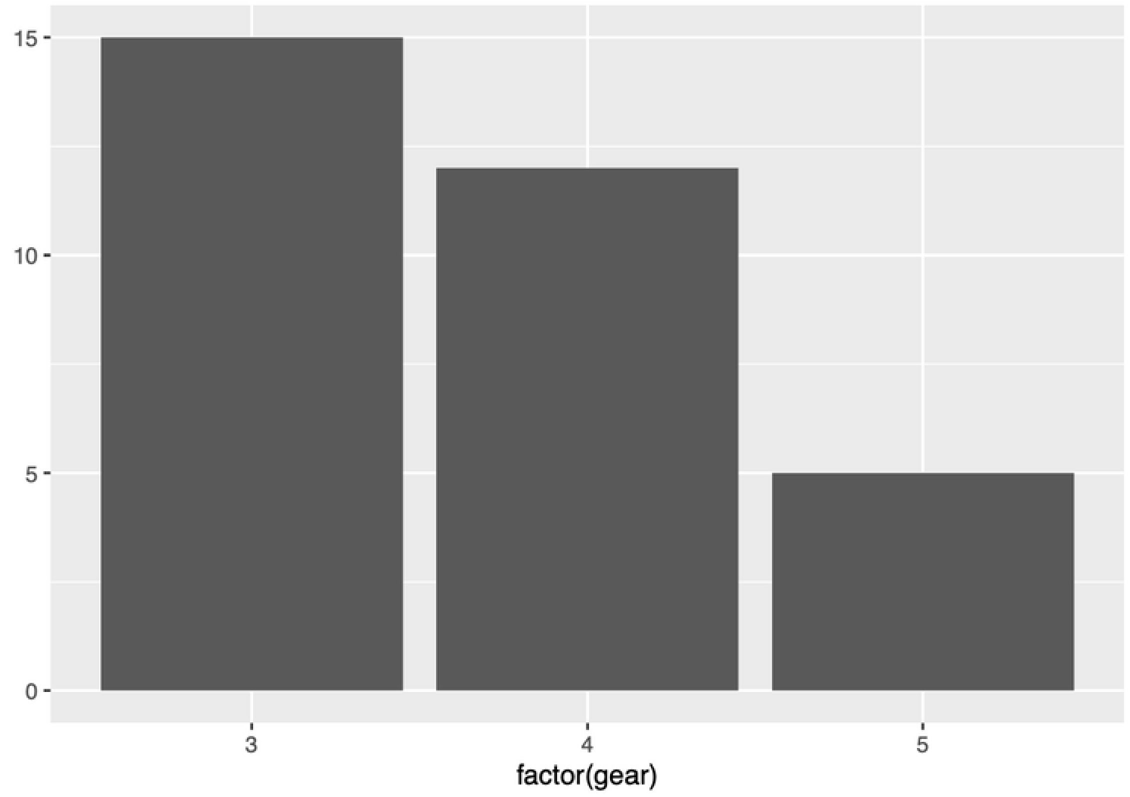
2. Review function syntax:

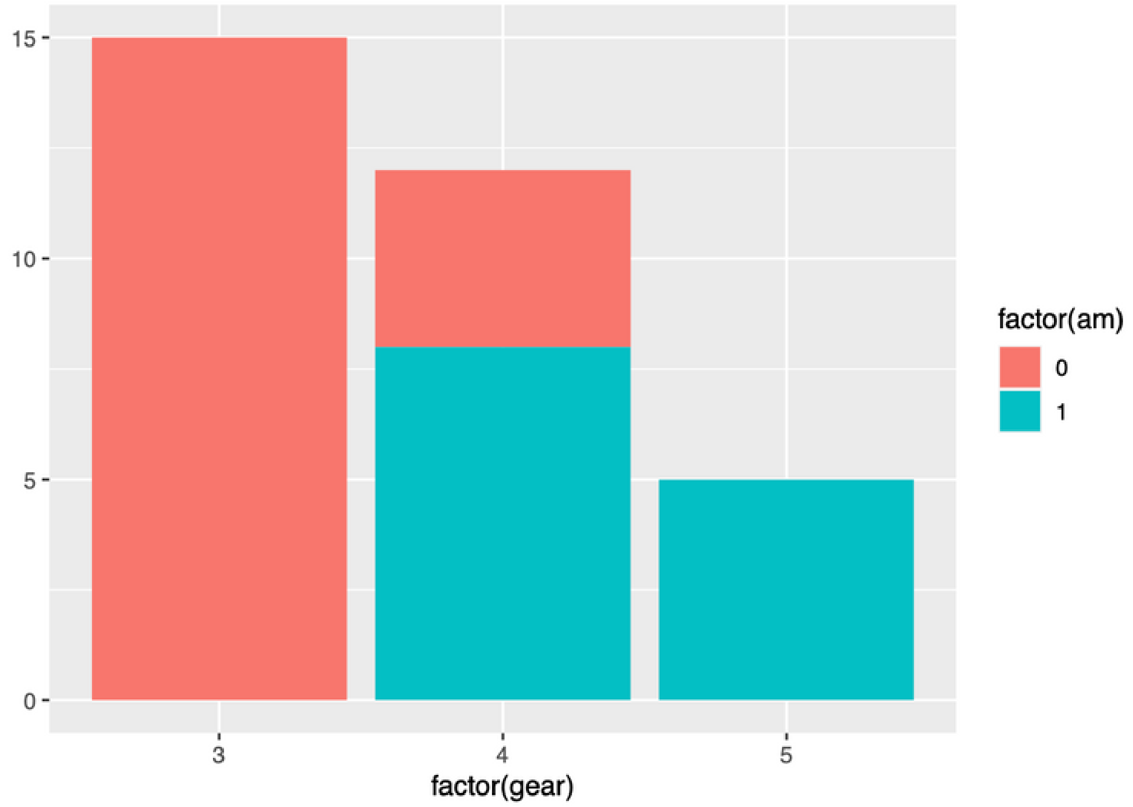


```
# Load the ggplot package
library(ggplot2)
# Review function syntax
?qplot
```

Histograms

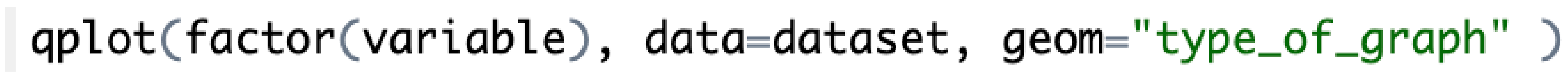
Histograms are generally used to display frequencies of certain categories, like number of people with different eye colors.





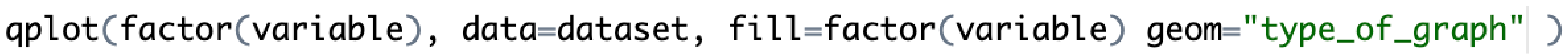
Histograms - Exercise

TASK 1:
report the number of cars with differing number of front gears



```
qplot(factor(variable), data=dataset, geom="type_of_graph" )
```

TASK 2:
report the number of cars with differing number of front gears and color the bars based on another variable.

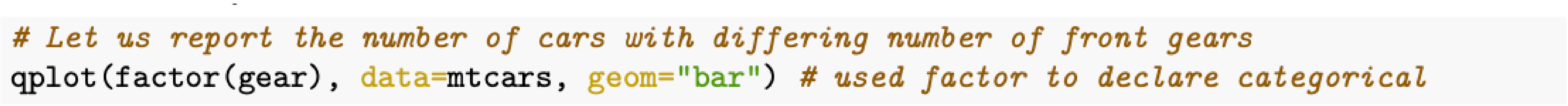


```
qplot(factor(variable), data=dataset, fill=factor(variable) geom="type_of_graph" )
```

Histograms - Exercise solved

TASK 3:

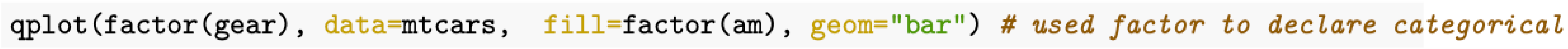
report the number of cars with differing number of front gears



```
# Let us report the number of cars with differing number of front gears
qplot(factor(gear), data=mtcars, geom="bar") # used factor to declare categorical
```

TASK 4:

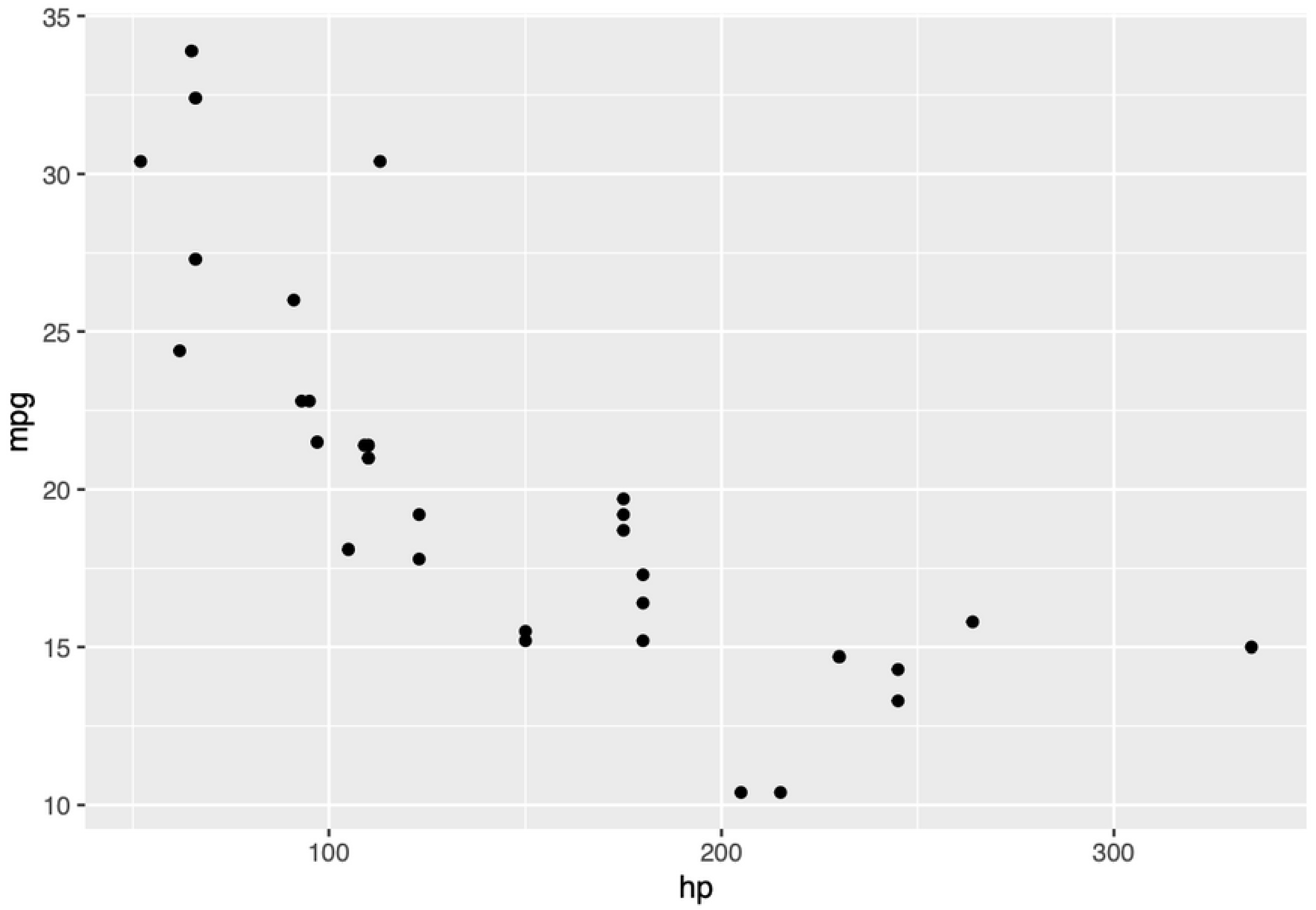
report the number of cars with differing number of front gears and color the bars based on another variable.

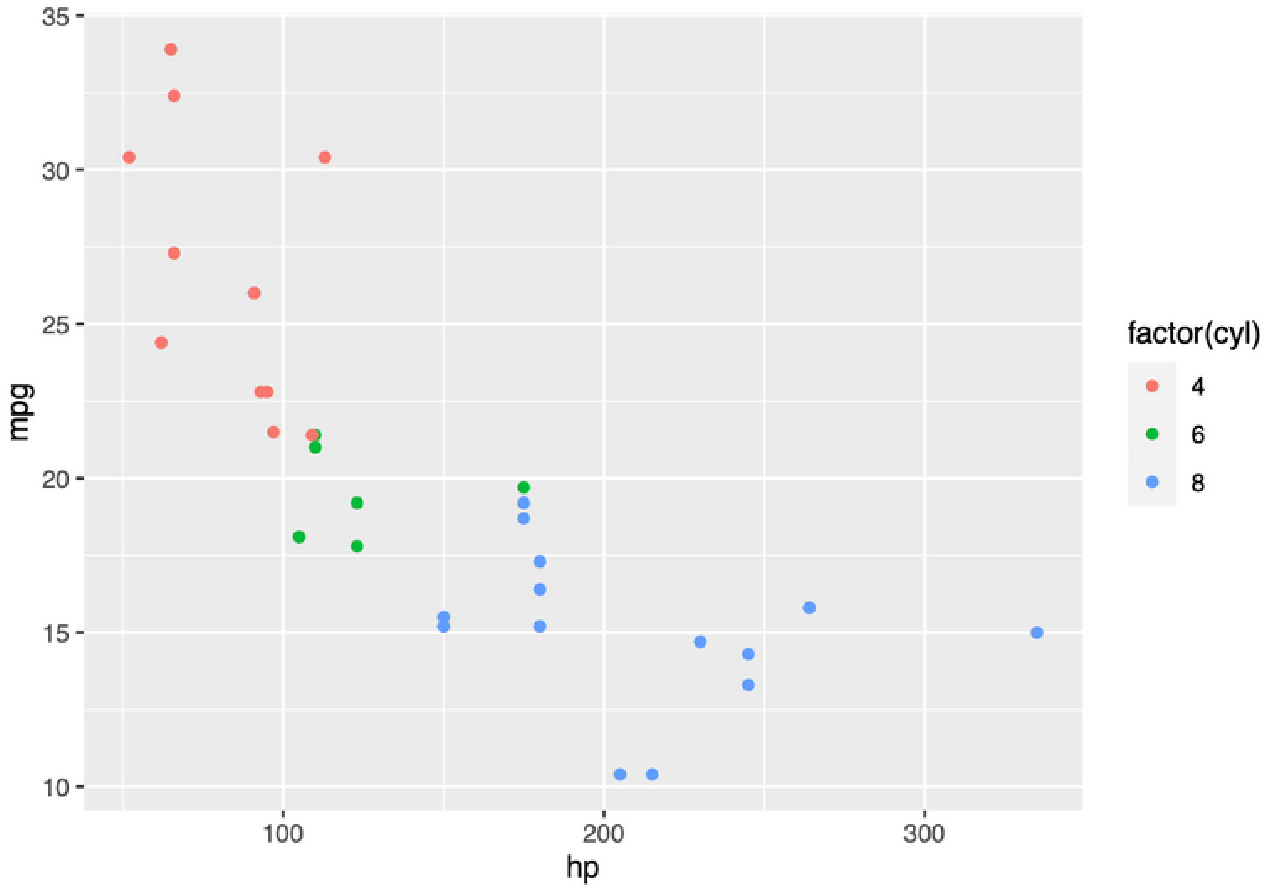


```
qplot(factor(gear), data=mtcars,  fill=factor(am), geom="bar") # used factor to declare categorical
```

Scatter plots

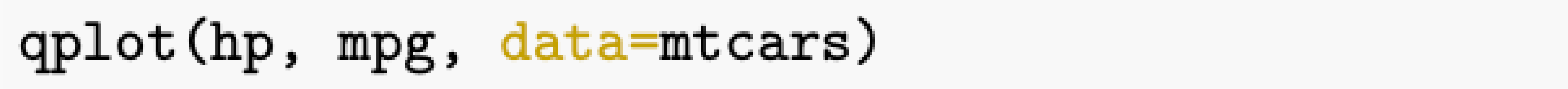
Scatter plots are generally used to display the relationship between two continuous variables.





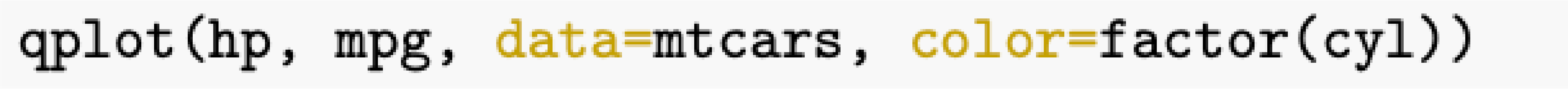
Scatter plots - Exercise

TASK 5:
Plot the relationship between mpg and hp:



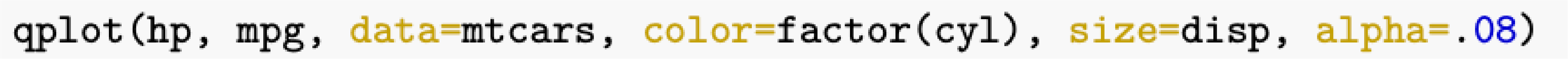
```
qplot(hp, mpg, data=mtcars)
```

TASK 6:
Plot the relationship between mpg and hp, and color the dots by the number of cylinders:



```
qplot(hp, mpg, data=mtcars, color=factor(cyl))
```

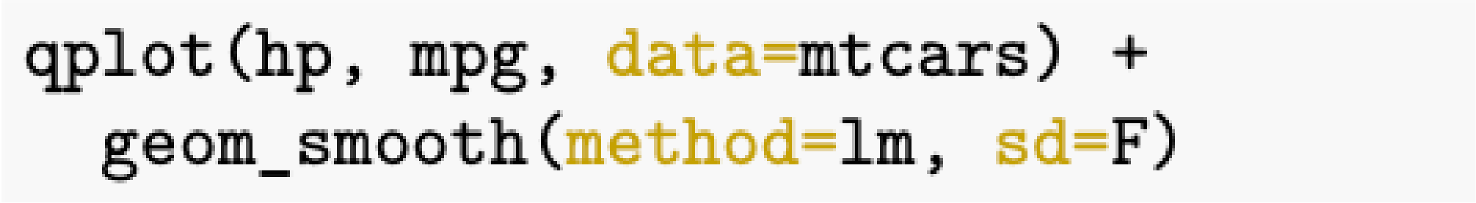
TASK 7:
Plot the relationship between mpg and hp, and color the dots by the number of cylinders and the size of dots dependent on a continuous variable (displacement).



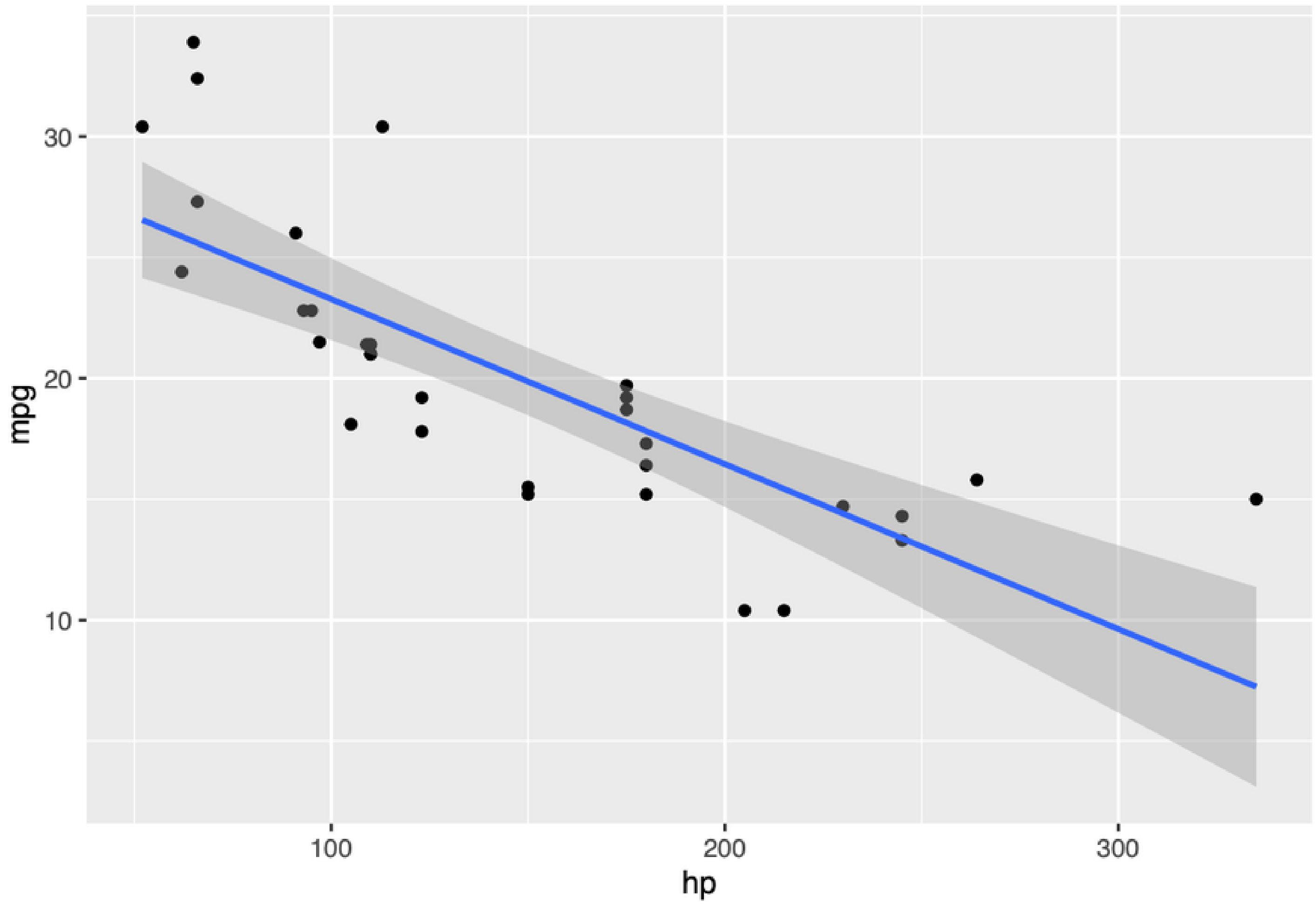
```
qplot(hp, mpg, data=mtcars, color=factor(cyl), size=disp, alpha=.08)
```

Regression line

TASK 8:
Let us fit a regression line:



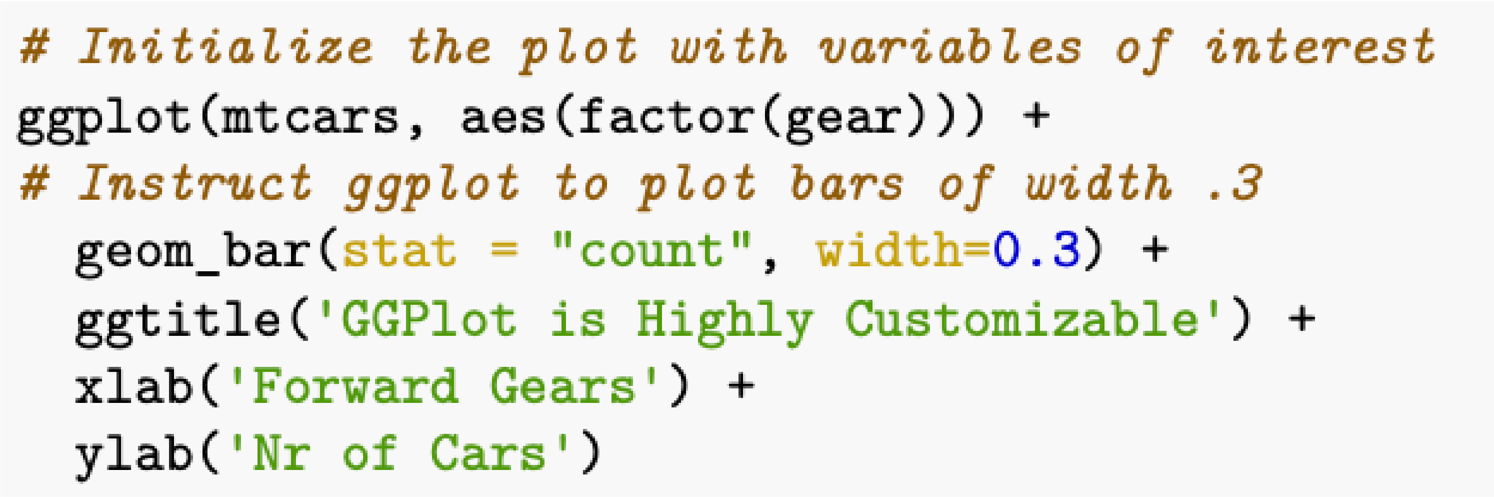
```
qplot(hp, mpg, data=mtcars) +
  geom_smooth(method=lm, sd=F)
```



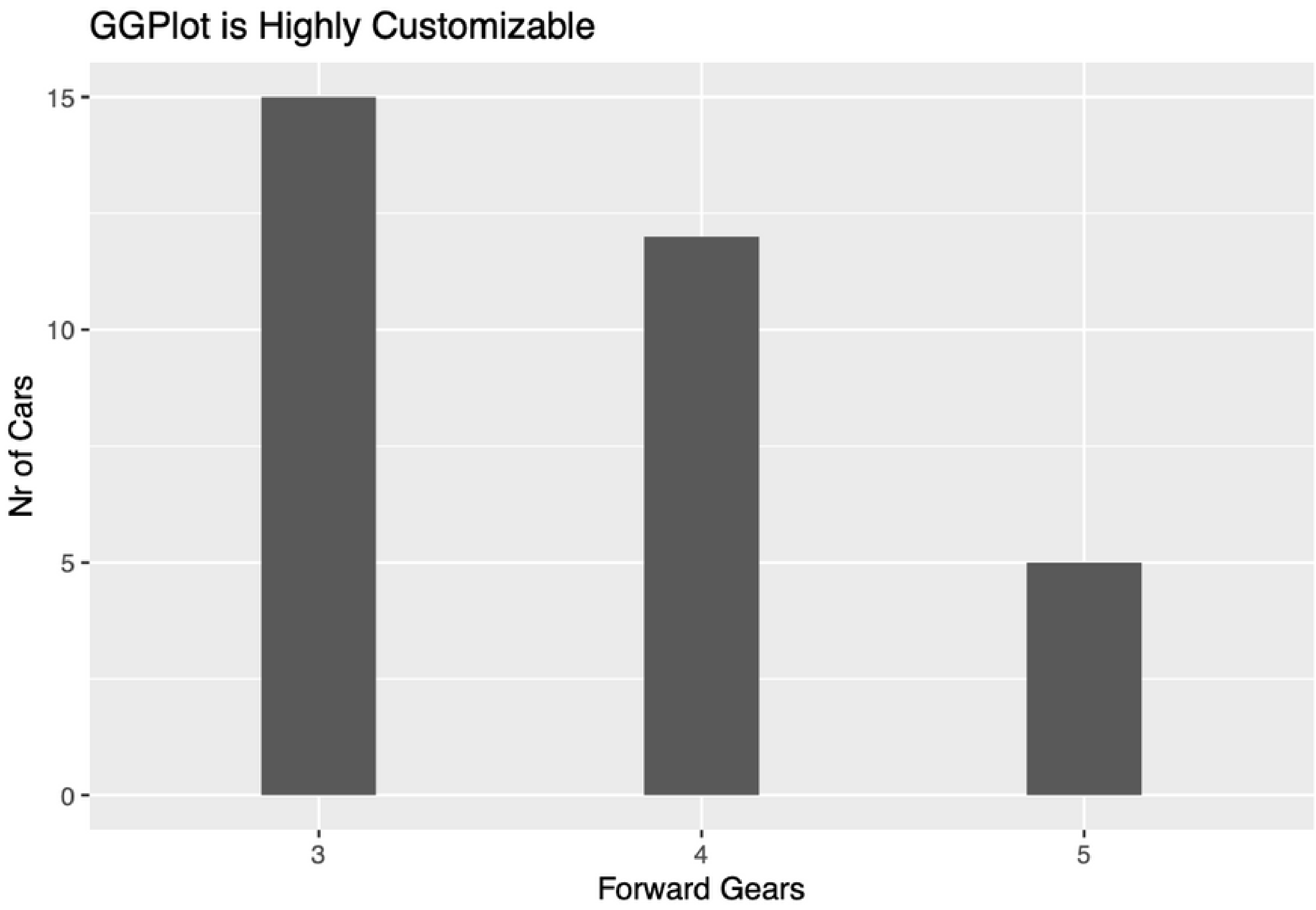
ggplot -histogram

TASK 9:

Plot the number of cars with differing number of front gears

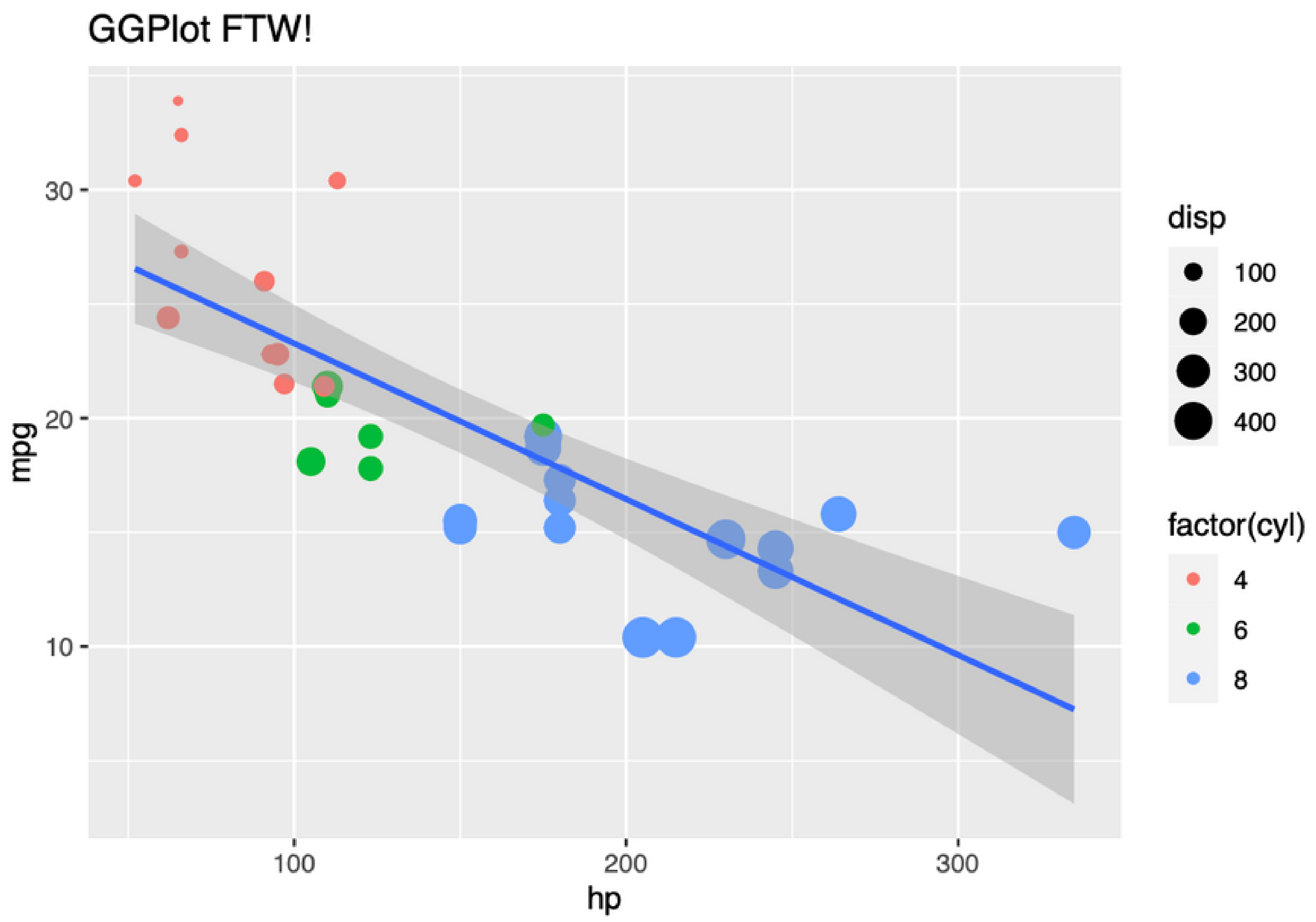


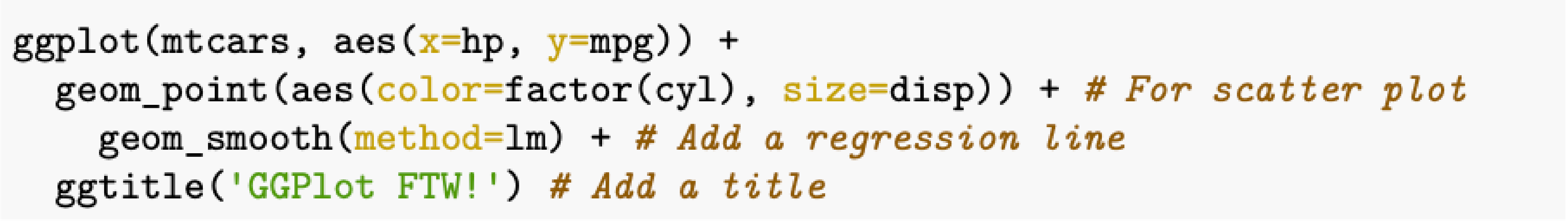
```
# Initialize the plot with variables of interest
ggplot(mtcars, aes(factor(gear))) +
# Instruct ggplot to plot bars of width .3
  geom_bar(stat = "count", width=0.3) +
  ggtitle('GGPlot is Highly Customizable') +
  xlab('Forward Gears') +
  ylab('Nr of Cars')
```



ggplot - scatter plot

TASK 10:
Plot the relationship between mpg and hp, and color the dots by the number of cylinders and the size of dots dependent on a continuous variable (displacement) as in task 7:

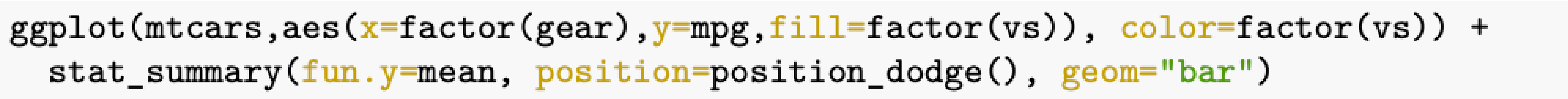




```
ggplot(mtcars, aes(x=hp, y=mpg)) +
  geom_point(aes(color=factor(cyl), size=disp)) + # For scatter plot
    geom_smooth(method=lm) + # Add a regression line
  ggtitle('GGPlot FTW!') # Add a title
```

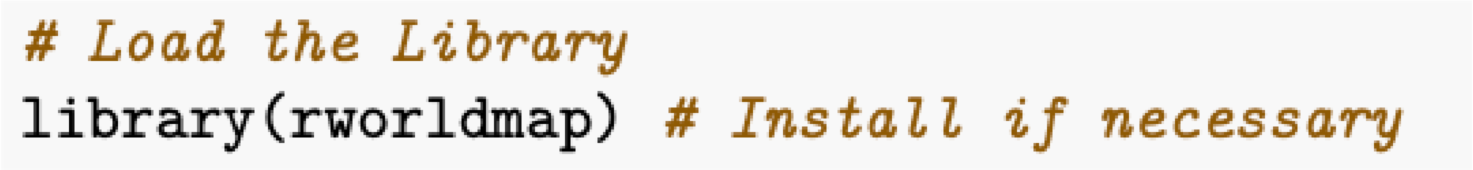
ggplot - bar charts

TASK 11:
Plot the relationship between mpg and two categorical variables:

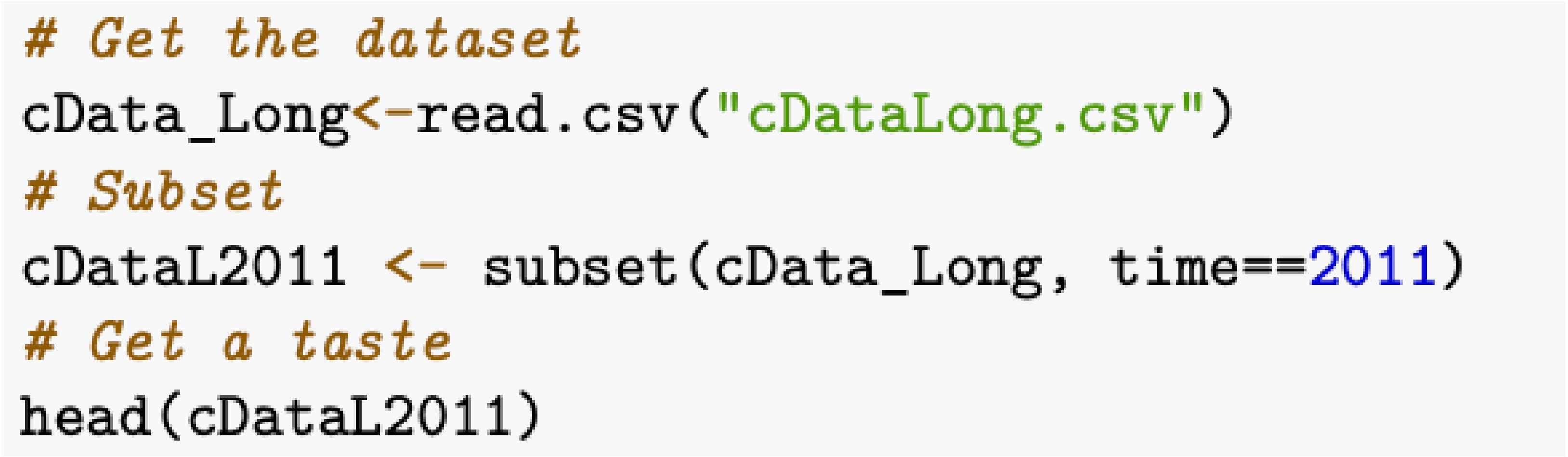


```
ggplot(mtcars,aes(x=factor(gear),y=mpg,fill=factor(vs)), color=factor(vs)) +
  stat_summary(fun.y=mean, position=position_dodge(), geom="bar")
```

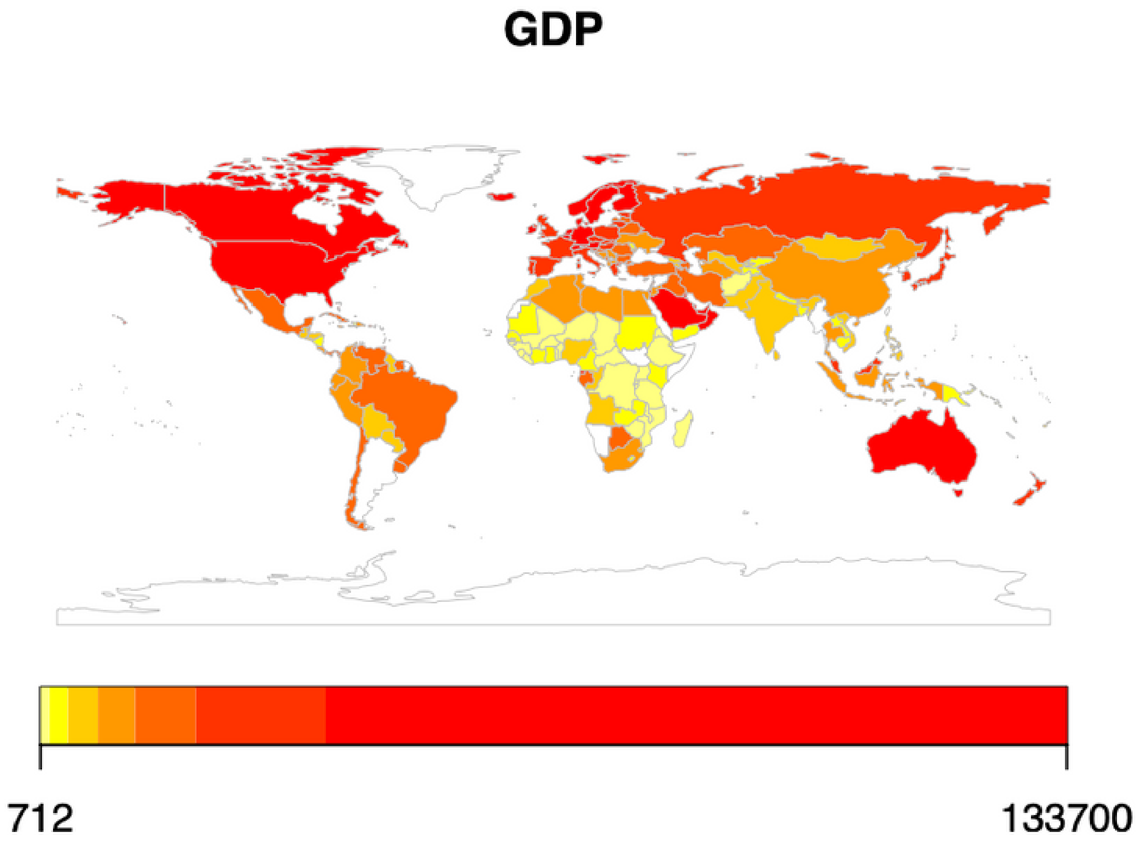
ggplot - World Map Visualizations

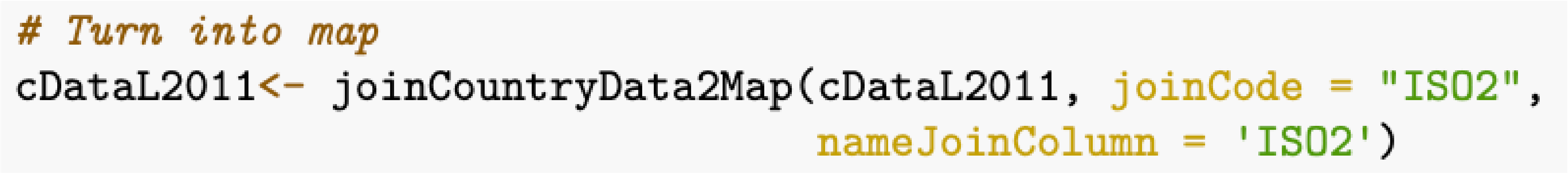


```r
# Load the Library
library(rworldmap) # Install if necessary
```

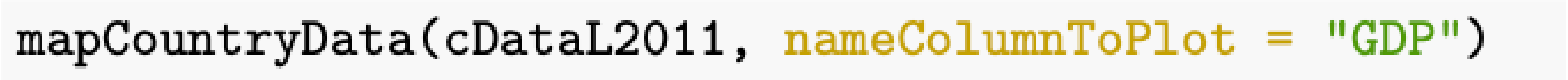


```r
# Get the dataset
cData_Long<-read.csv("cDataLong.csv")
# Subset
cDataL2011 <- subset(cData_Long, time==2011)
# Get a taste
head(cDataL2011)
```



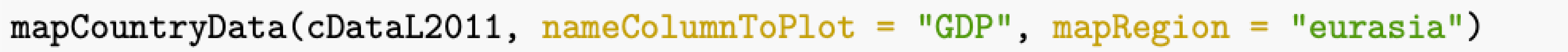


```r
# Turn into map
cDataL2011<- joinCountryData2Map(cDataL2011, joinCode = "ISO2",
                                 nameJoinColumn = 'ISO2')
```

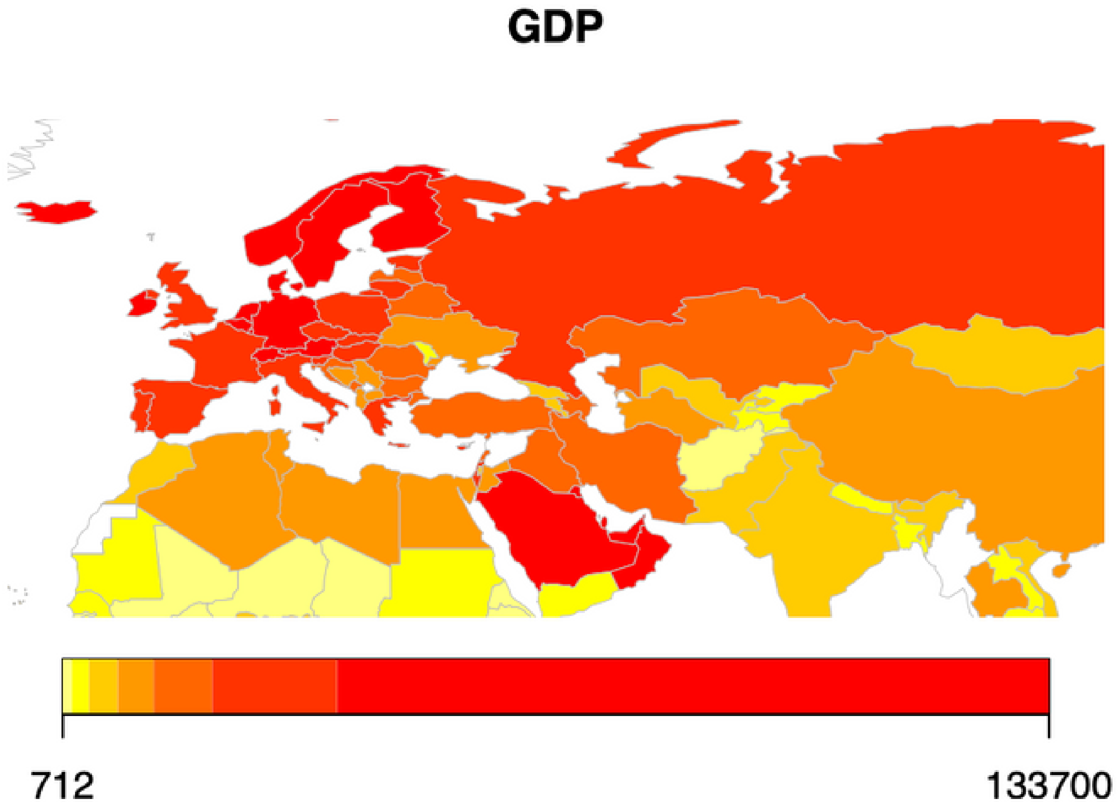


```r
mapCountryData(cDataL2011, nameColumnToPlot = "GDP")
```

ggplot - World Map Visualizations - Eurasia



```
mapCountryData(cDataL2011, nameColumnToPlot = "GDP", mapRegion = "eurasia")
```


GDP
712
133700