Project 3

Dipesh Poudel

12/2/2021

Project 3 - Data Visualization in R

Part 1: Data visualization with Base R graphics packages

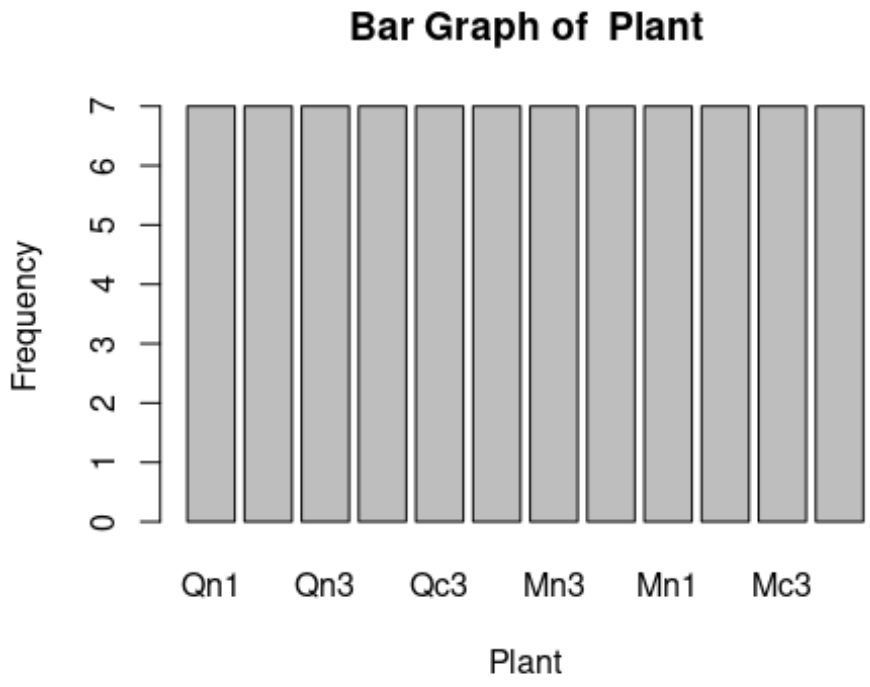
Use the built-in CO2 data and do as follows:

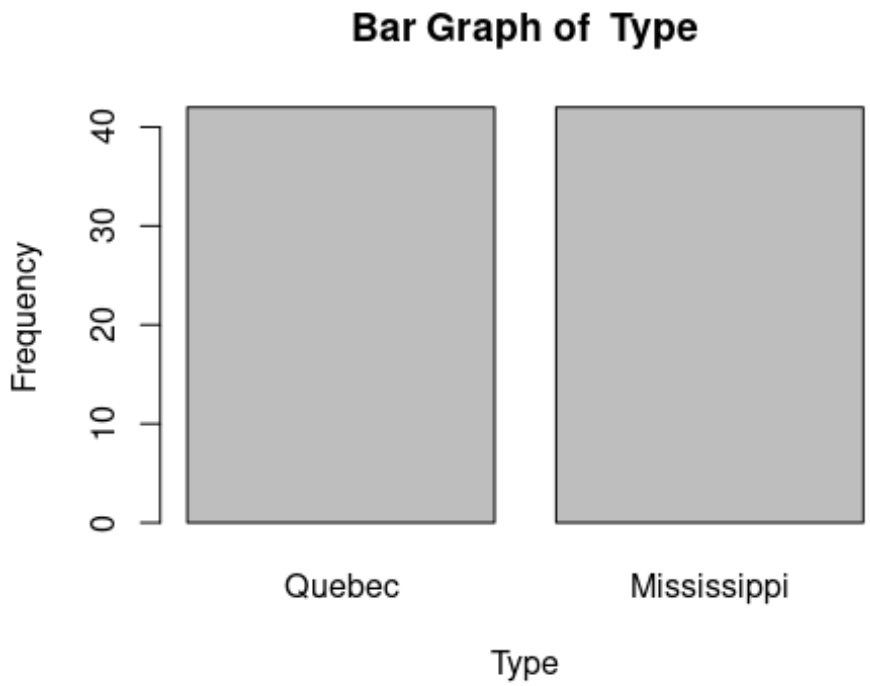
```
# Loading the CO2 data
co2_data<-CO2
str(co2_data)
```

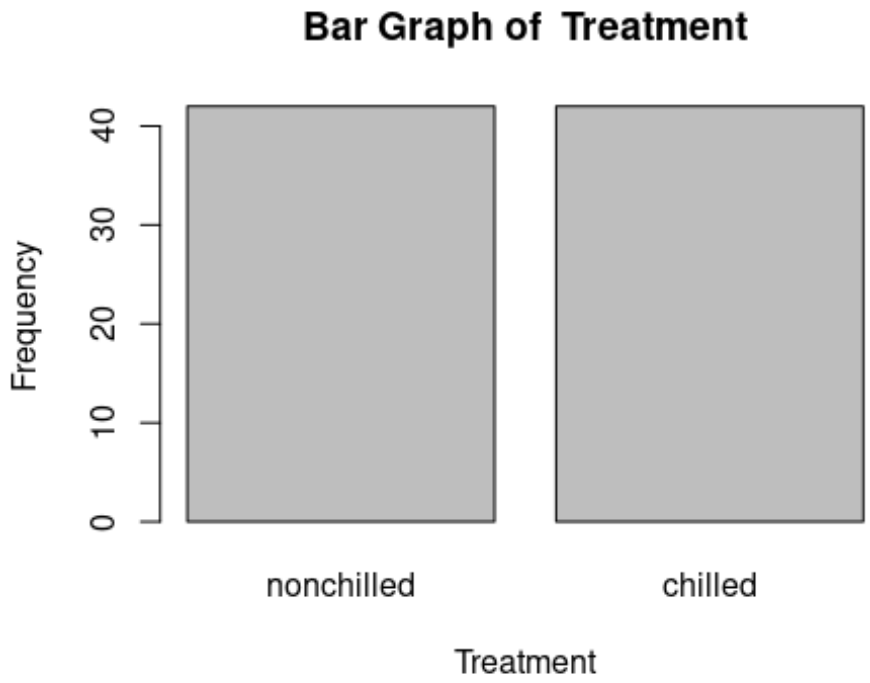
```
## Classes 'nfnGroupedData', 'nfGroupedData', 'groupedData' and
'data.frame':   84 obs. of  5 variables:
##  $ Plant    : Ord.factor w/ 12 levels "Qn1"<"Qn2"<"Qn3"<..: 1 1 1 1
1 1 1 2 2 2 ...
##  $ Type     : Factor w/ 2 levels "Quebec","Mississippi": 1 1 1 1 1
1 1 1 1 1 ...
##  $ Treatment: Factor w/ 2 levels "nonchilled","chilled": 1 1 1 1 1
1 1 1 1 1 ...
##  $ conc     : num  95 175 250 350 500 675 1000 95 175 250 ...
##  $ uptake   : num  16 30.4 34.8 37.2 35.3 39.2 39.7 13.6 27.3
37.1 ...
##  - attr(*, "formula")=Class 'formula'  language uptake ~ conc |
Plant
##   .. ..- attr(*, ".Environment")=<environment: R_EmptyEnv>
##  - attr(*, "outer")=Class 'formula'  language ~Treatment * Type
##   .. ..- attr(*, ".Environment")=<environment: R_EmptyEnv>
##  - attr(*, "labels")=List of 2
##   ..$ x: chr "Ambient carbon dioxide concentration"
##   ..$ y: chr "CO2 uptake rate"
##  - attr(*, "units")=List of 2
##   ..$ x: chr "(uL/L)"
##   ..$ y: chr "(umol/m^2 s)"
```

1. Create bar graph of plant, type and treatment variables

```
variables<-c('Plant','Type','Treatment')
for (var in variables){
barplot(table(co2_data[var]),main = paste("Bar Graph of ",var),
        xlab = var, ylab = "Frequency")
}
```

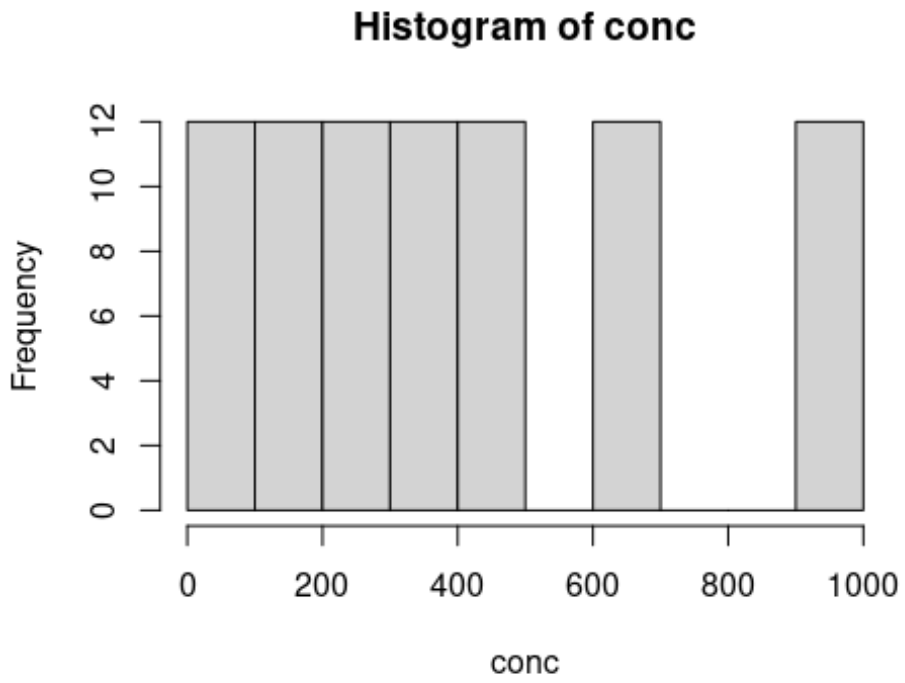

Bar Graph of Plant
Frequency
0
1
2
3
4
5
6
7
Qn1
Qn3
Qc3
Mn3
Mn1
Mc3
Plant


Bar Graph of Type
Frequency
0
10
20
30
40
Quebec
Mississippi
Type



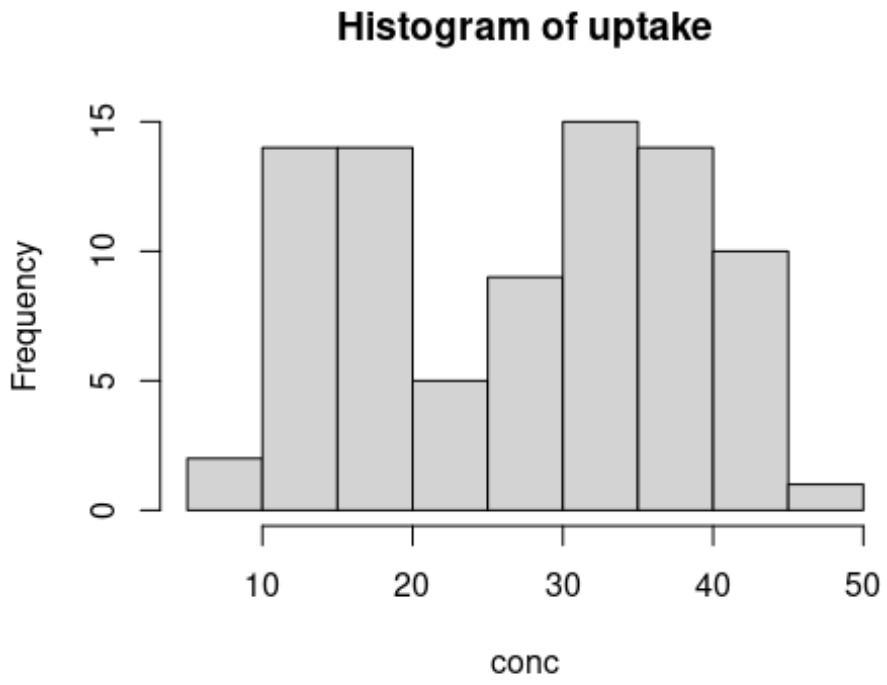
2. Create histogram of conc and uptake variables Histogram of conc variable

```
# Histogram of conc variable
hist(co2_data$conc,main = "Histogram of conc",xlab = "conc",ylab =
"Frequency")
```



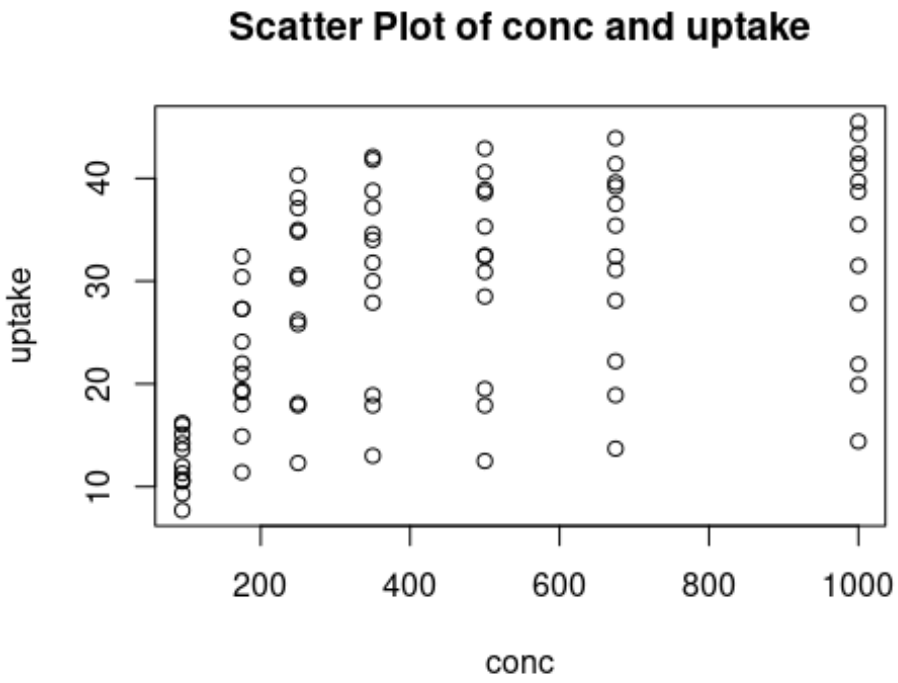
Histogram of uptake variable

```
# Histogram of uptake variable
hist(co2_data$uptake,main = "Histogram of uptake",xlab = "conc",ylab =
"Frequency")
```



3. Create scatterplot of conc and uptake variables

```
plot(co2_data$conc,co2_data$uptake,main = "Scatter Plot of conc and
uptake",
     xlab = "conc",ylab = "uptake")
```



4. Which measure of association is suitable for conc and uptake variables Since the relationship is not linear we have to use spearman correlation for association 5. Compute the best correlation coefficient for conc and uptake variables and interpret the result carefully.

```
cor(co2_data$conc,co2_data$uptake,method  = c("spearman"))
```

```
## [1] 0.5800041
```

Since the correlation coefficient is positive and greater than 0 we can say that the as the conc increases the uptake tends to increase but not in linear way.

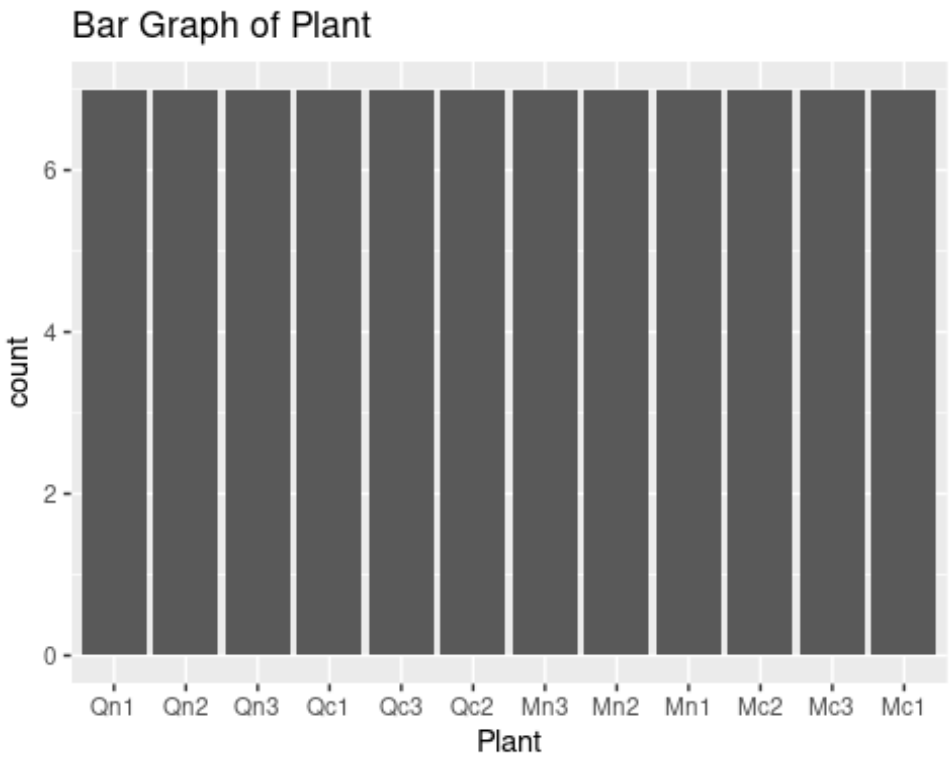
Part 2: Data visualization with ggplot2 package

Use the built-in CO2 data and do as follows:

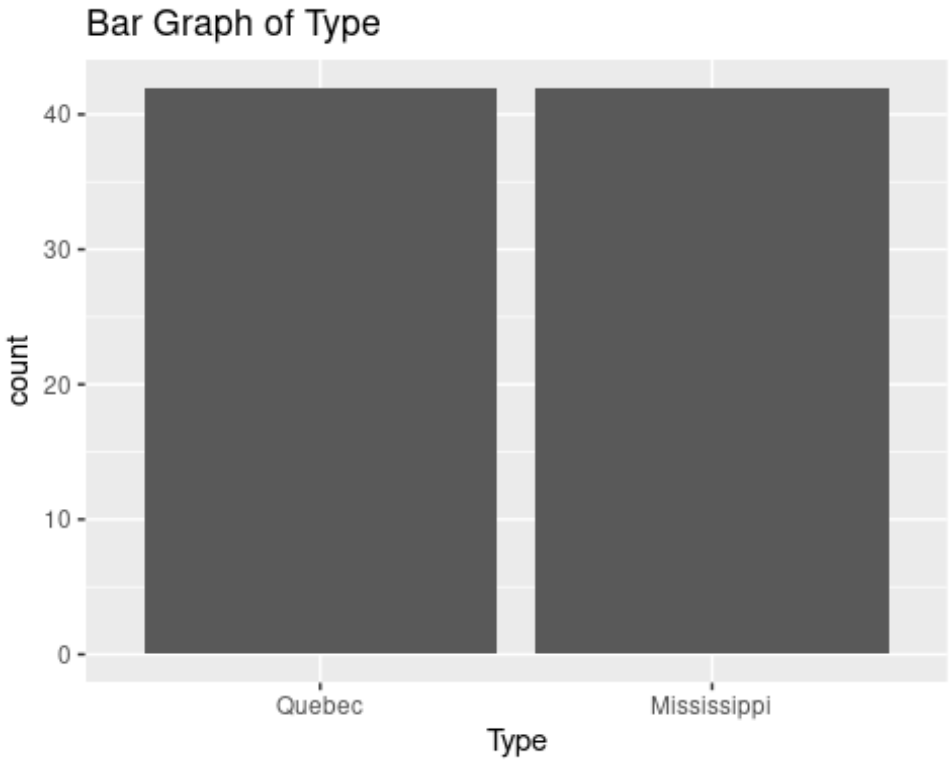
```
library(ggplot2)
```

1. Create bar graph of plant, type and treatment variables

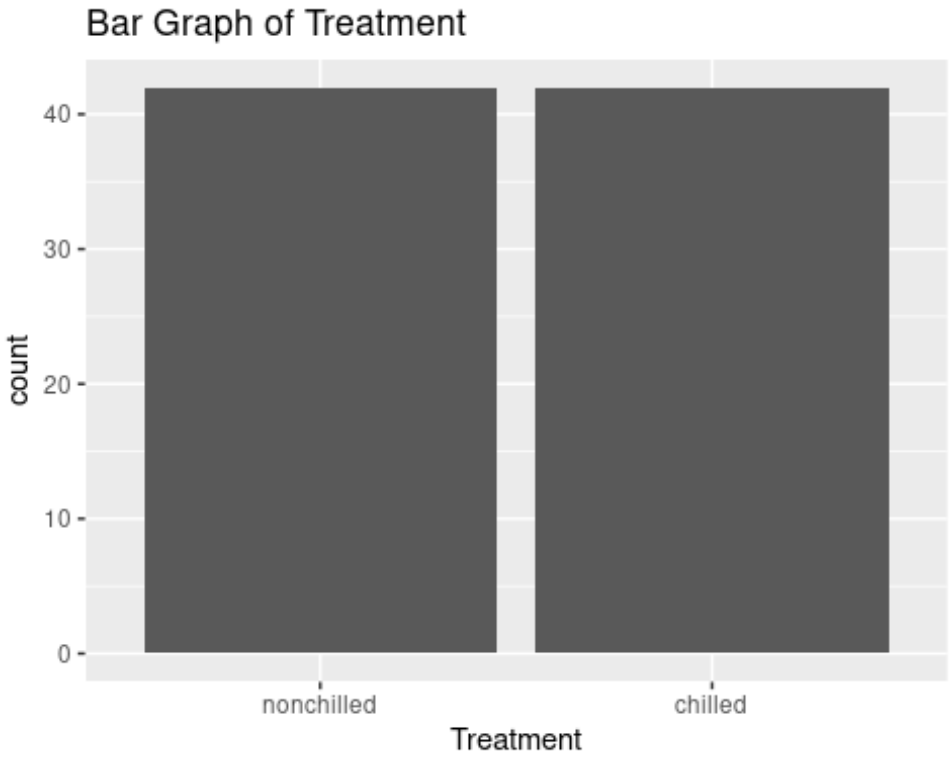
```
variables<-c('Plant','Type','Treatment')
ggplot(data = co2_data) + geom_bar(mapping = aes(x = Plant))
+ggtitle("Bar Graph of Plant")
```




```
ggplot(data = co2_data) + geom_bar(mapping = aes(x = Type))
+ggtitle("Bar Graph of Type")
```

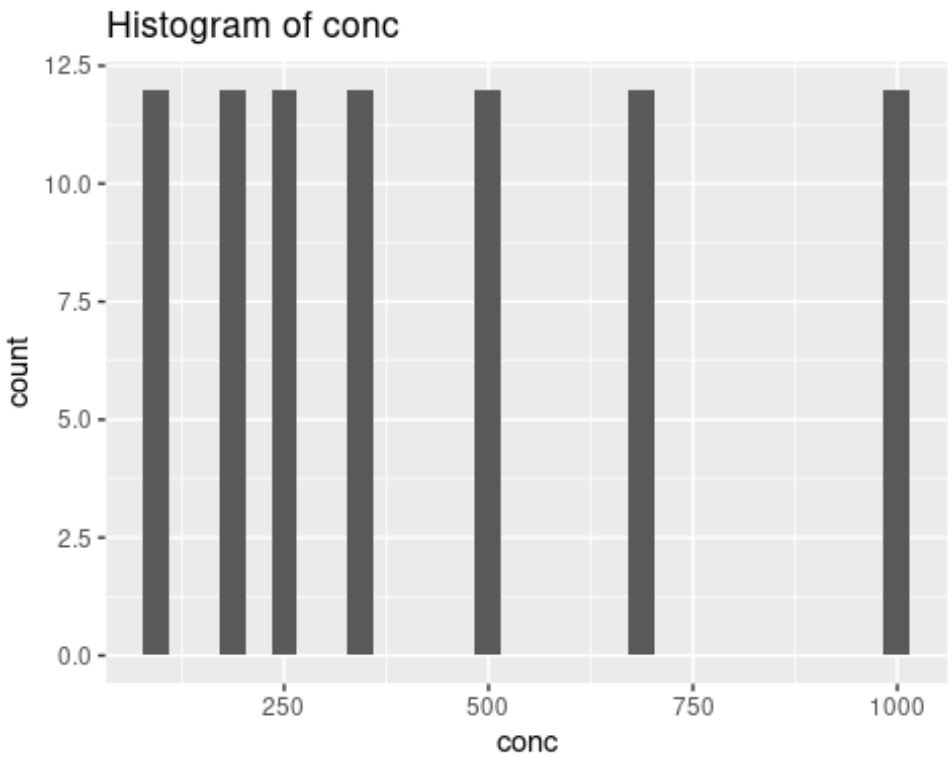


```
ggplot(data = co2_data) + geom_bar(mapping = aes(x = Treatment))
+ggtitle("Bar Graph of Treatment")
```

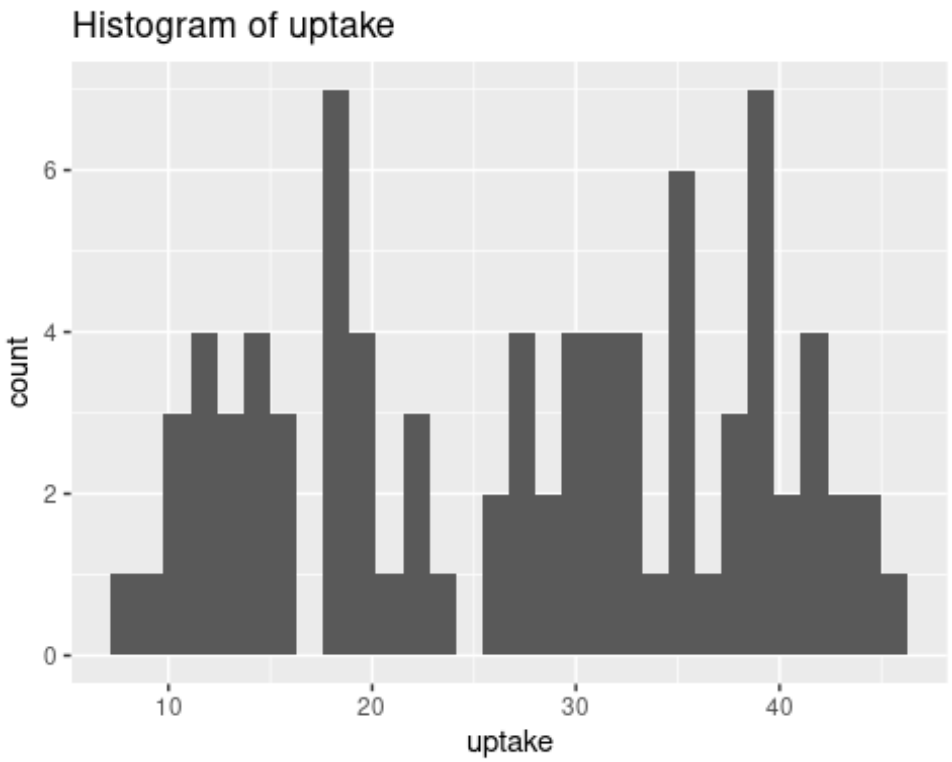


2. Create histogram of conc and uptake variables

```
ggplot(data = co2_data)+geom_histogram(mapping = aes(x=conc))
+ggtitle("Histogram of conc")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

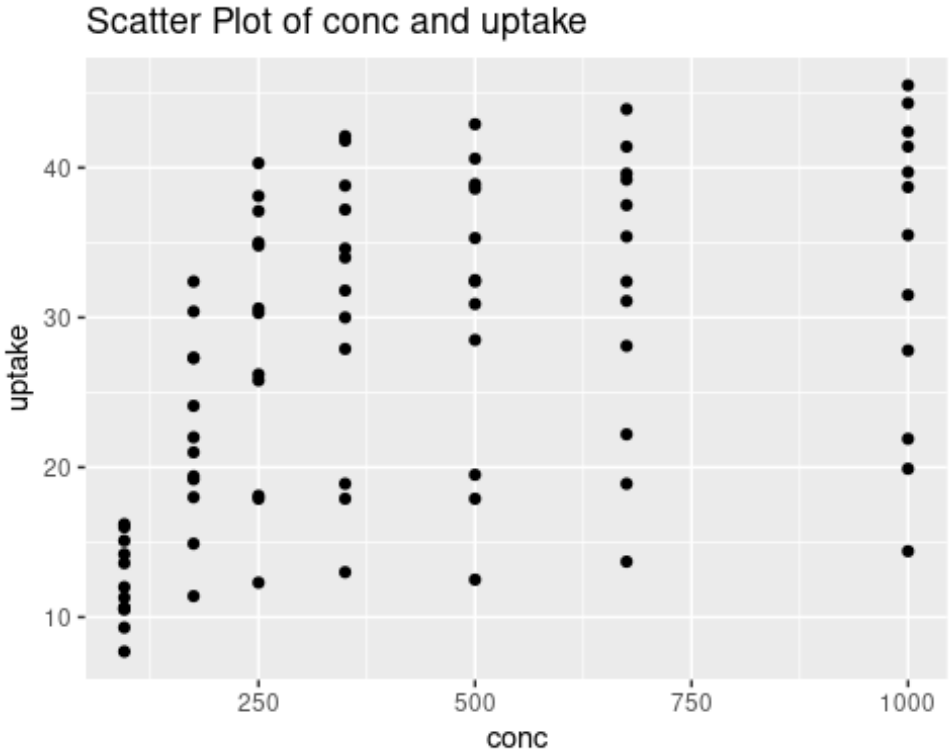



```
ggplot(data = co2_data)+geom_histogram(mapping = aes(x=uptake))
+ggtitle("Histogram of uptake")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



3. Create scatterplot of conc and uptake variables

```
ggplot(data=co2_data)+geom_point(mapping = aes(x=conc,y=uptake))
+ggtitle('Scatter Plot of conc and uptake')
```



Part 4: Text analysis with base/ggplot and Social Network Analysis with igraph package

Use/load the attached "termDocMatrix.rdata" file in R studio and do as follows:

```
file_path = 'data/termDocMatrix.rdata'
term_matrix_data<-load(file = file_path)
```

1. Covert this data as matrix

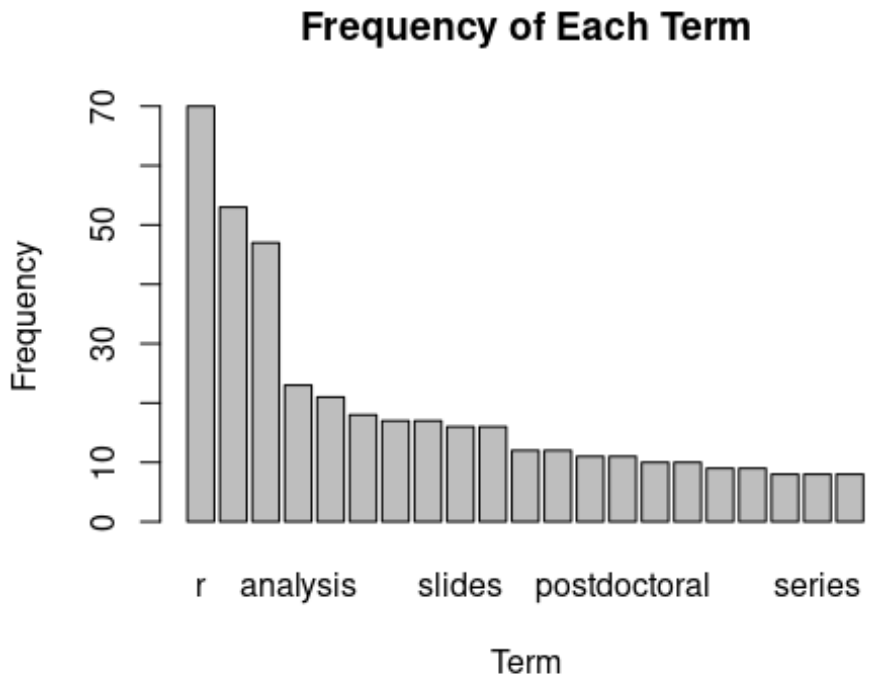
```
library(tm)

## Loading required package: NLP

##
## Attaching package: 'NLP'

## The following object is masked from 'package:ggplot2':
##
##     annotate

term_matrix_data<-
as.DocumentTermMatrix(termDocMatrix,weighting=weightBin)
term_matrix_data<-as.matrix(term_matrix_data)
```

2. Get the term frequencies

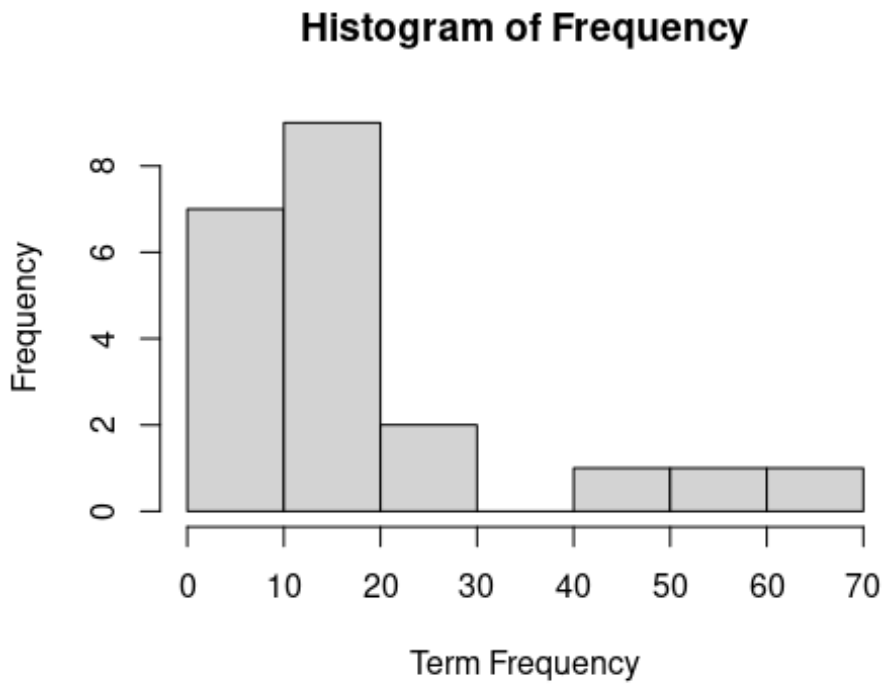
```
freq <- sort(rowSums(term_matrix_data), decreasing=T)
freq

##            r         data       mining     analysis      package
users
##           70           53           47           23           21
18
##     examples      network       slides     tutorial     research
social
##           17           17           16           16           12
12
##    positions postdoctoral    computing introduction applications
code
##           11           11           10           10            9
9
##     parallel       series         time
##            8            8            8
```

3. Create the histogram of the term frequencies

```
barplot(freq,main = "Frequency of Each Term",xlab = "Term",ylab =
"Frequency")
```

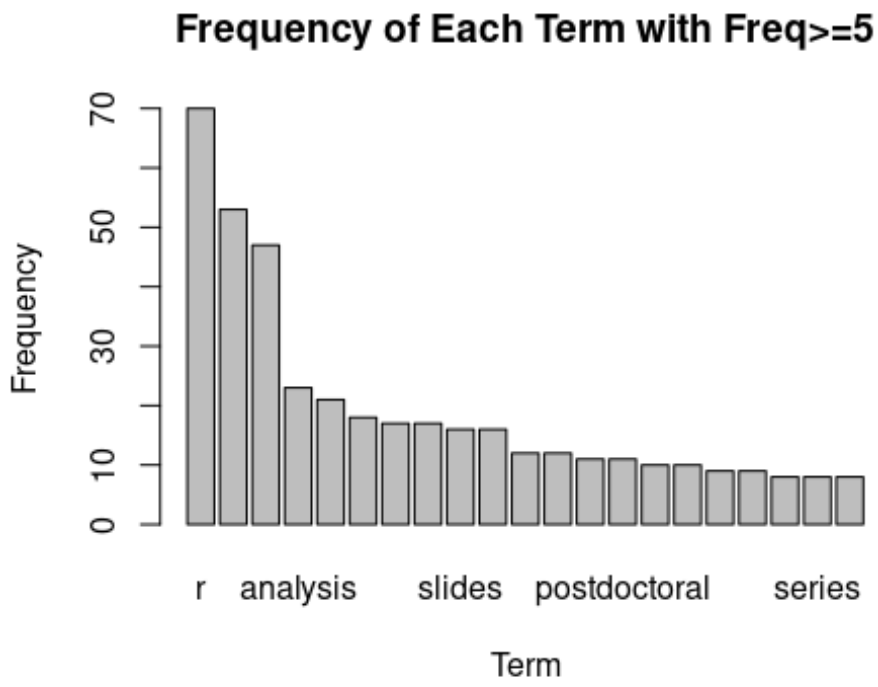



```
hist(freq,main = "Histogram of Frequency",xlab = "Term Frequency")
```

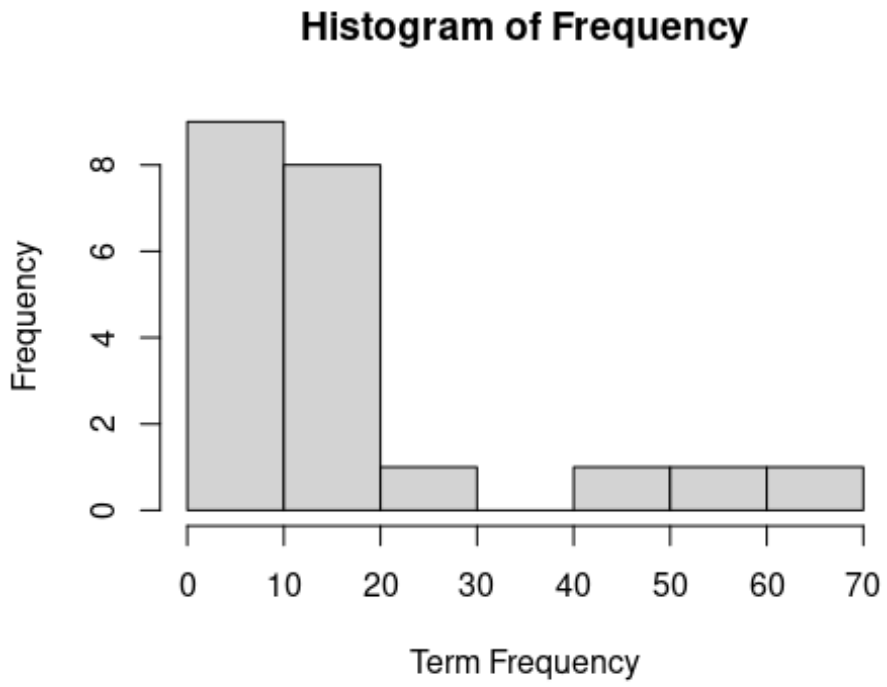


4. Create the histogram of the terms with frequencies of 5 and more

```r
freq_1<-subset(freq,freq>=5)
barplot(freq_1,main = "Frequency of Each Term with Freq>=5",xlab =
"Term",ylab = "Frequency")
```




```r
hist(freq-1,main = "Histogram of Frequency",xlab = "Term Frequency")
```



5. Create word cloud of the term frequencies

```
library(wordcloud)

## Loading required package: RColorBrewer

freq <- sort(rowSums(term_matrix_data), decreasing=T)
wordcloud(words=names(freq), freq=freq, min.freq=5,
random.order=F)
```

6. Perform social network analysis of the termDocumentMatrix data and interpret it carefully

```
library(igraph)
```

```
##
## Attaching package: 'igraph'
```

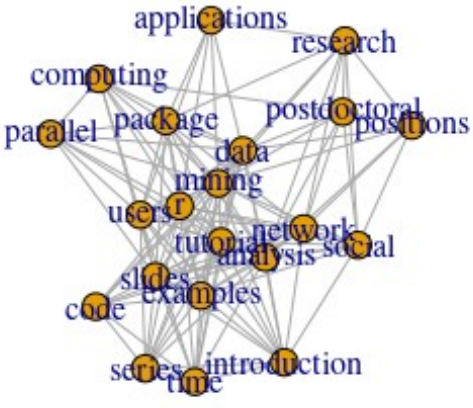
```
## The following objects are masked from 'package:stats':
##
##     decompose, spectrum
```

```
## The following object is masked from 'package:base':
##
##     union
```

```
#Transform Data into an Adjacency Matrix
termDocMatrix[termDocMatrix>=1] <- 1
# Transformation into term-term adjacency matrix
termMatrix <- termDocMatrix %*% t(termDocMatrix)
# Checking few terms in the adjacency matrix
termMatrix[1:5,1:5]
```

```
##               Terms
## Terms          analysis applications code computing data
##   analysis           23            0    1         0    4
##   applications        0            9    0         0    7
##   code                1            0    9         0    1
```

```
##   computing           0            0    0        10    1
##   data                4            7    1         1   53

# Creating a undirected graph
g <- graph.adjacency(termMatrix, weighted=T, mode = "undirected")
# Removing the loop in same term
g<-simplify(g)
plot(g)
```



In the graph above, we can see that the terms like ‘r’, ‘mining’, ‘data’ are at center and are frequently with other words. We can also see that ‘time’,‘series’,‘introduction’ have a cluster. We can also see the cluster of words ‘research’, ‘postdoctoral’,‘positions’ from cluster. It makes sense for these words to come together.