HW1

Jordan Hilton

April 2, 2019

I'll be doing these homeworks in RMDs and knitting to PDF- let me know if there's any problem with formatting you'd like me to fix.

1 Bar Chart

Let's load the data:

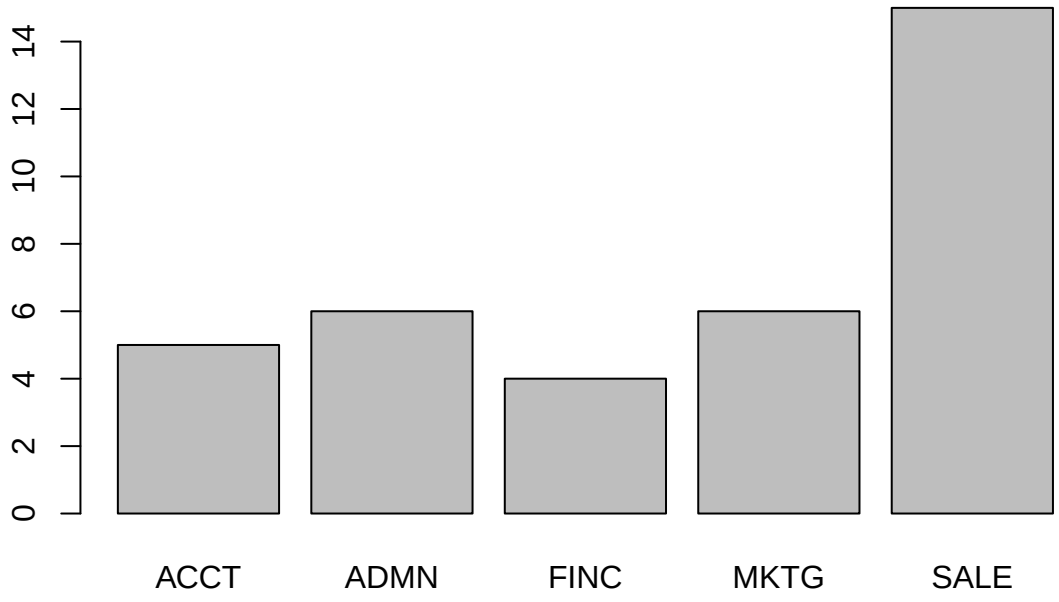
```
d <- rd("Employee", format="lessR")
```

```
##
## >>> Suggestions
## Details about your data, Enter:  details()  for d, or  details(name)
##
## Data Types
## ------------------------------------------------------------
## character: Non-numeric data values
## integer: Numeric data values, integers only
## double: Numeric data values with decimal digits
## ------------------------------------------------------------
##
##     Variable                  Missing  Unique
##         Name     Type  Values  Values  Values   First and last values
## ------------------------------------------------------------------------------------------
##  1     Years   integer     36       1      16   7  NA  15 ... 1  2  10
##  2    Gender character     37       0       2   M  M  M ... F  F  M
##  3      Dept character     36       1       5   ADMN  SALE  SALE ... MKTG  SALE  FINC
##  4    Salary    double     37       0      37   53788.26  94494.58 ... 56508.32  57562.36
##  5    JobSat character     35       2       3   med  low  low ... high  low  high
##  6      Plan   integer     37       0       3   1  1  3 ... 2  2  1
##  7       Pre   integer     37       0      27   82  62  96 ... 83  59  80
##  8      Post   integer     37       0      22   92  74  97 ... 90  71  87
## ------------------------------------------------------------------------------------------
```

a.

Here's a barplot of the number of employees in each department using the base R plot:

```
barplot(table(d$Dept))
```



b.

Here's the same data in table form:

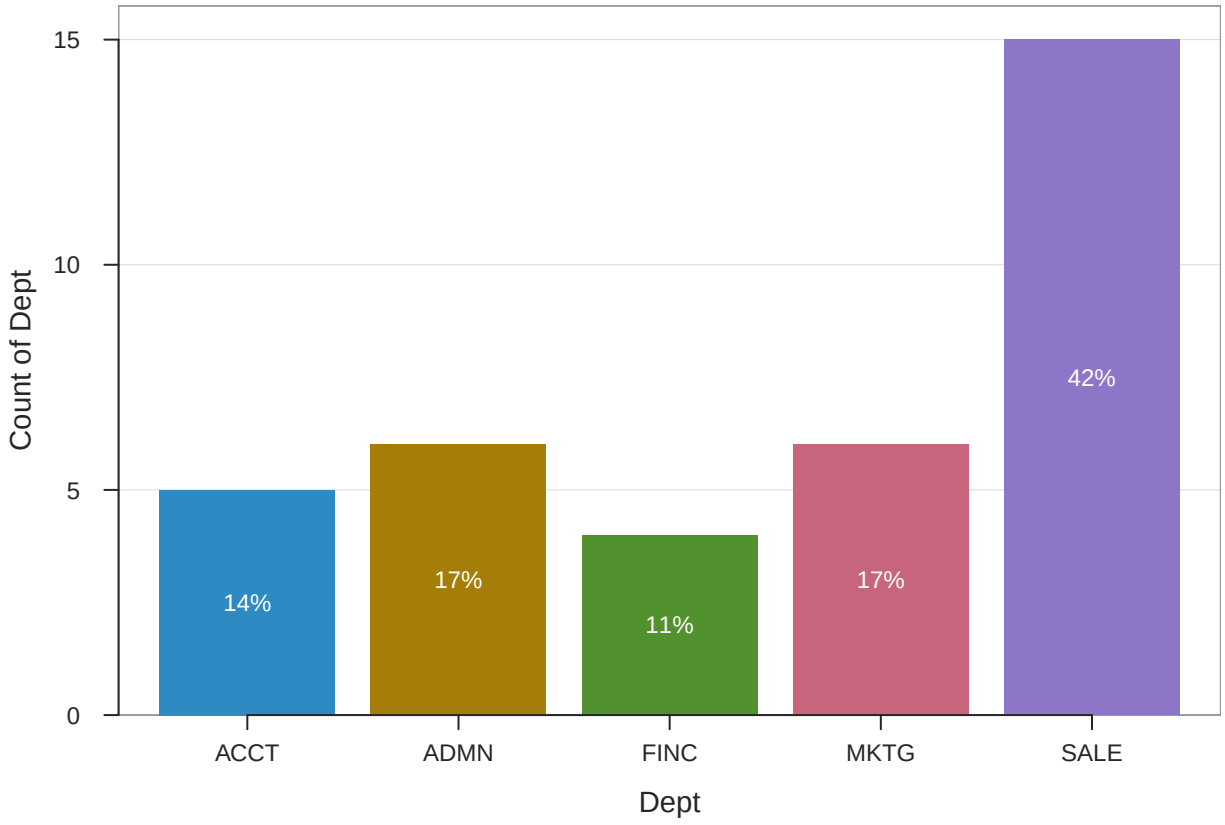
```
table(d$Dept)
```

```
##
## ACCT ADMN FINC MKTG SALE
##    5    6    4    6   15
```

c.

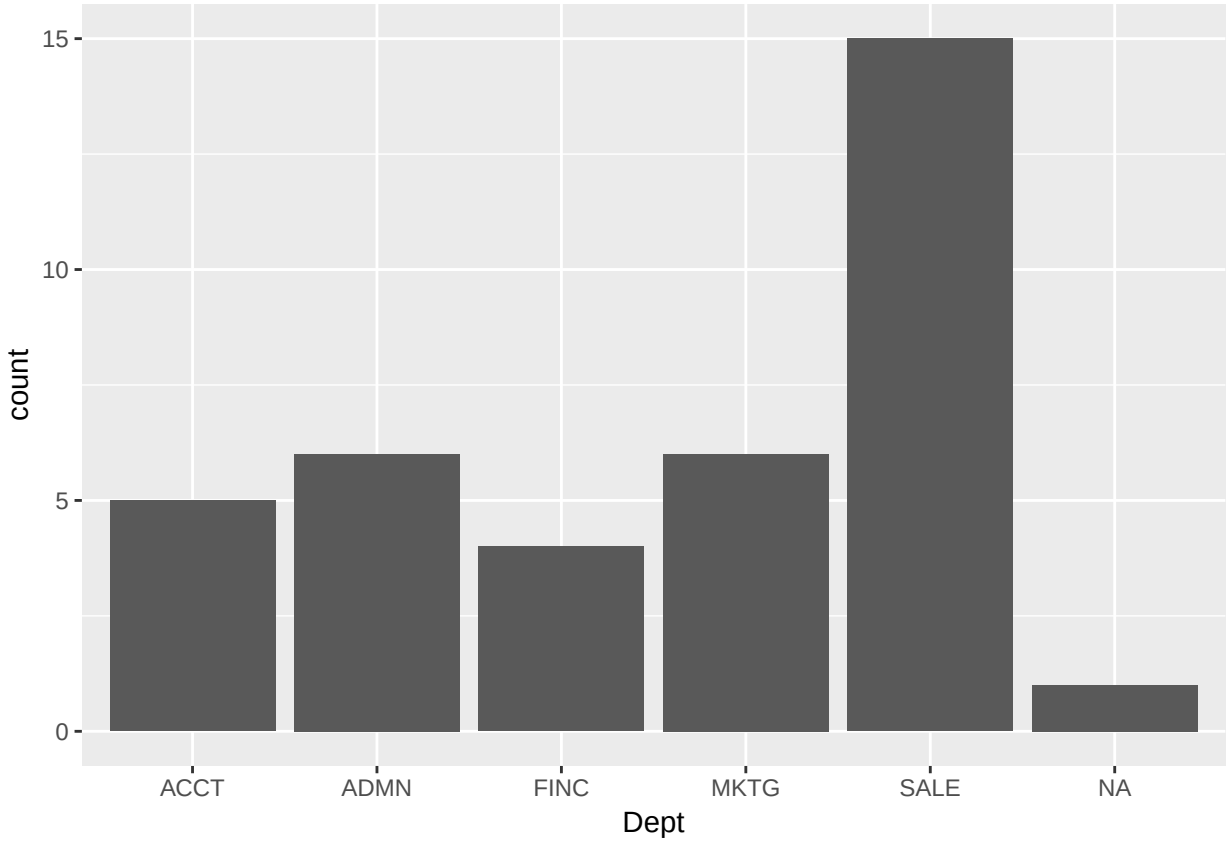
Here's the same chart in lessR:

```
BarChart(Dept, quiet=TRUE)
```



and ggplot2:

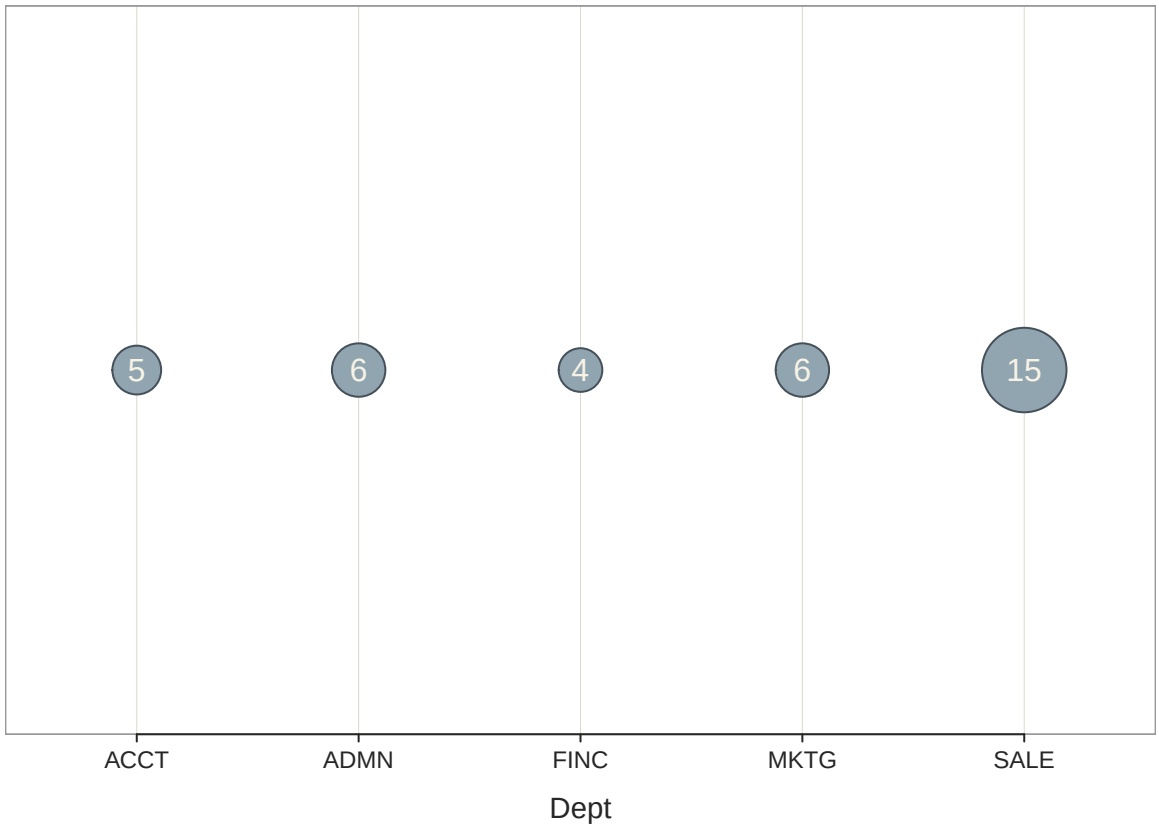
```
ggplot(d, aes(Dept))+geom_bar()
```



d.

Here's the lessR 1d bubble plot:

```
Plot(Dept, quiet=TRUE)
```



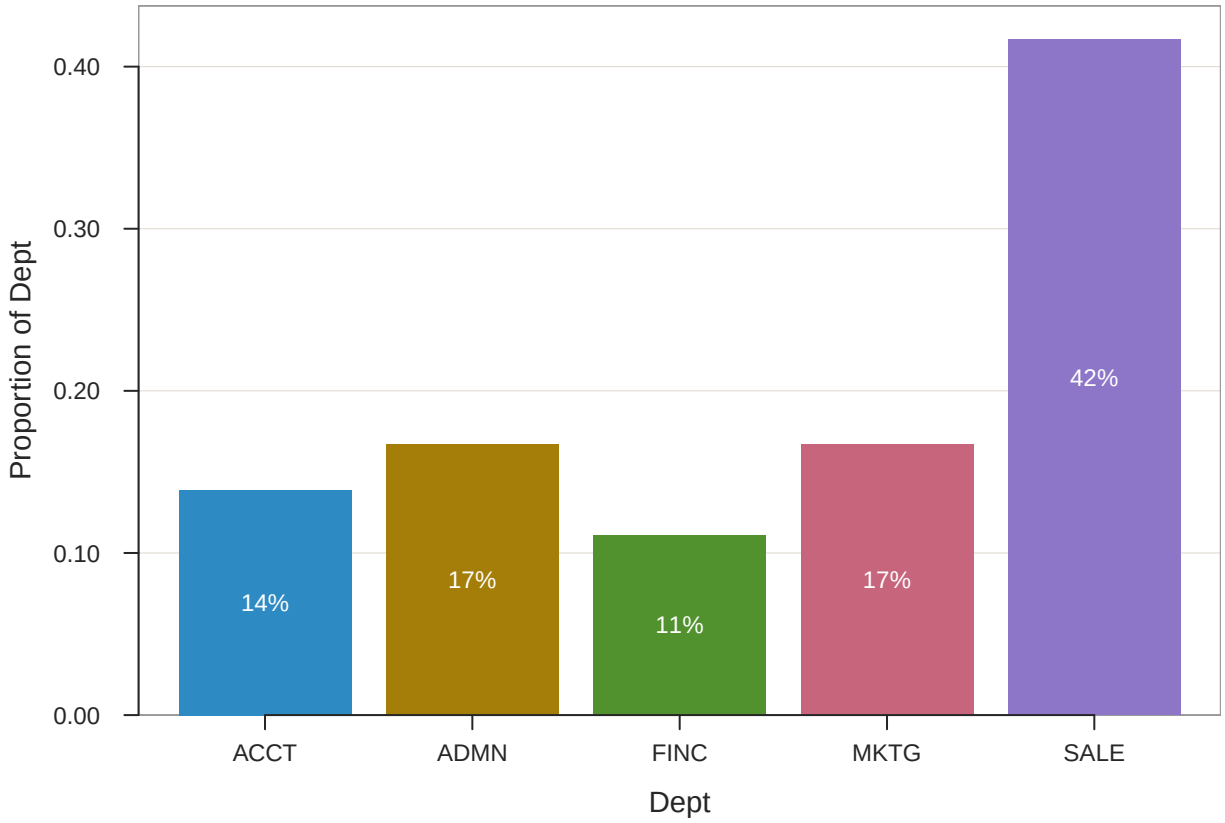
e.

The bubble chart is more compact and there could be applications where showing relative size as an area as opposed to a length is useful. The bar chart is more readable and more common, so it will make more sense to most readers.

h. (no f/g?)

Here's the bar chart with proportions instead of counts:

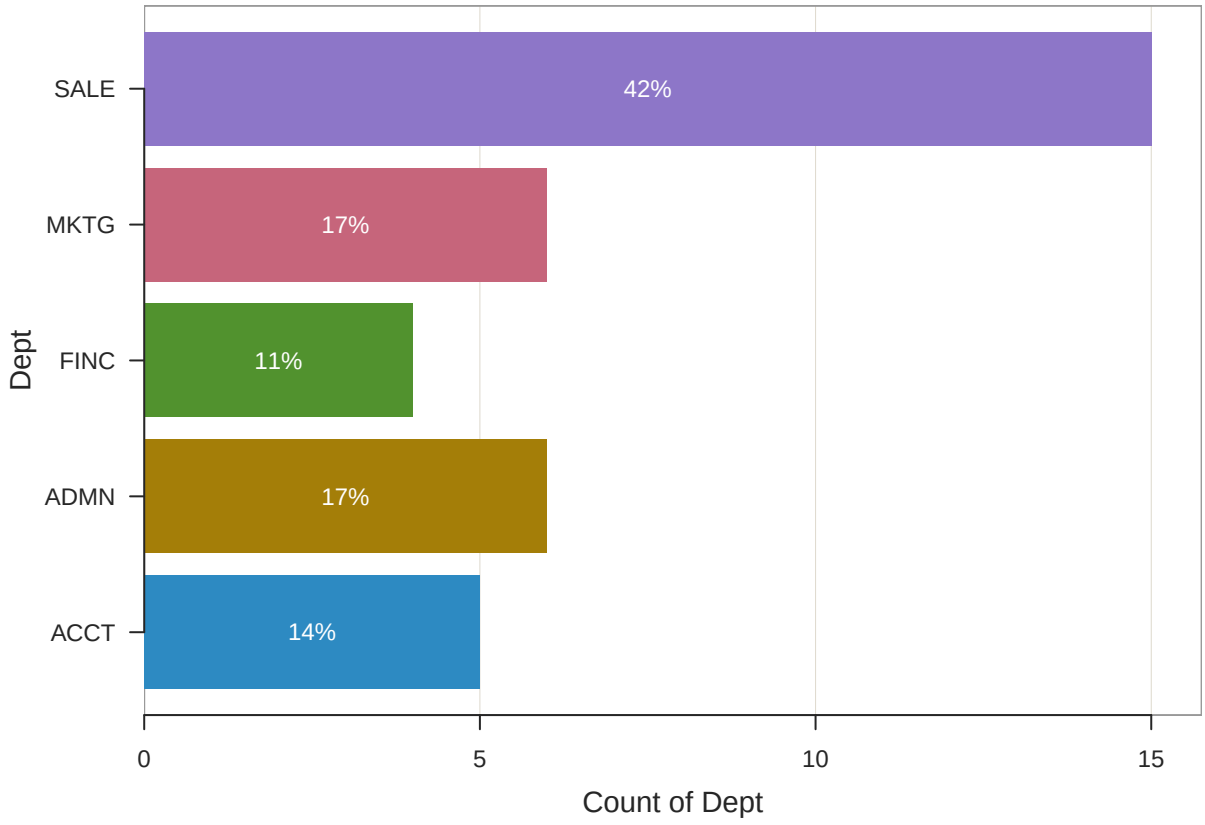
```
BarChart(Dept, quiet=TRUE, stat.x="proportion")
```



i.

With horizontal bars:

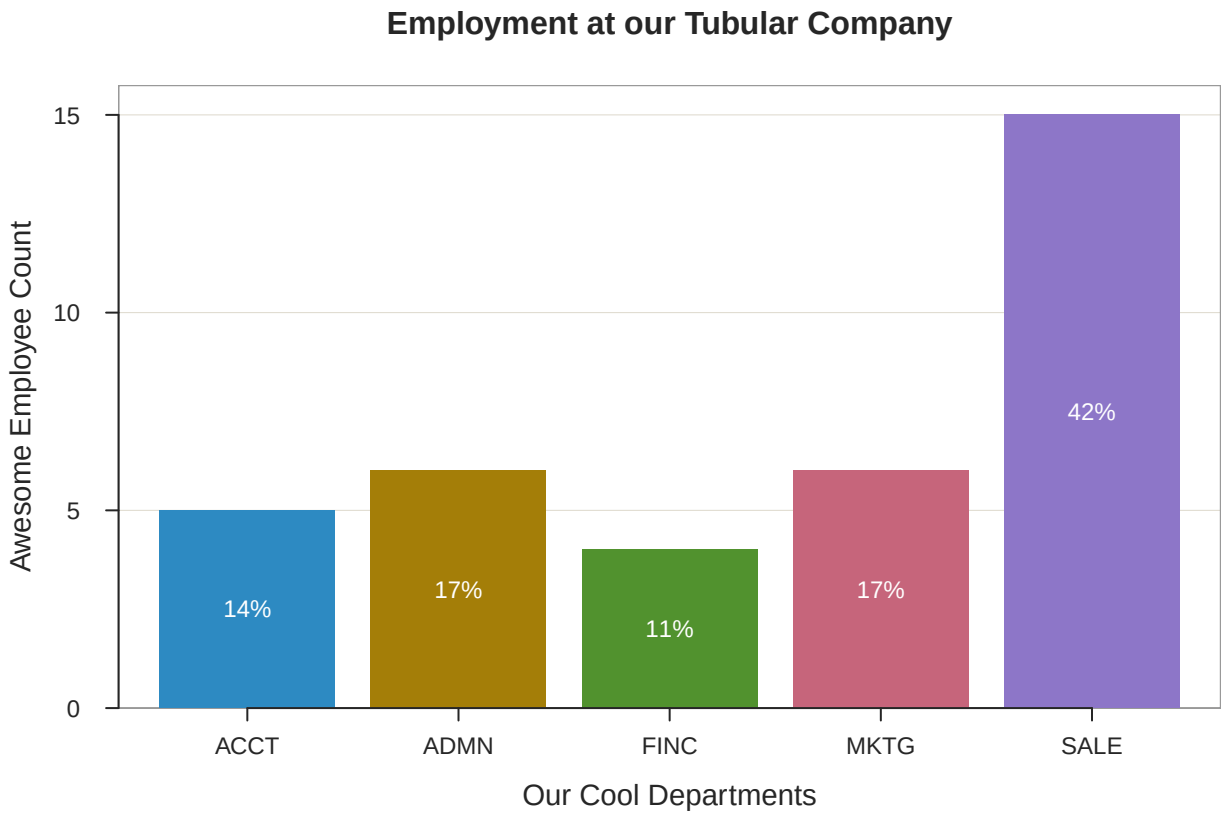
```
BarChart(Dept, quiet=TRUE, horiz=TRUE)
```



j.

Now providing a title and custom axis lables:

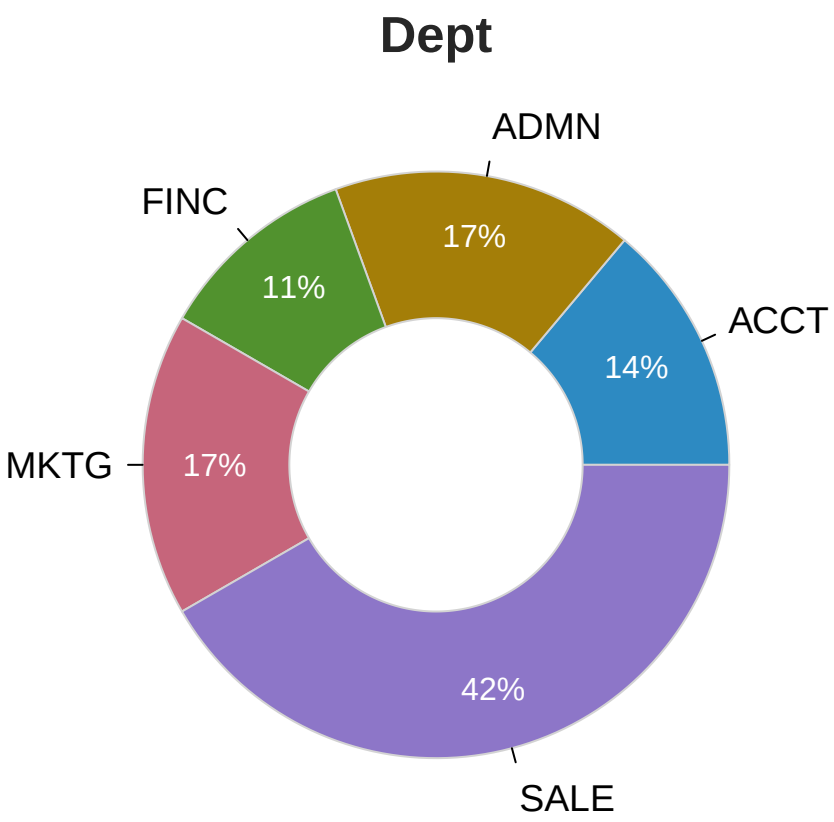
```
BarChart(Dept, quiet=TRUE, xlab="Our Cool Departments", ylab="Awesome Employee Count", main="Employment
```



k.

A ring chart, using lessR:

```
PieChart(Dept, hole=.5, quiet=TRUE)
```



l.

A waffle chart, using the "waffle" package as from the code examples:

```
waffle(table(d$Dept))
```

2. R Factors

Let's load the survey data:

```
surveydata<-rd("460S14.csv", quiet=TRUE)
```

```
head(surveydata)
```

```
##   Learn_1 Learn_2 Learn_3 Learn_4 Feel_1 Feel_2 Feel_3 Feel_4 Past_1
## 1       3       5       7       6      6      7      5      7      3
## 2       4       4       3       6      2      6      6      6      4
## 3       3       3       7       3      5      3      4      1      4
## 4       4       4       5       5      2      6      4      5      2
## 5       6       6       5       6      3      6      4      3      2
## 6       7       7       2       7      5      7      7      7      3
##   Past_2 Past_3 Past_4 Past2_1 Gender Class Learn2_1 Learn2_2 Learn2_3
## 1      4      3      4       6      2     2       59       78       95
## 2      4      4      4       6      1     2       30       50       60
## 3      2      3      2       7      2     2       NA       NA      100
## 4      2      1      2       6      1     2       50       39       70
## 5      3      4      3      13      2     2       60      100       50
```

```
## 6      1      2      2       7      1      2      100       100        10
##   Learn2_4
## 1       53
## 2       50
## 3       NA
## 4       60
## 5       91
## 6      100
```

```
length(surveydata$Learn_1)
```

```
## [1] 31
```

a.

We can see from the length of the first column that there are 31 rows of data, so that's probably the number of student responses we got.

b.

The four variables we're discussing are Past_1, Past_2, Past_3, and Past_4

c.

```
sum(is.na(surveydata))
```

```
## [1] 7
```

```
sum(is.na(surveydata$Past_1))
```

```
## [1] 0
```

```
sum(is.na(surveydata$Past_2))
```

```
## [1] 0
```

```
sum(is.na(surveydata$Past_3))
```

```
## [1] 0
```

```
sum(is.na(surveydata$Past_4))
```

```
## [1] 0
```

We can see that there are 7 total missing values in the dataset, but no missing values for any of these 4 questions. The missing values are in the columns "Past2_1" and "Learn2_1" so I"m guessing that those questions are either optional or only presented under some circumstances.

d.

Here are the first 3 rows of data for our four variables:

```
head(cbind(surveydata$Past_1, surveydata$Past_2, surveydata$Past_3,surveydata$Past_4),3)
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    3    4    3    4
## [2,]    4    4    4    4
## [3,]    4    2    3    2
```

e.

It looks like the responses are recorded as integers 1-4, presumably with 1 corresponding to "not at all" and 4 corresponding to "cannot remember"

f.

Let's convert these four variables to factors, and I'll use head to show the results for one variable:

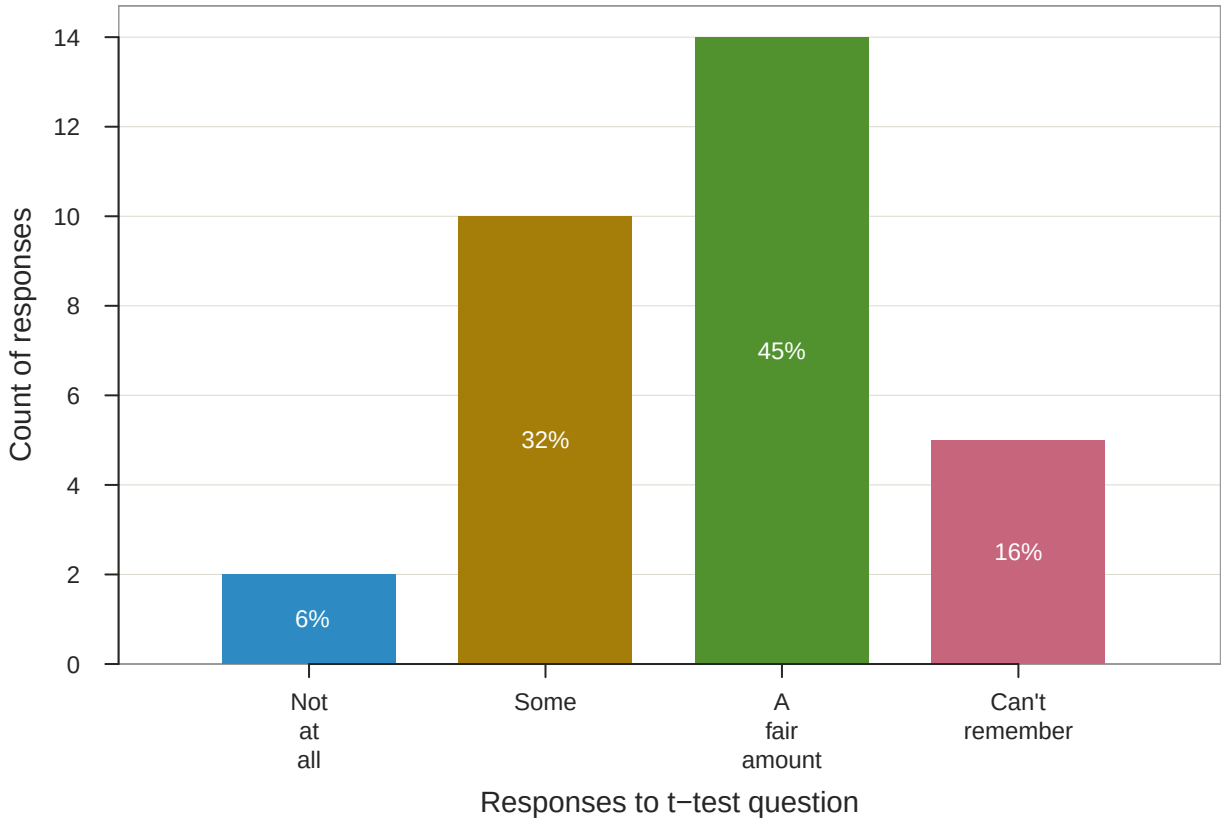
```
surveydata$Past_1 = factor(surveydata$Past_1, levels=1:4, labels=c("Not at all", "Some", "A fair amount
surveydata$Past_2 = factor(surveydata$Past_2, levels=1:4, labels=c("Not at all", "Some", "A fair amount
surveydata$Past_3 = factor(surveydata$Past_3, levels=1:4, labels=c("Not at all", "Some", "A fair amount
surveydata$Past_4 = factor(surveydata$Past_4, levels=1:4, labels=c("Not at all", "Some", "A fair amount
head(surveydata$Past_1, 3)
```

```
## [1] A fair amount  Can't remember Can't remember
## Levels: Not at all Some A fair amount Can't remember
```

g.

A bar chart for the t-test survey question:

```
BarChart(Past_1, quiet=TRUE, data=surveydata, xlab="Responses to t-test question", ylab="Count of respo
```



h.

Create a new, reordered factor variable, just for the t-test question: (I had to reload the data first in order to recreate the factor)

```
surveydata<-rd("460S14.csv", quiet=TRUE)
```

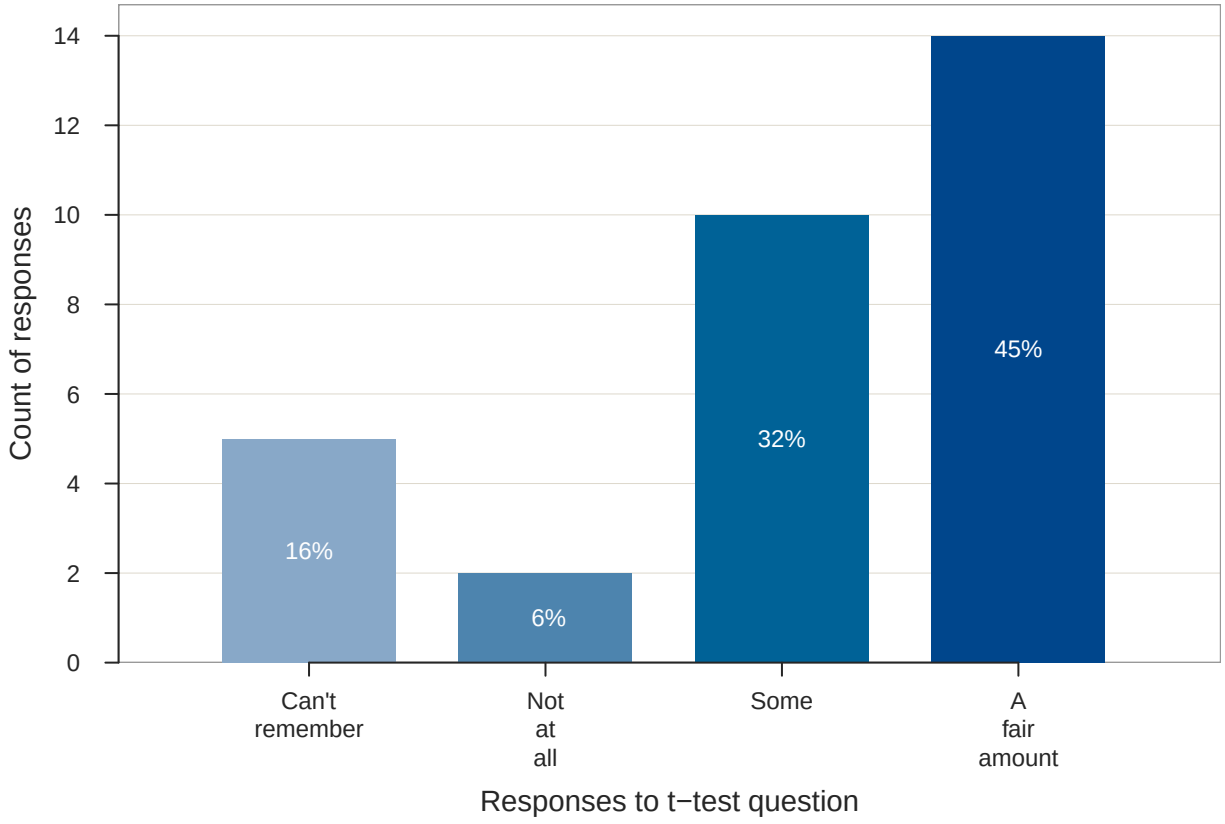
```
reordered = factor(surveydata$Past_1, levels=c(4,1,2,3), labels=c("Can't remember","Not at all", "Some"
```

i.

Chart the reordered factor variable: it's interesting how the default coloring changes for ordered factors in lessR

```
BarChart(reordered, quiet=TRUE, data=reordered, xlab="Responses to t-test question", ylab="Count of resp
```

```
## >>> Note: reordered is from the workspace, not in a data frame (table)
```



j.

Let's see if there are any cases where nobody picked one of the options for one of these 4 questions:

```
table(surveydata$Past_1)
```

```
##
##  1  2  3  4
##  2 10 14  5
```

```
table(surveydata$Past_2)
```

```
##
##  1  2  3  4
##  6 10 11  4
```

```
table(surveydata$Past_3)
```

```
##
##  1  2  3  4
##  3 13  6  9
```

```
table(surveydata$Past_4)
```

```
##
##  1  2  3  4
##  3 15  8  5
```

It looks like at least 2 people picked each option for all of these 4 questions. If we had missing cases, what we could do is create a factor with a level for the missing response, which would then show up as having 0 instances. For example if there was a 5th option that nobody had picked, we might run this pseudocode which I'll commment out so it doesn't break my RMD:

```
## missingvalues = factor(surveyresponse, levels=1:5, labels=c("Can't remember","Not at all", "Some", "
```