Code

Exploratory Data Analysis (EDA) for Four Year Graduation

Vanessa Gonzalez

2018-06-25

Data Preparation and EDA

First File to Run Open Libraries

Hide

```
library("ggplot2", lib.loc="/Library/Frameworks/R.framework/Versions/3.4/Resources/library")
library("graphics", lib.loc="/Library/Frameworks/R.framework/Versions/3.4/Resources/library")
library("gplots", lib.loc="/Library/Frameworks/R.framework/Versions/3.4/Resources/library")
library("dplyr")
library("lattice", lib.loc="/Library/Frameworks/R.framework/Versions/3.4/Resources/library")
library("data.table", lib.loc="/Library/Frameworks/R.framework/Versions/3.4/Resources/library")
library("datasets")
library("readxl")
library("stats")
```

Import data set

Hide

```
setwd("~/REGIS/Practicum I/RStudio")
dataSet <- read_excel("./dataSet.xlsx")
#View(dataSet)
```

Transform data set into data frame

Hide

```
dfDataSet <- as.data.frame(dataSet)
head(dfDataSet)
```

Year of OriginalMajorDate	GraduationStatus	YearsFromO...	CsGrad	...	...	...	1_(

	<chr>	<chr>	<chr>	<chr>	<chr≈chr≈chr>
1	2014	Graduated	4.00	Yes	Yes Yes Yes
2	2008	Graduated	9.84	OtherMajor	No Yes Yes
3	2008	InactiveReg	9.84	NG	No No No
4	2011	InactiveReg	6.84	NG	No No No
5	2008	Graduated	9.84	OtherMajor	Yes Yes Yes
6	2008	Graduated	9.84	Yes	Yes Yes Yes

6 rows | 1-9 of 24 columns

Hide

```
head(dfDataSet)
```

	Year of OriginalMajorDate <chr>	**GraduationStatus** <chr>	**YearsFromO...** <chr>	**CsGrad** <chr>	**...** **...** **...** **1_** <chr≈chr≈chr>
1	2014	Graduated	4.00	Yes	Yes Yes Yes
2	2008	Graduated	9.84	OtherMajor	No Yes Yes
3	2008	InactiveReg	9.84	NG	No No No
4	2011	InactiveReg	6.84	NG	No No No
5	2008	Graduated	9.84	OtherMajor	Yes Yes Yes
6	2008	Graduated	9.84	Yes	Yes Yes Yes

6 rows | 1-9 of 24 columns

Hide

```
summary(dfDataSet)
```

```
 Year of OriginalMajorDate GraduationStatus   YearsFromOMD          CsGrad
4YG                5YG
 Length:536                Length:536         Length:536         Length:536         L
ength:536          Length:536
 Class :character          Class :character   Class :character   Class :character   C
lass :character   Class :character
 Mode  :character          Mode  :character   Mode  :character   Mode  :character   M
ode  :character   Mode  :character
```

```
     6YG              1_CSCI101       1_MATH111      2_CSCI261       2_MATH112
2_MATH201       3_CSCI262      3_MATH213
 Length:536         Min.   :0.300   Min.   :0.30   Min.   :0.300   Min.   :0.300   Mi
n.   :0.300   Min.   :0.30   Min.   :0.300
 Class :character   1st Qu.:3.000   1st Qu.:3.00   1st Qu.:3.000   1st Qu.:2.000   1s
t Qu.:2.000   1st Qu.:3.00   1st Qu.:2.000
 Mode  :character   Median :4.000   Median :3.00   Median :4.000   Median :3.000   Me
dian :3.000   Median :4.00   Median :3.000
                    Mean   :3.419   Mean   :2.91   Mean   :3.405   Mean   :2.875   Me
an   :2.701   Mean   :3.24   Mean   :2.888
                    3rd Qu.:4.000   3rd Qu.:3.00   3rd Qu.:4.000   3rd Qu.:4.000   3r
d Qu.:3.300   3rd Qu.:4.00   3rd Qu.:4.000
                    Max.   :4.000   Max.   :4.00   Max.   :4.000   Max.   :4.000   Ma
x.   :4.000   Max.   :4.00   Max.   :4.000
                    NA's   :76      NA's   :12     NA's   :37      NA's   :26      NA
's   :108     NA's   :81     NA's   :49
   4_CSCI341       4_CSCI358       4_MATH225       5_CSCI306       5_CSCI403       5_
MATH332       6_CSCI406       7_CSCI370
 Min.   :0.300   Min.   :0.300   Min.   :0.300   Min.   :0.300   Min.   :0.300   Min.
:0.300   Min.   :0.300   Min.   :2.300
 1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:3.000   1st Qu.:3.000   1st
Qu.:2.000   1st Qu.:2.000   1st Qu.:4.000
 Median :3.000   Median :3.000   Median :3.000   Median :3.700   Median :4.000   Medi
an :3.000   Median :3.000   Median :4.000
 Mean   :2.854   Mean   :2.961   Mean   :2.763   Mean   :3.433   Mean   :3.565   Mean
:2.671   Mean   :2.795   Mean   :3.895
 3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd
Qu.:3.300   3rd Qu.:3.300   3rd Qu.:4.000
 Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.
:4.000   Max.   :4.000   Max.   :4.000
 NA's   :108     NA's   :108     NA's   :58      NA's   :132     NA's   :250     NA's
:121     NA's   :154     NA's   :179
   8_CSCI400       9_CSCI442
 Min.   :0.300   Min.   :0.300
 1st Qu.:2.925   1st Qu.:3.000
 Median :3.300   Median :3.300
 Mean   :3.175   Mean   :3.133
 3rd Qu.:4.000   3rd Qu.:4.000
 Max.   :4.000   Max.   :4.000
 NA's   :164     NA's   :165
```

Transform Variable's type to the appropriate data type

Hide

```
dfDataSet[1] <- lapply(dfDataSet[1], as.integer)
dfDataSet[3] <- lapply(dfDataSet[3], as.numeric)
dfDataSet[8:24] <- lapply(dfDataSet[8:24], as.numeric)
dfDataSet[2] <- lapply(dfDataSet[2], as.factor)
dfDataSet[4:7] <- lapply(dfDataSet[4:7], as.factor)
str(dfDataSet)
```

```
'data.frame':   536 obs. of  24 variables:
 $ Year of OriginalMajorDate: int  2014 2008 2008 2011 2008 2008 2008 2008 2008 2008
...
 $ GraduationStatus         : Factor w/ 3 levels "CurrentStudent",..: 2 2 3 3 2 2 2 2
3 2 ...
 $ YearsFromOMD             : num  4 9.84 9.84 6.84 9.84 9.84 9.84 9.84 9.84 9.84 ...
 $ CsGrad                   : Factor w/ 3 levels "NG","OtherMajor",..: 3 2 1 1 2 3 3
2 1 2 ...
 $ 4YG                      : Factor w/ 2 levels "No","Yes": 2 1 1 1 2 2 2 2 1 2 ...
 $ 5YG                      : Factor w/ 2 levels "No","Yes": 2 2 1 1 2 2 2 2 1 2 ...
 $ 6YG                      : Factor w/ 2 levels "No","Yes": 2 2 1 1 2 2 2 2 1 2 ...
 $ 1_CSCI101                : num  4 NA NA NA NA 4 4 NA NA NA ...
 $ 1_MATH111                : num  3 3 3 3 3 3 4 3 2 3 ...
 $ 2_CSCI261                : num  4 4 4 3.3 4 3 4 3 3 3 ...
 $ 2_MATH112                : num  2 2 2 4 3 3 4 4 2 2 ...
 $ 2_MATH201                : num  3 3 2 NA 3 2 4 NA NA 1 ...
 $ 3_CSCI262                : num  3 NA 1 3 NA 3 4 4 1 NA ...
 $ 3_MATH213                : num  4 3 2 2 3 4 4 4 1 2 ...
 $ 4_CSCI341                : num  2 NA 2 2 NA 3 4 NA 3 3 ...
 $ 4_CSCI358                : num  4 NA 3 2 NA 2 4 NA NA NA ...
 $ 4_MATH225                : num  4 3 1 4 4 3 4 4 NA 1 ...
 $ 5_CSCI306                : num  3 NA NA 3.7 NA 4 4 4 NA NA ...
 $ 5_CSCI403                : num  4 NA NA 3 NA 4 4 NA NA NA ...
 $ 5_MATH332                : num  3 NA 2 3 NA 3 4 NA NA NA ...
 $ 6_CSCI406                : num  2 NA NA 0.3 NA 2 4 NA NA NA ...
 $ 7_CSCI370                : num  3.3 NA NA NA NA 4 4 NA NA NA ...
 $ 8_CSCI400                : num  3.3 NA NA 3.3 NA 3 4 NA NA NA ...
 $ 9_CSCI442                : num  2.3 NA NA NA NA 3 4 NA NA NA ...
```

Hide

```
summary(dfDataSet)
```

```
 Year of OriginalMajorDate       GraduationStatus  YearsFromOMD          CsGrad     4
YG       5YG       6YG        1_CSCI101
 Min.   :2008              CurrentStudent: 32     Min.   :3.830   NG        :140   No
:252   No :169   No :150   Min.   :0.300
 1st Qu.:2009              Graduated     :396     1st Qu.:4.830   OtherMajor: 57   Ye
```

```
s:284   Yes:367   Yes:386   1st Qu.:3.000
 Median :2011              InactiveReg   :108     Median :6.840   Yes       :339
Median :4.000
 Mean   :2011                                    Mean   :6.706
Mean   :3.419
 3rd Qu.:2013                                    3rd Qu.:8.840
3rd Qu.:4.000
 Max.   :2014                                    Max.   :9.840
Max.   :4.000

NA's   :76
   1_MATH111      2_CSCI261       2_MATH112       2_MATH201       3_CSCI262      3_MA
TH213       4_CSCI341       4_CSCI358
 Min.   :0.30   Min.   :0.300   Min.   :0.300   Min.   :0.300   Min.   :0.30   Min.
:0.300   Min.   :0.300   Min.   :0.300
 1st Qu.:3.00   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:3.00   1st Qu
.:2.000   1st Qu.:2.000   1st Qu.:2.000
 Median :3.00   Median :4.000   Median :3.000   Median :3.000   Median :4.00   Median
:3.000   Median :3.000   Median :3.000
 Mean   :2.91   Mean   :3.405   Mean   :2.875   Mean   :2.701   Mean   :3.24   Mean
:2.888   Mean   :2.854   Mean   :2.961
 3rd Qu.:3.00   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:3.300   3rd Qu.:4.00   3rd Qu
.:4.000   3rd Qu.:4.000   3rd Qu.:4.000
 Max.   :4.00   Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.00   Max.
:4.000   Max.   :4.000   Max.   :4.000
 NA's   :12     NA's   :37      NA's   :26      NA's   :108     NA's   :81     NA's
:49      NA's   :108     NA's   :108
   4_MATH225       5_CSCI306       5_CSCI403       5_MATH332       6_CSCI406       7_
CSCI370       8_CSCI400       9_CSCI442
 Min.   :0.300   Min.   :0.300   Min.   :0.300   Min.   :0.300   Min.   :0.300   Min.
:2.300   Min.   :0.300   Min.   :0.300
 1st Qu.:2.000   1st Qu.:3.000   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:2.000   1st
Qu.:4.000   1st Qu.:2.925   1st Qu.:3.000
 Median :3.000   Median :3.700   Median :4.000   Median :3.000   Median :3.000   Medi
an :4.000   Median :3.300   Median :3.300
 Mean   :2.763   Mean   :3.433   Mean   :3.565   Mean   :2.671   Mean   :2.795   Mean
:3.895   Mean   :3.175   Mean   :3.133
 3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:3.300   3rd Qu.:3.300   3rd
Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000
 Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.
:4.000   Max.   :4.000   Max.   :4.000
 NA's   :58      NA's   :132     NA's   :250     NA's   :121     NA's   :154     NA's
:179     NA's   :164     NA's   :165
```

Transform data frame into a table

Hide

```
tbDataSet <- data.table(dfDataSet)
tbDataSet
```

Year of OriginalMajorDate	GraduationStatus	YearsFromO…	CsGrad	4…	5…	6…
<int>	<fctr>	<dbl>	<fctr>	<fctr>	<fctr>	<fctr>
2014	Graduated	4.00	Yes	Yes	Yes	Yes
2008	Graduated	9.84	OtherMajor	No	Yes	Yes
2008	InactiveReg	9.84	NG	No	No	No
2011	InactiveReg	6.84	NG	No	No	No
2008	Graduated	9.84	OtherMajor	Yes	Yes	Yes
2008	Graduated	9.84	Yes	Yes	Yes	Yes
2008	Graduated	9.84	Yes	Yes	Yes	Yes
2008	Graduated	9.84	OtherMajor	Yes	Yes	Yes
2008	InactiveReg	9.84	NG	No	No	No
2008	Graduated	9.84	OtherMajor	Yes	Yes	Yes

1-10 of 536 rows | 1-8 of 24 columns Previous **1** 2 3 4 5 6 … 54 Next

Create a table for variable “GraduationStatus”

Hide

```
tb <- table(dfDataSet$GraduationStatus)
tb
```

```
CurrentStudent      Graduated    InactiveReg
            32            396            108
```

Create table with percentages of total

Hide

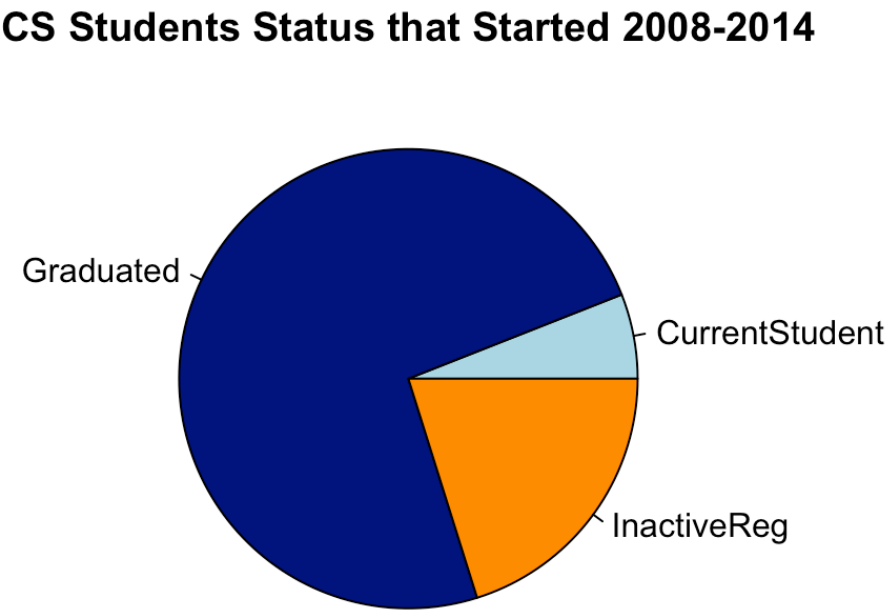
```
tb.prop <- dfDataSet$GraduationStatus %>%
   table() %>%
   prop.table() %>% {. * 100} %>%
   round(2)
tb.prop
```

```
.
CurrentStudent      Graduated    InactiveReg
          5.97          73.88          20.15
```

Create a Pie chart

Hide

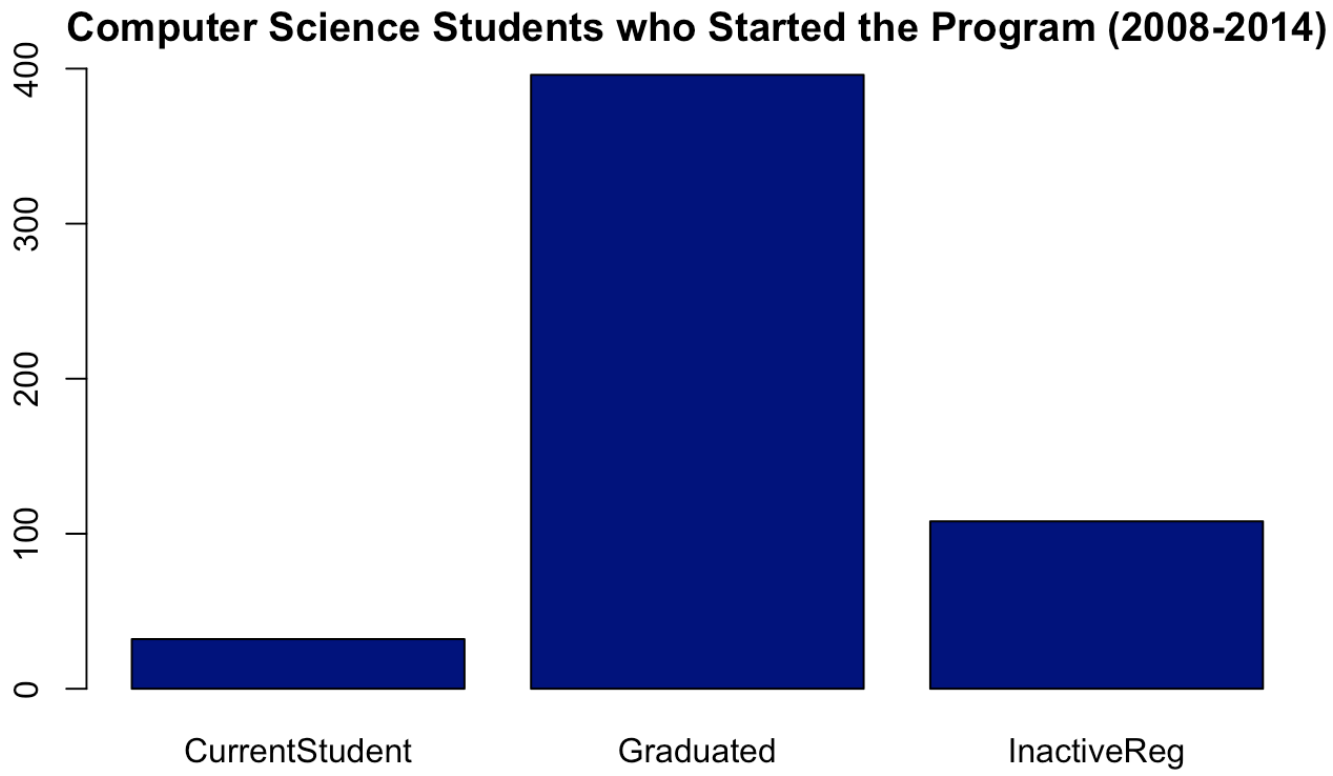
```
pie(tb, main = "CS Students Status that Started 2008-2014", col = c("Light Blue","Navy Blue","Dark Orange"))
```



Create a Bar chart

Hide

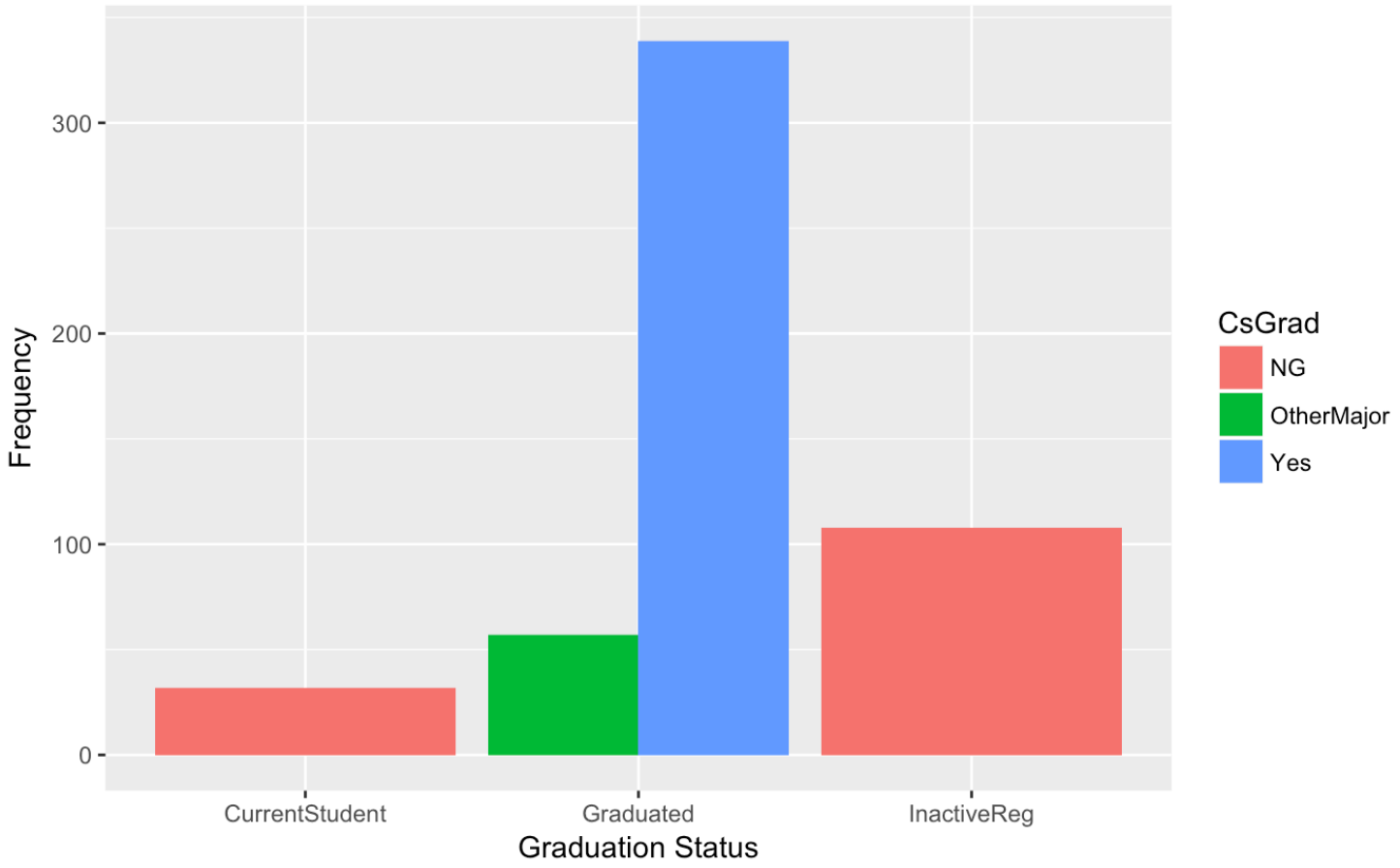
```
barplot(tb, col = "navy blue", ylim=c(0,400),main = "Computer Science Students who Started the Program (2008-2014)")
```



Create a Bar plot where the “graduated”" students are divided into CS and Other Major students

Hide

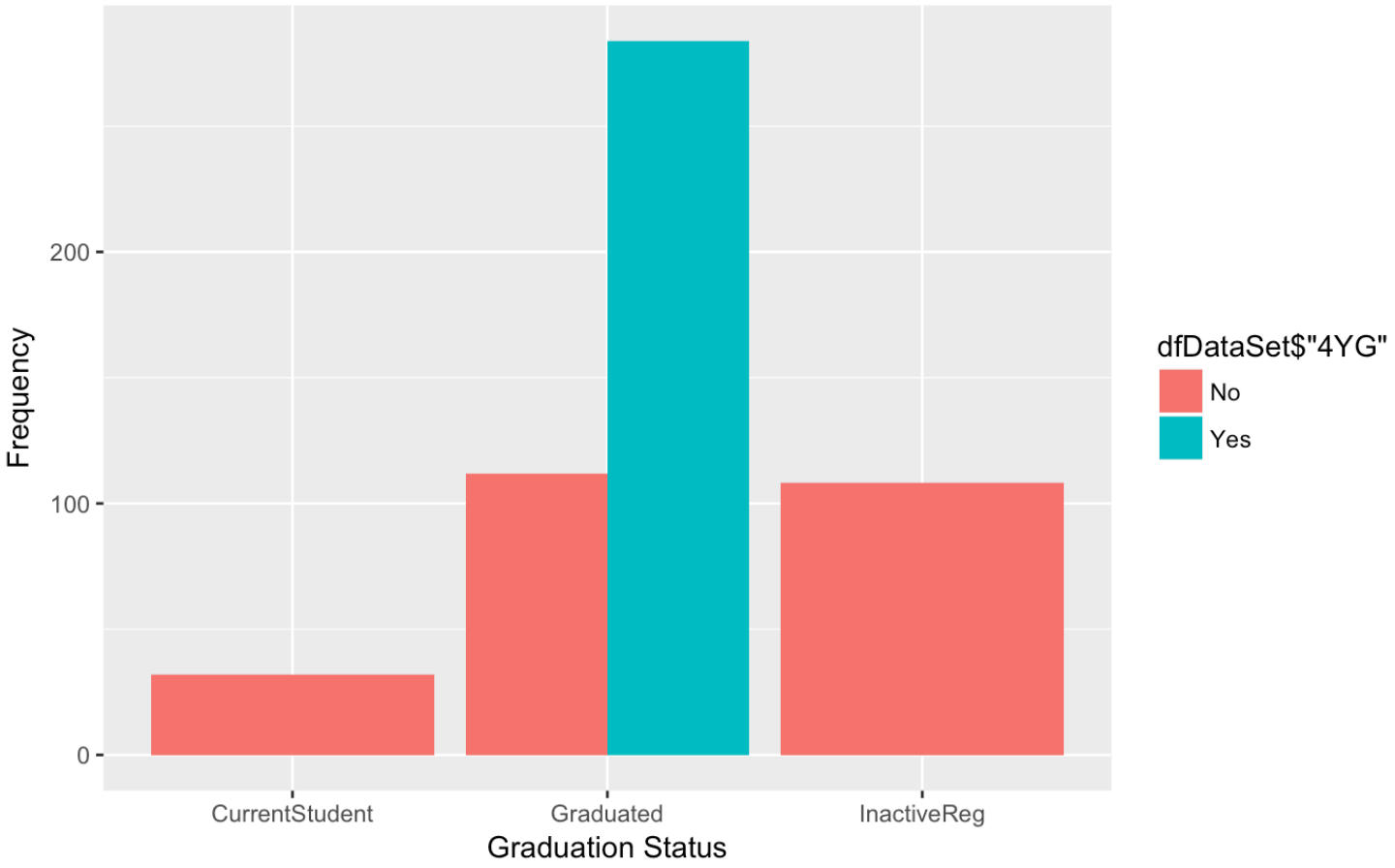
```
ggplot(dfDataSet) +
  geom_bar(aes(x = factor(GraduationStatus), fill = CsGrad), position = 'dodge') +
  xlab('Graduation Status') + ylab('Frequency')
```



Create a plot to represent the CS students that have graduated in 4 years

Hide

```
ggplot(dfDataSet) +
  geom_bar(aes(x = factor(GraduationStatus), fill = dfDataSet$"4YG"), position = 'dod
ge') +
  xlab('Graduation Status') + ylab('Frequency')
```



Create a table of numbers and percetages of students that graduate in 4 years

Hide

```
tb4YG <- table(dfDataSet$"4YG")
tb4YGpercent <- prop.table(tb4YG)
tb4YGpercent
```

```
       No       Yes 
0.4701493 0.5298507 
```

Hide

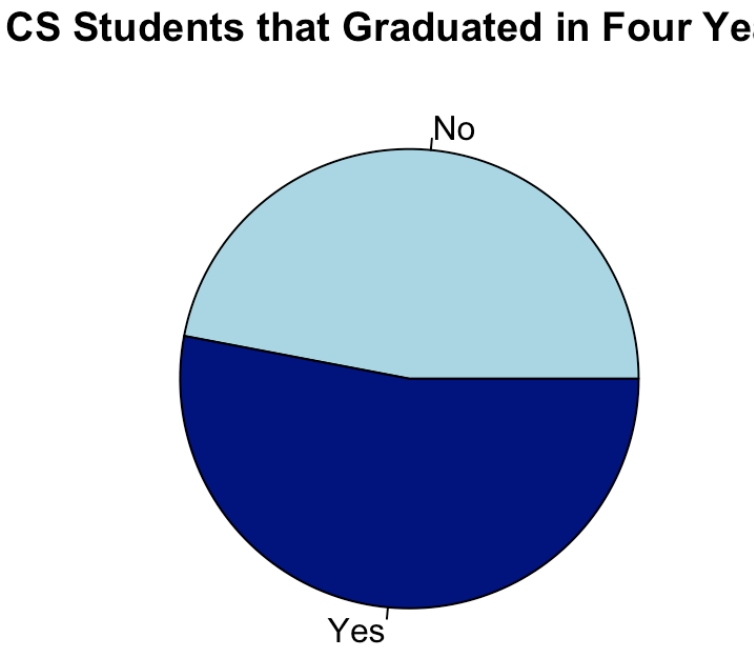
```
cbind(tb4YG,prop.table(tb4YG))
```

```
    tb4YG
No    252 0.4701493
Yes   284 0.5298507
```

Create a Pie Chart for Four-year graduation rate

Hide

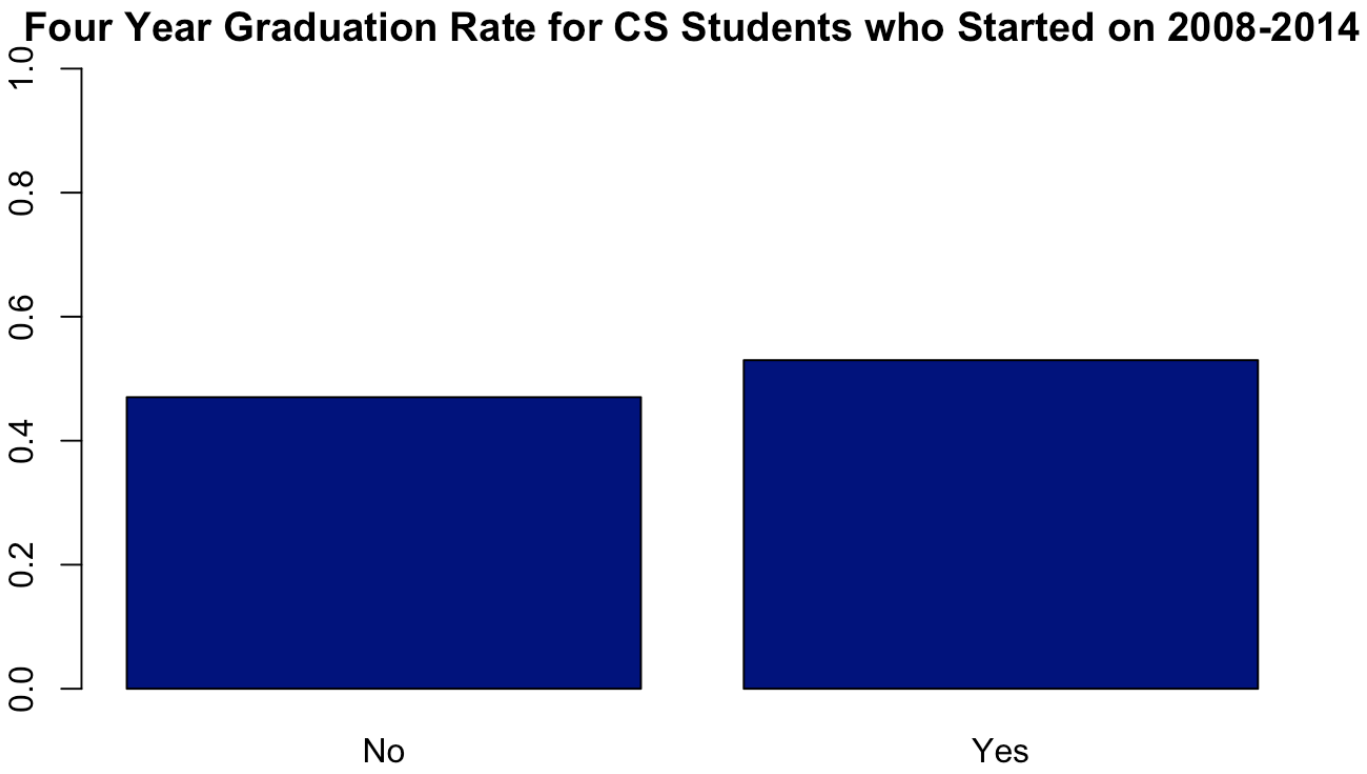
```
pie(tb4YG, main = "CS Students that Graduated in Four Years", col = c("Light Blue","Navy Blue","Dark Orange"))
```



Create a Bar chart for Four-year graduation rate

Hide

```
barplot(tb4YGpercent, col = "navy blue", ylim=c(0,1),main = "Four Year Graduation Rate for CS Students who Started on 2008-2014")
```



Create a table of numbers and percetages of students that graduate in 5 years

Hide

```
tb5YG <- table(dfDataSet$"5YG")
tb5YGpercent <- prop.table(tb5YG)
tb5YGpercent
```

```
       No       Yes 
0.3152985 0.6847015 
```

Hide

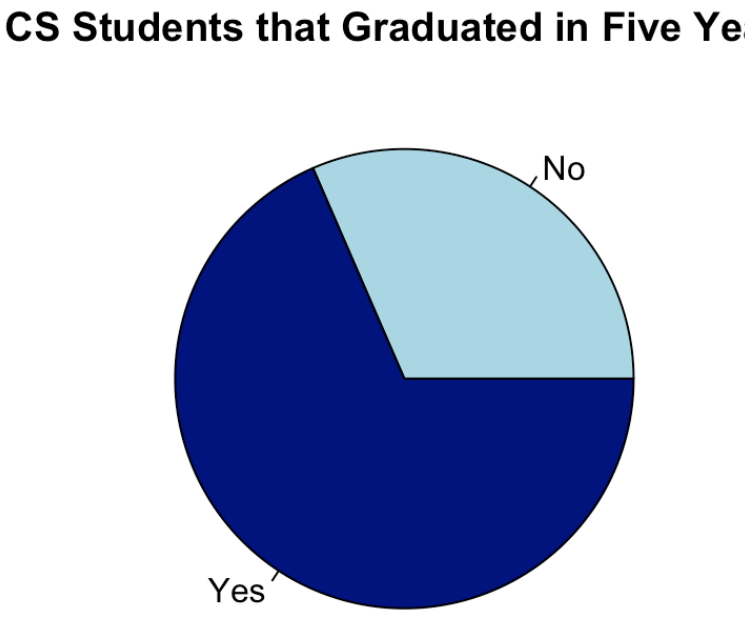
```
cbind(tb5YG,prop.table(tb5YG))
```

```
    tb5YG
No    169 0.3152985
Yes   367 0.6847015
```

Create a Pie Chart for Five-year graduation rate

Hide

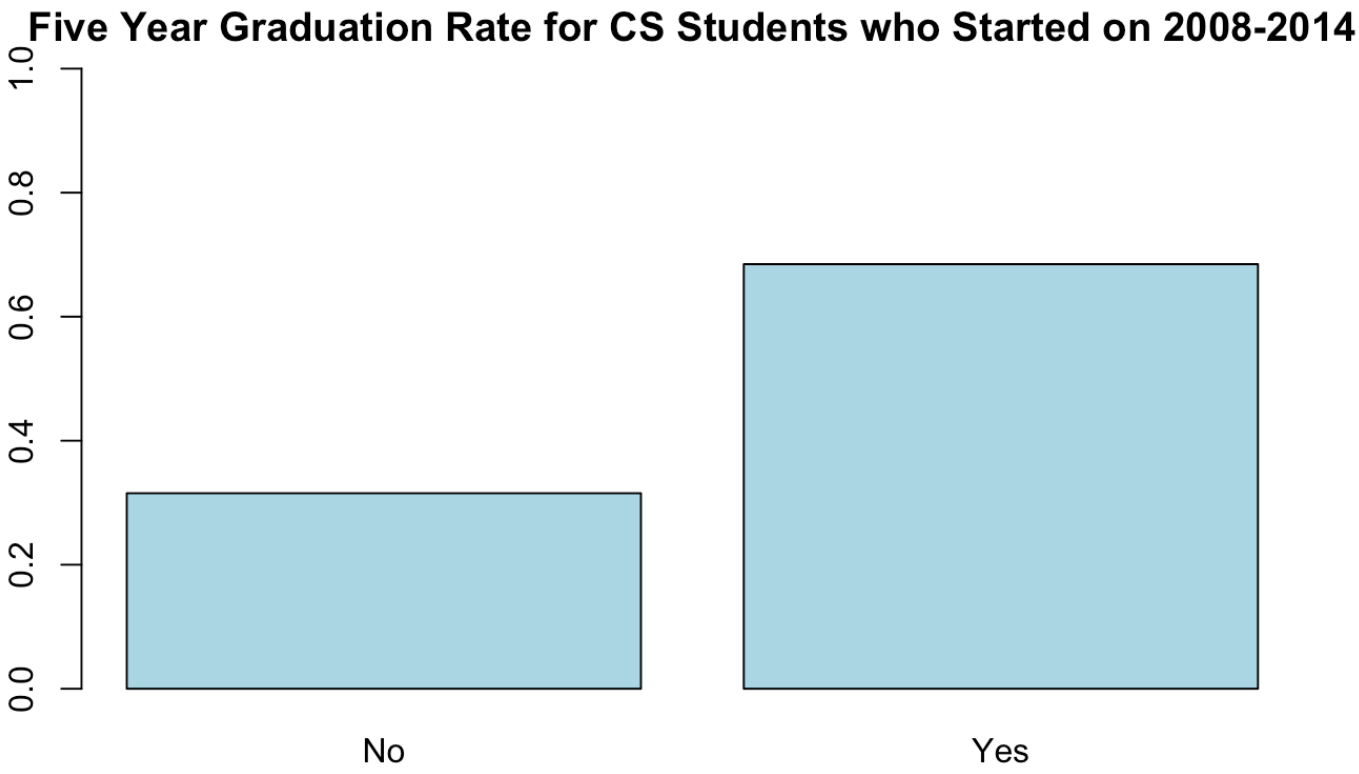
```
pie(tb5YG, main = "CS Students that Graduated in Five Years", col = c("Light Blue","Navy Blue","Dark Orange"))
```



Create a Bar Chart for Five-year graduation rate

Hide

```
barplot(tb5YGpercent, col = "lightblue", ylim=c(0,1),main = "Five Year Graduation Rate for CS Students who Started on 2008-2014")
```



Create a table of numbers and percetages of students that graduate in 6 years

Hide

```
tb6YG <- table(dfDataSet$"6YG")
tb6YGpercent <- prop.table(tb6YG)
tb6YGpercent
```

```
       No       Yes 
0.2798507 0.7201493 
```

Hide

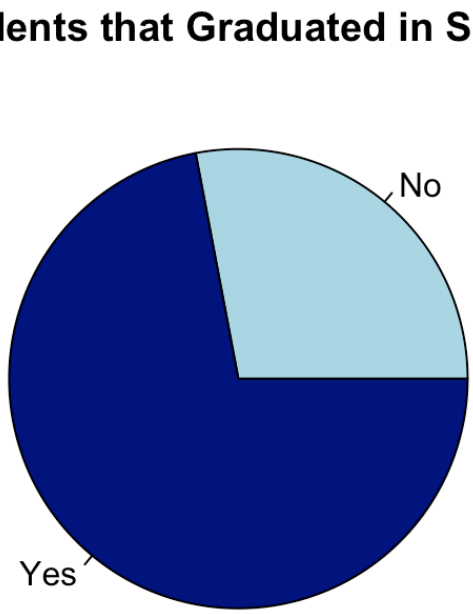
```
cbind(tb6YG,prop.table(tb6YG))
```

```
    tb6YG          
No    150 0.2798507
Yes   386 0.7201493
```

Create a Pie Chart for Five-year graduation rate

Hide

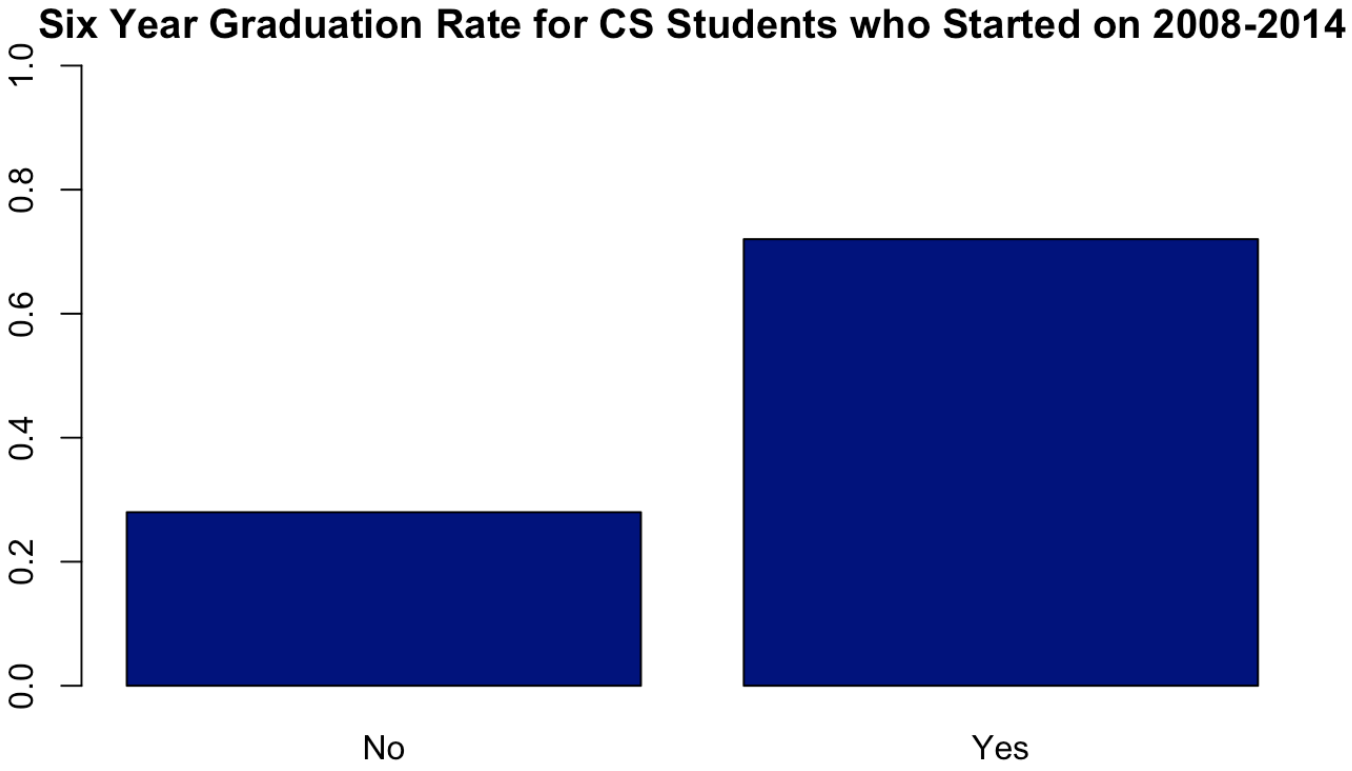
```
pie(tb6YG, main = "CS Students that Graduated in Six Years", col = c("Light Blue","Navy Blue","Dark Orange"))
```



Create a Bar Chart for Six-year graduation rate

Hide

```
barplot(tb6YGpercent, col = "navy blue", ylim=c(0,1), main = "Six Year Graduation Rate for CS Students who Started on 2008-2014")
```


Six Year Graduation Rate for CS Students who Started on 2008-2014
0.0
0.2
0.4
0.6
0.8
1.0
No
Yes