Exploratory Data Analysis (EDA) for Four Year Graduation

Vanessa Gonzalez

2018-06-24

Data Preparation and EDA

First File to Run Open Libraries

```
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")
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
library("dplyr")
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
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")
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
library("datasets")
library("readxl")
library("stats")
```

Import data set

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

Transform data set into data frame

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

```
##   Year of OriginalMajorDate GraduationStatus YearsFromOMD     CsGrad 4YG
## 1                      2014        Graduated         4.00        Yes Yes
## 2                      2008        Graduated         9.84 OtherMajor  No
## 3                      2008      InactiveReg         9.84         NG  No
## 4                      2011      InactiveReg         6.84         NG  No
## 5                      2008        Graduated         9.84 OtherMajor Yes
## 6                      2008        Graduated         9.84        Yes Yes
##   5YG 6YG 1_CSCI101 1_MATH111 2_CSCI261 2_MATH112 2_MATH201 3_CSCI262
## 1 Yes Yes         4         3       4.0         2         3         3
## 2 Yes Yes        NA         3       4.0         2         3        NA
## 3  No  No        NA         3       4.0         2         2         1
## 4  No  No        NA         3       3.3         4        NA         3
## 5 Yes Yes        NA         3       4.0         3         3        NA
## 6 Yes Yes         4         3       3.0         3         2         3
##   3_MATH213 4_CSCI341 4_CSCI358 4_MATH225 5_CSCI306 5_CSCI403 5_MATH332
## 1         4         2         4         4       3.0         4         3
## 2         3        NA        NA         3        NA        NA        NA
## 3         2         2         3         1        NA        NA         2
## 4         2         2         2         4       3.7         3         3
## 5         3        NA        NA         4        NA        NA        NA
## 6         4         3         2         3       4.0         4         3
##   6_CSCI406 7_CSCI370 8_CSCI400 9_CSCI442
## 1       2.0       3.3       3.3       2.3
## 2        NA        NA        NA        NA
## 3        NA        NA        NA        NA
## 4       0.3        NA       3.3        NA
## 5        NA        NA        NA        NA
## 6       2.0       4.0       3.0       3.0
```

```r
head(dfDataSet)
```

```
##   Year of OriginalMajorDate GraduationStatus YearsFromOMD     CsGrad 4YG
## 1                      2014        Graduated         4.00        Yes Yes
## 2                      2008        Graduated         9.84 OtherMajor  No
## 3                      2008      InactiveReg         9.84         NG  No
## 4                      2011      InactiveReg         6.84         NG  No
## 5                      2008        Graduated         9.84 OtherMajor Yes
## 6                      2008        Graduated         9.84        Yes Yes
##   5YG 6YG 1_CSCI101 1_MATH111 2_CSCI261 2_MATH112 2_MATH201 3_CSCI262
## 1 Yes Yes         4         3       4.0         2         3         3
## 2 Yes Yes        NA         3       4.0         2         3        NA
## 3  No  No        NA         3       4.0         2         2         1
## 4  No  No        NA         3       3.3         4        NA         3
## 5 Yes Yes        NA         3       4.0         3         3        NA
## 6 Yes Yes         4         3       3.0         3         2         3
##   3_MATH213 4_CSCI341 4_CSCI358 4_MATH225 5_CSCI306 5_CSCI403 5_MATH332
## 1         4         2         4         4       3.0         4         3
## 2         3        NA        NA         3        NA        NA        NA
## 3         2         2         3         1        NA        NA         2
## 4         2         2         2         4       3.7         3         3
## 5         3        NA        NA         4        NA        NA        NA
## 6         4         3         2         3       4.0         4         3
```

```
##   6_CSCI406 7_CSCI370 8_CSCI400 9_CSCI442
## 1       2.0       3.3       3.3       2.3
## 2        NA        NA        NA        NA
## 3        NA        NA        NA        NA
## 4       0.3        NA       3.3        NA
## 5        NA        NA        NA        NA
## 6       2.0       4.0       3.0       3.0
```

```
summary(dfDataSet)
```

```
##  Year of OriginalMajorDate GraduationStatus   YearsFromOMD
##  Length:536                Length:536         Length:536
##  Class :character          Class :character   Class :character
##  Mode  :character          Mode  :character   Mode  :character
##
##
##
##
##     CsGrad              4YG                5YG
##  Length:536         Length:536         Length:536
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##      6YG              1_CSCI101       1_MATH111      2_CSCI261
##  Length:536         Min.   :0.300   Min.   :0.30   Min.   :0.300
##  Class :character   1st Qu.:3.000   1st Qu.:3.00   1st Qu.:3.000
##  Mode  :character   Median :4.000   Median :3.00   Median :4.000
##                     Mean   :3.419   Mean   :2.91   Mean   :3.405
##                     3rd Qu.:4.000   3rd Qu.:3.00   3rd Qu.:4.000
##                     Max.   :4.000   Max.   :4.00   Max.   :4.000
##                     NA's   :76      NA's   :12     NA's   :37
##    2_MATH112       2_MATH201       3_CSCI262      3_MATH213
##  Min.   :0.300   Min.   :0.300   Min.   :0.30   Min.   :0.300
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:3.00   1st Qu.:2.000
##  Median :3.000   Median :3.000   Median :4.00   Median :3.000
##  Mean   :2.875   Mean   :2.701   Mean   :3.24   Mean   :2.888
##  3rd Qu.:4.000   3rd Qu.:3.300   3rd Qu.:4.00   3rd Qu.:4.000
##  Max.   :4.000   Max.   :4.000   Max.   :4.00   Max.   :4.000
##  NA's   :26      NA's   :108     NA's   :81     NA's   :49
##    4_CSCI341       4_CSCI358       4_MATH225       5_CSCI306
##  Min.   :0.300   Min.   :0.300   Min.   :0.300   Min.   :0.300
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:3.000
##  Median :3.000   Median :3.000   Median :3.000   Median :3.700
##  Mean   :2.854   Mean   :2.961   Mean   :2.763   Mean   :3.433
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.: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
##    5_CSCI403       5_MATH332       6_CSCI406       7_CSCI370
##  Min.   :0.300   Min.   :0.300   Min.   :0.300   Min.   :2.300
##  1st Qu.:3.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:4.000
##  Median :4.000   Median :3.000   Median :3.000   Median :4.000
##  Mean   :3.565   Mean   :2.671   Mean   :2.795   Mean   :3.895
```

```
##  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
##  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

```
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 ...
```

```
summary(dfDataSet)
```

```
##  Year of OriginalMajorDate       GraduationStatus  YearsFromOMD
##  Min.   :2008              CurrentStudent: 32     Min.   :3.830
##  1st Qu.:2009              Graduated     :396     1st Qu.:4.830
##  Median :2011              InactiveReg   :108     Median :6.840
##  Mean   :2011                                     Mean   :6.706
##  3rd Qu.:2013                                     3rd Qu.:8.840
##  Max.   :2014                                     Max.   :9.840
```

```
##
##         CsGrad     4YG       5YG       6YG        1_CSCI101
##  NG        :140   No :252   No :169   No :150   Min.   :0.300
##  OtherMajor: 57   Yes:284   Yes:367   Yes:386   1st Qu.:3.000
##  Yes       :339                                 Median :4.000
##                                                 Mean   :3.419
##                                                 3rd Qu.:4.000
##                                                 Max.   :4.000
##                                                 NA's   :76
##    1_MATH111      2_CSCI261       2_MATH112       2_MATH201
##  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
##  Median :3.00   Median :4.000   Median :3.000   Median :3.000
##  Mean   :2.91   Mean   :3.405   Mean   :2.875   Mean   :2.701
##  3rd Qu.:3.00   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:3.300
##  Max.   :4.00   Max.   :4.000   Max.   :4.000   Max.   :4.000
##  NA's   :12     NA's   :37      NA's   :26      NA's   :108
##    3_CSCI262      3_MATH213       4_CSCI341       4_CSCI358
##  Min.   :0.30   Min.   :0.300   Min.   :0.300   Min.   :0.300
##  1st Qu.:3.00   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :4.00   Median :3.000   Median :3.000   Median :3.000
##  Mean   :3.24   Mean   :2.888   Mean   :2.854   Mean   :2.961
##  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
##  NA's   :81     NA's   :49      NA's   :108     NA's   :108
##    4_MATH225       5_CSCI306       5_CSCI403       5_MATH332
##  Min.   :0.300   Min.   :0.300   Min.   :0.300   Min.   :0.300
##  1st Qu.:2.000   1st Qu.:3.000   1st Qu.:3.000   1st Qu.:2.000
##  Median :3.000   Median :3.700   Median :4.000   Median :3.000
##  Mean   :2.763   Mean   :3.433   Mean   :3.565   Mean   :2.671
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:3.300
##  Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000
##  NA's   :58      NA's   :132     NA's   :250     NA's   :121
##    6_CSCI406       7_CSCI370       8_CSCI400       9_CSCI442
##  Min.   :0.300   Min.   :2.300   Min.   :0.300   Min.   :0.300
##  1st Qu.:2.000   1st Qu.:4.000   1st Qu.:2.925   1st Qu.:3.000
##  Median :3.000   Median :4.000   Median :3.300   Median :3.300
##  Mean   :2.795   Mean   :3.895   Mean   :3.175   Mean   :3.133
##  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
##  NA's   :154     NA's   :179     NA's   :164     NA's   :165
```

Transform data frame into a table

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

```
##      Year of OriginalMajorDate GraduationStatus YearsFromOMD     CsGrad
##   1:                      2014        Graduated         4.00        Yes
##   2:                      2008        Graduated         9.84 OtherMajor
##   3:                      2008      InactiveReg         9.84         NG
##   4:                      2011      InactiveReg         6.84         NG
##   5:                      2008        Graduated         9.84 OtherMajor
##  ---
## 532:                      2014        Graduated         3.83        Yes
```

```
## 533:                    2014   CurrentStudent         3.83         NG
## 534:                    2014        Graduated         3.83        Yes
## 535:                    2014        Graduated         3.83        Yes
## 536:                    2014        Graduated         3.83        Yes
##      4YG 5YG 6YG 1_CSCI101 1_MATH111 2_CSCI261 2_MATH112 2_MATH201
##   1: Yes Yes Yes         4         3       4.0         2         3
##   2:  No Yes Yes        NA         3       4.0         2         3
##   3:  No  No  No        NA         3       4.0         2         2
##   4:  No  No  No        NA         3       3.3         4        NA
##   5: Yes Yes Yes        NA         3       4.0         3         3
##  ---
## 532: Yes Yes Yes         4         3       4.0         4         4
## 533:  No  No  No         4         4       3.7         4         3
## 534: Yes Yes Yes         4         4       4.0         4         3
## 535: Yes Yes Yes         3         3       4.0         3         3
## 536: Yes Yes Yes         4         4       4.0         4         4
##      3_CSCI262 3_MATH213 4_CSCI341 4_CSCI358 4_MATH225 5_CSCI306 5_CSCI403
##   1:         3         4         2       4.0         4       3.0         4
##   2:        NA         3        NA        NA         3        NA        NA
##   3:         1         2         2       3.0         1        NA        NA
##   4:         3         2         2       2.0         4       3.7         3
##   5:        NA         3        NA        NA         4        NA        NA
##  ---
## 532:         3         3         4       4.0         4       3.7         4
## 533:         3         4         3       4.0         3       4.0         4
## 534:         4         4         4       4.0         4       4.0         4
## 535:         3         3         2       2.7         2       3.3         4
## 536:         4         4         4       4.0         4       4.0         4
##      5_MATH332 6_CSCI406 7_CSCI370 8_CSCI400 9_CSCI442
##   1:       3.0       2.0       3.3       3.3       2.3
##   2:        NA        NA        NA        NA        NA
##   3:       2.0        NA        NA        NA        NA
##   4:       3.0       0.3        NA       3.3        NA
##   5:        NA        NA        NA        NA        NA
##  ---
## 532:       3.3       2.0       4.0       4.0       3.3
## 533:       3.0       3.0        NA       3.7        NA
## 534:       4.0       4.0       4.0       3.7       4.0
## 535:       3.0       2.0       4.0       2.7       2.3
## 536:       4.0       3.0       4.0       4.0       4.0
```

Create a table for variable "GraduationStatus"

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

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

Create table with percentages of total

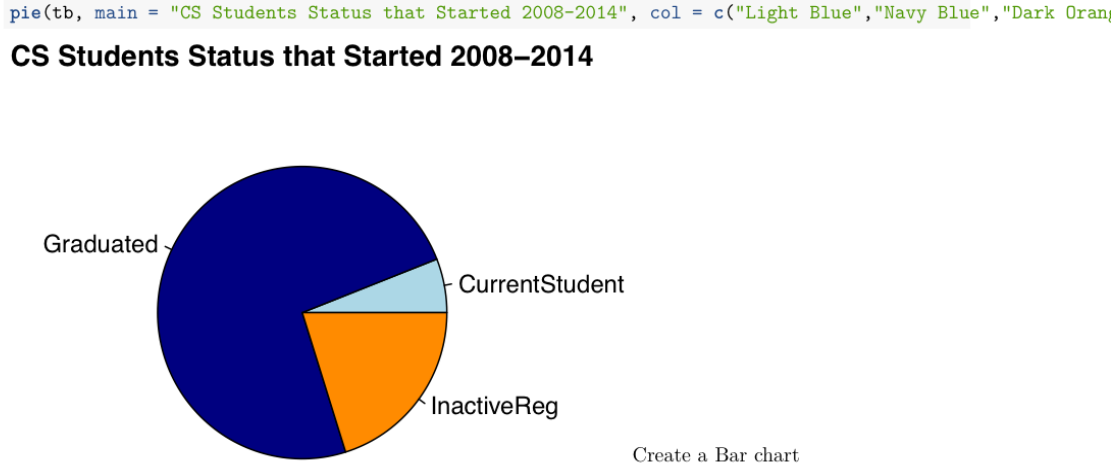
```
tb.prop <- dfDataSet$GraduationStatus %>%
   table() %>%
   prop.table() %>% {. * 100} %>%
   round(2)
```

```
tb.prop
```

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

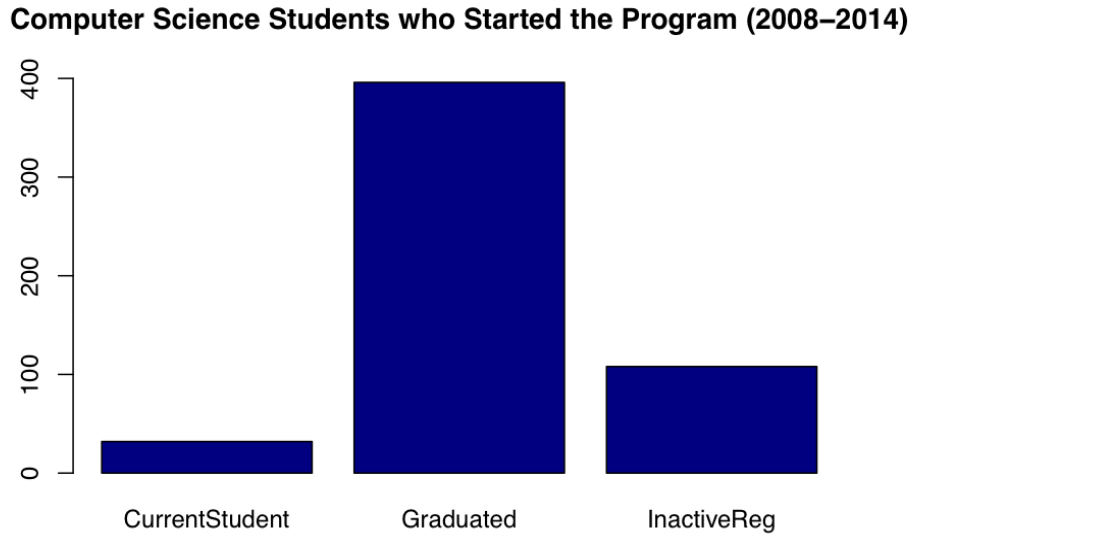
Create a Pie chart

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



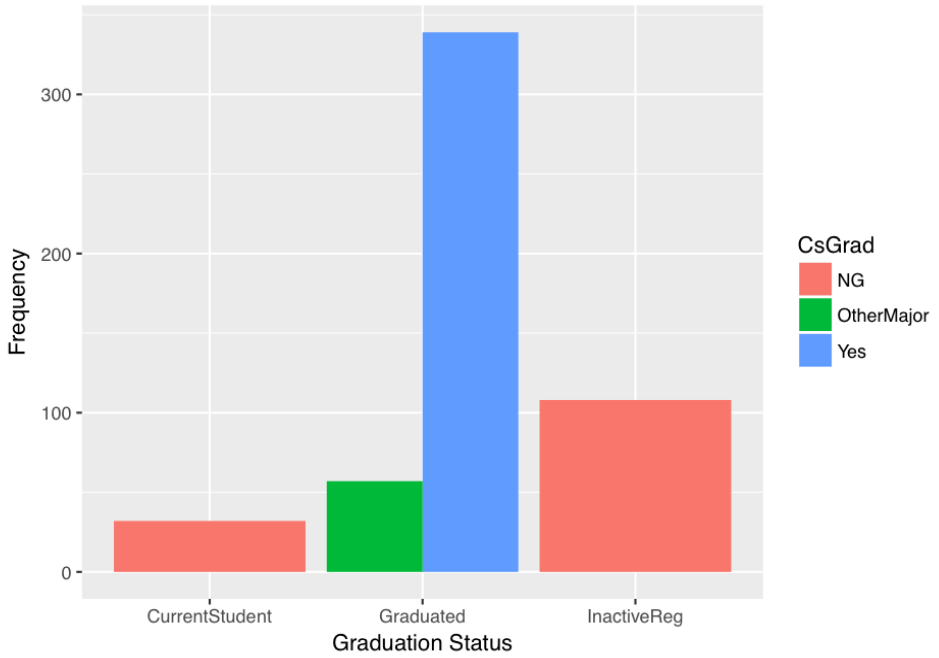
Create a Bar chart

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



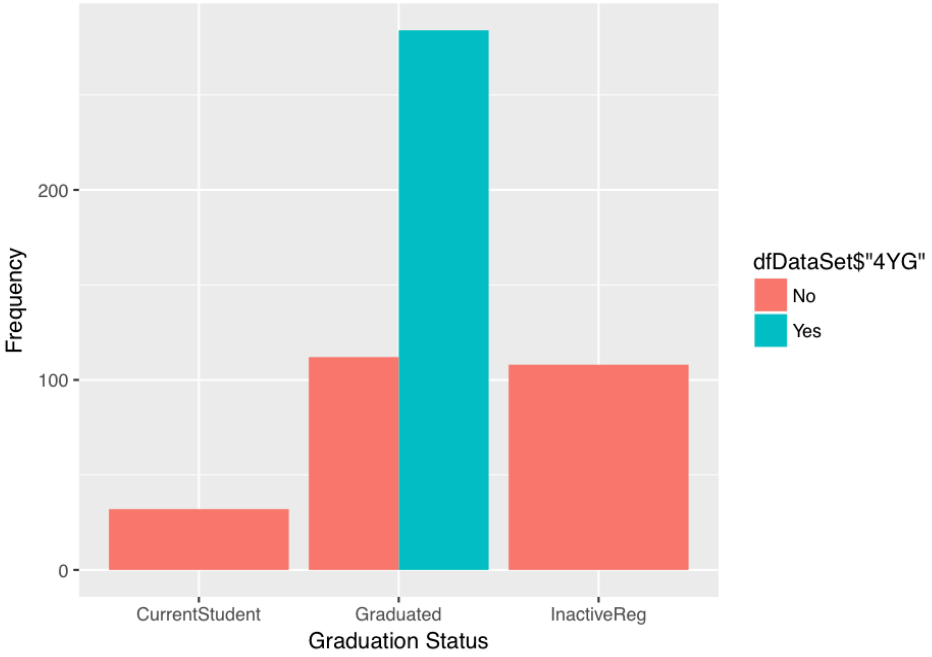
Create a Bar plot where the "graduated"" students are divided into CS and Other Major students

```
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

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



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

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

```
##
##        No       Yes
## 0.4701493 0.5298507
```

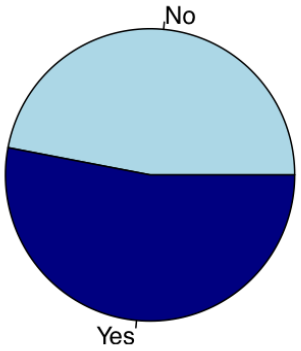
```
cbind(tb4YG,prop.table(tb4YG))
```

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

Create a Pie Chart for Four-year graduation rate

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

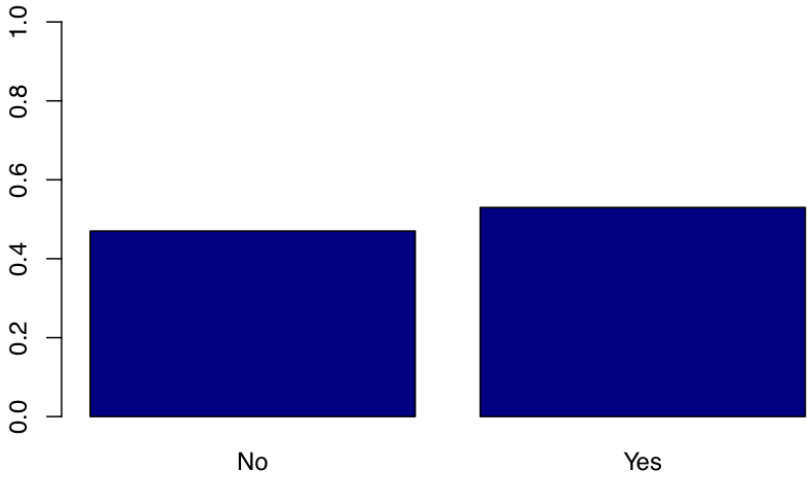
CS Students that Graduated in Four Years



Create a Bar chart for Four-year graduation rate

```
barplot(tb4YGpercent, col = "navy blue", ylim=c(0,1),main = "Four Year Graduation Rate for CS Students
```

Four Year Graduation Rate for CS Students who Started on 2008–20



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

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

```
##
##        No       Yes
## 0.3152985 0.6847015
```

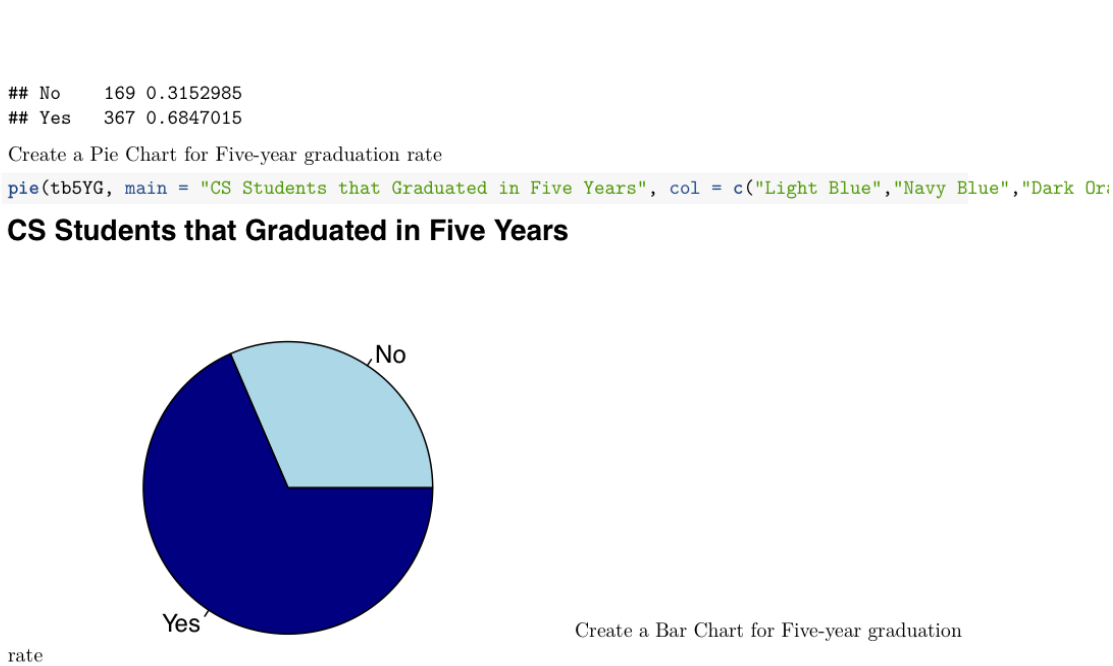
```
cbind(tb5YG,prop.table(tb5YG))
```

```
##     tb5YG
```

```
## No    169 0.3152985
## Yes   367 0.6847015
```

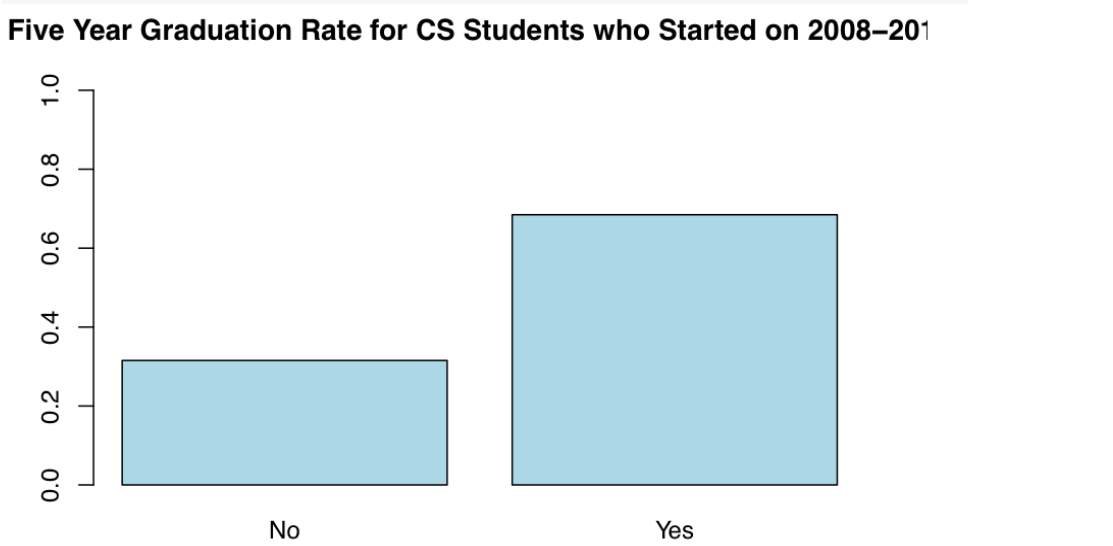
Create a Pie Chart for Five-year graduation rate

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



Create a Bar Chart for Five-year graduation rate

```
barplot(tb5YGpercent, col = "lightblue", ylim=c(0,1),main = "Five Year Graduation Rate for CS Students
```



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

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

```
##
```

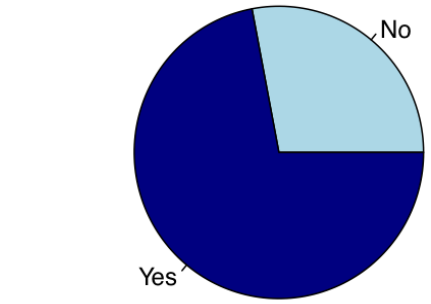
```
##        No       Yes
## 0.2798507 0.7201493
```

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

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

Create a Pie Chart for Five-year graduation rate

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



Create a Bar Chart for Six-year graduation rate

```
barplot(tb6YGpercent, col = "navy blue", ylim=c(0,1), main = "Six Year Graduation Rate for CS Students
```