

MDAIR Presentation-Data Manipulation and Visualization in R

Laura Walker, Institutional Research Data Analyst, University System of Maryland—April 27, 2018

Table of Contents

Download R Studio	3
Opening Packages & Setting the Working Directory.....	3
Uploading A Dataframe and a list.....	3
Head of a Data Frame	4
Viewing the nth Row of Data.....	4
First Three Rows of a Dataframe	5
Last Five Rows of a Dataframe	5
Summary Function.....	6
The Structure of our Dataframe	8
What Do We Do With NAs?.....	9
Adding Fields.....	9
Rearranging Fields By Field Names.....	9
Visualizations	10
Comparison–Column Chart	10
Tweaking the Column Chart Part 1	11
Tweaking the Column Chart Part 2	12
Distribution–Histogram	13
Tweaking the Histogram Part 1-Adding a Title	14
Tweaking the Histogram Part 2-Centering the Title	15
Tweaking the Histogram Part 3-Editing the Axis Names	16
Tweaking the Histogram Part 4-Editing the Y-Axis Limits.....	17
Association–Two-Variable Scatter Plot	18
Tweaking the Scatter Plot-Part 1-That Grey Background Though!.....	19
Tweaking the Scatter Plot-Part 2-Axis Lines.....	20
Tweaking the Scatter Plot-Part 3-Color– added “, color=Inst_Sector” to first layer	21
Tweaking the Scatter Plot-Part 4-Editing the Legend	22

Tweaking the Scatter Plot-Part 5-Wrapping the Legend, added “color=str_wrap(Inst_Sector,20)”	23
Tweaking the Scatter Plot-Part 6-Custom Colors.....	24
Tweaking the Scatter Plot-Part 7-Size	25
Review of Components	26
Scatter Plot-Final	27
Composition–Stacked 100% Chart	28
Stacked 100% Chart.....	29
Tweaking the Stacked 100% Chart-Part 1-Legend	30
Tweaking the Stacked 100% Chart-Part 2-The X Axis	31
Tweaking the Stacked 100% Chart-Part 3-Let’s Group this!	32
Review of Components	33
Stacked 100% Chart-Final.....	34
Helpful Websites.....	35
For Extra Challenges	35

Download R Studio

Opening Packages & Setting the Working Directory

```
# Opening Packages  
library(ggplot2) # visualizations  
library(knitr) # R Markdown files  
library(scales) # Stacked Percent Chart  
library(readxl) # Uploading Data  
library(stringr) # Legend title wrapping  
library(tidyr) # rearranging data for Stacked 100% Chart  
  
# Setting the Working Directory  
  
# 1 Open Windows Explorer  
# 2 Navigate to C Drive  
# 3 Select "Users"  
# 4 Select Desktop  
# 5 Copy the Address **BE SURE TO REPLACE THE "\" WITH "/"  
setwd("C:/Users/alwalker/Desktop")  
  
## REPLACE "alwalker" WITH YOUR MACHINE'S USERNAME
```

Uploading A Dataframe and a list

```
IPEDS <- read_excel("Hands-on Workshop 1-Laura Walker-Data Manipulation and Visualization  
in R-IPEDS Source Document.xlsx")  
manual_colors <- c("#101820", "#EAAA00", "#AF272F")
```

Head of a Data Frame

```
head(IPEDS)
```

```
## # A tibble: 6 x 21
##   UnitID Inst_Name      Inst_Sector  Count_Doctor_Deg Count_Masters_D~
##   <dbl> <chr>          <chr>          <dbl>          <dbl>
## 1 161688. Allegany Colle~ Public, 2-year      NA            NA
## 2 161767. Anne Arundel C~ Public, 2-year      NA            NA
## 3 161864. Baltimore City~ Public, 2-year      NA            NA
## 4 162007. Bowie State Un~ Public, 4-yea~      10.          337.
## 5 405872. Carroll Commun~ Public, 2-year      NA            NA
## 6 162104. Cecil College  Public, 2-year      NA            NA
## # ... with 16 more variables: Count_Bachelors_Deg <dbl>,
## #   Count_Assoc_Deg <dbl>, Count_PostBac_PostMasters_Deg <dbl>,
## #   Count_Cert <dbl>, Count_Cert_Less_Year <dbl>,
## #   Grand_Total_All_Students <dbl>, Grand_Total_Female <dbl>,
## #   Grand_Total_Amer_Ind_AK_Native <dbl>, Grand_Total_Asian <dbl>,
## #   Grand_Total_Black_AA <dbl>, Grand_Total_Hispanic <dbl>,
## #   Grand_Total_Native_Hawaiian_PI <dbl>, Grand_Total_White <dbl>,
## #   Grand_Total_Two_More <dbl>, Grand_Total_Race_Unknown <dbl>,
## #   Grand_Total_Nonresident_Alien <dbl>
```

Viewing the nth Row of Data

```
# Someone asked how one would view a specific row. Below is the code to see the "tenth" row
```

```
IPEDS[10,]
```

```
## # A tibble: 1 x 21
##   UnitID Inst_Name      Inst_Sector  Count_Doctor_Deg Count_Masters_D~
##   <dbl> <chr>          <chr>          <dbl>          <dbl>
## 1 162557. Frederick Commun~ Public, 2-y~      NA            NA
## # ... with 16 more variables: Count_Bachelors_Deg <dbl>,
## #   Count_Assoc_Deg <dbl>, Count_PostBac_PostMasters_Deg <dbl>,
## #   Count_Cert <dbl>, Count_Cert_Less_Year <dbl>,
## #   Grand_Total_All_Students <dbl>, Grand_Total_Female <dbl>,
## #   Grand_Total_Amer_Ind_AK_Native <dbl>, Grand_Total_Asian <dbl>,
## #   Grand_Total_Black_AA <dbl>, Grand_Total_Hispanic <dbl>,
## #   Grand_Total_Native_Hawaiian_PI <dbl>, Grand_Total_White <dbl>,
## #   Grand_Total_Two_More <dbl>, Grand_Total_Race_Unknown <dbl>,
## #   Grand_Total_Nonresident_Alien <dbl>
```

First Three Rows of a Dataframe

```
head(IPEDS, 3)

## # A tibble: 3 x 21
##   UnitID Inst_Name      Inst_Sector  Count_Doctor_Deg Count_Masters_D~
##   <dbl> <chr>          <chr>          <dbl>          <dbl>
## 1 161688. Allegany College~ Public, 2-y~      NA            NA
## 2 161767. Anne Arundel Com~ Public, 2-y~      NA            NA
## 3 161864. Baltimore City C~ Public, 2-y~      NA            NA
## # ... with 16 more variables: Count_Bachelors_Deg <dbl>,
## #   Count_Assoc_Deg <dbl>, Count_PostBac_PostMasters_Deg <dbl>,
## #   Count_Cert <dbl>, Count_Cert_Less_Year <dbl>,
## #   Grand_Total_All_Students <dbl>, Grand_Total_Female <dbl>,
## #   Grand_Total_Amer_Ind_AK_Native <dbl>, Grand_Total_Asian <dbl>,
## #   Grand_Total_Black_AA <dbl>, Grand_Total_Hispanic <dbl>,
## #   Grand_Total_Native_Hawaiian_PI <dbl>, Grand_Total_White <dbl>,
## #   Grand_Total_Two_More <dbl>, Grand_Total_Race_Unknown <dbl>,
## #   Grand_Total_Nonresident_Alien <dbl>
```

Last Five Rows of a Dataframe

```
tail(IPEDS, 5)

## # A tibble: 5 x 21
##   UnitID Inst_Name      Inst_Sector  Count_Doctor_Deg Count_Masters_D~
##   <dbl> <chr>          <chr>          <dbl>          <dbl>
## 1 163286. University of~ Public, 4-year~    628.         2795.
## 2 163204. University of~ Public, 4-year~    46.          3667.
## 3 162210. Washington Ad~ Private not-fo~    0.            86.
## 4 164216. Washington Co~ Private not-fo~    0.            3.
## 5 164313. Wor-Wic Commu~ Public, 2-year     NA            NA
## # ... with 16 more variables: Count_Bachelors_Deg <dbl>,
## #   Count_Assoc_Deg <dbl>, Count_PostBac_PostMasters_Deg <dbl>,
## #   Count_Cert <dbl>, Count_Cert_Less_Year <dbl>,
## #   Grand_Total_All_Students <dbl>, Grand_Total_Female <dbl>,
## #   Grand_Total_Amer_Ind_AK_Native <dbl>, Grand_Total_Asian <dbl>,
## #   Grand_Total_Black_AA <dbl>, Grand_Total_Hispanic <dbl>,
## #   Grand_Total_Native_Hawaiian_PI <dbl>, Grand_Total_White <dbl>,
## #   Grand_Total_Two_More <dbl>, Grand_Total_Race_Unknown <dbl>,
## #   Grand_Total_Nonresident_Alien <dbl>
```

Summary Function

summary(IPEDS)

```
##      UnitID      Inst_Name      Inst_Sector      Count_Doctor_Deg
##  Min.    :161688  Length:40      Length:40      Min.    : 0.00
##  1st Qu.:162489  Class :character  Class :character  1st Qu.: 0.00
##  Median :163125  Mode  :character  Mode  :character  Median : 5.50
##  Mean   :175867                      Mean   :115.88
##  3rd Qu.:163706                      3rd Qu.: 73.75
##  Max.   :434672                      Max.   :796.00
##
##                      NA's    :16
##  Count_Masters_Deg  Count_Bachelors_Deg  Count_Assoc_Deg
##  Min.    : 3.0      Min.    : 86.0      Min.    : 0.0
##  1st Qu.: 122.8    1st Qu.: 365.5    1st Qu.: 0.0
##  Median : 267.0    Median : 647.0    Median : 0.0
##  Mean   : 746.8    Mean   :1343.0    Mean   : 402.6
##  3rd Qu.: 665.2    3rd Qu.:1141.5    3rd Qu.: 611.5
##  Max.   :4953.0    Max.   :6748.0    Max.   :2519.0
##  NA's   :16        NA's   :16
##  Count_PostBac_PostMasters_Deg  Count_Cert  Count_Cert_Less_Year
##  Min.    : 0                      Min.    : 0.00  Min.    : 0.00
##  1st Qu.: 0                      1st Qu.: 0.00  1st Qu.: 0.00
##  Median : 18                      Median : 0.00  Median : 0.00
##  Mean   : 95                      Mean   : 58.95  Mean   : 49.38
##  3rd Qu.: 78                      3rd Qu.: 64.75  3rd Qu.: 54.75
##  Max.   :743                      Max.   :684.00  Max.   :479.00
##  NA's   :16
##  Grand_Total_All_Students  Grand_Total_Female
##  Min.    : 451                Min.    : 197
##  1st Qu.: 2607                1st Qu.: 1838
##  Median : 5078                Median : 3112
##  Mean   : 8800                Mean   : 4914
##  3rd Qu.: 8911                3rd Qu.: 5173
##  Max.   :50248                Max.   :24177
##
##  Grand_Total_Amer_Ind_AK_Native  Grand_Total_Asian  Grand_Total_Black_AA
##  Min.    : 1.00                Min.    : 2.0      Min.    : 11
##  1st Qu.: 7.75                 1st Qu.: 57.0     1st Qu.: 377
##  Median : 12.50                Median : 107.0     Median : 1088
##  Mean   : 28.65                Mean   : 604.5     Mean   : 2337
```

```

## 3rd Qu.: 26.50          3rd Qu.: 538.5    3rd Qu.: 2816
## Max.    :234.00        Max.    :5156.0    Max.    :14924
##
## Grand_Total_Hispanic Grand_Total_Native_Hawaiian_PI Grand_Total_White
## Min.    : 12.0         Min.    : 0.00         Min.    : 46
## 1st Qu.: 128.2        1st Qu.: 2.00         1st Qu.: 1092
## Median  : 219.0        Median   : 5.50         Median   : 2414
## Mean    : 701.5        Mean     : 20.75         Mean     : 3992
## 3rd Qu.: 691.5        3rd Qu.: 16.75        3rd Qu.: 4658
## Max.    :5732.0        Max.     :356.00        Max.     :19921
##
## Grand_Total_Two_More Grand_Total_Race_Unknown
## Min.    : 0.00         Min.    : 1.0
## 1st Qu.: 63.25        1st Qu.: 54.0
## Median  : 172.00       Median   : 94.0
## Mean    : 296.75       Mean     : 347.7
## 3rd Qu.: 441.25       3rd Qu.: 244.5
## Max.    :1873.00       Max.     :4617.0
##
## Grand_Total_Nonresident_Alien
## Min.    : 7.0
## 1st Qu.: 34.0
## Median  : 151.0
## Mean    : 470.9
## 3rd Qu.: 331.0
## Max.    :4531.0
##

```

The Structure of our Dataframe

```
str(IPEDS)

## Classes 'tbl_df', 'tbl' and 'data.frame':   40 obs. of  21 variables:
## $ UnitID          : num  161688 161767 161864 162007 405872 ...
## $ Inst_Name       : chr   "Allegany College of Maryland" "Anne Arundel C
ommunity College" "Baltimore City Community College" "Bowie State University" ...
## $ Inst_Sector     : chr   "Public, 2-year" "Public, 2-year" "Public, 2-y
ear" "Public, 4-year or above" ...
## $ Count_Doctor_Deg : num   NA NA NA 10 NA NA NA NA 0 NA ...
## $ Count_Masters_Deg : num   NA NA NA 337 NA NA NA NA 77 NA ...
## $ Count_Bachelors_Deg : num   NA NA NA 832 NA NA NA NA 464 NA ...
## $ Count_Assoc_Deg  : num   438 1717 423 0 611 ...
## $ Count_PostBac_PostMasters_Deg : num   NA NA NA 62 NA NA NA NA 0 NA ...
## $ Count_Cert       : num   91 218 37 0 20 5 14 684 0 70 ...
## $ Count_Cert_Less_Year : num   74 408 69 0 9 60 53 100 0 116 ...
## $ Grand_Total_All_Students : num  3091 14689 4726 5430 3542 ...
## $ Grand_Total_Female : num  2010 8721 3222 3414 2159 ...
## $ Grand_Total_Amer_Ind_AK_Native : num   6 60 7 5 9 11 26 35 1 23 ...
## $ Grand_Total_Asian : num   8 558 123 75 57 35 27 243 12 286 ...
## $ Grand_Total_Black_AA : num  353 2472 3634 4432 145 ...
## $ Grand_Total_Hispanic : num  45 972 134 155 139 129 92 489 62 662 ...
## $ Grand_Total_Native_Hawaiian_PI : num   3 42 6 8 1 0 3 32 2 7 ...
## $ Grand_Total_White : num  2557 8821 353 199 3039 ...
## $ Grand_Total_Two_More : num   58 541 86 184 79 94 45 448 52 244 ...
## $ Grand_Total_Race_Unknown : num   27 1072 92 110 66 ...
## $ Grand_Total_Nonresident_Alien : num   34 151 291 262 7 14 20 34 370 36 ...
```


What Do We Do With NAs?

```
# Structure of calling specific columns in R: Data_frame_Name$Variable_Name
# Changing NAs to 0s for a specific column
IPEDS$Count_Doctor_Deg[is.na(IPEDS$Count_Doctor_Deg)] <- 0
# Changing NAs to 0s for the entire dataframe
IPEDS[is.na(IPEDS)] <- 0
```

Adding Fields

```
IPEDS$Grand_Total_Male <- IPEDS$Grand_Total_All_Students - IPEDS$Grand_Total_Female
colnames(IPEDS)
```

```
## [1] "UnitID" "Inst_Name"
## [3] "Inst_Sector" "Count_Doctor_Deg"
## [5] "Count_Masters_Deg" "Count_Bachelors_Deg"
## [7] "Count_Assoc_Deg" "Count_PostBac_PostMasters_Deg"
## [9] "Count_Cert" "Count_Cert_Less_Year"
## [11] "Grand_Total_All_Students" "Grand_Total_Female"
## [13] "Grand_Total_Amer_Ind_AK_Native" "Grand_Total_Asian"
## [15] "Grand_Total_Black_AA" "Grand_Total_Hispanic"
## [17] "Grand_Total_Native_Hawaiian_PI" "Grand_Total_White"
## [19] "Grand_Total_Two_More" "Grand_Total_Race_Unknown"
## [21] "Grand_Total_Nonresident_Alien" "Grand_Total_Male"
```

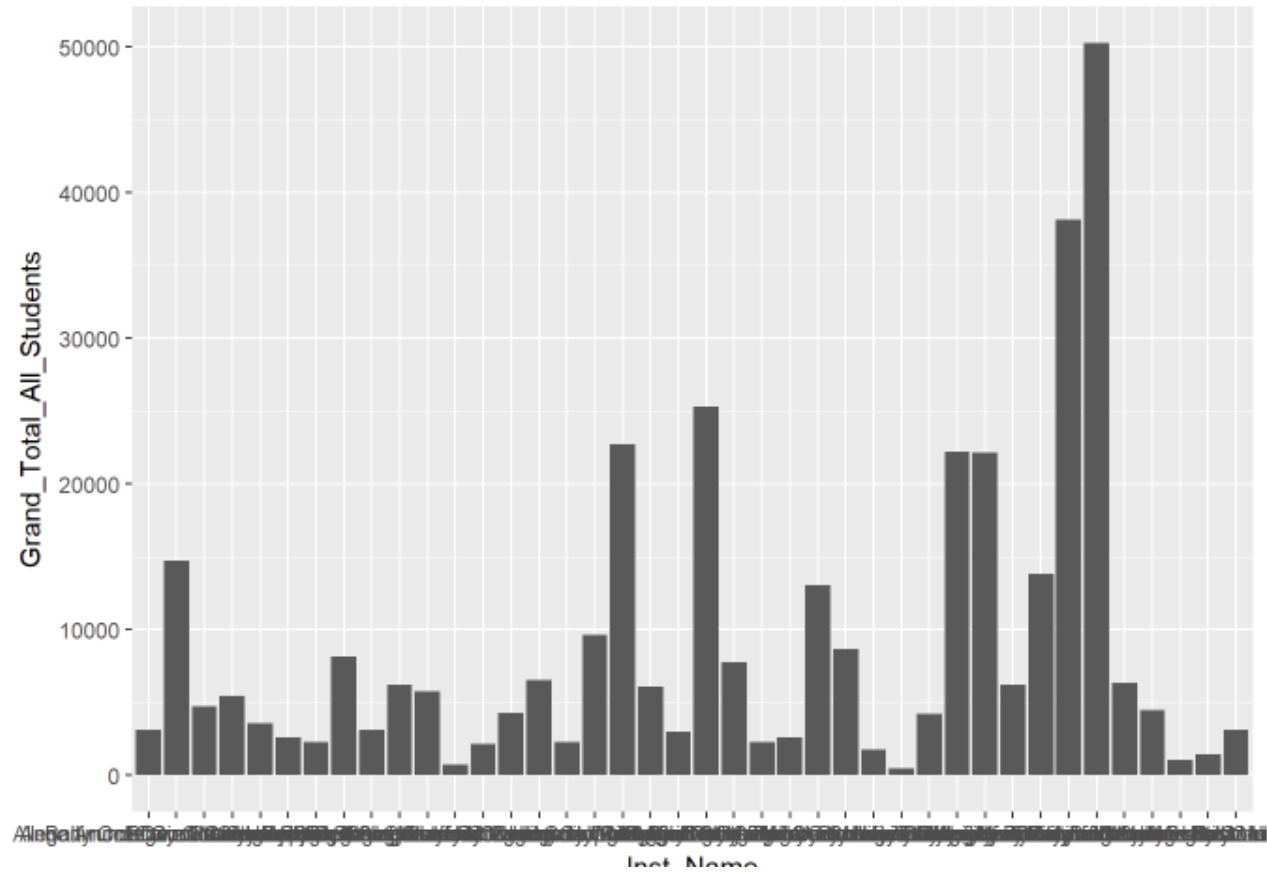
Rearranging Fields By Field Names

```
#Moving the new variable created, Grand_Total_Male in between "Grand_Total_All_Students"
& "Grand_Total_Female"
IPEDS <- IPEDS[c("UnitID", "Inst_Name", "Inst_Sector", "Count_Doctor_Deg", "Count_Masters
_Deg", "Count_Bachelors_Deg", "Count_Assoc_Deg", "Count_PostBac_PostMasters_Deg", "Count_
Cert", "Count_Cert_Less_Year", "Grand_Total_All_Students", "Grand_Total_Male", "Grand_Tot
al_Female", "Grand_Total_Amer_Ind_AK_Native", "Grand_Total_Asian", "Grand_Total_Black_AA"
, "Grand_Total_Hispanic", "Grand_Total_Native_Hawaiian_PI", "Grand_Total_White", "Grand_T
otal_Two_More", "Grand_Total_Race_Unknown", "Grand_Total_Nonresident_Alien")]
```

Visualizations

Comparison–Column Chart

```
vis_column_chart <-ggplot(IPEDS, aes(Inst_Name, Grand_Total_All_Students)) + geom_bar(sta  
t = "identity")  
vis_column_chart
```

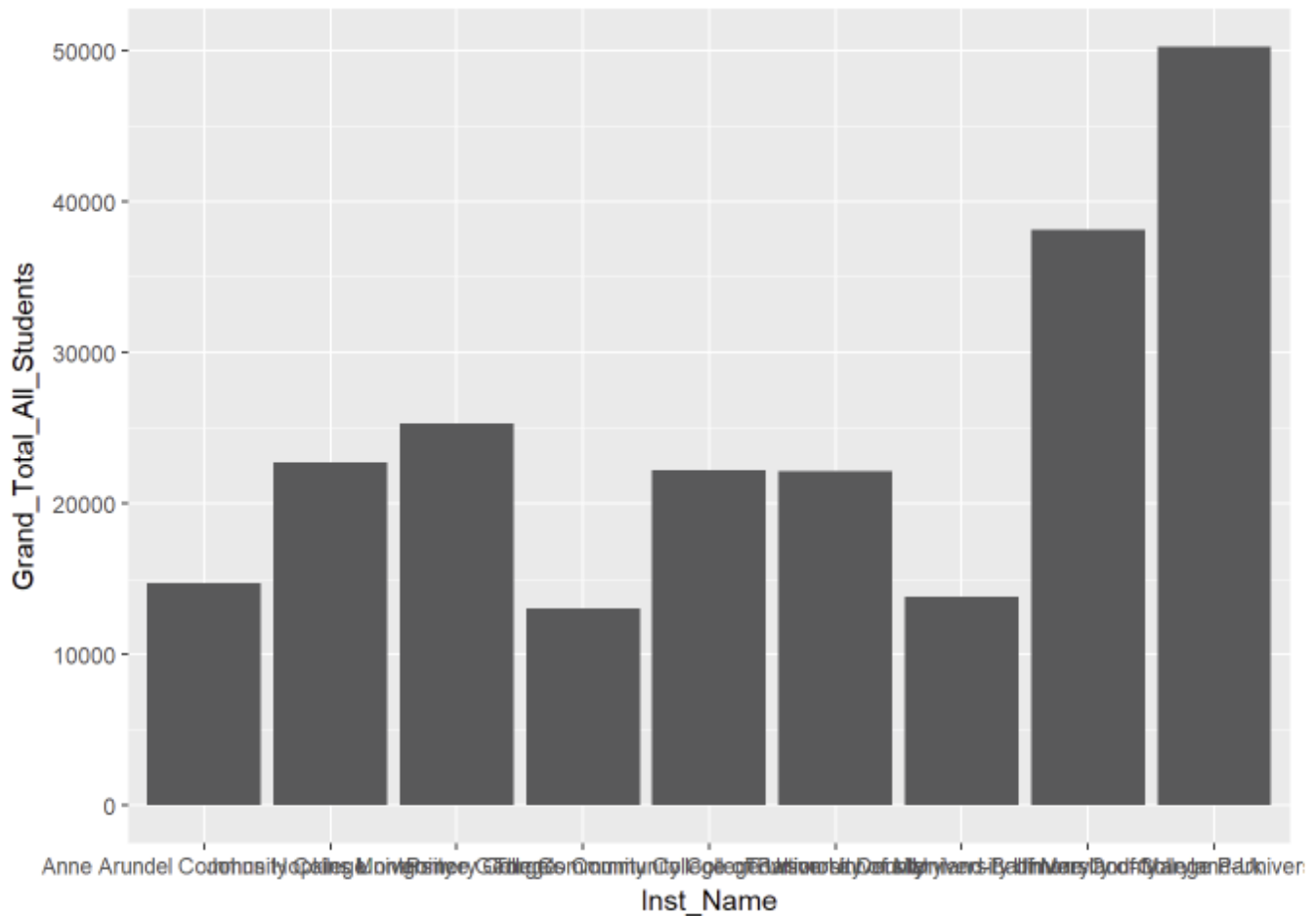


Tweaking the Column Chart Part 1

```
IPEDS_large_pops <- subset(IPEDS, Grand_Total_All_Students > 10000)

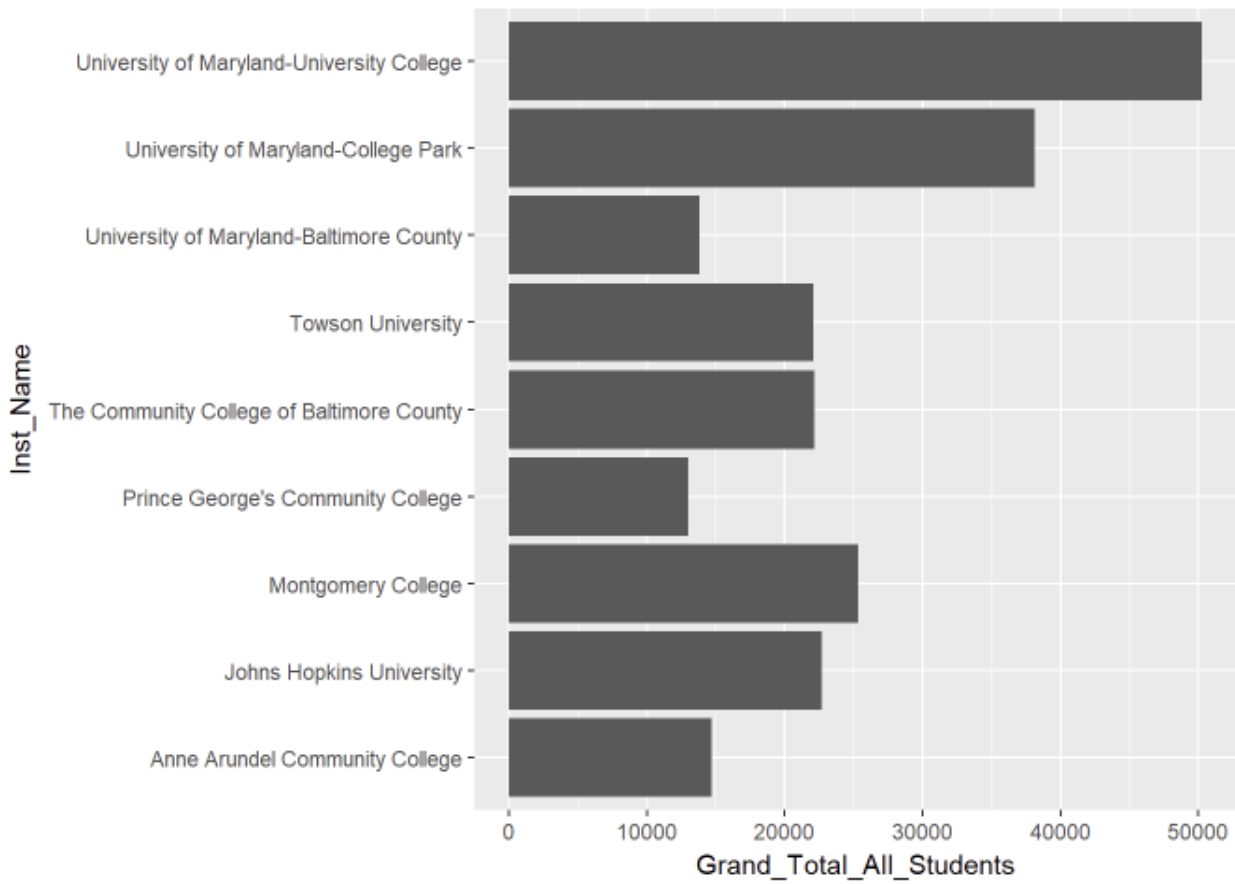
vis_column_chart_v2 <- ggplot(IPEDS_large_pops, aes(Inst_Name, Grand_Total_All_Students))
+ geom_bar(stat = "identity")

vis_column_chart_v2
```



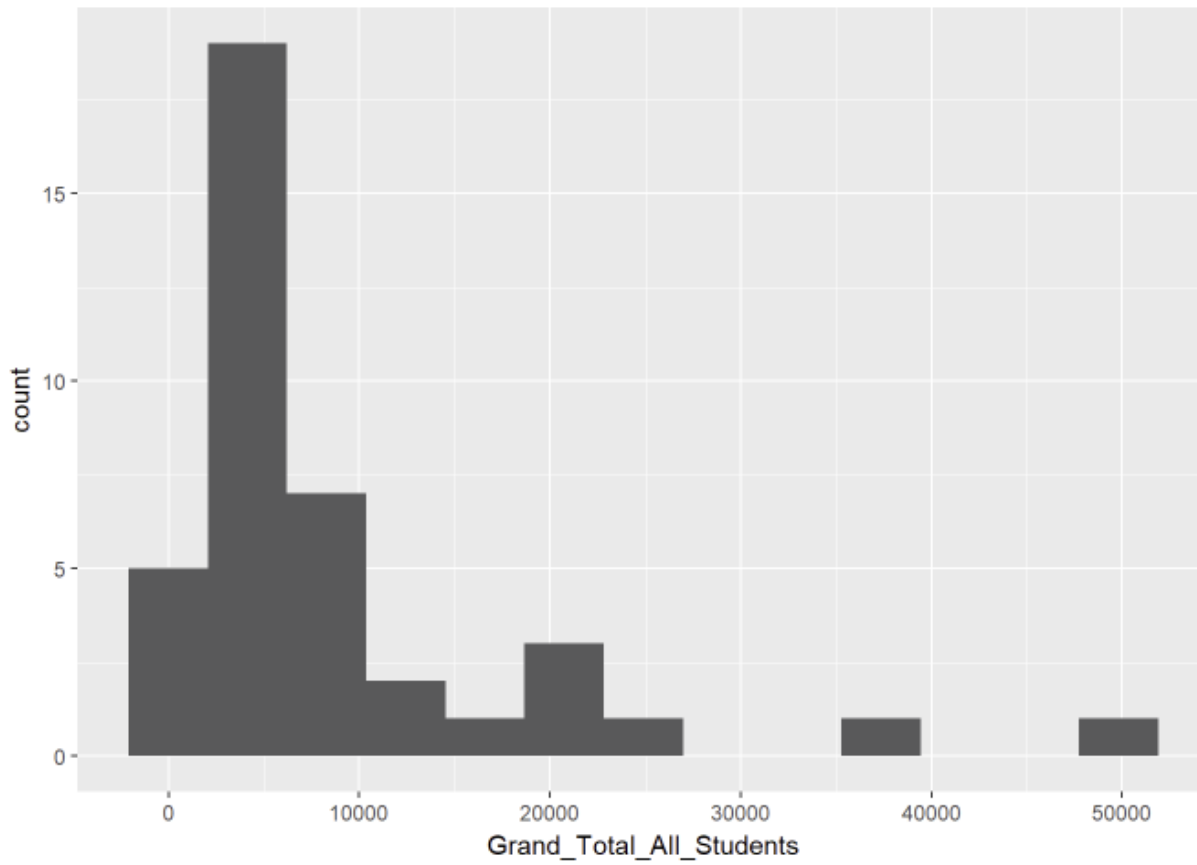
Tweaking the Column Chart Part 2

```
vis_column_chart_horiz <- vis_column_chart_v2 + coord_flip()  
vis_column_chart_horiz
```



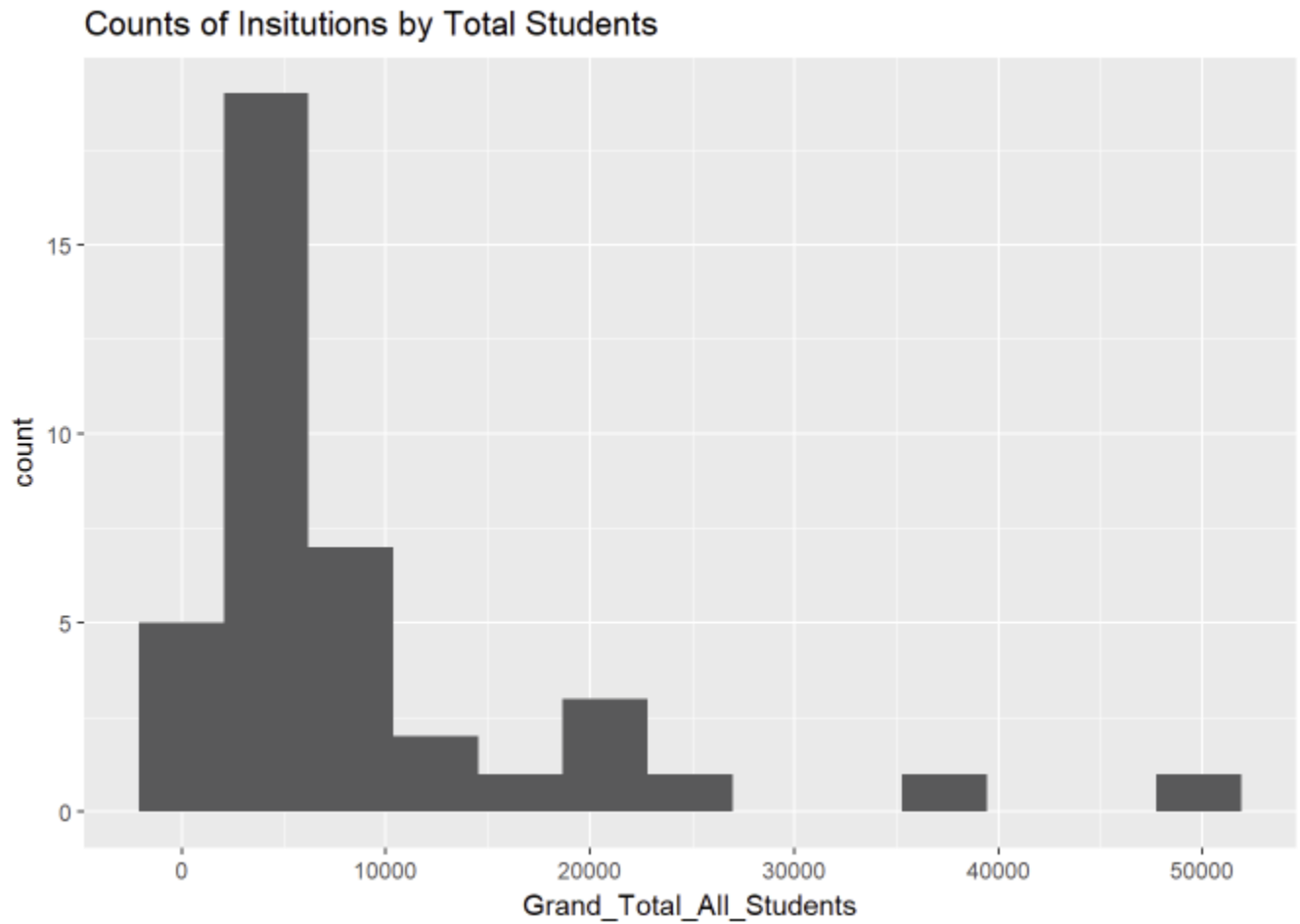
Distribution–Histogram

```
vis_histogram <- ggplot(IPEDS, aes(Grand_Total_All_Students)) + geom_histogram(bins = 13)  
vis_histogram
```



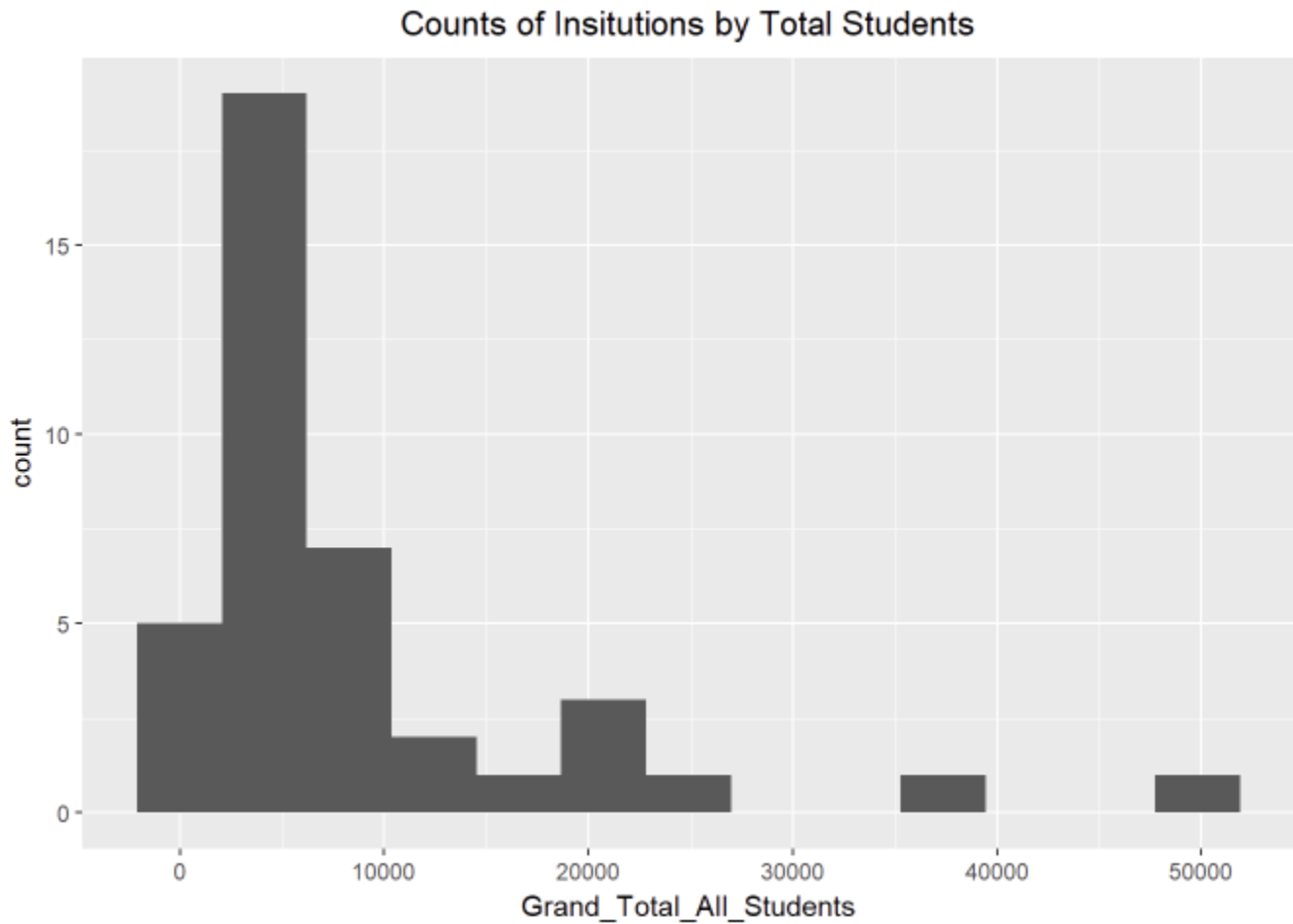
Tweaking the Histogram Part 1-Adding a Title

```
vis_histogram_w_title <- vis_histogram + labs(title="Counts of Insitutions by Total Students") # Adds a Title to the graphic  
vis_histogram_w_title
```



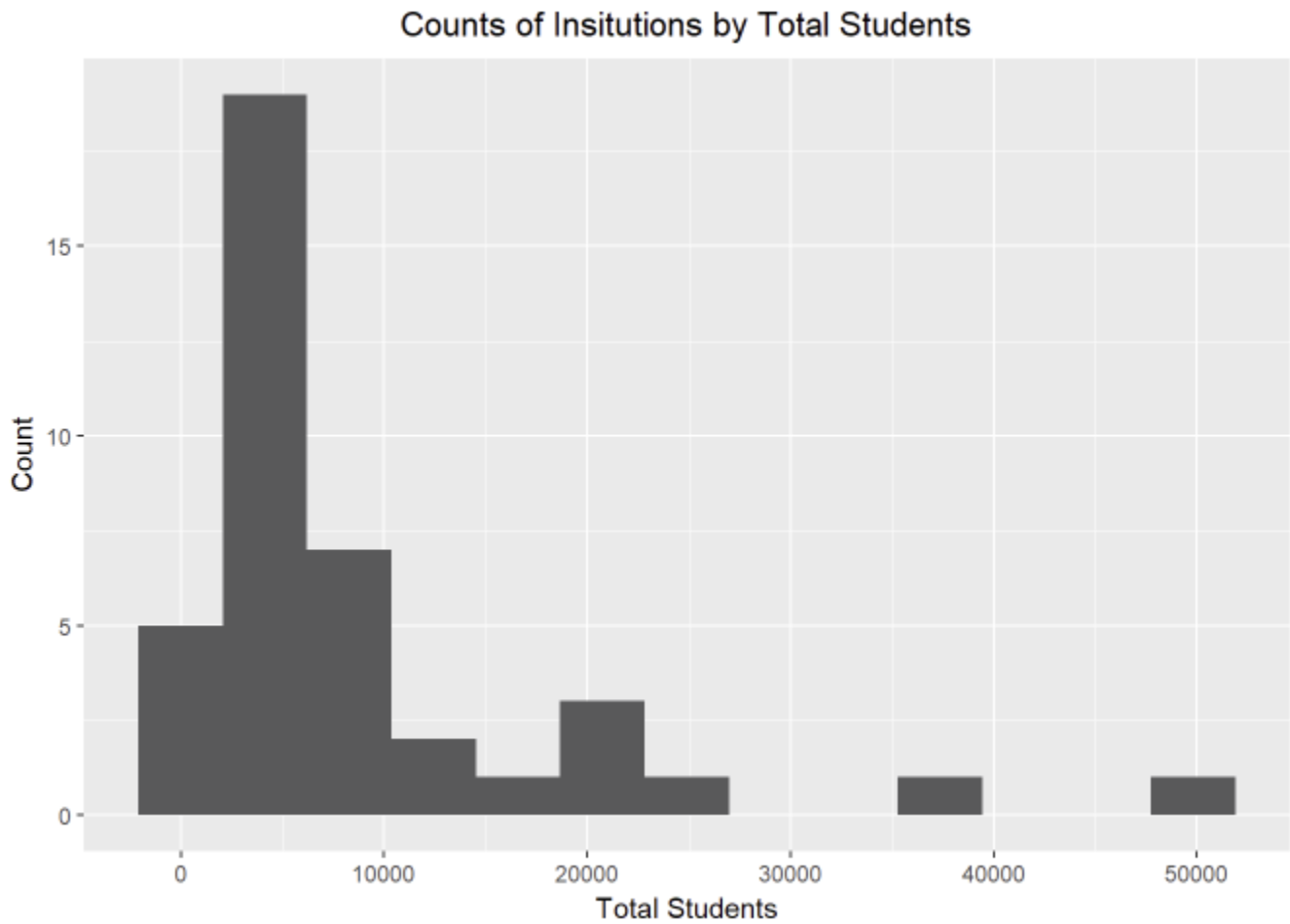
Tweaking the Histogram Part 2-Centering the Title

```
vis_histogram_w_centered_title <- vis_histogram_w_title + theme(plot.title = element_text  
(hjust = 0.5)) # Centers the Title  
vis_histogram_w_centered_title
```



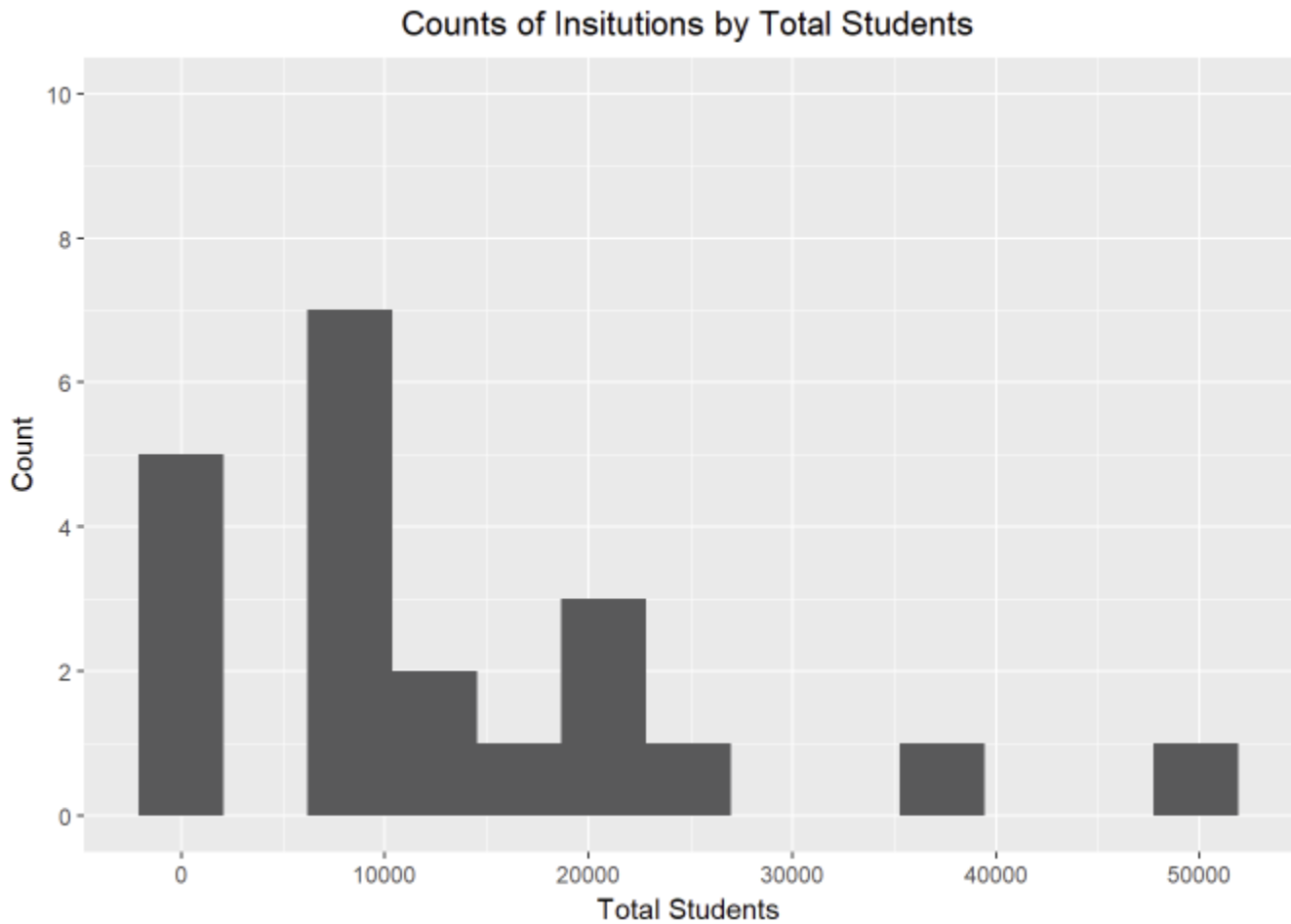
Tweaking the Histogram Part 3-Editing the Axis Names

```
vis_histogram_w_centered_title_edited_axes <- vis_histogram_w_centered_title + labs(x="Total Students",y="Count") # Changes axis names  
vis_histogram_w_centered_title_edited_axes
```



Tweaking the Histogram Part 4-Editing the Y-Axis Limits

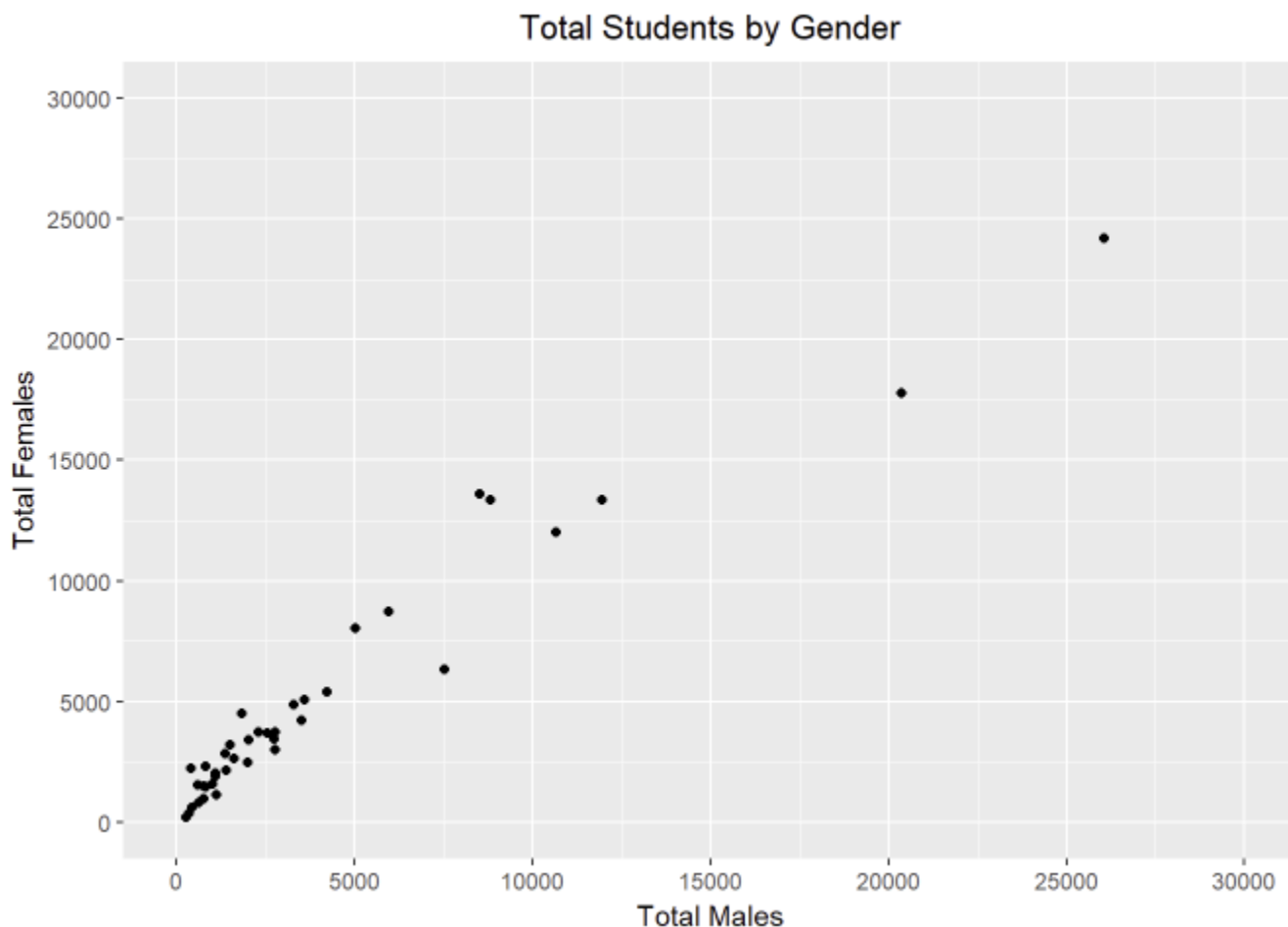
```
vis_histogram_w_edited_y_axis <- vis_histogram_w_centered_title_edited_axes + scale_y_continuous(breaks = seq(0, 10, by=2), limits=c(0,10)) # sets the y-axis from 0 to 10 in 2 increments  
vis_histogram_w_edited_y_axis
```



Association—Two-Variable Scatter Plot

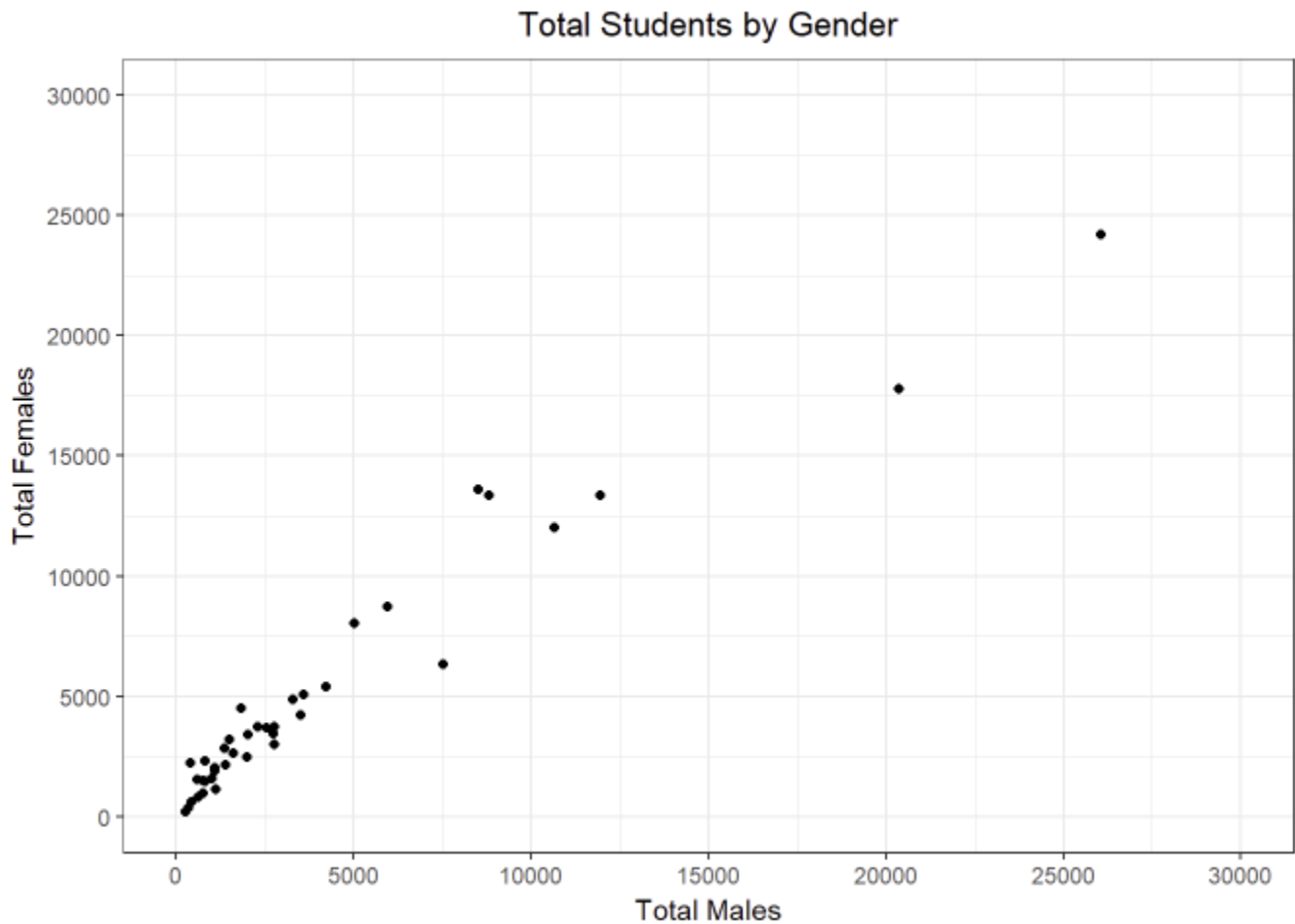
```
vis_scatter <- ggplot(IPEDS, aes(x=Grand_Total_Male, y=Grand_Total_Female)) + geom_point(  
  ) + labs(title="Total Students by Gender") + theme(plot.title = element_text(hjust = 0.5)  
  ) + labs(x="Total Males",y="Total Females") + scale_y_continuous(breaks = seq(0, 30000, b  
y=5000), limits=c(0,30000)) + scale_x_continuous(breaks = seq(0, 30000, by=5000), limits=  
c(0,30000))
```

```
vis_scatter
```



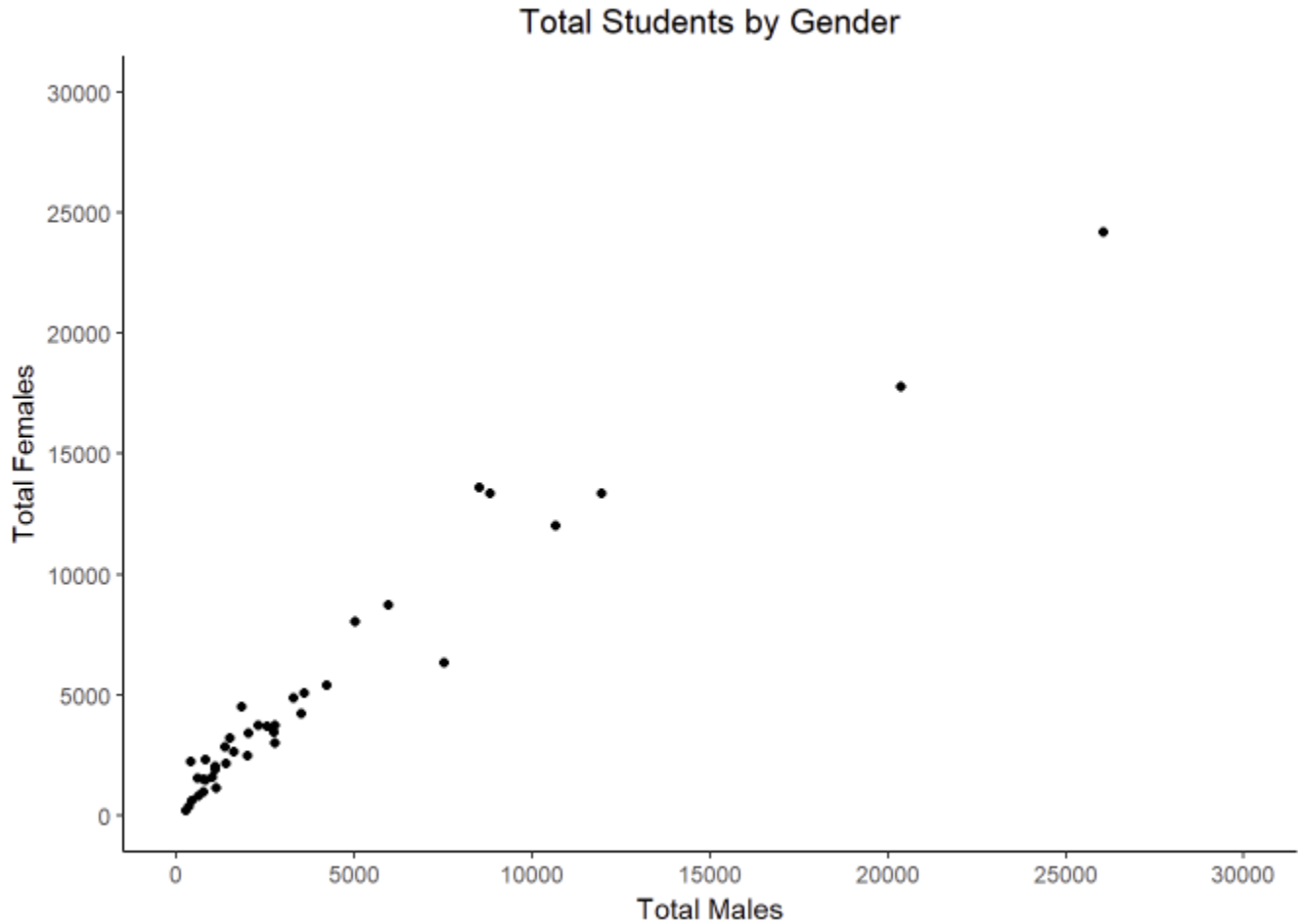
Tweaking the Scatter Plot-Part 1-That Grey Background Though!

```
# "theme_bw()" gets rid of the grey background
vis_scatter_background_edited <- vis_scatter + theme_bw() + theme(plot.title = element_text(hjust = 0.5))
vis_scatter_background_edited
```



Tweaking the Scatter Plot-Part 2-Axis Lines

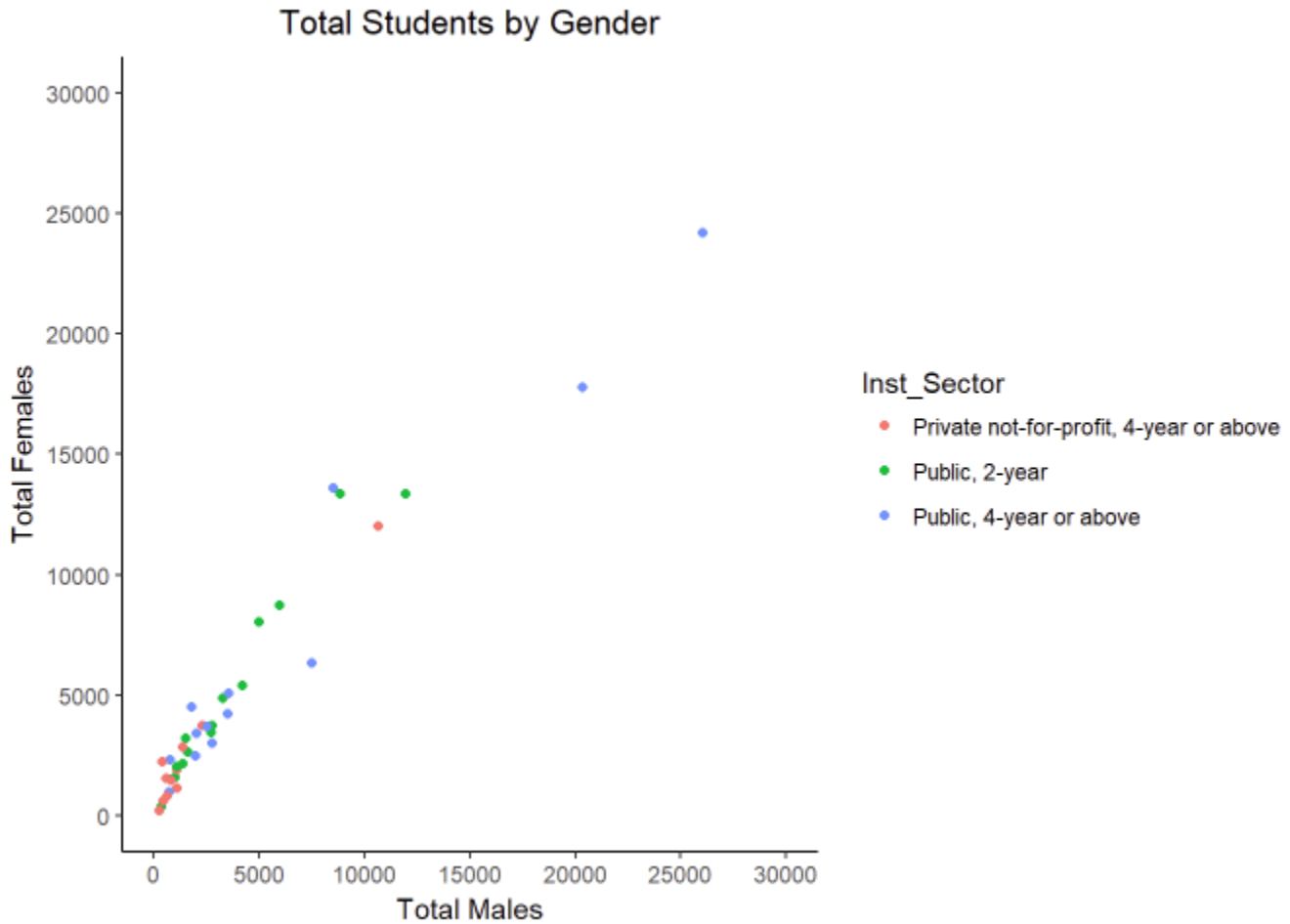
```
vis_scatter_axis_lines <- vis_scatter_background_edited + theme_bw() + theme(panel.border = element_blank(), panel.grid.major = element_blank(), panel.grid.minor = element_blank(), axis.line = element_line(colour = "black")) + theme(plot.title = element_text(hjust = 0.5))  
vis_scatter_axis_lines
```



Tweaking the Scatter Plot-Part 3-Color- added “, color=Inst_Sector” to first layer

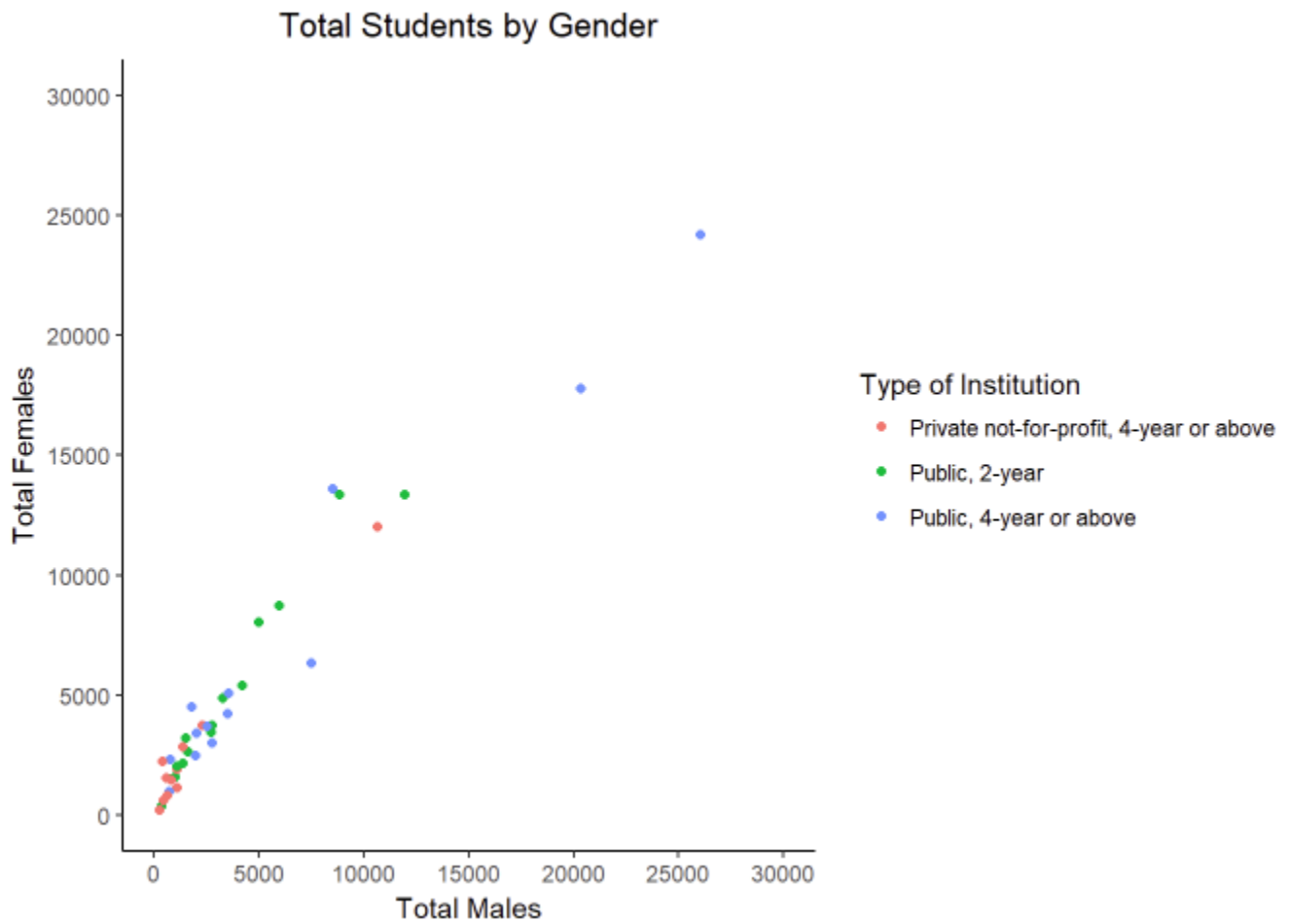
```
vis_scatter_point_colors <- ggplot(IPEDS, aes(x=Grand_Total_Male, y=Grand_Total_Female, c  
olor=Inst_Sector)) + geom_point() + labs(title="Total Students by Gender") + labs(x="Tota  
l Males", y="Total Females") + scale_y_continuous(breaks = seq(0, 30000, by=5000), limits=  
c(0, 30000)) + scale_x_continuous(breaks = seq(0, 30000, by=5000), limits=c(0, 30000)) + th  
eme_bw() + theme(panel.border = element_blank(), panel.grid.major = element_blank(), pane  
l.grid.minor = element_blank(), axis.line = element_line(colour = "black")) + theme(plot.  
title = element_text(hjust = 0.5))
```

```
vis_scatter_point_colors
```



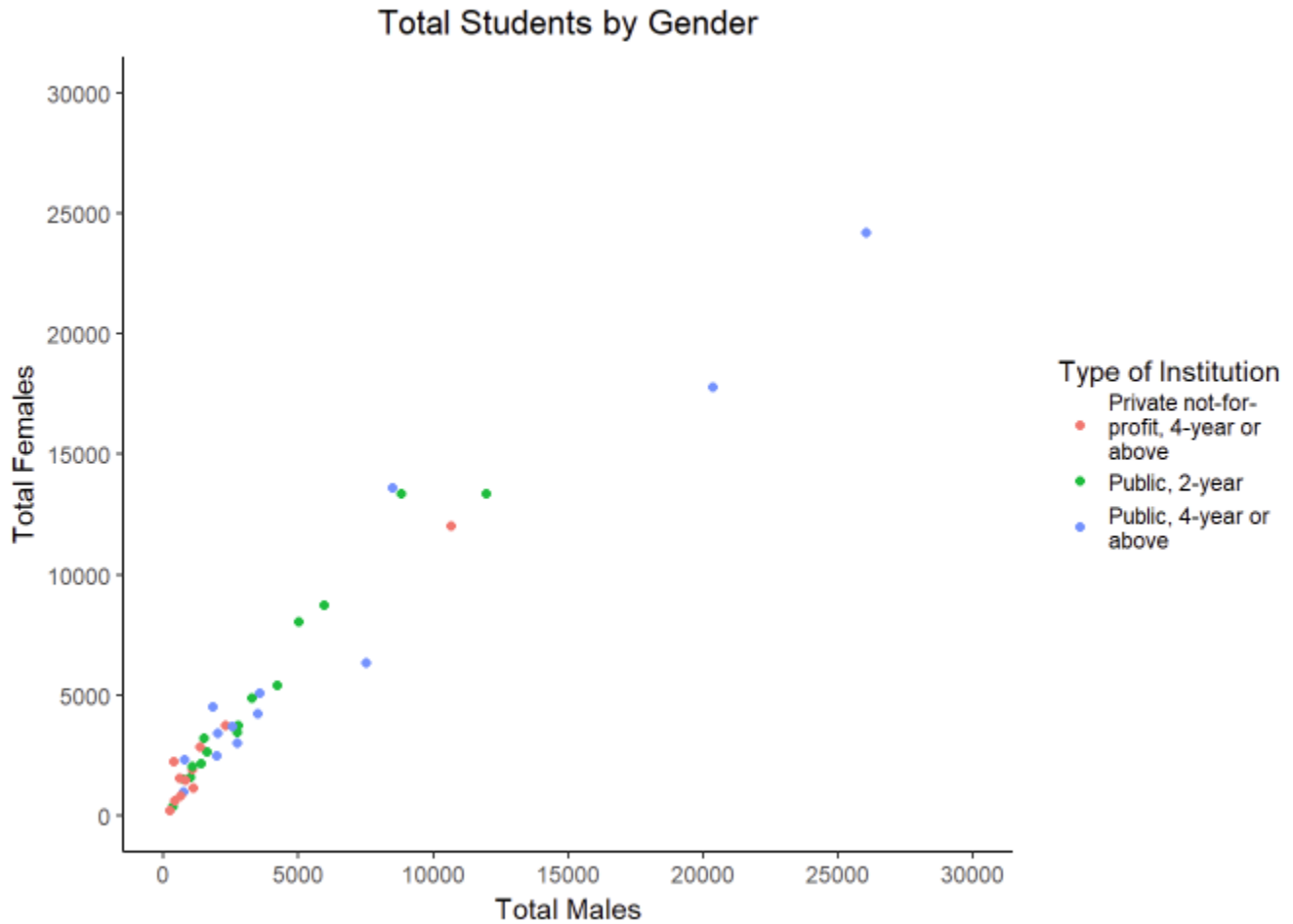
Tweaking the Scatter Plot-Part 4-Editing the Legend

```
vis_scatter_edited_legend <- vis_scatter_point_colors + labs(color = "Type of Institution")  
vis_scatter_edited_legend
```



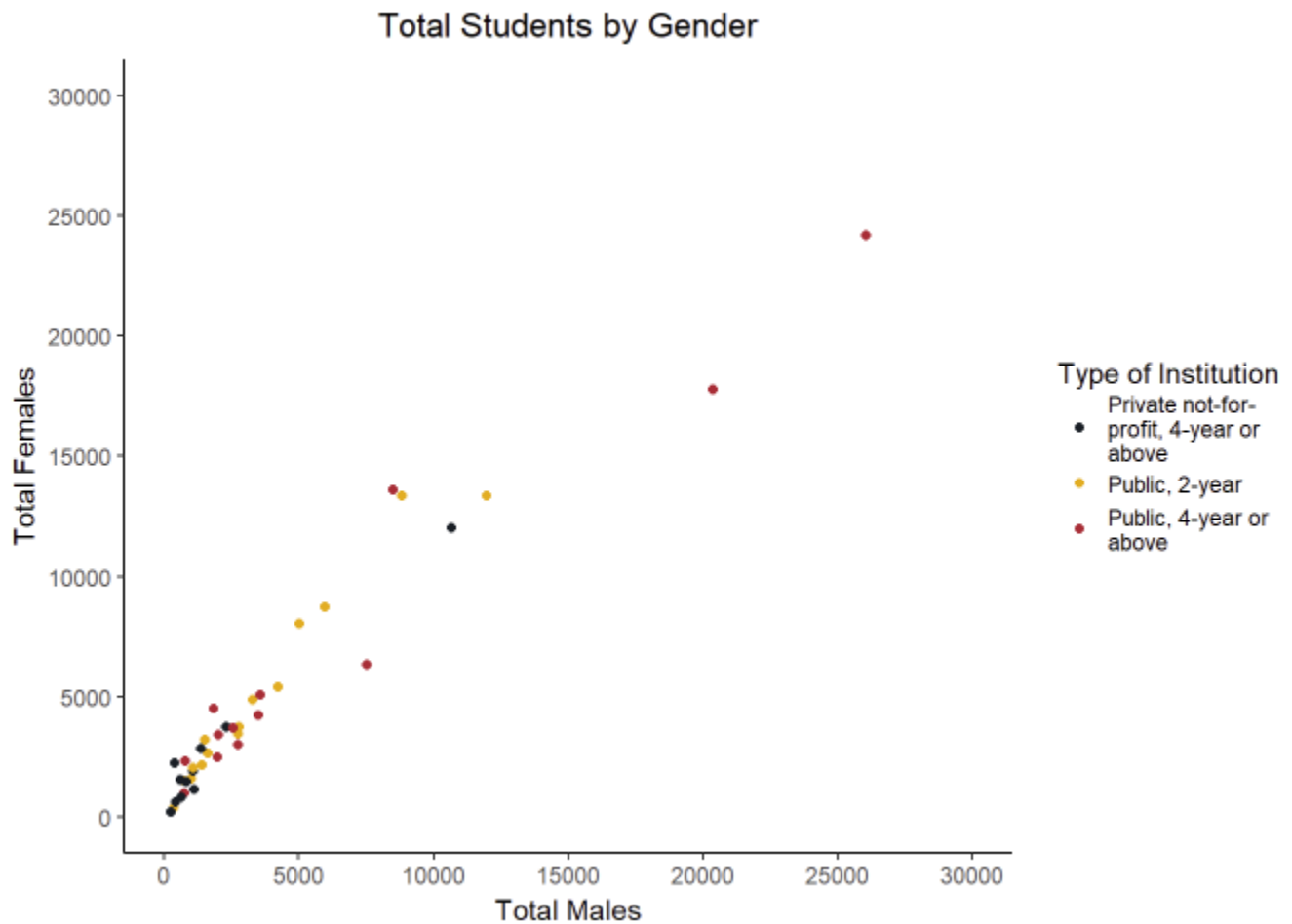
Tweaking the Scatter Plot-Part 5-Wrapping the Legend, added "color=str_wrap(Inst_Sector,20)"

```
vis_scatter_wrapped_legend <- ggplot(IPEDS, aes(x=Grand_Total_Male, y=Grand_Total_Female,  
  color=str_wrap(Inst_Sector,20))) + geom_point() + labs(title="Total Students by Gender")  
  + labs(x="Total Males",y="Total Females") + scale_y_continuous(breaks = seq(0, 30000, by  
=5000), limits=c(0,30000)) + scale_x_continuous(breaks = seq(0, 30000, by=5000), limits=c  
(0,30000)) + theme_bw() + theme(panel.border = element_blank(), panel.grid.major = elemen  
t_blank(), panel.grid.minor = element_blank(), axis.line = element_line(colour = "black")  
) + theme(plot.title = element_text(hjust = 0.5)) + labs(color = "Type of Institution")  
vis_scatter_wrapped_legend
```



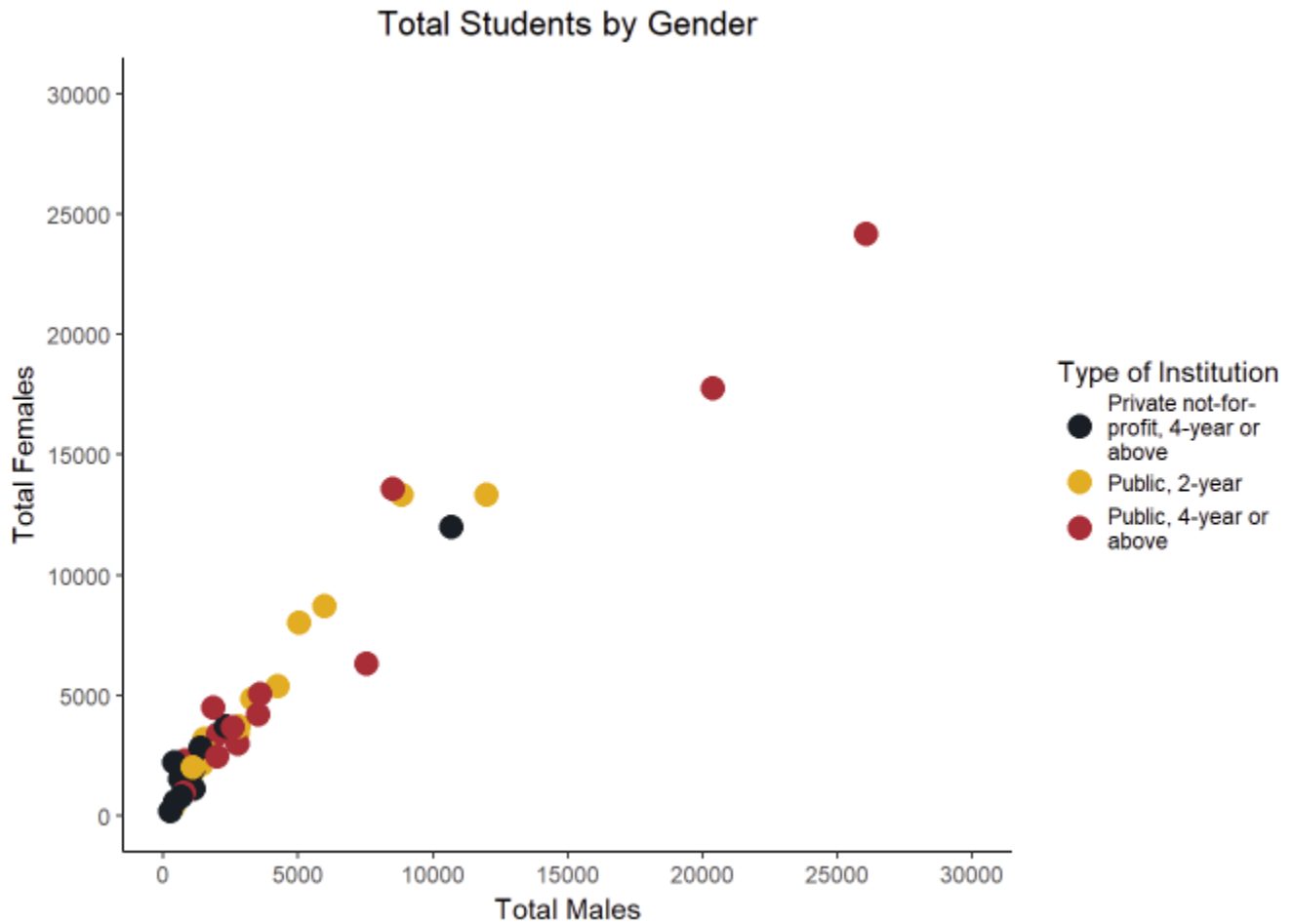
Tweaking the Scatter Plot-Part 6-Custom Colors

```
vis_scatter_custom_colors <- vis_scatter_wrapped_legend + scale_color_manual(values = manual_colors)
vis_scatter_custom_colors
```



Tweaking the Scatter Plot-Part 7-Size

```
vis_scatter_adj_size <- vis_scatter_custom_colors + geom_point(size = 4)  
vis_scatter_adj_size
```



Review of Components

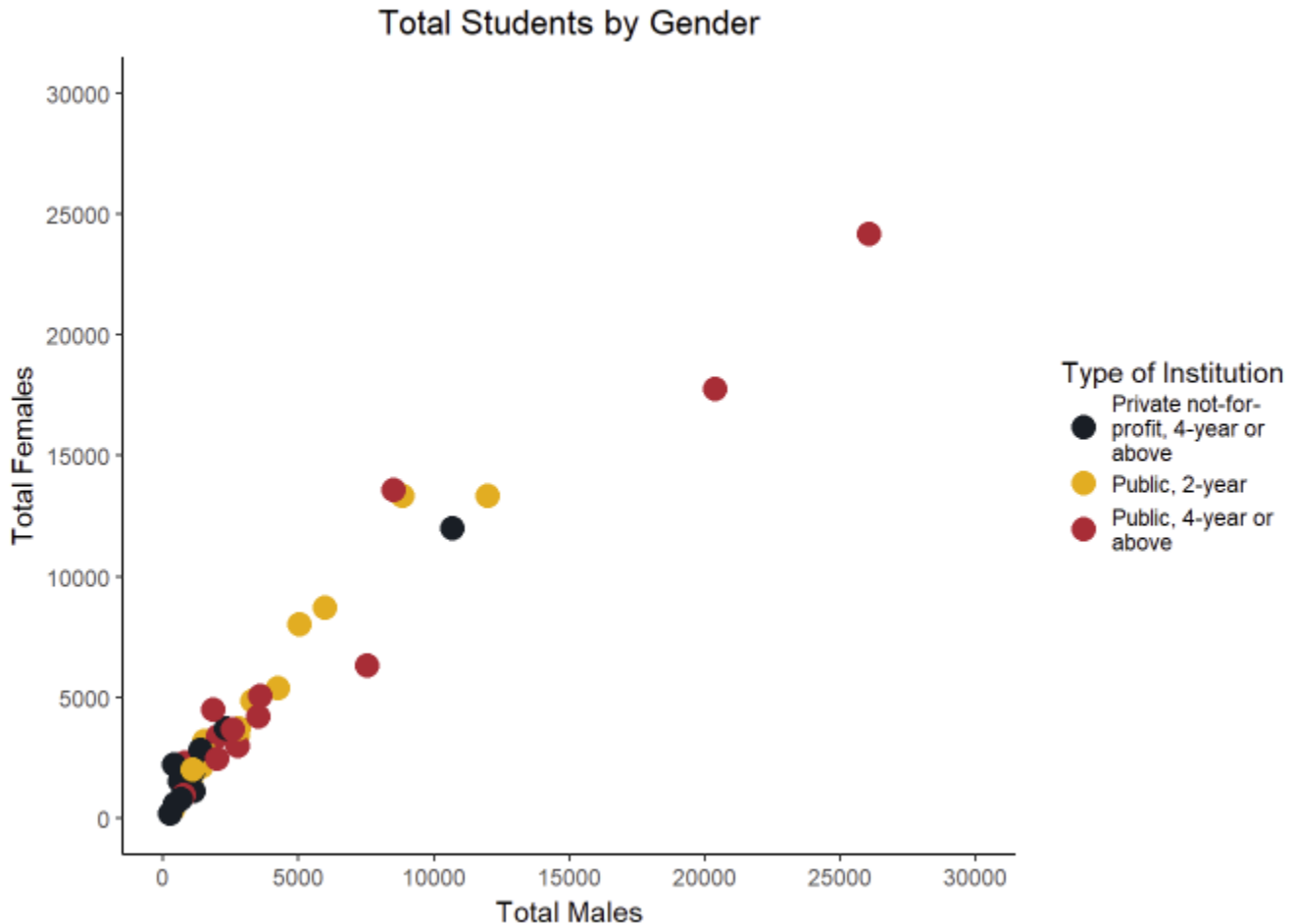
```
vis_scatter <- ggplot(IPEDS, aes(x=Grand_Total_Male, y=Grand_Total_Female, color=str_wrap
(Inst_Sector,20)))

#Sets what variables are on each axis
# str_wrap(Inst_Sector,20) - Allows the legend to be wrapped
+ geom_point() # Makes this plot a scatter plot
+ labs(title="Total Students by Gender") # Adds Plot Title
+ theme(plot.title = element_text(hjust = 0.5)) # Centers Plot Title
+ labs(x="Total Males",y="Total Females") # Sets Axis Names
+ scale_y_continuous(breaks = seq(0, 30000, by=5000), limits=c(0,30000)) # Sets Y Axis li
mits. Min=0 Max=30000 Increments=5000
+ scale_x_continuous(breaks = seq(0, 30000, by=5000), limits=c(0,30000)) # Sets X Axis li
mits. Min=0 Max=30000 Increments=5000
+ theme_bw() # Gets rid of grey background
+ theme(panel.border = element_blank(), panel.grid.major = element_blank(),
panel.grid.minor = element_blank(), axis.line = element_line(colour = "black"))
# panel.border = element_blank() - removes border around entire output
# panel.grid.major = element_blank() - removes major gridlines
# panel.grid.minor = element_blank() - removes minor gridlines
# axis.line = element_line(colour = "black") - adds black axis line
+ labs(color = "Type of Institution") # Changes the Legend Title
+ scale_color_manual(values = manual_colors) # Changes the dot colors to our custom color
s
+ geom_point(size = 4) # Increases the size of the dots
+ theme(plot.title = element_text(hjust = 0.5))
```

Scatter Plot-Final

```
vis_scatter_final <- ggplot(IPEDS, aes(x=Grand_Total_Male, y=Grand_Total_Female, color=st
r_wrap(Inst_Sector,20))) + geom_point() + labs(title="Total Students by Gender") + theme(
plot.title = element_text(hjust = 0.5)) + labs(x="Total Males",y="Total Females") + scale
_y_continuous(breaks = seq(0, 30000, by=5000), limits=c(0,30000))+ scale_x_continuou
s(breaks = seq(0, 30000, by=5000), limits=c(0,30000)) + theme_bw() + theme(panel.bor
der = element_blank(), panel.grid.major = element_blank(), panel.grid.minor = elem
ent_blank(), axis.line = element_line(colour = "black")) + labs(color = "Type of Institution")+ scale_color
_manual(values = manual_colors)+ geom_point(size = 4) + theme(plot.title = element_text(
hjust = 0.5))

vis_scatter_final
```



Composition–Stacked 100% Chart

```
IPEDS_small_pops <- subset(IPEDS, Grand_Total_All_Students < 5000)

# Goal: Construct a Stacked 100% Chart on Degree/Certificate Counts by Institution and,
eventually, sector.

# Step 1-Move our data from wide to short. We will need data transposed from wide to
short. Race and Gender Data is not needed for this graph.

# The arguments to gather():
# - data: the name of the dataframe that needs to be transposed.
# - key: Name of new key column (made from names of data columns), in this case the
various degree/certificate types
# - value: Name of new value column
# - ...: Names of source columns that contain values
# - factor_key: Treat the new key column as a factor (instead of character vector)

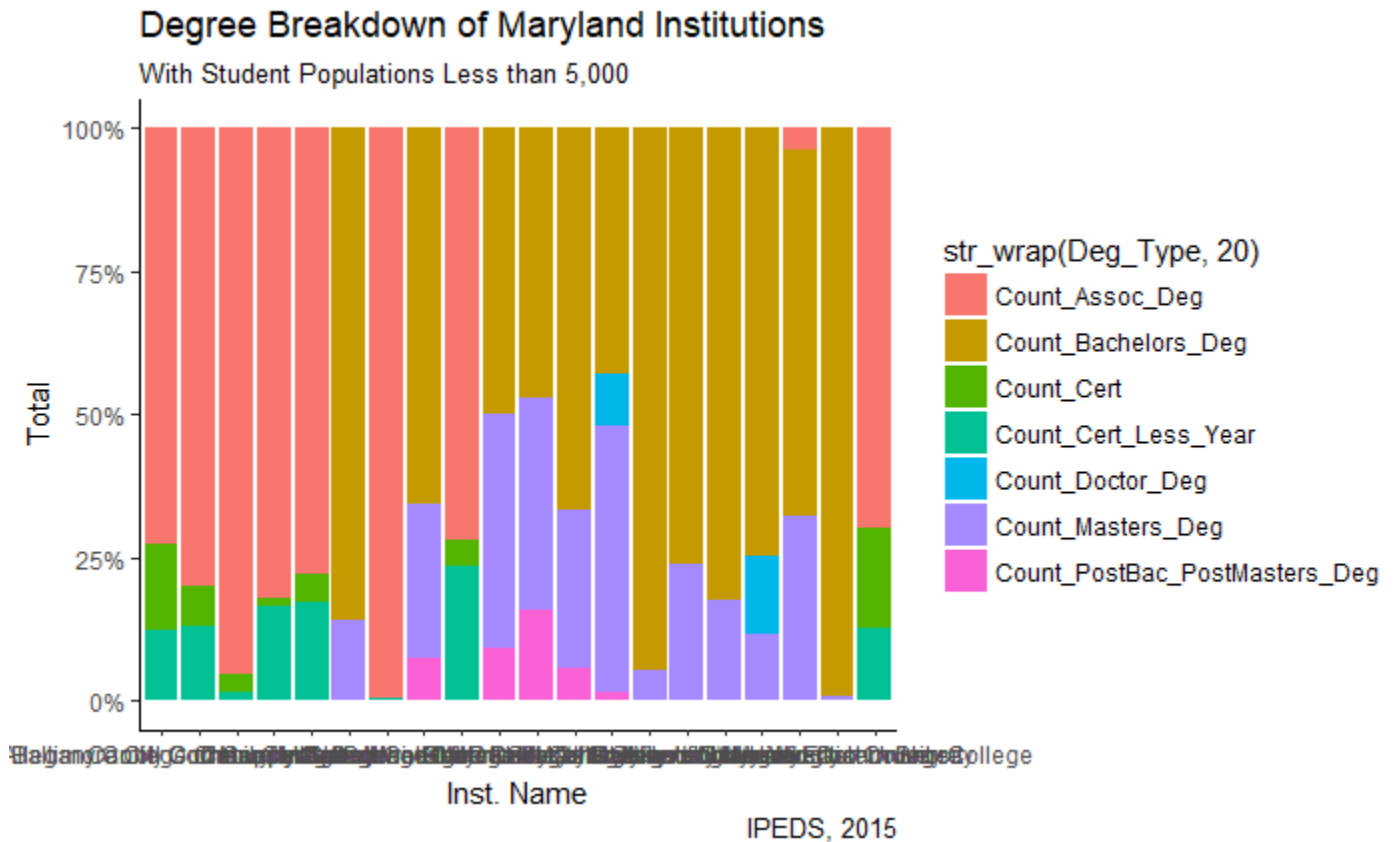
IPEDS_small_pops_long <- gather(IPEDS_small_pops, Deg_Type, Total, Count_Doctor_Deg:Count
_Cert_Less_Year, factor_key=TRUE)

# Step 2-Removing columns we don't need:

IPEDS_small_pops_long <- subset(IPEDS_small_pops_long, select=-c(Grand_Total_All_Students
, Grand_Total_Male, Grand_Total_Female, Grand_Total_Amer_Ind_AK_Native, Grand_Total_Asian
, Grand_Total_Black_AA, Grand_Total_Hispanic, Grand_Total_Native_Hawaiian_PI, Grand_Total
_White, Grand_Total_Two_More, Grand_Total_Race_Unknown, Grand_Total_Nonresident_Alien))
```

Stacked 100% Chart

```
vis_stacked <- ggplot(IPEDS_small_pops_long, aes(x=Inst_Name, y=Total, fill=str_wrap(Deg_Type, 20))) +
  geom_bar(stat="identity", position = "fill") +
  scale_y_continuous(labels = percent_format()) +
  theme_bw() +
  labs(title = "Degree Breakdown of Maryland Institutions",
  subtitle = "With Student Populations Less than 5,000",
  caption = "IPEDS, 2015",
  x = "Inst. Name",
  y = "Total") +
  theme(panel.border = element_blank(),
  panel.grid.major = element_blank(),
  panel.grid.minor = element_blank(),
  axis.line = element_line(colour = "black"))
vis_stacked
```



Tweaking the Stacked 100% Chart-Part 1-Legend

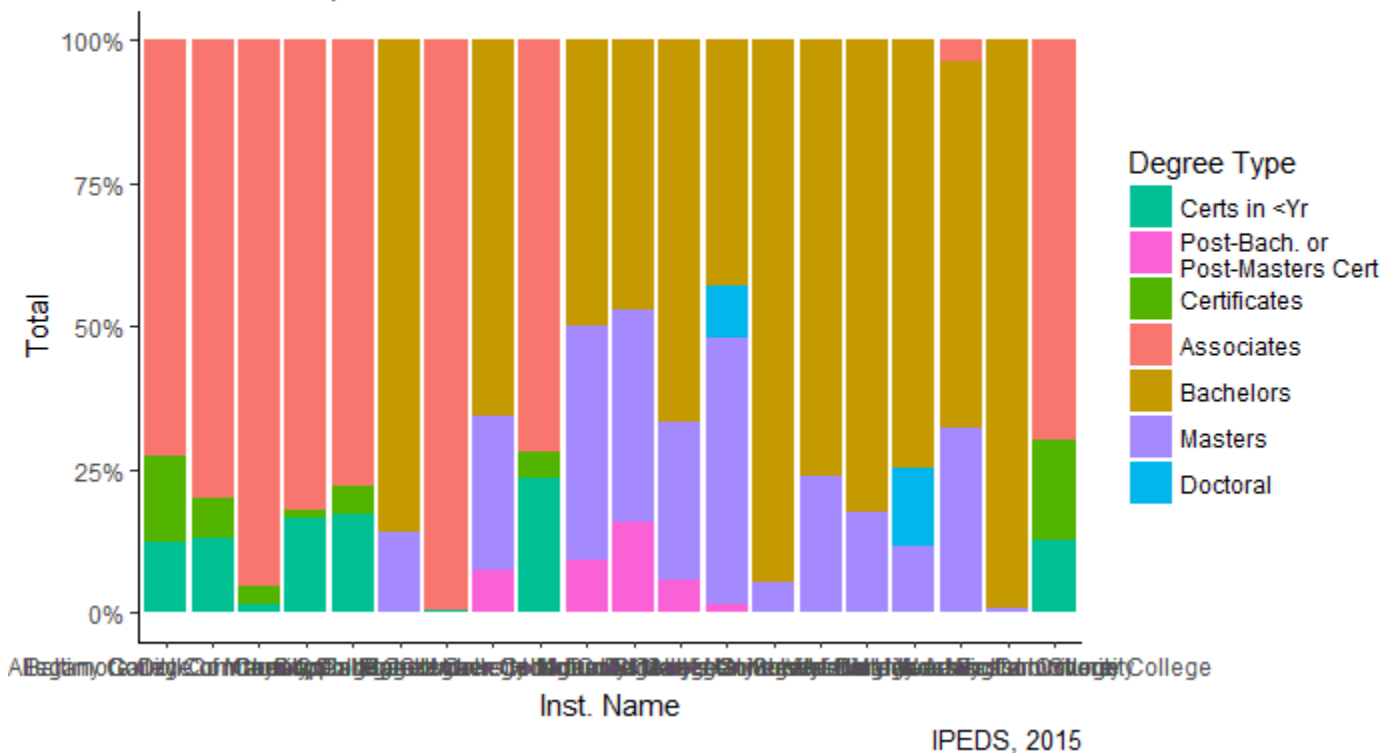
```
# getting unique types of degrees:
list <- unique(c(as.character(IPEDS_small_pops_long$Deg_Type)))
list

## [1] "Count_Doctor_Deg"          "Count_Masters_Deg"
## [3] "Count_Bachelors_Deg"      "Count_Assoc_Deg"
## [5] "Count_PostBac_PostMasters_Deg" "Count_Cert"
## [7] "Count_Cert_Less_Year"

# Note: "\n" adds lines breaks
vis_edited_legend <- vis_stacked + scale_fill_discrete(name="Degree Type",
  breaks=c("Count_Cert_Less_Year", "Count_PostBac_PostMasters_Deg", "Count_Cert",
  "Count_Assoc_Deg", "Count_Bachelors_Deg", "Count_Masters_Deg", "Count_Doctor_Deg"),
  labels=c("Certs in <Yr", "Post-Bach. or\nPost-Masters Cert", "Certificates",
  "Associates", "Bachelors", "Masters", "Doctoral"))
vis_edited_legend
```

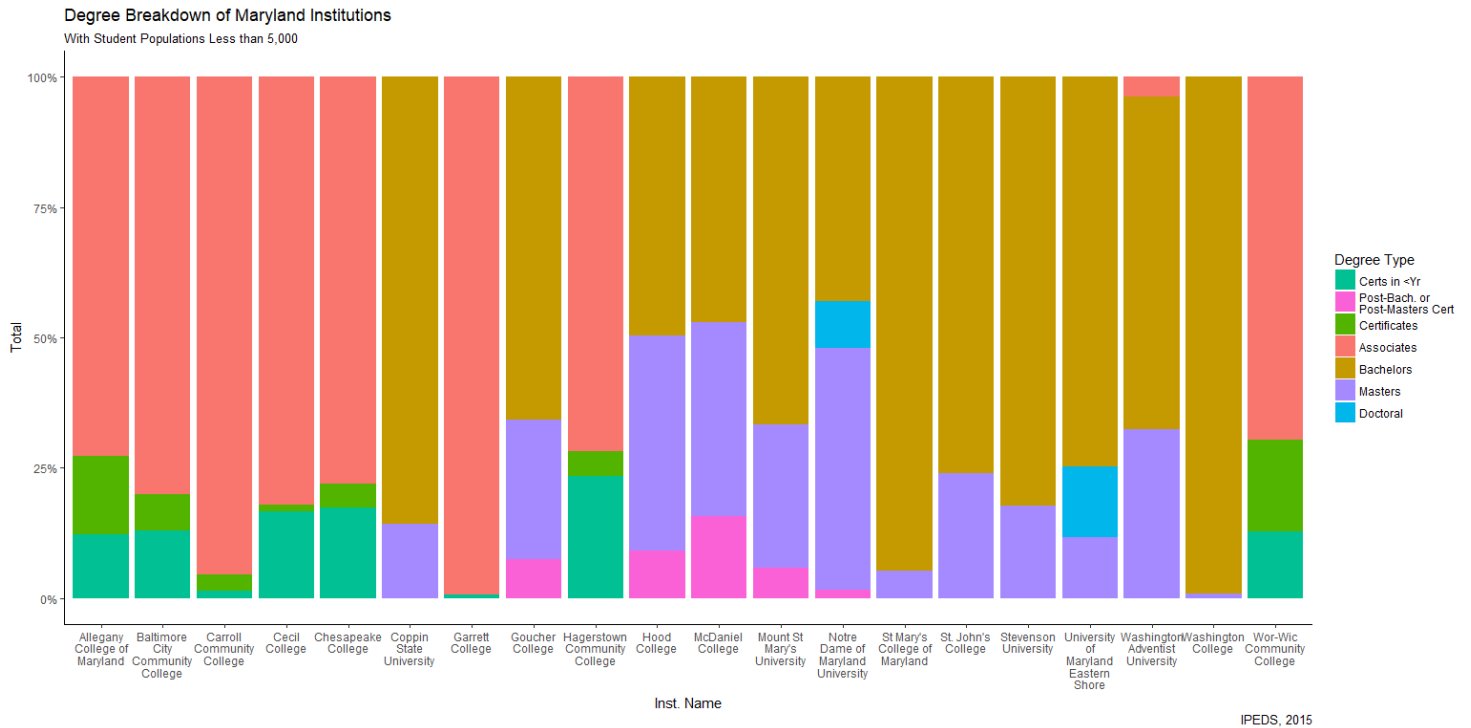
Degree Breakdown of Maryland Institutions

With Student Populations Less than 5,000



Tweaking the Stacked 100% Chart-Part 2-The X Axis

```
# "fig.width=16,fig.height=8" From above allows for the plot to be widened
# "+ scale_x_discrete(labels = function(x) str_wrap(x, width = 10))"
# allows for institution names to be moved to multiple lines
vis_edited_x_axis <- vis_edited_legend + scale_x_discrete(labels = function(x) str_wrap(x
, width = 10))
vis_edited_x_axis
```

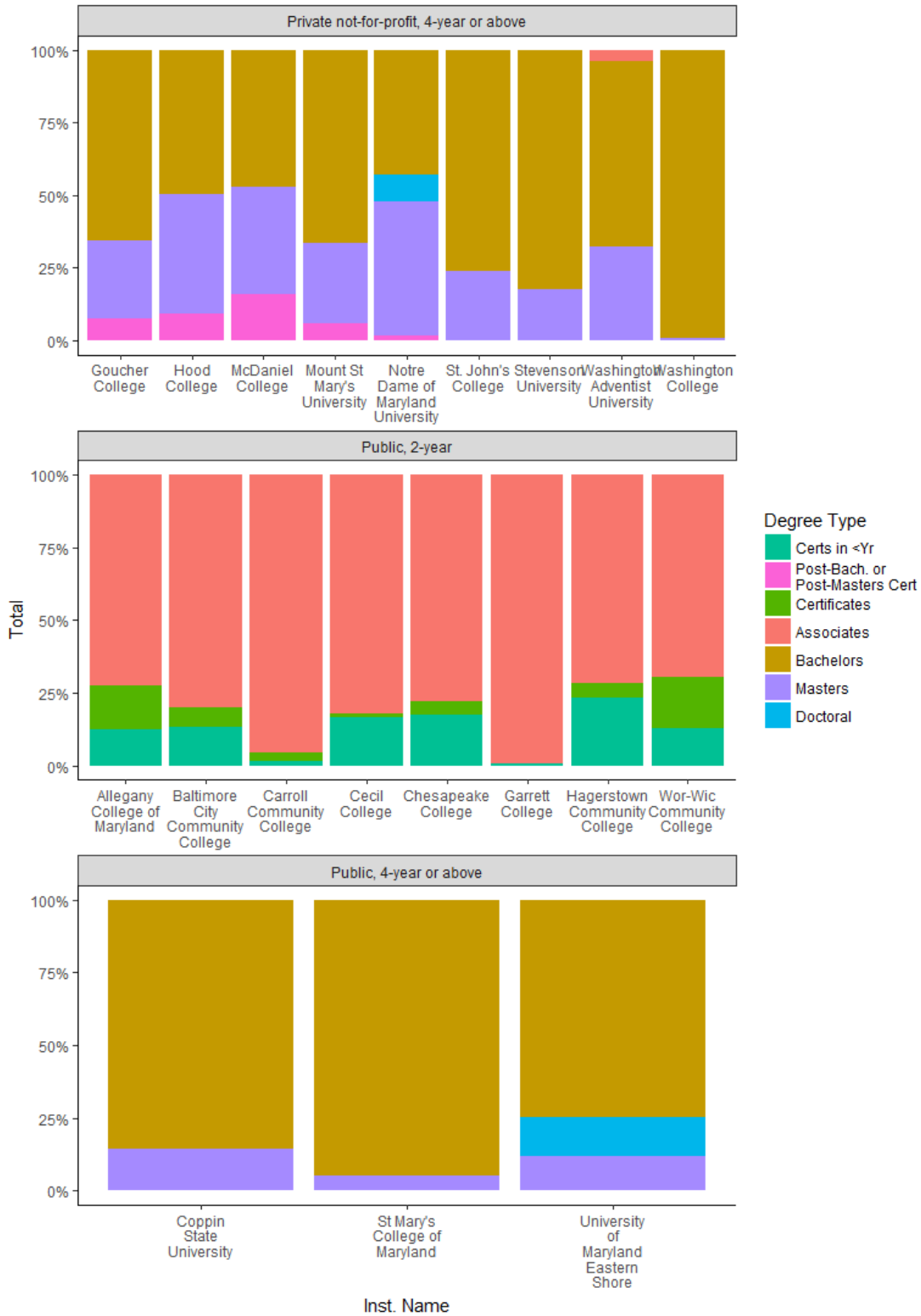


Tweaking the Stacked 100% Chart-Part 3-Let's Group this!

```
vis_edited_facet <- vis_edited_x_axis + facet_wrap(~Inst_Sector, scales="free", ncol=1)
vis_edited_facet
```

Degree Breakdown of Maryland Institutions

With Student Populations Less than 5,000



IPEDS, 2015

Review of Components

```
vis_stacked <-ggplot(IPEDS_small_pops_long, aes(x=Inst_Name, y=Total, fill=str_wrap(Deg_Type, 20))) #Sets what variables are on each axis

# str_wrap(Inst_Sector,20) - Allows the legend to be wrapped

+ geom_bar(stat="identity", position = "fill") #heights of the bars to represent values in the data

+ scale_y_continuous(labels = percent_format()) # graph is a stacked 100% chart

+ theme_bw() + labs(title = "Degree Breakdown of Maryland Institutions", subtitle = "With Student Populations Less than 5,000", caption = "IPEDS, 2015", x = "Inst. Name", y = "Total") # Title, Subtitle, Source Title and Axis Titles added

+ theme(panel.border = element_blank(), panel.grid.major = element_blank(), panel.grid.minor = element_blank(), axis.line = element_line(colour = "black"))

# panel.border = element_blank() - removes border around entire output

# panel.grid.major = element_blank() - removes major gridlines

# panel.grid.minor = element_blank() - removes minor gridlines

# axis.line = element_line(colour = "black") - adds black axis line

+ scale_fill_discrete(name="Degree Type",

                      breaks=c("Count_Cert_Less_Year", "Count_PostBac_PostMasters_Deg", "Count_Cert", "Count_Assoc_Deg", "Count_Bachelors_Deg", "Count_Masters_Deg", "Count_Doctor_Deg"),

                      labels=c("Certs in <Yr", "Post-Bach. or\nPost-Masters Cert", "Certificates", "Associates", "Bachelors", "Masters", "Doctoral"))

# Custom labels for the Legend based on Degree Types

+ scale_x_discrete(labels = function(x) str_wrap(x, width = 10))

# Allows for the Institution names to be multiple lines

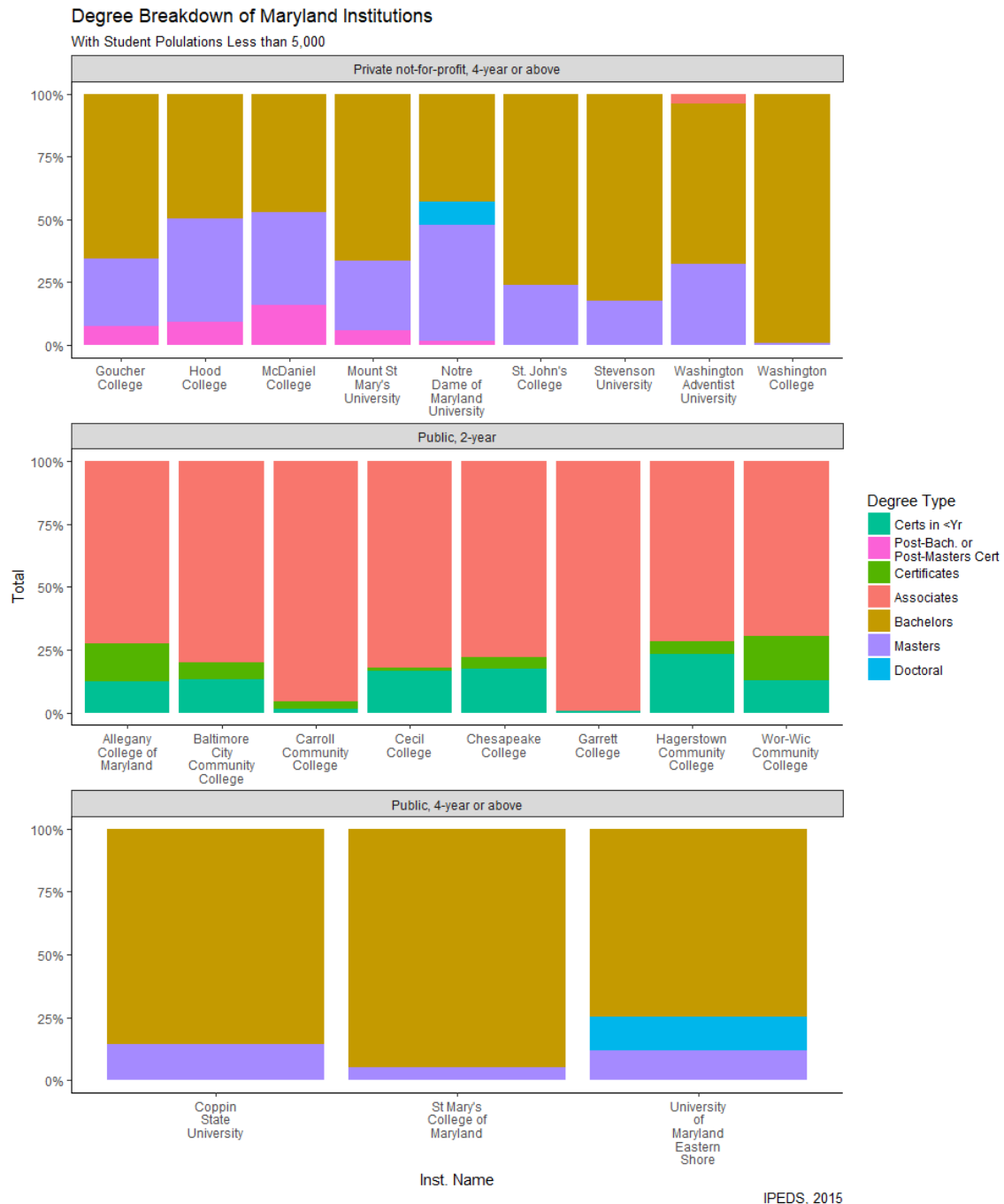
+ facet_wrap(~Inst_Sector, scales="free", ncol=1)

# Groups the plot by sector in one column
```

Stacked 100% Chart-Final

```
vis_stacked_final <-ggplot(IPEDS_small_pops_long, aes(x=Inst_Name, y=Total, fill=str_wrap(Deg_Type,
20))) + geom_bar(stat="identity", position = "fill") + scale_y_continuous(labels = percent_format(
)) + theme_bw() + labs(title = "Degree Breakdown of Maryland Institutions", subtitle = "With Studen
t Populations Less than 5,000", caption = "IPEDS, 2015", x = "Inst. Name", y = "Total") + theme(panel.border = element_blank(), panel.grid.major = element_blank(), panel.grid.minor = element_blank()
, axis.line = element_line(colour = "black")) + scale_fill_discrete(name="Degree Type",breaks=c("Co
unt_Cert_Less_Yr", "Count_PostBac_PostMasters_Deg", "Count_Cert", "Count_Assoc_Deg", "Count_Bache
lors_Deg", "Count_Masters_Deg", "Count_Doctor_Deg"), labels=c("Certs in <Yr", "Post-Bach. or\nPost-
Masters Cert", "Certificates", "Associates", "Bachelors", "Masters", "Doctoral")) + scale_x_discret
e(labels = function(x) str_wrap(x, width = 10)) + facet_wrap(~Inst_Sector, scales="free", ncol=1)

vis_stacked_final
```



Helpful Websites

[Datacamp Courses](#)

[R Cheat Sheets](#)

[Cookbook for R](#)

[GGPlot Book by Hadley Wickham Building Plots Layer by Layer](#)

[GGPlot Themes](#)

[Customizing Themes with R Package, “ggthemr”](#)

[Gallery of R Markdown Files](#)

For Extra Challenges

[Histograms Using R's 'airquality' Dataset](#)

[Scatter Plots-Corruption and Human Development](#)

[Stacked Bar Plots-Exports to China](#)