R Notebook

```
#This is an exploratory data analysis I created for a class. There were 4 things I did in this project:
#1. I cleaned up data.
#2. I transformed data to make my analysis possible.
#3. I ran data analysis on median income, gender breakdown, & unemployment rate.
#4. I graphed the results of my analysis.
#5. The purpose for doing this was to see which major group has the highest median wage, gender breakdown, and unemployment rate.

# The first thing I did here was load in extra packages to make the base version of R more capable.
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(ggplot2)
library(magrittr)
library(tidyr)
```

```
##
## Attaching package: 'tidyr'
```

```
## The following object is masked from 'package:magrittr':
##
##     extract
```

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────── tidyverse 1.3.0 ──
```

```
## ✓ tibble  3.0.1     ✓ stringr 1.4.0
## ✓ readr   1.3.1     ✓ forcats 0.5.0
## ✓ purrr   0.3.4
## ── Conflicts ────────────────────────── tidyverse_conflicts() ──
## x tidyr::extract()   masks magrittr::extract()
## x dplyr::filter()    masks stats::filter()
## x dplyr::lag()       masks stats::lag()
## x purrr::set_names() masks magrittr::set_names()
```

```
#First I had to load in the data file I'd be using. This data set was obtained from the
 website 538 and I got the csv file here: https://github.com/fivethirtyeight/data/tree/m
aster/college-majors. The cool part of R is that it can load in any data set from Excel
 with ease.

data <- read.csv("recent-grads.csv")

#Next, I had to filter my data set using the dplyr package in R. I used the filter() fun
ction to split up my Major categories from the larger data set.

engineering <- filter(data,Major_category == "Engineering")

business <- filter(data,Major_category == "Business")

Physical_sciences <- filter(data,Major_category == "Physical Sciences")

Law_and_Public_Policy   <- filter(data,Major_category == "Law & Public Policy")

Computers_and_Mathematics   <- filter(data,Major_category == "Computers & Mathematics")

Agriculture_and_Natural_Resources<- filter(data,Major_category == "Agriculture & Natural
Resources")

Industrial_Arts_and_Consumer_Services   <- filter(data,Major_category == "Industrial Art
s & Consumer Services")

Arts <- filter(data,Major_category == "Arts")

Health <- filter(data,Major_category == "Health")

Social_Science <- filter(data,Major_category == "Social Science")

Biology_and_Life_Science<- filter(data,Major_category == "Biology & Life Science")

Education<- filter(data,Major_category == "Education")

Humanities_and_Liberal_Arts <- filter(data,Major_category == "Humanities & Liberal Arts"
)

Psychology_and_Social_Work<- filter(data,Major_category == "Psychology & Social Work")

Communications_and_Journalism<- filter(data,Major_category == "Communications & Journali
sm")

Interdisciplinary<- filter(data,Major_category == "Interdisciplinary")
```

```r
#Next, I created a tribble with the information that I'd need to perform my calculations by combining all relevant data points into one massive tribble. This tribble includes the major group and it's median income, percentage of population that is male, and it's unemployment rate.

databygroup<-tribble(
  ~Major_Group, ~Median_Income, ~Percent_Male, ~Unemployment_Rate,
  "Engineering",mean(engineering$Median),sum(engineering$Men)/sum(engineering$Total),mean(engineering$Unemployment_rate),

  "Business",mean(business$Median),sum(business$Men)/sum(business$Total),mean(business$Unemployment_rate),
"Physical Sciences",mean(Physical_sciences$Median),sum(Physical_sciences$Men)/sum(Physical_sciences$Total),mean(Physical_sciences$Unemployment_rate),
"Law & Public Policy",mean(Law_and_Public_Policy$Median),sum(Law_and_Public_Policy$Men)/sum(Law_and_Public_Policy$Total),mean(Law_and_Public_Policy$Unemployment_rate),
"Computers & Mathematics",mean(Computers_and_Mathematics$Median),sum(Computers_and_Mathematics$Men)/sum(Computers_and_Mathematics$Total),mean(Computers_and_Mathematics$Unemployment_rate),
"Agriculture & Natural Resources",mean(Agriculture_and_Natural_Resources$Median),sum(Agriculture_and_Natural_Resources$Men,na.rm = TRUE)/sum(Agriculture_and_Natural_Resources$Total,na.rm = TRUE),mean(Agriculture_and_Natural_Resources$Unemployment_rate),
"Industrial Arts & Consumer Services",mean(Industrial_Arts_and_Consumer_Services$Median),sum(Industrial_Arts_and_Consumer_Services$Men)/sum(Industrial_Arts_and_Consumer_Services$Total),mean(Industrial_Arts_and_Consumer_Services$Unemployment_rate),
"Arts",mean(Arts$Median),sum(Arts$Men)/sum(Arts$Total),mean(Arts$Unemployment_rate),
"Health",mean(Health$Median),sum(Health$Men)/sum(Health$Total),mean(Health$Unemployment_rate),
"Social Science",mean(Social_Science$Median),sum(Social_Science$Men)/sum(Social_Science$Total),mean(Social_Science$Unemployment_rate),
"Biology & Life Science",mean(Biology_and_Life_Science$Median),sum(Biology_and_Life_Science$Men)/sum(Biology_and_Life_Science$Total),mean(Biology_and_Life_Science$Unemployment_rate),
"Education",mean(Education$Median),sum(Education$Men)/sum(Education$Total),mean(Education$Unemployment_rate),
"Humanities & Liberal Arts",mean(Humanities_and_Liberal_Arts$Median),sum(Humanities_and_Liberal_Arts$Men)/sum(Humanities_and_Liberal_Arts$Total),mean(Humanities_and_Liberal_Arts$Unemployment_rate),
"Psychology and Social Work",mean(Psychology_and_Social_Work$Median),sum(Psychology_and_Social_Work$Men)/sum(Psychology_and_Social_Work$Total),mean(Psychology_and_Social_Work$Unemployment_rate),
"Communications & Journalism",mean(Communications_and_Journalism$Median),sum(Communications_and_Journalism$Men)/sum(Communications_and_Journalism$Total),mean(Communications_and_Journalism$Unemployment_rate),
"Interdisciplinary",mean(Interdisciplinary$Median),sum(Interdisciplinary$Men)/sum(Interdisciplinary$Total),mean(Interdisciplinary$Unemployment_rate)
)

databygroup
```

Major_Group	Median_Income	Percent_Male	Unemployment_Rate
<chr>	<dbl>	<dbl>	<dbl>

Major_Group <chr>	Median_Income <dbl>	Percent_Male <dbl>	Unemployment_Rate <dbl>
Engineering	57382.76	0.7595236	0.06333388
Business	43538.46	0.5127951	0.07106354
Physical Sciences	41890.00	0.5142900	0.04651108
Law & Public Policy	42200.00	0.5087964	0.09080476
Computers & Mathematics	42745.45	0.6980582	0.08425599
Agriculture & Natural Resources	36900.00	0.5336816	0.05632831
Industrial Arts & Consumer Services	36342.86	0.4516302	0.04807134
Arts	33062.50	0.3763055	0.09017270
Health	36825.00	0.1630227	0.06592017
Social Science	37344.44	0.4846235	0.09572883

1-10 of 16 rows Previous **1** 2 Next

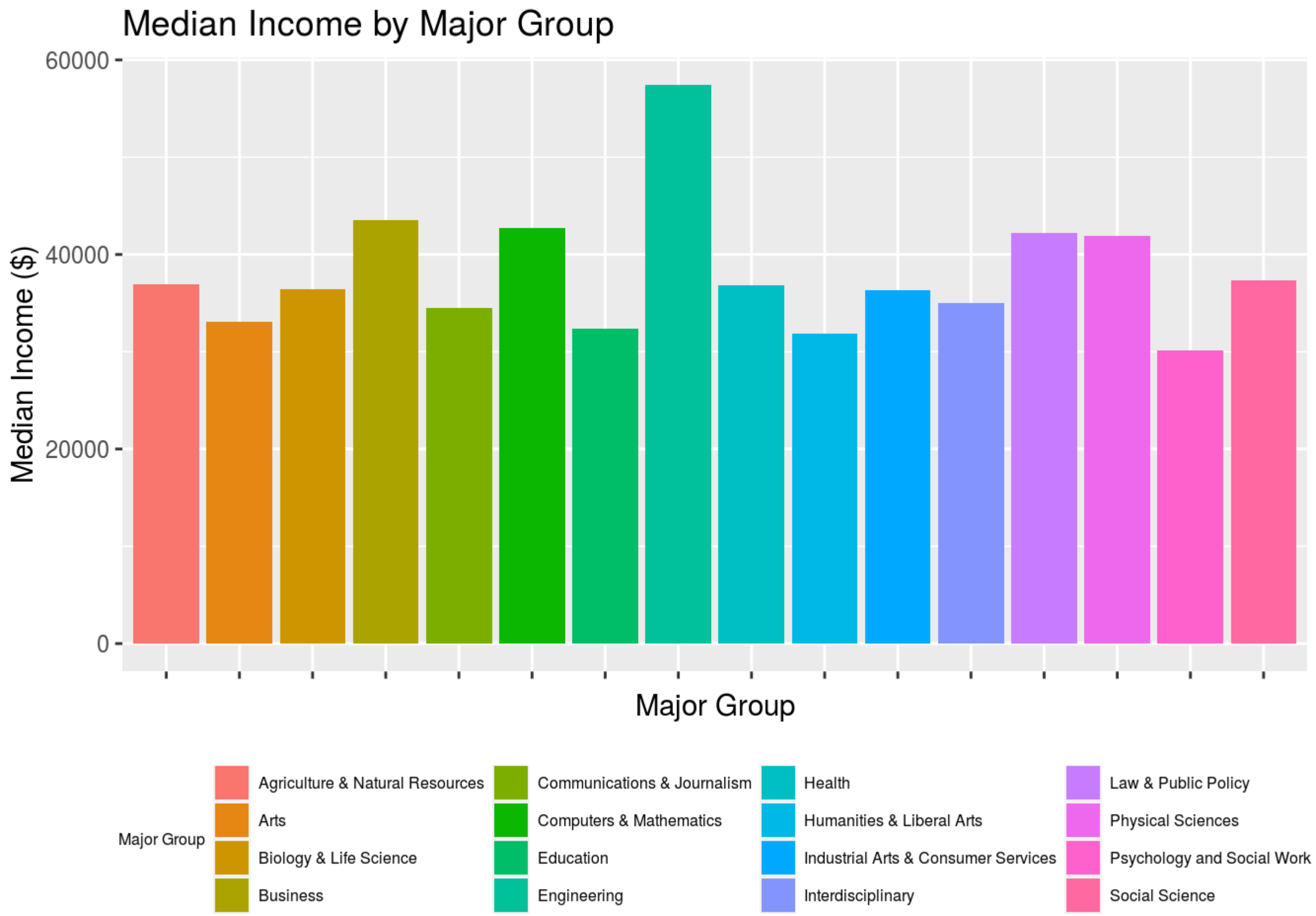
```
#Next, I was about to move on to plotting my charts, but then I realized it might be bet
ter to use the mutate() function to transform the percentages from Percent Male and Unem
ployment to full digit percentages rounded to the nearest hundreth. This would then make
them nicer on a plot.

newdatabygroup<-databygroup %>% mutate(
  Percent_Male = round(Percent_Male*100, digits = 2),
  Unemployment_Rate = round(Unemployment_Rate*100,digits = 2)
)

newdatabygroup
```

Major_Group <chr>	Median_Income <dbl>	Percent_Male <dbl>	Unemployment_Rate <dbl>
Engineering	57382.76	75.95	6.33
Business	43538.46	51.28	7.11
Physical Sciences	41890.00	51.43	4.65
Law & Public Policy	42200.00	50.88	9.08
Computers & Mathematics	42745.45	69.81	8.43
Agriculture & Natural Resources	36900.00	53.37	5.63
Industrial Arts & Consumer Services	36342.86	45.16	4.81
Arts	33062.50	37.63	9.02
Health	36825.00	16.30	6.59

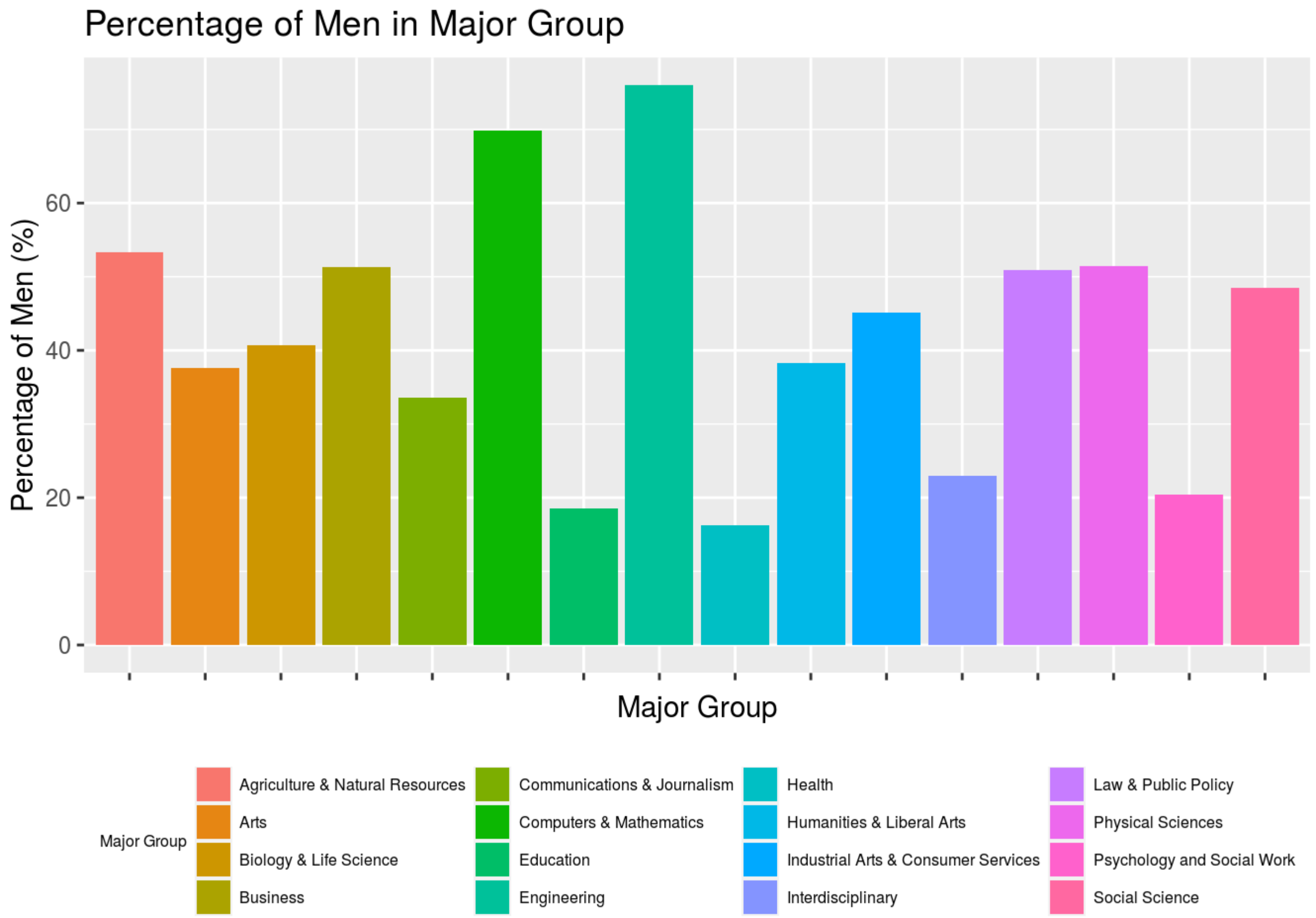
Major_Group	Median_Income	Percent_Male	Unemployment_Rate
<chr>	<dbl>	<dbl>	<dbl>
Social Science	37344.44	48.46	9.57

1-10 of 16 rows Previous **1** 2 Next

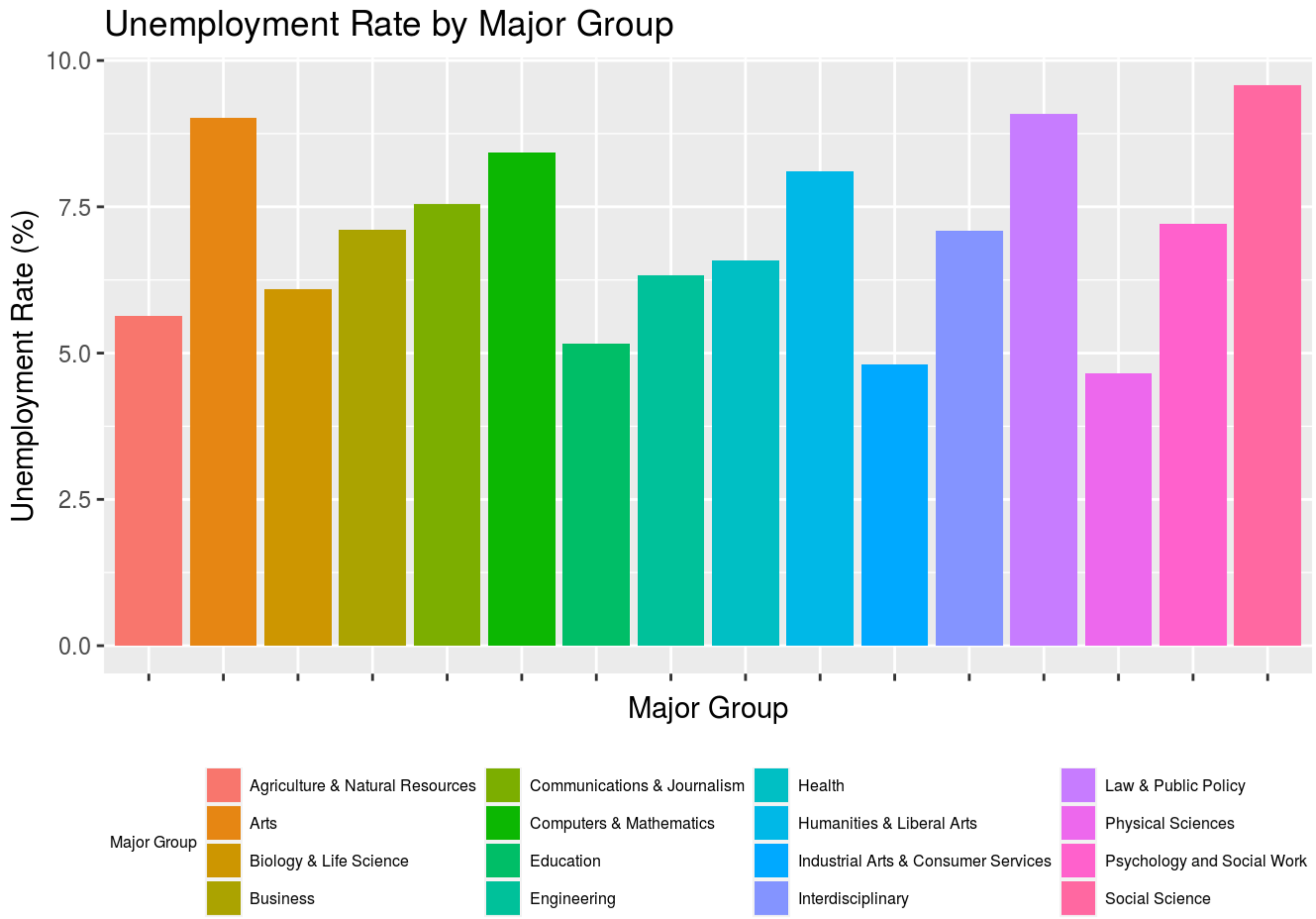
```
#With my new, nice round percentage numbers in hand, I was ready to plot. First, I creat
ed new variables for ease of inputting data into the ggplot function. Then, I made 3 bar
plots, one for median income, one for percentage of men, and one for unemployment rate

Major_group<-newdatabygroup$Major_Group
Median_Income<-newdatabygroup$Median_Income
Percent_Male<-newdatabygroup$Unemployment_Rate
Unemployment_Rate<-newdatabygroup$Unemployment_Rate


gi <- ggplot(newdatabygroup, aes(Major_group,Median_Income))

gi + geom_col(aes(fill=Major_group))+ theme(legend.position="bottom",legend.title = elem
ent_text(size = 6),
          legend.text = element_text(size = 6), legend.key.size = unit(0.5, "cm"),legen
d.key.width = unit(0.5,"cm"),axis.text.x= element_text(size=0))+
  labs(
    title= "Median Income by Major Group",
    x = "Major Group",
    y = "Median Income ($)",fill = "Major Group"
    )
```



```
gm <- ggplot(newdatabygroup, aes(Major_group,Percent_Male))

gm + geom_col(aes(fill=Major_group))+ theme(legend.position="bottom",legend.title = element_text(size = 6),
          legend.text = element_text(size = 6), legend.key.size = unit(0.5, "cm"),legend.key.width = unit(0.5,"cm"),axis.text.x= element_text(size=0))+
  labs(
    title= "Percentage of Men in Major Group",
    x = "Major Group",
    y = "Percentage of Men (%)",fill = "Major Group"
    )
```



```
gu <- ggplot(newdatabygroup, aes(Major_group,Unemployment_Rate))

gu + geom_col(aes(fill=Major_group))+ theme(legend.position="bottom",legend.title = elem
ent_text(size = 6),
          legend.text = element_text(size = 6), legend.key.size = unit(0.5, "cm"),legen
d.key.width = unit(0.5,"cm"),axis.text.x= element_text(size=0))+
  labs(
    title= "Unemployment Rate by Major Group",
    x = "Major Group",
    y = "Unemployment Rate (%)",fill = "Major Group"
    )
```



```
#After I had my plots I used the arrange() function in dplyr to arrange the data by my v
ariables in order to more easily see which major group had the highest and lowest value
s.

newdatabygroup %>% arrange(newdatabygroup$Median_Income)
```

Major_Group	Median_Income	Percent_Male	Unemployment_Rate
<chr>	<dbl>	<dbl>	<dbl>
Psychology and Social Work	30100.00	20.40	7.21
Humanities & Liberal Arts	31913.33	38.24	8.10
Education	32350.00	18.52	5.17
Arts	33062.50	37.63	9.02
Communications & Journalism	34500.00	33.60	7.55
Interdisciplinary	35000.00	22.91	7.09
Industrial Arts & Consumer Services	36342.86	45.16	4.81
Biology & Life Science	36421.43	40.74	6.09
Health	36825.00	16.30	6.59
Agriculture & Natural Resources	36900.00	53.37	5.63

1-10 of 16 rows Previous **1** 2 Next

```
newdatabygroup %>% arrange(newdatabygroup$Percent_Male)
```

Major_Group	**Median_Income**	**Percent_Male**	**Unemployment_Rate**
<chr>	<dbl>	<dbl>	<dbl>
Health	36825.00	16.30	6.59
Education	32350.00	18.52	5.17
Psychology and Social Work	30100.00	20.40	7.21
Interdisciplinary	35000.00	22.91	7.09
Communications & Journalism	34500.00	33.60	7.55
Arts	33062.50	37.63	9.02
Humanities & Liberal Arts	31913.33	38.24	8.10
Biology & Life Science	36421.43	40.74	6.09
Industrial Arts & Consumer Services	36342.86	45.16	4.81
Social Science	37344.44	48.46	9.57

1-10 of 16 rows Previous **1** 2 Next

```
newdatabygroup %>% arrange(newdatabygroup$Unemployment_Rate)
```

Major_Group	**Median_Income**	**Percent_Male**	**Unemployment_Rate**
<chr>	<dbl>	<dbl>	<dbl>
Communications & Journalism	34500.00	33.60	7.55
Humanities & Liberal Arts	31913.33	38.24	8.10
Computers & Mathematics	42745.45	69.81	8.43
Arts	33062.50	37.63	9.02
Law & Public Policy	42200.00	50.88	9.08
Social Science	37344.44	48.46	9.57

11-16 of 16 rows Previous 1 **2** Next

```
#What does this all mean? It turns out that the engineering group of majors has the highest median income ($57382.76), as well as the highest percentage of males (75.95%). The honor of having the lowest unemployment rate goes to the physical sciences major group, with an unemployment rate of 4.65%. The engineering major group comes in 6th place with an unemployment rate of 6.33%.
```