

Data 608 Final Project

Bryan Persaud

12/13/2020

```
library(tidyverse)

## -- Attaching packages ----- tidyverse 1.3.0 --

## v ggplot2 3.3.2      v purrr 0.3.4
## v tibble 3.0.4       v dplyr 1.0.2
## v tidyr 1.1.2        v stringr 1.4.0
## v readr 1.4.0        v forcats 0.5.0

## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
```

Introduction

For my final project I will be using data on college majors. The data is taken from FiveThirtyEight and can be found on their Github page: <https://github.com/fivethirtyeight/data/tree/master/college-majors>. The data is split into different .csv files. The one I will be working with is the all-ages.csv dataset.

What I want to show is the different majors by their categories and the median salary made and/or to show employment vs unemployment by major. I feel this data is important because it helps students decide on what major to pick and to see trends on what majors will get them a job. The data is a little old and it can be used to compare to recent data to see trends in how the job market has changed over the years.

Data Analysis

```
data <- read.csv("https://raw.githubusercontent.com/bpersaud104/Data608/master/all-ages.csv")
head(data, 50)
```

```
##   Major_code      Major
## 1      1100  GENERAL AGRICULTURE
## 2      1101 AGRICULTURE PRODUCTION AND MANAGEMENT
## 3      1102  AGRICULTURAL ECONOMICS
## 4      1103  ANIMAL SCIENCES
## 5      1104  FOOD SCIENCE
## 6      1105 PLANT SCIENCE AND AGRONOMY
```

## 7	1106	SOIL SCIENCE
## 8	1199	MISCELLANEOUS AGRICULTURE
## 9	1301	ENVIRONMENTAL SCIENCE
## 10	1302	FORESTRY
## 11	1303	NATURAL RESOURCES MANAGEMENT
## 12	1401	ARCHITECTURE
## 13	1501	AREA ETHNIC AND CIVILIZATION STUDIES
## 14	1901	COMMUNICATIONS
## 15	1902	JOURNALISM
## 16	1903	MASS MEDIA
## 17	1904	ADVERTISING AND PUBLIC RELATIONS
## 18	2001	COMMUNICATION TECHNOLOGIES
## 19	2100	COMPUTER AND INFORMATION SYSTEMS
## 20	2101	COMPUTER PROGRAMMING AND DATA PROCESSING
## 21	2102	COMPUTER SCIENCE
## 22	2105	INFORMATION SCIENCES
## 23	2106	COMPUTER ADMINISTRATION MANAGEMENT AND SECURITY
## 24	2107	COMPUTER NETWORKING AND TELECOMMUNICATIONS
## 25	2201	COSMETOLOGY SERVICES AND CULINARY ARTS
## 26	2300	GENERAL EDUCATION
## 27	2301	EDUCATIONAL ADMINISTRATION AND SUPERVISION
## 28	2303	SCHOOL STUDENT COUNSELING
## 29	2304	ELEMENTARY EDUCATION
## 30	2305	MATHEMATICS TEACHER EDUCATION
## 31	2306	PHYSICAL AND HEALTH EDUCATION TEACHING
## 32	2307	EARLY CHILDHOOD EDUCATION
## 33	2308	SCIENCE AND COMPUTER TEACHER EDUCATION
## 34	2309	SECONDARY TEACHER EDUCATION
## 35	2310	SPECIAL NEEDS EDUCATION
## 36	2311	SOCIAL SCIENCE OR HISTORY TEACHER EDUCATION
## 37	2312	TEACHER EDUCATION: MULTIPLE LEVELS
## 38	2313	LANGUAGE AND DRAMA EDUCATION
## 39	2314	ART AND MUSIC EDUCATION
## 40	2399	MISCELLANEOUS EDUCATION
## 41	2400	GENERAL ENGINEERING
## 42	2401	AEROSPACE ENGINEERING
## 43	2402	BIOLOGICAL ENGINEERING
## 44	2403	ARCHITECTURAL ENGINEERING
## 45	2404	BIOMEDICAL ENGINEERING
## 46	2405	CHEMICAL ENGINEERING
## 47	2406	CIVIL ENGINEERING
## 48	2407	COMPUTER ENGINEERING
## 49	2408	ELECTRICAL ENGINEERING
## 50	2409	ENGINEERING MECHANICS PHYSICS AND SCIENCE
##		Major_category Total Employed
## 1	Agriculture & Natural Resources	128148 90245
## 2	Agriculture & Natural Resources	95326 76865
## 3	Agriculture & Natural Resources	33955 26321
## 4	Agriculture & Natural Resources	103549 81177
## 5	Agriculture & Natural Resources	24280 17281
## 6	Agriculture & Natural Resources	79409 63043
## 7	Agriculture & Natural Resources	6586 4926
## 8	Agriculture & Natural Resources	8549 6392
## 9	Biology & Life Science	106106 87602

## 10	Agriculture & Natural Resources	69447	48228		
## 11	Agriculture & Natural Resources	83188	65937		
## 12	Engineering	294692	216770		
## 13	Humanities & Liberal Arts	103740	75798		
## 14	Communications & Journalism	987676	790696		
## 15	Communications & Journalism	418104	314438		
## 16	Communications & Journalism	211213	170474		
## 17	Communications & Journalism	186829	147433		
## 18	Computers & Mathematics	62141	49609		
## 19	Computers & Mathematics	253782	218248		
## 20	Computers & Mathematics	29317	22828		
## 21	Computers & Mathematics	783292	656372		
## 22	Computers & Mathematics	77805	66393		
## 23	Computers & Mathematics	39362	32366		
## 24	Computers & Mathematics	51771	44071		
## 25	Industrial Arts & Consumer Services	42325	33388		
## 26	Education	1438867	843693		
## 27	Education	4037	3113		
## 28	Education	2396	1492		
## 29	Education	1446701	819393		
## 30	Education	68808	47203		
## 31	Education	281661	193542		
## 32	Education	157079	113460		
## 33	Education	56477	36224		
## 34	Education	224262	129486		
## 35	Education	149689	108272		
## 36	Education	127022	78785		
## 37	Education	88067	58885		
## 38	Education	181445	111347		
## 39	Education	231861	155159		
## 40	Education	225553	126054		
## 41	Engineering	503080	359172		
## 42	Engineering	65734	44944		
## 43	Engineering	32748	24270		
## 44	Engineering	19587	13713		
## 45	Engineering	18347	12876		
## 46	Engineering	188046	131697		
## 47	Engineering	358593	262831		
## 48	Engineering	154160	128742		
## 49	Engineering	671647	489965		
## 50	Engineering	20582	14909		
##	Employed_full_time_year_round	Unemployed	Unemployment_rate	Median	P25th
## 1	74078	2423	0.02614711	50000	34000
## 2	64240	2266	0.02863606	54000	36000
## 3	22810	821	0.03024832	63000	40000
## 4	64937	3619	0.04267890	46000	30000
## 5	12722	894	0.04918845	62000	38500
## 6	51077	2070	0.03179089	50000	35000
## 7	4042	264	0.05086705	63000	39400
## 8	5074	261	0.03923042	52000	35000
## 9	65238	4736	0.05128983	52000	38000
## 10	39613	2144	0.04256333	58000	40500
## 11	50595	3789	0.05434128	52000	37100
## 12	163020	20394	0.08599113	63000	40400

## 13	50530	5525	0.06793896	46000	32000
## 14	595739	54390	0.06436031	50000	35000
## 15	235407	20754	0.06191675	50000	35000
## 16	125489	15431	0.08300476	48000	32000
## 17	111552	10624	0.06721626	50000	34000
## 18	37261	4609	0.08500867	50000	34500
## 19	189950	11945	0.05189124	65000	45000
## 20	18747	2265	0.09026422	60000	40000
## 21	561052	34196	0.04951866	78000	51000
## 22	57604	3704	0.05284106	68000	46200
## 23	28156	2626	0.07504572	55000	40000
## 24	35954	2748	0.05869412	55000	36000
## 25	25780	1941	0.05494070	40000	26200
## 26	591863	38742	0.04390352	43000	32000
## 27	2468	0	0.00000000	58000	44750
## 28	1093	169	0.10174594	41000	33200
## 29	501786	32685	0.03835916	40000	31000
## 30	29494	1610	0.03298302	43000	34000
## 31	136343	9389	0.04626696	48400	34000
## 32	71133	5890	0.04935065	35300	27000
## 33	24817	1596	0.04219989	46000	35000
## 34	88917	5925	0.04375568	45000	34000
## 35	71615	5357	0.04714466	42000	34000
## 36	51632	3800	0.04601320	45000	33000
## 37	37892	2032	0.03335686	40000	30000
## 38	67651	5624	0.04808029	42000	32000
## 39	94756	6629	0.04097337	42600	32000
## 40	91322	5145	0.03921524	50000	35600
## 41	312023	17986	0.04768824	75000	50000
## 42	38491	1969	0.04197131	80000	58000
## 43	18621	1521	0.05897406	62000	40000
## 44	11180	1017	0.06904277	78000	50000
## 45	9202	1105	0.07903583	65000	40000
## 46	109406	6388	0.04626136	86000	60000
## 47	220528	14823	0.05338659	78000	55000
## 48	111025	7456	0.05474383	80000	60000
## 49	422317	26064	0.05050879	88000	60000
## 50	12257	683	0.04380452	65000	45000
##	P75th				
## 1	80000				
## 2	80000				
## 3	98000				
## 4	72000				
## 5	90000				
## 6	75000				
## 7	88000				
## 8	75000				
## 9	75000				
## 10	80000				
## 11	75000				
## 12	93500				
## 13	71000				
## 14	80000				
## 15	80000				

```
## 16 70000
## 17 75000
## 18 75000
## 19 90000
## 20 85000
## 21 105000
## 22 95000
## 23 80000
## 24 80000
## 25 60000
## 26 59000
## 27 79000
## 28 50000
## 29 50000
## 30 60000
## 31 66500
## 32 45800
## 33 61000
## 34 60000
## 35 53000
## 36 64000
## 37 51000
## 38 54000
## 39 56000
## 40 71000
## 41 100000
## 42 110000
## 43 91000
## 44 102000
## 45 96000
## 46 120000
## 47 105000
## 48 107000
## 49 116000
## 50 100000
```

```
summary(data)
```

```
##      Major_code      Major      Major_category      Total
## Min.   :1100    Length:173    Length:173    Min.   : 2396
## 1st Qu.:2403    Class :character    Class :character    1st Qu.: 24280
## Median :3608    Mode  :character    Mode  :character    Median : 75791
## Mean   :3880                                Mean   : 230257
## 3rd Qu.:5503                                3rd Qu.: 205763
## Max.   :6403                                Max.   :3123510
##      Employed      Employed_full_time_year_round      Unemployed
## Min.   : 1492    Min.   : 1093                                Min.   : 0
## 1st Qu.: 17281    1st Qu.: 12722                                1st Qu.: 1101
## Median : 56564    Median : 39613                                Median : 3619
## Mean   : 166162    Mean   : 126308                                Mean   : 9725
## 3rd Qu.: 142879    3rd Qu.: 111025                                3rd Qu.: 8862
## Max.   :2354398    Max.   :1939384                                Max.   :147261
## Unemployment_rate      Median      P25th      P75th
## Min.   :0.00000    Min.   : 35000    Min.   :24900    Min.   : 45800
```

```
## 1st Qu.:0.04626 1st Qu.: 46000 1st Qu.:32000 1st Qu.: 70000
## Median :0.05472 Median : 53000 Median :36000 Median : 80000
## Mean :0.05736 Mean : 56816 Mean :38697 Mean : 82506
## 3rd Qu.:0.06904 3rd Qu.: 65000 3rd Qu.:42000 3rd Qu.: 95000
## Max. :0.15615 Max. :125000 Max. :78000 Max. :210000
```

The data consists of 11 different variables with 173 observations. Each observation is a different major, so the data shown consists of 173 different majors. The total number of students in each major can be seen as well as information on employment and the median salary made.

```
sapply(data, function(x)
  sum(is.na(x)))
```

```
##           Major_code           Major
##                0                0
## Major_category     Total
##                0                0
## Employed Employed_full_time_year_round
##                0                0
## Unemployed      Unemployment_rate
##                0                0
##           Median           P25th
##                0                0
##           P75th
##                0
```

Here we see that there is no missing data in the dataset.

```
unique(data$Major_category)
```

```
## [1] "Agriculture & Natural Resources" "Biology & Life Science"
## [3] "Engineering" "Humanities & Liberal Arts"
## [5] "Communications & Journalism" "Computers & Mathematics"
## [7] "Industrial Arts & Consumer Services" "Education"
## [9] "Law & Public Policy" "Interdisciplinary"
## [11] "Health" "Social Science"
## [13] "Physical Sciences" "Psychology & Social Work"
## [15] "Arts" "Business"
```

There are 16 different categories that the majors are split into. This is useful to help distinguish a major based on the subject it falls under.

```
data %>%
  select(Major, Total) %>%
  arrange(desc(Total)) %>%
  group_by(Major)
```

```
## # A tibble: 173 x 2
## # Groups:   Major [173]
## Major Total
## <chr> <int>
```

```
## 1 BUSINESS MANAGEMENT AND ADMINISTRATION 3123510
## 2 GENERAL BUSINESS 2148712
## 3 ACCOUNTING 1779219
## 4 NURSING 1769892
## 5 PSYCHOLOGY 1484075
## 6 ELEMENTARY EDUCATION 1446701
## 7 GENERAL EDUCATION 1438867
## 8 MARKETING AND MARKETING RESEARCH 1114624
## 9 ENGLISH LANGUAGE AND LITERATURE 1098647
## 10 COMMUNICATIONS 987676
## # ... with 163 more rows
```

The top 5 majors with the most students are Business Management and Administration, General Business, Accounting, Nursing, and Psychology.

```
data %>%
  select(Major, Employed) %>%
  arrange(desc(Employed)) %>%
  group_by(Major)
```

```
## # A tibble: 173 x 2
## # Groups:   Major [173]
##   Major Employed
##   <chr>      <int>
## 1 BUSINESS MANAGEMENT AND ADMINISTRATION 2354398
## 2 GENERAL BUSINESS 1580978
## 3 ACCOUNTING 1335825
## 4 NURSING 1325711
## 5 PSYCHOLOGY 1055854
## 6 MARKETING AND MARKETING RESEARCH 890125
## 7 GENERAL EDUCATION 843693
## 8 ELEMENTARY EDUCATION 819393
## 9 COMMUNICATIONS 790696
## 10 ENGLISH LANGUAGE AND LITERATURE 708882
## # ... with 163 more rows
```

```
data %>%
  select(Major, Unemployed) %>%
  arrange(desc(Unemployed)) %>%
  group_by(Major)
```

```
## # A tibble: 173 x 2
## # Groups:   Major [173]
##   Major Unemployed
##   <chr>      <int>
## 1 BUSINESS MANAGEMENT AND ADMINISTRATION 147261
## 2 GENERAL BUSINESS 85626
## 3 PSYCHOLOGY 79066
## 4 ACCOUNTING 75379
## 5 COMMUNICATIONS 54390
## 6 ENGLISH LANGUAGE AND LITERATURE 52248
## 7 MARKETING AND MARKETING RESEARCH 51839
```

```
## 8 POLITICAL SCIENCE AND GOVERNMENT      40376
## 9 GENERAL EDUCATION                     38742
## 10 BIOLOGY                             36757
## # ... with 163 more rows
```

The top 5 majors with the most employment are Business Management and Administration, General Business, Accounting, Nursing, and Psychology. However the top 5 majors with the most unemployment are Business Management and Administration, General Business, Psychology, Accounting, and Communications. Both are close slightly different but are close to the top majors with the most students.

```
data %>%
  select(Major, Median) %>%
  arrange(desc(Median)) %>%
  group_by(Major)
```

```
## # A tibble: 173 x 2
## # Groups:   Major [173]
##   Major                               Median
##   <chr>                               <int>
## 1 PETROLEUM ENGINEERING              125000
## 2 PHARMACY PHARMACEUTICAL SCIENCES AND ADMINISTRATION 106000
## 3 NAVAL ARCHITECTURE AND MARINE ENGINEERING           97000
## 4 METALLURGICAL ENGINEERING              96000
## 5 NUCLEAR ENGINEERING                   95000
## 6 MINING AND MINERAL ENGINEERING          92000
## 7 MATHEMATICS AND COMPUTER SCIENCE          92000
## 8 ELECTRICAL ENGINEERING                88000
## 9 CHEMICAL ENGINEERING                  86000
## 10 GEOLOGICAL AND GEOPHYSICAL ENGINEERING           85000
## # ... with 163 more rows
```

The top 5 majors with the most median salary are Petroleum Engineering, Pharmacy Pharmaceutical Sciences and Administration, Naval Architecture and Marine Engineering, Metallurgical Engineering, and Nuclear Engineering.

```
data %>%
  select(Major, Employed, Unemployed, Median) %>%
  arrange(desc(Employed)) %>%
  group_by(Major)
```

```
## # A tibble: 173 x 4
## # Groups:   Major [173]
##   Major                               Employed Unemployed Median
##   <chr>                               <int>     <int>   <int>
## 1 BUSINESS MANAGEMENT AND ADMINISTRATION 2354398    147261   58000
## 2 GENERAL BUSINESS                     1580978     85626   60000
## 3 ACCOUNTING                          1335825     75379   65000
## 4 NURSING                             1325711     36503   62000
## 5 PSYCHOLOGY                         1055854     79066   45000
## 6 MARKETING AND MARKETING RESEARCH      890125     51839   56000
## 7 GENERAL EDUCATION                     843693     38742   43000
## 8 ELEMENTARY EDUCATION                  819393     32685   40000
```



```
## 9 COMMUNICATIONS 790696 54390 50000
## 10 ENGLISH LANGUAGE AND LITERATURE 708882 52248 50000
## # ... with 163 more rows
```

```
data %>%
  select(Major, Employed, Unemployed, Median) %>%
  arrange(desc(Median)) %>%
  group_by(Major)
```

```
## # A tibble: 173 x 4
## # Groups:   Major [173]
##   Major Employed Unemployed Median
##   <chr>      <int>      <int>   <int>
## 1 PETROLEUM ENGINEERING 14002        617 125000
## 2 PHARMACY PHARMACEUTICAL SCIENCES AND ADMINISTRATI~ 124058        4414 106000
## 3 NAVAL ARCHITECTURE AND MARINE ENGINEERING 10690         449 97000
## 4 METALLURGICAL ENGINEERING 6939         326 96000
## 5 NUCLEAR ENGINEERING 7320         527 95000
## 6 MINING AND MINERAL ENGINEERING 7416         366 92000
## 7 MATHEMATICS AND COMPUTER SCIENCE 5874         150 92000
## 8 ELECTRICAL ENGINEERING 489965      26064 88000
## 9 CHEMICAL ENGINEERING 131697       6388 86000
## 10 GEOLOGICAL AND GEOPHYSICAL ENGINEERING 4120          0 85000
## # ... with 163 more rows
```

Here we compare the employment of a major to the median salary. Above we saw that the majors with the most employment were different from the majors with the most median salary. This tells us that there is no relationship between employment and median salary of a major.

Shiny App

A shiny app was created to visualize the dataset. The shiny app can be found here: <https://bpersaud104.shinyapps.io/Data608FinalProject/>

The visualization was made while taking into consideration the dataset. The dataset was taken from FiveThirtyEight and the data is from the American Community Survey 2010-2012 Public Use Microdata Series. The data consists of 173 different majors split into 16 different categories. There are also other information such as the total students in the major, the employment by major, and the median salary.

The shiny app consists of three different graphs, one graph to show the total students in a major, one graph to show the median salary of a major, and one graph to show employment vs unemployment of a major. The first two graphs were designed in a way where you select from the different major categories to help narrow down the different majors by the subject of the major. The third graph you select from a list of all the majors to see the employment vs unemployment. I think this is important because the shiny app was made in a way where a student can explore the data to see important information on the different majors. Whether the student knows what major they want to pursue or is undecided, this can help the student in picking a major that will help them to get a job.

References

<https://github.com/fivethirtyeight/data/tree/master/college-majors>

<https://bpersaud104.shinyapps.io/Data608FinalProject/>