

USF Project

Stephen Hsu

Source file ⇒ [USFSideProj.Rmd](#)

load the necessary libraries

```
library(ggplot2)
library(XML)
```

```
## Warning: package 'XML' was built under R version 3.2.4
```

```
library(RCurl)
```

```
## Loading required package: bitops
```

```
##
## Attaching package: 'RCurl'
```

```
## The following object is masked from 'package:tidyr':
##
## complete
```

```
library(dplyr)
library(tidyr)
```

Data Wrangling

collect the sites

```

AlumniUrl <- "https://www.usfca.edu/arts-sciences/graduate-programs/analytics/our-alumni"
Alumnitext <- getURLContent(AlumniUrl)
docstatAlumni <- htmlParse(Alumnitext)
nameAlumni <- xpathSApply(docstatAlumni, '//div[@class="field field-name-body field-type-text-with-summary field-label-hidden typography"]/h3', xmlValue)
degreeAlumni <- xpathSApply(docstatAlumni, '//div[@class="field field-name-body field-type-text-with-summary field-label-hidden typography"]/p', xmlValue)

StudentUrl <- "https://www.usfca.edu/arts-sciences/graduate-programs/analytics/our-students"
Studenttext <- getURLContent(StudentUrl)
docstatStudent <- htmlParse(Studenttext)
nameStudent <- xpathSApply(docstatStudent, '//div[@class="field field-name-body field-type-text-with-summary field-label-hidden typography"]/h3', xmlValue)
degreeStudent <- xpathSApply(docstatStudent, '//div[@class="field field-name-body field-type-text-with-summary field-label-hidden typography"]/p', xmlValue)

Allstudents <- append(degreeAlumni, degreeStudent)

```

Clean the names and data

```

formatnames <- function(nodes) {
  toupper(gsub("^ *(.*[ ])*$", "\\1", gsub("\\.", "", nodes)))
}

```

Working on Alumni Only

Create clean alumni table and turn it into a list

```

degreeAlumni <- degreeAlumni[degreeAlumni != ""]
Alumnilist <- as.list(degreeAlumni)

```

split list and maintain structure

```

splitAlumnilist <- strsplit(as.character(Alumnilist), ",")
max.length <- max(sapply(splitAlumnilist, length))
splitAlumnilist <- lapply(splitAlumnilist, function(x) { c(x, rep(NA, max.length-length(x))) })
AlumniTable <- do.call(rbind, splitAlumnilist)

```

Analysis

```

Completematrix <- matrix(data=NA, nrow=52, ncol=5)
newnames <- c("Degree", "Type", "College", "Country", "Year")
colnames(Completematrix) <- newnames

```

move columns and fix

```
AlumniTable <- gsub("\\.", "", AlumniTable)
Alumnidf <- as.data.frame(AlumniTable)
Degreepattern <- "(BA|BS|B Sc|B|MA|MS|M Sc|MBA|PhD)"
Completedf <- data.frame(Completematrix)
Completedf$Degree <- Alumnidf$V1
```

capture all the degrees

```
grepl(Degreepattern, Completedf$Degree)
```

```
## [1] TRUE TRUE FALSE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [12] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [23] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [34] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE FALSE TRUE TRUE
## [45] TRUE FALSE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

```
totalDegreepattern <- "(BA|BS|B Sc|B Tech|B Eng|MA|MS|M Ed|M Eng|M Phil|M Sc|MBA|PhD)"
grepl(totalDegreepattern, Completedf$Degree)
```

```
## [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [15] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [29] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [43] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

```
Completedf$Type <- gsub(totalDegreepattern, "", Completedf$Degree)
```

somewhat get the college

```
Completedf$College <- Alumnidf$V2
```

get last value

```
lastValue <- function(x) {
  tail(x[!is.na(x)], 1)
}
Year <- apply(Alumnidf, 1, lastValue)
Yeardf <- as.data.frame(Year)
Yeardf <- as.numeric(as.character(Yeardf$Year))
```

```
## Warning: NAs introduced by coercion
```

```
Completedf$Year <- Yeardf
```

copy alumni

```
Completedf$Degree <- substring(Completedf$Degree, 1, 3)
Completedf$Degree <- gsub(" ", "", Completedf$Degree)
Completedf$Degree <- as.character(Completedf$Degree)
```

delete bad values

```
CleanCompleteddf <- Completedf[-39,]
CleanCompleteddf <- CleanCompleteddf[-25,]
```

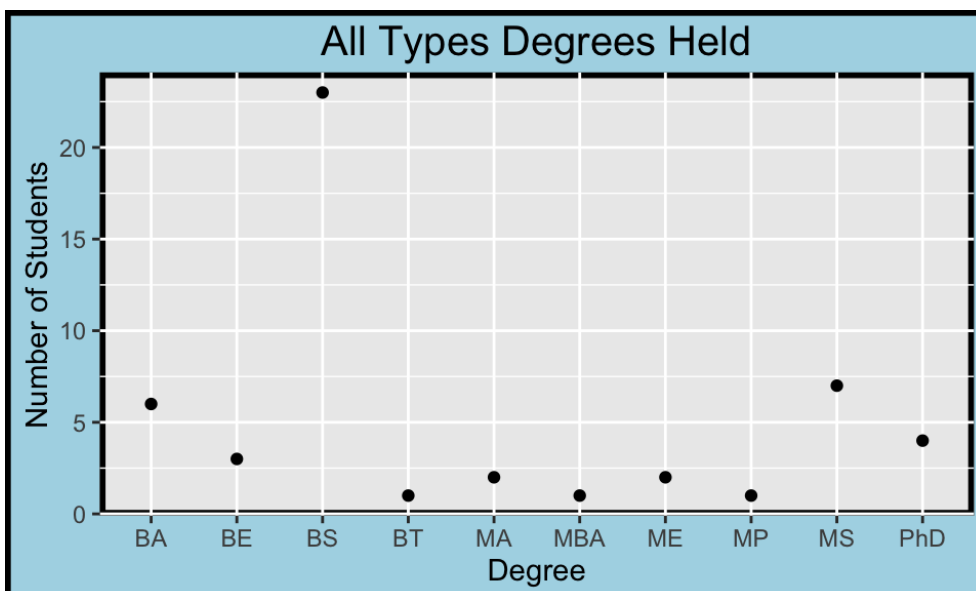
Graphing

create a new theme

```
abluetheme <- theme(plot.background = element_rect(fill = "lightblue", colour = "black", size = 2, linetype = "solid"),
                    legend.background=element_rect(colour = "black", size = 1, linetype = "solid"),
                    panel.background=element_rect(colour = "black", size = 2, linetype = "solid"),
                    plot.title=element_text(size=15))
```

find all unique degrees held by all students

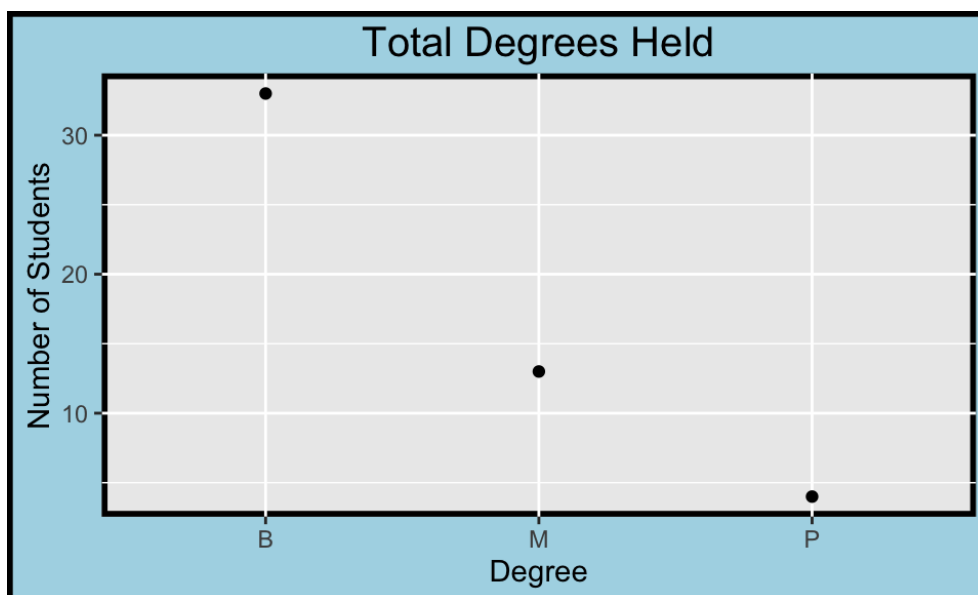
```
CleanCompleteddf %>%
  group_by(Degree) %>%
  summarize(tot=n()) %>%
  ggplot(aes(x=Degree, y = tot)) + geom_point() +
  labs(title="All Types Degrees Held") + ylab("Number of Students") + abluetheme
```



find total Bachelors, Masters, PhD held by students

```
Alldegreesdf <- CleanCompletedddf
Alldegreesdf <- Alldegreesdf %>%
  mutate(DegreeType = substring(CleanCompletedddf$Degree, 2,2))
Alldegreesdf$Degree <- substring(Alldegreesdf$Degree, 0,1)

OverviewDegrees <- Alldegreesdf %>%
  group_by(Degree) %>%
  summarize(tot=n()) %>%
  ggplot(aes(x=Degree, y = tot)) + geom_point() +
  labs(title="Total Degrees Held") +
  ylab("Number of Students") + abluetheme
OverviewDegrees
```

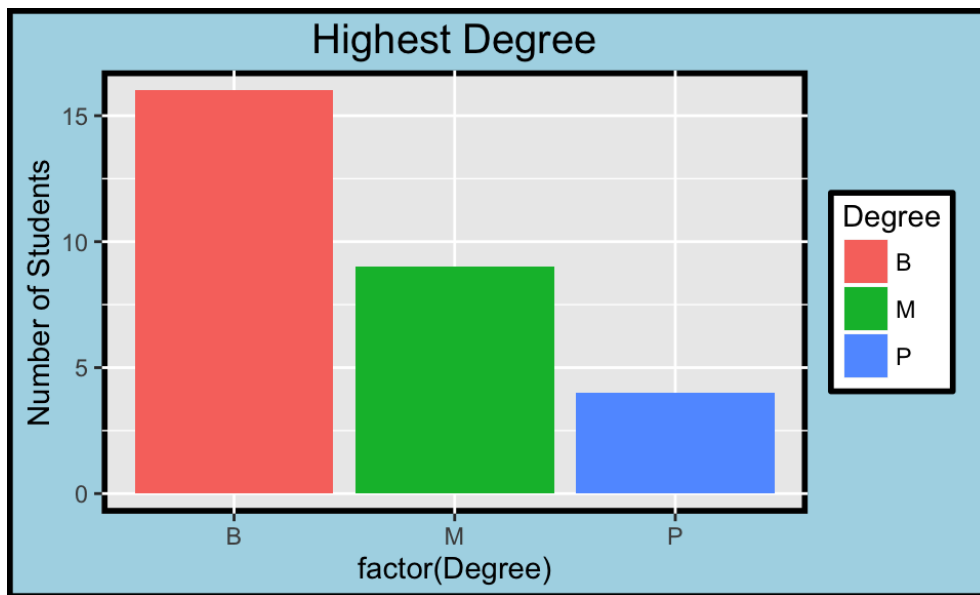


find unique degrees

```
Alldegrees <- Alldegreesdf %>%
  group_by(Degree) %>%
  summarize(tot=n())

Uniquedegrees <- Alldegrees
Uniquedegrees$tot <- c(Uniquedegrees$tot[1] - (Uniquedegrees$tot[2]+Uniquedegrees$tot[3]), Uniquedegrees$tot[2] - Uniquedegrees$tot[3], Uniquedegrees$tot[3])

Uniquedegrees %>%
  ggplot(aes(x=factor(Degree), y=tot)) +
  geom_bar(aes(fill=Degree), stat="identity") +
  labs(title="Highest Degree") +
  ylab("Number of Students") +
  abluetheme
```

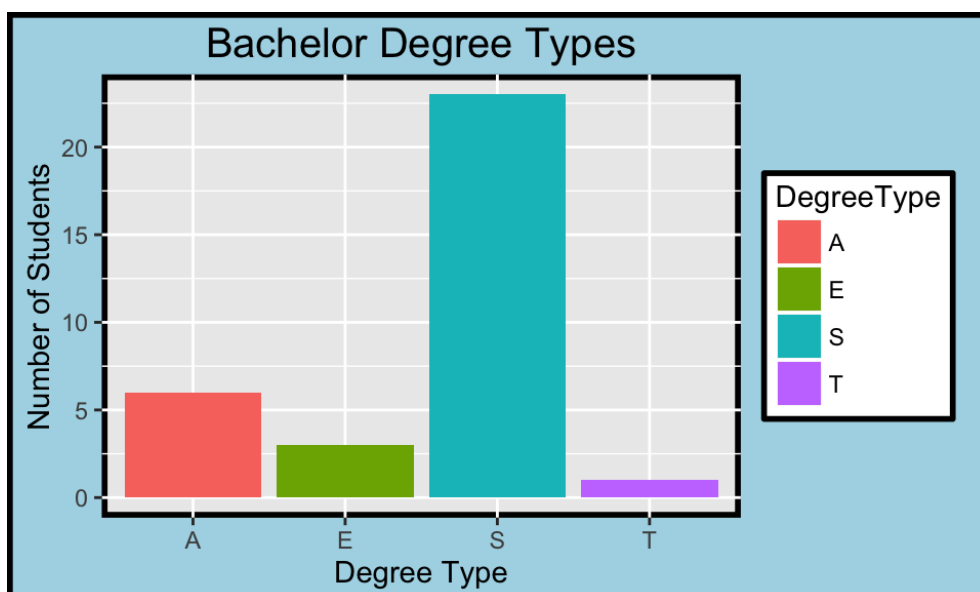


find bachelors differences only

```
Bacheloronlydf <- Alldegreesdf
Bacheloronlydf <- subset(Bacheloronlydf, Bacheloronlydf$Degree == "B")
```

Find the degree type graphs

```
Bachelordegrees <- Bacheloronlydf %>%
  group_by(DegreeType) %>%
  summarize(tot=n()) %>%
  ggplot(aes(x=factor(DegreeType), y=tot)) +
  geom_bar(aes(fill=DegreeType), stat="identity") +
  labs(title="Bachelor Degree Types") +
  xlab("Degree Type") +
  ylab("Number of Students") +
  abluetheme
Bachelordegrees
```

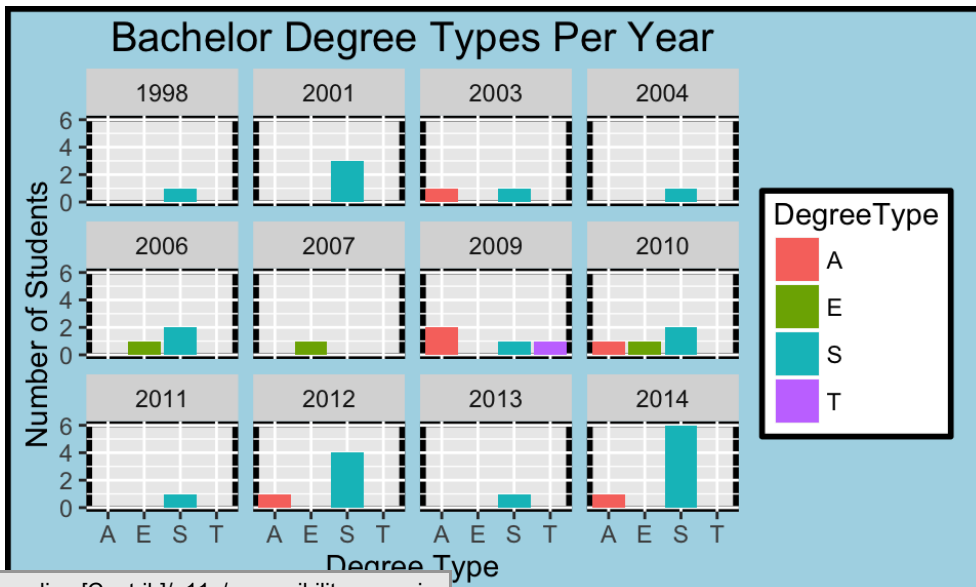


Find Bachelors by Year

```

Bacheloryear <- Bacheloronlydf %>%
  group_by(DegreeType, Year) %>%
  summarize(tot=n()) %>%
  ggplot(aes(x=factor(DegreeType), y=tot)) +
  facet_wrap(~Year) +
  geom_bar(aes(fill=DegreeType), stat="identity") +
  labs(title="Bachelor Degree Types Per Year") +
  xlab("Degree Type") +
  ylab("Number of Students") +
  abluetheme
Bacheloryear

```



Loading [Contrib]/a11y/accessibility-menu.js