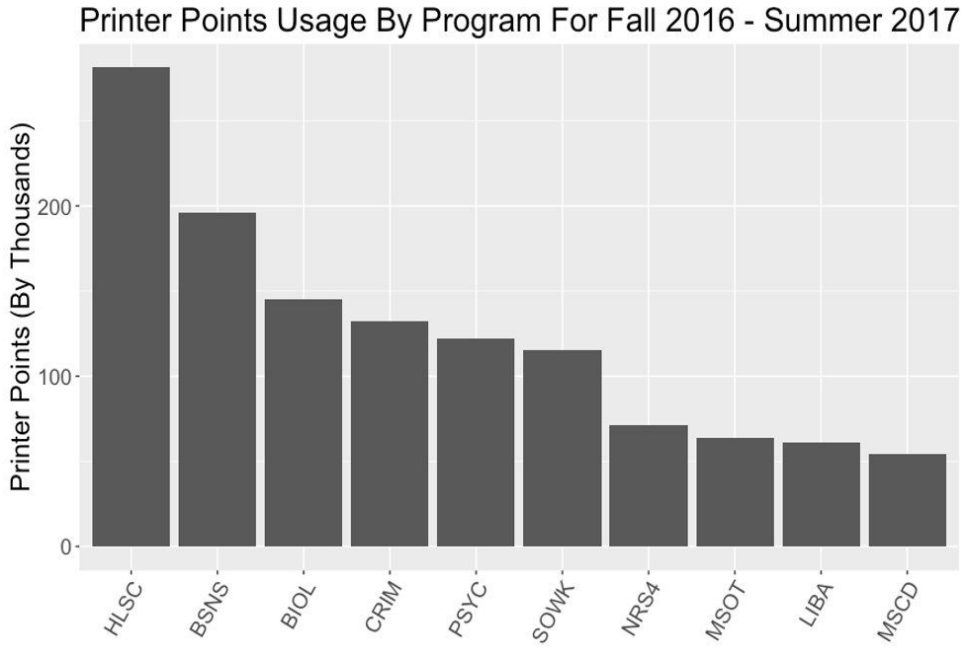
Gregory Walsh
DSSA-5101-091 - DATA EXPLORATION
Fall 2019
Final Project
Stockton Printer Data

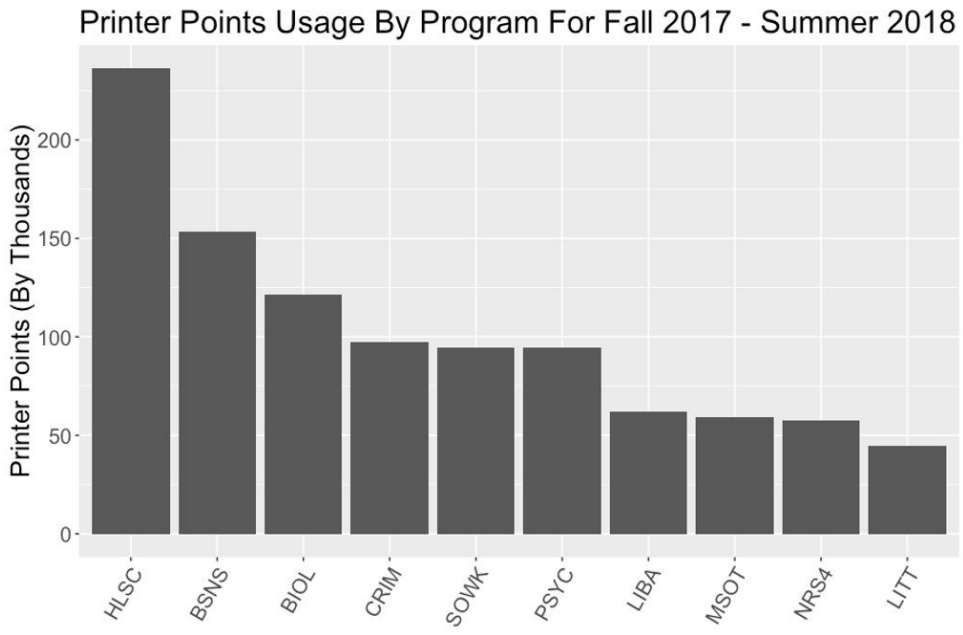
Stockton University does A LOT of printing. I am working with the CTO, Scott Huston of Stockton to look at printing data. Colored, Black and White, Letter, Legal, We have records of all prints done for the last few years and details of who, where, and when. This study will look at Student printing done for the last 6 years. The question I am specifically interested in is what is the cost of printing? What major costs the most to print? How has the cost of Printing changed overtime? It is interesting because it is a real world question Stockton is looking at right now.

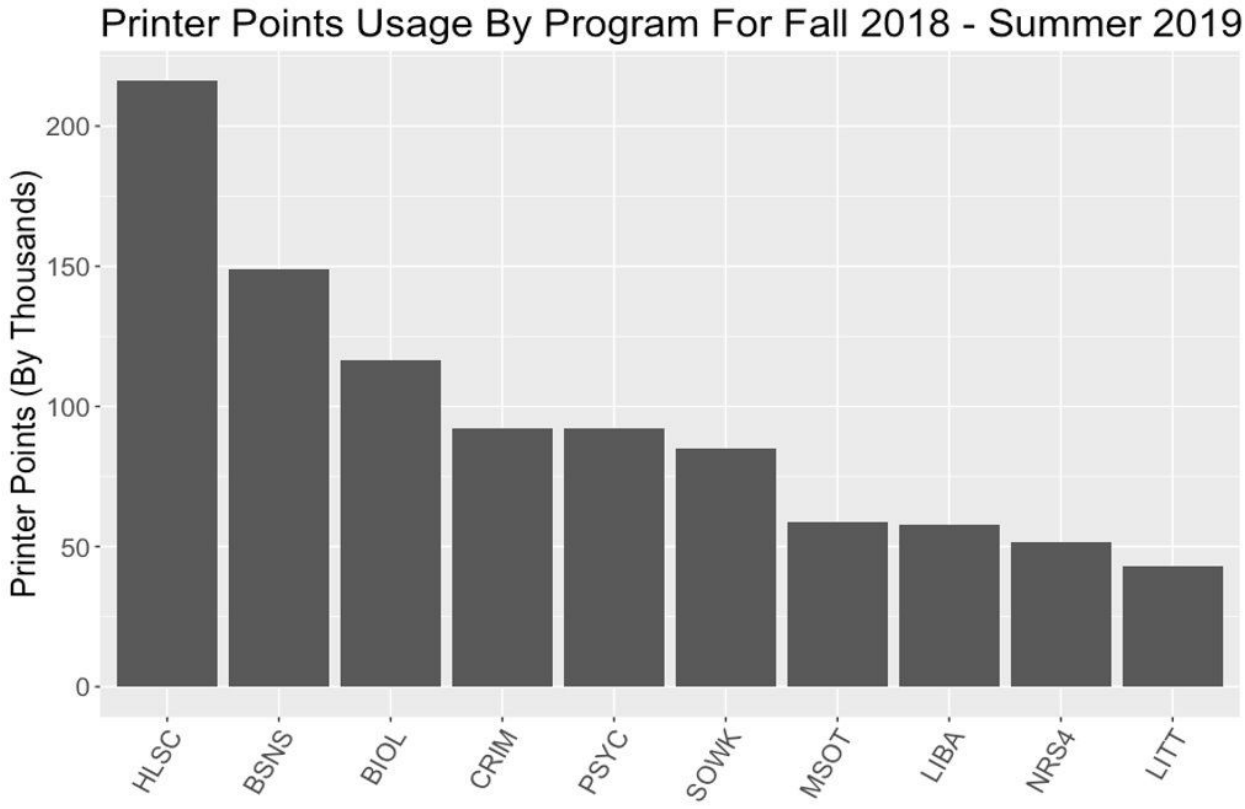
There were many limitations to the data that I needed to overcome. The Data that was originally collected is from our print queue stored in a csv file. In order to generate the CSV a powershell script needed to be run on Stocktons Print Servers. The csv contains the Time, User, Pages, Copies, Printer, Document Name, Client, Paper Size, Language, Height, Width, Duplex, Grayscale, and Size. From there I needed to get an IRB research approval to match the data with Students personal information. The data I asked for was to collect was Students Username, Semester Term, Class, Age, Sex, Major, and Minor. The Data needed to then be cleaned and the ppoints data (grabbed with powershell) needed to be merged with the Student data from the IRB. By using an Inner Join I was able to get the most crucial data from both sets into one. After the merge there were 2 csv's created. One for Black and White Prints and the other for Color Prints with about 2 million rows of data to look at. Processing the data took a good amount of time but saving it to an R Data Frame helped with further cleaning. Once the data was fully merged it needed to be cleaned. Lots of the print jobs came from service accounts used by IT for testing or assigning students points from paying tuition. Removing them allowed me to look exclusively at data printed by students over the years. There were also students that had Multiple Majors and Minors. To

solve this problem I duplicated the row so that the pages could be counted for both Majors / Minors. Finally I needed to change the dates to one uniform format

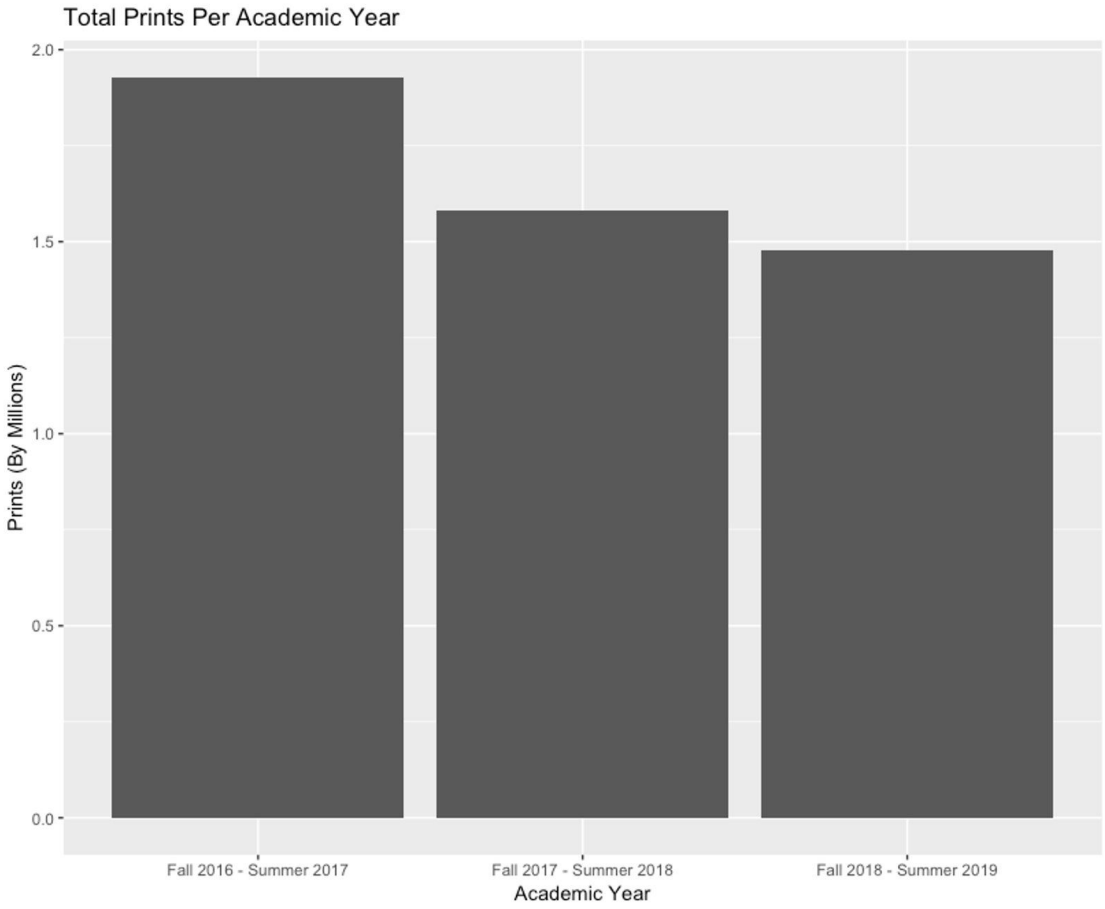
Now that I had a fully cleaned and merged dataset I could start my research. Once again I wanted to look at the cost of printing over the years. To Begin I looked at Black and White Printing. I wanted to see what majors printed the most:



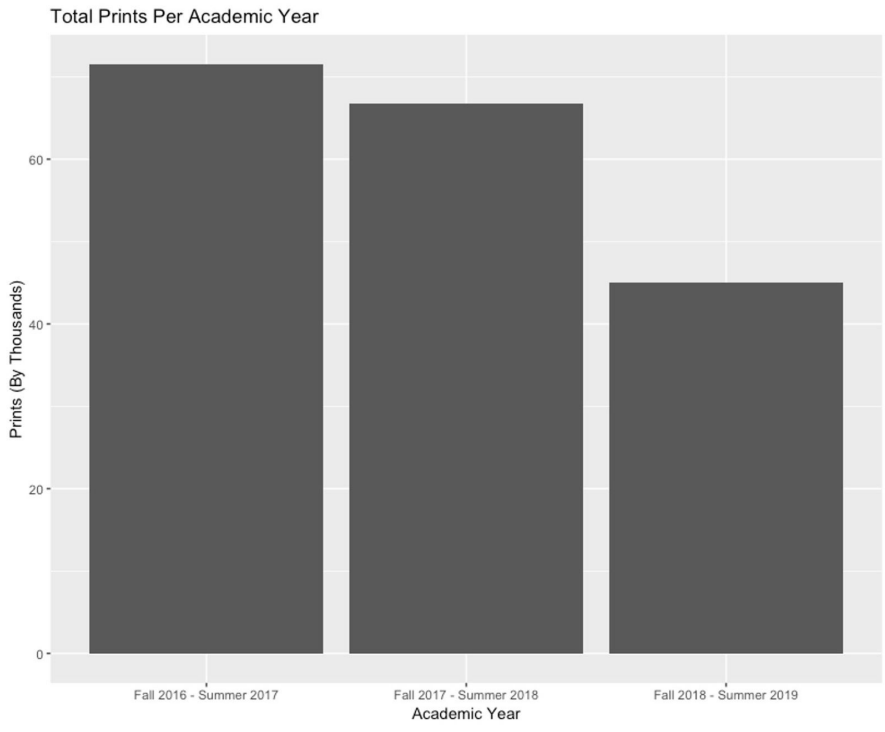




To my surprise Health Science (HLSC) Students printed a majority of total prints every year. Second Biggest being Business and Third being Biology. The more interesting observation about the data is the total prints done per academic year. I noticed although HLSC Majors printed the most each year, the total prints went down significantly (As seen by the HLSC bar getting closer to 200,000 prints). This observation intrigued me to look further into the amount printed each year. I originally thought the number of printers per year would increase and make printing more expensive. I predicted from the last observation that the printing actually decreased rather than increased as originally thought. My next Graph shows total prints per year at Stockton for Black and White Prints:

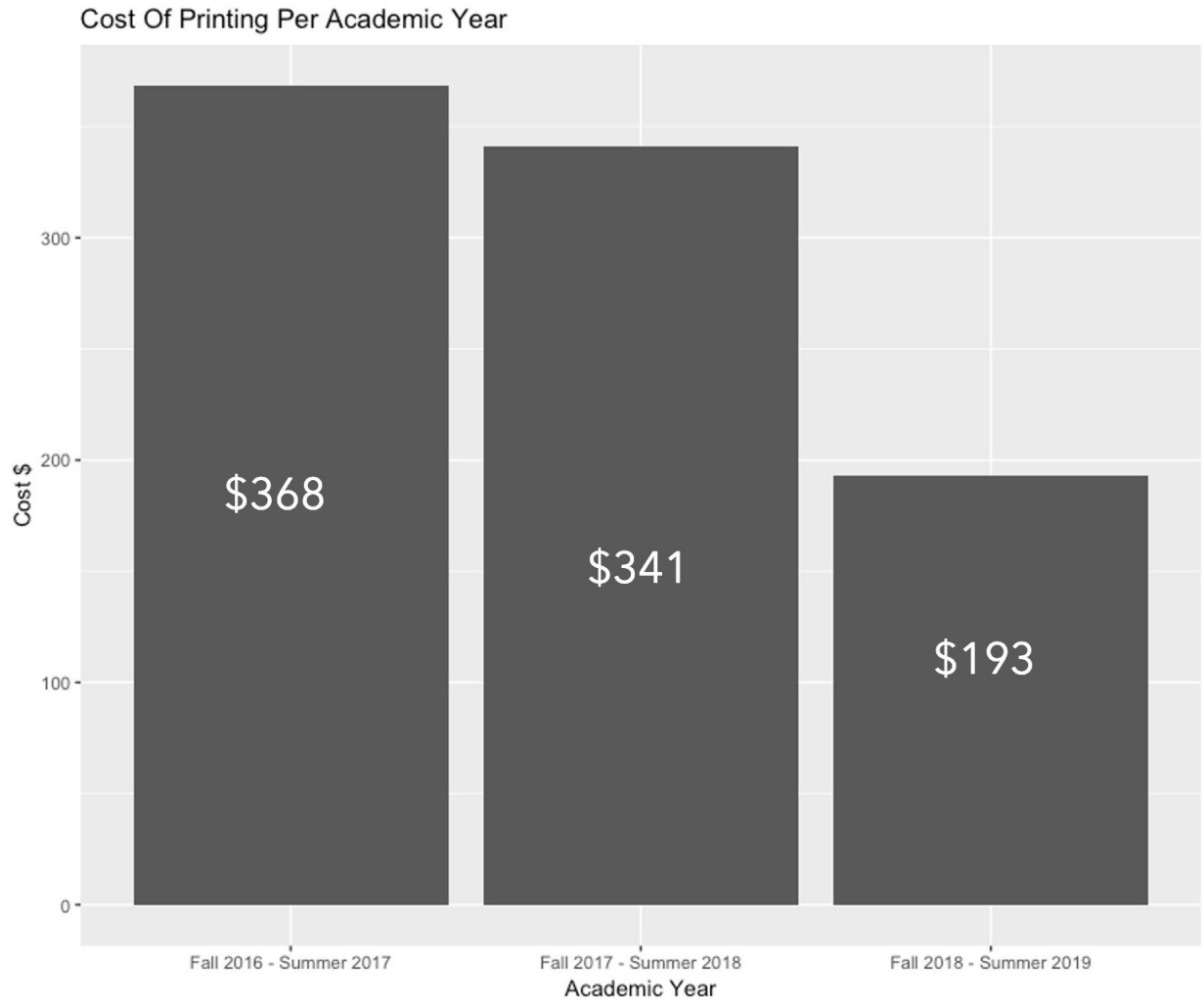


My Hypothesis was correct in that the total printers were going down significantly since Fall 2016. Stockton as a whole is printing less each year. The data was the same for Colored Prints as well but on a smaller total scale (Thousands rather than Millions):

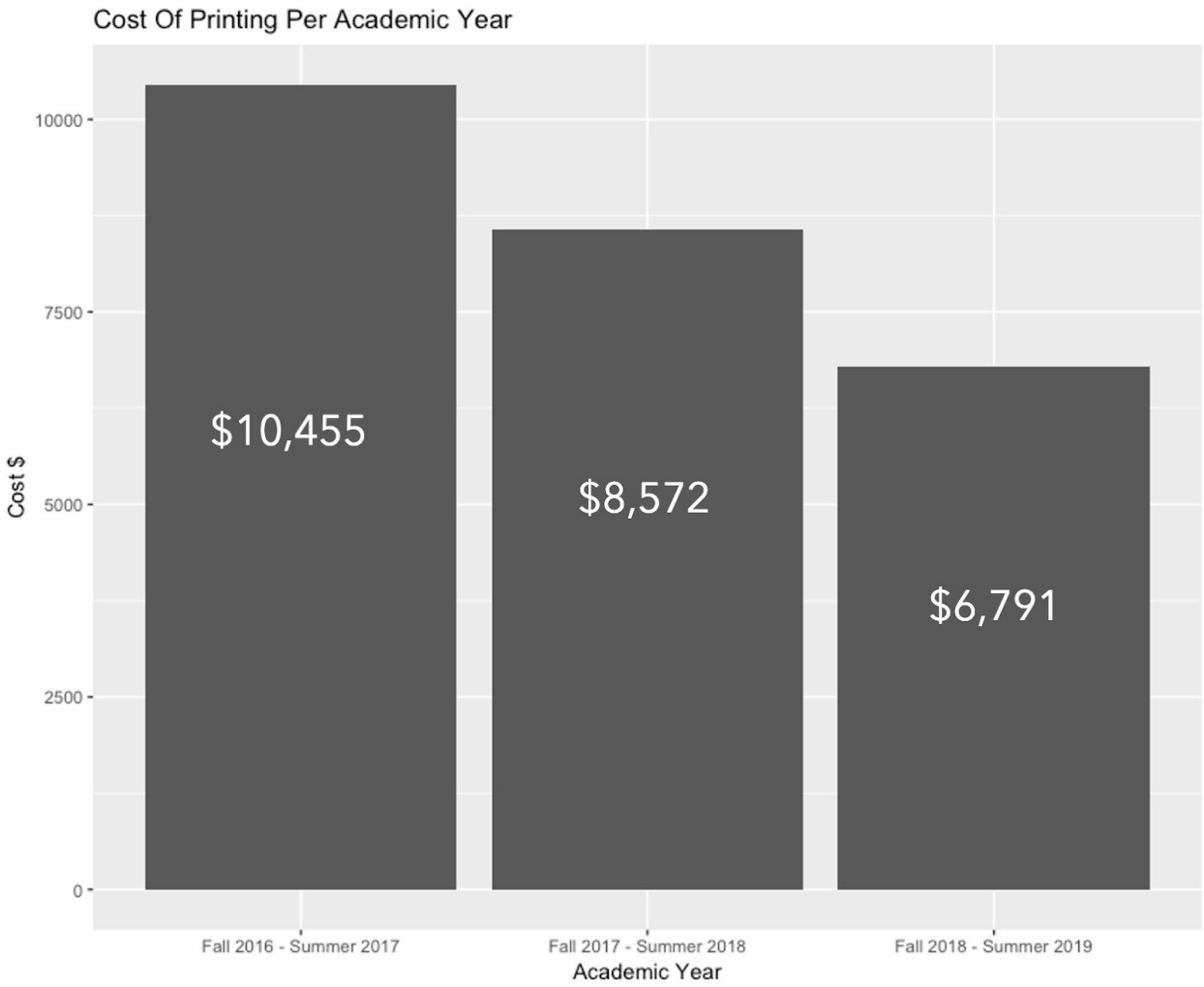


My next Goal was to look at the cost associated with printing. It is difficult to predict the amount of Toner we go through at the University because we have a contract with a company called Printer Tech where we are charged by the Page. They service the printers and replace toner when needed. Stockton does need to pay for the paper however. One case of paper contains 5000 sheets and from 2016 - 2018 Stockton Paid $27.10 a case ($ 0.00542/pg). This year Stockton changed over the type of paper used and is getting it for $22.99 ($ 0.00460/pg) The price change is only affected for Spring and Summer 2019. Looking at Pricing you can see the total cost of printer paper is on a down trend as predicted.

Color:



Black and White:



My original Question was to look at printing over time and the cost associated with it. I believed Stockton was printing more over the years however, to my surprise we are printing less. I believe this is because the use of blackboard has gone up tremendously over the years as well as the use of electronic devices in classrooms. More things are submitted online rather than in person. One of the hardships I ran into with this project is getting appropriate data on toner usage. I would like to refine my data collection to get more info from Printer Tech on how often Toner is replaced. With that data I can look at more costs other than the cost of paper. Another Area I would like to continue my research in is Canon Printers. Stockton is starting to switch over its contracts from HP Printers to Canon Printers. I would like to see the cost of the Canon Contract with Stockton and see how it compares to the Printer Tech contract for the

volume of printing that Stockton does. From a curiosity standpoint the data looked at is only of Student printing at Stockton. I am very curious to look into Printer Data of Faculty and Staff members. This would be more difficult because each department pays for their own printing bills. Getting all the data would be difficult but it is possible. Lastly I look forward to seeing how the data looks at the end of the summer. Once this academic year is over I plan to rerun the scripts and account for changing in data. My prediction is the cost has decreased a slight amount but is still about the same as last academic year. On a final note, Thank you to Stockton University for allowing me to do this research. Thank you to James Girrard and Evan Yeunge for helping with the data collection and merging as well as sharing their thoughts on the direction of my research.

Appendix.

Powershell Script To acquire printer data:

```
## Downloads a copy of printer data from PCounter and PaperCut servers, and
compresses them into an archive
## Version 1.00

$pcounter_printers = @("acprint2","acprint4","acprint6","acprint8")
$papercut_printers = @("fsprint2","fsprint3")

$dest = New-Item -Type Directory -Name "Printer Data RAW $(Get-Date -Format
MM-dd-yy)" -Force

#PaperCut PrintLogger
try {
    Foreach($printer in $papercut_printers) {
        Get-ChildItem -Filter "*.csv" -Path "\\$printer.stockton.edu\c$\Program
Files (x86)\PaperCut Print Logger\logs\csv\monthly" | foreach {
            Copy-Item -Path $_.FullName -Destination "$dest/$($printer)_$($_.Name)"
        }
        Write-Host "Finished $printer"
    }
} Catch {
    Write-Error "Error processing $printer"
}

#PCounter
Try {
    Foreach($printer in $pcounter_printers) {
        Get-ChildItem -Filter "PCOUNTER*.LOG" -Path
"\\$printer.ac.stockton\pcounterdata$" | foreach {
            Copy-Item -Path $_.FullName -Destination "$dest/$($printer)_$($_.Name)"
        }
        Write-Host "Finished $printer"
    }
} Catch {
    Write-Error "Error processing $printer"
}

Try {
    Compress-Archive -Path $dest -DestinationPath "$($dest.Name).zip"
-CompressionLevel Optimal
    Write-Host "Finished compressing archive"
} Catch {
    Write-Error "Error compressing archive"
}
```

R Script:

```
library(tidyverse)
library(ggplot2)


### Function Calls ###

acprint_merge <- function(source_csv, demographic_csv) {
  # Source
  source_df <- read_csv(
    source_csv,
    col_types = cols(
      Date = col_date("%m/%d/%Y") #convert col to date
    )
  )

  # Demographic data sent back
  demographic_df <- read_csv(
    demographic_csv,
    col_types = cols(
      PRINT_DATE = col_date("%m/%d/%Y") #convert col to date
    )
  ) %>% distinct(STUDENT_UNAME, PRINT_DATE, .keep_all = TRUE) #select distinct
names and dates


  merge_df <- merge(source_df, demographic_df, by.x = c("User","Date"), by.y =
c("STUDENT_UNAME","PRINT_DATE"), all.x = TRUE) %>%
    # Sanitize data, remove Username
    select(c("User","Document","Printer","Date","Time","Computer","Pages","Cost",
"STVTERM_CODE", "AGE", "SPBPERS_SEX", "MAJR_CODE_LIST", "MINR_CODE_LIST",
"CLASS_CODE", "RESIDENT")) %>%
    arrange(Date)

  # clean Resident data
  merge_df <- replace_na(merge_df, list(RESIDENT="N"))

  # clean Major list
  merge_df <- merge_df %>%
    # mark records that will be modified
    mutate(multiple_majors = ifelse(str_detect(MAJR_CODE_LIST, ":"), "Y", "N")) %>%
    # separate records that have multiple majors
    separate_rows(MAJR_CODE_LIST, sep = ":")

  # clean Minor list
  merge_df <- merge_df %>%
```

```
    # mark records that will be modified
    mutate(multiple_minors = ifelse(str_detect(MINR_CODE_LIST, ":"), "Y", "N")) %>%
    # separate records that have multiple majors
    separate_rows(MINR_CODE_LIST, sep = ":")

  return(merge_df)
}

# Function that takes a term code and converts it to a friendly string
term_tostring <- function(term) {
  year <- substr(term, 0, 4)
  #
https://stackoverflow.com/questions/28593265/is-there-a-function-like-switch-which-
works-inside-of-dplyrmutate
  semester <- recode(substr(term, 5, 6), "80" = "Fall", "50" = "Summer", "20" =
"Spring")


  return(paste(semester, year))
}

sort_printer_data <- function(df) {
  dfsort <- aggregate(df$Pages, by=list(Category=df$MAJR_CODE_LIST), FUN=sum)
  dfsort <- dfsort %>% arrange(desc(x))
  dfsort <- dfsort %>% head(10)
  return(dfsort)
}
sort_printer_data_color <- function(df) {
  dfsort <- aggregate(df$Cost, by=list(Category=df$MAJR_CODE_LIST), FUN=sum)
  dfsort <- dfsort %>% arrange(desc(x))
  dfsort <- dfsort %>% head(10)
  return(dfsort)
}

print_data_yearly <- function(df, year){
  ggplot(data.frame(df), aes(reorder(x=df$Category, -df$x), y=df$x / 1000)) +
    geom_bar(stat = "identity") + labs( x = NULL,
                                        y = "Printer Points (By Thousands)",
                                       title = paste("Printer Points Usage By
Program For", year )) +
    theme(axis.text.x=element_text(angle=60, hjust=1), text=element_text(size=18))
## Include for text labels + geom_text(label = df$x)
}

### Begin Data Merge ###

# Merge acprint4
```

```r
acprint4_merge <- acprint_merge("acprint4.csv", "DSSA_ACPRINT4_3.csv")
# Merge acprint6
acprint6_merge <- acprint_merge("acprint6.csv", "DSSA_ACPRINT6_3.csv")

# get unknown majors
enrollment_majors <-
c("AFST","ARTS","ARTV","COMM","HIST","LCST","LITT","MAAS","PHIL","BSNS","CMPT","CSC
I","CSIS","HTMS","INSY","MBA","CERT","EDOL","MAED","MAIT","TEDU","CERT","LIBA","MAH
G","CERT","DNP","DPT","EXSC","HLSC","MSCD","MSN","MSOT","NRS4","NURS","PUBH","SPAD"
,"BCMB","BIOL","CHEM","CPLS","DSSA","ENVL","GEOL","MARS","MATH","MSCP","PHYS","PSM"
,"SSTB","CERT","COUN","CRIM","ECON","MACJ","MSW","POLS","PSYC","SOCY","SOWK","NMAT"
,"UNDC")
acprint6_merge %>% filter(!MAJR_CODE_LIST %in% enrollment_majors) %>%
select(MAJR_CODE_LIST) %>% distinct()
acprint4_merge %>% filter(!MAJR_CODE_LIST %in% enrollment_majors) %>%
select(MAJR_CODE_LIST) %>% distinct()

#acprint6_merge <- acprint6_merge[is.na(acprint6_merge$Document),]
### Being Data Cleaning ###
## Remove IT Admins Username from Printer account
acprint6_merge <- acprint6_merge[!acprint6_merge$Printer == "AC\\gallag99"
                      & !acprint6_merge$Printer == "AC\\koppt"
                      & !acprint6_merge$Printer == "AC\\roubosd"
                      & !acprint6_merge$Printer == "AC\\yeunge"
                      & !acprint6_merge$Printer == "AC\\admin-kapmat"
                      & !acprint6_merge$Printer == "AC\\pestritm"
                      & !acprint6_merge$Printer == "AC\\Dan"
                      & !acprint6_merge$Printer == "AC\\mcadmin"
                      ,  ]

acprint4_merge <- acprint4_merge[!acprint4_merge$Printer == "AC\\gallag99"
                                 & !acprint4_merge$Printer == "AC\\koppt"
                                 & !acprint4_merge$Printer == "AC\\roubosd"
                                 & !acprint4_merge$Printer == "AC\\yeunge"
                                 & !acprint4_merge$Printer == "AC\\admin-kapmat"
                                 & !acprint4_merge$Printer == "AC\\pestritm"
                                 & !acprint4_merge$Printer == "AC\\Dan"
                                 & !acprint6_merge$Printer == "AC\\mcadmin"
                                 ,  ]

## Remove the Documents Col
acprint6_merge = acprint6_merge[-c(2)]

## Things to Look For
## What Major Prints the most
## Pages X Cost
## Acprint4
```

```
## Need to do this for AC Print 6 as well
## Acprint4
df1 <- filter(acprint4_merge, Pages > 0 )
df1 <- df1 %>% filter(MAJR_CODE_LIST %in% enrollment_majors)

## Semester FA 19 + SP 20 + SU 20

## 2016/04/

#Fall 2019 September 1 - Dec 20
#Summer 19 May 10 - Aug 10th
#Spring 19 Jan 12 - May 6
df1 <- df1[!is.na(df1$MAJR_CODE_LIST),]
subset2k16 <- subset(df1, format(as.Date(df1$Date),"%Y/%M/%D") >= "2016-09-01" &
df1$Date <= "2017-08-10")
subset2k17 <- subset(df1, format(as.Date(df1$Date),"%Y/%M/%D") >= "2017-09-01" &
df1$Date <= "2018-08-10")
subset2k18 <- subset(df1, format(as.Date(df1$Date),"%Y/%M/%D") >= "2018-09-01" &
df1$Date <= "2019-08-10")
cost2k18 <- subset(df1, format(as.Date(df1$Date),"%Y/%M/%D") >= "2018-09-01" &
df1$Date <= "2019-01-01")
cost2k19 <- subset(df1, format(as.Date(df1$Date),"%Y/%M/%D") >= "2019-01-02" &
df1$Date <= "2019-08-10")
sumsubset2k16 <- sum(subset2k16$Pages)
sumsubset2k17 <- sum(subset2k17$Pages)
sumsubset2k18 <- sum(subset2k18$Pages)

sum2k18 <- sum(cost2k18$Pages) * 0.004598
sum2k19 <- sum(cost2k19$Pages) * 0.004598
total <- sum2k18 + sum2k19

pagecost2k17 <- sumsubset2k17 * 0.00542
pagecost2k16 <- sumsubset2k16 * 0.00542

cost_sums <- data.frame(c(pagecost2k16, pagecost2k17, total), c("Fall 2016 - Summer
2017","Fall 2017 - Summer 2018","Fall 2018 - Summer 2019" ))
ggplot(data.frame(cost_sums), aes(y=cost_sums$c.pagecost2k16..pagecost2k17..total.
,
x=cost_sums$c..Fall.2016...Summer.2017....Fall.2017...Summer.2018....Fall.2018...Su
mmer.2019..)) +
  geom_bar(stat = "identity") +
  labs(x = "Academic Year",
       y = "Cost $",
       title = "Cost Of Printing Per Academic Year")


print_sums <- data.frame(c(sumsubset2k16, sumsubset2k17, sumsubset2k18 ), c("Fall
```

```
2016 - Summer 2017","Fall 2017 - Summer 2018","Fall 2018 - Summer 2019" ))
ggplot(data.frame(print_sums),
aes(y=print_sums$c.sumsubset2k16..sumsubset2k17..sumsubset2k18. / 1000000,
x=print_sums$c..Fall.2016...Summer.2017....Fall.2017...Summer.2018....Fall.2018...S
ummer.2019..)) +
  geom_bar(stat = "identity") +
  labs(x = "Academic Year",
       y = "Prints (By Millions)",
       title = "Total Prints Per Academic Year")

print_data_yearly(sort_printer_data(subset2k16), "Fall 2016 - Summer 2017")
print_data_yearly(sort_printer_data(subset2k17), "Fall 2017 - Summer 2018")
print_data_yearly(sort_printer_data(subset2k18), "Fall 2018 - Summer 2019")
ggplot(data.frame(top10), aes(reorder(x=top10$Category, -top10$x), y=top10$x)) +
  geom_bar(stat = "identity")
## Are people spending more to print over the years
## Are the majors changing over the years
ggplot(data.frame(acprint6_merge), aes(x=acprint6_merge$Date,
y=acprint6_merge$Cost)) +
  scale_x_date(limits = as.Date(c("2016-04-01","2016-04-29")), date_breaks = "1
day") +
  geom_bar(stat = "identity") + labs(x = "Date",
                                     y = "Printer Points",
                                     title = "Printer Points Usage") +
  theme(axis.text.x=element_text(angle=60, hjust=1))

ggplot(data.frame(acprint6_merge), aes(x=acprint6_merge$Date,
y=acprint6_merge$Cost)) +
  scale_x_date(limits = as.Date(c("2018-09-01","2018-12-31")), date_breaks = "1
weeks") +
  geom_bar(stat = "identity") + labs(x = "Date",
                                     y = "Printer Points",
                                     title = "Printer Points Usage") +
  theme(axis.text.x=element_text(angle=60, hjust=1))

## ACPRINT6

df2 <- filter(acprint6_merge, Cost > 0 )
df2 <- df2 %>% filter(MAJR_CODE_LIST %in% enrollment_majors)
df2 <- df2[!is.na(df2$MAJR_CODE_LIST),]

subset2k16color <- subset(df2, format(as.Date(df2$Date),"%Y/%M/%D") >= "2016-09-01"
& df2$Date <= "2017-08-10")
subset2k17color <- subset(df2, format(as.Date(df2$Date),"%Y/%M/%D") >= "2017-09-01"
& df2$Date <= "2018-08-10")
subset2k18color <- subset(df2, format(as.Date(df2$Date),"%Y/%M/%D") >= "2018-09-01"
& df2$Date <= "2019-08-10")
```

```
cost2k18color <- subset(df2, format(as.Date(df2$Date),"%Y/%M/%D") >= "2018-09-01" &
df2$Date <= "2019-01-01")
cost2k19color <- subset(df2, format(as.Date(df2$Date),"%Y/%M/%D") >= "2019-01-02" &
df2$Date <= "2019-08-10")
#subset2k19 <- subset(df1, format(as.Date(df1$Date),"%Y/%M/%D") >= "2016-09-01" &
df1$Date <= "2017-08-10")
sumsubset2k16color <- sum(subset2k16color$Cost)
sumsubset2k17color <- sum(subset2k17color$Cost)
sumsubset2k18color <- sum(subset2k18color$Cost)

sum2k18color <- sum(cost2k18color$Cost) * 0.004598
sum2k19color <- sum(cost2k19color$Cost) * 0.004598
totalcolor <- sum2k18color + sum2k19color

pagecost2k17color <- sumsubset2k17color * 0.00542
pagecost2k16color <- sumsubset2k16color * 0.00542

cost_sumscolor <- data.frame(c(pagecost2k16color, pagecost2k17color, totalcolor),
c("Fall 2016 - Summer 2017","Fall 2017 - Summer 2018","Fall 2018 - Summer 2019" ))
ggplot(data.frame(cost_sums),
aes(y=cost_sumscolor$c.pagecost2k16color..pagecost2k17color..totalcolor.,
x=cost_sumscolor$c..Fall.2016...Summer.2017....Fall.2017...Summer.2018....Fall.2018
...Summer.2019..)) +
  geom_bar(stat = "identity") +
  labs(x = "Academic Year",
       y = "Cost $",
       title = "Cost Of Printing Per Academic Year")

print_sumscolor <- data.frame(c(sumsubset2k16color, sumsubset2k17color,
sumsubset2k18color), c("Fall 2016 - Summer 2017","Fall 2017 - Summer 2018","Fall
2018 - Summer 2019" ))
ggplot(data.frame(print_sumscolor),
aes(y=print_sumscolor$c.sumsubset2k16color..sumsubset2k17color..sumsubset2k18color.
/ 1000,
x=print_sumscolor$c..Fall.2016...Summer.2017....Fall.2017...Summer.2018....Fall.201
8...Summer.2019..)) +
  geom_bar(stat = "identity") +
  labs(x = "Academic Year",
       y = "Prints (By Thousands)",
       title = "Total Prints Per Academic Year")
print_data_yearly(sort_printer_data_color(subset2k16color), "Fall 2016 - Summer
2017")
print_data_yearly(sort_printer_data_color(subset2k17color), "Fall 2017 - Summer
2018")
print_data_yearly(sort_printer_data_color(subset2k18color), "Fall 2018 - Summer
2019")
sum(subset2k16color$Cost)
```

```
sum(subset2k17color$Cost)
sum(subset2k18color$Cost)
ggplot(data.frame(top10), aes(reorder(x=top10$Category, -top10$x), y=top10$x)) +
  geom_bar(stat = "identity")
## Are people spending more to print over the years
ggplot(data.frame(acprint6_merge), aes(x=acprint6_merge$Date,
y=acprint6_merge$Cost)) +
  scale_x_date(limits = as.Date(c("2016-04-01","2016-04-29")), date_breaks = "1
day") +
  geom_bar(stat = "identity") + labs(x = "Date",
                                     y = "Printer Points",
                                     title = "Printer Points Usage") +
  theme(axis.text.x=element_text(angle=60, hjust=1))

ggplot(data.frame(acprint6_merge), aes(x=acprint6_merge$Date,
y=acprint6_merge$Cost)) +
  scale_x_date(limits = as.Date(c("2018-09-01","2018-12-31")), date_breaks = "1
weeks") +
  geom_bar(stat = "identity") + labs(x = "Date",
                                     y = "Printer Points",
                                     title = "Printer Points Usage") +
  theme(axis.text.x=element_text(angle=60, hjust=1))
```