So you want to be a data scientist?

An Exploration of Data Scientists in the 2013 US Census

Benjamin Hulet, Zhenshan Jin, and Yuqiu Yang

1 Introduction

As Sherlock Holmes says in *The Adventure of the Copper Beeches* "Data! Data! Data! I can't make bricks without clay.", data science is becoming one of the most fundamental building blocks of many new professions and academic areas. With more and more people investing their time and energy into exploring this exciting career, methods for finding patterns in big data, making valuable insights, etc. emerge successively. As potential data scientists eager to be better prepared for this exciting career, in our analysis, we investigated different factors that will influence our decision on which industry to enter and which degree to pursue. By using the 2013 US census data with 3,132,795 observations and 283 variables and selecting different variables, we are able to delve into the impact of English proficiency on total income and industries people will enter, glass ceiling in data science professions, and the value of different degrees.

2 Data Filtering

2.1 Variables summary

The following is the table of the variables we used and their corresponding explanations.

Variables	Explanation
ST	Name of States
AGEP	Age
CIT	Citizenship
ENG	Ability to speak English
MARHT	Number of times married
ESR	Employment status
SCHL	Educational attainment
SEX	Gender
FOD1P	First field of degree
FOD2P	Second field of degree
WKHP	Usual woking hours per week
WAOB	World area of birth
OCCP	Occupation

Table 1: Summary of variables

2.2 Data Scientist

To select Data Scientists, we focused on people with data science related degrees for either their first or second degree. We defined data science degrees as computer programming and data processing, computer science, mathematics, applied mathematics, statistics and decision science, mathematics and computer science and management information systems and statistics. We found that all people with data science degrees, as sampled by the census, had a bachelors degree or higher. After filtering the data for Data Science related degrees, we have 873,033 observations.

3 Where are we?

Before our investigation, we think it would be helpful for us to paint the "big picture" so as to get an overview of the distribution of data scientists in the U.S.. The following map shows the proportion of data scientists in each state.

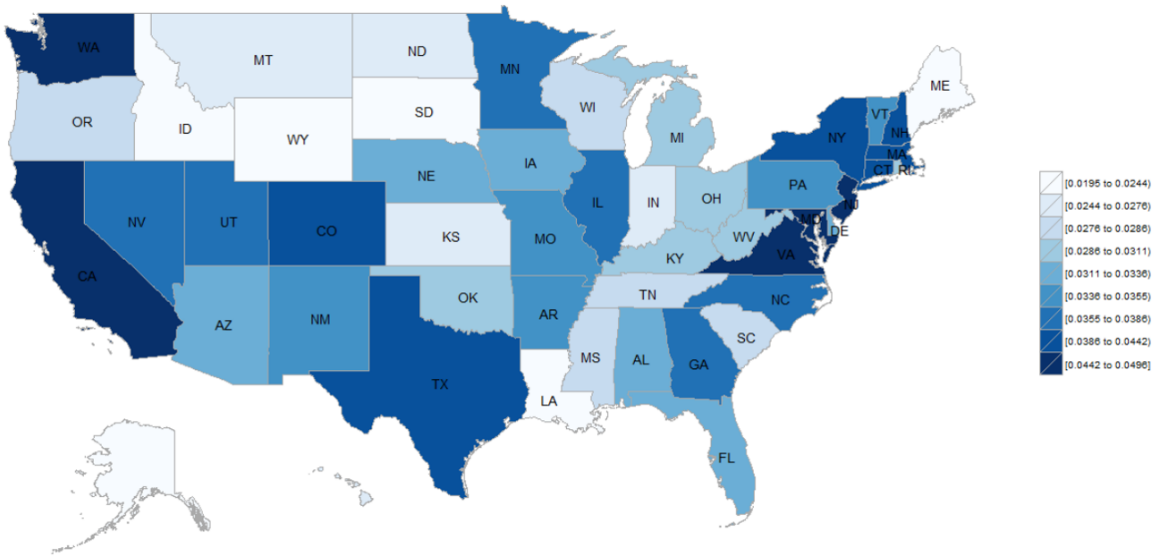


Figure 1: Distribution of Data Scientist

Looking at the map, we can see that California, Washington and Virginia have the highest density of data scientists and within these states approximately 5% of degree holders have a data science related degree. We can see a relationship between the type of industry the region is known for and the density of data science degree holders. For example Silicon Valley in California, the headquarters of Microsoft and Amazon in Washington, and the US government and related Defense industries in Virginia. And as for Texas, the proportion is also relatively high: around 4%, likely because oil companies are seeking tech savvy graduates to help reduce expenditures.

4 Were you born with right status?

4.1 Where were you born?

Since the U.S is a highly diversified country, we would like to check whether people's birth location has an impact on their total income.

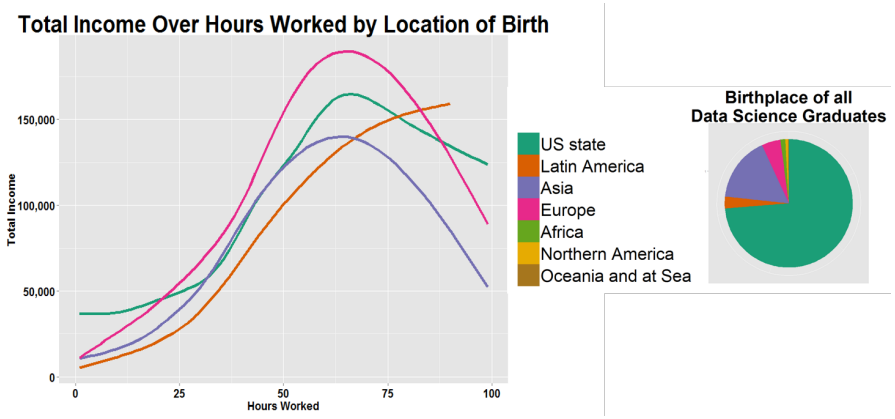


Figure 2: (Left) Total income against working hours by location of birth.
(Right) Proportion of people by location of birth

From the pie chart(right), we can find out that 25% of the data science degrees are held by people who were born in a country different than the US. What's more, besides U.S. citizens, the majority of people with a data science degree are from Asia, Europe and Latin America. In order to avoid overplotting, we focused on people with these four backgrounds. We excluded the small number of outliers who were born in Africa and U.S. Territories.

By plotting total income against working hours for these four backgrounds, we can see that there is an income disparity depending on location of birth. And if we check people working a traditional number of working hours per week, the greatest gap occurs between Europeans and Latin Americans which is approximately 25 thousand dollars. As a result, it seems that birth location plays an relatively significant role in prediction total income.

We, however, noticed the tendency that Asians and Latin Americans whose first language might not be English make less than Europeans and Americans, which leads us to think that maybe it is not just because where people were born that determines total income, what really matters might be how fluently people can speak English.

4.2 Does Communication Matter?

By plotting earning per hour against degrees split by levels of English proficiency, not only are we able to answer this question, but we can also get one step further: investigating the impact of the interaction between degree and English proficiency on earning per hour.

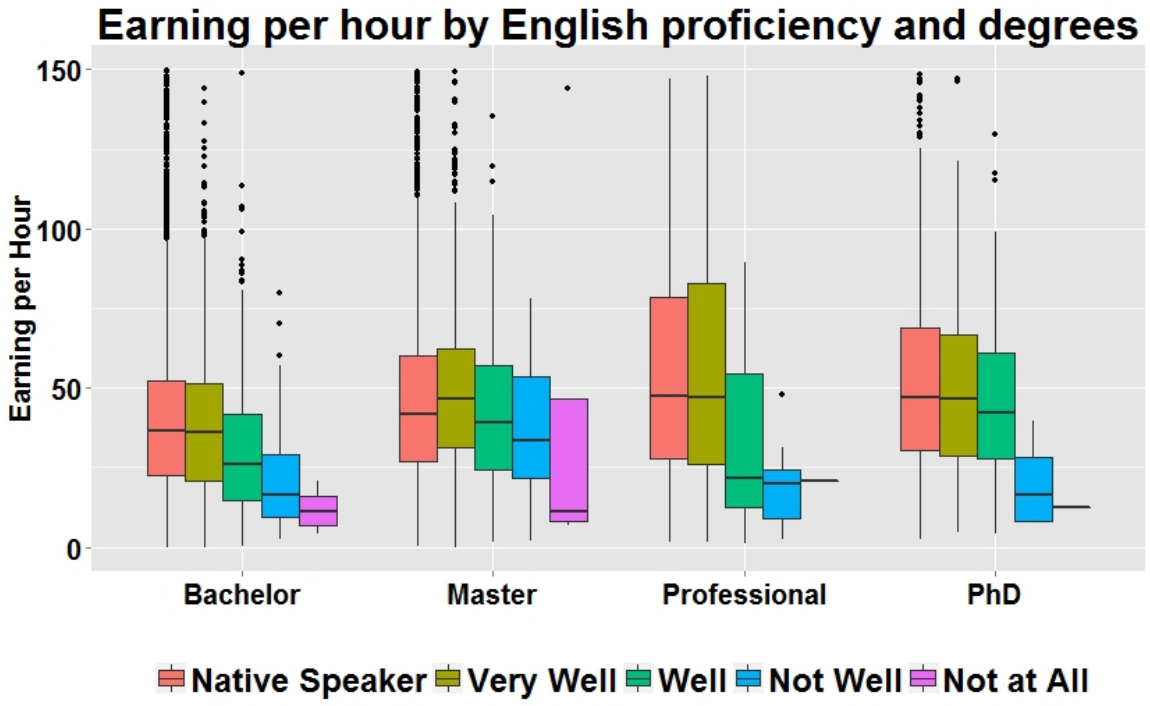
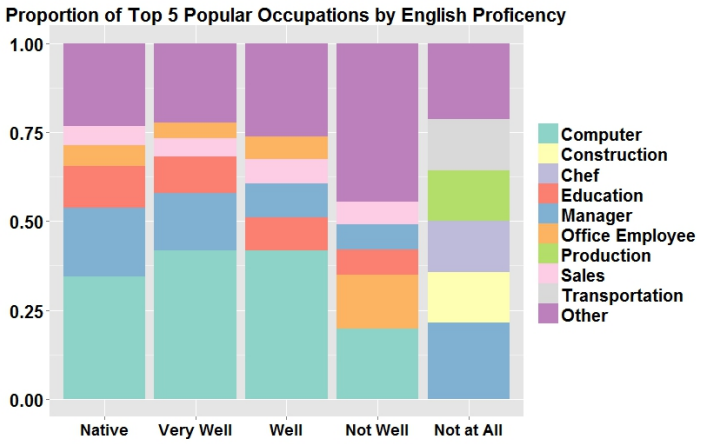


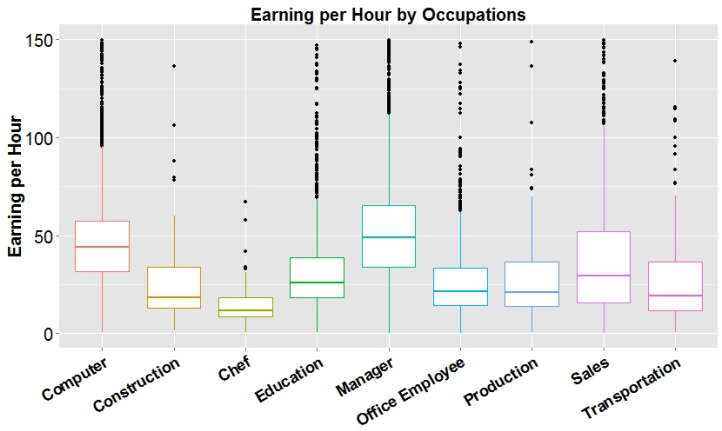
Figure 3: Earning per hour against degrees split by degrees of English proficiency

The graph shows a quite significant positive relationship between English proficiency and earning per hour in general. Moreover, by comparing different degrees, we find it interesting that getting a master or a PhD seems to be able to alleviate this impact.

And just out of curiosity, we summarized the data and tried to find out what occupations people with different English levels will likely to enter.



(a) Top 5 popular occupations by English proficiency



(b) Earning per hour by occupations

Figure 4: English proficiency and occupations

We can't help but notice that not only people who can not speak English at all actually wasted their data science degrees by becoming a chef, construction labor etc, but in general they also tend to earn less than their counterparts.

Within the rest of the English levels, the proportion of occupations which people who have a firm grasp on English will enter displays a similar pattern: the majority of them would enter computer occupations while the rest would likely to choose a career in education and management. It is worth noting, that occupations in computer technology, education, and management, are the occupations to most likely make use of the data science degree. whereas people who can not speak English well besides computer science, their alternative will be becoming office employees.

5 What is the influence of gender in data science professions?

In the past several years, many news outlets reported the relative sparsity of women entering Science Technology Engineering and Math (STEM) career paths and the pay discrepancy that exists for women who chose to enter a STEM profession.

5.1 Distribution of degrees based on gender

To evaluate these claims we wanted to see the distribution of degrees by gender. We found that there are in fact many more men than women earning STEM degrees, however, the proportion of men and women who are earning Bachelors degrees versus PhD's is comparable. There is a slight negative trend in the proportion of women to men who are earning the upper level degrees, however, this difference is not great.

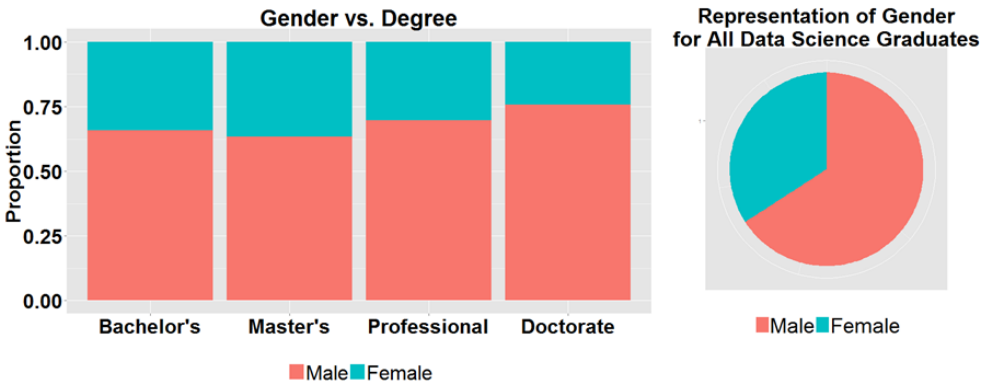


Figure 5: Distribution of Data Science Degrees

5.2 Is there a pay gap?

By plotting total income by hours worked we can see if men and women with similar educational backgrounds are earning different amounts.

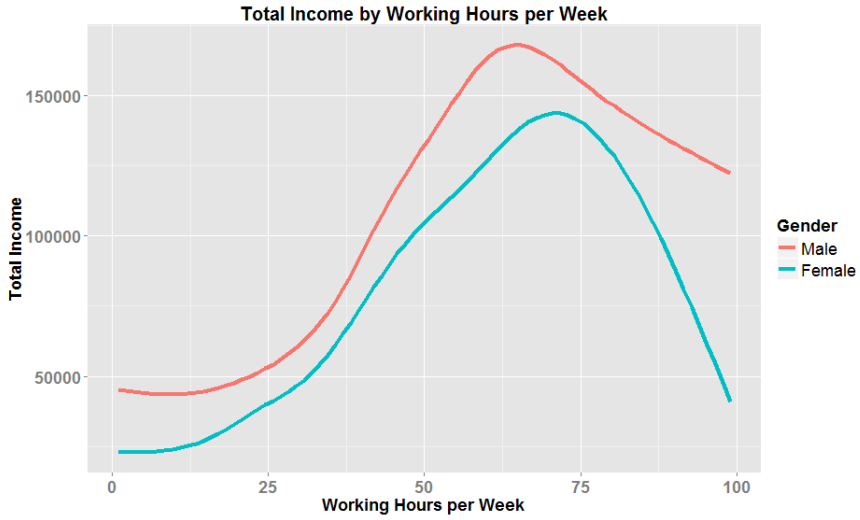


Figure 6: Total Income by Hours Worked per Week

For an average working week of 40 hours, women are earning significantly less than men. This difference amounts to approximately twenty thousand dollars.

5.3 Are men and women entering different occupations?

To understand why men and women with similar backgrounds are earning vastly different salaries, we explored which occupations they are entering.

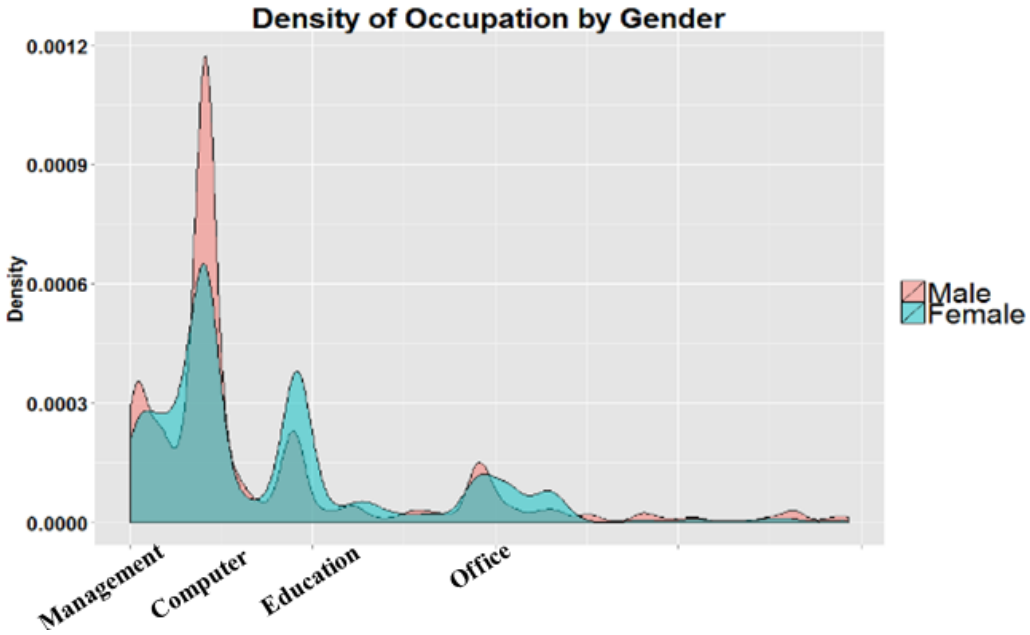


Figure 7: Density of Occupation by Gender

We see that men and women are working in similar occupations, however, there are a few subtle differences. Very generally, the density plot is more peaked for men than it is for women. This suggests that women see their roles more broadly than men and are more likely to work with diverse responsibilities. Additionally, for the highest paying occupations, we see a greater proportion of men than women in these professions. For example, on the far left of the plot you can see that there is a higher proportion of

men entering upper management positions which are generally well compensated. Similarly, for computer programming positions, we do see a higher proportion of men in this field than women. These observations are contrasted against general management, education, and office roles which may not be as highly compensated.

5.4 Is there a difference in income for specific occupations?

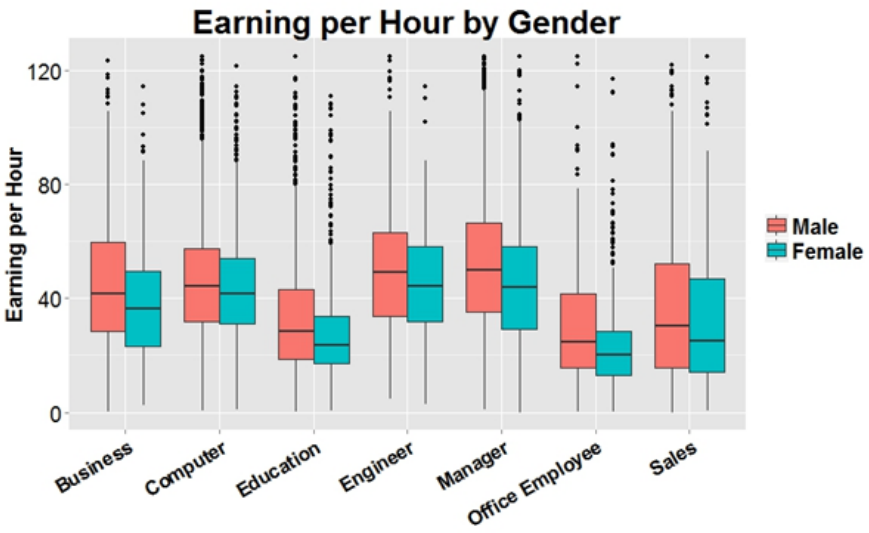


Figure 8: Income per hour by Gender

Because we see a higher density of women in lower paying occupations we have partially explained the income disparity within the data science graduates. However, from this plot we can also see that for every occupation that the data science graduates enter, men report a higher median income than women. We were not able to uncover any additional factors that might explain why this difference exists. From this exploration, we find that this is strong evidence of gender bias within data science.

6 What is the value of your degree?

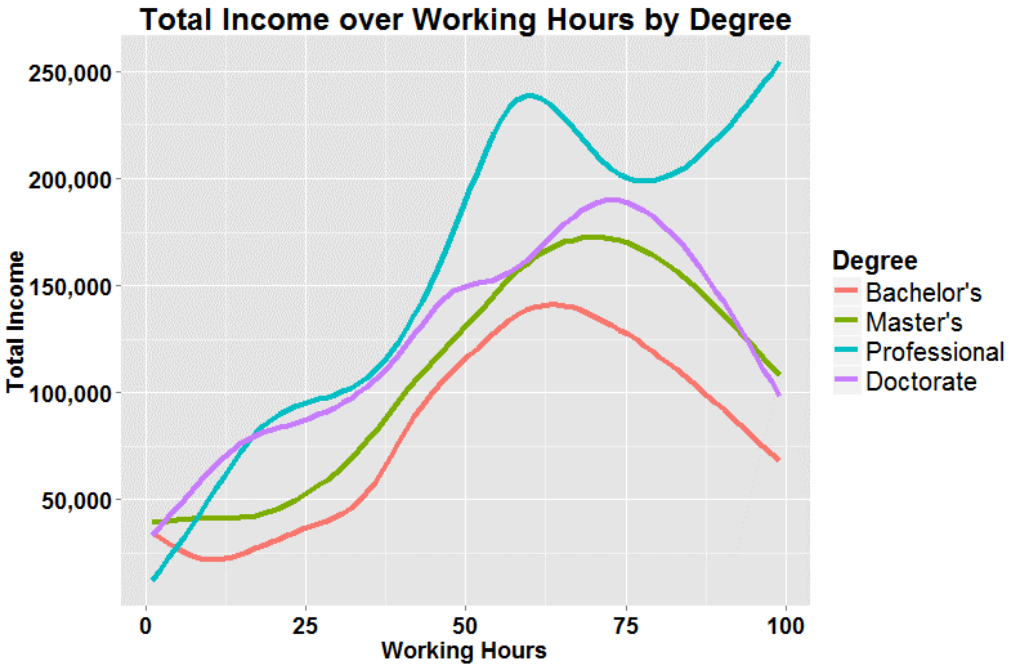


Figure 9: Total income over working hours by degree

The plot suggests in general a positive relationship between degree and earning per hour which means greater investments on average compensate more per hour. Also, we find it interesting that professional degrees outrun all the other degrees including PhD. So we buckled down and found out that professional degree albeit not being well represented within our data, means having both a data science background but also a business background. As a result, it appears that having a business background will likely increase earnings per hour.

7 Relationship between degree and occupation

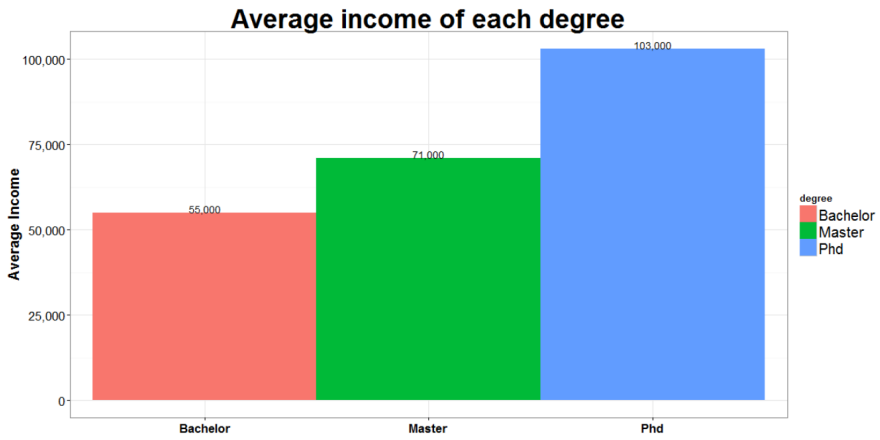


Figure 10: Income of each degree statistics

From the previous plot, we can find that there do have a huge gap of the average income among each degree

7.1 Reasons for income disparity

First, let's see the occupation proportion of each degree.

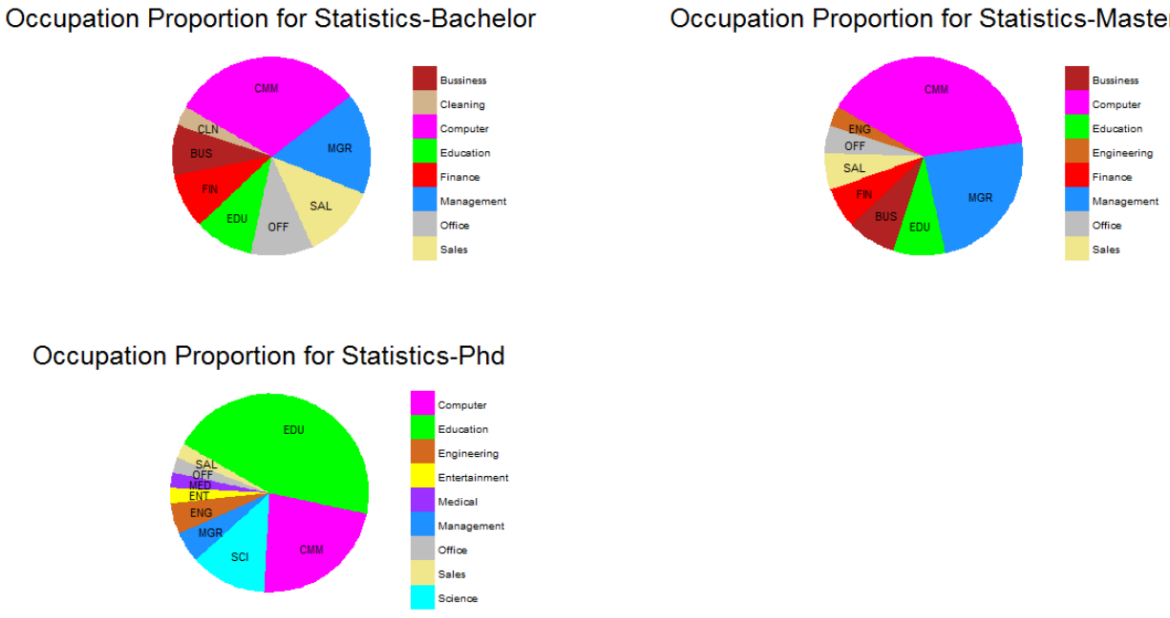


Figure 11: Occupation proportion of each degree

From the pie chart which is about the occupation proportion for statistics PhD, we find that almost half occupation is Education, and it's totally different from master and bachelor's occupation, in which computer and management make up the

most part.
Then we come up with two possible reasons. First is that does education occupation pay more than other occupations? The second is that do PhD earn more income in all occupations?

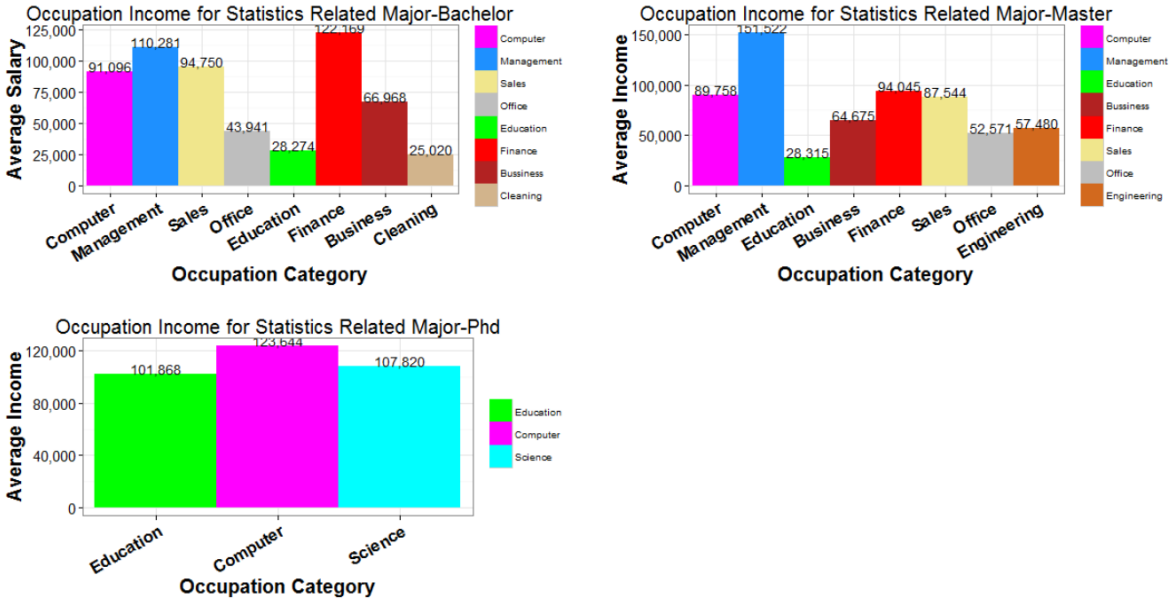


Figure 12: Occupation income for each degree

First, let's see that average income of every occupation in each degree. In these bar chart, we find several interesting fact that master can earn much money in management and PhD can earn much more than master and bachelor in education. However, it seems we still can't solve our two questions.
Then let's take the proportion of people's occupations into consideration.

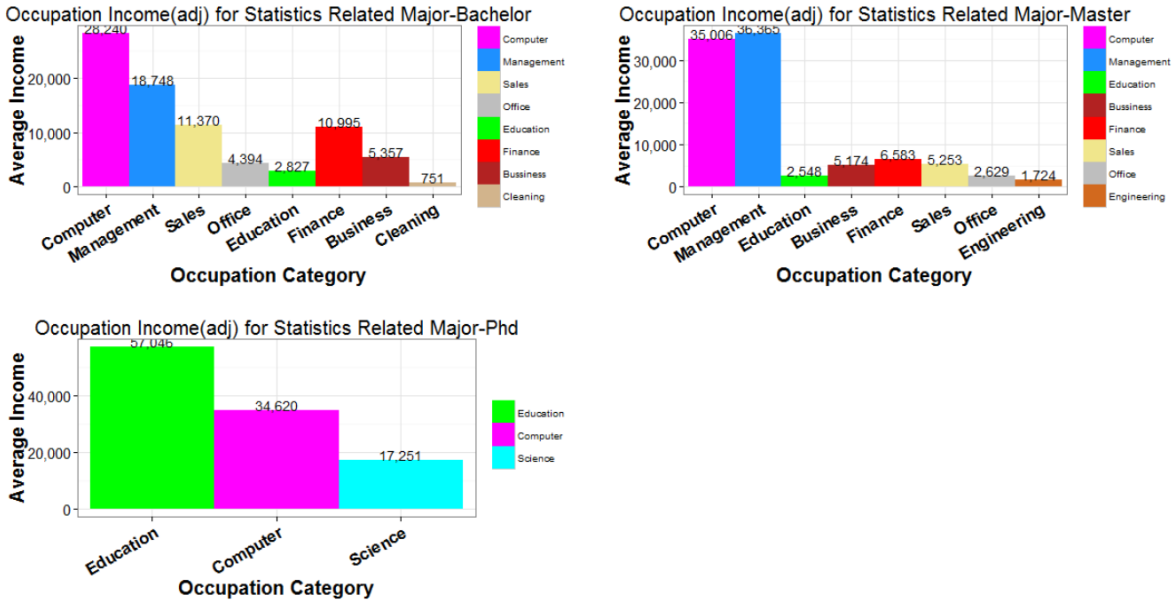


Figure 13: Adjusted occupation income for each degree

Here the way that we take occupation proportion into consideration:

$$Occupation_{income} * Occupation_{proportion}$$

From the adjusted occupation income, we find that education in PhD do have highest adjusted income over all occupation in each degree.

So the reason why PhD guys can earn more than master and bachelor is that education maybe doesn't pay the highest salary among all the occupations, However, since almost half of PhD go to education and the majority of master and bachelor students go to computer science occupations with relatively low income, PhD will earn more than masters and bachelors.

7.2 Tips: PhD or Master(Bachelor)?

If you want to go to education industry and also live a comfortable life, it's a good choice for you to have a PhD degree.
if you think master or bachelor degree is enough for you, you can go to management industry and still have a high income, but which means you get to practice you soft skill, like communication and leadership.

8 Conclusion

The U.S. Census dataset contains a wide spectrum of information that could be used for a variety of different analysis. To limit ourselves and allow for deeper insights we chose to explore only data science graduates. We have shown that English proficiency is positively correlated with income, that there is strong evidence of gender bias in the salaries of data science graduates, and that the value of each data science degree depends on the specific occupation one enters. However, it is still possible to derive further insights and connections from the data. In the future, we would like to explore other attributes in the life of the data scientist such as housing data, marital status, and family life.

9 Appendix

```
################################################################################
##
##   2013 - USA CENSUS
##   kaggle: https://www.kaggle.com/c/2013-american-community-survey
##
##   Download the four data files from Kaggle and save to your working directory
##
################################################################################

library(data.table)
library(sqldf)

## Create database and connection
dcon <- dbConnect(SQLite(), dbname="census.sqlite")

## read the first table into memory, write to database, remove from memory
houseA <- fread("ss13husa.csv")
dbWriteTable(conn = dcon, name = "houseA", value = houseA, row.names = FALSE)
rm(houseA)

## read the 2nd table into memory, write to database, remove from memory
houseB <- fread("ss13husb.csv")
dbWriteTable(conn = dcon, name = "houseB", value = houseB, row.names = FALSE)
rm(houseB)

## read the 3rd table into memory, write to database, remove from memory
peopleA <- fread("ss13pusa.csv")   ### possible error for column 126 #####
dbWriteTable(conn = dcon, name = "peopleA", value = peopleA, row.names = FALSE)
rm(peopleA)

## read the 4th table into memory, write to database, remove from memory
peopleB <- fread("ss13pusb.csv")
dbWriteTable(conn = dcon, name = "peopleB", value = peopleB, row.names = FALSE)
rm(peopleB)
```

```
## UNION houseA and houseB
dbSendQuery(dcon,"
            CREATE TABLE house AS
            SELECT*
            FROM houseA
            UNION ALL
            SELECT*
            FROM houseB;
            ")

## UNION peopleA and peopleB
dbSendQuery(dcon,"
            CREATE TABLE people AS
            SELECT*
            FROM peopleA
            UNION ALL
            SELECT*
            FROM peopleB;
            ")

## Removing the old tables
dbRemoveTable(dcon, "houseA")
dbRemoveTable(dcon, "houseB")
dbRemoveTable(dcon, "peopleA")
dbRemoveTable(dcon, "peopleB")

## create table with state codes
states<-read.table("statescode.txt",sep=".")
colnames(states)<-c("NO","region")
dbWriteTable(conn = dcon, name = "states", value = states, row.names = FALSE,append=FALSE,overwrite=TRUE)
rm(states)

## Close Connection
dbDisconnect(dcon)

#################################################################################################
##
##  THE LIFE OF A DATA SCIENTIST
##
##  Selecting All of the Statistics and Applied Math majors
##          - FOD1P (Field of First Degree) AND FOD2P (2nd Degree)
#################################################################################################

library(sqldf)
library(ggplot2)
library(maps)
library(stringr)
library(plyr)
library(dplyr)
library(grid)
library(scales)
library(choroplethr)
library(choroplethrMaps)

dcon <- dbConnect(SQLite(), dbname="census.sqlite")

res <- dbSendQuery(conn=dcon, "
                   SELECT SCHL, ST, AGEP, CIT, SEX, MAR, MARHT, MARHW, FOD1P, FOD2P, PINCP, SFR, WKHP, WAOB, OCC
```

```
                    FROM people
                    WHERE FOD1P  IN ('3702' , '6212', '3701', '4005', '3700', '2101', '2102')
                    OR FOD2P IN ('3702' , '6212', '3701', '4005', '3700', '2101', '2102');
                    ")
data <- fetch(res,-1)
dbClearResult(res)

#####################################################################
## preparing the data
#####################################################################

## coding gender
data$SEX <- factor(data$SEX, levels = c(1,2), labels = c("Male", "Female"))

## coding citizenship
data$CIT <- factor(data$CIT, levels = c(1,2,3,4,5), labels = c("Born in the U.S.",
                                                               "Born in U.S. Territories",
                                                               "Born abroad of American parent(s)",
                                                               "U.S. citizen by naturalization",
                                                               "Not a citizen of the U.S."))
## coding Where were you born
data$WAOB <- factor(data$WAOB, levels = c(1,2,3,4,5,6,7,8), labels = c("US state",
                                                                       "PR and US Island Areas",
                                                                       "Latin America",
                                                                       "Asia",
                                                                       "Europe",
                                                                       "Africa",
                                                                       "Northern America",
                                                                       "Oceania and at Sea"))



## Coding Occupation for all majors
occupation_code <- fread("occupation_code.txt")
occupation_code$OCCP = as.numeric(as.character(occupation_code$OCCP))
occupation_code$Occupation_Name<-str_extract(occupation_code$Occupation_Name,".{3}-")
occupation_code$Occupation_Name<-substr(occupation_code$Occupation_Name,1,nchar(occupation_code$Occupation_Name)


#####################################################################
#
# Creating a map of the proprtion of Data Science Degrees
#
#####################################################################

res <- dbSendQuery(conn=dcon, "
SELECT a.SERIALNO, a.ST,b.region,a.AGEP,a.CIT,a.ENG,a.LANX,a.MAR,a.MARHT,a.ESR,a.SCHL,a.SEX,a.FOD1P,a.PINCP,a.WK
FROM PROJECT AS a
INNER JOIN states AS b
ON a.ST = b.NO
ORDER BY b.region ASC
")
DATA<-fetch(res,-1)
dbClearResult(res)

DATA$STATEABBRE<-str_extract(DATA$region, "/.*")
DATA$STATEABBRE<-substr(DATA$STATEABBRE,2,nchar(DATA$STATEABBRE))
DATA$region<-substr(tolower(DATA$region),1,nchar(tolower(DATA$region))-3)
```

```
######################################################################
##
##  Using ggplot to create graphical summaries of the data
##
######################################################################

degree_major_income <- tbl_df(data)
save(degree_major_income, file="degree_major_income.RData")
################################################
## Add degree and major name to the table
################################################
load("degree_major_income.RData")
major_name <- "FOD1P,major
3700,Math
3701,Apllied Math
3702,Statistics
4005,Computer Science
2102,Computer Science"
major_name <- fread(major_name)

degree_name <- "SCHL, degree
21, Bachelor
22, Master
24, Phd"
degree_name <- fread(degree_name)

degree_major_income <- left_join(degree_major_income, major_name, by = c("FOD1P" = "FOD1P"))
degree_major_income <- left_join(degree_major_income, degree_name, by = c("SCHL" = "SCHL"))

################################################
#Data Science related major proportion in each state
################################################
query <- "Select FOD1P, ST from people;"
res <- dbSendQuery(dcon, query)
data_science_proportion <- dbFetch(res, -1)
dbClearResult(res)

data_science_pro_plus <- full_join(data_science_proportion, state_code, by = c("ST" = "ST"))
data_science_pro_specific <- data_science_pro_plus %>%
  na.omit() %>%
  filter(FOD1P %in% c(3702, 3700, 3701,4005,2102, 6212, 2101)) %>%
  group_by(region) %>%
  summarise(count = n())

data_science_pro_percent <- data_science_pro_plus %>%
  na.omit() %>%
  group_by(region) %>%
  summarise(count = n()) %>%
  mutate(value = data_science_pro_specific$count/count)

state_choropleth(data_science_pro_percent, title = "", num_colors=9)

## PIE CHART BIRTH
b <- data[data$WAOB != "PR and US Island Areas",]
ggplot(data = b) + aes(x = factor(1), fill = WAOB) +
  geom_bar(width = 1) +
  coord_polar(theta = "y", start = 0) +
```

```
  theme(axis.text.x=element_blank(), axis.title = element_text(size = 20, face = "bold"),
        plot.title = element_text(size = 55, face = "bold"),
        legend.text = element_text(size = 55),
        legend.title=element_text(size = 46, face = "bold"),
        legend.position = "left") +
  xlab("") +
  ylab("") +
  guides(fill = guide_legend(override.aes=list(size=25))) +
  labs(fill="") +
  ggtitle("Birthplace of all\nData Science Graduates") +
  scale_fill_brewer(palette="Dark2")
rm(b)

## total income by working hour and Location of BIRTH
dsmall <- data[,c("PINCP", "AGEP", "WAOB", "WKHP")]
dsmaller <- dsmall[dsmall$WAOB %in% c("US state", "Latin America", "Asia", "Europe"),]
ggplot(dsmaller)+
  aes(y=PINCP, x = WKHP, colour = WAOB) +
  geom_smooth(se = F, size = 1.5) +
  xlab("Hours Worked") +
  ylab("Total Income") +
  ggtitle("Total Income Over Hours Worked by Location of Birth") +
  scale_colour_discrete(name="Location of\nBirth") +
  scale_y_continuous(labels = comma) +
  theme(axis.text.x = element_text(size = 18, face = "bold", color = "black"),
        axis.text.y = element_text(size = 18, face = "bold", color = "black"),
        axis.title.x = element_text(size = 18, face = "bold"),
        axis.title.y = element_text(size = 18, face = "bold"),
        legend.text = element_text(size = 18), legend.title=element_text(size = 18, face = "bold"),
        plot.title = element_text(size = 36, face = "bold")) +
  scale_colour_brewer(palette="Dark2")
rm(dsmall, dsmaller)
## Earning per hour by English proficiency and degrees
ggplot(subset(data,data$ESR==1),aes(x=SCHL,y=PINCP/(WKHP*48),fill=factor(ENG,labels=c("Native","Very Well","Well
  geom_boxplot()+
  theme(legend.text=element_text(size=24,face="bold"),
        axis.text=element_text(size=20,face="bold",colour="black"),
        axis.title=element_text(size=20,face="bold",colour="black"))+
  theme(legend.title=element_blank())+
  scale_x_discrete(labels=c("21" = "Bachelor", "22" = "Master",
                            "23" = "Professional","24" = "PhD"))+
  xlab("")+ylab("Earning per Hour")+
  ylim(c(0,150))+labs(title="Earning per hour by English proficiency and degrees")+
  theme(plot.title = element_text(size = 30,colour="black",face='bold'),
        legend.position="bottom")
## Proportion of Top 5 Popular Occupations by English Proficiency
EngOcc<-ddply(subset(data,data$ESR==1), c("ENG","OCCPABBRE"),summarise,sum.n = length(region))
EngOcc1<-ddply(EngOcc,c("ENG"),summarise,n=sum(sum.n))
EngOcc<-merge(EngOcc,EngOcc1,by="ENG")
EngOcc$Prop<-EngOcc$sum.n/EngOcc$n
top5occ<-ddply(EngOcc,'ENG',function(dat)dat[order(dat$Prop,decreasing=TRUE)[1:5],])
top5occ<-top5occ[,-c(3,4)]
Other<-ddply(top5occ,'ENG',summarise,Prop=1-sum(Prop))
Other$OCCPABBRE<-c("Other")
Other<-Other[,c(1,3,2)]
top5occ<-rbind(top5occ,Other)
ggplot(top5occ, aes(x=ENG, y=Prop, fill=factor(OCCPABBRE,labels=c("Computer","Construction","Chef","Education","
  geom_bar(stat="identity")+
```

```
  theme(legend.text=element_text(size=20,face="bold"),
        axis.title=element_text(size=10,face="bold"))+
  scale_x_discrete(labels=c("0" = "Native", "1" = "Very Well",
                            "2" = "Well","3" = "Not Well","4"="Not at All"))+
  xlab("")+ylab("")+theme(legend.title=element_blank())+
  labs(title="Proportion of Top 5 Popular Occupations by English Proficiency")+
  theme(axis.text.x = element_text(face='bold',size=18, color = "black"),
        axis.text.y = element_text(size=20, color = "black",face="bold"),
        plot.title = element_text(size =22,colour="black",face='bold'))+
  scale_fill_brewer(palette="Set3")
##Earning per Hour by Occupations
ggplot(subset(data,data$ESR==1&data$OCCPABBRE %in% unique(top5occ$OCCPABBRE)),aes(x=OCCPABBRE,y=PINCP/(WKHP*48),
  geom_boxplot()+
  theme(legend.text=element_text(size=15),
        axis.title=element_text(size=20,face="bold"))+theme(legend.position='none',axis.text.x=element_text(angl
  xlab("")+ylab("Earning per Hour")+ylim(c(0,120))+
  scale_x_discrete(labels=c("EAT" = "Chef","CON"="Construction" ,"CMM" = "Computer",
                            "EDU" = "Education","PRD" = "Production","MGR"="Manager",
                            "OFF"="Office Employee","SAL"="Sales","TRN"="Transportation"))+
  theme(axis.text.x = element_text(face='bold',size=18, color = "black"),
        axis.text.y = element_text(size=20, color = "black",face="bold"),
        plot.title = element_text(size = 30,colour="black",face='bold'))+
  labs(title="Earning per Hour by Occupations")
## WORKING HOURS and Income by Degree
ggplot(data)+
  aes(y=PINCP, x = WKHP, colour = as.factor(SCHL)) +
  geom_smooth(se = F, size = 1.5) +
  xlab("Working Hours") +
  ylab("Total Income") +
  scale_y_continuous(labels = comma) +
  ggtitle("Total Income over Working Hours by Degree") +
  scale_colour_discrete(name="Degree",
                        breaks=c("20", "21", "22", "23", "24"),
                        labels=c("Associates degree", "Bachelor's", "Master's", "Professional", "Doctorate")) +
  theme(axis.text.x = element_text(size = 18, face = "bold", color = "black"),
        axis.text.y = element_text(size = 18, face = "bold", color = "black"),
        axis.title.x = element_text(size = 18, face = "bold"),
        axis.title.y = element_text(size = 18, face = "bold"),
        legend.text = element_text(size = 20), legend.title=element_text(size = 20, face = "bold"),
        plot.title = element_text(size = 24, face = "bold"))

## PIE CHART GENDER
ggplot(data = data) + aes(x = factor(1), fill = SEX) +
  geom_bar(width = 1) +
  coord_polar(theta = "y", start = 0) +
  theme(axis.text.x=element_blank(), axis.title = element_text(size = 20, face = "bold"),
        plot.title = element_text(size = 38, face = "bold"),
        legend.text = element_text(size = 38),
        legend.title=element_text(size = 32, face = "bold"),
        legend.position = "bottom") +
  xlab("") +
  ylab("") +
  labs(fill="") +
  ggtitle("Representation of Gender\nfor All Data Science Graduates")

## Total Income by Working Hours per Week MF
ggplot(data)+
  aes(y=PINCP, x = WKHP, colour = SEX) +
```

```
  geom_smooth(se = F, size = 1.5) +
  xlab("Working Hours per Week") +
  ylab("Total Income") +
  ggtitle("Total Income by Working Hours per Week") +
  scale_colour_discrete(name = "") +
  theme(axis.text.x = element_text(size = 24, face = "bold", color = "black"),
        axis.text.y = element_text(size = 24, face = "bold", color = "black"),
        axis.title.x = element_text(size = 24, face = "bold"),
        axis.title.y = element_text(size = 24, face = "bold"),
        legend.text = element_text(size = 34), legend.title=element_text(size = 40, face = "bold"),
        plot.title = element_text(size = 30, face = "bold"),
        legend.position="bottom")

## Total Income by Working Hours per Week
ggplot(data)+
  aes(y=PINCP, x = WKHP, colour = SEX) +
  geom_smooth(se = F, size = 1.5) +
  xlab("Working Hours per Week") +
  ylab("Total Income") +
  ggtitle("Total Income by Working Hours per Week") +
  scale_colour_discrete(name = "") +
  scale_y_continuous(labels = comma) +
  theme(axis.text.x = element_text(size = 18, face = "bold", color = "black"),
        axis.text.y = element_text(size = 18, face = "bold", color = "black"),
        axis.title.x = element_text(size = 18, face = "bold"),
        axis.title.y = element_text(size = 18, face = "bold"),
        legend.text = element_text(size = 18), legend.title=element_text(size = 18, face = "bold"),
        plot.title = element_text(size = 20, face = "bold"),
        legend.position="bottom")

## DENSITY OF oCCUPATION BY SEX
normHours <- data[data$WKHP >= 30 & data$WKHP <= 70,]
ggplot(data = normHours) +
  aes(OCCP, fill = SEX) +
  geom_density(alpha = I(1/2)) +
  ggtitle("Density of Occupation by Gender") +
  scale_fill_discrete(name = "") +
  ylab("Density") +
  xlab("Occupation Code") +
  theme(axis.text.x=element_blank(),
        axis.text.y = element_text(size = 20, face = "bold", color = "black"),
        axis.title.y = element_text(size = 20, face = "bold"),
        axis.title.x = element_blank(),
        legend.text = element_text(size = 30),
        legend.title=element_text(size = 30, face = "bold"),
        plot.title = element_text(size = 30, face = "bold"))
rm(normHours)

################################################
#Occupation for Statistics Major
################################################
degree_major_income <- left_join(degree_major_income, occupation_code, by = c("OCCP" = "OCCP"))


stat_occupation <- degree_major_income %>%
  filter(FOD1P == 3702) %>%
  mutate(occupation_cat = substr(Occupation_Name, 2,4))
```

```r
stat_occupation_percent <- stat_occupation %>%
  group_by(SCHL, occupation_cat) %>%
  tally  %>%
  group_by(SCHL) %>%
  mutate( percent = (100*n)/sum(n))

###########
# Bachelor Major Occupation
###########
stat_occupation_bach_percent13 <- stat_occupation_percent %>%
  filter(SCHL == 21, percent > 1, occupation_cat != "N/A")

stat_occupation_bach_percent8 <- stat_occupation_percent %>%
  filter(SCHL == 21, percent > 2, occupation_cat != "N/A")


percent_str_bach <- paste(round(stat_occupation_bach_percent8$percent / sum(stat_occupation_bach_percent8$percen
stat_occupation_bach_percent8$percent_str_bach <- percent_str_bach
stat_occupation_bach_percent8 <- stat_occupation_bach_percent8[c(3,6,8,7,4,5,1,2),]

occupation_stat_bach <- ggplot(stat_occupation_bach_percent8) +
  aes(x = "", y = percent, fill = occupation_cat) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar(theta = "y",  start = - pi / 3) +
  scale_fill_manual(name = "Occupation",
                    values =  c("CMM" = "magenta",
                                "MGR" = "dodgerblue",
                                "OFF" = "gray",
                                "SAL" = "khaki",
                                "BUS" = "firebrick",
                                "FIN" = "red",
                                "EDU" = "green",
                                "CLN" = "tan"),
                    labels=c("Bussiness", "Cleaning",
                             "Computer", "Education",
                             "Finance", "Management",
                             "Office", "Sales")) +
  geom_text(aes(x = 1.2, y = percent/2 + c(0, cumsum(percent)[-length(percent)]),
                label = occupation_cat),size = 6) +
  ggtitle("Occupation Proportion for Statistics-Bachelor") +
  ylab("") +
  xlab("") +
  theme(panel.background = element_rect(fill = 'white' ),
        plot.title = element_text(size = 40,colour="black"),
        axis.text = element_blank(),
        axis.ticks = element_blank(),
        panel.grid  = element_blank(),
        legend.key.size = unit(2, "cm"),
        legend.text = element_text(size = 20),
        legend.title = element_text(size = 13))


###########
# Master Major Occupation
###########
stat_occupation_mast_percent11 <- stat_occupation_percent %>%
  filter(SCHL == 22, percent > 1, occupation_cat != "N/A")
```

```r
stat_occupation_mast_percent7 <- stat_occupation_percent %>%
  filter(SCHL == 22, percent > 2, occupation_cat != "N/A")


percent_str_mast <- paste(round(stat_occupation_mast_percent7$percent / sum(stat_occupation_mast_percent7$percen
stat_occupation_mast_percent7$percent_str_mast <- percent_str_mast
stat_occupation_mast_percent7 <- stat_occupation_mast_percent7[c(2,6,3,1,5,8,7,4),]

occupation_stat_mast <- ggplot(stat_occupation_mast_percent7) +
  aes(x = "", y = percent, fill = occupation_cat) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar(theta = "y",  start = - pi / 3) +
  scale_fill_manual(name = "Occupation",
                    values =  c("CMM" = "magenta",
                                "MGR" = "dodgerblue",
                                "OFF" = "gray",
                                "SAL" = "khaki",
                                "BUS" = "firebrick",
                                "FIN" = "red",
                                "EDU" = "green",
                                "ENG" = "chocolate"),
                    labels=c("Bussiness", "Computer",
                             "Education", "Engineering",
                             "Finance", "Management",
                             "Office", "Sales")) +
  geom_text(aes(x = 1.2, y = percent/2 + c(0, cumsum(percent)[-length(percent)]),
                label = occupation_cat),size = 6) +
  ggtitle("Occupation Proportion for Statistics-Master") +
  ylab("") +
  xlab("") +
  theme(panel.background = element_rect(fill = 'white' ),
        plot.title = element_text(size = 40,colour="black"),
        axis.text = element_blank(),
        axis.ticks = element_blank(),
        panel.grid  = element_blank(),
        legend.key.size = unit(2, "cm"),
        legend.text = element_text(size = 20),
        legend.title = element_text(size = 13))
###########
# Phd Major Occupation
###########
stat_occupation_phd_percent12 <- stat_occupation_percent %>%
  filter(SCHL == 24, percent > 1, occupation_cat != "N/A")

stat_occupation_phd_percent8 <- stat_occupation_percent %>%
  filter(SCHL == 24, percent > 2, occupation_cat != "N/A")

percent_str_phd <- paste(round(stat_occupation_phd_percent8$percent / sum(stat_occupation_phd_percent8$percent)
stat_occupation_phd_percent8$percent_str_phd <- percent_str_phd
stat_occupation_phd_percent8 <- stat_occupation_phd_percent8[c(2,1,9,6,3,4,5,7,8),]

occupation_stat_phd <- ggplot(stat_occupation_phd_percent8) +
  aes(x = "", y = percent, fill = occupation_cat) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar(theta = "y",  start = - pi / 3) +
  scale_fill_manual(name = "Occupation",
                    values =  c("CMM" = "magenta",
                                "MGR" = "dodgerblue",
```

```
                                "OFF" = "gray",
                                "SAL" = "khaki",
                                "EDU" = "green",
                                "ENG" = "chocolate",
                                "MED" = "purple1",
                                "SCI" = "cyan",
                                "ENT" = "yellow"),
                     labels=c("Computer","Education",
                              "Engineering","Entertainment",
                              "Medical", "Management",
                              "Office", "Sales", "Science")) +
  geom_text(aes(x = 1.2, y = percent/2 + c(0, cumsum(percent)[-length(percent)]),
                label = occupation_cat),size = 6) +
  ggtitle("Occupation Proportion for Statistics-Phd") +
  ylab("") +
  xlab("") +
  theme(panel.background = element_rect(fill = 'white' ),
        plot.title = element_text(size = 40,colour="black"),
        axis.text = element_blank(),
        axis.ticks = element_blank(),
        panel.grid  = element_blank(),
        legend.key.size = unit(2, "cm"),
        legend.text = element_text(size = 20),
        legend.title = element_text(size = 13))


################
#(Mean)Occupation Income for Statistics Related Major - Bachelor
################
stat_occupation_bach <- stat_occupation %>%
  filter(SCHL == 21)
stat_occupation_bach_income <- left_join(stat_occupation_bach_percent8, stat_occupation_bach, by = c("occupation

stat_occupation_bach_income_field <- stat_occupation_bach_income %>%
  group_by(occupation_cat) %>%
  summarise(average_field = round(mean(PINCP)))

stat_occupation_bach_income_field <- stat_occupation_bach_income_field[c(3,6,8,7,4,5,1,2),]

stat_occupation_bach_income_field$occupation_cat <-factor(stat_occupation_bach_income_field$occupation_cat, leve

occupation_Income_stat_bach <- ggplot(stat_occupation_bach_income_field) +
  aes(x = occupation_cat, y = average_field, fill = occupation_cat, label = comma(average_field)) +
  geom_bar(width = 1, stat = "identity") +
  geom_text(vjust = 0) +
  scale_fill_manual(name = "Occupation",
                    values =  c("CMM" = "magenta",
                                "MGR" = "dodgerblue",
                                "OFF" = "gray",
                                "SAL" = "khaki",
                                "BUS" = "firebrick",
                                "FIN" = "red",
                                "EDU" = "green",
                                "CLN" = "tan"),
                    labels = c("Computer","Management",
                               "Sales","Office",
                               "Education","Finance",
                               "Bussiness", "Cleaning")) +
```

```
  ggtitle("Occupation Income for Statistics Related Major-Bachelor") +
  scale_y_continuous(labels = comma) +
  ylab("Average Salary") +
  xlab("Occupation Category") +
  scale_x_discrete("Occupation Category",
                   labels = c("EDU" = "Education","CMM" = "Computer",
                              "MGR" = "Management","SCI" = "Science", "OFF" = "Office",
                              "SAL" = "Sales", "BUS" = "Business", "FIN" = "Finance",
                              "ENG" = "Engineering", "CLN" = "Cleaning")) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 30, hjust = 1, face='bold',size=17, color = "black"),
        axis.text.y = element_text(hjust = 1,size=15, color = "black"),
        axis.title.x = element_text(face='bold',size=20, color = "black"),
        axis.title.y = element_text(face='bold',size=20, color = "black"),
        panel.background = element_rect(fill = 'white'),
        plot.title = element_text(size = 20,colour="black"),
        legend.background = element_rect(fill = 'white'))
#################
#(Mean)Bachelor Income Adjusted
#################
top_stat_percent_bach <- round(stat_occupation_bach_percent8$percent / sum(stat_occupation_bach_percent8$percent
stat_occupation_bach_percent8$top_stat_percent_bach <- top_stat_percent_bach

stat_occupation_bach <- stat_occupation %>%
  filter(SCHL == 21)

stat_occupation_bach_income <- left_join(stat_occupation_bach_percent8, stat_occupation_bach, by = c("occupation

stat_occupation_bach_income_field <- stat_occupation_bach_income %>%
  group_by(occupation_cat) %>%
  summarise(average_field = round(mean(PINCP)))

stat_occupation_bach_income_field <- left_join(stat_occupation_bach_income_field,  stat_occupation_bach_percent8
stat_occupation_bach_income_field <- stat_occupation_bach_income_field %>%
  mutate(adj_income = round(average_field * top_stat_percent_bach, 0))


stat_occupation_bach_income_field <- stat_occupation_bach_income_field[c(3,6,8,7,4,5,1,2),]

stat_occupation_bach_income_field$occupation_cat <-factor(stat_occupation_bach_income_field$occupation_cat, leve

occupation_Income_stat_bach_adj <- ggplot(stat_occupation_bach_income_field) +
  aes(x = occupation_cat, y = adj_income, fill = occupation_cat, label = comma(adj_income)) +
  geom_bar(width = 1, stat = "identity") +
  geom_text(vjust=0) +
  scale_fill_manual(name = "Occupation",
                    values =  c("CMM" = "magenta",
                                "MGR" = "dodgerblue",
                                "OFF" = "gray",
                                "SAL" = "khaki",
                                "BUS" = "firebrick",
                                "FIN" = "red",
                                "EDU" = "green",
                                "CLN" = "tan"),
                    labels = c("Computer","Management",
                               "Sales","Office",
                               "Education","Finance",
                               "Bussiness", "Cleaning")) +
```

```
  scale_y_continuous(labels = comma) +
  ylab("Adjusted Average Income") +
  ggtitle("Occupation Income(adj) for Statistics Related Major-Bachelor") +
  scale_x_discrete("Occupation Category", labels = c("EDU" = "Education","CMM" = "Computer",
                                                     "MGR" = "Management","SCI" = "Science", "OFF" = "Office",
                                                     "SAL" = "Sales", "BUS" = "Business", "FIN" = "Finance",
                                                     "ENG" = "Engineering", "CLN" = "Cleaning")) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 30, hjust = 1, face='bold',size=17, color = "black"),
        axis.text.y = element_text(hjust = 1,size=15, color = "black"),
        axis.title.x = element_text(face='bold',size=20, color = "black"),
        axis.title.y = element_text(face='bold',size=20, color = "black"),
        panel.background = element_rect(fill = 'white' ),
        plot.title = element_text(size = 20,colour="black"),
        legend.background = element_rect(fill = 'white'))

################
#(Mean)Occupation Income for Statistics Related Major - Master
################
stat_occupation_mast <- stat_occupation %>%
  filter(SCHL == 22)

stat_occupation_mast_income <- left_join(stat_occupation_mast_percent7, stat_occupation_mast, by = c("occupation

stat_occupation_mast_income_field <- stat_occupation_mast_income %>%
  group_by(occupation_cat) %>%
  summarise(average_field = round(mean(PINCP)))

stat_occupation_mast_income_field <- left_join(stat_occupation_mast_income_field,  stat_occupation_mast_percent8
stat_occupation_mast_income_field <- stat_occupation_mast_income_field %>%
  mutate(adj_income = round(average_field * top_stat_percent_mast, 0))

stat_occupation_mast_income_field <- stat_occupation_mast_income_field[c(2,6,3,1,5,8,7,4),]

stat_occupation_mast_income_field$occupation_cat <-factor(stat_occupation_mast_income_field$occupation_cat, leve

occupation_Income_stat_mast <- ggplot(stat_occupation_mast_income_field) +
  aes(x = occupation_cat, y = average_field, fill = occupation_cat, label = comma(average_field)) +
  geom_bar(width = 1, stat = "identity") +
  geom_text(vjust = 0) +
  scale_fill_manual(name = "Occupation",
                    values =  c("CMM" = "magenta",
                                "MGR" = "dodgerblue",
                                "OFF" = "gray",
                                "SAL" = "khaki",
                                "BUS" = "firebrick",
                                "FIN" = "red",
                                "EDU" = "green",
                                "ENG" = "chocolate"),
                    labels=c("Computer","Management",
                             "Education","Bussiness",
                             "Finance","Sales" ,
                             "Office","Engineering")) +
  scale_y_continuous(labels = comma) +
  ylab("Average Income") +
  ggtitle("Occupation Income for Statistics Related Major-Master") +
  scale_x_discrete("Occupation Category", labels = c("EDU" = "Education","CMM" = "Computer",
                                                     "MGR" = "Management","SCI" = "Science", "OFF" = "Office",
```

```
                                                    "SAL" = "Sales", "BUS" = "Business", "FIN" = "Finance",
                                                    "ENG" = "Engineering")) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 30, hjust = 1, face='bold',size=17, color = "black"),
        axis.text.y = element_text(hjust = 1,size=15, color = "black"),
        axis.title.x = element_text(face='bold',size=20, color = "black"),
        axis.title.y = element_text(face='bold',size=20, color = "black"),
        panel.background = element_rect(fill = 'white' ),
        plot.title = element_text(size = 20,colour="black"),
        legend.background = element_rect(fill = 'white'))

################
#(Mean)Master Income Adjusted
################
top_stat_percent_mast <- round(stat_occupation_mast_percent7$percent / sum(stat_occupation_mast_percent7$percent
stat_occupation_mast_percent7$top_stat_percent_mast <- top_stat_percent_mast

stat_occupation_mast <- stat_occupation %>%
  filter(SCHL == 22)

stat_occupation_mast_income <- left_join(stat_occupation_mast_percent7, stat_occupation_mast, by = c("occupation

stat_occupation_mast_income_field <- stat_occupation_mast_income %>%
  group_by(occupation_cat) %>%
  summarise(average_field = round(mean(PINCP)))

stat_occupation_mast_income_field <- left_join(stat_occupation_mast_income_field,  stat_occupation_mast_percent7
stat_occupation_mast_income_field <- stat_occupation_mast_income_field %>%
  mutate(adj_income = round(average_field * top_stat_percent_mast, 0))


stat_occupation_mast_income_field <- stat_occupation_mast_income_field[c(2,6,3,1,5,8,7,4),]

stat_occupation_mast_income_field$occupation_cat <-factor(stat_occupation_mast_income_field$occupation_cat, leve

occupation_Income_stat_mast_adj <- ggplot(stat_occupation_mast_income_field) +
  aes(x = occupation_cat, y = adj_income, fill = occupation_cat, label = comma(adj_income)) +
  geom_bar(width = 1, stat = "identity") +
  geom_text(vjust=0) +
  scale_fill_manual(name = "Occupation",
                    values =  c("CMM" = "magenta",
                                "MGR" = "dodgerblue",
                                "OFF" = "gray",
                                "SAL" = "khaki",
                                "BUS" = "firebrick",
                                "FIN" = "red",
                                "EDU" = "green",
                                "ENG" = "chocolate"),
                    labels=c("Computer","Management",
                             "Education","Bussiness",
                             "Finance","Sales" ,
                             "Office","Engineering")) +
  scale_x_discrete("Occupation Category", labels = c("EDU" = "Education","CMM" = "Computer",
                                                     "MGR" = "Management","SCI" = "Science", "OFF" = "Office",
                                                     "SAL" = "Sales", "BUS" = "Business", "FIN" = "Finance",
                                                     "ENG" = "Engineering")) +
  ggtitle("Occupation Income(adj) for Statistics Related Major-Master") +
  theme_bw() +
```

```
  ylab("Adjusted Average Income") +
  scale_y_continuous(labels = comma) +
  theme(axis.text.x = element_text(angle = 30, hjust = 1, face='bold',size=17, color = "black"),
        axis.text.y = element_text(hjust = 1,size=15, color = "black"),
        axis.title.x = element_text(face='bold',size=20, color = "black"),
        axis.title.y = element_text(face='bold',size=20, color = "black"),
        panel.background = element_rect(fill = 'white' ),
        plot.title = element_text(size = 20,colour="black"),
        legend.background = element_rect(fill = 'white'))


################
#(Mean)Occupation Income for Statistics Related Major - Phd
################

#for the rest 4 occupations, each of them only have 2 sample,
# In order to get rid of the outliers, we only take the first three occupations
stat_occupation_phd_percent8

stat_occupation_phd_percent4 <- stat_occupation_phd_percent8 %>%
  filter(percent > 8)

stat_occupation_phd <- stat_occupation %>%
  filter(SCHL == 24)
stat_occupation_phd_income <- left_join(stat_occupation_phd_percent4, stat_occupation_phd, by = c("occupation_ca

stat_occupation_phd_income_field <- stat_occupation_phd_income %>%
  group_by(occupation_cat) %>%
  summarise(average_field = round(mean(PINCP)))
stat_occupation_phd_income_field <- stat_occupation_phd_income_field[c(2,1,3),]
#[c(2,1,9,6,3,4,5,7,8),]
stat_occupation_phd_income_field$occupation_cat <-factor(stat_occupation_phd_income_field$occupation_cat, levels

occupation_Income_stat_phd <- ggplot(stat_occupation_phd_income_field) +
  aes(x = occupation_cat, y = average_field, fill = occupation_cat, label = comma(average_field)) +
  geom_bar(width = 1, stat = "identity") +
  geom_text(vjust = 0) +
  scale_fill_manual(name = "Occupation",
                    values =  c("CMM" = "magenta",
                                "MGR" = "dodgerblue",
                                "OFF" = "gray",
                                "SAL" = "khaki",
                                "BUS" = "firebrick",
                                "FIN" = "red",
                                "EDU" = "green",
                                "ENG" = "chocolate",
                                "SCI" = "cyan"),
                    labels=c("Education","Computer",
                             "Science","Engineering",
                             "Management")) +
  ggtitle("Occupation Income for Statistics Related Major-Phd") +
  ylab("Average Income") +
  scale_y_continuous(labels = comma) +
  scale_x_discrete("Occupation Category", labels = c("EDU" = "Education","CMM" = "Computer",
                                                     "MGR" = "Management","SCI" = "Science", "OFF" = "Office",
                                                     "SAL" = "Sales", "BUS" = "Business", "FIN" = "Finance",
                                                     "ENG" = "Engineering")) +
  theme_bw() +
```

```
  theme(axis.text.x = element_text(angle = 30, hjust = 1, face='bold',size=17, color = "black"),
        axis.text.y = element_text(hjust = 1,size=15, color = "black"),
        axis.title.x = element_text(face='bold',size=20, color = "black"),
        axis.title.y = element_text(face='bold',size=20, color = "black"),
        panel.background = element_rect(fill = 'white' ),
        plot.title = element_text(size = 20, colour="black"),
        legend.background = element_rect(fill = 'white'))
################
#(Mean)Phd Income Adjusted
################
stat_occupation_phd_percent4 <- stat_occupation_phd_percent8 %>%
  filter(percent > 8)

top_stat_percent_phd <- round(stat_occupation_phd_percent4$percent / sum(stat_occupation_phd_percent4$percent),2
stat_occupation_phd_percent4$top_stat_percent_phd <- top_stat_percent_phd

stat_occupation_phd <- stat_occupation %>%
  filter(SCHL == 24)

stat_occupation_phd_income <- left_join(stat_occupation_phd_percent4, stat_occupation_phd, by = c("occupation_ca

stat_occupation_phd_income_field <- stat_occupation_phd_income %>%
  group_by(occupation_cat) %>%
  summarise(average_field = round(mean(PINCP)))

stat_occupation_phd_income_field <- left_join(stat_occupation_phd_income_field,  stat_occupation_phd_percent4, b
stat_occupation_phd_income_field <- stat_occupation_phd_income_field %>%
  mutate(adj_income = round(average_field * top_stat_percent_phd, 0))


stat_occupation_phd_income_field <- stat_occupation_phd_income_field[c(2,1,3),]

stat_occupation_phd_income_field$occupation_cat <-factor(stat_occupation_phd_income_field$occupation_cat, levels

occupation_Income_stat_phd_adj <- ggplot(stat_occupation_phd_income_field) +
  aes(x = occupation_cat, y = adj_income, fill = occupation_cat, label = comma(adj_income)) +
  geom_bar(width = 1, stat = "identity") +
  geom_text(vjust=0) +
  scale_fill_manual(name = "Occupation",
                    values =  c("CMM" = "magenta",
                                "MGR" = "dodgerblue",
                                "OFF" = "gray",
                                "SAL" = "khaki",
                                "BUS" = "firebrick",
                                "FIN" = "red",
                                "EDU" = "green",
                                "ENG" = "chocolate",
                                "SCI" = "cyan"),
                    labels=c("Education","Computer",
                             "Science","Engineering",
                             "Management")) +
  ggtitle("Occupation Income(adj) for Statistics Related Major-Phd") +
  ylab("Adjusted Average Income")+
  scale_y_continuous(labels = comma) +
  scale_x_discrete("Occupation Category", labels = c("EDU" = "Education","CMM" = "Computer",
                                                     "MGR" = "Management","SCI" = "Science", "OFF" = "Office",
                                                     "SAL" = "Sales", "BUS" = "Business", "FIN" = "Finance",
                                                     "ENG" = "Engineering")) +
```

```
  theme_bw() +
  theme(axis.text.x = element_text(angle = 30, hjust = 1, face='bold',size=17, color = "black"),
        axis.text.y = element_text(hjust = 1,size=15, color = "black"),
        axis.title.x = element_text(face='bold',size=20, color = "black"),
        axis.title.y = element_text(face='bold',size=20, color = "black"),
        panel.background = element_rect(fill = 'white' ),
        plot.title = element_text(size = 20,colour="black"),
        legend.background = element_rect(fill = 'white'))


#################
#(Mean)Occupation Income Multiple graph(Bachelor, Master, Phd) in one plot
#################
multiplot(occupation_Income_stat_bach, occupation_Income_stat_phd,
          occupation_Income_stat_mast, cols=2)

multiplot(occupation_Income_stat_bach_adj, occupation_Income_stat_phd_adj,
          occupation_Income_stat_mast_adj, cols=2)

occupation_Income_stat
```