Data Cleaning Draft for GitHub

Team 7

December 13th, 2018

INFM 600

From the dataset PGEEOCS (PG County Equal Employment Opportunity Computer Science), we want to show the total computer related job occupation shared between male and female. Then, we want to select the data that shows occupational category in Computer and Information System. The second picture shows the occupation shared between male and female in Computer and Information System related field.

## MoCo MD tuition Assistance dataset

1 step:

For the department:

which departments are hiring or paying for IT related jobs (1. Filter out majors to represent IT, 2, degree, certificate, non-degree, others, )

For Major:

CS major, C system management, IT, Professional/Technical)

For degree:

Certificate, non-degree, other

2 steps

In r, show us filering in to majors, and degrees

Run how many departments are hiring from major we selected and degree we selected

How many departments, and how many times each department were asked to attend these classes

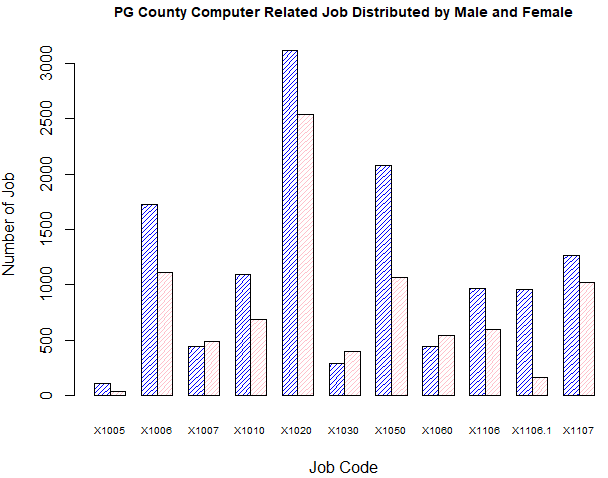
## MD Department of Planning, MD county Socialeconomic Characteristics

> PGEEOCS <- read.csv("C:/Users/dyang129/Desktop/INFM 600 Project/PGEEOCS.csv")

> View(PGEEOCS)

>

> > barplot(as.matrix(PGEEOCS),main = "PG County Computer Related Job Distributed by Male and Female", cex.main = 0.9, col = c("blue", "pink"), density = 30, ylab = "Number of Job", xlab = "Job Code", cex.names = 0.6, beside = T)



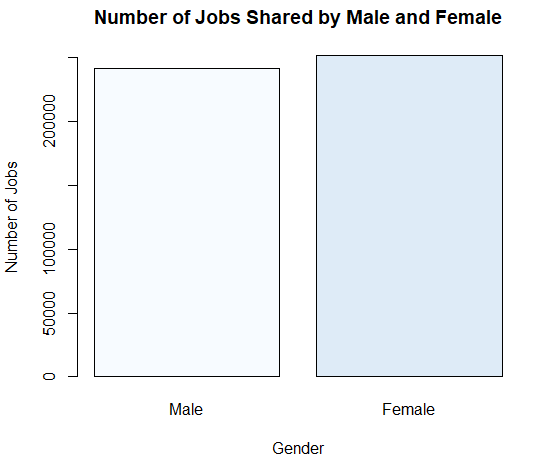
From the dataset PGEEO (PG County Equal Employment Opportunity), we want to show the total job occupation shared between male and female as a reference, to compare it with computer related job occupation. The select data shows total occupation in general. Blue bar represent number of jobs taken by male, and pink bar represent number of jobs taken by female.

setwd("/Volumes/DY JOB RELA/UM 2018 Classes/INFM 600")

> df <- read.table("PGEEO.csv")

> sharedJobs <- c(241625, 251445)

> barplot(sharedJobs, main = "Number of Jobs Shared by Male and Female", names.arg = c("Male", "Female"), beside = TRUE, col = blues9, xlab = "Gender", ylab = "Number of Jobs")



In Maryland Department of Planning, Maryland Counties Socioeconomic Characteristics dataset, we select a few columns that make help to illustrate the necessity of boot camp for lower income people and minority who do not have resources and would like to enhance their capability in job market.

This should probably just be a infographic about WOC in PG county. This will help us build a profile for the potential need in PG for WOC to do this Bootcamp classes.

Then we will compare those same factors (**Total Households, Employment Status of the Population 16 years and over, Employed, Unemployed, Unemployment Rate, Median Household Income ($), Percent Families in Poverty**) to the rest of the counties as a whole. This will help with our argument that PG is a great place to do this because of its the demorgaphics compared to all the rest of Maryland counties.

> mdSocioEc <- read.csv("/Volumes/DY JOB RELA/UM 2018 Classes/INFM 600/Team Project/USED in Plot/mdSocioEc.csv", header=TRUE)

> View(mdSocioEc)

# to calculate the mean for Median Household Income

> mean(mdSocioEc$MedianHouseholdIncome)

[1] 70159.42

> # to see relationship between HS diploma with Female

> x1 <- mdSocioEc$High.School.Diploma

> y1 <- mdSocioEc$Female

> plot(x1, y1, col="blue", main = "Female with HS Diploma Regression", abline(lm(y1~x1)), xlab = "HS Diploma", ylab = "Felame" )

> lm(formula = y1~x1)

Call:

lm(formula = y1 ~ x1)

Coefficients:

(Intercept) x1

-9234.675 3.198

> relation1 <- lm(formula = y1~x1)

> summary(relation1)

Call:

lm(formula = y1 ~ x1)

Residuals:

Min 1Q Median 3Q Max

-61180 -18678 -5205 -893 224991

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -9234.6748 15472.0194 -0.597 0.557

x1 3.1979 0.2475 12.918 9.53e-12 \*\*\*

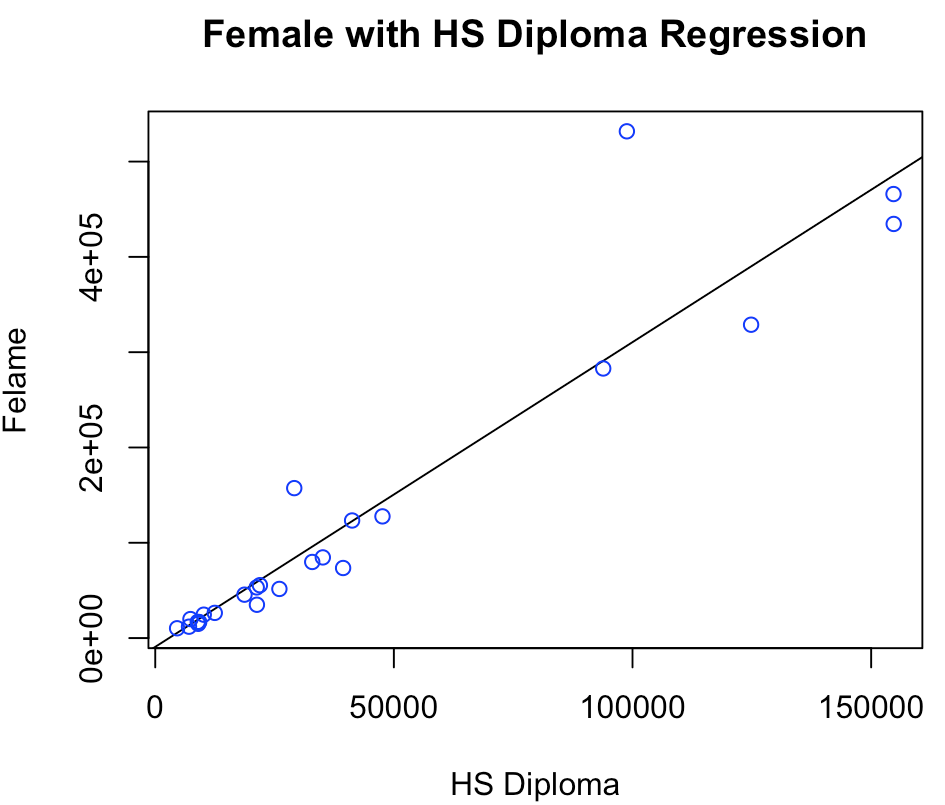
---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 55090 on 22 degrees of freedom

Multiple R-squared: 0.8835, Adjusted R-squared: 0.8782

F-statistic: 166.9 on 1 and 22 DF, p-value: 9.526e-12



> # here we want to see the relationship between unemployed and family in provity

> x2<- mdSocioEc$Unemployment.Rate

> y2 <- mdSocioEc$PercentFamiliesinPoverty

> relation2 <- lm(y2~x2)

> lm(formula = y2~x2)

Call:

lm(formula = y2 ~ x2)

Coefficients:

(Intercept) x2

-4.395 1.851

> summary(lm(y2~x2))

Call:

lm(formula = y2 ~ x2)

Residuals:

Min 1Q Median 3Q Max

-6.6392 -1.5311 -0.5254 0.8640 9.4505

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -4.3953 2.2947 -1.915 0.0685 .

x2 1.8506 0.3305 5.600 1.25e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

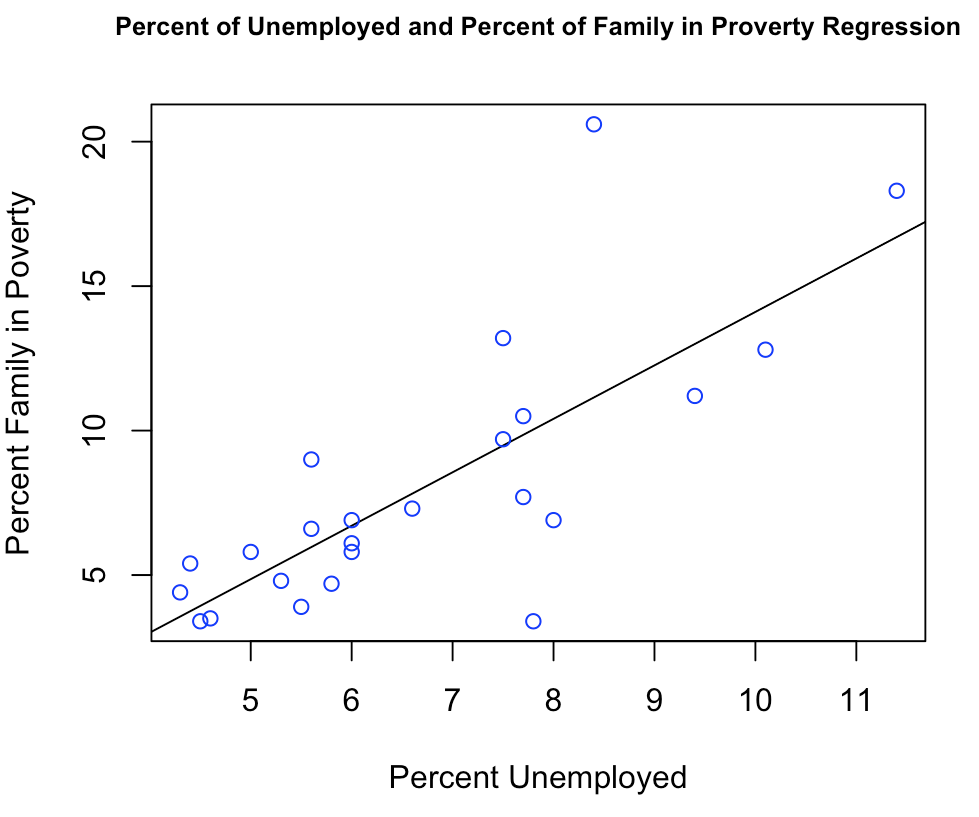
Residual standard error: 2.98 on 22 degrees of freedom

Multiple R-squared: 0.5877, Adjusted R-squared: 0.569

F-statistic: 31.36 on 1 and 22 DF, p-value: 1.249e-05

> plot(x2, y2, col="blue", main = "Percent of Unemployed and Percent of Family in Proverty Regression", abline(lm(y2~x2)),cex.main = 0.8, xlab = "Percent Unemployed", ylab= "Percent Family in Poverty")

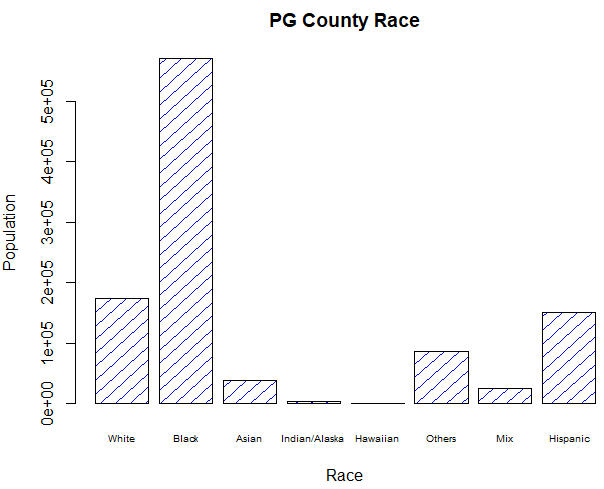
Since the P value is very small, so the X2 is very significant.



Race in PG County

> PGRace <- c(173881, 570138, 38063, 3449, 269, 86885, 25008, 150268)

> barplot(PGRace, main = "PG County Race", xlab = "Race", ylab = "Population", names.arg = c("White", "Black", "Asian", "Indian/Alaska", "Hawaiian", "Others", "Mix", "Hispanic"), col = "blue", density = 10, cex.names = 0.6)



MD IT Tuition Assistance

> MDITTuitionAssistance <- read.csv("E:/UM 2018 Classes/INFM 600/Team Project/USED in Plot/MDITTuitionAssistance.csv")

> View(MDITTuitionAssistance)

> CostAssistance <- as.numeric(MDITTuitionAssistance$Cost)

> mean(CostAssistance)

[1] 703.8842

> > MarylandTuitionAssistance <- read.csv("E:/UM 2018 Classes/INFM 600/Team Project/USED in Plot/MarylandTuitionAssistance.csv")

> View(MarylandTuitionAssistance)

> summary(MarylandTuitionAssistance)

Department Major

:1045760 :1045760

Correction & Rehabilitation: 2 Computer Science : 4

Health & Human Services : 2 Computer Systems Management: 3

Housing & Community Affairs: 1 Information Technology : 15

Police : 11 Professional/Technical : 2

Technology Services : 4

Transportation : 4

Degree Cost

:1045760 Min. : 0.0

Certificate: 13 1st Qu.: 335.3

Non-Degree : 4 Median : 412.0

Other : 7 Mean : 703.9

3rd Qu.: 984.0

Max. :2130.0

NA's :1045760

## For the State of Maryland, there are only 6 department offered 24 tuition assistantship for the computer science related major. Among these computer related training, 13 of them are for certificate, 4 of them are non-Degree, and 7 for all others. The mean of the tuition assistance is $703.90, which is very limited to employee career development. For people who live in PG county, there is highest number of minority live in the county and

> mdSocioEcMinSum <- read.csv("C:/Users/dyang129/Desktop/Team Project/USED in Plot/mdSocioEcMinSum.csv")

> View(mdSocioEcMinSum)

> summary(mdSocioEcMinSum)

Jurisdictions Total.Households MedianHouseholdIncome

Allegany County : 1 Min. : 7683 Min. : 35886

Anne Arundel County: 1 1st Qu.: 17459 1st Qu.: 52839

Baltimore city : 1 Median : 38074 Median : 68464

Baltimore County : 1 Mean : 90729 Mean : 70159

Calvert County : 1 3rd Qu.: 96841 3rd Qu.: 86873

Caroline County : 1 Max. :367764 Max. :113800

(Other) :18

Families PercentFamiliesinPoverty Male Female

Min. : 4654 Min. : 3.400 Min. : 9539 Min. : 10280

1st Qu.: 12622 1st Qu.: 4.775 1st Qu.: 22547 1st Qu.: 23405

Median : 27000 Median : 6.750 Median : 52890 Median : 54331

Mean : 60665 Mean : 7.996 Mean :120281 Mean :128049

3rd Qu.: 70418 3rd Qu.: 9.900 3rd Qu.:129332 3rd Qu.:135112

Max. :256128 Max. :20.600 Max. :494612 Max. :531759

WhiteAlone totalMinority

Min. : 13719 Min. : 1136

1st Qu.: 39742 1st Qu.: 10349

Median : 80771 Median : 23944

Mean :142010 Mean :129242

3rd Qu.:184673 3rd Qu.:102525

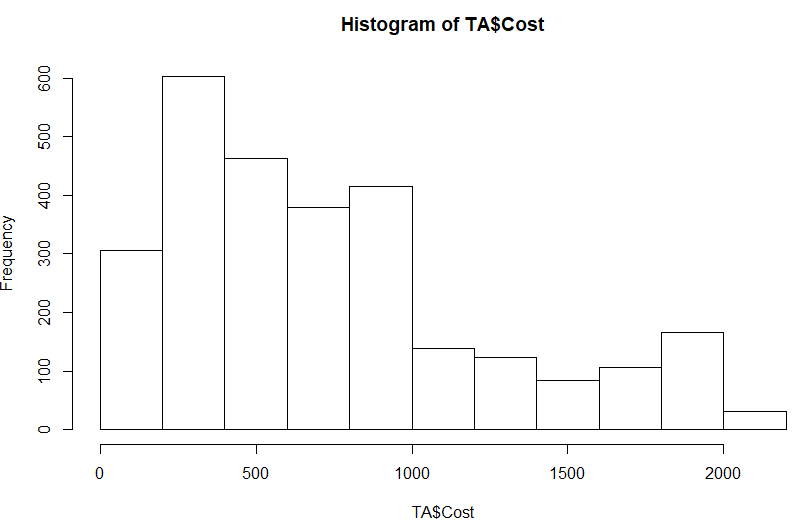
Max. :566239 Max. :874080

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

hist(TA$Cost)

This gives us the below histogram which shows that if cost of a major is higher then fewer people opt for it and vice versa.

This is a right-skewed distribution.



We use unique(TA$Major) to get infer that there are 79 different majors people are opting for

aggregate(TA$Cost, list(Major = TA$Major), mean)

Major x

1 Accounting (Business) 645.2757

2 Aerospace Engineering 940.0000

3 African-American Studies 366.0000

4 Agricultural General 168.0000

5 Animal Sciences: Animal Care and Management 403.3333

6 Architecture 325.6000

7 Art Education 372.0000

8 Art History 305.0000

9 Bioengineering 1830.0000

10 Biological Sciences: Ecology and Evolution 354.0000

11 Biological Sciences: General Biology 686.6667

12 Business/Admin./Mgmt. 775.9929

13 Central European, Russian and Eurasian Studies 482.5000

14 Civil and Environmental Engineering 452.5500

15 Communication 595.6875

16 Community Health 1343.6667

17 Computer Engineering 380.4000

18 Computer Science 674.9217

19 Computer Systems Management 588.6538

20 Criminal Justice 697.0933

21 Criminology and Criminal Justice 771.2214

22 Early Childhood Education 1013.7500

23 Education (Teacher:Undecided) 1435.2500

24 Electrical Engineering 452.0000

25 Emergency Management 654.0522

26 Engineering (Undecided) 639.6500

27 English Language and Literature 720.9350

28 Environmental Science and Policy (Policy) 1881.0000

29 Environmental Science and Policy (Science) 878.0000

30 Environmental, Educational and Park Management 1077.8571

31 Family Studies 1464.0000

32 Finance (Business) 1352.4667

33 Fire Protection Engineering 630.6167

34 Fire Science 502.8623

35 Forensic Science 1402.1429

36 General Business and Management 1306.6667

37 General Studies 383.8517

38 Geography 535.4250

39 Government and Politics 1273.7143

40 Health & Human Services 789.3700

41 History 654.3997

42 Human Resources 720.5636

43 HVAC 364.4750

44 Individual Studies 735.0000

45 Information Systems-Business 1150.6667

46 Information Technology 910.1266

47 International Business 813.0000

48 Job Related 954.9978

49 Law 1022.2646

50 Liberal Arts/Gen. Studies 185.3333

51 Linguistics 577.5000

52 Logistics, Transportation and Supply Chain Management (Business) 1374.0000

53 Marketing (Business) 749.0000

54 Mathematics 810.5000

55 Mechanical Engineering 1107.4375

56 None 745.8185

57 Not Listed 756.3678

58 Not Specified 686.5805

59 Nursing 746.9164

60 Nutritional Sciences 501.0833

61 Operations Management (Business) 1041.8000

62 Other/Misc. 769.2516

63 Paralegal Studies 442.3429

64 Physical Education 386.0000

65 Political Science 947.4545

66 Pre-Medical Technology 292.0000

67 Pre-Nursing 347.2222

68 Professional/Technical 520.7143

69 Psychology 704.3370

70 Public Administration 1196.2807

71 Public Safety 976.4076

72 Russian Language and Literature 1830.0000

73 Science Education 585.0000

74 Social Studies Education 1930.0000

75 Social Work 982.5535

76 Sociology 607.9000

77 Spanish Language and Literature 909.6860

78 Undecided (Letters and Sciences) 240.0000

79 Urban Forestry 528.0000

From this data set we subset the data amongst two excel sheets TA3 & TA4 such that TA3 has all non-computer/data science courses and TA4 has all computer and data science related courses.

Now using R script

mean(TA3$Cost) =

|  |
| --- |
| 791.5187351  mean(TA4$Cost) = 740.95376  It is astonishing that computer/data science related courses actually cost lesser |