DSC 520 Final Project

Lara Clasen

6/1/2019

First I read in my Excel document containing the data sat I obtained from the State of Minnesota. The data dictionary for this data can be found in the State of Minnesota Data Tool (<https://apps.deed.state.mn.us/lmi/etd/Results.aspx>).

mydata <- read.csv("schooldata.csv")  
head(mydata)

## Instructional.Program.Title  
## 1 All CIP codes  
## 2 All CIP codes  
## 3 All CIP codes  
## 4 All CIP codes  
## 5 Natural resources and conservation  
## 6 Natural resources and conservation  
## Award.Level PFT2 WFT2 PFT3  
## 1 Certificates (less than Bachelor’s Degrees) 35.00% $37,955 41.30%  
## 2 Associate Degree 36.10% $39,814 41.50%  
## 3 Bachelor's Degree 43.00% $44,162 49.10%  
## 4 Graduate (Certificates and Degrees) 49.60% $70,492 53.10%  
## 5 Certificates (less than Bachelor’s Degrees) 39.10% $36,664 45.00%  
## 6 Associate Degree 17.60% $32,659 23.00%  
## WFT3 PFT4 WFT4  
## 1 $37,955 44.90% $40,250   
## 2 $39,814 46.80% $41,756   
## 3 $44,162 53.30% $47,489   
## 4 $70,492 53.40% $74,474   
## 5 $36,664 57.10% $31,143   
## 6 $32,659 33.10% $36,247

I need to adjust my data types so that I can analyze them. I change the percentage values to numeric and remove the dollar signs and commas from the currency values.

mydata$WFT2 <- as.numeric(gsub('\\$|,', '', mydata$WFT2))  
mydata$WFT3 <- as.numeric(gsub('\\$|,', '', mydata$WFT3))  
mydata$WFT4 <- as.numeric(gsub('\\$|,', '', mydata$WFT4))

## Warning: NAs introduced by coercion

mydata$PFT2 <- as.numeric(mydata$PFT2)  
mydata$PFT3 <- as.numeric(mydata$PFT3)  
mydata$PFT4 <- as.numeric(mydata$PFT4)  
  
head(mydata)

## Instructional.Program.Title  
## 1 All CIP codes  
## 2 All CIP codes  
## 3 All CIP codes  
## 4 All CIP codes  
## 5 Natural resources and conservation  
## 6 Natural resources and conservation  
## Award.Level PFT2 WFT2 PFT3 WFT3 PFT4  
## 1 Certificates (less than Bachelor’s Degrees) 37 37955 42 37955 37  
## 2 Associate Degree 40 39814 43 39814 45  
## 3 Bachelor's Degree 56 44162 56 44162 56  
## 4 Graduate (Certificates and Degrees) 66 70492 64 70492 57  
## 5 Certificates (less than Bachelor’s Degrees) 48 36664 50 36664 68  
## 6 Associate Degree 4 32659 4 32659 9  
## WFT4  
## 1 40250  
## 2 41756  
## 3 47489  
## 4 74474  
## 5 31143  
## 6 36247

str(mydata)

## 'data.frame': 112 obs. of 8 variables:  
## $ Instructional.Program.Title: Factor w/ 37 levels "All CIP codes",..: 1 1 1 1 25 25 25 25 2 2 ...  
## $ Award.Level : Factor w/ 4 levels "Associate Degree",..: 3 1 2 4 3 1 2 4 1 2 ...  
## $ PFT2 : num 37 40 56 66 48 4 16 57 103 32 ...  
## $ WFT2 : num 37955 39814 44162 70492 36664 ...  
## $ PFT3 : num 42 43 56 64 50 4 32 54 92 39 ...  
## $ WFT3 : num 37955 39814 44162 70492 36664 ...  
## $ PFT4 : num 37 45 56 57 68 9 42 69 101 60 ...  
## $ WFT4 : num 40250 41756 47489 74474 31143 ...

dim(mydata)

## [1] 112 8

glimpse(mydata)

## Observations: 112  
## Variables: 8  
## $ Instructional.Program.Title <fct> "All CIP codes", "All CIP codes", ...  
## $ Award.Level <fct> Certificates (less than Bachelor’s...  
## $ PFT2 <dbl> 37, 40, 56, 66, 48, 4, 16, 57, 103...  
## $ WFT2 <dbl> 37955, 39814, 44162, 70492, 36664,...  
## $ PFT3 <dbl> 42, 43, 56, 64, 50, 4, 32, 54, 92,...  
## $ WFT3 <dbl> 37955, 39814, 44162, 70492, 36664,...  
## $ PFT4 <dbl> 37, 45, 56, 57, 68, 9, 42, 69, 101...  
## $ WFT4 <dbl> 40250, 41756, 47489, 74474, 31143,...

stat.desc(mydata)

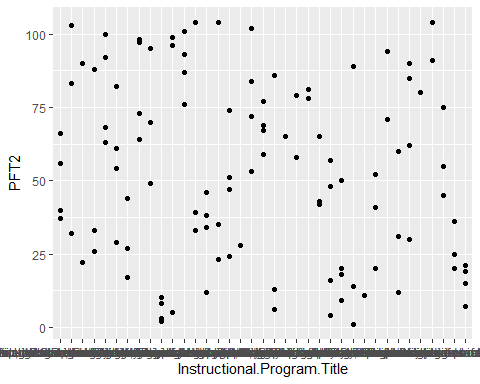
## Instructional.Program.Title Award.Level PFT2 WFT2  
## nbr.val NA NA 112.000000 1.120000e+02  
## nbr.null NA NA 0.000000 0.000000e+00  
## nbr.na NA NA 0.000000 0.000000e+00  
## min NA NA 1.000000 2.613900e+04  
## max NA NA 104.000000 2.513900e+05  
## range NA NA 103.000000 2.252510e+05  
## sum NA NA 5908.000000 5.336353e+06  
## median NA NA 52.500000 4.188600e+04  
## mean NA NA 52.750000 4.764601e+04  
## SE.mean NA NA 2.902355 2.255426e+03  
## CI.mean NA NA 5.751210 4.469278e+03  
## var NA NA 943.450450 5.697381e+08  
## std.dev NA NA 30.715639 2.386919e+04  
## coef.var NA NA 0.582287 5.009693e-01  
## PFT3 WFT3 PFT4 WFT4  
## nbr.val 112.0000000 1.120000e+02 112.0000000 1.110000e+02  
## nbr.null 0.0000000 0.000000e+00 0.0000000 0.000000e+00  
## nbr.na 0.0000000 0.000000e+00 0.0000000 1.000000e+00  
## min 1.0000000 2.613900e+04 1.0000000 2.859700e+04  
## max 101.0000000 2.513900e+05 106.0000000 2.517600e+05  
## range 100.0000000 2.252510e+05 105.0000000 2.231630e+05  
## sum 5763.0000000 5.336353e+06 5976.0000000 5.649092e+06  
## median 51.5000000 4.188600e+04 54.5000000 4.542300e+04  
## mean 51.4553571 4.764601e+04 53.3571429 5.089272e+04  
## SE.mean 2.7309916 2.255426e+03 2.8841548 2.287556e+03  
## CI.mean 5.4116423 4.469278e+03 5.7151453 4.533399e+03  
## var 835.3313224 5.697381e+08 931.6550837 5.808533e+08  
## std.dev 28.9020989 2.386919e+04 30.5230255 2.410090e+04  
## coef.var 0.5616927 5.009693e-01 0.5720513 4.735628e-01

summary(mydata)

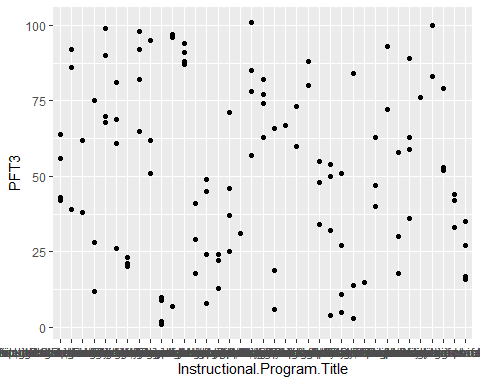
## Instructional.Program.Title  
## All CIP codes : 4   
## Business, management, marketing, and related support services: 4   
## Communication, journalism, and related programs : 4   
## Computer and information sciences and support services : 4   
## Education : 4   
## Engineering technologies and engineering-related fields : 4   
## (Other) :88   
## Award.Level PFT2   
## Associate Degree :25 Min. : 1.00   
## Bachelor's Degree :32 1st Qu.: 25.75   
## Certificates (less than Bachelor’s Degrees):27 Median : 52.50   
## Graduate (Certificates and Degrees) :28 Mean : 52.75   
## 3rd Qu.: 79.25   
## Max. :104.00   
##   
## WFT2 PFT3 WFT3 PFT4   
## Min. : 26139 Min. : 1.00 Min. : 26139 Min. : 1.00   
## 1st Qu.: 36823 1st Qu.: 26.75 1st Qu.: 36823 1st Qu.: 26.75   
## Median : 41886 Median : 51.50 Median : 41886 Median : 54.50   
## Mean : 47646 Mean : 51.46 Mean : 47646 Mean : 53.36   
## 3rd Qu.: 51593 3rd Qu.: 76.25 3rd Qu.: 51593 3rd Qu.: 78.25   
## Max. :251390 Max. :101.00 Max. :251390 Max. :106.00   
##   
## WFT4   
## Min. : 28597   
## 1st Qu.: 39560   
## Median : 45423   
## Mean : 50893   
## 3rd Qu.: 56105   
## Max. :251760   
## NA's :1

Next is to obtain some visual representation of the data.

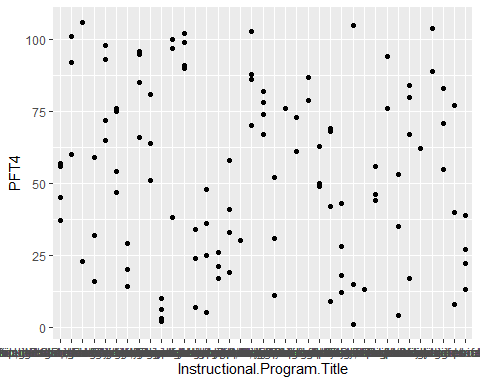
ggplot(mydata, aes(x = Instructional.Program.Title, y = PFT2)) + geom\_point()



ggplot(mydata, aes(x = Instructional.Program.Title, y = PFT3)) + geom\_point()



ggplot(mydata, aes(x = Instructional.Program.Title, y = PFT4)) + geom\_point()

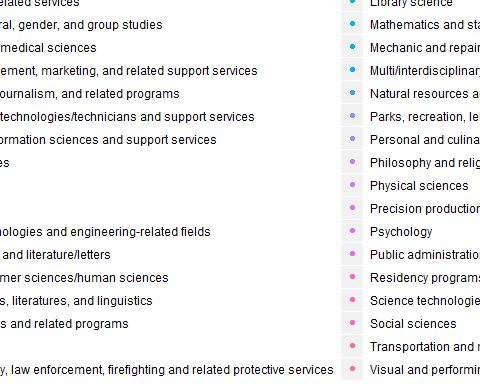


ggplot(data = mydata, mapping = aes(x = WFT2, y = WFT4)) +  
 geom\_point(mapping = aes(color = Instructional.Program.Title)) +  
 geom\_smooth()

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'

## Warning: Removed 1 rows containing non-finite values (stat\_smooth).

## Warning: Removed 1 rows containing missing values (geom\_point).



I want to fit a regression model on the data.

model1 <- lm(WFT4 ~ WFT2, data = mydata)  
model1

##   
## Call:  
## lm(formula = WFT4 ~ WFT2, data = mydata)  
##   
## Coefficients:  
## (Intercept) WFT2   
## 3261.3422 0.9995

summary(model1)

##   
## Call:  
## lm(formula = WFT4 ~ WFT2, data = mydata)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8763.0 -1469.5 -121.1 748.9 12968.7   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.261e+03 5.452e+02 5.982 2.85e-08 \*\*\*  
## WFT2 9.995e-01 1.023e-02 97.714 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2572 on 109 degrees of freedom  
## (1 observation deleted due to missingness)  
## Multiple R-squared: 0.9887, Adjusted R-squared: 0.9886   
## F-statistic: 9548 on 1 and 109 DF, p-value: < 2.2e-16

This translates to a regression of Wages after Year 4 on Wages after Year 2. According to our model, for every additional dollar amount earned after 2 years working full-time, we expect an additional 0.9995 earned in the 4th year after graduation while working full-time. This tells us that it does matter to obtain a higher-paying position just after graduation, but it may not matter all that much.

The Data Explorer report is also attached, which was produced using the following code: DataExplorer::create\_report(mydata)