Higher Education Outcomes

Sampritha Hassan Manjunath; Student ID 19232922

10/22/2019

Setup

```
# this setup chunk loads the tidyverse. no other libraries should be used
library("tidyverse")
```

```
## -- Attaching packages -------------------------------------------------------------- tidyverse 1.2.1
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ----------------------------------------------------------------- tidyverse_conflicts()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
# setting echo=TRUE will cause all R code to be included in the pdf output
knitr::opts_chunk$set(echo=TRUE)
```

Installing Latex

If you don't have Latex installed on your machine, you may see an error when `knit`ting this file, and no pdf will be produced. Try installing Latex as follows, and `knit` again.

```
# see https://yihui.name/tinytex/ for docs
> install.packages('tinytex')
> tinytex::install_tinytex()
```

Loading data

First, load the data, and convert to a tibble (i.e. a `dplyr` dataframe) named `earnings`, with column names "Years.since.graduation", "NFQ.Level", "Sex", "Field", "Statistic", and "Value".

```
# STUDENTS ADD CODE HERE
# Load csv and add column names to each column. Also fill ..(empty value) with NA
data <- read.csv2("earnings.csv", header = FALSE,
                  col.names = c("Years.since.graduation", "NFQ.Level", "Sex",
                                "Field", "Statistic", "Value"), na = "..")

# Convert data to a tibble (data frame)
```

```
earnings <- as_tibble(data)

# print data frame
earnings
```

```
## # A tibble: 1,600 x 6
##    Years.since.gradu~ NFQ.Level  Sex   Field           Statistic       Value
##                 <int> <fct>      <fct> <fct>           <fct>           <fct>
##  1                  1 NFQ Level~ Male  Education       Number of Grad~ 0
##  2                  1 NFQ Level~ Male  Education       P25 Earnings o~ <NA>
##  3                  1 NFQ Level~ Male  Education       P50 Earnings o~ <NA>
##  4                  1 NFQ Level~ Male  Education       P75 Earnings o~ <NA>
##  5                  1 NFQ Level~ Male  Arts and Huma~ Number of Grad~ 10
##  6                  1 NFQ Level~ Male  Arts and Huma~ P25 Earnings o~ 125.0
##  7                  1 NFQ Level~ Male  Arts and Huma~ P50 Earnings o~ 195.0
##  8                  1 NFQ Level~ Male  Arts and Huma~ P75 Earnings o~ 370.0
##  9                  1 NFQ Level~ Male  Social Scienc~ Number of Grad~ 0
## 10                  1 NFQ Level~ Male  Social Scienc~ P25 Earnings o~ <NA>
## # ... with 1,590 more rows
```

Reshaping and cleaning

We should change the NFQ Level values to integers. The following function will be useful:

```
convert_nfq <- function(s) {strtoi(substr(s, 11, 13))} # convert substring to int
```

Apply `convert_nfq` and check the result:

```
# STUDENTS ADD CODE HERE

# Replace  string values in NFQ.Level by Integer values
# with the help of confert_nfq function
earnings <- earnings %>% mutate(NFQ.Level = convert_nfq(NFQ.Level))

# Print the resulting data frame
earnings
```

```
## # A tibble: 1,600 x 6
##    Years.since.gradu~ NFQ.Level Sex   Field           Statistic       Value
##                 <int>     <int> <fct> <fct>           <fct>           <fct>
##  1                  1         6 Male  Education       Number of Grad~ 0
##  2                  1         6 Male  Education       P25 Earnings o~ <NA>
##  3                  1         6 Male  Education       P50 Earnings o~ <NA>
##  4                  1         6 Male  Education       P75 Earnings o~ <NA>
##  5                  1         6 Male  Arts and Human~ Number of Grad~ 10
##  6                  1         6 Male  Arts and Human~ P25 Earnings o~ 125.0
##  7                  1         6 Male  Arts and Human~ P50 Earnings o~ 195.0
##  8                  1         6 Male  Arts and Human~ P75 Earnings o~ 370.0
##  9                  1         6 Male  Social Science~ Number of Grad~ 0
## 10                  1         6 Male  Social Science~ P25 Earnings o~ <NA>
## # ... with 1,590 more rows
```

Let's rename the `Years.since.graduation` column since it's a long name:

```
# STUDENTS ADD CODE HERE

# Rename Years.since,graduation column to Years
earnings <- earnings %>% rename(Years = Years.since.graduation)

# Print the resulting data frame
earnings
```

```
## # A tibble: 1,600 x 6
##    Years NFQ.Level Sex   Field                    Statistic           Value
##    <int>     <int> <fct> <fct>                    <fct>               <fct>
##  1     1         6 Male  Education                Number of Graduate~ 0
##  2     1         6 Male  Education                P25 Earnings of Gr~ <NA>
##  3     1         6 Male  Education                P50 Earnings of Gr~ <NA>
##  4     1         6 Male  Education                P75 Earnings of Gr~ <NA>
##  5     1         6 Male  Arts and Humanities      Number of Graduate~ 10
##  6     1         6 Male  Arts and Humanities      P25 Earnings of Gr~ 125.0
##  7     1         6 Male  Arts and Humanities      P50 Earnings of Gr~ 195.0
##  8     1         6 Male  Arts and Humanities      P75 Earnings of Gr~ 370.0
##  9     1         6 Male  Social Sciences, Journa~ Number of Graduate~ 0
## 10     1         6 Male  Social Sciences, Journa~ P25 Earnings of Gr~ <NA>
## # ... with 1,590 more rows
```

Using `filter`, we discard all data where `Years` is not 1, because for some reason all that data is `NA`. Notice this reduces from 1600 rows to 400.

```
# STUDENTS ADD CODE HERE

# Filter data frame to select only rows where Years = 1
earnings <- earnings %>% filter(Years == 1)

# Print the resulting data frame
earnings
```

```
## # A tibble: 400 x 6
##    Years NFQ.Level Sex   Field                    Statistic           Value
##    <int>     <int> <fct> <fct>                    <fct>               <fct>
##  1     1         6 Male  Education                Number of Graduate~ 0
##  2     1         6 Male  Education                P25 Earnings of Gr~ <NA>
##  3     1         6 Male  Education                P50 Earnings of Gr~ <NA>
##  4     1         6 Male  Education                P75 Earnings of Gr~ <NA>
##  5     1         6 Male  Arts and Humanities      Number of Graduate~ 10
##  6     1         6 Male  Arts and Humanities      P25 Earnings of Gr~ 125.0
##  7     1         6 Male  Arts and Humanities      P50 Earnings of Gr~ 195.0
##  8     1         6 Male  Arts and Humanities      P75 Earnings of Gr~ 370.0
##  9     1         6 Male  Social Sciences, Journa~ Number of Graduate~ 0
## 10     1         6 Male  Social Sciences, Journa~ P25 Earnings of Gr~ <NA>
## # ... with 390 more rows
```

Our analysis is going to be based on Field, Sex, NFQ Level, Median Earnings, and Number of Graduates. We would like to have a column giving Median Earnings and another column giving Number of Graduates.

That would be *tidy data*. Instead, we have one column giving the `Statistic` name, and another giving that statistic's `Value`. We fix this using `spread`. Notice that in the result, there are several new columns. Some are shown directly, and the tibble says "2 more variables" at the bottom.

```
# STUDENTS ADD CODE HERE

# Separte column statistic and value to generate new columns
# (as Keys from 'statistic' value from 'value')
# Convert attribute, helps to get the type of new column created
earnings <- earnings %>% spread(Statistic, Value, convert = TRUE)

# Print the resulting data frame
earnings
```

```
## # A tibble: 100 x 8
##    Years NFQ.Level Sex   Field `Number of Grad~ `P25 Earnings o~
##    <int>     <int> <fct> <fct>            <int>            <dbl>
##  1     1         6 Fema~ Agri~               10              185
##  2     1         6 Fema~ Arts~               10              220
##  3     1         6 Fema~ Busi~              140              200
##  4     1         6 Fema~ Educ~                0               NA
##  5     1         6 Fema~ Engi~               10              215
##  6     1         6 Fema~ Heal~               90              210
##  7     1         6 Fema~ Info~                0               NA
##  8     1         6 Fema~ Natu~               20              195
##  9     1         6 Fema~ Serv~              100              280
## 10     1         6 Fema~ Soci~                0               NA
## # ... with 90 more rows, and 2 more variables: `P50 Earnings of Graduates
## #   (Euro)` <dbl>, `P75 Earnings of Graduates (Euro)` <dbl>
```

Now we can discard the 25th and 75th percentiles and rename the other columns:

```
# STUDENTS ADD CODE HERE

# Drop columns
earnings <- select(earnings, -c('P25 Earnings of Graduates (Euro)',
                                'P75 Earnings of Graduates (Euro)'))

# Rename columns
earnings <- earnings %>% rename(Number.grads = `Number of Graduates (Persons)`,
                                Median.Earnings = `P50 Earnings of Graduates (Euro)`)

# Rename can also be done as below. I have just preferred to use the above written one
#colnames(earnings)[5] <- "Number.grads"
#colnames(earnings)[6] <- "Median.grads"

# Print the resulting data frame
earnings
```

```
## # A tibble: 100 x 6
##    Years NFQ.Level Sex    Field                Number.grads Median.Earnings
##    <int>     <int> <fct>  <fct>                       <int>           <dbl>
##  1     1         6 Female Agriculture, Forest~           10             225
##  2     1         6 Female Arts and Humanities            10             255
##  3     1         6 Female Business, Administr~          140             250
##  4     1         6 Female Education                       0              NA
##  5     1         6 Female Engineering, Manufa~           10             260
##  6     1         6 Female Health and Welfare             90             290
##  7     1         6 Female Information and Com~            0              NA
##  8     1         6 Female Natural Sciences, M~           20             385
##  9     1         6 Female Services                      100             330
## 10     1         6 Female Social Sciences, Jo~            0              NA
## # ... with 90 more rows
```

Now, let's have a summary of what we've got:

```
summary(earnings)
```

```
##      Years      NFQ.Level       Sex    
##  Min.   :1   Min.   : 6   Female:50  
##  1st Qu.:1   1st Qu.: 7   Male  :50  
##  Median :1   Median : 8              
##  Mean   :1   Mean   : 8              
##  3rd Qu.:1   3rd Qu.: 9              
##  Max.   :1   Max.   :10              
##                                                                            
##                                                 Field     Number.grads   
##  Agriculture, Forestry, Fisheries and Veterinary:10   Min.   :   0.0  
##  Arts and Humanities                            :10   1st Qu.:  10.0  
##  Business, Administration and Law               :10   Median :  70.0  
##  Education                                      :10   Mean   : 256.8  
##  Engineering, Manufacturing and Construction    :10   3rd Qu.: 252.5  
##  Health and Welfare                             :10   Max.   :2550.0  
##  (Other)                                        :40                   
##  Median.Earnings
##  Min.   :195.0  
##  1st Qu.:355.0  
##  Median :460.0  
##  Mean   :478.9  
##  3rd Qu.:612.5  
##  Max.   :825.0  
##  NA's   :17
```

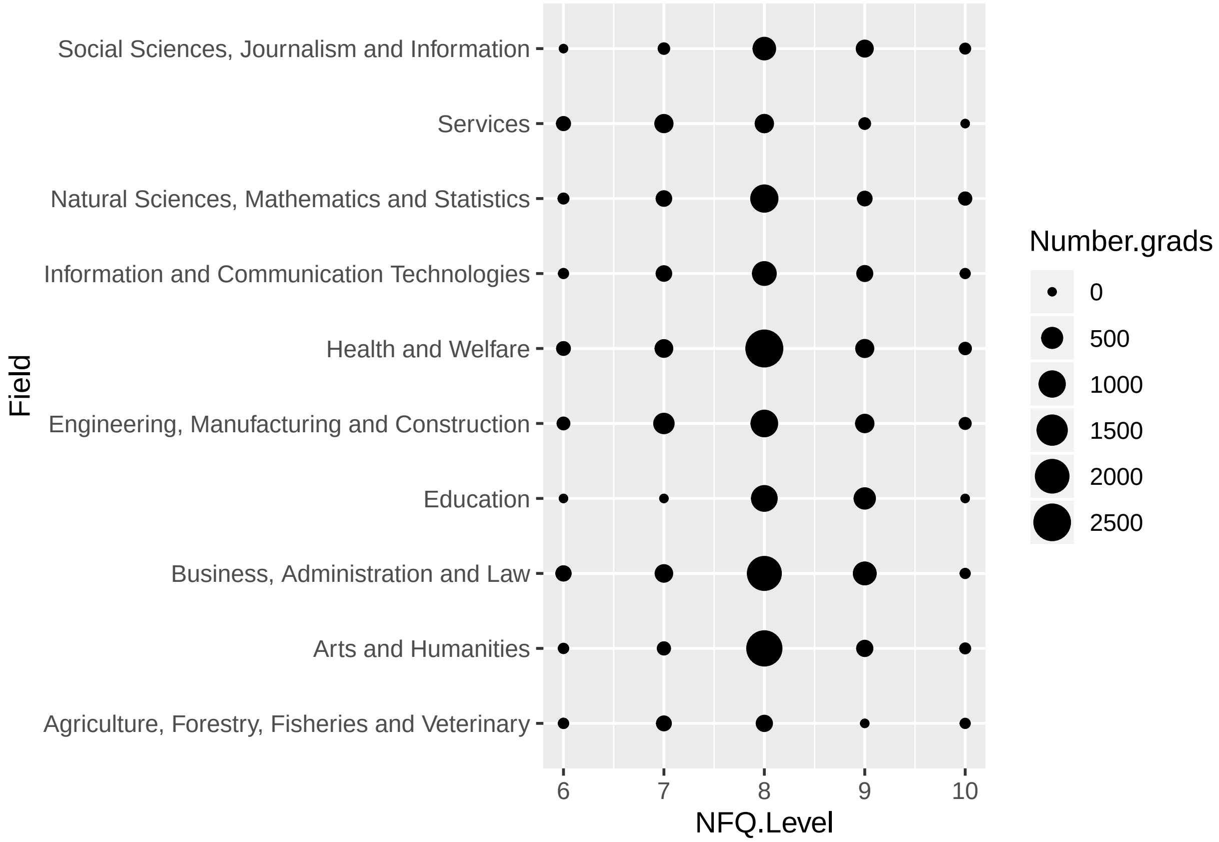
Plotting

Now we are ready to make a first plot. Let's look at the number of grads, by field and NFQ level.

```
# STUDENTS ADD CODE HERE

# Load the data frame to new variable
mpg <- earnings

# plot the graph with x and y as 'NFQ.Level' and 'Field' by taking Number.grads as the size
# Specifies how many grad are in what field and in which NFQ Level
ggplot(mpg) + geom_point(aes(x = NFQ.Level, y = Field, size = Number.grads))
```

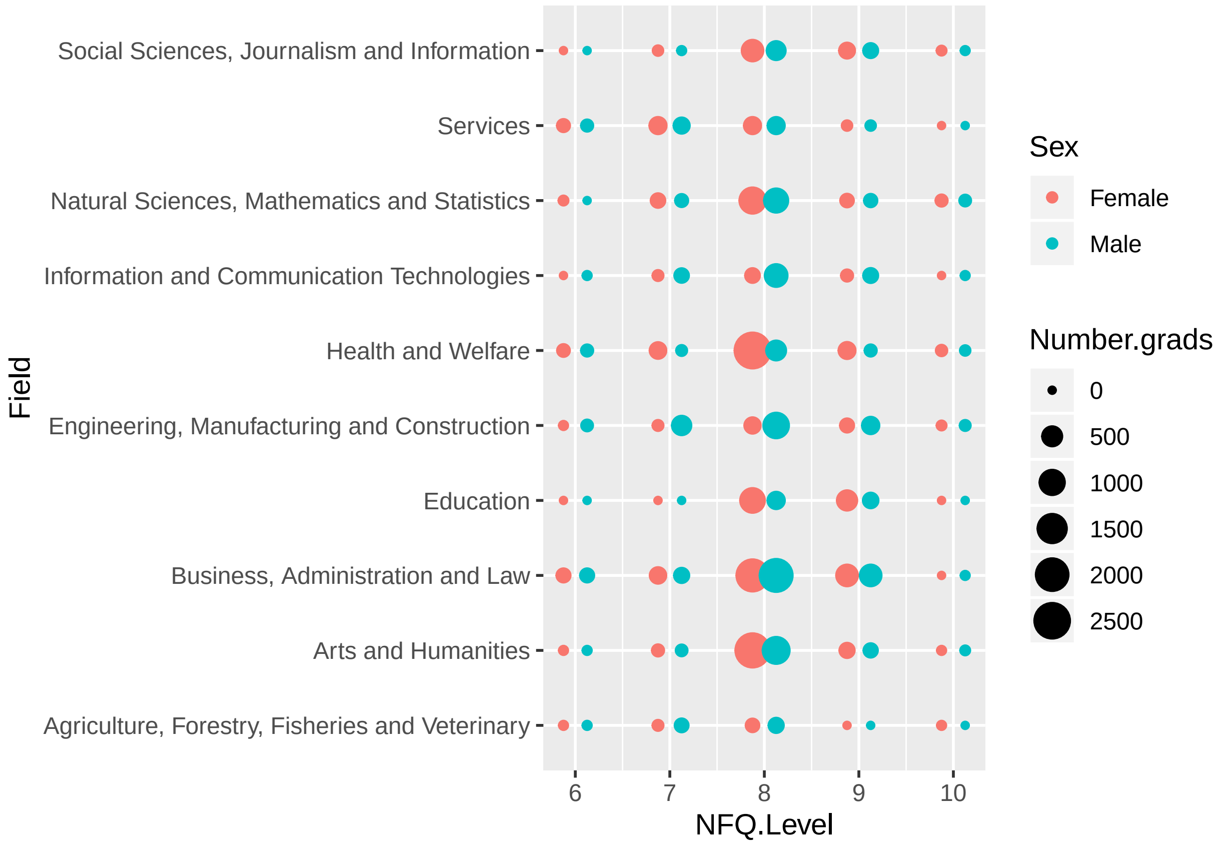


Now we'll analyse the data by Sex. Getting the male and female dots to appear correctly is tricky, so here is a snippet you can add to your `ggplot` call:

```
geom_point(position=position_nudge(x=0.25*(as.numeric(earnings$Sex) - 1.5)))
```

```
# STUDENTS ADD CODE HERE

# Filter out the dataframe as 'Females' and 'Males' and colour according to the Sex
mpg %>% filter(Sex == "Female" | Sex == "Male") %>%
  ggplot(mapping = aes(x=NFQ.Level, y=Field, size = Number.grads, colour = Sex)) +
  geom_point(position=position_nudge(x=0.25*(as.numeric(earnings$Sex) - 1.5)))
```

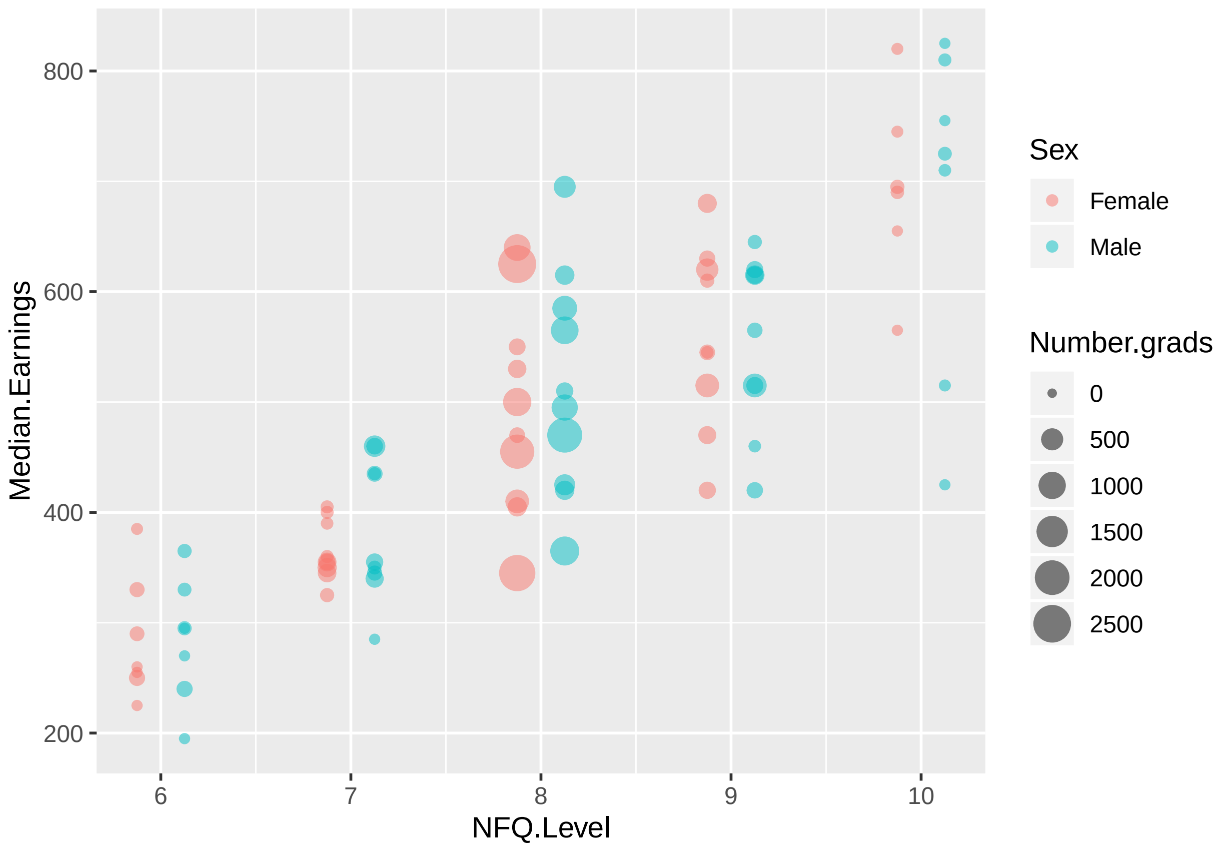


Here is a more traditional scatter plot, but bear in mind that what we see are distributions of median earnings, not distributions of earnings. We will see a Warning message "Removed 17 rows containing missing values (geom_point)." - this is correct, of course, as we do have NA values for earnings wherever there were no grads. We can ignore it.

```
# STUDENTS ADD CODE HERE

# Number if gradutes are now measured based on NFQ.Level and Median,Earnings
# alpha ia added to give the transparency and the overlapping picture in the graph
mpg %>% filter(Sex == "Female" | Sex == "Male") %>%
  ggplot(mapping = aes(x=NFQ.Level, y=Median.Earnings, size = Number.grads, colour = Sex)) +
  geom_point(position=position_nudge(x=0.25*(as.numeric(earnings$Sex) - 1.5)), alpha = 0.5)
```

```
## Warning: Removed 17 rows containing missing values (geom_point).
```



Join

Downloaded HEO02: Number of Graduates by NFQ Level, Sex, Type of Institute, Field of Study and Year, for year 2016

Loading and Processing the data as done earlier.

```
# Load csv and add column names to each column. Also fill ..(empty value) with NA
new_data <- read.csv("earnings_new.csv", header = FALSE,
                     col.names = c("NFQ.Level", "Type.Of.Institute", "Sex", "Field",
                                   "Number.grads"), na = "..")

# Convert data to a tibble (data frame)
new_earnings <- as_tibble(new_data)

# Replace  string values in NFQ.Level by Integer values with the help of confert_nfq function
new_earnings <- new_earnings %>% mutate(NFQ.Level = convert_nfq(NFQ.Level))

# filter columns where Number of graduates is zero
new_earnings <- new_earnings %>% filter(Number.grads != 0)

# Print the resulting data frame
new_earnings
```

```
## # A tibble: 148 x 5
##    NFQ.Level Type.Of.Institute    Sex   Field                  Number.grads
##        <int> <fct>                <fct> <fct>                         <int>
##  1         6 University           Male  Business, Administrat~           10
##  2         6 University           Male  Health and Welfare              180
##  3         6 University           Fema~ Business, Administrat~           10
##  4         6 University           Fema~ Natural Sciences, Mat~           10
##  5         6 University           Fema~ Health and Welfare              180
##  6         6 Institute of Techno~ Male  Arts and Humanities              10
##  7         6 Institute of Techno~ Male  Business, Administrat~          200
##  8         6 Institute of Techno~ Male  Natural Sciences, Mat~           10
##  9         6 Institute of Techno~ Male  Information and Commu~           30
## 10         6 Institute of Techno~ Male  Engineering, Manufact~           80
## # ... with 138 more rows
```

```r
summary(new_earnings)
```

```
##    NFQ.Level                    Type.Of.Institute     Sex
##  Min.   : 6.000   College                :15      Female:74
##  1st Qu.: 7.000   Institute of Technology:72      Male  :74
##  Median : 8.000   University             :61
##  Mean   : 8.162
##  3rd Qu.: 9.000
##  Max.   :10.000
##
##                                            Field     Number.grads
##  Health and Welfare                          :20   Min.   :  10.0
##  Arts and Humanities                         :18   1st Qu.:  20.0
##  Natural Sciences, Mathematics and Statistics:17   Median : 105.0
##  Business, Administration and Law            :16   Mean   : 253.4
##  Engineering, Manufacturing and Construction :15   3rd Qu.: 300.0
##  Information and Communication Technologies  :14   Max.   :2180.0
##  (Other)                                     :48
```

Perform full_join of earnings and new_earnings and save the result to new data frame for further use

```r
# perform full join
join_earnings <- full_join(earnings, new_earnings, by = c("NFQ.Level", "Sex", "Field", "Number.grads"))

join_earnings
```

```
## # A tibble: 233 x 7
##    Years NFQ.Level Sex   Field Number.grads Median.Earnings
##    <int>     <int> <fct> <fct>        <int>           <dbl>
##  1     1         6 Fema~ Agri~           10             225
##  2     1         6 Fema~ Arts~           10             255
##  3     1         6 Fema~ Busi~          140             250
##  4     1         6 Fema~ Educ~            0              NA
##  5     1         6 Fema~ Engi~           10             260
##  6     1         6 Fema~ Heal~           90             290
##  7     1         6 Fema~ Info~            0              NA
##  8     1         6 Fema~ Natu~           20             385
##  9     1         6 Fema~ Serv~          100             330
## 10     1         6 Fema~ Soci~            0              NA
## # ... with 223 more rows, and 1 more variable: Type.Of.Institute <fct>
```

```r
# Remove rows where Type of Institute is null
join_earnings <- join_earnings %>% filter(is.na(Type.Of.Institute) == FALSE)

join_earnings
```

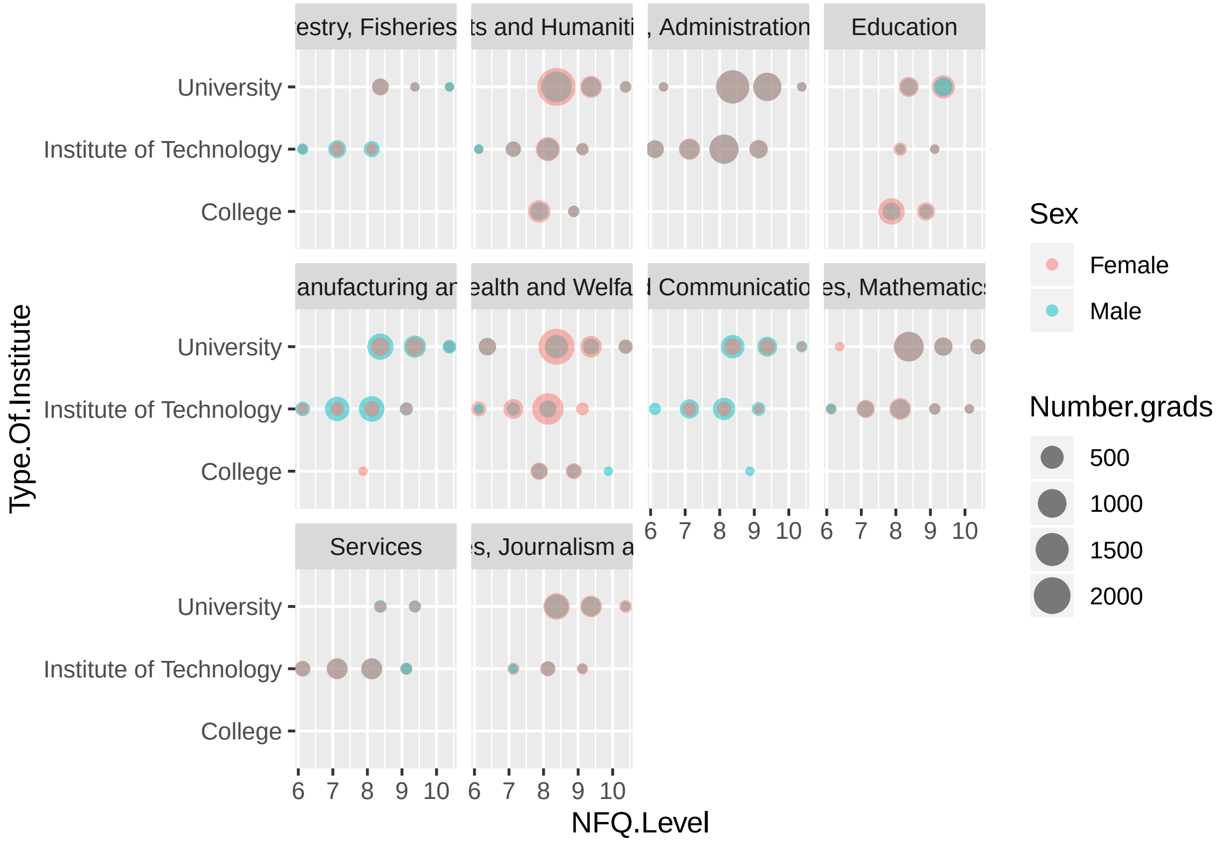
```
## # A tibble: 148 x 7
##    Years NFQ.Level Sex   Field Number.grads Median.Earnings
##    <int>     <int> <fct> <fct>        <int>           <dbl>
##  1     1         6 Fema~ Agri~           10             225
##  2     1         6 Fema~ Arts~           10             255
##  3     1         6 Fema~ Heal~           90             290
##  4     1         6 Fema~ Natu~           20             385
##  5     1         6 Male  Arts~           10             195
##  6     1         7 Fema~ Soci~           30             390
##  7     1         7 Male  Soci~           10             285
##  8     1         9 Fema~ Educ~          520             620
##  9     1         9 Fema~ Serv~           30             545
## 10     1         9 Male  Serv~           30             460
## # ... with 138 more rows, and 1 more variable: Type.Of.Institute <fct>
```

——————————————Analysis———————————————-

The intention behind the below analysis to check which 'Type of Insititute' graduates choose for different fields of study, along with diversity ratio.

```r
mpg <- join_earnings

# Check how many graduates are involved in each field along with diversity ratio
# Also plots which Type of Institute students prefer for respective Fields
mpg %>% ggplot(mapping = aes(x = NFQ.Level, y = Type.Of.Institute, size = Number.grads,
                             colour = Sex)) +
  geom_point(position=position_nudge(x=0.25*(as.numeric(join_earnings$Type.Of.Institute) - 1.5)),
             alpha = 0.5) +
  facet_wrap(~Field)
```



Based on above plotting, it can be concluded that most of the students prefer ‘University’ for thier field of study. To narrow it down, it can be observed that most of the Master’s and Ph.D students prefer ‘University’ above the other two types of Institute.

————————————End of Analysis————————————