Higher Education Outcomes

Swati; Student ID 19233301

10/25/2019

Setup

```
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()
```

```
knitr::opts_chunk$set(echo=TRUE)
```

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

#Source:https://readr.tidyverse.org/reference/read_delim.html

# Reading the csv as a tibble
earnings <- read_delim("earnings.csv", delim = ";",col_names = c("Years.since.graduation",
"NFQ.Level", "Sex", "Field", "Statistic", "Value"),
col_types = cols(Years.since.graduation = "i",
NFQ.Level = "f", Sex = "f", Field = "f", Statistic = "f", Value = "d"), na = c("..", "NA"))


# Displaying the data in tibble
earnings
```

```
## # A tibble: 1,600 x 6
##    Years.since.gradu~ NFQ.Level  Sex   Field          Statistic       Value
##                 <int> <fct>      <fct> <fct>          <fct>           <dbl>
##  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
##  7                  1 NFQ Level~ Male  Arts and Huma~ P50 Earnings o~   195
##  8                  1 NFQ Level~ Male  Arts and Huma~ P75 Earnings o~   370
##  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

# Convert function for converting column factor to integer
convert_nfq <- function(s) {strtoi(substr(s, 11, 13))}

# Selecting the NFQ.Level column from tibble
Test <- earnings %>% select(NFQ.Level)

# Applys convert function to column NFQ.Level
earnings1 <- apply(Test,2,convert_nfq)

# Unpacking the column
NFQ.Level <- earnings1[,1]

# Drops the existing NFQ.Level(fact) column
Test1 <- select(earnings, -c(2))

# Adds NFQ.Level(int) column back to tibble
earnings <- add_column(Test1, NFQ.Level, .after = "Years.since.graduation")

# Displays the tibble
earnings
```

```
## # A tibble: 1,600 x 6
##    Years.since.gradu~ NFQ.Level Sex   Field           Statistic       Value
##                 <int>     <int> <fct> <fct>           <fct>           <dbl>
##  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
##  7                  1         6 Male  Arts and Human~ P50 Earnings o~   195
##  8                  1         6 Male  Arts and Human~ P75 Earnings o~   370
```

```
##  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

#Source: https://medium.com/@HollyEmblem/renaming-columns-with-dplyr-in-r-55b42222cbdc

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

#Displays the tibble
earnings
```

```
## # A tibble: 1,600 x 6
##    Years NFQ.Level Sex   Field                    Statistic           Value
##    <int>     <int> <fct> <fct>                    <fct>               <dbl>
##  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
##  7     1         6 Male  Arts and Humanities      P50 Earnings of Gr~   195
##  8     1         6 Male  Arts and Humanities      P75 Earnings of Gr~   370
##  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

# Source: https://blog.exploratory.io/filter-data-with-dplyr-76cf5f1a258e

# Discards the rows with years != 1
earnings2 <- earnings %>%
  select(Years, NFQ.Level, Sex, Field, Statistic, Value) %>%
  filter(Years == "1")

#Display the tibble
earnings2
```

```
## # A tibble: 400 x 6
##    Years NFQ.Level Sex   Field                    Statistic           Value
##    <int>     <int> <fct> <fct>                    <fct>               <dbl>
##  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
##  7     1         6 Male  Arts and Humanities      P50 Earnings of Gr~   195
##  8     1         6 Male  Arts and Humanities      P75 Earnings of Gr~   370
##  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
```

options(tinytex.verbose = TRUE)

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

# Source:http://www.sthda.com/english/wiki/tidyr-crucial-step-re shaping-data-with-r-for-easier-analyse

# spreads 1 column into multiple columns
earnings <- spread(earnings2, Statistic,Value
                   )

# sorts Sex column in Lexicographical  order
earnings <- earnings %>% arrange(desc(Sex))

# displays the tibble
earnings
```

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

# Drops column 6 and 8
earnings3 <- select (earnings,-c(6,8))
```

```r
# Renames column Number of Graduates (Persons) & P50 Earnings of Graduates (Euro)
earnings <- rename(earnings3,
                   Number.grads = 'Number of Graduates (Persons)',
                   Median.Earnings = 'P50 Earnings of Graduates (Euro)')

# Displays the tibble
earnings
```

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

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

```r
# Displays the summary
summary(earnings)
```

```
##      Years     NFQ.Level      Sex
##  Min.   :1   Min.   : 6   Male  :50
##  1st Qu.:1   1st Qu.: 7   Female:50
##  Median :1   Median : 8
##  Mean   :1   Mean   : 8
##  3rd Qu.:1   3rd Qu.: 9
##  Max.   :1   Max.   :10
##
##                                          Field     Number.grads
##  Education                                  :10   Min.   :   0.0
##  Arts and Humanities                        :10   1st Qu.:  10.0
##  Social Sciences, Journalism and Information :10   Median :  70.0
##  Business, Administration and Law           :10   Mean   : 256.8
##  Natural Sciences, Mathematics and Statistics:10   3rd Qu.: 252.5
##  Information and Communication Technologies  :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
```

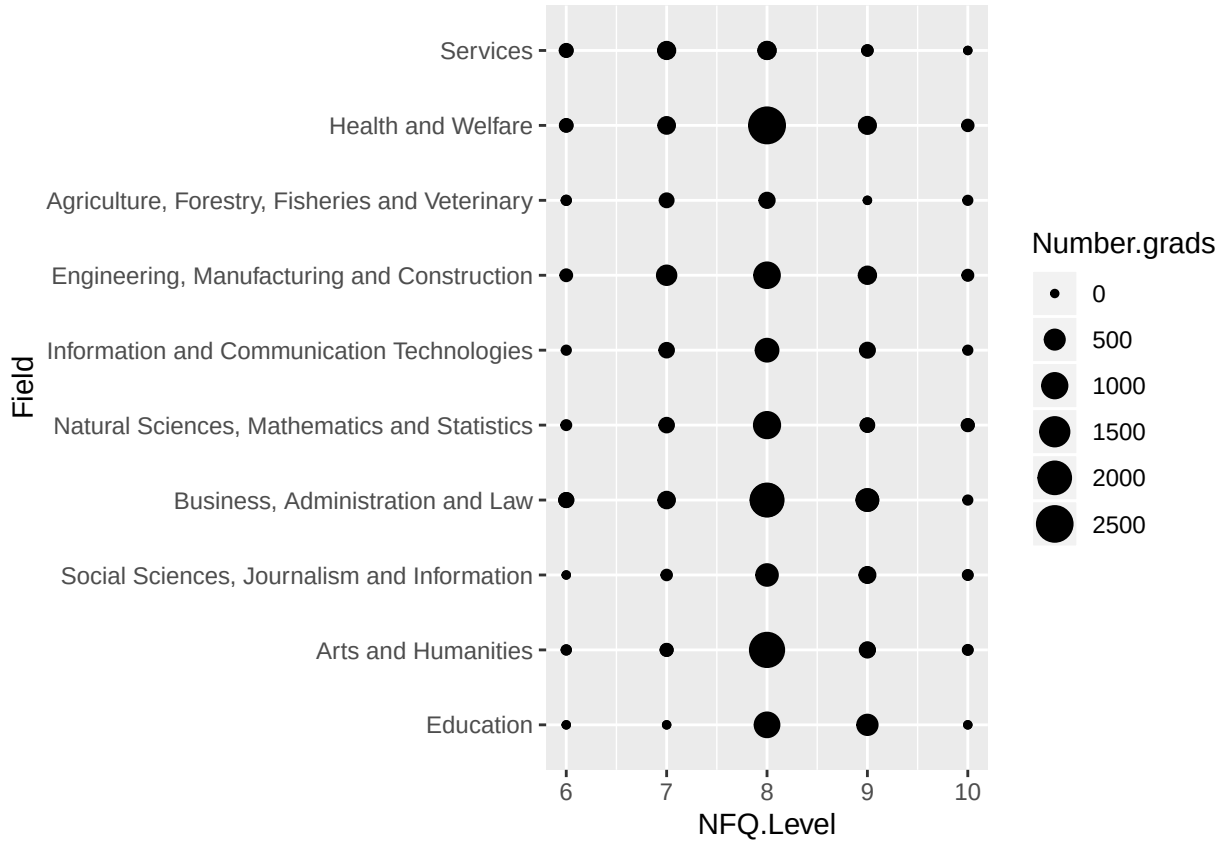
Plotting

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

# Source: https://uoftcoders.github.io/rcourse/lec04-dplyr.html

# plots a graph with NFQ Level as x axis, Field as y axis
ggplot(earnings, aes(x = NFQ.Level, y = Field, size = Number.grads)) + geom_point()
```



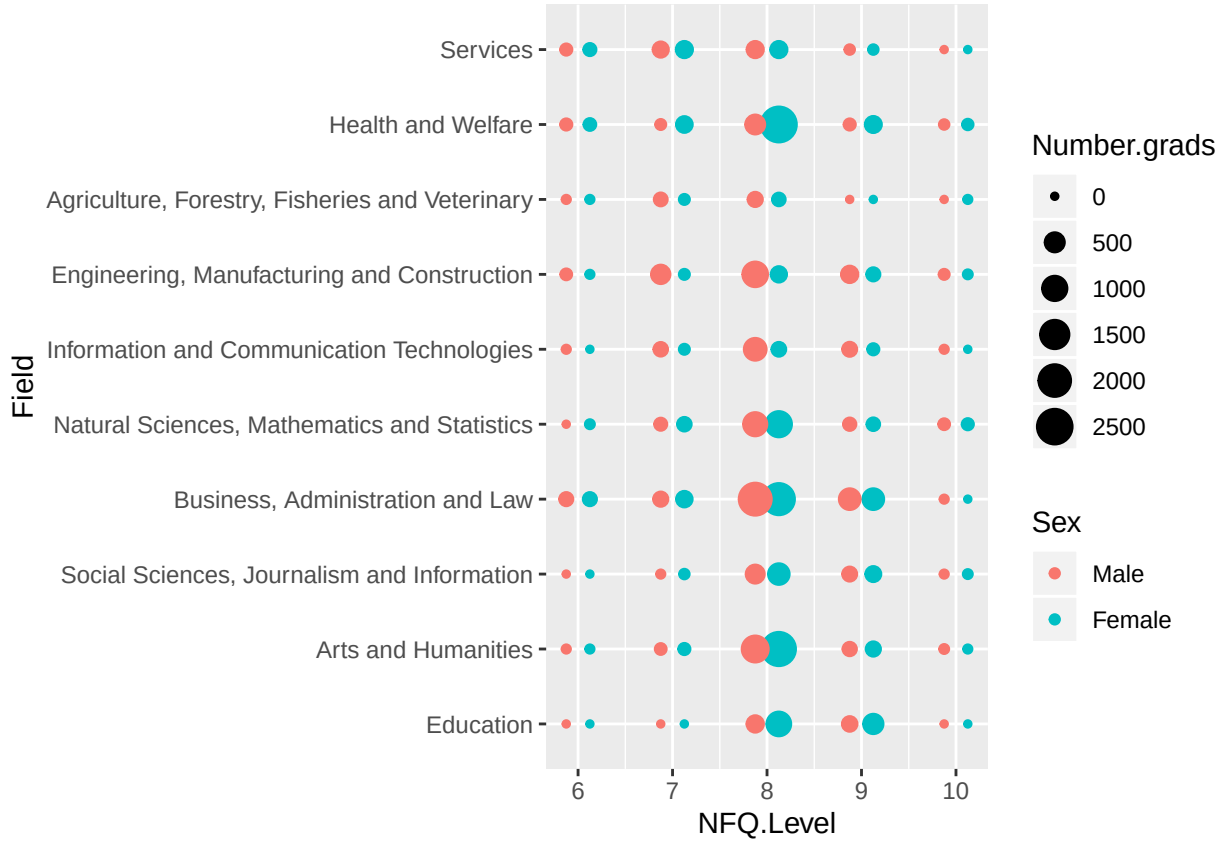
Now we'll analyse the data by Sex. Getting the male and female dots to appear correctly is tricky, so here is a snippet you can add to your `ggplot` call:

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

```
# STUDENTS ADD CODE HERE

# plots a scatter plot displaying male and female
ggplot(earnings, aes(x = NFQ.Level, y = Field, size = Number.grads, color = Sex)) +

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



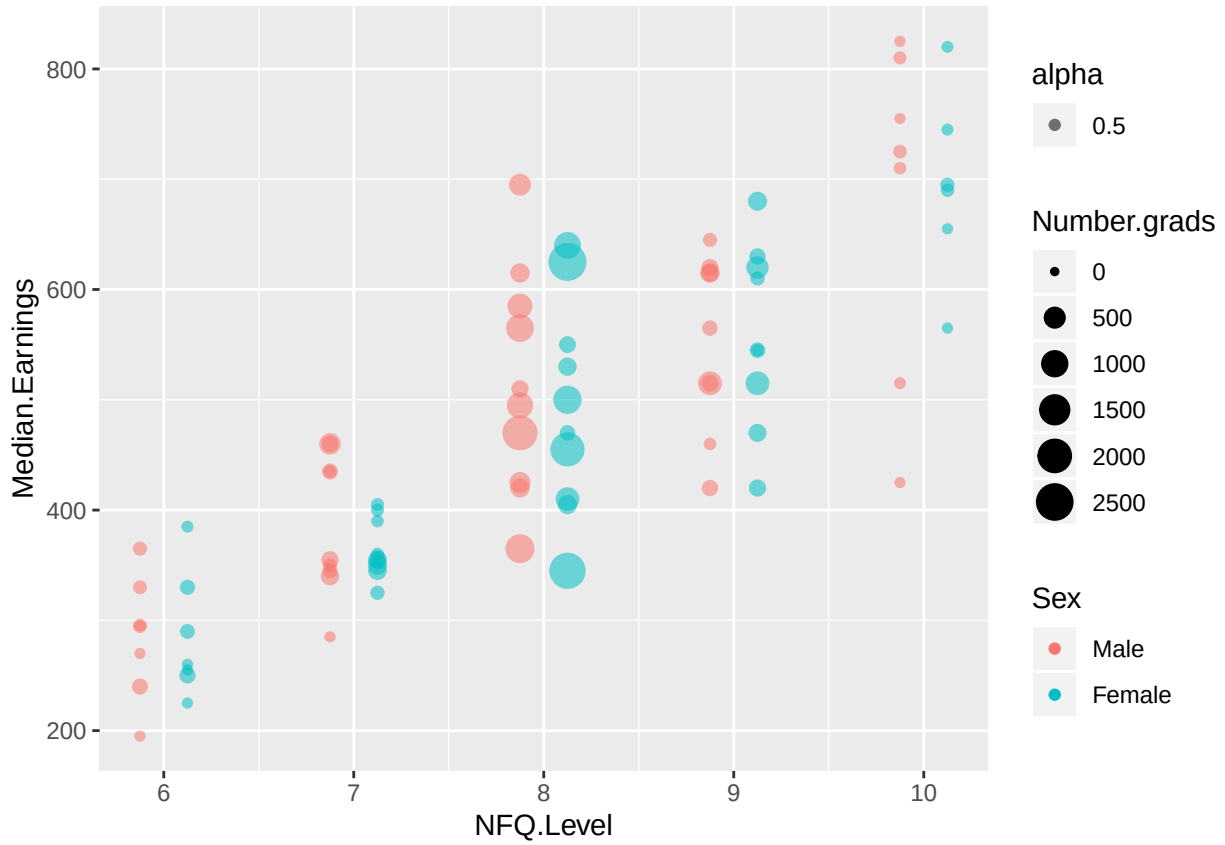
Here is a more traditional scatter plot, but bear in mind that what we see are distributions of median earnings, not distributions of earnings. We will see a Warning message "Removed 17 rows containing missing values (geom_point)." - this is correct, of course, as we do have NA values for earnings wherever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

```
# STUDENTS ADD CODE HERE

# plots a scatter plot with y axis as median earnings
ggplot(earnings, aes(alpha = 0.5,x = NFQ.Level,
                     y = Median.Earnings,
                     size = Number.grads, color = Sex)) +

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

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




```
# Joining 2 tibbles


# Reading TypeOfInstitute.csv as a tibble
TypeOfInstitute <- read_delim("TypeOfInstitute.csv",
    delim = ";",col_names = c("Number of Graduates by NFQ Level","Type of Institute",
    "Sex", "Field of Study", "Year"),
    col_types = cols('Number of Graduates by NFQ Level' = "c", 'Type of Institute' = "f",
                     Sex = "f", 'Field of Study' = "f"))

# Displays tibble
TypeOfInstitute
```

```
## # A tibble: 300 x 5
##    `Number of Graduate~ `Type of Institu~ Sex   `Field of Study`       Year
##    <chr>                <fct>             <fct> <fct>                 <dbl>
##  1 NFQ Level 6          University        Male  Education                 0
##  2 NFQ Level 6          University        Male  Arts and Humanities       0
##  3 NFQ Level 6          University        Male  Social Sciences, Jou~     0
##  4 NFQ Level 6          University        Male  Business, Administra~    10
##  5 NFQ Level 6          University        Male  Natural Sciences, Ma~     0
##  6 NFQ Level 6          University        Male  Information and Comm~     0
##  7 NFQ Level 6          University        Male  Engineering, Manufac~     0
##  8 NFQ Level 6          University        Male  Agriculture, Forestr~     0
##  9 NFQ Level 6          University        Male  Health and Welfare      180
## 10 NFQ Level 6          University        Male  Services                  0
```

```
## # ... with 290 more rows
```

```
# Renaming some columns of tibble
TypeOfInstitute <- TypeOfInstitute %>%
  rename(Field = "Field of Study", Institute = "Type of Institute",
         NumberGraduates = "Number of Graduates by NFQ Level")

# Dropping column Sex (manipulating data)
TypeOfInstitute <- select(TypeOfInstitute, -c(3))


# Dropping columns Years & NFQ.Level from earnings
earnings <- select(earnings, -c(1,2))


# Two tibbles to join

earnings  # tibble 1
```

```
## # A tibble: 100 x 4
##    Sex    Field                                Number.grads Median.Earnings
##    <fct>  <fct>                                       <dbl>           <dbl>
##  1 Female Education                                       0              NA
##  2 Female Arts and Humanities                            10             255
##  3 Female Social Sciences, Journalism and Inf~            0              NA
##  4 Female Business, Administration and Law              140             250
##  5 Female Natural Sciences, Mathematics and S~           20             385
##  6 Female Information and Communication Techn~            0              NA
##  7 Female Engineering, Manufacturing and Cons~           10             260
##  8 Female Agriculture, Forestry, Fisheries an~           10             225
##  9 Female Health and Welfare                             90             290
## 10 Female Services                                      100             330
## # ... with 90 more rows
```

```
TypeOfInstitute # tibble 2
```

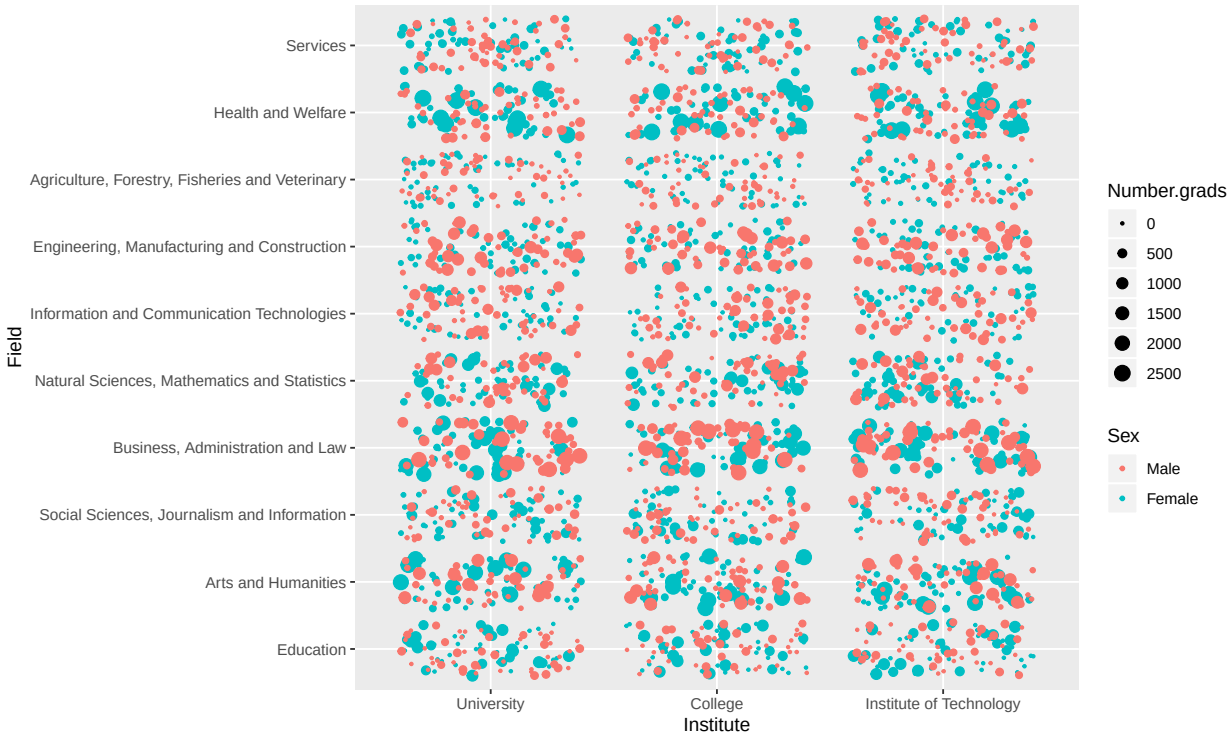
```
## # A tibble: 300 x 4
##    NumberGraduates Institute  Field                                    Year
##    <chr>           <fct>      <fct>                                   <dbl>
##  1 NFQ Level 6     University Education                                   0
##  2 NFQ Level 6     University Arts and Humanities                         0
##  3 NFQ Level 6     University Social Sciences, Journalism and Inform~     0
##  4 NFQ Level 6     University Business, Administration and Law           10
##  5 NFQ Level 6     University Natural Sciences, Mathematics and Stat~     0
##  6 NFQ Level 6     University Information and Communication Technolo~     0
##  7 NFQ Level 6     University Engineering, Manufacturing and Constru~     0
##  8 NFQ Level 6     University Agriculture, Forestry, Fisheries and V~     0
##  9 NFQ Level 6     University Health and Welfare                        180
## 10 NFQ Level 6     University Services                                     0
## # ... with 290 more rows
```

```
# Using full_join to join by column Field
Joined <- dplyr::full_join(earnings, TypeOfInstitute, by = "Field")

# Displays the joined tibble
Joined
```

```
## # A tibble: 3,000 x 7
##    Sex   Field Number.grads Median.Earnings NumberGraduates Institute  Year
##    <fct> <fct>        <dbl>           <dbl> <chr>           <fct>     <dbl>
##  1 Fema~ Educ~            0              NA NFQ Level 6     Universi~     0
##  2 Fema~ Educ~            0              NA NFQ Level 6     Universi~     0
##  3 Fema~ Educ~            0              NA NFQ Level 6     College       0
##  4 Fema~ Educ~            0              NA NFQ Level 6     College       0
##  5 Fema~ Educ~            0              NA NFQ Level 6     Institut~     0
##  6 Fema~ Educ~            0              NA NFQ Level 6     Institut~     0
##  7 Fema~ Educ~            0              NA NFQ Level 7     Universi~     0
##  8 Fema~ Educ~            0              NA NFQ Level 7     Universi~     0
##  9 Fema~ Educ~            0              NA NFQ Level 7     College       0
## 10 Fema~ Educ~            0              NA NFQ Level 7     College       0
## # ... with 2,990 more rows
```

```
# Plots a jitter plot for 1 column from TypeOfInstitute, 1 from earnings with Field
ggplot(Joined, aes(x = Institute,
                   y = Field, size = Number.grads, color = Sex))+geom_jitter(shape = 20)
```



INTERPRETATION OF GRAPH ABOVE

The plot is produced after joining two tibble by column Field with a full_join. This method is a union as it joins all the rows and places NA for the missing values. X-axis gives the type of institutions, Y-axis gives

the Field of study. It is a jitter plot which is plotted on the basis of Number of graduates and the type of institution. The colour defines the sex of the graduates.

Health and Welfare have a large number of female graduates from all the 3 institutions. Business administration & Law has same number of male and female graduates from all 3 institutions. Arts and Humanitues have large number of female graduates. Information and Communication Technologies has very few graduates. So, the total number of graduates from all the different type of institutions is more or less the same as seen in the graph