(Fast) Introduction to R

Jump into a notebook

Joana Cima

11 outubro 2023

My beamer

BlaBlaBla

Outline

1. Motivation
2. Data
3. Conceptual discussion

3. Import data (from an excel file)

Load your data using point and click

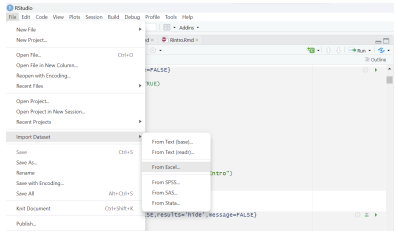


Figure 1: Point and click

which corresponds to the following code

```
nlswork <- as.data.frame(read_excel("nlswork.xlsx"))
# nlswork <- read_dta("nlswork.dta") # in case you have a Stata data source
```

4. Data manipulation – check the pipe operator, %>%

4.1. Select a subset of variables

```
nlswork_s<- nlswork %>%
  select(idcode, ln_wage)
```

4.2. Rename variables

```
nlswork_r <- nlswork %>%
  rename(cae = ind_code)
```

4.3. Filter a subset of observations

```
nlswork_f<- nlswork %>%
  filter(age > 40)
```

4.4. Mutate: create variables

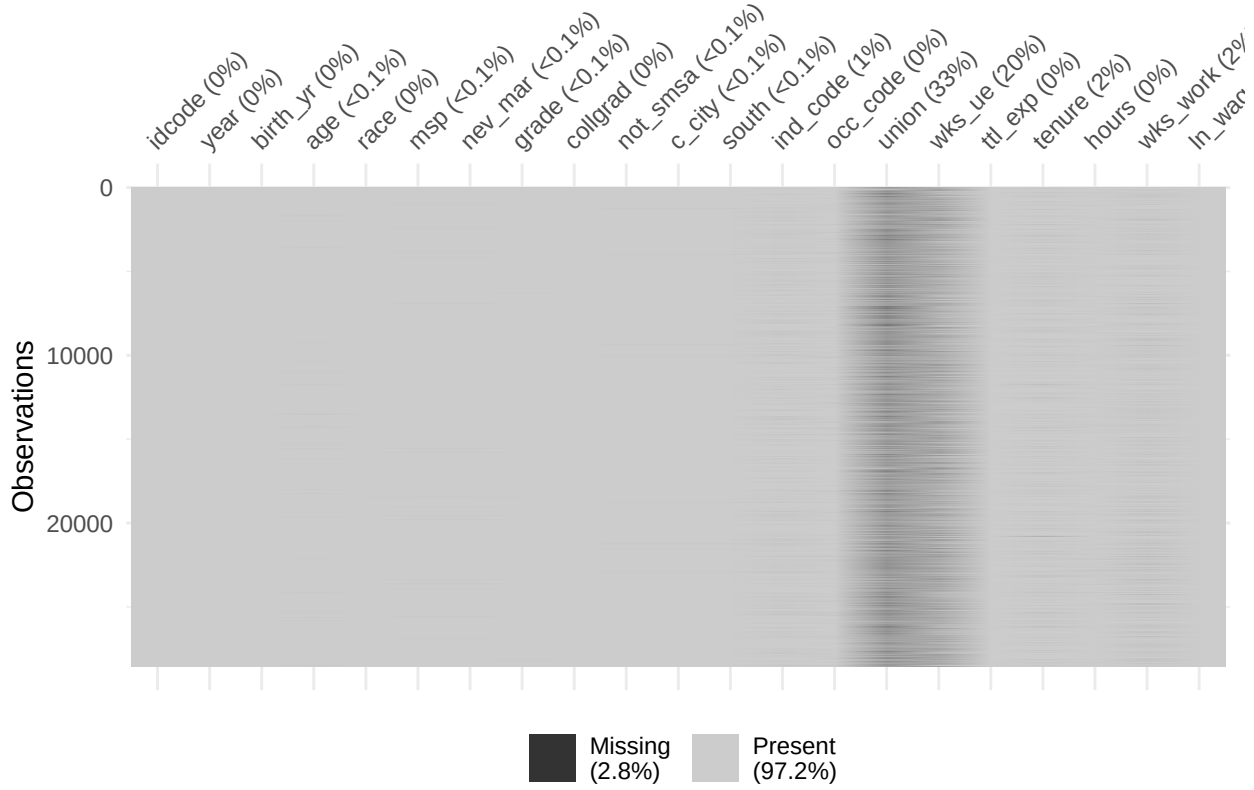
```
 nlswork_m <- nlswork %>%
  mutate(ln_asd=log(age))
```

4.5. Manipulate the data in a single sequence

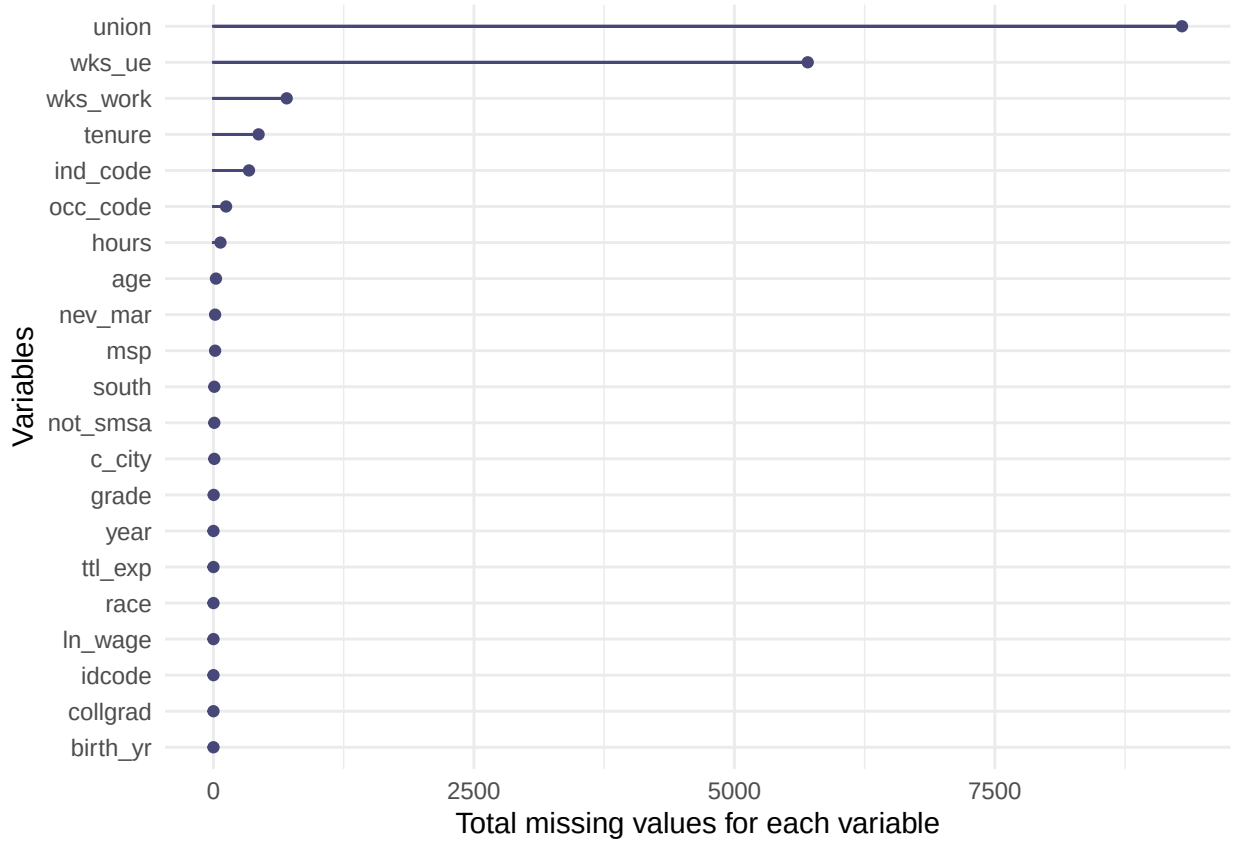
```
nlswork1<- nlswork %>%
  rename(cae = ind_code) %>%
  select(idcode, ln_wage, age) %>%
  filter(age > 40) %>%
  mutate(age2=age^2)
```

5. Visualize missing information:

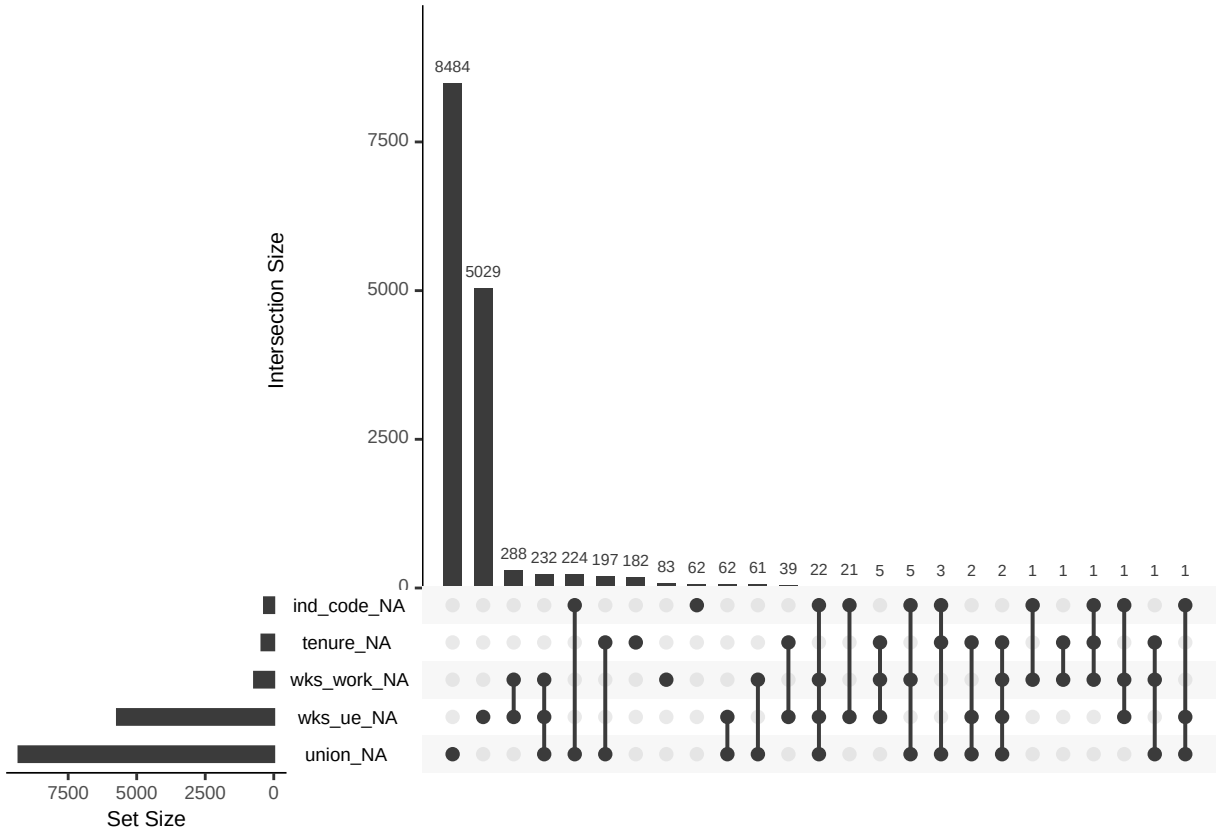
```
vis_miss(nlswork)
```



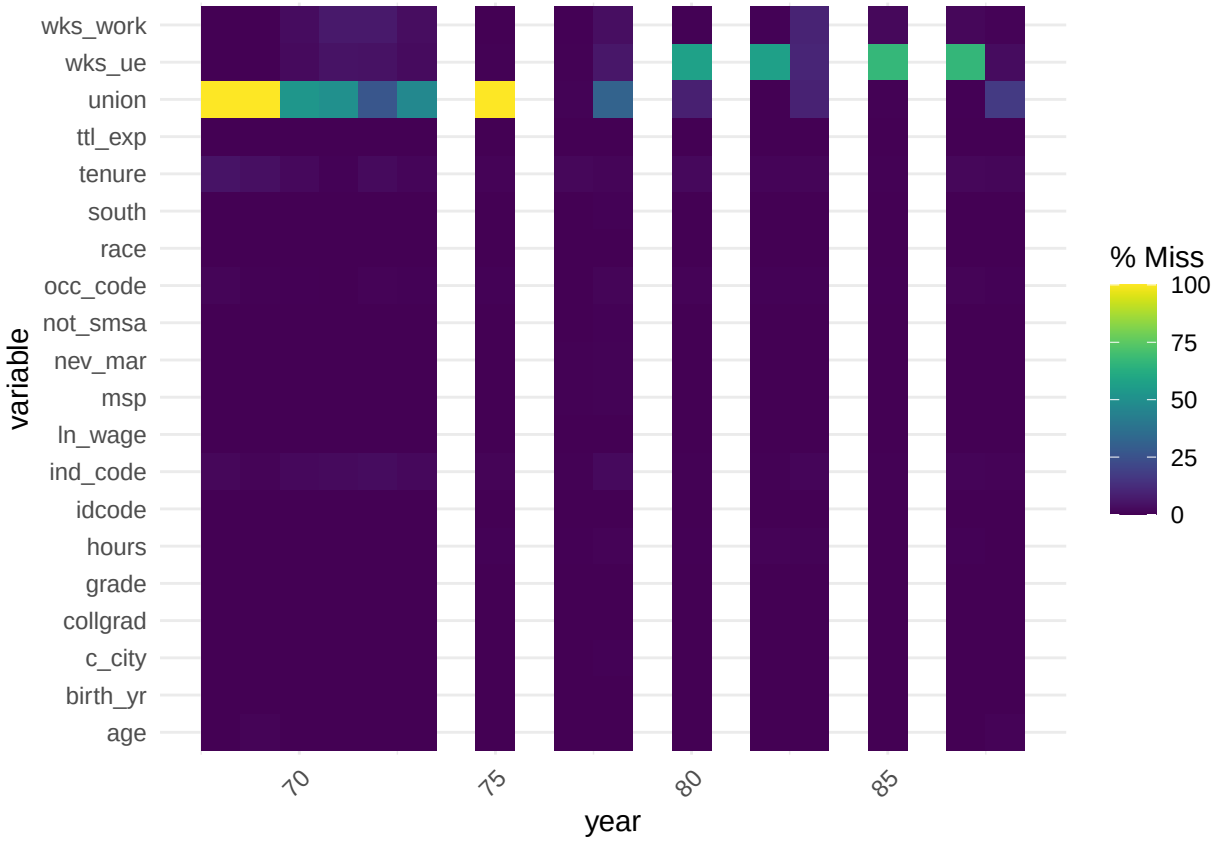

```
gg_miss_var(nlswork) + labs(y = "Total missing values for each variable")
```




```
gg_miss_upset(nlswork)
```

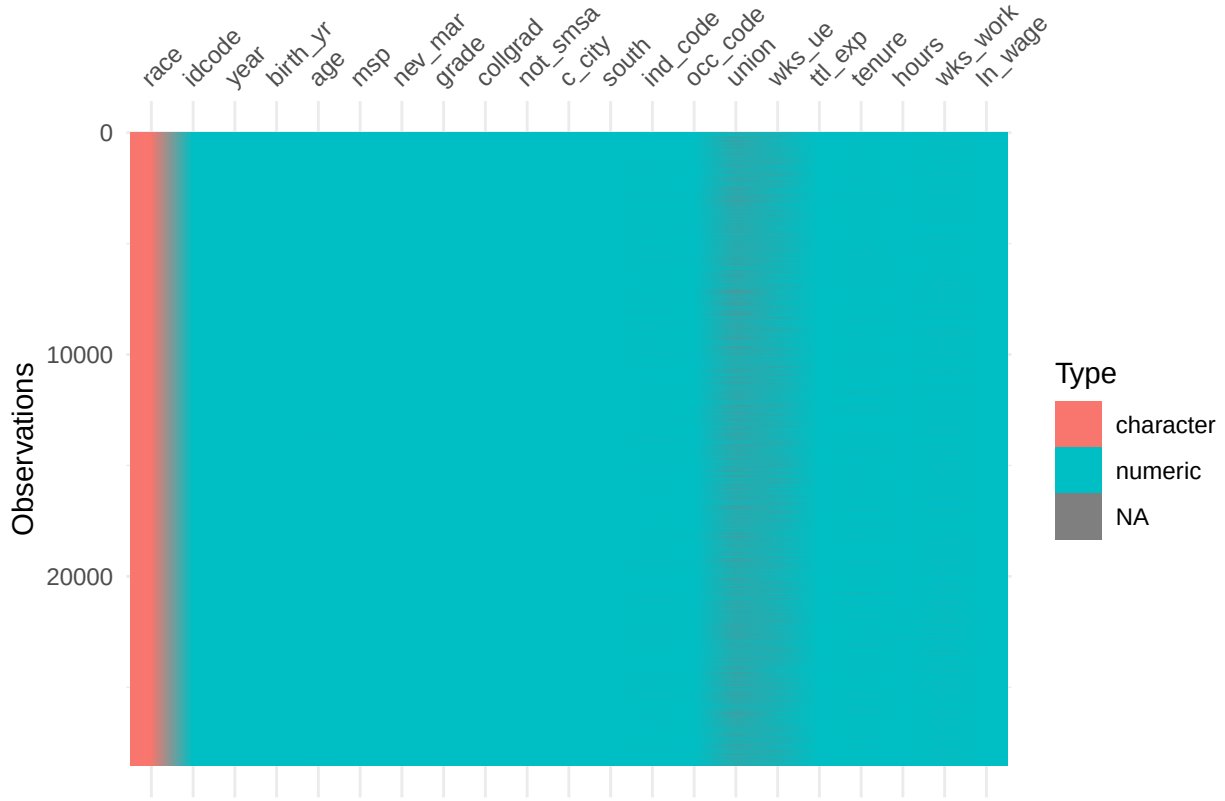



```
gg_miss_fct(x = nlswork,fct = year)
```



Alternative

```
vis_dat(nlswork)
```



6. Handling Missing Data

Handling missing data is a crucial step in the exploratory data analysis. Depending on the nature and mechanism of the missingness, we might decide to impute missing values or to exclude the observations with missing data.

6.1 Filling Missing Data

In some situations, we may opt to fill in the missing data. For instance, one common method involves replacing missing values with the mean of the variable.

```
# Filling Missing Data (with the average - this is an example)
nlswork_filled <- nlswork %>%
  mutate(across(c("union"), ~ ifelse(is.na(.), mean(., na.rm = TRUE), .)))
```

6.2 Excluding rows with missing data

```
# Or excluding rows with missing data

nlswork_no_na <- na.omit(nlswork)
```

7. Descriptive statistics

```
summary(nlswork_no_na)
```

```
##      idcode           year          birth_yr          age
##  Min.   :   1   Min.   :70.00   Min.   :41.00   Min.   :16.0
##  1st Qu.:1280   1st Qu.:73.00   1st Qu.:46.00   1st Qu.:25.0
##  Median :2594   Median :78.00   Median :48.00   Median :30.0
##  Mean   :2589   Mean   :79.12   Mean   :48.11   Mean   :30.2
##  3rd Qu.:3859   3rd Qu.:83.00   3rd Qu.:51.00   3rd Qu.:35.0
##  Max.   :5159   Max.   :88.00   Max.   :54.00   Max.   :46.0
##      race                msp            nev_mar           grade
##  Length:13452       Min.   :0.0000   Min.   :0.0000   Min.   : 0.00
##  Class :character   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:12.00
##  Mode  :character   Median :1.0000   Median :0.0000   Median :12.00
##                     Mean   :0.6257   Mean   :0.2081   Mean   :12.68
##                     3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:14.00
##                     Max.   :1.0000   Max.   :1.0000   Max.   :18.00
##     collgrad         not_smsa         c_city           south
##  Min.   :0.0000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.000   Median :0.0000   Median :0.0000
##  Mean   :0.1887   Mean   :0.284   Mean   :0.3417   Mean   :0.4081
##  3rd Qu.:0.0000   3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.000   Max.   :1.0000   Max.   :1.0000
##     ind_code         occ_code          union            wks_ue
##  Min.   : 1.000   Min.   : 1.000   Min.   :0.0000   Min.   : 0.000
##  1st Qu.: 5.000   1st Qu.: 3.000   1st Qu.:0.0000   1st Qu.: 0.000
##  Median : 7.000   Median : 3.000   Median :0.0000   Median : 0.000
##  Mean   : 7.842   Mean   : 4.839   Mean   :0.2286   Mean   : 2.112
##  3rd Qu.:11.000   3rd Qu.: 6.000   3rd Qu.:0.0000   3rd Qu.: 0.000
##  Max.   :12.000   Max.   :13.000   Max.   :1.0000   Max.   :75.000
##     ttl_exp           tenure            hours          wks_work
##  Min.   : 0.000   Min.   : 0.0000   Min.   :  1.0   Min.   :  0.00
##  1st Qu.: 3.417   1st Qu.: 0.8333   1st Qu.: 35.0   1st Qu.: 43.00
##  Median : 5.635   Median : 2.0833   Median : 40.0   Median : 52.00
##  Mean   : 6.773   Mean   : 3.4475   Mean   : 36.2   Mean   : 50.73
##  3rd Qu.: 9.263   3rd Qu.: 4.5000   3rd Qu.: 40.0   3rd Qu.: 58.00
##  Max.   :28.885   Max.   :25.9167   Max.   :168.0   Max.   :103.00
##     ln_wage
##  Min.   :0.000
##  1st Qu.:1.397
##  Median :1.690
##  Mean   :1.714
##  3rd Qu.:2.001
##  Max.   :5.264
```

```
summary(nlswork_no_na[,c("grade","union","ln_wage")])
```

```
##      grade           union           ln_wage
##  Min.   : 0.00   Min.   :0.0000   Min.   :0.000
##  1st Qu.:12.00   1st Qu.:0.0000   1st Qu.:1.397
##  Median :12.00   Median :0.0000   Median :1.690
##  Mean   :12.68   Mean   :0.2286   Mean   :1.714
##  3rd Qu.:14.00   3rd Qu.:0.0000   3rd Qu.:2.001
```

```
##  Max.   :18.00   Max.   :1.0000   Max.   :5.264
```

```
str(nlswork_no_na)
```

```
## 'data.frame':    13452 obs. of  21 variables:
##  $ idcode  : num  1 1 1 1 1 1 2 2 2 2 ...
##  $ year    : num  72 77 80 85 87 88 71 77 78 83 ...
##  $ birth_yr: num  51 51 51 51 51 51 51 51 51 51 ...
##  $ age     : num  20 25 28 33 35 37 19 25 26 31 ...
##  $ race    : chr  "black" "black" "black" "black" ...
##  $ msp     : num  1 0 0 0 0 0 1 1 1 1 ...
##  $ nev_mar : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ grade   : num  12 12 12 12 12 12 12 12 12 12 ...
##  $ collgrad: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ not_smsa: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ c_city  : num  1 1 1 1 0 0 1 1 1 1 ...
##  $ south   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ ind_code: num  4 12 5 5 5 5 4 4 4 4 ...
##  $ occ_code: num  6 8 6 6 6 6 3 6 6 6 ...
##  $ union   : num  1 0 1 1 1 1 0 1 1 1 ...
##  $ wks_ue  : num  0 0 0 0 0 0 19 0 0 12 ...
##  $ ttl_exp : num  2.26 3.78 5.29 7.16 8.99 ...
##  $ tenure  : num  0.917 1.5 1.833 1.917 3.917 ...
##  $ hours   : num  40 32 45 42 45 48 40 40 40 38 ...
##  $ wks_work: num  51 52 75 97 95 70 13 52 52 37 ...
##  $ ln_wage : num  1.59 1.78 2.55 2.61 2.54 ...
##  - attr(*, "na.action")= 'omit' Named int [1:15082] 1 2 4 5 7 9 14 15 16 19 ...
##   ..- attr(*, "names")= chr [1:15082] "1" "2" "4" "5" ...
```

7.1. Export descriptive statistics table to html, with 2 digits

Shorter statistics

Statistic N Mean St. Dev. Min Max

age 13,452 30.20 6.41 16 46
collgrad 13,452 0.19 0.39 0 1
ttl_exp 13,452 6.77 4.41 0.00 28.88 union 13,452 0.23 0.42 0 1
hours 13,452 36.20 10.03 1 168

7.2. Export descriptive statistics table to txt, with 3 digits

Shorter statistics

Statistic N Mean St. Dev. Min Max

age 13,452 30.203 6.414 16 46
collgrad 13,452 0.189 0.391 0 1
ttl_exp 13,452 6.773 4.409 0.000 28.885 union 13,452 0.229 0.420 0 1
hours 13,452 36.199 10.034 1 168

7.3. Transposing the descriptive statistics table

Shorter statistics

Statistic age collgrad ttl_exp union hours

N 13,452 13,452 13,452 13,452 13,452 Mean 30.203 0.189 6.773 0.229 36.199 St. Dev. 6.414 0.391 4.409 0.420 10.034 Min 16 0 0.000 0 1
Max 46 1 28.885 1 168

7.4. Export to pdf

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: qua, out 11, 2023 - 20:10:10

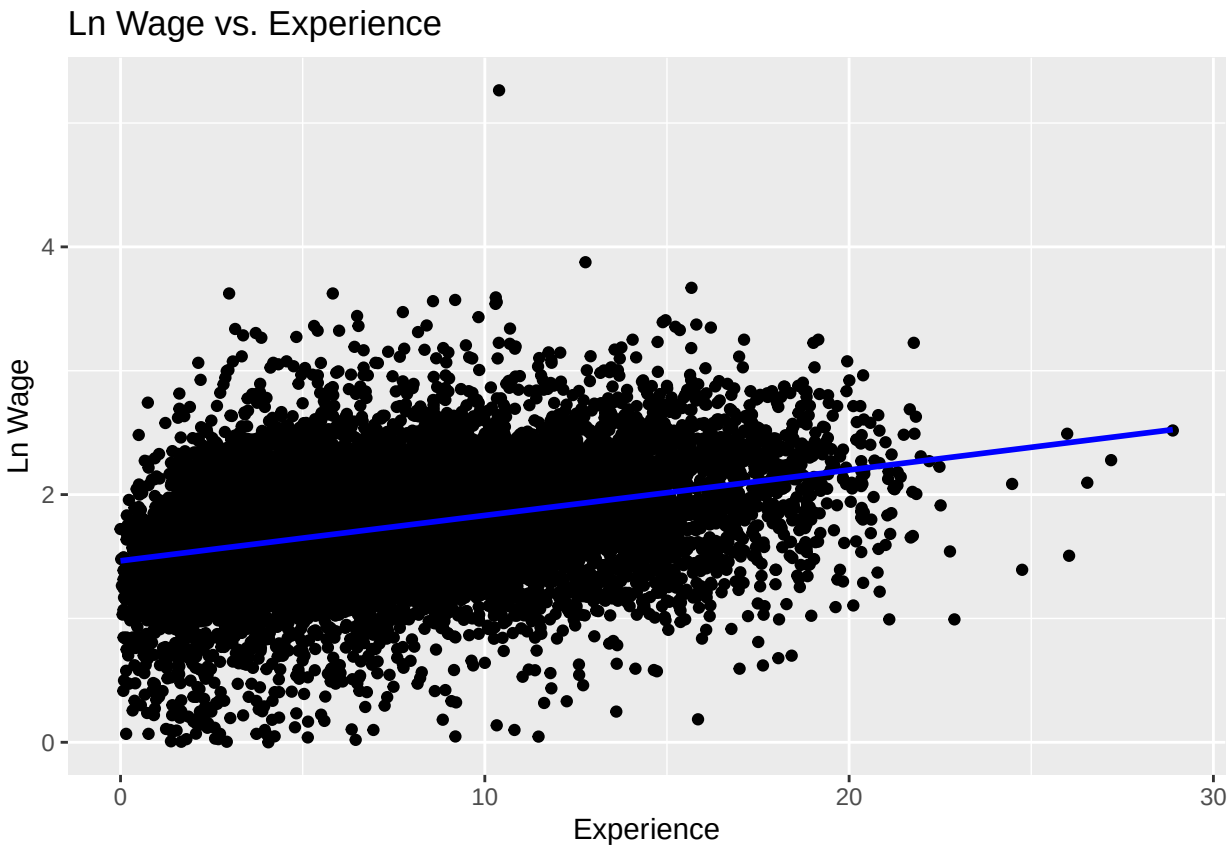
Table 1: Shorter statistics

Statistic	age	collgrad	ttl_exp	union	hours
N	13,452	13,452	13,452	13,452	13,452
Mean	30.203	0.189	6.773	0.229	36.199
St. Dev.	6.414	0.391	4.409	0.420	10.034
Min	16	0	0.000	0	1
Max	46	1	28.885	1	168

8. Visualisation to explore your data

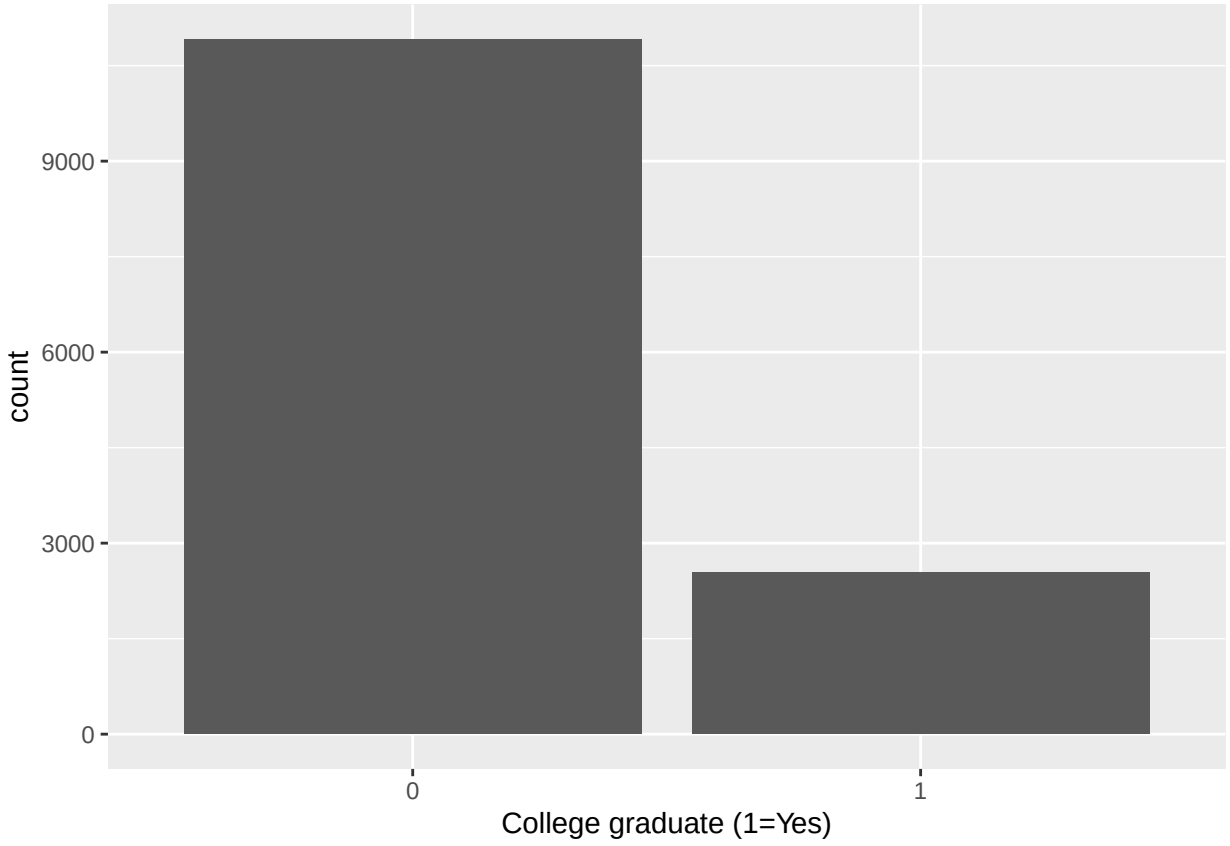
8.1. Relationships Between Continuous Variables

```
## `geom_smooth()` using formula = 'y ~ x'
```



8.2. Categorical variable

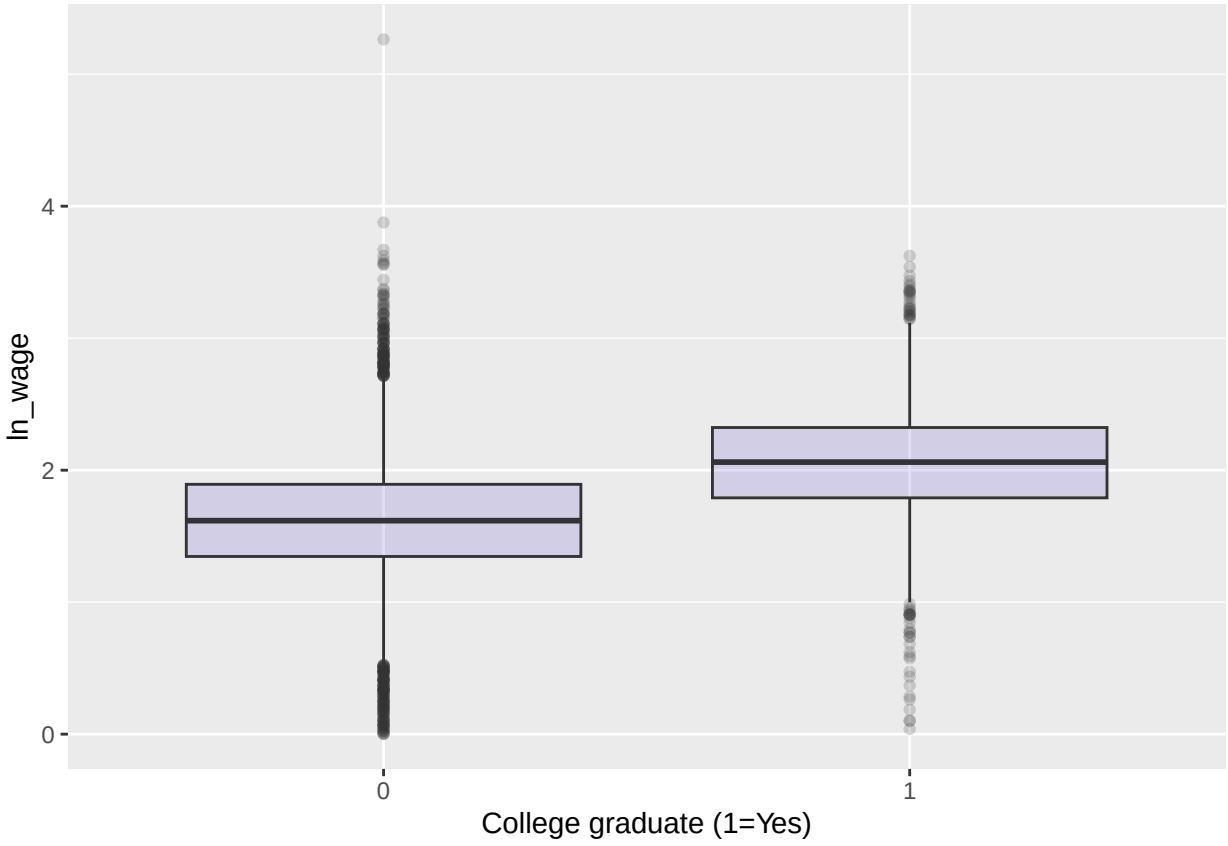
```
ggplot(data = nlswork_no_na) +
  geom_bar(mapping=aes(x=as.factor(collgrad))) +
  xlab("College graduate (1=Yes)")
```



8.3. Continuous Variable Distributions

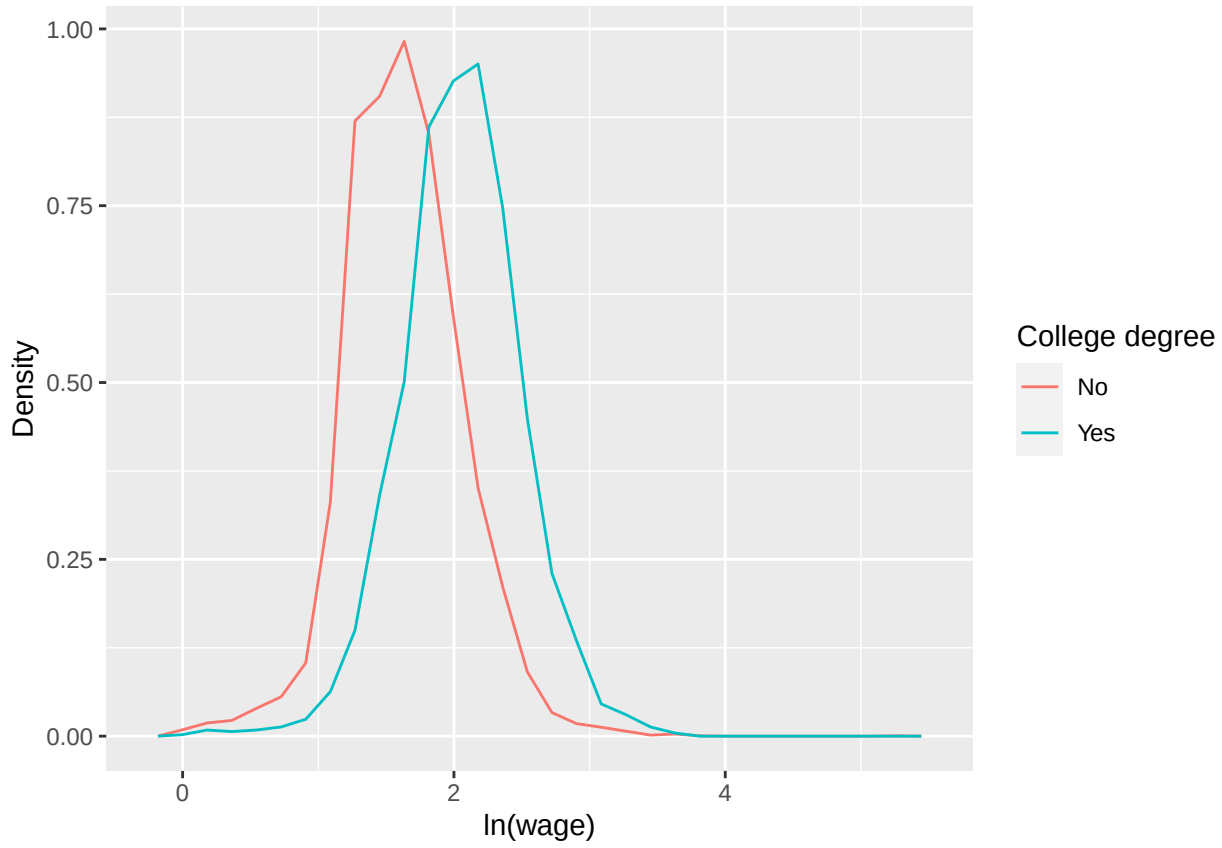
8.4 Categorical and continuous variables

```
nlswork_no_na %>% ggplot(aes(x=as.factor(collgrad), y=ln_wage)) +
  geom_boxplot(fill="slateblue", alpha=0.2) +
  xlab("College graduate (1=Yes)")
```



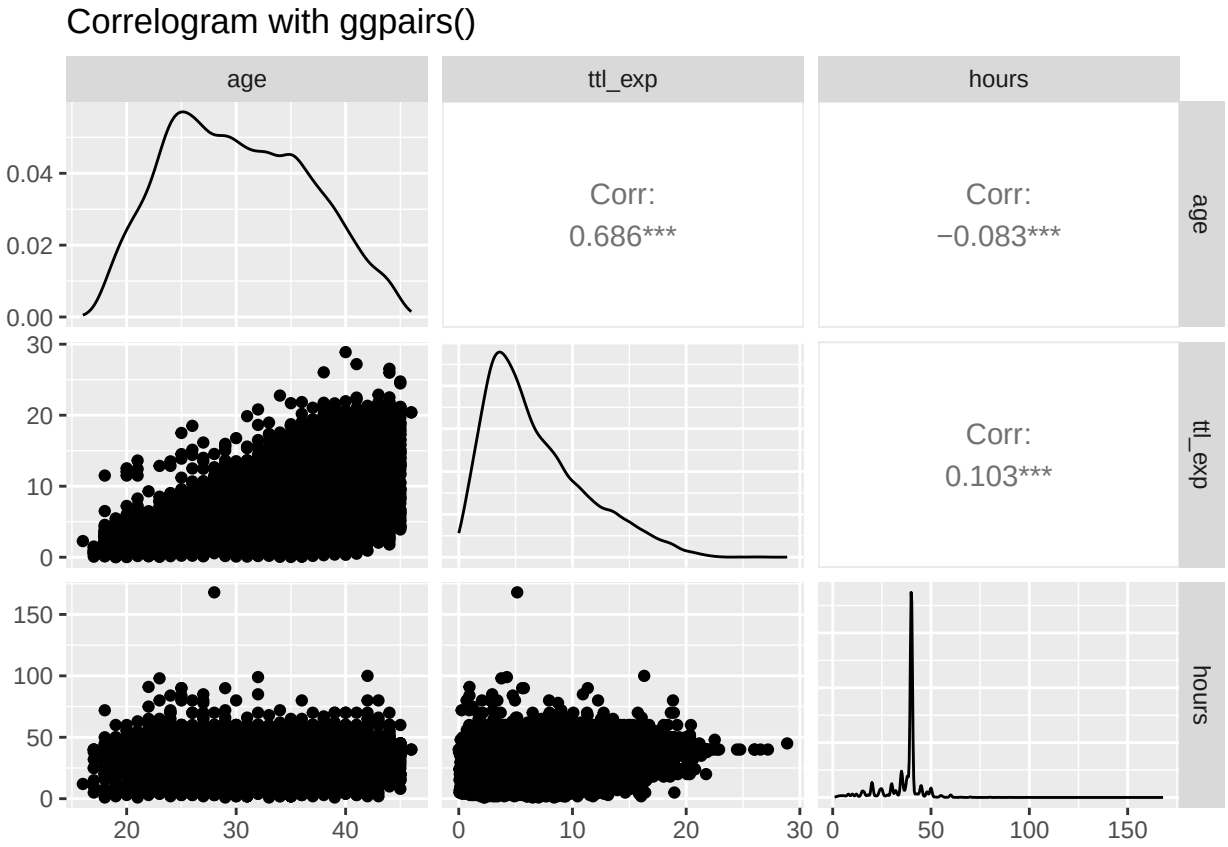

```
nlswork_no_na %>% ggplot(mapping = aes(x = ln_wage, y = ..density..)) +
    xlab("ln(wage)") +
    ylab("Density") +
    geom_freqpoly(mapping = aes(colour = factor(collgrad, labels=c("No", "Yes")))) +
  labs(color ="College degree")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



9. Correlation

```
ggpairs(nlswork_no_na[, c("age","ttl_exp","hours")], title="Correlogram with ggpairs()")
```



10. Assessment

Problem 1: Data Importing

Import the "card" dataset.

```
#BEGIN SOLUTION

#END SOLUTION
```

Problem 2: Visualizing Missing Data

Graphically show which variables have the most missing values.

```
#BEGIN SOLUTION

#END SOLUTION
```

Problem 3: Handling Missing Data

Adopt a strategy to handle the missing values.How many observations were lost?

```
#BEGIN SOLUTION

#END SOLUTION
```

Problem 4: Descriptive Statistics after Missing Data Handling

Present statistics of the dataset that has been treated for missing values.

```
#BEGIN SOLUTION

#END SOLUTION
```

Problem 5: Relationship Visualization

Graphically show the relationship between age and salary.Does the relationship between the variables make sense?

```
#BEGIN SOLUTION

#END SOLUTION
```

Problem 6: Age Distribution

Display the distribution of age.

```
#BEGIN SOLUTION

#END SOLUTION
```

Problem 7: Correlation

What is the correlation value between age and salary?

```
#BEGIN SOLUTION

#END SOLUTION
```

Problem 8:

In the nlswork_no_na dataset, can you identify any patterns or trends in the data related to unionized workers and their salaries?

```
#BEGIN SOLUTION

#END SOLUTION
```