(Fast) Introduction to R

Jump into a notebook

Joana Cima

03 dezembro 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

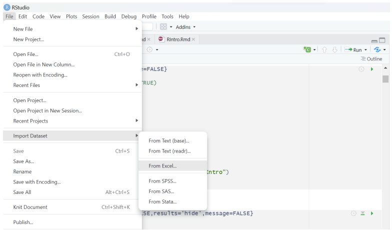


Figure 1: Point and click

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

head(nlswork)
```

```
##   idcode year birth_yr age  race msp nev_mar grade collgrad not_smsa c_city
## 1      1   70       51  18 black   0       1    12        0        0      1
## 2      1   71       51  19 black   1       0    12        0        0      1
## 3      1   72       51  20 black   1       0    12        0        0      1
## 4      1   73       51  21 black   1       0    12        0        0      1
## 5      1   75       51  23 black   1       0    12        0        0      1
## 6      1   77       51  25 black   0       0    12        0        0      1
##   south ind_code occ_code union wks_ue  ttl_exp     tenure hours wks_work
## 1     0        6        3    NA      2 1.083333 0.08333334    20       27
## 2     0        4        6    NA     22 1.275641 0.08333334    44       10
## 3     0        4        6     1      0 2.256410 0.91666669    40       51
## 4     0        4        6    NA      0 2.314102 0.08333334    40        3
## 5     0        5        6    NA      0 2.775641 0.16666667    10       24
## 6     0       12        8     0      0 3.775641 1.50000000    32       52
##    ln_wage
## 1 1.451214
## 2 1.028620
## 3 1.589977
## 4 1.780273
## 5 1.777012
## 6 1.778681
```

```
colnames(nlswork)
```

```
##  [1] "idcode"   "year"     "birth_yr" "age"      "race"     "msp"     
##  [7] "nev_mar"  "grade"    "collgrad" "not_smsa" "c_city"   "south"   
## [13] "ind_code" "occ_code" "union"    "wks_ue"   "ttl_exp"  "tenure"  
## [19] "hours"    "wks_work" "ln_wage"
```

```
str(nlswork)
```

```
## 'data.frame':    28534 obs. of  21 variables:
##  $ idcode  : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ year    : num  70 71 72 73 75 77 78 80 83 85 ...
##  $ birth_yr: num  51 51 51 51 51 51 51 51 51 51 ...
##  $ age     : num  18 19 20 21 23 25 26 28 31 33 ...
##  $ race    : chr  "black" "black" "black" "black" ...
##  $ msp     : num  0 1 1 1 1 0 0 0 0 0 ...
##  $ nev_mar : num  1 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 1 1 1 1 1 1 ...
##  $ south   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ ind_code: num  6 4 4 4 5 12 5 5 5 5 ...
##  $ occ_code: num  3 6 6 6 6 8 6 6 6 6 ...
##  $ union   : num  NA NA 1 NA NA 0 NA 1 1 1 ...
##  $ wks_ue  : num  2 22 0 0 0 0 7 0 NA 0 ...
##  $ ttl_exp : num  1.08 1.28 2.26 2.31 2.78 ...
##  $ tenure  : num  0.0833 0.0833 0.9167 0.0833 0.1667 ...
##  $ hours   : num  20 44 40 40 10 32 52 45 49 42 ...
##  $ wks_work: num  27 10 51 3 24 52 4 75 101 97 ...
##  $ ln_wage : num  1.45 1.03 1.59 1.78 1.78 ...
```

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 > 20)
```

4.4. Mutate: create variables

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

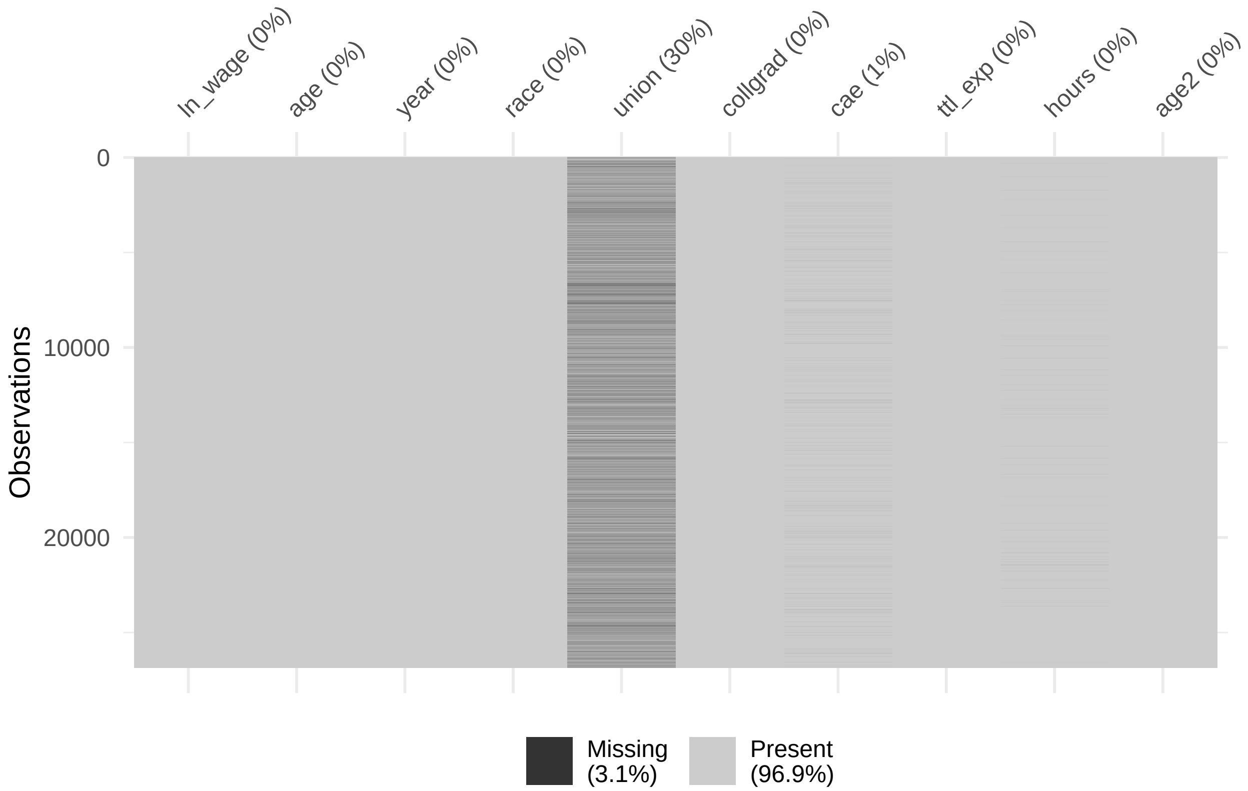
4.5. Manipulate the data in a single sequence

```
nlswork_new <- nlswork %>%
  rename(cae = ind_code) %>%
  select(ln_wage, age, year, race, union, collgrad, cae, ttl_exp, hours ) %>%
  filter(age>=20) %>%
  mutate(age2=age^2)
```

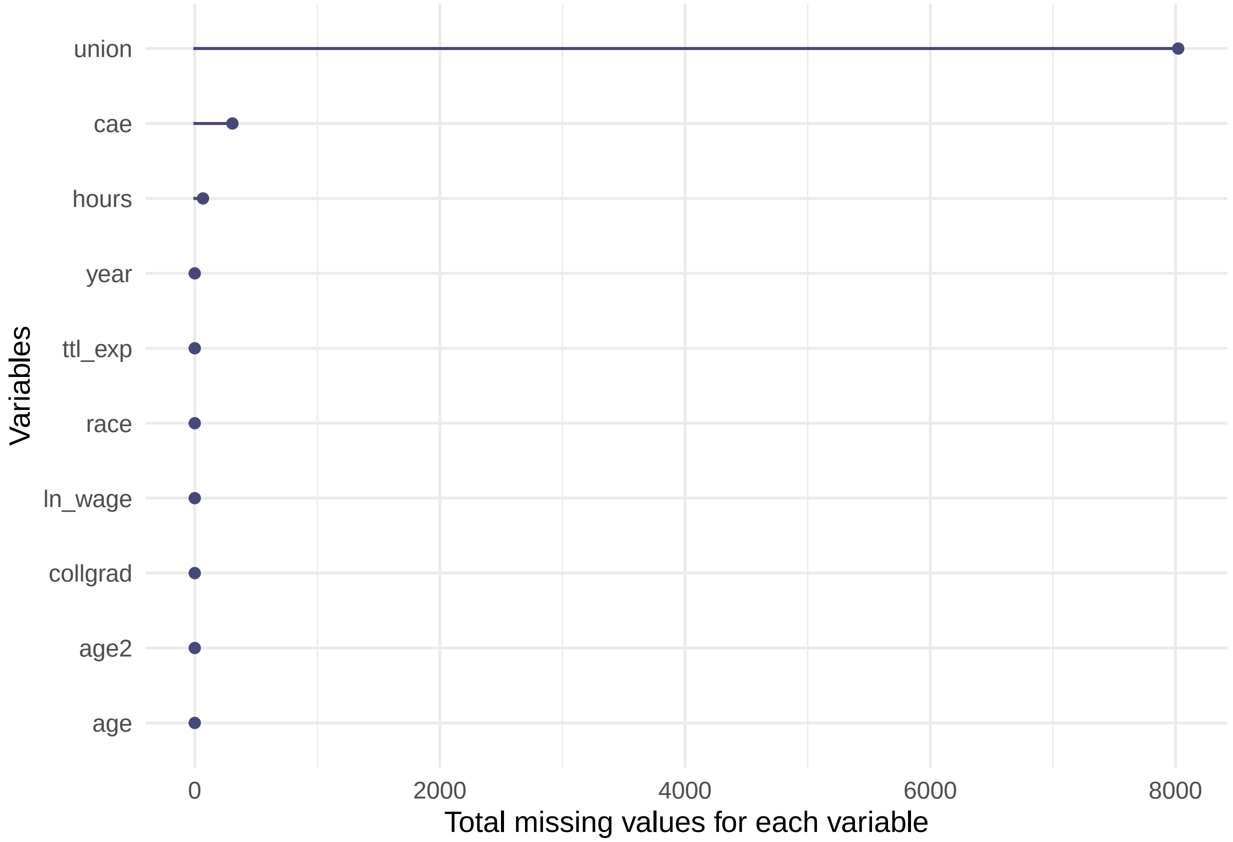
5. Detecting and Handling Missing Data

5.1 Detect Missing Data

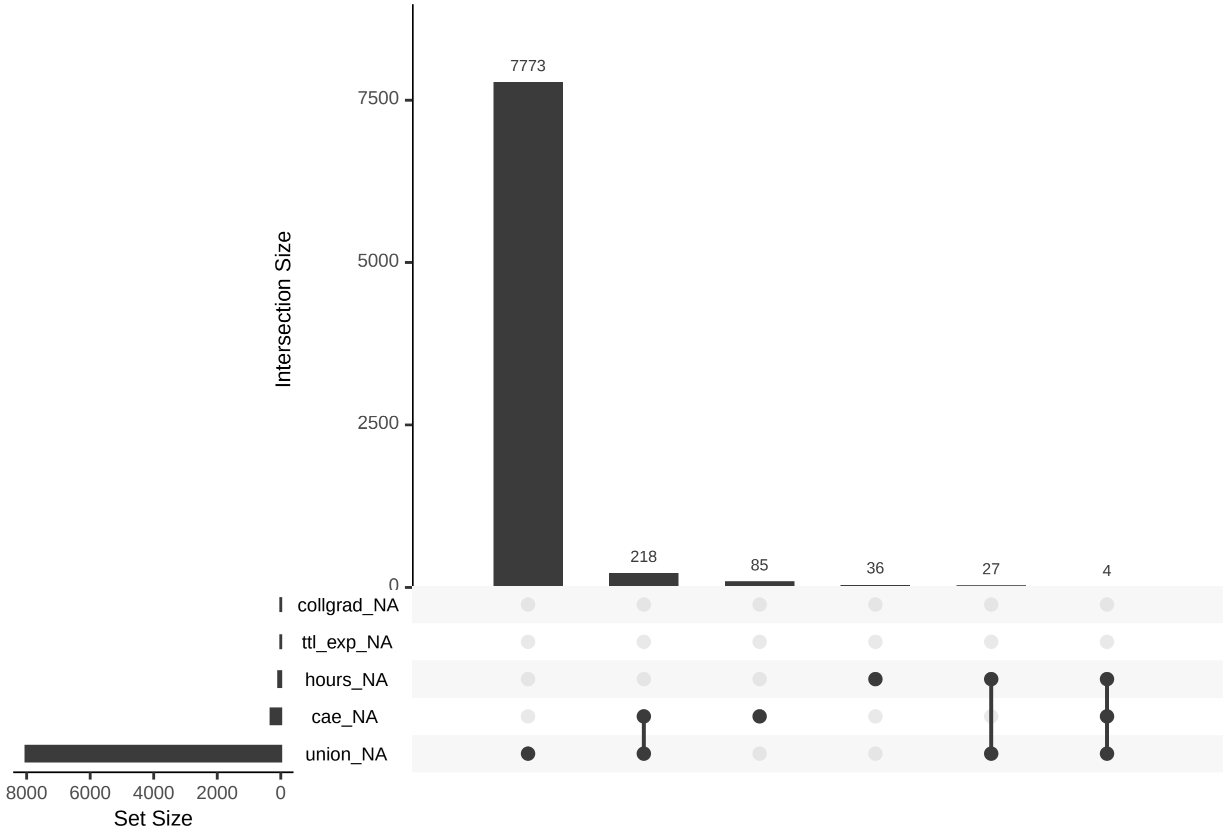
```
vis_miss(nlswork_new)
```



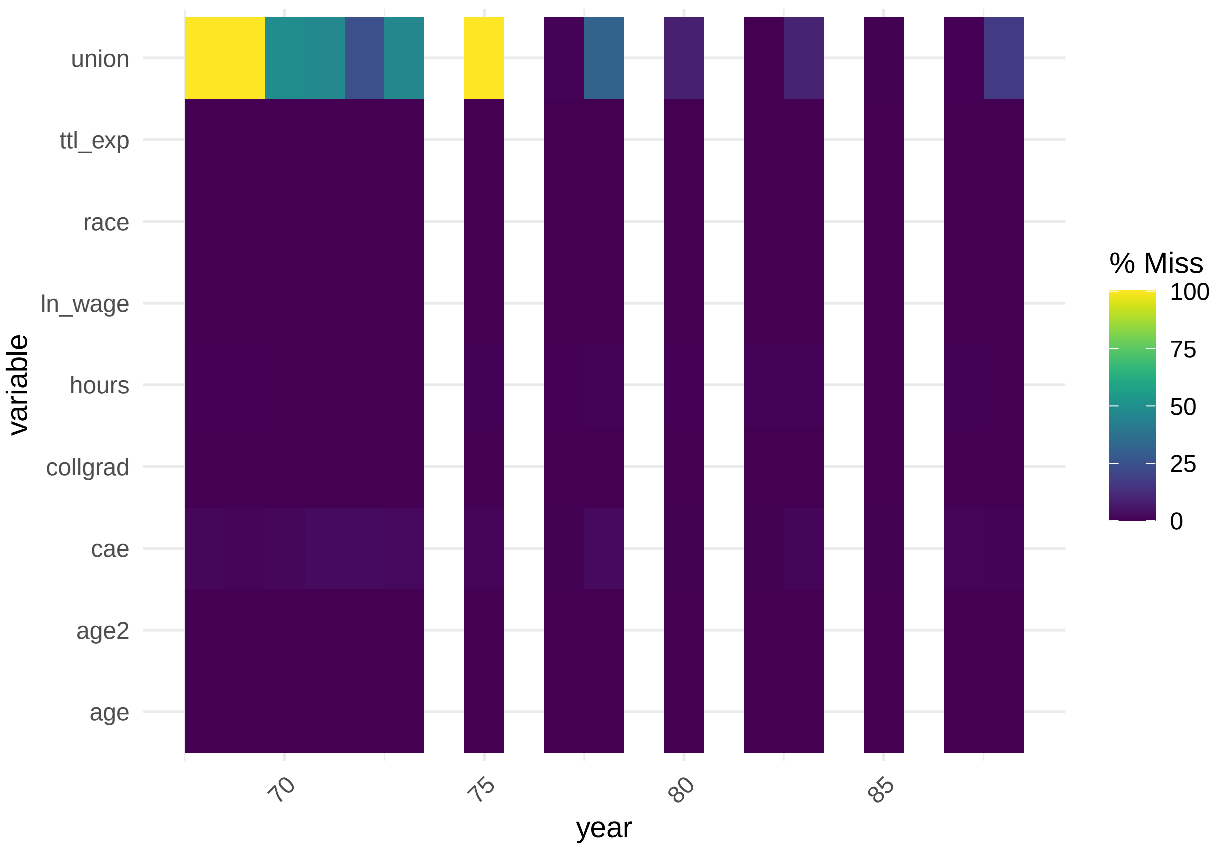
```r
gg_miss_var(nlswork_new) + labs(y = "Total missing values for each variable")
```



```r
gg_miss_upset(nlswork_new)
```

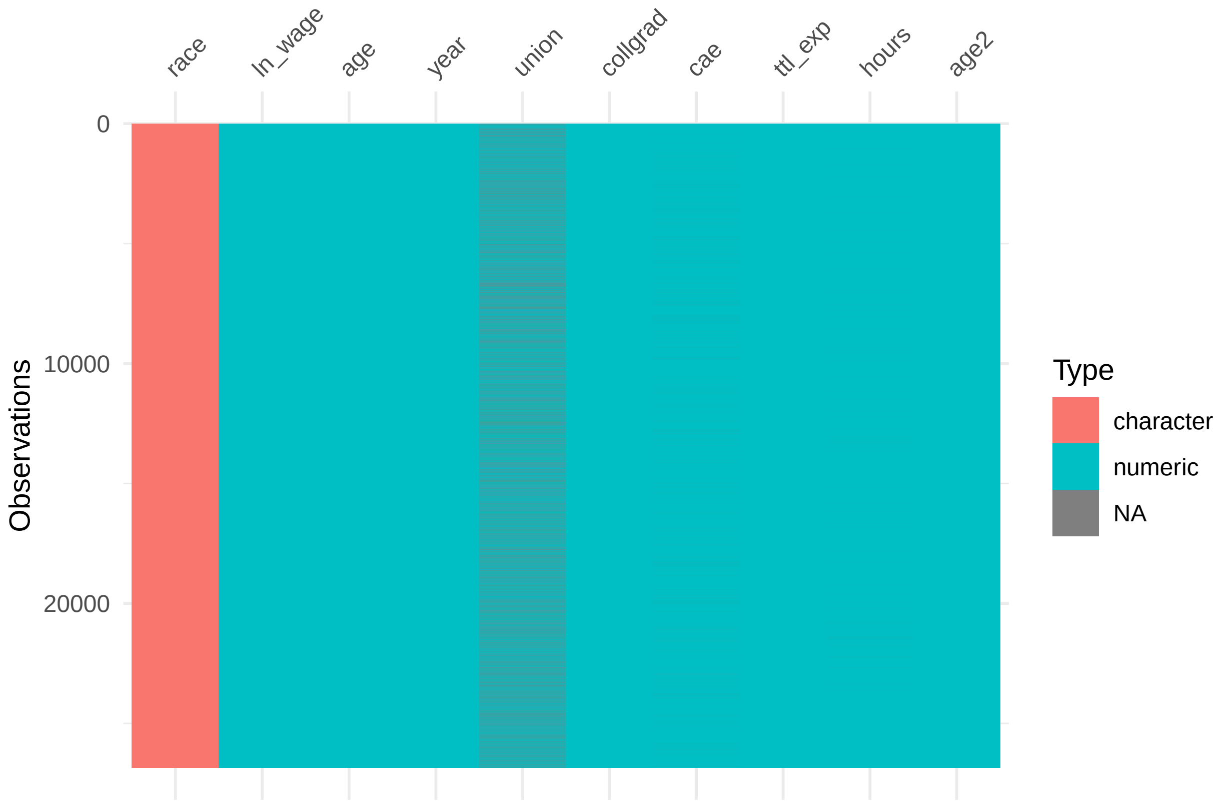


```r
gg_miss_fct(x = nlswork_new,fct = year)
```



Alternative

```
vis_dat(nlswork_new)
```



5.2. 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.

5.2.1 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.

```
library(tidyverse)
# Filling Missing Data

## (with the average - this is an example - it does not make sense in this case)
nlswork_filled <- nlswork %>%
  mutate(across(c("union"), ~ ifelse(is.na(.), mean(., na.rm = TRUE), .)))

## (with the mode)
    ### Create a function to compute mode

mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}

nlswork_filled2 <- nlswork
```

```
union_mode <- mode(nlswork$union[!is.na(nlswork$union)])
nlswork_filled2$union[is.na(nlswork$union)] <- union_mode
```

5.2.2 Excluding rows with missing data

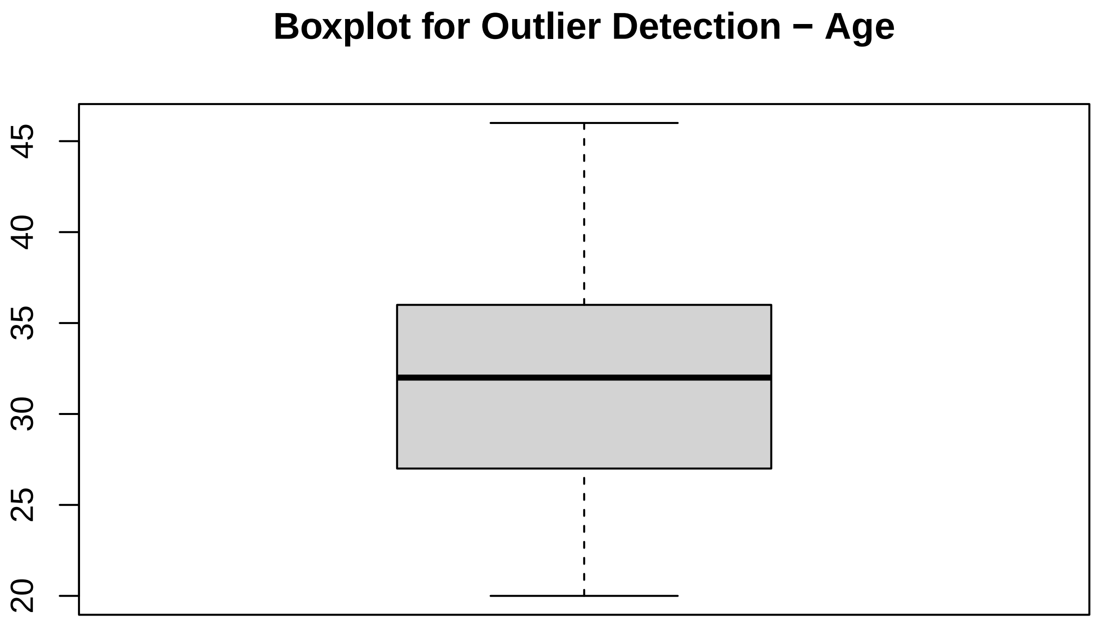
```
nlswork_no_na <- na.omit(nlswork_new)
```

6 Detecting and Handling Outliers

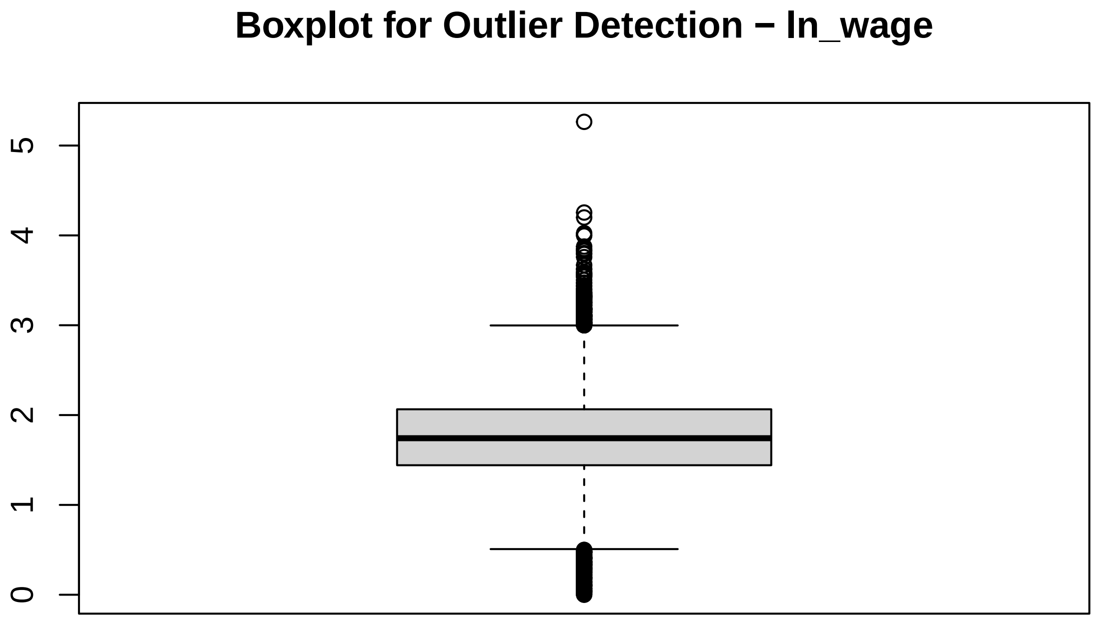
6.1 Detecting Outliers

6.1.1. Using Boxplot (example: age and ln_wage)

```
boxplot(nlswork_no_na$age, main="Boxplot for Outlier Detection - Age")
```



```
boxplot(nlswork_no_na$ln_wage, main="Boxplot for Outlier Detection - ln_wage")
```



### 6.1.2. Detecting Outliers using "identify_outliers" (example: ln_wage)

```
outliers <- identify_outliers(as.data.frame(nlswork_no_na$ln_wage))
extreme_outliers <- outliers[outliers$is.extreme, ]
extreme_outliers
```

```
##     nlswork_no_na$ln_wage is.outlier is.extreme
## 73               4.025415       TRUE       TRUE
## 149              4.005049       TRUE       TRUE
## 159              5.263916       TRUE       TRUE
## 235              4.254619       TRUE       TRUE
## 249              3.997510       TRUE       TRUE
## 254              4.199647       TRUE       TRUE
```

6.2 Handling Outliers

6.2.1 Removing the outliers from the original dataframe nlswork_no_na

```
extreme_values_to_remove <- extreme_outliers$`nlswork_no_na$ln_wage`
nlswork_no_outliers <- nlswork_no_na[!nlswork_no_na$ln_wage %in% extreme_values_to_remove, ]
```

6.2.2 Replacing the outliers using winsorize

```
nlswork_no_na$ln_wage_winsorized <- Winsorize(nlswork_no_na$ln_wage,
                                              probs = c(0, 0.99))
```

7. Descriptive statistics

```
summary(nlswork_no_na)
```

```
##     ln_wage           age             year           race          
##  Min.   :0.000   Min.   :20.00   Min.   :70.00   Length:18703      
##  1st Qu.:1.442   1st Qu.:27.00   1st Qu.:77.00   Class :character  
##  Median :1.742   Median :32.00   Median :82.00   Mode  :character  
##  Mean   :1.763   Mean   :31.63   Mean   :80.57                     
##  3rd Qu.:2.065   3rd Qu.:36.00   3rd Qu.:85.00                     
##  Max.   :5.264   Max.   :46.00   Max.   :88.00                     
##      union           collgrad           cae            ttl_exp        
##  Min.   :0.0000   Min.   :0.0000   Min.   : 1.000   Min.   : 0.01923  
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 5.000   1st Qu.: 4.13462  
##  Median :0.0000   Median :0.0000   Median : 7.000   Median : 7.14103  
##  Mean   :0.2349   Mean   :0.1999   Mean   : 7.892   Mean   : 7.85462  
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:11.000   3rd Qu.:10.96795  
##  Max.   :1.0000   Max.   :1.0000   Max.   :12.000   Max.   :28.88462  
##      hours             age2      ln_wage_winsorized
##  Min.   :  1.00   Min.   : 400   Min.   :0.000     
##  1st Qu.: 35.00   1st Qu.: 729   1st Qu.:1.442     
##  Median : 40.00   Median :1024   Median :1.742     
##  Mean   : 36.77   Mean   :1036   Mean   :1.760     
##  3rd Qu.: 40.00   3rd Qu.:1296   3rd Qu.:2.065     
##  Max.   :168.00   Max.   :2116   Max.   :2.963
```

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

Shorter statistics

Statistic N Mean St. Dev. Min Max

age 18,703 31.63 5.96 20 46
collgrad 18,703 0.20 0.40 0 1
ttl_exp 18,703 7.85 4.54 0.02 28.88 union 18,703 0.23 0.42 0 1
hours 18,703 36.77 9.59 1 168 ln_wage_winsorized 18,703 1.76 0.46 0.00 2.96

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

Shorter statistics

Statistic N Mean St. Dev. Min Max

age 18,703 31.634 5.960 20 46
collgrad 18,703 0.200 0.400 0 1
ttl_exp 18,703 7.855 4.536 0.019 28.885 union 18,703 0.235 0.424 0 1
hours 18,703 36.772 9.586 1 168
ln_wage_winsorized 18,703 1.760 0.458 0.000 2.963

7.3. Transposing the descriptive statistics table

Shorter statistics

Statistic age collgrad ttl_exp union hours ln_wage_winsorized

N 18,703 18,703 18,703 18,703 18,703 18,703

Mean 31.634 0.200 7.855 0.235 36.772 1.760

St. Dev. 5.960 0.400 4.536 0.424 9.586 0.458

Min 20 0 0.019 0 1 0.000

Max 46 1 28.885 1 168 2.963

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: dom, dez 03, 2023 - 00:20:09

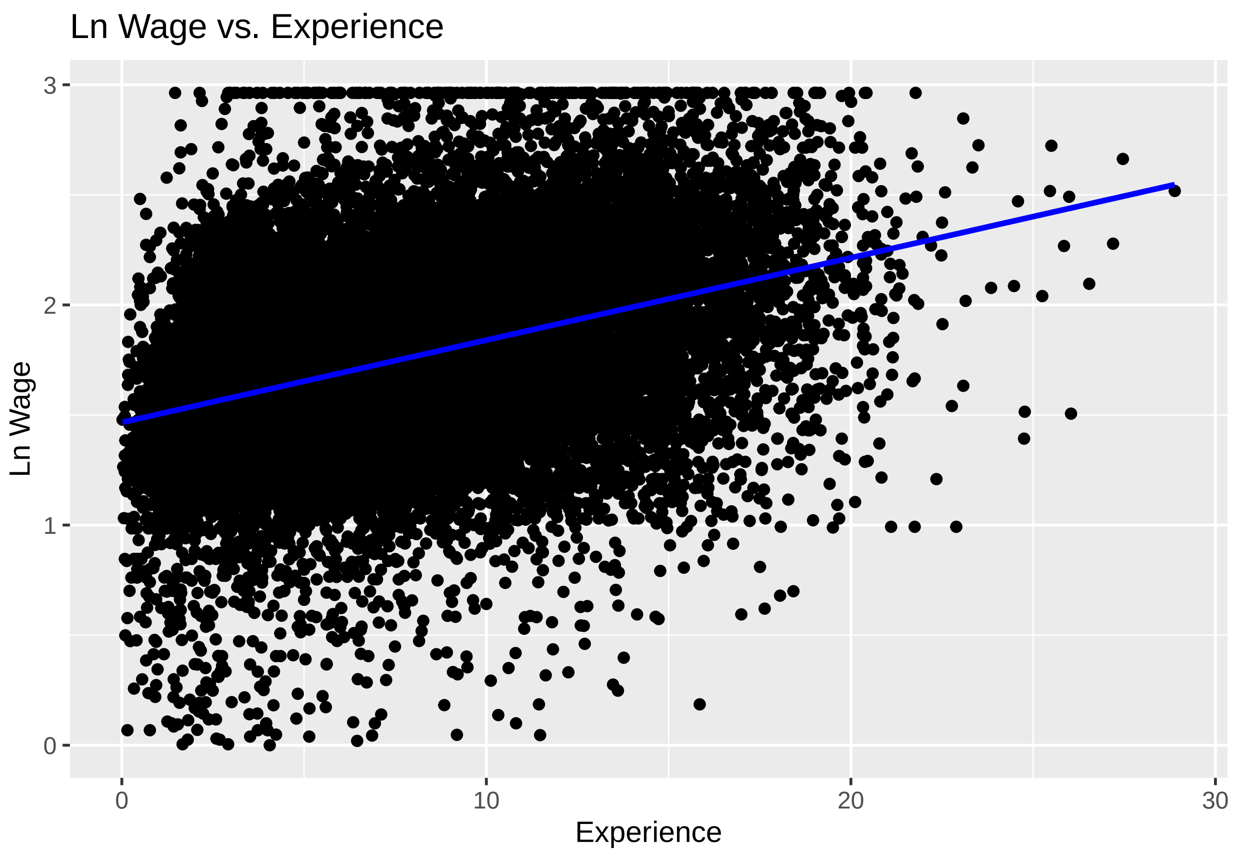
Table 1: Shorter statistics

Statistic	age	collgrad	ttl_exp	union	hours
N	18,703	18,703	18,703	18,703	18,703
Mean	31.634	0.200	7.855	0.235	36.772
St. Dev.	5.960	0.400	4.536	0.424	9.586
Min	20	0	0.019	0	1
Max	46	1	28.885	1	168

8. Visualisation to explore your data

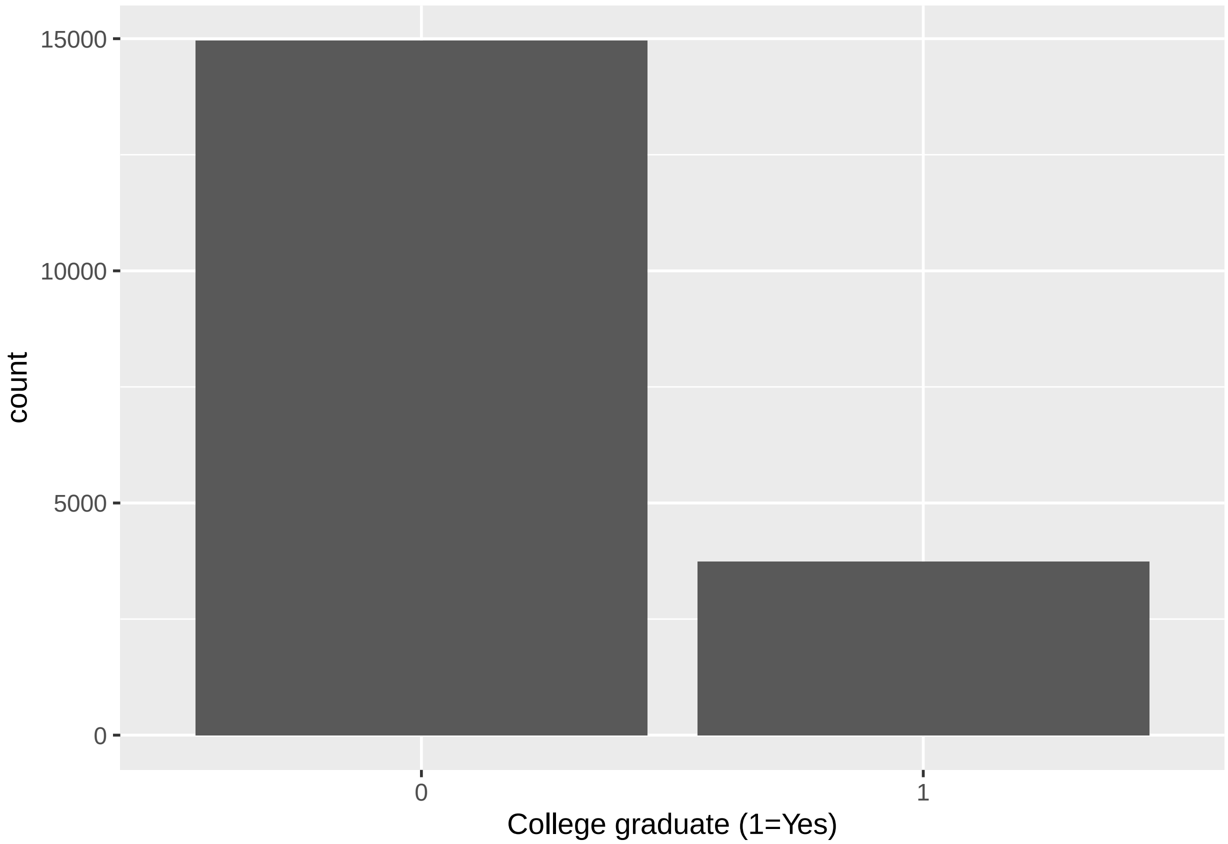
8.1. Relationship Between Continuous Variables

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



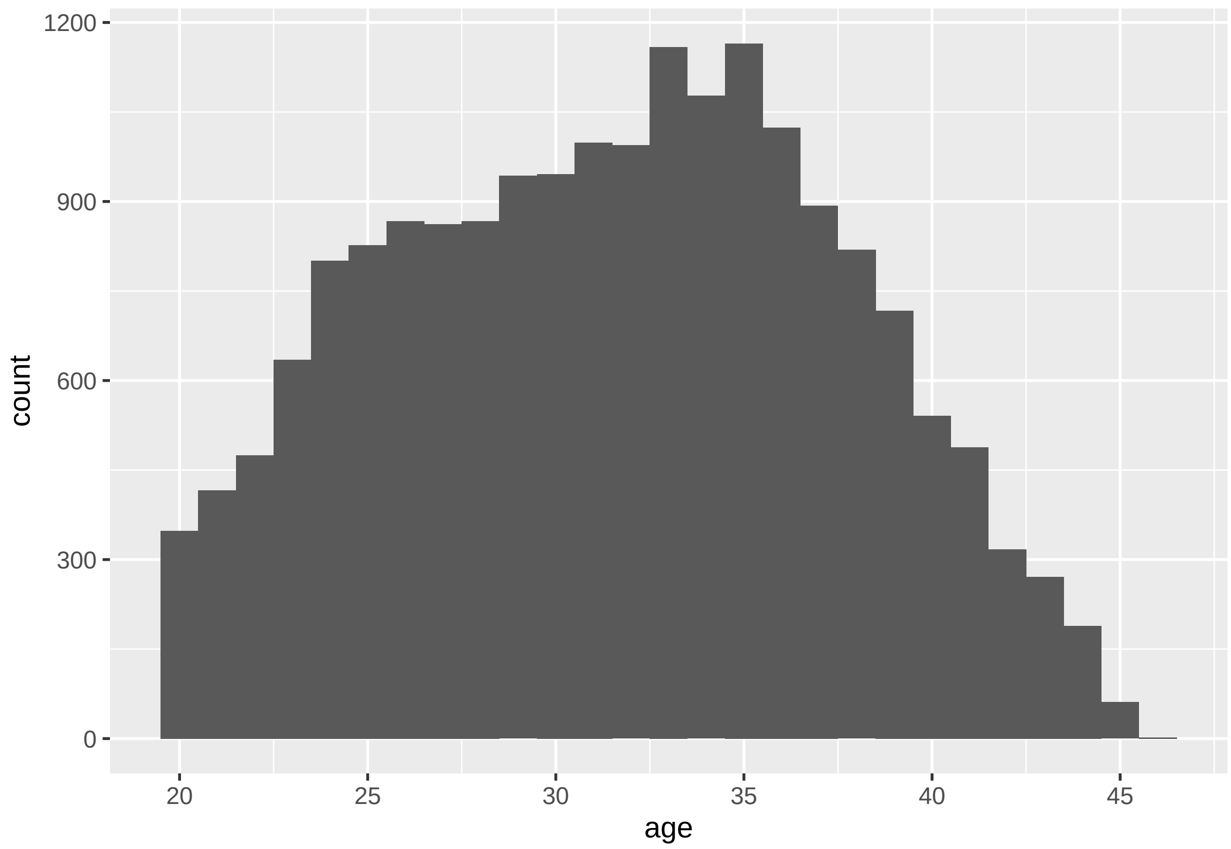
8.2. Categorical variable

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



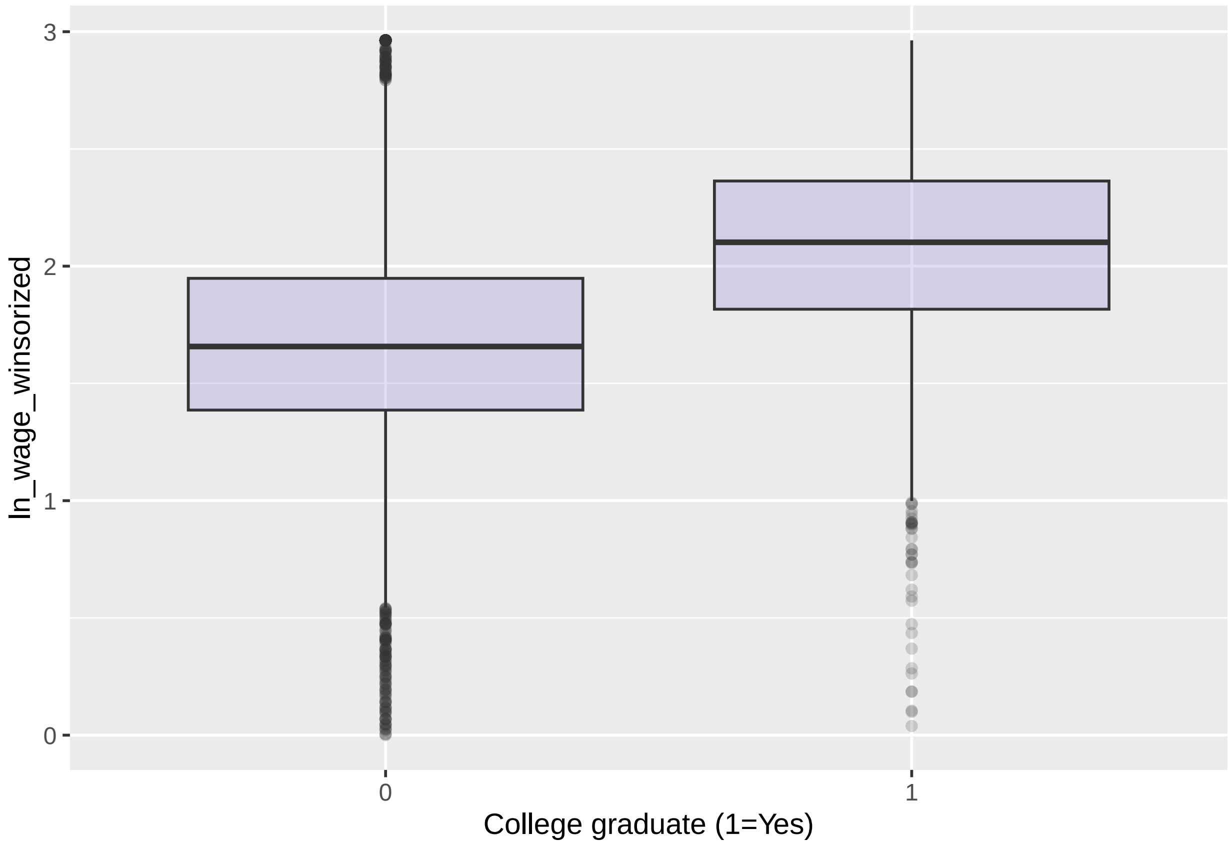
8.3. Continuous Variable Distributions

```
ggplot(data = nlswork_no_na) + geom_histogram(mapping = aes(x = age), binwidth = 1) +
  scale_x_continuous(breaks = seq(20, 50, by = 5))
```



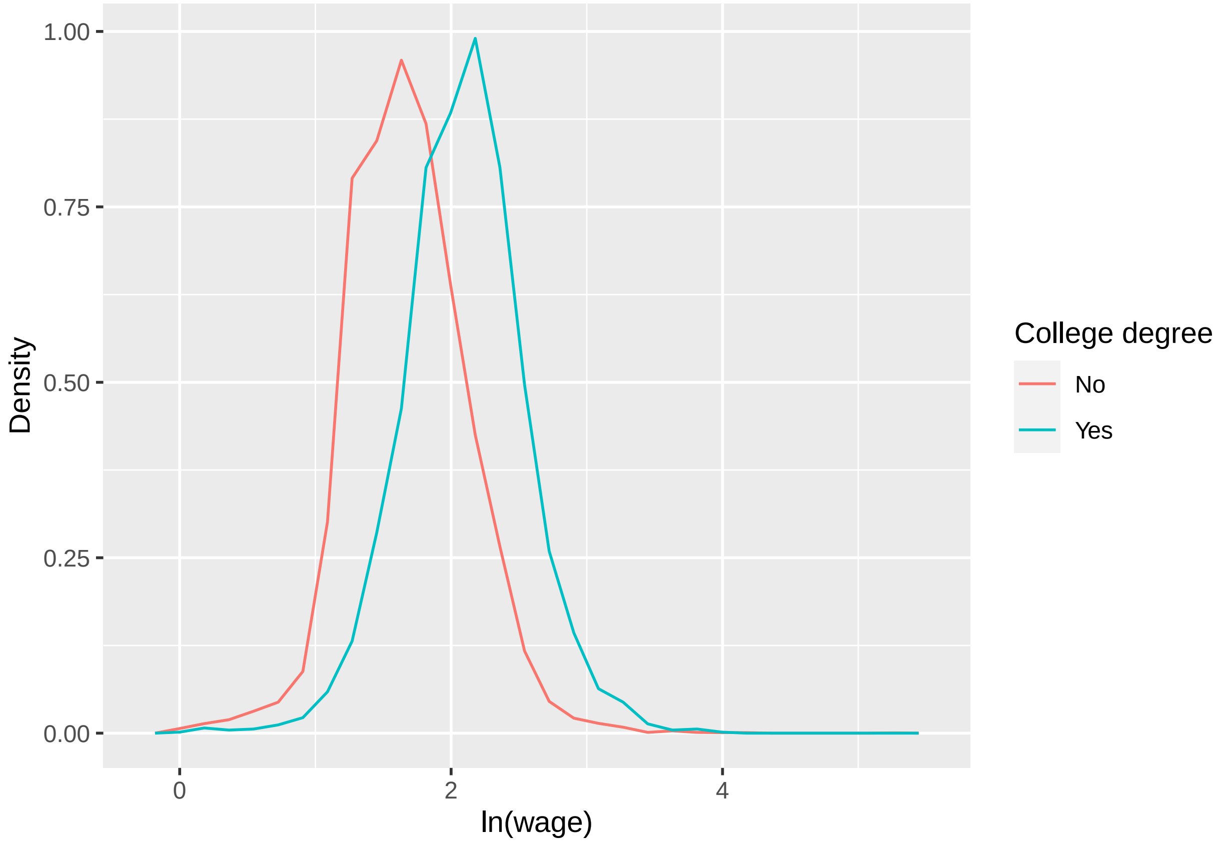
8.4 Categorical and continuous variables

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



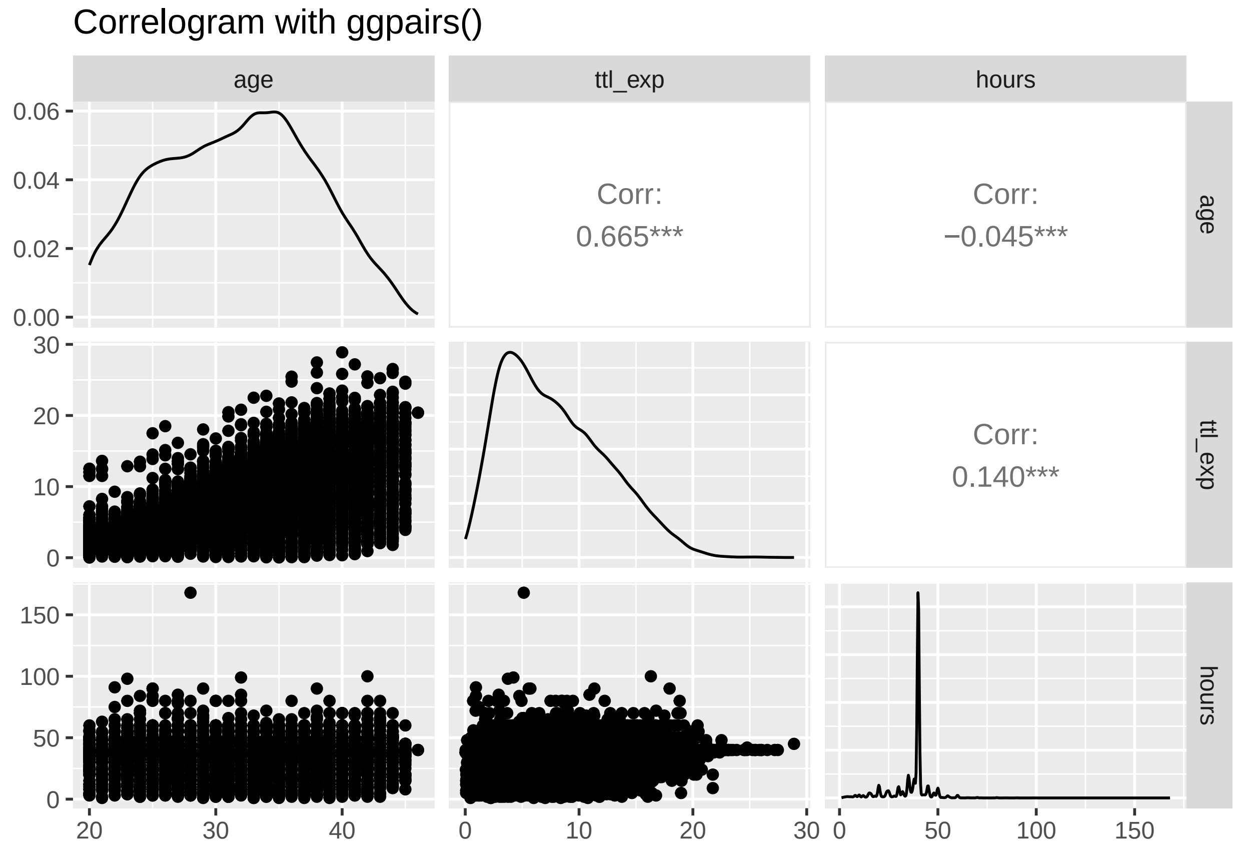
```r
nlswork_no_na %>% ggplot(mapping = aes(x = ln_wage, y = ..density..)) +
    xlab("ln(wage)") +
    ylab("Density") +
    geom_freqpoly(mapping = aes(colour = factor(collgrad, labels=c("No", "Yes")))) +
  labs(color ="College degree")
```

```
## `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()")
```


Correlogram with ggpairs()
age
ttl_exp
hours
Corr:
0.665***
Corr:
−0.045***
Corr:
0.140***
age
ttl_exp
hours
0.06
0.04
0.02
0.00
30
20
10
0
150
100
50
0
20
30
40
0
10
20
30
0
50
100
150

“‘