Bivariate analysis

2025-02-04

M1 MIDS/MFA/LOGOS
Université Paris Cité
Année 2024
Course Homepage
Moodle

> **Objectives**
>
> In Exploratory analysis of tabular data, bivariate analysis is the second step. It consists in exploring, summarizing, visualizing pairs of columns of a dataset.

Setup

```
stopifnot(
  require(glue),
  require(magrittr),
  require(lobstr),
  require(arrow),
  require(ggforce),
  require(vcd),
  require(ggmosaic),
  require(httr),
  require(patchwork),
  require(corrr),
  require(gapminder),
  require(slider),
  require(tidyverse)
)
```

Bivariate techniques depend on the types of columns we are facing.

For *numerical/numerical* samples

- Scatter plots
- Smoothed lineplots (for example linear regression)
- 2-dimensional density plots

For *categorical/categorical* samples : mosaicplots and variants

For *numerical/categorical* samples

- Boxplots per group

- Histograms per group
- Density plots per group

Dataset

Once again we rely on the `Recensement` dataset.

> Since 1948, the US Census Bureau carries out a monthly Current Population Survey, collecting data concerning residents aged above 15 from 150000 households. This survey is one of the most important sources of information concerning the american workforce. Data reported in file `Recensement.txt` originate from the 2012 census.

Load the data into the session environment and call it `df`. Take advantage of the fact that we saved the result of our data wrangling job in a self-documented file format. Download a `parquet` file from the following URL:

`https://stephane-v-boucheron.fr/data/Recensement.parquet`

Question

Download a `parquet` file from the following URL:
`https://stephane-v-boucheron.fr/data/Recensement.parquet`

- Use `httr::GET()` and `WriteBin()`
- Use `download.file()`
- Use `fs` to handle files and directories

Solution

Manage the `DATA` sub-directory

```
if (fs::dir_exists('DATA')){
  datapath <- "DATA"
} else {
  datapath <- "../DATA"
}
```

Using `httr::get` and `writeBin`

```
fname <- "Recensement.parquet"

fpath <- paste(datapath, fname, sep="/")

url <- 'https://stephane-v-boucheron.fr/data/Recensement.parquet'

if (!file.exists(fpath)) {
  tryCatch(
    expr = {

      rep <- httr::GET(url)
      stopifnot(rep$status_code==200)

      con <- file(fpath, open="wb")
      writeBin(rep$content, con)
      close(con)
    },
    warning = function(w) {
      glue("Successful download but {w}")
    },
    error = function(e) {
      stop("Houston, we have a problem!")    # error-handler-code
    },
    finally = {
      if (exists("con") && isOpen(con)){
        close(con)
      }
    }
  )
}
```

```
if (!file.exists(fpath)) {
  tryCatch(
    expr = {
      download.file(url, fpath, mode="wb", quiet=T)
      print(glue::glue('>>> file downloaded at {fpath}\n'))
    },
    warning = function(w) {
      glue::glue("Successful download but {w}")
    },
    error = function(e) {
      stop("Houston, we have a problem!")    # error-handler-code
```

Question

Load the data contained in the downloaded file into the session environment and call it `df`

Solution

```
df <- arrow::read_parquet(fpath)
```

```
df |>
  glimpse()
## Rows: 599
## Columns: 11
## $ AGE         <dbl> 58, 40, 29, 59, 51, 19, 64, 23, 47, 66, 26, 23, 54, 44, 56,~
## $ SEXE        <fct> F, M, M, M, M, M, F, F, M, F, M, F, F, F, F, F, M, M, F,~
## $ REGION      <fct> NE, W, S, NE, W, NW, S, NE, NW, S, NE, NE, W, NW, S, S, NW,~
## $ STAT_MARI   <fct> C, M, C, D, M, C, M, C, M, D, M, C, M, C, M, C, S, M, S, C,~
## $ SAL_HOR     <dbl> 13.25, 12.50, 14.00, 10.60, 13.00, 7.00, 19.57, 13.00, 20.1~
## $ SYNDICAT    <fct> non, non, non, oui, non, non, non, non, oui, non, non, non,~
## $ CATEGORIE   <fct> "Administration", "Building ", "Administration", "Services"~
## $ NIV_ETUDES <fct> "Bachelor", "12 years schooling, no diploma", "Associate de~
## $ NB_PERS     <fct> 2, 2, 2, 4, 8, 6, 3, 2, 3, 1, 3, 2, 6, 5, 4, 4, 3, 2, 3, 2,~
## $ NB_ENF      <fct> 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,~
## $ REV_FOYER   <fct> [35000-40000), [17500-20000), [75000-1e+05), [17500-20000),~

df |>
  head()
## # A tibble: 6 x 11
##     AGE SEXE  REGION STAT_MARI SAL_HOR SYNDICAT CATEGORIE     NIV_ETUDES NB_PERS
##   <dbl> <fct> <fct>  <fct>       <dbl> <fct>    <fct>         <fct>      <fct>
## 1    58 F     NE     C            13.2 non      "Administrat~ Bachelor   2
## 2    40 M     W      M            12.5 non      "Building "   12 years ~ 2
## 3    29 M     S      C            14   non      "Administrat~ Associate~ 2
## 4    59 M     NE     D            10.6 oui      "Services"    12 years ~ 4
## 5    51 M     W      M            13   non      "Services"    9 years s~ 8
## 6    19 M     NW     C             7   non      "Services"    12 years ~ 6
## # i 2 more variables: NB_ENF <fct>, REV_FOYER <fct>
```

Categorical/Categorical pairs

Question

Project the dataframe on categorical columns

Solution

```
df |>
  select(where(is.factor)) |>
  head()
```

```
# A tibble: 6 x 9
  SEXE  REGION STAT_MARI SYNDICAT CATEGORIE  NIV_ETUDES NB_PERS NB_ENF REV_FOYER
  <fct> <fct>  <fct>     <fct>    <fct>      <fct>      <fct>   <fct>  <fct>
1 F     NE     C         non      "Administ~ Bachelor   2       0      [35000-4~
2 M     W      M         non      "Building~ 12 years ~ 2       0      [17500-2~
3 M     S      C         non      "Administ~ Associate~ 2       0      [75000-1~
4 M     NE     D         oui      "Services" 12 years ~ 4       1      [17500-2~
5 M     W      M         non      "Services" 9 years s~ 8       1      [75000-1~
6 M     NW     C         non      "Services" 12 years ~ 6       0      [1e+05-1~
```

Question

- Explore the connection between `CATEGORIE` and `SEX`.
- Compute the 2-ways contingency table using `table()`, and `count()` from `dplyr`.

- Use `tibble::as_tibble()` to transform the output of `table()` into a dataframe/tibble.
- Use `tidyr::pivot_wider()` so as to obtain a wide (but messy) tibble with the same the same shape as the output of `table()`.
- Can you spot a difference?

Solution

```
tb <- df |>
  dplyr::select(CATEGORIE, SEXE) |>
  table()

# tb
```

```
tb2 <- df |>
  count(CATEGORIE, SEXE)

tb2
```

```
# A tibble: 18 x 3
   CATEGORIE                          SEXE      n
   <fct>                              <fct> <int>
 1 "Business, Management and Finance" F        23
 2 "Business, Management and Finance" M        23
 3 "Liberal profession"               F        82
 4 "Liberal profession"               M        51
 5 "Services"                         F        75
 6 "Services"                         M        50
 7 "Selling"                          F        30
 8 "Selling"                          M        18
 9 "Administration"                   F        72
10 "Administration"                   M        22
11 "Agriculture, Fishing, Forestry"   F         2
12 "Agriculture, Fishing, Forestry"   M         8
13 "Building "                        M        36
14 "Repair and maintenance"           M        32
15 "Production"                       F         9
16 "Production"                       M        30
17 "Commodities Transport"            F         4
18 "Commodities Transport"            M        32
```

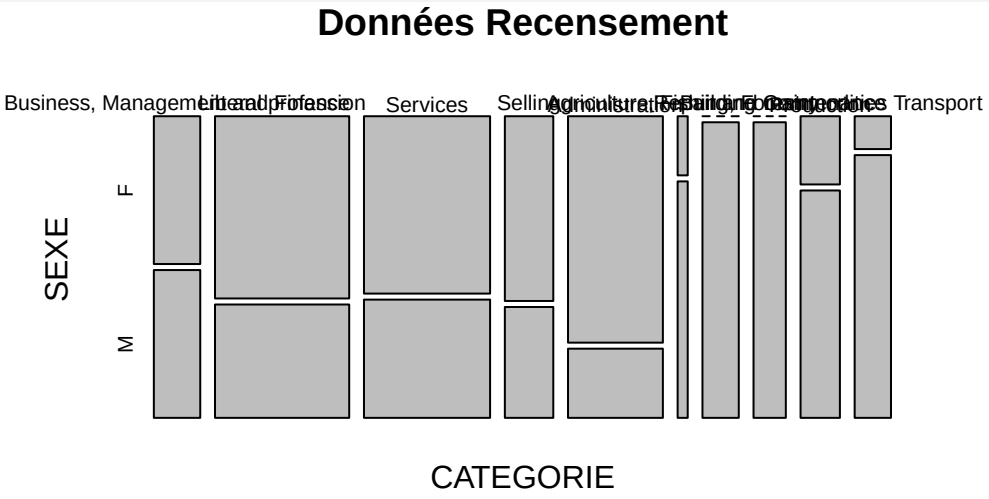
```
tb2 |>
  pivot_wider(id_cols=CATEGORIE,
              names_from=SEXE,
              values_from=n)
```

```
# A tibble: 10 x 3
   CATEGORIE                             F     M
   <fct>                             <int> <int>
 1 "Business, Management and Finance"    23    23
 2 "Liberal profession"                  82    51
 3 "Services"                            75    50
 4 "Selling"                             30    18
 5 "Administration"                      72    22
 6 "Agriculture, Fishing, Forestry"       2     8
 7 "Building "                           NA    36
 8 "Repair and maintenance"              NA    32
 9 "Production"                           9    30
10 "Commodities Transport"                4    32
```

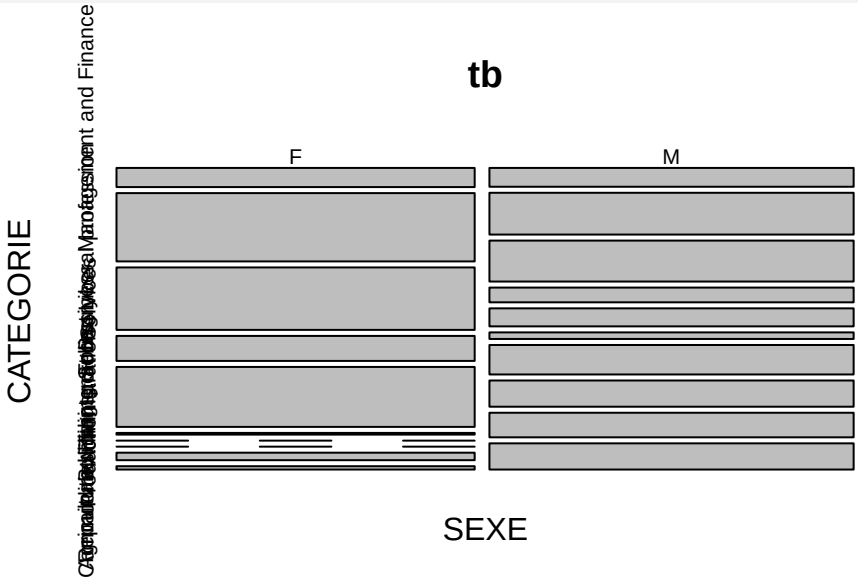
Question

Use `mosaicplot()` from base `R` to visualize the contingency table.

```
mosaicplot(~ CATEGORIE + SEXE,
           tb,
           main="Données Recensement")
```




```
mosaicplot(~ SEXE + CATEGORIE, tb)
```



Question

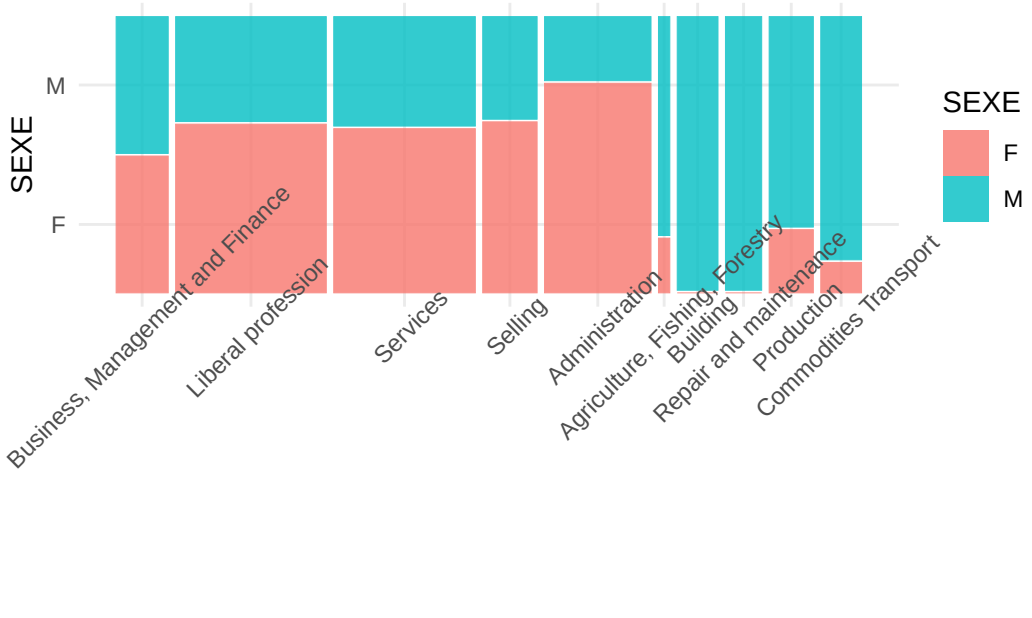
Use `geom_mosaic` from `ggmosaic` to visualize the contingency table

- Make the plot as readable as possible
- Reorder `CATEGORIE` acccording to counts

```
rot_x_text <- theme(
  axis.text.x = element_text(angle = 45)
)
```

```
df |>
  ggplot() +
  geom_mosaic(aes(x=product(SEXE, CATEGORIE), fill=SEXE)) +
  rot_x_text
```

```
Warning: The `scale_name` argument of `continuous_scale()` is deprecated as of ggplot2
3.5.0.
Warning: The `trans` argument of `continuous_scale()` is deprecated as of ggplot2 3.5.
i Please use the `transform` argument instead.
Warning: `unite_()` was deprecated in tidyr 1.2.0.
i Please use `unite()` instead.
i The deprecated feature was likely used in the ggmosaic package.
  Please report the issue at <https://github.com/haleyjeppson/ggmosaic>.
```



Question

- Collapse rare levels of `CATEGORIE` (consider that a level is rare if it has less than 40 occurrences). Use tools from `forcats`.

Solution

```
df |>
  count(CATEGORIE) |>
  arrange(desc(n))
```

```
# A tibble: 10 x 2
   CATEGORIE                              n
   <fct>                              <int>
 1 "Liberal profession"                 133
 2 "Services"                           125
 3 "Administration"                      94
 4 "Selling"                             48
 5 "Business, Management and Finance"    46
 6 "Production"                          39
 7 "Building "                           36
 8 "Commodities Transport"               36
 9 "Repair and maintenance"              32
10 "Agriculture, Fishing, Forestry"      10
```

```
rare_categories <- df |>
  count(CATEGORIE) |>
  filter(n<=40)

rare_categories
```

```
# A tibble: 5 x 2
  CATEGORIE                            n
  <fct>                            <int>
1 "Agriculture, Fishing, Forestry"    10
2 "Building "                         36
3 "Repair and maintenance"            32
4 "Production"                        39
5 "Commodities Transport"             36
```

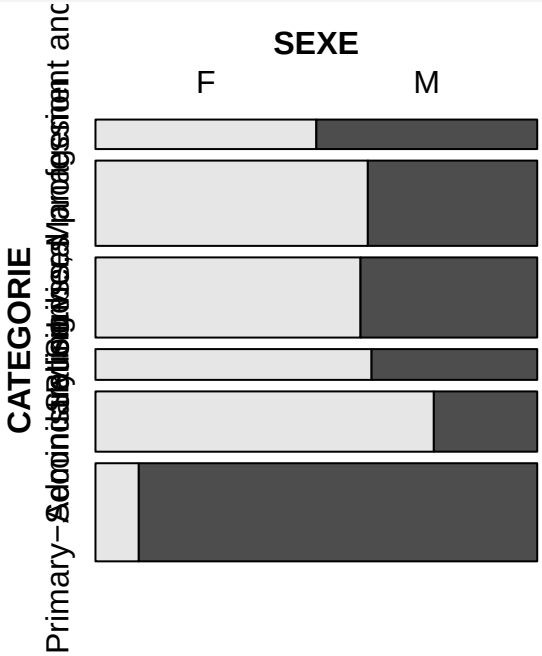
```
df <- df |>
  mutate(CATEGORIE=fct_lump_min(CATEGORIE,
                                min=40,
                                other_level = "Primary-Secondary"))
```

Question

Same as above with `vcd::mosaic`

Solution

```
vcd::mosaic(formula=SEXE~CATEGORIE,
            data=table(select(df, CATEGORIE, SEXE)))
```



Testing association

Chi-square independence/association test

https://statsfonda.github.io/site/content/ch4_2.html#test-dindépendance

Question

- Compute the chi-square association statistic between `CATEGORIE` and `SEXE`.
- Display the output of `chisq.test()` as a table, using `broom::tidy()`

```r
test_1 <- df |>
  select(CATEGORIE, SEXE) |>
  table() |>
  chisq.test()

# test_1

test_1 |>
  broom::tidy() |>
  knitr::kable()
```

statistic	p.value	parameter	method
140.6717	0	5	Pearson's Chi-squared test

Question

Compute the Chi-square statistics from the contingeny table

```r
rowcounts <- apply(tb, MARGIN = 1, FUN = sum)
colcounts <- apply(tb, MARGIN = 2, FUN = sum)

expected <- (rowcounts %*% t(colcounts))/sum(colcounts)

# norm((tb - expected) / sqrt(expected), type = "F")^2

expected |>
  as_tibble() |>
  knitr::kable()
```

F	M
22.80801	23.19199
65.94491	67.05509
61.97830	63.02170
23.79967	24.20033
46.60768	47.39232
75.86144	77.13856

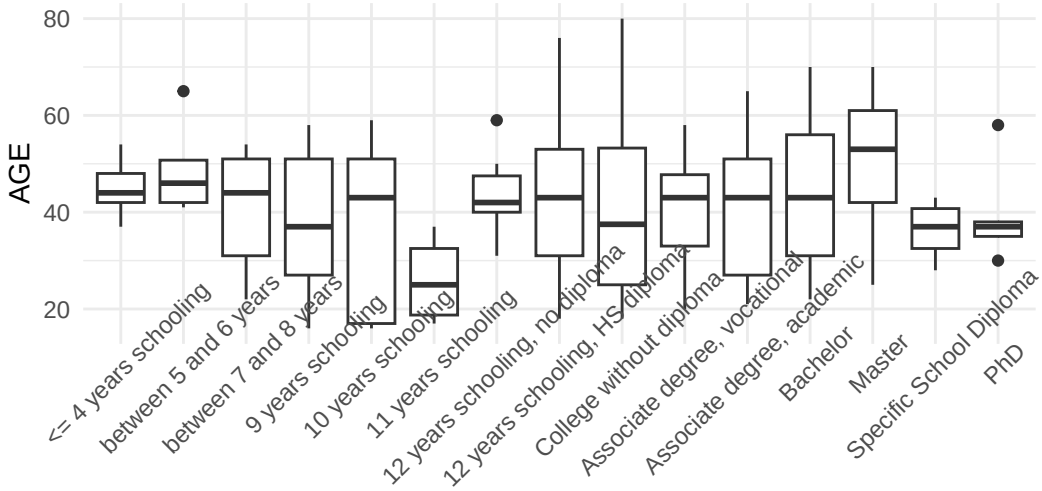
Categorical/Numerical pairs

Grouped boxplots

Question

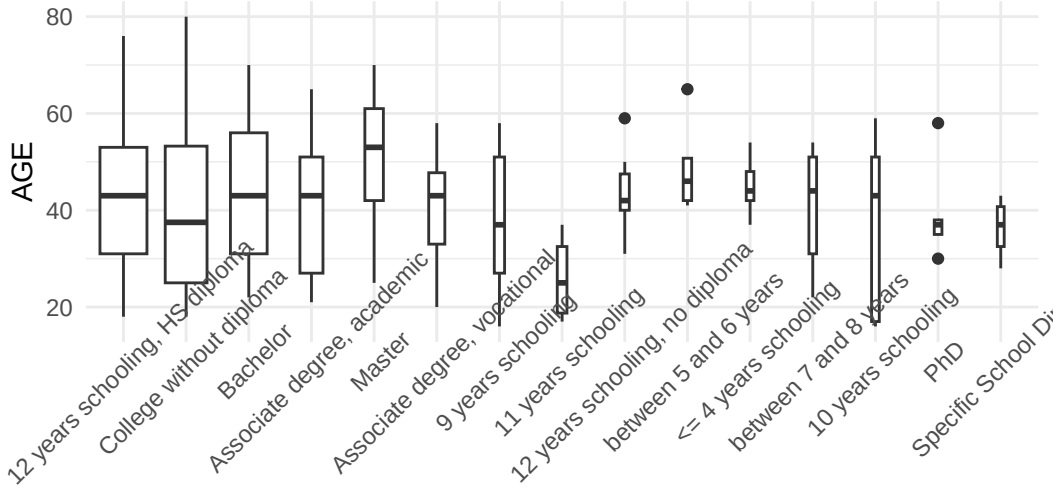
Plot boxplots of `AGE` according to `NIV_ETUDES`

```
df |>
  ggplot() +
  aes(x=NIV_ETUDES, y=AGE) +
  geom_boxplot() +
  rot_x_text
```



NIV_ETUDES

```
df |>
  ggplot() +
  aes(x=fct_infreq(NIV_ETUDES), y=AGE) +
  geom_boxplot(varwidth = T) +
  rot_x_text
```



fct_infreq(NIV_ETUDES)

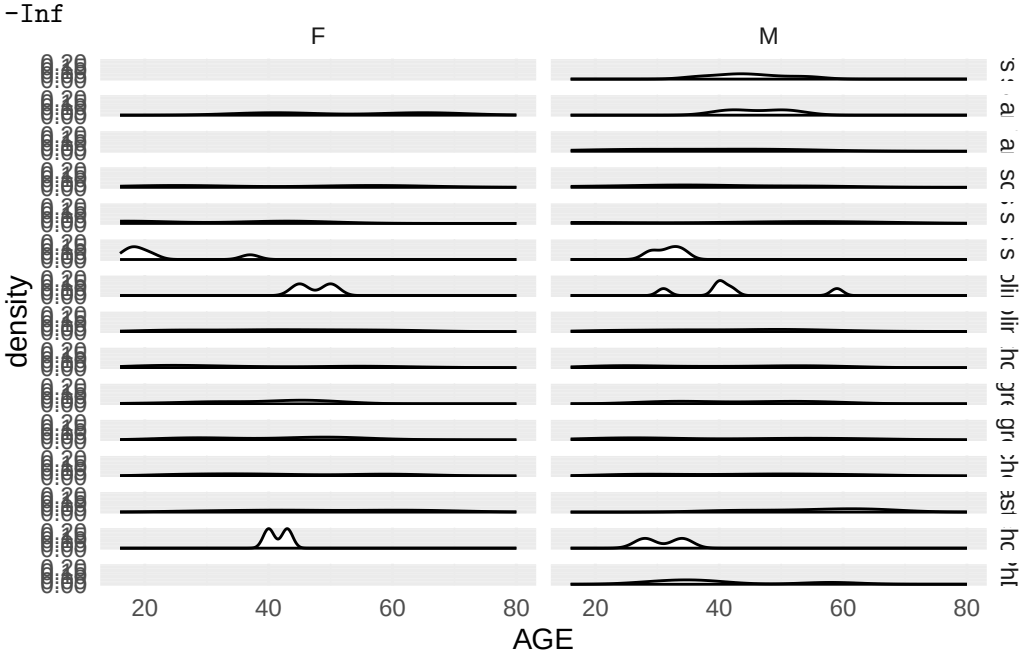
Question

Draw density plots of `AGE`, facet by `NIV_ETUDES` and `SEXE`

```
p <- df |>
  ggplot() +
  aes(x=AGE) +
  stat_density(fill="white", color="black") +
  facet_grid(rows=vars(NIV_ETUDES),
             cols=vars(SEXE))

p
```

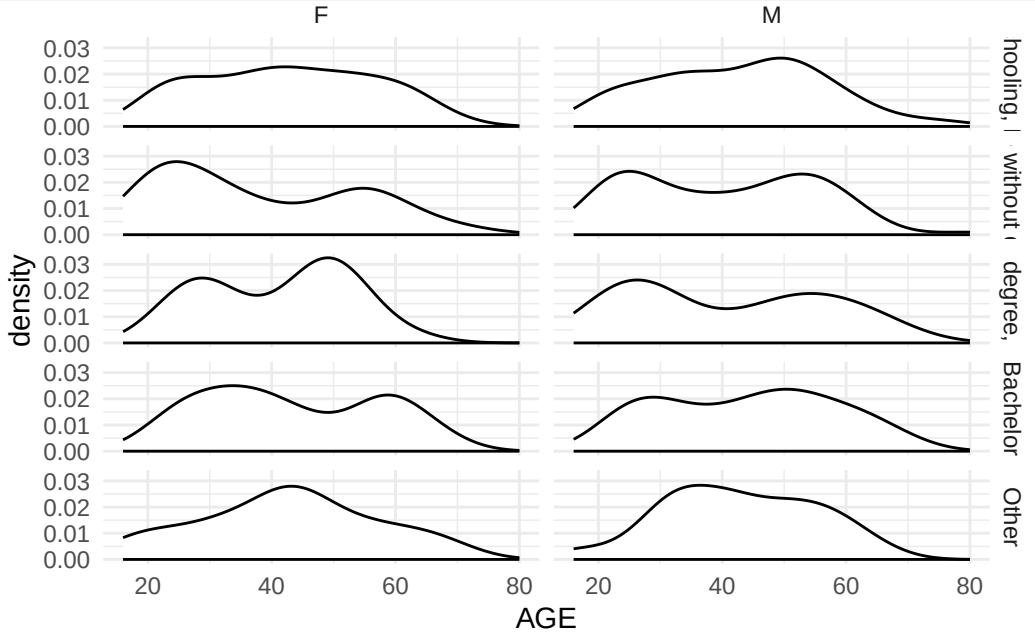
```
Warning: Groups with fewer than two data points have been dropped.
Groups with fewer than two data points have been dropped.
Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
-Inf
Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
-Inf
```



Question

Collapse rare levels of `NIV_ETUDES` and replay.

```
p %+% (
  df |>
    mutate(NIV_ETUDES = fct_lump_min(NIV_ETUDES, min=30))
)
```



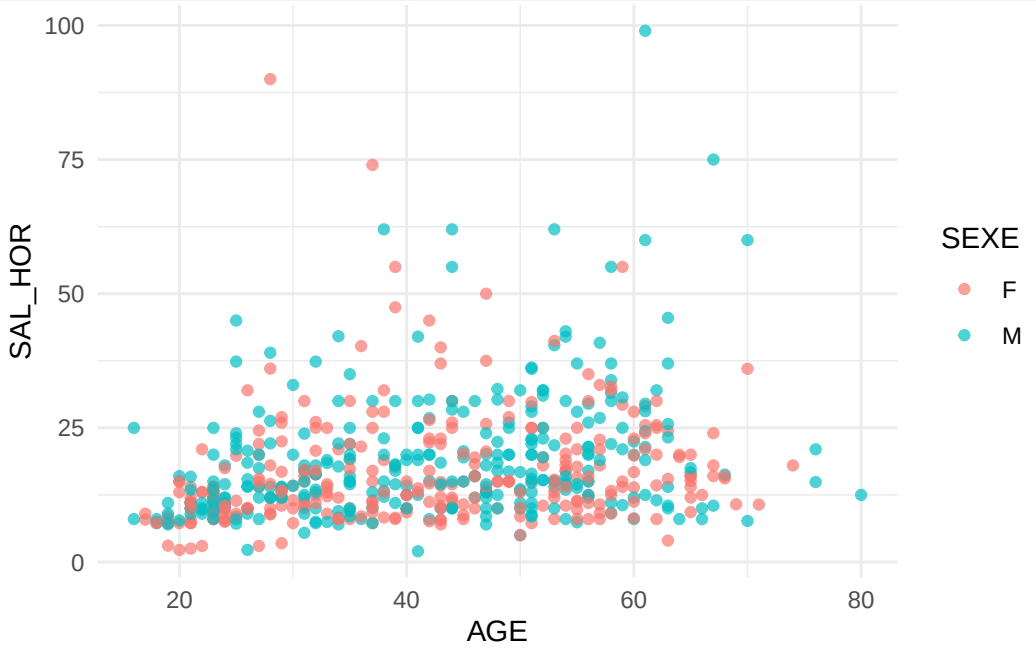
Numerical/Numerical pairs

Scatterplots

Question

Make a scatterplot of `SAL_HOR`with respect to `AGE`

```r
df |>
  ggplot() +
  aes(x=AGE, y=SAL_HOR, color=SEXE) +
  geom_point(alpha=.7)
```



Correlations

- Linear correlation coefficient (Pearson)
- Linear rank correlation coefficient (Spearman ρ, Kendall τ)
- ξ rank correlation coefficient (Chatterjee)

Linear correlation coefficient

Question

Compute the Pearson, Spearman and Kendall correlation coefficients between `AGE` and `SAL_HOR` using function `cor()` from base `R`

Solution

```
df |>
  summarise(
    pearson=cor(AGE, SAL_HOR),
    spearman=cor(AGE, SAL_HOR, method = "spearman"),
    kendall=cor(AGE, SAL_HOR, method="kendall")) |>
  gt::gt() |>
  gt::fmt_number(decimals=2) |>
  gt::tab_caption(
    "Correlation coefficients between SAL_HOR and AGE\nRecensement dataset"
    )
```

pearson	spearman	kendall
0.25	0.32	0.22

Rank based methods

Spearman's rho ( ) and Kendall's tau ( ) are both non-parametric correlation coefficients used to measure the strength and direction of a *monotonic* relationship between two variables.

Spearman's rho ( ) Based on *rank differences*. Defined as the *Pearson correlation coefficient* between the *ranked variables*.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where d_i is the difference between the ranks of each pair, and n is the number of observations.

Kendall's tau ( ) Based on *concordant and discordant pairs*. Measures the *proportion of pairs that have the same order* in both variables compared to the total number of pairs.

$$\tau = \frac{(C - D)}{\frac{1}{2}n(n-1)}$$

where C is the number of *concordant pairs*, and D is the number of *discordant pairs*.

When to Use Which?

Factor	Spearman's rho ( )	Kendall's tau ( )
Large differences in ranks	More sensitive	Less sensitive
Small sample sizes	Less reliable	More reliable
Outlier resistance	Moderate	High
Computational efficiency	Faster	Slower (due to pairwise comparisons)
Interpretation	Similar to Pearson's correlation	More intuitive (proportion of concordance)

Chatterjee's correlation coefficient (Chatterjee's ξ)

The three most popular classical measures of statistical association are Pearson's correlation coefficient, Spearman's , and Kendall's . These coefficients are very powerful for detecting linear or monotone associations, and they have well-developed asymptotic theories for calculating P-values. However, the big problem is that they are not effective for detecting associations that are not monotonic, even in the complete absence of noise.

Let (X, Y) be a pair of random variables, where Y is not a constant. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. pairs with the same law as (X, Y), where $n \geq 2$. The new coefficient has a simpler formula if the X_i's and the Y_i 's have no ties. This simpler formula is presented first, and then the general case is given. Suppose that the X_i's and the Y_i 's have no ties. Rearrange the data as $(X_{(1)}, Y_{(1)}), \ldots, (X_{(n)}, Y_{(n)})$ such that $X_{(1)} \leq \cdots \leq X_{(n)}$. Since the X_i's have no ties, there is a unique way of doing this. Let ri be the rank of $Y_{(i)}$, that is, the number of j such that $Y_{(j)} \leq Y_{(i)}$. The new correlation coefficient is defined as

$$\xi_n(X, Y) := 1 - 3 \sum_{i=1}^{n-1} \frac{|r_{i+1} - r_i|}{n^2 - 1}$$

In the presence of ties, ξ_n is defined as follows. If there are ties among the X_i's, then choose an increasing rearrangement as above by breaking ties uniformly at random. Let r_i be as before, and additionally define l_i to be the number of j such that $Y_{(j)} \geq Y_{(i)}$. Then define

$$\xi_n(X, Y) := 1 - 3n \sum_{i=1}^{n-1} \frac{|r_{i+1} - r_i|}{2 \sum_{i=1}^{n} l_i(n - l_i)}$$

When there are no ties among the Y_i 's, $l_1, \ldots, l_n$ is just a permutation of $1, \ldots, n$, and so the denominator in the above expression is just $n(n^2 - 1)/3$, which reduces this definition to the earlier expression.

From Sourav Chatterjee: A new correlation coefficient

> **Question**
>
> Write a `dplyr` pipeline from computing the ξ correlation coefficient between `Y=lifeExp` and `X=gdpPercap` in the `gapminder` dataset, per `year` and `continent`.

```
gapminder::gapminder |>
  group_by(year, continent) |>
  arrange(gdpPercap) |>
  mutate(rnk= row_number(lifeExp),
         lnk=rank(desc(lifeExp), ties.method = "max"),
         N=n()) |>
  mutate(fol=lead(rnk), dd=abs(fol-rnk)) |>
  summarise(Xi=1-n()*sum(dd, na.rm = T)/(2*sum(lnk*(N-lnk), na.rm = T))) |>
  print(n=50)
```

```
`summarise()` has grouped output by 'year'. You can override using the
`.groups` argument.
```

```
# A tibble: 60 x 3
# Groups:   year [12]
```

	year	continent	Xi
	<int>	<fct>	<dbl>
1	1952	Africa	0.0377
2	1952	Americas	0.135
3	1952	Asia	0.244
4	1952	Europe	0.610
5	1952	Oceania	0
6	1957	Africa	0.0466
7	1957	Americas	0.240
8	1957	Asia	0.330
9	1957	Europe	0.516
10	1957	Oceania	0
11	1962	Africa	-0.00777
12	1962	Americas	0.346
13	1962	Asia	0.377
14	1962	Europe	0.396
15	1962	Oceania	0
16	1967	Africa	0.0533
17	1967	Americas	0.168
18	1967	Asia	0.267
19	1967	Europe	0.383
20	1967	Oceania	0
21	1972	Africa	0.238
22	1972	Americas	0.188
23	1972	Asia	0.324
24	1972	Europe	0.316
25	1972	Oceania	0
26	1977	Africa	0.273
27	1977	Americas	0.135
28	1977	Asia	0.264
29	1977	Europe	0.379
30	1977	Oceania	0
31	1982	Africa	0.383
32	1982	Americas	0.303
33	1982	Asia	0.289
34	1982	Europe	0.0590
35	1982	Oceania	0
36	1987	Africa	0.271
37	1987	Americas	0.428
38	1987	Asia	0.297
39	1987	Europe	0.289
40	1987	Oceania	0
41	1992	Africa	0.363
42	1992	Americas	0.524
43	1992	Asia	0.429
44	1992	Europe	0.393
45	1992	Oceania	0
46	1997	Africa	0.306
47	1997	Americas	0.327
48	1997	Asia	0.498
49	1997	Europe	0.506
50	1997	Oceania	0

i 10 more rows

Using package corrr

https://corrr.tidymodels.org

Question

Solution

```
c <- df |>
  select(where(is.numeric)) |>
  corrr::correlate()
```

```
Correlation computed with
* Method: 'pearson'
* Missing treated using: 'pairwise.complete.obs'
```

```
c |>
  corrr::shave()
```

```
# A tibble: 2 x 3
  term       AGE SAL_HOR
  <chr>    <dbl>   <dbl>
1 AGE     NA          NA
2 SAL_HOR  0.250      NA
```

```
c |>
  corrr::fashion()
```

```
     term  AGE SAL_HOR
1     AGE          .25
2 SAL_HOR  .25
```

```
c |>
  corrr::shave() |>
  corrr::rplot()
```

`pairs` from base `R`

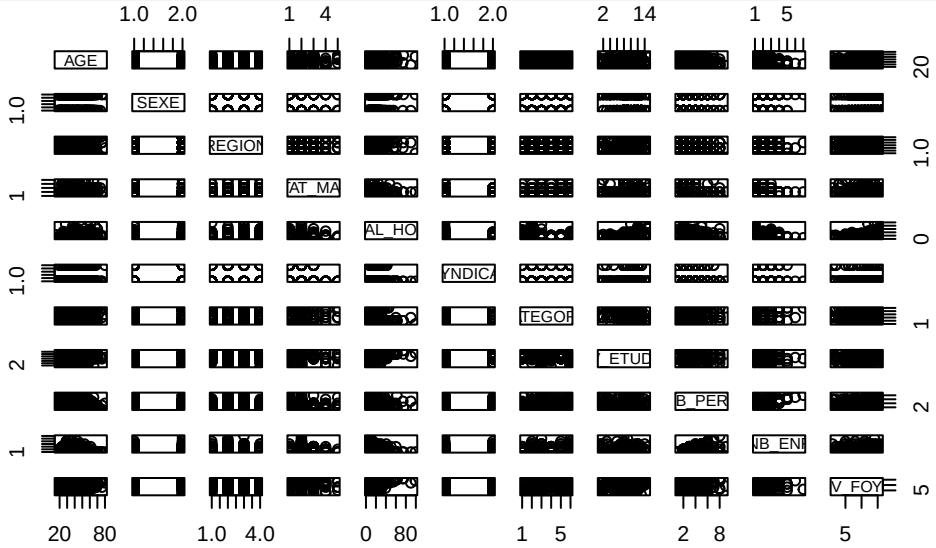
Just as function `skim` from package `skimr` allows us to automate univariate analysis, function `pairs` from base `R` allows us to automate bivariate analysis.

Question

- Apply function `pairs` to the `Recensement` dataset
- How does `pairs` handle pairs of categorical columns?
- How does `pairs` handle pairs of numerical columns?
- How does `pairs` handle categorical/numerical columns?
- Suggestions for improvements?

Solution

```
pairs(df)
```



`ggpairs()` from `GGally`

Documentation

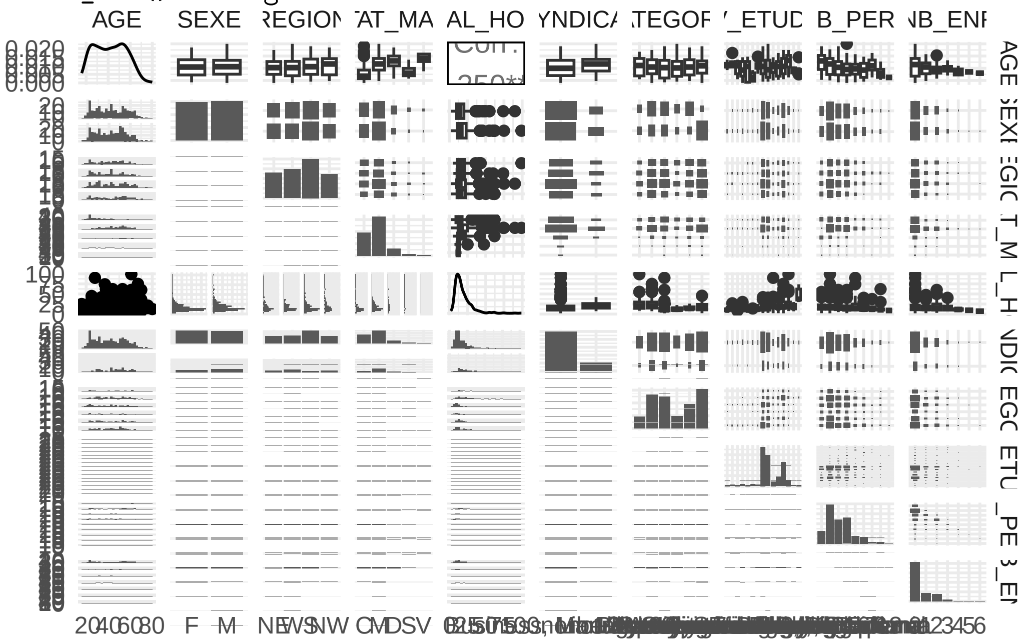
Question

- Apply function `GGally::ggpairs()` to the `Recensement` dataset
- How does `ggpairs` handle pairs of categorical columns?
- How does `ggpairs` handle pairs of numerical columns?
- How does `ggpairs` handle categorical/numerical columns?
- How does `ggpairs` handle diagonal diagrams?
- How can you modify layout, labels, ticks?
- Suggestions for improvements?

Solution

```
df |>
  select(-REV_FOYER) |>
  GGally::ggpairs()
```

```
Registered S3 method overwritten by 'GGally':
  method from
  +.gg   ggplot2
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Useful links

- rmarkdown
- dplyr
- ggplot2
- *R Graphic Cookbook*. Winston Chang. O' Reilly.
- A blog on ggplot object
- `skimr`
- `vcd`

- `ggmosaic`
- `ggforce`
- `arrow`
- `httr`