DS09_Capstone_CYO_project02_mc

Marc Camprodon

2023-11-22

Index-Introduction:

Index:

Introduction:

1. The SUBJECT of study. In this project we´ll study evolution of salaries in the USA through the last 40 years. In particular we'll be interested in analyzing and describing relationships between education and demographic variables and wages.
2. MEASURING EFFECTS. How do these variables impact wages, how do the potential effects of these variables evolve through time? Are there significant differences based on such variables?
3. DATA available. We'll use the data.set "wages-by-education-in-the-usa-1973-2022" from Kaggle. Using Data Wrangling and visualization, we will explore the data assess our first perceptions and define our working hypothesis.
4. In order to understand the test our working hypothesis, we'll utilize least square estimates,linear regression techniques.
5. Further to this, we'll implement a model which will take into account the accumulative effects of variables on outcomes,
6. Finally we'll draw some conclusions, impact of the findings (effect of variables) on the subject of study (pay- outcome).

Parts 1 and 2 - Subject of study and Measuring Effects.

Defining the focus of the study, as described in point 1 and 2 in the introduction above: We'll be interested in measuring effects of demographics (variables) on wages.

Part 3 - DATA analysis.

Loading packages:

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   1.0.0
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dslabs)
```

```
## Warning: package 'dslabs' was built under R version 4.2.3
```

```
library(dplyr)
library(ggplot2)
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(lubridate)
```

```
## Loading required package: timechange
##
## Attaching package: 'lubridate'
##
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(readr)
```

Reading the input dataset

-Read from web source: kaggle datasets download dasaniczka/ wages-by-education-in-the-usa-1973-2022

wages_edu <-read.csv ("https://www.kaggle.com/datasets/asaniczka/wages-by-education-in-the-usa-1973-2022/wages_by_education.csv")

Note "kaggle" might ask for signing-in. Once signed in, click on the download button to recover the data set. The data set can then be stored in the computer, and be read into R from a local folder:

or, -Read from local folder:

```
options(timeout = 120)
wages_edu <- read.csv("C:\\Users\\marc.camprodon\\Documents\\DS_09_Capstone\\Datasets for project 2\\PR
```

Reviewing DATA

List of unique columns

```
col_names_edu <- colnames(wages_edu)
col_names_edu
```

```
##  [1] "year"                          "less_than_hs"
##  [3] "high_school"                   "some_college"
##  [5] "bachelors_degree"              "advanced_degree"
##  [7] "men_less_than_hs"              "men_high_school"
##  [9] "men_some_college"              "men_bachelors_degree"
## [11] "men_advanced_degree"           "women_less_than_hs"
## [13] "women_high_school"             "women_some_college"
## [15] "women_bachelors_degree"        "women_advanced_degree"
## [17] "white_less_than_hs"            "white_high_school"
## [19] "white_some_college"            "white_bachelors_degree"
## [21] "white_advanced_degree"         "black_less_than_hs"
## [23] "black_high_school"             "black_some_college"
## [25] "black_bachelors_degree"        "black_advanced_degree"
## [27] "hispanic_less_than_hs"         "hispanic_high_school"
## [29] "hispanic_some_college"         "hispanic_bachelors_degree"
## [31] "hispanic_advanced_degree"      "white_men_less_than_hs"
## [33] "white_men_high_school"         "white_men_some_college"
## [35] "white_men_bachelors_degree"    "white_men_advanced_degree"
## [37] "black_men_less_than_hs"        "black_men_high_school"
## [39] "black_men_some_college"        "black_men_bachelors_degree"
## [41] "black_men_advanced_degree"     "hispanic_men_less_than_hs"
## [43] "hispanic_men_high_school"      "hispanic_men_some_college"
## [45] "hispanic_men_bachelors_degree" "hispanic_men_advanced_degree"
## [47] "white_women_less_than_hs"      "white_women_high_school"
## [49] "white_women_some_college"      "white_women_bachelors_degree"
## [51] "white_women_advanced_degree"   "black_women_less_than_hs"
## [53] "black_women_high_school"       "black_women_some_college"
## [55] "black_women_bachelors_degree"  "black_women_advanced_degree"
## [57] "hispanic_women_less_than_hs"   "hispanic_women_high_school"
## [59] "hispanic_women_some_college"   "hispanic_women_bachelors_degree"
## [61] "hispanic_women_advanced_degree"
```

Viewing the first part of the data.frame:

```
head(wages_edu)
```

```
##   year less_than_hs high_school some_college bachelors_degree advanced_degree
## 1 2022        16.52       21.94        24.81            41.60           53.22
## 2 2021        16.74       22.28        24.92            41.32           53.45
## 3 2020        17.02       22.70        25.44            41.65           53.74
## 4 2019        16.11       21.64        24.00            39.61           51.57
```

```
## 5 2018        15.94       21.50        23.70            38.87           51.03
## 6 2017        15.92       21.26        23.31            38.65           49.40
##   men_less_than_hs men_high_school men_some_college men_bachelors_degree
## 1            17.99           24.08            27.96                49.01
## 2            18.34           24.36            27.96                47.83
## 3            18.76           25.09            28.55                48.15
## 4            17.55           23.99            26.99                45.74
## 5            17.70           23.72            26.61                44.97
## 6            17.63           23.47            25.91                44.50
##   men_advanced_degree women_less_than_hs women_high_school women_some_college
## 1               63.51              14.33             18.93               21.76
## 2               63.52              14.36             19.36               21.97
## 3               62.70              14.40             19.35               22.35
## 4               59.93              13.96             18.48               21.09
## 5               59.73              13.36             18.49               20.91
## 6               56.77              13.39             18.31               20.83
##   women_bachelors_degree women_advanced_degree white_less_than_hs
## 1                  34.39                44.34              15.70
## 2                  35.08                44.80              16.20
## 3                  35.41                46.04              17.01
## 4                  33.80                44.22              15.89
## 5                  33.03                43.19              16.05
## 6                  33.01                42.75              15.96
##   white_high_school white_some_college white_bachelors_degree
## 1             23.31              26.28                  43.30
## 2             23.60              26.29                  43.06
## 3             24.00              26.90                  43.16
## 4             22.94              25.47                  41.06
## 5             23.02              25.15                  40.49
## 6             22.71              24.65                  40.07
##   white_advanced_degree black_less_than_hs black_high_school black_some_college
## 1                 53.30              15.19             19.39               21.34
## 2                 53.62              14.55             19.66               21.26
## 3                 53.80              14.55             19.66               21.79
## 4                 51.81              14.19             18.73               20.44
## 5                 51.80              13.31             18.14               19.98
## 6                 50.03              13.93             17.94               19.80
##   black_bachelors_degree black_advanced_degree hispanic_less_than_hs
## 1                  33.39                 44.67                 17.32
## 2                  32.62                 43.37                 17.60
## 3                  33.64                 45.76                 17.61
## 4                  31.83                 42.73                 16.71
## 5                  32.00                 42.22                 16.44
## 6                  31.49                 40.68                 16.34
##   hispanic_high_school hispanic_some_college hispanic_bachelors_degree
## 1                20.72                 22.96                     36.00
## 2                21.03                 23.34                     35.26
## 3                21.46                 23.24                     36.13
## 4                20.47                 22.01                     34.66
## 5                20.13                 21.74                     33.19
## 6                19.83                 21.59                     33.07
##   hispanic_advanced_degree white_men_less_than_hs white_men_high_school
## 1                    48.30                 17.14                25.92
## 2                    48.60                 18.13                26.03
```

```
## 3                   48.58                 18.97                 26.73
## 4                   46.69                 17.66                 25.58
## 5                   44.83                 17.96                 25.49
## 6                   43.92                 17.82                 25.14
##   white_men_some_college white_men_bachelors_degree white_men_advanced_degree
## 1                  29.93                      51.23                      63.86
## 2                  29.79                      50.06                      64.04
## 3                  30.40                      50.15                      63.50
## 4                  28.90                      47.76                      60.84
## 5                  28.41                      47.28                      61.27
## 6                  27.61                      46.44                      57.64
##   black_men_less_than_hs black_men_high_school black_men_some_college
## 1                  16.38                 20.73                 22.58
## 2                  15.38                 20.86                 22.63
## 3                  15.52                 21.22                 23.57
## 4                  15.07                 20.22                 22.02
## 5                  14.27                 19.36                 21.82
## 6                  15.00                 19.46                 21.09
##   black_men_bachelors_degree black_men_advanced_degree
## 1                      37.63                    52.91
## 2                      36.95                    49.01
## 3                      36.70                    51.67
## 4                      33.73                    46.87
## 5                      33.42                    46.42
## 6                      34.41                    44.62
##   hispanic_men_less_than_hs hispanic_men_high_school hispanic_men_some_college
## 1                     18.67                    22.32                    25.49
## 2                     18.98                    22.70                    25.58
## 3                     19.26                    23.33                    25.69
## 4                     17.93                    22.33                    24.35
## 5                     18.05                    22.04                    24.20
## 6                     17.91                    21.68                    23.59
##   hispanic_men_bachelors_degree hispanic_men_advanced_degree
## 1                         41.48                       57.08
## 2                         39.61                       55.60
## 3                         41.00                       53.58
## 4                         39.56                       52.24
## 5                         37.27                       51.27
## 6                         36.62                       50.58
##   white_women_less_than_hs white_women_high_school white_women_some_college
## 1                    13.84                   19.56                   22.52
## 2                    13.77                   20.08                   22.73
## 3                    14.20                   20.06                   23.31
## 4                    13.44                   19.22                   21.99
## 5                    13.45                   19.52                   21.90
## 6                    13.46                   19.34                   21.72
##   white_women_bachelors_degree white_women_advanced_degree
## 1                        35.31                       44.45
## 2                        36.11                       44.82
## 3                        36.20                       45.58
## 4                        34.47                       43.83
## 5                        33.78                       43.31
## 6                        33.76                       43.22
##   black_women_less_than_hs black_women_high_school black_women_some_college
```

```
## 1                    13.89                    17.83                    20.36
## 2                    13.73                    18.30                    20.18
## 3                    13.66                    17.93                    20.31
## 4                    13.30                    17.18                    19.17
## 5                    12.48                    16.81                    18.57
## 6                    12.99                    16.33                    18.76
##   black_women_bachelors_degree black_women_advanced_degree
## 1                        29.94                       39.41
## 2                        29.35                       40.07
## 3                        31.38                       42.44
## 4                        30.31                       40.42
## 5                        30.85                       39.64
## 6                        29.19                       38.26
##   hispanic_women_less_than_hs hispanic_women_high_school
## 1                       14.74                      18.18
## 2                       14.97                      18.34
## 3                       14.58                      18.50
## 4                       14.50                      17.71
## 5                       13.47                      17.28
## 6                       13.36                      17.02
##   hispanic_women_some_college hispanic_women_bachelors_degree
## 1                       20.64                           31.13
## 2                       21.14                           31.25
## 3                       20.69                           31.55
## 4                       19.69                           30.18
## 5                       19.29                           29.47
## 6                       19.60                           29.69
##   hispanic_women_advanced_degree
## 1                          40.64
## 2                          42.47
## 3                          44.15
## 4                          42.30
## 5                          39.35
## 6                          38.43
```

Displaying the internal strcuture of the data set:

```
str(wages_edu)
```

```
## 'data.frame':    50 obs. of  61 variables:
##  $ year                          : int  2022 2021 2020 2019 2018 2017 2016 2015 2014 2013 ...
##  $ less_than_hs                  : num  16.5 16.7 17 16.1 15.9 ...
##  $ high_school                   : num  21.9 22.3 22.7 21.6 21.5 ...
##  $ some_college                  : num  24.8 24.9 25.4 24 23.7 ...
##  $ bachelors_degree              : num  41.6 41.3 41.6 39.6 38.9 ...
##  $ advanced_degree               : num  53.2 53.5 53.7 51.6 51 ...
##  $ men_less_than_hs              : num  18 18.3 18.8 17.6 17.7 ...
##  $ men_high_school               : num  24.1 24.4 25.1 24 23.7 ...
##  $ men_some_college              : num  28 28 28.6 27 26.6 ...
##  $ men_bachelors_degree          : num  49 47.8 48.1 45.7 45 ...
##  $ men_advanced_degree           : num  63.5 63.5 62.7 59.9 59.7 ...
##  $ women_less_than_hs            : num  14.3 14.4 14.4 14 13.4 ...
##  $ women_high_school             : num  18.9 19.4 19.4 18.5 18.5 ...
```

```
##  $ women_some_college             : num  21.8 22 22.4 21.1 20.9 ...
##  $ women_bachelors_degree         : num  34.4 35.1 35.4 33.8 33 ...
##  $ women_advanced_degree          : num  44.3 44.8 46 44.2 43.2 ...
##  $ white_less_than_hs             : num  15.7 16.2 17 15.9 16.1 ...
##  $ white_high_school              : num  23.3 23.6 24 22.9 23 ...
##  $ white_some_college             : num  26.3 26.3 26.9 25.5 25.1 ...
##  $ white_bachelors_degree         : num  43.3 43.1 43.2 41.1 40.5 ...
##  $ white_advanced_degree          : num  53.3 53.6 53.8 51.8 51.8 ...
##  $ black_less_than_hs             : num  15.2 14.6 14.6 14.2 13.3 ...
##  $ black_high_school              : num  19.4 19.7 19.7 18.7 18.1 ...
##  $ black_some_college             : num  21.3 21.3 21.8 20.4 20 ...
##  $ black_bachelors_degree         : num  33.4 32.6 33.6 31.8 32 ...
##  $ black_advanced_degree          : num  44.7 43.4 45.8 42.7 42.2 ...
##  $ hispanic_less_than_hs          : num  17.3 17.6 17.6 16.7 16.4 ...
##  $ hispanic_high_school           : num  20.7 21 21.5 20.5 20.1 ...
##  $ hispanic_some_college          : num  23 23.3 23.2 22 21.7 ...
##  $ hispanic_bachelors_degree      : num  36 35.3 36.1 34.7 33.2 ...
##  $ hispanic_advanced_degree       : num  48.3 48.6 48.6 46.7 44.8 ...
##  $ white_men_less_than_hs         : num  17.1 18.1 19 17.7 18 ...
##  $ white_men_high_school          : num  25.9 26 26.7 25.6 25.5 ...
##  $ white_men_some_college         : num  29.9 29.8 30.4 28.9 28.4 ...
##  $ white_men_bachelors_degree     : num  51.2 50.1 50.1 47.8 47.3 ...
##  $ white_men_advanced_degree      : num  63.9 64 63.5 60.8 61.3 ...
##  $ black_men_less_than_hs         : num  16.4 15.4 15.5 15.1 14.3 ...
##  $ black_men_high_school          : num  20.7 20.9 21.2 20.2 19.4 ...
##  $ black_men_some_college         : num  22.6 22.6 23.6 22 21.8 ...
##  $ black_men_bachelors_degree     : num  37.6 37 36.7 33.7 33.4 ...
##  $ black_men_advanced_degree      : num  52.9 49 51.7 46.9 46.4 ...
##  $ hispanic_men_less_than_hs      : num  18.7 19 19.3 17.9 18.1 ...
##  $ hispanic_men_high_school       : num  22.3 22.7 23.3 22.3 22 ...
##  $ hispanic_men_some_college      : num  25.5 25.6 25.7 24.4 24.2 ...
##  $ hispanic_men_bachelors_degree  : num  41.5 39.6 41 39.6 37.3 ...
##  $ hispanic_men_advanced_degree   : num  57.1 55.6 53.6 52.2 51.3 ...
##  $ white_women_less_than_hs       : num  13.8 13.8 14.2 13.4 13.4 ...
##  $ white_women_high_school        : num  19.6 20.1 20.1 19.2 19.5 ...
##  $ white_women_some_college       : num  22.5 22.7 23.3 22 21.9 ...
##  $ white_women_bachelors_degree   : num  35.3 36.1 36.2 34.5 33.8 ...
##  $ white_women_advanced_degree    : num  44.5 44.8 45.6 43.8 43.3 ...
##  $ black_women_less_than_hs       : num  13.9 13.7 13.7 13.3 12.5 ...
##  $ black_women_high_school        : num  17.8 18.3 17.9 17.2 16.8 ...
##  $ black_women_some_college       : num  20.4 20.2 20.3 19.2 18.6 ...
##  $ black_women_bachelors_degree   : num  29.9 29.4 31.4 30.3 30.9 ...
##  $ black_women_advanced_degree    : num  39.4 40.1 42.4 40.4 39.6 ...
##  $ hispanic_women_less_than_hs    : num  14.7 15 14.6 14.5 13.5 ...
##  $ hispanic_women_high_school     : num  18.2 18.3 18.5 17.7 17.3 ...
##  $ hispanic_women_some_college    : num  20.6 21.1 20.7 19.7 19.3 ...
##  $ hispanic_women_bachelors_degree: num  31.1 31.2 31.6 30.2 29.5 ...
##  $ hispanic_women_advanced_degree : num  40.6 42.5 44.1 42.3 39.4 ...
```

'data.frame': 50 obs. of 61 variables

We'd like to have a more manageable data frame. We want to transform it to Tidy Data. Tidy format in which each row represents one observation and columns represent (combinations of) the different variables available for each of these observations. We could consolidate the information reshaping the data: we could

organize data in 5 columns, 4 selection variables, Education level (edu_level; edu_5), Race (race), Gender (gender), Time (year), and the result of the observation or fifth variable Wages (pay)(value USD per h).

We will iniciate the process of DATA WRANGLING, converting the data set to tidy form. We can observe that data can be organized in 12 groups: -All, All-Men, All-Women, All-Black,All-White,All-Hispanic; -Men-Black, Women-Black, Men-White, Women-White, Men-Hispanic, Women-Hispanic. All of these groups have 5 possible values for the education level.

We'll generate following these 12 groups as data.frames from "wages_edu", all with a Tidy format:

-6B.1_wages_edu_all: year=all ; gen=all ; race=all ; level_edu=all five ; pay=values

-6B.2_wages_edu_Men: year=all ; gen=Men ; race=all ; level_edu=all five ; pay=values

-6B.3_wages_edu_Women: year=all ; gen=Women ; race=all ; level_edu=all five ; pay=values

-6B.4_wages_edu_B: year=all ; gen=all ; race=Black ; level_edu=all five ; pay=values

-6B.5_wages_edu_W: year=all ; gen=all ; race=White ; level_edu=all five ; pay=values

-6B.6_wages_edu_H: year=all ; gen=all ; race=Hispanic ; level_edu=all five ; pay=values

-6A.1_wages_edu_M_B: year=all ; gen=Men ; race=Black ; level_edu=all five ; pay=values

-6A.2_wages_edu_M_W: year=all ; gen=Men ; race=White ; level_edu=all five ; pay=values

-6A.3_wages_edu_M_H: year=all ; gen=Men ; race=Hispanic ; level_edu=all five ; pay=values

-6A.4_wages_edu_W_B: year=all ; gen=Women ; race=Black ; level_edu=all five ; pay=values

-6A.5_wages_edu_W_W: year=all ; gen=Women ; race=White ; level_edu=all five ; pay=values

-6A.6_wages_edu_W_H: year=all ; gen=Women ; race=Hispanic ; level_edu=all five ; pay=values

Let's prepare the first of these 12 data.frames, 6B.1:

6B.1_wages_edu_all: year=all ; gen=all ; race=all ; level_edu=all five; pay=values

```
wages_edu_all <- wages_edu %>%      select(year,less_than_hs,high_school,some_college,bachelors_degree,a
head(wages_edu_all)
```

```
##   year less_than_hs high_school some_college bachelors_degree advanced_degree
## 1 2022        16.52       21.94        24.81            41.60           53.22
## 2 2021        16.74       22.28        24.92            41.32           53.45
## 3 2020        17.02       22.70        25.44            41.65           53.74
## 4 2019        16.11       21.64        24.00            39.61           51.57
## 5 2018        15.94       21.50        23.70            38.87           51.03
## 6 2017        15.92       21.26        23.31            38.65           49.40
```

```
tidy_wages_edu_all <- gather(wages_edu_all, edu_level, pay, `less_than_hs`:`advanced_degree`)
head(tidy_wages_edu_all)
```

```
##   year    edu_level   pay
## 1 2022 less_than_hs 16.52
## 2 2021 less_than_hs 16.74
## 3 2020 less_than_hs 17.02
## 4 2019 less_than_hs 16.11
## 5 2018 less_than_hs 15.94
## 6 2017 less_than_hs 15.92
```

```
tidy_wages_edu_all <-tidy_wages_edu_all %>% mutate(gender="All", race="All") %>% mutate(edu_5=c(1:250))
```

We'll prepare these 12 data.frames from the main source, and work on transforming them to tidy data. In the process, in order to improve and complete tidiness of data, we'll focus on the edu_level variable. This has 55 different values, while there are actually only 5 different levels of education being considered in the original data. The "edu_level" variable includes gender and race variables within in the original data frame structure. We'll reduce that complexity to 5 actual education level values (hereby referred as "a","b","c","d","e"),from the maximum education level (advanced degree), now level "a",to the minimum education level,"less than hs", hereby level "e".

Our new education level naming convention is then as following:

-"a" = "advanced_degree"

-"b" = "bachelors_degree"

-"c" = "some_college"

-"d" = "high_school"

-"e" = "less_than_hs"

```
    tidy_wages_edu_all$edu_5[1:50]="e"
    tidy_wages_edu_all$edu_5[51:100]="d"
    tidy_wages_edu_all$edu_5[101:150]="c"
    tidy_wages_edu_all$edu_5[151:200]="b"
    tidy_wages_edu_all$edu_5[201:250]="a"

    df6B1<-tidy_wages_edu_all

    #Review of the new data frame, structure and visualization:
    head(df6B1)
```

```
##   year    edu_level   pay gender race edu_5
## 1 2022 less_than_hs 16.52    All  All     e
## 2 2021 less_than_hs 16.74    All  All     e
## 3 2020 less_than_hs 17.02    All  All     e
## 4 2019 less_than_hs 16.11    All  All     e
## 5 2018 less_than_hs 15.94    All  All     e
## 6 2017 less_than_hs 15.92    All  All     e
```

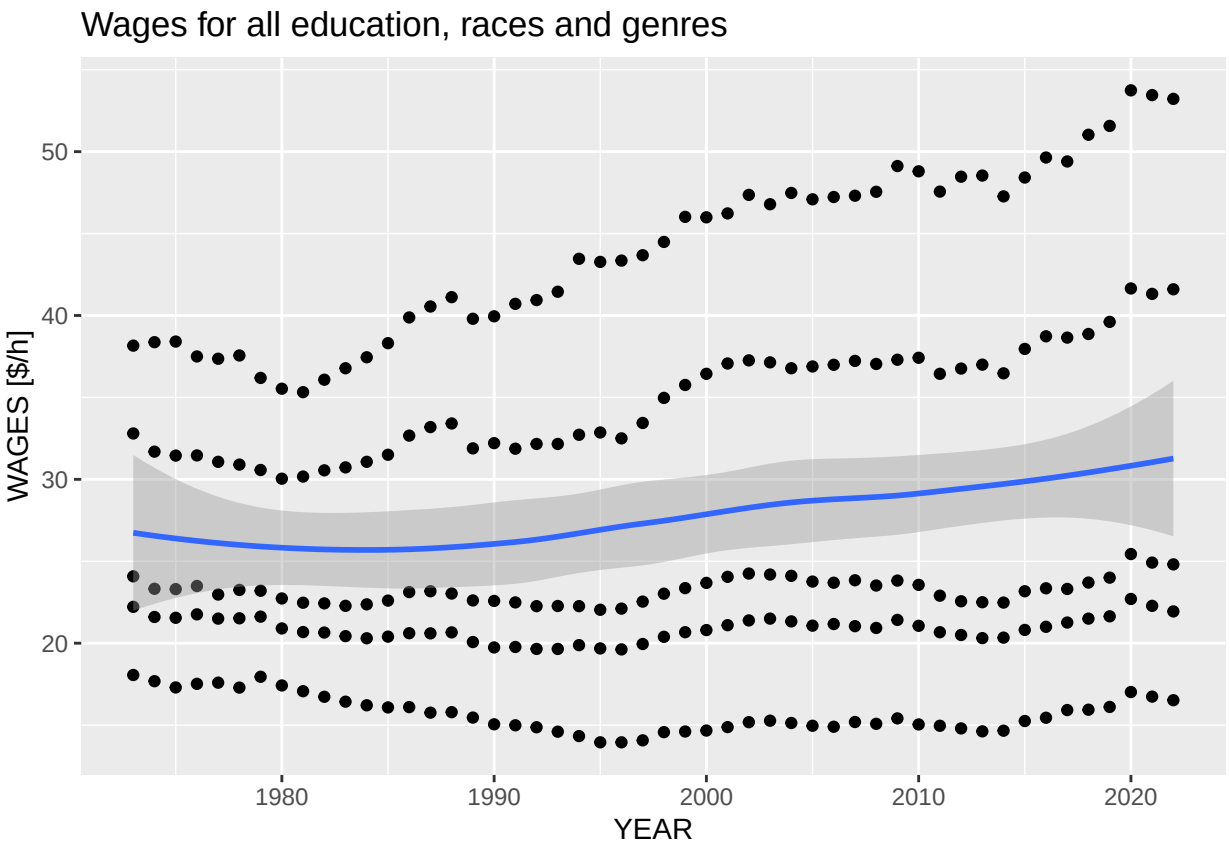
```
    str(df6B1)
```

```
## 'data.frame':    250 obs. of  6 variables:
##  $ year     : int  2022 2021 2020 2019 2018 2017 2016 2015 2014 2013 ...
##  $ edu_level: chr  "less_than_hs" "less_than_hs" "less_than_hs" "less_than_hs" ...
##  $ pay      : num  16.5 16.7 17 16.1 15.9 ...
##  $ gender   : chr  "All" "All" "All" "All" ...
##  $ race     : chr  "All" "All" "All" "All" ...
##  $ edu_5    : chr  "e" "e" "e" "e" ...
```

```
    qplot(year,pay,data=df6B1)+geom_smooth()+
      xlab("YEAR") +
      ylab("WAGES [$/h]") +
      ggtitle("Wages for all education, races and genres")
```

```
## Warning: ‘qplot()‘ was deprecated in ggplot2 3.4.0.
```

```
## ‘geom_smooth()‘ using method = ’loess’ and formula = ’y ~ x’
```



Preparing the second of the 12 data.frames, 6B.2:

6B.2_wages_edu_Men: year=all ; gen=Men ; race=all ; level_edu=all five; pay=values

```
    wages_edu_Men <- wages_edu %>%   select(year,men_less_than_hs,men_high_school,men_some_college,men_
    head(wages_edu_Men)
```

```
##   year men_less_than_hs men_high_school men_some_college men_bachelors_degree
## 1 2022            17.99           24.08            27.96                49.01
## 2 2021            18.34           24.36            27.96                47.83
## 3 2020            18.76           25.09            28.55                48.15
## 4 2019            17.55           23.99            26.99                45.74
## 5 2018            17.70           23.72            26.61                44.97
## 6 2017            17.63           23.47            25.91                44.50
##   men_advanced_degree
## 1               63.51
## 2               63.52
## 3               62.70
## 4               59.93
## 5               59.73
## 6               56.77
```

```
    tidy_wages_edu_Men <- gather(wages_edu_Men, edu_level, pay, `men_less_than_hs`:`men_advanced_degree`)
    head(tidy_wages_edu_Men)
```

```
##   year        edu_level   pay
## 1 2022 men_less_than_hs 17.99
## 2 2021 men_less_than_hs 18.34
## 3 2020 men_less_than_hs 18.76
## 4 2019 men_less_than_hs 17.55
## 5 2018 men_less_than_hs 17.70
## 6 2017 men_less_than_hs 17.63
```

```
    tidy_wages_edu_Men <-tidy_wages_edu_Men %>% mutate(gender="Men", race="All")%>% mutate(edu_5=c(1:250

    tidy_wages_edu_Men$edu_5[1:50]="e"
    tidy_wages_edu_Men$edu_5[51:100]="d"
    tidy_wages_edu_Men$edu_5[101:150]="c"
    tidy_wages_edu_Men$edu_5[151:200]="b"
    tidy_wages_edu_Men$edu_5[201:250]="a"

    df6B2<-tidy_wages_edu_Men

    #Review of the new data frame, structure and visualization:
    head(df6B2)
```

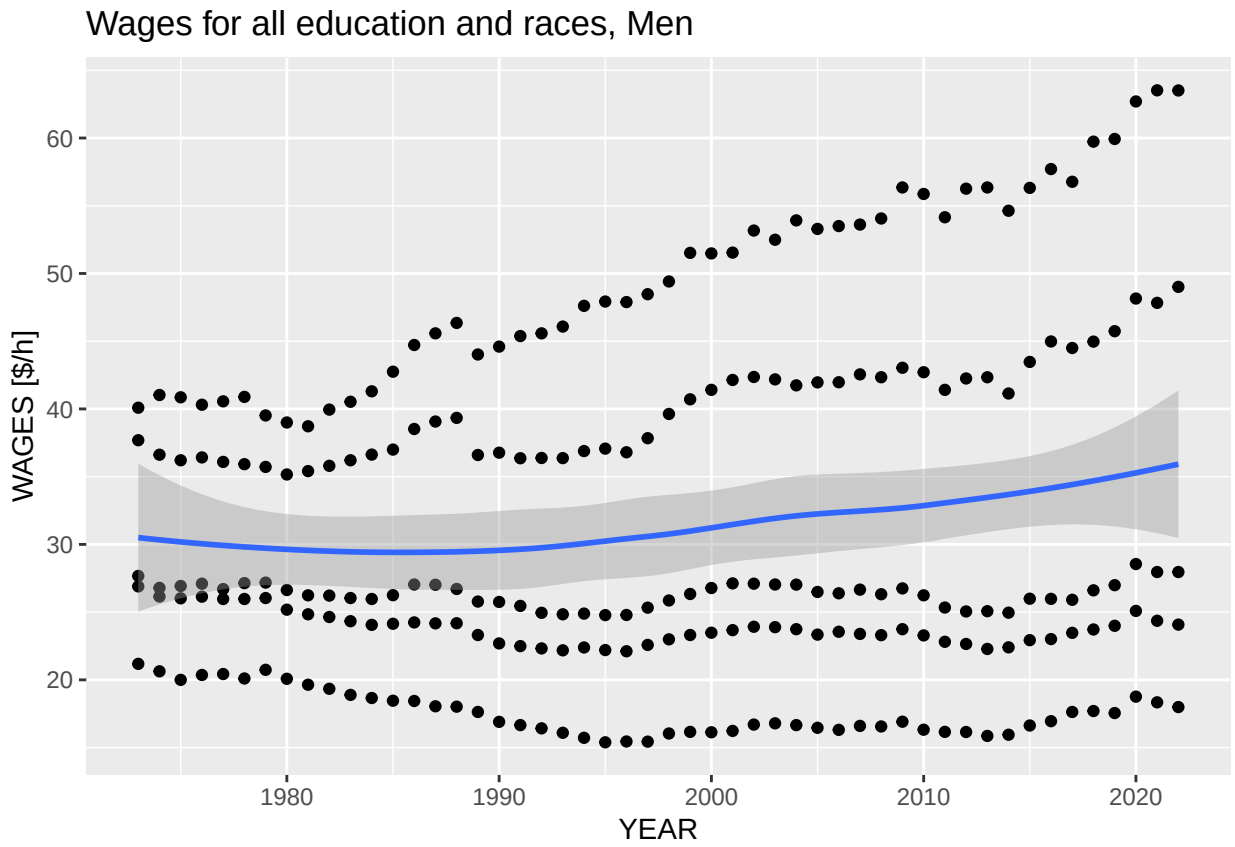
```
##   year        edu_level   pay gender race edu_5
## 1 2022 men_less_than_hs 17.99    Men  All     e
## 2 2021 men_less_than_hs 18.34    Men  All     e
## 3 2020 men_less_than_hs 18.76    Men  All     e
## 4 2019 men_less_than_hs 17.55    Men  All     e
## 5 2018 men_less_than_hs 17.70    Men  All     e
## 6 2017 men_less_than_hs 17.63    Men  All     e
```

```
    str(df6B2)
```

```
## 'data.frame':    250 obs. of  6 variables:
##  $ year     : int  2022 2021 2020 2019 2018 2017 2016 2015 2014 2013 ...
##  $ edu_level: chr  "men_less_than_hs" "men_less_than_hs" "men_less_than_hs" "men_less_than_hs" ...
##  $ pay      : num  18 18.3 18.8 17.6 17.7 ...
##  $ gender   : chr  "Men" "Men" "Men" "Men" ...
##  $ race     : chr  "All" "All" "All" "All" ...
##  $ edu_5    : chr  "e" "e" "e" "e" ...
```

```
    qplot(year,pay,data=df6B2)+geom_smooth()+
      xlab("YEAR") +
      ylab("WAGES [$/h]") +
      ggtitle("Wages for all education and races, Men")
```

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



Preparing the third of the 12 data.frames, 6B.3:

6B.3_wages_edu_Women: year=all ; gen=Women ; race=all ; level_edu=all five; pay=values

```
    wages_edu_Women <- wages_edu %>% select(year,women_less_than_hs,women_high_school,women_some_college
    head(wages_edu_Women)
```

```
##   year women_less_than_hs women_high_school women_some_college
## 1 2022              14.33             18.93              21.76
## 2 2021              14.36             19.36              21.97
## 3 2020              14.40             19.35              22.35
## 4 2019              13.96             18.48              21.09
## 5 2018              13.36             18.49              20.91
## 6 2017              13.39             18.31              20.83
##   women_bachelors_degree women_advanced_degree
## 1                  34.39                44.34
## 2                  35.08                44.80
## 3                  35.41                46.04
## 4                  33.80                44.22
## 5                  33.03                43.19
## 6                  33.01                42.75
```

```
    tidy_wages_edu_Women <- gather(wages_edu_Women, edu_level, pay, `women_less_than_hs`:`women_advance
    head(tidy_wages_edu_Women)
```

```
##   year          edu_level   pay
```

```
## 1 2022 women_less_than_hs 14.33
## 2 2021 women_less_than_hs 14.36
## 3 2020 women_less_than_hs 14.40
## 4 2019 women_less_than_hs 13.96
## 5 2018 women_less_than_hs 13.36
## 6 2017 women_less_than_hs 13.39
```

```
    tidy_wages_edu_Women <-tidy_wages_edu_Women %>% mutate(gender="Women", race="All")%>% mutate(edu_5=

    tidy_wages_edu_Women$edu_5[1:50]="e"
    tidy_wages_edu_Women$edu_5[51:100]="d"
    tidy_wages_edu_Women$edu_5[101:150]="c"
    tidy_wages_edu_Women$edu_5[151:200]="b"
    tidy_wages_edu_Women$edu_5[201:250]="a"

    df6B3<-tidy_wages_edu_Women

    #Review of the new data frame, structure and visualization:
    head(df6B3)
```

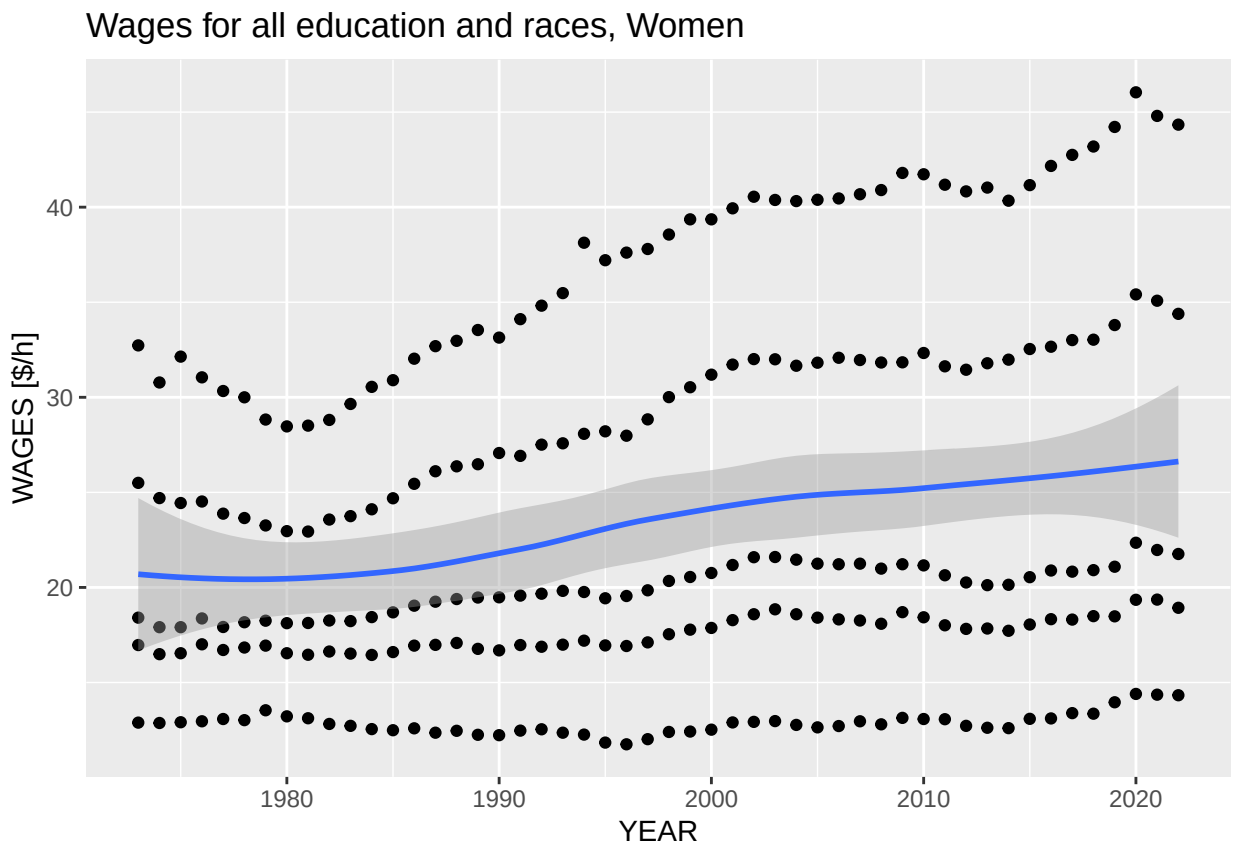
```
##   year          edu_level   pay gender race edu_5
## 1 2022 women_less_than_hs 14.33  Women  All     e
## 2 2021 women_less_than_hs 14.36  Women  All     e
## 3 2020 women_less_than_hs 14.40  Women  All     e
## 4 2019 women_less_than_hs 13.96  Women  All     e
## 5 2018 women_less_than_hs 13.36  Women  All     e
## 6 2017 women_less_than_hs 13.39  Women  All     e
```

```
    str(df6B3)
```

```
## 'data.frame':    250 obs. of  6 variables:
##  $ year     : int  2022 2021 2020 2019 2018 2017 2016 2015 2014 2013 ...
##  $ edu_level: chr  "women_less_than_hs" "women_less_than_hs" "women_less_than_hs" "women_less_than_hs
##  $ pay      : num  14.3 14.4 14.4 14 13.4 ...
##  $ gender   : chr  "Women" "Women" "Women" "Women" ...
##  $ race     : chr  "All" "All" "All" "All" ...
##  $ edu_5    : chr  "e" "e" "e" "e" ...
```

```
    qplot(year,pay,data=df6B3)+geom_smooth()+
      xlab("YEAR") +
      ylab("WAGES [$/h]") +
      ggtitle("Wages for all education and races, Women")
```

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



Preparing the fourth of the 12 data.frames, 6B.4:

6B.4_wages_edu_B: year=all ; gen=all ; race=Black ; level_edu=all five; pay=values

```
    wages_edu_B <- wages_edu %>% select(year,black_less_than_hs,black_high_school,black_some_college,bla
    head(wages_edu_B)
```

```
##   year black_less_than_hs black_high_school black_some_college
## 1 2022              15.19             19.39              21.34
## 2 2021              14.55             19.66              21.26
## 3 2020              14.55             19.66              21.79
## 4 2019              14.19             18.73              20.44
## 5 2018              13.31             18.14              19.98
## 6 2017              13.93             17.94              19.80
##   black_bachelors_degree black_advanced_degree
## 1                  33.39                 44.67
## 2                  32.62                 43.37
## 3                  33.64                 45.76
## 4                  31.83                 42.73
## 5                  32.00                 42.22
## 6                  31.49                 40.68
```

```
    tidy_wages_edu_B <- gather(wages_edu_B, edu_level, pay, `black_less_than_hs`:`black_advanced_degree
    head(tidy_wages_edu_B)
```

```
##   year          edu_level   pay
```

```
## 1 2022 black_less_than_hs 15.19
## 2 2021 black_less_than_hs 14.55
## 3 2020 black_less_than_hs 14.55
## 4 2019 black_less_than_hs 14.19
## 5 2018 black_less_than_hs 13.31
## 6 2017 black_less_than_hs 13.93
```

```
    tidy_wages_edu_B <-tidy_wages_edu_B %>% mutate(gender="All", race="Black")%>% mutate(edu_5=c(1:250))

    tidy_wages_edu_B$edu_5[1:50]="e"
    tidy_wages_edu_B$edu_5[51:100]="d"
    tidy_wages_edu_B$edu_5[101:150]="c"
    tidy_wages_edu_B$edu_5[151:200]="b"
    tidy_wages_edu_B$edu_5[201:250]="a"

    df6B4<-tidy_wages_edu_B

    #Review of the new data frame, structure and visualization:
    head(df6B4)
```

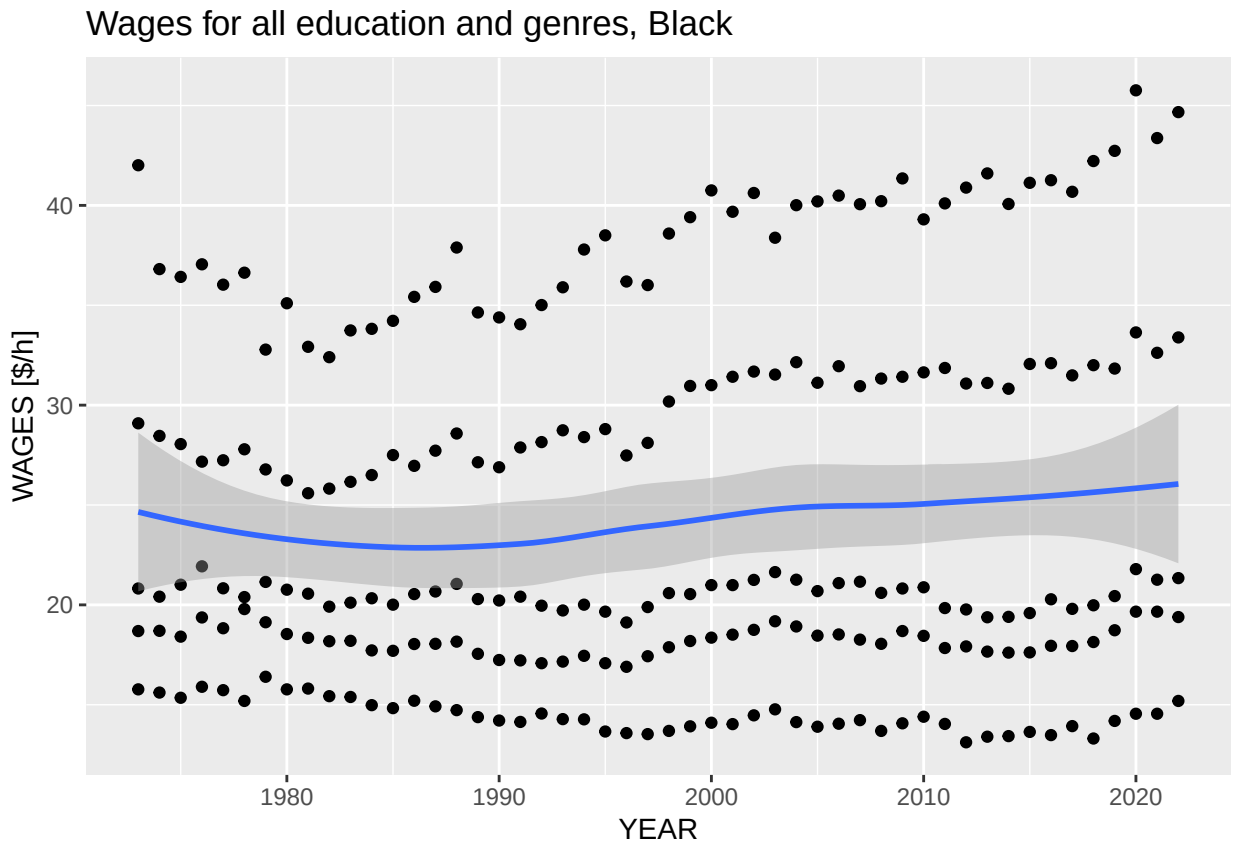
```
##   year          edu_level   pay gender  race edu_5
## 1 2022 black_less_than_hs 15.19    All Black     e
## 2 2021 black_less_than_hs 14.55    All Black     e
## 3 2020 black_less_than_hs 14.55    All Black     e
## 4 2019 black_less_than_hs 14.19    All Black     e
## 5 2018 black_less_than_hs 13.31    All Black     e
## 6 2017 black_less_than_hs 13.93    All Black     e
```

```
    str(df6B4)
```

```
## 'data.frame':    250 obs. of  6 variables:
##  $ year     : int  2022 2021 2020 2019 2018 2017 2016 2015 2014 2013 ...
##  $ edu_level: chr  "black_less_than_hs" "black_less_than_hs" "black_less_than_hs" "black_less_than_hs
##  $ pay      : num  15.2 14.6 14.6 14.2 13.3 ...
##  $ gender   : chr  "All" "All" "All" "All" ...
##  $ race     : chr  "Black" "Black" "Black" "Black" ...
##  $ edu_5    : chr  "e" "e" "e" "e" ...
```

```
    qplot(year,pay,data=df6B4)+geom_smooth()+
      xlab("YEAR") +
      ylab("WAGES [$/h]") +
      ggtitle("Wages for all education and genres, Black")
```

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



Preparing the fifth of the 12 data.frames, 6B.5:

6B.5_wages_edu_W: year=all ; gen=all ; race=White ; level_edu=all five ; pay=values

```
    wages_edu_W <- wages_edu %>% select(year,white_less_than_hs,white_high_school,white_some_college,wh
    head(wages_edu_W)
```

```
##   year white_less_than_hs white_high_school white_some_college
## 1 2022              15.70             23.31              26.28
## 2 2021              16.20             23.60              26.29
## 3 2020              17.01             24.00              26.90
## 4 2019              15.89             22.94              25.47
## 5 2018              16.05             23.02              25.15
## 6 2017              15.96             22.71              24.65
##   white_bachelors_degree white_advanced_degree
## 1                  43.30                53.30
## 2                  43.06                53.62
## 3                  43.16                53.80
## 4                  41.06                51.81
## 5                  40.49                51.80
## 6                  40.07                50.03
```

```
    tidy_wages_edu_W <- gather(wages_edu_W, edu_level, pay, `white_less_than_hs`:`white_advanced_degree
    head(tidy_wages_edu_W)
```

```
##   year          edu_level   pay
```

```
## 1 2022 white_less_than_hs 15.70
## 2 2021 white_less_than_hs 16.20
## 3 2020 white_less_than_hs 17.01
## 4 2019 white_less_than_hs 15.89
## 5 2018 white_less_than_hs 16.05
## 6 2017 white_less_than_hs 15.96
```

```
    tidy_wages_edu_W <-tidy_wages_edu_W %>% mutate(gender="All", race="White")%>% mutate(edu_5=c(1:250)
    tidy_wages_edu_W$edu_5[1:50]="e"
    tidy_wages_edu_W$edu_5[51:100]="d"
    tidy_wages_edu_W$edu_5[101:150]="c"
    tidy_wages_edu_W$edu_5[151:200]="b"
    tidy_wages_edu_W$edu_5[201:250]="a"

    df6B5<-tidy_wages_edu_W

    #Review of the new data frame, structure and visualization:
    head(df6B5)
```

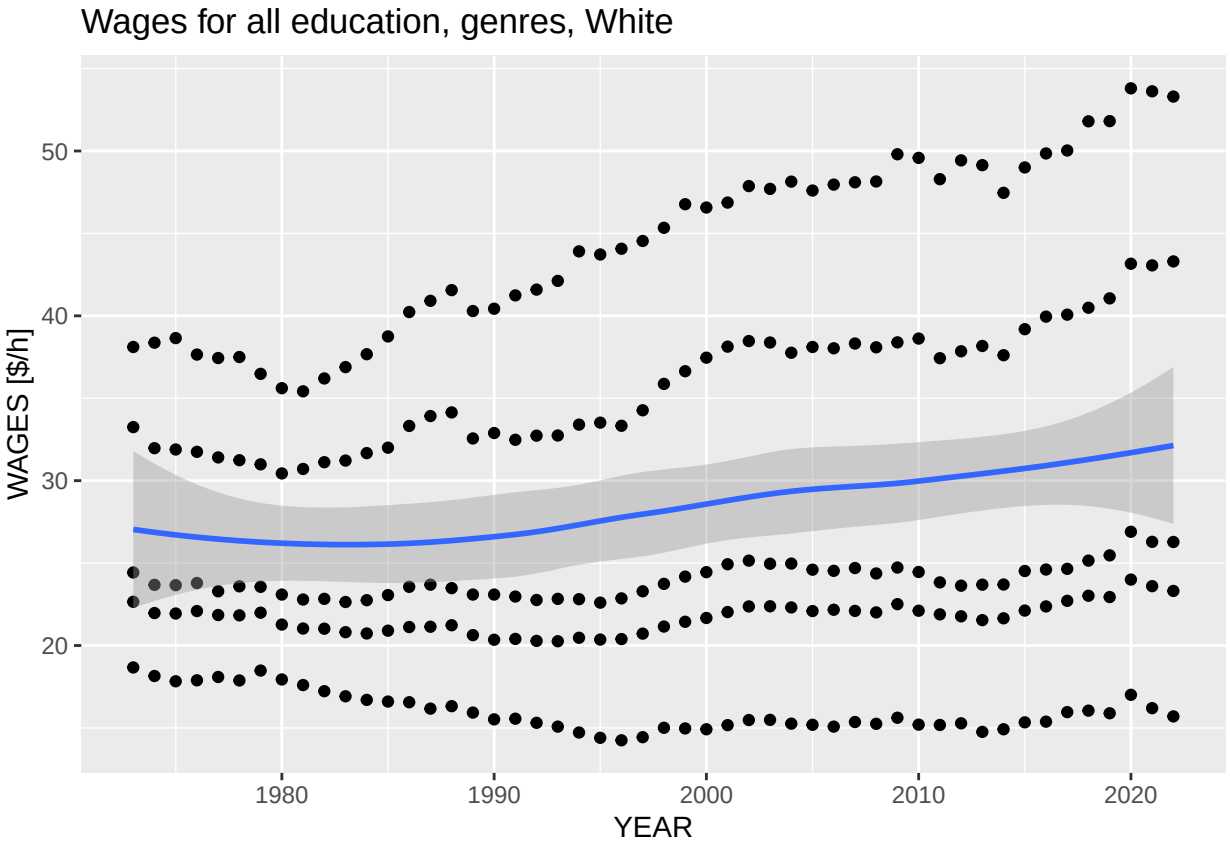
```
##   year          edu_level   pay gender  race edu_5
## 1 2022 white_less_than_hs 15.70    All White     e
## 2 2021 white_less_than_hs 16.20    All White     e
## 3 2020 white_less_than_hs 17.01    All White     e
## 4 2019 white_less_than_hs 15.89    All White     e
## 5 2018 white_less_than_hs 16.05    All White     e
## 6 2017 white_less_than_hs 15.96    All White     e
```

```
    str(df6B5)
```

```
## 'data.frame':    250 obs. of  6 variables:
##  $ year     : int  2022 2021 2020 2019 2018 2017 2016 2015 2014 2013 ...
##  $ edu_level: chr  "white_less_than_hs" "white_less_than_hs" "white_less_than_hs" "white_less_than_hs
##  $ pay      : num  15.7 16.2 17 15.9 16.1 ...
##  $ gender   : chr  "All" "All" "All" "All" ...
##  $ race     : chr  "White" "White" "White" "White" ...
##  $ edu_5    : chr  "e" "e" "e" "e" ...
```

```
    qplot(year,pay,data=df6B5)+geom_smooth()+
      xlab("YEAR") +
      ylab("WAGES [$/h]") +
      ggtitle("Wages for all education, genres, White")
```

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



Preparing the sixth of the 12 data.frames, 6B.6:

6B.6_wages_edu_H: year=all ; gen=all ; race=Hispanic ; level_edu=all five; pay=values

```
    wages_edu_H <- wages_edu %>% select(year,hispanic_less_than_hs,hispanic_high_school,hispanic_some_c
    head(wages_edu_H)
```

```
##   year hispanic_less_than_hs hispanic_high_school hispanic_some_college
## 1 2022                17.32               20.72                 22.96
## 2 2021                17.60               21.03                 23.34
## 3 2020                17.61               21.46                 23.24
## 4 2019                16.71               20.47                 22.01
## 5 2018                16.44               20.13                 21.74
## 6 2017                16.34               19.83                 21.59
##   hispanic_bachelors_degree hispanic_advanced_degree
## 1                     36.00                   48.30
## 2                     35.26                   48.60
## 3                     36.13                   48.58
## 4                     34.66                   46.69
## 5                     33.19                   44.83
## 6                     33.07                   43.92
```

```
    tidy_wages_edu_H <- gather(wages_edu_H, edu_level, pay, `hispanic_less_than_hs`:`hispanic_advanced_
    head(tidy_wages_edu_H)
```

```
##   year             edu_level   pay
```

```
## 1 2022 hispanic_less_than_hs 17.32
## 2 2021 hispanic_less_than_hs 17.60
## 3 2020 hispanic_less_than_hs 17.61
## 4 2019 hispanic_less_than_hs 16.71
## 5 2018 hispanic_less_than_hs 16.44
## 6 2017 hispanic_less_than_hs 16.34
```

```
    tidy_wages_edu_H <-tidy_wages_edu_H %>% mutate(gender="All", race="Hispanic")%>% mutate(edu_5=c(1:25

    tidy_wages_edu_H$edu_5[1:50]="e"
    tidy_wages_edu_H$edu_5[51:100]="d"
    tidy_wages_edu_H$edu_5[101:150]="c"
    tidy_wages_edu_H$edu_5[151:200]="b"
    tidy_wages_edu_H$edu_5[201:250]="a"

    df6B6<-tidy_wages_edu_H

    #Review of the new data frame, structure and visualization:
    head(df6B6)
```

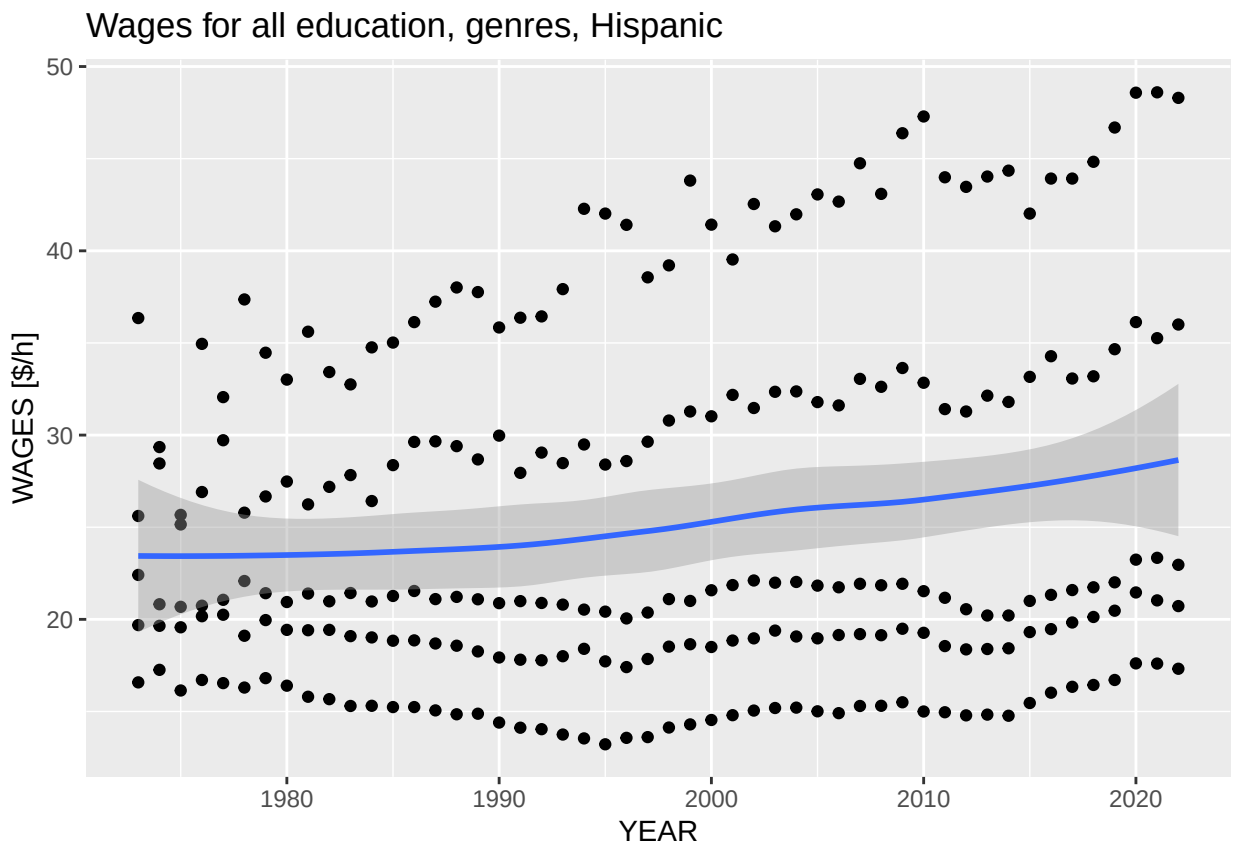
```
##   year             edu_level   pay gender     race edu_5
## 1 2022 hispanic_less_than_hs 17.32    All Hispanic     e
## 2 2021 hispanic_less_than_hs 17.60    All Hispanic     e
## 3 2020 hispanic_less_than_hs 17.61    All Hispanic     e
## 4 2019 hispanic_less_than_hs 16.71    All Hispanic     e
## 5 2018 hispanic_less_than_hs 16.44    All Hispanic     e
## 6 2017 hispanic_less_than_hs 16.34    All Hispanic     e
```

```
    str(df6B6)
```

```
## 'data.frame':    250 obs. of  6 variables:
##  $ year     : int  2022 2021 2020 2019 2018 2017 2016 2015 2014 2013 ...
##  $ edu_level: chr  "hispanic_less_than_hs" "hispanic_less_than_hs" "hispanic_less_than_hs" "hispanic_
##  $ pay      : num  17.3 17.6 17.6 16.7 16.4 ...
##  $ gender   : chr  "All" "All" "All" "All" ...
##  $ race     : chr  "Hispanic" "Hispanic" "Hispanic" "Hispanic" ...
##  $ edu_5    : chr  "e" "e" "e" "e" ...
```

```
    qplot(year,pay,data=df6B6)+geom_smooth()+
      xlab("YEAR") +
      ylab("WAGES [$/h]") +
      ggtitle("Wages for all education, genres, Hispanic")
```

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



Preparing the seventh of the 12 data.frames, 6A.1:

6A.1_wages_edu_M_B: year=all ; gen=Men ; race=Black ; level_edu=all five; pay=values

```
    wages_edu_M_B <- wages_edu %>% select(year,black_men_less_than_hs,black_men_high_school,black_men_s
    head(wages_edu_M_B)
```

```
##   year black_men_less_than_hs black_men_high_school black_men_some_college
## 1 2022                 16.38                20.73                 22.58
## 2 2021                 15.38                20.86                 22.63
## 3 2020                 15.52                21.22                 23.57
## 4 2019                 15.07                20.22                 22.02
## 5 2018                 14.27                19.36                 21.82
## 6 2017                 15.00                19.46                 21.09
##   black_men_bachelors_degree black_men_advanced_degree
## 1                      37.63                    52.91
## 2                      36.95                    49.01
## 3                      36.70                    51.67
## 4                      33.73                    46.87
## 5                      33.42                    46.42
## 6                      34.41                    44.62
```

```
    tidy_wages_edu_M_B <- gather(wages_edu_M_B, edu_level, pay, `black_men_less_than_hs`:`black_men_adva
    head(tidy_wages_edu_M_B)
```

```
##   year             edu_level   pay
```

```
## 1 2022 black_men_less_than_hs 16.38
## 2 2021 black_men_less_than_hs 15.38
## 3 2020 black_men_less_than_hs 15.52
## 4 2019 black_men_less_than_hs 15.07
## 5 2018 black_men_less_than_hs 14.27
## 6 2017 black_men_less_than_hs 15.00
```

```
    tidy_wages_edu_M_B <-tidy_wages_edu_M_B %>% mutate(gender="Men", race="Black")%>% mutate(edu_5=c(1:1

    tidy_wages_edu_M_B$edu_5[1:50]="e"
    tidy_wages_edu_M_B$edu_5[51:100]="d"
    tidy_wages_edu_M_B$edu_5[101:150]="c"
    tidy_wages_edu_M_B$edu_5[151:200]="b"
    tidy_wages_edu_M_B$edu_5[201:250]="a"

    df6A1<-tidy_wages_edu_M_B

    #Review of the new data frame, structure and visualization:
    head(df6A1)
```

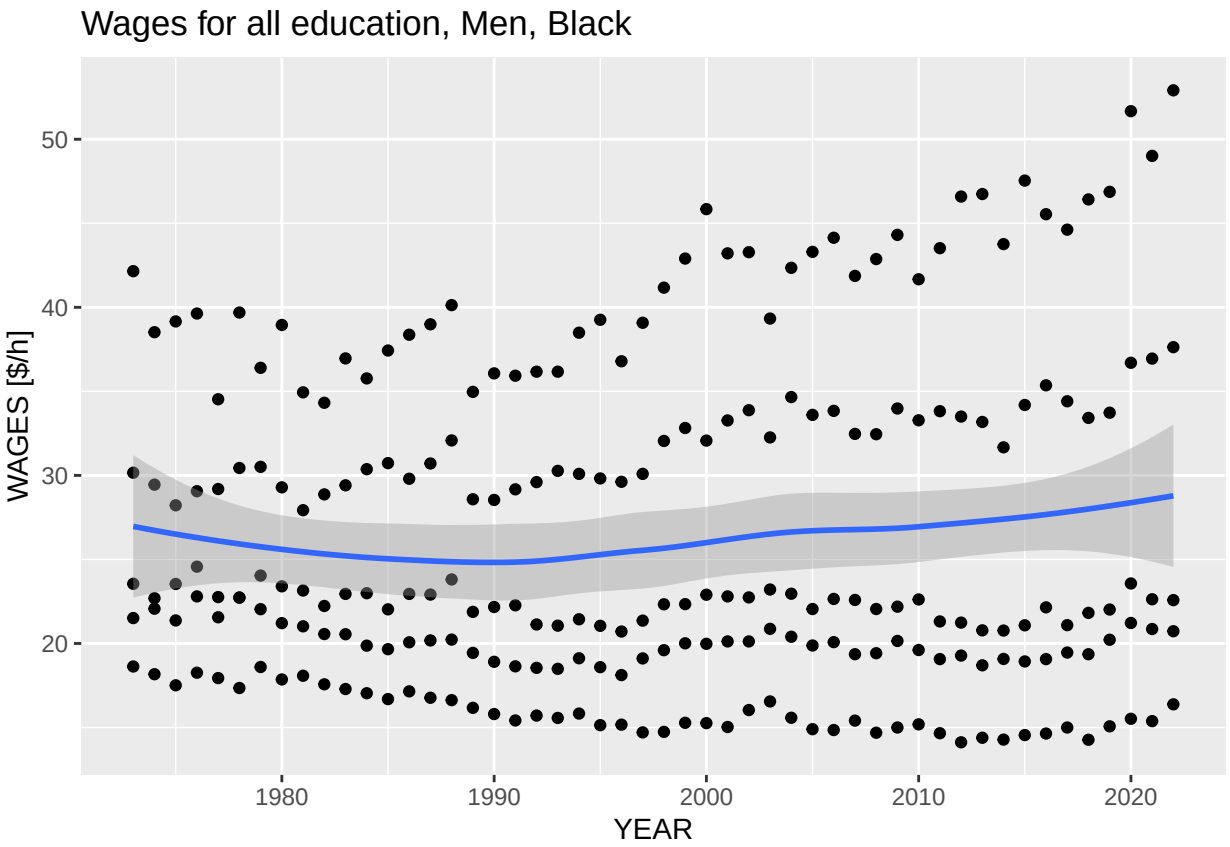
```
##   year              edu_level   pay gender  race edu_5
## 1 2022 black_men_less_than_hs 16.38    Men Black     e
## 2 2021 black_men_less_than_hs 15.38    Men Black     e
## 3 2020 black_men_less_than_hs 15.52    Men Black     e
## 4 2019 black_men_less_than_hs 15.07    Men Black     e
## 5 2018 black_men_less_than_hs 14.27    Men Black     e
## 6 2017 black_men_less_than_hs 15.00    Men Black     e
```

```
    str(df6A1)
```

```
## 'data.frame':    250 obs. of  6 variables:
##  $ year     : int  2022 2021 2020 2019 2018 2017 2016 2015 2014 2013 ...
##  $ edu_level: chr  "black_men_less_than_hs" "black_men_less_than_hs" "black_men_less_than_hs" "black
##  $ pay      : num  16.4 15.4 15.5 15.1 14.3 ...
##  $ gender   : chr  "Men" "Men" "Men" "Men" ...
##  $ race     : chr  "Black" "Black" "Black" "Black" ...
##  $ edu_5    : chr  "e" "e" "e" "e" ...
```

```
    qplot(year,pay,data=df6A1)+geom_smooth()+
      xlab("YEAR") +
      ylab("WAGES [$/h]") +
      ggtitle("Wages for all education, Men, Black")
```

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



Preparing the eighth of the 12 data.frames, 6A.2:

6A.2_wages_edu_M_W: year=all ; gen=Men ; race=White ; level_edu=all five ; pay=values

```
    wages_edu_M_W <- wages_edu %>% select(year,white_men_less_than_hs,white_men_high_school,white_men_s
    head(wages_edu_M_W)
```

```
##   year white_men_less_than_hs white_men_high_school white_men_some_college
## 1 2022                 17.14                25.92                 29.93
## 2 2021                 18.13                26.03                 29.79
## 3 2020                 18.97                26.73                 30.40
## 4 2019                 17.66                25.58                 28.90
## 5 2018                 17.96                25.49                 28.41
## 6 2017                 17.82                25.14                 27.61
##   white_men_bachelors_degree white_men_advanced_degree
## 1                      51.23                    63.86
## 2                      50.06                    64.04
## 3                      50.15                    63.50
## 4                      47.76                    60.84
## 5                      47.28                    61.27
## 6                      46.44                    57.64
```

```
    tidy_wages_edu_M_W <- gather(wages_edu_M_W, edu_level, pay, `white_men_less_than_hs`:`white_men_adva
    head(tidy_wages_edu_M_W)
```

```
##   year             edu_level   pay
```

```
## 1 2022 white_men_less_than_hs 17.14
## 2 2021 white_men_less_than_hs 18.13
## 3 2020 white_men_less_than_hs 18.97
## 4 2019 white_men_less_than_hs 17.66
## 5 2018 white_men_less_than_hs 17.96
## 6 2017 white_men_less_than_hs 17.82
```

```
    tidy_wages_edu_M_W <-tidy_wages_edu_M_W %>% mutate(gender="Men", race="White")%>% mutate(edu_5=c(1:1

    tidy_wages_edu_M_W$edu_5[1:50]="e"
    tidy_wages_edu_M_W$edu_5[51:100]="d"
    tidy_wages_edu_M_W$edu_5[101:150]="c"
    tidy_wages_edu_M_W$edu_5[151:200]="b"
    tidy_wages_edu_M_W$edu_5[201:250]="a"

    df6A2<-tidy_wages_edu_M_W

    #Review of the new data frame, structure and visualization:
    head(df6A2)
```

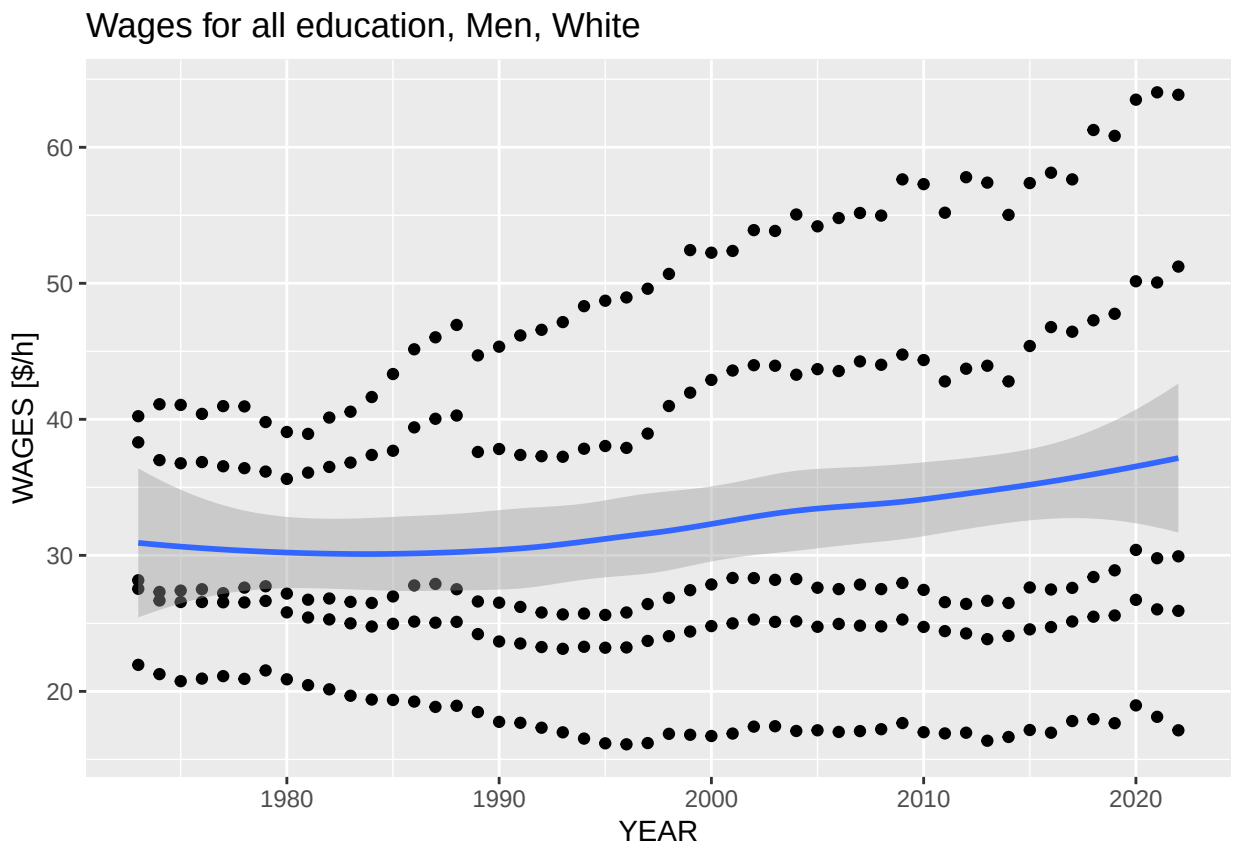
```
##   year              edu_level   pay gender  race edu_5
## 1 2022 white_men_less_than_hs 17.14    Men White     e
## 2 2021 white_men_less_than_hs 18.13    Men White     e
## 3 2020 white_men_less_than_hs 18.97    Men White     e
## 4 2019 white_men_less_than_hs 17.66    Men White     e
## 5 2018 white_men_less_than_hs 17.96    Men White     e
## 6 2017 white_men_less_than_hs 17.82    Men White     e
```

```
    str(df6A2)
```

```
## 'data.frame':    250 obs. of  6 variables:
##  $ year     : int  2022 2021 2020 2019 2018 2017 2016 2015 2014 2013 ...
##  $ edu_level: chr  "white_men_less_than_hs" "white_men_less_than_hs" "white_men_less_than_hs" "white
##  $ pay      : num  17.1 18.1 19 17.7 18 ...
##  $ gender   : chr  "Men" "Men" "Men" "Men" ...
##  $ race     : chr  "White" "White" "White" "White" ...
##  $ edu_5    : chr  "e" "e" "e" "e" ...
```

```
    qplot(year,pay,data=df6A2)+geom_smooth()+
      xlab("YEAR") +
      ylab("WAGES [$/h]") +
      ggtitle("Wages for all education, Men, White")
```

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



Preparing the ninth of the 12 data.frames, 6A.3:

6A.3_wages_edu_M_H: year=all ; gen=Men ; race=Hispanic ; level_edu=all five; pay=values

```
    wages_edu_M_H <- wages_edu %>% select(year,hispanic_men_less_than_hs,hispanic_men_high_school,hispa
    head(wages_edu_M_H)
```

```
##   year hispanic_men_less_than_hs hispanic_men_high_school
## 1 2022                    18.67                    22.32
## 2 2021                    18.98                    22.70
## 3 2020                    19.26                    23.33
## 4 2019                    17.93                    22.33
## 5 2018                    18.05                    22.04
## 6 2017                    17.91                    21.68
##   hispanic_men_some_college hispanic_men_bachelors_degree
## 1                     25.49                         41.48
## 2                     25.58                         39.61
## 3                     25.69                         41.00
## 4                     24.35                         39.56
## 5                     24.20                         37.27
## 6                     23.59                         36.62
##   hispanic_men_advanced_degree
## 1                        57.08
## 2                        55.60
## 3                        53.58
## 4                        52.24
```

```
## 5                        51.27
## 6                        50.58
```

```
    tidy_wages_edu_M_H <- gather(wages_edu_M_H, edu_level, pay, `hispanic_men_less_than_hs`:`hispanic_m
    head(tidy_wages_edu_M_H)
```

```
##   year                edu_level   pay
## 1 2022 hispanic_men_less_than_hs 18.67
## 2 2021 hispanic_men_less_than_hs 18.98
## 3 2020 hispanic_men_less_than_hs 19.26
## 4 2019 hispanic_men_less_than_hs 17.93
## 5 2018 hispanic_men_less_than_hs 18.05
## 6 2017 hispanic_men_less_than_hs 17.91
```

```
    tidy_wages_edu_M_H <-tidy_wages_edu_M_H %>% mutate(gender="Men", race="Hispanic")%>% mutate(edu_5=c

    tidy_wages_edu_M_H$edu_5[1:50]="e"
    tidy_wages_edu_M_H$edu_5[51:100]="d"
    tidy_wages_edu_M_H$edu_5[101:150]="c"
    tidy_wages_edu_M_H$edu_5[151:200]="b"
    tidy_wages_edu_M_H$edu_5[201:250]="a"

    df6A3<-tidy_wages_edu_M_H

    #Review of the new data frame, structure and visualization:
    head(df6A3)
```

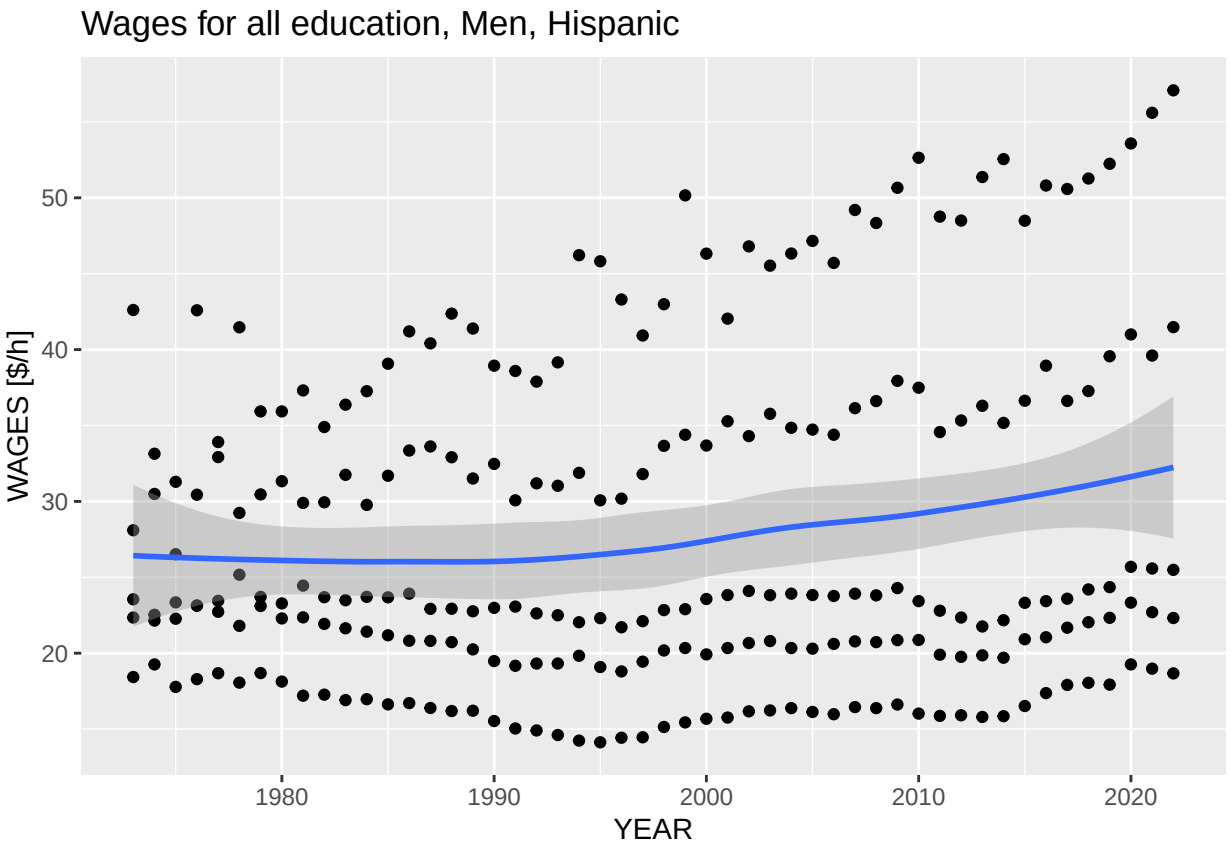
```
##   year                edu_level   pay gender     race edu_5
## 1 2022 hispanic_men_less_than_hs 18.67    Men Hispanic     e
## 2 2021 hispanic_men_less_than_hs 18.98    Men Hispanic     e
## 3 2020 hispanic_men_less_than_hs 19.26    Men Hispanic     e
## 4 2019 hispanic_men_less_than_hs 17.93    Men Hispanic     e
## 5 2018 hispanic_men_less_than_hs 18.05    Men Hispanic     e
## 6 2017 hispanic_men_less_than_hs 17.91    Men Hispanic     e
```

```
    str(df6A3)
```

```
## 'data.frame':    250 obs. of  6 variables:
##  $ year     : int  2022 2021 2020 2019 2018 2017 2016 2015 2014 2013 ...
##  $ edu_level: chr  "hispanic_men_less_than_hs" "hispanic_men_less_than_hs" "hispanic_men_less_than_hs
##  $ pay      : num  18.7 19 19.3 17.9 18.1 ...
##  $ gender   : chr  "Men" "Men" "Men" "Men" ...
##  $ race     : chr  "Hispanic" "Hispanic" "Hispanic" "Hispanic" ...
##  $ edu_5    : chr  "e" "e" "e" "e" ...
```

```
    qplot(year,pay,data=df6A3)+geom_smooth()+
      xlab("YEAR") +
      ylab("WAGES [$/h]") +
      ggtitle("Wages for all education, Men, Hispanic")
```

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



Preparing the tenth of the 12 data.frames, 6A.4:

6A.4_wages_edu_W_B: year=all ; gen=Women ; race=Black ;level_edu=all five; pay=values

```
    wages_edu_W_B <- wages_edu %>% select(year,black_women_less_than_hs,black_women_high_school,black_w
    head(wages_edu_W_B)
```

```
##   year black_women_less_than_hs black_women_high_school
## 1 2022                   13.89                   17.83
## 2 2021                   13.73                   18.30
## 3 2020                   13.66                   17.93
## 4 2019                   13.30                   17.18
## 5 2018                   12.48                   16.81
## 6 2017                   12.99                   16.33
##   black_women_some_college black_women_bachelors_degree
## 1                    20.36                         29.94
## 2                    20.18                         29.35
## 3                    20.31                         31.38
## 4                    19.17                         30.31
## 5                    18.57                         30.85
## 6                    18.76                         29.19
##   black_women_advanced_degree
## 1                       39.41
## 2                       40.07
## 3                       42.44
## 4                       40.42
```

```
## 5                       39.64
## 6                       38.26
```

```
    tidy_wages_edu_W_B <- gather(wages_edu_W_B, edu_level, pay, `black_women_less_than_hs`:`black_women
    head(tidy_wages_edu_W_B)
```

```
##   year               edu_level   pay
## 1 2022 black_women_less_than_hs 13.89
## 2 2021 black_women_less_than_hs 13.73
## 3 2020 black_women_less_than_hs 13.66
## 4 2019 black_women_less_than_hs 13.30
## 5 2018 black_women_less_than_hs 12.48
## 6 2017 black_women_less_than_hs 12.99
```

```
    tidy_wages_edu_W_B <-tidy_wages_edu_M_B %>% mutate(gender="Women", race="Black")%>% mutate(edu_5=c(

    tidy_wages_edu_W_B$edu_5[1:50]="e"
    tidy_wages_edu_W_B$edu_5[51:100]="d"
    tidy_wages_edu_W_B$edu_5[101:150]="c"
    tidy_wages_edu_W_B$edu_5[151:200]="b"
    tidy_wages_edu_W_B$edu_5[201:250]="a"

    df6A4<-tidy_wages_edu_W_B

    #Review of the new data frame, structure and visualization:
    head(df6A4)
```

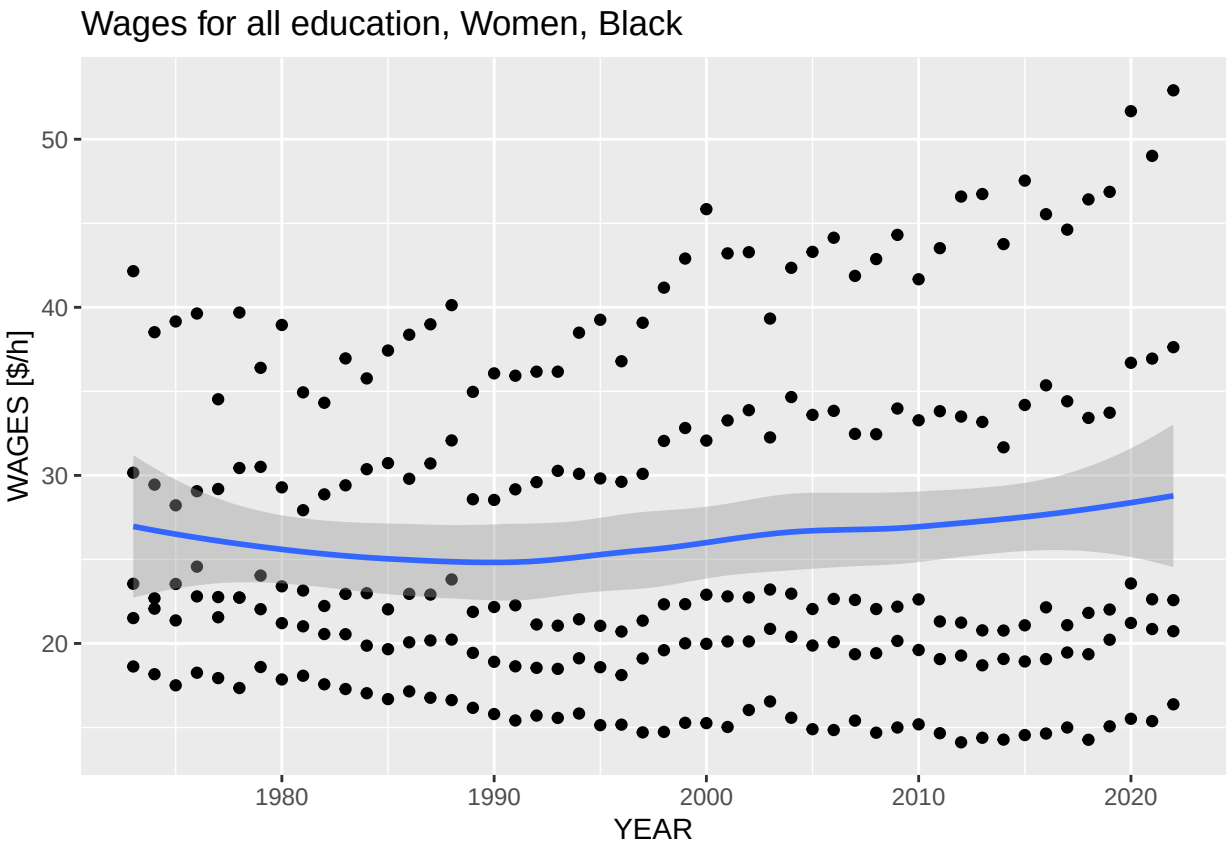
```
##   year              edu_level   pay gender  race edu_5
## 1 2022 black_men_less_than_hs 16.38  Women Black     e
## 2 2021 black_men_less_than_hs 15.38  Women Black     e
## 3 2020 black_men_less_than_hs 15.52  Women Black     e
## 4 2019 black_men_less_than_hs 15.07  Women Black     e
## 5 2018 black_men_less_than_hs 14.27  Women Black     e
## 6 2017 black_men_less_than_hs 15.00  Women Black     e
```

```
    str(df6A4)
```

```
## 'data.frame':    250 obs. of  6 variables:
##  $ year     : int  2022 2021 2020 2019 2018 2017 2016 2015 2014 2013 ...
##  $ edu_level: chr  "black_men_less_than_hs" "black_men_less_than_hs" "black_men_less_than_hs" "black
##  $ pay      : num  16.4 15.4 15.5 15.1 14.3 ...
##  $ gender   : chr  "Women" "Women" "Women" "Women" ...
##  $ race     : chr  "Black" "Black" "Black" "Black" ...
##  $ edu_5    : chr  "e" "e" "e" "e" ...
```

```
    qplot(year,pay,data=df6A4)+geom_smooth()+
      xlab("YEAR") +
      ylab("WAGES [$/h]") +
      ggtitle("Wages for all education, Women, Black")
```

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



Preparing the eleventh of the 12 data.frames, 6A.5:

6A.5_wages_edu_W_W: year=all ; gen=Women ; race=White ;level_edu=all five ; pay=values

```
    wages_edu_W_W <- wages_edu %>% select(year,white_women_less_than_hs,white_women_high_school,white_w
    head(wages_edu_W_W)
```

```
##   year white_women_less_than_hs white_women_high_school
## 1 2022                   13.84                   19.56
## 2 2021                   13.77                   20.08
## 3 2020                   14.20                   20.06
## 4 2019                   13.44                   19.22
## 5 2018                   13.45                   19.52
## 6 2017                   13.46                   19.34
##   white_women_some_college white_women_bachelors_degree
## 1                    22.52                         35.31
## 2                    22.73                         36.11
## 3                    23.31                         36.20
## 4                    21.99                         34.47
## 5                    21.90                         33.78
## 6                    21.72                         33.76
##   white_women_advanced_degree
## 1                       44.45
## 2                       44.82
## 3                       45.58
## 4                       43.83
```

```
## 5                       43.31
## 6                       43.22
```

```
    tidy_wages_edu_W_W <- gather(wages_edu_W_W, edu_level, pay, `white_women_less_than_hs`:`white_women
    head(tidy_wages_edu_W_W)
```

```
##   year               edu_level   pay
## 1 2022 white_women_less_than_hs 13.84
## 2 2021 white_women_less_than_hs 13.77
## 3 2020 white_women_less_than_hs 14.20
## 4 2019 white_women_less_than_hs 13.44
## 5 2018 white_women_less_than_hs 13.45
## 6 2017 white_women_less_than_hs 13.46
```

```
    tidy_wages_edu_W_W <-tidy_wages_edu_W_W %>% mutate(gender="Women", race="White")%>% mutate(edu_5=c(

    tidy_wages_edu_W_W$edu_5[1:50]="e"
    tidy_wages_edu_W_W$edu_5[51:100]="d"
    tidy_wages_edu_W_W$edu_5[101:150]="c"
    tidy_wages_edu_W_W$edu_5[151:200]="b"
    tidy_wages_edu_W_W$edu_5[201:250]="a"

    df6A5<-tidy_wages_edu_W_W

    #Review of the new data frame, structure and visualization:
    head(df6A5)
```

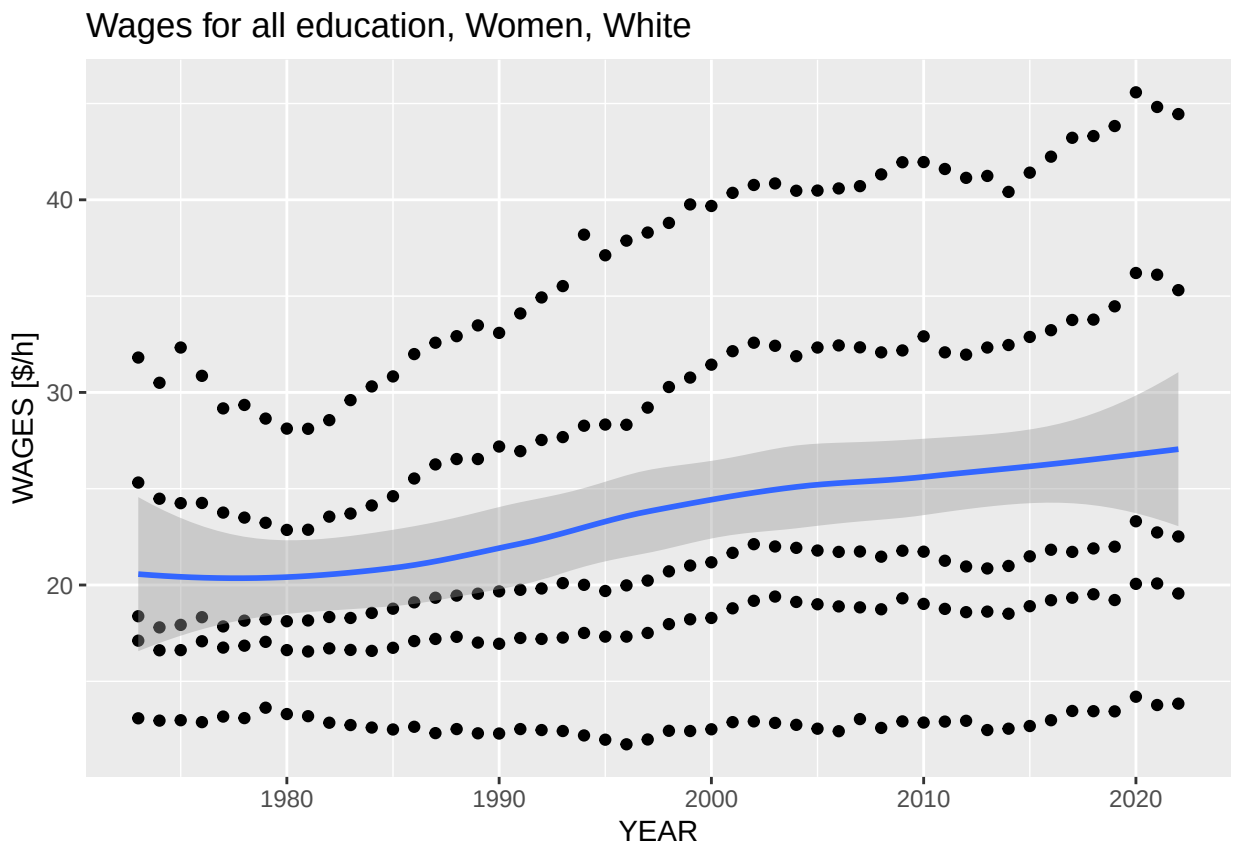
```
##   year               edu_level   pay gender  race edu_5
## 1 2022 white_women_less_than_hs 13.84  Women White     e
## 2 2021 white_women_less_than_hs 13.77  Women White     e
## 3 2020 white_women_less_than_hs 14.20  Women White     e
## 4 2019 white_women_less_than_hs 13.44  Women White     e
## 5 2018 white_women_less_than_hs 13.45  Women White     e
## 6 2017 white_women_less_than_hs 13.46  Women White     e
```

```
    str(df6A5)
```

```
## 'data.frame':    250 obs. of  6 variables:
##  $ year     : int  2022 2021 2020 2019 2018 2017 2016 2015 2014 2013 ...
##  $ edu_level: chr  "white_women_less_than_hs" "white_women_less_than_hs" "white_women_less_than_hs" "
##  $ pay      : num  13.8 13.8 14.2 13.4 13.4 ...
##  $ gender   : chr  "Women" "Women" "Women" "Women" ...
##  $ race     : chr  "White" "White" "White" "White" ...
##  $ edu_5    : chr  "e" "e" "e" "e" ...
```

```
    qplot(year,pay,data=df6A5)+geom_smooth()+
      xlab("YEAR") +
      ylab("WAGES [$/h]") +
      ggtitle("Wages for all education, Women, White")
```

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



Preparing the twelfth of the 12 data.frames, 6A.6:

6A.6_wages_edu_W_H: year=all ; gen=Women ; race=Hispanic ;level_edu=all five; pay=values

```
    wages_edu_W_H <- wages_edu %>% select(year,hispanic_women_less_than_hs,hispanic_women_high_school,hi
    head(wages_edu_W_H)
```

```
##   year hispanic_women_less_than_hs hispanic_women_high_school
## 1 2022                       14.74                       18.18
## 2 2021                       14.97                       18.34
## 3 2020                       14.58                       18.50
## 4 2019                       14.50                       17.71
## 5 2018                       13.47                       17.28
## 6 2017                       13.36                       17.02
##   hispanic_women_some_college hispanic_women_bachelors_degree
## 1                       20.64                           31.13
## 2                       21.14                           31.25
## 3                       20.69                           31.55
## 4                       19.69                           30.18
## 5                       19.29                           29.47
## 6                       19.60                           29.69
##   hispanic_women_advanced_degree
## 1                          40.64
## 2                          42.47
## 3                          44.15
## 4                          42.30
```

```
## 5                          39.35
## 6                          38.43
```

```
    tidy_wages_edu_W_H <- gather(wages_edu_W_H, edu_level, pay, `hispanic_women_less_than_hs`:`hispanic_
    head(tidy_wages_edu_W_H)
```

```
##   year                   edu_level   pay
## 1 2022 hispanic_women_less_than_hs 14.74
## 2 2021 hispanic_women_less_than_hs 14.97
## 3 2020 hispanic_women_less_than_hs 14.58
## 4 2019 hispanic_women_less_than_hs 14.50
## 5 2018 hispanic_women_less_than_hs 13.47
## 6 2017 hispanic_women_less_than_hs 13.36
```

```
    tidy_wages_edu_W_H <-tidy_wages_edu_W_H %>% mutate(gender="Women", race="Hispanic")%>% mutate(edu_5=

    tidy_wages_edu_W_H$edu_5[1:50]="e"
    tidy_wages_edu_W_H$edu_5[51:100]="d"
    tidy_wages_edu_W_H$edu_5[101:150]="c"
    tidy_wages_edu_W_H$edu_5[151:200]="b"
    tidy_wages_edu_W_H$edu_5[201:250]="a"

    df6A6<- tidy_wages_edu_W_H

    #Review of the new data frame, structure and visualization:
    head(df6A6)
```

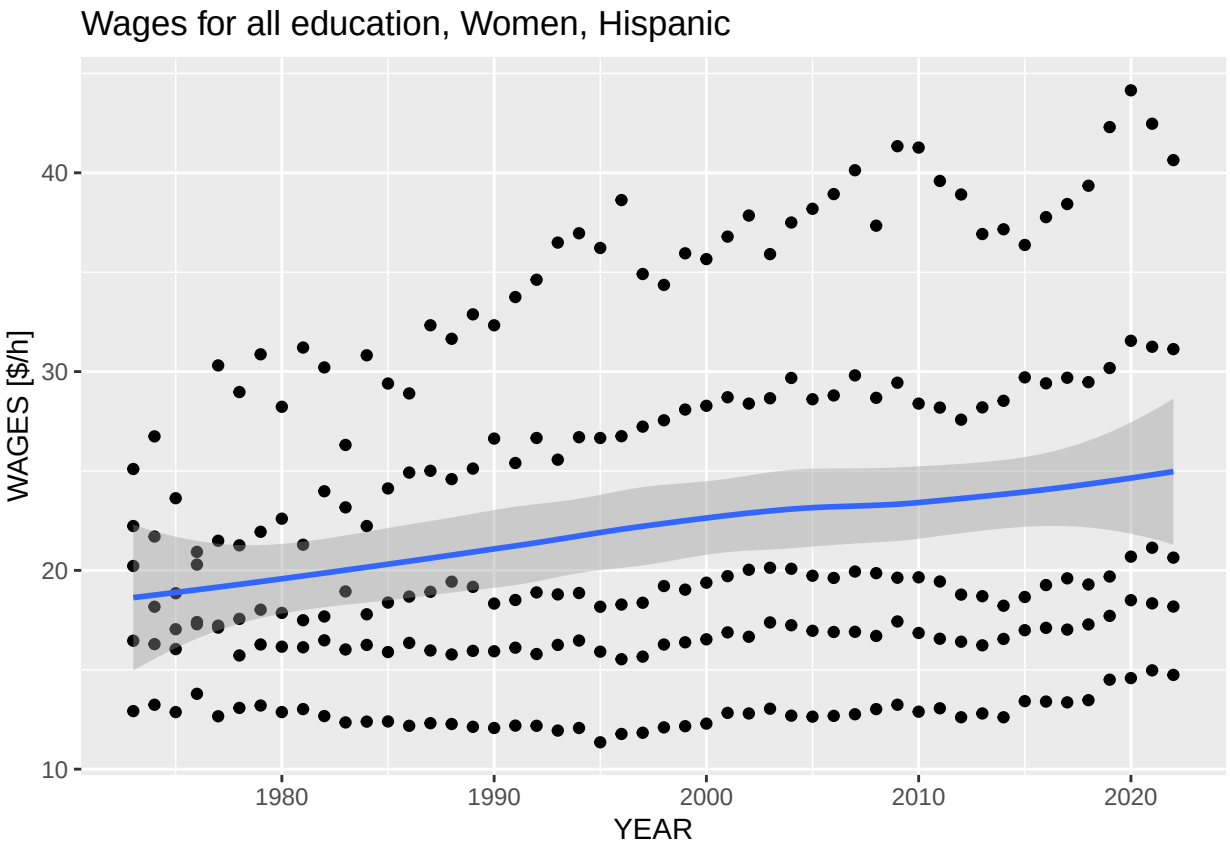
```
##   year                   edu_level   pay gender     race edu_5
## 1 2022 hispanic_women_less_than_hs 14.74  Women Hispanic     e
## 2 2021 hispanic_women_less_than_hs 14.97  Women Hispanic     e
## 3 2020 hispanic_women_less_than_hs 14.58  Women Hispanic     e
## 4 2019 hispanic_women_less_than_hs 14.50  Women Hispanic     e
## 5 2018 hispanic_women_less_than_hs 13.47  Women Hispanic     e
## 6 2017 hispanic_women_less_than_hs 13.36  Women Hispanic     e
```

```
    str(df6A6)
```

```
## 'data.frame':    250 obs. of  6 variables:
##  $ year     : int  2022 2021 2020 2019 2018 2017 2016 2015 2014 2013 ...
##  $ edu_level: chr  "hispanic_women_less_than_hs" "hispanic_women_less_than_hs" "hispanic_women_less_
##  $ pay      : num  14.7 15 14.6 14.5 13.5 ...
##  $ gender   : chr  "Women" "Women" "Women" "Women" ...
##  $ race     : chr  "Hispanic" "Hispanic" "Hispanic" "Hispanic" ...
##  $ edu_5    : chr  "e" "e" "e" "e" ...
```

```
    qplot(year,pay,data=df6A6)+geom_smooth()+
      xlab("YEAR") +
      ylab("WAGES [$/h]") +
      ggtitle("Wages for all education, Women, Hispanic")
```

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



Now we've made data tidy, and have created 12 data frames: 6A.1,6A.2,6A.3,6A.4,6A.5,6A.6; 6B.1,6B.2,6B.3,6B.4,6B.5,6B.6.

Let's combine them all together, so we can start data exploration-analysis. We need to add rows. We'll use the bind_rows function from tidyverse:

```
    df6A <- bind_rows(list(df6A1,df6A2,df6A3,df6A4,df6A5,df6A6))
    df6B <- bind_rows(list(df6B1,df6B2,df6B3,df6B4,df6B5,df6B6))
    df12 <- bind_rows(list(df6A,df6B))

    #Let's display the internal strcuture of the new formed data frame:
    head(df12)
```

```
##   year              edu_level   pay gender  race edu_5
## 1 2022 black_men_less_than_hs 16.38    Men Black     e
## 2 2021 black_men_less_than_hs 15.38    Men Black     e
## 3 2020 black_men_less_than_hs 15.52    Men Black     e
## 4 2019 black_men_less_than_hs 15.07    Men Black     e
## 5 2018 black_men_less_than_hs 14.27    Men Black     e
## 6 2017 black_men_less_than_hs 15.00    Men Black     e
```

```
    str(df12)
```

```
## 'data.frame':    3000 obs. of  6 variables:
##  $ year     : int  2022 2021 2020 2019 2018 2017 2016 2015 2014 2013 ...
##  $ edu_level: chr  "black_men_less_than_hs" "black_men_less_than_hs" "black_men_less_than_hs" "black
```

```
##  $ pay      : num  16.4 15.4 15.5 15.1 14.3 ...
##  $ gender   : chr  "Men" "Men" "Men" "Men" ...
##  $ race     : chr  "Black" "Black" "Black" "Black" ...
##  $ edu_5    : chr  "e" "e" "e" "e" ...
```

Level of education is now contained in two variables, "edu_level" and "edu_5". We'll create a data frame showing only one variable for education level, selecting "edu_5", since edu_5 does not include other additional (demographics) information:

```
    df12_t <- df12 %>% select(year,gender,race,edu_5,pay)

    head(df12_t)
```

```
##   year gender  race edu_5   pay
## 1 2022    Men Black     e 16.38
## 2 2021    Men Black     e 15.38
## 3 2020    Men Black     e 15.52
## 4 2019    Men Black     e 15.07
## 5 2018    Men Black     e 14.27
## 6 2017    Men Black     e 15.00
```

```
    str(df12_t)
```

```
## 'data.frame':    3000 obs. of  5 variables:
##  $ year  : int  2022 2021 2020 2019 2018 2017 2016 2015 2014 2013 ...
##  $ gender: chr  "Men" "Men" "Men" "Men" ...
##  $ race  : chr  "Black" "Black" "Black" "Black" ...
##  $ edu_5 : chr  "e" "e" "e" "e" ...
##  $ pay   : num  16.4 15.4 15.5 15.1 14.3 ...
```

We can compare "df12" and "df12_t", see how df12_t (Tidy) is the Tidiest version:

```
    unique_columns_count <-  df12 %>%
      summarise(n_year = n_distinct(year),
                n_edu_level = n_distinct(edu_level),
                n_gender = n_distinct(gender),
                n_race = n_distinct(race),
                n_edu_5=n_distinct(edu_5),
                n_pay = n_distinct(pay))
    print(unique_columns_count)
```

```
##   n_year n_edu_level n_gender n_race n_edu_5 n_pay
## 1     50          55        3      4       5  1857
```

```
    unique_columns_count <-  df12_t %>%
      summarise(n_year = n_distinct(year),
                n_gender = n_distinct(gender),
                n_race = n_distinct(race),
                n_edu_5=n_distinct(edu_5),
                n_pay = n_distinct(pay))
    print(unique_columns_count)
```

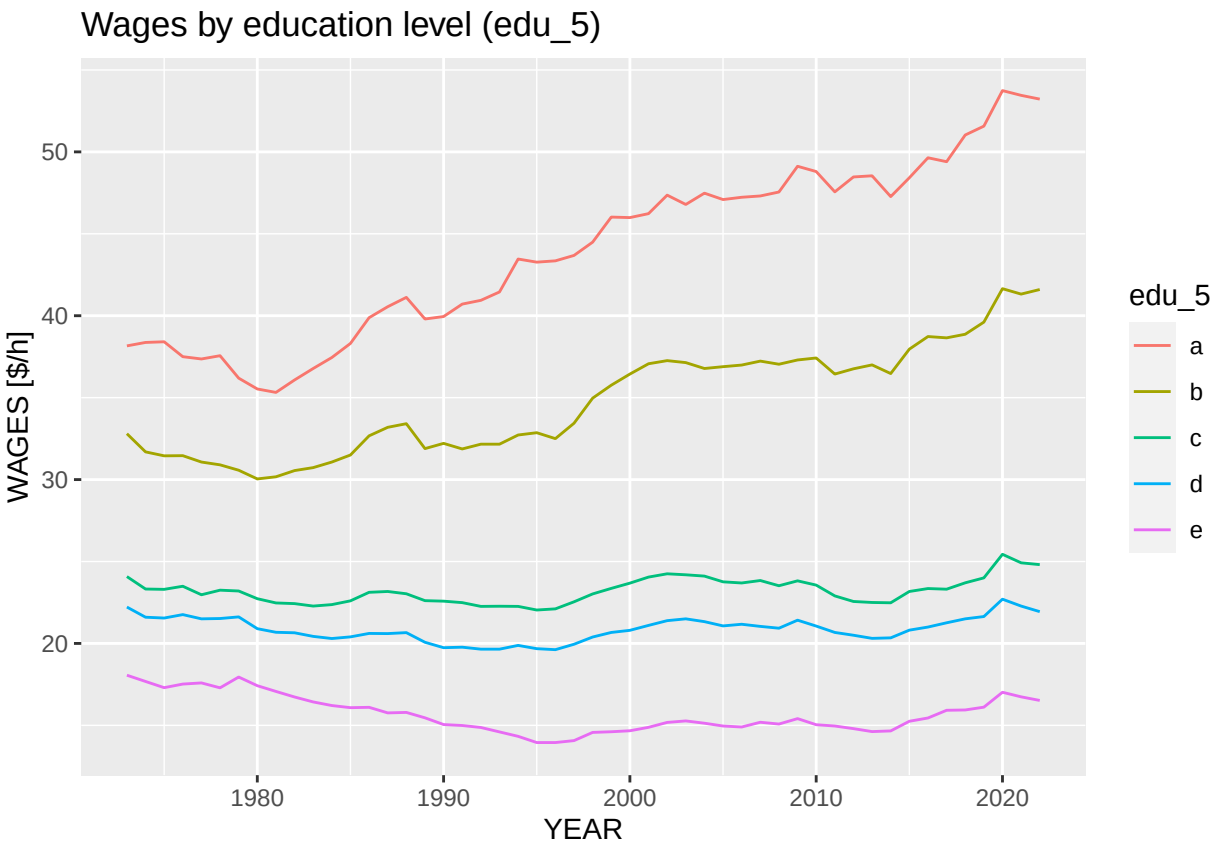
```
##   n_year n_gender n_race n_edu_5 n_pay
## 1     50        3      4       5  1857
```

Summary (df12_t -vs- df12): edu_level variable had 55 different values, while there are actually only 5 different levels of education being considered. "edu_level" included gender and race variables within. We have now simplified these 55 values, to the 5 actual education level values, without loosing the additional information contained in "edu_level".

Data exploration: once we've got data re-organized (Tidy), let's start with representing data through data visualization tools:

Starting with "generic data" (data including "All" races, "All" genders, "All" races and genders) compare avg values versus segmented data (by race, by gender). See how education level and year affect wages: We can select a race and gender, see how the level of education plays a role in the wages level (pay) through time:

```
df12_t %>%
  filter(race == "All",gender == "All") %>%
  ggplot(aes(year, pay, color = edu_5)) +
  geom_line()+
  xlab("YEAR") +
  ylab("WAGES [$/h]") +
  ggtitle("Wages by education level (edu_5)")
```



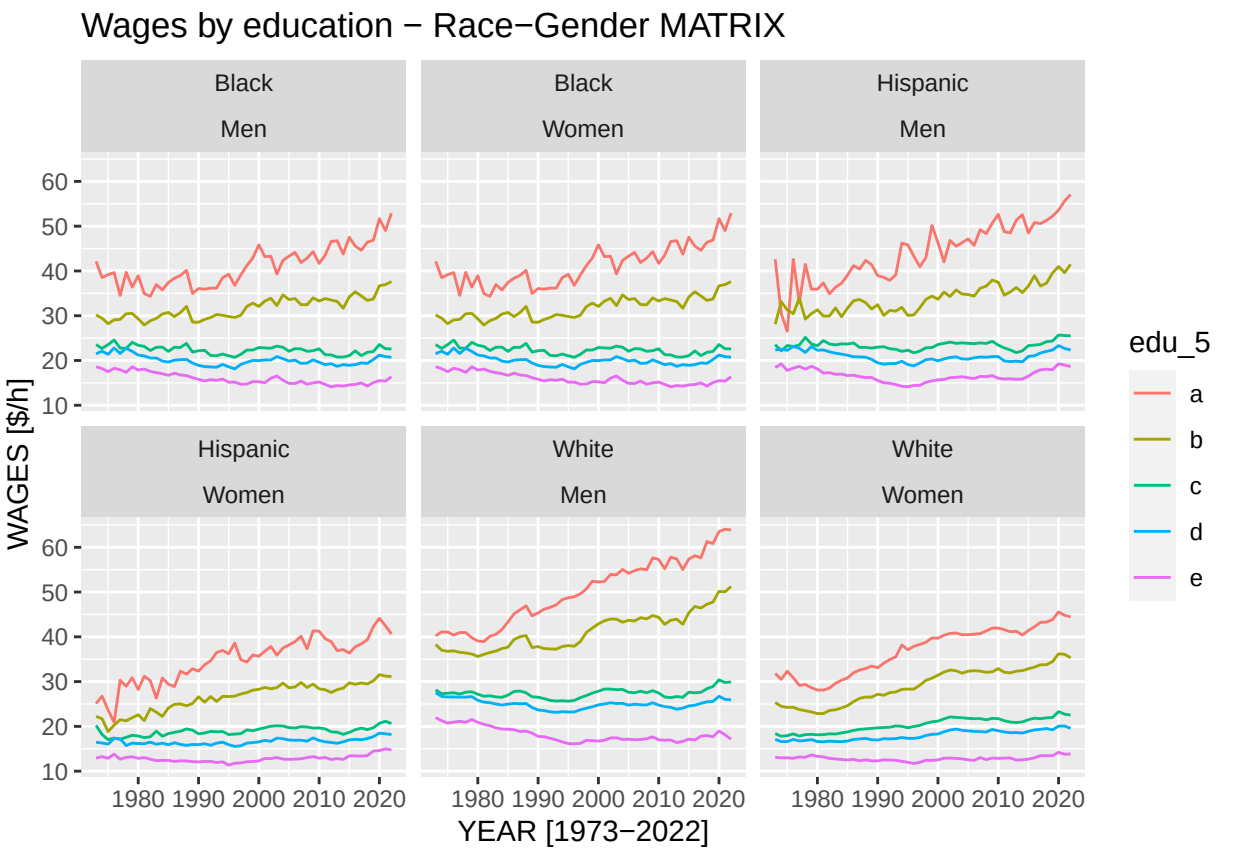
This plot provides already a visual overview of the evolution of wages, based on the 5 different education levels considered, through time. By education level, through time, we observe:

-Level “a”, advanced degree, has evolved from around 37 (dollars per hour) in 1973 to around 54 (dollars per hour) in 2022. Note we’ll use the abbreviation ‘dph’ for ‘dollars per hour’ onwards. -Level “b”, bachelor degree, has evolved from around 33dph in 1973 to around 42dph in 2022. -Level “c”, some college, advanced degree, has changed little, from around 24dph in 1973 to around 25$/h in 2022. -Level “d”, high-school, has stagnated around 22dph, from 1973 until 2022.Dropping 2 to 3dph through the mid 90s, recovering in the mid 2000s,just to drop adn recover again around 2020. -Level “e”, less than high-school, has dropped from around 18dph in 1973 to around 16dph in 2022.

Between education levels: we observe large dispersion on pay values, around 38dph wages difference between levels “e” and “a”(16-54) in 2022 (wages “a” being 2,3 times that of wages “e”). In 1973 dispersion was at 19dph (18-37) (wages “a” being 2 times that of wages “e”). An interesting finding is, wages for education levels (a,b) have grown through time, while wages in the c-d categories have stagnated (which means, purchasing power should be much lower in 2022 with that level of wages versus the one enjoyed in 1973 with the same wages). Category “e” has decreased.

We’ll now use a “facet wrap” plot to compare evolution of wages through time (for each of the 5 education levels), and for the 6 race-gender combinations: Black-Women, Black-Men; Hispanic-Women, Hispanic-Men; White-Women, White-Men: This might help us identify any significant differences among graphics:

```
df12_t %>%
  filter(race %in% c("Black", "White", "Hispanic") & gender %in% c("Women","Men")) %>%
  ggplot(aes(year, pay, color = edu_5)) +
  geom_line() +
  facet_wrap(~race ~gender)+
  xlab("YEAR [1973-2022]") +
  ylab("WAGES [$/h]") +
  ggtitle("Wages by education - Race-Gender MATRIX")
```



Observations:

1. On RACE and GENDER.

1.1.RACE-GENDER gap:

-White_Men are the best paid at equal level of education, for all 5 education levels; the best paid of any race-gender combination. Specially for higher education levels. -Hispanic_Women are the worst paid group at the highest education level. We can print a few values confirming these observations from the plot:

```
    which.max(df12_t$pay)
```

```
## [1] 452
```

```
    df12_t$pay[452]
```

```
## [1] 64.04
```

```
    df12_t$race[452]
```

```
## [1] "White"
```

```
    df12_t$gender[452]
```

```
## [1] "Men"
```

```
    #[1] 64.04 [1] "White" [1] "Men"
    which.min(df12_t$pay)
```

```
## [1] 1278
```

```
    df12_t$pay[1278]
```

```
## [1] 11.35
```

```
    df12_t$race[1278]
```

```
## [1] "Hispanic"
```

```
    df12_t$gender[1278]
```

```
## [1] "Women"
```

```
    #[1] 11.35 [1]"Hispanic" [1] "Women"
```

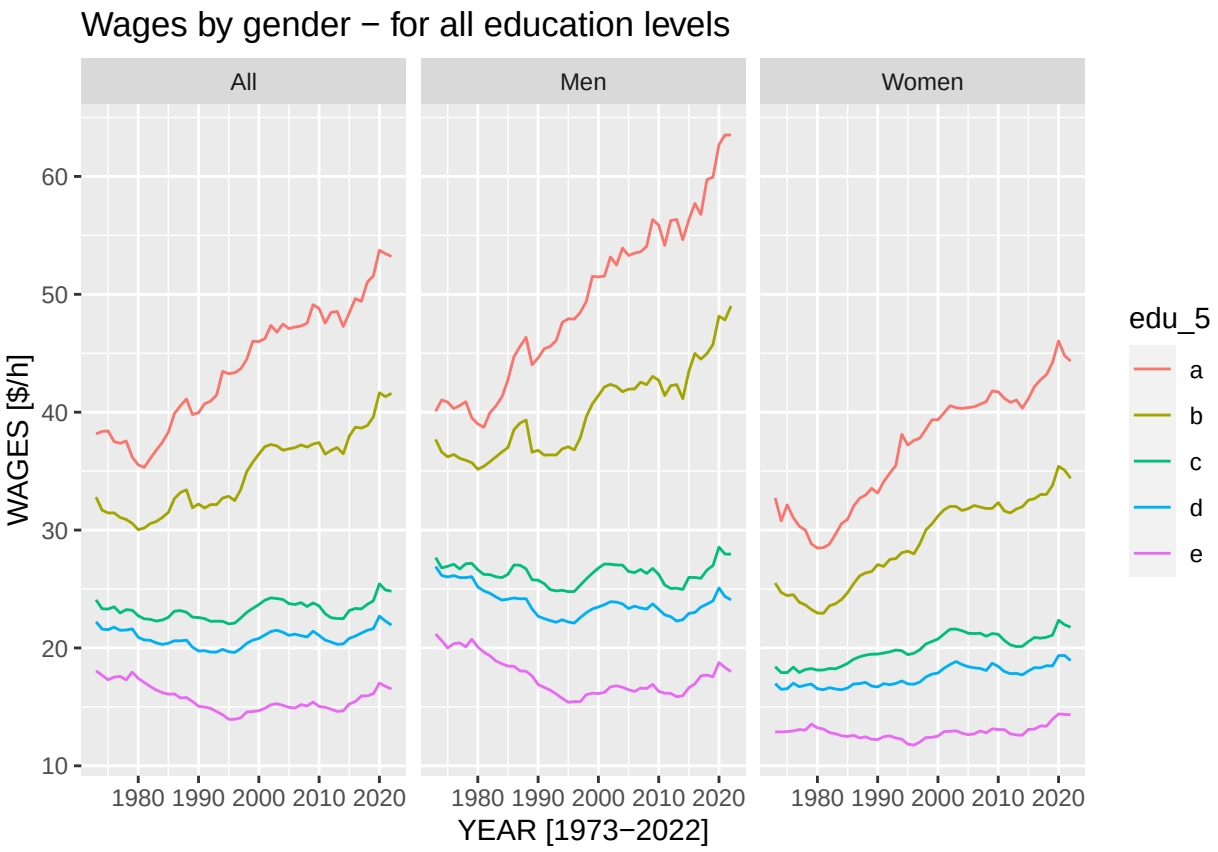
These are coherent with our perception: Best (pay) wages value of all across our data belongs to the White_Men group; Worst (pay) wages value of all across our data, belongs to the Hispanic_Women group.

1.2. GENDER gap:

-The lowest gender gap within a race seems to be within the Black group. Both gender distributions through time are similar, at similar values. -The White and Hispanic groups showing a larger gap between gender. -White group, gap between Women and Mean is around a 40% plus pay for Men, for all education levels. In absolute values,the more education level, the larger the absolute value differential. -Hispanic group: Men receive between 30 to 40% more pay than Women, for the same education level.

We can produce a plot considering only GENDER (race within), for a more generic view of the effect of gender on pay: Overview of pay evolution through time for the two considered genders(plus All):

```
df12_t %>%
  filter(race =="All" & edu_5 %in% c("a","b","c","d","e")) %>%
  ggplot(aes(year, pay, color = edu_5)) +
  geom_line() +
  facet_wrap(~gender)+
  xlab("YEAR [1973-2022]") +
  ylab("WAGES [$/h]") +
  ggtitle("Wages by gender - for all education levels")
```



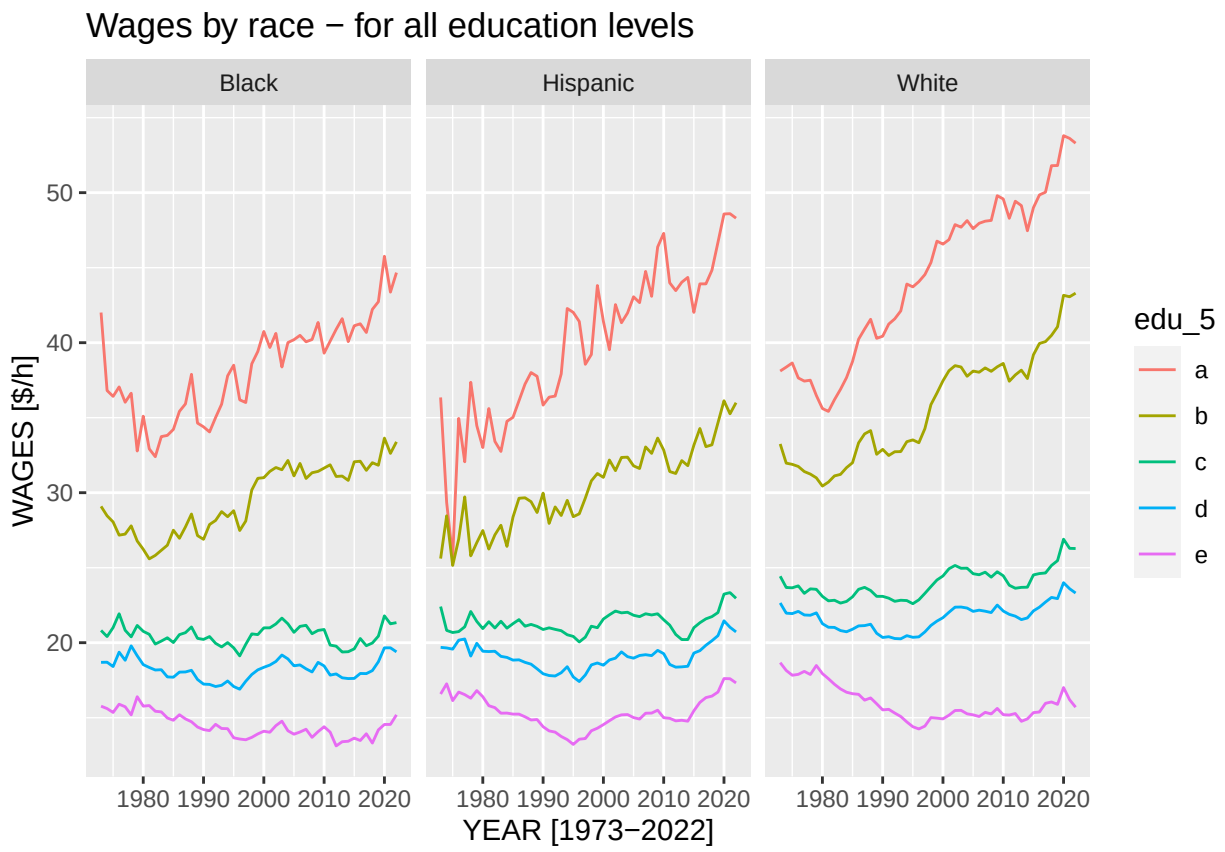
Overall (not splitting gender by race), Men wages are well above those of Women. All the way from 1973 until 2022. Furthermore, the gap (i.e."a" level, 1973: 33-40 to 2022: 44-63) seems to have increased (in absolute values) through time.

1.3. RACE gap:

-White_Men are the best paid group, while White_Women receive a worse wages than Hispanic_Men and Black_Men and Women, at high education levels. -Hispanic Men are better paid than Black Men and Women, however Hispanic Women receive lower wages than Black Mean and Women.

It appears that race and gender are key-factors, which combined result in the observed effects. In order to view a more generic picture of the race gap, we could represent a Race based plot, which combines genders within each race group:

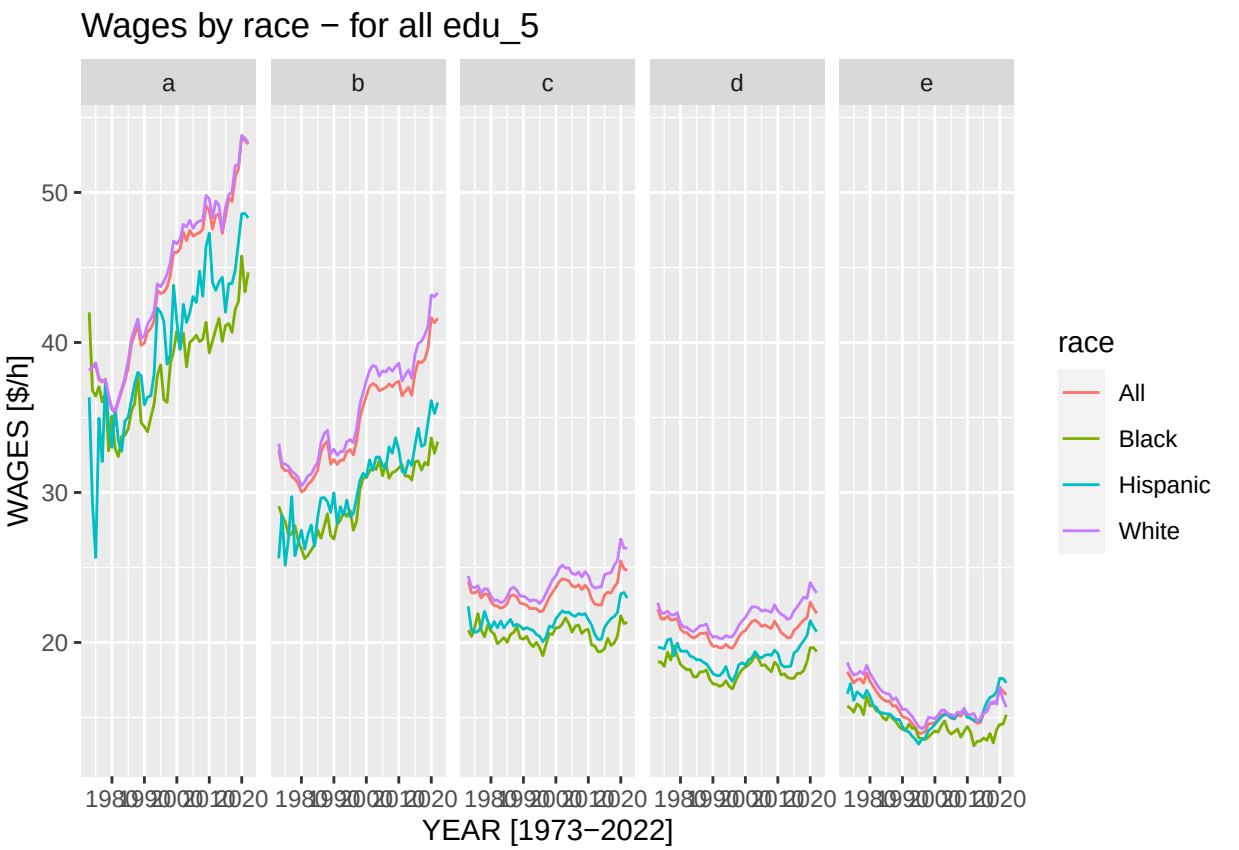
```
df12_t %>% filter(race %in% c("Black", "White", "Hispanic") & gender == "All") %>%
  ggplot(aes(year, pay, color = edu_5)) +
  geom_line() +
  facet_wrap(~race)+
  xlab("YEAR [1973-2022]") +
  ylab("WAGES [$/h]") +
  ggtitle("Wages by race - for all education levels")
```



This "Wages by race - for all education levels" helps compare wages among races. When not considering gender(race values combining here both genders) separately, then we can see that: -The best wages for education levels a,b,c,d have been through time for the White group. -Hispanic where paid less than Black and White in 1973, however Hispanic are paid more than Black for the top education levels a,b in 2022, however still worse than the White group. -The White group at the lowest education level "e", where paid more than Hispanic, and Hispanic in turn more than Black group members, in 1973. However in 2022 the White group is the second best paid at edu_5 level "e", behind Hispanic, remaining the Black group as the worst paid at level "e".

Similar to the former representations, but with focus on each EDUCATION level across RACE, following is a plot "Wages by race - for all edu_5":

```
df12_t %>% filter(edu_5 %in% c("a","b","c","d","e") & gender=="All") %>%
  ggplot(aes(year, pay, color=race)) +
  geom_line()+
  facet_grid(~ edu_5)+
  xlab("YEAR [1973-2022]") +
  ylab("WAGES [$/h]")+
  ggtitle("Wages by race - for all edu_5")
```



This approach reveals that, after the mid 80s:

-The White-group (genders mixed) perceive the highest wages for all education levels b,c,d through time.For level "a" the same it true after the mid 80s; for level "e", the White-Group (genders mixed) is in 2nd place,being the best paid the Hispanic-group, the worst the Back-group. -The Hispanic-group perceive the second highest wages (except a short periods i.e. in the early 2000s); -The Black-group perceive the lowest wages.

2. On TIME(year) and EDUCATION(edu_5).

From the plots we've viewed so far,

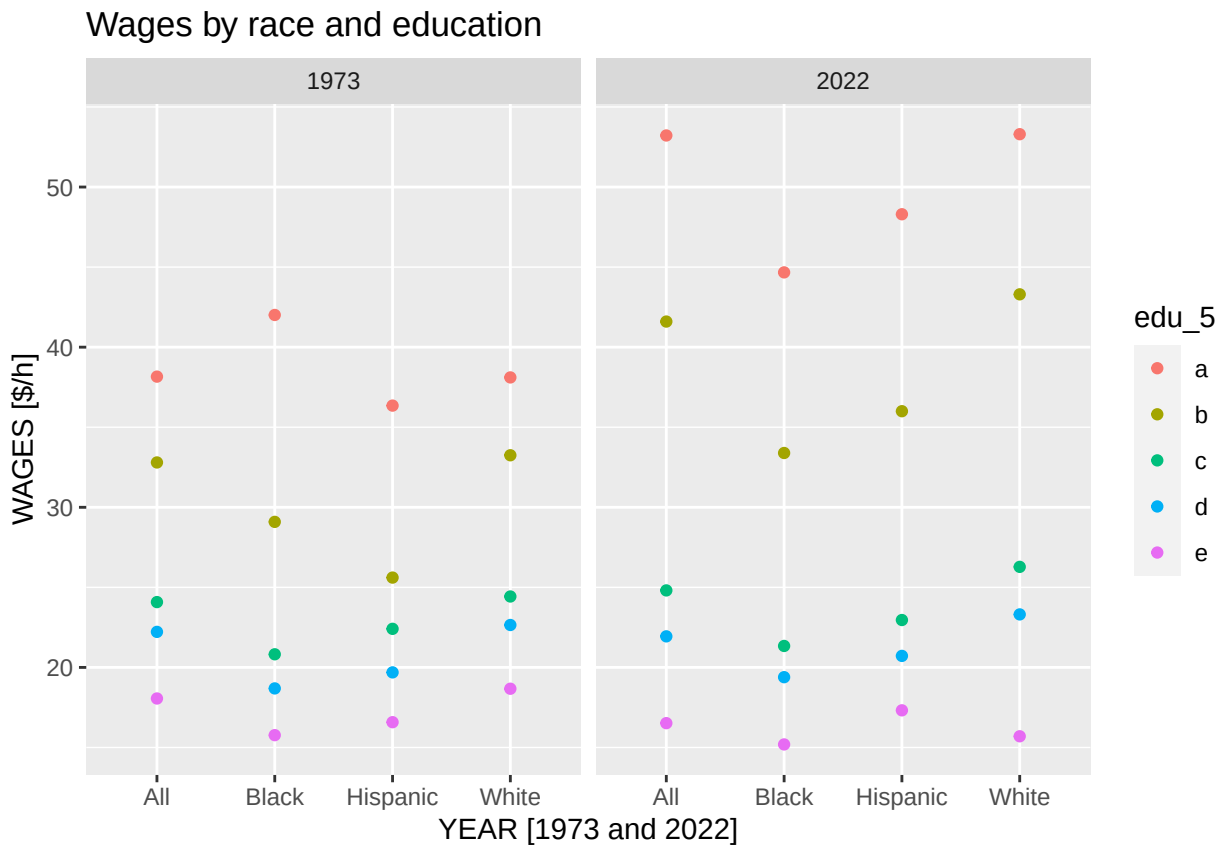
2.1. Time effect: consistently, -(a,b) GROWING wages: for "a" and "b" education levels, although oscillations (positive, negative slopes) are observed, a general positive trend is observed through time on wages. -(c,d) STAGNANT wages: For levels "c" and "d", wages are overall stagnant through time; -(e) DECREASING wages: For level "e", wages have decreased over time.

2.2. Education level effect: we've covered this variable before. As a generic consistent finding, for all other variable effect combinations, education remains a key variable, its effect on wages coherent across all visual representations of data. Conclusion: across time, the higher the education level, the higher the wages.

Finally, we can display the starting (year 1973) and end (year 2022) points of the time scale being considered, observe how pay-race fare (comparison between starting and end points):

Faceting:

```
filter(df12_t, year%in%c(1973, 2022) & gender=="All") %>%
  ggplot(aes(race, pay, col = edu_5)) +
  geom_point() +
  facet_grid(. ~ year)+
  xlab("YEAR [1973 and 2022]") +
  ylab("WAGES [$/h]")+
  ggtitle("Wages by race and education")
```

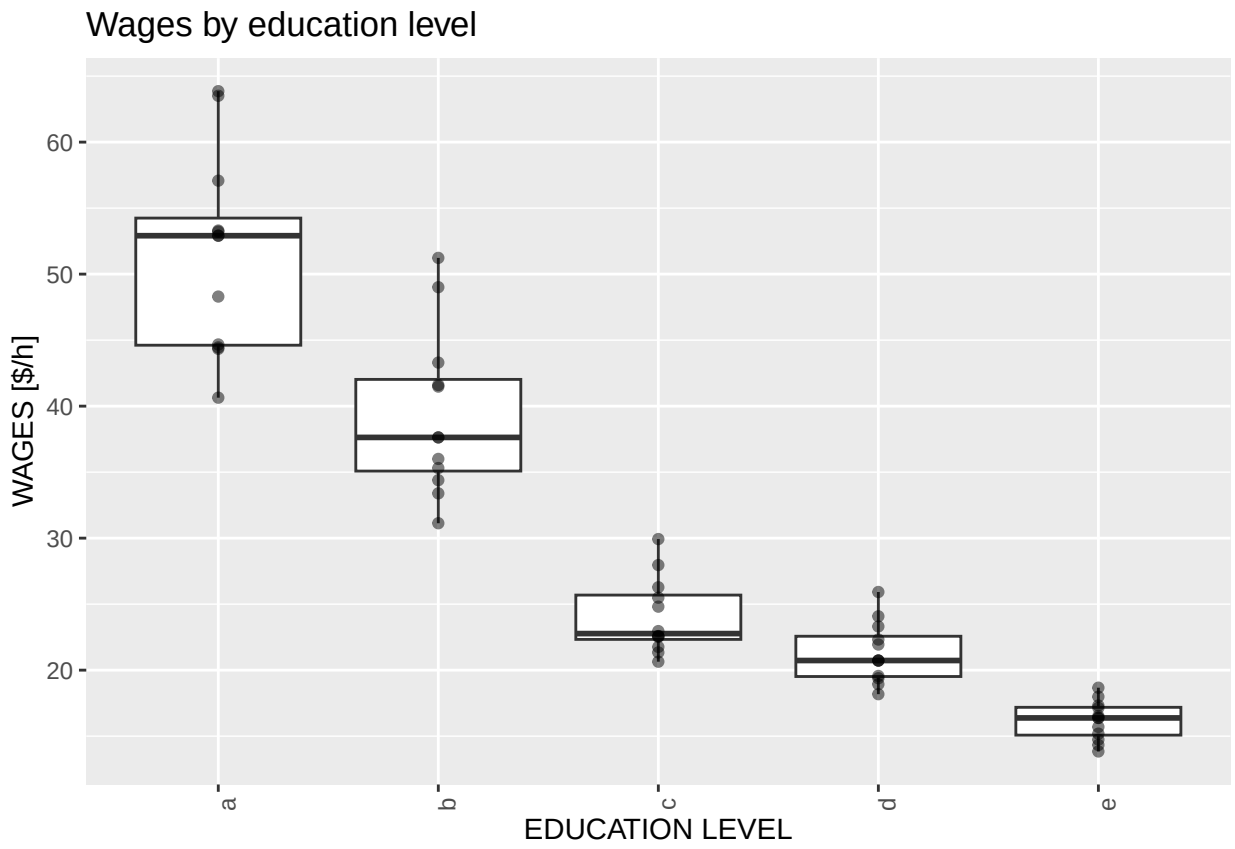


Overview of starting and ending points, for each education level for each race. We can appreciate coherence with former representations. It stands out that for level "e" the White-group comes in seond place, consistent with former findings.

Further visualization of the education level effects. We can add a boxplot for a visual representation of wages versus education, where time, race, gender explain the position of the box-plot defining values (mean, quartiles, range, plus outliers).

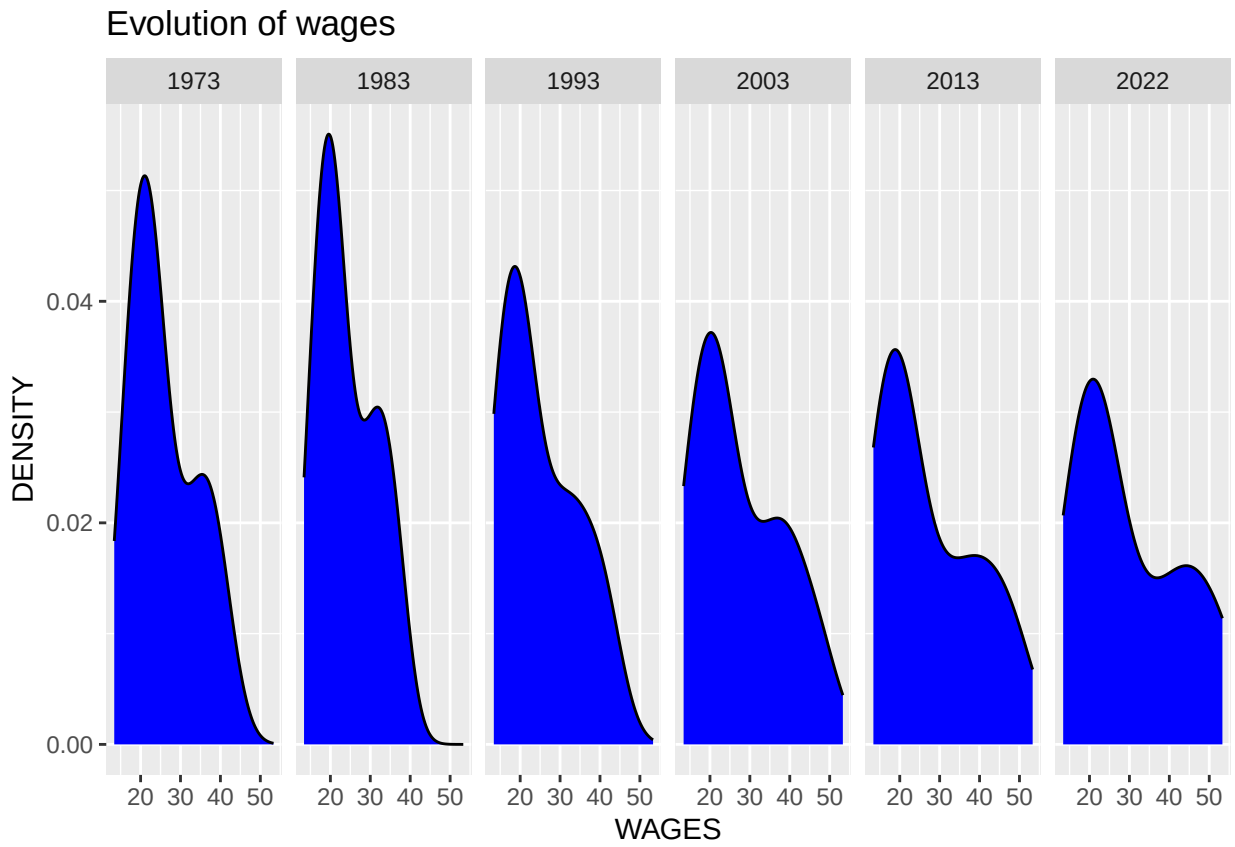
```
b_p <- df12_t %>%
filter(year == 2022 & !is.na(pay)) %>%
ggplot(aes(edu_5, pay)) +
geom_boxplot() +
theme(axis.text.x = element_text(angle = 90, hjust = 1))+
geom_point(alpha = 0.5)+
```

```
    xlab("EDUCATION LEVEL") +
    ylab("WAGES [$/h]")+
    ggtitle("Wages by education level")
    b_p
```



In addition to that, we can represent the generic effect of time on wages. Facet_grid density plots can help us with visualizing trends. We'll select four points in time (years 1983,1993,2003,2013) quite evenly distributed between the starting point in our time scale (1973) and the ending point (2022):

```
filter(df12_t, year %in% c(1973,1983, 1993,2003,2013, 2022) & gender=="All") %>%
    ggplot(aes(pay)) +
    geom_density(fill="blue")+
    facet_grid(. ~ year)+
    xlab("WAGES") +
    ylab("DENSITY")+
    ggtitle("Evolution of wages")
```



This graphic shows a growth of wages density towards higher ones through time, consistent with an increasing gap between education wages levels, growth of wages for "a" and "b" education levels through time, and stagnation for "c" and "d", along together with a wage decrease for the "e" level of education.

We can summarize our findings after data visualization work. GENERAL CONCLUSIONS after data wrangling and visualization:

1.level of education has the highest effect on pay;

2.time has an effect on pay, which varies according to edu_5: -2.1."a","b" levels growing; -2.2."c","d" stagnating; -2.3."e" levels decreasing.

2.Race consistently (through time) affects the level of pay for the same education level,being -in general- White the highest paid, Hispanic following, and Black last;

3.Gender has been and continues to be a factor affecting pay level. Men are paid more than Women across education level and race; 4.Through time wages are increasing for higher levels of education (a, b), and showing stagnation for the lower levels of education (c,d),and decreasing for the lowest (e), across races and genders.

Part 4. Initial solution.

Our working hypothesis will be based on the work we've done so far with Data Visualization.

We'll fit models which consider education, gender and race effects, separately, then a global model which considers effects from education, gender and race. We'll start considering race, gender, the race+gender, then a stronger variable, edu_5, and finally all of them combined, including year too, as following:

fit1 [pay ~ race] fit2 [pay ~ gender] fit3 [pay ~ gender + race] fit4 [pay ~ edu_5] fit5 [pay ~ edu_5 + gender + race] fit6 [pay ~ edu_5 + year + gender + race]

Least squares:

We will try first the race effect on wages (pay ~ race). Race effect _ pay - fitting a linear model:

```
    fit1 <- lm(pay ~ as.factor(race), data = df12_t)
    fit1
```

```
##
## Call:
## lm(formula = pay ~ as.factor(race), data = df12_t)
##
## Coefficients:
##             (Intercept)     as.factor(race)Black  as.factor(race)Hispanic
##                 27.4589                  -1.8798                  -2.4815
##    as.factor(race)White
##                  0.6016
```

```
    summary(fit1)
```

```
##
## Call:
## lm(formula = pay ~ as.factor(race), data = df12_t)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.331  -7.822  -3.429   7.846  36.061
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               27.4589     0.3811  72.046  < 2e-16 ***
## as.factor(race)Black      -1.8798     0.5390  -3.488 0.000495 ***
## as.factor(race)Hispanic   -2.4815     0.5390  -4.604 4.32e-06 ***
## as.factor(race)White       0.6016     0.5390   1.116 0.264472
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.44 on 2996 degrees of freedom
## Multiple R-squared:  0.01476,    Adjusted R-squared:  0.01377
## F-statistic: 14.96 on 3 and 2996 DF,  p-value: 1.15e-09
```

Fitting a linear model with pay as dependent and race as independent variables, data from df12_t: -we obtain an intercept of 27.4589 and slopes for Black (-1.8798), Hispanic(-2.4815) and White (0.6016). -According to this, we observe a positive effect on pay for the White group, and negative effects for then Black and Hispanic groups. -p-values are very low for Black and Hispanic (good predictors), but over 0.05 for White. -Residuals are large. -Multiple R squared is far from 1, the model can explain some 1,4% of the variability.

Then we'll study the gender effect on pay (gender ~ race):

```
    fit2 <- lm(pay ~ as.factor(gender), data=df12_t)
    summary(fit2)
```

```
##
## Call:
```

```
## lm(formula = pay ~ as.factor(gender), data = df12_t)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.349  -7.590  -3.539   7.470  34.571
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              26.3648     0.3241  81.355  < 2e-16 ***
## as.factor(gender)Men      3.1038     0.4583   6.772 1.52e-11 ***
## as.factor(gender)Women   -2.6412     0.4583  -5.763 9.10e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.25 on 2997 degrees of freedom
## Multiple R-squared:  0.04992,    Adjusted R-squared:  0.04929
## F-statistic: 78.74 on 2 and 2997 DF,  p-value: < 2.2e-16
```

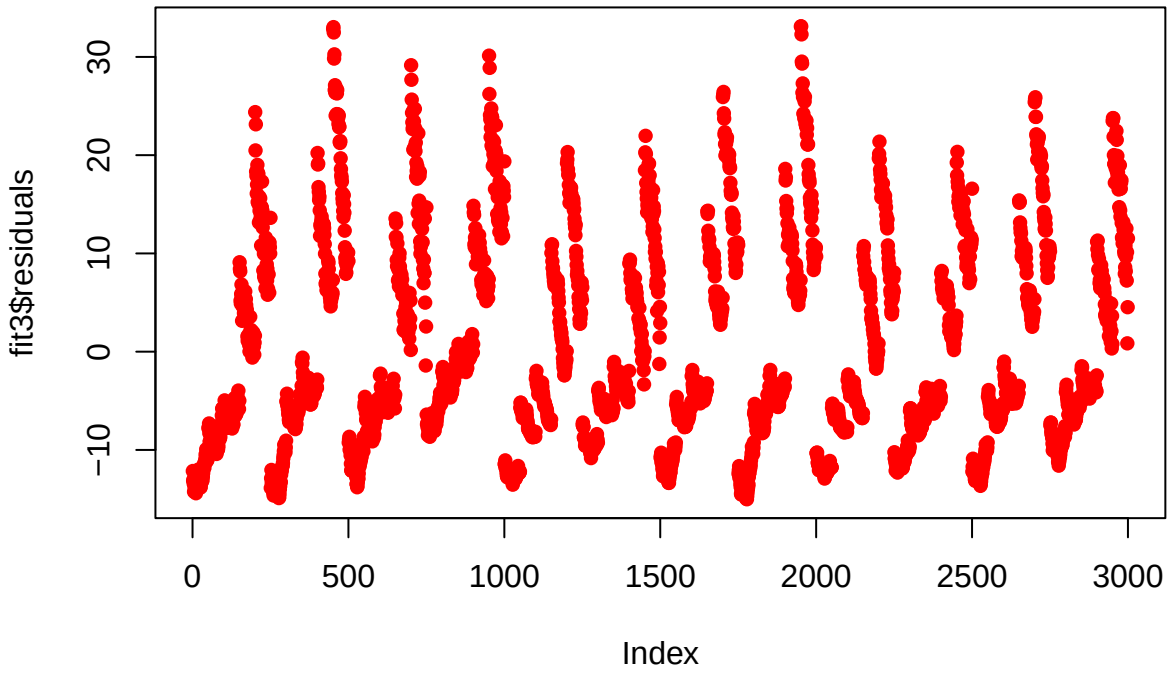
Fitting a linear model with pay as dependent, gender as independent variables, data from df12_t: -we obtain an intercept of 26.365 and slopes for Men (3.104) and Women (-2.641). -gender=Men adds to the pay value while gender=Women reduces pay. -p-values are very low for gender, both for Men and Women, thus the predictors are good. -However residuals are very high. -Multiple R squared is far from 1, the model can explain some 5% of the variability.

Now we'll combine gender and race, see how its combined effects improve the model:

```
fit3 <- lm(pay ~ as.factor(gender) + as.factor(race), data=df12_t)
summary(fit3)
```

```
##
## Call:
## lm(formula = pay ~ as.factor(gender) + as.factor(race), data = df12_t)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.019  -7.344  -3.813   7.402  33.111
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               27.3047     0.4550  60.016  < 2e-16 ***
## as.factor(gender)Men       3.1038     0.4550   6.822 1.08e-11 ***
## as.factor(gender)Women    -2.6412     0.4550  -5.805 7.09e-09 ***
## as.factor(race)Black      -1.8798     0.5253  -3.578 0.000351 ***
## as.factor(race)Hispanic   -2.4815     0.5253  -4.724 2.42e-06 ***
## as.factor(race)White       0.6016     0.5253   1.145 0.252256
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.17 on 2994 degrees of freedom
## Multiple R-squared:  0.06468,    Adjusted R-squared:  0.06312
## F-statistic: 41.41 on 5 and 2994 DF,  p-value: < 2.2e-16
```

```
plot(fit3$residuals, pch = 16, col = "red")
```



Fitting a linear model with pay as dependent and gender and race as independent variables, data from df12_t: -we obtain an intercept of 26.365 and slopes for Men (3.104) and Women (-2.641); -slopes for Black (-1.8798), Hispanic (-2.4815), and White (0.6016); -gender=Men and race=White adds to the pay (wages) while gender=Women and race=Hispanic or Black reduces pay; -p-values are very low for Men, Women, Black, Hispanic, making these good predictors, however White shows a 0.252 value (>0.05). -Residuals remain large. -Multiple R squared is still far from 1. The model can explain some 6,3% of the variability.

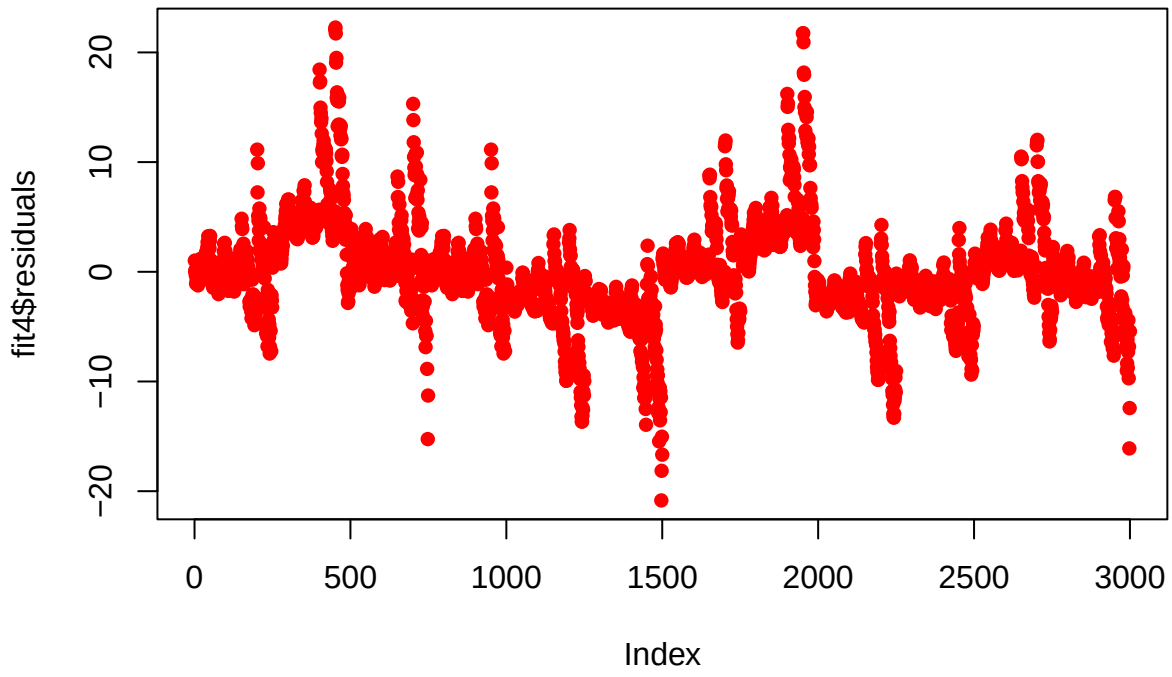
Following we study the effect of edu_5 on pay. Considering education level seems to be have the strongest effects on pay (hypothesis after perception from data visualization):

```
fit4 <- lm(pay ~ edu_5, data=df12_t)
summary(fit4)
```

```
##
## Call:
## lm(formula = pay ~ edu_5, data = df12_t)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.8410  -2.4267  -0.3572   1.9541  22.2690
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  41.7710     0.1818  229.71   <2e-16 ***
```

```
## edu_5b       -8.9748     0.2572  -34.90   <2e-16 ***
## edu_5c      -19.2598     0.2572  -74.89   <2e-16 ***
## edu_5d      -21.6126     0.2572  -84.04   <2e-16 ***
## edu_5e      -26.4129     0.2572 -102.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.454 on 2995 degrees of freedom
## Multiple R-squared:  0.8206, Adjusted R-squared:  0.8204
## F-statistic:  3426 on 4 and 2995 DF,  p-value: < 2.2e-16
```

```
    plot(fit4$residuals, pch = 16, col = "red")
```



Fitting a linear model with pay as dependent and edu_5 as independent variables, data from df12_t: -we obtain an intercept of 41.7710 and negative slopes for edu levels b,c,d and e. -p-values are low across the education levels, thus the predictor(s) are good. -However residuals keep being large. -Multiple R squared is now close to 1, the model can explain some 82% of the variability. -This is a KEY FINDING, this confirms the perception that EDUCATION LEVEL has the MOST EFFECT on PAY(wages).

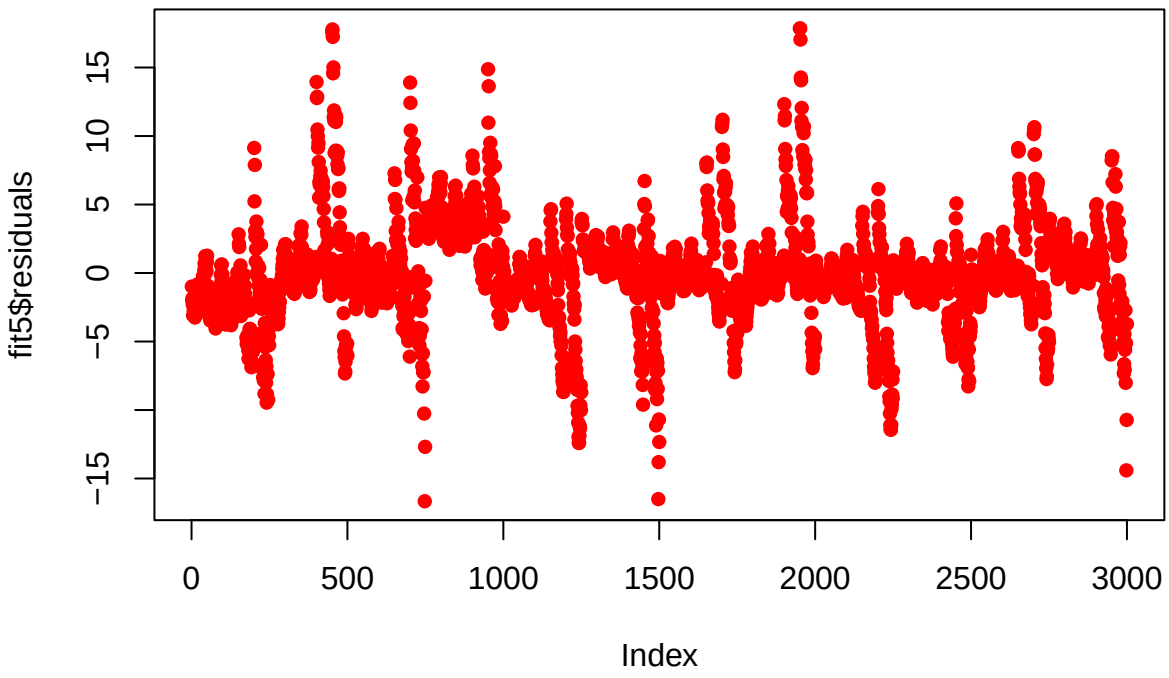
We'll now fit a model with "edu_5", "gender", and "race" as independent variables (pay ~ edu_5 + gender + race):

```
    fit5 <- lm(pay ~ as.factor(edu_5) + as.factor(gender) + as.factor(race), data=df12_t)
    summary(fit5)
```

```
##
```

```
## Call:
## lm(formula = pay ~ as.factor(edu_5) + as.factor(gender) + as.factor(race),
##     data = df12_t)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.6591  -1.6084  -0.1449   1.3394  17.8594
##
## Coefficients:
##                          Estimate Std. Error  t value Pr(>|t|)
## (Intercept)               42.5568     0.2058  206.783   <2e-16 ***
## as.factor(edu_5)b         -8.9748     0.2058  -43.608   <2e-16 ***
## as.factor(edu_5)c        -19.2598     0.2058  -93.583   <2e-16 ***
## as.factor(edu_5)d        -21.6126     0.2058 -105.016   <2e-16 ***
## as.factor(edu_5)e        -26.4129     0.2058 -128.340   <2e-16 ***
## as.factor(gender)Men       3.1038     0.1594   19.470   <2e-16 ***
## as.factor(gender)Women    -2.6412     0.1594  -16.568   <2e-16 ***
## as.factor(race)Black      -1.8798     0.1841  -10.212   <2e-16 ***
## as.factor(race)Hispanic   -2.4815     0.1841  -13.481   <2e-16 ***
## as.factor(race)White       0.6016     0.1841    3.268   0.0011 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.565 on 2990 degrees of freedom
## Multiple R-squared:  0.8853, Adjusted R-squared:  0.885
## F-statistic:  2565 on 9 and 2990 DF,  p-value: < 2.2e-16
```

```
    plot(fit5$residuals, pch = 16, col = "red")
```



Fitting a linear model with pay as dependent and edu_5, gender, race as independent variables, data from df12_t: -we obtain an intercept of 42.5568 and positive slopes for Men and White, and negative slopes for edu levels b,c,d and e, Women, Black, and Hispanic. -p-values are low across the education levels, thus the predictor(s) are good. -Residuals have decreased (-17 to +18). -Multiple R squared is now close to 1, the model can explain some 88.5% of the variability.

This is the best model version yet.

As a last potential improvement, let's add to it YEAR as independent variable:

Finally, we'll fit a model with "edu_5", "year", "gender", and "race" as independent variables (pay ~ edu_5 + year + gender + race):

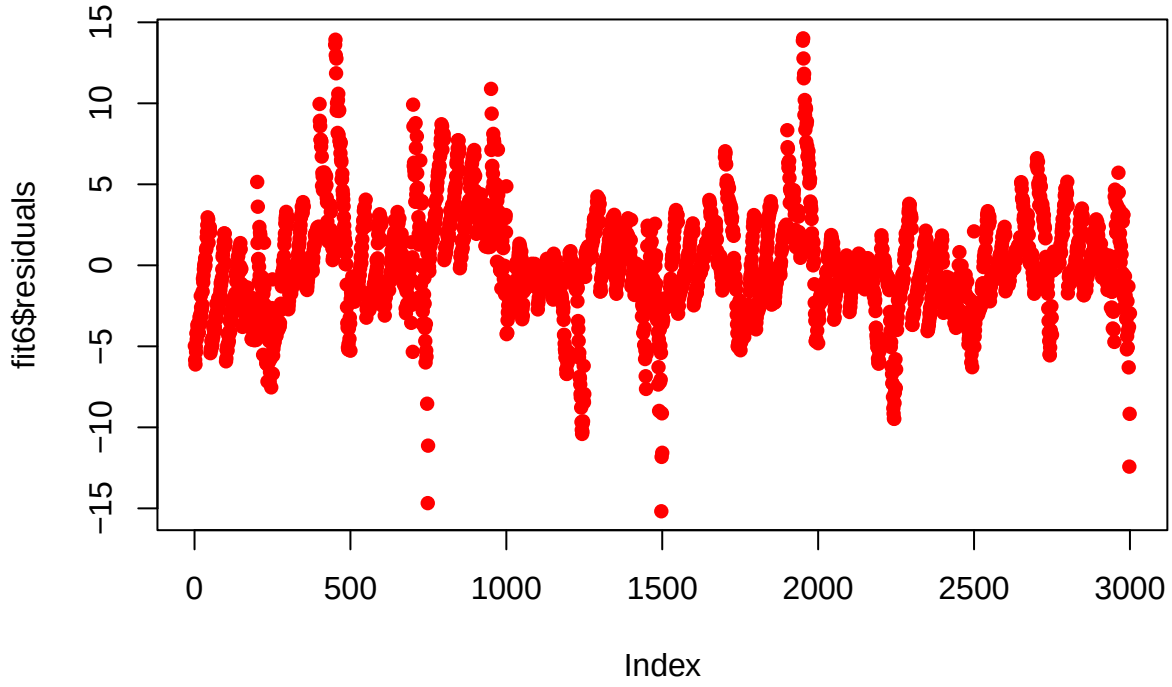
```
    fit6 <- lm(pay ~ as.factor(edu_5) + as.factor(year)+ as.factor(gender) + as.factor(race), data=df12_
    summary(fit6)
```

```
##
## Call:
## lm(formula = pay ~ as.factor(edu_5) + as.factor(year) + as.factor(gender) +
##     as.factor(race), data = df12_t)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.1794  -1.7902  -0.2027   1.7869  14.0083
##
## Coefficients:
##                          Estimate Std. Error  t value Pr(>|t|)
```

```
## (Intercept)             41.79359    0.43883   95.239  < 2e-16 ***
## as.factor(edu_5)b       -8.97478    0.18066  -49.677  < 2e-16 ***
## as.factor(edu_5)c      -19.25977    0.18066 -106.606  < 2e-16 ***
## as.factor(edu_5)d      -21.61263    0.18066 -119.630  < 2e-16 ***
## as.factor(edu_5)e      -26.41293    0.18066 -146.200  < 2e-16 ***
## as.factor(year)1974     -0.79700    0.57131   -1.395 0.163106
## as.factor(year)1975     -1.21750    0.57131   -2.131 0.033165 *
## as.factor(year)1976     -0.56150    0.57131   -0.983 0.325768
## as.factor(year)1977     -0.95433    0.57131   -1.670 0.094938 .
## as.factor(year)1978     -0.65100    0.57131   -1.139 0.254589
## as.factor(year)1979     -0.84500    0.57131   -1.479 0.139229
## as.factor(year)1980     -1.22767    0.57131   -2.149 0.031725 *
## as.factor(year)1981     -1.48783    0.57131   -2.604 0.009253 **
## as.factor(year)1982     -1.52133    0.57131   -2.663 0.007789 **
## as.factor(year)1983     -1.37467    0.57131   -2.406 0.016181 *
## as.factor(year)1984     -1.28883    0.57131   -2.256 0.024147 *
## as.factor(year)1985     -0.98117    0.57131   -1.717 0.086009 .
## as.factor(year)1986     -0.41367    0.57131   -0.724 0.469078
## as.factor(year)1987     -0.20933    0.57131   -0.366 0.714084
## as.factor(year)1988      0.04733    0.57131    0.083 0.933975
## as.factor(year)1989     -0.87617    0.57131   -1.534 0.125230
## as.factor(year)1990     -0.97117    0.57131   -1.700 0.089254 .
## as.factor(year)1991     -0.98733    0.57131   -1.728 0.084056 .
## as.factor(year)1992     -0.90283    0.57131   -1.580 0.114146
## as.factor(year)1993     -0.79500    0.57131   -1.392 0.164164
## as.factor(year)1994     -0.15383    0.57131   -0.269 0.787743
## as.factor(year)1995     -0.41350    0.57131   -0.724 0.469258
## as.factor(year)1996     -0.60567    0.57131   -1.060 0.289165
## as.factor(year)1997     -0.27633    0.57131   -0.484 0.628643
## as.factor(year)1998      0.57367    0.57131    1.004 0.315397
## as.factor(year)1999      1.27633    0.57131    2.234 0.025554 *
## as.factor(year)2000      1.42633    0.57131    2.497 0.012593 *
## as.factor(year)2001      1.57933    0.57131    2.764 0.005738 **
## as.factor(year)2002      2.02717    0.57131    3.548 0.000394 ***
## as.factor(year)2003      1.83583    0.57131    3.213 0.001326 **
## as.factor(year)2004      1.98050    0.57131    3.467 0.000535 ***
## as.factor(year)2005      1.77267    0.57131    3.103 0.001935 **
## as.factor(year)2006      1.85833    0.57131    3.253 0.001156 **
## as.factor(year)2007      1.99917    0.57131    3.499 0.000473 ***
## as.factor(year)2008      1.81417    0.57131    3.175 0.001511 **
## as.factor(year)2009      2.56200    0.57131    4.484 7.59e-06 ***
## as.factor(year)2010      2.26367    0.57131    3.962 7.60e-05 ***
## as.factor(year)2011      1.63250    0.57131    2.857 0.004300 **
## as.factor(year)2012      1.71250    0.57131    2.998 0.002745 **
## as.factor(year)2013      1.70367    0.57131    2.982 0.002887 **
## as.factor(year)2014      1.35917    0.57131    2.379 0.017421 *
## as.factor(year)2015      2.15383    0.57131    3.770 0.000166 ***
## as.factor(year)2016      2.60850    0.57131    4.566 5.18e-06 ***
## as.factor(year)2017      2.59267    0.57131    4.538 5.90e-06 ***
## as.factor(year)2018      3.00883    0.57131    5.267 1.49e-07 ***
## as.factor(year)2019      3.48767    0.57131    6.105 1.16e-09 ***
## as.factor(year)2020      5.03467    0.57131    8.813  < 2e-16 ***
## as.factor(year)2021      4.61433    0.57131    8.077 9.60e-16 ***
## as.factor(year)2022      4.74667    0.57131    8.308  < 2e-16 ***
```

```
## as.factor(gender)Men      3.10380    0.13994   22.179  < 2e-16 ***
## as.factor(gender)Women   -2.64123    0.13994  -18.874  < 2e-16 ***
## as.factor(race)Black     -1.87977    0.16159  -11.633  < 2e-16 ***
## as.factor(race)Hispanic  -2.48149    0.16159  -15.357  < 2e-16 ***
## as.factor(race)White      0.60157    0.16159    3.723 0.000201 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.129 on 2941 degrees of freedom
## Multiple R-squared:  0.9131, Adjusted R-squared:  0.9114
## F-statistic: 532.6 on 58 and 2941 DF,  p-value: < 2.2e-16
```

```
    plot(fit6$residuals, pch = 16, col = "red")
```



Fitting a linear model with pay as dependent and edu_5, year, gender, race as independent variables, data from df12_t: -we obtain an intercept of 41.79359 and positive slopes for Men and White, and years from 1998 to 2022; -and negative slopes for edu levels b,c,d and e, Women, Black, and Hispanic, and years 1973-1997; -p-values are low across all variables-values, expect for years 1973-1997. -Residuals have decreased (-15 to +14). -Multiple R squared is now close to 1, the model can explain some 91.1% of the variability.

Part 5. Solution.

Building a "Recommendation System" tool - Machine Learning Model -. Now we'll be interested on projecting expected wages (pay) through the period, 1973-2022. We can start with defining the overall pay average, through time, and across all relevant variables, education level, race, and gender.

We'll work with a "recommendation system" tool. We will create a simple model -algorithm- which we will improve through a sequence of steps (successive models). Using a training set (train_set) and a test set (test_set) to assess the accuracy of the models. The RMSE function (residual mean square error) will help us evaluate model accuracy. By calculating the error made between prediction of wages, and true wages.

Creating train and test sets:

```
    library(caret)
    set.seed(755)
    test_index <- createDataPartition(y = df12_t$pay, times = 1, p = 0.2,
                                      list = FALSE)
    #test_index
    train_set <- df12_t[-test_index,]
    test_set <- df12_t[test_index,]
    head(train_set)
```

```
##   year gender  race edu_5   pay
## 2 2021    Men Black     e 15.38
## 3 2020    Men Black     e 15.52
## 4 2019    Men Black     e 15.07
## 6 2017    Men Black     e 15.00
## 7 2016    Men Black     e 14.64
## 8 2015    Men Black     e 14.55
```

```
    head(test_set)
```

```
##    year gender  race edu_5   pay
## 1  2022    Men Black     e 16.38
## 5  2018    Men Black     e 14.27
## 10 2013    Men Black     e 14.39
## 11 2012    Men Black     e 14.12
## 13 2010    Men Black     e 15.19
## 16 2007    Men Black     e 15.41
```

```
    #To make sure we don't include RACE and GENDER in the test set that do not appear in the
    #training set, we remove these entries using the semi_join function:
    test_set <- test_set %>%
      semi_join(train_set, by = "race") %>%
      semi_join(train_set, by = "gender")
    head(test_set)
```

```
##   year gender  race edu_5   pay
## 1 2022    Men Black     e 16.38
## 2 2018    Men Black     e 14.27
## 3 2013    Men Black     e 14.39
## 4 2012    Men Black     e 14.12
## 5 2010    Men Black     e 15.19
## 6 2007    Men Black     e 15.41
```

```
    #The RMSE function
    RMSE <- function(true_pay, predicted_pay){
      sqrt(mean((true_pay - predicted_pay)^2))
```

```
    }
    #A FIRST MODEL:
    mu_hat <- mean(df12_t$pay)
    mu_hat
```

```
## [1] 26.51901
```

```
    #> #[1]  26.51901

    naive_rmse <- RMSE(test_set$pay, mu_hat)
    naive_rmse
```

```
## [1] 10.70215
```

```
    #> #[1] 10.70215
```

NOTE - CLAUSE: Regarding significance of the obtained values, considering the nature of our data: -The obtained mu_hat is the average wages(pay), through the considered period, and all variables. -We've got sets of data referring to ALL genders and ALL races. -and also sets of data, specific for a matrix of gender-race values.

Let's see if these two (generic-ALL ; matrix) types of data, are to be considered in this study, or a filter is to be applied. And study if values for ALL genders-ALL races, are somehow equivalent to the combination of the 3 races and 2 genders being considered:

```
    #df12_t.All <- df12_t%>% filter( race=="All" & gender=="All")
    #str(df12_t.All)
    #mu_hat.All <- mean(df12_t.All$pay)
    #mu_hat.All
    #[1] 27.6932

    #we've got 6 groups of 250 observations, each based on generic
    #information (including more than variable variation within, for example
    #all White observations, which include both genders within)
    #These are "All"; "All MEN"; "All WOMEN";"All White"; "All Black";
    #"All Hispanic".

    #if we only include observations with independent variables defining
    #the observation value (matrix of genders x races), we might get more
    #accurate variable -effectapproximations.

    #df12_t.matrix <- df12_t%>% filter( race %in%
          #c("Black","White","Hispanic") & gender %in% c("Women","Men"))
    #str(df12_t.matrix)
    #mu_hat.matrix <- mean(df12_t.matrix$pay)
    #mu_hat.matrix
    #[1] 26.34754

    #naive_rmse
    #[1] 10.34644 for df12_t.matrix, a little lower than the former
    #10.706  (for df12_t)
```

```
    #We might want to select df12_t.matrix from here. However the observed
    #difference between using one (all data)
    #or the other ("matrix" data only) approach is small (delta-mu = 0.17;
    #delta-rmse= 0.35 )
```

Conclusion on note: We will consider all available data for our study. END NOTE - END CLAUSE.

MODEL 1.

Our first model will return an average expected pay for any combination of education level, race, gender and year:

```
    mu_hat <- mean(df12_t$pay)
    mu_hat
```

```
## [1] 26.51901
```

```
    naive_rmse <- RMSE(test_set$pay, mu_hat)
    naive_rmse
```

```
## [1] 10.70215
```

```
    rmse_results <- data_frame(method = "Just the average", RMSE = naive_rmse)
```

```
## Warning: `data_frame()` was deprecated in tibble 1.1.0.
## i Please use `tibble()` instead.
```

```
    rmse_results
```

```
## # A tibble: 1 x 2
##   method            RMSE
##   <chr>            <dbl>
## 1 Just the average  10.7
```

MODEL 2.

Modeling EDUCATION (edu_5) effects

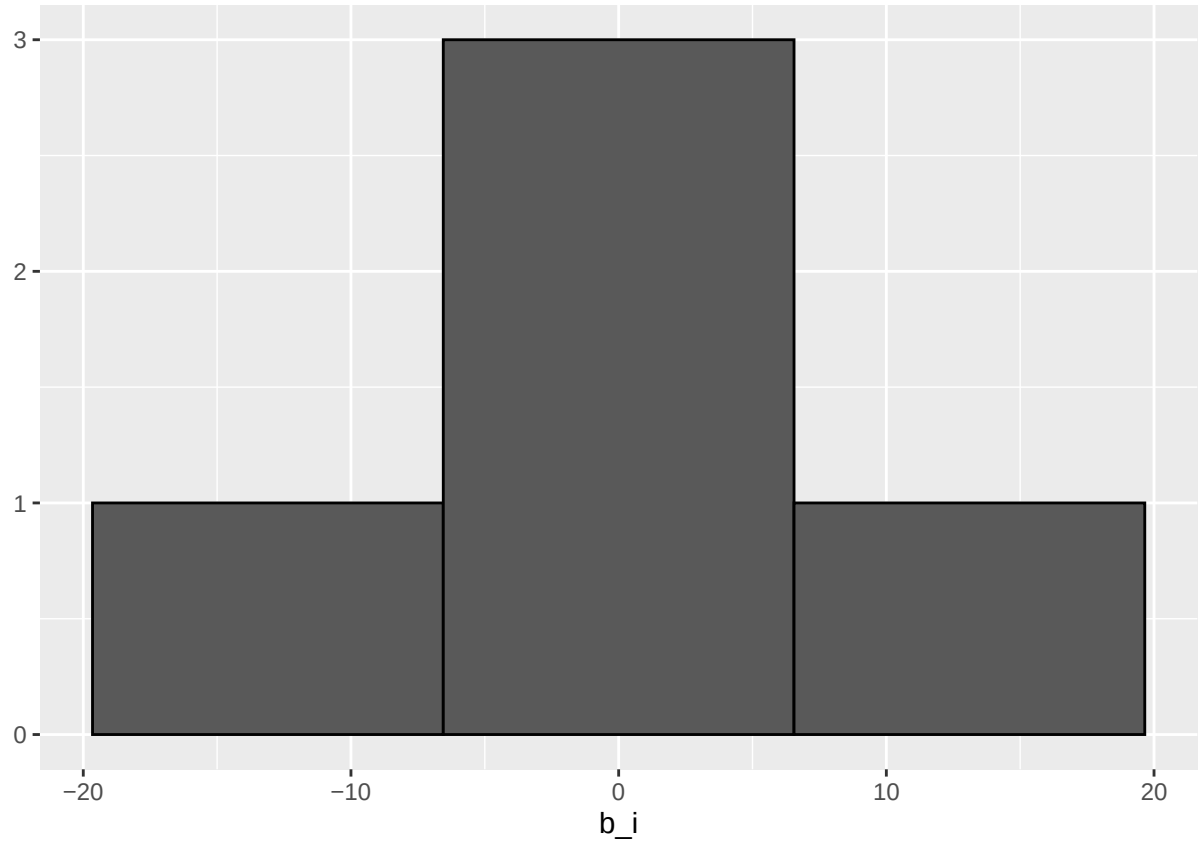
```
    mu_hat_2 <- mean(df12_t$pay)
    mu_hat_2
```

```
## [1] 26.51901
```

```
    edu_avgs <- train_set %>%
      group_by(edu_5) %>%
      summarize(b_i = mean(pay - mu_hat_2))
    edu_avgs
```

```
## # A tibble: 5 x 2
##   edu_5    b_i
##   <chr>  <dbl>
## 1 a      15.1
## 2 b       6.19
## 3 c      -4.06
## 4 d      -6.37
## 5 e     -11.2
```

```
    edu_avgs %>% qplot(b_i, geom ="histogram", bins = 3, data = ., color = I("black"))
```




```
    predicted_pay <- mu_hat_2 + test_set %>%
      left_join(edu_avgs, by='edu_5') %>%
      pull(b_i)
    head(predicted_pay)
```

```
## [1] 15.36634 15.36634 15.36634 15.36634 15.36634 15.36634
```

```
    model_1_rmse <- RMSE(predicted_pay, test_set$pay)
    model_1_rmse
```

```
## [1] 4.496685
```

```
    rmse_results <- bind_rows(rmse_results,
          data_frame(method="Educ Effect Model",
          RMSE = model_1_rmse ))
    rmse_results
```

```
## # A tibble: 2 x 2
##   method             RMSE
##   <chr>             <dbl>
## 1 Just the average  10.7
## 2 Educ Effect Model  4.50
```

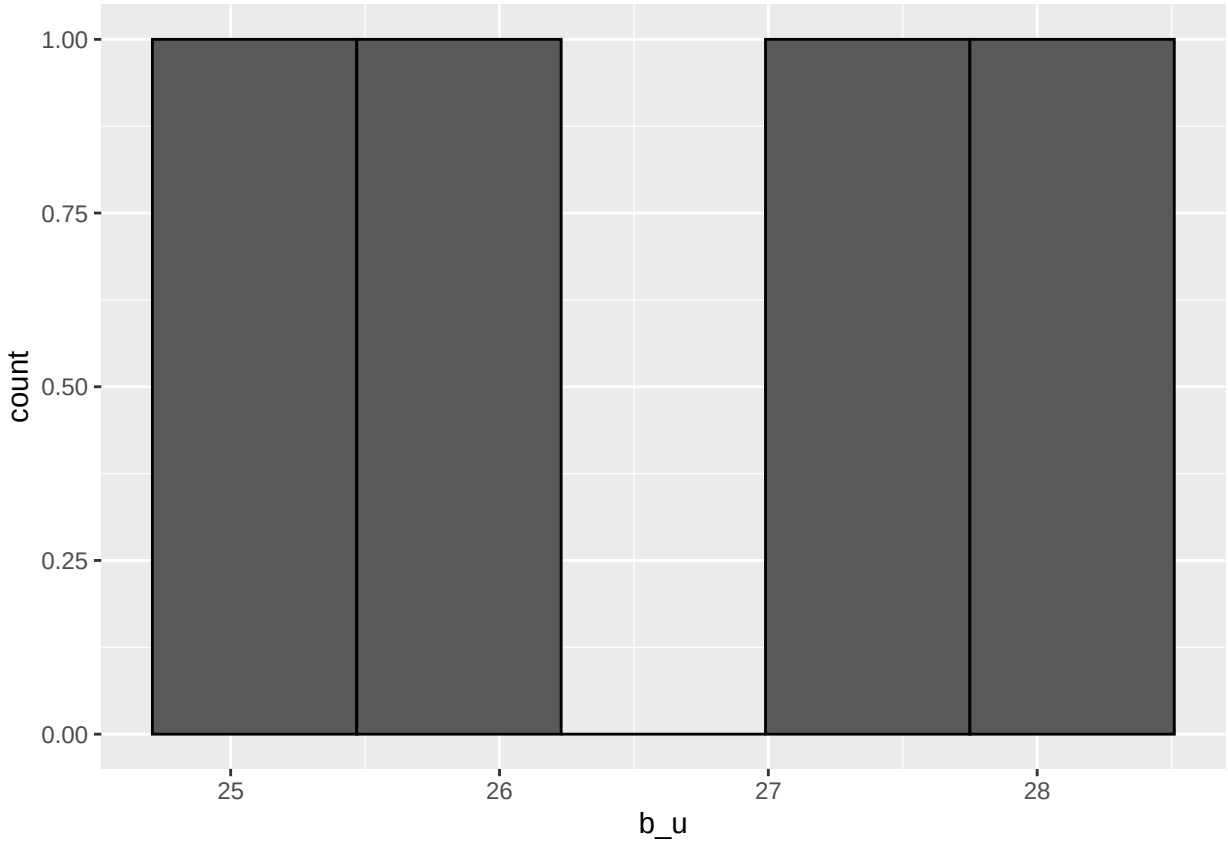
```
    rmse_results %>% knitr::kable()
```

method	RMSE
Just the average	10.702153
Educ Effect Model	4.496684

MODEL 3.

Modeling RACE effect on top of edu effect:

```
      train_set %>%
          group_by(race) %>%
          summarize(b_u = mean(pay)) %>%
          ggplot(aes(b_u)) +
          geom_histogram(bins = 5, color = "black")
```




```
      #graphic shows distribution around b_u of ca.26,5

  race_avgs <-train_set %>%
       left_join(edu_avgs, by='edu_5') %>%
       group_by(race) %>%
       summarize(b_u = mean(pay - mu_hat - b_i))
  race_avgs
```

```
## # A tibble: 4 x 2
##   race        b_u
##   <chr>     <dbl>
## 1 All       0.941
## 2 Black    -0.876
## 3 Hispanic -1.59
## 4 White     1.54
```

```
  predicted_pay_2 <- test_set %>%
      left_join(edu_avgs, by='edu_5') %>%
      left_join(race_avgs, by='race') %>%
      mutate(pred = mu_hat + b_i + b_u) %>%
      pull(pred)
  head(predicted_pay_2)
```

```
## [1] 14.49031 14.49031 14.49031 14.49031 14.49031 14.49031
```

```
      model_2_rmse <- RMSE(predicted_pay_2, test_set$pay)
      model_2_rmse
```

```
## [1] 4.318814
```

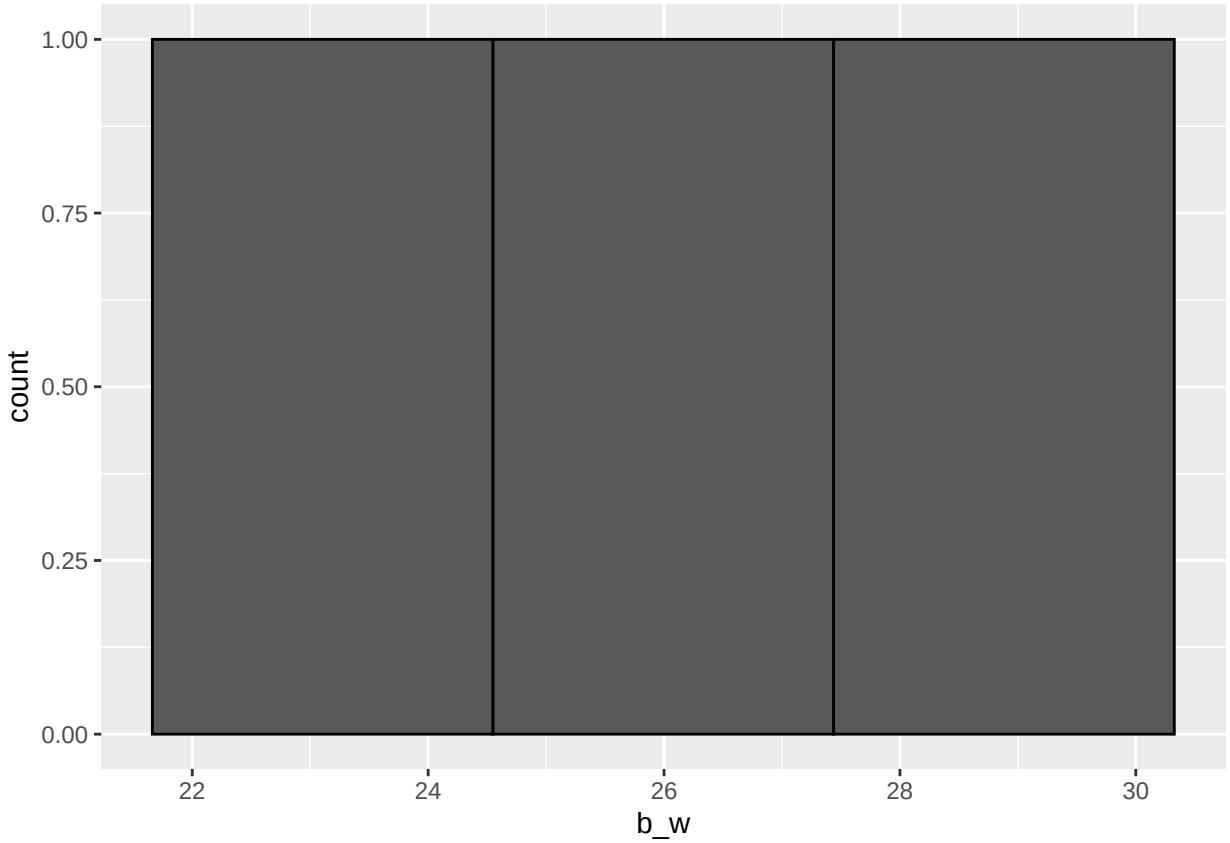
```
      rmse_results <- bind_rows(rmse_results,
            data_frame(method="Educ + Race Effects Model",
            RMSE = model_2_rmse ))
      rmse_results %>% knitr::kable()
```

method	RMSE
Just the average	10.702153
Educ Effect Model	4.496684
Educ + Race Effects Model	4.318814

MODEL 4.

Adding GENDER effect on top of race effect on top of edu effect:

```
      train_set %>%
        group_by(gender) %>%
        summarize(b_w = mean(pay)) %>%
        ggplot(aes(b_w)) +
        geom_histogram(bins = 3, color = "black")
```



```
        gender_avgs <-train_set %>%
          left_join(edu_avgs, by='edu_5') %>%
          left_join(race_avgs, by='race') %>%
          group_by(gender) %>%
          summarize(b_w = mean(pay - mu_hat - b_i - b_u))
        gender_avgs
```

```
## # A tibble: 3 x 2
##   gender    b_w
##   <chr>   <dbl>
## 1 All    -0.133
## 2 Men     2.99
## 3 Women  -2.81
```

```
        predicted_pay_03 <- test_set %>%
          left_join(edu_avgs, by='edu_5') %>%
          left_join(race_avgs, by='race') %>%
          left_join(gender_avgs, by='gender')%>%
          mutate(pred = mu_hat + b_i + b_u +b_w) %>%
          pull(pred)
        head(predicted_pay_03)
```

```
## [1] 17.47558 17.47558 17.47558 17.47558 17.47558 17.47558
```

```
        model_03_rmse <- RMSE(predicted_pay_03, test_set$pay)
        model_03_rmse
```

```
## [1] 3.644879
```

```
        #[1] 3.644879

        rmse_results <- bind_rows(rmse_results,
                                  data_frame(method="Educ + Race + Gender Effects Model",
                                             RMSE = model_03_rmse ))
        rmse_results %>% knitr::kable()
```

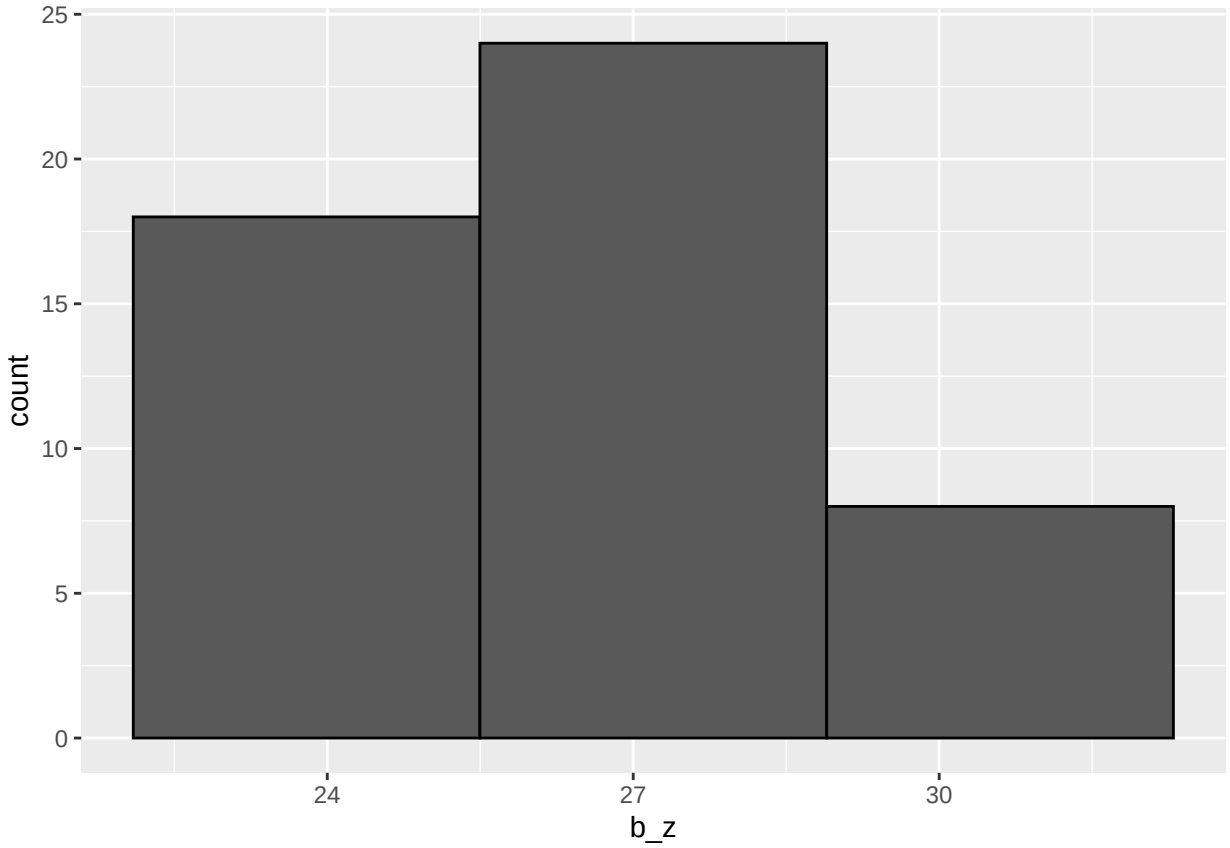
method	RMSE
Just the average	10.702153
Educ Effect Model	4.496684
Educ + Race Effects Model	4.318814
Educ + Race + Gender Effects Model	3.644879

MODEL 5.

Effect of education in the model is important. Also taking into account race and Gender effects help improve the model's results.

Adding YEAR effects:

```
train_set %>%
  group_by(year) %>%
  summarize(b_z = mean(pay)) %>%
  ggplot(aes(b_z)) +
  geom_histogram(bins = 3, color = "black")
```




```
year_avgs <-train_set %>%
  left_join(edu_avgs, by='edu_5') %>%
  left_join(race_avgs, by='race') %>%
  left_join(gender_avgs, by='gender')%>%
  group_by(year) %>%
  summarize(b_z = mean(pay - mu_hat - b_i - b_u - b_w))
year_avgs
```

```
## # A tibble: 50 x 2
##     year    b_z
##    <int>  <dbl>
##  1  1973 -0.680
##  2  1974 -1.50
##  3  1975 -2.41
##  4  1976 -1.28
##  5  1977 -1.86
##  6  1978 -1.09
##  7  1979 -1.09
##  8  1980 -1.94
```

```
##  9  1981 -2.12
## 10  1982 -2.83
## # ... with 40 more rows
```

```
        predicted_pay_04 <- test_set %>%
          left_join(edu_avgs, by='edu_5') %>%
          left_join(race_avgs, by='race') %>%
          left_join(gender_avgs, by='gender')%>%
          left_join(year_avgs, by='year')%>%
          mutate(pred = mu_hat + b_i + b_u +b_w +b_z) %>%
          pull(pred)
        head(predicted_pay_04)
```

```
## [1] 21.40197 19.97702 18.33277 18.67511 19.19940 19.08335
```

```
        model_04_rmse <- RMSE(predicted_pay_04, test_set$pay)
        model_04_rmse
```

```
## [1] 3.262347
```

```
        rmse_results <- bind_rows(rmse_results,
                                  data_frame(method="Educ + Race + Gender + Year Effects Model",
                                             RMSE = model_04_rmse ))
        rmse_results %>% knitr::kable()
```

method	RMSE
Just the average	10.702153
Educ Effect Model	4.496684
Educ + Race Effects Model	4.318814
Educ + Race + Gender Effects Model	3.644879
Educ + Race + Gender + Year Effects Model	3.262347

MODEL 6.

REGULARIZATION

-By considering the education, race, gender and year effect in our model, we have improved model accuracy, now well below the initial 10,7 value. -We could further improve model accuracy, by using regularization, removing effects of noisy estimates (penalizing large estimates coming from small sample sizes).Anomalies in pay curves through time.

Penalized least squares Choosing lambda equal 3.

```
       lambda <- 3
       mu <- mean(train_set$pay)
       edu_train_avgs <- train_set %>%
            group_by(edu_5) %>%
            summarize(b_i=sum(pay-mu)/(n()+lambda),n_i=n())
       edu_train_avgs
```

```
## # A tibble: 5 x 3
##   edu_5    b_i   n_i
##   <chr>  <dbl> <int>
## 1 a      15.0    483
## 2 b       6.19   482
## 3 c      -3.99   476
## 4 d      -6.29   474
## 5 e     -11.0    484
```

```
      #results
      predicted_pay_05 <- test_set %>%
        left_join(edu_train_avgs, by = "edu_5") %>%
        mutate(pred=mu + b_i) %>%
        pull(pred)
      head(predicted_pay_05)
```

```
## [1] 15.43481 15.43481 15.43481 15.43481 15.43481 15.43481
```

```
      sum(is.na(predicted_pay_05)) #[1] 0
```

```
## [1] 0
```

```
      model_5_rmse <- RMSE(predicted_pay_05, test_set$pay)
      model_5_rmse
```

```
## [1] 4.501724
```

```
         predicted_pay_05na <-replace_na(predicted_pay_05, mu_hat)
         head(predicted_pay_05na)
```

```
## [1] 15.43481 15.43481 15.43481 15.43481 15.43481 15.43481
```

```
         #RMSE(predicted_pay_05na,test_set$pay)
         model_5_rmse.na <- RMSE(predicted_pay_05na, test_set$pay)
         model_5_rmse.na
```

```
## [1] 4.501724
```

```
      rmse_results <- bind_rows(rmse_results,
                data_frame(method="Regularized Educ Effect Model",
                RMSE = model_5_rmse))
      rmse_results %>% knitr::kable()
```

method	RMSE
Just the average	10.702153
Educ Effect Model	4.496684
Educ + Race Effects Model	4.318814
Educ + Race + Gender Effects Model	3.644879

method	RMSE
Educ + Race + Gender + Year Effects Model	3.262347
Regularized Educ Effect Model	4.501724

MODEL 7.

Cross-validation - we'll use cross-validation for choosing a lambda:

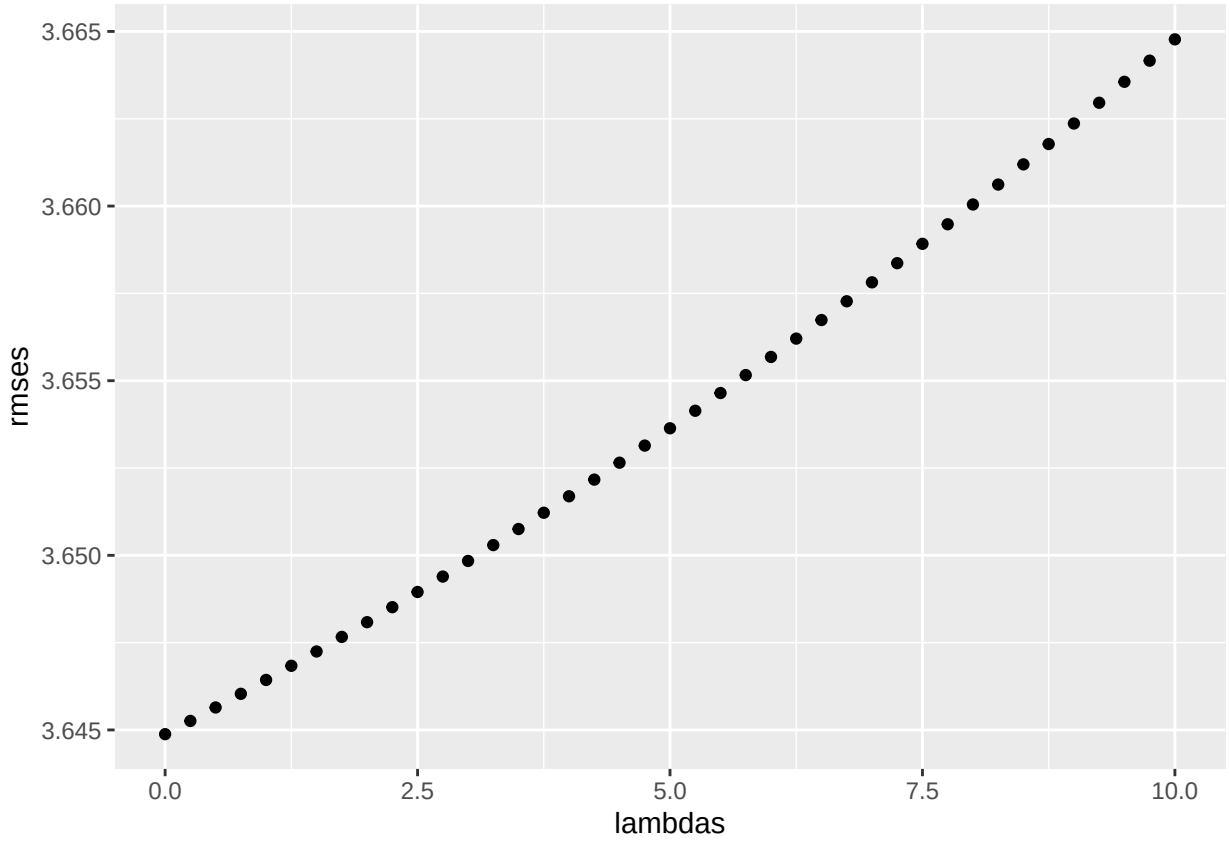
```
        lambdas <- seq(0, 10, 0.25)

        head(train_set$pay)
```

```
## [1] 15.38 15.52 15.07 15.00 14.64 14.55
```

```
        rmses <- sapply(lambdas, function(l){
          mu <- mean(train_set$pay)
          b_i <- train_set %>%
            group_by(edu_5) %>%
            summarize(b_i = sum(pay - mu)/(n()+l))
          b_u <- train_set %>%
            left_join(b_i, by="edu_5") %>%
            group_by(race) %>%
            summarize(b_u = sum(pay - b_i - mu)/(n()+l))
          b_w <- train_set %>%
            left_join(b_i, by="edu_5") %>%
            left_join(b_u, by="race") %>%
            group_by(gender) %>%
            summarize(b_w = sum(pay - b_i -b_u - mu)/(n()+l))
          predicted_pay_06 <-
            test_set %>%
            left_join(b_i, by = "edu_5") %>%
            left_join(b_u, by = "race") %>%
            left_join(b_w, by = "gender") %>%
            mutate(pred = mu + b_i + b_u + b_w) %>%
            pull(pred)
          return(RMSE(predicted_pay_06, test_set$pay))
        })

        qplot(lambdas, rmses)
```




```
        lambda <- lambdas[which.min(rmses)]
        lambda
```

```
## [1] 0
```

```
        #[1] 0
        lambda_REE <-0
        lambda_REE
```

```
## [1] 0
```

```
        rmse_lambda_REE <- sapply(lambda_REE, function(l){
          mu <- mean(train_set$pay)
          b_i <- train_set %>%
            group_by(edu_5) %>%
            summarize(b_i = sum(pay - mu)/(n()+l))
          b_u <- train_set %>%
            left_join(b_i, by="edu_5") %>%
            group_by(race) %>%
            summarize(b_u = sum(pay - b_i - mu)/(n()+l))
          b_w <- train_set %>%
            left_join(b_i, by="edu_5") %>%
            left_join(b_u, by="race") %>%
            group_by(gender) %>%
```

```
            summarize(b_w = sum(pay - b_i -b_u - mu)/(n()+l))
          predicted_pay_06 <-
            test_set %>%
            left_join(b_i, by = "edu_5") %>%
            left_join(b_u, by = "race") %>%
            left_join(b_w, by = "gender") %>%
            mutate(pred = mu + b_i + b_u + b_w) %>%
            pull(pred)
          return(RMSE(predicted_pay_06, test_set$pay))
        })

        rmse_lambda_REE
```

```
## [1] 3.644879
```

```
        #[1] 3.644879

    rmse_results <- bind_rows(rmse_results,
             data_frame(method="Regularized Educ+Race+Gender Effect Model",
                                                    RMSE = rmse_lambda_REE ))
    rmse_results %>% knitr::kable()
```

method	RMSE
Just the average	10.702153
Educ Effect Model	4.496684
Educ + Race Effects Model	4.318814
Educ + Race + Gender Effects Model	3.644879
Educ + Race + Gender + Year Effects Model	3.262347
Regularized Educ Effect Model	4.501724
Regularized Educ+Race+Gender Effect Model	3.644879

Regularization is not doing much here, we could conclude there isn't much "noise" to be neutralized. Our model including "year" produces the best RMSE. MODEL 5, considering edu_5, year, race, and gender variables, provides the lowest RMSE (3,2623). This is the best result - recommendation model- we've developed in this study.

Part 6 - CONCLUSIONS.

GENERAL CONCLUSIONS after data wrangling and visualization, and modeling:

1.All variables (4 independent), education level (edu_5), time (year), gender (gender), race (race) have effects on wages (pay).

2.Level of EDUCATION has the highest effect on pay. The higher the education level, the higher wages,through time and for any race and gender;

3.TIME has an effect on pay: -3.1."a" (advanced_degree) and "b" (bachelors_degree) levels growing through time; -3.2."c" (some_college) and "d" (high_school) stagnating; -3.3."e" (lower_than_high-school)level decreasing through time. Through time pay rates are increasing for higher levels of education (a, b), and showing stagnation for the lower levels of education (c,d),and decreasing for the lowest (e), across races and genders.

4.RACE consistently (through time) affects the level of pay for the same education level,being through time (in general, except a few anomalies) White the highest paid, Hispanic following, and Black last;

5.GENDER has been and continues to be a factor affecting pay level. Men are better paid than Women across education level and race;

6.Regarding combinations of RACE-GENDER: -White_Men are the best paid at equal level of education, for all 5 education levels; the best paid of any race-gender combination. Specially for higher education levels. -Hispanic_Women are the worst paid group at the highest education level.

Effects of the 4 independent variables are shown on the various modeling approaches we've utilized in this study, coherent with these observations.

Note this study is based on the available data (selected data frame), and is therefore limited by the extension and depth of the data. As an example ,three races are considered only, while population in the USA includes a larger number of ethnicity.

Nevertheless, this the study illustrates, within its limitations, the impact- effect, of the various considered demographics variables.The study could be further enriched by adding additional relevant complementary data to it.