Homework

CPES 2 - Fall 2022

Louis Sirugue

Guidelines:

- This homework can be done by **groups of 2 at most**.
- Write your **answers directly in this .Rmd** file, which must compile without errors.
- Make sure that your answers are unambiguous and that your code is annotated with comments.
- You can write either in English or in French.
- Save and knit your file regularly to avoid any last-minute technical inconvenience.
- Both the .Rmd and the knitted .html or .pdf should be sent in a **.zip by email** (louis.sirugue@psemail.eu)
  - If you are not familiar with .zip files, first download and install winRAR.
  - Then select your .Rmd and .html files > right click > Add to archive > select ''zip" and save.
- Deadline: **Sunday the 16th of October 2022, 23:59**.
- Late submissions will be penalized by half a point for each 30min beyond the deadline.
- The number of points associated with each question is indicative and may be subject to modifications.

This homework covers the material from lectures 1 to 5. I encourage you to start **working on it progressively** as we go through the course. Contrarily to the online quizzes, this homework is meant to be challenging. **Partial answers will be taken into account**, so if you cannot do everything, **write down your thought process**.

Context:

This homework consists in analyzing data from a survey conducted for the academic article available here. In this article, authors study individuals' **preferences regarding income redistribution**, and how these preferences may vary if individuals are more or less aware of the extent and consequences of income inequality.

Authors conducted an online survey in which respondents had to answer to two waves of questions:

- **Wave 1: General information:** gender, age, education, marital status, household income, ...
- **Wave 2: Questions on income inequality/redistribution:** if inequality is a problem, if high incomes deserve their income, whether income is due to effort or luck, ...

Each participant to the survey was allocated to one of **two groups:** either to the **control** group, or to the **treatment** group. Between the two waves of questions, **information** on the extent and consequences of income inequality **were given to individuals in the treatment group** but not to individuals in the control group.

Data:

The datasets for this homework are available at this webpage. To **download the data** you will have to create an ICPSR account. You can do it with your @psl.eu address. You will also find an **Online Appendix** containing documentation on the experiment and the dataset, including the questions and possible answers of the survey.

Questions:

1) Load the `haven` package and import the file `Inequality_AER_Omnibus.dta` using the dedicated function (/1)

Note: `.dta` *data imported with haven often come with labels. These labels can cause errors when reshaping the data with* `pivot_longer()` *and plotting the data with* `ggplot()`*. You can remove these labels at any point using the function* `zap_labels()`*.*

```
library(haven)
omnibus <- read_dta("Inequality_AER_Omnibus.dta")
```

2) Use the `filter()` function from `dplyr` to keep only individuals older than 18, who finished the survey, and who belong either to the `Treatment Group` or `Control` group. (/1)

```
library(dplyr)
omnibus <- omnibus %>%
  filter(age > 18 & finished == 1 & group %in% c("Treatment Group", "Control"))
```

3) Given the answers of the survey, do people with higher household income tend to be more or less satisfied with their income? (/1)

It is specified in the documentation of the survey that the answers to the satisfaction question are the following: *Very satisfied*, *Somewhat satisfied*, *Not too satisfied*, and *Not at all satisfied*. We can compute the share of individuals choosing each answer by household income group as follows.

```
library(knitr)
omnibus %>%
  filter(!(is.na(hhincome)|is.na(satisfied))) %>%
  group_by(hhincome) %>%
  summarise(Very_satisfied = 100 * mean(satisfied == 1),
            Somewhat_satisfied = 100 * mean(satisfied == 2),
            Not_too_satisfied = 100 * mean(satisfied == 3),
            Not_at_all_satisfied = 100 * mean(satisfied == 4)) %>%
  kable(digits = 2)
```

hhincome	Very_satisfied	Somewhat_satisfied	Not_too_satisfied	Not_at_all_satisfied
1	2.81	11.56	20.94	64.69
2	3.88	11.17	39.32	45.63
3	1.00	21.89	32.84	44.28
4	3.19	19.85	39.46	37.50
5	2.30	28.39	41.69	27.62
6	3.24	38.05	35.69	23.01
7	4.23	45.60	32.25	17.92
8	9.63	53.48	25.67	11.23
9	21.92	49.32	16.44	12.33
10	15.38	64.62	4.62	15.38
11	32.61	41.30	13.04	13.04
12	13.38	48.41	30.57	7.64

We can see that the higher the household income group, the higher the proportion of respondents who are satisfied with their income.

4) We are interested in knowing whether or not treated individuals answered differently to the questions *'Inequality is a problem?'*, *'Inequality has increased?'*, and *'Do you think high incomes deserve their income?'*. Find out a convenient way to summarise the 3 corresponding variables into their mean, their standard deviation, and their number of non-missing values, separately for the treatment and control group. Display your results in a good-looking table. (/2)

```
library(tidyr)
library(knitr)

mean_table <- omnibus %>%
  select(group, inequality_problem, inequality_increase, deserving) %>%
  zap_labels() %>%
  pivot_longer(!group, names_to = "variable", values_to = "value") %>%
  group_by(variable, group) %>%
  summarise(mean = mean(value, na.rm = T),
            sd = sd(value, na.rm = T),
            n = sum(!is.na(value)))

kable(mean_table, digits = 2)
```

variable	group	mean	sd	n
deserving	Control	2.00	0.61	2015
deserving	Treatment Group	2.16	0.62	1183
inequality_increase	Control	2.61	0.69	2015
inequality_increase	Treatment Group	2.77	0.56	1183
inequality_problem	Control	3.61	1.16	2014
inequality_problem	Treatment Group	3.89	1.15	1183

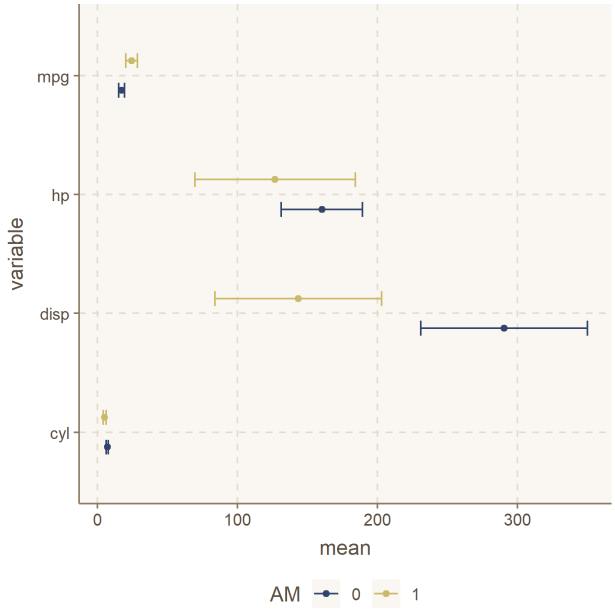
5) Write down the formula of the standard error of the mean using a LaTeX equation, and compute the 97% confidence interval of each mean. (/2)

$$\text{S.E.}(x) = \frac{\text{SD}(x)}{\sqrt{N}} = \frac{\sqrt{\text{Var}(x)}}{\sqrt{N}} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}}{\sqrt{N}}$$

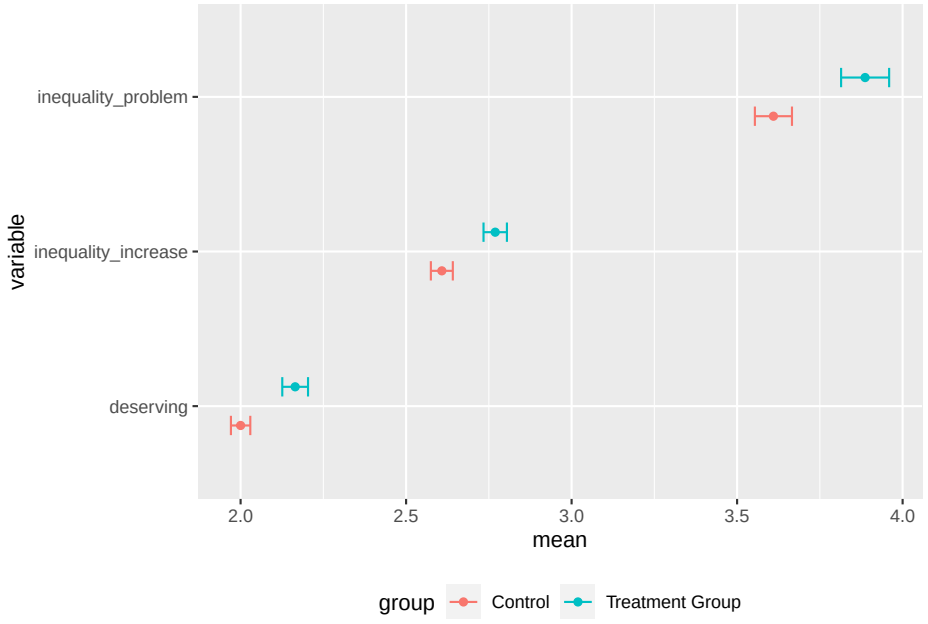
```
mean_table <- mean_table %>%
  mutate(se = sd / sqrt(n),
         t = qt(1 - ((1 - .97)/2), n - 1),
         lb = mean - t*se,
         ub = mean + t*se)

kable(mean_table, digits = 2)
```

variable	group	mean	sd	n	se	t	lb	ub
deserving	Control	2.00	0.61	2015	0.01	2.17	1.97	2.03
deserving	Treatment Group	2.16	0.62	1183	0.02	2.17	2.13	2.20
inequality_increase	Control	2.61	0.69	2015	0.02	2.17	2.57	2.64
inequality_increase	Treatment Group	2.77	0.56	1183	0.02	2.17	2.73	2.80
inequality_problem	Control	3.61	1.16	2014	0.03	2.17	3.55	3.67
inequality_problem	Treatment Group	3.89	1.15	1183	0.03	2.17	3.81	3.96

6) Produce a graph like the following one to represent the mean and confidence interval of the three variables separately for the two groups. Do you notice a specific pattern? (/3)




```
library(ggplot2)
ggplot(mean_table, aes(x = variable, y = mean, color = group)) +
  geom_point(position = position_dodge(.5)) +
  geom_errorbar(aes(xmin = variable, xmax = variable, ymin = lb, ymax = ub),
                position = position_dodge(.5), width = .25) +
  coord_flip() + theme(legend.position = "bottom", legend.direction = "horizontal")
```



The treatment group always has a higher mean than the control group, and the confidence intervals do not overlap. Being given the treatment seems to have an effect on the preferences for redistribution.

7) Consider the mean of the variable `inequality_increase` for the treatment group. What is the confidence level below which you would not consider, and above which you would consider, that this mean value could be equal to 2.75. (/3)

Hint: While `qt(proba, df)` *gives the t-stat corresponding to a given probability to fall below this t-stat in a Student t distribution with a given number of degrees of freedom,* `pt(t_stat, df)` *gives the probability to fall below a given t-stat in a Student t distribution with a given number of degrees of freedom.*

Given that the mean of the variable `inequality_increase` for the treatment group is larger than 2.75, we are interested in the lower bound of the confidence interval. The formula for the lower bound of the confidence interval writes as follows.

$$\bar{x} - t_{\text{df},\, 1-\frac{1-\text{CL}}{2}} \times \text{SE}(x),$$

Where CL denotes the confidence level. We can find the confidence level whose confidence interval lower bound equals 2.75 by solving the following equation.

$$\begin{aligned} \bar{x} - t_{\text{df},\, 1-\frac{1-\text{CL}}{2}} \times \text{SE}(x) &= 2.75 \\ t_{\text{df},\, 1-\frac{1-\text{CL}}{2}} &= \frac{\bar{x} - 2.75}{\text{SE}(x)} \end{aligned}$$

We can compute the value of this t-stat:

```
x <- omnibus$inequality_increase[omnibus$group == "Treatment Group"]

t_stat <- (mean(x) - 2.75) / (sd(x)/sqrt(length(x)))
t_stat
```

```
## [1] 1.182216
```

Then we can use `pt(t_stat, df)` to get the probability the fall below this t-stat with a Student t distribution with the corresponding number of degrees of freedom:

```
prob <- pt(t_stat, df = length(x) - 1)
prob
```

```
## [1] 0.8813212
```

And we can compute the corresponding confidence level:

$$\begin{aligned} 1 - \frac{1-\text{CL}}{2} &= 0.8813212 \\ \text{CL} &= 1 - 2 \times (1 - 0.8813212) \end{aligned}$$

```
1 - (2*(1-prob))
```

```
## [1] 0.7626423
```

The lower bound of the confidence interval equals 2.75 when the confidence level is 76%. Below this threshold, we would not consider that the mean could be equal to 2.75, but we would consider it plausible above this threshold.

8) We would like to know whether the difference in opinion on *'deserving'* between the control and the treatment group persisted over time. A followup survey has been conducted, and the results are stored in `Inequality_AER_Followup.dta`. In this dataset, variables from the follow up survey end with the suffix `_s2`. Import the followup dataset, keep only the variables you need, and join it to the omnibus data. (/1)

```
followup <- read_dta("Inequality_AER_Followup.dta") %>%
  select(responseid, deserving_s2)

omnibus <- omnibus %>%
  left_join(followup, by = "responseid")
```

9) Compute the fraction of the initial sample that appears in the followup dataset, and the response rate to the *'deserving'* question in the followup survey. (/1)

```
mean(omnibus$responseid %in% followup$responseid)
```

```
## [1] 0.3061288
```

```
mean(!is.na(followup$deserving_s2))
```

```
## [1] 0.1443696
```

10) Would you conclude that the difference in opinion on *'deserving'* between the control and the treatment group persisted over time? (/2)

```
omnibus %>%
  group_by(group) %>%
  summarise(mean = mean(deserving_s2, na.rm = T),
            sd = sd(deserving_s2, na.rm = T),
            n = sum(!is.na(deserving_s2))) %>%
  mutate(se = sd / sqrt(n),
         t = qt(1 - ((1 - .95)/2), n - 1),
         lb = mean - t*se,
         ub = mean + t*se)
```

```
## # A tibble: 2 x 8
##   group            mean    sd     n     se     t    lb    ub
##   <chr>           <dbl> <dbl> <int>  <dbl> <dbl> <dbl> <dbl>
## 1 Control          1.92 0.583    76 0.0669  1.99  1.79  2.05
## 2 Treatment Group  1.98 0.673    65 0.0835  2.00  1.82  2.15
```

Even though the mean value is still higher for the treatment group in the second wave, the precision is not sufficient to conclude that the two means are different at a reasonable confidence level.

11) Make the best graph you can using the variables of your choice. The graph should be preceded by a description of the variables used, both in words and using appropriate statistics. The aim of the graph can be to document a relationship that has not been explored in this homework, to investigate the heterogeneity of a relationship that was studied in the homework, etc. The graph should be clear and relevant. The interest of your graph could lie in how efficient it is to convey a story about your data, or in how it helps uncover underlying patterns in your data. (/3)

If you do not find inspiration with the variables available in the dataset, feel free to do this exercise using variables from another dataset (from the list below or other sources). Using another dataset will not grant extra points, but if it helps you making an interesting graph it can be worth it.

- data.gouv.fr
- data.gov
- data.oecd.org
- data.worldbank.org
- ec.europa.eu
- www.bfi.org.uk
- apps.who.int
- dataverse.harvard.edu
- archive.ics.uci.edu
- kaggle.com
- datahub.io
- wid.world
- opportunityinsights.org
- ...

Example with variables from the dataset The dataset contains information on individuals' income level, and perception on whether income is due to effort or luck. We can expect that the higher one's income, the more likely to believe that income is due to effort rather than luck. Given that the dataset also provides individuals' education level, it can be interesting to study how the relationship between the two variables vary with educational attainment.

Let's keep only the variables we need and clean the subset.

```
plot_dt <- omnibus %>%
  select(hhincome, worldbelief, education)
```

```
plot_dt %>%
  group_by(worldbelief) %>%
  tally()
```

World belief

```
## # A tibble: 3 x 2
##                                           worldbelief     n
##                                             <dbl+lbl> <int>
## 1  1 [One's individual effort]                          1302
## 2  2 [One's family background, luck, or health issue]   1894
## 3 NA                                                       2
```

```
plot_dt %>%
  group_by(education) %>%
  tally()
```

Education

```
## # A tibble: 9 x 2
##                                education     n
##                                <dbl+lbl> <int>
## 1 1 [Eighth Grade or less]                 3
## 2 2 [Some High School]                    40
## 3 3 [High School Degree/GED]             403
## 4 4 [Some College]                      1027
## 5 5 [2-year College Degree]              332
## 6 6 [4-year College Degree]              982
```

```
## 7 7 [ Master's Degree]                    315
## 8 8 [Doctoral Degree]                      40
## 9 9 [Professional Degree (JD, MD, MBA)]    56
```

```
plot_dt %>%
  group_by(hhincome) %>%
  tally()
```

Household income

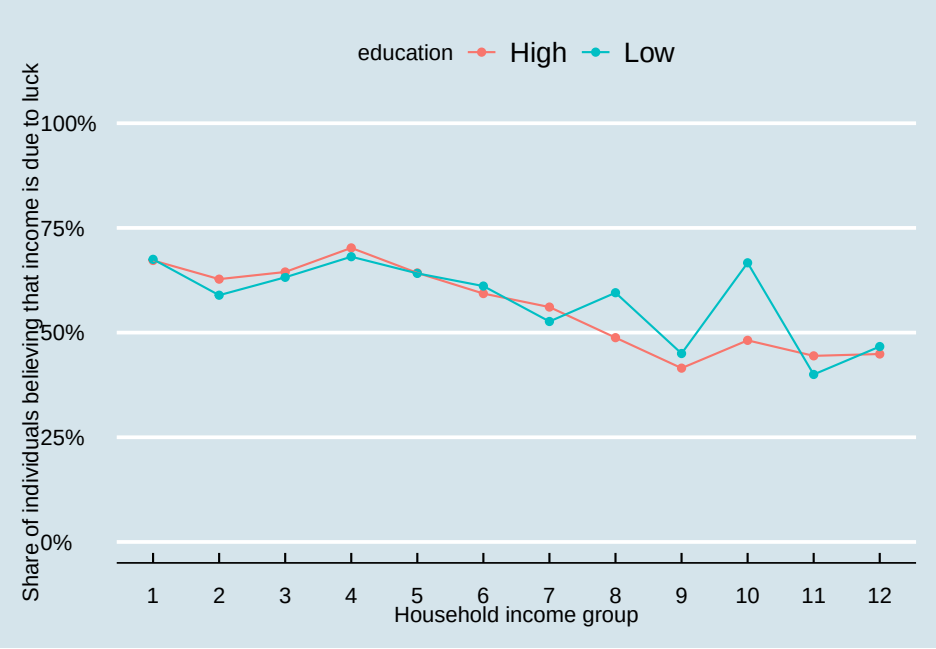
```
## # A tibble: 13 x 2
##                  hhincome     n
##                 <dbl+lbl> <int>
##  1  1 [$0-$9,999]           320
##  2  2 [$10,000-$14,999]     206
##  3  3 [$15,000-$19,999]     201
##  4  4 [$20,000-$29,999]     408
##  5  5 [$30,000-$39,999]     391
##  6  6 [40,000-$49,999]      339
##  7  7 [$50,000-$74,999]     614
##  8  8 [$75,000-$99,999]     374
##  9  9 [$100,000-$124,999]    73
## 10 10 [$125,000-$149,999]    66
## 11 11 [$150,001-$199,999]    46
## 12 12 [$200,000+]           157
## 13 NA                         3
```

We can remove individuals with missing values and recode the variables conveniently.

```
plot_dt <- plot_dt %>%
  filter(!(is.na(worldbelief) | is.na(hhincome))) %>%
  mutate(worldbelief = ifelse(worldbelief == 1, "Effort", "Luck"),
         education = ifelse(education > 4, "High", "Low"))
```

We can now plot how the average perception of income merit evolves with income, separately for the 2 education groups

```
library(ggthemes)

plot_dt %>%
  zap_labels() %>%
  group_by(hhincome, education) %>%
  summarise(bl = mean(worldbelief == "Luck")) %>%
  ggplot(aes(x = hhincome, y = bl, color = education)) +
  geom_point() + geom_line() +
  scale_x_continuous(name = "Household income group", limits = c(1, 12), breaks = 1:12) +
  scale_y_continuous(name = "Share of individuals believing that income is due to luck",
                     limits = c(0, 1), breaks = seq(0, 1, .25),
                     labels = c("0%", "25%", "50%", "75%", "100%")) +
  theme_economist()
```



It seems like indeed the higher the income level the more likely to believe that income is due to effort rather than luck, but this trend seems to be quite homogeneous across education groups.

Example with another data source The package `HSAUR` comes with a data set of **2 variables** on the **47 stars** of the CYGOB1 star cluster in the Cygnus constellation:

- Effective temperature: $\log(T_e)$
- Light intensity: $\log(L/L_0)$

```
library(HSAUR)
data("CYGOB1", package = "HSAUR")
kable(head(CYGOB1, 5), caption = "5 first observations")
```

Table 4: 5 first observations

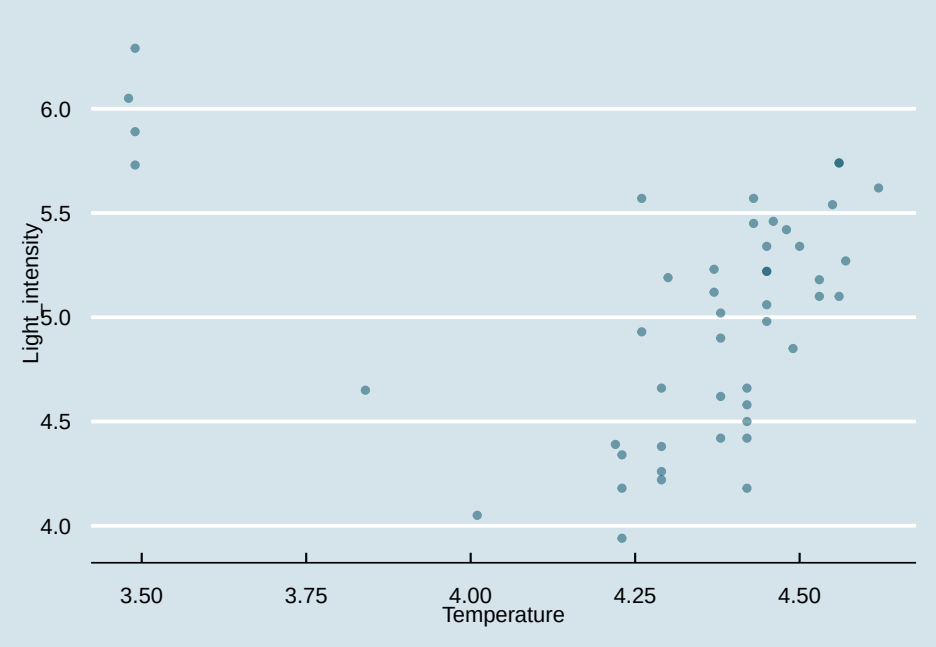
	logst	logli
[1,]	4.37	5.23
[2,]	4.56	5.74
[3,]	4.26	4.93
[4,]	4.56	5.74
[5,]	4.30	5.19

```
summary(CYGOB1)
```

```
##      logst           logli
##  Min.   :3.480   Min.   :3.940
##  1st Qu.:4.275   1st Qu.:4.540
##  Median :4.420   Median :5.100
##  Mean   :4.310   Mean   :5.012
##  3rd Qu.:4.455   3rd Qu.:5.435
##  Max.   :4.620   Max.   :6.290
```

We can use a scatterplot to represent the relationship between these two variables

```
CYGOB1 <- CYGOB1 %>%
  rename(Temperature = logst, Light_intensity = logli)

ggplot(CYGOB1, aes(x = Temperature, y = Light_intensity)) +
  geom_point(color = "#014D64", alpha = .5) +
  theme_economist()
```



The relationship is negative overall: the higher the temperature, the lower the light intensity. But this seems to be fallaciously driven by four stars at the top left. Indeed, the documentation of the data mentions that there are **two types of stars**:

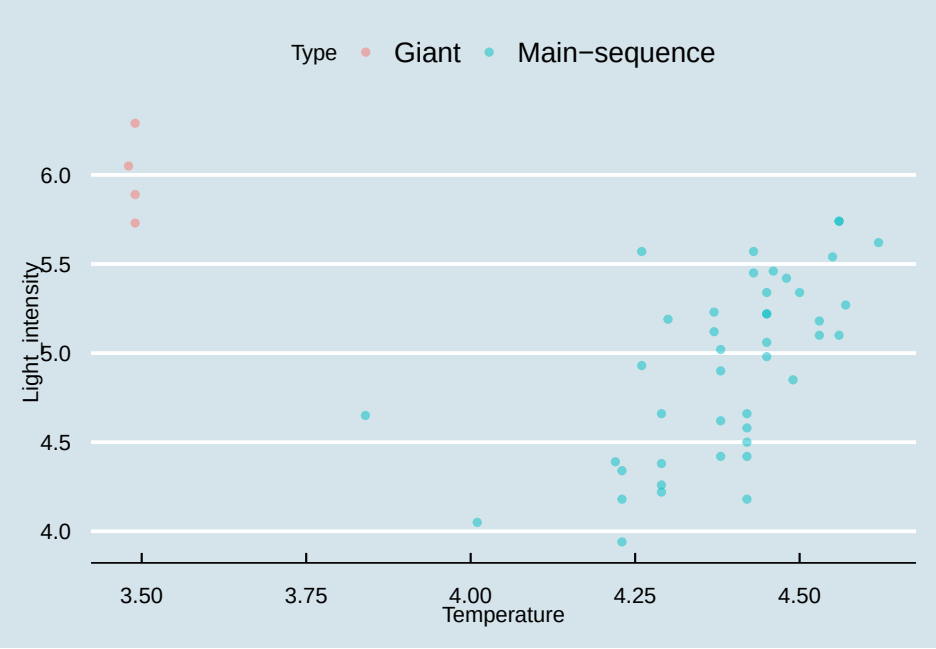
- Stars that lie on the main sequence
- Giants stars

There's no variable in the data to distinguish between these two groups, but the documentation indicates that Giants are located at the following rows:

- The 11th;
- The 20th;
- The 30th;
- The 34th.

Based on these information, a variable indicating the type of star can easily be created as follows, so that we can then distinguish the two types of stars in the plot by attributing them separate colors:

```
CYGOB1 <- CYGOB1 %>%
  mutate(Type = ifelse(row_number() %in% c(11, 20, 30, 34),
                       "Giant", "Main-sequence"))

ggplot(CYGOB1, aes(x = Temperature, y = Light_intensity, color = Type)) +
  geom_point(alpha = .5) + theme_economist()
```



By omitting the difference in type of stars, we would have fallaciously concluded that the relationship was somewhat negative. Actually we observe no relationship for the giants, and a strong positive relationship for stars that lie on the main sequence.