Take Home Assignment

From the original dataset of 26 variables, 14 were taken out.

- **NQUEST** and **nord** because they were IDs and so not at all relevant;
- **TIPOLAU, VOTOEDU, SUEDU, ANNOEDU** and **TIPODIP** because of the high number of NAs;
- **ANASC, NASCAREA, NASCREG, CLETA5, IREG, ACOM5** and **PARENT** because of the high correlation and redundancies.

The remaining and used variables being: **NCOMP, SEX, STACIV, STUDIO, ETA, PERC, PERL, NPERL, NPERC, AREA5, ACOM4C** and **LFP**.

After running a logistic regression with LFP as the dependent variable, I performed a stepwise model selection on the first regression:

	LFP		
Predictors	*Log-Odds*	*CI*	*p*
(Intercept)	-4.38 ***	-4.82 – -3.95	**<0.001**
NCOMP	0.27 ***	0.20 – 0.35	**<0.001**
SEX [2]	-0.50 ***	-0.65 – -0.35	**<0.001**
STACIV [2]	0.65 ***	0.47 – 0.84	**<0.001**
STACIV [3]	1.27 ***	0.89 – 1.64	**<0.001**
STACIV [4]	-0.37	-0.93 – 0.15	0.179
STUDIO	0.38 ***	0.34 – 0.42	**<0.001**
PERC	-1.76 ***	-2.00 – -1.52	**<0.001**
PERL	8.01 ***	7.70 – 8.33	**<0.001**
NPERL	-0.73 ***	-0.84 – -0.63	**<0.001**
AREA5	0.30 ***	0.24 – 0.36	**<0.001**
ACOM4C	-0.10 *	-0.18 – -0.02	**0.015**
Observations	14446		
R^2 Tjur	0.792		

** p<0.05 ** p<0.01 *** p<0.001*

Among the predictor variables, NCOMP, SEX, STACIV, STUDIO, PERC, PERL, NPERL, AREA5, and ACOM4C show statistically significant associations with LFP.

Education level (STUDIO), number of household components (NCOMP), having work income (PERL), marital status (STACIV) and the area of residence (AREA5) exhibits a positive association with labour force participation. This means that an increase in these variables (considering some of them ordinal categorical) leads to an increase of the log-odds of participating in the labour force.

On the other hand Gender (SEX), having personal income (PERC), the number of people in the household who have work income (NPERL), and the town size (ACOM4C) exhibits a negative association with labour force participation.

It seems that a higher education qualification increases one's employability while the negative coefficient for the female sex category (SEX [2]) indicates that women have a

lower likelihood of participating in the labor force compared to men. People with personal income aside from work also show a lower likelihood in LFP.

Surprisingly the age variable (ETA) was not statistically significant even in the first model before the stepwise model selection.

Based on these results I decided to split the dataset between male and female individuals to further analyze the differences of gender in LFP.

Male

	LFP		
Predictors	*Log-Odds*	*CI*	*p*
(Intercept)	-3.92 ***	-4.75 – -3.09	**<0.001**
NCOMP	0.28 ***	0.17 – 0.38	**<0.001**
STACIV [2]	0.14	-0.24 – 0.52	0.457
STACIV [3]	1.26 ***	0.69 – 1.82	**<0.001**
STACIV [4]	-1.03 *	-1.96 – -0.17	**0.026**
STUDIO	0.26 ***	0.20 – 0.33	**<0.001**
ETA	0.02 ***	0.01 – 0.03	**<0.001**
PERC	-3.27 ***	-3.69 – -2.86	**<0.001**
PERL	8.33 ***	7.91 – 8.76	**<0.001**
NPERL	-0.76 ***	-0.90 – -0.62	**<0.001**
AREA5	0.33 ***	0.26 – 0.41	**<0.001**
ACOM4C	-0.12 *	-0.22 – -0.01	**0.032**
Observations	7281		
R^2 Tjur	0.785		

*p<0.05 **p<0.01 ***p<0.001

Female

	LFP		
Predictors	*Log-Odds*	*CI*	*p*
(Intercept)	-4.68 ***	-5.58 – -3.79	**<0.001**
NCOMP	0.17 **	0.04 – 0.29	**0.009**
STACIV [2]	0.67 **	0.26 – 1.08	**0.001**
STACIV [3]	1.11 ***	0.52 – 1.67	**<0.001**
STACIV [4]	0.07	-0.70 – 0.78	0.847
STUDIO	0.52 ***	0.45 – 0.59	**<0.001**
ETA	-0.02 ***	-0.03 – -0.01	**<0.001**
PERC	-1.25 ***	-1.70 – -0.82	**<0.001**
PERL	8.35 ***	7.81 – 8.91	**<0.001**
NPERL	-0.80 ***	-1.02 – -0.58	**<0.001**
NPERC	0.19	-0.05 – 0.43	0.117
AREA5	0.25 ***	0.16 – 0.34	**<0.001**
Observations	7165		
R^2 Tjur	0.813		

*p<0.05 **p<0.01 ***p<0.001

Although many results are similar and show no important difference between the gender of the individuals, there are some that deserve some focus.
Firstly the age variable (ETA) which was not a statistically significant regressor in the complete dataset, comes into significance when male and female are split. In particular the coefficients suggest that labour force participation likelihood slightly lowers as age increases in women while it increases in men. This led me to believe that an interaction

between gender and other variables may be in place and that is why I built the following interaction final model from the first complete one:

	LFP		
Predictors	*Log-Odds*	*CI*	*p*
(Intercept)	-3.79 ***	-4.51 – -3.07	**<0.001**
NCOMP	0.24 ***	0.16 – 0.32	**<0.001**
SEX [2]	-0.91 *	-1.72 – -0.09	**0.029**
STACIV [2]	0.12	-0.25 – 0.50	0.514
STACIV [3]	1.23 ***	0.67 – 1.77	**<0.001**
STACIV [4]	-1.05 *	-1.98 – -0.20	**0.022**
STUDIO	0.26 ***	0.20 – 0.32	**<0.001**
ETA	0.02 ***	0.01 – 0.03	**<0.001**
PERC	-3.27 ***	-3.65 – -2.90	**<0.001**
PERL	8.28 ***	7.96 – 8.62	**<0.001**
NPERL	-0.73 ***	-0.83 – -0.62	**<0.001**
AREA5	0.34 ***	0.26 – 0.41	**<0.001**
ACOM4C	-0.11 *	-0.19 – -0.02	**0.010**
SEX [2] × ETA	-0.04 ***	-0.05 – -0.03	**<0.001**
SEX [2] × STACIV [2]	0.60 *	0.08 – 1.13	**0.024**
SEX [2] × STACIV [3]	-0.12	-0.90 – 0.66	0.769
SEX [2] × STACIV [4]	1.07	-0.08 – 2.22	0.070
SEX [2] × STUDIO	0.26 ***	0.17 – 0.35	**<0.001**
SEX [2] × PERC	2.20 ***	1.77 – 2.64	**<0.001**
SEX [2] × AREA5	-0.09	-0.20 – 0.03	0.132
Observations	14446		
R^2 Tjur	0.802		

** p<0.05 ** p<0.01 *** p<0.001*

The interaction term between the female gender and the age demonstrates a negative coefficient, indicating that the effect of age on LFP leads to the statistically significant decrease of being part of the labour force among women as the age goes up.
On the other hand, women with higher qualifications and who are single (STACIV2) tend to participate more in the labour force.

This final one also shows the lowest AIC equal to 5378.8 whereas the one given by the first complete model was equal to 5546.2.