Question A

	Mean	Std Dev	Median	Maximum	Minimum	Skewness	Kurtosis
age	42,60236511	11,56349166	43	63	18	-0,18904	-0,93401
race	1,227827051	0,564797618	1	3	1	2,370998	4,252488
earnwke	848,4194322	586,292696	700	2884,61	0	1,388146	2,033279
employed	0,875461936	0,330216557	1	1	0	-2,27419	3,171928
unemployed	0,044161122	0,20547691	0	1	0	4,437404	17,69055
married	0,650591254	0,476821631	1	1	0	-0,6317	-1,60096
union	0,126940131	0,332935721	0	1	0	2,241233	3,023126
ne_states	0,210458234	0,407675594	0	1	0	1,420596	0,018094
so_states	0,292128593	0,454778165	0	1	0	0,914241	-1,16416
ce_states	0,249815226	0,432936519	0	1	0	1,155839	-0,66404
we_states	0,247597933	0,431646347	0	1	0	1,169565	-0,63212
government	0,17627494	0,38108483	0	1	0	1,699103	0,886953
private	0,705654085	0,455799669	1	1	0	-0,90249	-1,18551
self	0,118070953	0,32272321	0	1	0	2,367142	3,603362
educ_lths	0,067997046	0,251769394	0	1	0	3,432125	9,779481
educ_hs	0,275868446	0,446992695	0	1	0	1,002936	-0,99412
educ_somecol	0,196969703	0,397750944	0	1	0	1,523878	0,322206
educ_aa	0,109756097	0,3126176	0	1	0	2,496878	4,234398
educ_bac	0,220805615	0,414833099	0	1	0	1,346197	-0,18775
educ_adv	0,128603101	0,334794074	0	1	0	2,218884	2,923445
female	0,488359213	0,499902934	0	1	0	0,046576	-1,99783

Table 1. Table of descriptive statistics for the variables sample average, standard deviation, median, maximum and minimum value, skewness and kurtosis.

The variable age indicates how old the individuals in the sample are, with the oldest being 63 and the youngest 18 years old. The median age is 43, meaning half the sample is younger than 43, and half is older. The average age of individuals in the sample, however, is 42.6 years and the age standard deviation of 11.6, indicates the extent to which individual ages deviate from the mean. The higher the standard deviation, the more are ages spread out. The slightly negative skewness suggests that the age distribution is slightly left-skewed, meaning there are more individuals younger than the mean age than older ones. The negative kurtosis value indicates a relatively flat distribution with fewer extreme ages than a normal distribution.

Since the variable employed is binary (1 for employed, 0 for not), which translates to the maximum and minimum. The mean reflects the proportion of employed individuals. The average value of the employed variable is 0.875, suggesting that 87.5% of individuals in the sample were employed. The standard deviation of 0.33 indicates some variation in employment status, though a large majority are employed. The median is 1, meaning more than half of the sample is employed. This result is also supported by the negative skewness that indicates that the distribution is heavily skewed toward employment. The positive kurtosis suggests a distribution with sharper peaks than a normal distribution. This is expected due to the predominant outcome of 1 (being employed).

Question B

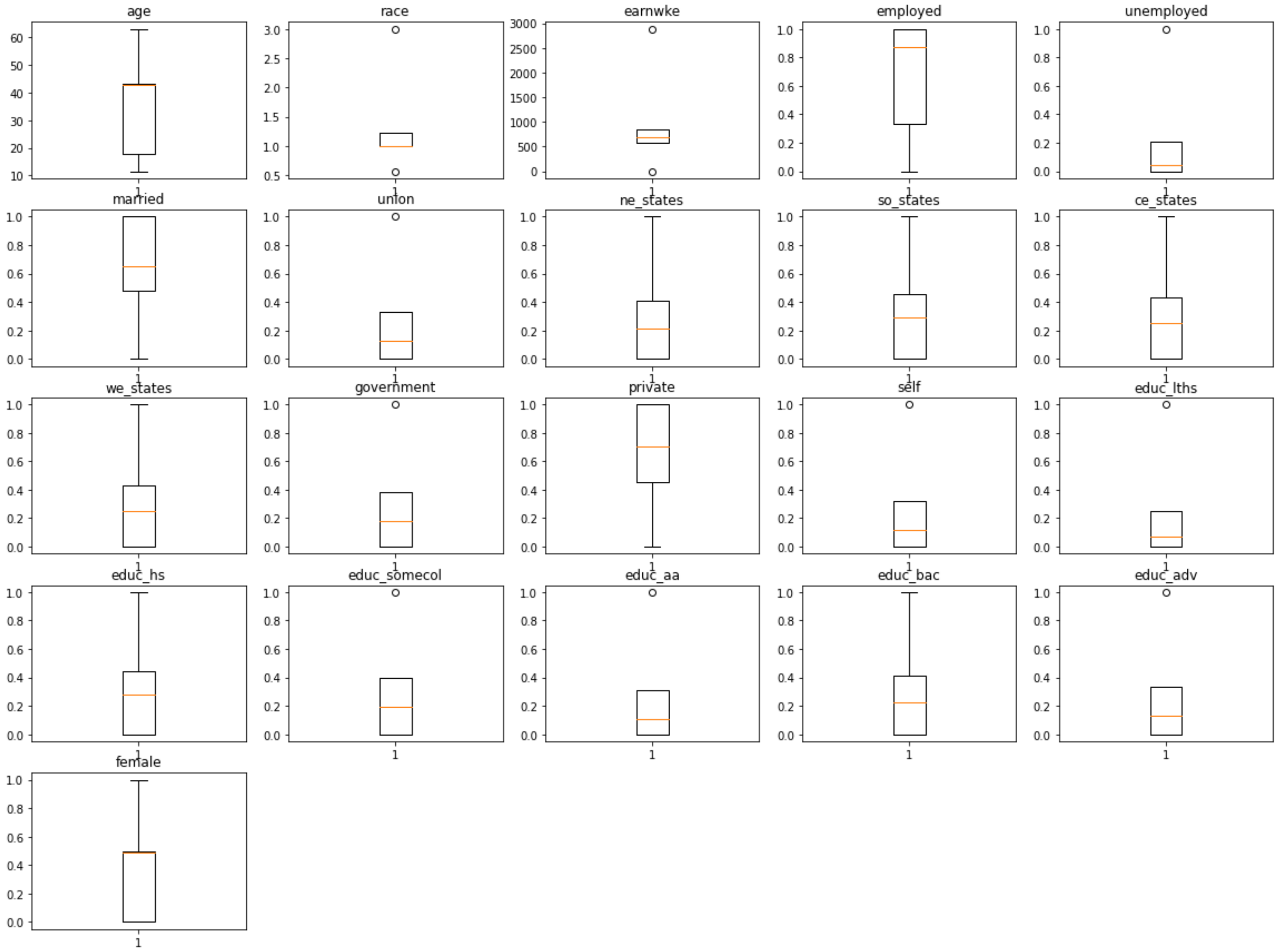

Figure 1. Boxplots for the variables.

Possible outliers can be seen as circles above or below the plotted boxes. This applies to the variables race, earnwke, unemployed, union, government, self, educ_lths, educ_somecol, educ_aa, and educ_adv. Outliers can indicate unusual values in these variables. They might represent either meaningful observations, like exceptionally high earnings or individuals with rare characteristics or data errors.

Question C

I knew from the description pdf file that 5421 workers were employed in April 2008. Then, I identified how many of these workers were still employed in April 2009 using Python, which results in 4738 employed workers. By dividing the 4738 employed workers in April 2009 by the 5421 workers that were in employed in April 2008, I obtain the fraction of workers in the sample that were employed in April 2009.

87.55% of the workers in the sample were employed in April 2009.

The 95% confidence interval for the probability that a worker was employed in April 2009, conditional on being employed in April 2008 is [0.8667, 0.8843].

Question D

```
                      OLS Regression Results
Dep. Variable:    ,employed        ,  R-squared:          ,   0.020
Model:            ,OLS             ,  Adj. R-squared:     ,   0.019
Method:           ,Least Squares   ,  F-statistic:        ,   54.22
Date:             ,Fri, 08 Nov 2024,  Prob (F-statistic):,4.83e-24
Time:             ,21:43:00        ,  Log-Likelihood:     , -1628.7
No. Observations:,  5412           ,  AIC:                ,   3263.
Df Residuals:     ,  5409          ,  BIC:                ,   3283.
Df Model:         ,     2          ,                      ,
Covariance Type: ,nonrobust        ,                      ,
           ,   coef   , std err ,     t    ,P>|t| ,  [0.025 ,  0.975]
const      ,    0.3075,    0.055,    5.619, 0.000,    0.200,    0.415
age        ,    0.0283,    0.003,   10.293, 0.000,    0.023,    0.034
age_squared,   -0.0003, 3.28e-05,   -9.971, 0.000,   -0.000,   -0.000
Omnibus:      ,2193.252,  Durbin-Watson:      ,   1.911
Prob(Omnibus):, 0.000  ,  Jarque-Bera (JB):   ,6593.720
Skew:         ,-2.215  ,  Prob(JB):           ,    0.00
Kurtosis:     , 6.100  ,  Cond. No.           ,2.68e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
[2] The condition number is large, 2.68e+04. This might indicate that
there are strong multicollinearity or other numerical problems.
```

Table 2. Regression table of regressing *Employed* on *Age* and *Age*^{2}, using a linear probability model.

Question D1

Since the p-value of the variable age is smaller than 0.05, age is a statistically significant determinant of employment in April 2009.

Question D2

Since the p-value of the variable age_squared is smaller than 0.05, there is evidence of a nonlinear effect of age on employment probability.

Question D3

The predicted employment probability for a 20-year-old is 0.7422841476413921, so 74.23%.

The predicted employment probability for a 40-year-old is 0.9157684560856655, so 91.58%.

The predicted employment probability for a 60-year-old is 0.8279458447195374, so 82.79%.

Question E

```
                    Probit Regression Results
Dep. Variable:  ,employed        ,  No. Observations:  ,  5412
Model:          ,Probit          ,  Df Residuals:      ,  5409
Method:         ,MLE             ,  Df Model:          ,     2
Date:           ,Fri, 08 Nov 2024,  Pseudo R-squ.:     ,0.02325
Time:           ,21:43:00        ,  Log-Likelihood:    , -1986.9
converged:      ,True            ,  LL-Null:           , -2034.2
Covariance Type:,nonrobust       ,  LLR p-value:       ,2.895e-21
           ,   coef   , std err ,    z    ,P>|z| ,  [0.025 ,  0.975]
const      ,   -1.2579,    0.246,   -5.121, 0.000,   -1.739,   -0.776
age        ,    0.1217,    0.013,    9.667, 0.000,    0.097,    0.146
age_squared,   -0.0014,    0.000,   -9.341, 0.000,   -0.002,   -0.001
```

Table 3. Regression table of regressing *Employed* on *Age* and *Age*^{2}, using a Probit model.

Question E1

Since the p-value of the variable age is smaller than 0.05, age is a statistically significant determinant of employment in April 2009.

Question E2

Since the p-value of the variable age_squared is smaller than 0.05, there is evidence of a nonlinear effect of age on employment probability.

Question E3

The predicted employment probability for a 20-year-old is 0.7295815090502125, so 72.96%.

The predicted employment probability for a 40-year-old is 0.9116617319682818, so 91.17%.

The predicted employment probability for a 60-year-old is 0.8316235459476327, so 83.16%.

Question F

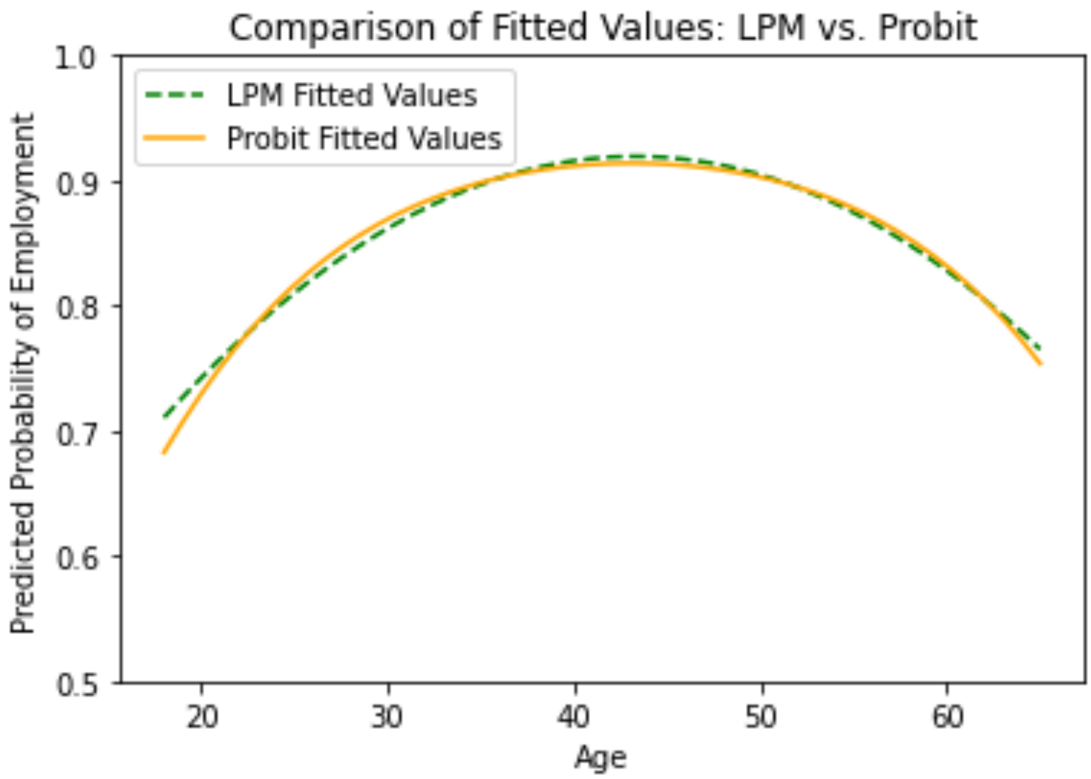

Figure 2. Plot of the fitted values using the Probit and the Linear Probability model.

There are no important differences in my answers to (D) and (E). The only minor difference, that can be observed in the graph as well as in my answers to (D3) and (E3), is that the probability of being employed is overall a little greater when using the Probit model compared to the Linear Probability model.

For both models the probability of employment increases up to the age of 40, after that it slowly starts to decrease. Nevertheless, the probability of employment for a 60-year-old is still greater than for a 20-year-old.

Question G1

	LPM	Logit	Probit
const	0.584 (0.000)	1.055 (nan)	0.654 (1.000)
ce_states	0.168 (0.000)	0.482 (1.000)	0.274 (1.000)
earnwke	0.000 (0.000)	0.001 (0.000)	0.000 (0.000)
educ_aa	0.129 (0.000)	0.456 (1.000)	0.261 (1.000)
educ_adv	0.114 (0.000)	0.297 (1.000)	0.179 (1.000)
educ_bac	0.105 (0.000)	0.195 (1.000)	0.129 (1.000)
educ_hs	0.097 (0.000)	0.140 (1.000)	0.091 (1.000)
educ_lths	0.031 (0.053)	-0.277 (1.000)	-0.151 (1.000)
educ_somecol	0.108 (0.000)	0.243 (1.000)	0.146 (1.000)
female	-0.000 (1.000)	0.007 (0.939)	-0.004 (0.928)
married	0.019 (0.059)	0.162 (0.077)	0.088 (0.073)
ne_states	0.141 (0.000)	0.212 (1.000)	0.141 (1.000)
race	-0.009 (0.299)	-0.073 (0.330)	-0.040 (0.329)
so_states	0.147 (0.000)	0.270 (1.000)	0.167 (1.000)
we_states	0.127 (0.000)	0.090 (1.000)	0.072 (1.000)

Table 2. Regression table comparing a Linear Probability, Logit, and Probit model. It shows the coefficients and p-values in parenthesis for every relevant variable. The p-value for the Logit model is indicated as "nan", but I could not identify the reason for that issue.

Hypothesis tests for LPM model:

For the variables educ_hs, educ_somecol, educ_aa, educ_bac, educ_adv, ne_states, so_states, ce_states, we_states, and earnwke, we reject the null hypothesis, meaning that every one of these variables is significant.

For the variables educ_lths, female, race, and married, we fail to reject the null hypothesis, meaning that none of these variables are significant.

Hypothesis tests for Logit model:

For the variable earnwke, we reject the null hypothesis, meaning that the variable is significant.

For the variables educ_lths, educ_hs, educ_somecol, educ_aa, educ_bac, educ_adv, female, race, married, ne_states, so_states, ce_states, and we_states, we fail to reject the null hypothesis, meaning that none of these variables are significant.

Hypothesis tests for Probit model:

For the variable earnwke, we reject the null hypothesis, meaning that the variable is significant.

For the variables educ_lths, educ_hs, educ_somecol, educ_aa, educ_bac, educ_adv, female, race, married, ne_states, so_states, ce_states, and we_states, we fail to reject the null hypothesis, meaning that none of these variables are significant.

Question G2

Characteristics of workers hurt most by the Great Recession in LPM model:

- Workers with higher educ_hs were less hurt by the Great Recession.
- Workers with higher educ_somecol were less hurt by the Great Recession.
- Workers with higher educ_aa were less hurt by the Great Recession.
- Workers with higher educ_bac were less hurt by the Great Recession.
- Workers with higher educ_adv were less hurt by the Great Recession.
- Workers with higher ne_states were less hurt by the Great Recession.
- Workers with higher so_states were less hurt by the Great Recession.
- Workers with higher ce_states were less hurt by the Great Recession.
- Workers with higher we_states were less hurt by the Great Recession.
- Workers with higher earnwke were less hurt by the Great Recession.

Characteristics of workers hurt most by the Great Recession in Logit model:

- Workers with higher earnwke were less hurt by the Great Recession.

Characteristics of workers hurt most by the Great Recession in Probit model:

- Workers with higher earnwke were less hurt by the Great Recession.

When using the Linear Probability model more variables, so characteristics, lead to workers being hurt by the Great Recession. The only similarity between the models is that workers with higher average weekly earnings are hurt by the Great Recession. However, none of these variables lead to workers being severely hurt by it.