Contents

Introduction:

In this analysis, we examined two datasets: EOTST2112023_GROW.csv and EOTST2112023_PUSA.csv. We found that the best model could be created using EOTST2112023_GROW.csv, thus our analysis focuses on a model created using this dataset.

Our datasets were taken from the Next Steps Study (NS), A study tracking young individuals in England from age 14 until age 25. The individuals in the study were surveyed periodically to collect a wide variety of data on them as they progressed through life.

The first data in the dataset were collected in 2004 when individuals in the survey were aged 14, the survey then surveyed individuals periodically, until the latest sweep, which was done when participants were aged 25.

Our research question for this report is to determine the variables which have significant predictive power for an individual's earnings at age 25. We view this question as important because if we know the factors which contribute towards an individual's earnings at 25, we can then make policy recommendations that would influence these factors, with the hope that changing these factors could lead to more people earning a good income and raising living standards.

The variable of interest in the regression is thus W8GROW (Gross Pay per Week at age 25). We identified several independent regressors which were statistically significant for predicting W8GROW. Among them were: The young person's level of education, occupation, sexual orientation, and socio-economic standing. This analysis' findings imply that enhancing young people's financial security, educational opportunities, and work prospects may improve their earnings.

The variable we investigated in the PUSA regression was W8PUSA (Partner's monthly income after tax). We identified several independent regressors which were statistically significant for predicting W8PUSA. Among them were: the young person's parent's employment status, socioeconomic standing, degree level and personal employment status at 25. Our analysis shows that the young person's personal attributes and choices between have significant predictive power for their eventual spouse's earnings.

Exploratory Analysis:

To start our analysis, after checking the summary of W8GROW, we found there were a lot of missing values. We removed all missing values from the dependent variable since the relevant independent variables might not be missing. They may correspond to some valid independent variables that, if not removed, would result in a significant bias in the model.

Then we created box plots with all the categorical predictors against our dependent variable for both data sets (W8GROW: Gross pay per week and W8PUSA: Partner's monthly income after tax). From our initial plots, it was clear that the upper extreme values were making it hard to predict which variables could affect our dependent variable.

Subsequently, we regressed all the predictors and the residual plot showed significant funnelling, leading us to use a logarithmic transformation. We added 1 to all the variables in our dependent variable to avoid taking the natural logarithm of zero. The transformation was well received, reducing the heteroskedasticity. This reduced our skewness from 26.54698 to -1.771938 and kurtosis from 928.2237 to 18.57653, creating a data set that resembles a normal distribution. The model's future predictive power increases with the variance being more constant.

Following the transformation, it became much clearer which predictors could affect our dependent variables. With that said, the majority of the categorical variables do not seem to affect weekly income; this is later shown when they were not significant at the 5% level (the baseline significance level used for all of our regressions) within our initial regression.

After we analysed the box plots of log(W8GROW), we found variables W1truantYP, W1alcever and W1bulrc share the same box plots for the values lower than -96, potentially leading to collinearity in the regression. When verified with a correlation matrix, these 3 variables ranged from 0.5 to 0.6 within each other, indicating a significant level of correlation. However, these 3 predictors were not significant in the initial regression too, so were disregarded before continuing.

Overall, it was unclear which predictors to eliminate at this stage without seeing their significance in the regression. The plots evaluated were not conclusive, therefore, we decided to proceed and include all predictors within our initial regression after the transformation.

After running the last regression, we removed all predictors that were not significant at the 5% level, leaving us with 8 predictors for W8GROW and 7 for W8PUSA. When evaluating the collinearity of the significant variables, we found W1hiqualdad and W1empsdad seemed that they could be dependent on each other, with the father's qualification level affecting their employment status. With the correlation coefficient being 0.94, there is a clear link between the two, potentially causing collinearity within the model. This was later verified through VIF. W1hiqualdad had a VIF value of 10.2 and W1empsdad had a VIF value of 8.6, leading to these variables being removed.

Since in W8GROW the definition of negative values in these significant variables is ambiguous, for example, in W1nssecfam, -91, -92, -94, and -99 all represent data not collected and account for a smaller percentage of the whole data, we then treat all the negative values of all significant variables as missing values. While treating the W8PUSA, we removed the missing values, which accounted for 10% of the overall values. We did not remove all negative values because the data set was too small and negative values accounted for a larger proportion. In W8GROW, W8DDEGP is the only exception to this rule, accounting for 30%. In W8PUSA we also kept W8DAGEYC, which accounted for 70% of all data, in addition to W8DDEGP.

We decided to merge a substantial number of the categories for our categorical predictors (appendix) to increase the significance of some of the subcategories, but also make it easier to interpret.

First to final model:

We present the regressions we ran to arrive at our final model in the table below, with * representing a 5% significance level, ** representing 1% and *** representing 0.1%.

Table 1 W8GROW

Coefficients	Model 1	Model 2	Model 3	Model 4	Model 5	Final Model
W8GROW						
W1hiqualdad	*	*				
W1disabYP	***					
W2disc1YP	*	*				
W1nssecfam		*	***	***	***	***
W1empsdad		*				
W6DebtattYP	*	*	***	***	***	***
W8DDEGP		***	***	***	***	***
W8DACTIVITYC	**	***	***	***	***	***
W6JobYP		**				
W8CMSEX		***	***	***	***	***
W8QMAFI		***	***	***	***	***
W6JobYP*W8DACTIVITYC				**	**	**
R2/Adj R2	0.06017/0.006335	0.2804/0.2392	0.1812/0.1787	0.1829/0.1801	0.1829/0.1803	0.1936/0.1909
P-value	0.1145	< 2.2e-16	< 2.2e-16	< 2.2e-16	<2.2e-16	< 2.2e-16
F-statistic	1.118	6.802	71.25	66.03	72.04	72.57
Std Error	732.9	0.5804	0.5842	0.5837	0.5837	0.3862

Table 2 W8PUSA

Coefficients	Model 1	Model 2	Model 3	Model 4	Model 5(Final Model)	Model 6
W8PUSA						
W1hiqualmum	*					
W1ethgrpYP	*					
W2disc1YP	**					
W8DGHQSC	*					
W1wrk1aMP		*	***	**	**	
W60wnchiDV		**	**	**	**	
W8DAGEYCH		*	***	***	***	***
W8DDEGP		*	***	***	***	***
W8DACTIVITYC	**	*	*		*	
W8TENURE	**	*	*			
W8DGHQSC		*				
W8DACTIVITYC* W8TENURE				*	*	
R2/Adj R2	0.1386/0.01644	0.2001/0.08665	0.06362/0.0564	0.07107/0.0627	0.06973/0.06196	0.07498/0.06488
P-value	0.1095	4.923e-09	< 2.2e-16	< 2.2e-16	<2.2e-16	< 2.2e-16
F-statistic	1.135	1.764	8.815	8.497	8.971	7.426
Std Error	10020	1.355	1.384	1.38	1.38	0.4259

In our initial model for W8GROW (Model 1), we ran a regression with all the predictors before the transformation, there were only six variables significant at the 5% level (the standard used for the rest of the regressions). After we applied the log transformation to

W8GROW, we reran the regression and found 10 significant predictors. Following the data clean mentioned in the exploratory analysis (missing values and merging), we ran a collinearity check, subsequently removing W1hiqualdad and W1empsdad because of their VIF values. This led to our third model with 8 significant predictors.

After examining Model 3 further, we decided to see if there was a potential interaction. W6jobYP and W8DACTIVITYC had the same sign of effect on W8GROW, and after drawing scatter plots and box plots to visualize the relationship, we discovered that plotting W8GROW against each variable separately does not fully capture the relationship between the variables, indicating a potential interaction. This ultimately led to the interaction between W6JobYP and W8DACTIVITYC being added. After incorporating the interaction (Model 4), we tested the model by ANOVA and concluded it was highly significant, while also producing a better fit (r^2 value shown in Table 1).

Using a stepwise method, we came to Model 5 with 6 statistically significant predictors out of the 7. W6JobYP is not significant at the 5% level, however, it is part of the interaction. For this reason, we decided to keep this predictor.

From the residuals vs fitted, Cook's distance and residuals vs leverage plot, we can see there are outliers. From our evaluation of these points, we identified 215 outliers and finally reached a data set with 3338 observations. Our Final Model has an increased r^2 value and a reduced standard error, indicating that the removal of outliers has produced a better fit.

We then evaluate the precision of the models with and without the outliers (Final Model) by using the difference between the in-sample and the out-sample mse in the two models and found that both results were quite small. Thereby, we introduced the 10-fold cross validation including RMSE, R-squared and MAE. From the results, we found that for the Final Model, the RMSE value is 0.3866316, the R-squared value is 0.1909592, and the MAE value is 0.3014485; for Model 5, the RMSE value is 0.5807969, the R-squared value is 0.1819642, and the MAE value is 0.3818511. Finally, we could concluded that Final Model is better than Model 5. (Results can be seen by Table 3).

For the W8PUSA data set (Table 2), we applied the same initial approach to arrive at Model 3. We included an interaction between W8DACTIVITYC and W8TENURE, which proved to be significant by ANOVA. However, when we removed the outliers to create Model 6, we found the variables became less significant. This may be caused by the small sample sizes. We decided to include the outliers, with our Final Model being Model 5.

Table 3 10-fold cross-validation

	Model 5	Final Model
RMSE	0.5807969	0.3866316
R-squared	0.1819642	0.1909592
MAE	0.3818511	0.3014485

Results:

The tables below show the results for the relevant regressions for W8GROW and W8PUSA:

Table 4 W8GROW

	Final Model	P-values
Intercept	6.047642	<2e-16
(W1nssecfam)other_employ	-0.079197	1.34e-08
(W1nssecfam)W1unemploy	-0.190832	2.04e-07
W6DebtattYP	0.010368	1.80e-06
(W8DDEGP)first or high	0.149523	<2e-16
(W8DACTIVITYC)unemployed	-0.480526	4.51e-14
(W6JobYP)no	0.010330	0.46484
(W8CMSEX)female	-0.127902	<2e-16
(W8QMAFI)Doing all right	-0.149911	<2e-16
(W8QMAFI)Other	-0.327793	<2e-16
(W8DACTIVITYC)unemployee* (W6JobYP)no	0.310674	0.00177

Table 5 W8PUSA

	Final Model	P-values
Intercept	6.12099	<2e-16
(W1wrk1aMP)full_time_employee	0.30493	0.009285
(W1wrk1aMP)part_time_employee	0.29629	0.012879
(W1wrk1aMP)look_after_family	0.20193	0.141290
(W6OwnchiDV)with_child	0.80353	0.037822
(W6OwnchiDV)without_child	1.03752	0.002191
(W8DAGEYCH)with_child	-0.33752	0.000137
(W8DDEGP)first or high	0.44028	6.91e-07
(W8TENURE)own_with_loan	0.02695	0.807524
(W8TENURE)rent_inc_housing	-0.10805	0.327116
(W8DACTIVITYC)unemployee	0.445294	0.049860
(W8TENURE)rent_inc_housing* (W8DACTIVITY)unemployee	-0.54788	0.035611

Table 4 investigates the association between W8GROW, the study respondents' income level at age 25, and many other characteristics. The results show a significant relationship between W8GROW and W8DDEGP. Individuals with a first-class degree or higher at the age

of 25 have a higher wage. Gender has a substantial influence as well, with female salaries being roughly 12% lower than male wages holding all else constant, as seen in the table. Furthermore, there are strong correlations between W8DACTIVITYC (employment status at age 25), W8QMAFI (asset management at age 25) and income level. The Family's NS-SEC class at 14 years old also affects income, which shows that children are likely to have higher earnings if their parents are management executives.

The employment status at age 17(W6Job) did not reach a significance level of 5%, but there is a strong interaction between employment status at age 17 and age 25, implying that being jobless at age 17 may have an influence on employment status at age 25. Individuals who were unemployed at the age of 17 are more likely to be unemployed at the age of 25 according to the model and affected the level of income.

Table 5 depicts the link between a spouse's income and other characteristics at the age of 25. It can be observed that, like W8GROW, W8PUSA, and W8DDEGP are highly associated, demonstrating a probable positive association between a first-class degree and future earnings for oneself and one's spouse, with an effective value of 0.44 (a 44% increase holding all else constant). This table also contains two variables relating to children (W6OwnchiDV and W8DAGEYCH). It can be shown that having children at age 17 has a positive relationship with the spouse's income level at age 25; however, having children at age 25 is inversely associated with the spouse's wage level.

W8PUSA, like W8GROW, is connected to parental employment (W1wrk1aMP), and the table illustrates that W8PUSA is greater when the parent is a full-time employee. Unlike W8GROW, the variable W8DACTIVITYC demonstrates that the spouse's wage is higher if the study subject is unemployed, with an impact value of around 0.45 (45% holding all else constant).

W8TENURE (i.e. housing status at age 25) does not reach a significant level of 5%, and there is an interaction between W8TENURE(housing status at age 25) and W8DACTIVITYC (employment status at age 25), that has an effect on spouse earnings. We discover that spouses of the subject who are jobless at age 25 and own a home earn more, but those who are unemployed but rent earn less, which possibly affected the results.

Commentary on Data/Analysis:

There are several potential drawbacks to our W8GROW analysis due to limitations in the data available. For example, there were a significant number of missing values in the regression which we removed. We would expect the removal of these missing values to cause some degree of bias in the regression coefficients which may lead to an under or over-estimation of the effects of certain variables on income growth.

The final regression model also has a low R2 value of 0.1936. This means that although the model provides insights into certain associations which are very important for predicting an individual's earnings at 25, the data do not allow the prediction of earnings with a high degree of accuracy.

There were also potentially some issues with extreme outliers in the earnings data. As noted in exploratory analysis, certain individuals had extremely high earnings relative to others, since linear regression uses the smallest sum of squared errors possible, it is highly sensitive to outliers and thus these few data points may hinder our analysis with regards to the general population or with regards to attempting to predict outcomes for the "median individual". Based on this we removed the outliers from the GROW data.

In the PUSA dataset, we tried to remove 365 outliers but found that almost all the predictors lost their significance, indicating that these outliers had a significant impact on the model. However, since the database is small and the outliers account for about 23% of the total data, it would not have made sense to exclude these influential points.

W8DDEGP, which is shared between datasets W8GROW and W8PUSA plays a crucial role in the regression model's analysis and prediction, displaying a high level of significance in both datasets. However, the data in W8DDEGP has a missing value percentage of over 30%, which may have impacted the results to some extent. Unfortunately, we cannot analyse this missing value further as we do not have enough information about it.

Furthermore, there is the potential issue of confounders, which are not controlled for, meaning that we should be careful when using the results to recommend policy decisions, as the coefficients in the regression may simply represent associations, rather than causal links.

A potential confounder in our regression is work ethic. Those with better work ethics are more likely to obtain a first degree, as harder work generally leads to better academic outcomes. A better work ethic should also cause individuals to earn more, as people who work harder are often rewarded with promotions, raises, and bonuses.

Thus, the potential for confounders like work ethic in our regression is a limitation in the data (since we do not have access to data on all possible confounders), which may prevent our regression from having a causal interpretation.

Interpretations and Conclusions for a lay audience:

We analysed a large data set with the intention to determine which information is important for determining an individual's earnings at age 25. For our analysis, we used data from the Next Steps Study, a study tracking various characteristics of young individuals over time. The study data begins when individuals are aged 14 and the most recent data was taken when the individuals were aged 25.

We found several variables which were shown to be accurate predictors of an individual's future earnings. Amongst these were the individuals' type of employment, their attitude to debt, whether they got a first degree/higher or not, whether they were doing paid work at age 17, whether they are unemployed at 25, their sex, and the level of financial security which they feel.

An interesting finding in the data was that having a first degree is associated with a reasonable increase in income. Having a first degree was associated with an increase in income of roughly 12% compared to no degree. The analysis also showed evidence for a difference in earnings between men and women, with women having expected earnings about 13% less than their male counterparts.

As an example of what the model we created predicts: A male with a first degree would be expected to earn logged earnings of about 6.2, whereas if it were a female with a first degree, expected logged earnings would be about 6.05 (Making certain basic assumptions about the individuals, which are kept the same between the two). Logged earnings are simply a more technical way of expressing earnings, using a mathematical technique, we can then convert these logged earnings to normal earnings. In this case, the model predicts a weekly income for the male of roughly 493 Pounds and an expected weekly income for the female of roughly 424 Pounds.

In our analysis, we also looked at our ability to predict an individual's spouse's earnings. Ultimately, our model was able to explain less than ten percent of someone's spouse's earnings based on the data we used. 10 Percent is an exceedingly small amount, meaning that a significant portion of what determines someone's spouse's earnings is either up to random chance or explained by other data.

We found that the strongest predictors of how much a person's spouse earns in the data are: Information about the individuals working status, whether the respondents have children, whether the individual achieved a first or high degree, the individual's method of financing their home, and their employment status.

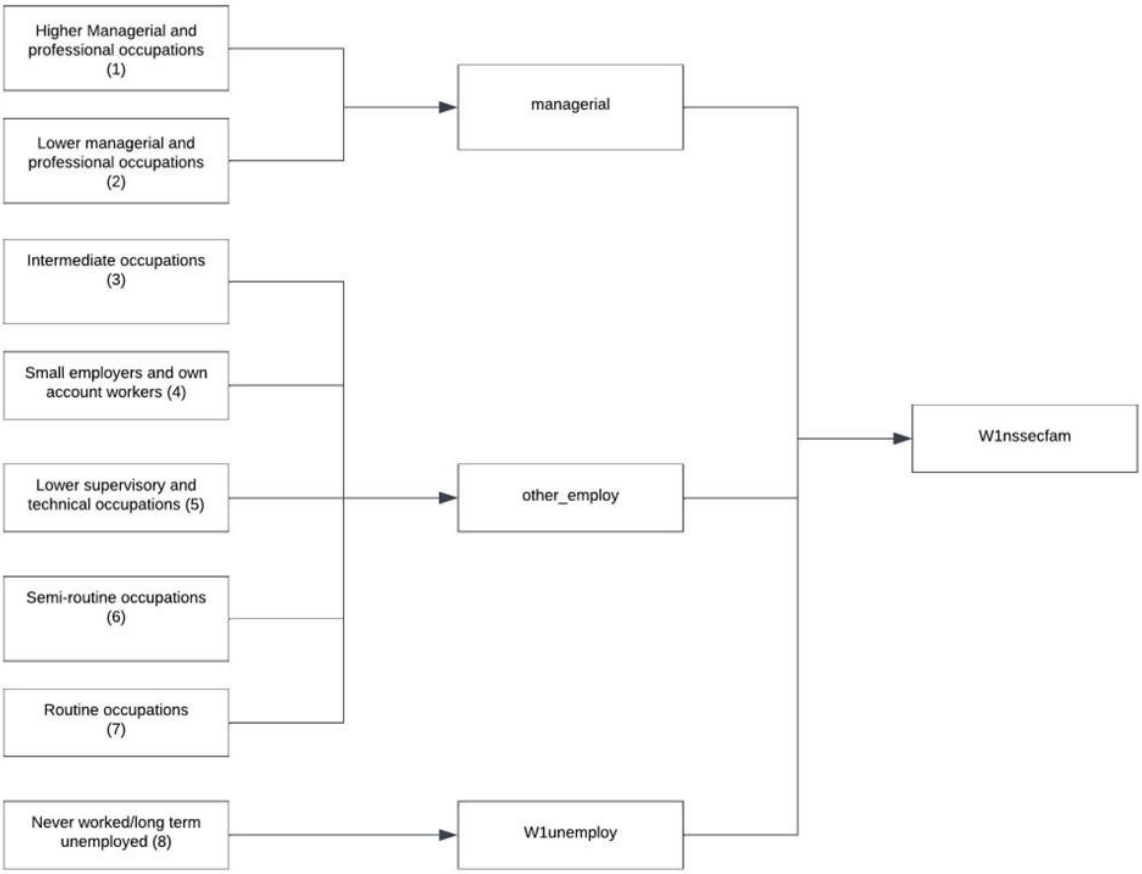
These variables seemed to indicate that people tend to have spouses of a roughly equivalent economic background to themselves.

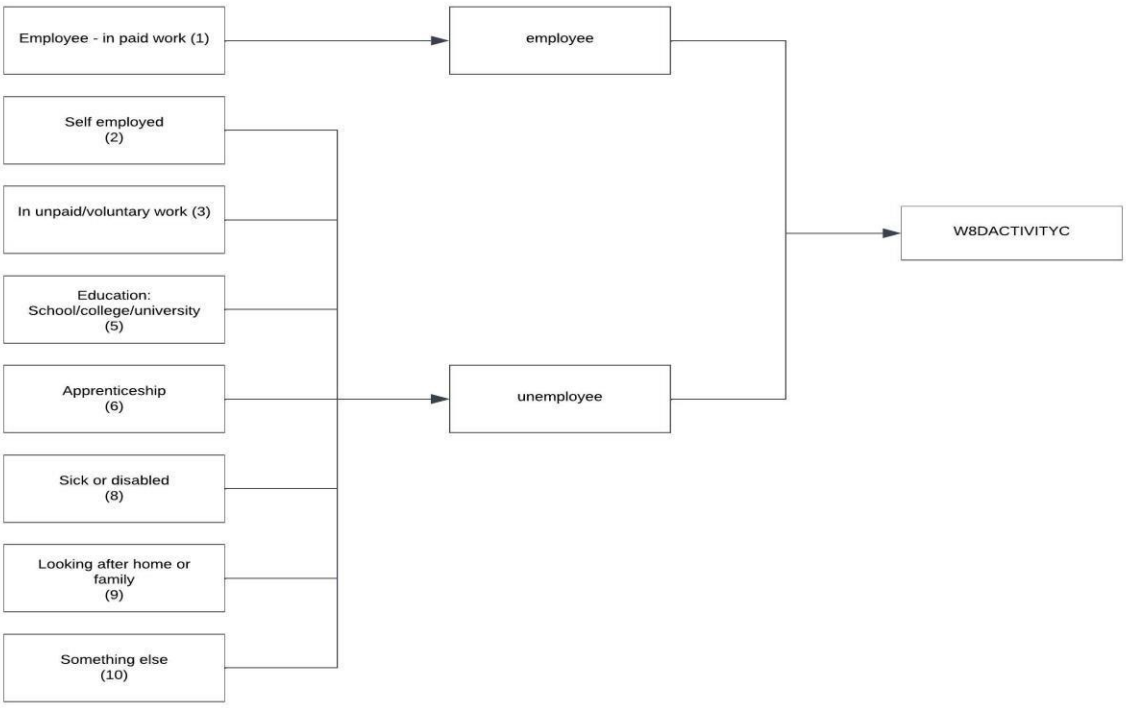
Despite some interesting findings in the data, both our models are flawed in some ways due to limitations in the data we used, so we would hesitate to make policy recommendations based on them. If we had to give some recommendations, we would say that the government should encourage more people to attain first degrees, due to the associated increase in income. Furthermore, the fact that women are earning less than men may be

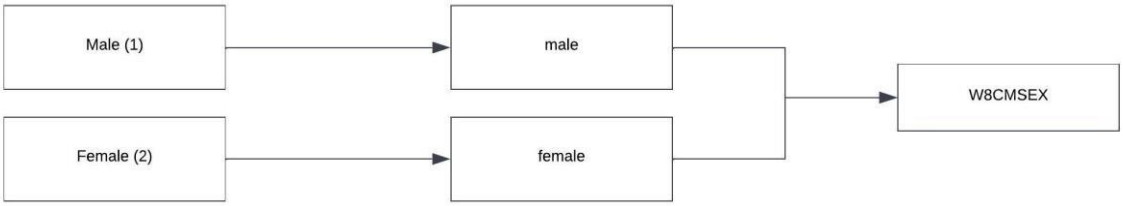
evidence of sexism in the workplace. Thus, we would encourage the government to have schools implement educational programs to teach boys and girls the correct way to interact with one another in the workplace at a young age.

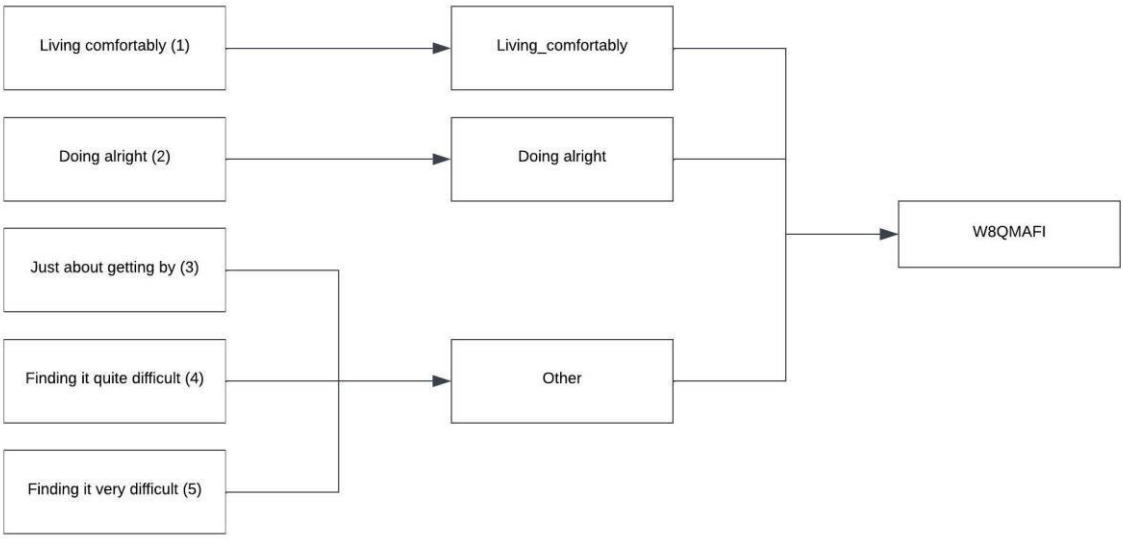
Appendix: The variable merging we did is displayed below

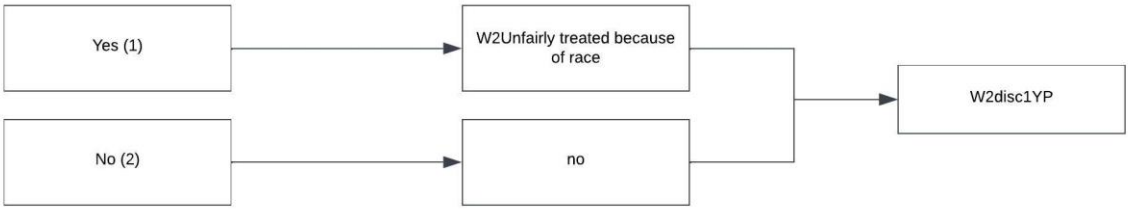
W8GROW

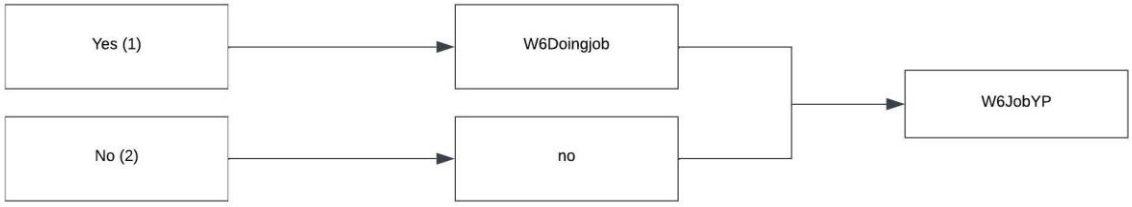


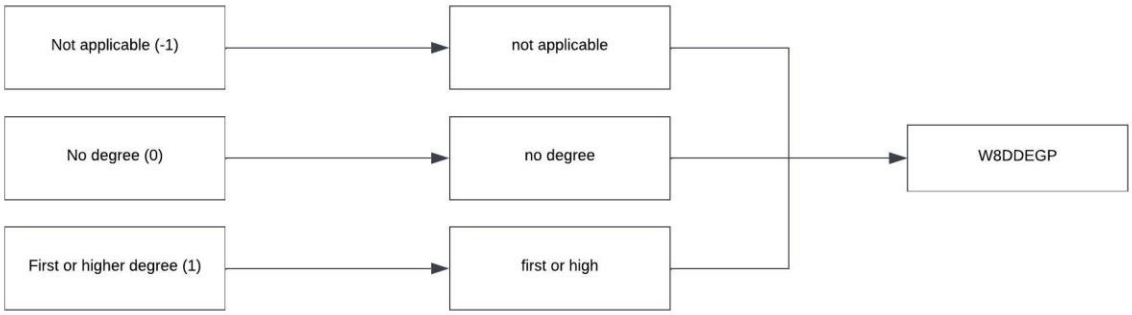



Male (1)
male
Female (2)
female
W8CMSEX

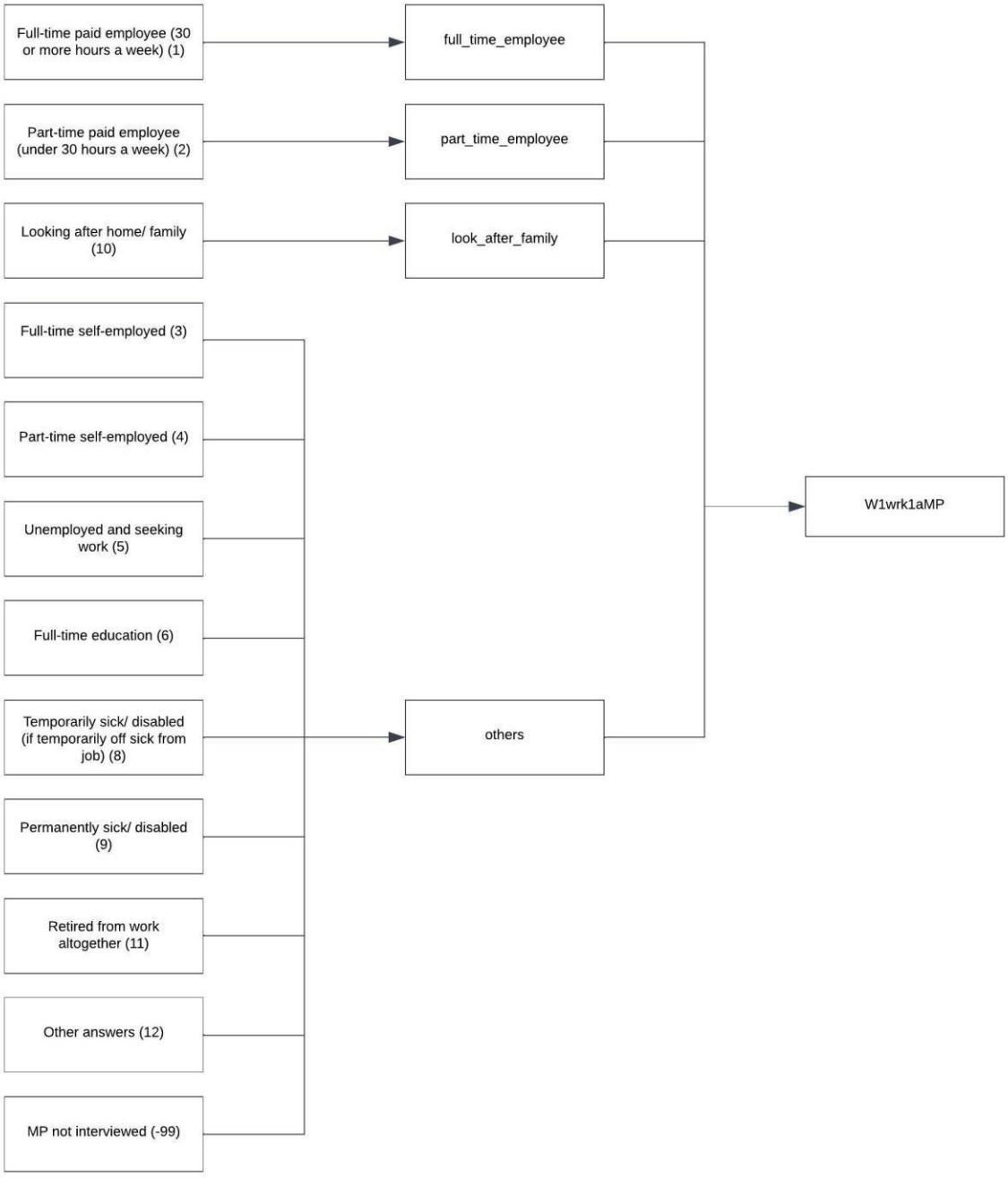

Living comfortably (1)
Living_comfortably
Doing alright (2)
Doing alright
Just about getting by (3)
Finding it quite difficult (4)
Other
Finding it very difficult (5)
W8QMAFI

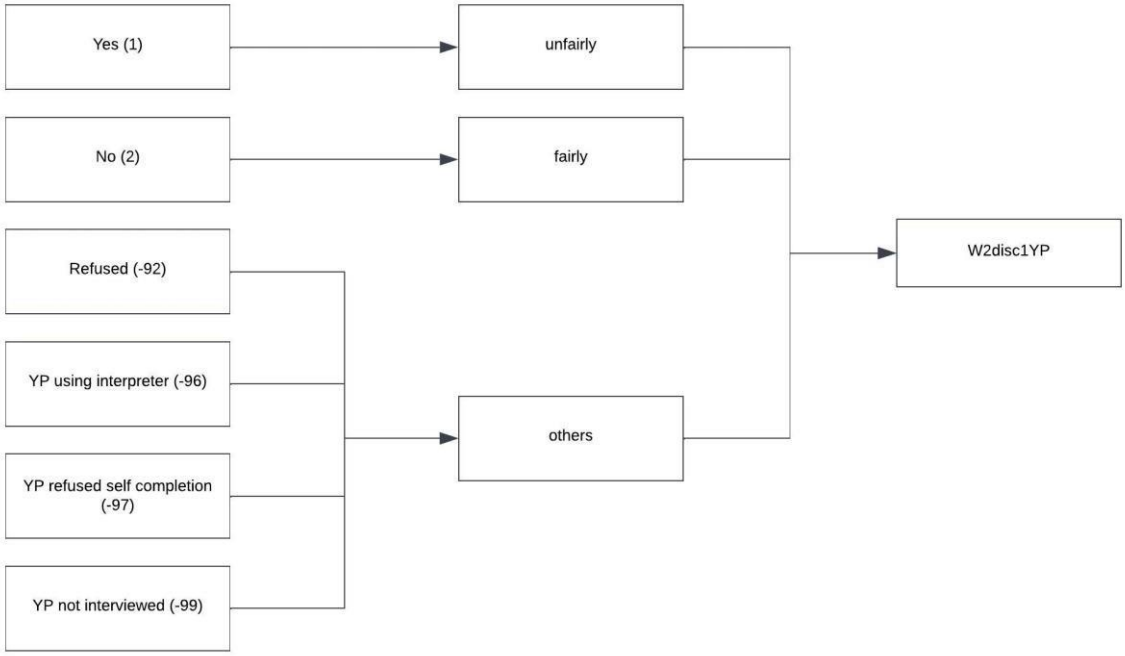

Yes (1)
W2Unfairly treated because of race
No (2)
no
W2disc1YP

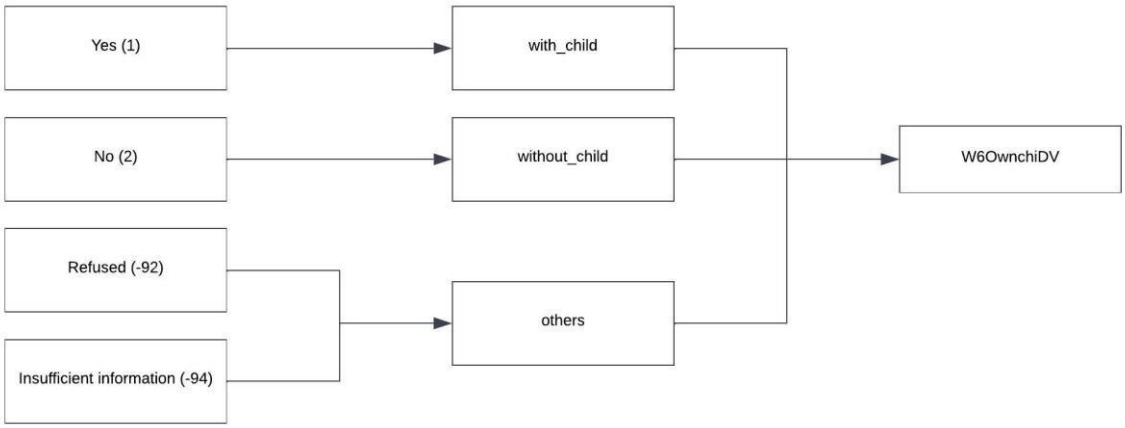

Yes (1)
No (2)
W6Doingjob
no
W6JobYP

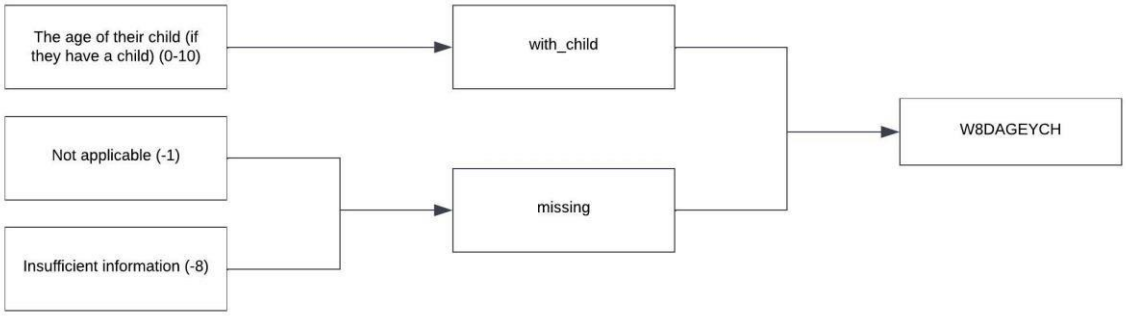

Not applicable (-1)
No degree (0)
First or higher degree (1)
not applicable
no degree
first or high
W8DDEGP

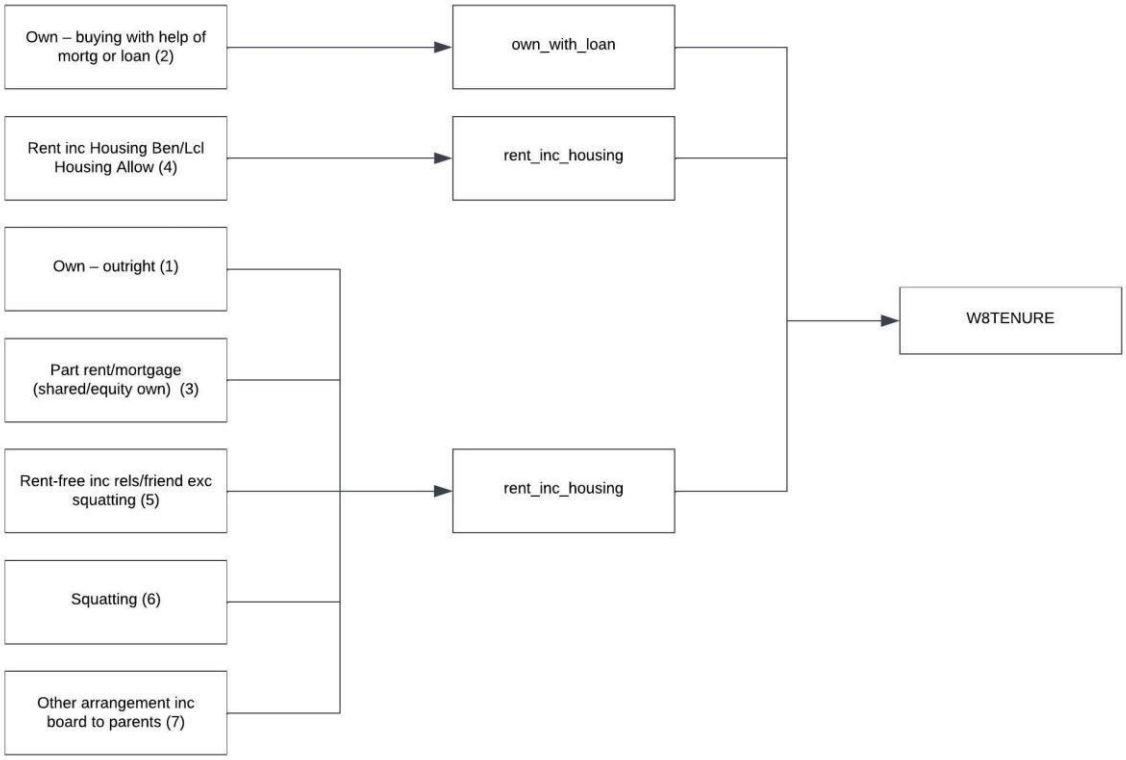
W8PUSA

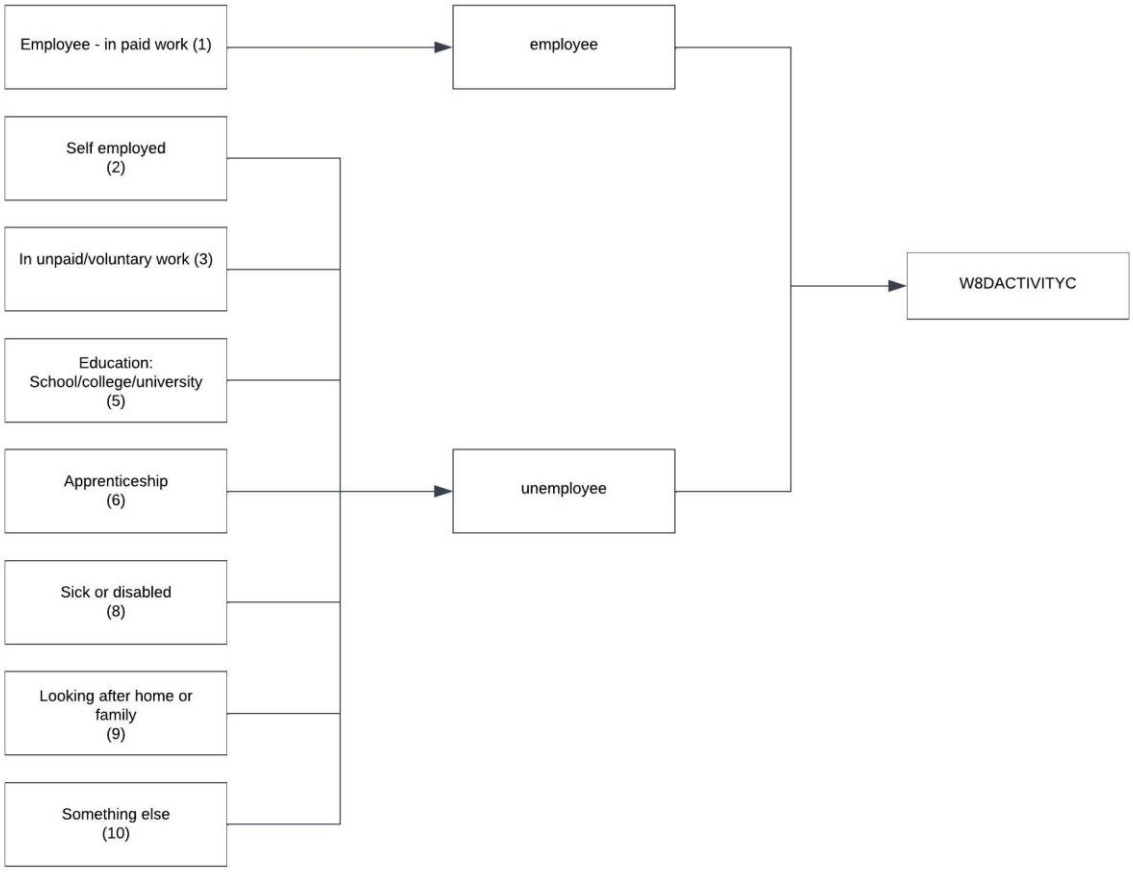

Full-time paid employee (30 or more hours a week) (1)
Part-time paid employee (under 30 hours a week) (2)
Looking after home/ family (10)
Full-time self-employed (3)
Part-time self-employed (4)
Unemployed and seeking work (5)
Full-time education (6)
Temporarily sick/ disabled (if temporarily off sick from job) (8)
Permanently sick/ disabled (9)
Retired from work altogether (11)
Other answers (12)
MP not interviewed (-99)
full_time_employee
part_time_employee
look_after_family
others
W1wrk1aMP


Yes (1)
No (2)
Refused (-92)
YP using interpreter (-96)
YP refused self completion (-97)
YP not interviewed (-99)
unfairly
fairly
others
W2disc1YP


Yes (1)
No (2)
Refused (-92)
Insufficient information (-94)
with_child
without_child
others
W6OwnchiDV


The age of their child (if they have a child) (0-10)
Not applicable (-1)
Insufficient information (-8)
with_child
missing
W8DAGEYCH


Own – buying with help of mortg or loan (2)
Rent inc Housing Ben/Lcl Housing Allow (4)
Own – outright (1)
Part rent/mortgage (shared/equity own) (3)
Rent-free inc rels/friend exc squatting (5)
Squatting (6)
Other arrangement inc board to parents (7)
own_with_loan
rent_inc_housing
rent_inc_housing
W8TENURE


Employee - in paid work (1)
employee
Self employed (2)
In unpaid/voluntary work (3)
Education: School/college/university (5)
Apprenticeship (6)
unemployee
Sick or disabled (8)
Looking after home or family (9)
Something else (10)
W8DACTIVITYC