COP529 Data Mining Assignment by Keunwoo Kim(F429147)

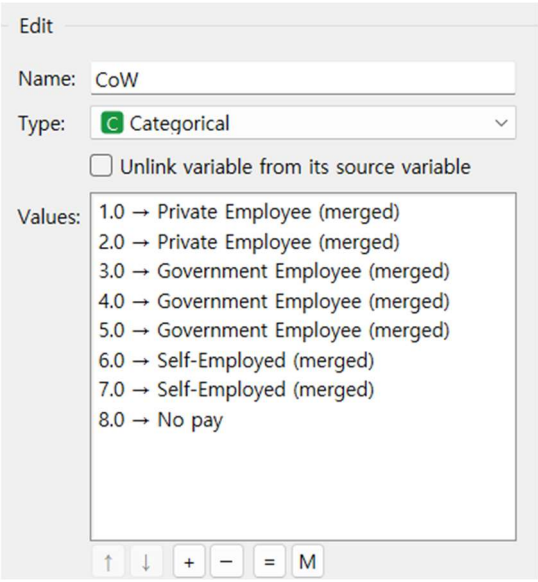
Part 1.

To transform the census data into a more meaningful form, I have created Cleaned_census_data.csv, as presented in Figure 0. The Cleaned_census_data.csv includes the transform of nine columns and added Industry column, which explains the industry type of the occupation column numeric codes.

Before performing exploratory data analysis to understand the dataset's structure and key characteristics, I would like to briefly explain the data transformation process.

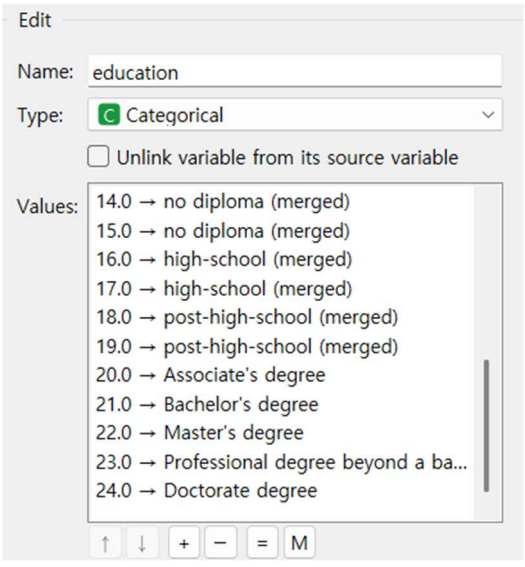
The CoW numeric code has been categorised using the Edit Domain widget (Figure 1).

Figure 1.



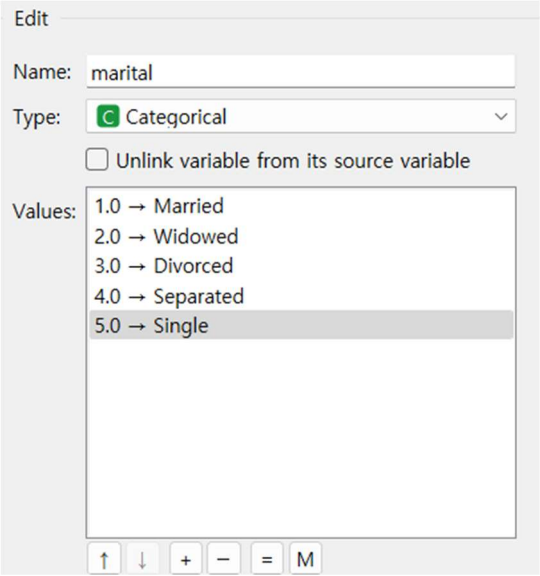
The Edit Domain widget categorises the education numeric code (Figure 2).

Figure 2.



The marital numeric codes followed the same process (Figure 3).

Figure 3.



Due to its massive variable, the occupation data was first merged as Left Join with the Attribute_Values data file.

Figure 4. Selected the Occupation column

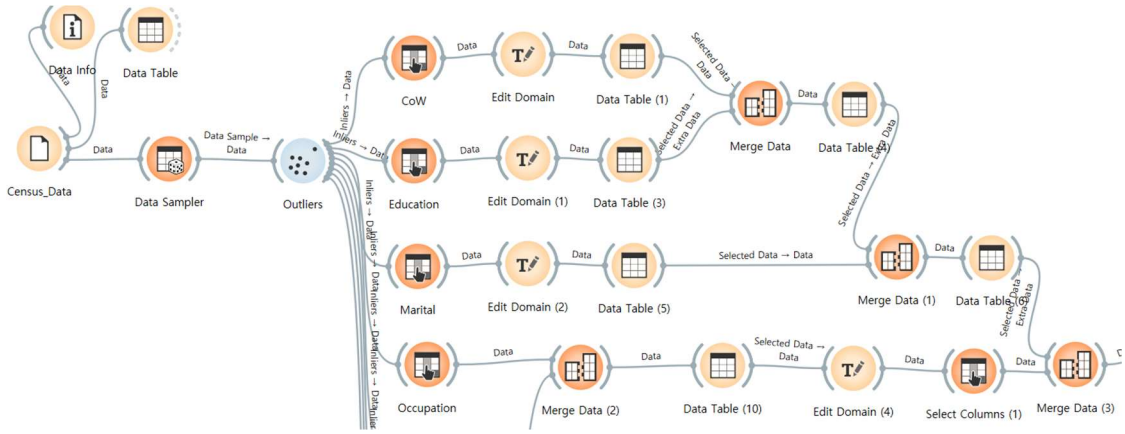
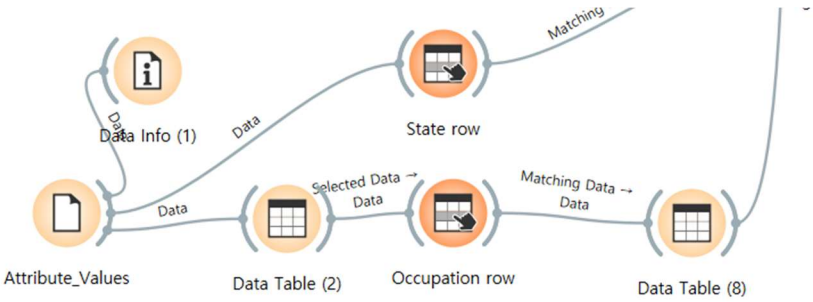


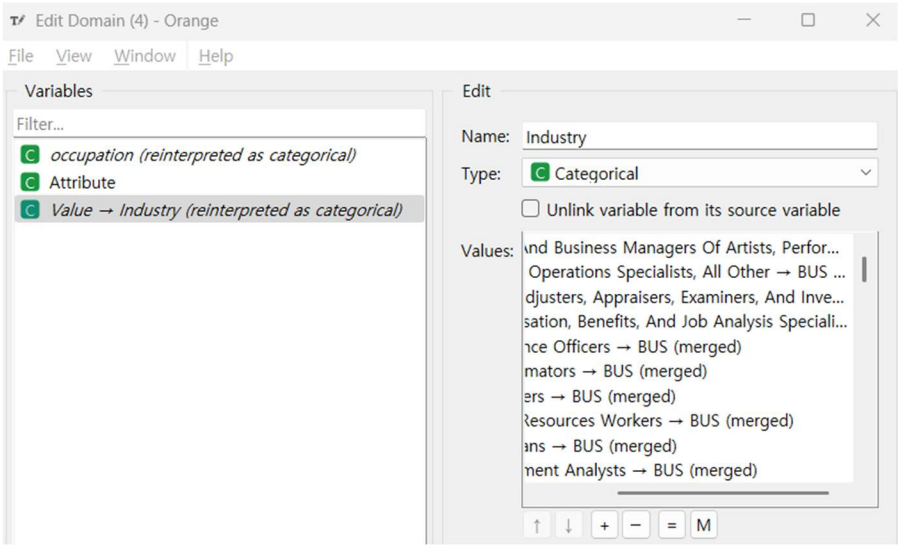
Figure 5. Selected Occupation row



The Occupation column and row from each data file have been merged (Figures 4 and 5).

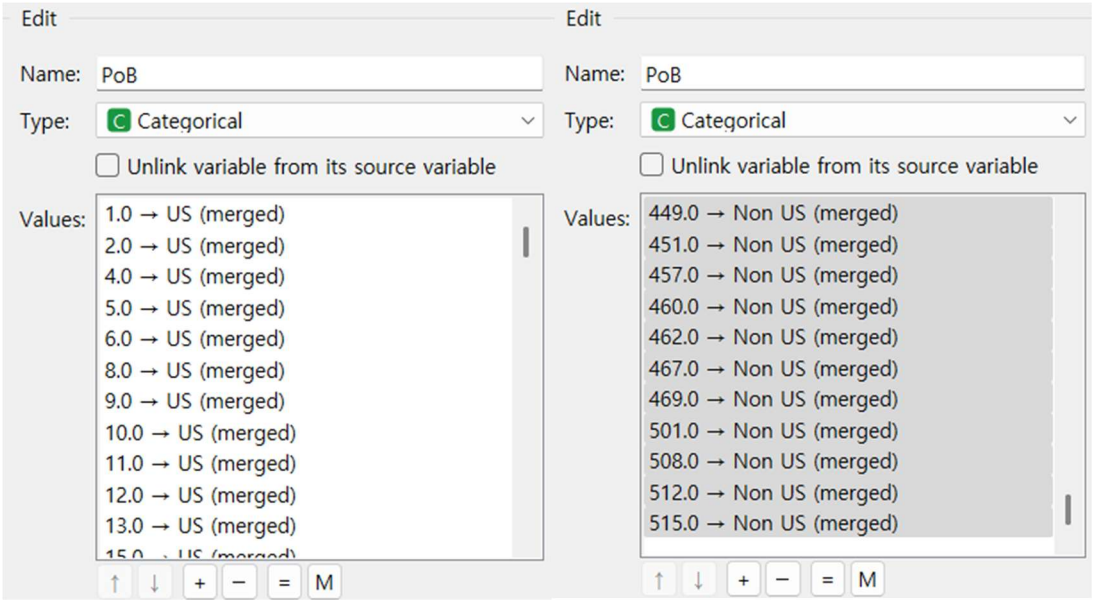
Afterwards, using the Edit Domain widget, the Value column was reinterpreted as categorical data, and its value was based on the first three characters (Figure 6) and displayed as a pair of numeric codes referred to earlier.

Figure 6.



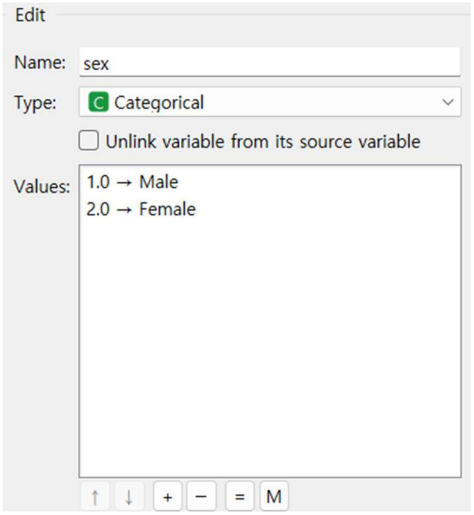
The PoB column followed a transformation process similar to the Marital column by categorising it into US and non-US. (Figure 7).

Figure 7.



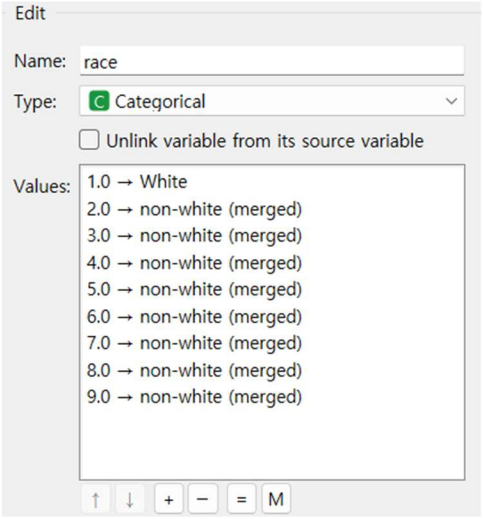
The sex column has been categorised into Male and Female (Figure 8).

Figure 8.



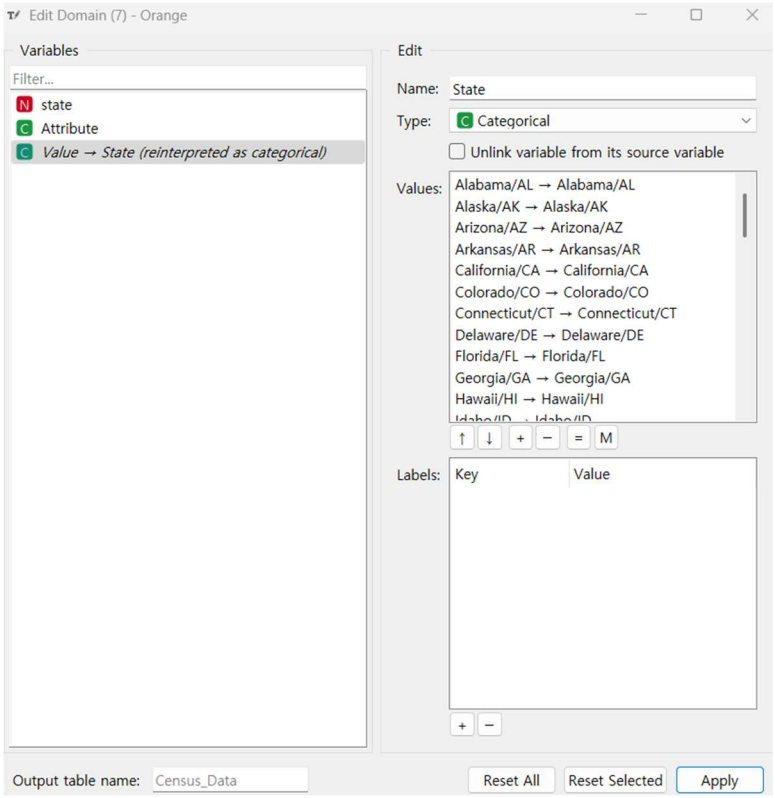
The Race column has been categorised into White and non-White (Figure 9).

Figure 9.



The State column was merged as Left Join with the Attribute_Values data file due to its massive variable, like the Occupation column (Figure 5). Replaced its numeric code with an actual state name (Figure 10).

Figure 10.



As is visible in Figures 11 and 12, each transformed Column has been merged and composed of the Cleaned_census_data.csv (Figure 13).

Figure 11. The whole process of data transformation

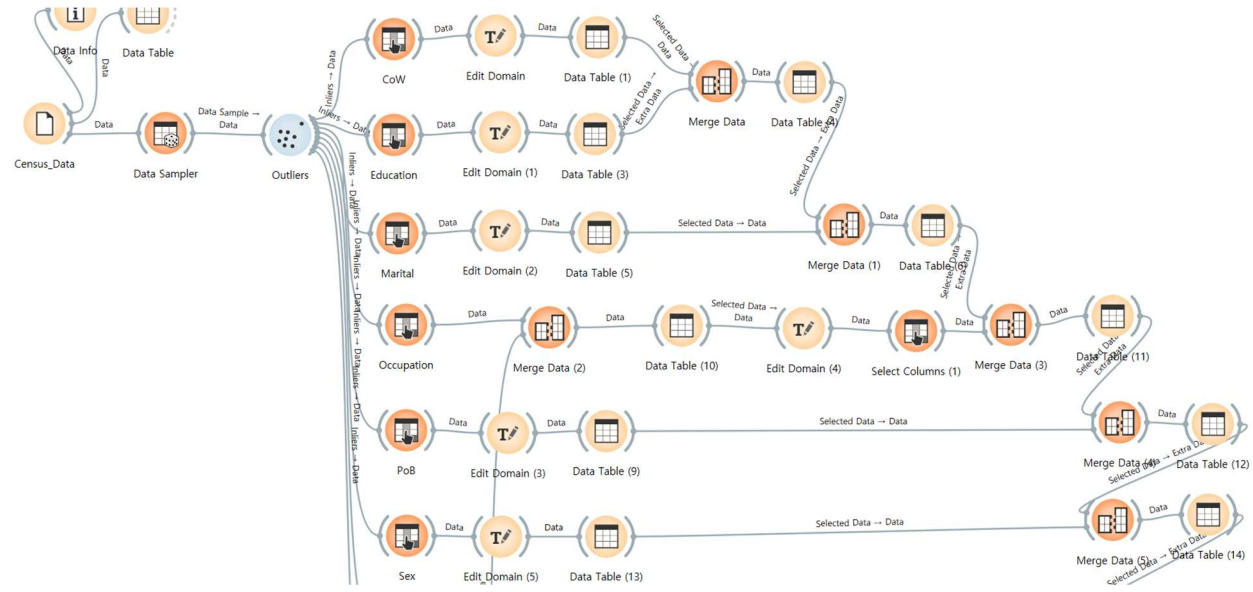


Figure 12. The whole process of data transformation (Continued image of Figure 11)

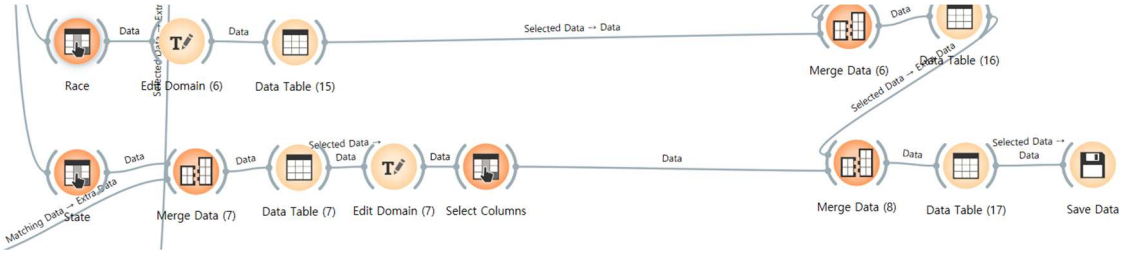
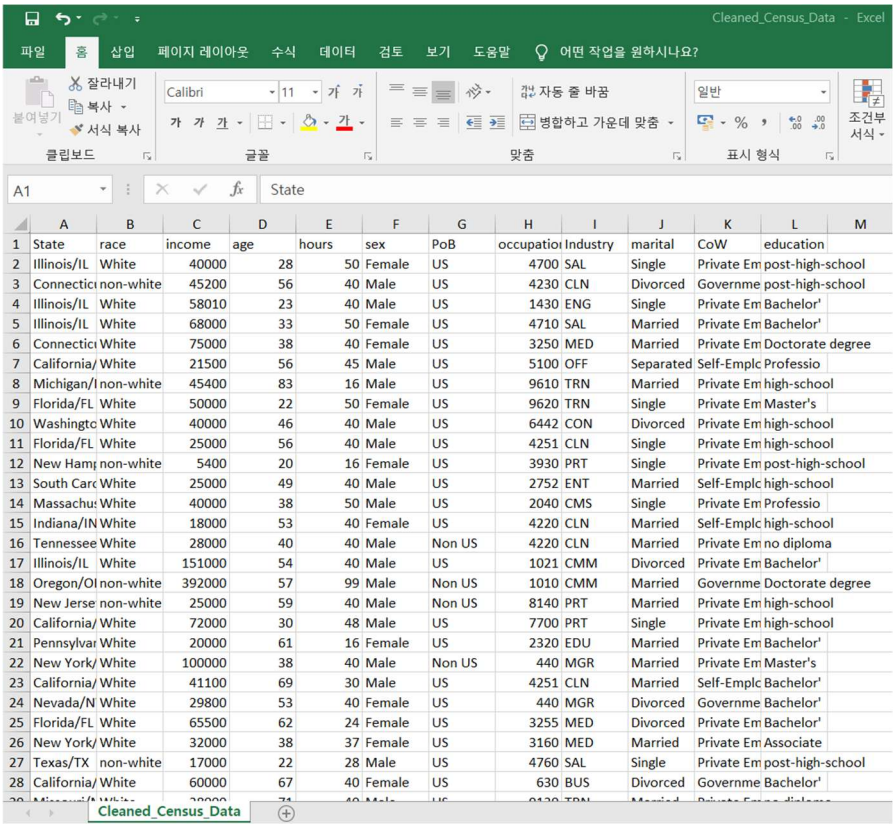


Figure 13. Cleaned_census_data.csv



Cleaned_Census_Data - Excel

State

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	State	race	income	age	hours	sex	PoB	occupatio	Industry	marital	CoW	education	
2	Illinois/IL	White	40000	28	50	Female	US	4700	SAL	Single	Private Em	post-high-school	
3	Connectici	non-white	45200	56	40	Male	US	4230	CLN	Divorced	Governme	post-high-school	
4	Illinois/IL	White	58010	23	40	Male	US	1430	ENG	Single	Private Em	Bachelor'	
5	Illinois/IL	White	68000	33	50	Female	US	4710	SAL	Married	Private Em	Bachelor'	
6	Connectici	White	75000	38	40	Female	US	3250	MED	Married	Private Em	Doctorate degree	
7	California/	White	21500	56	45	Male	US	5100	OFF	Separated	Self-Emplc	Professio	
8	Michigan/I	non-white	45400	83	16	Male	US	9610	TRN	Married	Private Em	high-school	
9	Florida/FL	White	50000	22	50	Female	US	9620	TRN	Single	Private Em	Master's	
10	Washingtc	White	40000	46	40	Male	US	6442	CON	Divorced	Private Em	high-school	
11	Florida/FL	White	25000	56	40	Male	US	4251	CLN	Single	Private Em	high-school	
12	New Hamp	non-white	5400	20	16	Female	US	3930	PRT	Single	Private Em	post-high-school	
13	South Carc	White	25000	49	40	Male	US	2752	ENT	Married	Self-Emplc	high-school	
14	Massachus	White	40000	38	50	Male	US	2040	CMS	Single	Private Em	Professio	
15	Indiana/IN	White	18000	53	40	Female	US	4220	CLN	Married	Self-Emplc	high-school	
16	Tennessee	White	28000	40	40	Male	Non US	4220	CLN	Married	Private Em	no diploma	
17	Illinois/IL	White	151000	54	40	Male	US	1021	CMM	Divorced	Private Em	Bachelor'	
18	Oregon/OI	non-white	392000	57	99	Male	Non US	1010	CMM	Married	Governme	Doctorate degree	
19	New Jerse	non-white	25000	59	40	Male	Non US	8140	PRT	Married	Private Em	high-school	
20	California/	White	72000	30	48	Male	US	7700	PRT	Single	Private Em	high-school	
21	Pennsylvar	White	20000	61	16	Female	US	2320	EDU	Married	Private Em	Bachelor'	
22	New York/	White	100000	38	40	Male	Non US	440	MGR	Married	Private Em	Master's	
23	California/	White	41100	69	30	Male	US	4251	CLN	Married	Self-Emplc	Bachelor'	
24	Nevada/N	White	29800	53	40	Female	US	440	MGR	Divorced	Governme	Bachelor'	
25	Florida/FL	White	65500	62	24	Female	US	3255	MED	Divorced	Private Em	Bachelor'	
26	New York/	White	32000	38	37	Female	US	3160	MED	Married	Private Em	Associate	
27	Texas/TX	non-white	17000	22	28	Male	US	4760	SAL	Single	Private Em	post-high-school	
28	California/	White	60000	67	40	Female	US	630	BUS	Divorced	Governme	Bachelor'	

At the beginning of the workflow, the outlier widget uses the Local Outlier Factor. However, according to the box plot result of the Income column, the Standard deviation is exceptionally high by around 67,645 (Figure 14). Therefore, I excluded incomes over 500,000.

To briefly mention the exploratory data analysis, according to the cleaned census data, regarding race, white ethnicity was superior by 78.23%.

Sex data is more balanced by 51.79% (Male) and 48.21%(Female).

The review of the correlation between income and continuous variables, such as age and hours, showed 0.24 and 0.33, which is a bit weak and at a medium level of correlation (Figures 14 and 15).

Figure 14.

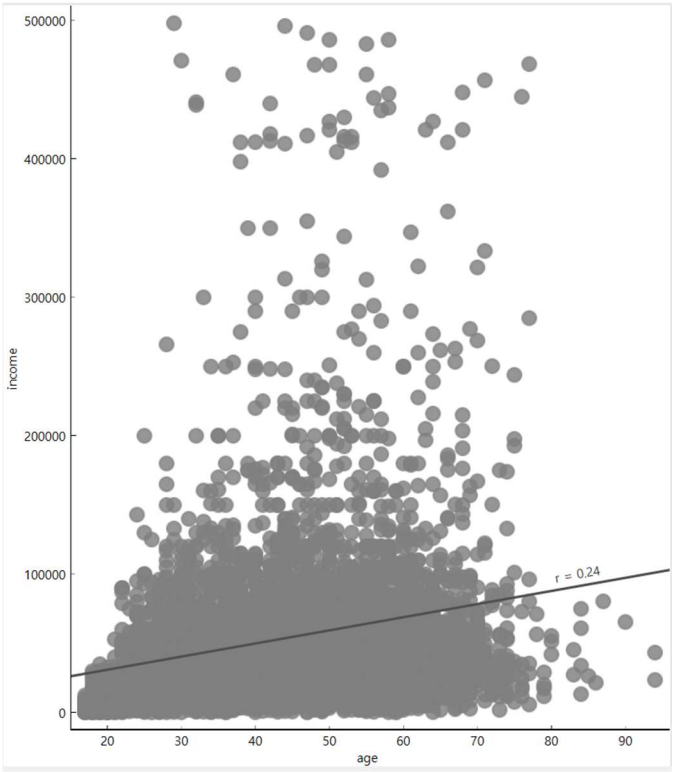
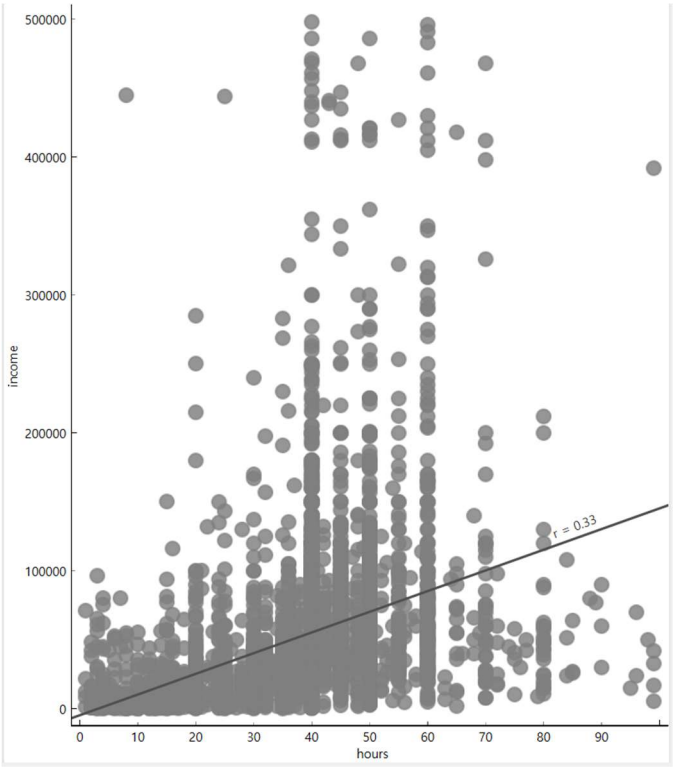

500000
400000
300000
200000
100000
0
income
r = 0.24
20
30
40
50
60
70
80
90
age

Figure 15.


500000
400000
300000
200000
100000
0
income
r = 0.33
0
10
20
30
40
50
60
70
80
90
hours

Part 2.

By applying the log transformation, the income histogram became more of a normalised distribution from the right-skewed graph (Figures 16 and 17).

Figure 16.

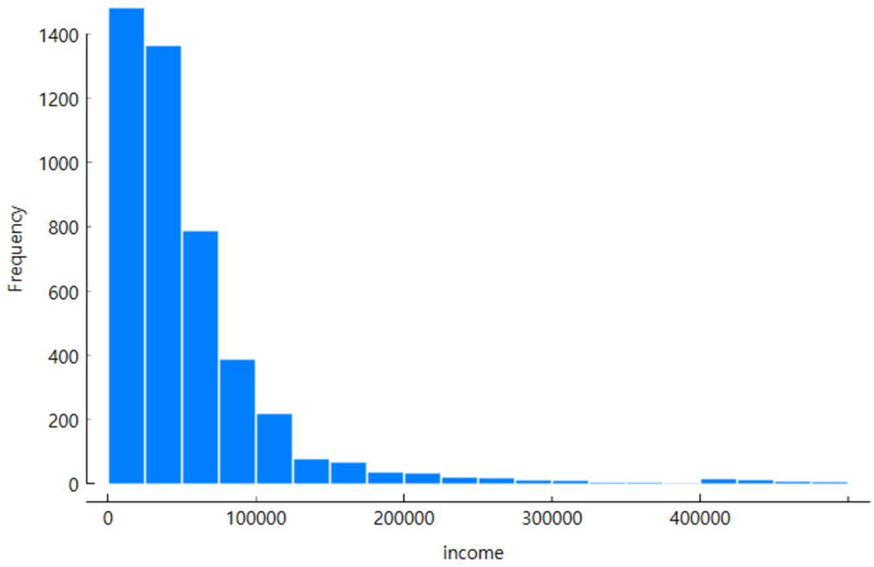
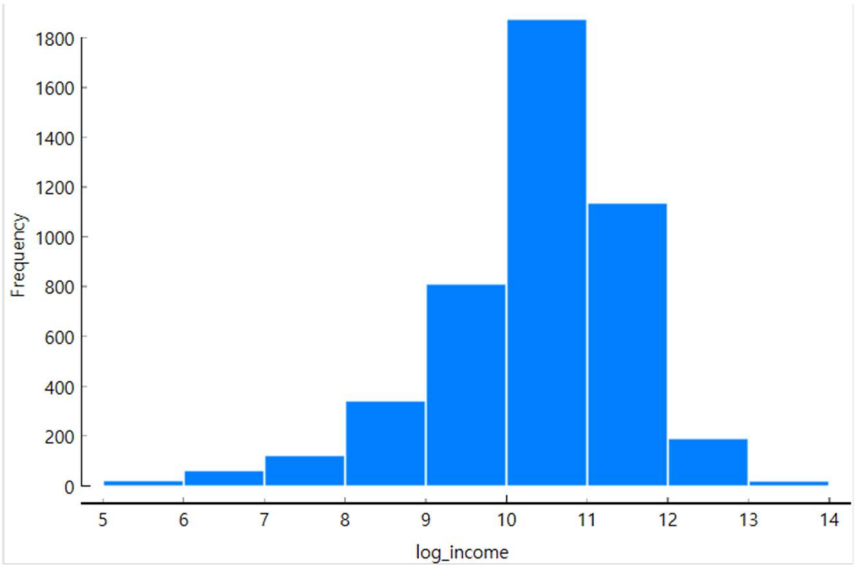
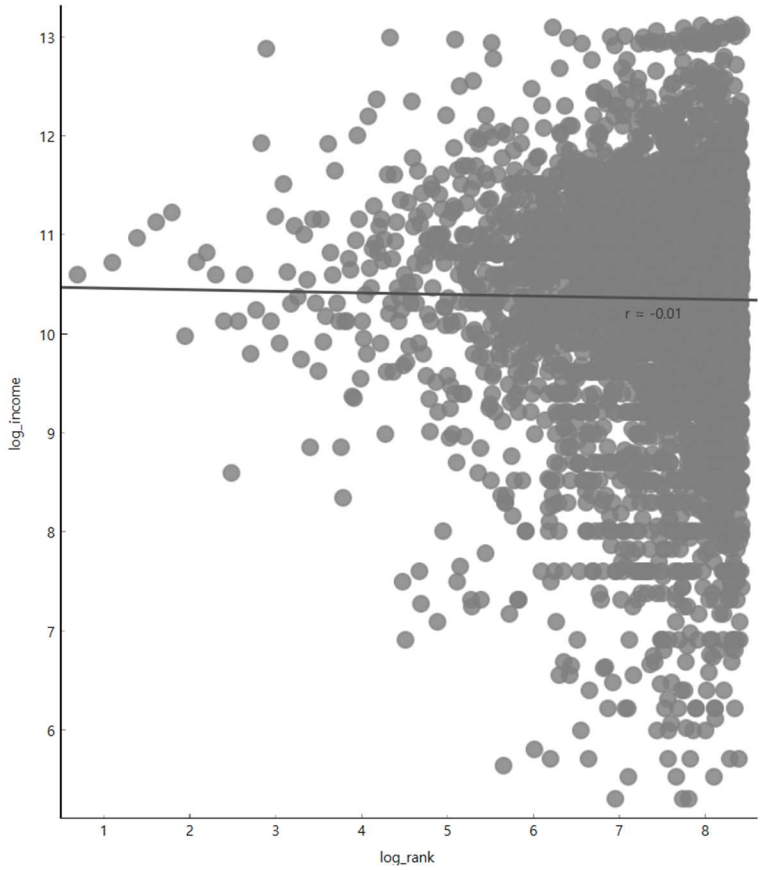


Figure 17.



Creating a rank list using Excel could create a Zipf plot. However, the scatter plot does not follow Zipf's law as the R-value is not close to -1 by -0.01 (Figure 18).

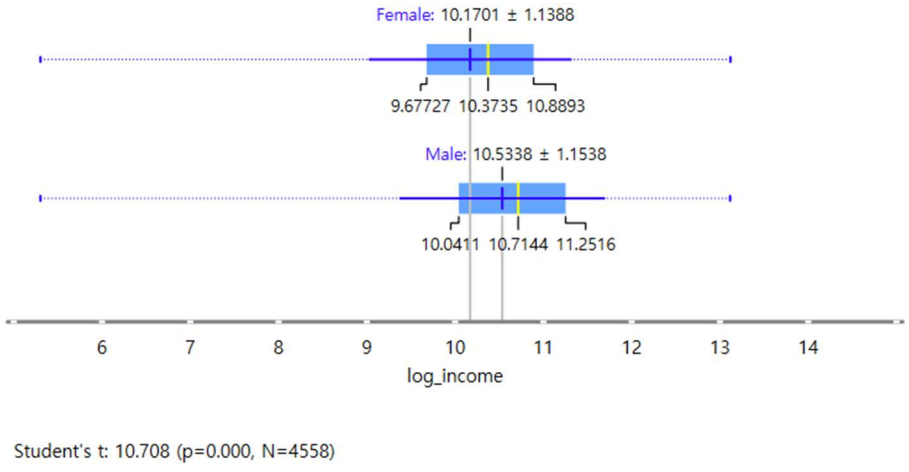
Figure 18.



Analysis by Box plots are adequate as they provide various results, including min/max, average, T-value, etc.

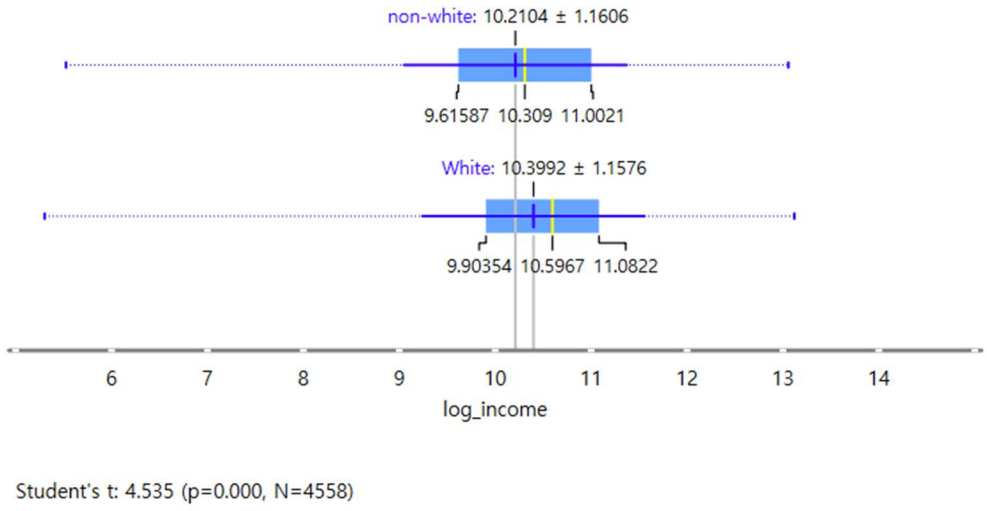
According to Figure 19, there is a difference in income by gender. The male's average log income is higher by around 10.53, while the female's is around 10.17. Furthermore, the T-value is significantly high by 10.708, which means that the average difference between the two genders is relatively high. Likewise, the P-value was very low at 0.000, meaning there is a 0% possibility of getting a value higher than the current T-value.

Figure 19.



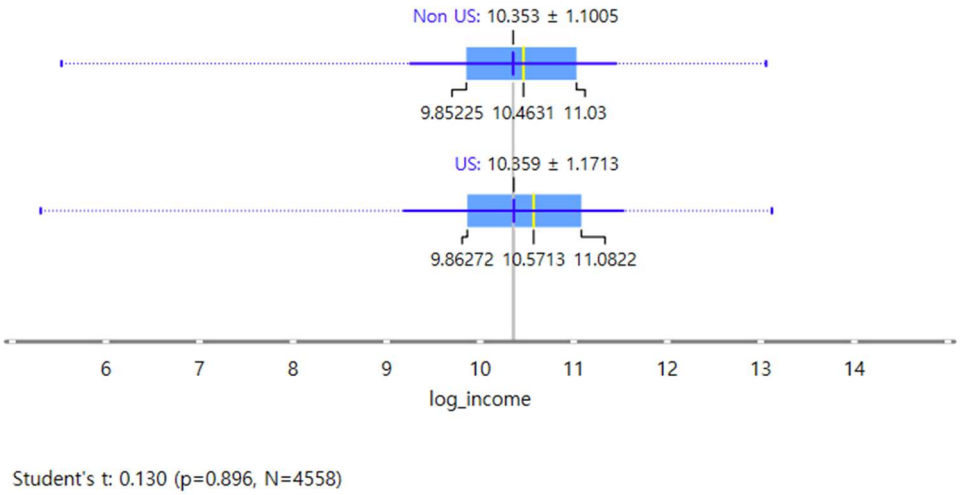
In terms of race, there was an income difference between Whites and non-whites. The White group's average log income was relatively higher by around 10.40, while the non-white group averaged around 10.21, according to Figure 20. Like the previous result, the P-value of the race attribute is 0.000. However, the T-value is less extreme than the sex attribute by 4.535 (Figure 20).

Figure 20.



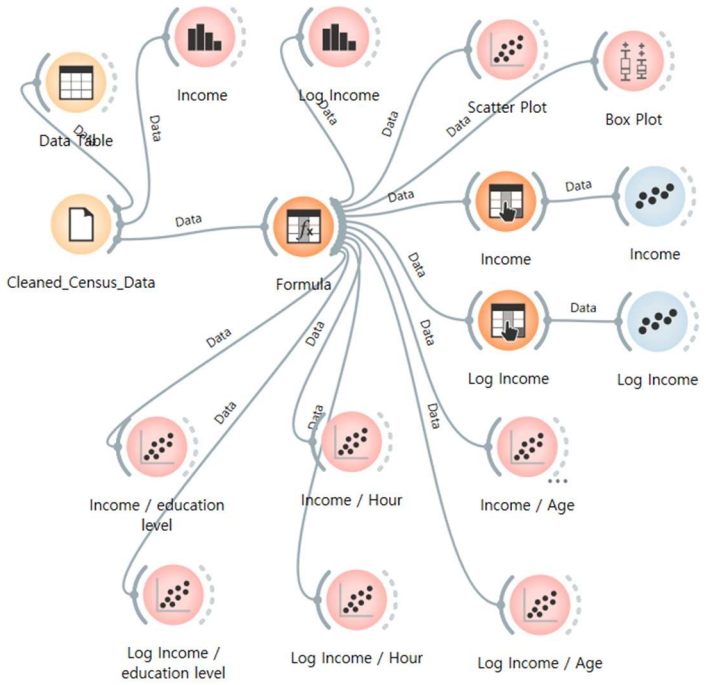
On the other hand, the PoB showed no significant difference between the US and non-US groups. The T-value was low by 0.130, which led to an increase in the P-value by 0.896, meaning there is almost a 90% probability of occurrence value over the T-value (Figure 21).

Figure 21.



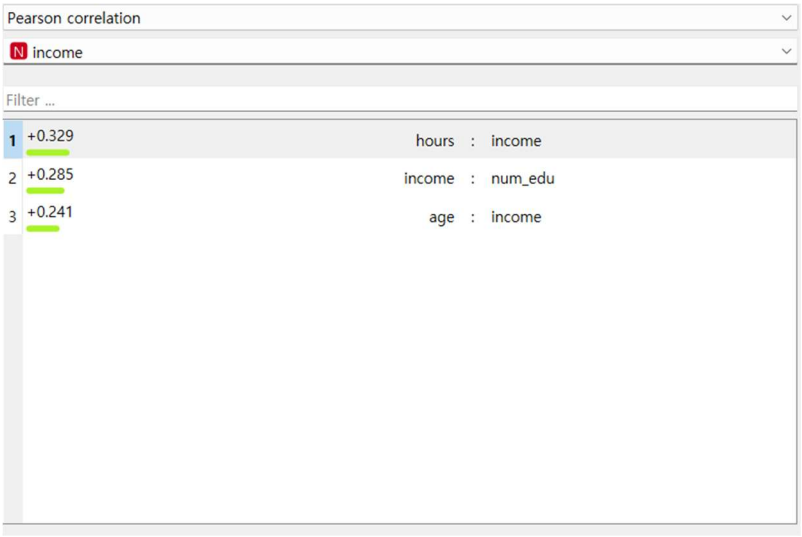
The following workflow in Figure 22 has been processed to investigate the correlation.

Figure 22.



As previously mentioned in Part 1, the correlations between hours and age showed a slightly weak correlation at each R-value of 0.329 and 0.241 (Figure 23). Furthermore, the correlation between educational level and education is somewhat weak by 0.285 (Figure 23).

Figure 23.

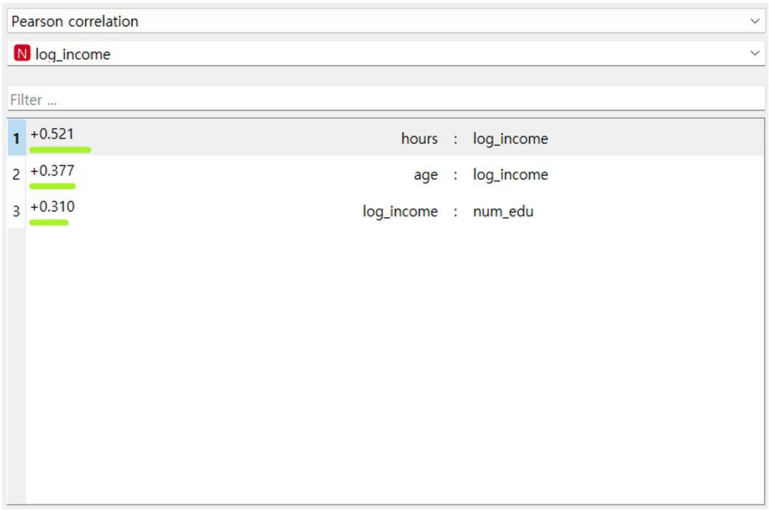


However, applying log transformation increased the R-value (Figure 24).

The previous R-value was expected to be distorted due to the extraordinarily high-income earning population. The log transformation decreased the impact of outliers and formed a better linear relationship.

As a reward for that, age and educational level show distinct correlations, and hours show strong correlations (Figure 24).

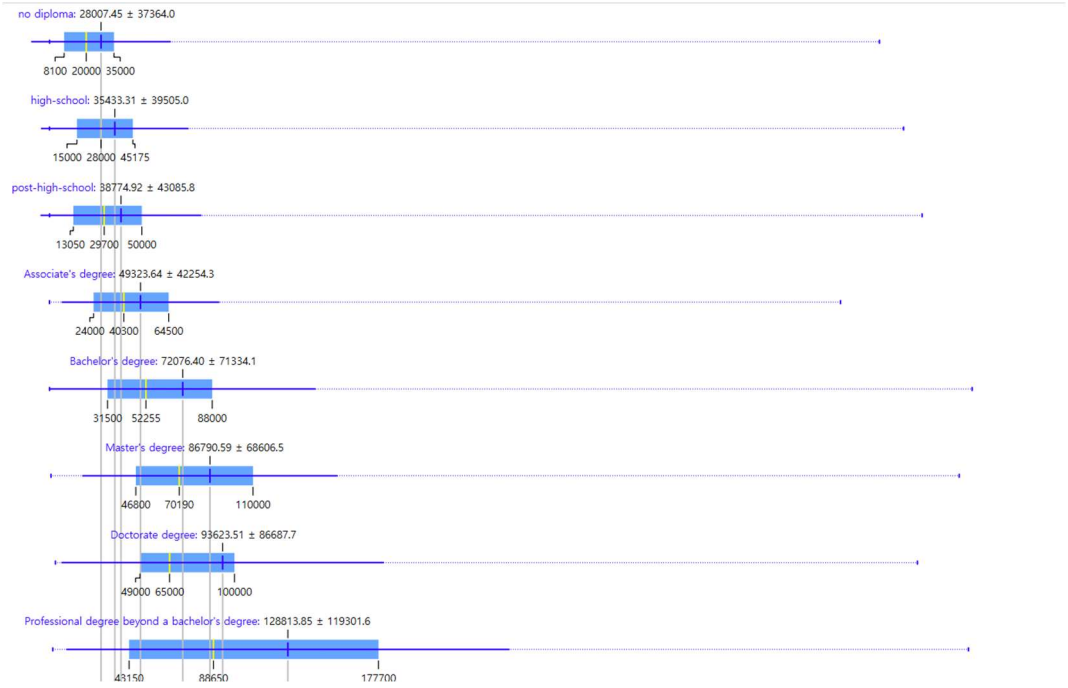
Figure 24.



Part 3.

This part of the report will focus on Income prediction. Plotting income against education level showed that the mean increased as the education level increased (Figure 25).

Figure 25.



By applying linear regression (Figure 26), the monetary value of education was estimated to be around $273 per increase in numeric education level (Figure 27). However, this result does not contain the effect of the level of education, as it is concentrated on the rise of the education yearly. Therefore, it is hard to explain how education level impacts income fully. Furthermore, the linear model's performance is extremely low, as shown in Figure 28. Considering MSE and RMSE, it interprets that the model's prediction is far different from the actual value, which led to a low R^2 value, meaning it is more of a predicting random value. To increase model performance, I have tried to apply log transformation, change attributes, and fit hyper-parameters, but it still wasn't possible to get meaningful results using the income attribute as a continuous value.

Figure 26.

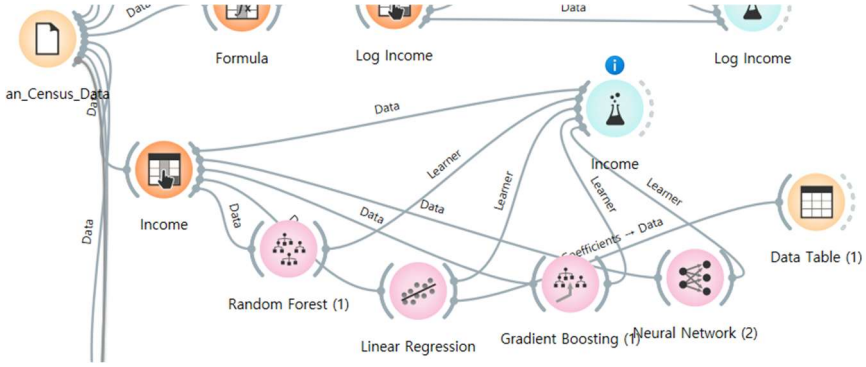


Figure 27. Income Increase by Increase of Education Level

	name	coef
1	intercept	-5656.79
2	num_edu	273.404
3	education=Ass...	-16928.4
4	education=Bac...	984.848
5	education=Doc...	16020.7
6	education=Ma...	14704.3
7	education=Pro...	43804.8
8	education=hig...	-20806.9
9	education=no ...	-20834.6
10	education=pos...	-16944.7

Figure 28. Test & Score Result

Model	MSE	RMSE	MAE	MAPE	F
inear Regression	3335383405.006	57752.778	35001.315	3.578	0.079
andom Forest (1)	3135027757.341	55991.319	33140.458	3.099	0.135
radient Boosting (1)	3131709963.405	55961.683	33151.771	3.114	0.136
leural Network (2)	6140438644.187	78360.951	50780.145	0.940	-0.695

The model's performance was reinforced by binary classifying Income into “Low Income” and “High Income” groups based on the median.

Test several classification models, as displayed in Figure 29. The gradient-boosting model scored the highest among the six models, scoring 0.876 in AUC. The AUC represents the classification model's overall performance higher as it gets closer to 1 (Figure 30).

Figure 29.

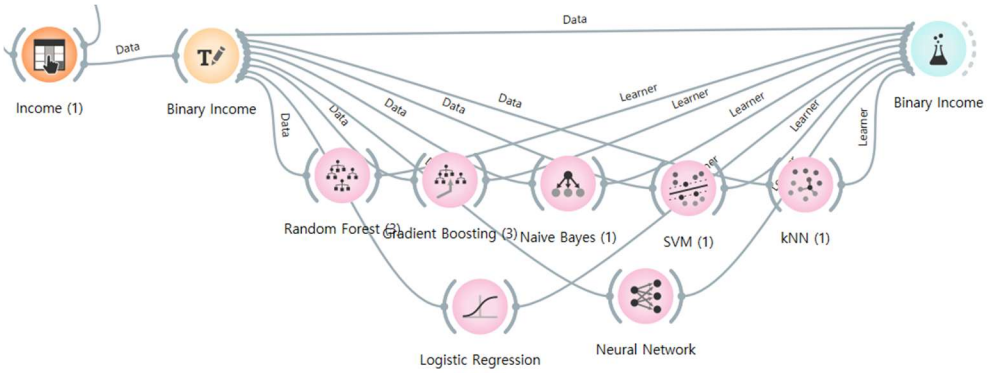


Figure 30.

Evaluation results for target (None, show average over classes)

Model	AUC	CA	F1	Prec	Recall	MCC
Random Forest (3)	0.850	0.765	0.765	0.765	0.765	0.530
Gradient Boosting (3)	0.876	0.787	0.787	0.788	0.787	0.576

The rank widget in Figure 31 shows that hours, occupation, age, and num_edu (education level in numeric code) are crucial attributes, as they scored in the top four in importance for the Gradient Boosting model (Figure 31).

Figure 31.

		#	Info. gain	Gain ratio	Gini	χ^2	Gradi...g (3)
1	N hours		0.151	0.085	0.097	627.249	0.369
2	N occupation		0.081	0.040	0.053	270.281	0.202
3	N age		0.071	0.036	0.048	244.184	0.147
4	N num_edu		0.101	0.052	0.067	486.669	0.124
5	C Industry	24	0.156	0.038	0.099	25.094	0.023
6	C sex	2	0.017	0.017	0.012	52.746	0.023
7	C marital	5	0.065	0.041	0.044	205.074	0.022
8	C education	8	0.109	0.041	0.072	527.237	0.003

Regarding the importance of the analysis by SHAP and Feature importance, it was spotted that both results prioritise age over occupation, unlike the feature ranking (Figures 32, 33, and 34).

In conclusion, the three analyses agreed that hours, age, and occupation are the key attributes. However, age and occupation were prioritised differently.

Figure 32. High Income by SHAP

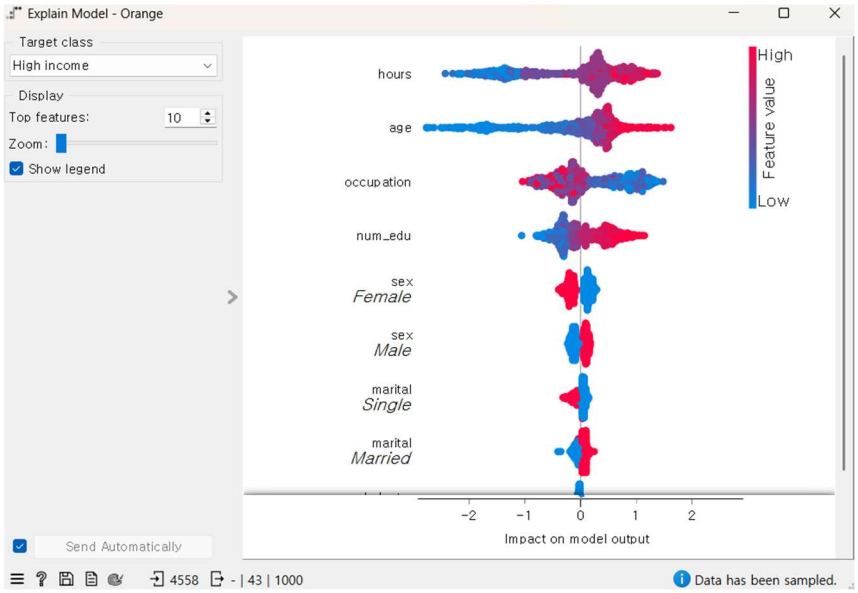


Figure 33. Low Income by SHAP

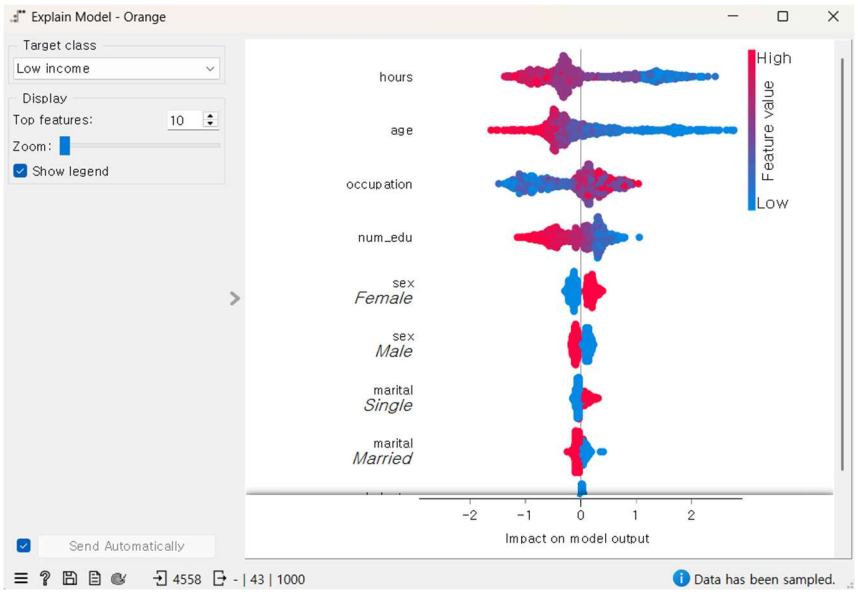
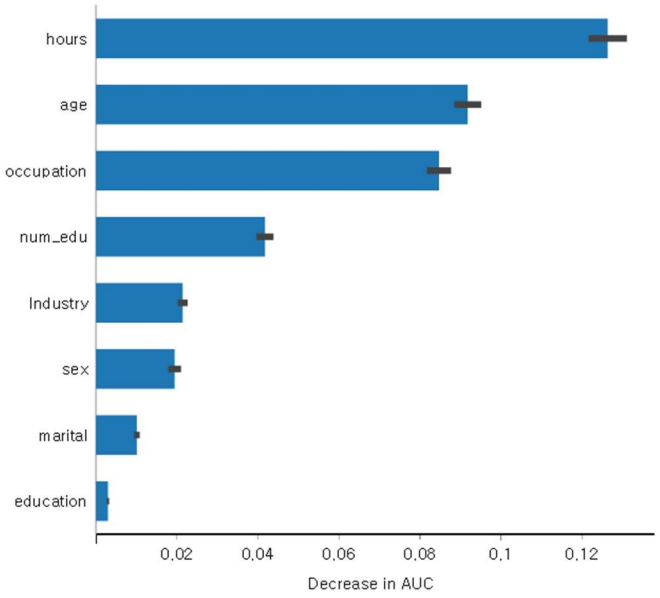


Figure 34.



Part 4.

This part of the report analyses and discusses the 2020 US presidential Election.

The mean income, educational levels, and election results by state will be displayed on the US map. The result will examine the hypothesis that “low-income states voted for Trump” and that “states with high educational attainment voted for Biden.” To simplify the visualisation interpretation, the mean income and education level were binary and classified using the median as a standard. Categorising the attribute and using the median as a standard minimises the effect of outliers and improves understandability when visualised.

Biden was elected president during the 2020 election. According to Figure 35, Biden showed dominance against Trump in 30 out of 51 states.

Regarding Figure 36, the States with low average income are red. 26 States were classified as low-income, and Trump was dominant in 10 States among low-income states (Figure 37).

Figure 38 shows that 25 states were classified as having a high education level, and Biden was dominant in 15 States (Figure 39).

In conclusion, according to the simple comparison by visualisation, the first hypothesis should be rejected as the low-income states voted for Trump less than 50%. However, the second hypothesis is relevant as states with high educational attainment voted for Biden over 50%.

Figure 35.

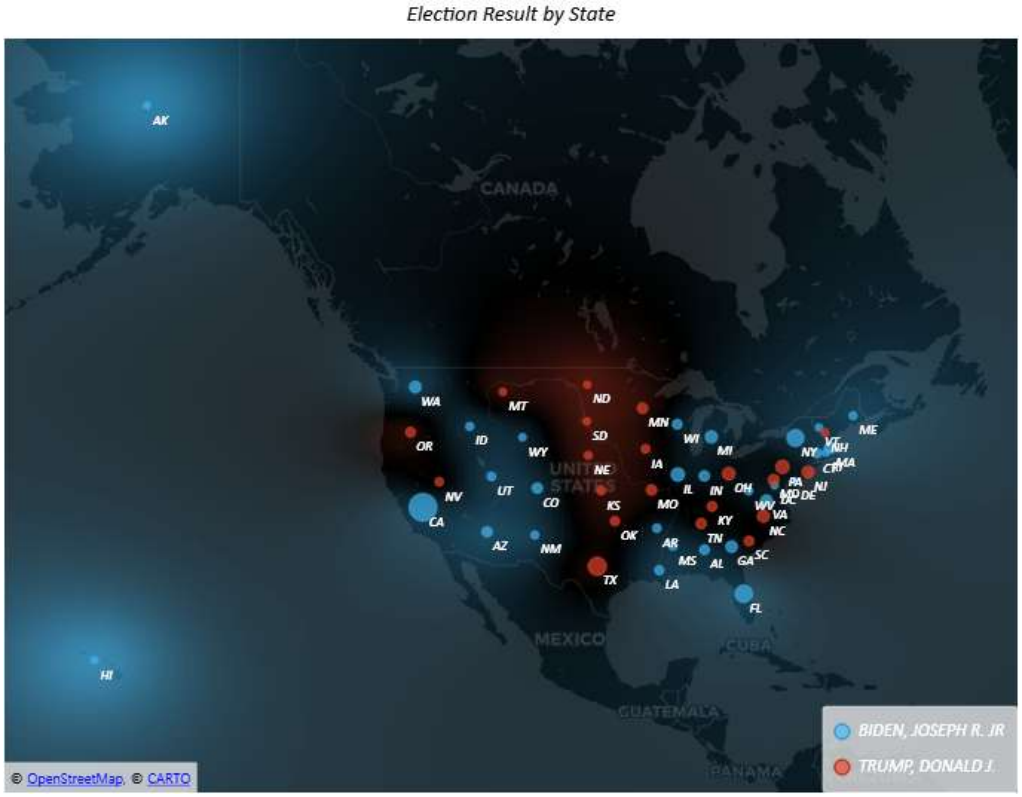


Figure 36.

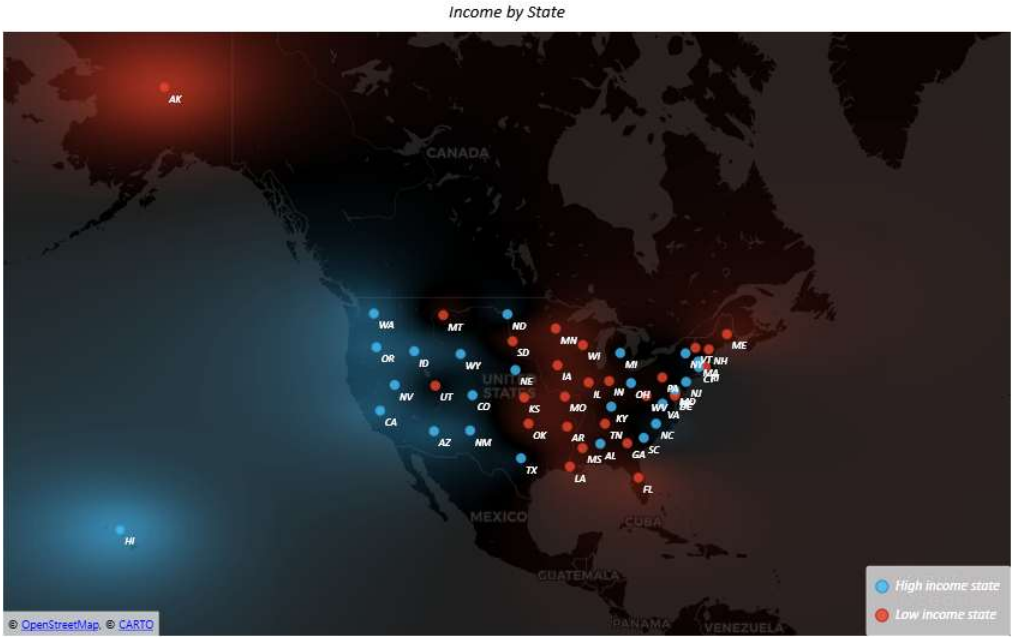


Figure 37. Trump Won in Low-income States

	state_po	ndidate - First val	datevotes - Max.	Mean income	ean education lev
1	IA	TRUMP, DONA...	897672	Low income st...	High edu level ...
2	KS	TRUMP, DONA...	771406	Low income st...	Low edu level ...
3	MN	TRUMP, DONA...	1717077	Low income st...	High edu level ...
4	MO	TRUMP, DONA...	1718736	Low income st...	Low edu level ...
5	MT	TRUMP, DONA...	343602	Low income st...	Low edu level ...
6	NH	TRUMP, DONA...	424921	Low income st...	High edu level ...
7	OK	TRUMP, DONA...	1020280	Low income st...	Low edu level ...
8	PA	TRUMP, DONA...	3458229	Low income st...	Low edu level ...
9	SD	TRUMP, DONA...	261043	Low income st...	High edu level ...
10	TN	TRUMP, DONA...	1852475	Low income st...	Low edu level ...

Figure 38.

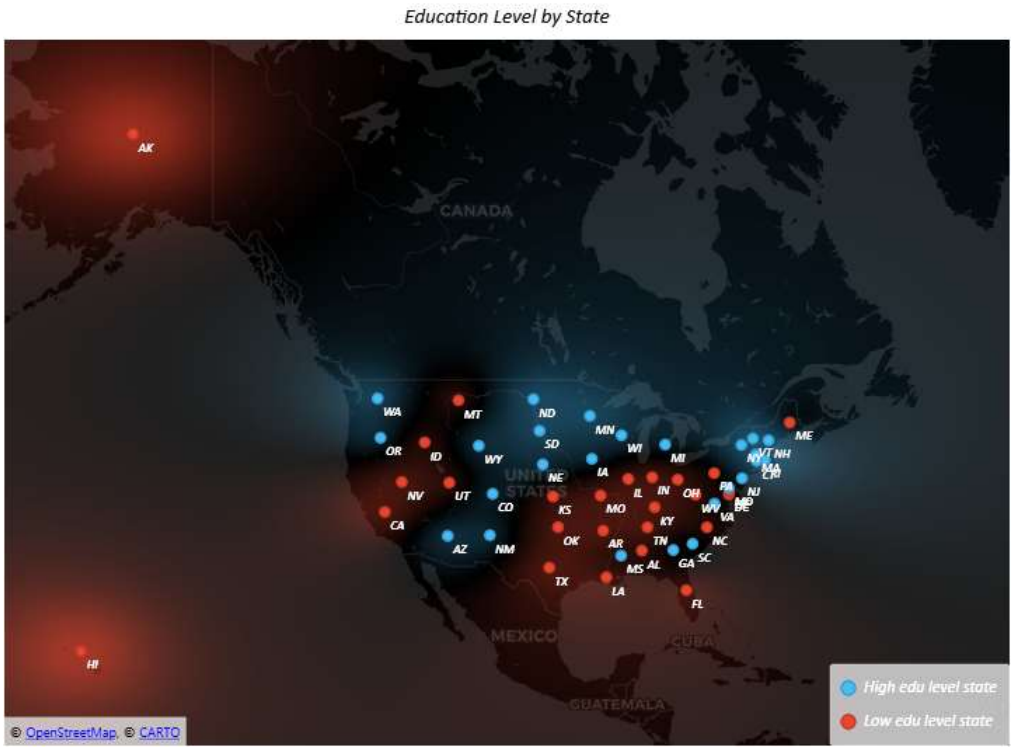


Figure 39. Biden Won in High-education Level States

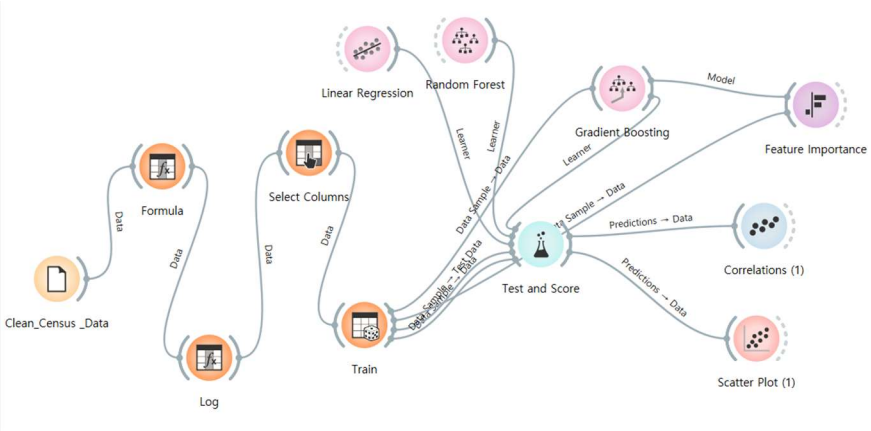
	state_po	candidate - First value	candidatevotes - Max. value	Mean income	Mean education level
1	AZ	BIDEN, JOSEPH R. JR	1672143	High income state	High edu level state
2	CO	BIDEN, JOSEPH R. JR	1804352	High income state	High edu level state
3	CT	BIDEN, JOSEPH R. JR	1080831	High income state	High edu level state
4	GA	BIDEN, JOSEPH R. JR	2473633	Low income state	High edu level state
5	MA	BIDEN, JOSEPH R. JR	2382202	High income state	High edu level state
6	MI	BIDEN, JOSEPH R. JR	2804040	High income state	High edu level state
7	MS	BIDEN, JOSEPH R. JR	756764	Low income state	High edu level state
8	NM	BIDEN, JOSEPH R. JR	501614	High income state	High edu level state
9	NY	BIDEN, JOSEPH R. JR	5230985	High income state	High edu level state
10	RI	BIDEN, JOSEPH R. JR	307486	Low income state	High edu level state
11	VA	BIDEN, JOSEPH R. JR	2413568	High income state	High edu level state
12	VT	BIDEN, JOSEPH R. JR	242820	Low income state	High edu level state
13	WA	BIDEN, JOSEPH R. JR	2369612	High income state	High edu level state
14	WI	BIDEN, JOSEPH R. JR	1630866	Low income state	High edu level state
15	WY	BIDEN, JOSEPH R. JR	193559	High income state	High edu level state

Part 5.

This part of the report examines the hypothesis that “higher education levels lead to higher income per hour worked.” Through the examination process, there will be a review of the correlation between education level and income per working hour.

First, the derived variable “Income_per_hour” was added using the formula income/hours. Then, we applied log transformation to better spot the linear relation.

Figure 40.



According to the scatter plot and Pearson correlation, it was able to find the positive correlation between Log_income_per_hour and num_edu by scoring 0.32 (Figures 40 and 41).

Furthermore, according to Gradient Boosting model’s feature importance, the num_edu was the second important variable that would decrease the model performance (Figure 42).

However, according to test & score in Figure 43, the model has low performance. Meaning that it is hard to predict income per hour only using education level.

In conclusion, there was positive correlation spotted, however it is a weak correlation and it is expected that other attributes will have better influence on the income per hour.

Figure 40.

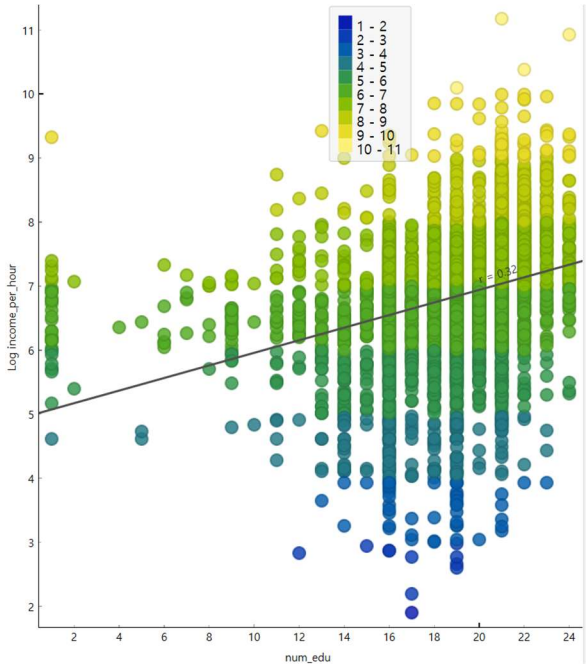


Figure 41.

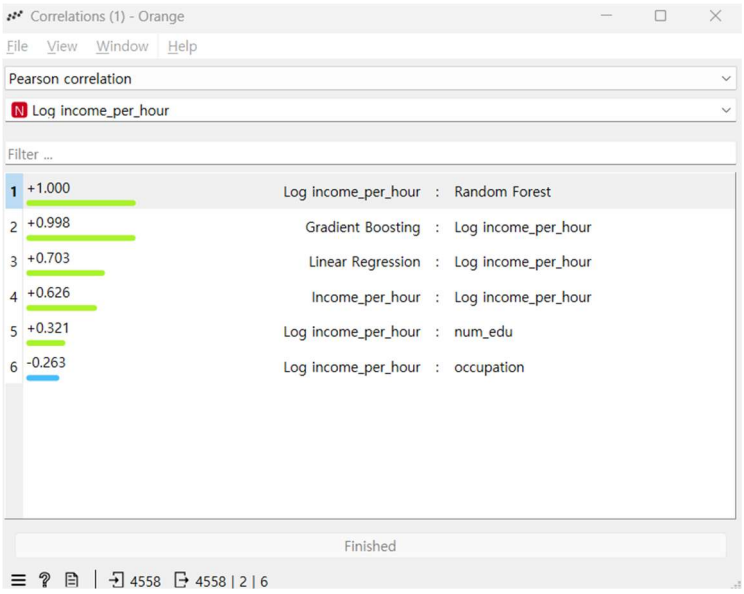


Figure 42.

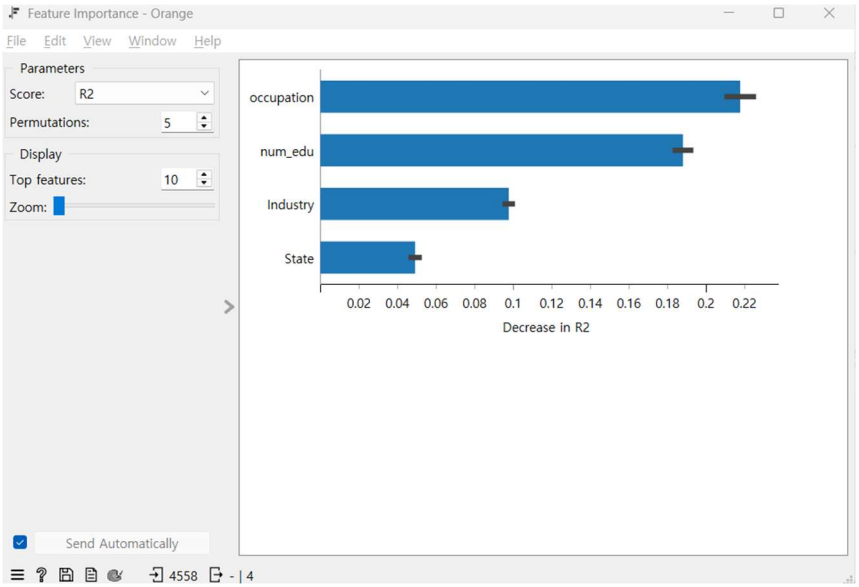
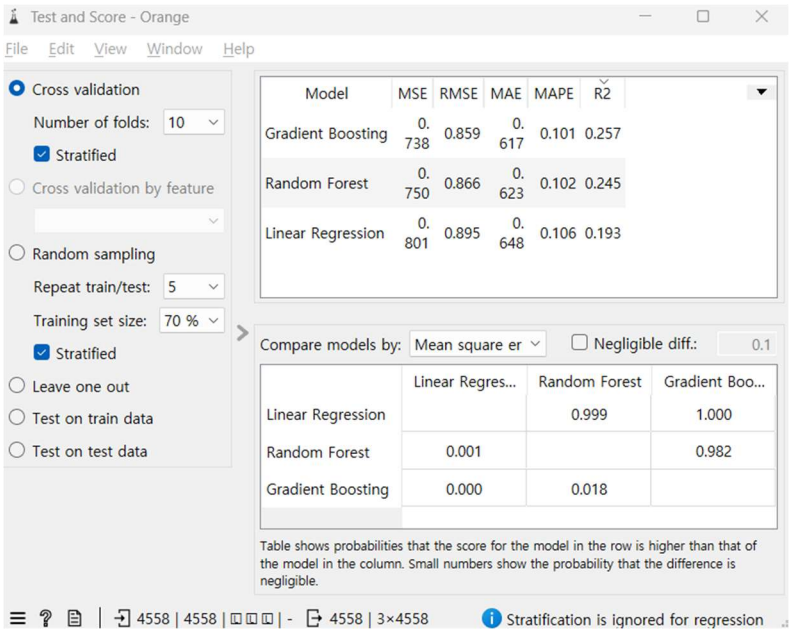

Feature Importance - Orange
File Edit View Window Help
Parameters
Score: R2
Permutations: 5
Display
Top features: 10
Zoom:
Send Automatically
occupation
num_edu
Industry
State
0.02 0.04 0.06 0.08 0.1 0.12 0.14 0.16 0.18 0.2 0.22
Decrease in R2
4558 - | 4

Figure 43.


Test and Score - Orange
File Edit View Window Help
Cross validation
Number of folds: 10
Stratified
Cross validation by feature
Random sampling
Repeat train/test: 5
Training set size: 70 %
Stratified
Leave one out
Test on train data
Test on test data
Model MSE RMSE MAE MAPE R2
Gradient Boosting 0. 738 0.859 0. 617 0.101 0.257
Random Forest 0. 750 0.866 0. 623 0.102 0.245
Linear Regression 0. 801 0.895 0. 648 0.106 0.193
Compare models by: Mean square er
Negligible diff.: 0.1
Linear Regres... Random Forest Gradient Boo...
Linear Regression 0.999 1.000
Random Forest 0.001 0.982
Gradient Boosting 0.000 0.018
Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.
4558 | 4558 | - 4558 | 3×4558
Stratification is ignored for regression