An Analysis of the Impact of Socioeconomic Factors on Personal Income Levels: A global inspective

Group 29: Yang Jiateng, Lin Gaoli, Bi Yanhan, Wang Qiyue, Yan Jingfei
(School of Mathematics and Statistics)

Introduction

- **Research Background**

According to the dataset of the US 1994 Census database, it contains data on individuals regarding their income level, and various socioeconomic factors:

Age, Education, Marital_Status, Occupation, Sex, Hours_Pw, Nationality and **Income.**

- **The main objective of this study:**

Based on the census data, determine:

- which characteristics will influence an individual's income
- which factors will lead to a situation where a person's annual income exceeds 50,000 US dollars.

Exploratory data analysis

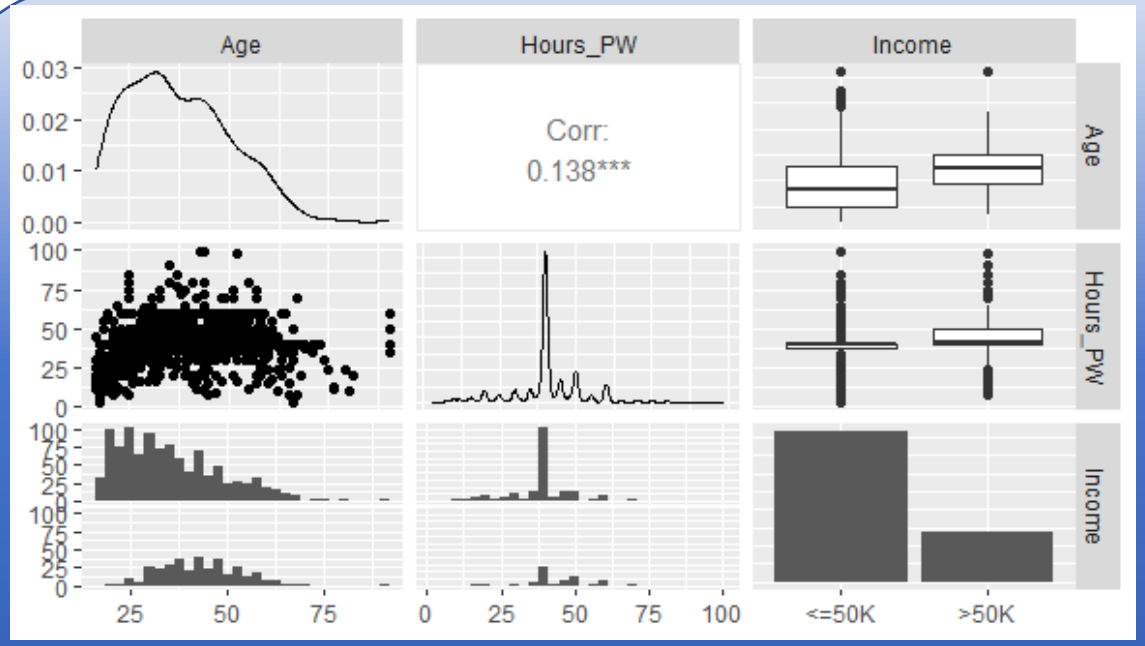


Fig.1. Numeric Data Correlation

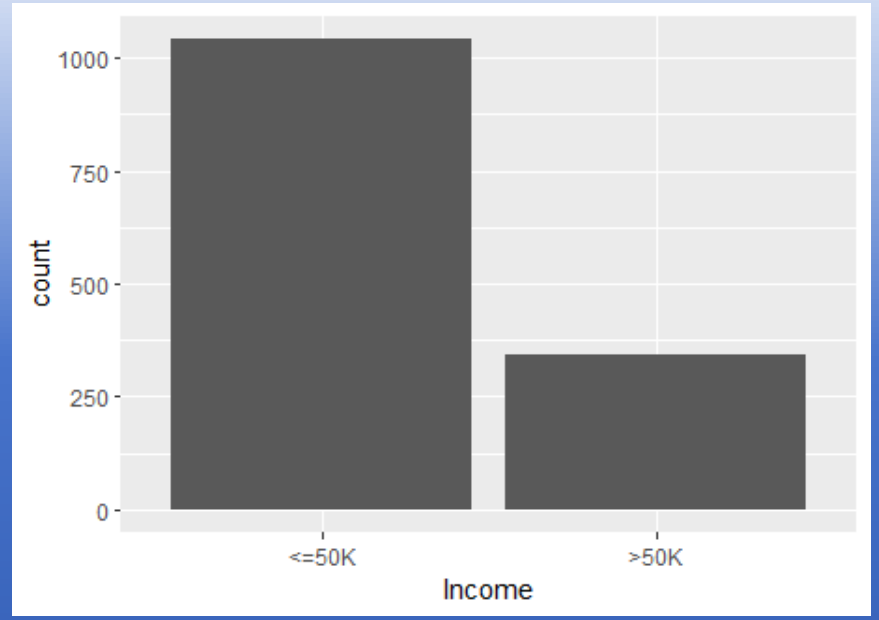


Fig.2. Income Distribution

From fig.1, there is a weak positive correlation between Age and Hours_PW (0.138). The scatter plot shows that older people are slightly more inclined to work longer hours.
The box plots show that groups with higher incomes tend to be older and work longer hours.

From fig.2, there is a significant class imbalance that can bias models toward the majority class and impair predictive accuracy for high-income individuals if not properly addressed.

Exploratory data analysis

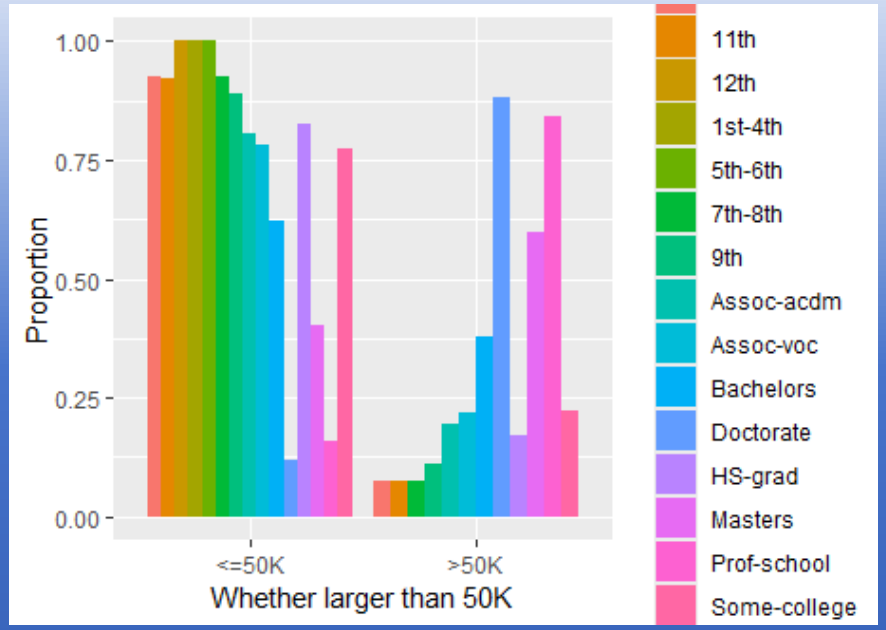


Fig.3. Distribution of Education Levels by income Group

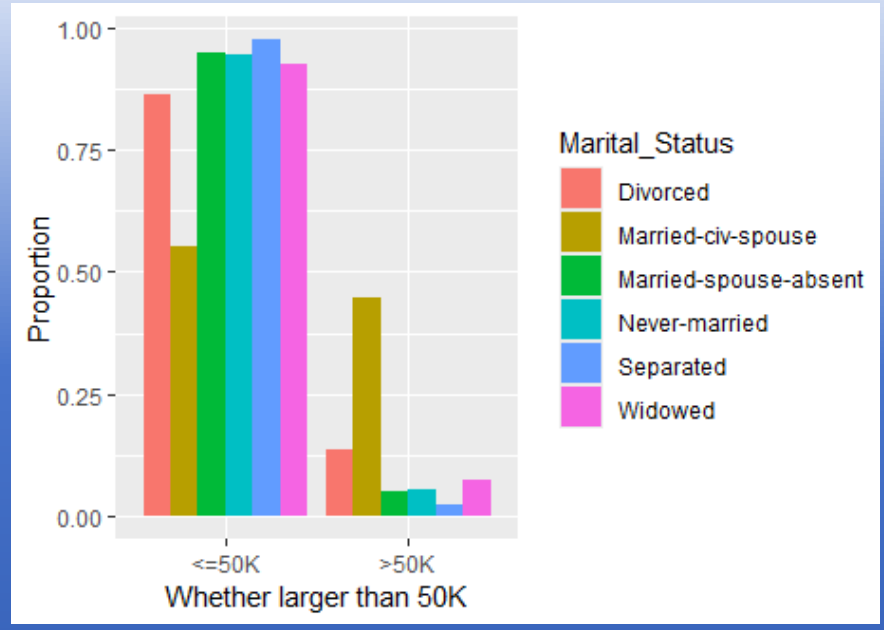


Fig.4. Distribution of Marital_Status Levels by income Group

Fig.3. shows a clear correlation between higher education and increased income potential: Lower educational levels predominate among individuals earning “<=50K”, whereas higher educational attainments (like Bachelors, Masters, and above) are more prevalent among those earning “>50K”.

This chart(fig.4.) shows that “Never-married” is the most prevalent marital status among individuals earning “<=50K”, whereas “Married-civ-spouse” overwhelmingly dominates among those earning “>50K”.

Exploratory data analysis

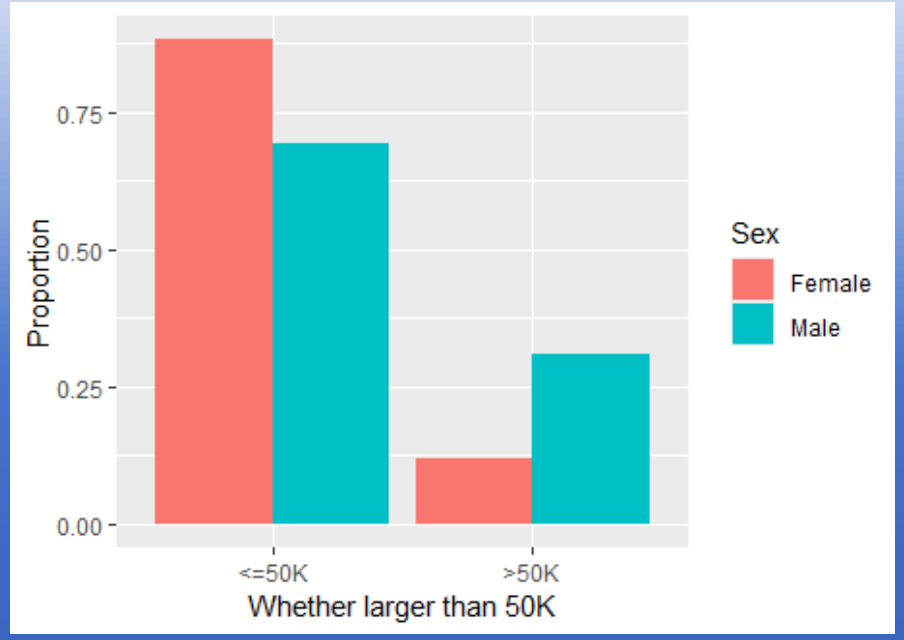


Fig.5. Distribution of Sex Levels by income Group

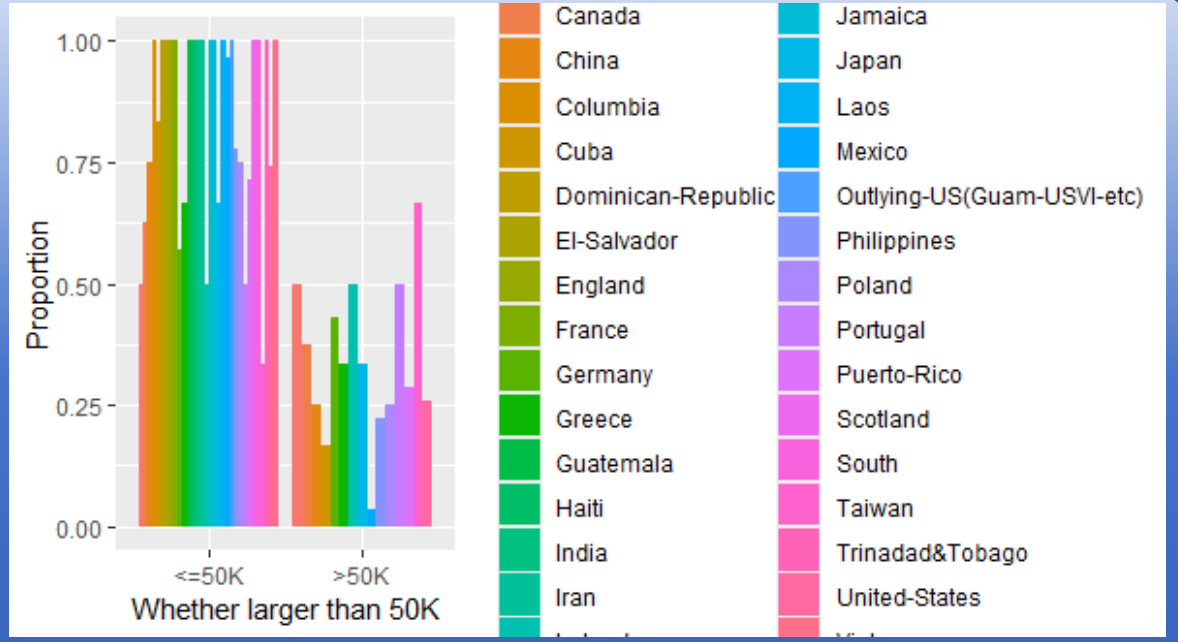


Fig.6. Distribution of Nationality Levels by income Group

The chart(fig.5.) shows that females dominate the “<=50K” income group, while males are more prevalent in the “>50K” income group.

Across nearly all nationalities, the majority of individuals earn “<=50K”, with relatively small proportions exceeding 50K. However, 1255 (90.7%) of the sample in the dataset had American citizenship. Because of this highly unbalanced distribution, nationality does not fit well as an explanatory variable.

Exploratory data analysis

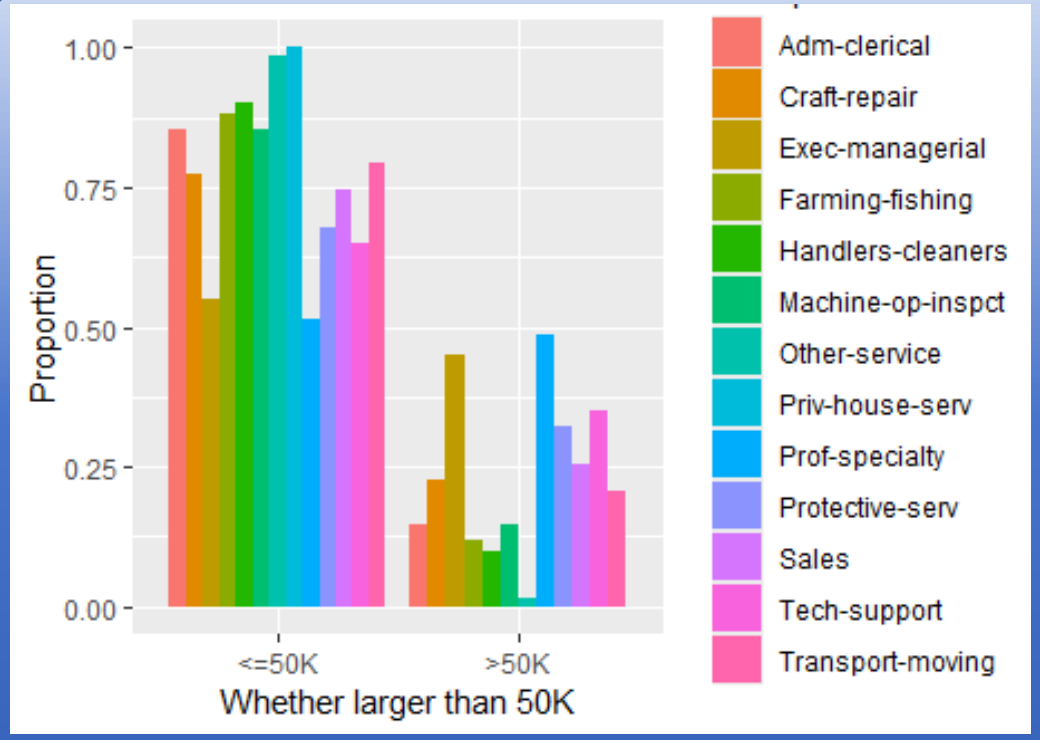


Fig.7. Distribution of occupation Levels by income Group

The chart shows that lower-paying occupations (e.g., Other-service, Adm-clerical) are more prevalent among individuals earning“<=50K”, whereas higher-paying roles (like Exec-managerial and Prof-specialty) dominate in the “>50K” income group.

After the above analysis, we observed that some categorical variables, such as **education and occupation**, have too many unbalanced categories - and sometimes too few observations, or even completely separate - leading to convergence problems. In subsequent model improvements, we will try to **merge them into broader**, representative groups to improve the robustness and interpretability of the model and ensure there is sufficient data and clearer insights into the impact of revenue.

Exploratory data analysis

Use the Chi_Square Test to test the correlation of category variable with response.

Vs . Income	X-squared	df	P-value
Education	193.07	14	<2.2e-16
Marital_Status	263.24	5	<2.2e-16
Occupation	167.57	12	<2.2e-16
Sex	57.101	1	4.139e-14
Nationality	28.476	31	0.5965

The chart shows that the p-value of Education, Marital_Status, Occupation, Sex are lower than 0.05, it shows that there is a statistically significant correlation. However, the p-value of Nationality is larger than 0.05, so they may be independent.

The Full Generalized Linear Model

The model selection process starts with fitting a Generalized Linear Model (GLM) with a binomial logistic regression to predict whether an individual's income falls into the >50K or <=50K category.

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \cdot Age + \beta_2 \cdot Education + \beta_3 \cdot Sex + \beta_4 \cdot Hours\_PW$$
$$+\beta_5 \cdot Marital\_Status + \beta_6 \cdot Occupation + \beta_7 \cdot Nationality$$

$where$:

$\cdot\ p = P(Income \geq 50K)\ is\ the\ probability\ that\ an\ individual\ earns\ more\ than\ 50K;$

$\cdot\ \beta_0\ is\ the\ intercept, and\ \beta_1, \beta_2, \cdots, \beta_7\ are\ the\ coefficients\ corresponding\ to\ each\ explanatory\ variable.$

There are some partial outputs.

Education11th	1.105e+00	9.856e-01	1.121	0.262436
Education12th	-1.448e+01	1.244e+03	-0.012	0.990718
Education1st-4th	-2.981e+01	1.875e+03	-0.016	0.987315
Education5th-6th	-1.524e+01	1.475e+03	-0.010	0.991753
Education7th-8th	-1.686e-01	1.027e+00	-0.164	0.869615
Education9th	2.560e-01	9.462e-01	0.271	0.786769
EducationAssoc-acdm	1.612e+00	7.895e-01	2.042	0.041190
EducationAssoc-voc	8.442e-01	7.523e-01	1.122	0.261769
EducationBachelors	2.034e+00	7.010e-01	2.902	0.003707
EducationDoctorate	4.593e+00	1.098e+00	4.183	2.88e-05
EducationHS-grad	1.042e+00	6.809e-01	1.531	0.125784
EducationMasters	2.290e+00	7.394e-01	3.097	0.001952
EducationProf-school	3.103e+00	9.911e-01	3.131	0.001741
EducationSome-college	1.450e+00	6.857e-01	2.115	0.034466

OccupationCraft-repair	-9.979e-02	3.548e-01	-0.281	0.778491
OccupationExec-managerial	3.437e-01	3.422e-01	1.004	0.315155
OccupationFarming-fishing	-1.698e+00	6.081e-01	-2.792	0.005240
OccupationHandlers-cleaners	-1.772e-01	5.794e-01	-0.306	0.759728
OccupationMachine-op-inspct	-5.569e-01	4.414e-01	-1.261	0.207134
OccupationOther-service	-2.477e+00	7.982e-01	-3.103	0.001913
OccupationPriv-house-serv	-1.633e+01	2.648e+03	-0.006	0.995080
OccupationProf-specialty	6.458e-01	3.595e-01	1.796	0.072471
OccupationProtective-serv	-2.791e-02	5.369e-01	-0.052	0.958538
OccupationSales	-5.323e-02	3.641e-01	-0.146	0.883781
OccupationTech-support	1.037e+00	5.441e-01	1.905	0.056736
OccupationTransport-moving	-3.092e-01	4.362e-01	-0.709	0.478456

In the output of the full_model, age, working hours, higher education level (such as college, bachelor, doctor, master, professional school and some universities), some occupations and married status have significant positive effects on income, **while** gender and some lower-level education categories do not reach statistical significance. It shows that there are significant differences in the importance of different variables in predicting income.

Model Optimization-StepAIC Selection Model

Stepwise Akaike Information Criterion (AIC) selection is used to iteratively remove the least significant predictors to find the best-performing model.

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \cdot Age + \beta_2 \cdot Education + \beta_3 \cdot Hours\_PW + \beta_4 \cdot Marital\_Status + \beta_5 \cdot Occupation$$

$where$:

$\cdot\ p = P(Income \geq 50K)\ is\ the\ probability\ that\ an\ individual\ earns\ more\ than\ 50K;$

$\cdot\ \beta_0\ is\ the\ intercept, and\ \beta_1, \beta_2, \cdots, \beta_5\ are\ the\ coefficients\ corresponding\ to\ each\ explanatory\ variable.$

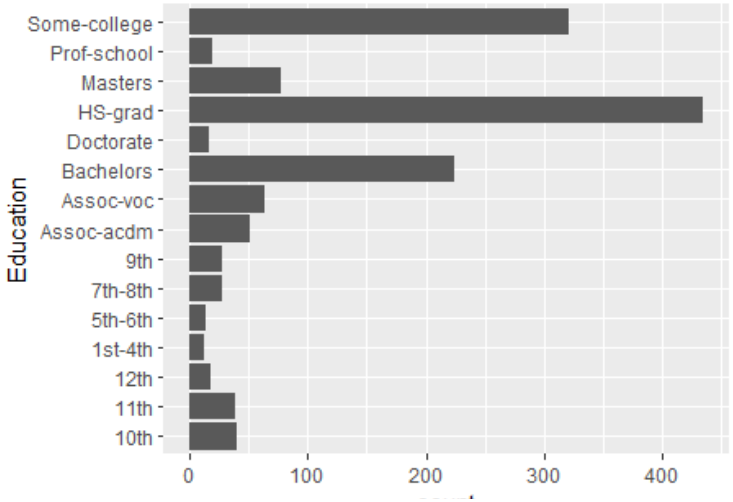
There are some partial outputs.

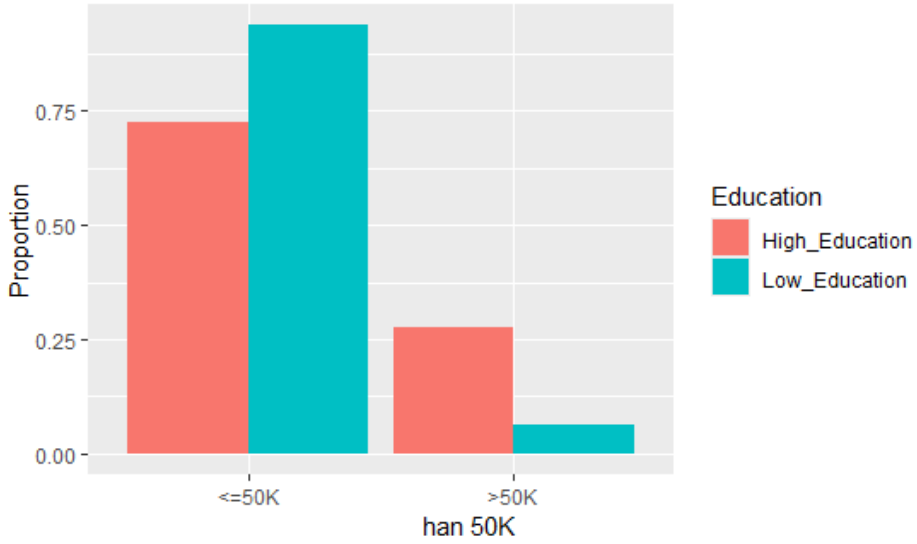
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-6.433e+00	9.215e-01	-6.982	2.92e-12	***
Age	3.793e-02	7.413e-03	5.117	3.10e-07	***
Education11th	1.288e+00	9.495e-01	1.357	0.17493	
Education12th	-1.362e+01	7.644e+02	-0.018	0.98578	
Education1st-4th	-1.459e+01	9.286e+02	-0.016	0.98747	
Education5th-6th	-1.447e+01	9.224e+02	-0.016	0.98748	
Education7th-8th	-1.669e-01	1.019e+00	-0.164	0.86992	
Education9th	3.944e-01	9.297e-01	0.424	0.67141	
EducationAssoc-acdm	1.560e+00	7.818e-01	1.996	0.04596	*
EducationAssoc-voc	8.264e-01	7.447e-01	1.110	0.26717	
EducationBachelors	1.943e+00	6.912e-01	2.811	0.00494	**
EducationDoctorate	4.400e+00	1.094e+00	4.022	5.77e-05	***
EducationHS-grad	1.043e+00	6.714e-01	1.554	0.12023	
EducationMasters	2.264e+00	7.285e-01	3.107	0.00189	**
EducationProf-school	3.036e+00	9.789e-01	3.102	0.00192	**
EducationSome-college	1.422e+00	6.767e-01	2.102	0.03557	*

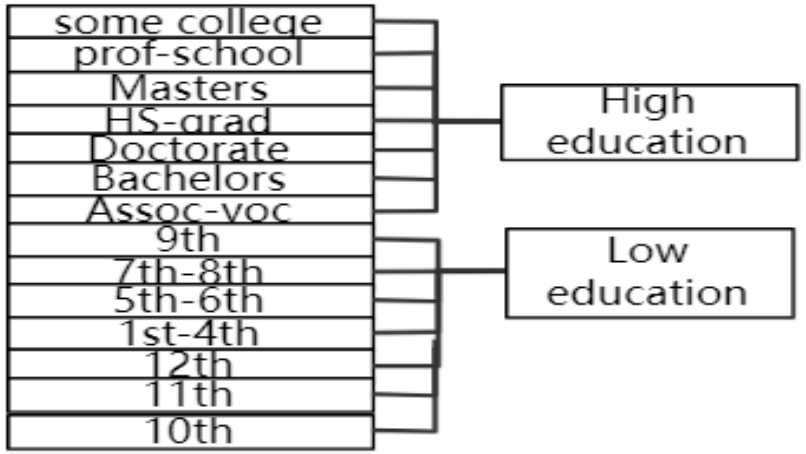
Stepwise AIC selection helps refine the model by retaining only the most informative variables, reducing overfitting and improving interpretability. The final optimized model suggests that **Age, Education, Work Hours, Marital Status, and Occupation** are the strongest predictors of income. Nationality and Sex are removed in the AIC-selected model, indicating they may not have a significant impact on predicting income in this dataset. The model suffers from perfect separation issues, with the explanatory variables Education and Occupation exhibiting perfect separation.

Model Optimization-Levels Merging

To resolve the perfect separation issue, merge the categories of some explanatory variables.



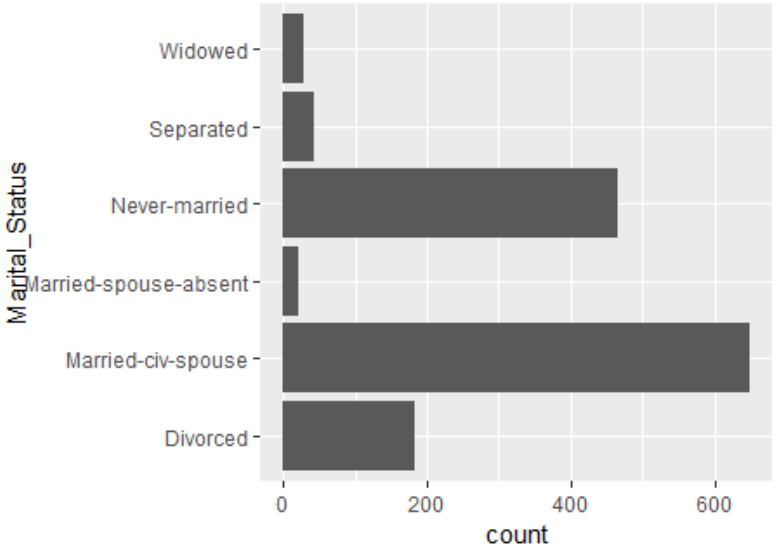


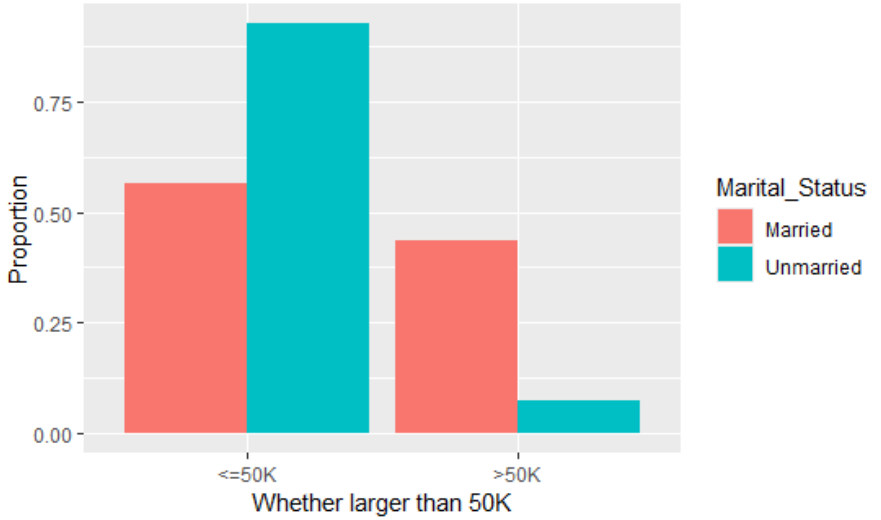


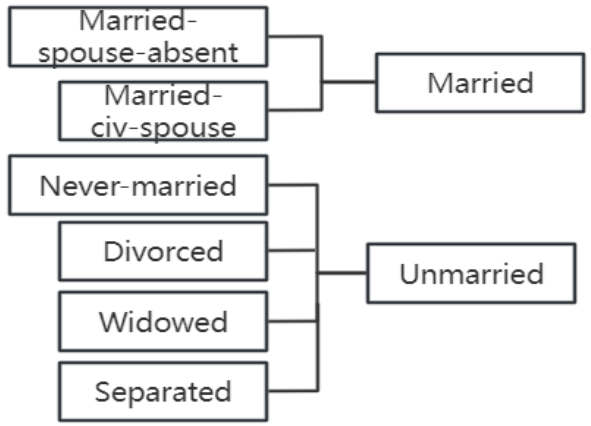
Merge the **Education**:
The newly created **Low Education and High Education** categories allow for a clearer comparison of income distribution between different education levels. Individuals with higher education levels are more likely to earn>50K compared to those in the Low Education category.

Model Optimization-Levels Merging

To resolve the perfect separation issue, merge the categories of some explanatory variables.



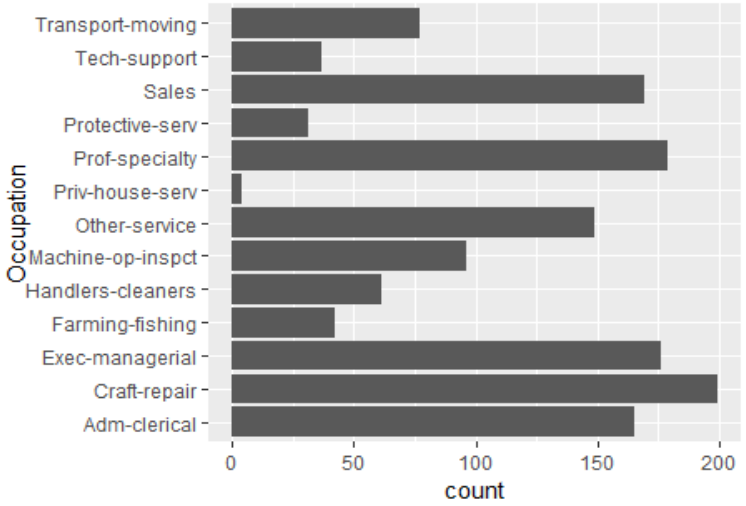


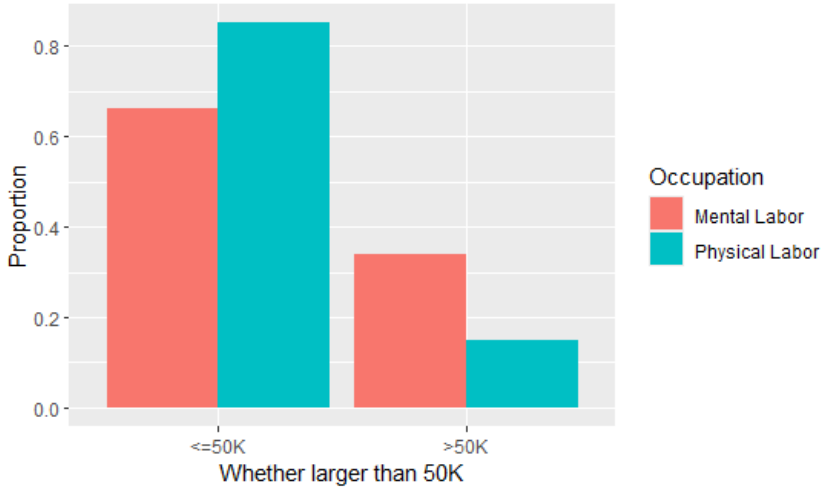


Merge the **Marital_Status**:
The distribution of income across the newly grouped **Married and Unmarried** categories is displayed. Married individuals show a higher proportion of high-income earners, reinforcing previous findings about marital status and financial stability.

Model Optimization-Levels Merging

To simplify occupational categories and address class imbalance, occupations have been grouped into two broader categories: Mental Labor and Physical Labor.



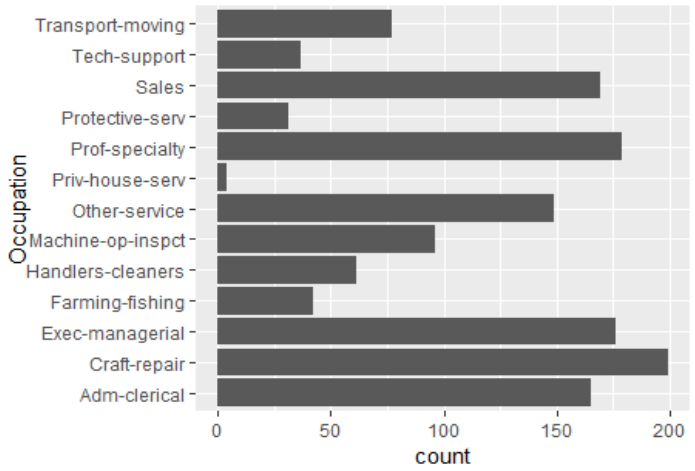


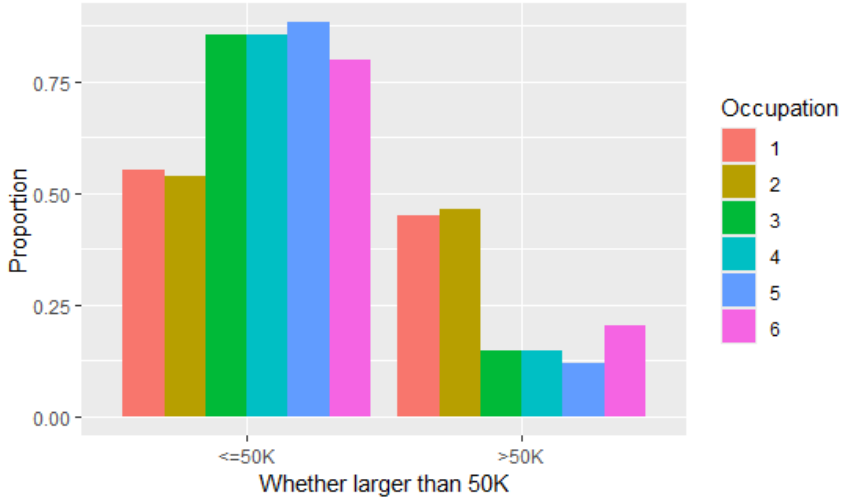
Merge the **Occupation(Mental-Physical)**:

Mental labor jobs have a higher proportion of individuals earning >50K, reinforcing the idea that cognitive and managerial roles tend to offer better salaries. **Physical labor** jobs predominantly fall in the <=50K category, suggesting that manual labor occupations generally provide lower wages. The proportion of high-income earners in mental labor jobs is significantly higher than in physical labor jobs, highlighting the economic advantage of cognitive and executive occupations.

Model Optimization-Levels Merging

To further refine the occupation categories, we classify jobs according to the PRC Job Classification List. This classification system groups occupations based on skill levels and job nature, and provides a structured way to analyze income disparities across different occupational groups.



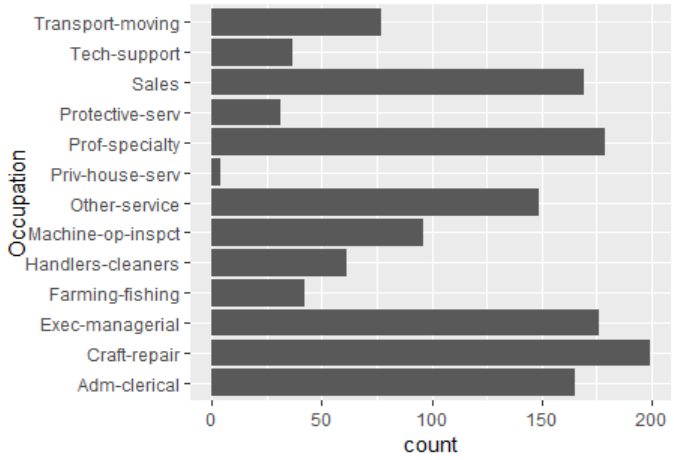


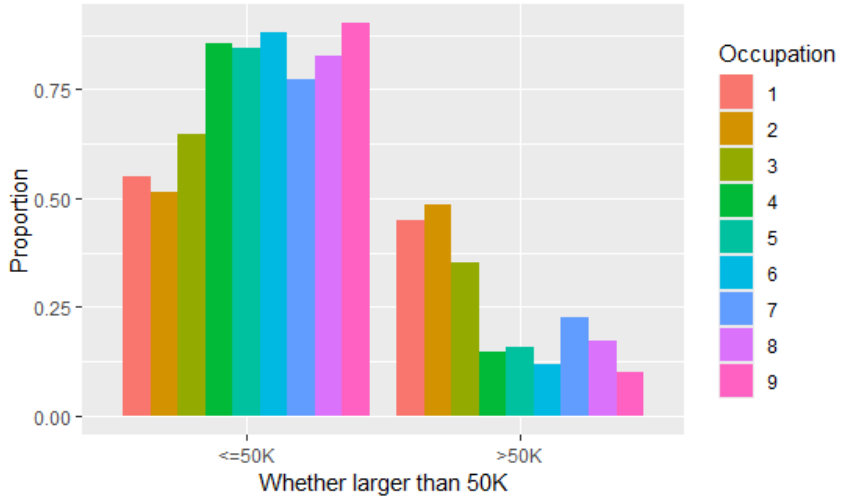
Merge the **Occupation(by PRC Job Classification List)**:
Management and technical jobs tend to have higher salaries, while sales, service, and agricultural jobs have a greater share of low-income earners. Category 1 (Senior Management) has the highest proportion of individuals earning >50K, reinforcing the idea that executive roles are highly paid. Category 2 (Specialists & Technical Support) also has a notable presence in the high-income group, indicating that specialized skills lead to better salaries. Category 4 (Sales & Service) and Category 5 (Agriculture & Fishing) have the lowest share of high-income earners, highlighting the financial struggles in these job sectors.

ref: https://zchweb.oss-cn-beijing.aliyuncs.com/contract/temp/20211221165413633304.pdf

Model Optimization-Levels Merging

the International Standard Classification of Occupations (ISCO-08) is also used to categorize occupations into structured groups based on job function and skill level. This classification allows for a globally standardized approach to analyzing income distribution by profession.





Merge the **Occupation(by ISCO-08)**: Category 1 (Managers) and Category 2 (Professionals) have the highest share of individuals earning >50K, emphasizing that managerial and specialized roles offer better earnings. Category 5 (Service & Sales) and Category 6 (Agricultural & Fishery) exhibit the lowest proportion of high-income earners, indicating the financial constraints faced by these workers. Category 7 (Craft Workers) and Category 8 (Machine Operators) have an intermediate income distribution, suggesting that skilled manual labor provides moderate earnings.

ref: https://ilostat.ilo.org/methods/concepts-and-definitions/classification-occupation/

Model Optimization-Mental-Physcial Model

Stepwise Akaike Information Criterion (AIC) selection is **STILL** used to iteratively remove the least significant predictors to find the best-performing model.

$$log\left(\frac{p}{1-p}\right) = \hat{\alpha} + \hat{\beta}_{Age} \cdot Age + \hat{\beta}_{Low-ducation} \cdot I_{Low-ducation}(x)$$
$$+\hat{\beta}_{Phycial-abor} \cdot I_{Phycial-Labor}(x) + \hat{\beta}_{Unmarried} \cdot I_{Unmarried}(x) + \hat{\beta}_{Male} \cdot I_{Male}(x) + \hat{\beta}_{hours\_PW} \cdot Hours\_PW$$

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.966040	0.445192	-6.662	2.69e-11	***
Age	0.037715	0.006356	5.934	2.96e-09	***
EducationLow_Education	-1.441793	0.349529	-4.125	3.71e-05	***
SexMale	0.585571	0.204248	2.867	0.00414	**
Hours_PW	0.030770	0.006887	4.468	7.89e-06	***
Marital_StatusUnmarried	-1.990257	0.187060	-10.640	< 2e-16	***
OccupationPhysical Labor	-1.320743	0.163077	-8.099	5.55e-16	***

The model shows that **older age, higher education, male status, longer working hours, married status, and brainwork occupations** are all significantly associated with higher income. For example, the coefficient of Education is negative and significant, indicating that higher education has a significant impact on higher income. A negative and significant manual labor coefficient indicates that it is more difficult for people who engage in manual labor to enter the higher income group than those who occupation physical_labor.

Model Optimization-PRC Model

Stepwise Akaike Information Criterion (AIC) selection is **STILL** used to iteratively remove the least significant predictors to find the best-performing model.

$$log\left(\frac{p}{1-p}\right) = \hat{\alpha} + \hat{\beta}_{Age} \cdot Age + \hat{\beta}_{Low-Education} \cdot I_{Low-Education}(x) + \sum_{i=1}^{6} \hat{\beta}_{Occupation,i} \cdot I_{Occupation,i}(x)$$
$$+\hat{\beta}_{Unmarried} \cdot I_{Unmarried}(x) + \hat{\beta}_{Male} \cdot I_{Male}(x) + \hat{\beta}_{hours\_PW} \cdot Hours\_PW$$

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.252612	0.510402	-6.373	1.86e-10	***
Age	0.040978	0.006535	6.270	3.60e-10	***
EducationLow_Education	-1.436365	0.346423	-4.146	3.38e-05	***
SexMale	0.488605	0.215606	2.266	0.0234	*
Hours_PW	0.035850	0.007234	4.956	7.19e-07	***
Marital_StatusUnmarried	-2.093813	0.195001	-10.737	< 2e-16	***
Occupation2	0.707103	0.253029	2.795	0.0052	**
Occupation3	-0.486564	0.319354	-1.524	0.1276	
Occupation4	-0.995424	0.244856	-4.065	4.80e-05	***
Occupation5	-2.613814	0.550179	-4.751	2.03e-06	***
Occupation6	-1.054506	0.237695	-4.436	9.15e-06	***

Age and Hours_PW have a significant positive impact on income, and the older the age and the longer the working hours, the easier it is to enter the high-income group. Both the Low_Education and Unmarried groups coefficients are negative and significant, indicating that low education and unmarried status significantly reduce the possibility of high income. Occupation2 has a significantly positive effect on income, while Occupation4, 5, and 6 are all significantly negative, indicating that Occupation2 is more likely to be of higher income than control occupations, while Occupation4, 5, and 6 are more closely associated with lower income.

Model Optimization-ISCO-08 Model

Stepwise Akaike Information Criterion (AIC) selection is **STILL** used to iteratively remove the least significant predictors to find the best-performing model.

$$log\left(\frac{p}{1-p}\right) = \hat{\alpha} + \hat{\beta}_{Age} \cdot Age + \hat{\beta}_{Low-Education} \cdot I_{Low-Education}(x) + \sum_{i=1}^{9} \hat{\beta}_{Occupation,i} \cdot I_{Occupation,i}(x)$$
$$+\hat{\beta}_{Unmarried} \cdot I_{Unmarried}(x) + \hat{\beta}_{Male} \cdot I_{Male}(x) + \hat{\beta}_{hours\_PW} \cdot Hours\_PW$$

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.262143	0.514533	-6.340	2.30e-10	***
Age	0.041171	0.006597	6.241	4.35e-10	***
EducationLow_Education	-1.431424	0.347718	-4.117	3.84e-05	***
SexMale	0.476856	0.216949	2.198	0.027948	*
Hours_PW	0.036079	0.007286	4.952	7.35e-07	***
Marital_StatusUnmarried	-2.094087	0.195332	-10.721	< 2e-16	***
Occupation2	0.730937	0.261676	2.793	0.005217	**
Occupation3	0.562765	0.487845	1.154	0.248675	
Occupation4	-0.489309	0.319379	-1.532	0.125506	
Occupation5	-0.999851	0.250967	-3.984	6.78e-05	***
Occupation6	-2.614898	0.550520	-4.750	2.04e-06	***
Occupation7	-0.931695	0.268166	-3.474	0.000512	***
Occupation8	-1.212247	0.291251	-4.162	3.15e-05	***
Occupation9	-0.951597	0.516129	-1.844	0.065224	.

The analysis results of other factors except occupation were similar to the previous two models.
Occupation 2 has a significantly positive impact (more of a higher income goal), while Occupation 5, 6, 7, 8, 9 are all significantly negative (more of a lower income goal);
Occupation 3 and 4 are not significant, indicating little difference between their occupationtime and the control group.

Models Summary

	df	AIC	Residual Deviance
Full Model	1319	1106.3	974.31
stepAIC Model	1351	1066.6	998.6
Mental-Physical Model	1378	1123.9	1109.9
PRC Model	1374	1101.5	1079.5
ISCO-08 Model	1371	1106.4	1078.4

The minimum AIC fit of the **PRC-model** best indicates that this occupation classification is desirable. To continue to improve prediction accuracy and account for potential interdependencies, we will continue to introduce interaction terms between variables in this logistic regression model. In this way, we can assess how a combination of factors affects the income classification (<=50,000 vs >50,000).

Model Optimization-Interaction Model

In order to improve prediction accuracy and take into account potential interdependence, we consider the interaction between two variables, introduce interaction terms into the model, and continue to use AIC iterative optimal model. There we show the outcome of the PRC$_6$Model with interaction:

$$log\left(\frac{p}{1-p}\right) = \hat{\alpha} + \hat{\beta}_{Age} \cdot Age + \hat{\beta}_{Low-Education} \cdot I_{Low-Education}(x) + \sum_{i=1}^{6} \hat{\beta}_{Occupation,i} \cdot I_{Occupation,i}(x) + \hat{\beta}_{Unmarried} \cdot I_{Unmarried}(x) + \hat{\beta}_{Male} \cdot I_{Male}(x) + \hat{\beta}_{hours\_PW} \cdot Hours\_PW + \hat{\beta}_{Male,Unmarried} \cdot I_{Male} \cdot I_{Unmarried}(x) + \sum_{i=1}^{6} \hat{\beta}_{Male.Occupation,i} \cdot I_{Male} \cdot I_{Occupation,i}(x)$$

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.703368	0.599795	-4.507	6.57e-06	***
Age	0.045978	0.006855	6.708	1.98e-11	***
EducationLow_Education	-1.513441	0.348787	-4.339	1.43e-05	***
SexMale	-0.708754	0.489645	-1.447	0.14776	
Hours_PW	0.040908	0.007421	5.513	3.53e-08	***
Marital_StatusUnmarried	-3.585524	0.449124	-7.983	1.42e-15	***
Occupation2	0.781470	0.570534	1.370	0.17078	
Occupation3	-1.266576	0.563022	-2.250	0.02447	*
Occupation4	-1.558491	0.662737	-2.352	0.01869	*
Occupation5	2.604163	1.505891	1.729	0.08375	.
Occupation6	-1.190624	0.729529	-1.632	0.10267	
SexMale:Marital_StatusUnmarried	1.956976	0.501180	3.905	9.43e-05	***
SexMale:Occupation2	0.032557	0.638509	0.051	0.95933	
SexMale:Occupation3	1.269024	0.704213	1.802	0.07154	.
SexMale:Occupation4	0.747952	0.713819	1.048	0.29472	
SexMale:Occupation5	-5.390792	1.625243	-3.317	0.00091	***
SexMale:Occupation6	0.283107	0.768359	0.368	0.71253	

Including interaction terms improves model interpretability and accuracy. Gender interacts with both Marital Status and Occupation, indicating income disparities linked to societal roles. ROC analysis suggests that models with interactions perform better than those without, validating the importance of capturing interdependencies between variables.

Model Check

To ensure the reliability and effectiveness of the logistic regression models, multiple validation techniques are applied, including coefficient visualization, ROC curve analysis, and ANOVA testing.

Coefficient Visualization

By visualizing the coefficients of the model, one can intuitively understand the direction and strength of the impact of each independent variable on the dependent variable. If the confi-dence interval of a variable lies entirely to the right of 1 (OR >1), it indicates a positive relationship between the variable and the dependent variable. If the confid-ence interval lies entirely to the left of 1 (OR < 1), it suggests a negative relationship. If the confid-ence interval spans across 1, the positive or negative effect cannot be significantly distinguished.

ROC Curve & AUC Values

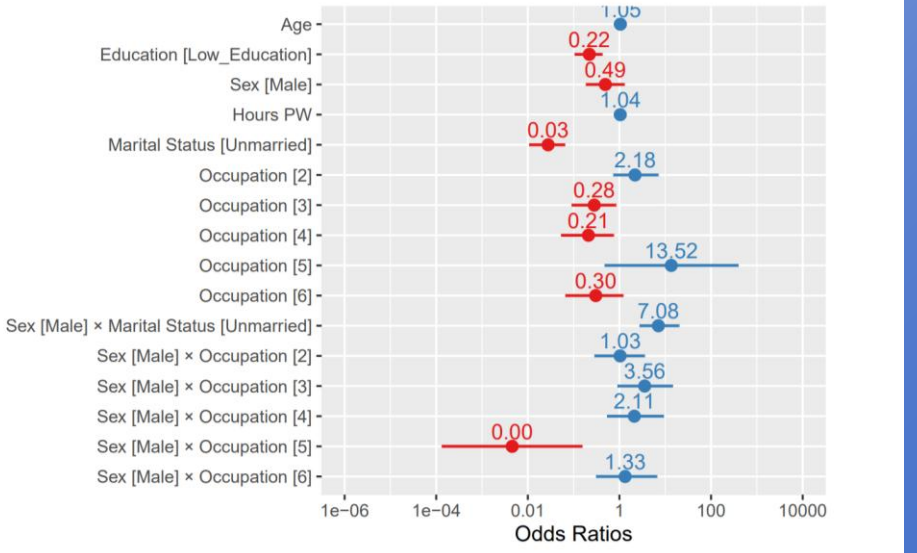
The ROC curve (Receiver Operating Characteristic curve) is plotted for each model to assess its classification performance. AUC (Area Under the Curve) values are calculated: A higher AUC (closer to 1) indicates a better-performing model. A lower AUC (closer to 0.5) suggests poor classification performance. The AUC values of different models are compared, helping to determine which occupational classification method improves predictive accuracy..

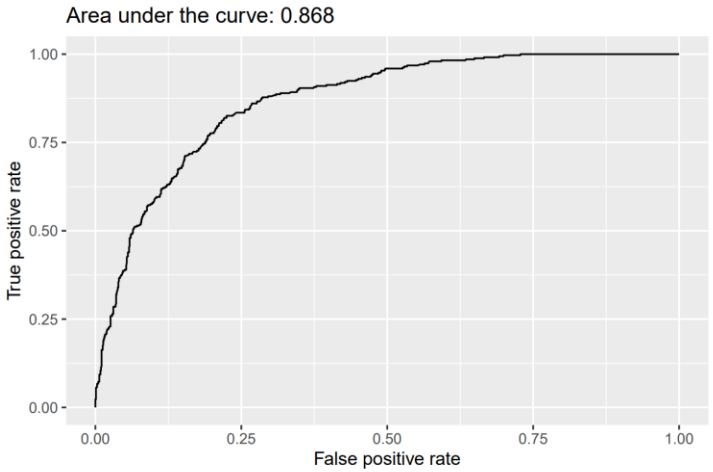
ANOVA

ANOVA Model Comparison ANOVA (Analysis of Variance) tests compare model fits: A significant p-value (< 0.05) indicates that additional predictors improve model performance. If models have similar p-values, a simpler model may be preferred to avoid overfitting. The results help in deciding whether stepAIC_model_new.1, stepAIC_model_new.2, or stepAIC_model_new.3 should be used for the final analysis.

Model Check

We interpret the odds ratios as follows: men's odds of higher income were 0.49 times those of women, second class Occupation's odds of higher income were 2.18 times those of first class Occupation, men's unmarried odds of higher income were 7.08 times those of women unmarried, and men in Occupation 2 were 1.03 time those women in Occupation 1. Finally, for each year increase in the individual age, their odds of higher income increase (by a factor of 1.05), for the increase in the individual age, their odds of higher income increase(by a factor of 1.04).





The AUC =0.868 is larger than the value of 0.5, this means that its classification ability is outstanding, able to correctly distinguish between positive and negative samples 86.8% of the time, far better than random classification (AUC = 0.5).

Model Check

	df	AIC	Residual Deviance	P-value of Hosmer-Lemeshow test
Full Model	1319	1106.3	974.31	0.898
stepAIC Model	1351	1066.6	998.6	0.8927
Mental-Physical Model	1378	1123.9	1109.9	0.4618
PRC Model	1374	1101.5	1079.5	0.5786
ISCO-08 Model	1371	1106.4	1078.4	0.2499
PRC interaction Model	1368	1084.7	1050.7	0.3321

The residual deviance of PRC interaction Model is 1050.7 which is lower than 1455.159 (X^2(1368,0.95)), the P-value if Hosmer-Lemeshow test(0.3321) is larger than 0.05, and the value of AIC is 1084.7 is the lowest. This means PRC interaction Model is well-fitted model.

Conclusion

Age: Higher age groups are more likely to achieve annual income >50k, reflecting the cumulative effect of seniority and work experience.

Higher education level: Having a college degree, bachelor's degree or above signi-ficantly increases the probability of high income.

.

Married status: Marriage generally leads to a more stable family or two sources of income, and is positively associated with higher income.

High-paying occupations: such as management, professional technology (Exec-managerial, Prof-specialty, etc.) are easier to cross the 50k threshold.

.

Longer work hours : An increase in the number of hours worked per week also increases the probability of a high income, but the nature of the occupation is equally critical.

Gender: Men are more likely to be in the >50k income group overall, but whether this is significant in the final model depends on the specific categorical combination and control variables.

Conclusion

- Older, higher education level, married, longer work hours, high-paying occupations: such as management, professional technology (Exec-managerial, Prof-specialty, etc.) groups are more likely to achieve annual income >50k.
- Men are more likely to be in the >50k income group overall, but whether this is significant in the final model depends on the specific categorical combination and control variables.
- Since most of the samples in the data are American nationality, and the sample size of other nationalities is very small, the overall impact on income is not influence.

Future Directions