Analysis of individual earns

Group 29

1 Introduction

Dataset 29 were collected from the US 1994 Census database. This dataset contain data on individuals regarding their income level, and various socioeconomic factors. You will have access to the following variables:

- **Age** - The age of the individual in years
- **Education** - Highest education level obtained by the individual
- **Marital_Status** - The marital status of the individual
- **Occupation** - The occupation of the individual
- **Sex** - The sex of the individual
- **Hours_Pw** - Number of hours worked per week by the individual
- **Nationality** - The nationality of birth of the individual
- **Income** - A factor variable with two levels: `>50k` if the individual earns more than $50k per year, or `<=50k` if the individual earns less than or equal to $50k per year

---

1.1 Task

Imagine you have been tasked by the government to identify which features based on the census data impact the income an individual makes: - Which factors influence whether an individual earns more than $50k per year?

You should:

1. Conduct an analysis to answer your question using a **Generalised Linear Model (GLM)**
2. Summarise your results in the form of a presentation

2 Exploratory Data Analysis

Summary statistics of **Income** are presented in the following for each column separately.

Observe the correlation of the numeric of data:

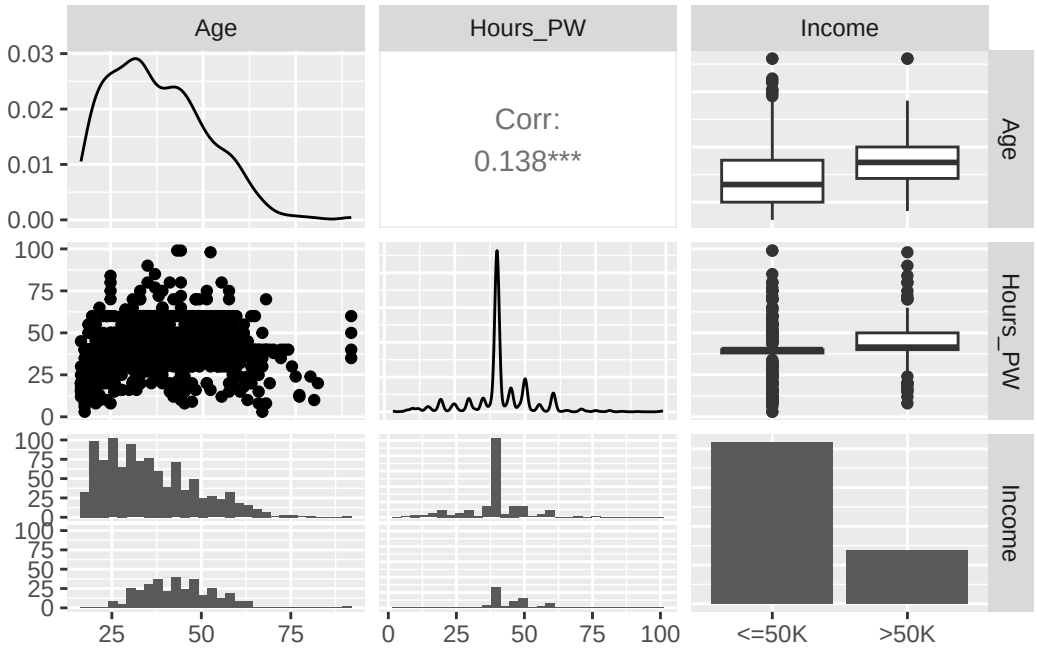


Figure 1: Numeric Data Correlation

2.1 Data levels

Observe the levels and correlations of categorical data:

```
 [1] "10th"         "11th"         "12th"         "1st-4th"      "5th-6th"
 [6] "7th-8th"      "9th"          "Assoc-acdm"   "Assoc-voc"    "Bachelors"
[11] "Doctorate"    "HS-grad"      "Masters"      "Prof-school"  "Some-college"
```

```
[1] "Divorced"              "Married-civ-spouse"    "Married-spouse-absent"
[4] "Never-married"         "Separated"             "Widowed"

 [1] "Adm-clerical"      "Craft-repair"      "Exec-managerial"
 [4] "Farming-fishing"   "Handlers-cleaners" "Machine-op-inspct"
 [7] "Other-service"     "Priv-house-serv"   "Prof-specialty"
[10] "Protective-serv"   "Sales"             "Tech-support"
[13] "Transport-moving"

[1] "Female" "Male"

 [1] "Cambodia"                  "Canada"
 [3] "China"                     "Columbia"
 [5] "Cuba"                      "Dominican-Republic"
 [7] "El-Salvador"               "England"
 [9] "France"                    "Germany"
[11] "Greece"                    "Guatemala"
[13] "Haiti"                     "India"
[15] "Iran"                      "Ireland"
[17] "Italy"                     "Jamaica"
[19] "Japan"                     "Laos"
[21] "Mexico"                    "Outlying-US(Guam-USVI-etc)"
[23] "Philippines"               "Poland"
[25] "Portugal"                  "Puerto-Rico"
[27] "Scotland"                  "South"
[29] "Taiwan"                    "Trinadad&Tobago"
[31] "United-States"             "Vietnam"

Chi-square test for Education vs  Income :

    Pearson's Chi-squared test

data:  contingency_table
X-squared = 193.07, df = 14, p-value < 2.2e-16


Chi-square test for Marital_Status vs  Income :

    Pearson's Chi-squared test

data:  contingency_table
```

```
X-squared = 263.24, df = 5, p-value < 2.2e-16


Chi-square test for Occupation vs  Income :

    Pearson's Chi-squared test

data:  contingency_table
X-squared = 167.57, df = 12, p-value < 2.2e-16


Chi-square test for Sex vs  Income :

    Pearson's Chi-squared test with Yates' continuity correction

data:  contingency_table
X-squared = 57.101, df = 1, p-value = 4.139e-14


Chi-square test for Nationality vs  Income :

    Pearson's Chi-squared test

data:  contingency_table
X-squared = 28.476, df = 31, p-value = 0.5965
```

2.2 Income Distribution

```
Income distribution:


<=50K  >50K
 1041   344
```

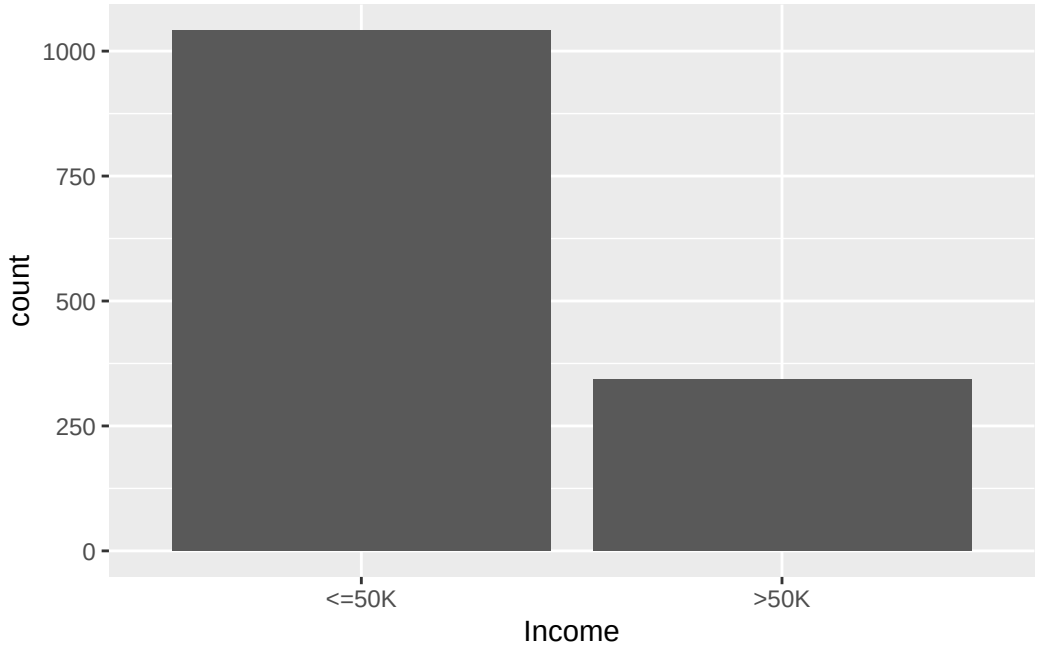


Figure 2: Income Distribution.

The income distribution analysis reveals that 75.2% of individuals earn <=50K, while only 24.8% earn >50K. This indicates a significant class imbalance, where the majority of the dataset consists of low-income individuals.

This imbalance can impact predictive modeling, as machine learning algorithms may become biased towards the dominant class (<=50K), leading to poor classification performance for high-income individuals. If a classification model is trained without addressing this issue, it may struggle to accurately predict the >50K category.

2.3 Occupation by Income

Income	Adm-clerical	Craft-repair	Exec-managerial	Farming-fishing
<=50K	13.5% (141)	14.8% (154)	9.3% (97)	3.6 (37)
>50K	7.0% (24)	13.1% (45)	23.0% (79)	1.5% (5)

Income	Handlers-cleaners	Machine-op-inspct	Other-service	Priv-house-serv
<=50K	5.3% (55)	7.9% (82)	14.1% (147)	0.4 (4)
>50K	1.7% (6)	4.1% (14)	0.6% (2)	0.0 (0)

Income	Prof-specialty	Protective-serv	Sales	Tech-support	Transport-moving
<=50K	8.8% (92)	2.0% (21)	12.1% (126)	2.3% (24)	5.9 (61)
>50K	25.3% (87)	2.9% (10)	12.5% (43)	3.8% (13)	4.7 (16)

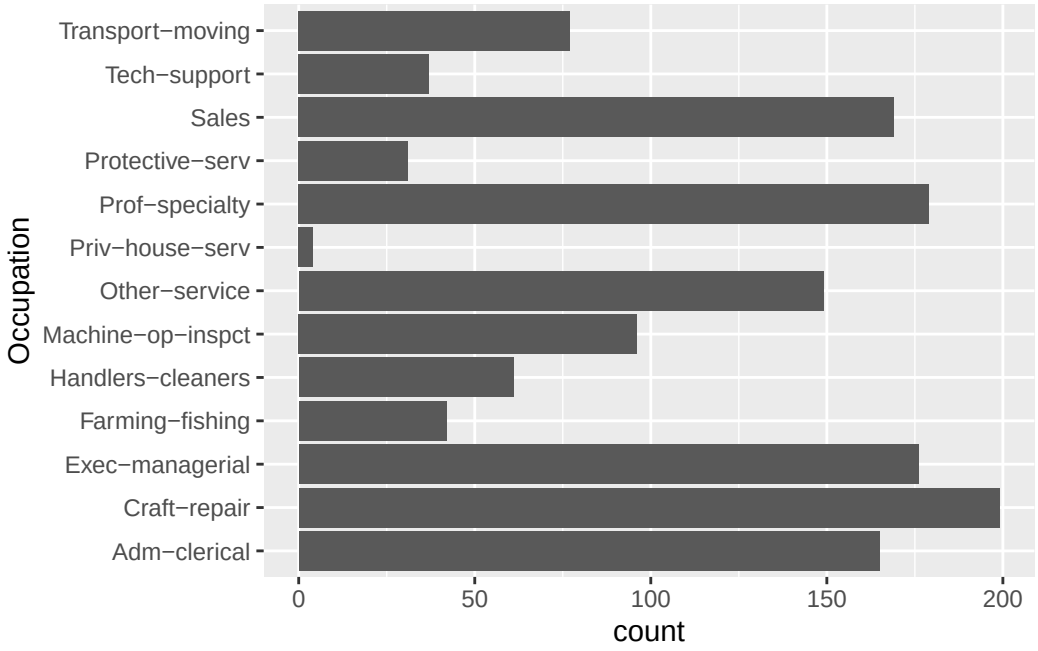


Figure 3: Occupation Distribution.

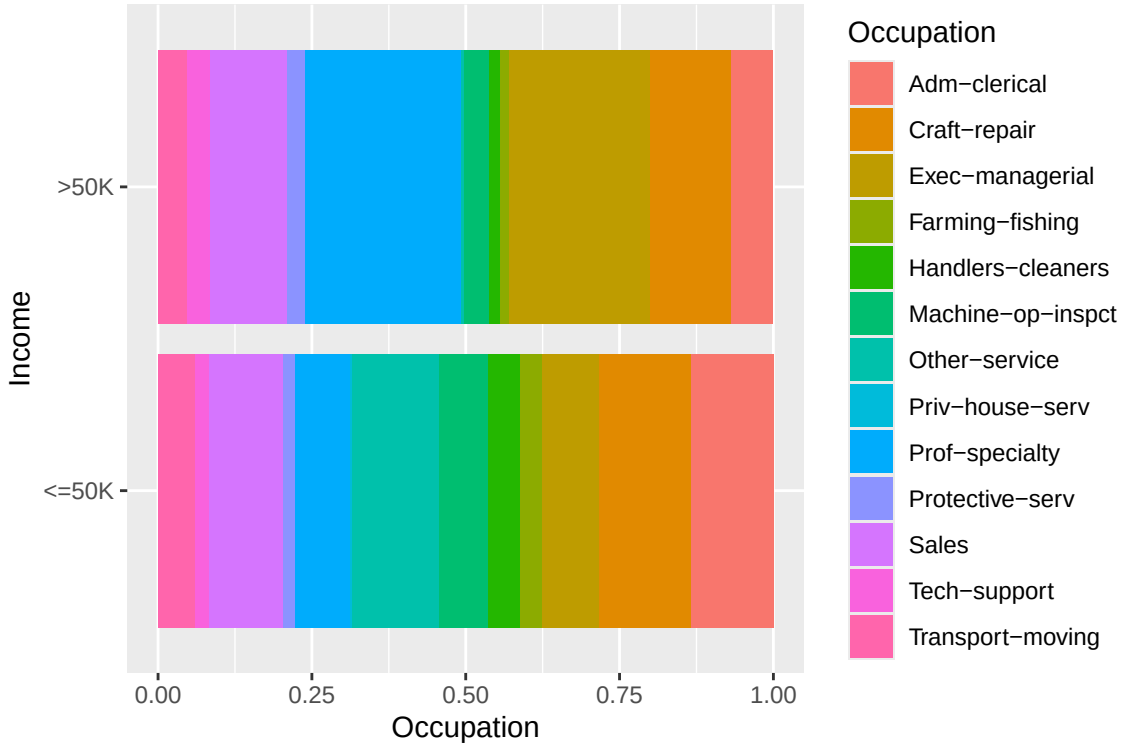


Figure 4: Income by Occupation.

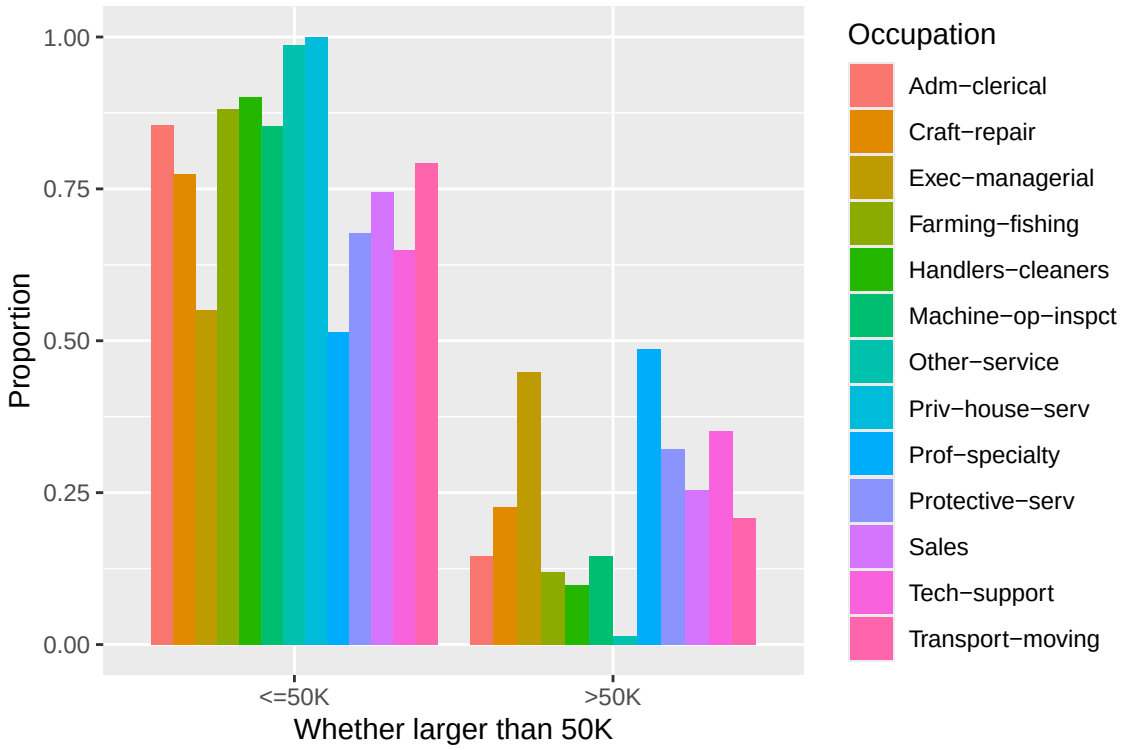


Figure 5: Income by Occupation.

The analysis of occupation by income shows significant variation across different job categories.

Administrative/Clerical roles: Approximately 85.6% of individuals in this category earn <=50K, while only 14.4% earn >50K, indicating that these jobs are predominantly lower-income. Craft and repair jobs: 76.4% of workers earn <=50K, with 23.6% earning >50K, suggesting a slightly better income distribution compared to clerical jobs. Executive/Managerial roles: This category has the highest proportion of high-income earners, with 44.9% earning >50K, showing that leadership positions significantly increase the likelihood of higher income. Farming/Fishing occupations: 88.1% of individuals earn <=50K, and only 11.9% earn >50K, reinforcing that agricultural jobs tend to be lower-paying. Handlers/Cleaners: The lowest income distribution, with 90.3% earning <=50K and only 9.7% earning >50K, indicating very limited access to higher salaries. This distribution highlights that occupational type is a key factor in determining income, with executive roles offering the highest proportion of high-income earners, while farming, cleaning, and clerical jobs predominantly fall into the lower-income category. These findings suggest that education, experience, and industry type play crucial roles in income disparities.

2.4 Education by Income

Income	10th	11th	12th	1st-4th	5th-6th	7th-8th	9th
<=50K	3.6% (37)	3.5% (36)	1.7% (18)	1.2% (12)	1.3% (14)	2.4% (25)	2.3(24)
>50K	0.9% (3)	0.9% (3)	0.0% (0)	0.0% (0)	0.0% (0)	0.6% (2)	0.9% (3)

Assoc-acdm	Assoc-voc	Bachelors	Doctorate	HS-grad	Masters	Prof-school
3.9% (41)	4.8% (50)	13.4% (139)	0.2% (2)	34.6% (360)	3.0% (31)	0.3% (3)
2.9% (10)	4.1% (14)	24.7% (85)	4.4% (15)	21.8% (75)	13.4% (46)	4.7(16)

Some-college
23.9 (249)
20.9% (72)

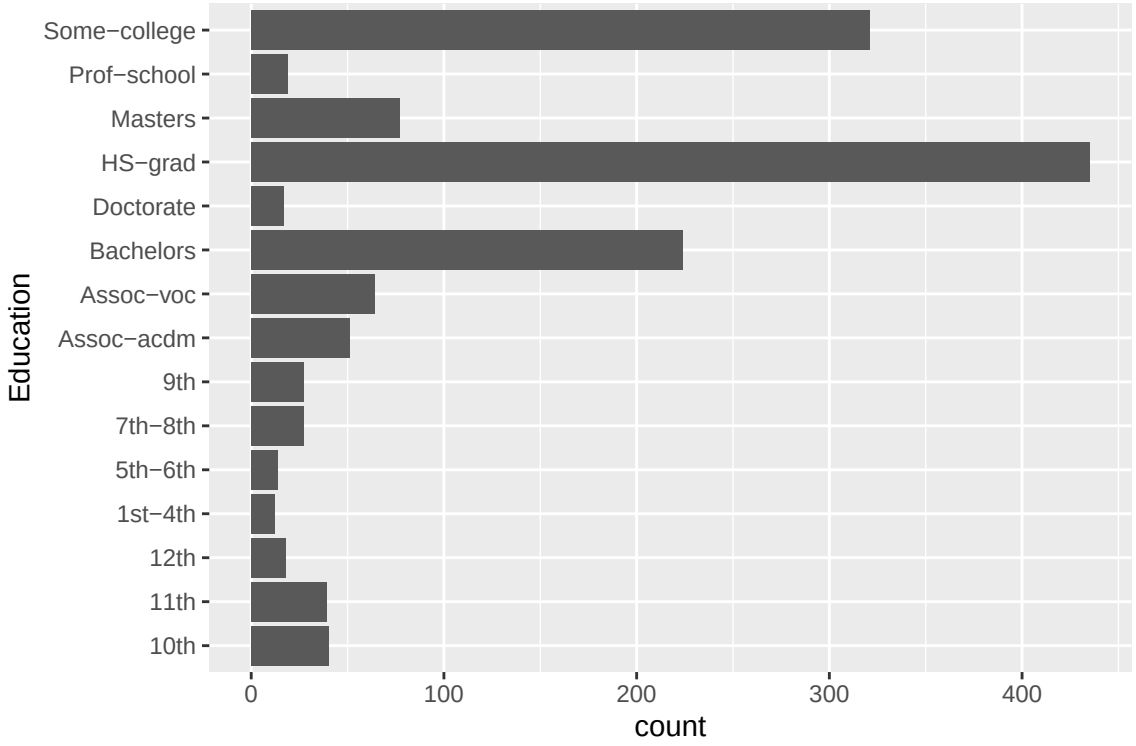


Figure 6: Education Distribution.

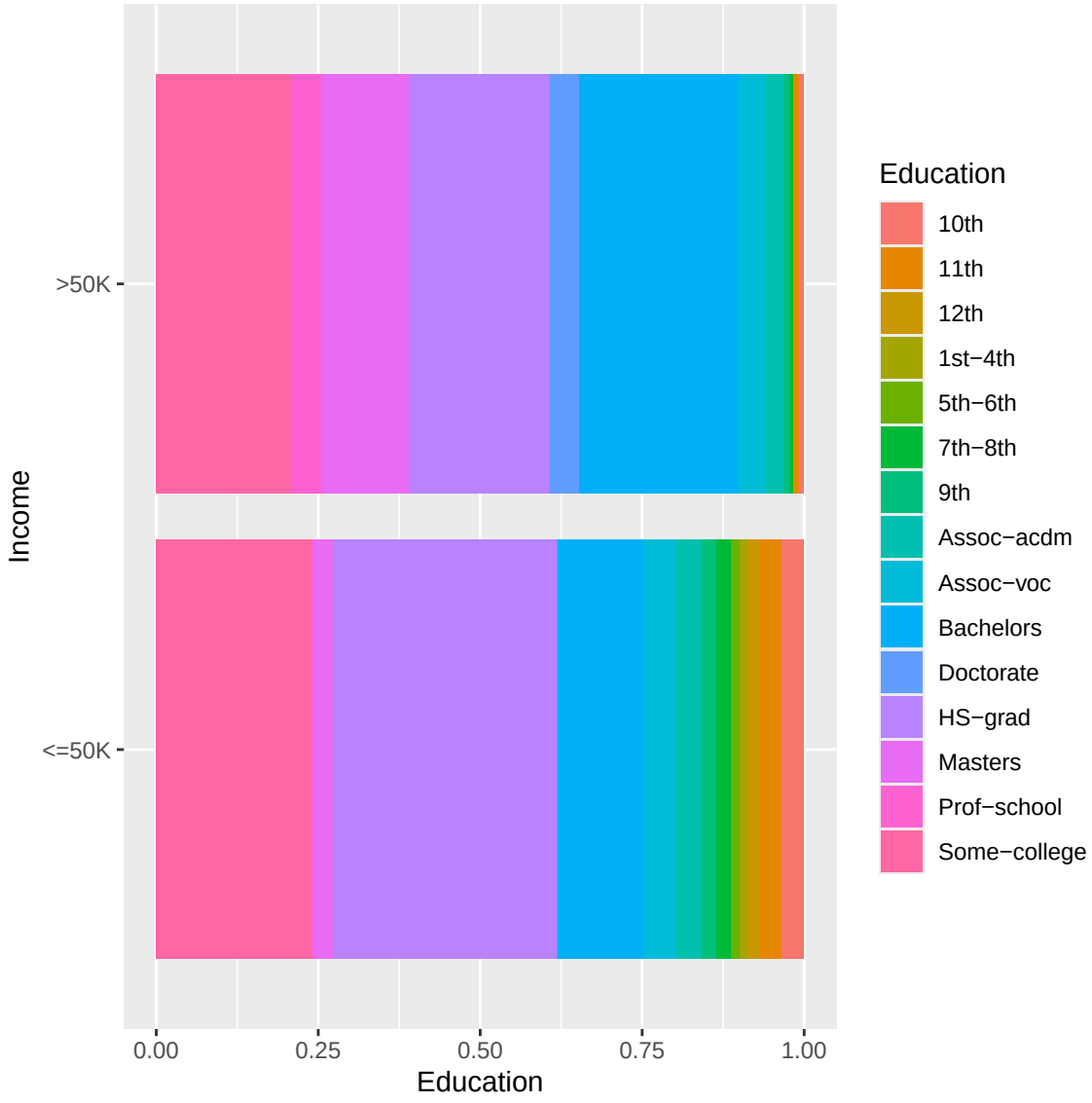


Figure 7: Income by Education.

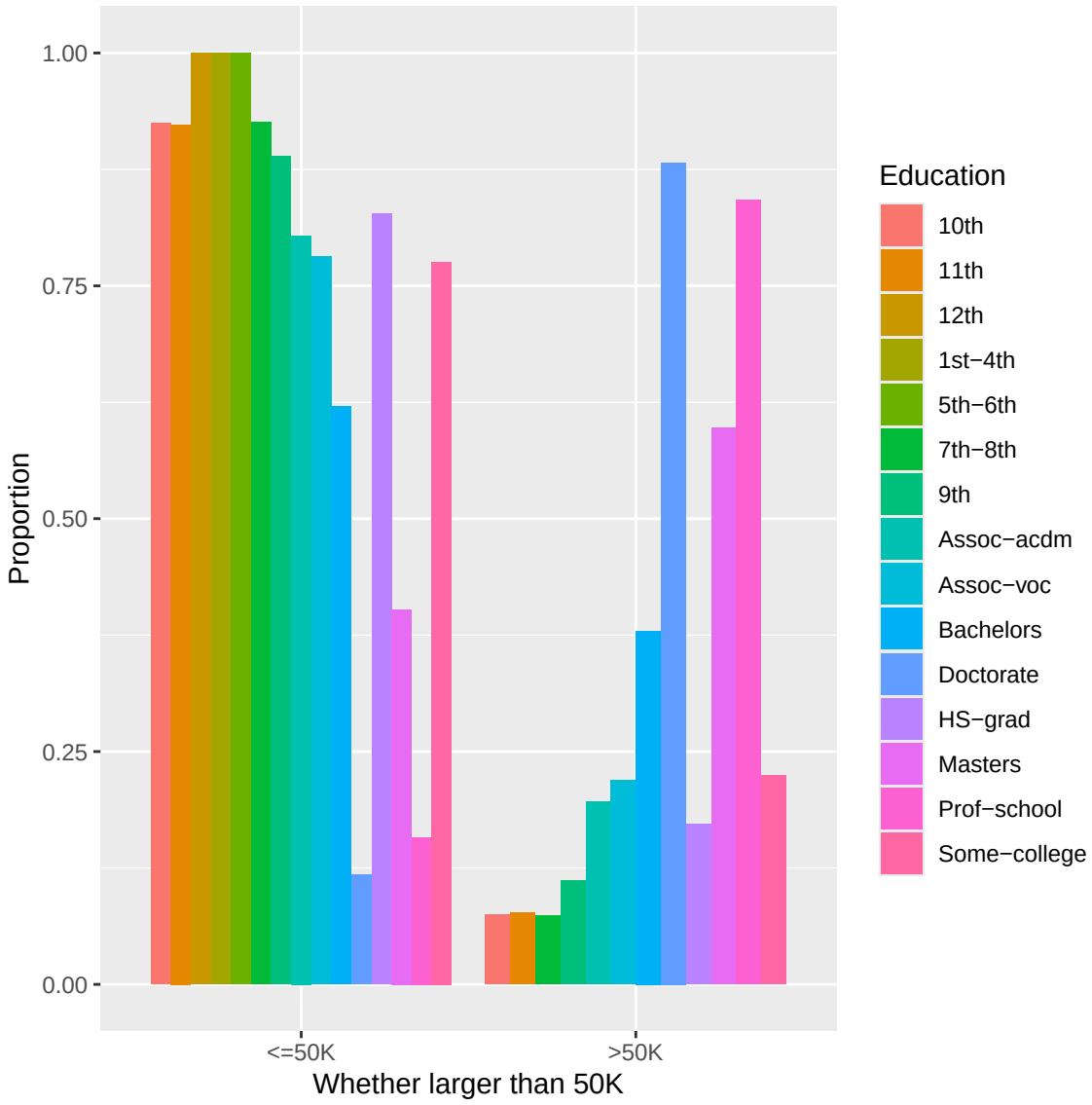


Figure 8: Income by Education.

The distribution of income across different education levels shows a clear correlation between higher education and increased income potential:

Lower education levels (1st-4th, 5th-6th, 12th grade):

100% of individuals in these categories earn <=50K, with no individuals earning >50K. This indicates that individuals with minimal education have almost no access to high-income jobs. 10th and 11th grade education:

93.8% of individuals with a 10th-grade education earn <=50K, while only 6.3% earn >50K. 92.9% of individuals with an 11th-grade education earn <=50K, with 7.1% in the high-income group. Although slightly better than primary education, these groups still have a very low likelihood of earning above 50K.

Key Takeaways: Education level is a strong predictor of income. Individuals without a high school diploma have almost no chance of reaching the high-income category. These findings highlight the importance of higher education in increasing earning potential and access to well-paying jobs.

2.5 Martial_Status by Income

Table 1: Martial_Status Distribution.

Income	Divorced	Married-civ-spouse	Married-spouse-absent	Never-married	Separated	Widowed
<=50K	15.1% (157)	34.5% (359)	1.8% (19)	42.3% (440)	3.9% (41)	2.4% (25)
>50K	7.3% (25)	84.3% (290)	0.3% (1)	7.3% (25)	0.3% (1)	0.6% (2)

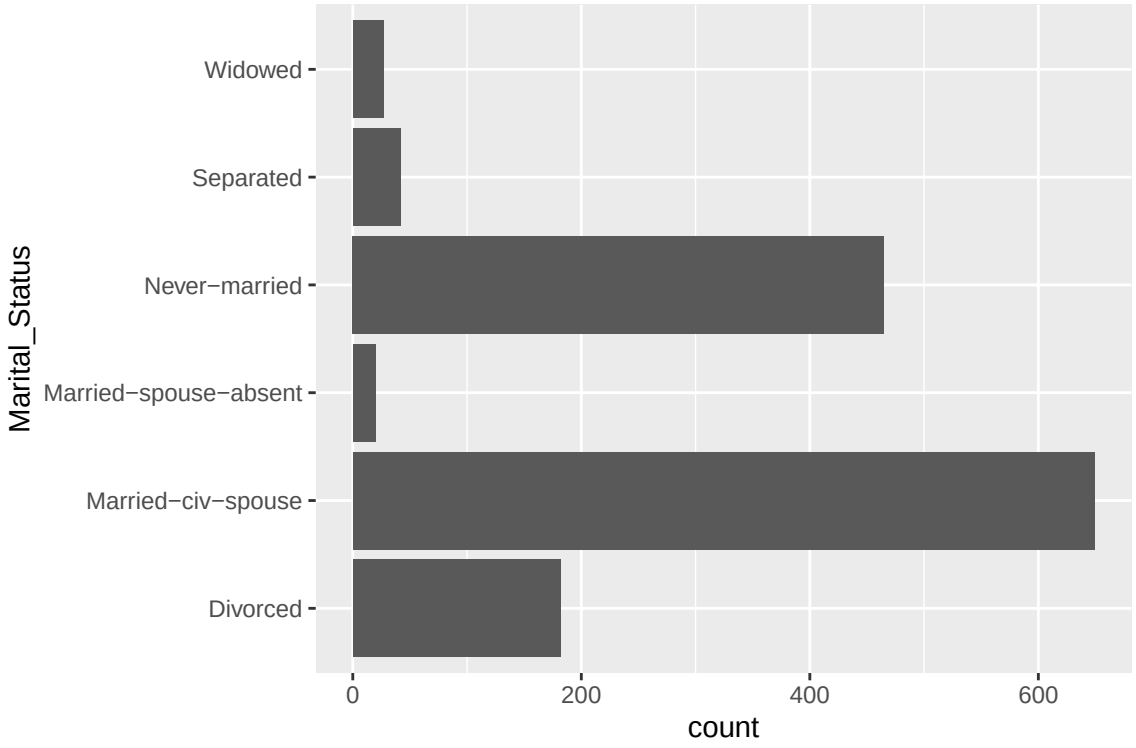


Figure 9: Marital_Status Distribution.

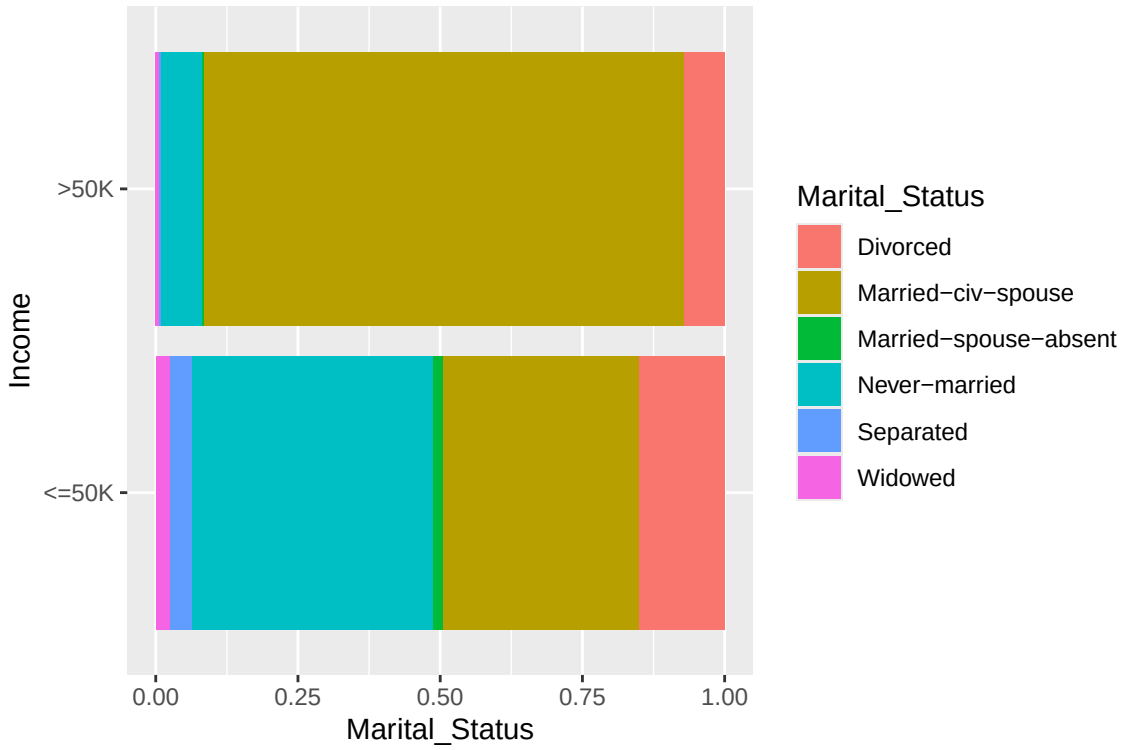


Figure 10: Income by Marital_Status.

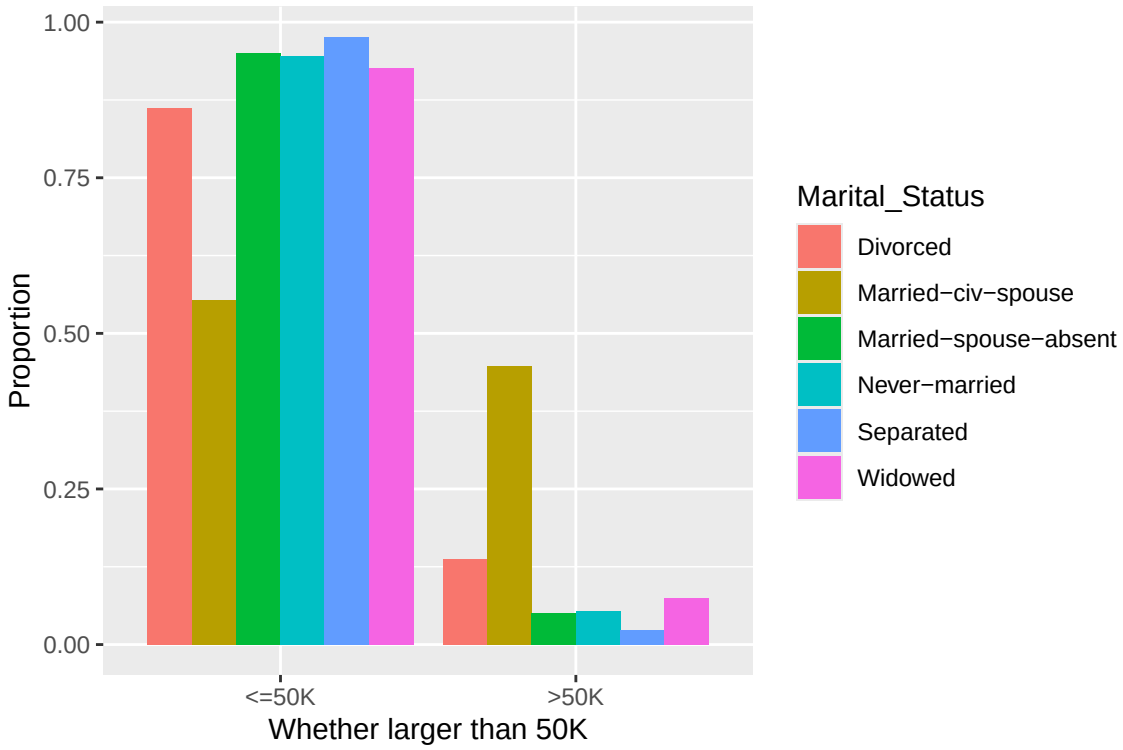


Figure 11: Income by Marital_Status.

The distribution of income by marital status reveals a strong correlation between marriage and income level:

Married individuals (civil spouse):

44.3% earn >50K, the highest proportion among all marital statuses. This suggests that marriage (especially in stable relationships) is associated with higher earnings, possibly due to dual-income households or increased financial stability. Divorced individuals:

85.8% earn <=50K, while 14.2% earn >50K. Although lower than married individuals, this group has a higher proportion of high-income earners than single or separated individuals. Never-married and separated individuals:

95.5% of never-married individuals and 97.9% of separated individuals earn <=50K, with only a very small fraction entering the >50K category. This indicates that single individuals are more likely to belong to lower-income groups, possibly due to younger age demographics and lower accumulated financial resources. Key Takeaways: Being married is strongly associated with higher income levels. Single and separated individuals are less likely to earn above 50K, which may reflect differences in career stability, accumulated wealth, or household income. This analysis suggests that marital status can play a significant role in economic well-being.

2.6 Sex by Income

Table 2: Sex Distribution.

Income	Female	Male
<=50K	37.2% (387)	62.8% (654)
>50K	15.1% (52)	84.9% (292)

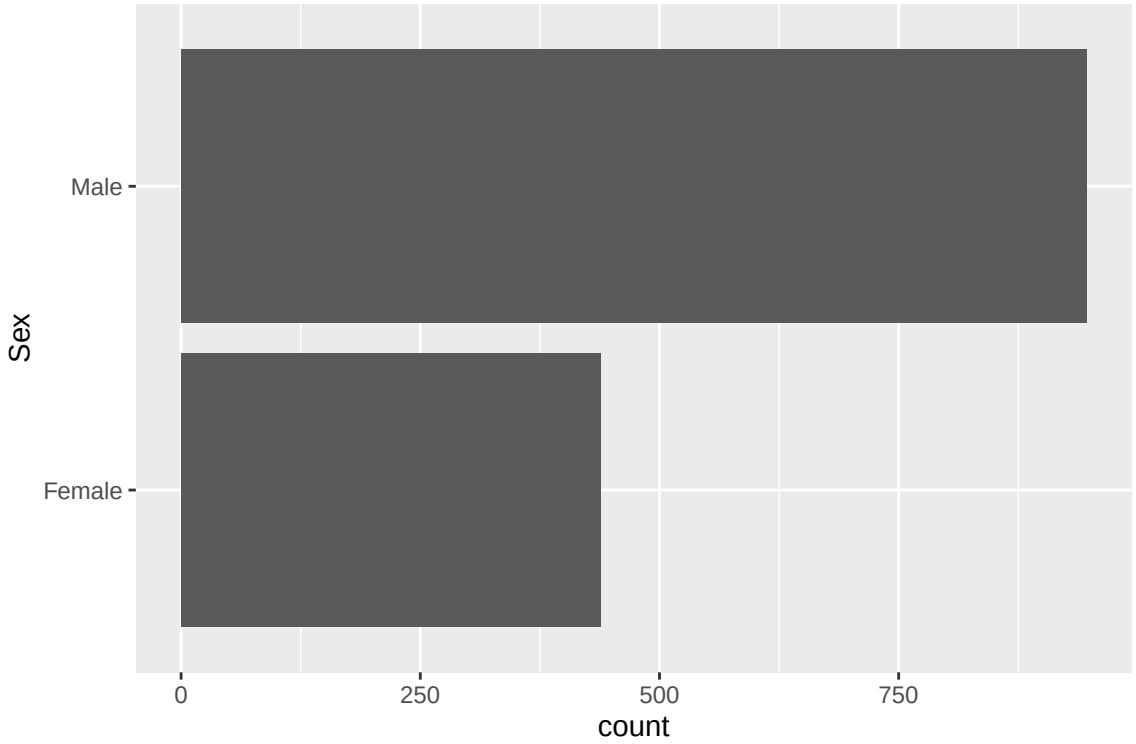


Figure 12: Sex Distribution.

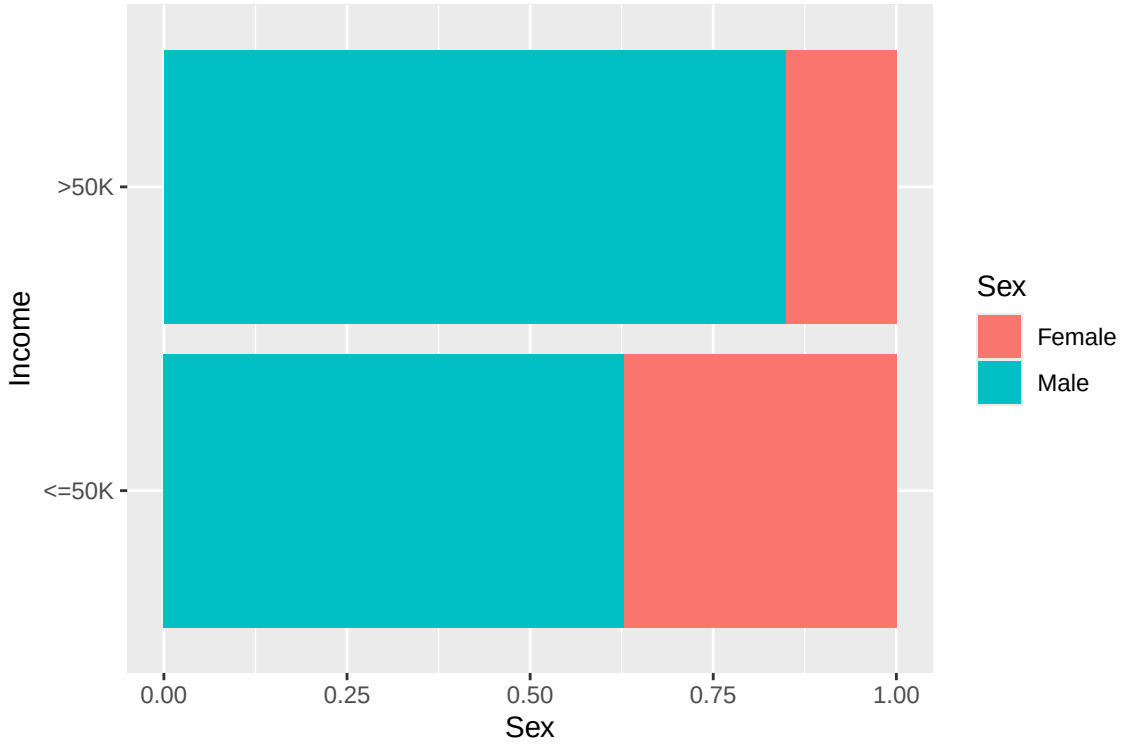


Figure 13: Income by Sex.

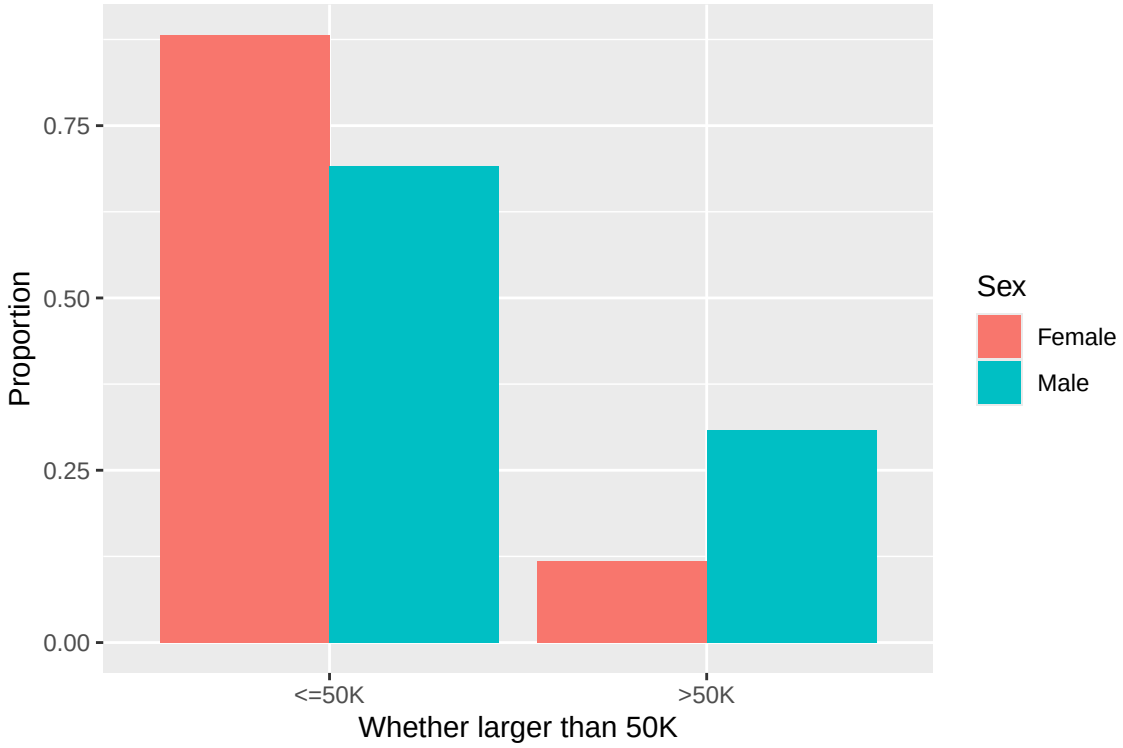


Figure 14: Income by Sex.

The distribution of income by gender reveals a significant disparity between males and females:

Females:

88.2% earn <=50K, while only 11.8% earn >50K. This indicates that women are much less likely to be in the high-income category. Males:

69.6% earn <=50K, whereas 30.4% earn >50K. The proportion of males earning above 50K is significantly higher than that of females. Key Takeaways: Men are significantly more likely to earn >50K compared to women. The income gap suggests possible gender disparities in wages, career advancement, or occupational roles. These findings highlight the need to investigate underlying causes, such as differences in industries, job roles, or social factors affecting earnings.

2.7 Nationality by Income

Income	Cambodia	Canada	China	Columbia	Cuba	Dominican-Republic
<=50K	0.1% (1)	0.5% (5)	0.3% (3)	0.1% (1)	0.5% (5)	0.4 (4)

```
  >50K 0.3% (1) 0.9% (3) 0.3% (1) 0.0% (0) 0.3% (1)           0.0      (0)
El-Salvador  England   France  Germany   Greece Guatemala    Haiti    India
   0.5% (5) 0.2% (2) 0.1% (1) 0.4% (4) 0.2% (2)  0.5% (5) 0.1% (1) 0.3 (3)
   0.0% (0) 0.0% (0) 0.0% (0) 0.9% (3) 0.3% (1)  0.0% (0) 0.0% (0) 0.0 (0)
    Iran  Ireland    Italy  Jamaica    Japan     Laos    Mexico
0.1% (1) 0.1% (1) 0.2% (2) 0.2% (2) 0.2% (2) 0.2% (2) 2.8             (29)
0.0% (0) 0.3% (1) 0.0% (0) 0.0% (0) 0.3% (1) 0.0% (0) 0.3%  (1)
Outlying-US(Guam-USVI-etc) Philippines   Poland Portugal Puerto-Rico Scotland
                  0.1% (1)    0.7% (7) 0.3% (3) 0.1% (1)    0.5% (5) 0.1(1)
                  0.0% (0)    0.6% (2) 0.3% (1) 0.3% (1)    0.6% (2) 0.0(0)
   South   Taiwan Trinadad&Tobago United-States  Vietnam
0.3% (3) 0.1% (1)        0.1% (1)   89.5% (932) 0.5                   (5)
0.0% (0) 0.6% (2)        0.0% (0)   93.9% (323) 0.0                   (0)
```

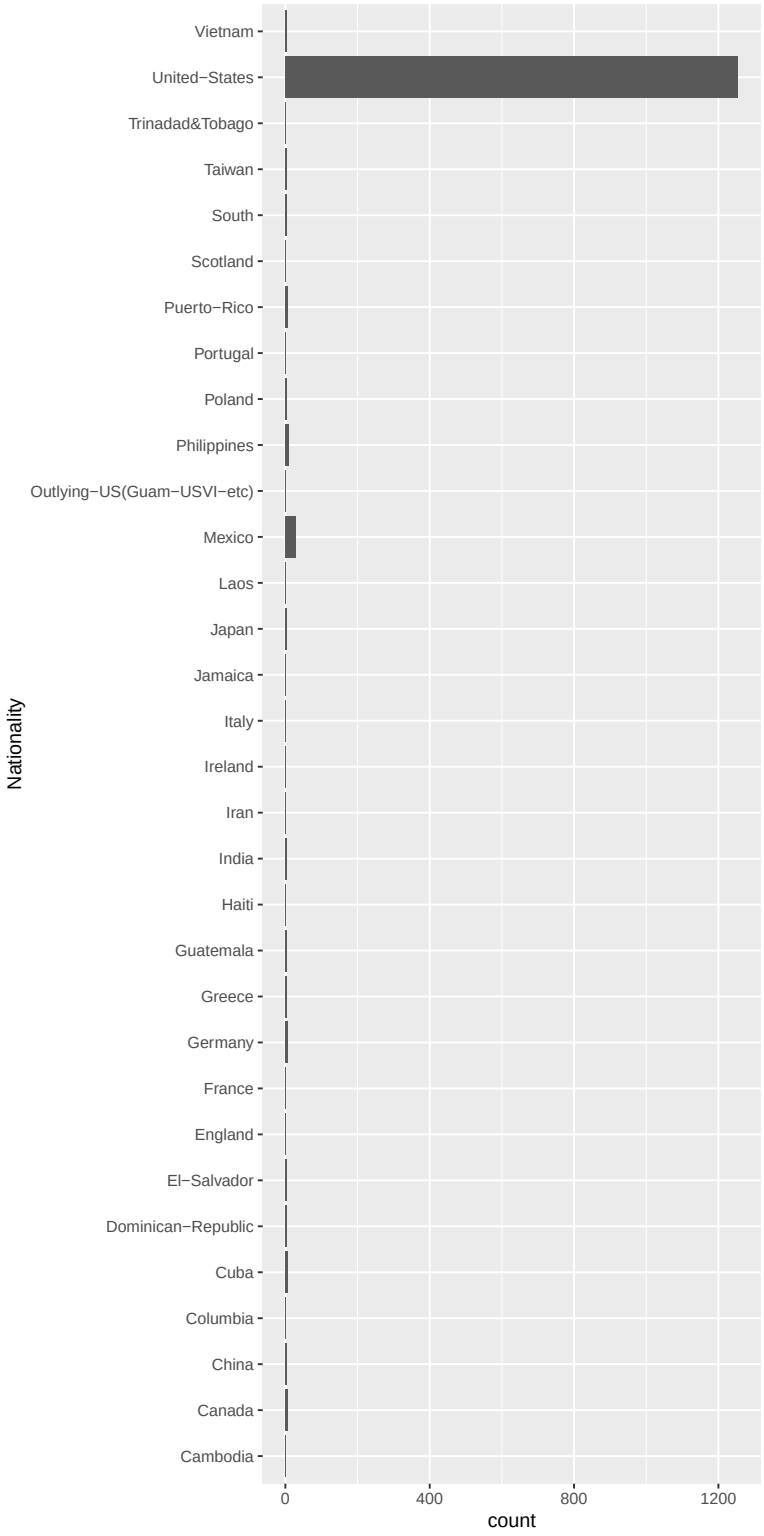


Figure 15: Nationality Distribution.

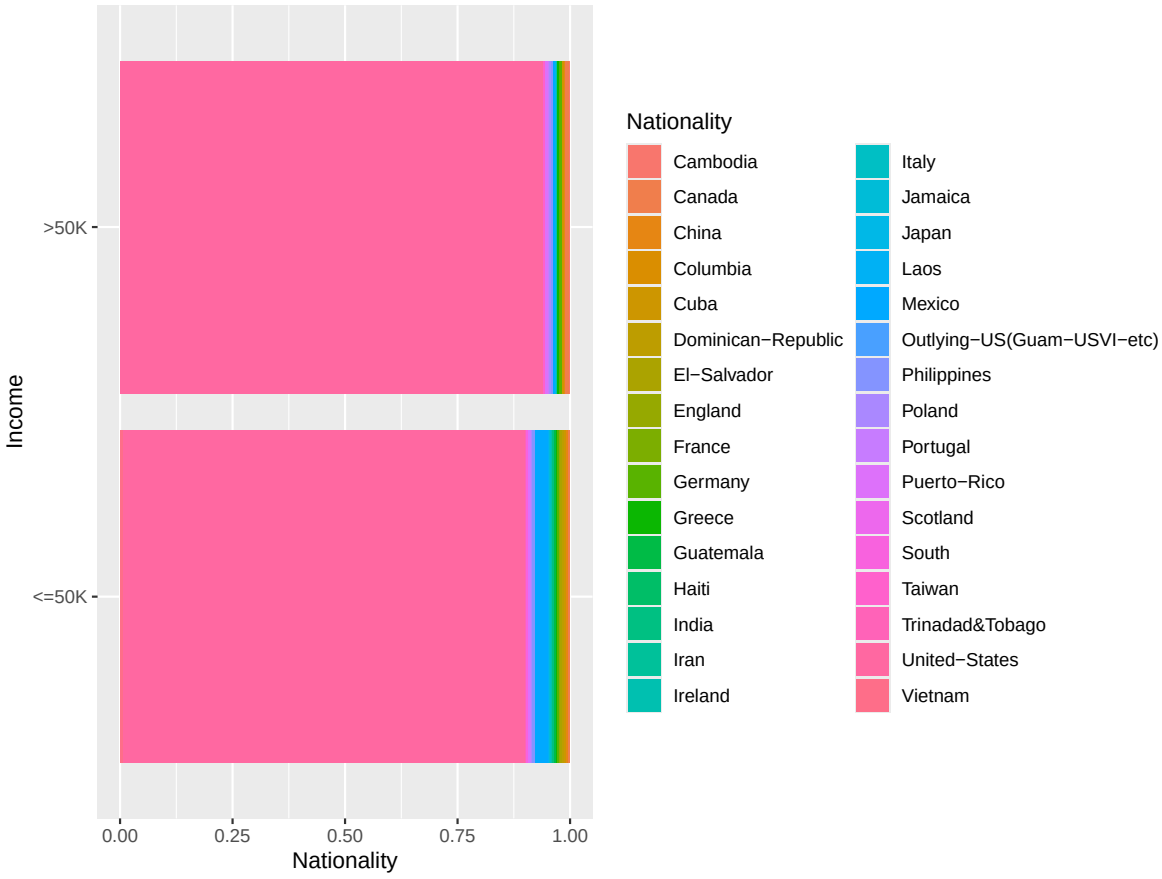


Figure 16: Income by Nationality.

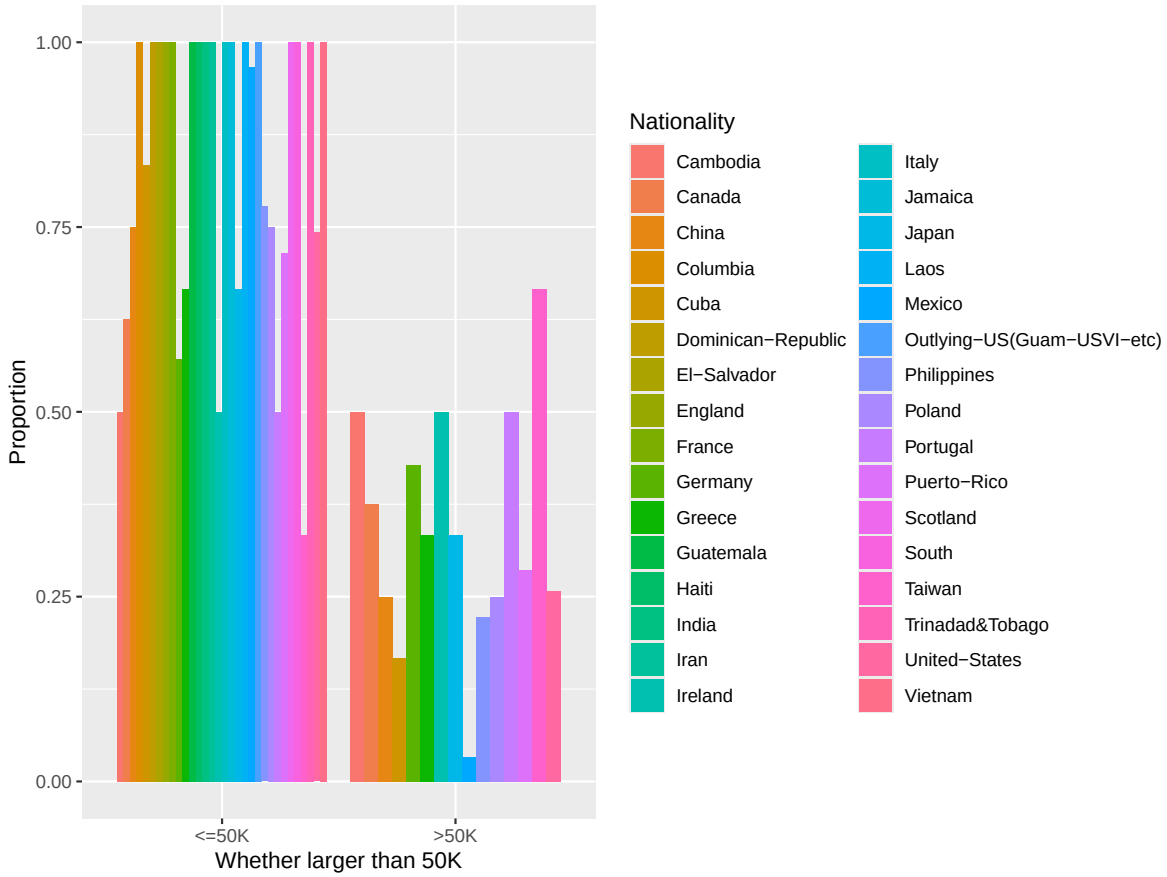


Figure 17: Income by Nationality.

The income distribution across different nationalities varies significantly:

Cambodia:

50% earn <=50K, and 50% earn >50K. This is the most balanced distribution among the nationalities observed. Canada:

62.5% earn <=50K, while 37.5% earn >50K. A relatively high proportion of individuals from Canada belong to the high-income group. China:

75% earn <=50K, and 25% earn >50K. While some individuals reach the high-income group, the majority remain below 50K. Colombia:

100% earn <=50K, with no individuals in the >50K category. This suggests that people from this nationality are more likely to be in lower-income jobs. Cuba:

83.3% earn <=50K, while 16.7% earn >50K. The majority still earn less than 50K, but a small proportion reaches higher income levels. Key Takeaways: Nationality appears to be a factor influencing income distribution. Some nationalities, such as Canada and Cambodia, have higher proportions of high-income individuals, while others, like Colombia and Cuba, have a stronger concentration in the low-income group. These differences could be attributed to factors such as job opportunities, skill levels, or immigration status affecting income potential.

However, in the dataset, there are 1255 samples with American nationality, accounting for 90.7% of the total. Due to this highly imbalanced distribution, nationality is not suitable as an explanatory variable.

2.8 Age by Income

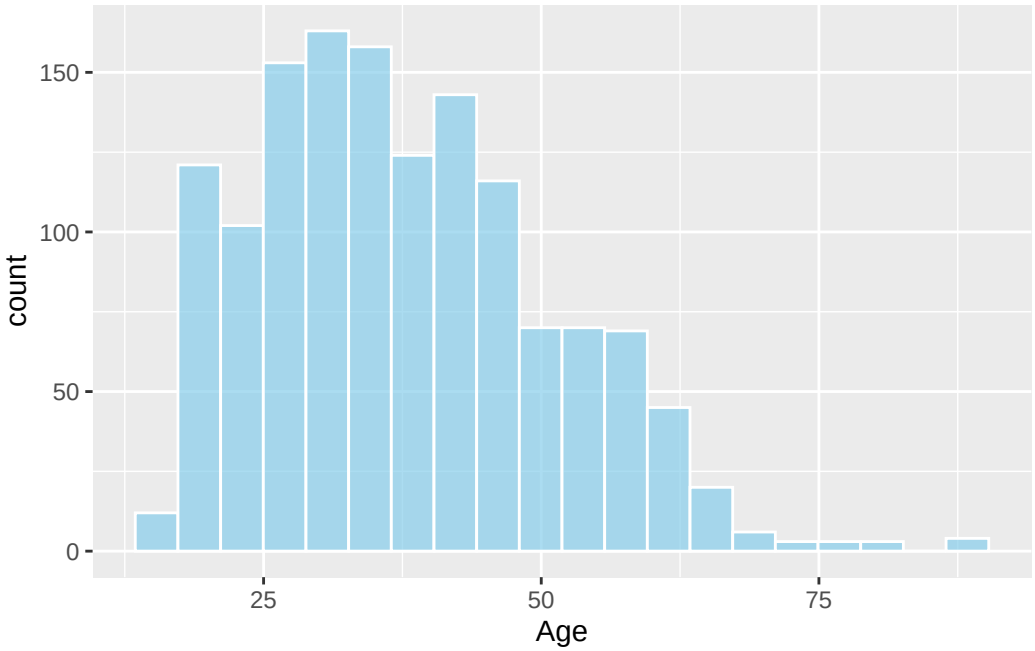


Figure 18: Age Distribution.

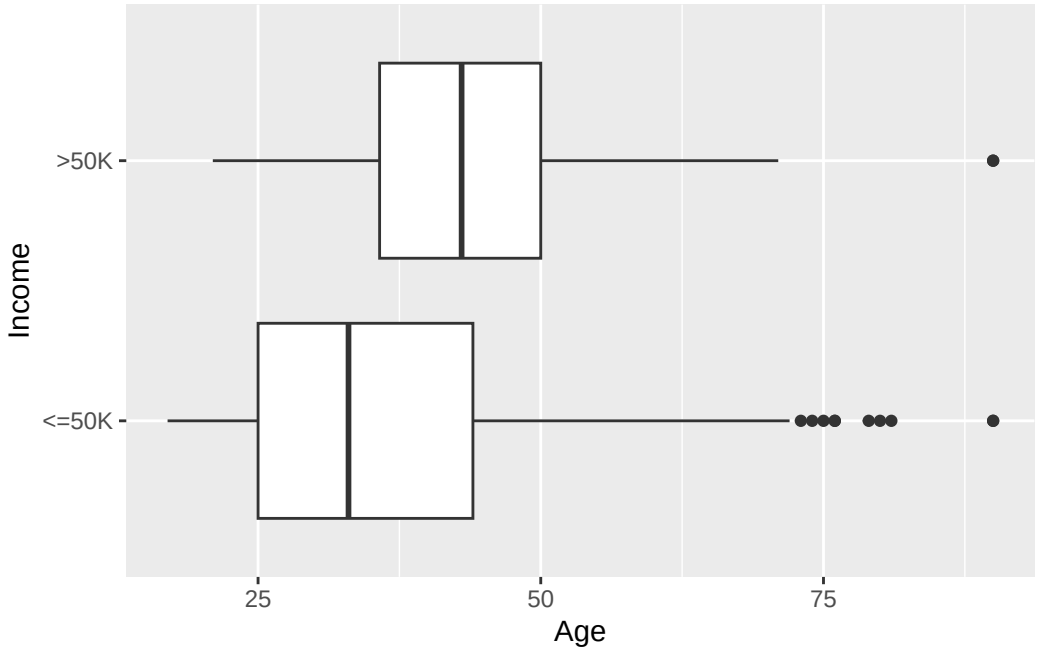


Figure 19: Income by Age.

The analysis of age distribution and its relationship to income reveals the following trends:

Age Distribution:

The dataset contains individuals aged 17 to 90 years old. The age distribution appears roughly normal, with most individuals concentrated in their 30s and 40s. Income and Age Statistics:

Low-income group (<=50K): Average age: 36.1 years 25th percentile: 25 years 50th percentile (median): 33 years 75th percentile: 45 years High-income group (>50K): Average age: 43.9 years 25th percentile: 35 years 50th percentile (median): 43 years 75th percentile: 51 years Key Takeaways: Older individuals are more likely to be in the high-income group. The median age of high-income earners (43 years) is significantly higher than that of low-income earners (33 years). This suggests that experience, seniority, and career progression contribute to higher earnings over time. Younger individuals tend to belong to the low-income group. This is likely because they are in the early stages of their careers, earning entry-level salaries.

2.9 Hours_pw by Income

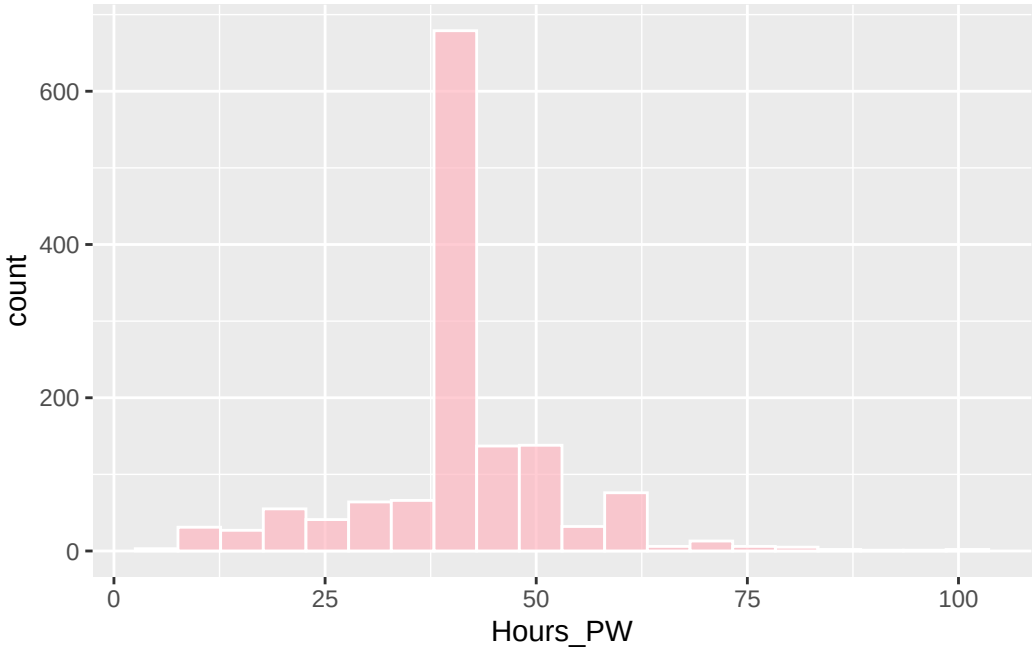


Figure 20: Hours_PW Distribution.

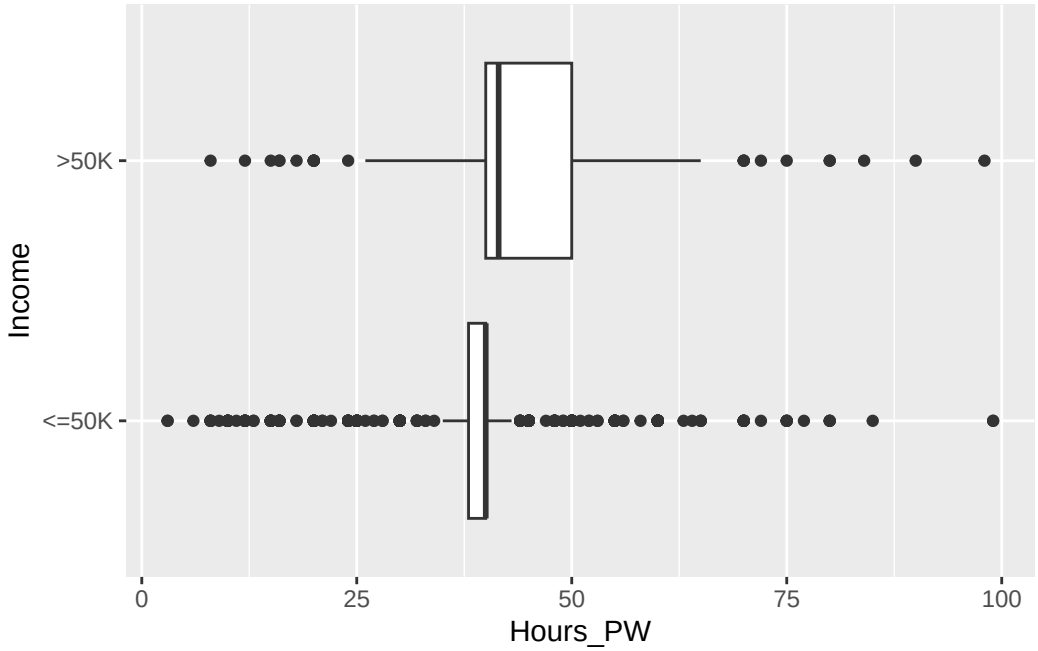


Figure 21: Income by Hours_PW.

The analysis of weekly work hours and its relationship to income reveals key differences between low and high earners:

Work Hours Distribution:

The dataset contains a variety of work hours, with most individuals working around 40 hours per week. The distribution appears right-skewed, meaning some individuals work significantly longer hours. Work Hours Statistics by Income:

Low-income group (<=50K): The most common work hours are 40 hours per week, with 546 individuals working this amount. There are 58 unique work hour values, indicating a wide range of working schedules. High-income group (>50K): The most common work hours are also 40 hours per week, but only 142 individuals in this category work this amount. There are 39 unique work hour values, suggesting slightly less variability in working hours compared to low-income earners. Key Takeaways: Most individuals work 40 hours per week, regardless of income level. High-income earners tend to work more stable hours, while low-income earners exhibit greater variation in work schedules. Longer working hours do not necessarily guarantee higher income, suggesting that job type and skill level are more influential factors in determining earnings.

3 Model Selection

fitting full model.

```
Implementation: ROI | Solver: lpsolve
Separation: TRUE
Existence of maximum likelihood estimates
                          (Intercept)                                   Age
                                    0                                     0
                         Education11th                         Education12th
                                    0                                  -Inf
                     Education1st-4th                      Education5th-6th
                                 -Inf                                  -Inf
                     Education7th-8th                          Education9th
                                    0                                     0
                   EducationAssoc-acdm                   EducationAssoc-voc
                                    0                                     0
                    EducationBachelors                   EducationDoctorate
                                    0                                     0
                      EducationHS-grad                     EducationMasters
                                    0                                     0
                  EducationProf-school                EducationSome-college
                                    0                                     0
                               SexMale                              Hours_PW
                                    0                                     0
      Marital_StatusMarried-civ-spouse   Marital_StatusMarried-spouse-absent
                                    0                                     0
           Marital_StatusNever-married                Marital_StatusSeparated
                                    0                                     0
                 Marital_StatusWidowed                OccupationCraft-repair
                                    0                                     0
            OccupationExec-managerial              OccupationFarming-fishing
                                    0                                     0
          OccupationHandlers-cleaners           OccupationMachine-op-inspct
                                    0                                     0
              OccupationOther-service              OccupationPriv-house-serv
                                    0                                  -Inf
             OccupationProf-specialty              OccupationProtective-serv
                                    0                                     0
                       OccupationSales                 OccupationTech-support
                                    0                                     0
           OccupationTransport-moving                      NationalityCanada
```

```
                                0                                 0
                NationalityChina                 NationalityColumbia
                                0                              -Inf
                 NationalityCuba       NationalityDominican-Republic
                                0                              -Inf
           NationalityEl-Salvador                  NationalityEngland
                             -Inf                              -Inf
                NationalityFrance                  NationalityGermany
                             -Inf                                 0
                NationalityGreece                NationalityGuatemala
                                0                              -Inf
                 NationalityHaiti                    NationalityIndia
                             -Inf                              -Inf
                  NationalityIran                   NationalityIreland
                             -Inf                                 0
                 NationalityItaly                   NationalityJamaica
                             -Inf                              -Inf
                 NationalityJapan                     NationalityLaos
                                0                              -Inf
                NationalityMexico NationalityOutlying-US(Guam-USVI-etc)
                                0                              -Inf
           NationalityPhilippines                    NationalityPoland
                                0                                 0
              NationalityPortugal               NationalityPuerto-Rico
                                0                                 0
              NationalityScotland                    NationalitySouth
                             -Inf                              -Inf
                NationalityTaiwan           NationalityTrinadad&Tobago
                                0                              -Inf
         NationalityUnited-States                   NationalityVietnam
                                0                              -Inf
0: finite value, Inf: infinity, -Inf: -infinity


Call:
glm(formula = Income ~ Age + Education + Sex + Hours_PW + Marital_Status +
    Occupation + Nationality, family = binomial(link = "logit"),
    data = data)

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -4.710e+00  2.198e+00  -2.142 0.032163
```

Age	3.840e-02	7.626e-03	5.036	4.76e-07
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
SexMale	1.914e-01	2.490e-01	0.769	0.442085
Hours_PW	2.978e-02	7.914e-03	3.762	0.000168
Marital_StatusMarried-civ-spouse	1.936e+00	2.892e-01	6.695	2.15e-11
Marital_StatusMarried-spouse-absent	2.740e-01	1.158e+00	0.237	0.813013
Marital_StatusNever-married	-5.338e-01	3.586e-01	-1.489	0.136554
Marital_StatusSeparated	-1.594e+00	1.068e+00	-1.493	0.135565
Marital_StatusWidowed	-8.949e-01	8.426e-01	-1.062	0.288201
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
NationalityCanada	-1.789e+00	2.313e+00	-0.773	0.439233
NationalityChina	-2.918e+00	3.035e+00	-0.961	0.336342
NationalityColumbia	-1.680e+01	6.523e+03	-0.003	0.997945
NationalityCuba	-1.802e+00	2.365e+00	-0.762	0.445958
NationalityDominican-Republic	-1.660e+01	2.909e+03	-0.006	0.995448
NationalityEl-Salvador	-1.838e+01	2.495e+03	-0.007	0.994122
NationalityEngland	-1.812e+01	4.607e+03	-0.004	0.996862
NationalityFrance	-1.804e+01	6.523e+03	-0.003	0.997793
NationalityGermany	-1.414e+00	2.206e+00	-0.641	0.521713

```
NationalityGreece                       -1.153e+00  3.135e+00  -0.368 0.713110
NationalityGuatemala                    -1.548e+01  2.162e+03  -0.007 0.994285
NationalityHaiti                        -1.707e+01  6.523e+03  -0.003 0.997912
NationalityIndia                        -2.097e+01  3.539e+03  -0.006 0.995273
NationalityIran                         -1.884e+01  6.523e+03  -0.003 0.997696
NationalityIreland                       6.005e-01  2.728e+00   0.220 0.825742
NationalityItaly                        -1.874e+01  4.184e+03  -0.004 0.996426
NationalityJamaica                      -1.688e+01  4.609e+03  -0.004 0.997078
NationalityJapan                        -1.502e+00  2.608e+00  -0.576 0.564588
NationalityLaos                         -1.838e+01  4.187e+03  -0.004 0.996497
NationalityMexico                       -2.297e+00  2.282e+00  -1.006 0.314240
NationalityOutlying-US(Guam-USVI-etc)   -1.694e+01  6.523e+03  -0.003 0.997928
NationalityPhilippines                  -2.274e+00  2.375e+00  -0.957 0.338412
NationalityPoland                       -2.433e+00  2.397e+00  -1.015 0.310036
NationalityPortugal                      1.404e+01  1.326e+03   0.011 0.991550
NationalityPuerto-Rico                   4.044e-01  2.361e+00   0.171 0.863991
NationalityScotland                     -1.343e+01  6.523e+03  -0.002 0.998357
NationalitySouth                        -1.776e+01  3.246e+03  -0.005 0.995635
NationalityTaiwan                       -3.320e+00  2.486e+00  -1.336 0.181695
NationalityTrinadad&Tobago              -1.855e+01  6.523e+03  -0.003 0.997730
NationalityUnited-States                -1.815e+00  2.002e+00  -0.906 0.364805
NationalityVietnam                      -1.958e+01  2.253e+03  -0.009 0.993067

(Intercept)                             *
Age                                     ***
Education11th
Education12th
Education1st-4th
Education5th-6th
Education7th-8th
Education9th
EducationAssoc-acdm                     *
EducationAssoc-voc
EducationBachelors                      **
EducationDoctorate                      ***
EducationHS-grad
EducationMasters                        **
EducationProf-school                    **
EducationSome-college                   *
SexMale
Hours_PW                                ***
Marital_StatusMarried-civ-spouse        ***
Marital_StatusMarried-spouse-absent
```

```
Marital_StatusNever-married
Marital_StatusSeparated
Marital_StatusWidowed
OccupationCraft-repair
OccupationExec-managerial
OccupationFarming-fishing             **
OccupationHandlers-cleaners
OccupationMachine-op-inspct
OccupationOther-service               **
OccupationPriv-house-serv
OccupationProf-specialty              .
OccupationProtective-serv
OccupationSales
OccupationTech-support                .
OccupationTransport-moving
NationalityCanada
NationalityChina
NationalityColumbia
NationalityCuba
NationalityDominican-Republic
NationalityEl-Salvador
NationalityEngland
NationalityFrance
NationalityGermany
NationalityGreece
NationalityGuatemala
NationalityHaiti
NationalityIndia
NationalityIran
NationalityIreland
NationalityItaly
NationalityJamaica
NationalityJapan
NationalityLaos
NationalityMexico
NationalityOutlying-US(Guam-USVI-etc)
NationalityPhilippines
NationalityPoland
NationalityPortugal
NationalityPuerto-Rico
NationalityScotland
NationalitySouth
NationalityTaiwan
```

```
NationalityTrinadad&Tobago
NationalityUnited-States
NationalityVietnam
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1552.71  on 1384  degrees of freedom
Residual deviance:  974.31  on 1319  degrees of freedom
AIC: 1106.3

Number of Fisher Scoring iterations: 17



Call:
glm(formula = Income ~ Age + Education + Hours_PW + Marital_Status +
    Occupation, family = binomial(link = "logit"), data = data)

Coefficients:
                                  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 *
Hours_PW                         3.046e-02  7.687e-03   3.963 7.40e-05 ***
Marital_StatusMarried-civ-spouse 1.941e+00  2.657e-01   7.306 2.76e-13 ***
Marital_StatusMarried-spouse-absent -4.485e-02  1.120e+00  -0.040  0.96806
Marital_StatusNever-married     -5.497e-01  3.490e-01  -1.575  0.11531
Marital_StatusSeparated         -1.611e+00  1.066e+00  -1.511  0.13083
```

```
Marital_StatusWidowed               -9.765e-01  8.396e-01  -1.163  0.24481
OccupationCraft-repair              -6.118e-02  3.259e-01  -0.188  0.85109
OccupationExec-managerial            4.468e-01  3.265e-01   1.368  0.17124
OccupationFarming-fishing           -1.621e+00  5.942e-01  -2.728  0.00638 **
OccupationHandlers-cleaners         -9.311e-02  5.548e-01  -0.168  0.86672
OccupationMachine-op-inspct         -5.220e-01  4.147e-01  -1.259  0.20812
OccupationOther-service             -2.445e+00  7.761e-01  -3.150  0.00163 **
OccupationPriv-house-serv           -1.542e+01  1.606e+03  -0.010  0.99234
OccupationProf-specialty             7.109e-01  3.459e-01   2.055  0.03985 *
OccupationProtective-serv            8.969e-02  5.165e-01   0.174  0.86213
OccupationSales                      3.198e-02  3.426e-01   0.093  0.92563
OccupationTech-support               1.079e+00  5.209e-01   2.072  0.03827 *
OccupationTransport-moving          -2.654e-01  4.154e-01  -0.639  0.52289
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1552.7  on 1384  degrees of freedom
Residual deviance:  998.6  on 1351  degrees of freedom
AIC: 1066.6

Number of Fisher Scoring iterations: 16


Implementation: ROI | Solver: lpsolve
Separation: TRUE
Existence of maximum likelihood estimates
                        (Intercept)                                 Age
                                  0                                   0
                      Education11th                       Education12th
                                  0                                -Inf
                   Education1st-4th                    Education5th-6th
                               -Inf                                -Inf
                   Education7th-8th                        Education9th
                                  0                                   0
                EducationAssoc-acdm                 EducationAssoc-voc
                                  0                                   0
                 EducationBachelors                 EducationDoctorate
                                  0                                   0
                    EducationHS-grad                    EducationMasters
                                  0                                   0
               EducationProf-school               EducationSome-college
```

```
                                0                                 0
                         Hours_PW    Marital_StatusMarried-civ-spouse
                                0                                 0
Marital_StatusMarried-spouse-absent         Marital_StatusNever-married
                                0                                 0
            Marital_StatusSeparated              Marital_StatusWidowed
                                0                                 0
             OccupationCraft-repair           OccupationExec-managerial
                                0                                 0
          OccupationFarming-fishing         OccupationHandlers-cleaners
                                0                                 0
        OccupationMachine-op-inspct             OccupationOther-service
                                0                                 0
          OccupationPriv-house-serv            OccupationProf-specialty
                             -Inf                                 0
          OccupationProtective-serv                    OccupationSales
                                0                                 0
             OccupationTech-support          OccupationTransport-moving
                                0                                 0
0: finite value, Inf: infinity, -Inf: -infinity
```

The model selection process involves 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.

1. Full Model Fitting The initial model (full_model) includes all available predictor variables: Age, Education, Sex, Work Hours per Week (Hours_PW), Marital Status, Occupation, and Nationality. A separation detection test (detect_separation_full_model) is applied to check if certain variables lead to perfect separation, which may cause convergence issues in the model. The summary of the full model provides insights into which variables are statistically significant.
2. Model Optimization using Stepwise AIC Stepwise Akaike Information Criterion (AIC) selection is applied (stepAIC_model), which iteratively removes the least significant predictors to find the best-performing model. The final optimized model retains only the most relevant predictors: Age, Education, Work Hours per Week, Marital Status, and Occupation. A separation detection test (detect_separation_stepAIC_model) is performed again to check whether the optimized model exhibits perfect separation issues.
3. Key Takeaways: The full model includes all predictors but may contain unnecessary variables that do not contribute significantly. 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.

3.1 Option choose of model

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

Table 3: Martial_Status Distribution.

Income	Married	Unmarried
<=50K	36.3% (378)	63.7% (663)
>50K	84.6% (291)	15.4% (53)

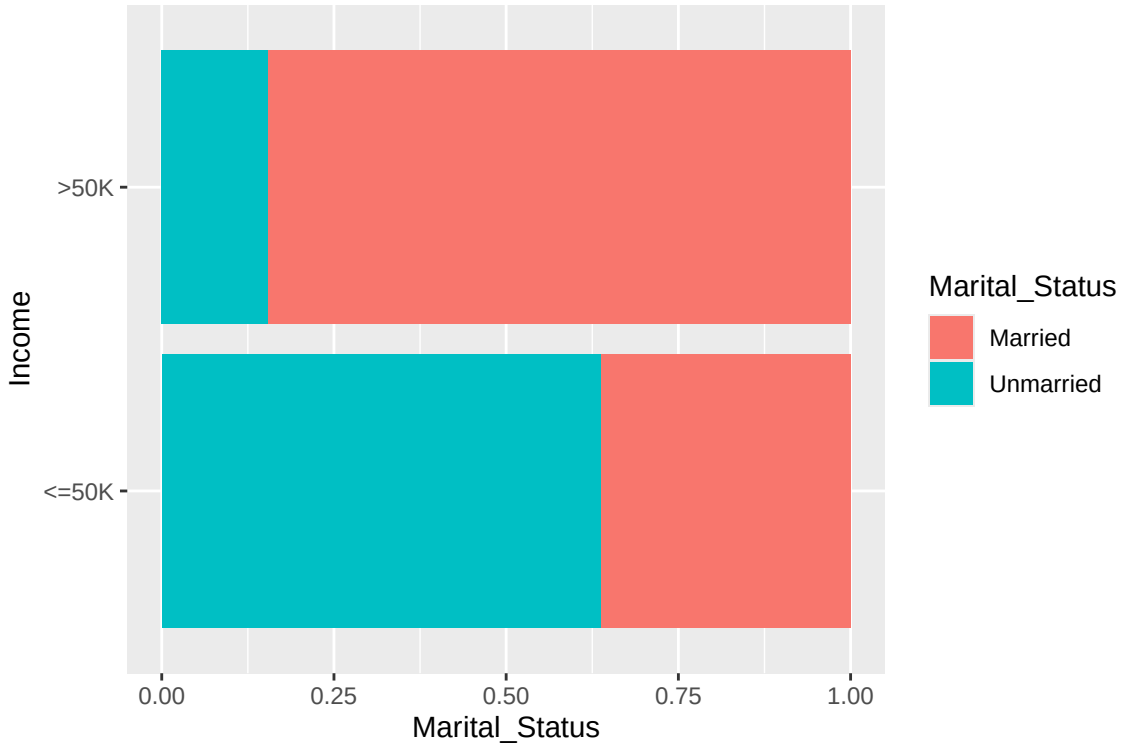


Figure 22: Income by Marital_Status.

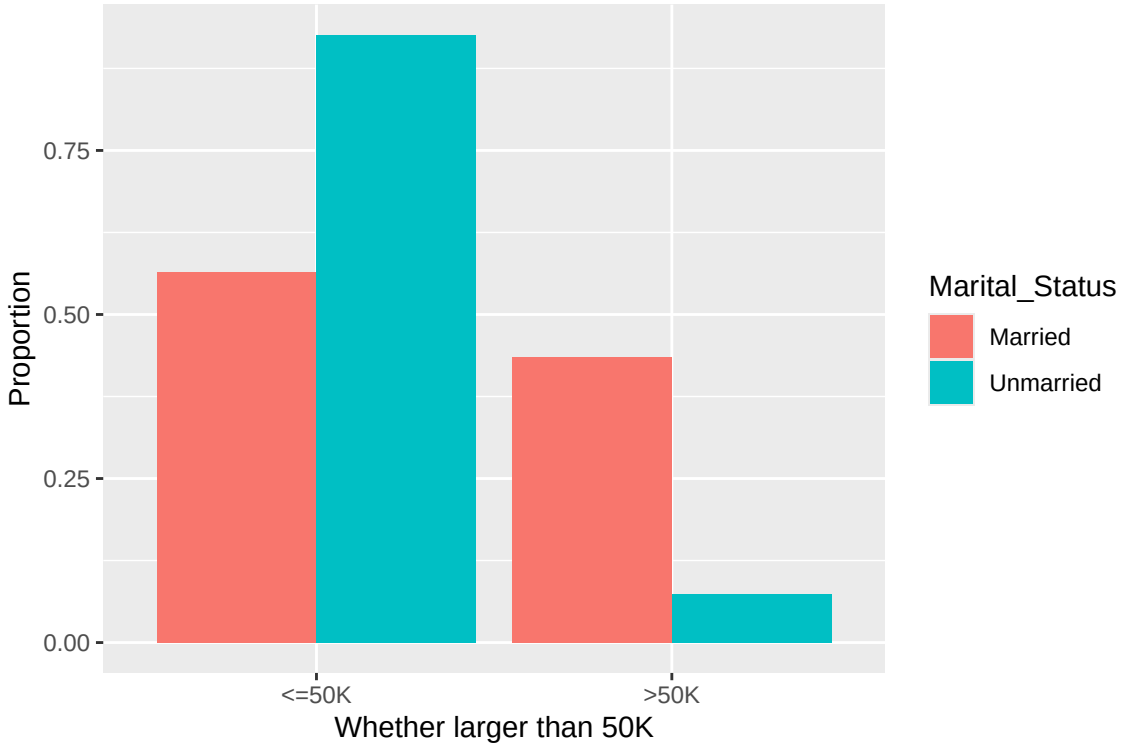


Figure 23: Income by Marital_Status.

Table 4: Education Distribution.

Income	High_Education	Low_Education
<=50K	84.1% (875)	15.9% (166)
>50K	96.8% (333)	3.2% (11)

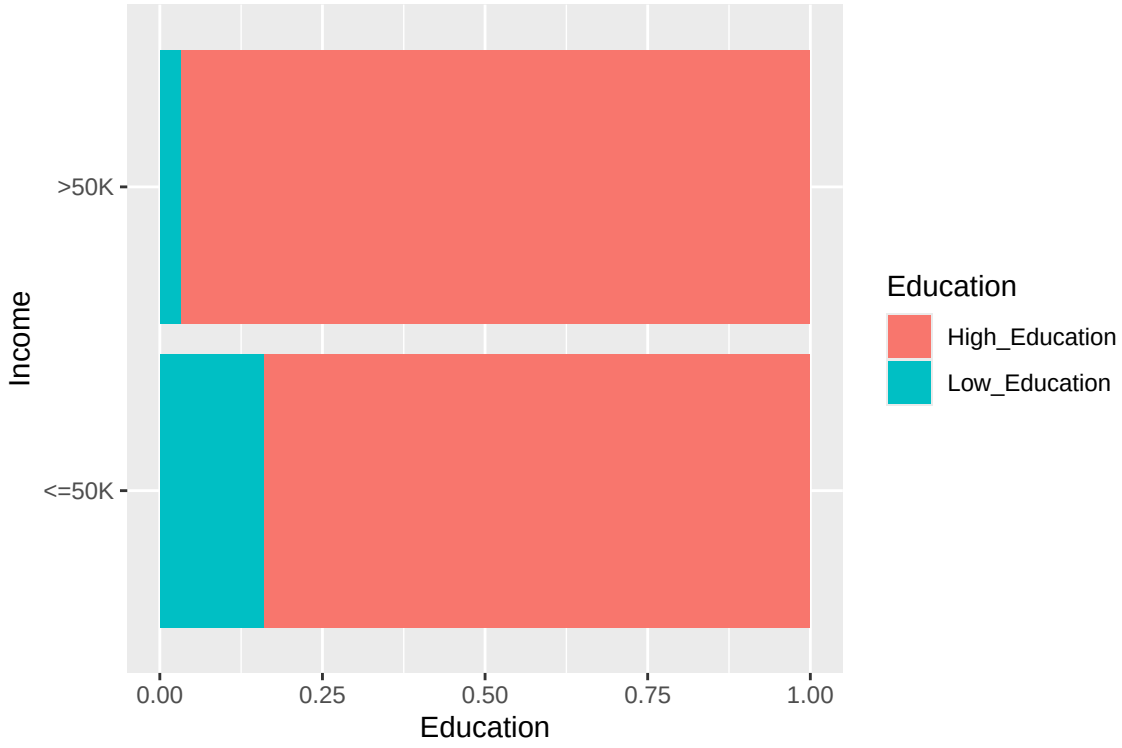


Figure 24: Income by Education.

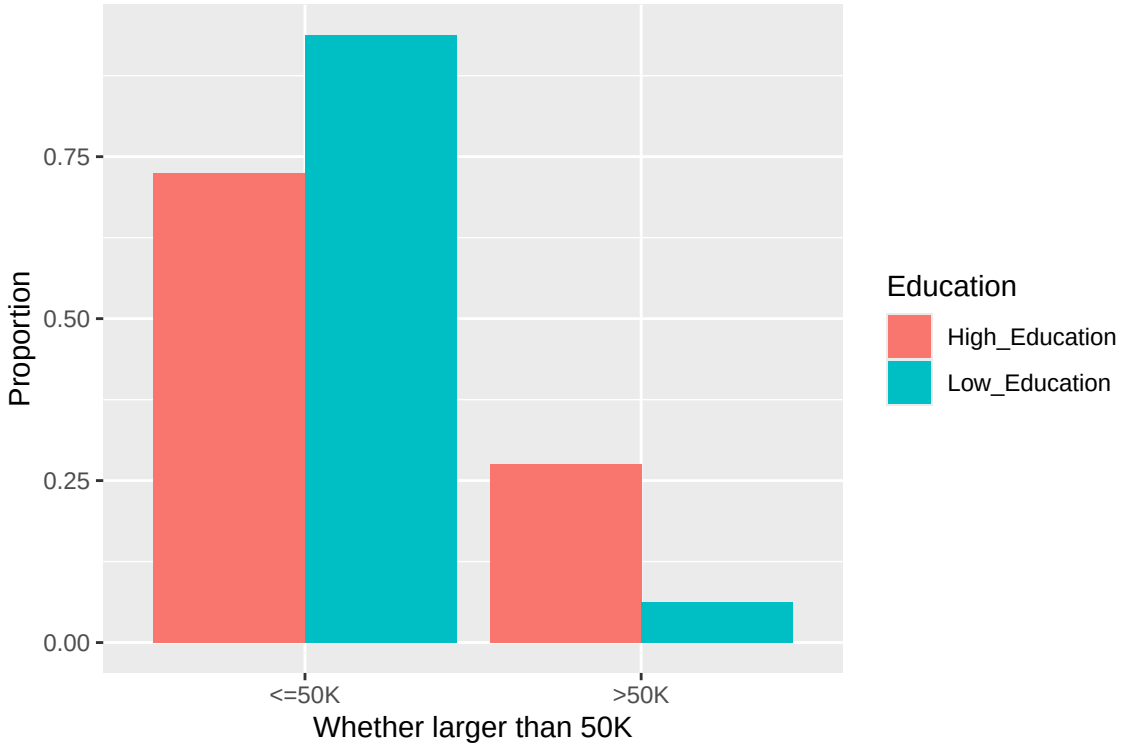


Figure 25: Income by Education.

To address the complete separation issue, some explanatory variables have been merged into broader categories. This helps ensure that each category has sufficient data points, improving the stability of the logistic regression model.

1. Education Level Merging Low Education:

Groups together individuals with education levels from 1st grade to 12th grade (e.g., 1st-4th, 5th-6th, 7th-8th, 9th, 10th, 11th, 12th). These individuals generally have no formal higher education and are more likely to be in low-income jobs. High Education:

Includes individuals with some form of post-secondary education (e.g., Associate degree, Bachelor's, Master's, Doctorate, HS-grad, Prof-school, Some-college). These individuals have better earning potential and are more likely to be in high-income categories.

2. Marital Status Merging Unmarried:

Combines Separated, Widowed, Never-married, and Divorced into a single category. These individuals tend to have lower financial stability compared to married individuals. Married:

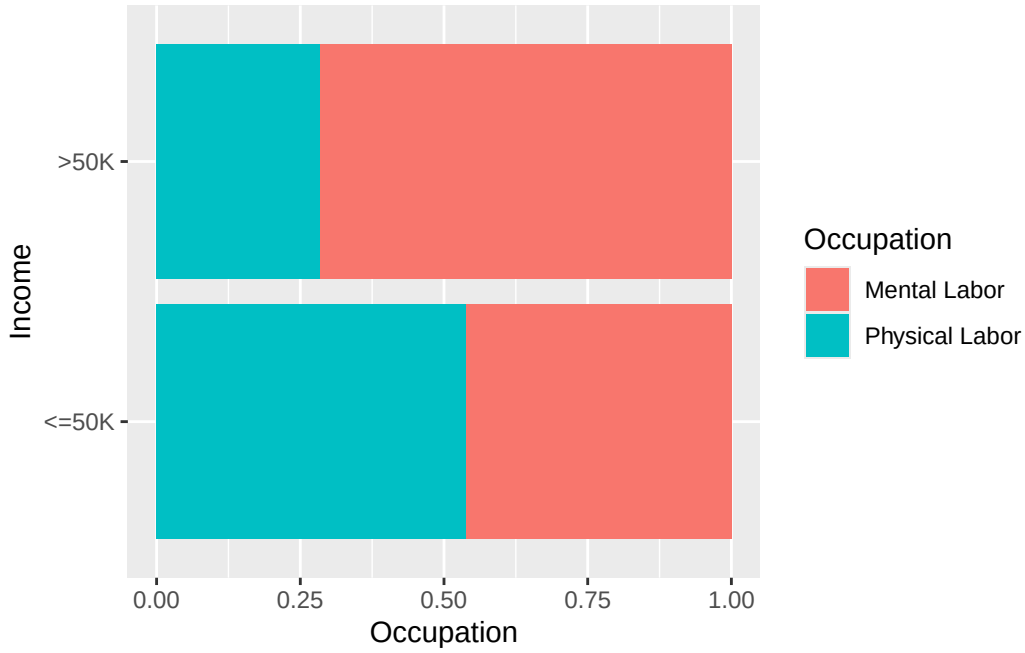
Combines Married-spouse-absent and Married-civ-spouse into a single category. Married individuals generally have more financial stability, often benefiting from dual-income households.

3. Visualizing the Effects of Merging

   - Marital Status vs. Income:
     The distribution of income across the newly grouped Married and Unmarried categories is displayed.
   - Education vs. Income:
     The newly created Low Education and High Education categories allow for a clearer comparison of income distribution between different education levels. Key

Takeaways: Merging categories helps reduce model complexity and prevent complete separation issues. Individuals with higher education levels are more likely to earn >50K compared to those in the Low Education category. Married individuals show a higher proportion of high-income earners, reinforcing previous findings about marital status and financial stability. These transformations improve the model's ability to generalize and make more accurate predictions.

3.2 Combine occupation types (physical labor and mental labor)



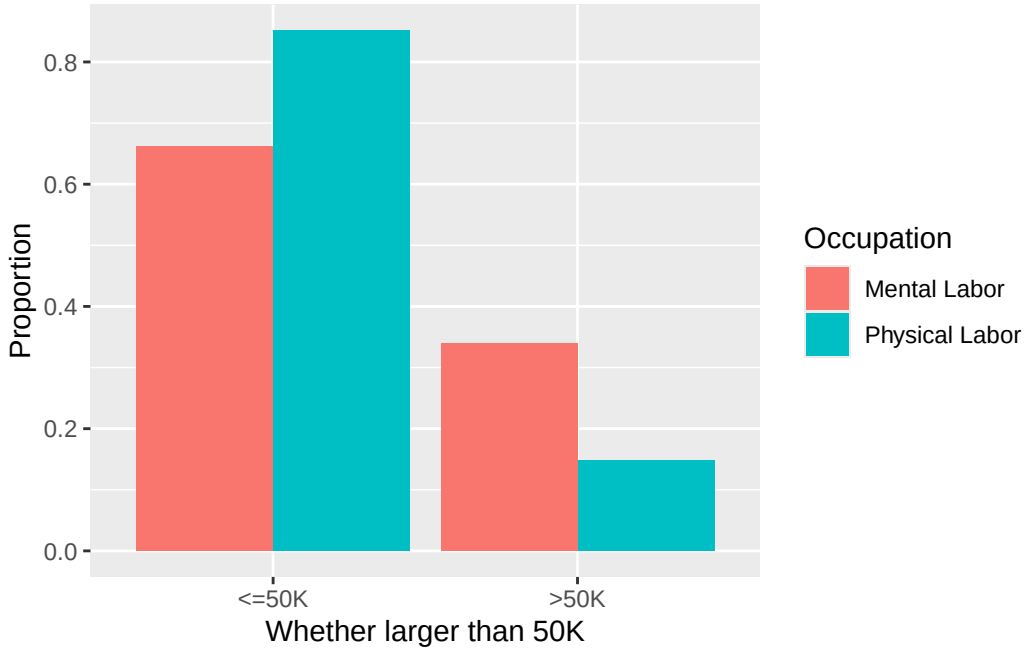


Table 5: Occupation Distribution (physical labor and mental labor).

Income	Mental Labor	Physical Labor
<=50K	46.1% (480)	53.9% (561)
>50K	71.5% (246)	28.5% (98)

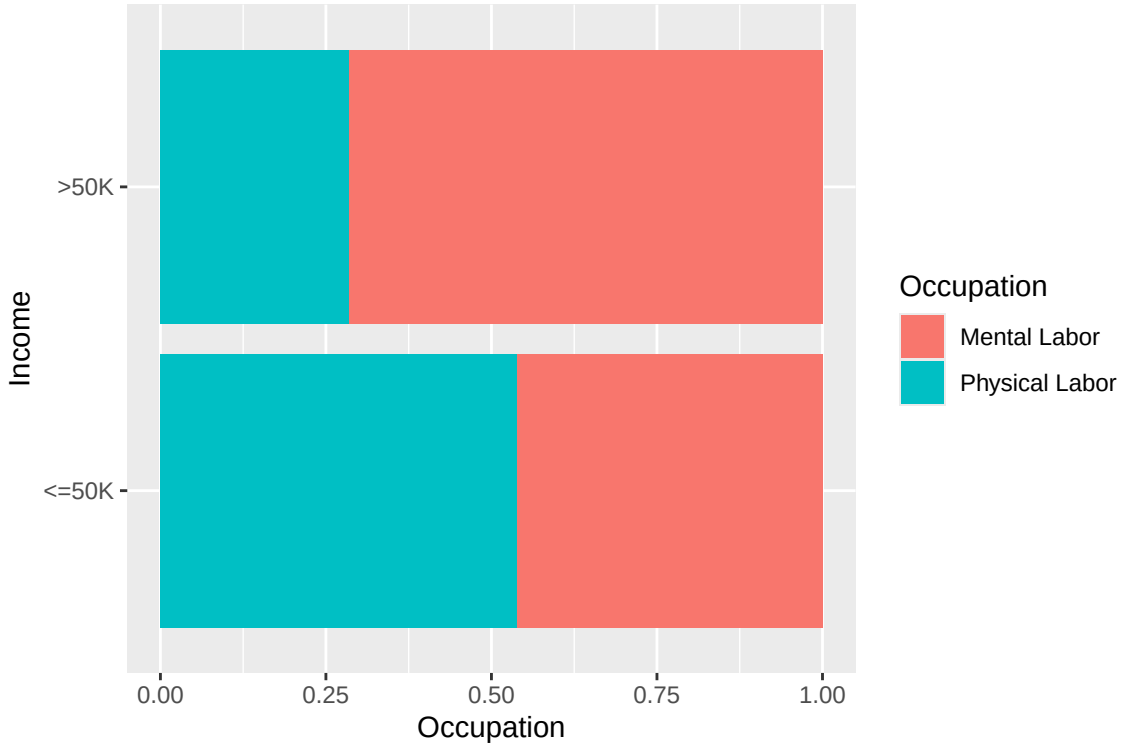


Figure 26: Income by Occupation

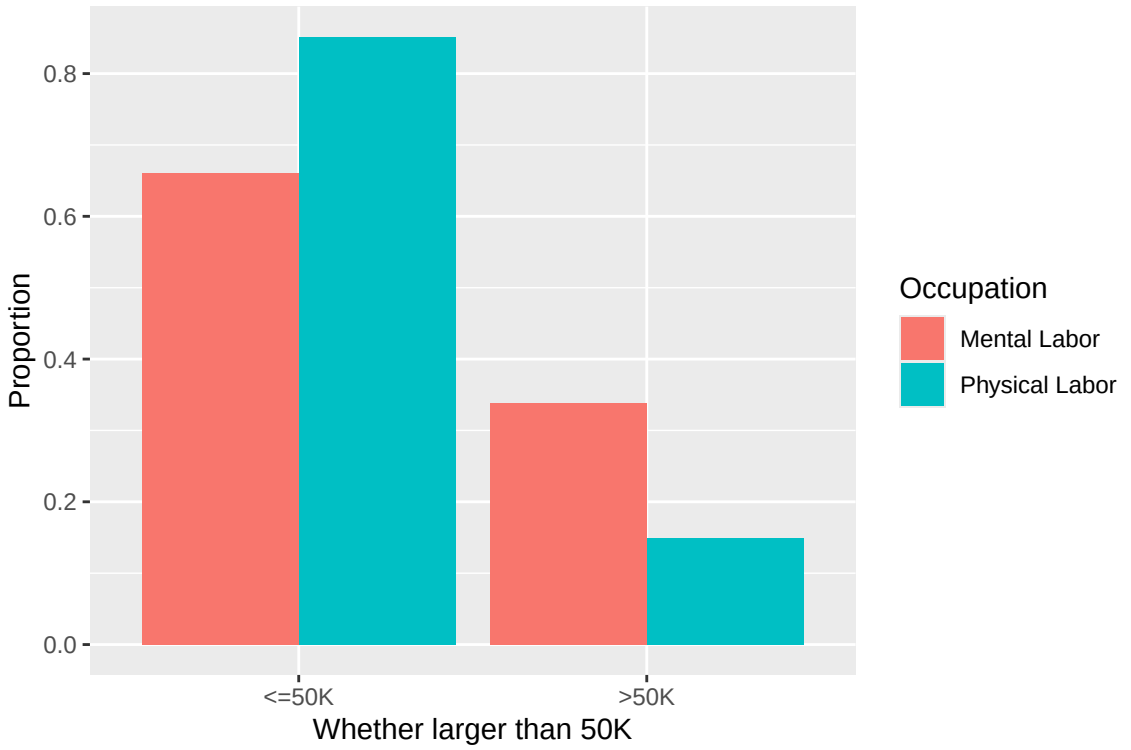


Figure 27: Income by Occupation

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

Mental Labor (White-collar jobs):

Includes occupations such as Administrative clerical, Executive managerial, Professional specialty, Sales, and Technical support. These jobs typically involve cognitive work, decision-making, and problem-solving.

Physical Labor (Blue-collar jobs):

Includes Craft repair, Farming and fishing, Handlers and cleaners, Machine operators, Transport workers, Other services, Private household services, and Protective services. These jobs are more labor-intensive and require physical exertion.

Analysis of Income Distribution after Merging:

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.

Key Takeaways: Merging occupations into Physical vs. Mental labor simplifies the analysis while preserving meaningful insights. Mental labor positions are more likely to be associated with higher salaries. Physical laborers predominantly fall into the low-income category, likely due to industry pay standards and job requirements. These findings emphasize the importance of education and skill specialization in securing higher-paying jobs.

3.3 Combine occupation types (by PRC Job Classification List)

ref:https://zchweb.oss-cn-beijing.aliyuncs.com/contract/temp/2021122116541363304.pdf

Table 6: Occupation Distribution (PRC Job Classification List).

Income	1	2	3	4	5	6
<=50K	9.3% (97)	11.1% (116)	13.5% (141)	33.9% (353)	3.6% (37)	28.5% (297)
>50K	23.0% (79)	29.1% (100)	7.0% (24)	17.7% (61)	1.5% (5)	21.8% (75)

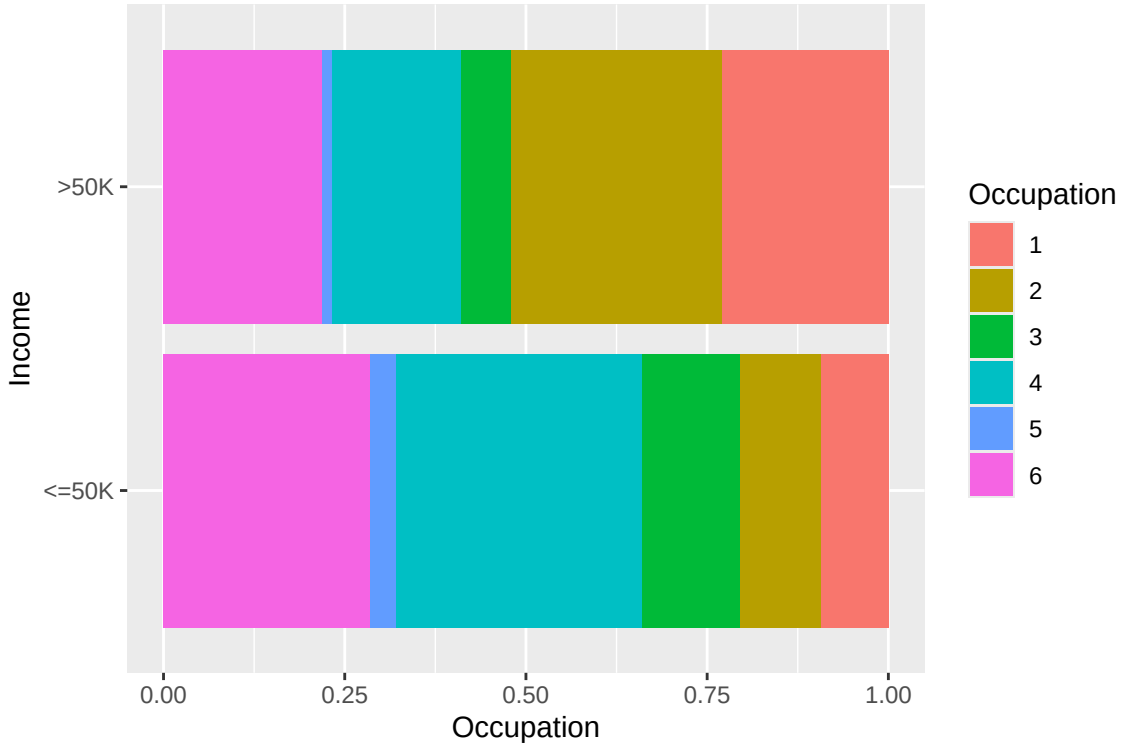


Figure 28: Income by Occupation

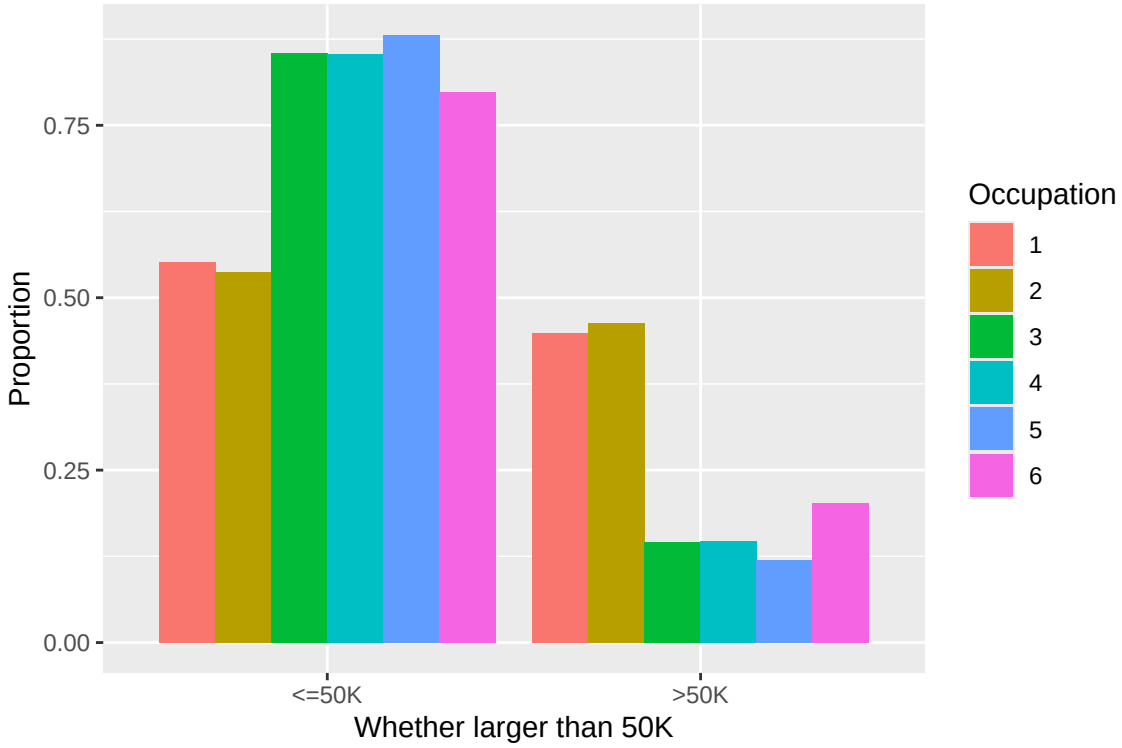


Figure 29: Income by Occupation

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, which allows for a more structured income comparison.

Income Distribution Analysis After Merging:

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.

Key Takeaways: The PRC Job Classification List provides a structured way to analyze income disparities across different occupational groups. Management and technical jobs tend to have higher salaries, while sales, service, and agricultural jobs have a greater share of low-income earners. Grouping occupations in this way allows for more precise policy recommendations and workforce planning.

3.4 Combine occupation types (by International Standard Classification of Occupations (ISCO-08))

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

Table 7: Occupation Distribution (ISCO-08).

Income	1	2	3	4	5	6	7	8	9
<=50K	9.3% (97)	8.8% (92)	2.3% (24)	13.5% (141)	28.6% (298)	3.6% (37)	14.8% (154)	13.7% (143)	5.3% (55)
>50K	23.0% (79)	25.3% (87)	3.8% (13)	7.0% (24)	16.0% (55)	1.5% (5)	13.1% (45)	8.7% (30)	1.7% (6)

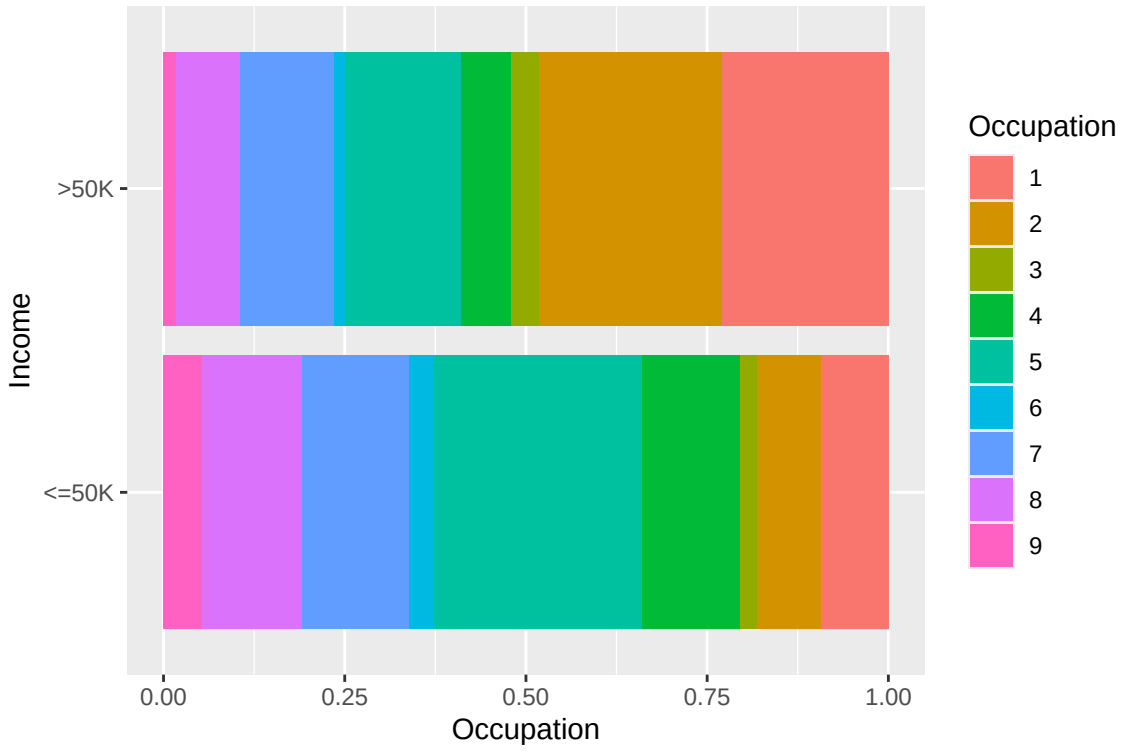


Figure 30: Income by Occupation

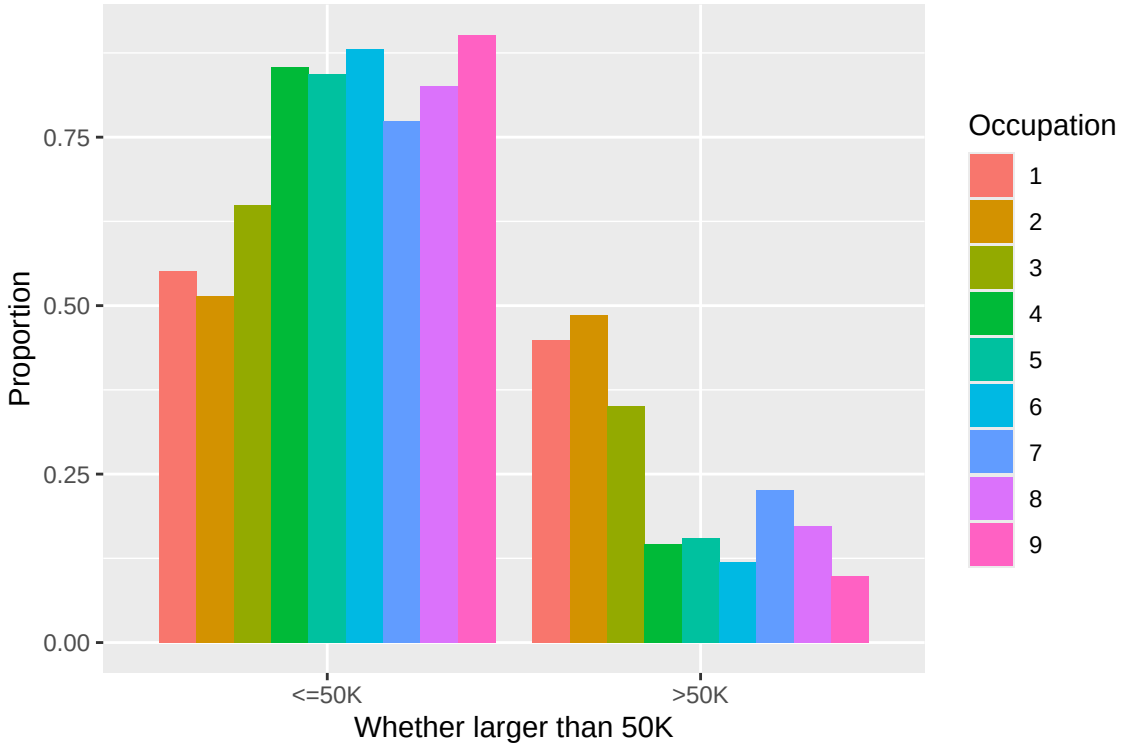


Figure 31: Income by Occupation

```
Call:
glm(formula = Income ~ Age + Education + Sex + Hours_PW + Marital_Status +
    Occupation, family = binomial(link = "logit"), data = data.new.1)

Coefficients:
                           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 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 1552.7  on 1384  degrees of freedom
Residual deviance: 1109.9  on 1378  degrees of freedom
AIC: 1123.9

Number of Fisher Scoring iterations: 6


Implementation: ROI | Solver: lpsolve
Separation: FALSE
Existence of maximum likelihood estimates
             (Intercept)                       Age   EducationLow_Education
                       0                         0                       0
                 SexMale  Marital_StatusUnmarried OccupationPhysical Labor
                       0                         0                       0
                Hours_PW
                       0
0: finite value, Inf: infinity, -Inf: -infinity



Call:
glm(formula = Income ~ Age + Education + Sex + Hours_PW + Marital_Status +
    Occupation, family = binomial(link = "logit"), data = data.new.2)

Coefficients:
                         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 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1552.7  on 1384  degrees of freedom
```

```
Residual deviance: 1079.5  on 1374  degrees of freedom
AIC: 1101.5

Number of Fisher Scoring iterations: 6


Implementation: ROI | Solver: lpsolve
Separation: FALSE
Existence of maximum likelihood estimates
            (Intercept)                     Age  EducationLow_Education
                      0                       0                       0
                SexMale Marital_StatusUnmarried             Occupation2
                      0                       0                       0
            Occupation3             Occupation4             Occupation5
                      0                       0                       0
            Occupation6                Hours_PW
                      0                       0
0: finite value, Inf: infinity, -Inf: -infinity



Call:
glm(formula = Income ~ Age + Education + Sex + Hours_PW + Marital_Status +
    Occupation, family = binomial(link = "logit"), data = data.new.3)

Coefficients:
                         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 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1552.7  on 1384  degrees of freedom
Residual deviance: 1078.4  on 1371  degrees of freedom
AIC: 1106.4

Number of Fisher Scoring iterations: 6


Implementation: ROI | Solver: lpsolve
Separation: FALSE
Existence of maximum likelihood estimates
            (Intercept)                      Age  EducationLow_Education
                      0                        0                       0
                SexMale Marital_StatusUnmarried             Occupation2
                      0                        0                       0
            Occupation3             Occupation4             Occupation5
                      0                        0                       0
            Occupation6             Occupation7             Occupation8
                      0                        0                       0
            Occupation9                Hours_PW
                      0                        0
0: finite value, Inf: infinity, -Inf: -infinity
```

The International Standard Classification of Occupations (ISCO-08) is 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.

Income Distribution Analysis After Merging:

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.

Model Selection and Performance Comparison:

The stepwise AIC model selection process was applied separately to datasets using different occupational classification methods (Mental vs. Physical Labor, PRC Job Classification, and ISCO-08). All models showed improvement in fitting compared to the full model, but the best performance was observed with the ISCO-08 classification. Nationality was removed in the AIC-selected models, reinforcing the earlier observation that nationality has little impact on income prediction.

Key Takeaways: ISCO-08 classification provides a structured way to assess income distribution across occupations. Managers and professionals dominate the high-income group, while service, sales, and agriculture workers struggle with lower wages. Machine operators and craft workers occupy a middle ground, earning more than service workers but less than professionals. The final model using ISCO-08 classification achieves a better balance between accuracy and interpretability, making it a robust choice for workforce and economic analysis.

4 Model Check

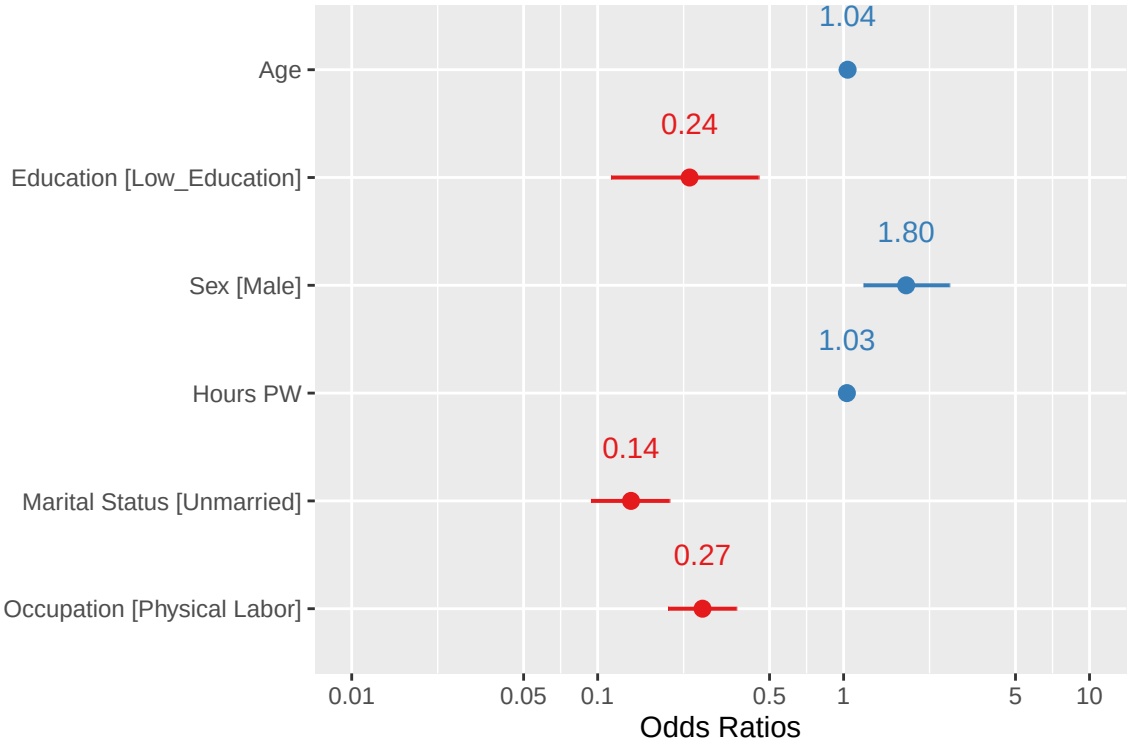


Figure 32: Model Plot for stepAIC_model_new.1

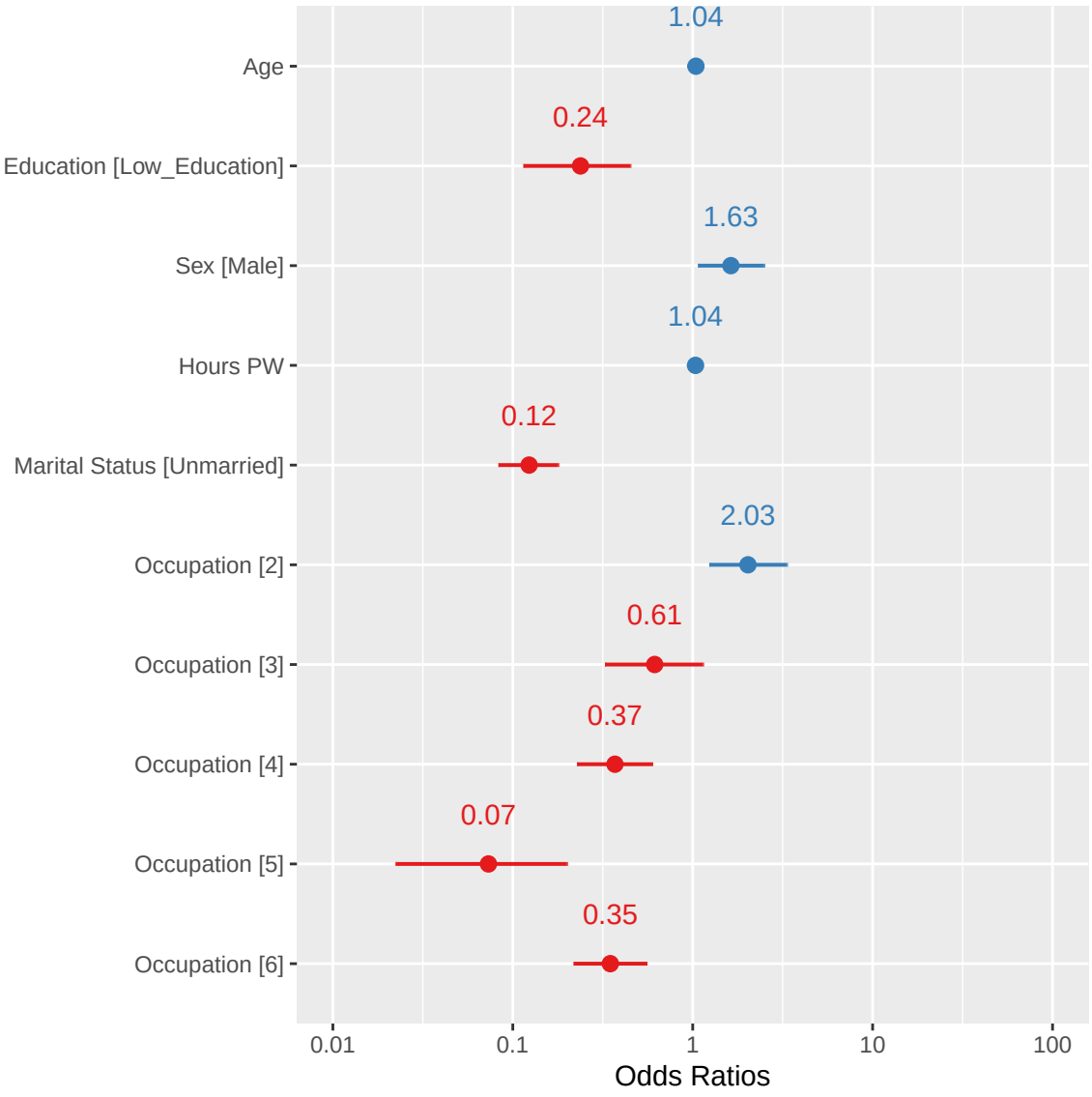


Figure 33: Model Plot for stepAIC_model_new.2

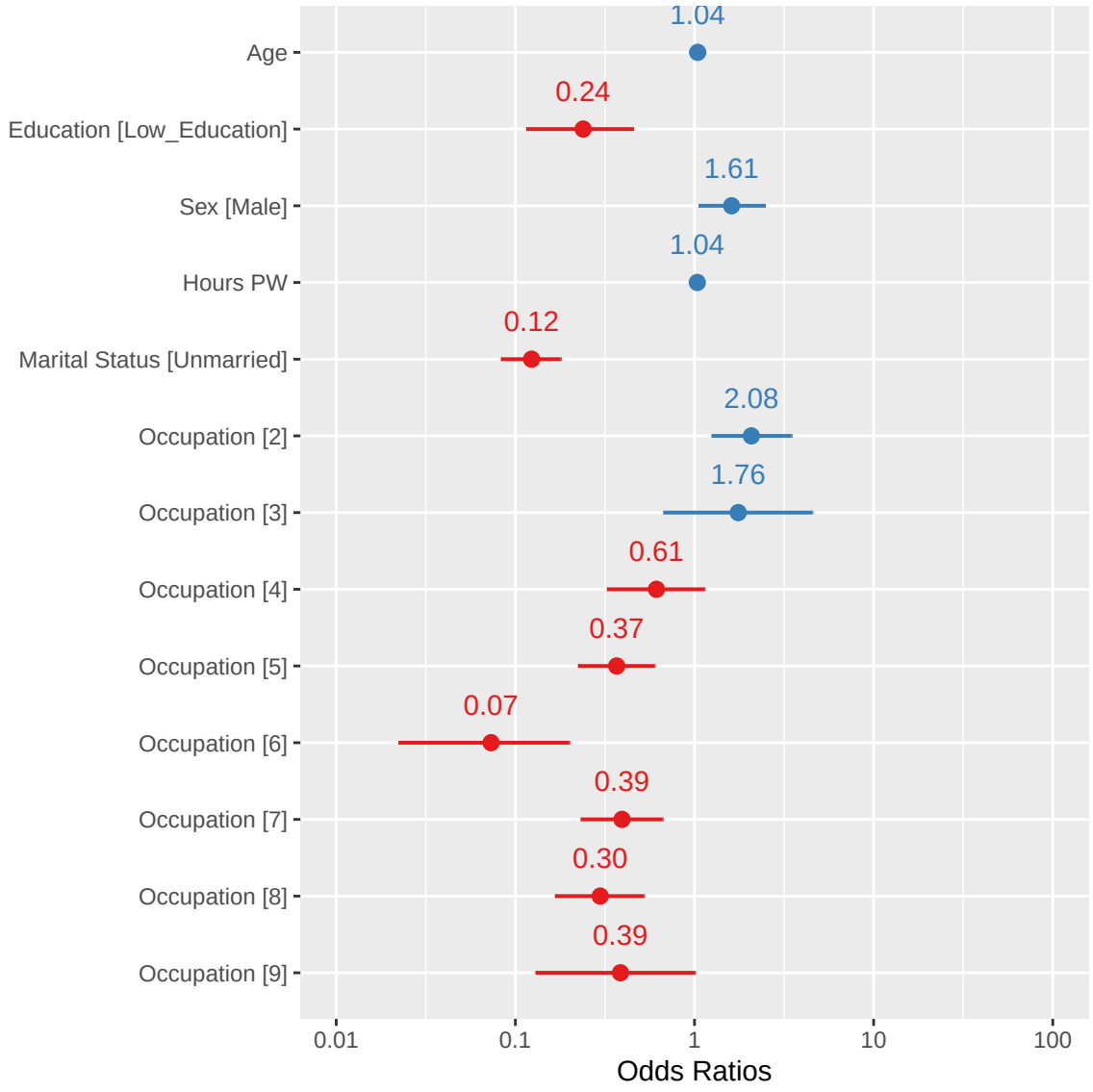


Figure 34: Model Plot for stepAIC_model

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: Income

Terms added sequentially (first to last)
```

```
               Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                           1384     1552.7
Age             1   90.112      1383     1462.6 < 2.2e-16 ***
Education       1   53.385      1382     1409.2 2.742e-13 ***
Sex             1   62.449      1381     1346.8 2.734e-15 ***
Hours_PW        1   36.563      1380     1310.2 1.478e-09 ***
Marital_Status  1  129.706      1379     1180.5 < 2.2e-16 ***
Occupation      1   70.560      1378     1109.9 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Analysis of Deviance Table

Model: binomial, link: logit

Response: Income

Terms added sequentially (first to last)

               Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                           1384     1552.7
Age             1   90.112      1383     1462.6 < 2.2e-16 ***
Education       1   53.385      1382     1409.2 2.742e-13 ***
Sex             1   62.449      1381     1346.8 2.734e-15 ***
Hours_PW        1   36.563      1380     1310.2 1.478e-09 ***
Marital_Status  1  129.706      1379     1180.5 < 2.2e-16 ***
Occupation      5  101.032      1374     1079.5 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Analysis of Deviance Table

Model: binomial, link: logit

Response: Income

Terms added sequentially (first to last)

               Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                           1384     1552.7
Age             1   90.112      1383     1462.6 < 2.2e-16 ***
```

```
Education       1   53.385      1382     1409.2 2.742e-13 ***
Sex             1   62.449      1381     1346.8 2.734e-15 ***
Hours_PW        1   36.563      1380     1310.2 1.478e-09 ***
Marital_Status  1  129.706      1379     1180.5 < 2.2e-16 ***
Occupation      8  102.097      1371     1078.4 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

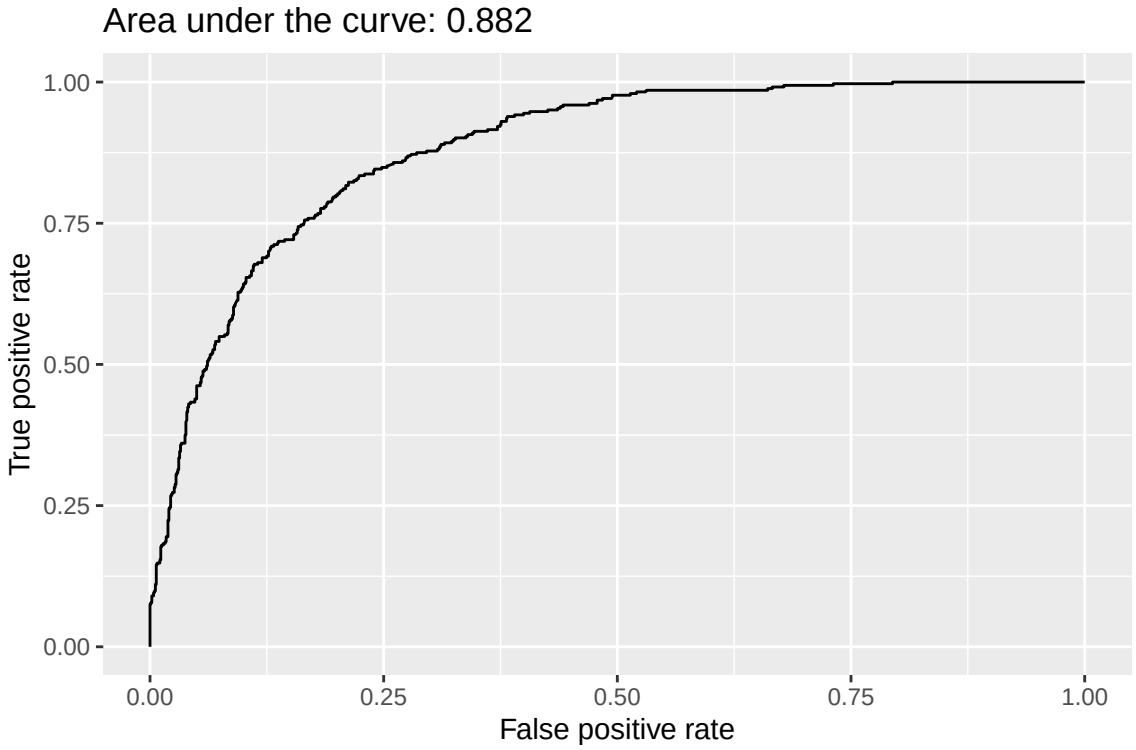


Figure 35: ROC curve for stepAIC_model

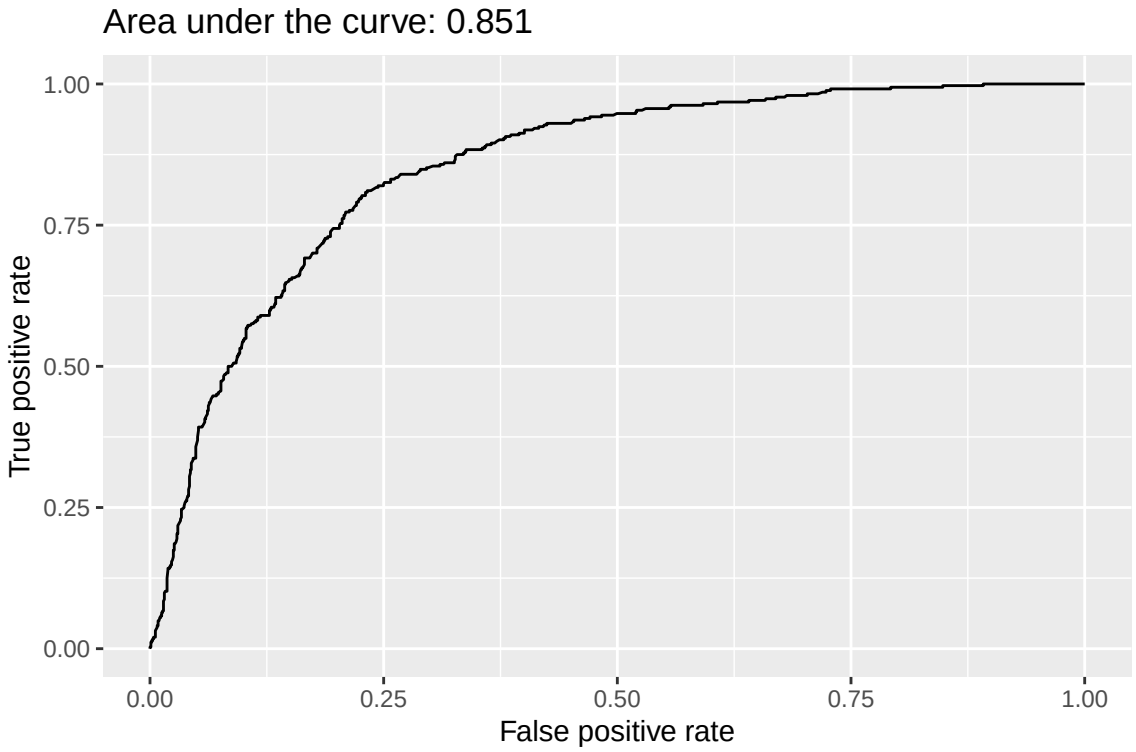


Figure 36: ROC curve for stepAIC_model_new.1

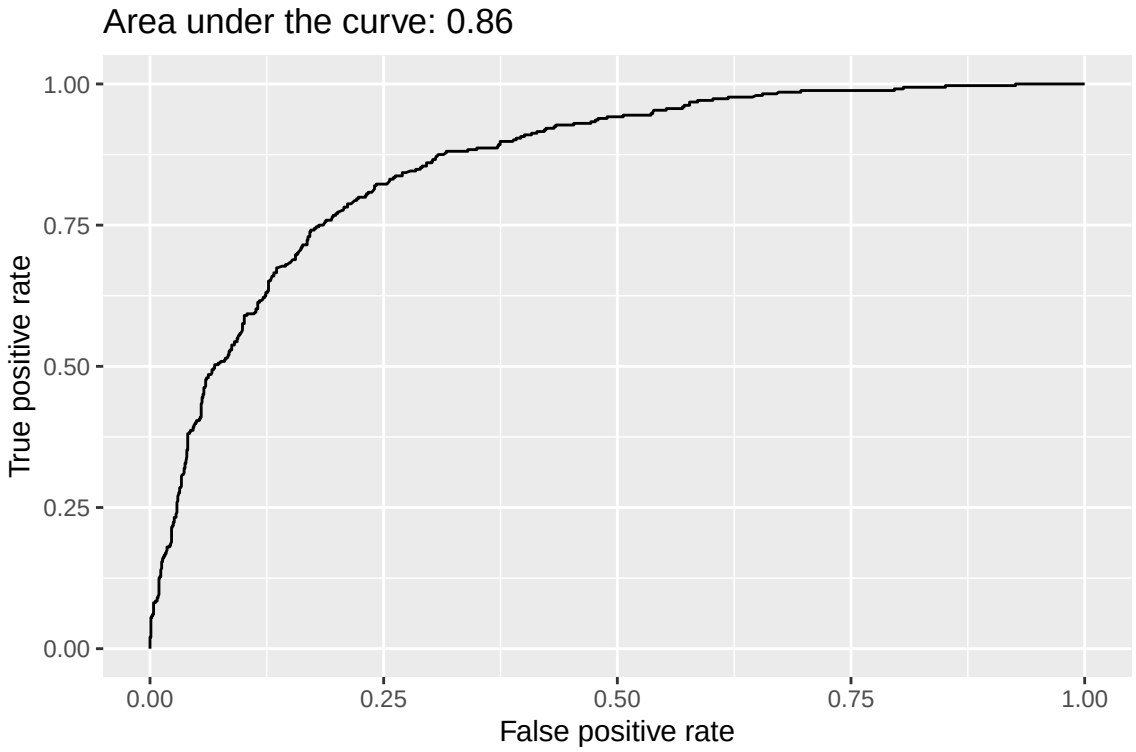


Figure 37: ROC curve for stepAIC_model_new.2

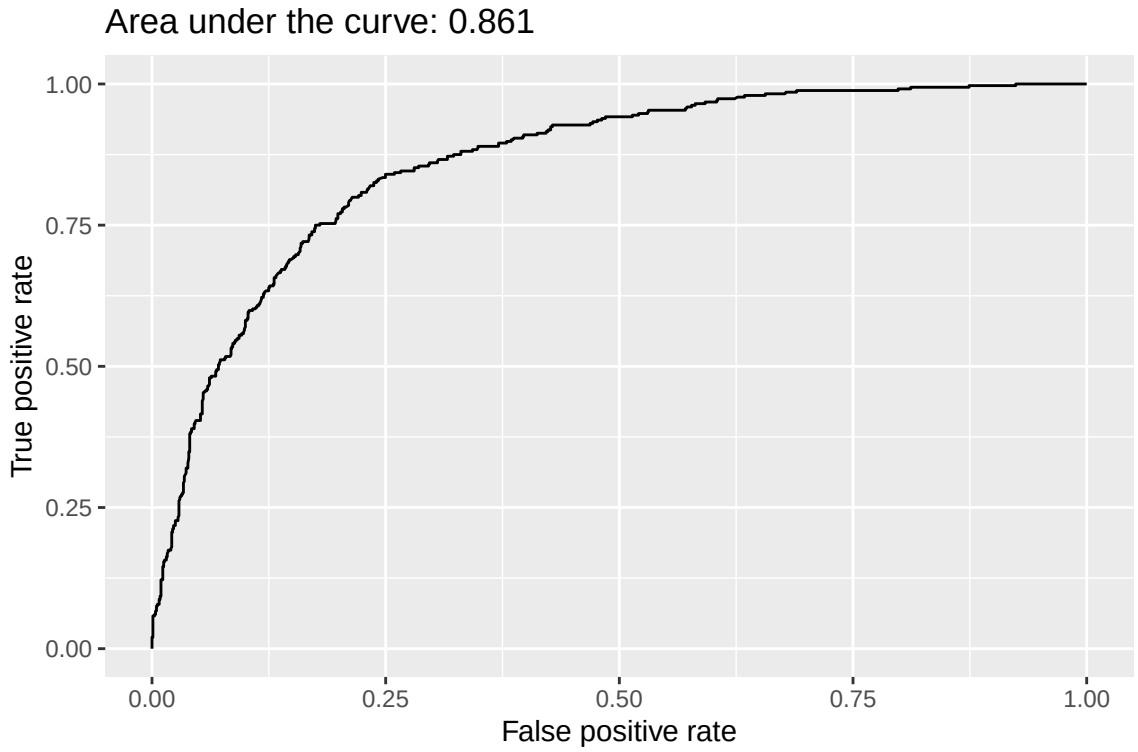


Figure 38: ROC curve for stepAIC_model_new.3

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 confidence 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 confidence interval lies entirely to the left of 1 (OR $<$ 1), it suggests a negative relationship. If the confidence interval spans across 1, the positive or negative effect cannot be significantly distinguished.

ROC Curve Analysis & AUC Comparison: 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 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.

Key Takeaways: The ROC curve and AUC values help determine the best occupational classification method in terms of prediction accuracy. Model coefficients provide insights into the most influential factors affecting income. ANOVA helps verify whether additional variables significantly improve prediction performance. The final model selection should balance accuracy and interpretability, preferring models with a high AUC while avoiding excessive complexity.

4.1 Consider the interaction effects

```
Call:
glm(formula = Income ~ Age + Education + Sex + Hours_PW + Marital_Status +
    Occupation + Age:Marital_Status + Education:Sex + Sex:Marital_Status +
    Marital_Status:Occupation, family = binomial(link = "logit"),
    data = data.new.1)

Coefficients:
                                                Estimate Std. Error z value
(Intercept)                                    -2.189254   0.495876  -4.415
Age                                             0.031676   0.008176   3.874
EducationLow_Education                        -15.313498 481.667910  -0.032
SexMale                                        -0.060678   0.272610  -0.223
Hours_PW                                        0.032659   0.007021   4.651
Marital_StatusUnmarried                        -4.514939   0.707999  -6.377
OccupationPhysical Labor                       -1.420146   0.180064  -7.887
Age:Marital_StatusUnmarried                     0.030029   0.013723   2.188
EducationLow_Education:SexMale                 14.012202 481.668031   0.029
SexMale:Marital_StatusUnmarried                 1.546335   0.444445   3.479
Marital_StatusUnmarried:OccupationPhysical Labor  0.633460   0.383444   1.652
                                               Pr(>|z|)
(Intercept)                                    1.01e-05 ***
Age                                            0.000107 ***
EducationLow_Education                         0.974637
SexMale                                        0.823862
Hours_PW                                       3.30e-06 ***
Marital_StatusUnmarried                        1.81e-10 ***
```

```
OccupationPhysical Labor                          3.10e-15 ***
Age:Marital_StatusUnmarried                       0.028653 *
EducationLow_Education:SexMale                    0.976792
SexMale:Marital_StatusUnmarried                   0.000503 ***
Marital_StatusUnmarried:OccupationPhysical Labor 0.098529 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1552.7  on 1384  degrees of freedom
Residual deviance: 1084.2  on 1374  degrees of freedom
AIC: 1106.2

Number of Fisher Scoring iterations: 16


Call:
glm(formula = Income ~ Age + Education + Sex + Hours_PW + Marital_Status +
    Occupation + Age:Marital_Status + Sex:Marital_Status, family = binomial(link = "logit"),
    data = data.new.1)

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -2.270129   0.491722  -4.617 3.90e-06 ***
Age                              0.031818   0.008089   3.933 8.38e-05 ***
EducationLow_Education          -1.462369   0.346401  -4.222 2.43e-05 ***
SexMale                         -0.037620   0.264656  -0.142   0.8870
Hours_PW                         0.032681   0.007023   4.654 3.26e-06 ***
Marital_StatusUnmarried         -4.369683   0.705494  -6.194 5.87e-10 ***
OccupationPhysical Labor        -1.292511   0.160702  -8.043 8.77e-16 ***
Age:Marital_StatusUnmarried      0.028900   0.013623   2.121   0.0339 *
SexMale:Marital_StatusUnmarried  1.692187   0.431822   3.919 8.90e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1552.7  on 1384  degrees of freedom
Residual deviance: 1090.2  on 1376  degrees of freedom
AIC: 1108.2

Number of Fisher Scoring iterations: 6
```

```
Implementation: ROI | Solver: lpsolve
Separation: FALSE
Existence of maximum likelihood estimates
                     (Intercept)                              Age
                               0                                0
          EducationLow_Education                         SexMale
                               0                                0
                       Hours_PW          Marital_StatusUnmarried
                               0                                0
        OccupationPhysical Labor       Age:Marital_StatusUnmarried
                               0                                0
SexMale:Marital_StatusUnmarried
                               0
0: finite value, Inf: infinity, -Inf: -infinity
```

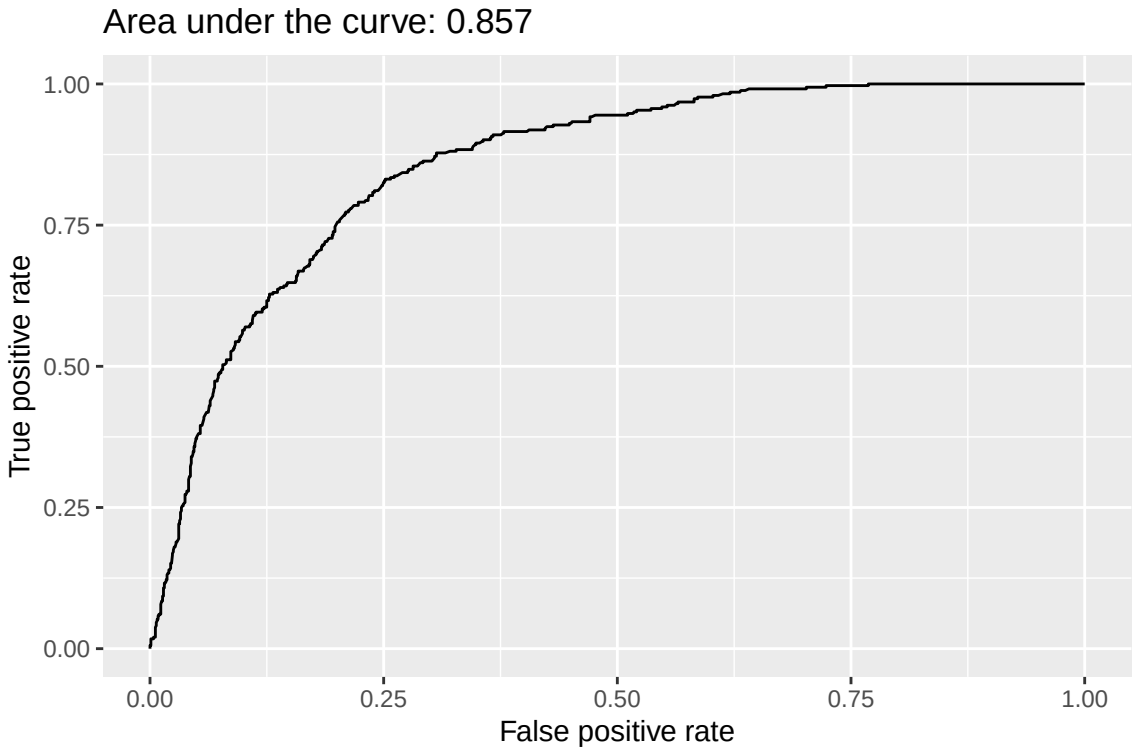


Figure 39: ROC curve for model_new.1.interaction

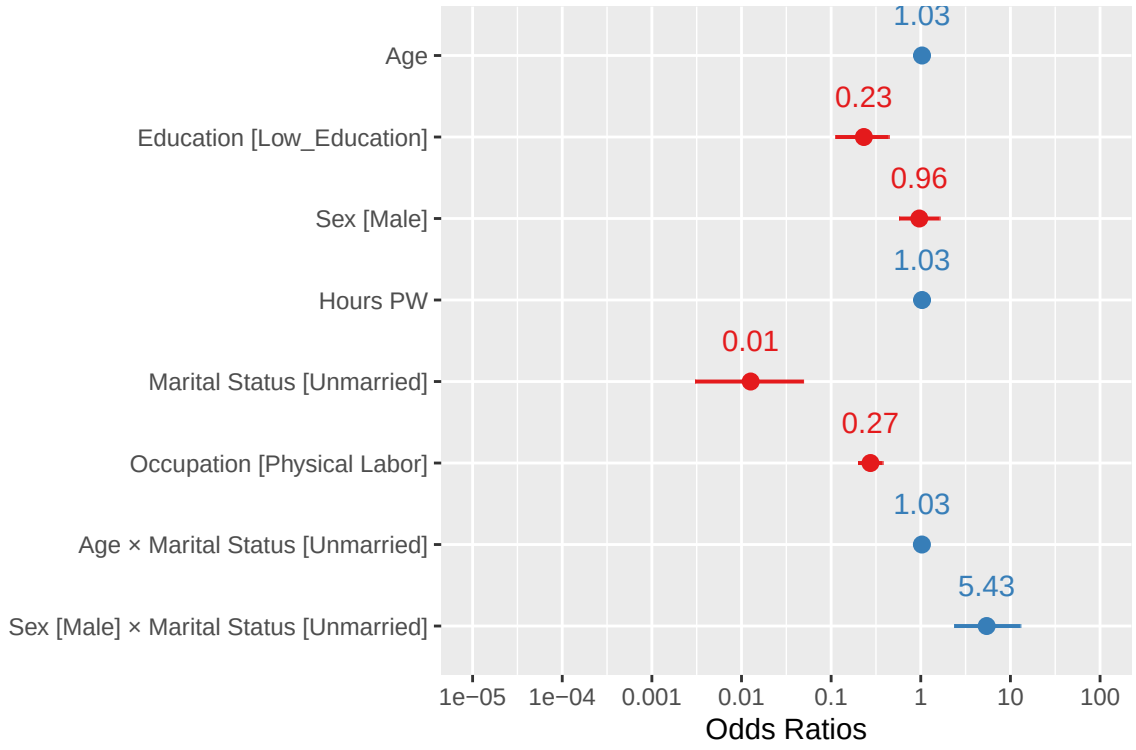


Figure 40: Model Plot for model_new.1.interaction

```
Call:
glm(formula = Income ~ Age + Education + Sex + Hours_PW + Marital_Status +
    Occupation + Age:Marital_Status + Education:Sex + Education:Hours_PW +
    Sex:Marital_Status + Sex:Occupation + Marital_Status:Occupation,
    family = binomial(link = "logit"), data = data.new.2)

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -2.073e+00  6.763e-01  -3.065  0.00218 **
Age                            3.958e-02  8.521e-03   4.645 3.41e-06 ***
EducationLow_Education        -1.704e+01  4.614e+02  -0.037  0.97054
SexMale                       -7.620e-01  5.285e-01  -1.442  0.14933
Hours_PW                       3.742e-02  7.619e-03   4.911 9.04e-07 ***
Marital_StatusUnmarried       -5.198e+00  9.421e-01  -5.517 3.44e-08 ***
Occupation2                    5.189e-01  6.916e-01   0.750  0.45305
Occupation3                   -1.388e+00  6.260e-01  -2.218  0.02656 *
Occupation4                   -2.057e+00  7.192e-01  -2.860  0.00423 **
```

```
Occupation5                            1.500e+01  1.069e+03   0.014  0.98880
Occupation6                           -1.430e+00  7.775e-01  -1.839  0.06588 .
Age:Marital_StatusUnmarried            2.268e-02  1.443e-02   1.572  0.11594
EducationLow_Education:SexMale         1.384e+01  4.614e+02   0.030  0.97608
EducationLow_Education:Hours_PW        4.163e-02  2.911e-02   1.430  0.15264
SexMale:Marital_StatusUnmarried        1.647e+00  5.299e-01   3.107  0.00189 **
SexMale:Occupation2                    2.714e-01  7.078e-01   0.383  0.70137
SexMale:Occupation3                    1.591e+00  7.707e-01   2.064  0.03899 *
SexMale:Occupation4                    1.022e+00  7.444e-01   1.373  0.16987
SexMale:Occupation5                   -1.790e+01  1.069e+03  -0.017  0.98664
SexMale:Occupation6                    2.372e-01  7.994e-01   0.297  0.76666
Marital_StatusUnmarried:Occupation2    6.281e-01  6.740e-01   0.932  0.35144
Marital_StatusUnmarried:Occupation3   -3.286e-01  9.822e-01  -0.335  0.73795
Marital_StatusUnmarried:Occupation4    1.450e+00  6.810e-01   2.129  0.03329 *
Marital_StatusUnmarried:Occupation5   -1.209e+01  1.069e+03  -0.011  0.99097
Marital_StatusUnmarried:Occupation6    1.684e+00  6.627e-01   2.541  0.01104 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1552.7  on 1384  degrees of freedom
Residual deviance: 1030.1  on 1360  degrees of freedom
AIC: 1080.1

Number of Fisher Scoring iterations: 16



Call:
glm(formula = Income ~ Age + Education + Sex + Hours_PW + Marital_Status +
    Occupation + Sex:Marital_Status + Sex:Occupation, family = binomial(link = "logit"),
    data = data.new.2)

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
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1552.7  on 1384  degrees of freedom
Residual deviance: 1050.7  on 1368  degrees of freedom
AIC: 1084.7

Number of Fisher Scoring iterations: 6

Implementation: ROI | Solver: lpsolve
Separation: FALSE
Existence of maximum likelihood estimates
                     (Intercept)                              Age
                               0                                0
         EducationLow_Education                          SexMale
                               0                                0
                        Hours_PW         Marital_StatusUnmarried
                               0                                0
                     Occupation2                      Occupation3
                               0                                0
                     Occupation4                      Occupation5
                               0                                0
                     Occupation6 SexMale:Marital_StatusUnmarried
                               0                                0
            SexMale:Occupation2              SexMale:Occupation3
                               0                                0
            SexMale:Occupation4              SexMale:Occupation5
                               0                                0
            SexMale:Occupation6
                               0
0: finite value, Inf: infinity, -Inf: -infinity
```

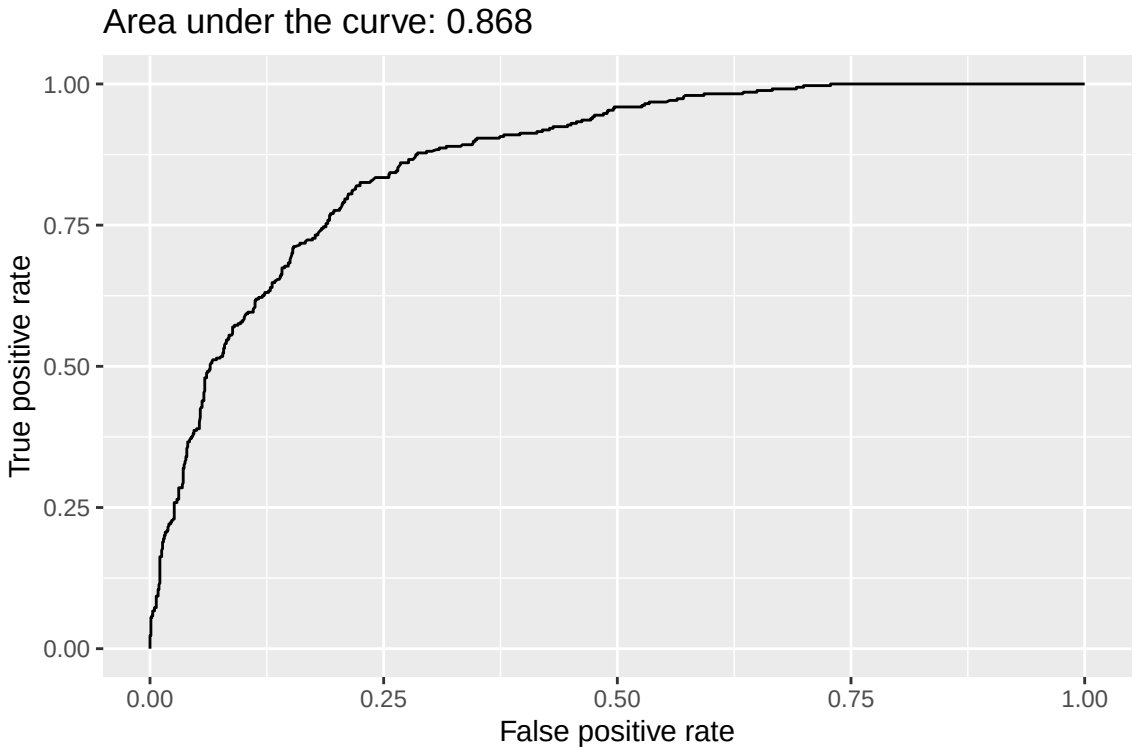


Figure 41: ROC curve for model_new.2.interaction

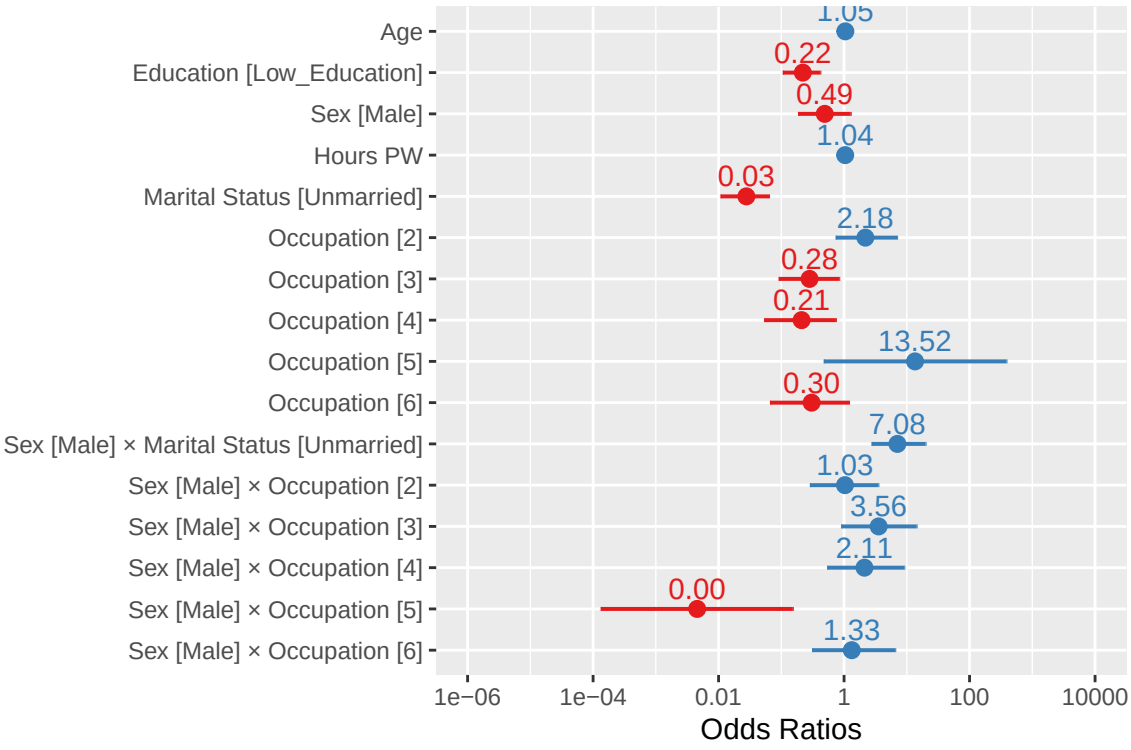


Figure 42: Model Plot for model_new.2.interaction

```
Call:
glm(formula = Income ~ Age + Education + Sex + Hours_PW + Marital_Status +
    Occupation + Age:Occupation + Education:Sex + Education:Marital_Status +
    Sex:Marital_Status, family = binomial(link = "logit"), data = data.new.3)

Coefficients:
                                      Estimate Std. Error z value
(Intercept)                         -3.518053   0.856736  -4.106
Age                                  0.058001   0.017461   3.322
EducationLow_Education             -15.965073 510.132302  -0.031
SexMale                             -0.339964   0.294255  -1.155
Hours_PW                             0.041293   0.007539   5.478
Marital_StatusUnmarried             -3.377304   0.397452  -8.497
Occupation2                          1.729139   1.031589   1.676
Occupation3                         -2.075271   2.716100  -0.764
Occupation4                         -0.462935   1.188057  -0.390
Occupation5                          0.510018   0.950263   0.537
```

```
Occupation6                                          1.765975   2.241941   0.788
Occupation7                                         -0.117523   1.030062  -0.114
Occupation8                                         -1.747957   1.218622  -1.434
Occupation9                                         -5.786074   2.476922  -2.336
Age:Occupation2                                     -0.022482   0.024172  -0.930
Age:Occupation3                                      0.073706   0.065199   1.130
Age:Occupation4                                     -0.001889   0.027509  -0.069
Age:Occupation5                                     -0.033699   0.021804  -1.546
Age:Occupation6                                     -0.096113   0.051625  -1.862
Age:Occupation7                                     -0.016641   0.023975  -0.694
Age:Occupation8                                      0.013741   0.027381   0.502
Age:Occupation9                                      0.123809   0.056975   2.173
EducationLow_Education:SexMale                      14.141329 510.132309   0.028
EducationLow_Education:Marital_StatusUnmarried       1.354707   0.750523   1.805
SexMale:Marital_StatusUnmarried                      1.663952   0.456993   3.641
                                                   Pr(>|z|)
(Intercept)                                        4.02e-05 ***
Age                                                0.000895 ***
EducationLow_Education                             0.975034
SexMale                                            0.247951
Hours_PW                                           4.31e-08 ***
Marital_StatusUnmarried                             < 2e-16 ***
Occupation2                                        0.093701 .
Occupation3                                        0.444830
Occupation4                                        0.696790
Occupation5                                        0.591466
Occupation6                                        0.430873
Occupation7                                        0.909164
Occupation8                                        0.151466
Occupation9                                        0.019492 *
Age:Occupation2                                    0.352339
Age:Occupation3                                    0.258277
Age:Occupation4                                    0.945240
Age:Occupation5                                    0.122222
Age:Occupation6                                    0.062640 .
Age:Occupation7                                    0.487613
Age:Occupation8                                    0.615767
Age:Occupation9                                    0.029776 *
EducationLow_Education:SexMale                     0.977885
EducationLow_Education:Marital_StatusUnmarried     0.071072 .
SexMale:Marital_StatusUnmarried                    0.000271 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1552.7  on 1384  degrees of freedom
Residual deviance: 1036.0  on 1360  degrees of freedom
AIC: 1086

Number of Fisher Scoring iterations: 16


Call:
glm(formula = Income ~ Age + Sex + Education + Hours_PW + Marital_Status +
    Occupation + Marital_Status:Occupation, family = binomial(link = "logit"),
    data = data.new.3)

Coefficients:
                                   Estimate Std. Error z value Pr(>|z|)
(Intercept)                       -3.304748   0.529508  -6.241 4.34e-10 ***
Age                                0.044727   0.006817   6.561 5.34e-11 ***
SexMale                            0.445923   0.222749   2.002  0.04529 *
EducationLow_Education            -1.446628   0.346257  -4.178 2.94e-05 ***
Hours_PW                           0.037683   0.007366   5.116 3.12e-07 ***
Marital_StatusUnmarried           -2.915217   0.531143  -5.489 4.05e-08 ***
Occupation2                        0.722076   0.336026   2.149  0.03164 *
Occupation3                        0.697480   0.692564   1.007  0.31389
Occupation4                       -0.208668   0.380242  -0.549  0.58316
Occupation5                       -1.264080   0.288458  -4.382 1.17e-05 ***
Occupation6                       -3.060971   0.609949  -5.018 5.21e-07 ***
Occupation7                       -1.186279   0.297107  -3.993 6.53e-05 ***
Occupation8                       -1.623163   0.330524  -4.911 9.07e-07 ***
Occupation9                       -1.035350   0.586947  -1.764  0.07774 .
Marital_StatusUnmarried:Occupation2  0.525588   0.648670   0.810  0.41779
Marital_StatusUnmarried:Occupation3  0.213914   1.136972   0.188  0.85076
Marital_StatusUnmarried:Occupation4 -0.879872   0.941130  -0.935  0.34983
Marital_StatusUnmarried:Occupation5  1.301111   0.658313   1.976  0.04811 *
Marital_StatusUnmarried:Occupation6  3.306006   1.320659   2.503  0.01230 *
Marital_StatusUnmarried:Occupation7  1.645927   0.717551   2.294  0.02180 *
Marital_StatusUnmarried:Occupation8  2.161565   0.714046   3.027  0.00247 **
Marital_StatusUnmarried:Occupation9  0.676033   1.272549   0.531  0.59525
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1552.7  on 1384  degrees of freedom
Residual deviance: 1054.8  on 1363  degrees of freedom
AIC: 1098.8

Number of Fisher Scoring iterations: 6


Implementation: ROI | Solver: lpsolve
Separation: FALSE
Existence of maximum likelihood estimates
                          (Intercept)                                   Age
                                    0                                     0
                              SexMale                EducationLow_Education
                                    0                                     0
                            Hours_PW               Marital_StatusUnmarried
                                    0                                     0
                           Occupation2                           Occupation3
                                    0                                     0
                           Occupation4                           Occupation5
                                    0                                     0
                           Occupation6                           Occupation7
                                    0                                     0
                           Occupation8                           Occupation9
                                    0                                     0
Marital_StatusUnmarried:Occupation2 Marital_StatusUnmarried:Occupation3
                                    0                                     0
Marital_StatusUnmarried:Occupation4 Marital_StatusUnmarried:Occupation5
                                    0                                     0
Marital_StatusUnmarried:Occupation6 Marital_StatusUnmarried:Occupation7
                                    0                                     0
Marital_StatusUnmarried:Occupation8 Marital_StatusUnmarried:Occupation9
                                    0                                     0
0: finite value, Inf: infinity, -Inf: -infinity
```

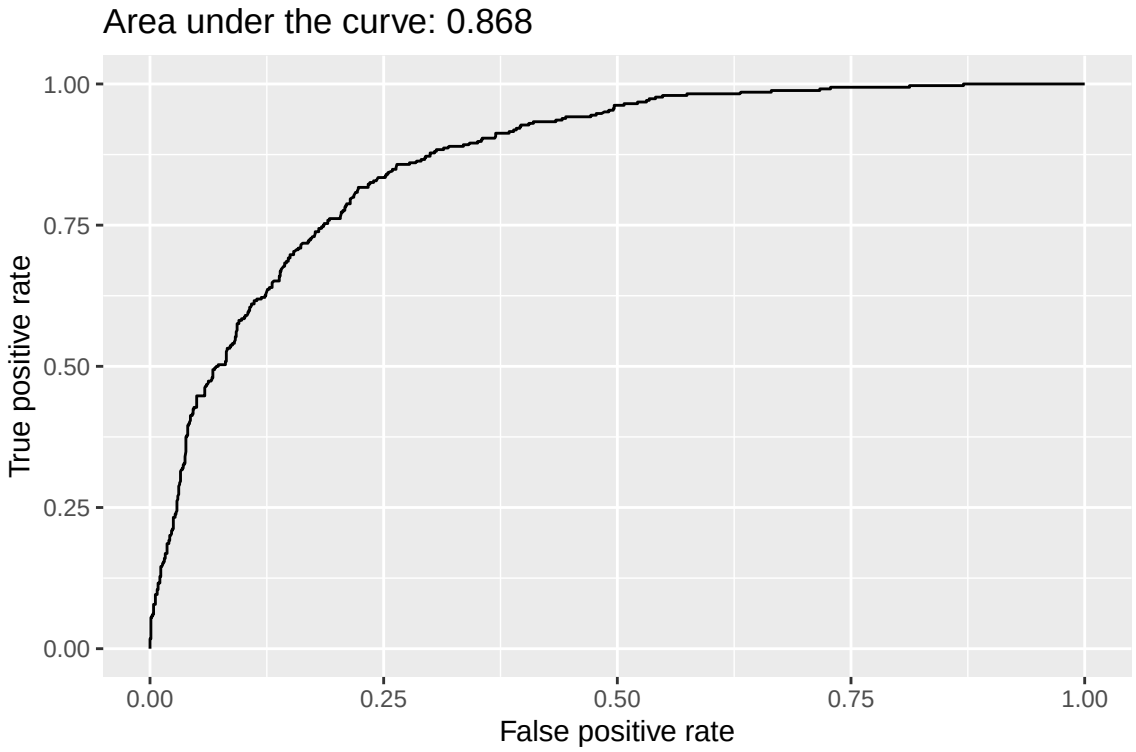


Figure 43: ROC curve for model_new.3.interaction

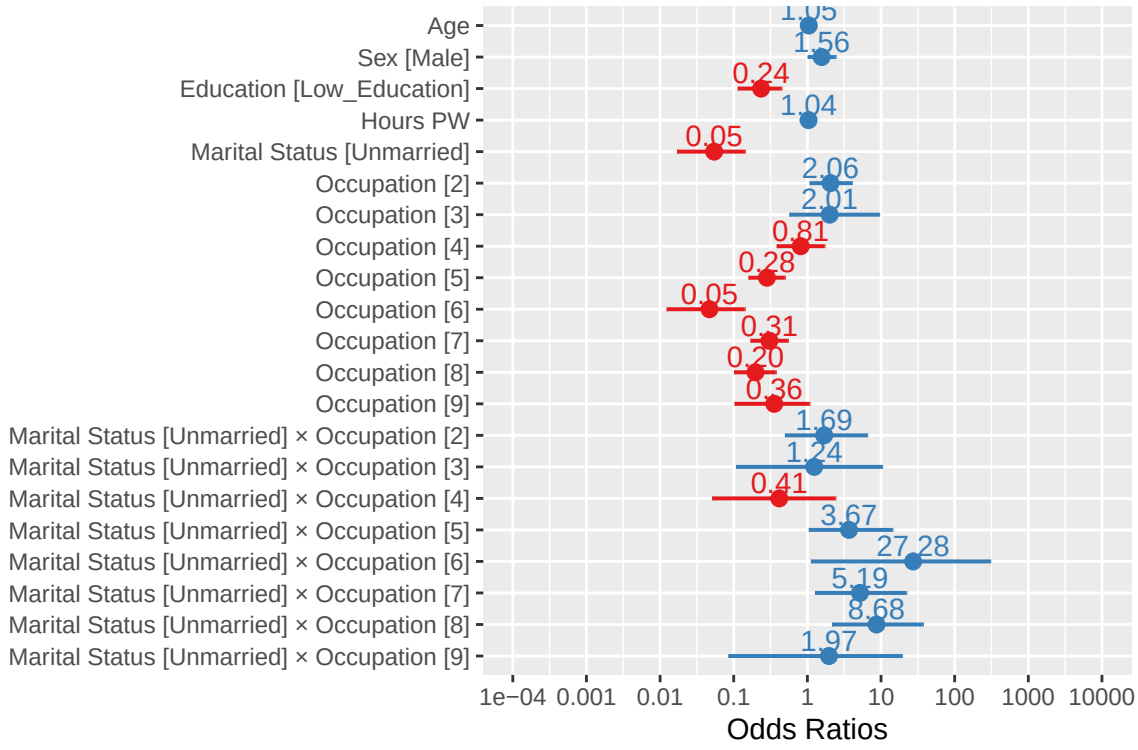


Figure 44: Model Plot for model_new.3.interaction

To enhance predictive accuracy and account for potential interdependencies, interaction terms between variables are introduced into the logistic regression model. This allows us to evaluate how combinations of factors influence income classification (<=50K vs. >50K).

Key Interaction Effects Identified Model using Physical vs. Mental Labor (model_new.1.interaction): Significant interactions found: Age × Marital Status Sex × Marital Status This suggests that age influences the impact of marital status on income, and gender differences in income depend on marital status. Model using PRC Job Classification (model_new.2.interaction):

Additional significant interactions: Sex × Marital Status Sex × Occupation Marital Status × Occupation This highlights that gender effects on income vary across job categories, and marital status plays a role in occupational income disparities. Model using ISCO-08 Job Classification (model_new.3.interaction):

The most significant interaction: Marital Status × Occupation This suggests that marital status interacts with occupation type to influence income, possibly due to differences in career stability or dual-income households.

Model Performance Assessment ROC curve analysis: The AUC values are compared for interaction models, showing whether incorporating interaction effects improves classification

performance. Detect separation tests: Ensures that models do not suffer from convergence issues caused by perfect separation. Stepwise AIC optimization: Helps reduce overfitting by removing redundant interaction terms.

Key Takeaways: Including interaction terms improves model interpretability and accuracy. Marital Status and Occupation consistently interact, meaning these two factors jointly affect income potential. 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.

5 Model Assumption

The Hosmer-Lemeshow test p-values for all seven models are greater than 0.05, indicating a good model fit. The predicted probabilities do not show a significant difference from the actual data distribution, confirming that the models are suitable for use.

```
    Hosmer and Lemeshow goodness of fit (GOF) test

data:  stepAIC_model$y, fitted(stepAIC_model)
X-squared = 3.5824, df = 8, p-value = 0.8927
```

```
    Hosmer and Lemeshow goodness of fit (GOF) test

data:  stepAIC_model_new.1$y, fitted(stepAIC_model_new.1)
X-squared = 7.7147, df = 8, p-value = 0.4618
```

```
    Hosmer and Lemeshow goodness of fit (GOF) test

data:  stepAIC_model_new.2$y, fitted(stepAIC_model_new.2)
X-squared = 6.6156, df = 8, p-value = 0.5786
```

```
    Hosmer and Lemeshow goodness of fit (GOF) test

data:  stepAIC_model_new.3$y, fitted(stepAIC_model_new.3)
X-squared = 10.221, df = 8, p-value = 0.2499
```

```
    Hosmer and Lemeshow goodness of fit (GOF) test

data:  model_new.1.interaction$y, fitted(model_new.1.interaction)
X-squared = 9.4961, df = 8, p-value = 0.3022


    Hosmer and Lemeshow goodness of fit (GOF) test

data:  model_new.2.interaction$y, fitted(model_new.2.interaction)
X-squared = 9.1216, df = 8, p-value = 0.3321


    Hosmer and Lemeshow goodness of fit (GOF) test

data:  model_new.3.interaction$y, fitted(model_new.3.interaction)
X-squared = 6.8584, df = 8, p-value = 0.552
```

When evaluating the prediction accuracy of the seven models, all models achieved an accuracy greater than 0.8, with a difference of no more than 0.02 between them. This indicates that the models can effectively distinguish between high-income (>50K) and low-income (<=50K) groups.

```
         Actual
Predicted <=50K >50K
    <=50K   958  154
    >50K     83  190

[1] "Accuracy1: 0.829"

         Actual
Predicted <=50K >50K
    <=50K   942  165
    >50K     99  179

[1] "Accuracy2: 0.809"

         Actual
Predicted <=50K >50K
    <=50K   961  169
    >50K     80  175
```

```
[1] "Accuracy3: 0.82"

         Actual
Predicted <=50K >50K
    <=50K   965  169
    >50K     76  175

[1] "Accuracy4: 0.823"

         Actual
Predicted <=50K >50K
    <=50K   938  150
    >50K    103  194

[1] "Accuracy5: 0.817"

         Actual
Predicted <=50K >50K
    <=50K   958  161
    >50K     83  183

[1] "Accuracy6: 0.824"

         Actual
Predicted <=50K >50K
    <=50K   957  171
    >50K     84  173

[1] "Accuracy7: 0.816"
```

The AIC of `model_new.2.interaction` is the smallest, indicating the best model fit. Additionally, `model_new.2.interaction` achieves the highest AUC of 0.868, demonstrating the strongest ability to classify high-income (>50K) and low-income (<=50K) groups.

```
                        df      AIC
stepAIC_model           34 1066.601
stepAIC_model_new.1      7 1123.930
stepAIC_model_new.2     11 1101.458
stepAIC_model_new.3     14 1106.393
model_new.1.interaction  9 1108.156
model_new.2.interaction 17 1084.720
model_new.3.interaction 22 1098.772
```

The X-axis ($0 \rightarrow 1$) represents the predicted probability, where 0 indicates low income, 1 indicates high income, and 0.5 represents model uncertainty. The Y-axis (residuals) represents the prediction error. A residual value close to 0 indicates that the predicted result is close to the actual value, with a small error, while a residual value far from 0 indicates a larger discrepancy between the predicted and actual values.
At points where the X-axis is close to 0 and 1, residual values are larger, indicating higher prediction errors in these regions. However, this is a normal phenomenon in logistic regression. When the X-axis is near 0.5, the residual values are smaller, suggesting that the model fits better in the middle probability range.

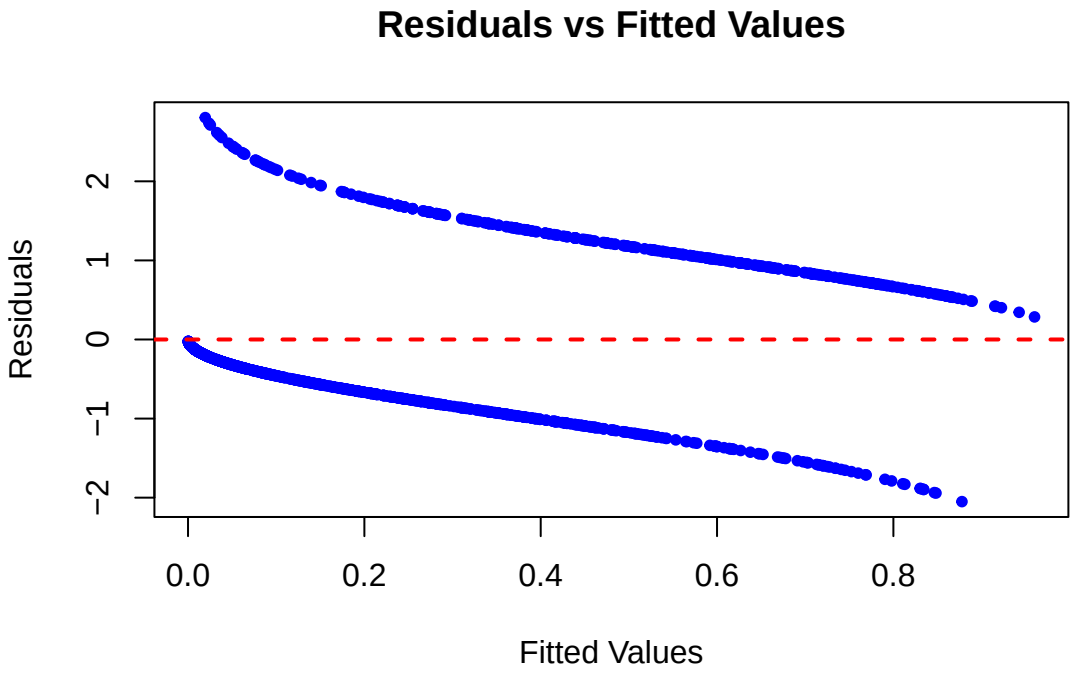


Figure 45: Residuals vs Fitted Values.

The residuals follow a normal distribution in the middle region but deviate from the diagonal line at both ends, especially on the right side. However, slight tail deviations are common in logistic regression.

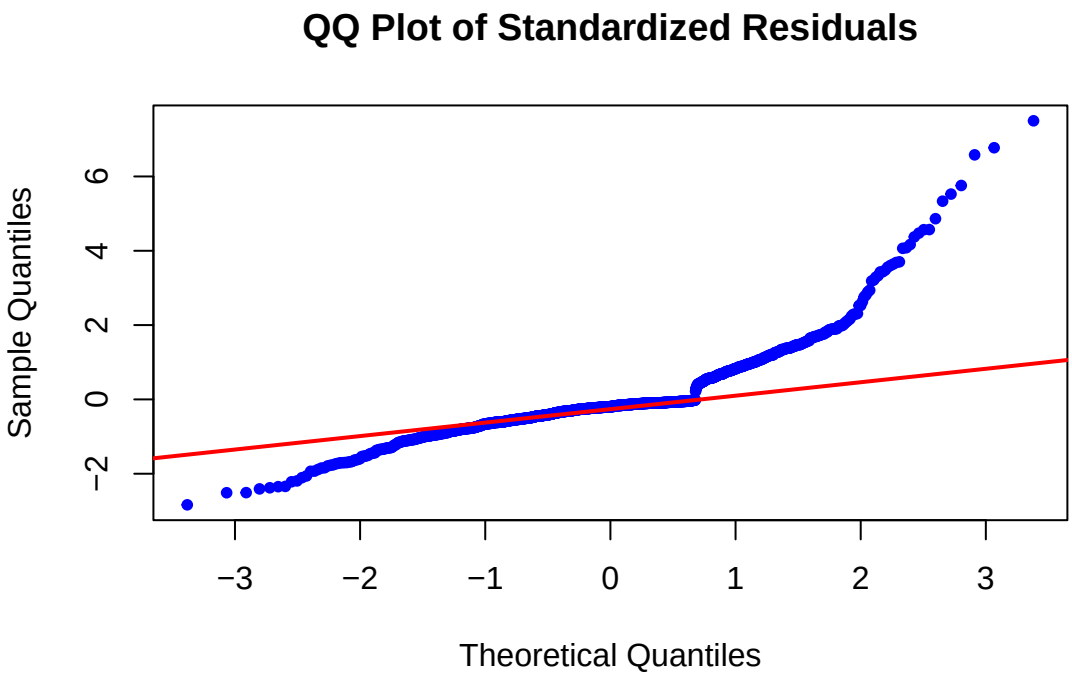


Figure 46: QQ Plot of Standardized Residuals.