

COP259: Data Mining Coursework

F128607

Part 1. Preprocessing

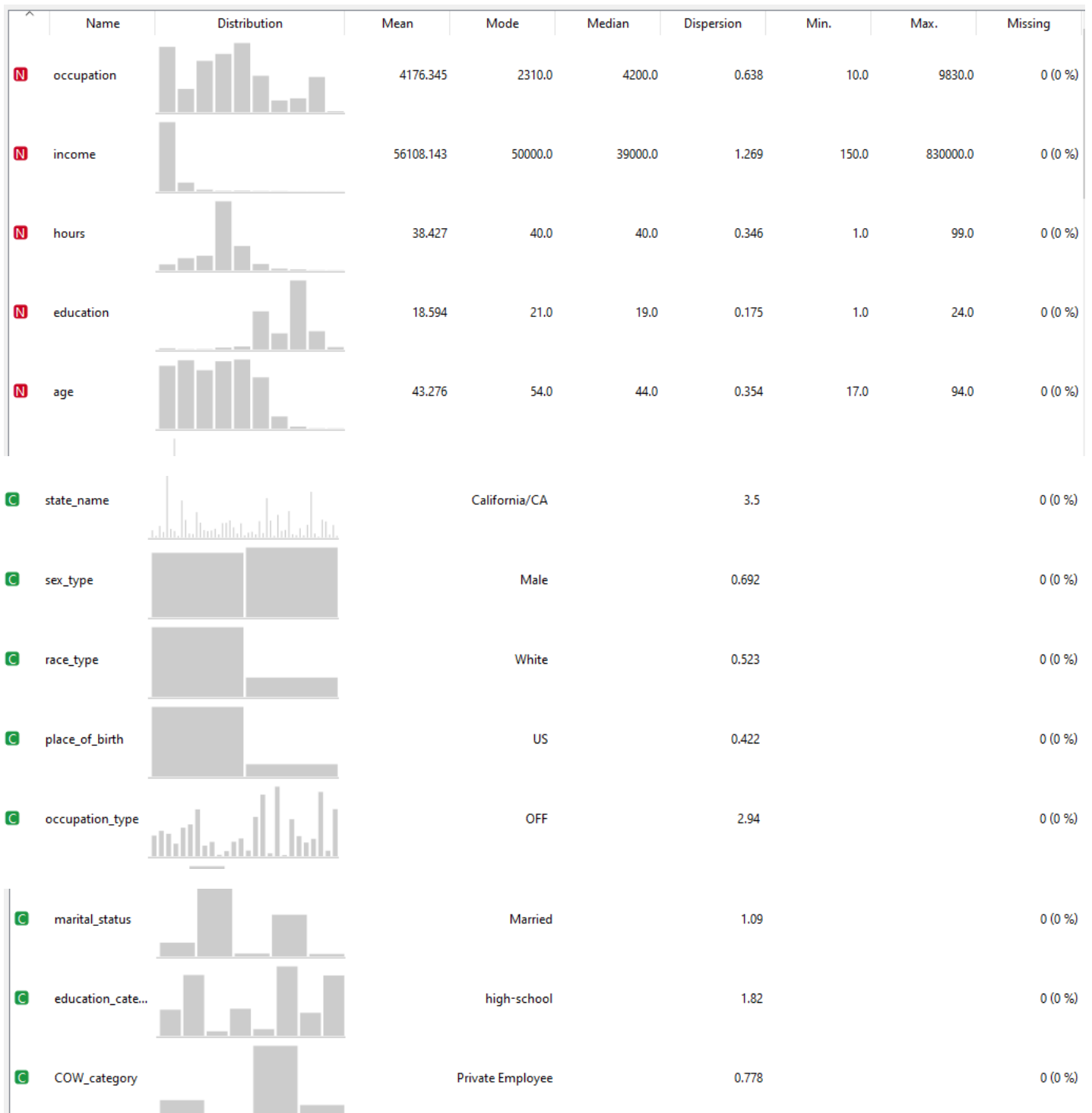


Figure 1 - exploratory data analysis to understand the dataset's structure and key characteristics
From the 5000 data-sample there was no noticeable outliers and missing data.

Part 2. Fairness in income distribution

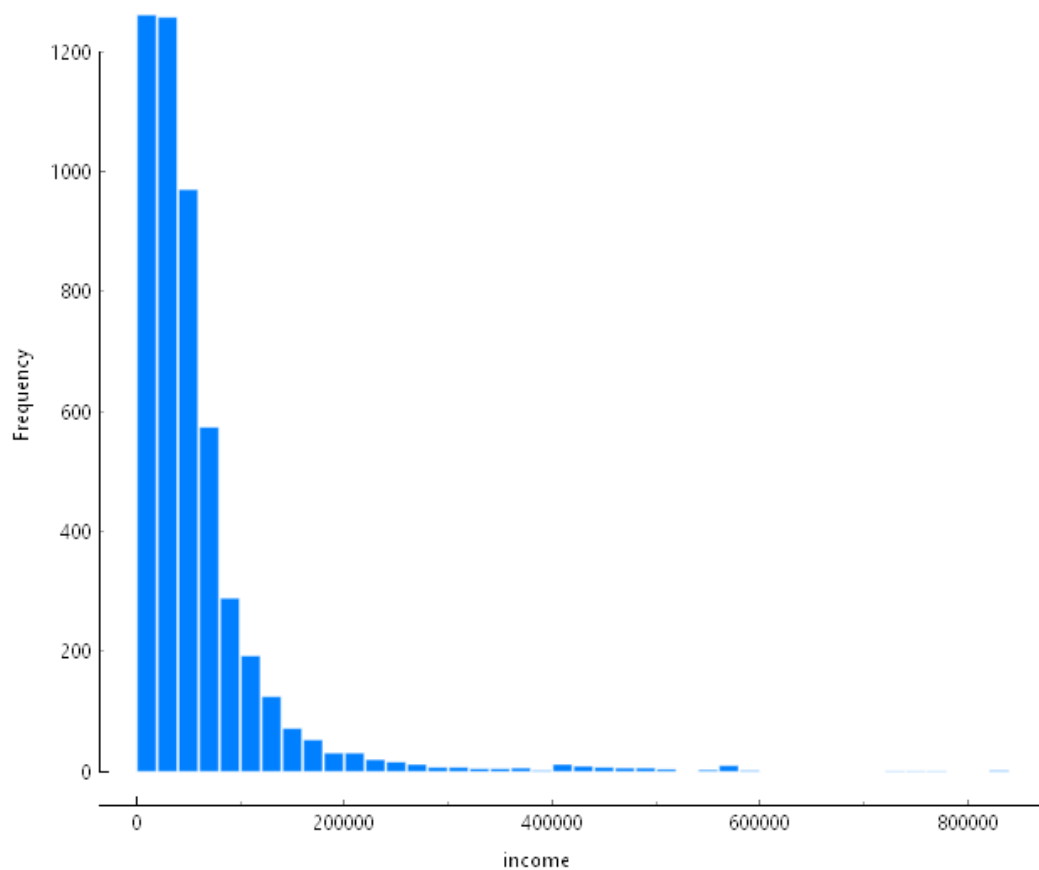


Figure 2 – histogram of income attribute

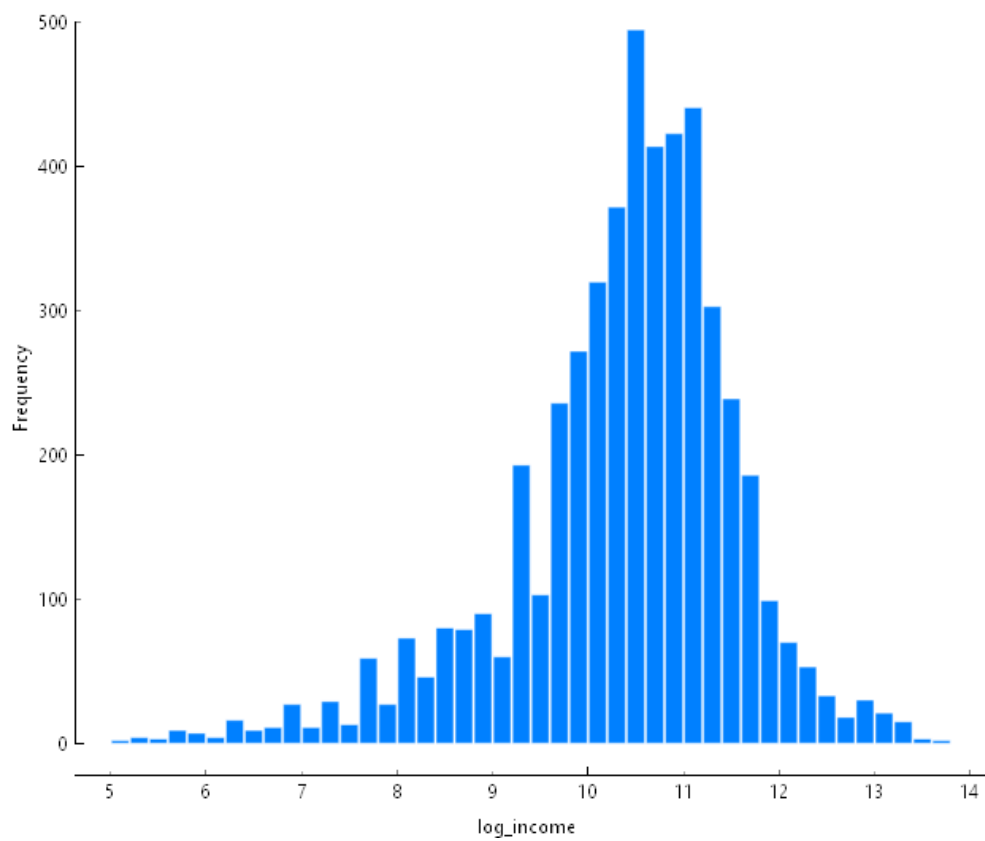


Figure 3 – histogram of log income attribute

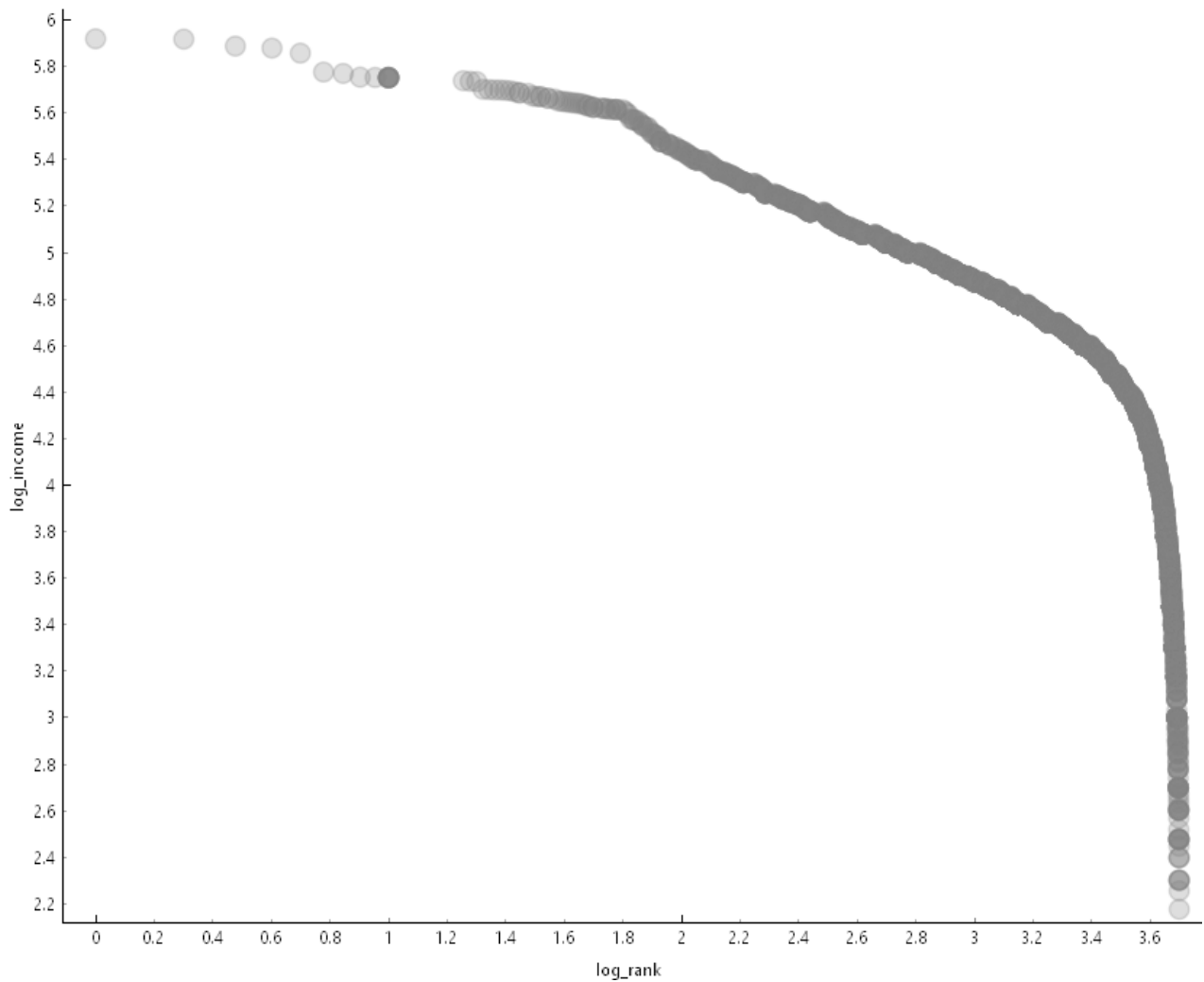


figure 4 – Zipf plot for income

Interpretation

Figure 2 shows a right-skewed income distribution, with most individuals earning below £100,000 and the highest concentration in the £20,000–£40,000 range. High incomes exist but are much less common. Figure 3's log-income histogram follows a bell-shaped curve, peaking at $\log(\text{Income}) \approx 10.6$ (£40,000), indicating that logarithmic transformation normalizes income distribution. This reinforces that most people earn within a middle-income range, while extreme values are rare. The Zipf plot reveals a gradual decline in log-income against log-rank, followed by a steep drop at $\log(\text{rank}) = 3.2$ and $\log(\text{income}) = 4.6$, highlighting sharp income disparities among top earners. This aligns with a power-law distribution, where the highest earners hold a

disproportionately large share of total wealth.

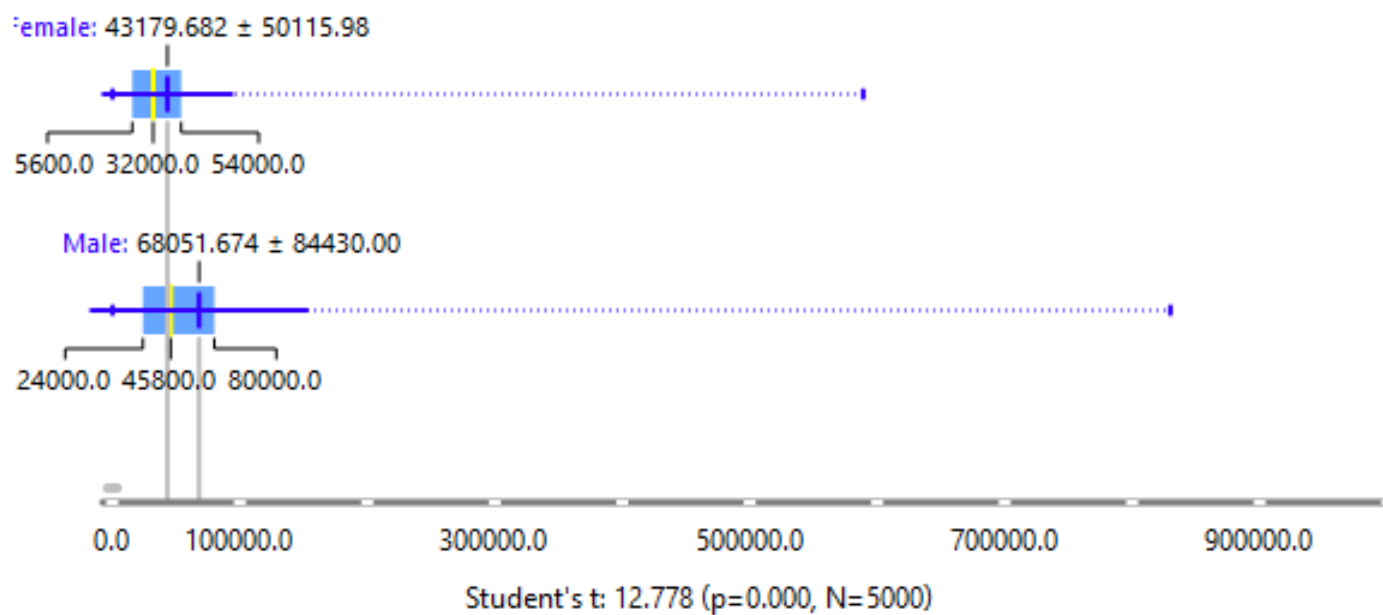


Figure 5 – boxplot income distribution broken down by sex plus t-test

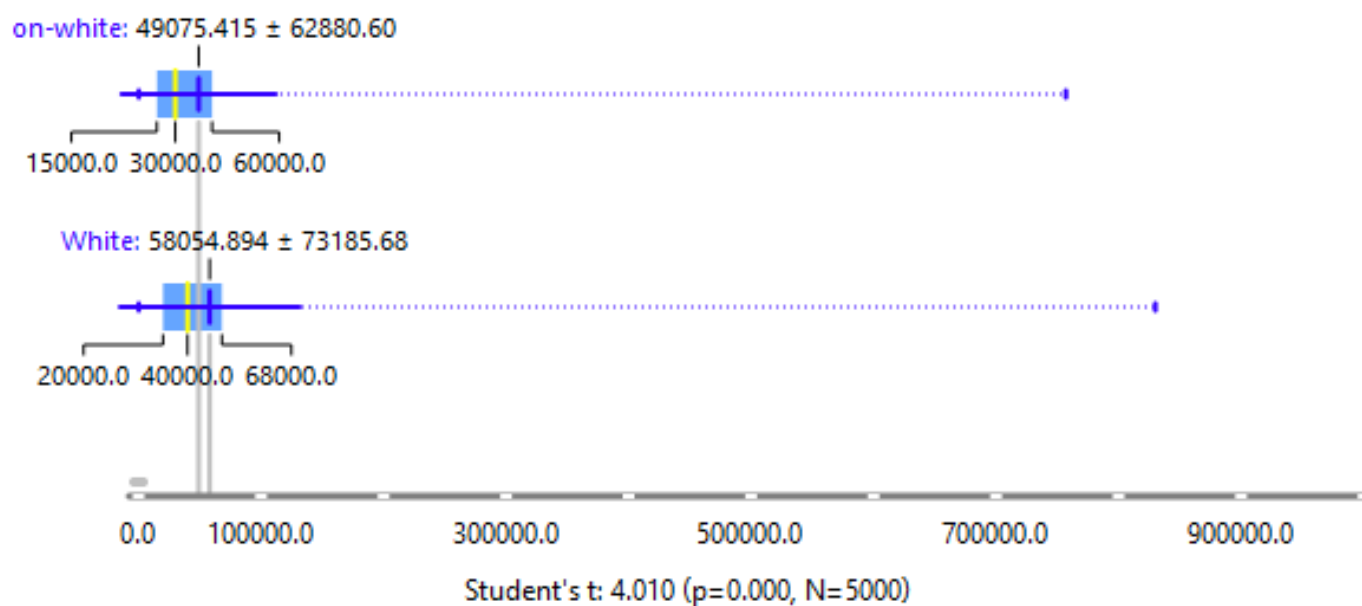


Figure 6 – boxplot income distribution broken down by race plus t-test

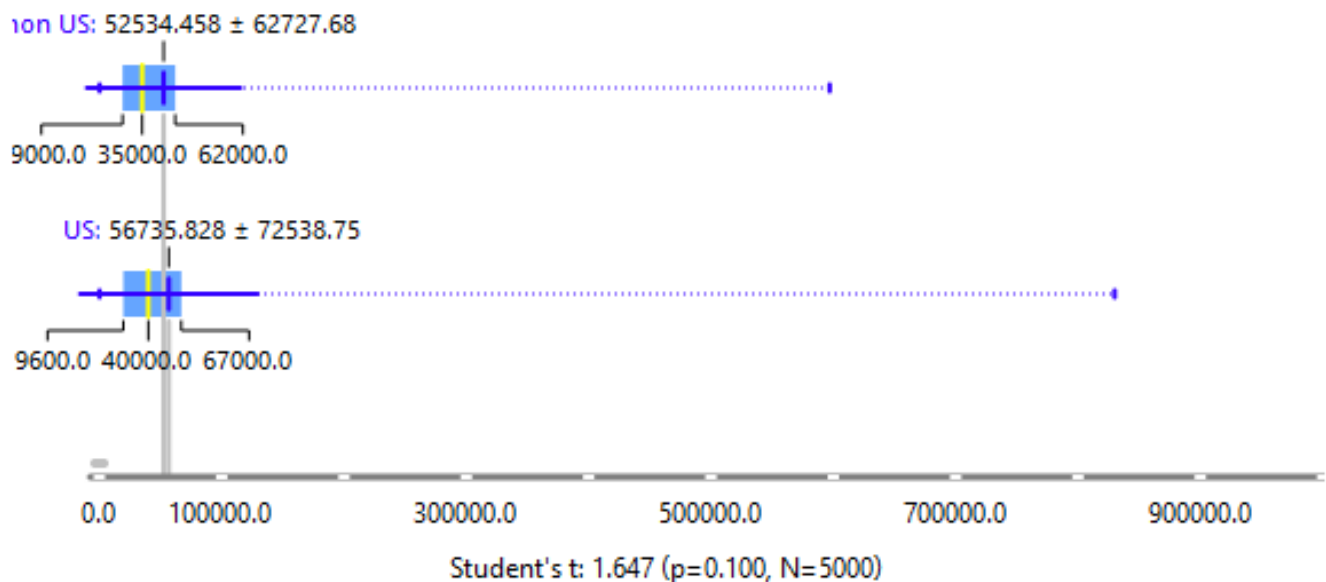


Figure 7 – boxplot income distribution broken down by place of birth plus t-test

Comments

First, males have a significantly higher mean income (\$68,051.67) than females (\$43,179.68), with greater variability (\$84,430.00 vs. \$50,115.98). The median income for males (\$45,800) also exceeds that of females (\$32,000), indicating a gender pay gap. The t-test ($t = 12.778$, $p = 0.000$) confirms this difference is statistically significant, suggesting systemic factors such as occupational differences. Second, White individuals have a higher mean income (\$58,054.89) than Non-White individuals (\$49,075.42), with greater variability (\$73,185.68 vs. \$62,880.60). The median income for White individuals (\$40,000) is also higher, reflecting income disparity. The t-test ($t = 4.010$, $p = 0.000$) confirms the difference is statistically significant, suggesting socio-economic factors at play. Finally, US-born individuals have a slightly higher mean income (\$56,735.83) compared to non-US-born individuals (\$52,534.46), with greater income variability. However, the t-test ($t = 1.647$, $p = 0.100$) shows the difference is not statistically significant, indicating weak evidence against the null hypothesis and no strong conclusion about the effect of place of birth on income.

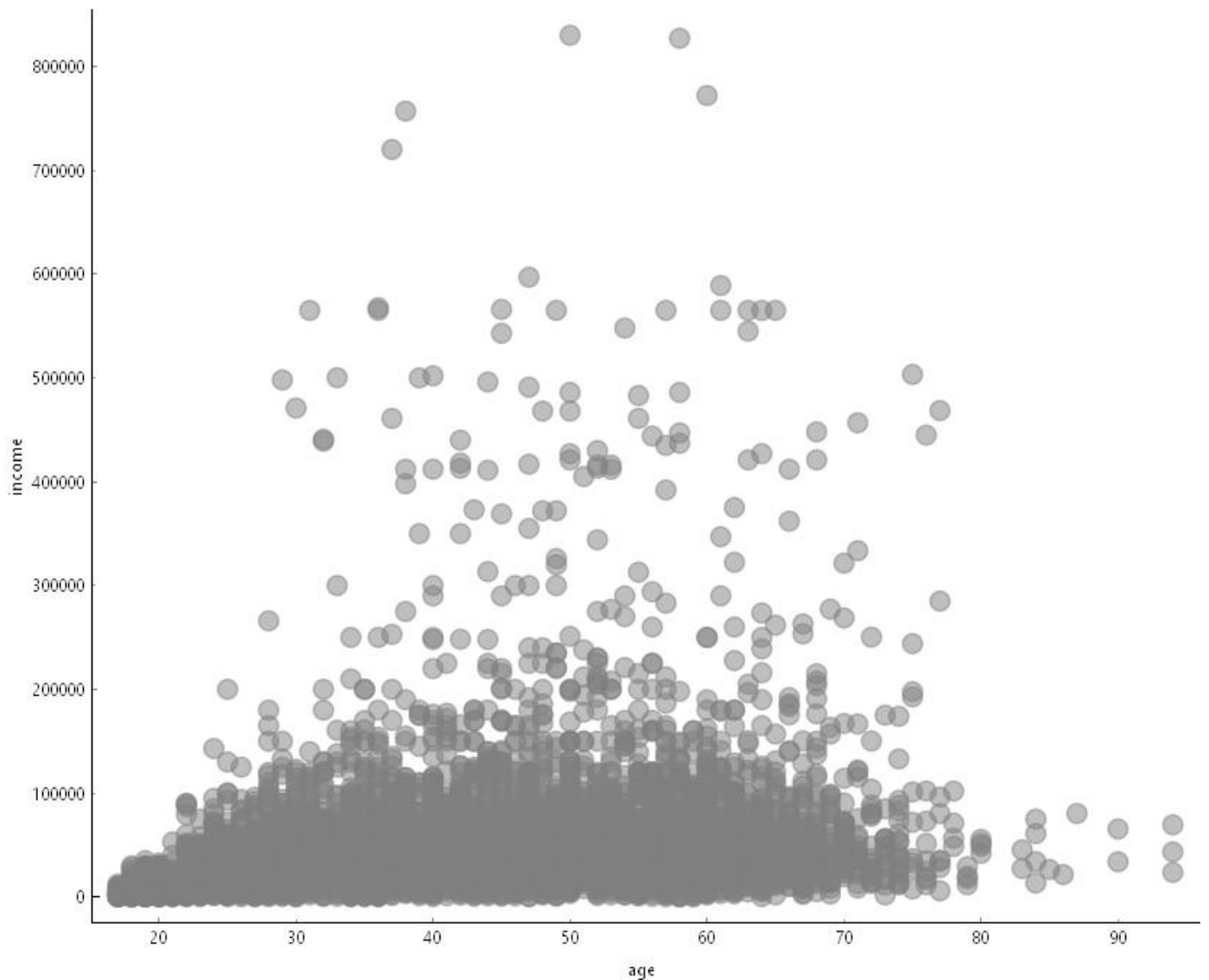


Figure 9 – scatter plot of age attribute against income attribute

Interpretation

In Figure 9, the scatter plot shows a weak correlation between age and income, with widely dispersed data points and no strong linear relationship. Income variance increases with age, particularly between ages 40 and 70, with some high incomes exceeding \$500,000. This suggests age is not a strong predictor of income, though middle-aged individuals may have higher earning potential.

In Figure 10, the scatter plot reveals a weak positive correlation between hours worked per week and income. Higher incomes are generally associated with more hours worked, but the relationship is not strictly linear. Income variability increases between 30 and 60 hours per week, with extreme outliers over \$500,000, suggesting other factors beyond hours worked affect income.

In Figure 11, the scatter plot shows a positive correlation between education (years) and income, with individuals having more education generally earning higher incomes. However, the relationship is not strictly linear, and earnings vary significantly at each education level. While education is a strong predictor of income, other factors are also influential, with some extreme outliers exceeding \$500,000.

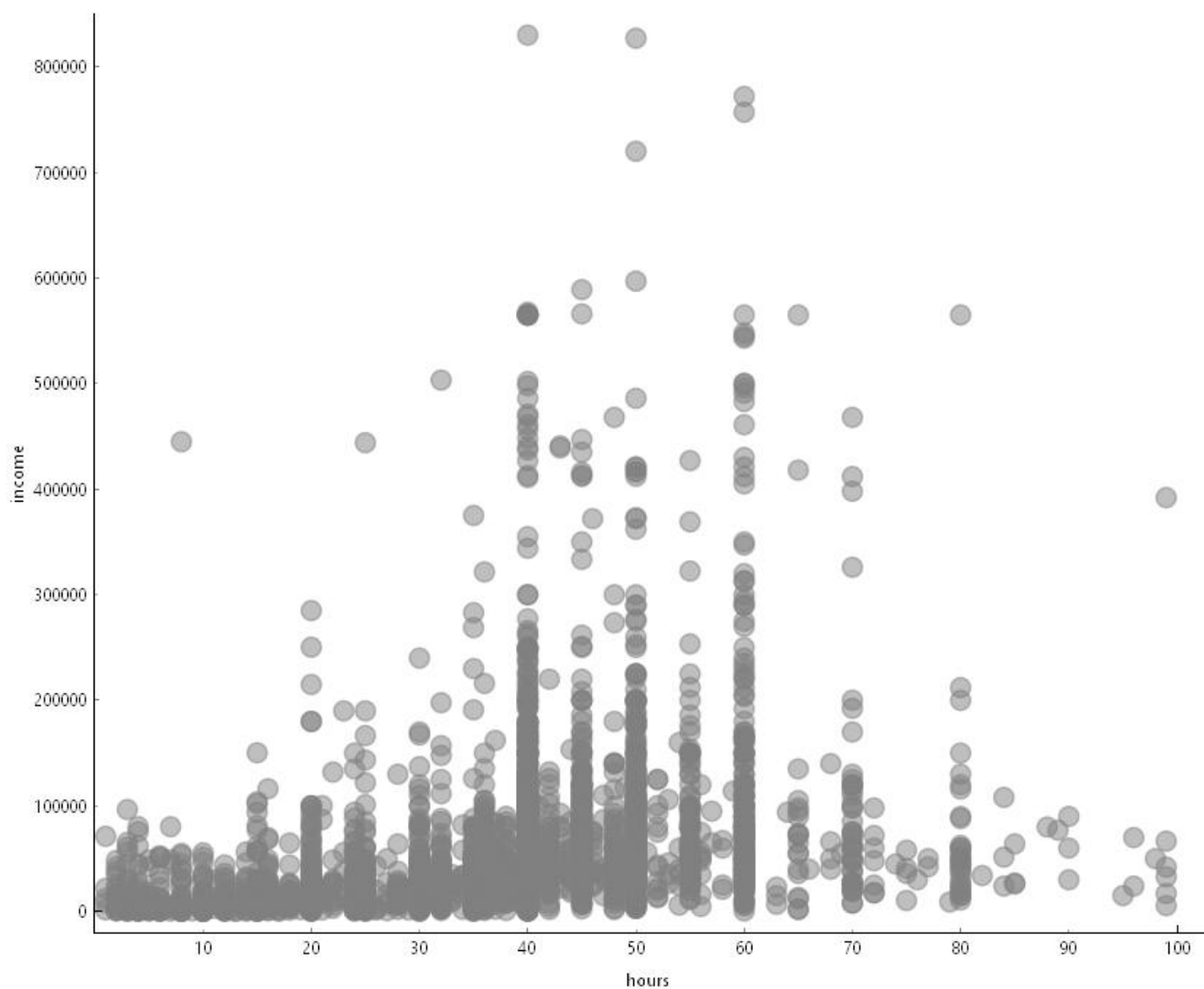


Figure 10 – scatter plot of hours worked against income attribute

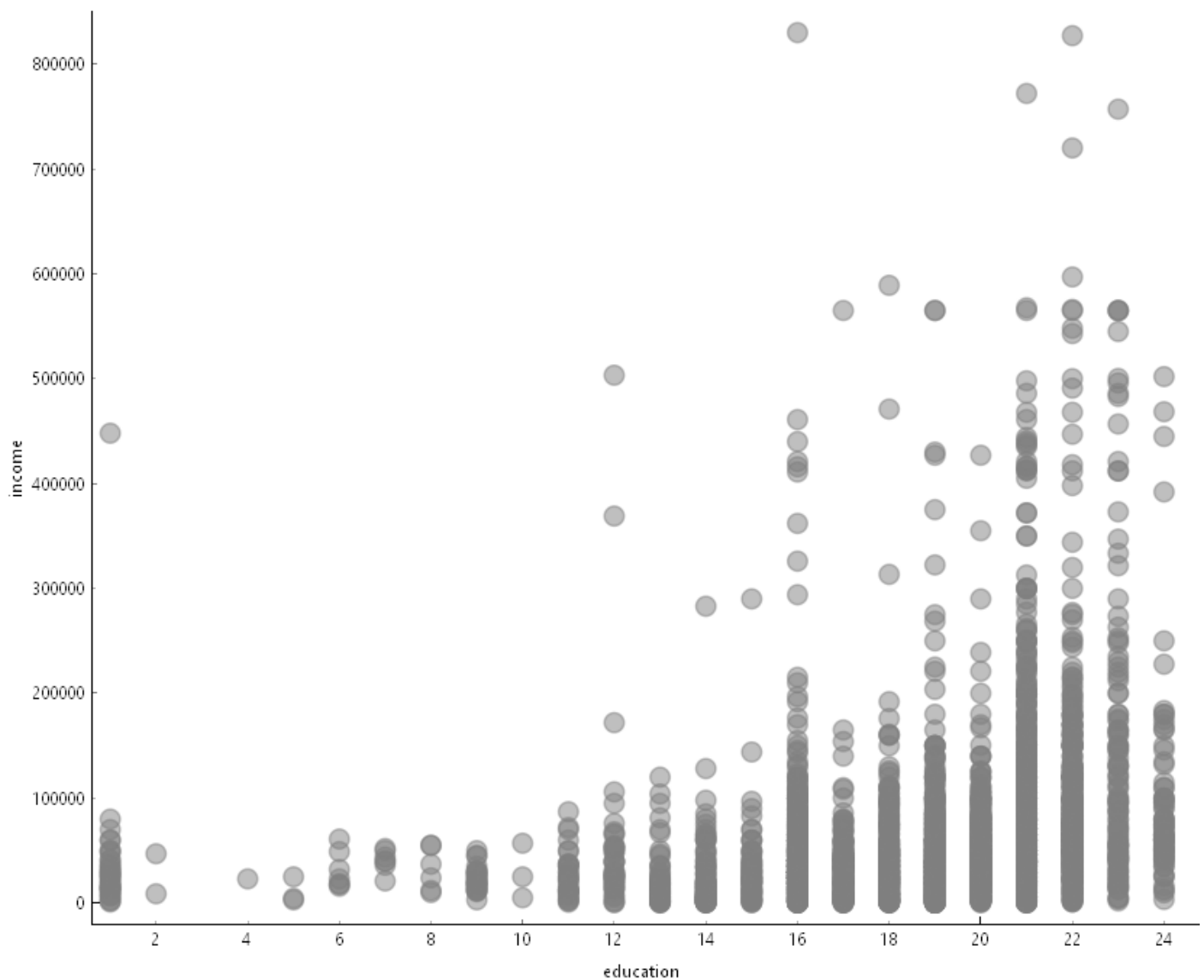


Figure 11 – scatter plot of years in education against income

1	+0.313	hours	:	income
2	+0.262	education	:	income
3	-0.241	income	:	occupation
4	+0.225	age	:	income

Figure 12 – computed Pearson correlation for income

The Pearson correlation coefficients reveal the strength and direction of relationships between income and various variables: hours worked per week (+0.313) shows the strongest positive correlation, indicating that more hours worked moderately increase income; education level (+0.262) also has a positive correlation, suggesting that higher education is linked to higher income, but the relationship is weaker; age (+0.225) shows the weakest positive correlation, indicating a slight increase in income with age; and occupation (-0.241) reveals a negative correlation, suggesting that certain occupations may be associated with lower incomes. Overall, income is most influenced by hours worked and education, while age has a weaker effect.

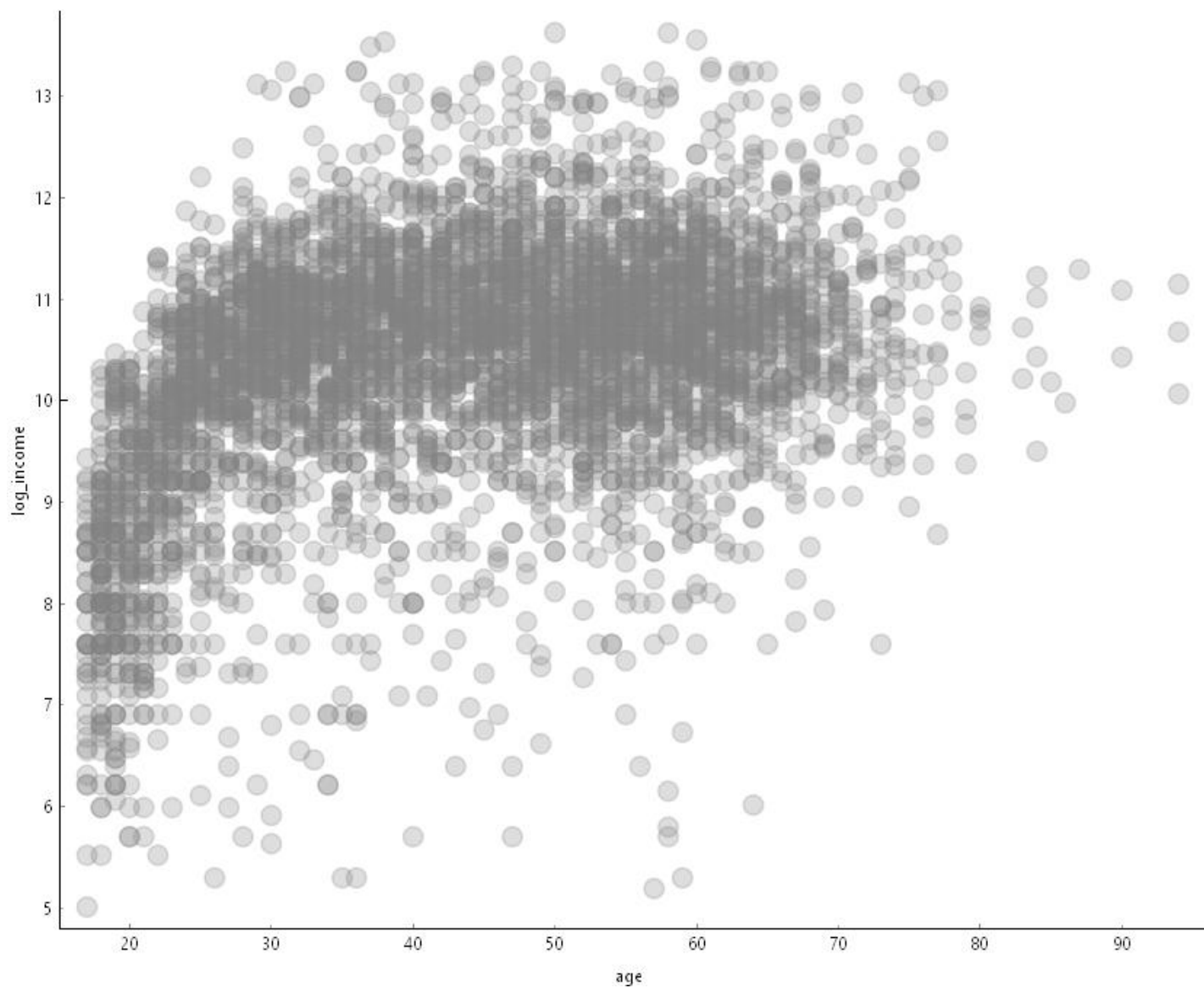


Figure 13 - scatter plot of age attribute against log-income attribute

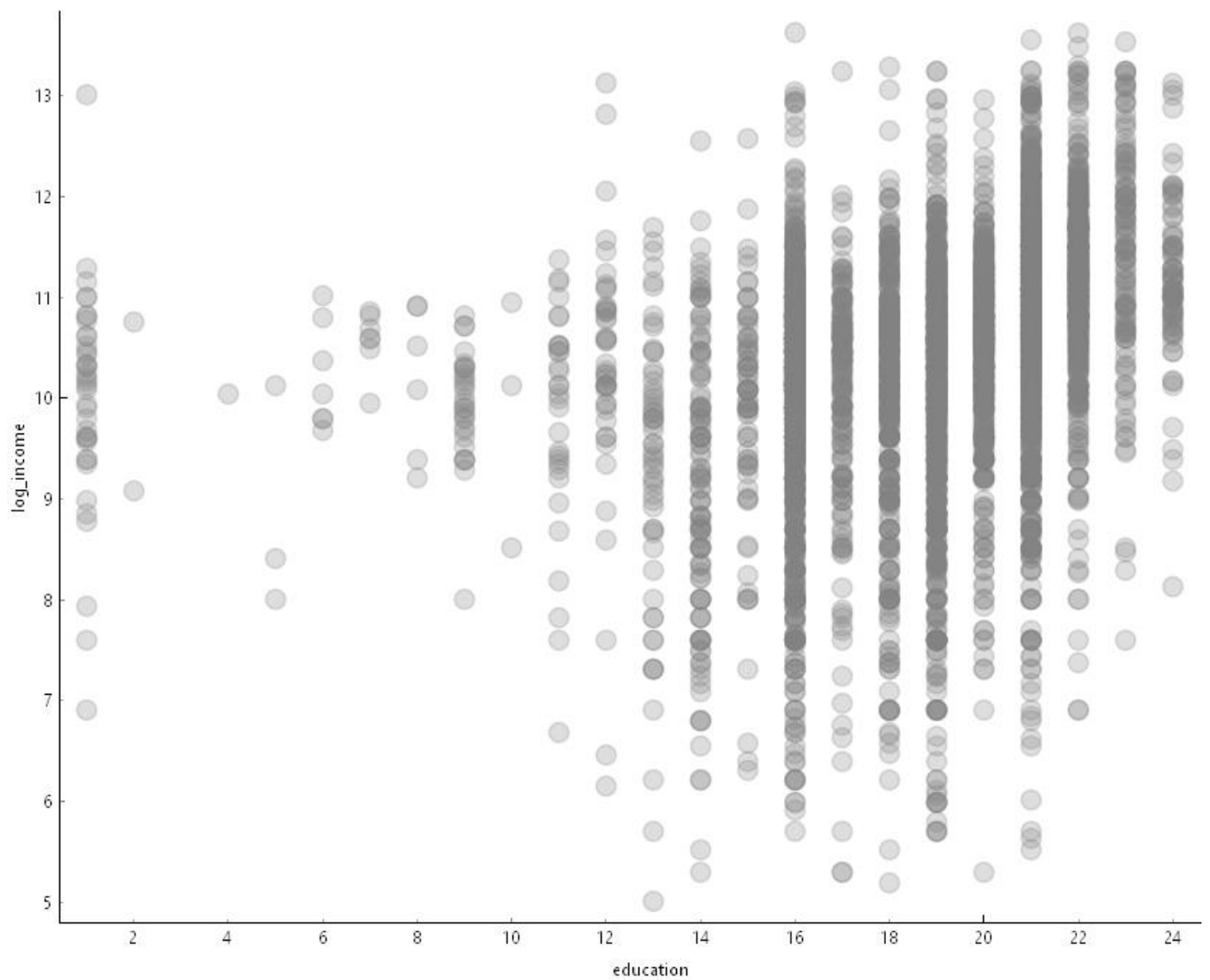


Figure 14 – scatter plot of hours worked against log-income attribute

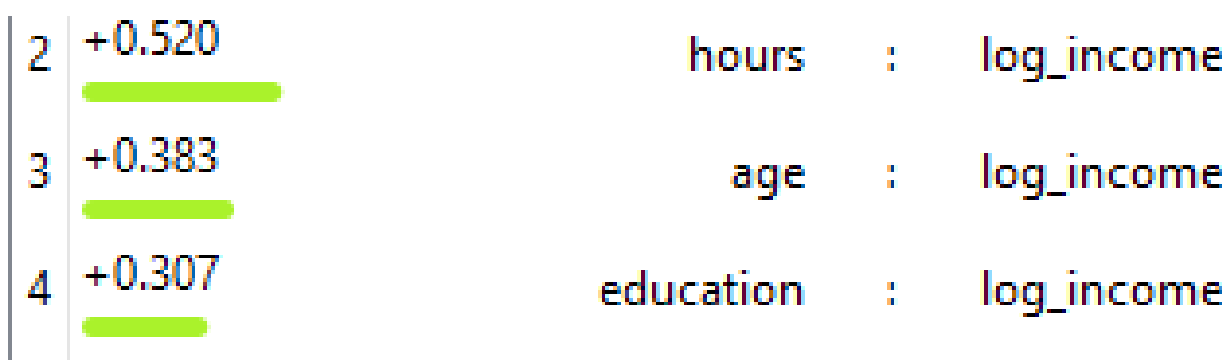


Figure 15 – Pearson correlation for log-income

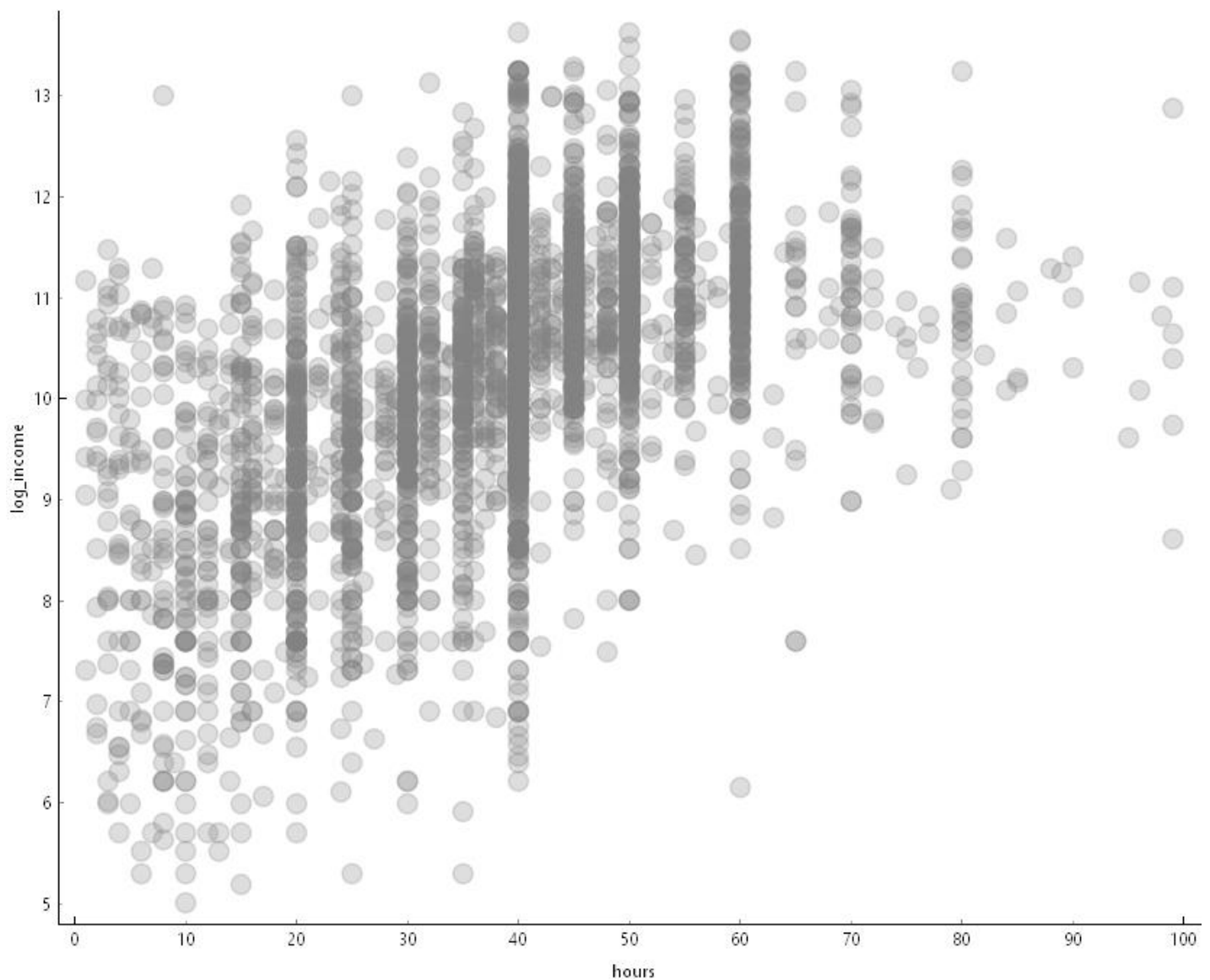


Figure 16 – scatter plot of years in education against log-income

Log-transformed income reveals clearer trends than raw income. For age, log-income rises in early career years, plateaus in middle age, and declines slightly later, reducing skewness and extreme values seen in raw income. Similarly, log-income shows a positive but diminishing correlation with education, with clearer clustering at key milestones. Compared to raw income, log transformation normalizes distribution, minimizes outliers, and enhances interpretability, highlighting structured income progression over time.

Part 3. Predicting income

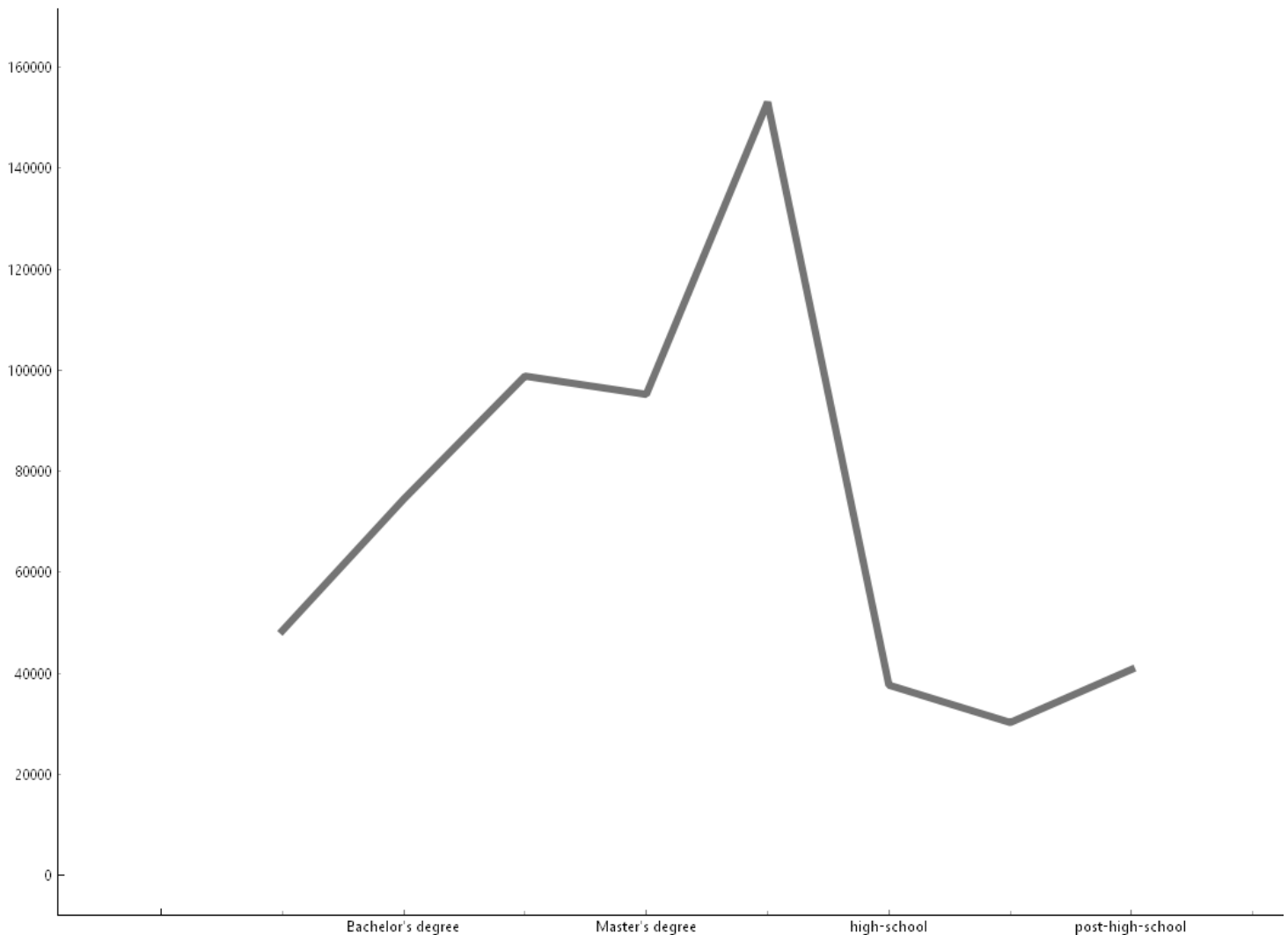


Figure 17 - mean income against education level

	name	coef
1	intercept	-50529.7
2	education	5735.19

Figure 18 - results of liner regression for income against education

Linear regression was used to estimate ow much one's income increases per year of education (\$5735.19)

4. Selection Bias

- Higher education may be **more accessible to wealthier individuals**, who might already have better financial opportunities.
- The dataset may not include **self-employed individuals or those in informal work**, whose income is harder to measure.

5. Data Quality Issues

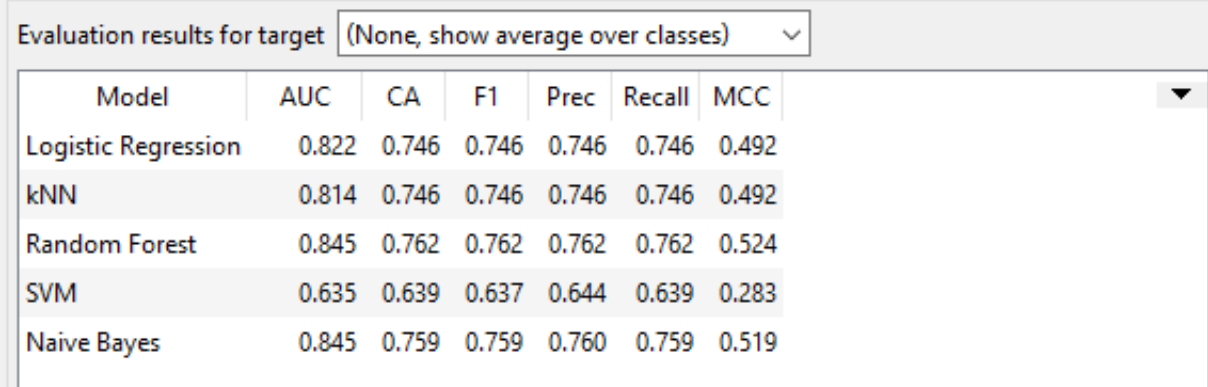
- **Sampling Bias**: Since the dataset was **sampled down to 5,000 rows**, it may not fully represent the overall population.
- **Errors in Reporting**: If income data is self-reported, people might **overestimate or underestimate** their earnings.

6. Ignoring Outliers

- Extreme incomes (e.g., CEOs, entrepreneurs) could **skew** the results, making education appear more or less valuable than it actually is.
- A **log-transformed income variable** might help reduce skewness.

Potential Issues with the Analysis of Education's Impact on Income:

The analysis only shows correlation, not causation. Other factors (e.g., industry, experience, location) may influence income, not just education. Higher education might be correlated with other income-boosting factors (e.g., networking opportunities). Also, there's a potential for election Bias. Higher education may be more accessible to wealthier individuals, who might already have better financial opportunities. The dataset may not include self-employed individuals or those in informal work, whose income is harder to measure.



Evaluation results for target (None, show average over classes) ▾						
Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression	0.822	0.746	0.746	0.746	0.746	0.492
kNN	0.814	0.746	0.746	0.746	0.746	0.492
Random Forest	0.845	0.762	0.762	0.762	0.762	0.524
SVM	0.635	0.639	0.637	0.644	0.639	0.283
Naive Bayes	0.845	0.759	0.759	0.760	0.759	0.519

Figure 19 - classifier models test and score results

Based on the test and score results in Orange, the best model would be Random Forest or Naïve Bayes, as both achieve the highest AUC (0.845), CA (0.762), F1-score (0.762/0.759), Precision (0.762/0.760), Recall (0.762/0.759), and MCC (0.524/0.519). However, Random Forest slightly outperforms Naïve Bayes in F1-score and MCC, making it the best overall choice.

Ranks					
		#	Info. gain	Gain ratio	Gini
1	N	hours	0.149	0.084	0.096
2	N	education	0.104	0.054	0.069
3	N	occupation	0.099	0.049	0.065
4	N	age	0.073	0.037	0.049

Figure 20 - feature ranking results

Hours ranks highest with the strongest information gain (0.149), gain ratio (0.084), and Gini index (0.096), indicating it's the most predictive feature for your target variable, likely due to its significant impact on reducing uncertainty and splitting data effectively.

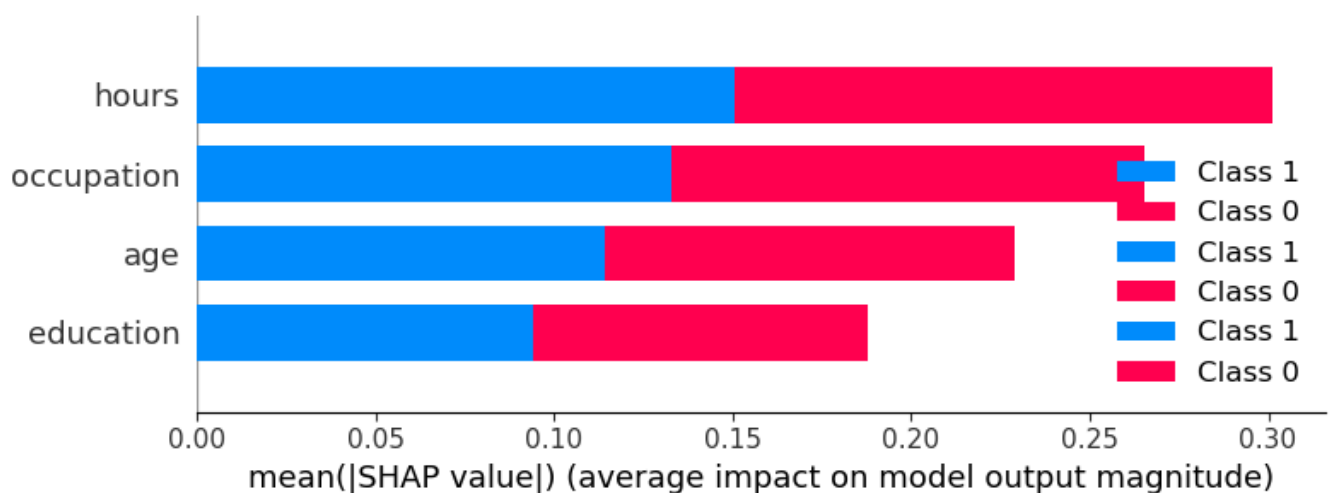


Figure 21 - SHAP technique results

```
Running script:
Permutation-Based Feature Importance:
      Feature  Importance      Std
2  occupation  0.265314  0.006290
|3      age    0.246343  0.006109
1      hours   0.243800  0.006077
0  education  0.194857  0.002631
...

```

Figure 22 – permutation-based feature importance results

The high SHAP value for hours in both classes indicate a strong positive impact on the model's prediction, as also shown in the feature ranking (Figure 20).

Part 4 - Demographics of US elections

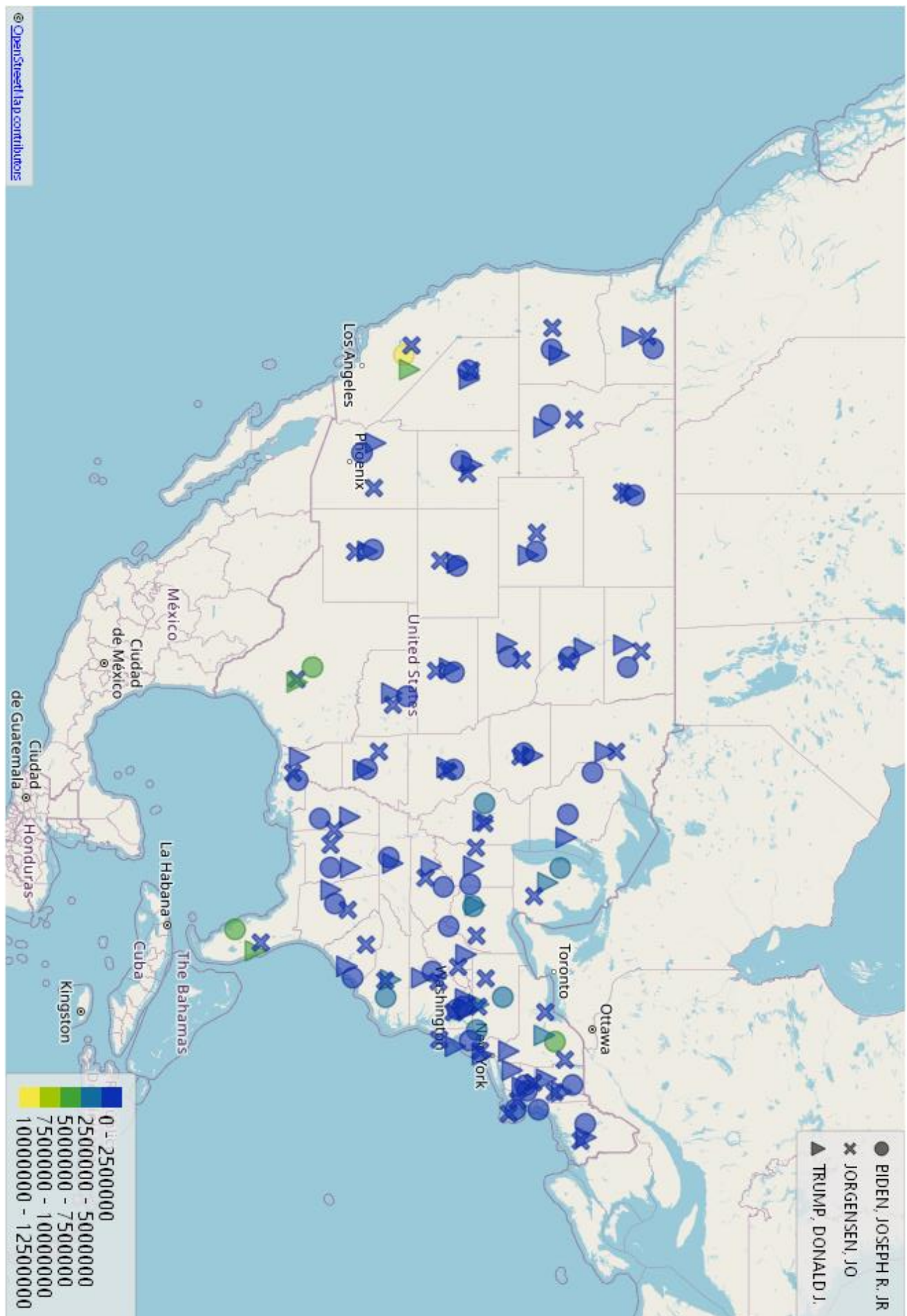


Figure 23 – election result for Biden, Trump and Jorgensen on the US map

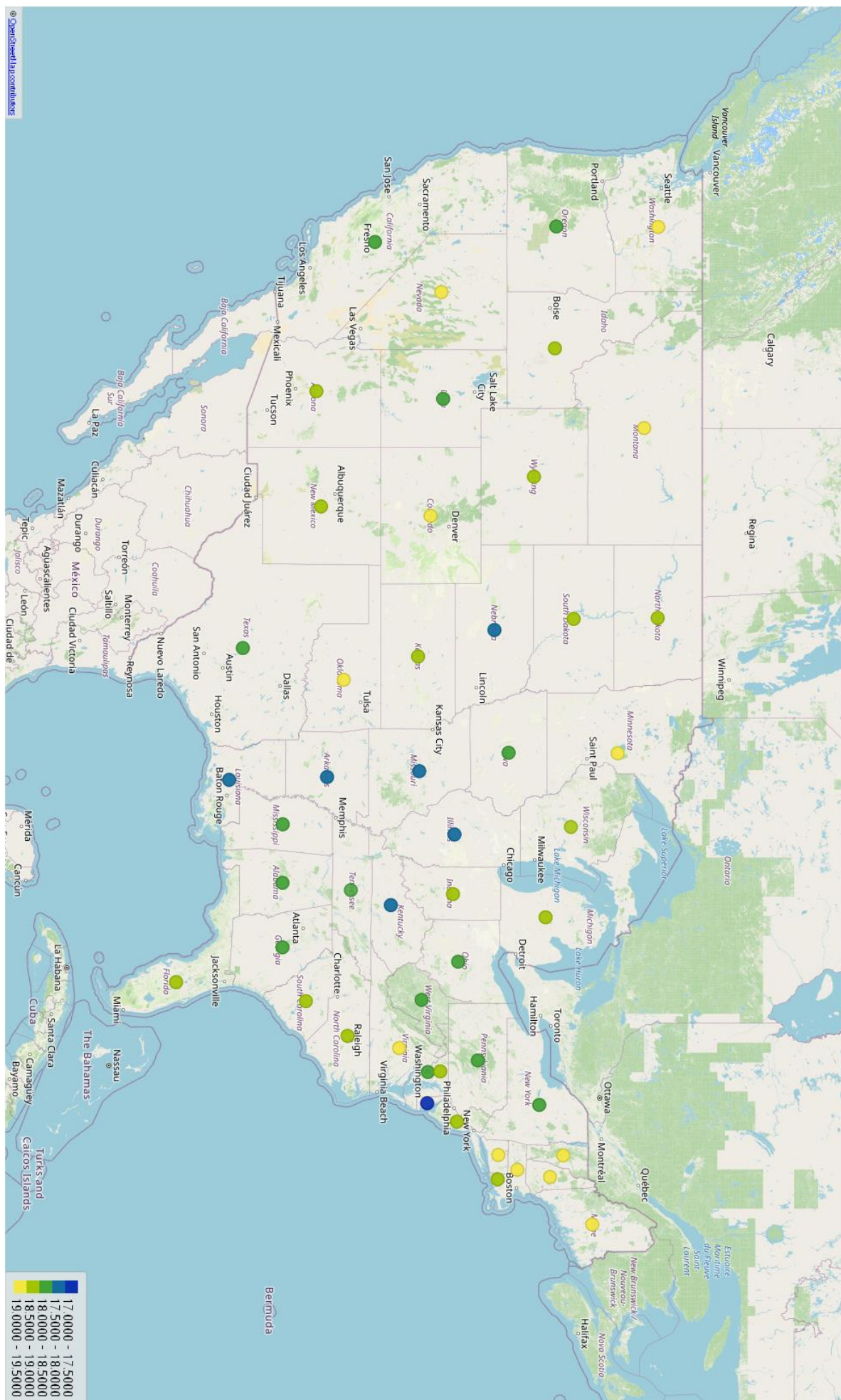


Figure 24 - Mean education attainment level in years of each US state on the US Map

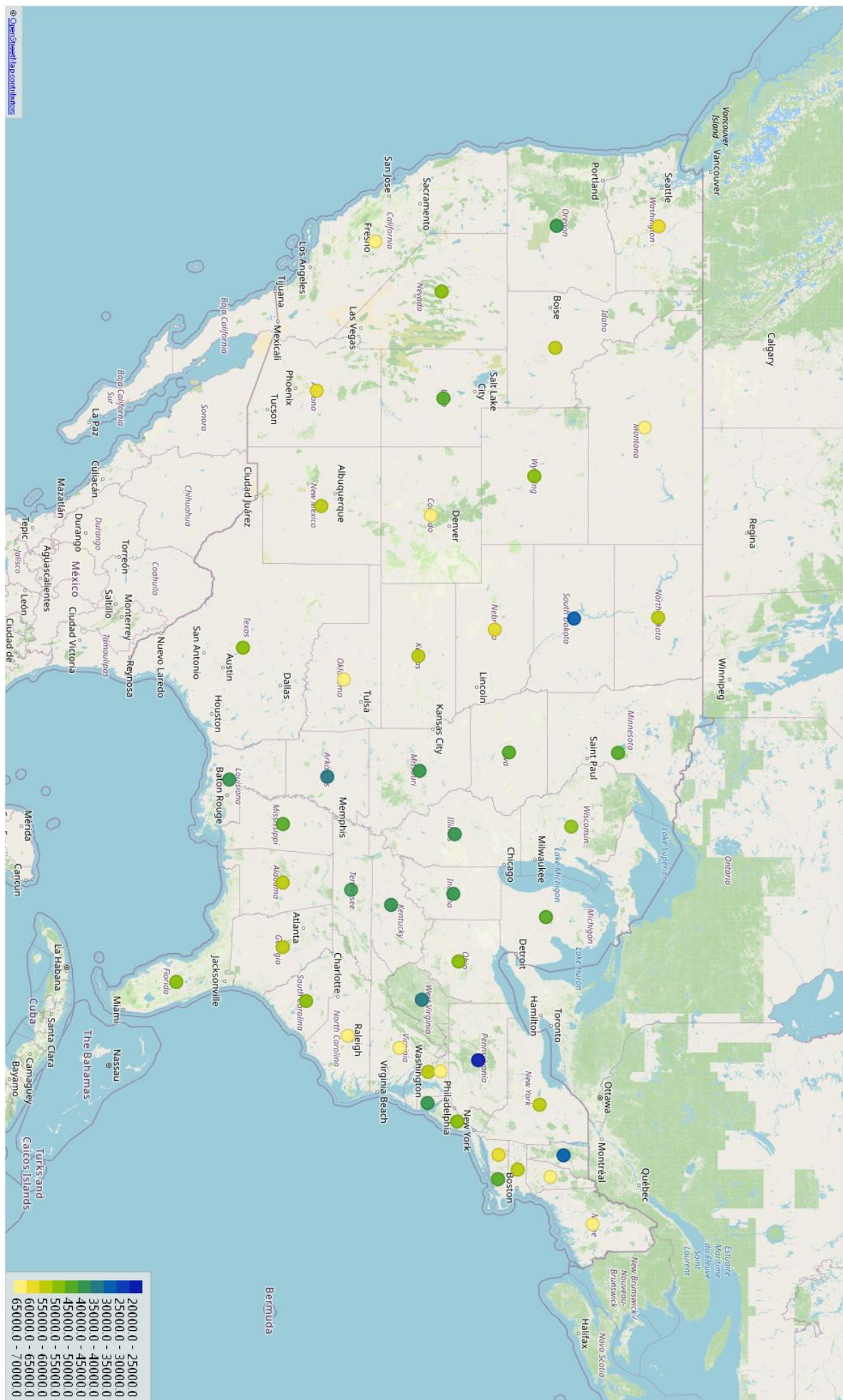


Figure 25 - mean income in dollars of each US state on the US Map

The comparison of the three maps—election results, mean income, and educational attainment—reveals strong correlations between socio-economic factors and voting patterns in the 2020 U.S. Presidential election. Biden’s support was concentrated in urban, high-income, and highly educated regions, while Trump performed well in rural, lower-income, and lower-education areas. Wealthier coastal cities and regions with a higher percentage of college graduates overwhelmingly voted for Biden, whereas lower-income and less-educated areas, particularly in the South and Midwest, favoured Trump. This highlights a clear urban-rural divide, where economic and educational disparities play a key role in shaping political preferences in the U.S.

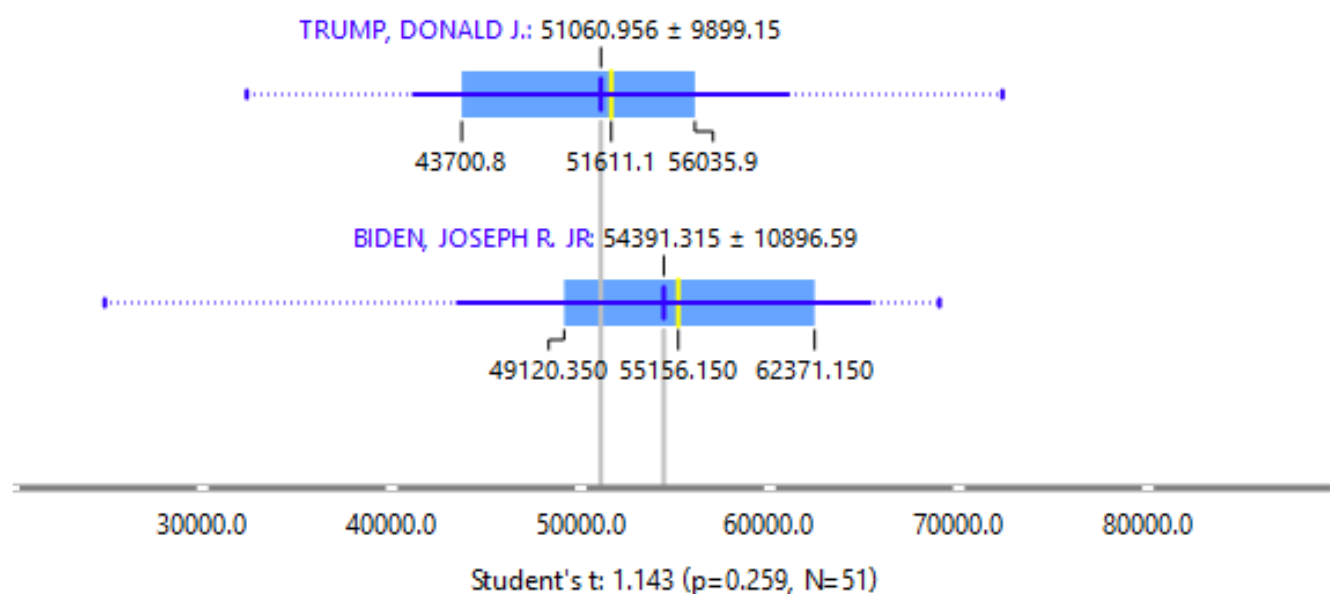


Figure 26 – boxplot of mean income of Trump winning and Biden winning states

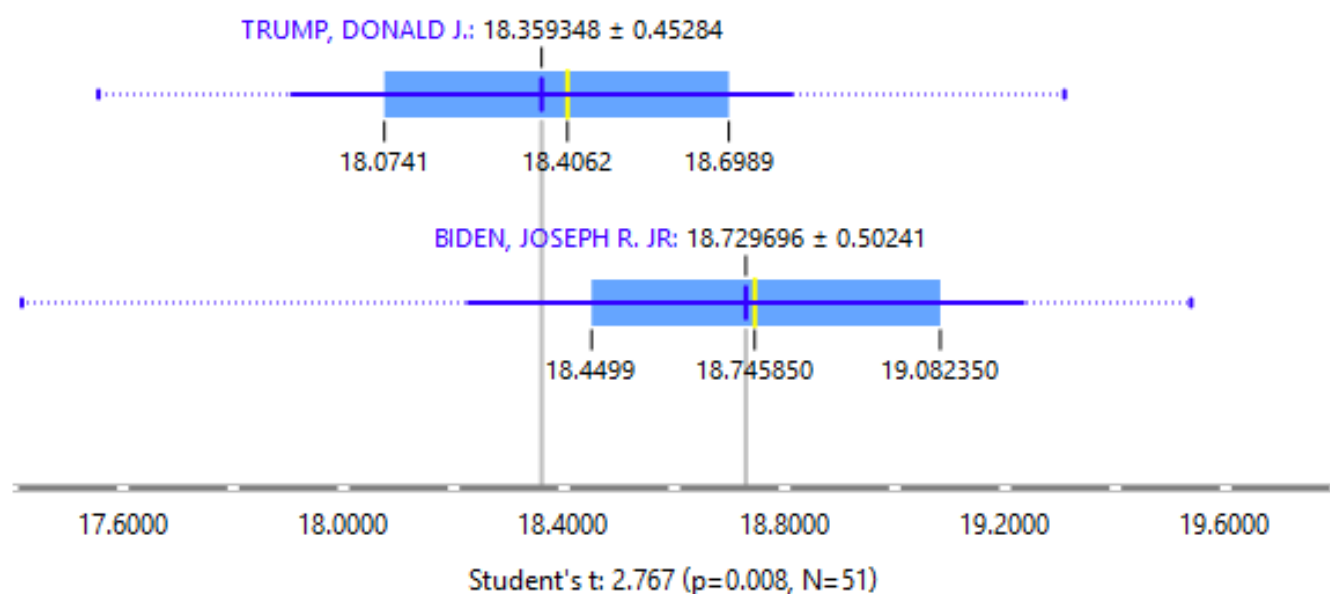


Figure 27 - boxplot of mean education entry level of Trump winning and Biden winning states

The boxplot in figure 26 compares the mean state-level vote counts for Trump and Biden in the 2020 election. Biden's average vote count ($54,391.31 \pm 10,896.59$) is slightly higher than Trump's ($51,060.96 \pm 9,899.15$), but the overlapping distributions suggest no significant difference. The student's t-test ($t = 1.143$, $p = 0.259$) confirms that the difference is not statistically significant, indicating that the vote counts per state for both candidates were relatively similar in distribution. The boxplot in figure 27 indicates that Biden's mean value (18.73 ± 0.50) is slightly higher than Trump's (18.36 ± 0.45). Unlike the first hypothesis, the t-test ($t = 2.767$, $p = 0.008$) shows a statistically significant difference, suggesting Biden's values are consistently higher across states. The minimal overlap in distributions further supports this distinction.

Part 5 – My own data mining

My Hypothesis: "The impact of weekly working hours on income is moderated by the type of occupation."

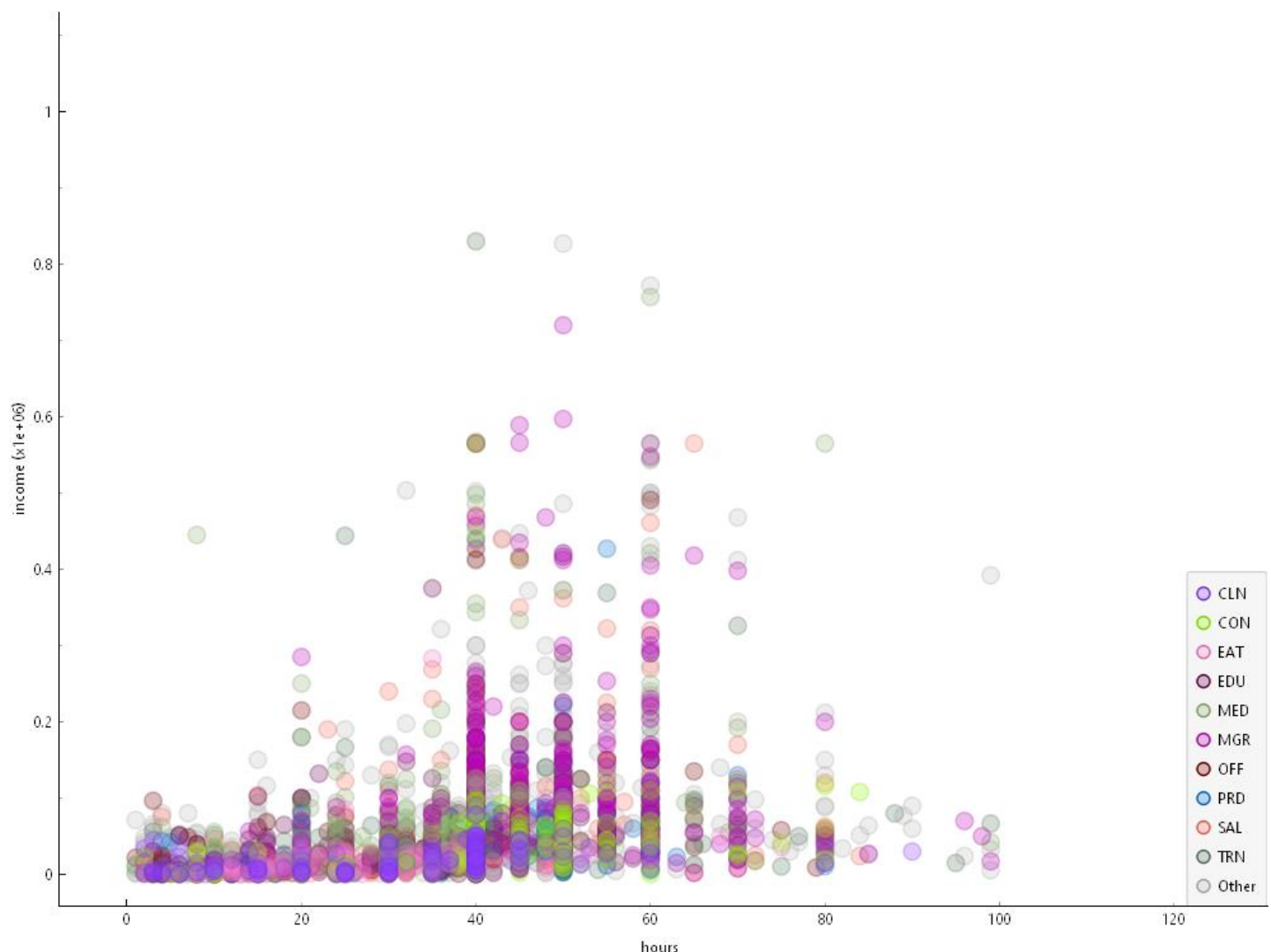


Figure 28 – scatter plot of hours against income with colour gradient for each occupation.

Through this visualisation, no pattern/trend is explicitly shown and deeper statistical test is needed. Statistical test used: Linear regression with dependent feature: income, independent feature: Hours worked, occupation (dummy variables used), hours*occupation (interaction term).

Model Performance

R^2 (coefficient of determination) = 0.196. This means that 19.6% of the variation in income is explained by hours and occupation. The low R^2 suggests that other factors outside this model also influence income.

Adjusted R^2 = 0.191. Since it's close to R^2 , it confirms that adding variables (occupations) still provides meaningful improvement.

The p-value for hours (hours = 1324.55, $p < 0.0001$) which means it's strongly significant. An increase in work hours is associated with a higher income. Many occupation_ variables have p-values < 0.05 , meaning they significantly impact income. The smallest eigenvalue = $4.19e-27$ and condition number = $4.44e+16$. This suggests strong multicollinearity, meaning some occupations might be highly correlated. This could make coefficient estimates unreliable.

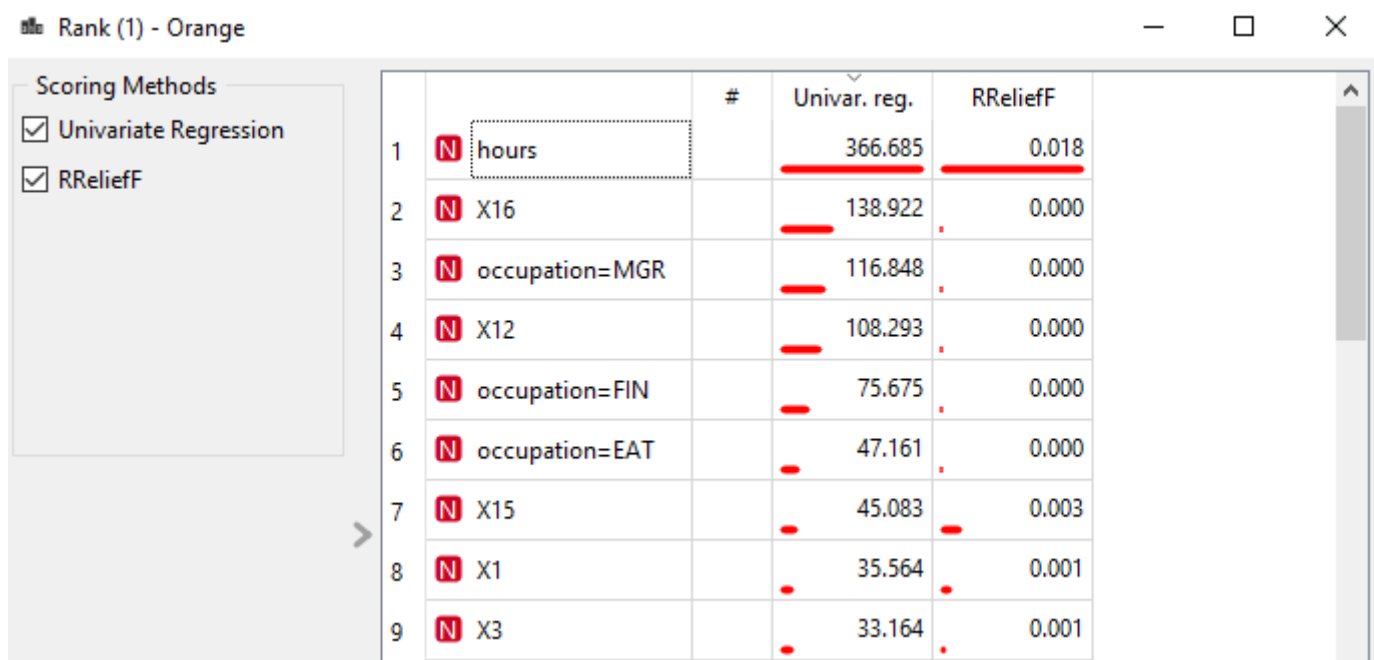


Figure 29 – Rank of Linear regression test

Given that both occupation and interaction data have been shown to influence the outcome, we can somewhat support the hypothesis that occupation plays a role in determining income per hour worked.

Key Components of ANOVA Output

Source	sum_sq (Sum of Squares)	df (Degrees of Freedom)	F (F-statistic)	PR(>F) (p-value)
C(occupation)	2.47×10^{12}	24.0	25.15	2.72×10^{-105}
hours	1.39×10^{12}	1.0	339.52	2.14×10^{-73}
Residual (Error)	2.04×10^{13}	4974.0	NaN	NaN

Figure 30 - ANOVA test output

Occupation

The sum of squares (SS) value is 2.47×10^{12} , meaning occupation explains a significant portion of the variation in income. The F-statistic (25.15) is quite large, suggesting that occupation

significantly influences income. The p-value (2.72×10^{-105}) is extremely small (almost zero), meaning the effect of occupation on income is statistically significant.

Hours

The sum of squares is 1.39×10^{12} is also large, indicating hours worked explains income variation.

The F-statistic (339.52) is very high, showing a strong effect of hours worked on income.

The p-value (2.14×10^{-73}) is almost zero, confirming that hours worked significantly impacts income.

Residual (Error Term)

This represents unexplained variation in income. The sum of squares (2.04×10^{13}) is much higher than the other values, meaning other factors (not included in the model) still explain a significant portion of income variation. No F-statistic or p-value is shown for residuals since it's the baseline for comparison.