Protective service occupations

Protective Service Occupations (Occupation codes 3700-3955) from the CPS-Earnings dataset were selected and relevant positions were merged. After filtering the data, I focused on the columns necessary for building predictive models and calculated the earnings per hour. Missing values were removed, and outliers were excluded using boxplots. This resulted in a dataset of 3248 observations (Descriptive statistic Table 2).

Four linear regression models were developed, each with increasing complexity, using earnings per hour as the dependent variable. The following models were tested:

$$\textbf{Model 1:} \quad \text{earnings per hour} = \beta_0 + \beta_1 \text{age} + \beta_2 \text{sex} + \beta_3 \text{race} + \epsilon$$

$$\textbf{Model 2:} \quad \text{earnings per hour} = \beta_0 + \beta_1 \text{age} + \beta_2 \text{sex} + \beta_3 \text{race} + \beta_4 \text{grade92} + \beta_5 \text{marital} + \epsilon$$

$$\textbf{Model 3:} \quad \text{earnings per hour} = \beta_0 + \beta_1 \text{age} + \beta_2 \text{sex} + \beta_3 \text{race} + \beta_4 \text{grade92} + \beta_5 \text{marital} + \\ + \beta_6 \text{state} + \beta_7 \text{ind02} + \epsilon$$

$$\textbf{Model 4:} \quad \text{earnings per hour} = \beta_0 + \beta_1 \text{age} + \beta_2 \text{sex} + \beta_3 \text{race} + \beta_4 \text{grade92} + \beta_5 \text{marital} + \\ + \beta_6 \text{state} + \beta_7 \text{ind02} + \beta_8 \text{occ2012} + \beta_9 \text{class} + \beta_{10} \text{unionmme} + \epsilon$$

The selection of predictors was based on both theoretical and empirical considerations. Age, sex, and race are fundamental demographic factors influencing earnings. Education (grade92) and marital status were added due to their known impact on labor market outcomes. State and industry (ind02) capture regional and sectoral variations, while occupation-specific variables (occ2012, class, and unionmme) in Model 4 account for job-type differences, employment class, and union membership.

Model	RMSE	CV RMSE	BIC
Model 1	9.796	9.808	24073.369
Model 2	9.307	9.326	23757.101
Model 3	8.633	9.158	24667.305
Model 4	8.480	9.015	24599.547

Table 1: Model Performance Metrics

The performance of the four models was evaluated using RMSE, cross-validated RMSE, and BIC. Model 1, the simplest model, had the highest RMSE, indicating weak predictive power. Model 2 improved upon this by incorporating education and marital status, which lowered both RMSE and BIC, meaning a better fit without unnecessary complexity. Model 3 added state and industry variables, further reducing RMSE, but the increase in CV RMSE and a significant rise in BIC indicated potential overfitting.

Model 4, the most complex model, had the lowest RMSE, meaning the best predictive accuracy. However, its high BIC implies that the additional predictors might not contribute enough value to justify the increased complexity. Considering both accuracy and model parsimony, Model 2 is the most balanced choice, avoiding excessive complexity. If the priority is maximum predictive accuracy, Model 4 could be preferred despite its higher BIC.

As shown in Table 1, increasing model complexity generally improves RMSE, suggesting a better fit to the training data. However, the BIC does not consistently decrease, indicating that adding more variables may introduce unnecessary complexity without a proportional increase in explanatory power.

Residual plots (Figure 8) show that increasing complexity reduces residual variance, improving predictions. However, Model 4 still exhibits some heteroscedasticity, suggesting possible non-linearity or missing factors. While Model 4 offers the best predictive accuracy, Model 2 strikes the best balance between simplicity and performance.

Appendix

	age	grade92	sex	race	marital	earn_per_hour
Count	3248.00	3248.00	3248.00	3248.00	3248.00	3248.00
Mean	40.21	40.74	1.23	1.51	3.38	20.37
Std	12.19	1.85	0.42	1.61	2.72	10.23
Min	16.00	34.00	1.00	1.00	1.00	0.00
25%	30.00	39.00	1.00	1.00	1.00	12.00
50%	41.00	40.00	1.00	1.00	1.00	18.25
75%	50.00	43.00	1.00	1.00	7.00	26.92
Max	64.00	46.00	2.00	21.00	7.00	50.00

Table 2: Descriptive Statistics of the Selected Variables

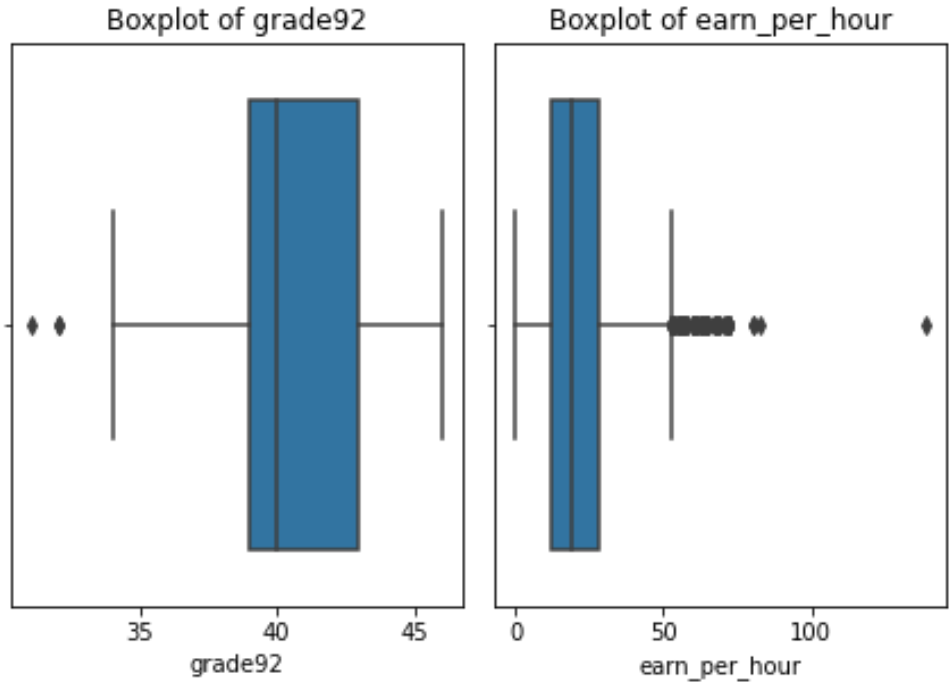

Figure 1: Box plot

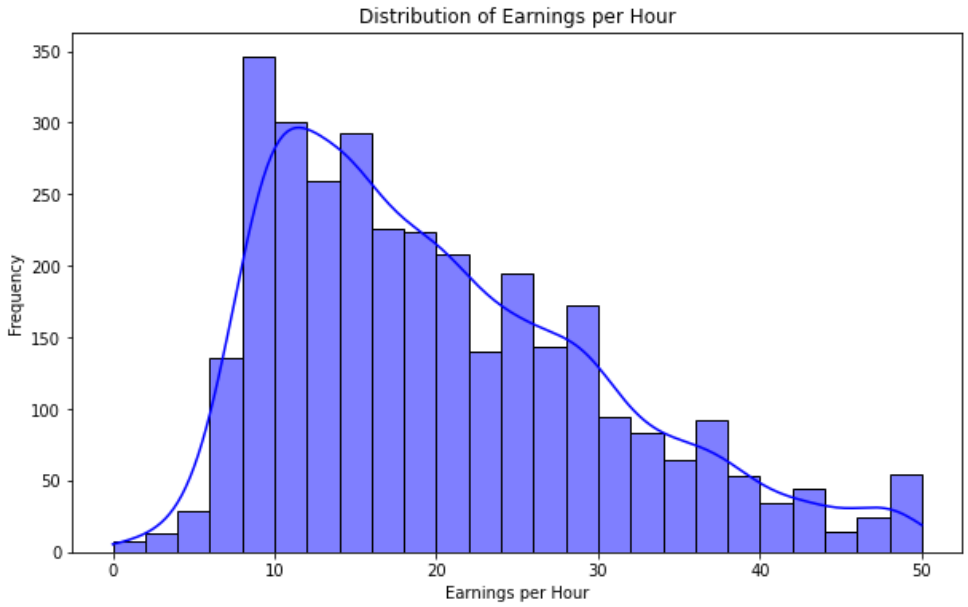

Figure 2: Distribution for Earnings per Hour

Figure 3: Average Earnings per Hour by Age

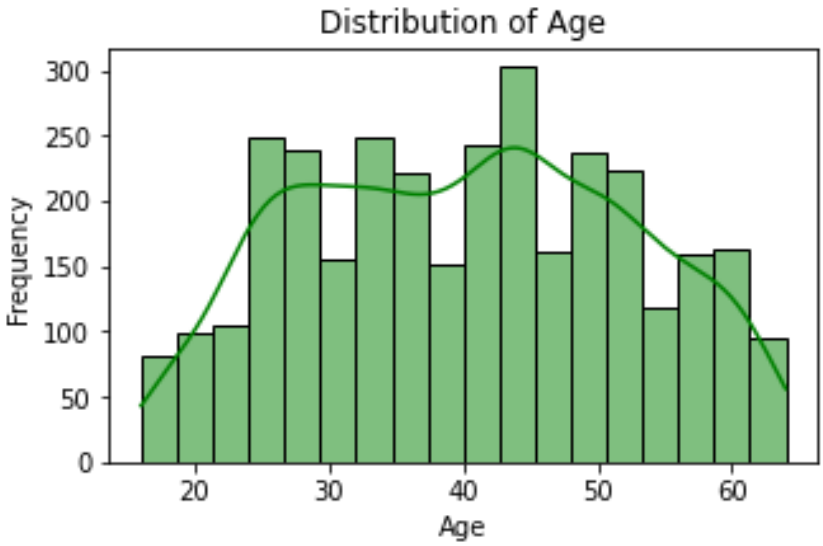

Figure 4: Distribution plots

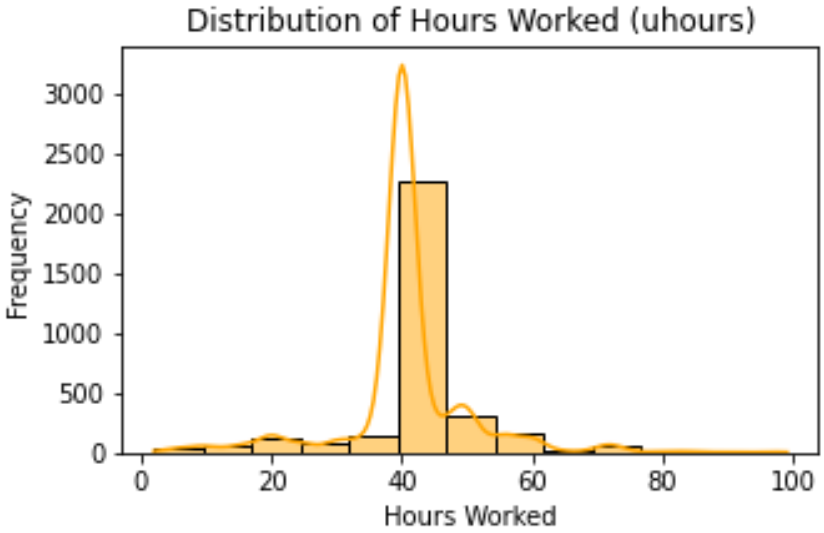

Figure 5: Distribution plots

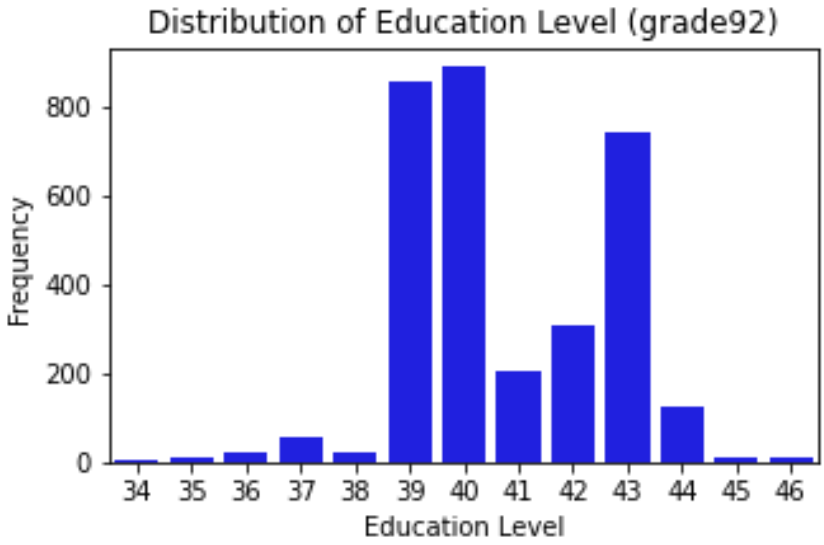

Figure 6: Distribution plots

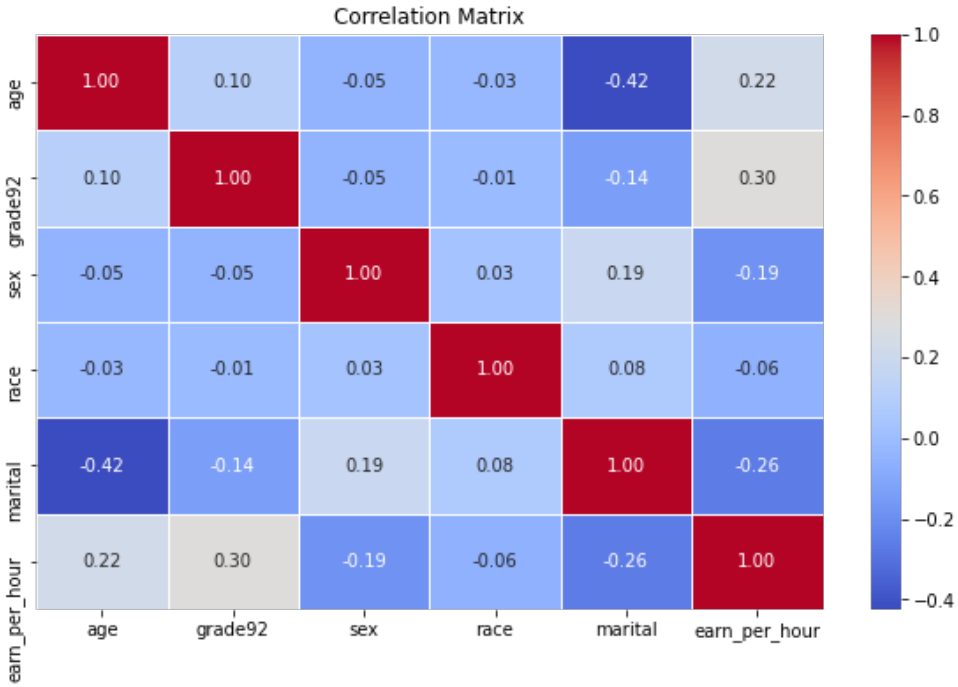

Figure 7: Correlation Matrix

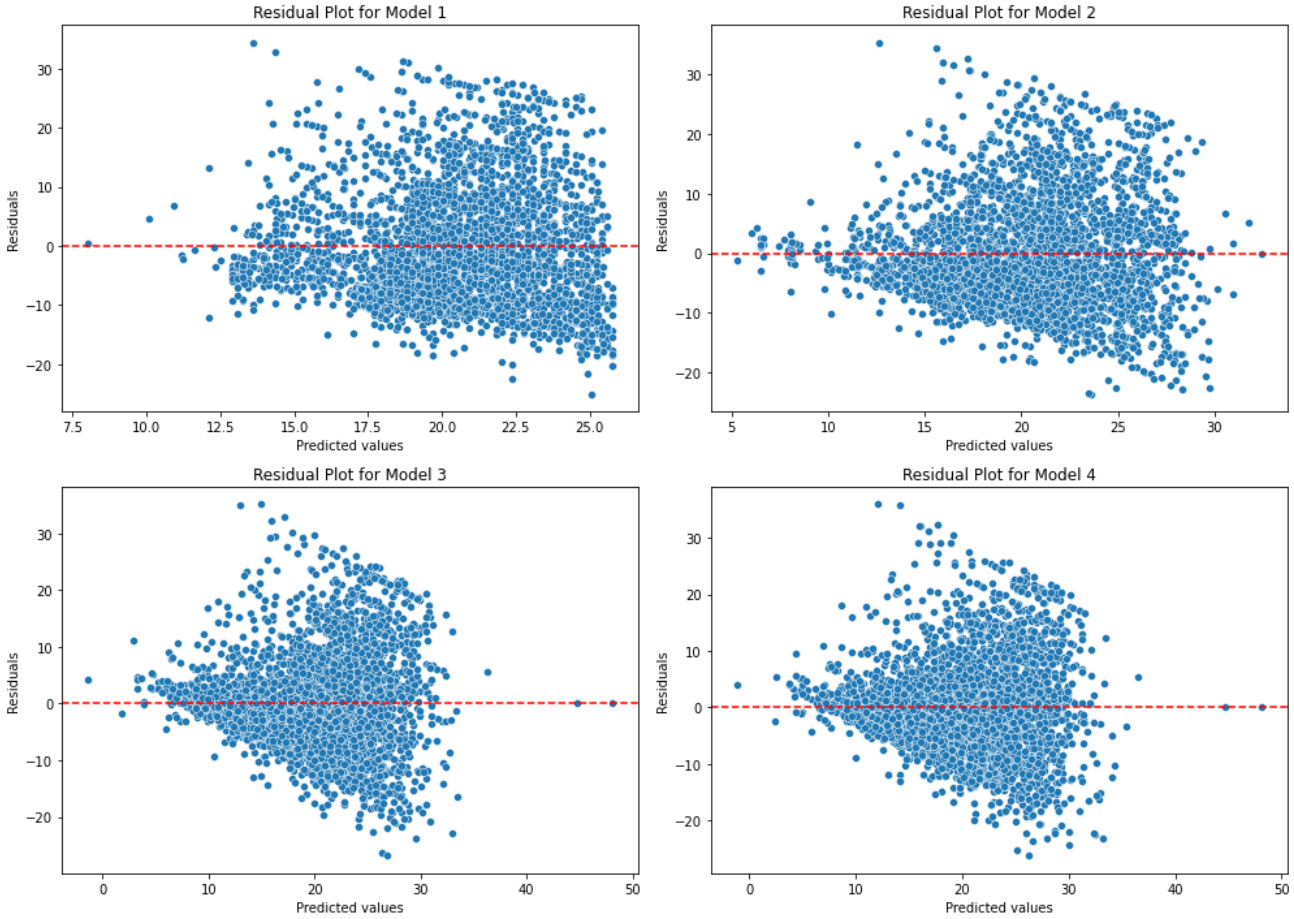

Figure 8: Residual plots for models 1-4

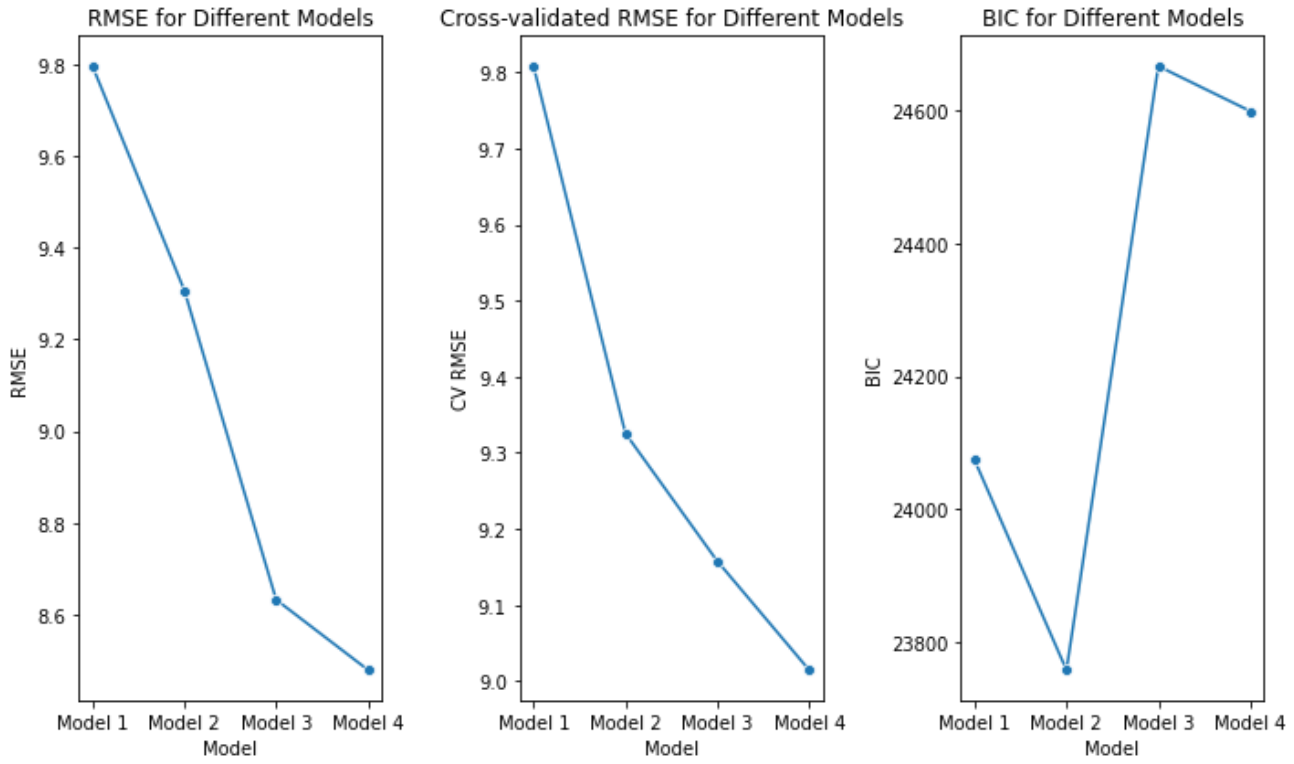

Figure 9: Model comparison