First_assignment

Akos Almasi

2023-01-25

I decided to investigate the surgeon and physician occupation (occupation number: 3060). I chose it because it's a profession that requires a high level of qualification, so my assumption is that it pays well and as the years go by, as the experience increases, so does the salary. In order to work as a surgeon or physician, one needs a university degree, so I limited my analysis to data where the individual had at least a bachelor's degree.

```
# Filter the data to surgeons and physicians
df <- df %>% filter(occ2012 == '3060')
```

I eliminated extreme values that would have significantly distorted my results. In order to see what are the extreme values, I used the data summary function. Based on the result filtered out the unrealistically low and high values of earnhourly column.

	P0	P01	P25	P50	P75	P99	P100	Mean	SD
earnhourly	0.00	8.06	21.26	39.11	57.69	96.15	432.50	41.12	26.63

Figure 1 Shows the distribution of the hourly earnings of surgeons and physicians.

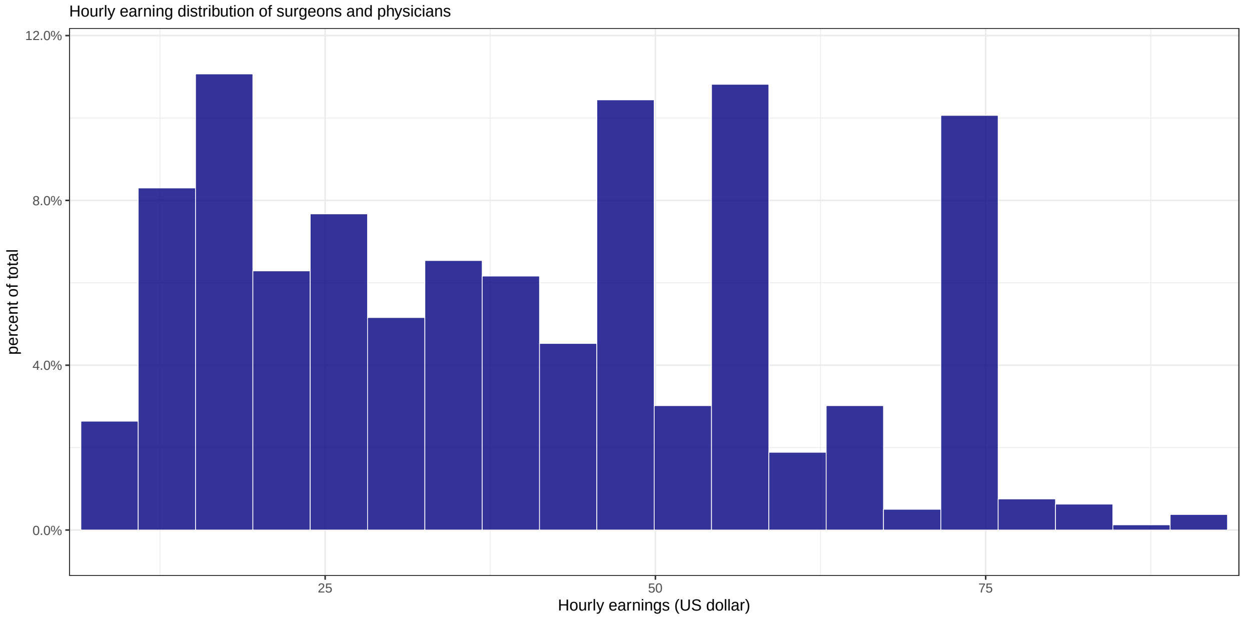
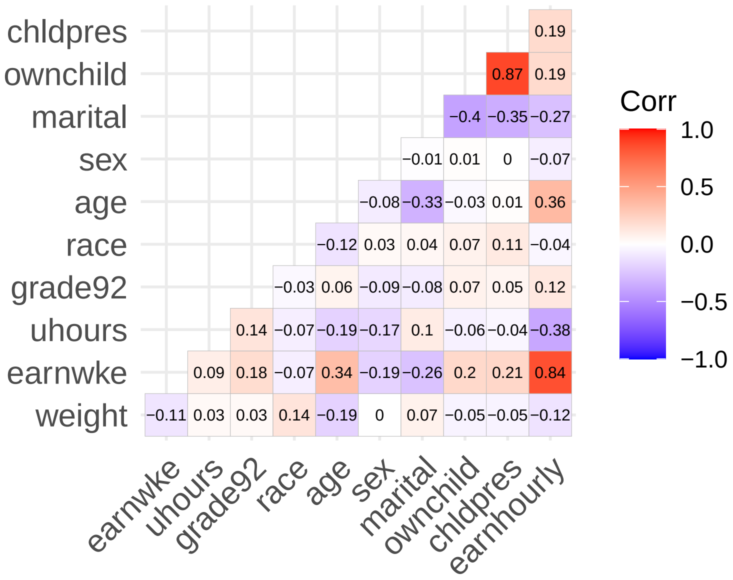
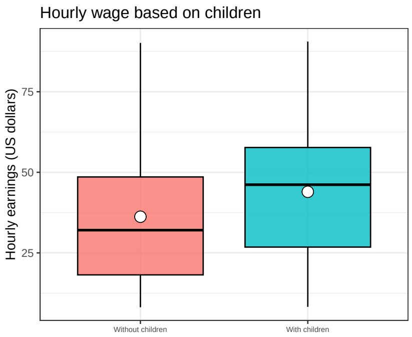
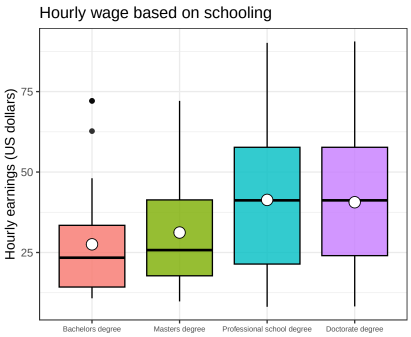


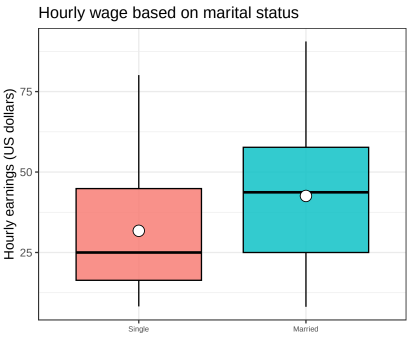
Figure 2 shows a correlation matrix of the numeric variables. Based on the correlation matrix I chose the most important variables that could explain the earnhourly column, and included them in the models. The age, whether they have children or not and schoolings are the ones which have the highest positive correlation. The marital status has the highest negative correlation according to the graph.



The graph shows that the range of hourly wages differs between the childbearing and childless groups, with childbearing surgeons and doctors could have higher hourly wages than childless ones If you have a higher education, you might have a higher hourly wage, only a doctorate degree does not increase your hourly wage compared to the Professional school degree. For marital status,we can see married individuals can have higher hourly wages.


Hourly wage based on children


Hourly wage based on schooling


Hourly wage based on marital status

Evaluation of the models

We can see that even though **Model4** has the highest R-squared, it has the lowest Training RMSE value, and the highest BIC value with 13 predictors. The **Model2** has the lowest BIC value. BIC estimates the likelihood of a model to predict, if we would decide which model to pick solely on this we would choose **Model2**

Table 2: Model Comparision

Model	N predictors	R-squared	Training RMSE	BIC
(1)	2	0.1894898	17.98901	6877.546
(2)	3	0.1963539	17.91268	6877.463
(3)	7	0.2167286	17.68415	6883.761
(4)	13	0.2245693	17.59542	6915.833

Cross validation

Cross validation is a procedure used to avoid overfitting and estimate the skill of the model on unseen data. I would choose **Model3**, since the cross validated method is the most robust evaluation criterion, and it has the lowest RMSE value on average.

Table 3: Cross validation results

Resample	Model1	Model2	Model3	Model4
Fold1	18.26986	18.14020	17.88318	17.88666
Fold2	18.01383	17.93787	17.80219	17.98006
Fold3	17.64035	17.51443	17.55561	17.58440
Fold4	18.24344	18.31719	18.16770	18.12207
Average	18.04363	17.97991	17.85351	17.89438