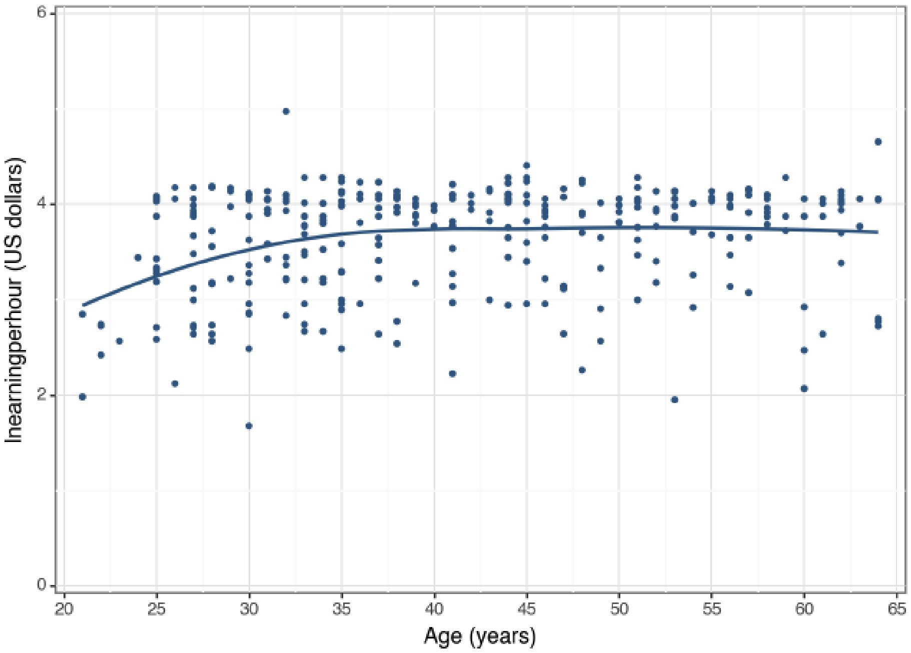
DA3 ASSIGNMENT I

In this article, it is aimed to evaluate the performance of 4 models established with our data set using 1) RMSE in the full sample, (2) cross-validated RMSE and (c) BIC in the full sample. Models will go from simple to more complex, and the complexity will be determined by the number of variables in the model. In the models, the target variable will be earnings per hour, while other variables will be used as predictors. Afterwards, the article will be completed by analyzing the complexity and performance of the models. For this, the data set at the link https://osf.io/g8p9j/ was used and occupation Pharmacist (code: 3050) was selected.

After the data set is uploaded to the Jupyer notebook, the missing values are first filled with NaN object values. First of all, our target variable is earnings per hour; It is obtained by dividing earnwke and uhours values. Since using relative differences produces more inferential results, this value is obtained as the log value lnearnperhour and the female dummy variable to look at the gender distribution of earnings.

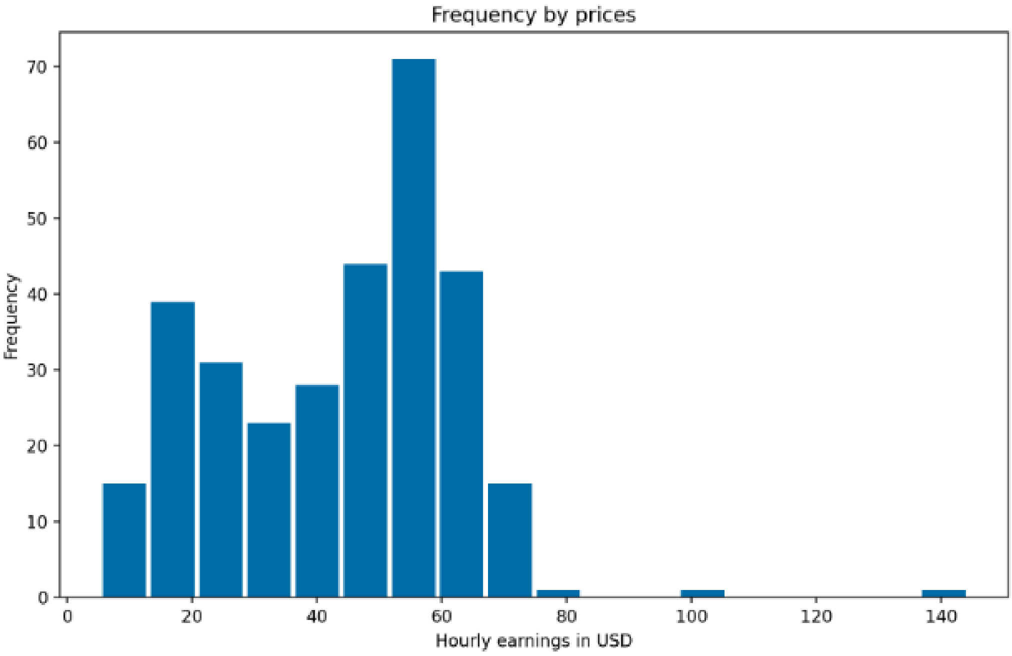
As a result of the initial analysis, it was seen that the average hourly wage of pharmacists was 43,598 USD, and 123 of these people were men and 189 were women.

From the earning distribution plots, it was determined that the distribution was nonlinear, and in order to understand this nonlinearity, the following graph was obtained using the Loess method.



According to this chart; Logearnings are slightly higher for older ages among young adult pharmacists: 0.3 log point higher for 35-year-old analysts than for 20-year-old analysts. Also age differences are positive but get smaller and smaller as we compare older people; little or any difference by age [35,60]; Those older than 60 earn less than average.

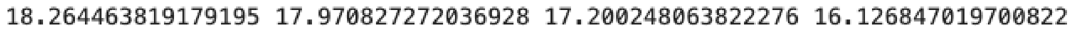
The frequencies of hourly wages can be easily seen with the "Frequency by prices" chart below. According to the chart, it can be concluded that the highest frequency of hourly earnings varies between 50-60 USD and 100 and 140 USD can be considered outliers.



Many dummy variables were created by determining the predictors to be used in the models. These are education level, marriage status, child ownership, etc. It covers topics such as. The aim is to select the most optimum model by establishing relationships between models.

1) RMSE in the Full Sample

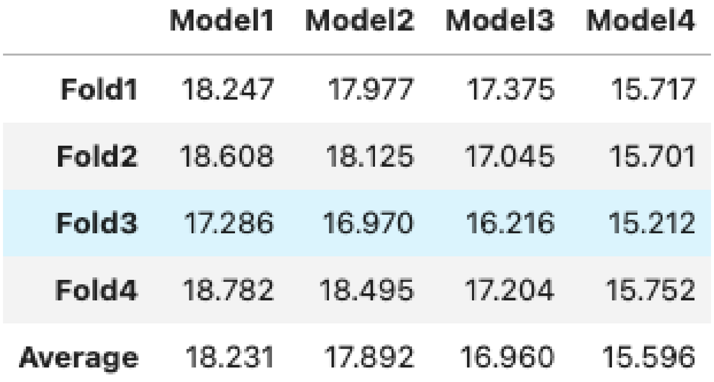
In the first model we obtained using 4 regressions, the RMSE in the full sample, the lowest RMSE was the 4th regression with **16.12**. Since the lowest RMSE is more selectable, it can be chosen as the 4th regression model in this model. However, it can be said that the 3rd model is more suitable because there is not much difference between it and the 3rd model and the 3rd model is less complex.



```
18.264463819179195 17.970827272036928 17.200248063822276 16.126847019700822
```

2) Cross Validated RMSE

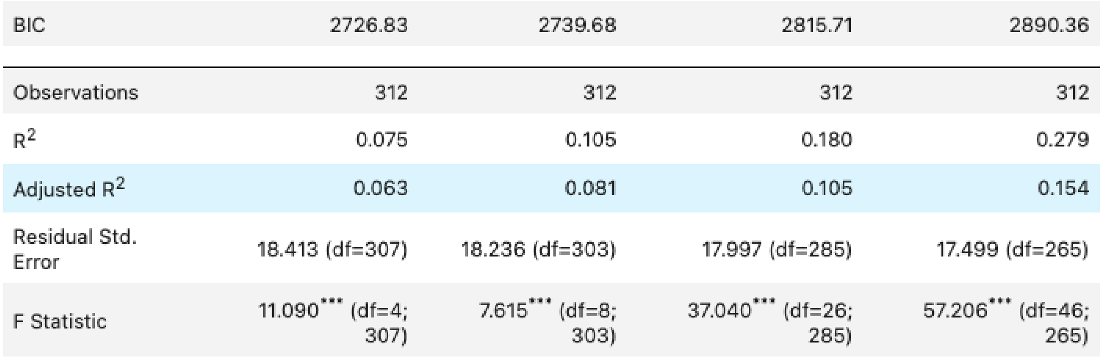
With the Cross Validated RMSE obtained by creating the same regressions and 4 folds, model 4 has a value of **15.596**, less than the others. However, since the 3rd model is the second lowest value with **16,960** and is less complex, the 3rd model can be said to be more suitable.



	Model1	Model2	Model3	Model4
Fold1	18.247	17.977	17.375	15.717
Fold2	18.608	18.125	17.045	15.701
Fold3	17.286	16.970	16.216	15.212
Fold4	18.782	18.495	17.204	15.752
Average	18.231	17.892	16.960	15.596

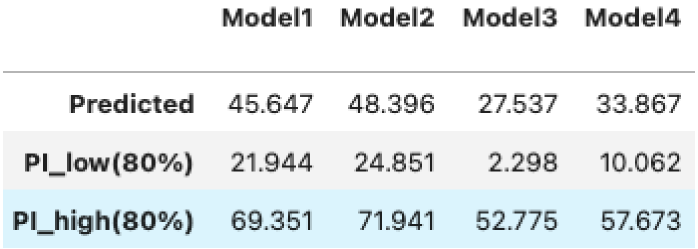
3) Creating BIC in the Full Sample

With this model, it seems that model 1 has the lowest value with 2726.83. However, when the first model is evaluated with an R2 value of 0.075, it may not be a correct interpretation to see it as the best model.



BIC	2726.83	2739.68	2815.71	2890.36
Observations	312	312	312	312
R^2	0.075	0.105	0.180	0.279
Adjusted R^2	0.063	0.081	0.105	0.154
Residual Std. Error	18.413 (df=307)	18.236 (df=303)	17.997 (df=285)	17.499 (df=265)
F Statistic	11.090*** (df=4; 307)	7.615*** (df=8; 303)	37.040*** (df=26; 285)	57.206*** (df=46; 265)

Additionally, when we look at the 80% PI range chart, we see that each model has high values. This can be explained by the fact that we have too many variables and too little data.



	Model1	Model2	Model3	Model4
Predicted	45.647	48.396	27.537	33.867
PI_low(80%)	21.944	24.851	2.298	10.062
PI_high(80%)	69.351	71.941	52.775	57.673

Summary

When all models are evaluated in terms of low RMSE, low BIC, high R2 and less complexity (fewer variables), choosing the 3rd model will be the right decision.

Abdullah DUMAN