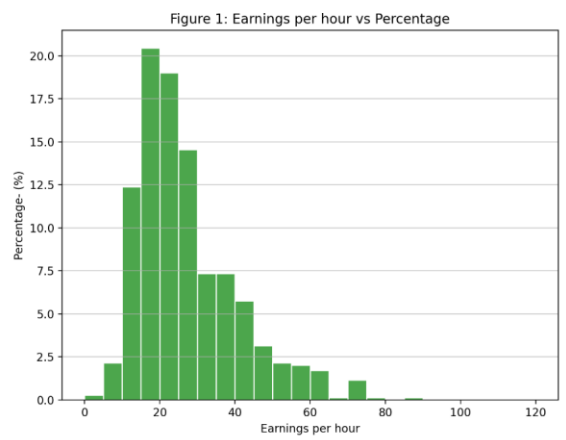
DA3 ASSIGNMENT

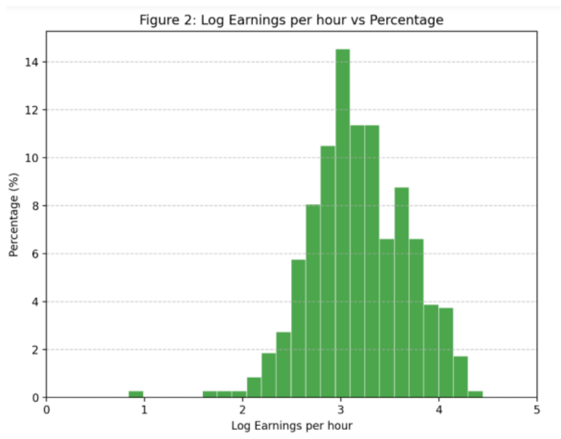
Introduction:

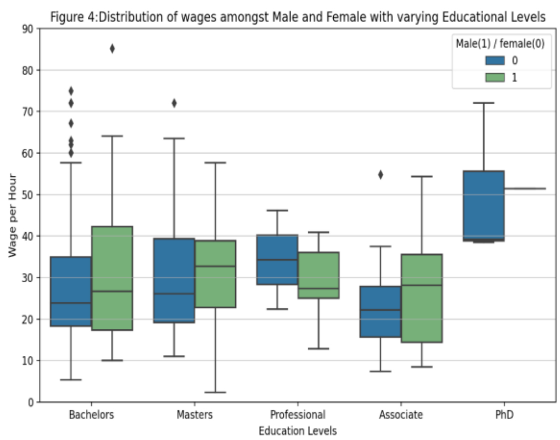
The aim of this assignment is to uncover building models for regression analysis and how we can achieve the desired outcome using other variables in the data. And also, to understand the factors we consider while creating a good model. In addition, we also see how a model could move towards overestimation if too many variables are used, and it might be underestimated if too few variables are used. In this assignment, we are analyzing cps-earnings dataset and will be building 4 models from the simplest one to the more complex ones. It will give us a better understanding of how to choose the right model.

Data Cleaning:

The occupation code we chose was ‘0630’ and we filtered the data for people above the age of 20 years. The dummy variables were created since we had qualitative values as well, so true/false variables were created and used in the analysis. The male variable was also created, and other variables were used like marital status, whether the person has a child, their ethnicity, how educated they are etc. The graphs have been included below to better understand our data and statistics. We can see our wages and how they were skewed and what the distribution looks like after taking a log. We can also see the age vs wages loess graph, and lastly we see the boxplot with education categories.







Mohammad Ahmed

Model Comparison:

The simplest, model 1 will have education as main variable. We will set the base at 'No Diploma' and take all the categories above it. We will be noticing that how the wages differ between men and women and our results show that.

The second model will have the age variable added. This way we will keep increasing the variables in each model and understand the statistics it provides. Lastly, model 4 will have the most complexity with the most variables.

We will keep in mind that a lower BIC will be preferred in the statistics because it helps in preventing overfitting. To assess our model, we will also check RMSE and would prefer the one with the lowest RMSE value. In this we create 4 folds and perform the calculations.

	Model1	Model2	Model3	Model4
Fold1	12.919500	11.877732	11.784141	11.627399
Fold2	13.160009	12.103995	11.996320	11.799814
Fold3	13.187572	12.093038	12.051339	11.827412
Fold4	13.193208	11.915244	11.806503	11.687276
Average	13.115072	11.997502	11.909576	11.735475

With these statistics, we can see that with RMSE values, the Model 4 which has the highest number of variables performs better with an average of 11.73. Also keeping in mind that model 2 and model 3 aren't too far ahead in terms of performance with RMSE. Now we will run the regressions and see all stats and assess which model performance was the best.

```
Regression: reg1
                            OLS Regression Results
==============================================================================
Dep. Variable:                  wages   R-squared:                       0.062
Model:                            OLS   Adj. R-squared:                  0.056
Method:                 Least Squares   F-statistic:                     9.568
Date:                Fri, 19 Jan 2024   Prob (F-statistic):           7.71e-09
Time:                        19:03:28   Log-Likelihood:                -2771.7
No. Observations:                 694   AIC:                             5555.
Df Residuals:                     688   BIC:                             5583.
Df Model:                           5
Covariance Type:                  HC1
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]

------------------------------------------------------------------------------------------
Intercept                 23.3977      0.806     29.027      0.000      21.815      24.980
Associate[T.True]          0.5117      1.453      0.352      0.725      -2.340       3.364
Bachelors[T.True]          5.3039      1.143      4.640      0.000       3.060       7.548
Masters[T.True]            7.5261      1.744      4.316      0.000       4.103      10.950
Professional[T.True]       6.7195      4.141      1.623      0.105      -1.411      14.850
PhD[T.True]               26.9066      6.874      3.914      0.000      13.410      40.403
==============================================================================
Omnibus:                      126.362   Durbin-Watson:                   1.840
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              208.706
Skew:                           1.140   Prob(JB):                     4.79e-46
Kurtosis:                       4.420   Cond. No.                         15.0
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)

Regression: reg2
                            OLS Regression Results
==============================================================================
Dep. Variable:                  wages   R-squared:                       0.214
Model:                            OLS   Adj. R-squared:                  0.205
Method:                 Least Squares   F-statistic:                     25.34
Date:                Fri, 19 Jan 2024   Prob (F-statistic):           2.31e-34
Time:                        19:03:28   Log-Likelihood:                -2710.4
No. Observations:                 694   AIC:                             5439.
Df Residuals:                     685   BIC:                             5480.
Df Model:                           8
Covariance Type:                  HC1
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]

------------------------------------------------------------------------------------------
Intercept              -26.9112      5.515     -4.880      0.000     -37.739     -16.083
Associate[T.True]        0.4814      1.302      0.370      0.712      -2.076       3.038
Bachelors[T.True]        8.2056      1.138      7.212      0.000       5.972      10.440
Masters[T.True]          8.6881      1.695      5.127      0.000       5.361      12.016
Professional[T.True]     5.7174      4.170      1.371      0.171      -2.470      13.905
PhD[T.True]             24.9108      6.460      3.856      0.000      12.227      37.594
age                      2.0432      0.288      7.095      0.000       1.478       2.609
agesq                   -0.0197      0.004     -5.590      0.000      -0.027      -0.013
male                     3.8196      1.187      3.219      0.001       1.490       6.150
==============================================================================
Omnibus:                       94.690   Durbin-Watson:                   1.833
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              148.984
Skew:                           0.897   Prob(JB):                     4.45e-33
Kurtosis:                       4.390   Cond. No.                     2.85e+04
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

```
Regression: reg3
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                  wages   R-squared:                       0.224
Model:                            OLS   Adj. R-squared:                  0.212
Method:                 Least Squares   F-statistic:                     19.84
Date:                Fri, 19 Jan 2024   Prob (F-statistic):           6.71e-35
Time:                        19:03:28   Log-Likelihood:                -2705.8
No. Observations:                 694   AIC:                             5436.
Df Residuals:                     682   BIC:                             5490.
Df Model:                          11                                         
Covariance Type:                  HC1                                         
=========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]

-----------------------------------------------------------------------------------------
Intercept               -26.8292      6.186     -4.337      0.000     -38.974     -14.684
Associate[T.True]         0.4263      1.290      0.330      0.741      -2.107       2.959
Bachelors[T.True]         8.0135      1.141      7.023      0.000       5.773      10.254
Masters[T.True]           8.5359      1.699      5.025      0.000       5.201      11.871
Professional[T.True]      6.1226      3.989      1.535      0.125      -1.710      13.955
PhD[T.True]              23.6612      6.468      3.658      0.000      10.962      36.361
male                      3.9151      1.167      3.355      0.001       1.624       6.207
age                       1.8814      0.330      5.709      0.000       1.234       2.528
agesq                    -0.0179      0.004     -4.465      0.000      -0.026      -0.010
white                     2.9005      1.156      2.509      0.012       0.631       5.170
child                     1.0417      1.122      0.929      0.353      -1.160       3.244
marital_status            1.0617      1.048      1.013      0.311      -0.996       3.120
==============================================================================
Omnibus:                       96.880   Durbin-Watson:                   1.823
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              155.562
Skew:                           0.904   Prob(JB):                     1.66e-34
Kurtosis:                       4.452   Cond. No.                     3.08e+04
==============================================================================
```

```
Regression: reg4
                            OLS Regression Results
==============================================================================
Dep. Variable:                  wages   R-squared:                       0.246
Model:                            OLS   Adj. R-squared:                  0.225
Method:                 Least Squares   F-statistic:                     290.9
Date:                Fri, 19 Jan 2024   Prob (F-statistic):          1.69e-309
Time:                        19:03:28   Log-Likelihood:                -2696.2
No. Observations:                 694   AIC:                             5432.
Df Residuals:                     674   BIC:                             5523.
Df Model:                          19
Covariance Type:                  HC1
=============================================================================================
                                coef    std err          t      P>|t|      [0.025      0.975]

---------------------------------------------------------------------------------------------
Intercept                   -28.5772      6.456     -4.426      0.000     -41.254     -15.900
Associate[T.True]             0.7287      1.296      0.562      0.574      -1.816       3.273
Bachelors[T.True]             9.0935      1.288      7.061      0.000       6.565      11.622
Masters[T.True]              11.2560      1.949      5.775      0.000       7.429      15.083
Professional[T.True]         13.7779      6.922      1.990      0.047       0.187      27.369
PhD[T.True]                  25.7202      8.598      2.992      0.003       8.839      42.602
male                          3.4591      2.573      1.344      0.179      -1.593       8.511
male:Associate[T.True]       -0.9642      3.654     -0.264      0.792      -8.139       6.211
male:Bachelors[T.True]       -3.2340      2.850     -1.135      0.257      -8.830       2.362
male:Masters[T.True]         -9.3977      3.797     -2.475      0.014     -16.853      -1.943
male:Professional[T.True]   -12.2790      8.223     -1.493      0.136     -28.425       3.867
male:PhD[T.True]            -10.2181      8.949     -1.142      0.254     -27.790       7.353
age                           1.9617      0.335      5.852      0.000       1.303       2.620
agesq                        -0.0189      0.004     -4.648      0.000      -0.027      -0.011
white                         3.0248      1.170      2.586      0.010       0.728       5.321
child                         0.6691      3.891      0.172      0.864      -6.970       8.308
marital_status               -0.6420      1.091     -0.588      0.557      -2.785       1.501

marital_status:male          7.8976      2.994      2.638      0.009       2.019      13.777
male:child                  -3.7118      2.982     -1.245      0.214      -9.567       2.144
age:child                    0.0188      0.100      0.187      0.852      -0.179       0.216
==============================================================================
Omnibus:                       93.710   Durbin-Watson:                   1.825
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              147.597
Skew:                           0.888   Prob(JB):                     8.91e-33
Kurtosis:                       4.396   Cond. No.                     6.69e+04
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

Choosing the model:

As we look at the BIC values from all models, we can see that the lowest value is of Model 2, but Model 3 also has a value pretty close to the lowest one. If we consider our RMSE values, we can remember that Model 4 had the lowest. With BIC, that is not the case as Model 2 performs the best here. So, what will we conclude and how will we select a model. I believe that due to the penalty term in BIC when the number of variables increases, that causes our Model 4 to have a higher value. As we can see that it isn't the highest value, 5583 of Model 1 being the highest BIC and Model 4's BIC is 5523. Therefore, it seems like the best model to use with the live data would be Model 4 since it performs the best with RMSE, and the BIC values aren't bad either.