## Content for quick navigation

[1](#_br9raxkz4fnw)

[**Purpose 2**](#_gm69eetptfcm)

[**Few words on dataset 2**](#_9ltswk3bdl5o)

[**Gender 2**](#_ixlsvqxaok4r)

[**Experience & Education by gender 3**](#_wim24h2uf4n5)

[**Industry 5**](#_ea2pcrcyrulx)

[**Location 7**](#_mcq7i2jwlado)

[**Occupation 7**](#_ubce0vt3angp)

[**Model development 7**](#_by3z41j4lrvu)

[**Results & summary 8**](#_pe2axo40zrn)

[**Recommendations on what could improve the model 11**](#_9tw4n7snbmff)

## 

## Purpose

The purpose of this task is to identify the key variables that impact wages and develop a model to predict individual hourly wage.

## Few words on dataset

The data set contains 526 rows of data which correspond to 526 individuals in the United States.

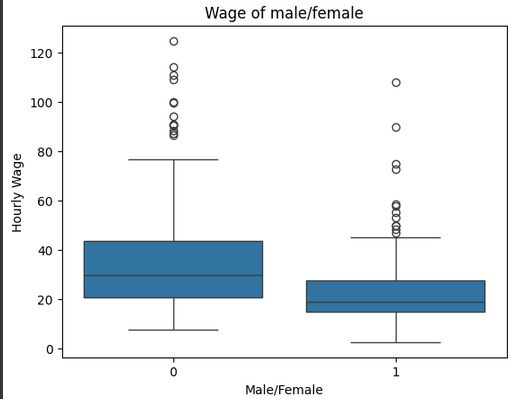
The dataset features further information: hourly wage, education, experience & tenure of individuals in years, number of dependents, economic area and region, occupation & industry.

A more detailed breakdown of each variable is available in the workbook.

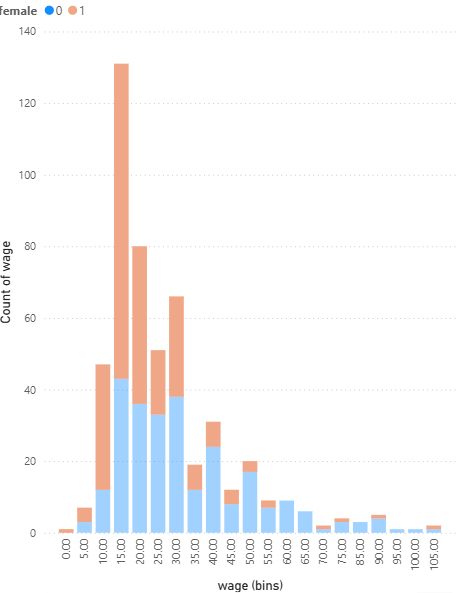
16 lines of data were excluded in the process of data verification. Wage outliers which were examined for validity in the workbook but were not excluded from data as those turned to be valid data points.

## Gender

There is a significant difference in median wage between male & female as shown below:

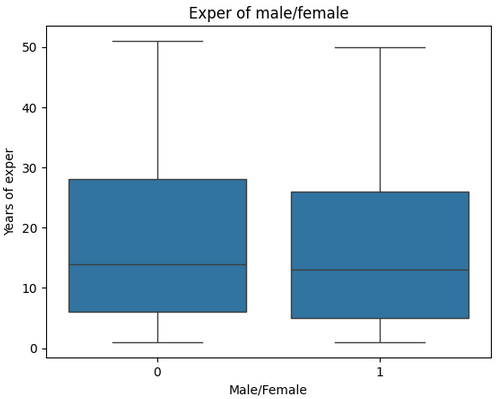


There are more females in lower paying wage bins while the proportion of male is higher in higher wage bins.



## Experience & Education by gender

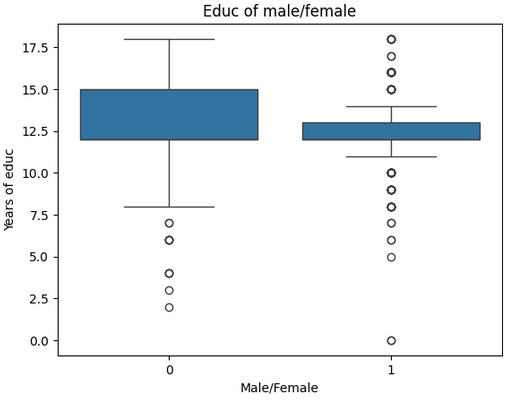
Both groups in a given dataset have similar experience distribution.



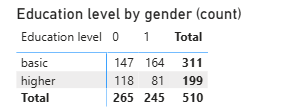
For male the spread of wage increases with years of experience peaking between 15 and 30 years of experience but after 30 years the wage spread has a tendency to drop. For female - the spread seems to increase quicker at lower experience (10 - 15 years) and decline with some spike around 35-40 years of experience as well.



The amount of years spent on education is a little higher on average within male group compared to females.



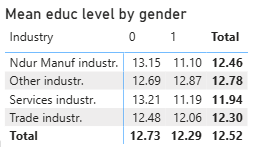
If we take 12 years of education as a basic education - there are more males who take higher (13 and more years) education compared to females.



Increase in wage is positively correlated with education increase in general for both male and female - however the upper band of wage spread increases with higher education more sharply and steadily for male; for females this increase is less sharp and smaller overall in terms of wage level.



Average education level is higher among male across all the industries except for “Other” which is somewhat generic category to judge by:



## Industry

The categories of industries available from the sample population suggest further spread of wage:



The manufacturing industry has the highest middle pay, followed by trade and then services.

“Other” category has top middle pay among all - this seems correct given the other defined industries 3 are not an exhaustive list of industries and more detailed classification might be beneficial in order to make predictions better.

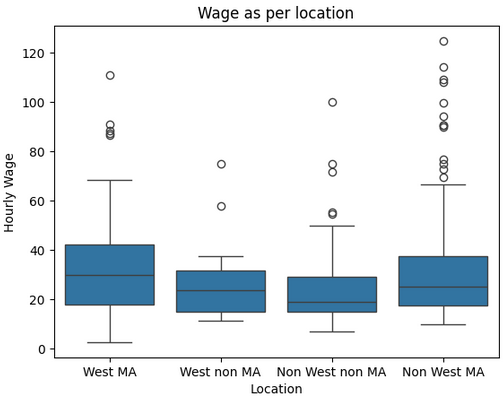
Females are paid less across all industries.



This can correlate with the fact that education is higher across all industries except for “Other” within male group.

Overall - average education level is higher within “Other” and “ Non dur manufacturing” industries - which again correlated with higher pay in those industries compared to other 2 industries.

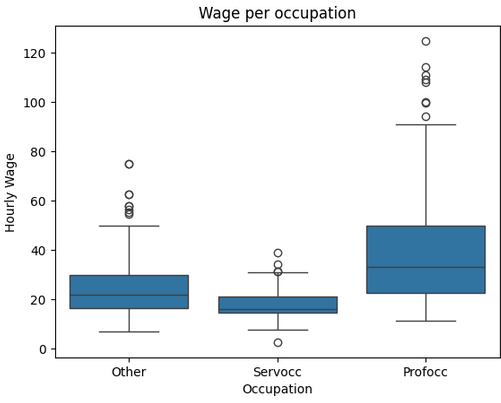
## Location



Middle wages in metropolitan areas are higher compared to non metropolitan areas, whereas Western regions’s middle pay is higher in both MA and non MA areas compared to non Western regions respective areas. The spread (and outliers) are higher in non west MA then West MA though, suggesting that there are more high earners in non west MA.

## Occupation

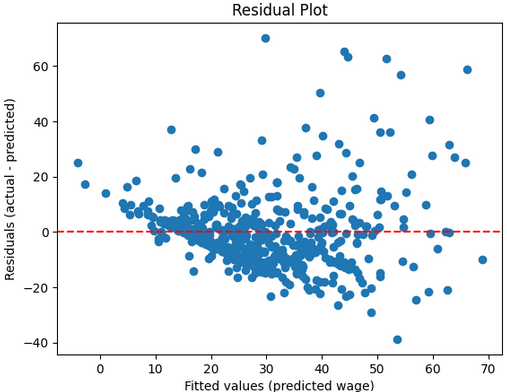
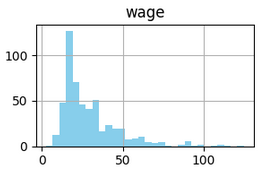
Professional occupation has higher middle pay then service occupation or the rest not defined.



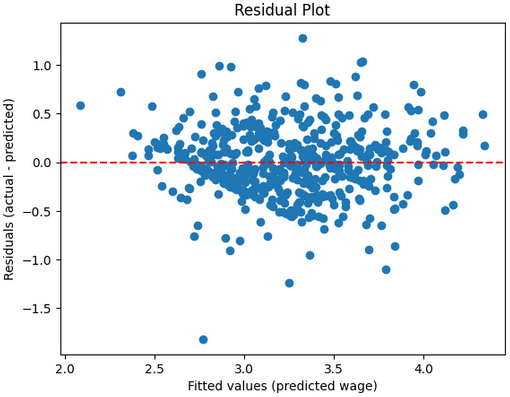
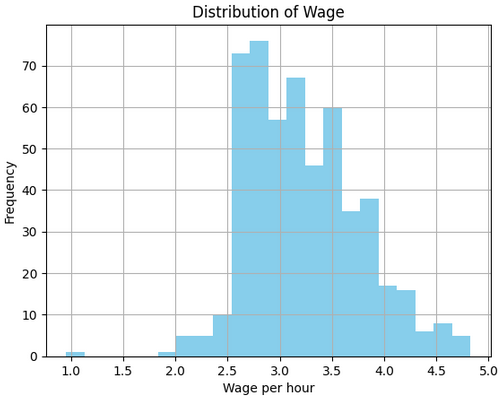
## Model development

The wage distribution is right skewed, so log(wage) was applied for further model building which has improved the residual plot:

Before:



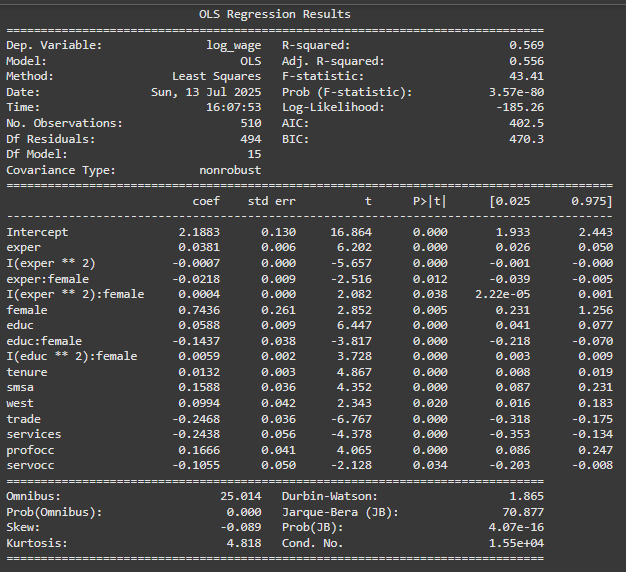
After:



There is a non linear relation between wage & educ for female as well as wage & exper so edu\*\*2 & exper\*\*2 were used in the model.

## Results & summary

Below are the results of the model with best adj r^2 achieved given the stat significance of the separate variables:



Each year of education increase wage by 5% for male, for female the year of education have negative impact - (-14%) but do increase at higher education levels (edu\*\*2):female.

Each year of experience increases wages by 3% for male and decreases by 2% for females but the higher the experience gets - the smaller effect it will have on wages - negative for men and positive for females.

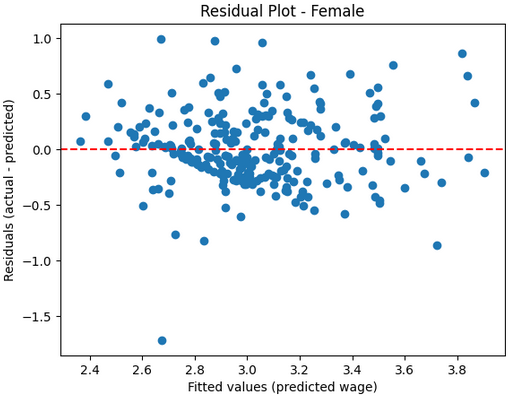
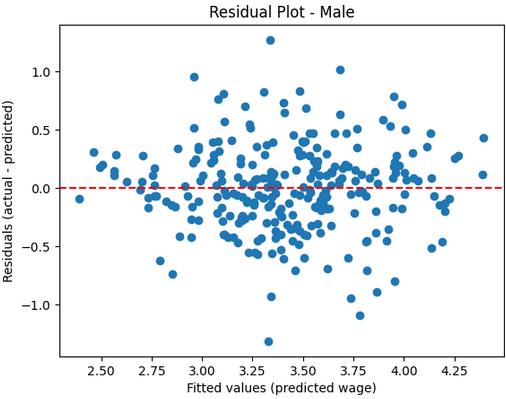
Wages in Metropolitan areas are higher by 15%, in Western regions wage is higher by 9%.

Being in professional occupation increases wages by 16% whereas being in service occupation has a negative effect on salary of 10%.

Wages in both trade and services industries are lower by 24%.

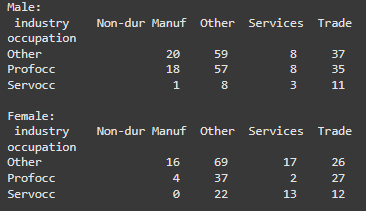
Residual plots compared between male & female group suggest that there is a pattern in how the model underestimates the wage for females.

This could be due to the fact that female’s middle 50% earners are more tightly clustered whereas for male the middle 50% earnings have wider spread as well as some outliers.

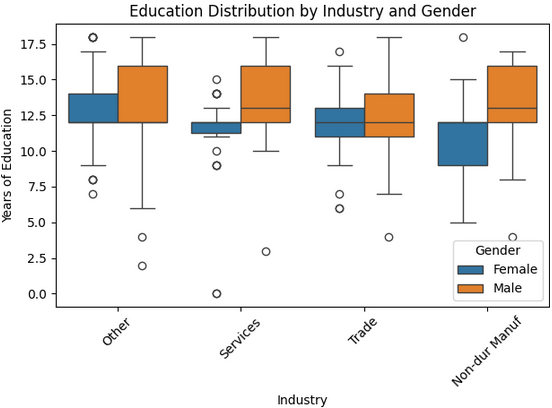


Women are paid less on average across all industries presented - as mentioned earlier.

It can correlate with the fact that less females are in prof occ among industries then male.



As well as the fact that educ is higher among male across industries:



## Recommendations on what could improve the model

* Level of career (Manager/Non managing - jun, mid, sen)
* The number of years in a current career could help, as experience might include total working experience not always relevant to current role and tenure is limited to 1 employer and does not explain either.
* Maternity break can be an important factor to consider for the female population.
* The breakdown of industries is limited and expanding it might be helpful for better modelling - “other” industries’ groups are quite big in a given data.
* Other locations to be specified.
* Age of participants or year of birth.