MA334-SP-7: Final Project (2024-25)

Maha Ejaz

2025-04-17

1. Data exploration

The no.of variables in the data set are 12.

The no.of observations in the data set are 1113.

The data types of the variables in the dataset are integer, numeric, character.

The average age of individuals is 42.55, years with most falling between 80 and 18.

People worked on an average of 41.89 hours last week, with some variation.

People work on average of 41.89 hours last week, but some report working a maximum of 80 hours.

The wage distribution suggests that while some earn as little as 2.75 per hour, the highest earners make up to 99 per hour with an average wage of 23.05 per hour.

The table below shows information about the gender distribution:

Table 1: Gender Distribution

Gender	Count
Male	625
Female	488

Using the above information, we can conclude that male makes up a larger portion of the sample, but there is still a substantial number of females represented in the data set.

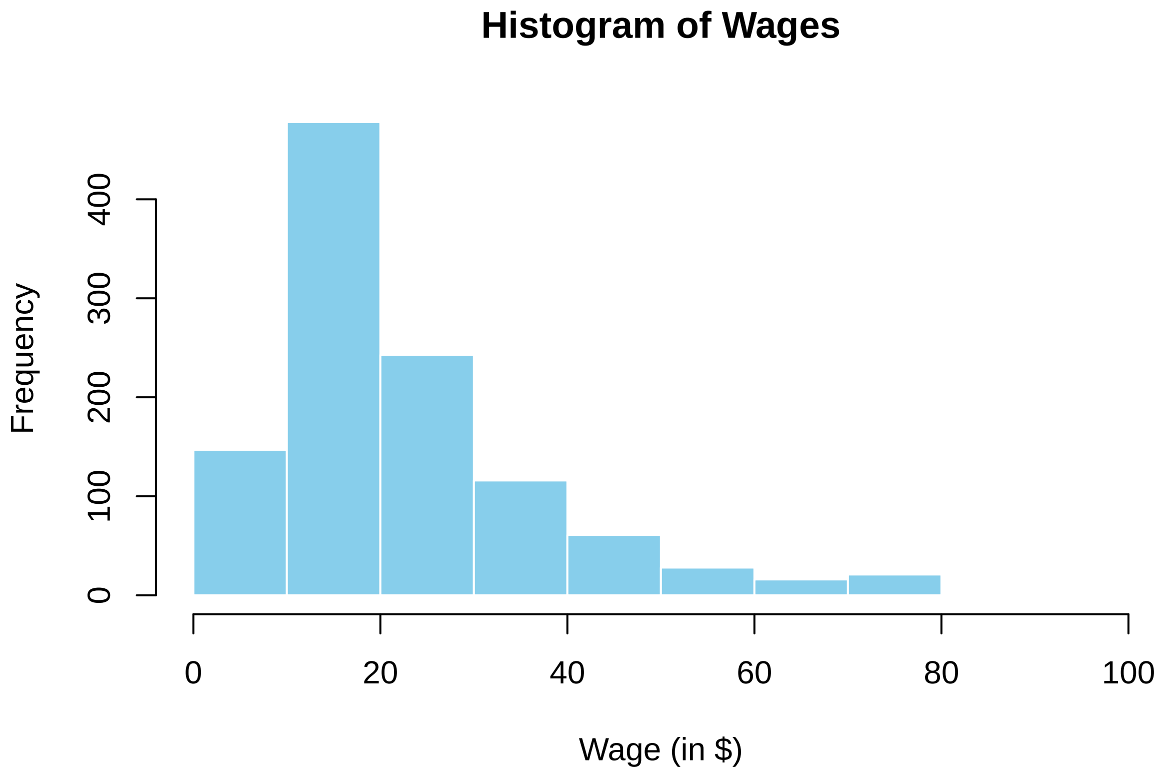
The table below shows the no.of people in different education levels

Table 2: Education Level Distribution

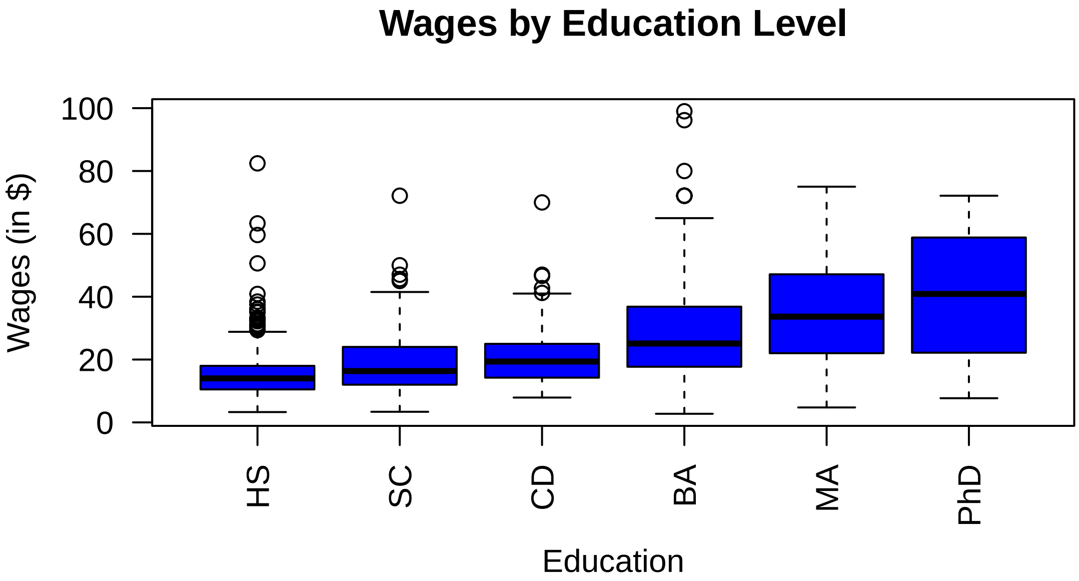
Education Level	Count
High School	356
College No Degree	194
College Degree	125
BA	287
MA	135
Phd	16

- The most common education level in the data set is High School, followed by BA and College No Degree, which is consistent with the general distribution in many regions.
- PHD holders are significantly underrepresented, which may be expected as Phds are a relatively rare qualification.

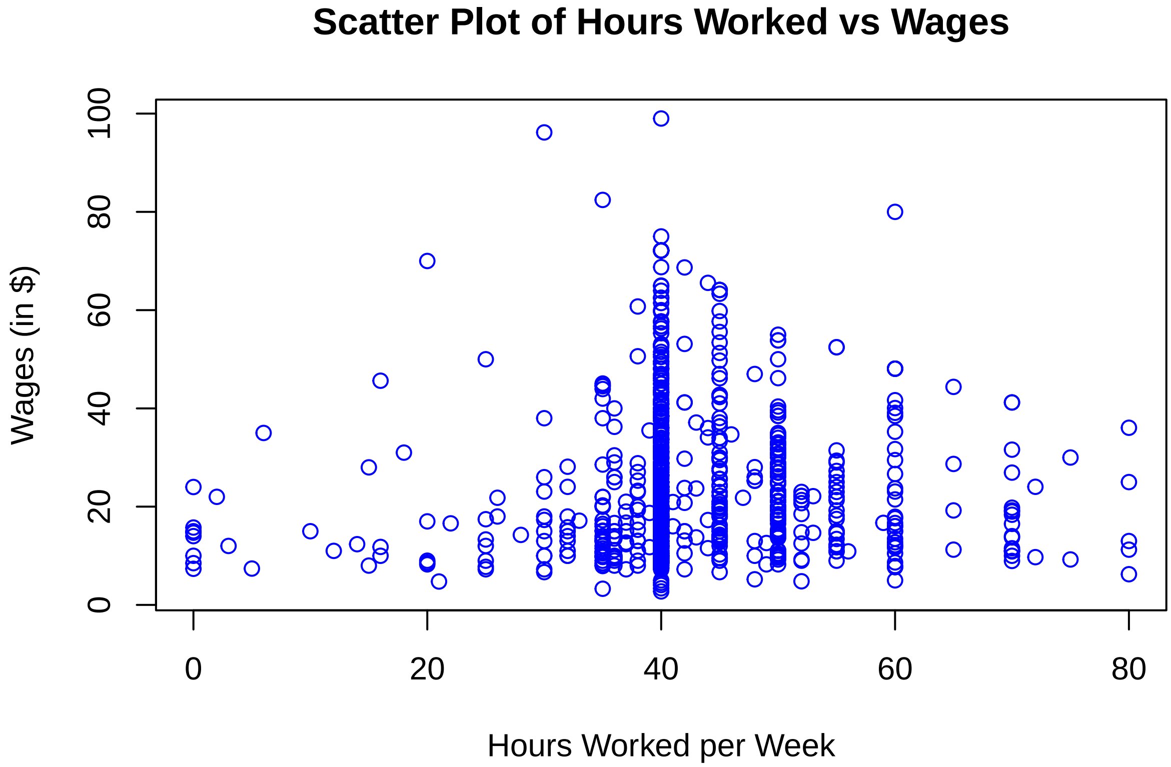
- The data shows that College Degree, BA, MA & PhD are less common compared to high school and some college education, suggesting that a large portion of the population in the dataset might not have completed their education



- The histogram above shows that the distribution of wages is right-skewed, meaning that most individuals earn lower wages, with a few earning significantly higher wages.
- The majority of people earn between $10 and $30 per hour, as seen by the tallest bars in that range.
- Very few people earn above $60, which makes those values potential outliers as they can be termed as high-income earners
- This suggests income inequality between the dataset, with a large portion of the population earning on the lower end of the wage scale.



- The box plot above shows that wage tends to increase with higher levels of education, with PhD holders earning the most on average, and High School graduates the least.
- It also reveals greater wage variability among those with advanced degrees(BA, MA, PhD) as shown by the wider spread and more outliers in these groups



- The scatter plot above shows the relationship between hours worked per week and wages.

- Most individuals work between 35 and 50 hours per week.

- There are several outliers working above $60, even for fewer working hours.
- No linear pattern is observed - suggesting that wages do not consistently increase with more hours worked.

Correlations between variables

- The correlation between education and wage is 0.17, indicating a weak positive correlation. It has a slight tendency for wages to increase with age but its not a strong effect.
- The correlation between age and education is very weak which is 0.09 as older individuals may have slightly more education, but its almost negligible.
- Moderate to strong positive correlation between education and wage which is 0.49, which means that as education increases, wage tends to increase.
- Relationship between hrs worked and wage rate is very weak which is 0.05 may be due to variations in wage types.

2. Probability, probability distributions and confidence intervals

Solution

The probability that 1 or more out of the 5 will NOT be covered by private health insurance is 0.62.

The probability that a person selected at random from your data set will have 1 or more children, given that they are married is 0.602.

The probability distribution for the variable 'nchild' is:

Table 3: No.of Children

No.of Children	Count	Probability
0	622	0.5588
1	193	0.1734
2	201	0.1806
3	74	0.0665
4	16	0.0144
5	5	0.0045
6	2	0.0018

- The mean of nchild is 0.82.
- The variance of nchild is 1.22.
- The probability that nchild is greater than or equal to 3 is 0.087.

3. Point Estimates, Confidence Intervals and Hypothesis Tests

- The point estimate for households with 2 children is 25.07.
- The confidence interval on the mean population wage with 2 children is [22.98 , 27.16].
- The point estimate for households with 5 or more children is is 19.62.
- The 95% CI on the mean volume of population wage is [5.73 , 33.51].

- A point estimate for the mean wage in households with 5 or more children is calculated as 19.62. However, since the sample size is only 7, the 95% confidence interval may be unreliable. The results should be interpreted with caution due to the limited data available in this subgroup.

The following table shows the distribution of insurance status by gender.

Table 4: Insurance by Gender

Gender	Insurance	Count
Female	Insured	416
Male	Insured	502
Female	Not Insured	72
Male	Not Insured	123

- A Chi-square test is being used to determine the association between two categorical variables.
- Null Hypothesis(Ho): Insurance status is independent of gender.
- Alternative Hypothesis(H1): Insurance status is not independent of gender.
- The p value is 0.039.
- As p value is 0.039 “which is <0.05 so we conclude that there is a statistically significant association between insurance status and gender”.

4. Simple Linear Regression

Regression Model for individuals below the age of 35

- The equation of the regression model for young individuals is: log(wage) = 1.4847 + 0.0485 x age.
- The intercept represents the log(wage) when age=0.This shows the baseline value of log(wage).
- The slope represents the change in log(wage), so for every one year increase in age, the log(wage) increases by 0.0485.
- The p-value for age coefficient is 6.11e-14 which is extremely small as it is less than 0.05, which means that age is a significant factor in predicting(wage). Therefore, we reject the null hypothesis.
- The Multiple R-square value is 0.1521 which indicates that about 15.21% of the variability in log(wage) is explained by age. This suggests that age is a predictor of log(wage), but because other factors are not included in the model likely explaining the remaining 84.79%.
- The adjusted R-squared value is 0.1496 which accounts for the no.of predictors in the model. Since this is very close to R-squared, it suggests that adding age does not significantly increase the explanatory power of model
- The F-statistic tests, whether the overall regression model is a good fit for the data. Since the p-value associated with the F-statistic is very small, we can conclude that age is actually a significant factor explaining log(wage)

Regression Model for individuals above the age of 35

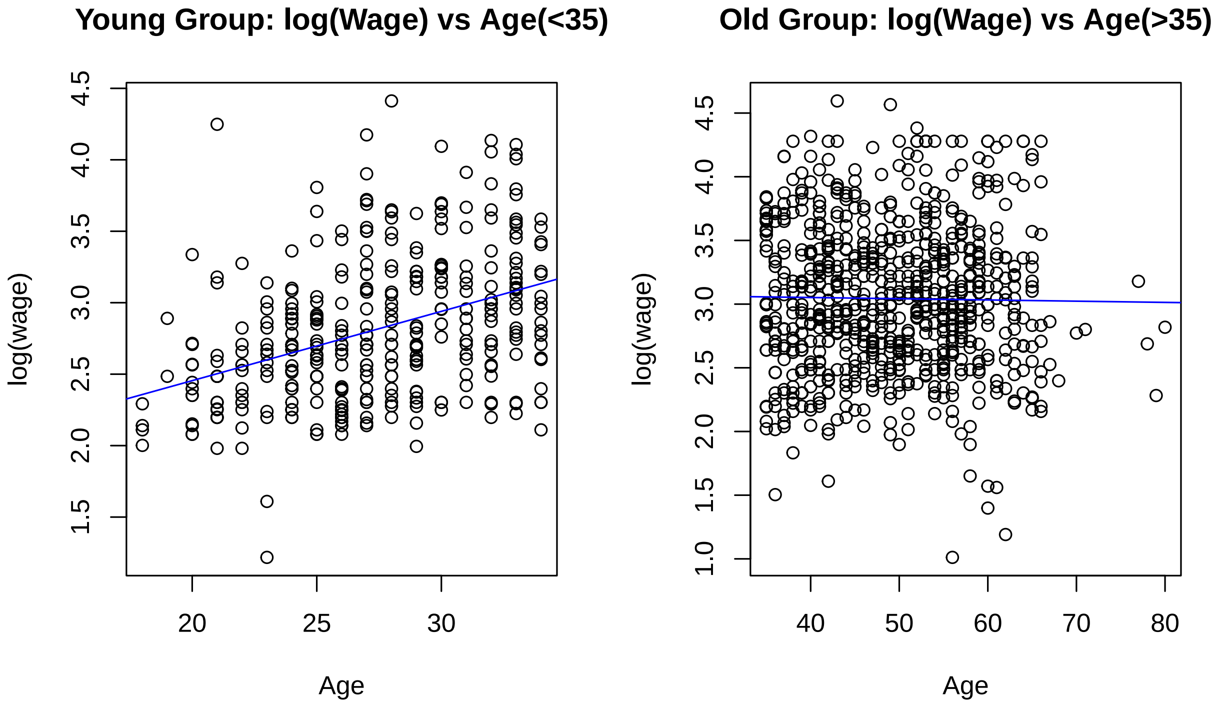
- The equation of the regression model for old individuals is: log(wage) = 3.0903 - 0.00096 x age.

- The intercept represents the log(wage) when age=0.This shows the baseline value of log(wage).
- The slope represents the change in log(wage), so for every one year increase in age, the log(wage) decreases by - 0.00096.
- The p-value which is 0.682 for age coefficient is larger than 0.05, which means that age is not significant factor in predicting(wage) for people above the age of 35.
- The Multiple R-square value is 0.0002189 which indicates that about 0.022% of the variability in log(wage) is explained by age. This suggests that age is not a predictor of log(wage) as the value is quite low, explaining that age has no explanatory power in predicting log(wage).
- The adjusted R-squared value is -0.001085 which indicates that the model is a very poor fit and age does not explain the variability in log(wage).
- The F-statistic tests, whether the overall regression model is a good fit for the data. The p-value associated with the F-statistic is very high, indicating that the model is not significant. This suggests that age does not provide any meaningful explanation for variation in log(wage) for older people .

Conclusion - In the young group, wage increases significantly with age. - In the old group, age is not a significant predictor of wage.

This may suggest that wage growth levels off or stabilizes as individuals get older, possibly due to career plateaus or other non-age-related wage factors in later life.

Scatter Plot for individuals below and above the age of 35 respectively



Explanation of Coefficient of Young Group - The R-squared value is 0.1521 which indicates that about 15.21% of the variability in log(wage) is explained by age. The R-squared value is relatively low, but it still indicates a positive trend, meaning age has some, though limited, explanatory power in determining the wages.

Explanation of Coefficient of Old Group - The R-squared value is 0.0002189 which indicates that about 0.022% of the variability in log(wage) is explained by age which is almost negligible. The regression line in the above diagram is very flat, as the relationship between age and log(wage) for Older group is weak.

5. Multiple Linear Regression

Explanation on what has been done to fit multiple linear regression models against all the remaining variables

- We have split the dataset into two subsets: individuals younger than 35 (young_df) & individuals over the age of 35 and old (old_df). To prepare for the regression model, we have converted the categorical variables (gender,education level, insurance covered, metropolitan area, union member, race & marital status) into factors, so that lm() handles them using dummy variables.
- We then fitted a multiple linear regression model for each group using log(wage) as the response variable and all other variables as explanatory variable. The formula ~. was used to include the remaining variables in the model. This approach allows us to assess how different factors effect the logarithm of wages in different age groups.

Explanation of Multiple Regression Model for Young Group

Adjusted R2 :This is 0.915 which indicates that 91.5% of variability in log(wage) is explained by predictors in the model.

Significant Predictors ($p<0.05$) are as follows:

- Age (positive effect).
- Multiple levels of education (especially BA & MA).
- Female Gender (gender1) is associated with lower wage.
- Being insured has positive effect.
- Living in a metropolitan areas has a positive effect.
- Being a union member has a positive effect.

Conclusion:

For younger individuals, the model suggests that age, education, union membership, insurance and urban location significantly contributes to higher wage. The positive and significant age coefficient confirms the earlier simpler regression finding that as young individuals age, their wages tend to increase.

Explanation of Multiple Regression Model for Old Group

Adjusted R2 : This is 0.890 (89%) which is slightly lower than the younger group.

Significant Predictors ($p <0.05$) are as follows:

- Multiple levels of education (1.College with no degree, 2. College with degree, 3.BA)
- Being insured has positive effect
- Being a union member has a positive effect

Age is not a significant factor in this group as $p =0.261$

Conclusion:

For older individuals, age no longer plays a significant role in predicting wage. This contrasts with the young group and supports the idea that wage growth with age plateau at a certain point. Education, union affiliation and insurance continues to have a significant impact.

Comparing Multiple Regression to Simple Linear Regression

Comparison for Younger Individuals:

Conclusion: For younger individuals, the effect of age on wage remains significant even after controlling other factors, which strengthens the evidence that age independently contributes to wage.

Comparison for Older Individuals:

Conclusion: For older individuals, age does not significantly impact wage and this holds true even when other variables are included. This reflects a weight plateau to age in older generations.

Summary

Both models have extremely small p-values for the F-statistic which suggests that the models as a whole are statistically significant. This indicates that at least one or more of the independent variable contributes meaningfully to predicting the response variable, log(wage).

Explaining why a reduced model with fewer variables may be preferable to the full model

Including too many variables can lead to over fitting, where the model performs well on training data but poorly on new data (test data). This eventually produces less reliable predictions than those from a model that picks out only the underlying trend.