Wage Determinants Analysis: A Regression and Exploratory Study on Young and Old Workers

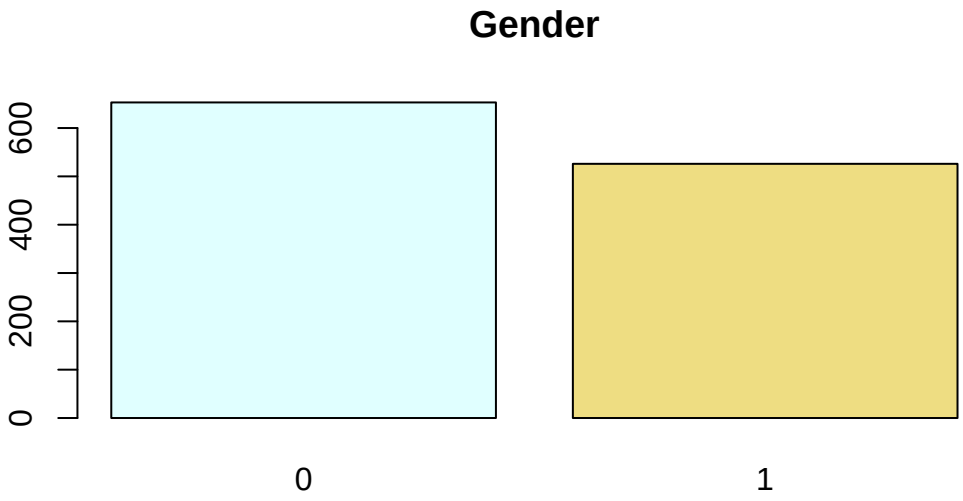
Mada Sai Kiran

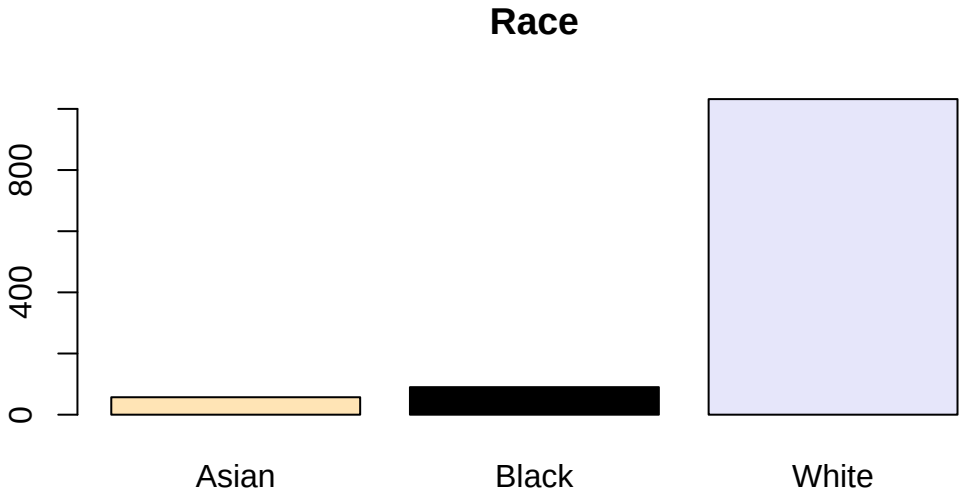
1 Data Exploration

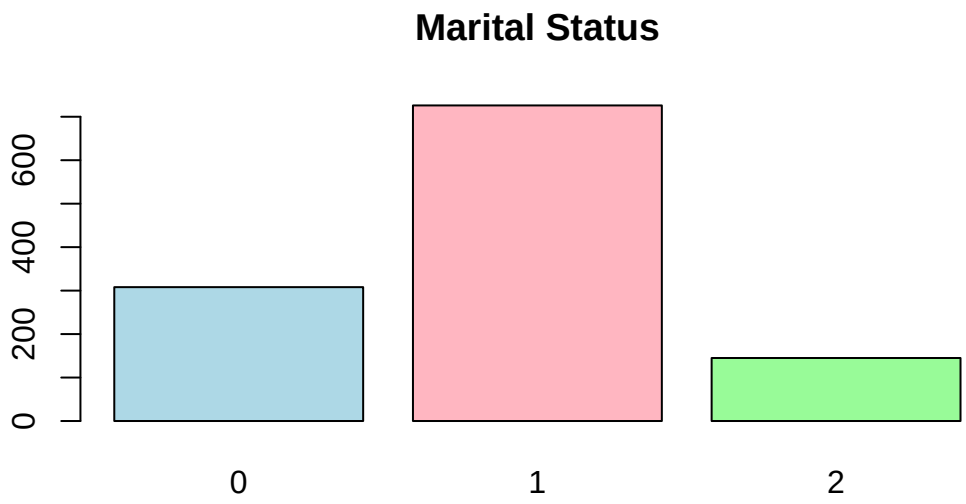
```
##       age             educ           gender          hrswork
##  Min.   :18.00   Min.   :0.000   Min.   :0.0000   Min.   : 0.00
##  1st Qu.:32.00   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:40.00
##  Median :42.00   Median :2.000   Median :0.0000   Median :40.00
##  Mean   :42.23   Mean   :1.768   Mean   :0.4461   Mean   :41.79
##  3rd Qu.:52.00   3rd Qu.:3.000   3rd Qu.:1.0000   3rd Qu.:45.00
##  Max.   :85.00   Max.   :5.000   Max.   :1.0000   Max.   :75.00
##       wage           nchild
##  Min.   : 3.46   Min.   :0.0000
##  1st Qu.:13.01   1st Qu.:0.0000
##  Median :19.62   Median :0.0000
##  Mean   :23.22   Mean   :0.8694
##  3rd Qu.:29.80   3rd Qu.:2.0000
##  Max.   :72.13   Max.   :6.0000
```

The dataset contains a variety of variables related to individuals, including age, education level, gender, hours worked per week, wage, and number of children. These summary statistics provide a first look at the distribution of each variable and reveal that wages and hours worked have a wide range of values.

From the histograms we got the information like most people earn lower wages, but only a few people make significantly more.Similarly, while most people work a standard 40 hours a week, there are some people who put in much longer hours. Age seems to be more evenly distributed across the sample.


Gender
0
200
400
600
0
1


Race
0
400
800
Asian
Black
White



- Above bar plots show that
  - There are more males when compared to females.
  - Larger number of individuals belong to White race group among all.
  - Marital status distribution shows that a significant proportion are married.

Table 1: Correlation Matrix of Selected Variables

	age	hrswork	wage	nchild
age	1.0000000	0.1606775	0.2400348	-0.0181092
hrswork	0.1606775	1.0000000	0.0862780	0.0497726
wage	0.2400348	0.0862780	1.0000000	0.0348539
nchild	-0.0181092	0.0497726	0.0348539	1.0000000

This correlation matrix helps us to know that the relationship between hours work and wage is positively correlation. Which mean if individuals work more they will get more Wages. Additionally, relationship between age and wage is weaker correlation which mean there is no that much relation between both elements.

2 Probability & Confidence Intervals

As per the given data set, Approximately 16.9% of individuals are not covered by private insurance. This gives a probability of 0.603 that at least one out of five randomly selected individuals lacks private insurance.

Among married individuals, 60.9% have at least one child.

Table 2: Frequency of Number of Children

Number_of_Children	Frequency
0	647
1	191
2	234
3	75
4	21
5	9
6	2

This is the table for sum of number children per individual as per the given data set

The mean number of children in a household is 0.87, with a variance of 1.31. The probability that someone has three or more children is 9.1%.

3 Estimates & Hypothesis Tests

```
##
##  One Sample t-test
##
## data:  children2$wage
## t = 25.888, df = 233, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  22.83952 26.60227
## sample estimates:
## mean of x
##   24.7209
```

```
##
##  One Sample t-test
##
## data:  children5$wage
## t = 5.3515, df = 10, p-value = 0.0003229
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  13.20935 32.05610
## sample estimates:
## mean of x
##  22.63273
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  insurance_gender_table
## X-squared = 2.1079, df = 1, p-value = 0.1465
```

The mean wage for people with two children amounts to approximately 24.72. The confidence interval shows an estimated range for the actual population mean.

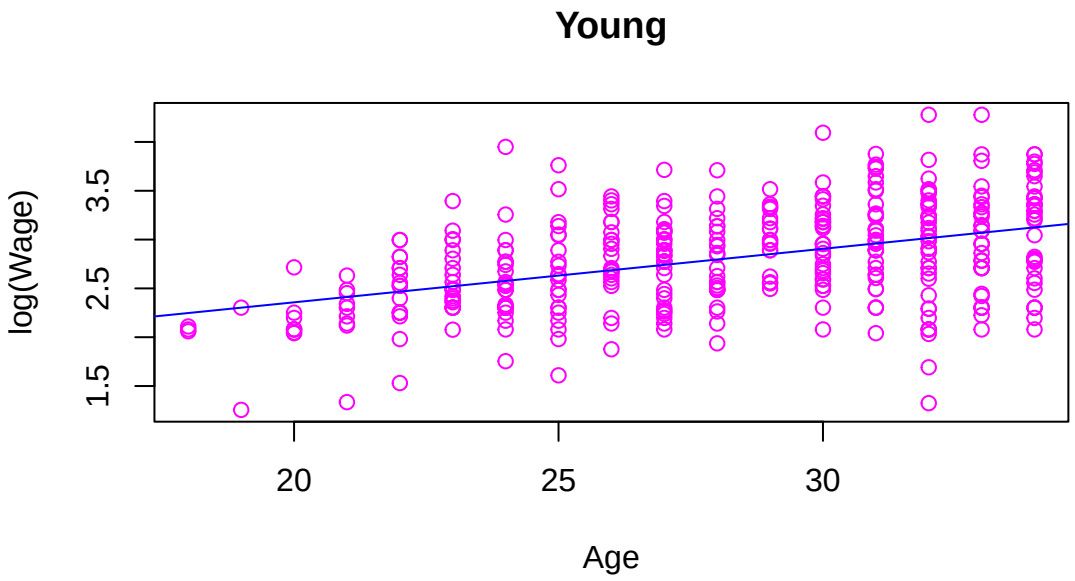
The mean wage for people with 5 children or above amounts to approximately 22.63. The confidence interval shows an estimated range for the actual population mean. We have only 11 individuals having children 5 or more in the given data set.

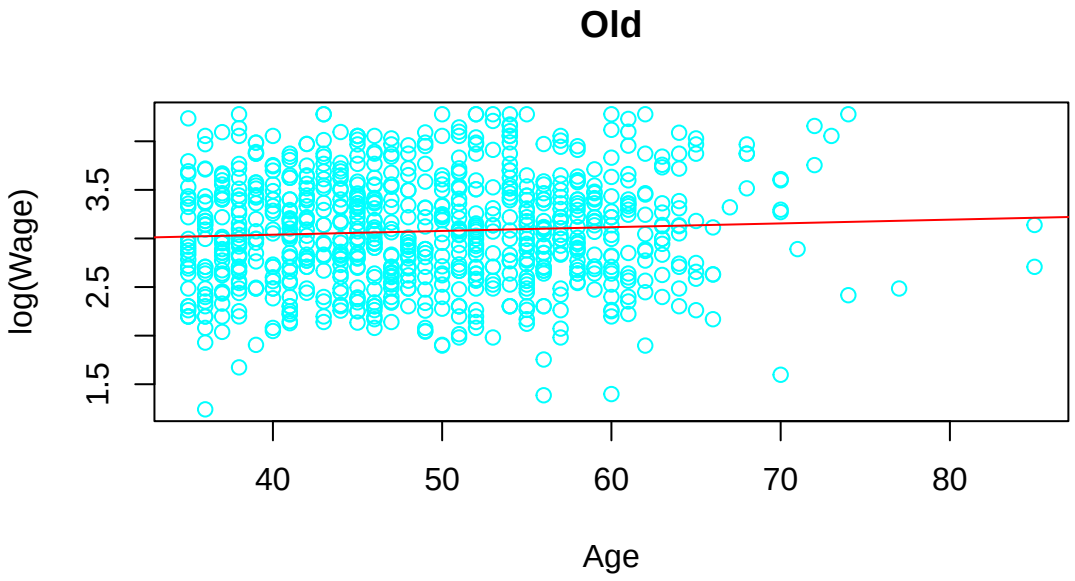
I evaluated chi-squared test for the connection between gender and insurance status. So, p-value determines whether we should accept or reject the statistical independence between the two variables.

4 Simple Linear Regression

To know how age affects wages, I first applied a log transformation to the wage data. Then I split age into young(<35) and old(>=35) age groups. Then I applied the linear regression model to see how age affects log-transformed wages within age groups.

- Young age group: The Model shows that there are slight changes in increasing wages as age increases, so younger people gain early career experience.
- Old age group: This showed effect of age on wages is smaller or even flat, it suggests that wage growth or stabilizes later in life.





I visualized scatter plots for better understanding of young age group and old age group effects log(wage)

- Here we can clearly see that in young scatter plot Blue line is increasing so as age increases wage also increasing
- where as in old scatter plot Red line is almost flat or it is slight increasing so life is stabilizing

5 Multiple Linear Regression

Table 3: Summary of the Young Model

term	estimate	std.error	statistic	p.value
(Intercept)	1.4564813	0.1975076	7.3743053	0.0000000
age	0.0361677	0.0062240	5.8110234	0.0000000
educ	0.1418885	0.0156098	9.0897179	0.0000000
gender	-0.1172011	0.0420879	-2.7846723	0.0056344
hrswork	-0.0033088	0.0020746	-1.5949151	0.1115873
insure	0.1625222	0.0521826	3.1144921	0.0019867
metro	0.0140579	0.0540372	0.2601529	0.7948910
nchild	-0.0109960	0.0237840	-0.4623282	0.6441185
union	0.0763979	0.0696682	1.0965959	0.2735332
raceBlack	0.0012933	0.1257443	0.0102848	0.9917996
raceWhite	0.1221127	0.0945972	1.2908710	0.1975566
marital	0.0200583	0.0402973	0.4977584	0.6189508
regionnortheast	0.0566703	0.0648955	0.8732543	0.3830924
regionsouth	-0.0019528	0.0540883	-0.0361046	0.9712185
regionwest	0.0127151	0.0601226	0.2114861	0.8326248

Table 4: Summary of the Old Model

term	estimate	std.error	statistic	p.value
(Intercept)	2.4656856	0.1692568	14.5677227	0.0000000
age	0.0033701	0.0020105	1.6762674	0.0940866
educ	0.1622046	0.0113822	14.2506783	0.0000000
gender	-0.2680010	0.0343054	-7.8122192	0.0000000
hrswork	-0.0022008	0.0021855	-1.0070103	0.3142421
insure	0.3448563	0.0518231	6.6544840	0.0000000
metro	0.0864879	0.0429073	2.0156900	0.0441749
nchild	-0.0040365	0.0150313	-0.2685400	0.7883547
union	-0.0112621	0.0470218	-0.2395082	0.8107745
raceBlack	-0.0129243	0.1008272	-0.1281832	0.8980370
raceWhite	-0.0603766	0.0810944	-0.7445223	0.4567848
marital	0.0450227	0.0319087	1.4109832	0.1586483
regionnortheast	0.0588051	0.0493690	1.1911338	0.2339634
regionsouth	-0.0275303	0.0470557	-0.5850591	0.5586772
regionwest	0.0462964	0.0499243	0.9273308	0.3540416

To explore what factors effects wages differently for younger and older individuals, I built two separate multiple linear regression models — one for people under 35 (young) and another for those aged 35 and above (old). In both models, I used log(wage) as the response variable, with all other available variables (except wage) as predictors. These are referred to as the full models.

Before fitting the models, I converted the categorical variables — gender, race, marital status, region, and education — into factors. Because, this step ensures R handles them properly in the regression analysis.

- Comparing the Models

When comparing these full models to the simple linear regression models, I found that the full models had higher R-squared values. This means they do a better job explaining the variation in wages because they consider multiple factors rather than just one.

This Full models also showed that the impact of different variables on wages changes with age. For example, education or marital status might play a bigger role in one group than the other, highlighting how wage determinants can shift across different life stages.

- Why Not Always Use the Full Model?

Even though the full models fit the data better, including all variables isn't always the good idea. When we add too many predictors, then model try to capture noise rather than the patterns. Here we it can leads to risk of overfitting. This make the model not to recognize properly.

That's why it can be useful to build a reduced model, keeping only the most meaningful variables. This can be done using selection techniques like AIC, BIC, or even just logical reasoning based on what makes sense. A simple model is usually easier to understand and interpret, and often performs better in the long run.