

Data Analysis 3 : Assignment 1

Ghazal

1. Introduction

The purpose of this assignment is to build four predictive models using linear regression. The source of data for this purpose is CPS-earning data set which can be found here. The chosen occupation for this project is *Human Resource worker* with the census occupation code of *630* which has 701 observations.

2. Data Munging and Transformation

Predicting the weekly earning is a quantitative prediction exercise. For the process of data munging and transformation, hourly wage_**(w)**_ is calculated by dividing the weekly earnings (earnwke) by the number of hours (uhours) and the log of the mentioned variable (**lnw**) is also created. To model other variables are created and transformed which are as following: education level, marital status, if an individual owns a child, working sector dummies are created. Other character variables such as Gender and education levels are created. To model non-linearity in age for the regression with earning per week as a dependent variable, quadratic age predictor is created. The sample is varied with high number of females (521) compared to men (180). For further data filtering process, education is included from high school to PhD. The education levels below college are added into *Non-degree* category, and both associate certificates of vocational and academic are added to *associate*. The data contains observation who are 18 years old or above and less than 64 years. Moreover, the selected data set is transformed to observations who earns more than one USD per week.

3. Variables, Interactions, and Regressions

Variables : Education is likely a strong predictor of earning per hour. Table 4, in the Appendix indicates that the mean difference between of a human resource worker with no degree to a human resource worker who holds a PhD degree is 27 USD. Moreover, other variables such as age and age square are added to the predication models. Gender plays an important role, Table 7 shows that female human resource workers earn 4 USD less than their male counterparts. Moreover, a binary variable capturing if human resource worker is married or otherwise, indicates interesting summary in the Appendix Table 10. It shows that married human resource workers earn more than the otherwise. **Interactions :** To further capture the interplay of independent variables, interactions are used. Interactions such as gender and education, marital status and gender, gender times owning a child and education are used to understand the interplay of variables. **Regressions :** Four linear regression models are built to prediction analysis. As Table 1, in the appendix shows, Model 1 is the simplest containing education dummies, as following Model 2 has the Model 1 explanatory variables along with age and age squared. Moreover, in the Model 3, more explanatory variables such as gender, working sector, marital status, and own-child are added. Model 4, is the most complex among all model. This mentioned model contains all the mentioned independent variables and the respective interactions.

4. Model Performance

BIC is the measure of the fit of a model using all the original data and it penalizes the model complexity and helps to avoid over-fitting. Models with lower BIC are generally preferred. Among the models Model 2 has the lowest BIC, however it has minimal difference with Model 3. The second measurement to evaluate model performance is RMSE, which is the average squared loss across several target observations. RMSE is the lowest for the models 3 and 4. By looking to the Table 2 for the result of cross validation of the models, it suggests that Model 2 and Model 3 have the best properties. Model 3 has the lowest for BIC and cross validation RMSE average. Thus, Model 3 is selected for the purpose of the project which contains 14 variables.

Table 1: Regression Models for Earning per hour

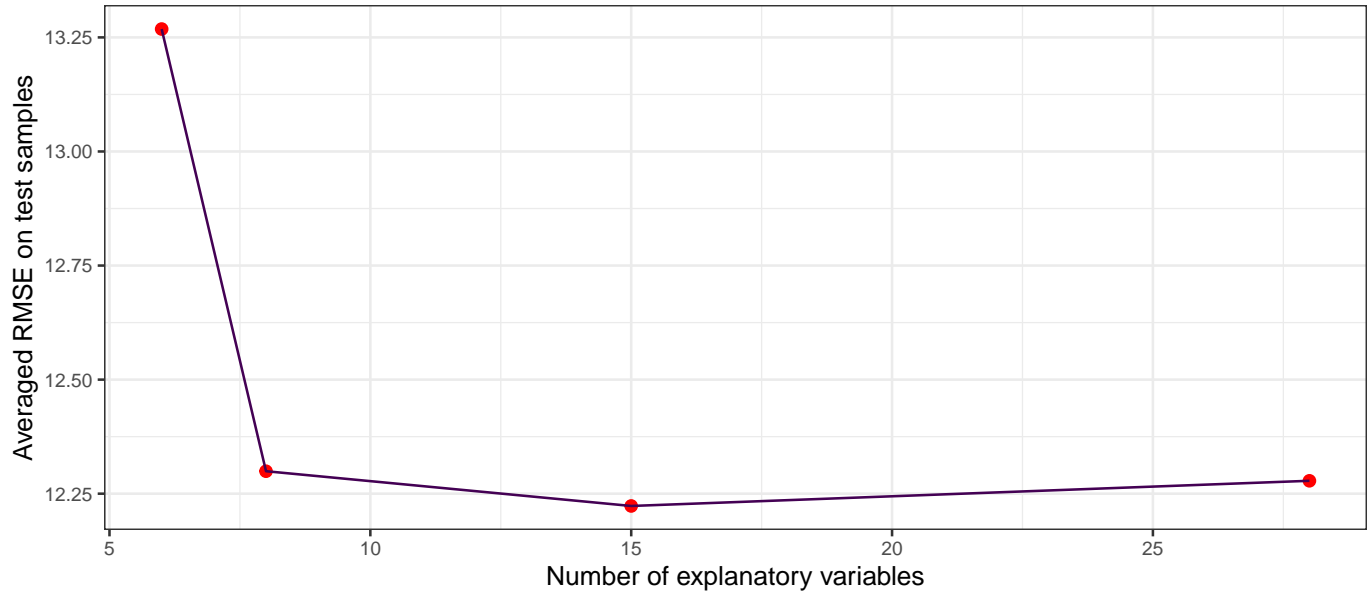
| | M1 | M2 | M3 | M4 |
|------------------------------------|---------------------|---------------------|---------------------|---------------------|
| Dependent Var.: | Hourly wage | Hourly wage | Hourly wage | Hourly wage |
| Intercept | 23.10*** (0.8037) | -22.81*** (5.378) | -20.52** (6.539) | -13.63* (6.572) |
| Associate | 0.8083 (1.451) | 0.1552 (1.345) | -0.2556 (1.306) | -3.622 (4.754) |
| BA Degree | 5.600*** (1.141) | 7.851*** (1.142) | 7.588*** (1.137) | 3.540 (3.069) |
| MA Degree | 7.823*** (1.742) | 8.596*** (1.678) | 8.142*** (1.653) | -4.287 (4.167) |
| Professional Degree | 7.016. (4.140) | 7.325. (3.946) | 6.013 (4.052) | -2.634 (6.896) |
| PhD | 27.20*** (6.873) | 24.81*** (6.406) | 23.16*** (6.629) | 15.16*** (2.980) |
| Age | | 1.908*** (0.2840) | 1.762*** (0.3245) | 1.850*** (0.3298) |
| Age Squared | | -0.0182*** (0.0035) | -0.0166*** (0.0040) | -0.0175*** (0.0040) |
| genderMale | | | 3.902** (1.198) | |
| classGovernment-Local | | | -3.165 (2.657) | -3.549 (2.742) |
| classGovernment-State | | | -6.105* (2.573) | -7.255** (2.675) |
| classPrivate,ForProfit | | | -0.6463 (2.225) | -1.622 (2.311) |
| classPrivate,Nonprofit | | | -1.131 (2.551) | -2.011 (2.603) |
| Has Child | | | 0.8024 (1.111) | 5.244 (5.690) |
| Married | | | 1.513 (1.022) | 1.080 (1.066) |
| Female | | | | 5.972 (9.845) |
| Associate x Female | | | | -10.21 (10.73) |
| BA Degree x Female | | | | -9.481 (10.09) |
| MA Degree x Female | | | | -0.3416 (10.60) |
| Female x Has Child | | | | -5.955 (6.029) |
| Female x educNoDegree | | | | -14.29 (10.07) |
| Female x educPhD | | | | -2.911 (13.29) |
| Has Child x educBachelors | | | | -5.111 (6.622) |
| Has Child x educMasters | | | | 3.798 (7.628) |
| Has Child x educNoDegree | | | | -8.396 (6.780) |
| Has Child x educProfessional | | | | 2.046 (9.128) |
| Female x Has Child x educBachelors | | | | 6.810 (7.124) |
| Female x Has Child x educMasters | | | | 0.8014 (8.659) |
| Female x Has Child x educNoDegree | | | | 8.834 (7.268) |
| S.E. type | Heteroskedast.-rob. | Heteroskedast.-rob. | Heteroskedast.-rob. | Heteroskedast.-rob. |
| AIC | 5,614.2 | 5,508.0 | 5,491.3 | 5,498.8 |
| BIC | 5,641.6 | 5,544.4 | 5,559.6 | 5,626.3 |
| RMSE | 13.157 | 12.162 | 11.898 | 11.743 |
| R2 | 0.06554 | 0.20151 | 0.23578 | 0.25562 |
| Observations | 701 | 701 | 701 | 701 |
| No. Variables | 5 | 7 | 14 | 27 |

Table 2: 4-Fold Cross Validation and RMSE

| Resample | Model1 | Model2 | Model3 | Model4 |
|----------|----------|----------|----------|----------|
| Fold1 | 13.56982 | 12.33827 | 12.53872 | 12.59866 |
| Fold2 | 12.44817 | 11.49459 | 11.30927 | 11.40808 |
| Fold3 | 12.54938 | 12.22487 | 11.92689 | 11.96471 |
| Fold4 | 14.40869 | 13.08738 | 13.04787 | 13.07657 |
| Average | 13.26833 | 12.29924 | 12.22310 | 12.27824 |

Appendix

Prediction performance and model complexity



Cross-Validation RMSE in the graphs shows lowest result for the Model 3. This Model contains 14 Variables

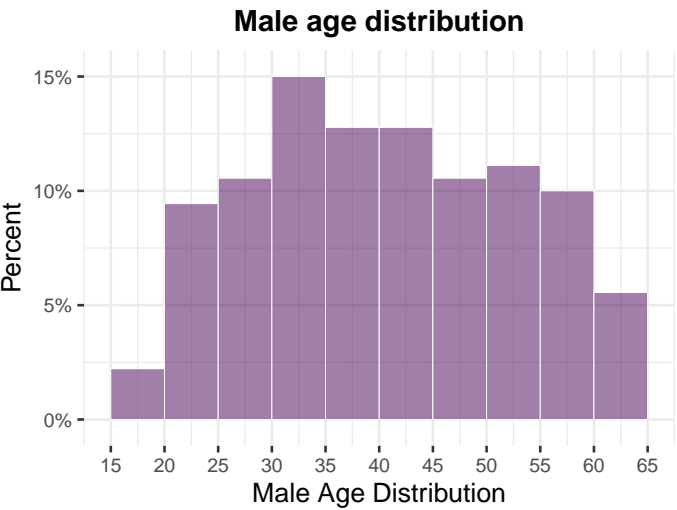
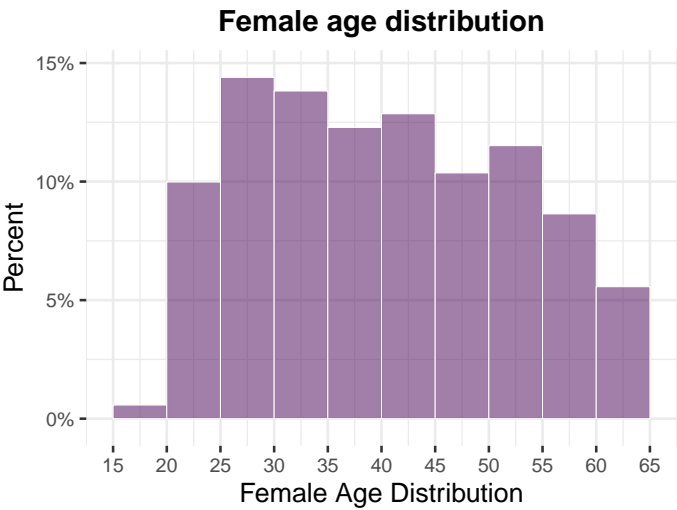
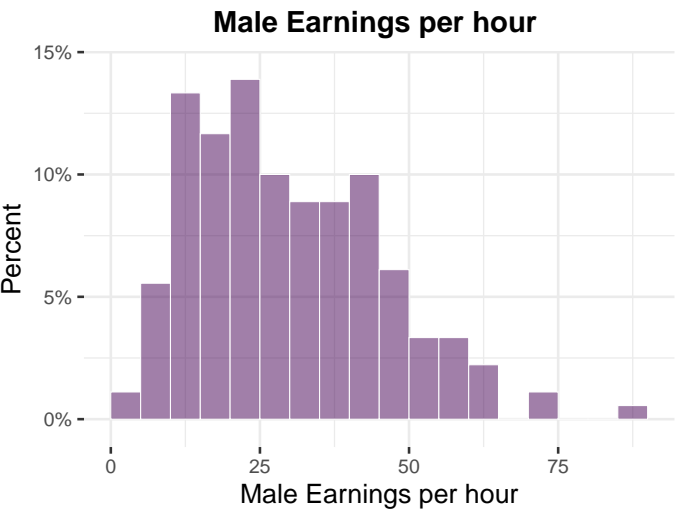
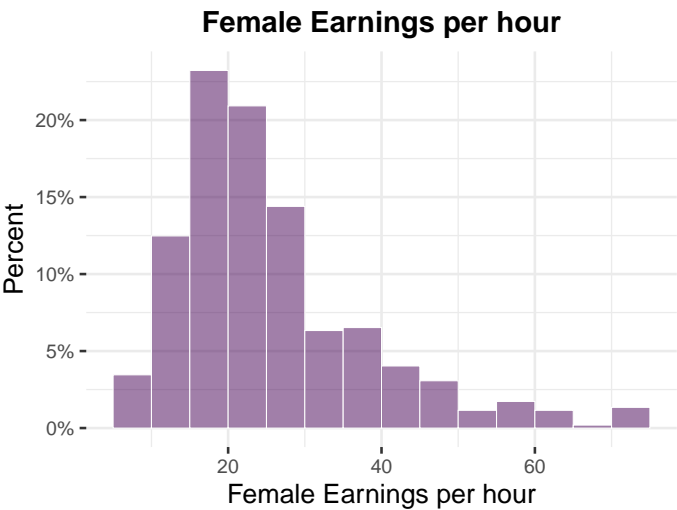
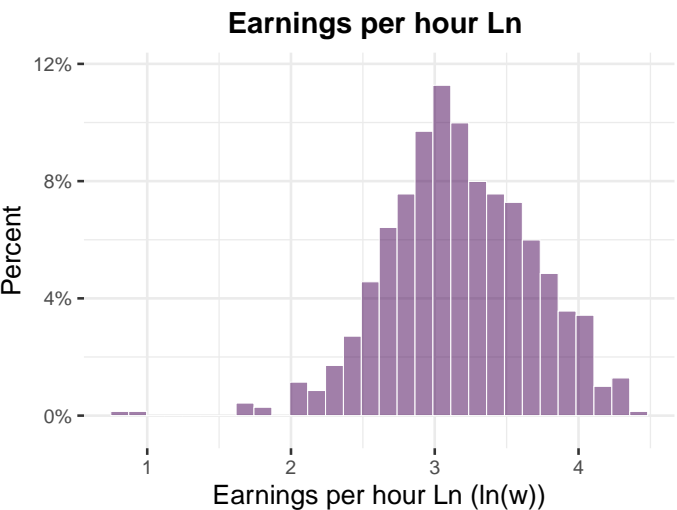
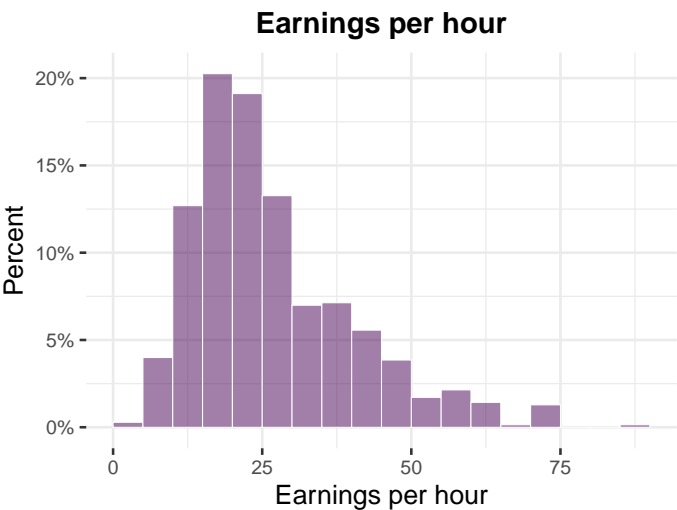
Descriptive Data Summary

Table 3: Descriptive Summary Statistics

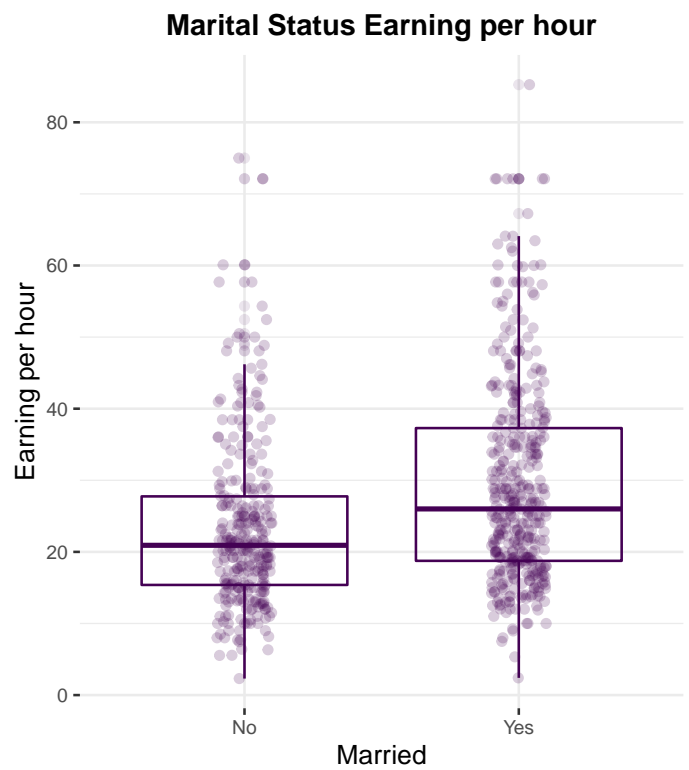
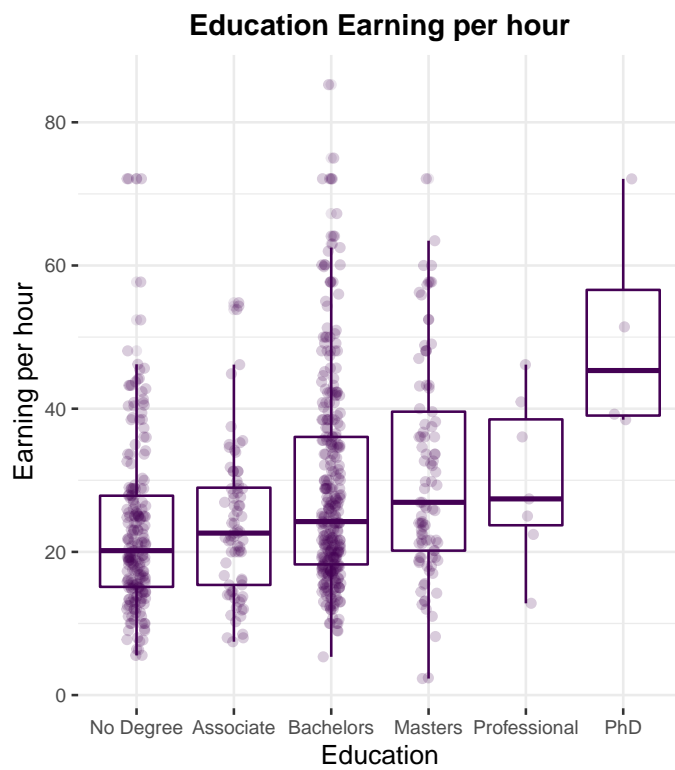
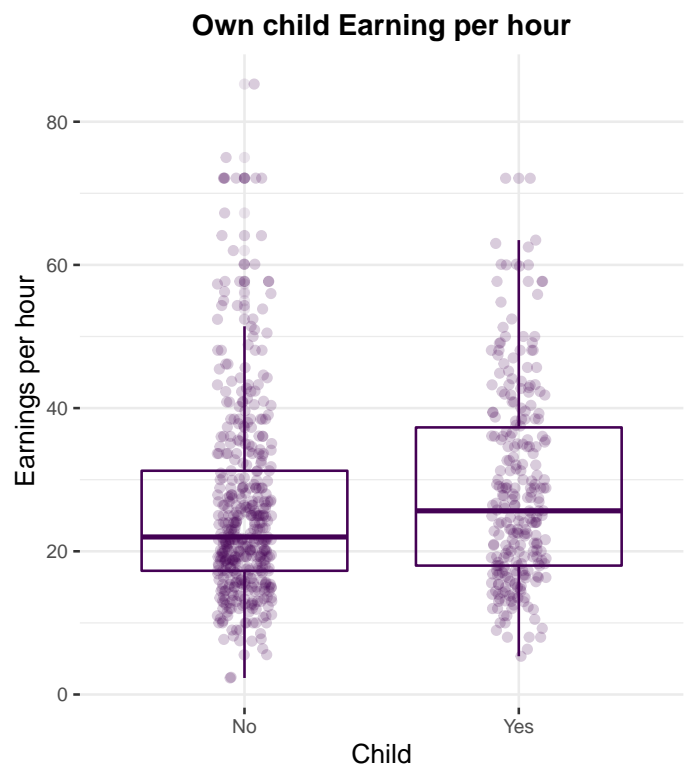
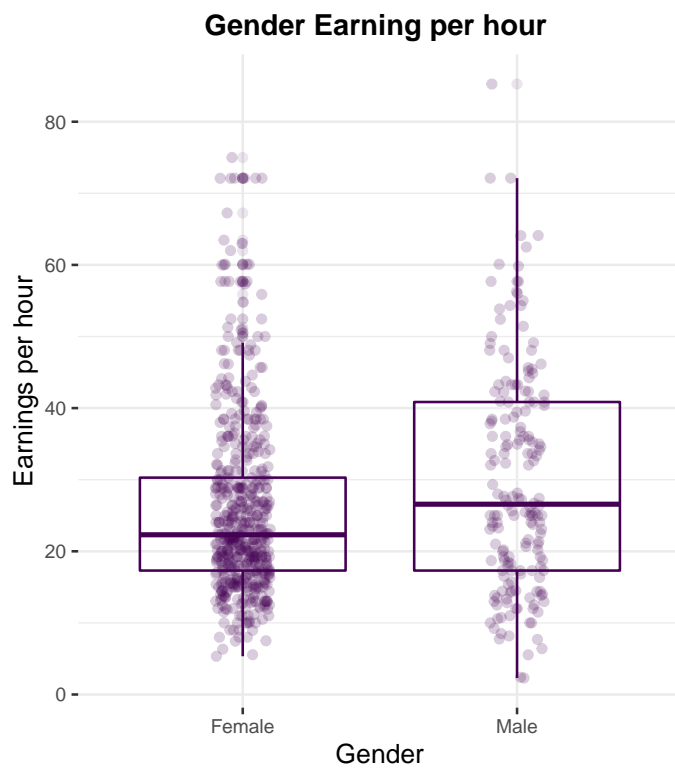
| | Median | Mean | SD | Min | Max | P05 | P95 | N |
|------------------------|--------|---------|--------|-------|---------|--------|---------|-----|
| Weekly earnings | 961.53 | 1111.08 | 605.24 | 75.00 | 2884.61 | 400.00 | 2394.00 | 701 |
| Weekly hours worked | 40.00 | 40.88 | 7.44 | 6.00 | 65.00 | 28.00 | 55.00 | 701 |
| Earning per hour | 23.50 | 26.92 | 13.62 | 2.31 | 85.26 | 11.00 | 55.88 | 701 |
| Female | 1.00 | 0.74 | 0.44 | 0.00 | 1.00 | 0.00 | 1.00 | 701 |
| No degree | 0.00 | 0.30 | 0.46 | 0.00 | 1.00 | 0.00 | 1.00 | 701 |
| Associate | 0.00 | 0.11 | 0.31 | 0.00 | 1.00 | 0.00 | 1.00 | 701 |
| BA Degree | 0.00 | 0.44 | 0.50 | 0.00 | 1.00 | 0.00 | 1.00 | 701 |
| MA Degree | 0.00 | 0.13 | 0.34 | 0.00 | 1.00 | 0.00 | 1.00 | 701 |
| Professional Degree | 0.00 | 0.01 | 0.10 | 0.00 | 1.00 | 0.00 | 0.00 | 701 |
| PhD | 0.00 | 0.01 | 0.08 | 0.00 | 1.00 | 0.00 | 0.00 | 701 |
| Age | 40.00 | 40.83 | 11.91 | 19.00 | 64.00 | 23.00 | 61.00 | 701 |
| Work in Private Sector | 1.00 | 0.79 | 0.41 | 0.00 | 1.00 | 0.00 | 1.00 | 701 |
| Has child | 0.00 | 0.37 | 0.48 | 0.00 | 1.00 | 0.00 | 1.00 | 701 |
| Marital Status | 1.00 | 0.57 | 0.49 | 0.00 | 1.00 | 0.00 | 1.00 | 701 |

Descriptive summary of the main variables in the data set can be in the above table. From the table it can be inferred that because of the presence of high hourly wage values like USD 2884, mean tends to be to the right of the median thus making the sample distribution rightly skewed. Similarly, we see that there are certain people who work more than 40 hours (maximum value of 65 hours a week) which is also the cause for skewness. Moreover, there is also the presence of extreme values. For example, the minimum wage value is computed out to be USD 2.31

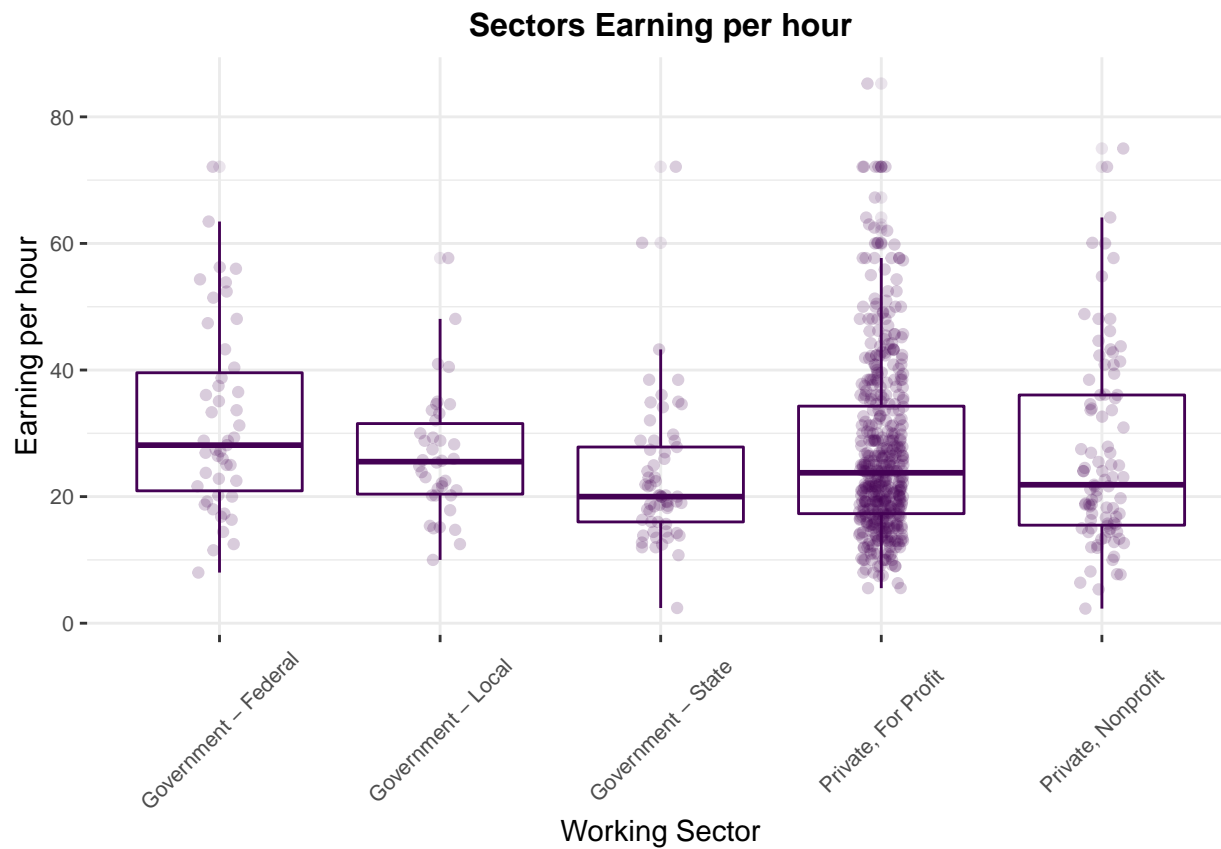
Histograms



One variable plots

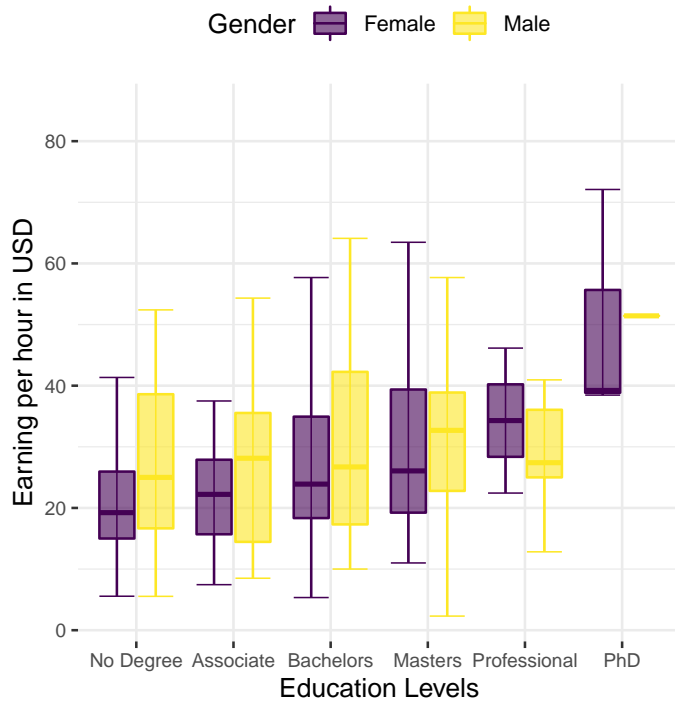


One variable plot

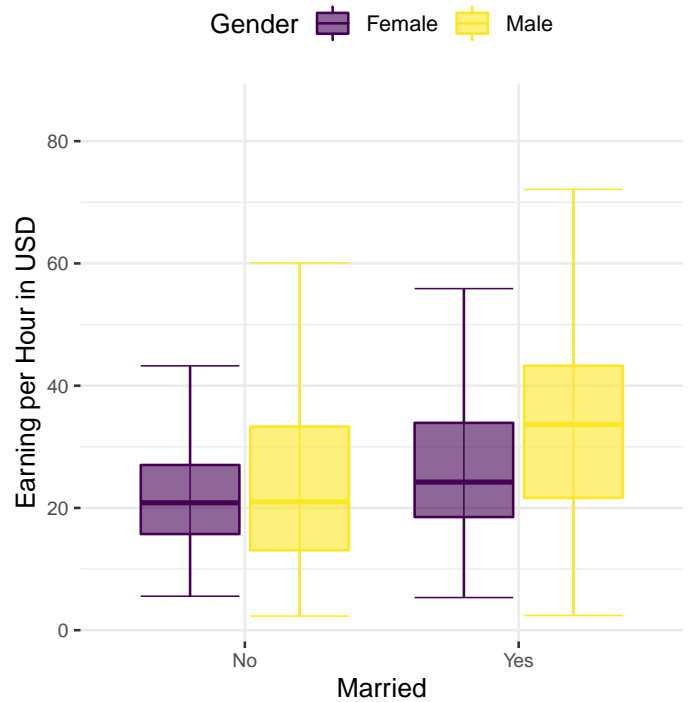


Two Variable plots

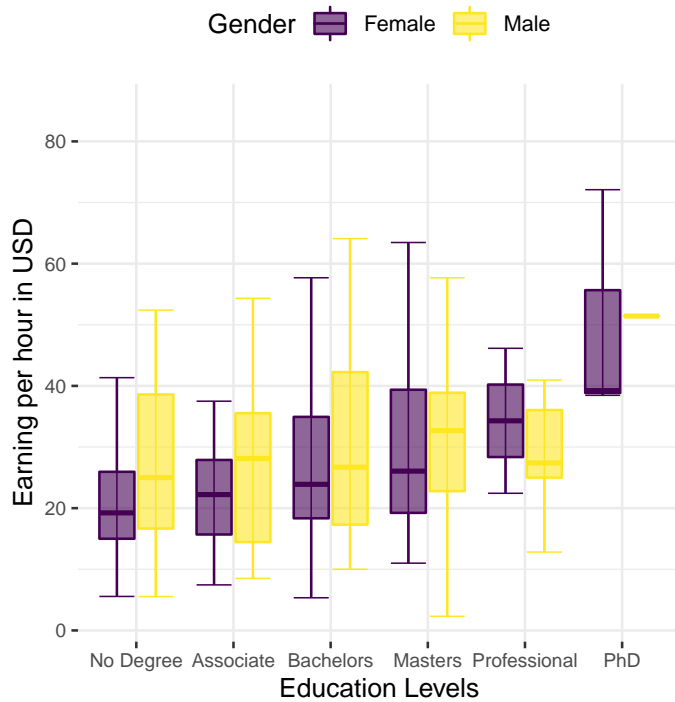
Gender and education levels Earning per hour



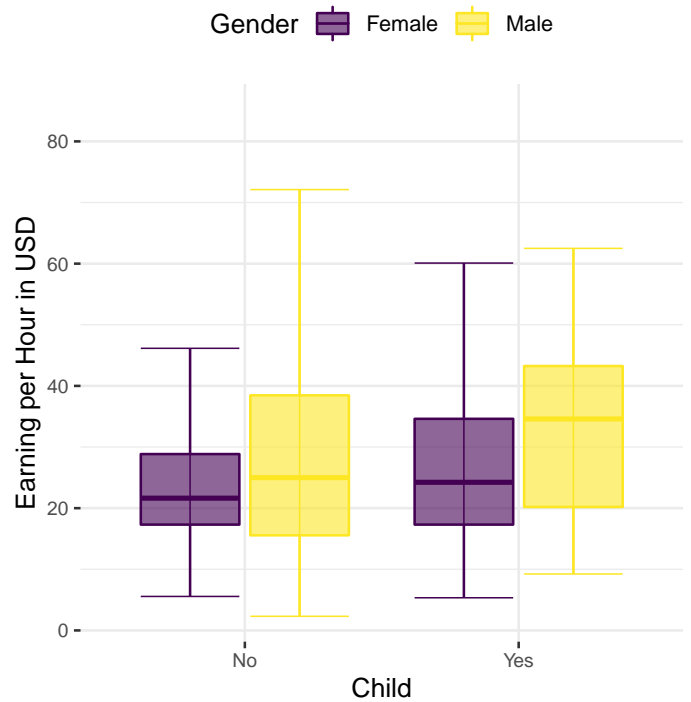
Gender and marital status Earning per hour



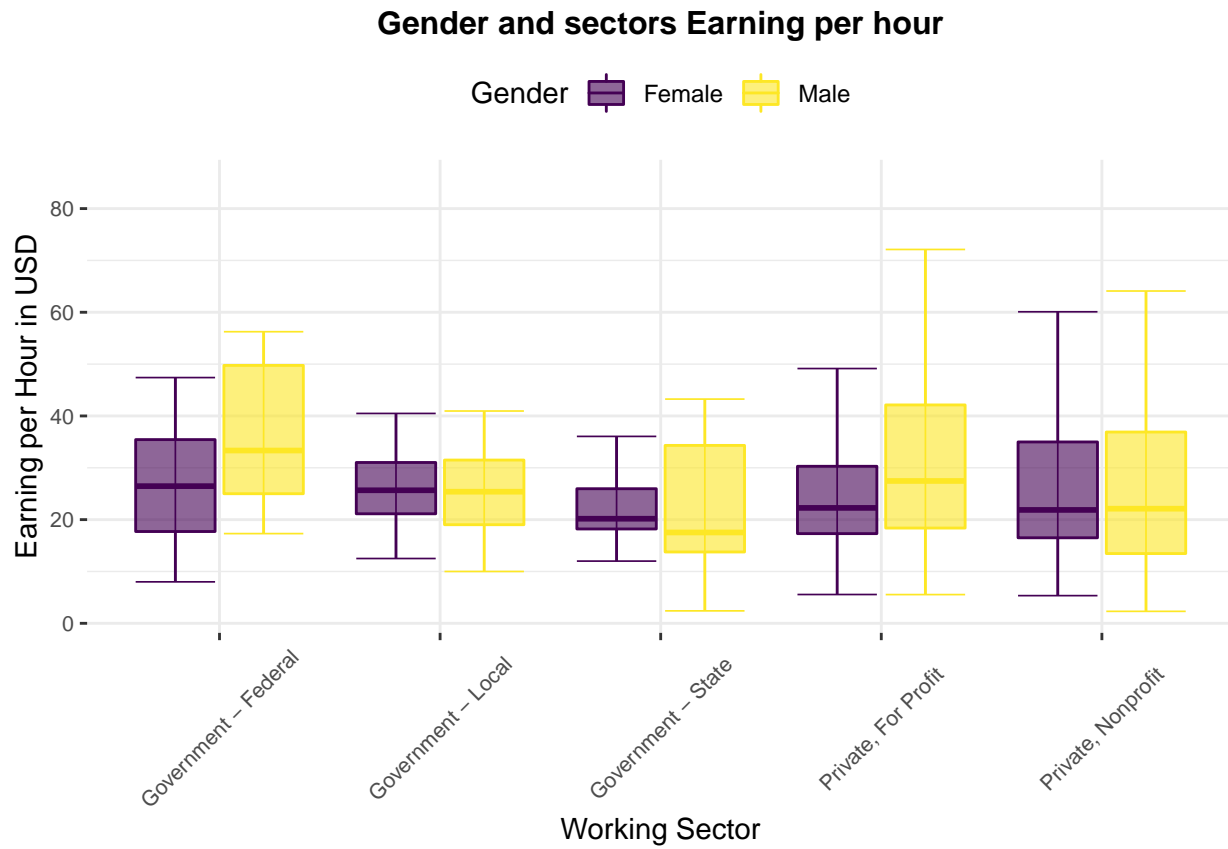
Marital status and education level Earning per hour



Gender and owning a child Earning per hour



Two Variable plot



Loess

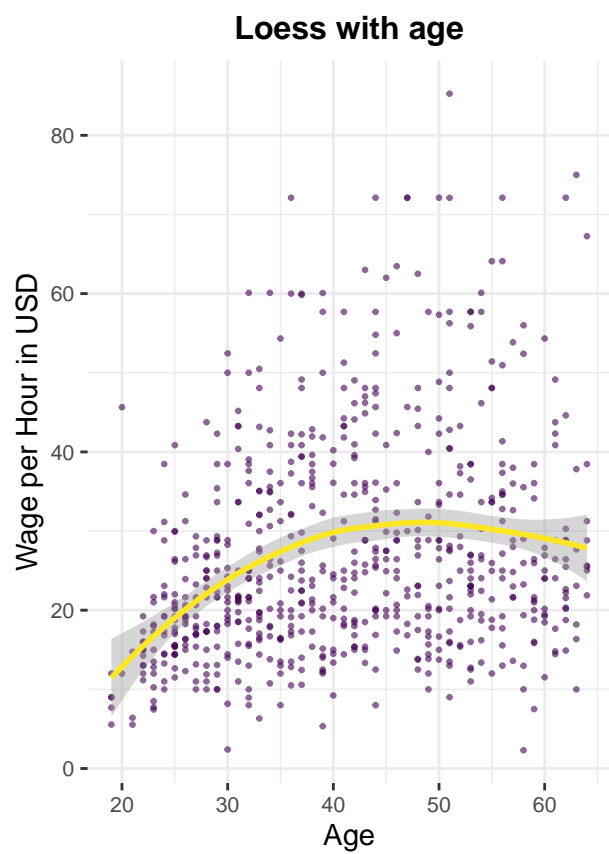
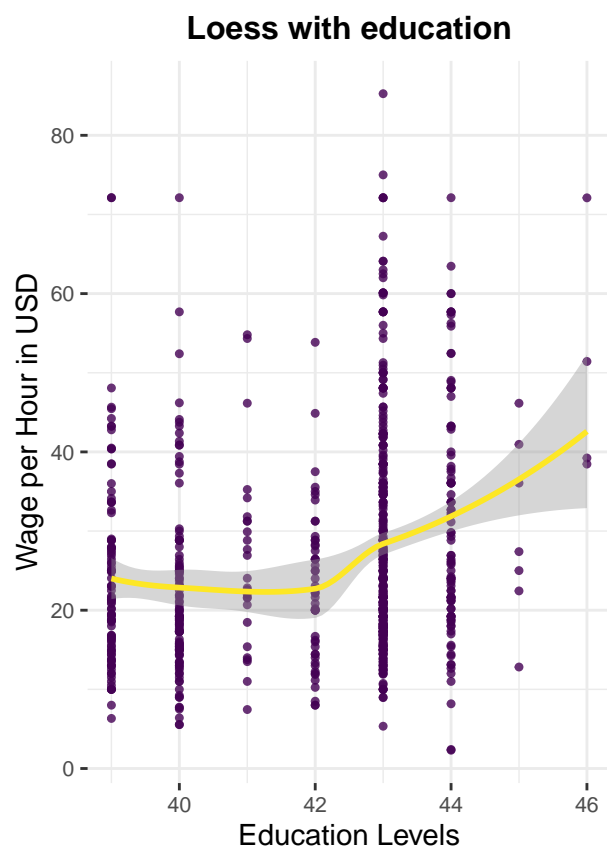


Table 4: Wage and education summary

| | Education | N | Percent | mean |
|---|--------------|-----|---------|-------|
| w | Associate | 76 | 10.84 | 23.91 |
| | Bachelors | 309 | 44.08 | 28.70 |
| | Masters | 93 | 13.27 | 30.92 |
| | No Degree | 212 | 30.24 | 23.10 |
| | PhD | 4 | 0.57 | 50.30 |
| | Professional | 7 | 1.00 | 30.12 |

Table 5: Wage and class summary

| | class | N | Percent | mean |
|---|----------------------|-----|---------|-------|
| w | Government - Federal | 47 | 6.70 | 32.00 |
| | Government - Local | 38 | 5.42 | 26.51 |
| | Government - State | 61 | 8.70 | 23.21 |
| | Private, For Profit | 464 | 66.19 | 26.95 |
| | Private, Nonprofit | 91 | 12.98 | 26.79 |

Table 6: Wage and Own a child summary

| | Own a child | N | Percent | mean |
|---|-------------|-----|---------|-------|
| w | 0 | 439 | 62.62 | 25.91 |
| | 1 | 262 | 37.38 | 28.61 |

Table 7: Wage and Gender summary

| | Gender | N | Percent | mean |
|---|--------|-----|---------|-------|
| w | Female | 521 | 74.32 | 25.86 |
| | Male | 180 | 25.68 | 29.98 |

Table 8: Wage, Gender and own a child summary

| | Gender | Own a child | N | Percent | mean |
|---|--------|-------------|-----|---------|-------|
| w | Female | 0 | 320 | 45.65 | 24.87 |
| | | 1 | 201 | 28.67 | 27.43 |
| | Male | 0 | 119 | 16.98 | 28.70 |
| | | 1 | 61 | 8.70 | 32.49 |

Table 9: Wage, education and Gender summary

| | Education | Gender | N | Percent | mean |
|---|--------------|--------|-----|---------|-------|
| w | Associate | Female | 59 | 8.42 | 22.39 |
| | | Male | 17 | 2.43 | 29.17 |
| | Bachelors | Female | 234 | 33.38 | 27.78 |
| | | Male | 75 | 10.70 | 31.59 |
| | Masters | Female | 66 | 9.42 | 31.00 |
| | | Male | 27 | 3.85 | 30.73 |
| | No Degree | Female | 157 | 22.40 | 21.58 |
| | | Male | 55 | 7.85 | 27.44 |
| | PhD | Female | 3 | 0.43 | 49.93 |
| | | Male | 1 | 0.14 | 51.42 |
| | Professional | Female | 2 | 0.29 | 34.29 |
| | | Male | 5 | 0.71 | 28.45 |

Table 10: Wage and Marital status summary

| | Married | N | Percent | mean |
|---|---------|-----|---------|-------|
| w | 0 | 298 | 42.51 | 23.49 |
| | 1 | 403 | 57.49 | 29.45 |

Table 11: Wage, education and marital status summary

| | Education | Married | N | Percent | mean |
|---|--------------|---------|-----|---------|-------|
| w | Associate | 0 | 30 | 4.28 | 19.69 |
| | | 1 | 46 | 6.56 | 26.66 |
| | Bachelors | 0 | 136 | 19.40 | 25.85 |
| | | 1 | 173 | 24.68 | 30.94 |
| | Masters | 0 | 29 | 4.14 | 24.63 |
| | | 1 | 64 | 9.13 | 33.78 |
| | No Degree | 0 | 100 | 14.27 | 21.05 |
| | | 1 | 112 | 15.98 | 24.93 |
| | PhD | 0 | 0 | 0.00 | |
| | | 1 | 4 | 0.57 | 50.30 |
| | Professional | 0 | 3 | 0.43 | 25.40 |
| | | 1 | 4 | 0.57 | 33.65 |

Table 12: Wage, gender and marital status summary

| | Gender | Married | N | Percent | mean |
|---|--------|---------|-----|---------|-------|
| w | Female | 0 | 228 | 32.52 | 23.35 |
| | | 1 | 293 | 41.80 | 27.81 |
| | Male | 0 | 70 | 9.99 | 23.95 |
| | | 1 | 110 | 15.69 | 33.82 |