Projekt US Bank Wages

Is there a wage gap between genders?

Farzan Mirzada

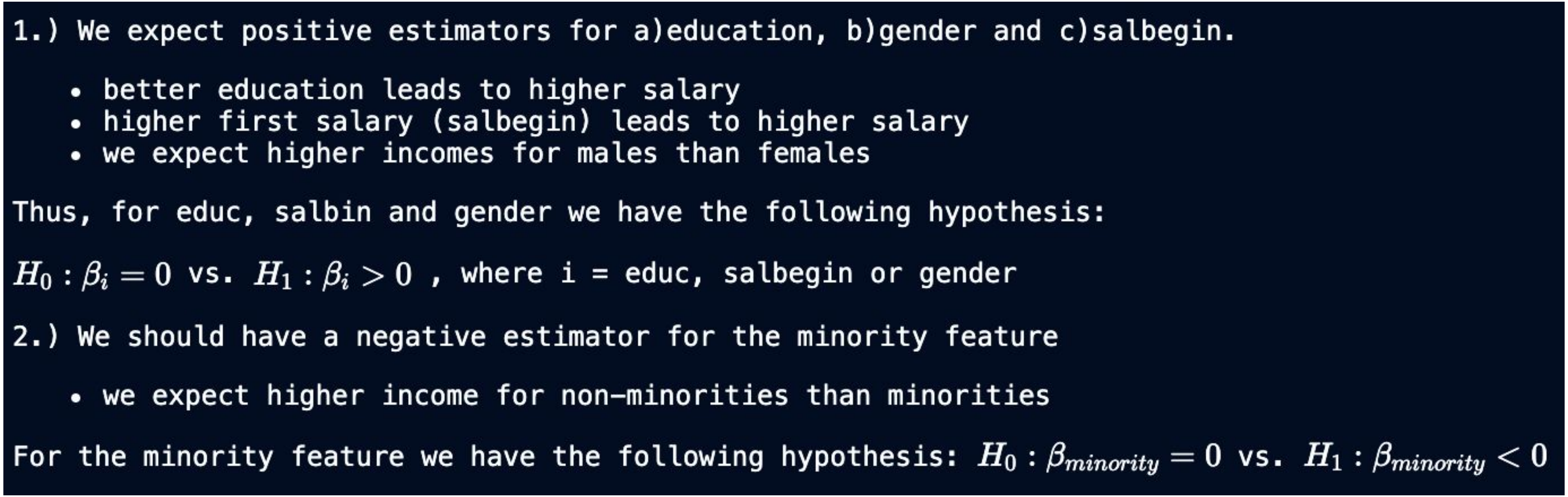
Background / Motivation

- Policymakers are thinking about implementing new regulations for the finance sector.
- One policymaker claims that the finance sector is male-dominated and needs to close the wage gap between males and females. He also believes that minorities get payed less.
- Now he needs quantitative evidence to persuade other policymakers.

Data

- We analyse a sample from a us bank with 474 observations.
- The response/target variable is yearly salary in $
- Features = education, yearly salary in first year, gender(dummy), minority (dummy) and the jobcatagory of the employee (administrative position, custodial position, management position)

Hypothesis



1.) We expect positive estimators for a)education, b)gender and c)salbegin.

- better education leads to higher salary
- higher first salary (salbegin) leads to higher salary
- we expect higher incomes for males than females

Thus, for educ, salbin and gender we have the following hypothesis:

$H_0 : \beta_i = 0$ vs. $H_1 : \beta_i > 0$, where i = educ, salbegin or gender

2.) We should have a negative estimator for the minority feature

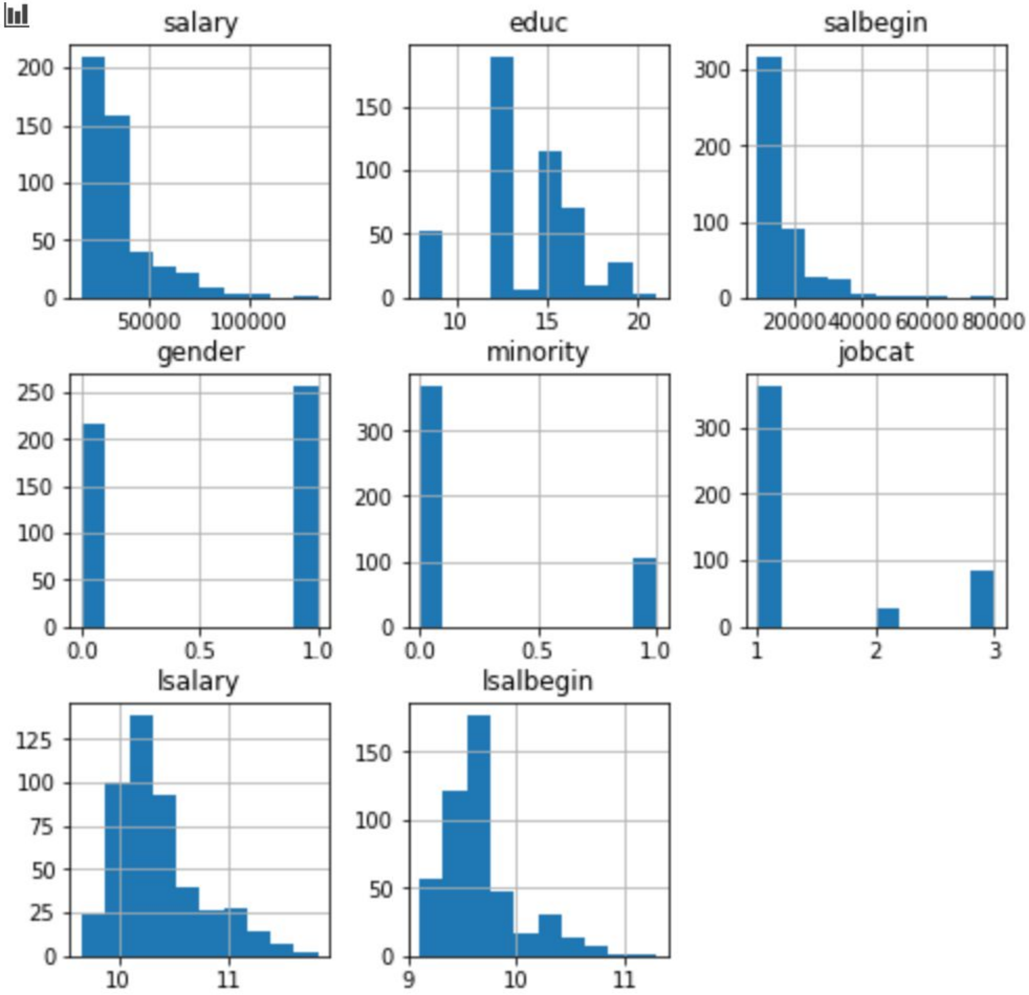
- we expect higher income for non-minorities than minorities

For the minority feature we have the following hypothesis: $H_0 : \beta_{minority} = 0$ vs. $H_1 : \beta_{minority} < 0$

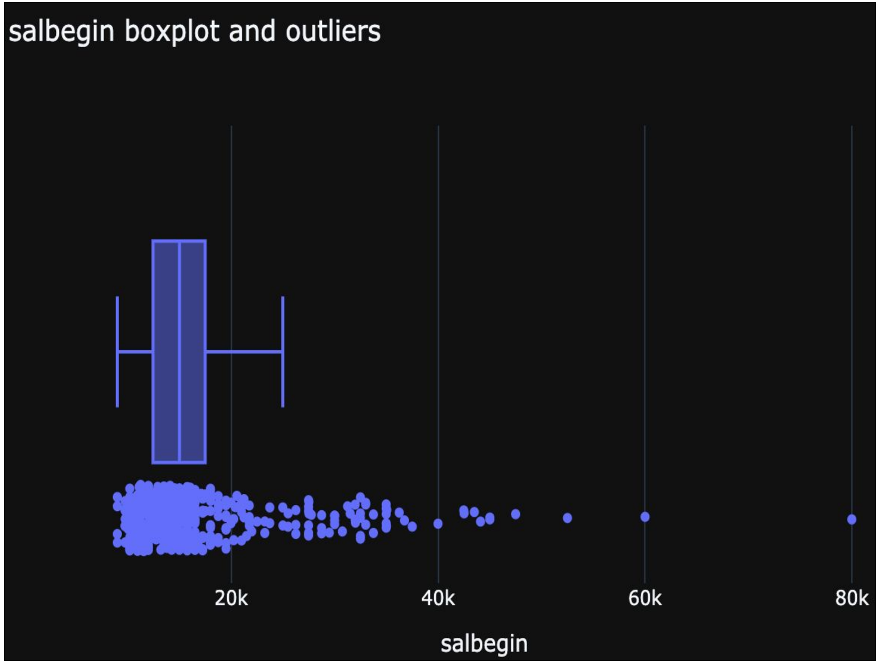
Visualisation

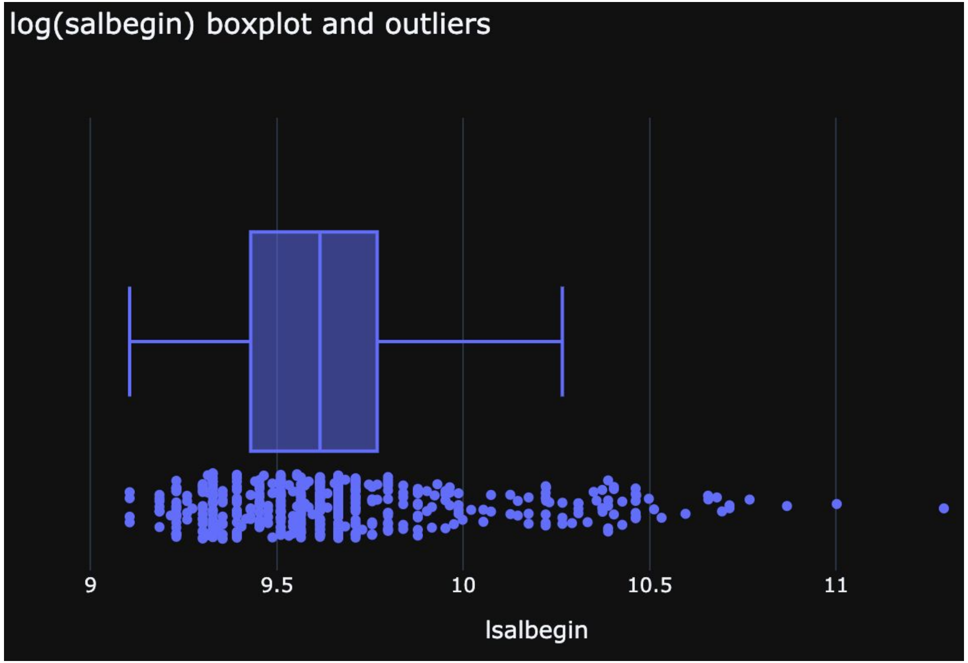
Histograms

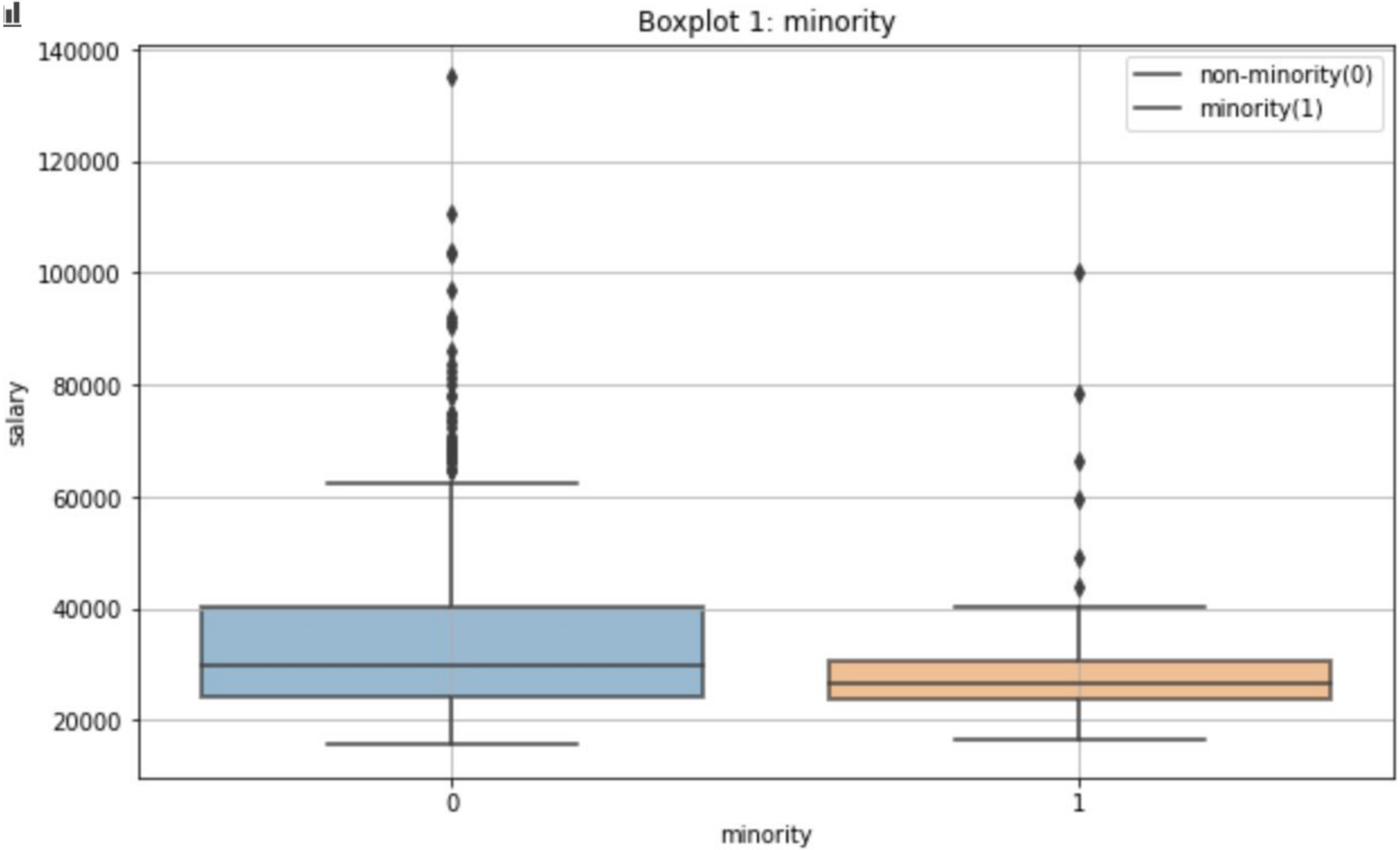
- right-taled distributions for salary and salbegin.
- hard to assume normal distribution for any feature.
- salbegin might have a non-linear relation with salary.
- **Better use natural logarithm for salary values.**

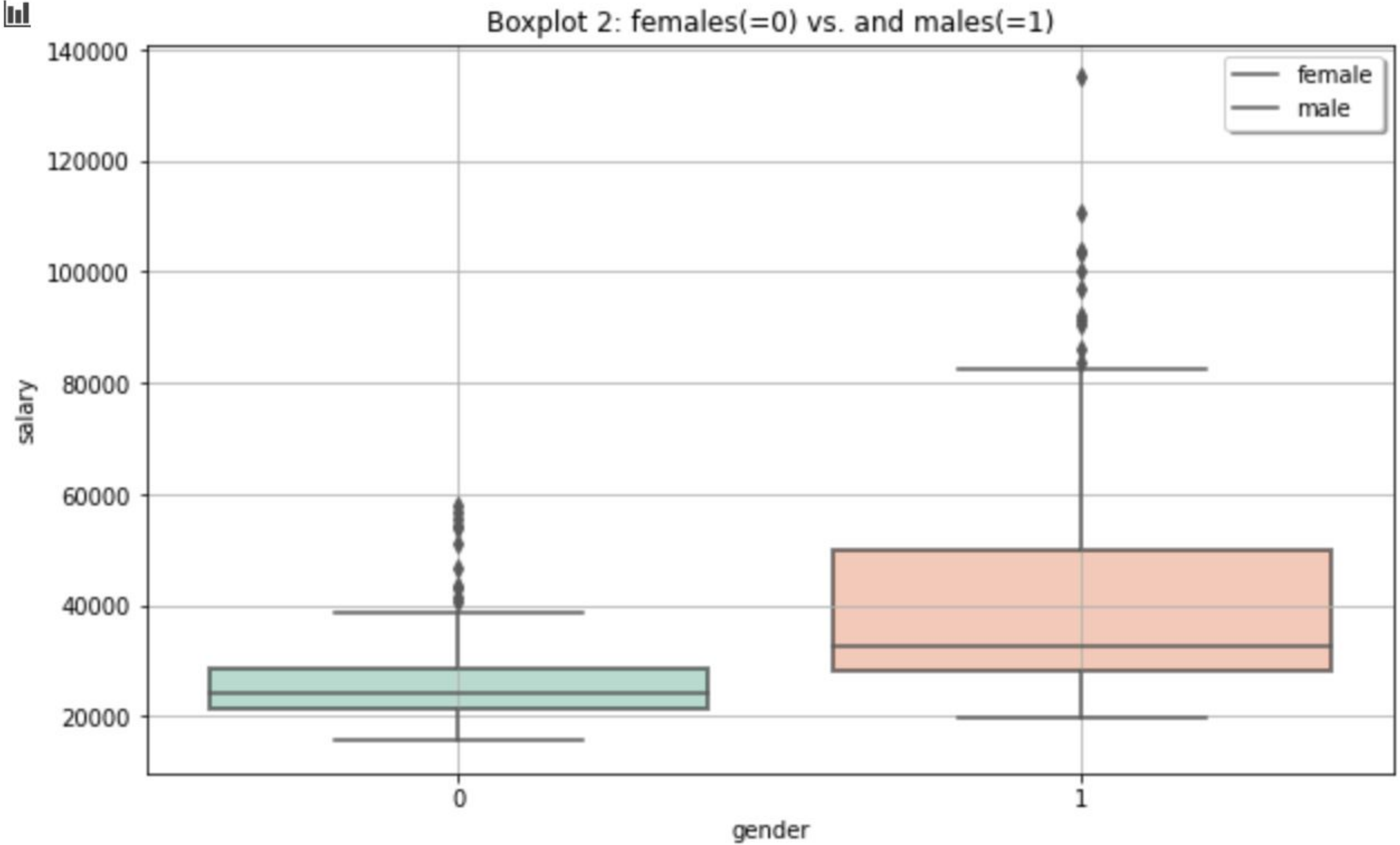


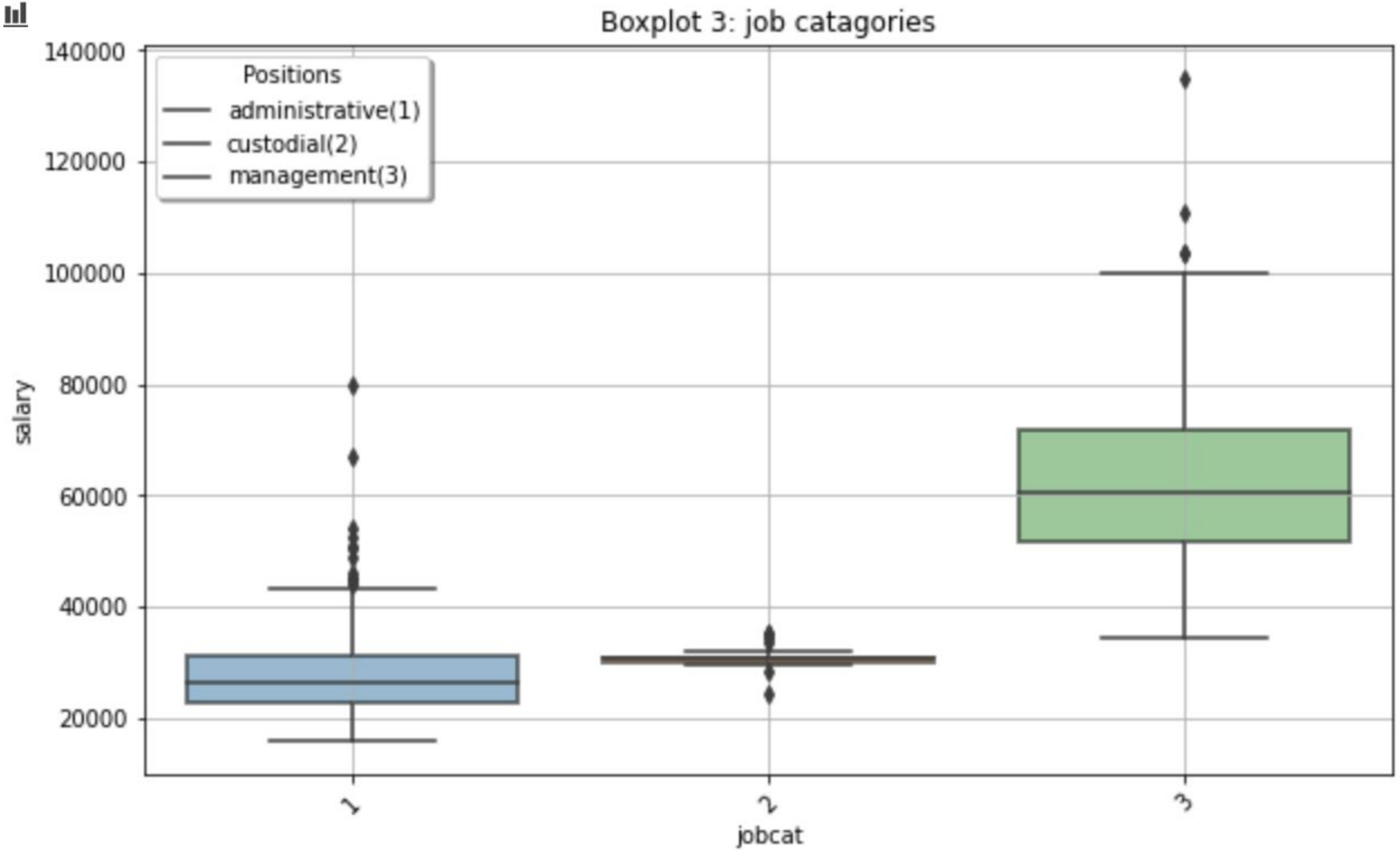
log reduces outliers






Boxplot 1: minority
non-minority(0)
minority(1)
salary
140000
120000
100000
80000
60000
40000
20000
0
1
minority

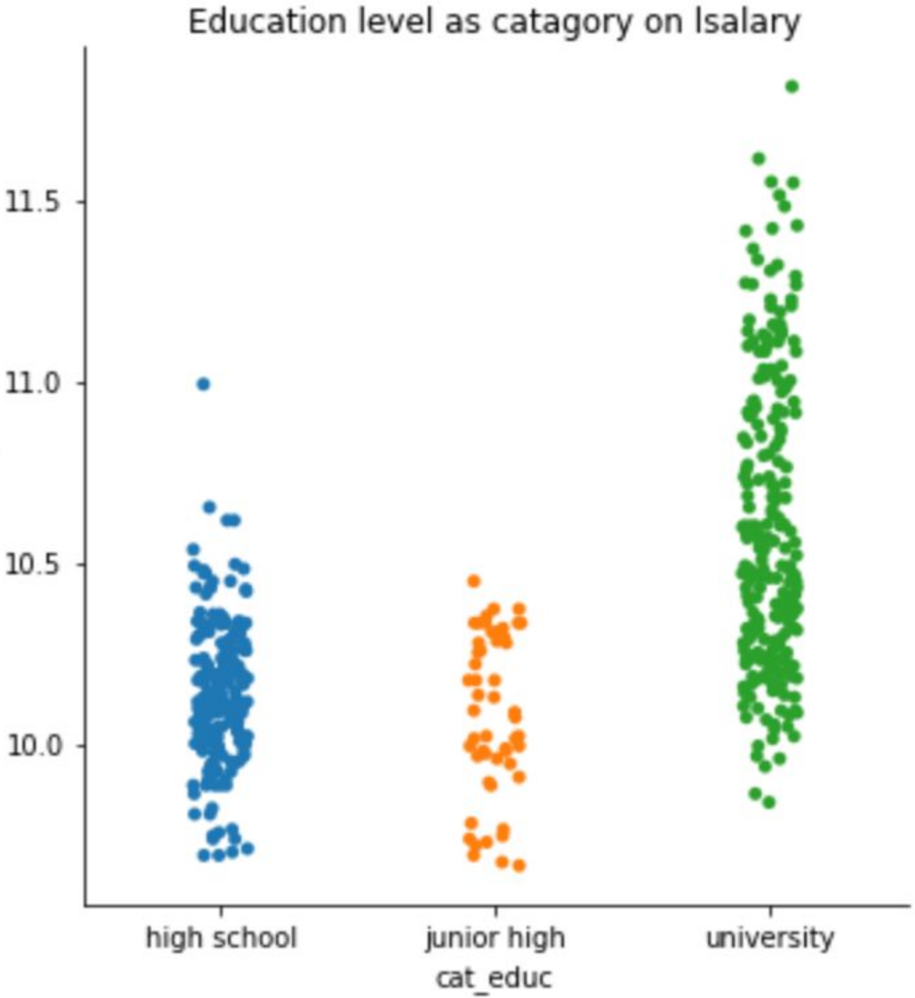

Boxplot 2: females(=0) vs. and males(=1)
female
male
salary
140000
120000
100000
80000
60000
40000
20000
0
1
gender

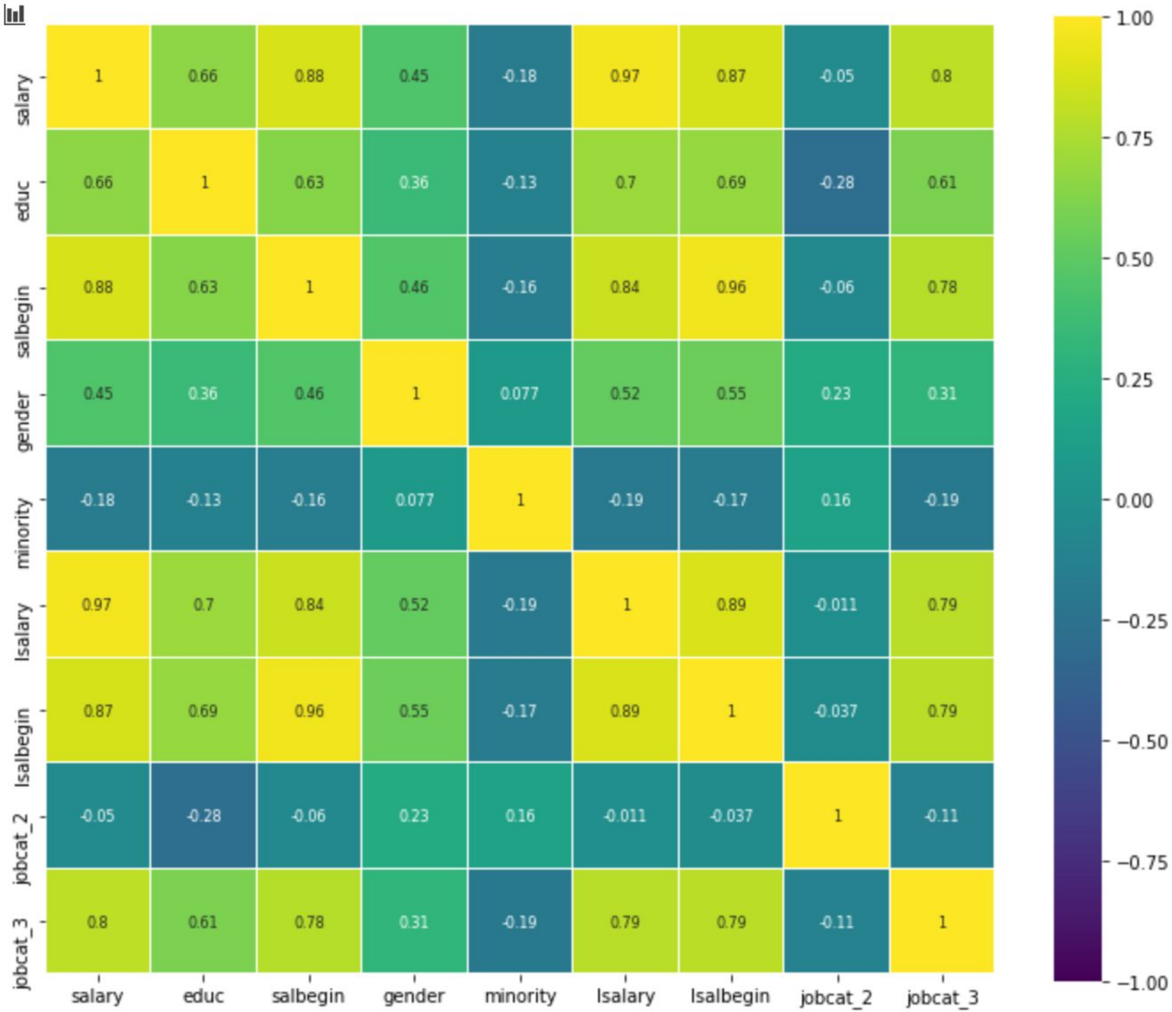


Boxplot 3: job catagories

Categorizing Education

- We can split education values into degree categories
- A university degree seems to have a higher impact on the salary than a high school or junior high degree.





	salary	educ	salbegin	gender	minority	lsalary	lsalbegin	jobcat_2	jobcat_3
salary	1	0.66	0.88	0.45	-0.18	0.97	0.87	-0.05	0.8
educ	0.66	1	0.63	0.36	-0.13	0.7	0.69	-0.28	0.61
salbegin	0.88	0.63	1	0.46	-0.16	0.84	0.96	-0.06	0.78
gender	0.45	0.36	0.46	1	0.077	0.52	0.55	0.23	0.31
minority	-0.18	-0.13	-0.16	0.077	1	-0.19	-0.17	0.16	-0.19
lsalary	0.97	0.7	0.84	0.52	-0.19	1	0.89	-0.011	0.79
lsalbegin	0.87	0.69	0.96	0.55	-0.17	0.89	1	-0.037	0.79
jobcat_2	-0.05	-0.28	-0.06	0.23	0.16	-0.011	-0.037	1	-0.11
jobcat_3	0.8	0.61	0.78	0.31	-0.19	0.79	0.79	-0.11	1

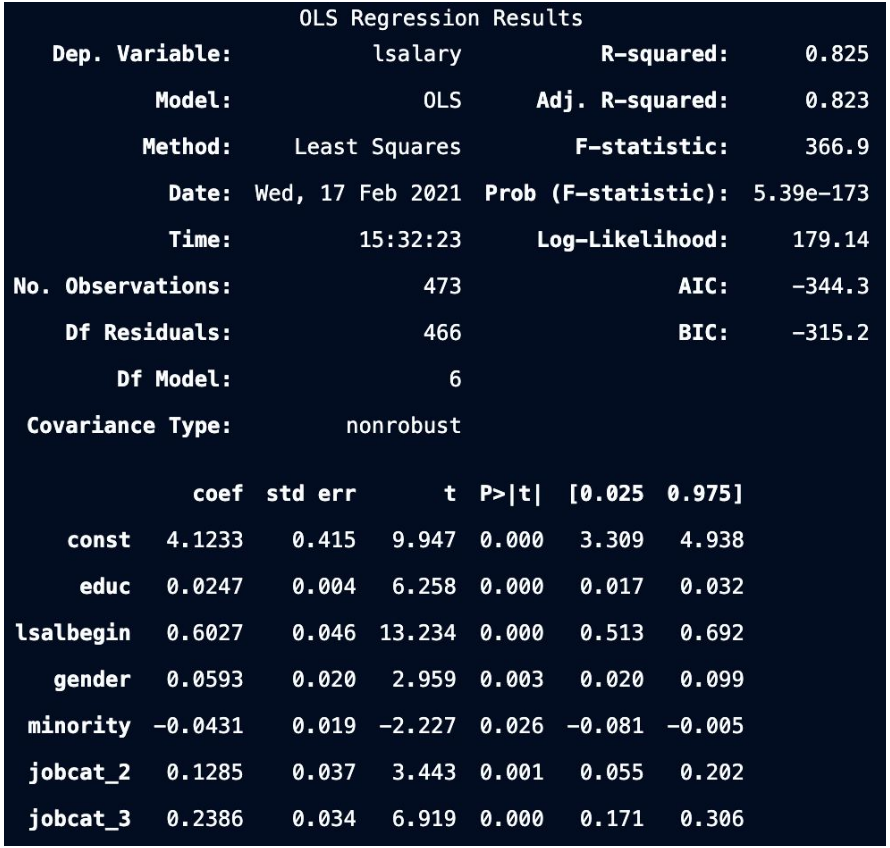
Correlation Matrix

lsalary row is crucial for us:

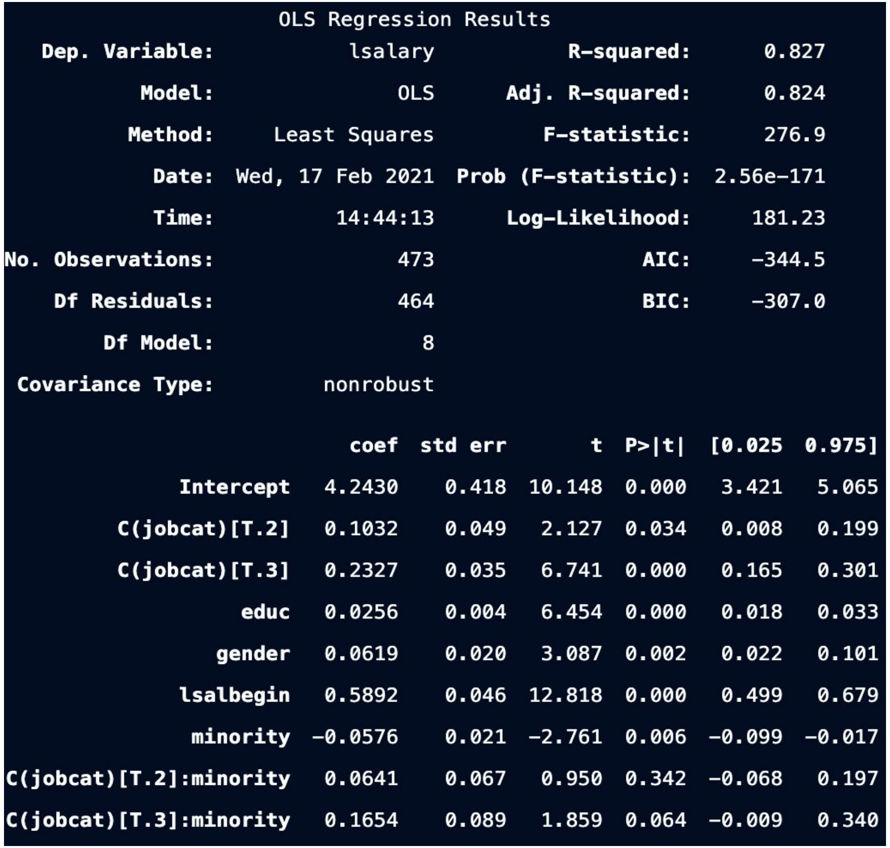
- minority and employees with custodial position have a negative correlation with log(salary)
- educ, log(salbegin) and management position is positively correlated with log(salary)
- support for hypothesis 1) and 2)

Linear Regression

Summary statistics using OLS

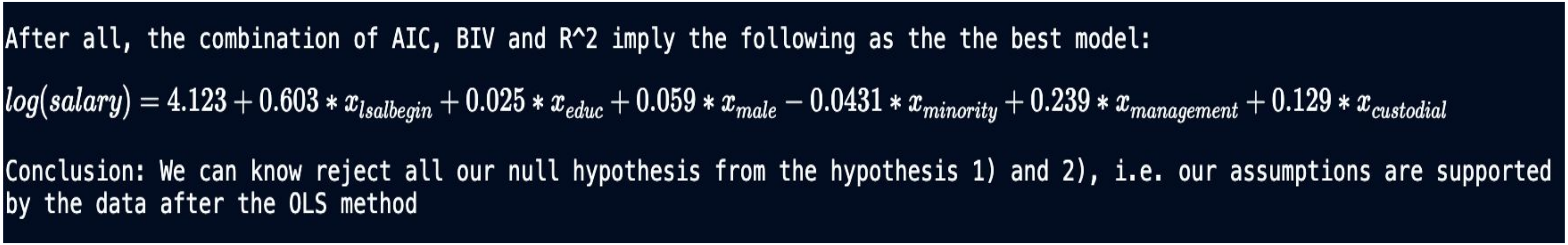


```
                     OLS Regression Results
   Dep. Variable:            lsalary          R-squared:      0.825
           Model:                OLS     Adj. R-squared:      0.823
          Method:      Least Squares         F-statistic:      366.9
            Date:   Wed, 17 Feb 2021 Prob (F-statistic):  5.39e-173
            Time:           15:32:23      Log-Likelihood:     179.14
No. Observations:                473                 AIC:     -344.3
    Df Residuals:                466                 BIC:     -315.2
        Df Model:                  6
 Covariance Type:          nonrobust

               coef  std err       t  P>|t|  [0.025  0.975]
    const    4.1233    0.415   9.947  0.000   3.309   4.938
     educ    0.0247    0.004   6.258  0.000   0.017   0.032
lsalbegin    0.6027    0.046  13.234  0.000   0.513   0.692
   gender    0.0593    0.020   2.959  0.003   0.020   0.099
 minority   -0.0431    0.019  -2.227  0.026  -0.081  -0.005
 jobcat_2    0.1285    0.037   3.443  0.001   0.055   0.202
 jobcat_3    0.2386    0.034   6.919  0.000   0.171   0.306
```



```
                        OLS Regression Results
   Dep. Variable:          lsalary          R-squared:     0.827
           Model:              OLS     Adj. R-squared:     0.824
          Method:    Least Squares         F-statistic:     276.9
            Date: Wed, 17 Feb 2021 Prob (F-statistic): 2.56e-171
            Time:         14:44:13      Log-Likelihood:    181.23
No. Observations:              473                 AIC:    -344.5
    Df Residuals:              464                 BIC:    -307.0
        Df Model:                8
 Covariance Type:        nonrobust

                              coef  std err       t  P>|t|  [0.025  0.975]
                 Intercept  4.2430    0.418  10.148  0.000   3.421   5.065
             C(jobcat)[T.2] 0.1032    0.049   2.127  0.034   0.008   0.199
             C(jobcat)[T.3] 0.2327    0.035   6.741  0.000   0.165   0.301
                      educ  0.0256    0.004   6.454  0.000   0.018   0.033
                    gender  0.0619    0.020   3.087  0.002   0.022   0.101
                 lsalbegin  0.5892    0.046  12.818  0.000   0.499   0.679
                  minority -0.0576    0.021  -2.761  0.006  -0.099  -0.017
C(jobcat)[T.2]:minority     0.0641    0.067   0.950  0.342  -0.068   0.197
C(jobcat)[T.3]:minority     0.1654    0.089   1.859  0.064  -0.009   0.340
```

Inference

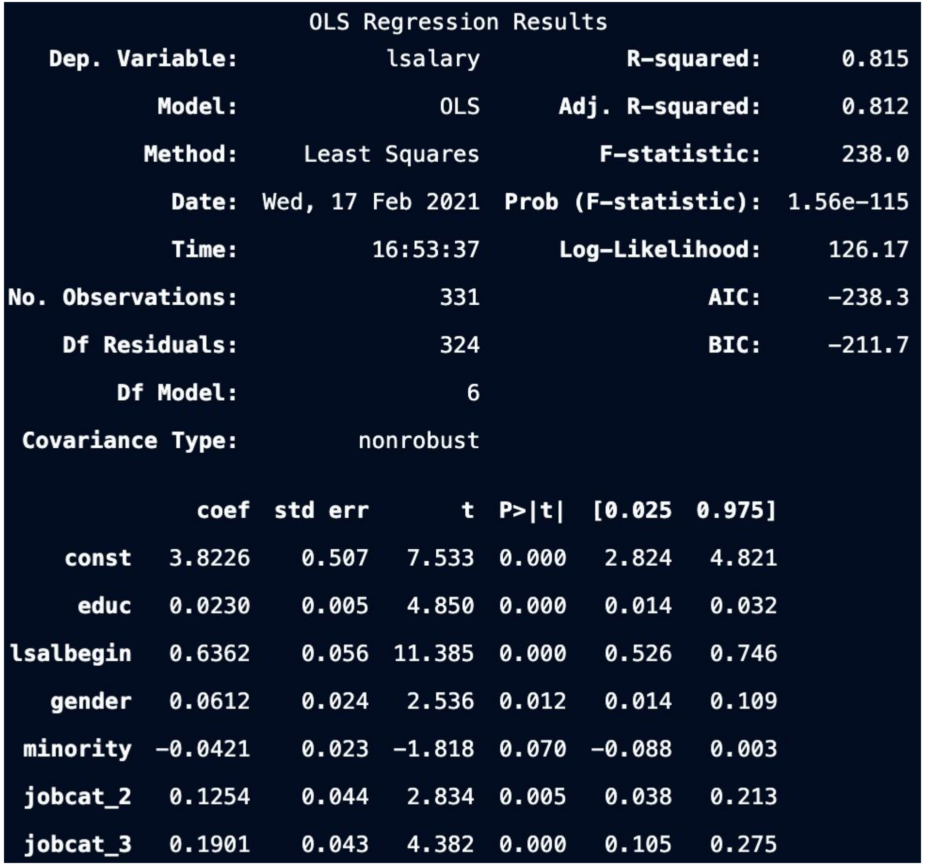


After all, the combination of AIC, BIV and R^2 imply the following as the the best model:

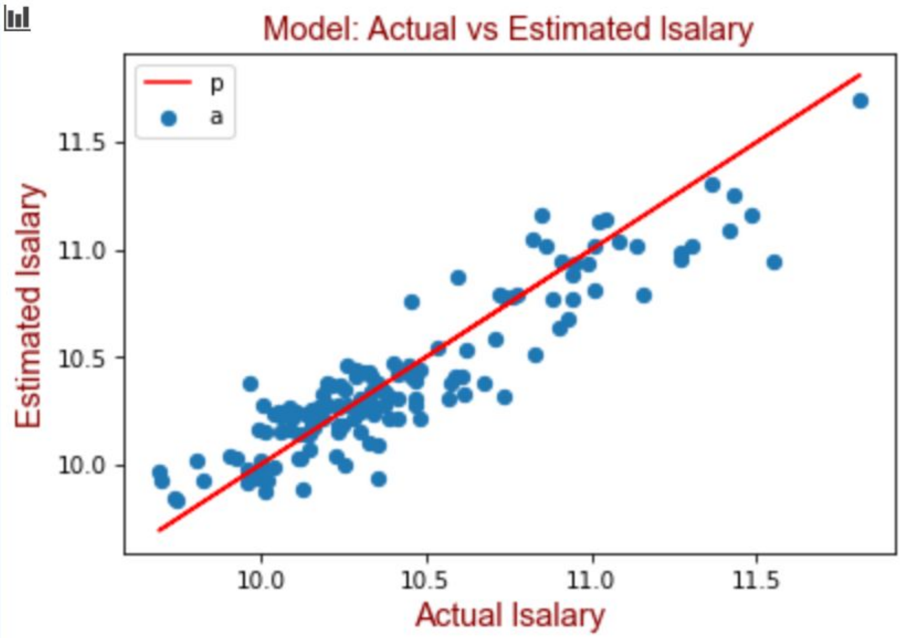
$$log(salary) = 4.123 + 0.603 * x_{lsalbegin} + 0.025 * x_{educ} + 0.059 * x_{male} - 0.0431 * x_{minority} + 0.239 * x_{management} + 0.129 * x_{custodial}$$

Conclusion: We can know reject all our null hypothesis from the hypothesis 1) and 2), i.e. our assumptions are supported by the data after the OLS method

Forecasting Model 1

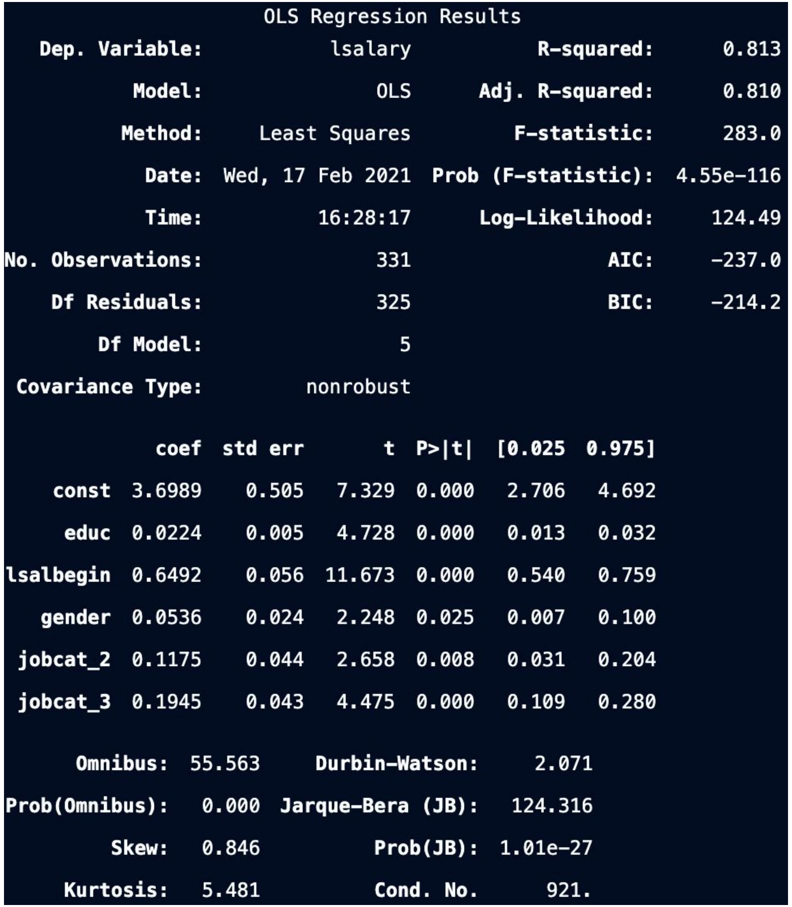


```
                    OLS Regression Results
   Dep. Variable:           lsalary         R-squared:      0.815
           Model:               OLS     Adj. R-squared:      0.812
          Method:     Least Squares        F-statistic:      238.0
            Date:  Wed, 17 Feb 2021  Prob (F-statistic):  1.56e-115
            Time:          16:53:37     Log-Likelihood:     126.17
No. Observations:               331                AIC:     -238.3
    Df Residuals:               324                BIC:     -211.7
        Df Model:                 6
 Covariance Type:         nonrobust

              coef  std err        t  P>|t|  [0.025  0.975]
    const   3.8226    0.507    7.533  0.000   2.824   4.821
     educ   0.0230    0.005    4.850  0.000   0.014   0.032
lsalbegin   0.6362    0.056   11.385  0.000   0.526   0.746
   gender   0.0612    0.024    2.536  0.012   0.014   0.109
 minority  -0.0421    0.023   -1.818  0.070  -0.088   0.003
 jobcat_2   0.1254    0.044    2.834  0.005   0.038   0.213
 jobcat_3   0.1901    0.043    4.382  0.000   0.105   0.275
```

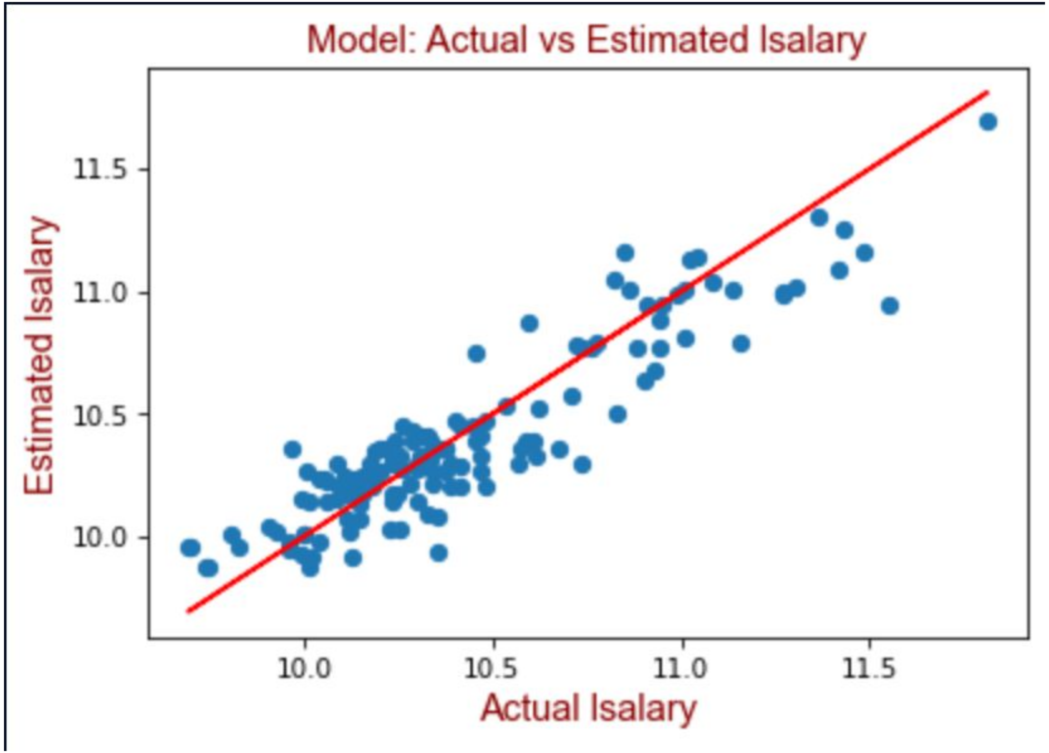


holding other var constant, males earn about 6,12 % more/ …. minorities earn about 4,21 % less

Forecasting Model 2 (minority is dropped)



```
                          OLS Regression Results
  Dep. Variable:           lsalary          R-squared:      0.813
         Model:                OLS     Adj. R-squared:      0.810
        Method:      Least Squares        F-statistic:      283.0
          Date:   Wed, 17 Feb 2021  Prob (F-statistic):  4.55e-116
          Time:           16:28:17     Log-Likelihood:     124.49
No. Observations:              331                AIC:     -237.0
  Df Residuals:                325                BIC:     -214.2
      Df Model:                  5
 Covariance Type:        nonrobust

           coef  std err       t  P>|t|  [0.025  0.975]
    const 3.6989   0.505   7.329  0.000   2.706   4.692
     educ 0.0224   0.005   4.728  0.000   0.013   0.032
lsalbegin 0.6492   0.056  11.673  0.000   0.540   0.759
   gender 0.0536   0.024   2.248  0.025   0.007   0.100
 jobcat_2 0.1175   0.044   2.658  0.008   0.031   0.204
 jobcat_3 0.1945   0.043   4.475  0.000   0.109   0.280

      Omnibus:  55.563      Durbin-Watson:    2.071
Prob(Omnibus):   0.000  Jarque-Bera (JB):   124.316
         Skew:   0.846           Prob(JB):  1.01e-27
     Kurtosis:   5.481           Cond. No.      921.
```



Model: Actual vs Estimated lsalary

holding other var constant, males earn about 5,3% more

Inference

- Model is better according to the RMSE
  - Root Mean Squared Error Model 1(RMSE) : 0.16899098463207346
  - Root Mean Squared Error Model 2(RMSE) : 0.16993043057879956
- AIC is smaller in model 1
- R-squared is larger in model 1