Exploratory Data Analysis - Gender Inequality

Data: US Bank Wages

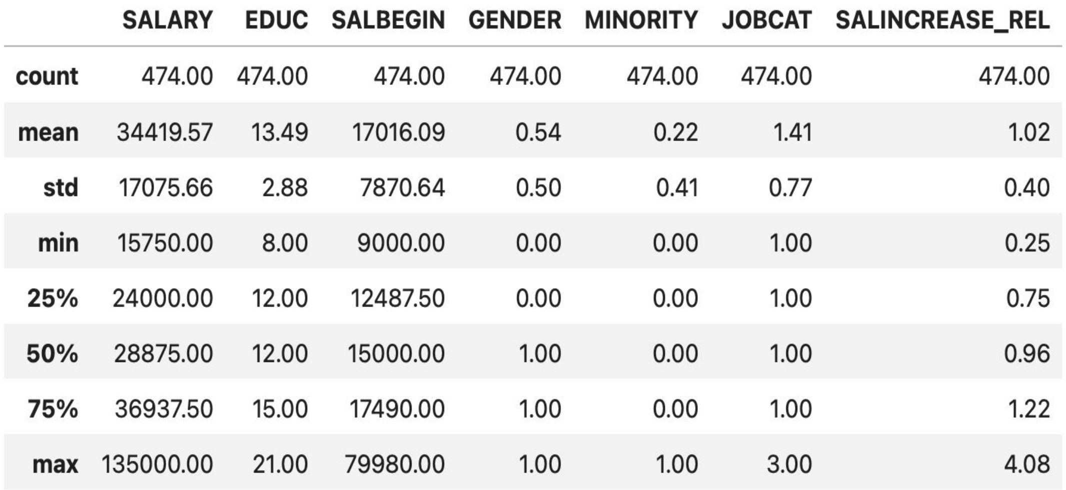
Goals

- Perform iterative EDA
- Predict target variable as accurately as possible
- Highlight interesting insights from EDA
- Present findings to non-technical audience

Data Preparation

- Dataset containing information about 457 employees of an US Bank
- Information on:
  - Current Salary: Annual salary in US-$
  - Education: Duration of (successful) education in years
  - Salary Begin: Annual salary at beginning of employment
  - Gender: Observations’ gender (0=female; 1=male)
  - Minority: Observations’ minority status (0=non-minority; 1=minority)
  - Job Category: Category of employment (1=administration; 2=custodial; 3=management)
- No missing values

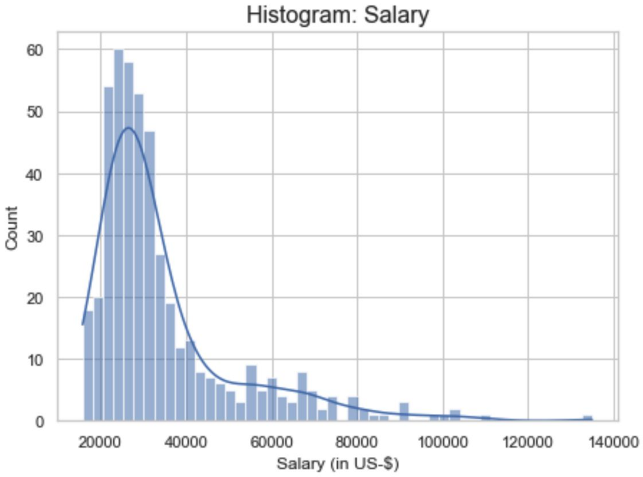
Univariate Analysis - Descriptive Statistics

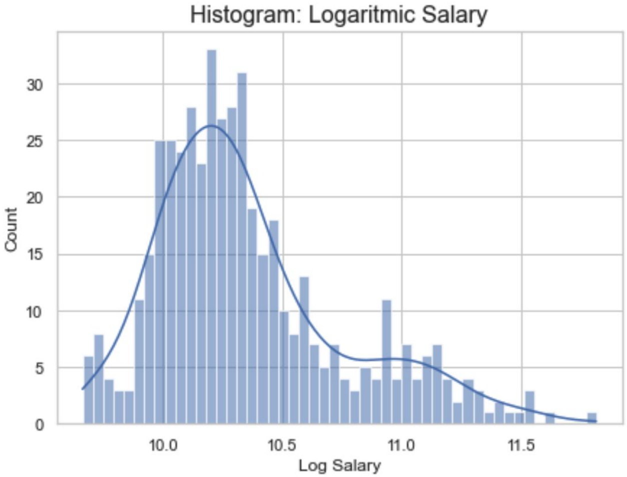


	SALARY	EDUC	SALBEGIN	GENDER	MINORITY	JOBCAT	SALINCREASE_REL
count	474.00	474.00	474.00	474.00	474.00	474.00	474.00
mean	34419.57	13.49	17016.09	0.54	0.22	1.41	1.02
std	17075.66	2.88	7870.64	0.50	0.41	0.77	0.40
min	15750.00	8.00	9000.00	0.00	0.00	1.00	0.25
25%	24000.00	12.00	12487.50	0.00	0.00	1.00	0.75
50%	28875.00	12.00	15000.00	1.00	0.00	1.00	0.96
75%	36937.50	15.00	17490.00	1.00	0.00	1.00	1.22
max	135000.00	21.00	79980.00	1.00	1.00	3.00	4.08

- SALINCREASE_REL = (SALARY - SALBEGIN) / SALBEGIN
- It is the relative increase in salary since first employment
- **Makes out-of-sample predictions impossible**

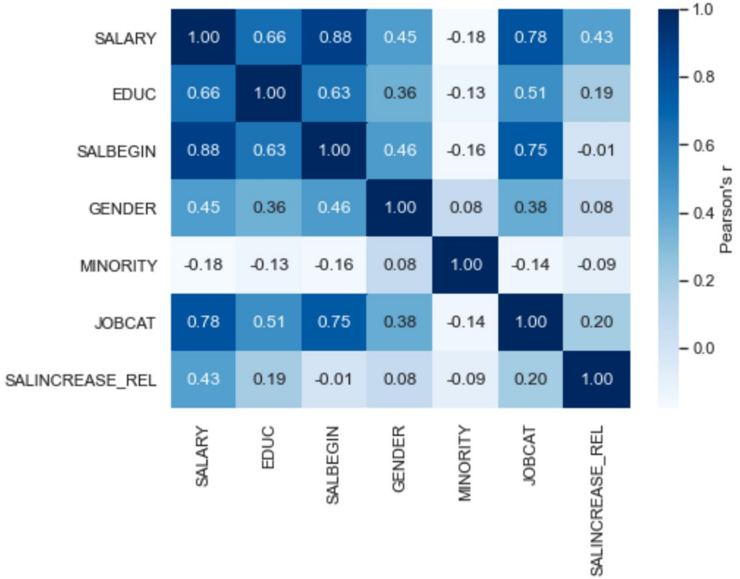
Univariate Analysis - Salary





- Log-transformation reduces the influence of outliers by “drawing them in”

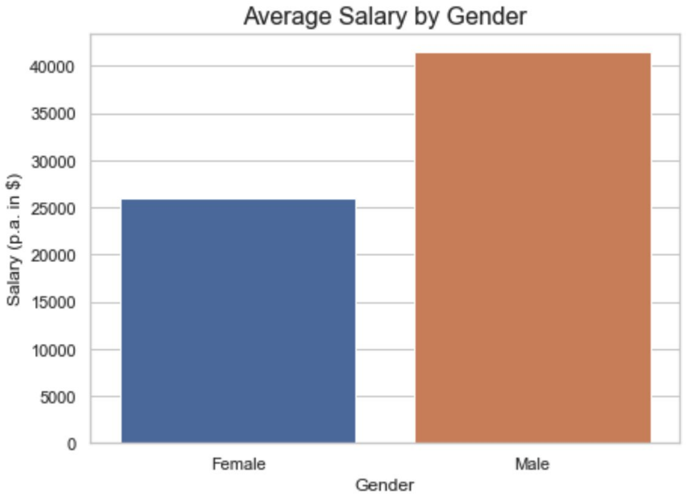
Bivariate Analysis - Heatmap

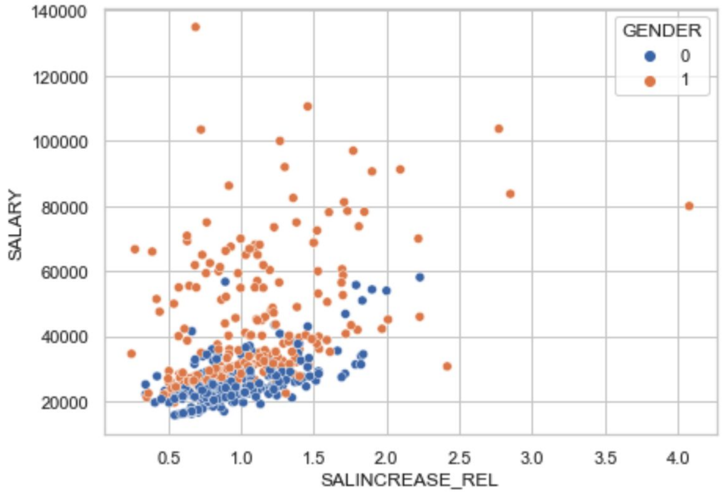


	SALARY	EDUC	SALBEGIN	GENDER	MINORITY	JOBCAT	SALINCREASE_REL
SALARY	1.00	0.66	0.88	0.45	-0.18	0.78	0.43
EDUC	0.66	1.00	0.63	0.36	-0.13	0.51	0.19
SALBEGIN	0.88	0.63	1.00	0.46	-0.16	0.75	-0.01
GENDER	0.45	0.36	0.46	1.00	0.08	0.38	0.08
MINORITY	-0.18	-0.13	-0.16	0.08	1.00	-0.14	-0.09
JOBCAT	0.78	0.51	0.75	0.38	-0.14	1.00	0.20
SALINCREASE_REL	0.43	0.19	-0.01	0.08	-0.09	0.20	1.00

- Features with highest correlation with target variable SALARY:
  - SALBEGIN: r = 0.88
  - JOBCAT: r = 0.78
  - EDUC: r = 0.66
- Direction of correlations as expected

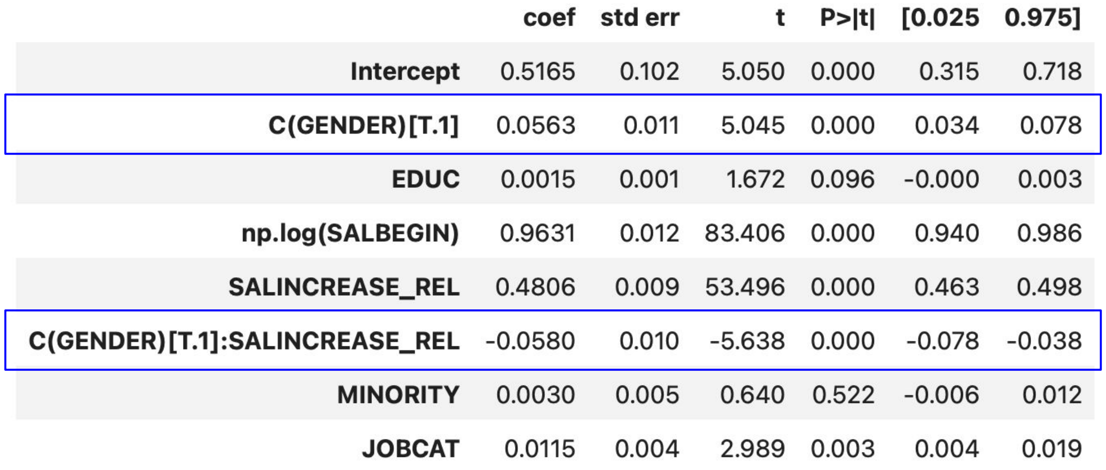
Bivariate Analysis - Inequality by Gender





- On average women earn less and their salary increase is lower compared to men

Multivariate Analysis - Model 3



	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.5165	0.102	5.050	0.000	0.315	0.718
C(GENDER)[T.1]	0.0563	0.011	5.045	0.000	0.034	0.078
EDUC	0.0015	0.001	1.672	0.096	-0.000	0.003
np.log(SALBEGIN)	0.9631	0.012	83.406	0.000	0.940	0.986
SALINCREASE_REL	0.4806	0.009	53.496	0.000	0.463	0.498
C(GENDER)[T.1]:SALINCREASE_REL	-0.0580	0.010	-5.638	0.000	-0.078	-0.038
MINORITY	0.0030	0.005	0.640	0.522	-0.006	0.012
JOBCAT	0.0115	0.004	2.989	0.003	0.004	0.019

- On average women earn 1 - exp(0.0563) = 6% less than men
- On average women's salary increases 1 - exp(-0.0580) = 5% less than men's
- R^2 = 0.994
- exp(RMSE)= 1432

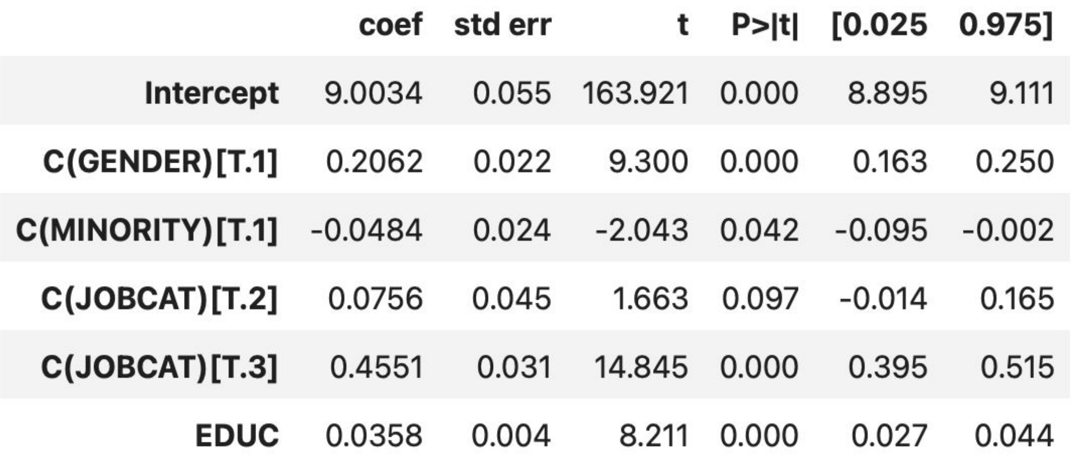
- **Exploratory** regression analysis confirms the hypotheses from bivariate analysis

Summary

- Data
  - High data quality
  - Small number of features
- Limitations:
  - R^2 ridiculously high by design (overfitted)
  - Model 3 not suited for out-of-sample predictions
  - Causality of Model 3 not clear
- Recommendations:
  - Diversity management
  - More time/budget to further investigate causality
  - Gather more explanatory information (e.g. working hours, performance)

Appendix

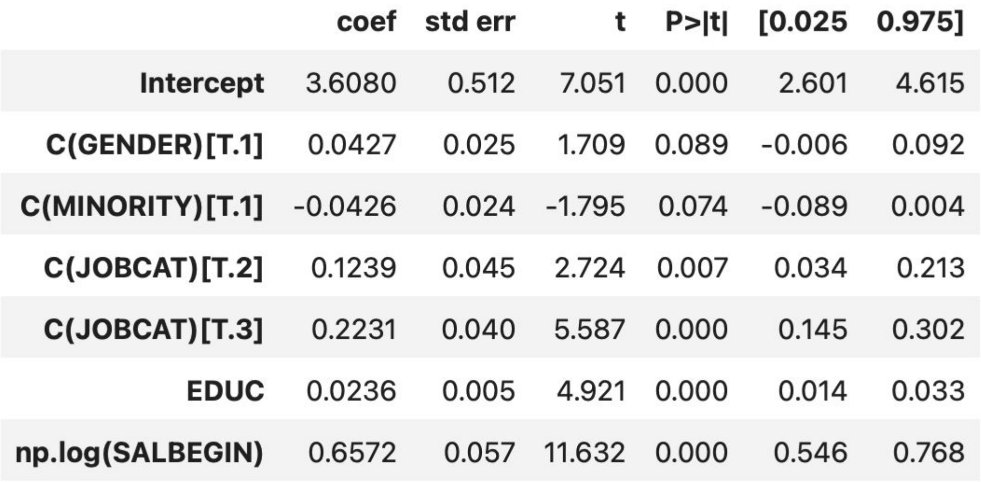
Multivariate Analysis - Model 1: Applicant



	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	9.0034	0.055	163.921	0.000	8.895	9.111
C(GENDER)[T.1]	0.2062	0.022	9.300	0.000	0.163	0.250
C(MINORITY)[T.1]	-0.0484	0.024	-2.043	0.042	-0.095	-0.002
C(JOBCAT)[T.2]	0.0756	0.045	1.663	0.097	-0.014	0.165
C(JOBCAT)[T.3]	0.4551	0.031	14.845	0.000	0.395	0.515
EDUC	0.0358	0.004	8.211	0.000	0.027	0.044

- Model to predict starting salary (log(SALBEGIN))
- exp(RMSE)= 3811
- Adj. R^2 = 0.780

Multivariate Analysis - Model 2: Employer



	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.6080	0.512	7.051	0.000	2.601	4.615
C(GENDER)[T.1]	0.0427	0.025	1.709	0.089	-0.006	0.092
C(MINORITY)[T.1]	-0.0426	0.024	-1.795	0.074	-0.089	0.004
C(JOBCAT)[T.2]	0.1239	0.045	2.724	0.007	0.034	0.213
C(JOBCAT)[T.3]	0.2231	0.040	5.587	0.000	0.145	0.302
EDUC	0.0236	0.005	4.921	0.000	0.014	0.033
np.log(SALBEGIN)	0.6572	0.057	11.632	0.000	0.546	0.768

- Model to predict current salary (log(SALARY))
- exp(RMSE)= 7553
- Adj. R^2 = 0.836
- RMSE not comparable with Model 1 (different target variable)