EDA-project:
US Bank Wages

Sandra Groß

Structure

- Data Overview
- Hypothesis
- Choosing the model based on R^2
- The Model
- Train and prediction

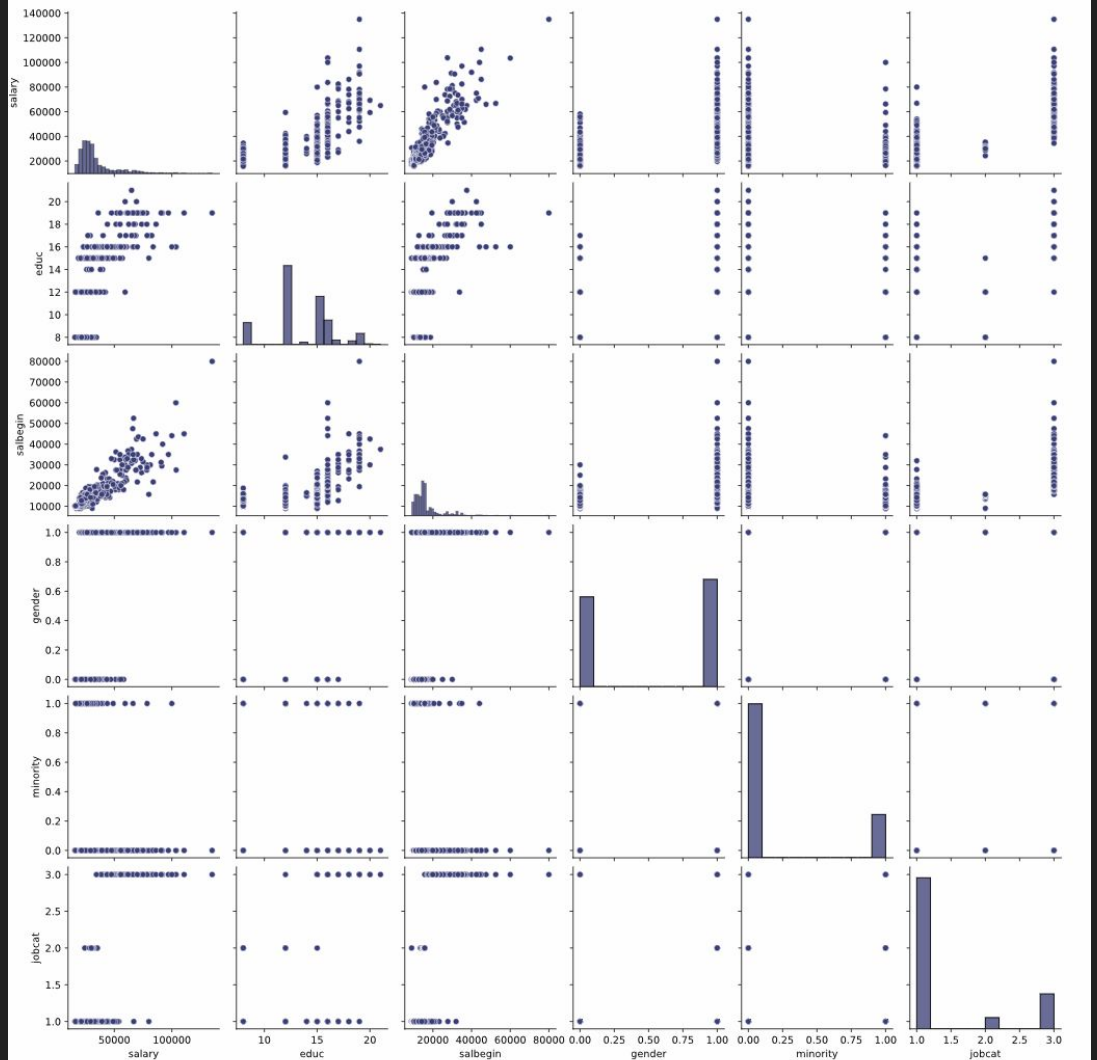
Goal

+ find a model to describe and predict the salary

Data Overview

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 474 entries, 0 to 473
Data columns (total 6 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   SALARY     474 non-null    int64
 1   EDUC       474 non-null    int64
 2   SALBEGIN   474 non-null    int64
 3   GENDER     474 non-null    int64
 4   MINORITY   474 non-null    int64
 5   JOBCAT     474 non-null    int64
dtypes: int64(6)
memory usage: 42.1 KB
```

- no NaNs
- all Dtype: int64



- Correlation between:
  - salary & education
  - salary and salbegin
- Dumie variables:
  - gender
  - minority
  - jobcat

Hypothesis

1. Salary depends on education level:

$$H_0 = \beta_{edc} = 0$$

$$H_1 = \beta_{edc} > 0$$

2. Salary depends on the first salary:

$$H_0 = \beta_{salbegin} = 0$$

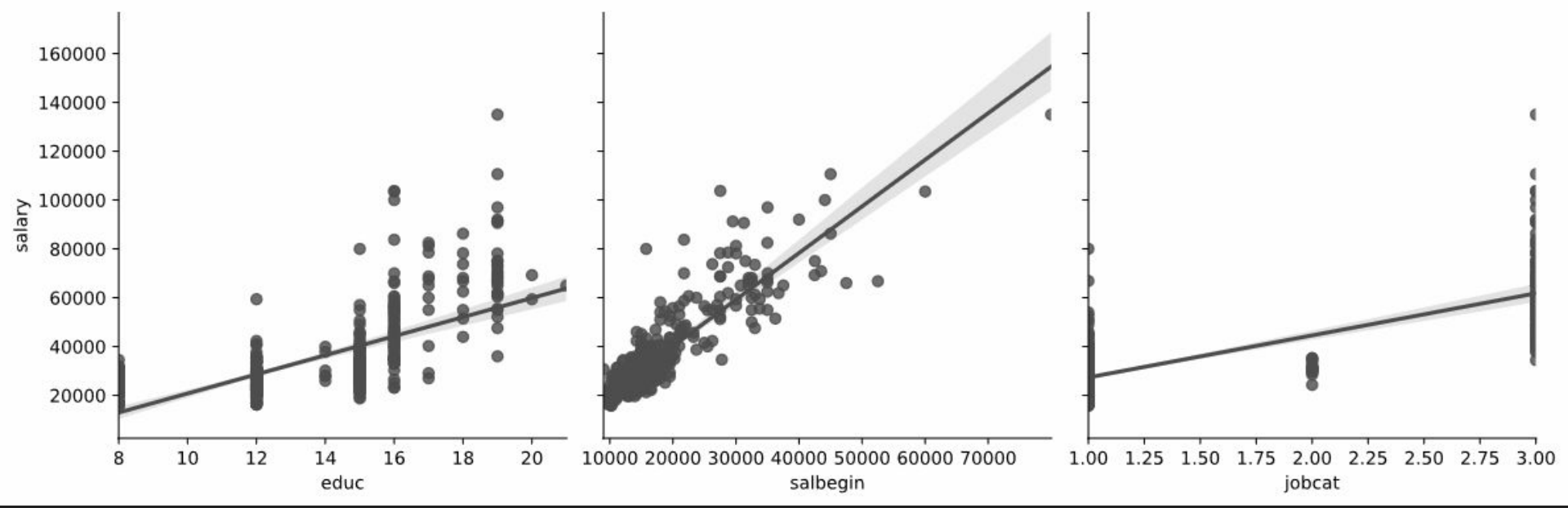
$$H_1 = \beta_{salebegin} > 0$$

3. Salary depends on job category:

$$H_0 = \beta_{jobcat} = 0$$

$$H_1 = \beta_{jobcat} > 0$$

Visualization for the linear relation



Choosing the model based on R^2

- calculated with OLS
- best found:

  $R^2 = 0.821$

  for logarithmized values

The Model

The multiple regression model explains about 82% of the variation in the salary:

$$\widehat{log(salary)} = 2.9346 + 0.2901 \text{ x } \log(\text{educ}) + 0.6851 \text{ x } \log(\text{salbegin}) + 0.2113 \text{ x } \log(\text{jobcat})$$

Train and prediction

- Method: train-test-split
- Division: 30% / 70%

- RMSE ≃ 0.178
- Accuracy ≃ 77,13%