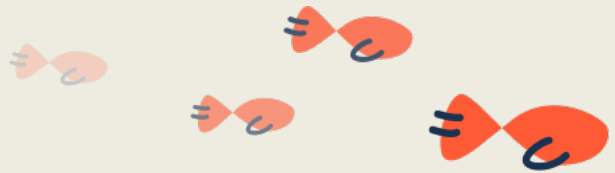
EDA-Project

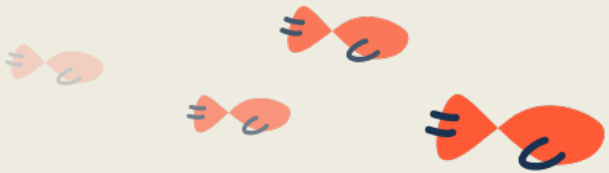
US Bank Wages

by Daniel Müller



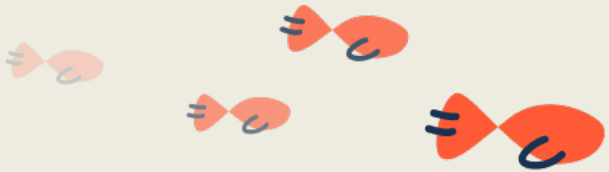
Overview

- Questions
- Impressions data set
- Procedure EDA
- Multivariate Linear Regression



Questions

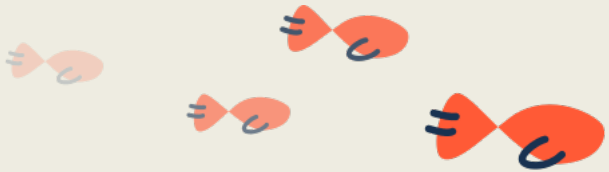
- Do men earn more than women just because they are men?
- Is the starting salary of a college graduate higher on average than of a high school graduate?
- Is there a significant correlation between starting salary and current salary?



Impressions data set

	Unnamed: 0	SALARY	EDUC	SALBEGIN	GENDER	MINORITY	JOBCAT
0	0	57000	15	27000	1	0	3
1	1	40200	16	18750	1	0	1
2	2	21450	12	12000	0	0	1

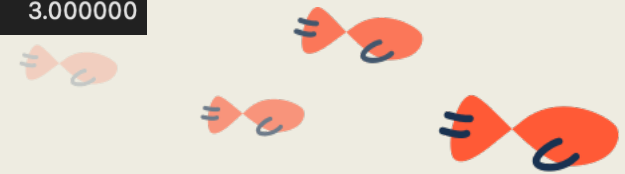
- 474 rows, 6 columns
- all columns type “int64“
- a few categorical variables



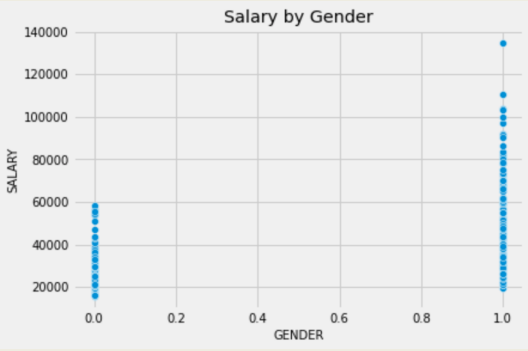
Impressions data set

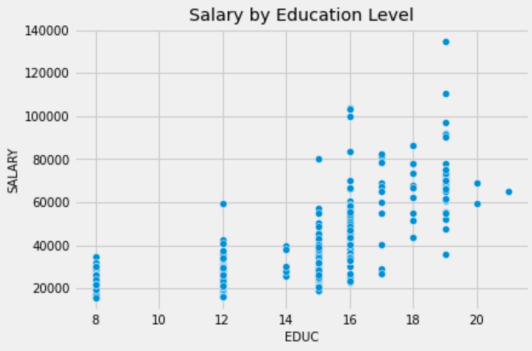
- Drop index column
- Get a statistical overview:

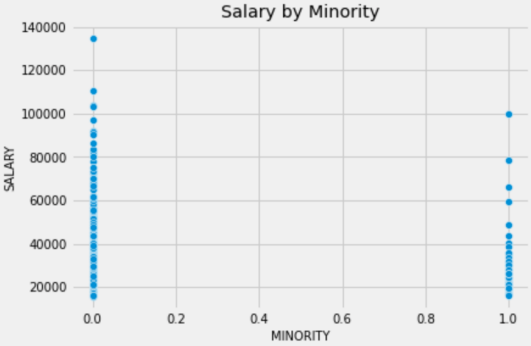
	SALARY	EDUC	SALBEGIN	GENDER	MINORITY	JOBCAT
count	474.000000	474.000000	474.000000	474.000000	474.000000	474.000000
mean	34419.567511	13.491561	17016.086498	0.544304	0.219409	1.411392
std	17075.661465	2.884846	7870.638154	0.498559	0.414284	0.773201
min	15750.000000	8.000000	9000.000000	0.000000	0.000000	1.000000
25%	24000.000000	12.000000	12487.500000	0.000000	0.000000	1.000000
50%	28875.000000	12.000000	15000.000000	1.000000	0.000000	1.000000
75%	36937.500000	15.000000	17490.000000	1.000000	0.000000	1.000000
max	135000.000000	21.000000	79980.000000	1.000000	1.000000	3.000000

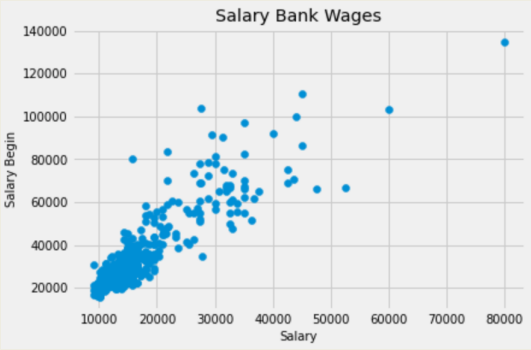


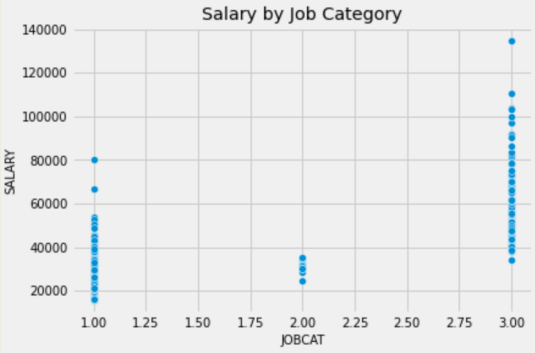
Impressions data set







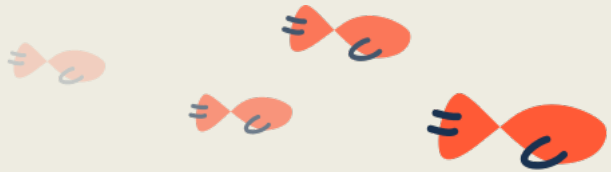




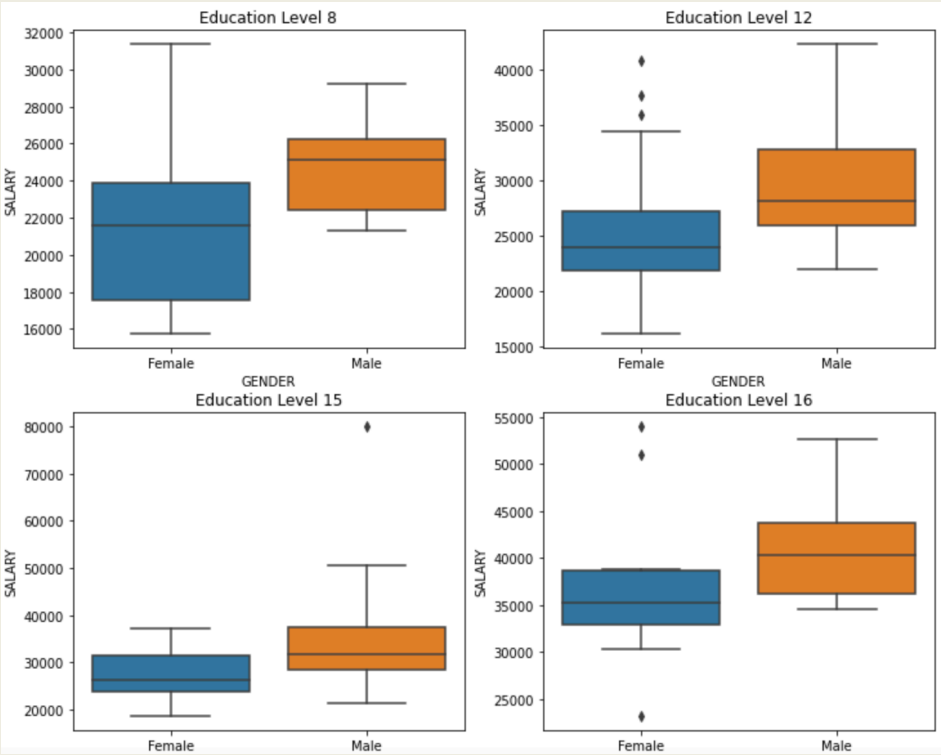
Do men earn more than women just because they are men?

Do men earn more than women just because they are men?

- Keep variables constant
- Selection of the largest possible group
- Distinction by education level
- Selection of the largest remaining groups

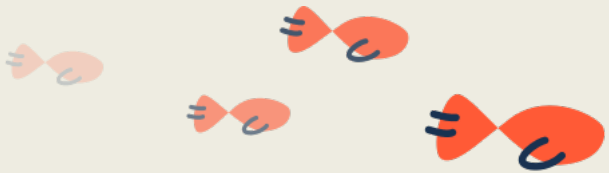


Do men earn more than women just because they are men?

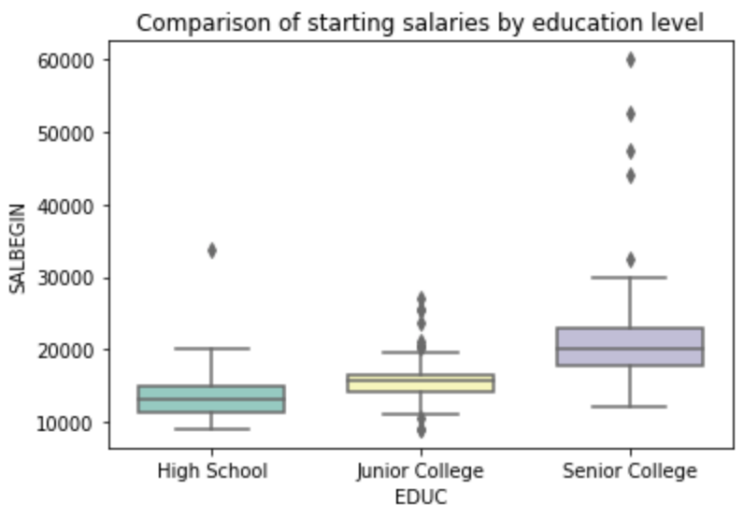


Is the starting salary of a college graduate higher on average than of a high school graduate?

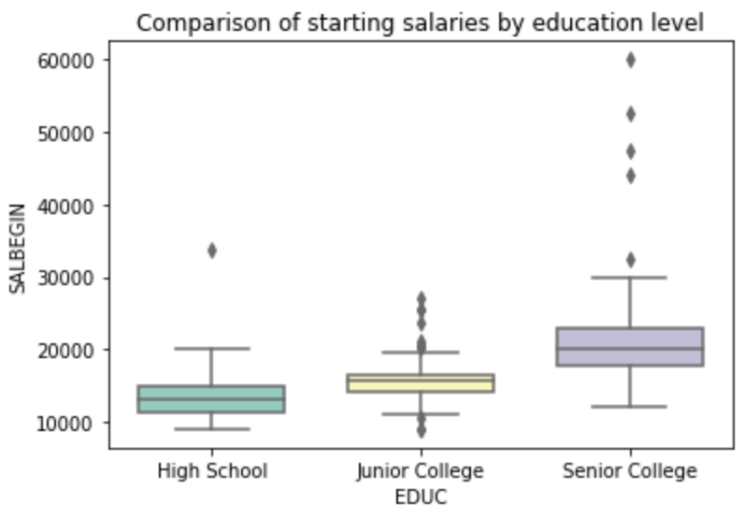
- Make assumptions about education level
- Assignment of degrees to different charecteristics
- Filter the required data

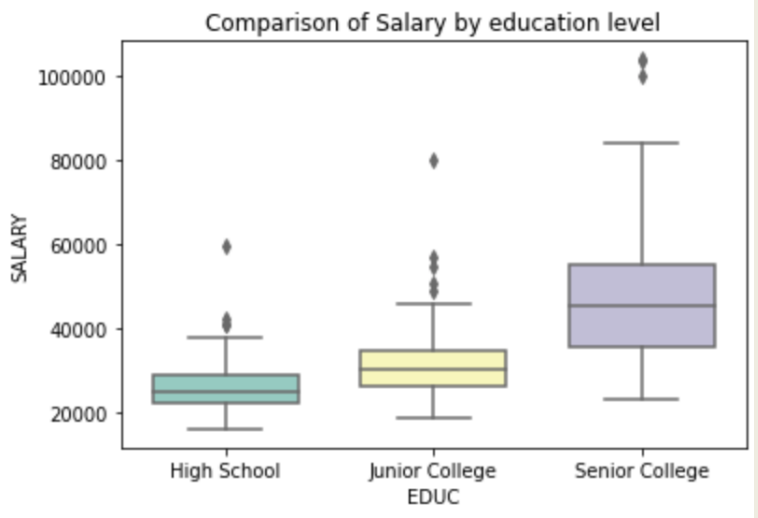


Is the starting salary of a college graduate higher on average than of a high school graduate?



Is the starting salary of a college graduate higher on average than of a high school graduate?





Is there a significant correlation between starting salary and current salary?

Is there a significant correlation between starting salary and current salary?

- Salary as target
- Starting Salary as feature
- Create OLS model
- Calculate Intercept & Slope

Is there a significant correlation between starting salary and current salary?

OLS Regression Results

Dep. Variable:	SALARY	**R-squared:**	0.775
Model:	OLS	**Adj. R-squared:**	0.774
Method:	Least Squares	**F-statistic:**	1622.
Date:	Thu, 18 Feb 2021	**Prob (F-statistic):**	8.20e-155
Time:	07:44:50	**Log-Likelihood:**	-4938.3
No. Observations:	474	**AIC:**	9881.
Df Residuals:	472	**BIC:**	9889.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1928.2058	888.680	2.170	0.031	181.947	3674.464
SALBEGIN	1.9094	0.047	40.276	0.000	1.816	2.003

Omnibus:	199.258	**Durbin-Watson:**	1.830
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1534.909
Skew:	1.630	**Prob(JB):**	0.00
Kurtosis:	11.191	**Cond. No.**	4.47e+04

- R-squared & Adj. R-squared near 1
- P-Values < 0.05

- $y = 1928.2058 + 1.9094x$

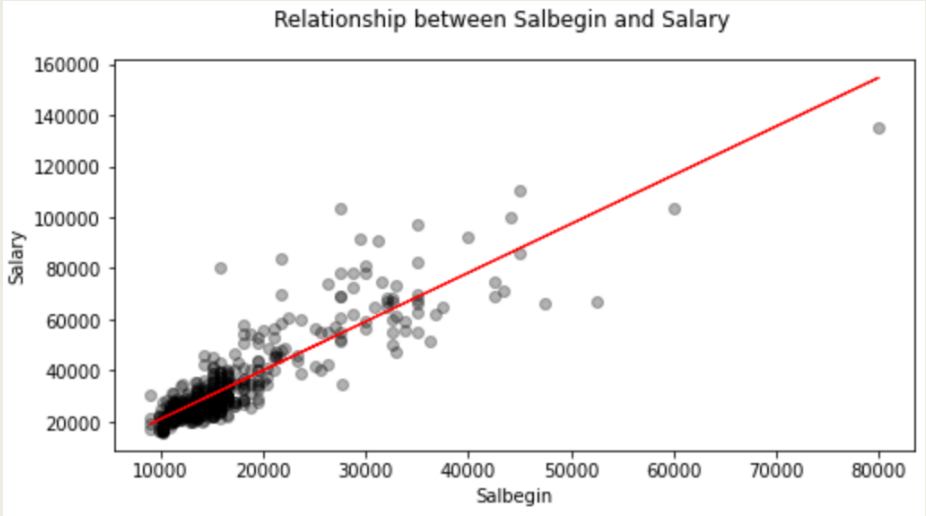
Is there a significant correlation between starting salary and current salary?

OLS Regression Results

Dep. Variable:	SALARY	**R-squared:**	0.775
Model:	OLS	**Adj. R-squared:**	0.774
Method:	Least Squares	**F-statistic:**	1622.
Date:	Thu, 18 Feb 2021	**Prob (F-statistic):**	8.20e-155
Time:	07:44:50	**Log-Likelihood:**	-4938.3
No. Observations:	474	**AIC:**	9881.
Df Residuals:	472	**BIC:**	9889.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1928.2058	888.680	2.170	0.031	181.947	3674.464
SALBEGIN	1.9094	0.047	40.276	0.000	1.816	2.003

Omnibus:	199.258	**Durbin-Watson:**	1.830
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1534.909
Skew:	1.630	**Prob(JB):**	0.00
Kurtosis:	11.191	**Cond. No.**	4.47e+04



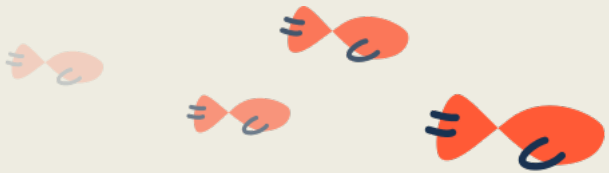
Multivariate Linear Regression

- Create dummie variables
- Drop origin columns
- Add dummie columns

	SALARY	SALBEGIN	edu_12	edu_14	edu_15	edu_16	edu_17	edu_18	edu_19	edu_20	edu_21	gd_1	mino_1	jcat_2	jcat_3
0	57000	27000	0	0	1	0	0	0	0	0	0	1	0	0	1
1	40200	18750	0	0	0	1	0	0	0	0	0	1	0	0	0
2	21450	12000	1	0	0	0	0	0	0	0	0	0	0	0	0
3	21900	13200	0	0	0	0	0	0	0	0	0	0	0	0	0
4	45000	21000	0	0	1	0	0	0	0	0	0	1	0	0	0

Multivariate Linear Regression

- Test size: 0.2
- Features: all 12 variables
- R-squared: 0.826
- Adj. R-squared: 0.820
- RMSE (train data): 7460.17
- RMSE (test data): 6366.35



Thanks for your Attention!