US Bank Wages

7th of May, 2021

Setting

A client needs help with the data analysis of a data set. The newly hired equal opportunities officer of the US bank is interested in understanding which points can be improved in his new company.

He wants to get insights concerning following topics:

- How is the gender distribution in the US bank?
- Is there a gender pay gap?
- Do people from minority groups have a disadvantage?
- Is the education directly related to the starting salary?

Data Source

Data source is a text file: “us_bank_wages.txt”

Content of this text files is data concerning the following items:

- job category
- starting salary
- current salary
- years of education
- gender
- information concerning whether the person is part of a minority or not

Data cleaning

- delete / add columns
- use lowercase column names without spaces
- rename columns
- check for duplicate entries
- check for categorical parameters
- check data types

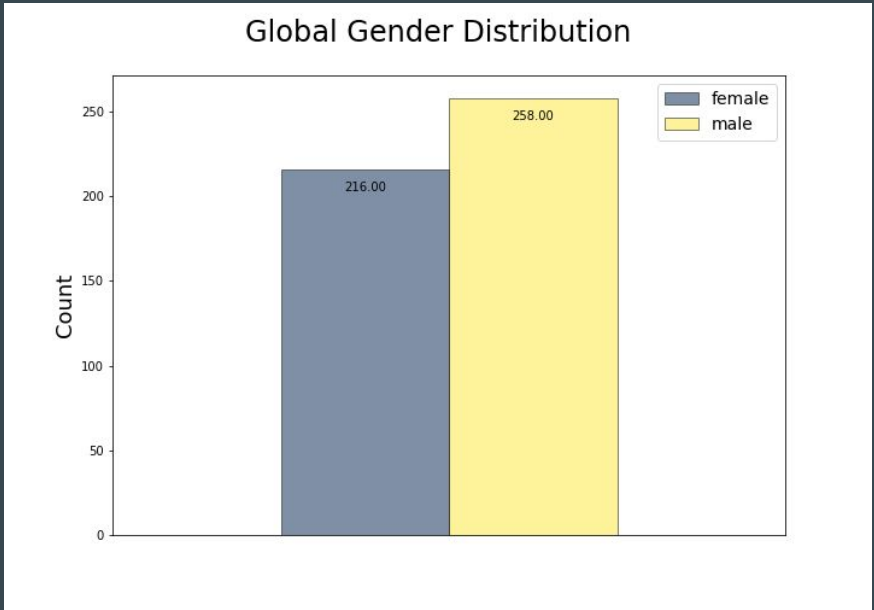
	Unnamed: 0	SALARY	EDUC	SALBEGIN	GENDER	MINORITY	JOBCAT
0	0	57000	15	27000	1	0	3
1	1	40200	16	18750	1	0	1
2	2	21450	12	12000	0	0	1
3	3	21900	8	13200	0	0	1
4	4	45000	15	21000	1	0	1
...	...	...	...	...	...	...	...
469	469	26250	12	15750	1	1	1
470	470	26400	15	15750	1	1	1
471	471	39150	15	15750	1	0	1
472	472	21450	12	12750	0	0	1
473	473	29400	12	14250	0	0	1

474 rows × 7 columns

	start_s	current_s	jobcat	educ	gender	minority
0	27000	57000	3	15	1	0
1	18750	40200	1	16	1	0
2	12000	21450	1	12	0	0
3	13200	21900	1	8	0	0
4	21000	45000	1	15	1	0
...	...	...	...	...	...	...
469	15750	26250	1	12	1	1
470	15750	26400	1	15	1	1
471	15750	39150	1	15	1	0
472	12750	21450	1	12	0	0
473	14250	29400	1	12	0	0

474 rows × 6 columns

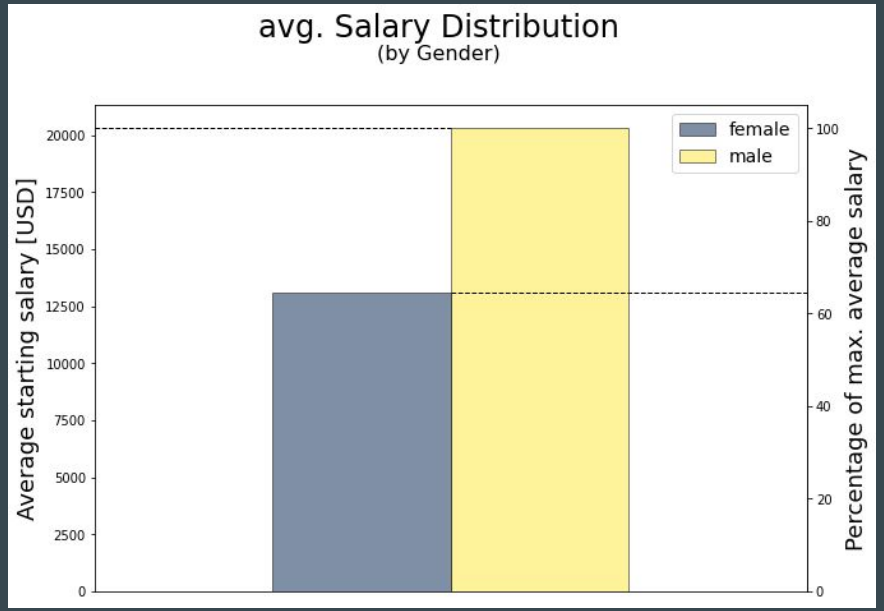
Q: How is the gender distribution in the US Bank?



First impression:

- quite equal

Q: Is there a gender pay gap? Focus on starting salary

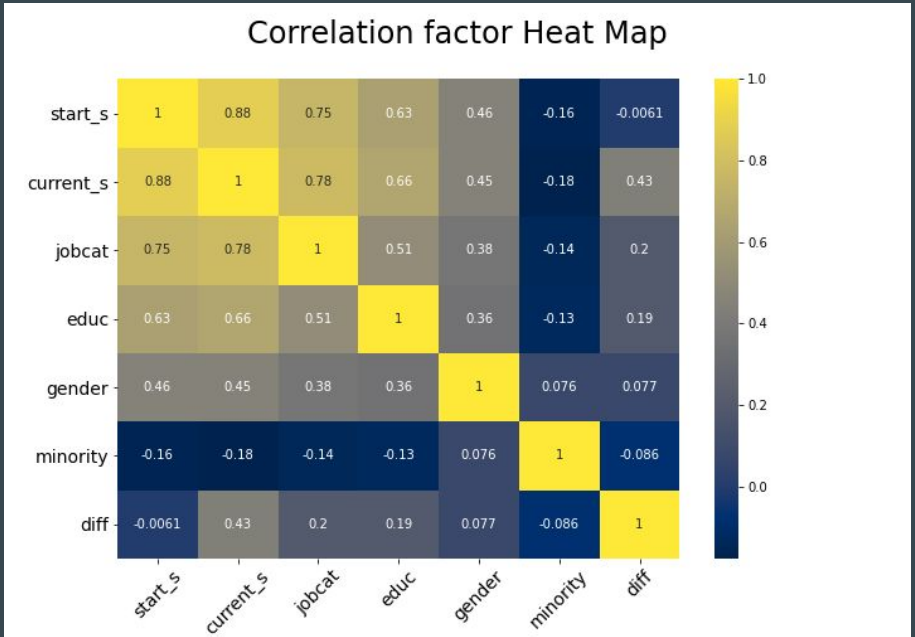


First impression:

- not equal
- female salary only 64.5% of male salary

What are the driving parameters?

Q: Is there a gender pay gap? Focus on starting salary

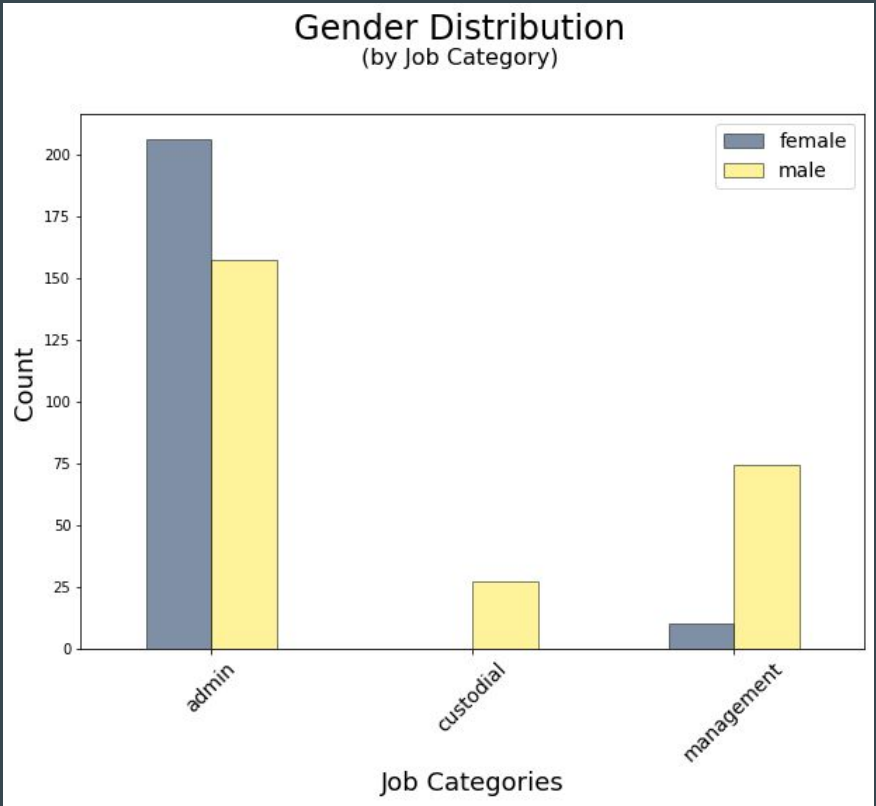


Correlated parameters:

- job category
- educational years
- gender

Let's take a look at those!

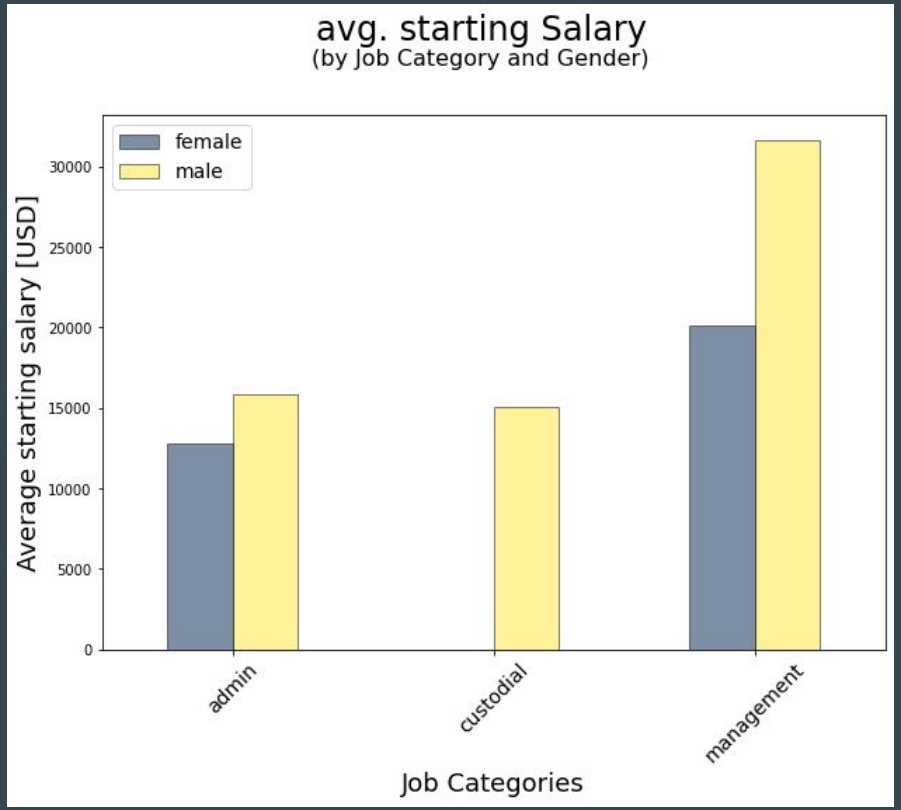
Q: Is there a gender pay gap? Parameter: job category



Observations:

- no females in job category “custodial”
- most females employed in administration department

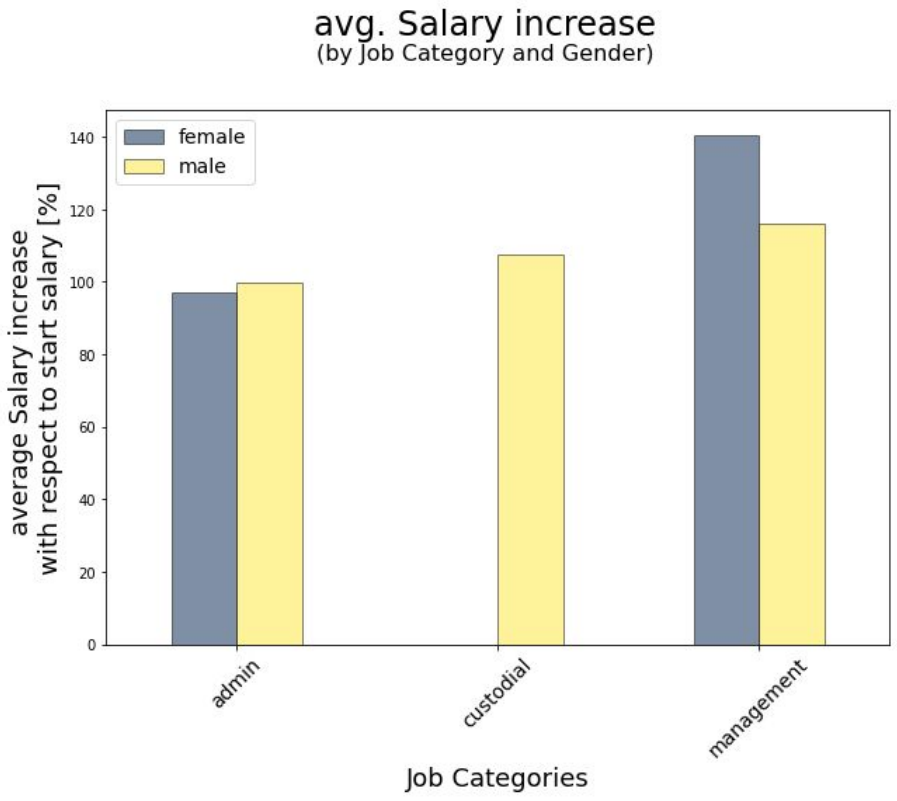
Q: Is there a gender pay gap? Parameter: job category



Observations:

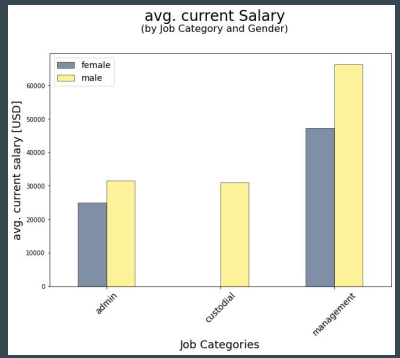
- most salary inequality in management department

Q: Is there a gender pay gap? Parameter: salary increase



Observations:

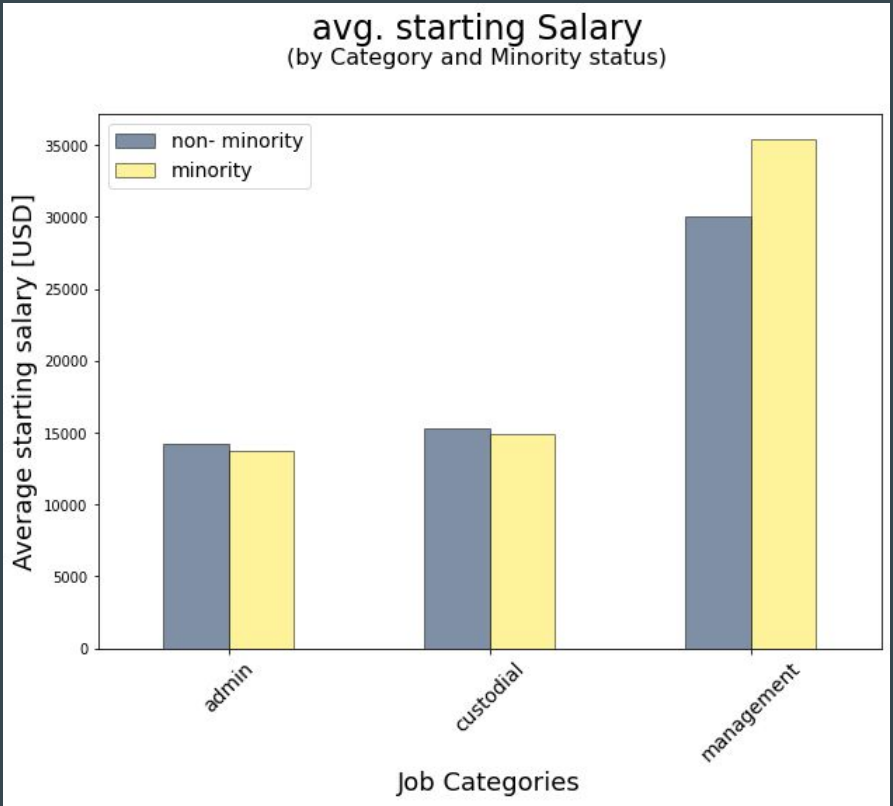
- no significant inequality in salary increase. Females have a slight advantage in the management department.



current salary

Investigation on minority groups

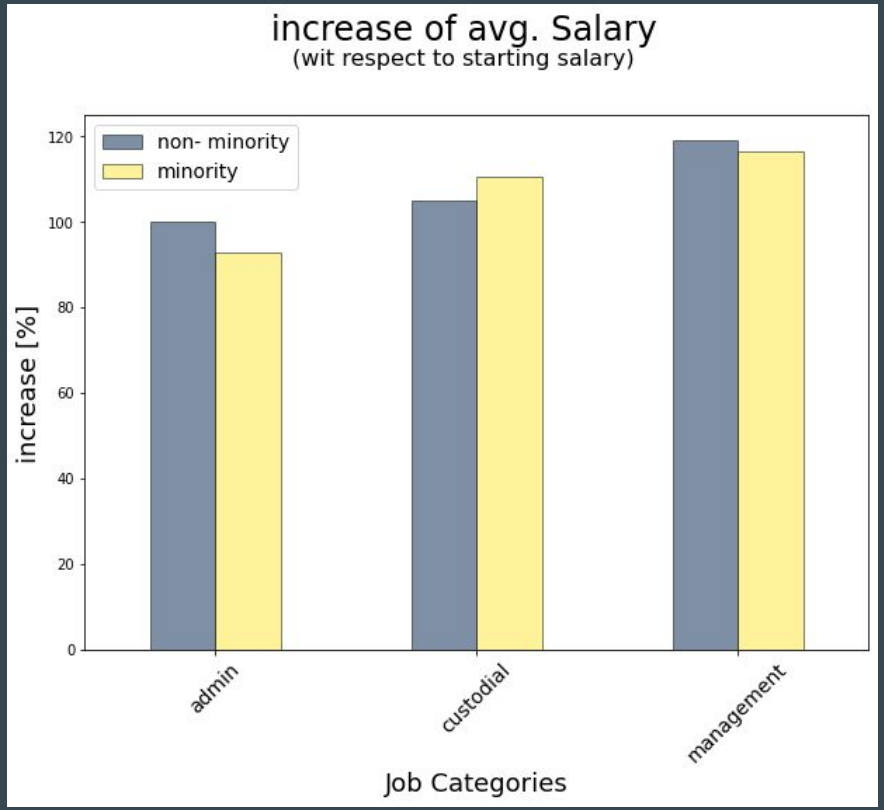
Q: Do people from minority groups have disadvantages?



Observations:

- no obvious structural disadvantages
- slight advantage in management department

Q: Do people from minority groups have disadvantages?

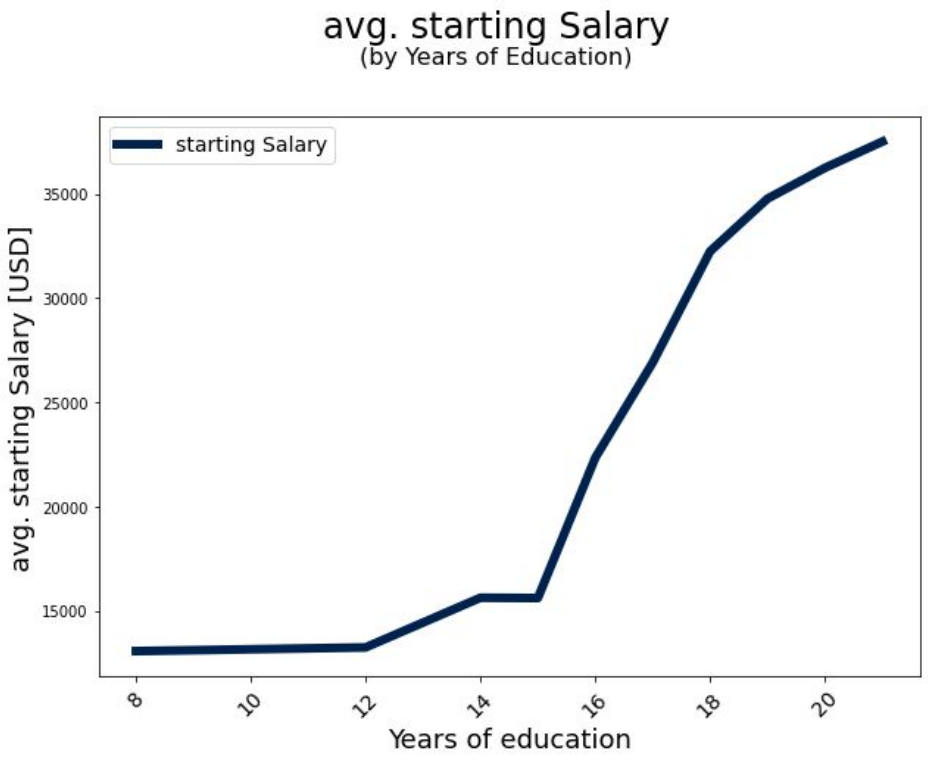


Observations:

- no obvious structural disadvantages

Educational structure

Q: Relation between education and starting salary

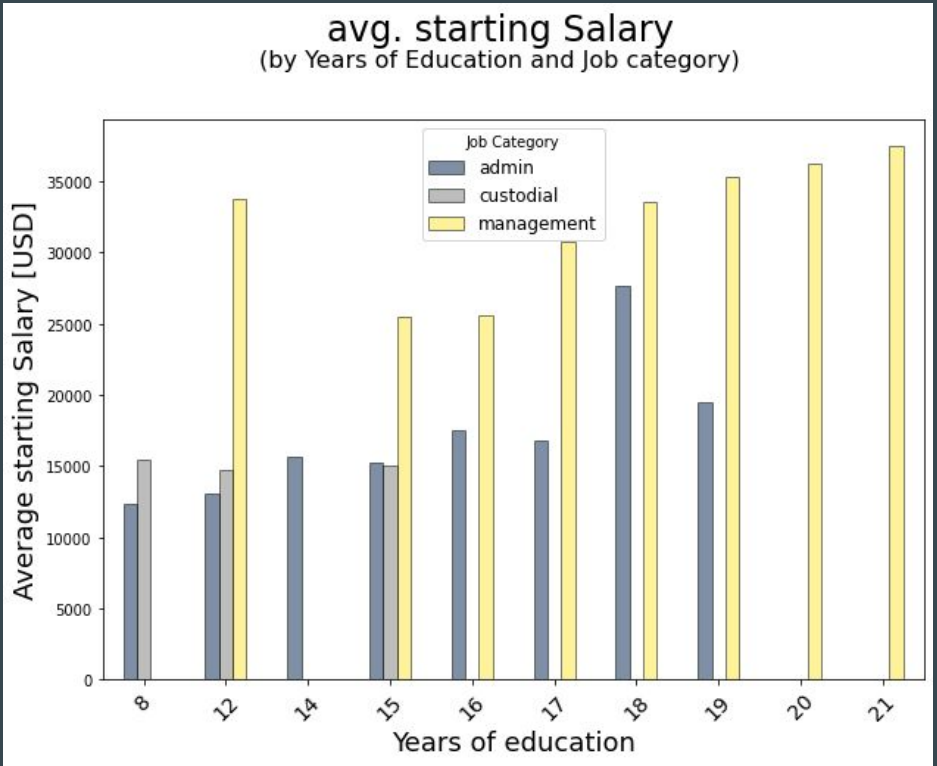


Observations:

- there is a relation, but not a linear one

Let’s take a look at the job categories!

Q: Relation between education and starting salary



Observations:

- In order to work in the management department, at least 12 years of education is required
- years of education alone do not guarantee a substantial higher starting salary

Regression model

Regression model

```
X_1 = df_LR[['educ', 'gender', 'minority', 'jobcat']]
y_1 = df_LR['start_s_log']

X_train_1, X_test_1, y_train_1, y_test_1 = train_test_split(X_1, y_1, test_size=0.25, random_state=42, shuffle=True)

# Merge datasets after test split for formula notation
X_train_1 = X_train_1.merge(y_train_1, left_index = True, right_index=True)

# Create and train an OLS model
results_1 = smf.ols(formula='start_s_log ~ educ + C(gender) + C(minority) + C(jobcat)', data=X_train_1).fit()

# Return output of the model
results_1.summary()
```

OLS Regression Results

Dep. Variable:	start_s_log	**R-squared:**	0.773
Model:	OLS	**Adj. R-squared:**	0.770
Method:	Least Squares	**F-statistic:**	238.1
Date:	Mon, 07 Jun 2021	**Prob (F-statistic):**	4.03e-110
Time:	11:48:48	**Log-Likelihood:**	129.75
No. Observations:	355	**AIC:**	-247.5
Df Residuals:	349	**BIC:**	-224.3
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	9.0298	0.053	171.283	0.000	8.926	9.133
C(gender)[T.1]	0.2001	0.021	9.419	0.000	0.158	0.242
C(minority)[T.1]	-0.0519	0.023	-2.271	0.024	-0.097	-0.007
C(jobcat)[T.2]	0.0541	0.044	1.242	0.215	-0.032	0.140
C(jobcat)[T.3]	0.4786	0.030	15.913	0.000	0.419	0.538
educ	0.0334	0.004	7.946	0.000	0.025	0.042

Omnibus:	91.128	**Durbin-Watson:**	1.916
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	290.545
Skew:	1.138	**Prob(JB):**	8.11e-64
Kurtosis:	6.804	**Cond. No.**	90.2

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

- Split Data Frame into Test/Train Data Frame
- Set up multiple linear regression model

Questions we had

- How is the gender distribution in the US bank?
- Is there a gender pay gap?
- Do people from minority groups have a disadvantage?
- Is the education directly related to the starting salary?

Answers we found

- Quite equal, 45,5% female, 54,5% male
- Yes, due to the starting salary being not equal
- No obvious disadvantages
- It is, but the job category is more dominant