# My First EDA Project

Data Sience Course Q1 DS-21-1 (Colon) powered by neuefische ™

# There are two topics available

1. King County Housing Data: This dataset contains information about home sales in King County (USA).

2. US Bank Wages: This dataset contains information about the wages of employees of a US bank.

**I chose: US Bank Wages**

# Tasks for you

1. Create a new repo and a new virtual environment.

2. Through EDA/statistical analysis above please come up with AT LEAST 3 insights/recommendations for your stakeholder. If you use linear regression in the exploration phase remember that R2 close to 1 is good.

3. Then, model this dataset with a multivariate linear regression to predict b. Note you can take either the perspective of an applicant or company.
    a. Split the dataset into a train and a test set. (use the sklearn split method
    b. Use Root Mean Squared Error (RMSE) as your metric of success and try to minimize this score on your test data.

# Stakeholder

Is a NGO that works on the integration of female war-refugees into society.

**The Task**

- The NGO is suspecting that a certain bank does not follow their own code of conduct.

**The Questions**

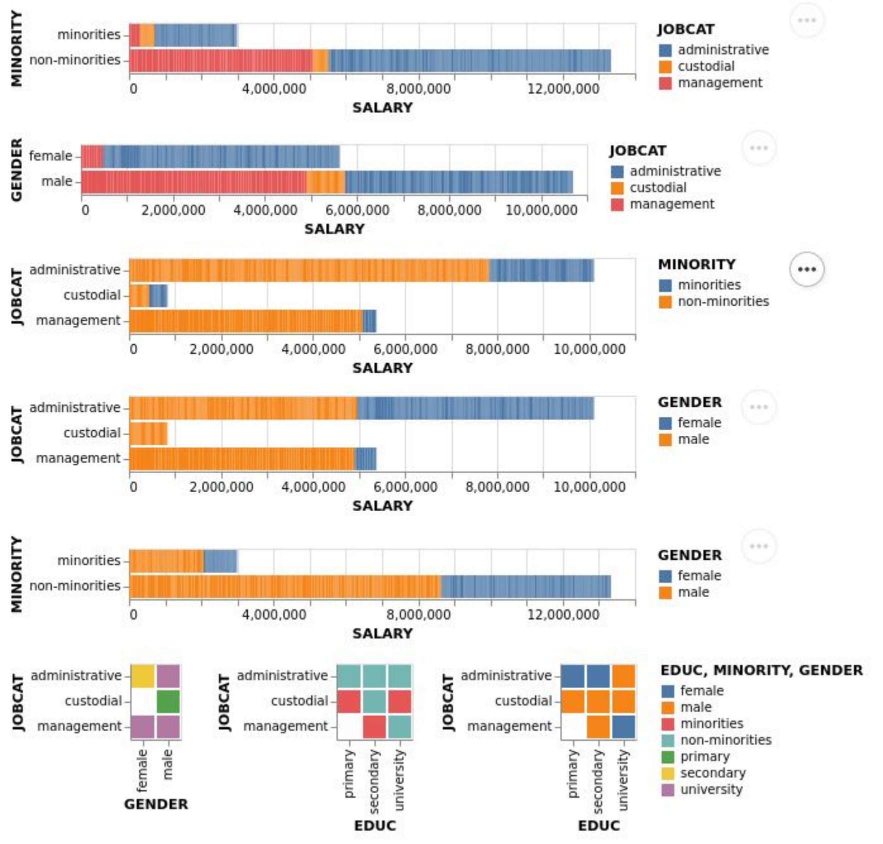
- What is the salary based on in this bank?
- Is there an difference in pay between male and female employees?
- Would females with a minority background have the same chances for a fair salary as non-minority women?

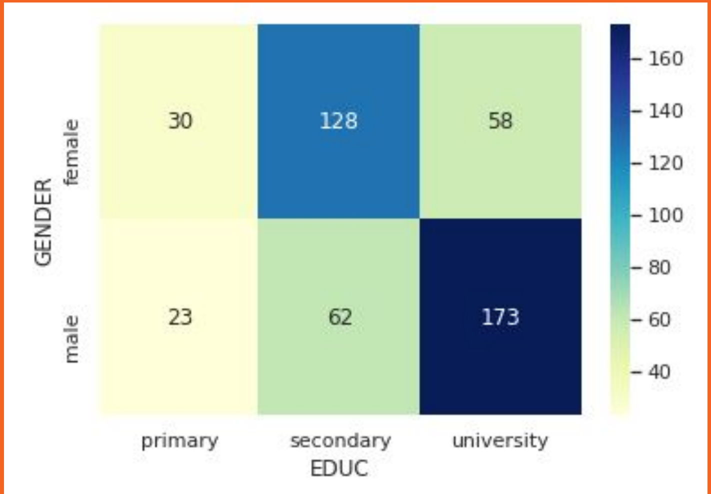
# My Approach

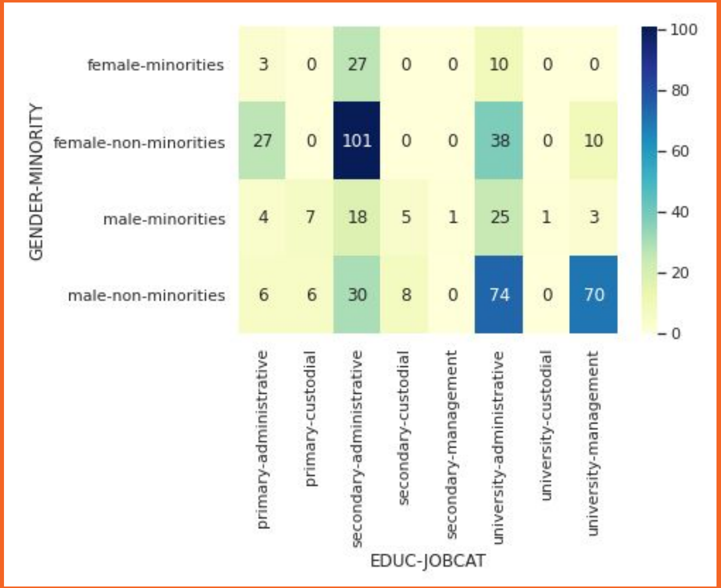
1. First I  set-up my environment and cleaned the data.
2. Then I started my EDA cycles, focusing on graphical representation of the data.
3. Next I focused on LRM model creation based on R^2_adj.
4. I went back to run more EDA cycles, before I generated a LRM model.
5. Finally I computed targets and the RMSE.
6. Last not least, I created functions for data storage.

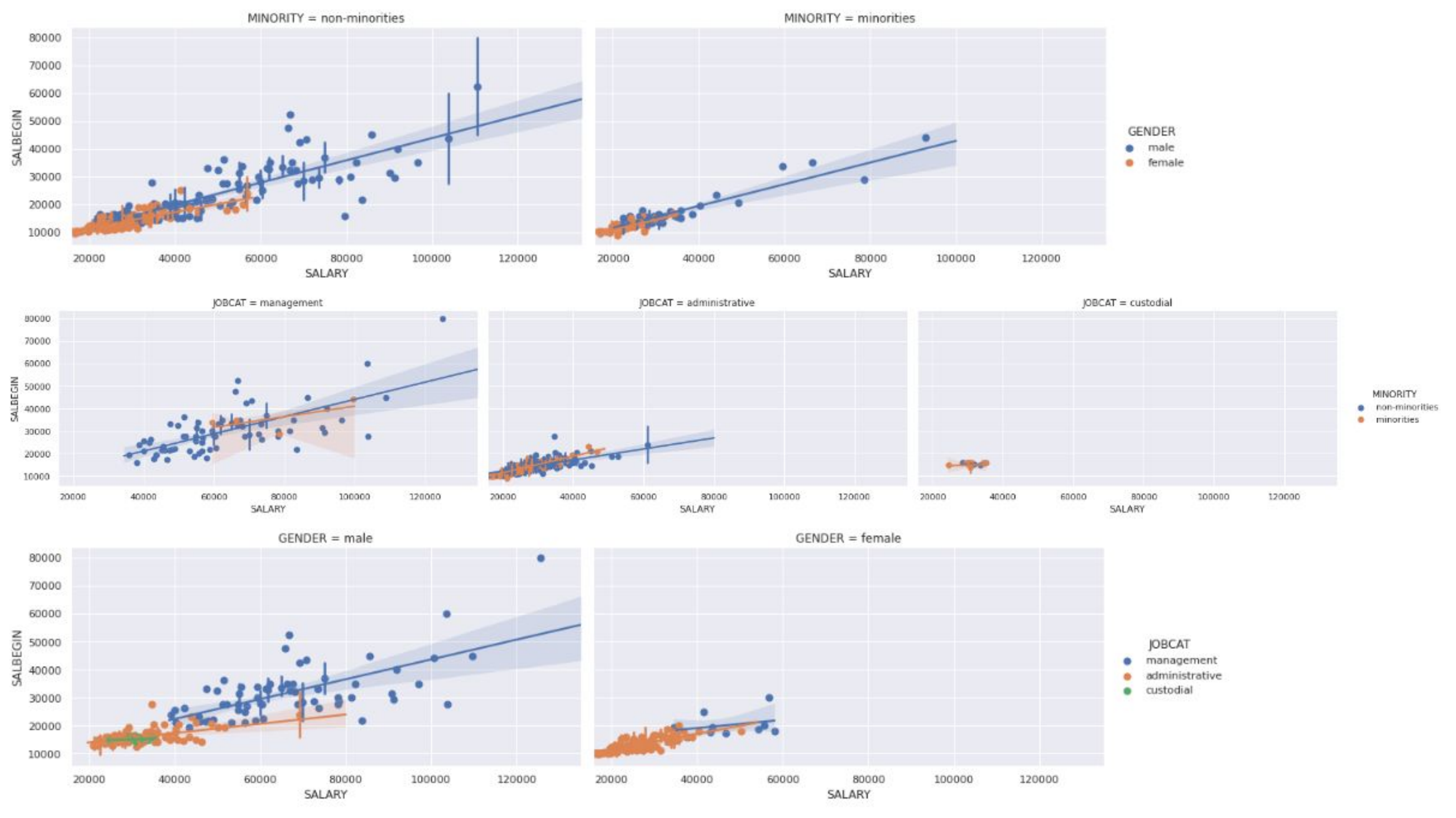
# EDA

| | EDUC | primary | | secondary | | | university | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| | JOBCAT | administrative | custodial | administrative | custodial | management | administrative | custodial | management | |
| GENDER | MINORITY | | | | | | | | | |
| female | minorities | 3 | 0 | 27 | 0 | 0 | 10 | 0 | 0 | 40 |
| | non-minorities | 27 | 0 | 101 | 0 | 0 | 38 | 0 | 10 | 176 |
| male | minorities | 4 | 7 | 18 | 5 | 1 | 25 | 1 | 3 | 64 |
| | non-minorities | 6 | 6 | 30 | 8 | 0 | 74 | 0 | 70 | 194 |
| Total | | 40 | 13 | 176 | 13 | 1 | 147 | 1 | 83 | 474 |


MINORITY
minorities
non-minorities
0
4,000,000
8,000,000
12,000,000
SALARY
JOBCAT
administrative
custodial
management
GENDER
female
male
0
2,000,000
4,000,000
6,000,000
8,000,000
10,000,000
SALARY
JOBCAT
administrative
custodial
management
JOBCAT
administrative
custodial
management
0
2,000,000
4,000,000
6,000,000
8,000,000
10,000,000
SALARY
MINORITY
minorities
non-minorities
JOBCAT
administrative
custodial
management
SALARY
GENDER
female
male
MINORITY
minorities
non-minorities
SALARY
GENDER
female
male
JOBCAT
female
male
GENDER
primary
secondary
university
EDUC
EDUC, MINORITY, GENDER
female
male
minorities
non-minorities
primary
secondary
university

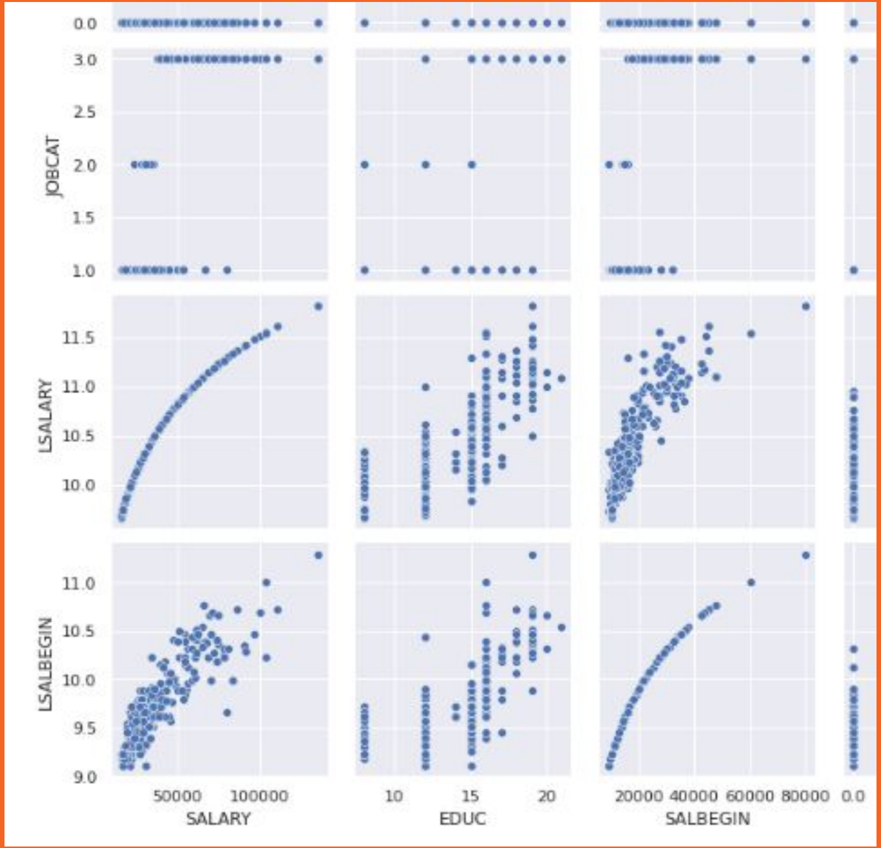

GENDER
female
male
30
128
58
23
62
173
primary
secondary
university
EDUC
160
140
120
100
80
60
40

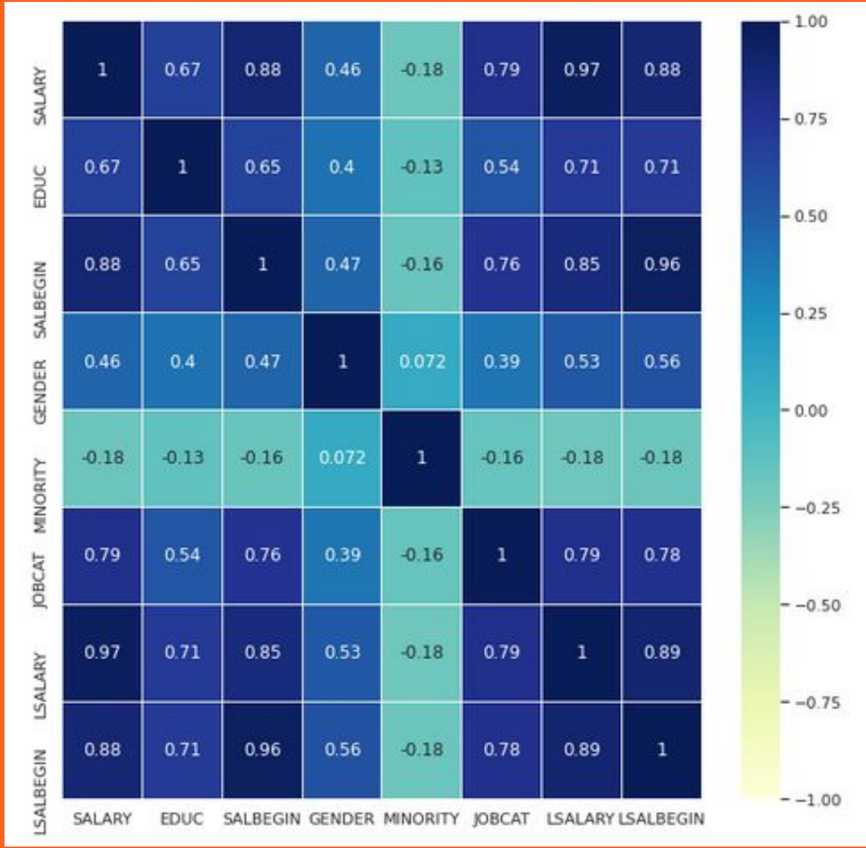

GENDER-MINORITY
female-minorities
female-non-minorities
male-minorities
male-non-minorities
3 0 27 0 0 10 0 0
27 0 101 0 0 38 0 10
4 7 18 5 1 25 1 3
6 6 30 8 0 74 0 70
primary-administrative
primary-custodial
secondary-administrative
secondary-custodial
secondary-management
university-administrative
university-custodial
university-management
EDUC-JOBCAT
100
80
60
40
20
0



MINORITY = non-minorities
MINORITY = minorities
SALBEGIN
SALARY
GENDER
male
female
JOBCAT = management
JOBCAT = administrative
JOBCAT = custodial
MINORITY
non-minorities
minorities
GENDER = male
GENDER = female
JOBCAT
management
administrative
custodial

# LRM


JOBCAT
LSALARY
LSALBEGIN
SALARY
EDUC
SALBEGIN


SALARY
EDUC
SALBEGIN
GENDER
MINORITY
JOBCAT
LSALARY
LSALBEGIN
1 0.67 0.88 0.46 -0.18 0.79 0.97 0.88
0.67 1 0.65 0.4 -0.13 0.54 0.71 0.71
0.88 0.65 1 0.47 -0.16 0.76 0.85 0.96
0.46 0.4 0.47 1 0.072 0.39 0.53 0.56
-0.18 -0.13 -0.16 0.072 1 -0.16 -0.18 -0.18
0.79 0.54 0.76 0.39 -0.16 1 0.79 0.78
0.97 0.71 0.85 0.53 -0.18 0.79 1 0.89
0.88 0.71 0.96 0.56 -0.18 0.78 0.89 1

```
[258]: # generate column combinations - to check various models

var = []
columns = ['SALBEGIN','LSALBEGIN','GENDER','C(MINORITY)','C(JOBCAT)','C(EDUC)']
for i in range(len(columns)):
    # remove first
    cols = columns[i+1:]
    # add first to end
    cols += columns[:i+1]
    #print(cols)

    for i in range(len(cols)):
        if not [cols[i]] in var:
            var.append([cols[i]])

        c = cols.copy()
        c.remove(cols[i])
        while len(c):
            if not sorted([x for x in c]) in var:
                var.append(sorted([x for x in c]))
            x = c.pop() # x is only used to have less print-output in JL

sorted(var)
```

```
[258]: [['C(EDUC)'],
 ['C(EDUC)', 'C(JOBCAT)'],
 ['C(EDUC)', 'C(JOBCAT)', 'C(MINORITY)'],
 ['C(EDUC)', 'C(JOBCAT)', 'C(MINORITY)', 'GENDER'],
 ['C(EDUC)', 'C(JOBCAT)', 'C(MINORITY)', 'GENDER', 'LSALBEGIN'],
 ['C(EDUC)', 'C(JOBCAT)', 'C(MINORITY)', 'GENDER', 'SALBEGIN'],
 ['C(EDUC)', 'C(JOBCAT)', 'C(MINORITY)', 'LSALBEGIN'],
 ['C(EDUC)', 'C(JOBCAT)', 'C(MINORITY)', 'LSALBEGIN', 'SALBEGIN'],
 ['C(EDUC)', 'C(JOBCAT)', 'C(MINORITY)', 'SALBEGIN'],
 ['C(EDUC)', 'C(JOBCAT)', 'GENDER'],
 ['C(EDUC)', 'C(JOBCAT)', 'GENDER', 'LSALBEGIN'],
 ['C(EDUC)', 'C(JOBCAT)', 'GENDER', 'LSALBEGIN', 'SALBEGIN'],
 ['C(EDUC)', 'C(JOBCAT)', 'GENDER', 'SALBEGIN'],
 ['C(EDUC)', 'C(JOBCAT)', 'LSALBEGIN'],
 ['C(EDUC)', 'C(JOBCAT)', 'LSALBEGIN', 'SALBEGIN'],
 ['C(EDUC)', 'C(JOBCAT)', 'SALBEGIN'],
 ['C(EDUC)', 'C(MINORITY)'],
 ['C(EDUC)', 'C(MINORITY)', 'GENDER'],
 ['C(EDUC)', 'C(MINORITY)', 'GENDER', 'LSALBEGIN'],
 ['C(EDUC)', 'C(MINORITY)', 'GENDER', 'LSALBEGIN', 'SALBEGIN'],
 ['C(EDUC)', 'C(MINORITY)', 'GENDER', 'SALBEGIN'],
 ['C(EDUC)', 'C(MINORITY)', 'LSALBEGIN', 'SALBEGIN'],
 ['C(EDUC)', 'C(MINORITY)', 'SALBEGIN'],
 ['C(EDUC)', 'GENDER', 'LSALBEGIN'],
 ['C(EDUC)', 'GENDER', 'LSALBEGIN', 'SALBEGIN'],
 ['C(EDUC)', 'GENDER', 'SALBEGIN'],
 ['C(EDUC)', 'LSALBEGIN'],
 ['C(EDUC)', 'LSALBEGIN', 'SALBEGIN'],
 ['C(EDUC)', 'SALBEGIN'],
 ['C(JOBCAT)'],
 ['C(JOBCAT)', 'C(MINORITY)'],
 ['C(JOBCAT)', 'C(MINORITY)', 'GENDER'],
 ['C(JOBCAT)', 'C(MINORITY)', 'GENDER', 'LSALBEGIN'],
 ['C(JOBCAT)', 'C(MINORITY)', 'GENDER', 'LSALBEGIN', 'SALBEGIN'],
```

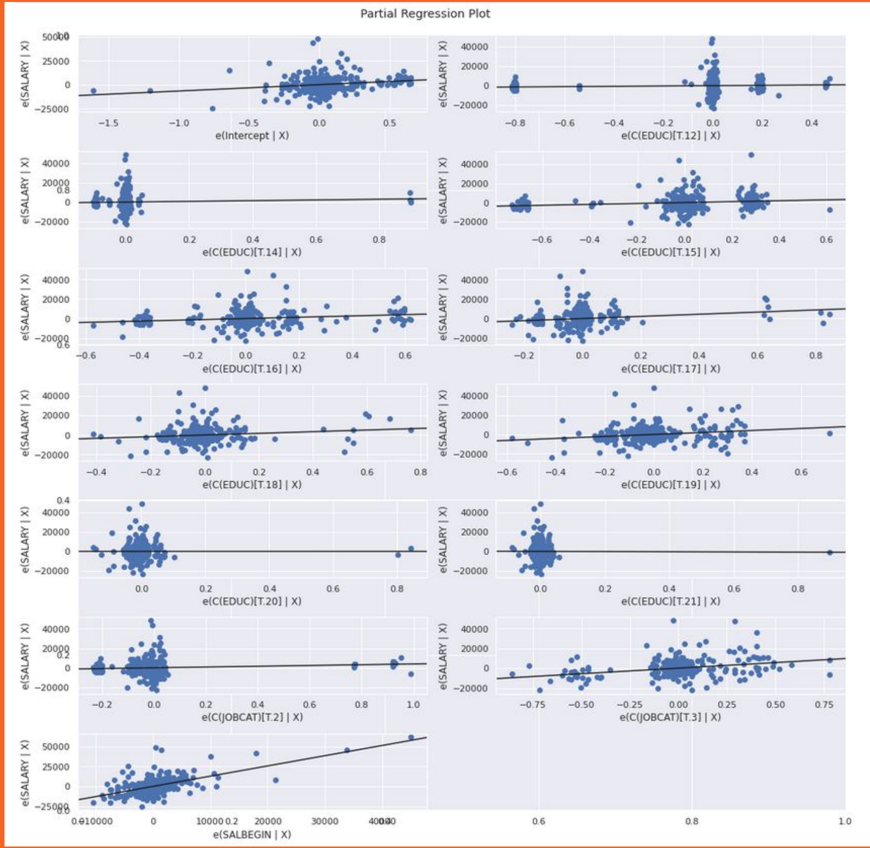
```
[260]: # run brute force model computation
model_fit_result = find_best_model(var,False)

# list model fittings - top 10
model_fit_result_list = sorted(model_fit_result.keys())[-11:]
for fit in model_fit_result_list:
    print('rsquared_adj:', fit, '\t<-', model_fit_result[fit])
```

```
rsquared_adj: 0.8182763850083632      <- SALARY ~ C(JOBCAT) + LSALBEGIN + SALBEGIN
rsquared_adj: 0.8187965885289206      <- SALARY ~ C(JOBCAT) + C(MINORITY) + GENDER + SALBEGIN
rsquared_adj: 0.8190180154354085      <- SALARY ~ C(JOBCAT) + C(MINORITY) + GENDER + LSALBEGIN + SALBEGIN
rsquared_adj: 0.8191689261595675      <- SALARY ~ C(JOBCAT) + GENDER + LSALBEGIN + SALBEGIN
rsquared_adj: 0.8251541213776562      <- SALARY ~ C(EDUC) + C(JOBCAT) + C(MINORITY) + LSALBEGIN + SALBEGIN
rsquared_adj: 0.8252442092330109      <- SALARY ~ C(EDUC) + C(JOBCAT) + LSALBEGIN + SALBEGIN
rsquared_adj: 0.8255631713443903      <- SALARY ~ C(EDUC) + C(JOBCAT) + C(MINORITY) + SALBEGIN
rsquared_adj: 0.8256204234943031      <- SALARY ~ C(EDUC) + C(JOBCAT) + SALBEGIN
rsquared_adj: 0.8258417789266599      <- SALARY ~ C(EDUC) + C(JOBCAT) + GENDER + LSALBEGIN + SALBEGIN
rsquared_adj: 0.8262946570863963      <- SALARY ~ C(EDUC) + C(JOBCAT) + GENDER + SALBEGIN
rsquared_adj: 0.8264595455648863      <- SALARY ~ C(EDUC) + C(JOBCAT) + C(MINORITY) + GENDER + SALBEGIN
```

I am choosing a feature selection that seems suitable for the target prediction. For reasons of units and ease, I am only using **non-log()** combination for now.

```
[261]: # choosing: the 2nd best, because is has one coef. less
#   reason: there seems to be some influence of the GENDER but it appears to be quite small (per thousand range)
# rsquared_adj: 0.8256204234943031 *<- SALARY ~ C(EDUC) + C(JOBCAT) + SALBEGIN
# rsquared_adj: 0.8262946570863963 *<- SALARY ~ C(EDUC) + C(JOBCAT) + GENDER + SALBEGIN

f1 = 'SALARY ~ C(EDUC) + C(JOBCAT) + SALBEGIN'    # 0.8256
f2 = 'LSALARY ~ C(EDUC) + C(JOBCAT) + LSALBEGIN' # 0.8318

# !!! f1 - for now !!!

train_model = smf.ols(formula=f1, data=train)
train_model_fit = train_model.fit()

train_model_fit.params
```

```
[261]: Intercept           6628.670206
C(EDUC)[T.12]       1643.783204
C(EDUC)[T.14]       3633.815913
C(EDUC)[T.15]       4707.838260
C(EDUC)[T.16]       6615.923640
C(EDUC)[T.17]      10941.830944
C(EDUC)[T.18]       8430.570609
C(EDUC)[T.19]      10289.739317
C(EDUC)[T.20]        -10.270960
C(EDUC)[T.21]       -937.772070
C(JOBCAT)[T.2]      3966.086703
C(JOBCAT)[T.3]     11243.592625
SALBEGIN               1.281747
dtype: float64
```


Partial Regression Plot
e(SALARY | X)
e(Intercept | X)
e(C(EDUC)[T.12] | X)
e(C(EDUC)[T.14] | X)
e(C(EDUC)[T.15] | X)
e(C(EDUC)[T.16] | X)
e(C(EDUC)[T.17] | X)
e(C(EDUC)[T.18] | X)
e(C(EDUC)[T.19] | X)
e(C(EDUC)[T.20] | X)
e(C(EDUC)[T.21] | X)
e(C(JOBCAT)[T.2] | X)
e(C(JOBCAT)[T.3] | X)
e(SALBEGIN | X)

First I will computed and plot the targets, using the **train data set**.

```
[280]: train_model_predict = train_model_fit.predict(train);
```

```
[281]: train_model_actual = train['SALARY'].astype(float);
```

```
[282]: plt.figure(dpi = 75);
plt.scatter(train_model_actual, train_model_predict);
plt.plot(train_model_actual, train_model_actual, color="red");
plt.xlabel("Actual Scores");
plt.ylabel("Estimated Scores");
plt.title("Train Data: Actual vs Estimated Scores");
plt.show();
```


Then I will compute the RMSE and other solution quantifiers, using the **train data set**.

```
[278]: print("Mean Absolute Error (MAE)         : {}".format(mean_absolute_error(train_model_actual, train_model_predict)))
print("Mean Squared Error (MSE)          : {}".format(mse(train_model_actual, train_model_predict)))
print("Root Mean Squared Error (RMSE)    : {}".format(rmse(train_model_actual, train_model_predict)))
print("Mean Absolute Perc. Error (MAPE)  : {}".format(np.mean(np.abs((train_model_actual - train_model_predict) / train_model_actual)) * 100))
```

```
Mean Absolute Error (MAE)         : 4762.818258860782
Mean Squared Error (MSE)          : 53924465.39180361
Root Mean Squared Error (RMSE)    : 7343.3279507185025
Mean Absolute Perc. Error (MAPE)  : 13.143136437260466
```

# Stackholder

**The Questions**

- What is the salary based on in this bank?
- Is there an difference in pay between male and female employees?
- Would females with a minority background have the same chances for a fair salary as non-minority women?

# Conclusion to Stakeholder

It could be generalized that the salary is determinant by:

- Group Affiliation
- Gender
- Education

...is this order.

# Conclusion

**What is the salary based on, in this bank?**

- The salary at this bank, is closely correlated to the education and to the type of job.

- There is a also correlation between the salary and the gender as well as the group affiliation of a person.

- Most management position are occupied by male individuals with a non-minority background and an university degree.

- Most of the women are working in administrative positions, where we also find a lower minority quota.

# Conclusion

**Is there an difference in pay between male and female employees?**

- Yes, it was found that the best paid positions are in management, where the quota for females and minorities is very low.

- There is also a group of badly paid males. They occupy custodial jobs, with a high quota of minorities and no women.

- It could be generalized that women get paid less then the man and that minorities get paid less as well.

# Conclusion

**Would females with a minority background have the same chances for a fair salary as non-minority women?**

- No, the possibilities is lower for women to have the same pay as there male colleagues.

- If the female is part of a minority group, the chances for equal pay would become even lower.

# Q & A

Thank you for you attention!