NeueFische - First Project: EDA

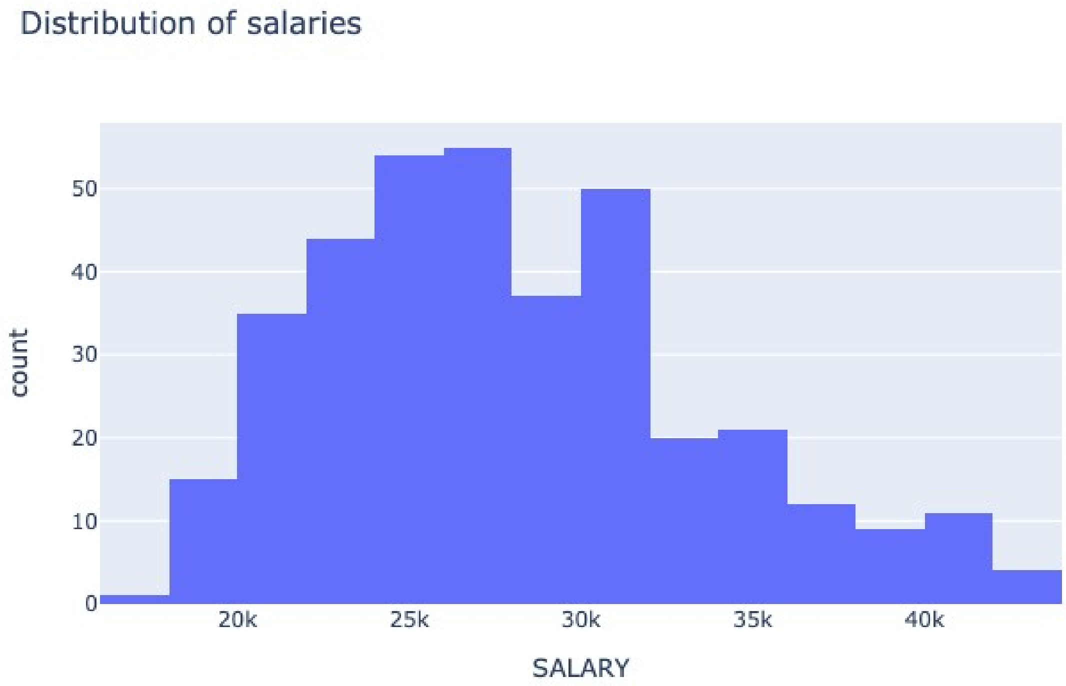
Data

- Target:
  - *Salary*
- *Features:*
  - *Education degree*
  - *Entry wage*
  - *Gender*
  - *Minority*
  - *Job category*

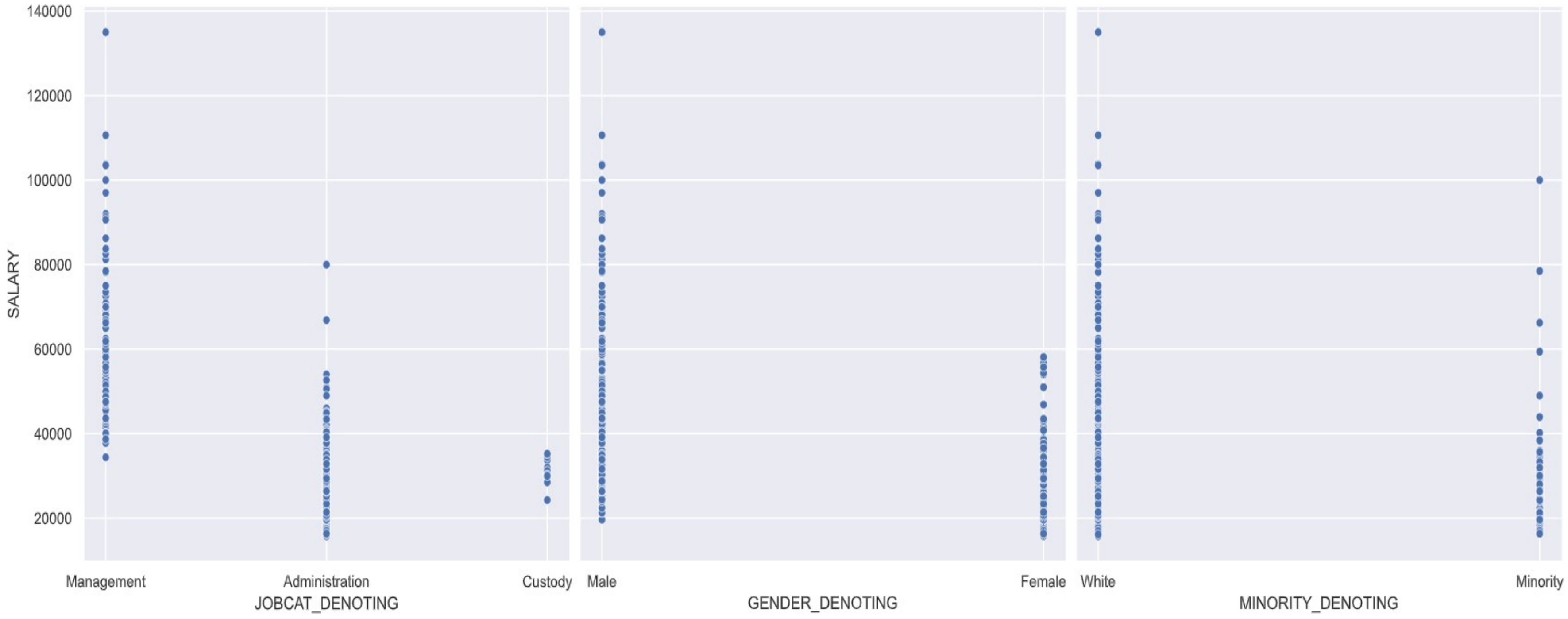
Observations:

- 474

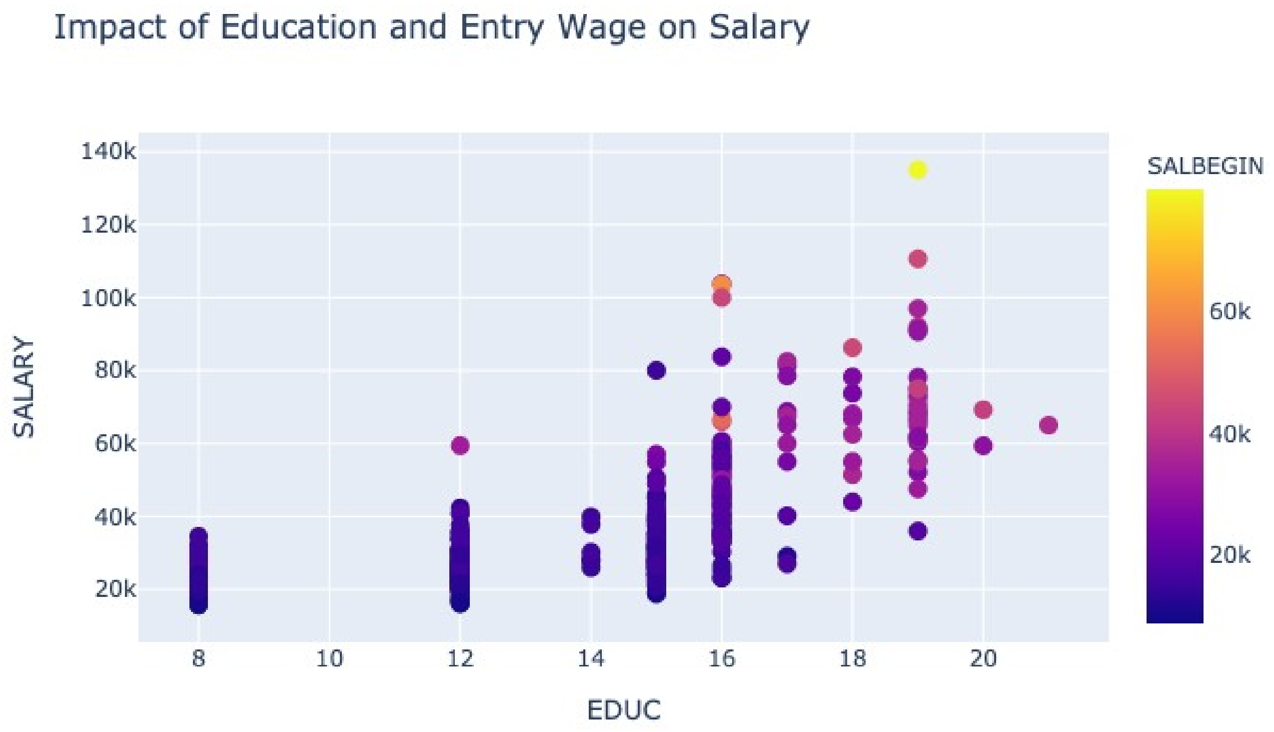
Visualization



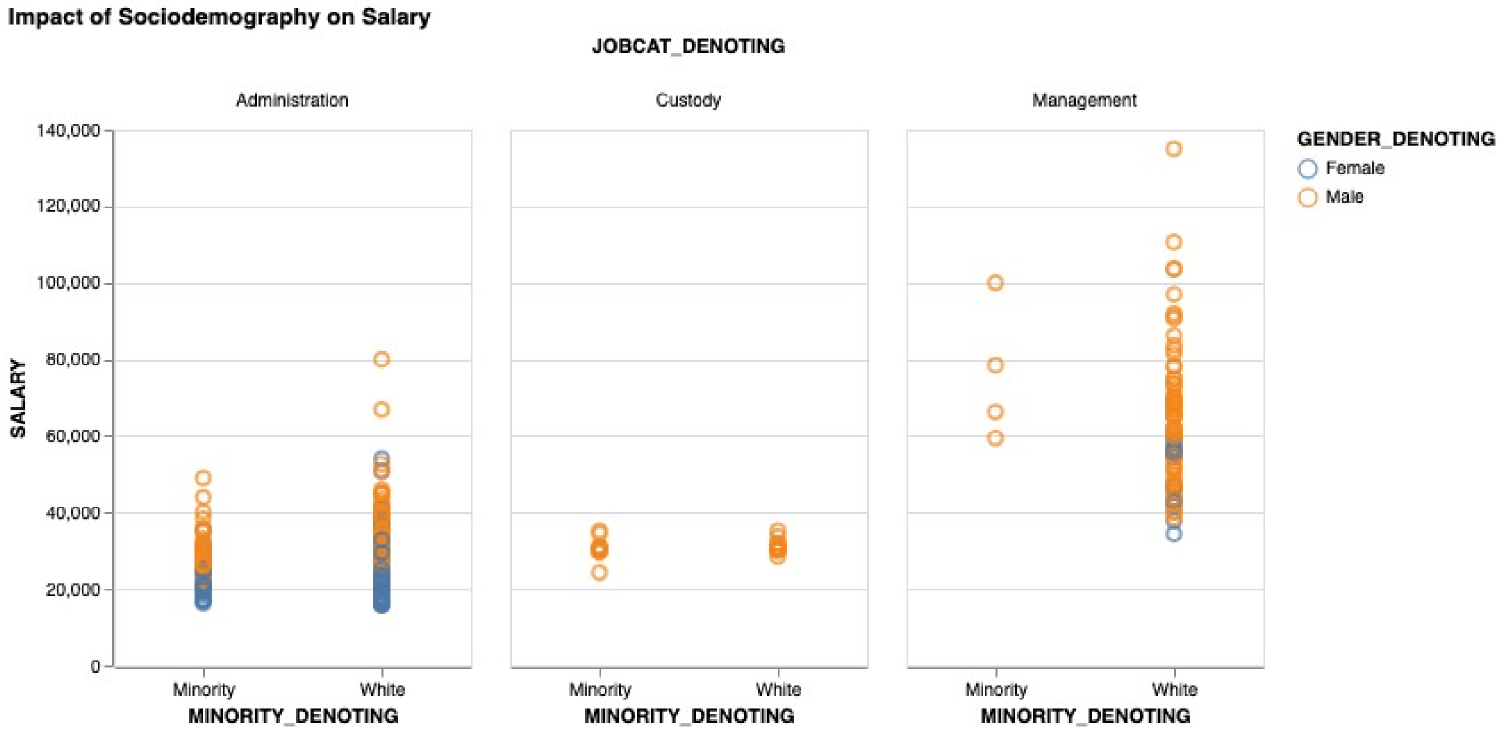
Visualization



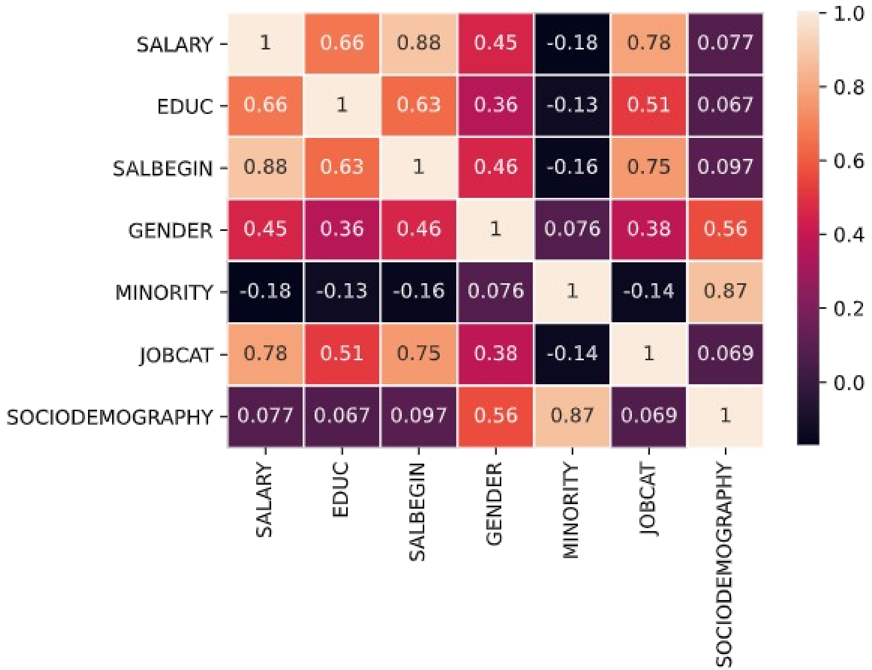
Visualization



Visualization



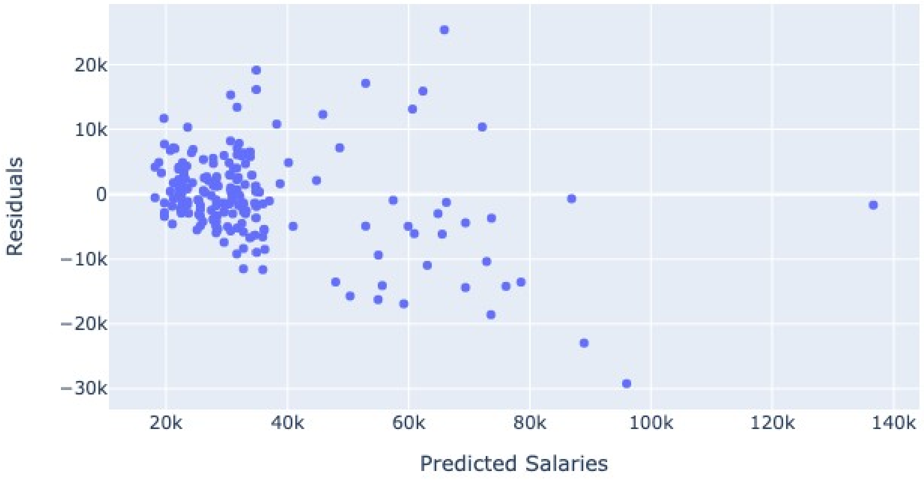
Correlation



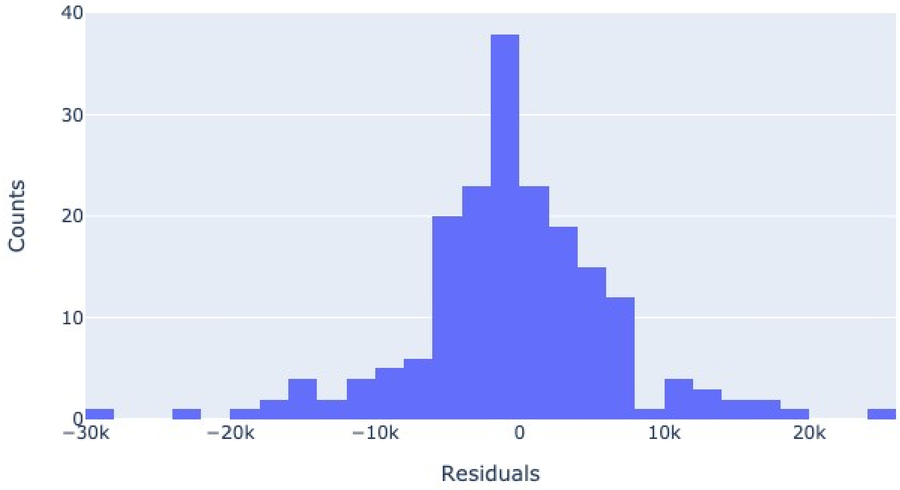
Linear Regression

```
df_X = df[
          ["EDUC",
           "SALBEGIN",
           "GENDER",
           "MINORITY",
           "JOBCAT",
           "SOCIODEMOGRAPHY"
          ]
]

Y = df["SALARY"]

X_train, X_test, y_train, y_test = train_test_split(df_X, Y, test_size=0.4, random_state=17)

model = LinearRegression()

model.fit(X_train,y_train)

predictions = model.predict(X_test)

mean_squared_error(y_test, predictions)**0.5
```

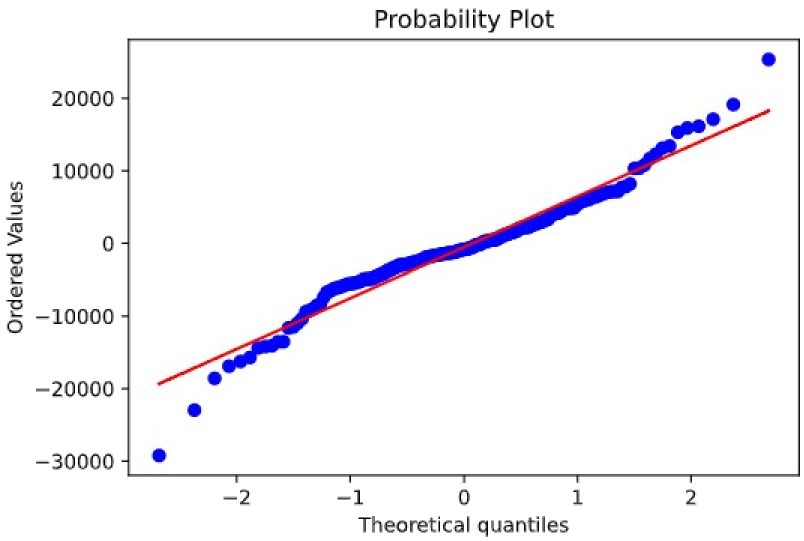
Linear Regression



Linear Regression



Linear Regression



Linear Regression

OLS Regression Results

==

Dep. Variable:	SALARY	R-squared:	0.548
Model:	OLS	Adj. R-squared:	0.538
Method:	Least Squares	F-statistic:	51.90
Date:	Fri, 04 Jun 2021	Prob (F-statistic):	4.31e-35
Time:	16:22:17	Log-Likelihood:	-2133.0
No. Observations:	220	AIC:	4278.
Df Residuals:	214	BIC:	4298.
Df Model:	5		
Covariance Type:	nonrobust		

strong multicollinearity problems or that the design matrix is singular.

RMSE: 7115.8

Linear Regression

	coef	std err	t	P>|t|	[0.025	0.975]
const	5379.0675	1935.169	2.780	0.006	1564.634	9193.501
EDUC	475.6058	130.125	3.655	0.000	219.116	732.096
SALBEGIN	0.9647	0.130	7.402	0.000	0.708	1.222
GENDER	2122.4826	633.179	3.352	0.001	874.417	3370.549
MINORITY	-1232.9478	347.817	-3.545	0.000	-1918.533	-547.362
JOBCAT	2307.3362	768.728	3.002	0.003	792.088	3822.584
SOCIODEMOGRAPHY	-343.4130	216.306	-1.588	0.114	-769.776	82.950

Omnibus:	14.624	Durbin-Watson:	1.851
Prob(Omnibus):	0.001	Jarque-Bera (JB):	15.935
Skew:	0.657	Prob(JB):	0.000346
Kurtosis:	3.114	Cond. No.	8.52e+19

Clean The Data

```
q1 = df.describe().loc["75%", "SALARY"]
q3 = df.describe().loc["25%", "SALARY"]

iqr = q1 - q3

mask = (q1 - 1.5*iqr) <= df["SALARY"]
mask &= df["SALARY"] <= (q3 + 1.5*iqr)
```

Sociodemography

```
GENDER_DENOTING
Female    26031.921296
Male      41441.782946
Name: SALARY, dtype: float64
```

```
GENDER_DENOTING
Female    24300.0
Male      32850.0
Name: 50%, dtype: float64
```

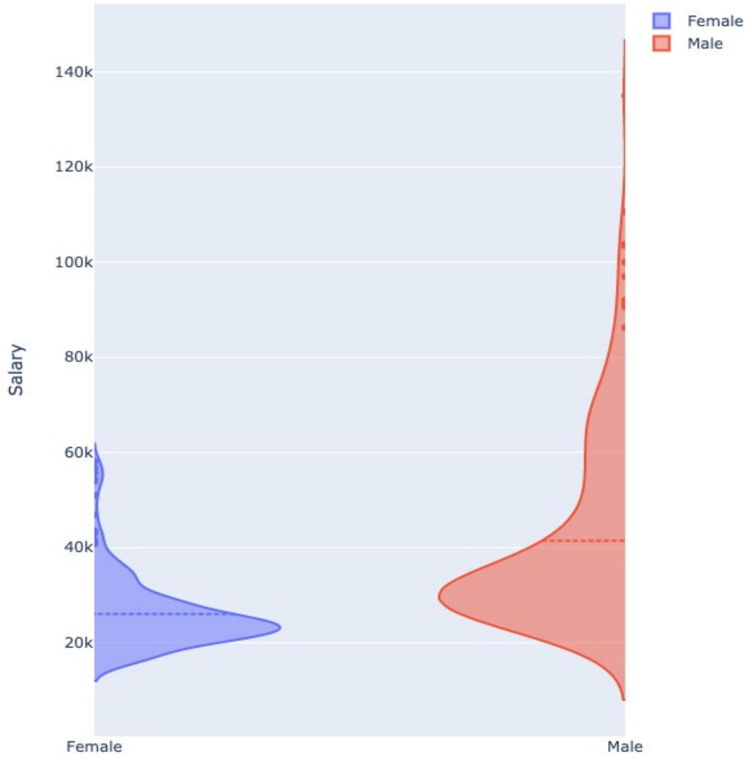
```
MINORITY_DENOTING
Minority    28713.942308
White       36023.310811
Name: SALARY, dtype: float64
```

```
MINORITY_DENOTING
Minority    26625.0
White       29925.0
Name: 50%, dtype: float64
```

```
SOCIODEMOGRAPHY_DENOTING
Minority_Female    23062.500000
Minority_Male      32246.093750
White_Female       26706.789773
White_Male         44475.412371
Name: SALARY, dtype: float64
```

```
SOCIODEMOGRAPHY_DENOTING
Minority_Female    23775.0
Minority_Male      29025.0
White_Female       24450.0
White_Male         36000.0
Name: 50%, dtype: float64
```

Sociodemography



Sociodemography

- RMSE:

SALARY~<allFeatures>	=	7,115.8
SALARY~SALBEGIN	=	8,510.1
SALARY~GENDER	=	15,114.4

Insights

- Management earns the highest salary
- The vast majority of the management ain't female nor a minority

Insights

- The lesser the education degree
  -> the more narrow the constricted range

- If You already start with a higher salary
  -> You will later get an even higher salary

- The linear regression model is not the best model to explain gender or minority pay gaps.

Merci!