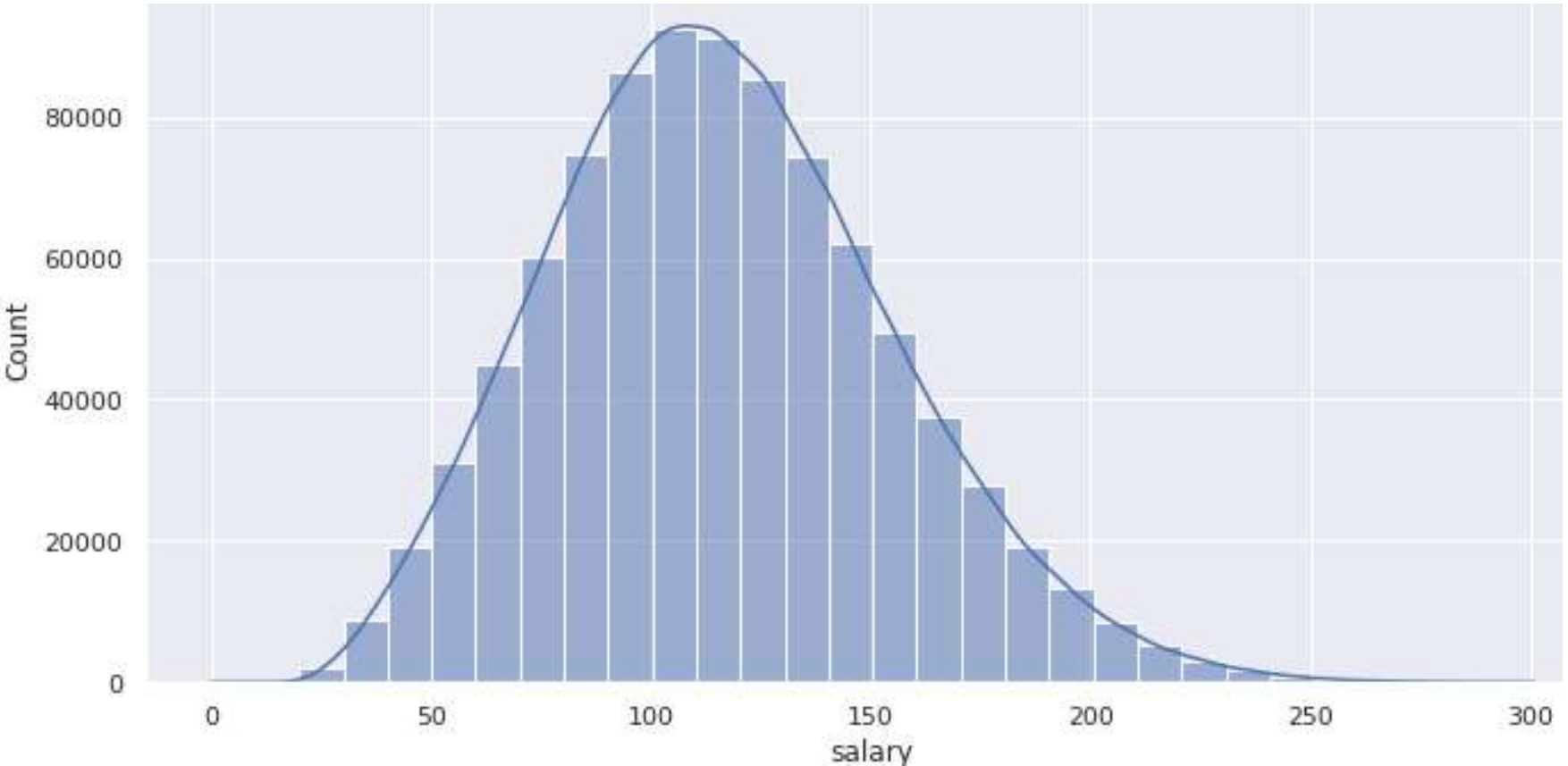
Job Posting Salaries Predictions

Submitted By : MHD KHAIR SULTAN

Data Exploration & preparation

Data set Distribution



Steps taken in Data Preprocessing

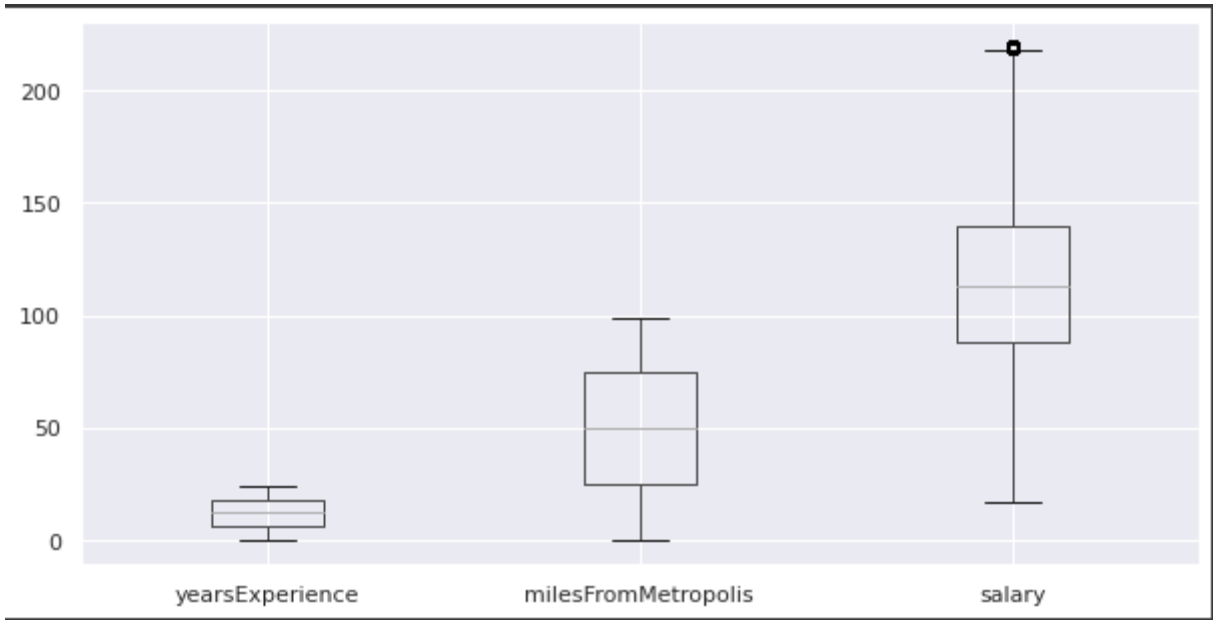
- Performing EDA
- handling missing values
- Handling outliers
- Check for the correlations
- Encoding nominal features using one hot encoding
- Encoding categorical features using label encoding
- Remove unrelated features
- Saving the cleaned csv file

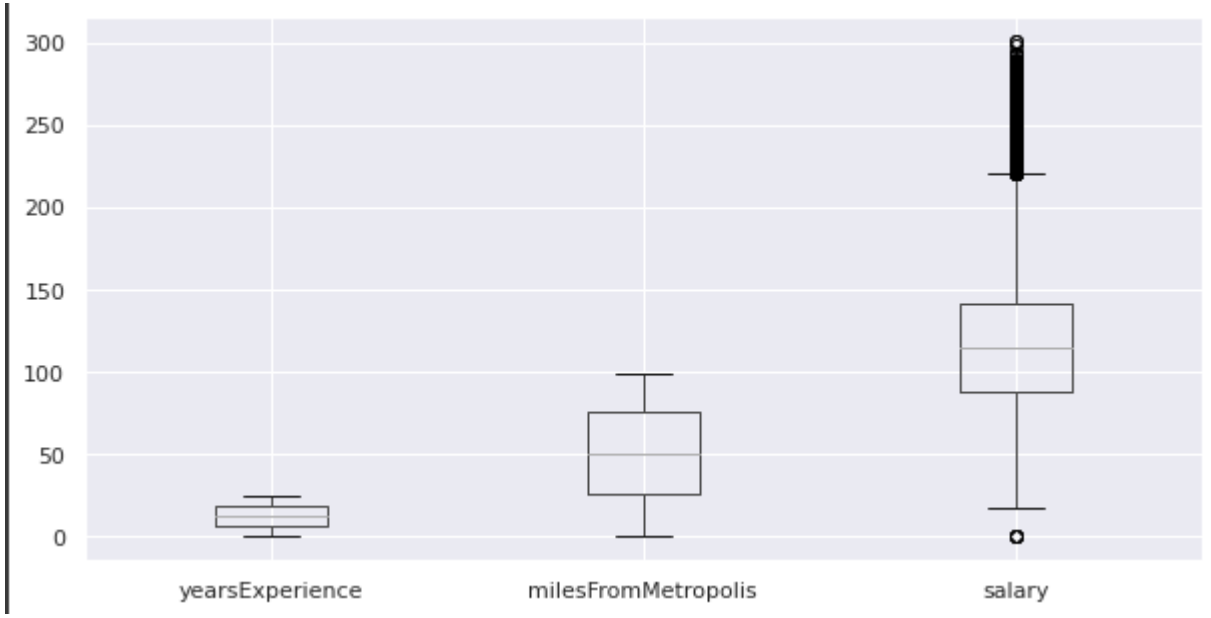
Column	jobId	companyId	jobType	degree	major	industry	yearsExperience	milesFromMetropolis	salary
Unique Values	900,000	63	8	5	9	7	25	100	280
Is Categorical	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No

Statistics	yearsExperience	milesFromMetropolis	salary
count	900000	900000	900000
mean	11.991183	49.525906	116.067520
std	7.211222	28.883348	38.717799
min	0	0	0
max	24	99	301
25%	6	25	88
50%	12	50	114
75%	18	75	141

Shape of the Data Set	Outliers	Missing Values	Categorical Features
(900000, 9)	6,406	0	6

Handling Outliers using Inter Quartile Range approach





IQR = Q3 - Q1

Where:
Q_3 = median of the n largest values
Q1= median of the n smallest values

Handling the Categorical Features

- Using one hot encoding for nominal Features
- Using Label Encoder for Ordinal Features

Attribute	jobId	companyId	jobType	degree	major	industry
Ordinal	No	No	No	Yes	NO	NO

- Drop the following columns after preprocessing

Removed Attribute	jobId	companyId	major	industry	jobType

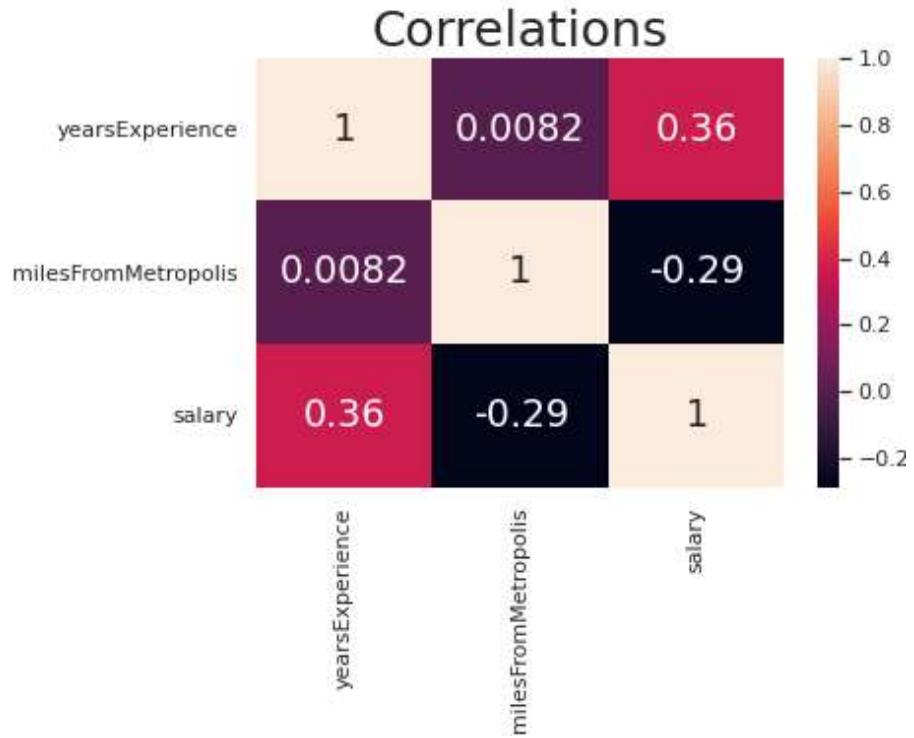
Data Set New Shape
(893594, 28)

Columns After Date Preprocessing

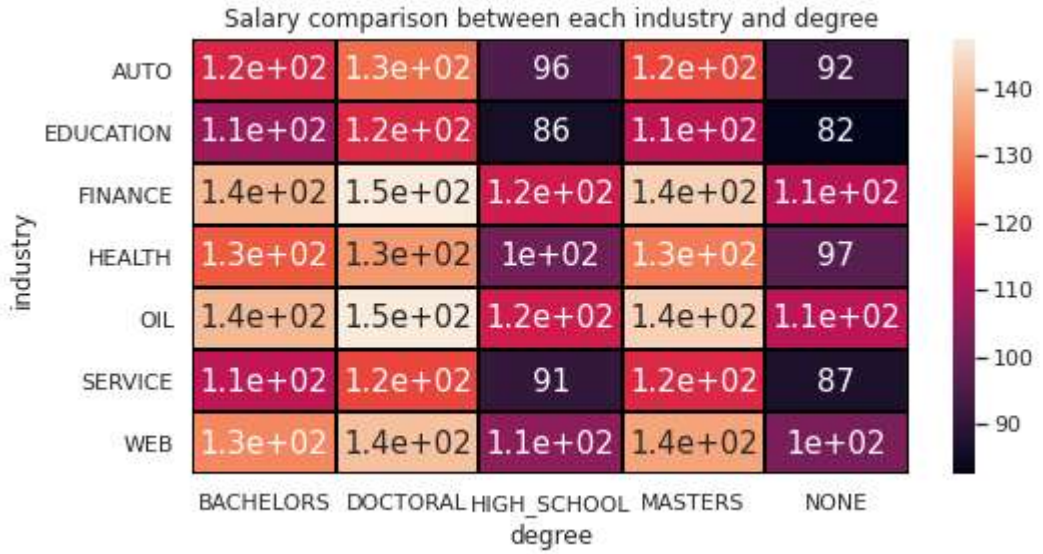
```
 #   Column
---  ------
 0   degree
 1   yearsExperience
 2   milesFromMetropolis
 3   salary
 4   major_BIOLOGY
 5   major_BUSINESS
 6   major_CHEMISTRY
 7   major_COMPSCI
 8   major_ENGINEERING
 9   major_LITERATURE
 10  major_MATH
 11  major_NONE
 12  major_PHYSICS
 13  industry_AUTO
 14  industry_EDUCATION
 15  industry_FINANCE
 16  industry_HEALTH
 17  industry_OIL
 18  industry_SERVICE
 19  industry_WEB
 20  jobType_CEO
 21  jobType_CFO
 22  jobType_CTO
 23  jobType_JANITOR
 24  jobType_JUNIOR
 25  jobType_MANAGER
 26  jobType_SENIOR
 27  jobType_VICE_PRESIDENT
```

Variables Relationships

- There is a negative correlation between salary and miles from metro



- The most paid people those whom obtain Masters or Doctoral Degrees specially in Oil and Finance Industries we can say they earn more than 110 Money unit



Data Modeling

Data Modeling

Steps :

1. Split the data set into training and testing sets
2. Check multi-correlanity
3. Normalizing data set`s values
4. Hyper parameters tuning using (cross validation and Grid search methods)
5. Evaluation
6. results

Set	X_train	X_test	Y_train	Y_test
Shape	(714875, 27)	(178719, 27)	(714875,)	(178719,)

Results	Lasso	Ridge	Normal Linear Regression with regularization
R^2 Training	0.738887	0.738888	0.738885
R^ Testing	0.738106	0.738107	0.738116
MSE	-367.537787	-368.59774	-368.585335
Alpha	0.01	7.96	NA
Intercept	115.230999	115.230999	-830351156556627.11

OLS Regression Results

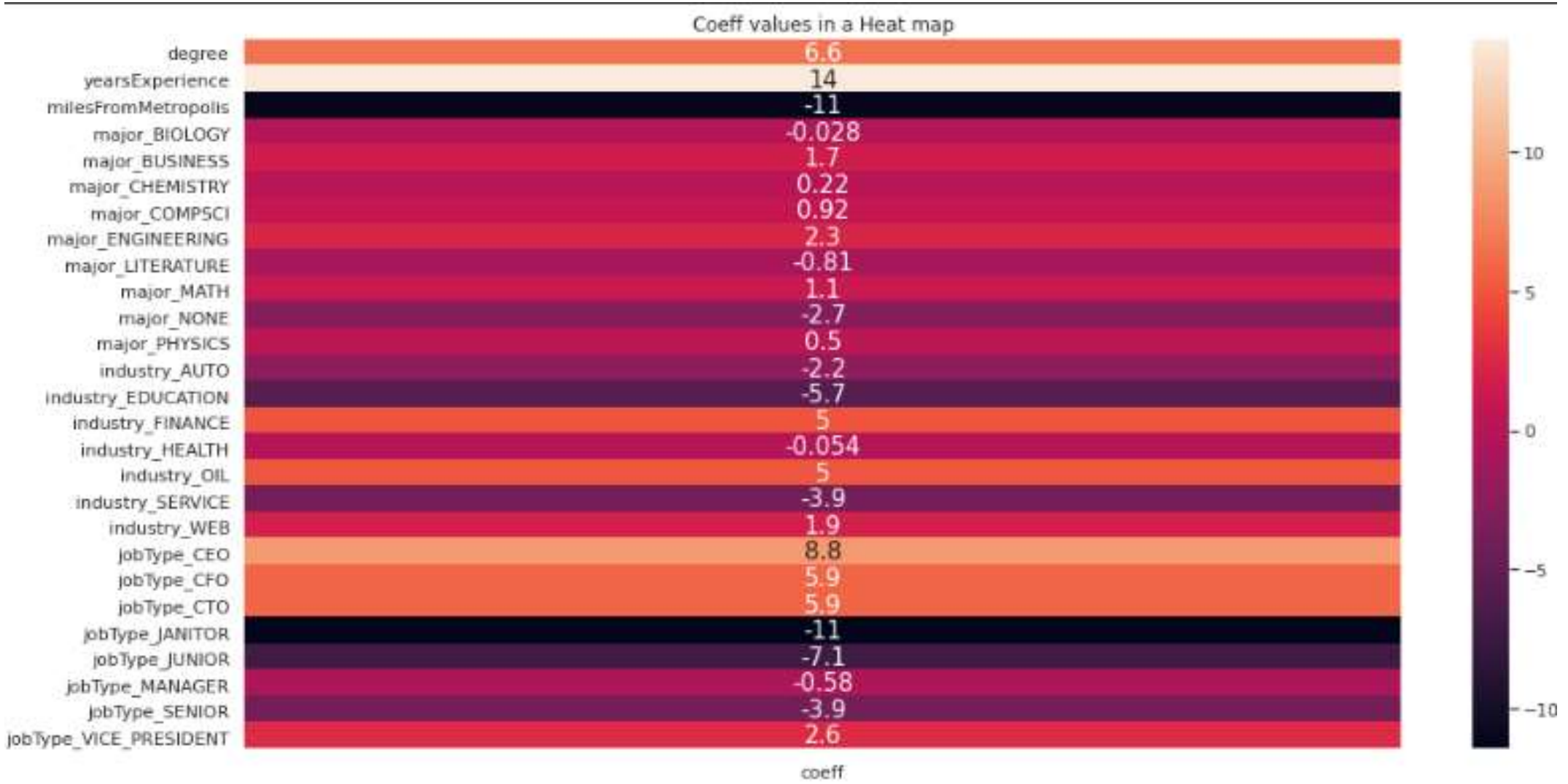
Dep. Variable:	salary	R-squared:	0.739
Model:	OLS	Adj. R-squared:	0.739
Method:	Least Squares	F-statistic:	1.053e+05
Date:	Sun, 15 Aug 2021	Prob (F-statistic):	0.00
Time:	05:51:37	Log-Likelihood:	-3.9073e+06
No. Observations:	893594	AIC:	7.815e+06
Df Residuals:	893569	BIC:	7.815e+06
Df Model:	24		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	77.2680	0.060	1277.510	0.000	77.149	77.387
degree	4.6711	0.023	203.701	0.000	4.626	4.716
yearsExperience	1.9536	0.003	693.094	0.000	1.948	1.959
milesFromMetropolis	-0.3864	0.001	-549.412	0.000	-0.388	-0.385
major_BIOLOGY	6.3521	0.080	79.740	0.000	6.196	6.508
major_BUSINESS	13.4682	0.080	168.508	0.000	13.312	13.625
major_CHEMISTRY	7.3702	0.079	92.874	0.000	7.215	7.526
major_COMPSCI	10.3160	0.080	129.214	0.000	10.160	10.473
major_ENGINEERING	16.0836	0.080	200.835	0.000	15.927	16.241
major_LITERATURE	2.9141	0.079	36.727	0.000	2.759	3.070
major_MATH	11.2135	0.080	139.741	0.000	11.056	11.371
major_NONE	0.9463	0.054	17.614	0.000	0.841	1.052
major_PHYSICS	8.6038	0.080	107.905	0.000	8.448	8.760
industry_AUTO	4.7187	0.050	93.815	0.000	4.620	4.817
industry_EDUCATION	-5.1745	0.050	-102.804	0.000	-5.273	-5.076
industry_FINANCE	25.2228	0.051	498.460	0.000	25.124	25.322
industry_HEALTH	10.8669	0.050	215.610	0.000	10.768	10.966
industry_OIL	25.3191	0.051	499.836	0.000	25.220	25.418
industry_SERVICE	-0.2140	0.050	-4.251	0.000	-0.313	-0.115
industry_WEB	16.5291	0.050	327.768	0.000	16.430	16.628
jobType_CEO	36.4325	0.055	665.248	0.000	36.325	36.540
jobType_CFO	27.3159	0.055	501.027	0.000	27.209	27.423
jobType_CTO	27.3415	0.054	502.717	0.000	27.235	27.448
jobType_JANITOR	-25.0298	0.058	-431.276	0.000	-25.144	-24.916
jobType_JUNIOR	-12.0397	0.054	-222.034	0.000	-12.146	-11.933
jobType_MANAGER	7.7899	0.054	143.700	0.000	7.684	7.896
jobType_SENIOR	-2.0757	0.054	-38.431	0.000	-2.182	-1.970
jobType_VICE_PRESIDENT	17.5333	0.054	323.441	0.000	17.427	17.640

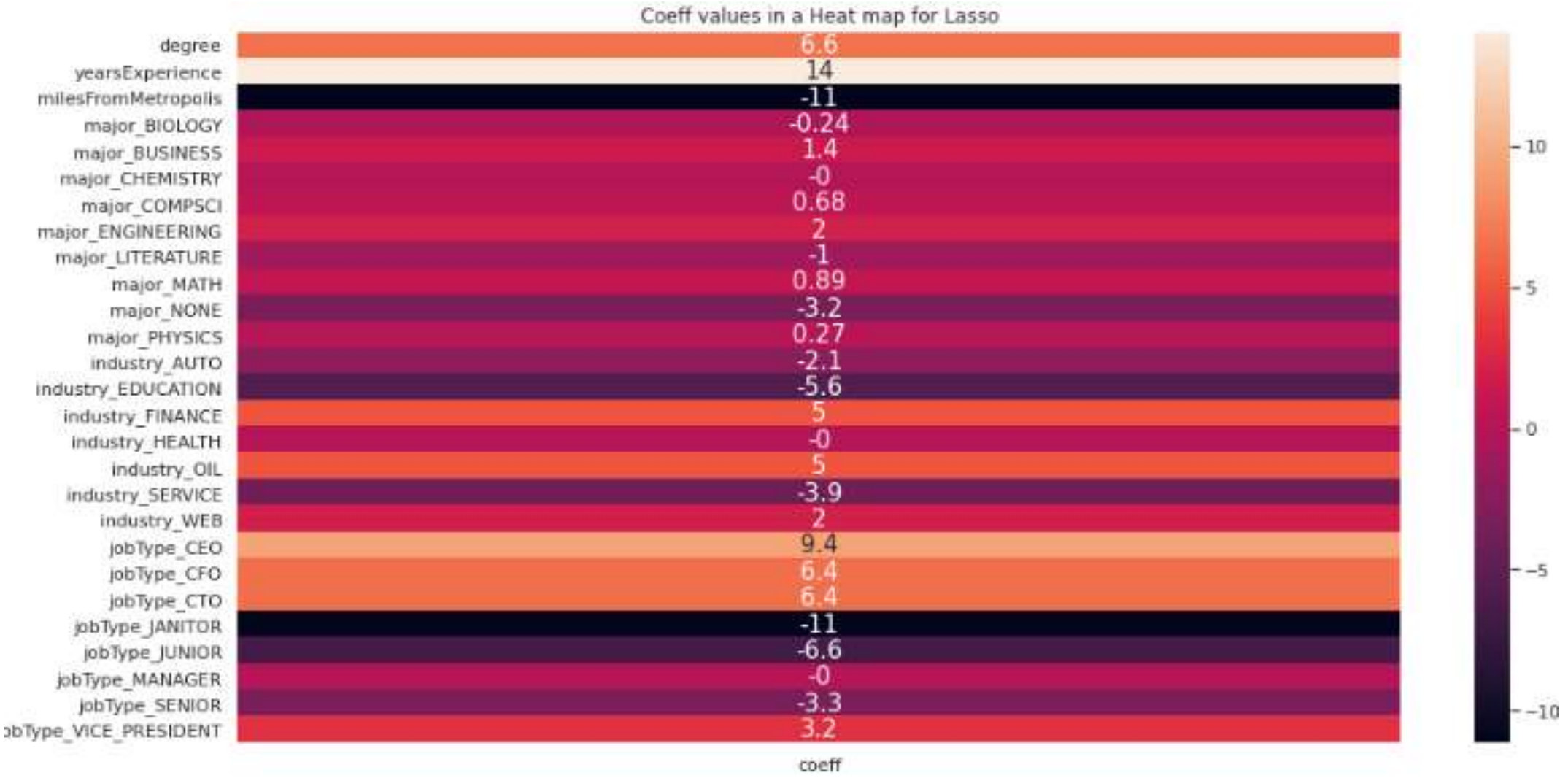
Omnibus:	9345.955	Durbin-Watson:	2.001
Prob(Omnibus):	0.000	Jarque-Bera (JB):	7882.359
Skew:	0.169	Prob(JB):	0.00
Kurtosis:	2.688	Cond. No.	2.08e+17

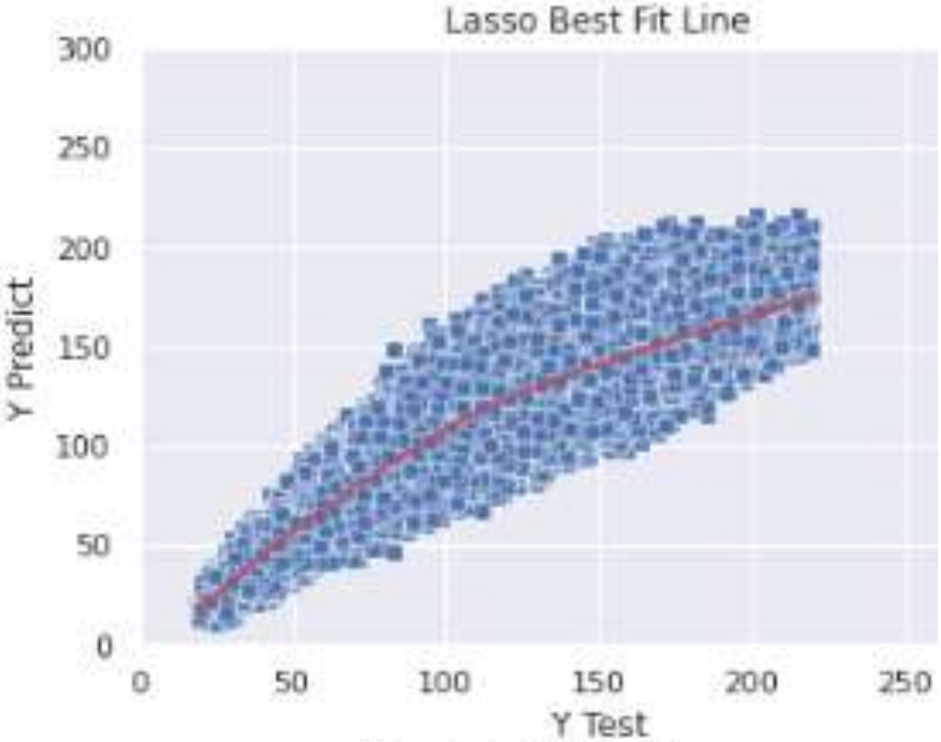
Checking multi-Correnality

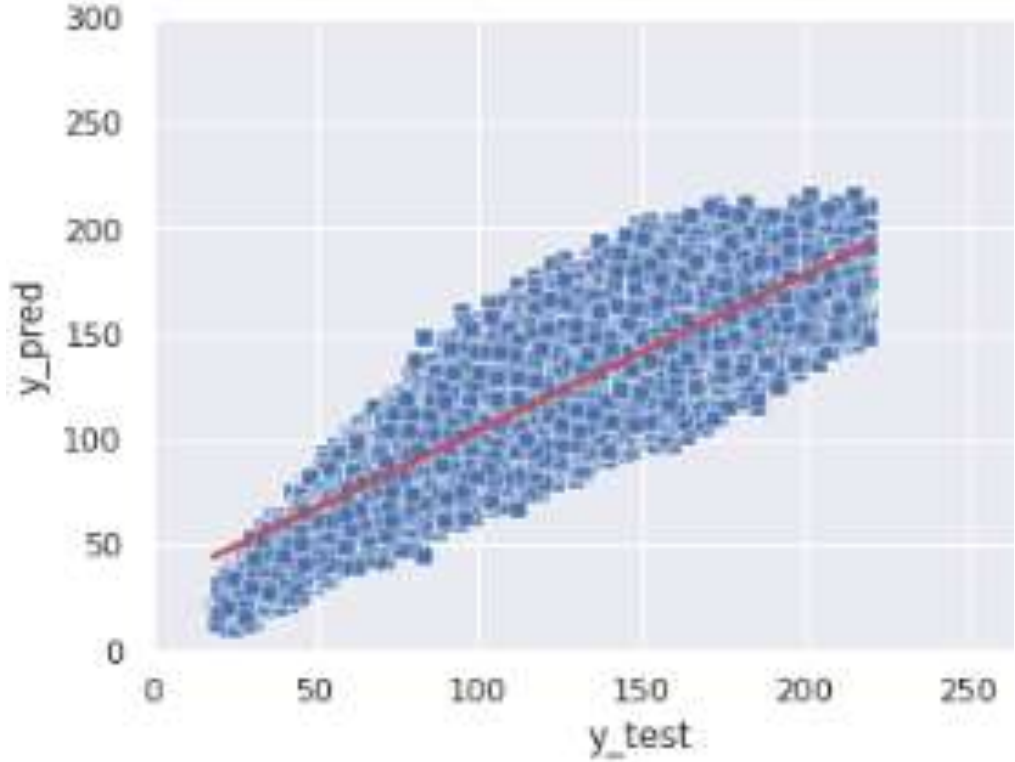
Coefficients comparisons (Ridge)

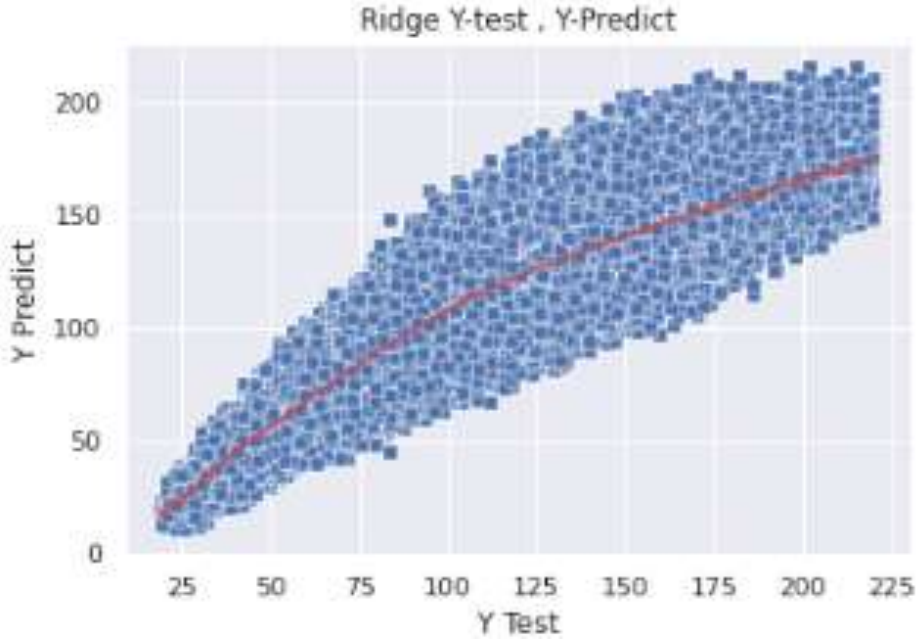


Coefficients comparisons (Lasso)

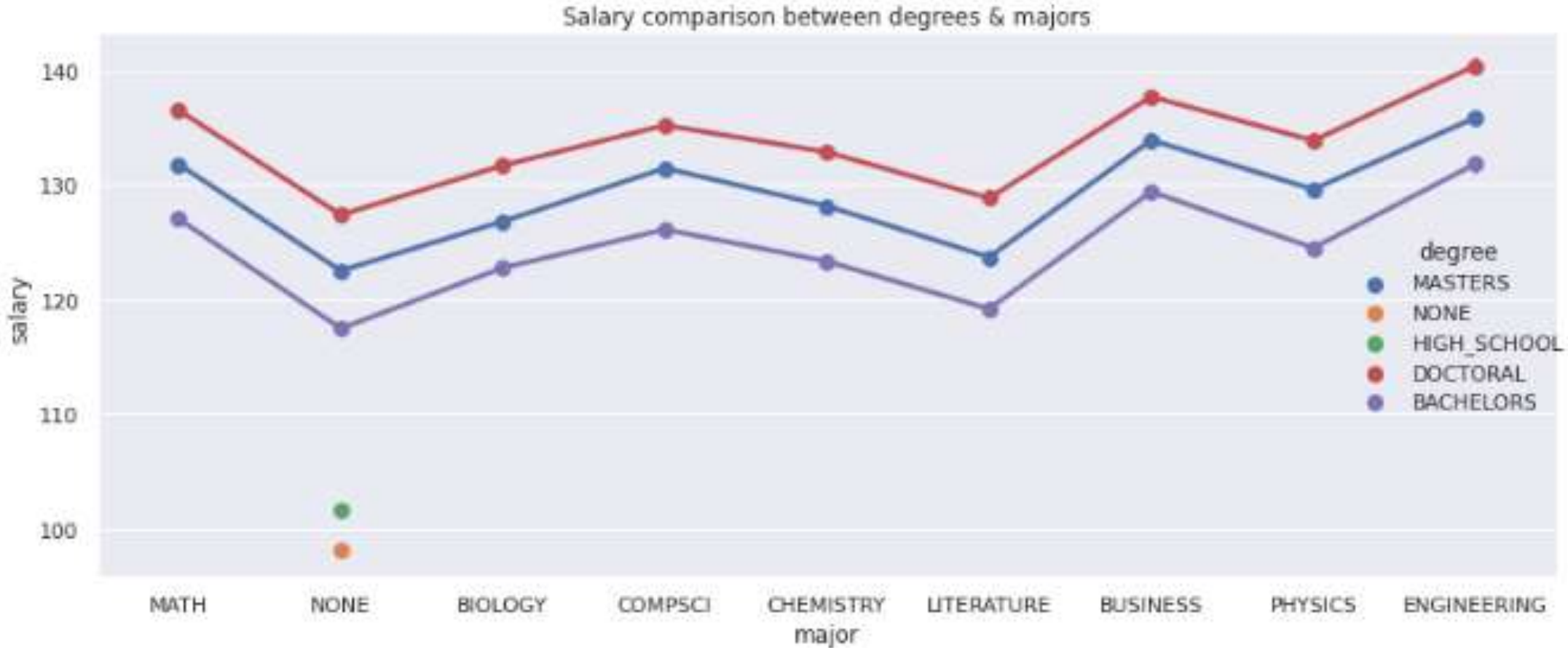



Lasso Best Fit Line
Y Predict
Y Test
0
50
100
150
200
250
300

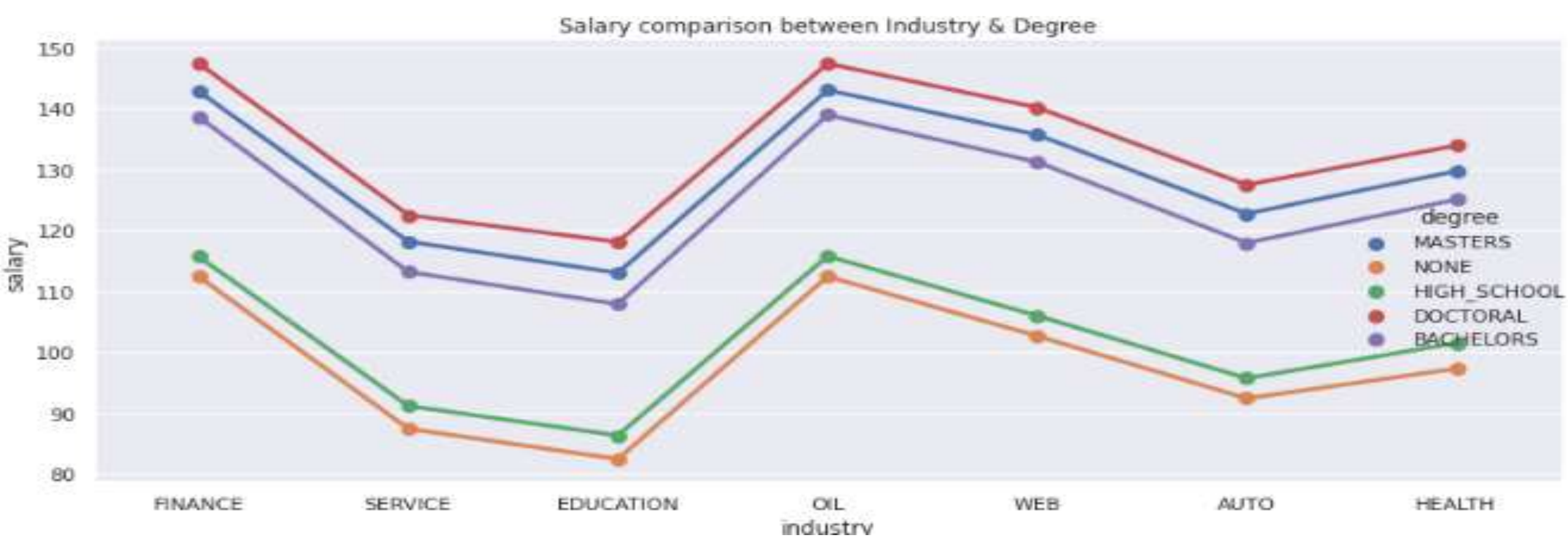

y_pred
y_test
0
50
100
150
200
250
300


Ridge Y-test , Y-Predict
Y Predict
Y Test
0
50
100
150
200
25
75
125
175
225

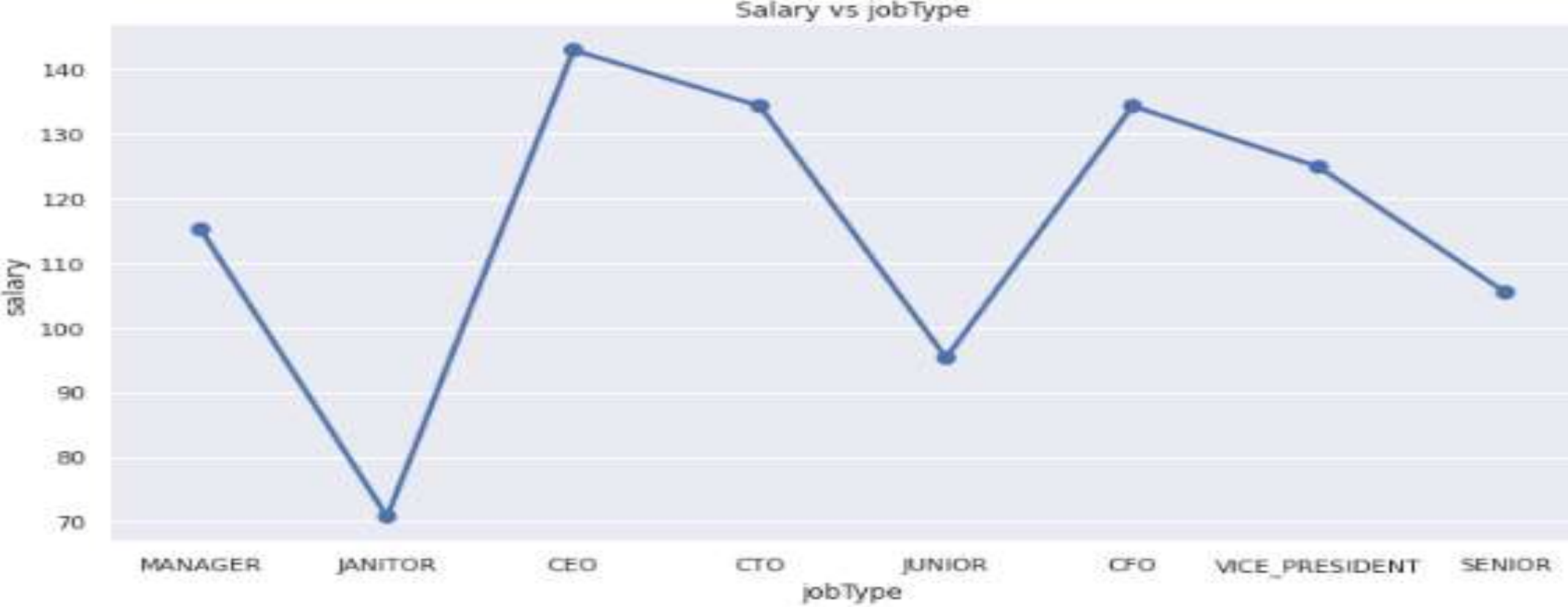
Key findings



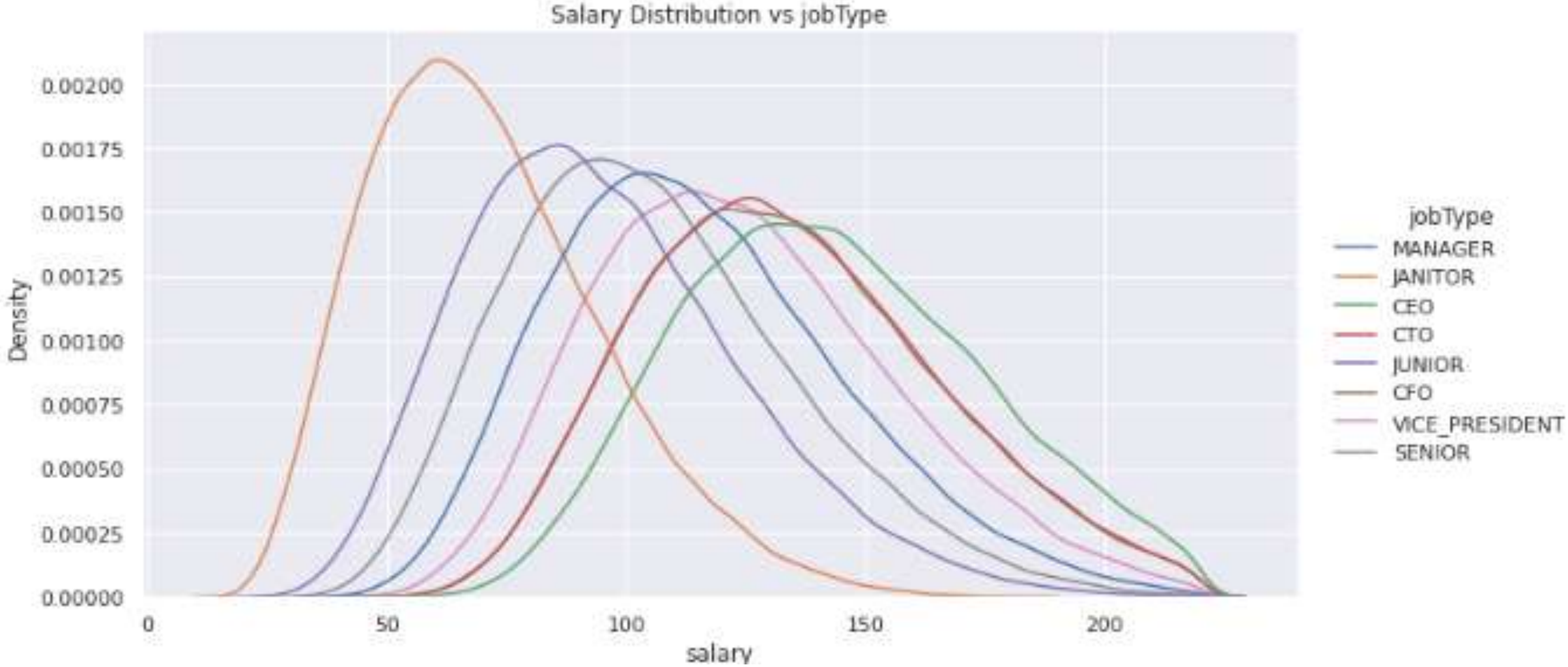
- People with none or high school degrees can only work for None Major and they earn less than 110 salary unit
- People with masters and doctoral earns can work for any major and earns more than 125 salary unit
- They make more money in business and engineering majors
- People with bachelors degree can work for any major and they earn more than 110 salary unit and they earn less
- Than people whome holds master or doctoral degrees



- People with none or high school can earn more money in oil industry and less money in education
- People with bachelors ,masters and doctoral earns the most between others whom hold less degrees and
- they make more money In oil and finance industries.
- People whom works in oil and finance and web earns more money



- People with job type CEO earns the most money
- People with job type janitor earns the least money between all job types
- People with job type CTO and CFO they are earning almost the same amount of money



- most of people works as janitors and junior
- Few people can work as CEO

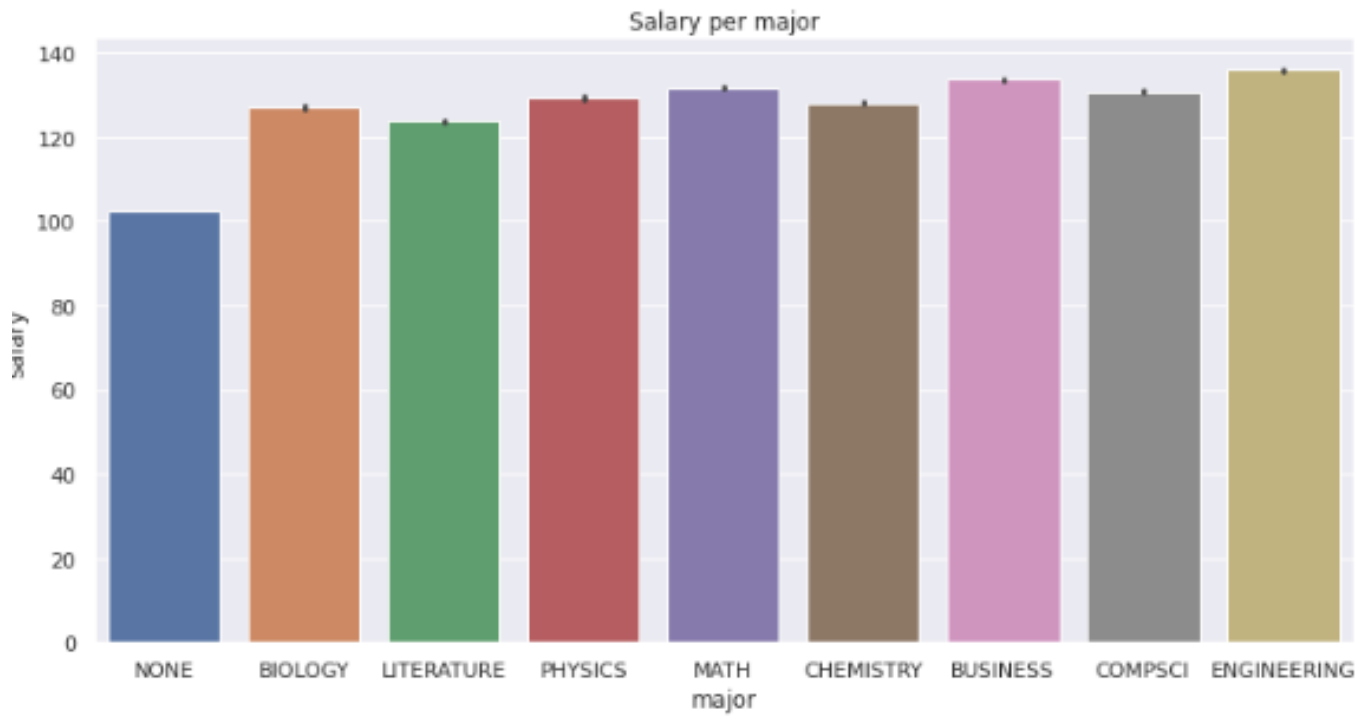
Working in the biology major can make more money than Literature

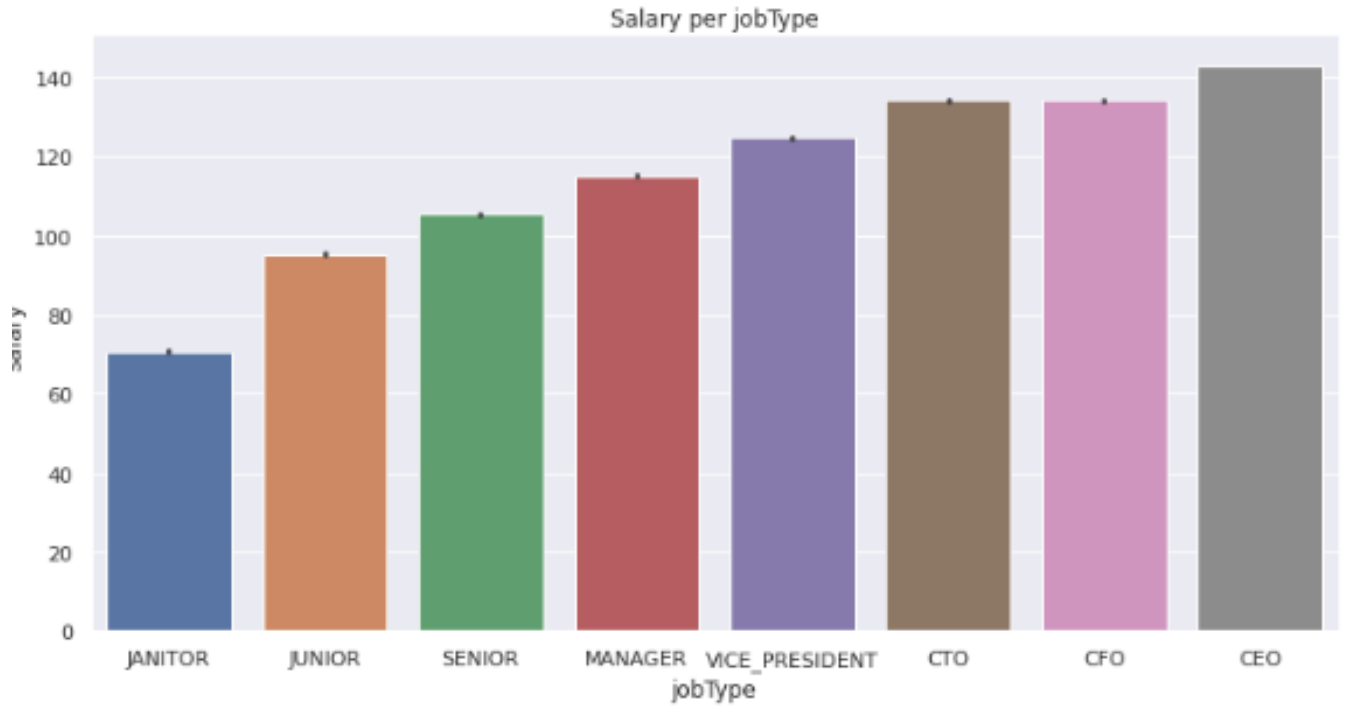
The most three paid people by major are:

1. Engineering
2. Business
3. Computer Science

The most three paid people by job type are:

1. CEO
2. CFO
3. CTO





Next Step

Steps to improve the models accuracy :

- **Using Principle component analysis to reduce number of features**
- **Using more parameters and iterations to improve the models using Grid search methods**
- **Using none linear model**
- **Using SGD Regressor or another approach of gradient decent like Stochastic Gradient Descent**

Thank You