CLARISSA CHEVALIER

Salary Analysis

VERSION 1.0 - 2020

The assigment was to build a model to predict the salaries for the job postings contained in test_features.csv.

The output of the system should be a CSV file entitled test_salaries.csv where each row has the following format: jobId, salary

01

Clean data

First we expect to learn if our data has been sanitized or are there any missing values, corrupt data or correct data types that we are supposed to deal with

02

EDA

Our presentation will focus on the discoveries we'll have during the exploratory data analsys. Our main goal is to extract intel to make the best possible model and discuss throughtly all the details that make up for a good salary predictor with the info we have access

03

Hypotesis

We end the first part of our project and therefore our presentation outlining some hypotesis that could be gather during EDA, baseline metrics and other comments on things that could help our modeling

FEATURES OF THE DATAFRAME

	jobId	companyId	jobType	degree	major	industry	yearsExperience	milesFromMetropolis	salary
0	JOB1362684407687	COMP37	CFO	MASTERS	MATH	HEALTH	10	83	130
1	JOB1362684407688	COMP19	CEO	HIGH_SCHOOL	NONE	WEB	3	73	101
2	JOB1362684407689	COMP52	VICE_PRESIDENT	DOCTORAL	PHYSICS	HEALTH	10	38	137
3	JOB1362684407690	COMP38	MANAGER	DOCTORAL	CHEMISTRY	AUTO	8	17	142
4	JOB1362684407691	COMP7	VICE_PRESIDENT	BACHELORS	PHYSICS	FINANCE	8	16	163

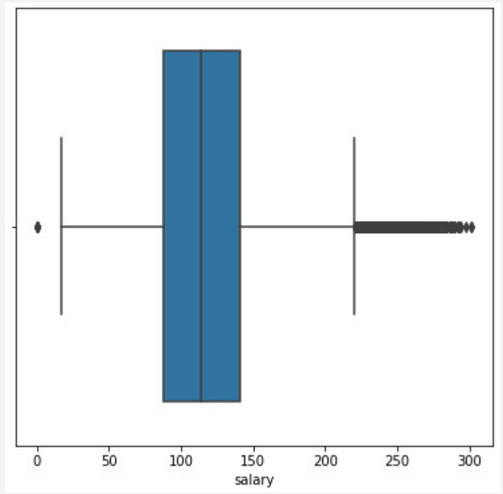
The heading containing all the features on the raw data

```
jobId                1000000 non-null object
companyId            1000000 non-null object
jobType              1000000 non-null object
degree               1000000 non-null object
major                1000000 non-null object
industry             1000000 non-null object
yearsExperience      1000000 non-null int64
milesFromMetropolis  1000000 non-null int64
salary               1000000 non-null int64
```

The datatypes sugest that there's more categorical variables than the numerical ones, wich has to be noted

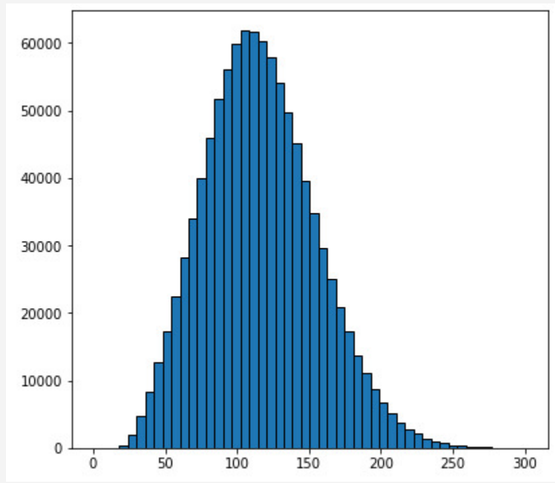
Also, our findings shown no initial problems of null values or invalid data

THE LABEL



01

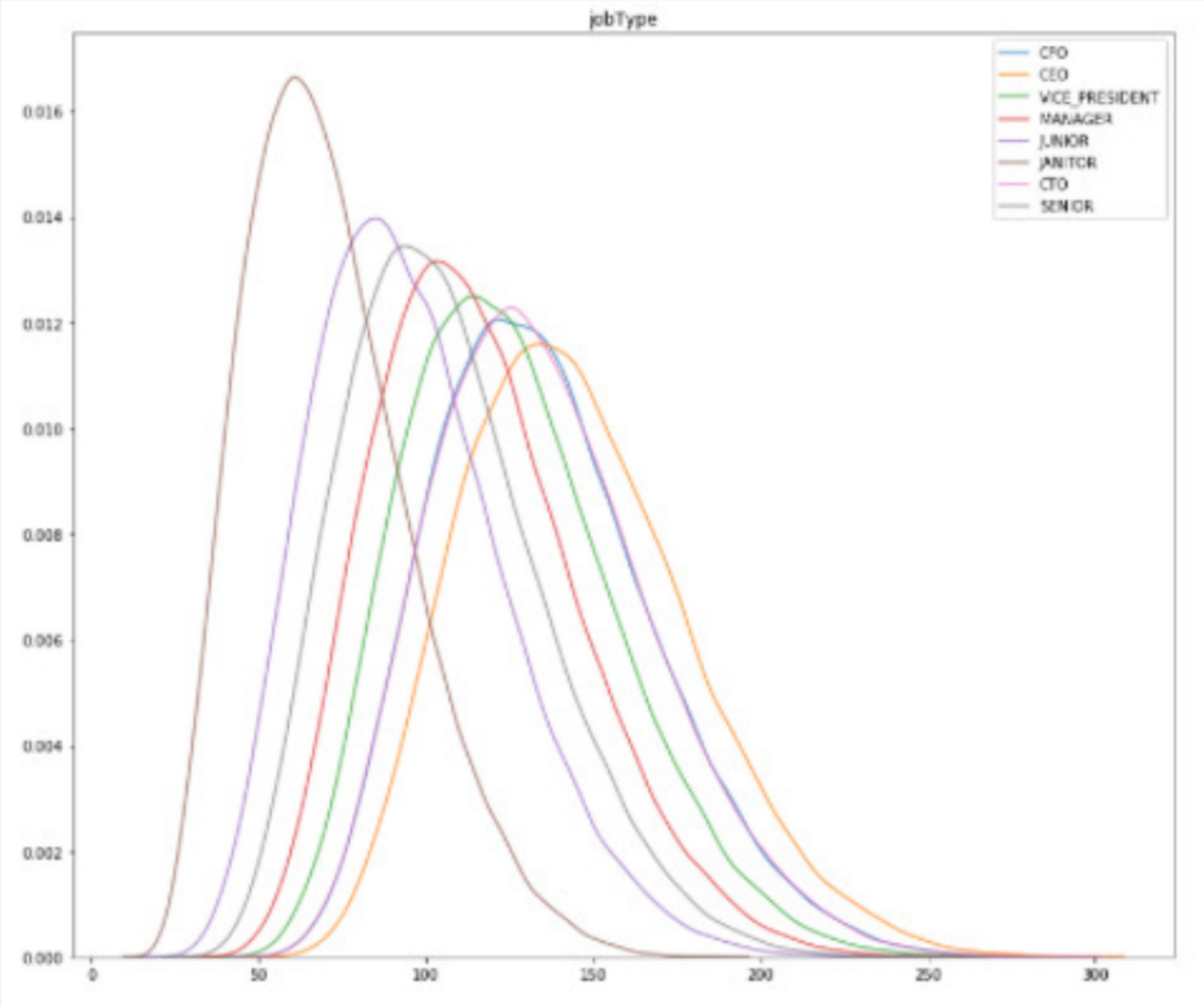
Boxplot showing a lot of dispersion inside the distribution of the output



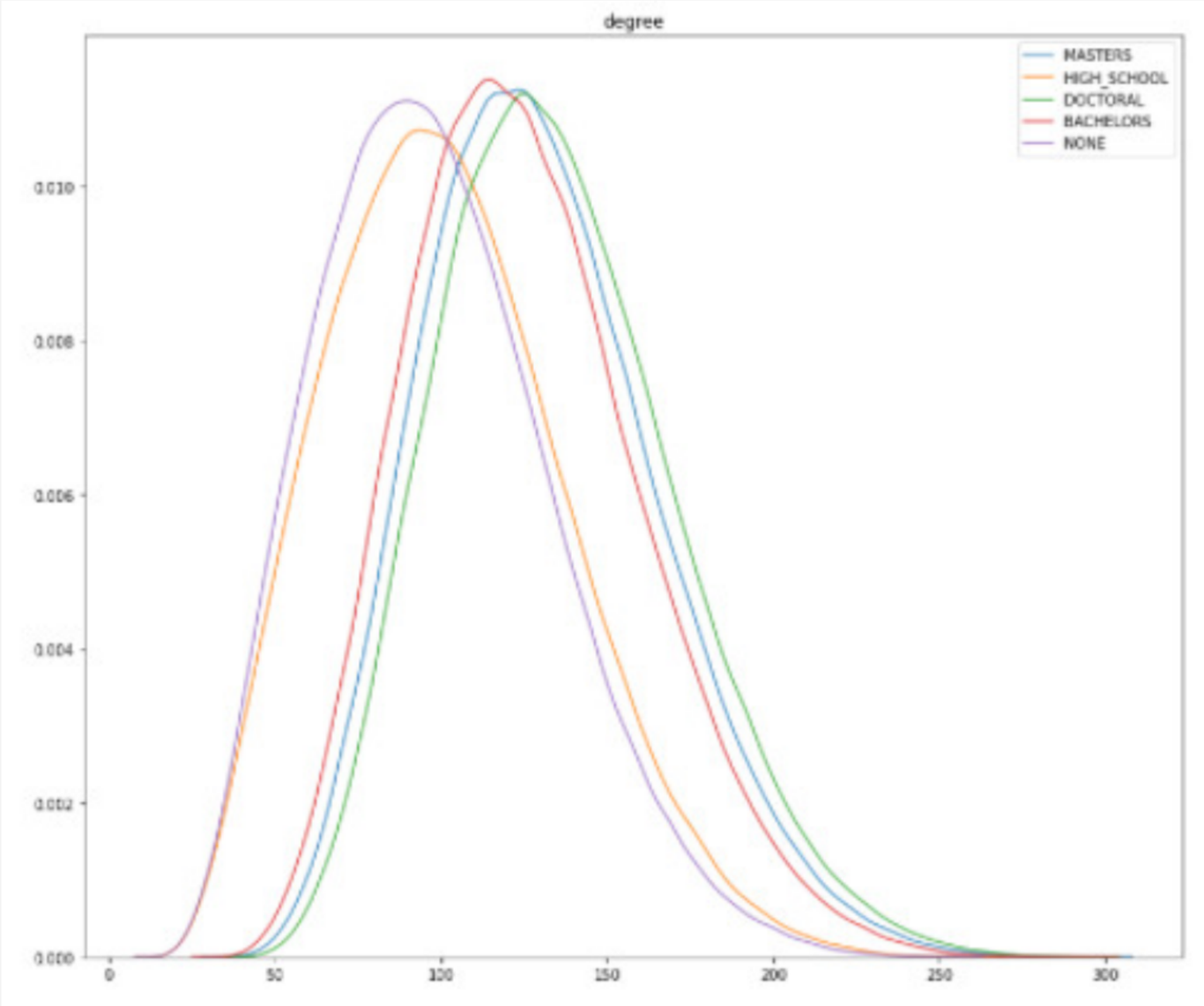
02

The histogram shows a right-skewed inclination a mean close to 100 a year

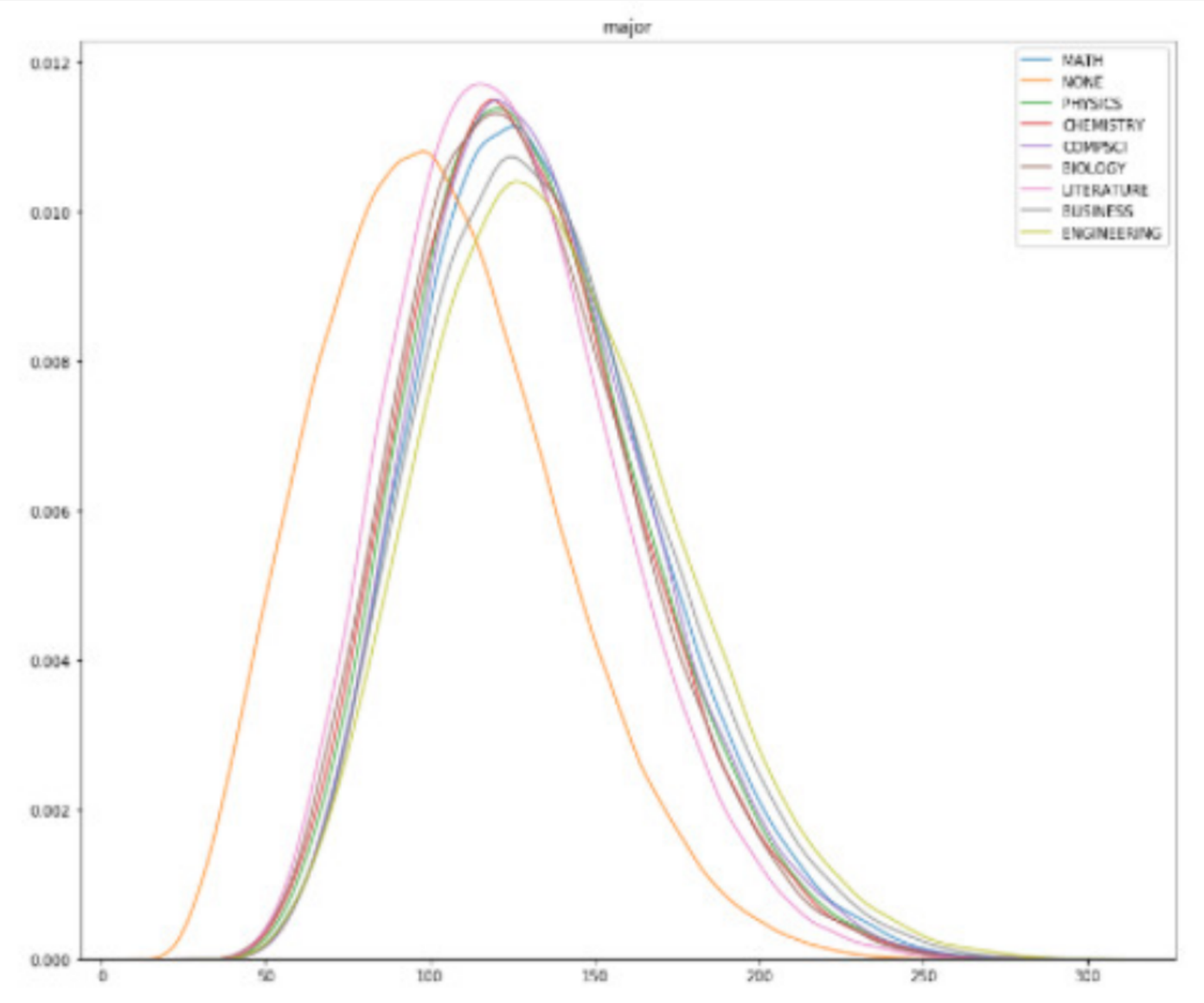
AFTER CAREFULLY EXAMINING THE VALUES OF THE TARGET THAT STOOD OUTSIDE THE SUPERIOR AND INFERIOR MARGIN FOR QUARTILES, NO CONCLUSIONS COULD BE DRAWN FROM THE SUPERIOR BOUND SINCE MULTIPLE FACTORS APPLY AND DISCARDING THE DATA SEEMED PRECOCIOUS. HOWEVER IT'S CLEAR THAT SALARIES OF 0 ARE JUST CORRUPT DATA AND THEREFORE SHOULD BE EXCLUDED FROM OUR ANALYSIS.



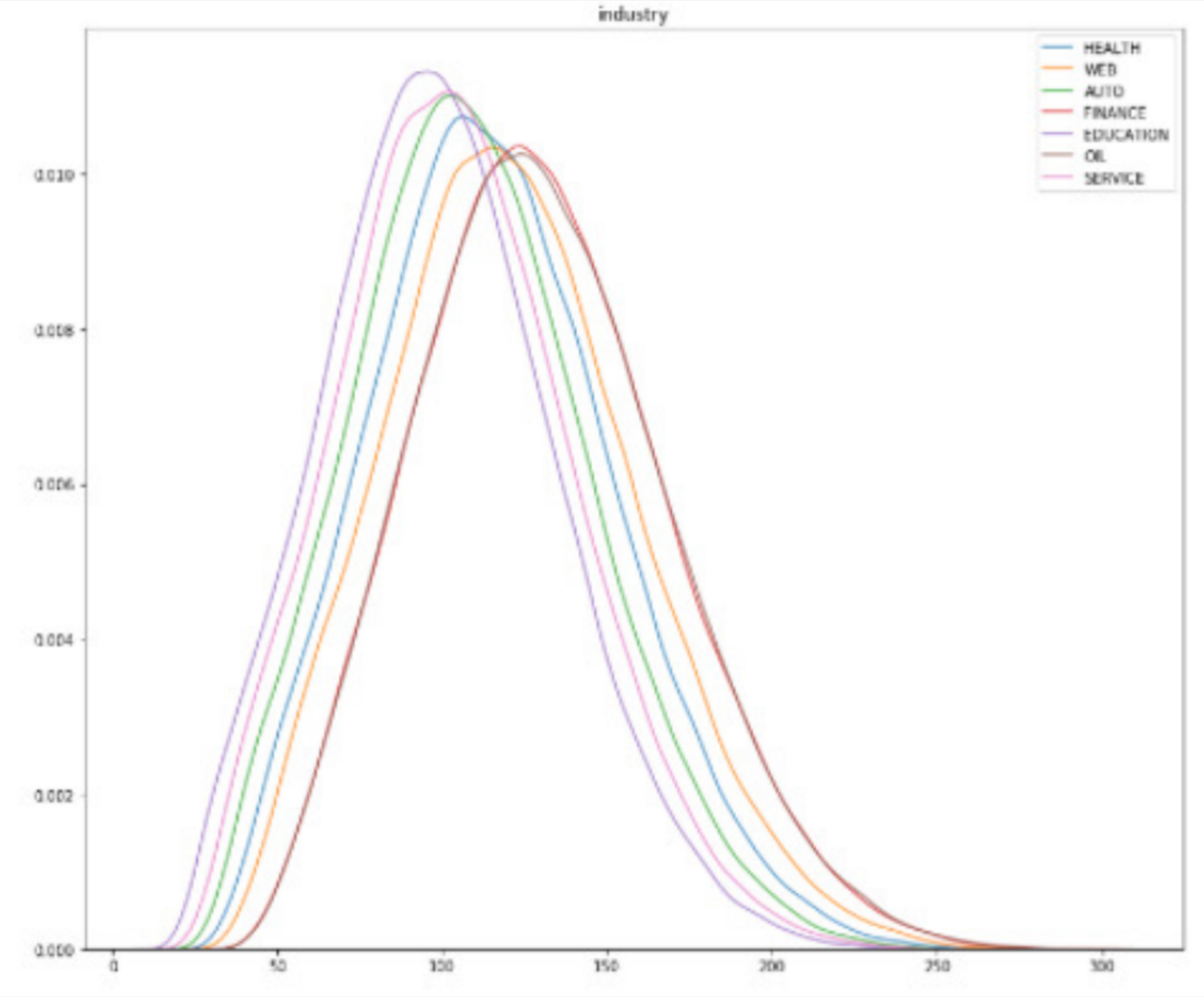
It's possible to infer that the label "janitor" usually have smaller paying averages. And the highest paying jobs are the positions for CEO, CTO along with CFO



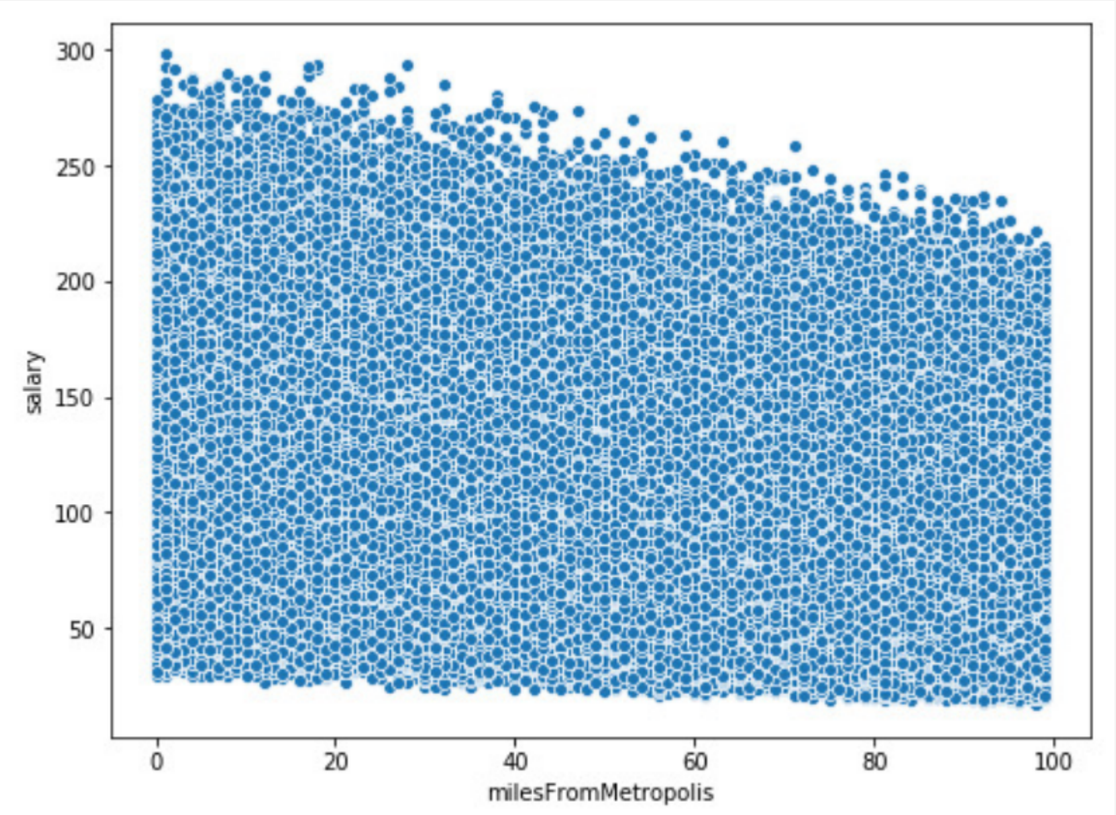
There seems to be a distinction between the average salaries of the higher skilled job positions and lesser skilled with educations None or high school

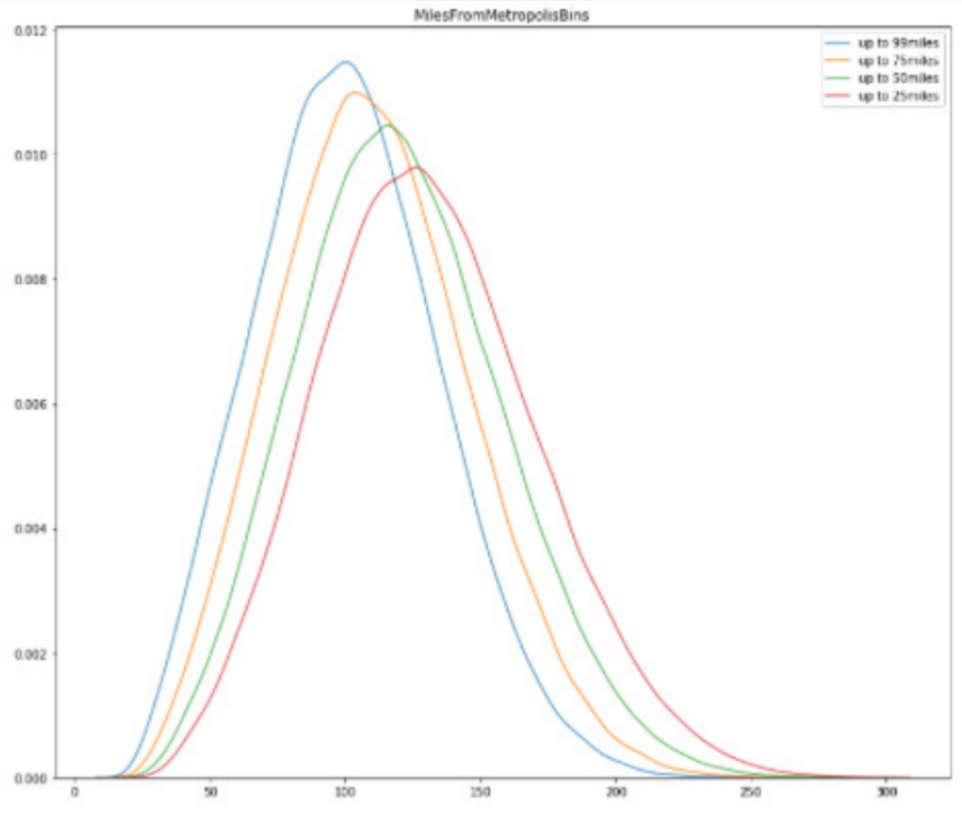


The major groups of fields there are among the most popular seem to have not a big difference between them meaning the diferences of salaries inside of them, however some present smaller contigency and slightly higher pay such as Engineering, Business and Math

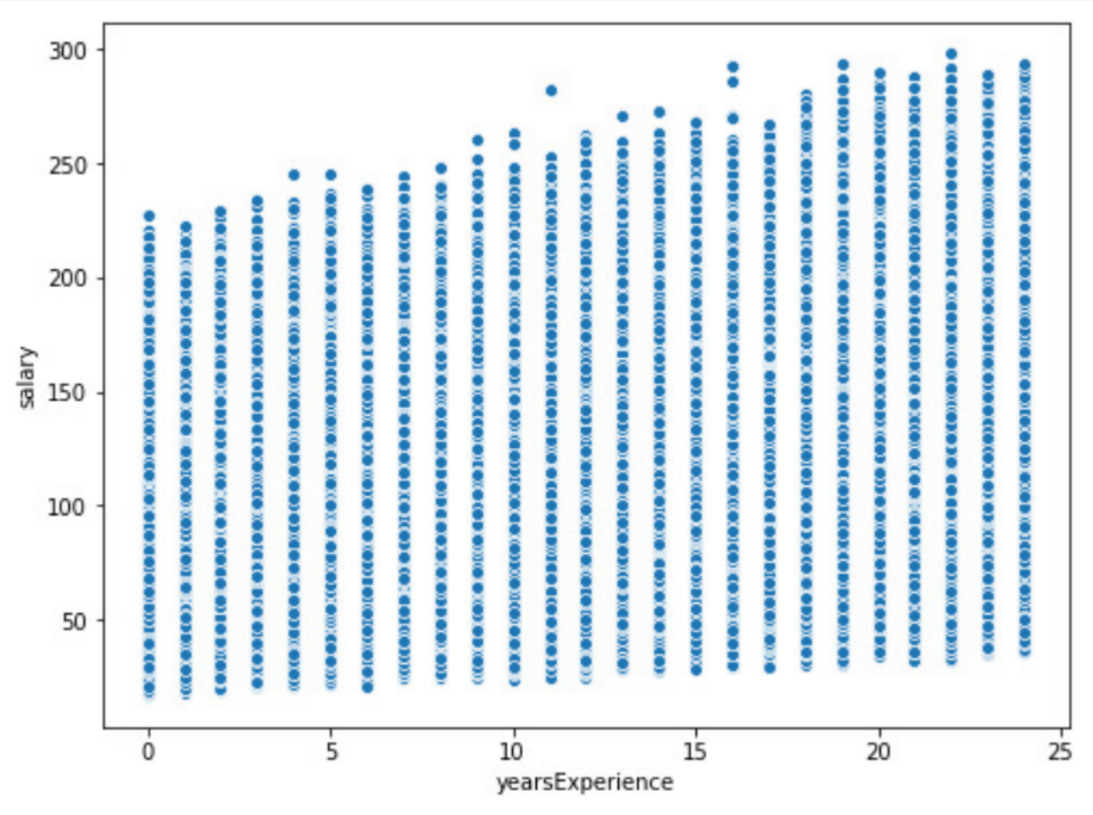


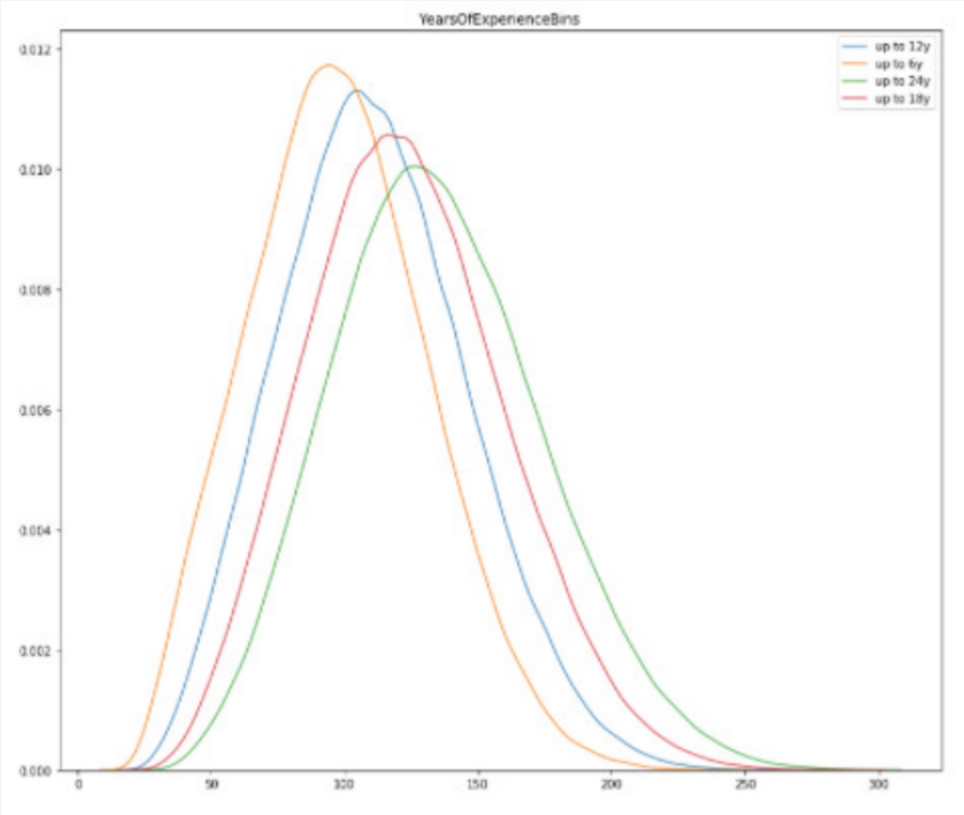
Also, similiar to the field of study, alll industries seemed to have very broad normal distributions but Education having smaller salary averages and Oil and Finance as leaders





Here it shows a binned abstraction of the "Miles from Metropolis" variable. And it's very clear the difference in numbers of individuals working close to a big city and those living afar, meaning that there are less people working close to the city and with higher pay, and the opposite is also true. Also, the scatterplot (though it's very disperse) there's a clear downward trend - the further from a metropoly the lower the pay





Similarly, there's a binned abstraction of the "Years of Experience" variable. Between 18 and 24y are the salaries with the highest average, and the scatterplot clearly shows a ascending trend evidencing a positive correlation between the feature ~years of experience" and the target.

02

In conclusion

As for job types, the positions of high scaler are the ones with higher salaries (CEO, CFO, CTO)
Larger salaries happen also for higher skilled educations (Doctors, Masters, bachelors)
The fields that have the largest salaries are for Engineering, Business and Math
As for industries: Oil and Finance are the higher payers
And for the numerical variables, those closer to the metropolis and the positions that require more experience also tends to pay more

03

Hypotesis

Based on the features we've seen, the models that would be most accurate would be ensable models.
I would like to try first the linear regression due to its plots that indicate strong correlation between the features and also two tree based models that are simple and have a very good predictive power
Also, even though the dataframe is not small at all, the features that are most categorical need some adjustments and might be a good fit to try label enconding, so we won't end up with a matrix so sparse.