31.05.2023

IT 258 Data Science

SALARY PREDICTION OF DATA SCIENCE JOBS

Gulshan Goyal
Siddharth Tyagi

CONTENT

INTRODUCTION

- Existing resources for salary information in data science are often outdated or sparse, leading to uncertainty in compensation discussions.
- Accurate salary estimation in data science is important for fair compensation, empowering job seekers, and promoting the growth of the field.
- Traditional wage surveys and websites like Glassdoor provide some insights, but they have limitations in terms of granularity, subjectivity, and data recency.
- By incorporating job-specific variables and leveraging machine learning, this approach provides more precise and unbiased salary estimations compared to existing alternatives like self-reported data and traditional surveys.

PROBLEM STATEMENT & OBJECTIVES

The demand for skilled data scientists is increasing rapidly, but determining appropriate salaries for data science positions can be challenging for both job seekers and employers.

***OBJECTIVES*:**

1. This project aims to leverage the existing data on Glassdoor.com to develop a predictive model that accurately estimates compensation for data science positions.
2. By utilizing the available data from Glassdoor.com, this project aims to provide a reliable and valuable resource for salary estimation, benefiting the data science community as a whole.

LITERATURE SURVEY

Title	Authors
Predicting data science job salaries using regression models.	John Smith, Emily Johnson, Michael Brown.
Hybrid approach for salary estimation in the technology industry.	Sarah Johnson, Ling Zhang, Jessica Lee.
Factors influencing salaries in data science: An analysis of qualification, programming language, sectors and experience.	David Lee, Rachel Kim, Andrew Chen.
Web scraping and Machine Learning for salary prediction in job markets.	Jennifer Chen, Richard Gupta, Samantha Sharma.

DATASET

- ***Data Collection:*** A pre-made **Selenium web scraper** was utilized to collect data with the parameters adjusted to meet the project's requirements. The data collection took place over a span of two days in April.

- ***Data Source:*** **glassdoor.com**, a widely recognized website for salary and job information, was utilized as the data source for this study, providing a substantial and up-to-date database. Its advantages include a large volume of data, user transparency in sharing salary information, and comprehensive coverage of data science job opportunities.

- ***Data Fields:*** Job Title, Job Description, Company Information, Salary Information, Location, Industry, Experience Level, Educational Qualification, Skills,.

- ***Data Representation & Size:*** The data used in this study is structured as a table with rows representing data points and columns representing attributes. It comprises approximately **955 rows** and **15 columns** in a comma-separated values (.csv) file format.

- ***Data Limitations:*** glassdoor.com, while valuable, has limitations as a data source due to potential biases and inaccuracies from self-reporting and incomplete salary information in some job postings. These factors should be considered when analyzing the data.

METHODOLOGY

***A. Data Collection*:**
In this study, data was collected from glassdoor.com using a pre-made Selenium web scraper. The scraper's parameters were customized to meet the project's requirements. The data collection process took place over two days in April.

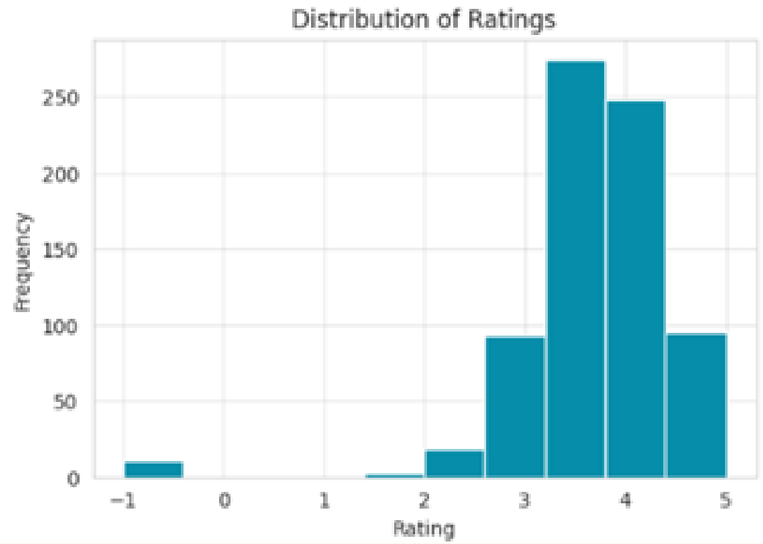
***B. Pre-processing and Feature Engineering*:**
In this study, only data points with complete job details were collected, and no missing value imputation technique was employed. The dynamic nature of the market and the potential complexity of the data made it necessary to ensure data completeness without imputation.

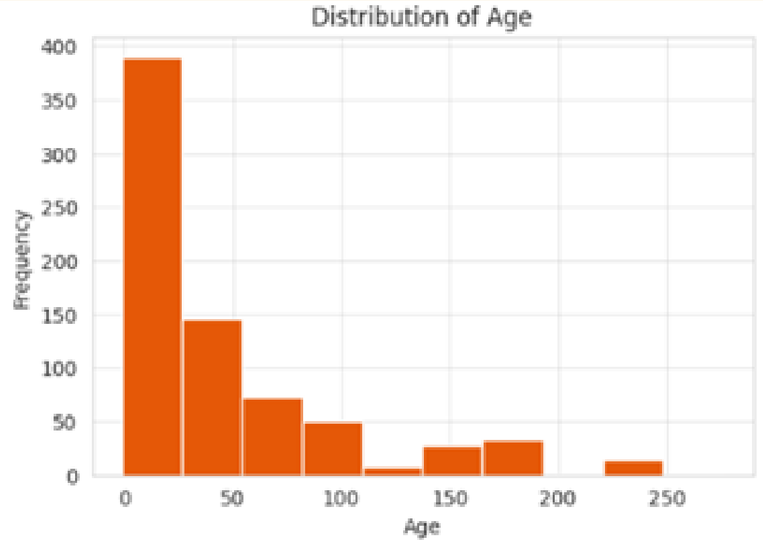
- Parsing salaries to generate new features.
- Creating new features from company names and location.
- Salary cleaning.
- Parsing Job Description.

METHODOLOGY

C. EDA

i) Histograms:





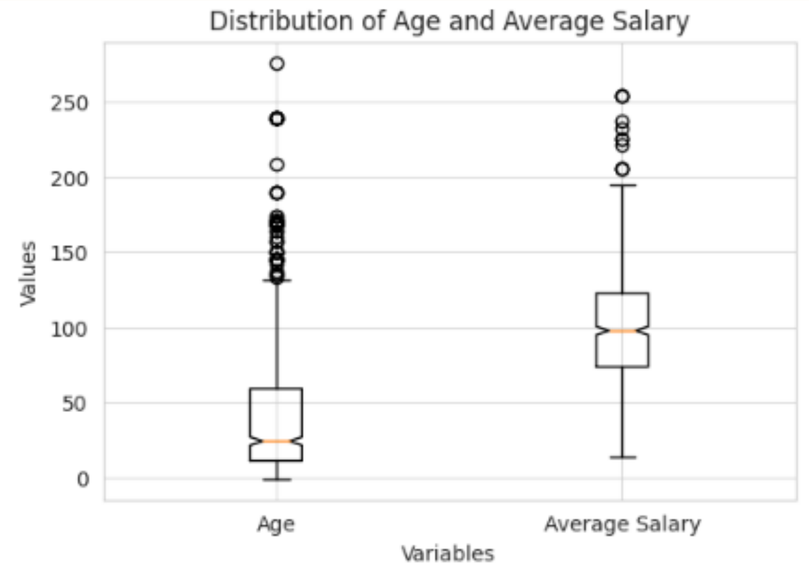
Observations:

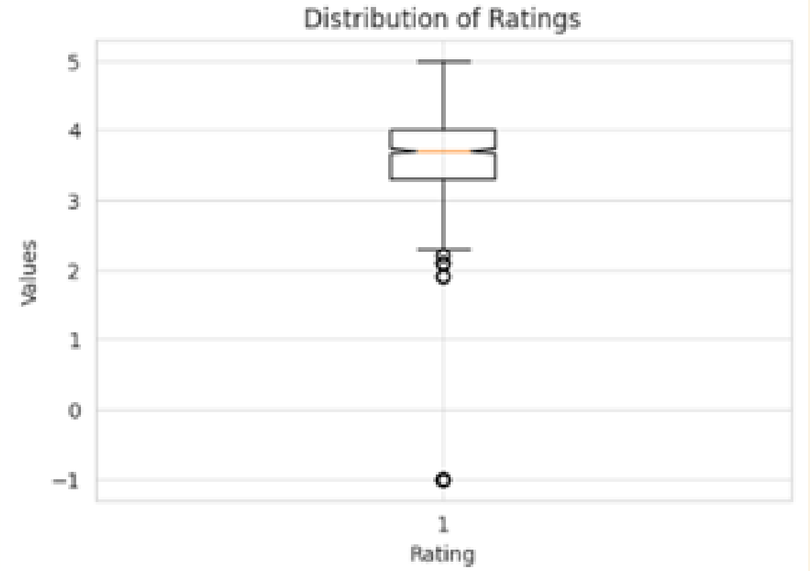
- Most of the companies are rated 3.5.
- Most of the people have an average salary $100K per year.
- Most of the companies are 0 to 25 years old indicating that data science jobs are recent.

METHODOLOGY

C. EDA

ii) Boxplots:





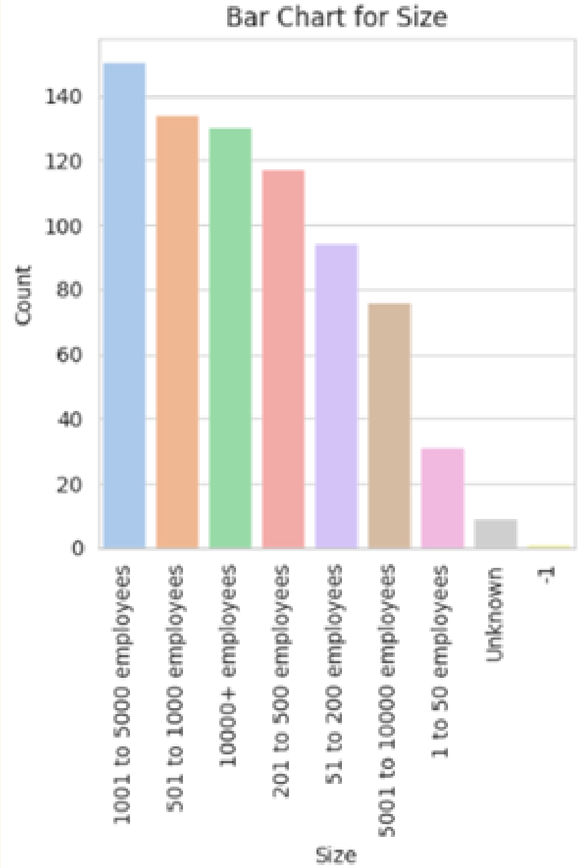
Observations:

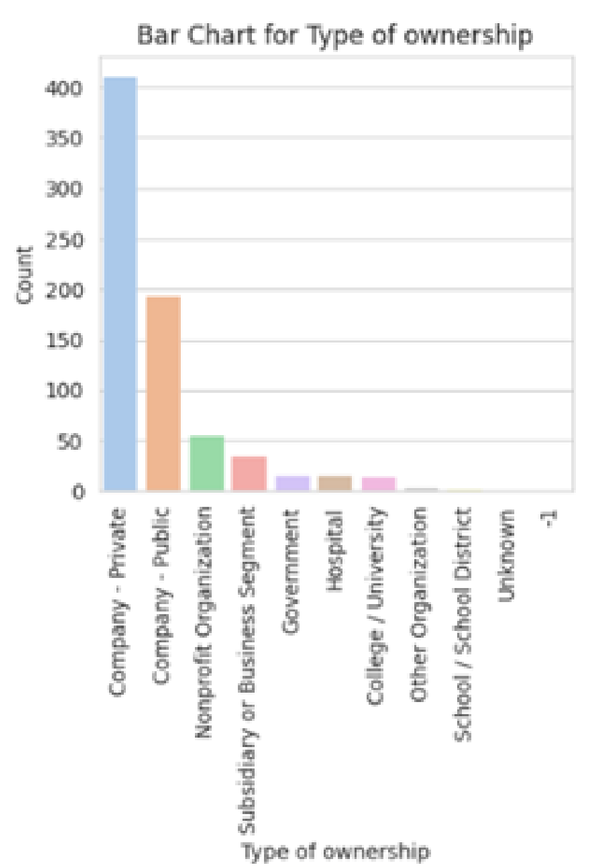
- There are several companies with relatively higher payscale and several other with a huge history.
- There are very few companies with exceptionally great or worst ratings barring one or two with worst ratings.

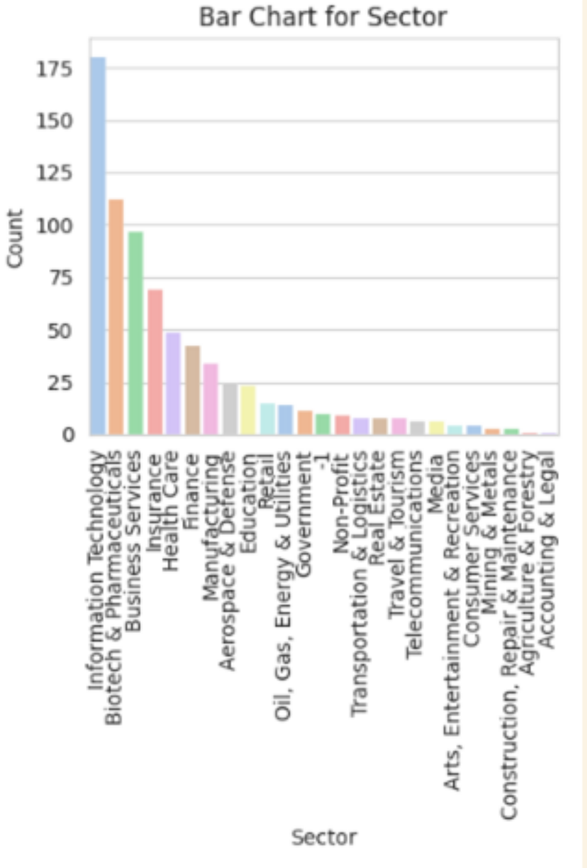
METHODOLOGY

C. EDA

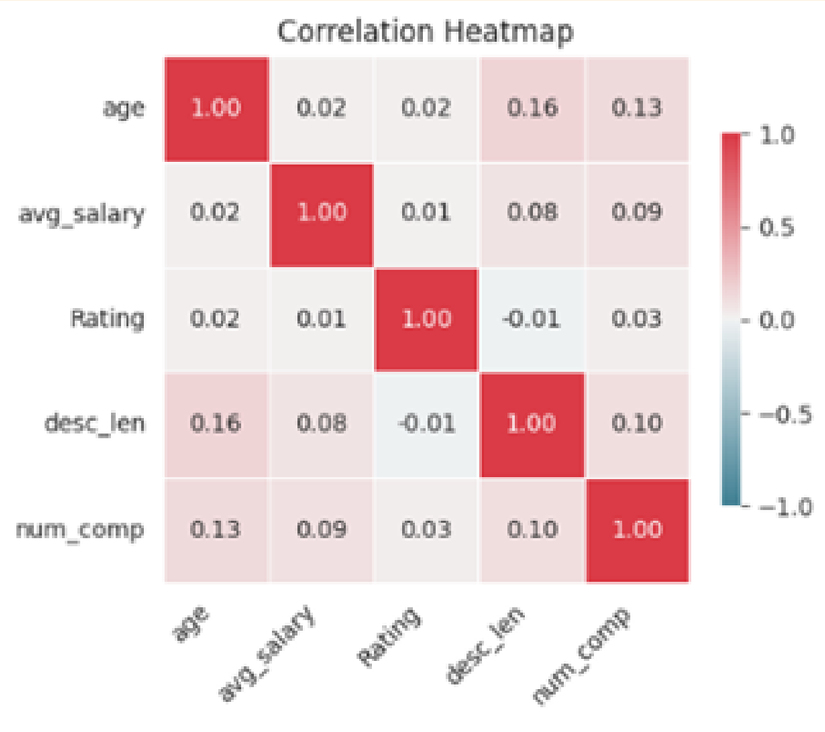
iii) Bar Graphs:



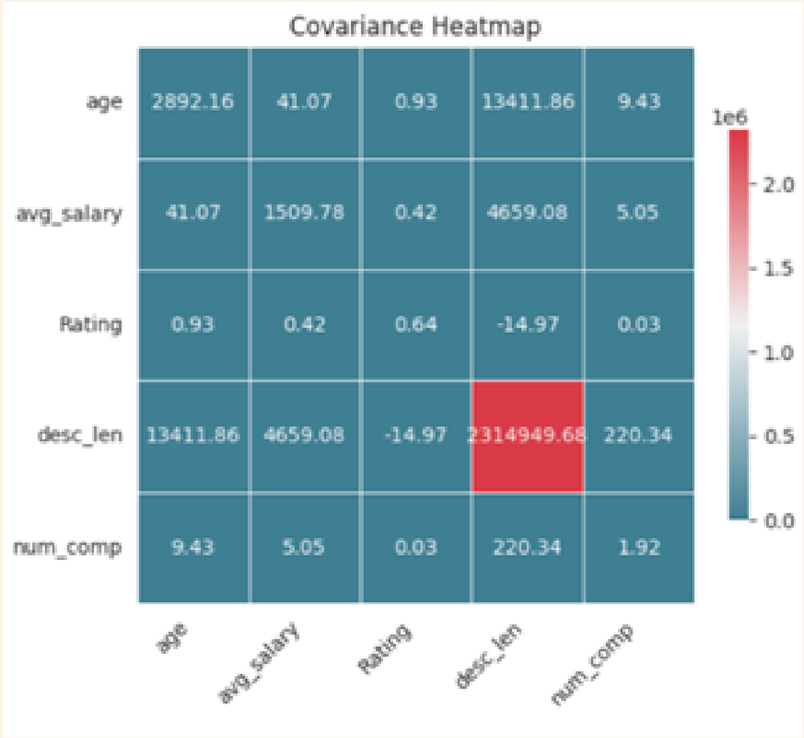




CORRELATION MAP



COVARIANCE MAP



METHODOLOGY

***D. Feature Extraction*:**

- 5 models were chose: Linear Regression, Lasso Regression, Decision Tree Regressor, Random Forest Tree, Gradient Boost Regressor.
- These models were applied on all 178 features (after hot encoding). Three types of error scores were recorded.
- Now top 10 features were selected through PCA and the same procedure was followed.
- Now a random feature was made and added to the dataFrame. Using RFT to find the feature scores we observe that the random feature is ranks 9th.
- The same procedure was performed on the top 8 features.
- After this we follow the below procedure for final results:

1. Iteratively choosing 1 to 100 features at a time.
2. Applying the aforementioned 5 models.
3. Recording the 3 errors.

RESULTS & ANALYSIS

The error scores after running the 5 models on all features:

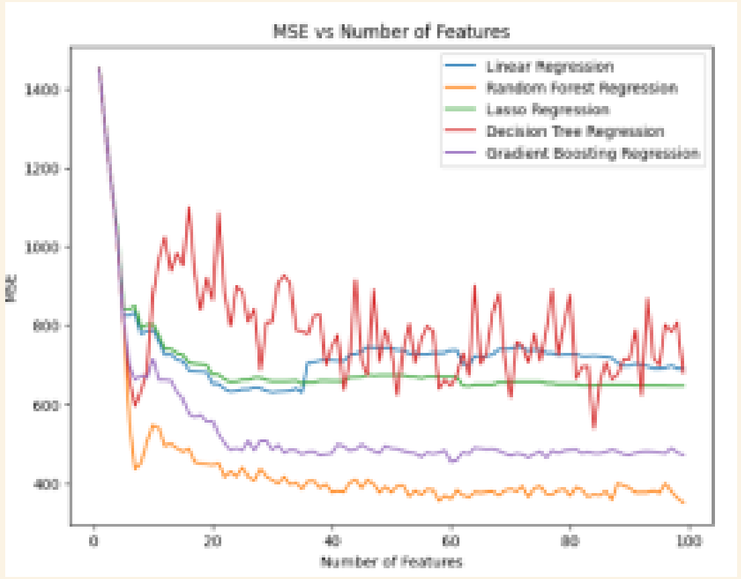
	Model	MAE	R-squared	MSE
2	Random Forest	11.257047	0.783398	354.037703
4	Gradient Boosting	15.614721	0.733845	435.033135
3	Decision Tree	8.949664	0.665866	546.145973
1	Lasso Regression	19.665304	0.604216	646.912823
0	Linear Regression	18.855190	0.578427	689.065222

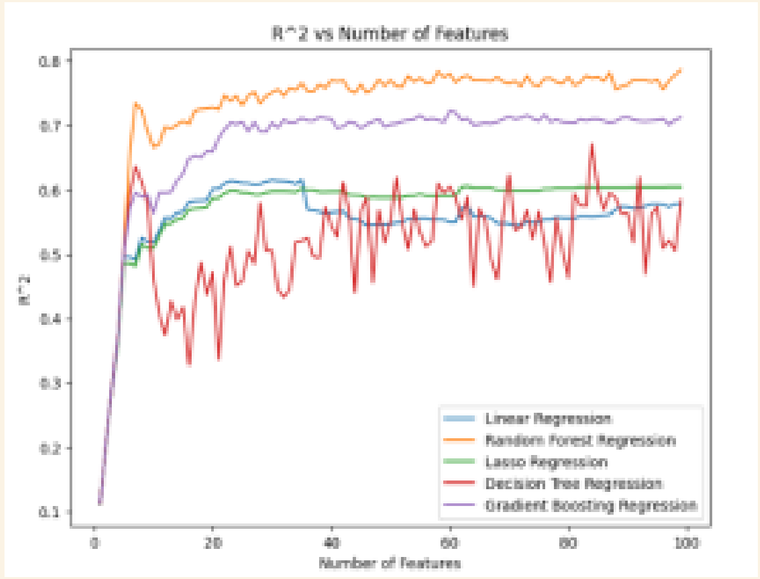
The error scores of 10 features obtained after PCA:

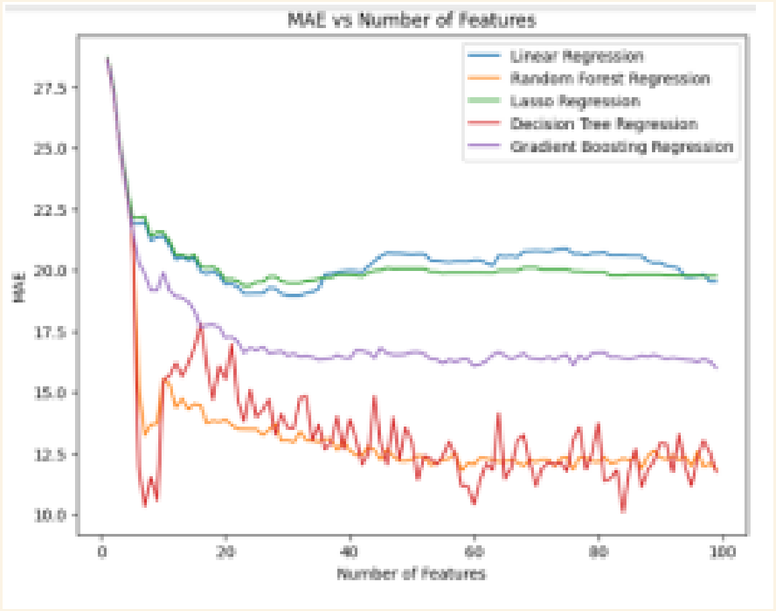
	Model	MAE	R-squared	MSE
2	Random Forest	12.537987	0.751718	405.819861
4	Gradient Boosting	16.494993	0.673062	534.383329
3	Decision Tree	11.761745	0.587261	674.625839
1	Lasso Regression	23.636611	0.363115	1040.995742
0	Linear Regression	23.667187	0.362016	1042.792072

	Model	MAE	R-squared	MSE
2	Random Forest	13.618993	0.730159	441.057859
3	Decision Tree	10.587248	0.716153	463.951342
4	Gradient Boosting	19.165290	0.589885	670.337173
0	Linear Regression	21.226592	0.524963	776.453147
1	Lasso Regression	21.405763	0.515178	792.447117

RESULTS & ANALYSIS







CONCLUSION

The below observations were made after all this:
1. Random Forest Tree came out to be the best in the first two trains (all features and 10 features with PCA) on the basis of MSE and R^2.
2. The accuracy in the latter case decreased.
3. The train for the features better than the random feature also resulted in a decreased accuracy.
4. In the 3 graphs, the values of MSE and MAE keep on decreasing and that of R^2 keeps on increasing as the number of features increases.

This leads us to the conclusion that using maximum features is beneficial to our model. One reason for this could be a weak correlation among all the features. By this we mean, no two features stand out with a strong correlation.

THANK YOU !!