LARGE-SCALE DATA MANAGEMENT

ENRIK MAZAI - *18390080*

1. *INTRODUCTION*

Having conducted an extensive research and reviewed various available websites with large-scale data for analysis, we have come to the conclusion that the data sheet titled "Data Science Salaries 2023" is the most appropriate and interesting for our analysis. This dataset refers to the salary trends of employees in the field of data science.

Using the Python programming language and various libraries,
We carry out data mining and analysis. We perform various analysis techniques such as graph and graph visualization, we identify correlations between
variables and we perform predictions using classification and regression algorithms.

2. *PROBLEM DEFINITION AND MOTIVATION*

The motivations that led us to choose the specific topic "DATA SCIENCE SALARIES 2023" for our work are enough. Among others, some of the main reasons for selection include:

1. There is a general interest and will to discover data science from students, professors and professionals who are part of this field.

2. Job prospects: Data science is a growing and vibrant industry. Professionals in this field often have good career prospects and can look for opportunities to advance and increase their income.

3. Need to investigate salary trends: Understanding salary trends in the field of data science can help professionals or students understand how competitive the sector is and what are the areas where they can develop their skills to improve their chances of greater income.

4. Contributing to the community: Analyzing salary data can help inform other professionals or stakeholders about market trends and provide useful data for career decisions.

Obviously, data analysis of the salary trends of data science employees is extremely useful for:

1.Students and learners : For those studying or training in the field of data science, analysis of salary trends can provide information
on career prospects and labour market competitiveness. It can help them understand which skills and knowledge are particularly in demand and what their reward in the labour market might be.

2. Data science professionals: Data science professionals can use salary data analysis to understand

pay trends in the sector and understand what skills and experiences they can enhance to increase their value in the job market.

3. Companies and employers: For companies active in the field of data science, analyzing wage data can help them understand the
current market trends and attract and retain talented professionals. In addition, it can help them design pay and subsidy policies for their employees.

4. Simply interested: People interested in the general field of data science can find interesting information about fees, market
Career trends and possibilities in this field through the analysis of salary data

3. BRIEF DESCRIPTION OF THE DATA WE USED

About the dataset we use, the dataset contains 11 columns of data, which is:

work_year: The year the salary was paid. (2021-2022-2023)

experience_level: The level of experience at work during the year. It contains the values EN, MI, SE, EX which translate to entry, intermediate, superior, executive respectively.

job_type: The type of employment during the year.It mainly contains PT-FT values for part-time and full-time.

job_title: The job title during the year.

salary: The total gross salary paid.

salary_currency: The currency of the salary paid as a currency code.

salary_in_usd: Ο μισθός σε δολάρια

employee_residence: The main country of residence of the employee as ISO 3166 country code.

remote_ratio: The total amount of work done remotely. The values are 100 for full remote work, 50 for hybrid work model and 0 for full on-site
employment.

company_location: The country of the employer's head office or branch.

company_size: The median number of people who worked for the company during the year. The values found are S,M,L for small, medium and large size respectively.

4. DESCRIPTION OF THE METHOD OF DATA ANALYSIS

The analysis of the dataset was initially done with some graphs in order to understand our data better and the connection of the columns to each other. These graphs will also help us to better approximate our conclusions. The column that interests us most is the column with the salaries of employees (in dollars), whenever we display in boxplot form the relationships of salaries with the percentage of telework (remote ratio) and the size of the company they work (company size). Then we show the average salary in countries where we have more than 15 registrations and the relationship of the salary to the employee's experience level.

Once we have performed the necessary steps of cleaning our data in case of blank data, we proceed to our experiment. We will use the analysis technique classification & regression.

For classification we use column salary_in_usd and divide it into 3 classes with the help of quadrants (0.25 , 0.75) where we have: low (0 to $ 95,000), medium ($ 95,000- 175,000), high (+ $ 175000). Then we keep the columns that interest us and have the greatest connection to the final salary where they are 'work_year', 'experience_level', 'employment_type', 'job_title','remote_ratio','company_location', 'company_size'.
With the help of onehot encoder we can also recognize non-numeric data by converting them to binary formats. When the pre-processing of the
Of categorical data, we divide the selected dataset into training and testing by percentages of 80% and 20% respectively. This means we use 80% of the dataset to train the machine learning system and then evaluate it with the 20% that was not used. The preferred method for the creation of the classification model is Logistic Regression due to better results.

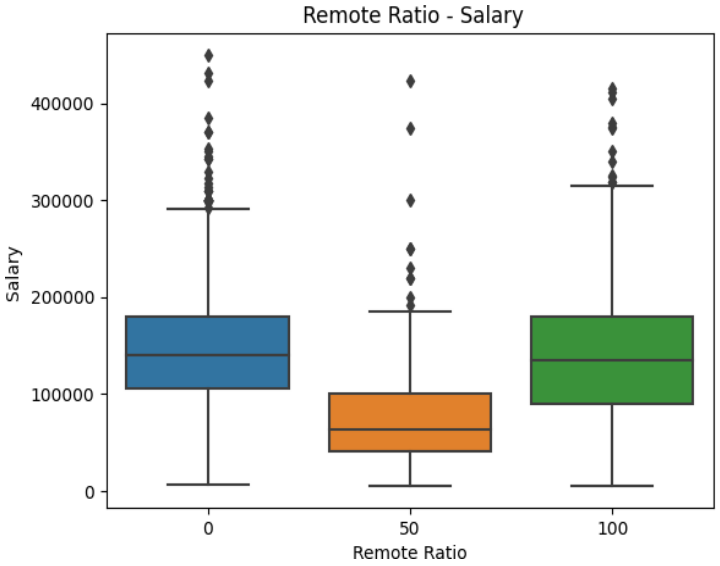
For **regression** we follow similar steps alongside classification in our code and use the Linear Regression method for our model. The dataset is divided into training - testing and we use the same columns as the classification model.

The purpose of classification and Logistic Regression is to categorize input data by predicting which economic class it belongs to. While the purpose of regression and
Linear Regression is to predict as an expense the exact salary of the employee depending on the entry prices.

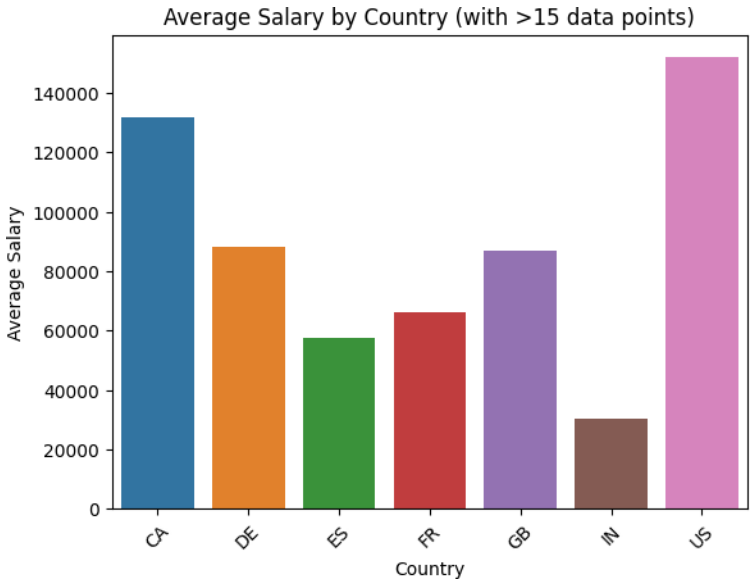
5. EXPERIMENTAL RESULTS

Initially from the visualization of the data and the relationships between them we have the following graphs that we mentioned

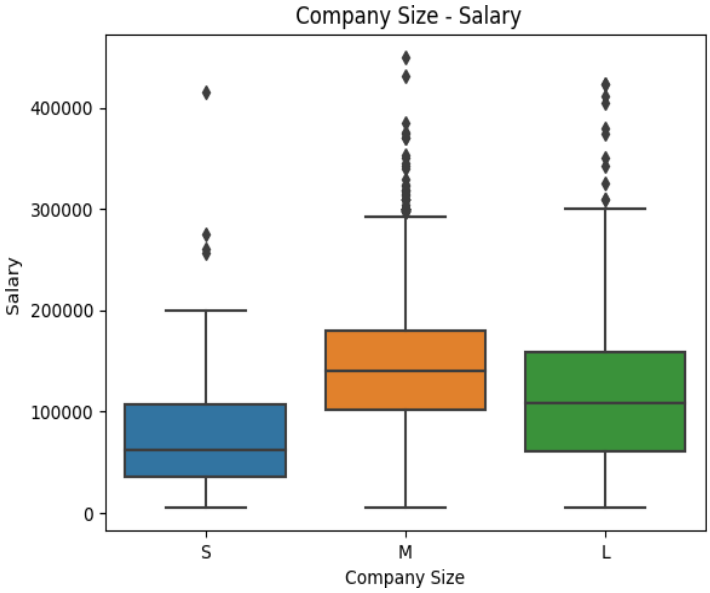
remote ratio-salary



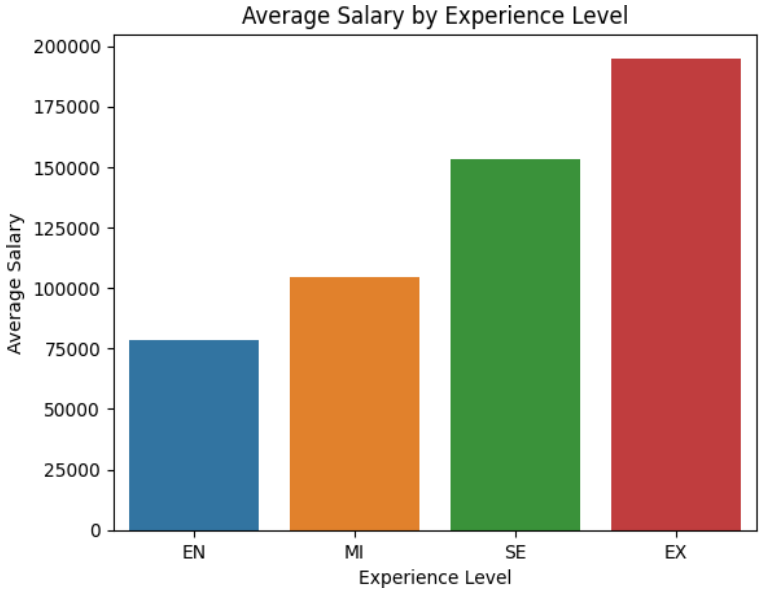
Average salaries countries with 15> records



company size -salary



Average salary Experience base



For the results of our experimental models we have created in the code a
New instance variable where we enter the data we want our employee to have and with the output the models classify him in a class (classification model) and calculate his salary (regression model). Here are some examples:

1st I/O example

```
# Example instance
new_instance = [[2023,'EN', 'FT', 'ML Engineer',100,'US', 'M']]
```

```
Predicted Salary Class: ['medium'] Range: 95000.00-175000.00
Predicted Salary (in number): 148048.92
```

2nd Example

```
# Example instance
new_instance = [[2023,'EX', 'FT', 'ML Engineer',100,'US', 'M']]
```

```
Predicted Salary Class: ['high'] Range: 175000.00+
Predicted Salary (in number): 235727.69
```

3rd Example

```
# Example instance
new_instance = [[2023,'MI', 'FT', 'ML Engineer',100,'IN', 'M']]
```

```
Predicted Salary Class: ['low'] Range: 0-95000.00
Predicted Salary (in number): 82726.00
```

6. *CRITICAL EVALUATION OF RESULTS*

At first glance at the outputs of the models we can say that it responds correctly to our inputs. In more detail by looking at the graphs we created for the relationships of
columns with salaries we can see how in the first example a Machine Learning Engineer in entry position, full time and based in America is correctly placed in the
medium class. Although the employee is in an entry position, his classification in medium class and not in low can be justified based on the graph that shows us the ranking of wages by country that we see America first with a significant difference.

In the second example we change our input about the employee's experience where we enter him EX-executive keeping the rest of the values the same. We observe that the system ranks him in the high class, a prediction justified by the graph on the average salary per experience level where executives are first by a significant margin. Also
We see the regression model predicts a salary of about 236,000 which is well above average due to the remaining prices being favorable conditions for large income.

In the third example we change the position to MI-Intermediate and the company's headquarters to India. The system again provides a justified prediction and places the worker in the Low class, as India is the country with the lowest median income in all jobs based on the dataset data.

For the most accurate approximation of the results we have introduced a classification report with the sklearn.metrics library where we calculate performance metrics of our model. Some values it calculates are precision, accuracy, support, f1-score etc

Classification Report:

	precision	recall	f1-score	support
high	0.63	0.24	0.35	164
low	0.85	0.64	0.73	164
medium	0.65	0.90	0.75	371
accuracy			0.68	699
macro avg	0.71	0.59	0.61	699
weighted avg	0.69	0.68	0.65	699

Our model has a prediction accuracy of 68% which translates into 68 correct predictions per 100. Overall, the model appears to perform well for the "low" and "medium" categories, but has difficulty correctly predicting the "high" category.

7. CONCLUSIONS

Based on the outputs we received, the results can be characterized as successful, of course with the help of metric calculation tools we can see how the systems can be significantly improved.

In general, given our purpose of forecasting and classifying salaries on an international scale, the results of metrics appear to be of the order of 68% due to its complexity. For example, the dataset contains many job titles relative to its total size, so due to limited records in some positions or countries, accuracy drops. A more limited analysis of the models only in countries, such as America, would be considerably more successful because of the amount of data we hold.

To improve systems and increase the accuracy of results, several changes can be made. One of them is to add additional data to our dataset as this will improve our training model and clarify data where there are few records and the results may not be recognized correctly.

We can also investigate more algorithms for the techniques we followed to find the most suitable for our dataset, such as support vector machines- SVM or Decision Trees for classification.

Finally, an extensive correlation matrix could be used that would better show us the weight of the columns that affect the salary column. With this
The selection of columns for the training model would be based on gravity and could increase the accuracy of the systems.

Bibliography

1. *Data Science Salaries 2023* (n.d.). Ιούλιος 1, 2023, από https://www.kaggle.com/datasets/arnabchaki/data-science-salaries-2023