JOB MARKET ANALYSIS KEY FINDINGS

INTRODUCTION

The intention of this project is to provide a thorough analysis of a dataset regarding it-related job offerings, where we analyze relevant aspects such as the job title, location and gross average salary.

METHODOLOGY

1. the project consists of an analysis of IT industry job openings from a kaggle datasets that has a sample size of 3,755 elements
2. we provide information on the gross salary, irrespective of other benefits (contributions to pensions schema, health insurance, travel allowances, etc.)
3. we do not take into account the inflation effect / consumer price index
4. this dataset provides information as of April 2023, for updated information we might have to refer to a dataset that contains the most recent job openings

Note: before we get into the project, we would like to highlight that we have also created another jupyter notebook where we perform the entire machine learning after removing outliers, following the iqr method. File is available in the following link:

https://github.com/carlosruiz-stack/ironhack_final_version/blob/600f7441f4b314ed7409ccefd8cee210c6bfc965/it_job_offerings_removed_outliers.ipynb

HIGHLIGHTS

Compared to the initial distribution histogram (figure 6), where the distribution was right-skewed, this time the mean is shifted to the center and takes more the shape of a normal distribution (avg. salary, figure 10).

We perform a hypothesis testing where we want to prove that the average salary for an entry level job (experience_level = 'EN') is at least (higher than or equal to) 40.000 (salary_in_usd).

Note: this is a one-tail t-test where we take a significance level of 0.05 for our hypothesis testing.

Result for the hypothesis testing: Since the p-value is smaller than alpha, we keep the null hypothesis, which means that we can estimate, with a confidence of 95%, that the average salary for an entry-level job, given the dataset at hand, is higher than 40,000 usd a year. We provide a graphical representation for this (t-testing) in figure 11.

perform a hypothesis testing where we assume that there is no significant difference between the average salary for an senior and middle management position ('SE', 'ME'). with 99% of confidence, we can assume that there is not significant difference in the avg. salary between senior and middle level manager positions for the job openings provided in the dataset.

Based on the initial dataset, we want to find if there has been any significant difference in the avg. salary for the years 2021, 2022 and 2023.

In the analysis performed, we find that there is a significant difference in the average salary in usd for the years 2021-2023. In this case, we obtained a F-statisic of 94.39761605417166 and a p-value of 1.051205815597408e-40.

If we wanted to assess this condition individually, It would require us to analyse the following pairs:

2021-2022
2022-2023
2021-2023

Due to time constraints we will not proceed any further from here, however, it is an interesting detail for future analyses.

DATA VISUALISATION

Performed with Tableau workbook:
https://public.tableau.com/views/ITJOBOPENINGSDASHBOARD/Sheet2?:language=es-ES&:display_count=n&:origin=viz_share_link

SQL QUERIES

CREATE DB IN SQL;

```
use job_openings_db;

CREATE TABLE JOB_OPENING (
  job_opening_id INT PRIMARY KEY,
  job_title VARCHAR(255),
  salary DECIMAL(10, 2),
  location VARCHAR(255),
  currency VARCHAR(255),
  experience_level VARCHAR(255),
  employment_type VARCHAR(255)
);

CREATE TABLE COUNTRY (
  country_id INT PRIMARY KEY,
  country_name VARCHAR(255)
);

CREATE TABLE CURRENCY (
  currency_id INT PRIMARY KEY,
  currency_name VARCHAR(255)
);
```

Created the following DB and performed the corresponding queries:

```
SELECT * FROM job_openings_db.`ds_salaries (1)`;

# Provide a top 20 of the entry-level positions with the highest pay in the year 2022

SELECT job_title, salary_in_usd FROM job_openings_db.`ds_salaries (1)`
WHERE experience_level = 'EN' AND work_year = 2022
ORDER BY salary_in_usd DESC
LIMIT 20;

# Calculate the average salary of a machine learning engineer
```

```
SELECT AVG(salary) AS average_salary
FROM job_openings_db.`ds_salaries (1)`
WHERE job_title = 'Machine Learning Engineer' AND experience_level = 'MI';

# Select the job openings in Spain that are not 'experienced level' (EX) and provide the first 20 positions, order by job title

SELECT salary, job_title, experience_level, company_size FROM job_openings_db.`ds_salaries (1)`
WHERE company_location = 'ES' AND experience_level !='EX'
ORDER BY job_title ASC
LIMIT 20;
```

STREAMLIT

In simple words, we are going to deploy a Streamlit user interface where we select the values for the input variables of our model ('experience_level','employment_type','job_title', 'remote_ratio') and the user eventually gets an estimate of the average salary (expressed in usd, irrespective of the job location) on how much he is going to get paid on average. Note that this is just an estimate which does not consider other aspects from the package benetis, such as travel allowance, insurance, overtime payment, etc.

URL to the application deployed:

https://job-market-analysis.streamlit.app/

Github repository:

https://github.com/carlosruiz-stack/streamlit_app

LIMITATIONS

Due to time constraints, and the technical difficulties encountered in previous web scraping attempts, it has been withdrawn the option to scrap data from infojobs, since the general T&C of

the website do not allow for web scrapping, plus, they have controls in place to block any web scrapping action.

We instead take the data from a .csv dataset obtained from kaggle, whose link is provided in the initial section.

There is ambiguity in some variables, for example, when they mention small, medium and large companies they do not provide a criteria for that categorisation (number of employees / yearly revenue, etc.), the same applies to contract type (FT, PT, FL, CT), does not provide clear information on what each of them mean and had to figure out by ourselves, using our best judgement.

We encountered some problems with the libraries for partial dependence plots, after several attempts in installing and uninstalling python and sk.learn packages we decided to skip this visualisation plot to move further with the rest of the analysis.

The same way, we encountered several problems in running the Streamlit application code in Visual Studio. Eventhought we managed to launch the user interface based on VS code, we could not get the application 100 % operational as it does not return the value for the target variable, average salary. It is thought that this has to do with the sk.learn library for linear regression. No successful attempt has been made to resolve this issue.

FURTHER ANALYSIS

Some considerations that can be made for future related analysis on the field:

- Scrap the data from a job website that allows for websraping in their T&C and does not block web scraping attempts using captcha validation requests. Examples include www.tecnoempleo.com
- Obtain a dataset with a larger number of observations and wider variety of IT-related roles, other than data analytics and machine-learning related positions (e.g. software engineer, QA, beta-tester, etc.). Obtain dataset with more columns to get a model with more variables
- It would also be interesting to analyse, whenever this type of information is provided, the number of applicants to that job opening and the amount of time that the position has been opened, as applicable in the job posting information site
- Another aspect that we might consider interesting is to create a regression model where we predict the ‘salary’ based on each local currency and the candidate location (irrespective of the company’s location), and evaluating those results

- Categorize the job openings by groups, for example, Data Analyst, Data Scientist, Machine Learning, etc., using the most representative criterion for grouping the job openings in an effcient way
- We obtained a dataset for job openings across worldwide locations, however, it would also be interesting to get a larger dataset where we analyse positions based on lower aggregation levels for each location (per region or town in a country, postcode where available, etc.).

ANNEXED CONTENT

Performance evaluation of the machine learning model (csv)

https://docs.google.com/spreadsheets/d/1q2hXf5aacgEXfMfpvyseP_63ZCkGvQSyEvoD_IW0ZxU/edit?usp=sharing

Fig. 1.- Average salary by experience level

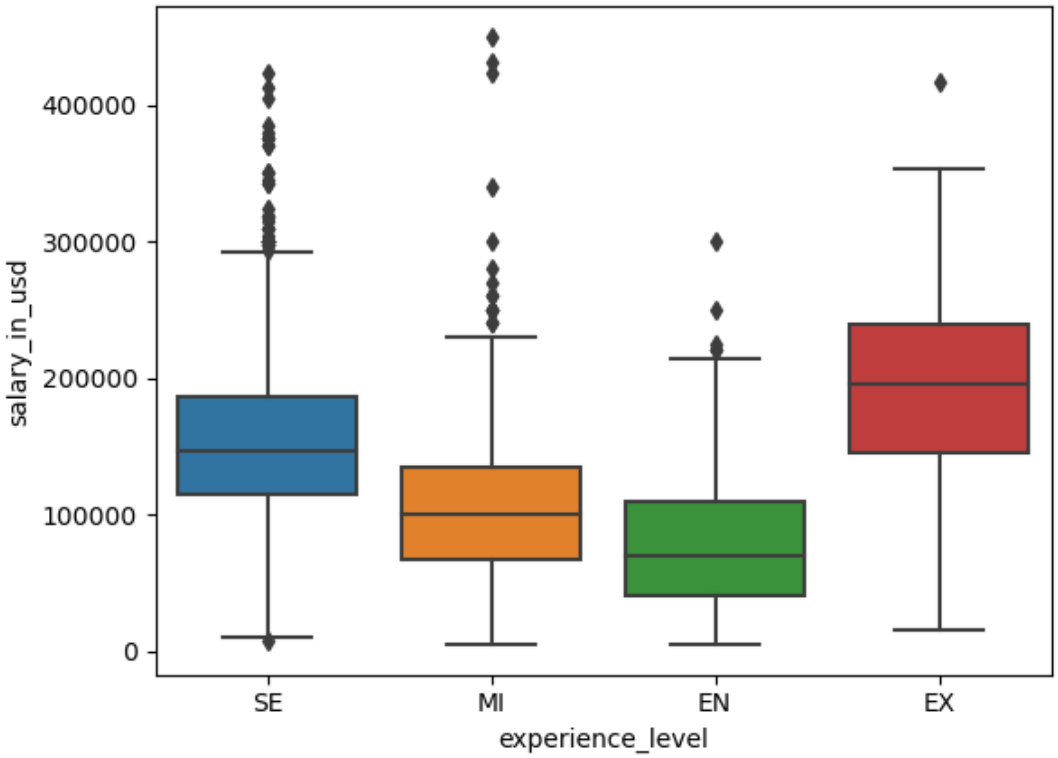


Fig. 2.- Average salary by employment type

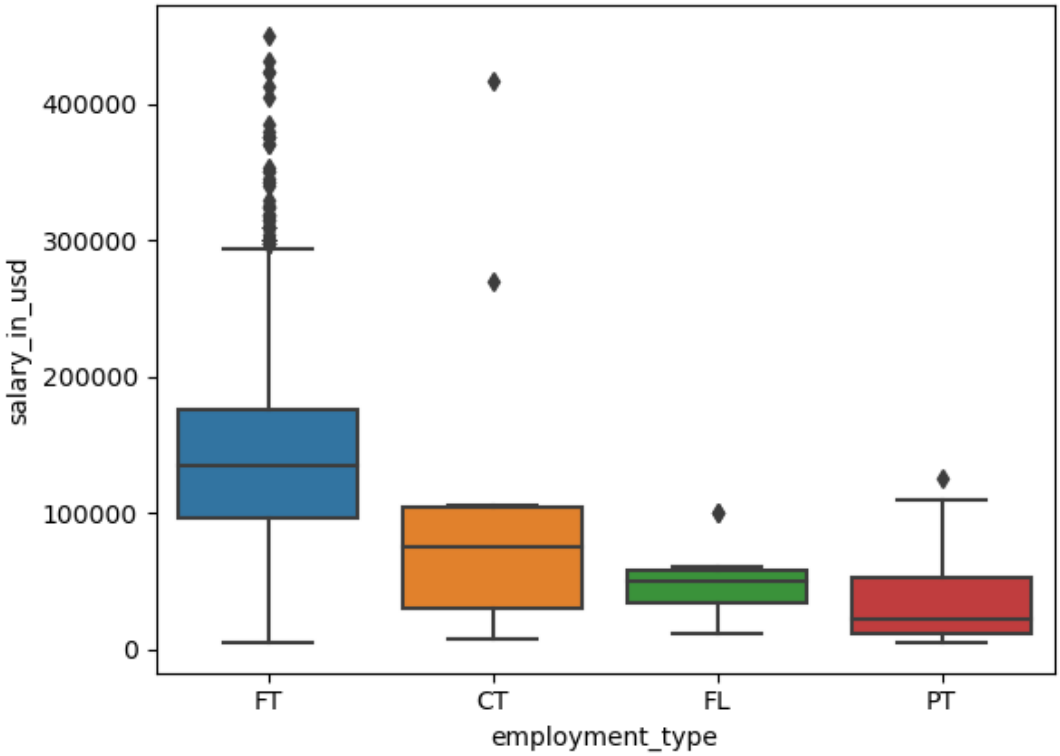

400000
300000
200000
100000
0
salary_in_usd
FT
CT
FL
PT
employment_type

Fig. 3.- Average salary by job title

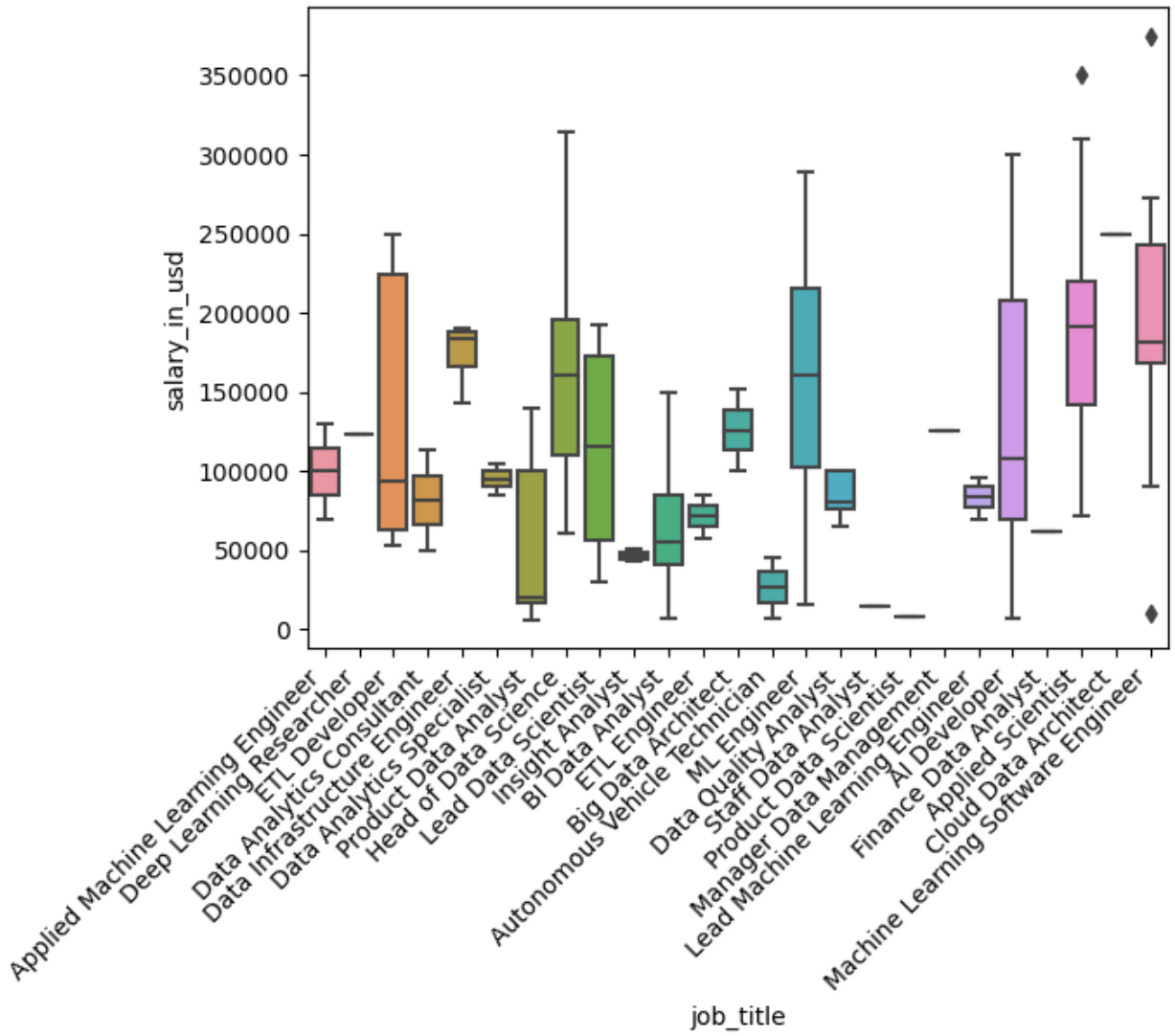


Fig. 4.- Average salary by company location

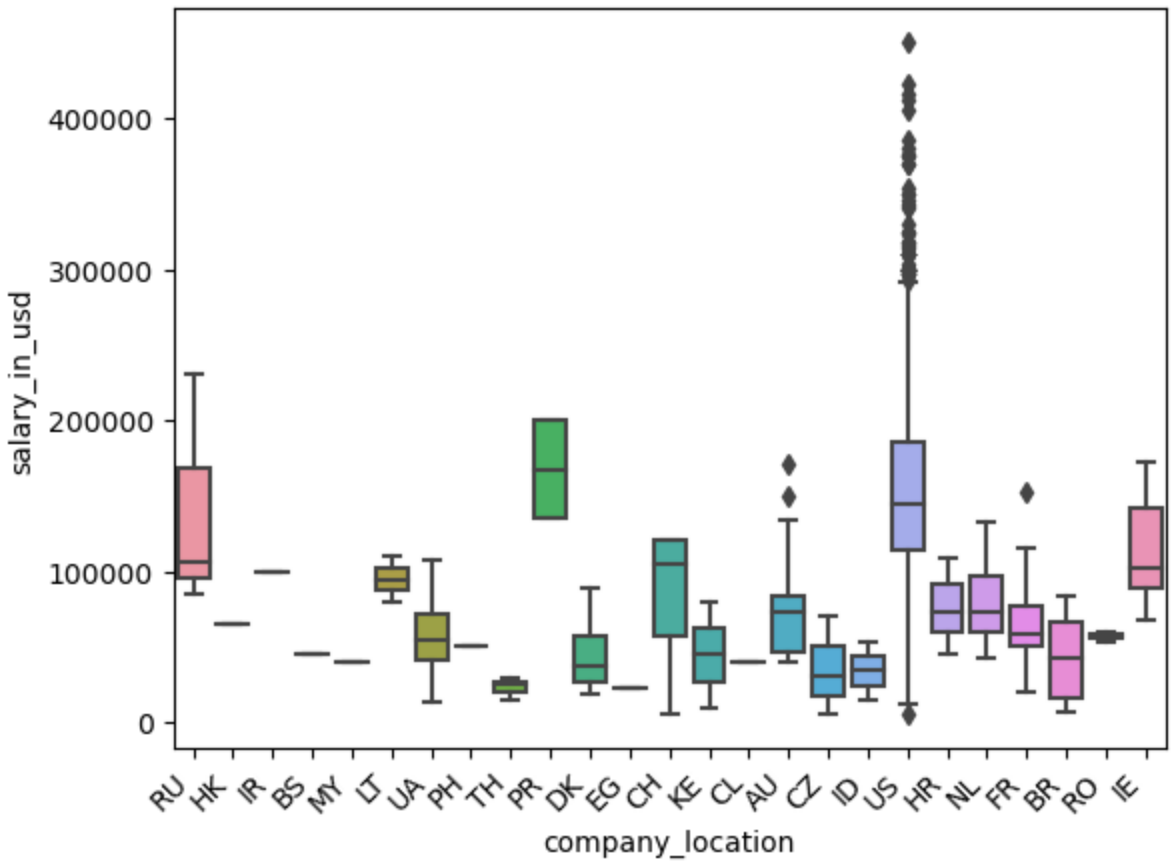


Fig. 5.- Correlations heatmap: salary in usd and remote ratio

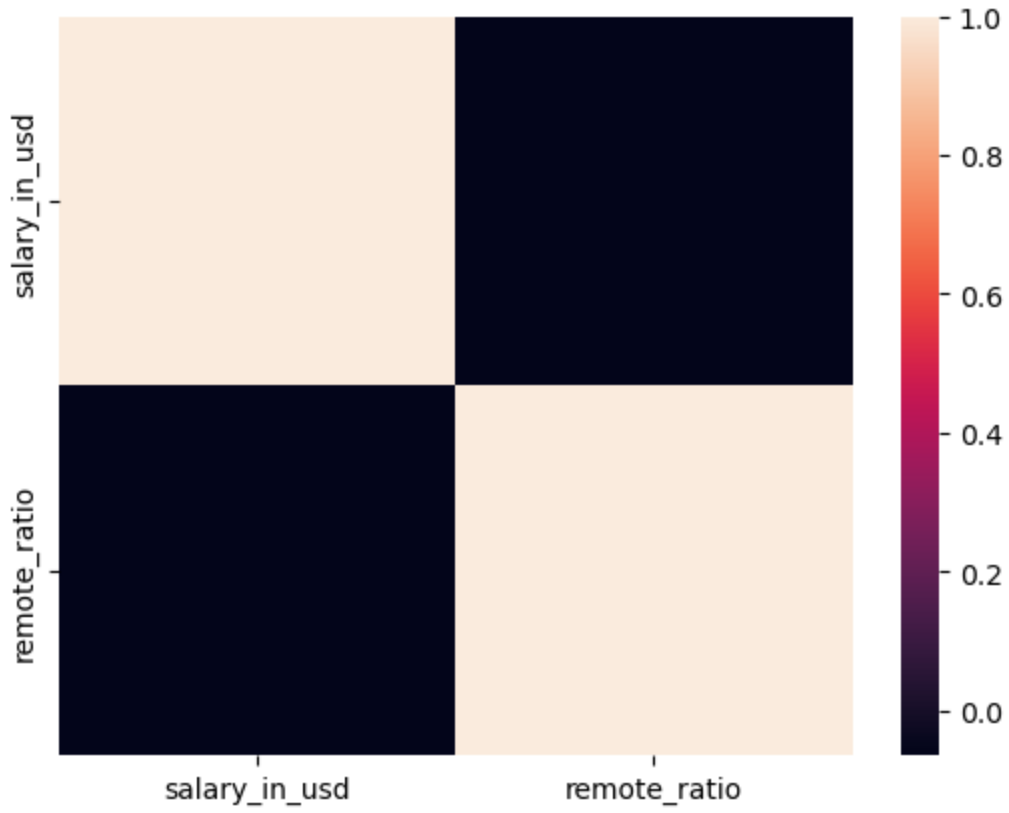

salary_in_usd
remote_ratio
salary_in_usd
remote_ratio
1.0
0.8
0.6
0.4
0.2
0.0

Fig. 6.- Average salary histogram

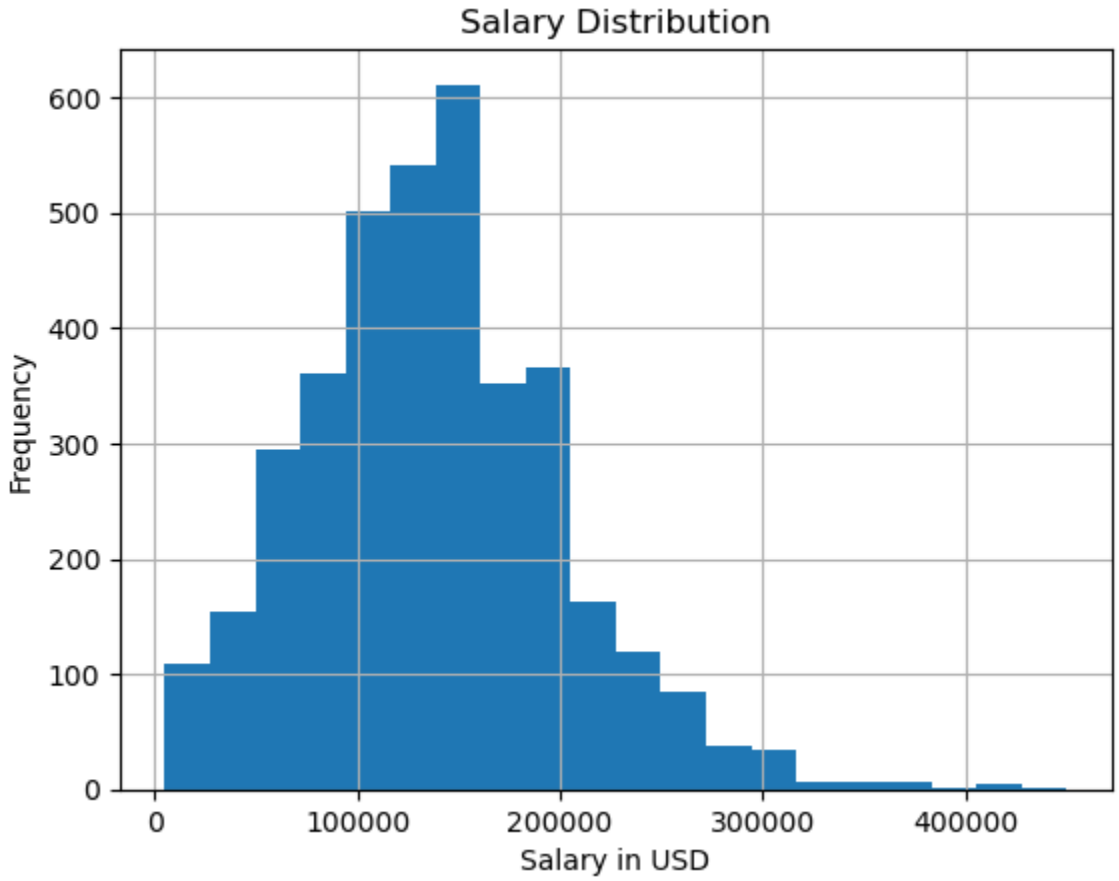


Fig. 7.- Salary by experience level boxplot after removing outliers

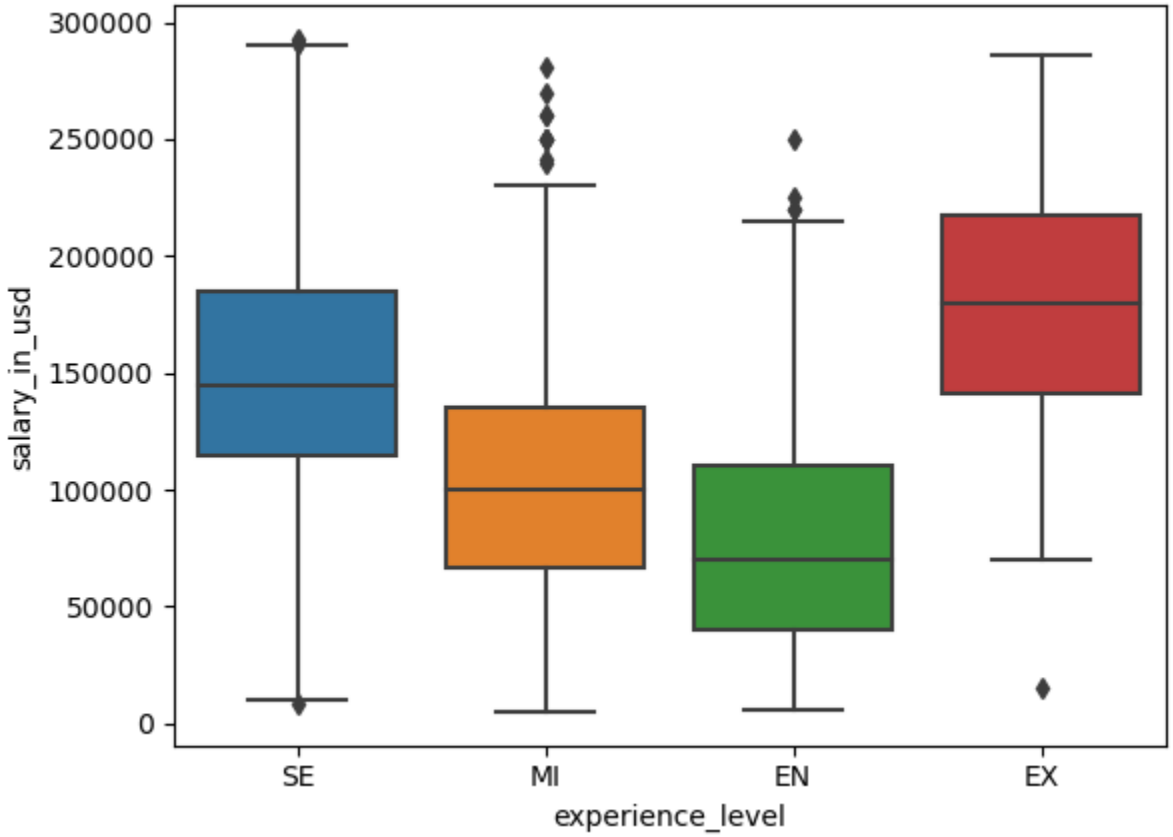


Fig. 8.- Salary by employment type boxplot after removing outliers

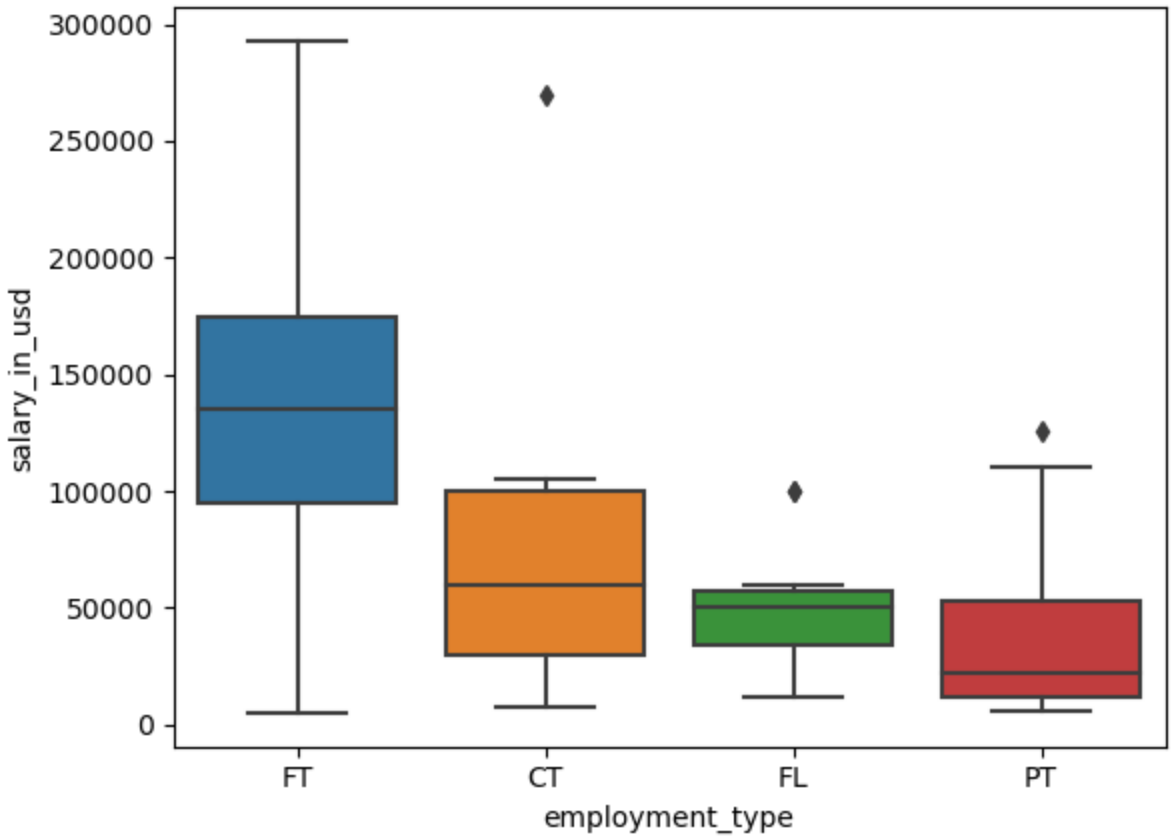

300000
250000
200000
150000
100000
50000
0
salary_in_usd
FT
CT
FL
PT
employment_type

Fig. 9.- Salary by job title after removing outliers

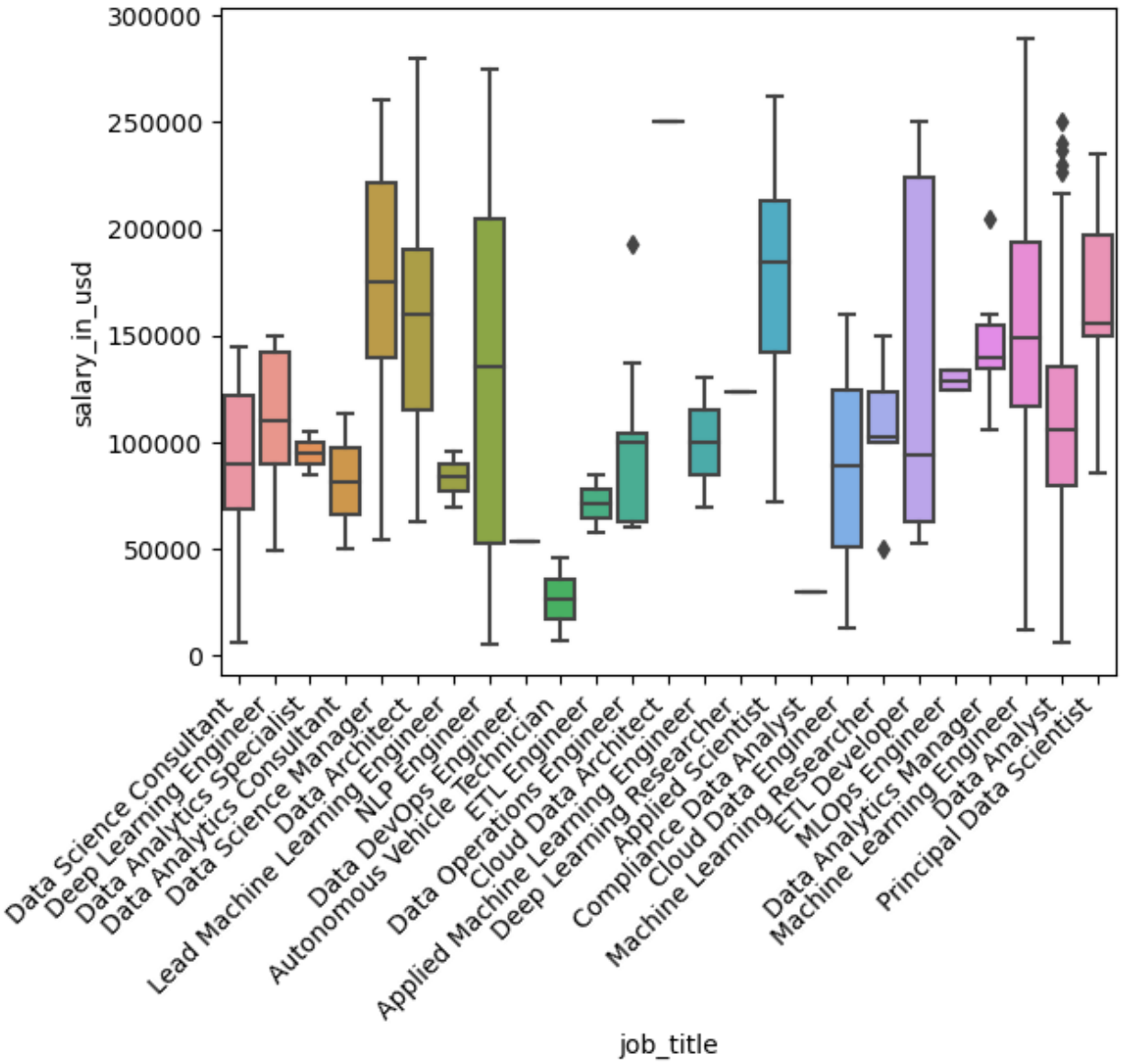


Fig. 10.- Average salary histogram after removing outliers

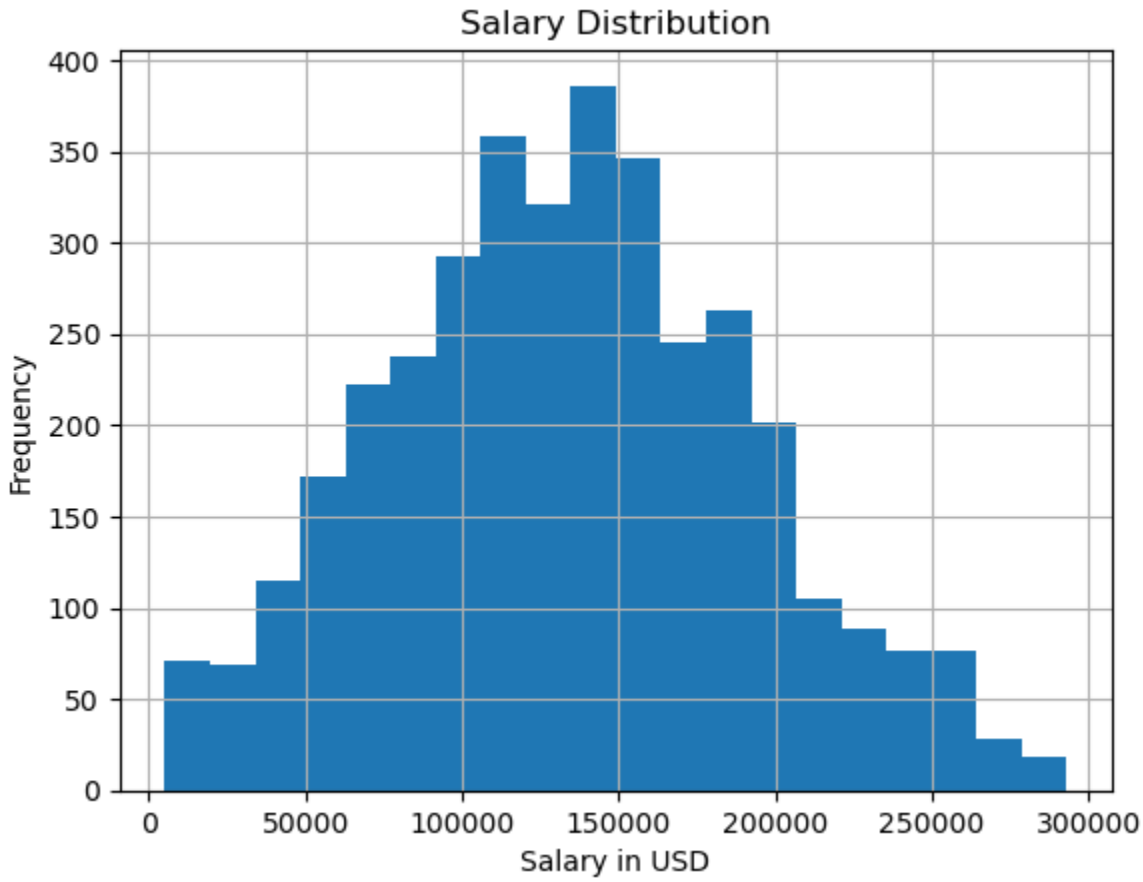


Fig. 11.- Salary distribution for entry-level positions (hypothesis testing)

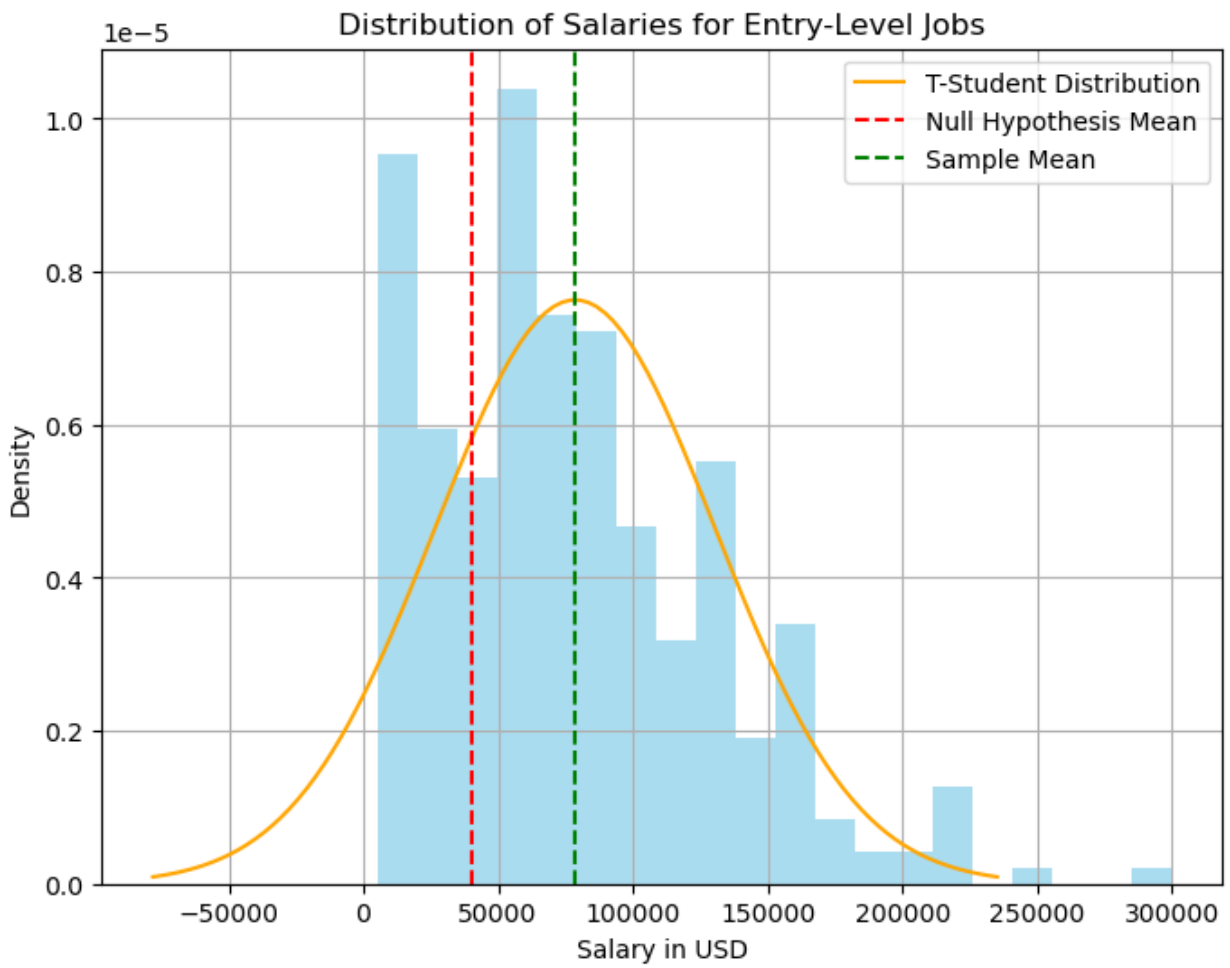


Fig. 12.- Streamlit visual studio code

The code below shows the python script used for launching the streamlit user interface, which we launched in the main_test.py:

```
# from sklearn.models.regression import LinearRegression
import streamlit as st
import pandas as pd
import numpy as np

# dataset = pd.read_csv("¨name_of_file.csv")
# X = dataset[['experience_level', 'employment_type', 'job_title', 'remote_ratio', 'company_size']]
```

```
# y = dataset['salary_in_usd']
# model = LinearRegression()
# model.fit(X, y)

def main():
    st.title("Salary Estimator")

    experience_level = st.selectbox("Select Experience Level", ['SE','MI','EN','EX'])
    employment_type = st.selectbox("Select Employment Type", ['FT','CT','FL','PT'])

    job_title = st.selectbox("Enter Job Title", ['Principal Data Scientist', 'ML Engineer', 'Data
Scientist', 'Applied Scientist',                    'Data Analyst', 'Data Modeler',
'Research Engineer', 'Analytics Engineer',                    'Business Intelligence
Engineer', 'Machine Learning Engineer', 'Data Strategist',
                    'Data Engineer', 'Computer Vision Engineer', 'Data Quality Analyst',
'Compliance Data Analyst',                    'Data Architect', 'Applied Machine
Learning Engineer', 'AI Developer', 'Research Scientist',                    'Data
Analytics Manager', 'Business Data Analyst', 'Applied Data Scientist', 'Staff Data Analyst',
                    'ETL Engineer', 'Data DevOps Engineer', 'Head of Data', 'Data
Science Manager', 'Data Manager',                    'Machine Learning Researcher',
'Big Data Engineer', 'Data Specialist', 'Lead Data Analyst',                    'BI Data
Engineer', 'Director of Data Science', 'Machine Learning Scientist', 'MLOps Engineer',
                    'AI Scientist', 'Autonomous Vehicle Technician', 'Applied Machine
Learning Scientist',                    'Lead Data Scientist', 'Cloud Database
Engineer', 'Financial Data Analyst', 'Data Infrastructure Engineer',
'Software Data Engineer', 'AI Programmer', 'Data Operations Engineer', 'BI Developer', 'Data
Science Lead',
                    'Deep Learning Researcher', 'BI Analyst', 'Data Science
Consultant', 'Data Analytics Specialist',
                    'Machine Learning Infrastructure Engineer', 'BI Data Analyst',
'Head of Data Science', 'Insight Analyst',
```

```
                                'Deep Learning Engineer', 'Machine Learning Software Engineer',
'Big Data Architect', 'Product Data Analyst',
                                'Computer Vision Software Engineer', 'Azure Data Engineer',
'Marketing Data Engineer', 'Data Analytics Lead',
                                'Data Lead', 'Data Science Engineer', 'Machine Learning Research
Engineer', 'NLP Engineer', 'Manager Data Management',
                                'Machine Learning Developer', '3D Computer Vision Researcher',
'Principal Machine Learning Engineer',
                                'Data Analytics Engineer', 'Data Analytics Consultant', 'Data
Management Specialist', 'Data Science Tech Lead',
                                'Data Scientist Lead', 'Cloud Data Engineer', 'Data Operations
Analyst', 'Marketing Data Analyst', 'Power BI Developer',
                                'Product Data Scientist', 'Principal Data Architect', 'Machine
Learning Manager', 'Lead Machine Learning Engineer',
                                'ETL Developer', 'Cloud Data Architect', 'Lead Data Engineer',
'Head of Machine Learning', 'Principal Data Analyst',
                                'Principal Data Engineer', 'Staff Data Scientist', 'Finance Data
Analyst'])

    company_location = st.selectbox("Company Location",
                            ['ES', 'US', 'CA', 'DE', 'GB', 'NG', 'IN', 'HK', 'NL', 'CH',
                             'CF', 'FR', 'FI', 'UA', 'IE', 'IL', 'GH', 'CO', 'SG', 'AU',
                             'SE', 'SI', 'MX', 'BR', 'PT', 'RU', 'TH', 'HR', 'VN', 'EE',
                             'AM', 'BA', 'KE', 'GR', 'MK', 'LV', 'RO', 'PK', 'IT', 'MA',
                             'PL', 'AL', 'AR', 'LT', 'AS', 'CR', 'IR', 'BS', 'HU', 'AT',
                             'SK', 'CZ', 'TR', 'PR', 'DK', 'BO', 'PH', 'BE', 'ID', 'EG',
                             'AE', 'LU', 'MY', 'HN', 'JP', 'DZ', 'IQ', 'CN', 'NZ', 'CL',
                             'MD', 'MT'])

    remote_ratio = st.slider("Remote Work Ratio (%)", min_value=0, max_value=100, value=50,
step=50)
```

```
    user_input = pd.DataFrame({
        'experience_level': [experience_level],
        'employment_type': [employment_type],
        'job_title': [job_title],
        'location': [company_location],
        'remote_ratio': [remote_ratio]

    })

    # predicted_salary = model.predict(user_input)[0]

    # st.write(f"Estimated Average Salary: ${predicted_salary:.2f}")

if __name__ == '__main__':
    main()
```

When declaring the user output variable, which is a result of all the input values, we encounter the following error, which has also been documented on the git hub repository:

https://github.com/carlosruiz-stack/ironhack_final_version/blob/main/VS%20CODE%20ERROR%20OUTPUT.txt

This might be caused by the execution of the sk.lean library for linear regression in visual studio, however, we have verified that the system has the most recent version of python and numpy, the same for the sk.learn library.

Output error:

Please note and check the following:

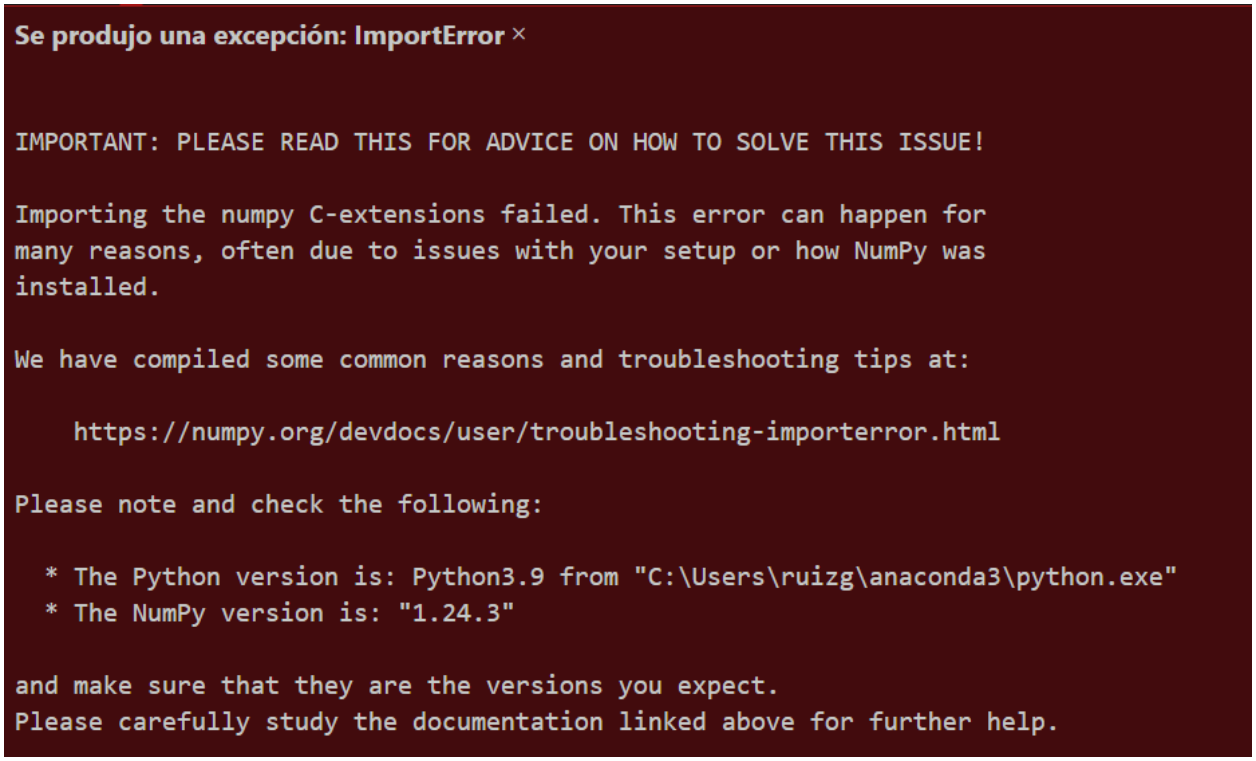
* The Python version is: Python3.9 from "C:\Users\ruizg\anaconda3\python.exe"
* The NumPy version is: "1.24.3"

and make sure that they are the versions you expect.
Please carefully study the documentation linked above for further help.

Original error was: DLL load failed while importing _multiarray_umath: No se puede encontrar el módulo especificado.
ImportError: DLL load failed while importing _multiarray_umath: No se puede encontrar el módulo especificado.

During handling of the above exception, another exception occurred:

File "C:\Users\ruizg\OneDrive\Escritorio\IRONHACK FINAL PROJECT CONTENT\VISUAL STUDIO CODE\main_test.py", line 2, in <module>
import streamlit as st



REFERENCE

Some of the websites with information that we have turned to, in order to validate our procedure for data analysis and machine learning include:

Digitalocean.com
Kaggle
Datacamp
geeks for geeks

Medium.com
Towardsdatascience.com
Educative.io
Appsloveworld
Datatechnotes
Projectpro
askpython.com

LINKS

Link to the Github repo: https://github.com/carlosruiz-stack/ironhack_final_version

Presentation with summarised information:
https://docs.google.com/presentation/d/1Fkux5Rm5JF0YTFGpBE3wzlHvlDtJzIyf4e35iOTWJB8/edit?usp=sharing

Model performance by scenario and transformation performed. MAE, MSE, RMSE, R2

https://docs.google.com/spreadsheets/d/1zO-v0txTPsPpe7QUkAH9HvqeNoJQ-td7DdO8VOylvxw/edit?usp=sharing

Tableau data visualisation
https://public.tableau.com/views/ITJOBOPENINGSDASHBOARD/Sheet2?:language=es-ES&:display_count=n&:origin=viz_share_link