ĐẠI HỌC KHOA HỌC TỰ NHIÊN
UNIVERSITY OF SCIENCES

BÁO CÁO ĐỒ ÁN CUỐI KÌ

NHẬP MÔN KHOA HỌC DỮ LIỆU
LỚP 20KHDL2
HKI (2022-2023)

Sinh viên thực hiện: 18127100 - Võ Chí Hiếu
19127395 - Phan Đức Hiển
20127279 - Trần Thị Thanh Phú
20127631 - Thái Văn Thiên

Giáo viên hướng dẫn: GVTH - Kiều Vũ Minh Đức
GVTG - Lê Nhựt Nam

Thành phố Hồ Chí Minh, tháng 1 năm 2023

I. Thu thập dữ liệu

1. Dữ liệu mang chủ đề gì ? Được lấy từ nguồn nào ?

- Dữ liệu có chủ đề liên quan đến công việc về IT.
- Được crawl từ web itviec.com và xử lý để đưa về dạng file csv.

2. Thu thập dữ liệu như thế nào ?

- Dùng thư viện selenium và BeautifulSoup
- Đăng nhập vào website itviec.com và lấy link các bài đăng tuyển dụng theo level và theo các skill.
- Xử lý đưa về dạng mảng các dữ liệu.
- Viết dữ liệu vào file csv.
- Xuất file csv để thực hiện làm việc với dữ liệu trên file này.

II. Khám phá dữ liệu

1. Dữ liệu ban đầu có số hàng và số cột là bao nhiêu ?

- Số hàng: 1200
- Số cột: 13

2. Ý nghĩa các cột như thế nào ?

- Các cột về kỹ năng (java, ..., angular): thể hiện kĩ năng mà một công việc yêu cầu.
- Cột về kinh nghiệm (level): thể hiện trình độ mà một công việc yêu cầu.
- Cột về lương (salary): thể hiện mức lương có thể đạt được của một công việc.

3. Trong bảng dữ liệu có bao nhiêu hàng trùng nhau ?

- Số hàng trùng: 679

=> *Thực hiện lọc trùng.*

4. Xử lý dữ liệu:

-Vì khi thể hiện cột dữ liệu về mức lương, thấy có nhiều dữ liệu khác nhau không thể hiện đúng yêu cầu dữ liệu về mức lương (là một con số hoặc khoảng dữ liệu số). Nên can thực hiện xử lý ở bước này, chỉ lấy những dữ liệu có số và những dữ liệu này đều được đưa về đơn vị USD.

- Sẽ có những hàng, giá trị về mức lương nằm trong khoảng nhất định. Nên sẽ được xử lý cột salary về dạng mức lương trung bình.
- Chuyển đổi các mức kinh nghiệm về dạng số: fresher thành 0, junior thành 1 và senior thành 2.
- Bảng dữ liệu sau khi xử lý ở bước này sẽ còn 377 hàng và 13 cột. Bảng dữ liệu này sẽ được dùng để phân tích và mô hình hóa vào những bước sau.

5. Kiểu dữ liệu của các cột dữ liệu là gì ?

- Các cột dữ liệu đều có kiểu dữ liệu là object.
- Riêng cột salary có kiểu dữ liệu không phù hợp nên cần đưa về kiểu dữ liệu float để xử lý.

6. Sự phân bổ dữ liệu như thế nào ?

-Salary:

	average_salary
missing_ratio	0.0
min	400.0
lower_quartile	1150.0
median	1500.0
upper_quartile	2000.0
max	4500.0

-Skills:

	java	nodejs	reactjs	ruby	android	ios	php	python	c++	golang	angular
missing_ratio	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
num_diff_vals	2	2	2	2	2	2	2	2	2	2	2
diff_vals	[1, 0]	[0, 1]	[0, 1]	[0, 1]	[0, 1]	[0, 1]	[0, 1]	[0, 1]	[0, 1]	[0, 1]	[0, 1]

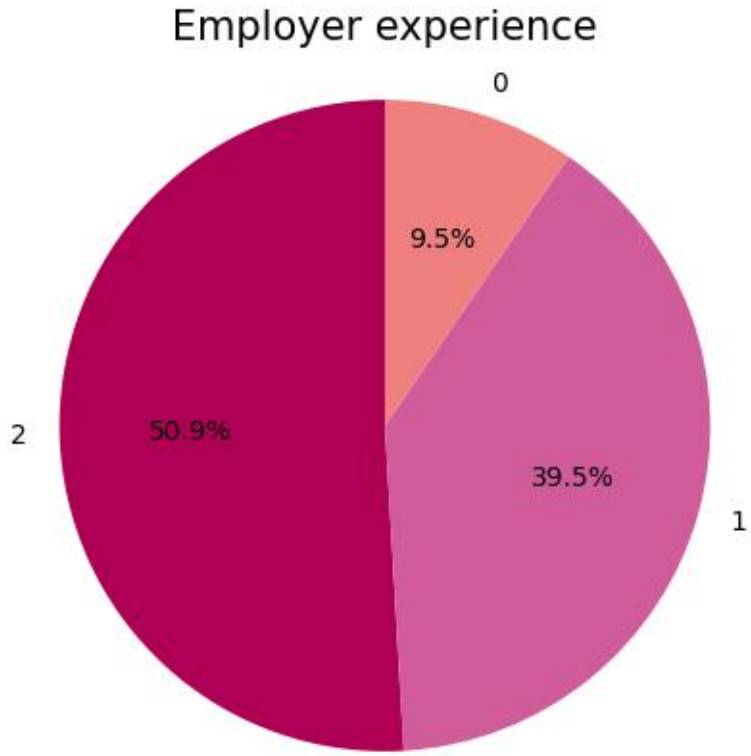
-Level:

```
missing_ratio                0.0
num_diff_vals                  3
diff_vals          [0.0, 1.0, 2.0]
Name: level, dtype: object
```

III. Phân tích và trực quan hóa dữ liệu

Ở phần này, các câu hỏi có ý nghĩa về dữ liệu sẽ được đặt ra và sau đó sẽ được trực quan hóa để tìm ra câu trả lời.

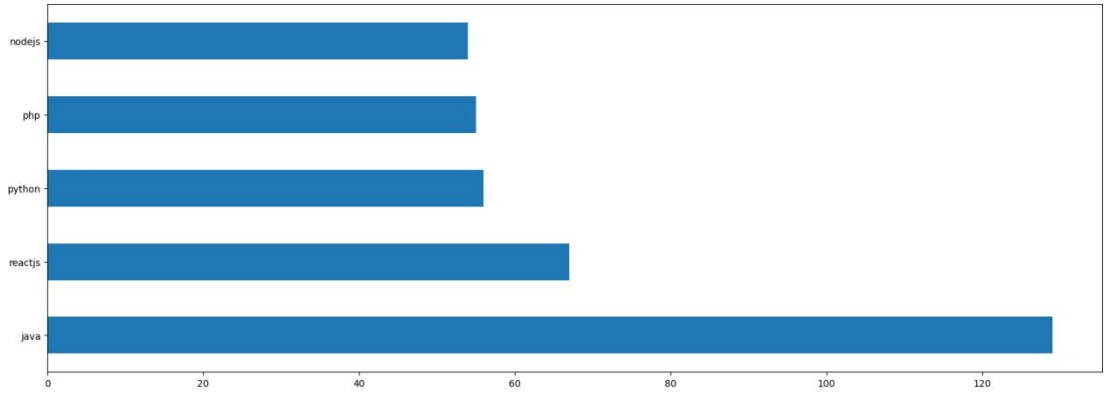
1. Tỉ lệ của các mức kinh nghiệm được phân bố như thế nào ? Mức kinh nghiệm nào chiếm đa số ?



Trả lời:

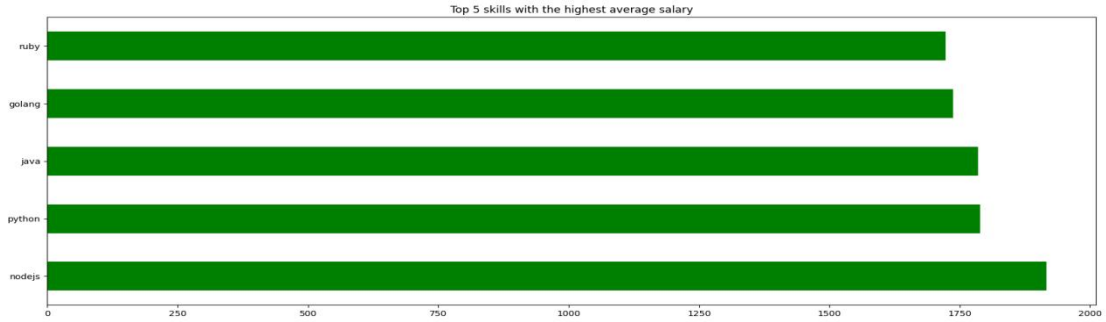
- Các level “senior”, “junior”, “fresher” lần lượt chiếm: 50.9 %, 39.5 % và 9.5 %
- Senior là loại level chiếm tỉ lệ nhiều nhất, tiếp đến là junior và cuối cùng là fresher.

2. Năm loại kĩ năng phổ biến nhất là ?



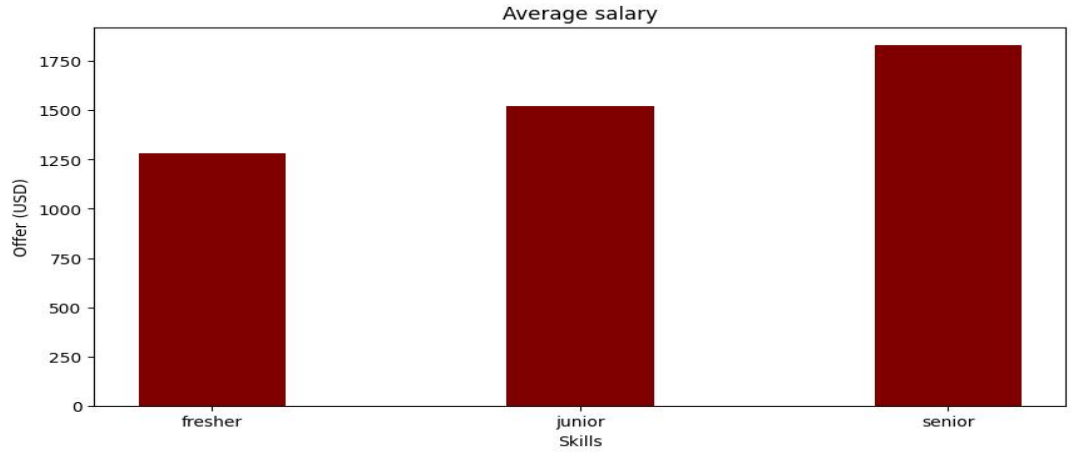
Trả lời:
- 5 loại kỹ năng phổ biến nhất lần lượt là: java, nodejs, pyhton, php và reactjs.

3. Năm loại kĩ năng với mức lương trung bình cao nhất là ?



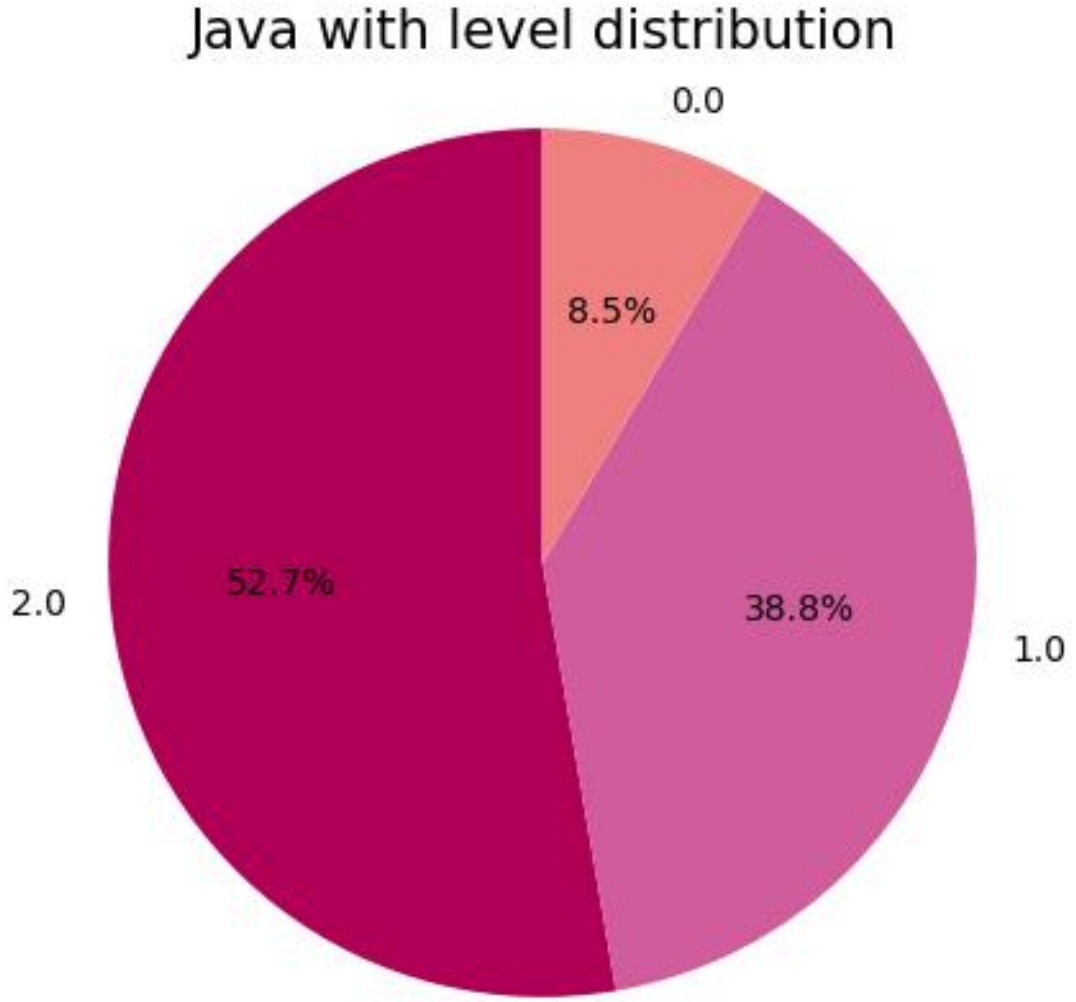
Trả lời:
- 5 loại kỹ năng có mức lương trung bình cao nhất lần lượt là: nodejs, python, java, golang và ruby.

4. Mỗi level có mức lương trung bình là bao nhiêu ? Level nào có mức lương cao nhất ?



Trả lời:
- Mức lương trung bình của các "level" lần lượt là: 1283 USD, 1520 USD, 1831 USD tương ứng với fresher, junior, senior.
- Senior có mức lương cao nhất, tiếp đến là junior và fresher có mức lương thấp nhất.

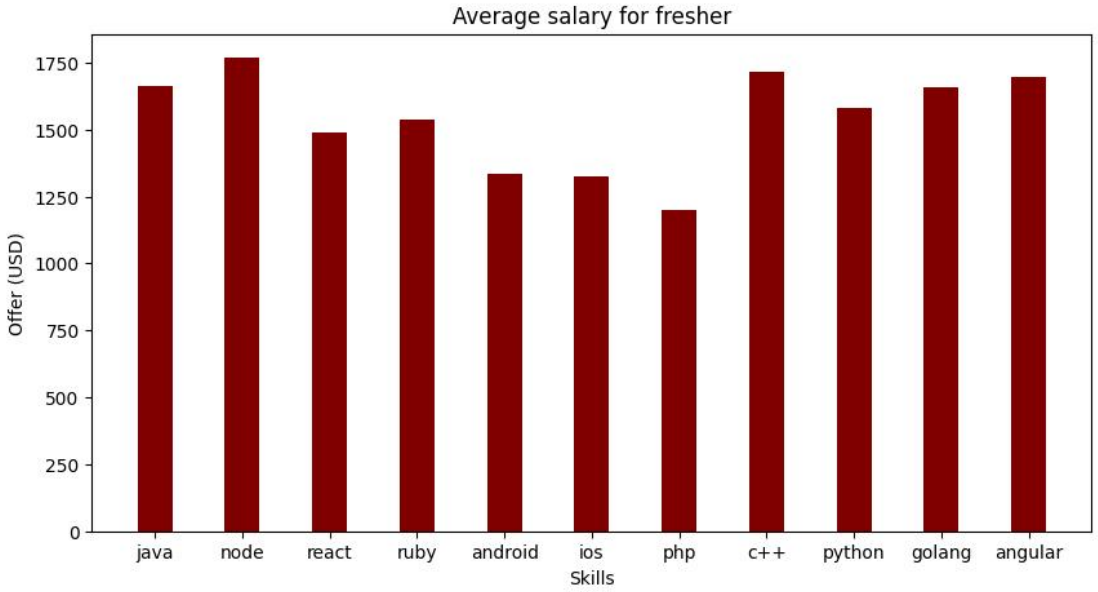
5. Thực hiện phân tích cột kĩ năng "java", tỉ lệ của các mức level được phân bố như thế nào ? Level nào chiếm ưu thế ?



Trả lời:

- Các level "senior", "junior", "fresher" lần lượt chiếm: 52.7 %, 38.8 % và 8.5 %
- Senior là loại level chiếm tỉ lệ nhiều nhất, tiếp đến là junior và cuối cùng là fresher.

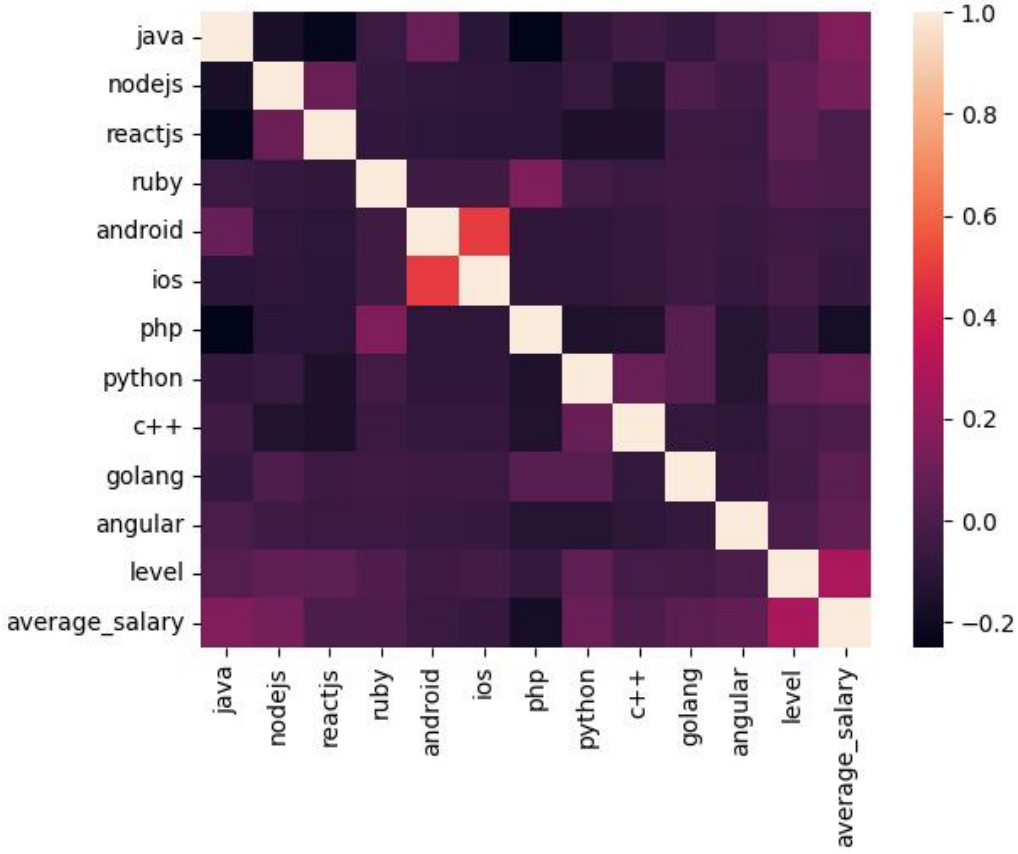
6. Với kinh nghiệm là “fresher” thì sẽ có mức lương trung bình như thế nào giữa các kĩ năng ?



Trả lời:

- Mức lương cao nhất, thấp nhất là đối với kĩ năng Nodejs và PHP.

7. Mức độ tương quan giữa các cột ?



Trả lời:
- Kĩ năng Android và IOS có sự tương quan lớn với nhau.
- Level và average salary cũng có sự tương quan lớn với nhau.

=> Điều này thể hiện:
- Khi một công việc yêu cầu các kĩ năng nào đó thì android và ios sẽ đi cùng với nhau. Điều này đúng với thực tế, vì 2 kĩ năng này thuộc về lập trình app.
- Mức lương và kinh nghiệm sẽ có sự liên quan trực tiếp, kinh nghiệm càng nhiều thì mức lương càng cao và ngược lại.

IV. Mô hình dự đoán mức lương

- Chọn 2 tập dữ liệu X và Y. Trong đó X thể hiện các kỹ năng, Y thể hiện mức lương trung bình.
- Chia tập train và tập test của 2 tập dữ liệu X và Y. Train là 70% và test là 30%
- Sử dụng mô hình Linear Regression để tìm ra được 2 giá trị Intercept và Coefficients.
- Thay các giá trị tính được vào hàm dự đoán lương.
- Thực hiện dự đoán lương thực tế dựa theo các kỹ năng.