

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH  
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN  
KHOA CÔNG NGHỆ THÔNG TIN



# BÁO CÁO CUỐI KỲ TRỰC QUAN HÓA DỮ LIỆU

| CHỦ ĐỀ |  
**PHÂN TÍCH TÌNH HÌNH VIỆC LÀM Ở VIỆT NAM**

| GIẢNG VIÊN HƯỚNG DẪN |  
**TS. BÙI TIẾN LÊN**

| NHÓM 1 |

21120242 – Huỳnh Thị Kiều Hoa  
21120247 – Nguyễn Văn Quang Hưng  
21120254 – Lưu Chấn Huy  
21120298 – Chiêm Bình Nguyên  
21120302 – Huỳnh Trí Nhân

**Thành phố Hồ Chí Minh - 2024**

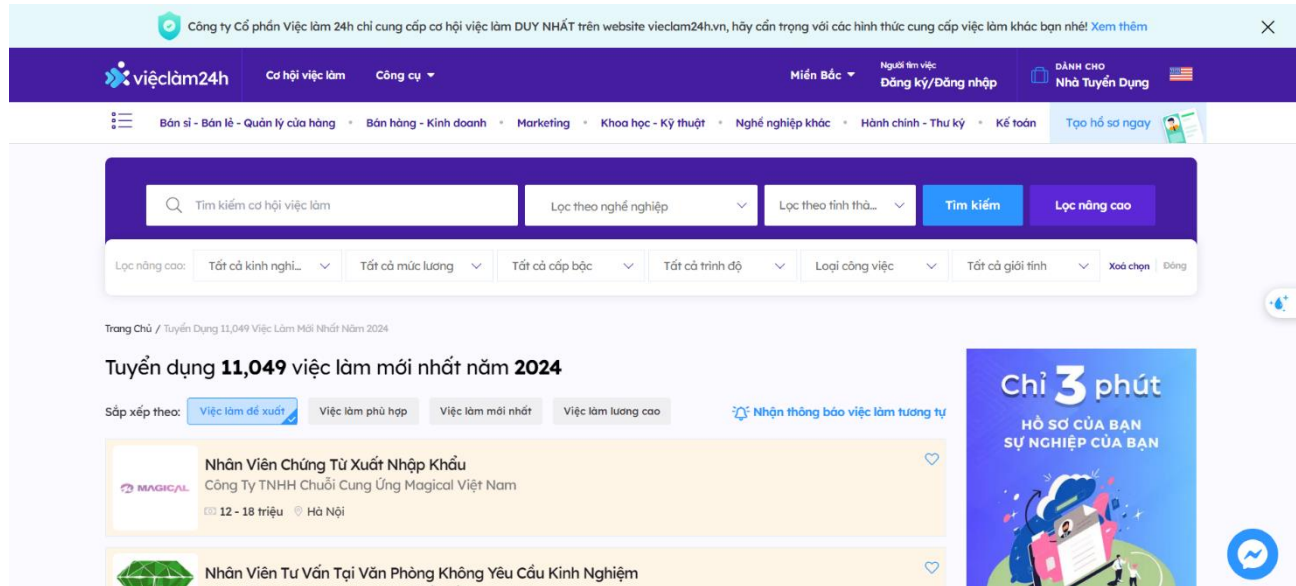
# MỤC LỤC

<b>MỤC LỤC.....</b>	<b>2</b>
<b>1. GIỚI THIỆU ĐỀ TÀI.....</b>	<b>4</b>
1.1 NGUỒN DỮ LIỆU .....	4
1.2 THU THẬP VÀ TIỀN XỬ LÝ DỮ LIỆU .....	5
1.3 CÔNG CỤ TRỰC QUAN HÓA DỮ LIỆU .....	7
<b>2. TRỰC QUAN HÓA BẰNG BIỂU ĐỒ .....</b>	<b>8</b>
2.1 Biểu đồ trang 1 .....	8
2.1.1 Biểu đồ Histogram thể hiện lương trung bình .....	8
2.1.2 Biểu đồ tròn thể hiện số lượng công việc theo bằng cấp .....	9
2.1.3 Biểu đồ thể hiện số lượng công việc theo nhóm tuổi.....	9
2.1.4 Bản đồ thể hiện số lượng công việc theo khu vực tuyển .....	9
2.2 Biểu đồ trang 2 .....	10
2.2.1 Histogram thể hiện sự phân bố mức lương trung bình theo giới tính .....	11
2.2.2 Boxplot phân bố mức lương theo giới tính .....	12
2.2.3 Số lượng công việc theo cấp bậc .....	13
2.2.4 Hình thức làm việc.....	14
2.2.5 Yêu cầu về bằng cấp .....	15
2.2.6 Yêu cầu kinh nghiệm .....	16
2.2.7 Lọc dữ liệu phân tích .....	17
2.3 Biểu đồ trang 3 .....	18
2.3.1 Biểu đồ Top 10 ngành nghề có tỉ lệ công việc lương cao.....	18

2.3.2 Số bài đăng theo ngày .....	19
2.3.3 Những từ khoá phổ biến.....	20
2.3.4 Những công việc được quan tâm nhiều nhất .....	21
2.4 Trang tìm kiếm .....	22
3. MÔ HÌNH HÓA DỮ LIỆU VÀ TÍCH HỢP AI.....	24
3.1 MÔ HÌNH HÓA DỮ LIỆU .....	24
3.1.1 Trích xuất đặt trung .....	24
3.1.2 Chuẩn bị dữ liệu.....	24
3.1.3 Mô hình hóa dữ liệu .....	25
3.1.4 Mô hình dự đoán mức lương .....	26
3.1.5 Mô hình dự đoán mức độ quan tâm .....	27
3.1.6 Kết quả mô hình hóa .....	28
3.2 TÍCH HỢP AI .....	29
3.2.1 Mô hình sử dụng .....	29
3.2.2 Hướng dẫn sử dụng.....	29
4. ĐÁNH GIÁ DASHBOARD QUA CÁC TIÊU CHÍ.....	31

# 1. GIỚI THIỆU ĐỀ TÀI

## 1.1 NGUỒN DỮ LIỆU



- Tại Việt Nam hiện nay, rất nhiều nhà tuyển dụng lựa chọn đăng thông tin tuyển dụng lên các trang web việc làm như: <https://vieclam24h.vn/>, <https://www.topcv.vn/viec-lam>, <https://careerbuilder.vn/tim-viec-lam.html>, <https://www.vieclamtot.com/viec-lam>,... Nhóm quyết định thu thập dữ liệu từ trang web [Việc làm 24h](https://vieclam24h.vn/) cho đề án cuối kỳ này.
- Công cụ sử dụng:
  - o Python
  - o Các thư viện: request\_html, BeautifulSoup
- Quá trình thu thập dữ liệu được mô tả cụ thể trong file [...\notebooks\1\\_0\\_data\\_collecting.ipynb](#), dữ liệu thu thập được được lưu tại [...\data\raw\\_dataset.csv](#)
- Sau đó, nhóm thực hiện tiền xử lý và có những khám phá sơ lược về dữ liệu. Quá trình này cụ thể trong [...\notebooks\2\\_0\\_data\\_preprocessing.ipynb](#), dữ liệu sau tiền xử lý được lưu tại [...\notebooks\cleaned\\_dataset.csv](#)

## 1.2 THU THẬP VÀ TIỀN XỬ LÝ DỮ LIỆU

```
Data columns (total 19 columns):
#      Column      Non-Null Count  Dtype
---  -
0      Liên kết      14994 non-null  object
1      Tên công ty     14994 non-null  object
2      Tên công việc   14994 non-null  object
3      Ngày cập nhật   14994 non-null  object
4      Lượt xem        14974 non-null  float64
5      Mức lương       14994 non-null  object
6      Khu vực tuyển   14994 non-null  object
7      Yêu cầu giới tính 4404 non-null  object
8      Cấp bậc         14994 non-null  object
9      Thời gian thử việc 6233 non-null  object
10     Số lượng tuyển  14993 non-null  float64
11     Hình thức làm việc 14994 non-null  object
12     Độ tuổi         5126 non-null  object
13     Yêu cầu bằng cấp 9164 non-null  object
14     Yêu cầu kinh nghiệm 14994 non-null  object
15     Ngành nghề      14994 non-null  object
16     Từ khóa         14994 non-null  object
17     Địa chỉ công ty 13193 non-null  object
18     Quy mô công ty  14994 non-null  object
dtypes: float64(2), object(17)
```

Dữ liệu ban đầu gồm 19 cột với 14994 dòng dữ liệu. Tiếp theo chúng em sẽ tiến xử lý dữ liệu qua các trường

- Xóa những dòng dữ liệu bị thiếu ở Lượt xem, định dạng lại cột Ngày cập nhật.
- Tên công ty: Phân tách thành “Loại công ty” Trách nhiệm hữu hạn hay Công ty cổ phần.
- Mức lương: “18-25 triệu” thì chúng ta chỉ lấy lương trung bình để xem xét. Bỏ đi các outliers của Lương trung bình như khoảng các quá lớn giữa lương cao nhất và lương thấp nhất.
- Độ tuổi: Dữ liệu thuộc các khoảng như cột mức lương nên chúng ta sẽ lấy độ tuổi trung bình và xếp vào các nhóm tuổi Tuổi trẻ, Trưởng thành, Trung niên, tuổi cao, và Không yêu cầu đặc biệt đối với những dữ liệu thiếu.
- Yêu cầu bằng cấp: Điền những giá trị thiếu bằng Không yêu cầu cụ thể.
- Ngành nghề có định dạng là 'Công việc chính/Công việc liên quan 1/Công việc liên quan 2' nên chúng ta tách thành 3 cột.
- Sau đó xóa những cột không cần thiết.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 13704 entries, 0 to 13703
Data columns (total 16 columns):
```

#	Column	Non-Null Count	Dtype
0	Ngày cập nhật	13704 non-null	object
1	Lượt xem	13704 non-null	float64
2	Khu vực tuyển	13704 non-null	object
3	Yêu cầu giới tính	13704 non-null	object
4	Cấp bậc	13704 non-null	object
5	Số lượng tuyển	13704 non-null	float64
6	Hình thức làm việc	13704 non-null	object
7	Yêu cầu bằng cấp	13704 non-null	object
8	Yêu cầu kinh nghiệm	13704 non-null	object
9	Từ khóa	13704 non-null	object
10	Loại công ty	13704 non-null	object
11	Lương trung bình	13704 non-null	float64
12	Nhóm tuổi	13704 non-null	object
13	Công việc chính	13704 non-null	object
14	Công việc liên quan 1	13704 non-null	object
15	Công việc liên quan 2	13704 non-null	object

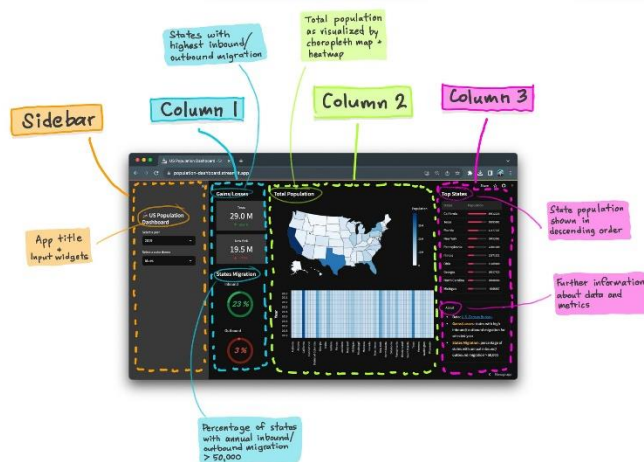
dtypes: float64(3), object(13)

Sau khi thu thập và tiền xử lý dữ liệu, nhóm thu được dữ liệu gồm các trường sau:

- Ngày cập nhật
- Lượt xem
- Khu vực tuyển
- Yêu cầu giới tính
- Cấp bậc
- Số lượng tuyển
- Hình thức làm việc
- Yêu cầu bằng cấp
- Yêu cầu kinh nghiệm
- Từ khóa
- Loại công ty
- Lương trung bình
- Nhóm tuổi
- Công việc chính
- Công việc liên quan 1
- Công việc liên quan 2

Dữ liệu gồm tất cả 16 cột và 13704 dòng dữ liệu.

## 1.3 CÔNG CỤ TRỰC QUAN HÓA DỮ LIỆU



**Streamlit** là một framework mã nguồn mở dành cho Python, được thiết kế đặc biệt xây dựng các ứng dụng web tương tác. Được ra mắt vào năm 2019, Streamlit đã nhanh chóng trở thành một công cụ phổ biến trong cộng đồng khoa học dữ liệu nhờ vào sự đơn giản và mạnh mẽ của nó.

Một số thành phần của Streamlit có thể nói đến như là:

- **Sidebar** : Giúp ta tạo ra một thanh bên để chứa các yếu tố giao diện người dùng. Chúng ta có thể sử dụng sidebar để tạo các điều khiển như thanh trượt, hộp chọn, nút bấm, văn bản nhập liệu và nhiều hơn nữa.
- **Columns** : Giúp ta chia bố cục của ứng dụng thành nhiều cột, cho phép sắp xếp các yếu tố giao diện theo hàng ngang.
- Ngoài ra Streamlit có thể tích hợp mạnh mẽ với Plotly để tạo ra các biểu đồ đẹp mắt và tương tác trực tiếp trên ứng dụng web của mình.

### Ưu điểm của Streamlit:

- Dễ dàng sử dụng.
- Không yêu cầu kinh nghiệm lập trình web.
- Tương tác mạnh mẽ với nhiều thư viện Python.
- Tùy chỉnh dễ dàng.

Với tất cả những ưu điểm trên, **Streamlit** là một công cụ mạnh mẽ và tiện lợi để xây dựng các ứng dụng trực quan hóa dữ liệu và học máy, giúp các ta dễ dàng tạo ra các sản phẩm hữu ích và hấp dẫn.

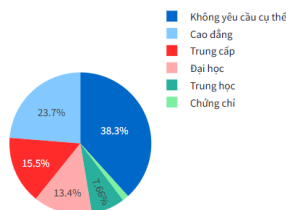
## 2. TRỰC QUAN HÓA BẰNG BIỂU ĐỒ

### 2.1 Biểu đồ trang 1

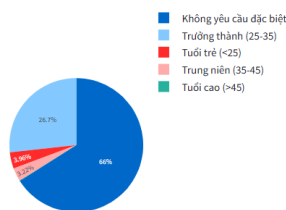
#### Tổng quan



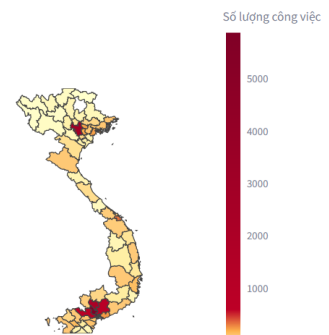
#### Số lượng công việc theo bằng cấp



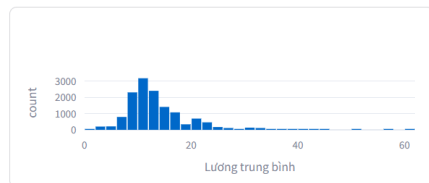
#### Số lượng công việc theo nhóm tuổi



#### Số lượng công việc tuyển theo

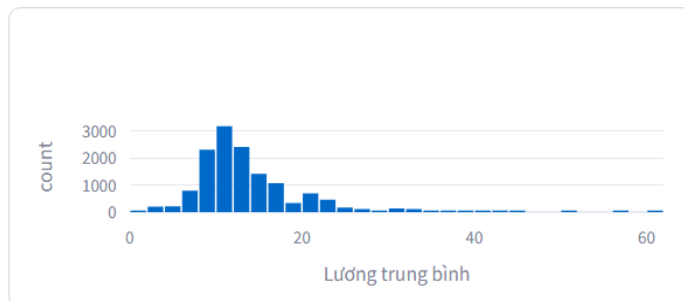


#### Lương trung bình



#### 2.1.1 Biểu đồ Histogram thể hiện lương trung bình

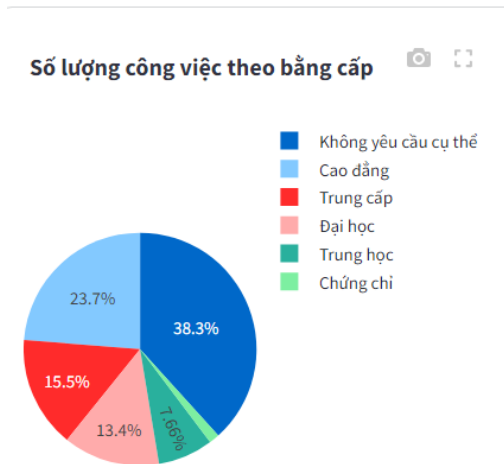
##### Lương trung bình



- Biểu đồ Histogram thể hiện lương trung bình
- Từ biểu đồ ta có thể thấy mức lương nhiều công việc tuyển nhất là từ 10-11 triệu

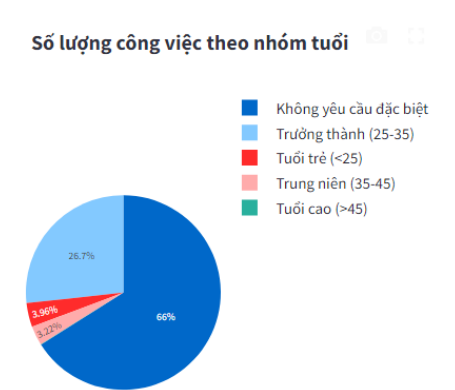


### 2.1.2 Biểu đồ tròn thể hiện số lượng công việc theo bằng cấp



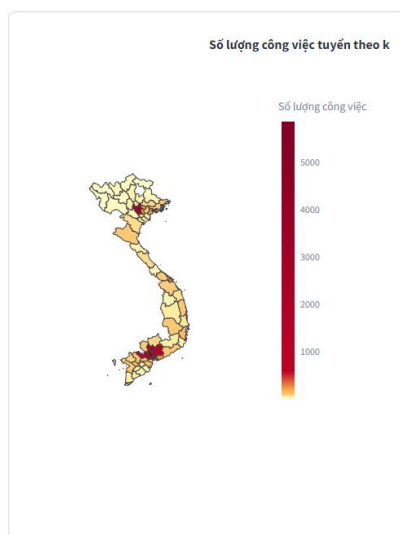
- Biểu đồ tròn thể hiện số lượng công việc theo bằng cấp
- Dễ thấy, các công việc hầu hết không có yêu cầu bằng cấp cụ thể
- Nhiều công việc yêu cầu bằng từ cao đẳng trở lên
- Rất ít thấy các công việc chỉ yêu cầu chứng chỉ

### 2.1.3 Biểu đồ thể hiện số lượng công việc theo nhóm tuổi



- Biểu đồ tròn thể hiện số lượng công việc theo nhóm tuổi
- Dễ thấy, các công việc hầu hết không có yêu cầu tuổi tác
- Ngoài nhóm công việc không yêu cầu ra, phần lớn công việc tuyển ứng viên trong độ tuổi 25-35, đây là nhóm tuổi đã có kinh nghiệm làm việc nhưng cũng không quá lớn tuổi
- Cực kì ít công việc tuyển người đã trên 45 tuổi

### 2.1.4 Bản đồ thể hiện số lượng công việc theo khu vực tuyển



- Có thể thấy được có hai khu vực chính có số lượng tuyển vượt trội so với những khu vực còn lại là Tp.Hà Nội, Tp.HCM và các tỉnh lân cận hai thành phố này.

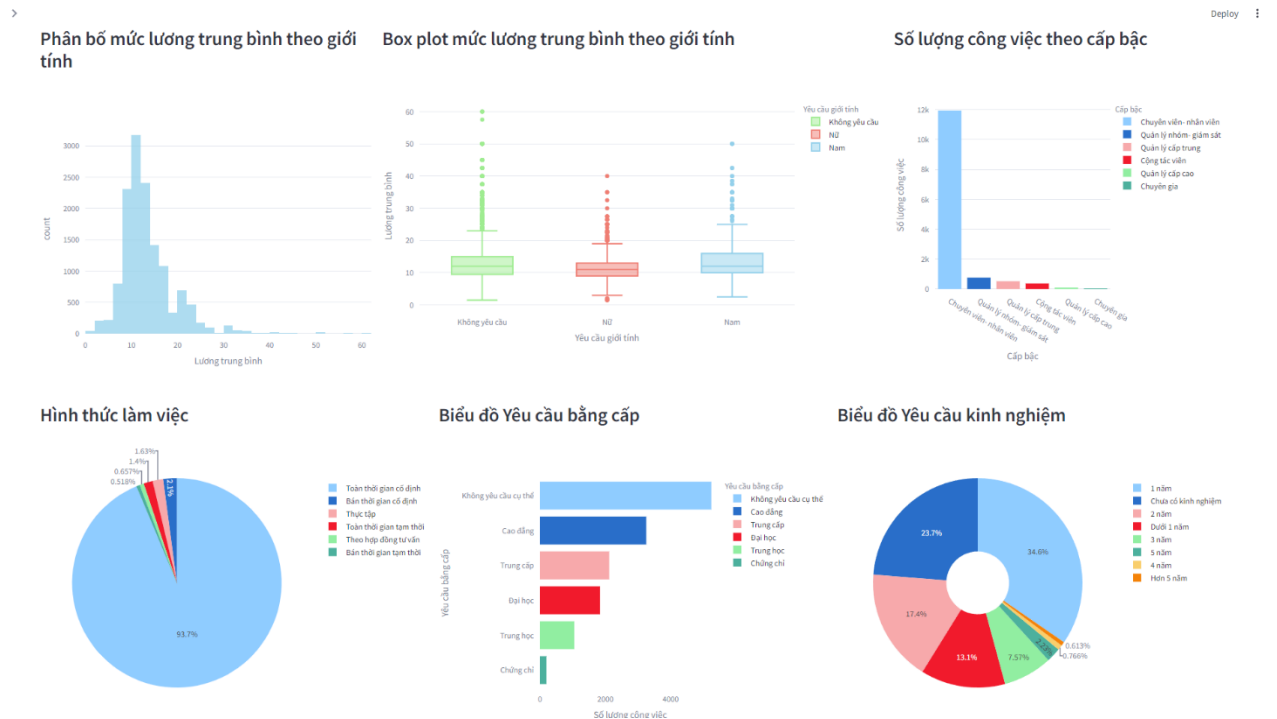
- Điều này cho thấy nhu cầu tuyển dụng ở hai miền Nam và Bắc có phần cao hơn so với miền Trung khi màu sắc chủ yếu của khu vực này là màu vàng và cam nhạt

➔ Dựa vào biểu đồ trên ta có thể thấy được sự khác biệt rõ rệt về số lượng công việc tuyển dụng giữa các tỉnh thành qua đó cung cấp cái nhìn tổng quan về thị trường lao động ở Việt Nam giúp người tuyển dụng hay người lao động có thể có những chiến lược, cạnh tranh phù hợp tìm kiếm cơ hội việc làm. Biểu đồ trang 2

Hình 1. Bản đồ nhu cầu việc làm

## 2.2 Biểu đồ trang 2

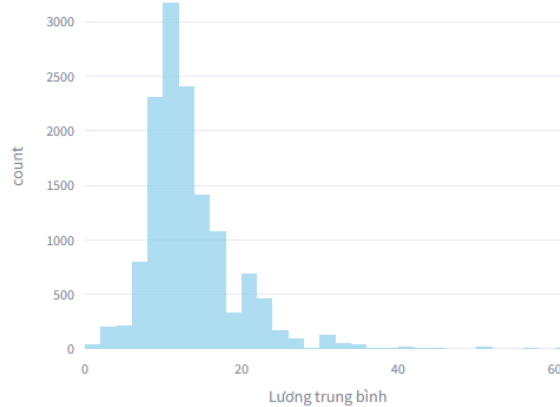
Trang 2 của Dashboard sẽ đi sâu vào phần phân tích về lương của các công việc



Hình 2. Page 2

### 2.2.1 Histogram thể hiện sự phân bố mức lương trung bình theo giới tính

Phân bố mức lương trung bình

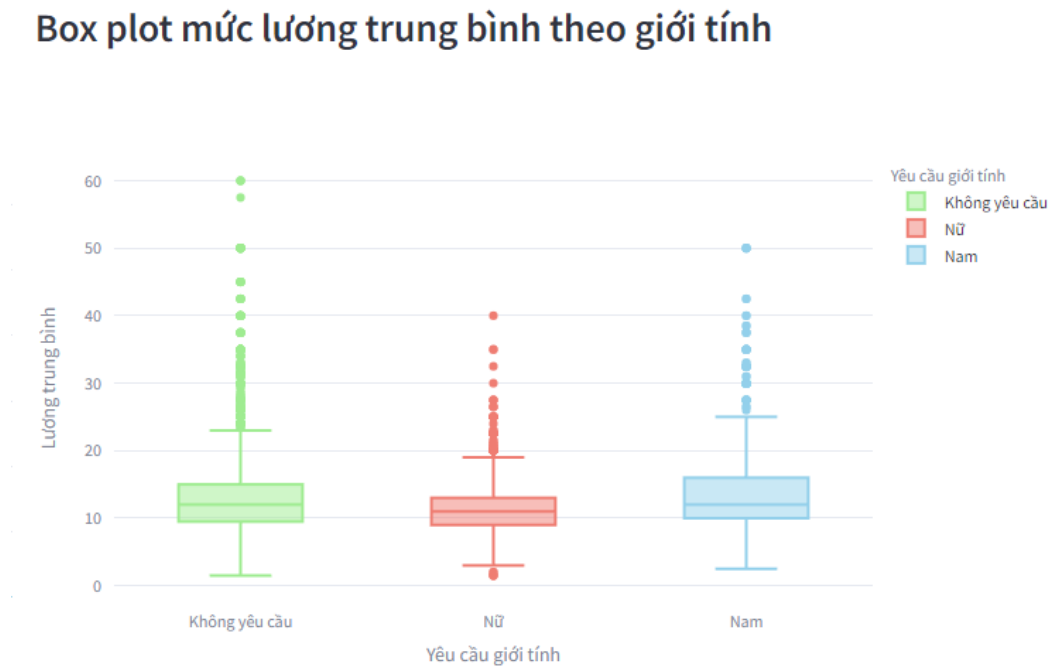


Hình 3. Phân bố mức lương

#### Nhận xét:

- Biểu đồ lệch phải, đuôi dài cho thấy những công việc với mức lương từ 30 triệu trở lên xuất hiện rất ít
  - Tập trung nhiều trong khoảng từ 10 đến 20 triệu
  - Có sự đa dạng về mức lương khi mức lương trải dài từ 10 đến hơn 50 triệu/tháng
- ➔ **Phân bố này phản ánh khá rõ thực tế thị trường lao động, nơi các vị trí như nhân viên – chuyên viên phổ biến hơn nhiều với mức lương trung bình trong khoảng 10 – 20 triệu trong khi đó các vị trí như quản lý, giám đốc với mức lương cao hơn thì ít phổ biến hơn.**

## 2.2.2 Boxplot phân bố mức lương theo giới tính



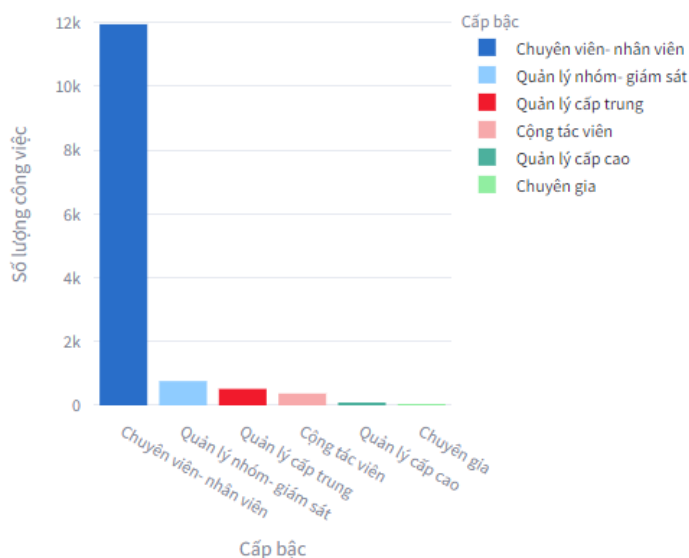
Hình 4. Mức lương trung bình theo giới tính

### Nhận xét:

- Mức lương trung bình của các công việc yêu cầu giới tính nữ thấp hơn và phân bố hẹp hơn so với hai nhóm còn lại
  - Nhóm nam và không yêu cầu có độ phân bố rộng, mức lương trung bình cao, đặc biệt xuất hiện nhiều dữ liệu ngoại lai với mức lương rất cao so với phần còn lại của bộ dữ liệu.
  - Nhìn chung thì mức lương của cả ba nhóm đều nằm trong khoảng từ 10 – 20 triệu.
- ➔ Mức lương trung bình của các công việc yêu cầu giới tính nữ thấp hơn so với hai nhóm còn lại thể hiện có sự chênh lệch về mức lương dựa theo giới tính tại một vài ngành nghề, vị trí nhất định.

### 2.2.3 Số lượng công việc theo cấp bậc

Số lượng công việc theo cấp bậc



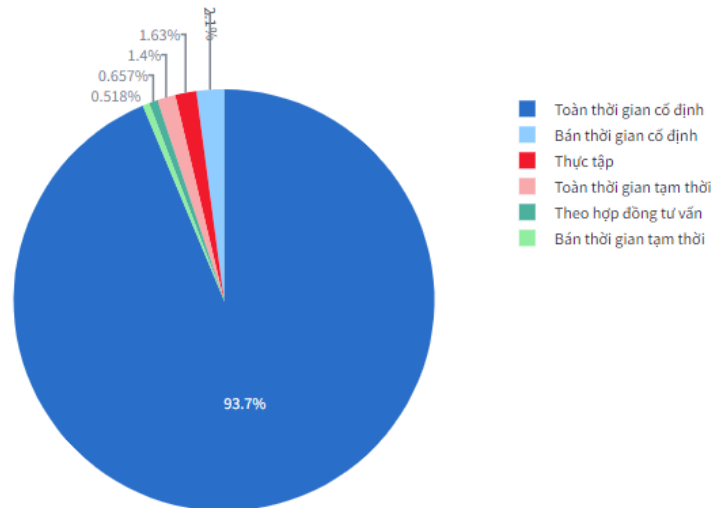
Hình 5. Số lượng công việc theo cấp bậc

#### Nhận xét:

- Biểu đồ cho thấy sự phân bố số lượng công việc dựa theo các cấp bậc khác nhau. Phần lớn các vị trí tuyển dụng đều là Chuyên viên – Nhân viên.
  - Cấp bậc "Chuyên viên - nhân viên" chiếm số lượng lớn nhất, với hơn 12,000 công việc. Điều này cho thấy rằng các công ty có nhu cầu cao đối với các vị trí ở cấp bậc này, thường là các vị trí cơ bản hoặc trung cấp yêu cầu kỹ năng chuyên môn.
  - Các vị trí như Quản lý – Chuyên gia,... có số lượng tuyển ít hơn khá nhiều thể hiện độ ít phổ biến của các vị trí này cũng như việc các công ty có khá ít nhu cầu tuyển dụng cho các vị trí này.
- ➔ Các ứng viên tìm kiếm công việc ở vị trí Chuyên viên – Nhân viên sẽ có nhiều cơ hội hơn. Đối với các vị trí như Quản lý – Chuyên gia với ít cơ hội việc làm hơn thì sẽ cần có chiến lược tìm việc cụ thể hơn hay sẽ phải cạnh tranh nhiều hơn.

## 2.2.4 Hình thức làm việc

Hình thức làm việc



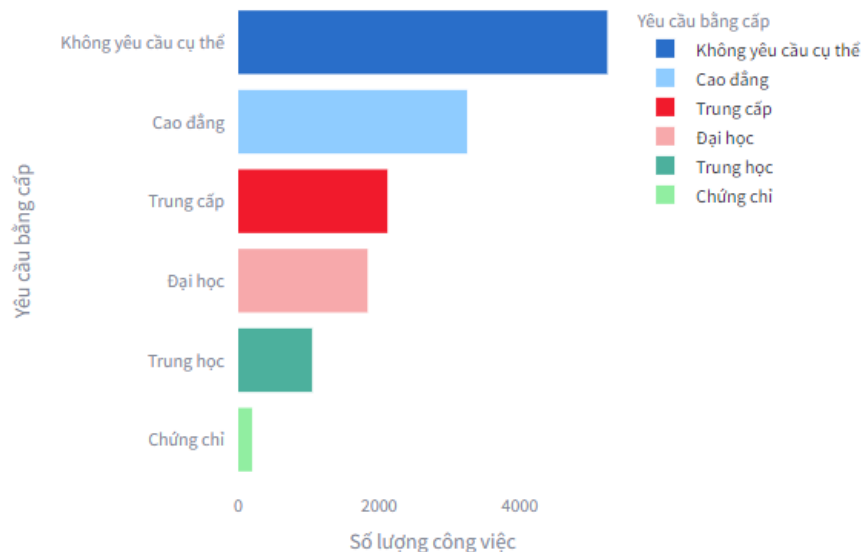
Hình 6. Hình thức làm việc

### Nhận xét:

- Hình thức Toàn thời gian cố định là hình thức làm việc chủ yếu và chiếm 93.7% số vị trí tuyển dụng
  - Hình thức "Toàn thời gian cố định" là lựa chọn chủ yếu của các công ty khi tuyển dụng nhân viên. Điều này cho thấy các công ty ưu tiên tuyển dụng nhân viên có sự cam kết dài hạn và ổn định.
- ➔ Có thể cân nhắc mở rộng các hình thức làm việc khác để thu hút thêm nhân tài và đáp ứng các nhu cầu linh hoạt hơn của ứng viên.

### 2.2.5 Yêu cầu về bằng cấp

Biểu đồ Yêu cầu bằng cấp



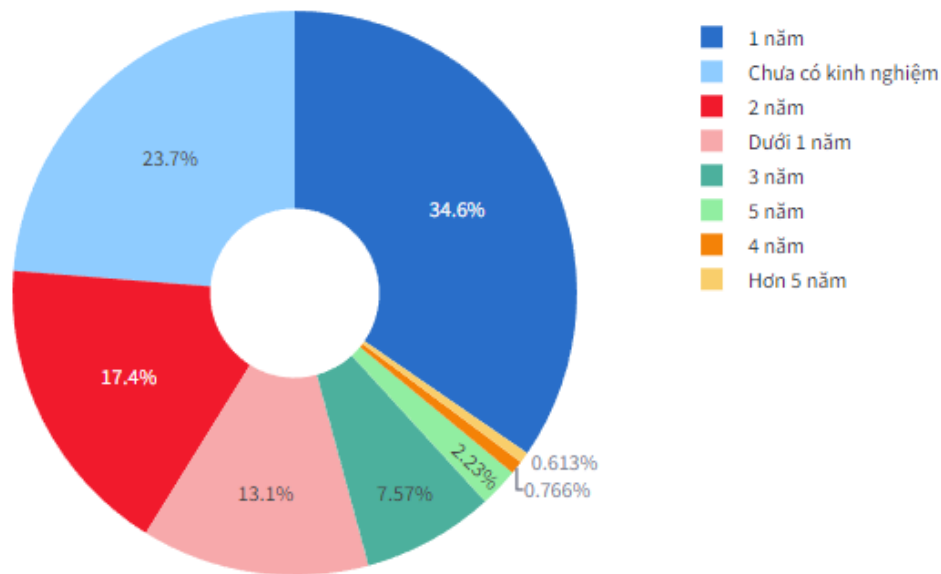
Hình 7. Yêu cầu bằng cấp

#### Nhận xét:

- "Không yêu cầu cụ thể" chiếm số lượng lớn nhất, với hơn 5000 công việc. Điều này cho thấy rằng nhiều công ty không đặt nặng yêu cầu về bằng cấp mà có thể chú trọng hơn vào kinh nghiệm và kỹ năng của ứng viên.
  - "Cao đẳng" là yêu cầu bằng cấp phổ biến thứ hai, với khoảng 3000 công việc.
  - "Trung học" và "Chứng chỉ" có số lượng yêu cầu ít nhất.
- ➔ Điều này thể hiện bằng cấp không phải là một yêu cầu quá quan trọng đối với các công ty nhưng bên cạnh đó vẫn nên có bằng cấp đại học hay cao đẳng thì cơ hội việc làm sẽ lớn hơn so với các ứng viên khác.

### 2.2.6 Yêu cầu kinh nghiệm

Biểu đồ Yêu cầu kinh nghiệm



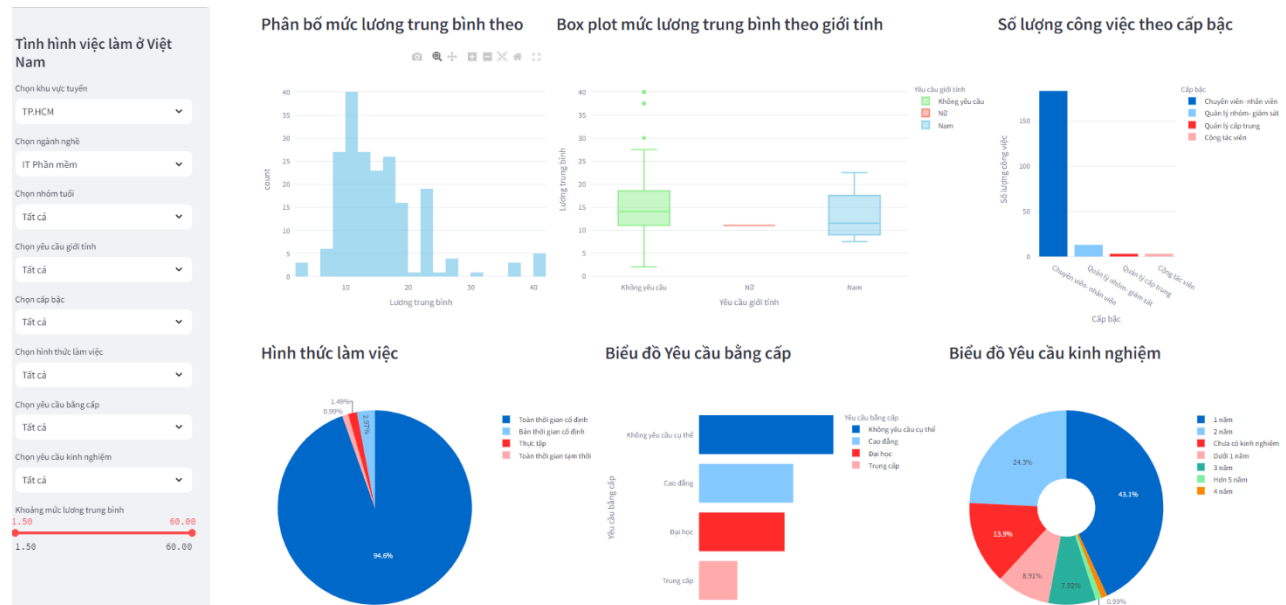
Hình 8. Yêu cầu kinh nghiệm

#### Nhận xét:

- 1 năm kinh nghiệm chiếm tỷ lệ cao nhất (34.6%), cho thấy rằng các công ty thường yêu cầu ứng viên có ít nhất một năm kinh nghiệm làm việc.
  - Chưa có kinh nghiệm chiếm tỷ lệ lớn thứ hai (23.7%), cho thấy các công ty cũng mở cửa cho các ứng viên mới ra trường hoặc chưa có kinh nghiệm làm việc.
  - Nhu cầu chủ yếu tập trung vào các ứng viên có 1 năm kinh nghiệm hoặc chưa có kinh nghiệm.
  - Đối với các vị trí yêu cầu kinh nghiệm từ 2 năm trở lên, số lượng công việc ít hơn, điều này có thể phản ánh rằng các công ty có thể đang tìm kiếm ứng viên có kỹ năng cụ thể hoặc cho các vị trí cao cấp hơn.
- ➔ **Ứng viên chưa có kinh nghiệm cũng có nhiều cơ hội tuyển dụng, đặc biệt là trong các công ty có chương trình đào tạo hoặc phát triển nhân viên mới. Ứng viên có kinh nghiệm từ 2 năm trở lên cần tìm kiếm các vị trí cụ thể yêu cầu kỹ năng và kinh nghiệm tương ứng.**



## 2.2.7 Loại dữ liệu phân tích

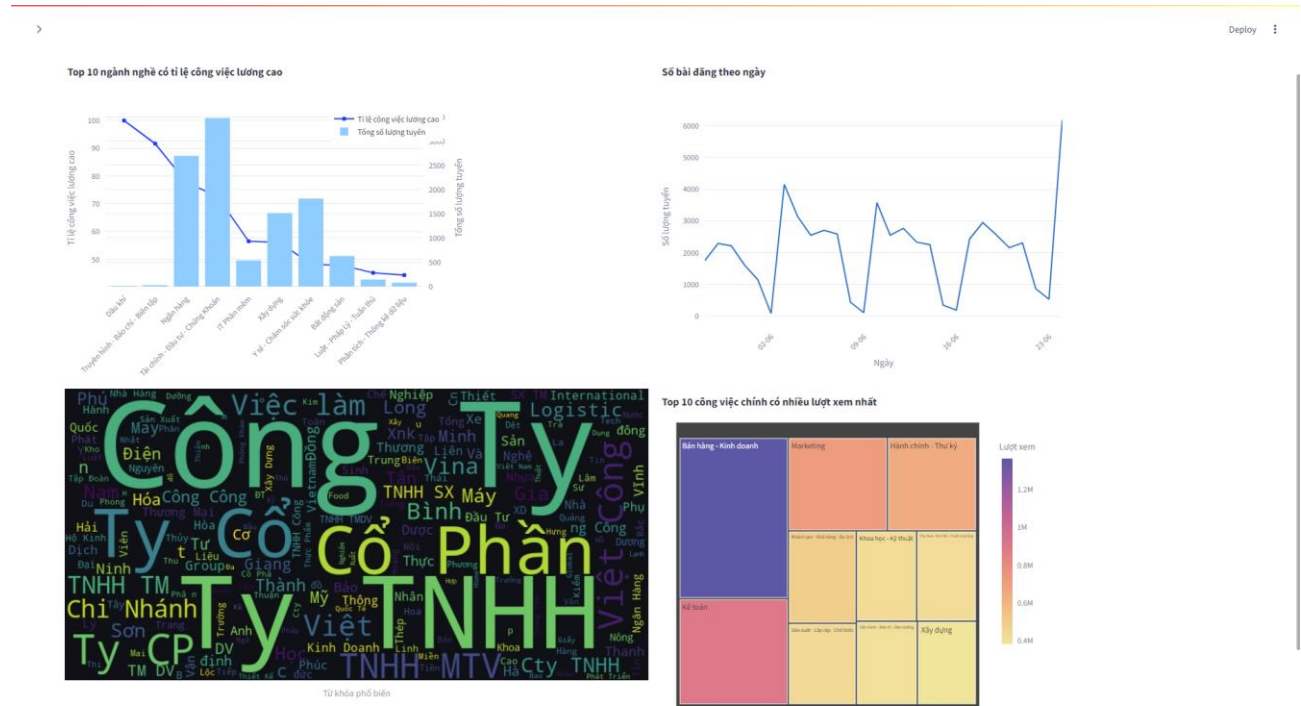


### Nhận xét:

- ➔ Ngoài ra Streamlit hỗ trợ sidebar có chức năng lọc, trích xuất dữ liệu theo yêu cầu từng trường. Như trong biểu đồ, có thể thấy mức lương của Ngành nghề IT phần mềm có mức lương trung bình lên đến 15 triệu so với các ngành nghề khác như những phân tích trên.
- ➔ Từ chức năng này của Streamlit chúng ta có thể tìm hiểu những ngành nghề hot nhất hiện nay và có mức thu nhập cao.

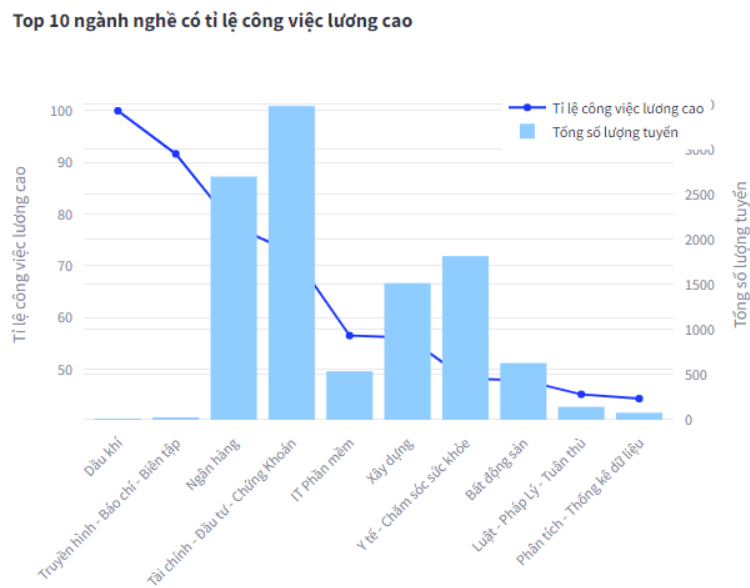
## 2.3 Biểu đồ trang 3

Trang 3 của Dashboard sẽ đi vào phân tích về các công việc.



Hình 9. Page 3

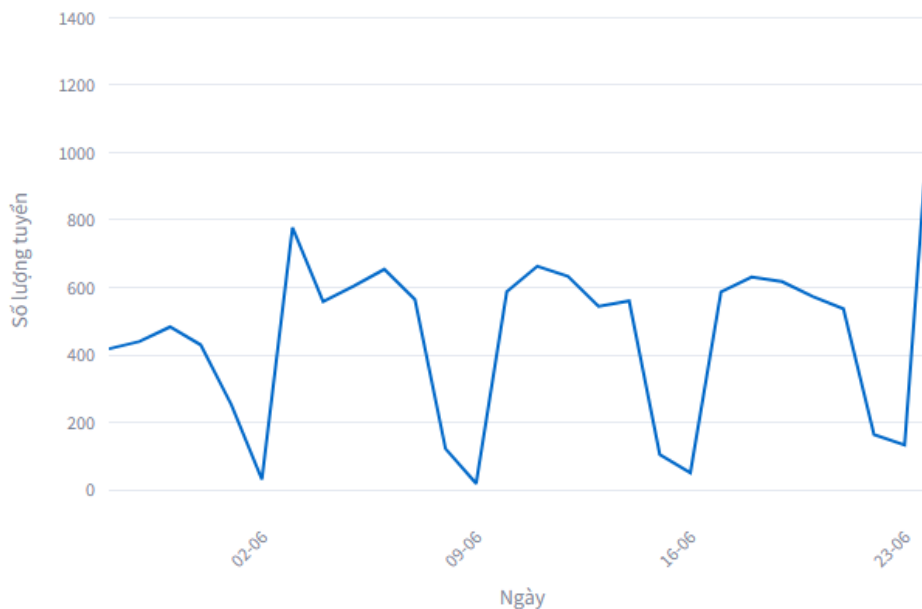
### 2.3.1 Biểu đồ Top 10 ngành nghề có tỉ lệ công việc lương cao



Hình 10. Top 10 công việc có tỉ lệ lương cao nhất

**Nhận xét:**

- Ngành Dầu khí với tỉ lệ tuyển lương cao nhất với 100% điều này thể hiện tính đặc thù của công việc với số lượng tuyển ít (2 vị trí) nhưng lương sẽ rất cao so với các ngành khác.
- Bên cạnh đó là các ngành Truyền hình – Báo chí – Biên tập hay Ngân hàng với tỉ lệ trên 50% cũng là những công việc rất hấp dẫn về mặt thu nhập
- Có thể thấy có các ngành tuyển lương cao nhưng số lượng tuyển rất ít và có một số ngành phổ biến hơn với mức lương rất ổn.

**2.3.2 Số bài đăng theo ngày****Số bài đăng theo ngày**

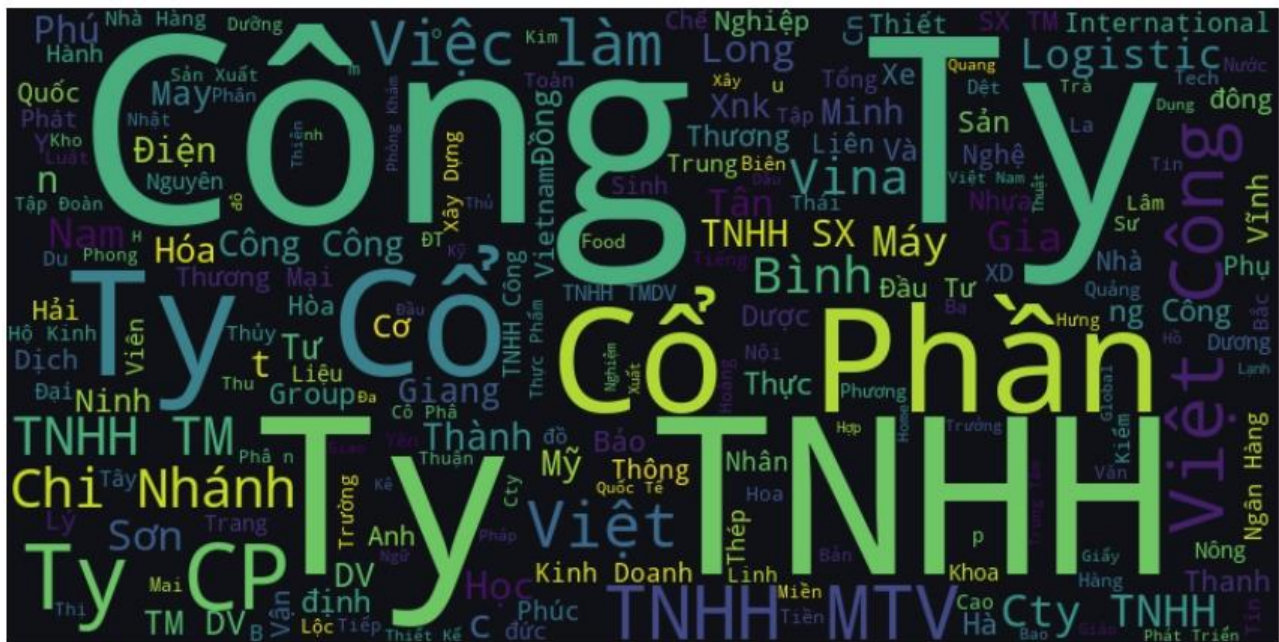
Hình 11. Số bài đăng theo ngày

**Nhận xét:**

- Có xu hướng tăng giảm luân phiên, với các đỉnh cao và đáy thấp rõ rệt.
- Cứ sau một vài ngày số lượng bài đăng giảm, sẽ có một ngày số lượng bài đăng tăng mạnh trở lại. Điều này có thể do các công ty thường đăng tuyển vào các ngày nhất định trong tuần.
- Số lượng bài đăng tăng giảm theo chu kì, giảm dần theo các ngày trong tuần. Nhiều nhất vào các ngày thứ 2 đầu tuần và thấp nhất vào ngày chủ nhật.
- Có những ngày cao nhất lên đến hơn gần 800 bài đăng tuyển dụng, và có những ngày thấp nhất chỉ đạt vào khoảng 19 bài.

➔ **Biểu đồ giúp người xem có thể đưa ra các chiến lược đăng bài vào các thời điểm phù hợp, để tăng khả năng tiếp cận với ứng viên. Bên cạnh đó cũng giúp các ứng viên tăng khả năng nhận được phản hồi.**

### 2.3.3 Những từ khoá phổ biến

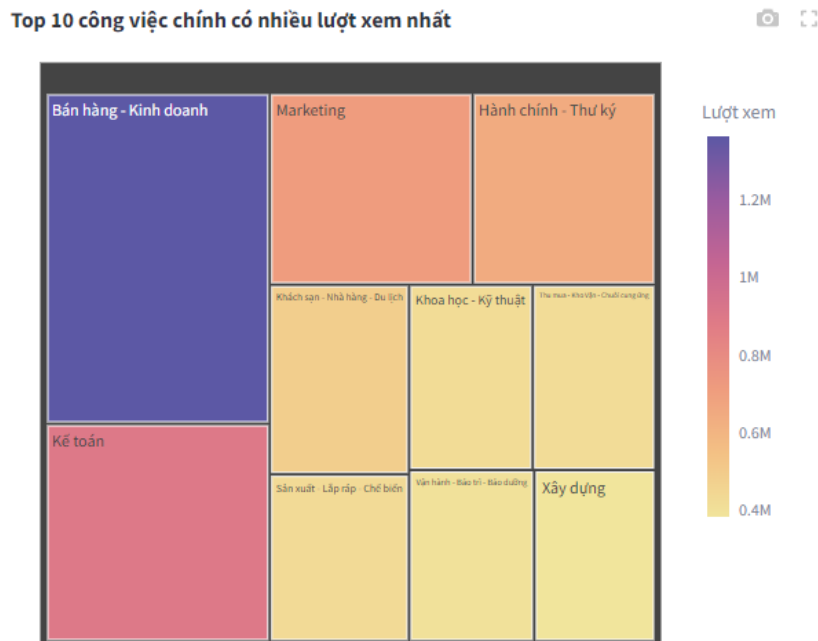


Từ khóa phổ biến

**Nhận xét:**

- "Công ty", "Cổ phần", "TNHH": Đây là các từ khóa phổ biến nhất, cho thấy rằng nhiều bài đăng tuyển dụng thuộc về các công ty có cấu trúc tổ chức này.
  - "Việt", "Việc làm", "Chi nhánh": Các từ khóa này cũng xuất hiện rất nhiều, phản ánh sự phân bố rộng rãi của các công ty và vị trí tuyển dụng tại Việt Nam.
  - "Logistic", "Hóa", "Điện": Các từ khóa này phản ánh các ngành nghề phổ biến hoặc các lĩnh vực mà các công ty đang tuyển dụng.
  - "Công nghệ", "Thiết kế", "Xây dựng": Các từ khóa này thể hiện các lĩnh vực chuyên môn khác nhau mà các công ty đang tìm kiếm ứng viên.
- ➔ Hiểu rõ các từ khóa phổ biến có thể giúp tối ưu hóa bài đăng tuyển dụng, sử dụng các từ khóa này để tăng khả năng tìm kiếm và thu hút ứng viên phù hợp. Tạo ra các từ khóa cụ thể và hấp dẫn có thể giúp bài đăng tuyển dụng nổi bật hơn.

### 2.3.4 Những công việc được quan tâm nhiều nhất



Hình 12. Những công việc được quan tâm nhiều nhất

#### Nhận xét:

- Bán hàng - Kinh doanh: Đây là công việc có lượt xem cao nhất, vượt trên 1.2 triệu lượt xem, cho thấy ngành này rất thu hút sự quan tâm của ứng viên.
  - Marketing: Cũng có số lượt xem rất cao, cho thấy sự phổ biến và nhu cầu lớn đối với các vị trí marketing.
  - Hành chính - Thư ký: Đứng thứ ba về lượt xem, phản ánh sự phổ biến của các công việc hành chính và thư ký.
- ➔ Số lượt xem của các ngành nghề cho thấy sự quan tâm và nhu cầu lớn đối với các vị trí bán hàng, kinh doanh, marketing và hành chính. Các ngành kỹ thuật, xây dựng và chuỗi cung ứng cũng có sự quan tâm đáng kể, cho thấy nhu cầu tuyển dụng và sự hấp dẫn của các ngành này.

## 2.4 Trang tìm kiếm

Trang dùng để tìm kiếm công việc theo nhu cầu

### Tìm kiếm công việc theo nhu cầu

Khu vực tuyển

Toàn bộ

Yêu cầu giới tính

Toàn bộ

Cấp bậc

Toàn bộ

Hình thức làm việc

Toàn bộ

Yêu cầu bằng cấp

Toàn bộ

Yêu cầu kinh nghiệm

Toàn bộ

Loại công ty

Toàn bộ

Nhóm tuổi

Toàn bộ

Ngành nghề

Toàn bộ

Customize Salary Range

Start of Range

1.50

End of Range

60.00

Lương trung bình

1.50

60.00

Các trường dữ liệu có thể dùng để lọc

Các cột giá trị có kiểu object như 'Khu vực tuyển', 'Yêu cầu giới tính', 'Cấp bậc', 'Hình thức làm việc', 'Yêu cầu bằng cấp', 'Yêu cầu kinh nghiệm', 'Loại công ty', 'Nhóm tuổi' thì sẽ dùng selection box để chọn giá trị theo yêu cầu, giá trị mặc định là toàn bộ, nghĩa là áp dụng với mọi giá trị của các cột kiểu object

Cột mức 'Lương trung bình' là kiểu numerical dùng slider để chọn khoảng lương muốn tìm, giá trị min và max của slider tương ứng với min max của cột 'Lương trung bình' trong bộ dữ liệu, có thể nhập liệu cho khoảng trên và khoảng dưới của lương nếu muốn

Giá trị trong Selection box 'Ngành nghề' sẽ lọc các cột Công việc chính, công việc liên quan 1, công việc liên quan 2 và xuất ra các công việc theo mong muốn

Chọn các trường để hiển thị:

- |   |   |   |   |
|---|---|---|---|
| <input checked="" type="checkbox"/> Ngày cập nhật       | <input checked="" type="checkbox"/> Lướt xem        | <input checked="" type="checkbox"/> Khu vực tuyển         | <input checked="" type="checkbox"/> Yêu cầu giới tính     |
| <input checked="" type="checkbox"/> Cấp bậc             | <input checked="" type="checkbox"/> Số lượng tuyển  | <input checked="" type="checkbox"/> Hình thức làm việc    | <input checked="" type="checkbox"/> Yêu cầu bằng cấp      |
| <input checked="" type="checkbox"/> Yêu cầu kinh nghiệm | <input checked="" type="checkbox"/> Từ khóa         | <input checked="" type="checkbox"/> Loại công ty          | <input checked="" type="checkbox"/> Lương trung bình      |
| <input checked="" type="checkbox"/> Nhóm tuổi           | <input checked="" type="checkbox"/> Công việc chính | <input checked="" type="checkbox"/> Công việc liên quan 1 | <input checked="" type="checkbox"/> Công việc liên quan 2 |

Có thể lựa chọn các cột dữ liệu muốn xem xét bằng việc tick vào các cột mà mình muốn

Ví dụ về tìm kiếm theo nhu cầu:

## Tìm kiếm công việc theo nhu cầu

Khu vực tuyển

Bình Dương

Yêu cầu giới tính

Nam

Cấp bậc

Chuyên viên- nhân viên

Hình thức làm việc

Toàn bộ

Yêu cầu bằng cấp

Toàn bộ

Yêu cầu kinh nghiệm

Toàn bộ

Loại công ty

Toàn bộ

Nhóm tuổi

Toàn bộ

Ngành nghề

Toàn bộ

### Customize Salary Range

Start of Range

22.00

-

+

End of Range

60.00

-

+

Lương trung bình

22.00

60.00

1.50

60.00

Số lượng công việc phù hợp với yêu cầu của bạn: 6

Có 6 công việc ở khu vực Bình Dương , có yêu cầu giới tính Nam, cấp bậc Chuyên viên – Nhân viên và lương từ 22-60 triệu

Chọn các trường để hiển thị:

- ☐ Ngày cập nhật
 ☐ Lượt xem
 ☒ Khu vực tuyển
 ☒ Yêu cầu giới tính
- ☒ Cấp bậc
 ☐ Số lượng tuyển
 ☐ Hình thức làm việc
 ☐ Yêu cầu bằng cấp
- ☐ Yêu cầu kinh nghiệm
 ☐ Từ khóa
 ☐ Loại công ty
 ☒ Lương trung bình
- ☐ Nhóm tuổi
 ☒ Công việc chính
 ☒ Công việc liên quan 1
 ☒ Công việc liên quan 2

Các công việc phù hợp với yêu cầu của bạn:

	Khu vực tuyển	Yêu cầu giới tính	Cấp bậc	Lương trung bình	Công việc chính	Công việc liên quan 1	Công việc liên quan 2
2,008	Bình Dương	Nam	Chuyên viên- nhân viên	22.5	Điện - Điện tử - Điện lạnh	Cơ khí - Ô tô - Tự động hóa	Sản xuất - Lắp ráp - Chế biến
5,258	Bình Dương	Nam	Chuyên viên- nhân viên	30	Xây dựng	Vận Tải - Lái xe - Giao nhận	Không yêu cầu thêm
6,357	Bình Dương	Nam	Chuyên viên- nhân viên	30	Bán hàng - Kinh doanh	Khoa học - Kỹ thuật	Khai thác năng lượng - Khoáng sản - Địa chất
8,470	Bình Dương	Nam	Chuyên viên- nhân viên	22.5	Bán hàng - Kinh doanh	Chăm sóc khách hàng	Quản lý dự án
10,752	Bình Dương	Nam	Chuyên viên- nhân viên	22.5	Điện - Điện tử - Điện lạnh	Cơ khí - Ô tô - Tự động hóa	Sản xuất - Lắp ráp - Chế biến
12,306	Bình Dương	Nam	Chuyên viên- nhân viên	22.5	Điện - Điện tử - Điện lạnh	Cơ khí - Ô tô - Tự động hóa	Sản xuất - Lắp ráp - Chế biến

Bảng kết quả sau khi đã lọc và chọn các cột dữ liệu muốn xem xét



## 3. MÔ HÌNH HÓA DỮ LIỆU VÀ TÍCH HỢP AI

### 3.1 MÔ HÌNH HÓA DỮ LIỆU

- Mục tiêu của mô hình hóa dữ liệu chúng em muốn hướng tới là dự đoán Mức lương trung bình mà nhà tuyển dụng và nhân viên có thể trao đổi và mức độ cạnh tranh (số người quan tâm công việc) của công việc đó. Đây đều là những dữ liệu số nên mô hình hồi quy có thể thực hiện tốt việc dự đoán này.

#### 3.1.1 Trích xuất đặt trung

- Loại bỏ những cột không cần thiết hoặc định dạng không thống nhất, cuối cùng ta có 11 cột đặt trung được chọn để cho mô hình phân tích: Khu vực tuyển, Yêu cầu giới tính, Cấp bậc, Hình thức làm việc, Yêu cầu Kinh nghiệm, Loại công ty, Nhóm tuổi, Công việc chính, Công việc liên quan 1, Công việc liên quan 2.
- Những cột dữ liệu đặt trung đều là những cột phân loại nên chúng em thực hiện phân tích ANOVA của thư viện ols tức là phân tích một chiều xem từng cột có ảnh hưởng đến cột mục tiêu hay không.

ANOVA for Recruitment\_Area:

	sum_sq	df	F	PR(>F)
Recruitment_Area	29128.769226	124.0	7.85389	2.488064e-127
Residual	493813.216287	16510.0	NaN	NaN

- Dựa vào giá trị “PR(>F)” để quyết định bác bỏ giả thuyết hay không. Nếu giá trị này  $< 0.05$  thì cột đang xét có ảnh hưởng đến cột mục tiêu của chúng ta.

Kết quả thu được thì 11/11 cột đều ảnh hưởng đến cột Mức lương.

#### 3.1.2 Chuẩn bị dữ liệu

- Vì các cột đặt trung đều là các cột phân loại nên chúng ta chuyển hết sang dữ liệu số bằng phương pháp Label Encoder.

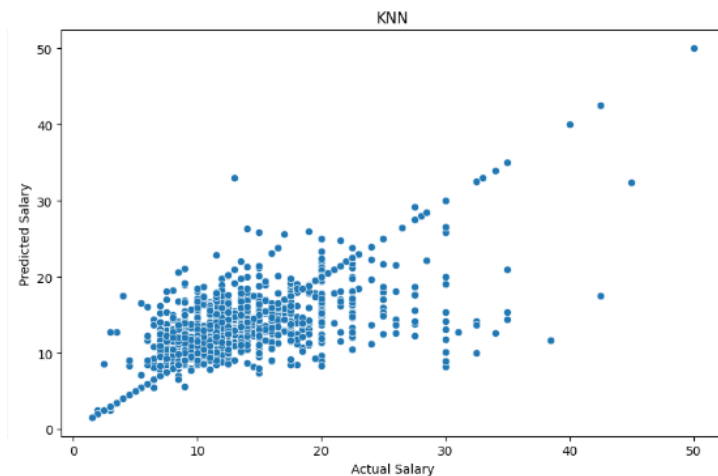
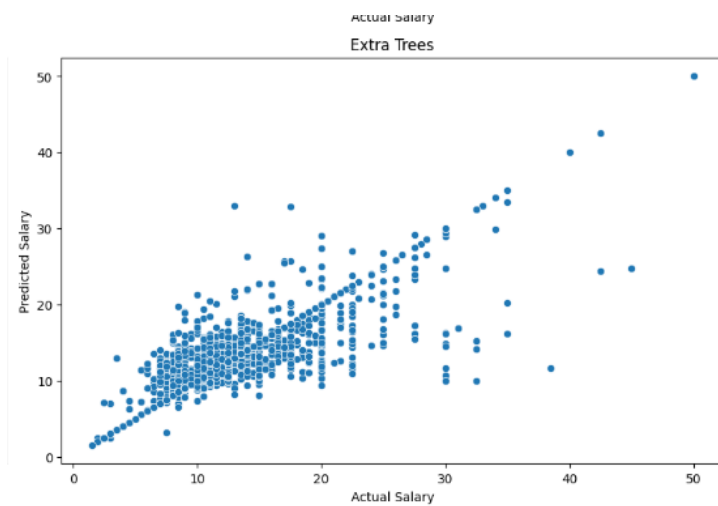


### 3.1.3 Mô hình hóa dữ liệu

- Sử dụng **Cross-Validation (CV)** dùng để đánh giá hiệu suất của mô hình. Mục đích của nó là kiểm tra độ chính xác và khả năng tổng quát hóa của mô hình khi áp dụng dữ liệu test.
- Nguyên tắc hoạt động K-fold Cross-Validation là
  - o Chia dữ liệu thành K phần (folds) bằng nhau.
  - o Quá trình huấn luyện và kiểm tra được lặp lại K lần. Mỗi lần, một fold được giữ lại làm tập kiểm tra, và K-1 folds còn lại được sử dụng để huấn luyện mô hình.
  - o Kết quả cuối cùng là trung bình của các kết quả kiểm tra từ K lần lặp lại.
- Lợi ích:
  - o Giúp đánh giá mô hình một cách đáng tin cậy bằng cách giảm thiểu hiện tượng overfitting (quá khớp) và underfitting (khớp kém).
  - o Đảm bảo rằng mô hình được kiểm tra trên nhiều phần khác nhau của dữ liệu, cung cấp một cái nhìn toàn diện về hiệu suất của mô hình.
- Áp dụng GridSearchCV để tìm ra mô hình với tham số tốt nhất. Đây là một kỹ thuật tìm kiếm tham số tối ưu cho mô hình học máy bằng cách thử nghiệm một cách hệ thống tất cả các kết hợp có thể của các tham số được xác định trước (grid of parameters). Nguyên tắc:
  - o GridSearchCV sẽ huấn luyện và đánh giá mô hình với từng tổ hợp của các tham số trong grid.
  - o Sử dụng Cross-Validation (thường là K-fold) để đánh giá mỗi tổ hợp tham số. Điều này đảm bảo rằng kết quả là đáng tin cậy và không phụ thuộc vào một phần dữ liệu cụ thể.
  - o GridSearchCV sẽ chọn bộ tham số có hiệu suất tốt nhất dựa trên tiêu chí đánh giá (scoring metric)

### 3.1.4 Mô hình dự đoán mức lương

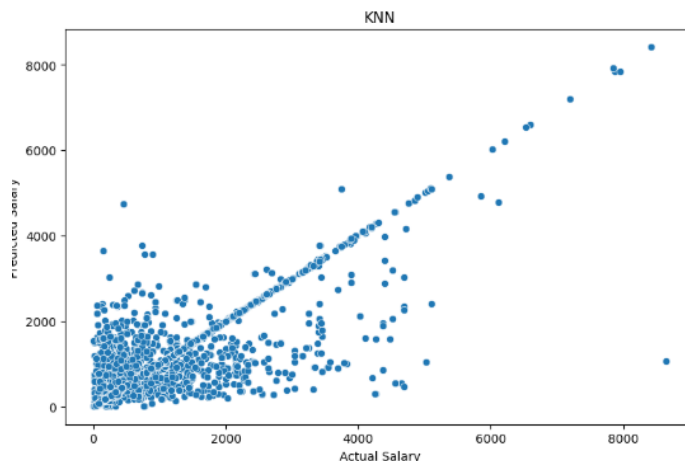
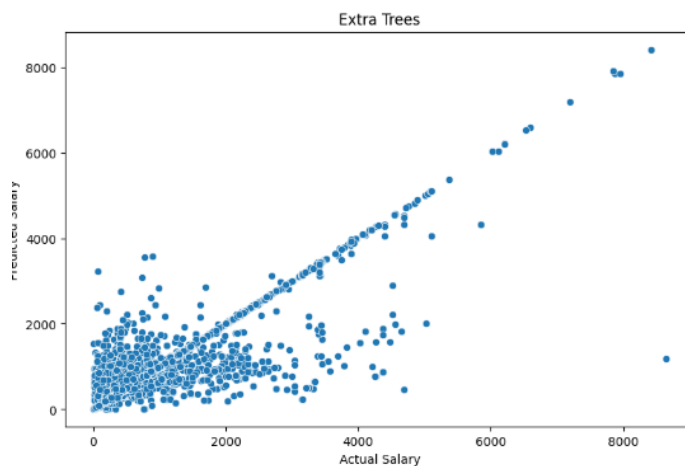
	Model	Best Parameters	MSE	MAE	R2
0	Random Forest	{'max_depth': 8, 'max_features': 1.0, 'n_estim...	15.256665	2.718641	0.513913
1	Gradient Boosting	{'learning_rate': 0.1, 'max_depth': 5, 'n_esti...	9.825167	2.082631	0.686964
2	XGBoost	{'learning_rate': 0.1, 'max_depth': 5, 'n_esti...	9.900396	2.087818	0.684567
3	Extra Trees	{'max_features': 'log2', 'n_estimators': 500}	6.563752	1.034874	0.790875
4	Decision Tree	{'max_depth': 5, 'min_samples_split': 2}	21.677160	3.309255	0.309352
5	KNN	{'n_neighbors': 7, 'p': 1, 'weights': 'distance'}	9.117886	1.266560	0.709498



- Đây là 2 mô hình tốt nhất độ chính xác 70% cho KNN và 79% cho mô hình Extra Trees với (MSE và MAE thấp nhất và R2 cao nhất). Điều này có thể cho thấy các mô hình có thể giải thích được dữ liệu.

### 3.1.5 Mô hình dự đoán mức độ quan tâm

	Model	Best Parameters	MSE	MAE	R2
0	Random Forest	{'max_depth': 8, 'max_features': 1.0, 'n_estim...	654351.880060	549.023464	0.357923
1	Gradient Boosting	{'learning_rate': 0.1, 'max_depth': 5, 'n_esti...	500240.419243	461.822381	0.509144
2	XGBoost	{'learning_rate': 0.1, 'max_depth': 5, 'n_esti...	511808.053124	471.237169	0.497793
3	Extra Trees	{'max_features': 'log2', 'n_estimators': 500}	301621.189475	252.960627	0.704037
4	Decision Tree	{'max_depth': 5, 'min_samples_split': 3}	851459.881307	627.664158	0.164513
5	KNN	{'n_neighbors': 7, 'p': 1, 'weights': 'distance'}	380730.908216	287.592460	0.626411



- Cũng giống như mô hình dự đoán lương thì mô hình dự đoán mức độ người quan tâm cũng cho kết quả tốt ở 2 mô hình Extra Trees và KNN với độ chính xác thấp hơn một chút là trên 60%.
- Thứ tự các mô hình theo mức độ giải thích dữ liệu là: Extra Trees > KNN > Gradient Boosting > XGBoost > Random Forest > Decision Tree

### 3.1.6 Kết quả mô hình hóa

## Mức Lương và Mức độ cạnh tranh của công việc mà bạn đang quan tâm

Hãy cho chúng tôi thông tin về công việc của bạn để dự đoán

Khu vực bạn muốn làm việc là?

An Giang

Yêu cầu giới tính của bạn là?

Không yêu cầu

Cấp bậc của bạn là?

Chuyên gia

Hình thức làm việc của bạn là?

Bán thời gian cố định

Yêu cầu bằng cấp của bạn là?

Cao đẳng

Yêu cầu kinh nghiệm của bạn là?

1 năm

Loại hình công ty bạn muốn làm việc?

Công ty cổ phần

Nhóm tuổi của bạn?

Không yêu cầu đặc biệt

Công việc chính của bạn là?

An ninh - Bảo vệ

Công việc có thể liên quan 1 của bạn là?

An ninh - Bảo vệ

Công việc có thể liên quan 2 của bạn là?

An ninh - Bảo vệ

Các công việc không được chọn trùng nhau.

## 3.2 TÍCH HỢP AI

### 3.2.1 Mô hình sử dụng

Ở đây nhóm sử dụng 1 mô hình miễn phí rất nổi tiếng là gemini-1.5-flash. Đây là mô hình cho phép gửi kèm file.

### 3.2.2 Hướng dẫn sử dụng

Giao diện chatbot với AI:



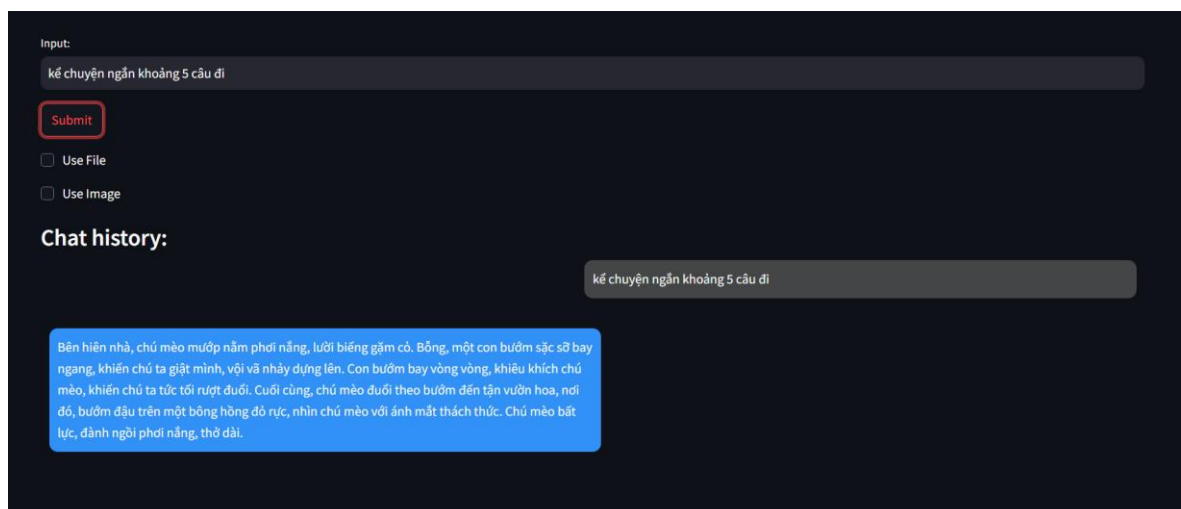
The screenshot shows a dark-themed chatbot interface. At the top, there is an 'Input:' label followed by a text input field. Below the input field is a 'Submit' button. Underneath the button are two checkboxes: 'Use File' and 'Use Image', both of which are currently unchecked. At the bottom of the interface, there is a section labeled 'Chat history:'.

Ở đây, chatbot cung cấp 3 chế độ chính

- General: Người dùng có thể hỏi AI bất cứ điều gì, tuy nhiên AI sẽ không có tri thức về bộ dữ liệu được sử dụng trong Dashboard
- File: AI có tri thức về bộ dữ liệu
- Image: AI có tri thức về Dashboard

#### 3.2.2.1 General

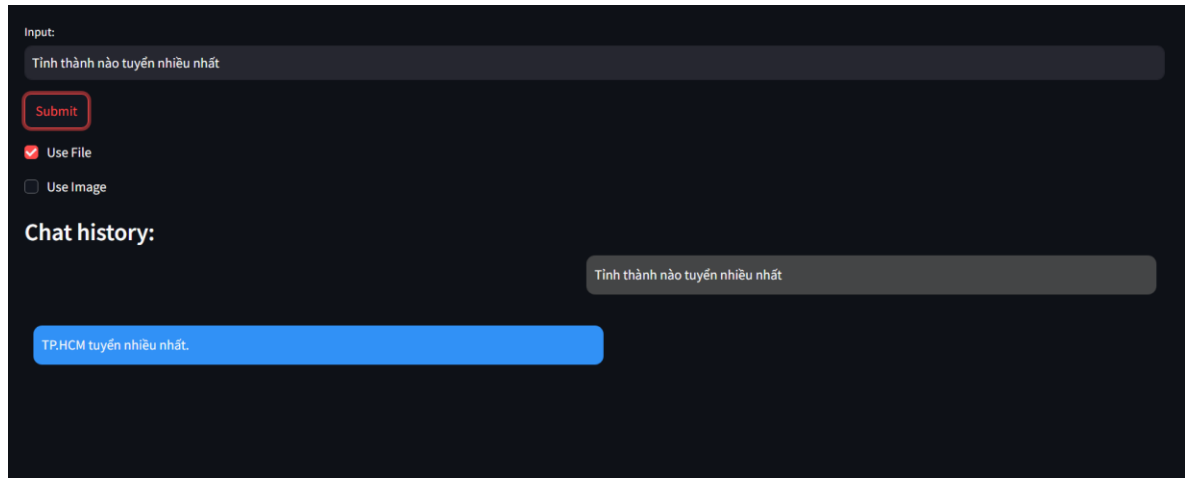
Hình dưới là một câu hỏi của người dùng, AI trả lời trong khoảng vài giây



The screenshot shows the same chatbot interface as before, but now with a question entered in the input field: 'kể chuyện ngắn khoảng 5 câu đi'. The 'Submit' button is highlighted with a red border. Below the input field, the 'Chat history:' section is visible, showing the AI's response in a blue box: 'Bên hiên nhà, chú mèo mập nằm phơi nắng, lưới biếng gặm cỏ. Bỗng, một con bướm sắc sỡ bay ngang, khiến chú ta giật mình, vội vã nhảy dựng lên. Con bướm bay vòng vòng, khiêu khích chú mèo, khiến chú ta tức tối núp đuôi. Cuối cùng, chú mèo đuổi theo bướm đến tận vườn hoa, nơi đó, bướm đậu trên một bông hồng đỏ rực, nhìn chú mèo với ánh mắt thách thức. Chú mèo bất lực, đành ngồi phơi nắng, thở dài.'

### 3.2.2.2 File

Chế độ này tốn rất nhiều thời gian để có reply nên cần cân nhắc trước khi sử dụng, nếu chỉ cần hỏi các thông tin cơ bản được thể hiện trong Dashboard thì khuyến khích sử dụng chế độ Image



Input:

Tỉnh thành nào tuyển nhiều nhất

Submit

☒ Use File

☐ Use Image

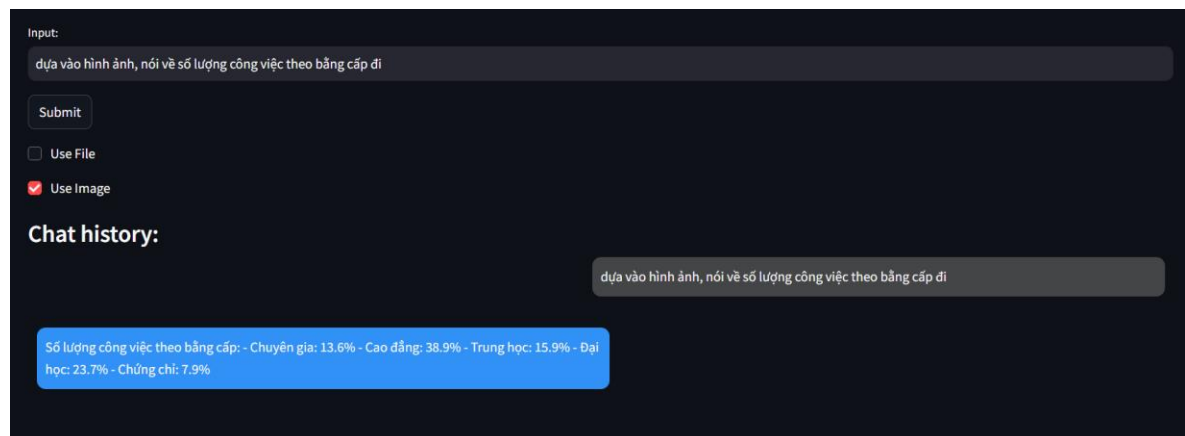
Chat history:

Tỉnh thành nào tuyển nhiều nhất

TP.HCM tuyển nhiều nhất.

### 3.2.2.3 Image: Chế độ khuyên dùng

Người dùng có thể hỏi về các thông tin liên quan tới dashboard



Input:

dựa vào hình ảnh, nói về số lượng công việc theo bảng cấp độ

Submit

☐ Use File

☒ Use Image

Chat history:

dựa vào hình ảnh, nói về số lượng công việc theo bảng cấp độ

Số lượng công việc theo bảng cấp: - Chuyên gia: 13.6% - Cao đẳng: 38.9% - Trung học: 15.9% - Đại học: 23.7% - Chứng chỉ: 7.9%

---

## 4. ĐÁNH GIÁ DASHBOARD QUA CÁC TIÊU CHÍ

---

### 1. Kết hợp nguồn dữ liệu đáng tin cậy:

- Nguồn dữ liệu được lấy từ trang web [vieclam24h.vn](http://vieclam24h.vn), trang web về tuyển dụng và tìm kiếm việc làm uy tín tại Việt Nam
- Dữ liệu được công khai trực tiếp để cho mọi người có thể tham khảo và khai thác, Nhóm dùng các kỹ thuật để thu thập dữ liệu từ trang web về một cách hợp lệ
- Dữ liệu sau khi thu thập được tiền xử lý qua để đảm bảo rằng dữ liệu sạch, không có missing value, các trường dữ liệu thống nhất, và loại bỏ các trường dữ liệu thừa

### 2. Các biểu đồ được sử dụng phù hợp với mục đích, đúng chức năng, như biểu đồ cột để so sánh số lượng công việc ở các tỉnh thành, biểu đồ histogram để xem phân phối lương, biểu đồ tròn để xem phân phối của nam và nữ trong tổng thể. Đồng thời biểu đồ phù hợp cũng phù hợp với các đối tượng người xem

### 3. Rõ ràng và dễ hiểu: Các biểu đồ giúp người xem nhanh chóng hiểu được thông điệp muốn truyền đạt

### 4. Sự tích hợp và liên kết: Các biểu đồ được liên kết về thông tin trong cùng 1 dashboard, người xem có thể dễ dàng tổng hợp được thông tin trực quan nhất của dashboard qua các biểu đồ này. Ví dụ ở MainDashBoard, chúng ta có ngay lập tức biết được tình hình việc làm ở Việt Nam, qua số lượng tuyển, worldmap dựa trên số lượng tuyển và phân phối về mức lương, độ tuổi, giới tính

### 5. Tương tác và điều hướng: Sự tương tác thể hiện qua sidebar, người xem có thể dùng sidebar để lọc dữ liệu theo mong muốn, các biểu đồ sẽ thay đổi theo bộ lọc đó

### 6. Thiết kế hấp dẫn: Thiết kế đơn giản, màu sắc được thể hiện nhẹ nhàng, đồng thời cũng tạo nên sự tương phản ở từng biểu đồ giúp người xem nhận thấy sự khác biệt, đồng thời biểu đồ cũng đa dạng, thu hút được người xem

### 7. Phân tích dữ liệu:

- Vì dữ liệu về công việc là rất lớn và thay đổi liên tục qua từng ngày, đồng thời khi cào dữ liệu sẽ tốn nhiều tài nguyên, đặc biệt là về thời gian, nên dữ liệu cho dự án cuối kỳ được thu thập tại 1-2 ngày trước ở thời điểm cào. Do đó không có các biểu đồ xu hướng theo thời gian.
- Các cột trong dữ liệu (hoặc gọi là biến) biểu thị cho các đặc trưng riêng trong dữ liệu, trong phân tích có áp dụng mô hình dùng các trường dữ liệu để dự đoán cho lượt xem (mức độ quan tâm) và mức lương.

### 8. Tích hợp AI: Dự án có tích hợp AI – mô hình Gemini 1.5 để thực hiện việc dựa vào AI để tổng hợp, trích xuất thông tin từ dữ liệu