

**BỘ CÔNG THƯƠNG**  
**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI**

-----□□□□□-----



**BÁO CÁO BÀI TẬP LỚN**  
**HỌC PHẦN: Phân Tích Dữ Liệu Lớn**

**Phân tích mô tả mức lương ngành công nghiệp  
phần mềm dựa trên dữ liệu về mức lương**

GVHD: T.S Nguyễn Mạnh Cường

Lớp: 20241IT7227001

Nhóm 16:

Vũ Văn Phúc – 2024700024

**Hà Nội 2024**

# MỤC LỤC

LỜI CẢM ƠN.....	5
LỜI NÓI ĐẦU .....	6
CHƯƠNG 1 : TỔNG QUAN VỀ ĐỀ TÀI .....	8
1.1. Giới thiệu về ngành công nghiệp phần mềm.....	8
1.2. Tình hình nhân lực trong ngành công nghiệp phần mềm.....	8
1.3. Mức lương trong ngành công nghiệp phần mềm.....	8
1.4. Vấn đề bất bình đẳng về mức lương .....	9
1.5. Nguồn dữ liệu và công cụ phân tích.....	9
1.6. Mục tiêu và phạm vi của nghiên cứu .....	10
CHƯƠNG 2: CÁC PHƯƠNG PHÁP PHÂN TÍCH MÔ TẢ.....	11
2.1. Phương pháp phân tích mô tả .....	11
2.2. Phương pháp xử lý và làm sạch dữ liệu .....	11
2.3. Các biện pháp thống kê mô tả .....	12
2.4. Biểu đồ trực quan hóa dữ liệu .....	13
2.5 Các công việc chính của phân tích mô tả .....	14
2.5.1 Tính các đại lượng thống kê, đo lường sự tập trung và phân tán của dữ liệu.....	14
2.5.2 Vẽ các loại biểu đồ.....	23
CHƯƠNG 3: THỰC NGHIỆM VÀ ĐÁNH GIÁ.....	30
3.1 Dữ liệu thực nghiệm.....	30
3.2 Công cụ sử dụng để thực hiện bài toán.....	31
3.3 Các thư viện sử dụng.....	31
3.4 Quy trình thực nghiệm .....	32
3.4.1 Làm sạch dữ liệu .....	33
3.4.2 Data Analysis and Visualization .....	36
KẾT LUẬN.....	43
TÀI LIỆU THAM KHẢO.....	45

## **Danh Mục Hình Ảnh**

Hình 2.1 Hàm Min() và Max() trong python .....	15
Hình 2.2 Hàm tính MAX của tập dữ liệu trong Excel .....	15
Hình 2.3 Hàm mean() trong python .....	16
Hình 2.4 Hàm AVERAGE trong Excel .....	17
Hình 2.5 Hàm median() trong python .....	17
Hình 2.6 Hàm MEDIAN() trong Excel .....	18
Hình 2.7 Bảng tần suất xuất hiện của các giá trị .....	18
Hình 2.8 Hàm mode() trong python .....	19
Hình 2.9 Hàm MODE() trong Excel .....	19
Hình 2.10 Dữ liệu đối xứng .....	20
Hình 2.11 Dữ liệu lệch trái .....	20
Hình 2.12 Dữ liệu lệch phải .....	20
Hình 2.13 Công thức tính Phương sai .....	21
Hình 2.14 Hàm var() trong python .....	21
Hình 2.15 Hàm VAR() trong Excel .....	22
Hình 2.16 Công thức tính Standard Deviation .....	22
Hình 2.17 Hàm std() trong python .....	23
Hình 2.18 Hàm STDEV() trong Excel .....	23
Hình 2.20 Bộ dữ liệu lương .....	24
Hình 2.21 Bảng thống kê tần suất xuất hiện của lương .....	25
Hình 2.22 Hướng dẫn tạo biểu đồ Histogram .....	25
Hình 2.23 Biểu đồ Histogram theo lương .....	26
Hình 2.24 Bảng dữ liệu vẽ biểu đồ Box & whisker .....	27

Hình 2.25 Các bước tạo biểu đồ box plot .....	27
Hình 2.26 Biểu đồ box plot của bảng dữ liệu .....	28
Hình 2.27 Bảng dữ liệu với biến độc lập x và biến phụ thuộc y .....	29
Hình 2.28 Chọn biểu đồ scatter plot .....	30
Hình 2.29 Biểu đồ scatter plot của bảng dữ liệu.....	30
Hình 3.1 Một số bản ghi trong bộ dữ liệu .....	31
Hình 3.2 Một số thư việc xử lý dữ liệu trong python .....	32
Hình 3.3 Quy trình thực hiện .....	33
Hình 3.4.1 Tải bộ dữ liệu trong python .....	34
Hình 3.4.2 Hình dạng bộ dữ liệu bằng python .....	34
Hình 3.4.3 Hiển thị columns trong bộ dữ liệu bằng python .....	35
Hình 3.4.4 Đoạn code kiểm tra giá trị null trong bộ dữ liệu .....	36
Hình 3.4.5 Đoạn code hiển thị location trong python .....	36
Hình 3.4.6 Hiển thị Mean, Median, Max và Min của lương .....	36
Hình 3.4.7 Biểu đồ barplot cho số lượng việc theo location .....	38
Hình 3.4.8 Biểu đồ barplot thể hiện mức lương trung bình .....	39
Hình 3.4.9 Biểu đồ barplot cho chức danh công việc.....	40
Hình 3.4.10 Biểu đồ Pie.....	59
Hình 3.4.11 Biểu đồ Histogram.....	62
Hình 3.4.12 Biểu đồ Scatter Plot .....	64

## LỜI CẢM ƠN

Để hoàn thành bài tiểu luận này, em đã nhận được sự giúp đỡ và hỗ trợ quý báu từ nhiều người. Trước tiên, em xin bày tỏ lòng biết ơn sâu sắc đến thầy cô trong khoa đã tận tình giảng dạy, truyền đạt kiến thức và định hướng cho em trong suốt quá trình học tập và nghiên cứu. Những bài giảng của thầy cô không chỉ mang đến cho em những kiến thức chuyên sâu mà còn mở rộng tầm nhìn và khơi dậy niềm đam mê trong lĩnh vực mà em đang theo đuổi.

Đặc biệt, em xin gửi lời cảm ơn chân thành tới Thầy, TS Nguyễn Mạnh Cường, người đã luôn sát cánh, hướng dẫn và hỗ trợ em trong suốt quá trình thực hiện tiểu luận này. Những góp ý chuyên môn quý báu, sự tận tình chỉ bảo và những lời động viên của thầy đã giúp em vượt qua nhiều khó khăn và hoàn thiện bài tiểu luận này một cách tốt nhất. em thật may mắn và trân trọng vì có được sự chỉ bảo từ thầy trong suốt quá trình học tập và nghiên cứu này.

Em cũng không quên cảm ơn những người bạn của mình – những người đã luôn đồng hành, chia sẻ và hỗ trợ em trong suốt quá trình học tập. Sự giúp đỡ và những lời khuyên từ các bạn đã giúp em rất nhiều trong việc hoàn thiện bài tiểu luận này.

Cuối cùng, em xin bày tỏ lòng biết ơn sâu sắc đến gia đình, những người đã luôn bên cạnh, tạo mọi điều kiện tốt nhất và động viên em vượt qua những thử thách trong suốt quá trình học tập và nghiên cứu.

Sinh viên thực hiện

Vũ Văn Phúc

## LỜI NÓI ĐẦU

Ngành công nghiệp phần mềm đang trở thành một trong những lĩnh vực phát triển nhanh chóng và có sức ảnh hưởng lớn đến nền kinh tế toàn cầu. Với sự bùng nổ của công nghệ thông tin, nhu cầu về nhân lực trong ngành phần mềm ngày càng tăng cao, dẫn đến sự cạnh tranh mạnh mẽ và sự biến động lớn trong mức lương của các chuyên gia công nghệ. Nghiên cứu về mức lương trong ngành công nghiệp phần mềm không chỉ mang lại cái nhìn tổng quan về thị trường lao động mà còn giúp các nhà quản lý nhân sự, các chuyên gia và sinh viên định hướng đúng đắn trong việc lựa chọn nghề nghiệp và nâng cao kỹ năng.

Bài tiểu luận này được thực hiện với mục tiêu phân tích mô tả mức lương trong ngành công nghiệp phần mềm, dựa trên dữ liệu thực tế. Thông qua việc phân tích các yếu tố ảnh hưởng như kinh nghiệm làm việc, vị trí địa lý, trình độ học vấn và các yếu tố khác, bài tiểu luận sẽ cung cấp cái nhìn rõ ràng về sự khác biệt trong mức lương của các chuyên gia phần mềm, từ đó giúp hiểu rõ hơn về tình hình nhân lực trong ngành này.

Quá trình thực hiện bài tiểu luận đã mang đến cho tôi nhiều kiến thức bổ ích và kinh nghiệm quý giá, giúp tôi có cơ hội áp dụng những lý thuyết đã học vào thực tiễn phân tích. Tôi hy vọng rằng bài tiểu luận này không chỉ đáp ứng được yêu cầu của đề tài mà còn góp phần làm phong phú thêm hiểu biết về một lĩnh vực đang phát triển mạnh mẽ..

Bài tiểu luận này sẽ tập trung vào việc phân tích mô tả mức lương ngành công nghiệp phần mềm dựa trên dữ liệu thực tế. Nội dung của bài tiểu luận được chia thành ba chương:

### **Chương 1:** Tổng quan về đề tài

Chương này giới thiệu về ngành công nghiệp phần mềm, các xu hướng phát

triển của ngành và lý do tại sao việc nghiên cứu mức lương lại quan trọng. Ngoài ra, chương này sẽ trình bày các yếu tố ảnh hưởng chính đến mức lương như kinh nghiệm, vị trí địa lý, trình độ học vấn, và kỹ năng chuyên môn.

**Chương 2:** Các phương pháp trong phân tích mô tả:

Chương này sẽ nói về cách sử dụng các công cụ để phân tích mô tả cũng như các công việc cần phải làm.

**Chương 3:** Thực nghiệm và đánh giá:

Chương này là trình bày cụ thể quá trình thực hiện bài toán và đưa ra kết quả phân tích cũng như những đánh giá về chúng

# CHƯƠNG 1 : TỔNG QUAN VỀ ĐỀ TÀI

## 1.1. Giới thiệu về ngành công nghiệp phần mềm

- Khái quát về ngành công nghiệp phần mềm: Ngành công nghiệp phần mềm đã phát triển mạnh mẽ trong vài thập kỷ qua và trở thành một trong những ngành có tốc độ tăng trưởng nhanh nhất trên toàn cầu. Các công ty công nghệ, từ các tập đoàn lớn đến các công ty khởi nghiệp, đều đóng góp vào việc phát triển công nghệ mới, phục vụ cho mọi khía cạnh của đời sống.
- Tốc độ phát triển và xu hướng hiện nay: Xu hướng như trí tuệ nhân tạo, điện toán đám mây, blockchain và phát triển phần mềm di động đã và đang thay đổi cục diện ngành công nghiệp phần mềm, từ đó ảnh hưởng đến nhu cầu lao động và mức lương của các chuyên gia phần mềm.

## 1.2. Tình hình nhân lực trong ngành công nghiệp phần mềm

- Cầu và cung lao động: Với nhu cầu ngày càng tăng về nhân lực trong ngành công nghiệp phần mềm, thị trường lao động trở nên cạnh tranh hơn bao giờ hết. Việc thiếu hụt nguồn nhân lực chất lượng cao đã đẩy mức lương của các chuyên gia phần mềm lên mức đáng kể, đặc biệt ở các quốc gia phát triển.
- Vai trò của các vị trí công việc trong ngành: Từ các vị trí phát triển phần mềm, quản lý dự án, kỹ sư DevOps, đến chuyên gia dữ liệu, mỗi vị trí đều có những yêu cầu khác nhau về kỹ năng và kinh nghiệm, từ đó ảnh hưởng trực tiếp đến mức lương.

## 1.3. Mức lương trong ngành công nghiệp phần mềm

- Khái niệm về mức lương trong ngành phần mềm: Mức lương trong ngành phần mềm phụ thuộc vào nhiều yếu tố như vị trí công việc, kinh nghiệm, kỹ năng chuyên môn, và địa lý. Định nghĩa mức lương trung bình, mức lương khởi điểm, và các yếu tố khác liên quan đến thu nhập sẽ được trình bày trong phần này.



- Các yếu tố ảnh hưởng đến mức lương: Các yếu tố chính ảnh hưởng đến mức lương sẽ được thảo luận, bao gồm kinh nghiệm làm việc, vị trí địa lý, trình độ học vấn, kỹ năng cụ thể (như lập trình ngôn ngữ cụ thể hoặc quản lý dự án), và sự phát triển của các công nghệ mới.

#### 1.4. Vấn đề bất bình đẳng về mức lương

- Sự khác biệt về mức lương theo giới tính: Có sự khác biệt đáng kể về mức lương giữa nam và nữ trong ngành công nghệ, một hiện tượng được ghi nhận qua nhiều nghiên cứu. Sự bất bình đẳng này không chỉ phản ánh ở mức lương mà còn ở cơ hội thăng tiến trong nghề nghiệp.
- Sự chênh lệch về mức lương theo khu vực: So sánh mức lương giữa các quốc gia hoặc khu vực khác nhau, bao gồm các thị trường công nghệ lớn như Mỹ, châu Âu và châu Á.
- Các yếu tố văn hóa và xã hội khác: Văn hóa doanh nghiệp và xã hội có thể ảnh hưởng đến sự phân bổ lương không công bằng trong ngành công nghiệp phần mềm.

#### 1.5. Nguồn dữ liệu và công cụ phân tích

- Dataset sử dụng: Dataset được sử dụng trong tiểu luận này lấy từ Kaggle, cụ thể là từ dự án “Software Professional Salaries”. Dataset này chứa thông tin về mức lương, kinh nghiệm, vị trí công việc và các yếu tố khác liên quan đến nhân sự trong ngành công nghiệp phần mềm.
- Công cụ phân tích: Để phân tích dữ liệu, hai công cụ chính được sử dụng là Microsoft Excel và Python.
  - Excel sẽ được sử dụng để xử lý dữ liệu cơ bản, như làm sạch dữ liệu, tính toán các số liệu thống kê cơ bản (mean, median, variance) và tạo biểu đồ.

- Python (với các thư viện như pandas, matplotlib, seaborn và scikit-learn) sẽ được sử dụng cho các phân tích nâng cao như hồi quy, phân tích phân phối và trực quan hóa dữ liệu.
- Lý do sử dụng Excel và Python: Sự kết hợp giữa Excel và Python không chỉ mang lại sự tiện lợi trong việc xử lý dữ liệu một cách trực quan mà còn cung cấp khả năng phân tích mạnh mẽ với các mô hình thống kê và trực quan hóa linh hoạt.

### **1.6. Mục tiêu và phạm vi của nghiên cứu**

- Mục tiêu nghiên cứu: Mục tiêu của nghiên cứu này là phân tích mô tả mức lương của các chuyên gia phần mềm trên toàn cầu, xác định các yếu tố chính ảnh hưởng đến mức lương, và tìm hiểu sự khác biệt giữa các vùng địa lý và nhóm nhân khẩu học khác nhau.
- Phạm vi nghiên cứu: Nghiên cứu sẽ tập trung vào các thông tin từ dataset Kaggle, giới hạn trong một khoảng thời gian nhất định. Phạm vi nghiên cứu sẽ bao gồm các yếu tố như kinh nghiệm, vị trí công việc, và khu vực địa lý.

## **CHƯƠNG 2: CÁC PHƯƠNG PHÁP PHÂN TÍCH MÔ TẢ**

### **2.1. Phương pháp phân tích mô tả**

- Định nghĩa phân tích mô tả: Phân tích mô tả là quá trình sử dụng các biện pháp thống kê cơ bản để tóm tắt và mô tả các đặc điểm chính của một tập dữ liệu. Các chỉ số phổ biến như trung bình, trung vị, độ lệch chuẩn, phân phối tần suất, và các đồ thị trực quan sẽ được sử dụng để hiểu rõ hơn về các xu hướng và mẫu trong dữ liệu.
- Mục tiêu của phân tích mô tả: Trong tiểu luận này, mục tiêu của phân tích mô tả là khám phá và hiểu rõ các yếu tố ảnh hưởng đến mức lương trong ngành công nghiệp phần mềm, bao gồm kinh nghiệm làm việc, vị trí địa lý, trình độ học vấn, và kỹ năng chuyên môn.

### **2.2. Phương pháp xử lý và làm sạch dữ liệu**

- Thu thập dữ liệu từ Kaggle: Dataset được thu thập từ Kaggle chứa thông tin về mức lương của các chuyên gia phần mềm cùng với các biến số liên quan như vị trí công việc, kinh nghiệm, và khu vực địa lý.
- Làm sạch dữ liệu: Trước khi tiến hành phân tích, dữ liệu cần được làm sạch để đảm bảo tính chính xác. Quá trình làm sạch dữ liệu bao gồm:
  - Xử lý giá trị thiếu (missing values): Loại bỏ hoặc điền giá trị cho các mục bị thiếu trong dữ liệu.
  - Loại bỏ dữ liệu ngoại lai (outliers): Xác định và loại bỏ các giá trị ngoại lai có thể làm sai lệch kết quả phân tích.
  - Chuẩn hóa dữ liệu: Điều chỉnh dữ liệu để đảm bảo rằng tất cả các giá trị đều ở cùng một đơn vị đo hoặc định dạng nhất quán, chẳng hạn như chuẩn hóa các số liệu về lương sang cùng một loại tiền tệ nếu cần.

Công cụ thực hiện:

- Excel: Sử dụng Excel để thực hiện bước đầu trong làm sạch dữ liệu, kiểm tra tính hợp lệ và chuẩn hóa định dạng.
- Python: Sử dụng Python (với thư viện pandas) để xử lý dữ liệu thiếu, lọc ngoại lệ và chuẩn hóa dữ liệu một cách tự động và chính xác hơn.

### 2.3. Các biện pháp thống kê mô tả

Trung bình (Mean): Trung bình là thước đo phổ biến nhất được sử dụng để xác định mức lương trung bình của các chuyên gia phần mềm trong dataset. Trung bình giúp xác định xu hướng tổng thể của mức lương.

- Trung vị (Median): Trung vị là giá trị ở giữa của một tập hợp dữ liệu đã được sắp xếp. Nó được sử dụng để xác định mức lương giữa trong khi loại bỏ ảnh hưởng của các giá trị ngoại lai.
- Độ lệch chuẩn (Standard Deviation): Độ lệch chuẩn đo lường sự phân tán của dữ liệu xung quanh giá trị trung bình. Nó giúp hiểu rõ sự biến động trong mức lương của các chuyên gia phần mềm.
- Phân phối tần suất (Frequency Distribution): Phân phối tần suất được sử dụng để biểu diễn mức độ phổ biến của các mức lương khác nhau trong dataset. Số liệu này giúp xác định những khoảng lương phổ biến nhất.

Công cụ thực hiện:

- Excel: Excel có thể được sử dụng để tính toán các giá trị trung bình, trung vị và độ lệch chuẩn một cách dễ dàng. Các bảng phân phối tần suất cũng có thể được tạo ra bằng cách sử dụng các hàm và công cụ như PivotTable.
- Python: Python, với các thư viện pandas và numpy, có thể thực hiện tính toán các chỉ số thống kê này một cách tự động và tối ưu hóa hơn cho các tập dữ liệu lớn.

## 2.4. Biểu đồ trực quan hóa dữ liệu

- Biểu đồ hộp (Box Plot): Biểu đồ hộp được sử dụng để trực quan hóa phân phối của mức lương, giúp hiển thị các giá trị trung vị, giá trị tối đa, tối thiểu và các ngoại lệ. Điều này đặc biệt hữu ích trong việc nhận diện sự phân bố và bất bình đẳng trong mức lương.
- Biểu đồ histogram: Histogram sẽ được sử dụng để trực quan hóa phân bố tần suất của các mức lương trong dataset. Nó cho thấy mức độ tập trung của dữ liệu ở các khoảng lương khác nhau.
- Biểu đồ tán xạ (Scatter Plot): Biểu đồ tán xạ sẽ giúp hiển thị mối quan hệ giữa các biến số, chẳng hạn như mức lương và kinh nghiệm làm việc, hoặc mức lương và vị trí địa lý.

Công cụ thực hiện:

- Excel: Excel có khả năng tạo các biểu đồ như biểu đồ hộp và histogram. Tuy nhiên, khả năng trực quan hóa của Excel có thể hạn chế đối với các biểu đồ phức tạp hoặc cần sự tùy biến cao.
- Python: Python, với các thư viện như matplotlib và seaborn, là công cụ mạnh mẽ để trực quan hóa dữ liệu với khả năng tùy chỉnh cao. Python có thể tạo ra các biểu đồ tán xạ, biểu đồ hộp, và histogram với độ chính xác và thẩm mỹ cao hơn.

## 2.5 Các công việc chính của phân tích mô tả

### 2.5.1 Tính các đại lượng thống kê, đo lường sự tập trung và phân tán của dữ liệu

**Min và Max:** Là giá trị nhỏ nhất và lớn nhất trong tập dữ liệu. Cách tính:

Trong Python: Dùng hàm *max()* và *min()*

Ví dụ:

```
salary = [1000, 2000, 1500, 3000, 4000, 4500, 3500];
print(min(salary)) # 1000
print(max(salary)) # 4500
```

Hình 2.1 Hàm Min() và Max() trong python

Trong Excel: Dùng hàm *MIN()* và *MAX()*

Ví dụ:

=MAX(D2:D5391)								
D	E	F	G	H	I	J	K	L
Salary	Salaries R	Location						
400000	3	Bangalore						
400000	3	Bangalore						
1000000	3	Bangalore						
300000	3	Bangalore						
600000	3	Bangalore						
100000	3	Bangalore					Lương cao nhất:	10000000
192000	3	Bangalore						
400000	3	Bangalore						

Hình 2.2 Hàm tính MAX của tập dữ liệu trong Excel

Nhập công thức *=MAX(D2:D5391)* vào ô hiển thị kết quả.

**Mean:** Giá trị trung bình Mean cho biết trung bình giá trị của biến nằm ở mức độ nào so với ngưỡng giá trị nhỏ nhất, lớn nhất, được tính bằng tổng các giá trị trong một phân phối chia cho tổng số giá trị

Cách tính:

Đối với 10 số sau: 2 10 8 2 13 12 0 6 11 4 => trung bình cộng số học là  $(2 + 10 + 8 + 2 + 13 + 12 + 0 + 6 + 11 + 4)/10 = 6.8$

**Trimmed mean:** Trung bình cộng đã cắt bớt có thể được sử dụng với các mẫu lớn và tương tự như giá trị trung bình cộng số học (arithmetic mean) nhưng có một số giá trị nhỏ nhất và lớn nhất bị loại bỏ trước khi tính toán. Thông thường, 5% giá trị dưới cùng và trên cùng bị loại bỏ và giá trị trung bình được tính trên 90% giá trị còn lại. Hiệu ứng là giảm thiểu ảnh hưởng của quan sát ngoại lệ cực trị trong tính toán giá trị trung bình

Ví dụ:

3, 13, 15, 16, 19, 20, 21, 25, 40

⇒ Trimmed mean

Trong Python: Dùng hàm *mean()*

Ví dụ:

```
import numpy as np
import pandas as pd
numlist = [3, 6, 9, -1, 0, -10, 18]

# tính mean trên 1 list
print(np.mean(numlist)) #>> 3.5714285714285716

# tính mean trên từng cột dữ liệu trong 1 data frame
mean_col = data_frame.mean()
```

*Hình 2.3 Hàm mean() trong python*

Trong Excel: Nhập hàm *AVERAGE*(*number1*; [*number2*]; ...)

Ví dụ:

```
=AVERAGE(D2:D5391)
```

D	E	F	G	H	I	J	K	L
Salary	Salaries R	Location						
400000	3	Bangalore						
400000	3	Bangalore						
1000000	3	Bangalore						
300000	3	Bangalore						
600000	3	Bangalore						
100000	3	Bangalore				Average("mean")		605163.9
192000	3	Bangalore						
400000	3	Bangalore						

*Hình 2.4 Hàm AVERAGE trong Excel*

**Median:** Là giá trị trung vị (vị trí trung tâm, đối với dãy đã được sắp)

3, 13, 15, 16, 19, 20, 21, 25, 40

13, 15, 16, 19, 20, 21, 25, 40

Cách tính:

Trong Python: Dùng hàm *median()*

Ví dụ:

```
import numpy as np
import pandas as pd
numlist = [3, 6, 9, -1, 0, -10, 18]

# tính median trên 1 list
print(np.median(numlist)) #>> 3.0

# tính median trên từng cột dữ liệu trong 1 data frame
median_col = data_frame.median()
```

*Hình 2.5 Hàm median() trong python*



Trong Excel: Nhập hàm *MEDIAN()*

Ví dụ:

Nhập hàm *=MEDIAN(D2:D5391)* nhấn 'Enter'

	B	C	D	E	F	G	H	I	J	K	L
g	Company	Job Title	Salary	Salaries R	Location						
3.8	Sasken	Android D	400000	3	Bangalore						
4.5	Advanced	Android D	400000	3	Bangalore						
4	Unacadem	Android D	1000000	3	Bangalore						
3.8	SnapBizz	(Android D	300000	3	Bangalore						
4.4	Appoids T	Android D	600000	3	Bangalore						
4.2	Freelance	Android D	100000	3	Bangalore				Median=		400000
3.7	SQUARE N	Android D	192000	3	Bangalore						
3.1	Samsung	Android D	400000	3	Bangalore						
3.7	DXMinds	Android D	300000	3	Bangalore						
3.6	Endeavou	Android D	600000	3	Bangalore						

Hình 2.6 Hàm *MEDIAN()* trong Excel

**Mode:** Là giá trị xuất hiện thường xuyên nhất trong một phân phối

3, 1, 3, 6, 1, 8, 1, 6, 4		3, 1, 3, 6, 1, 8, 1, 3, 4		3, 1, 3, 6, 1, 6, 4, 4	
Giá trị	Tần suất	Giá trị	Tần suất	Giá trị	Tần suất
1	3	1	3	1	2
3	2	3	3	3	2
4	1	4	1	4	2
6	2	6	1	6	2
8	1	8	1		
Unimodal		Multimodal			

Hình 2.7 Bảng tần suất xuất hiện của các giá trị

Trong một phân phối nếu có 1 giá trị xuất hiện thường xuyên nhất thì gọi là *Unimodal* nếu có nhiều hơn 1 giá trị xuất hiện thường xuyên nhất thì gọi là *Multimodal*

Cách tính:

Trong Python: dùng hàm *mode()*

Ví dụ:

```
import numpy as np
import pandas as pd
from scipy.stats import mode
numlist = [3, 6, 9, -1, 0, -10, 18, 9]

# tính mode trên 1 list
print(mode(numlist)) #>> ModeResult(mode=9, count=2)

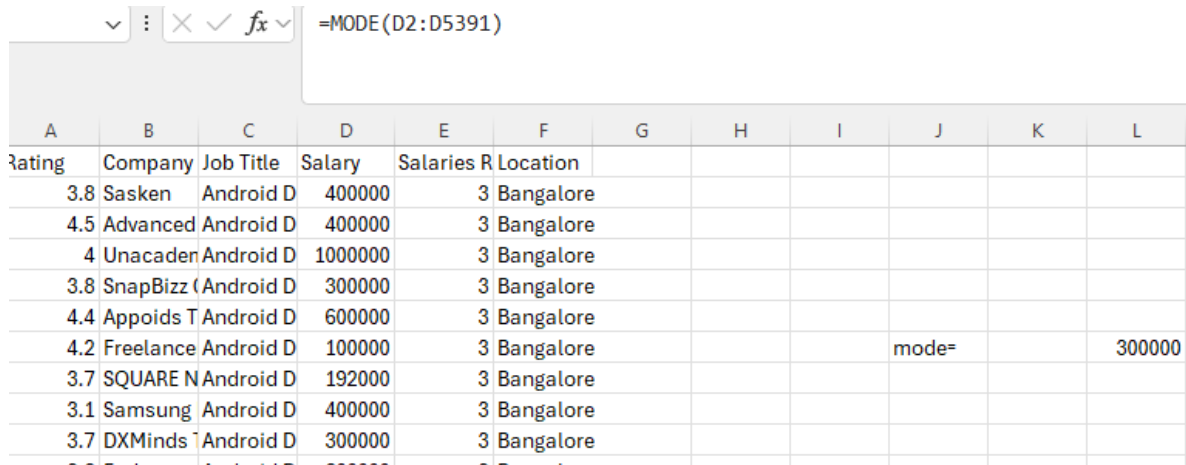
# tính mode trên từng cột dữ liệu trong 1 data frame
mode_col = data_frame.mode()
```

Hình 2.8 Hàm mode() trong python

Trong Excel: Nhập hàm *MODE()*

Ví dụ:

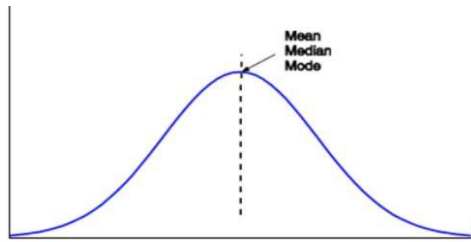
Nhập *=MODE(D2:D5391)* nhấn 'Enter'



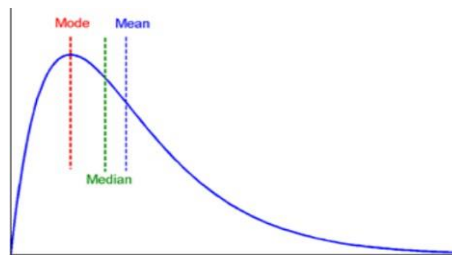
A	B	C	D	E	F	G	H	I	J	K	L
Rating	Company	Job Title	Salary	Salaries R	Location						
3.8	Sasken	Android D	400000	3	Bangalore						
4.5	Advanced	Android D	400000	3	Bangalore						
4	Unacadem	Android D	1000000	3	Bangalore						
3.8	SnapBizz	(Android D	300000	3	Bangalore						
4.4	Appoids T	Android D	600000	3	Bangalore						
4.2	Freelance	Android D	100000	3	Bangalore				mode=		300000
3.7	SQUARE N	Android D	192000	3	Bangalore						
3.1	Samsung	Android D	400000	3	Bangalore						
3.7	DXMinds	Android D	300000	3	Bangalore						

Hình 2.9 Hàm MODE() trong Excel

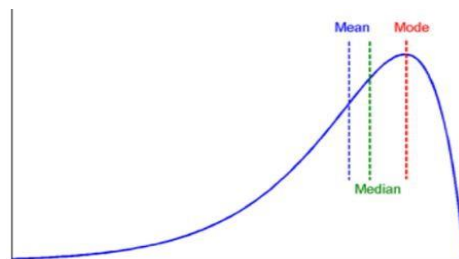
Một số đặc điểm về mức độ tập trung của dữ liệu:



Hình 2.10 Dữ liệu đối xứng



Hình 2.11 Dữ liệu lệch trái



Hình 2.12 Dữ liệu lệch phải

Dữ liệu đối xứng: Khi bộ 3 *Mean*, *Median*, *Mode* cùng mang 1 giá trị thì khi đó dữ liệu được phân bố dạng đồ thị đối xứng như hình vẽ.

Dữ liệu lệch trái: Khi 3 giá trị *Mean*, *Median*, *Mode* cùng lệch về phía bên trái của đồ thị và khi đó dữ liệu cũng được phân bố tập trung đa phần về phía bên trái của đồ thị.

Dữ liệu lệch phải: Khi 3 giá trị *Mean*, *Median*, *Mode* cùng lệch về phía bên phải của đồ thị và khi đó dữ liệu cũng được phân bố tập trung đa phần về phía bên phải của đồ thị.

**Phương sai (Variance):** là thước đo độ biến thiên của các giá trị xung quanh giá trị trung bình số học của chúng, nó cho biết các giá trị đó ở cách giá trị kỳ vọng bao xa. Một cách dễ hiểu hơn, phương sai sẽ cho biết mức độ chênh lệch trong tập dữ liệu.

Cách tính:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

Hình 2.13 Công thức tính Phương sai

Trong đó:  $x_i$  là giá trị của quan sát thứ  $i$

$\mu$  là giá trị trung bình tổng thể

$N$  là tổng số quan sát của tổng thể

Trong Python: dùng hàm *var()*

Ví dụ:

```
import numpy as np
import pandas as pd
from scipy.stats import mode
numlist = [3, 6, 9, -1, 0, -10, 18, 9]

# tính variance trên 1 list
print(np.var(numlist)) #>> 60.9375

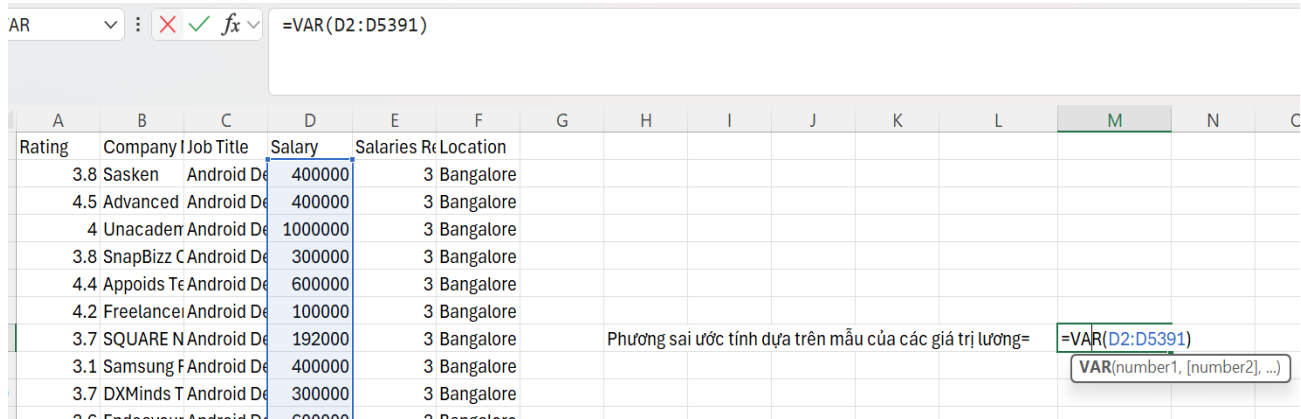
# tính variance trên từng cột dữ liệu trong 1 data frame
var_col = data_frame.var()
```

Hình 2.14 Hàm *var()* trong python

Trong Excel: Nhập hàm VAR()

Ví dụ:

Nhập =VAR(D2:D5391) và ấn 'Enter'



Rating	Company	Job Title	Salary	Location
3.8	Sasken	Android Developer	400000	3 Bangalore
4.5	Advanced	Android Developer	400000	3 Bangalore
4	Unacademy	Android Developer	1000000	3 Bangalore
3.8	SnapBizz	Android Developer	300000	3 Bangalore
4.4	Appoids	Android Developer	600000	3 Bangalore
4.2	Freelancer	Android Developer	100000	3 Bangalore
3.7	SQUARE	Android Developer	192000	3 Bangalore
3.1	Samsung	Android Developer	400000	3 Bangalore
3.7	DXMinds	Android Developer	300000	3 Bangalore

Hình 2.15 Hàm VAR() trong Excel

**Standard Deviation (stdev):** là thước đo độ phân tán của các giá trị trong một tập dữ liệu đã cho từ giá trị trung bình của chúng. Nó cho biết trung bình mỗi giá trị nằm bao xa so với giá trị trung bình.

Cách tính:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$$

Hình 2.16 Công thức tính Standard Deviation

## Trong Python: dùng hàm std()

Ví dụ:

```
import numpy as np
import pandas as pd
from scipy.stats import mode
numlist = [3, 6, 9, -1, 0, -10, 18, 9]

# tính stdev trên 1 list
print(np.std(numlist)) #>> 7.806247497997997

# tính standard deviation trên từng cột dữ liệu trong 1 data frame
dev_col = data_frame.dev()
```

Hình 2.17 Hàm `std()` trong python

Trong Exel: Nhập hàm *STDEV()*

Ví dụ:

Nhập  $=STDEV(D2:D5391)$  và ấn 'Enter'

Rating	Company	Job Title	Salary	Salaries R	Location
3.8	Sasken	Android D	400000	3	Bangalore
4.5	Advanced	Android D	400000	3	Bangalore
4	Unacademy	Android D	1000000	3	Bangalore
3.8	SnapBizz (	Android D	300000	3	Bangalore
4.4	Appoids T	Android D	600000	3	Bangalore
4.2	Freelance	Android D	100000	3	Bangalore
3.7	SQUARE N	Android D	192000	3	Bangalore
3.1	Samsung	Android D	400000	3	Bangalore
3.7	DXMinds T	Android D	300000	3	Bangalore
3.6	Endeavour	Android D	600000	3	Bangalore
3.6	Craft Silico	Android D	300000	3	Bangalore
3.9	Baronford	Android D	240000	2	Bangalore
3.7	Wibmo	Android D	900000	2	Bangalore
4.8	Retail Puls	Android D	24000	2	Bangalore
3.9	Bookmysh	Android D	600000	2	Bangalore
3.9	Knowledge	Android D	228000	2	Bangalore
3.6	Novopay S	Android D	600000	2	Bangalore

Standard deviation =STDEV(D2:D5391)

STDEV(number1, [number2], ...)

Hình 2.18 Hàm STDEV() trong Excel

### 2.5.2 Về các loại biểu đồ

**Biểu đồ Histogram:** Biểu đồ Histogram là một biểu đồ cột sử dụng để biểu diễn phân phối tần suất của một tập dữ liệu liên tục. Thông qua biểu đồ này, chúng ta có thể dễ dàng nhận ra xu hướng phân phối, biên độ, độ tập trung của dữ liệu và sự phân tán của chúng.

Tạo biểu đồ histogram với Excel: Tạo biểu đồ histogram lương cho bộ dữ liệu chấm công bên dưới.

Rating	Company	Job Title	Salary	Salaries Reported	Location
3.8	Sasken	Android D	300000	3	Bangalore
4.5	Advanced	Android D	400000	3	Bangalore
4	Unacadem	Android D	400000	3	Bangalore
3.8	SnapBizz	Android D	300000	3	Bangalore
4.4	Appoids T	Android D	400000	3	Bangalore
4.2	Freelance	Android D	400000	3	Bangalore
3.7	SQUARE N	Android D	300000	3	Bangalore
3.1	Samsung	Android D	300000	3	Bangalore
3.7	DXMinds	Android D	350000	3	Bangalore
3.6	Endeavou	Android D	300000	3	Bangalore
3.6	Craft Silic	Android D	300000	3	Bangalore
3.9	Baronford	Android D	300000	2	Bangalore
3.7	Wibmo	Android D	300000	2	Bangalore
4.8	Retail Pul	Android D	400000	2	Bangalore
3.9	Bookmysf	Android D	550000	2	Bangalore
3.9	Knowledg	Android D	300000	2	Bangalore
3.6	Novopay	Android D	300000	2	Bangalore
3.7	WealthEng	Android D	500000	2	Bangalore
4	J.P. Morga	Android D	400000	2	Bangalore
3.6	Acviss	Android D	300000	2	Bangalore
4.1	Fresher	Android D	400000	2	Bangalore
4.2	MedOnGo	Android D	400000	2	Bangalore
4	Nuclei	Android D	400000	2	Bangalore
4.4	eSecForte	Android D	400000	2	Bangalore
4.3	Moveinsyr	Android D	400000	2	Bangalore
3.6	Tech Mahi	Android D	300000	2	Bangalore
4	ThiDiff Tec	Android D	400000	2	Bangalore
4.9	Retranz In	Android D	400000	2	Bangalore
4	FicusLot	Android D	400000	2	Bangalore
3.7	KrazyBee	Android D	300000	2	Bangalore
5	powerplay	Android D	600000	2	Bangalore
4.2	Dcoder	Android D	400000	2	Bangalore

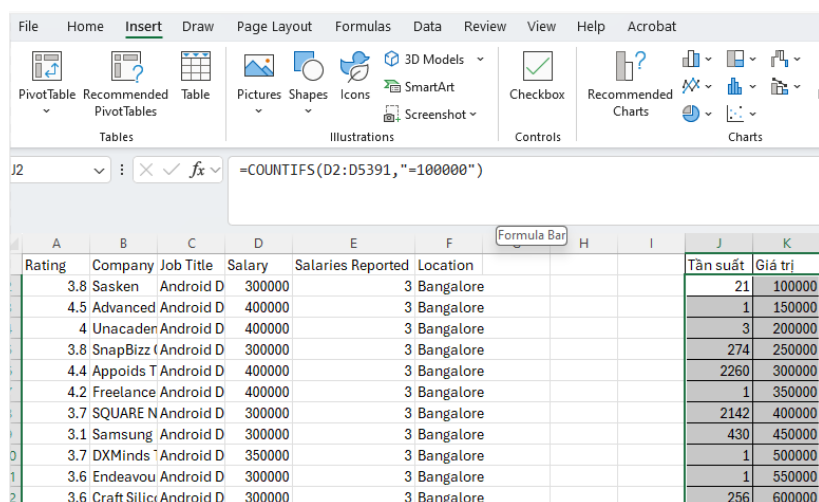
Hình 2.20 Dữ liệu lương

B1: Tạo bảng thống kê tần suất của từng giá trị trong cột thuộc tính ‘Lương’ bằng lệnh `=COUNTIFS(D2:B5391,"=100000")` trong đó ‘D5:D5391’ là phạm vi dãy dữ liệu, ‘=100000’ là giá trị cần tính tần suất, tương tự đối với các giá trị còn lại: 150000, 200000, 250000, 300000, 350000, 400000, 450000, 500000, 550000 và 600000.

Tần suất	Giá trị
21	100000
1	150000
3	200000
274	250000
2260	300000
1	350000
2142	400000
430	450000
1	500000
1	550000
256	600000

Hình 2.21 Bảng thống kê tần suất của từng giá trị trong cột thuộc tính

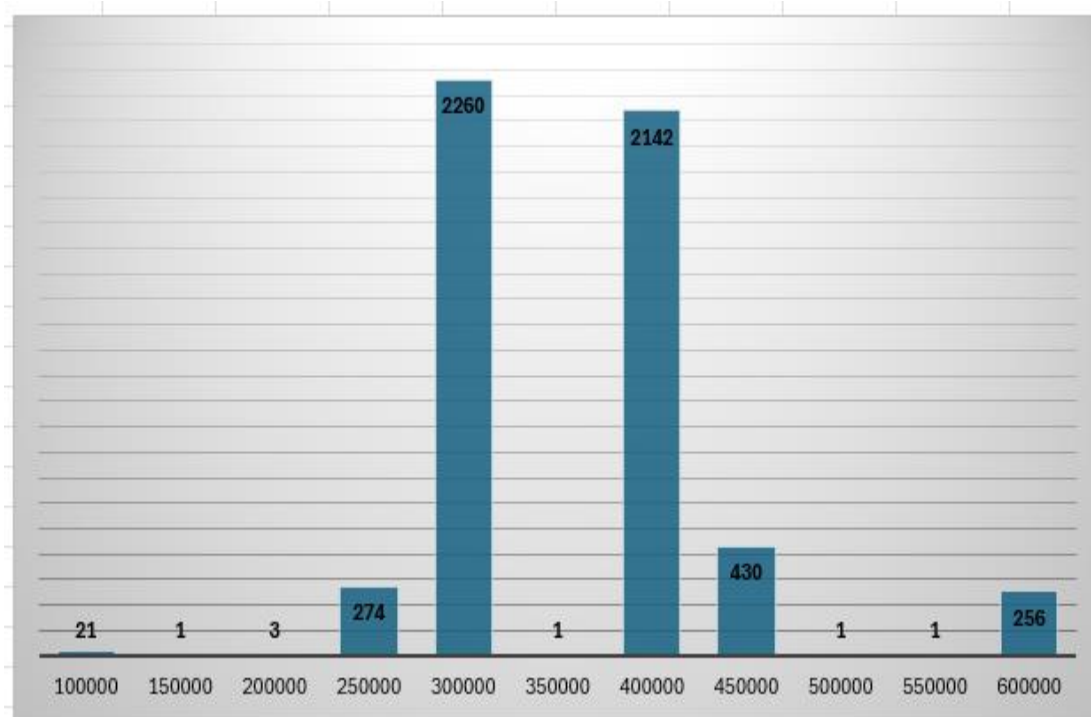
B2: Chọn ‘Insert’ và chọn vào icon có biểu tượng đồ thị histogram như hình vẽ.



Hình 2.22 Chọn điều đồ histogram



Kết Quả:



*Hình 2.23 biểu đồ histogram của lương*

Từ biểu đồ ta thấy rõ được tần suất của từng số lương xuất hiện : có 21 người có lương là 100.000, 1 người có có lương là 150000, 3 người có lương 200000, 274 người có lương là 250000, 2260 người có lương là 300000 ...

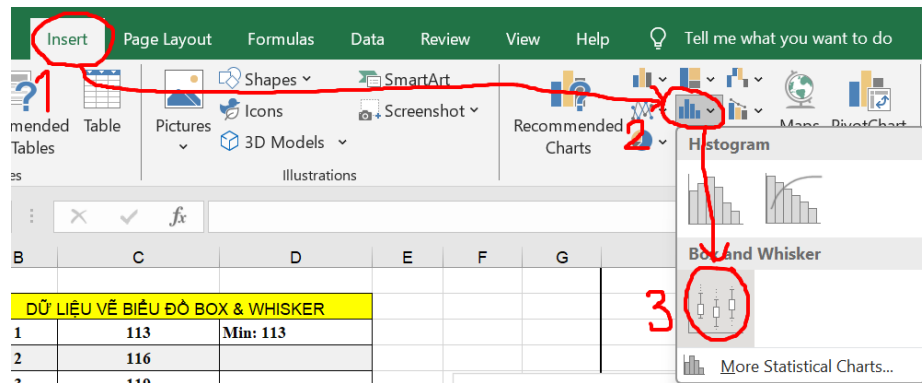
**Biểu đồ Box & whisker:** Biểu đồ hộp (Box plot) hay còn gọi là biểu đồ hộp và râu (Box and whisker plot) là biểu đồ diễn tả 5 vị trí phân bố của dữ liệu, đó là: giá trị nhỏ nhất (min), tứ phân vị thứ nhất (Q1), trung vị (median), tứ phân vị thứ 3 (Q3) và giá trị lớn nhất (max).

Tạo biểu đồ Box plot với Excel:

DỮ LIỆU VỀ BIỂU ĐỒ BOX & WHISKER		
1	113	Min: 113
2	116	
3	119	
4	121	
5	124	Q1: 124
6	124	
7	125	
8	126	
9	126	
10	126	Median (Q2): 126.5
11	127	
12	127	
13	128	
14	129	
15	130	Q3: 130
16	130	
17	131	
18	132	
19	133	
20	136	Max: 136

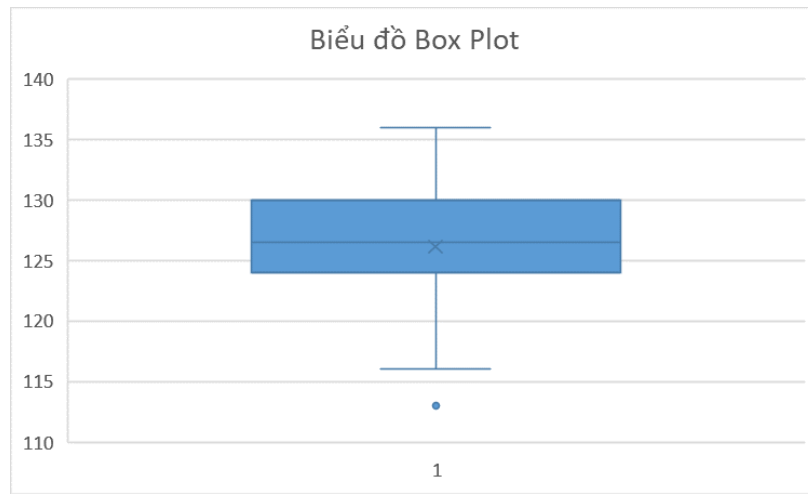
Hình 2.24 Bảng dữ liệu về biểu đồ Box & whisker

Đầu tiên ta chọn vùng dữ liệu cần để vẽ biểu đồ sau đó chọn ‘Insert’ và chọn vào icon có biểu tượng biểu đồ box plot như hình vẽ



Hình 2.25 Chọn biểu đồ box plot

Kết quả:



Hình 2.26 biểu đồ box plot của bảng dữ liệu

Từ kết quả trên ta có thể thấy xuất hiện một điểm dữ liệu ngoại lệ (*outliers*), đó chính là điểm dữ liệu  $min(113)$  trên toàn bộ dữ liệu. Một điểm dữ liệu được gọi là ngoại lệ khi nó nằm cách điểm mang giá trị Q1 về phía bên dưới hoặc nằm cách điểm mang giá trị Q3 về phía bên trên một đoạn quá 1.5 lần độ trải giữa ( $1.5 \times IQR$ ). Trong trường hợp này điểm mang giá trị 113 nằm cách điểm mang giá trị Q1 về phía bên dưới một đoạn 11 ( $124 - 113$ ) lớn hơn so với  $1.5 \times (130 - 124) = 9$ .

**Biểu đồ Scatter:** Biểu đồ phân tán (scatter plot) là loại biểu đồ trực quan hóa mối tương quan giữa 2 biến số dựa vào các tọa độ toán học. Mối tương quan này được biểu diễn dưới dạng các dấu chấm tròn đại diện cho 2 biến, với một biến phụ thuộc chạy cố định trên trục tung và một biến độc lập chạy cố định dựa vào trục hoành.

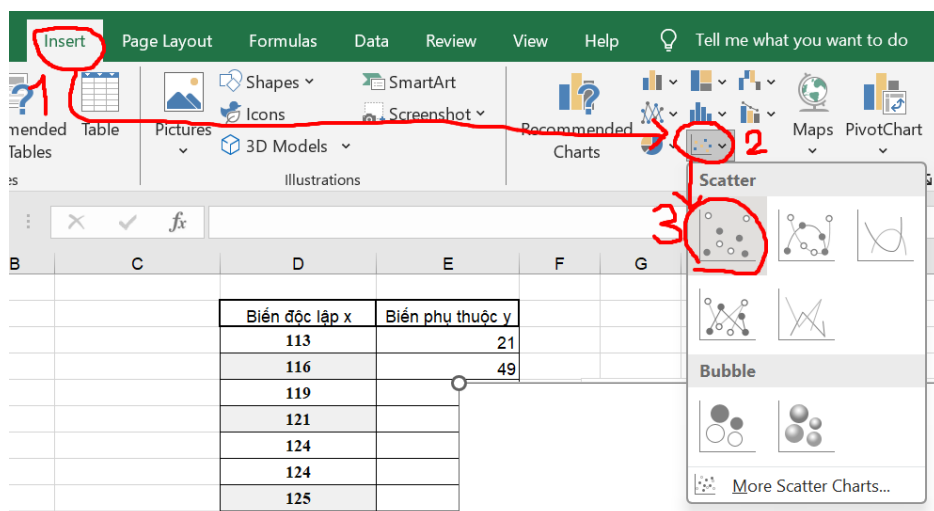
Tạo biểu đồ scatter plot với Excel:

Với bộ dữ liệu gồm 2 thuộc tính là biến độc lập x và biến phụ thuộc y

Biến độc lập x	Biến phụ thuộc y
113	21
116	49
119	32
121	57
124	32
124	90
125	12
126	94
126	56
126	87
127	120
127	5
128	19
129	43
130	67
130	4
131	76
132	36
133	9
136	11

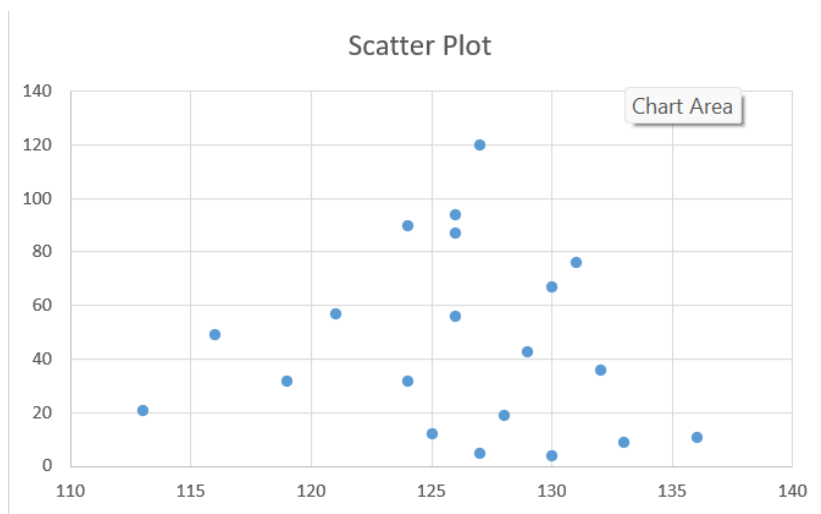
*Hình 2.27 Bảng dữ liệu với biến độc lập x và biến phụ thuộc y*

Sau khi kéo chuột chọn toàn bộ vùng dữ liệu cần để vẽ biểu đồ ta chọn ‘*Insert*’ và chọn vào icon có biểu tượng biểu đồ scatter plot như hình vẽ



Hình 2.28 Chọn biểu đồ scatter plot

Kết quả:



Hình 2.29 biểu đồ scatter plot của bảng dữ liệu

Mỗi một chấm xanh tượng trưng cho 1 điểm dữ liệu và biểu diễn 2 giá trị của 2 thuộc tính: biến độc lập (trục hoành) và biến phụ thuộc (trục tung)

## CHƯƠNG 3: THỰC NGHIỆM VÀ ĐÁNH GIÁ

### 3.1 Dữ liệu thực nghiệm

Dữ liệu của bài toán phân tích mô tả hành vi mua hàng của khách hàng được thu thập từ Kaggle - một nền tảng trực tuyến cho cộng đồng Machine Learning (ML) và Khoa học dữ liệu, cho phép người dùng chia sẻ, tìm kiếm các bộ dữ liệu; tìm hiểu và xây dựng models, tương tác với những nhà khoa học và kỹ sư ML trên toàn thế giới. Tên của bộ dữ liệu trên Kaggle là ‘Software Professional Salaries’ chứa thông tin những chuyên gia phần mềm trong năm 2022. Cụ thể bộ dữ liệu được lưu dưới dạng file csv bao gồm 5391 bản ghi và 6 thuộc tính chứa các thông tin về các chuyên gia phần mềm. Đó là:

**Rating** - Xếp hạng của Công ty do Nhân viên đưa ra

**Company Name** - Tên công ty

**Job Title** - Vị trí công việc hoặc Chức danh

**Salary** - Tổng tiền lương bằng USD

**Salaries Reported** - Số lần lương được báo cáo

**Location** - Địa điểm Công ty

Rating	Company Name	Job Title	Salary	Salaries Reported	Location
3.8	Sasken	Android Developer	300000		3 Bangalore
4.5	Advanced Millennium Technologies	Android Developer	400000		3 Bangalore
4	Unacademy	Android Developer	400000		3 Bangalore
3.8	SnapBizz Cloudtech	Android Developer	300000		3 Bangalore
4.4	Appoids Tech Solutions	Android Developer	400000		3 Bangalore
4.2	Freelancer	Android Developer	400000		3 Bangalore
3.7	SQUARE N CUBE	Android Developer	300000		3 Bangalore
3.1	Samsung R&D Institute India - Bangalore	Android Developer	300000		3 Bangalore
3.7	DXMinds Technologies	Android Developer	350000		3 Bangalore
3.6	Endeavour Software Technologies	Android Developer	300000		3 Bangalore
3.6	Craft Silicon	Android Developer	300000		3 Bangalore
3.9	Baronford & Associates	Android Developer	300000		2 Bangalore
3.7	Wibmo	Android Developer	300000		2 Bangalore
4.8	Retail Pulse	Android Developer - Intern	400000		2 Bangalore
3.9	Bookmyshow	Android Developer	550000		2 Bangalore
3.9	Knowledge Flex	Android Developer	300000		2 Bangalore
3.6	Novopay Solutions	Android Developer	300000		2 Bangalore
3.7	WealthEngine	Android Developer	500000		2 Bangalore
4	J.P. Morgan	Android Developer	400000		2 Bangalore
3.6	Acviss	Android Developer	300000		2 Bangalore
4.1	Fresher	Android Developer	400000		2 Bangalore
4.2	MedOnGo	Android Developer	400000		2 Bangalore
4	Nuclei	Android Developer	400000		2 Bangalore
4.4	eSecForte Technologies	Android Developer	400000		2 Bangalore
4.3	Moveinsync Technology Solutions	Android Developer	400000		2 Bangalore

Hình 3.1 Một số bản ghi trong bộ dữ liệu

### 3.2 Công cụ sử dụng để thực hiện bài toán

Trong bài tập lớn này em sử dụng công cụ thực hiện là PyCharm - là một Môi trường Phát triển Tích hợp (IDE) mạnh mẽ và chuyên dụng dành cho lập trình Python. Được phát triển bởi JetBrains, PyCharm cung cấp nhiều tính năng hữu ích giúp các lập trình viên viết, kiểm thử, gỡ lỗi và quản lý mã Python một cách hiệu quả và tiện lợi.

### 3.3 Các thư viện sử dụng

Một số thư viện cần thiết cho việc xử lý dữ liệu

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
import plotly.graph_objects as go
from PIL import Image
from IPython.display import Image
```

*Hình 3.2 Một số thư viện xử lý dữ liệu*

Những thư viện trên cung cấp các công cụ mạnh mẽ cho việc khám phá và xử lý dữ liệu, cũng như trực quan hóa thông tin một cách hiệu quả, cụ thể:

**Seaborn (sns) và Matplotlib (plt):** Seaborn là một thư viện trực quan hóa dữ liệu dựa trên Matplotlib, giúp tạo ra các biểu đồ thống kê và trực quan hóa dữ liệu một cách dễ dàng và đẹp mắt.

**Pandas** là một thư viện mạnh mẽ dành cho thao tác và phân tích dữ liệu trong Python. Nó đặc biệt hữu ích cho việc làm việc với dữ liệu dạng bảng (như Excel, CSV, SQL, hoặc các loại dữ liệu tương tự).

**NumPy (Numerical Python)** là thư viện cơ bản cho toán học khoa học trong Python, đặc biệt mạnh trong việc xử lý và tính toán trên mảng số học đa chiều (arrays).

**Pillow (Python Imaging Library, còn gọi là PIL)** là một thư viện xử lý hình

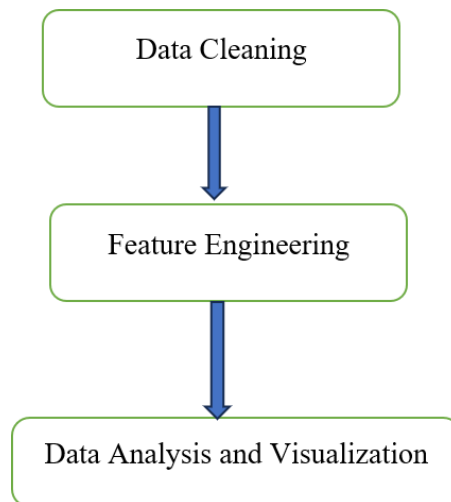
ảnh phổ biến trong Python. Nó là phiên bản được cải tiến từ PIL (thư viện ban đầu) và hỗ trợ nhiều chức năng xử lý hình ảnh.

**IPython.display** là một module trong IPython, được sử dụng để hiển thị nội dung đa phương tiện (như hình ảnh, âm thanh, video, HTML) trong các môi trường tương tác như Jupyter Notebook.

**Plotly Express (px), Plotly Graph Objects (go) và Make Subplots:** Plotly là một thư viện mạnh mẽ cho việc tạo các biểu đồ tương tác. Plotly Express (px) cung cấp cách dễ dàng để tạo các biểu đồ phức tạp, trong khi Plotly Graph Objects (go) cho phép tùy chỉnh chi tiết hơn. Make Subplots giúp tạo ra các subplot để so sánh và hiển thị nhiều biểu đồ trên cùng một hình ảnh.

### 3.4 Quy trình thực nghiệm

Quy trình thực hiện bài toán bao gồm 3 bước lớn:



*Hình 3.3 Quy trình thực hiện*



### 3.4.1 Làm sạch dữ liệu

Làm sạch dữ liệu là một trong những bước tiền xử lý quan trọng mà em sẽ áp dụng cho bài tập lớn này sau khi tiến hành khảo sát dữ liệu.

#### 3.4.1.1 Tải tập dữ liệu

Đoạn code thực hiện:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
import plotly.graph_objects as go
from PIL import Image
from IPython.display import Image

df = pd.read_csv("Software_Professional_Salaries.csv")
print(df.head())
```

Hình 3.4.1 Tải bộ tập dữ liệu

Kết quả:

	Rating	Company Name	...	Unnamed: 9	Unnamed: 10
0	3.8	Sasken	...	NaN	NaN
1	4.5	Advanced Millennium Technologies	...	NaN	NaN
2	4.0	Unacademy	...	NaN	NaN
3	3.8	SnapBizz Cloudtech	...	NaN	NaN
4	4.4	Appoids Tech Solutions	...	NaN	NaN

#### 3.4.1.2 Hình dạng của bộ dữ liệu

```
df = pd.read_csv("Software_Professional_Salaries.csv")
print(df.shape)
```

Hình 3.4.2 Hình dạng của bộ dữ liệu

Kết Quả: (5390, 11)

### 3.4.1.3 Hiển thị các columns có trong bộ dữ liệu

```
df = pd.read_csv("Software_Professional_Salaries.csv")
print(df.columns)
```

Hình 3.4.3 Hiển thị columns trong bộ dữ liệu

Kết Quả:

```
Index(['Rating', 'Company Name', 'Job Title', 'Salary', 'Salaries Reported',
      'Location', 'Unnamed: 6', 'Unnamed: 7', 'Unnamed: 8', 'Unnamed: 9',
      'Unnamed: 10'],
      dtype='object')
```

### 3.4.4 Hiển thị thông tin của bộ dữ liệu

```
df = pd.read_csv("Software_Professional_Salaries.csv")
print(df.info)
```

Kết Quả:

```
<bound method DataFrame.info of      Rating
0      3.8      Sasken ...
1      4.5  Advanced Millennium Technologies ...
2      4.0      Unacademy ...
3      3.8      SnapBizz Cloudtech ...
4      4.4      Appoids Tech Solutions ...
...    ...      ... ...
5385    4.7      Expert Solutions ...
5386    4.0      Nextgen Innovation Labs ...
5387    4.1      Fresher ...
5388    4.1      Accenture ...
5389    3.8      Thomson Reuters ...

[5390 rows x 11 columns]>
```

#### 3.4.1.5 Kiểm tra nếu có bất kì giá trị null trong bộ dữ liệu

```
df = pd.read_csv("Software_Professional_Salaries.csv")
print(df.isnull().sum())
```

Hình 3.4.5 Đoạn code kiểm tra dữ liệu null

Kết Quả:

```
Rating      0
Company Name 0
Job Title    0
Salary       0
Salaries Reported 0
Location     0
```

#### 3.4.1.6 Hiển thị location xuất hiện trong bộ dữ liệu

```
df = pd.read_csv("Software_Professional_Salaries.csv")
print(df["Location"].unique())
```

Hình 3.4.6 Đoạn code hiển thị locations

Kết Quả:

```
['Bangalore' 'Chennai' 'Hyderabad' 'New Delhi' 'Pune']
```

#### 3.4.1.7 Kiểm tra Trung bình, Trung vị, Tối đa và Tối thiểu của Lương

```
df = pd.read_csv("Software_Professional_Salaries.csv")
print("Mức lương trung bình:", round(df["Salary"].mean()))
print("Mức lương trung vị:", round(df["Salary"].median()))
print("Mức lương cao nhất:", round(df["Salary"].max()))
print("Mức lương thấp nhất:", round(df["Salary"].min()))
```

Hình 3.4.7 Đoạn code hiển thị Trung bình, Trung vị,

## Tối đa và Tối thiểu của Lương

*Kết Quả:*

```
Mức lương trung bình: 362644  
Mức lương trung vị: 400000  
Mức lương cao nhất: 600000  
Mức lương thấp nhất: 100000
```

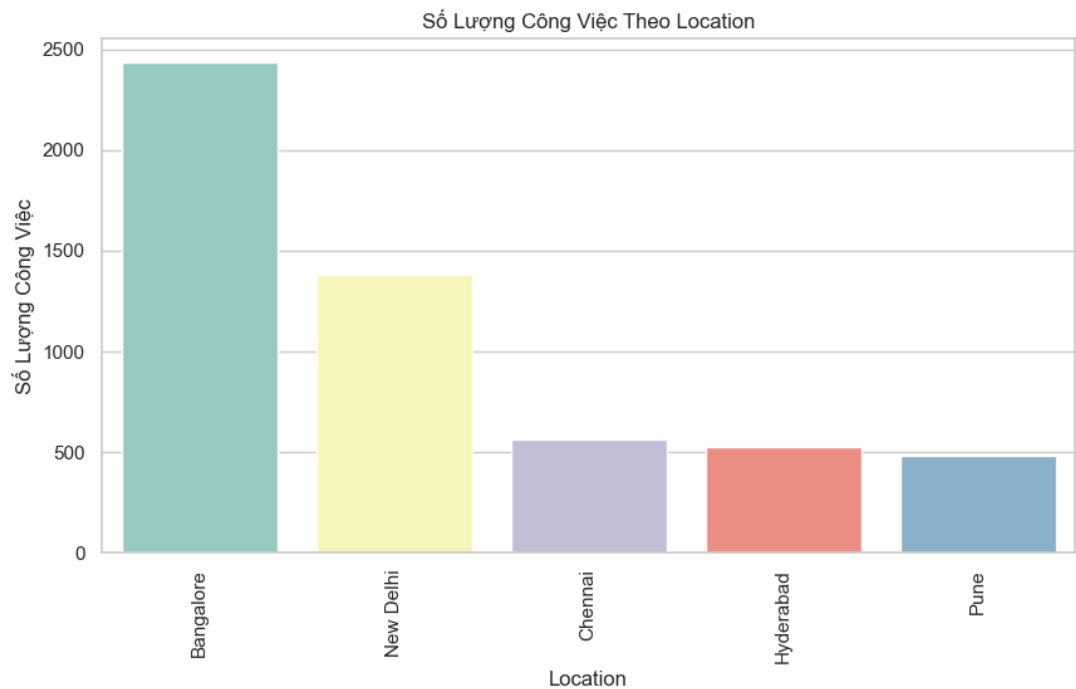
### 3.4.2 Data Analysis and Visualization

#### 3.4.2.1 Biểu đồ Barplot cho vị trí

Đoạn code:

```
df = pd.read_csv("Software_Professional_Salaries.csv")  
# Đếm số lượng công việc ở mỗi location  
location_counts = df['Location'].value_counts()  
  
# Tạo barplot cho số lượng công việc ở mỗi location  
plt.figure(figsize=(10, 6))  
sns.set(style="whitegrid")  
  
# Vẽ Barplot cho location và số lượng công việc  
sns.barplot(x=location_counts.index, y=location_counts.values, palette="Set3")  
  
# Xoay nhãn trục X để dễ đọc hơn nếu tên location dài  
plt.xticks(rotation=90)  
  
# Thêm tiêu đề và nhãn trục  
plt.title("Số Lượng Công Việc Theo Location")  
plt.xlabel("Location")  
plt.ylabel("Số Lượng Công Việc")  
  
plt.show()
```

*Kết Quả:*



*Hình 3.4.8 Biểu đồ Barplot cho số lượng công việc theo location*

Nhìn vào biểu đồ ta thấy được sự phân bố số lượng công việc ở mỗi location .

### *3.4.2.2 Biểu đồ Barplot thể hiện mức lương trung bình tại mỗi location*

Đoạn code:

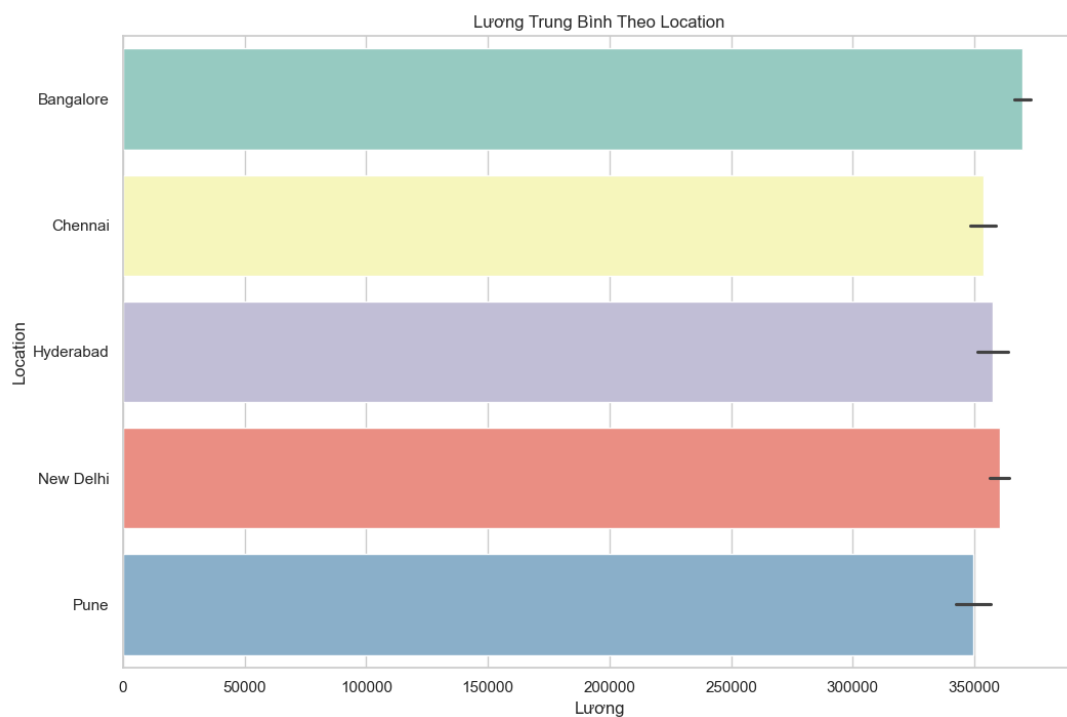
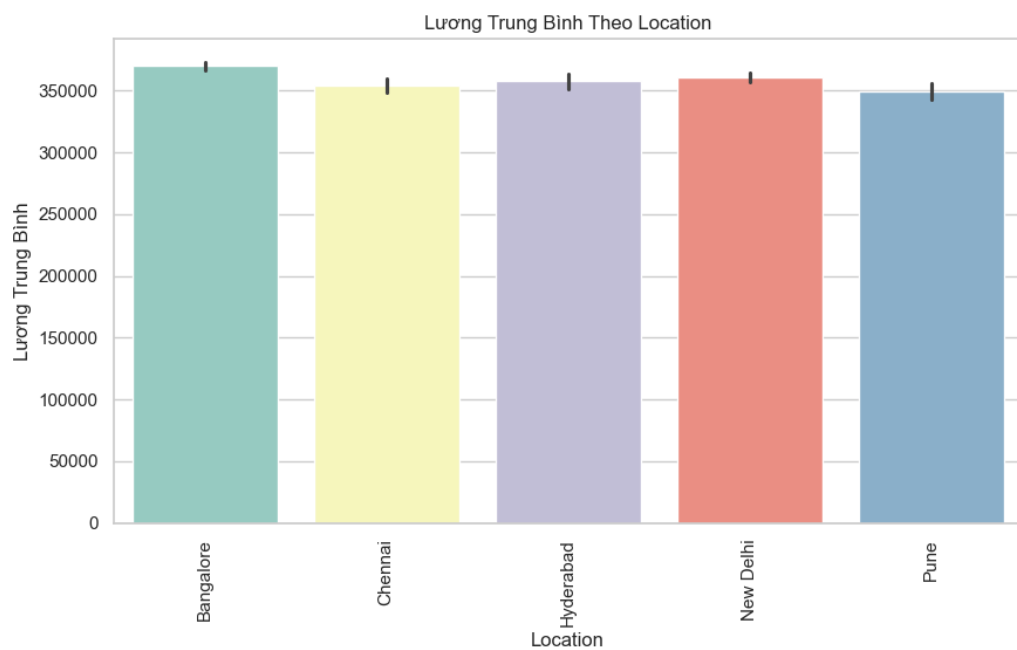
```
df = pd.read_csv("Software_Professional_Salaries.csv")
# Tạo barplot để thể hiện lương trung bình tại mỗi location
plt.figure(figsize=(10, 6))
sns.set(style="whitegrid")

# Vẽ Barplot cho Location và mức lương trung bình
sns.barplot(x="Location", y="Salary", data=df, palette="Set3")

# Xoay nhãn trục X để dễ đọc hơn nếu tên location dài
plt.xticks(rotation=90)

# Thêm tiêu đề và nhãn trục
plt.title("Lương Trung Bình Theo Location")
plt.xlabel("Location")
plt.ylabel("Lương Trung Bình")
plt.show()
```

*Kết Quả:*



*Hình 3.4.9 Biểu đồ Barplot thể hiện mức lương trung bình tại mỗi location*

Nhìn vào biểu đồ ta có thể thấy mức lương trung bình ở Bangalore là cao nhất

### 3.4.2.3 Biểu đồ Barplot cho số lượng các chức danh công việc khác nhau

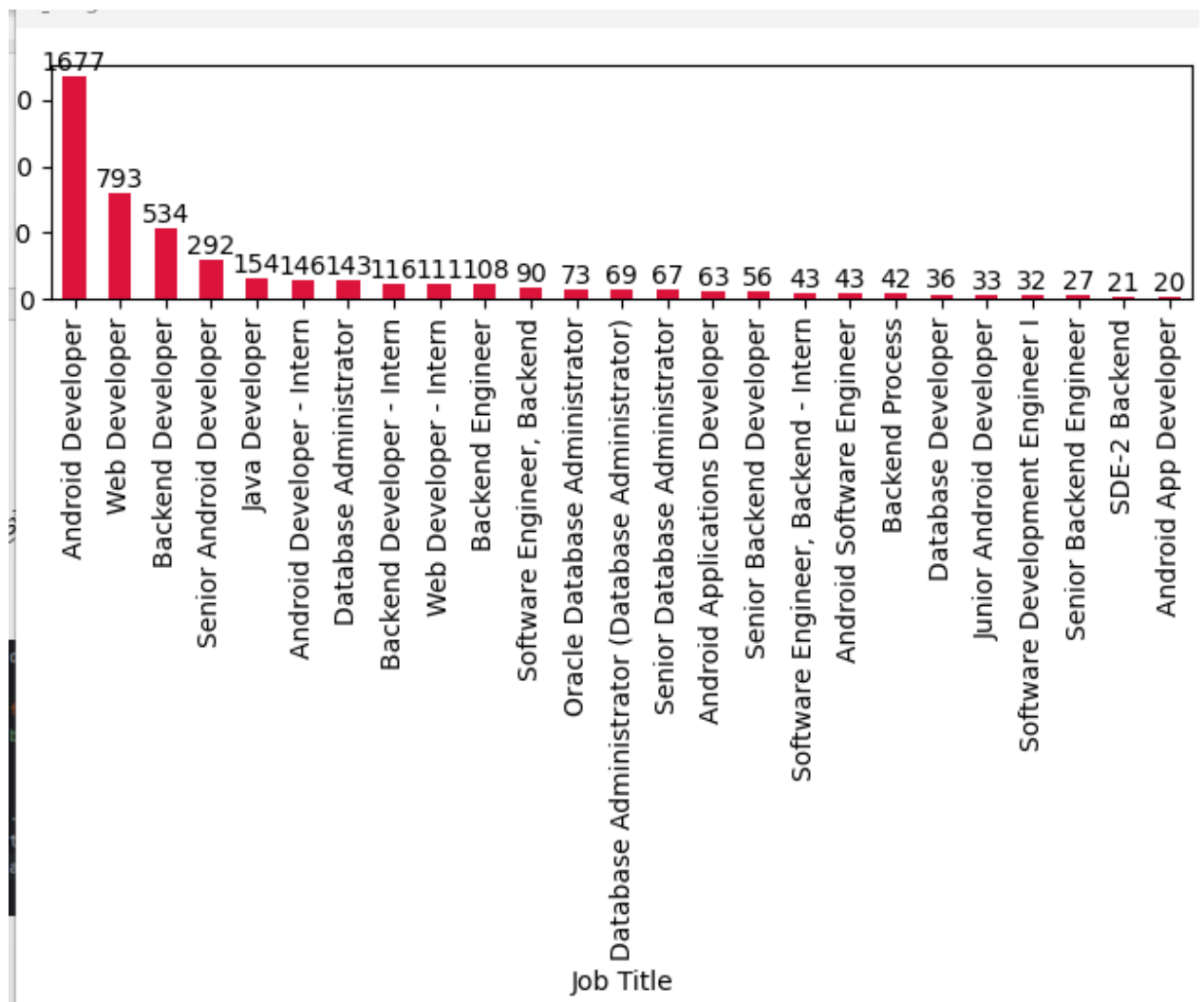
Đoạn code:

```
df = pd.read_csv("Software_Professional_Salaries.csv")

plt.figure(figsize = (20, 5))
ax = df["Job Title"].value_counts()[:25].plot(kind = 'bar',
                                              color = 'crimson')

for p in ax.patches:
    ax.annotate(int(p.get_height()), (p.get_x() + 0.25, p.get_height() + 1), ha = 'center', va = 'bottom', color = 'black')
plt.tight_layout()
plt.show()
```

Kết Quả



Hình 3.4.10 Biểu đồ Barplot cho số lượng các chức danh công việc khác nhau  
Biểu đồ cho thấy vị trí Android Developer chiếm số lượng lớn nhất

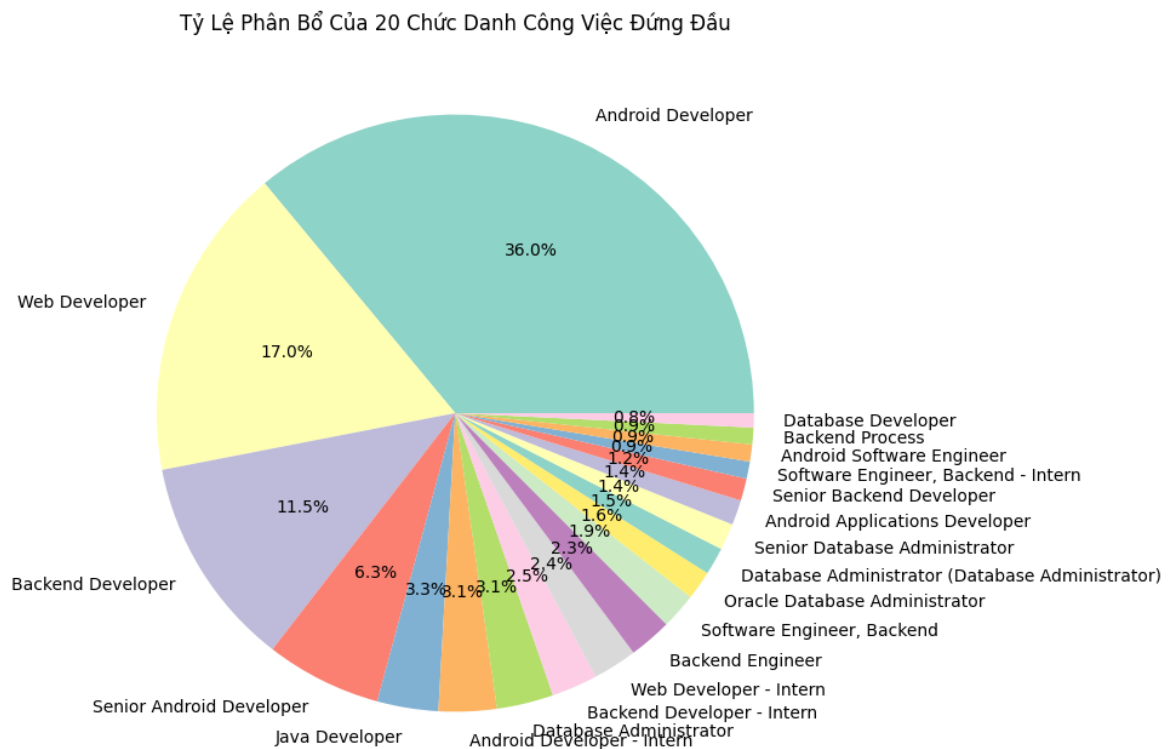
#### 3.4.2.4 Biểu đồ Pie cho 20 chức danh công việc đứng đầu

Đoạn code:

```
df = pd.read_csv("Software_Professional_Salaries.csv")
df["Job Title"].value_counts()[:20].plot.pie(figsize = (7, 7),
                                              autopct = '%1.0f%%')

plt.title("Pie Chart")
plt.xticks(rotation = 90)
plt.show()
```

*Kết Quả*



*Hình 3.4.11 Biểu đồ Pie cho 20 chức danh công việc*

Nhìn vào biểu đồ chúng ta có thể thấy chức danh Android Developer chiếm phần lớn.



### 3.4.2.5 Biểu đồ histogram cho company rating

Đoạn code:

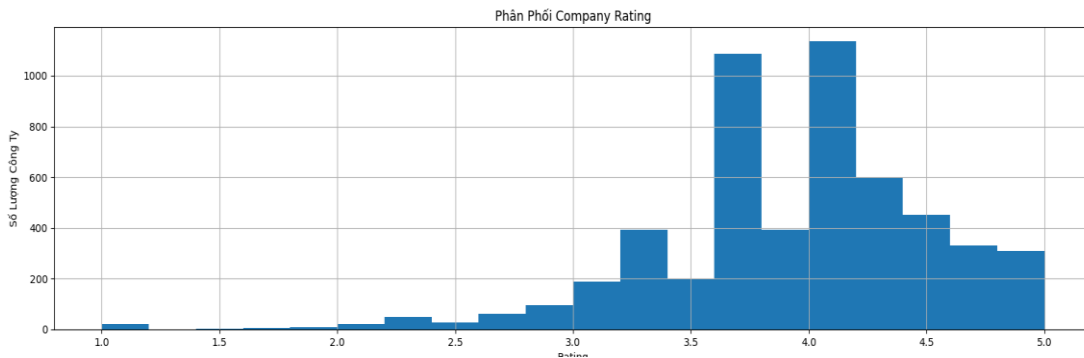
```
df = pd.read_csv("Software_Professional_Salaries.csv")
# Đặt số lượng bins
n_bins = 20

# Tạo biểu đồ histogram với 20 bins
plt.figure(figsize=(20, 5))
df["Rating"].hist(bins=n_bins)

# Thêm tiêu đề và nhãn trục
plt.title("Phân Phối Company Rating")
plt.xlabel("Rating")
plt.ylabel("Số Lượng Công Ty")

# Hiển thị biểu đồ
plt.show()
```

Kết Quả



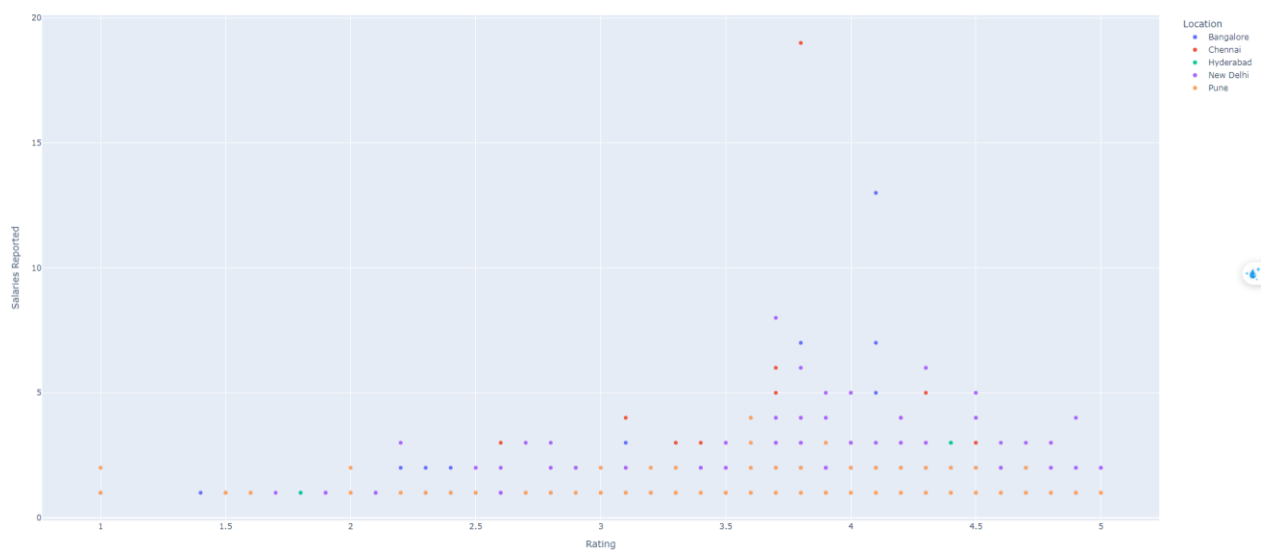
Hình 3.4.12 Biểu đồ histogram cho company rating

### 3.4.2.6 Biểu đồ phân tán về xếp hạng công ty so với mức lương được báo cáo so với vị trí

Đoạn code:

```
df = pd.read_csv("Software_Professional_Salaries.csv")  
fig = px.scatter(df, x = "Rating", y = "Salaries Reported", color = "Location")  
fig.show()
```

Kết Quả



Hình 3.4.13 Biểu đồ phân tán về xếp hạng công ty so với mức lương được báo cáo so với vị trí

## KẾT LUẬN

Trong bối cảnh phát triển mạnh mẽ của ngành công nghiệp phần mềm, việc phân tích và hiểu rõ mức lương của các chuyên gia trong lĩnh vực này là điều cần thiết để đánh giá thị trường lao động và đưa ra những chiến lược phù hợp cho cả người lao động và nhà tuyển dụng. Thông qua bài tiểu luận này, chúng tôi đã thực hiện phân tích mô tả mức lương trong ngành công nghiệp phần mềm dựa trên dữ liệu thực tế, từ đó rút ra một số kết luận quan trọng.

### **Những phát hiện chính:**

- Sự biến động về mức lương giữa các vị trí: Qua phân tích, chúng tôi nhận thấy có sự chênh lệch đáng kể giữa các vị trí công việc trong ngành công nghiệp phần mềm. Các vị trí như Software Engineer, Data Scientist, và DevOps Engineer thường có mức lương cao hơn so với các vị trí khác như QA Engineer hay Technical Support. Điều này phản ánh nhu cầu cao và sự khan hiếm nhân lực có kỹ năng chuyên sâu ở những vị trí này.
- Ảnh hưởng của vị trí địa lý: Phân tích cho thấy sự khác biệt lớn về mức lương giữa các khu vực địa lý. Các thị trường phát triển như Mỹ, Châu Âu và Ấn Độ thường có mức lương cao hơn so với các khu vực khác. Điều này có thể do sự khác biệt về chi phí sinh hoạt, mức độ phát triển của nền công nghiệp phần mềm ở mỗi quốc gia, và sự cạnh tranh trên thị trường lao động.
- Tỷ lệ lương và đánh giá công ty: Qua biểu đồ histogram và các phân tích khác, chúng tôi cũng nhận thấy mối liên hệ giữa mức lương và đánh giá công ty. Các công ty có rating cao thường cung cấp mức lương cạnh tranh hơn để thu hút nhân tài, trong khi những công ty có rating thấp hơn có xu hướng trả lương thấp hơn.
- Phân bổ công việc và lương: Biểu đồ phân bố số lượng công việc theo từng location và company name cho thấy rằng có một số công ty và khu vực nổi bật hơn về số lượng công việc, từ đó dẫn đến sự khác biệt trong mức lương và yêu

cầu về kỹ năng.

### **Hạn chế và hướng nghiên cứu tiếp theo:**

Mặc dù phân tích đã cung cấp một cái nhìn tổng quan về mức lương trong ngành công nghiệp phần mềm, vẫn còn một số hạn chế nhất định. Chúng tôi chỉ tập trung vào dữ liệu có sẵn và chưa mở rộng ra các yếu tố khác như phúc lợi, lương thưởng hay yêu cầu kỹ năng đặc thù của mỗi vị trí. Do đó, trong các nghiên cứu tiếp theo, việc mở rộng phạm vi phân tích để bao quát toàn diện hơn về các yếu tố ảnh hưởng đến mức lương và sự phát triển nghề nghiệp sẽ mang lại cái nhìn sâu sắc hơn.

### **Kết luận tổng quan:**

Bài tiểu luận đã làm rõ một số khía cạnh quan trọng về mức lương trong ngành công nghiệp phần mềm qua các phân tích mô tả và biểu đồ. Thông tin này không chỉ hữu ích cho các nhà quản lý nhân sự trong việc đưa ra các chính sách đãi ngộ hợp lý, mà còn giúp các chuyên gia trong ngành công nghiệp phần mềm hiểu rõ hơn về thị trường lao động và điều chỉnh chiến lược phát triển nghề nghiệp của mình. Trong tương lai, ngành công nghiệp phần mềm được dự đoán sẽ tiếp tục phát triển mạnh mẽ, dẫn đến những thay đổi trong mức lương và yêu cầu kỹ năng của các vị trí công việc. Việc liên tục cập nhật và phân tích dữ liệu sẽ giúp cả nhà tuyển dụng và người lao động thích ứng tốt hơn với xu hướng mới của thị trường.

## TÀI LIỆU THAM KHẢO

Tài liệu tham khảo là website:

[1] Chat GPT – <https://chat.openai.com/>

[2] Big DATA - <https://itnavi.com.vn/blog/big-data>

[3] Descriptive Analytics - <https://viblo.asia/p/descriptive-analytics-phan-tich-mo-ta-la-gi-0gdJzxl3Vz5>

[4] Data Visualization - <https://viblo.asia/p/data-visualization-voi-seaborn-oOVIYP9vZ8W>