

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN  
ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**



**ĐỒ ÁN CUỐI KỲ  
TRỰC QUAN HÓA DỮ LIỆU**

**Nhóm: 14**

**Giảng viên: Bùi Tiến Lên**

**TPHCM, tháng 7 năm 2023**

### **Mức độ hoàn thành các yêu cầu của mỗi thành viên**

<b>STT</b>	<b>Tên thành viên</b>	<b>MSSV</b>	<b>Công việc</b>	<b>Mức độ hoàn thành</b>
1	Nguyễn Quang Gia Bảo	20120040	Trực quan hoá các biểu đồ lên Dashboard bằng công cụ Tableau	100%
2	Nguyễn Việt Khoa	20120120	Viết báo cáo và làm Slide	60%
3	Huỳnh Tuấn Nam	20120136	Viết báo cáo và hỗ trợ ý tưởng cho các biểu đồ.	100%
4	Trần Hoàng Anh Phi	20120158	Trực quan hoá các biểu đồ lên Dashboard bằng công cụ Tableau.	100%

## MỤC LỤC

<b>1. Giới thiệu về Dataset.</b>	<b>4</b>
<b>2. Tiền xử lý dữ liệu.</b>	<b>4</b>
<b>2.1. Khám phá dữ liệu.</b>	<b>4</b>
<b>2.2. Tiền xử lý.</b>	<b>5</b>
<b>3. Trực quan hóa dữ liệu bằng Tableau.</b>	<b>8</b>
<b>4. Reference.</b>	<b>11</b>

## 1. Giới thiệu về Dataset.

- **Chủ đề:** Khảo sát việc sử dụng thời gian của người Việt Nam năm 2022 (Vietnamese Time-Use Survey 2022).
- **Nguồn:** [World Bank Group - International Development, Poverty, & Sustainability](#)
- **Điều khoản truy cập:** Đây là dữ liệu truy cập công cộng, được trích dẫn bởi các thông tin sau đây.
  - + **Điều tra viên chính:** Indochina Research (Viet Nam) Ltd.
  - + **Thời gian thu thập:** Từ 13/10/2022 đến 10/01/2023.
  - + **Nguồn dữ liệu:** [Vietnam - Time-Use Survey 2022 \(worldbank.org\)](#)
  - + **Số tham chiếu khảo sát sử dụng:** 3.
- **Mô tả chung:**

Dự án được triển khai nhằm hỗ trợ Ngân hàng thế giới thu thập dữ liệu phân chia theo giới tính để giám sát khoảng cách giới:

  - + Đóng góp vào dữ liệu sử dụng thời gian vào công việc theo giới tính ở Việt Nam.
  - + Đưa ra các chỉ số đo lường gánh nặng và sự khác biệt về thời gian dành cho các công việc phụ giúp gia đình, chăm sóc người già và trẻ em, theo giới tính, thành thị và dân tộc.

Dự án gồm có 6000 vấn đáp viên ở 6 vùng kinh tế - xã hội, thuộc mọi dân tộc, nghề nghiệp hay mức thu nhập.
- **Mục đích sử dụng:** Nhóm sử dụng bộ dữ liệu để phân tích, đánh giá việc sử dụng thời gian vào các công việc phụ giúp gia đình của nhiều đối tượng từ độ tuổi, dân tộc, nghề nghiệp (thu nhập), ...

## 2. Tiền xử lý dữ liệu.

### 2.1. Khám phá dữ liệu.

Tập thư mục Time-Use Survey 2022 gồm 3 tập dữ liệu là:

- **Cover\_id\_main:**

Gồm có 6001 dòng và 29 cột. Thông tin cơ bản của cuộc phỏng vấn và phân tài sản của 6001 hộ gia đình.

- **Individual\_id\_main:**

Gồm có 6001 dòng và 33 cột. Chứa thông tin giáo dục và việc làm của 6001 người được chọn.

- **Diary main:**

Gồm có 118933 dòng và 26 cột. Chứa nhật ký sử dụng thời gian và hoạt động của 6001 người được phỏng vấn.

## 2.2. Tiền xử lý.

Sử dụng Python3 và thư viện Pandas để thực hiện tiền xử lý dữ liệu trong file “Exploratory.ipynb”. Sử dụng python giúp xác định nhanh các thành phần bị thiếu hay lặp tốt hơn, loại bỏ, đổi tên các cột cũng như thay đổi định dạng nhằm dễ hiểu, dễ phân tích. Các bước tiền xử lý được thực hiện như sau:

- Cover\_id\_main:

Kiểm tra kiểu dữ liệu ta biết được các cột đều có dạng *int64* tức *numerical*. Riêng cột cuối **C10** có dạng *categorical*. Nhưng khi phân tích, cột này thiếu 5654 dữ liệu và theo như ý nghĩa của cột **C9** và **C10** là các câu hỏi cảm nhận về ngày hôm đó. Dường như nó không có quá nhiều ý nghĩa khi phân tích. Ta sẽ xem xét loại bỏ các cột này.

```
assets_18      0
assets_19      0
C9             0
C10            5654
dtype: int64
```

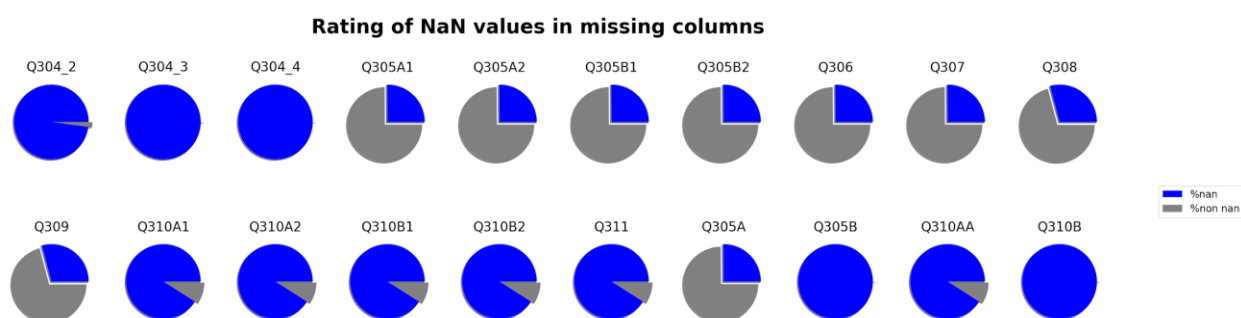
Đối với tên các cột còn lại, các cột có tên mặc định rất khó xác định để phân tích. Ta tiến hành đổi tên các cột cho dễ hiểu hơn. Đồng thời thay đổi kiểu dữ liệu trong data nhằm tăng độ nhận diện, trực quan hóa tốt hơn.

```
Index(['ID', 'MATINH', 'Urban_Rural', 'age', 'C1', 'C2', 'TOTAL_HH_member',
      'assets_1', 'assets_2', 'assets_3', 'assets_4', 'assets_5', 'assets_6',
      'assets_7', 'assets_8', 'assets_9', 'assets_10', 'assets_11',
      'assets_12', 'assets_13', 'assets_14', 'assets_15', 'assets_16',
      'assets_17', 'assets_18', 'assets_19'],
      dtype='object')
```

```
Index(['ID', 'MATINH', 'Urban_Rural', 'age', 'Ethnicity',
      'Num of 15-64y.o HH members', 'TOTAL_HH_member', 'Automobile',
      'Motorbike', 'Bike', 'Ship/boats', 'Pumping machine',
      'Electricity generators', 'Moblie phone', 'Sewing machine', 'TV',
      'Computer/Laptop', 'Refrigerator', 'Air conditioner', 'Washing machine',
      'Gas stove', 'Elctric worker', 'Rice cooker', 'Baking oven',
      'Citrus juicer', 'Piano/Keyboard', 'TENTINH'],
      dtype='object')
```

#### - Individual\_id\_main:

Tương tự như **cover\_id\_main**, các trường dữ liệu đều được đơn giản hóa ở dạng *numerical*, ta sẽ thay đổi các giá trị tương ứng rõ ràng hơn. Nhưng trước khi thực hiện, có vẻ như dữ liệu bị thiếu rất nhiều giá trị:



Có rất nhiều cột thiếu giá trị, những cột thiếu quá nhiều không thể sử dụng để phân tích ta buộc phải loại bỏ. (Các cột có tỉ lệ thiếu trên 50%)

Các cột bị thiếu còn lại, nhìn biểu đồ có cảm quan như chúng bị thiếu với số lượng gần như nhau chứng tỏ chúng có liên hệ với nhau. Ta cần phải phân tích ý nghĩa.

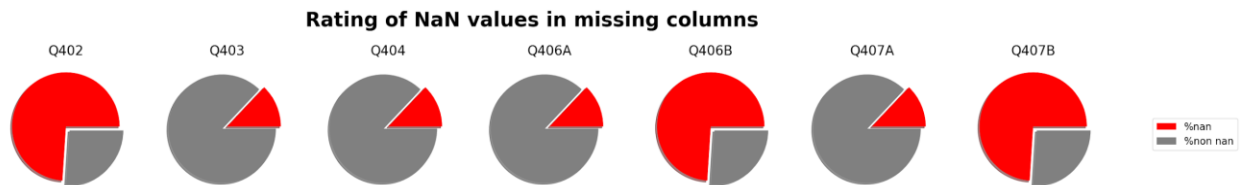
Dựa vào ý nghĩa cột cũng như câu trả lời, các câu này thuộc về loại mô tả công việc, mức lương, nơi làm việc. Có liên quan trực tiếp tới câu **Q304**: “Activities to generate income” và có **1506** câu trả lời là “No”, con số hoàn toàn khớp với dữ liệu bị thiếu ở cột khác. Ta tiến hay lấp vào các dữ liệu bị thiếu số “0” với ý nghĩa là không có công việc.

Tiến hành thay đổi giá trị như với dữ liệu trên nhằm trực quan hóa.

#### - Diary\_main:

Các cột “int1-6” có ý nghĩa quan trọng trong phân tích khi nó thể hiện việc người được khảo sát có hay không giúp đỡ gia đình. Dễ dàng nhận ra nếu không chọn thì tức là không. Ta thay thế dữ liệu thiếu bằng 0.

Thực hiện tương tự như individual\_id\_main xóa các cột bị thiếu, các trường dữ liệu có giá trị thiếu về mặt ý nghĩa không quá quan trọng, ta loại bỏ các cột thiếu trên 60%.



Ta lại thấy các trường bị thiếu còn lại lượng giá trị bị thiếu giống nhau. Giữa chúng có liên quan đến trường dữ liệu Q401 với ý nghĩa là “Main activity”. Nhưng dữ liệu này lại khá khó hiểu nên rất khó để điền một giá trị cụ thể nào đó vào vị trí thiếu. Cũng như việc giá trị thiếu chỉ chiếm phần nhỏ. Có thể suy xét loại bỏ những dòng bị thiếu này vì đây là nhật ký của một người trong nhiều ngày, việc loại bỏ nó có lẽ tốt hơn là thay thế với một giá trị bất kỳ.

### 3. Trực quan hóa dữ liệu bằng Tableau.

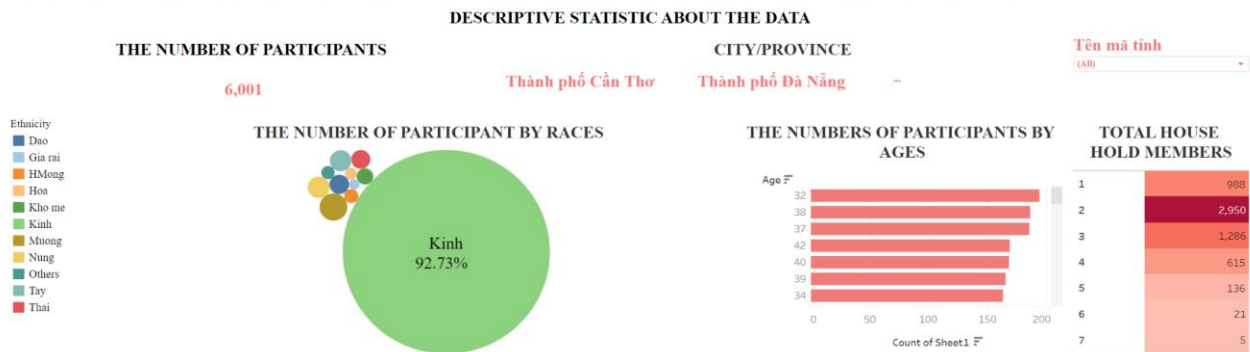
#### Thông kê mô tả

#### Chủ đề: Khảo sát việc sử dụng thời gian của người Việt Nam năm 2022

Dự án được triển khai nhằm hỗ trợ Ngân hàng thế giới thu thập dữ liệu phân chia theo giới tính để giảm sát khoảng cách giới:

+ Đóng góp vào dữ liệu sử dụng thời gian vào công việc theo giới tính ở Việt Nam.

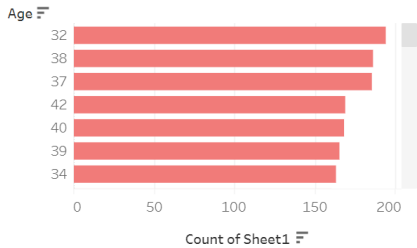
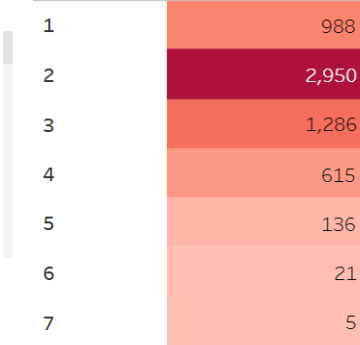
+ Đưa ra các chỉ số đo lường gánh nặng và sự khác biệt về thời gian dành cho các công việc phụ giúp gia đình, chăm sóc người già và trẻ em, theo giới tính, thành thị và dân tộc.



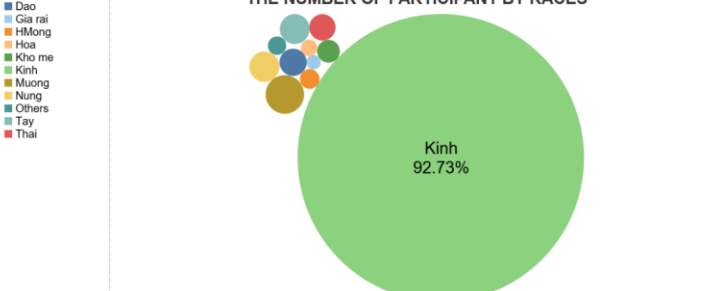
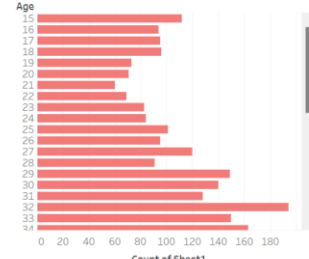
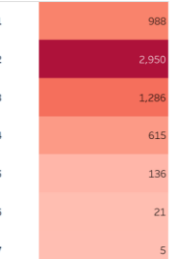
Dựa vào kết quả trực quan, có thể thấy: có 6001 ứng viên tham gia vào cuộc khảo sát.

<p>THE NUMBER OF PARTICIPANTS</p> <p>6,001</p>	Quan sát số lượng ứng viên tham gia ở mỗi tỉnh hoặc cả nước tham gia vào cuộc khảo sát.
<p>THE NUMBER OF PARTICIPANT BY RACES</p> <p>Kinh 92.73%</p>	Quan sát số lượng dân tộc tham gia ở mỗi tỉnh khi tham gia vào cuộc khảo sát.



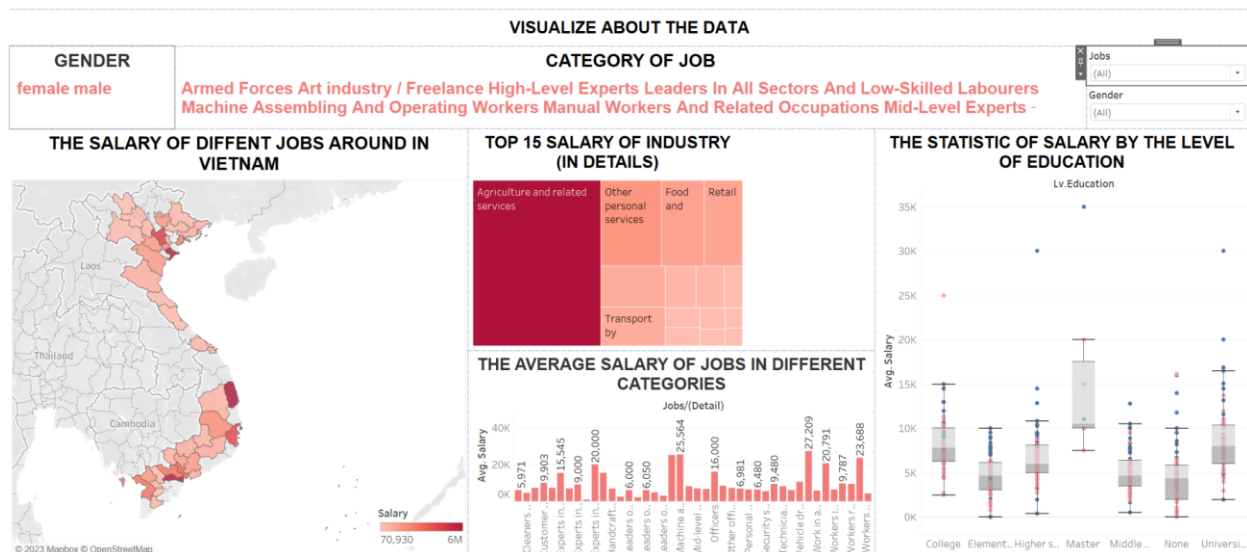
<p><b>THE NUMBERS OF PARTICIPANTS BY AGES</b></p> 	<p>Số lượng tuổi tác của các ứng viên tham gia cuộc khảo sát.</p>
<p><b>TOTAL HOUSE HOLD MEMBERS</b></p> 	<p>Số thành viên trong gia đình của ứng viên tham gia khảo sát.</p>
<p><b>Tên mã tỉnh</b></p> <p>(All)</p>	<p>Sử dụng bảng chọn để thay đổi mã vùng mong muốn để quan sát.</p>

Thay đổi mã vùng “**Thành phố Hồ Chí Minh**”

THE NUMBER OF PARTICIPANTS	CITY/PROVINCE	Tên mã tỉnh	
6,001	Thành phố Cần Thơ Thành phố Đà Nẵng Thành phố Hà Nội Thành phố Hải Phòng Thành Phố Hồ Chí Minh Tỉnh An Giang Tỉnh Bà Rịa - Vũng Tàu Tỉnh Bắc Giang Tỉnh Bạc Liêu Tỉnh Bình Định -	(All)	
<p><b>THE NUMBER OF PARTICIPANT BY RACES</b></p> 	<p><b>THE NUMBERS OF PARTICIPANTS BY AGES</b></p> 	<p><b>TOTAL HOUSE HOLD MEMBERS</b></p> 	

- Dân tộc: Với mỗi dân tộc ở mỗi tỉnh, họ sẽ có sử dụng thời gian vào những việc khác nhau dựa vào văn hóa của mỗi dân tộc.
- Tuổi: Mức tuổi tham gia khảo sát trải dài từ 15 đến 64 tuổi. Trung bình tham gia nhiều nhất ở tầm 30 đến 50.
- Thành viên trong gia đình: Với những người tham gia sống cô đơn (1 mình) thì việc sử dụng thời gian sẽ khác đối với những người tham gia có gia đình ( từ 2 người trở lên như anh em, vợ chồng, ...).

### Mô tả thông tin công việc, số lương trung bình, của các ứng viên.

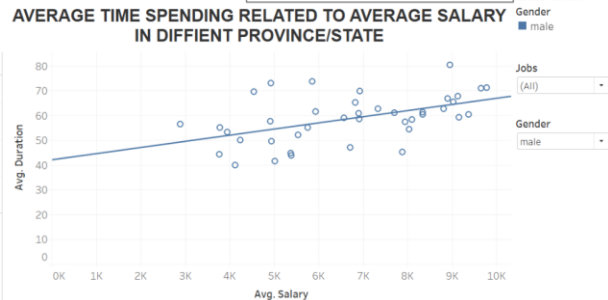
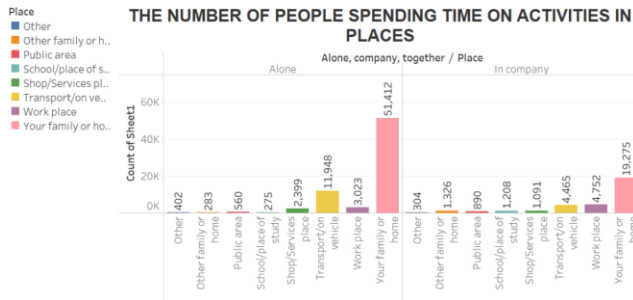
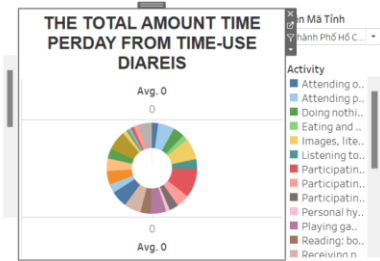
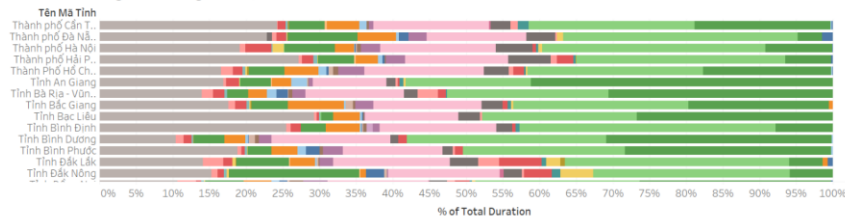


- **Bản đồ:** mô tả thông tin tiền lương của từng công việc phân bố trên cả nước Việt Nam, hoặc ở một tỉnh nào đó được chọn. Phần vùng tô màu cho thấy mức lương trung bình cao thấp ở các khu vực.
- Cũng như mô tả mức trung bình các mức lương theo từng nghề nghiệp, và theo trình độ học vấn của mỗi ID.
- Phân phối mức lương ở từng trình độ học vấn.
- Quan sát bảng “The Statistic of salary by the level of education” cho thấy, mức lương của “**Master**” cao nhất là hoàn toàn chính xác. Tuy nhiên có một số dữ liệu “khai” không đúng - diễn hình việc là còn đi học những mức lương lại cao hơn đáng kể so với “**Master**”.

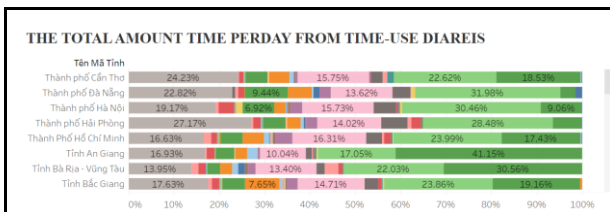
## Thông kê về thời gian sử dụng:

### VISUALIZE ABOUT THE DATA

Phân bố thời gian của người Việt Nam ở các tỉnh



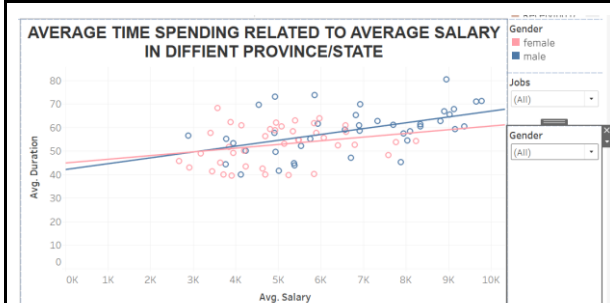
Biểu đồ thống kê mô tả việc sử dụng thời gian ở từng mã tỉnh của các ứng viên



Biểu đồ mô tả % các hoạt động thời gian cho từng mã tỉnh.



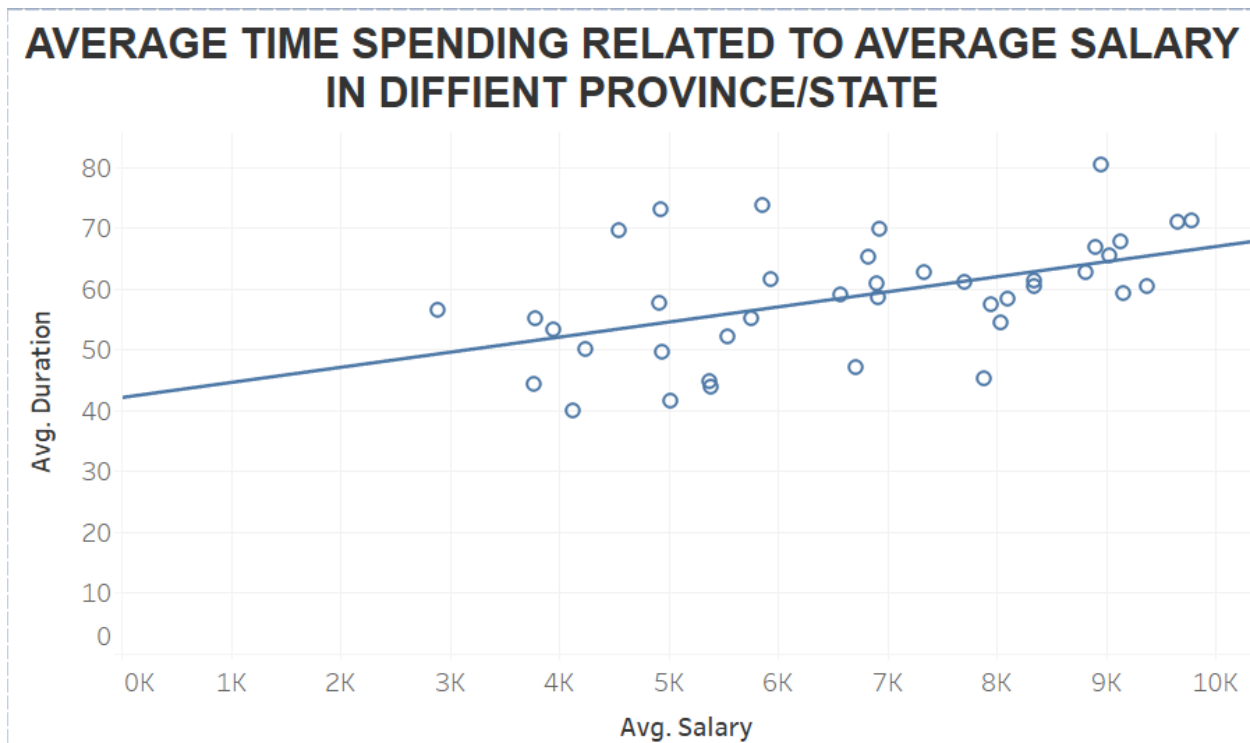
Biểu đồ quan sát địa điểm mà ứng viên tham gia dành thời gian.



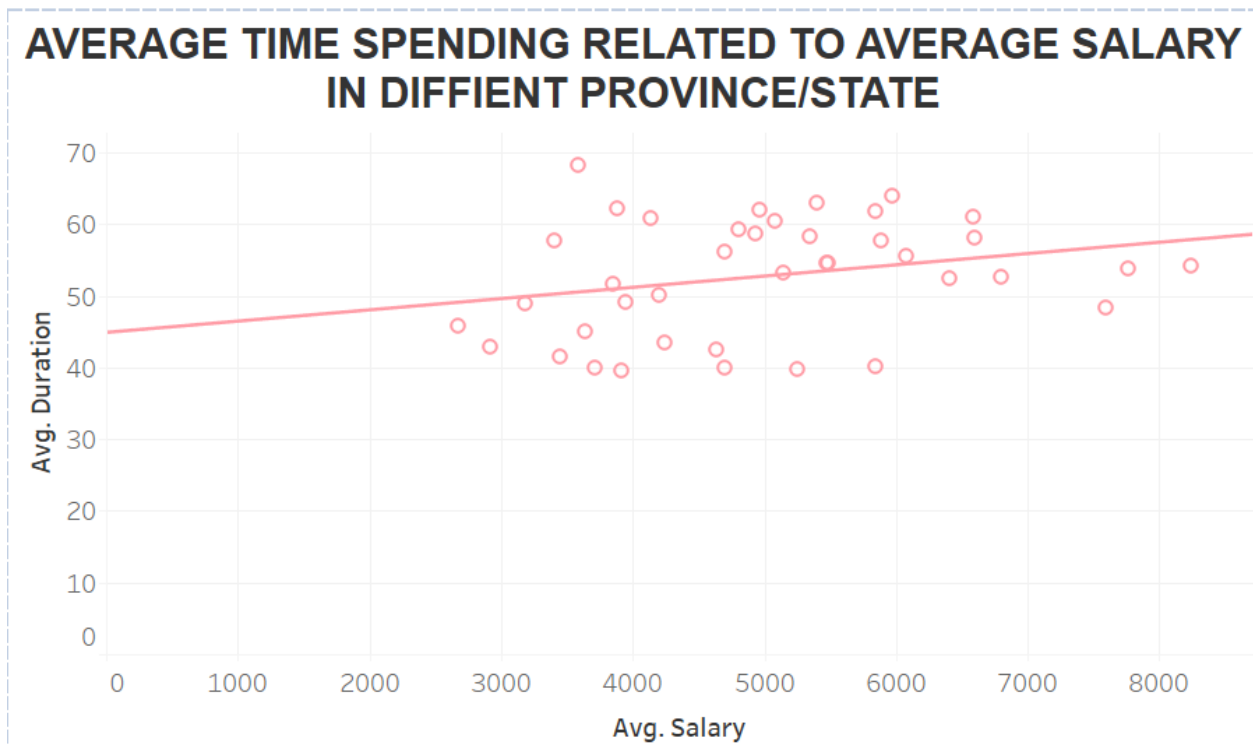
Biểu đồ phân bố hồi quy các giá trị thời gian ứng viên dành ra theo mức lương trung bình ở các vùng.

- Đối với biểu đồ địa điểm mà ứng viên tham gia dành thời gian, có thể thấy được hầu như đối với người “Alone” hoặc “In Company” đều dành hầu hết thời gian cho các hoạt động trong gia đình.

- Biểu đồ hồi quy giữa tiền lương trung bình và thời gian sử dụng thời gian trung bình:
  - Đối với “nam”: việc sử dụng đồng lương của họ cho các công việc thường ngày khá cao. Do họ có những công việc ngoại giao với các đối tác, hoặc “ăn ngoài” với những người bạn.



- Đối với “nữ”: Họ sử dụng những đồng lương ít hơn so với “nam”.



Việc phân tích dữ liệu, hay đưa ra suy đoán về tỉnh nào đó dựa vào mối liên hệ giữa hai Dashboard 2 và 3 (tức là mức lương trung bình của từng nghề nghiệp, và bảng sử dụng thời gian cho các hoạt động của ứng viên). Dựa vào mức lương đầu ra của các ứng viên, ở từng việc làm, hoặc việc phân bố mức lương ở các tỉnh thành và các hoạt động thường diễn ra ở tỉnh thành đó. Ta có thể biết được rằng đối với người dân ở một tỉnh nào đó, với tính chất công việc nổi trội ở đó thì việc sinh hoạt sẽ có gì khác với tỉnh khác.

### **Đánh giá chung:**

- Dựa vào biểu đồ mô tả % thời gian, có thể thấy đa số các công việc mà họ sử dụng chủ yếu để “Xem phim, nghe nhạc, rèn luyện thể dục thể thao, ăn uống, và thực hiện những ngày giải lao”. Chủ yếu là cùng với gia đình, người thân.
- Các tỉnh có mức lương cao như “Nam Định, Bình Định, Tiền Giang”, sử dụng chủ yếu cho việc ăn uống bên ngoài, gặp gỡ đối tác, ...

#### 4. Reference.

- [Vietnam - Time-Use Survey 2022 \(worldbank.org\)](https://worldbank.org)