

**BÁO CÁO HỌC PHẦN
KHAI KHOÁNG DỮ LIỆU**

CÀO VÀ PHÂN TÍCH DỮ LIỆU TRANG UPWORK.COM

**SINH VIÊN THỰC HIỆN
TRẦN ANH KHOA B1913240**

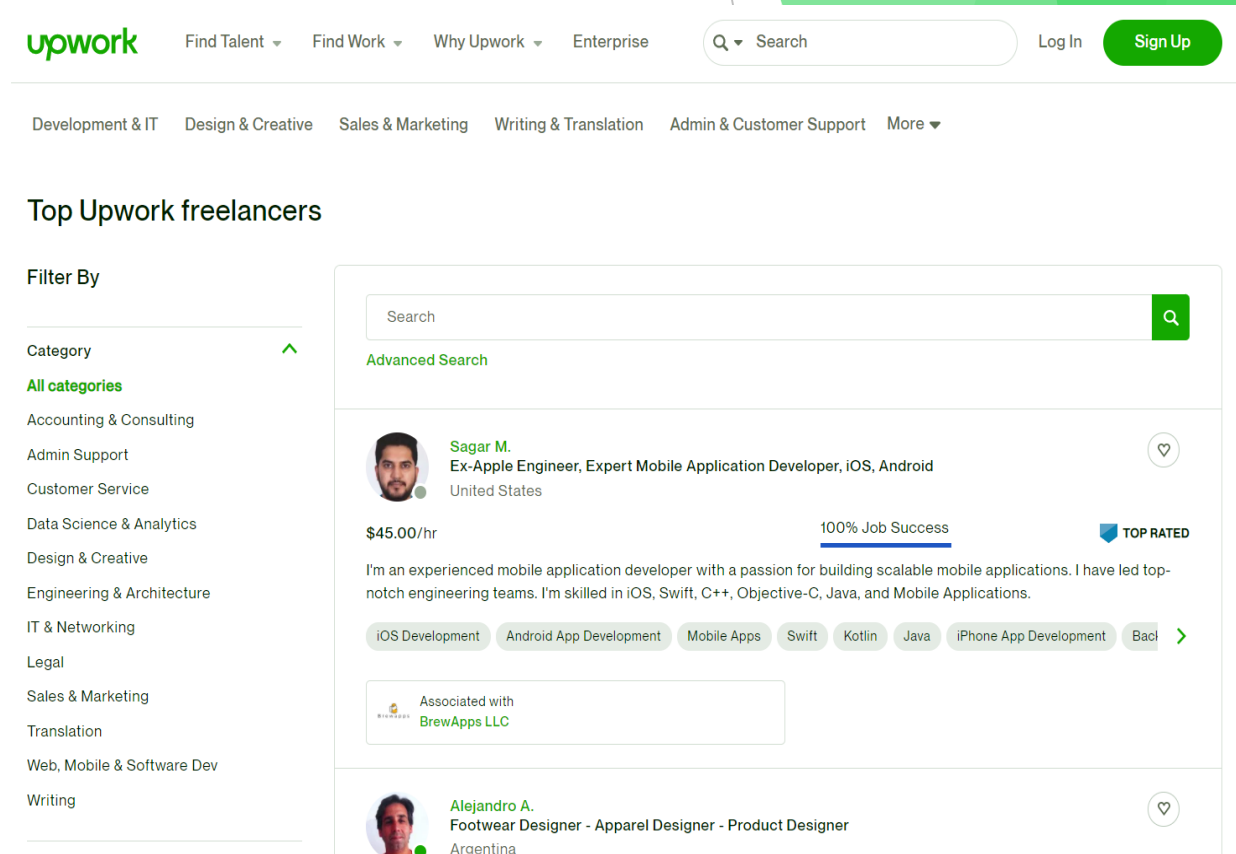
**GIẢNG VIÊN HƯỚNG DẪN
TS. LƯU TIẾN ĐẠO**

1. GIỚI THIỆU
2. PHƯƠNG PHÁP CÀO DỮ LIỆU
3. PHÂN TÍCH DỮ LIỆU
4. KẾT LUẬN

1. GIỚI THIỆU

Mô tả về trang web upwork.com

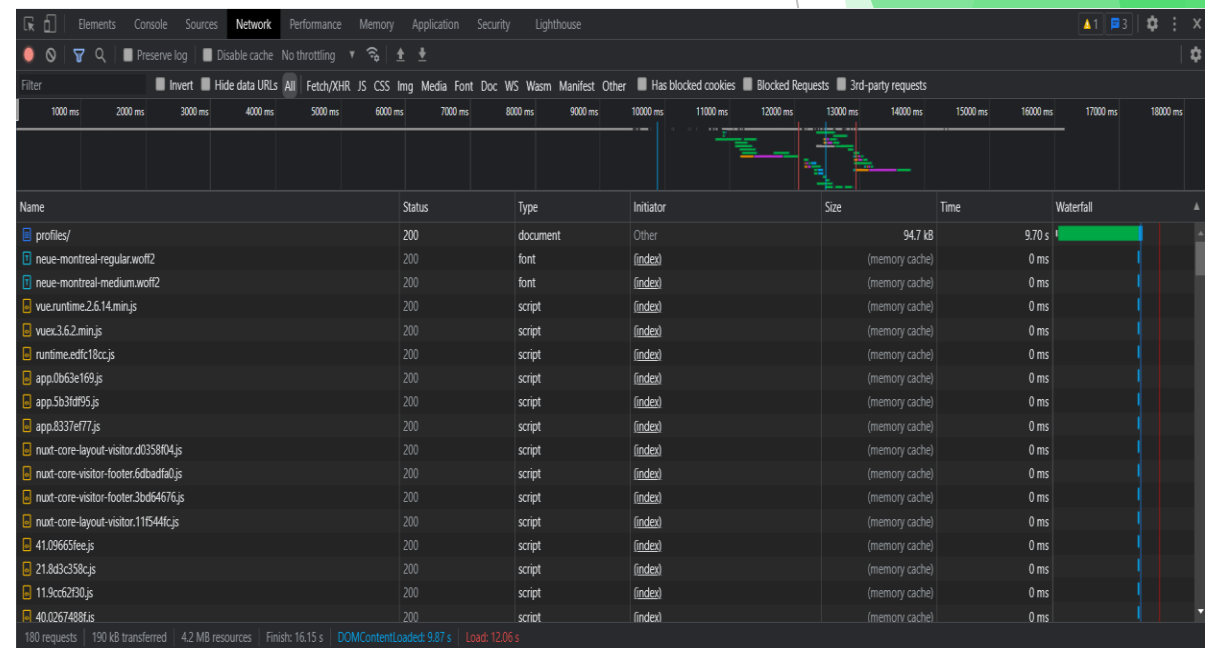
- ▶ Upwork là một nền tảng tìm kiếm việc làm trực tuyến của Mỹ, nơi các doanh nghiệp và cá nhân kết nối với nhau để thỏa thuận nội dung công việc.
- ▶ Trang web này sử dụng API để gửi các thông tin của người tìm việc từ máy chủ về phía máy khách khi được gửi yêu cầu.
- ▶ Do đó để lấy được các thông tin cần thiết chúng ta cần phải tiến hành tìm kiếm API của trang web này.



The screenshot displays the Upwork website interface. At the top, the Upwork logo is on the left, and navigation links for 'Find Talent', 'Find Work', 'Why Upwork', and 'Enterprise' are on the right. A search bar and 'Log In'/'Sign Up' buttons are also present. Below the header, a horizontal menu lists various service categories: 'Development & IT', 'Design & Creative', 'Sales & Marketing', 'Writing & Translation', 'Admin & Customer Support', and 'More'. The main section is titled 'Top Upwork freelancers'. On the left, a 'Filter By' sidebar lists categories such as 'Accounting & Consulting', 'Admin Support', 'Customer Service', 'Data Science & Analytics', 'Design & Creative', 'Engineering & Architecture', 'IT & Networking', 'Legal', 'Sales & Marketing', 'Translation', 'Web, Mobile & Software Dev', and 'Writing'. The main content area shows a search bar, an 'Advanced Search' link, and a list of top freelancers. The first freelancer is Sagar M., an Ex-Apple Engineer and Expert Mobile Application Developer from the United States, with a 100% job success rate and a 'TOP RATED' badge. He is associated with BrewApps LLC. The second freelancer is Alejandro A., a Footwear Designer and Apparel Designer from Argentina.

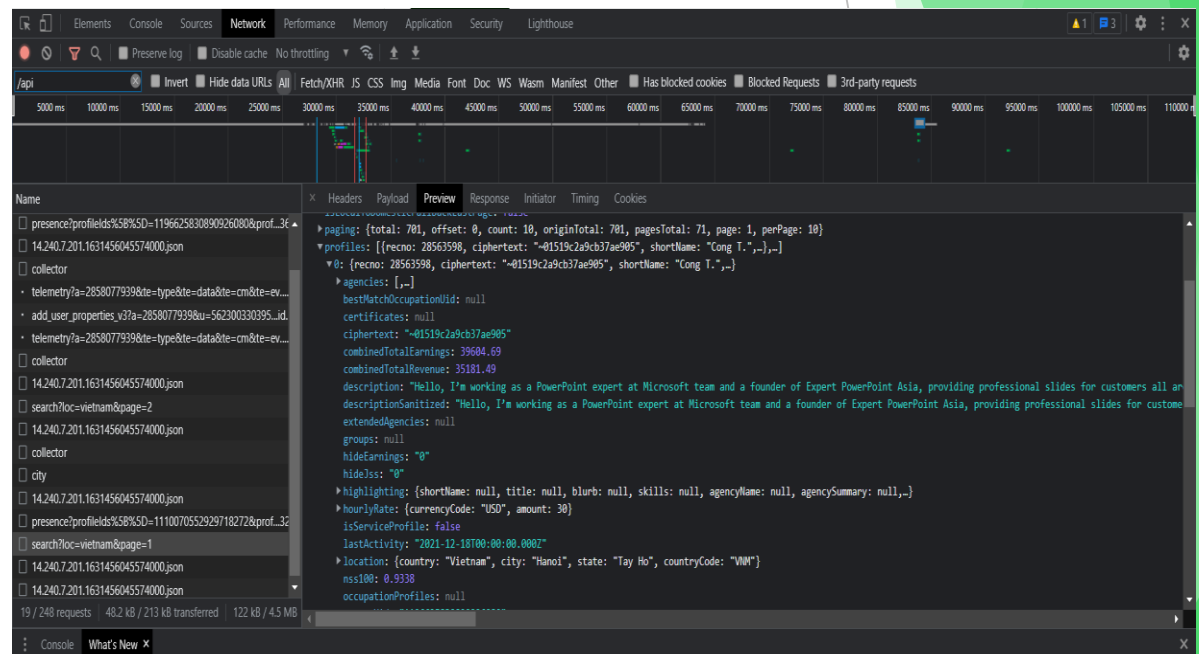
2. PHƯƠNG PHÁP CÀO DỮ LIỆU

- ▶ Để tiến hành dò ra được API của trang web ta sử dụng công cụ devtool của trình duyệt Google Chrome, một công cụ hỗ trợ mạnh mẽ trong việc tương tác giữa các developer với các hệ thống website.
- ▶ Để mở được công cụ devtool ta nhấn phím F12 và chuyển sang tab network để có được giao diện như hình minh họa phía bên phải.
- ▶ Sau khi đã tiến hành các bước ở trên ta tiến hành việc tìm API.



2. PHƯƠNG PHÁP CÀO DỮ LIỆU

- ▶ Sau khi tiến hành dò tìm API, ta thấy rằng API phía bên dưới trả về đầy đủ thông tin của một người tìm việc ở khu vực Việt Nam.
- ▶ <https://www.upwork.com/search/profiles/?loc=vietnam&page=1>
- ▶ Trong đó thuộc tính quan trọng nhất là ciphertext dùng để phân biệt những người dùng với nhau.



2. PHƯƠNG PHÁP CÀO DỮ LIỆU

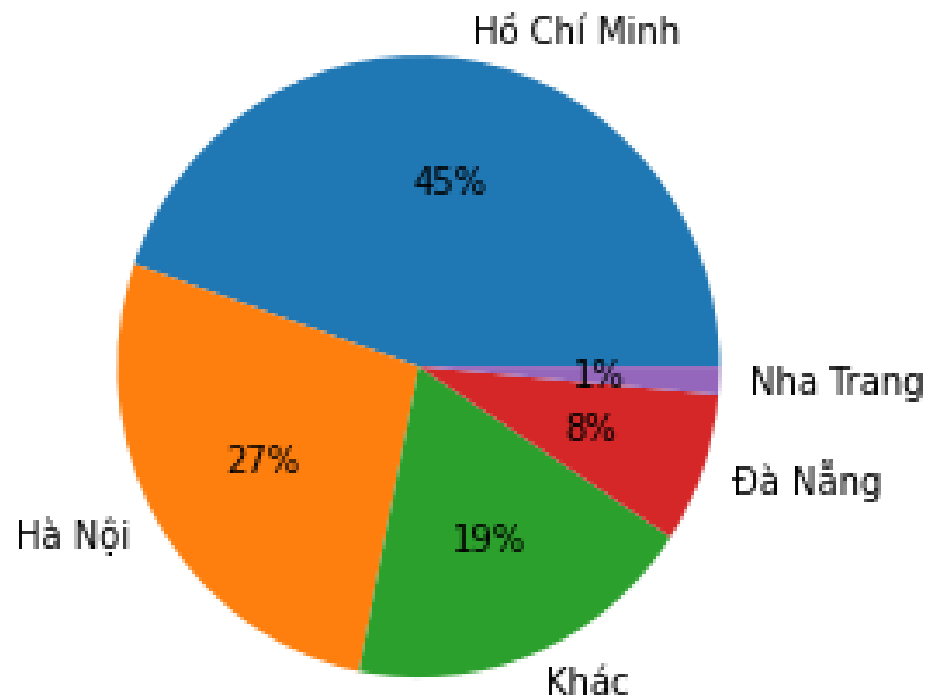
Ảnh minh họa về các trường dữ liệu được API trả về.

```
▼ {profiles: [{recno: 28563598, ciphertext: "~01519c2a9cb37ae905", shortName: "Cong T.",-,-,-}],-}
  ▶ facetCounters: {category_uid: [], pt: [], rate: [], hrs: [], nss: [], last: [], revenue: [], top Rated status: []}
  ▶ facets: {loc: ["vietnam"], page: 1, country: ["vietnam"]}
    fallbackPaging: false
    fallbackProfiles: false
  ▶ flags: {isPortfolioSearch: false}
    isDomesticEnabled: false
    isLocalToDomesticFallbackFirstPage: false
    isLocalToDomesticFallbackLastPage: false
  ▶ paging: {total: 700, offset: 0, count: 10, originTotal: 700, pagesTotal: 70, page: 1, perPage: 10}
  ▶ profiles: [{recno: 28563598, ciphertext: "~01519c2a9cb37ae905", shortName: "Cong T.",-,-,-}]
    ▼ 0: {recno: 28563598, ciphertext: "~01519c2a9cb37ae905", shortName: "Cong T.",-,-}
      ▶ agencies: [-,-]
        bestMatchOccupationUId: null
        certificates: null
        ciphertext: "~01519c2a9cb37ae905"
        combinedTotalEarnings: 39604.69
        combinedTotalRevenue: 35181.49
        description: "Hello, I'm working as a PowerPoint expert at Microsoft team and a founder of Expert PowerPoint Asia, providing professional slides for customers all an
        descriptionSanitized: "Hello, I'm working as a PowerPoint expert at Microsoft team and a founder of Expert PowerPoint Asia, providing professional slides for custome
        extendedAgencies: null
        groups: null
        hideEarnings: "0"
        hideJss: "0"
      ▼ highlighting: {shortName: null, title: null, blurb: null, skills: null, agencyName: null, agencySummary: null,-}
        agencyDescription: null
        agencyName: null
        agencySummary: null
        attributeSkillNames: null
        blurb: null
        shortName: null
        skills: null
        title: null
      ▼ hourlyRate: {currencyCode: "USD", amount: 30}
```

```
▼ hourlyRate: {currencyCode: "USD", amount: 30}
  amount: 30
  currencyCode: "USD"
  isServiceProfile: false
  lastActivity: "2021-12-18T00:00:00.000Z"
  ▶ location: {country: "Vietnam", city: "Hanoi", state: "Tay Ho", countryCode: "VNM"}
    nss100: 0.9338
    occupationProfiles: null
    personUId: "1196625830890926080"
    portfolioItems: null
    portrait: "https://www.upwork.com/profile-portraits/c1JVMsLq3Br5amSMktIfjLAuYSn3ZLsyVCnuQpRGso3k-nDWLwv5gIpBvatd-1U-0"
  ▶ rankInfo: {iv: "+J12CMekacDfiAXAM53tvge=-",-}
    recno: 28563598
    shortName: "Cong T."
    shortNameSanitized: "Cong T."
  ▶ skills: [-,-]
    title: "Expert for PowerPoint | Microsoft"
    titleSanitized: "Expert for PowerPoint | Microsoft"
    topRatedStatus: "eligible"
    topRatedStatusEx: null
    topTalentGroup: "tr_plus_eligible"
    totalFpJobs: 33
    totalHourlyJobs: 9
    totalHoursBilled: 361.66666667
    totalPassedTests: 0
    totalPortfolioItems: 3
    uid: "1196625830890926080"
    url: "/o/profiles/users/~01519c2a9cb37ae905/"
  ▶ 1: {recno: 4351688, ciphertext: "~01d1198187badeb489", shortName: "Bui H.",-,-}
  ▶ 2: {recno: 23829868, ciphertext: "~0198631759486bf6b5", shortName: "Huong N.",-,-}
  ▶ 3: {recno: 27537040, ciphertext: "~01b44d703b4f443c47", shortName: "Anh Nguyen Ngoc D.",-,-}
  ▶ 4: {recno: 24306928, ciphertext: "~01fe95b1afaacf52d4", shortName: "Dat L.", title: "architect",-}
  ▶ 5: {recno: 18409970, ciphertext: "~015bfdb6374910f984", shortName: "Vo N.",-,-}
  ▶ 6: {recno: 30013714, ciphertext: "~01ea599f137414da5b", shortName: "Jenny N.",-,-}
  ▶ 7: {recno: 5920447, ciphertext: "~01b9f0487938dccc2c3", shortName: "Quynh N.",-,-}
```

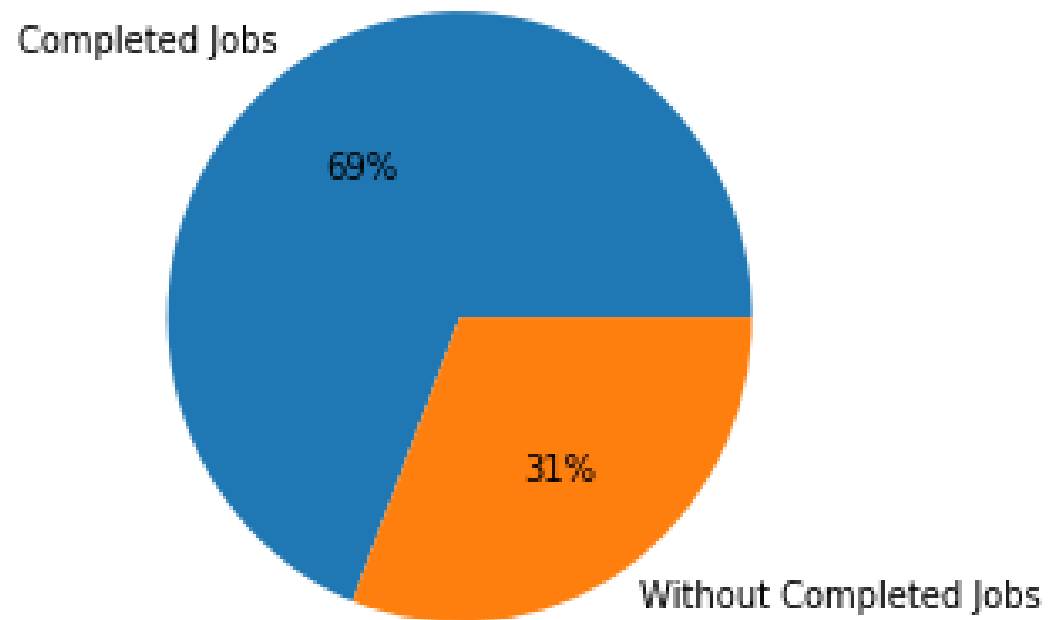
3. PHÂN TÍCH DỮ LIỆU

Tỉ lệ phân bố người tìm việc ở khu vực Việt Nam



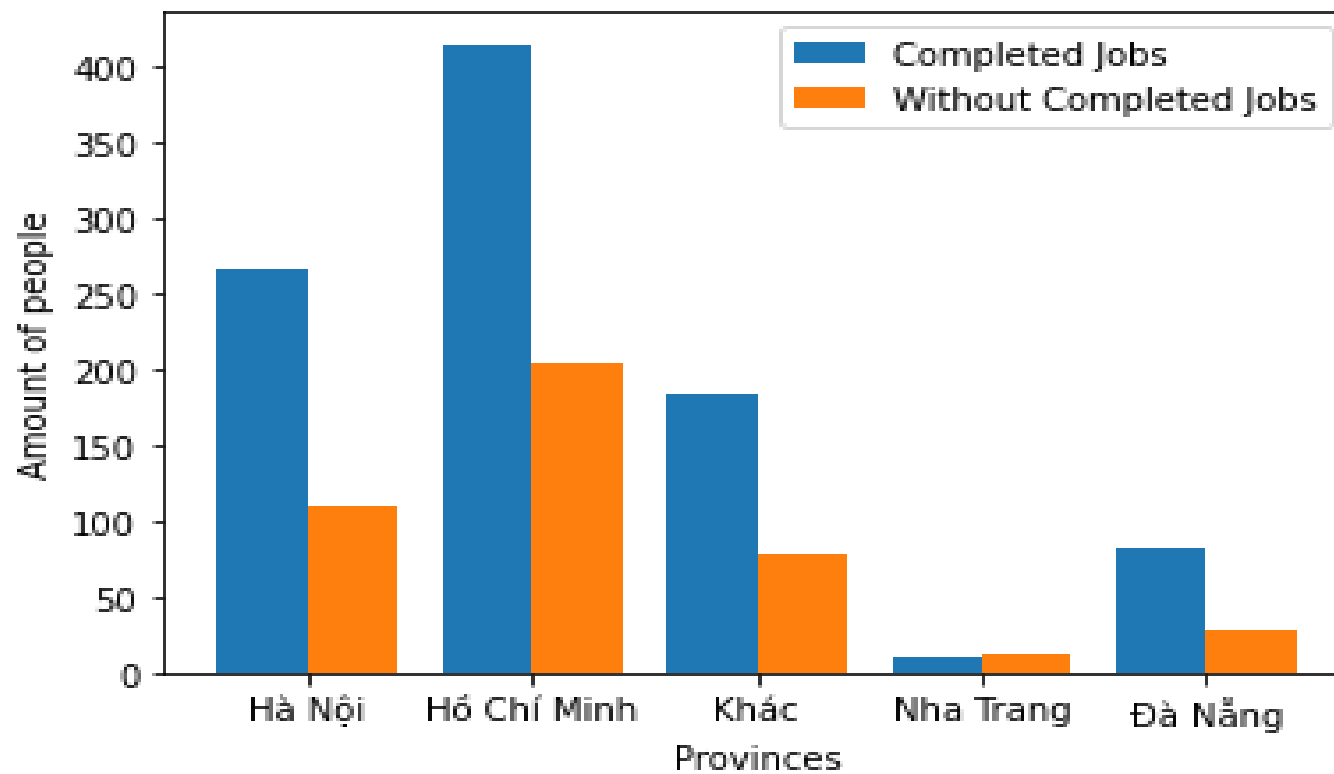
3. PHÂN TÍCH DỮ LIỆU

Tỉ lệ người tìm được việc làm trên cả cả nước



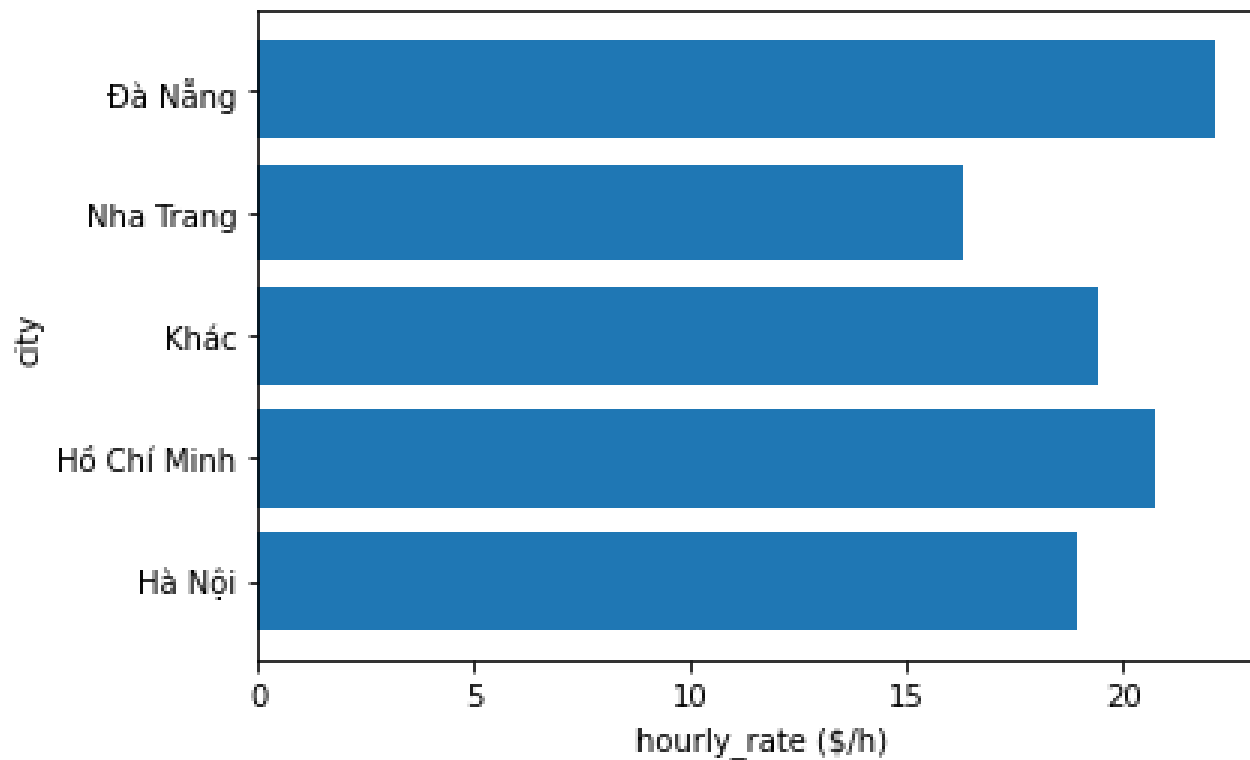
3. PHÂN TÍCH DỮ LIỆU

Tỉ lệ tìm được việc làm giữa các tỉnh thành của Việt Nam



3. PHÂN TÍCH DỮ LIỆU

Thu nhập bình quân đầu người sau một giờ làm việc giữa các tỉnh thành của Việt Nam



4. KẾT LUẬN

- ▶ Xây dựng được chương trình “Cào và phân tích dữ liệu trang web upwork.com” tạo thành tập dữ liệu bao gồm 1.386 dữ liệu về các freelancer ở các tỉnh thành của Việt Nam.
- ▶ Các dữ liệu được phân tích và trực quan hoá bằng các biểu đồ.
- ▶ Đúc kết được hướng phát triển.

4. KẾT LUẬN

Hướng phát triển

- ▶ Cần khai thác thêm các thuộc tính chưa được phân tích. (vd: description, top_skills)
- ▶ Tăng cường tốc độ cào dữ liệu bằng cách tối ưu hoá source code.
- ▶ Sử dụng các thuật toán xử lý ngôn ngữ tự nhiên để gợi ý kỹ năng dựa trên mô tả.

Cảm ơn thầy cô và các bạn đã
lắng nghe