**TRƯỜNG CAO ĐẲNG THỰC HÀNH FPT POLYTECHNIC**

----------🕮----------

A black background with orange and blue text

AI-generated content may be incorrect.

BÁO CÁO DỰ ÁN TỐT NGHIỆP

Chuyên ngành Xử Lý Dữ Liệu

**HÀNH TRÌNH MUA NHÀ CỦA THẾ HỆ TRẺ LÀ MỘT GIẤC MƠ “Bất Khả Thi”**

Các thành viên:

1. Trần Xuân Gia Vũ - PS40819 (Nhóm trưởng)
2. Nguyễn Duy Thành - PS39877
3. Phan Cảnh Nam - PS41473
4. Kiều Minh Bảo - PS30658
5. Lê Văn Giang - PS32984
6. Hồ Thị Ái Trinh – PS41068

GVHD: Thầy Văn Công Khanh

MỤC LỤC

1. Giới thiệu dự án2

1.1 Bối cảnh3

1.2 Mục tiêu dự án3

* 1. Phạm vi dự án3
  2. Bảng kế hoạch dự án3

1. Phân tích yêu cầu khách hàng3

2.1 Phân tích dữ liệu đầu vào3

2.2 Yêu cầu phân tích3

2.3 Câu chuyện dữ liệu 4

* 2.3.1 Đặt vấn đề4
* 2.3.2 Xác định câu chuyện4
* 2.3.3 Xác định rõ đối tượng4
* 2.3.4 Xác định câu chuyện chi tiết4
* 2.3.5 Trình bày dữ liệu4
* 2.3.6 Những điều cần lưu ý4

1. Thu thập và xử lý dữ liệu5

3.1 Danh sách các file thu thập dữ liệu5

4.2 Phân tích luồng crawl dữ liệu5

* 1. Tiền xử lý và làm sạch5
  2. Tính toán biến mới5

1. Phân tích và trực quan hóa5

5.1 Hồ sơ thế hệ trẻ6

5.2 Biến động giá nhà6

* 1. Price-to-Income Ratio6
  2. Chi phí sinh hoạt và tiết kiệm6
  3. Xu hướng vay mua nhà6
  4. Thái độ và hành vi tiêu dùng6
  5. So sánh quốc tế6

1. Xây dựng báo cáo và dashboard6
2. Kết luận và hướng phát triển7

**1. GIỚI THIỆU DỰ ÁN**

**1.1. Bối cảnh**

* Trong bối cảnh giá bất động sản tại các đô thị lớn liên tục tăng cao, việc sở hữu nhà của thế hệ trẻ (20–35 tuổi) đang trở thành thách thức lớn. "Liệu việc mua nhà của thế hệ trẻ hiện nay có phải là một giấc mơ bất khả thi?"
* Từ năm 2010 đến nay, giá bất động sản tại các thành phố lớn như TP.HCM, Hà Nội đã tăng trung bình trên 10% mỗi năm, vượt xa tốc độ tăng thu nhập của người trẻ (<6%/năm). Mức giá trung bình/m² tại quận trung tâm TP.HCM lên tới 70-100 triệu VND, trong khi thu nhập bình quân của người trẻ chỉ đạt 10-15 triệu VND/tháng.

**1.2. Mục tiêu dự án**

* + Phân tích: Đánh giá mức giá trung bình/m² và xu hướng biến động qua các khu vực.
  + So sánh: Xác định khu vực nội thành, ngoại thành và tỉnh lân cận có giá phù hợp nhất.
  + Trực quan: Xây dựng dashboard giúp người trẻ hoặc nhà đầu tư cá nhân ra quyết định.

**1.3. Phạm vi dự án**

* + Vùng thu thập: TP.HCM, Hà Nội,…
  + Dữ liệu: Tiêu đề, giá, diện tích, địa chỉ, ngày đăng, loại hình (căn hộ, nhà riêng).
  + Công cụ: Python (Selenium, Pandas), Power BI, DAX.

**1.4. Bảng kế hoạch dự án**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **TT** | **Hạng mục** | **Bắt đầu** | **Kết thúc** | **Kết quả** |
| 1 | Khảo sát yêu cầu & bối cảnh | 15/05/2025 | 01/06/2025 | Tài liệu yêu cầu chi tiết, phạm vi, nguồn dữ liệu |
| 2 | Thu thập dữ liệu (Crawl) | 01/06/2025 | 05/06/2025 | Dataset thô chứa >50.000 bản ghi |
| 3 | Làm sạch & chuyển đổi dữ liệu | 06/06/2025 | 22/06/2025 | Dataset sạch chuẩn hóa, tính toán price/m2, tách quận/huyện |
| 4 | Mô hình hóa & phân tích DAX | 23/06/2025 | 28/06/2025 | Measures DAX, data model star schema |
| 5 | Trực quan hóa & dashboard | 29/06/2025 | 07/06/2025 | Dashboard interactive với 5 page: Overview, Detail, Heatmap, Trend, Comparison |
| 6 | Viết báo cáo & hoàn thiện | 08/06/2025 | 14/07/2025 | Báo cáo Word hoàn chỉnh |

**2. PHÂN TÍCH YÊU CẦU KHÁCH HÀNG**

**2.1 Phân tích dữ liệu đầu vào**

* Nguồn dữ liệu: Website Batdongsan.com.vn, sử dụng Selenium để tự động thu thập.
* Trường dữ liệu:
  + title: Tiêu đề tin đăng
  + price\_raw: Giá hiển thị kèm đơn vị
  + area\_raw: Diện tích hiển thị kèm đơn vị
  + address\_raw: Địa chỉ đầy đủ
  + date\_posted: Ngày đăng tin
  + type: Loại hình bất động sản

**2.2 Yêu cầu phân tích**

Khách hàng (nhóm nghiên cứu bất động sản) mong muốn:

* + Báo cáo thống kê giá trung bình nhà/m² theo khu vực.
  + So sánh giá giữa các loại hình (căn hộ, nhà riêng).
  + Xu hướng biến động giá theo thời gian.
  + Dashboard tổng quan kết hợp bản đồ.

**2.3. Câu chuyện dữ liệu (Data Storytelling)**

**2.3.1 Đặt vấn đề**

Giá nhà tại các đô thị lớn gia tăng nhanh, vượt xa tốc độ tăng thu nhập.

**2.3.2 Xác định câu chuyện**

* Khu vực có giá nhà hợp lý nhất.
* Xu hướng biến động giá theo thời gian.

**2.3.3 Xác định rõ đối tượng**

Người trẻ (20–35 tuổi), thu nhập trung bình, tại TP.HCM, Hà Nội.

**2.3.4 Xác định câu chuyện chi tiết**

* So sánh giá trung bình/m² giữa các quận.
* Phân tích loại hình: căn hộ vs nhà riêng.

**2.3.5 Trình bày dữ liệu**

* Line chart cho xu hướng giá.
* Bar chart cho so sánh vùng miền.
* Map cho phân bố địa lý.

**2.3.6 Những điều cần lưu ý**

* Dữ liệu có thể thiếu hoặc lỗi định dạng.
* Cần loại bỏ bản ghi trùng và giá trị ngoại lai.

**3. THU THẬP & XỬ LÝ DỮ LIỆU**

**4.1. Danh sách file thu thập dữ liệu**

|  |  |
| --- | --- |
| **Tên file** | **Chức năng** |
| Crawldata.py | Crawl dữ liệu nhà riêng tại TP.HCM |
| CrawldataHN.py | Crawl dữ liệu nhà riêng tại Hà Nội |
| CrawldataHN\_APT.py | Crawl dữ liệu căn hộ chung cư Hà Nội |
| Crawldata\_APT.PY | Crawl dữ liệu căn hộ TP.HCM |
| Crawdatajob.py | Crawl dữ liệu tuyển dụng bất động sản |
| careerviet\_crawler.py | Crawl trang careerviet.vn tổng quát |
| careerviet\_crawler\_job.py | Crawl thông tin cụ thể từng công việc |
| data\_cleaner.py | Làm sạch dữ liệu: chuẩn hóa, tính toán, tách trường |

A screenshot of a computer

AI-generated content may be incorrect.

**4.2 Phân tích luồng crawl dữ liệu**

Trong giai đoạn này, sử dụng tập lệnh Python với Selenium để tự động truy cập các trang Batdongsan.com.vn:

*Tên file mã nguồn: Crawldata.py  
Nguồn dữ liệu:* [*https://batdongsan.com.vn*](https://batdongsan.com.vn) *Số lượng trang cần crawl: 935 trang  
Kết quả đầu ra: File Excel nha\_dat\_batdongsan.xlsx chứa thông tin nhà đất tại TP.HCM*

**4.2.1 Crawldata.py – Crawl nhà riêng TP.HCM**

A screenshot of a computer

AI-generated content may be incorrect.

***- Tạo trình duyệt Chrome chạy ở chế độ ẩn (headless)***

* Cấu hình trình duyệt với các tùy chọn:
  + Chạy ẩn (không hiển thị cửa sổ trình duyệt).
  + Tắt thông báo trình duyệt bị kiểm soát tự động (AutomationControlled).
  + Thiết lập kích thước và user-agent giống người thật.
* Chống bị phát hiện bot bằng thư viện selenium-stealth.

def create\_browser():  
 chrome\_options = Options()  
 chrome\_options.add\_argument("--headless=new")  
 chrome\_options.add\_argument("--disable-blink-features=AutomationControlled")  
 chrome\_options.add\_argument("--window-size=1920,1080")  
 chrome\_options.add\_argument("--disable-gpu")  
 chrome\_options.add\_argument("user-agent=Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/124.0.0.0 Safari/537.36")  
  
 # 👉 CHỈNH lại đường dẫn chromedriver.exe ở đây  
 service = Service(r"D:\DATT\_XLDL\DATN\chromedriver-win64\chromedriver.exe")  
 #service = Service(r"E:\chromedriver-win64\chromedriver-win64\chromedriver.exe") D:\DATT\_XLDL\DATN  
 browser = webdriver.Chrome(service=service, options=chrome\_options)  
  
 stealth(browser,  
 languages=["vi-VN", "vi"],  
 vendor="Google Inc.",  
 platform="Win32",  
 webgl\_vendor="Intel Inc.",  
 renderer="Intel Iris OpenGL Engine",  
 fix\_hairline=True,  
 )  
 return browser

***- Crawl dữ liệu từ website Batdongsan.com.vn***

* **Tạo vòng lặp duyệt qua các trang**:
  + Duyệt từ p1 đến p2 (đang chạy thử 2 trang đầu).
  + Với mỗi trang:
    - Mở URL.
    - Chờ ngẫu nhiên từ 6–9 giây cho trang tải xong.
    - Cuộn trang như người dùng thật.
* **Trích xuất thông tin từ từng tin rao:**
  + Tìm tất cả các thẻ div chứa thông tin nhà đất (re\_\_card-info-content).
  + Với mỗi tin:
    - Trích xuất:
      * Tên dự án|Giá|Diện tích|Vị trí|Phòng ngủ|Nhà vệ sinh|ngày đăng
    - Nếu có lỗi thì bỏ qua tin đó.

# Crawl 20 trang  
for page in range(1, 659):  
 print(f"📄 Đang xử lý trang {page}...")  
  
 url = f"https://batdongsan.com.vn/ban-can-ho-chung-cu-tp-hcm/p{page}"  
 browser = create\_browser()  
 browser.get(url)  
 time.sleep(random.uniform(6, 9))  
  
 # Scroll như người dùng  
 browser.execute\_script("window.scrollTo(0, document.body.scrollHeight \* 0.5);")  
 time.sleep(random.uniform(1, 2))  
 browser.execute\_script("window.scrollTo(0, document.body.scrollHeight);")  
 time.sleep(random.uniform(2, 4))  
  
 cards = browser.find\_elements(By.XPATH, '//div[@class="re\_\_card-info-content"]')

# Test 2 trang  
  
for page in range(1, 3): # test 2 trang  
 print(f"📄 Đang xử lý trang {page}...")  
  
 url = f"https://batdongsan.com.vn/ban-nha-rieng-tp-hcm/p{page}"  
 browser = create\_browser()  
 browser.get(url)  
 time.sleep(random.uniform(6, 9))  
  
 browser.execute\_script("window.scrollTo(0, document.body.scrollHeight \* 0.5);")  
 time.sleep(random.uniform(1, 2))  
 browser.execute\_script("window.scrollTo(0, document.body.scrollHeight);")  
 time.sleep(random.uniform(2, 4))  
  
 cards = browser.find\_elements(By.XPATH, '//div[@class="re\_\_card-info-content"]')

📄 Đang xử lý trang 1...  
📄 Đang xử lý trang 2...

for ele in cards:  
 try:  
 name\_project = ele.find\_element(By.XPATH, './/span[contains(@class, "js\_\_card-title")]').text  
 price\_project = ele.find\_element(By.XPATH, './/span[contains(@class, "re\_\_card-config-price")]').text  
 area\_project = ele.find\_element(By.XPATH, './/span[contains(@class, "re\_\_card-config-area")]').text   
 # Vị trí  
 try:  
 location = ele.find\_element(By.XPATH, './/div[@class="re\_\_card-location"]/span[last()]').text  
 except:  
 location = ""  
 # Phòng ngủ  
 try:  
 bedroom = ele.find\_element(By.XPATH, './/span[contains(@class, "re\_\_card-config-bedroom")]').get\_attribute("aria-label")  
 except:  
 bedroom = ""  
 # WC  
 try:  
 toilet = ele.find\_element(By.XPATH, './/span[contains(@class, "re\_\_card-config-toilet")]').get\_attribute("aria-label")  
 except:  
 toilet = ""  
 # Ngày đăng (từ `aria-label`)  
 try:  
 posted\_date = ele.find\_element(By.XPATH, './/span[contains(@class, "re\_\_card-published-info-published-at")]').get\_attribute("aria-label")  
 except:  
 posted\_date = ""  
  
 data.append({  
 "Tên dự án": name\_project,  
 "Giá": price\_project,  
 "Diện tích": area\_project,  
 "Vị trí": location,  
 "Phòng ngủ": bedroom,  
 "Nhà vệ sinh": toilet,  
 "Ngày đăng": posted\_date  
 })  
  
 except Exception as e:  
 print("❌ Bỏ qua tin bị lỗi:", e)

***- Lưu dữ liệu vào file Excel***

# Lưu Excel  
df = pd.DataFrame(data)  
df.columns.name = "Thông tin Chung cư"  
df.to\_excel("Chung\_Cu\_batdongsan.xlsx", index=False)  
print("✅ Đã crawl xong trang và lưu vào file Excel.")

**4.2.2 CrawldataHN.py – Crawl dữ liệu nhà riêng tại Hà Nội**

*Tên file mã nguồn: CrawldataHN.py  
Nguồn dữ liệu:* [*https://batdongsan.com.vn*](https://batdongsan.com.vn) *Số lượng trang cần crawl: 1.022 trang  
Kết quả đầu ra: File Excel nha\_dat\_batdongsan\_HN.xlsx chứa thông tin nhà đất tại Hà Nội*

***- Tạo trình duyệt Chrome ẩn (headless)***

A screenshot of a computer

AI-generated content may be incorrect.

* *Tính năng nổi bật*: Sử dụng selenium-stealth để trình duyệt không bị hệ thống phát hiện là bot tự động.

***- Duyệt qua từng trang chứa danh sách tin rao nhà đất***

Sử dụng vòng lặp for page in range(1, 1023) để đi qua tất cả các trang từ 1 đến 1022 trên trang:



Giả lập thao tác người dùng:

* Chờ trang tải trong 6–9 giây.
* Cuộn xuống giữa và đáy trang giống hành vi người dùng thật.
* Điều này giúp các phần tử “lazy load” (chưa tải ngay ban đầu) có cơ hội hiển thị.

### **- Trích xuất dữ liệu từ mỗi tin rao**

Dùng XPATH để tìm từng phần tử chứa thông tin:

|  |  |
| --- | --- |
| **Trường dữ liệu** | **Mô tả** |
| Tên dự án | Tên của căn nhà hoặc tiêu đề tin rao |
| Giá | Giá bán nhà |
| Diện tích | Diện tích căn nhà |
| Vị trí | Quận, phường nơi nhà được rao bán |
| Phòng ngủ | Số lượng phòng ngủ |
| Nhà vệ sinh | Số lượng nhà vệ sinh |
| Ngày đăng | Thời gian tin được đăng |

A screenshot of a computer code

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

***- Lưu dữ liệu ra file Excel***

Sau khi duyệt qua toàn bộ trang và thu thập dữ liệu thành công, dữ liệu sẽ được lưu vào file Excel như sau:

A screenshot of a computer code

AI-generated content may be incorrect.

**4.2.3 CrawldataHN\_APT.py – Crawl dữ liệu căn hộ chung cư tại Hà Nội**

**4.2.4 Crawldata\_APT.py – Crawl dữ liệu căn hộ TP.HCM**

**4.2.5 careerviet\_crawler.py – Crawl trang careeviet.vn tổng quát**

*Tên file mã nguồn: careerviet\_crawler.py  
Nguồn dữ liệu: https://www.careerlink.vn/tim-viec-lam-tai/ho-chi-minh/HCM*

***- Chi tiết các bước xử lý:***

A screen shot of a computer

AI-generated content may be incorrect.

* *Bước 1: Khởi tạo trình duyệt*
* Sử dụng Chrome headless để chạy ẩn.
* Thêm proxy để ẩn IP.
* Dùng selenium-stealth để tránh bị phát hiện là bot.



* *Bước 2: Giải captcha tự động*
* Một số trang yêu cầu vượt captcha Google reCAPTCHA.
* Dùng API của [2Captcha.com](https://2captcha.com) để tự động giải.
* Inject mã captcha vào HTML và kích hoạt lại form.

*Khó khăn khi gặp phải cào dữ liệu công việc làm trên web*

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer screen

AI-generated content may be incorrect.

*Tiếp tục xử dụng phương án fix proxy và sử dụng proxy động để chạy nhưng 1 thời gian lại lỗi địa chỉ máy chủ*

A screenshot of a computer program

AI-generated content may be incorrect.

A close-up of a number

AI-generated content may be incorrect.

*Cuối cùng phải sử dụng kèm thêm phương án v2 capcha để tự động xác minh robot tránh bức tường của web + kèm chạy đổi proxy liên tục để cào dữ liệu thành công.*

A screenshot of a computer program

AI-generated content may be incorrect.

**4.3 Làm sạch dữ liệu**

**A screenshot of a computer

AI-generated content may be incorrect.**

**A screenshot of a computer

AI-generated content may be incorrect.**

**4. PHÂN TÍCH & TRỰC QUAN HÓA**

* **5.1. Hồ sơ thế hệ trẻ**
* Phân bố độ tuổi, giới tính, nghề nghiệp.
* Thu nhập trung bình theo nhóm.
* **5.2. Biến động giá nhà**
* So sánh TP.HCM vs Hà Nội (2010–2024).
* Biểu đồ đường.
* **5.3. Price‑to‑Income Ratio**
* Diễn biến PIR qua các năm.
* Heatmap thể hiện chênh lệch.
* **5.4. Chi phí sinh hoạt & tiết kiệm**
* Cơ cấu chi phí: nhà ở, ăn uống, di chuyển.
* Boxplot & stacked bar.
* **5.5. Xu hướng vay mua nhà**
* Lãi suất vay trung bình.
* Area chart chuỗi thời gian.
* **5.6. Thái độ & hành vi tiêu dùng**
* Kết quả Likert scale.
* Wordcloud lý do ưu tiên.
* **5.7. So sánh quốc tế**
* PIR tại Việt Nam, Thái Lan, Singapore, Mỹ.
* Radar chart.

**5. XÂY DỰNG BÁO CÁO & DASHBOARD**

1. **Dashboard Tableau**: Trang tổng quan, phân tích chi tiết, mô phỏng dự báo.
2. **Slide PowerPoint**: Dàn bài rõ ràng, mỗi thành viên phụ trách một phần.

**6. KẾT LUẬN & HƯỚNG PHÁT TRIỂN**

**Kết luận chính**:

* **Phân tích** cho thấy giá trung bình/m² tại TP.HCM (~30 triệu VND) cao gấp 2–3 lần so với khu vực Hà Nội (~12 triệu VND).
* Khu vực ngoại thành như Bình Chánh, Hóc Môn có giá mềm hơn (10–15 triệu VND).
* **Kết luận**: Giấc mơ sở hữu nhà trong nội thành với thu nhập trung bình thực sự khó tiếp cận; người trẻ nên cân nhắc chọn khu vực ven đô hoặc các tỉnh lân cận.

**7.1 Ưu điểm**

* Quy trình tự động hóa hiệu quả, dễ bảo trì.
* Báo cáo trực quan sinh động, hỗ trợ quyết định đầu tư.

**7.2 Khó khăn**

* Thời gian crawl dài, cần tối ưu.
* Dữ liệu trang web không đồng nhất, nhiều lỗi định dạng.

**7.3 Hướng phát triển**

* Triển khai API crawl tự động hàng tuần.
* Mở rộng tập dữ liệu, kết hợp dữ liệu thu nhập, lãi suất.
* Phân tích dự báo giá trong tương lai.