BÁO CÁO KỸ THUẬT

DỰ ÁN PHÂN TÍCH DỮ LIỆU AIRBNB

A. THÔNG TIN CƠ BẢN

- **Tên dự án:** Inside Airbnb - Phân tích dữ liệu đặt phòng khách sạn theo thời gian / thành phố.
- **Nhóm thực hiện:** Nhóm 2526-LTXLDL-Project4-AIT2006-2-4.4.
- **Thành viên:**

1. Lê Việt Phú - 24022426.
2. Hoàng Huy Hoàng - 24022336.
3. Trần Trung Hiếu - 24022330.

- **Các thành phố được phân tích:**

1. **Brussels**, Belgium.
2. **Berlin**, Germany.
3. **Paris**, France.

- **Danh sách các snapshot sử dụng:**

1. **Brussels**, Belgium: 22/12/2024, 16/03/2025, 21/06/2025.
2. **Berlin**, Germany: 21/12/2024, 15/03/2025, 20/06/2025, 23/09/2025.
3. **Paris**, France: 06/12/2024, 03/03/2025, 06/06/2025, 12/09/2025.

- **Nguồn dữ liệu:**

1. Trang tải dữ liệu chính thức: https://insideairbnb.com/get-the-data/
2. Trang tải tài liệu khám phá: https://insideairbnb.com/explore

- **Ngày nộp:**
- **Commit hash:**

B. GIỚI THIỆU & DỮ LIỆU

B.1 Giới thiệu

- Dự án sử dụng dữ liệu từ **Inside Airbnb** - một nền tảng cung cấp các snapshots dữ liệu Airbnb theo thành phố và theo thời điểm thu thập (collection date).
- **Dữ liệu sau khi được xử lý của các thành phố sẽ được dùng để phản ánh:**

1. Nguồn cung và giá cho thuê biến động như thế nào theo thời gian?
2. Khu vực nào đắt đỏ nhất và khu vực nào tập trung nhiều phòng nhất?
3. Tỷ lệ lấp đầy và hiệu suất doanh thu ra sao?
4. Mức độ chuyên nghiệp hóa của chủ nhà.
5. Đánh giá của khách hàng.
6. Thông tin về vị trí và khu vực.

- **Mục tiêu:** So sánh sự biến động của thị trường Airbnb theo **thời gian** và **không gian** giữa 3 thành phố: **Brussels**, **Berlin** và **Paris**.

B.2 Dữ liệu

- **Phạm vi dữ liệu:** Dự án sử dụng dữ liệu được thu thập của mỗi thành phố. Mỗi snapshot bao gồm các tệp: `listings.csv.gz`, `calendar.csv.gz`, `reviews.csv.gz`, `neighbourhoods.csv`.
- **Cấu trúc các tệp dữ liệu đầu vào:**

1. `listings.csv.gz`: Thông tin chỗ ở và giá thuê.
2. `calendar.csv.gz`: Tính sẵn sàng và giá theo ngày.
3. `reviews.csv.gz`: Đánh giá của khách hàng.
4. `neighbourhoods.csv`: Ranh giới khu vực GeoJSON/CSV.

- **Các trường chính được sử dụng từ các tệp dữ liệu đầu vào:**

1. `listings.csv.gz`: `id`, `host_id`, `host_name`, `neighbourhood_cleansed`, `latitude`, `longitude`, `room_type`, `price`, `availability_90`, `host_since`.
2. `calendar.csv.gz`: `date`, `available`, `price`, `listing_id`.
3. `reviews.csv.gz`: `date`, `listing_id`, `comments`.
4. `neighbourhoods.csv`: `neighbourhood`.

C. QUY TRÌNH QA & LÀM SẠCH DỮ LIỆU

- **Chuẩn hóa giá:**

1. Tách giá trị số từ chuỗi (loại bỏ `$`, `,`, ký tự không phải số).
2. Chuyển về dạng số thực `price_numeric`, bản ghi không thể chuyển sẽ thành `NaN`.
3. Áp dụng cho cả `listings.csv.gz` và `calendar.csv.gz`.

- **Loại snapshot lỗi:**

1. Một số snapshot cuối bị lỗi toàn giá = 0 hoặc không parse được.
2. Snapshot không hợp lệ sẽ bị loại khỏi pipeline để tránh sai biểu đồ.

- **Kiểm tra giá bất thường - QA001:**

1. Điều kiện: `price_numeric ≤ 0`, hoặc `price_numeric` bị `NaN` sau parse.
2. Xử lý: Không xoá bản ghi, chỉ **gắn cờ** (`qa_flag_price_zero = True`).

- **Chuẩn hóa thời gian và tọa độ:**

1. Chuẩn hoá định dạng ngày (datetime): `host_since`, `calendar.date`, `reviews.date`.
2. Chuyển `latitude`, `longitude` sang số (`float`), xử lý lỗi bằng `NaN`.

- **Kiểm tra tọa độ ngoài ranh giới thành phố - QA002:**

1. Áp dụng bounding box cho từng thành phố.
2. Điều kiện: Các bản ghi nằm ngoài vùng quy định của thành phố.
3. Xử lý: Không xoá bản ghi, chỉ **gắn cờ** (`qa_flag_out_of_city = True`).

- **Kiểm tra ID trùng lặp - QA003:**

1. Điều kiện: Phát hiện bản ghi có `id` trùng nhau.
2. Xử lý: Giữ bản đầu tiên, xoá các bản ghi còn lại, số lượng trùng được ghi lại trong báo cáo QA.

- **Làm sạch dữ liệu khu vực - QA004:**

1. Xóa cột thừa `neighbourhood_group`.
2. Lấy dach sách khu vực hợp lệ từ file `neighbourhoods.csv`.
3. Điều kiện: Bản ghi có `neighbourhood_cleansed` không nằm trong danh sách khu vực chuẩn.
4. Xử lý: Không xóa bản ghi, chỉ **gắn cờ** (`qa_flag_invalid_neigh = True`).

- **Làm sạch dữ liệu đánh giá - QA005:**

1. Chuẩn hóa thời gian (`date`).
2. Điều kiện: Các review không có nội dung (`comments = NaN`), hoặc không tồn tại `listing_id` trong các bản ghi đã sạch.
3. Xử lý: Các review sẽ bị xóa bỏ.

- **Lưu dữ liệu sạch:**

1. Lưu lại 4 file đã làm sạch trong thư mục: `processed/<city_name>/<snapshot_date>/`
2. Chú thích: `<city_name>` - tên thành phố, `<snapshot_date>` - ngày thu thập dữ liệu.

- **Xuất báo cáo QA tổng hợp:**

1. Xuất file: `reports/qa_summary_<city_name>.csv`.
2. Ghi lại số bản ghi bị ảnh hưởng bởi từng quy tắc QA.
3. Chú thích: `<city_name>` - tên thành phố.

- **Tóm tắt kết quả QA:**

City	Snapshot Date	Rule ID	Records Affected	Handling Decision
Brussels	16/03/2025	QA001_price_zero	982	Gắn cờ
Brussels	16/03/2025	QA002_coords_out_of_bounds	36	Gắn cờ
Brussels	16/03/2025	QA003_duplicate_ids	0	Xoá dòng trùng
Brussels	16/03/2025	QA004_invalid_neighbourhood	0	Gắn cờ
Brussels	16/03/2025	QA005_orphaned_or_empty_reviews	23	Xoá bỏ
Brussels	21/06/2025	QA001_price_zero	1,020	Gắn cờ
Brussels	21/06/2025	QA002_coords_out_of_bounds	39	Gắn cờ
Brussels	21/06/2025	QA003_duplicate_ids	0	Xoá dòng trùng
Brussels	21/06/2025	QA004_invalid_neighbourhood	0	Gắn cờ
Brussels	21/06/2025	QA005_orphaned_or_empty_reviews	23	Xoá bỏ
Brussels	22/12/2024	QA001_price_zero	848	Gắn cờ
Brussels	22/12/2024	QA002_coords_out_of_bounds	38	Gắn cờ
Brussels	22/12/2024	QA003_duplicate_ids	0	Xoá dòng trùng
Brussels	22/12/2024	QA004_invalid_neighbourhood	0	Gắn cờ
Brussels	22/12/2024	QA005_orphaned_or_empty_reviews	22	Xoá bỏ
Berlin	15/03/2025	QA001_price_zero	5,047	Gắn cờ
Berlin	15/03/2025	QA002_coords_out_of_bounds	0	Gắn cờ
Berlin	15/03/2025	QA003_duplicate_ids	0	Xoá dòng trùng
Berlin	15/03/2025	QA004_invalid_neighbourhood	0	Gắn cờ
Berlin	15/03/2025	QA005_orphaned_or_empty_reviews	40	Xoá bỏ

City	Snapshot Date	Rule ID	Records Affected	Handling Decision
Berlin	20/06/2025	QA001_price_zero	5,004	Gắn cờ
Berlin	20/06/2025	QA002_coords_out_of_bounds	0	Gắn cờ
Berlin	20/06/2025	QA003_duplicate_ids	0	Xoá dòng trùng
Berlin	20/06/2025	QA004_invalid_neighbourhood	0	Gắn cờ
Berlin	20/06/2025	QA005_orphaned_or_empty_reviews	46	Xoá bỏ
Berlin	21/12/2024	QA001_price_zero	4,994	Gắn cờ
Berlin	21/12/2024	QA002_coords_out_of_bounds	0	Gắn cờ
Berlin	21/12/2024	QA003_duplicate_ids	0	Xoá dòng trùng
Berlin	21/12/2024	QA004_invalid_neighbourhood	0	Gắn cờ
Berlin	21/12/2024	QA005_orphaned_or_empty_reviews	39	Xoá bỏ
Berlin	23/09/2025	QA001_price_zero	5,010	Gắn cờ
Berlin	23/09/2025	QA002_coords_out_of_bounds	0	Gắn cờ
Berlin	23/09/2025	QA003_duplicate_ids	0	Xoá dòng trùng
Berlin	23/09/2025	QA004_invalid_neighbourhood	0	Gắn cờ
Berlin	23/09/2025	QA005_orphaned_or_empty_reviews	50	Xoá bỏ
Paris	03/03/2025	QA001_price_zero	30,409	Gắn cờ
Paris	03/03/2025	QA002_coords_out_of_bounds	78	Gắn cờ
Paris	03/03/2025	QA003_duplicate_ids	0	Xoá dòng trùng
Paris	03/03/2025	QA004_invalid_neighbourhood	0	Gắn cờ
Paris	03/03/2025	QA005_orphaned_or_empty_reviews	125	Xoá bỏ
Paris	06/12/2024	QA001_price_zero	30,938	Gắn cờ
Paris	06/12/2024	QA002_coords_out_of_bounds	63	Gắn cờ
Paris	06/12/2024	QA003_duplicate_ids	0	Xoá dòng trùng
Paris	06/12/2024	QA004_invalid_neighbourhood	0	Gắn cờ
Paris	06/12/2024	QA005_orphaned_or_empty_reviews	134	Xoá bỏ
Paris	06/06/2025	QA001_price_zero	30,092	Gắn cờ
Paris	06/06/2025	QA002_coords_out_of_bounds	74	Gắn cờ
Paris	06/06/2025	QA003_duplicate_ids	0	Xoá dòng trùng

City	Snapshot Date	Rule ID	Records Affected	Handling Decision
Paris	06/06/2025	QA004_invalid_neighbourhood	0	Gắn cờ
Paris	06/06/2025	QA005_orphaned_or_empty_reviews	140	Xoá bỏ
Paris	12/09/2025	QA001_price_zero	81,853	Gắn cờ
Paris	12/09/2025	QA002_coords_out_of_bounds	71	Gắn cờ
Paris	12/09/2025	QA003_duplicate_ids	0	Xoá dòng trùng
Paris	12/09/2025	QA004_invalid_neighbourhood	0	Gắn cờ
Paris	12/09/2025	QA005_orphaned_or_empty_reviews	140	Xoá bỏ

D. QUY TRÌNH TÍNH TOÁN KPI

- **Tổng số listing hợp lệ (Total Listings):**

1. Định nghĩa: Số lượng listings không bị gắn cờ QA (không lỗi giá và không lệch tọa độ).
2. Công thức:

$$\text{TotalListings} = |\text{Listing}_{\text{valid}}|$$

3. Ý nghĩa: Đo lường quy mô thị trường thực tế sau QA.

- **Tỷ lệ chủ nhà có nhiều hơn một căn (Multi-Host Rate):**

1. Định nghĩa: Tỷ lệ listings thuộc về các host có nhiều hơn 2 bất động sản cho thuê.
2. Công thức:

$$\text{MultiHostRate} = \frac{\sum \text{Host}_{\text{có } \geq 2 \text{ listing}}}{\text{TotalListings}} \times 100$$

3. Ý nghĩa: Dựa vào tỷ lệ để biết được thị trường đang thiên hướng chuyên nghiệp hay cá nhân.

- **Giá thuê trung vị toàn thành phố (Median Price):**

1. Định nghĩa: Giá cho thuê trung vị từ cột `price_numeric`.
2. Công thức:

$$\text{MedianPrice} = \text{median}(\text{price\_numeric})$$

3. Ý nghĩa: Đại diện cho mặt bằng giá thực, không bị nhiễu bởi outliers.

- **Số ngày trống trong 3 tháng tới (Median Availability 90):**

1. Định nghĩa: Giá trị trung vị của cột `availability_90`.
2. Công thức:

$$\text{MedianAvailability90} = \text{median}(\text{availability\_90})$$

3. Ý nghĩa: Dựa vào giá trị để biết được nhu cầu của khách hàng.

- **Tỷ lệ lấp đầy theo ngày (Daily Occupancy Rate):**

1. Định nghĩa: Dựa trên `calendar.csv.gz` để tính tỷ lệ lấp đầy theo ngày.
2. Công thức:

$$\text{OccRate}(d) = \frac{\text{số listings booked trong ngày d}}{\text{tổng listing hợp lệ}}$$

3. Ý nghĩa: Tạo chuỗi thời gian để phân tích theo mùa.

- **Tỷ lệ loại phòng (Room Type Distribution):**

1. Định nghĩa: Tỷ lệ từng loại phòng (Entire home/apt, Private room, Shared room, Hotel room).
2. Công thức:

$$\text{Pct}(\text{room\_type}) = \frac{\text{count}(\text{room\_type})}{\text{TotalListings}} \times 100$$

3. Ý nghĩa: Giúp mô tả cấu trúc lưu trú của mỗi thành phố.

- **Số lượng listings và giá thuê trung vị theo khu vực của từng thành phố (Neighbourhood Statistics):**

1. Định nghĩa: Dùng dữ liệu từ cột `neighbourhood_cleansed` (nếu có), hoặc `neighbourhood` (nếu không) để tính số lượng listings và giá thuê trung vị theo khu vực của từng thành phố.
2. Công thức: Giống công thức tính số lượng listing hợp lệ và giá thuê trung vị.
3. Ý nghĩa: Biết được quy mô thị trường và giá thuê theo từng khu vực của thành phố.

- **Xu hướng đánh giá của khách hàng (Review Trends):**

1. Định nghĩa: Lấy dữ liệu từ `reviews_processed.csv`, chuyển `date` thành datetime để tính toán xu hướng đánh giá của khách hàng.
2. Công thức:

$$ReviewCount_{month} = \text{size}(reviews_{tháng})$$

3. Ý nghĩa: Phản ánh xu hướng nhu cầu thị trường.

- **Lưu dữ liệu KPI:**

1. Xuất file `processed/<city_name>/kpi_summary_general_<city_name>.csv` (Total Listings, Multi-Host Rate, Median Price, Median Availability 90).
2. Xuất file `processed/<city_name>/kpi_seasonality_<city_name>.csv` (Daily Occupancy Rate).
3. Xuất file `processed/<city_name>/kpi_room_type_<city_name>.csv` (Room Type Distribution).
4. Xuất file `processed/<city_name>/kpi_neighbourhood_<city_name>.csv` (Neighbourhood Statistics).
5. Xuất file `processed/<city_name>/kpi_reviews_trend_<city_name>.csv` (Review Trends)
6. Chú thích `<city_name>` - tên thành phố.

- **Tóm tắt kết quả KPI:**

1. KPI tổng quan (General Summary) - file `kpi_summary_general_<city_name>.csv`:

City	Snapshot Date	Total Listings	Multi-host Rate (%)	Median Price (€)	Median Avail 90 (Days)
Brussels	16/03/2025	5,526	52.44%	85.0	44.0
Brussels	21/06/2025	5,664	50.05%	87.0	49.0
Brussels	22/12/2024	5,640	52.43%	95.0	61.0
Berlin	15/03/2025	8,898	52.33%	94.0	46.0
Berlin	20/06/2025	9,183	53.41%	108.0	44.0
Berlin	21/12/2024	8,990	50.46%	97.0	65.0

City	Snapshot Date	Total Listings	Multi-host Rate (%)	Median Price (€)	Median Avail 90 (Days)
Berlin	23/09/2025	9,264	54.75%	104.0	54.0
Paris	03/03/2025	55,601	39.48%	146.0	32.0
Paris	06/12/2024	60,046	38.23%	150.0	59.0
Paris	06/06/2025	53,912	40.29%	161.0	44.0
Paris	12/09/2025	0	0.00%	-	-

2. KPI theo mùa (Seasonality) - file `kpi_seasonality_<city_name>.csv`:

City	Date	Occupancy Rate	Snapshot Date
Brussels	16/03/2025	80.64%	16/03/2025
Brussels	17/03/2025	79.26%	16/03/2025
Brussels	18/03/2025	77.23%	16/03/2025
Brussels	19/03/2025	74.32%	16/03/2025
Brussels	20/03/2025	73.09%	16/03/2025
Brussels	21/03/2025	77.31%	16/03/2025
Brussels	22/03/2025	77.67%	16/03/2025
Brussels	23/03/2025	58.51%	16/03/2025
Brussels	24/03/2025	55.70%	16/03/2025
...	...	...	...
Berlin	15/03/2025	82.76%	15/03/2025
Berlin	16/03/2025	69.01%	15/03/2025
Berlin	17/03/2025	56.68%	15/03/2025
Berlin	18/03/2025	54.21%	15/03/2025
Berlin	19/03/2025	54.64%	15/03/2025
Berlin	20/03/2025	55.06%	15/03/2025
Berlin	21/03/2025	57.48%	15/03/2025
Berlin	22/03/2025	56.86%	15/03/2025
Berlin	23/03/2025	47.66%	15/03/2025
...	...	...	...

City	Date	Occupancy Rate	Snapshot Date
Paris	03/03/2025	83.82%	03/03/2025
Paris	04/03/2025	76.63%	03/03/2025
Paris	05/03/2025	71.98%	03/03/2025
Paris	06/03/2025	75.82%	03/03/2025
Paris	07/03/2025	77.03%	03/03/2025
Paris	08/03/2025	72.97%	03/03/2025
Paris	09/03/2025	62.86%	03/03/2025
Paris	10/03/2025	61.91%	03/03/2025
Paris	11/03/2025	62.36%	03/03/2025
...	...	...	...

3. KPI theo loại phòng (Room Type Distribution) - file `kpi_room_type_<city_name>.csv`:

City	Snapshot Date	Entire home/apt	Private room	Hotel room	Shared room
Brussels	16/03/2025	74.54%	24.77%	0.31%	0.38%
Brussels	21/06/2025	74.06%	25.46%	0.16%	0.32%
Brussels	22/12/2024	74.24%	25.41%	0.27%	0.09%
Berlin	15/03/2025	73.80%	24.39%	1.00%	0.81%
Berlin	20/06/2025	73.59%	24.31%	1.03%	1.07%
Berlin	21/12/2024	74.45%	24.07%	1.03%	0.44%
Berlin	23/09/2025	73.47%	24.40%	1.09%	1.05%
Paris	03/03/2025	90.14%	8.83%	0.81%	0.22%
Paris	06/12/2024	89.96%	8.73%	0.92%	0.39%
Paris	06/06/2025	90.26%	8.64%	0.85%	0.25%

4. KPI theo khu vực (Neighbourhood) - file `kpi_neighbourhood_<city_name>.csv`:

City	Neighbourhood	Listing Count	Median Price (€)	Snapshot Date
Brussels	Anderlecht	339	79.0	16/03/2025
Brussels	Auderghem	74	85.5	16/03/2025

City	Neighbourhood	Listing Count	Median Price (€)	Snapshot Date
Brussels	Berchem-Sainte-Agathe	34	87.5	16/03/2025
Brussels	Bruxelles (Center)	1,654	95.0	16/03/2025
Brussels	Etterbeek	285	82.0	16/03/2025
Brussels	Evere	72	75.0	16/03/2025
Brussels	Forest	308	81.0	16/03/2025
Brussels	Ganshoren	32	80.5	16/03/2025
Brussels	Ixelles	786	86.0	16/03/2025
...	...	...	...	...
Berlin	Adlershof	25	97.0	15/03/2025
Berlin	Albrechtstr.	48	76.0	15/03/2025
Berlin	Alexanderplatz	674	121.0	15/03/2025
Berlin	Allende-Viertel	3	42.0	15/03/2025
Berlin	Alt Treptow	54	89.5	15/03/2025
Berlin	Alt-Hohenschönhausen Nord	13	60.0	15/03/2025
Berlin	Alt-Hohenschönhausen Süd	17	102.0	15/03/2025
Berlin	Alt-Lichtenberg	29	84.0	15/03/2025
Berlin	Altglienicke	18	60.0	15/03/2025
...	...	...	...	...
Paris	Batignolles-Monceau	3,735	140.0	03/03/2025
Paris	Bourse	2,224	176.0	03/03/2025
Paris	Buttes-Chaumont	2,668	108.0	03/03/2025
Paris	Buttes-Montmartre	5,503	120.0	03/03/2025
Paris	Entrepôt	3,619	135.0	03/03/2025
Paris	Gobelins	1,672	115.0	03/03/2025
Paris	Hôtel-de-Ville	1,951	182.0	03/03/2025
Paris	Louvre	1,459	212.0	03/03/2025
Paris	Luxembourg	1,880	200.0	03/03/2025
...	...	...	...	...

5. KPI theo đánh giá khách hàng (Review Trends) - file `kpi_reviews_trend_<city_name>.csv`:

City	Date	Review Count	Snapshot Date
Brussels	30/11/2010	2	16/03/2025
Brussels	31/12/2010	0	16/03/2025
Brussels	31/01/2011	0	16/03/2025
Brussels	28/02/2011	1	16/03/2025
Brussels	31/03/2011	3	16/03/2025
Brussels	30/04/2011	0	16/03/2025
Brussels	31/05/2011	11	16/03/2025
Brussels	30/06/2011	9	16/03/2025
Brussels	31/07/2011	1	16/03/2025
...	...	...	...
Berlin	30/06/2009	1	15/03/2025
Berlin	31/07/2009	0	15/03/2025
Berlin	31/08/2009	0	15/03/2025
Berlin	30/09/2009	0	15/03/2025
Berlin	31/10/2009	0	15/03/2025
Berlin	30/11/2009	0	15/03/2025
Berlin	31/12/2009	0	15/03/2025
Berlin	31/01/2010	0	15/03/2025
Berlin	28/02/2010	1	15/03/2025
...	...	...	...
Paris	30/06/2009	1	03/03/2025
Paris	31/07/2009	2	03/03/2025
Paris	31/08/2009	0	03/03/2025
Paris	30/09/2009	1	03/03/2025
Paris	31/10/2009	0	03/03/2025
Paris	30/11/2009	0	03/03/2025

City	Date	Review Count	Snapshot Date
Paris	31/12/2009	1	03/03/2025
Paris	31/01/2010	1	03/03/2025
Paris	28/02/2010	2	03/03/2025
...	...	...	...

E. TRỰC QUAN HÓA

E.1 Thành phố Brussels

E.2 Thành phố Berlin

1. H1. XU HƯỚNG TỔNG NGUỒN CUNG THEO THỜI GIAN
2. H2. XU HƯỚNG GIÁ THUÊ TRUNG VỊ THEO THỜI GIAN
3. H3. TỶ LỆ CHỦ NHÀ CHUYÊN NGHIỆP THEO THỜI GIAN
4. H4. XU HƯỚNG ĐỘ SẴN SÀNG TRUNG VỊ THEO THỜI GIAN
5. H5. XU HƯỚNG TỶ LỆ LẤP ĐẦY THEO SNAPSHOT MỚI NHẤT
6. H6. CƠ CẤU LOẠI PHÒNG THEO THỜI GIAN
7. H7. TOP 10 KHU VỰC CÓ TỔNG NGUỒN CUNG CAO NHẤT THE0 SNAPSHOT MỚI NHẤT
8. H8. TOP 10 KHU VỰC CÓ GIÁ THUÊ TRUNG VỊ CAO NHẤT THEO SNAPSHOT MỚI NHẤT
9. H9. PHÂN PHỐI GIÁ THUÊ DỰA TRÊN TỔNG NGUỒN CUNG THEO SNAPSHOT MỚI NHẤT
10. H10. TOP 10 LOẠI HÌNH BẤT ĐỘNG SẢN PHỔ BIẾN NHẤT THEO SNAPSHOT MỚI NHẤT
11. H11. PHÂN KHÚC THỊ TRƯỜNG GIÁ THUÊ THEO SNAPSHOT MỚI NHẤT
12. H12. XU HƯỚNG SỐ LƯỢNG REVIEW THEO SNAPSHOT MỚI NHẤT

E.3 Thành phố Paris

F. SO SÁNH

G. NÂNG CAO

H. DIỄN GIẢI & KẾT LUẬN

H.1 Các phát hiện chính dựa trên kết quả phân tích dữ liệu từ 3 thành phố

- **Sự phân cực rõ rệt về quy mô, giá cả và sức hấp dẫn:**

1. **Paris** là thị trường vượt trội với quy mô khổng lồ (~55.000 - 60.000 listings), gấp hơn 6 lần **Berlin** (~9.000 - 9.300 listings) và 10 lần **Brussels** (~5.000 - 5.700 listings). (H1)
2. **Paris** đồng thời cũng là thị trường đắt đỏ nhất với giá thuê trung vị (Median Price) dao động từ 146€ - 161€/đêm, cao gấp rưỡi so với **Berlin** (97€ - 104€/đêm) và **Brussels** (85€ - 95€/đêm). (H2)
3. Tuy nhiên, thị trường của **Paris** đắt đỏ nhưng lại cực kỳ đắt khách, điều đáng chú ý là trong mùa cao điểm (Tháng 3), độ sẵn sàng trung vị (Median Availability 90) của **Paris** lại thấp nhất (32 ngày), so với 2 thành phố **Berlin** và **Brussels** (~45 ngày). (H3)

⇒ Kết luận: Điều này cho thấy **Paris** là một "Thị trường của người bán" (Seller's Market) - nơi nhu cầu du lịch khổng lồ chấp nhận mức giá cao, giúp người bán có lợi thế định giá. Trong khi **Berlin** và **Brussels** có sự cân bằng tốt hơn giữ cung và cầu.

- **Mức độ chuyên nghiệp hóa và cấu trúc thị trường:**

1. Một điểm tương phản đáng chú ý là dù **Paris** có quy mô lớn nhất, tỷ lệ chuyên nghiệp hóa (Multi-Host Rate - tỷ lệ chủ nhà quản lý từ 2 căn trở lên) lại thấp nhất (~38.2% - 40.3%). Điều này gợi ý rằng phần lớn nguồn cung tại **Paris** đến từ các các nhân hoặc hộ gia đình tận dụng căn hộ trống. Trong khi đó, **Berlin** và **Brussels** có tỷ lệ chuyên nghiệp hóa cao hơn hẳn (Berlin: ~50.5% - 54.8%, Brussels: ~50.05% - 52.4%), cho thấy sự tham gia mạnh mẽ và chi phối các đơn vị quản lý bất động sản chuyên nghiệp hoặc các nhà đầu tư tại 2 thành phố này. (H4)
2. Tỷ lệ chuyên nghiệp hóa ở 2 thành phố **Berlin** và **Brussels** cao cho thấy thị trường tại nơi đây đã dịch chuyển sang mô hình "Khách sạn hóa". Nguồn cung không chỉ đến từ người dân chia sẻ nhà thừa mà còn bị chi phối bởi các công ty quản lý bất động sản hoặc các nhà đầu tư chuyên nghiệp. Ngược lại, tỷ lệ chuyên nghiệp hóa thấp tại **Paris** phản ánh một thị trường vẫn giữ được bản chất của một "Nền kinh tế chia sẻ" (Sharing Economy). Điều này có thể là hệ quả của các quy định pháp lý chặt chẽ tại **Paris** khiến các cá nhân, tổ chức thâu tóm nhiều căn hộ để kinh doanh ngắn hạn trở nên khó khăn hơn.

⇒ Kết luận: Người thuê tại **Berlin** và **Brussels** có xác suất cao gặp các "Chủ nhà công nghiệp" vận hành bài bản như doanh nghiệp, trong khi tại **Paris**, cơ hội trải nghiệm ở tại nhà của người dân địa phương vẫn chiếm đa số.

- **Xu hướng tỷ lệ lấp đầy:**

1. Tại **Paris**, tỷ lệ lấp đầy duy trì ở mức cao và ổn định nhất trong 3 thành phố. Sau giai đoạn giảm nhẹ vào giữa năm, tỷ lệ lấp đầy của **Paris** phục hồi rõ rệt từ cuối năm 2025 và ổn định quanh mức 55% - 60%, cho thấy nhu cầu lưu trú mạnh và mang tính bền vững. Bên cạnh đó, **Brussels** lại thể hiện mức lấp đầy trung bình, với biến động vừa phải theo mùa. Giai đoạn đầu có dao động tương đối lớn, sau đó xu hướng trở nên ổn định hơn, đặc biệt từ đầu năm 2026 khi tỷ lệ lấp đầy duy trì quanh 50% - 55%. Điều này cho thấy thị trường có quy mô nhỏ hơn nhưng tương đối ổn định. Ngoài ra, **Berlin** là thành phố có biến động mạnh nhất. Tỷ lệ lấp đầy ban đầu khá cao nhưng giảm đáng kể vào cuối năm 2025, sau đó phục hồi dần trong năm 2026. Mức lấp đầy của **Berlin** nhìn chung thấp hơn **Paris** và **Brussels**, phản ánh sự bất ổn hoặc ảnh hưởng từ các yếu tố chính sách và cung và cầu. (H5)
2. Điểm chung ở 3 thành phố là đều khởi đầu với tỷ lệ lấp đầy rất cao (> 80%) rồi trượt dần xuống và ổn định ở mức 50% - 60%.

⇒ Kết luận: Dữ liệu lấp đầy củng cố nhận định rằng **Paris** thể hiện hiệu suất cao nhất và ổn định nhất, trong khi **Brussels** duy trì mức trung bình nhưng tương đối ổn định. **Berlin** cho thấy mức độ biến động lớn hơn, cho thấy thị trường này nhạy cảm hơn trước các yếu tố bên ngoài như chính sách quản lý, tính mùa vụ hoặc các điều chỉnh của thị trường.

- **Cấu trúc loại phòng đặc thù và phân khúc khách hàng mục tiêu:**

1. Tại **Paris**, loại hình "Entire home/apt" (Căn hộ nguyên căn) chiếm ưu thế tuyệt đối (~90%), trong khi "Private room" (Phòng riêng) chỉ chiếm chưa đầy 9%. Ngược lại, **Berlin** và **Brussels** có cơ cấu tương đồng nhau, duy trì khoảng 73% - 74% thị phần cho "Entire home/apt" và 24% - 25% thị phần cho "Private room". Đặc biệt, điểm chung của 3 thành phố này là đều dành rất ít thị phần cho 2 loại hình còn lại là "Shared room" (Phòng chung) và "Hotel room" (Phòng khách sạn). (H6)
2. Sự chênh lệch này định hình phân khúc khách hàng rõ rệt: **Paris** tập trung vào khách du lịch nhóm, gia đình hoặc phân khúc cao cấp đề cao sự riêng tư. Trong khi đó, **Berlin** và **Brussels** vẫn duy trì được phân khúc giá rẻ, phổ thông phục vụ tốt cho đối tượng khách hàng du lịch bụi (backpackers), sinh viên hoặc người đi công tác đơn lẻ muốn tiết kiệm chi phí. (H8)

⇒ Kết luận: Airbnb tại **Paris** dường như đang cạnh tranh trực tiếp với thị trường khách sạn truyền thống (về tính riêng tư và giá cả), trong khi tại **Berlin** và **Brussels**, nó đóng vai trò bổ trợ quan trọng cho thị trường lưu trú giá rẻ.

- **Xu hướng phản hồi từ khách hàng:**

1. **Paris** ghi nhận số lượng review cao vượt trội và tăng trưởng mạnh theo thời gian (đạt đỉnh > 70.000 review / tháng vào giữa năm 2025), phản ánh quy mô thị trường lớn và mức độ hoạt động cao của các listings Airbnb. **Berlin** thể hiện mức tăng ổn định hơn (đạt đỉnh ~15.000 review / tháng vào giữa năm 2025), với số lượng review thấp hơn **Paris** nhưng cao hơn **Brussels** trong hầu hết các giai đoạn. Trong khi đó, **Brussels** có quy mô nhỏ nhất (đạt đỉnh ~12.000 review / tháng vào giữa năm 2025), với số lượng review thấp hơn đáng kể, mặc dù vẫn cho thấy xu hướng tăng dần theo thời gian. (H12)
2. Đáng chú ý, cả ba thành phố đều ghi nhận sự sụt giảm mạnh số lượng review vào giai đoạn 2020–2021, nhiều khả năng liên quan đến tác động của đại dịch COVID-19. Sau giai đoạn này, số lượng review phục hồi rõ rệt, đặc biệt tại **Paris**, cho thấy sự quay trở lại mạnh mẽ của nhu cầu du lịch và lưu trú ngắn hạn.

⇒ Kết luận: **Paris** luôn dẫn đầu về số lượng review, tiếp theo là **Berlin** và **Brussels**. Giai đoạn 2020–2021 ghi nhận sự sụt giảm mạnh của cả 3 thành phố, nhưng sau đó là quá trình phục hồi rõ rệt, mạnh mẽ.

H.2 Đánh giá tác động của quy trình QA

- **Ngăn chặn sai lệch thống kê quan trọng:**

1. Quy tắc `QA001_price_zero` đã phát hiện sự cố bất thường trong snapshot ngày 12/09/2025 tại **Paris**, với 81.853 bản ghi bị lỗi giá (price = 0 hoặc null).
2. Tác động: Việc kiên quyết loại bỏ các bản ghi này dẫn đến việc snapshot **Paris** tháng 9/2025 không còn dữ liệu để phân tích (Total Listing = 0). Thay vào đó, ta coi snapshot mới nhất của thành phố **Paris** là snapshot gần nhất trước snapshot lỗi (6/6/2025). Việc loại bỏ snapshot lỗi là quyết định đánh đổi cần thiết, nếu giữ lại các chỉ số cơ bản sẽ bị kéo xuống thấp một cách sai lệch dẫn đến các kết luận sai về thị trường.

- **Tăng độ chính xác về không gian địa lí:**

1. Quy tắc `QA002_coords_out_of_bounds` đã lọc bỏ các listings nằm ngoài ranh giới hành chính thực tế của thành phố.
2. Tác động: Giúp biểu đồ phân bố hiển thị chính xác mật độ phòng tại ác quận trung tâm, loại bỏ các điểm nhiễu (outliers).

- **Tối ưu hóa hiệu suất xử lý:**

1. Quy tắc `QA003_duplicate_ids`, `QA005_orphaned_or_empty_reviews` đã lọc bỏ các listings có `id` trùng và các review của các listings không tồn tại.
2. Tác động: Giúp giảm kích thước bộ dữ liệu, tăng tốc độ tính toán KPI mà không ảnh hưởng đến bức tranh tổng thể.

H.3 Hạn chế

- **Khoảng trống dữ liệu:** Do sự cố tại nguồn dữ liệu đầu vào, toàn bộ dữ liệu của **Paris** vào tháng 9/2025 bị khuyết thiếu. Điều này gây khó khăn cho việc phân tích chuỗi thời gian liên tục và làm mất đi cái nhìn về giai đoạn chuyển giao giữa mùa hè và mùa thu tại thị trường quan trọng nhất.
- **Thiếu chỉ số doanh thu thực:** Các phân tích hiện tại về "Giá" và "Doanh thu ước tính" hoàn toàn dựa trên giá niêm yết (Listing Price) và lịch trống/bận. Do không có dữ liệu giao dịch thực tế từ Airbnb, chúng ta không thể biết chính xác giá cuối cùng khách hàng phải trả (sau khi áp mã giảm giá, đàm phán) hay liệu một ngày "bận" trên lịch là do khách đặt hay do chủ nhà tự khóa phòng.
- **Dữ liệu đánh giá có độ trễ:** Số lượng review chỉ phản ánh những khách hàng đã hoàn thành chuyến đi và viết đánh giá, do đó chỉ số này luôn có độ trễ so với nhu cầu đặt phòng thực tế.

H.4 Đề xuất cải tiến

- **Cải thiện quy trình thu thập dữ liệu:** Thiết lập cơ chế cảnh báo sớm: Nếu tỷ lệ dữ liệu lỗi trong một snapshot vượt quá ngưỡng an toàn, hệ thống cần gửi cảnh báo ngay lập tức để các kỹ sư có thể kiểm tra lại nguồn thu thập, tránh tình trạng mất dữ liệu.
- **Xây dựng KPI nâng cao:**

1. Phát triển chỉ số RevPAR (Revenue Per Available Room - Doanh thu trên mỗi phòng có sẵn) ước tính. Bằng cách kết hợp Giá niêm yết và Tỷ lệ lấp đầy, ta có thể đo lường hiệu quả kinh doanh của từng khu vực tốt hơn là chỉ nhìn vào giá cao hay thấp.
2. Phân tích Superhost vs. Regular Host: Đánh giá xem danh hiệu "Superhost" có thực sự mang lại tỷ lệ lấp đầy và mức giá cao hơn đáng kể hay không.

- **Tích hợp dữ liệu sự kiện:** Kết hợp dữ liệu Airbnb với lịch sự kiện địa phương (thế vận hội, tuần lễ thời trang, hội nghị thượng đỉnh...). Điều này sẽ giúp giải thích các đợt tăng giá đột biến (Spikes) một cách thuyết phục hơn và hỗ trợ dự báo nhu cầu (Demand Forecasting) chính xác hơn.

I. TÀI LIỆU THAM KHẢO & CÔNG CỤ SỬ DỤNG

- **Nguồn dữ liệu và tài liệu tham khảo, khám phá:** Nguồn dữ liệu chính được sử dụng trong dự án lấy từ **Inside Airbnb**, bao gồm các tập dữ liệu Listings, Calendar, Reviews, Neighbourhoods của **Brussels**, **Berlin** và **Paris**.

1. Trang tải dữ liệu chính thức: https://insideairbnb.com/get-the-data/
2. Trang tải tài liệu khám phá: https://insideairbnb.com/explore
3. Từ điển dữ liệu: https://insideairbnb.com/data_dictionary

- **Thư viện và ngôn ngữ lập trình:** Dự án được thực hiện trên ngôn ngữ Python, sử dụng môi trường Jupyter Notebook cùng các thư viện mã nguồn mở.

1. Xử lý dữ liệu: Numpy, Pandas.
2. Trực quan hóa: Matplotlib, Seaborn.
3. Tiện ích khác: Glob / OS, Gzip.