**1.3. Phân tích đơn biến và hai biến:**

**1.3.1. Ôn lý thuyết**

**a**. **Phân tích đơn biến (univariate analysis) là gì? Nó khác gì với phân tích hai biến (bivariate analysis) trong khám phá dữ liệu?**

**-Phân tích đơn biến (Univariate analysis)** là phương pháp thống kê dùng để phân tích một biến số đơn lẻ trong tập dữ liệu. Mục đích là hiểu rõ đặc tính của biến đó thông qua các chỉ số thống kê tóm tắt (trung bình, trung vị, mode, độ lệch chuẩn…), bảng phân phối tần suất hoặc các biểu đồ trực quan (histogram, bar chart, pie chart...).

**-Phân tích hai biến (Bivariate analysis)** là phương pháp thống kê dùng để phân tích đồng thời hai biến, nhằm khám phá mối quan hệ hoặc sự liên kết giữa chúng. Kỹ thuật này giúp xác định biến này thay đổi như thế nào khi biến kia thay đổi, đo lường cường độ và ý nghĩa thống kê của mối quan hệ. Các công cụ thường dùng gồm scatter plot, boxplot, heatmap, hệ số tương quan, bảng chéo.

-**Khác biệt chính**:

**+ Univariate**: tập trung vào **một biến duy nhất** để hiểu rõ đặc điểm riêng của nó.

**+ Bivariate**: xem xét **hai biến cùng lúc** để tìm hiểu **mối quan hệ, sự tương quan hoặc ảnh hưởng qua lại**.

**b. Các thước đo thống kê nào thường được sử dụng trong phân tích đơn biến (ví dụ: trung bình, trung vị, mode, độ lệch chuẩn)?**

**-Thước đo xu hướng trung tâm (Measures of Central Tendency):**

+ **Trung bình (Mean):**  
Giá trị trung bình cộng của tất cả các giá trị trong tập dữ liệu.

+ **Trung vị (Median):**  
Giá trị ở giữa khi dữ liệu được sắp xếp theo thứ tự tăng dần; nó ít bị ảnh hưởng bởi các giá trị ngoại lai hơn trung bình.

+ **Yếu vị (Mode):**  
Giá trị xuất hiện nhiều nhất trong tập dữ liệu; có thể có một hoặc nhiều mode (đa mode).

**-Thước đo độ phân tán (Measures of Dispersion / Variability)**

**+ Độ lệch chuẩn (Standard Deviation):**Đo lường sự phân tán của dữ liệu so với trung bình; giá trị càng nhỏ chứng tỏ dữ liệu tập trung gần trung bình, giá trị càng lớn cho thấy dữ liệu phân tán rộng.

**+ Phương sai (Variance):**  
Bình phương của độ lệch chuẩn, đo lường sự phân tán trung bình của các giá trị so với trung bình.

**+ Khoảng biến thiên (Range):**

Sự khác biệt giữa giá trị lớn nhất và nhỏ nhất trong tập dữ liệu.

**+ Khoảng tứ phân vị (Interquartile Range – IQR):**  
Sự khác biệt giữa tứ phân vị thứ ba (Q3)(Q\_3)(Q3​) và tứ phân vị thứ nhất (Q1)(Q\_1)(Q1​), cho biết sự phân tán của 50% dữ liệu ở giữa.

**-Thước đo vị trí và hình dạng phân phối**

**+ Phân vị (Percentiles / Quartiles):**

* Phân vị (Percentiles): chia tập dữ liệu thành 100 phần bằng nhau.
* Tứ phân vị (Quartiles): chia tập dữ liệu thành 4 phần bằng nhau.

**+ Độ xiên (Skewness):**

Đo lường mức độ bất đối xứng của phân phối dữ liệu.

**+ Độ nhọn (Kurtosis):**

Đo lường mức độ "nhọn" hoặc "bằng phẳng" của đỉnh phân phối so với phân phối chuẩn.

**c. Trong phân tích hai biến, làm thế nào để xác định mối quan hệ giữa hai biến (ví dụ: tương quan, nhân quả)?**

Để xác định mối quan hệ giữa hai biến, ta sử dụng các phương pháp như phân tích tương quan để đo lường sự cùng biến động, và phân tích nhân quả (thường thông qua thí nghiệm ngẫu nhiên có kiểm soát hoặc mô hình cao cấp) để tìm ra quan hệ nguyên nhân - kết quả. Phân tích tương quan chỉ ra rằng hai biến thay đổi cùng nhau, còn nhân quả chứng minh một biến gây ra sự thay đổi ở biến kia.

* Xác định sự tương quan:

**+ Khái niệm:** Tương quan đo lường mức độ mà hai biến có xu hướng thay đổi cùng nhau.

**+ Cách thực hiện:**

* + **Phân tích trực quan:** Vẽ biểu đồ phân tán (scatter plot) để xem liệu có một xu hướng chung giữa hai biến hay không.
  + **Sử dụng hệ số tương quan:** Tính toán hệ số tương quan (như [Pearson](https://www.google.com/search?cs=0&sca_esv=a69f5a9c38853b18&sxsrf=AE3TifOUOGmjJrVjm1WGlxpBSfYe_akGvw%3A1759591293557&q=Pearson&sa=X&ved=2ahUKEwiWsNSv7IqQAxXH1zQHHT8EOnkQxccNegQIFBAB&mstk=AUtExfCFCO43RGlCbJlOXvihQesaCENiwabW94tfnuoXS9n1s3H-ZUupZLGc8adFd7hdDTwXZhKZARvy_vjFHbm50xz9ILwWfDF5MSTRit2xE0c9991XO6eY4ni7MaeKaW_41anVc6d1yGY0bFFDBkyVj_vfNZYrreXv94_LlFZIXK2jURc&csui=3)) để đo lường sức mạnh và hướng của mối quan hệ tuyến tính. Hệ số này dao động từ -1 (tương quan nghịch hoàn toàn) đến +1 (tương quan thuận hoàn toàn), với 0 nghĩa là không có tương quan tuyến tính.
* Xác định nhân quả:

**+ Khái niệm:**

Nhân quả cho thấy một biến (biến độc lập) gây ra sự thay đổi ở biến kia (biến phụ thuộc).

**+ Cách thực hiện:**

* + **Thí nghiệm ngẫu nhiên có kiểm soát (RCT):** Là phương pháp “tiêu chuẩn vàng” để thiết lập nhân quả. Trong RCT, người tham gia được phân ngẫu nhiên vào nhóm đối chứng và nhóm can thiệp để loại bỏ ảnh hưởng của các yếu tố gây nhiễu.
  + **Thiết kế nghiên cứu khác:** Khi RCT không khả thi, có thể sử dụng các thiết kế nghiên cứu quan sát phức tạp hơn, như phân tích hồi quy có kiểm soát các biến nhiễu hoặc các mô hình [ca](https://www.google.com/search?cs=0&sca_esv=a69f5a9c38853b18&sxsrf=AE3TifOUOGmjJrVjm1WGlxpBSfYe_akGvw%3A1759591293557&q=ca&sa=X&ved=2ahUKEwiWsNSv7IqQAxXH1zQHHT8EOnkQxccNegQIKhAB&mstk=AUtExfCFCO43RGlCbJlOXvihQesaCENiwabW94tfnuoXS9n1s3H-ZUupZLGc8adFd7hdDTwXZhKZARvy_vjFHbm50xz9ILwWfDF5MSTRit2xE0c9991XO6eY4ni7MaeKaW_41anVc6d1yGY0bFFDBkyVj_vfNZYrreXv94_LlFZIXK2jURc&csui=3)o cấp (causal inference models).

**+ Tiêu chí để thiết lập nhân quả (theo**[**Bradford Hill**](https://www.google.com/search?cs=0&sca_esv=a69f5a9c38853b18&sxsrf=AE3TifOUOGmjJrVjm1WGlxpBSfYe_akGvw%3A1759591293557&q=Bradford+Hill&sa=X&ved=2ahUKEwiWsNSv7IqQAxXH1zQHHT8EOnkQxccNegQIIxAB&mstk=AUtExfCFCO43RGlCbJlOXvihQesaCENiwabW94tfnuoXS9n1s3H-ZUupZLGc8adFd7hdDTwXZhKZARvy_vjFHbm50xz9ILwWfDF5MSTRit2xE0c9991XO6eY4ni7MaeKaW_41anVc6d1yGY0bFFDBkyVj_vfNZYrreXv94_LlFZIXK2jURc&csui=3)**):**

Các tiêu chí như tính nhất quán, tính mạnh, tính đặc hiệu, tính thời gian (nguyên nhân phải xảy ra trước kết quả) và mối quan hệ liều lượng-phản ứng có thể hỗ trợ việc suy luận nhân quả.

**d. Sự khác biệt giữa tương quan (correlation) và hiệp biến (covariance) trong phân tích hai biến là gì?**

- Hiệp biến (Covariance):

Hiệp biến được dùng để đo lường mức độ mà hai biến thay đổi cùng nhau. Nếu giá trị hiệp biến lớn hơn 0, điều đó có nghĩa là khi một biến tăng thì biến còn lại cũng có xu hướng tăng theo (hai biến thay đổi cùng chiều). Ngược lại, nếu giá trị nhỏ hơn 0, thì khi một biến tăng, biến kia lại có xu hướng giảm (hai biến thay đổi ngược chiều). Tuy nhiên, giá trị hiệp biến phụ thuộc vào đơn vị đo lường của dữ liệu, nên khó dùng để so sánh trực tiếp giữa các cặp biến khác nhau.

- Tương quan (Correlation**):**

Tương quan có thể xem là dạng chuẩn hóa của hiệp biến, giúp khắc phục nhược điểm phụ thuộc vào đơn vị đo. Giá trị của hệ số tương quan luôn nằm trong khoảng từ -1 đến 1, nên dễ dàng so sánh giữa các cặp biến khác nhau. Nếu hệ số tương quan bằng +1, hai biến có mối quan hệ tuyến tính dương hoàn hảo (một biến tăng thì biến kia cũng tăng theo đúng tỉ lệ). Nếu bằng -1, hai biến có mối quan hệ tuyến tính âm hoàn hảo (một biến tăng thì biến kia giảm theo tỉ lệ). Còn nếu hệ số gần bằng 0, tức là gần như không tồn tại mối quan hệ tuyến tính rõ ràng giữa hai biến.

- Khác biệt nhau chính:

**+ Covariance**: cho biết hai biến thay đổi cùng/khác chiều, **nhưng khó so sánh** vì phụ thuộc đơn vị đo.

**+ Correlation**: là hiệp biến đã chuẩn hóa, **dễ so sánh và trực quan hơn**, luôn nằm trong [-1, 1]

**e. Khi nào nên sử dụng biểu đồ trực quan hóa trong phân tích đơn biến so với phân tích hai biến?**

Trong phân tích dữ liệu, biểu đồ trực quan hóa được sử dụng để giúp nhìn rõ đặc điểm phân phối hoặc mối quan hệ giữa các biến. Với phân tích đơn biến, ta thường dùng biểu đồ khi muốn mô tả đặc trưng của một biến duy nhất, chẳng hạn như histogram để quan sát phân phối giá trị, bar chart hoặc pie chart để xem tần suất và tỷ lệ của các nhóm, hoặc boxplot để phát hiện ngoại lệ. Ngược lại, với phân tích hai biến, biểu đồ được dùng để khám phá mối quan hệ giữa hai biến, ví dụ scatter plot để xem xu hướng tuyến tính giữa hai biến số, heatmap để hiển thị ma trận tương quan, hay boxplot/violin plot để so sánh phân phối của một biến số theo từng nhóm của một biến phân loại.

**f. Đoạn code mẫu để tạo biểu đồ scatter plot hoặc heatmap để phân tích mối quan hệ giữa hai biến?**

**-**Tạo dữ liệu:

A screenshot of a computer code

AI-generated content may be incorrect.

**-**Mẫu code tạo biểu đồ scatter plot:

A screen shot of a computer code

AI-generated content may be incorrect.

A graph with blue dots

AI-generated content may be incorrect.

- Mẫu code tạo biểu đồ heatmap:

A computer screen shot of text

AI-generated content may be incorrect.

A red squares with white text

AI-generated content may be incorrect.

**g. Làm thế nào để trực quan hóa mối quan hệ giữa một biến số và một biến phân loại bằng biểu đồ boxplot hoặc violin plot trong Python?**

Để trực quan hóa mối quan hệ giữa một biến số (numeric variable) và một biến phân loại (categorical variable) trong Python, ta thường sử dụng boxplot hoặc violin plot với thư viện seaborn. Biểu đồ boxplot giúp thể hiện trung vị, khoảng tứ phân vị và các giá trị ngoại lệ của biến số theo từng nhóm. Trong khi đó, violin plot không chỉ hiển thị đặc điểm của boxplot mà còn cho thấy mật độ phân phối dữ liệu trong mỗi nhóm.

**1.3.2. Bài làm mẫu**

**Bài toán 1:** Thực hiện các nhiệm vụ trong bài toán 1 để làm quen với các hàm và thư viện hỗ trợ phân tích dữ liệu đơn biến. Bài toán này được thực hiện trên 2 tập dữ liệu là tập dữ liệu về chim cánh cụt và tập dữ liệu giá nhà.

**Nhiệm vụ 1:** Phân tích dữ liệu đơn biến trên dữ liệu về chim cánh cụt

**1. Giới thiệu dữ liệu:**

* Bộ dữ liệu được sử dụng trong nghiên cứu này là Palmer Archipelago (Antarctica) Penguin Data, thu thập tại quần đảo Palmer, Nam Cực và được công bố công khai trên Kaggle. Đây là một tập dữ liệu phổ biến trong giảng dạy và nghiên cứu khoa học dữ liệu, bao gồm thông tin về ba loài chim cánh cụt cùng nhiều đặc trưng sinh học khác (ví dụ: chiều dài mỏ, chiều rộng mỏ, khối lượng cơ thể, giới tính...).
* Trong phạm vi phân tích đơn biến, nghiên cứu tập trung vào hai biến chính:
  + Species: Loài chim cánh cụt (Adelie, Chinstrap, Gentoo) – biến định tính, dùng để phân tích tần suất và tỷ lệ phân bố.
  + Culmen Length (mm): Chiều dài mỏ – biến định lượng liên tục, dùng để khảo sát phân phối giá trị bằng histogram.
* Mục tiêu phân tích hai biến này nhằm làm rõ đặc điểm phân bố số lượng loài và xu hướng biến thiên chiều dài mỏ trong quần thể chim cánh cụt.

**2. Phân tích đơn biến bằng Histogram**

A graph of a number of blue bars

AI-generated content may be incorrect.

Hình 1.1. Biểu đồ phân phối của biến Culmen Length (mm).

Quan sát histogram cho thấy phân phối chiều dài mỏ tập trung chủ yếu trong khoảng 35–50 mm, chiếm phần lớn mẫu quan sát. Đỉnh tần suất rơi vào khoảng 40–45 mm với trên 60 cá thể, phản ánh giá trị phổ biến nhất trong tập dữ liệu.

Ngược lại, số cá thể có chiều dài mỏ dưới 35 mm hoặc trên 55 mm là khá ít, được coi là ngoại biên (outliers). Đặc biệt, phần đuôi phải của phân phối có kéo dài nhẹ, với một số cá thể đạt chiều dài mỏ vượt quá 55 mm.

Điều này gợi ý rằng phân phối của biến Culmen Length gần cân đối quanh trung tâm nhưng có độ lệch phải nhẹ (slight right skew). Kết quả này hàm ý rằng đa số cá thể có mỏ với chiều dài trung bình, trong khi chỉ một nhóm nhỏ sở hữu mỏ dài vượt trội.

Hiện tượng này phản ánh đa dạng sinh học trong quần thể, đồng thời cho thấy chiều dài mỏ có thể đóng vai trò phân tách loài. Ví dụ, Gentoo Penguin thường có chiều dài mỏ lớn hơn đáng kể so với Adelie Penguin, từ đó tạo nên những giá trị ngoại biên ở phía đuôi phải. Ngoài đặc điểm loài, giới tính và tuổi sinh học cũng có thể ảnh hưởng đến đặc điểm hình thái này. Một số giá trị ngoại biên hiếm gặp có thể bắt nguồn từ sai số đo đạc hoặc điều kiện thu thập.

**3. Phân tích đơn biến bằng bar chart**

Biểu đồ Bar Chart được sử dụng để phân tích biến định tính *Species* (loài chim cánh cụt). Biểu đồ thể hiện số lượng quan sát của ba loài chính trong tập dữ liệu: **Adelie Penguin (Pygoscelis adeliae)**, **Chinstrap Penguin (Pygoscelis antarctica)** và **Gentoo Penguin (Pygoscelis papua)**.

A graph of blue squares

AI-generated content may be incorrect.

Hình 1.2. Biểu đồ tần suất loài chim cánh cụt (Species).

Qua biểu đồ cột cho thấy sự phân bố tần suất không đồng đều giữa ba loài được khảo sát. Loài **Adelie Penguin**  là loài chiếm ưu thế rõ rệt trong tập dữ liệu với số lượng **cá thể** cao nhất, xấp xỉ **150** **cá thể** đã được thu thập. Loài **Gentoo Penguin**  đứng ở vị trí thứ hai với khoảng **125 cá thể**, cho thấy mật độ thu thập dữ liệu khá gần với loài Adelie. Ngược lại, loài **Chinstrap Penguin** được ghi nhận với tần suất thấp nhất, chỉ khoảng **70 cá thể**.

Sự chênh lệch tần suất này báo hiệu một mức độ **mất cân bằng lớp (class imbalance)** đáng kể trong tập dữ liệu. Tỷ lệ số lượng của loài Adelie so với loài Chinstrap là hơn 2:1, điều này cần được xem xét nghiêm túc trong các giai đoạn phân tích tiếp theo, đặc biệt là khi xây dựng mô hình phân loại. Sự mất cân bằng này có thể dẫn đến **thiên vị mô hình (model bias)**, trong đó mô hình học máy có xu hướng dự đoán chính xác hơn cho các lớp đa số (Adelie và Gentoo) và kém hiệu quả hơn đối với lớp thiểu số (Chinstrap).

Do đó, ở các bước xử lý tiếp theo, cần cân nhắc áp dụng các kỹ thuật cân bằng dữ liệu như **oversampling**, **undersampling**, hoặc sử dụng các thuật toán thích ứng với dữ liệu mất cân bằng, nhằm đảm bảo tính công bằng và nâng cao độ chính xác của mô hình trên toàn bộ các lớp.

**4. Phân tích đơn biến bằng biểu đồ tròn (Pie-chart)**

Biểu đồ tròn (Pie Chart) được sử dụng để thể hiện tỷ lệ phần trăm quan sát của từng loài chim cánh cụt (*Species*) trong tập dữ liệu. Biểu đồ này giúp dễ dàng nhận diện tỷ trọng tương đối giữa các nhóm, thay vì số lượng tuyệt đối như Bar Chart.

A pie chart with different colored circles

AI-generated content may be incorrect.

Hình 1.3. Biểu đồ tròn thể hiện tỷ lệ phân bố loài.

Kết quả cho thấy loài **Adelie Penguin (Pygoscelis adeliae)** chiếm tỷ lệ cao nhất, khoảng **44%** tổng số mẫu. Tiếp đến là loài **Gentoo Penguin (Pygoscelis papua)** với khoảng **36%**, trong khi **Chinstrap Penguin (Pygoscelis antarctica)** có tỷ lệ thấp nhất, chỉ khoảng **20%**.

Quan sát này cho thấy dữ liệu tập trung chủ yếu ở hai loài Adelie và Gentoo, trong khi số lượng mẫu của loài Chinstrap ít hơn đáng kể. Thông tin từ biểu đồ tròn góp phần củng cố nhận định từ biểu đồ cột, rằng phân bố mẫu giữa các loài không đồng đều.

**5. Nhận xét và giải thích**

Kết quả phân tích đơn biến cho thấy chiều dài mỏ (Culmen Length) của chim cánh cụt chủ yếu tập trung trong khoảng 35–50 mm, với đỉnh tần suất ở khoảng 40–45 mm. Phân phối có dạng gần cân đối quanh trung tâm nhưng vẫn tồn tại một số giá trị ngoại biên, chủ yếu nằm ở phía đuôi phải trên 55 mm. Hiện tượng này có thể bắt nguồn từ sự khác biệt hình thái giữa các loài, trong đó loài Gentoo thường có chiều dài mỏ lớn hơn đáng kể so với loài Adelie, từ đó tạo nên những cá thể có mỏ dài vượt trội so với trung bình. Ngoài đặc trưng loài, các yếu tố sinh học như giới tính, độ tuổi cũng có khả năng tác động đến sự biến thiên này, trong khi một số ngoại biên có thể xuất phát từ sai số đo đạc hoặc điều kiện thu thập dữ liệu thực địa.

Đối với biến định tính Species, kết quả từ cả Bar Chart và Pie Chart đều phản ánh sự phân bố mẫu không đồng đều giữa ba loài. Adelie là loài chiếm ưu thế với khoảng 44% số mẫu, tiếp theo là Gentoo với khoảng 36%, trong khi Chinstrap chỉ chiếm 20%. Sự mất cân bằng này có thể giải thích từ hai khía cạnh: thứ nhất, yếu tố sinh thái tự nhiên khi Adelie và Gentoo có phạm vi phân bố rộng hơn, mật độ quần thể cao hơn tại khu vực khảo sát; thứ hai, yếu tố thiết kế nghiên cứu và thiên lệch trong quá trình lấy mẫu, khi dữ liệu có xu hướng tập trung nhiều hơn vào các loài phổ biến. Việc nhận diện sự mất cân bằng này đặc biệt quan trọng vì nó có thể ảnh hưởng trực tiếp đến các phân tích tiếp theo, chẳng hạn làm tăng nguy cơ thiên lệch trong mô hình dự đoán hoặc giảm độ chính xác khi phân loại loài thiểu số như Chinstrap.

Phân tích đơn biến đã cung cấp cái nhìn tổng quan về đặc điểm phân bố của các biến trong tập dữ liệu, đồng thời gợi ý những hướng phân tích sâu hơn ở mức nhị biến hoặc đa biến nhằm làm rõ hơn mối quan hệ giữa chiều dài mỏ và loài chim cánh cụt.

**Nhiệm vụ 2**: Phân tích dữ liệu đơn biến trên dữ liệu giá nhà

**1. Import thư viện, nạp dữ liệu giá nhà và phân tích đơn biến dựa vào boxplot**

Trong nghiên cứu này, nhóm tiến hành phân tích dữ liệu giá nhà tại Amsterdam với mục tiêu khảo sát đặc trưng phân bố và nhận diện các giá trị bất thường. Sau khi nạp dữ liệu và lựa chọn các biến số quan trọng như mã vùng, giá, diện tích và số phòng, nhóm xây dựng biểu đồ **boxplot** để mô tả sự phân bố giá trị của biến **Price**.A graph of a house price

AI-generated content may be incorrect.

Hình 2.1. Biểu đồ boxplot thể hiện phân bố giá nhà tại Amsterdam.

Biểu đồ boxplot cho thấy phần lớn giá nhà tập trung trong khoảng dưới một triệu (theo đơn vị dữ liệu), thể hiện qua phần hộp chính của biểu đồ. Tuy nhiên, bên cạnh đó xuất hiện rất nhiều điểm ngoại lai (outliers) nằm rải rác về phía bên phải, kéo dài đến mức giá hơn sáu triệu. Hiện tượng này cho thấy dữ liệu giá nhà có phân bố lệch phải (right-skewed), trong đó đa số bất động sản thuộc nhóm giá trung bình hoặc thấp, nhưng tồn tại một số ít bất động sản có giá trị đặc biệt cao, vượt trội so với mặt bằng chung.

Phát hiện này phản ánh đặc điểm thực tế của thị trường bất động sản tại Amsterdam: phần lớn giao dịch tập trung ở phân khúc phổ thông, trong khi phân khúc cao cấp tuy hiếm nhưng có giá trị rất lớn, góp phần tạo nên sự chênh lệch rõ rệt trong phân phối dữ liệu. Đồng thời, sự tồn tại của nhiều ngoại lai cũng cho thấy các thống kê mô tả như trung bình hoặc phương sai có thể bị ảnh hưởng mạnh, do đó cần được cân nhắc khi sử dụng ở các bước phân tích sau.

**2. Phân tích dữ liệu đơn biến dựa vào violin plot**

A blue line graph with black lines

AI-generated content may be incorrect.

Hình 2.2. Biểu đồ violin plot thể hiện mật độ phân bố giá nhà.

Quan sát từ violin plot, ta nhận thấy phần lớn giá nhà tập trung rất dày ở khoảng dưới 1 triệu. Đây là phân khúc chiếm ưu thế trên thị trường, với mật độ cao và trải rộng trong vùng giá thấp đến trung bình. Khi giá tăng dần, mật độ dữ liệu giảm rõ rệt và kéo dài về phía bên phải đến mức hơn 6 triệu. Phần “đuôi” bên phải của biểu đồ khá dài nhưng mật độ rất thưa, điều này cho thấy chỉ có một số rất ít bất động sản có giá trị vượt trội, đóng vai trò là các điểm ngoại lai.

Violin plot giúp minh họa trực quan sự **lệch phải (right-skewed)** trong phân bố giá nhà: đa số bất động sản tập trung ở vùng giá vừa phải, trong khi một nhóm nhỏ thuộc phân khúc cao cấp có giá cực kỳ cao, tạo ra sự phân hóa rõ rệt. So với boxplot, violin plot mang lại cái nhìn trực quan hơn về hình dạng phân bố, giúp ta thấy rõ không chỉ sự tồn tại của ngoại lai mà còn cả mật độ dữ liệu giữa các mức giá khác nhau.

**3. Phân tích dữ liệu đơn biến dựa vào bản tóm tắt dữ liệu**

A screenshot of a graph

AI-generated content may be incorrect.

Bảng 2.3. Thống kê mô tả biến Price trong dữ liệu giá nhà tại Amsterdam.

Dữ liệu mô tả về biến **Price** cho thấy một sự phân bố giá trị đáng chú ý trong tập dữ liệu gồm **920** quan sát. Giá trị **trung bình** (Mean) của giá nhà là , trong khi **trung vị** (50th Percentile) chỉ là . Sự chênh lệch lớn này—trung bình cao hơn trung vị tới hơn 155,000—là một dấu hiệu rõ ràng cho thấy sự phân bố **lệch phải dương (positively skewed)**. Điều này ngụ ý rằng tồn tại một số lượng nhỏ các căn nhà có giá rất cao đang kéo giá trị trung bình lên. Hơn nữa, với **độ lệch chuẩn (Standard Deviation)** là , gần bằng giá trị trung bình, cho thấy sự **biến động và phân tán giá rất lớn**. Phạm vi giá trị trải dài từ mức **tối thiểu** là đến mức **tối đa** là . Giá trị tối đa này cao gấp hơn 12 lần giá trị trung vị, cung cấp bằng chứng mạnh mẽ về sự tồn tại của **các giá trị ngoại lai (outliers)** cực kỳ đắt đỏ. Phân tích tứ phân vị cho thấy **50%** giá nhà tập trung trong khoảng từ (Q1) đến (Q3), xác nhận rằng đại đa số các giao dịch nhà nằm dưới mức , trong khi phần còn lại của dữ liệu bị chi phối bởi các bất động sản cao cấp.

4. Nhân xét và giải thích

Kết quả phân tích từ boxplot, violin plot và thống kê mô tả đều nhất quán, khẳng định rằng **phân phối giá nhà tại Amsterdam có dạng lệch phải rõ rệt**. Phần lớn giao dịch diễn ra ở mức giá trung bình và phổ thông (dưới 1 triệu), trong khi các bất động sản cao cấp hiếm hoi nhưng có giá trị vượt trội đã kéo dài đuôi phân phối.

Hiện tượng này phản ánh **bối cảnh thực tiễn của thị trường bất động sản Amsterdam**. Là một trung tâm kinh tế – tài chính quan trọng của châu Âu với mật độ dân cư cao nhưng quỹ đất hạn chế, Amsterdam có thị trường nhà ở phổ thông chiếm ưu thế về số lượng giao dịch. Ngược lại, các căn hộ hạng sang hoặc biệt thự cao cấp thường phục vụ giới thu nhập rất cao hoặc mục đích đầu tư quốc tế, dẫn đến sự xuất hiện của những giá trị ngoại lai cực lớn.

Ngoài ra, **chính sách quy hoạch đô thị** cũng góp phần định hình phân bố này: việc hạn chế phát triển dự án quy mô lớn tại trung tâm thành phố làm gia tăng mật độ giao dịch trong phân khúc vừa túi tiền, trong khi nguồn cung nhà ở cao cấp tuy ít nhưng vẫn tồn tại và tạo ra sự chênh lệch mạnh mẽ trong phân phối giá.

Tổng hợp lại, phân tích đơn biến không chỉ giúp mô tả đặc trưng của dữ liệu mà còn đưa ra những gợi ý quan trọng cho các bước tiếp theo. Đặc biệt, sự tồn tại của **ngoại lai và phân phối lệch** là những yếu tố cần được cân nhắc kỹ lưỡng khi xây dựng các mô hình dự báo giá nhà, nhằm tránh sai lệch trong ước lượng và cải thiện tính chính xác.