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

**1.3.1. Ôn lý thuyết**

* **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**.
* **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.

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

* **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]
* **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.

* **Đoạn code mẫu để tạo biểu đồ scatter plot hoặc heatmap để phân tích mối quan hệ giữa hai biến?**
* **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. Import thư viện và nạp dữ liệu

2. Phân tích đơn biến bằng Histogram

3. Phân tích đơn biến bằng bar chart

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

**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

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

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