### **Biểu đồ nâng cao 1**

### **Chi tiết bài đọc**

Trong bài trước, chúng ta đã tìm hiểu cơ bản về Seaborn. Trong bài này, chúng ta sẽ tìm hiểu về 3 loại biểu đồ nâng cao trong Seaborn, đó là biểu đồ xu hướng, biểu đồ phân bố, biểu đồ tần số.

**Biểu đồ xu hướng – lm plot:**

Biểu đồ xu hướng là một loại biểu đồ phân tích dữ liệu Hồi quy tuyến tính, đây là một phương pháp rất phổ biến trong lĩnh vực Business Intelligence và Khoa học dữ liệu nói riêng.

Mọi biểu đồ trong Seaborn đều có một tập hợp các tham số cố định. Đối với biểu đồ xu hướng, ta có ba tham số chính và phần còn lại là tùy chọn mà chúng ta có thể sử dụng theo yêu cầu của mình. 3 tham số này là giá trị cho trục X, giá trị cho trục Y và tham chiếu đến tập dữ liệu. 3 tham số này có thể nhìn thấy trước ở hầu hết các biểu đồ của Seaborn.

Tiếp theo chúng ta sẽ thực hành trên bộ dữ liệu “tips”. Đây là bộ dữ liệu thu thập thông tin tiền tips cho người phục vụ đồ ăn trong nhà hàng theo các thông tin như giới tính, tổng bill, có hút thuốc hay không,…

*# khai báo thư viện và set style*

**import** **numpy** **as** **np**

**import** **pandas** **as** **pd**

**import** **matplotlib.pyplot** **as** **plt**

**import** **seaborn** **as** **sns**

%matplotlib inline

sns.set(style="whitegrid", palette="hsv")

*# load bộ dữ liệu tips có sẵn trong seaborn:*

tips = sns.load\_dataset("tips")

# vẽ biểu đồ xu hướng giữa tổng bill và tiền tip

sns.lmplot(x="total\_bill", y="tip", data=tips)

Chart, scatter chart

Description automatically generated

Đường thẳng xuyên qua biểu đồ là đường thẳng phù hợp nhất với xu hướng tiền tip mà khách hàng thường đưa ra đối với tổng hóa đơn được tạo. Và các điểm dữ liệu mà chúng ta nhìn thấy ở phía trên cùng bên phải nằm cách xa dòng này được gọi là các điểm ngoại lai trong tập dữ liệu.

Mục tiêu của Khoa học dữ liệu là dự đoán sự phù hợp nhất để hiểu xu hướng trong hành vi của khách hàng ghé thăm và thuật toán của chúng ta sẽ luôn được thiết kế phù hợp. Có thể thấy đây là một tình huống phổ biến khi áp dụng các thuật toán hồi quy Logistic trong Học máy. Có thể thấy, phần bóng hội tụ ở trung tâm, nơi có một phần dữ liệu của chúng ta. Điểm cơ bản này thực sự là trung bình thống kê hay nói cách đơn giản hơn là dự đoán tổng quát về giá trị tiền boa tại nhà hàng này hàng ngày.

Trong trường hợp này, nhìn vào biểu đồ này, chúng ta có thể nói rằng nếu tổng hóa đơn là khoảng $ 20,00, thì nó sẽ nhận được tiền boa khoảng $ 3,00.

Chúng ta sẽ tinh chỉnh biểu đồ này hơn nữa bằng cách thêm nhiều tính năng hơn, chúng ta hãy cố gắng hiểu xem một người hút thuốc thì tip ít hay nhiều hơn:

sns.lmplot(x="total\_bill", y="tip", hue="smoker", data=tips)

Chart, scatter chart

Description automatically generated

Biểu đồ trên cho thấy rằng

Chúng ta sẽ xem xét một vài tham số khác thường được dùng để tinh chỉnh biểu đồ.

sns.lmplot(x="total\_bill", y="tip", hue="smoker", data=tips, markers=["o", "x"], palette="Set1", legend=**False**)

Chart, scatter chart

Description automatically generated

Ở đây, chúng ta đặt kiểu đánh dấu điểm dữ liệu, thay đổi màu và quyết định xóa chú thích theo mặc định luôn ở đó.

Có thể thây rằng dữ liệu về người hút thuốc và không hút thuốc đều nằm trên một biểu đồ, ta có thể tách chúng ra như sau.

sns.lmplot(x="total\_bill", y="tip", col="smoker", data=tips)

Chart, scatter chart

Description automatically generated

sns.lmplot(x="total\_bill", y="tip", palette="magma", row="sex", col="time", data=tips, size=3)

Chart, scatter chart

Description automatically generated

Biểu đồ này theo 4 khía cạnh riêng biệt đi sâu hơn vào việc trực quan hóa dữ liệu. Chúng ta vẫn hiển thị tiền tip được đưa ra so với tổng hóa đơn nhưng giờ đây cũng được phân đoạn thành liệu đó có phải là Giờ ăn trưa hay không cùng với sự phụ thuộc vào Giới tính.

Còn rất nhiều cách tinh chỉnh biểu đồ nữa, chúng ta có thể tham khảo thêm tại [*https://seaborn.pydata.org/generated/seaborn.lmplot.html#seaborn.lmplot*](https://seaborn.pydata.org/generated/seaborn.lmplot.html#seaborn.lmplot)

**Biểu đồ phân bố – Violin plot:**

Biểu đồ phân bố mô tả sự phân bố dữ liệu số cho một hoặc nhiều nhóm bằng cách sử dụng đường cong mật độ. Chiều rộng của mỗi đường cong tương ứng với tần số gần đúng của các điểm dữ liệu trong mỗi vùng.

Biểu đồ phân bố violin splot được sử dụng khi ta muốn quan sát sự phân bố của dữ liệu số và đặc biệt hữu ích khi ta muốn so sánh sự phân bố giữa nhiều nhóm. Các đỉnh, thung lũng và đuôi của đường cong mật độ của mỗi nhóm có thể được so sánh để xem các nhóm giống nhau hay khác nhau ở đâu.

Chúng ta sẽ sử dụng lại bộ dữ liệu “tips như phần trước”.

Vẽ biểu đồ phân bố cho thuộc tính total\_bill:

sns.violinplot(y = tips["total\_bill"], palette="coolwarm")

A picture containing arrow

Description automatically generated

Giải thích:

Thanh dày này mà ở trung tâm (giữa các giá trị 12 và 25) đại diện cho dải phân vị.

Đường mảnh ngay trên đường dày thể hiện Khoảng tin cậy 95%.

Chấm trắng nhỏ ở trung tâm của đường kẻ dày là trung vị.

Sự lan truyền tạo nên hình dạng cây đàn Violin thể hiển hình dạng phân phối của dữ liệu. Các phần rộng hơn thể hiện xác suất cao hơn của các phần tử từ không gian mẫu (dữ liệu quan sát). Các phần hẹp hơn đại diện cho một xác suất thấp hơn.

Tiếp theo, chúng ta sẽ tinh chỉnh biểu đồ này thêm.

Vẽ violin plot được nhóm theo biến 1 biến phân loại là ngày:

ax = sns.violinplot(x="day", y="total\_bill", data=tips)

Chart

Description automatically generated

Vẽ violin plot được nhóm theo biến 2 biến phân loại:

ax = sns.violinplot(x="day", y="total\_bill", hue="smoker",

data=tips, palette="muted")

Chart, shape

Description automatically generated

Vẽ violin plot được chia 2 phần để so sánh giữa các biến:

ax = sns.violinplot(x="day", y="total\_bill", hue="smoker",

data=tips, palette="muted", split=True)

Chart

Description automatically generated

Kiểm soát thứ tự các violin plot bằng cách truyền một thứ tự nhãn cụ thể:

ax = sns.violinplot(x="time", y="tip", data=tips,

order=["Dinner", "Lunch"])

Chart

Description automatically generated

Còn rất nhiều cách biểu diễn khác nữa với violin plot ta có thể tham khảo thêm tại https://seaborn.pydata.org/generated/seaborn.violinplot.html

**Biểu đồ tần số – count plot:**

Biểu đồ tần số hiển thị số lượng quan sát trong mỗi nhóm phân loại bằng cách sử dụng các thanh.

Chúng ta sẽ tiếp tục sử dụng bộ dữ liệu “tips” để ví dụ.

Hiển thị tần số của mỗi loại giới tính trong bộ dữ liệu:

sns.countplot(x ='sex', data = tips)

Chart, bar chart, treemap chart

Description automatically generated

Hiển thị số lượng giá trị cho hai biến phân loại và sử dụng tham số hue:

sns.countplot(x ='sex', hue = "smoker", data = tips)

Chart, bar chart

Description automatically generated

Vẽ các thanh theo chiều ngang

sns.countplot(y ='sex', hue = "smoker", data = tips)

Chart, bar chart

Description automatically generatedq

### **Tổng kết**

Qua bài đọc này, chúng ta đã tìm hiểu về 3 loại biểu đồ nâng cao trong Seaborn, đó là biểu đồ xu hướng, biểu đồ phân bố, biểu đồ tần số.