**EDA – Phân tích dữ liệu thăm dò**

**Khái niệm về EDA**

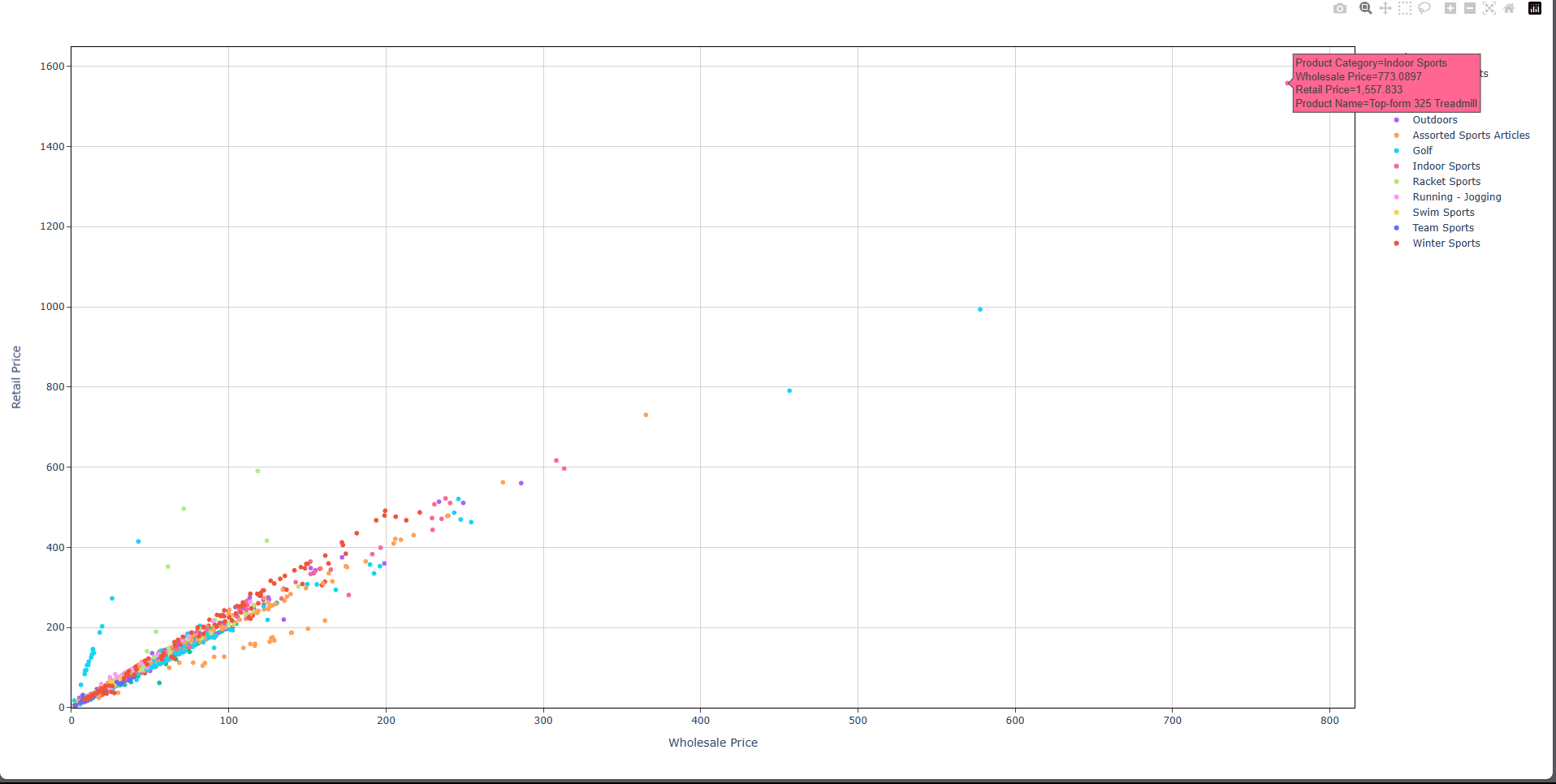
Phân tích dữ liệu thăm dò là quá trình mô tả dữ liệu bằng các kỹ thuật thống kê và trực quan hoá nhằm tập trung vào các khía cạnh quan trọng của dữ liệu để tiếp tục phân tích. Điều này bao gồm cả việc kiểm tra tập dữ liệu từ nhiều góc độ, mô tả và tóm tắt nó mà không đưa ra bất kỳ giả định nào khác về nội dung của nó. Trong đề tài của nhóm chúng em sẽ tập trung hiển thị các đặc trưng của dữ liệu trong bộ dữ liệu đã tiền xử lý ở phần trước

Để có thể hiểu hơn về dữ liệu mà nhóm chúng em đang thực nghiện phân tích, nhóm chúng em tiến hành phân tích và tổng hợp dữ liệu một lần dưới dạng mô hình nữa nhằm đưa ra cái nhìn trực quan hơn về dữ liệu mà chúng em đang làm việc.

* *Hiển thị danh mục mặt hàng với giá bán sỉ - lẻ*



Như hình vẽ trên, thì các danh mục phân loại mặt hàng đã được sắp xếp và hiển thị trên biểu đồ theo mức giá của các mặt hàng. Do đó, theo mô hình hiển thị thì mặt hàng “Top-form 325 Treadmill” đang có mức giá dao động là cao nhất so với toàn bộ mặt hàng.



* *Tổng tiền thu được từ các danh mục mặt hàng*



Biểu đồ trên cho biết tổng tiền thu được từ các danh mục mặt hàng. Trong đó cho biết danh mục quần áo (Clothes) là danh mục chiếm tỷ trọng tổng thu nhập là lớn nhất và danh mục quần áo thể thao trong nhà (Indoor Sports) chiếm tỷ trọng bé nhất trong tất cả danh mục sản phẩm.

* Tổng tiền thu được từ danh mục quần áo (Clothes) là 38.953k
* Tổng tiền thu được từ danh mục quần áo thể thao trong nhà là 2995
* *Số lượng sản phẩm bán được dao động trong các khoảng mức giá khác nhau*
* *Bán sỉ*



Biểu đồ trên cho biết tổng tiền thu được từ các sản phẩm có các khoảng mức giá bán sỉ khác nhau. Ở biểu đồ trên thì các sản phẩm có mức giá từ 5-10 có tổng số lượng bán được 23 ngàn sản phẩm, chiếm tỷ trọng lớn nhất trong toàn bộ sản bộ sản phẩm. Do đó, ta có thể ngầm kết luận rằng ở các khoảng mức giá bản sỉ từ 5-10 chiếm phần lớn người mua nhất và khoảng mức giá từ 770-775 thì chiếm số lượng người mua là ít nhất.

Ngoài ra, bảng nhỏ phía trên cũng mô tả thông tin các sản phẩm nằm trong khoảng mức giá ấy. Bên cạnh đó, ta cũng có kết luận từ một số thống kê cơ bản của các cột như sau:



* Tổng các cột (count): 3124
* Giá trị trung bình (mean): 34.1039767
* Độ lệch chuẩn (std): 40.766120
* Giá trị nhỏ nhất (min): 0.2
* Giá trị phân vị thứ 25 (25%): 11.996711
* Giá trị phân vị thứ 50 (50%): 22.36443
* Giá trị phân vị thứ 75 (75%): 40.180476
* Giá trị lớn nhất (max): 773.089744
* *Bán lẻ*



Biểu đồ trên cho biết tổng tiền thu được từ các sản phẩm có các khoảng mức giá bán lẻ khác nhau. Ở biểu đồ trên thì các sản phẩm có mức giá từ 20-29.99 có tổng số lượng bán được 21.387 sản phẩm, chiếm tỷ trọng lớn nhất trong toàn bộ sản bộ sản phẩm. Do đó, ta có thể ngầm kết luận rằng ở các khoảng mức giá bán lẻ từ 20-29.99 chiếm phần lớn người mua nhất và khoảng mức giá từ 1555-1559.99 thì chiếm số lượng người mua là ít nhất.

Ngoài ra, bảng nhỏ phía trên cũng mô tả thông tin các sản phẩm nằm trong khoảng mức giá ấy. Bên cạnh đó, ta cũng có kết luận từ một số thống kê cơ bản của các cột như sau:



* Tổng các cột (count): 3124
* Giá trị trung bình (mean): 73.575585
* Độ lệch chuẩn (std): 84.644969
* Giá trị nhỏ nhất (min): 0.625000
* Giá trị phân vị thứ 25 (25%): 26.658506
* Giá trị phân vị thứ 50 (50%): 48.963529
* Giá trị phân vị thứ 75 (75%): 86.635305
* Giá trị lớn nhất (max): 1557.833333