**BỘ MÔN HỆ THỐNG THÔNG TIN – KHOA CÔNG NGHỆ THÔNG TIN**

**ĐẠI HỌC KHOA HỌC TỰ NHIÊN THÀNH PHỐ HỒ CHÍ MINH, ĐẠI HỌC QUỐC GIA TP HCM**

Mã bài tập: Lab#1.3

Sinh viên thực hiện:

18127206 – Trịnh Văn Tấn

18127174 – Nguyễn Quang Pháp

18127037 – Phan Ngọc Thiên Ân

GV phụ trách: Hồ Bảo Quốc, Hồ Thị Hoàng Vy

HỌC KỲ I – NĂM HỌC 2021-2022

**ỨNG DỤNG PHÂN TÍCH DỮ LIỆU**

Logo

Description automatically generated

|  |  |
| --- | --- |
| **BẢNG PHÂN CÔNG CÔNG VIỆC** | |
| **Trịnh Văn Tấn** | - Mô tả ý nghĩa thuộc tính của tập dữ liệu  **-** Cho biết danh sách các column trong tập dữ liệu  - Mô tả thông tin của từng cột (kiểu dữ liệu, tỷ lệ null, số record…)  - Vẽ biểu đồ quan sát sự phân bố dữ liệu của từng thuộc tính, có outlier không? Nếu có hãy nêu cách xử lý |
| **Phan Ngọc Thiên Ân** | **-** Có tồn tại duplicate data không? Nếu có xử lý loại bỏ dòng trùng  - Phân đều Age thành 5 nhóm độ tuổi sao cho mỗi nhóm chiếm 20% số quan sát  - Vẽ biểu đồ barchart thống kê số lượng khách hàng theo nhóm độ tuổi vừa tạo được |
| **Nguyễn Quang Pháp** | **-** Mô tả thông tin của từng cột: count, min, max, mean…nhận xét  - Vẽ biểu đồ thể hiện sự tương quan giữa các thuộc tính  - Thống kê dữ liệu missing và xử lý  **-** Tính trung bình Annual income theo Gender, theo customer |

1. **Mô tả ý nghĩa thuộc tính của tập dữ liệu**

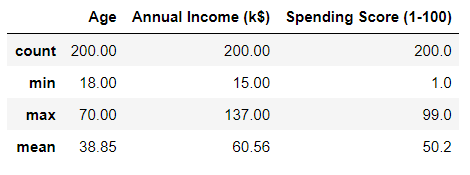
|  |  |
| --- | --- |
| **Tên thuộc tính** | **Ý nghĩa** |
| Customer | STT khách hàng |
| Gender | Giới tính |
| Age | Tuổi |
| Annual Income (k$) | Thu nhập hàng năm (đvt: nghìn đô) |
| Spending Score (1-000) | Điểm chi tiêu (1-100) |

1. **Danh sách các column trong tập dữ liệu**

|  |
| --- |
| **Tên cột** |
| Customer |
| Gender |
| Age |
| Annual Income (k$) |
| Spending Score (1-000) |

1. **Mô tả thông tin của từng cột: count, min, max, mean…nhận xét**

* Thông tin thống kê các cột Age, Annual Income, Spending Score:

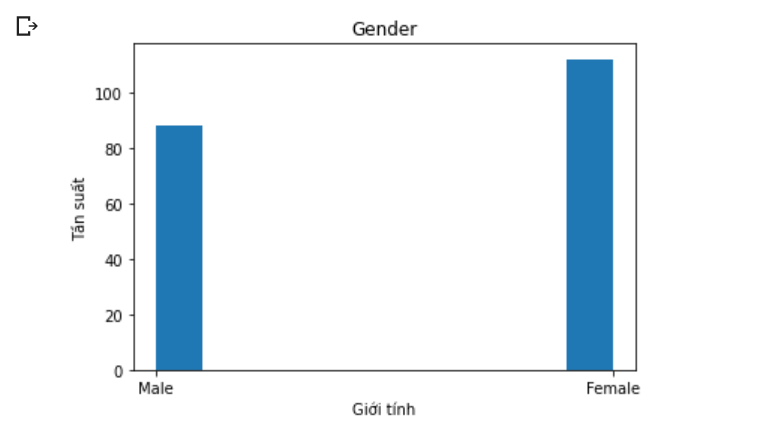


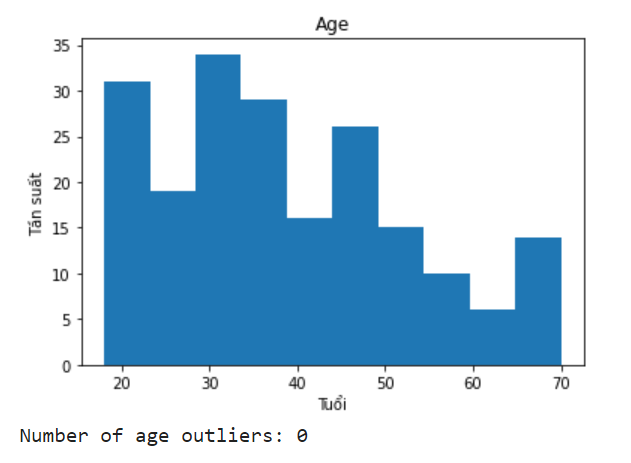
* Nhận xét: Độ tuổi khách hàng thuộc trong khoảng từ 18 đến 70 tuổi, độ tuổi trung bình là 70 tuổi. Thu nhập hằng năm của khách hàng từ 15k$ đến 137k$, trung bình thu nhập là 60.56k$.

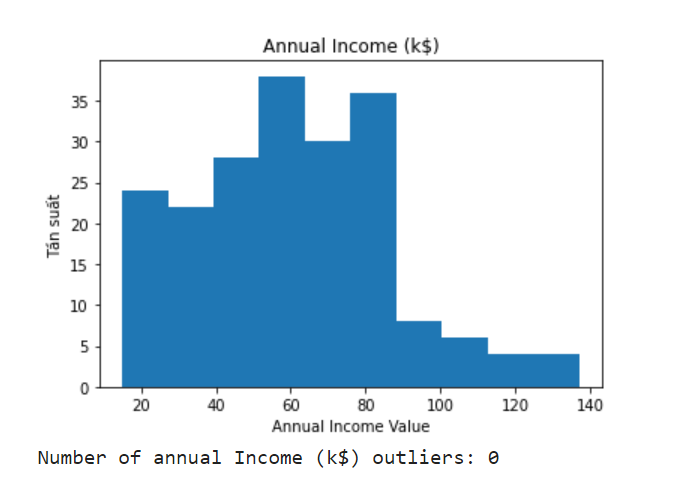
1. **Mô tả thông tin của từng cột (kiểu dữ liệu, tỷ lệ null, số record…)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Tên cột** | **Kiểu dữ liệu** | **Tỷ lệ null** | **Số record** |
| Customer | int64 | 0% | 200 |
| Gender | object | 0% | 200 |
| Age | int64 | 0% | 200 |
| Annual Income | int64 | 0% | 200 |
| Spending Score | int64 | 0% | 200 |

1. **Vẽ biểu đồ quan sát sự phân bố dữ liệu của từng thuộc tính, có outlier không? Nếu có hãy nêu cách xử lý**





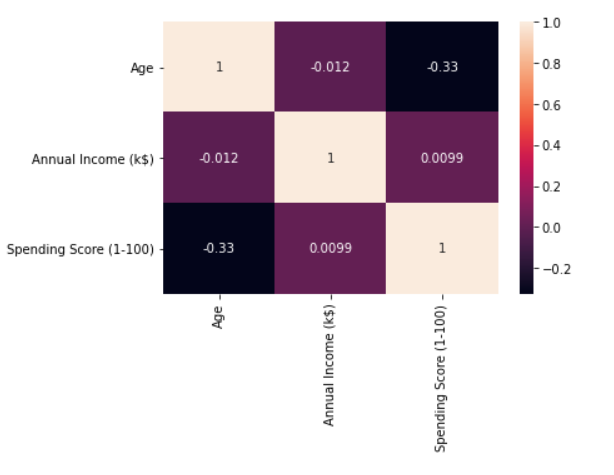




Nhận xét: không có outliers ở những thuộc tính trong dữ liệu Mall Customer

1. **Vẽ biểu đồ thể hiện sự tương quan giữa các thuộc tính:**

* Biểu đồ:

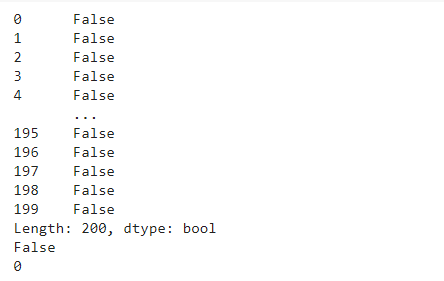


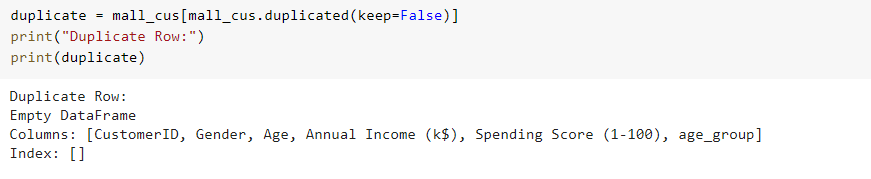
* Nhận xét: dựa vào biểu đồ ta thấy độ tương quan của Age và Spending Score là rất thấp. Trong dữ liệu này chỉ có Annual Income và Spending Score là có sự tương quan nhất.

1. **Thống kê dữ liệu missing và xử lý**

Nhận xét: không có dữ liệu missing.

1. **Có tồn tại duplicate data không? Nếu có xử lý loại bỏ dòng trùng**





-> Nhận xét: Trong bộ dữ liệu đã cho không có thuộc tính trùng lặp.

1. **Phân đều Age thành 5 nhóm độ tuổi sao cho mỗi nhóm chiếm 20% số quan sát**

Sau khi thực hiện gom tuổi vào thành từng nhóm tuổi

Ảnh có chứa bàn

Mô tả được tạo tự động

Số lượng người trong từng nhóm tuổi chiếm xấp xỉ 20%

Ảnh có chứa bàn

Mô tả được tạo tự động

1. **Vẽ biểu đồ barchart thống kê số lượng khách hàng theo nhóm độ tuổi vừa tạo được.**

