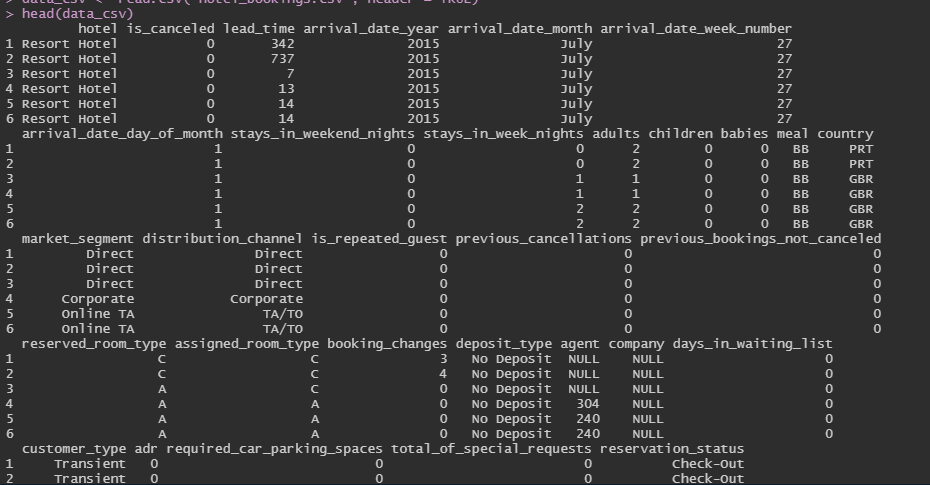
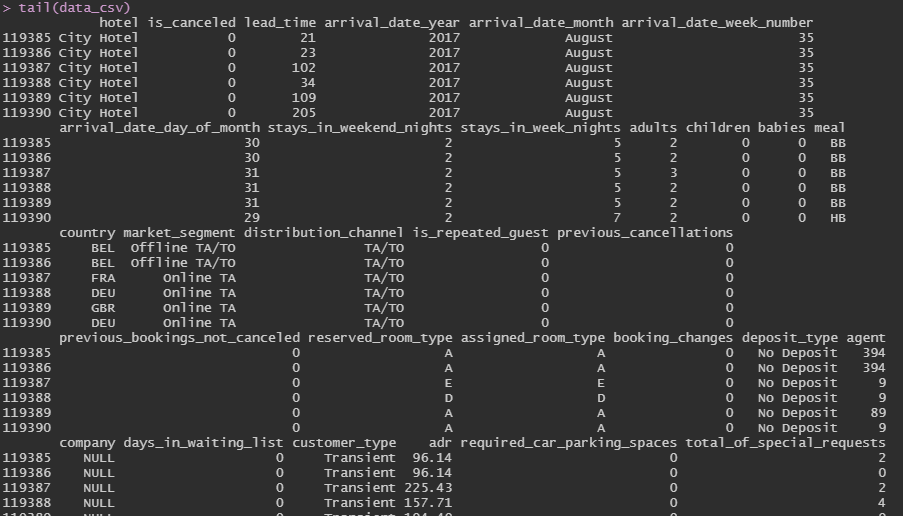
Tên: Lê Nguyễn Anh Nhật

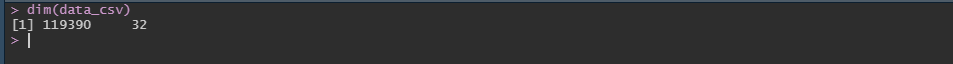
MSSV: 2186400330

Đầu tiên là file hotel\_booking.csv

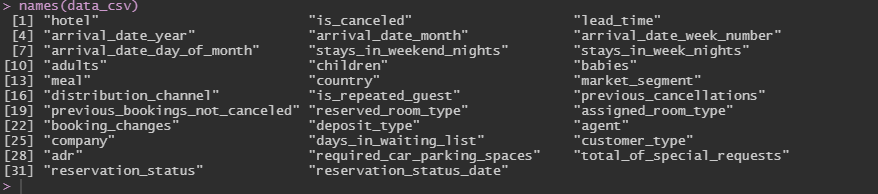




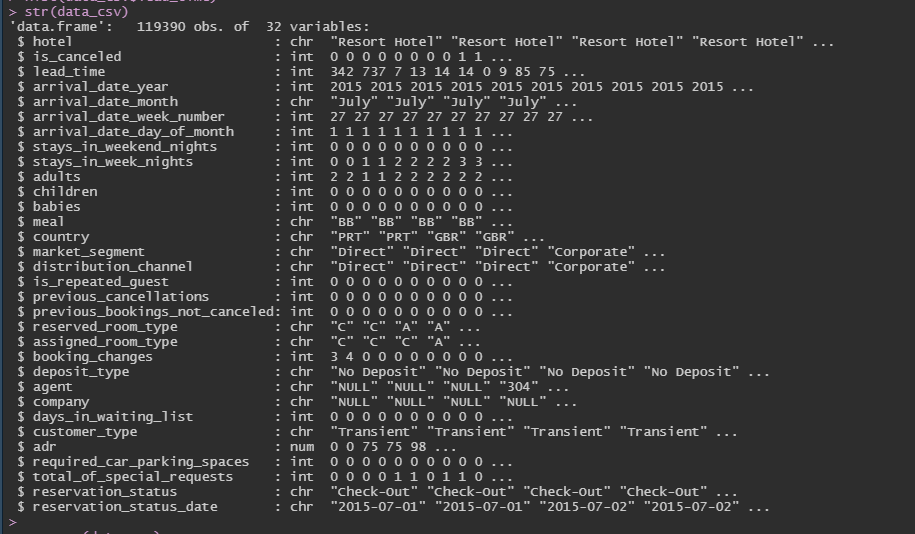
Nhìn vào dữ liệu đầu và cuối vài cột của tập dữ liệu ta thấy rằng dữ liệu này là dữ liệu thu thập về các khách sạn trong đó có những điều quan trọng như là khách hàng của họ là ai, quê quán ở đâu, đi vào ngày tháng nào và số lượng đặt phòng.



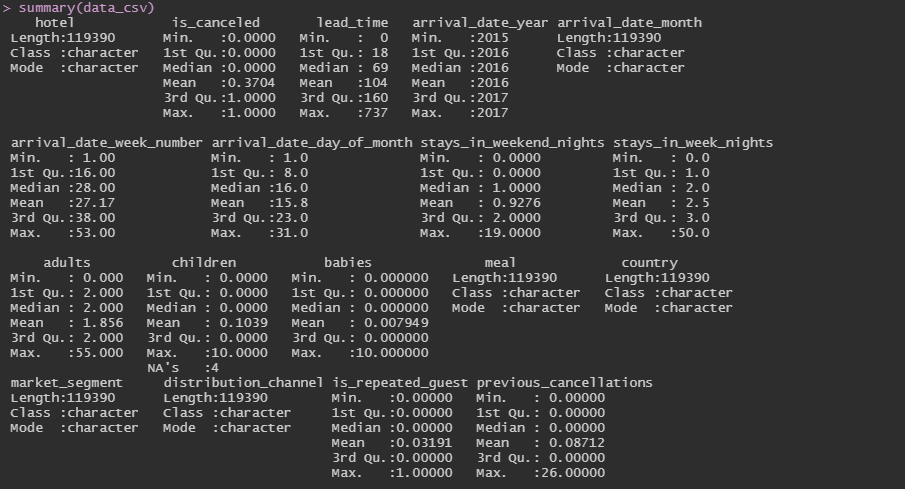
Số lượng của tập dữ liệu có tổng là 119390 và số cột là 32.

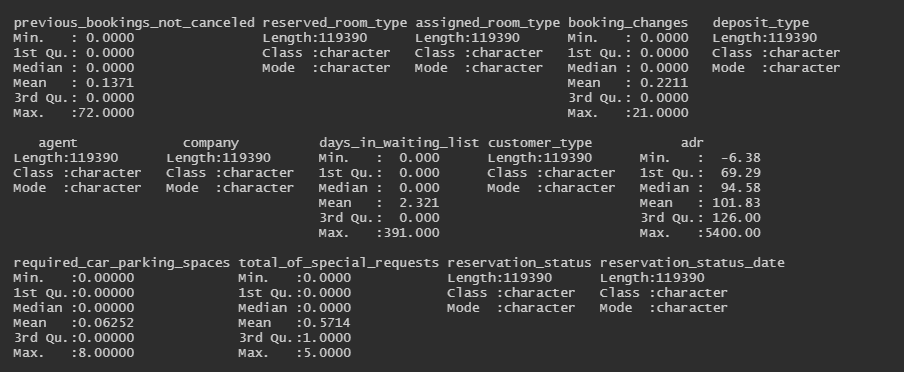


Tên các cột của dữ liệu là rất nhiều và chúng hầu như đều nói về một điều là biết được thời gian đi du lịch khách sạn nào đó, bao nhiêu loại khách hàng như là trường thành, người chưa trưởng thành và trẻ em sơ sinh dữ liệu còn xem xét về quê quán và số lượng đặt phòng.



Hình ảnh thể hiện kiểu dữ liệu đa số các dữ liệu về ngày tháng, tần suất , đối tượng, số lượng booking đều là kiểu int. Còn lại các thông tin như hotel, các biến mang yếu tố phân loại như customer\_type và country đều ở kiểu char.





Bảng tóm tắt ngắn gọn về các cột gồm có Min, median, max, phân vị 1 , phân vị 3 và median.



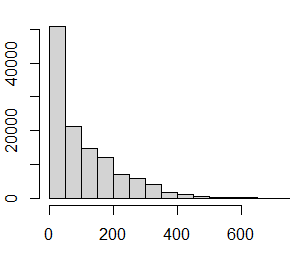
Giá trị trung bình của lead\_time là thời gian khách hàng đặt đến lúc nhận phòng là 104.0114. Và phương sai là 11419.72 cho thấy mức độ phân tán rất cao của lead\_time điều này cho thấy rằng mức độ ở khách sạn một đêm hay vài ngày và đặt trước đi việc du lịch hay lễ là cách nhau rất xa.

+ Dữ liệu có tổng cộng là 4 ô bị thiếu điều này là rất nhỏ so với gần 12 ngàn dữ liệu mà tập dữ liệu đang có.

+Phân vị 25% là 18| 50% là 69|

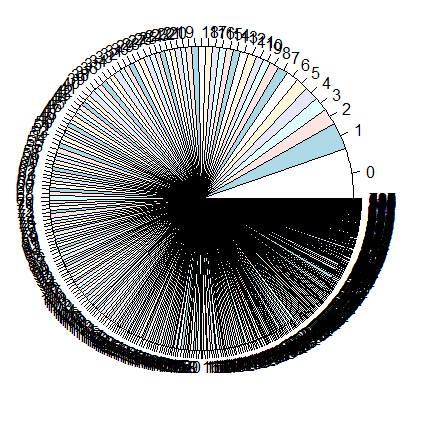
75% là 160 và 100% là 737.





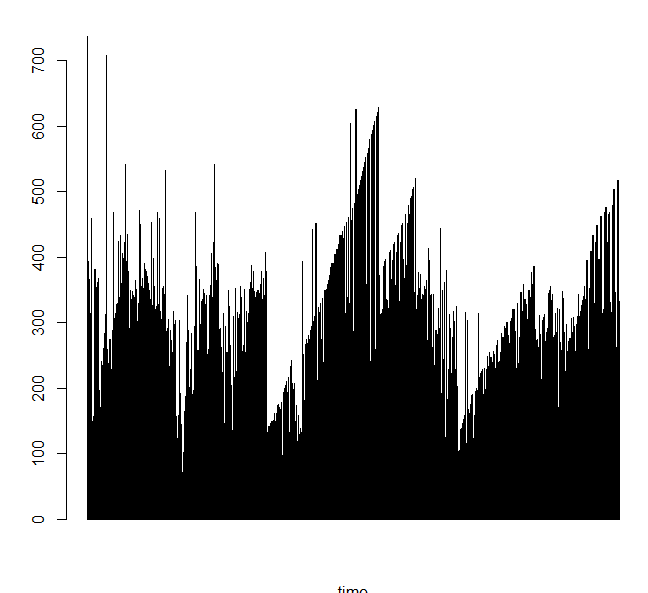
Nhìn vào đồ thị histogram ta thấy rằng khách hàng có xu hướng đặt phòng và nhận sớm là chủ yếu từ 1- 50 ngày là chiếm phân phối nhiều nhất và từ 50-200 ngày nó trở nên đều đặng và xu hướng giảm dần những khách hàng đặt phòng từ 600 ngày có thể là những người ở lâu dài chiếm rất ít và hầu như có thể xem nó là outliner.





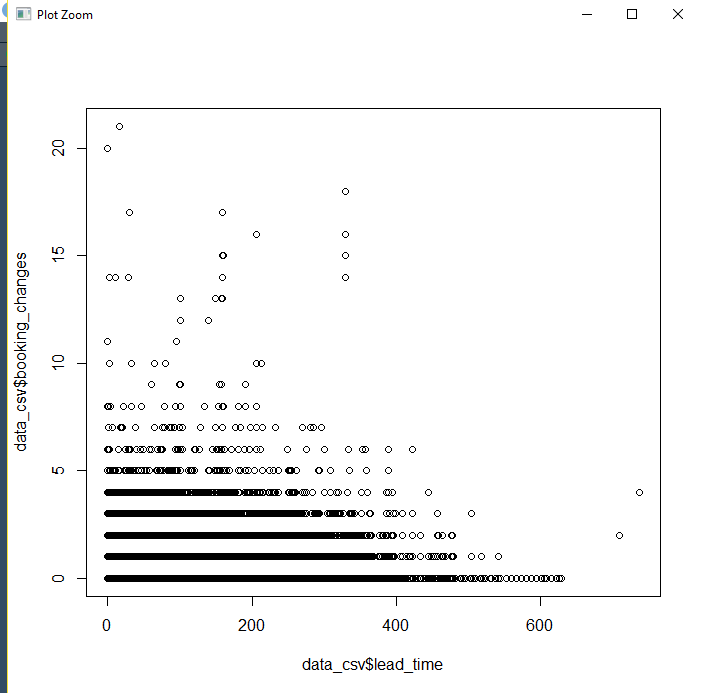
Mặc dù hình bị chèn rất nhiều số nhưng ta có thể thấy được nhu cầu ở khách sạn là rất ngắn đặt và nhận ngay trong vài ngày là chủ yếu do đó cho thấy nó lệch xa rất nhiều so với giá trị trung bình cụ thể ở đây khách hàng thườn đặt và lấy trong từ 1-30 ngày.





Đồ thị cột bổ sung thêm cho thây sự dày đặt của dữ liệu rơi vào khoảng từ 50 đến 150 và sự thưa thớt của số liệu cao hơn.





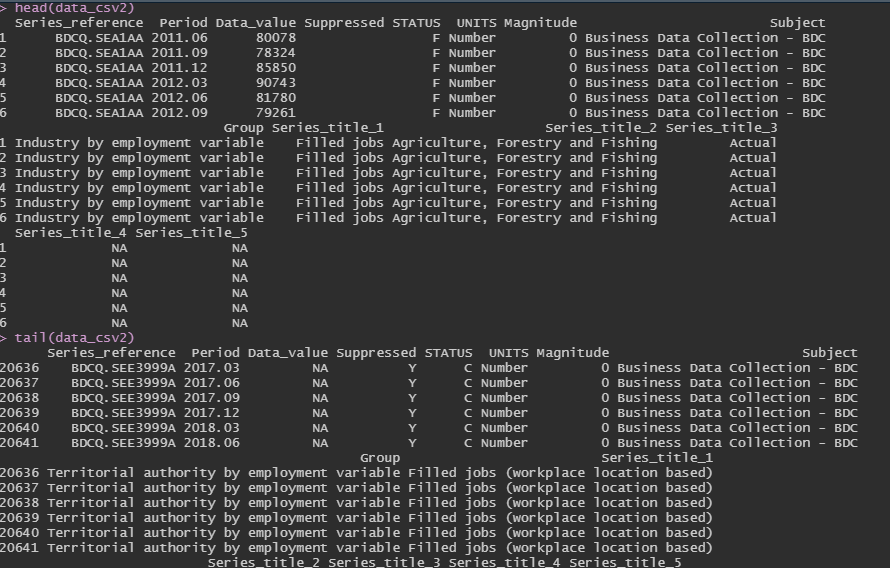
Biểu đồ scatter giữa hai biến là lead\_time và booking\_changes

Cho thấy lượt thay đổi hay hủy đặt phòng rất nhiều ở mức từ 1-5 tập trung dày đặt trong khoảng từ 1-400 ngày nhận phòng các giá trị outliner hầu như không đáng kể và ở mức thưa thớt. Cho thấy khách hàng có xu hướng thay đổ lịch nhận phòng là rất lớn khi trong khoảng 200 đến 400 ngày.

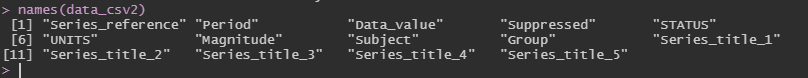
Dữ liệu của Business-employment-data-september-2022-quarter-csv.csv



Các dữ liệu đầu và cuối của data này.



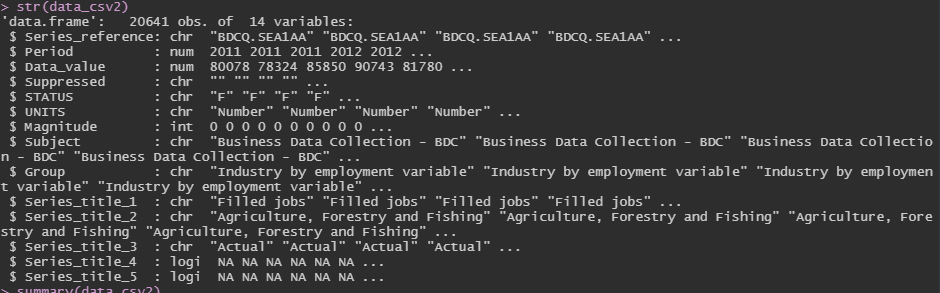
Nhìn vào dữ liệu ta thấy nó đang nói về dữ liệu nhân sự có mã nhân sự của phòng ban nào đó group nào và ngày tháng vào công ty và data values có thể là giá trị nhân sự hay group tạo ra cho công ty khi nhân sự còn làm việc.



Tên các cột thể hiện sự hiệu quả của nhân viên kinh doanh gồm các thông tin của nhân viên mã hay là ngày vào làm những đóng góp có giá trị gì, …



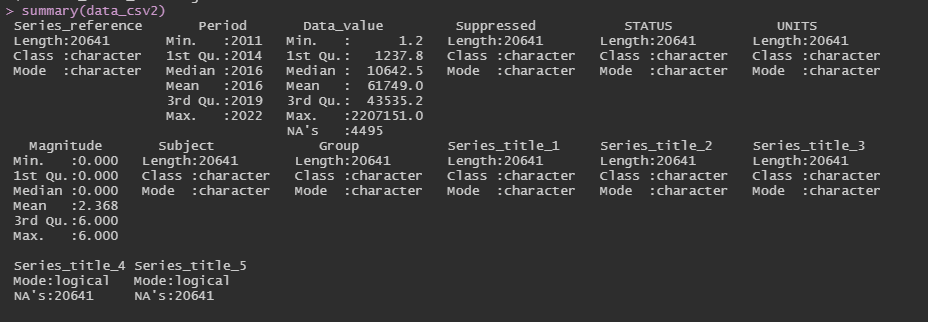
Có tổng là 20641 dữ liệu và trong đó có 14 cột



Các cột như peroid, data\_value, Magnitude là kiểu int

Có Series\_title\_4 và Series\_title\_4 là kiểu logical.

Còn lại tất cả là kiểu char



Bảng tóm tắt ngắn gọn về các cột gồm có Min, median, max, phân vị 1 , phân vị 3 và median.

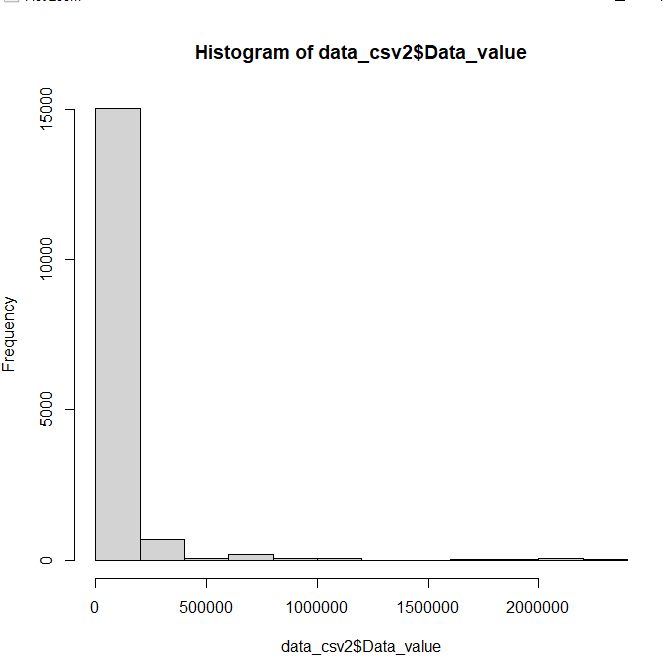
Ở đây ta quan tâm giá trị nhân sự tạo cho công ty trung bình là 61749 USD, có 4495 dữ liệu bị khuyết.



TỔng số lượng giá trị bị thiếu à 45777 một phần rất lớn dữ liệu đã bị khuyết.

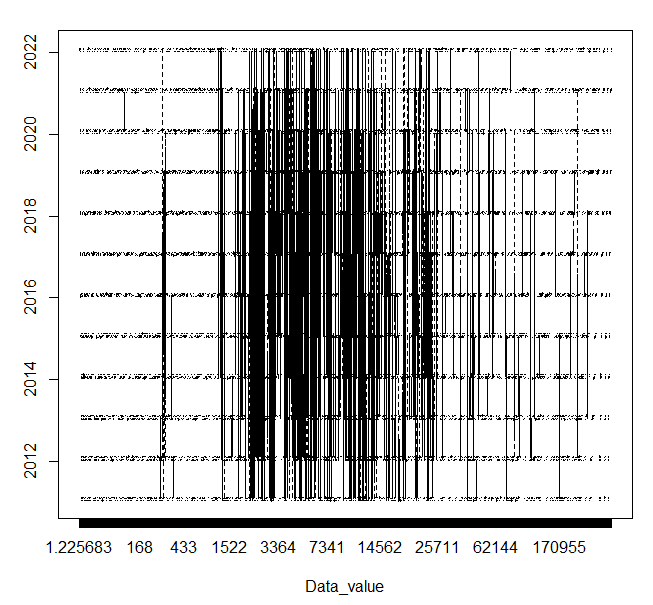
Và vì số lượng dữ liệu bị khuyết của data\_values chiếm gần 1/5 tổng dữ liệu nên việc kiểm định hay tính phương sai có thể gây ra thiết xót rất nhiều khi phân tích.



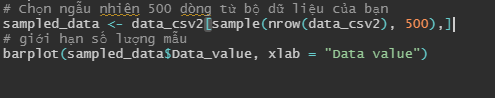


Đồ thị Histogram của Data\_values cho thấy rằng dữ liệu có phân phối tập trung rất nhiều vào khoản từ 0-150000 và điều này cho thấy sự trên không quá nhiều vẫn xoay quanh rất gần giá trị trung bình và các outlier không quá nhiều để ảnh hưởng đến dữ liệu, đồ thị hist này còn cho ta thấy dữ liệu từ Data\_values có thể mang phân phối chuẩn.

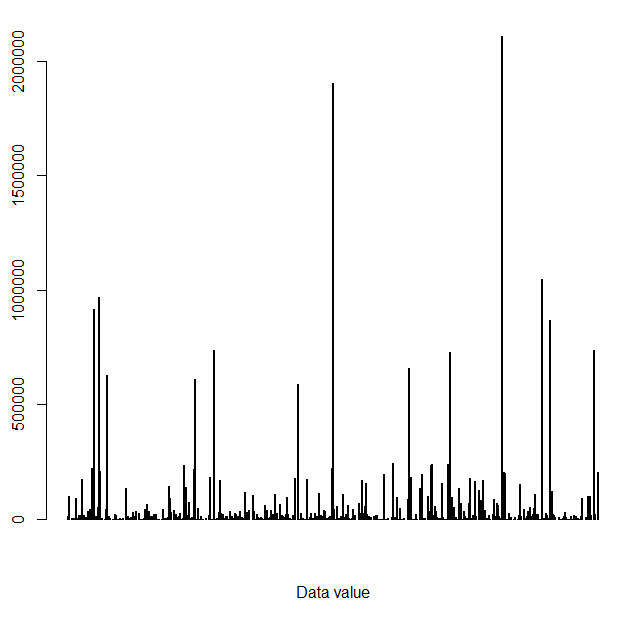




Đồ thị boxplot qua các năm và giá trị của nhân sự kinh doanh hầu như đều trải dài từ năm này đến năm khác không có quá nhiều sự tiến bộ đi lên của giá trị nhân sự.



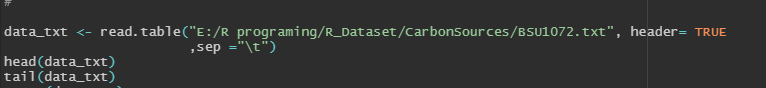
Đồ thị bar sau khi giới hạn lại dữ liệu

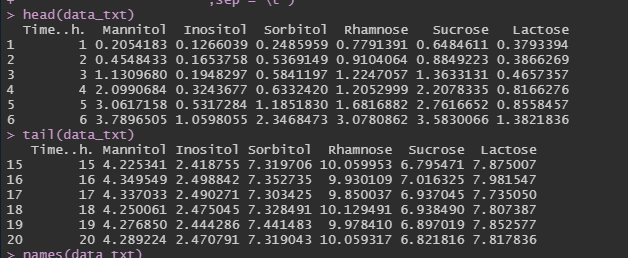


Nó vẫn thể hiện đầy đủ của dữ liệu ở mức dưới 500000 USD tạo ra là quá nhiều còn những người trên 500000 USD chiếm số lượng quá nhỏ khi nhìn vào đây ta thấy dự biến động của dữ liệu là rất ít.

Làm với 1 file từ thư mục CarbonSources

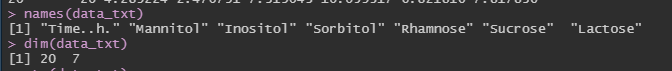
Ở đây ta sẽ chọn file BSU1072.txt từ thư mục CarbonSources





Nhìn vào dòng đầu và cuối của dữ liệu ta thấy dữ liệu này có thể đang nói lên một phản ứng nào đó có thể là thuốc chữa bệnh theo thời gian có thể dùng dữ liệu này để phân tích sinh học hay dược lý

Và bản trên là thể hiện nồng độc của đường có thể là ở cơ thể con người khi tác dụng với thuốc

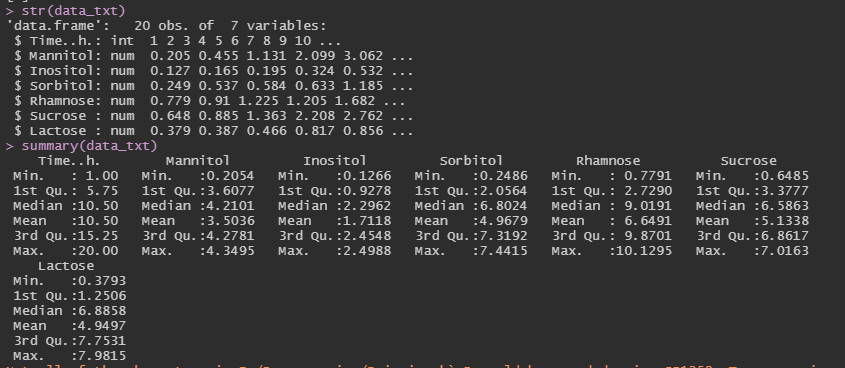


Cột có Time..h là thời gian tác dụng thuốc, Mannitol, Inositol, Sorbitol, Rhamnose, Sucrose, Lactose đều là đường do tác dụng của thuốc tạo ra để làm việc gì đó.

Và số lượng dữ liệu là 20 và có tổng là 7 cột.



Không có cột nào bị khuyết dữ liệu, dữ liệu này được thu thập đầy đủ .



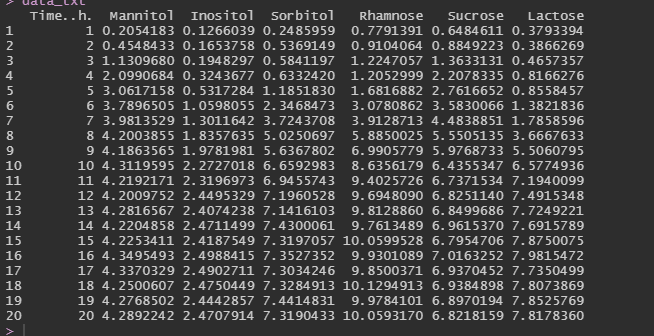
Toàn bộ các biến về đường đều mang kiểu dữ liệu numerical còn riêng biến thời gian nó được tính theo giờ nên dữ liệu là int.

Khi xem xét bảng summary xong ta thấy rằng dữ liệu của các cột về đường như Mannitol, Sorbitol,… có giá trị mean xoay quay giá trị max và min lệch không quá nhiều dữ liệu có mức ổn định khá cao

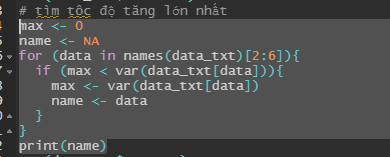


Giá trị phương sai 8.39 đối với giá trị trung bình này cho thấy mức biến động của dữ liệu không quá nhiều

Vì dữ liệu ít



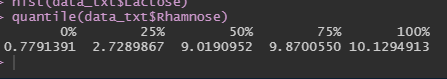
Nên khi nhìn vào nó ta thấy các biến tăng cùng nhau và chúng có thể hiện mối quan hệ tuyến tính thuận.





Khi xem xét dữ liệu trên muốn xem dữ liệu nào có tính biến động lớn nhất và tốc độ lớn nhất trong 6 biến về đường khi kiểm tra xong ta thấy rằng Rhamnose có tốc độ tăng cao nhất trong 6 biến về đường của tập dữ liệu (sữ dụng phương sai để xem).

Bây giờ ta sẽ phân tích Rhamnose



Kết quả phân tích quantile của biến Rhamnose trong dữ liệu cho thấy:

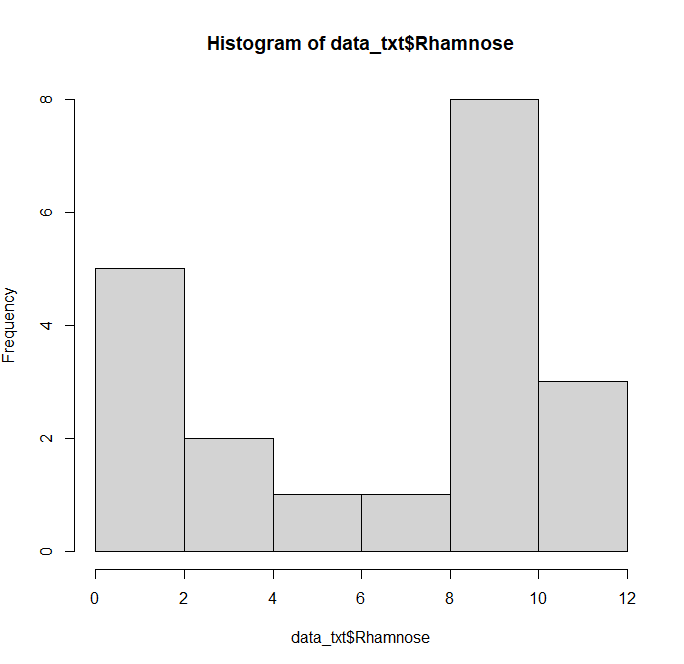
+Giá trị thấp nhất (0%) của Lactose là 0.779.

+Khoảng giá trị từ 25% đến 75% (tức là giá trị trung tâm) của Lactose là từ 2.7 đến 9.8 .

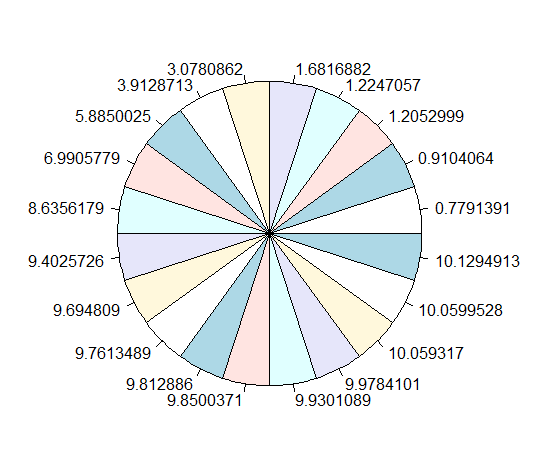
+Giá trị cao nhất (100%) của Lactose là10.12.

Điều này cho thấy:

Hầu hết các giá trị của biến Rhamnose tập trung trong khoảng từ 2.7 đến 9.8.

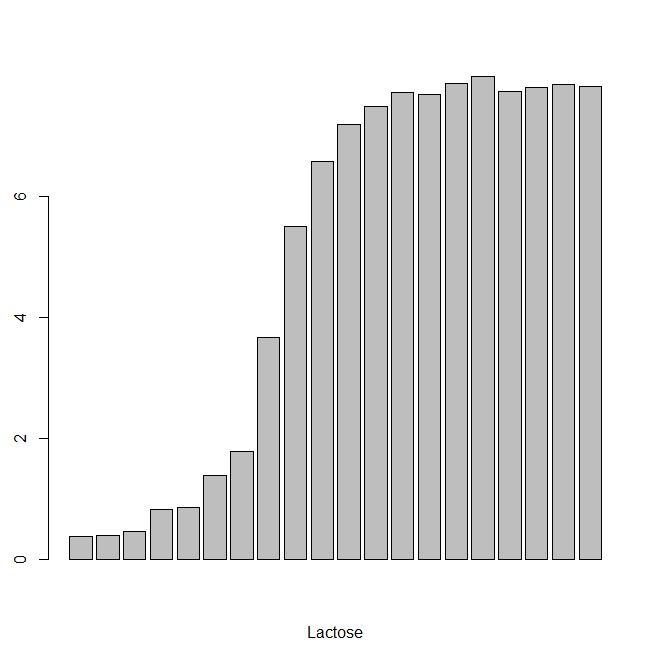


Đồ thị cho thấy phân phối tập trung nhiều vào từ 8 -10 của biến Rhamnose cho thấy khi liều thuốc có tác dụng càng lâu thì loại đường này sẽ tập trung ở mức 8-10 vào lúc đầu tác dụng thuốc sẽ ở mức 0-4 ở mức ổn định lâu dần sẽ tăng lên và có dấu hiệu bảo hòa về lượng đường.



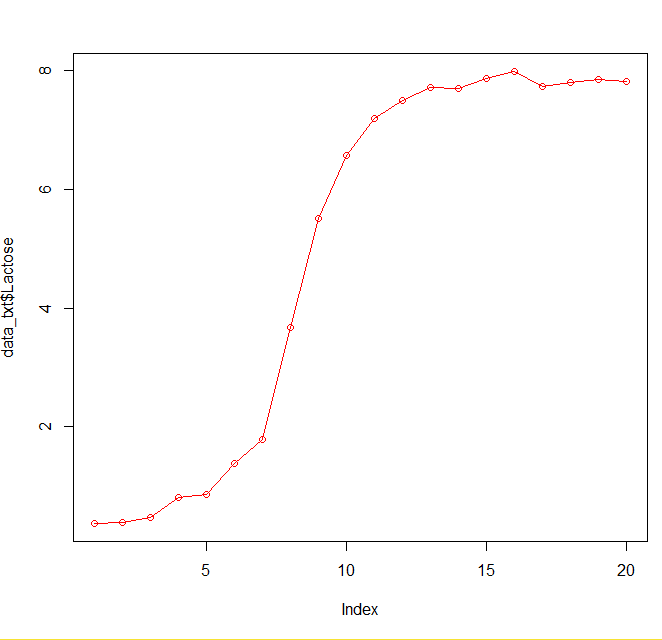
Sự biểu hiện của việc tăng lượng đường Rhamnose khi nhìn vào đây ta thấy chúng chiếm tỉ lệ là như nhau theo thời gian kết hợp với đồ thị hist ở trên ta thấy rõ việc cứ theo mối giờ sẽ có khoảng thời gian lượng đường giảm để cung cấp cho cơ thể khi dùng thuốc và chúng chiếm mỗi giờ là như nhau





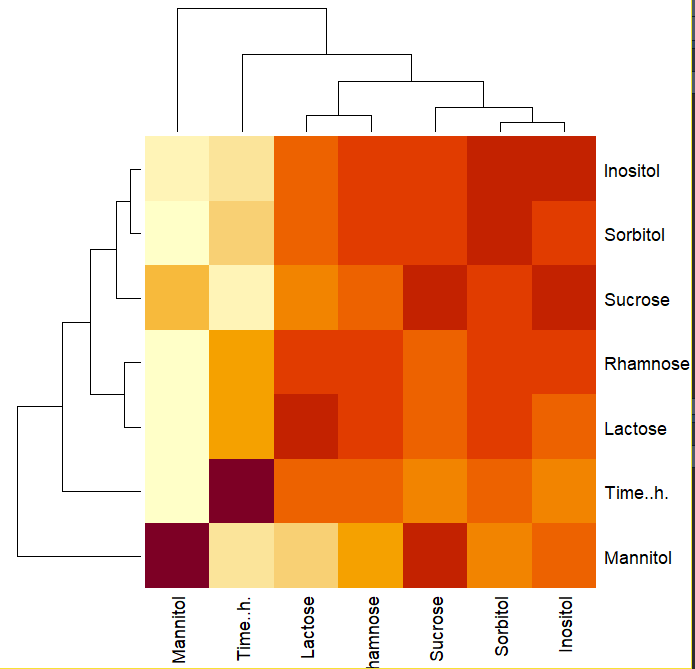
Biểu đồ cột cho ta biết rằng cứ giờ tăng thì lượng đường lactose sẽ tăng và cũng giống như Rhamnose thì sẽ có khoảng thời gian mà lượng đường ở mức bảo hòa tăng giảm hầu như quá ít.





Đồ thị đường cho thấy sự đi lên của đường lactose theo thời gian đến ở ngưỡng 7.6 tốc độ lên có phần chậm lại và sau đó là ở mức chung chung là 7.9 hay 8 càng về sau có phần đi xuống.

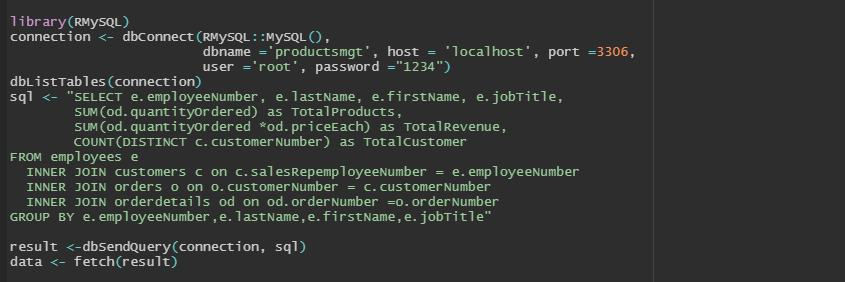




Biểu đồ heatmap biểu thị tính tương quan của dữ liệu nhìn vào ta thấy được các điểm đỏ mang tính tương quan mạnh với nhau như:

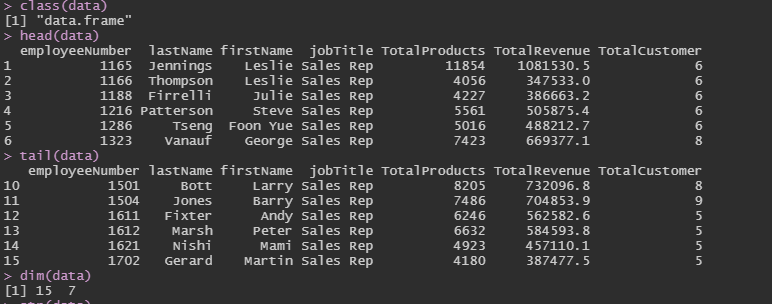
Sorbitol và inositol, inositol với Sucrose.

Tiếp theo phân tích Cơ sở dữ liệu mysql với database là productsmgt



Kết nối và truy vấn đến Mysql

Câu lệnh SQL để lấy ra thông tin nhân viên và tổng số đơn hàng, tổng số khách hàng, và tổng doanh thu mà nhân viên đó kiếm được.



Dữ liệu được lấy ra và đưa vào data frame khi nhìn vào dữ liệu head và tail ta thấy rằng chúng cho biết mã nhân viên, tên, họ, vị trí công việc, tổng sản phẩm bán được, tổng giá trị đơn hàng và tổng số khách hàng nhân viên đó bán được.

Số lượng dữ liệu là 15 và có tổng là 7 cột.



Có các cột là employeeNumber, lastName, firstName, jobTitle, TotalProducts, TotalRevenue, TotalCustomer.



Dữ liệu của employeeNumber thuộc kiểu int.

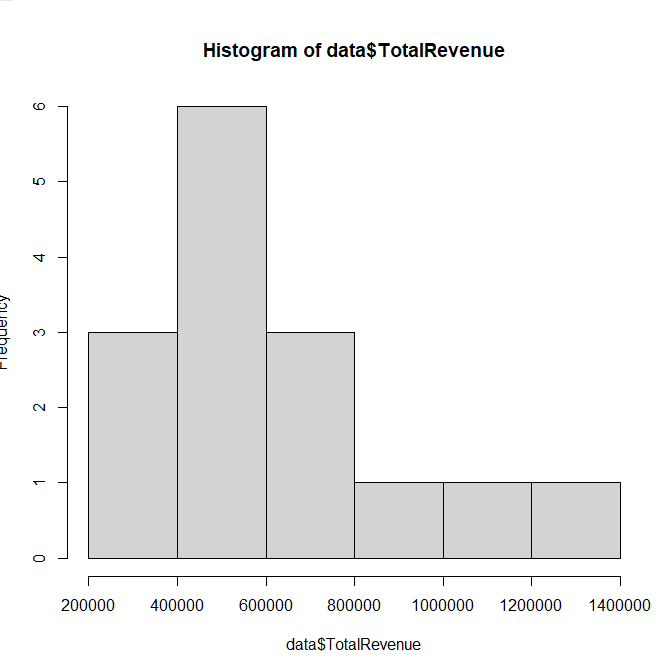
Dữ liệu của lastName, firstName, jobTitle thuộc kiểu char.

Dữ liệu của TotalProducts, TotalRevenue, TotalCustomer thuộc kiểu numerical.

Nhìn vào bảng Summary không có dữ liệu bị khuyết cột TotalProducts, TotalRevenue có giá trị trung bình đều gần trong khoảng phân vị 1 và 3. Cho thấy dữ liệu xoay quay giá trị trung bình rất nhiều.

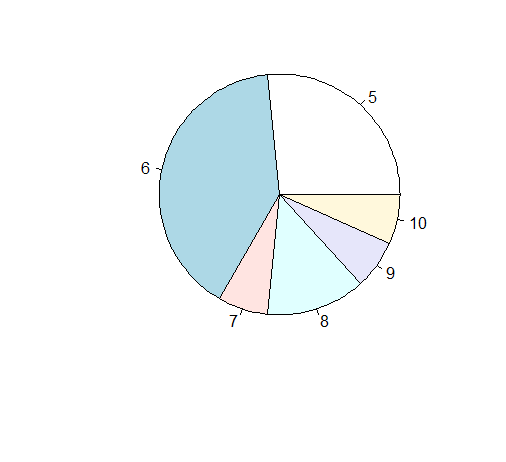


Hai giá trị phương sai của TotalProducts, TotalRevenue ở mức biến động cao.



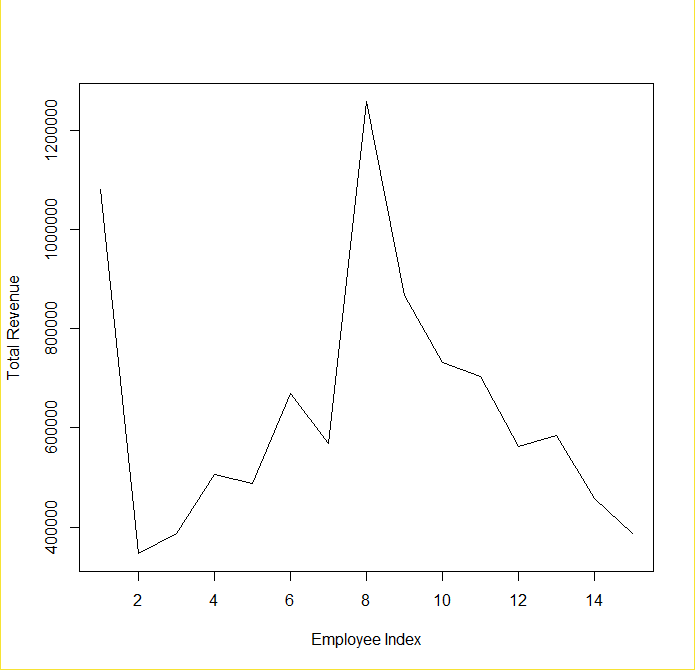
Biểu đồ hist của TotalRevenue thể hiện phôi phối dữ liệu tập trung nhiều từ 400000-600000 và từ 200000-400000 và từ 600000-800000 nó ở có xự đối xứng với nhau và càng về giá trị đơn cao hơn thì số nhân viên bán được chỉ là 1.



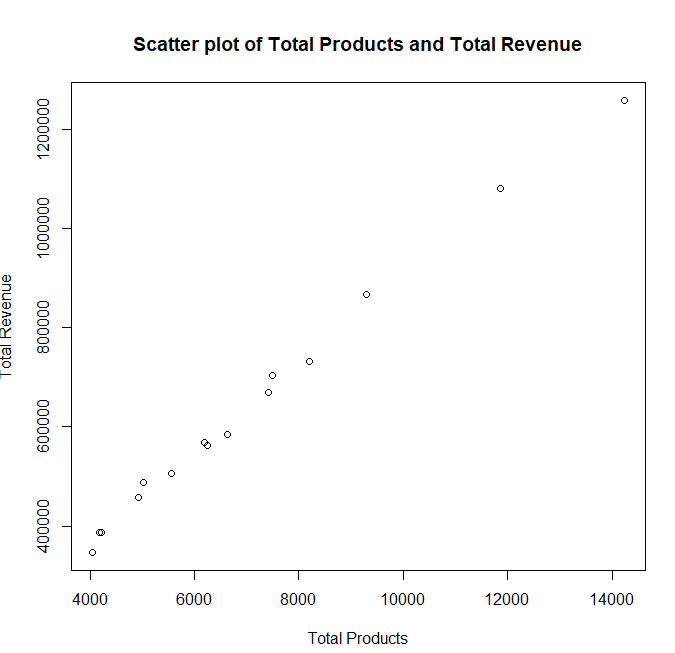


Tổng số khách hàng chiếm tỷ lệ nhiều nhất là 6 người chiếm đên 8 người cũng tương đối nhiều trong tổng số 15 nhân viên của tập dữ liệu.

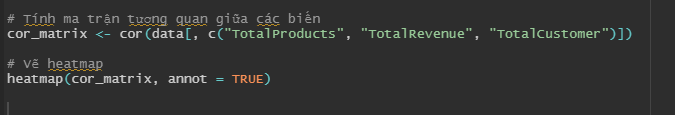


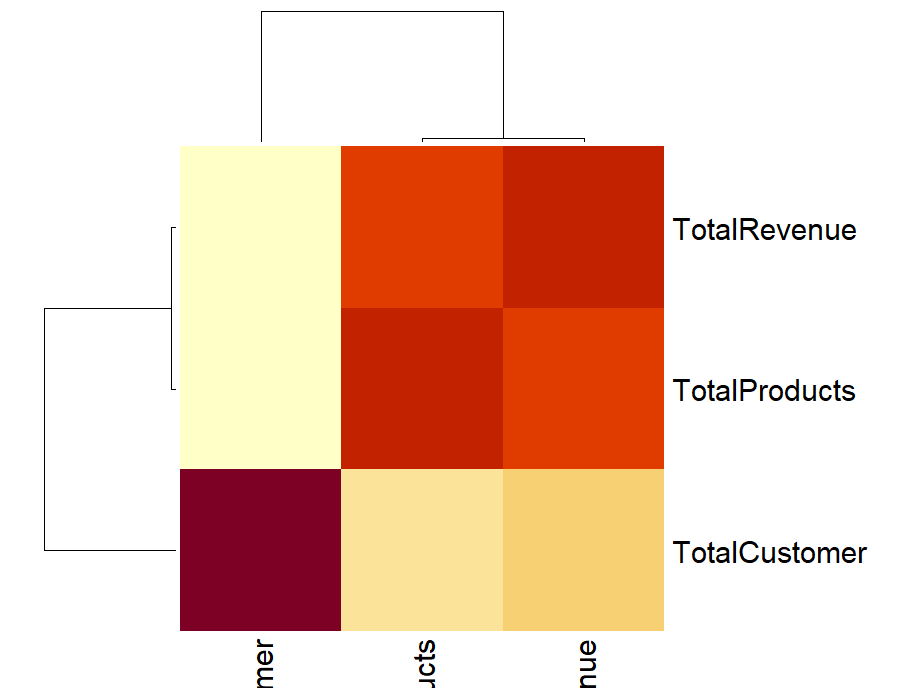


Biểu đồ đường cho thấy rằng Các nhân viên khác nhau có sự lên xuống đều đặng với giá trị đơn hàng chứ không mang tính tuyến tính nó có tính rời rạc.



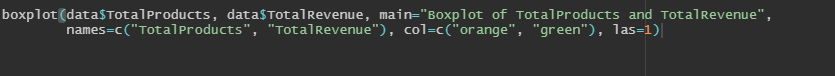
Đồ thị scatter TotalReneveu và TotalProduct cho thấy sự gia tăng sản phẩm thì sẽ gia tăng tổng doanh thu thể hiện sự tuyến tính thuận của hai biến này.

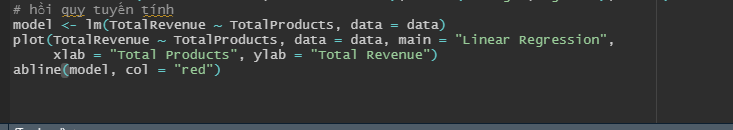


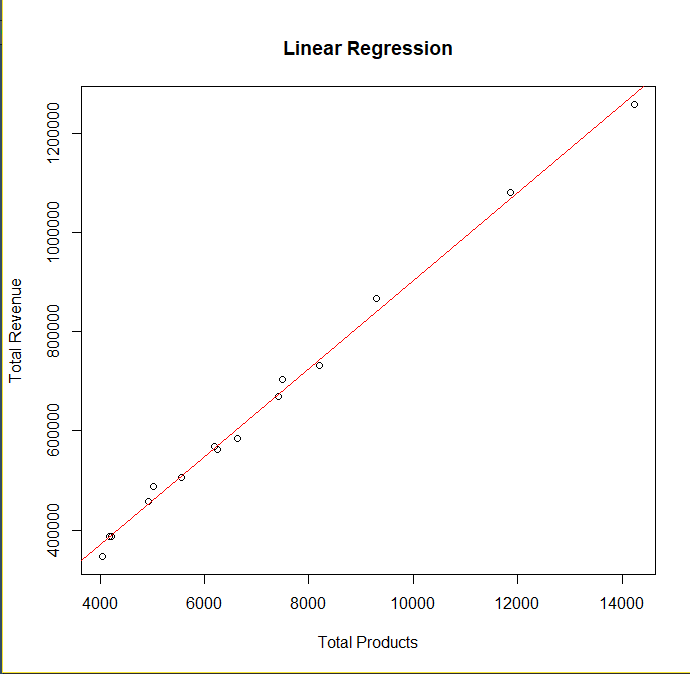


Nhìn vào heatmap ta thấy rằng hai biến TotalCustomer và TotalRenevue không có sự tương quan đến nhau ngược lại TotaRenevue có sự tương quan mạnh với TotalProducts.

Cho thấy sự gia tăng khách hàng không đồng nghĩa với việc gia tăng doanh thu.







Các biến này đều gần như nằm gần đường thẳng hồi quy và không có giá trị outlier nào xuất hiện ở đây