**­­**

**BỘ MÔN PHẦN CỨNG & MẠNG MÁY TÍNH**

----------- 🕮 ----------



**BÁO CÁO TỰ TÌM HIỂU**

**Lab 6: Data Analyst**

**………………………………………….…………**

**.....................................................................................**

**Giảng viên hướng dẫn : HỒ DIÊN TUẤN ANH**

**Nhóm thực hiện: SELF – DISCIPLINE**

**Các thành viên nhóm:**

**+ TRẦN THỊ NGỌC CẨN- MSSV: 0306221404**

**+ NGUYỄN KHẮC NGUYÊN– MSSV: : 0306221449**

**+ NGÔ TRỌNG NGHĨA – MSSV: 0306221443**

**Sinh viên thực hiện: TRẦN THỊ NGỌC CẨN**

**Lớp : CĐ TH 22DĐE**

**Khoá : 2022 – 2025**

**TP. HỒ CHÍ MINH, tháng 5 năm 2024**

**MỞ ĐẦU**

Trong khuôn khổ học phần môn Lập trình Python, nhóm chúng em đã thực hiện báo cáo tìm hiểu với chủ đề Data AnaLyst.

Mục tiêu chính của đồ án này là áp dụng các kiến thức về lập trình Python vào việc xây dựng một ứng dụng khả thi nhằm giải quyết nhanh việc báo cáo dữ liệu thông tin, tổ chức phân tích và vẽ biểu đồ trực quan biểu diễn các số liệu.

Trong thời đại số như hiện tại thì việc thực hiện phần tích và giải quyết các số liệu trong các văn phòng là vấn đề trực quan nhất cần quan tâm. Vì hầu như tất cả thông tin đều được biến thành số liệu, vấn đề của công ty là dựa vào các số liệu mà đánh giá tình hình thị trường đưa ra giải pháp cho các vấn đề như giá cả, tình hình mua sắm của khách hàng, phân tích số lượng mua tăng hay giảm, doanh thu đạt về có cần cải thiện điều gì và phát huy điều gì để mang lại hiệu quả khách hàng tốt nhất.

Phân tích từ các con số đưa ra sơ đồ và chiến lược bán hàng phù hợp với thị trường và khách hàng. Đưa công ty hay nhà phát triển vững mạnh.Quá trình thực hiện đồ án đã giúp nhóm chúng em hiểu sâu hơn về các khái niệm, kỹ thuật và công cụ trong lập trình Python, đồng thời rèn luyện kỹ năng phân tích, thiết kế và triển khai một dự án phần mềm.

Trong bài báo cáo này, chúng em sẽ trình bày chi tiết về quá trình thiết kế, triển khai và kết quả của báo cáo từ quá trình thực hiện.

**LỜI CẢM ƠN**

Kính thưa thầy Hồ Diên Tuấn Anh,

Thay mặt nhóm, chúng em xin chân thành cảm ơn thầy đã hướng dẫn và giúp đỡ

chúng em trong quá trình thực hiện báo cáo môn Lập trình Python.

Chúng em đã học hỏi rất nhiều kiến thức và kỹ năng quý báu từ môn học này, bao gồm cả các kiến thức lập trình nâng cao cũng như các kỹ năng phân tích, thiết kế và triển khai hệ thống phần mềm bằng ngôn ngữ Python. Những kiến thức và kỹ năng này sẽ rất hữu ích cho chúng em trong công việc và nghiên cứu sau này.

Bài báo cáo này đã giúp chúng em áp dụng và thực hành những lý thuyết đã học, từ đó nâng cao hiểu biết và kỹ năng lập trình Python. Quá trình thực hiện báo cáo cũng giúp chúng em rèn luyện khả năng làm việc nhóm, giao tiếp và trình bày kết quả.

Trong bài không thể thiếu những điều sai sót, vì vậy nếu thầy có ý kiến và đóng góp xin hãy liên hệ cùng chúng em. Chúng em xin chân thành nhận ý kiến và chỉnh sửa để bài làm được hoàn chỉnh nhất có thể.

Một lần nữa, chúng em xin chân thành cảm ơn thầy vì sự hướng dẫn, góp ý và hỗ trợ quý báu trong suốt quá trình thực hiện đồ án.

Trân trọng,

SELF – DISCIPLINE

Người thực hiện báo cáo: Trần Thị Ngọc Cẩn

1. **Công nghệ sử dụng trong bài làm:**
2. AnaConda:

Anaconda là một nền tảng phân phối miễn phí của ngôn ngữ lập trình Python và R cho tính toán khoa học (khoa học dữ liệu, machine learning, xử lý dữ liệu lớn, phân tích dự đoán, v.v.), nhằm mục đích đơn giản hóa việc quản lý và triển khai gói. Nó có trên cả Windows, MacOS và Linux.

1. Os (Operating System):

**Module os trong Python** cung cấp các chức năng được sử dụng để tương tác với hệ điều hành và cũng có được thông tin liên quan về nó. OS đi theo các Module tiện ích tiêu chuẩn của Python. Module này cung cấp một cách linh động sử dụng chức năng phụ thuộc vào hệ điều hành.

1. Pandas:

Pandas là một **thư viện phần mềm** được viết riêng cho ngôn ngữ lập trình Python để thực hiện các thao tác quản lý và và phân tích dữ liệu. Đặc biệt, nó cung cấp các cấu trúc dữ liệu và các phép toán để thao tác với các bảng số và chuỗi thời gian.

1. Matplotlib:

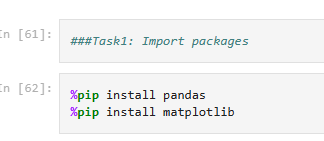
Matplotlib là một trong những thư viện Python phổ biến nhất được sử dụng để trực quan hóa dữ liệu. Nó là một thư viện đa nền tảng để tạo các đồ thị 2D từ dữ liệu trong các mảng. Matplotlib được viết bằng Python và sử dụng NumPy, phần mở rộng toán học của Python.

1. **Database:**

Mô tả: Dữ liệu được thu thập 12 tháng năm 2019 của một công ty.

Nguồn: <https://github.com/SuichiDaito/Report_DataAnaLyst>

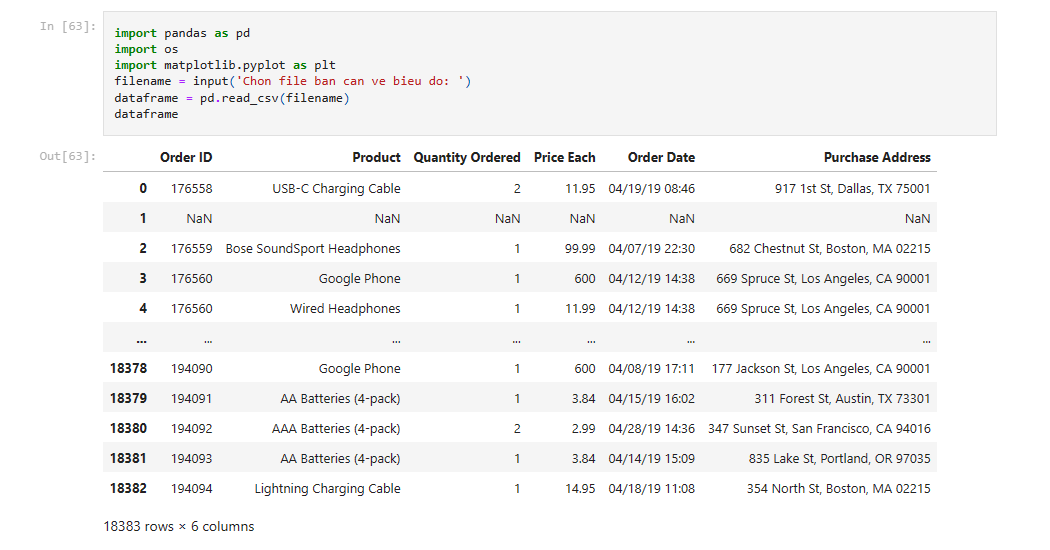
1. **Kết quả làm được:**
2. **Cài đặt Pandas và Matplotlib:**

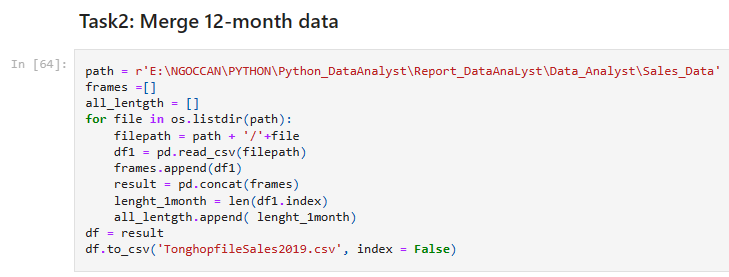
Có thể cài đặt bằng cmd của AnaConda hoặc sử dụng %pip của cmd máy tính hoặc Terminal.

1. **Import thư viện và tiến hành đọc file:**

Trong pandas có hai cấu trúc quan trọng là: **DataFrame và Series**  và trong bài này thì chúng em sử dụng DataFrame chính cho việc đọc dữ liệu từ file.

Và trong pandas có thể đọc rất nhiều định dạng file như: \*.csv,\*.json,\*.html,\*.xml …v.v…Tuỳ vào từng dạng file thì có từng kiểu hàm đọc tương ứng sử dụng câu lệnh read\_csv(filename) và in dataframe ra ngoài để kiểm tra.

Và kết quả trả về là một bảng trong file truyền vào trong hàm.

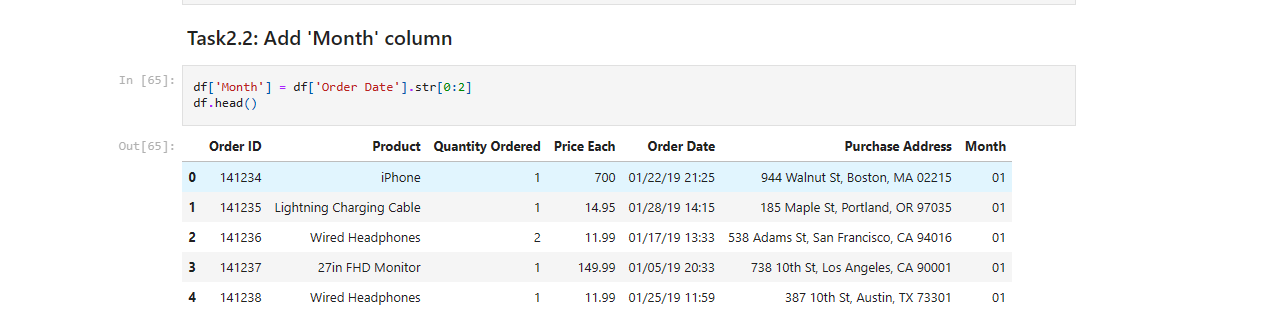
1. **Gộp tất cả các file \*.csv có trong thư mục Sales\_Data có trong 12 tháng thành 1 file tổng hợp để xử lí.**

Biến path dùng để lưu trữ đường dẫn tuyệt đối từ root đến file chứa các file csv. Sử dụng os để thực hiện đọc từng file có trong thư mục mà path trỏ tới.

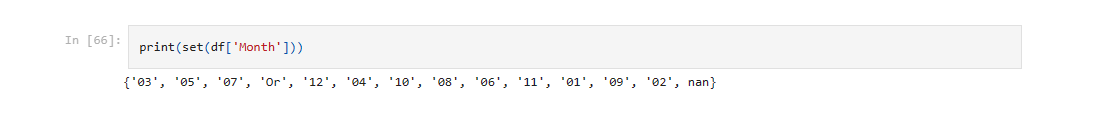
Dùng 1 list frame để lưu trữ tên các file có trong thư mục Sales\_Data. Và dùng hàm concat() để nối các nội dung có trong các đường dẫn của từng file thành 1 file tổng hợp tên **“TonghopfileSales2019”** .

1. **Thêm cột “Month” vào trong file tổng hợp và lấy tháng ra để dễ phân tích và so sánh.**

Có rất nhiều cách để cắt ra tháng trong file dữ liệu.Nhưng ngắn gọn nhất là sử dụng str[0:2]. Vừa tạo cột “Month” vừa thêm giá trị vừa cắt vào bên trong. Và dùng hàm head() để in ra kết quả.

****

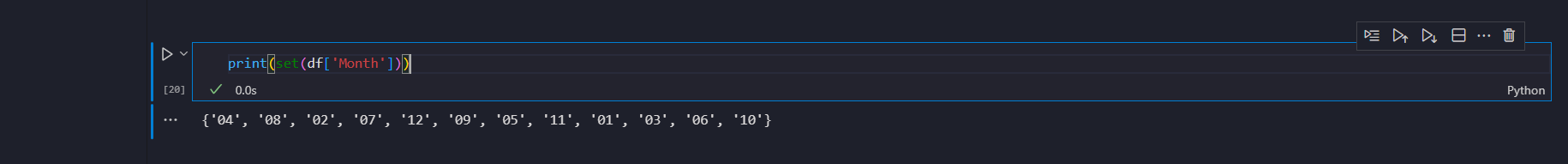
1. **Kiểm tra các giá trị trong cột Month có hợp lệ.**

****

Khi in ra sẽ có hai giá trị không hợp lệ. Là “Or” và “Nan”.

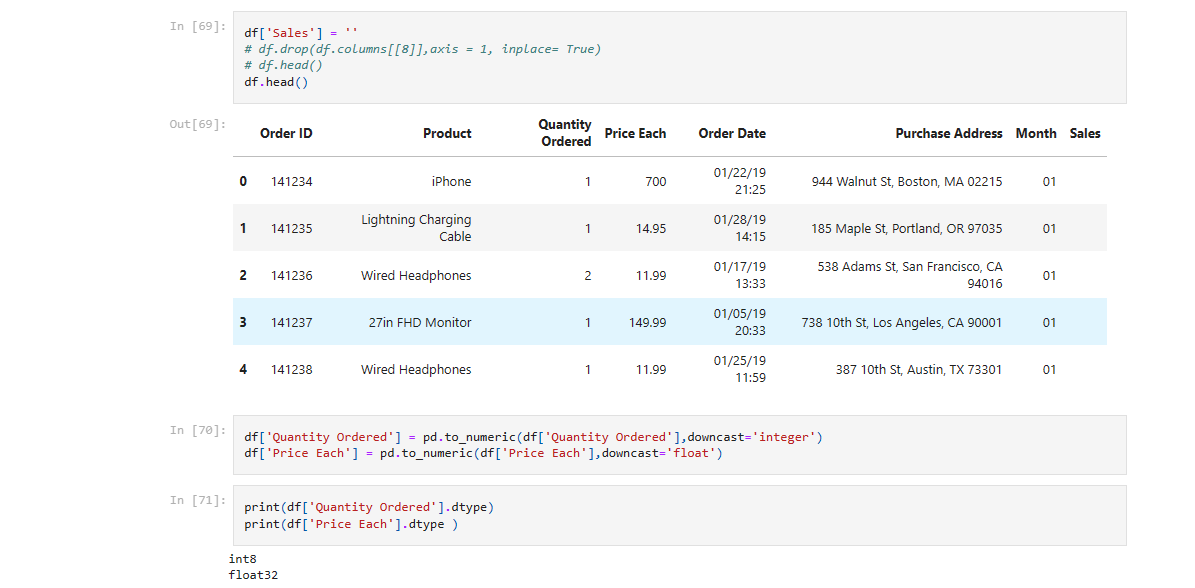
“Nan” là do có một số hàng không có dữ liệu.Còn “Or” là bởi vì khi nối các nội dung trong các file csv lại thì nó sẽ gộp lun cả tiêu đề. Vì vậy nên phải loại bỏ 2 giá trị không đúng này.

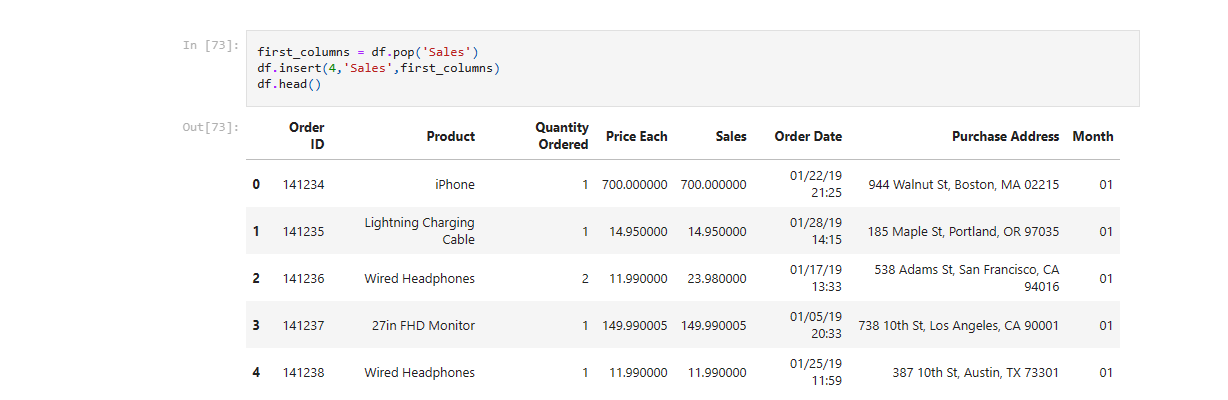
Sử dụng hàm df**.**dropna(how**=** 'all') để loại bỏ tất cả giá trị ”Nan” trong tất cả các hàng. Và lọc các kết quả trong cột Month khác “Or” bằng cách df[‘Month’] != 'Or'.

Và chạy lại để có được kết quả:

1. **Thực hiện phân tích trả lời từng câu hỏi:** 
   1. **Tháng nào bán được nhiều nhất? Và tháng đó bán được bao nhiêu?**

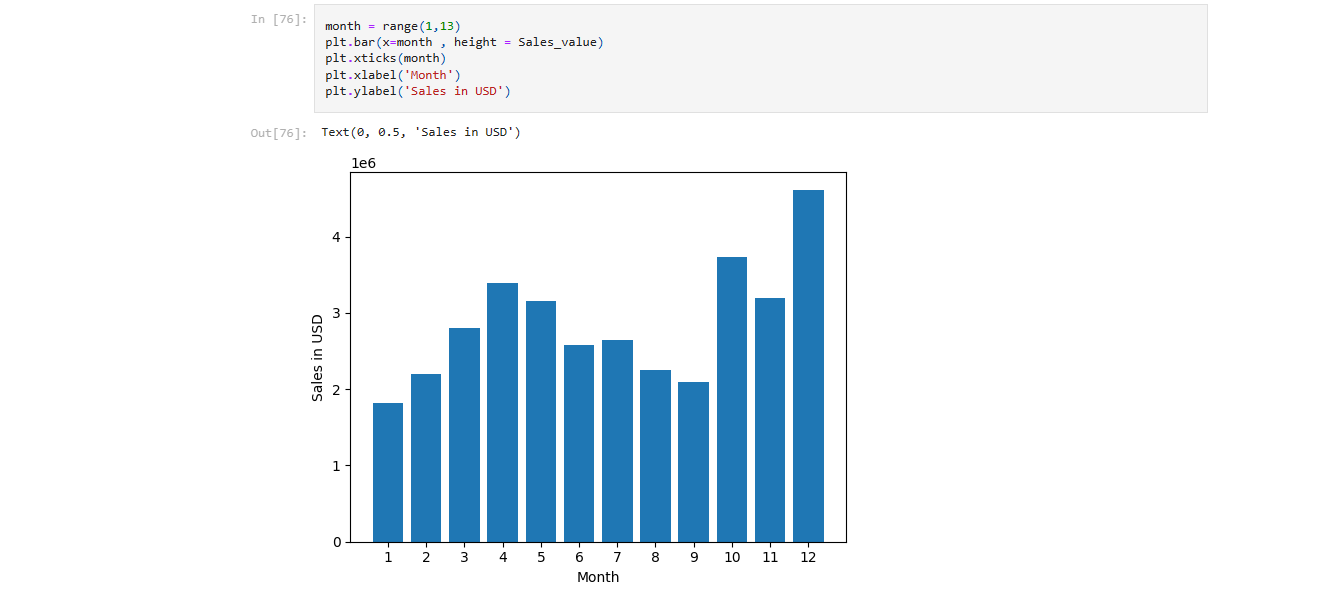
Khi đó tạo thêm 1 cột “Sales” để lưu doanh thu được tính từ 2 giá trị là **Prices Each và Quantity Ordered.**

Nhưng mà khi thực hiện sẽ phát hiện lỗi, vì vậy nên phải ép kiểu để đưa về kiểu dữ liệu phù hợp là int và float theo thứ tự sau đó kiểm tra lại kiểu dữ liệu của hai phần tử này.

Sau đó chỉ thêm dữ liệu vào trong cột “Sales” vừa mới tạo. Muốn cho dễ nhìn thì di chuyển cột “Sales” vào sau cột “Order Date” và kết quả là:

****Thực hiện “Groupby” gom nhóm theo tháng và tính tổng trong tháng đó. Và tìm ra giá trị lớn nhất là 4613443.5

Sau khi tìm được thì vẽ biểu đồ biểu diễn số liệu cột x là tháng, cột y là doanh thu tính theo đơn vị là USD.

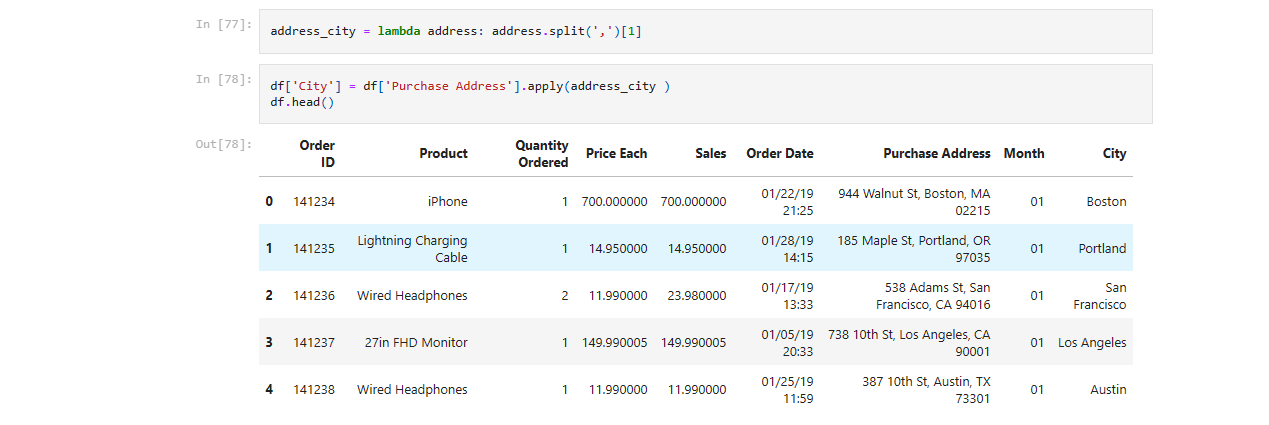


**­**

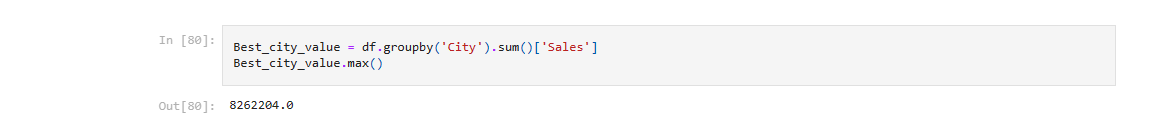
* 1. **Thành phố nào thì bán được nhiều nhất?**

Thực hiện tách tên thành phố ra bằng cách viết 1 hàm lambda và sử dụng hàm split() tách theo dấu “,”.

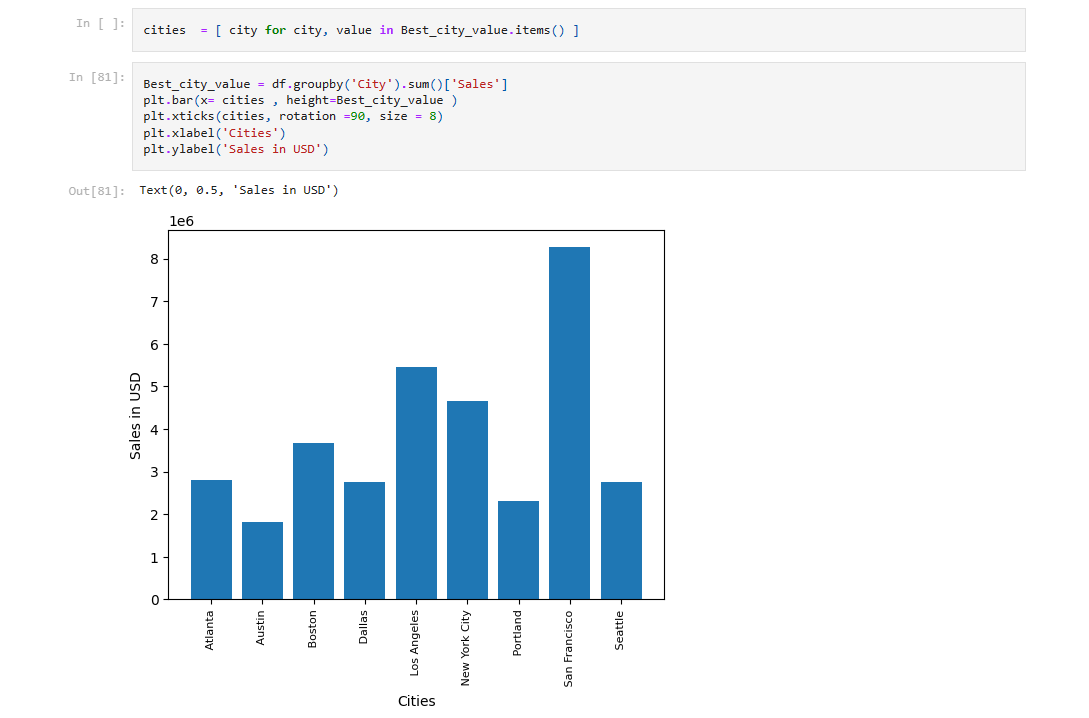
Tương tự như cột “**Sales**” thì tạo thêm 1 cột tên là “**City**” lưu tên thành phố vừa được tách.



Sau đó cũng “**gom nhóm**” và tính sum() theo thành phố đó và tìm ra giá trị lớn nhất.

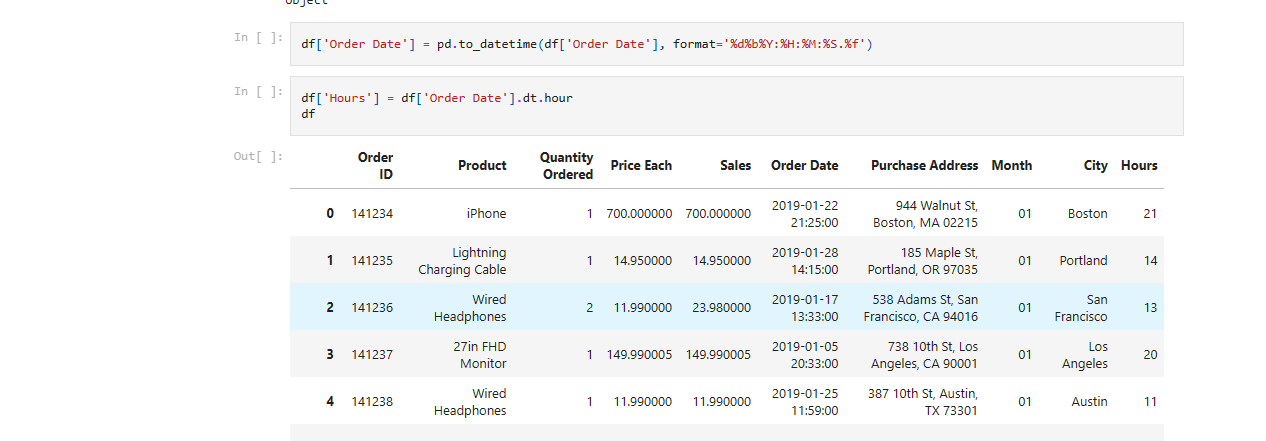


Dùng vòng lặp duyệt qua biến Best\_city\_value để có thứ tự đúng của từng City. Sau đó dùng từng City trong vòng lặp để vẽ biểu đồ để có sự chính xác giữa số liệu và thành phố đúng.

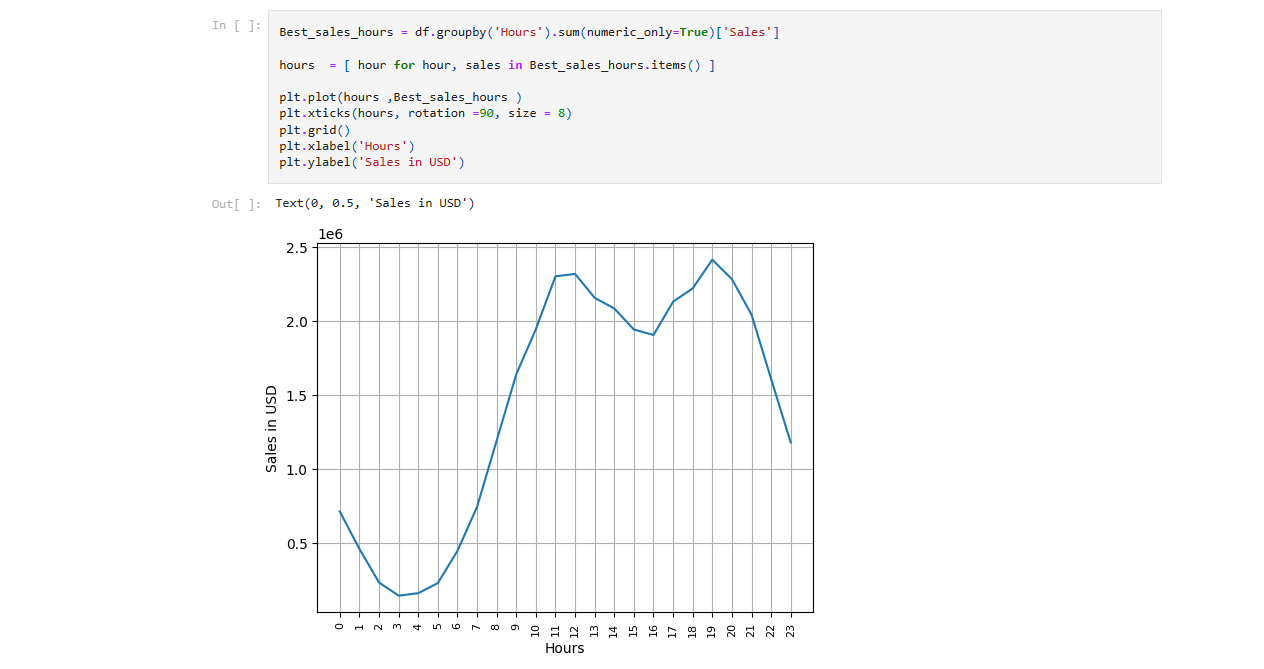


* 1. **Khung thời gian nào thì được mua hàng nhiều nhất cho khách hàng?**

Thực hiện đưa cột “**Order Date**” về đúng định dạng DateTime. Tạo thêm cột “**Hours**” để tách giờ trong “**Order Date**”ra theo đúng định dạng.

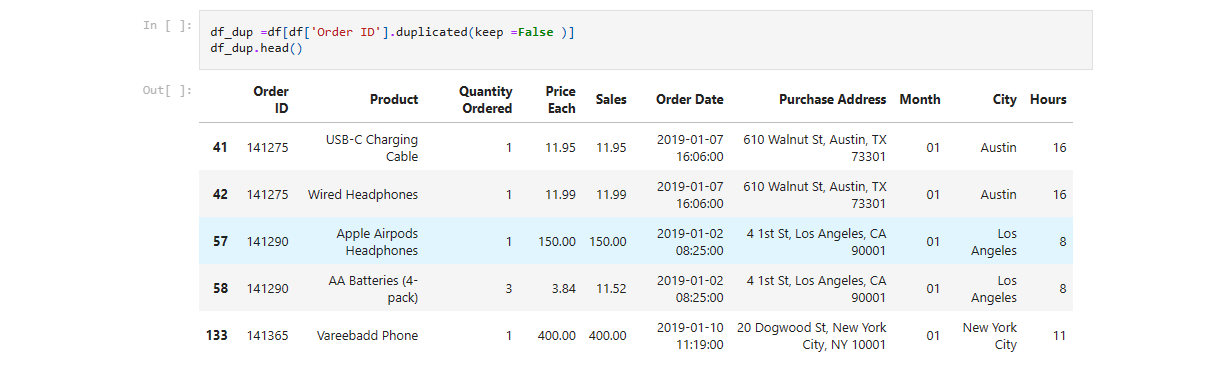


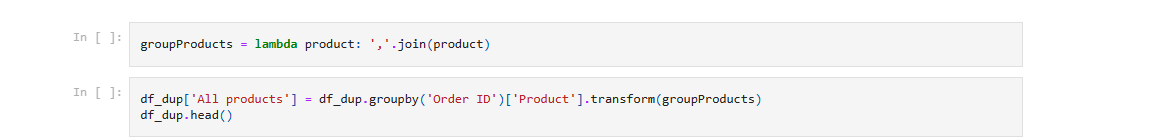
Tương tự như trong tìm thành phố nào có doanh thu cao nhất.



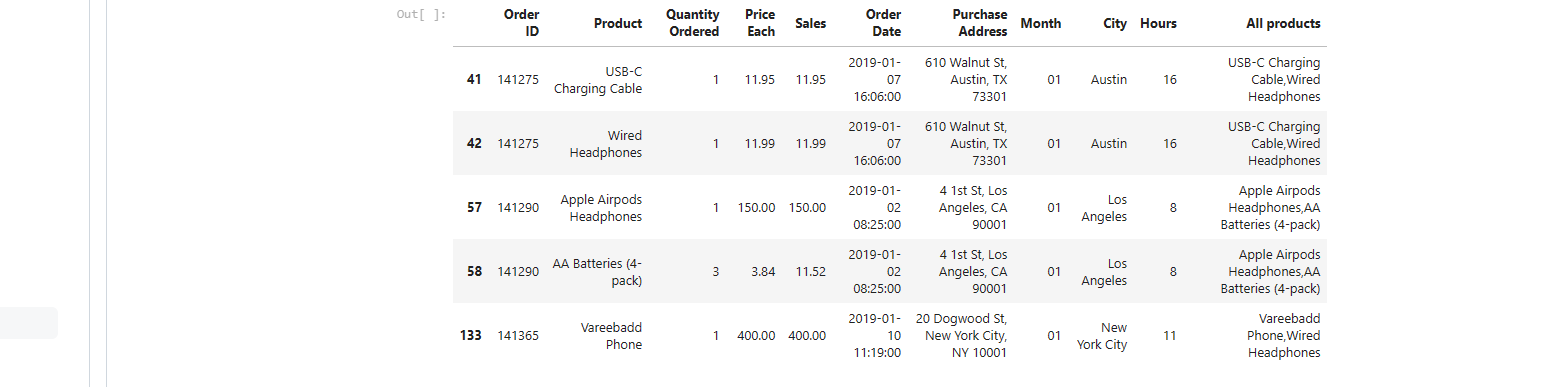
* 1. **Sản phẩm nào được bán cùng nhau nhiều nhất?**

Tìm những “**Order ID**” bị trùng lặp in ra để kiểm tra.

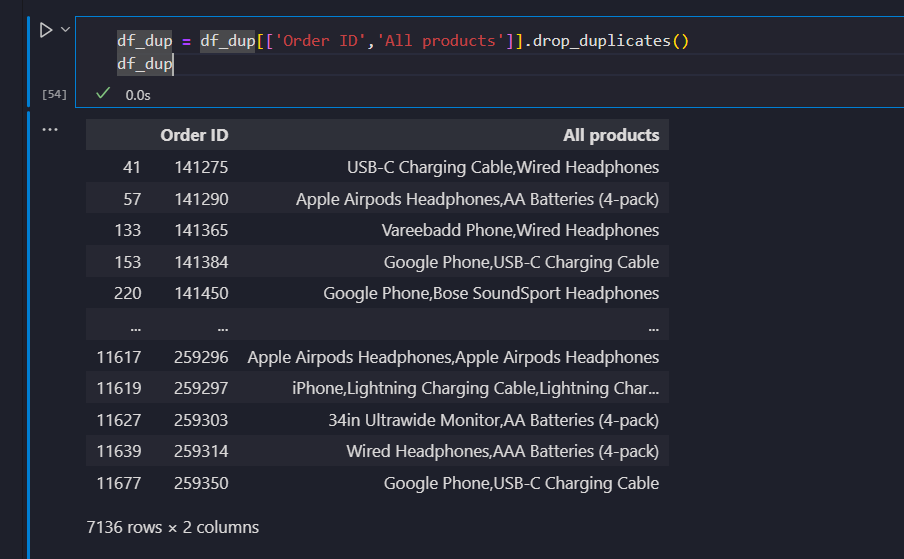
Tạo ra một cột “**All Products**” để thực hiện gom nhóm cột ID và cột Product. Sau đó dùng hàm transform() để chuyển đổi dữ liệu thành dict().

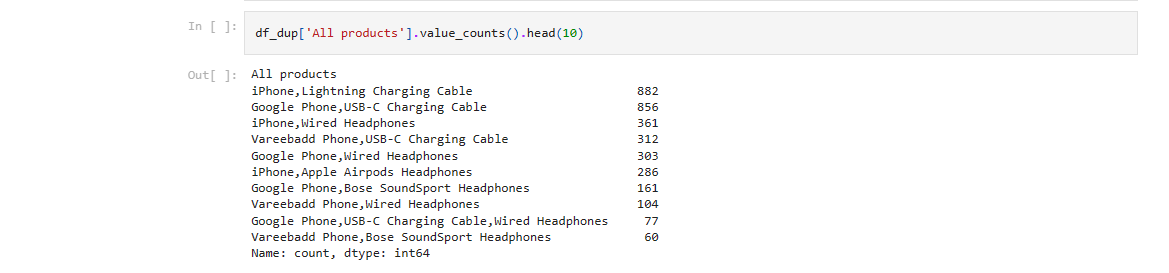


Kết quả là:

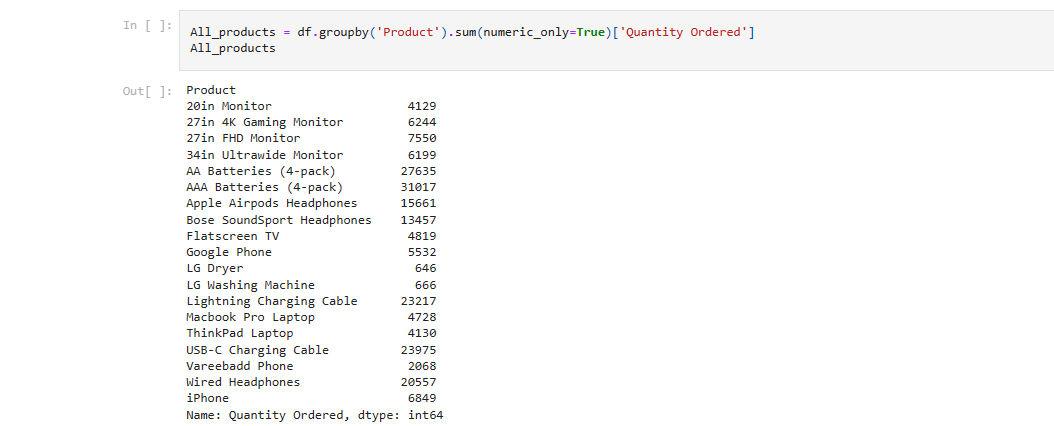


Sau đó xoá các hàng trùng lặp ra khỏi Data Frame để thực hiện đếm từng sản phẩm còn lại.

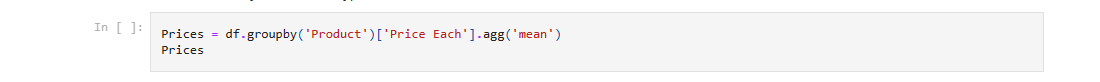


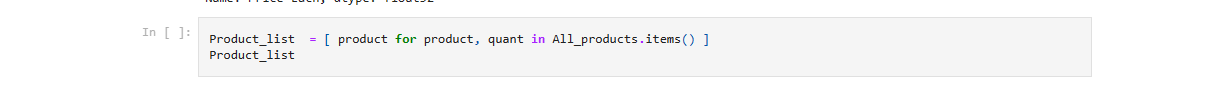
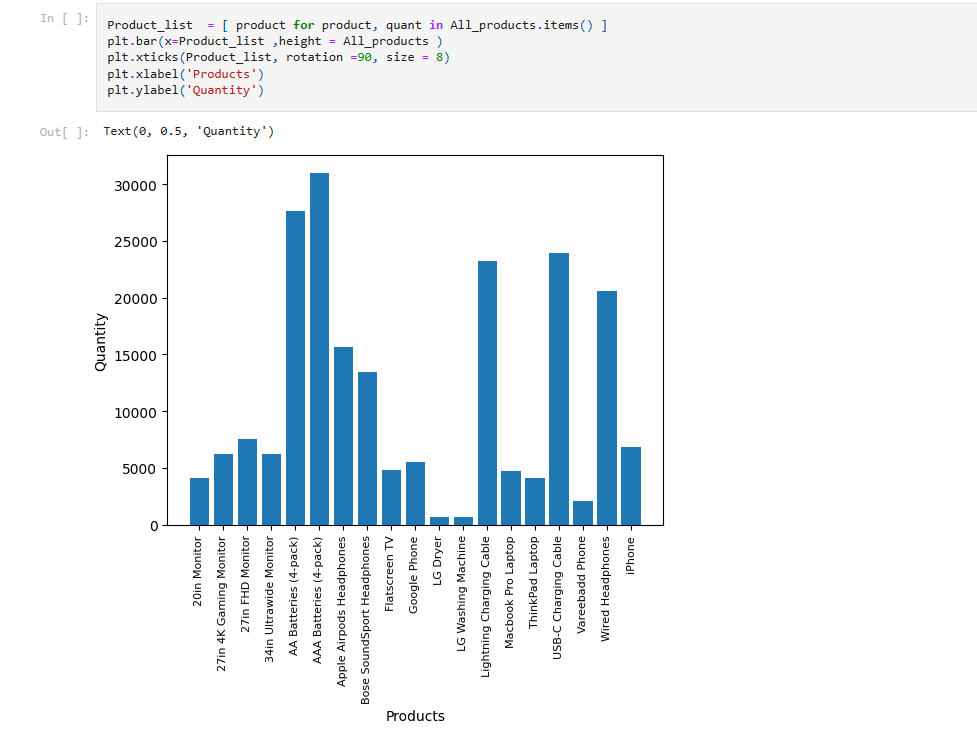
Thực hiện đếm từng cặp sản phẩm còn lại trong Data Frame còn lại để tìm cặp sản phẩm nào bán được nhiều nhất.

Thực hiện tính tổng các sản phẩm theo số lượng sản phẩm đó có trong Data Frame.

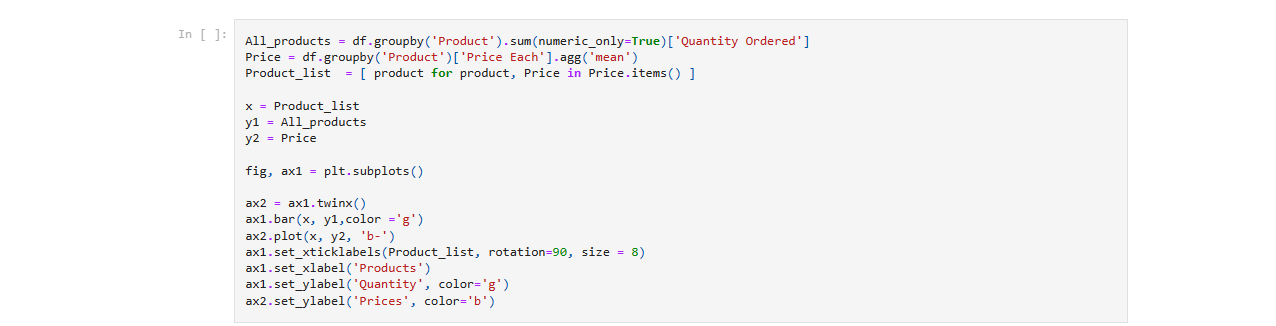
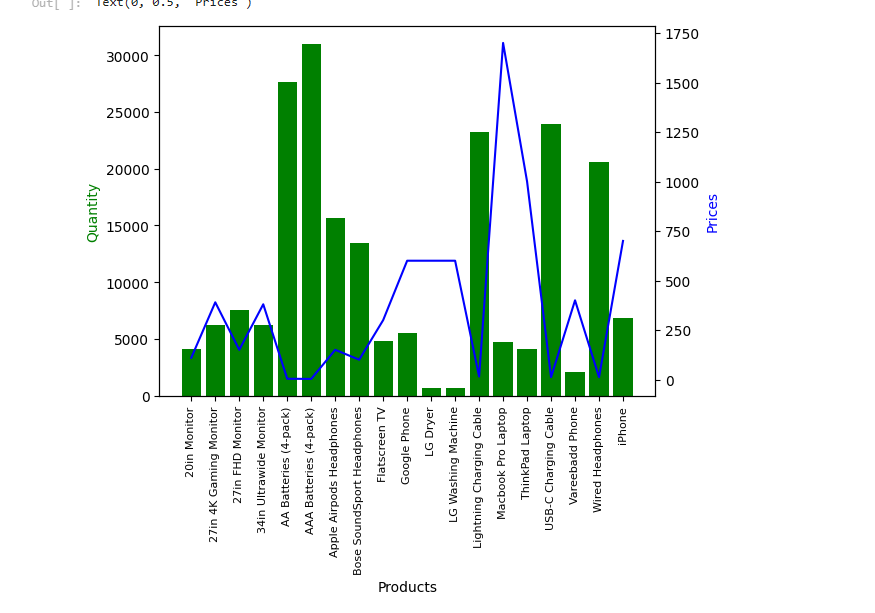


Gom nhóm từng sản phẩm và tính trung bình theo giá của từng loại sản phẩm.

Thực hiện duyệt qua vòng lặp để lấy tên từng sản phẩm và vẽ biểu đồ.

Kết quả:

Vẽ biểu đồ có các cột “Quantity”,”Prices”,”Products” để biểu diễn sự tương quan giữa sản phẩm, giá và số lượng của từng loại sản phẩm.

Kết quả: