Chapter 2 - Ex3: Predicting Customer Spend

Part1:

- Cho dữ liệu retail_transactions.csv, gồm các thông tin như sau:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	546729	22775	PURPLE DRAWERKNOB ACRYLIC EDWARDIAN	12	2011-03-16 11:36:00	1.25	18231.0	United Kingdom
1	559898	21868	POTTING SHED TEA MUG	6	2011-07-13 12:18:00	1.25	16225.0	United Kingdom
2	548648	71459	HANGING JAM JAR T-LIGHT HOLDER	24	2011-04-01 13:20:00	0.85	12949.0	United Kingdom
3	540543	22173	METAL 4 HOOK HANGER FRENCH CHATEAU	4	2011-01-09 15:23:00	2.95	14395.0	United Kingdom
4	561390	20726	LUNCH BAG WOODLAND	10	2011-07-27 09:52:00	1.65	17068.0	United Kingdom

- Hãy chuẩn hóa dữ liệu này và lưu vào tập tin wrangled_transactions.csv, gồm các thông tin như sau:

CustomerID	2010 revenue	days_since_first_purchase	days_since_last_purchase	number_of_purchases	avg_order_cost	2011 revenue
12347.0	711.79	23.0	23.0	1.0	711.79	3598.21
12348.0	892.80	14.0	14.0	1.0	892.80	904.44
12370.0	1868.02	16.0	13.0	2.0	934.01	1677.67
12377.0	1001.52	10.0	10.0	1.0	1001.52	626.60
12383.0	600.72	8.0	8.0	1.0	600.72	1249.84

Thông tin cung cấp:

- Trong dataset này có một số dữ liệu giao dịch lịch sử từ năm 2010 và 2011. Với mỗi giao dịch, có customer identifier (CustomerID), số lượng hàng đã mua (Quantity), ngày mua (InvoiceDate), đơn giá (Unitprice), và một số thông tin khác về mặt hàng đã mua.
- Cần tiền xử lý dữ liệu này thành dữ liệu giao dịch của khách hàng từ năm 2010 so với chi tiêu năm 2011. Vì vậy, cần tạo ra các feature từ dữ liệu cho năm 2010 và tính toán mục tiêu (số tiền đã chi) cho năm 2011.
- Khi xây dựng tạo mô hình này, nó sẽ khái quát cho những năm tới. Nhờ đó, doanh nghiệp có thể sử dụng dữ liệu năm 2020 để dự đoán trước hành vi chi tiêu vào năm 2021 (trừ khi thị trường hoặc doanh nghiệp đã thay đổi đáng kể kể từ khoảng thời gian dữ liệu được sử dụng để fit mô hình.

Part2:

- Hãy áp dụng thuật toán Linear Regression để xây dựng model dự đoán customer spend dựa vào dữ liệu wrangled_transactions.csv vừa có ở Part1. Đánh giá model. Trực quan hóa kết quả.
- Với '2010 revenue': [1000], 'days_since_last_purchase': [20], 'number_of_purchases': [2], 'avg_order_cost': [500] thì '2011 revenue' là bao nhiêu?