Your task was to use the dataset to provide answers to the following three questions:

1. Which stores are performing best and which ones are performing worst
2. The regions with the largest opportunity for growth if they were to open a new store
3. What top 5 products they should sell in a new store to maximise profit when they first open

1. Which stores are performing best and which ones are performing worst

To calculate the answer for this, you need to first define what the performance will be based on. In real life, usually some sort of metric such as profit, revenue or customer satisfaction will be used. In this case, the metric was not provided so it is your choice which metric to use. For the model answer, revenue was used. To calculate the revenue for each store, you simply had to aggregate the data by 'store' and summing the value of 'revenue'. This will give you the answer provided in 'store_performance.csv'.

2. The regions with the largest opportunity for growth if they were to open a new store

Opportunity for growth was once again not provided as a tangible metric, so you needed to do some thinking about what a good metric would be to measure the opportunity for growth. In this case, the model answer used 'profit' as a metric to measure opportunity for growth. This was because a largely profitable region may present an opportunity to open new stores and to take advantage of this profitability. Profit was not given as a column in the dataset, so it had to be calculated. It can be calculated with the following formula: revenue - cost_of_goods_sold.

Once you have your profit column, you simply aggregate the data by the 'region' values and summing the profit for the regions. This will give you the answer provided in 'region_opportunity.csv'.

3. What top 5 products they should sell in a new store to maximise profit when they first open

For this question, we need to make use of profit again. However, it will require a slightly different calculation. This is because the current profit calculation is biased by the quantity of the product item that was sold.

To answer this question, we need to use the individual price of a product and the cost of an individual product being sold. We can calculate the latter by dividing the 'cost_of_goods_sold' column by 'quantity'. Call this 'cost_of_goods_sold_individual'. Now we can calculate a per product sold profit, being: product_item_price - cost_of_goods_sold_individual. This new column will give us the profit of each individual product being sold, even if the quantity was greater than 1. This is important to know because this will help us understand which individual product item is the most profitable.

To get the final solution, you simply aggregate the transformed data by product_item_id and take the mean of the newly calculated profit value. This will give us the average profit expected by each individual sale of a product. This will give you the answer provided in 'maximise_profit.csv'.