Data Scientist Analytics Test

Summary

The public data set provided for this test covers e-commerce transactions over a two year period.

Your task is to produce a report on the data, including the scripts, code and visualised findings.

You are free to use either R or Python for this task and please return your finished work as a PDF exported from R Markdown, Jupyter or equivalent.

Your Report

The following should be included in your report:

1. Describe and summarise the data set. This should include how much missing data there is, how the data can be grouped, how unique the data is, any errors in the data and what should be done with those errors, and the visualisation of important features.

2. A description of what new features could be created using the data provided as a starting point, with examples.

3. Describe any trends that can be found in the data that could potentially be useful in a commercial sense.

4. Choose a class of modelling to perform on the data, be that something predictive or something concerning classification/clustering. Provide two different approaches to whichever one you choose and include results of the two approaches.

5. Include the following in your analysis:

    - Your rationale for test and training set splits.
    - What do the models tell us about features and their importance?
    - Compare the performance of the two models and describe what could be done, and what expectations should be, for future development of them.

The Data – Ecommerce Transactions

1. **InvoiceNo –** Numeric, invoice relating to the purchase
2. **StockCode –** String, unique per product
3. **Description –** String, unique per product
4. **Quantity –** Numeric, quantity of product on the invoice
5. **InvoiceDate –** Date
6. **UnitPrice –** Numeric, price per item
7. **CustomerID –** Numeric, unique identifier per customer
8. **Country –** String, country that the customer is from