DATA ANALYST TECHNICAL ASSESSMENT ANSWERS

July 2020

Dr Dimitar Dimov
Candidate at Laing O'Rourke

Question 1

- In terms of representativeness of real-world datasets – this is a typical SQL export dataset of multiple joined tables from a retailer keeping records of their sales activity
- It is a time-series data – each new entry is recorded with a time stamp
  - *This is good for identifying cyclical trends based on hours / days / months (seasonality)*
  - *Once a trend is found, you can make recommendations for optimisation, cost cutting and increasing revenue*
  - *A supervised machine learning model can be applied data to forecast dependent variables (ie profits)*
- The InvoiceNo column is the unique identifier (primary key in SQL) so it would make sense to collect and store the data in a relational database with multiple tables (you can have multiple invoices for one customer but not multiple customers on one invoice)
- There are 500k+ rows – even after cleaning the missing values this still will be considered a large dataset, hence, the machine learning prediction models will have good degree of accuracy and the conclusions made about purchasing trends and customers' behaviour will also have high degree of certainty
- We have number of units sold (Quantity) and Unit Price so I would immediately create a new column labelled Total Sales
- Following these calculations, we can visualise daily, weekly or monthly revenues and categorise them by product, returning customers etc. This would be best done if we load the cleaned and processed data in Tableau or Power BI
- We can group and sort the data to find the outliers – the best-selling and the worst selling products, clients with most purchases and countries generating the highest revenue

Question 2

Before performing any analysis, I always spend some quality time looking at the data in Excel to better understand the content, data types, scroll down to find any immediate / easy to spot irregularities. If the dataset has more than 10 columns there is no point in this visual observation and I would go straight to coding.

Moving onto Python, I loaded the xlsx file in a pandas dataframe. Since it is a time series data, it makes sense to use the InvoiceDate for parsing the dates. The first two checks are data.info() and data.describe() to better understand the types of data in each column and find where the missing values are.

There is one key observation which has to be analysed for this specific dataset.

- *Most of the columns have ~541,000 entries but the CustomerID column has ~406,000 entries. This means that ~25% of the CustomerIDs are empty.*

We will take two approaches. The first one is PRODUCT-focused and the second is CUSTOMER-focused.

1) ***Product focused analysis -*** *we can keep the empty CustomerID cells and clean the data by removing invoices with 0 products or other irregularities arising from the InvoiceNo and Description columns. This will give more accurate information and analytics for your* ***products, volume of sales and trends,*** *but you will not be able to derive conclusions for specific customers*
2) ***Customer focused analysis*** *- we will remove all empty cells (and rows) from the CustomerIDs column. This is a must if you are going to perform machine learning classification or regression models and if you intend to visualise your data based grouping by CustomerID*

The next step is to dive into the StockCode and Quantity columns and find outliers and irregularities. As you'd notice in the python code, I removed a few StockCode names which are clearly not related to product sales and few rows from Quantity where the values are clearly out of the standard deviation. I also checked the Descriptions column for missing values and removed those as well. Also dropped the duplicates.

The last step of the pre-processing work is to save two sub-datasets –

- ***products_dataset*** *for the Products-focused analysis*
- ***customers_dataset*** *for the Customers-focused analysis and future machine learning models*

Question 3

After all pre-processing and cleaning outliers / irregularities, I have added a new column – TotalSales – which is simply the number of items sold multiplied by their unit price. I have exported two separate databases as xlsx for visualisations in Tableau.

I have also written a python code to calculate the following:

1. *Recency – a measure of how recent one purchase is, benchmarked against the very first transaction recorded on the database*
2. *Frequency – a measure of how many times has each customer purchased stock*
3. *Monetary Value – a measure of how much has each customer spent in total over the recorded timeframe*

Which are part of the commonly used RFM analysis and give customer-focused insights.

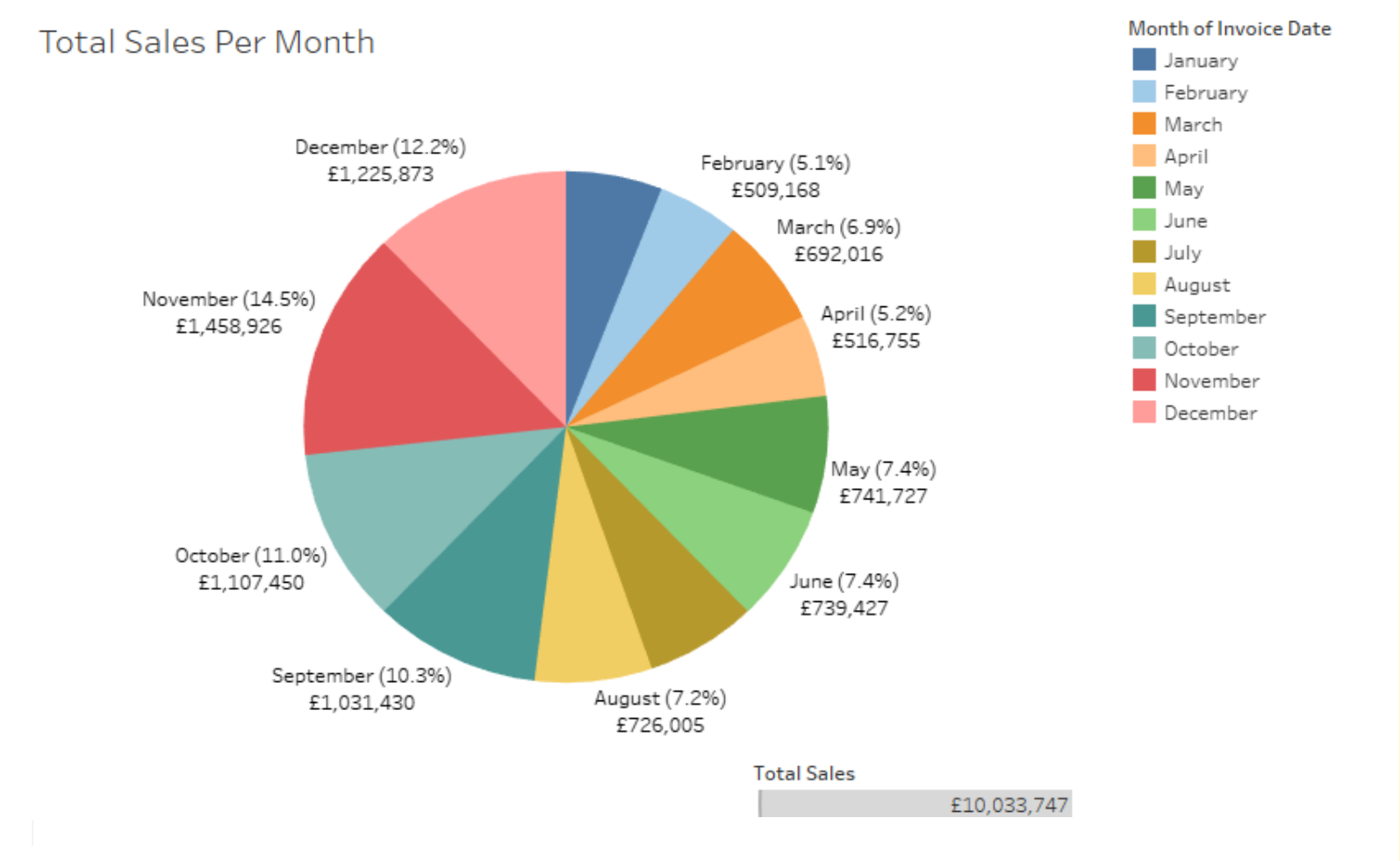
Question 4

Let's start by visualising basic business information – top 10 selling products, total weekly sales with an average value per week and a trendline, a pie chart with total sales breakdown on a monthly basis, and total sales per country. All of the above will come from the products database information.

I will include a summary of all key findings in Question 5.

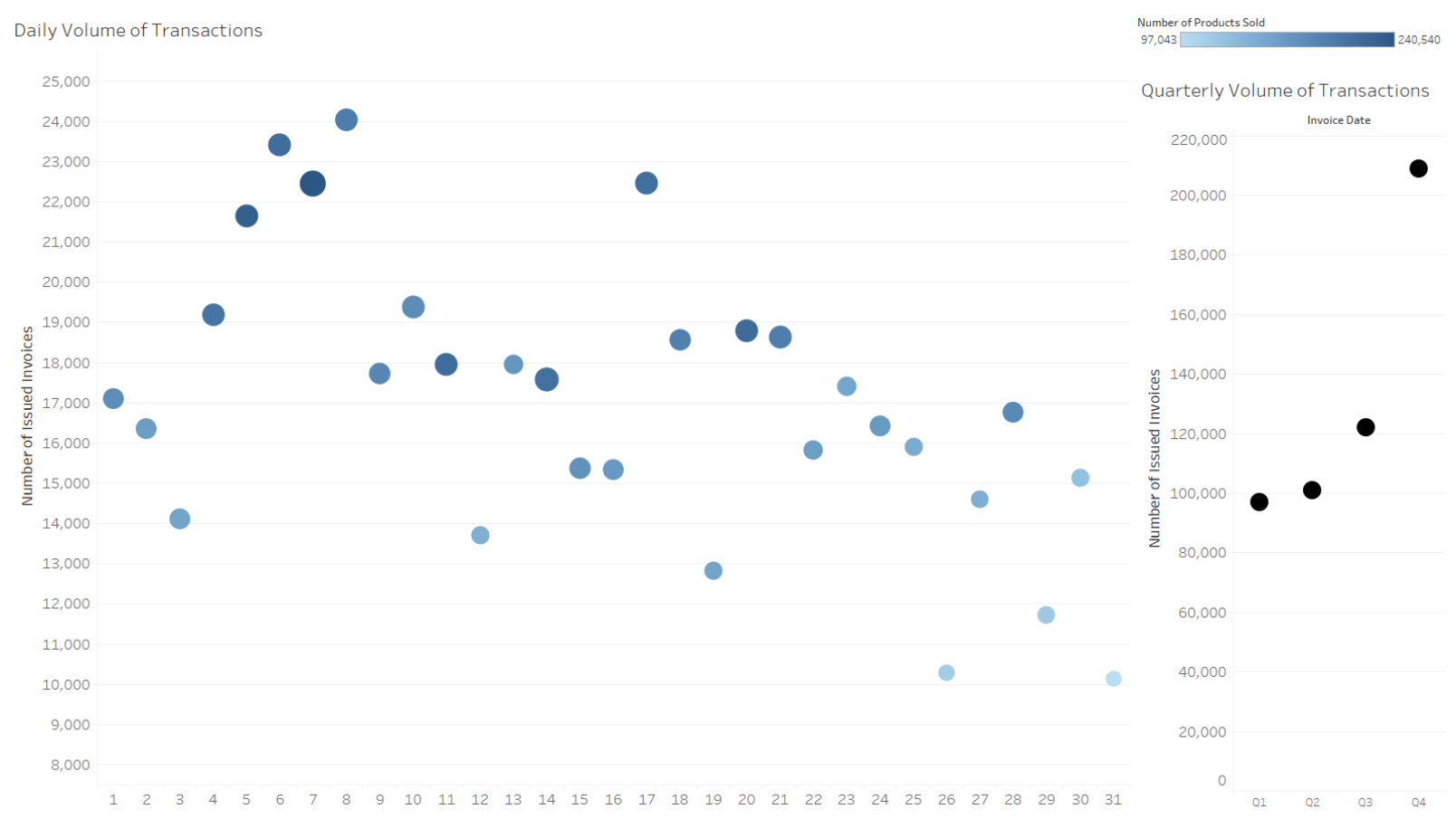
The second step is to create visualisations for the Customer-focused database and perform the RFM analysis – the simplest assumption one can prove is the Pareto 80-20 rule. The assumption made is that 80% of the revenue comes from 20% of all customers.

Question 4 – cont'd



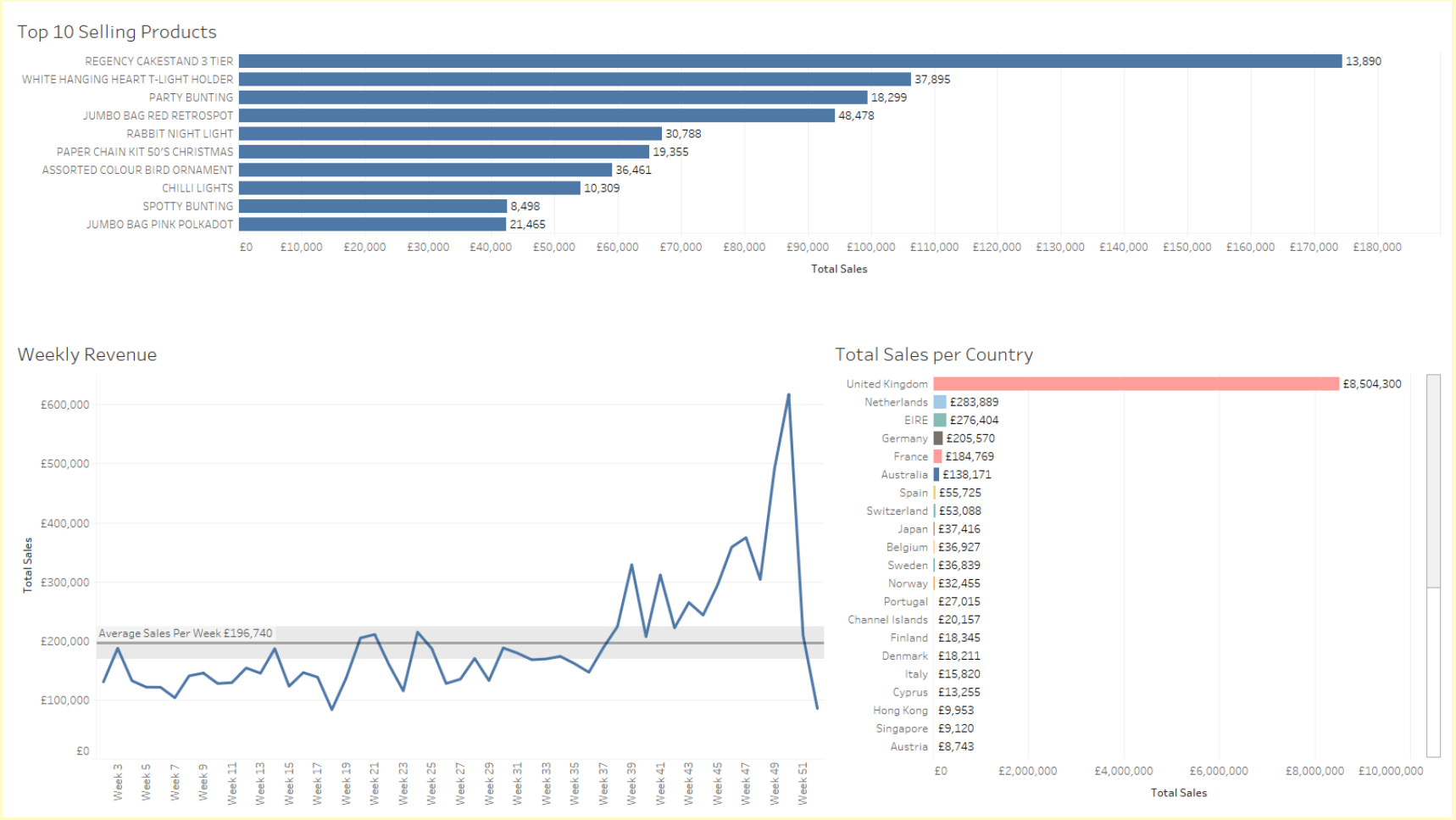
- November generated the largest revenue (14.5%)
- Q4 performance was significantly better than the rest
- Autumn / Pre-Christmas season is the best for the company so a strategy for optimising the resources should be mapped out
- It would be useful to collect data for the cost of shipping and returns

Question 4 – cont'd



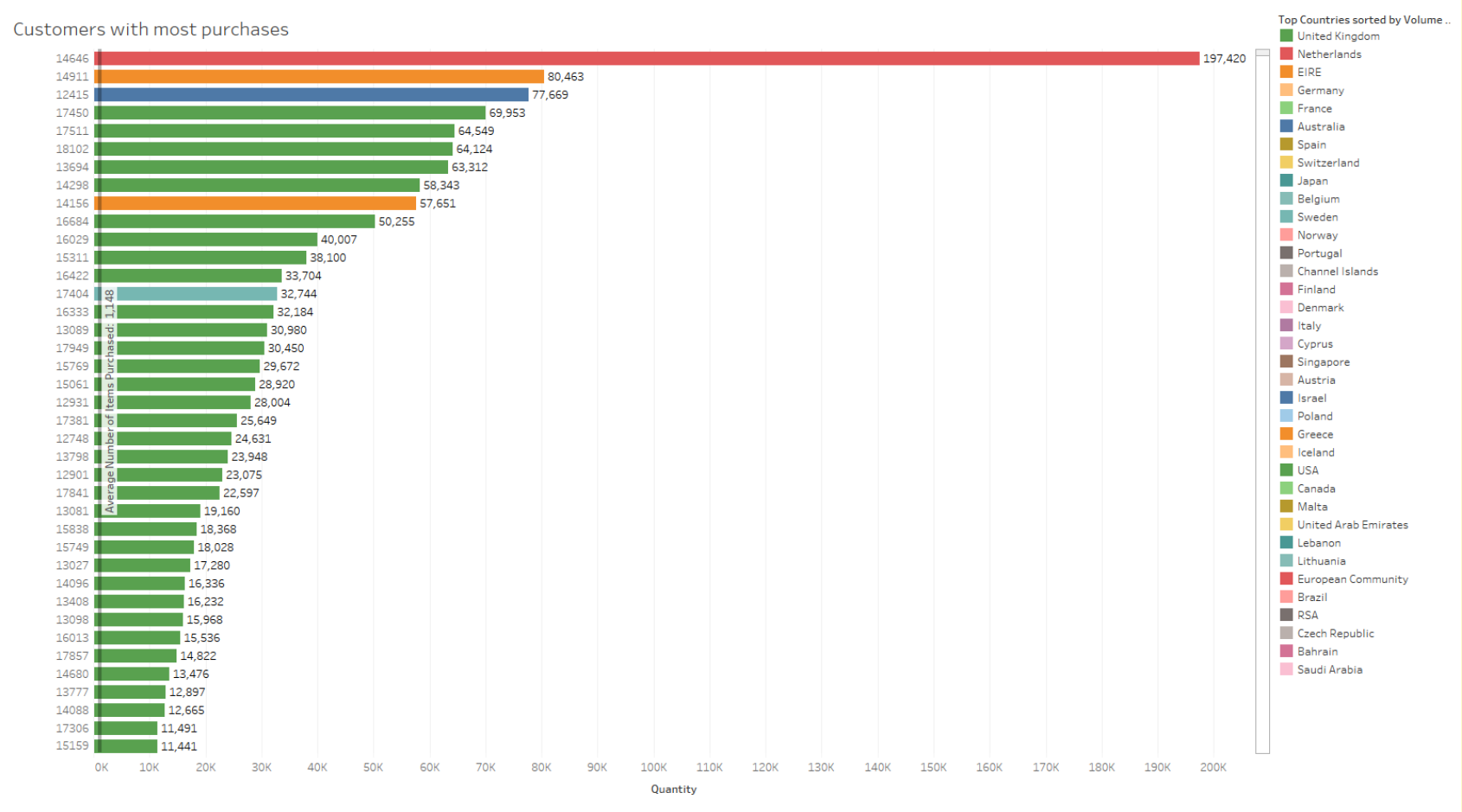
- The chart on the left shows the volume of all transactions grouped by the day of the month
- It is clear that the highest volume occurs between the 4th and 9th of each month – what is the reason behind this? Soon after pay checks?
- The colour of the data points shows the number of products sold – the darker the blue, the higher the number
- There is little activity towards the end of the month – perhaps incentivise the client by offering month end promotions and discounts?
- The 17th day of the month is an outlier – why is it special? Worth checking which products are sold on that day and why
- The chart on the right shows the clear trend of increased volume of transactions in Q3 and Q4 – what changed for the company in this period? How can they further optimise / leverage in the future?

Question 4 – cont'd



- Chart on the top shows the top 10 selling products – their total sales on the x axis and the total number of products sold against each bar
- It is interesting to note that the first product did not sell in large quantities but was expensive enough to generate large revenue – ideas how to further maximise and optimise this?
- Chart bottom left – weekly revenue – what caused the big fall in weeks 50-52? Is it missing data or post-Christmas time?
- Consistent average sales from January through September – reoccurring clients, seasonal products or limited marketing budget?
- What caused the sudden increase from Week 43 onwards?
- Chart bottom right – clearly the UK is the country with the highest volume of sales, followed by other European Countries and EIRE
- How can the company increase sales outside the UK? Perhaps marketing ads in German and French language?

Question 4 – cont'd



- The average number products purchased by a customer is 1,148
- This chart shows that the top 20 customers (out of ~4300 in total) have purchased more than 20x or even 60x the average
- Worth checking customer 14646 – is there a problem with the data? I dropped the duplicates but perhaps some are left?
- Clearly less than 10% of total customers are generating 80%+ of the revenue
- What about returns?
- How much actual profit are the customers generating?
- Is there a customer registration option and what does the company do to incentivise their highest spenders to spend more? Loyalty programs perhaps?

Question 5

The key findings from the **products_dataset** analysis:

- Total Revenue for the year - £10,033,747
- The average weekly revenue is £196,740
- Best performing month – November (14.5% of total sales)
- Most transactions occur between days 7 and 10 of each month
- The top 5 countries where customers purchase from
    - *UK, Netherlands, EIRE, Germany, France*

Products which generated the highest revenue:

Product Name	Total Products Sold	Total Revenue
REGENCY CAKESTAND 3 TIER	13,890	£174,485
WHITE HANGING HEART T-LIGHT HOLDER	37,895	£106,293
PARTY BUNTING	18,299	£99,504
JUMBO BAG RED RETROSPOT	48,478	£94,340
RABBIT NIGHT LIGHT	30,788	£66,964

The key findings from the **customers_dataset** analysis:

- There are 4334 unique customers
- The most active customer – number 14646, purchased total of 197,420 products and is from the Netherlands
- The remaining 9 from the top 10 customers (ranked by number of products purchased) have purchased an average of 65,257 products
- The average number of products purchased for all 4334 customers is 1,148 so this further highlights the importance of the top 10

Question 6

Everything related to communications skills.

Doing the entire data analytics cycle would lose its true value if you can't present it well to the key stakeholders, both internal and external.

The delivery and presentation of your work is key for success – the way you present your data, the way you articulate your findings, the way you answer the questions from the audience and consequent logical arguments to support your key statements – that's what this technical assessment can't assess and I strongly believe is a very important skill set of every Data / Business Intelligence Analyst.

Another important skill not demonstrated through this assessment is team work – had I done this analysis in the office with my colleagues, I would've interacted a lot with the data architect, the data scientist, would've bounced off ideas with my managers and would've had informal conversations with any colleague who frequently purchases online (I myself am not a big online shopper)

The latter is crucial for me to understand what the customers like and dislike from their e-shopping experience which information I will then use as recommendations to my retail client (or set target KPIs)

Question 7

I would like to start by stating the definition of key performance indicators (KPIs) – a measurable value that demonstrates how effectively a company is achieving key business objectives. Organisations use KPIs to measure their success at reaching targets.

So, if we deconstruct this, the most important first step, prior to discussing the approach for effective design, is to understand that KPIs are form of communication. When we design a strategy, we need to understand what the organisational objectives are and how are we going to achieve them. Every KPI should be related to a specific business outcome with a performance measure.

If we take the Online Retail database for example and put ourselves in the shoes of the Retail owner, we can define our goal as Increase Sales or Sales Growth. Then, we have to come up with specific KPIs which can quantitatively measure the performance. For example:

- *To increase sales by 10% this year*
- *The Sales department team is responsible – should take the lead and report back to the wider business on a monthly basis*
- *Hire additional sales staff*
- *Will review this target on a quarterly basis*
- *Progress will be measured as an increase in revenue measured in pounds spent*

The approach that I would take when defining the KPIs is always communicate with the wider team of stakeholders, both internally and externally, to understand and define our goals and then design a set of criteria points which can measure the performance of achieving these goals.

Question 8

Let's begin by briefly outlining the definition of linear regression (LR) – this is a predictive analytics tool / technique which uses one or more independent variables to predict the outcome of a dependent variable. Therefore, one can have single variable LR or multivariate LR.

The approach I would take when validating a linear regression measure how accurately is the predicted data. First, I will split the original data at 80-20% where 80% will be for training the linear regression model and 20% will be used for testing, where I will apply the LR model and predict real values. The next step will be to calculate the mean squared or mean absolute errors of both the training and testing data and compare the scores. The lower the values, the better the fitting was (ie the dependent variables are as close to the independent as possible)

MSE – mean squared error or mean squared deviation – measures the average squared difference between the estimated values and the actual value. Always non-negative (since it is a square) so the smaller the number the better (in terms of evaluating the quality of the fitted line)

The **pros of MSE are:**

- *Due to its simplicity, it is parameter free and inexpensive to compute*
- *Does not require memory allocation – the squared error is evaluated at each sample, independent of other samples*
- *It provides a consistent comparison due to its nonnegativity*

The **cons of MSE are** concerned with Variance – since the iterative process performs the exact same calculation on all data points, it is inherently prone to outliers. The bigger the outlier, the bigger the error in your MSE score. This is further amplified when one uses RMSE (the root version of MSE) because the calculation squares the difference first and then takes the root of it, hence amplifying the error margin. Conversely, if most of the outliers are removed during the pre-processing stage, then RMSE will yield a result with higher degree of accuracy.

MAE – mean absolute error measures the average magnitude of the errors in a set of predictions, without considering their direction. It's the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight. It is negatively oriented, so the lower the value the better. Similarly to MSE, **the pros** are related to the low computational power, **and the cons** are that since the differences between actual and forecast are weighted equally in the final average, big outliers will have bad impact on the final score.

Question 9

Linear regression (LR) is a statistical model which assesses the relationship between two variables and predicts the dependent variable values as a function of the independent variables. It can be a simple regression (with only one independent variable) or multivariate regression where multiple features are selected to predict the values of the explan

Before jumping on the pre-processing and cleaning the dataset I would brainstorm whether a logical correlation between the two variables exists in the first place.

Then, there are couple of mathematical assumptions that need to be taken into account:

- *Linearity and additivity – the expected value of the dependent variable is a straight-line function of each independent variable*
- *Statistical independence of the errors – specifically for time-series data, no correlation between two consecutive errors should be present*
- *Homoscedasticity – constant variance of the errors exists*

From programming perspective, one should not start training a LR model prior to cleaning the data. All missing cells should be either removed or filled with either the average value of the sample, or the value of the preceding cell, otherwise the evaluation metric (MSE or MAE) will contain errors. The drawbacks of LR are:

- *it is useful when the relationship between the variables is naturally linear but LR would be an inaccurate model for curved relationship (for example wealth and age have an exponential type of curve, it is rarely linear)*
- *it only looks at the mean of the dependent variable – this does not factor in low quantiles (extremes) which in some cases might be very relevant for the relationship between the variables*
- *it is very prone to outliers so unless those are completely removed the overall evaluation metric will have a high degree of error*
- *variables should be strictly independent – in many cases, especially with large datasets, clusters with centroids are often found so LR can not be applied*

Question 10

The three data roles – architect, scientist and analyst, have much in common but also differ a lot mainly in the output of their work.

The data architect is responsible for creating the infrastructure of the data flow, processes, tools, making sure that the data pipeline works well and the data is fed into the lake consistently without errors.

The data analyst takes the raw data, applies exploratory analysis by inspecting and cleaning the dataset as well as adding new features of needed. The next step is key – finding trends, conclusions, outliers, creating a story (powerful visualisations) based on the data and crucially, taking the client on a journey and explaining what this data means for their business.

On the back of this data visualisation and presentation, the data analyst works together with the client to create and define KPIs and sets new criteria for data collection which will be useful for future KPI analysis and benchmarking. This person has to have strong commercial acumen and excellent client relationship skills.

The data scientist takes the clean and pre-processed data set and applies complex mathematical and statistical models to predict the future. She/he has the freedom to combine or create new features, explore various machine learning supervised or unsupervised models, even deep neural networks, aiming to create meaningful predictions with minimised error.

The three roles fit together in the sense that they take the same dataset and despite working towards a different output, they still need to correlate their findings and analysis and work towards a common goal. Both the analyst and the scientist rely on the architect to maintain the data infrastructure which can support the influx of large amounts of information. The analyst and scientist often work together to exchange ideas about interesting trends, irregularities and outliers which can prove to be useful both for the visualisations and for the prediction models.

Communication and team work are probably the two most important skills that the professionals in all three roles should be comfortable with. Sharing ideas, experience and knowledge is key for driving team productivity, growth and increasing efficiency.