In this project, I am attempting to analyze the online retail dataset using Jupyter, Python (Pandas), and Python (Pandas, NumPy, and Matplotlib). The framework was around five specific questions. Which products drive the most revenue? How will demand move over a certain time frame? Which non-UK countries contribute the most? What relationship is there between quantity sold and unit price? And lastly, how certain customers can be segmented by RFM, or recency, frequency, and monetary value. To start, I implemented a short upload cell so I can load my raw CSV file, followed by a very long cleaning pipeline in order to remove credit notes and non-positive prices and quantities, standardize column names, and lastly, write a cleaned file for analysis later.

Through my coding, the results given are very clear and actionable. Specifically, a small set of SKUs concentrates the revenue from the products. 23843 ($168,469.60) is first, followed by 22423 ($142,264.75), 85123A ($100,547.45), 85099B ($85,040.54), and so on. Now, moving to seasonal analysis, which was my second question, November is the best month on average, indicating a pronounced peak late in the year. Moving to the third question, internationally speaking, the best or top non-UK markets are the Netherlands ($285,446.34), EIRE ($265,262.46), Germany ($228,678.40), and France ($208,934.31). Regarding the pricing and demand in quarter 4, the scatter of 385,081 transactions will show us a negative quantity-price relationship with the correlation being r equals -0.25. A log-to-log fit would yield elasticity at approximately -0.51. This means a 1% increase in price would be associated with an approximate 0.51 fewer units. Sales are mostly concentrated below $5.00 and under 10 units, while other items above $10.00 rarely are moved in those high quantities. Lastly, for customer value, RFM segmentation shows us 4,338 customers split into Occasional (53.9%, or 2,340), At-Risk (24.6%, or 1,065), and High-Value (21.5%, or 933). In my personal opinion, if I were advising a client, I would stock and promote the best SKUs while at the same time verifying the SKU-level margins before scaling spend. We have to put importance on building inventory as well as staffing ahead of the November peak and localize currency, shipping, and offers for EIRE, the Netherlands, Germany, and France. Moving on to the price sensitivity, it is quite modest in my opinion, so therefore I would avoid any kind of blanket discounting, instead trying out the small price increases on the most stable SKUs and using the bundles and free shipping thresholds to lift units. Especially for the larger or higher-priced items. As for customers, it's important to protect the margin on the highest-value buyers with the VIP perks as well as early access. Run personalized 14 two-month-long win-back flows for at-risk customers. And most importantly, guide the occasional buyers to their next purchase, where they wouldn't normally do that with starter bundles.

Future work will only help these insights by modeling and forecasting the seasonality and volatility with holiday indicators, estimating the elasticities at the SKU and international levels while being able to control their promotions, upgrading RFM to a more forward model, and running more controlled hypotheses and experiments to lift from VIP, win-back, and other types of promotional programs.