TAYLOR'S UNIVERSITY

Wisdom · Integrity · Excellence

ITS 65704 Data Science Principles

ASSIGNMENT

Report writing

HAND IN DATE: 5 November 2023

WEIGHTAGE: 30%

Student declaration:	
I declare that: § I understand what is meant by plagiarism § The implication of plagiarism have been explained to us by our lecturer This project is all our work and I have acknowledged any use of the published or unpublished works of other people.	
Names of Group Members	
Name	Student ID
1. Andrea Monoli	0364981
2. Linda Lischetti	0364980
3. Felicia Dossou	0364975
4. Mohammed Tariq Iqnat	0359101
5. Fathimath Amna Riznee	0356573

Avoid copy and paste job in your report and it is considered as plagiarism. Plagiarism in all forms is forbidden. Students who submit plagiarized assignment will deserved a 0 marks.

I. Introduction

Data science, considered as an intersection of math, business knowledge and computer science, has been, since its first application, a useful analytic procedure for getting insights from large clusters of data and forecasting matters. For this reason, the following document will be keen on a real world dataset, which regards e-commerce data, and will apply Data science structured path, in order to get valuable insights and predictions about customers' behaviors and transaction patterns.

A. Background of study

The landscape of commerce has been reshaped by e-commerce thanks to globalization and more digitalized mindsets, offering new avenues for business expansion and market growth. For this reason, existing and rising business realities can extend their market goals and boost their visibility by focusing into e-commerce, for buying and selling of goods and services online, which provides more efficient and cost-effective distribution channels for their products/services. E-commerce has also triggered a transformation in customer behavior, with many individuals now turning to their computers or smart devices to order goods comfortably, and enjoying the convenience of doorstep deliveries, nowadays more usual than ever.

The data we chose to analyze comes from a sales transaction dataset of a UK-based e-commerce, who has been selling gifts and homewares for adults and children through the website since 2007. Their customers come from all over the world and usually make direct purchases for themselves. There are also small businesses that buy in bulk and sell to other customers through retail outlet channels.

General purpose and motivation

By conducting this research, our team is trying to utilize the tools and techniques of data science to gain valuable insights and make predictions regarding customer behavior and transaction patterns in the context of e-commerce. As more customers turn to online platforms for their shopping needs, e-commerce businesses are presented with the opportunity to collect vast amounts of sales data, which can serve as valuable information to gain strategic advantages on the market.

This research is motivated by the need to address critical issues faced by online businesses, including the relationships between pricing strategies, sales frequency, and product selection. We aim, for instance, to uncover insights, patterns, and trends that can guide businesses in optimizing their strategies and making informed decisions.

One of the key problems we aim to address in this research is the need for e-commerce businesses to adapt and grow in an ever-evolving market. With the introduction of e-commerce, traditional businesses needed a shift in strategies and approaches to effectively serve advanced customers and remain competitive.

Addressed problems:

- Pricing Strategies: Understanding the relationship between pricing and customer behavior is a critical challenge. Incorrect pricing can lead to lost revenue or customer dissatisfaction.
- Sales Frequency: Inconsistencies in sales frequency can lead to financial instability.

- Product Selection: The choice of products or services offered can significantly impact the success of an e-commerce business. We aim to investigate how businesses can make informed decisions about their product selection, considering market trends and customer preferences.

General methods and materials

The core methods and materials we use include:

- Import of libraries such as Pandas, Google.collab, Matplotlib, Numpy, Sklearn, Folium and Seaborn
- Missing data detection and visualization on heating map
- Exploratory data analysis (EDA)
- Selection of three regression models: Linear Regression, Decision Tree Regressor and Gradient Boosting Regressor
- Application of hyperparameter tuning to find the best hyperparameters for each model
- Comparison of models using cross-validation techniques
- Prediction making with the trained model

- Visualize model performance through scatter plots, histogram and residual plots
- K-Fold Cross-Validation usage to evaluate model's performance

B. Problem statement

Is there a clear correlation between product pricing and the frequency of sales, and does allocating capital strictly to the best sellers yield fruitful results long term?

E-commerce stores struggle to establish pricing strategies that both attract customers and drive repeat purchases, particularly in a dynamic market characterized by shifting customer preferences and evolving trends. Neglecting this alignment can result in decreased sales frequency and reduced revenue, while also dealing with the challenge of selecting the right products. Failing to conduct thorough product selection and investment analyses can also lead to issues like unsold inventory, capital constraints, poor customer retention, and competitive disadvantages, endangering long-term growth.

Related questions:

- How can e-commerce businesses effectively align pricing strategies with changing customer preferences and market dynamics to boost sales frequency and revenue?
- What strategies can be employed to make informed product selection and investment decisions?

C. Objective of study

The objective of this study is to utilize data science tools and techniques, on an existing e-commerce dataset, in order to gain valuable insights and make predictions regarding customer behavior and transaction patterns.

The research, in fact, aims to address critical issues faced by online businesses, including the relationship between pricing strategies, sales frequency, and product selection.

II. Literature review

In this paper, we will base our literature review on five papers which attempt to answer similar questions to those that our research also attempts to answer. It is important to conduct research in these areas as it can vastly help companies or retailers understand consumer behavior and increase their revenue.

In the first paper (Zhao, H., Yao, X. et al. (2021), the authors seek to explain the relationship between product pricing and packaging and with consumer satisfaction as a mediating factor. There were several key findings that this paper generated. Some of the key findings are as follows: product prices significantly correlate with consumer buying behavior, the product information available on packaging influences the consumer's buying behavior and pricing of the product plays an essential role in customer satisfaction. The data collected for this research comes from 350 university students based in China. Similarly to our paper, the authors suggest putting most focus on better pricing strategies but also reiterate the importance of product information and packaging as well for better success.

The second paper focuses on the effect of discounting on sales. This paper (Kopalle, P. K., Mela, C. F., et al. (1999), explains that discounting can in fact reduce sales at baseline lines, thereby reducing sales at normal price. This research was conducted using a a basic model which follows the SCAN*PRO model of store sales posited by Wittink et al. (1988) For optimization, the authors present a normative model for dynamic pricing in a retail context. The model is based on insights from interviews with a retailer and a manufacturer. The data used for estimating the model is based on 124 weeks of A. C. Nielsen store-level data for liquid dishwashing detergent. The key findings addressed in this paper are: The results suggest that profit can be increased by as much as 7% to 31% over their current practices by balancing the trade-off between increasing sales in the current period and the corresponding reduction in baseline sales in future periods . The study introduces the concept of the "triple jeopardy," whereby promotions may lead to three negative effects on price (reduced baseline sales, increased price sensitivity, and reduced effectiveness in stealing sales from competing brands). They also conclude that excessive discounting can compromise national brands' pricing advantage.

The third paper (Gaur, V., & Fisher, M. L. (2005) is solely focused on the impact of price on sales. However in this paper, rather than using the traditional methods collecting data via interviews and surveys, they conduct their own experiment in a controlled environment. The authors state that they worked with data from 53 stores in their store selection model and that the process was essential to ensure the validity of the experiment. It was concluded to have clusters of three. Three products were chosen for the experiment. They are as follows: a family game center (unbranded, simple), ‘phonics traveler’ (branded) and a walkie-talkie (unbranded, complex). The authors note that the sales of the family game center and the phonics traveler show a downward-sloping trend with decreasing prices and conclude that price is an indicator of quality, which also supports our hypotheses that incorrect pricing strategies can lead to loss of revenue.

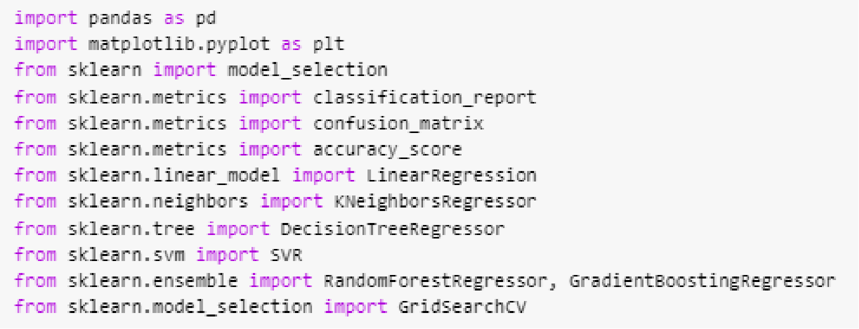
The fourth paper (Ali FRM., Diaz MC., et al. (2020) is based on research about e-cigarette unit sales by product and flavor. The data collected was licensed by IRI Inc., which included Universal Product Code sales from convenience stores, gas stations, grocery stores and more. This research found that while pre-filled cartridges remained the leading product type purchased, disposable vape sales increased as well to 19.8% of total sales between August 2019 and May 2020. By May 2020, menthol (61.8%) and tobacco (37.1%) flavors dominated the market. November 2016–August 2019 was mostly dominated by the sale of prefilled cartridges, which made up about 90% of the market by the end of this period. Previous research shows that this increase in total sales was mostly based on JUUL (a type of e-cigarette). These findings also support our hypothesis that product selection is extremely important in gaining customers. By understanding the products and which models customers prefer, and keeping on top of trends, revenue is likely to grow. As this paper stated, ‘During November 2016–August 2019, total e-cigarette unit sales in the U.S. increased nearly 300%’.

The fifth paper focuses on the impact of brand loyalty on sales (Shirin Jamal., Dr. Khurrm Sultan. (2021) This research is based on a clothing brand. The sample chosen consists of manufacturing and services both. The authors of this paper selected three different types of customer satisfaction, which are: satisfaction with the clothes; satisfaction with the sales service and satisfaction with the after-sales service with the expectation that all three types would be influenced by brand loyalty. The findings of the analysis of the results generated are that customer satisfaction with the clothes is hugely determined by brand loyalty. This is also in line with our hypothesis that customers are influenced by product selection. If the selection of products include those by brands that are popular among the target consumers, it is likely that the revenue generated will also increase.

III. Methodology

A. Tools

We mainly used the Google Colab web-browser IDE, as for the language we exclusively used Python 3.13.
We also used the following libraries: *Pandas, matplot, sklearn*



```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn import model_selection
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.linear_model import LinearRegression
from sklearn.neighbors import KNeighborsRegressor
from sklearn.tree import DecisionTreeRegressor
from sklearn.svm import SVR
from sklearn.ensemble import RandomForestRegressor, GradientBoostingRegressor
from sklearn.model_selection import GridSearchCV
```

We first imported the dataset published by *Gabriel Ramos* on *kaggle.com* which houses data for an online ecommerce store since the year 2007 into our IDE, in this case it was *Google Colab* with the preferred language being *Python* for access to its various useful libraries, in this instance *Pandas, Numpy, and Matplot* alongside others.

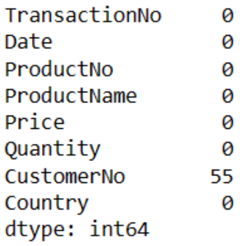
B. Data extraction

We started off by researching and exploring extensive open source datasets from online venues, until we landed on an effective dataset for the transactional sales of a UK-based e-commerce store selling gifts and homewares for adults and children through their website since the year 2007, which provided us with an extensive data set that contains 500k rows and 8 columns, which we found sufficient for our research paper.

C. Cleaning the data

We then proceeded cleaning the dataset by first handling the missing values that were on there, from our observations only CustomerNo had missing values.

output:

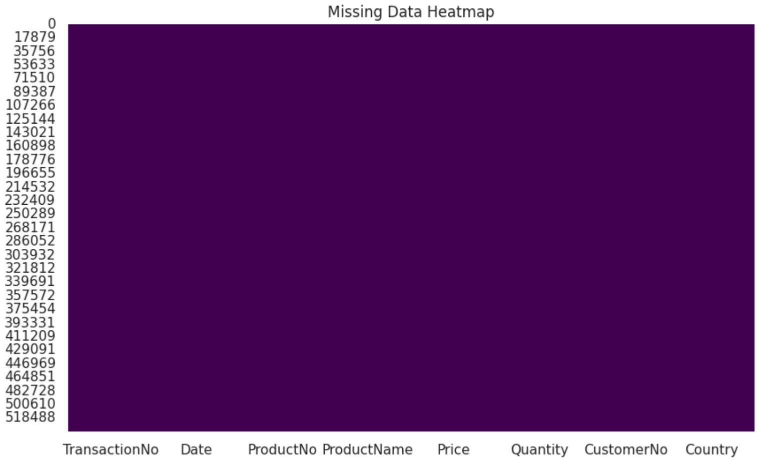


```
TransactionNo     0
Date              0
ProductNo         0
ProductName       0
Price             0
Quantity          0
CustomerNo       55
Country           0
dtype: int64
```

From the output, it's clear from a basic observation that the number of missing values is miniscule (55) relative to the overall number of rows (536350); the only column with missing values is CustomerNo, and so we deleted these rows after concluding their irrelevant weight to the overall output.

After that, we took extra precaution to make sure that there aren't any more missing values by visualizing it with a heatmap using the seaborn library.

Output:



After we've fully concluded that there are no missing values in our dataset, then we moved forward with the cleaning process by handing duplicate rows.

We've found that the number of duplicate rows in the data set were 5200, and due to the insignificance relative to the overall number of rows we've again decided to drop them before moving on to the next step, which in this case is handling absurd values.

To prevent unexpected results in our analysis we chose to remove rows with values that don't align with their intended representation. We pinpointed incoherent values as follows:

Quantity <= 0: This signifies a cancellation, as noted by the dataset author and since our analysis is centered on actual sales, we considered this as irrelevant.

Price *< 0:* This is simply an anomaly.

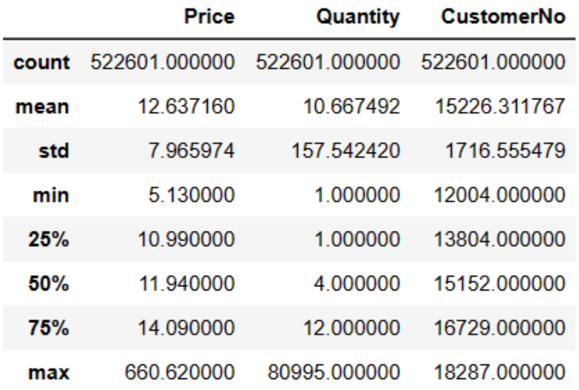
TransactionNo beginning with `C`: using the same rational as *Quantity* <= 0, cancellations transactions are irrelevant to our study.

D. Exploration and observations

After cleaning up the data and ensuring that it's fit we then move forward to exploring and preparing the data, we start off by structuring and shaping our investigation.

Our dataset covers 305 days of transactions for 100 products in 38 different countries.

Overall information shows that our DataFrame contains 1000 entries and 5 columns. It also displays the data type of each column, the count of non-null values, and the memory usage, which is useful for managing and optimizing memory when working with large datasets.

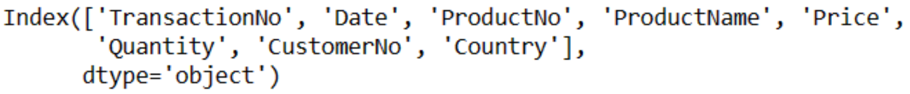


	Price	Quantity	CustomerNo
count	522601.000000	522601.000000	522601.000000
mean	12.637160	10.667492	15226.311767
std	7.965974	157.542420	1716.555479
min	5.130000	1.000000	12004.000000
25%	10.990000	1.000000	13804.000000
50%	11.940000	4.000000	15152.000000
75%	14.090000	12.000000	16729.000000
max	660.620000	80995.000000	18287.000000

After that we went ahead and utilized the *describe()* function on the *pandas* library to display a clear statistical summary of our dataset, which we thought would help provide us with valuable insight into its distribution structure.

A simple *.columns()* function to present a dictionary of all of the column names in our dataset.

Output:

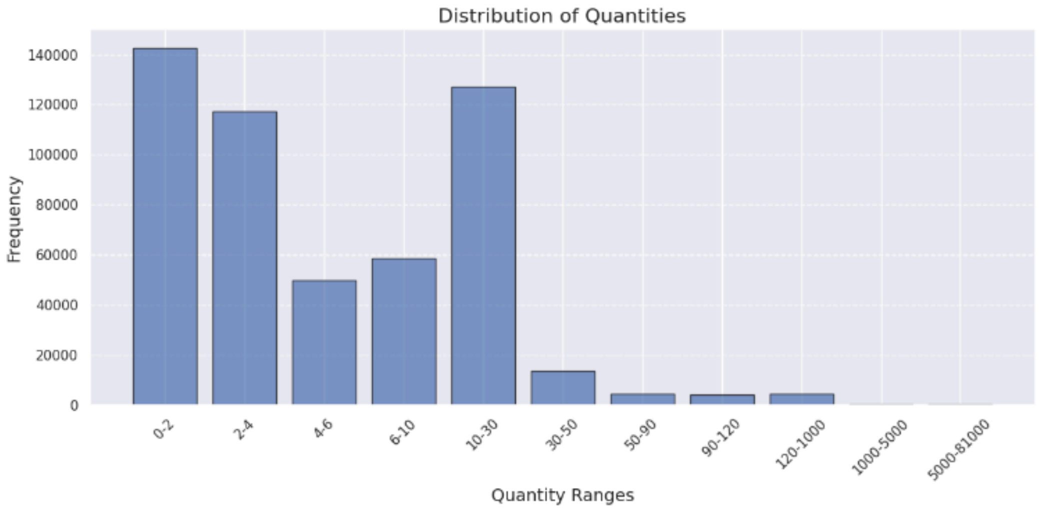


```
Index(['TransactionNo', 'Date', 'ProductNo', 'ProductName', 'Price',
       'Quantity', 'CustomerNo', 'Country'],
      dtype='object')
```

From an observation of the output we could see that our dataset contains 305 days of transactions for 100 products in 38 different countries.

Afterwards, we move to exploring the variables within our dataset, first by visualizing the distribution of quantities.

Output:

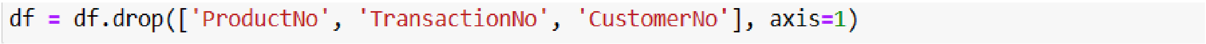


As we can see from the Histogram output above, a quick observation would enough to see that the majority of transactions are in done in tranches of small quantity with transactions being done orders of 1 or 2, followed by transactions done in batches of 10 to 30 items with the lowest being transactions of 50 items and above.

After that we moved forward to cleaning our data further by reducing it to the top 100 of most sold products, we did that to reduce error provoking data that was bound to occur due to the density and size of the dataset.

After sorting in ascending order than filtering, we found that the top 3 most sold products are *Peper Craft Little Birdie, Medium Ceramic Top Storage Jar, and Popcorn Holders,*

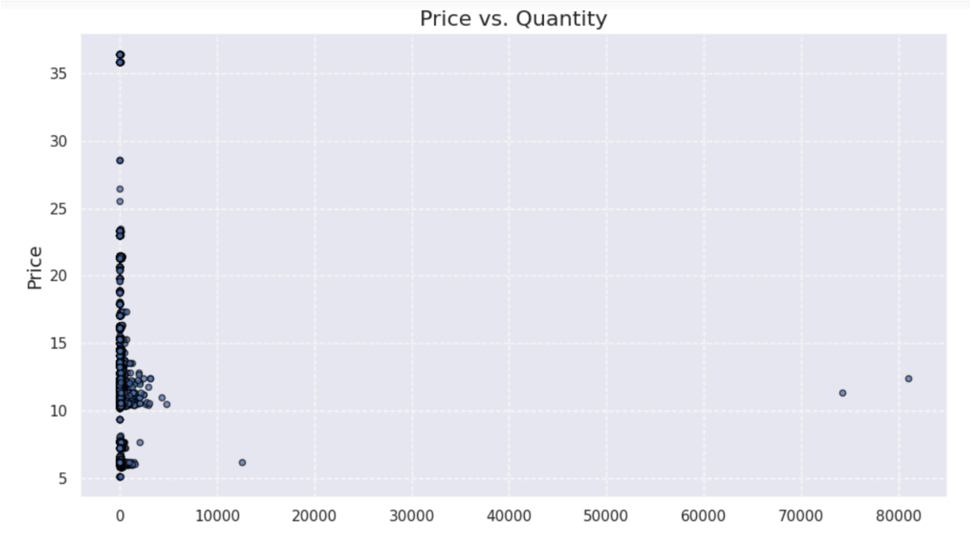
Hereafter, we proceed with cleaning the dataset further by deleting irrelevant attributes to our study.



```
df = df.drop(['ProductNo', 'TransactionNo', 'CustomerNo'], axis=1)
```

Carrying on with the cleaning process we then start filtering outliers in the dataset due to their insignificance from our desired model, as followed.

Output:

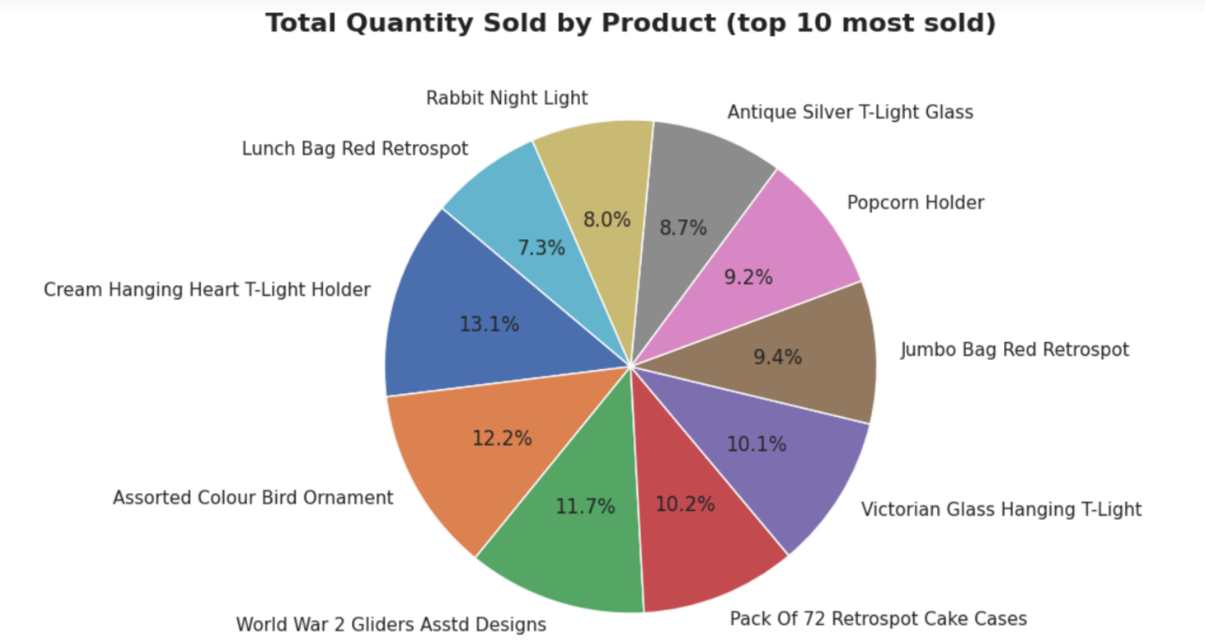


As we can clearly see from the scatter chart above, we’ve got a few outliers in quantity within our dataset, and due to the nature of our analysis being based around understanding customer behavior, we’ve came to the conclusion that it’ll be valuable to remove price and quantity outliers to gain more accurate insight into business operations, therefore, we’ve decided to remove the highest 3% of values in both `Price` and `Quantity`.

Visualizations

After we were done cleaning the data and getting it ready for our analysis we started visualizing the desired aspects of the dataset so we could get a clear picture of what is actually going on within the ecommerce store; we started that off by getting the 10 most sold products on the store visualized in a pie chart:

Output:

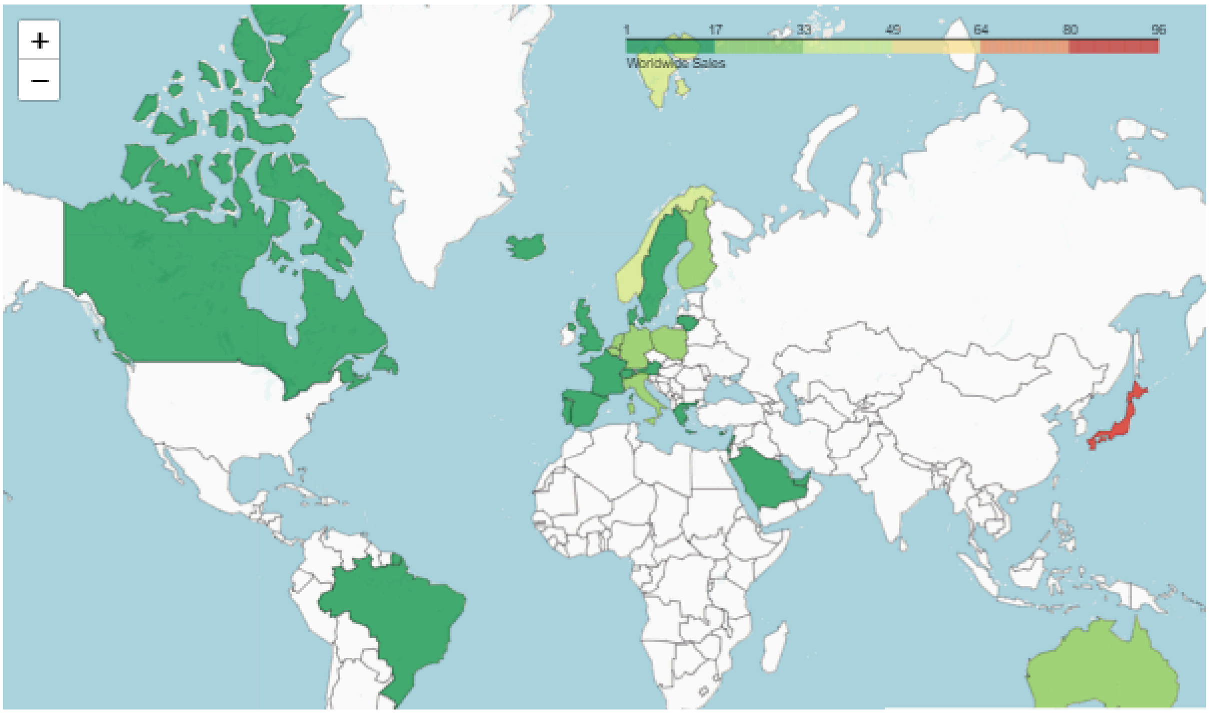


As we can clearly see from the pie chart above, there’s a near even distribution amongst the top 10 sold items on the store, with the highest being *Cream Hanging Heart T-Light Holder* coming in at 13.1% and the lowest being *Lunch Bag Red Retrospot* coming in at 7.3%, from a first glance observation, we’ve see clear merit in focusing on items that make up ~76% of sales in the top 10 bracket and discarding the rest.

We moved forward with visualizing the variables, this time, we went ahead and visualized items sold by geographic location to the detriment where items are mostly sold and we're least sold.

We utilized the Folium library to create an interactive map visualizing worldwide sales data from our dataset. We used a GeoJSON world map as the geographical base and created a choropleth map to represent sales quantities by country. The geo_data parameter pointed to the GeoJSON file representing country boundaries, while the data parameter referenced the dataset 'df' containing the sales data. We specified the 'Country' column in the DataFrame as the country identifier and the 'Quantity' column as the data to visualize. The map's coloring was determined by the 'RdYlGn_r' color scheme, with opacity settings for fill and lines. The map also featured a legend denoting the range of sales quantities.

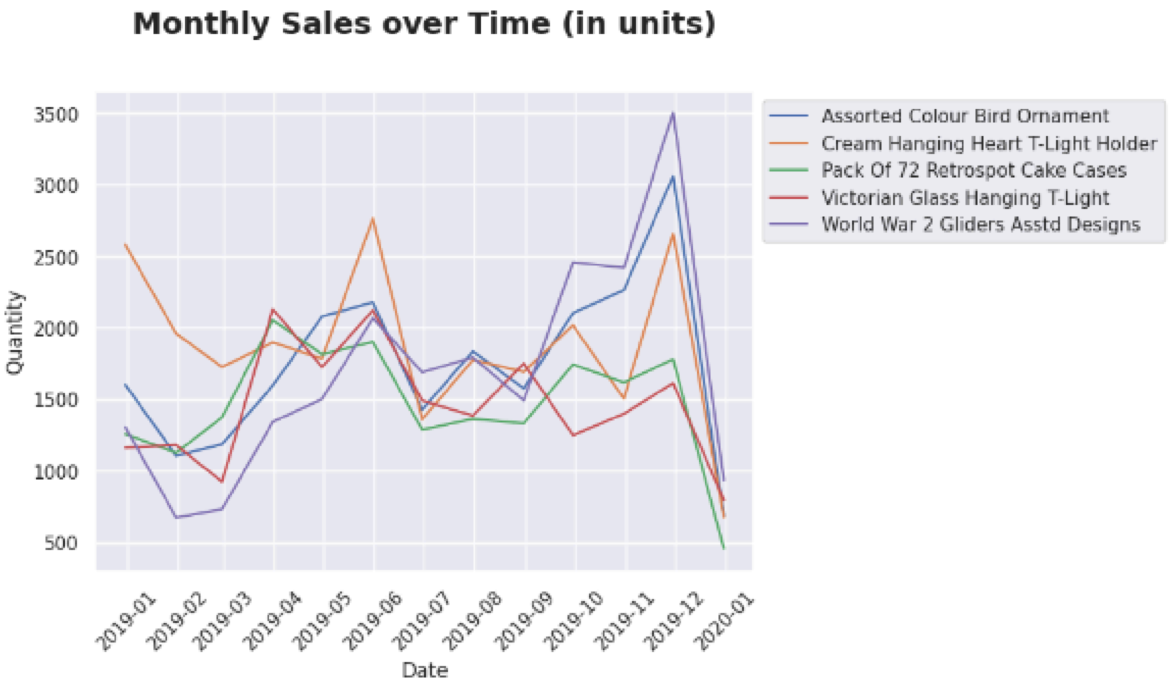
Output:



As we can clearly observe from our map illustration, the majority of sales volume comes from Canada in NA, Brazil in South America and Spain, France, Switzerland, Austria, Greece, Lithuania and the United Kingdom as well as Greenland and Sweden in the EU, and finally Saudi Arabia in the Middle east, we noted that down then we moved forward to visualizing sales volume over time.

Then we draw focus on the top 5 products with the highest total quantity in the dataset. We then extracted the names of these top products and created a new DataFrame, containing only rows where the 'ProductName' matches one of these top product names. Finally, we printed the unique product names to help us narrow our analysis to the top-performing products. In the second code snippet, we prepared the data for a time series analysis of monthly sales. First, we converted the 'Date' column to datetime format with the specified date format. Then, we set the 'Date' column as the index of our dataframe to enable time-based analysis. We created a line plot to visualize the monthly sales trends for each of the top 5 products. The plot shows the quantity of each product sold over time, with the x-axis indicating the date and the y-axis representing the quantity. The legend provides labels for each product's sales trend. This code allows us to visually explore how the top products' sales have evolved over time, identifying patterns or trends.

Output:



As we can clearly see in the chart above sales throughout the year fluctuate in a stable manner within a very tight bound between 1400 and 2300 units leading up to holiday season, thanksgiving in NOV and Christmas in DEC where we saw a large surge in sales volume all the way up to a peak of 3500 items followed by 2900 items for WW2 gliders and Colour Bird Ornaments before a considerable crash to all time lows when fear started spreading about the Covid-19 virus and countries worldwide started implementing mandatory quarantine, that chart is a perfect visual representation of how much seasonality affects ecommerce businesses and how the effects of unexpected events could often lead to extreme decline in sales and sometimes even bankruptcies due to inventory tied capital.

F. Developing Machine Learning

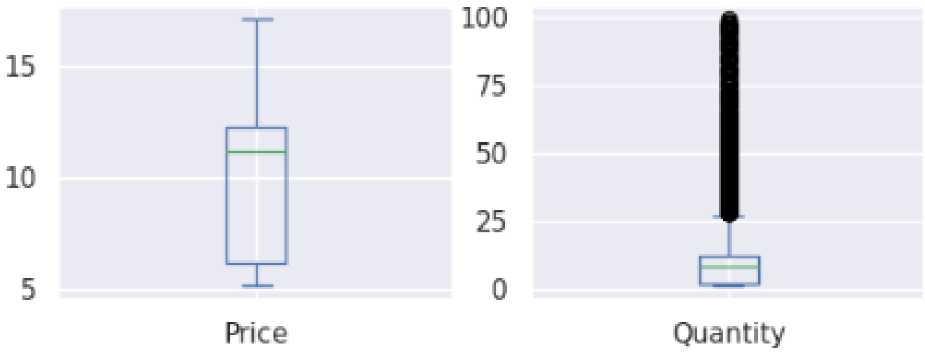
After we felt that the visual representation of our dataset was sufficient for our study, we proceeded to develop our Machine Learning and Prediction Models after conducting feature engineering. Additionally, we split out a validation dataset to evaluate and validate our models effectively. This process allowed us to ensure our data was properly structured and ready for the next stages of the machine learning pipeline.

We prepared the data for machine learning and deep learning prediction models by first categorizing the products into four distinct categories. We start by providing a summary of the statistical description of the dataset using the describe() function, which offers insights into the distribution of numerical features.

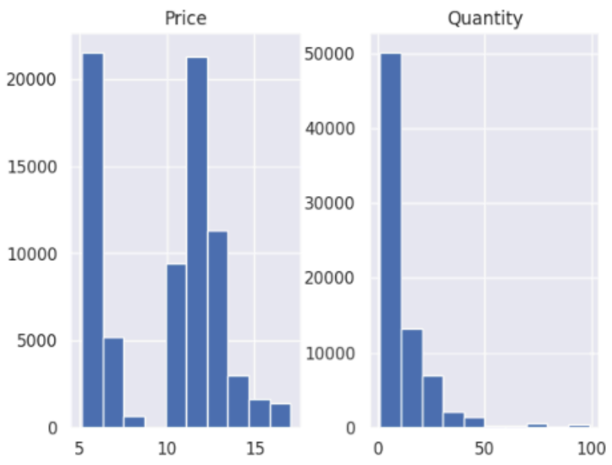
We then proceed to create product categories based on the 'ProductName' column. After analyzing the product names, we identify four categories:

- Kitchen and Dining Products,
- Home Decor and Accessories,
- Stationery and Craft Supplies,
- Fashion and Accessories.

To visualize the data further, we create box and whisker plots to examine the distribution of prices and quantities within the dataset. These plots provide insights into the spread of values, medians, and the presence of outliers.



The price plot shows a spread of values greater above the median compared to below it. For quantities, the majority of data falls between 10 and slightly above 75, with many values significantly higher than the central bulk, indicating variations in the data.

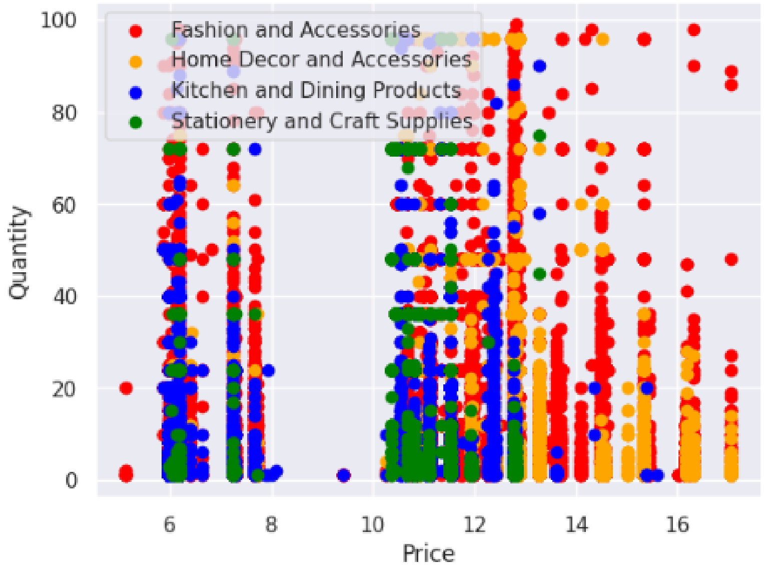


Lastly, we generate histograms to visualize the overall distribution of data.

These histograms reveal the frequency distribution of different numerical features within the dataset, in this instance we can see clear disparity in quantity distribution relative to price where the majority of orders go from highest amount to lowest amount as the quantity increases and that the majority of orders fall between 10 and 14 when it comes to price.

Next, through each of the four categories, we created a scatter plot for each category using 'Price' on the x-axis and 'Quantity' on the y-axis. The c parameter in the scatter function specifies the color of the data points for each category. The label parameter assigns a label to each category for the legend. The legend at the top right of the plot distinguishes between the four categories using the specified colors. This multivariate plot allows for a visual comparison of how price and quantity vary across different product categories.

Output:

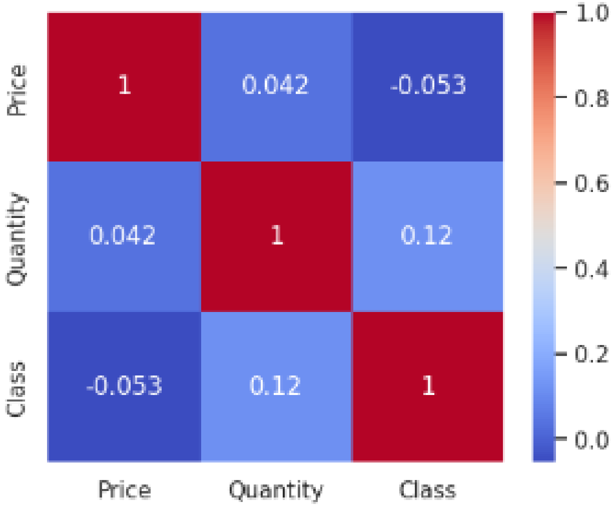


And as we can see from the scatter chart above, we get a clearer image of the order aggregation being mostly housed between 10 and 14 in terms of price reaching its highest quantity at between at 13 dominated by Fashion and Accessories.

Afterwards, we went ahead and created a correlation matrix to analyze the relationships between different numerical features in the dataset using the LabelEncoder from the sklearn.preprocessing module to convert the categorical values in the 'Class' column into numerical labels. The unique values for 'Class' are printed, showing the mapping of categories to numerical labels.

We proceeded with correlation analysis.

Output:



In the correlation analysis, we observed that the relationships between Price and Quantity, Price and Class, and Quantity and Class are all characterized by very weak to weak linear correlations. The minimal and often negligible correlation coefficients indicate that any associations between these variables are not practically significant for making predictions or drawing meaningful conclusions about the data.

Consequently, we moved forward with building a prediction model for our dataset and we started that off by first preparing the data.

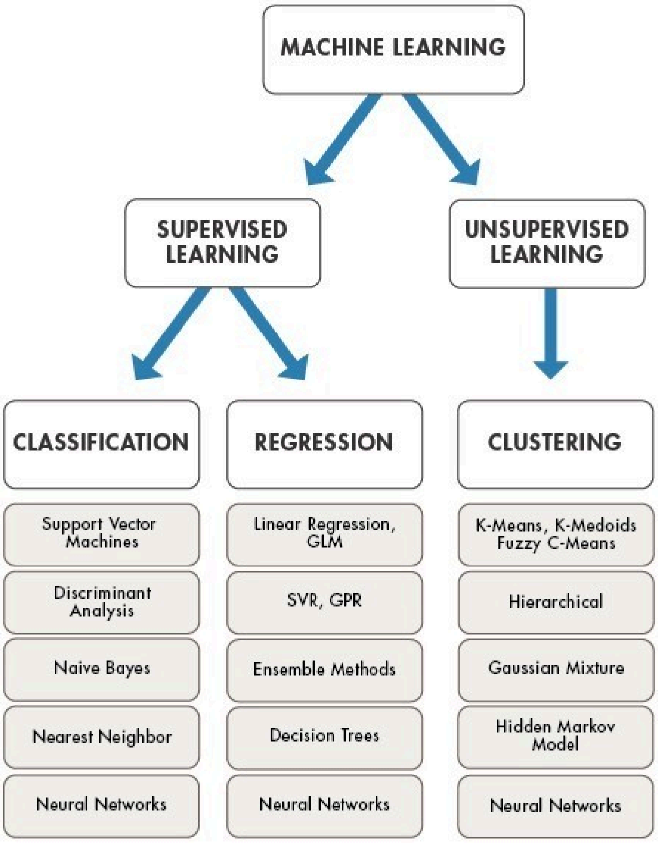
First, because of the model we're using to predict sales, we're doing one-hot encoding which means that we convert categorical variables into a numerical format on our categories and next we split the date columns into three parts.

To efficiently test and validate our sales prediction models, we initially dealt with our large dataset by taking a 10% sample. This smaller subset served as a quick testing ground to identify the most suitable model. Once the optimal model was determined on the sample data, we planned to leverage the entire dataset for training and comprehensive evaluation.

Modelling

For this purpose, we first created a 10% sample of our dataset, named "sampled_df," using a random seed to ensure reproducibility (in this case, random_state=42). We then separated our dataset into two main components: features (X) and the target variable (y). Features (X) consist of all columns except "Price," which is the target variable.

Subsequently, we split the "sampled_df" into training and validation sets. The training set, represented by "X_train" and "y_train," contains 80% of the data, while the validation set, denoted as "X_validation" and "y_validation," holds the remaining 20%. This partition allows us to evaluate the model's performance by comparing its predictions to the actual prices on the validation data. Once we have successfully identified the best model on the sample dataset, we will apply it to the entire dataset for more comprehensive predictions and insights.

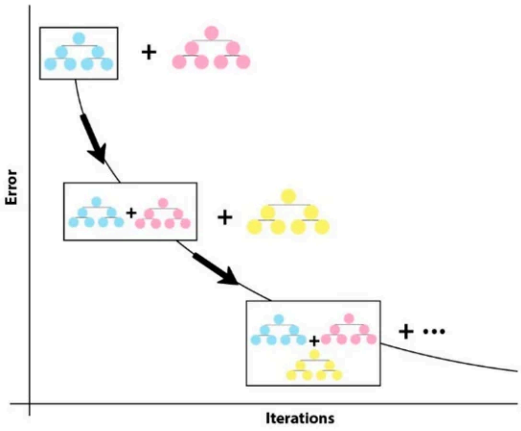


After we were done with that we then started on the actual model construction. In choosing our machine learning algorithms for this project, we considered the nature of our dataset and the specific goals we wanted to achieve. We focused on supervised machine learning algorithms because our data was labeled, and we aimed to predict a continuous value, which aligns with regression tasks.

We evaluated and selected three primary algorithms: Linear Regression, Decision Tree Regressor, and Gradient Boosting Regressor. Here's why we chose each of them:

Linear Regression: Linear regression is a simple yet effective algorithm for predicting continuous values. Given that we wanted to predict product prices, a regression-based approach was appropriate. Linear regression works well when there's a linear relationship between input features and the target variable. In our case, it's reasonable to assume that certain product attributes would have a linear impact on pricing.

Decision Tree Regressor: Decision trees are versatile and capable of capturing complex relationships within the data. They are well-suited when the data's structure is not strictly linear. In our dataset, various product attributes, categorical or continuous, could interact in complex ways to influence prices. Decision trees can discover these non-linear patterns, making them a valuable choice.

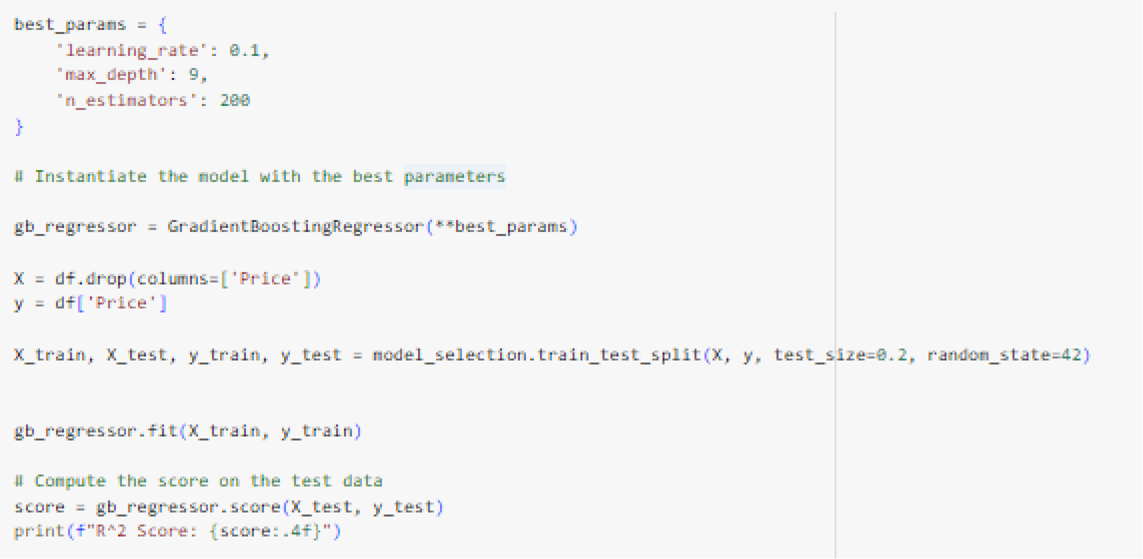


Gradient Boosting Regressor: Gradient boosting is a supervised ensemble algorithm used for regression and classification. It is based on the boosting technique. It is an algorithm based on a weak learner tree. In the decision tree, a model is created for the weak learner and a prediction is made. Calculation errors are passed on to the next weak learning tree, as shown in **Figure 4**.

This last model is the strong learner and is the weighted average of all models. Apache Spark MLlib is only supported for binary classification.

To ensure optimal performance, we employed hyperparameter tuning using grid search and conducted a 3-fold cross-validation to evaluate and compare each model's performance.

The results favored the Gradient Boosting Regressor, with a mean squared error of approximately -0.419, indicating its superior predictive capabilities. This model excelled with a learning rate of 0.1, a maximum depth of 9 and 200 estimators.



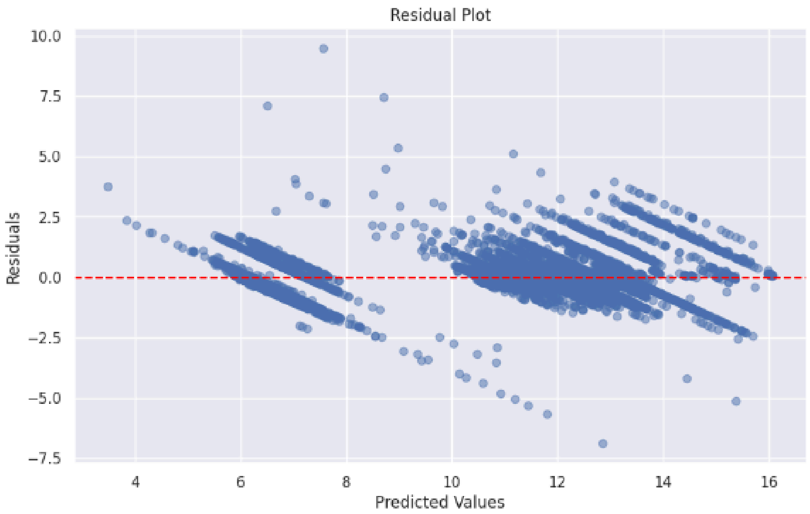
```
best_params = {
    'learning_rate': 0.1,
    'max_depth': 9,
    'n_estimators': 200
}

# Instantiate the model with the best parameters

gb_regressor = GradientBoostingRegressor(**best_params)

X = df.drop(columns=['Price'])
y = df['Price']

X_train, X_test, y_train, y_test = model_selection.train_test_split(X, y, test_size=0.2, random_state=42)


gb_regressor.fit(X_train, y_train)

# Compute the score on the test data
score = gb_regressor.score(X_test, y_test)
print(f"R^2 Score: {score:.4f}")
```

We trained the Gradient Boosting Regressor on the entire dataset, attaining an impressive R^2 score of 0.9563. This score signifies that the model is performing exceptionally well, as it explains a significant portion of the variance in the response variable, coming close to the perfect value of 1.

Predictions

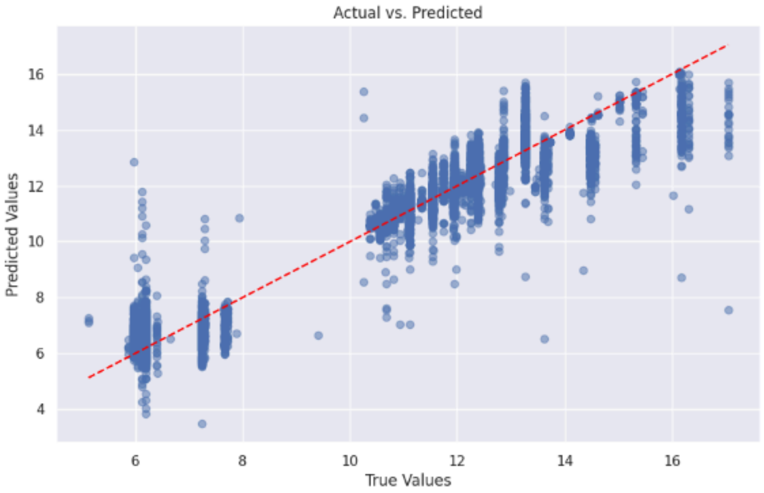
After that we then moved forward to training our different models and choosing whichever one has superior parameters, so that we could use it to predict optimal prices.

Output:



"Actual vs. Predicted" plot is a comparison between the true values (y_test) and the predicted values (y_pred). Each point in this scatter plot represents an actual sales value against its corresponding predicted value. The red dashed line serves as a reference, indicating perfect predictions when all points align along this line.

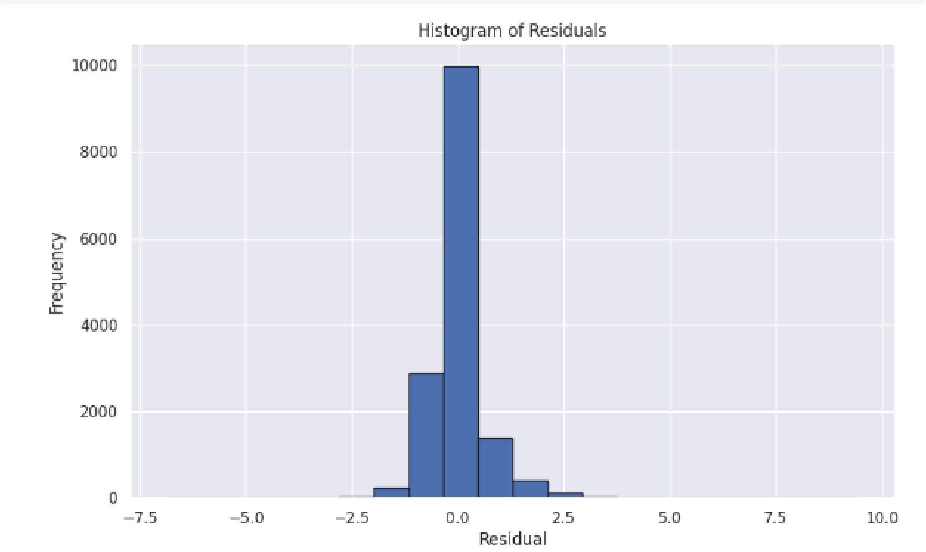
Output:



As we can see from our prediction model above, we have a clear scatter chart that indicates predicted sales being heavily aggregated between 10 and 14 which provides really strong confluence with the visualized charts presented above, which give a highly strong intuitive notion that indicates focusing on sales within expected ranges would prove highly fruitful for e-commerce stores modeling their data set the same way we've done above. The scatter points show how the residuals vary concerning the predicted values. Ideally, we want the residuals to be randomly scattered around the horizontal line at y = 0. The red dashed line represents this ideal condition, indicating that the model's predictions are unbiased.

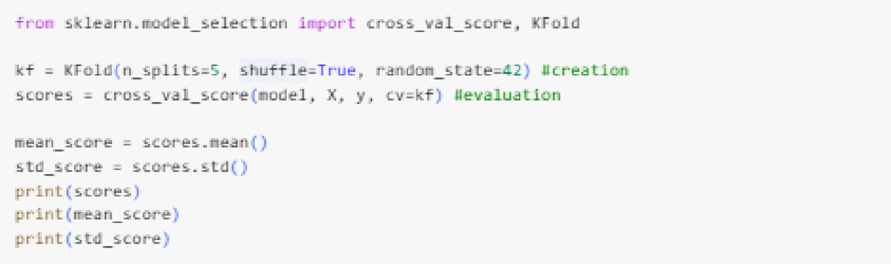
We also computed a residual plot in order to see the distribution of the model's residuals.

Output:



The histogram of residuals reveals the frequency distribution of the model's prediction errors. Ideally, we want the residuals to be normally distributed and centered around zero, indicating that the model is making consistent and unbiased predictions.

After we finished working on our predictive module we then went ahead and validated it's accuracy/assessed its performance using K-fold Cross-Validation as follows:



```
from sklearn.model_selection import cross_val_score, KFold

kf = KFold(n_splits=5, shuffle=True, random_state=42) #creation
scores = cross_val_score(model, X, y, cv=kf) #evaluation

mean_score = scores.mean()
std_score = scores.std()
print(scores)
print(mean_score)
print(std_score)
```

K-Fold Cross-Validation is a technique we use to assess the performance of our machine learning model. It's like a test we give the model to see how well it can handle different parts of the data. In this case, we're doing a 5-fold cross-validation, which means we divide the data into five parts, shuffle it for randomness, and set a random state for consistency.

Output:

```
[0.70650086     0.72738301     0.72553696     0.72501641     0.72317314]
0.7215220745180935 0.007629458801148264
```

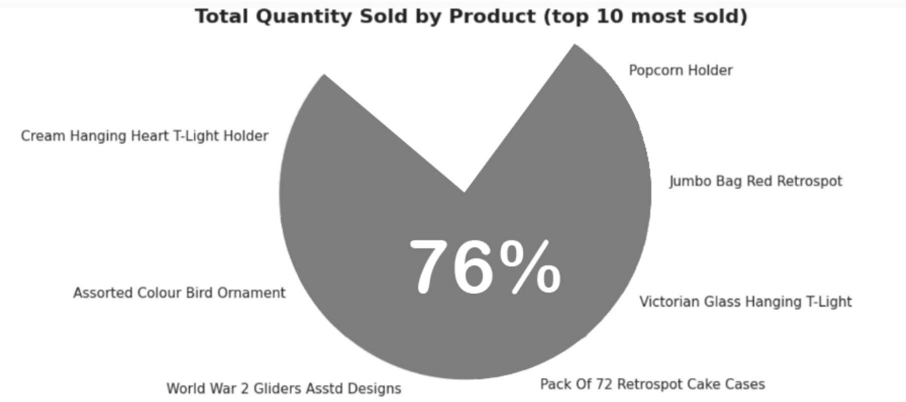
The results we get are individual scores for each fold, representing how well the model performs on different data subsets. The average score, which is about 0.7215, gives us an overall measure of how good our model is. The standard deviation, approximately 0.0076, tells us how consistent or variable the scores are.

Therefore, our model shows a consistent performance across different data subsets, with an average score of 0.7215. This means we can trust it to work well on new, unseen data. The low standard deviation of 0.0076 indicates that the model's performance doesn't change much between different data subsets. This cross-validation process helps us understand how reliable and consistent our model is.

IV. Results

In a brief conclusion, our analysis and predictive model have provided valuable insights into the business operations of e-commerce stores that broadly operate within the same market dynamics as the one we choose.

From our observations, the results suggest that aggregating capital and attention on top-selling products in this instance *`Cream Hanging Heart T-Light Header`, `Assorted Color Bird Ornament`*, `*World War 2 Asstd Designs*`, `*Retrospot Cake Cases,*` `*Glass Hanging T-Light*`, `*Jumbo bag Red Retrospect*`, `*Popcorn Holder*` which represent a total of (~76%) of the top 10 items sold on the store while discarding the lowest sold products completely, would prove quite fruitful for the companies expenses, cutting off costs, increasing profitability by focusing on demanded merchandise while also decreasing the risk of intrapping capital in unsold and unwanted products, our results also show clear disparity in demand based on seasonality, leaving us with the conclusion that aggregating advertisement capital near top selling seasons in this case from NOV to DEC while cutting off advertisement the rest of the year would save the company a large amount of capital and would provide a substantially higher return on investment for the funds deployed.



Our results also highlight the importance of risk prevention and mitigation when it comes to tail curve events like COVID-19, amongst others, which more often than not result in a significant drop in sales leaving ecommerce companies with large amounts of capital tied up in unsold merchandise, which when not accounted for by having capital saved up for emergencies leads to them being forced to file for bankruptcy and close the business down.

V. Conclusion and recommendations

E-commerce has grown increasingly more prevalent in recent years due to digitalization. It is understood that e-commerce has transformed customer behavior making it imperative for online retailers to conduct research into how exactly they can generate more revenue and attract more customers online.

In this paper, we attempted to analyze and understand purchase behavior so as to provide valuable insight into pricing strategies, what attracts customers and product selection.

Therefore, using our dataset, we have gone through all the steps of data processing and analyzation and proceeded to develop our Machine Learning and Prediction Models. We discovered that The results favored the Gradient Boosting Regressor. This model excelled with a learning rate of 0.1, a maximum depth of 9, and 200 estimators. Our model has an accuracy rate of 0.7215.

Our results suggest that the top selling 10 products make up 76% of the revenue generated, thereby answering our question as to how important product selection can be. If the retailer focuses more on these products and discards lower selling products, they’re more likely to generate more profit and cut down on losses. Having top-selling items in stock will also attract more customers. Results also show a vast difference in demand based on seasonal changes. We can conclude that retailers should take note of what sells most in what season and focus on those products accordingly.

VI. References

- *GeoJSON Maps of the globe*. (n.d.). https://geojson-maps.ash.ms/
- Python, R. (2023, February 1). *Python Folium: Create web maps from your data*. https://realpython.com/python-folium-web-maps-from-data/
- *Simple Plot — Matplotlib 3.8.0 documentation*. (n.d.). https://matplotlib.org/stable/gallery/lines_bars_and_markers/simple_plot.html#sphx-glr-gallery-lines-bars-and-markers-simple-plot-py
- Zhao, H., Yao, X., Liu, Z., & Qin, Y. (2021). *Impact of pricing and product information on consumer buying behavior with customer satisfaction in a mediating role. Frontiers in Psychology, 12.* https://doi.org/10.3389/fpsyg.2021.720151
- *E-commerce business transaction. (2022, May 14). Kaggle.* https://www.kaggle.com/datasets/gabrielramos87/an-online-shop-business/
- Dener, M., Ok, G., & Orman, A. (2022). *Malware detection using memory analysis data in big data environment. Applied Sciences, 12(17), 8604.* https://doi.org/10.3390/app12178604
- Admin. (2021, September 3). Emerging Technologies - *Machine Learning | IT Consulting Service Provider Company In San Antonio. IT Consulting Service Provider Company in San Antonio.* https://ylconsulting.com/machine-learning/
- Friedman, J. H. (1999). *Greedy Function Approximation: a gradient boosting machine. IMS 1999 Reitz Lecture.* https://jerryfriedman.su.domains/ftp/trebst.pdf
- Kopalle, P. K., Mela, C. F., & Marsh, L. C. (1999). *The dynamic Effect of discounting on sales: Empirical analysis and normative pricing implications. Marketing Science, 18(3), 317–332.* https://doi.org/10.1287/mksc.18.3.317
- Gaur, V., & Fisher, M. L. (2005). In‐Store experiments to determine the impact of price on sales. *Production and Operations Management*, *14*(4), 377–387. https://doi.org/10.1111/j.1937-5956.2005.tb00227
- Ali, F. R. M., Diaz, M. C., Vallone, D., Tynan, M. A., Cordova, J., Seaman, E. L., Trivers, K. F., Schillo, B., Talley, B., & King, B. A. (2020). E-cigarette unit sales, by product and flavor type — United States, 2014–2020. *Morbidity and Mortality Weekly Report*, *69*(37), 1313–1318. https://doi.org/10.15585/mmwr.mm6937e2
- Shirin Jamal, Dr. Khurrm Sultan, (2021) Impact of Brand Loyalty on Customer Satisfaction (An Empirical Analysis of Clothing Brands), *Turkish Journal of Computer and Mathematics Education Vol.12 No. 10 (2021), 7085-7093*