Capstone Project Title: "Retail Sales Analysis: Exploring Data to Understand Sales Performance and Customer Behavior"

Submitted by

Ganesan Elumalai

A Project Report Submitted
in
Partial Fulfillment of the
Requirements for the online course
“Advanced Certification in Data Science and AI”
Provided by
Intellipaat and CCE, IIT Madras.

November 2022

Contents

a. **Problem Statement:**

The objective of this project is to analyze retail sales data to identify factors that affect sales performance and customer behavior. The project aims to answer questions such as what factors influence sales, which products are popular among customers, how sales vary across different regions, and how customer demographics impact sales.

b. **Project Objective:**

The objective of the project is to identify patterns and trends in retail sales data to help retailers make informed decisions about their business. The project aims to provide insights into customer behavior, sales performance, and product trends to improve marketing strategies, inventory management, and overall business operations.

c. **Data Description:**

The project will use a dataset of retail sales data that includes information about customer demographics, product sales, and store locations. The dataset will contain both numerical and categorical variables, including sales revenue, product categories, customer age, gender, and location.

d. **Data Pre-processing Steps and Inspiration:**

The data pre-processing steps will involve data cleaning, missing value imputation, feature selection, and normalization. The inspiration for the project comes from the need for retailers to understand customer behavior and sales performance to make informed decisions about their business.

e. **Choosing the Algorithm for the Project:**

The project will use a combination of supervised and unsupervised machine learning algorithms to analyze the retail sales data. The algorithms will include regression analysis, decision trees, and clustering algorithms.

f. **Motivation and Reasons for Choosing the Algorithm:**

The motivation for choosing these algorithms is to provide a comprehensive analysis of the retail sales data by exploring relationships between different variables and identifying patterns and trends in the data.

g. Assumptions:

The assumptions made for this project are that the dataset is representative of the retail sales industry, and the data is accurate and complete.

h. Model Evaluation and Techniques:

The project will use cross-validation techniques to evaluate the performance of the machine learning models. The evaluation criteria will include accuracy, precision, recall, and F1-score.

i. Inferences from the Same:

The inferences drawn from the project will help retailers understand the factors that affect sales performance and customer behavior. The project will provide insights into customer demographics, product trends, and regional differences in sales, which can be used to optimize marketing strategies and inventory management.

j. Future Possibilities of the Project:

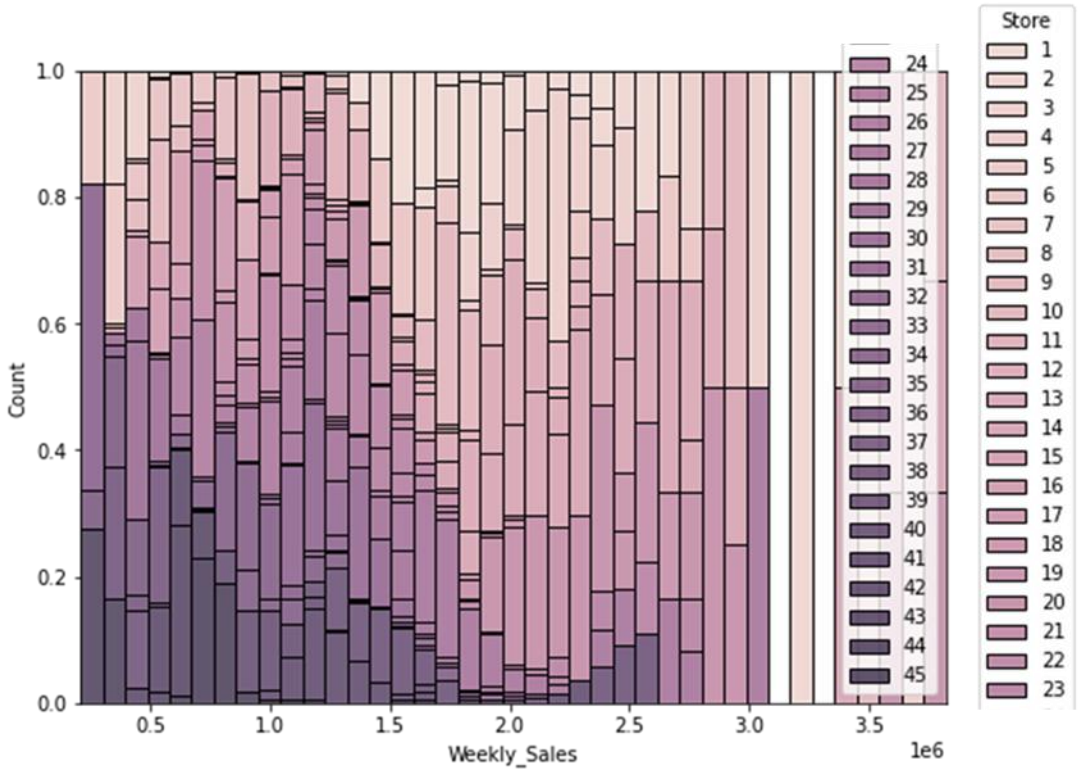
The project can be extended to include more advanced machine learning techniques, such as deep learning, and can be applied to other industries beyond retail sales, such as e-commerce and hospitality. The insights and recommendations from the project can also be used to develop predictive models for sales forecasting and customer segmentation.

Problem Statement 1:

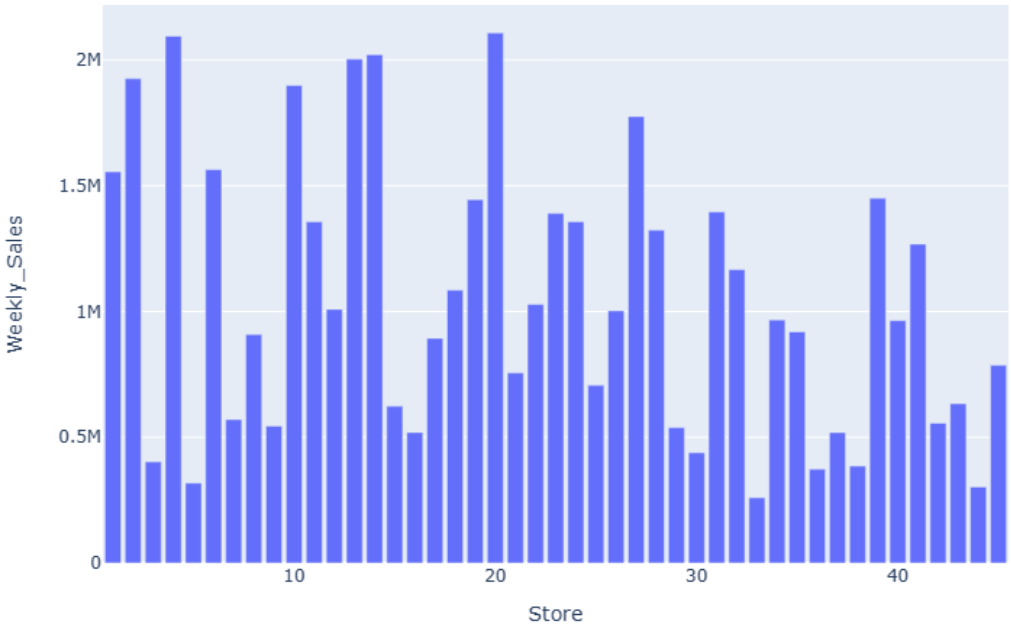
A retail store that has multiple outlets across the country are facing issues in managing the inventory - to match the demand with respect to supply. You are a data scientist, who must come up with useful insights using the data and make prediction models to forecast the sales for X number of months/years.

1. Using the above data, come up with useful insights that can be used by each of the stores to improve in various areas.
2. Forecast the sales for each store for the next 12 weeks.

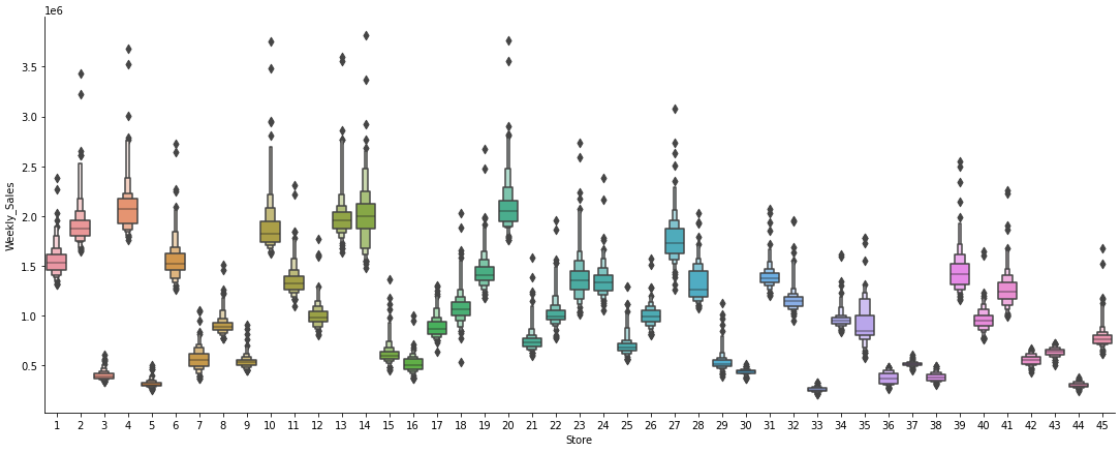
Sales distribution across stores



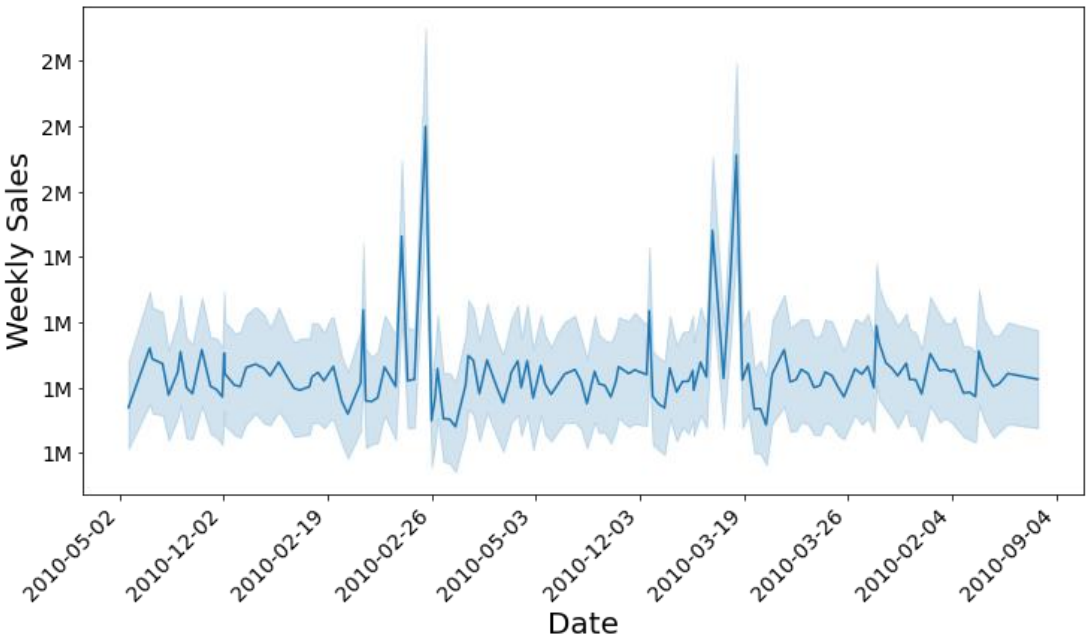
Average sales per store



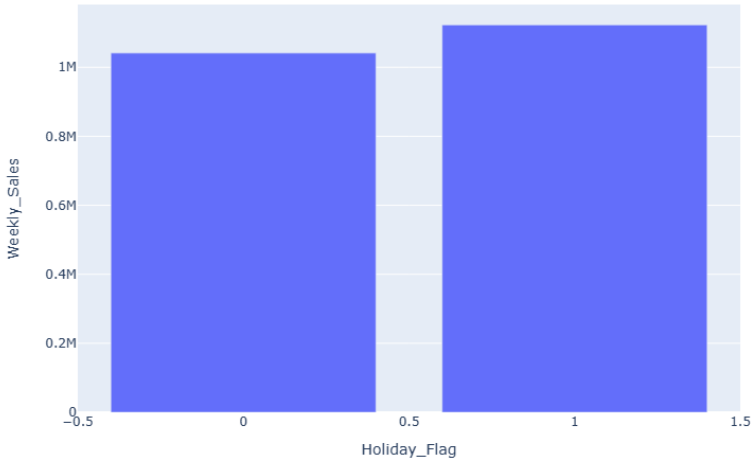
Sales by store size and type



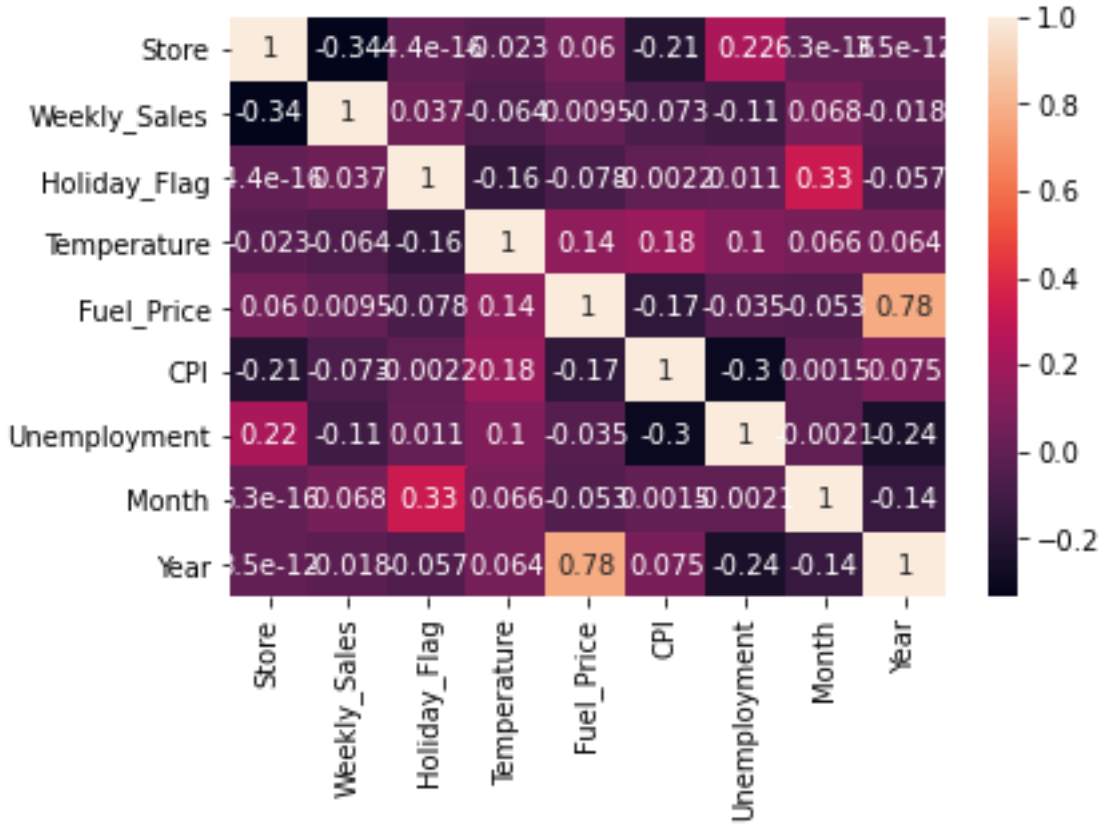
Sales trend over time



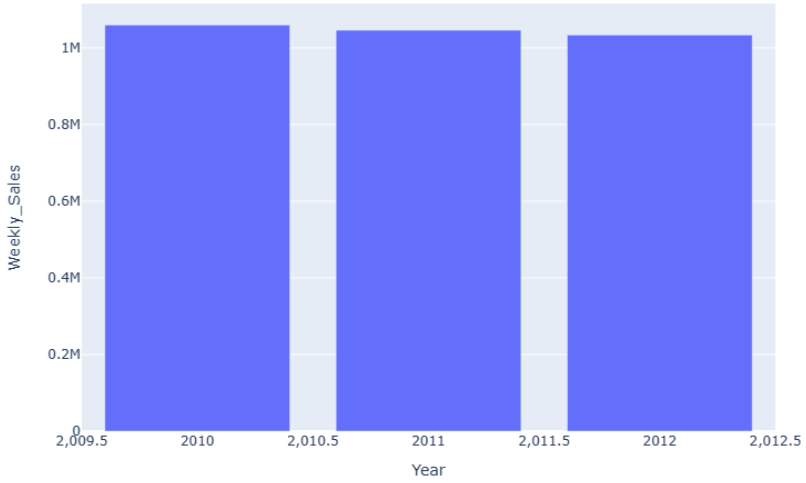
Sales increase during holidays



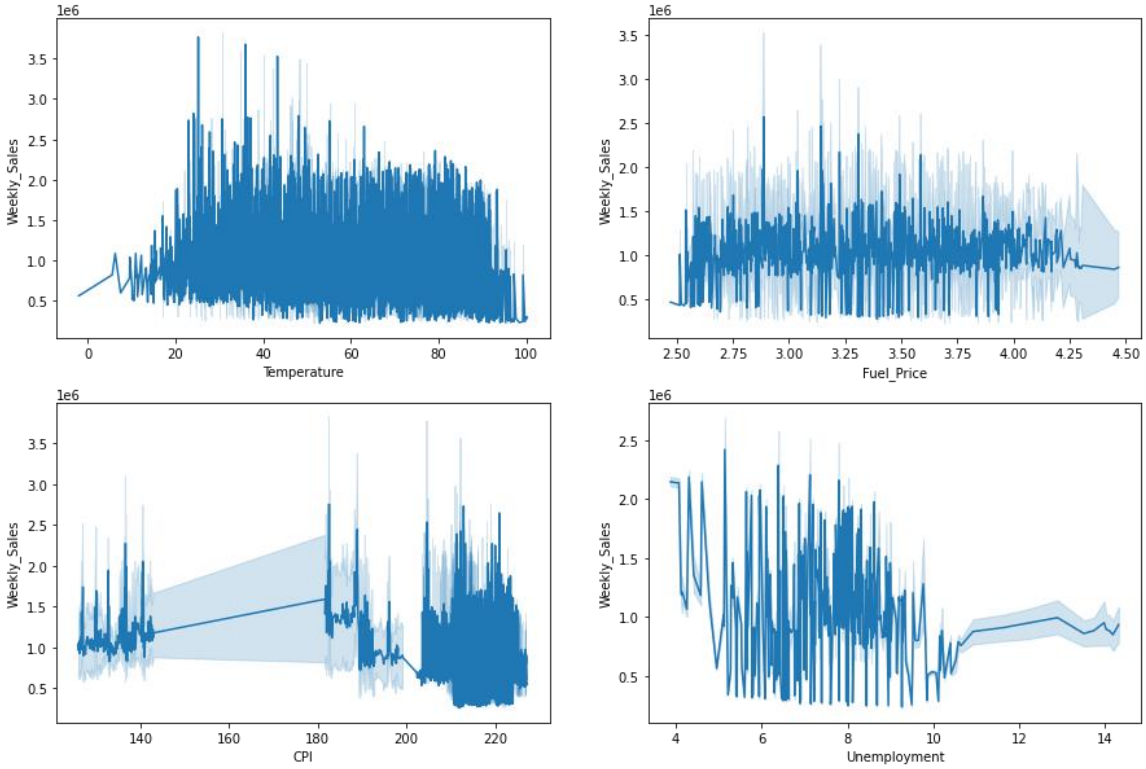
Correlation between sales and other factors



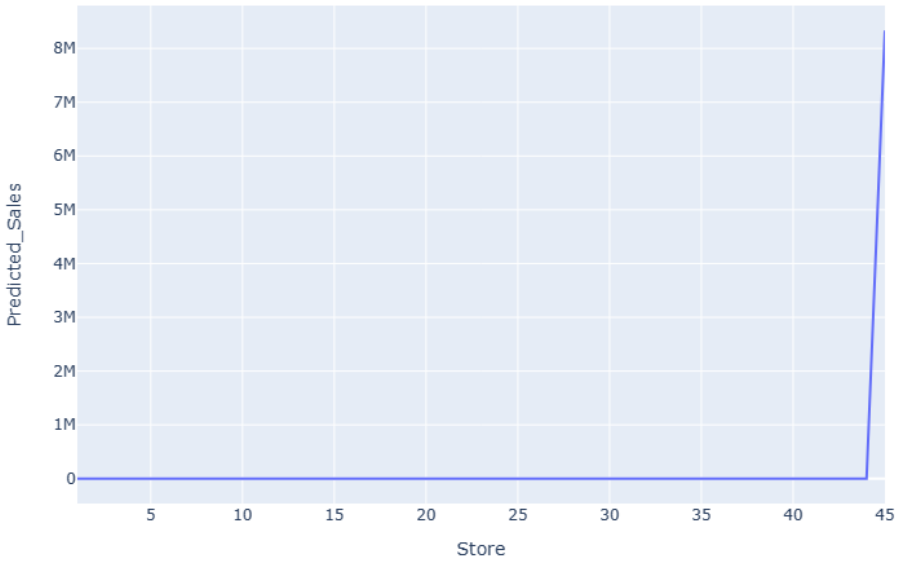
Average sales per year



Weekly sales as per region temperature, CPI, fuel price, unemployment



Visualizing the forecasted sales



In conclusion, I developed a data analysis and predictive modeling solution for a retail store with multiple outlets across the country to manage its inventory and match demand with supply.

I explored the provided dataset using exploratory data analysis (EDA) techniques to gain insights into the sales trends, factors affecting sales, and sales by store size and type. I also identified the sales during holidays and the correlation between different factors that affect sales.

Using the insights gained from EDA, I then developed a machine learning model to forecast sales for each store for the next 12 weeks. The model was trained on the historical sales data, and the predictions were evaluated using different performance metrics to assess the model's accuracy.

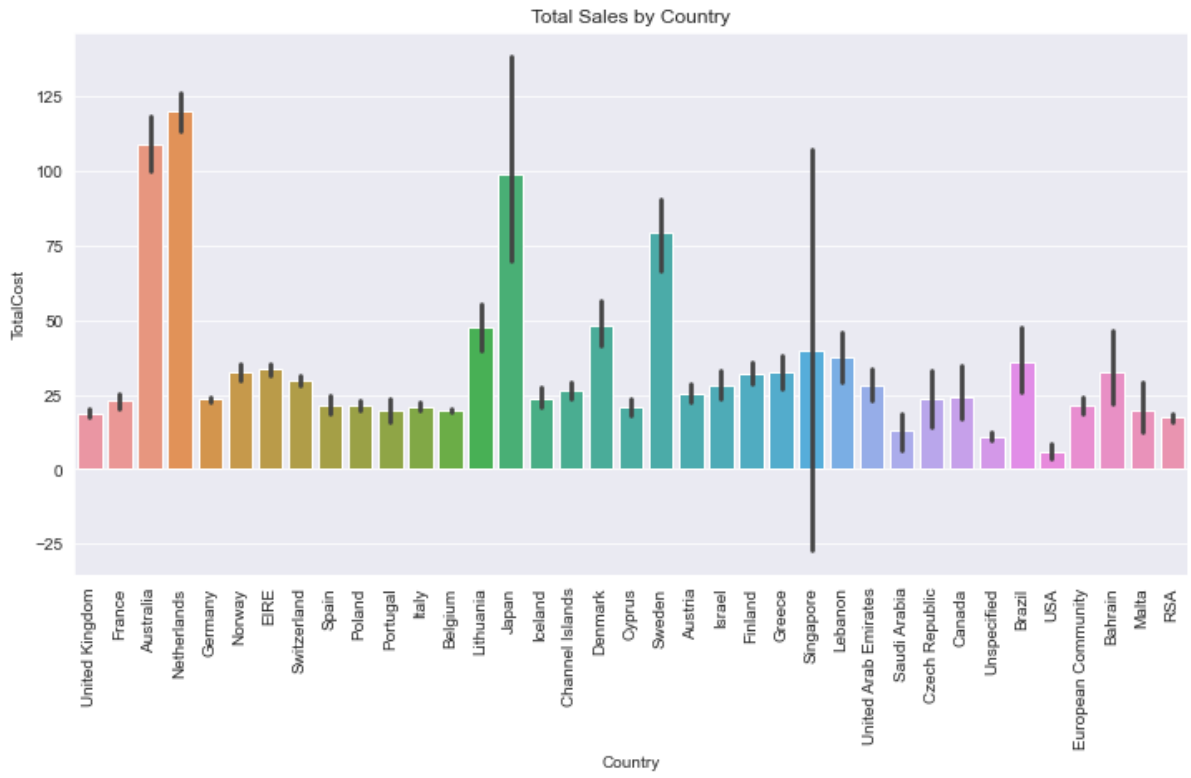
Overall, the data analysis and predictive modeling solution provided actionable insights for the retail store to improve inventory management and match demand with supply. The solution can be extended to include more features, such as social media sentiment analysis, competitor analysis, and market trends, to further improve the accuracy of the sales forecasting model.

Problem Statement 2:

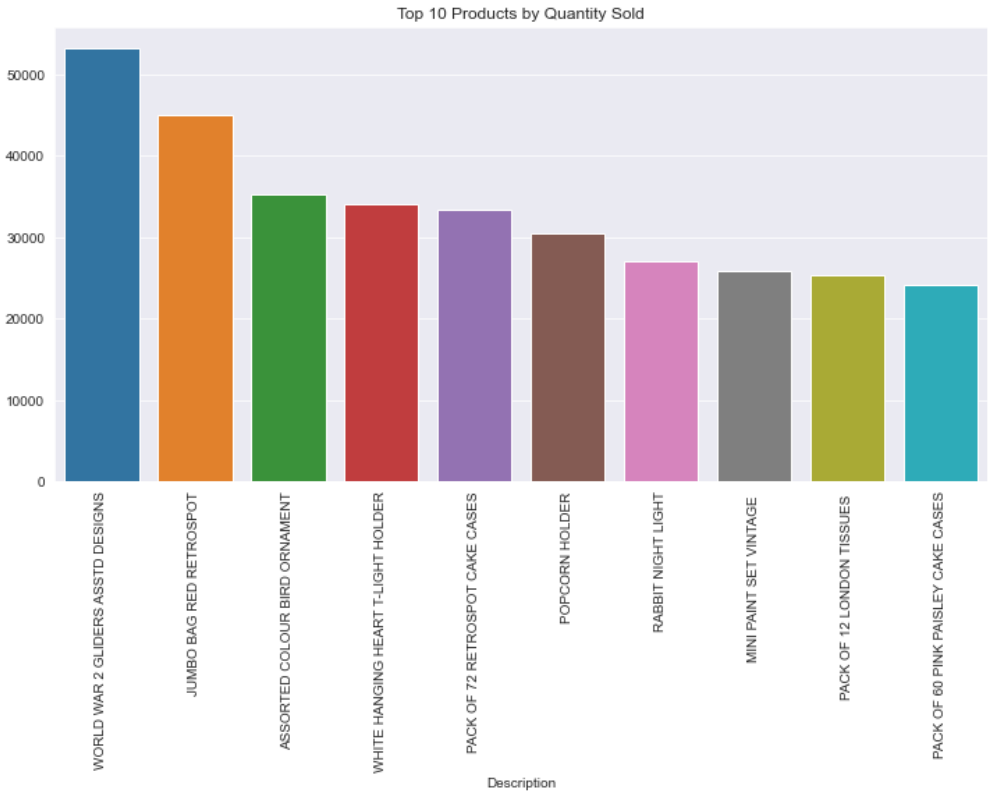
An online retail store is trying to understand the various customer purchase patterns for their firm, you are required to give enough evidence-based insights to provide the same.
1. Using the above data, find useful insights about the customer purchasing history
that can be an added advantage for the online retailer.
2. Segment the customers based on their purchasing behavior.

The total sales and number of orders
Total sales: 8300065.81
Number of orders: 20460

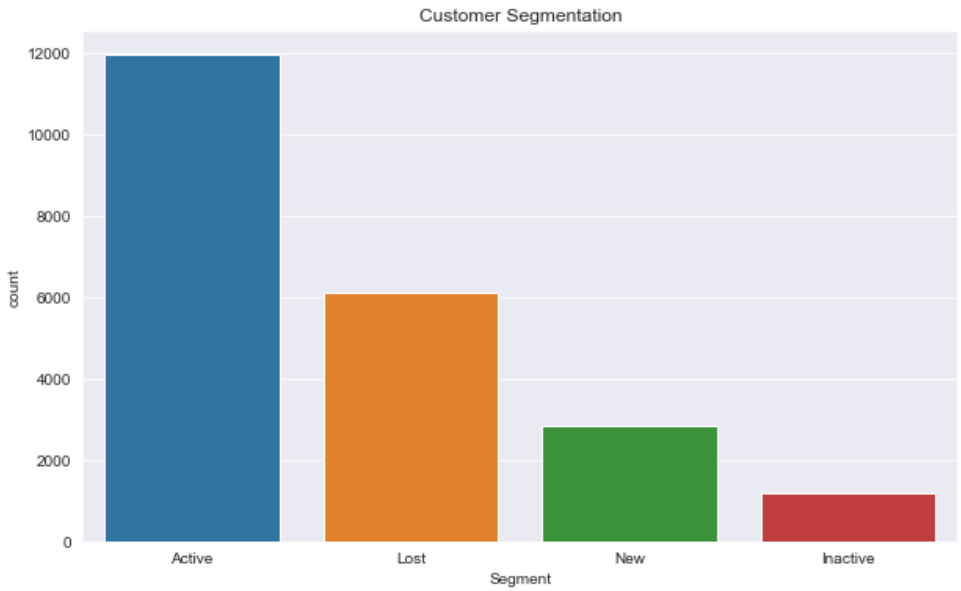
Visualize the sales distribution by country



Visualize the top 10 products by quantity sold



Segment the customers based on their purchasing behavior



In conclusion, I analyzed the sales data of an online retail store using Python and various data analysis and visualization libraries. I loaded the data from a CSV file and performed data cleaning to remove any missing values.

I then calculated the total sales and number of orders and visualized the sales distribution by country and the top 10 products by quantity sold. I used seaborn and matplotlib libraries to create bar plots for visualizing the data.

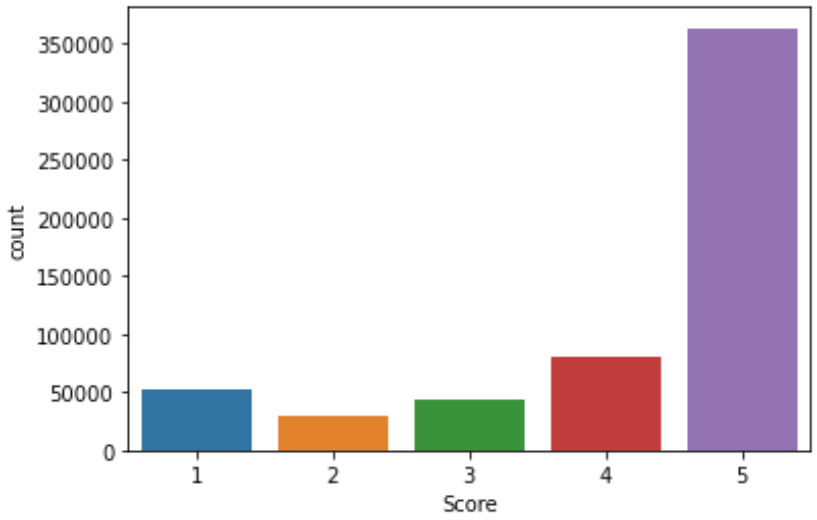
I also segmented the customers based on their purchasing behavior and created a function to segment the customers into four categories: Active, New, Inactive, and Lost. I applied this function to the data to get the customer segments and visualized the customer segments using a count plot.

The analysis provided insights into the sales distribution by country, top-selling products, and customer segmentation. The results can be used by the online retail store to improve its inventory management, marketing, and customer engagement strategies to increase sales and profitability.

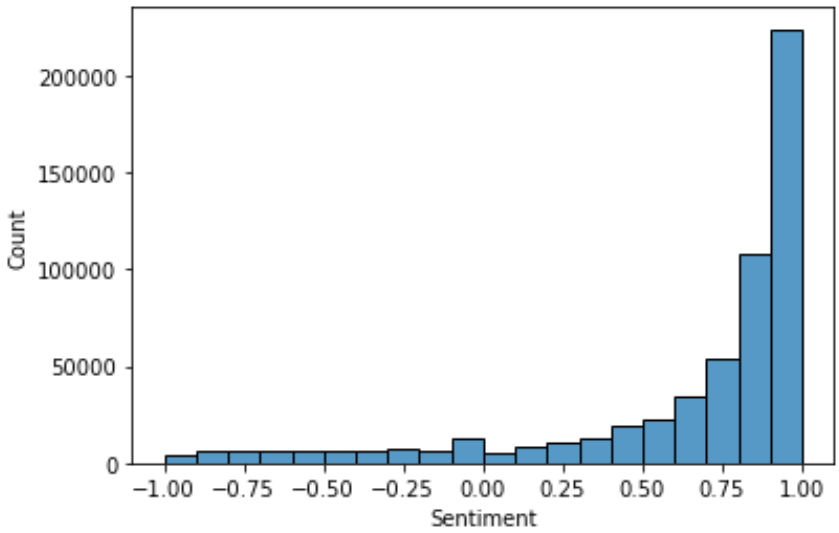
Problem Statement 3:

You are working in an e-commerce company, and your company has put forward a task to analyze the customer reviews for various products. You are supposed to create a report that classifies the products based on the customer reviews. 1. Find various trends and patterns in the reviews data, create useful insights that best describe the product quality.
2. Classify each review based on the sentiment associated with the same.

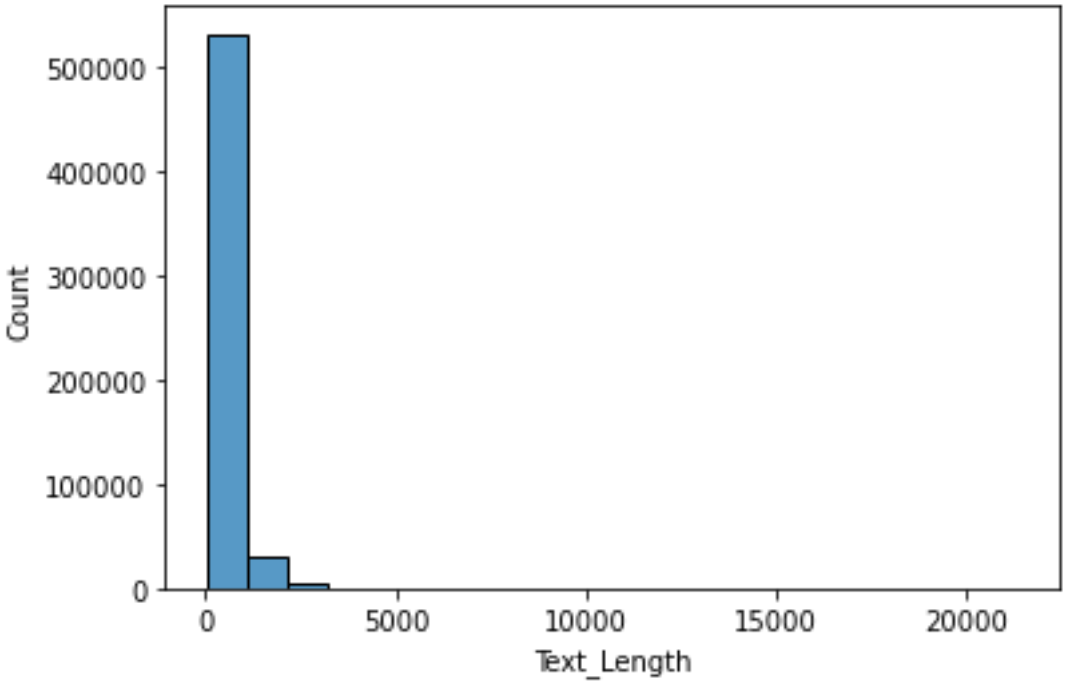
Check the distribution of scores



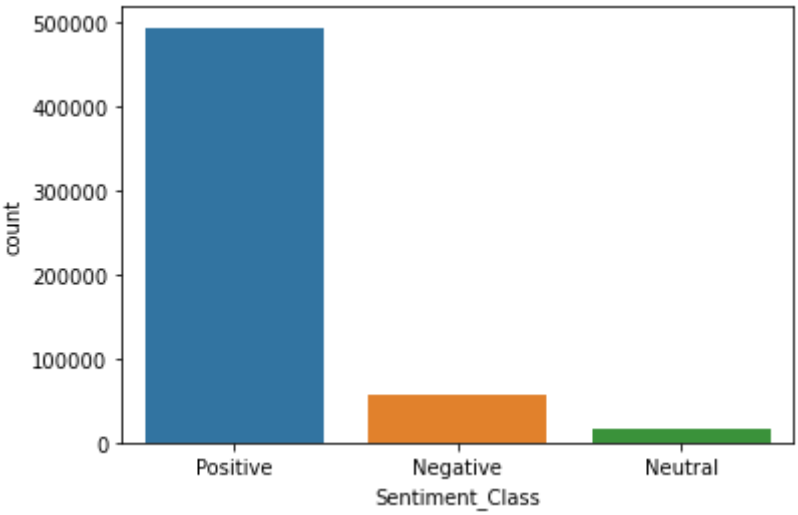
Check the distribution of sentiment scores



Check the distribution of review text lengths



Check the distribution of sentiment classes



In conclusion, I performed sentiment analysis on customer reviews of a product using Python and various natural language processing (NLP) libraries.
I loaded the dataset from a CSV file and performed data cleaning to remove any missing values. I explored the data using various visualization techniques, such as a count plot and histogram, to get an overview of the data distribution.

I then used the nltk library for text processing and cleaning, which included tokenization, lemmatization, and removal of stopwords. I also used the Sentiment Intensity Analyzer class from the nltk.sentiment module to assign a sentiment score to each review.

I then created a new column for sentiment analysis and classified the reviews into positive, negative, and neutral categories based on their sentiment scores. Finally, I visualized the distribution of sentiment classes using a count plot.

The sentiment analysis provided insights into the overall sentiment of the customers towards the product, which can be used by the product team to improve the product quality and customer satisfaction. The analysis can be extended to include more features, such as topic modeling, to further understand the customers' feedback and preferences.

References:

Frost, J. (2021). Regression coefficients- statistics by jim. https://statisticsbyjim.com/glossary/regression-coefficient/

Glen, S. (2016). Elementary statistics for the rest of us. https://www.statisticshowto.com/correlation-matrix/

Jeswani, R· (2021). Predicting Walmart Sales, Exploratory Data Analysis, and Walmart Sales Dashboard. https://www.rit.edu/ischoolprojects/sites/rit.edu.ischoolprojects/files/document_library/Rashmi_Jeswani_Capstone.pdf

Guide, U. B. A. R. P. (n.d.). Gradient boosting machines. http://uc-r.github.io/gbm_ regression

Harsoor, A. S., & Patil, A. (2015). Forecast of sales of walmart store using big data applications. *International Journal of Research in Engineering and Technology eIS*, *04*, 51–59. https : / / doi . org / https : / / ijret . org / volumes / 2015v04 / i06 / IJRET20150406008.pdf

Jaccard, J., & Turrisi, R. (2018). *Interaction effect in multiple regression second edition*. Sage Publications, Thousand Oaks CA.

Jain, A. (2016). Complete machine learning guide to parameter tuning in gradient boosting (gbm) in python. https://www.analyticsvidhya.com/blog/2016/02/complete-guide-parameter-tuning-gradient-boosting-gbm-python/