

FORECASTING OF CUSTOMER BEHAVIOUR BY PREDICTIVE ANALYSIS OF SALES DATA

Thesis Submitted in fulfilment of the Requirement for the degree

Of

Master of Technology (M.Tech)

in

Department of Computer Science and Engineering

By

DEBASISH DHAR

(Roll No. 502120011002 Reg. No. 201430411210006)

Under the guidance of

Dr. SANGEETA BHATTACHARYA

Computer Science & Engineering Department



Guru Nanak Institute of Technology, Kolkata-700110

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany the successful completion of task would be incomplete without the mention of the people who made it possible, whose constant guidance and encouragement always boosted the morale. We take a great pleasure in presenting a project, which is the result of a studied of both research and knowledge.

We first take the privilege to thank the Head of the Department, Dr. Sangeeta Bhattacharya, for permitting us in laying the first stone of success and providing the lab facilities, we would also like to thank the other staff in our department and lab assistant who directly or indirectly helped us in successful completion of the project.

We again feel extremely thankful to our project guide Dr. Sangeeta Bhattacharya who has shared her valuable knowledge with me and made me understand the real essence of the topic and created interest in me to work rigorously for the project and the support and encouragement provided for the topic of the thesis.

We also thank all the staff members of CSE department who extended part of support in the successful completion of the project.

Date: 26-05-2022
Debasish Dhar

GURUNANAK INSTITUTE OF TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the thesis entitled, “Forecasting of Customer Behaviour by Predictive Analysis of Sales Data” which is going to be submitted by Debasish Dhar, Roll no-502120011002, for the award of the degree of partial fulfilment of Master of Technology in the branch Computer Science and Engineering of Guru Nanak Institute of Technology, is a record of research work carried out under the supervision and guidance of the undersigned. Dr. Sangeeta Bhattacharya has worked for nearly one year on the subject at the Department of Computer Science and Engineering, Guru Nanak Institute of Technology, Kolkata and this has reached the standard fulfilment of the requirements and the regulation relating to the degree.

The contents of this thesis, in full or part, have not been submitted to any other university or institution for the award of any degree.

Supervisor

Dr. SANGEETA BHATTACHARYA

Department of Computer Science and Engineering
GNIT, Kolkata.

Head of the Department

Dr. SANGEETA BHATTACHARYA

Head of the Department
Department of Computer Science and Engineering,
GNIT, Kolkata

CONTENTS

	Page No.
ABSTRACT	01
CHAPTER 1: INTRODUCTION	02
CHAPTER 2: LITERATURE REVIEW	03
CHAPTER 3: PROBLEM STATEMENT & PROPOSED APPROACH	04
CHAPTER 4: THEORETICAL FRAMEWORK	09
CHAPTER 5: IMPLEMENTATION	10
CHAPTER 6: EXPERIMENTAL RESULTS	16
CHAPTER 7: CONCLUSION & FUTURE STUDY	26
REFERENCES	27

ABSTRACT

Since data can be in huge volumes, estimating the sales data and predicting the end customer behaviour can be quite challenging. Considering this problem, we have proposed the analysis and visualization of the sales data of an Electronics store and thereby predicting future buying behaviours of potential customers. We have used exploratory data analysis (EDA) for this project where data interpretations have been in row and column format. We have used Exploratory Data Analysis (EDA) to summarize the data by taking their main characteristics and visualize the data with proper representations. EDA focuses more narrowly on checking assumptions required for model fitting and hypothesis testing and handling missing values and making transformations of variables as needed. EDA quickly describes the data sets number of rows/columns, missing data, data types and preview. We have then cleaned the unstructured data, handled missing data, invalid data types and incorrect values present in the data set. We have used this approach of analysing data sets to summarize the main characteristics using various visual methods. In summary, this work has shown us the hidden relationships and attributes present in our data. We have used python programming for data analysis. The programming language Python, with its English commands and easy-to-follow syntax, offers an amazingly powerful open-source alternative to traditional techniques and applications. Data analysis and visualization programs used in this thesis work has allowed us to reach deeper understanding of the customer buying behaviour and to forecast using predictive analysis of sales data.

1. INTRODUCTION

Common data forecasting project workflows consist of five main stages as presented in the below figure. During our course of the thesis work, we have performed these procedures in many cycles until the goal is achieved. A significant number of business problems require decisions to be made based on a large amount of historical data and human decision-makers have difficulty integrating these data; thus, their decisions are susceptible to several biases. Machine learning algorithms and data mining have been employed in recent times to help organizations and businesses combat and tackle the problem of harvesting data and using the knowledge found within them to make predictions for sales. With a good data mining infrastructure in place and the right application of predictive algorithms, companies stand a better chance of making better informed decisions. The main reason a company does a forecast is to balance marketing resources and sales against supply capacity planning. With the right implementation of forecasting, companies and corporations no doubt can address fundamental questions such as "can we drive demand with our current rise of price, promotion, or marketing?", "are the resources currently at our disposal adequate to measure up to demand?", "Is there enough personnel to match the volume of budgeted sales?". In this regard, companies and organizations allocate and invest a reasonable number of resources in both human and finance to obtain adequate and genuine prediction results. Therefore, Sales forecasting is deemed of paramount importance for companies with intentions of entering new markets or adding new services, products, or experiencing high growth.

2. LITERATURE REVIEW

SL No	Reference Study	Methodology Used	Dataset Used	Drawback
1	Forecasting sales in the supply chain: Consumer analytics in the big data era - Tonya Boone, Ram Ganeshan, Aditya Jain, Nada R. Sanders (2018)	Time Series Analysis (TSA) were typically used for estimating patterns from the past sales data, which are then extrapolated for forecasting sales in the supply chain,	Granular data were gathered from the POS systems and used as the primary data source for the thesis.	The data gathered from POS systems for the purpose of the thesis were sparse and non-repetitive and required multiple methodologies for forecasting. This limits the usability of the slows the data & widespread acceptance of the results driven.
2	Predictive Analysis Sales for Corporate Services Telecommunications Company using Gradient Boost Algorithm - Oryza Wisesa, Andi Adriansyah & Osamah Ibrahim Khalaf (2021)	Sales Data of last three years has been analysed using Gradient boosting Algorithm. GBA combines the predictions from multiple decision trees to generate the final predictions.	For the three consecutive years of sales data, the data used for this analysis is based on the B2B revenue. To forecast B2B revenue, historical record revenue for three years was obtained. The data gathered covered category, region, item type and opportunity ID, quarter, product name, product sub-component, service product (MIDI) and sales revenue.	The Gradient Boosting Algorithm used for the purpose of this thesis can be computationally expensive and thus takes long time to train, especially on CPUs. Thereby, making it hard to interpret the final models in real life.
3	Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities - Mahya Seyedan and Fereshteh Mafakheri (2020)	Big Data Analytics using clustering analysis has been used to forecast future demands. The big data set has been grouped into data objects & subgroups based on their similarities.	For the purpose of the thesis, various studies have been collected and analysed with respect to methods and techniques used in demand prediction.	Since a multitude of sensors are used to gather information about customer browsing and purchase behaviours, it is imperative that firms address and articulate a clear privacy policy. This adds to the challenge that firms potentially may have to deal with privacy laws in different geographical locations. Finally, firms that make use of big data also need a proper clear big data strategy in order to keep their decisions free of bias.
4	A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector - Ullah, Raza, Malik, Imran, Islam, Kim (2019)	Using Random Forest Algorithm, many decision trees on various subsets of the given dataset were created and the average was taken in order to improve the predictive accuracy of the dataset.	User-generated data from Internet searches and social media were collected & then analysed using Random Forest Algorithm.	The size and unstructured content of the enormous data sets used can be daunting. Secondly, in real word, any failure to keep customer data, especially sensitive data (such as social security and credit card numbers) secure can be illegal and open to relevant government oversight agencies. Application of the results of the thesis can be limited since random forest is highly complex when compared to decision trees. Also, training time is more compared to other models due to its complexity. Whenever it must make a prediction each decision tree has to generate output for the given input data.
5	Churn Prediction and Retention in Banking, Telecom and IT Sectors using Machine Learning Techniques - Himani Jain , Garima Yadav , and R. Manooov(Springer Nature Singapore Pte Ltd. 2021)	Exploratory Data Analysis was performed on the Datasets & then Logistic Regression, Random Forest, SVM, XG Boost Model was applied on the dataset. The results were then compared using confusion matrix. The best algorithm was then identified. Following are the steps used for the purpose of the thesis: 1. Data Collection: - Collection of datasets for banking, telecom, and IT sectors. - Data cleansing and pre-processing 2. Exploratory Data Analysis: - Find the proportion of customer churned and retained. - Find correlation of other attributes with the target - Box plot to visualize outliers - Feature Engineering - Data Preparation for model fitting 3. Model Fitting and Selection Apply following models on each dataset: - Logistic Regression - Random Forest - SVM - XG Boost 4. Compare all the algorithms using confusion	End user data & customer behaviour data was collected for three different domains banking, telecom, and IT.	Although there are no domain constraints in this approach, no implementation has been done so it will not be practical to say for sure whether the results driven are actually feasible or not. The Results are not satisfying enough because the accuracy rate, measured by F-score is low and not much impressive.

		matrix 5. Analyse ROC curve and bar graphs to compare algorithms. 6. Identify the best algorithm for each domain. 7. Formulate retention strategies for each domain.		
6	Prediction of University Examination Results with Machine Learning Approach - Prediction of University Examination Results with Machine Learning Approach(Springer Nature Singapore Pte Ltd. 2021)	ML approach was used to find out the average marks based on the difficulty level of questions included in the question paper. The prediction model to predict average marks was designed using python programming language and several of its machine learning packages like pandas & NumPy, etc.	The dataset used was of the form of a csv file that is read into a pandas data frame easily using the readfromCSV command.	The drawback of this model lies in the fact that there are still several unanswered issues regarding data being spread across multiple organizations and multiple domains.
7	Prediction of Stock Market Prices of Using Recurrent Neural Network—Long Short-Term Memory by Haritha Harikrishnan and Siddhaling Urolagin(Springer Nature Singapore Pte Ltd. 2021)	The stock prediction system used in this thesis was based on the LSTM Model (Long short-term memory) & involved the following steps: 1. Obtaining dataset and pre-processing 2. Construction of the model 3. Prediction, error calculation and accuracy evaluation.	The stock market dataset used for the purpose of the thesis was obtained from Quandl, which is a premier source for financial, economic, and alternative datasets serving professionals interested in investments. The obtained dataset from Quandl contained twelve features, out of which, date of the observation, the opening price, the highest intra-day price reached, the lowest intra-day price reached, and the closing price of the stock was taken into consideration.	Although the approach used could predict the end of the day stock price accurately, the LSTM Model takes longer to train & require more memory to train. Also, the approach used is sensitive to different random weight initializations.

3. PROBLEM STATEMENT & PROPOSED APPROACH

3.1 Objective:

The main objective of the research is to use exploratory data analysis (EDA) to analyse the sales data of an electronics store and to forecast the end customers buying behaviour and thereby enabling the decision makers to take data driven decisions for the organization.

3.2 Problem Statement:

An Electronics store has last 12 months of sales data which contains numerous electronics store purchases broken down by month, product type, cost, customer name, purchase address, etc. The corporation needs data scientists to analyse the sales data and forecast the key buying decisions of potential customers.

3.3 Timeline:

The work process and the timeline of this thesis can be divided into the following different phases:

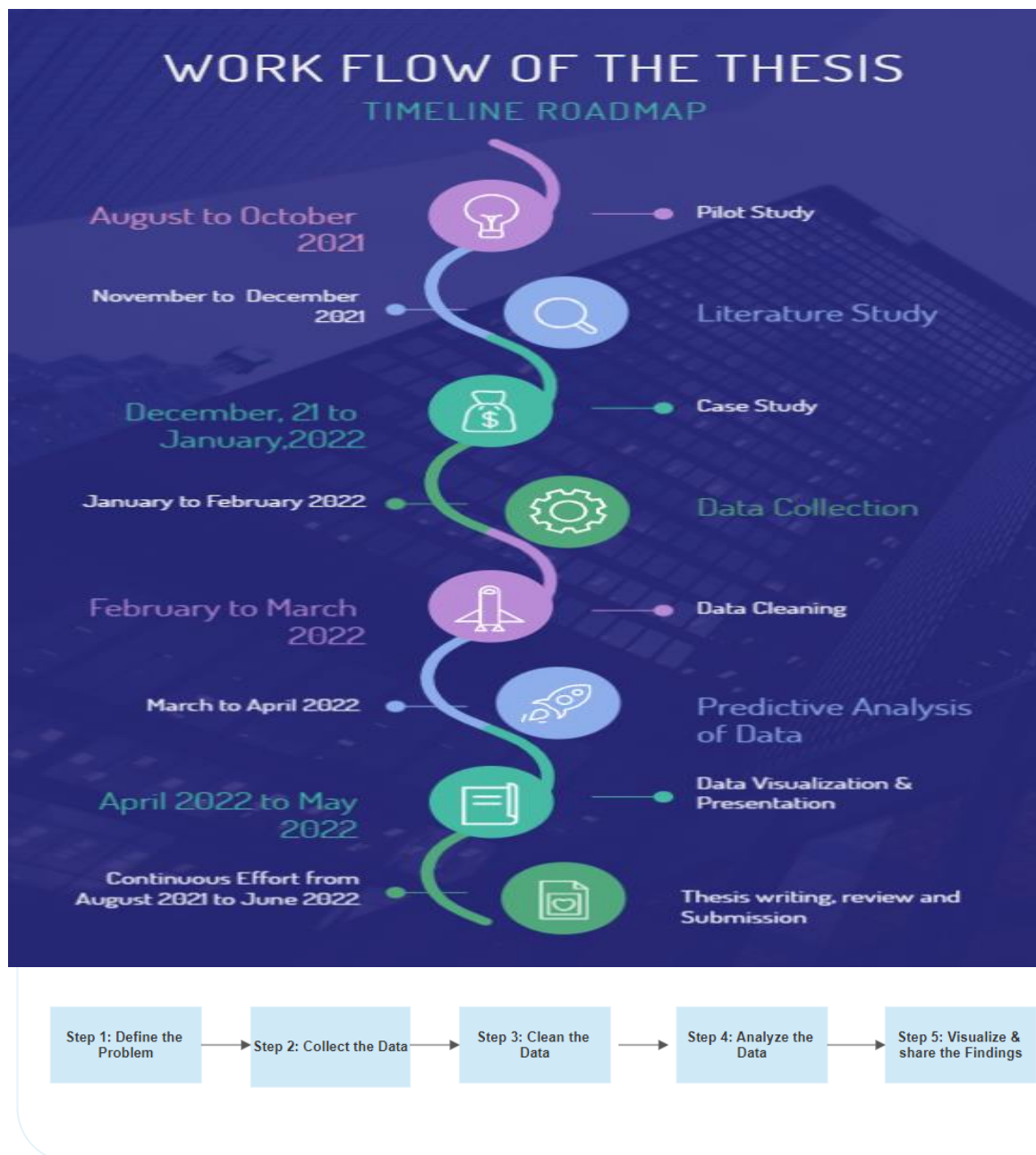


Fig 1: The Data Forecasting Process Workflow Diagram

Step 1: Defining the problem

The first step in any data analysis process is to define your objective. In data analytics jargon, this is sometimes called the ‘problem statement’.

Defining your objective means coming up with a hypothesis and figuring how to test it. Start by asking: What business problem am I trying to solve? While this might sound straightforward, it can be trickier than it seems. For instance, your organization’s senior management might pose an issue, such as: “Why are we

losing customers?” It’s possible, though, that this doesn’t get to the core of the problem. A data analyst’s job is to understand the business and its goals in enough depth that they can frame the problem the right way.

SL No	Business questions to be answered	How the answer to this question can help grow future profits?
1	What was the best month for sales? How much was earned in that month?	By knowing the best month for sales, the business leaders can be making data driven decisions like increasing prices during peak season, providing bundle discounts, coupons.
2	Which city sold the most products?	By knowing the city which sold the most products, business leaders can decide on which city needs more advertisements to drive sales.
3	What time should we display advertisements to maximize the likelihood of customer’s buying product?	By knowing the best time to advertise, business leaders can better manage the advertisement budget and publish add in the best possible time.
4	Which products are most often sold together?	By knowing which products customers buys together, business leaders can design an effective pricing strategy for the bundled product or provide bundled discounts as needed.
5	What product sold the most?	By knowing the most sold products, business leaders can decide which products have pricing power and decide on an effective pricing strategy based on the product.
6	Why do you think that product sold the most?	Business can find out whether there is a correlation between price and sales.

Step 2: Collecting the data

A key part of this is determining which data we need for further analysis. This might be quantitative (numeric) data, e.g., sales figures, or qualitative (descriptive) data, such as customer reviews. All data fit into one of three categories: first-party, second-party, and third-party data.

First-party data are data that you, or your company, have directly collected from customers. It might come in the form of transactional tracking data or information from our company’s customer relationship management (CRM) system. Whatever its source, first-party data is usually structured and organized in a clear, defined way. Other sources of first-party data might include customer satisfaction surveys, focus groups, interviews, or direct observation.

Second-party data is the first-party data of other organizations. This might be available directly from the company or through a private marketplace. The main benefit of second-party data is that they are usually structured, and although they will be less relevant than first-party data, they also tend to be quite reliable. Examples of second-party data include website, app or social media activity, like online purchase histories, or shipping data.

Third-party data is data that has been collected and aggregated from numerous sources by a third-party organization. Often (though not always) third-party data contains a vast amount of unstructured data points (big data). Many organizations collect big data to create industry reports or to conduct market research. The research and advisory firm Gartner is a good real-world example of an organization that collects big data and sells it on to other companies. Open data repositories and government portals are also sources of third-party data.

For the purpose of the thesis, we have collected Second-party data source (publicly available) for analysis and findings.

Step 3: Cleaning the data

Once we have collected our data, the next step is to get it ready for analysis. This means cleaning, or 'scrubbing' it, and is crucial in making sure that we are working with high-quality data. Key data cleaning tasks include removing major errors, duplicates, and outliers—all of which are inevitable problems when aggregating data from numerous sources.



Step 4: Analysing the data

The type of data analysis we carry out largely depends on what our goal is. But there are many techniques available. Univariate or bivariate analysis, time-series analysis, and regression analysis are just a few you might have heard of. More important than the different types, though, is how you apply them. This depends on what insights we are hoping to gain. Broadly speaking, all types of data analysis fit into one of the following four categories.

Predictive analysis

Predictive analysis allows business to identify future trends based on historical data. In business, predictive analysis is commonly used to forecast future growth. Predictive analysis has grown increasingly sophisticated in recent years. The speedy evolution of machine learning allows organizations to make surprisingly accurate forecasts. Take the insurance industry. Insurance providers commonly use past data to predict which customer groups are more likely to get into accidents. As a result, they will hike up customer insurance premiums for those groups. Likewise, the retail industry often uses transaction data to predict where future trends lie, or to determine seasonal buying habits to inform their strategies.

Step 5: Visualize & share the findings

The final step of the data analytics process is to share these insights with the wider world or with the organization's stakeholders. This is more complex than simply sharing the raw results of your work—it involves interpreting the outcomes and presenting them in a manner that is digestible for all types of audiences. Since we will often present information to decision-makers, it is very important that the insights you present are 100% clear and unambiguous. For this reason, data analysts commonly use reports, dashboards, and interactive visualizations to support their findings.

4. THEORETICAL FRAMEWORK

This section contains the theoretical components that form the basis of the thesis and was applied to the empirical data.

Data Analytics

Data analytics is a method for analysing data sets to find patterns and develop conclusions about the information they contain. Data analytics is increasingly being used with the aid of specialised tools and software. In commercial industries, various data analytics technologies, tools, and approaches are frequently used to allow firms to make better informed business choices. Scientists and researchers use it to confirm or refute scientific models, ideas, and hypotheses.

In this thesis, we employed data analytics in our application to gain an understanding of sales data for items sold over time. The application also helps the owner to get the visualization of product sales in form of charts and graphs, and sales comparison of each salesperson. The end user can even filter the visualization based on date and time of sales, for example, he can see the graph for sales of products in a particular month of a particular year.

The end user would be able to observe many characteristics of the transaction, such as the total number of sales, the highest profit per sale, the minimum profit per sale, the mean profit from all sales, the median profit from all sales, and so on. Businesses may enhance profits, improve productivity, and optimise marketing campaigns and customer service efforts by using data analytics programmes. It also aids organisations in recognising upcoming industry trends and gaining a competitive advantage over competitors.

Predictive Data Analysis:

Exploratory data analysis is primarily a method of determining what the data can tell us outside of formal modelling or hypothesis testing. EDA aids in the analysis of data sets in order to describe their statistical features, concentrating on four main aspects: measures of central tendency (mean, mode, and median), measures of dispersion (standard deviation and variance), distribution shape, and the presence of outliers.

Predictive Data Analysis in python:

In this thesis, for exploratory data analysis, we have used Python. Python has a large number of libraries. The amount of data that can be handled is substantially greater. It is a free and open source programming language. It has a wide range of libraries, some of which have excellent visualisation tools. The visualisation method can aid in the creation of a clear report.

Pandas

It is the most powerful data analysis programme available. The data may be cleaned, transformed, and analysed. In a computer, data may be saved in CSV format. It is possible to clean, visualise, and save data. It is based on the NumPy Python module. Matplotlib plotting utilities with the Scikit-learn machine learning method.

Jupyter Notebook

It enables the execution of code in a specific cell. It provides a computation method based on a console. It facilitates the application procedure using the internet. It contains the computation's input and output. It displays the object in a rich media format.

5. IMPLEMENTATION

Methodology:

Over the years, numerous data mining and machine learning technologies have been applied to aid researchers and analysts in solving various prediction challenges. One of the tools that should be used is product sales forecast, which is easily available to assist a company or group in making better informed growth decisions.

This presentation will focus on using the CRPIS-DM as a data mining guide to generate a good prediction that aligns with corporate goals and objectives. This method provides a foundation for achieving more significant and faster results. The CRISP-DM method divides research into six steps, making the process easier to understand and giving a road map for planning and conducting the study.

CRPIS-DM Methodology:

Data mining is a cutting-edge technique that necessitates a wide variety of abilities and knowledge. There is presently no standard framework for conducting data mining initiatives, meaning that the success or failure of a data mining project is highly dependent on the individual or team performing it. Data mining necessitates a standard methodology that can assist in the conversion of business or organisational problems into data mining tasks, the recommendation of appropriate data transformations and data mining techniques, and the provision of tools for evaluating the feasibility of findings and documenting the knowledge.

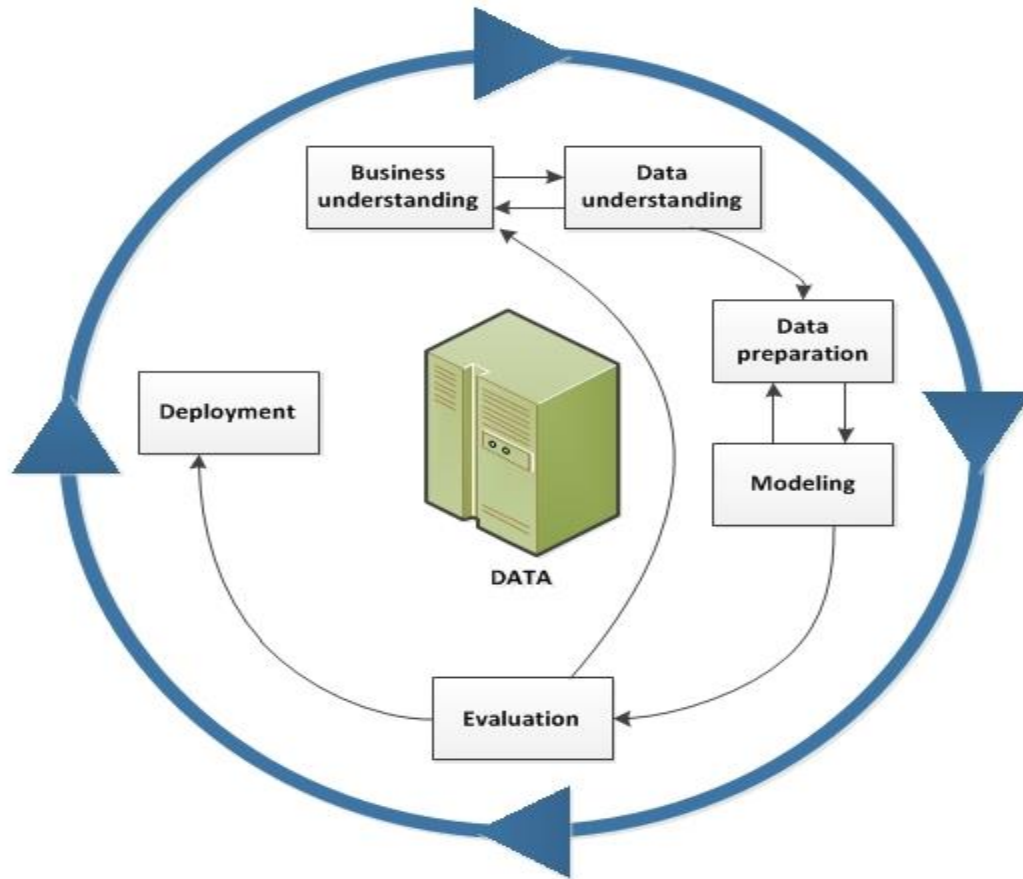
Some of these problems were addressed by the Cross Industry Standard Process for Data Mining (CRISP-DM), which defined a process model that provided a framework for the execution of data mining projects that was independent of both the industry and the technology utilised. CRISPDM approach has found to be beneficial by data mining analysts in a variety of ways. The approach gives assistance to beginners, assists in project planning, and provides guidance on each job or step of the process.

Furthermore, seasoned analysts employ checklists for each job to ensure that nothing is forgotten. The most important purpose of the CRISP-DM technique, however, is to communicate and log results. It aids in the integration of multiple tools and individuals with varying levels of competence and experience into a cohesive and productive project.

The Generic CRISP-DM Reference Model:

The CRISP-DM data mining reference model offers an outline of a data mining project's life cycle. It includes the stages, their respective tasks and results of a project. A data mining project's life cycle is split into six phases.

During the project and the approach deployed, the lessons learned will cause new, often more oriented business concerns, and subsequent data mining processes can benefit from previous experience. The CRISP-DM model is shown in the figure below.



The various stages of the CRISP-DM model have been employed in the cause of this project, and the stages will be explained with respect to sales forecasting.

Business Understanding:

This first stage stresses a company-led knowledge of project aims and ambitions, which is subsequently translated into a problem description for data mining and a draught project plan. In this regard, the business challenge for this project is to do a sales prediction for an off-license retail outlet in order to determine which of the goods sells the most by looking at previous sales data. This is required for company managers to make better informed judgments about overstocking or understocking in order to maximise profit and customer happiness. The goal of this study is to compare the best machine learning algorithms in order to identify the most accurate prediction model for the shop's two main product classes and the factors that impact them.

Data Understanding:

Data comprehension begins with data collection and progresses to being more familiar with the data in order to identify data consistency concerns, get first insight into the data, or spot intriguing subsets in order to draw inferences about hidden information. This initiative took use of information gathered from an off-license retail store's sale. The dataset contains statistics on over 50 different items sold in the business during the month of January, which are divided into various categories. There are 20892 rows and 16 columns in the data set. The dataset's columns are listed below.

- Year: showing the year of sales in review.
- Month: showing the month in review.
- Supplier: showing the supplier of the product to the store.
- Item code: showing the code for each item.
- Item Description: Describing which type of item is.
- Item Type: showing which class the item belongs
- Damaged units: showing how many of the units got damaged.
- Retail transfers: showing how many items transferred from the store.
- Item weight: showing the weight of each item.
- Item visibility: position on store shelf.
- Quantity: showing the quantity sold.
- Discount: showing a discount percentage for each product.
- Returned: showing amount returned.
- Unit price: showing the price for each item.
- Payment method: indicating if the payment is made in cash or via online platforms.
- Total sales: total amount sold for the product.

	A	B	C	D	E	F
1	Year	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
2	176558	USB-C Charging Cable	2	11.95	04/19/19 08:46	917 1st St, Dallas, TX 75001
3						
4	176559	Bose SoundSport Headphones	1	99.99	04-07-2019 22:30	682 Chestnut St, Boston, MA 02215
5	176560	Google Phone	1	600	04-12-2019 14:38	669 Spruce St, Los Angeles, CA 90001
6	176560	Wired Headphones	1	11.99	04-12-2019 14:38	669 Spruce St, Los Angeles, CA 90001
7	176561	Wired Headphones	1	11.99	04/30/19 09:27	333 8th St, Los Angeles, CA 90001
8	176562	USB-C Charging Cable	1	11.95	04/29/19 13:03	381 Wilson St, San Francisco, CA 94016
9	176563	Bose SoundSport Headphones	1	99.99	04-02-2019 07:46	668 Center St, Seattle, WA 98101
10	176564	USB-C Charging Cable	1	11.95	04-12-2019 10:58	790 Ridge St, Atlanta, GA 30301
11	176565	Macbook Pro Laptop	1	1700	04/24/19 10:38	915 Willow St, San Francisco, CA 94016
12	176566	Wired Headphones	1	11.99	04-08-2019 14:05	83 7th St, Boston, MA 02215
13	176567	Google Phone	1	600	04/18/19 17:18	444 7th St, Los Angeles, CA 90001
14	176568	Lightning Charging Cable	1	14.95	04/15/19 12:18	438 Elm St, Seattle, WA 98101
15	176569	27in 4K Gaming Monitor	1	389.99	04/16/19 19:23	657 Hill St, Dallas, TX 75001
16	176570	AA Batteries (4-pack)	1	3.84	04/22/19 15:09	186 12th St, Dallas, TX 75001
17	176571	Lightning Charging Cable	1	14.95	04/19/19 14:29	253 Johnson St, Atlanta, GA 30301
18	176572	Apple AirPods Headphones	1	150	04-04-2019 20:30	149 Dogwood St, New York City, NY 10001
19	176573	USB-C Charging Cable	1	11.95	04/27/19 18:41	214 Chestnut St, San Francisco, CA 94016
20	176574	Google Phone	1	600	04-03-2019 19:42	20 Hill St, Los Angeles, CA 90001
21	176574	USB-C Charging Cable	1	11.95	04-03-2019 19:42	20 Hill St, Los Angeles, CA 90001
22	176575	AAA Batteries (4-pack)	1	2.99	04/27/19 00:30	433 Hill St, New York City, NY 10001
23	176576	Apple AirPods Headphones	1	150	04/28/19 11:42	771 Ridge St, Los Angeles, CA 90001
24	176577	Apple AirPods Headphones	1	150	04-04-2019 19:25	260 Spruce St, Dallas, TX 75001
25	176578	Apple AirPods Headphones	1	150	04-09-2019 23:35	513 Church St, Boston, MA 02215
26	176579	AA Batteries (4-pack)	1	3.84	04-11-2019 10:23	886 Jefferson St, New York City, NY 10001
27	176580	USB-C Charging Cable	1	11.95	04-05-2019 00:35	886 Willow St, Los Angeles, CA 90001
28	176581	iPhone	1	700	04-09-2019 21:38	84 Jackson St, Boston, MA 02215
29	176582	Bose SoundSport Headphones	1	99.99	04/27/19 12:20	178 Lincoln St, Atlanta, GA 30301
30	176583	AAA Batteries (4-pack)	2	2.99	04/20/19 12:00	146 Jackson St, Portland, OR 97035
31	176584	Flatscreen TV	1	300	04/24/19 20:39	936 Church St, San Francisco, CA 94016
32	176585	Bose SoundSport Headphones	1	99.99	04-07-2019 11:31	823 Highland St, Boston, MA 02215
33	176585	Bose SoundSport Headphones	1	99.99	04-07-2019 11:31	823 Highland St, Boston, MA 02215
34	176586	AAA Batteries (4-pack)	2	2.99	04-10-2019 17:00	365 Center St, San Francisco, CA 94016
35	176586	Google Phone	1	600	04-10-2019 17:00	365 Center St, San Francisco, CA 94016
36	176587	27in FHD Monitor	1	149.99	04/29/19 19:38	557 5th St, Los Angeles, CA 90001

Figure: Screenshot of the Original dataset

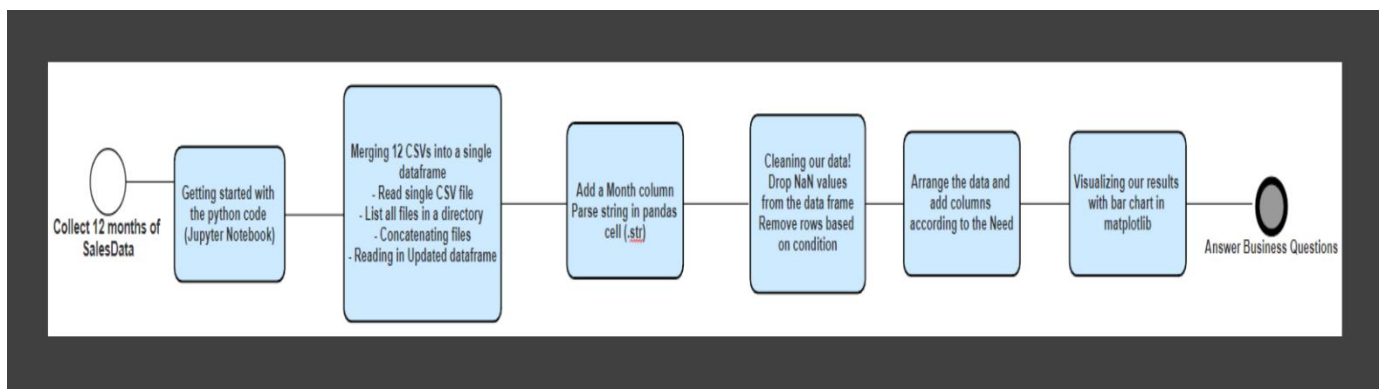
Data Preparation

All operations that result in the final data (data supplied from the original raw data in the modelling tool(s)) are included in the data preparation process. Procedures for data processing can be conducted several times and in any order. Table selection, document and attribute generation, data cleansing, new attribute creation, and data conversion for modelling tools are among these responsibilities. It includes properly analysing the

data acquired in order to construct the forecast's final dataset. The year, month, supplier, and item code columns were removed after initial data cleaning and preparation, and the item type was transformed to binomial. The dataset is shown in the diagram below before and after first cleaning and feature selection.

Modelling

This step involves selecting and implementing various modelling strategies, as well as fine-tuning their parameters to get best results. Typically, there are several approaches for tackling the same data mining problem, and each methodology has its own format. Data planning and modelling are inextricably linked, and many times, when modelling, one discovers data issues or makes proposals for additional data. This research examines five distinct types of prediction methods.















6. EXPERIMENTAL RESULTS

The goal of the thesis was to figure out how to use predictive analysis for sales forecasting. The collecting of data and the technologies that are used to do so to generate a forecast, the first stage in the forecasting process is to acquire information about the sales projection. It is important that companies take time to analyse what kind of information is needed to be collected to create an accurate sales forecast. With a shared database all departments within the company have access to all information within the company. However, the information that is put into the database must be sorted and simplified so others can understand and use the information.

Data Collection:

Step1: Researchers collect the monthly sales data of the electronics store for the last year from January to December. The datasets were publicly available for the various analysts to collect and perform various data cleaning.

Step 2: The collected Data sets were individual CSV files and had sales figures of each month starting from January to December. These publicly available data sets were second party sales data and was used to predict customer behaviour and sales forecasting.

Name	Date modified	Type	Size
 1_Sales_January_2021.csv	18-12-2021 01:38	Microsoft Excel C...	824 KB
 2_Sales_February_2021.csv	18-12-2021 01:38	Microsoft Excel C...	1,022 KB
 3_Sales_March_2021.csv	18-12-2021 01:38	Microsoft Excel C...	1,293 KB
 4_Sales_April_2021.csv	18-12-2021 01:38	Microsoft Excel C...	1,559 KB
 5_Sales_May_2021.csv	18-12-2021 01:38	Microsoft Excel C...	1,411 KB
 6_Sales_June_2021.csv	18-12-2021 01:38	Microsoft Excel C...	1,155 KB
 7_Sales_July_2021.csv	18-12-2021 01:38	Microsoft Excel C...	1,220 KB
 8_Sales_August_2021.csv	18-12-2021 01:38	Microsoft Excel C...	1,020 KB
 9_Sales_September_2021.csv	18-12-2021 01:38	Microsoft Excel C...	992 KB
 10_Sales_October_2021.csv	18-12-2021 01:38	Microsoft Excel C...	1,729 KB
 11_Sales_November_2021.csv	18-12-2021 01:38	Microsoft Excel C...	1,499 KB
 12_Sales_December_2021.csv	18-12-2021 01:38	Microsoft Excel C...	2,131 KB

Step 3: The researchers first action item was to Merge last 12 Months sales data to a single file for analysis. Merging into a single file will enable the researcher to perform the data cleaning actions better on the merged file. The researcher used Jupyter notebook and python language in conjunction to merge all files.

```
In [20]: import pandas as pd
import os
```

Task 1: Merge 12 Months sales data to a single file for analysis

```
In [2]: df = pd.read_csv("C:/Users/debdhar/Downloads/Merging/Sales_Data/Sales_April_2019.csv")

files = [file for file in os.listdir('C:/Users/debdhar/Downloads/Merging/Sales_Data')]
all_months_data = pd.DataFrame()  ## all_months_data is an empty data frame

for file in files:
    df = pd.read_csv("C:/Users/debdhar/Downloads/Merging/Sales_Data/" + file)
    all_months_data = pd.concat([all_months_data, df])

all_months_data.head()
all_months_data.to_csv("C:/Users/debdhar/Downloads/all_data.csv", index=False)
all_data = pd.read_csv("C:/Users/debdhar/Downloads/all_data.csv")
```

```
In [22]: all_data = pd.read_csv("C:/Users/debdhar/Downloads/all_data.csv")
all_data.head()
```

Data Cleaning:

Step 1: After the files were merged to a single file, the researchers next job was to view the data of any blank row or NAN values in any columns.

```
In [4]: ### To View NAN Rows

nan_df = all_data[all_data.isna().any(axis=1)]
nan_df.head()
```

```
Out[4]:
```

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
1	NaN	NaN	NaN	NaN	NaN	NaN
356	NaN	NaN	NaN	NaN	NaN	NaN
735	NaN	NaN	NaN	NaN	NaN	NaN
1433	NaN	NaN	NaN	NaN	NaN	NaN
1553	NaN	NaN	NaN	NaN	NaN	NaN

Step 2: After the NAN/blank values were present, the researchers next job was to clean the merged file of any blank row or NAN values in any columns.

```
In [5]: ### Delete the NAN rows

all_data = all_data.dropna(how='all')
all_data.head()
```

```
Out[5]:
```

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
0	176558	USB-C Charging Cable	2	11.95	04/19/19 08:46	917 1st St, Dallas, TX 75001
2	176559	Bose SoundSport Headphones	1	99.99	04/07/19 22:30	682 Chestnut St, Boston, MA 02215
3	176560	Google Phone	1	600	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001
4	176560	Wired Headphones	1	11.99	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001
5	176561	Wired Headphones	1	11.99	04/30/19 09:27	333 8th St, Los Angeles, CA 90001

Step 3: Next, step-in data cleaning process was to correct the format of the data present in the order date column.

In [6]: `### Find 'Or' & delete it`

```
all_data = all_data[all_data['Order Date'].str[0:2] != 'Or']
all_data.head()
```

Out[6]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
0	176558	USB-C Charging Cable	2	11.95	04/19/19 08:46	917 1st St, Dallas, TX 75001
2	176559	Bose SoundSport Headphones	1	99.99	04/07/19 22:30	682 Chestnut St, Boston, MA 02215
3	176560	Google Phone	1	600	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001
4	176560	Wired Headphones	1	11.99	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001
5	176561	Wired Headphones	1	11.99	04/30/19 09:27	333 8th St, Los Angeles, CA 90001

Step 4: Next, the researcher made sure all the columns had the same data type.

In [7]: `#### Convert Columns to correct data type`

```
all_data['Quantity Ordered'] = pd.to_numeric(all_data['Quantity Ordered'])
all_data['Price Each'] = pd.to_numeric(all_data['Price Each'])
all_data.head()
```

Out[7]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
0	176558	USB-C Charging Cable	2	11.95	04/19/19 08:46	917 1st St, Dallas, TX 75001
2	176559	Bose SoundSport Headphones	1	99.99	04/07/19 22:30	682 Chestnut St, Boston, MA 02215
3	176560	Google Phone	1	600.00	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001
4	176560	Wired Headphones	1	11.99	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001
5	176561	Wired Headphones	1	11.99	04/30/19 09:27	333 8th St, Los Angeles, CA 90001

Test Results & Screenshots

Task 1: To find out what was the best months for sales and how much money was earned by the company in that month.

Step 1: The first step of this task was to add a separate month column in the merged csv file & then to convert the data in the month column to integer format. The data in the month column needed to be extracted from the order date column.

Task 1 : What was the best months for sales? How much money was earned that month?

In [8]: `### Add Month Column and convert to integer`

```
all_data['Month'] = all_data['Order Date'].str[0:2]
all_data['Month'] = all_data['Month'].astype('int32')
all_data.head()
```

Out[8]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month
0	176558	USB-C Charging Cable	2	11.95	04/19/19 08:46	917 1st St, Dallas, TX 75001	4
2	176559	Bose SoundSport Headphones	1	99.99	04/07/19 22:30	682 Chestnut St, Boston, MA 02215	4
3	176560	Google Phone	1	600.00	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	4
4	176560	Wired Headphones	1	11.99	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	4
5	176561	Wired Headphones	1	11.99	04/30/19 09:27	333 8th St, Los Angeles, CA 90001	4

Step 2: Next step is to add a separate sales column in the merged excel sheet. The data in the sales column needs to be the product of Quantity Ordered & Price Each. Sales = (Quantity Ordered * Price Each).

```
In [18]: ### Add a Sales Column
all_data['Sales'] = all_data['Quantity Ordered']*all_data['Price Each']
all_data.head()
```

```
Out[18]:
```

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month	Sales
0	176558	USB-C Charging Cable	2	11.95	04/19/19 08:46	917 1st St, Dallas, TX 75001	4	23.90
2	176559	Bose SoundSport Headphones	1	99.99	04/07/19 22:30	682 Chestnut St, Boston, MA 02215	4	99.99
3	176560	Google Phone	1	600.00	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	4	600.00
4	176560	Wired Headphones	1	11.99	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	4	11.99
5	176561	Wired Headphones	1	11.99	04/30/19 09:27	333 8th St, Los Angeles, CA 90001	4	11.99

Step 3: Next step in this task is to group by the Month according to the sales made.

```
In [9]: ### Group by Month according to sales
all_data.groupby('Month').sum()
```

```
Out[9]:
```

	Quantity Ordered	Price Each
Month		
1	10903	1811768.38
2	13449	2188884.72
3	17005	2791207.83
4	20558	3367671.02
5	18667	3135125.13
6	15253	2562025.61
7	16072	2632539.56
8	13448	2230345.42
9	13109	2084992.09
10	22703	3715554.83
11	19798	3180600.68
12	28114	4588415.41

So, December was the best month for sales with approximately 4 Million 613 thousand, January was the worst month in sales.

Step 4: Final step in this task is to plot the sales data in graph using matplotlib lib to visualize the results better. This is an important step in data visualization.

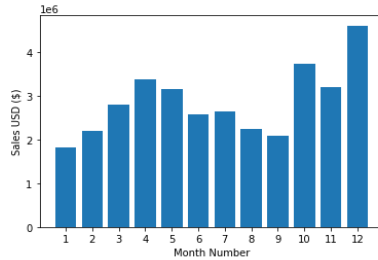
```
In [ ]: ##### Question: Plot the sales data in graph to visualize better
```

```
In [17]: import matplotlib.pyplot as plt

months = range(1,13)
print(months)

plt.bar(months,all_data.groupby(['Month']).sum()['Sales'])
plt.xticks(months)
plt.ylabel('Sales USD ($)')
plt.xlabel('Month Number')
plt.show()

range(1, 13)
```



Result Interpretation:

As evident in the results, December was the best month for sales with approximately 4 Million 613 thousand USD while January was the worst month in sales. This information can help business leaders interpret their sales data more effectively which in turn will drive profit making decisions. Business leaders can decide which month is the best for advertisement, or launching a new product, or select the best month for giving discounts to the end customer.

Task 2: To find out which city had the highest number of sales

Step 1: The first step to this task was to add a city and a state column. The data in the state and the city columns need to be extracted from the Purchase Address of the customer.

```
In [12]: ## Add a City & State Column
```

```
def get_city(address):
    return address.split(',')[1]

def get_state(address):
    return address.split(',')[2].split(' ')[1]

all_data['City'] = all_data['Purchase Address'].apply(lambda x: get_city(x))
all_data['State'] = all_data['Purchase Address'].apply(lambda x: get_state(x))

all_data.head()
```

```
Out[12]:
```

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month	Sales	City	State
0	176558	USB-C Charging Cable	2	11.95	04/19/19 08:46	917 1st St, Dallas, TX 75001	4	23.90	Dallas	TX
2	176559	Bose SoundSport Headphones	1	99.99	04/07/19 22:30	682 Chestnut St, Boston, MA 02215	4	99.99	Boston	MA
3	176560	Google Phone	1	600.00	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	4	600.00	Los Angeles	CA
4	176560	Wired Headphones	1	11.99	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	4	11.99	Los Angeles	CA
5	176561	Wired Headphones	1	11.99	04/30/19 09:27	333 8th St, Los Angeles, CA 90001	4	11.99	Los Angeles	CA

Step 2: The second step is to arrange the city in relation to the sales they made.

```
In [13]: results = all_data.groupby('City').sum()
results
```

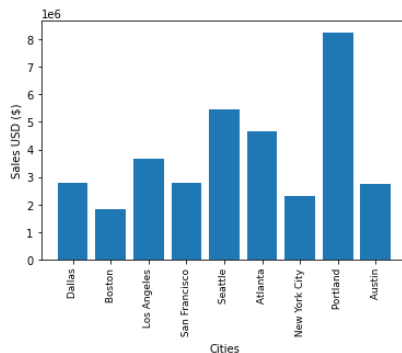
Out[13]:

City	Quantity Ordered	Price Each	Month	Sales
Atlanta	16602	2779908.20	104794	2795498.58
Austin	11153	1809873.61	69829	1819581.75
Boston	22528	3637409.77	141112	3661642.01
Dallas	16730	2752627.82	104620	2767975.40
Los Angeles	33289	5421435.23	208325	5452570.80
New York City	27932	4635370.83	175741	4664317.43
Portland	14053	2307747.47	87765	2320490.61
San Francisco	50239	8211461.74	315520	8262203.91
Seattle	16553	2733296.01	104941	2747755.48

Step 3: The final step of this task is to plot the results in graph using mat plot lib.

```
In [14]: # Plot in Graph
```

```
import matplotlib.pyplot as plt
cities = all_data['City'].unique()
plt.bar(cities, results['Sales'])
plt.xticks(cities, rotation='vertical', size=9)
plt.ylabel('Sales USD ($)')
plt.xlabel('Cities')
plt.show()
```



Therefore, San Francisco is the best city for sales.

Result Interpretation:

As evident in the results, San Francisco was the best city for sales while Boston was the worst city for sales. This useful information can help business leaders understand which city helps drives the company's revenue. Business leaders can now act upon their advertisement budget on each city, hires more or churn employees in a particular city, decide to give out discounts to the residents of a particular city.

Task 3: To find out what time should we display advertisements to maximize the like hood of the customer's buying the product.

Step 1: The first step in this task is to extract the hour and minute data from the order date column and then to create two separate columns namely hour and minute.

In [16]: `##Make an hour & minute column`

```
all_data['Hour'] = all_data['Order Date'].dt.hour
all_data['Minute'] = all_data['Order Date'].dt.minute
all_data.head()
```

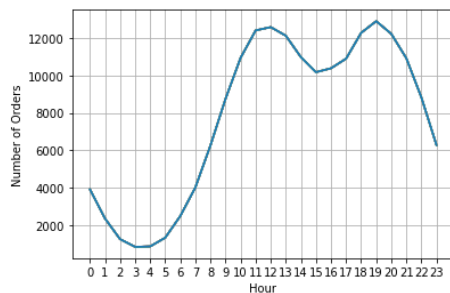
Out[16]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month	Sales	City	State	Hour	Minute
0	176558	USB-C Charging Cable	2	11.95	2019-04-19 08:46:00	917 1st St, Dallas, TX 75001	4	23.90	Dallas	TX	8	46
2	176559	Bose SoundSport Headphones	1	99.99	2019-04-07 22:30:00	682 Chestnut St, Boston, MA 02215	4	99.99	Boston	MA	22	30
3	176560	Google Phone	1	600.00	2019-04-12 14:38:00	669 Spruce St, Los Angeles, CA 90001	4	600.00	Los Angeles	CA	14	38
4	176560	Wired Headphones	1	11.99	2019-04-12 14:38:00	669 Spruce St, Los Angeles, CA 90001	4	11.99	Los Angeles	CA	14	38
5	176561	Wired Headphones	1	11.99	2019-04-30 09:27:00	333 8th St, Los Angeles, CA 90001	4	11.99	Los Angeles	CA	9	27

Step 2: Next step of this task is to plot the graph of Number of Orders placed vs. Hour, so that we can find which time is the best time to show advertisements.

```
In [19]: hours = [hour for hour, df in all_data.groupby('Hour')]
plt.plot(hours, all_data.groupby(['Hour']).count())
plt.xticks(hours)
plt.xlabel('Hour')
plt.ylabel('Number of Orders')
plt.grid()
plt.show()
```

We should advertise right before 11 AM or 6PM



Result Interpretation:

As evident in the results, right before 11 AM or before 6PM is the best time to advertise the products for the company. As analysed from the buying pattern of the end customers, it is evident that if the products of the company are advertised in the best time, then the advertisements are more likely to be seen by the potential end customers. Business Leaders can now collaborate with the sales & advertisements team and make sure that the advertisements are driven in the correct time, as this is the time when the end customers are more likely to view and buy the products.

Task 4: To find out which products are most often sold together?

Step 1: While analysing the merged sales data, the researcher came to know that when a customer orders any item or many items, that order is assigned a n unique order ID. So, if the order IDs of the products match, they were sold together.

Task 4: What products are most often sold together?

```
In [20]: ###So, if the order IDs of the products match they were sold together.
all_data.head()
```

Out[20]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month	Sales	City	State	Hour	Minute
0	176558	USB-C Charging Cable	2	11.95	2019-04-19 08:46:00	917 1st St, Dallas, TX 75001	4	23.90	Dallas	TX	8	46
2	176559	Bose SoundSport Headphones	1	99.99	2019-04-07 22:30:00	682 Chestnut St, Boston, MA 02215	4	99.99	Boston	MA	22	30
3	176560	Google Phone	1	600.00	2019-04-12 14:38:00	669 Spruce St, Los Angeles, CA 90001	4	600.00	Los Angeles	CA	14	38
4	176560	Wired Headphones	1	11.99	2019-04-12 14:38:00	669 Spruce St, Los Angeles, CA 90001	4	11.99	Los Angeles	CA	14	38
5	176561	Wired Headphones	1	11.99	2019-04-30 09:27:00	333 8th St, Los Angeles, CA 90001	4	11.99	Los Angeles	CA	9	27

Step 2: The next step is to check out all the cells in the order ID column and to list out the order IDs which are duplicate. This will help to identify the items, which the customer order together or in bundle.

```
In [23]: ### check all the cells in the order ID column and check which are duplicated
df = all_data[all_data['Order ID'].duplicated(keep=False)]
df.head(20)
```

Out[23]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month	Sales	City	State	Hour	Minute
3	176560	Google Phone	1	600.00	2019-04-12 14:38:00	669 Spruce St, Los Angeles, CA 90001	4	600.00	Los Angeles	CA	14	38
4	176560	Wired Headphones	1	11.99	2019-04-12 14:38:00	669 Spruce St, Los Angeles, CA 90001	4	11.99	Los Angeles	CA	14	38
18	176574	Google Phone	1	600.00	2019-04-03 19:42:00	20 Hill St, Los Angeles, CA 90001	4	600.00	Los Angeles	CA	19	42
19	176574	USB-C Charging Cable	1	11.95	2019-04-03 19:42:00	20 Hill St, Los Angeles, CA 90001	4	11.95	Los Angeles	CA	19	42
30	176585	Bose SoundSport Headphones	1	99.99	2019-04-07 11:31:00	823 Highland St, Boston, MA 02215	4	99.99	Boston	MA	11	31
31	176585	Bose SoundSport Headphones	1	99.99	2019-04-07 11:31:00	823 Highland St, Boston, MA 02215	4	99.99	Boston	MA	11	31
32	176586	AAA Batteries (4-pack)	2	2.99	2019-04-10 17:00:00	365 Center St, San Francisco, CA 94016	4	5.98	San Francisco	CA	17	0
33	176586	Google Phone	1	600.00	2019-04-10 17:00:00	365 Center St, San Francisco, CA 94016	4	600.00	San Francisco	CA	17	0
119	176672	Lightning Charging Cable	1	14.95	2019-04-12 11:07:00	778 Maple St, New York City, NY 10001	4	14.95	New York City	NY	11	7
120	176672	USB-C Charging Cable	1	11.95	2019-04-12 11:07:00	778 Maple St, New York City, NY 10001	4	11.95	New York City	NY	11	7

Step 3: The next step is to group the products sold together in a separate column.

```
In [28]: ### We have to group the products sold together in a separate column

df = all_data[all_data['Order ID'].duplicated(keep=False)]
df['Grouped'] = df.groupby('Order ID')['Product'].transform(lambda x: ', '.join(x))
df = df[['Order ID', 'Grouped']].drop_duplicates()
df.head()
```

Out[28]:

	Order ID	Grouped
3	176560	Google Phone,Wired Headphones
18	176574	Google Phone,USB-C Charging Cable
30	176585	Bose SoundSport Headphones,Bose SoundSport Hea...
32	176586	AAA Batteries (4-pack),Google Phone
119	176672	Lightning Charging Cable,USB-C Charging Cable

Step 4: Next step, is to list out the most common 3 items sold together.

In [29]: *### List the most common 3 items sold together*

```

from itertools import combinations
from collections import Counter

count = Counter()

for row in df['Grouped']:
    row_list = row.split(',')
    count.update(Counter(combinations(row_list,3)))

count.most_common(10)

```

```

Out[29]: [('Google Phone', 'USB-C Charging Cable', 'Wired Headphones'), 87),
          ('iPhone', 'Lightning Charging Cable', 'Wired Headphones'), 62),
          ('iPhone', 'Lightning Charging Cable', 'Apple AirPods Headphones'), 47),
          ('Google Phone', 'USB-C Charging Cable', 'Bose SoundSport Headphones'), 35),
          ('Vareebadd Phone', 'USB-C Charging Cable', 'Wired Headphones'), 33),
          ('iPhone', 'Apple AirPods Headphones', 'Wired Headphones'), 27),
          ('Google Phone', 'Bose SoundSport Headphones', 'Wired Headphones'), 24),
          ('Vareebadd Phone', 'USB-C Charging Cable', 'Bose SoundSport Headphones'),
          16),
          ('USB-C Charging Cable', 'Bose SoundSport Headphones', 'Wired Headphones'),
          5),
          ('Vareebadd Phone', 'Bose SoundSport Headphones', 'Wired Headphones'), 5)]

```

Result Interpretation:

As evident in the results, the list of the top 10 items often sold together are listed above. This useful information can help the business leaders identify which items the customers often bundles together, will help to better understand the customers buying preference. As a result, business can decide which products can be sold as a bundle in future. Company can also target its existing customers to cross sell or lure new customers with bundled promotions.

Task 5: To find out what products were sold the most.

Step 1: The first step in this task is to group by the product name with the Quantity ordered.

Task 5: What product sold the most?

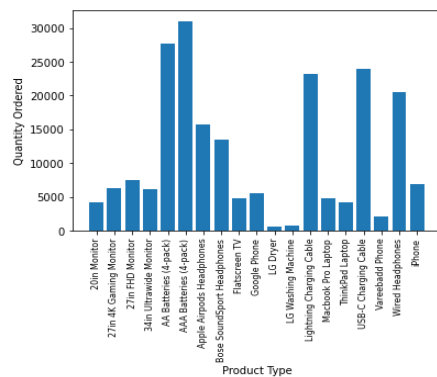
```
In [30]: product_group = all_data.groupby('Product')
quantity_ordered = product_group.sum()['Quantity Ordered']
product_group.sum()
```

Out[30]:

Product	Quantity Ordered	Price Each	Month	Sales	Hour	Minute
20in Monitor	4129	451068.99	29336	454148.71	58764	122252
27in 4K Gaming Monitor	6244	2429637.70	44440	2435097.56	90916	184331
27in FHD Monitor	7550	1125974.93	52558	1132424.50	107540	219948
34in Ultrawide Monitor	6199	2348718.19	43304	2355558.01	89076	183480
AA Batteries (4-pack)	27635	79015.68	145558	106118.40	298342	609039
AAA Batteries (4-pack)	31017	61716.59	146370	92740.83	297332	612113
Apple AirPods Headphones	15661	2332350.00	109477	2349150.00	223304	455570
Bose SoundSport Headphones	13457	1332366.75	94113	1345565.43	192445	392603
Flatscreen TV	4819	1440000.00	34224	1445700.00	68815	142789
Google Phone	5532	3315000.00	38305	3319200.00	79479	162773
LG Dryer	646	387600.00	4383	387600.00	9326	19043
LG Washing Machine	666	399600.00	4523	399600.00	9785	19462
Lightning Charging Cable	23217	323787.10	153092	347094.15	312529	634442
Macbook Pro Laptop	4728	8030800.00	33548	8037600.00	68261	137574
ThinkPad Laptop	4130	4127958.72	28950	4129958.70	59746	121508
USB-C Charging Cable	23975	261740.85	154819	286501.25	314645	647586
Vareebadd Phone	2068	826000.00	14309	827200.00	29472	61835
Wired Headphones	20557	226395.18	133397	246478.43	271720	554023
iPhone	6849	4789400.00	47941	4794300.00	98657	201688

Step 2: In this step, we have plotted the results in a graph, to better visualize which products were most often sold.

```
In [32]: product_group = all_data.groupby('Product')
quantity_ordered = product_group.sum()['Quantity Ordered']
products = [product for product,df in product_group]
plt.bar(products,quantity_ordered)
plt.ylabel('Quantity Ordered')
plt.xlabel('Product Type')
plt.xticks(products,rotation='vertical',size=8)
plt.show()
```



```
In [ ]: # So, AA batteries got sold the most, then AA batteries, USB C Charging Cable.
```

Result Interpretation:

As evident in the results, the AAA batteries were sold the most, then comes the AA batteries and then the USB C Cable. By knowing this key information, business leaders can better understand their business model, under their key revenue drivers, and better manage their advertisement budget.

Task 6: To understand the reason behind the product which got sold the most.

Step 1: The first step is to list out the prices of the products that were sold by the company.

Task 6: Why do you think those products sold the most?

```
In [33]: ### List the prices of the Products

prices = all_data.groupby('Product').mean()['Price Each']
print (prices)
```

Product	
20in Monitor	109.99
27in 4K Gaming Monitor	389.99
27in FHD Monitor	149.99
34in Ultrawide Monitor	379.99
AA Batteries (4-pack)	3.84
AAA Batteries (4-pack)	2.99
Apple AirPods Headphones	150.00
Bose SoundSport Headphones	99.99
Flatscreen TV	300.00
Google Phone	600.00
LG Dryer	600.00
LG Washing Machine	600.00
Lightning Charging Cable	14.95
Macbook Pro Laptop	1700.00
ThinkPad Laptop	999.99
USB-C Charging Cable	11.95
Vareebadd Phone	400.00
Wired Headphones	11.99
iPhone	700.00

Name: Price Each, dtype: float64

Step 2: The second step is to plot a graph to understand how the customers buying preference depends on the price of the product.

```
In [39]: ### check & plot the prices of the Products

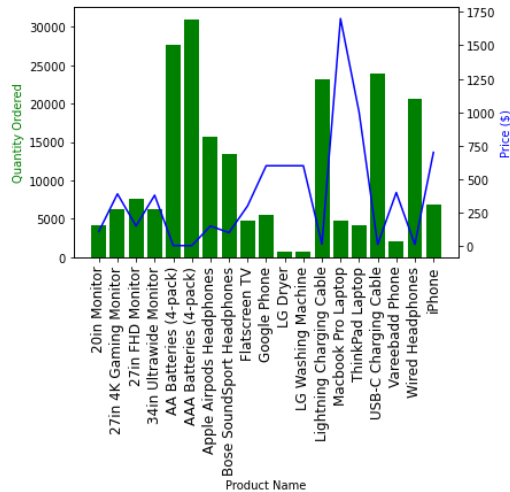
prices = all_data.groupby('Product').mean()['Price Each']

fig, ax1 = plt.subplots()

ax2 = ax1.twinx()
ax1.bar(products, quantity_ordered , color='g')
ax2.plot(products,prices, 'b-')

ax1.set_xlabel('Product Name')
ax1.set_ylabel('Quantity Ordered',color='g')
ax2.set_ylabel('Price ($)',color='b')
ax1.set_xticklabels(products,rotation='vertical',size=12)

plt.show()
```



Result Interpretation:

As evident in the graphical plot of Quantity ordered, Price & Product name, we can say that the quantity ordered & the price of the item are inversely proportional in most cases. This buying preference of the end customers can help the business leaders better formulate a pricing strategy of the products based on the customers buying preference.

7. CONCLUSION & FUTURE STUDY

We have gone into great length on explorative data analysis in this thesis. For the implementation, we utilised the Python programming language. For detailed investigation, we utilised a jupyter notebook. Different Python library packages have been used for data analysis. Using various parameters, we were able to get the desired result. We will utilise additional data sets and more functions in the future to acquire a more vivid picture of predictive data analysis. Product sales prediction systems are required by organisations to manage huge amounts of data, according to the findings of this project's research. Decision-making is based on the speed and precision of data processing technology. To be competent, businesses must equip themselves with specific strategies that allow them to account for various types of customer behaviour by forecasting product sales and demand in order to enhance profits. For the initial algorithm comparison, around 16,000 records of data instances were gathered. However, due to the lengthy implementation time and the difficulty of managing such a large collection of data, some records were rejected during the analysis and data processing phase. At the same time, the variables and characteristics employed in this study were not suitable for further investigation. During the research, it proved to be a significant challenge.

Because this is simply the initial idea of an application, it can be enhanced using far more advanced machine learning techniques. This work may also be put to the test by applying it to more diverse industries, and the system's performance can be evaluated as a result. Future work directions are numerous as a result of this work, and it is up to the user's interest to use it in the appropriate direction. Further, large-scale experiments utilising larger datasets and alternative modelling methodologies (using different types of basic functions) can be carried out in the future with the goal of developing a trustworthy model that might be used in clinical practise.

REFERENCES

1. Food Industry Sales Prediction by Maja Lindström, June 9, 2021
2. Consumer Packaged Goods (CPG) Predictive Analytics Models by Mantzoufa Ioanna, November 2018.
3. The contribution of Data Analytics in predicting the future purchase intentions of consumers by Pinakshi Kalita, August 2019
4. Prediction of user behaviour on the web by Nikolay Burlutskiy, 2017
5. Determining predictive metabolomic biomarkers of meniscal injury in dogs with cranial cruciate ligament disease using stifle joint synovial fluid by Pye, Christine (2021)