**CHAPTER – I**

**INTRODUCTION**

* 1. **INTRODUCTION OF STUDY**

**E-commerce**

Ecommerce (electronic commerce) refers to all online activity that involves the buying and selling of products and services. In other words, ecommerce is a process for conducting transactions online.

When you go to your favourite online retailer to buy a new pair of shoes, you’re engaging in ecommerce. If you pay online to attend a music concert or buy a plane ticket through the airline’s website, that’s ecommerce, too.

Ecommerce doesn’t only occur on desktop, though. In fact, most ecommerce traffic is driven through mobile commerce. Spurred by the influence of smartphones and the comfort of online shopping, mobile commerce sales are expected to account for more than half of all ecommerce sales by 2021. The term was coined and first employed by Dr. Robert Jacobson, Principal Consultant to the California State Assembly's Utilities & Commerce Committee, in the title and text of California's Electronic Commerce Act, carried by the late Committee Chairwoman Gwen Moore (D-L.A.) and enacted in 1984.

E-commerce typically uses the web for at least a part of a transaction's life cycle although it may also use other technologies such as e-mail. Typical e-commerce transactions include the purchase of products (such as books from Amazon) or services (such as music downloads in the form of digital distribution such as iTunes Store).[1] There are three areas of e-commerce: online retailing, electronic markets, and online auctions. E-commerce is supported by electronic business.[2] The existence value of e-commerce is to allow consumers to shop online and pay online through the Internet, saving the time and space of customers and enterprises, greatly improving transaction efficiency, especially for busy office workers, but also saving a lot of valuable time.

E-commerce businesses may also employ some or all of the following:

* Online shopping for retail sales direct to consumers via web sites and mobile apps, and conversational commerce via live chat, chatbots, and voice assistants;[4]
* Providing or participating in online marketplaces, which process third-party business-to-consumer (B2C) or consumer-to-consumer (C2C) sales;
* Business-to-business (B2B) buying and selling;
* Gathering and using demographic data through web contacts and social media;
* B2B electronic data interchange;
* Marketing to prospective and established customers by e-mail or fax (for example, with newsletters);
* Engaging in retail for launching new products and services;
* Online financial exchanges for currency exchanges or trading purposes.

**History of Ecommerce**

Most of us have shopped online for something at some point, which means we've taken part in ecommerce. So it goes without saying that ecommerce is everywhere. But very few people may know that ecommerce has a history that goes back before the internet began.

Ecommerce actually goes back to the 1960s when companies used an electronic system called the Electronic Data Interchange to facilitate the transfer of documents. But it wasn't until 1994 that the very first transaction. took place. This involved the sale of a CD between friends through an online retail website called NetMarket.33

The industry has gone through so many changes since then, resulting in a great deal of evolution. Traditional brick-and-mortar retailers were forced to embrace new technology in order to stay afloat as companies like Alibaba, Amazon, eBay, and Etsy became household names. These companies created a virtual marketplace for goods and services that consumers can easily access.

New technology continues to make it easier for people to do their online shopping. People can connect with businesses through smartphones and other devices and by downloading apps to make purchases. The introduction of free shipping, which reduces costs for consumers, has also helped increase the popularity of the ecommerce industry.

**1.2 E-Commerce Sales Prediction**

At present, the research results on product sales forecasting are relatively rich, and the research methods are different. Bi and Wei improved BP neural network from the two aspects of sample quality and initial weight by using principal component analysis method and particle swarm optimization algorithm [1];established a neural network prediction model for cigarette sales by using the improved BP neural network Levenberg–Marquardt algorithm . Research results using time series prediction method are also common. For example, Peng and Yu use RBF neural network to predict product sales based on time series analysis and optimize the prediction model [3]; Wang extracted product clusters according to product sales commonness and established a product reclassification time series sales prediction model based on sales data [4]. Some scholars have also adopted the support vector machine prediction method, such as Wu and Lin. Taking the cigarette sales of specific tobacco enterprises as the research object, they have proposed a hybrid method for cigarette sales prediction based on support vector machine . The research methods in the above literature have their own advantages, but there are also many disadvantages: first, online product data samples often have diversified characteristics, while most models do not have diversified data processing ability; second, with the increasing scale of online product sales, the resulting massive sales data not only is the basic basis for sales forecasting, but also reflects the deficiency of traditional forecasting methods in dealing with large-scale data. For example, Bi et al. used shallow neural network, which has advantages in big data processing, but the prediction accuracy needs to be improved. Liu et al. established crown model based on deep learning algorithm, on the basis of fully considering the characteristics of agricultural e-commerce sales data, and used this model to realize the classified prediction of online agricultural product sales [6]. Deep learning algorithm has its unique advantages in online product sales forecasting. Firstly, deep learning improves the training algorithm based on BP neural network, and the gradient disappearance problem is effectively solved, so that the effective time of training is longer. Secondly, online product sales forecasting needs high generalization model support. The deep learning model with high complexity capacity has good generalization in the big data environment. Thirdly, compared with the general model, deep learning can extract more and more effective information from massive data. Finally, deep learning has the feature of building layer by layer, which can extract higher-level features from the existing data, decompose the influencing factors of interaction into independent and more effective factors, and improve the prediction accuracy of the model. Based on the above advantages, this paper aims to establish a relatively perfect index system of influencing factors of online product sales and use deep learning algorithm to build a sales prediction model of all kinds of online products. Because the online product sales forecasting model based on deep learning algorithm usually classifies products and designs the model according to the characteristics of a certain kind of products, such a model has poor adaptability. Once the product type changes, the influencing factor index and model must be redesigned. Therefore, this paper not only evaluates the prediction accuracy and generalization ability of the model, but also focuses on the adaptability of the model.

**1.3 Data Science**

Data science is the domain of study that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions. Data science uses complex machine learning algorithms to build predictive models. The data used for analysis can come from many different sources and presented in various formats. Now that you know what data science is, let’s see why data science is essential to today’s IT landscape. Now that you know what is data science, next up let us focus on the data science lifecycle. Data science’s lifecycle consists of five distinct stages, each with its own tasks:

Capture: Data Acquisition, Data Entry, Signal Reception, Data Extraction. This stage involves gathering raw structured and unstructured data.

Maintain: Data Warehousing, Data Cleansing, Data Staging, Data Processing, Data Architecture. This stage covers taking the raw data and putting it in a form that can be used.

Process: Data Mining, Clustering/Classification, Data Modeling, Data Summarization. Data scientists take the prepared data and examine its patterns, ranges, and biases to determine how useful it will be in predictive analysis.

Analyze: Exploratory/Confirmatory, Predictive Analysis, Regression, Text Mining, Qualitative Analysis. Here is the real meat of the lifecycle. This stage involves performing the various analyses on the data.

Communicate: Data Reporting, Data Visualization, Business Intelligence, Decision Making. In this final step, analysts prepare the analyses in easily readable forms such as charts, graphs, and reports.

Prerequisites for Data Science Here are some of the technical concepts you should know about before starting to learn what is data science.

**Machine Learning**

Machine learning is the backbone of data science. Data Scientists need to have a solid grasp of ML in addition to basic knowledge of statistics.

Modeling

Mathematical models enable you to make quick calculations and predictions based on what you already know about the data. Modeling is also a part of Machine Learning and involves identifying which algorithm is the most suitable to solve a given problem and how to train these models.

Statistics

Statistics are at the core of data science. A sturdy handle on statistics can help you extract more intelligence and obtain more meaningful results.

Programming

Some level of programming is required to execute a successful data science project. The most common programming languages are Python, and R. Python is especially popular because it’s easy to learn, and it supports multiple libraries for data science and ML.

Databases

A capable data scientist needs to understand how databases work, how to manage them, and how to extract data from them.

**1.4 DATA PROCESSING**

Data, when initially obtained, must be processed or organized for analysis. For instance, these may involve placing data into rows and columns in a table format (known as structured data) for further analysis, often through the use of spreadsheet or statistical software.

**1.5 DATA CLEANING**

Once processed and organized, the data may be incomplete, contain duplicates, or contain errors. The need for data cleaning will arise from problems in the way that the datum is entered and stored. Data cleaning is the process of preventing and correcting these errors. Common tasks include record matching, identifying inaccuracy of data, overall quality of existing data, deduplication, and column segmentation. Such data problems can also be identified through a variety of analytical techniques. For example; with financial information, the totals for particular variables may be compared against separately published numbers that are believed to be reliable. Unusual amounts, above or below predetermined thresholds, may also be reviewed. There are several types of data cleaning, that are dependent upon the type of data in the set; this could be phone numbers, email addresses, employers, or other values. Quantitative data methods for outlier detection, can be used to get rid of data that appears to have a higher likelihood of being input incorrectly. Textual data spell checkers can be used to lessen the amount of mis-typed words. However, it is harder to tell if the words themselves are correct

**1.6 Data Visualization**

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. Our eyes are drawn to colours and patterns. We can quickly identify red from blue, square from circle. Our culture is visual, including everything from art and advertisements to TV and movies. Data visualization is another form of visual art that grabs our interest and keeps our eyes on the message. When we see a chart, we quickly see trends and outliers. If we can see something, we internalize it quickly. It’s storytelling with a purpose. If you’ve ever stared at a massive spreadsheet of data and couldn’t see a trend, you know how much more effective a visualization can be. Data visualization is the practice of translating information into a visual context, such as a map or graph, to make data easier for the human brain to understand and pull insights from. The main goal of data visualization is to make it easier to identify patterns, trends and outliers in large data sets. The term is often used interchangeably with others, including information graphics, information visualization and statistical graphics.

Data visualization is one of the steps of the data science process, which states that after data has been collected, processed and modeled, it must be visualized for conclusions to be made. Data visualization is also an element of the broader data presentation architecture (DPA) discipline, which aims to identify, locate, manipulate, format and deliver data in the most efficient way possible.

Data visualization is important for almost every career. It can be used by teachers to display student test results, by computer scientists exploring advancements in artificial intelligence (AI) or by executives looking to share information with stakeholders. It also plays an important role in big data projects. As businesses accumulated massive collections of data during the early years of the big data trend, they needed a way to quickly and easily get an overview of their data. Visualization tools were a natural fit.

Visualization is central to advanced analytics for similar reasons. When a data scientist is writing advanced predictive analytics or machine learning algorithms, it becomes important to visualize the outputs to monitor results and ensure that models are performing as intended. This is because visualizations of complex algorithms are generally easier to interpret than numerical outputs.

1.7 Need of Study

* To drive our point home, here are some of the benefits that an accurate sales forecast brings to the table.
* An accurate sales forecast allows companies to efficiently allocate resources for future growth and manage their cash flow.
* Sales forecasts help set benchmarks for future trends and allow leaders to course correct early. Revenue leaders can align sales quotas and revenue expectations and optimize for more wins.
* Conveys confidence to the board, and the management team that your business is supported by a reliable forecasting machine that will scale well in the future.
* Sales projections facilitate strategic planning and tell you how soon you will be ready for executing and implementing your plans.

1.8 Objectives of the study

1. To Analyse the Distribution of the Data
2. To Analyse the pattern of the Sales
3. To check the Trend and seasonality
4. To predict the sales by a model

1.9 Scope of the Study

* To analyse the popular product
* To analyse the common stocks
* To analyse the Revenue by Month
* To analyse the top Customers who purchase and visit frequently
* To analyse the Order data

1.10 Limitations of the study

* The Study will have only one model that is linear regression and its error rate
* This Study will only contains about the seasonality and trends Visualization
* In this study the as the data is insufficient it can be hard to predict the yearly analysis

**CHAPTETR II**

**Review of literature**

E-commerce is a platform where people are able to buy and sell goods. The main purpose of e-commerce is to provide convenience to the customers where they do not have to go to a physical store to make a purchase. As the will be able to make the purchase online and the item will be in their door step in the following days. In 2019, a total of $603 billion worth of sales were done via e-commerce in the United States compared to 3.17 billion in retail sales in the United States. The purpose of this study was to build machine learning algorithms which are able to forecast the sales of the e-commerce platform. A research was being done to understand the literature reviews based on similar systems and similar studies that relates to the researcher project. The purpose of doing this literature review is to understand which machine learning model was being used by other studies so the researcher will be able to select some of the best machine learning models for this study. Once the researcher has selected the models, he will them build the models and test their accuracy, error and performance. At the end, the researcher will compare all of the model’s accuracy and errors to get the best model which have low error and high accuracy for forecasting sales. The model which have been fulfil the criteria, will be integrated into the system which is being built by the researcher. The system will give a view of the current and forecasted sales.

**CHAPTER III**

**Research methodology**

3.1 Research Design

**Exploratory Data Analysis**

In statistics, exploratory data analysis is an approach of analysing datasets to summarize their main characteristics, often using statistical graphics and other data visualization methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal Modeling and thereby contrasts traditional hypothesis testing. Exploratory data analysis has been promoted by John Tusky since 1970 to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. EDA is different from initial data analysis (IDA). Which 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 encompasses IDA

Data analysis is a process of inspecting, [cleansing](https://en.wikipedia.org/wiki/Data_cleansing), [transforming](https://en.wikipedia.org/wiki/Data_transformation), and [modelling](https://en.wikipedia.org/wiki/Data_modelling) [data](https://en.wikipedia.org/wiki/Data) with the goal of discovering useful information, informing conclusions, and supporting decision-making

Analysis, refers to dividing a whole into its separate components for individual examination. Data analysis, is a [process](https://en.wikipedia.org/wiki/Process_theory) for obtaining [raw](https://en.wikipedia.org/wiki/Raw_data) data, and subsequently converting it into information useful for decision-making by users. The data is necessary as inputs to the analysis, which is specified based upon the requirements of those directing the analysis

Tukey defined data analysis in 1961 as: "Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data."[3]

Tukey's championing of EDA encouraged the development of statistical computing packages, especially S at Bell Labs.[4] The S programming language inspired the systems S-PLUS and R. This family of statistical-computing environments featured vastly improved dynamic visualization capabilities, which allowed statisticians to identify outliers, trends and patterns in data that merited further study.

Tukey's EDA was related to two other developments in statistical theory: robust statistics and nonparametric statistics, both of which tried to reduce the sensitivity of statistical inferences to errors in formulating statistical models. Tukey promoted the use of five number summary of numerical data—the two extremes (maximum and minimum), the median, and the quartiles—because these median and quartiles, being functions of the empirical distribution are defined for all distributions, unlike the mean and standard deviation; moreover, the quartiles and median are more robust to skewed or heavy-tailed distributions than traditional summaries (the mean and standard deviation). The packages S, S-PLUS, and R included routines using resampling statistics, such as Quenouille and Tukey's jackknife and Efron's bootstrap, which are nonparametric and robust (for many problems).

Exploratory data analysis, robust statistics, nonparametric statistics, and the development of statistical programming languages facilitated statisticians' work on scientific and engineering problems. Such problems included the fabrication of semiconductors and the understanding of communications networks, which concerned Bell Labs. These statistical developments, all championed by Tukey, were designed to complement the analytic theory of testing statistical hypotheses, particularly the Laplacian tradition's emphasis on exponential families.

**3.2 Data Collection**

Secondary Data From : <https://www.kaggle.com/code/allunia/e-commerce-sales-forecast>

Context

Typically e-commerce datasets are proprietary and consequently hard to find among publicly available data. However, The UCI Machine Learning Repository has made this dataset containing actual transactions from 2010 and 2011. The dataset is maintained on their site, where it can be found by the title "Online Retail".

Content

"This is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers."

InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction.

StockCode: code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.

Description: Product name. Nominal.

Quantity: The quantities of each product (item) per transaction. Numeric.

InvoiceDate: Invoice Date and time. Numeric, the day and time when each transaction was generated.

UnitPrice: Unit price. Numeric, Product price per unit in sterling.

CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.

Country: Country name. Nominal, the name of the country where each customer resides.

Reference: This is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts.

3.3 Period of Study

The period of Study is from March 15 2022 to May 15 2022

3.4 Tools Used

* Python
* Packages Used
* numpy , pandas, seaborn, matplotlib, plotly.express, warnings,warnings.filterwarnings('ignore')

**CHAPTER IV**

**ANALYSIS AND INTERPRETATION**

* 1. Data Preview

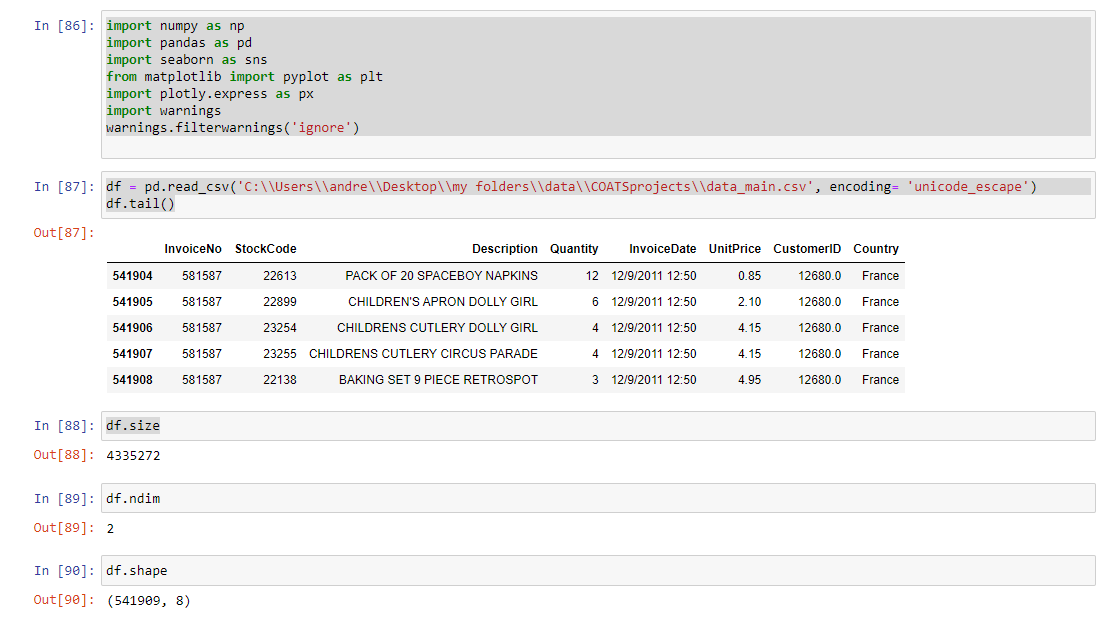


Fig 1.1

Interpretation:

Attributes of the Data set

InvoiceNo –invoice Number (Order number or bill Number)

Stockcode – Stock

Description – Product

Quantity – Buying Quantity

InvoiceDate - Bill Date

UnitPrice - Price Per Unit

Customer ID – Unique Customer

Country – Residency of the Customer

Shape of the Data – 541909,8

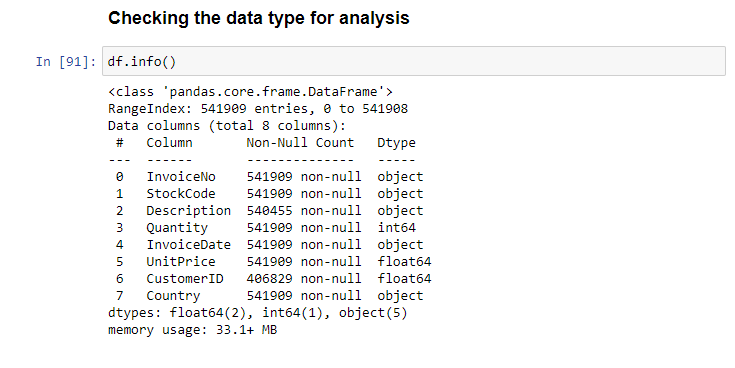


Fig1.2

Interpretation:

In the above we can see that the Invoice Date is in Object format we should convert it into a appropriate data type (date\_time)

And the customerID is in float value It should be changed to a string value, because it can cause unnecessary disturbance in the analysis when using mathematical functions

Checking for Null values

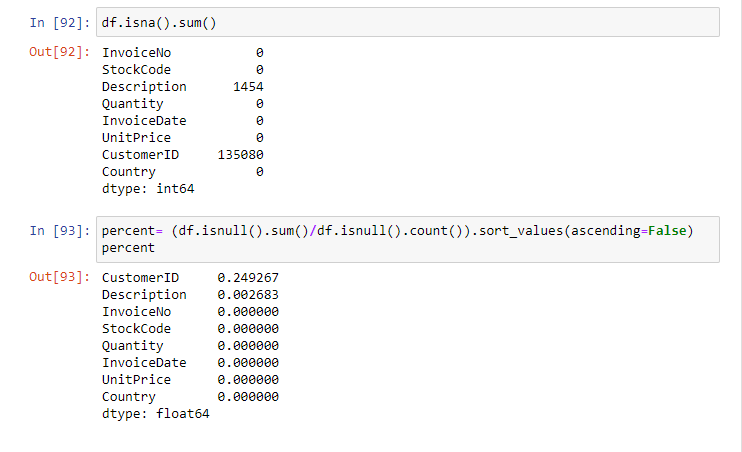


Fig 1.3

Interpretation:

Here the Nan values are Description and CustomerID

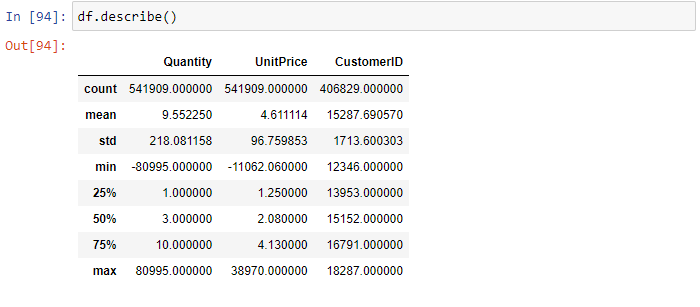


Fig 1.4

Interpretation:

In here we can see there are negative values in UnitPrice and Quantity



Fig 1.5

4.2 .Data Preparation & Process

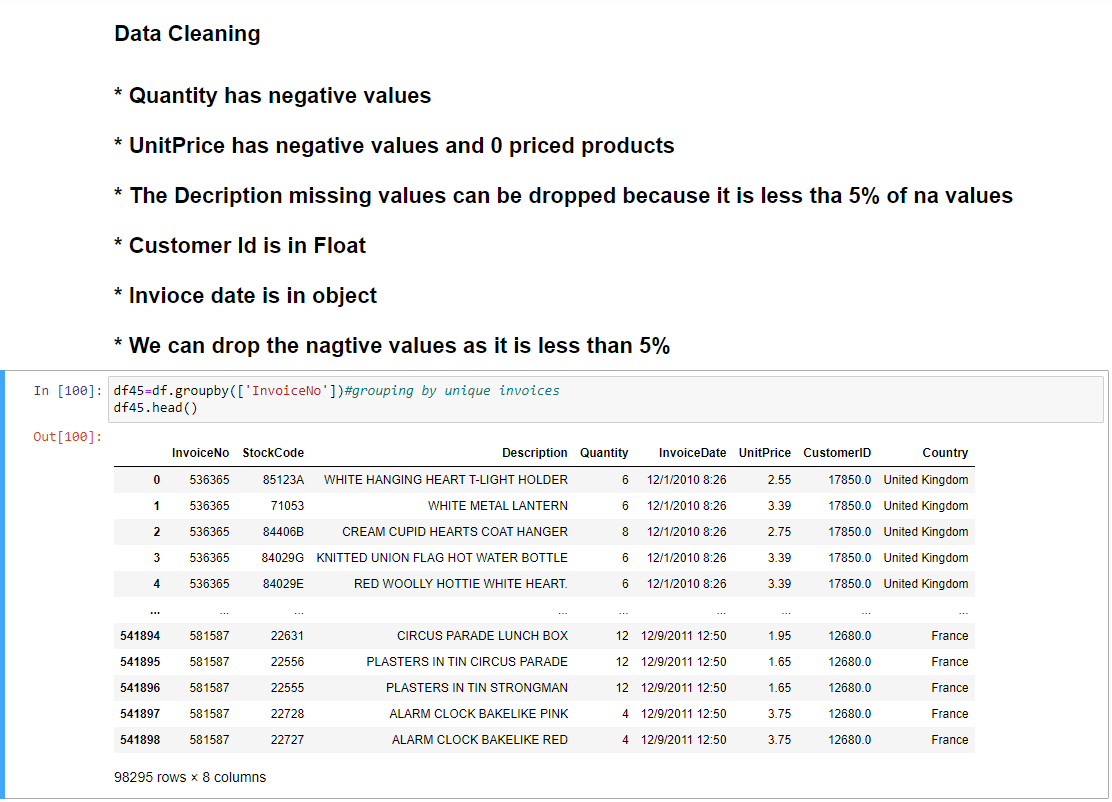


Fig 1.6

Interpretation:

We are grouping by the InvoiceNo to eliminate the repeating Invoice Number

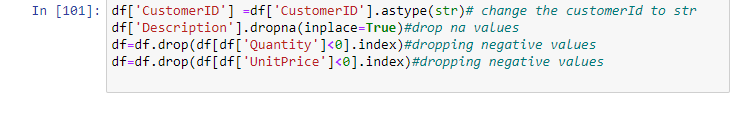


Fig 1.7

Interpretation:

In here we are converting the customer ID to a string value

Dropping Null values because its missing percentage is low we can drop the rows it does not effect the analysis

Dealing with negative values, It should be dropped because it does not make sense when quantity and UnitPrice have negative values

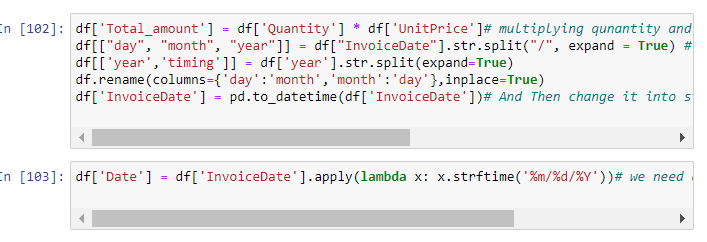


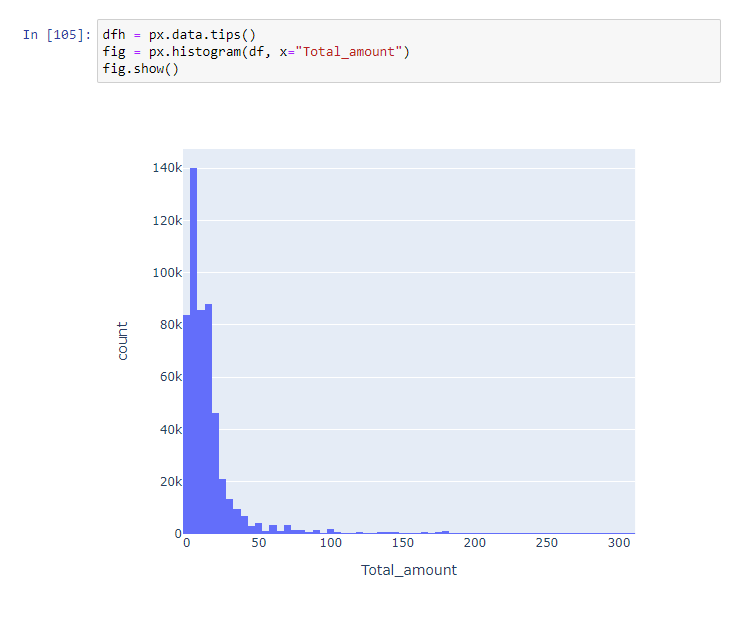
Fig 1.8

Interpretation:

Creating a total amount column by multiplying the Quantity and Unit Price Splitting the day, month, year and timing and create separate column for that Indexing and renaming the columns

Changing Invoice Object Data type to Datetime DataFrame

4.3 Frequency Distribution analysis

Fig 1.9

Interpretation:

Total amount distribution is left skewed (The mean is less than the median, which is often less than the mode)

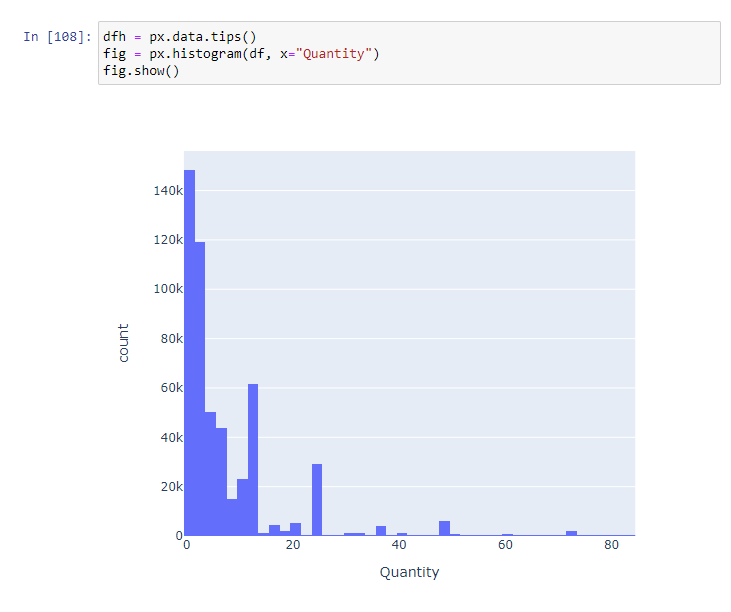


Fig 1.10

Interpretation:

Quantity distribution is left skewed (The mean is less than the median, which is often less than the mode)

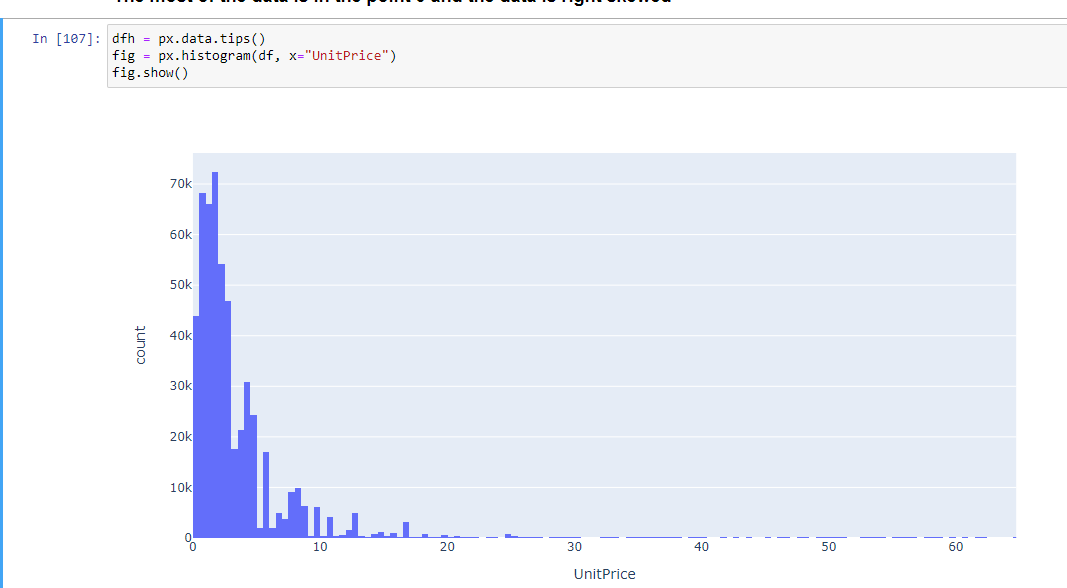


Fig 1.11

Interpretation:

UnitPrice distribution is left skewed (The mean is less than the median, which is often less than the mode)

4.4 Exploratory Data Analysis

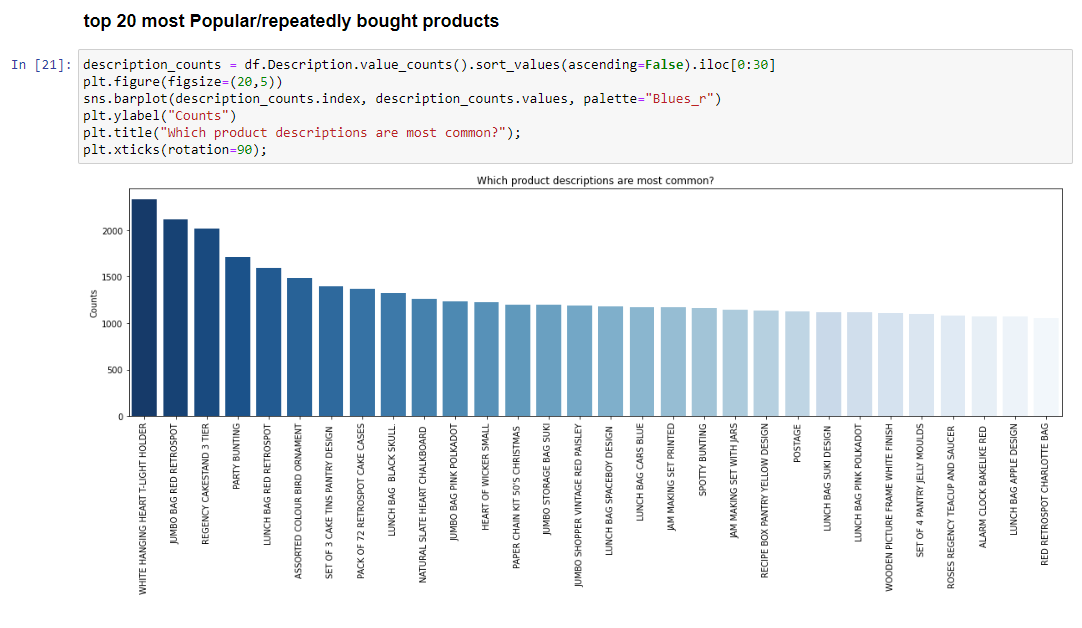


Fig 1.12

Interpretation:The most popularly bought product is WHITE HANGING T-LIGHT HOLDER

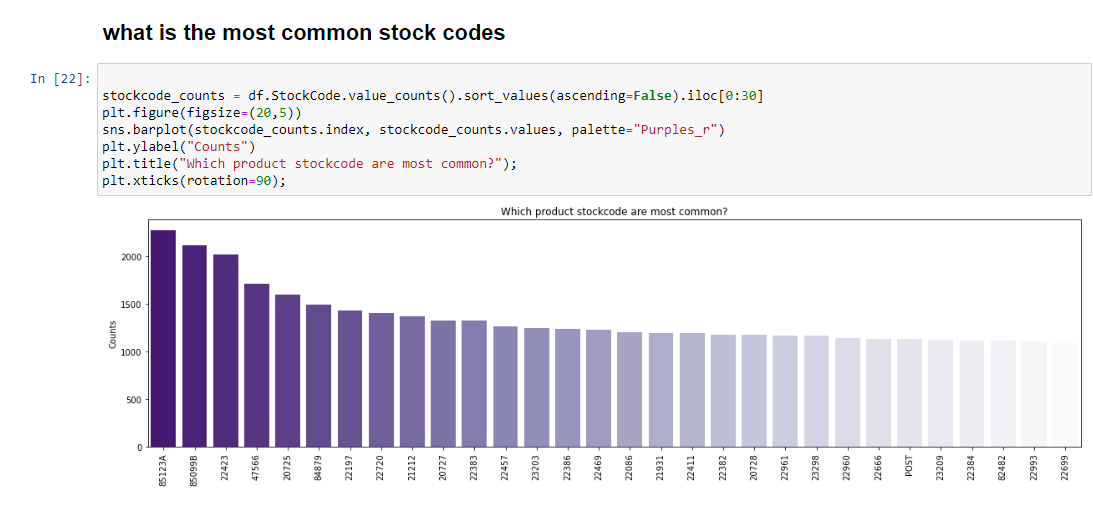


Fig 1.13

Interpretation:

The most common stock code is 85123A

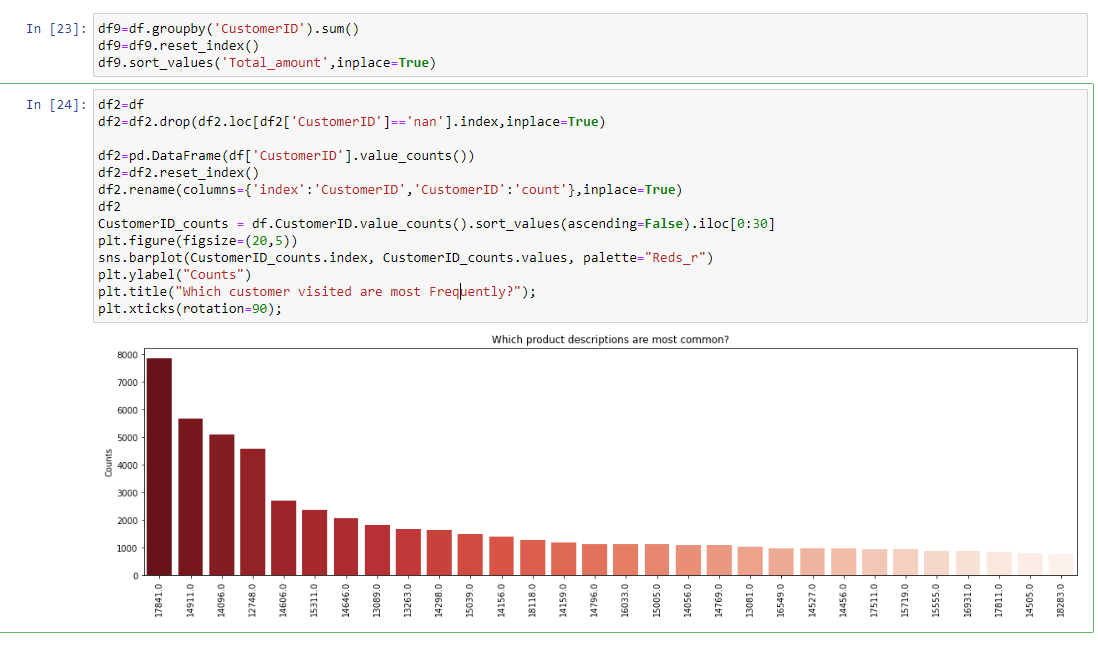


Fig1.14

Interpretation:

The customer visited the most times is CustomerID No:17841

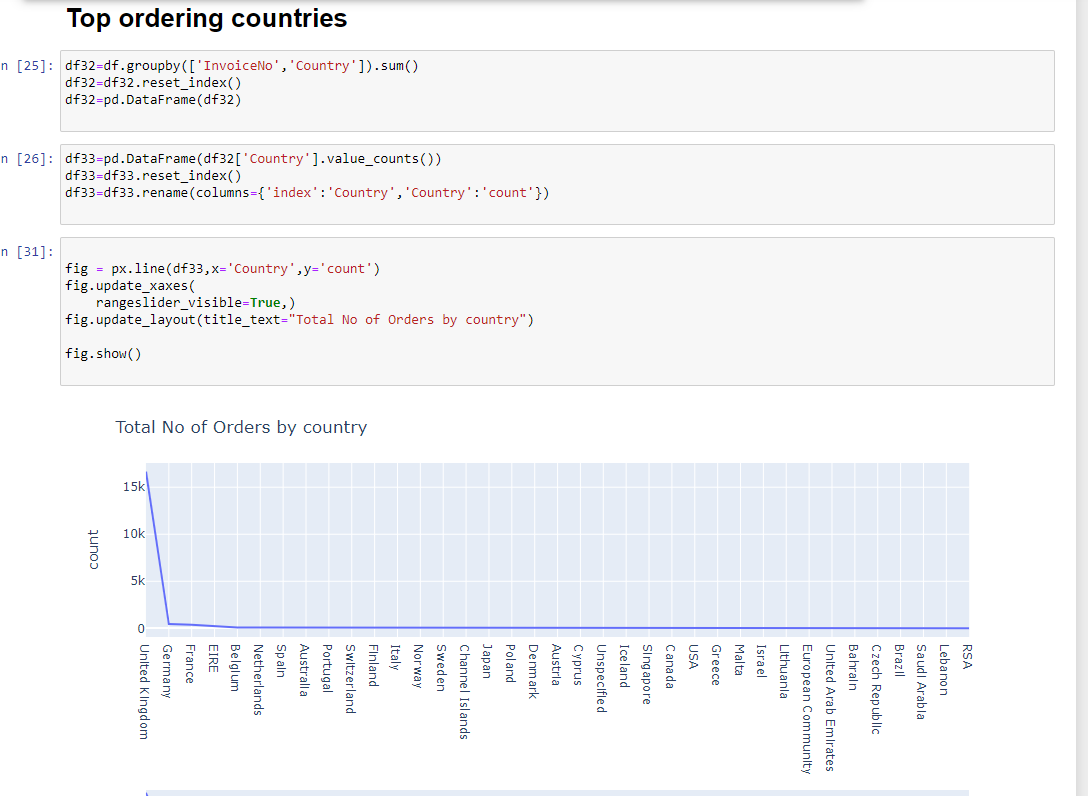


Fig 1.15

Interpretation:

The country which ordered the most orders is Untited knight with the total order of 16.49K

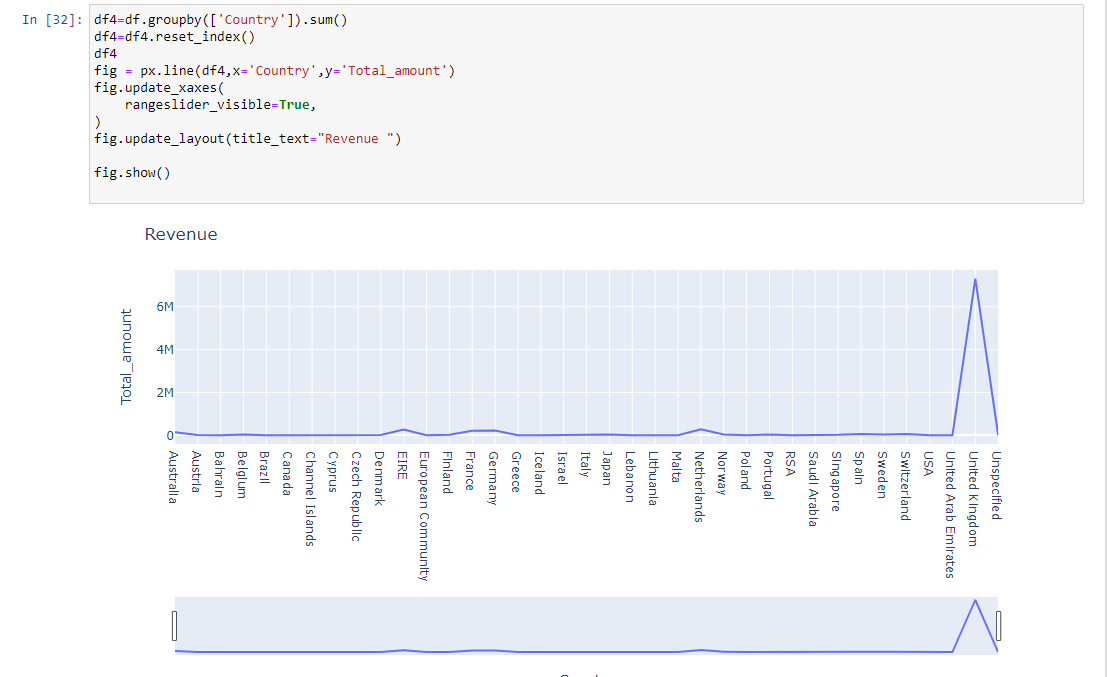


Fig 1.16

Interpretation:

The country which has the highest Total Revenue is United Kingdom with the amount of 7.3 million

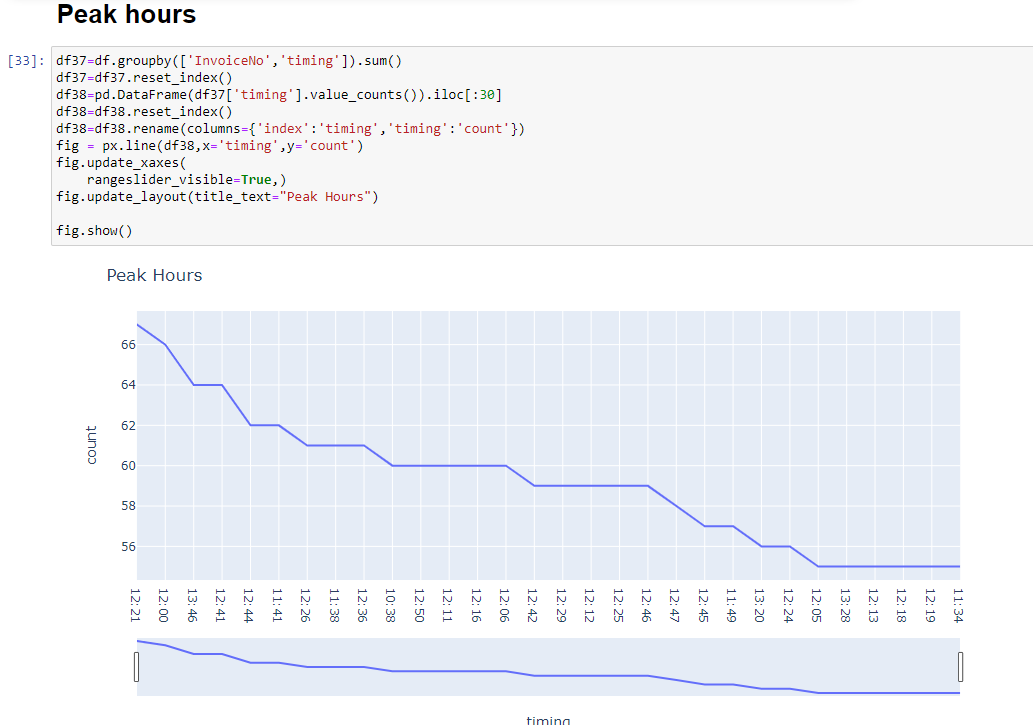


Fig 1.17

Interpretation:

The peak hours are from 11am to 1pm on average

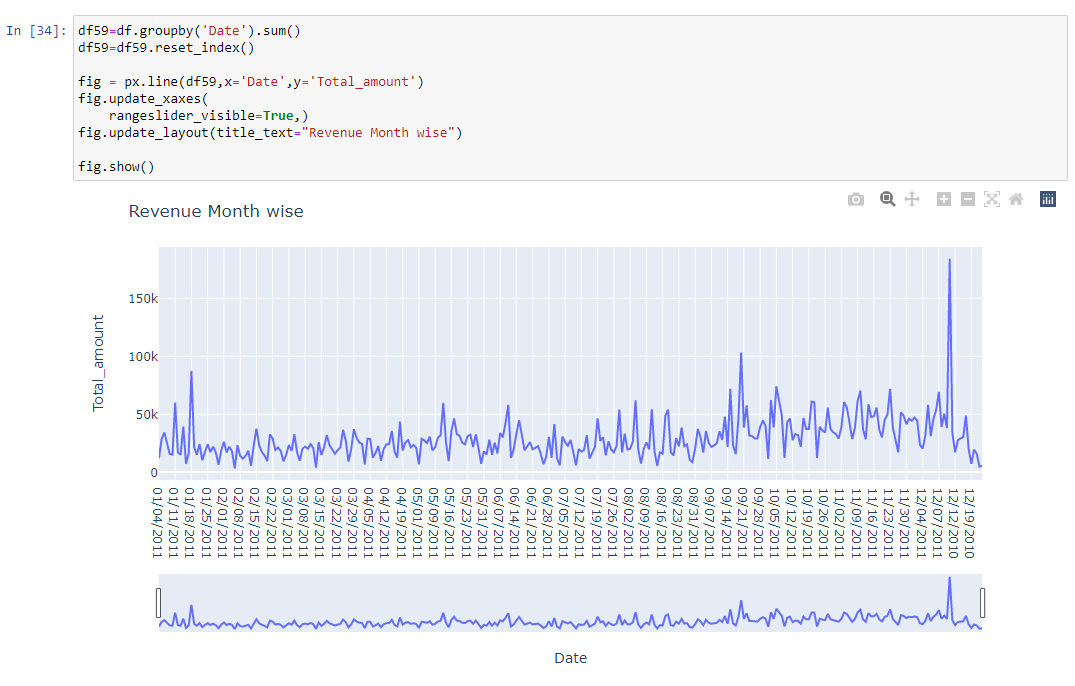


Fig 1.18

Interpredation: 184.34k is the over all Highest Revenue (on the 12th Nov 2011)

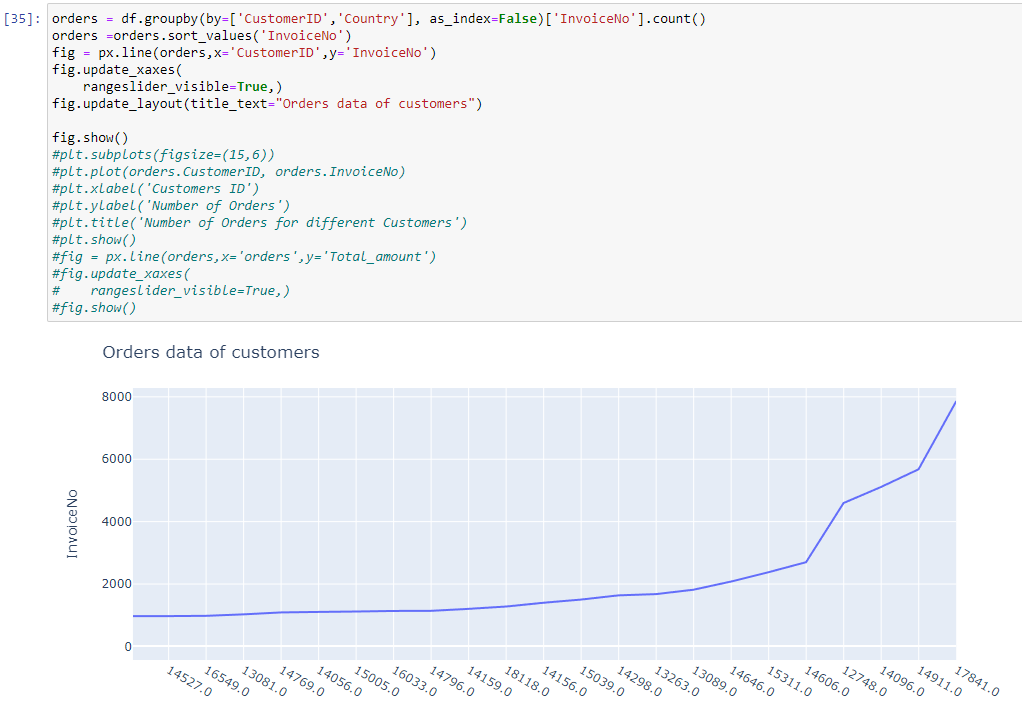


Fig 1.20

Interpretation: The customer who orders the most orders is CustomerID 17841.



Fig 1.21

Interpretation:

we create a new dataframe by splitting from the main dataframe

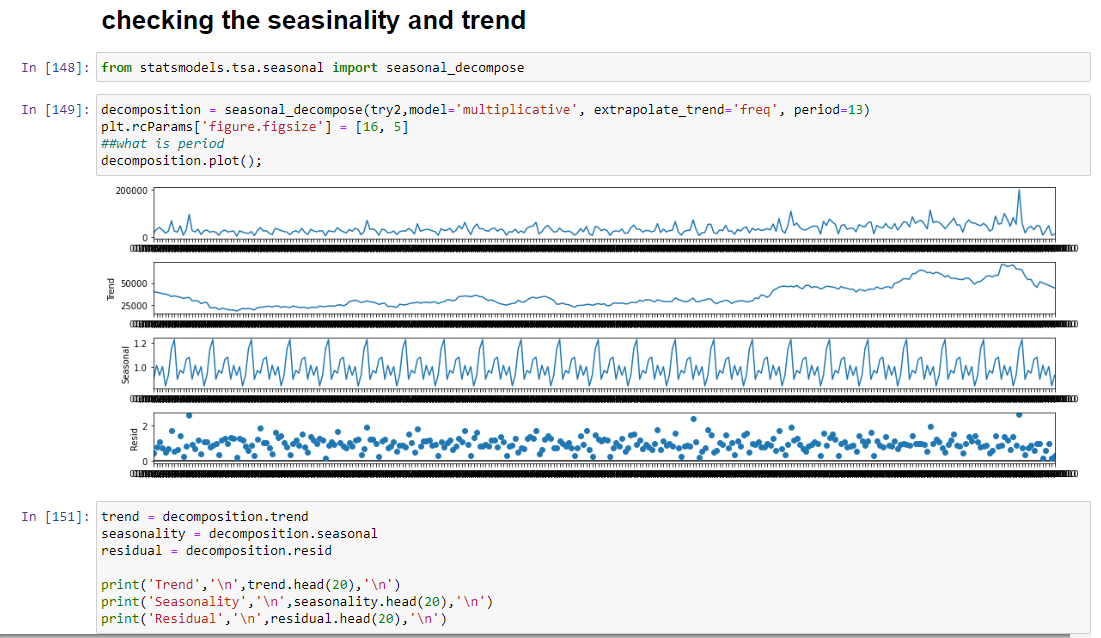
4.5 Checking for seasonality and trend 

Fig 1.22

Interpretation:

In here we are checking if there is seasonality and trend in the data As you can see there is trend pattern and seasonality pattern in the data

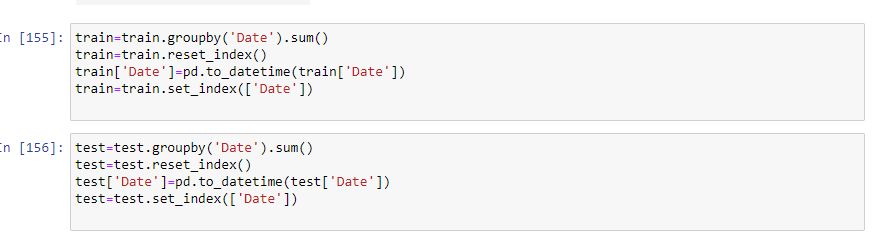
4.6 Linear Regression

Creating new data frame for linear regression



Fig 1.23

Testing and Train the data by splitting the data by date before 07-2011 and after 07-2011



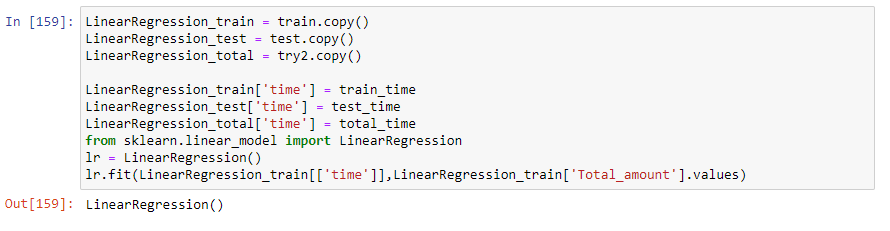


Fig 1.24

Interpretation:

Importing the time and revenue data from the test and train values

Training the data and after training, I will test the data and it will check its accuracy by the below graph

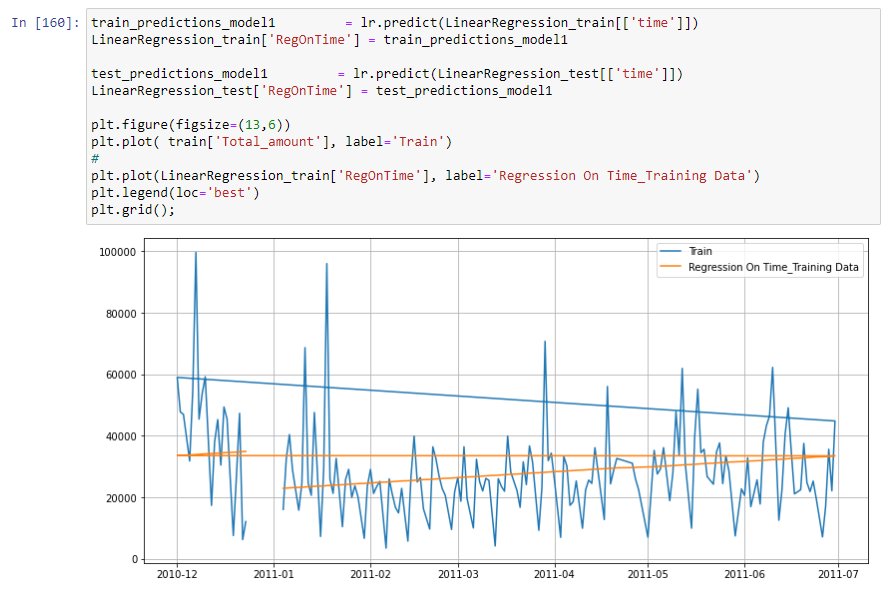
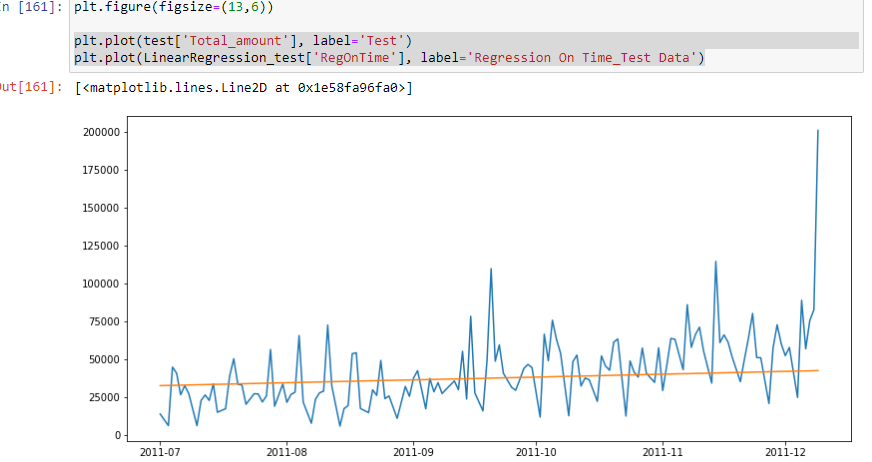


Fig 1.26

In the above its is the trained data

In the below it is the tested data and the yellow line is the prediction value



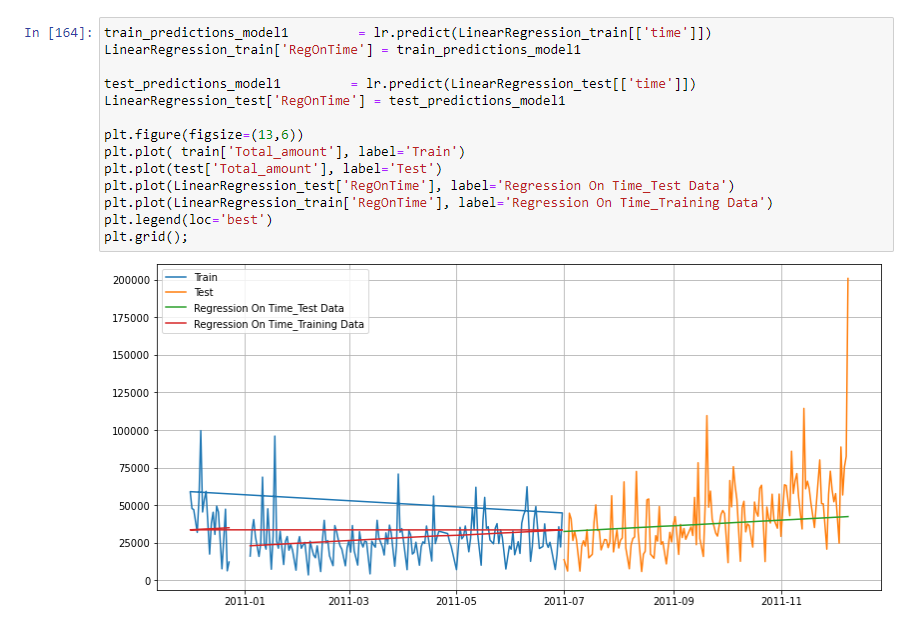


Fig 1.27

Interpretation:

it’s the combined chart of both charts above

**4.7 Model evaluation**

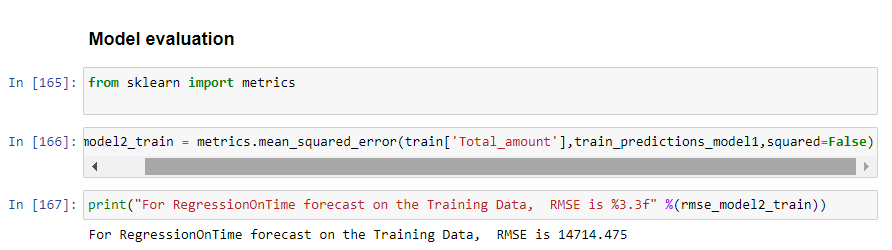


Fig 1.28

Interpretation: In here we are checking the error of the linear regression model The mean\_squared\_value is should be closer to zero but in here the value is high So this model wouldn’t fit the data

**CHAPTER V**

**Findings Suggestion And Conclusion**

**5.1 summary of Findings**

* The most popularly bought product is WHITE HANGING T-LIGHT HOLDER
* The most common stock code is 85123A
* The customer visited the most times is CustomerID No:17841
* The country which ordered the most orders is Untited knight with the total order of 16.49K
* The country which has the highest Total Revenue is United Kingdom with the amount of 7.3 million
* The peak hours are from 11am to 1pm on average
* The customer who orders the most orders is CustomerID 17841
* Have trend pattern and seasonality pattern in the data

**5.2 Suggestion**

* To perform more analysis types like outlier analysis
* To Try arima model
* To Try Sarima model

**5.3 Conclusion**

The United Kingdom is the highest product buying country by Quantity, Unit Price and Total\_amount The company have a good customer audience in United Kingdom so they can do further more ads and try new products to keep them engaging. Deeper Analysis must be performed by deeper analysis and model building

**REFERENCE**

* Machine Learning reference - <https://www.displayr.com/what-is-linear-regression/#:~:text=If%20we%20use%20advertising%20as,sales%20of%20168%20million%20Euros>.
* Data visualization reference -<https://plotly.com/python/getting-started/#:~:text=The%20plotly%20Python%20library%20is,3%2Ddimensional%20use%2Dcases>.
* Exploratory Data Analysis reference - <https://www.simplilearn.com/tutorials/data-analytics-tutorial/exploratory-data-analysis>