**SUMMER INTERNSHIP PROJECT**

**(SIP)**

**TITLE OF THE PROJECT**

Demand Forecasting and Inventory Management in Artificial Intelligence and Machine Learning

Submitted in the partial fulfilment for the award of the degree of Post Graduate Diploma in Management (PGDM) (AICTE)

**Batch Year**: 2021-23

**Submitted By**

Himanshu Chetnarayan Singh

(Roll No.133)

**Specialization**: Business Analytics

**Under The Guidance of**

Prof. Nikita Ramrakhiani

**Name of the Industry**:

Compunnel Digital

**Industry Mentor**:

Dr. Ravi Changale

**Note**: Summer Internship Project (SIP) Starts on: 6th of June

 Ends on: 29th of July

**DECLARATION**

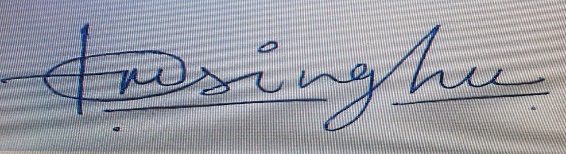
I hereby declare that this Internship Report submitted by me to the

VES Business School

is a bonafide work undertaken by me and it is not submitted to any other University or institution for the award of any degree diploma certificate or published any time before.

Name: Himanshu Chetnarayan Singh

Roll No. 133



 (Signature of the Student)

**CERTIFICATE**

This is to certify that project titled Demand Forecasting and Inventory Management in Artificial Intelligence and Machine Learning

successfully completed by during the 1st Year, in partial fulfilment of the PGDM recognised by AICTE for the academic year 2022 through Compunnel Digital Ltd.

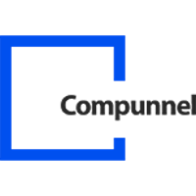
This project work is original and has not been submitted earlier for the award of any degree or diploma or associate ship of any other university/institution.

**Name:** Prof. Nikita Ramrakhiani

**Date:**

**(Signature of the Faculty mentor)**

**CERTIFICATE OF COMPLETION**



Ref No.: CD/ICL/2022/006

Himanshu Singh

Vivekanand Business School

Mumbai

[**www.compunnel.com**](http://www.compunnel.com/)

We wish him all the very best for his future endeavors.

For Compunnel Technology India Pvt. Ltd.

Sonal Arora

Vice President – Human Resources

**Subject: Internship Completion Letter**

This is to certify that Mr. Himanshu Singh, a student of Vivekanand Business School enrolled in

PGDM Business Analytics (2nd Semester), has successfully completed his 2 months internship with

Compunnel Technology Pvt. Ltd. Duration of his internship was from **June 6th, 2022, till July 29th,**

**2022.** He was part of **Team Alena** and has done project on **“Demand Forecasting and Inventory**

**Management”.**

During the tenure with us, we found Mr. Himanshu Singh, sincere, punctual, and inquisitive.



**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sr No.** | **Contents** | **Page No.** |
| **I** | **Internship Report** | **7** |
| **1** | **Executive Summary** | **8** |
| **2** | **Industry Analysis** | **9** |
| **2.1** | Major Players in The Industry |  |
| **2.2** | Market Share of Major Players |  |
| **2.3** | Industry Financials |  |
| **2.4** | Challenges Faced by the Industry |  |
| **2.5** | Government Regulations |  |
| **3** | **About Compunnel Digital** | **18** |
| **3.1** | Vision and Mission |  |
| **3.2** | Products and Services |  |
| **3.3** | Stakeholders and Financial Details of Company |  |
| **3.4** | Operational Details |  |
| **3.5** | SWOT Analysis of the Company |  |
| **4** | **On The Job Training** | **23** |
| **4.1** | Target Assigned |  |
| **4.2** | Target Achieved |  |
| **4.3** | Technical / Soft Skills Acquired |  |
| **5** | **Conclusion and Recommendations** | **34** |
| **6** | **Annexure** | **35** |
| **6.1** | References |  |
| **6.2** | Lists | 36 |
|  | **Industry Mentor Evaluation** | **37** |
|  | **Student’s Weekly Diary/Weekly Log** | **38** |

**I**

**Internship Report:**

**Demand forecasting and Inventory Management**

**1. EXECUTIVE SUMMARY**

This document describes all the work that has been done for the past 7 weeks and how it has been done. Interning for Compunnel Digital Ltd. Has been quite interesting. I was able to learn a number of new terms and concepts in Artificial Intelligence and Machine Learning. Businesses run on data and data has very little utility without analytics.

Demand forecasting using Machine Learning.

Demand forecasting is a technique for estimating probable demand for a product or service. It means making an estimation in the present for future occurring event.

Predicting Sales: Time Series Analysis & Forecasting with Python.

One of the crucial tasks for any retail store company is to analyse the performance of its stores. The main challenge faced by any retail store is predicting in advance the sales and inventory required at each store to avoid over- stocking and under –stocking .This helps the business to provide the best customer experience and avoid getting into losses.

What makes this analysis unique is that ‘Date’ is one of the important factors influencing sales and it acts as an independent variable. To put it simply, this is a time- series data i.e. a series of data points ordered in time.

In this age of information, the value of data as a business asset is essential.

How we use Python to forecast demand.

We will first pull and plot our data. Then using automated methods we estimate the best fit model parameters.

Applying ADF method to statistically test a time series.

Estimate the number of parameters for a SARIMA Model.

Test your models and begin to forecasts.

Interpret and export our forecasts.

Forecast works best when there is a lot of historical data.

Fitting your SARIMA Model.

We will be using automated methods to estimate the best fit model parameters we will now estimate the number of parameters for our SARIMA Model.

**2. INDUSTRIAL ANALYSIS**

Information Technology in India is a vast industry which comprises IT services, outsourcing and consulting. The IT industry accounted for 8% of India’s GDP in 2020. The IT Industry employed almost 2.8% million employees in FY 2021.

The IT BPM sector overall employs 4.5 million people as of March 2021. The Indian IT – BPM Sector industry has the highest employee attrition rate.

India’s IT services industry was born in Mumbai in 1967 with the establishment of TCS who in 1977 partnered which began India’s export of IT industries services.

Major Information Technology Hubs

1) Bangalore- is a global technology hub and is India’s biggest tech hub. Bangalore accounted for 38% of total IT exports from India ; employing 10 lakh people directly and 30 lakh indirectly.

2) Hyderabad – Notable tech companies include – Microsoft, Apple, Amazon, Google, Accenture, AMD, Deloitte, Cape Gemini, Inter TCS, HCL Technologies , Oracle, Corporation, Dell, Cognizant, J.P Morgan, UBS.

3) Chennai – India’s third largest exporter of IT after Bangalore and Hyderabad.

India’s IT market, including e-commerce, is worth a total of US $ 235 Billion. The hardware industry was left behind. In 2021, the market grew by more than 2% on an annualized basis. The sector contributes 51% of export services and 9% of the country’s GDP.

**2.1 Major players in the industry:**

The major players in the data analytics industry are as follows:

* Microsoft
* AWS
* Alteryx
* IBM

Microsoft – is one of the major players in enterprise BI and Analytics. The company’s flagship platform, Power BI, is cloud Power- BI is a unique tool which enables users to do data preparation, data discovery, and dashboards with the same design tool. The platform integrates with Excel and Office 365, and has a very active user community as well.

AWS - Amazon Quick Sight is a serverless and embeddable business intelligence service for the cloud featuring built-in machine learning. The product lets you create and publish interactive BI dashboards that can be queried using natural language. It can automatically scale to thousands of users without any infrastructure. QuickSight also touts pay-per-session pricing so customers only pay when users access dashboards or reports. Dashboards can be accessed from any device.

Alteryx is a self-service data analytics software company that specializes in data preparation and data blending. Alteryx Analytics allows users to organize, clean, and analyze data in a repeatable workflow. Business analysts find this tool particularly useful for connecting to and cleansing data from data warehouses, cloud applications, spreadsheets and other sources. The platform features tools to run a variety of analytic jobs (predictive, statistical, and spatial) inside a single interface.

IBM - The collection and analysis of data is fundamental to business analytics. In the 1990s, computer programming skills were needed to identify the available data in an organization, connect to myriad data sources, convert raw data it into a usable form, and generate periodic reports.

In the next decade, innovative companies began to analyze business data using programming languages like Python and R. This made it possible to see what was happening within their organization, and perhaps even why it was happening.

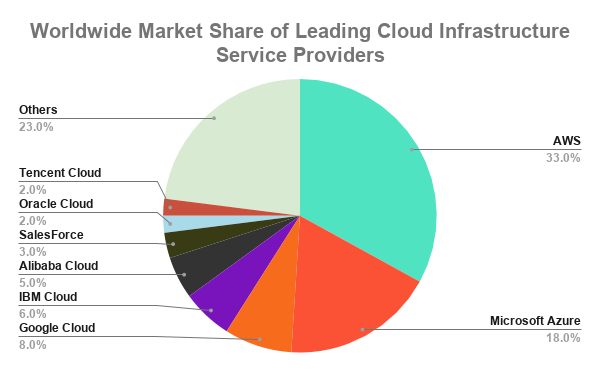
Over the last decade, self-service business analytics programs, AI, and cloud data management software have made it possible for almost anyone (regardless of skill level) to analyze and visualize trends in real time, pinpoint business problems, and make informed business decisions. Business analysis is no longer limited to the domain of computer science professionals. It's a common tool for finance, manufacturing, healthcare, sales, marketing, supply chain, and operations, among others. If you are looking for insights from your data, you need business analytics.

Today, business analytics skills are taught at leading business schools and are an important part of many degree programs. Business students learn to apply their skills in real-world situations like business operations, where they identify key metrics and take a data-driven approach to problem solving.

**2.2 Market Share of Major Players**

AWS leads in the share of major players by about 33% of the market. While

Microsoft stands at second in cloud computing worldwide share by about18%.



Cloud Computing Market Share Breakdown.

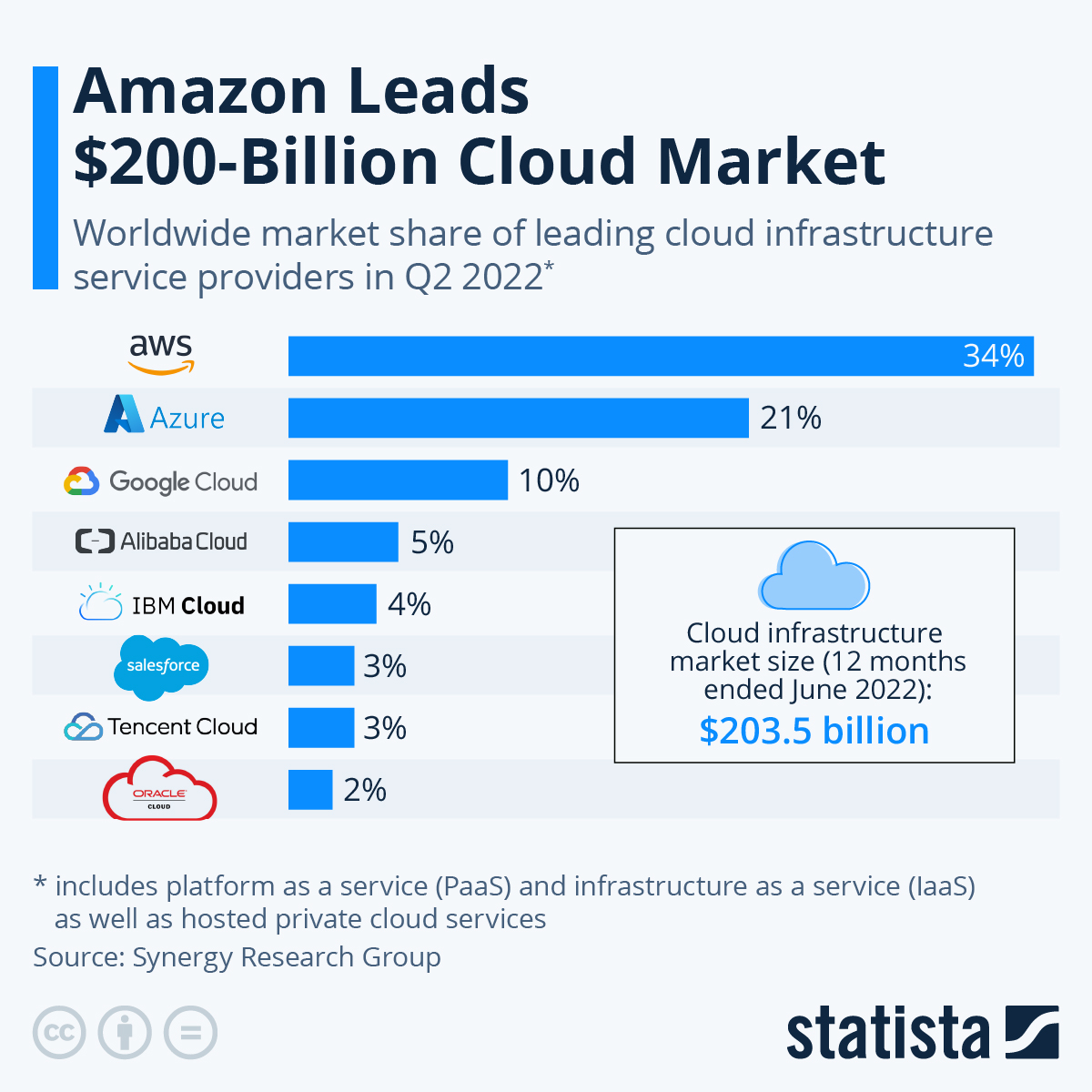
2.3 **Industry Financials**

The emergence cloud computing has undergone several phases before making a go-go in its present form. Cloud computing is a general term for anything that involves delivering hosted services over the internet. These services are divided into three main categories or types of cloud computing:

Infrastructure as a service (laaS), platform as a service (PaaS) and software as a

Service (SaaS).

A Cloud can be public or private. A Public cloud sells services to anyone on the internet. A private cloud is a proprietary network or a data centre that supplies hosted services to a limited number of people, with certain access to computing is to provide easy, scalable access to computing resources and IT services.



**2.4 Challenges faced by the Industry**

1**)** Collecting meaningful data:

There is data of every aspect of business, which more than often overwhelms employees. The plethora of data from different channels make it difficult for employees to drill down and determine the critical insights.

And they end up analyzing the data that’s readily available and not the one that truly adds value to the business.

2) Selecting the right tool

The second most common challenge in data analytics arises with the vast number of tools available in the market.

Is Cassandra the best for storing data or Mongo DB?

Is Rapid Miner a better option for data analytics, or should you opt for Microsoft’s Power BI?

If these questions are not answered correctly, the chances are that you will end up investing your time, effort and money in inappropriate tools.

Solution: The best option is to seek professional help. Experts who have experience working on different tools can help you to select the right one.

You can also start using the trial versions of the tools to check out the features by yourself.

3. Quality of Data collected-

Nothing can be more dangerous in data analytics than incorrect data. If the input quality of data is flawed and erroneous, the output can never be reliable.

One of the primary reasons behind inaccurate data is errors made during data entry i.e. manual errors.

Another reason for poor quality of data is the disparity in data. Suppose your data operator makes changes in one system and forgets to make the exact change in others; it will create asymmetric data.

4. Building a data culture among employees:

According to a research, the biggest hindrance a data driven company lies in an organization’s culture and not technologies. Only a meagre 9.1% of executives have pointed out technology as a challenge in the path of data analysis.

Many times, through top level understand the importance of data analysis, they do not extend the desired support to the employees. Constant pressure and lack of support from the top and lower- level employees are among the most significant data analytics challenges.

Solution: As Albert Einstein has said “The world cannot be changed without changing our thinking.” you need to build a culture that understands data and supports it. To make data analysis a success at every level, educate your employees and help them to upgrade themselves.

5. Data Security:

Once businesses realize the importance of Big data, they start focusing on storing, understanding and analyzing it. They tend to overlook the potential risks that come with the privacy and security of the large number of datasets collected.

Security of your company’s data is necessity and one of the scariest challenges in data analytics. Unprotected data sources can become an easy entry point for hackers.

Solution:

The moment you decide to use data analytics in your business, make sure to take care of security issues. Some steps that you can implement are:

* Hire cybersecurity professionals to protect your data.
* Conduct corporate training programs on big data for your managers and business owners.
* Use Big data analytics tools.
* Control access rights
* Encrypt data with secured login credentials.

6. Data Visualization:

Data Analytics holds no meaning for you or your stakeholders until the numbers tell a story. After all, the time, money and effort you invest in collecting and securing the data are to help you make informed decisions and meet your ROIs. So data visualization is very critical in data analytics and challenging too.

Solution: Use data visualization tools like Power BI, Tableau, Google Data Studio, which are easy to learn and have wide range of features. These tools have drag- and drop features and can also connect to various data sources. They come with intuitive graphs and charts, thus helping you to visualize your data.

Final thoughts on the business data challenges

When we are aware of the problems, it becomes easier for us to deal with them. Now that you know the data analytics issues faced by businesses and their solutions, you can start implementing them in a more structured form.

And in the end, none of the challenges in data analytics are critical enough to stop you from utilizing the benefits of big data!

**2.5 Government Regulations**

Cloud computing allows users to access and work with data stored on

remote networks. The growth of high speed Internet access and advances

in data storage and retrieval technologies has led to cloud computing

becoming the service model of choice for both business critical services,

like banking and payments.

a.) Standards apply , but no license is needed to provide cloud services. Cloud services can be provided in India on a cross- border basis. There is also no restriction or foreign investment for setting up cloud operations in India.

b.) Data privacy and security – Privacy can be a major bottleneck in a cloud environment. Remember that the more valuable the data, the more likely it faces cyber – attacks illegal access attempts.

The right to privacy is widely considered one of the basic human rights and the same is stated under Article 12 of the 1948 Universal Declaration of Human Rights.

a) It was held that privacy concerns in this day and age of technology can arise from both the state as well as non-state entities and such, a claim of violation of privacy lies against both of them.

b) The court also held that informational privacy in the age of the internet is not an absolute right and when an individual exercises his rights to control over his data, it may lead to the violation of his privacy to a considerable extent.

**3**. **ABOUT THE COMPANY**

Compunnel Digital has been identified among the most promising Data Analytics companies in 2021. Compunnel Digital has been defined an accelerated approach to help clients achieve their desired future state for their data projects. It includes agile approaches to data design, modelling with industry domain, dashboards and reports that can be implemented swiftly.  
Giving insights on the data analytics practice at Compunnel Digital, CEO,

Siraj Mallick.

The company’s focus is clients first.

The company is driven to create and deliver business value by orchestrating a seamless experience for our clients. Its aim is to help optimize our customer experience, improve digital business operations, and help us optimize our customer experience, improve digital business operations, and help our teams adapt and learn everything they need to succeed in today’s digital economy.

Curious- In this rapidly changing world, there are two types of people, the curious and the resistors.

Applying advances digital tools and techniques to drive the journey from creation to execution. It stays connected through advanced technology with our stakeholders and internal teams to deliver an uninterrupted, seamless and personalized experience. The company creates and leads the best practices in digital transformation with native and third party tools and technologies.

Compunnel Digital has defined an accelerated approach to help clients achieve their desired future state for their data projects. It includes agile approaches to data design, modelling with domain driven data blue prints, and pre- defined dashboards and reports that can customized and implementation rapidly.

Data and Analytics solutions covering key technologies like predictive and prescriptive analytics, decision sciences, AI/ML Visualization, provides organizations with 360 degree view of the customer, enriching customer experience, driving operational excellence, and enabling them to take faster and informed business decisions.

**3.2 Products and services of the company**

Services of Compunnel Digital-

As Technology is evolving, businesses today demand increased agility, minimised disruption, low operational cost, and faster delivery. To facilitate this IT stakeholders are struggling with their age- old legacy applications to adopt unified user experience upgraded technology platforms, business agility, mobile, social and cloud advantages.

Services of Compunnel Digital are as follows-

1) New Services- AI/ML- represents an important evolution in computer science and data processing that is quickly transforming a vast array of industries.

As Businesses and other organizations undergo digital transformation, they are faced with a growing tsunami of data that is at once incredibly valuable and increasingly burdensome to collect, process and analyze. New tools and methodologies are needed to manage the vast quantity of data being collected, to mine it for insights and to act on those insights when they are discovered.

2) Digital Application- A digital application is simply a computer program created to carry out or facilitate a task on a computing device. It should be noted that although all applications are programs, not all programs are applications. Digital application means an Internet- connected software application that a person uses to obtain or provide a remote service.

3) Cloud Transformation – the process of transitioning a business’s data, apps, and software to the cloud. Although this sounds straightforward, the process of transitioning itself can become complicated very quickly. One common area of confusion is how cloud transformation differs from cloud migration.

4) E- Commerce – Here at Compunnel Digital e- commerce experts works for utilizing their domain and industry expertise to assess, evaluate, and industry expertise to assess, evaluate , and deliberate in designing customized solutions for the success of your etail business. They create custom codebases that integrate with market- leading product management platforms to achieve streamlined global delivery and flawless customer experience.

5) Enterprise Data Analytics - Compunnel's digital experience in data analytics leverages all of the next generation of

Disruptive data technologies and can deliver multiple benefits to your business. It helps you

conduct situational analysis, build a strong data foundation, and capitalize on data

commercialization opportunities to drive business performance.

7) Robotics Process Automation- Compunnel Digital offers comprehensive RPA services that enable businesses to automate a

Wide variety of organizational activities. His AI and ML expertise helps companies unlock

the cognitive power of his RPA-enabled bots to deliver intelligent operations.

**3.3 SWOT Analysis**

**Strengths**

1) Organisation’s workplace technology has been updated within the past year.

2) Compunnel has a highly skilled and innovative IT Team.

**Weaknesses**

Weaknesses can also be assets, resources or attributes your company is missing. For instance:

1) Employees have to navigate between too many different workplace apps throughout the day.

2) The company doesn’t have an easy way to manage conference room scheduling.

**Opportunities**

While strengths and weaknesses are considered internal factors, opportunities and threats are external elements that can influence the success of the change or initiative. Opportunities are factors your company can leverage to gain an advantage over competitors, such as market shifts, new cultural trends, changes to the industry or upcoming tech launches.

Within the context of a digital transformation, your greatest opportunities might be:

* Moving to cloud-based, software-as-a-service (SaaS) applications
* Consolidating multiple applications into one to simplify our technology stack

**Threats**

Threats include anything outside the organization that could negatively impact your ability to implement a digital transformation strategy. These are the elements that are out of your control. Examples of threats are new competitors entering the marketplace, service providers increasing prices, changes in the economy and [new government regulations](https://www.iofficecorp.com/blog/how-the-new-fasb-lease-accounting-standard-impacts-other-departments).

Examples might include:

* [New data security regulations](https://www.forbes.com/sites/forbestechcouncil/2019/03/04/will-data-privacy-regulations-upend-your-information-security-program/#1843c74545b8) could impact our IT department
* A greater reliance on contractors in the gig economy will make it more difficult for us to predict our future office space needs

**4.**

**On the Job Training**

**Task 1: Machine Learning Landscape, Types of ML, Under fitting, Over fitting, basics of Hyper parameter Tuning**.

Objective:

To get basic knowledge of Machine Learning

What is Machine Learning?

Machine Learning is a subset of AI that uses computer algorithms to analyze data and make intelligent decisions based on what it has learned, without being explicitly programmed. Machine Learning algorithms use historical data as input to predict new output values.

Machine Learning is important because it gives enterprises a view of trends in customer behaviour and business operational patterns, as well as supports the development of new products. Many of today’s leading companies, such as Facebook, Google and Uber, make machine learning has become a significant competitive diffrentiator for many companies.

TYPES OF MACHINE LEARNING

1) Supervised Learning- In this type of machine learning, data scientists supply algorithms with labelled training data and define the variables they want the algorithm to assess for correlations. Both the input and the output of the algorithm is specified.

2) Unsupervised Learning – This type of machine learning involves algorithms that train on unlabelled data. The algorithm scans through data sets looking for any meaningful connection. The data that algorithms train on as well as the predictions or recommendations they output are predetermined.

3) Semi- Supervised Learning – This approach to machine learning involves a mix of the two preceding types. Data scientists may feed an algorithm mostly labelled training data, but the model is free to explore data on its own and develop its own understanding of the data set.

4) Reinforcement Learning- Data Scientists typically use reinforcement learning to teach a machine to complete a multi-step process for which there are clearly defined rules. Data Scientists program an algorithm to complete a task and give it positive or negative cues as it works out how to complete a task. But for the most part, the algorithm decides on its own what steps to take along the way.

CONCEPT OF BIAS AND VARIANCE

An important theoretical result of statistics and Machine Learning is the fact that a model’s generalization error can be expressed as the sum of the three very different errors.

1) BIAS – This part of the generalization error is due to wrong assumptions, such as assuming that the data is linear when it is actually quadratic. A high bias model is most likely to underfit the data.

2) VARIANCE – This part is due to the model’s excessive sensitivity to small variations in the training data. A model with many degrees of freedom (such as high degree polynomial model) is likely to have high variance and thus overfit the training data.

What is Underfitting?

Underfitting, on the other hand, occurs when the model is not complex enough.

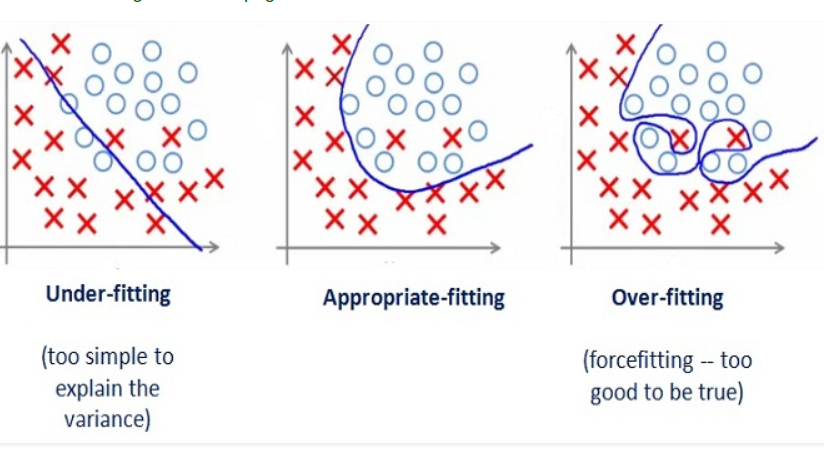
While there can be different causes, it is often the case that the model is not given enough data to learn from. As a result, the model is not able to learn the generalizable patterns in the data and ends up performing poorly on the both training data set and the new data points.

An underfitted model is “too simple”, with too few features and insufficient data to build an effective model. It shows a high bias and low variance. Adding more features to a model that is too simple can help to limit bias.

What is Overfitting?

Overfitting is one of the biggest challenges that developers face.

Overfitting occurs when a model is too complex for the data it is too complex for the data it is supposed to be modelling. This can happen for a variety of reasons , but one of the most common is that the model is simply trying to learn too much from the data. When this occurs, the model ends up memorizing the data set instead of learning generalizable patterns .As a result, the model performs well on the training dataset but does not generalize well to new data.



Perfect fit— This is the sweet spot, this is highly desirable and occurs when our model can produce good accuracies on both testing and validation data. This can happen with correct set of all hyperparameters. Relating it to the above example this happens when our child is able to learn as well as score better marks in exams.

Task 2: Data Pre- Processing and cleaning, feature scaling, handling missing value, one hot encoding

1) Data Pre- Processing- a component of data preparation, describes any type of processing performed on raw data to prepare it for another data processing procedure. It has traditionally been an important preliminary step for the data mining process.

Importance of Data Cleaning

Data Cleaning implies the way toward distinguishing the erroneous, deficient, mistaken, immaterial or missing piece of the data and afterwards changing, supplanting or erasing them as per the need. Data cleaning is considered an essential component of fundamental data science.

Data is the most important thing for Analytics and Machine learning. In processing or Business data is required all over the place. With regards to genuine data, it isn’t impossible that data may contain fragmented, conflicting or missing qualities. Assuming the data is defiled, it might prevent the interaction or give erroneous outcomes.

2) Feature scaling is the process of normalising the range of features in a dataset.

Real-world datasets often contain features that are varying in degrees of magnitude, range and units. Therefore, in order for machine learning models to interpret these features on the same scale, we need to perform feature scaling.

3) Handling Missing Values

The real-world data often has a lot of missing values. The cause of missing values can be data corruption or failure to record data. The handling of missing data is very important during the pre-processing of the dataset as many machine learning algorithms do not support missing values.

Missing values can be handled by deleting the rows or columns having null values. If columns have more than half of the rows as null then the entire column can be dropped. The rows which are having one or more columns values as null can also be dropped.

**Pros:**

* A model trained with the removal of all missing values creates a robust model.

**Cons:**

* Loss of a lot of information.
* Works poorly if the percentage of missing values is excessive in comparison to the complete dataset.

4) One Hot Encoding-

One hot encoding is useful for data that has no relationship to each other. Machine learning algorithms treat the order of numbers as an attribute of significance. In other words, they will read a higher number as better as or more important than a lower number.

While this is helpful for some ordinal situations, some input data **does not have any ranking**for category values, and this can lead to issues with predictions and poor performance. That’s when one hot encoding saves the day.

One hot encoding makes our training data more useful and expressive, and it can be rescaled easily. By using numeric values, we more easily determine a probability for our values. In particular, one hot encoding is used for our output values, since it provides more nuanced predictions than single labels.

Task 3: Linear Regression, Ridge Regression, Lasso Regression, Gradient Descent

What is Linear Regression- Linear Regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things

(1) Does a set of predictor variables do a good job in predicting an outcome (dependent) variable?

(2) These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of regression equation with one dependent and one independent variable is defined by the formula y = c + bx, where y is estimated dependent variable score, c = constant, b= regression coefficient and x is score on the independent variable.

Naming the variables. There are many names for a regression’s dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, predictor variables, or regressor variables.

Three major uses for regression analysis are –

(1) Determining the strength of predictors

(2) Forecasting an effect

(3) Trend Forecasting

Logistic Regression- Logistic Regression is a classification algorithm. It is used to predict a binary outcome based on a set of independent variables.

Logistic regression is used to calculate the probability of a binary event occurring, and to deal with issues of classification. For Example- predicting if an incoming email is spam or not spam. OR predicting if a credit card transaction is fraudulent or not fraudulent. In a medical context, logistic regression may be used to predict whether a tumour is benign or malignant. In

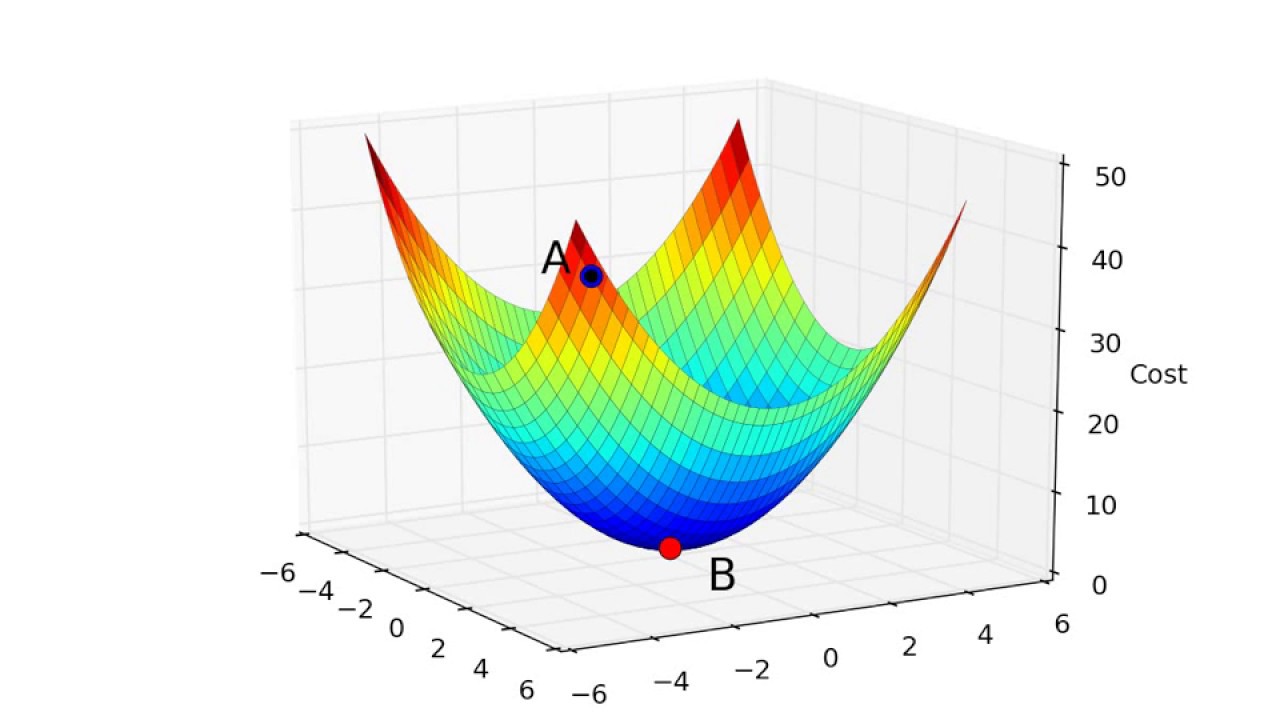
Marketing it may be used to predict if a given user (or group of users) will buy a certain product or not. An online education company might use logistic regression to predict whether a student will complete their course or not.

As you can see, logistic regression is used to predict the likelihood of all kinds of “yes” or “no” outcomes. By predicting such outcomes, logistic regression helps data analysts to make informed decisions. In the grand scheme of things, this helps to both minimize the risk of loss and to optimize spending in order to maximize profits.

Ridge Regression – modifies the cost function by adding a penalty equal to the square of the size of coefficients. Lasso regression can lessen over- fitting while also assisting in feature selection.

Gradient Descent – Gradient Descent is a generic optimization algorithm capable of finding optimal solutions to a wide range of problems. The general idea of gradient descent is to make parameters iteratively in order to minimize a cost function.

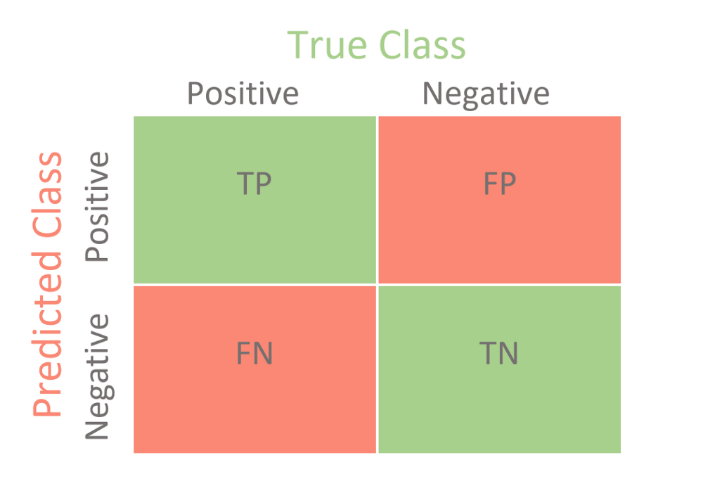
In fact the cost function has the shape of a bowl, but it can be an elongated bowl if the features have very different scales.



Simple Gradient Descent

Task 4: Confusion Matrix and Cross Validation

Confusion Matrix – Confusion matrix is a very popular measure and while solving classification problems. It can be applied to binary classification as well as for multiclass classification problem. An example of a confusion matrix for binary classification is:



Confusion matrices represents counts from predicted and actual values. The output TN stands for True Negative which shows number of negative examples classified accurately. Similarly TP stands for True Positive which indicates the number of positive examples classified accurately. The term FP shows False Positive values i.e. the number of actual negative examples classified as positive. FN means False Negative value which is the number of actual positive examples classified as negative. One of the most commonly used metrics while performing classification is accuracy. The accuracy of a model is calculated through the given formula:

Accuracy = TN + TP

TN + TP + FP + FN

The most frequently used performance metrics for classification according to these values are accuracy, precision, sensitivity, specificity and F-score

Precision = TP/ TP+ FP

Sensitivity = TP/ TP + FN

Specificity = TN/ TN + FP

F- Score = 2 \* P\*Sn/ P+ Sn

Cross Validation – Cross Validation is a technique for evaluating ML models for training several ML models on subsets of the available input data and evaluating them on the complementary subset of the data. Use cross validation to detect over fitting, i.e. failing to generalize a pattern.

One way to overcome this problem is not to use the entire data set when training a learner. Some of the data is removed before training begins. Then when training is done the data was removed can be used to test the performance of the learned model. This is the basic idea for a whole class of model evaluation methods called cross validation.

Steps involved in Cross Validation

1) Divide the data set into training and test sets at random.

2) Fit model working out on training set.

3) To test the model on test set.

4) Calculate and record the fit statistic using the test data.

5) Numerous times, repeat 1 through 4 , then average the outcomes.

Task 5: Support Vector Machines, Data Leakage and API Basics

Support vector machines (SVMs) are powerful yet flexible supervised machine learning algorithms which are used both for classification and regression. But generally, they are used in classification problems. In 1960s, SVMs were first introduced but later they got refined in 1990. SVMs have their unique way of implementation as compared to other machine learning algorithms. Lately, they are extremely popular because of their ability to handle multiple continuous and categorical variables.

## Working of SVM

An SVM model is basically a representation of different classes in a hyper plane in multidimensional space. The hyper plane will be generated in an iterative manner by SVM so that the error can be minimized. The goal of SVM is to divide the datasets into classes to find a maximum marginal hyper plane (MMH).

The followings are important concepts in SVM −

* **Support Vectors** − Data points that are closest to the hyper plane is called support vectors. Separating line will be defined with the help of these data points.
* **Hyper plane** − As we can see in the above diagram, it is a decision plane or space which is divided between a set of objects having different classes.
* **Margin** − It may be defined as the gap between two lines on the closet data points of different classes. It can be calculated as the perpendicular distance from the line to the support vectors. Large margin is considered as a good margin and small margin is considered as a bad margin.

Data leaking occurs when the model is built using information from sources other than the training dataset. The expected performance of the mode that is being developed may be invalidated if the model learns or understands something that it otherwise would not be able to.

You can use the two following strategies to lessen data leaks while developing predictive models:

• Prepare your cross-validation folds' data.

• Keep a validation dataset on hand for a last-minute check of the models you've built.

API Basics:

An application programming interface or API is a server that you can programmatically

access to retrieve and deliver data. Retrieving data is most commonly done through an API.

To get data from an API, you need to make a request to the API. Queries are frequently used on the Internet.

Task 6: Decision Tree, Entropy, and Information gain, Gini Impurity and Post Pruning

Decision Tree – Decision tree is the most powerful and popular tool for classification and prediction. A decision tree is a flowchart tree like structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.



Entropy – Entropy is a metric for the unpredictability of a random variable and it describes the impurity of a random sample of samples. Entropy causes an increase in the amount of information content.

Information gain is the reduction in entropy or surprise by transforming a dataset and is calculated by comparing the entropy of the dataset before and after a transformation.

Gini Impurity is a measurement of the likelihood of an incorrect classification of a new instance of a random variable, if that new instance were randomly classifiedaccording to the distribution of class labels from the data set.

Gini impurity is lower bounded by 0, with 0 occurring if the data set contains only one class.

Post Pruning - This approach is used after a decision tree has been created. This approach is used when

a decision tree's depth is anticipated to be very high and the model over fits.

It is also known as reverse pruning. This technique is applied when a decision tree has

Infinitely extended. Here, the maximum depth and minimum sample split branches of the

Decision tree will be controlled via cost complexity pruning.

Task 7: Random Forest Algorithm, Bagging, Boosting, Voting Classifier.

Random Forest is a supervised Machine Learning Algorithm that is widely used in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

One of the most important features of the random Forest algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.

Bagging – It creates a different training subset from sample training data with replacement and the final output is based on majority voting. For Example, Random Forest.

Boosting – It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For Example, ADA Boost and XGG Boost.

A Voting Classifier is a machine learning model that trains on an ensemble of numerous models and predicts an output (class) based on their highest probability of chosen class as the output.  
It simply aggregates the findings of each classifier passed into Voting Classifier and predicts the output class based on the highest majority of voting. The idea is instead of creating separate dedicated models and finding the accuracy for each them, we create a single model which trains by these models and predicts output based on their combined majority of voting for each output class.

Voting Classifier supports two types of votings.

1. **Hard Voting:**In hard voting, the predicted output class is a class with the highest majority of votes i.e the class which had the highest probability of being predicted by each of the classifiers. Suppose three classifiers predicted the output class *(A, A, B)*, so here the majority predicted *A* as output. Hence *A* will be the final prediction.
2. **Soft Voting:**In soft voting, the output class is the prediction based on the average of probability given to that class. Suppose given some input to three models, the prediction probability for class *A = (0.30, 0.47, 0.53)* and *B = (0.20, 0.32, 0.40)*. So the average for class *A is 0.4333* and *B is 0.3067*, the winner is clearly class *A* because it had the highest probability averaged by each classifier.

Task 8: Implementation

Objective: To predict Inventory after 3 months.

We were provided data to test our analytical and data visualisation theories.

Task 8 : Implementation

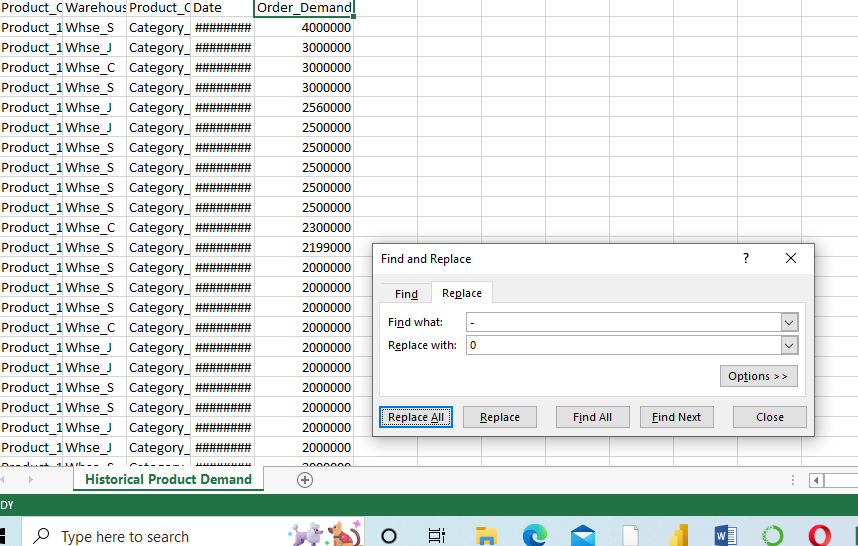
Data Analysis:

1. In MS Excel, we opened the file, and we changed every negative value to a 0 (i.e.

i. Using Ctrl + F, we first selected the "Order Demand" Column.

ii. Next, we inserted the Negative (-) sign in the "Find what" dialogue box and the number

"0" in the "Replace with" field.



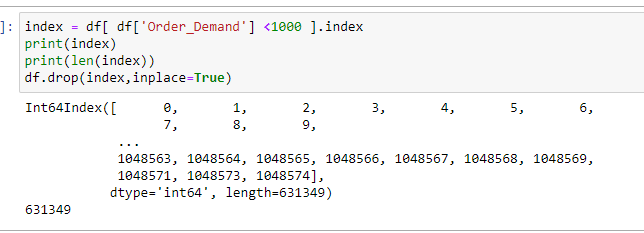
iii. 10,462 substitutions were made after selecting "Replace All," and all of the negative

numbers were changed to zero.

2. Import Numpy, Pandas, Matplotlib, and Seaborn, along with any other required libraries.

3. Use the "pd read.csv" function in Jupyter Notebook to read the Excel file.

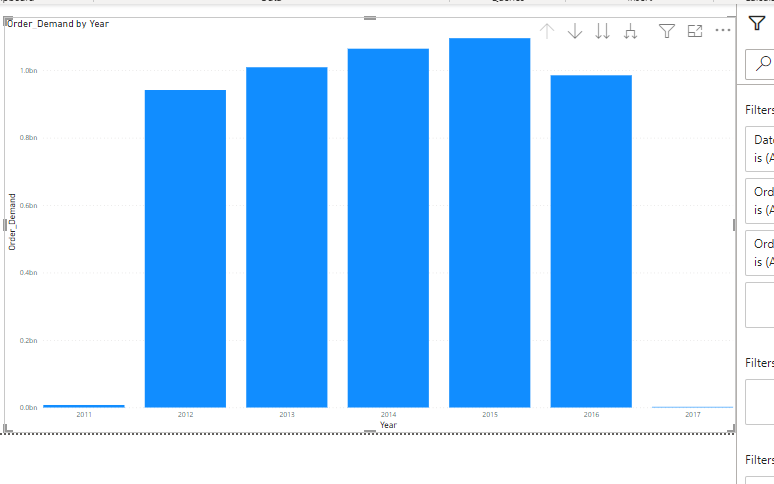
4. We opted to remove the numbers below 1000 because Order Demand cannot be less than 1000.



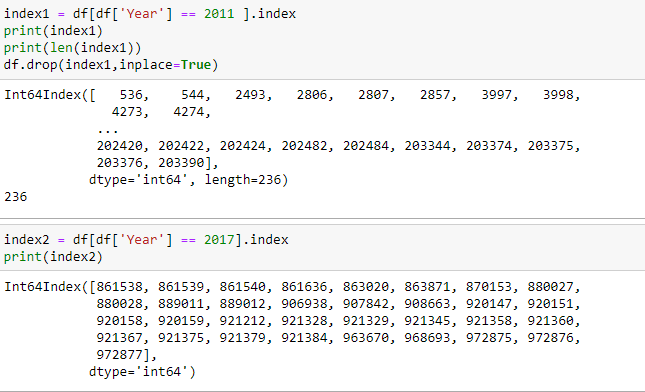
5. We used Power BI to better visualise the data for the exploratory investigation.

We have placed the year on the X-axis of a column chart and the count of orders/demand on

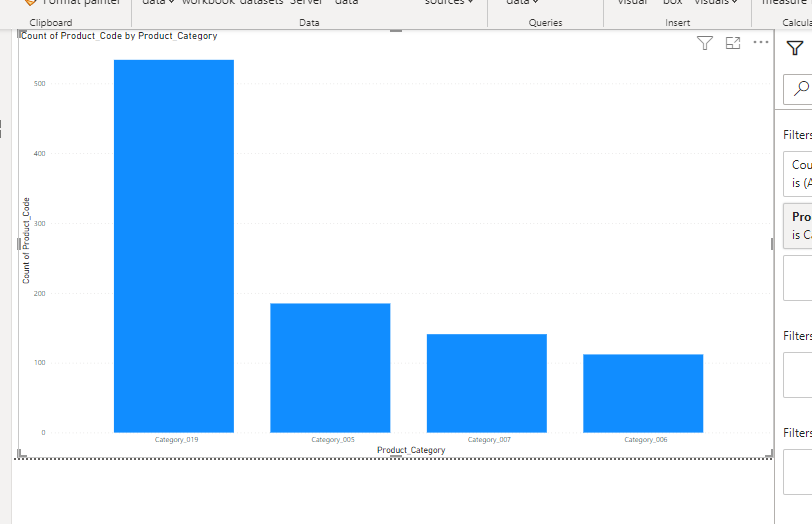
the Y-axis.



After visualizing, it is seen that the Year 2011 & 2017 have contributed the least for Order Demand, so we decide to drop the years 2011 & 2017.



6. Further analysis is done on Power BI,



With the lowest product code count of 112 in the 6th Product Category, the 19th Product

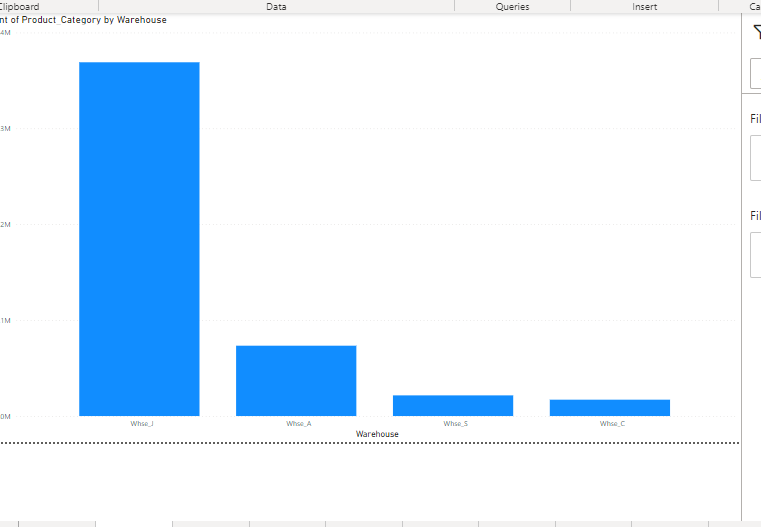
Category had a greater product code count of 376.79%.

At 534, category 19 continues to have the most product codes, followed by categories 5, 7,

and finally 6.

When taking into account all 4 Product Categories, the number of Product Codes ranged from

122 to 534.

ii) 

Count of Product Category by Warehouse.

There are four warehouses in total: A, C, J, and S.

With 3, 69, 078 different product categories, Warehouse J had the biggest number of items.

Warehouse D's lowest product category count of 17,067 was 52% higher than Warehouse

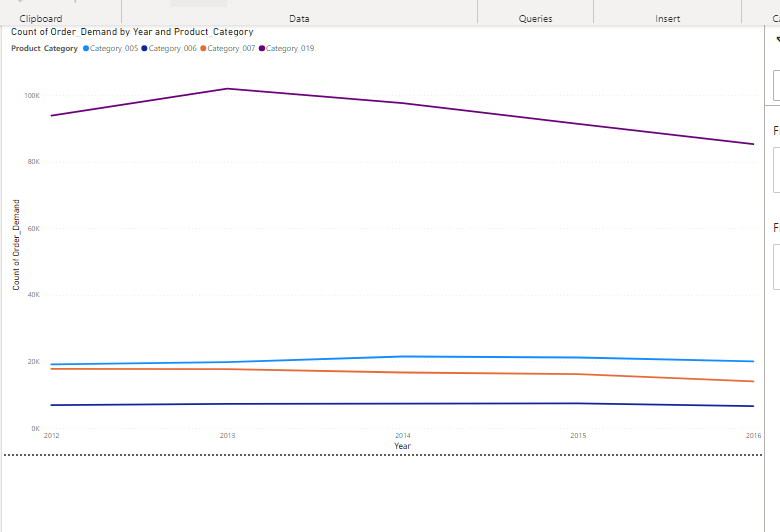
C's.

Warehouse J, Warehouse A, Warehouse S, and Warehouse C were in order of highest

Product category count.

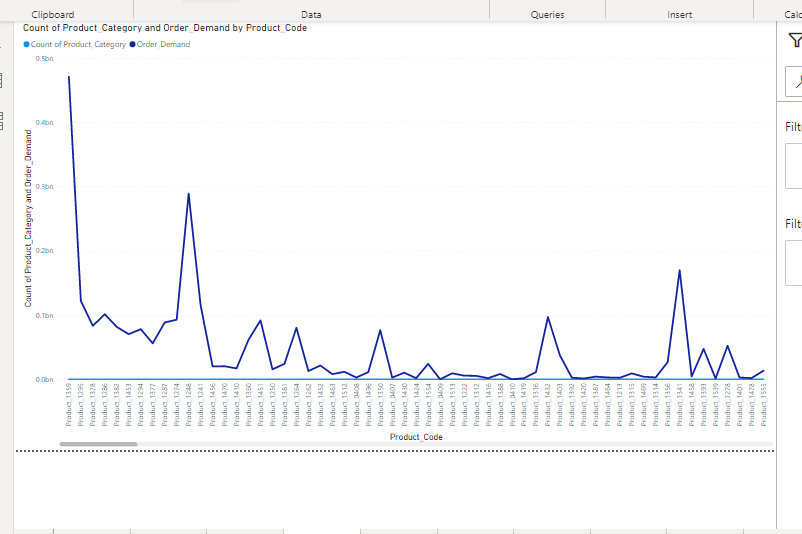
Warehouse J contributed 76.72% of the total count for the Product Category.

In total, the 4 Warehouses' product categories ranged from 17,067 to 3,69,078.

iii) 

The most intriguing recent trend in Product Category was Category 019, which started

moving downward in 2012 and fell by 9.13% (8,571) in four years.



When compared to Product Code 1525, which had the lowest Product Category count of 9,

Product Code 1359 had the highest Product Category count of 16,936 and was greater at a

rate of 1,88,077.78%.

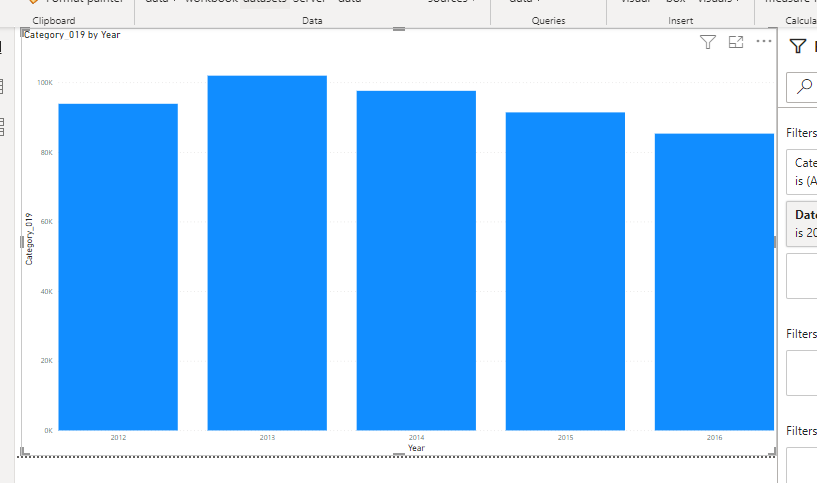
The sum of Order Demand and the count of Product Category have a negative correlation

with one another.

Product Code 1359 represented a rate of 3.52% for the total count of Product Category.

Order Demand and Product Category count deviated the highest when the Product Code was

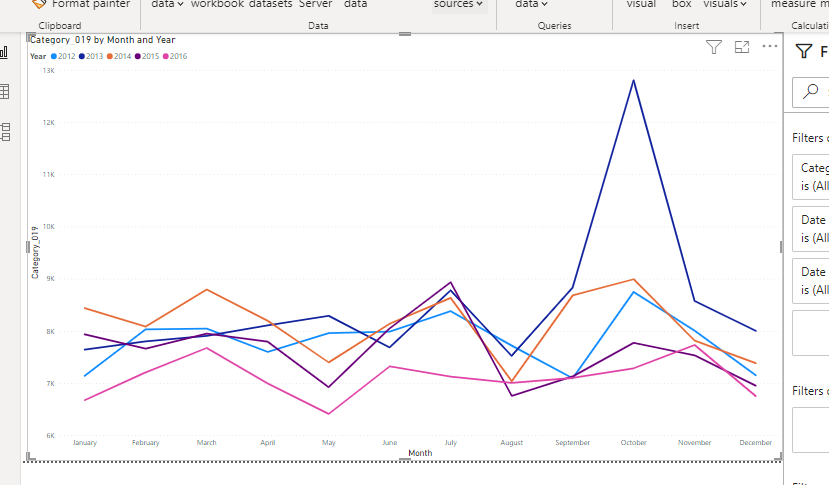
1359 and when it was 47,06,93,064 higher than the count of the Product Category.

v.) 

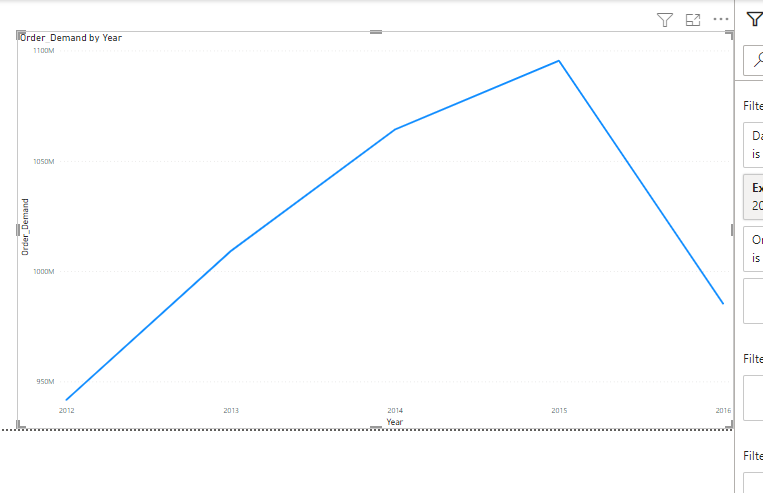
Category\_019 trended down, resulting in a 9.13% decrease between 2012 and 2016.

Category\_019 started trending down on 2012, falling by 9.13% (8571) in 4 years.

Category\_019 dropped from 93864 to 85293 during its steepest decline between 2012 and 2016

vi) 

October 2013 represented 2.72 percent of Category 019.

vii) 

Order Demand increased by 4.63% between 2012 and 2016 due to an upward trend.

Order Demand began to trend upward in 2012, increasing by 4.63% (43625222) during the

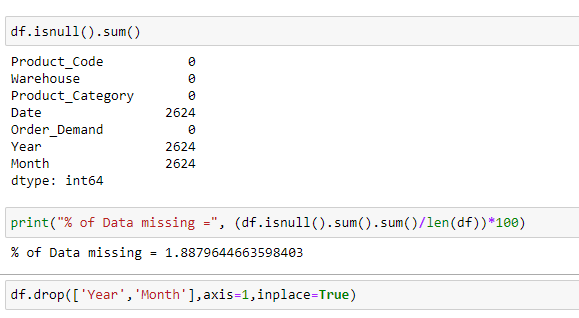
following four years.

Between 2012 and 2016, Order Demand had a spike from 941742907 to 985368129 during

Its highest ascent.

7.) The df.isnull ().sum () tool allowed us to locate 1% of the dataset's null values.

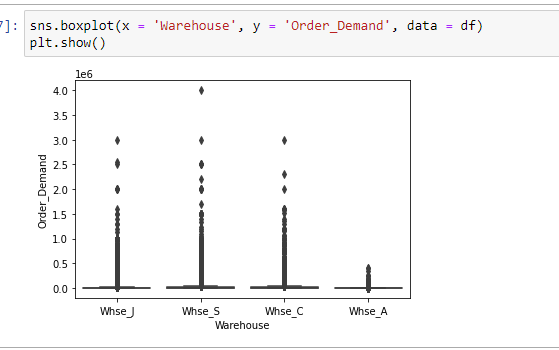
We have used the below command to remove the missing values:

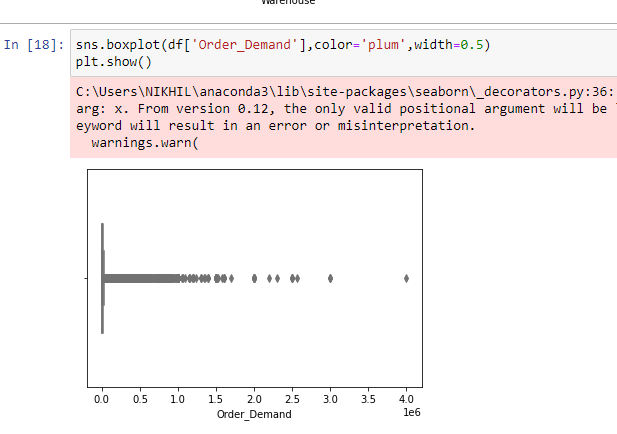


The command mentioned above allowed us to remove all instances of null values:

8. We have employed the sns.boxplot () tool to find outliers. We learned that the dataset

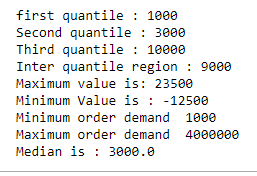
Contained outliers as a result.





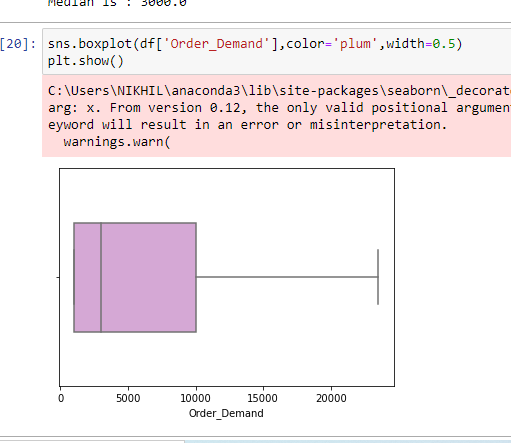
We therefore utilised the Quantile function to eliminate the outliers:





Data points below the lower limit were replaced with the lower limit, while those above the

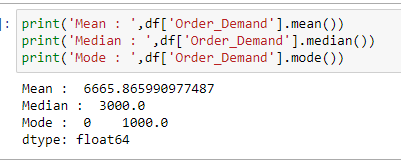
Upper limit were replaced with the upper limit.

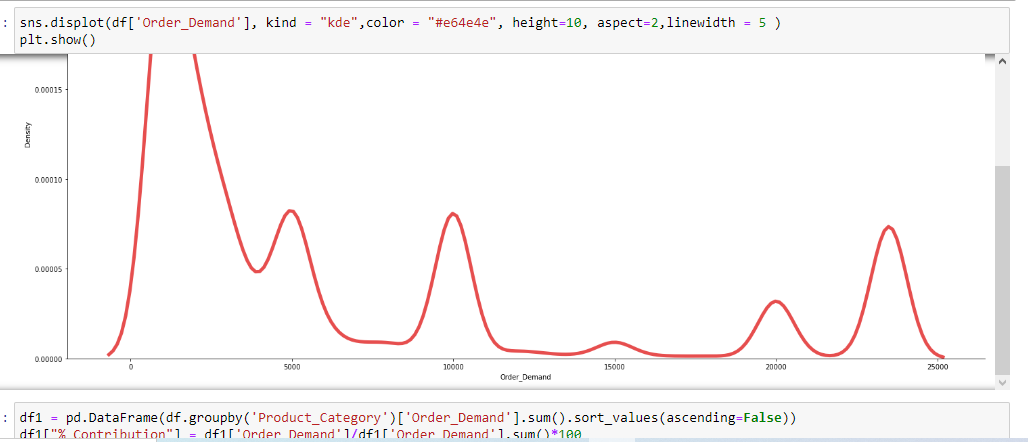


As you can see, the Quantile function was used to handle the outliers.

Since we eliminated values below 1000, the Lower Limit and Quantile1 (q1) are identical.

The Mode<Median<Mean. The statistical graph is therefore right-skewed.





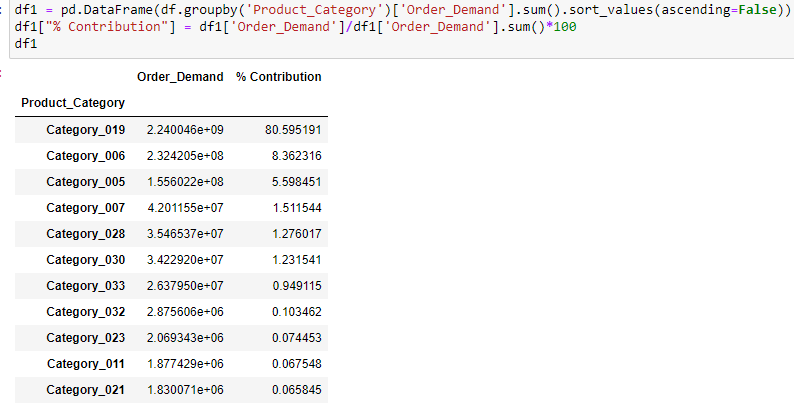
10. On Order Demand & Product Category, the Python GroupBy Clause has been

implemented. The Order Demand is thus divided into Product Categories. The percentage

contributions for each category have also been charted, and Category019 having the largest

% contribution.

The results are as follows:



12. Additionally, we used the following code to resample because we required the date wise

Order demand for Category 019

df19 = df [df.Product\_Category=='Category\_019']

df19=df19.set\_index ('Period')

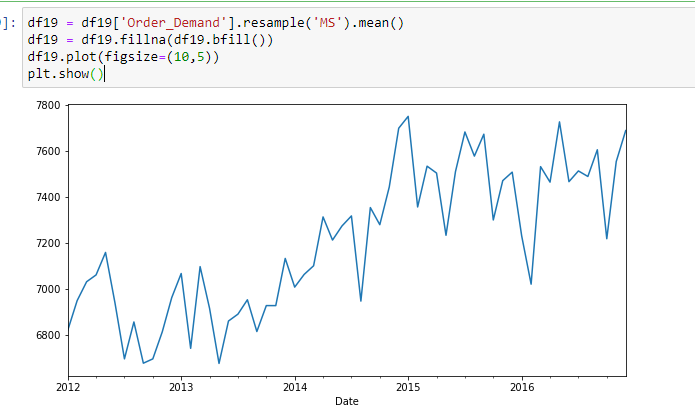
df19=df19.drop (['Product Category'], axis=1)

df19

13. Additionally, we resampled the data every month. After finishing, we plot the results of

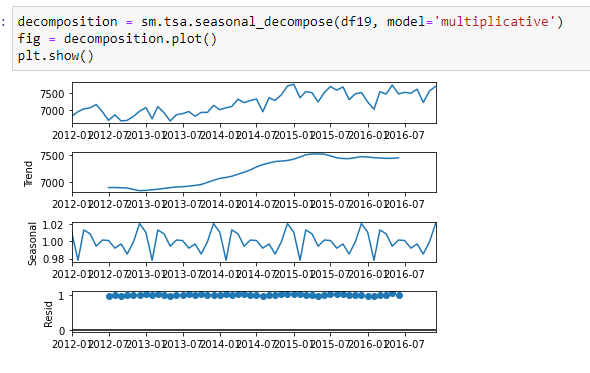
the time series analysis. In df19, we'll receive 60 entries. This is as a result of our monthly

data resampling.



df19.shape (60, 0)

14. We must perform the seasonal decomposition function as stated below in order to determine which model should be applied.



As we can see that the Seasonal signal clearly tells us that we will have to use the SARIMAX Model for the project.

15. ACF: Auto Correlation Factor.

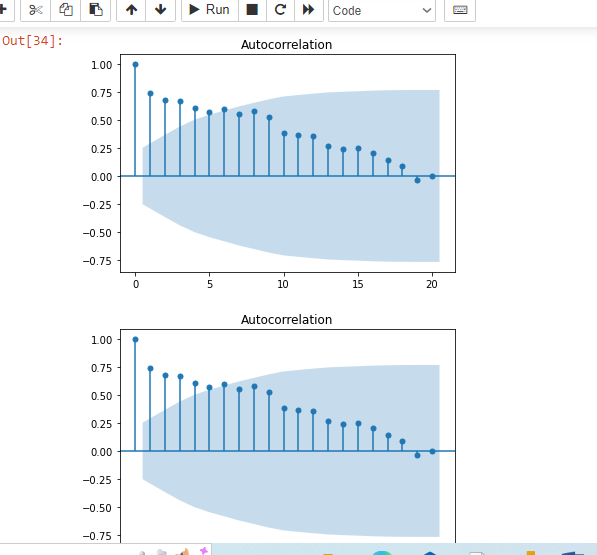
To determine which month is having an impact on the current month's demand.

16. PACF: Partial Correlation Factor

17. Lags: Number of time periods for which the PACF/ACF are being considered.



PACF & ACF are useful for entering the values of p and q in the SARIMAX Model.



Another value, "d," sometimes known as the differencing value, is present in addition to these

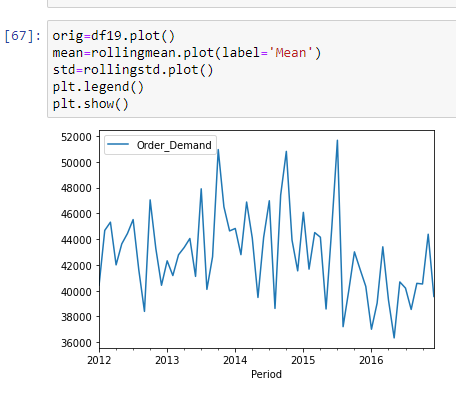
and indicates how long the differencing process took.

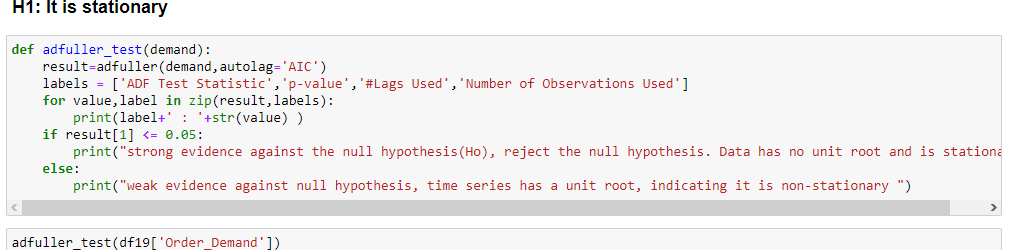
The data are differentiated to make them stationary.

We utilised the "Adfuller" test to determine whether or not the dataset is steady. This test is

based on the null hypothesis, which states that the data are not stationary, and the alternate

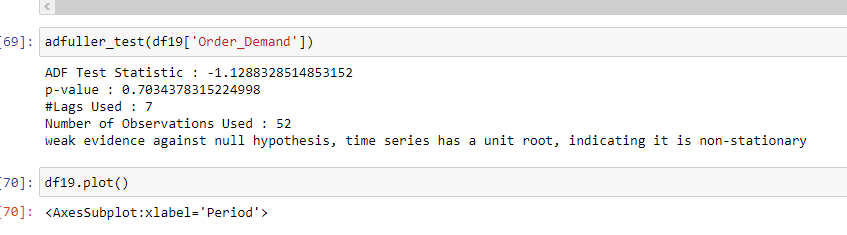
Hypothesis, which states that the data are stationary.





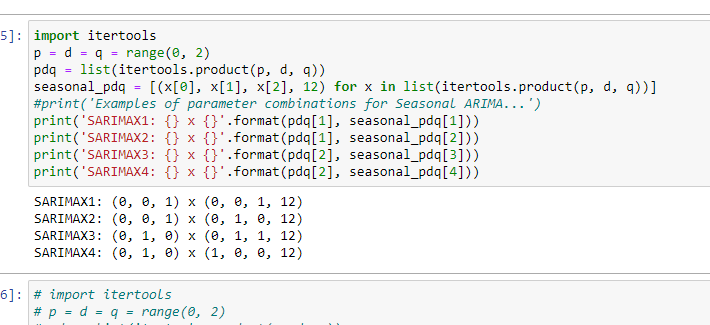
In this test, we find the sig value of the entire dataset to check. If the sig value is more than .05 then, the null hypothesis will be accepted & the data will be not stationary. If the sig value is less than .05 then, the alternate hypothesis will be accepted & the data will be stationary.

We can observe that data is not steady in this instance:



Now, as shown in the graphic below, we have used the itertools library to determine the

Values for p, d, and q for our dataset:



We shall choose the minimal AIC Score for p, d, and q.

P :Factor of Partial Correlation

D: Difficult to Compare Factor

Q: Autocorrelation Factor

for param in pdq:

for param\_seasonal in seasonal\_pdq:

try:

mod = sm.tsa.statespace.SARIMAX(df19,

order=param,

seasonal\_order=param\_seasonal,enforce\_stationarity=False,

enforce\_invertibility=False)

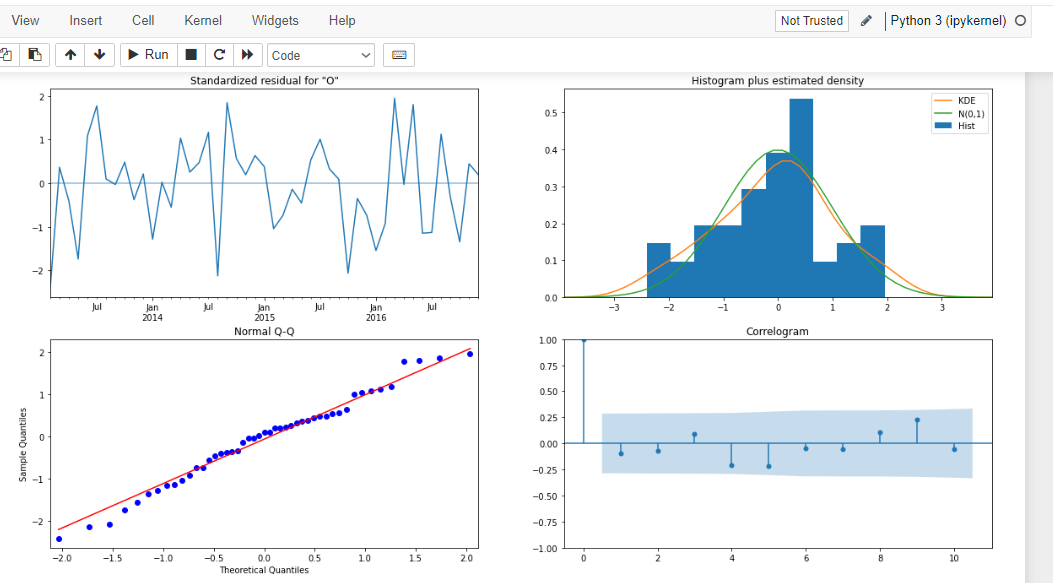
results = mod.fit()

print('SARIMA{}x{}12 - AIC:{}'.format(param, param\_seasonal, results.aic))

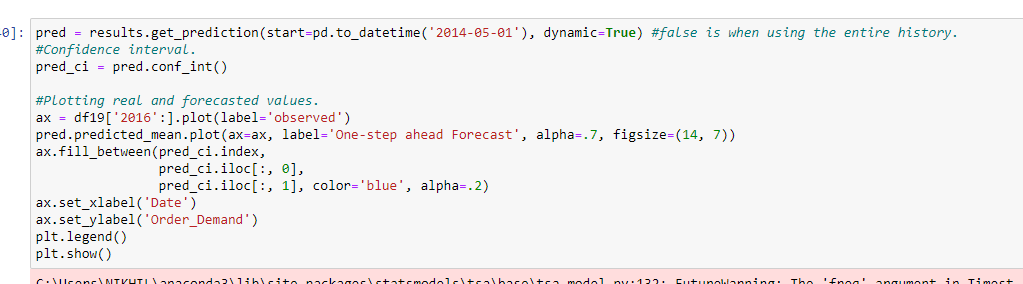
except:

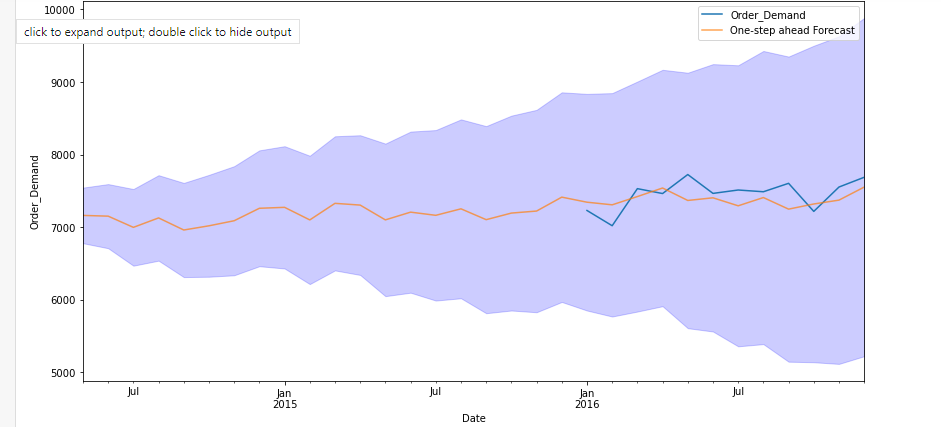
continue

After deciding on the minimal AIC Score, we put SARIMAX into practise.



The output prediction for the SARIMAX Model is shown in the graphic below:





We used the expected output from the SARIMAX Model for inventory management. We

then input the forecast into a list and carried out the reasoning as follows:

1) list1=[200,250,300,350,50]

2) stock=0

li=[]

# for i in range(len(list1)):

# print(i)

for x in list1:

if stock<=(x\*1.2):

#then placedOrder

Extra\_order=(x\*1.2)-stock

stock=stock+Extra\_order

print("refill\_stock",stock)

stock=stock-x#balancedStock

print("balanced\_Stock",stock)

print('orderPlaced',Extra\_order)

li.append(Extra\_order)

else:

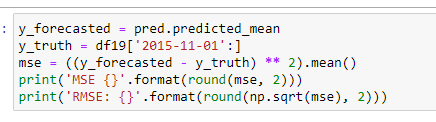
#dnt placeOrder

print('dontplace')

# print(x)

print(li)

We have used streamlit to deploy the model:



mae = np.mean(np.abs(results.resid))

MAE: 328.041

5. Conclusion:

SARIMAX Model was utilised by me to forecast inventory for the next six months (Seasonal

Auto-Regressive Moving Average with Exogenous Factors). So, this model provides a

general understanding of inventory management. For instance, the demand for Warehouse J

Category is higher than it is for the other Warehouse Categories. Therefore, the project

clearly explains which Warehouse Category will produce more inventory in the near future.

6. References:

1) <https://www.kaggle.com/datasets/felixzhao/productdemandforecasting>

2)<https://paperswithcode.com/datasets?q=demand+forecasting+and+inventory+management&>

3.) <https://towardsdatascience.com/demand-forecast-with-different-data-science-approaches->

4.) <https://github.com/kcngnn/Product-DemandForecasting/blob/master/Analysis%20and%20Model.ipynb>

5.) <https://projectsinventory.com/code-of-inventory-management-system-project-in-python/>

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Vivekanand Education Society's Business School (VBS)** | | | | | | | | |
| **Evaluation Sheet for Industry Mentor PGDM Batch 2021-2023** | | | | | | | | |
| **Sr.**  **No.** | **Roll No.** | **Name of the**  **Student** | **Name of**  **the**  **Industry** | **Attendance** | **Dedication Toward Assigned Work** | **Successfully Conducting the task**  **/Project** | **Overall Behavior** | **Total Marks** |
| **(5)** | **(5)** | **(5)** | **(5)** | **(20)** |
| 3 | 133 | Himanshu Chetnarayan Singh | Compunnel Digital | 2 | 2 | 3 | 3 | 10 |

Name of the Industry Mentor- Dr. Ravi Changale

Signature of the Industry Mentor – Dr. Ravi Changale

**ANNEXURE XIV**

**STUDENT’S WEEKLY DIARY/ WEEKLY LOG**

|  |  |  |  |
| --- | --- | --- | --- |
| WEEK No- 1 (Induction Session) | | | DATE: From 06/06/2022 To- 13/06/2022 |
|  | |  |  | |
| Dept. / Division | | Demand Forecasting and Inventory  Management. | Percentage target achieved- 100% | |
|  |  | |  |
| Name of HOD/  Industry Mentor  With e-mail id | Dr. Ravi Changale | |  |
|  | |  |
| Ravi.Changle@compunneldigital.com | |  |

**List down main point of the week**

1) Induction Program helped us to know about the organization’s knowledge and culture.

2) Explored about Artificial Intelligence and its Applications on the Internet.

**Signature of Industry Mentor**

**ANNEXURE XIV**

**STUDENT’S WEEKLY DIARY/ WEEKLY LOG**

|  |  |  |
| --- | --- | --- |
| WEEK No- 2 Manufacturing Sector Deck | | DATE: From- 13/06/2022 To- 19.06.2022 |
| Dept. / Division | Demand Forecasting and Inventory  Management. | Percentage target achieved- 100% |
|  |  |  |
| Name of HOD/ Supervisor  With e-mail id | Dr. Ravi Changle |  |
|  |  |
| Ravi.Changle@compunneldigital.com |  |

**List down main point of the week**

1) Researched about the Manufacturing Sector on the Internet.

2) Explored about the use cases.

3) Understood how broad AI domain is.

4) The goal of AI in Manufacturing.

**Signature of Industry Mentor**

**ANNEXURE XIV**

**STUDENT’S WEEKLY DIARY/ WEEKLY LOG**

|  |  |  |
| --- | --- | --- |
| WEEK No- 3 ( Machine Learning Training) | | DATE: From- 20/06/2022 To- 26/06/2022 |
| Dept. / Division | Demand Forecasting and Inventory  Management. | Percentage target achieved- 100% |
|  |  |  |
| Name of HOD/ Supervisor  With e-mail id | Dr. Ravi Changle |  |
|  |  |
| Ravi.Changle@compunneldigital.com |  |

**List down main point of the week**

1) Understood how machine learning uses algorithms to imitate the

humans learn.

2) How it uses statistical methods to train algorithms and make accurate predictions

3) The accuracy of these predictions improves over time.

4) Supervised Learning and Linear Regression.

5) Classification and Logistic Regression.

6) Decision Tree and Random Forest.

7) Naïve Bayes and Support Vector Machine.

8) Unsupervised Learning.

9) Natural Language Processing and Text Mining.

10) Introduction to Deep Learning.

11) Time Series Analysis.

**Signature of Industry Mentor**

**ANNEXURE XIV**

**STUDENT’S WEEKLY DIARY/ WEEKLY LOG**

|  |  |  |
| --- | --- | --- |
| WEEK No- 4 (Power BI Assignment) | | DATE: From- 27/06/2022 To- 02/07/2022 |
| Dept. / Division | Demand Forecasting and Inventory Management | Percentage target achieved- 100% |
|  |  |  |
| Name of HOD/ Supervisor  With e-mail id | Dr. Ravi Changle |  |
|  |  |
| Ravi.Changle@compunneldigital.com |  |

**List down main point of the week**

1) Completed Power BI Assessment.

2) How Data Visualization tools provide an accessible way to see and understand trends, Outliers and patterns in data.

3) Understood the common flow of activity in Power BI-

Bring Data into Power BI desktop and create a report, share it to Power BI Service, view

and interact with reports and dashboards in the service and Power BI Mobile.

4) One can create visualizations in a page sheet.

5) Advantages of Power BI

1) Information can be visualized using powerful templates and visualization.

2) Users get cutting intelligence technologies and powerful algorithms that are updated

Regularly.

3) Users can have personalized dashboards which are easy to access and understand.

4) Users can perform queries on reports using DAX Language. DAX is a type of data

Modelling in Power BI that uses Data Analysis Expressions and similar to Excel formulas.

6) Data Visualization helps users in analyzing a large amount of data in a simpler way.

7) Data Visualization makes complex data more accessible, understandable, and usable.

8) Data Visualization is a graphical Representation of data.

9) BI Stands for Business Intelligence.

10) A Dashboard is a single page, often called a canvas that tells a story through Visualizations.

**Signature of Industry Mentor**

**ANNEXURE XIV**

**STUDENT’S WEEKLY DIARY/ WEEKLY LOG**

|  |  |  |
| --- | --- | --- |
| WEEK No- 5 Power BI Quiz | | DATE: From- 04/07/2022 To- 09/07/2022 |
| Dept. / Division | Demand Forecasting and Inventory Management | Percentage target achieved – 100% |
|  |  |  |
| Name of HOD/ Supervisor  With e-mail id | Dr. Ravi Changle | Attempted |
|  |  |
| Ravi.Changle@compunneldigital.com |  |

**List down main point of the week**

1) Attempted the Power BI Quiz.

**Signature of Industry Mentor**

**ANNEXURE XIV**

**STUDENT’S WEEKLY DIARY/ WEEKLY LOG**

|  |  |  |
| --- | --- | --- |
| WEEK No- 6 ( Recording of Deployment) | | DATE: From- 11/07/2022 To- 16/07/2022 |
| Dept. / Division | Demand Forecasting and Inventory  Management. | Percentage target achieved -100% |
|  |  |  |
| Name of HOD/ Supervisor  With e-mail id | Dr.Ravi Changle |  |
|  |  |
| Ravi.Changle@compunneldigital.com |  |

**List down main point of the week**

1) Completed all the given tasks assigned in the given weeks.

2) Completed watching all the Recordings.

**Signature of Industry Mentor**

**STUDENT’S WEEKLY DIARY/ WEEKLY LOG**

|  |  |  |
| --- | --- | --- |
| WEEK No- 7 Project assigned on Demand Forecasting and Inventory Management | | DATE: From- 18/07/2022 To- 23/07/2022 |
| Dept. / Division |  | Percentage target achieved- 100% |
|  |  |  |
| Name of HOD/ Supervisor  With e-mail id |  |  |
|  |  |
|  |  |

**List down the main points of the week**

1) Worked on finding a Dataset from various sources provided.

2) Finalized a Dataset and started working on it.

3) Started working on Time Series Analysis.

4) Researched about Inventory Management.

5) Found various sources for Inventory Management.

6) Completed the Inventory Management Coding.

7) Completed the Project.

