***Product demand prediction with***

***MACHINE learning***

Batch Member

712221121015:SUGASRI.K

Phase 3 Submission Document

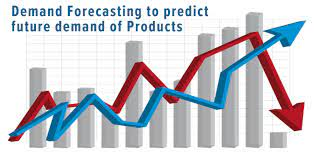
Title: Product Demand Prediction with Machine Learning

Phase 3: Development part 1

Topic : Start Build the project product demand prediction by

loading and preprocessing the dataset



**** Product Demand Prediction

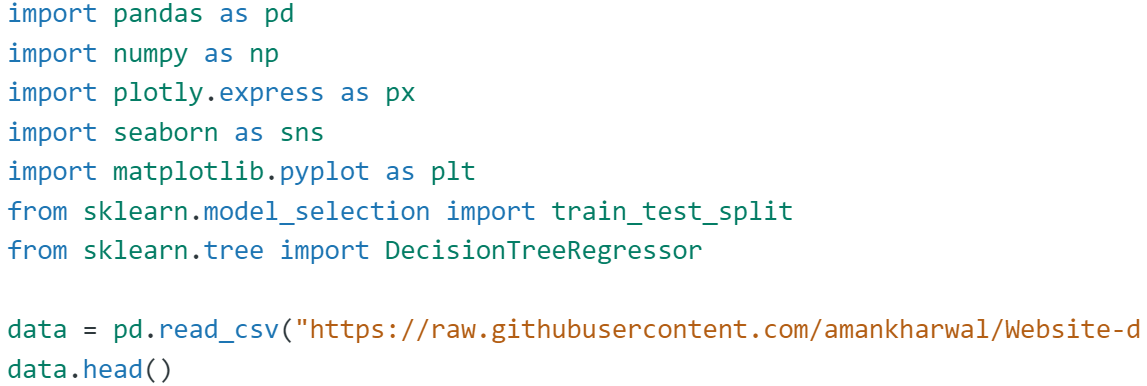
**INTRODUCTION**

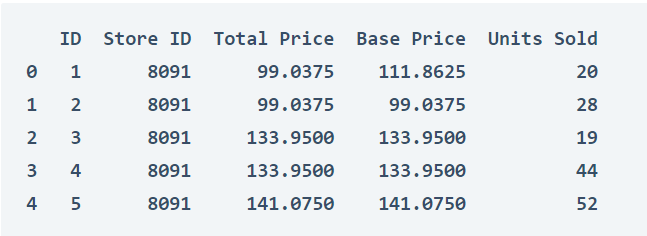
Demand forecasting is the process of using predictive analysis of historical data to estimate and predict customers' future demand for a product or service. Demand forecasting helps the business make better-informed supply decisions that estimate the total sales and revenue for a future period of time. Machine learning (ML) in demand forecasting makes it possible to avoid traditional challenges associated with planning such as long delivery lead times, high transport costs, high inventory and waste levels, and incorrect decision making due to inaccurate.

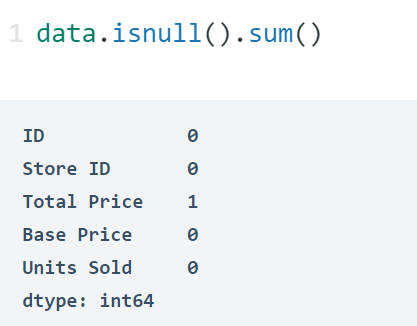
For example, Machine Learning algorithms incorporate a range of techniques including time-series analysis, autoregressive integrated moving average (ARIMA) and exponential smoothing methods to predict demand, accounting for seasonality and other time-based patterns incorrect decision making due to inaccurate forecasts.

***1.Import the dataset:***

In this step, you need to import the dataset/s that you have gathered for the ML project at hand. Importing the dataset is one of the important steps in data preprocessing in machine learning. However, before you can import the dataset/s, you must set the current directory as the working directory. importing the necessary Python libraries and the dataset we need for the task of product demand prediction

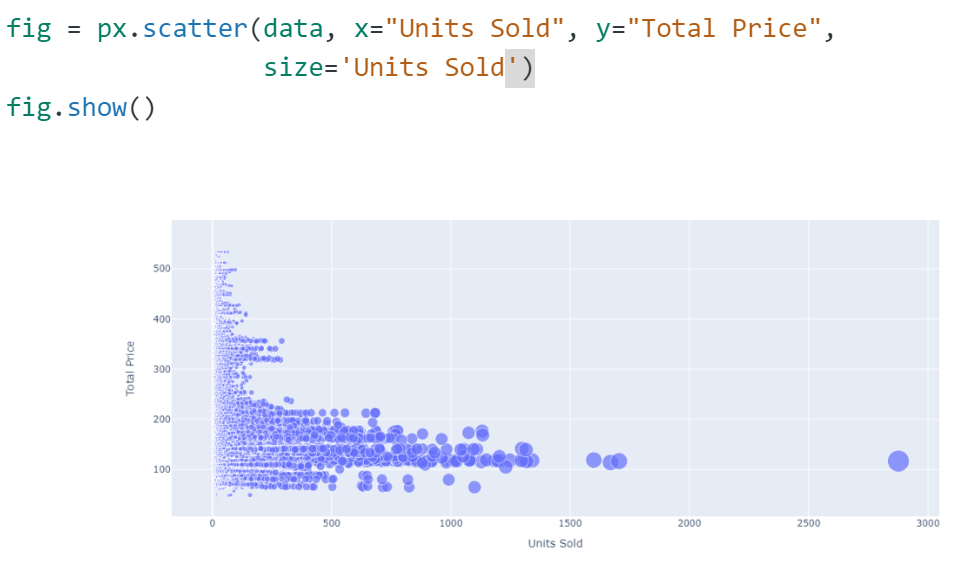






So the dataset has only one missing value in the Total Price column, I will remove that entire row for now:

data = data.Dropna()



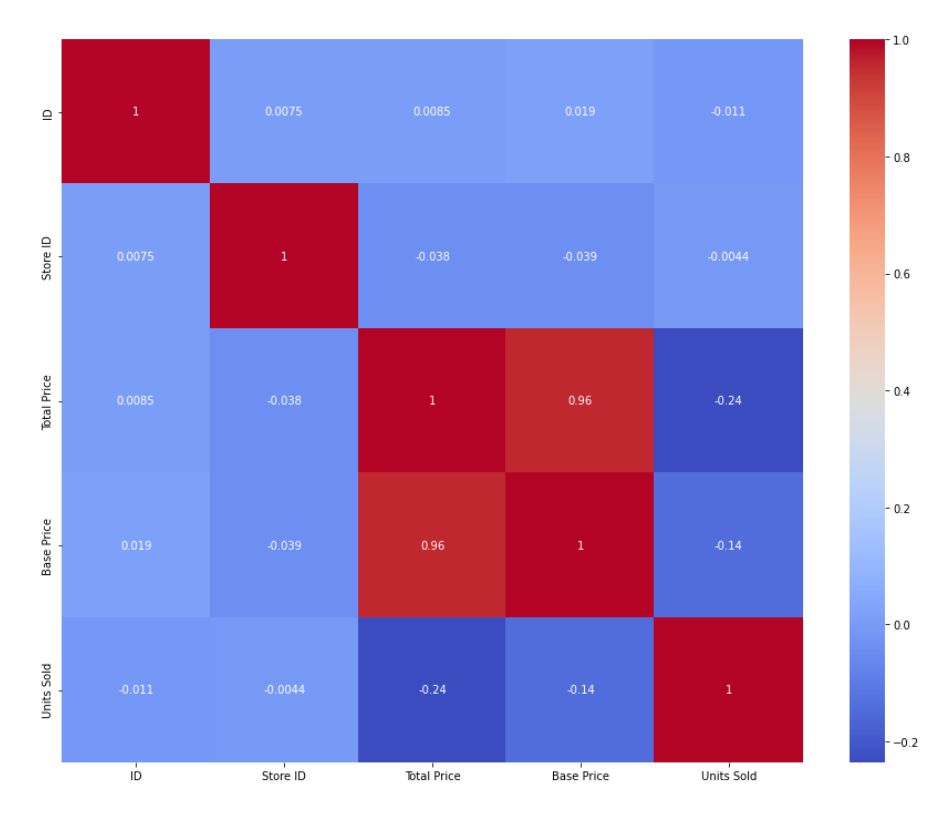


correlations = data.corr(method='pearson')

plt.figure(figsize=(15, 12))

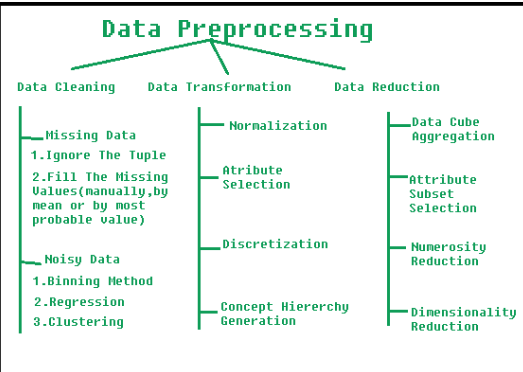
sns.heatmap(correlations, cmap="coolwarm", annot=True)

plt.show()



***Data preprocessing:***

Data preprocessing is the process of transforming raw data into an understandable format. It is also an important step in data mining as we cannot work with raw data. The quality of the data should be checked before applying machine learning or data mining algorithms.



***Data Cleaning***:

Data cleaning, one of the major preprocessing steps in machine learning, locates and fixes errors or discrepancies in the data. From duplicates and outliers to missing numbers, it fixes them all. Methods like transformation, removal, and imputation help ML professionals perform data cleaning seamless.

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

**1. Missing Data**:

This situation arises when some data is missing in the data. It can be handled in various ways.

Ignore the tuples:

This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.

Fill the Missing values:

There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.

**2.Noisy Data:**

Noisy data is a meaningless data that can’t be interpreted by machines.It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :

Binning Method:

This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.

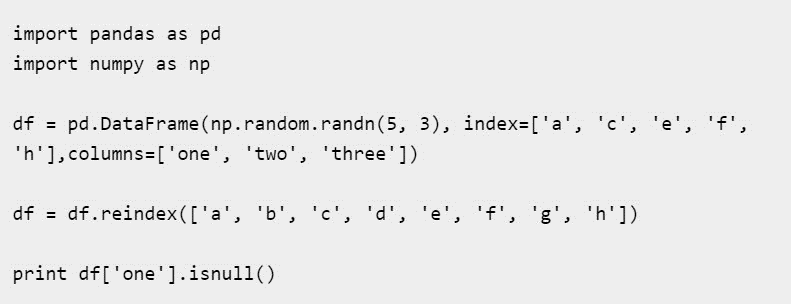
Regression:

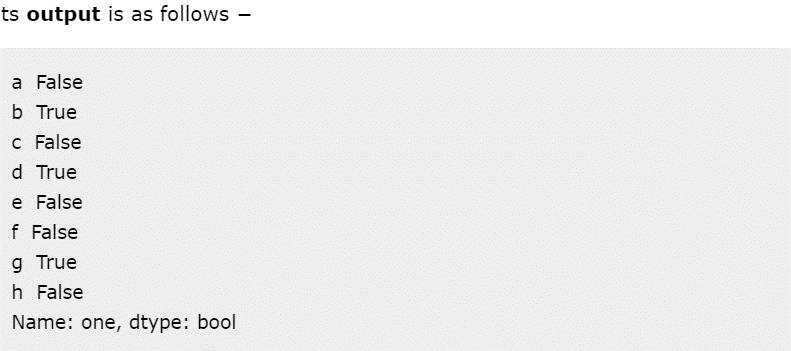
Here data can be made smooth by fitting it to a regression function.The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).

Clustering:

This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

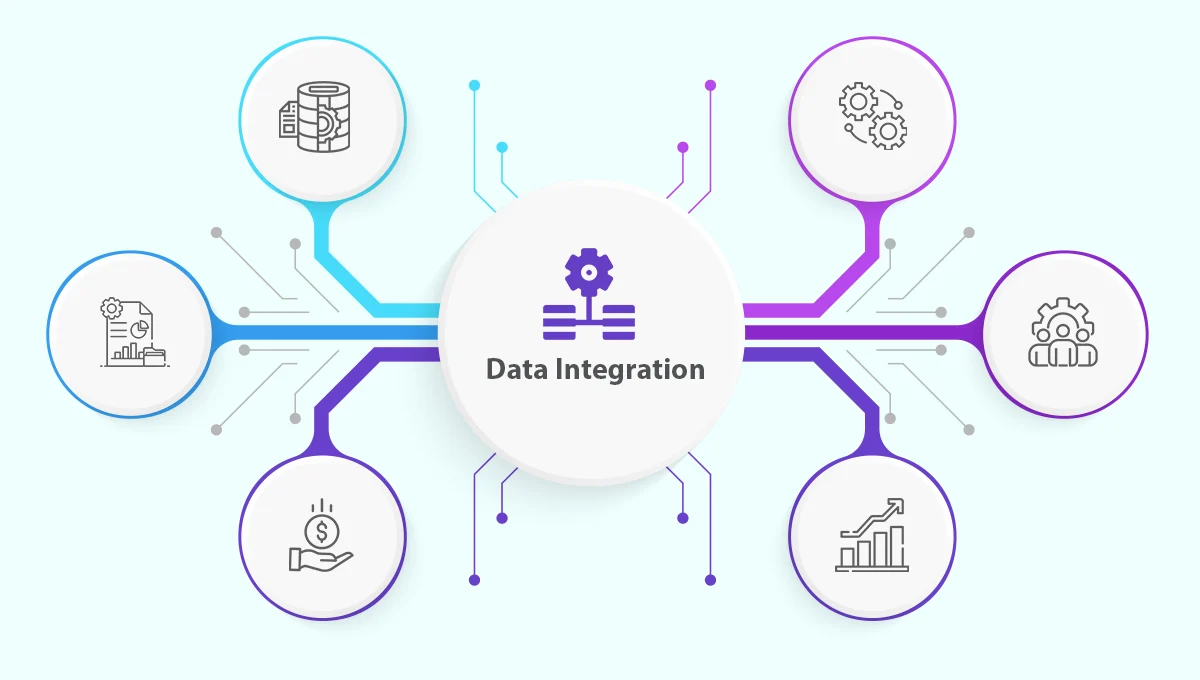
**Example :**





***Data Integration***

Data integration is among the major responsibilities of data preprocessing in machine learning. This process integrates (merges) information extracted from multiple sources to outline and create a single dataset. The fact that you need to handle data in multiple forms, formats, and semantics makes data integration a challenging task for many ML developers.



***Data Transformation***

ML programmers must pay close attention to data transformation when it comes to data preprocessing steps. This process entails putting the data in a format that will allow for analysis. Normalization, standardization, and discretisation are common data transformation procedures. While standardization transforms data to have a zero mean and unit variance, normalization scales data to a common range. Continuous data is discretized into discrete categories using this technique.

**1. Smoothing** - Data Smoothing is used to remove noise in the dataset, and it helps identify important features and detect patterns. Therefore, it can help in predicting trends or future events.

**2.Aggregation** - Data Aggregation is the process of transforming large volumes of data into an organized and summarized format that is more understandable and comprehensive. For example, a company may look at monthly sales data of a product instead of raw sales data to understand its performance better and forecast future sales.

**3.Discretization** - Data Discretization is a process of converting numerical or continuous variables into a set of intervals/bins. This makes data easier to analyze. For example, the age features can be converted into various intervals such as (0-10, 11-20, ..) or (child, young, …).

**4.Normalization** - Data Normalization is a process of converting a numeric variable into a specified range such as [-1,1], [0,1], etc. A few of the most common approaches to performing normalization are Min-Max Normalization, Data Standardization or Data Scaling, etc.



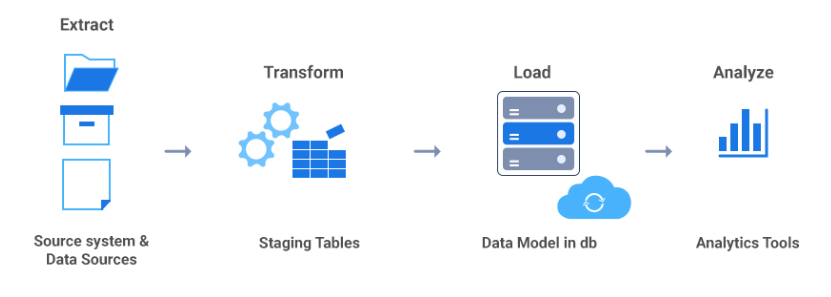
***Data Reduction :***

Data reduction is the process of lowering the dataset’s size while maintaining crucial information. Through the use of feature selection and feature extraction algorithms, data reduction can be accomplished. While feature extraction entails translating the data into a lower-dimensional space while keeping the crucial information, feature selection requires choosing a subset of pertinent characteristics from the dataset.

**1.Dimensionality Reduction** - It is the process of reducing the number of features in the input dataset. It can be performed in various ways, such as selecting features with the highest importance, Principal Component Analysis (PCA), etc.

**2.Numerosity Reduction** - In this method, various techniques can be applied to reduce the volume of data by choosing alternative smaller representations of the data. For example, a variable can be approximated by a regression model, and instead of storing the entire variable, we can store the regression model to approximate it.

**3.Data Compression** - In this method, data is compressed. Data Compression can be lossless or lossy depending on whether the information is lost or not during compression.



***Applications of Data Preprocessing :***

* Data Preprocessing is important in the early stages of a Machine Learning and AI application development lifecycle. A few of the most common usage or application include -
* Improved Accuracy of ML Models - Various techniques used to preprocess data, such as Data Cleaning, Transformation ensure that data is complete, accurate, and understandable, resulting in efficient and accurate ML models.
* Reduced Costs - Data Reduction techniques can help companies save storage and compute costs by reducing the volume of the data
* Visualization - Preprocessed data is easily consumable and understandable that can be further used to build dashboards to gain valuable insights.

***Conclusion***

* Data Preprocessing is a process of converting raw datasets into a format that is consumable, understandable, and usable for further analysis. It is an important step in any Data Analysis project that will ensure the input datasets's accuracy, consistency, and completeness.
* The key steps in this stage include - Data Cleaning, Data Integration, Data Reduction, and Data Transformation.
* It can help build accurate ML models, reduce analysis costs, and build dashboards on raw data.

THANK YOU