MARKET BASKET INSIGHTS

Phase 3 Submission Document

Project: Market basket Insights

Topic: building the market basket insights model to loading and pre-processing in given data set



**Introduction:**

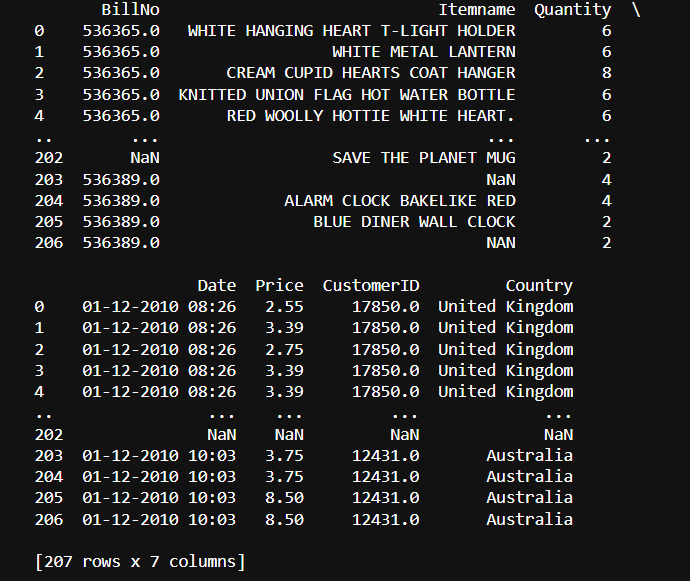
1. Market basket analysis is a technique that helps retailers understand the buying patterns of their customers and optimize their sales strategies accordingly. It involves analysis large data sets of purchase history to find out which products are frequently bought together, and how likely a customer is to buy one product given that they have bought another product.
2. retailers use data mining algorithms that extract patterns and rules from large data sets. One of the most common algorithms is the Apriori algorithm.

**market basket insights:**

**Dataset link:**

[**https://www.kaggle.com/datasets/aslanahmedov/market-basket-analysis**](https://www.kaggle.com/datasets/aslanahmedov/market-basket-analysis)

model dataset:



Using Libraries:

Program:

import pandas as pd

import numpy as np

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import StandardScaler

Data set loading and pre-processing:

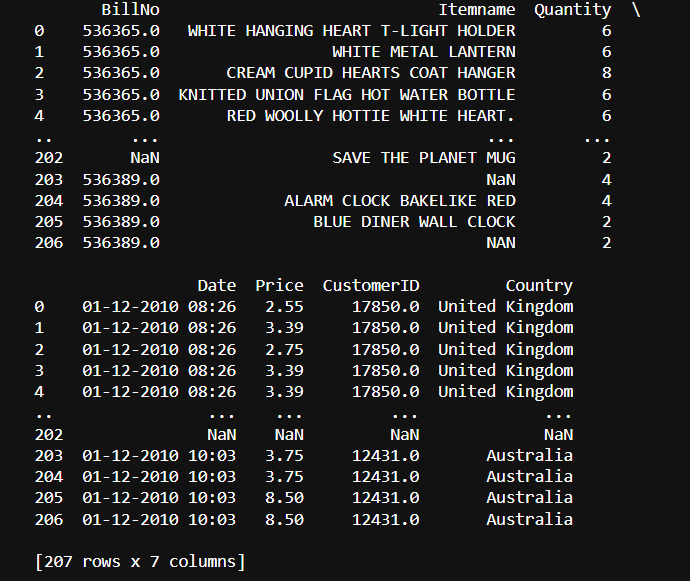
1. Clean and preprocess the data, removing any irrelevant information, handling missing values, and converting the data into a suitable format for analysis. You can use tools like Excel, Python, or R to perform these tasks.
2. Apply data mining algorithms that extract patterns and rules from large data sets.

1.Loading Data set:

Program:

df = pd.read\_csv(' E:\market basket insights.csv ')

print(df)

output: 

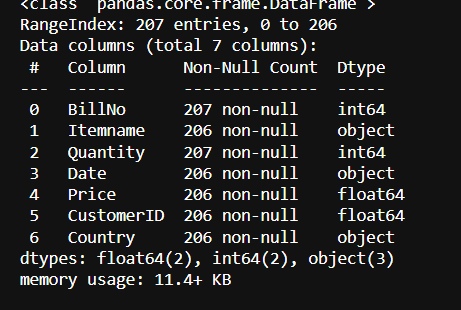
2.Display the data types:

Display the information about the DataFrame which is to provide an overview of the DataFrame's structure and column data types.

Program:

df.info()

output:



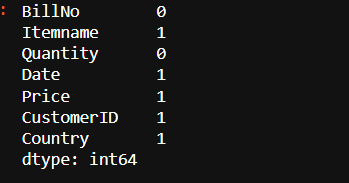
3.Calculate missing values:

Calculate the number of missing values for each column and sort them in descending order

Program:

data.isnull().sum()

output:



4.Split the Data:

Split your dataset into training and testing sets. This helps you evaluate

your model's performance later.

X = df.drop('Itemsname', axis=1)

y = df['Itemsname']

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2,

random\_state=42)

Important of data loading and preprocess:

Loading and preprocessing the dataset is an important first step in building any machine learning model. However, it is especially important for market basket insight , as remove the incorrect data and error values

By loading and preprocessing the dataset, we can ensure that the machine learning algorithm is able to learn from the data effectively and accurately.

Process of loading and preprocessing in market basket insights:

There are a number of challenges involved in loading and preprocessing

1.Handling missing values:

datasets often contain missing values, which can be due to a variety of factors, such as human error or incomplete data collection. Common methods for handling missing values include dropping the rows with missing values, imputing the missing values with the mean or median of the feature, or using a more sophisticated method such as multiple imputation.

**2.Unacceptable format:**

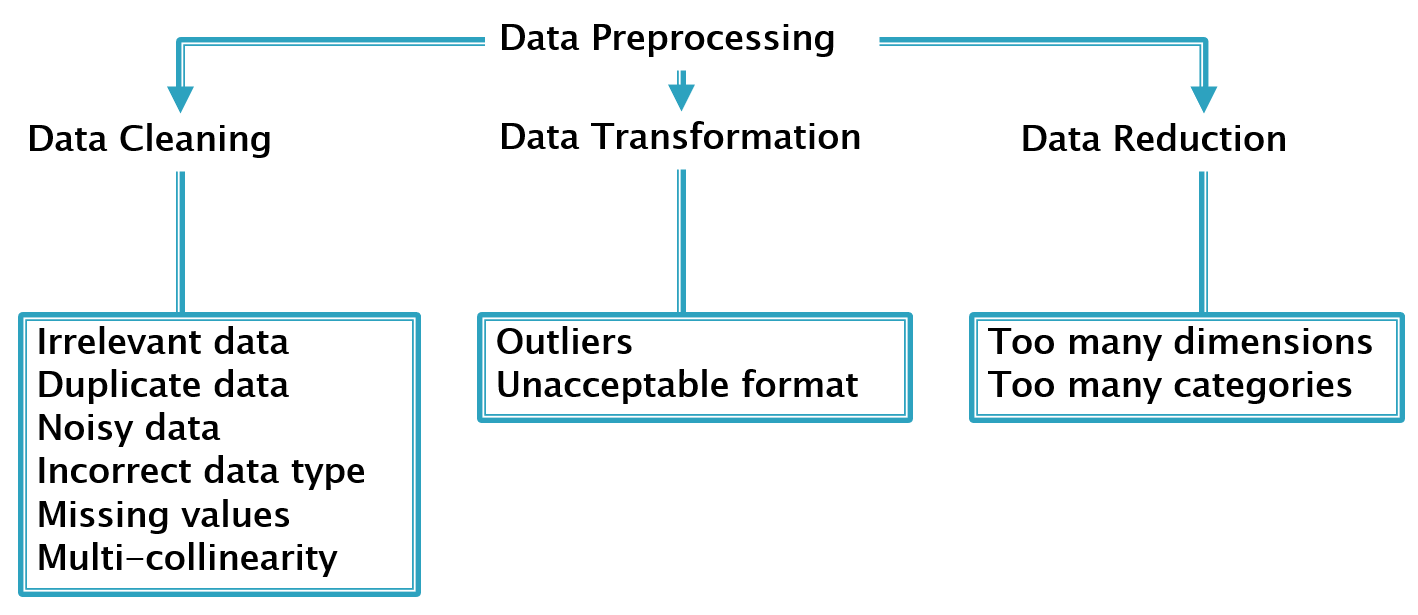
The data may be in a format that is not acceptable to the machine learning algorithm to be applied later in the data mining stage.

**3. Duplicate data:**

Due to data replication, duplicate rows and columns would hamper the analysis process. It should be detected and removed at the onset of the data cleaning process. In case of a data value conflict for redundant data, the original data source can be checked to resolve the problem and perform data integration.

**4.Clustering:**

This involves grouping similar data points together into clusters. Clustering is often used to reduce the size of the dataset by replacing similar data points with a representative centroid. It can be done using techniques such as k-means, hierarchical clustering, and density-based clustering.



Loading dataset:

The specific steps involved in loading the dataset will vary depending

on the machine learning library or framework that is being used.

However, there are some general steps that are common to most

machine learning frameworks

Program:

import pandas as pd

import numpy as np

df=pd.read\_csv("E:/market basket insightscsv")

**Types:**

**Numpy.loadtxt function:**

This is a built-in function in Numpy, a famous numerical library in Python. It is a really simple function to load the data. It is very useful for reading data which is of the same datatype.

When data is more complex, it is hard to read using this function, but when files are easy and simple, this function is really powerful.

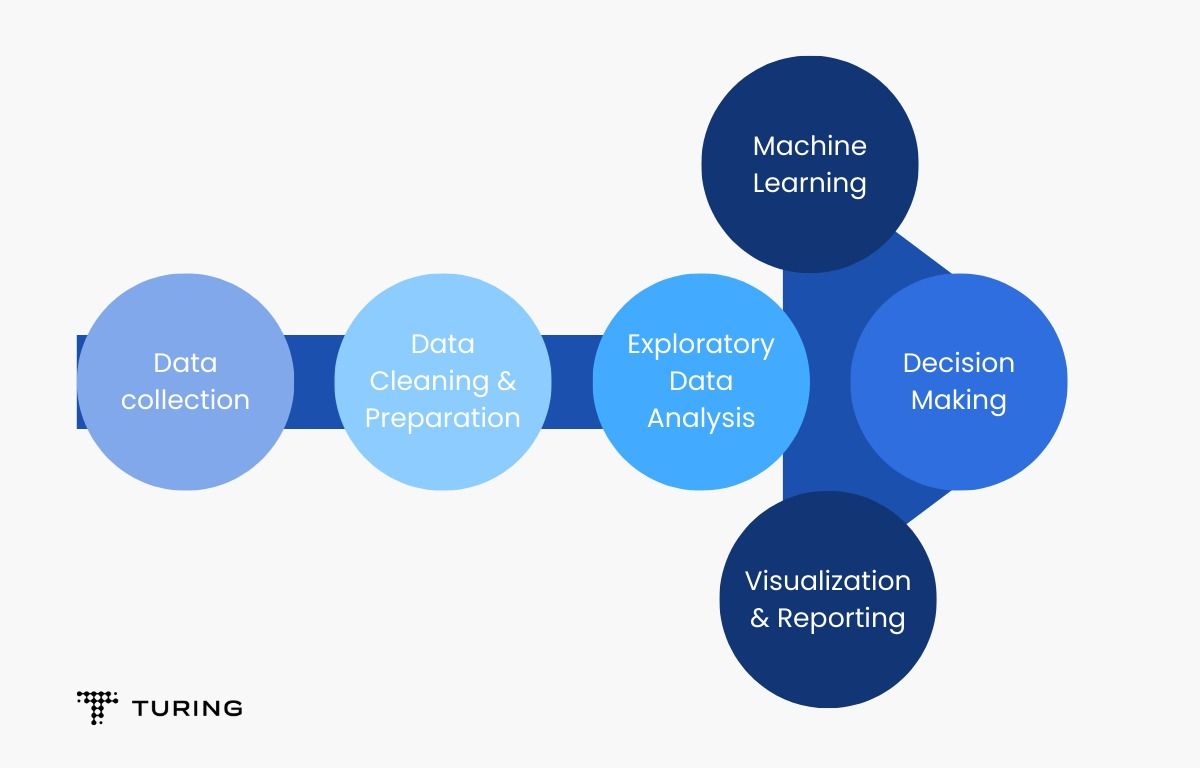
**Pandas.read\_csv():**

Pandas is a very popular data manipulation library, and it is very commonly used. One of it’s very important and mature functions is *read\_csv()* which can read any .csv file very easily and help us manipulate it. Let’s do it on our 100-Sales-Record dataset.

This function is very popular due to its ease of use. You can compare it with our previous codes, and you can check it.

**Pickle:**

When your data is not in a good, human-readable format, you can use pickle to save it in a binary format. Then you can easily reload it using the pickle library.



Programs:(1.Loading an dataset)

Import pandas as pd

df = pd.read\_csv(' E:\market basket insights.csv ')

print(df)

output:



2.Preprocessing the dataset:

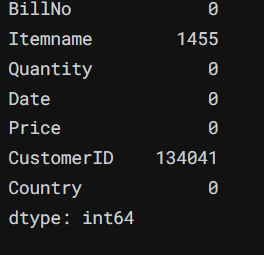
Data preprocessing is the process of cleaning, transforming, and

integrating data in order to make it ready for analysis.

1.Calculate the null values:

dataset.isnull().sum()

Output:



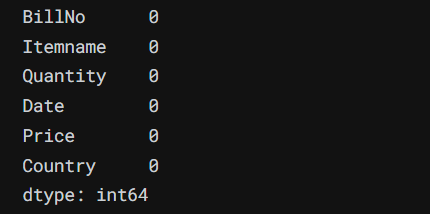
2.Remove null values:

dataset.dropna(axis=0, subset=['Itemname'], inplace = True)

dataset = dataset.drop(columns= ['CustomerID'])

dataset.isnull().sum()

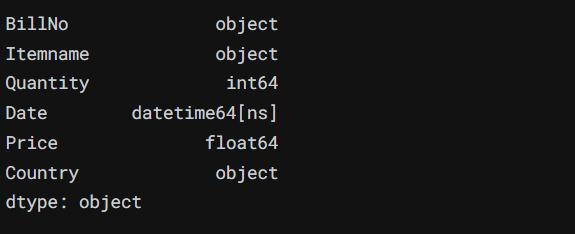
Output:



3.Datatypes:

dataset.dtypes

Output:



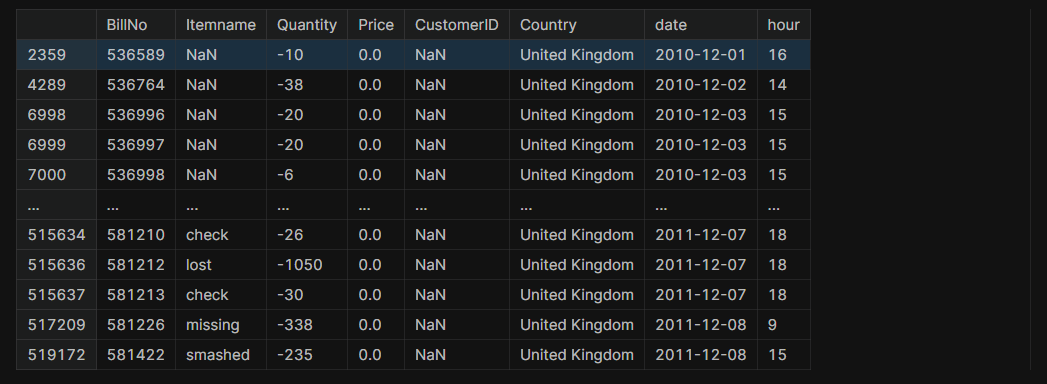
4.Remove duplicate

data.drop\_duplicates(inplace=True)

5.Find best quality of data:

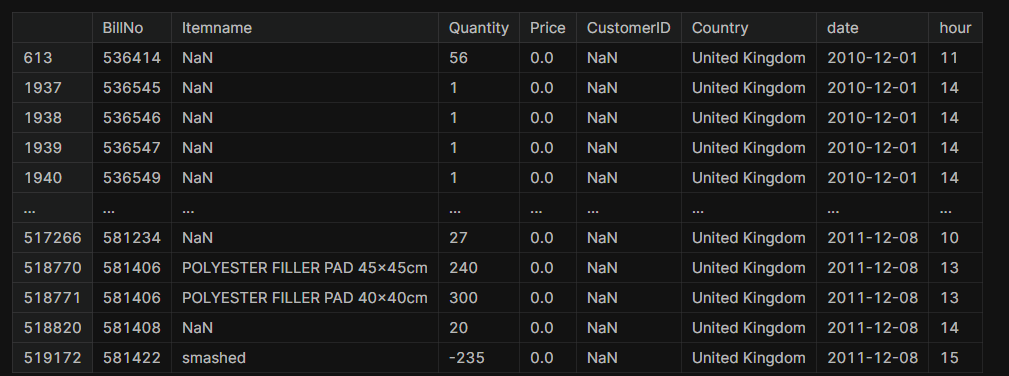
data[data['Quantity']<=0]

output:



data[data['Price']<=0]

output:



remove the rows which has the buyed quality is small or equal to zero:

data=data[data['Quantity']>0]

data=data[data['Price']>0]

data.shape

output:

(51234,8)

Program:

plt.figure(figsize=(22,7))

plt.subplot(1,2,1)

data.Price.plot()

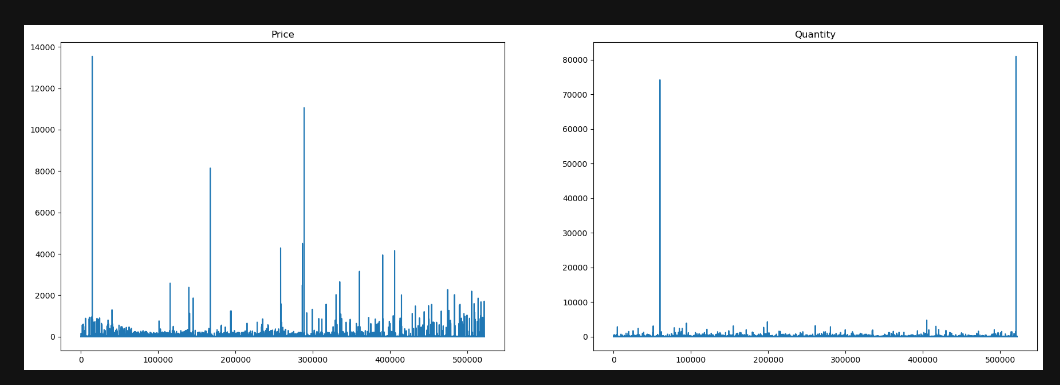
plt.title("Price")

plt.subplot(1,2,2)

data.Quantity.plot()

plt.title("Quantity")

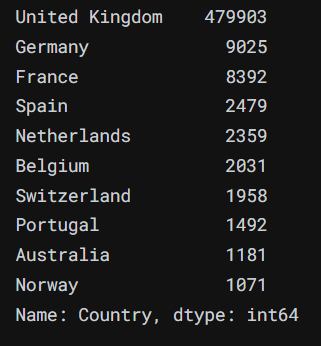
Output:



Program:

data.Country.value\_counts().head(10)

Output:



Program:

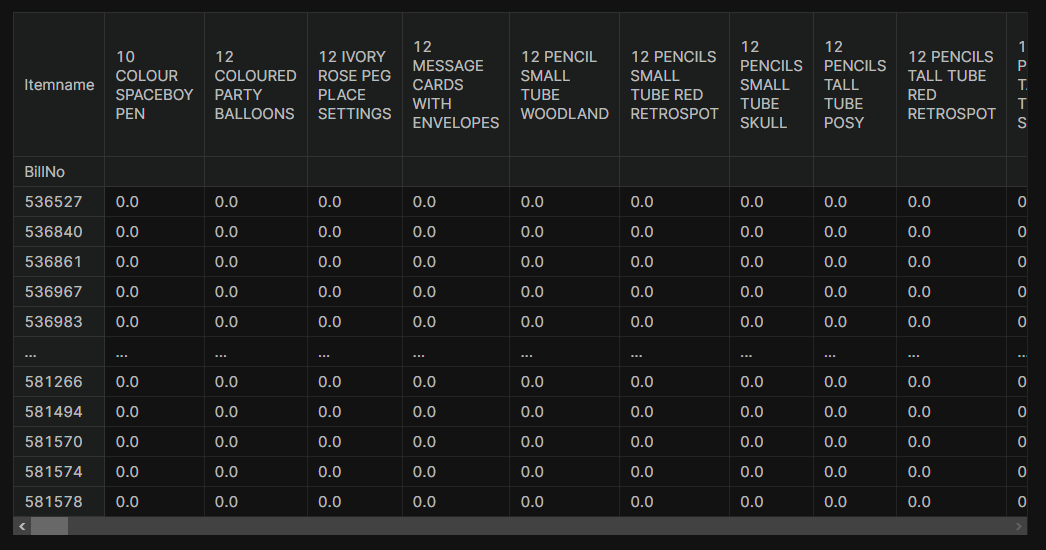
mybasket= (data[data['Country'] =="Germany"]

.groupby(['BillNo', 'Itemname'])['Quantity']

.sum().unstack().reset\_index().fillna(0)

.set\_index('BillNo'))

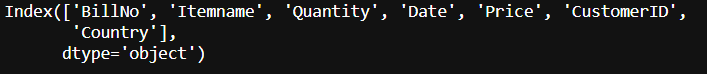
Output:



Program:

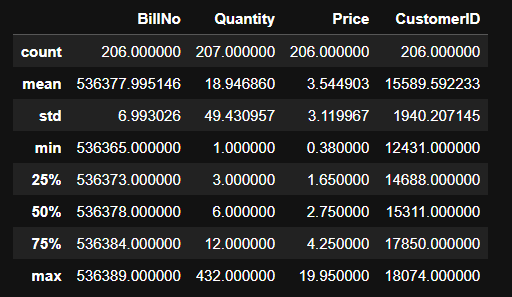
data.columns

output:



data.describe()

Output:



Conclusion:

1.Market basket insights is a powerful technique that can help businesses understand customer behaviour and preferences

2. By applying machine learning algorithms such as Apriori, AIS, SETM, or FP Growth, businesses can discover patterns and associations among the items that customers buy together.

3. Market basket insights is not limited to the retail industry; it can also be applied to other domains such as banking, healthcare, education, or e-commerce. Market basket analysis is a valuable tool for gaining insights into customer behaviour and enhancing business performance.