Data Collection and Preprocessing Phase

Date	June 2024
Team ID	739971
Project Title	Estimating the stock keeping units using Machine Learning
Maximum Marks	6 Marks

Preparation Template

The images will be preprocessed by resizing, normalizing, augmenting, denoising, adjusting contrast, detecting edges, converting color space, cropping, batch normalizing, and whitening data. These steps will enhance data quality, promote model generalization, and improve convergence during neural network training, ensuring robust and efficient performance across various computer vision tasks.

Section	Description
Data Overview	There are many popular open sources for collecting the data. Eg: kaggle.com, UCI repository, etc. In this project we have used .csv data.
Data Preparation	These are the general steps of pre-processing the data before using it for machine learning
Handling missing values	We use Handling missing values For checking the null values
Handling categorical data	As we can see our dataset has categorical data we must convert the categorical data to integer encoding or binary encoding
Handling Outliers in Data	With the help of boxplot, outliers are visualized. And here we are going to find upper bound and lower bound of numerical features with some mathematical formula.

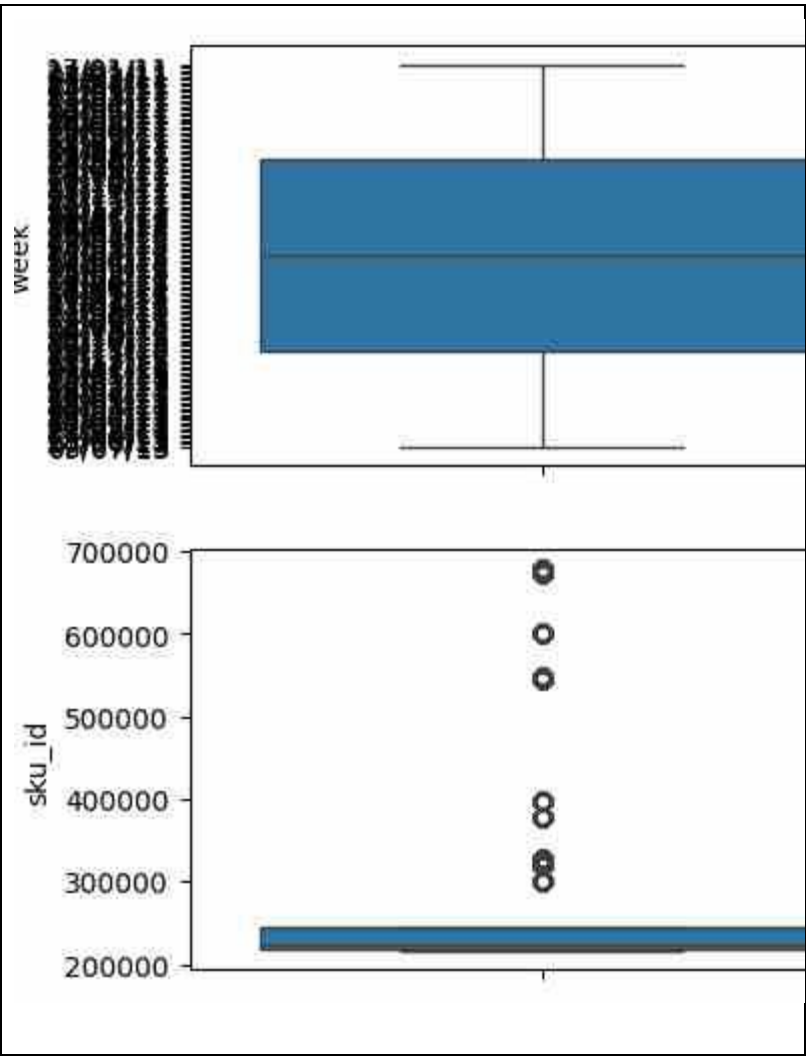
Data Preparation

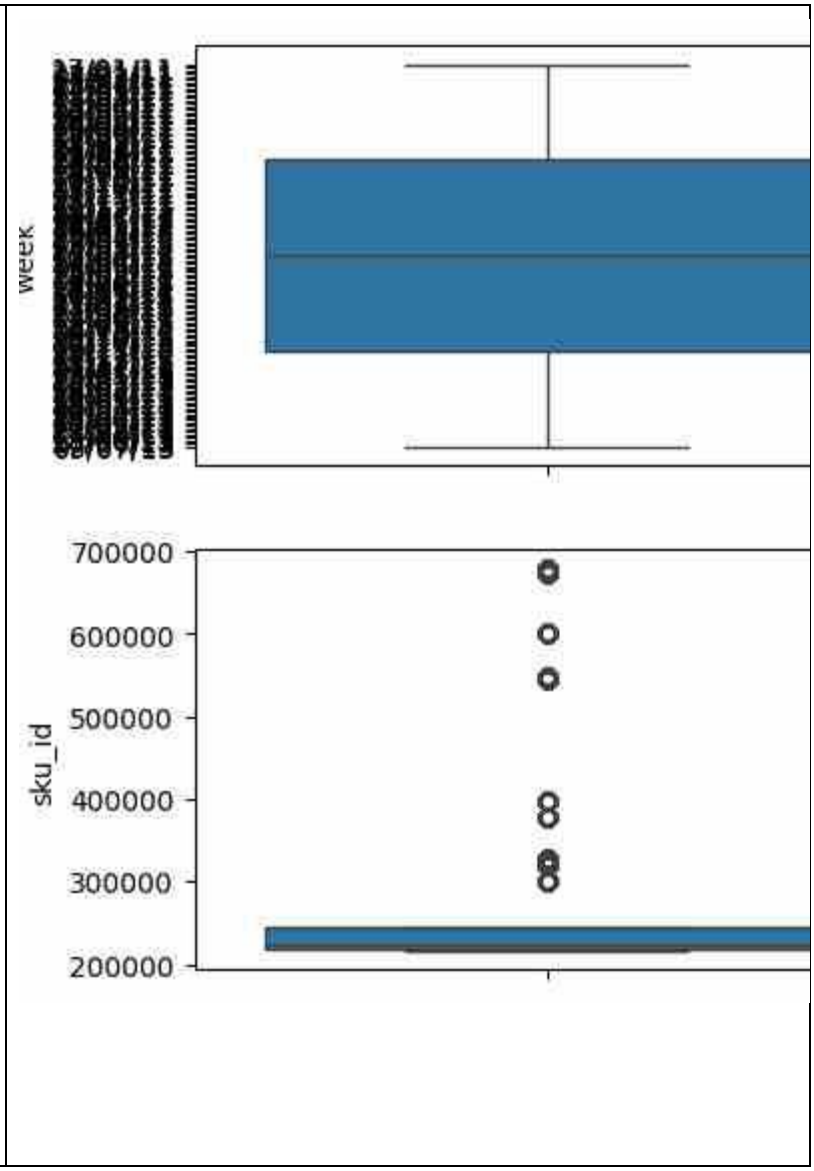
Collect the dataset	Please refer to the link given below to download the dataset. Link: https://www.kaggle.com/datasets/aswathrao/demand-forecasting

Importing the libraries	`import numpy as np`
	`import pandas as pd`
	`import matplotlib.pyplot as plt`
	`import seaborn as sns`
	`from sklearn.model_selection import train_test_spl`
	`from sklearn.linear_model import LinearRegression`
	`from sklearn.metrics import mean_squared_error,r2_`
	`from sklearn.ensemble import RandomForestRegressor`
	`from sklearn.tree import DecisionTreeRegressor`
	`from sklearn.metrics import accuracy_score`

Loading Data	We use the code `df=pd.read_csv('/content/train_0irEZ2H.csv')` For reading the dataset
Handling missing values	`df.isnull().sum()` record_ID 0 week 0 store_id 0 sku_id 0 total_price 1 base_price 0 is_featured_sku 0 is_display_sku 0 units_sold 0 dtype: int64

Finding Categorical values and numerical values	`categorical_cols=df.select_dtypes(include=['objec`
	`print(categorical_cols)`
	`Index(['week'], dtype='object')`
	`numerical_cols=df.select_dtypes(include=['int64',`
	`print(numerical_cols)`
	`Index(['record_ID', 'store_id', 'sku_id', 'total_`
	`'is_featured_sku', 'is_display_sku', 'unit`
	`dtype='object')`

Handling Outliers	


week
sku_id
700000
600000
500000
400000
300000
200000