PRUDENT TECHOLOGIES AND CONSULTING

HACKATHON 2024

DATA SCIENCE

USE CASE 1.1 Retail Sales Forecasting

By:

R.K.R. Sai Tarun

Innomatics Research Labs

Hyderabad

**Problem Statement:**

Developing a Regression model for retail sales forecasting.

Note:

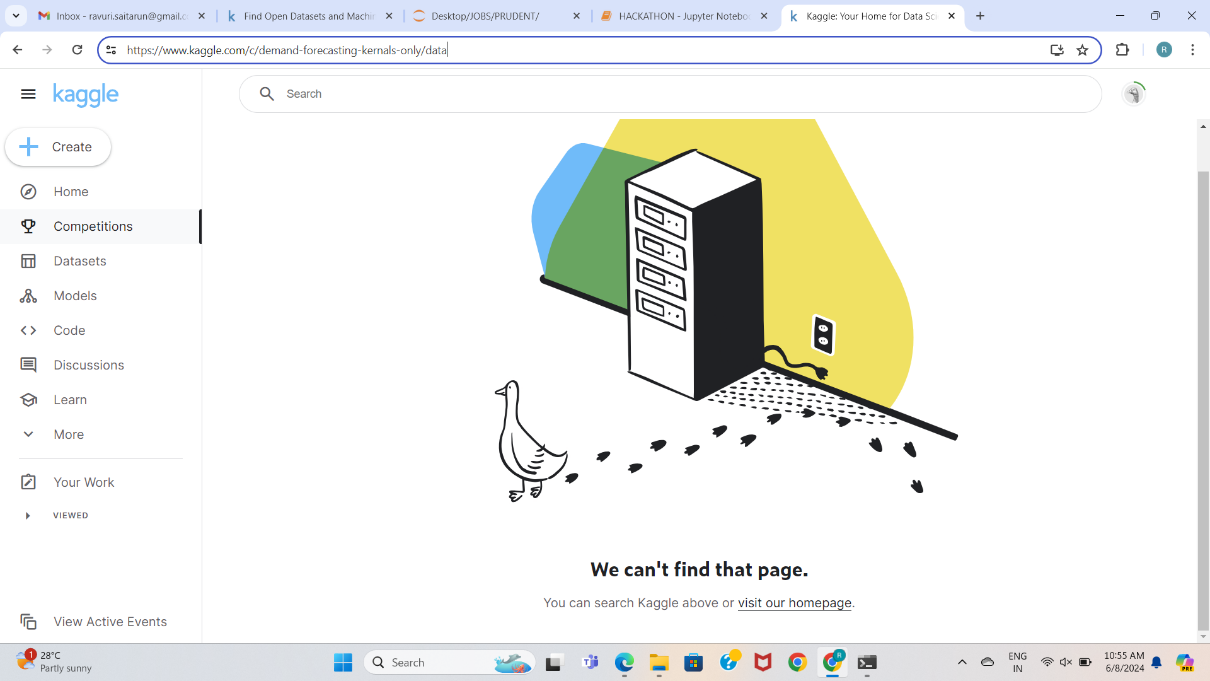
Platform: Jupyter Notebook

File name: Prudent\_Hackathon.ipynb

**1.Data Collection:**

Import all the required packages for the project.

As the dataset from the link provided by the company is not available on Kaggle I choose another dataset from Kaggle and performed the analysis and model building which consists of 5000rows and 14columns.



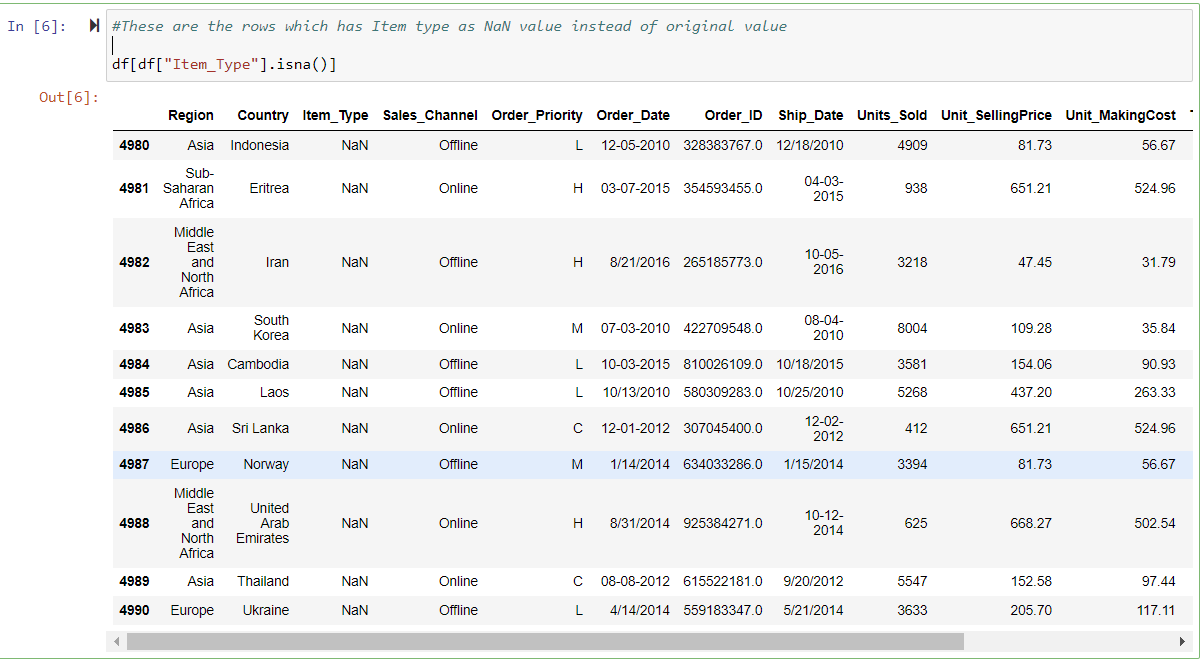
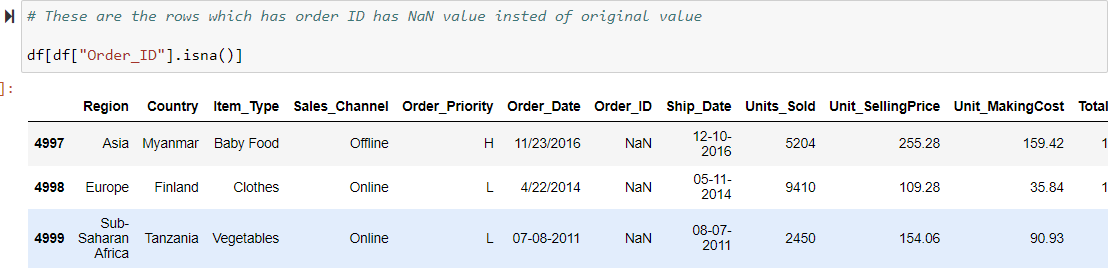
I also attached the dataset(.csv) file in the drive folder for your reference.

Reading the dataset(.csv) file by using pandas module and performing some basic operations like info, describe and column names. Where i found out that the data set is not processed and it has some null values.

**2.Data Cleaning (Preprocessing):**

1. Finding Null Values

Seperating the rows which has null values and droping the rows from the original dataset and preparing the final dataframe named df\_final.



After Cleaning the dataset the size of data is 4986rows and 14 columns.

1. Finding Duplicate Values

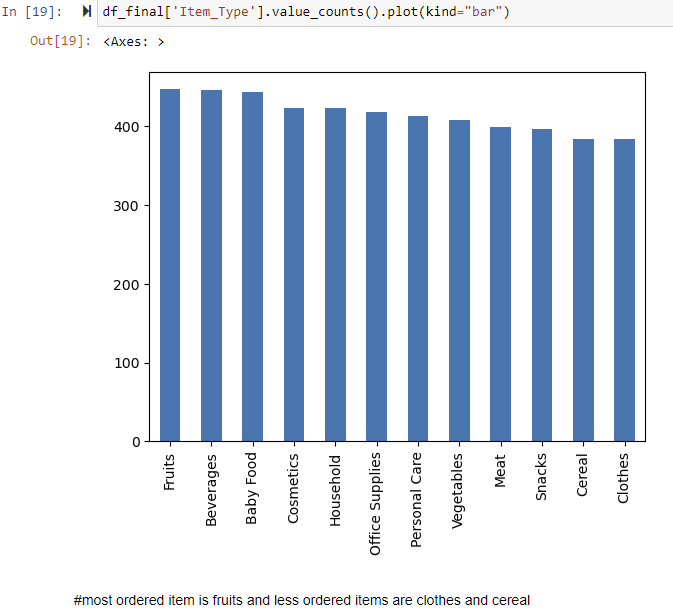
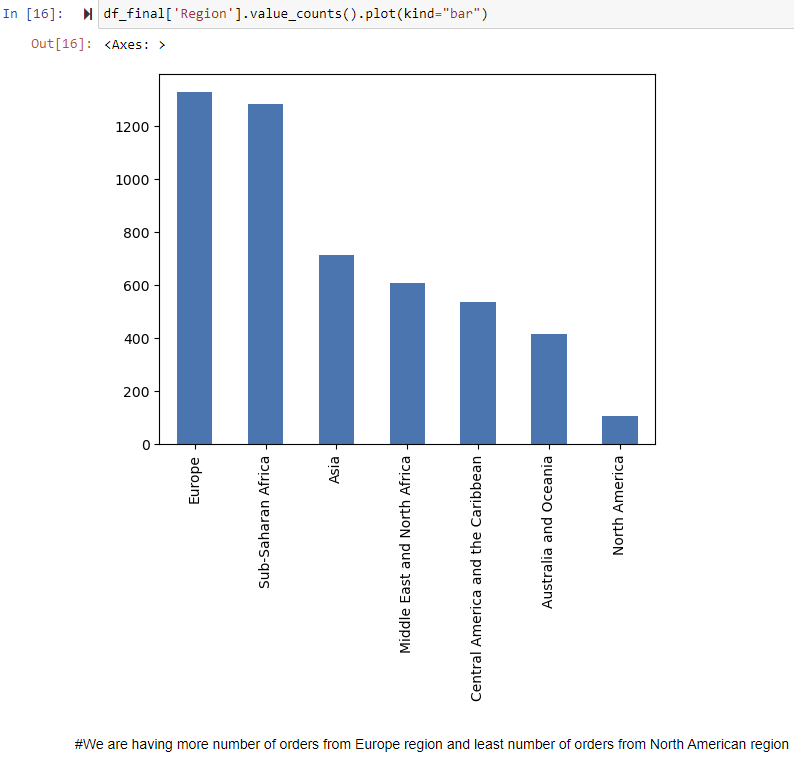
There are no duplicate values in the dataset

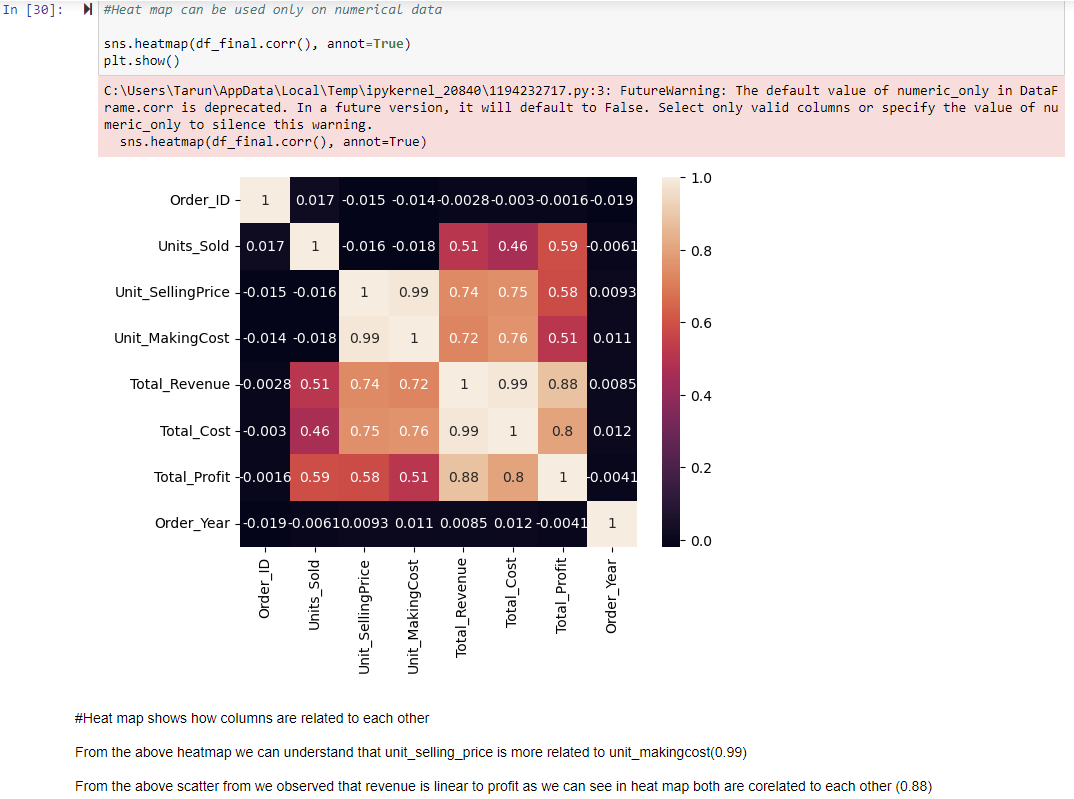
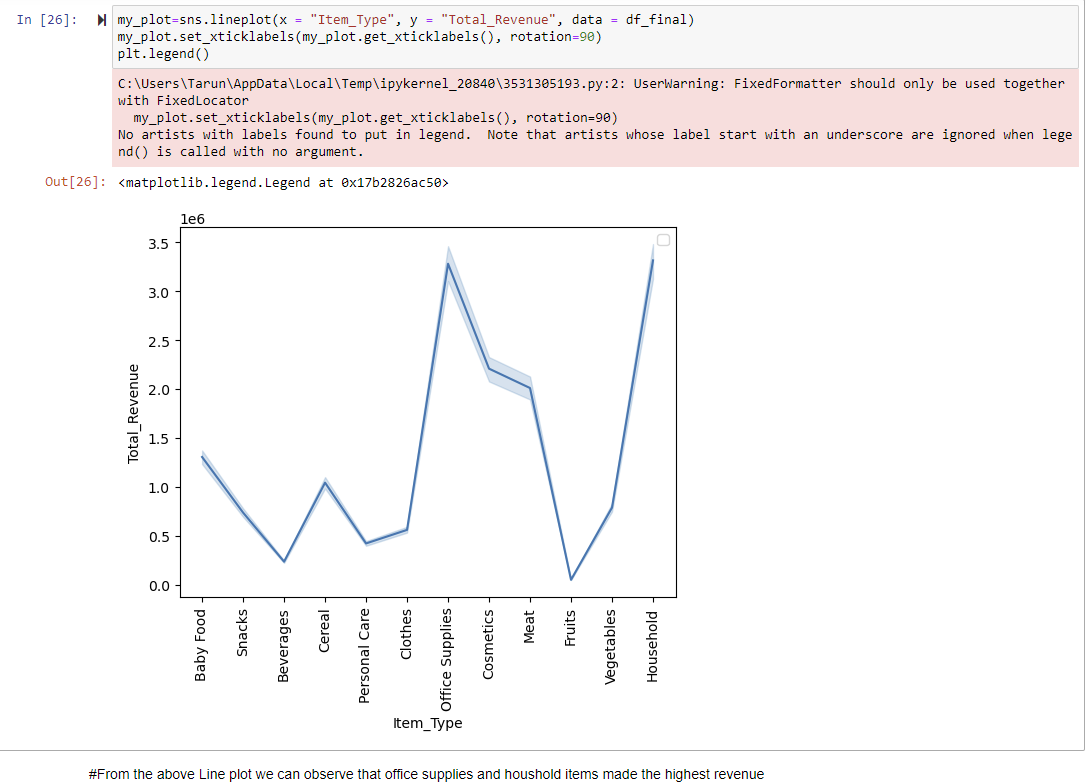
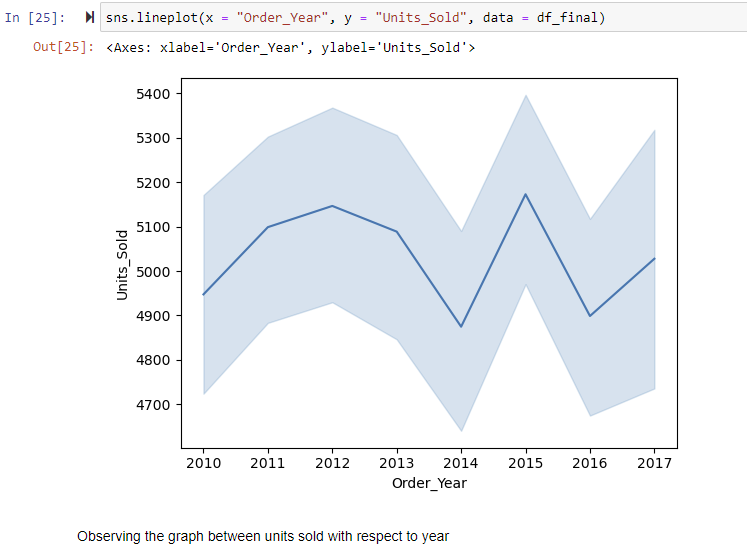
1. Handling Outliers

After the data cleaning step here comes the analysis part

**3.Exploratory Data Analysis (EDA):**

In this section the analysis comes in three parts

* Univariate Analysis : Analysing a single column
* Bivariate Analysis : Analysing two columns at a time
* MultiVariate Analysis : Analysis multiple columns at a time.



**4. Regression Model Building:**

Firstly we are dividing the final dataframe into two different dataframes which consists all the numerical data in on dataframe and categorical data into another dataframe

For training a model we need that dat only in numerical format that is why we divided in two seperate dataframes and then we convert the categorical dataframe into numerical with standard scalar ,onehotencoder and simple imputation methods and creating pipelines for easy use.

Then we combine both the dataframes with a pipeline and then train the model and complete the evaluation metrics

For building the model we split the data into 3 parts x\_train,x\_cv,x\_test

x\_train=training model

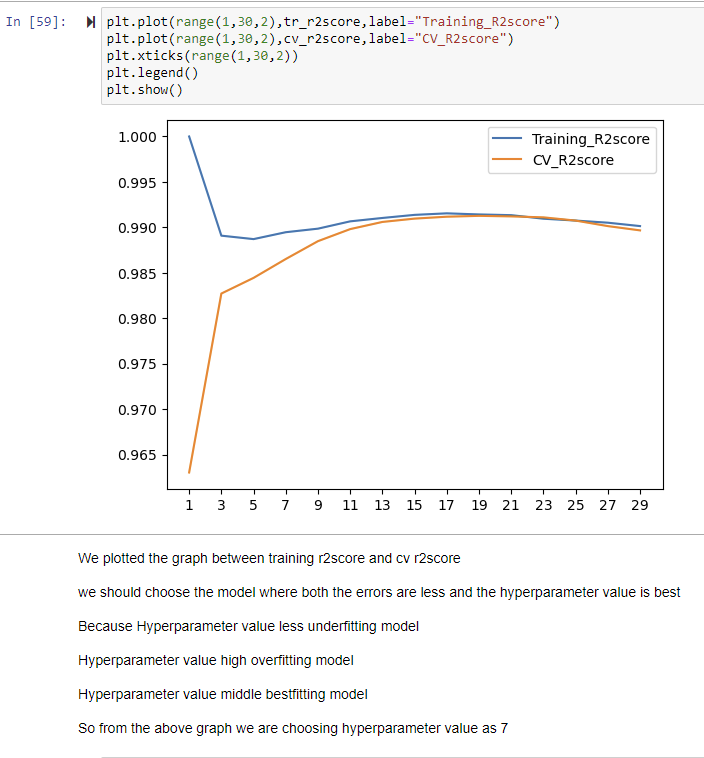
x\_cv=for testing performance of model

x\_test = for finding final model R2 Score

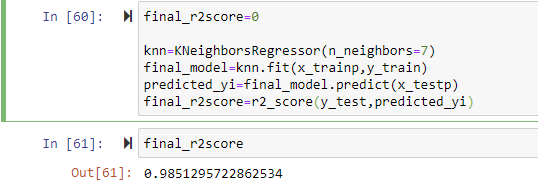
For finding the Genaralized R2 score the model shouldnt seen the data before for that we are splitting the data into two parts first x\_train and x\_test

Then we never use x\_test data till the final model r2 score calculation or best model

For finding best model we again split the x\_train data into x\_train and x\_cv

Ploting the graph between training error and cross validation error for finding the best hyperparameter value 

By selecting the hyperparameter value we are building the final model and finding its R2score



After the evalution metrics making the model ready for deploying by creating pickle files.