JOB-A-THON Report

DATA ANALYSIS AND PREDICTION REPORT

On the sales of WOMart

Prepared by: - Dipankar Medhi

Date: - 19th September

Overview

This is a detailed data analysis and Job-a-thon problem solving approach report of the sales of WOMart that includes all the information gained from the study of the dataset provided to us. This report also includes the method that has been used to solve the problem to obtain maximum accuracy score in predicting the target variables using the Test dataset.

Tools libraries used:

- Numpy
- Pandas
- Scikit-learn
- Matplotlib
- Seaborn
- Xgboost

Dataset

There are two dataset and both are csv (comma separate values) files – **Train.csv** and **Test_Final.csv**.

There are total **188340 rows** and **9 columns** (features) in the Train dataset and **22265 rows** and **7 columns** (features) in the Test dataset.

The features of Train dataset are –

'ID', 'Store_id', 'Store_Type', 'Location_Type', 'Region_Code', 'Date', 'Holiday', 'Discount', '#Order', 'Sales', 'Year', 'Month'.

The features of Test dataset are –

'ID', 'Store_id', 'Store_Type', 'Location_Type', 'Region_Code', 'Date', 'Holiday', 'Discount'.

Exploratory Data Analysis

The dataset is imported in the notebook using Pandas library.

Following observations were made upon analysing the dataset:

The distribution of sales throughout the given time-period.

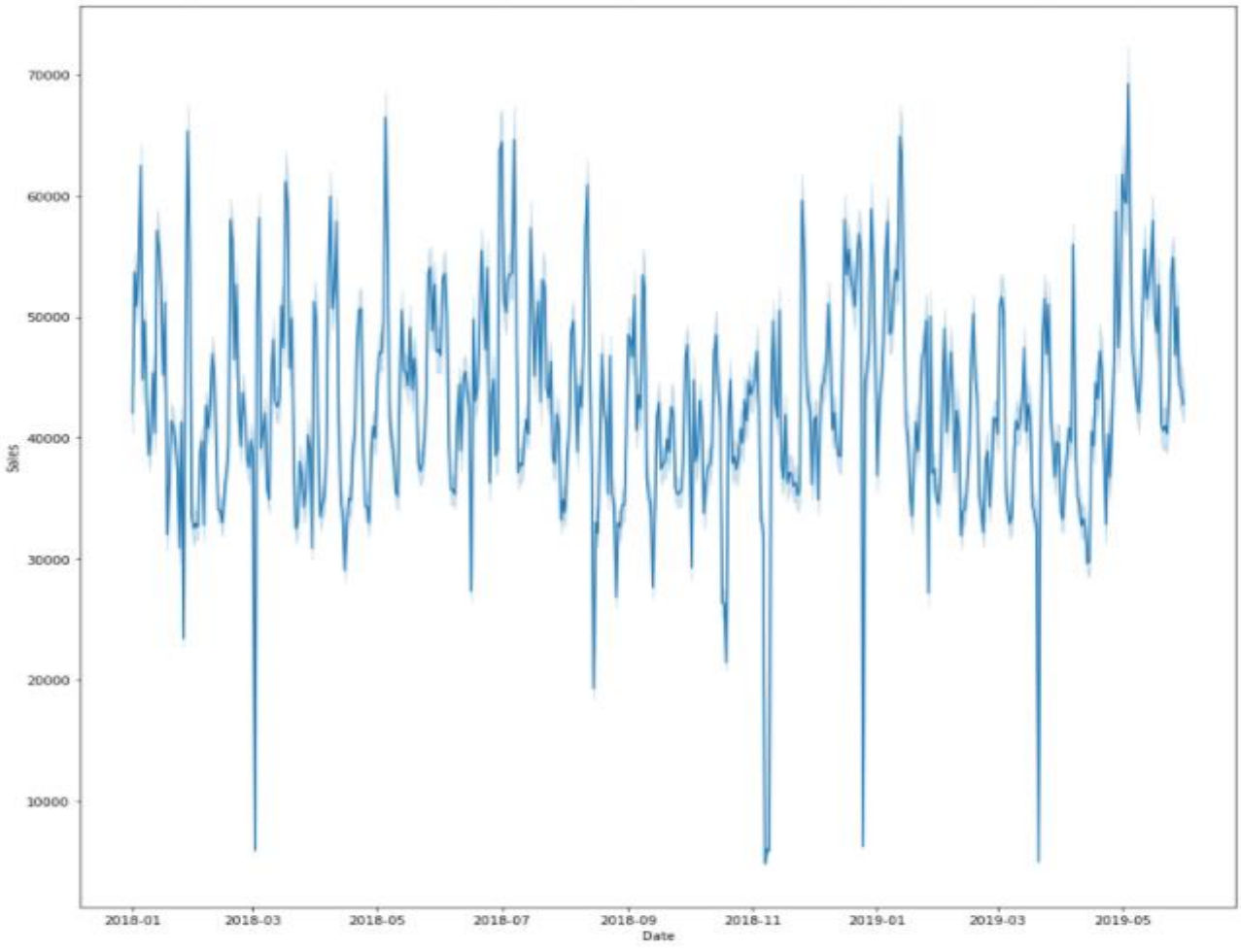


Fig1: sales vs date

The distribution of sales with respect to the number of days. We can see that during the initial days of a month, the sales are highest. The reason for this kind of behaviour is may be due to the fact that most of the people get their salaries at the end of the month and sometimes at the beginning of the month. So, they buy stuffs that they need when they have money.

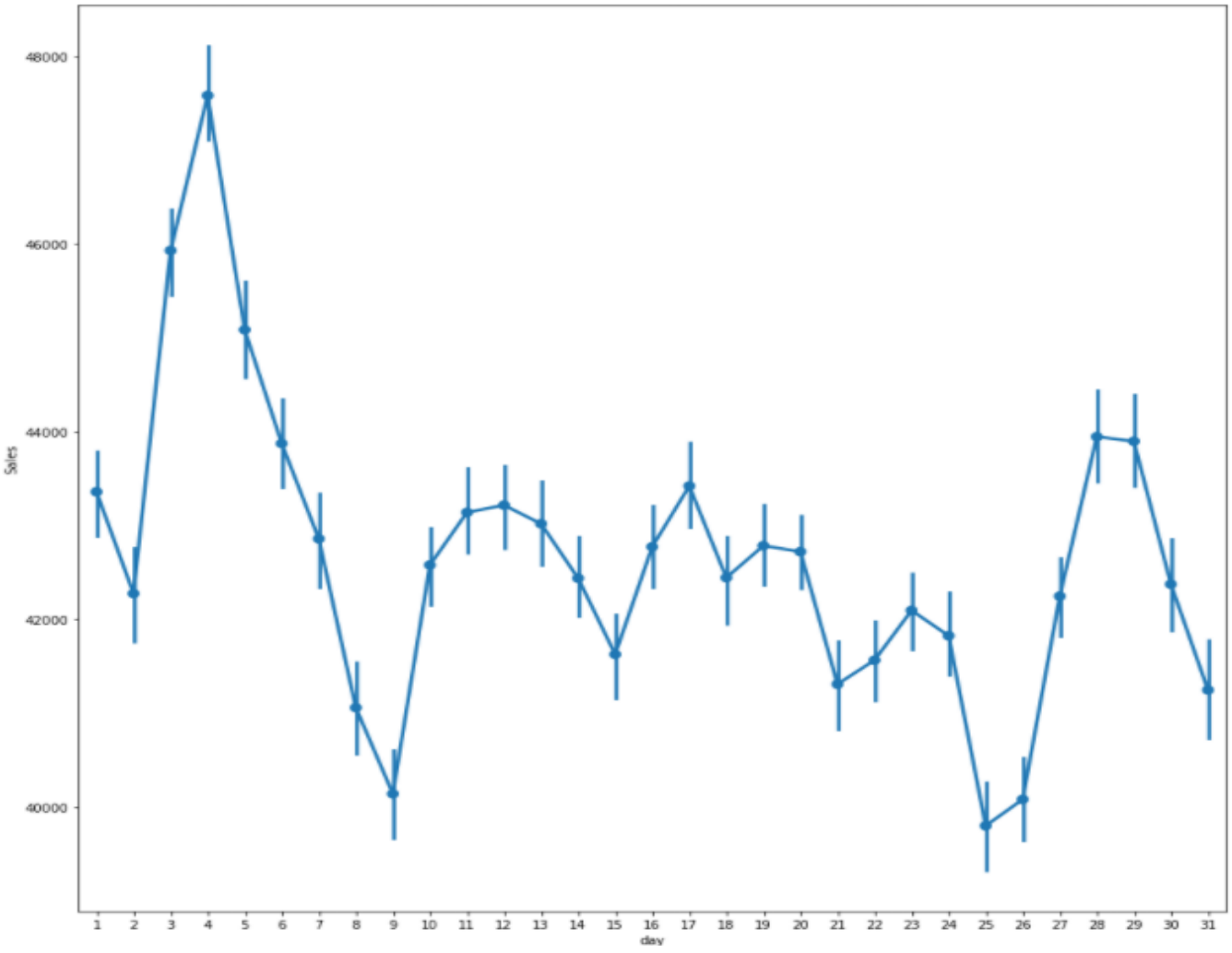
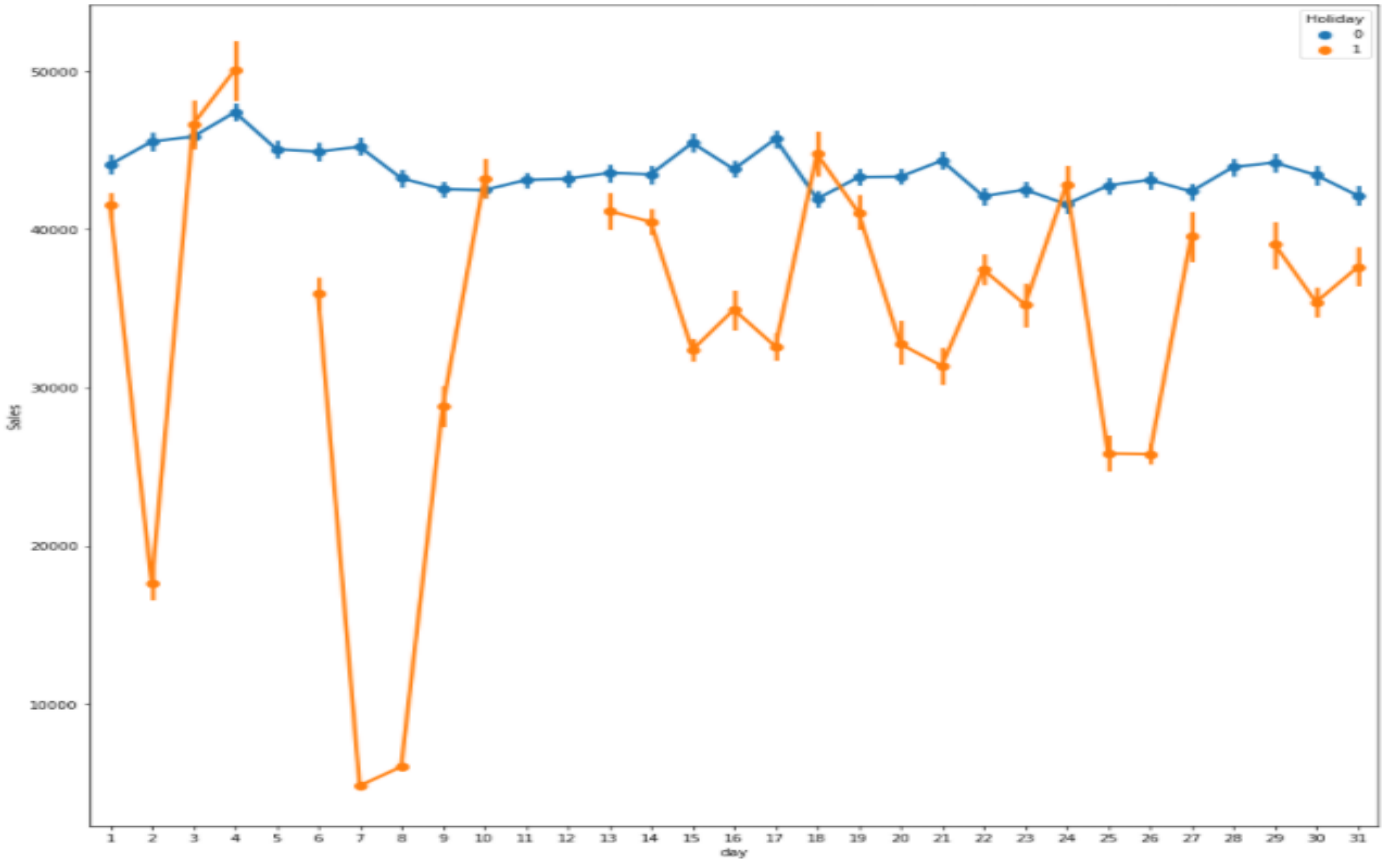


Fig2: sales vs day of month

Here we can see the difference of sales as per holidays and working days. During Holidays, the sales have a lot of variance but during the working days sales are mostly constant.

The orange line denotes the holiday sales and blue line denotes the working day sales. Working day sales mostly resides between 40000 to 50000 sales per day, which is quite good. But during holidays, the sales goes down. The reason may be that during holiday, people prefer to stay with their family at home or go for joy rides. This causes the sales to fluctuates. We can see that the maximum sales are during holidays, which means that when people go out with their families and friends during holidays, they spend more and buy more items from the shops.



Now let us plot and visualize the sales as per the given categorical features.

So, using group by method, the ‘Sales VS Store Type’ , ‘Sales VS Location type’ and ‘Sales VS Region code’ bar plots are plotted.

- Fig1: Sales vs Store type
- Fig2: Sales vs Location type
- Fig3: Sales vs Region code

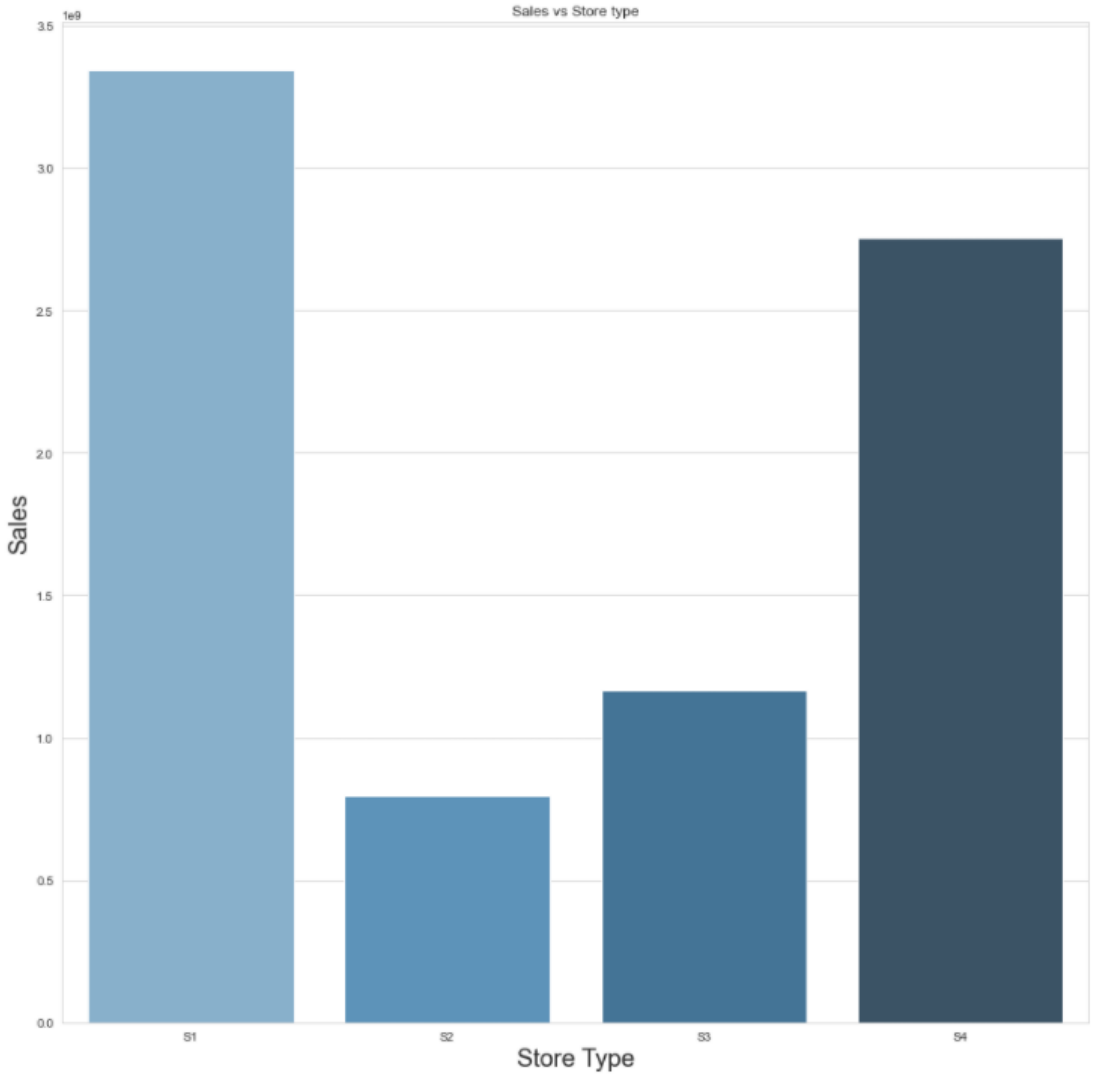


Fig1

From the above graph(fig1), we can see that store type S1 has the highest sales compared to other store types.

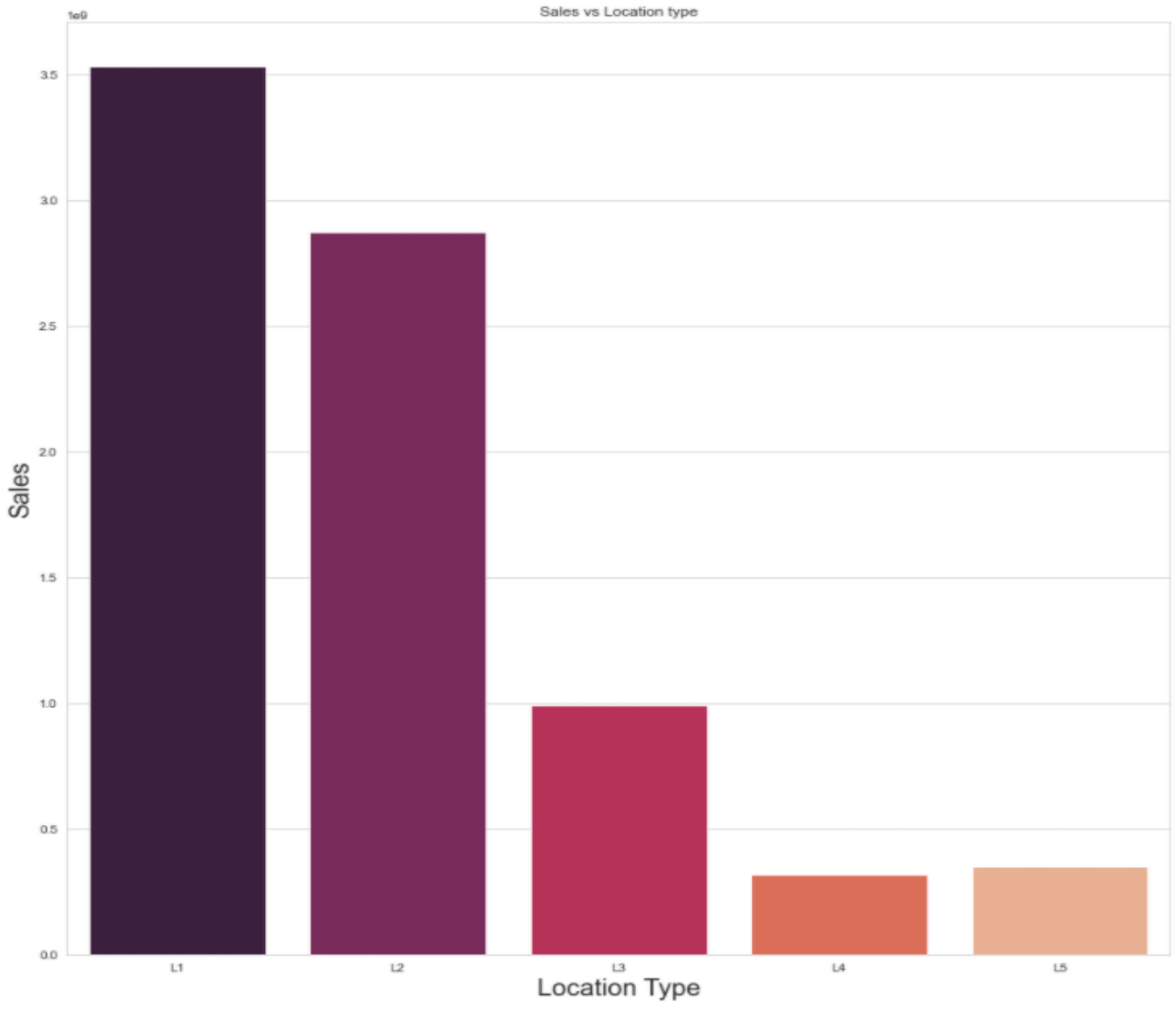


Fig2

From the above bar plot(fig2), it is clear that location L1 has the highest sales followed by L2, L3, L5 and L4 respectively.

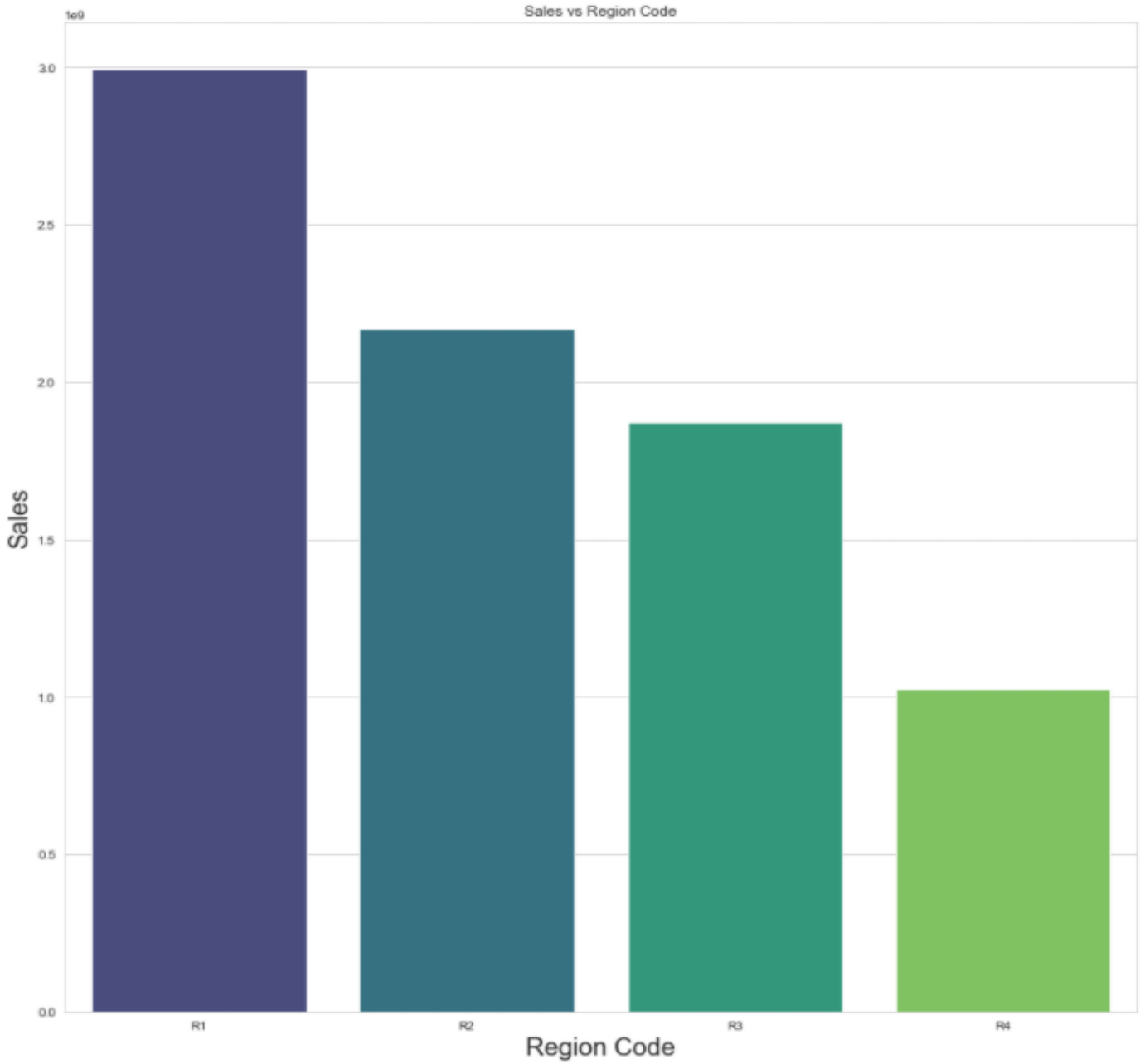


Fig3

From the fig3, we can say that region R1 has the highest sales followed by R2, R3 and R4 respectively.

Data Preprocessing

Before jumping straight into model training and prediction it is very important to pre-process the dataset. Pre-processing includes choosing the appropriate features for training and handling missing values and outliers. It also includes encoding the categorical features into numeric features.

Luckily there are no missing data in our both the training and testing dataset.

The outliers are handled using the quantile method by ignoring or removing the extreme values from the dataset. (More detailed approach is present on the notebook)

After handling the missing values and dealing with the outliers, the next step is to encode the categorical features.

The categorical features are handled using LabelEncoder method from scikit-learn and pd.dummies method from Pandas.

Feature Engineering:

Then, comes the feature engineering. It is done to select the necessary as appropriate features for training the model.

In this case, the ‘Date’ column is divided into separate columns/features of ‘Year’ and ‘Month’. Doing this will help in proper training of the model.

Also, we can see that there is no ‘#Order’ column in the test dataset. But upon correlation of the features, it is been found that orders and sales highly correlate. So first it is important to predict the number of future orders for proper and effective prediction of the number of sales.

Model used: XGBoost (based on Gradient Boosting technique)

Using XGBoost, the future sales are predicted. These sales are then concatenated into the testing dataset for the final sales predictions.

Model Training and Evaluation

Now for the final training, the train dataset is split into training and testing dataset for model training and evaluation using train_test_split method from scikit-learn library.

Then with XGBoost model, the sales are predicted.

The final R squared score obtained: 0.956367350603446

The mean squared error obtained: 14337205.958033673

Conclusion

In this case general regression method is used for predicting the sales. But with Time series analysis more accurate sales values may be predicted.