CECS 551 Advanced Artificial Intelligence

Final Project

Fall 2022

Artificial Intelligence (AI) approach in Retail Market Analysis and Growth

Group 1

11.04.2022

Supervisors
Dr. Mahshid Fardadi
Dr. Allen Bolourchi

FALL 2022

Contents

Abstract

In Retail Industry and chain of stores one of the biggest issues they face are supply chain management. The component of supply chain management (SCM) involved with determining how best to fulfill the requirements created from the Demand Plan.

Its objective is to balance supply and demand in a manner that achieves the financial and service objectives of the enterprise.

If we investigate the case of a retail chain stores one of the basic cases is to know the demand of products that are sold in the store. If the decision-making authority know what's the demand of each product for a week or month, they would be able to plan the supply chain accordingly. If that is possible this would save a lot of money for them because they don't have to overstock or can plan their Logistics accordingly.

The dataset deals with an international retail business which has 30 stores spread across many geographical locations. The dataset is designed to mimic a real tech company software department and machine learning environment.

We have visualized the sales figures and sales pattern against many different features, to analyze significant factors that contribute to the change in the weekly sales across all stores.

1 Introduction

In the team of six, we divided the task into two team, each having 3 individuals. Team 1 worked on data visualization of dataset1 and plotting the heatmap, correlation matrix etc., while team 2 worked on dataset 2 plotting and visualizing the dataset and creating the tableau dashboard.

1.1 Problem Statement

The final project is designed to implement three-week sprints of the scrum process, mimicking a real tech company machine learning or software development team environment. The dataset describes the weekly sales of 35 stores across all the departments. Different exploratory data analysis methods have been applied to visualize the data. This data visualization will help in prediction of sales resulting in a more efficient operation.

1.2 Dataset Description

The table below describes the features of the dataset for retail store. We will be considering only 10 stores for the initial analysis.

Features	Description
Store	The number of stores date - MMDDYYY format
Temperature	Temperature in Fahrenheit
Gas price	Price per gallon
Discounts	discounts discount clearance
CPI	The Consumer Price Index (CPI)
Unemployment	Unemployment rate in the region where store is present
IsHoliday	Yes or No

Table 1: Store Feature data description.

1.3 Proposed Workflow

We perform the analysis in three phases.

1. Exploratory Data analysis.
2. Create a machine leaning model for sales prediction.
3. Inference and recommendation to maximize the profit.

2. Exploratory Data Analysis (Dataset_01)

Each phase is covered in individual chapters. The first phase tires to understand the data by visualizing it and defining relationship among them using correlation matrix and SHAP feature interactions.

- Rank the features based on their influence on weekly sales, identify where/when the sales are affected most by the feature and perform what-if analysis
- Understand if the seasonal change impacts the sales of certain stores.
- Draw conclusions and suggest a recommendation to optimize the sales

Data

Store: The store number. Range from 1–45.
Type: Three types of stores 'A', 'B' or 'C'.
Size: Sets the size of a Store would be calculated by the no. of products available in the store ranging from 34,000 to 210,000.

We will perform detailed EDA and gather useful insights

2.1 Identifying key variables for the model using correlation plots, heatmaps, histograms, feature importance (SHAP)

2.1.1 Heatmaps:

The heatmap above depicts the analysis by visually representing the data between store, dept, Weekly_Sales, and IsHoliday. From the heat map, we can see from the pattern in cell colors across weekly shows that sales are maximum in Dept during weekly sales, and least in store during weekly sales.

As we can observe from the heat map, weekly_sales are closely related to size and discounts and loosely correlated to all the features, i.e., store, dept and IsHoliday.

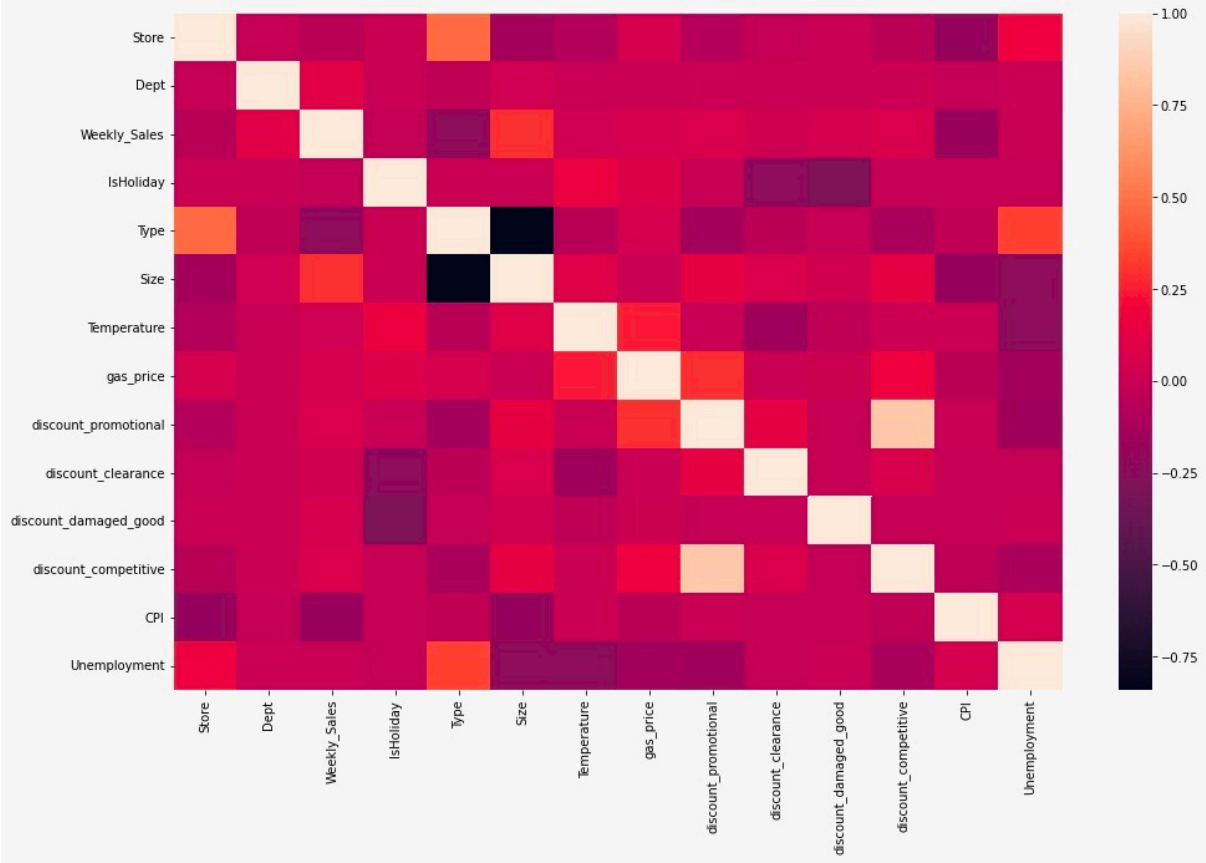

Store
Dept
Weekly_Sales
IsHoliday
Type
Size
Temperature
gas_price
discount_promotional
discount_clearance
discount_damaged_good
discount_competitive
CPI
Unemployment
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

Fig 1:

2.1.2 Histogram:

A histogram is basically used to represent data provided in a form of some groups. It is an accurate method for the graphical representation of numerical data distribution. In the below histogram the bar represents the frequency of store, department, size, temperature, gas_price, discount and weekly sales in the dataset.

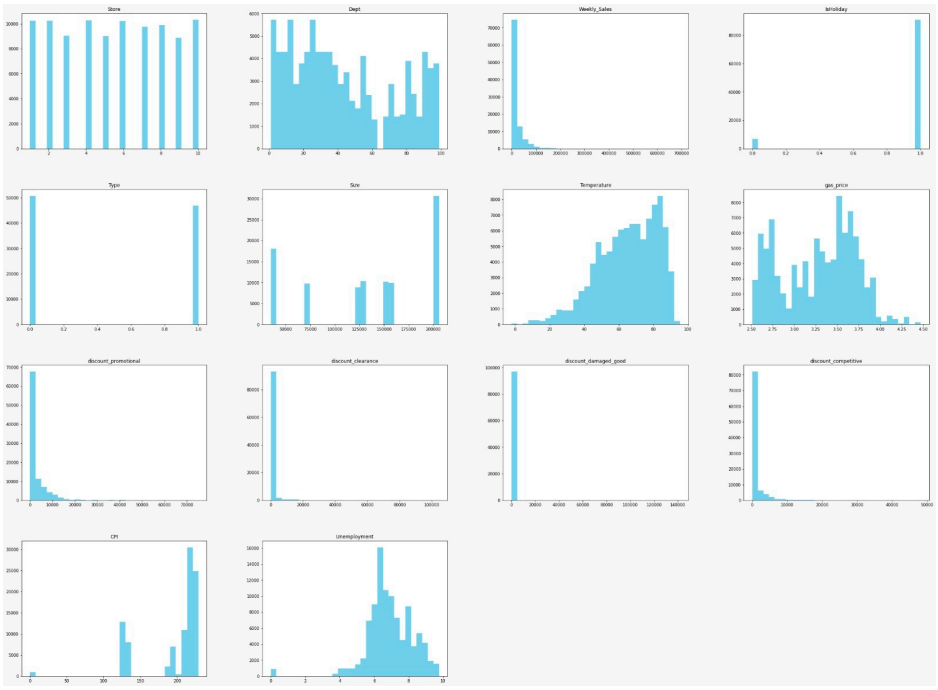


Fig 2:

2.1.3 SHAP:

The below plot shows the importance of the following features.

As visible from the graph unemployment, CPI, discounts is a feature of high importance, also we can see that the department is also important.

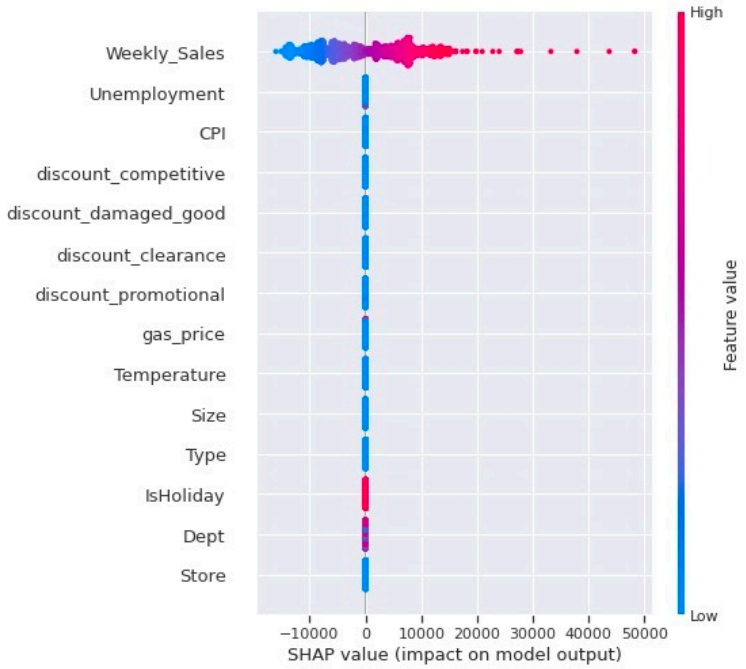


Fig 3: SHAP Value Graph

2.2 **Top 35% of the department sales for the first 10 stores**

2.2.1. Identify the best department across the first ten stores.

The below image represents the top 35 departments for the first ten stores along with the total sales across all weeks. As seen, department 38 is the best department as it has the highest sales total across all weeks.

	Dept	Weekly_Sales
36	38	1.063737e+08
74	95	1.052439e+08
71	92	1.037589e+08
58	72	8.489495e+07
38	40	7.399350e+07

Fig 4: Best department across first 10 stores

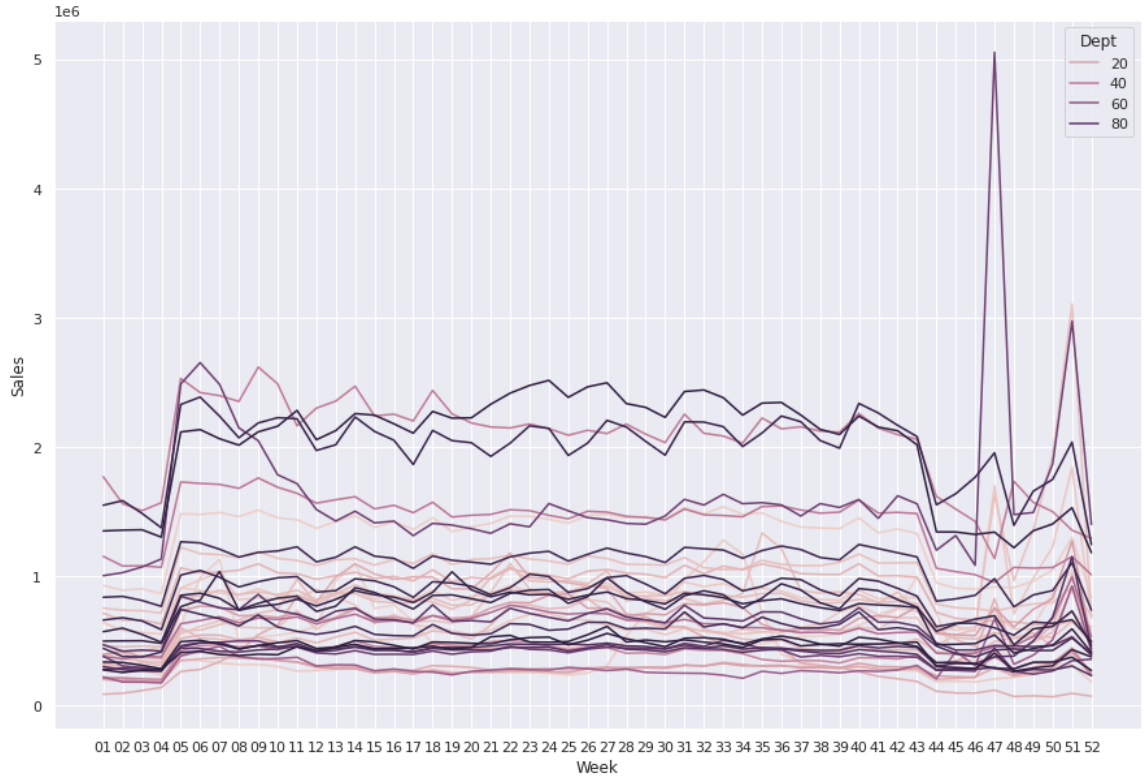


Fig 5: Weekly sales across all weeks

The above graph represents the weekly sales pattern for the first ten stores for the top 35 departments across all weeks.

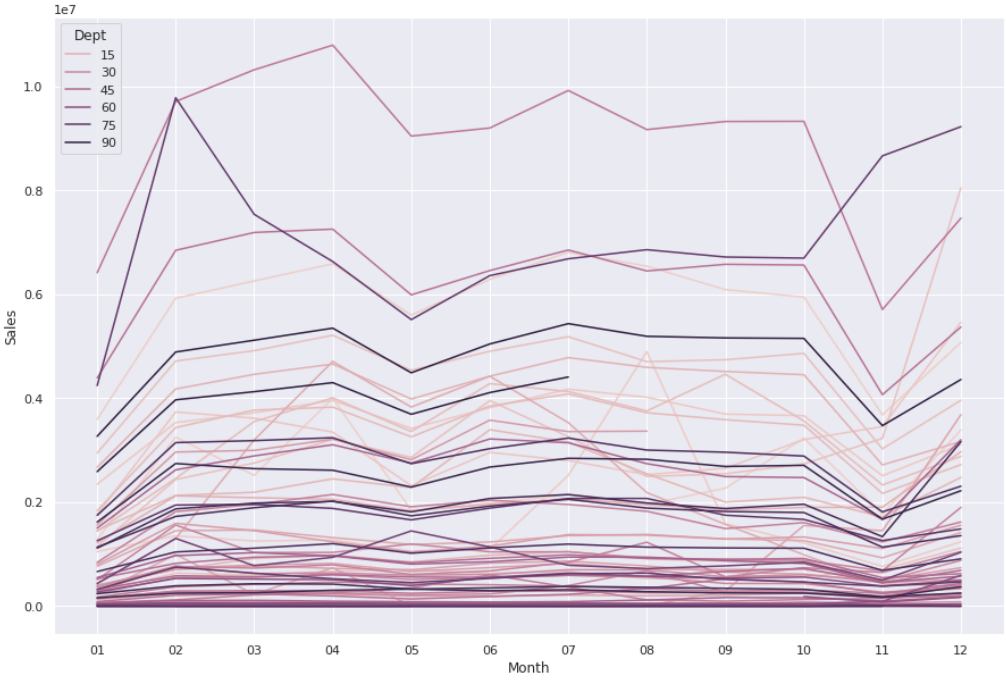


Fig 6: Weekly sales across all months

The above graph represents the weekly sales pattern for the first ten stores for the top 35 departments across all months.

2.3 Relationship between weekly sales over CPI and unemployment for first 10 stores

The below diagram represents the correlation heatmap of the first 10 stores. We can see that weekly sales are loosely correlated to CPI and unemployment.

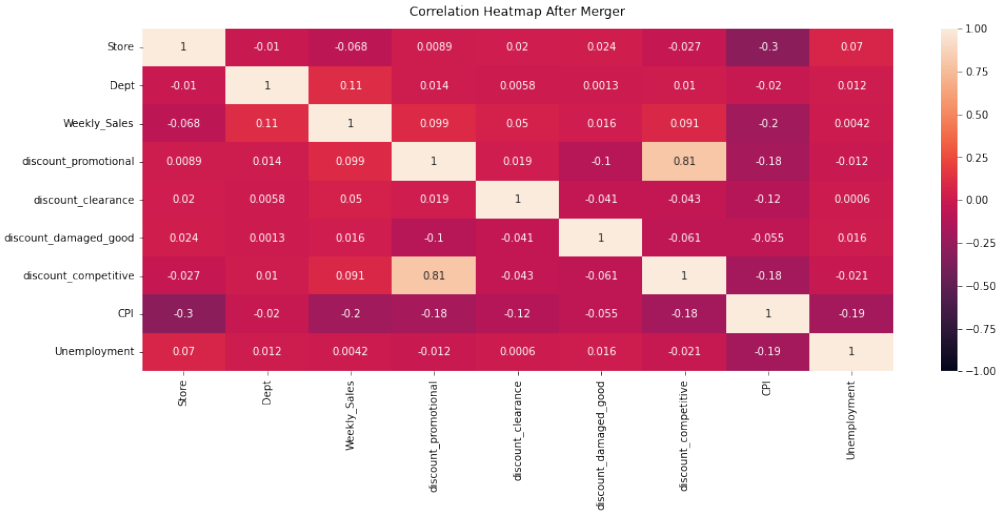


Figure 7: Heatmap for various features

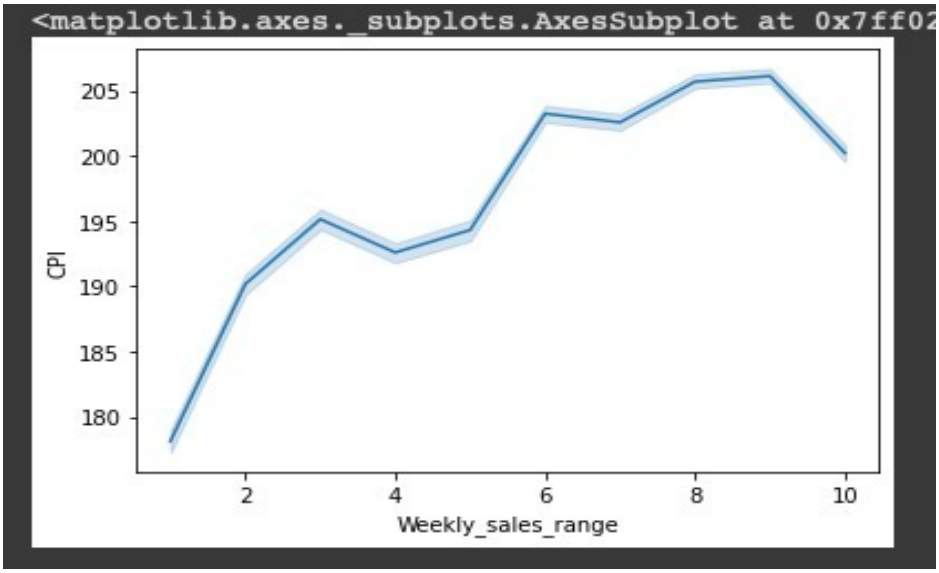


Figure 8: Line plot of CPI

The above graph shows changes in CPI with respect to weekly sales ranges.

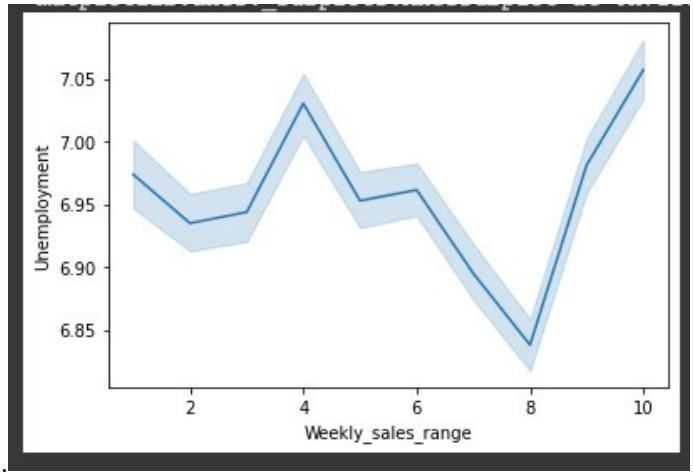


Figure 9: Line plot of Unemployment

The above graph shows changes in unemployment with respect to weekly sales ranges.

2.4 Investigate the impact of various types of discounts

2.4.1. Type of discount helpful in increasing the sales, considering the top 30% of the best performing stores (sales per 1000 square feet).

- Discount Promotional (DP) is the type of discount which is helpful in increasing the sales of the top 30% of best performing stores per 1000 square feet. Discount promotion plays a vital role in increasing the sales of stores across multiple departments and DP is directly proportional to the weekly sales of the top performing stores.
- It is important to understand the relationship of discount promotion, discount clearance, discount damaged goods and discount competitive vs weekly sales. The correlation coefficient is a statistical measure that quantifies the relationship between two variables, seaborn in python provides an option of correlation coefficient to draw relation between variables.

- Correlation map: From the below heatmap, it can be inferred that weekly sales of top 30% stores is directly proportional to discount promotional and possess a strong relationship. Discount competitive has the second highest coefficient measure with weekly sales, followed by discount clearance and discount damaged goods.

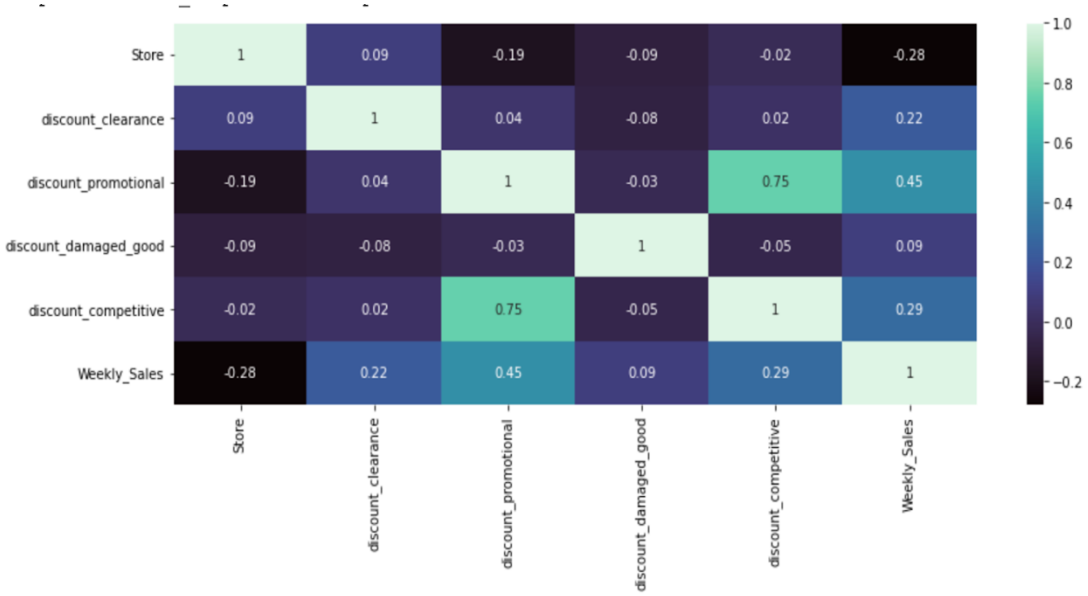


Figure 10: Heatmap correlation of discount vs weekly sales

From the above heatmap we can draw strong correlation of promotional discount, clearance discount and competitive discount.

- **Bar graph**: The graph below shows correlation of different discounts to total weekly sales for top 30% stores per 1000 square feet. It can be inferred from the graph that by providing more promotional discounts on the stores, the weekly sales of the stores grow linearly.

Figure 11: Bar graph of Sales vs Different discounts for top 30% stores

2.4.2. Considering the bottom 30% of the least performing store (sales per 1000 square feet) does the observed behavior hold true for all stores

- The observed behavior is true for all the stores when considering bottom 30% of the stores that Discount promotional(DP) is strongly co related when the DP increases the weekly sales also increases.
- Correlation map:From the below heatmap, it can be inferred that weekly sales of bottom 30% stores is directly proportional to discount promotional and possess a strong relationship. Discount competitive has the second highest coefficient measure with weekly sales, followed by discount clearance and discount damaged goods.

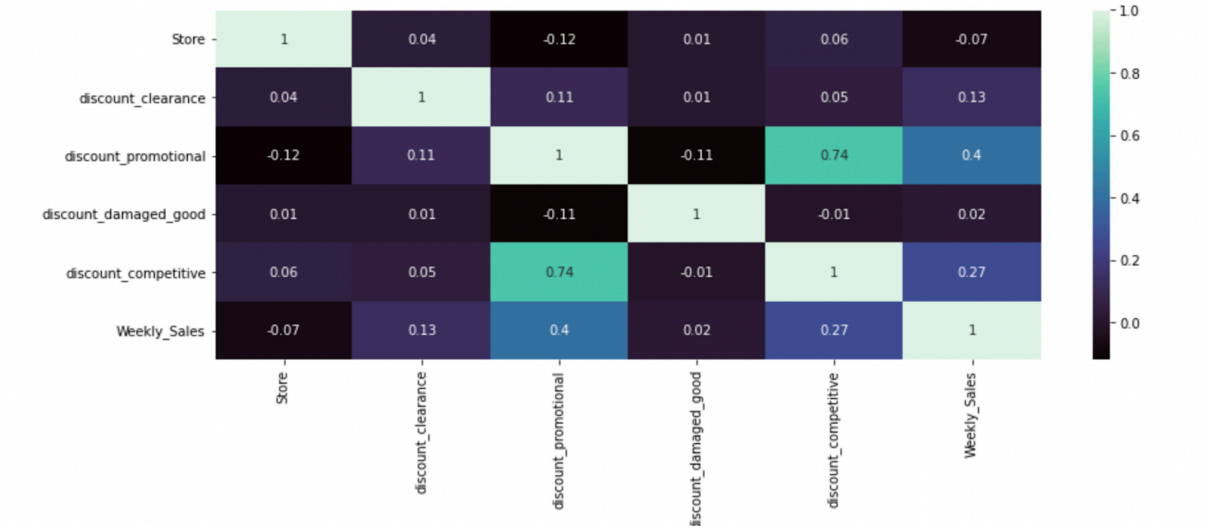


Figure 12: Estimated Loss of Sales in Volume

- **Bar graph:** The graph below shows correlation of different discounts to total weekly sales for bottom 30% stores per 1000 square feet. It can be inferred from the graph that by providing more promotional discounts on the stores, the weekly sales of the stores grow linearly.

Figure 13: Sales vs Different discounts for top 30%

2.5 Determining if there is any correlation between overall sales and holiday

After merging data for the first 35 stores and 10 departments in train.csv with stores_features files to get the following data frame.

	Store	Dept	Date	Weekly_Sales	IsHoliday_x	Temperature	gas_price
332179	35	10	2012-08-24	13727.58	False	72.93	3.834
332180	35	10	2012-08-31	13173.20	False	75.00	3.867
332181	35	10	2012-09-07	14630.52	True	76.00	3.911
332182	35	10	2012-09-14	12744.10	False	68.72	3.948
332183	35	10	2012-09-21	14567.14	False	66.10	4.038
332184	35	10	2012-09-28	14057.75	False	64.92	3.997
332185	35	10	2012-10-05	14270.55	False	64.50	3.985
332186	35	10	2012-10-12	14821.31	False	55.40	4.000
332187	35	10	2012-10-19	15732.96	False	56.53	3.969
332188	35	10	2012-10-26	15071.72	False	58.99	3.882

Figure 14:

Weekly sales range from [-4988.94: 406988.63]. It was divided into 10 equal groups to get a weekly sales range. From the following graph we can see that for different departments the trends of weekly sales range in comparison to the temperature. For

example, in Department 06 weekly sales increases with the increase in temperature whereas for Department 10 for weekly sales increases with decrease in temperature.

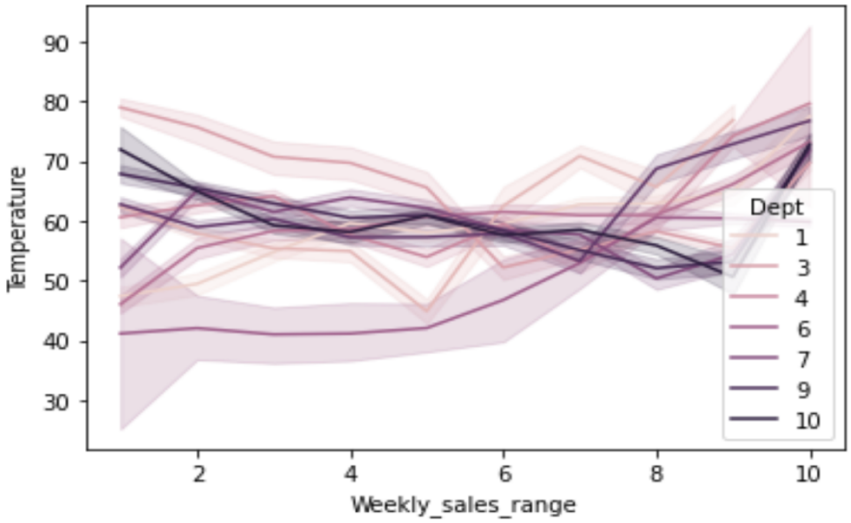


Fig 15: Weekly_sales_range vs Temperature Graph

In the graph below for different departments the trends of weekly sales range in comparison to the gas prices is shown. For example, in Department 01 weekly sales increases with the increase in gas prices whereas for Department 10 for weekly sales increases with decrease in gas prices.

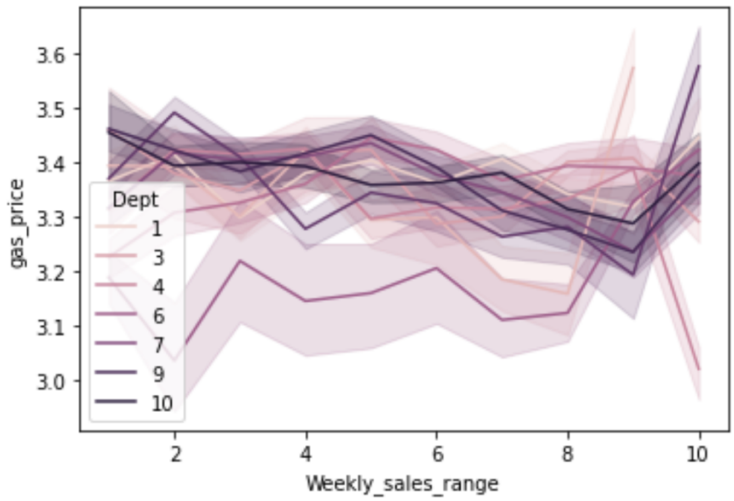


Fig 16: Weekly_sales_range vs gas price graph

For Holidays we can see that overall sales vary more Department range (30 - 50).

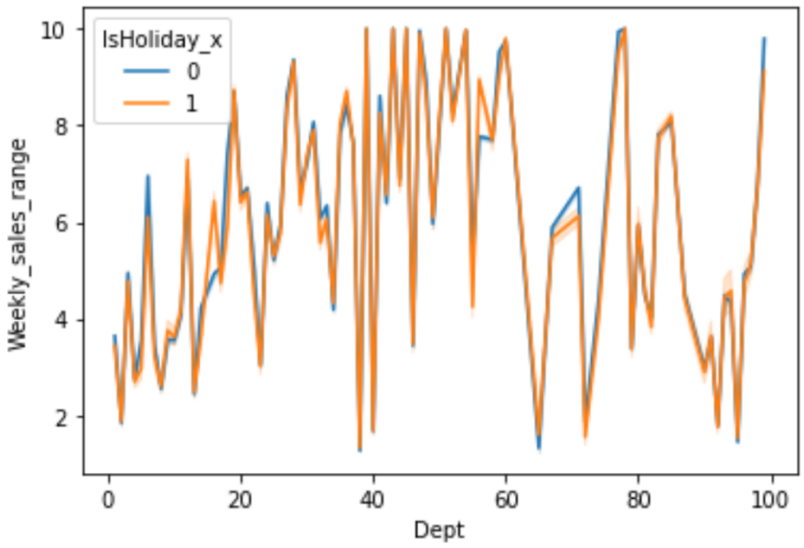


Fig 17: Dept vs weekly_sales_range graph

2.6 Data preprocessing:

Data preprocessing is a step in the data mining and data analysis process that takes raw data and transforms it into a format that can be understood and analyzed by computers and machine learning.

Data cleaning:

3 data frames are created because we are handling missing values with different techniques and creating a model for the same and comparing the output of the models.

Techniques:

1. Replace missing and NaN values with median value.
2. Filling discount NaN values with zero in the entire data frame.
3. Splitting date column into day, month and year
4. Creating a Total additional discount by adding all the discount columns

Since the date column Since the date column is in string format, typecasting it to date format and divide them to individual columns for day, month and year.

One hot Encoding is performed on columns such as **type**, **department** and **store** as it is necessary to convert the categorical data variables to be provided to machine and deep learning algorithms which in turn improve predictions as well as classification accuracy of a model.

How weekly sales are related to other features:

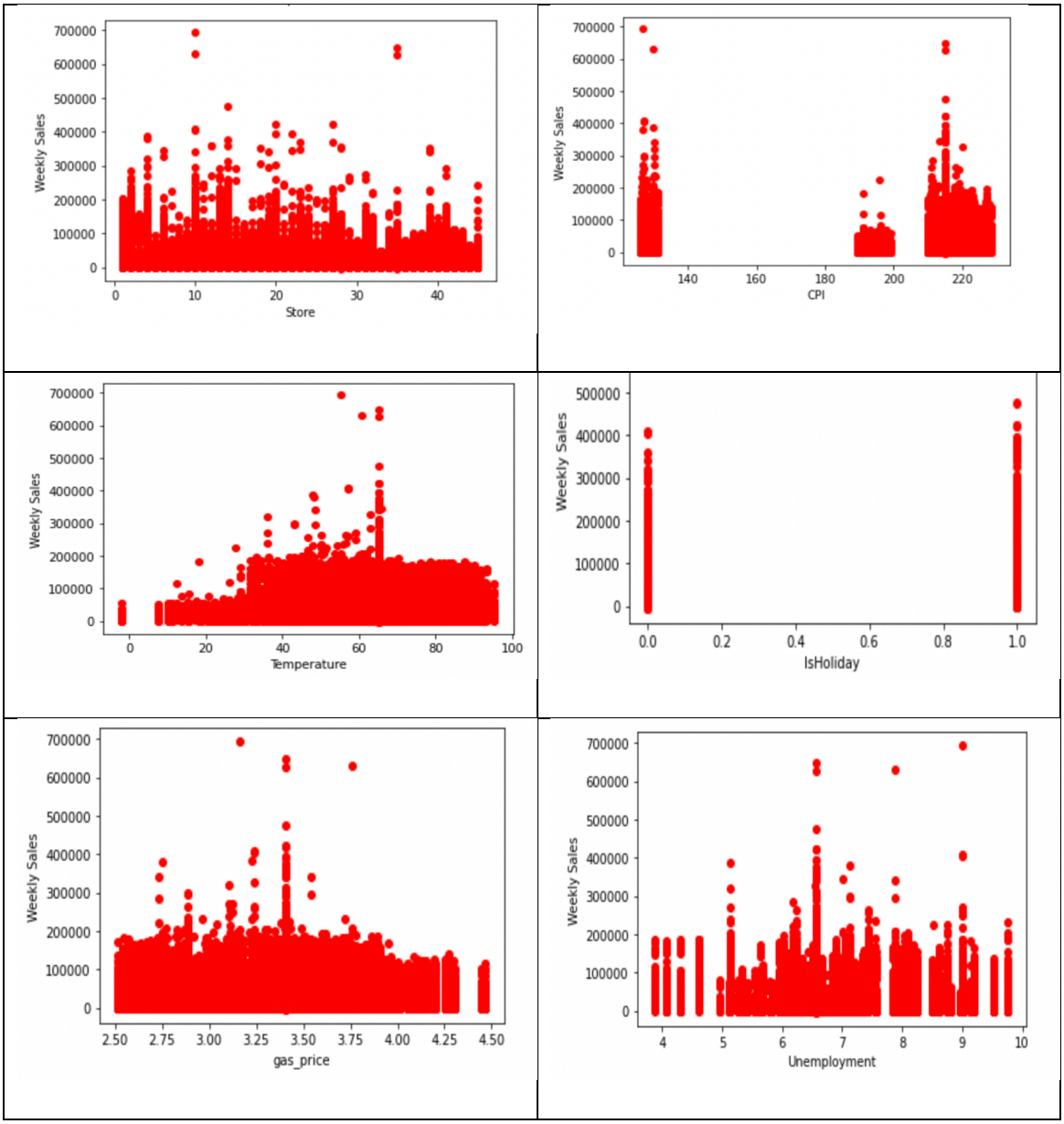


Fig 18:

Key observations from the above graphs:

- When the gas_price is in the range of 2.5 - 3.75 there is more sales, for >3.75 the sales is comparatively lower.
- Type A and Type B have more weekly sales than type C stores.
- There are more sales when there is a holiday than non-holiday days.
- When the temperature is 30 units of temperature there is less sales than >30 units of temperature.
- Highest sale is for the store whose size is 125000 sq ft.

Data Normalization:

Data Scaling is a data preprocessing step for numerical features. Many machine learning algorithms like KNN algorithm, linear and logistic regression, etc. require data scaling to produce good results.

MinMax Scaler:

In this method the features are made equal to zero and the maximum of features equal to one. MinMax Scaler shrinks the data within the given range, usually of 0 to 1. It transforms data by scaling features to a given range.

Below is the image after normalizing the data using MinMax Scaler.

Date	Weekly_Sales	IsHoliday	Size	Temperature	gas_price	CPI	Unemployment	Year	Month	Week	...	Dept_93	Dept_94	Dept_95	Dept_96	Dept_97	Dept_9
2010-02-05	0.035961	1	0.630267	0.630158	0.432958	0.950609	0.501869	2010	2	5	...	0	0	0	0	0	
2010-02-05	0.000964	1	0.492338	0.411312	0.029683	0.862911	0.430853	2010	2	5	...	0	0	0	0	1	
2010-02-05	0.001001	1	0.492338	0.411312	0.029683	0.862911	0.430853	2010	2	5	...	0	0	0	0	0	
2010-02-05	0.012487	1	0.650636	0.371587	0.029683	0.861118	0.411145	2010	2	5	...	0	0	0	0	0	
2010-02-05	0.016049	1	0.492338	0.411312	0.029683	0.862911	0.430853	2010	2	5	...	0	0	0	0	0	

5 rows x 145 columns

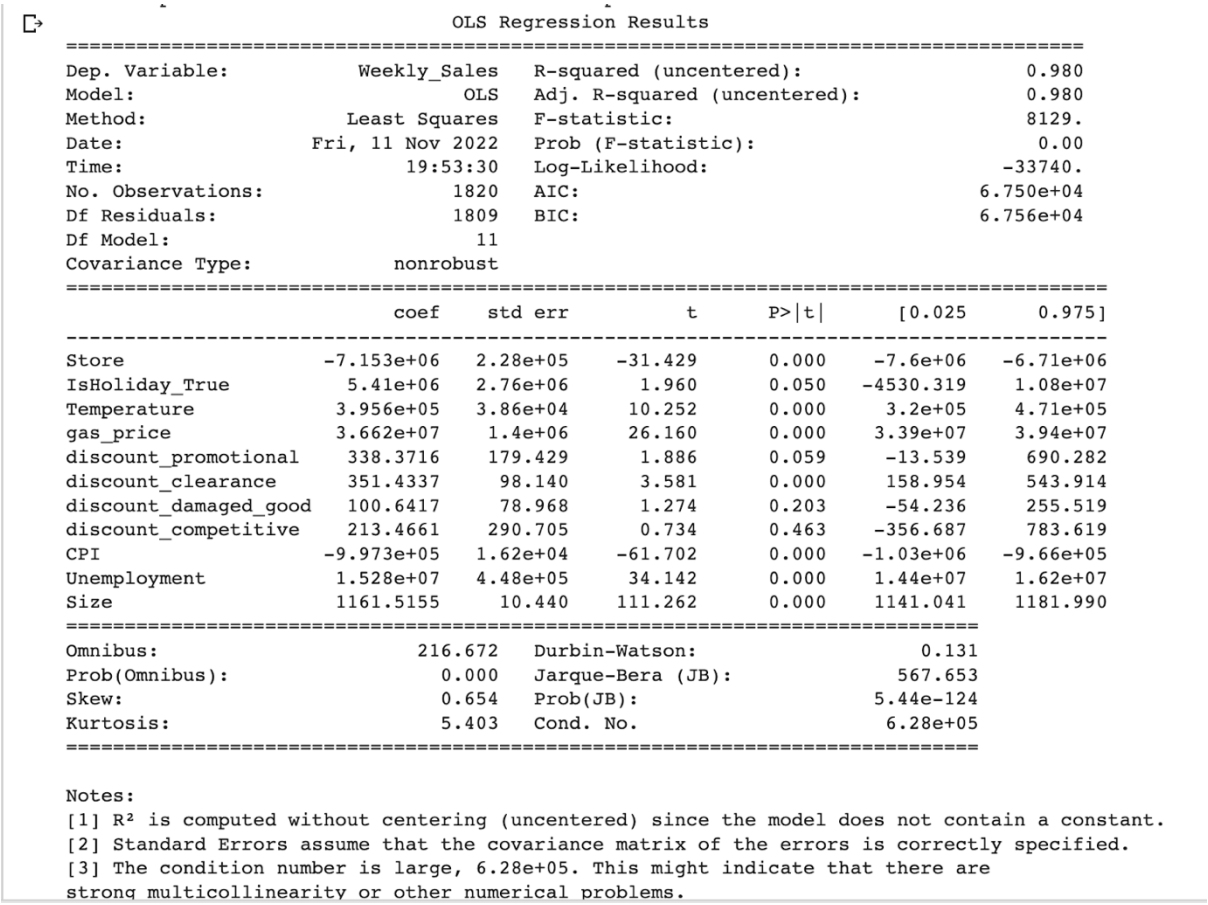
Fig 19:

2.7 Feature Selection:

Feature Selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data. It is the process of automatically choosing relevant features for your machine learning model based on the type of problem you are trying to solve.

Method 1: OLS

The ordinary least squares (OLS) method is a linear regression technique that is used to estimate the unknown parameters in a model. The method relies on minimizing the sum of squared residuals between the actual and predicted values. The OLS method can be used to find the best-fit line for data by minimizing the sum of squared errors or residuals between the actual and predicted values.



OLS Regression Results

Dep. Variable:	Weekly_Sales	R-squared (uncentered):	0.980
Model:	OLS	Adj. R-squared (uncentered):	0.980
Method:	Least Squares	F-statistic:	8129.
Date:	Fri, 11 Nov 2022	Prob (F-statistic):	0.00
Time:	19:53:30	Log-Likelihood:	-33740.
No. Observations:	1820	AIC:	6.750e+04
Df Residuals:	1809	BIC:	6.756e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Store	-7.153e+06	2.28e+05	-31.429	0.000	-7.6e+06	-6.71e+06
IsHoliday_True	5.41e+06	2.76e+06	1.960	0.050	-4530.319	1.08e+07
Temperature	3.956e+05	3.86e+04	10.252	0.000	3.2e+05	4.71e+05
gas_price	3.662e+07	1.4e+06	26.160	0.000	3.39e+07	3.94e+07
discount_promotional	338.3716	179.429	1.886	0.059	-13.539	690.282
discount_clearance	351.4337	98.140	3.581	0.000	158.954	543.914
discount_damaged_good	100.6417	78.968	1.274	0.203	-54.236	255.519
discount_competitive	213.4661	290.705	0.734	0.463	-356.687	783.619
CPI	-9.973e+05	1.62e+04	-61.702	0.000	-1.03e+06	-9.66e+05
Unemployment	1.528e+07	4.48e+05	34.142	0.000	1.44e+07	1.62e+07
Size	1161.5155	10.440	111.262	0.000	1141.041	1181.990

Omnibus:	216.672	Durbin-Watson:	0.131
Prob(Omnibus):	0.000	Jarque-Bera (JB):	567.653
Skew:	0.654	Prob(JB):	5.44e-124
Kurtosis:	5.403	Cond. No.	6.28e+05

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[3] The condition number is large, 6.28e+05. This might indicate that there are strong multicollinearity or other numerical problems.

Fig 20:

Results after OLS:

After 8 iterations of OLS fitting the mode, the linear regression model achieved an accuracy of 75.28.

Method 2: PCA

Principal Component Regression (PCR) is a regression technique that serves the same goal as standard linear regression — model the relationship between a target variable and the predictor variables.
The idea is that the smaller number of principal components represents most of the variability in the data and (presumptively) the relationship with the target variable. Therefore, instead of using all the original features for regression, we only utilize a subset of the principal components.

Cross-validation and visual analysis are generally used to calculate the number of main components (k). K is essentially a hyperparameter that needs to be tuned. We evaluate the RMSE scores that occur as we iterate over an increasing number of principal components to include in regression modeling. The training set performance of the PCR increases with more principal components, as expected, when we look at the plot of training set cross-validation RMSE vs. the number of principal components employed. The baseline standard linear regression model with all of the original features' RMSE benchmark is represented by the green line.

After determining the best number of principal components to use (i.e., M=17), we proceed to run PCR on our test dataset. Accuracy 78.02 % after performing PCR was not satisfactory.

Method 3: RandomForest Regressor

Feature selection using random forest regressor technique improves performance, reduces overfitting and increases interpretability. It is an embedded technique that is a combination of filter and wrapper methods. A random extraction of the observations from the dataset and a random extraction of the features are used to build each of the 4–12 hundred decision trees that make up a random forest.

Every tree also consists of a series of yes-or-no questions depending on a single or several attributes. The dataset is split into two buckets at each node (i.e., each question) by the three, with each bucket containing observations that are more similar to one another and distinct from those in the other bucket. The significance of each feature is therefore determined by how "pure" each of the buckets is. For regression the measure of impurity is variance.

	rank	feature	importance
0	1	mean	9.132935e-01
1	2	Week	3.065128e-02
2	3	min	1.926021e-02
3	4	Month	7.084516e-03
4	5	Size	3.576628e-03
...	...	...	...
128	129	Dept_51	1.958067e-10
129	130	Dept_45	1.483400e-10
130	131	Dept_78	6.163958e-12
131	132	Dept_39	4.336655e-13
132	133	Dept_43	2.364089e-16

133 rows × 3 columns

Fig 21:

Linear Regression:

A variable's value can be predicted using linear regression analysis based on the value of another variable. The dependent variable is the one you want to be able to forecast. The independent variable is the one you're using to make a prediction about the value of the other variable.
Following scatter plot shows actual vs predicted output after Linear Regression:

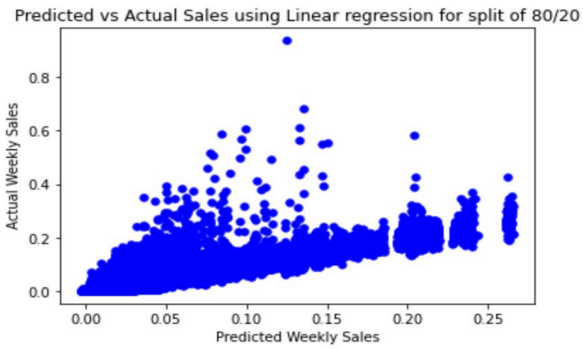


Fig 22:

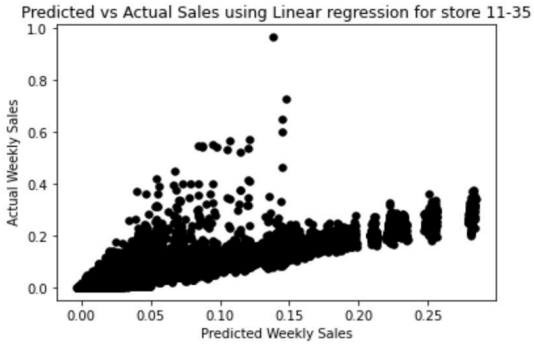


Fig 23:

We get best accuracy with linear regression after feature selection with random forest regressor:

Approach 1: 80 - 20 Test Train split on Stores 1 – 35

Approach	**Accuracy**
Linear Regression	91.098
Random Forest Regression	96.759
Ridge Regression	91.091
XGBoost	96.945

Approach 2: Train on Store 1- 10 and Test on Store 11- 35

Approach	**Accuracy**
Linear Regression	89.379

Random Forest Regression	95.701
Ridge Regression	89.368
XGBoost	97.146

ARIMA:

ARIMA is an acronym that stands for AutoRegressive Integrated Moving Average. It is a class of models that captures a suite of different standard temporal structures in time series data.

Here we will be performing the below steps to calculate time series.

1. Visualize the weekly series

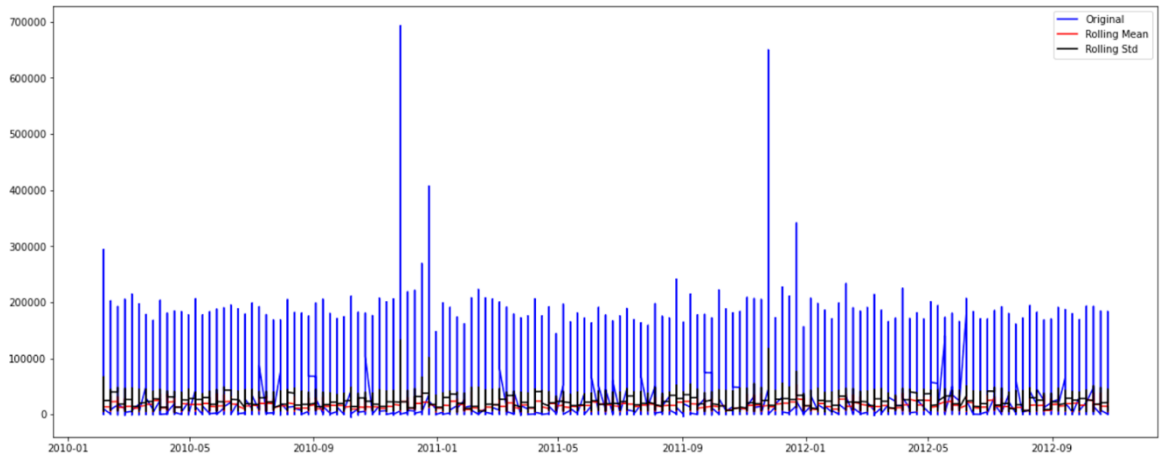


Fig 24:

2. Make sure the variance of weekly sales is non stationary
3. Test stationary of data using ADF Test

```
Test Statistic                -50.864267
p-value                         0.000000
#Lags Used                     97.000000
No of Observesations Used    421472.000000
Critical Value (1%)             -3.430366
Critical Value (5%)             -2.861547
Critical Value (10%)            -2.566774
dtype: float64
```

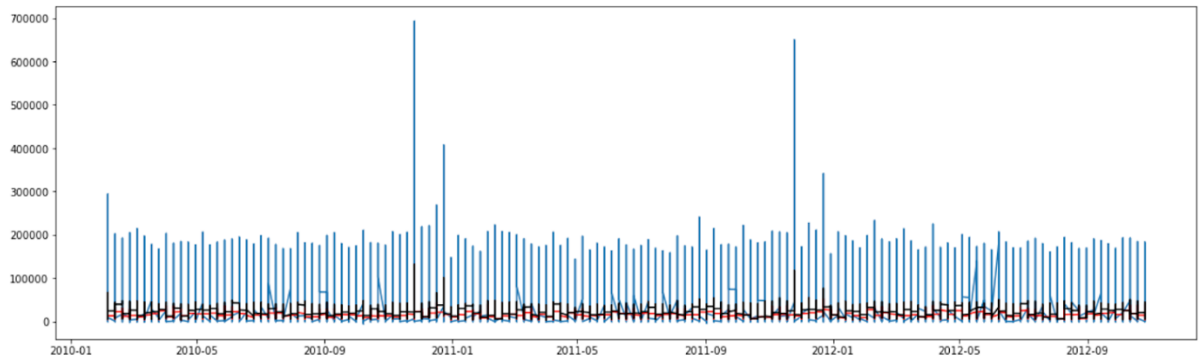
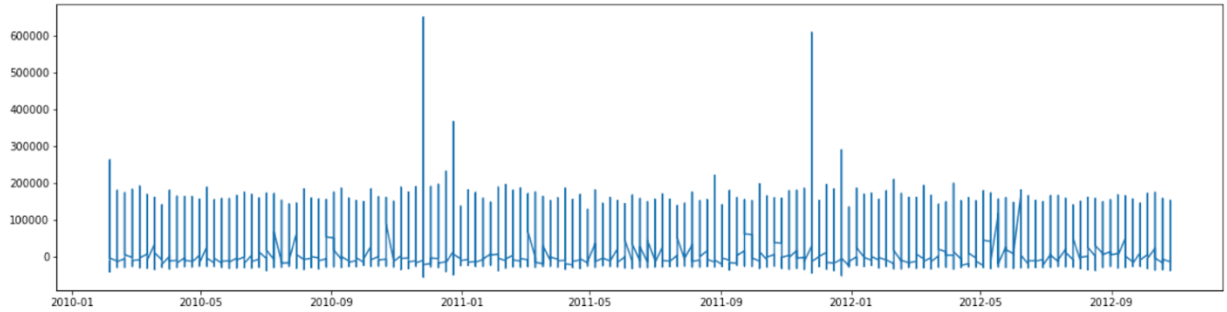


Fig 25:

Weak evidence against null hypothesis, time series has a unit root, indicating it is non-stationary.




```
Test Statistic               -101.91134
p-value                         0.00000
#Lags Used                     97.00000
No of Observesations Used    421472.00000
Critical Value (1%)             -3.43037
Critical Value (5%)             -2.86155
Critical Value (10%)            -2.56677
dtype: float64
```

Fig 26:

Strong evidence against the null hypothesis, reject the null hypothesis. Data has no unit root and is stationary.

4. Convert data to stationary using first order differencing and calculate correlation function.

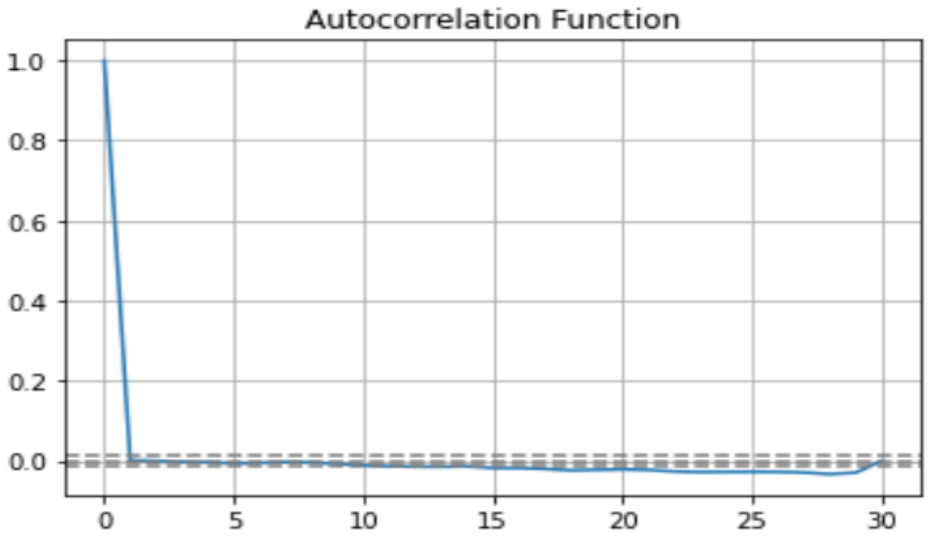


Fig 27:

5. Predict weekly sales over a time period. Here we are predicting weekly sales from May 2010 to Sept 2010 the blue mark shows forecast vs actual values in blue.

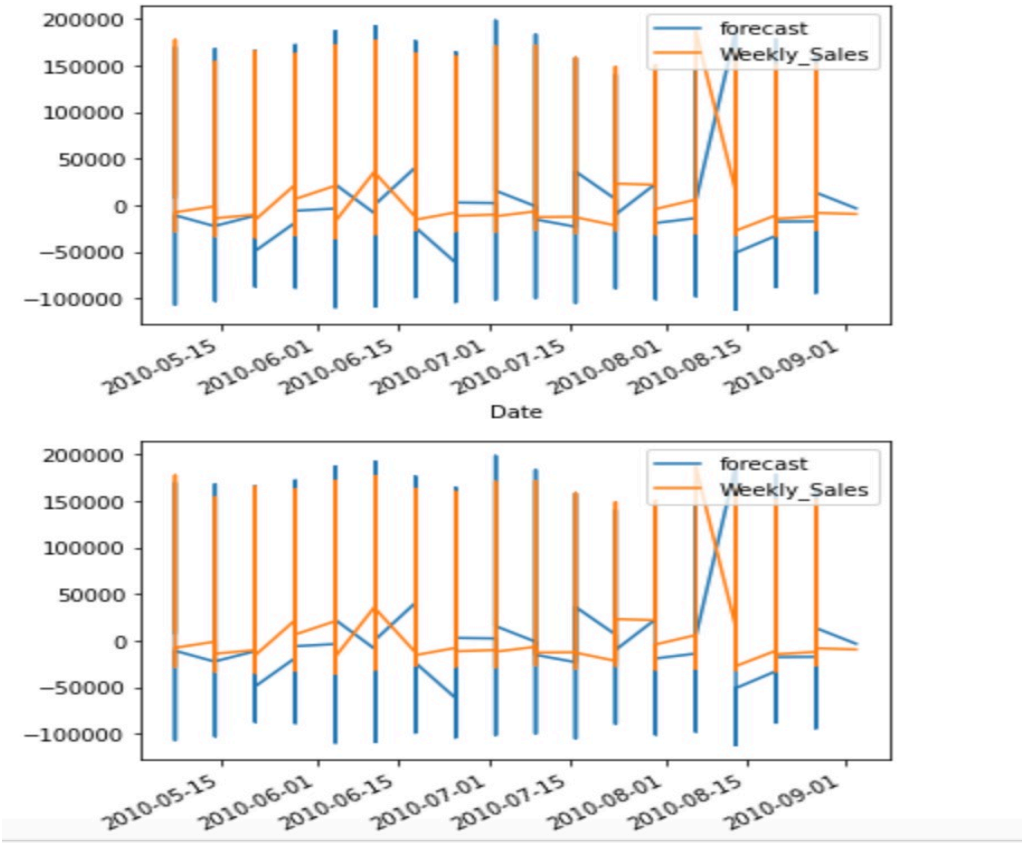


Fig 28:

2.8 Ensemble modeling

Ensemble Modeling is the process of running two or more related but different analytical models and then synthesizing the results into a single score or spread in order to improve the accuracy of predictive analytics and data mining applications.

2.8.1 Classification and Prediction

The problem required us to merge the following dataframes

Train
Stores
Store_feature

train_data_expanded

	Store	Dept	IsHoliday	Type	Size	datetime	Temperature	gas_price	discount_promotional	discount_clearance	discount_damaged_good	discount_competitive	CPI	Unemployment
0	1	1	1	0	151315	2010-02-05	59.33	3.360	9667.50	268.29	0.60	8368.15	223.659114	6.833
1	1	1	0	0	151315	2010-02-12	51.65	3.409	8687.47	1594.87	2.20	2144.87	223.753643	6.833
2	1	1	1	0	151315	2010-02-19	52.39	3.510	2706.87	3128.74	1.88	2396.68	223.917015	6.833
3	1	1	1	0	151315	2010-02-26	60.12	3.555	6129.28	1802.84	0.00	301.48	224.132020	6.833
4	1	1	1	0	151315	2010-03-05	61.65	3.630	3552.58	601.32	0.00	2666.22	224.347025	6.833
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
97834	10	98	1	1	126512	2012-09-28	82.52	3.966	6091.96	62.82	2.82	3350.88	131.043000	7.170
97835	10	98	1	1	126512	2012-10-05	80.88	4.132	8975.95	0.00	46.24	9546.75	131.075667	6.943
97836	10	98	1	1	126512	2012-10-12	76.03	4.468	2674.51	0.00	10.32	1390.15	131.108333	6.943
97837	10	98	1	1	126512	2012-10-19	72.71	4.449	3067.64	0.00	53.60	967.02	131.149968	6.943
97838	10	98	1	1	126512	2012-10-26	70.50	4.301	9657.93	63.30	100.00	1925.87	131.193097	6.943

97839 rows × 14 columns

Fig 29:

Adjustment made to the dataset

1. train_data_expanded = train_data_expanded.fillna(0)
2. train_data_expanded['IsHoliday'] = train_data_expanded['IsHoliday'].astype('str').map({'True':0,'False':1})
3. train_data_expanded['Type'] = train_data_expanded['Type'].astype('str').map({'A':0,'B':1,'C':2})

1.Filled Nan values with 0
2.for Conditioned input for below prediction models we the Encoded Categorical data to numerical features

RandomForestClassifier
XGBClassifier
LogisticRegression

2.8.2 Ensemble Modeling:

Ensemble modeling is the process of running two or more related but different analytical models and then synthesizing the results into a single score or spread in order to improve the accuracy of predictive analytics and data mining applications.

1. **Averaging method:**

It is mainly used for regression problems. The method consists of building multiple models independently and returning the average of the prediction of all the models

mean_squared_error = 0.02

```
/usr/local/lib/python3.7/dist-packages/sklearn/linear_model/_logistic.py:818: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  extra_warning_msg=_LOGISTIC_SOLVER_CONVERGENCE_MSG,
0.022396327951883514
```

Fig 30:

2. **Max voting:**

It is mainly used for classification problems. The method consists of building multiple models independently and getting their individual output called 'vote'. The class with maximum votes is returned as output

For out model calculated log_loss is 9.99

```
Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  extra_warning_msg=_LOGISTIC_SOLVER_CONVERGENCE_MSG,
9.992007221626413e-16
```

Fig 31:

3. **Bagging:**

It is also known as a bootstrapping method. Base models are run on bags to get a fair distribution of the whole dataset. A bag is a subset of the dataset along with a replacement to make the size of the bag the same as the whole dataset. The final output is formed after combining the output of all base models.

2.9 Recurrent Neural Network (RNN):

A Neural Network consists of different layers connected to each other, it learns from huge volumes of data and uses complex algorithms to train a neural net.

RNN will do the following:

RNN converts the independent activations into dependent activations by providing the same weights and biases to all the layers, thus reducing the complexity of increasing parameters and memorizing each previous output by giving each output as input to the next hidden layer.

Hence these three layers can be joined together such that the weights and bias of all the hidden layers is the same, into a single recurrent layer.

They recognize patterns and clusters and use them to give us powerful insights and terrific applications of different kinds of levels. RNN or Recurrent Neural Networks, as the name suggests, is a repeating neural network. They are the kind whose output from the previous step is fed as input to the current step.

Here we have used RNN to predict the Store type:

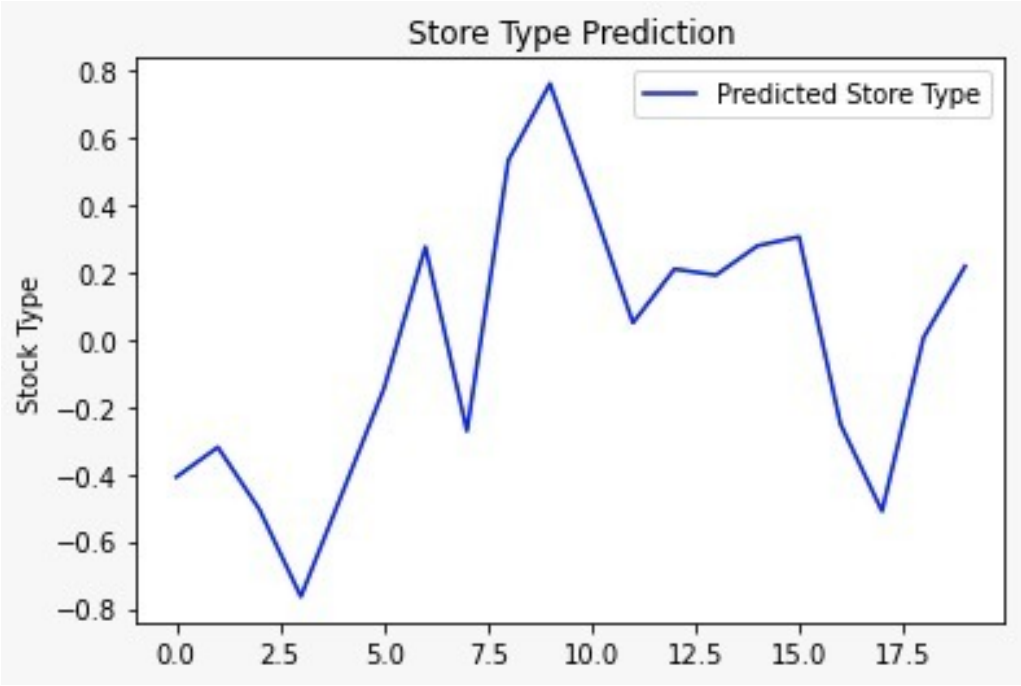


Fig 32:

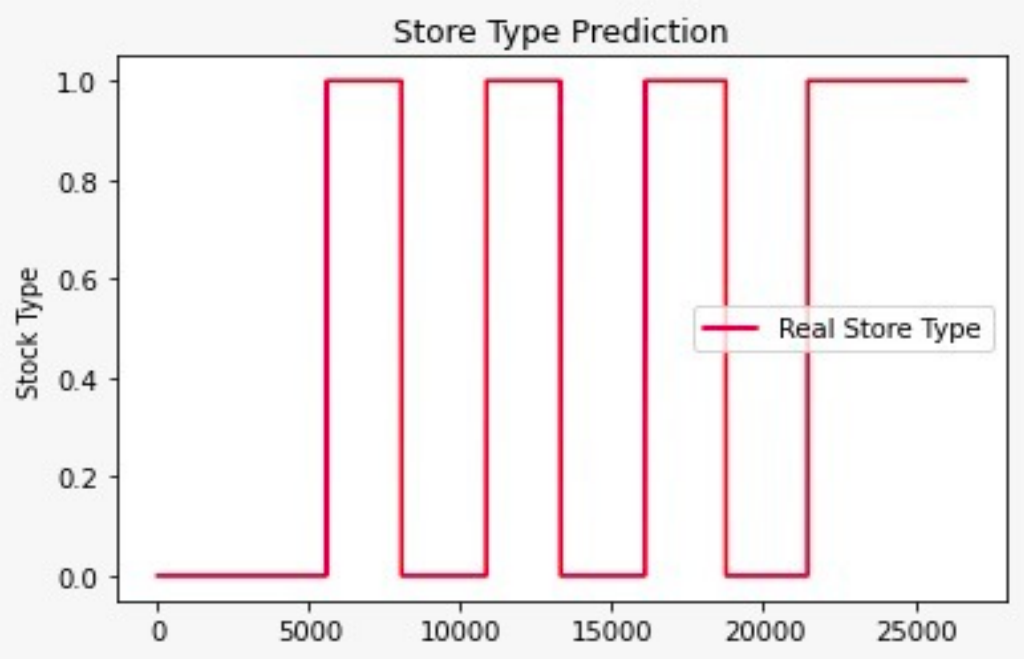


Fig 33:

While RNNs are suitable for handling temporal or sequential data, CNNs are suitable for handling spatial data (images). Though both models work a bit similarly by introducing sparsity and reusing the same neurons and weights over time (in case of RNN) or over different parts of the image (in case of CNN).

2.10 Convolutional Recurrent Network (CNN):

Convolutional Neural Networks (CNNs) are designed to map image data (or 2D multidimensional data) to an output variable (1 dimensional data). They have proven so effective that they are the ready to use method for any type of prediction problem involving image data as an input.

The different layers of a CNN:
There are four types of layers for a convolutional neural network: the convolutional layer, the pooling layer, the ReLU correction layer and the fully connected layer.

The convolutional layer:
The convolutional layer is the key component of convolutional neural networks and is always at least their first layer. Its purpose is to detect the presence of a set of features in the images received as input.

The Pooling Layer:
The pooling layer reduces the number of parameters and calculations in the network. This improves the efficiency of the network and avoids over-learning.

Thus, it reduces the number of parameters to learn and the amount of computation performed in the network.

The ReLU Correction Layer:
ReLU (Rectified Linear Units) refers to the real non-linear function defined by ReLU(x)=max(0, x). Visually, it looks like the following:

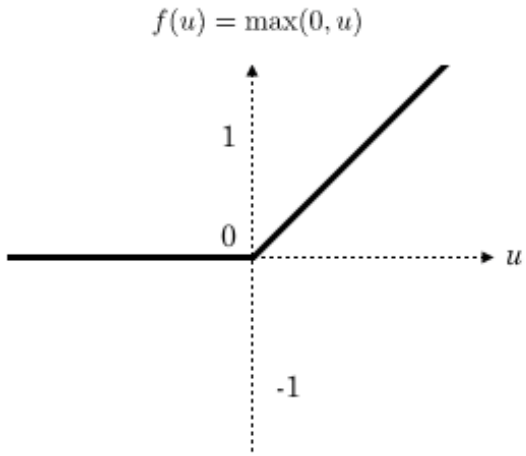


Fig 34:
The ReLU correction layer replaces all negative values received as inputs by zeros. It acts as an activation function.

The fully-connected Layer:
The fully-connected layer is always the last layer of a neural network, convolutional or not — so it is not characteristic of a CNN.

This type of layer receives an input vector and produces a new output vector. To do this, it applies a linear combination and then possibly an activation function to the input values received.

```
Model: "sequential_5"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d_15 (Conv2D)          (None, 60, 1, 8)          664

 conv2d_16 (Conv2D)          (None, 58, 1, 8)          200

 max_pooling2d_10 (MaxPoolin  (None, 29, 1, 8)         0
 g2D)

 conv2d_17 (Conv2D)          (None, 27, 1, 8)          200

 max_pooling2d_11 (MaxPoolin  (None, 13, 1, 8)         0
 g2D)

 flatten_5 (Flatten)         (None, 104)               0

 dropout_5 (Dropout)         (None, 104)               0

 dense_5 (Dense)             (None, 1)                 105

=================================================================
Total params: 1,169
Trainable params: 1,169
Non-trainable params: 0
```

Fig 35:

2.11 Performance matrix

Performance metrics are measure of calculation of the quality of the prediction models
Following the subcategories in performance matrix

- Confusion Matrix
- Precision
- Recall
- Roc Curve
- Auc curve
- F Score
- Sensitivity

Now let's understand each of the categories one -by- one and let me walk us through how our code performed as per the mentioned performance matrices

What is a confusion matrix?

It is a matrix of size 2×2 for binary classification with actual values on one axis and predicted on another.
Like in our case we have actual values on y-axis and predicted on x-axis.

True Positive (TP) -- model correctly predicts the positive class (prediction and actual both are positive).
For the given problem set **13870** were made true and they were actually true also

True Negative (TN) -- model correctly predicts the negative class (prediction and actual both are negative). In the above plot, of all stores predicted Type B 11836 were actually type B, from test data.

False Positive (FP) — model gives the wrong prediction of the negative class . In the above plot,
Of all the stores predicted to be type A **, 970** stores are actually type B

False Negative (FN) — model wrongly predicts the positive class (predicted-negative, actual-positive). In the above plot , of all the stores predicted to be type B all of them were actually type b. There is '0' false negative here.

Precision -

Out of all the values predicted positive how many are actually positive
So, we have 13870 Stores predicted positive , and all of them actually turned out to be A
Thus, precision is **1.**

Recall -

Out of the total positive, what percentage are predicted positive. It is the same as TPR (true positive rate).

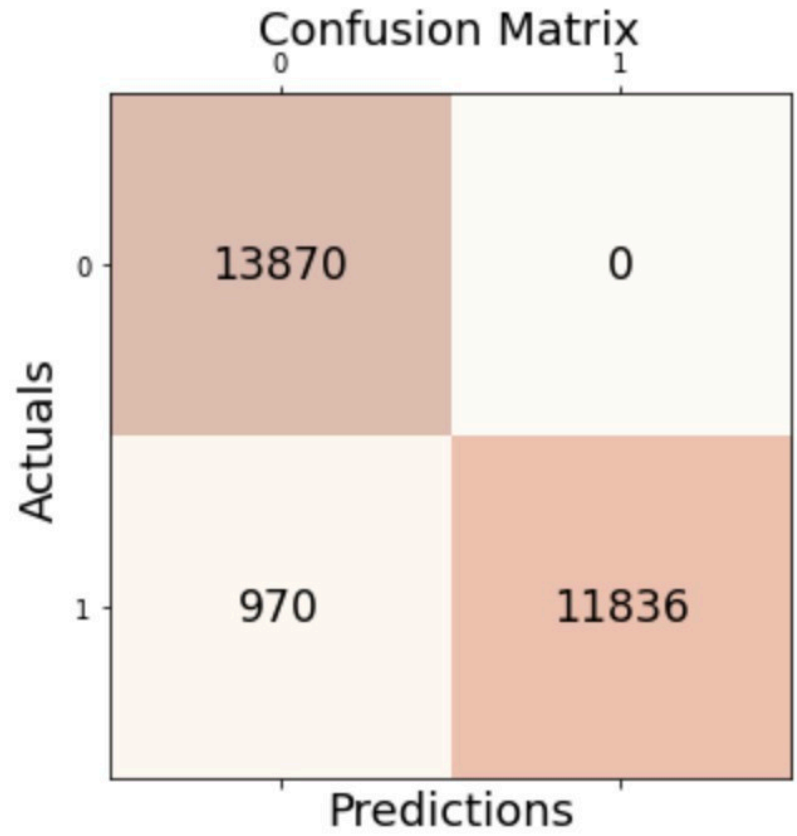
Accuracy -

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations.

F1 Score –

It is the harmonic mean of precision and recall. It takes both false positive and false into account. Therefore, it performs well on an imbalanced dataset.

Below is the confusion matrix as asked for given dataset.




```
Precision: 1.000
Recall: 0.924
Accuracy: 0.964
F1 Score: 0.961
```

Fig 36: Confusion Matrix

ROC curve

An **ROC curve** (**receiver operating characteristic curve**) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

- True Positive Rate
- False Positive Rate

AUC: Area Under the ROC Curve

AUC stands for "Area under the ROC Curve." That is, AUC measures the entire two-dimensional area underneath the entire ROC curve from (0,0) to (1,1).

AUC provides an aggregate measure of performance across all possible classification thresholds. One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example.

AUC = 0.96

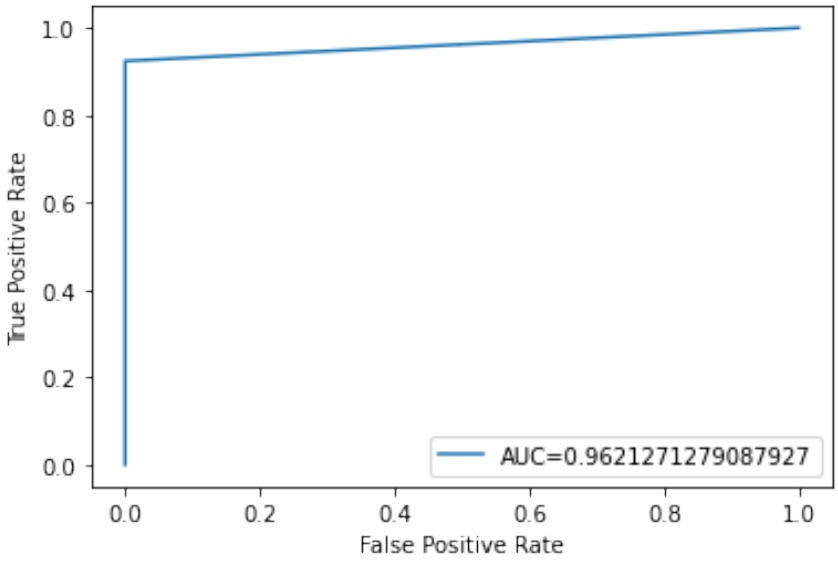


Fig 37:

3 Dataset_02

3.1 Exploratory Data Analysis (Dataset_02)

Datasets contain 2 files:

1) Calendar.csv

2) Sell_prices.csv

We have data from California, Wisconsin, Texas. And we have items which are sold in each state containing numerous stores in them.

Yearly sales of 2011 shows that there is an increase in sales in month of April, July, December 2011.

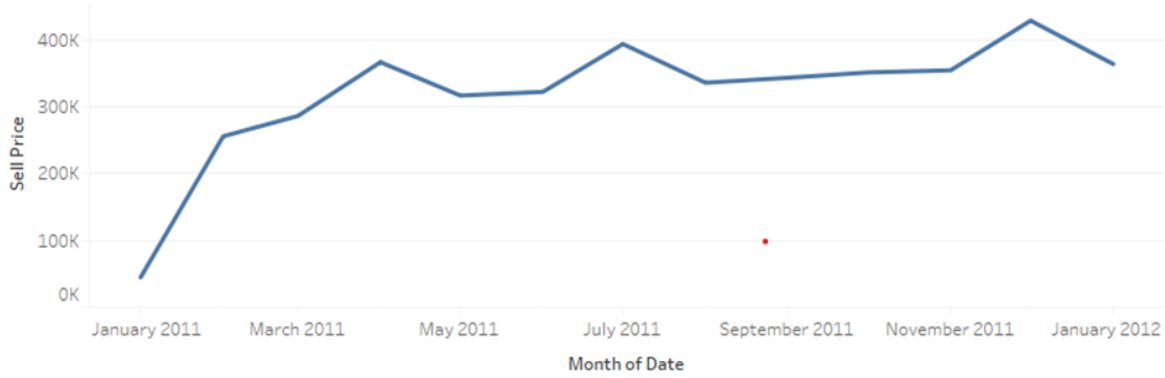


Fig 38: Monthly sales

Which Item in each group has highest and lowest sales?

Ans: Food_3 has highest sales and Food_1 has lowest sales. Hobbies_2 has highest sales and Hobbies_1 has lowest sales. Household_2 has highest sales and Household_1 has lowest sales. Out of all Household_2 is highest selling of 7,115,264.

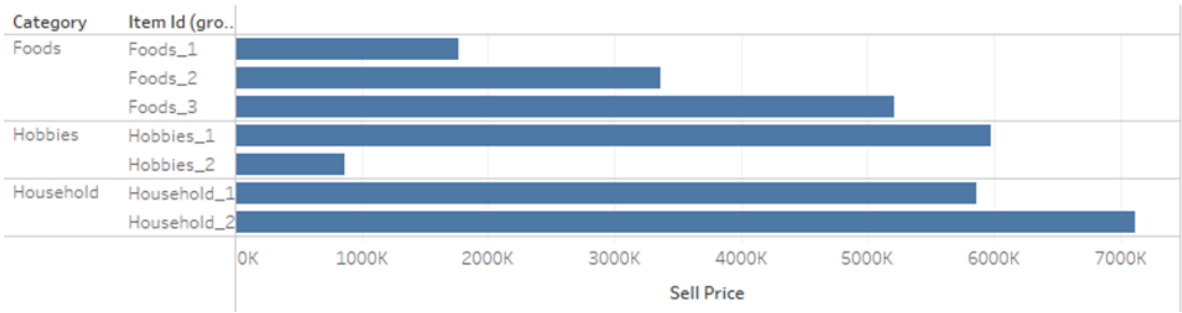


Fig 39: Categorical sales in a year

Which state has maximum sales?

Ans: California has maximum sales in all the years. And Wisconsin has least sales for all the years.

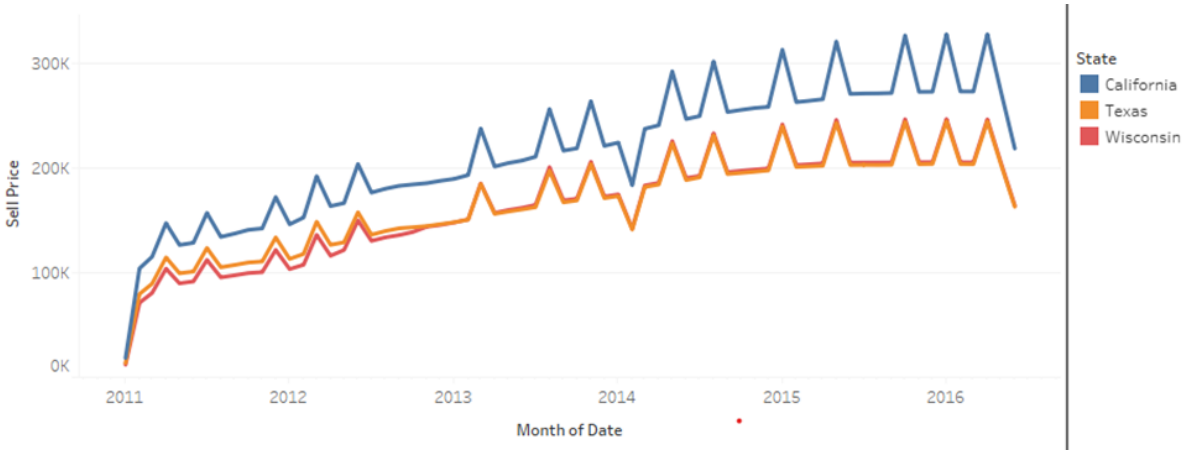


Fig 40: State wise sales over the years

What do we interpret from Pie-Chart?

Ans: Out of 3 item categories, household account for 43.04%, Food account for 34.28%, and Hobbies account for 22.69%.

Fig 41: Pie chart for categorical sales percentage

What are our sales averages on Event days?

Ans: Average sales is $4 Million. But most of our sales are being done on normal routine days. In terms of event most sales are done on National Events.

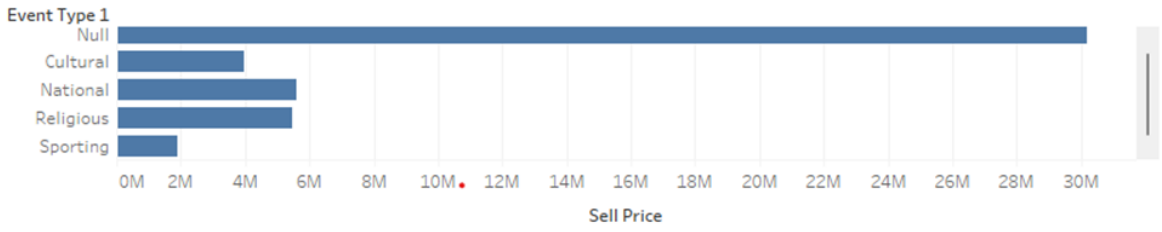


Fig 42: sales over the event_type period

This chart shows that daily sales based on weekdays are very similar. Average sales in all the states are equal.

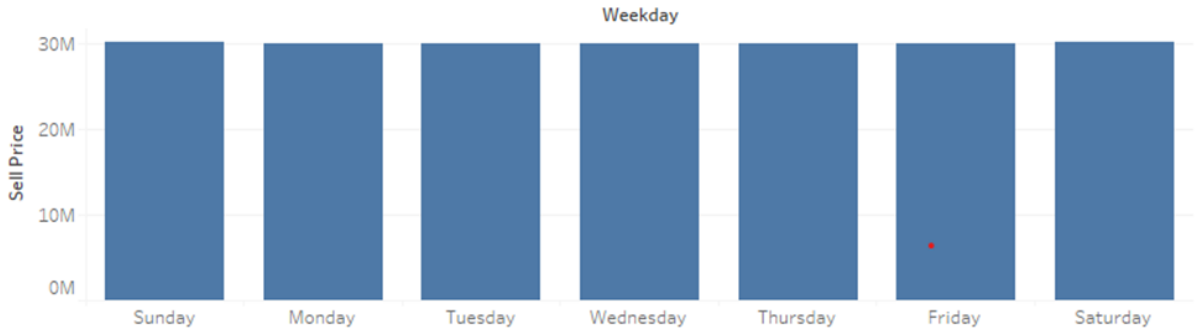


Fig 43: Individual Weekday Sales

What are our findings from this graph?

Ans: We get to know that sales gradually increase from year 2011 to 2016 gradually. And most of the sales amount is made from household and food categories.

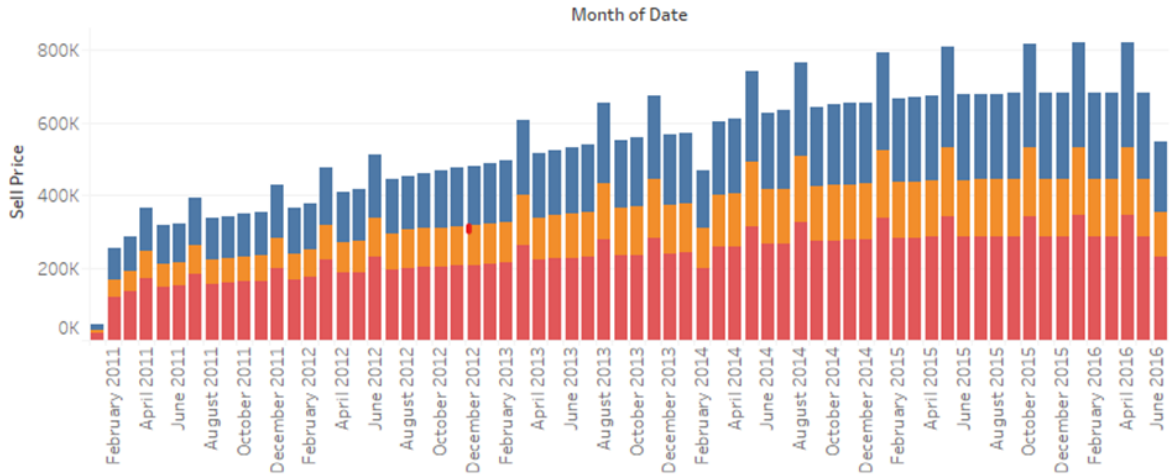


Fig 44: Sales of each category in each month

Which is the most sold item in all the stores?

Ans: Household_2_446 is the most sold item across all the stores.

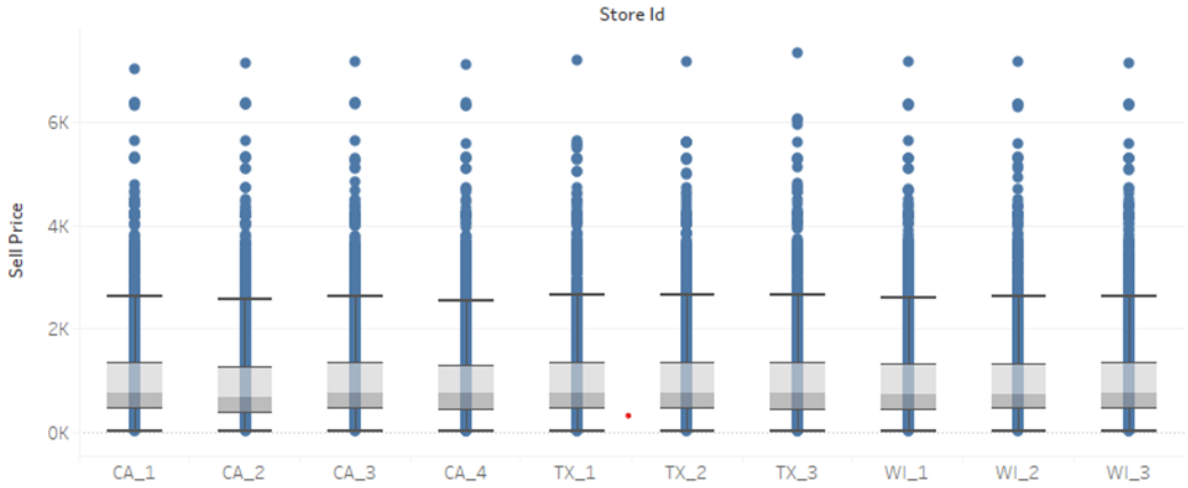


Fig 45: Highest selling item in each store

How does our sales flare over the period of year?

Ans: We interpret from the graph that there is sudden increase in graph from February to May, then the sales decrease for every store in all the states.

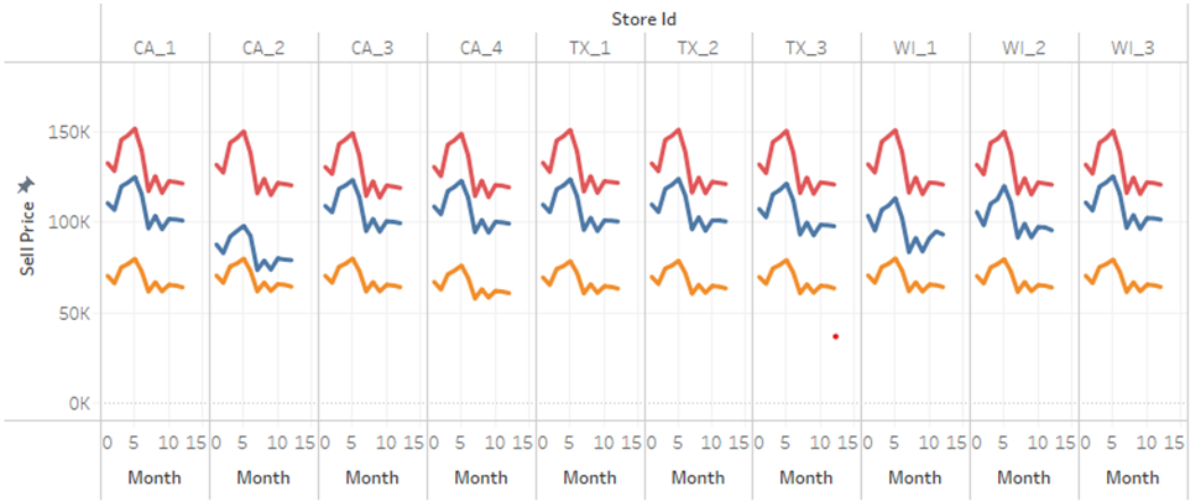


Fig 46: Sales of each category in every store

What does this graph represent?

Ans: We see that most sales are done on normal days. One of the reasons can be that normal days are more than event days.

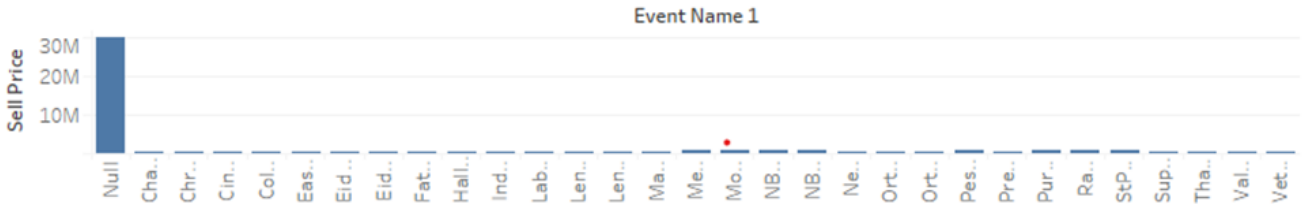


Fig 47: Sales during event day

3.2 Modeling:

- Large Files
  - sales_train_evaluation is 116 MB and has 30490 rows, 1947 columns
  - sell_prices are 193 MB and has 1048575 rows, 4 columns
  - calendar is 101 KB and has 1969 rows, 14 columns

Downcasting: Shrink Pandas DataFrames with precision safe schema inference. Pandas-downcast finds the minimum viable type for each column, ensuring that resulting values are within tolerance of original values.

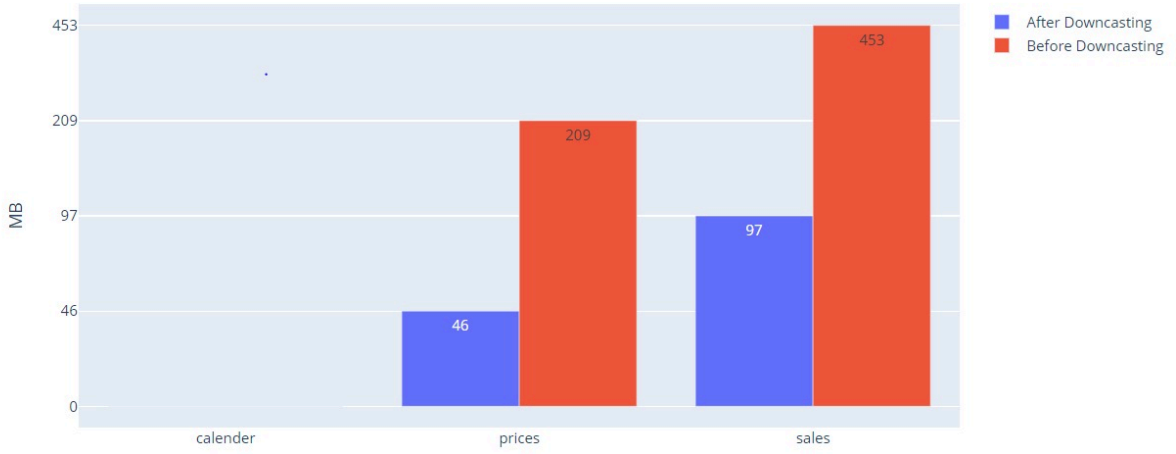


Fig 48: File sizes after down casting

EDA:

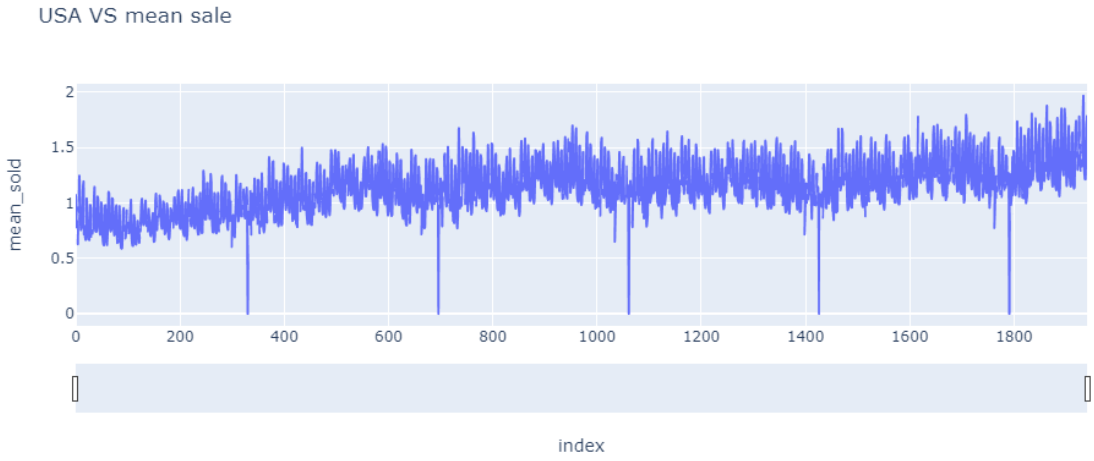


Fig 49: Variation in Mean sales vs Sales in USA

From this graph, the variation in mean sales can be seen vs the sales in USA

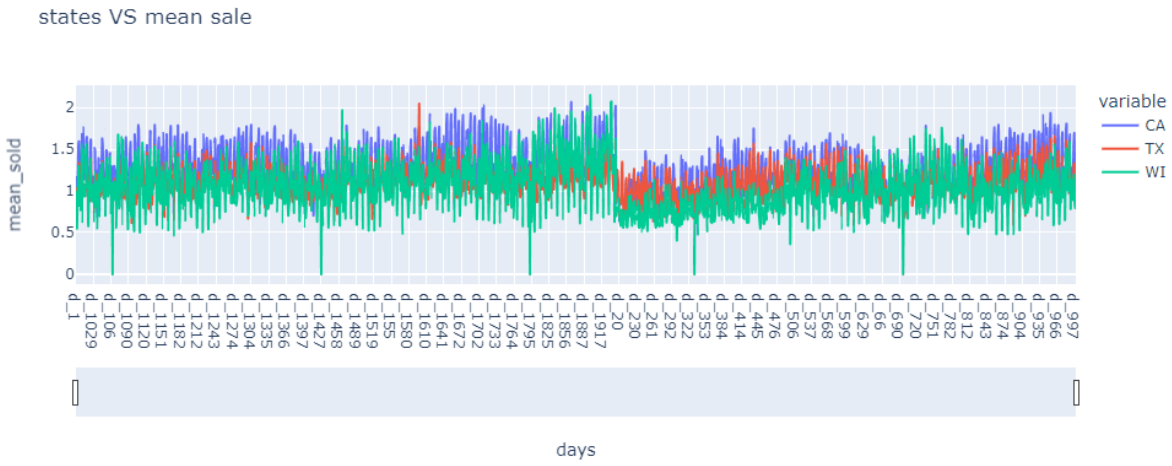


Fig 50 : State Sales vs Mean sales

The variation in sales state wise vs mean sales can be observed from this graph

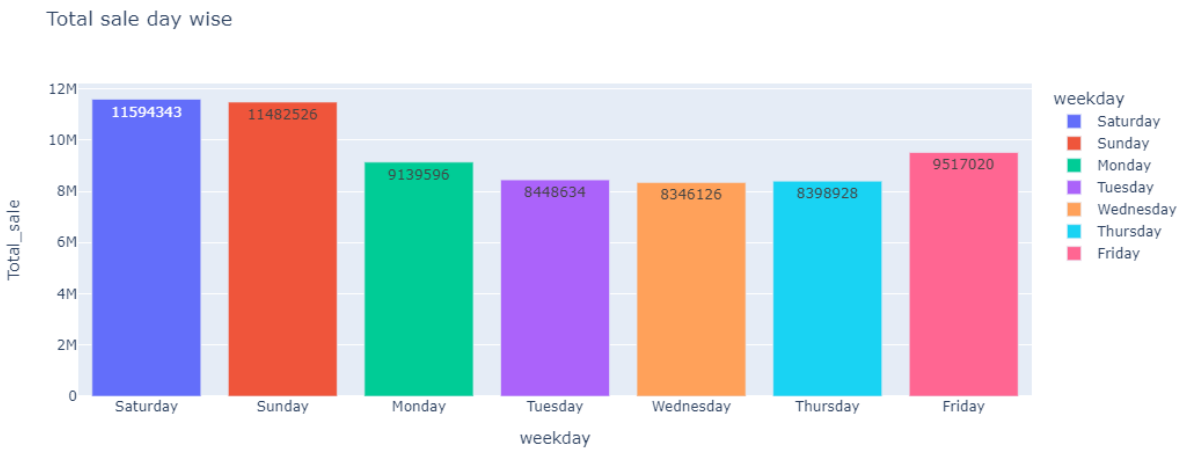


Fig 51: Day-wise Total sales

The variation in total sales according to the day can be observed. As can be seen, sales are higher on Friday, Saturday and Sunday. A new feature called is_high_sale_day is introduced corresponding to this.

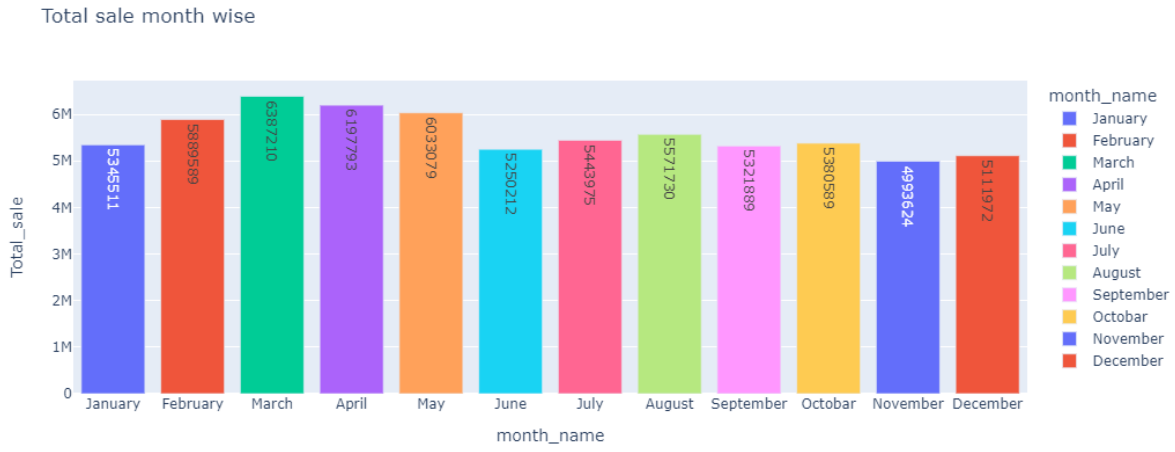


Fig 52: Month-wise Total Sales

The variation in total sales according to the month can be observed. As can be seen, sales are higher in the months of March, April, and May. A new feature called is_high_sale_months is introduced corresponding to this. This corresponds to seasonality in the data.

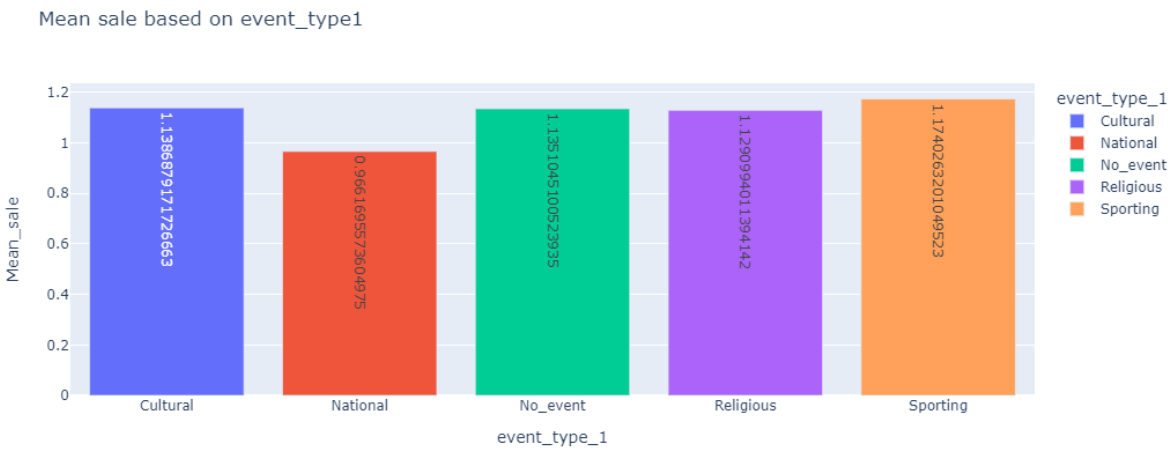


Fig 53: Dales bases on event_type1

As observed in this graph, there is no significant change in sales based on event_type1. Therefore event_type1 can be dropped.

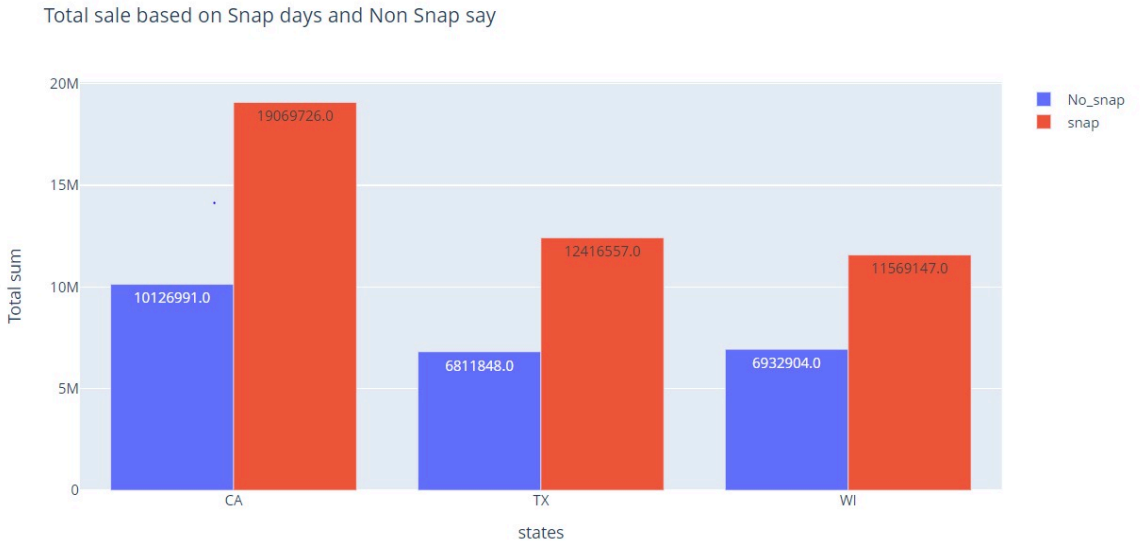


Fig 54: Total sale based on snap days and non-snap days

The variation in sales in the three states on Snap and non-Snap days can be seen here. Clearly, Snap days have higher sales than non-snap days.

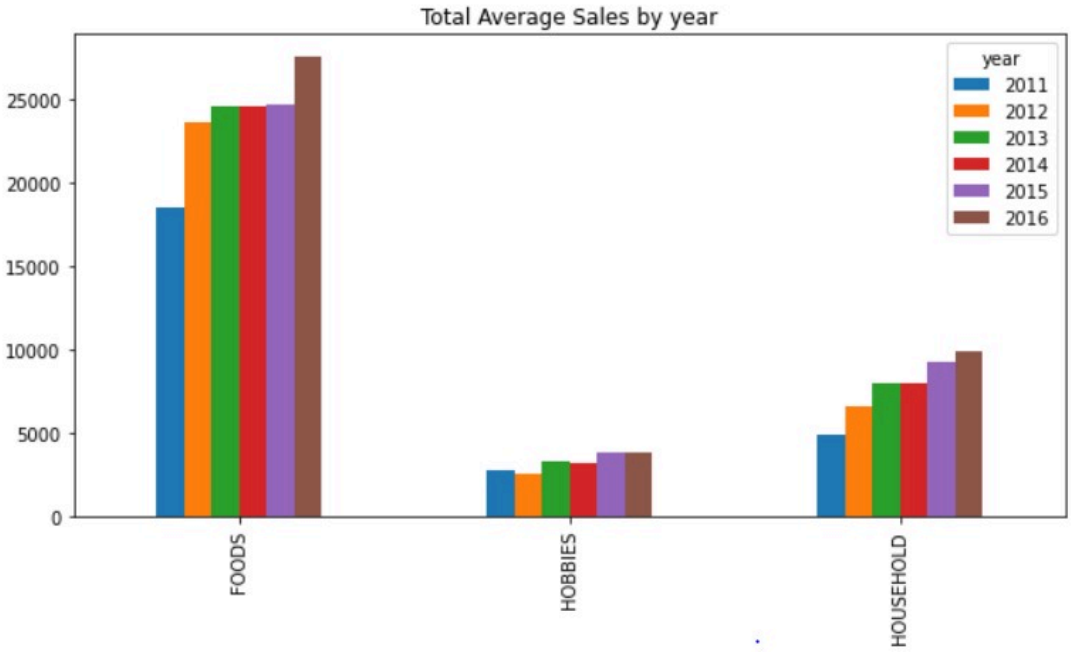


Fig 55: Average Sales for each category

The average sales for each category for each year are plotted. As observed, the sales in all three categories are increasing year-on-year.

Feature Engineering:

Weather data was obtained for the 4 Zip Codes in California, 3 Zip Codes in Texas, and 3 Zip Codes in Wisconsin. This data includes Precipitation, Snow and Temperature. The days on which Precipitation or Snow is high, sales are down compared to other days. This weather data was added as one of the features.

Median Income was also fetched from the internet for all 4 California Zip Codes, 3 Texas Zip Codes, and 3 Wisconsin Zip Codes. The median income was also added as one of the features.

Models:

ARIMA

ARIMA is an acronym for Auto Regressive Integrated Moving Average. It is a model used for analyzing and forecasting time series data.

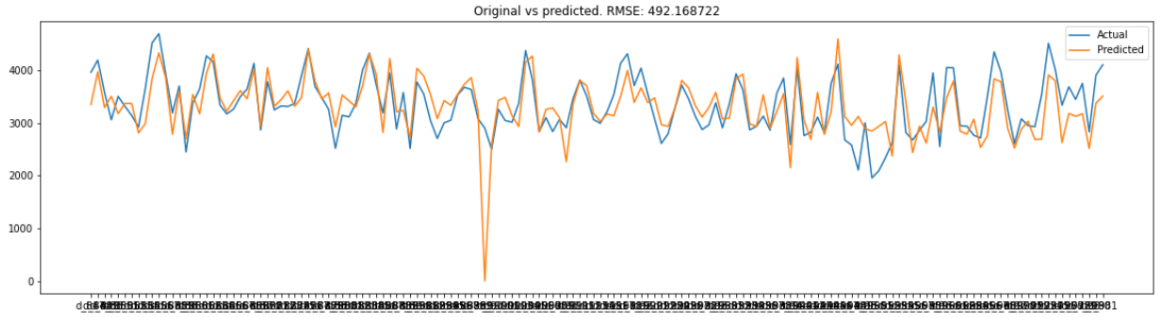


Fig 56: Prediction for HOBBIES

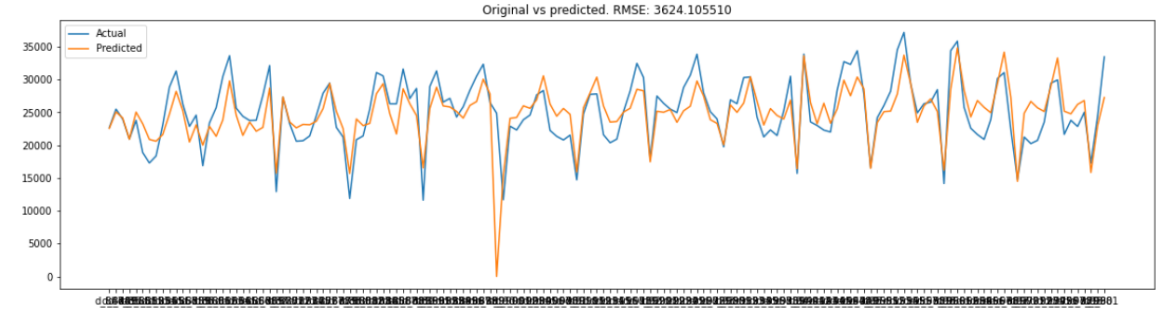


Fig 57: Prediction for FOOD

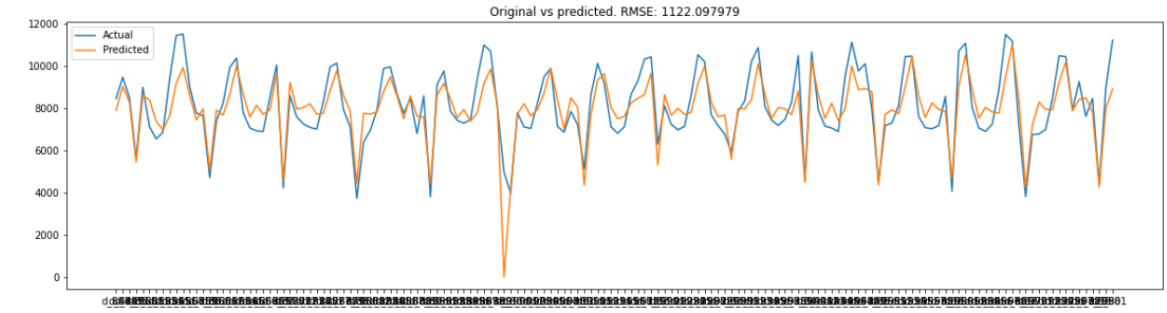


Fig 58: Prediction for HOUSEHOLD

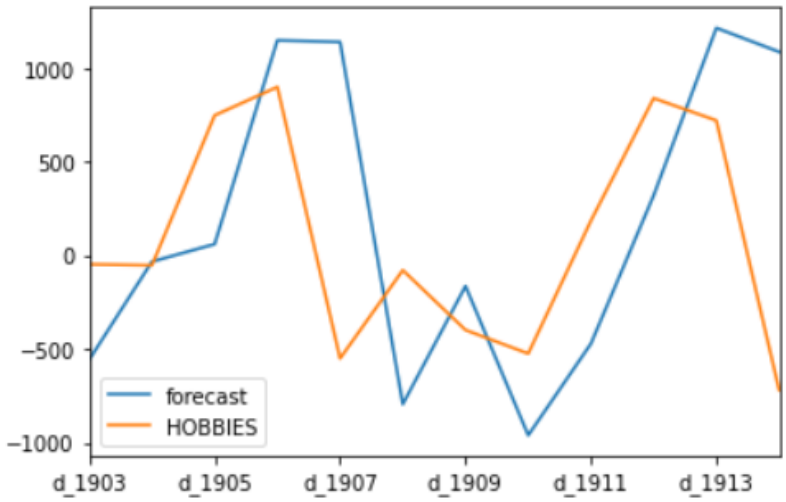


Fig 59: Prediction for HOBBIES for 10 days relative to the average sale

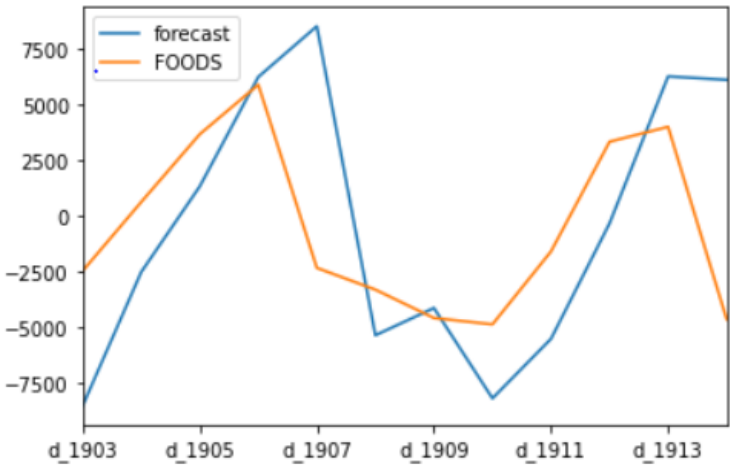


Fig 60: Prediction for FOODS for 10 days relative to the average sale

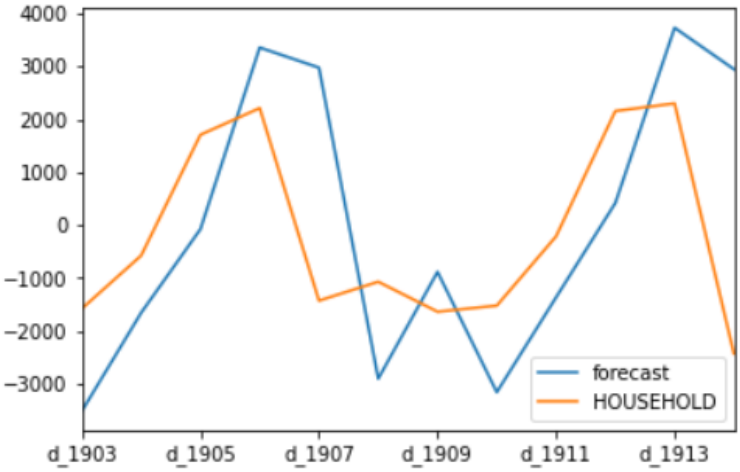


Fig 61: Prediction for HOUSEHOLD for 10 days relative to the average sale

Category	RMSE without External Features	RMSE with External Features
HOBBIES	492.16	550.25
FOOD	3624.14	2877.21
HOUSEHOLD	1122.09	1015.2

LSTM

Long Short Term Memory networks(LSTMs) are a special kind of Recurrent Neural Networks, capable of learning long-term dependencies.

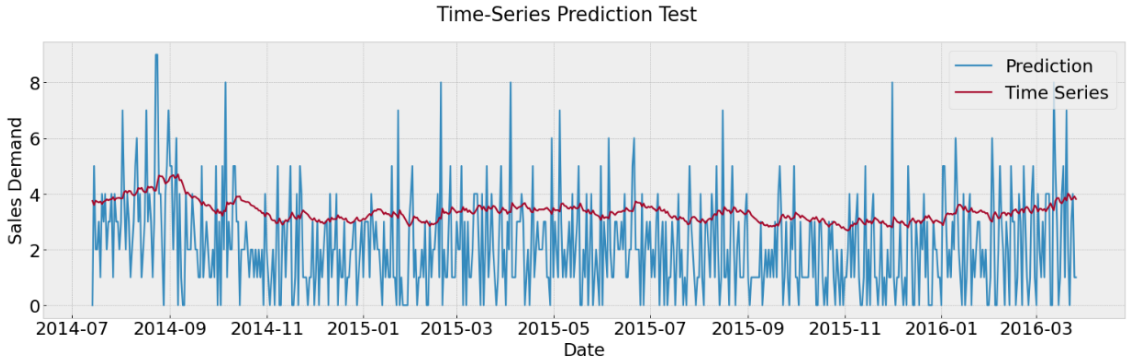


Fig 62: Prediction without external features

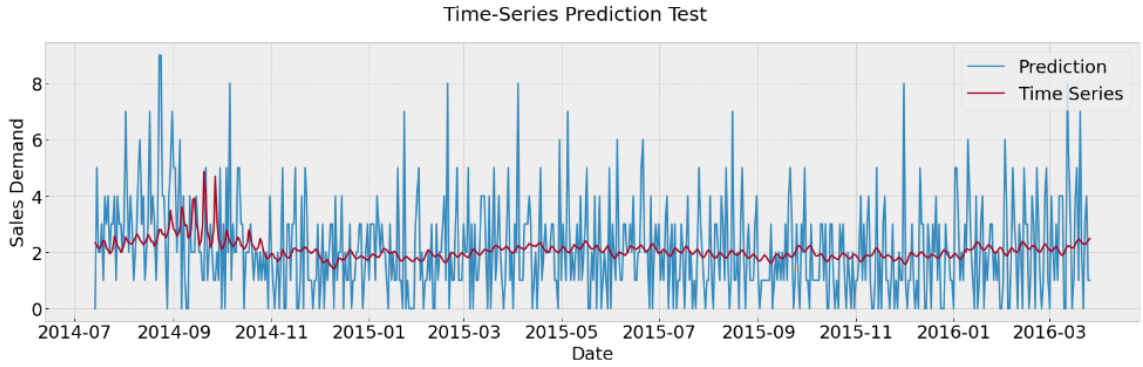


Fig 63: Prediction with external features

RMSE without external features	RMSE with external features
2.025	1.687

Model Comparison:

Model	RMSE without external features	RMSE with external features
ARIMA	1746.13	1480.90
LSTM	2.025	1.687

4 Model Deployment and Business Recommendation

The job of a data scientist goes beyond designing an accurate model on the available data, the deployment of data for the consumption of the end user is also the part of job. In this section, we are deploying the data using Streamlit.

4.1 Deploying the data using Streamlit.

Streamlit is an open-source app framework in Python language. It enables us to develop intricate web applications for data science and machine learning. Major Python libraries like scikit-learn, Keras, PyTorch, SymPy (latex), NumPy, pandas, and Matplotlib are all compatible with it.Since Streamlight has all the major python libraries that we have used it is an apt framework for this project.

We deployed the model on the local system. The features of the deployed data are sales categories, i.e., Food, household, and hobbies. The model predicts the sales of different categories of items, also the user can set the start and end date to analyze the sales in the long term and short term.

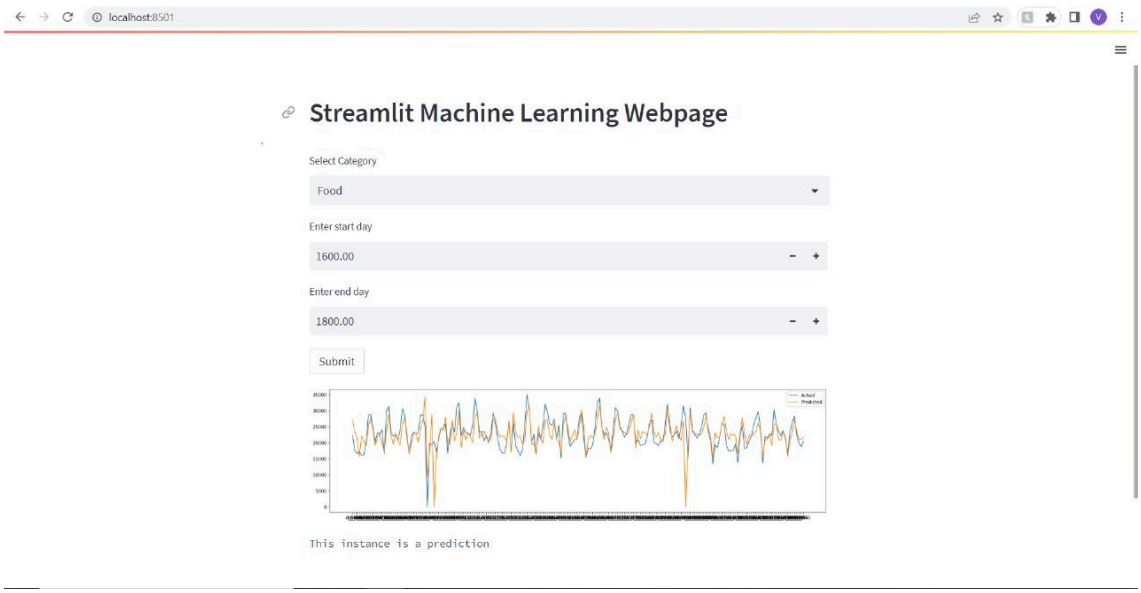


Fig 64. Streamlight representation of food sales in 1600 - 1800 days

We deployed the ARIMA model for the consideration of the above Streamlit webpage. We have considered data from d_1 to d_1969 for the above model. The features from the dropboxes are fed to the code and the plots are obtained respectively using pyplot.
A link to the demo video has been provided in the “Dashboard for Model” in the end of the document.

4.2 Product Segmentation based on demand variability

A method for categorizing inventory goods based on their consumption values is ABC analysis. The entire cost of an item over a predetermined period, such as a year, is the consumption value. The strategy is used in this situation and is based on the Pareto principle to help manage what matters.

Below mentioned are the references that are driving most of the sales

- Class A: Very Fast Movers: top 5%
- Class B: The following 15% of fast movers
- Class C: The remaining 80% of very slow movers

Below are the tasks created for product segmentation and business recommendation

4.2.1 Use first-year data of the Household category to create ABC Analysis and interpret the graph. How stable is the customers' demand? (Coefficient of Variation)

```
A    292
C    185
B    172
Name: CLASS, dtype: int64
Cost of Class A : 5972687.670000002
Cost of Class B : 1122231.7800000012
Cost of Class C : 376338.1000000001
Percent of Cost of Class A : 0.79942200225717
Percent of Cost of Class B : 0.15020654454617222
Percent of Cost of Class C : 0.05037145319665762
```

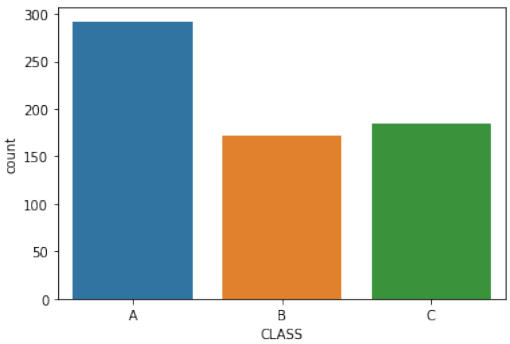


Fig 65: Total product item count

The below inferences can be drawn from the above outpu

Class	Total product item count	Total Cost ($)	Quantity percentage	Cost Percentage
A	292	5972687.67	44.99 %	79.94 %
B	185	1122231.78	28.51 %	15.02 %
C	172	376338.1	26.50 %	5.04 %
Total	649	7471257.55		

4.2.2 To understand which products will bring planning and distribution challenges, compute the coefficient of variation of the yearly distribution of sales of each reference.

The above table gives information about the distribution of products into categories. Category A items represent 45% of an overall inventory by item and represent 80% of the value of an inventory. Category B items represent 28.5% of an overall inventory by item and represent 15% of the value of an inventory. Category C items represent 26.5% of an overall inventory by item and represent 5% of the value of an inventory.

The total revenue generated by Category item A is 5972687.670000002, Category B is 1122231.7800000012, and Category C is 376338.100000000.

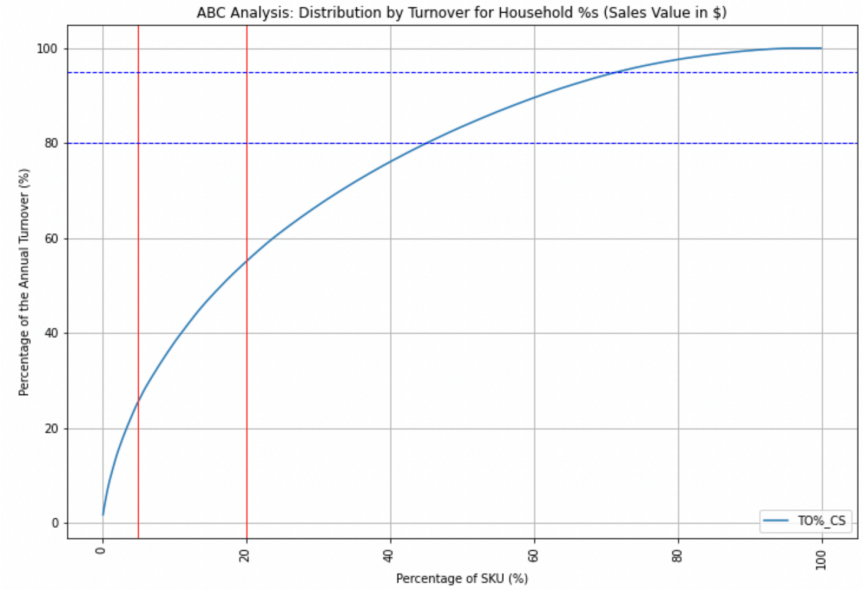


Fig 66:ABC Analysis: Distribution by turnover for household %s

The above graph shows how the percentage of turnover of products spans across the percentage of products sold over a period for the category of household

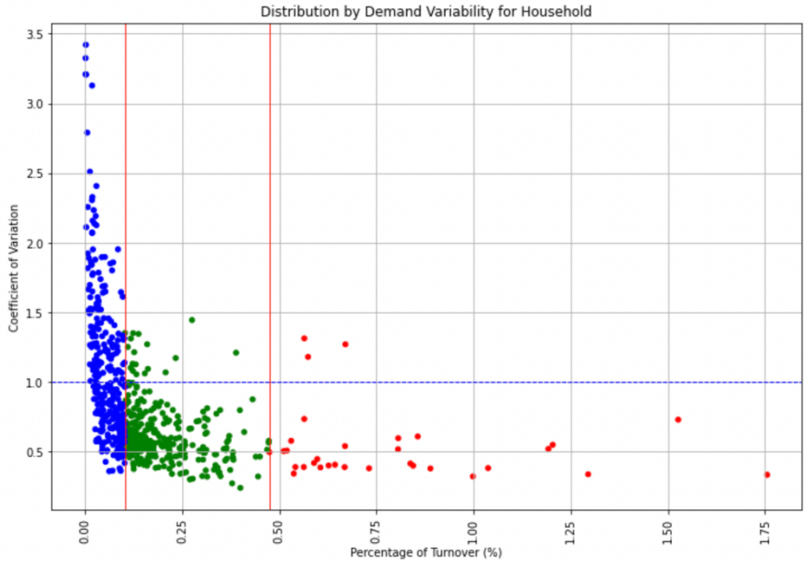


Fig 67: Distribution by Demand Variability for Household

The coefficient of variation is a way to measure how spread-out values are in a dataset relative to the mean. In the above graph, there are many items in the category household whose COV varies from 0.5 being the lowest to 3.5 being the highest.

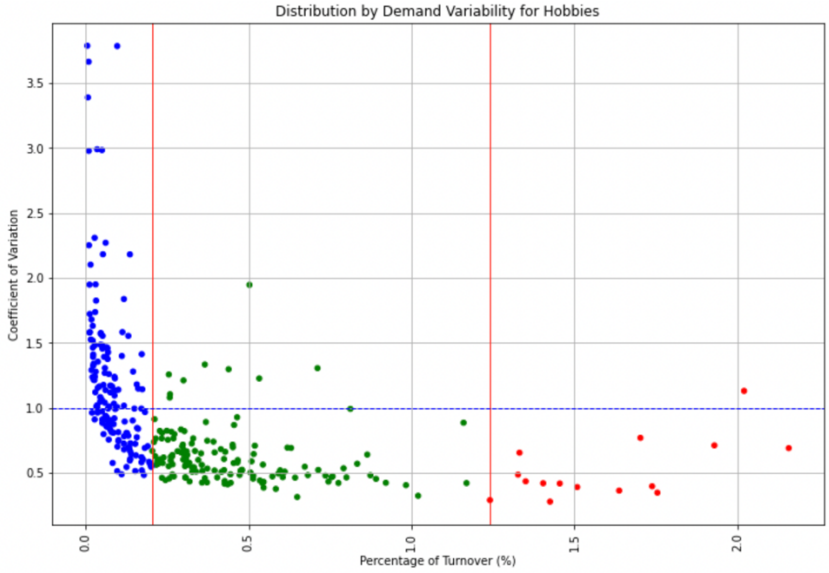


Fig 68: Distribution by Demand Variability for Hobbies

In the above graph, there are many items in the category Hobbies where it has items whose turnover lies in the first 50 percentiles has more no of items in this category and

the values are not very scattered making the majority chunk fall in the first quadrant of the segment.

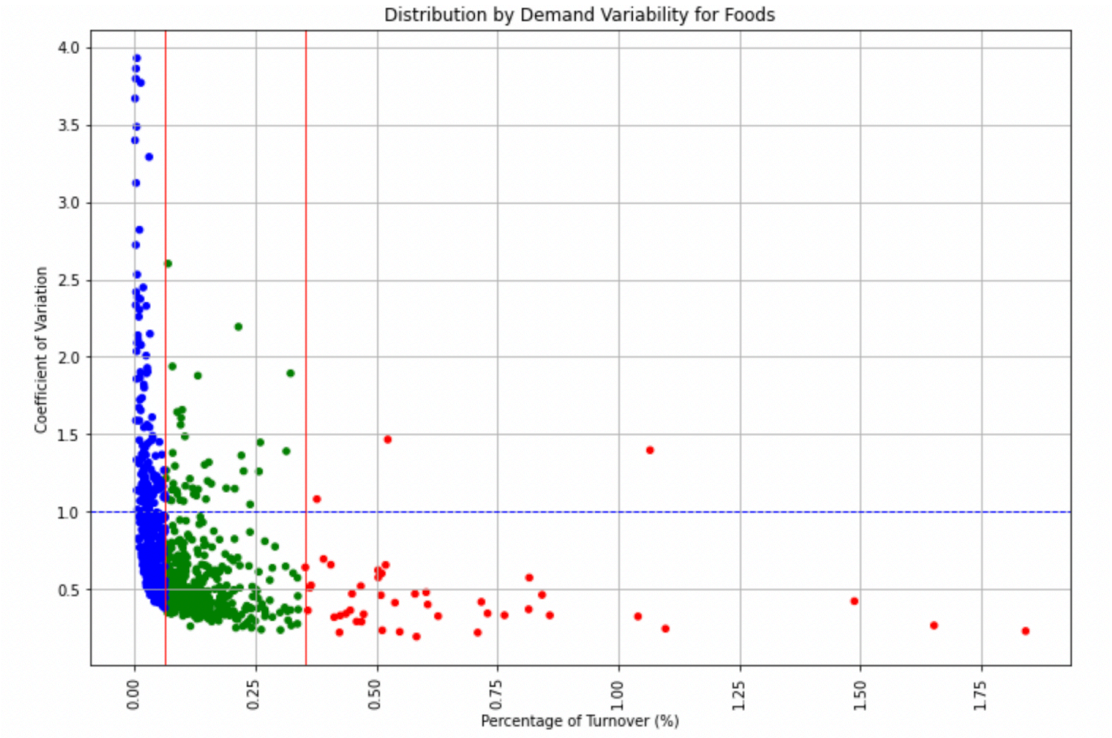


Fig 69: Distribution by Demand Variability for Foods

In the above graph, there are many items in the category Hobbies where it has items whose turnover lies in the first 50 percentiles has more than 85 % of items in this category and the values are not very scattered making the majority chunk fall in the first quadrant of the segment.

4.2.3 Discuss a few initiatives and recommendations for improving the retail business for dataset_02.

It is very important to draw out conclusions from the analysis of the prediction model, and what makes it beneficial for the businesses is, if the recommendations drawn can increase the sales figures.

In these sections, using different graphs and plot we provide recommendations for different scenarios

Fig 69: Mean sales vs. Store name

The highest sales are for CA_3 and the lowest are for CA_4. Even though both of them are in the same state but their sales vary. The possible reason for the above could be the high median income of residents of CA_3 relative to the residents of CA_4

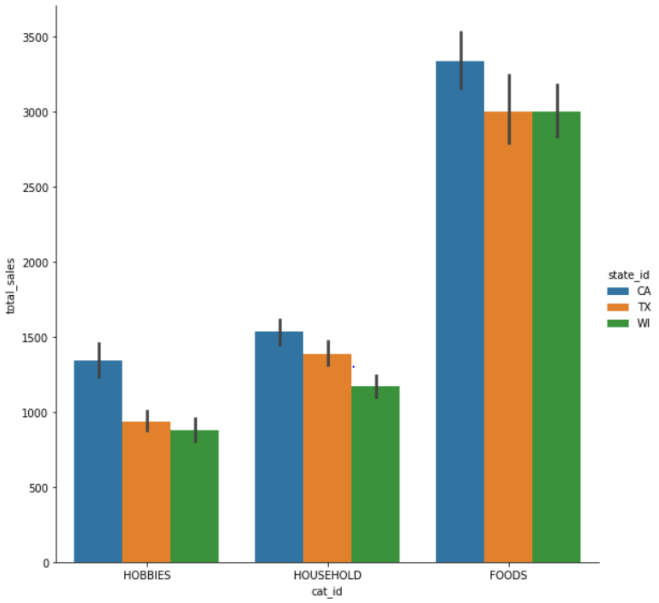


Fig 70: Sales across different categories in all states

Sales in all the states are very similar, the population of Wisconsin is 1/9th of California and 1/7th of texas still their sales are the same.

So more offers should promote in California and Texas as the market over there hasn't reached its potential. And we can see through the weather too that it is very good in California and Texas compared to Wisconsin.

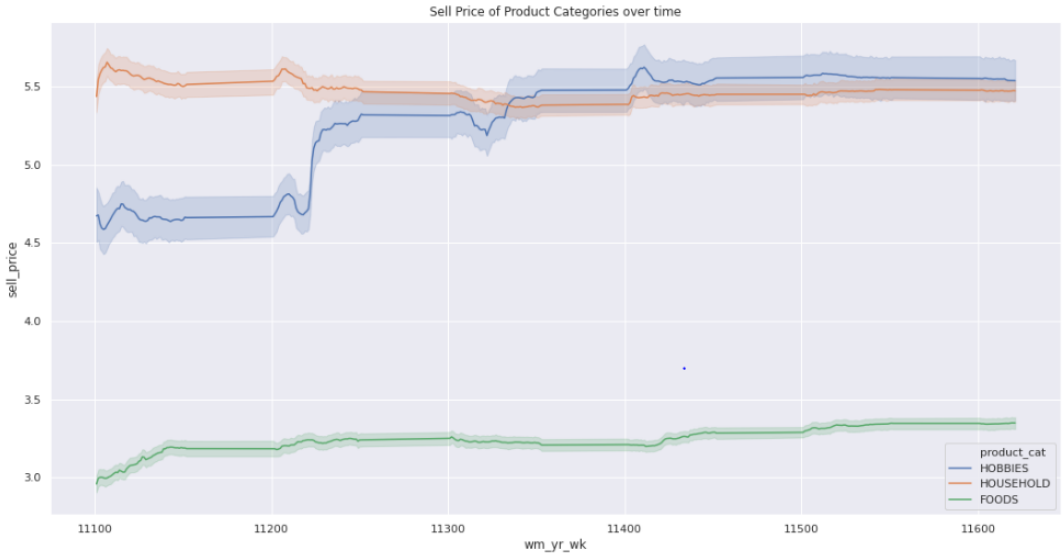


Fig 71: Sell Price of Product Categories over time

The below graph shows that prices of each commodity rise each year. But the prices of hobbies rose maximum with a sudden jump. It might be because the raw material for hobbies products got expensive. So, stores should provide discounts on hobbies products so sales of hobbies can also be boosted.

Fig 72: Total Sales across all years

We can see a sudden dip in Christmas. So people must not be going out to shop on those days. Shops should give maximum promotion on that days as there is the probability of getting good sales as everyone is having holidays and everyone can shop freely.

Fig 72: Yearly Sale Value vs Year

The sales graph shows an increase yearly but after 2015 it looks constant.Might be because of rising competition in this segment. So they should come up with some new products to improve their sales

5 Conclusion:

5.1 DataSet_01

The following inferences and conclusions can be drawn from the analysis of the data:
Type 'A' stores are more popular than 'B' and 'C' types.
Type 'A' stores outclass the 'B' and 'C' types in terms of size and average weekly sales.
Weekly Sales are affected by the week of the year. Holiday weeks witnessed more sales than non-holiday weeks.
The size of the store is a major contributing factor in the weekly sales.
Sales are also dependent on the department of the store as different departments showed different levels of weekly sales.
Among the trained models for predicting future sales, Arima with tuned hyperparameters performs the best.

5.2 DataSet_02

So, at first, looking at the datasets we found out there are 3 categorical data. And sales of stores in 3 states are given. In a particular dataset, five key innovations are emerging directly or indirectly from the valuable feedback we can get.
The use of high-frequency (daily) data for the entire set is considered. The data we had sales for individual stores for an individual day.

The use of data that are intermittent in nature. In the problem, there is a high degree of intermittence for the majority of the time series which translates to zero values (zero sales) for some observations. This creates an additional challenge in terms of forecasting: one needs to forecast not only “how much”, but also “when” to achieve the best performance.

This problem also offers a clear hierarchical structure, where the most granular level corresponds to the sales of one product at a particular store. Then, such sales can be aggregated in terms of stores and states or products, departments, and categories.

This dataset has individual value for each product and sales on special days too. That gave us much insight regarding the suggestions and reasons behind each sale in each store.

Despite the external data featured, we believe this problem can be regarded as a “work in progress” and that any results are limited by the data used and, in this case, the particular retail sales application. Regardless, we also believe it has a lot to offer to our

understanding of how forecasts perform under different settings and how uncertainty behaves.

Model:

Now on part of our model, we can improve It using Hyperparameter tuning adjustments. And also we can try to use other models like LGBM. And also other features as weather features didn't help us as the weather in Texas and California are very favorable throughout the year. We can try features like transportation around the area to the store. The average population of people near that store. We can also consider the Average age of customers in that state.

Key Findings

• The model(s) is conservative in predicting sales particularly the peak and low changes. We found defining some initiatives/stimulation patterns is a very helpful tactic to stimulate the model to take/recognize the pattern of peak and low (by realizing such feature patterns) of market demand. Due to the conservative nature of the model, increasing inventory by a flat rate (e.g. 20% in our initiative 3) drives the model to be closer to the market reality. Weekend/holiday can be combined with a flat raise to form a new initiative.

• Temperature is a complicated pattern, but we do think it is a helpful pattern to define a new initiative. It can be combined with all other initiatives together to reflect the real market demand.

6 Source Code (Dataset_01)

siddharth-01/Retail-market-analysis (github.com)

7 Dashboard for Model (Dataset_02)

https://public.tableau.com/app/profile/varun.shah1550/viz/Group_1_Shah_Sharma_Patil_Patel_Lokhande_Katkamwar/Dashboard1?publish=yes

8 Acknowledgement:

Throughout this project's problem-solving and reporting process, we received a great deal of invaluable insight and guidance from Dr. Allen Bolourchi. We would like to thank him for sharing his expertise with us. He utilized his private time throughout the semester to discuss with us how to approach the problems, providing his professional insights in the ML/DL/Time Series field, and giving us feedback to improve our findings and reporting skills.

The project, to all of us, is huge and new. This is a tremendous learning ex- experience to tackle an industry's real problems and we appreciate very much the opportunity to work with Dr. Bolourchi throughout this semester.

Meantime, we thank Dr. Mahshid Fardadi for giving us this opportunity in her class to tackle this real-world industry problem, and for connecting us with Dr. Bolourchi. We thank her support and assistance during the project and reporting process this semester.

9 References

Sprint 01:

Dataset_01: EDA URL:
https://colab.research.google.com/drive/1F1kAHjIxbO-LwYp8MZizqiOY_HHxH-UR#scrollTo=1CwY5eXgPp2c

Dataset_02: EDA URL
https://public.tableau.com/app/profile/varun.shah1550/viz/Group_1_Shah_Sharma_Patil_Patel_Lokhande_Katkamwar/Dashboard1?publish=yes

Sprint 02:

Dataset_01: Modelling URL:
https://colab.research.google.com/drive/1syT7wveZu3_cxB6G9LwLE8XmnW1VJVS#scrollTo=5MsZsC6VZSa3

https://colab.research.google.com/drive/1Mh92KwAcsKRKrP8YYeOZjGo9XcThCoEu#scrollTo=OG4z3CqGGNK7

https://colab.research.google.com/drive/1XG4frXMhrZ4F9g_W9-SWVbDbeUywN5cW?usp=sharing

Dataset_02: Modelling URL:
https://colab.research.google.com/drive/1GyVAuLOOLJN8qJ9eOCUGqe9JoOnTRZAt?usp=sharing

Sprint 03:

Dataset_02: URL:
https://colab.research.google.com/drive/1_NDHfmFBcRlNT6CZYk7mwJD2Ag7exgMu#scrollTo=AFHTeOOKKn95