|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| **NAME:Bhavani.P**  **. REGISTER NO:312521104005**  **. COLLEGE Name:TJ institute of technology** | | | | | |



The ’test.csv’ data file that is a part of this dataset is only being used to predict values derived from the model with the lowest WMAE score. Because, the dataset contains no target variable, in our case ’Weekly Sales’, it cannot be used for testing

for this analysis. Instead, the training dataset ’train.csv’ is being split into training and

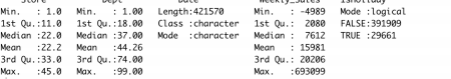
validation datasets for this study.

The main goal of this study is going to be to predict the department-wide

weekly sales for each of these stores.

The dataset is already divided into separate training and testing data; the testing

data is identical to the training dataset apart from the weekly sales information. The training dataset contains weekly sales information from 2010-02-05 to 2012-11-01 about the stores and departments. It also contains a column that suggests whether a particu- lar date falls on a holiday or not. In total, there are 4,21,570 rows in the training dataset and 1,15,064 rows in the testing dataset. (Figure 1)



A summary of the Training dataset

There is another dataset called ‘stores.csv’ that contains some more detailed infor

- mation about the type and size of these 45 stores used in this study.

Another big aspect of this study is to determine whether there is an increase in the weekly store sales because of changes in temperature, fuel prices, holidays, mark- downs, unemployment rate, and fluctuations in consumer price indexes, The

file ‘fea- tures.csv’ contains all necessary information about these factors and is used in

the anal- ysis to study their impact on sale performances.

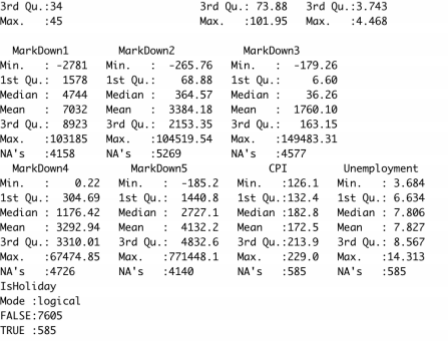
The holiday information listed in the study is:

List of holidays from the dataset

Holiday Name Date 1 Date 2 Date 3 Date 4

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Super Bowl | 12-Feb-10 | 11-Feb-11 | 10-Feb-12 | 8-Feb-13 |
| Labor Day | 10-Sep-10 | 9-Sep-11 | 7-Sep-12 | 6-Sep-13 |
| Thanksgiving | 26-Nov-10 | 25-Nov-11 | 23-Nov-12 | 29-Nov-13 |
| Christmas | 31-Dec-10 | 30-Dec-11 | 28-Dec-12 | 27-Dec-13 |

A summary of the features dataset is displayed in the image below.



A summary of the Features dataset

The final file called ‘sampleSubmission.csv’ contains two main columns: dates for each of the weeks in the study as well as a blank column that should be utilized to record predicted sales for that week based on the different models and techniques applied.

The results of the most accurate and efficient model have been recorded in this file and the final Power BI dashboard has been created based on these predicted values, in conformity with the ‘stores’ and ‘features’ dataset.



It is crucial to have an in-depth understanding of the dataset that is used in this anal- ysis to understand the models that would give the most accurate prediction. Several times there are underlying patterns or trends in the data that would not be identified as easily, hence the need for an extensive exploratory data analysis. This thorough exam- ination is necessary to understand the underlying structure of the dataset and to draw conclusions or insight about the validity of our analysis.

The study is going to begin with a brief analysis of the available dataset to get a sense of the main characteristics and components that are relevant to the research. An exploratory data analysis is crucial to this study considering the numerous attributes that are a part of the dataset that will be essential when trying to draw insights and making predictions. As part of the exploratory data analysis, several visualizations have been created that will help us understand what it is that we are trying to achieve and to keep in mind the various attributes that we can use to

improve results.

The EDA is like a primary investigation and tries to look at the relationships and

nature of the different columns available to us. As part of this, the ‘ inspectdf’ pack

* age (Ellis, 2019) and the ‘glimpse’ package (Sullivan,2019) have been used and imple
* mented in R that will answer questions related to the number and nature of columns and rows in the dataset, missing values, distribution of numeric and categorical vari- ables, correlation coefficients, etc.

Several other packages like ‘ggplot2’, ‘matplotlib’, ‘seaborn’, and ‘plotly’ have also been used in this study to create visualizations that provide information about weekly sales by store and department, weekly sales on holidays versus on normal days, weekly sales based on region, store type and store size, average sales per year, change in sales as a result of factors like CPI, fuel price, temperature, and unemployment, etc in the form of heatmaps , correlation matrix (Kedia et al . , 2013 ) , histograms , scatterplots and several more. These visualizations are accompanied by brief descriptions that will discuss the findings and scope for potential modeling that will be performed in the next stages of this project.



The ‘inspectdf’ package in R does basically what the name suggests: it inspects crucial components of a dataframe under study. For this study, it was essential to get a sense of the several datasets before they were used to create complex models.

This package inspects, compares, and visualizes the different components associ

* ated with the above-mentioned datasets. It gives a brief visualized column-wise sum- mary about missing or out of range values, distribution of values, types of columns, the correlation between numeric columns, etc. Starting with the initial explorations, it is imperative to understand the data types and ranges of the values in each column; the ‘inspecttypes()’ function from the package helps explore the data type for columns in the dataset.

The dataset contains the following five main CSV files: train, test, features, stores,

saonmd esaomf tphleeSimubpmoritsasnitoonn. eEsaacrhe fdiliesccuosnsteadinbselcorwuc: ial information relevant to this study,

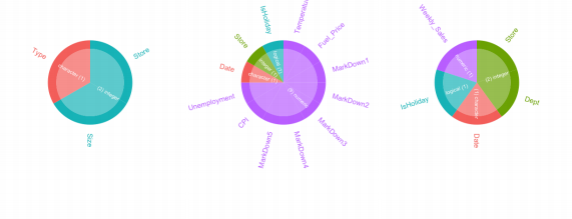
Table 2. Description of Columns

Column Name Column Type Column Description

|  |  |  |
| --- | --- | --- |
| Store Department Date  Weekly Sales  IsHoliday Fuel Price CPI  Unemployment | Categorical Categorical Categorical  Numerical Continuous Categorical Binary Numerical Continuous Numerical Continuous Numerical Continuous | 45 stores each with 143 observations  99 departments each  Weekly Sales data from 2010 until 2012  Sales Ranging from 2,00,000 to 38,00,000  0 and 1 values associated with date Prices ranging from 2 . 4 to 4 . 4 Values ranging from 126 to  227 Values ranging from 3 to  14 |

A visualization of the ’inspect df()’ package for the features, train, and

stores datasets can also be observed in Figure 3.

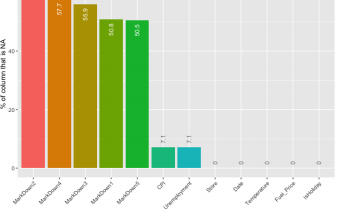


* 1. Data Type: Stores Dataset (b) Data Type: Features Dataset (c) Data Type: Training Dataset

Data Types

Another important function from the package is the ‘ inspect na()’ function that highlights the missing values in each column of the dataset used for this study. Apart from the ‘features.csv’ file, it was found that no other dataset used had any missing val- ues in any of the columns. The major missing values in the features dataset come from

the markdown columns that contain information about the different promotional ac- tivities happening at different Walmart stores. A reason behind such a massive amount of missing values in these columns is due to the seasonal promotional prices set by the stores during holidays (that mostly happen to start from November until January) .



Missing Values in the Features dataset

The next function from the package looks at the distribution of the numeric vari

* ables using histograms created from identical bins. Considering that the features dataset has the most numeric variables, that is the only one that will be looked at in detail. According to the package website, ‘The hist column is a list whose elements are tibbles each containing the relative frequencies of bins for each feature. These tibbles are used to generate the histograms when showplot = ‘TRUE’.

The histograms are represented through heat plot comparisons using Fisher’s exact

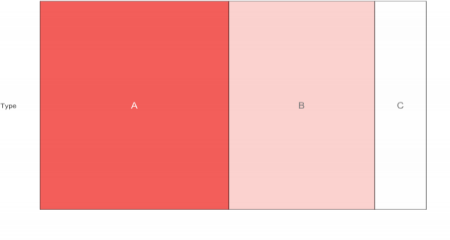
test to highlight the significance of values within the column; the higher the signifi

* cance, the redder the data label (Rushworth, n.d.).



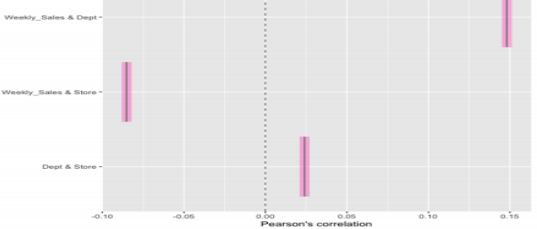
Distribution of Numerical attributes in the Features dataset

After looking at the distribution of numeric variables, the study proceeds to look at the distribution of categorical variables. This is done using ‘ inspect cat()’; this function looks at the complete distribution of values in the categorical columns. Considering there are not many categorical variables in the datasets used for this study (as seen above through ’inspect types()’), the only relevant information that was gathered using the function was the distribution of types of stores in the ‘stores’ dataset.



Distribution of Store Types

From the image above, it is clear that a majority of the Walmart stores included in this study belong to Type ‘A’ . This will be briefly discussed in the coming sections of the study when advanced EDA will be used to answer some important questions.

The last significant function from the ‘inspectdf()’ package is called ‘inspect cor()’. This function, in a nutshell, enables users to look at Pearson’s correlation between dif- ferent variables in a dataframe. Understanding if there is an association between vari- ables beforehand will answer a lot of questions about what variables

.

Correlation between attributes of the training dataset

A look at the correlation between the variables of the training dataset depicts that while there is a slight association between the weekly sales and department, it is still not as significant (higher the R-value, higher the significance).



The second section under this Exploratory Data Analysis looks at advanced and exten

* sive visualizations that answer some crucial questions about the Walmart

dataset, as listed in the purpose statement.

After inspecting crucial elements in each of the data frames about the types of vari

* ables, their distribution, correlation, and association, etc. using ‘inspectdf’, more de- tailed and summarized information about the weekly sales for each department/store and the effect of various factors on the weekly sales are studied here. This is performed using a combination of R and Python packages like ‘ggplot2’, ‘matplotlib’, ‘seaborn’, ‘plotly’, and several others.

This section will aim at looking at the following aspects of the Walmart dataset and also possibly look at some more crucial information that stems out from the below- mentioned criteria:

Day, etc.

There is an evident hike in sales in weeks 4 7 and 5 1 that correspond to Thanksgiving

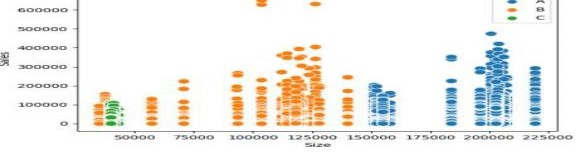
and Christmas respectively, proving again that sales rise during the holiday season. Due to the insufficiency of data for the year 2012, these conclusions have only been made based on the data available from 2010 and 2011. This graph also tells that there is a distinguished pattern of decline immediately following Christmas and New Year’ s .

After studying the overall sales summaries of different components of the Walmart dataset, this report will now throw light upon the effect of different factors (such as holidays, markdowns, CPIs, unemployment, etc.) on the weekly sales. It has always been an integral part of this study to understand the effect that these factors have on Walmart’s sales performance. I will also create several visualizations that shed light on the difference in Walmart store sales on holidays versus non-holiday days, the impact of store size and type on weekly sales, and finally create a correlation matrix to examine the correlation between the many factors included in the study.



It is known from the previous visualization that Type ‘A’ is the biggest store type fol

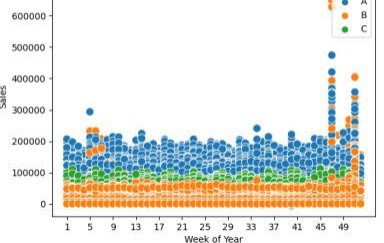
* lowed by Type ‘ B’ and ‘C’ . The graph below shows a linear relationship between the size of a store and their consequent sales, with some exceptions. A few Type B stores, as depicted below, acquire more average sales than Type A stores, going against the gen- eral idea of the bigger the size of the store, the higher the sales. But in general, type A stores still show a high amount of sales while Type C stores show a significantly small amount of sales. To summarize, sales generally increase with an increase in the size of the store, with some minor exceptions.



Impact of Size of Store on Sales



This visualization is similar to Figure 14 in the way that it shows a summarized view of the average sales for each week of a year. The difference with this scatter plot is that it shows the average weekly sales for each store type and helps understand if sales go up for each store at the end of the year. With this plot, it is clear to understand that, unlike Store Type A and B, the average sales do not necessarily go up for Type C at the end of the year around Thanksgiving and Christmas. It also shows that Type A stores typically have higher weekly sales as compared to the other two stores, proving once again that a bigger store size signifies higher sales.





It has widely been known in the retail sector that weather has a profound effect on sales. While warmer weather promotes sales, cold/harsh or extremely hot weather is generally not a great encouragement for shoppers to get outdoors and spend money. Generally speaking, temperatures between 4 0 to 7 0 degrees Fahrenheit are considered as favorable for humans to live in considering they are not as hot or cold.

As seen below, the highest sales occur for most store types between the range of 40 to 80 degrees Fahrenheit, thus proving the idea that pleasant weather encourages higher sales. Sales are relatively lower for very low and very high temperatures but seem to be adequately high for favorable climate conditions.

The heatmap/correlation matrix in Figure 22, created using the seaborn library in Python (Szabo, 2020) gives the following information:

* There is a slight correlation between weekly sales and store size, type, and de- partment
* There seems to be a negative correlation between weekly sales and temperature, unemployment, CPI , and fuel price. This could suggest that sales are not im- pacted by changes in these factors
* Markdowns 1-5 also seem to have no distinct correlation with weekly sales, thus they are not as

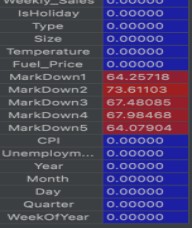
important a factor in the study



The data contains 421 , 570 rows , with some store -specific departments missing a few too many weeks of sales. As observed in Figure 4, some columns in the features dataset contain missing values, however, after the features dataset is merged with the training dataset, the only missing values that exist are in the Markdown columns (as shown in figure 23).

After the extensive EDA, it was determined that these five markdown files, with missing values,have barely any correlation to the weekly sales for Walmart, hence these five columns have been eliminated from the subsequent training and testing dataset. Because the source already provides training and testing datasets, there is no need to create them for our study.

Because the main focus of this study is to accurately predict weekly sales for dif- ferent Walmart stores, the previously modified ‘ Date’, ‘Month’, ‘Quarter’, and ‘ Day’ columns have been dropped and only the ‘Week of Year’ column has been used in the upcoming models.



Data has been checked for inaccuracies, missing or out of range values using the ‘ inspectdf’ package in R as part of the initial EDA. Columns with missing values have been dropped. The dataset contains information about weekly sales which was ini- tially broken down to acquire information about monthly as well as quarterly sales for our analysis, however, that information is not going to be utilized during the modeling process.

The boolean ‘ isHoliday’ column in the dataset contains information about whether the weekly date was a holiday week or not. As observed in the EDA above, sales have been higher during the holiday season as compared to non-holiday season sales, hence the ‘ isHoliday’ column has been used for further analysis.

Furthermore, as part of this data preprocessing step, I have also created input and target data

frames along with the training and validation datasets that help accurately measure the

performance

of applied models. In addition, as part of this data prepro- cessing, feature scaling (Vashisht, 2021) has been applied to normalize different data attributes. This has primarily been done to unionize the independent variables in the training and testing datasets so that these variables will be centered around the same range (0,1) and provide more accuracy.

Also referred to as normalization, this method uses a simple min-max scaling tech- nique

(implemented in Python using the Scikit- learn (Sklearn) library (Pedregosa et al.,

2011).

Lastly, as part of this competition, Walmart wished that participants assess the ac- curacies of models using the Weighted Mean Absolute Error (WMAE) (“Walmart Re- cruiting - Store Sales Forecasting , ” 2014 ) , a brief description of which is displayed as follows .

where

WMAE =

⼭i |yi − |

i

⼭

* n is the number of rows

•  is the predicted sales

* yi is the actual sales
* ⼭i are weights. ⼭ = 5 if the week is a holiday week and 1 otherwise

TheWeighted Mean Absolute Error is one of the most common metrics used to measure accuracy for

continuous variables (JJ, 2016).

AWMAE function has been created that provides a measure of success for the dif- ferent models applied. It is the average of errors between prediction and actual ob- servations, with a weighting factor. In conclusion, the smaller the WMAE, the more efficient the model.



Trying to find and implement the most effective model is the biggest challenge of this study. Selecting a model will depend solely on the kind of data available and the anal- ysis that has to be performed on the data (UNSW, 2020).

Several models have been studied as part of this study that were selected based on different aspects of our dataset; the main purpose of creating such models is to predict the weekly sales for different Walmart stores and departments, hence, based on the nature of models that should be created, the following four machine learning models have been used:

* Linear Regression
* Lasso Regression
* Gradient Boosting Machine
* Random Forest

Each of these methods have been discussed briefly in the upcoming report. For each of the models, why they were chosen, their implementation and their success rate (through WMAE) have been included.