Srilasya Garigipaty

Superstore Sales Forecasting Project Analysis

Student # 100822953

DATA 1202

**Project Introduction:**

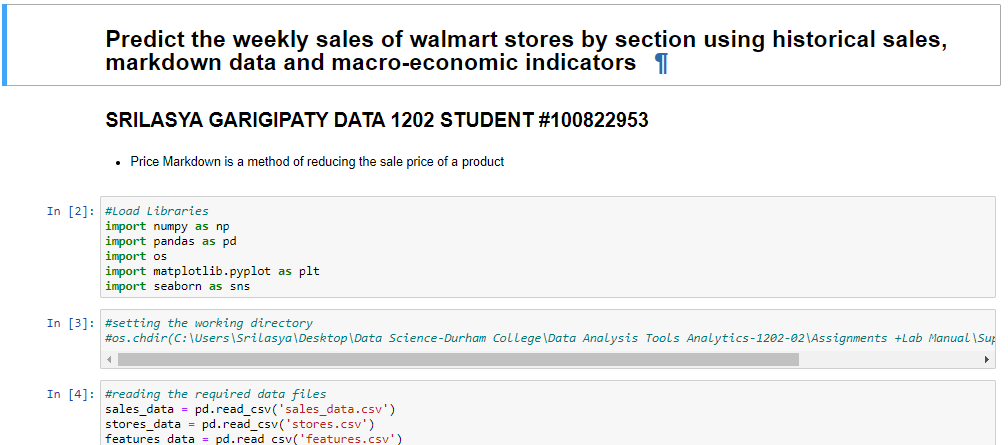
Every department store chain wants to predict the store sales in the nearby future so that they can plan for inventory. Sales prediction helps increase/decrease store staff base on the rush. More sales indicate that there are more customers coming to the store. Sales and revenue forecasting will help to better understand a company’s cash flow and overall growth.

To plan for inventory, it is important to know what products or category of products will be utilized more. Understocking can occur, and sales are affected. Overstock items such as perishables and loss can occur due to loss and product expiry. Sales Prediction is done at a combination of store and department level and also at product level for high selling products.

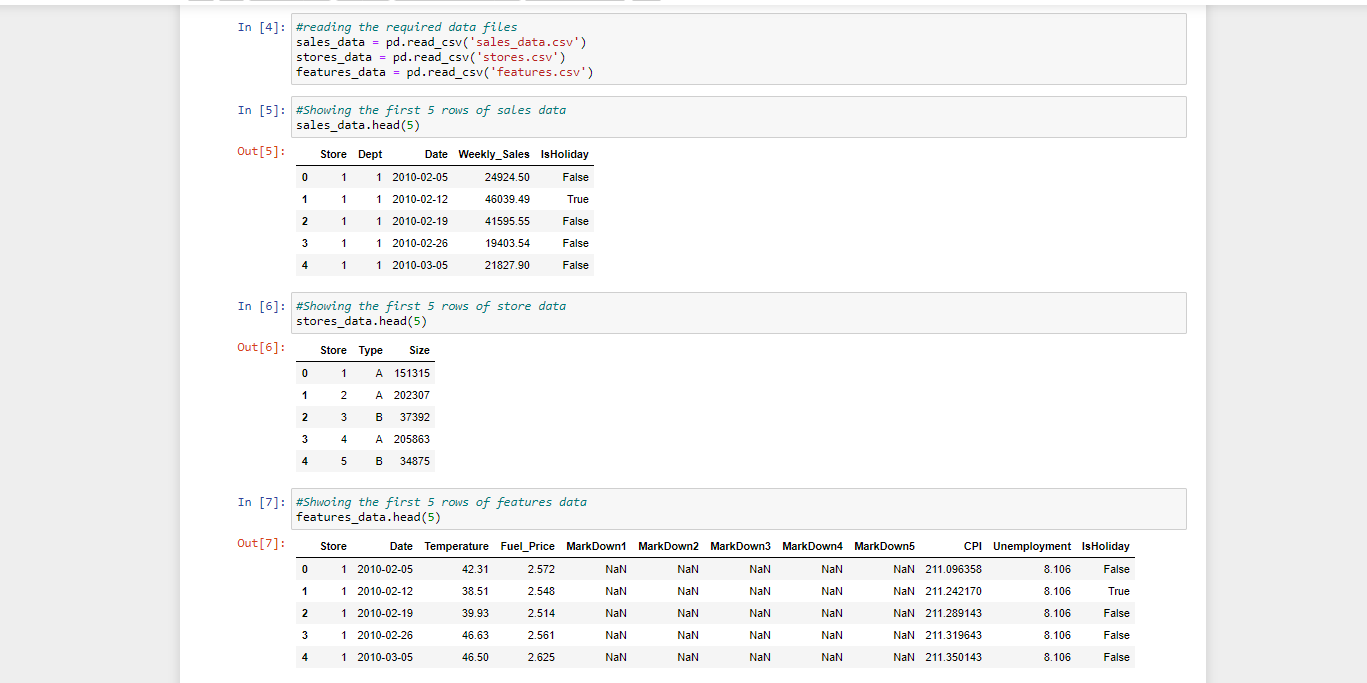
In this project (walmart\_sales.ipynb) the sales data includes 45 stores based on store, department, and week. The size and type of each store is provided. Holiday weeks have been marked. Along with this, price markdown data (discount data) has been given. A few macro indicators such as CPI, Unemployment rate, Fuel Price etc. are also provided.

**Project Concepts:**

* Understand the problem Statement clearly and how it is applicable to the scenario of Sales Forecasting.
* Perform basic EDA to familiarize with the data.
* Take care of missing values and datatype issues in the data.
* Understand unique key in different data and Merging the data.
* Perform Univariate analysis for both numeric and categorical variables.
* Plot trend of each predictor with target variable
* Do an in-depth analysis on the impact of Date/Week on Sales
* Create new features that might add value to model.
* Define a function for each set of code that might need to be repeated again.
* Prepare the data for modelling.
* Make prediction using statistical techniques.
* Make model using machine learning techniques.
* Create time series ARIMA models and learn to give their parameters.
* Perform Hyper-parameter tuning to get the best parameters.
* Learn how to make predictions where data is sparse.
* Compare the performance of different models using multiple metrics.
* **Load the Libraries necessary to begin to tackle the project problem.**

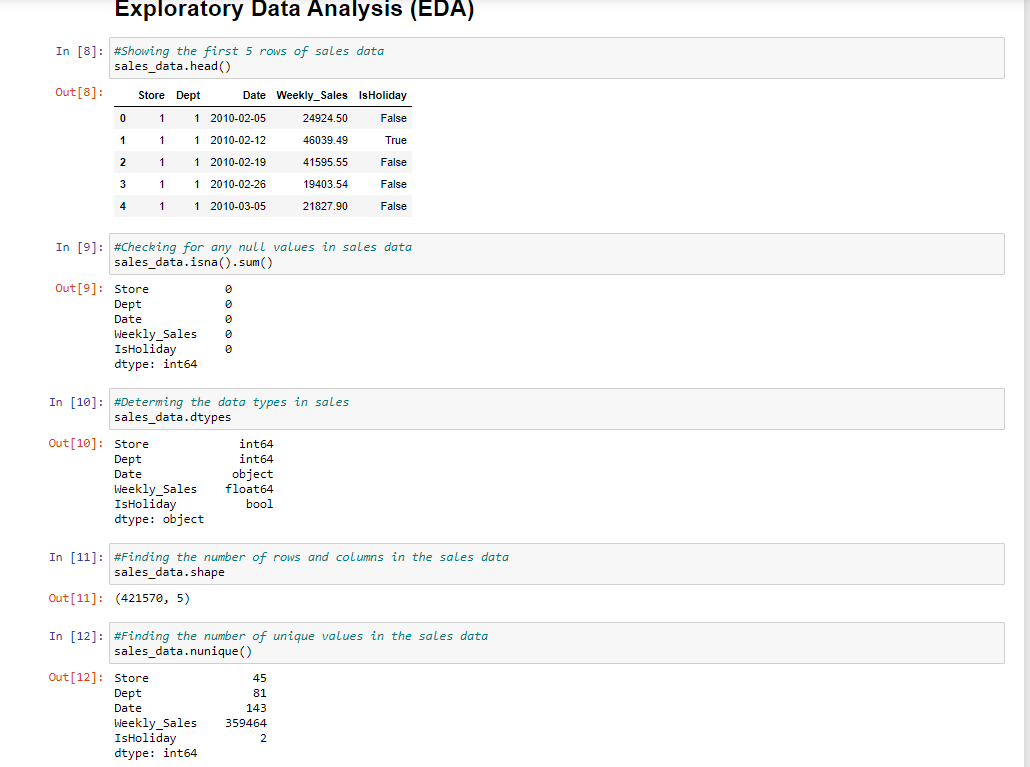


* **Open the sales\_data, stores\_data,and features\_data csv files and reading them to get an understanding of what data they contain and their columns.**

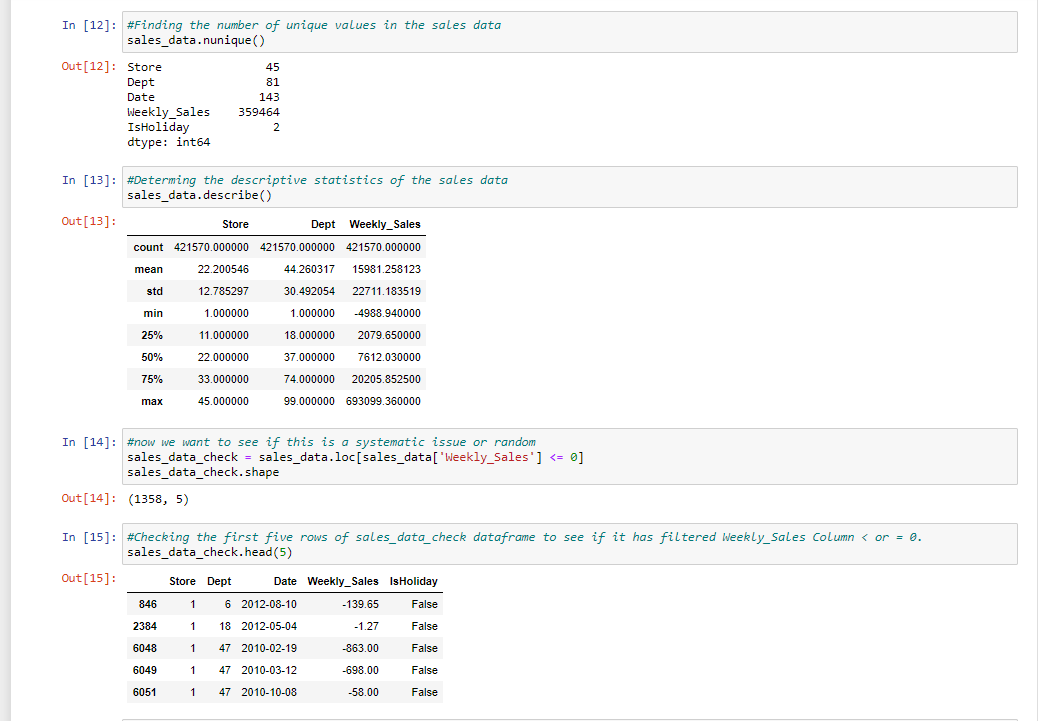


* **Exploratory Data Analysis (EDA)**

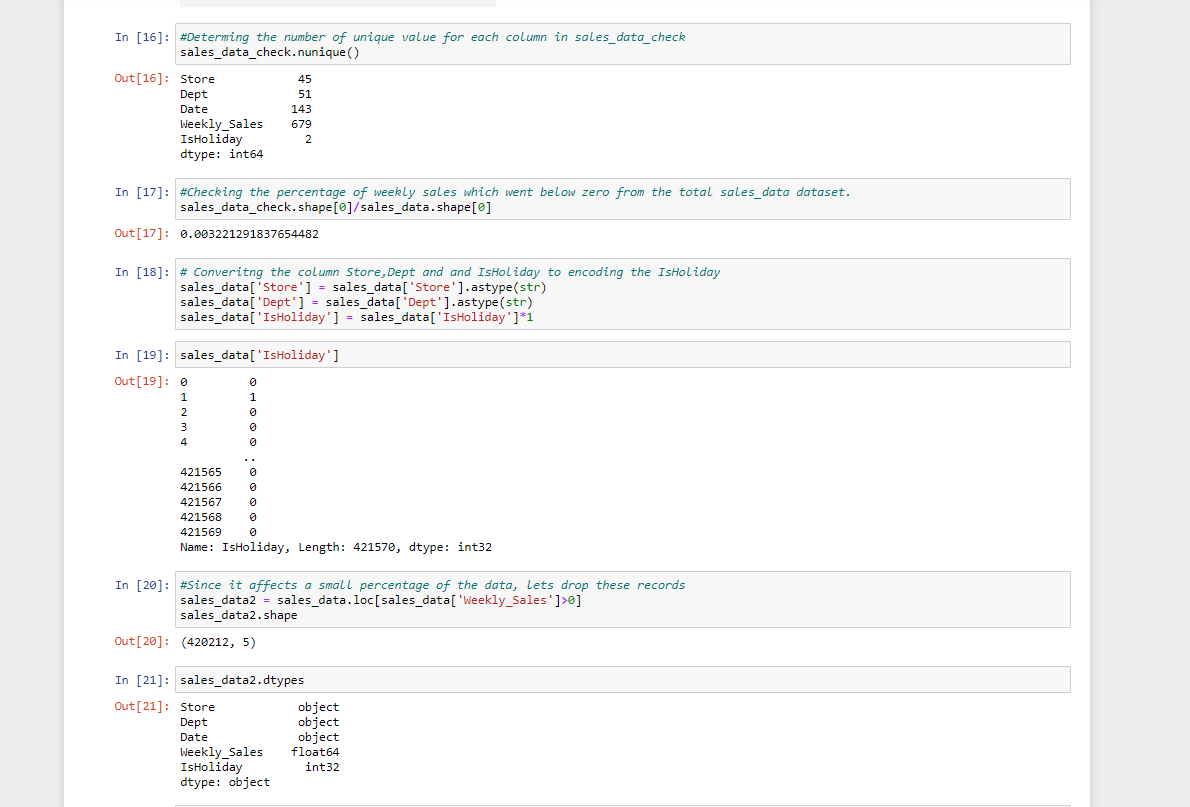
**This step is used to analyze and investigate data sets and summarize their main characteristic. This step often employs visualization methods and helps determine how to best manipulate the data sources to get the answers, discover patterns, spot anomalies, test a hypothesis, or check assumptions.**



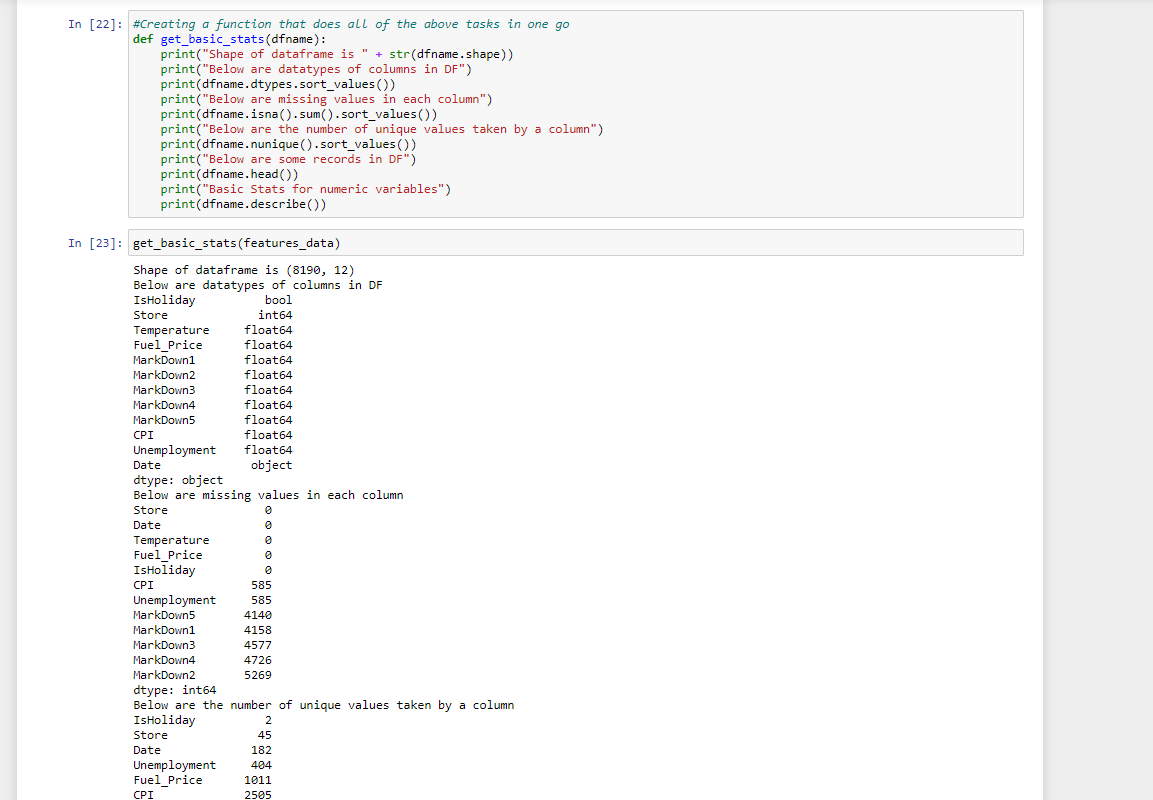
* **Checking the Descriptive Statistics of the data is important because the descriptive statistics summarizes the data and are broken down into measures of central tendency (mean, median, and mode) and measures of variability (standard deviation, minimum/maximum values, range, kurtosis and skewness).**
* **Sales Data was Also Filtered to Separate negative weekly sales drop, and positive weekly sales increase from the sales data.**



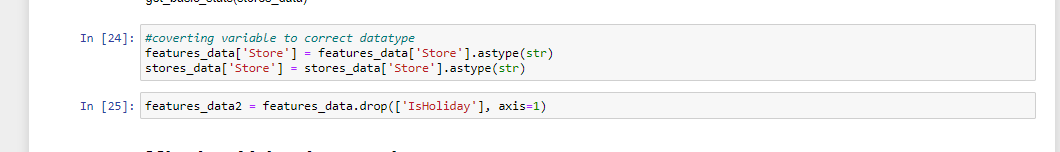
* **Categorical Variables are encoded, and data types are changed for easier data manipulation. In addition, the percentage of weekly sales that went below zero compared to the total weekly sales was calculated and found to be .32% only. This means that the superstore is not taking much of a loss despite weekly sales dropping sometimes.**



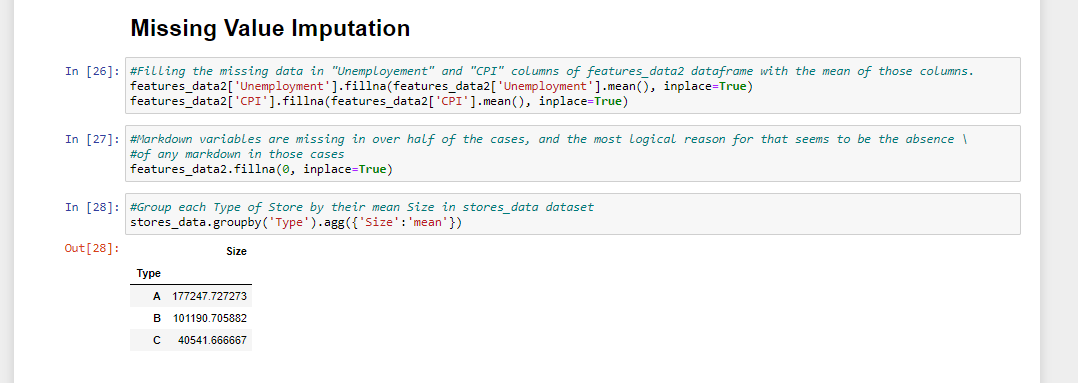
* **A function was created that does all of the EDA processing in one go. This is an efficient and quick method to get all the information needed about the data. It tells the shape, datatypes, missing value count, number of unique values in a column, basic statistics for numerical variables, and a list of records in DataFrame.**



* **Data Types are changed to “string” format and the categorical column of the features data is dropped for further data manipulation.**

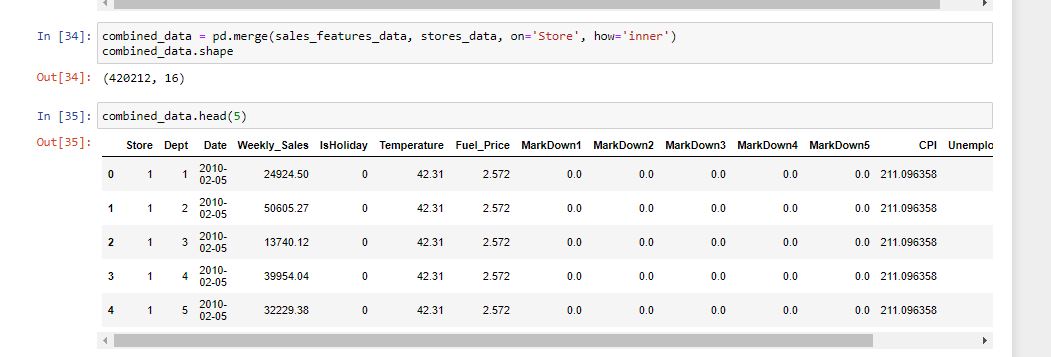


* **Missing Value Imputation**
* **Many real-world datasets contain missing values, encoded as blanks, NaNs or other placeholders. A missing value imputation allows for a better machine learning algorithm quality, and a good way to deal with missing values is to infer them from part of the data rather than get rid of them altogether as this comes with the risk of losing data that may be valuable though it is incomplete.**



* **Data Preparation**
* **Data Preparation is the process of cleaning and transforming the raw data prior to process and analysis. It is an important step that occurs prior to processing the data and involves reformatting the data, making correction to the data, and combining data sets to enrich the data.**
* **Data Preparation Process is lengthy and involves standardizing data formats, enriching source data, and/or removing outliers.**

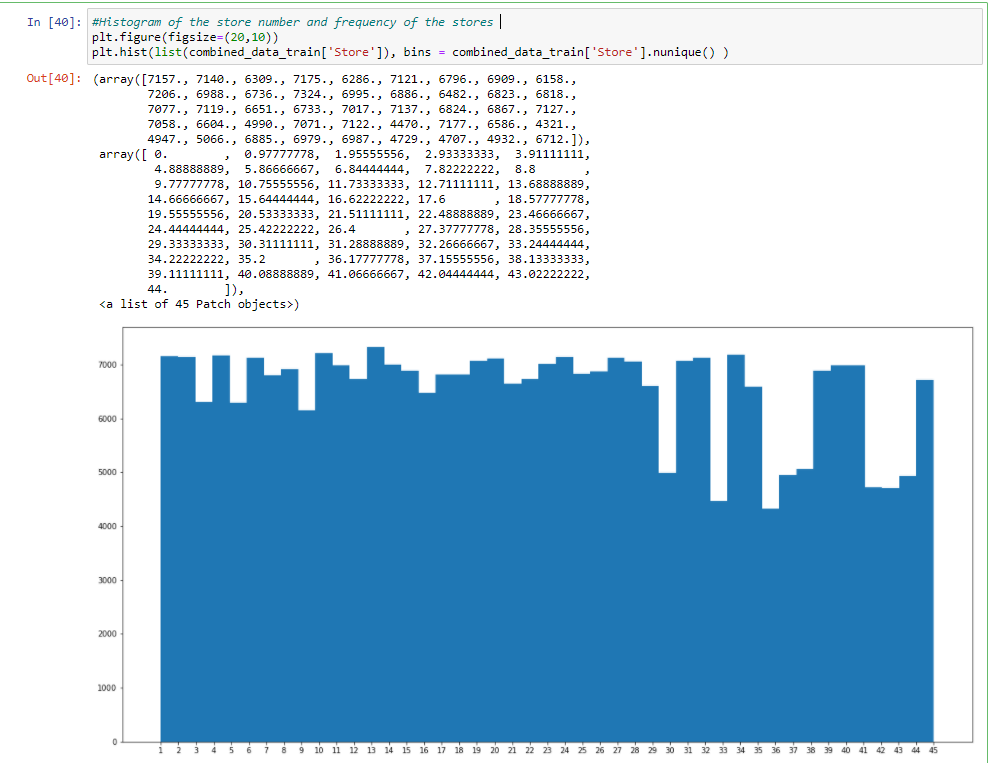


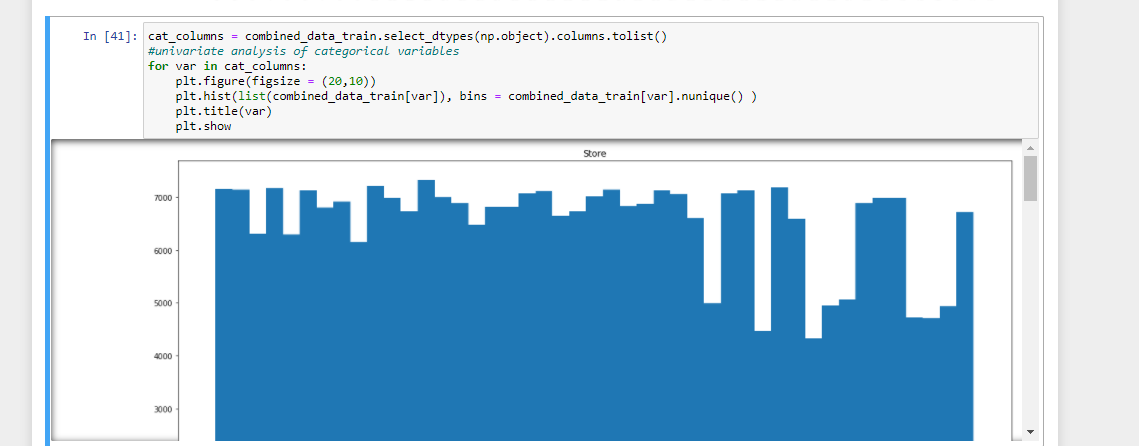


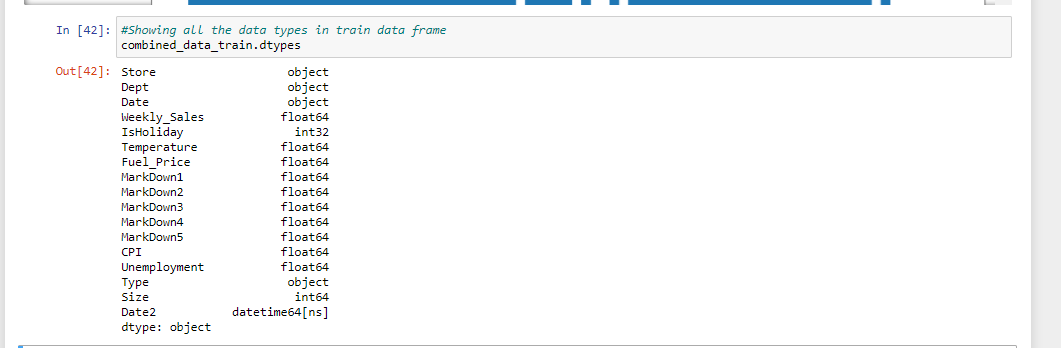
* **Dividing the Data into Train and Test. The train / and test procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model. Train/Test Split is a method used to measure the accuracy of the model. It is called Train/Test because the data is split into two sets: a training set and a testing set.**
* **The train-test split procedure is appropriate when you have a very large dataset, a costly model to train, or require a good estimate of model performance quickly.**
* **How to use the scikit-learn machine learning library to perform the train-test split procedure.**
* **Here a train test split of 70 to 30 ratio is taken.**

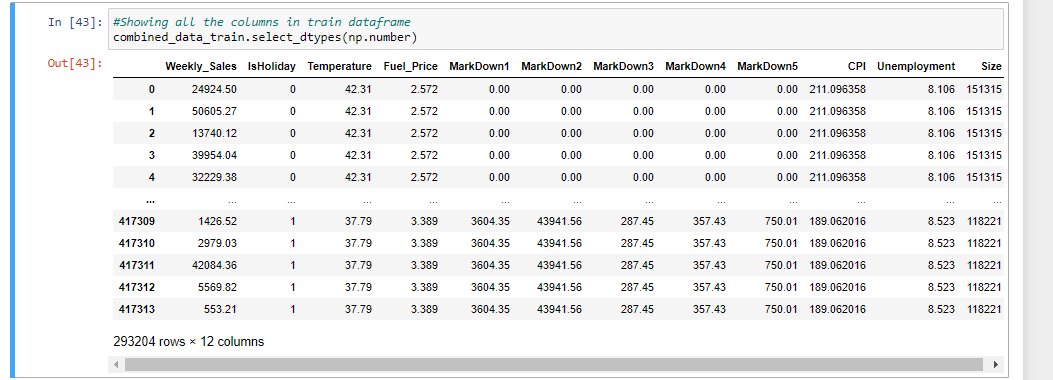


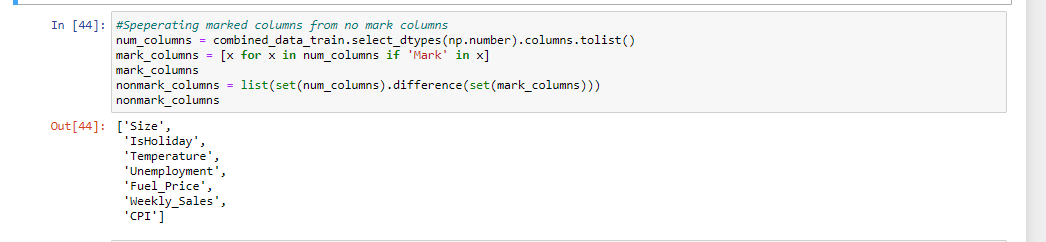
* **Univariate Analysis**
* **Univariate analysis refers to the analysis of one variable. The purpose of univariate analysis is to understand the distribution of values for a single variable. The most common way to perform univariate analysis is through performing univariate analysis, frequency distributions, and charts.**

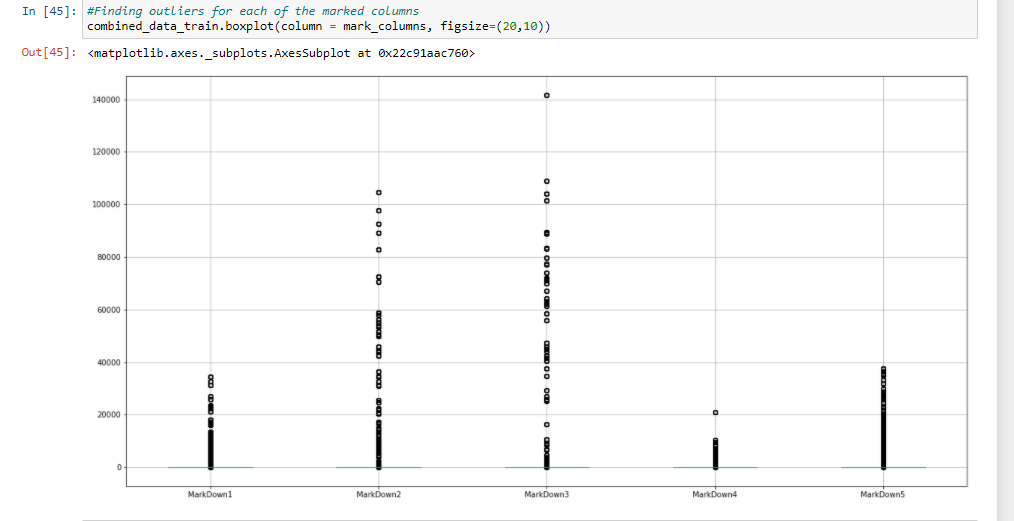


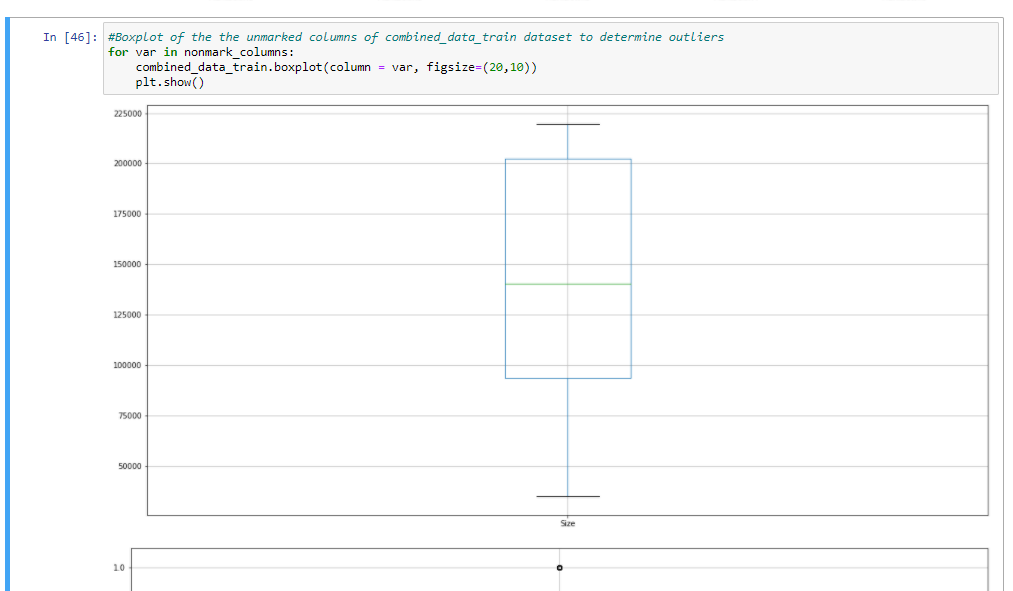




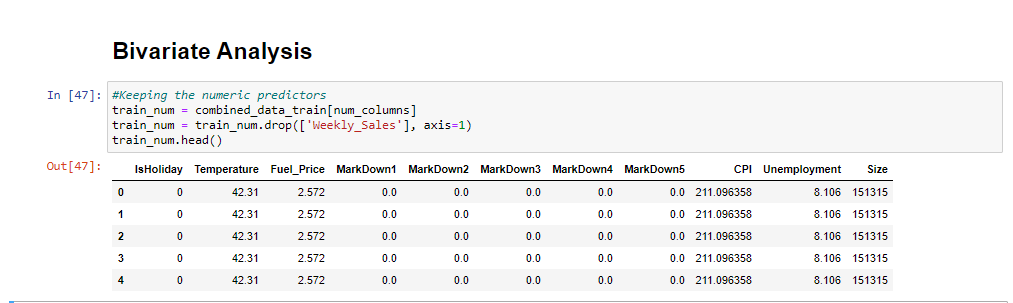


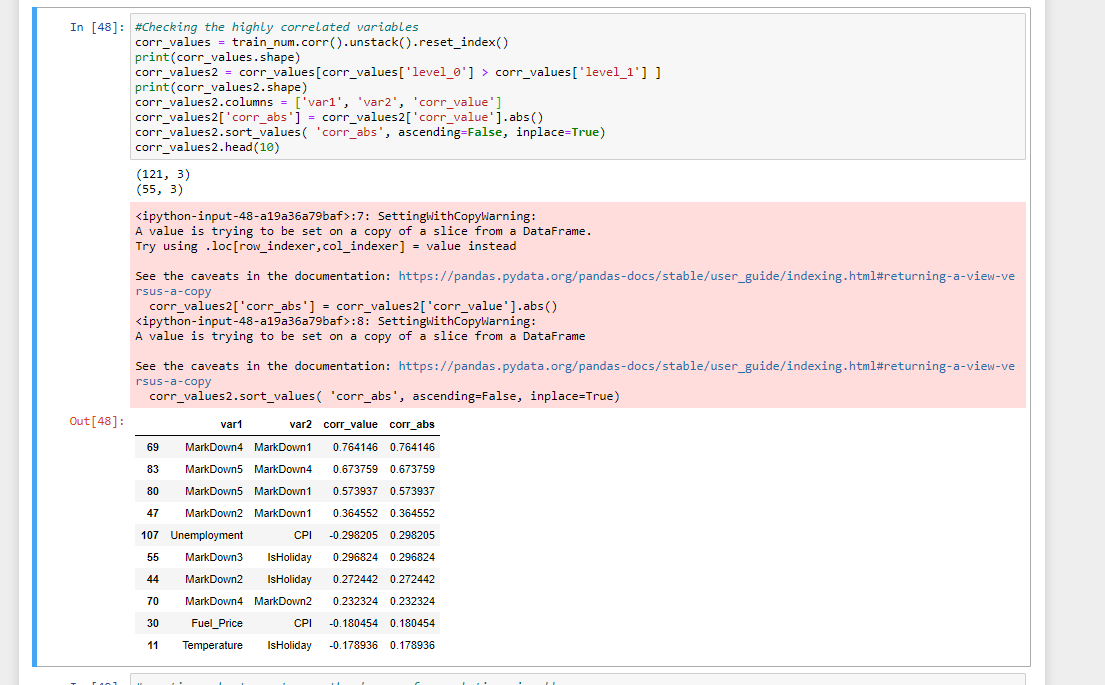


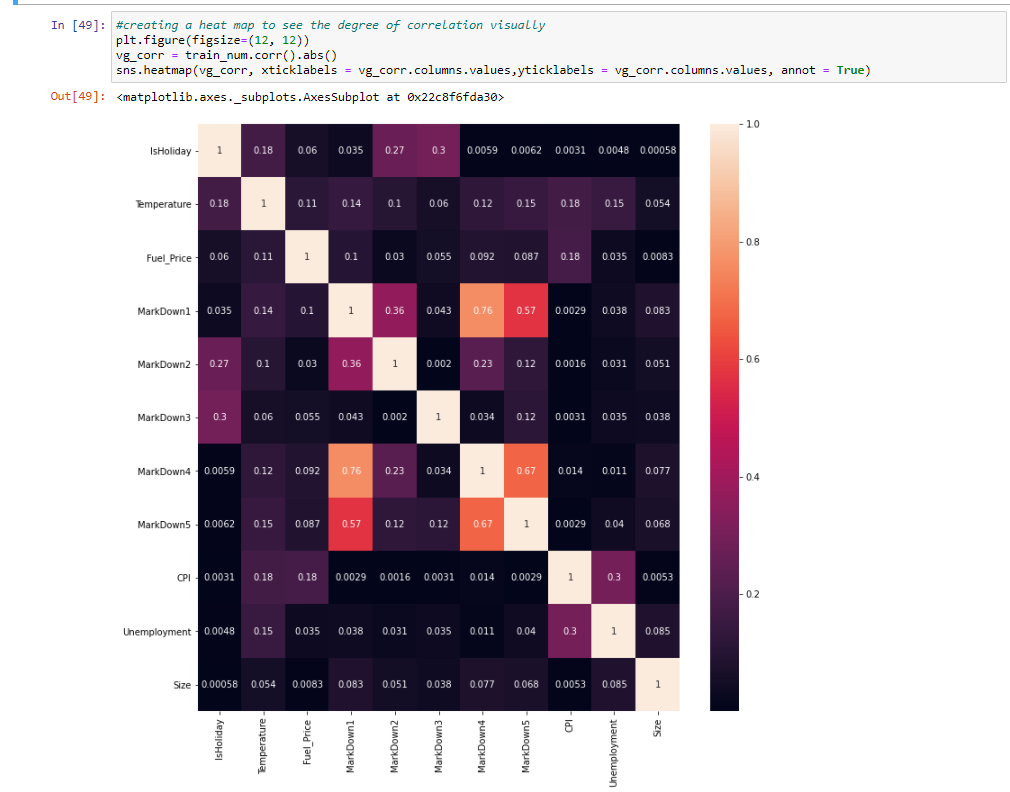


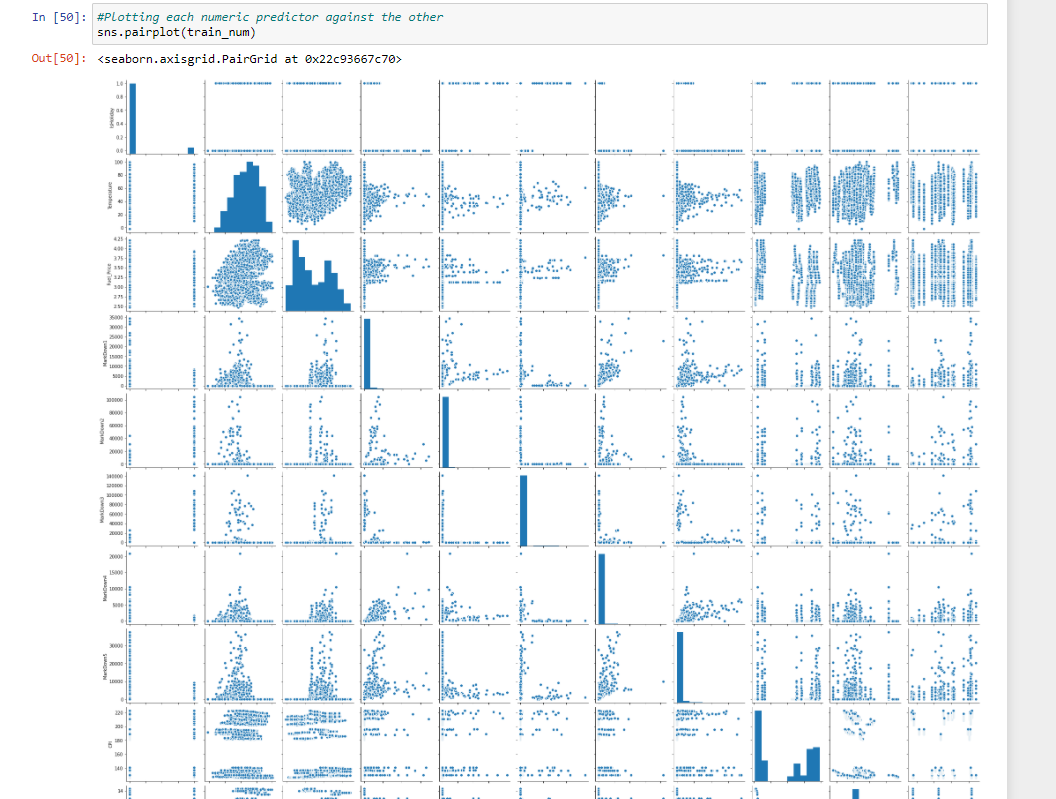


* **Bivariate Analysis**
* **Bivariate Analysis is a statistical analysis where two variables are involved. One variable is dependent while the other variable is independent. The variables are usually denoted by x and y. In Bivariate Analysis, changes between two variables are analyzed and to what extent those changes occur.**

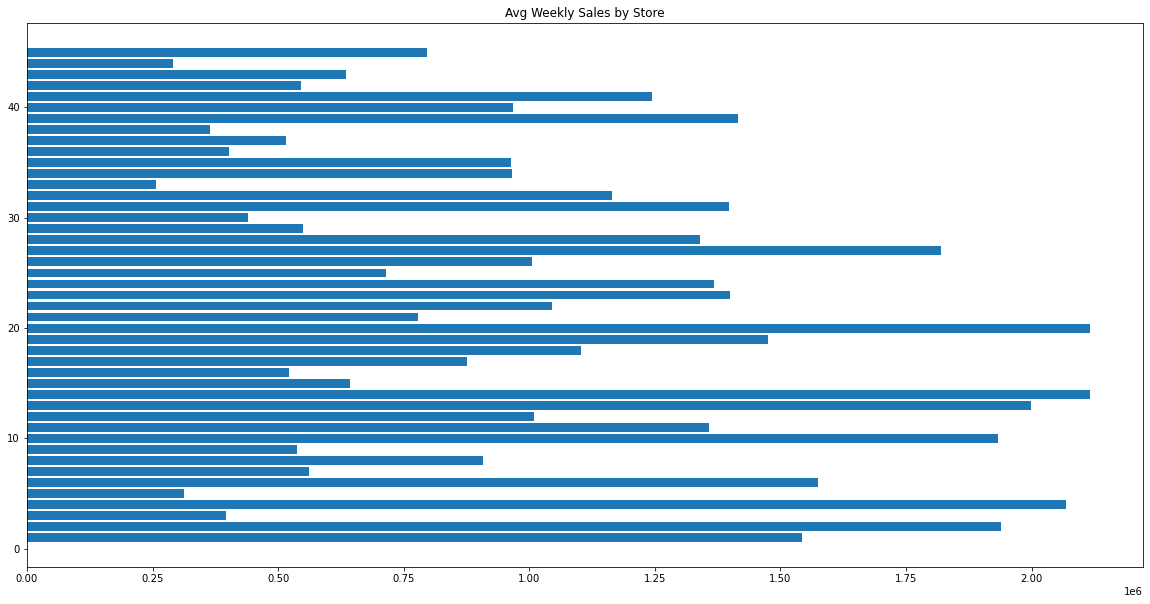


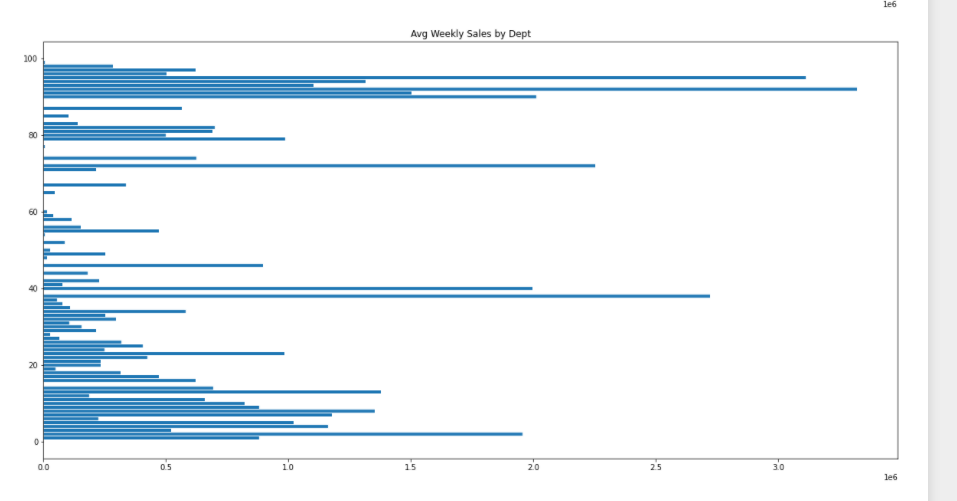


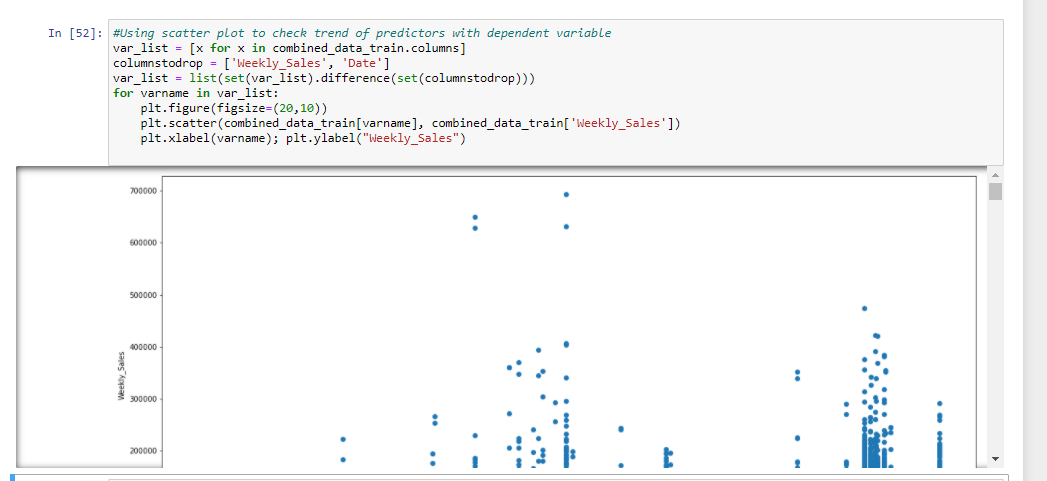




* **Variable Trend with the dependent variable is done to determine how different independent variable effect the dependent variable of Weekly Sales. The Average Weekly Sales are plotted against the independent variables of Store, Dept, Size, Markdown2, Markdown3,IsHoliday,Unemployment, Fuel\_Price,Type,Markdown5,CPI,DeptTemperature,Markdown4,Date2,Markdown1.**

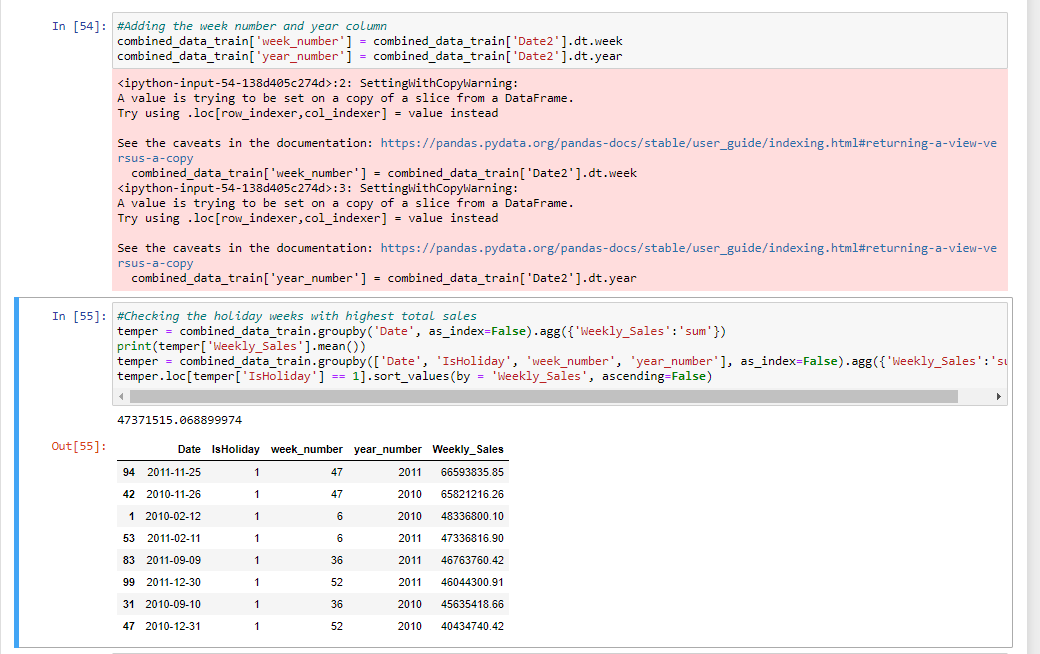
****

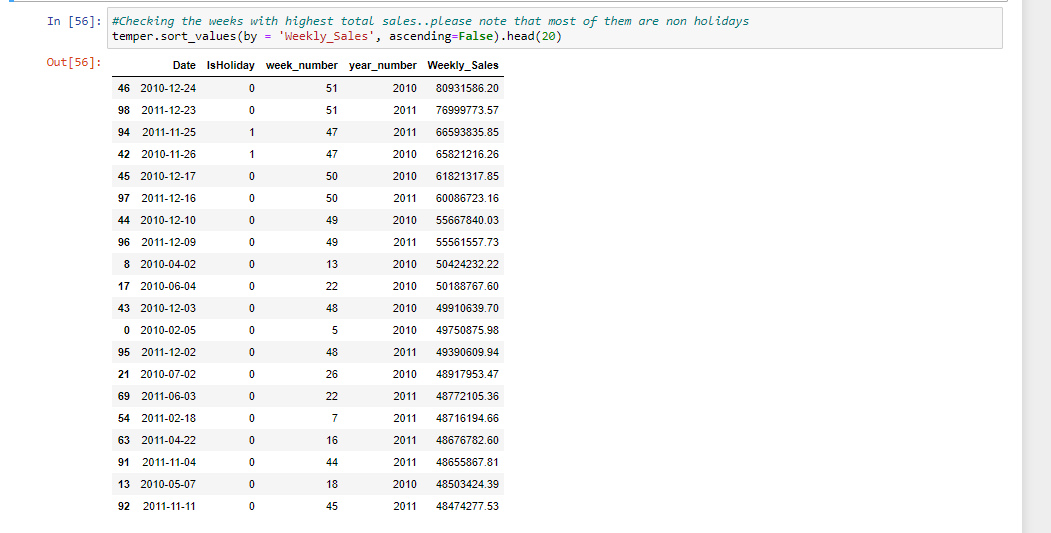


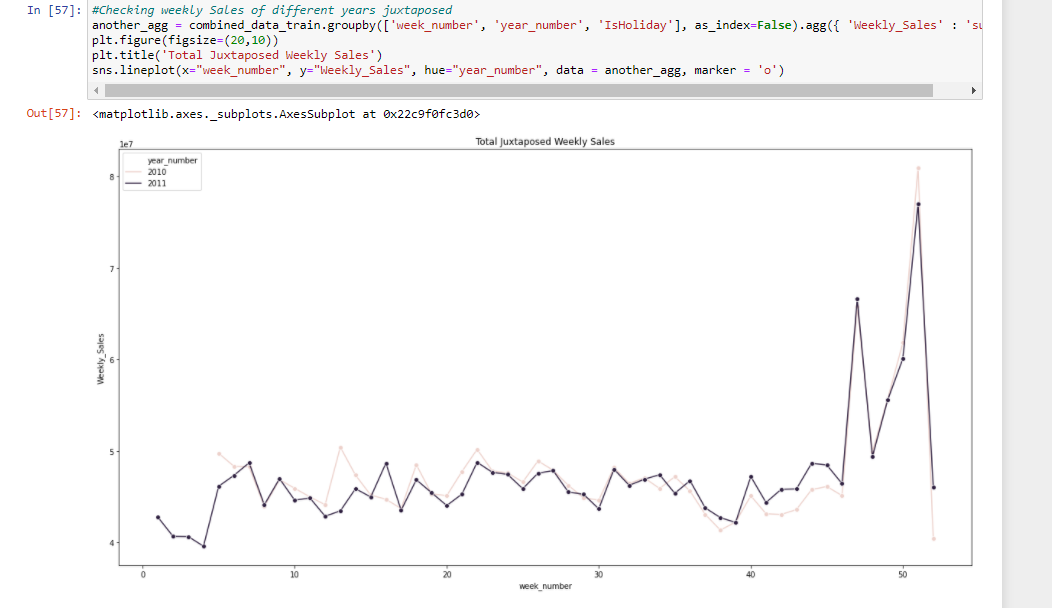


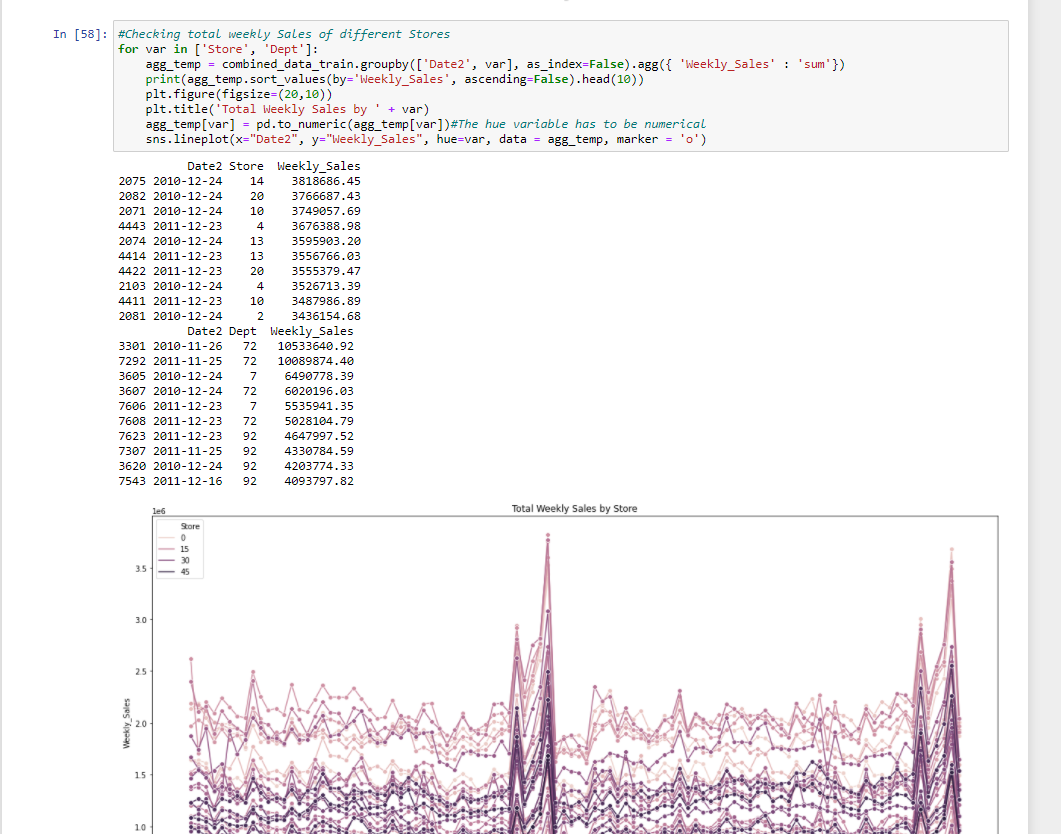


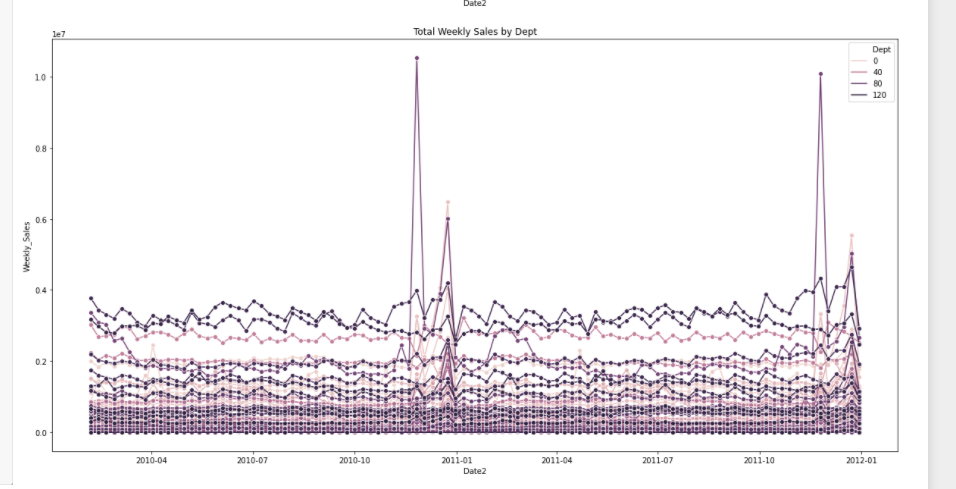
* **Date Trends**
* **In this step the week number and year columns are added. The dates are grouped and analyzed based on weekly sales per week number, total weekly sales by store, total weekly sales by department, etc.**

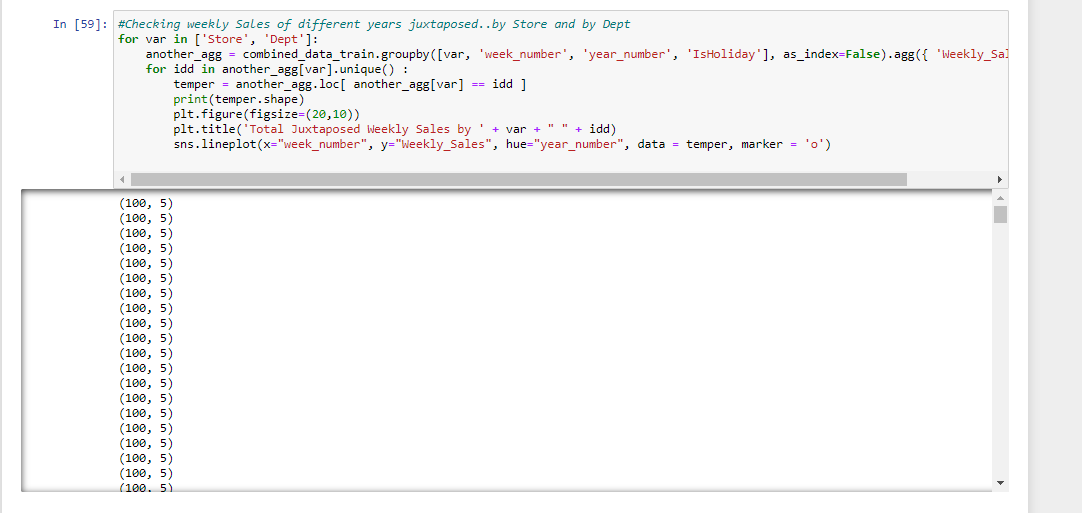








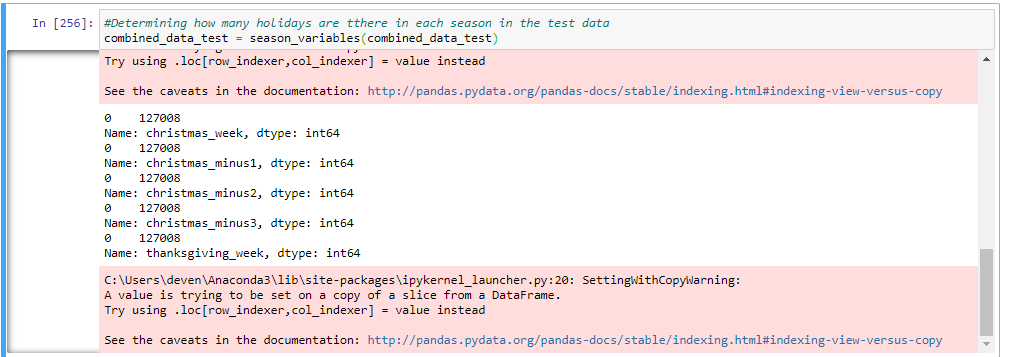


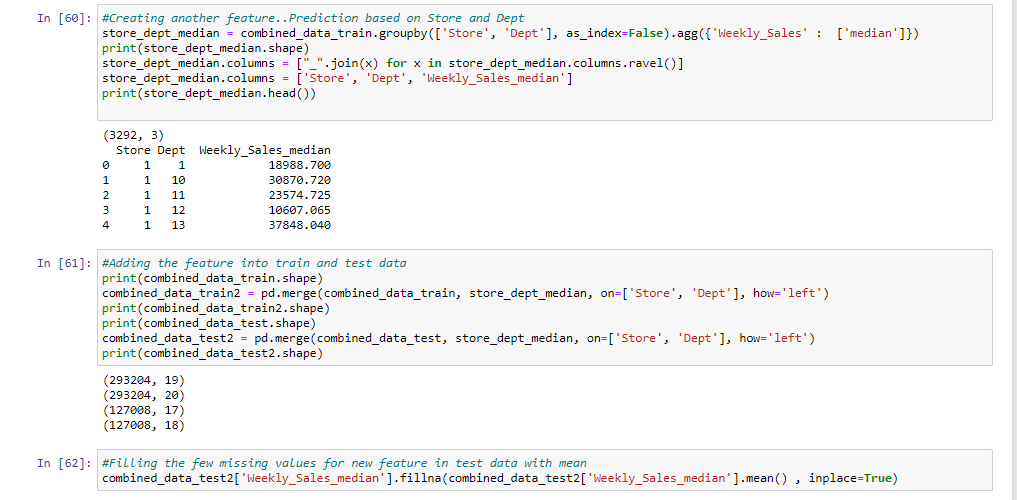


* **Feature Creation is the creation of new input or target features from existing features. The objective of Feature Creation is to create features that represent a machine learning problem to the model that can improve its accuracy. Good Feature Creation and Feature Engineering can be the difference between a poor model and a good model.**

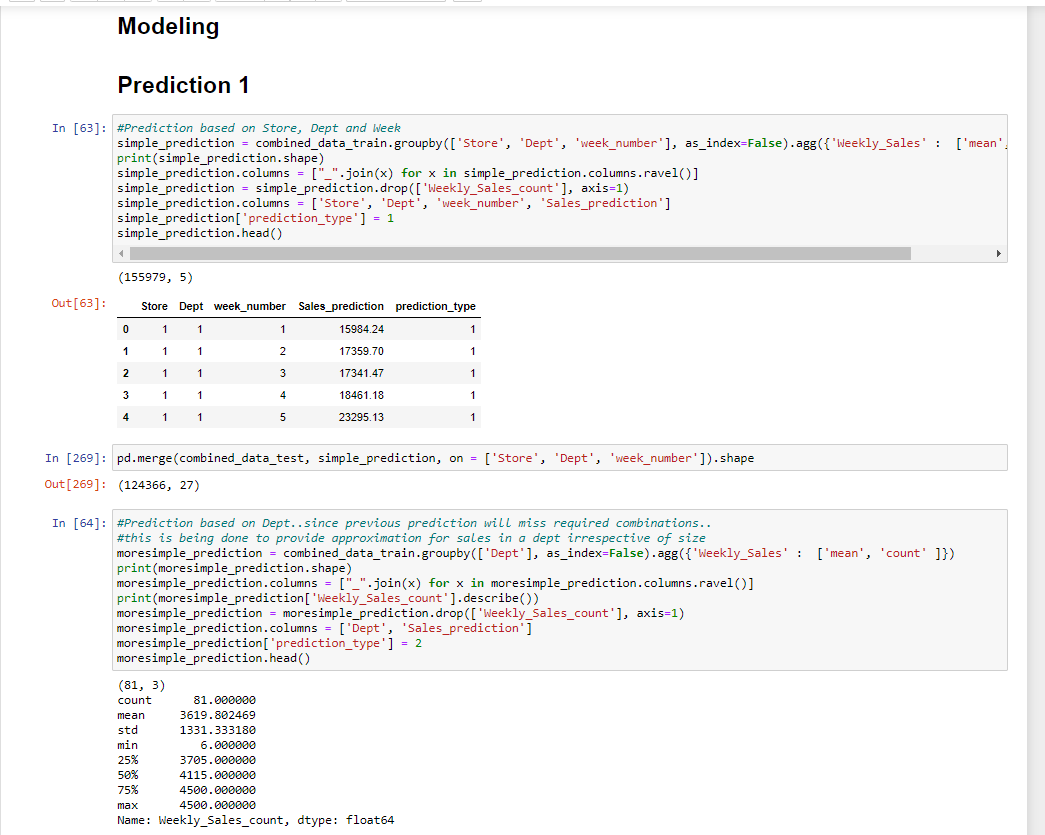


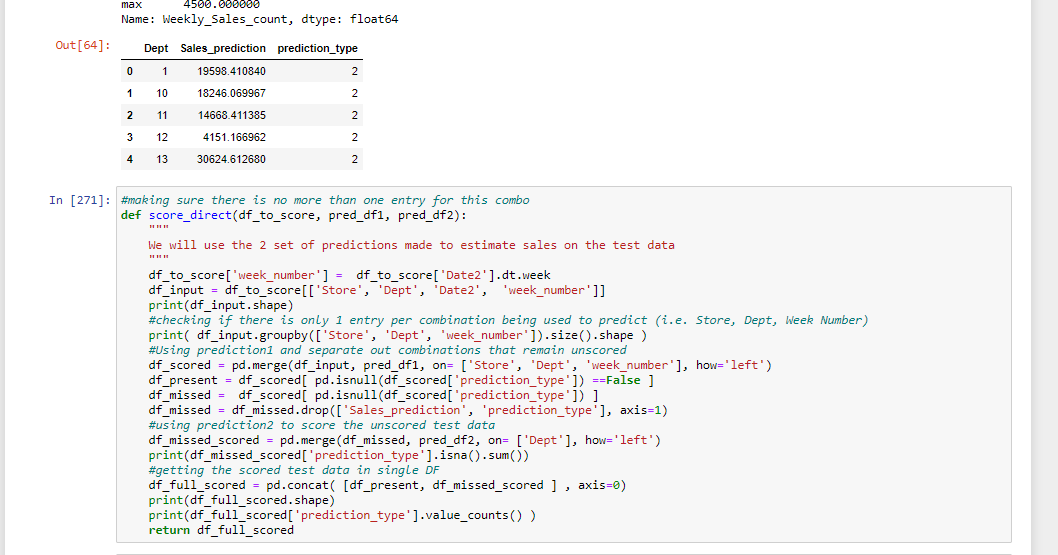


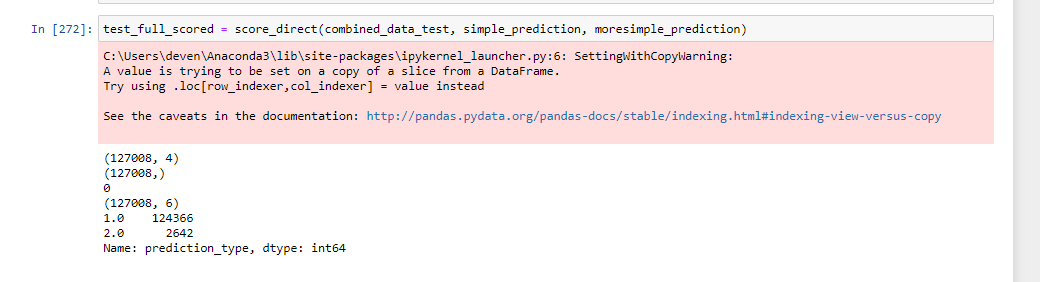


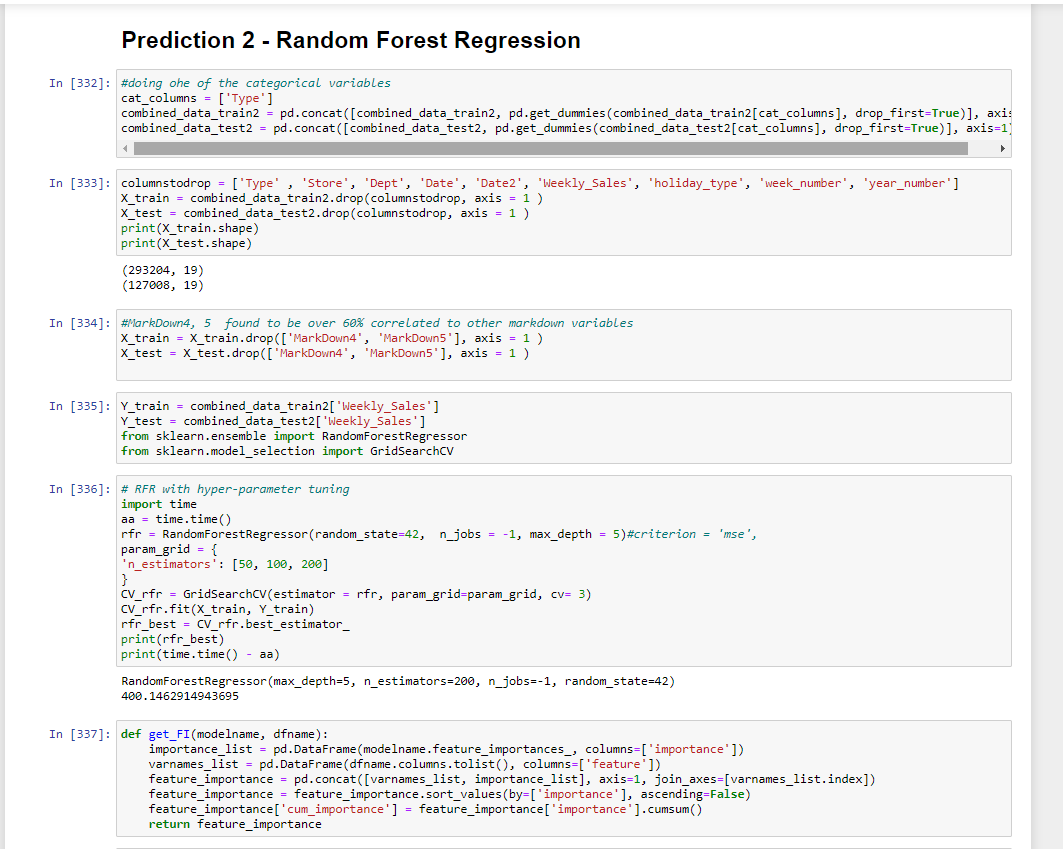
data

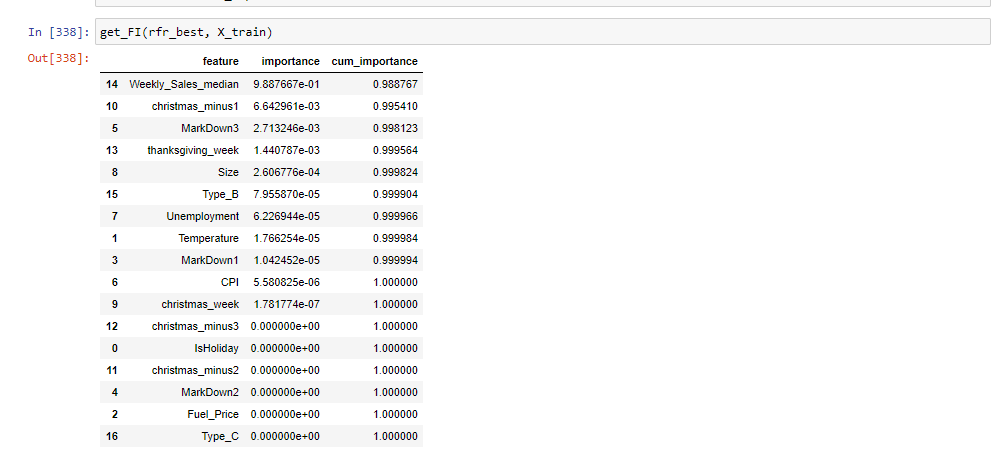
* **Modeling**
* **In This Superstore Project 3 Model Techniques were used. Statistical, RFR, and Arima.**
* **Modeling is the process of creating a visual or a whole information system or parts of it to communicate connections between data points and structures. The goal of modeling is to illustrate the types of data used and stored within the system, and the relationship between the data types. Data models are built around the business needs.**
* **In the Statistical Modeling two prediction types to determine sales were used. The first prediction type determined sales based on store, department, and week number. The second prediction type determined sales based on the department only.**
* **Random Forest is a powerful and versatile supervised machine learning algorithm that grows and combines multiple decision trees to create a “forest.” It can be used for both classification and regression problems.**
* **An**[**ARIMA model**](https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average)**is a class of statistical models for analyzing and forecasting time series data.**
* **It explicitly caters to a suite of standard structures in time series data, and as such provides a simple yet powerful method for making skillful time series forecasts.**
* **ARIMA is an acronym that stands for Auto Regressive Integrated Moving Average. It is a generalization of the simpler Auto Regressive Moving Average and adds the notion of integration.**

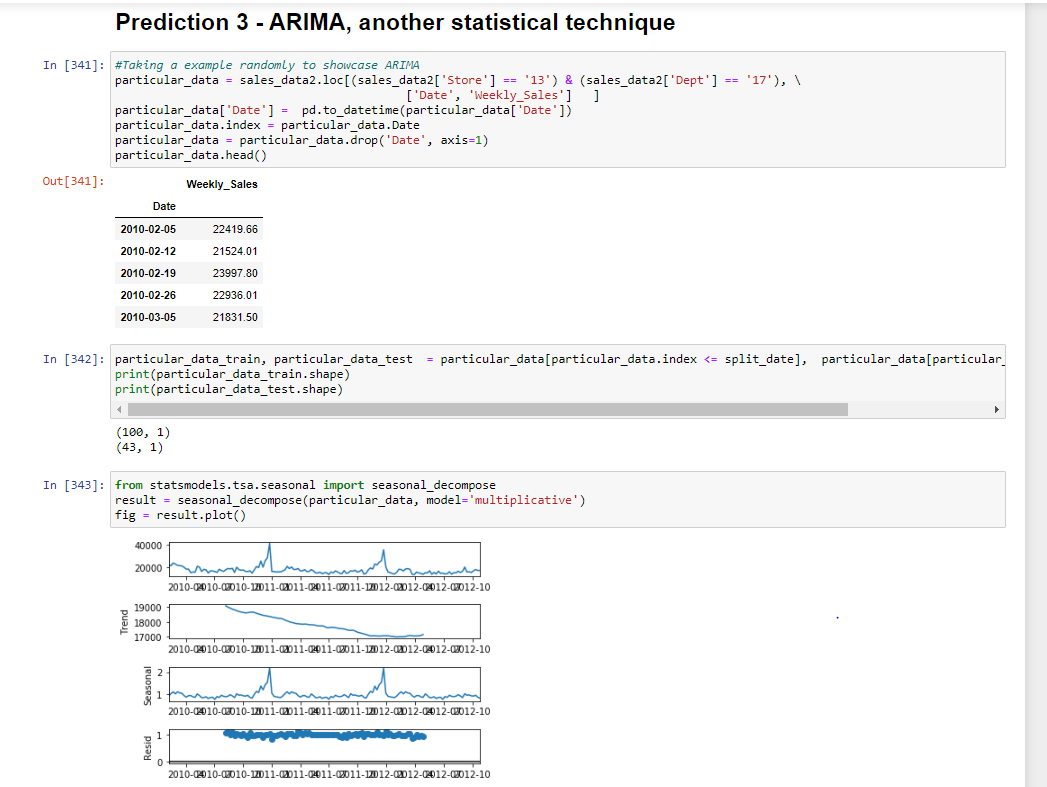


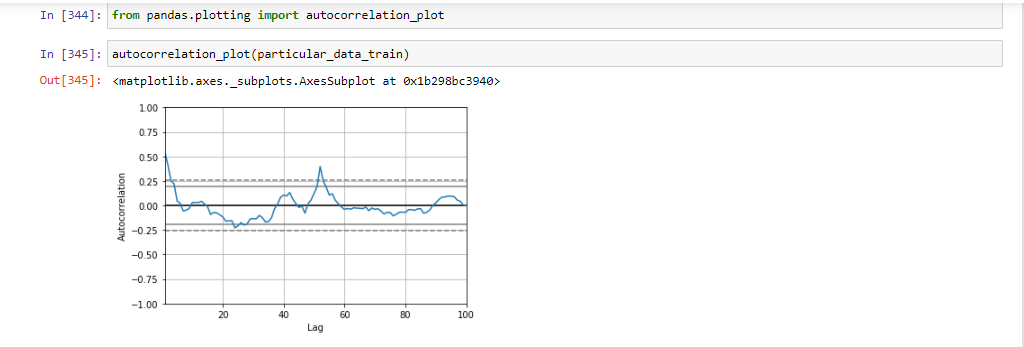


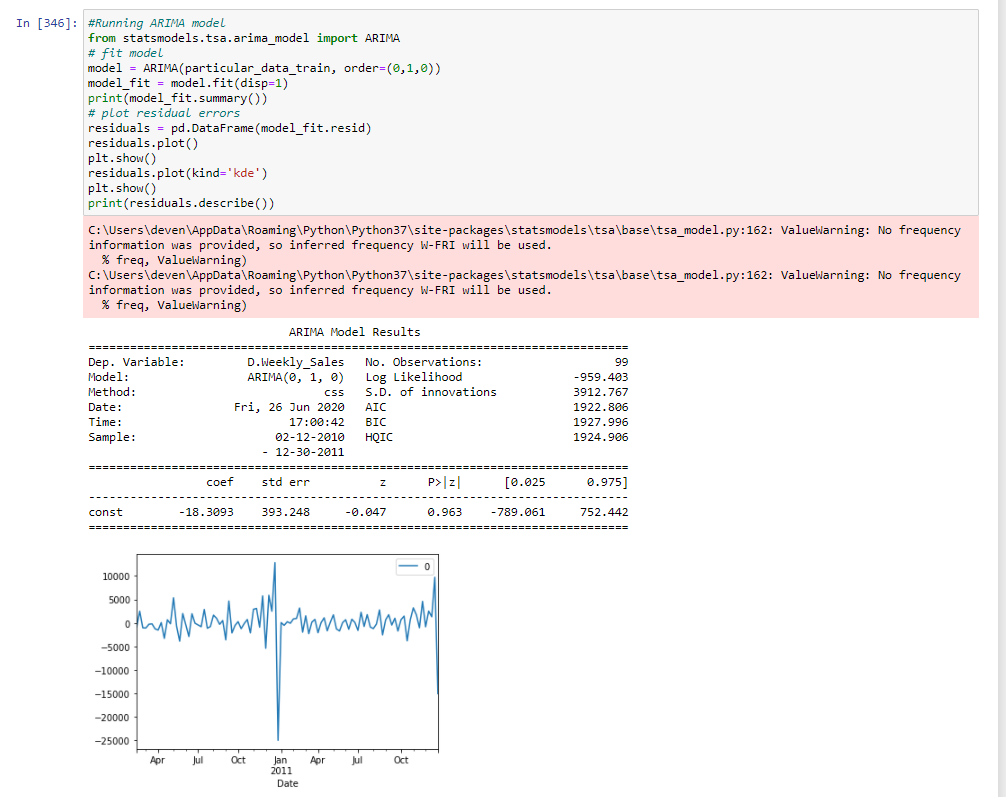


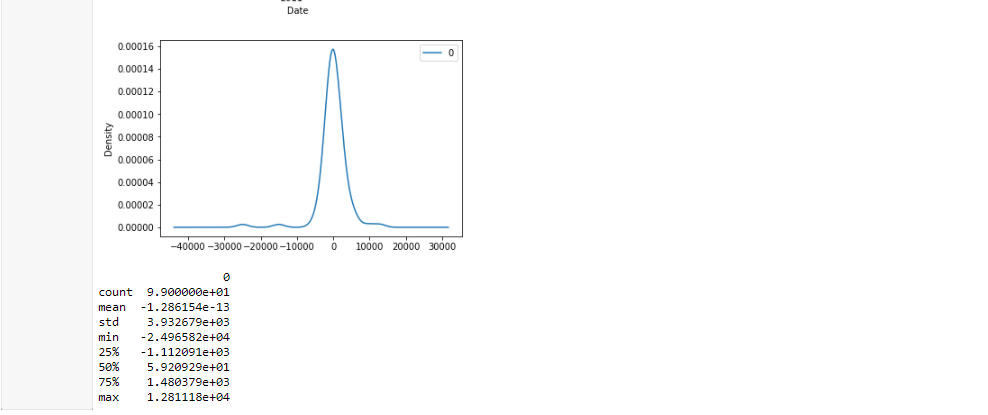


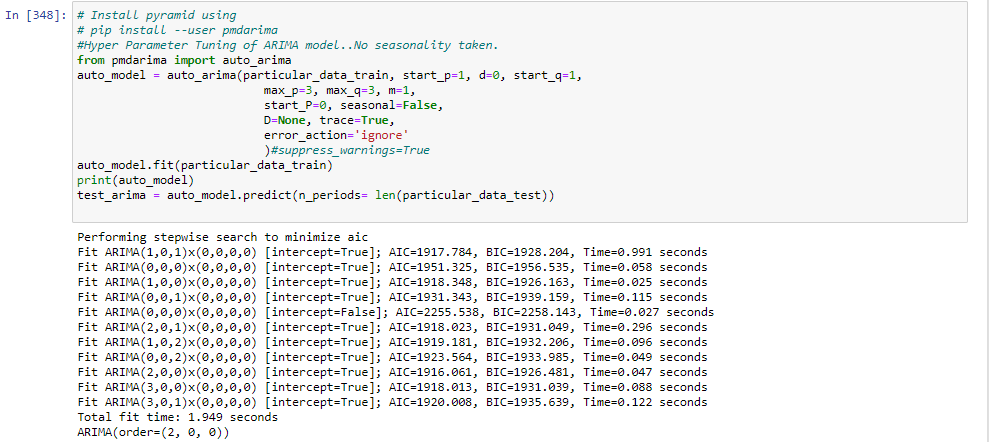


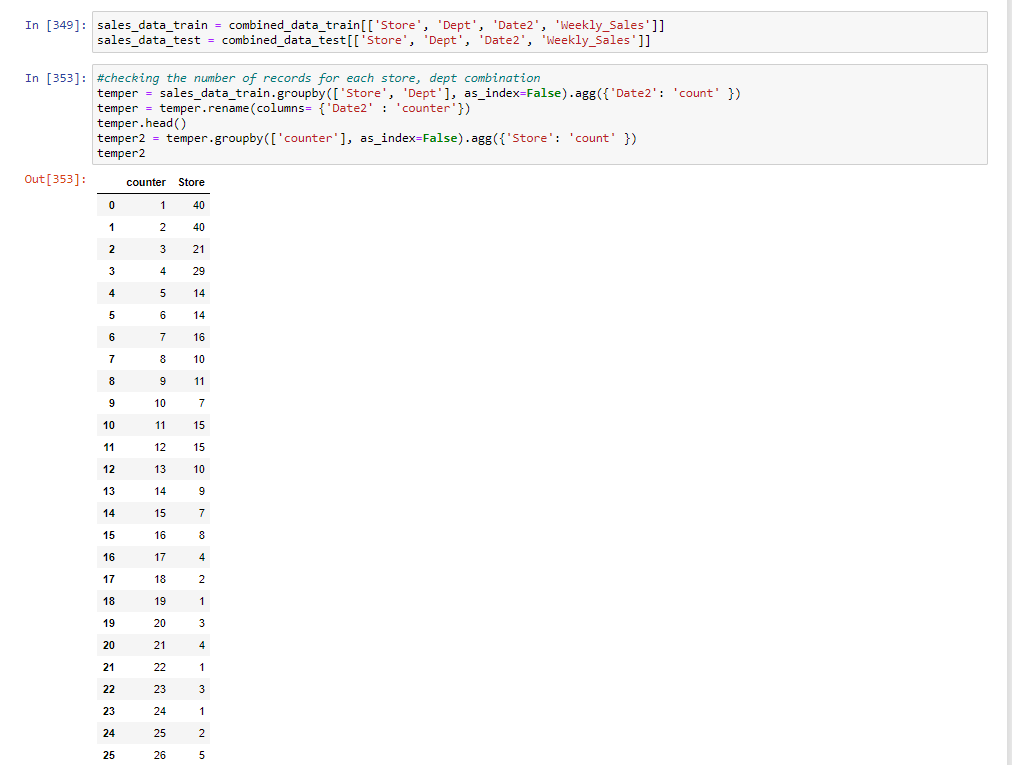


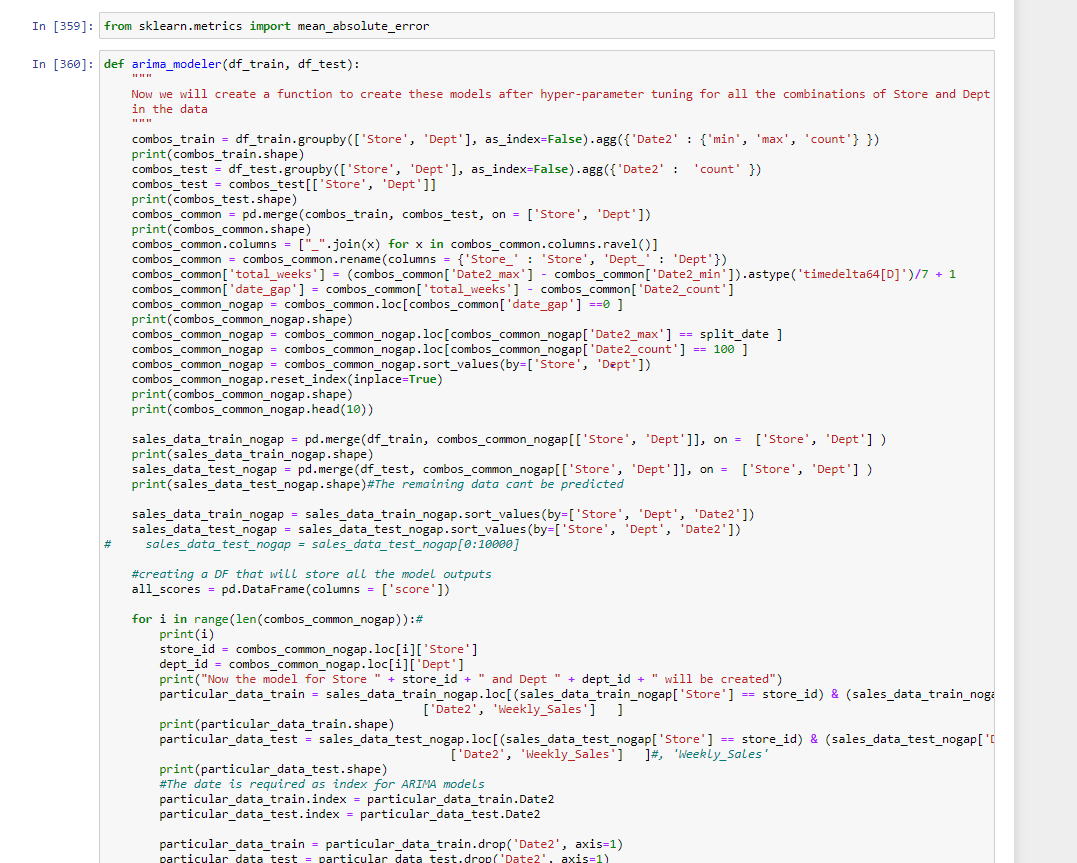


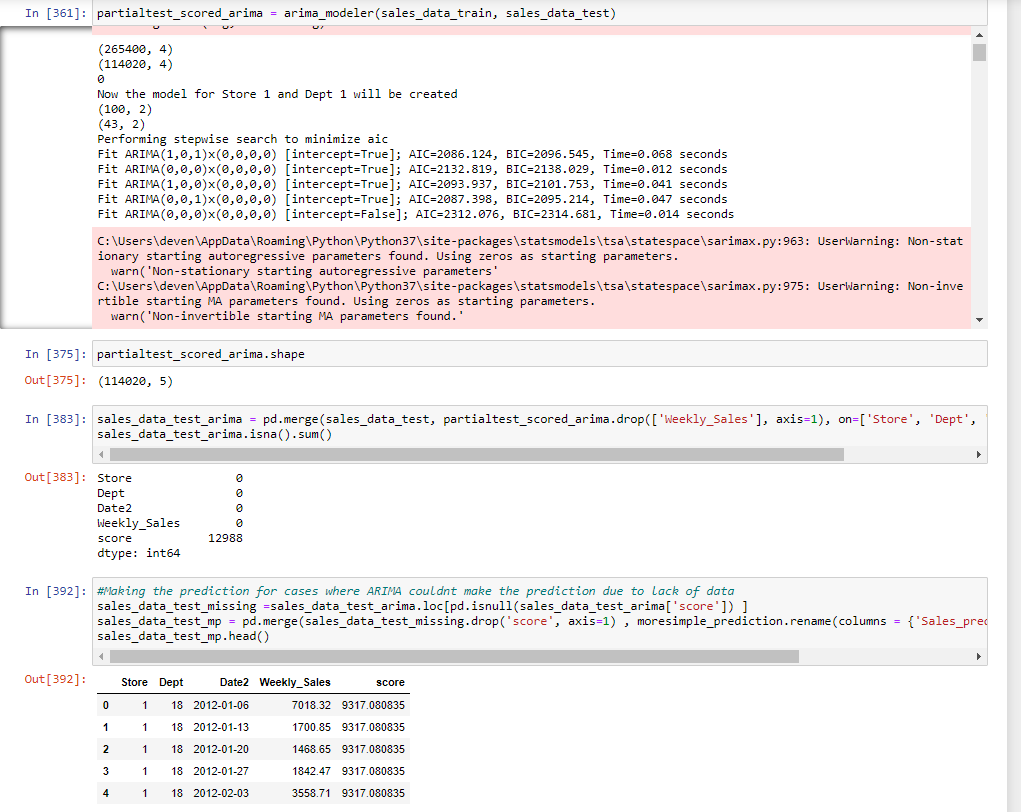






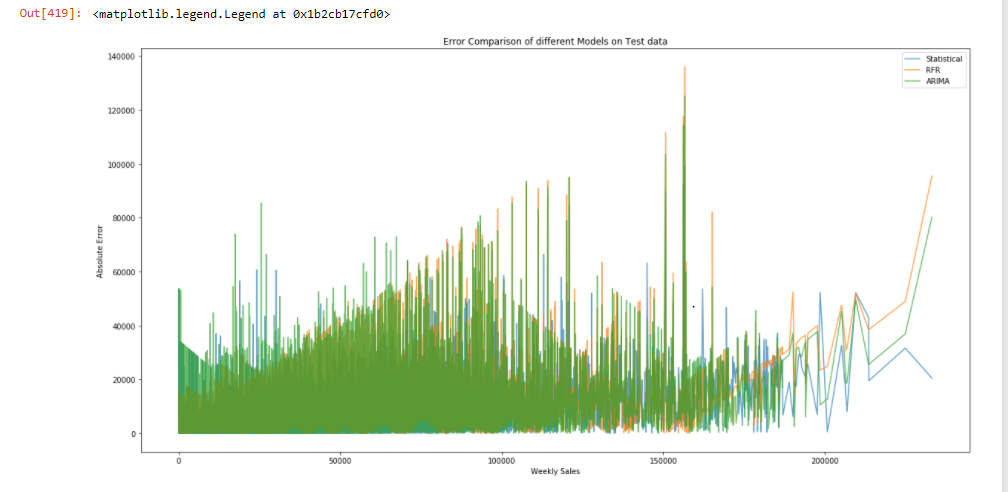












* **Discussion and Conclusion**
* **Macro Economic Indicators did not help in prediction of weekly sales. The sales data is very similar over different years and this consistency helped to predict the sales.**
* **When Model Comparison was done, the best model was found to be the statistical model as it has the lowest rmse (root mean squared error) and mae (mean absolute error) compared to the other models and would most accurately predict the sales with the lowest error.**