RETAIL STORE SALES ANALYSIS

Abstract:

Analysis and processing the retail store data to detect the factors impacting sales, and key factors which disturb the trends in sales and revenue of stores.

Methodology:

Data collection:

I have work on three datasets, one is of retail store data with main features like [‘categories’,’store id’,’Date’,’Units sold’,’Discount’] etc. The next is of Holiday dataset with columns like type of holiday, holiday date, which day of the week. Finally, the data is about the weather in US it consists of many columns with wind, temp, events, Date etc. Looking on a perfect features and size of the datasets I have considered those three from free data platform called “Kaggle”.

Data Preprocessing:

Before working with data for solutions. Initially, I have process and make the data ready by checking the null values, considering only required columns and manage size for fasters processing and efficient results.

After carefully clean and process the three datasets which are retail_data, holiday_data and weather_data.

By using the ‘Date’ column in all three datasets and I integrated the data to form a final dataset to work with, labelled as “retail_holiday_weather” using merge function. I use python programming language and Jupiter notebook for this analytical process. To read and handle the data I use ‘pandas’ library.

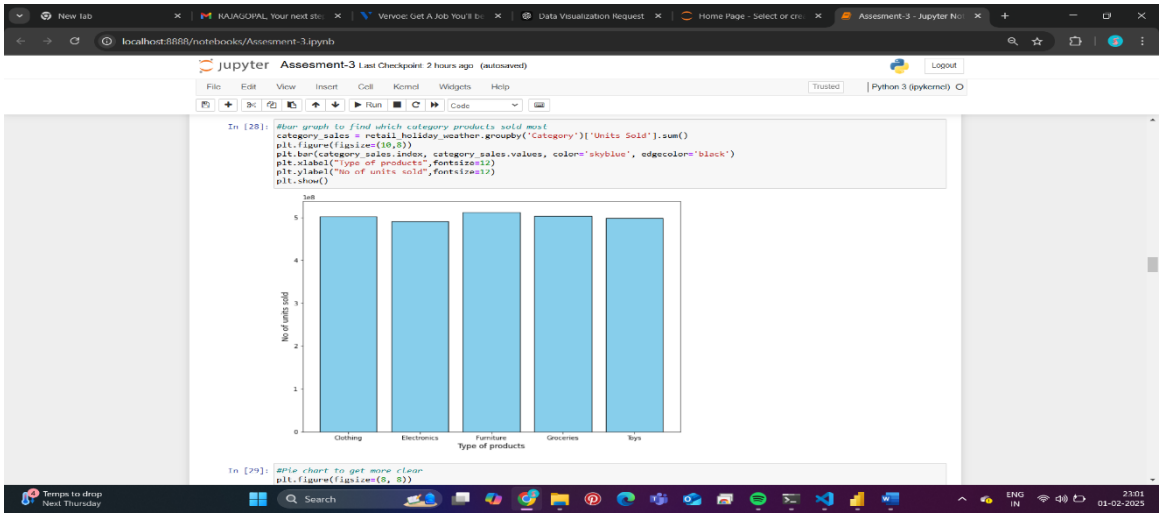
Retail_holiday_weather features = ['Date', 'Store ID', 'Product ID', 'Category', 'Region', 'Inventory Level', 'Units Sold', 'Units Ordered', 'Demand Forecast', 'Price', 'Discount', 'Competitor Pricing', 'Holiday', 'WeekDay','Month', 'Day', 'Max TemperatureC', 'Mean TemperatureC','Min TemperatureC', ' Max Wind SpeedKm/h', ' Mean Wind SpeedKm/h',' CloudCover', ' Events', 'WindDirDegrees', 'Revenue']

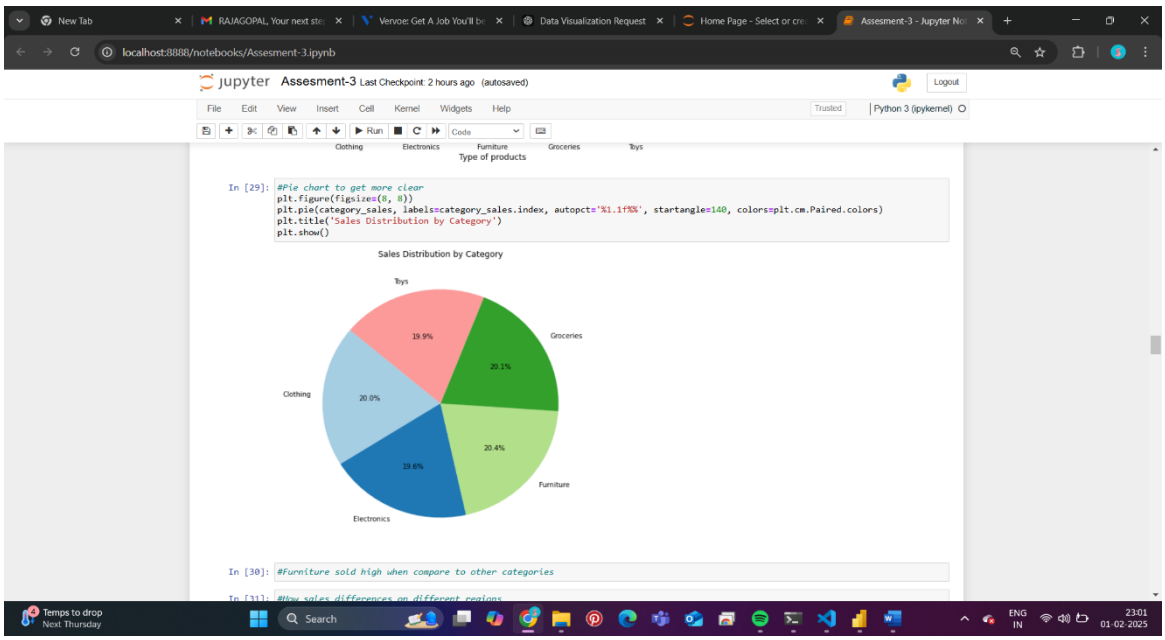
I also perform feature engineering to data by alter and add some columns to data for better analysis, like by adding revenue column by multiplying price of the unit with number of units sold.

Data Analysis:

Firstly, I use matplotlib and seaborn libraries in python for creating visualisations which helps for better understanding of data and helps in finding required impacts.

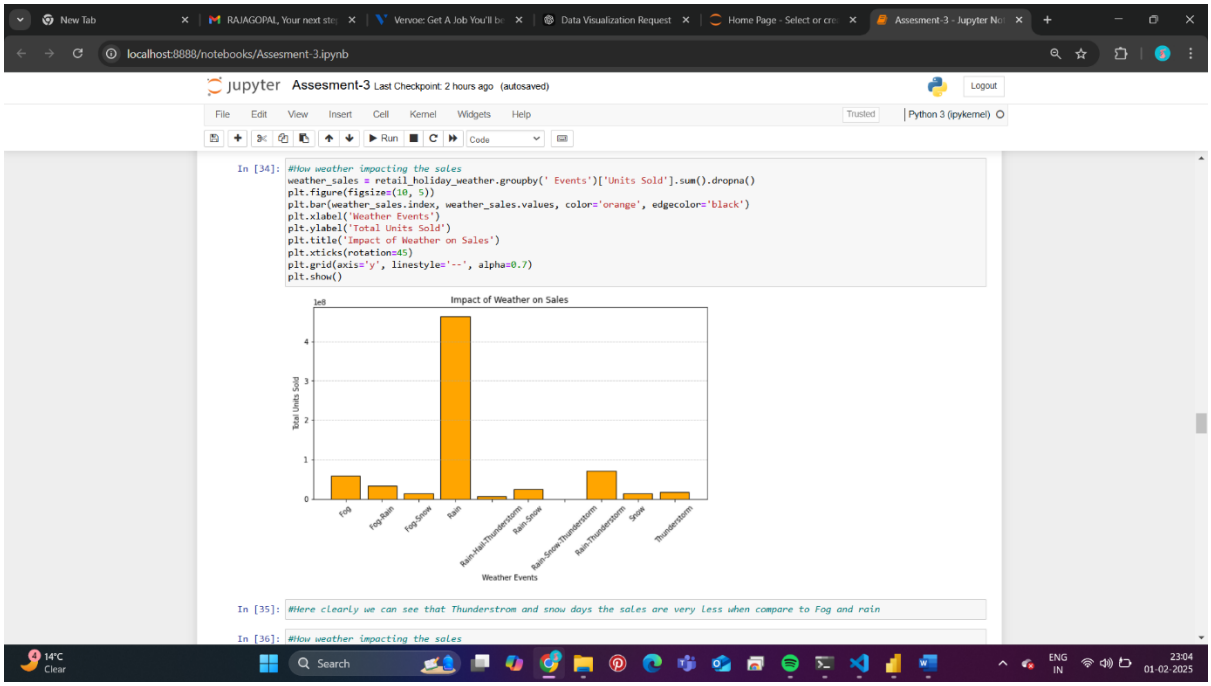
I have analysed number of units sold for each category of products using bar graph and pie chart.





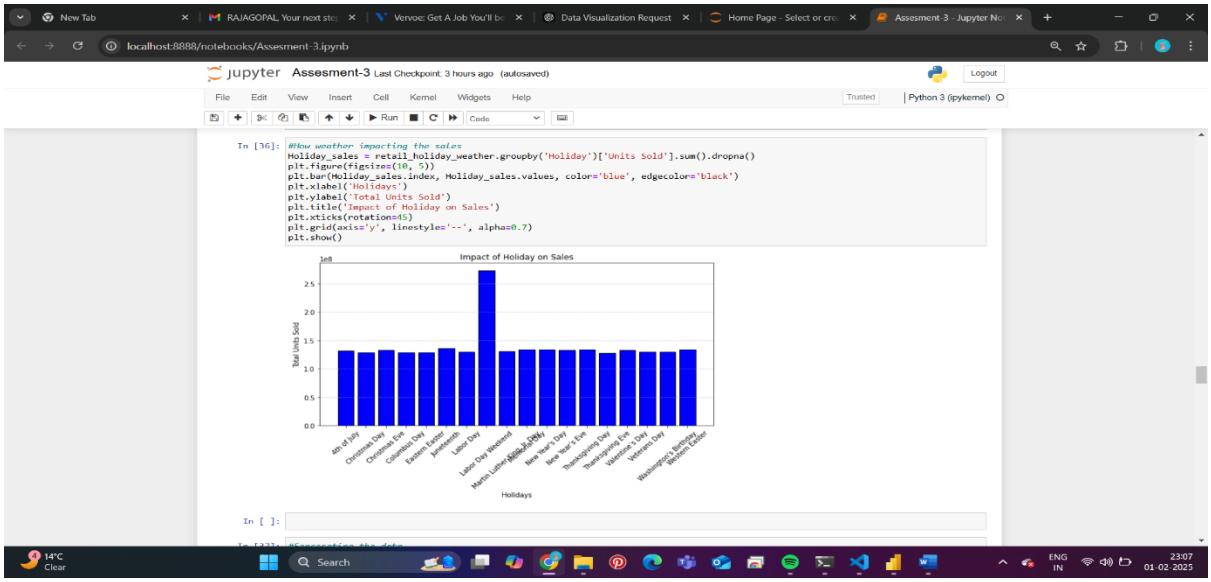
Furniture sales are more and electronic sales are less when compare to other categories in retail market. Represent in bar graph.

How weather impact on sales, for this I made comparison between weather and units sold.



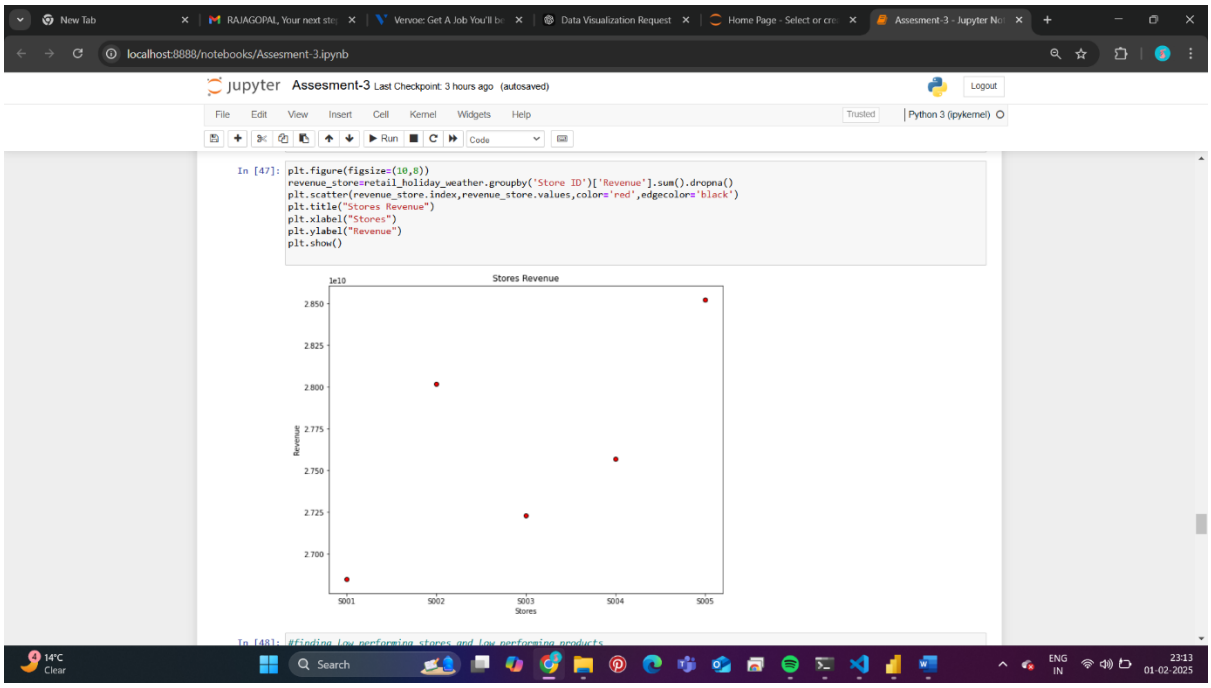
When weather is snow or thunderstorm the sales are bad, apart from sunny days the sales are good on rainy day than other conditions.

How Holiday impacting the retail sales, for this I compare the type of holiday and units sold, represented in bar graph.



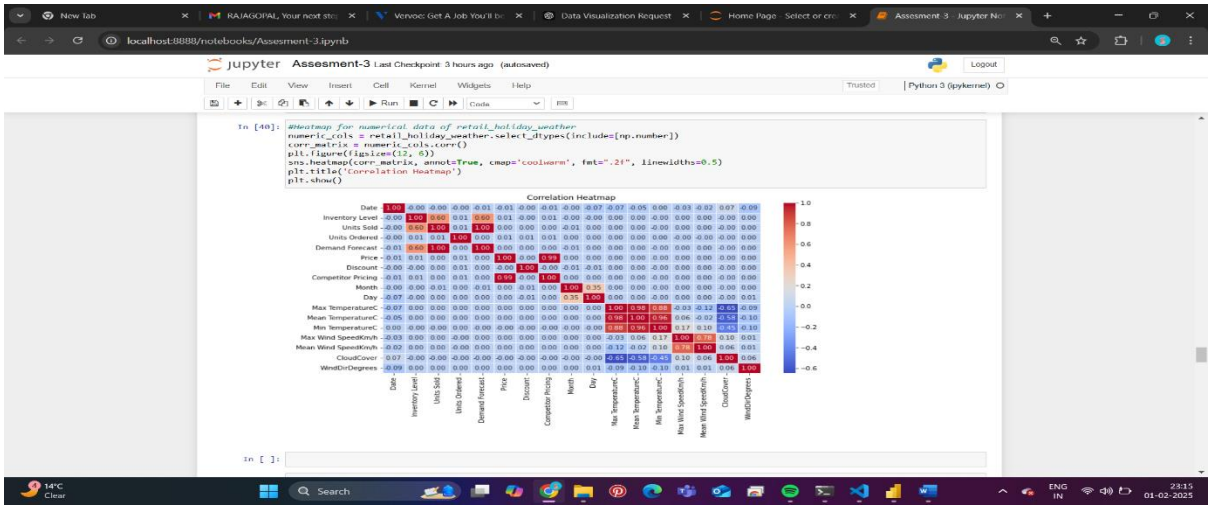
In x-axis there are type of holidays like 4th of July, labour day etc. The Number of units sold on the holidays are less than 15, only the labour day was excluded and maintain a good-sales among all the holidays.

To find the revenue of stores, I have analysis that which store has more revenue and which store has less revenue.

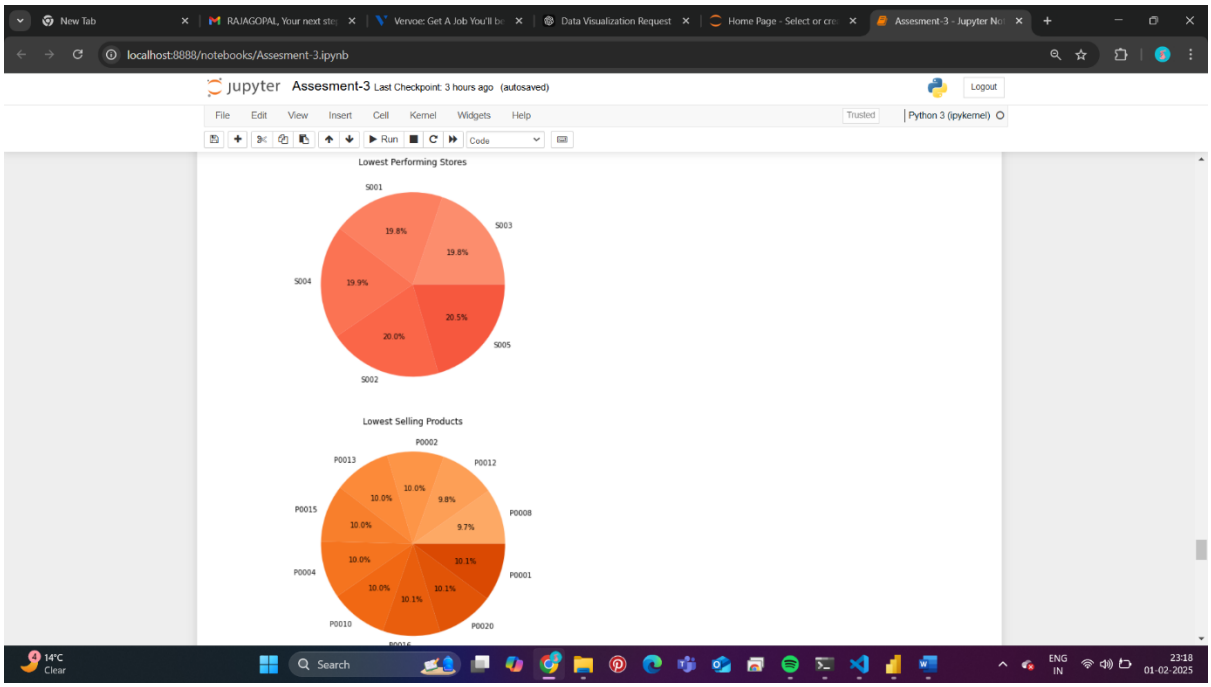


I used scatter plot for the comparison, I have found that store-5 has greater revenue and store-1 has least revenue.

Using all numerical data in the retail_holiday_weather data, I have made a heatmap for brief explaining of the data.



Lastly, using pie chart I made two visualizations to find the least performance columns of the data like lowest selling products and lowest selling stores.

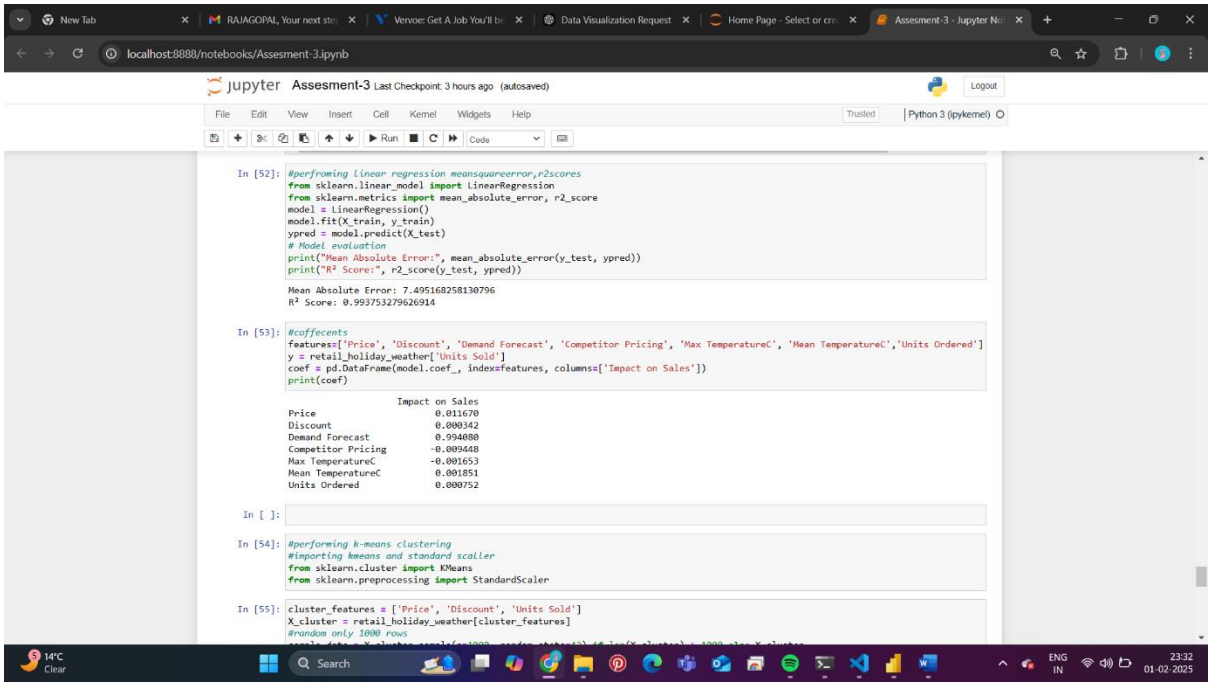


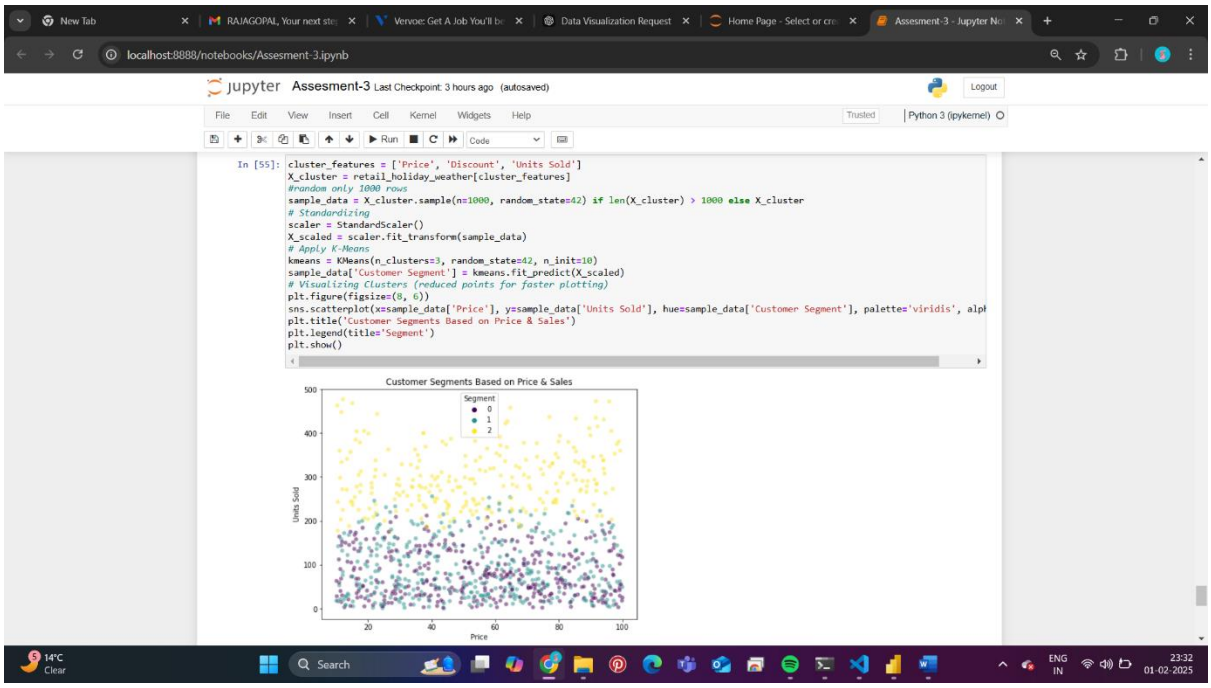
Advanced Analysis:

To make model prediction and advanced visualizations of the data, I have used some ML techniques like sklearn linear models for regression and clustering for KMeans to evaluate the accuracy of the model and finding the advanced impacts of the data with other columns.

Firstly, I have train_test_split the data, to fit the LinearRegression. Using the model I have calculated the r2_score and mean square error values of the data, later on using coefficients I printed the some features impact on sales.

Secondly, I use clustering algorithm KMeans clustering, to cluster the data and make visualizations with random 1000 rows of the data.





Hypothesis Testing:

I have tested by taking the date column, considering a year of 2010 and testing with scale of before campaign year and after campaign year. On this process I have got a results/output as:

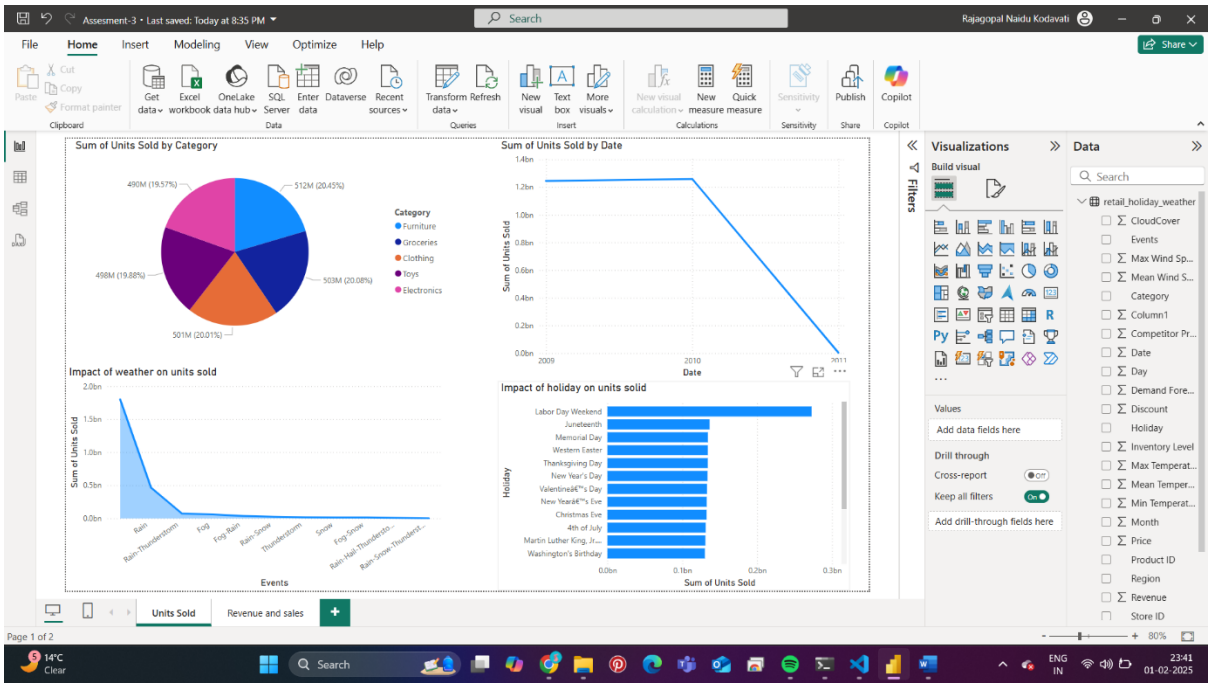
T-Statistic: 21.1706, P-Value: 0.0000

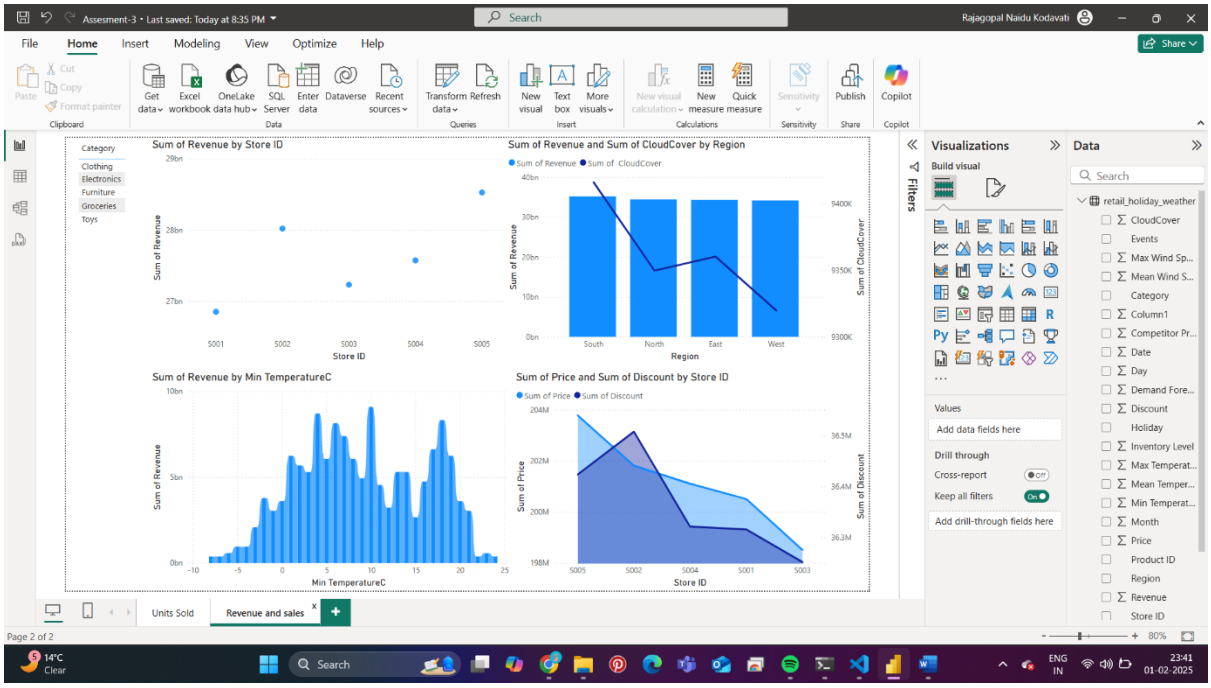
The marketing campaign had a significant impact on sales!

Visualization Dashboard:

Apart from python, I have used a power full tool called PowerBI to make high efficient visualizations and created two dashboards for better and clear understanding for clients/audience:

For this I have exported our final data from python to csv using .to_csv format, after than I loaded the data and using columns I have created 1st dashboard with impact of sales on other factors. The second dashboard is about the revenue and different conditions impacting the retail stores.





Conclusion:

After successfully implemented all required algorithms and analysis on the retail data, I have got to note that most of the retail sales are impacted by poor weather conditions majorly like snow, thunderstorm, floods etc, also the holiday dates making the unit selling down. In our data from 2009-2015 margin all type of categories product sales almost similar with little differences in percentages. The other factors like temperature, wind, cloud cover has no much impact on sales, and there is unique revenue and unit sales for every store, mostly they are dependent on the inventory and region of the stores. As I make good visualizations, I better understand that competitor pricing also shows some difference in sales of the stores. My brief advice is to make less inventory on bad weather conditions and holiday dates, apart from that competitor prices and discounts prices is continuous considering factor in retail market. Always analysing the data by comparing the trends with past years make the understanding of market gets easy and it get to know the demand of the products to increase the inventory and impact areas to make necessary changes. As we see that with final visualization on dashboard the stores sales are raised and decreased corresponding to the discount made by the stores. Also when we see weather conditions that rainy days and in holiday dates labour day is not considered as low sale days.

Rajagopal Naidu Kodavati

Rajagopal.naidu@aogjob.com

+1 940-977-2260

Denton, Texas

76201