

TITLE & PROJECT TAGLINE:

AI for Business reporting and analyses: Turning raw sales data into actionable insights and automated reports

OVERVIEW / EXECUTIVE SUMMARY:

This report aims to leverage AI and machine learning techniques to uncover patterns in customer behavior, identify key drivers of sales decline, and recommend actionable strategies for recovery.

Backed up with this documentation to help stakeholders and users understand the business problem and solution invented for this problem.

DATA COLLECTION PHASE:

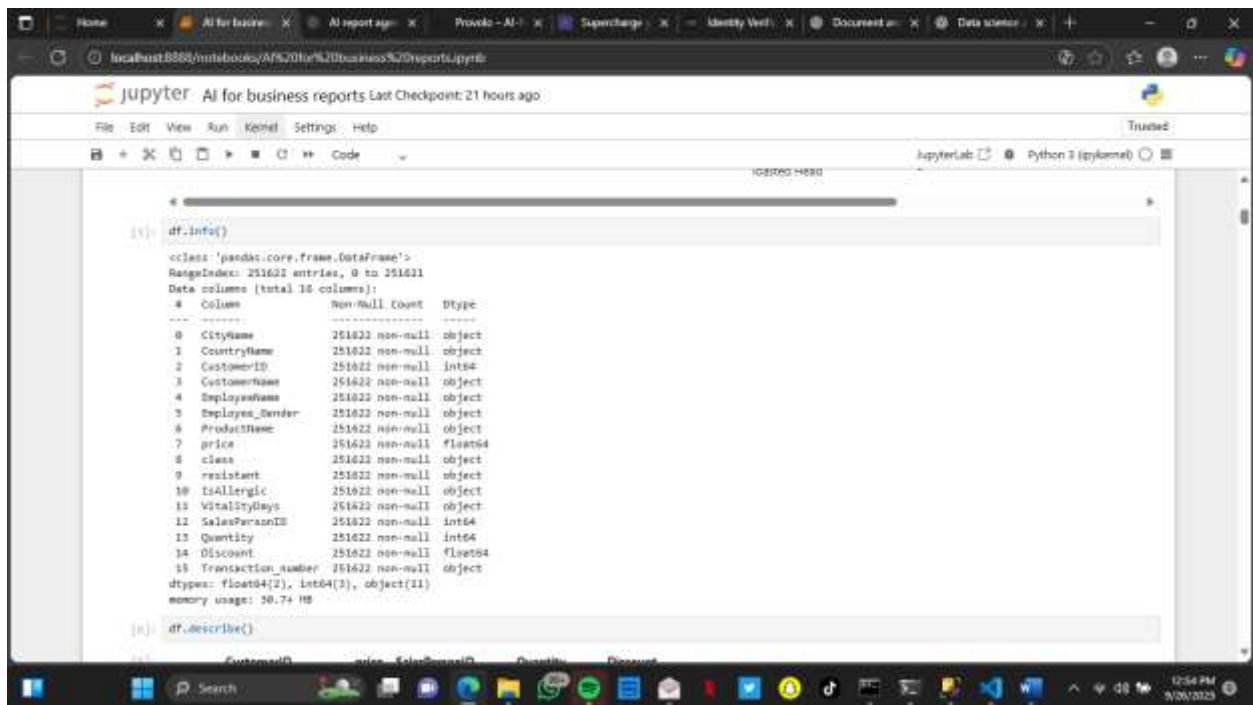
A robust ETL pipeline was implemented using SQL for data extraction and pandas for cleaning and transformation. SQLAlchemy facilitated seamless database interaction, ensuring data integrity throughout the process.

EXPLORATORY DATA ANALYSIS PHASE:

Descriptive statistics and distribution plots were employed to explore customer engagement, product performance, and seasonal trends. Visualizations, highlight key anomalies and

AI For Business Reporting

behavioral shifts.



```
[1]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 251622 entries, 0 to 251621
Data columns (total 15 columns):
 #   Column                Non-Null Count  Dtype  
---  --   ---
 0   CityName              251622 non-null object  
 1   CountryName           251622 non-null object  
 2   CustomerID            251622 non-null int64   
 3   CustomerName          251622 non-null object  
 4   EmployeeName          251622 non-null object  
 5   EmployeeGender        251622 non-null object  
 6   ProductName           251622 non-null object  
 7   price                 251622 non-null float64  
 8   class                 251622 non-null object  
 9   assistant             251622 non-null object  
10  isAllergic            251622 non-null object  
11  VitalityDays          251622 non-null object  
12  SalesPersonID         251622 non-null int64   
13  Quantity              251622 non-null int64   
14  Discount              251622 non-null float64  
15  TransactionNumber      251622 non-null object  
dtypes: float64(2), int64(3), object(11)
memory usage: 10.74 MB

[2]: df.describe()
```

After going through the descriptive phase of the data, Here are the findings discovered

~~~~~

### -EXPLORATORY REPORT

-The dataset comprises 251,622 rows and 8 key features, including customer ID, product category, purchase frequency, and transaction dates.

-Data types are 11 objects, 3 integers and 2 float numbers

### -COLUMN DESCRIPTIONS

-CustomerID:

we have 251622 rows of non-missing data

IDs are sequential identifiers and do not have business meaning

Customer with the unique identifier 7 is the least customer that bought from us in this sales data. Jay Grocery Store has recently faced a noticeable decline in sales. This trend prompts a deeper investigation into customer behavior particularly the absence of previously frequent buyers such as Customer IDs 1 to 6.

## AI For Business Reporting

And the customer with identifier 98742 is the highest identifier in our sales data and this raises question for the remaining missing customers that haven't bought from us

-Price:

we have 251622 rows of non-missing data

On average, product costs around 50.84

Standard deviation is giving 28.54, which means most prices of products are scattered around 22.3 to 79.38. Information can be used as promotional strategy

Minimum price is 0.045, given it non significance, this can be an error or a product that cant be bought per unit and requires large bulk purchase.

25% of the products bought fall below 26.58

median price is 52.64 indicating most product bought are around this price

75% of the product fall below 75.25

And the highest unit price of this sales data is 99.88

-SalesPersonID:

we have 251622 rows of non-missing data

Although we have a total of 23 employees, mean of this dataset is 12 and the standard deviation is 6 with means most of our employees that are present in this sales data fall between 6 and 18 and also given the minimum identifier integer for employee is 1 and the maximum is 23, it looks like most of our employees participated in the sales

-Quantity

we have 251622 rows of non-missing data

The mean is 13 and standard deviation is 7 which means most quantity bought is around 6 to 20, but we dont know which of the product is bought the most till we do further analysis

The minimum quantity bought is 1 unit and maximum is 25

## AI For Business Reporting

-Discount

we have 251622 rows of non-missing data

75% of data has 0 discount, the highest amount of discount is 20 and most sales has no discount.

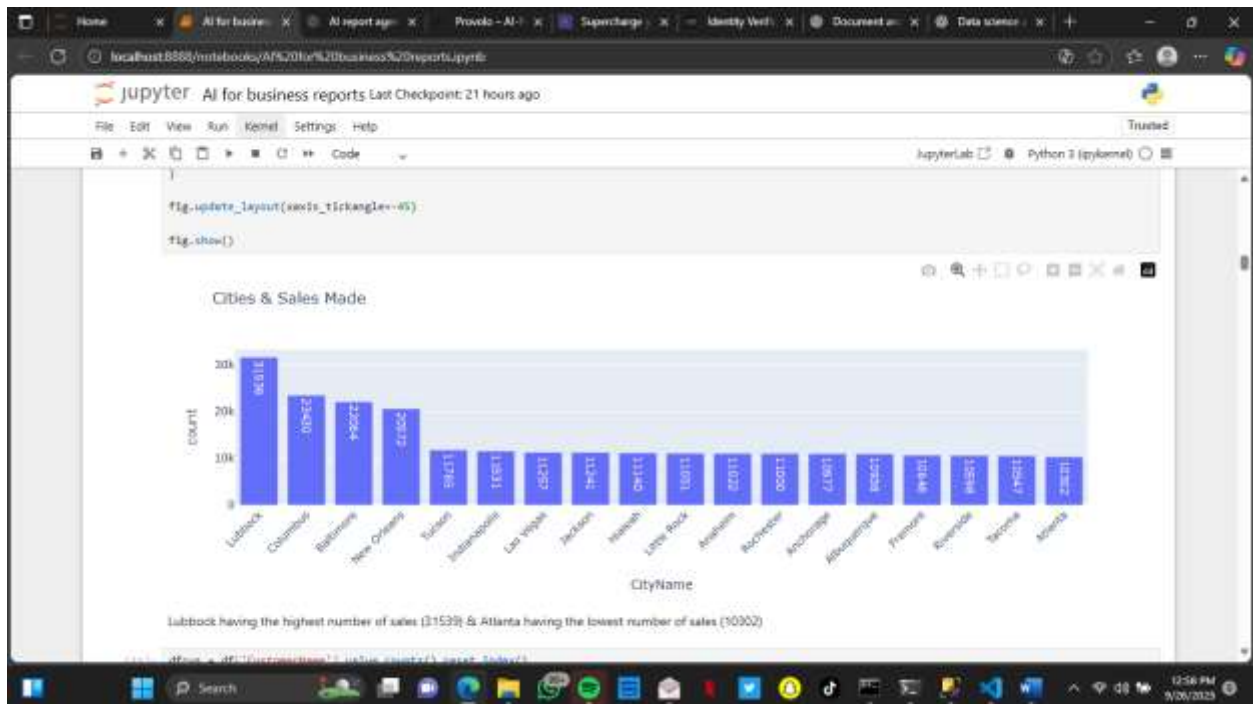
-IMPORTANT NOTE

There is no sign of duplicated or null values in the data

~~~~~

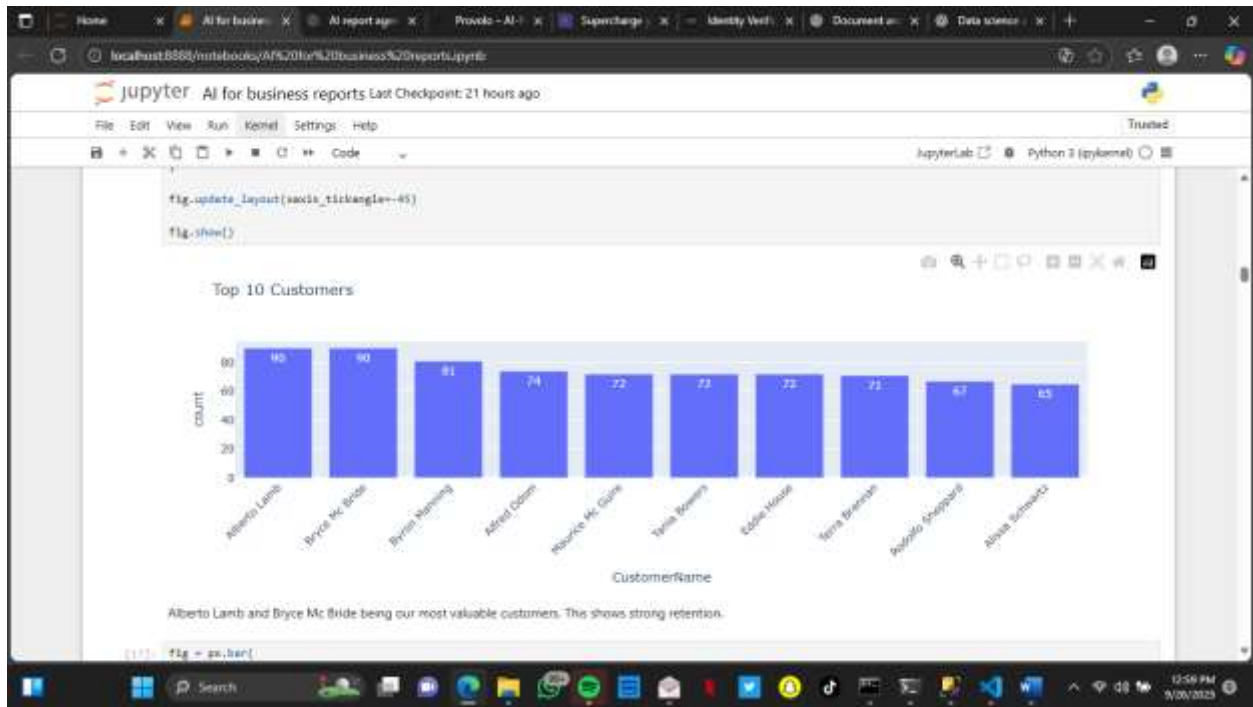
And the above gave us a bit of insight in the data.

Moved on to doing some plots if we could get hidden patterns in this data

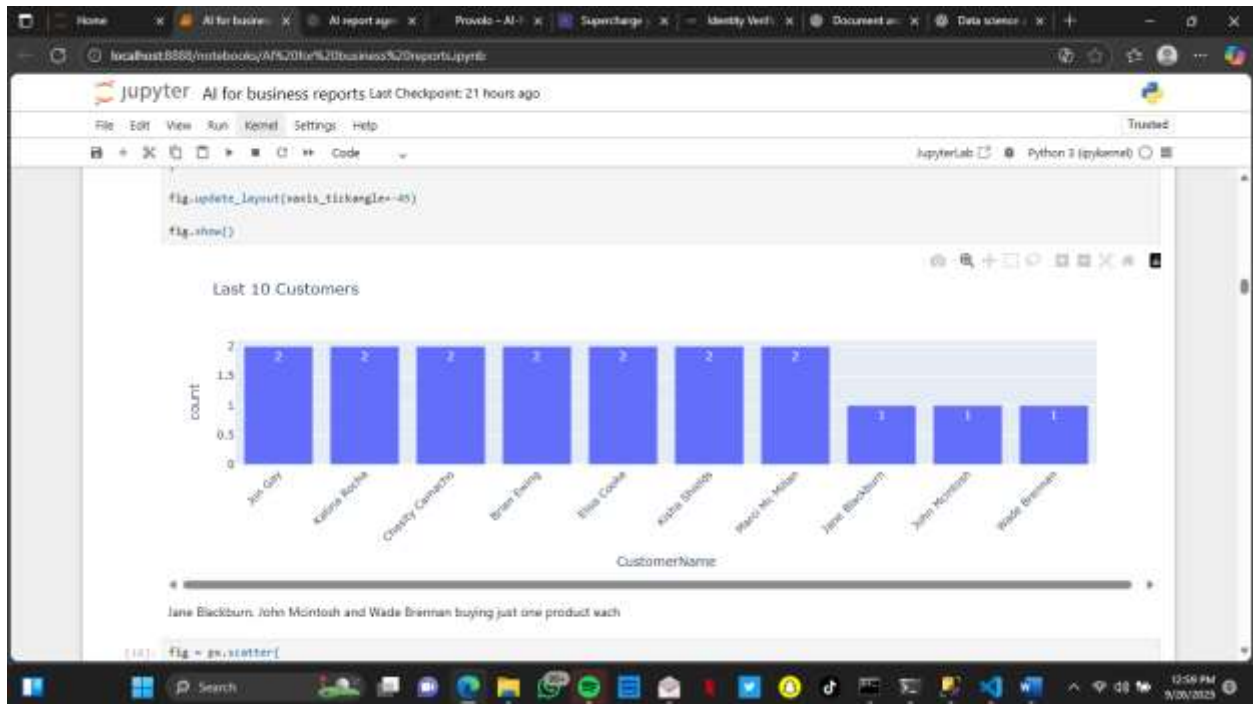


The above plot shows cities and how frequently they made purchase

AI For Business Reporting

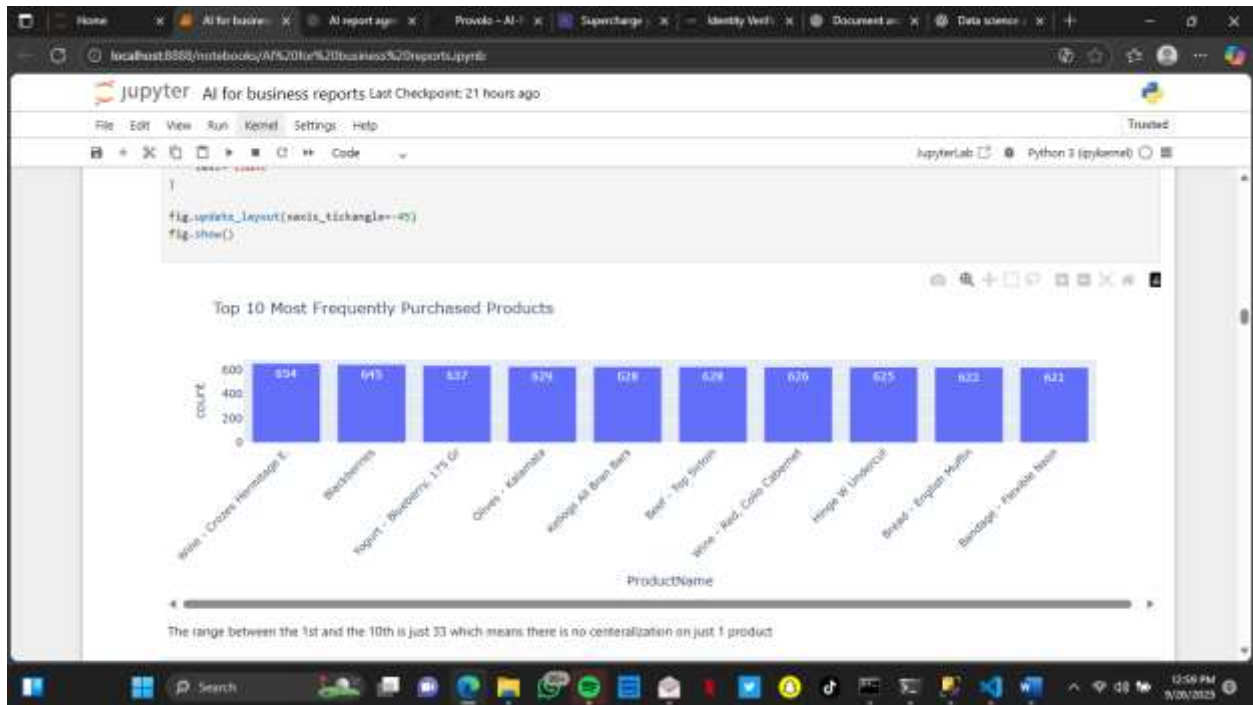


The above plot shows our top Ten Customers by revenue

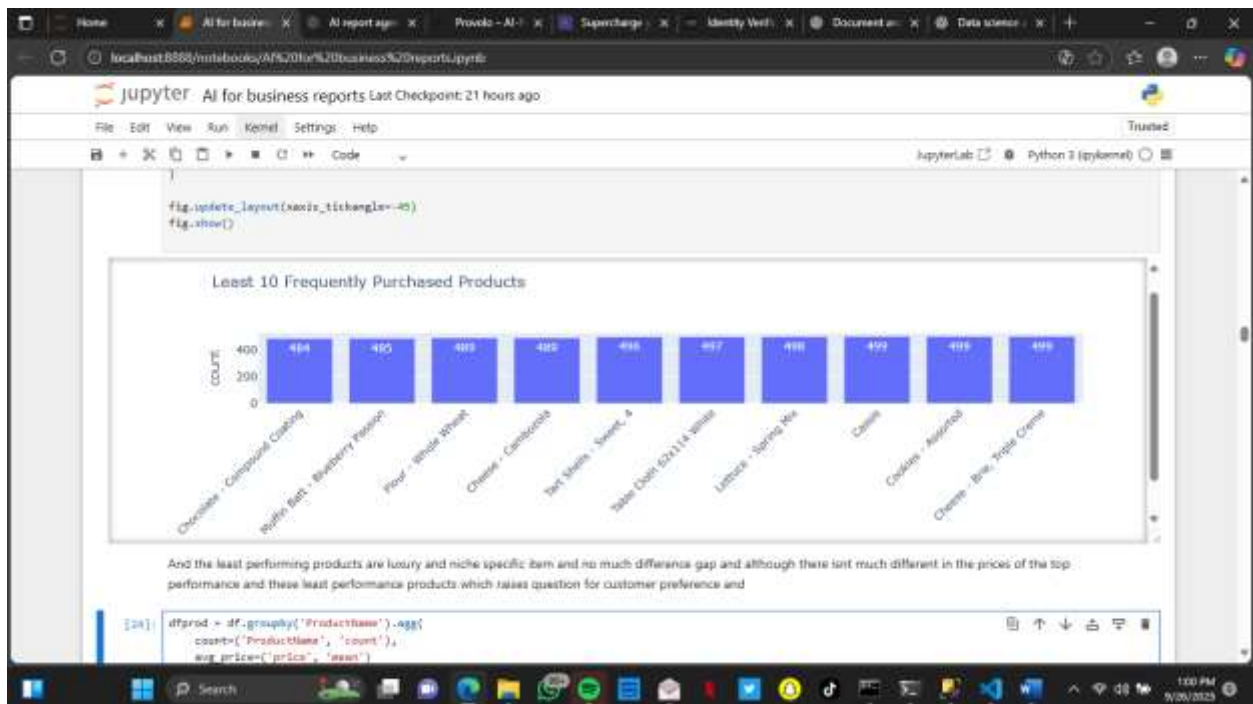


The above shows least ten customers by revenue

AI For Business Reporting

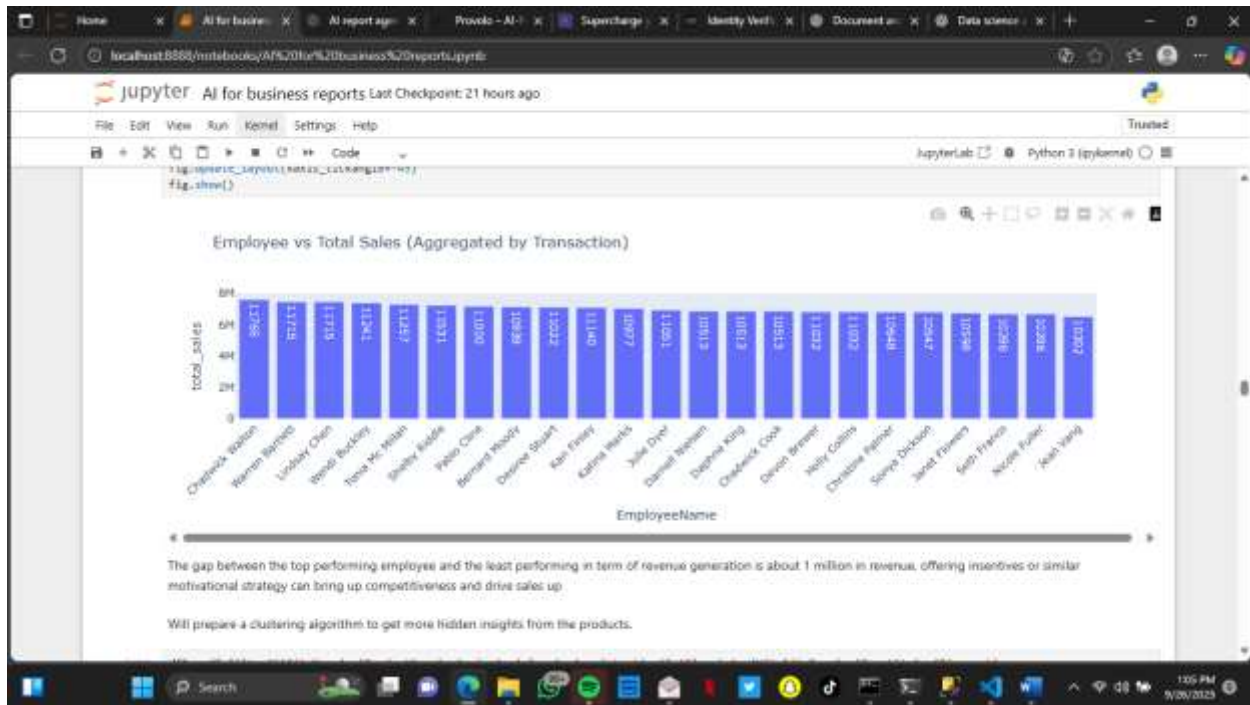


The above shows top 10 of the most frequently purchased Goods



The above shows goods with minimal engagement

AI For Business Reporting



The above shows employees and their engagements in sales

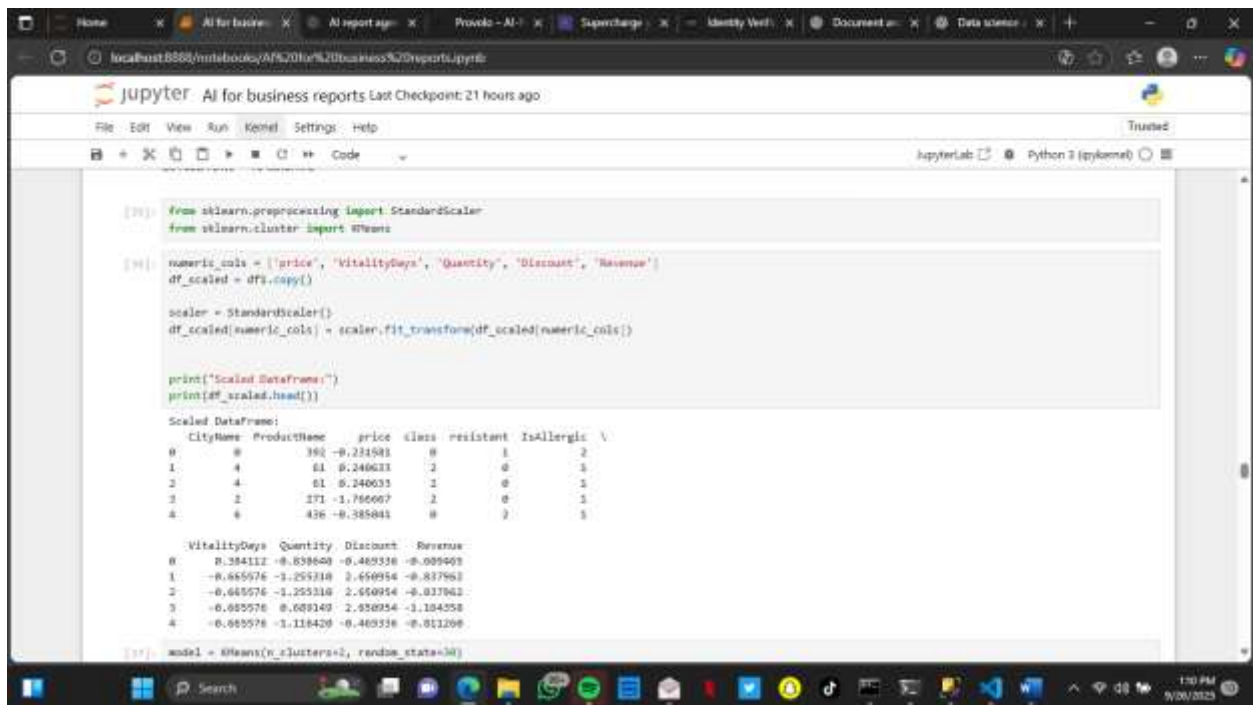
After all analyses have been done the data that gave more insights were the products purchasing pattern and city engagement in purchase, although that didn't give us sufficient insight to make a decision. Then that led us to building a Machine Learning Model (KMeans)

MODEL BUILDING PHASE:

KMeans clustering, enhanced by Principal Component Analysis (PCA), was applied to segment customers into distinct behavioral groups. Silhouette scores and the Elbow method validated

AI For Business Reporting

the optimal number of clusters.



```
[17]: from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans

[18]: numeric_cols = ['price', 'VitalityDays', 'Quantity', 'Discount', 'Revenue']
df_scaled = df.copy()

scaler = StandardScaler()
df_scaled[numeric_cols] = scaler.fit_transform(df_scaled[numeric_cols])

print("Scaled DataFrame:")
print(df_scaled.head())

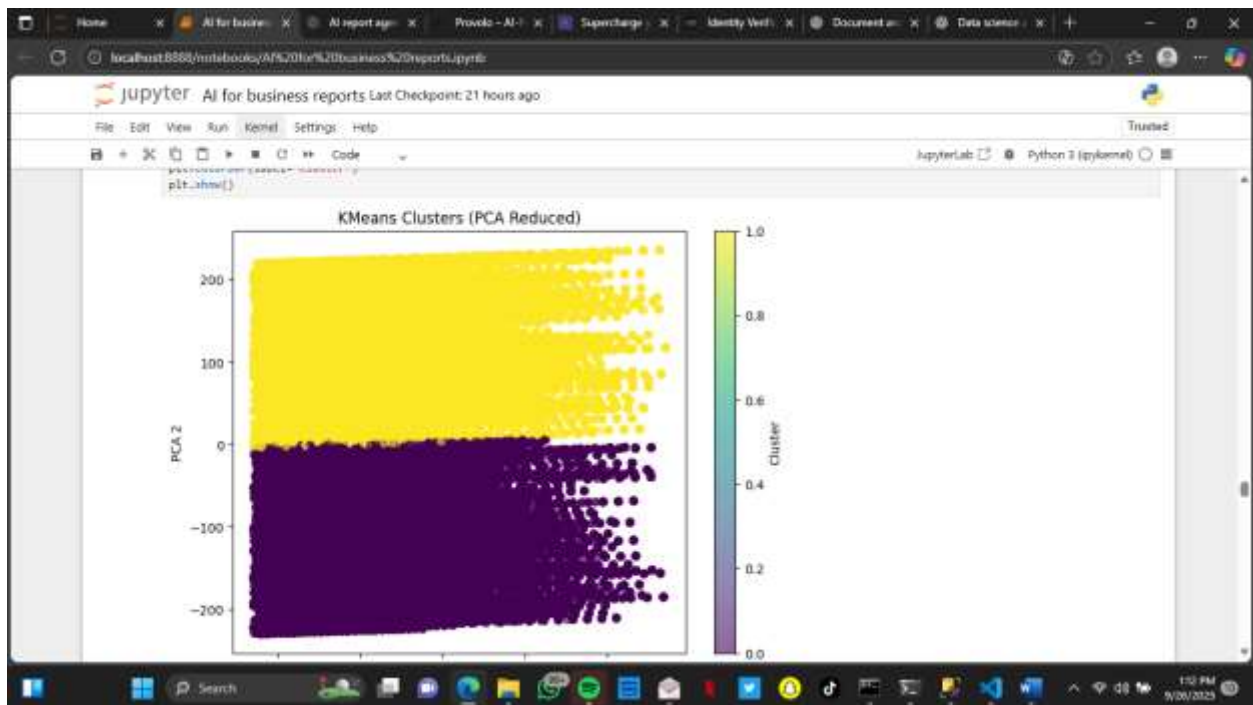
Scaled DataFrame:
  CityName  ProductName  price  class  resistant  IsAllergic  V
0      0             0    392 -0.231583      0         1         2
1      4             4     61  0.240633      2         0         1
2      4             4     61  0.240633      2         0         1
3      2             2    271 -1.766067      2         0         1
4      6             6    436 -0.385041      0         2         1

  VitalityDays  Quantity  Discount  Revenue
0  0.284112 -0.838640 -0.469338 -0.009409
1 -0.665576 -1.252310  2.650954 -0.837863
2 -0.665576 -1.252310  2.650954 -0.837863
3 -0.665576  0.609140  2.850954 -1.104558
4 -0.665576 -1.118420 -0.469338 -0.811208

[19]: model = KMeans(n_clusters=2, random_state=30)
```

Started and tested this model and data with 2 cluster to check performance, and did a pca plot.

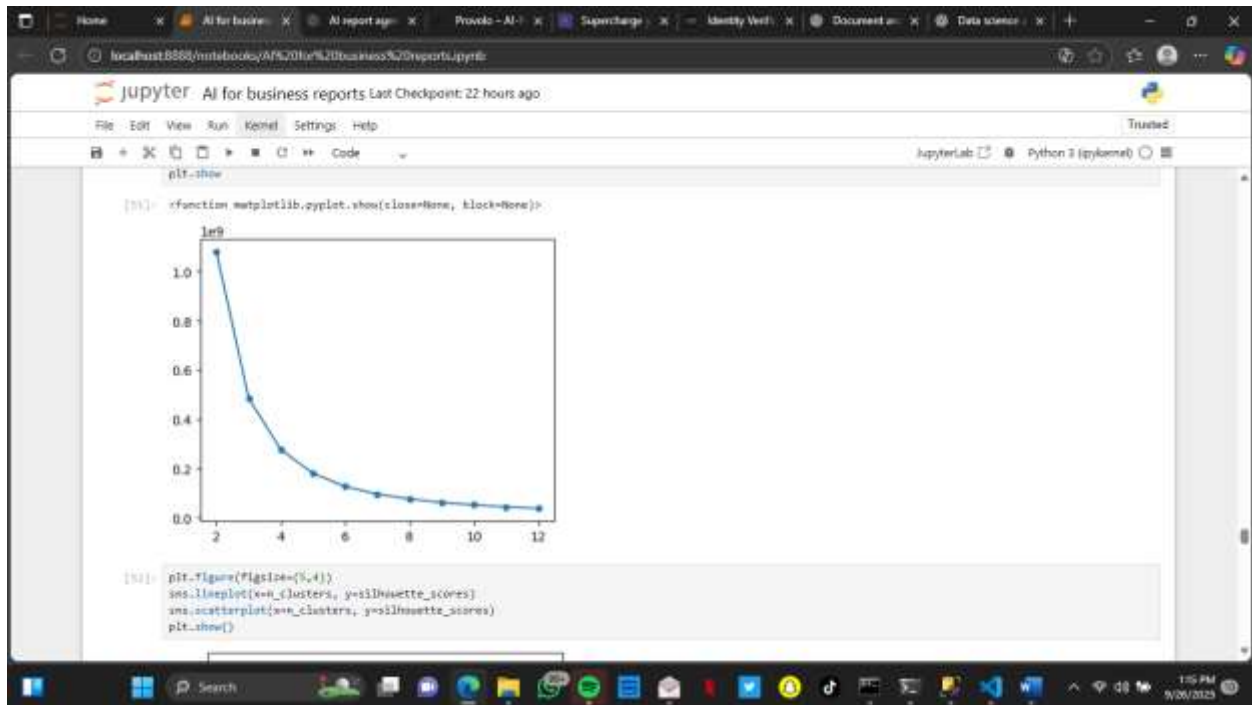
""image below""



Got a 62% silhouette score, and a further metric evaluation was done using elbow plot

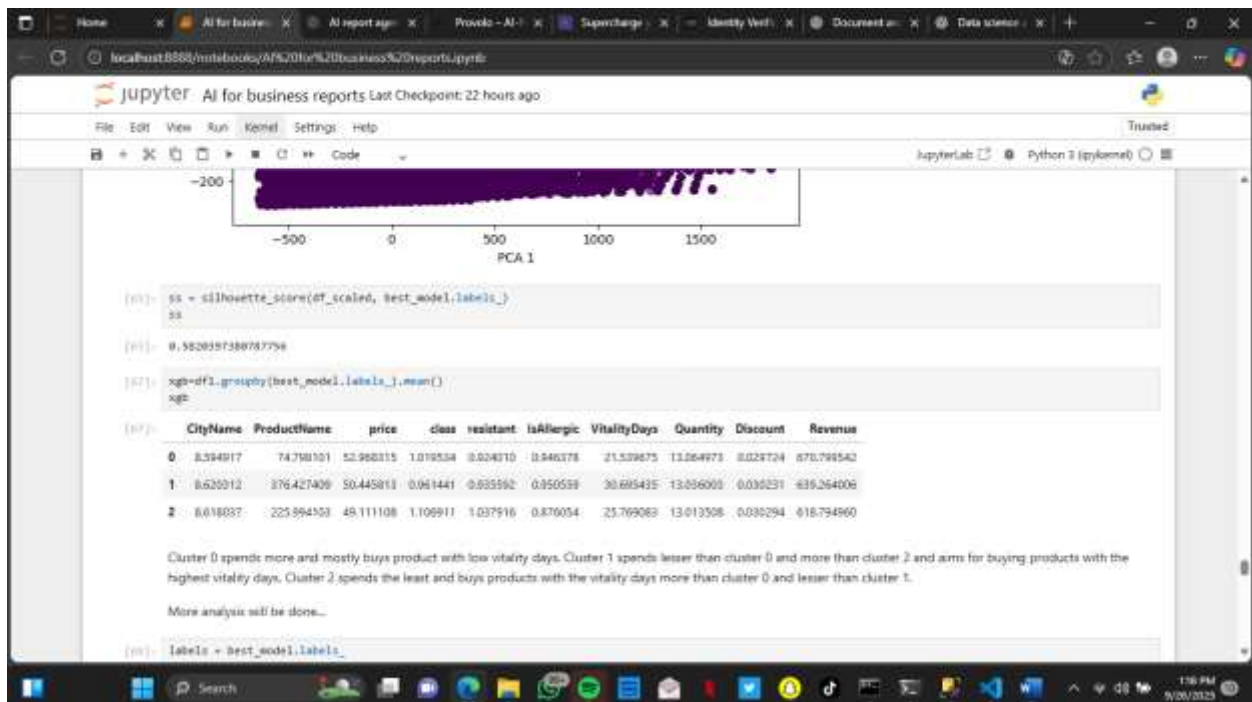
""image below""

AI For Business Reporting



As the name implies, elbow plot is a way to check where the best cluster falls and decided to move on to 3 clusters. Although silhouette score dropped to 58% but it was still a bit close to the score of the 2 cluster and then decision was based on the 3 cluster

Went ahead in doing an analysis on what makes the 3 clusters different using the groupby function



AI For Business Reporting

Result of the initial analyses:

~~~~~

Cluster 0 spends more and mostly buys product with low vitality days.

Cluster 1 spends lesser than cluster 0 and more than cluster 2 and aims for buying products with the highest vitality days.

Cluster 2 spends the least and buys products with the vitality days more than cluster 0 and lesser than cluster 1.

~~~~~

More analyses were then done and these key insights were discovered:

FOR CLUSTER 0:

The mean price spent is around \$52.97 and they are the highest spenders among the 3 clusters; most preferred products are product with 22 days vitality on average and this is the least among all the clusters. The quantity bought in this cluster is 13 on average although the difference in discount compared to other clusters was insignificant. And this cluster also generated the highest revenue which is \$670.80.

FOR CLUSTER 1:

The mean price is around \$50.45 which is the second highest and product vitality days of 31 which means customers in this category prefer products around this number

The quantity also ordered is 13 on average, with average discount of 3%, revenue from this cluster is of \$639.26 on average.

FOR CLUSTER 2:

The mean price is \$49.11 which is the least and vitality days of 26 which means they mostly purchase product of long-term value average discount of 3% and revenue of \$618.79 making them the least cluster in revenue generation.

THE DECISION:

After further analyses, we discovered a big significant pattern in each of this cluster:

AI For Business Reporting

CLUSTER 0: They are luxury buyers, that buys products of premium and high quality no matter the price and their location are almost evenly distributed among cities

CLUSTER 1: Buys more of health focused product and foods that adds to the wellness of their health, and also wines too.

CLUSTER 2: Buys more of household items and mostly family related products

THE STRATEGY:

A tailored personalized strategy was suggested to improve growth between the clusters:

CLUSTER 0: Consider Pairing complementary luxury products with high class values to boost sales of 2 products, and promotional similar product recommendations to be made.

CLUSTER 1: More of health-conscious products with promotions of the value they deliver and also complementary niche products to be paired with similar purchased products

CLUSTER 2: Tailored promotional household promotion to be made and complementary product shelf placements to be made.

AUTOMATED MONTHLY REPORTS ON BUSINESS PERFORMANCE:

An automated reporting system generates monthly performance summaries using AI-driven insights. These reports are distributed to stakeholders to support data-informed decision-making and performance tracking.

CONCLUSION:

This report demonstrates the power of AI and machine learning in diagnosing sales challenges and guiding strategic interventions. By understanding customer behavior, Jay Grocery Store can implement targeted solutions to drive growth.