Forecasting Unit Sales (Task 1)

Name : Saptharishee
20MIA1150
saptharisheemuthu@gmail.com
91+7708709518

Checkout the Code on :
https://github.com/Saptharishee/Anarix---Forecasting-Unit-Sales-Task-1-

Preprocessing :
https://colab.research.google.com/drive/172n0Roz9pDHO4o5Nqn5KsHJlXCeQmn-_?usp=sharing

Modelling:
https://colab.research.google.com/drive/11WDe0JdAD7zZpwDcnqEduwik2AtFrZ7a?usp=sharing

Introduction

This report provides an analysis of the dataset obtained for the forecasting unit sales task. The focus of this report includes data imputation, visualization, preliminary analysis and modelling to understand the underlying patterns and trends in the data and forecast the unit sales .

Handling Missing Data :

ID	0
date	0
Item Id	2
Item Name	1832
ad_spend	24187
anarix_id	0
units	17898
unit_price	0

Approach 1 : Impute from repeated values

Idea was to identify columns which have repeated values for example we have most of the data from Item Id but vast number of data were missing in Item Name . Idea was to check if there was atleast one same Item Id that had Item Name , but no change was observed .

Approach 2:

ID and Item Id even though are 2 different columns there were some minimal value of Item Id that was missing . meanwhile ID had all the data . It was noted that Id was in the format of " DATE_ITEMID" . Used regex to handle this and fill Item Id

Approach 4 :

If Item name was missing 55% of the time Units the target variable was also missing . This set up ends up creating more noise . Therefore cleared those values where Item Name and Units both were missing . 1023 Rows were removed.

Approach 5:

See if more than 50 % row or Column is missing . In that case scrap that row

Approach 6 :
Impute the columns using KNN imputer . Median is considered if it's a numeric value and most-frequent element is considered in case of categorical value .

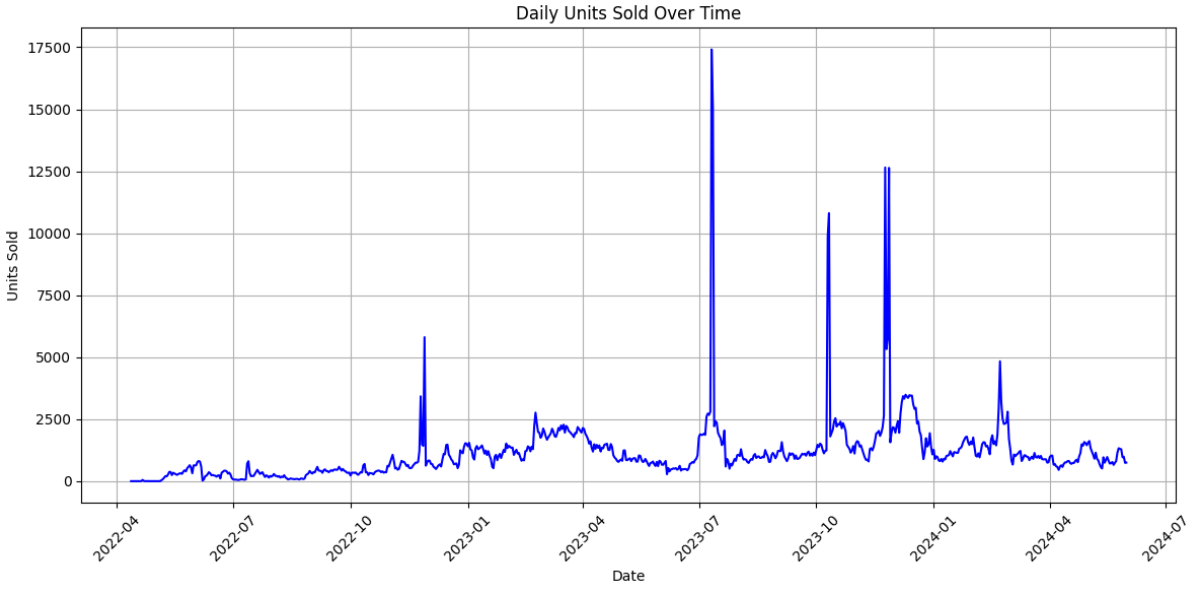
After Handling

ID	0
date	0
Item Id	0
Item Name	0
ad_spend	0
anarix_id	0
units	0
unit_price	0

Visualization

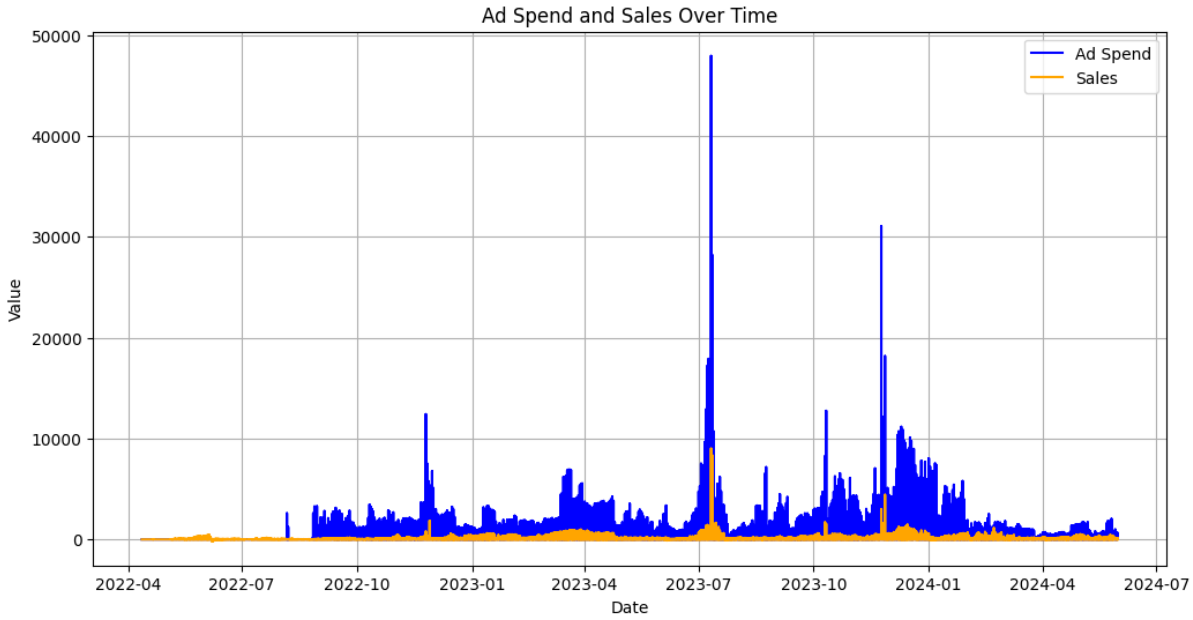
The Time Range of provided data ranges between Earliest Date: 2022-04-12 00:00:00 Latest Date: 2024-05-31 00:00:00

1. **Daily Units Sold Over Time**



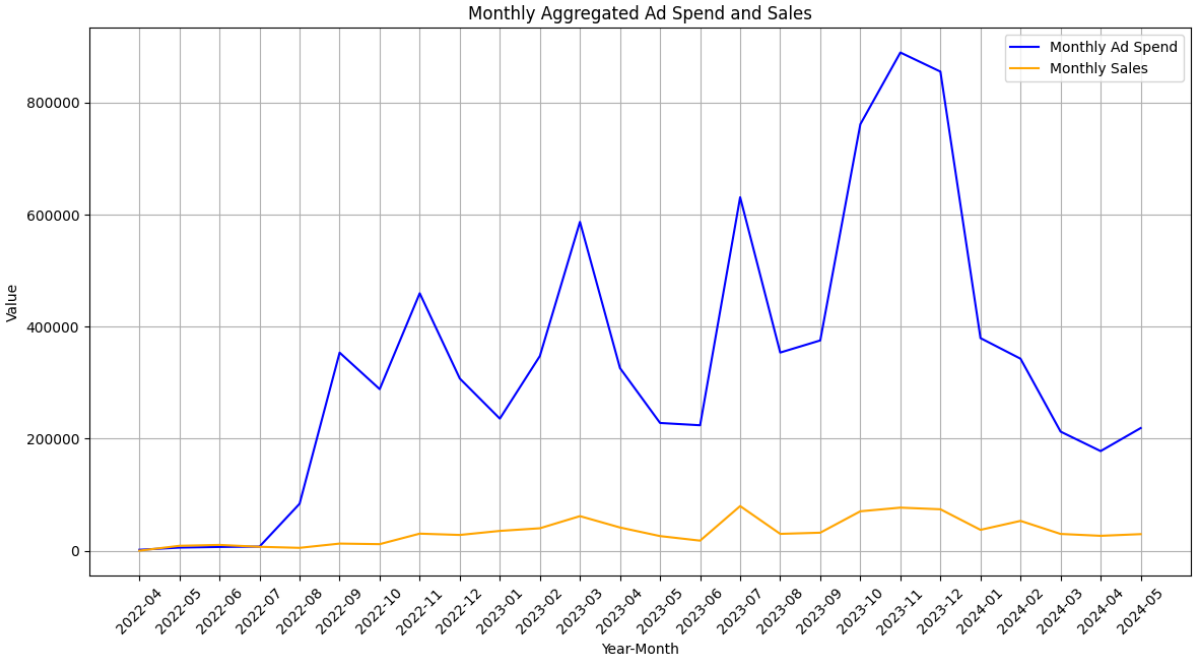

```
High sales spikes, particularlyaround October 2022, July 2023,
October 2023, and another smaller spike around January 2024.
```

2. Ad Spend and Sales Over Time



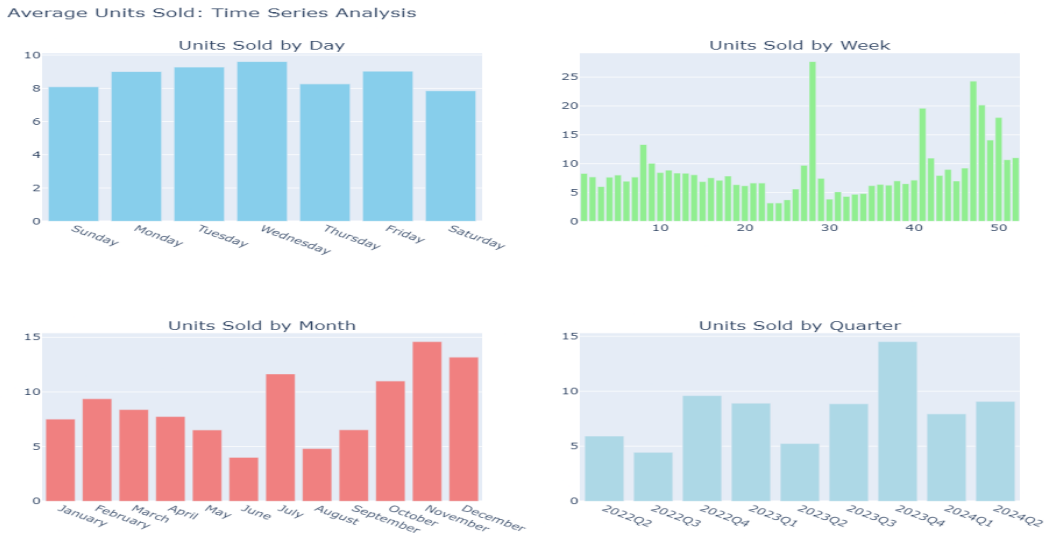
The fluctuations in ad spend do **not** highly correlate with changes in sales but there is a slight relation which has to be looked further . Ad spend is significantly higher than sales which is not a good sign.

3. Monthly Aggregated Ad Spend and Sales



This clearly shows that spending in ad has not really helped and there is a high ad sending from mid 2023 to end of 2023 .

4. Time Series Graphs



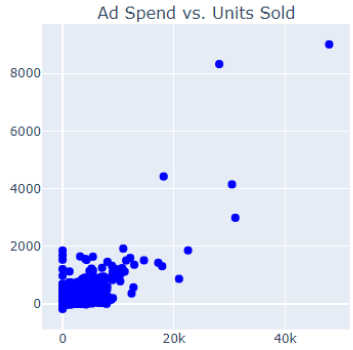
* Units Sold by Day: Sales exhibit fluctuations, with mid-week (Wednesday and Thursday) showing higher average sales.

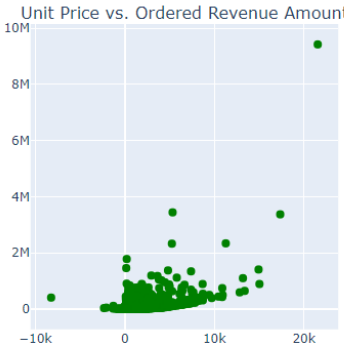
* Units Sold by Month:Holiday Seson of October-November-December have a high sales simialarly Independence day seson of US July also has high sales

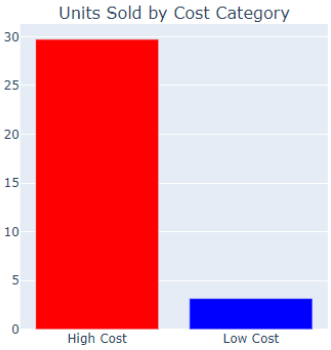
* Units Sold by Week: Sales exhibit variability across weeks, with some weeks showing higher average sales than others. Consider exploring weekly patterns or events that impact sales.

* 2023 Q4 has the best sales of all . Usualy Q4s perform well . Followed by Q2s and Q3s

5. Other Visualizations







* Ad Spend vs. Units Sold:

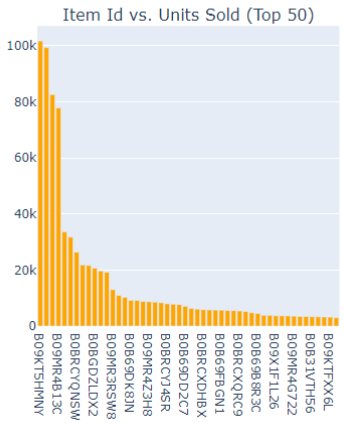
As ad spend increases, units sold tend to increase. There's a positive correlation between advertising investment and sales volume.

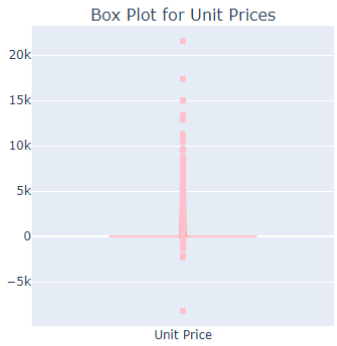
* Unit Price vs. Ordered Revenue Amount:

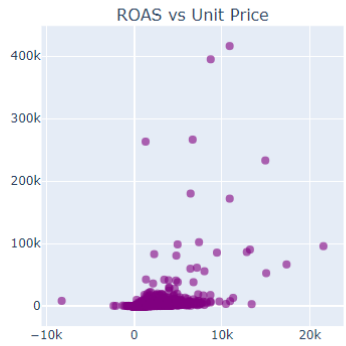
Lower unit prices may lead to higher ordered revenue amounts, although the relationship is not strictly linear. Pricing strategy impacts revenue.

* Units Sold by Cost Category:

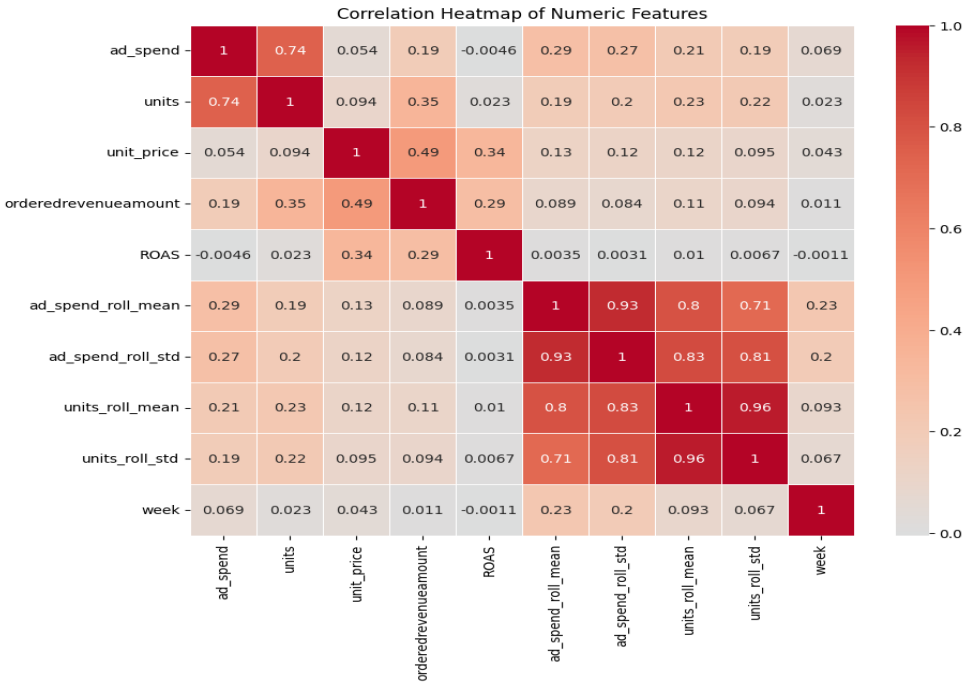
Significantly more units are sold in the 'High Cost' category compared to the 'Low Cost' category. Consider analyzing profitability and market positioning.





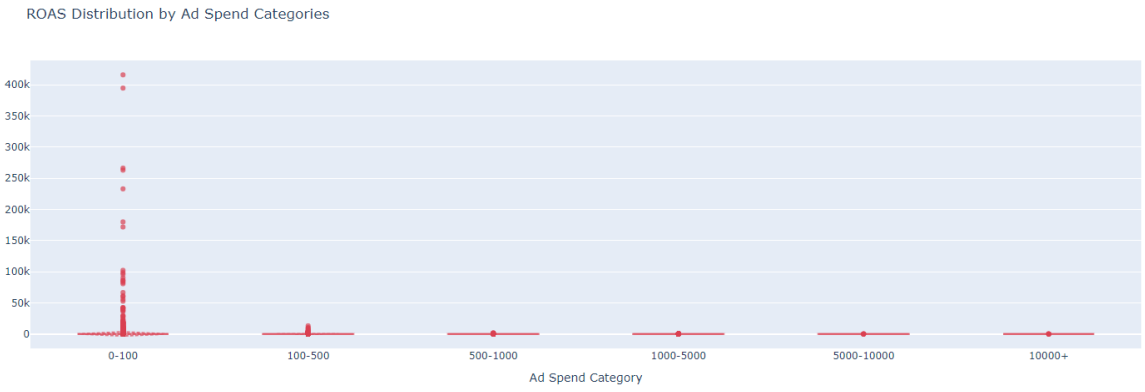


6. Correlation Plot



7.Top Selling Item Name

8.Return on Ad Spending categorized



Return on Ad spent on product higher than 100 is very much low and non existent on products greater than 500

Feature Importance :

Feature Importances:
ad_spend: 0.15199066845270626
unit_price: 0.0373598933141927
orderedrevenueamount: 0.1575621837832993
ROAS: 0.06319847248175957
ad_spend_per_unit: 0.05155252059197297
revenue_per_unit: 0.027237971523698213
Item Id_ohe: 0.0
Item Name_ohe: 0.0
week_ohe: 0.0
month_ohe: 0.0
quarter_ohe: 0.0
cost_category_ohe: 0.0

Note ROAS, ad_spend_per_unit, orderedrevenueamount, revenue_per_unit are all dependent on units therefore they can be useful to gain insights from visualization but not useful in modelling context . Therefore only ad_spend and unit_price are numerical coulums that are of importance.

Py Sprak .

Even though the data is big the simplicity of columns prompted me to check if it will be able to run on a regular Model . Trained the features on RF model without pyspark . Then When trying a stacking approach for this , I realized we might need parallel processing .

Modelling

Model	RMSE	MSE	MAE	R2 Score
ARIMA	35.32	1247.67	9.52	-0.0091
LSTM	2145511.41	4603219228018.12	897936.22	-0.0026
Bi-LSTM	1903348.21	3622734393735.93	604193.89	-0.0016
Bi-GRU	1905364.42	3630413567107.20	597320.35	-0.0038
Random Forest	5.20	27.05	2.28	0.0993
Gradient Boosting	5.26	27.63	2.37	0.0800
Decision Tree	5.26	27.62	2.36	0.0804
Stacking	0.83	0.69	0.06	-0.0049

- Approaches made include hyper parameter tuning in LSTM which ran for 5+hrs for 41 Trial runs but was still not able to decide on a model .

- Statistical Model Arima and sequential models like LSTM,GRU and their bi directional variant did not perform well.

- Tried a Stacking approach of ML models which was not able to be reflected in submission .

- The metrics for LSTM, Bi-LSTM, and Bi-GRU are unusually high, which might indicate an issue with the model setup or data scaling.
- The stacking model, which combines multiple models, shows significantly lower RMSE, MSE, and MAE compared to individual models, but a similar R2 score