Data Science Challenge

Prepared by; Robert Gembe Marcely

Contents

Objectives

- Develop a forecasting model to predict future demand for healthcare product sales in Tanzania for the next 6 months starting from June 2024.
- The goal is to help local pharmacies make informed decisions and improve health outcomes by anticipating healthcare needs.

Dataset

- A dataset containing historical data on healthcare product sales, including variables such as:

➢ Pharmacy name

➢ Product code

➢ Product name

➢ Timeline(Month, year)

➢ Sales

Data Exploration and Cleaning

- It includes the following;

1. Understand the Data: Through Knowing the context of the data and Identify variables and their types.
2. Initial Data Exploration: Through Loading the data, Displaying a sample of the data and Generate basic descriptive statistics.

Data Exploration and Cleaning

- Displaying a sample of the data

Out[2]:

	Pharmacy Name	Product Code	Product Name	Month	Year	Sales
0	TEMEKE PHARMACY	10010194AC	LEVONORGESTREL IMPLANT 75MG	October	2023	577098.0
1	TEMEKE PHARMACY	10010106AC	LEVONORGESTREL 0.15MG + ETHINYLESTRADIOL 0.03 ...	February	2024	1005058.0
2	UBUNGO PHARMACY	10010194AC	LEVONORGESTREL IMPLANT 75MG	February	2023	436704.0
3	ilala pharmacy	10010353AC	LEVONORGESTREL TABLETS 0.75 mg (2TB)	March	2023	NaN
4	TEMEKE PHARMACY	10010106AC	LEVONORGESTREL 0.15MG + ETHINYLESTRADIOL 0.03 ...	August	2023	NaN
...	...	...	...	...	...	...
445	KINONDONI PHARMACY	10010106AC	LEVONORGESTREL 0.15MG + ETHINYLESTRADIOL 0.03 ...	January	2023	552827.0
446	KINONDONI PHARMACY	10010353AC	LEVONORGESTREL TABLETS 0.75 mg (2TB)	January	2023	434411.0
447	KINONDONI PHARMACY	10010194AC	LEVONORGESTREL IMPLANT 75MG	January	2023	NaN
448	Kigamboni Pharmacy	40030134AC	Copper T IUD	January	2023	27.4
449	Kigamboni Pharmacy	10010108AC	CONDOMS	January	2023	NaN

450 rows × 6 columns

Data Exploration and Cleaning

- Generate basic descriptive statistics.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 450 entries, 0 to 449
Data columns (total 6 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Pharmacy Name   450 non-null    object
 1   Product Code    445 non-null    object
 2   Product Name    444 non-null    object
 3   Month           450 non-null    object
 4   Year            450 non-null    int64
 5   Sales           351 non-null    float64
dtypes: float64(1), int64(1), object(4)
memory usage: 21.2+ KB
```

Data Exploration and Cleaning

3. Data Cleaning: Through Handle missing values(Remove or impute), Address inconsistencies(Correct errors), Standardize formats, Remove duplicates(Handle outliers).

- Checking for a missing values

```
Pharmacy Name     0
Product Code      5
Product Name      6
Month             0
Year              0
Sales            99
dtype: int64
```

Data Exploration and Cleaning

- Columns with missing values were as shown in above figure
- Not a Number(NaN) for Sales Column were replaced by Zero(0) because I assumed that; no sales were made at that time, hence setting it to zero.
- Combining Month and Year to Timeline; It provides a unique identifier for each time period, which is essential for time series analysis and ensures clarity when referencing specific periods.

Data Exploration and Cleaning

- Check for rows where both 'Product Name' and 'Product Code' are NaN.

```
Empty DataFrame
Columns: [Pharmacy Name, Product Code, Product Name, Month, Year, Sales]
Index: []
```

- If the output is an empty DataFrame, it means that there are no rows in the DataFrame where both 'Product Name' and 'Product Code' are NaN.

Data Exploration and Cleaning

- Filter rows where Product name is NaN and collect Product codes and then Create a dictionary mapping Product codes to Product names. Replace NaN Product names with Corresponding Product names
- The procedure was repeated also for Product Code which were NaN

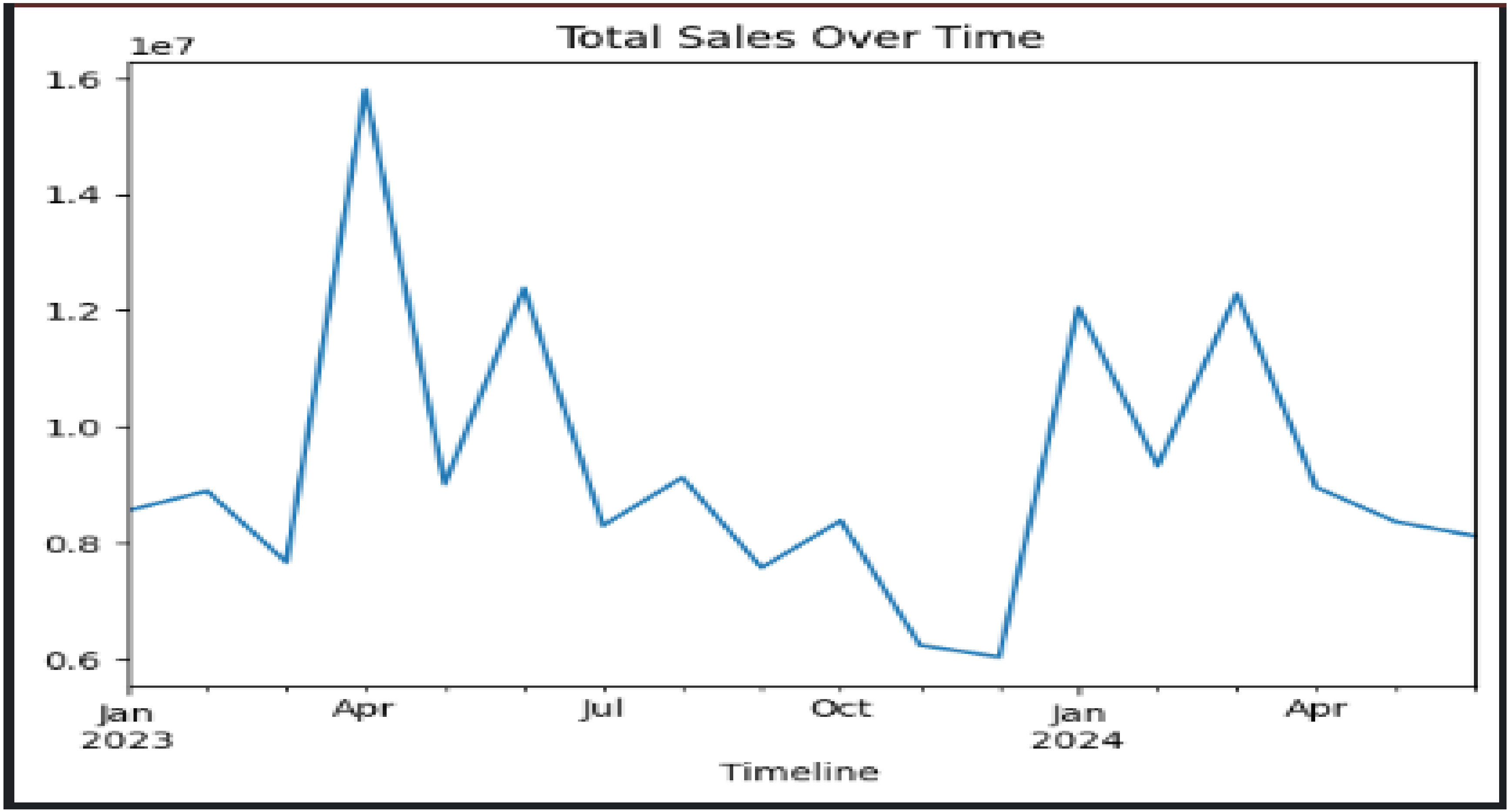
Data Exploration and Cleaning

- After Handling all missing values

```
Pharmacy Name    0
Product Code     0
Product Name     0
Month            0
Year             0
Sales            0
dtype: int64
```

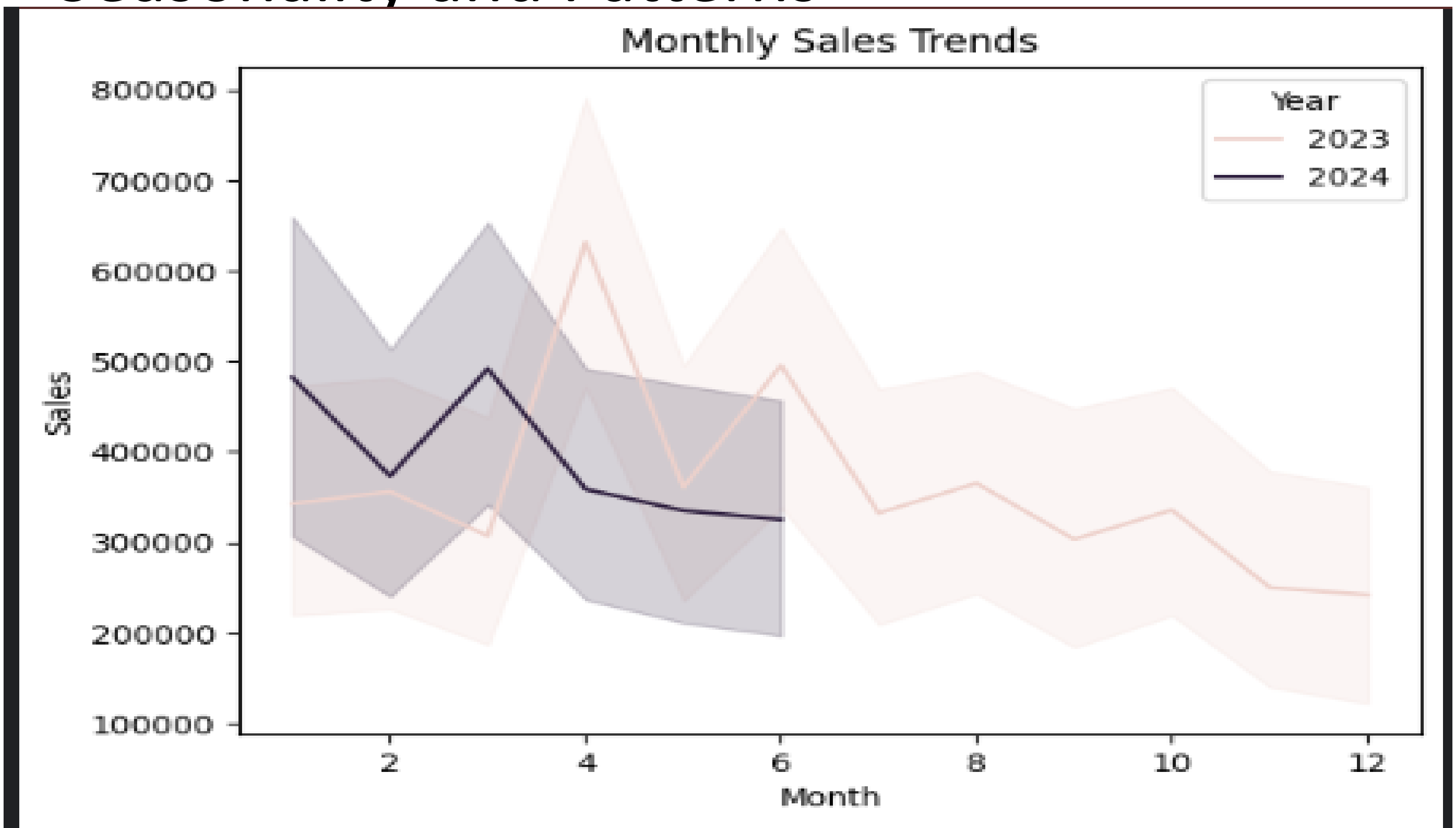
Data Exploration and Cleaning

- Visualize Sales trends over time



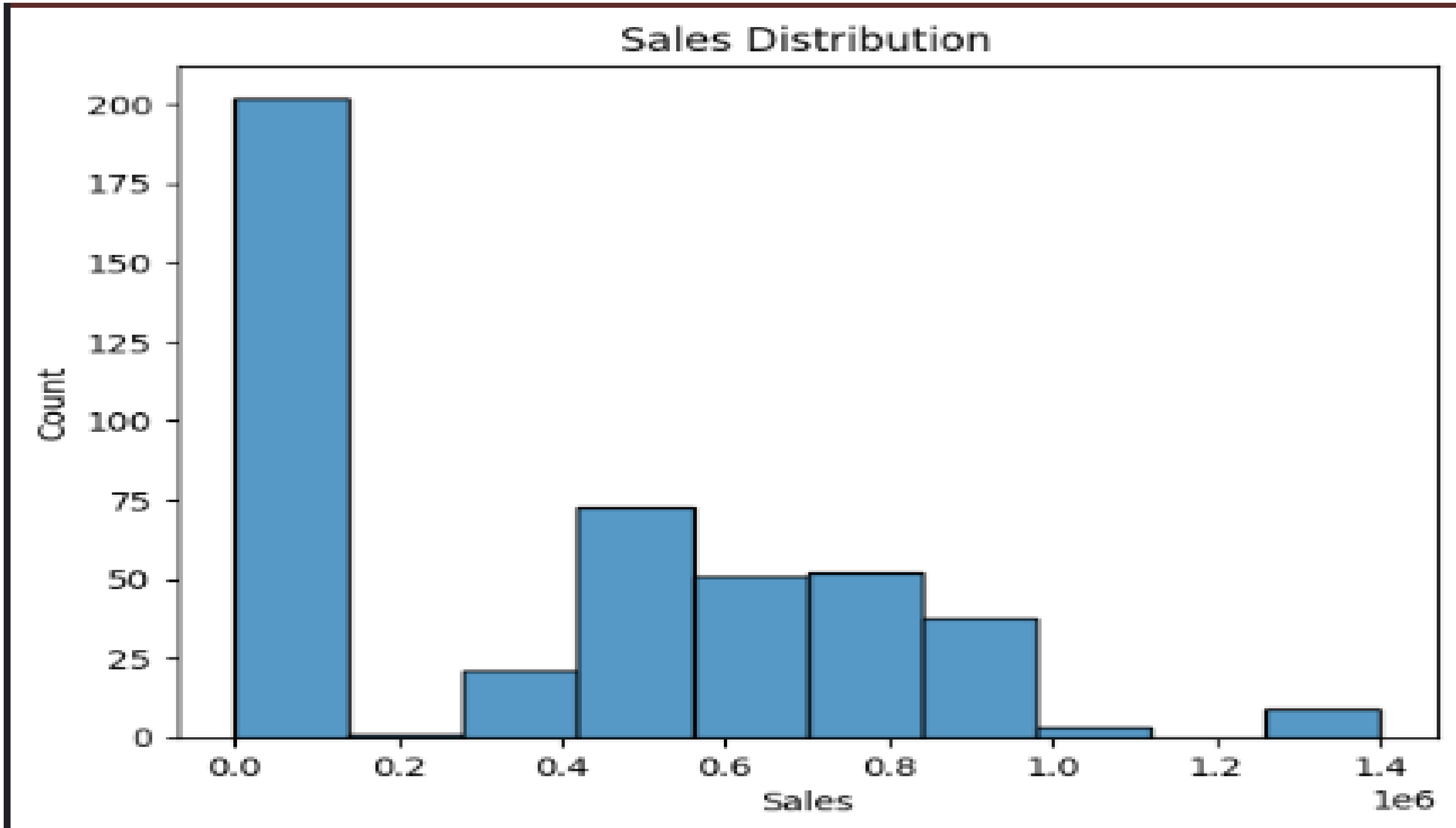
Data Exploration and Cleaning

- Seasonality and Patterns



Data Exploration and Cleaning

- Sales Distribution



Data Exploration and Cleaning

- Other Data Manipulations

```
Unique counts in each column:
 Pharmacy Name        5
Product Code          5
Product Name          5
Month                12
Year                  2
Sales               334
dtype: int64

Unique Pharmacies: 5
Unique Product Codes: 5
Unique Product Name: 5
```

Data Exploration and Cleaning

- Other Data Manipulations

```
Occurrences of each Pharmacy Name:
 Pharmacy Name
TEMEKE PHARMACY        90
UBUNGO PHARMACY        90
ilala pharmacy         90
Kigamboni Pharmacy     90
KINONDONI PHARMACY     90
Name: count, dtype: int64

Occurrences of each Product Code:
 Product Code
10010106AC    120
10010108AC     95
40030134AC     95
10010194AC     70
10010353AC     70
Name: count, dtype: int64

Occurrences of each Product Name:
 Product Name
LEVONORGESTREL IMPLANT 75MG                                                                                        135
LEVONORGESTREL TABLETS 0.75 mg (2TB)                                                                                85
CONDOMS                                                                                                             85
LEVONORGESTREL 0.15MG + ETHINYLESTRADIOL 0.03 MG + FERROUS FUMEARATE 75 MG (MICROGYNON) TABLETS 0.1 + 0.3 + 75 mg/mg/mg (1CY)     85
Copper T IUD                                                                                                        60
Name: count, dtype: int64
```

Model Development

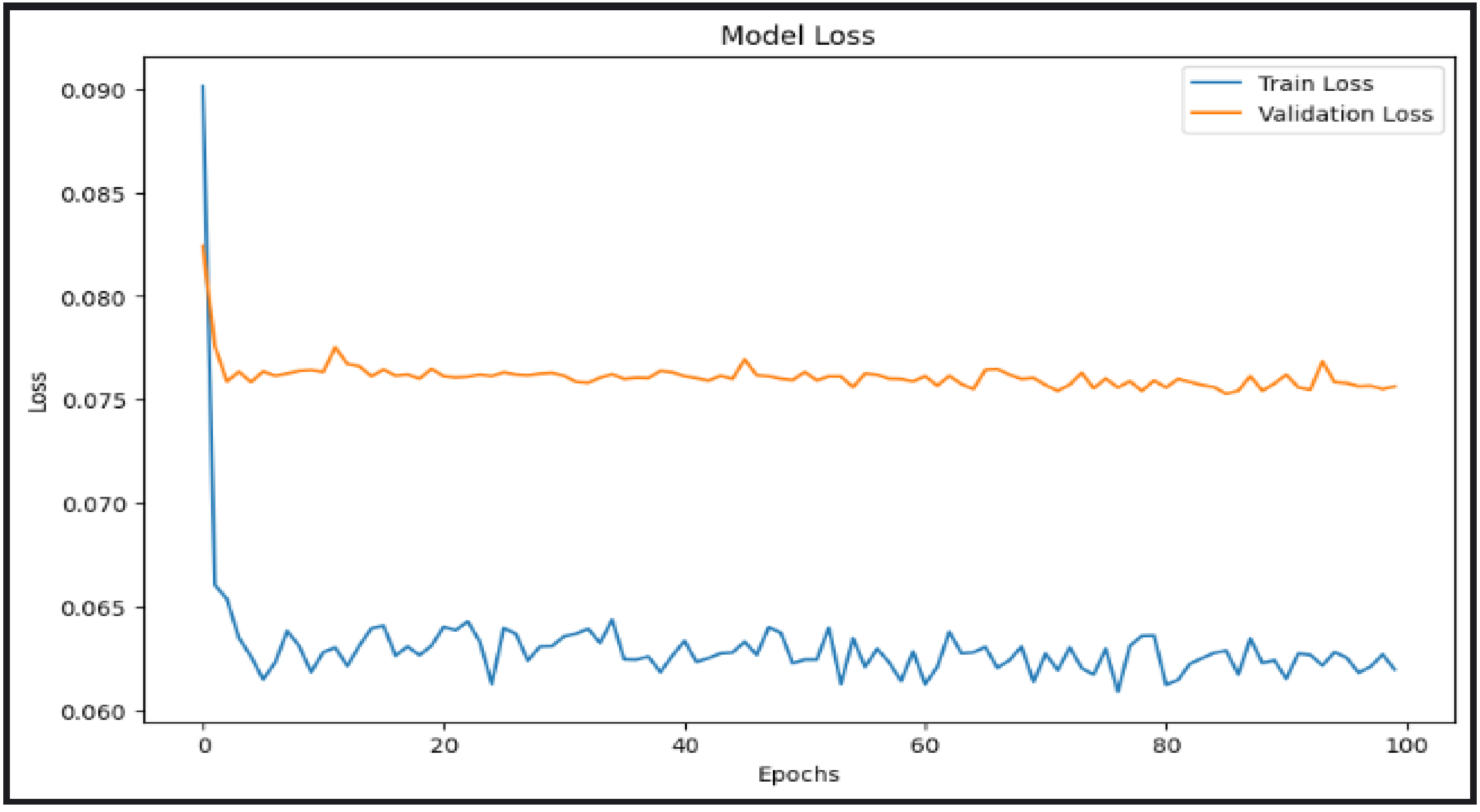
- Long Short-Term Memory (LSTM) networks were used.
- Long Short-Term Memory (LSTM) networks are a type of recurrent neural network (RNN) that are particularly well-suited for time series forecasting due to their ability to capture long-term dependencies and patterns in sequential data.

Model Development

- In summary, LSTM networks are a powerful tool for time series forecasting due to their ability to capture long-term dependencies, handle complex temporal patterns, and adapt to various types of time series data, all while mitigating common issues associated with traditional RNNs.

Model Evaluation

- Model Evaluation



Model Evaluation

- A model was trained for 100 epoch. Our model trained a data for accuracy around 60’s percent and validate the data around 70’s percent.
- No overfiting or underfiting at all.

Model Evaluation

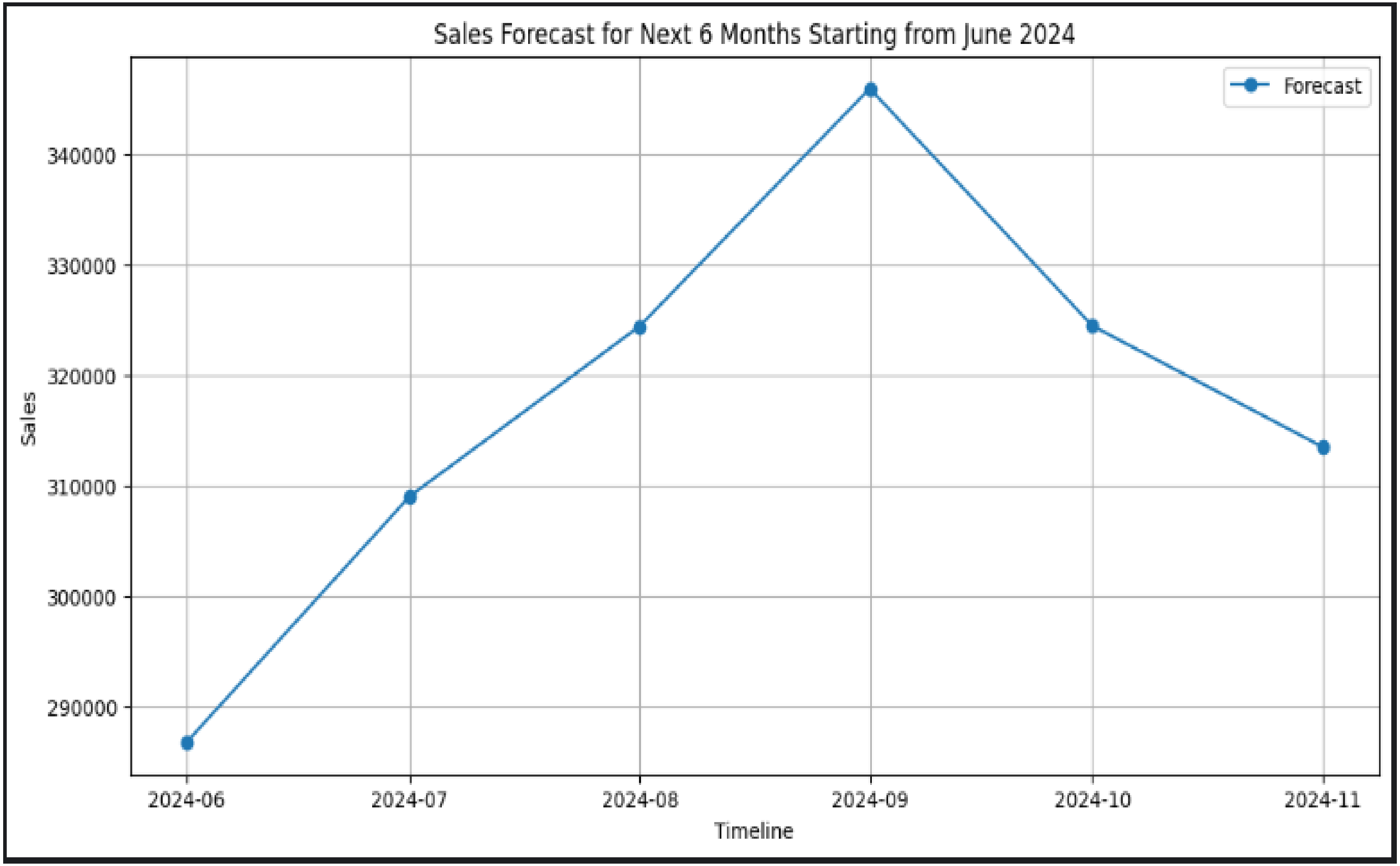
- Evaluation Metrics

```
11/11 ━━━━━━━━━━━━━━━━━━━━ 1s 32ms/step
3/3 ━━━━━━━━━━━━━━━━━━━━ 0s 8ms/step
Train RMSE: 346526.6977766928
Test RMSE: 385075.135878
```

Results/Output

- Prediction for future demand for healthcare product sales in Tanzania for the next 6 months starting from June 2024 is shown below;

Results/Output



Recommendation

- From the graph which predicts the future demands, it seems that on September the demand will be high since sales is at a climax/at the peak.
- I recommend all pharmacy to have enough stock for their products in order to accommodate that demands.

Conclusion

- All Documentation about this project is found in my github repository;

[GitHub - Robert-Xsa/DATA-SCIENCE-CHALLENGE-AFYA-INTELLIGENCE]

- Jupyter notebook, Processed dataset and all python code for this project is found in my kaggle platform for data scientist;

https://www.kaggle.com/code/robertgembe/data-science-challenge-afya-intelligence-tanzania/

Conclusion

My Email; gemberobert1@gmail.com

My Contact; +255757685690

My Linkedin; https://www.linkedin.com/in/robert-gembe-marcely/

Thanks in Advance