AZUBI AFRICA

Capstone Project

Grocery Store Forecasting Challenge For Azubian Presentation.

Seoul Group

AZUBI AFRICA

WHAT THE PROJECT IS

The goal of this project was to develop a model to predict the volume of goods that will be bought each week by each store throughout the course of the next eight weeks, for grocery stores spread across various regions of the same nation.

IMPORTANCE

Inventory Management
Manufacturing
Planning Resource
Distribution Financial
Preparation fulfillment

BENEFITS

Data-driven
assessments, Decisions
lower costs Operating
Effectiveness
Competitive Benefit

DESCRIPTION OF DATA

AZUBI AFRICA

This time series forecasting project involves daily sales data for 4 years, and the dataset is comprised of several files:

01 **train.csv**

Contains the target. This is the dataset that you used to train the model.

`date store_id category_id target onpromotion nbr_of_transactions`

02 **holidays.csv**

Information about holidays

`date type`

03 **test.csv**

Resembles Train.csv but without the target-related columns. This is the dataset on which the trained model will be applied to

`date store_id category_id onpromotion`

04 **stores.csv**

Information about the different stores such as their locations

`store_id city type cluster`

05 **dates.csv**

Information about the time periods with their associated date features

`date year month dayofmonth dayofweek dayofyear weekofyear quarter is_month_start is_month_end is_quarter_start is_q`

06 **SampleSubmission.csv**

Shows the submission format for this competition, with the 'ID' column mirroring that of Test.csv.

THE ISSUES	THE SOLUTIONS
♦ The data type for the date column in train, test, holidays, dates datasets are in numerical format	♦ Convert to Datetime format
♦ stores dataset: the city, type & cluster are in numerical format	♦ Convert to string and make the categories more descriptive
♦ The type column in holidays dataset is in a numerical format and the types of holidays do not have an ordinal Relationship	♦ Convert to string and make it more descriptive
♦ After checking for unique values in year of day column in the dates dataset, we found 366 unique values which showed that there was a leap year.	♦ Create two new columns called "sin(dayofyear)" & "cos(dayofyear)". These new columns will help our machine learning models understand the cyclic nature of a year.

ANALYSIS & CLEANING OF THE DATA

AZUBI AFRICA

Observations

- The date column in dates data frame is from 365 - 1684 which covers the date column in the train and test datasets.
- There are no null values in any of the datasets
- There are no missing values
- There is no holiday column in any of the other datasets.

Data Cleaning

- The Year column in dates dataframe has 4 unique vales. Replace them with [2001, 2002, 2003, 2004] by creating a dictionary called replacements and using the replace method.
- Combine the year, month and day columns in dates data frame to create a fulldate column. This column will be used as the index for dates, train, test and holidays data frames.
- Create a new column in the train DataFrame called "holiday_type". The value of this column is the type of holiday in the holidays DataFrame for the corresponding date in the train DataFrame. If the date is not a holiday, the value of the column is "WD".

HYPOTHESIS & QUESTIONS

AZUBI AFRICA

HYPOTHESIS

H0: Holidays do not affect Sales, therefore the sales data is stationary.

H1: Holidays affects Sales, therefore the sales data is seasonal.

QUESTIONS ABOUT THE DATA

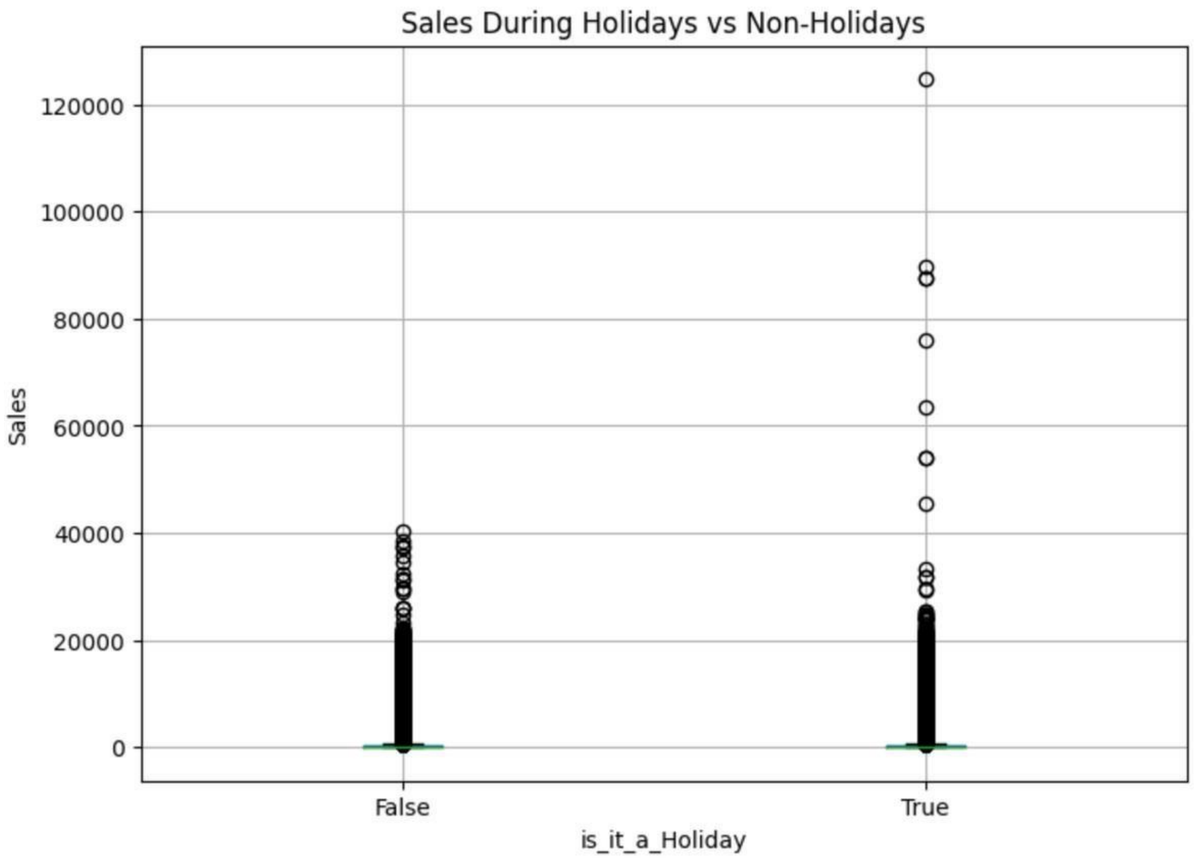
1. What is the distribution of sales?
2. What is the average sales for each category?
3. How do sales vary by promotion status?
4. Is there a relationship between sales and the number of transactions?
5. How do sales vary during holidays compared to non-holidays?

QUESTIONS ABOUT THE DATA

6. How do sales vary by holiday type?
7. What is the trend in sales over time?
8. How do sales vary across different store IDs?
9. Are there any seasonal patterns in sales?
10. How do sales vary across different combinations of category and promotion?

HYPOTHESIS VALIDATION

AZUBI AFRICA



Conclusion

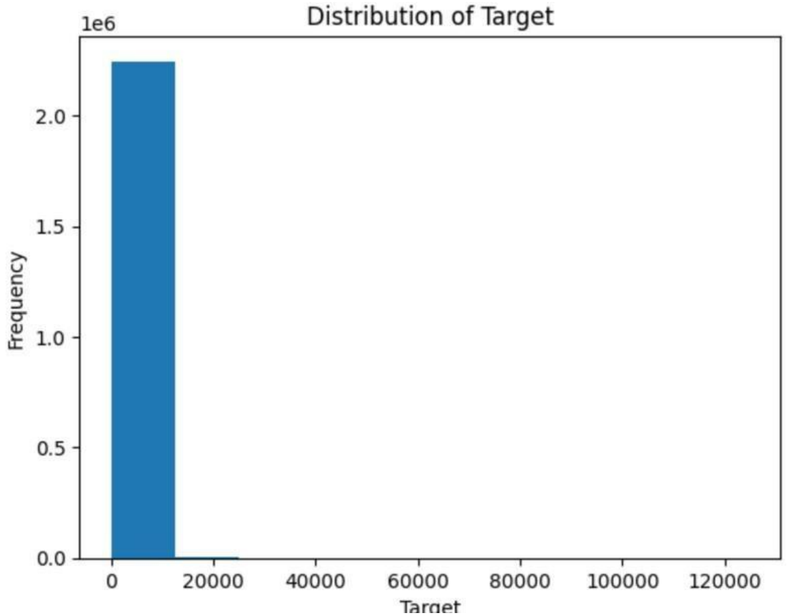
The hypothesis H1, which states that holidays affect sales and the sales data is seasonal, is more likely to be true

Null Hypothesis REJECTED

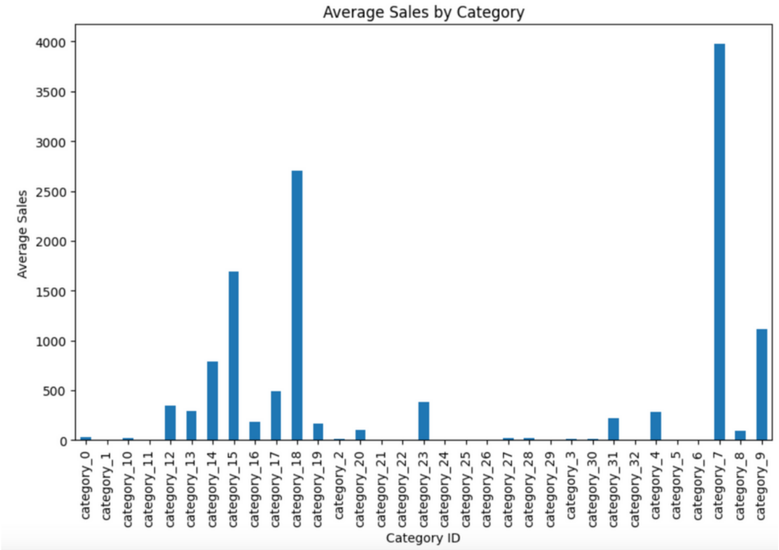
ANSWERING THE QUESTIONS

AZUBI AFRICA

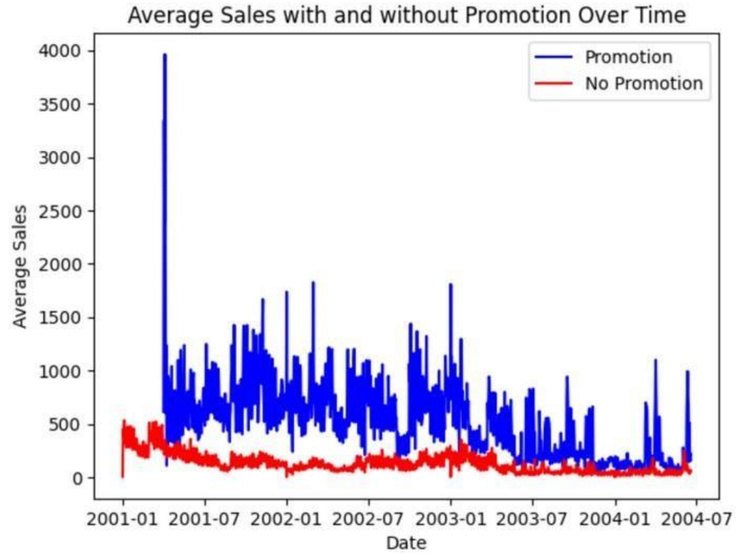
What is the distribution of Sales?



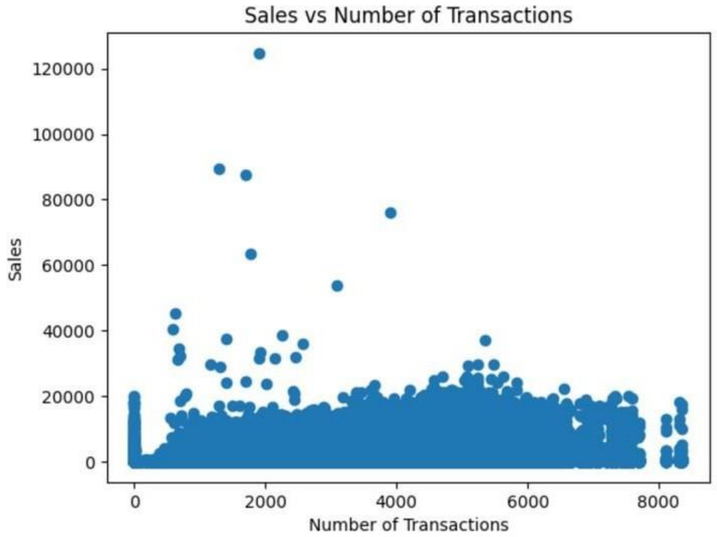
What is the average sales for each category?



How do sales vary by promotion status?



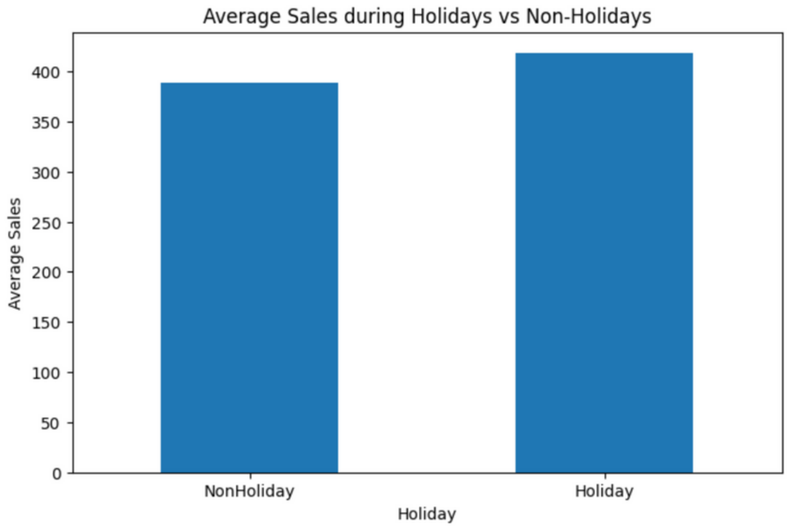
4. Is there a relationship between sales and the number of transactions?

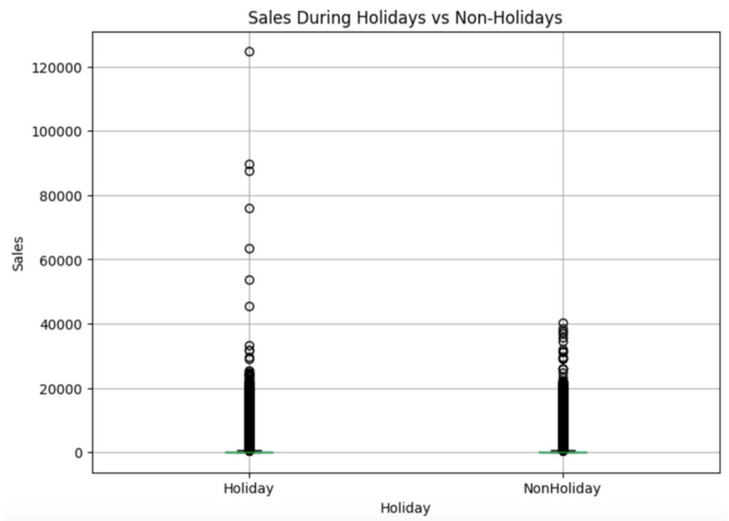


ANSWERING THE QUESTIONS

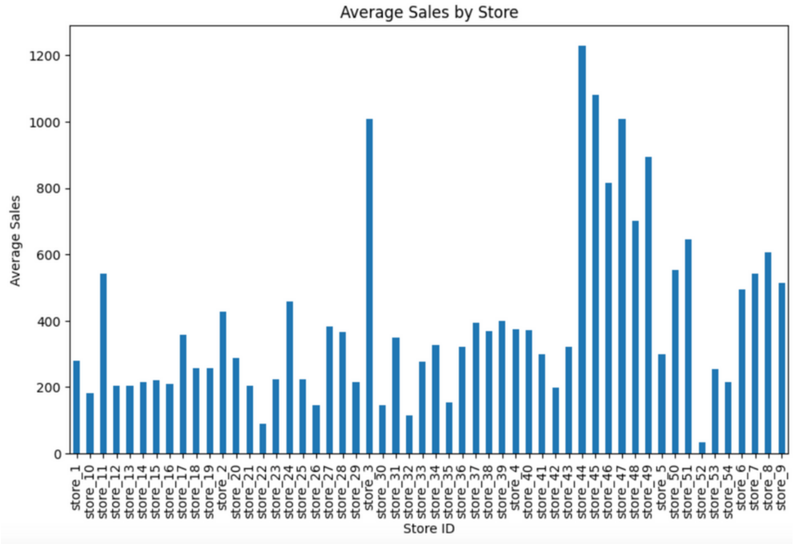
AZUBI AFRICA

How does sales vary during holidays compared to non-holidays?

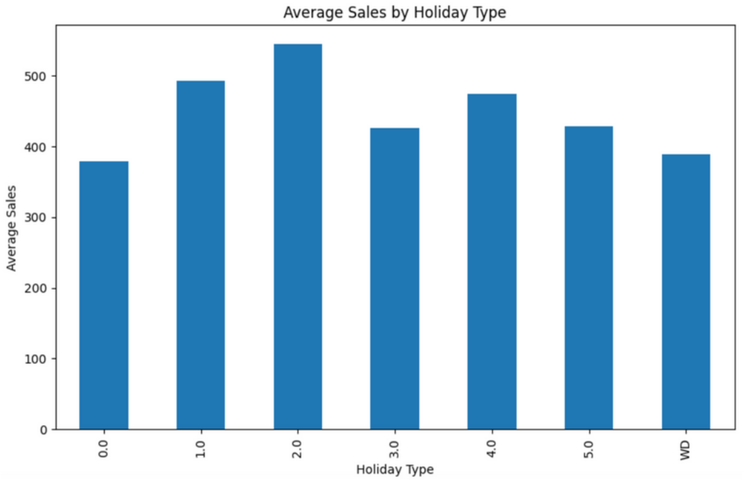




How do sales vary across different store IDs?



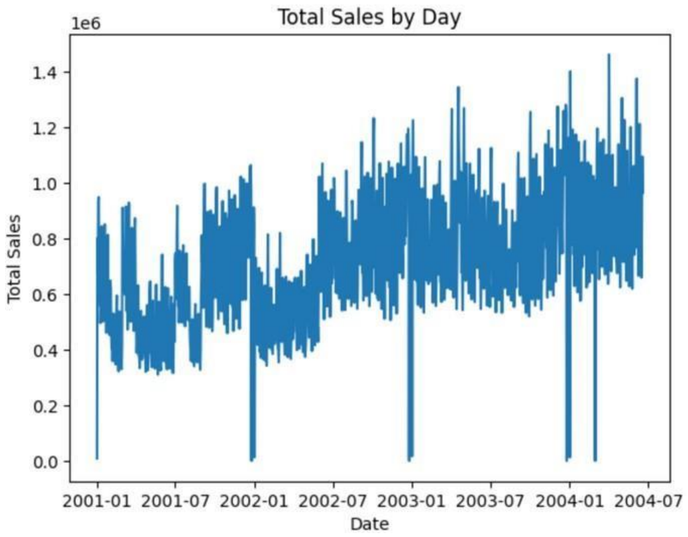
How do sales vary by holiday type?

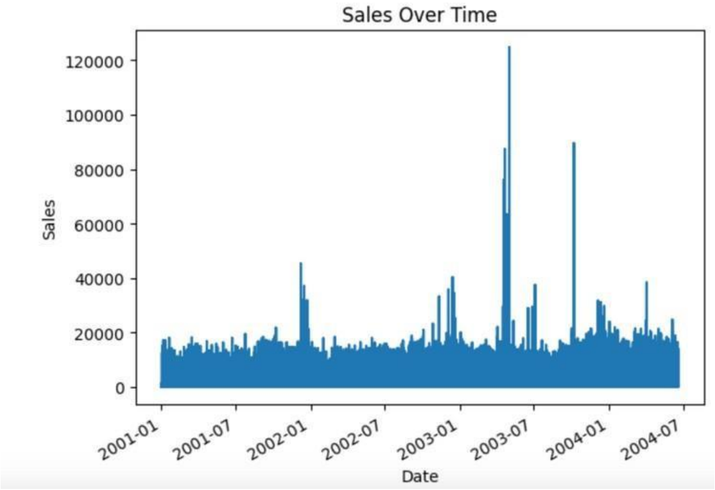


ANSWERING THE QUESTIONS

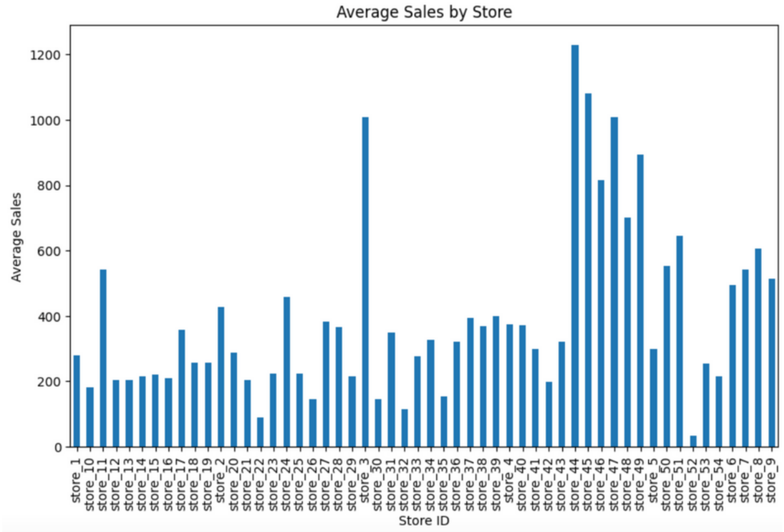
AZUBI AFRICA

What is the trend in sales over time?

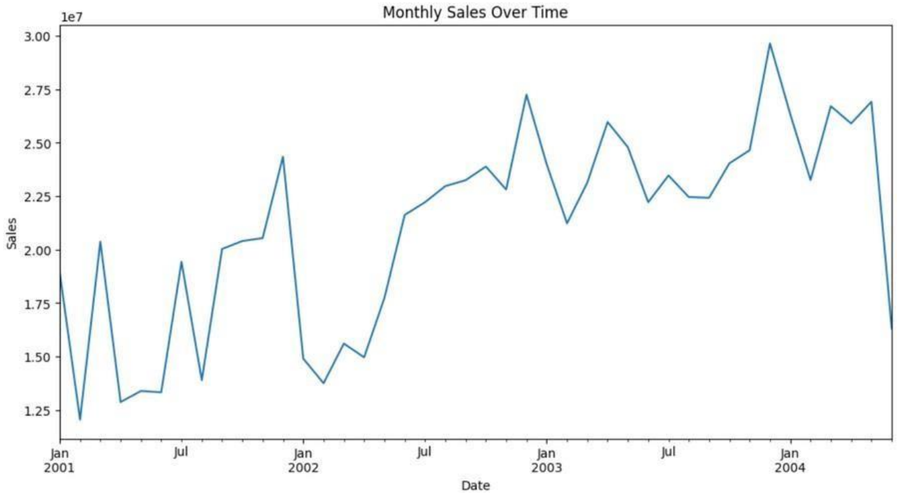


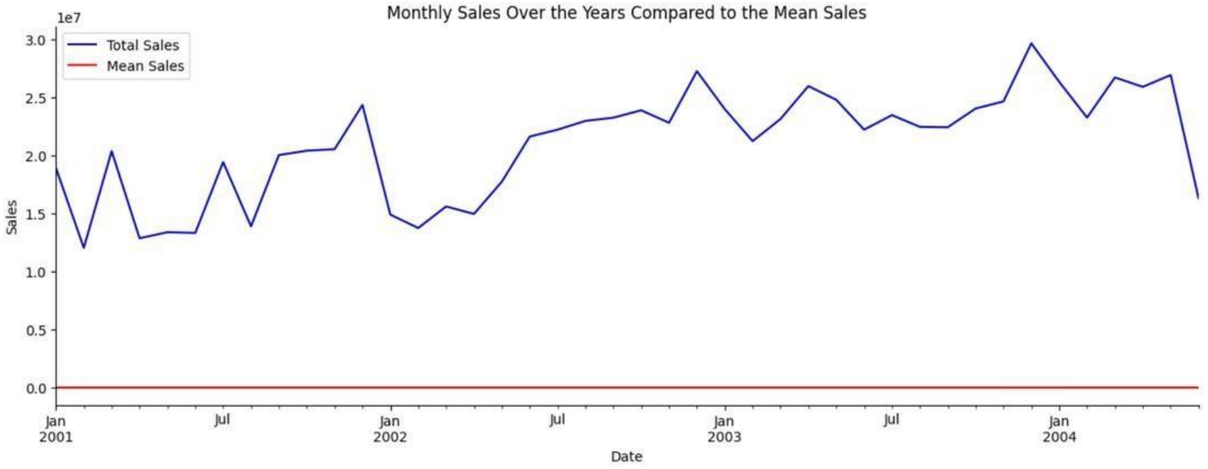


How do sales vary across different store IDs?



Are there any seasonal patterns in sales?

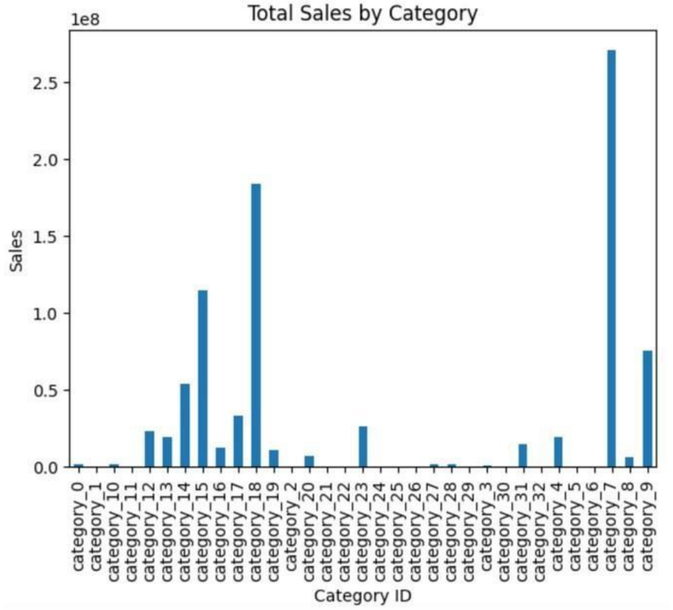


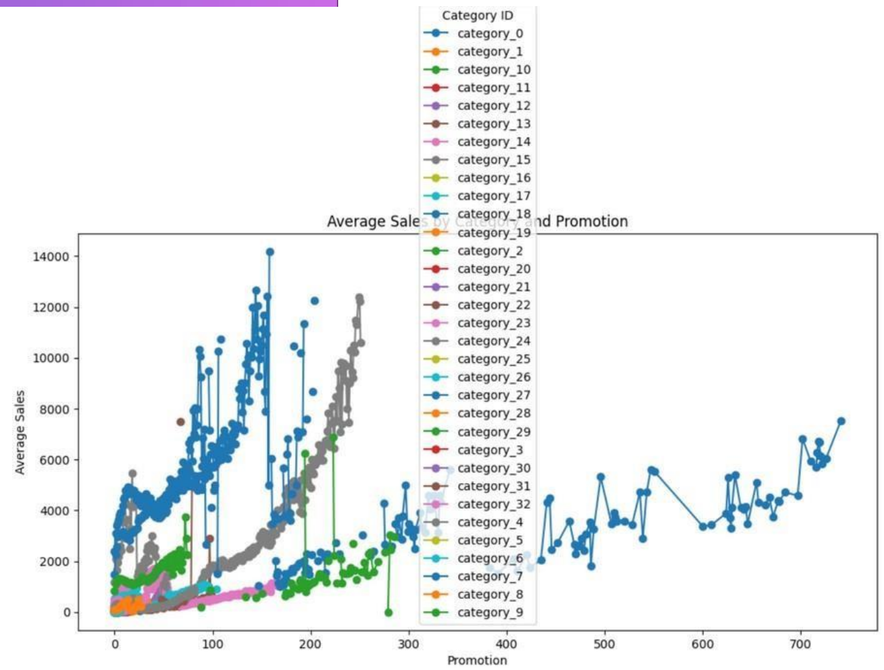


ANSWERING THE QUESTIONS

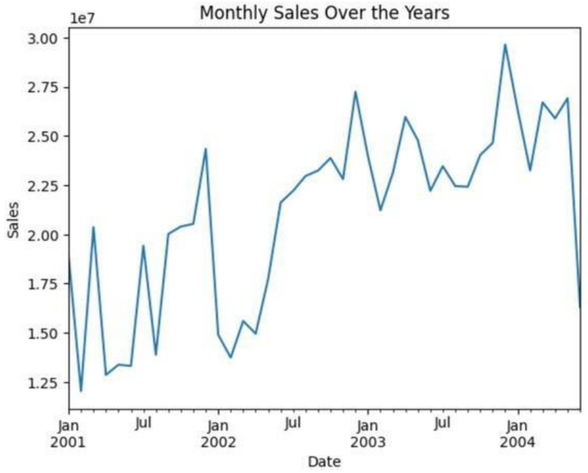
AZUBI AFRICA

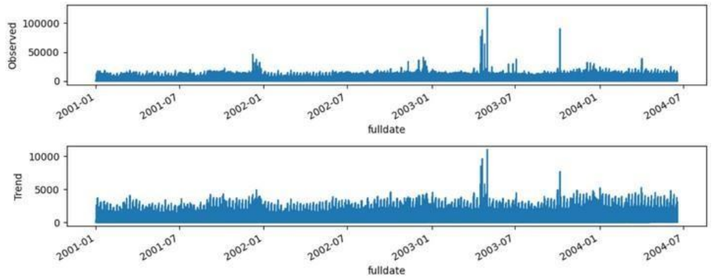
How do sales vary across different combinations of category and promotion?

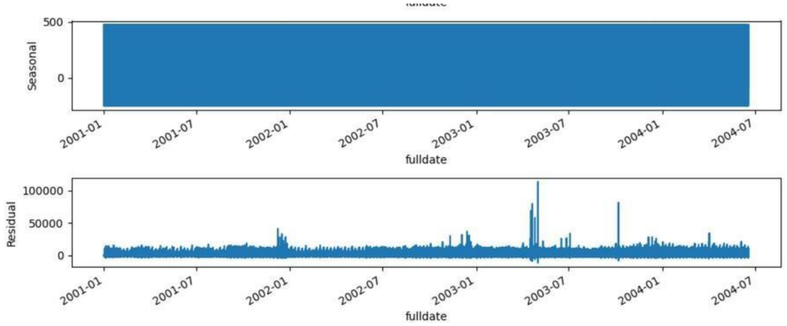




Monthly Statistics







MACHINE LEARNING MODELS

(TIME SERIES)

AZUBI AFRICA

MODELS

- DecisionTreeRegressor
- K-Nearest Neighbors (KNN)
- RandomForestRegressor
- Support Vector Regression (SVN)
- Gradient Boosting
- XGBoost
- Linear Regression

MODELS COMPARISON

	Model	MSE	MSLE	RMSE	RMSLE
0	DecisionTree	17878.36	0.12	133.71	0.35
1	KNN	15157.08	0.11	123.11	0.33
2	Random Forest	18626.54	0.13	136.48	0.36
3	SVR	16025.20	0.11	126.59	0.34
4	Gradient Boosting	18096.48	0.12	134.52	0.35
5	XGBoost	21284.22	0.14	145.89	0.37
6	Linear Regression	1043641.46	1.65	1021.59	1.28

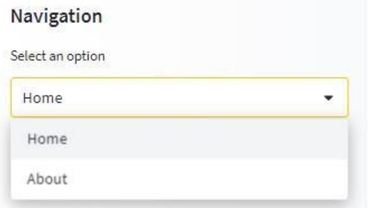
CONCLUSION

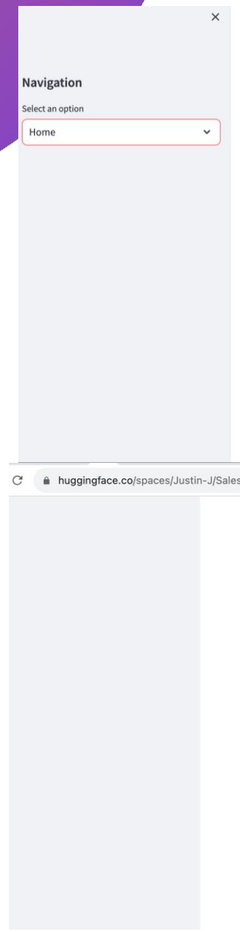
Based on the provided results, the best performing model was the **KNN model** with an **RMSE score of 123.11**

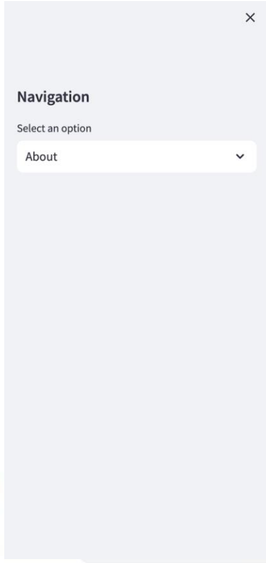
The worst performing model was **Linear Regression** with an **RMSE score of 1029.59**

THE APP

AZUBI AFRICA







SEER

A Streamlit Sales Forecasting / Prediction App

SEER is a powerful tool designed to assist businesses in making accurate and data-driven sales predictions. By leveraging advanced algorithms and machine learning techniques, our app provides businesses with valuable insights into future sales trends. With just a few input parameters, such as distance and average speed, our app generates reliable sales forecasts, enabling businesses to optimize their inventory management, production planning, and resource allocation. The user-friendly interface and intuitive design make it easy for users to navigate and obtain actionable predictions. With our Sales Forecasting App, businesses can make informed decisions, mitigate risks, and maximize their revenue potential in an ever-changing market landscape.

Meet Our Team

Justin Jabo

Wycliffe Omondi Ayodo

Stephen Arhin-Aidoo

Lionel Boris Rene Bizo Mendome

Alex Saruni Lodaru

Marydiana Njeri Njoroge

AZUBI AFRICA

Quotes Today

"The goal of forecasting is not
to predict the future but
to tell you what you need to know
to take meaningful action in the present."

AZUBI AFRICA

Thank You

Seoul Group