**Wooly Transactions Analysis Report**

Introduction

This report details the analysis of the Wooly Transactions dataset, including exploratory data analysis (EDA), feature engineering, predictive modeling, and clustering of transactions.

Dataset Overview

The dataset consists of the following columns:

Date: The date of the transaction.

Suburb: The suburb where the transaction took place.

Address: The address of the store where the transaction occurred.

Invoice Number: The invoice number for the transaction.

Item Description: Description of the item purchased.

Unit Price: The price per unit of the item (NaN values present).

Item Price: The total price for the item.

Exploratory Data Analysis (EDA)

Summary Statistics:

The dataset contains 318 transactions with 62 unique dates, 16 unique suburbs, 19 unique addresses, 64 unique invoice numbers, and 230 unique item descriptions.

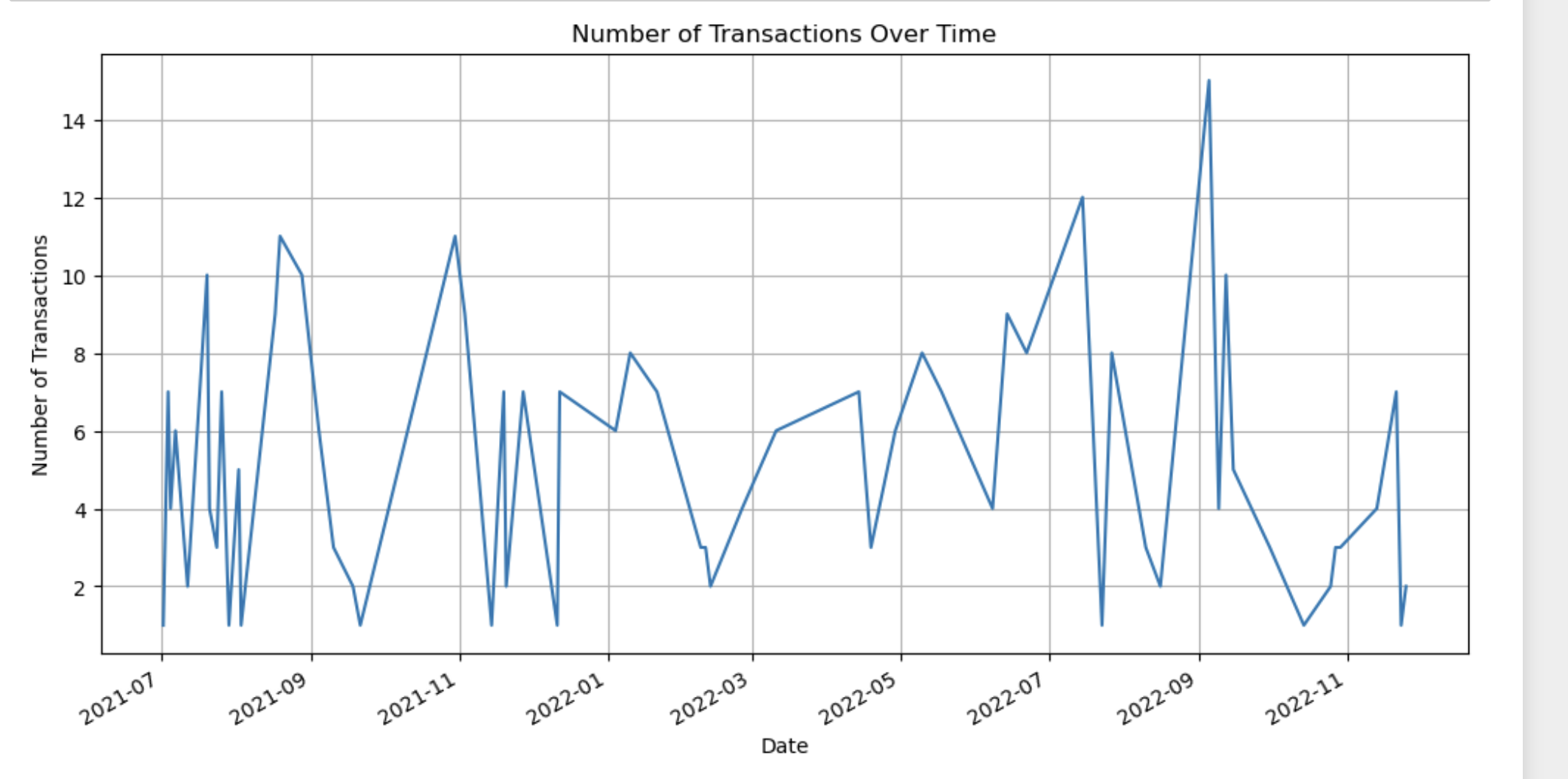
The average item price is $4.53 with a maximum of $50.00.

Missing Values:

The Unit Price column has 256 missing values.

Transactions Over Time:

A line plot shows the distribution of transactions over the given dates, indicating patterns such as peaks and troughs.



Most Frequently Purchased Items:

The most frequently purchased item is Colgate MW Plax Fresh Tea 500ml, followed by Cauliflower Whole and Pumpkin Butternut Cut.

Transactions by Suburb and Address:

The top suburb with the highest number of transactions is 3191 Lilydale, and the top address is 33 Hutchinson Street.

Feature Engineering

Date Features:

Extracted day, month, and year from the Date column to create new features.

Categorical Encoding:

Encoded the Suburb, Address, and Item Description columns using Label Encoding.

Predictive Modeling

Model Training:

Trained four different models: Linear Regression, Decision Tree Regressor, Random Forest Regressor, and Gradient Boosting Regressor.

Evaluated models using Mean Absolute Error (MAE).

Model Performance:

Decision Tree Regressor performed the best with an MAE of 2.60, followed by Gradient Boosting Regressor (MAE = 2.71) and Random Forest Regressor (MAE = 2.74).

Clustering

Dimensionality Reduction:

Applied PCA to reduce the dimensions of the dataset for clustering.

KMeans Clustering:

Performed KMeans clustering to categorize transactions into 3 clusters.

Visualized the clusters using a scatter plot.

A diagram of different colored dots

Description automatically generated

Results

The dataset was successfully analysed, and useful insights were derived through EDA. Predictive models were built to estimate item prices, with the Decision Tree Regressor showing the best performance.

Transactions were clustered into 3 distinct groups, providing additional insights into transaction patterns.

Conclusion

This analysis provides a comprehensive understanding of the Wooly Transactions dataset, including key trends, predictive modelling for item prices, and clustering of transactions. These insights can be valuable for business decision-making and further analysis.