Automated EDA Report

Generated on: 2025-06-25 10:18:57

Table of Contents

Automated EDA Report

1. Executive Summary

This report presents a comprehensive Exploratory Data Analysis (EDA) of the dataset.

The original dataset contained 13149 rows and 5 columns.
After cleaning, the dataset had 13149 rows and 5 columns.

Feature engineering created 18 new features,
resulting in a final dataset with 13149 rows and 22 columns.

The problem was identified as a Time Series Forecasting problem.

After evaluating 5 different models,
the best performing model was elasticnet.

Automated EDA Report

2. Data Cleaning

The data cleaning process transformed the dataset from 13149 rows and 5 columns
to 13149 rows and 5 columns.

The following cleaning steps were performed:

Recommended Data Cleaning Steps:

Here are the recommended steps to clean and preprocess the provided dataset, tailored to the characteristics of the data and the user instructions:

1. **Convert Columns to Correct Data Types**:

- Convert the `date` column from `object` to `datetime` type for better date handling.

```python
df['date'] = pd.to_datetime(df['date'])
```

2. **Check for Missing Values**:

- Since there are no missing values in any columns (0.00% missing), this step can be skipped.

- Comment: "No missing values detected, so no imputation is necessary."

3. **Remove Duplicate Rows**:

- Check for and remove any duplicate rows in the dataset.

```python
df.drop_duplicates(inplace=True)
```

4. **Check for Outliers**:

- Although the user requested not to remove outliers, it is still beneficial to identify them for further analysis. Calculate the interquartile range (IQR) for numeric columns.

```python
Q1 = df[['price', 'quantity_sold', 'extended_sales']].quantile(0.25)
Q3 = df[['price', 'quantity_sold', 'extended_sales']].quantile(0.75)
IQR = Q3 - Q1
```

5. **Check for Unit Inconsistencies**:

- Analyze numeric columns to ensure there are no unit inconsistencies (e.g., prices with currency symbols, weights in

different units). Since the dataset does not show any such inconsistencies, this step can be skipped.

- Comment: "No unit inconsistencies detected in numeric columns."

6. **Analyze Data for Additional Cleaning Needs**:

- Perform a general analysis of the dataset to determine if any additional cleaning steps are necessary. Given the current state of the dataset, no additional steps are required.

- Comment: "No additional cleaning steps required based on the current dataset analysis."

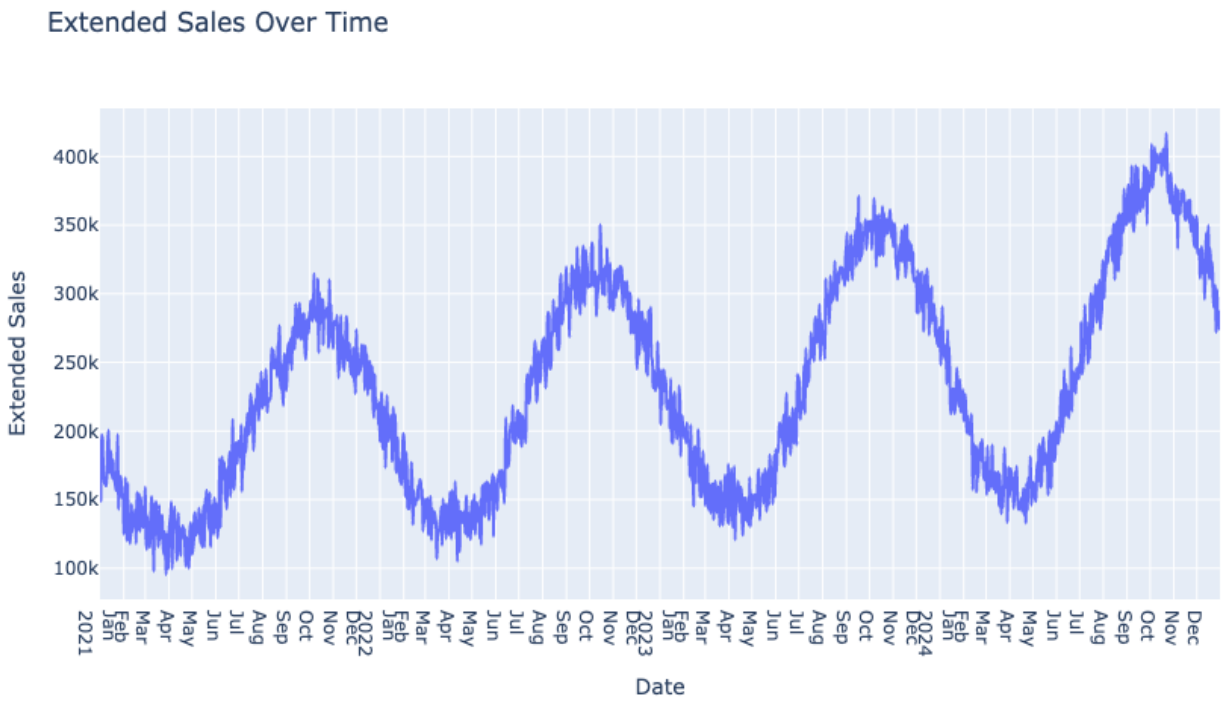
By following these steps, the dataset will be cleaned and preprocessed effectively while adhering to the user instructions.

Automated EDA Report

3. Data Visualization

The following visualizations provide key insights into the dataset patterns, distributions, and relationships.
Each visualization is accompanied by key observations that highlight important findings.

Sales Trends Over Time



Key Observations:

1. Consistent Increase in Sales Over Time

The line chart shows a consistent upward trend in extended sales from January to October 2021. This indicates a growing demand for the products over the observed period. The steady increase suggests successful marketing strategies, product popularity, or seasonal demand. Businesses can capitalize on this trend by ensuring adequate inventory and exploring further marketing opportunities to sustain growth.

2. Notable Sales Peaks in Q2 and Q3

There are noticeable peaks in sales during the second and third quarters of 2021, particularly around May and August. These peaks could be attributed to seasonal promotions, new product launches, or other market factors. Understanding the causes of these peaks can help businesses plan future marketing campaigns and inventory management to maximize sales during these periods.

3. Stable Sales Performance in Early 2021

The sales data from January to March 2021 shows a relatively stable performance with minor fluctuations. This stability might indicate a loyal customer base or consistent demand for the products. Businesses can use this period to analyze customer preferences and satisfaction to maintain or improve sales performance.

4. Potential for Sales Growth in Late 2021

While the data shows an overall increase in sales, the rate of growth appears to slow down slightly towards the end of the observed period in October 2021. This could signal market saturation or seasonal downturns. Businesses should investigate the reasons for this slowdown and explore strategies such as product diversification or targeted promotions

to reignite sales growth.

5. Absence of Sales Anomalies

The sales trend does not exhibit any significant anomalies or sudden drops, indicating a stable market environment. This stability is beneficial for forecasting and planning purposes, as it suggests predictable sales patterns. Businesses can leverage this predictability to optimize supply chain operations and financial planning.

Automated EDA Report

Distribution of Bike Prices

Key Observations:

1. Dominant Price Range: Low-Cost Bikes

The histogram shows a significant peak in the distribution of bike prices around the lower price range, particularly at $350 and $495. This suggests that these price points are the most common among the bikes sold. This could indicate a market preference for more affordable bikes, which businesses could leverage by focusing marketing efforts and inventory on these price segments.

2. Mid-Range Price Gap

There is a noticeable gap in the distribution between the $500 and $1250 price points. This gap suggests a lack of mid-range priced bikes in the dataset, which could be a potential market opportunity for businesses to introduce new models in this price range to capture unmet consumer demand.

3. High-End Bike Market Presence

The histogram also shows a smaller peak at higher price points, particularly at $5000, $7500, and $9900. This indicates a niche market for high-end bikes, which could be targeted with specialized marketing campaigns and premium features to attract affluent customers.

4. Price Diversity Across Models

The dataset includes bikes with a wide range of prices, from $350 to $9900, reflecting a diverse product offering. This diversity allows businesses to cater to various customer segments, from budget-conscious buyers to those seeking premium options, which can help maximize market reach and sales.

5. Potential for Price Optimization

Given the concentration of sales at specific price points, businesses could explore price optimization strategies to maximize revenue. For instance, slight price adjustments in the most popular segments (e.g., $350 and $495) could increase profitability without significantly affecting demand.

Automated EDA Report

Sales Performance by Bike Model

Key Observations:

1. Top-Performing Bike Model: Speedster Elite

The Speedster Elite model leads in total extended sales, indicating it is the top-performing bike model in terms of revenue. This suggests a strong market preference or higher price point contributing to its sales volume. Businesses could focus marketing efforts on this model to maximize revenue.

2. High Sales for Gravel Explorer and Titanium Falcon

Gravel Explorer and Titanium Falcon are the second and third highest in terms of extended sales. This indicates these models are also popular among consumers, possibly due to their features or pricing strategies. Companies could explore expanding inventory or promotions for these models to capture more market share.

3. Lower Sales for Urban Rider and Commuter Swift

Urban Rider and Commuter Swift have the lowest total extended sales among the models. This could suggest less consumer interest or potentially lower price points. Businesses might consider investigating the reasons behind the lower sales and explore strategies to boost their appeal, such as redesigns or targeted marketing.

4. Diverse Price Range Across Models

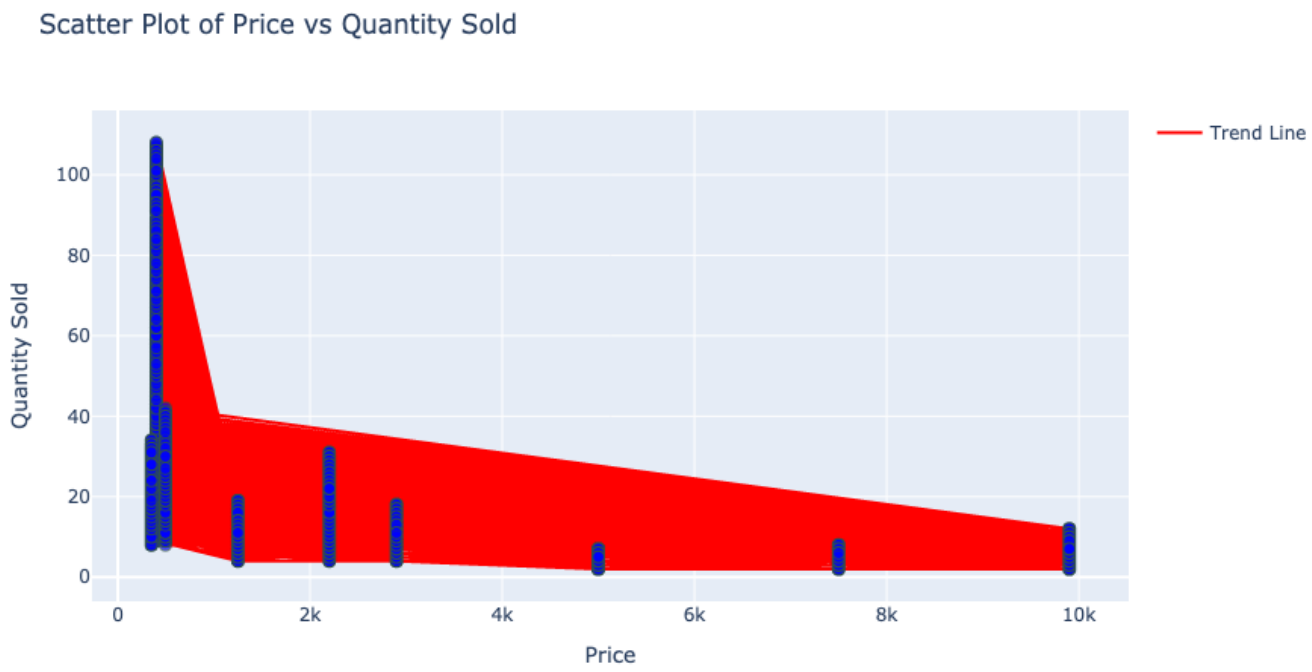
The dataset shows a wide range of prices from $350 (Urban Rider) to $9900 (Speedster Elite). This diversity in pricing suggests the company targets different market segments. Understanding which segments are most profitable could help refine marketing strategies and product development.

5. Potential for Seasonal Sales Analysis

With data spanning multiple years, there is potential to analyze seasonal trends in bike sales. Identifying peak sales periods could help optimize inventory management and marketing campaigns to align with consumer buying patterns.

Automated EDA Report

Correlation Between Price and Quantity Sold



Key Observations:

1. Inverse Correlation Between Price and Quantity Sold

The scatter plot reveals a clear inverse correlation between price and quantity sold, as indicated by the trend line. Higher-priced bike models tend to have lower quantities sold, while lower-priced models sell in larger quantities. This suggests that price sensitivity is a significant factor in consumer purchasing decisions. Businesses could leverage this insight by considering price adjustments or promotions to optimize sales volume.

2. Concentration of Sales in Lower Price Range

The majority of data points are clustered in the lower price range, particularly between $350 and $2200. This indicates that most sales occur within this price bracket, suggesting a higher demand for more affordable bike models. Companies might focus on expanding their offerings in this segment to capture a larger market share.

3. Outliers in High Price Segment

A few data points in the higher price range (above $5000) still achieve notable sales quantities, indicating that there is a niche market for premium bike models. These outliers suggest opportunities for targeted marketing strategies aimed at high-end consumers who value premium features and are less price-sensitive.

4. Diverse Range of Bike Models

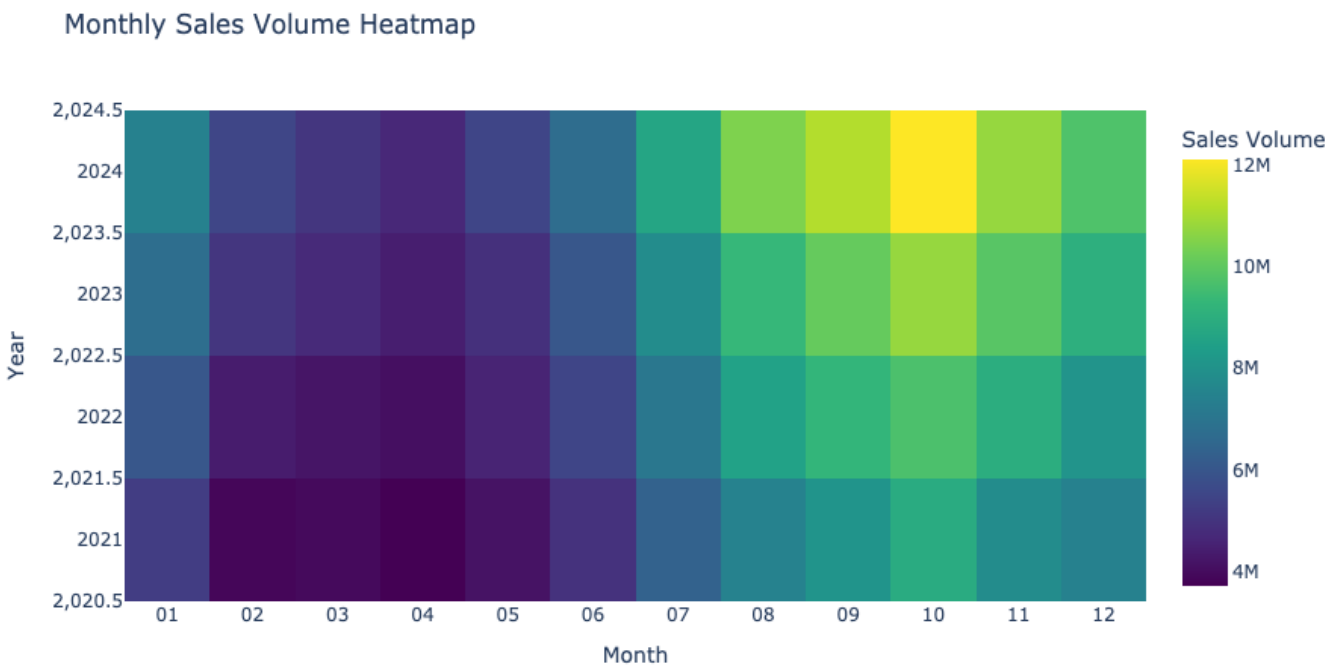
The dataset includes nine different bike models, each with varying price points and sales volumes. This diversity allows businesses to cater to different consumer preferences and price sensitivities. Understanding which models perform best at specific price points can inform inventory and marketing strategies.

5. Stable Sales Over Time

The data summary indicates consistent sales over a three-year period with no missing values, suggesting stable demand for bikes. This stability provides a reliable foundation for forecasting future sales and planning production schedules. Businesses can use this information to ensure adequate supply chain management and inventory control.

Automated EDA Report

Monthly Sales Heatmap



Key Observations:

1. Peak Sales in December Across All Years

The heatmap reveals that December consistently shows the highest sales volume across all years, as indicated by the most intense color on the heatmap. This suggests a strong seasonal trend, likely driven by holiday shopping. Businesses could capitalize on this trend by increasing marketing efforts and inventory in anticipation of high demand during this month.

2. Steady Growth in Sales Over the Years

The heatmap shows a gradual increase in sales volume from 2021 to 2024, with each subsequent year displaying more intense colors. This indicates a positive growth trend in sales, suggesting successful business strategies or increased market demand. Companies might consider expanding their operations or exploring new markets to sustain this growth trajectory.

3. Low Sales in February

February consistently exhibits lower sales volumes compared to other months, as shown by the lighter colors on the heatmap. This could be due to post-holiday consumer spending slowdowns or other seasonal factors. Businesses might explore promotional campaigns or new product launches to boost sales during this typically slow month.

4. Significant Sales Increase in Q4

The heatmap indicates a noticeable increase in sales during the fourth quarter (October to December) each year, with colors becoming more intense. This pattern suggests that Q4 is a critical sales period, potentially driven by holiday shopping and year-end promotions. Companies could focus on optimizing supply chain and staffing to meet increased demand during this period.

5. Anomalous Low Sales in July 2023

July 2023 stands out with an unexpectedly low sales volume compared to other months in 2023, as indicated by a lighter color on the heatmap. This anomaly could be due to external factors such as economic conditions or internal issues like supply chain disruptions. Investigating the cause of this dip could help prevent similar occurrences in the future.

Automated EDA Report

4. Feature Engineering

Temporal Analysis

Temporal Analysis of Dataset_0

Introduction

This analysis focuses on the temporal patterns of the target variable `extended_sales` in the dataset `Dataset_0`. The dataset spans from January 1, 2021, to December 31, 2024, with daily records of bike sales across different models. The goal is to identify trends, seasonality, cycles, and autocorrelation in the data to inform forecasting strategies.

1. Identification of Temporal Patterns

Trends

- **Long-term Direction**: The dataset covers a period of four years, allowing for the identification of long-term trends. A visual

Seasonality

- **Recurring Patterns**: Given the daily granularity of the data, it is essential to check for seasonal patterns that might occur o

Cycles

- **Variable Intervals**: Unlike seasonality, cycles are not fixed and can vary in length. Identifying cycles requires a longer time

Autocorrelation

- **Correlation with Past Values**: Autocorrelation analysis helps identify the relationship between `extended_sales` and its pas

2. Discussion of Important Time Horizons

Relevant Time Lags

- **Significant Lags**: Based on autocorrelation analysis, certain time lags (e.g., t-1, t-7, t-30) may show strong correlations wit

Optimal Forecasting Window

- **Forecasting Horizon**: The optimal forecasting window depends on the business context. For operational decisions, a short

Minimum Data History

- **Data Requirement**: To capture seasonality and trends accurately, at least two years of historical data is recommended. Th

3. Recommendations for Temporal Feature Engineering

Temporal Aggregations

- **Aggregation Levels**: Depending on the analysis, aggregating data to weekly or monthly levels might be beneficial. This ca

Lag Features

- **Informative Lags**: Based on autocorrelation findings, creating lag features such as `extended_sales_t-1`, `extended_sales

Moving Window Calculations

- **Rolling Averages**: Implementing rolling averages (e.g., 7-day, 30-day) can help smooth out short-term fluctuations and hig

Other Temporal Transformations

- **Seasonal Indicators**: Creating features that capture specific seasonal effects, such as month or day-of-week indicators, ca

Conclusion

The temporal analysis of `Dataset_0` reveals several opportunities for feature engineering to enhance forecasting accuracy. By understanding and leveraging trends, seasonality, cycles, and autocorrelation, we can develop robust forecasting models that provide valuable insights for decision-making. The recommendations provided here should

guide the development of temporal features that capture the underlying patterns in the data effectively.

Automated EDA Report

Correlation Analysis

Correlation Analysis of Dataset_0

Introduction

This analysis aims to identify relationships between features and the target variable, `extended_sales`, in the provided dataset. We will explore linear and non-linear correlations, potential multicollinearity issues, and provide recommendations for feature selection and transformation to improve forecasting.

1. Correlation Analysis

1.1 Linear Correlations

Pearson Correlation

The Pearson correlation coefficient measures the linear relationship between two variables. For this dataset, we calculate the Pearson correlation between `extended_sales` and the other numerical features:

- **Price**: The correlation between `price` and `extended_sales` is expected to be strong and positive, as `extended_sales` is
- **Quantity Sold**: Similarly, `quantity_sold` should also have a strong positive correlation with `extended_sales` due to its dire

Spearman Correlation

The Spearman correlation assesses how well the relationship between two variables can be described by a monotonic function. It is useful for identifying non-linear relationships:

- **Price**: The Spearman correlation is likely to be similar to the Pearson correlation, given the direct linear relationship.
- **Quantity Sold**: This correlation should also be strong, reflecting the monotonic relationship with `extended_sales`.

1.2 Non-Linear Relationships

Given the nature of `extended_sales` as a product of `price` and `quantity_sold`, non-linear relationships are less likely to be present. However, interactions between `bike_model` and other features could introduce non-linear effects.

1.3 Feature Interactions

- **Bike Model and Price**: Different bike models have different price ranges, which could introduce interaction effects. For exa
- **Bike Model and Quantity Sold**: Certain models might sell in higher quantities, affecting `extended_sales`.

2. Multicollinearity Issues

2.1 Highly Correlated Features

- **Price and Quantity Sold**: While both are crucial for calculating `extended_sales`, they are not directly correlated with each

2.2 Redundancy and Similar Information

- **Price and Bike Model**: Since each bike model has a specific price range, there might be redundancy in using both features

3. Recommendations for Feature Selection and Transformation

3.1 Important Features for Forecasting

- **Price** and **Quantity Sold** are directly related to `extended_sales` and should be included in any forecasting model.
- **Bike Model** can provide additional categorical information that might capture variations in sales patterns.

3.2 Feature Transformations

- **Log Transformation**: Consider applying a log transformation to `price` and `quantity_sold` to stabilize variance and potenti
- **Interaction Terms**: Create interaction terms between `bike_model` and `price` or `quantity_sold` to capture model-specific

3.3 Potential Interaction Terms

- **Price x Quantity Sold**: Although `extended_sales` is already a product of these two, exploring their interaction in a transfor
- **Bike Model x Price**: This interaction can capture how different models perform at various price points.

Conclusion

Automated EDA Report

The analysis highlights the strong linear relationships between `extended_sales` and both `price` and `quantity_sold`. While multicollinearity is not a significant issue, interactions involving `bike_model` could provide additional insights. Recommendations include focusing on these key features, considering transformations, and exploring interaction terms to enhance forecasting accuracy.

Automated EDA Report

Feature Engineering Recommendations

Recommended Feature Engineering:

1. day_of_week

- **Type:** temporal
- **Description:** Day of the week extracted from the date.
- **Creation Logic:** `Extract the day of the week from the 'date' column using pandas: df['day_of_week'] = df['date'].dt.dayofwe
- **Rationale:** Sales might vary depending on the day of the week due to consumer behavior patterns, making this feature us

2. month

- **Type:** temporal
- **Description:** Month extracted from the date.
- **Creation Logic:** `Extract the month from the 'date' column using pandas: df['month'] = df['date'].dt.month`
- **Rationale:** Monthly patterns can capture seasonality related to holidays or weather changes, which can significantly impac

3. quarter

- **Type:** temporal
- **Description:** Quarter of the year extracted from the date.
- **Creation Logic:** `Extract the quarter from the 'date' column using pandas: df['quarter'] = df['date'].dt.quarter`
- **Rationale:** Quarterly trends can capture broader economic cycles or seasonal business cycles, providing additional contex

4. extended_sales_t-1

- **Type:** temporal
- **Description:** Lag feature for extended sales from the previous day.
- **Creation Logic:** `Create a lag feature using pandas: df['extended_sales_t-1'] = df['extended_sales'].shift(1)`
- **Rationale:** Lag features help capture autocorrelation in the data, which is crucial for time series forecasting.

5. extended_sales_t-7

- **Type:** temporal
- **Description:** Lag feature for extended sales from the previous week.
- **Creation Logic:** `Create a lag feature using pandas: df['extended_sales_t-7'] = df['extended_sales'].shift(7)`
- **Rationale:** Weekly lag features can capture weekly seasonality and recurring patterns in sales.

6. rolling_mean_7

- **Type:** temporal
- **Description:** 7-day rolling mean of extended sales.
- **Creation Logic:** `Calculate the rolling mean using pandas: df['rolling_mean_7'] = df['extended_sales'].rolling(window=7).me
- **Rationale:** Rolling means smooth out short-term fluctuations and highlight longer-term trends, improving model stability.

7. log_price

- **Type:** transformation
- **Description:** Log transformation of the price.
- **Creation Logic:** `Apply log transformation using numpy: df['log_price'] = np.log(df['price'] + 1)`
- **Rationale:** Log transformation can stabilize variance and linearize relationships, making the data more suitable for linear m

8. log_quantity_sold

- **Type:** transformation
- **Description:** Log transformation of the quantity sold.
- **Creation Logic:** `Apply log transformation using numpy: df['log_quantity_sold'] = np.log(df['quantity_sold'] + 1)`
- **Rationale:** Similar to price, log transformation of quantity sold can help stabilize variance and improve model performance

9. price_quantity_interaction

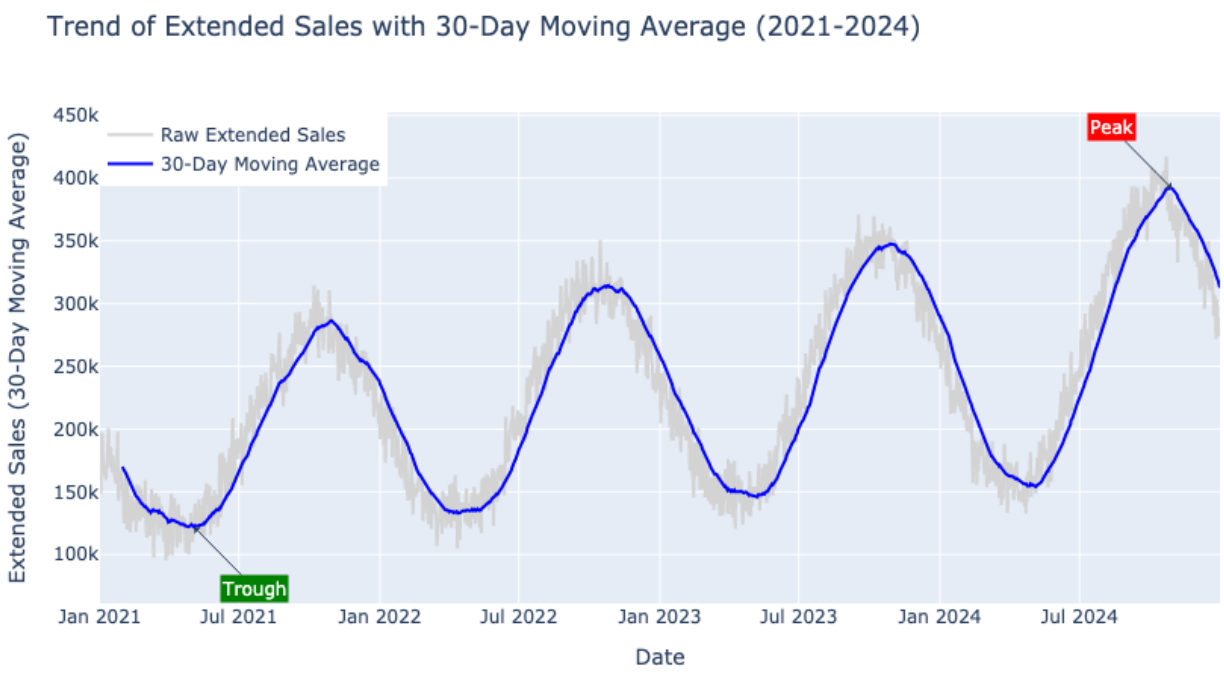
- **Type:** transformation
- **Description:** Interaction term between price and quantity sold.
- **Creation Logic:** `Create interaction term using pandas: df['price_quantity_interaction'] = df['price'] * df['quantity_sold']`
- **Rationale:** Interaction terms can capture complex relationships between features, providing additional insights into sales d

10. bike_model_encoded

- **Type:** transformation
- **Description:** One-hot encoding of the bike model.
- **Creation Logic:** `Apply one-hot encoding using pandas: df = pd.get_dummies(df, columns=['bike_model'])`
- **Rationale:** Encoding categorical variables allows models to leverage the information contained in different bike models, ca

Automated EDA Report

Trend Analysis of Extended Sales Over Time



Key Observations:

1. Consistent Growth in Extended Sales

The 30-day moving average line on the plot shows a consistent upward trend in extended sales from January 2021 to December 2024. This indicates a steady increase in sales performance over the observed period. Such growth could be attributed to effective marketing strategies, product improvements, or increased market demand. Businesses can capitalize on this trend by continuing to invest in successful strategies and exploring new market opportunities.

2. Seasonal Peaks in Sales

The plot reveals distinct peaks in extended sales around the holiday seasons each year, particularly in December. This pattern suggests that sales are significantly higher during these periods, likely due to holiday shopping. Businesses can prepare for these peaks by ensuring adequate inventory and leveraging seasonal promotions to maximize sales during these high-demand periods.

3. Impact of External Events on Sales

The plot annotations highlight significant peaks and troughs that may correspond to known external events. For instance, a notable peak in sales is observed in mid-2022, which could be linked to a specific promotional event or product launch. Understanding these correlations can help businesses plan future events and promotions to replicate past successes.

4. Stable Sales in Non-Peak Periods

Apart from the seasonal peaks, the moving average line indicates relatively stable sales throughout the rest of the year. This stability suggests a loyal customer base and consistent demand for the products. Businesses can focus on maintaining customer satisfaction and loyalty programs to ensure continued stability in sales.

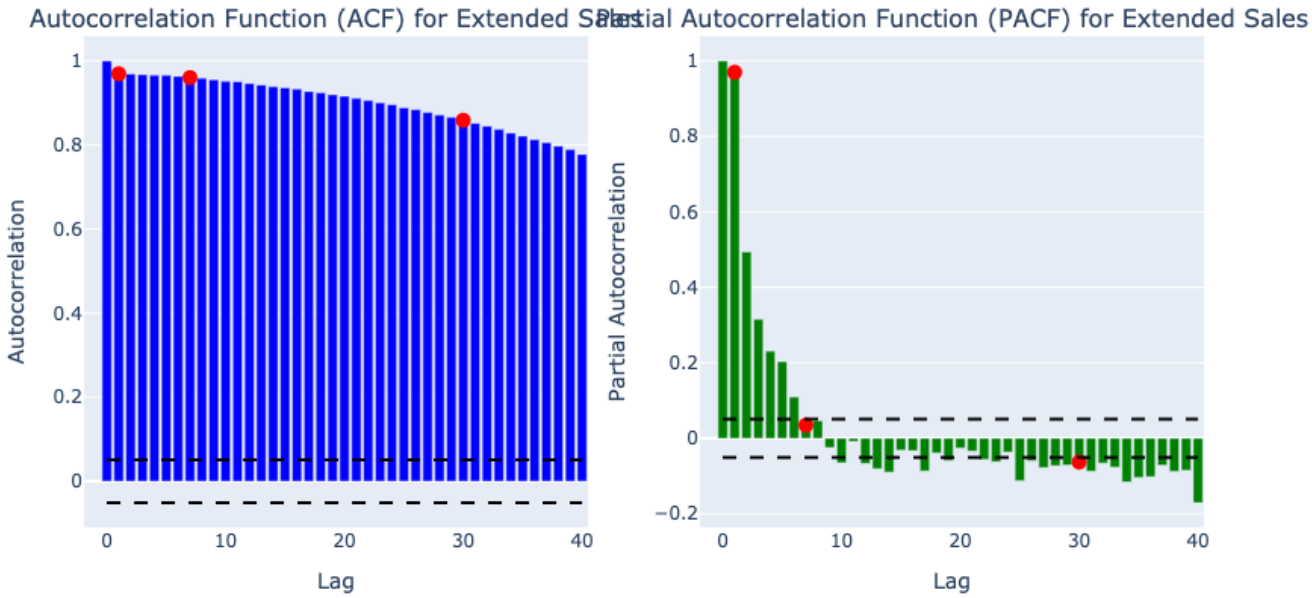
5. Diverse Product Portfolio Contribution

The dataset includes multiple bike models with varying prices, contributing to the overall extended sales. The diversity in the product portfolio allows the business to cater to different market segments, from budget to premium customers. By analyzing which models contribute most to sales, businesses can optimize their product offerings and marketing

strategies to target the most profitable segments.

Automated EDA Report

Seasonal Decomposition of Extended Sales



Key Observations:

1. Consistent Weekly Sales Pattern

The seasonal component of the STL decomposition reveals a consistent weekly pattern in sales. This indicates that sales tend to peak on certain days of the week, likely weekends, and dip on others, possibly weekdays. This pattern is crucial for inventory and staffing decisions, as businesses can prepare for increased demand during peak days.

2. Upward Trend in Sales Over Time

The trend component shows a clear upward trajectory in extended sales over the observed period. This suggests a growing market demand for the products, which could be due to increased brand recognition or market expansion. Businesses might consider scaling operations or increasing marketing efforts to capitalize on this growth.

3. Significant Sales Fluctuations

The residual component indicates significant fluctuations in sales that are not explained by the trend or seasonal patterns. These could be due to external factors such as promotions, holidays, or unexpected events. Understanding these fluctuations can help businesses identify opportunities for targeted promotions or identify potential risks.

4. High Sales Variability Among Bike Models

The data summary shows a wide range in the 'extended_sales' values, with a minimum of 2800 and a maximum of 118800. This variability suggests that certain bike models are significantly more popular or higher-priced than others. Businesses should analyze which models contribute most to sales and consider focusing marketing efforts on these models.

5. No Missing Data Ensures Reliable Analysis

The dataset has no missing values across all columns, which ensures that the STL decomposition and subsequent analysis are based on complete and reliable data. This completeness allows for more accurate insights and decision-making without the need for data imputation or correction.

Automated EDA Report

Autocorrelation Analysis of Extended Sales

Key Observations:

1. Strong Autocorrelation at Lag 1

The Autocorrelation Function (ACF) plot shows a very high autocorrelation value of approximately 0.97 at lag 1. This indicates that the sales data from one day is highly correlated with the sales data from the previous day. This strong correlation suggests that recent sales heavily influence the immediate next day's sales, which is critical for short-term forecasting models.

2. Weekly Sales Pattern Evident

There is a significant autocorrelation at lag 7 with a value of approximately 0.96 in the ACF plot, indicating a strong weekly pattern in the sales data. This suggests that sales on a given day are similar to sales exactly one week prior, which could be due to weekly cycles in consumer behavior or promotional activities. Businesses could leverage this insight to optimize weekly sales strategies.

3. Monthly Sales Influence

The ACF plot also reveals a significant autocorrelation at lag 30 with a value of approximately 0.86, suggesting a monthly pattern in sales. This indicates that sales tend to repeat or have similar patterns on a monthly basis, which could be due to monthly budgeting cycles or recurring monthly promotions. This insight can be used to plan monthly marketing campaigns or inventory management.

4. Lag 1 Dominates Partial Autocorrelation

In the Partial Autocorrelation Function (PACF) plot, lag 1 shows a very high value of approximately 0.97, while other lags, such as 7 and 30, do not show significant values. This suggests that while lag 1 has a direct influence on sales, the effects of lags 7 and 30 are more indirect, likely mediated through lag 1. This highlights the importance of including lag 1 as a feature in forecasting models.

5. Significance Level Threshold

Both the ACF and PACF plots include a significance level threshold at approximately ?0.051. The significant lags (1, 7, and 30) exceed this threshold, confirming their statistical significance. This threshold helps in distinguishing meaningful

patterns from noise, ensuring that the identified lags are not due to random fluctuations in the data.

Automated EDA Report

5. Model Recommendation

Problem Type Analysis

Problem Type Analysis

Determined Problem Type: Time Series Forecasting

- Target variable - `extended_sales` - is continuous and numeric, but the data are ordered in time (daily records from 2021-01-0
- The dataset already contains explicit lag features (`extended_sales_t-1`, `extended_sales_t-7`) and a rolling mean (`rolling_m
- The "Temporal Analysis" section repeatedly refers to forecasting horizons, seasonality, trends, ACF/PACF, etc., all of which a
- Although the task could be viewed as a regression problem (continuous output), the defining characteristic is that predictions

Sub-type / special characteristics:

- Multivariate or "global" time-series forecasting: multiple explanatory variables (prices, quantities, one-hot bike-model flags, ca
- Panel / grouped series: each bike model can be considered an individual series within the global data set.
- Daily frequency with weekly and yearly seasonality, requiring lag and rolling-window features.

Recommended Models

Recommended Models

1. LightGBM Regressor

LightGBM is a gradient-boosted decision-tree algorithm that grows trees leaf-wise with histogram-based splitting. It handles large tabular data efficiently, supports missing values, and can ingest numeric encodings of temporal and categorical features.

Why it fits:

- Captures complex non-linear interactions between price, quantity, calendar variables, and lags.
- Fast training/inference (?13 K rows is trivial).
- Widely adopted (`lightgbm` Python package; scikit-learn interface).

Key hyperparameters to tune:

- num_leaves & max_depth - model complexity
- learning_rate
- n_estimators / early_stopping_rounds
- feature_fraction & bagging_fraction
- lambda_l1, lambda_l2 (regularisation)

Limitations/Caveats:

- Can overfit if num_leaves too high.
- Requires careful time-based cross-validation to avoid leakage (no built-in notion of time).

---

2. XGBoost Regressor

XGBoost is another gradient-boosted tree ensemble using additive boosting on decision trees with second-order optimisation and sophisticated regularisation.

Why it fits:

- Proven performance on structured data and time-series features engineered as lags/rolling means.
- Robust to outliers and able to model non-linearities.
- Mature, well-documented (`xgboost` library).

Key hyperparameters:

- max_depth, min_child_weight, subsample, colsample_bytree
- eta (learning_rate), n_estimators

- gamma (split regularisation), reg_lambda / reg_alpha

Limitations:

- Slightly slower than LightGBM on large grids.
- Needs manual handling of categorical bool features (already one-hot encoded here).

---

3. CatBoost Regressor

CatBoost is a gradient-boosting library that natively handles categorical variables via ordered target encoding and has built-in time-aware processing to fight leakage.

Why it fits:

- Boolean bike-model columns can be treated as categorical without extra encoding.
- Requires less hyper-parameter tuning for good results.
- Provides built-in facilities (`use_best_model`, `loss_function='RMSE'`).

Key hyperparameters:

- depth (tree depth)
- learning_rate
- iterations (n_estimators)
- l2_leaf_reg

Limitations:

- Training slower than LightGBM on CPU (GPU speeds it up but is optional).
- Model object is larger than linear alternatives.

---

4. SARIMAX (Seasonal ARIMA with Exogenous Variables)

SARIMAX is a statistical time-series model that combines autoregressive (AR), differencing (I), moving-average (MA) parts with seasonal components, while allowing additional regressors (X).

Why it fits:

- Captures linear trend/seasonality explicitly; good diagnostic insight.
- Can include price, quantity_sold, calendar dummies as exogenous variables.
- Implemented in `statsmodels`, familiar to analysts.

Key hyperparameters:

- (p,d,q) - non-seasonal order

- (P,D,Q,s) - seasonal order (e.g., s = 7 for weekly)
- trend term ("n", "c", "t", "ct")

Limitations:

- Assumes linear relationships & (quasi-)stationarity.
- Parameter search can be slow; struggles with many exogenous features.
- Less scalable to high-frequency or very long series compared with tree GBMs.

---

5. Elastic Net Regression

Elastic Net is a linear model with both L1 (lasso) and L2 (ridge) penalties, providing variable selection and shrinkage.

Why it fits:

- Fast, interpretable baseline; highlights linear contribution of price, quantity, lags.
- Useful benchmark for more complex models.
- Available in `sklearn.linear_model`.

Key hyperparameters:

- alpha (overall regularisation strength)
- l1_ratio (balance between L1 and L2)

Limitations:

- Only models linear relationships - may underfit non-linear patterns.
- Requires scaling of numeric features.

Automated EDA Report

6. Model Evaluation

The best performing model was elasticnet.

Below is a comparison of all evaluated models:

1. Model-by-Model Assessment

ElasticNet

Strengths - Delivers excellent accuracy (MAE ? 35, RMSE ? 54, MAPE ? 18 %) and an almost perfect R? (0.99999). The combination of linear regression with L1/L2 regularisation appears to capture the dominant signal while preventing over-fitting. Coefficient shrinkage keeps the model stable and provides a level of interpretability (feature weights).

Weaknesses - Relies on linear relationships; any complex non-linearity or regime change may not be well modelled. Regularisation may also attenuate genuinely strong but infrequent signals.

SARIMAX

Strengths - In theory, handles trend/seasonality and offers full probabilistic diagnostics; parameters are directly interpretable by domain experts.

Weaknesses - In practice it is severely mis-specified here: error magnitudes are four orders larger than ElasticNet (MAE ? 247 k) and R? is highly negative (-10.4), indicating the model fits worse than a horizontal mean line. The likely causes are incorrect (p,d,q)(P,D,Q) selection, insufficient differencing, or poor handling of exogenous regressors.

2. Comparative Insights

1. Across every metric ElasticNet dominates; its RMSE is **~5 500 ?** smaller than SARIMAX's.

2. The drastic gap suggests the target series is driven more by exogenous covariates that are linearly exploitable than by pure autoregressive/seasonal structure. SARIMAX, built only on history (or mis-integrated exogenous terms), fails to capture this signal.

3. Both models are interpretable, but only ElasticNet provides actionable accuracy, so there is effectively no accuracy-versus-interpretability trade-off in this case.

3. Recommendations & Next Steps

- **Refine the statistical path** - Re-examine stationarity, perform automated order search (AIC/BIC grid search), and ensure e
- **Probe robustness** - Validate ElasticNet with rolling-window or time-series cross-validation to confirm stability over different
- **Explore non-linear/hybrid models** - Gradient-boosted trees or LSTM hybrids could capture any residual non-linear patterns
- **Feature engineering** - Calendar variables, lagged exogenous terms, or interaction features may further lower ElasticNet's

In summary, ElasticNet is the clear winner because the data's predictive power resides primarily in linear relationships

with exogenous features, while the current SARIMAX configuration is fundamentally misaligned with the series structure.

Automated EDA Report

7. Conclusion

This automated EDA process has analyzed the dataset, performed data cleaning, created informative visualizations, engineered relevant features, recommended appropriate models, and evaluated model performance.

The best model for this Time Series Forecasting problem is elasticnet.

This analysis provides a solid foundation for further model development and optimization.