

Chapter 3: Methodology

by Fuyuan Cai

Submission date: 24-Jul-2024 06:20PM (UTC-0700)

Submission ID: 2422035210

File name: cypypaste.txt (2.85K)

Word count: 379

Character count: 2516

2

3.1 Research design

This study uses a quantitative research design to analyze figures and data visualization to produce results to give recommendations. Quantitative methods are well suited to the application of this study as it involves analyzing large data sets to identify patterns, trends and relationships in supply chain operations. Specific methods include descriptive statistics, predictive modeling and machine learning techniques.

3.2 Data collection and pre-processing

Data source: supply store dataset from Github

<https://github.com/drshahizan/dataset/tree/main/mongodb/01-sales> Data cleaning: data was imported and cleaned to ensure accuracy and completeness, resolve missing values, inconsistencies, and outliers, standardize date formats and ensure consistency of categorical data, and create relevant variables for analysis.

data

3.3 Exploratory Data Analysis (EDA)

Descriptive statistics: use Python libraries such as Pandas and NumPy to calculate metrics such as mean, median, standard deviation, and distribution for key variables (e.g., sales, volume, customer satisfaction). Visualization: Use histograms, bar charts, and box-and-line plots using Matplotlib and Seaborn to visualize other key metrics such as sales trends, seasonal patterns, product popularity, and customer demographics. Correlation analysis: the Pearson correlation coefficient is used to

identify relationships between variables such as sales volume, customer satisfaction and buying patterns to inform the predictive model.

3.4 Forecasting Modeling

Demand forecasting: time series analysis using ARIMA⁴ and exponential smoothing models to forecast future sales based on historical data. These analyses are performed using Python's statsmodels library. Inventory optimization: machine learning algorithms such as Random Forest and XGBoost are applied to recommend optimal inventory levels and reorder points. Scikit-learn and XGBoost libraries are used. Customer Segmentation : Cluster analysis using K-means or hierarchical clustering to develop customer segmentation models by clustering customers based on purchase behavior and demographics. Scikit-learn library is used for clustering.

3.5 Model evaluation and validation

³ Cross-validation: k-fold cross-validation is used to assess the model performance to ensure robustness and prevent overfitting.¹ Evaluation metrics: metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Accuracy are used to evaluate model predictions. These are calculated using Python⁷ libraries such as Scikit-learn and Statsmodels.

3.6 Software and analytic frameworks

Python: the main programming language used for data analysis, containing⁵ libraries such as Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, Statsmodels and XGBoost.

Jupyter Notebooks: for interactive data analysis and visualization. Power BI: for data visualization of analysis results.

Chapter 3: Methodology

ORIGINALITY REPORT

17%
SIMILARITY INDEX

16%
INTERNET SOURCES

4%
PUBLICATIONS

0%
STUDENT PAPERS

PRIMARY SOURCES

1	ejournal.seminar-id.com Internet Source	3%
2	www.researchgate.net Internet Source	3%
3	ia601805.us.archive.org Internet Source	2%
4	www.fastercapital.com Internet Source	2%
5	www.ijraset.com Internet Source	2%
6	businessyield.com Internet Source	2%
7	Ton Duc Thang University Publication	1%

Exclude quotes On
Exclude bibliography On

Exclude matches Off