**Title**

**Exploring Factors Affecting Walmart Sales**

**Summary**

The purpose of this paper is to forecast Walmart sales by considering several variables, including temperature, fuel costs, the consumer price index (CPI), unemployment, and holidays. Historical weekly sales data from multiple Walmart locations is included in the Kaggle dataset. Predictive modeling techniques are used in the project to examine how various factors affect sales performance. The study uses machine learning models including XGBOOST, DECISION TREES, KNN, random forest, and linear regression to find important trends and offer suggestions. To improve future sales forecasts, the model that best fits Walmart's sales data is being considered.

**Introduction**

Specifically, this study falls under the Predictive Analytics and Machine Learning subfield of Data Science and Applied Analytics. Because of its emphasis on forecasting sales in the retail sector using machine learning and historical data, it also has connections to business analytics and retail analytics.

For big stores like Walmart, retail sales forecasting is essential to their success since it allows for improved inventory control, demand forecasting, and promotional strategy optimization. Effective sales forecasting can result in better decision-making and more effective use of resources given Walmart's size and varied clientele. Several external factors impact sales, such as market pressures like fuel prices, environmental elements like temperature, and economic indices like the Consumer Price Index (CPI) and unemployment rates. Holidays and other special occasions can sometimes cause notable increases in sales, which makes forecasting both difficult and crucial for the long-term viability of the business.

This study expands on earlier research, including that conducted by Dr. Lutful Islam et al. (Islam, M. B., & Islam, K. S. (2023)), which looked at machine learning models for retail sales prediction. Using comparable methods, this study attempts to forecast Walmart's weekly sales using a Kaggle dataset that include variables including temperature, fuel costs, CPI, unemployment, and holiday occurrence. The study aims to analyze the efficacy of several predictive models, such as random forest and linear regression, and assess how these characteristics affect sales. The results of this investigation will contribute to the larger field of retail analytics and predictive modeling by shedding light on important sales drivers and making suggestions for better forecasting techniques.

**Theoretical Foundations**

**Data Structure**

**Store**:

* Type: Numeric
* Description: Identifier for each Walmart store. There are 45 unique stores in the dataset.

**Weekly\_Sales**:

* Type: Numeric (Continuous)
* Description: The total sales for the given week for the store. This is the target variable for prediction.

**Holiday\_Flag**:

* Type: Categorical (Binary)
* Description: A binary indicator (0 or 1) to mark whether a given week had a holiday (1) or not (0). This variable captures the impact of holidays on sales performance.

**Temperature**:

* Type: Numeric (Continuous)
* Description: The average temperature during the week in the store’s location, measured in degrees Fahrenheit. Weather can have a direct effect on consumer behavior.

**Fuel\_Price**:

* Type: Numeric (Continuous)
* Description: The average fuel price during the week. Fuel prices are a critical economic factor affecting consumer purchasing behavior.

**CPI (Consumer Price Index)**:

* Type: Numeric (Continuous)
* Description: The CPI for the week, representing inflationary trends in the economy. CPI is often used to gauge the economic environment and purchasing power.

**Unemployment**:

* Type: Numeric (Continuous)
* Description: The unemployment rate in the region of each store. This economic indicator can have a negative correlation with sales, as higher unemployment often leads to reduced consumer spending.

**Number of Observations:**

The dataset contains **6435 rows** of data, each corresponding to a particular week's sales at one of the 45 stores. These records span from 201o - 2012, capturing the temporal fluctuations and trends in sales data.

**Descriptive Statistics:**

The key variables and their respective descriptive statistics are as follows:

1. **Store**: The store variable has a mean value of 23, with a minimum value of 1 and a maximum value of 45, indicating the dataset covers 45 different Walmart stores.
2. **Weekly\_Sales**: The weekly sales data ranges from a minimum of 209,986 to a maximum of 3,818,686, with an average weekly sale of 1,046,965. The standard deviation for weekly sales is 564,366, reflecting significant variation in sales across the stores. The 25th percentile is 553,350, the median is 960,746, and the 75th percentile is 1,420,159.
3. **Holiday\_Flag**: This binary variable has a mean of 0.0699, with values of 0 (non-holiday weeks) and 1 (holiday weeks). This suggests that holidays make up approximately 7% of the total weeks in the dataset.
4. **Temperature**: The temperature variable has an average of 60.66°F, with a standard deviation of 18.44°F, indicating considerable variation in temperatures. The minimum temperature recorded is -2.06°F, while the maximum is 100.14°F. The 25th percentile is 47.46°F, the median is 62.67°F, and the 75th percentile is 74.94°F.
5. **Fuel\_Price**: The fuel price ranges from a minimum of 2.47 USD to a maximum of 4.47 USD, with an average price of 3.36 USD. The standard deviation is 0.46 USD, indicating moderate fluctuation in fuel prices. The 25th percentile is 2.93 USD, the median is 3.45 USD, and the 75th percentile is 3.74 USD.
6. **CPI (Consumer Price Index)**: The CPI has an average value of 171.58, with a minimum of 126.06 and a maximum of 227.23. The standard deviation is 39.36, showing significant variation in CPI values across the dataset. The 25th percentile is 131.74, the median is 182.62, and the 75th percentile is 212.74.
7. **Unemployment**: The unemployment rate has a mean value of 7.99%, with a minimum of 3.88% and a maximum of 14.31%. The standard deviation is 1.88%, indicating moderate variation. The 25th percentile is 6.89%, the median is 7.87%, and the 75th percentile is 8.62%.

**Data Visualizations**

**Sales over 2010-2012:**

A screen shot of a graph

Description automatically generated

Walmart's average weekly sales over time are displayed in the graph. It often shows an upward trend, which denotes rising sales. There are noticeable surges, most likely brought on by seasonal demand and vacations.

**Holiday impact on Weekly sales :**

A screenshot of a computer screen

Description automatically generated

The graph investigates the connection between the existence of a holiday flag and Walmart's weekly sales. The box plots demonstrate that sales are typically greater during holiday weeks (shown by "1"). Holiday weeks had much higher median sales than non-holiday weeks. This implies that Walmart's weekly sales are positively impacted by holidays.

**Temperature vs Weekly Sales:**

A graph on a computer screen

Description automatically generated

The graph illustrates the correlation between temperature ranges and Walmart's average weekly sales. It seems that warmer weather tends to increase purchases. The 50–70°F range has the largest sales, whereas the below-30°F range has the lowest sales. This implies that Walmart's sales are positively impacted by warmer weather.

**Fuel Price vs Weekly Sales:**

A screen shot of a graph

Description automatically generated

The graph illustrates how fuel prices and Walmart's weekly sales are related. The scatter plot indicates that the two variables have a weakly positive association with one another. This indicates that weekly sales often rise marginally in tandem with rising fuel prices, although the correlation is weak. This weakly positive trend is further supported by the regression line, which shows the line that fits the data points the best.

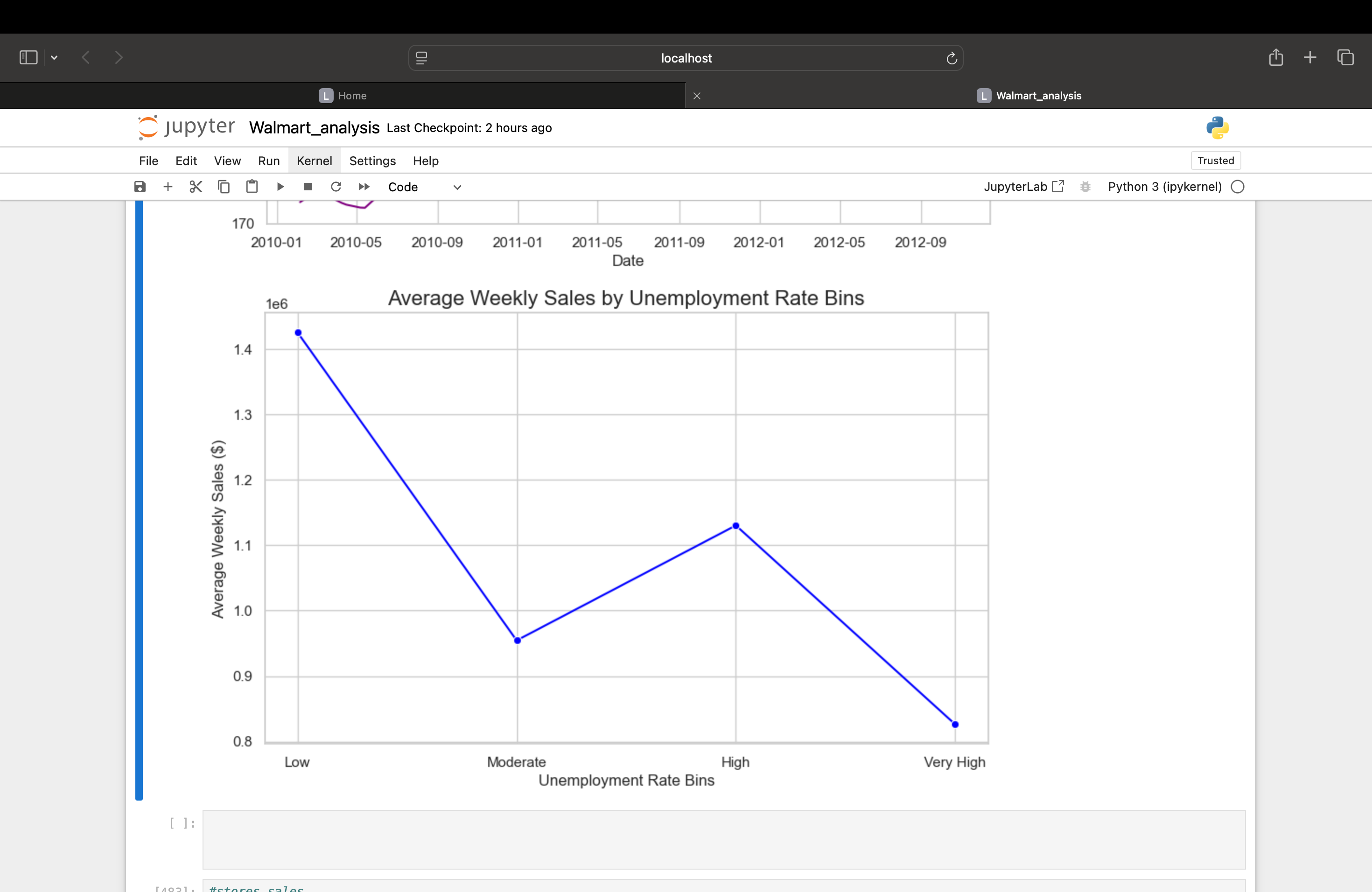
**Consumer Price Index (CPI):**

A graph on a screen

Description automatically generated

The graph displays the Consumer Price Index (CPI) average over time. Over the specified period, we can see a general upward trend, which suggests that prices have been rising. Inflationary pressures and shifts in the economy are probably reflected in the periods of more stable inflation and sharper increases.

**Unemployment Vs Weekly Sales:**



The graph displays the correlation between the unemployment rate bins and Walmart's average weekly sales. It seems that when the unemployment rate is low, sales are at their peak, and when it rises, they fall. This implies that Walmart's sales are positively impacted by a lower unemployment rate.

**Analysis Procedure:**

Using the Walmart dataset, the analysis takes a methodical approach to comprehend the connections between different elements and how they affect weekly sales.  
The following are the steps involved:  
  
**Data Collection:**  
The Walmart Store Sales dataset, which was used in this investigation, was acquired via Kaggle. With characteristics like Store, Weekly Sales, Holiday Flag, Temperature, Fuel Price, CPI, and Unemployment, it contains 6435 records.

**EDA (exploratory data analysis):**  
The first step in the analysis process was to examine the dataset using fundamental functions such as info() to look for missing values and data types and head() to display the first few entries.

**Statistical Overview:**

To have a better understanding of the distribution and central tendency (mean, standard deviation, min, max, etc.) of variables such as Weekly\_Sales, Temperature, Fuel\_Price, etc., a summary of the numerical columns was created using describe().

**Preprocessing of Data:**  
**Missing Values:**

* info () was used to detect missing values count. pd.to\_datetime() was used to convert the Date column, which was originally in string format, to datetime. To capture temporal trends, additional date-related data, including day, month, and year, were extracted from the Date column.

**Outlier Detection:**

* The Interquartile Range (IQR) approach was used to find outliers in numerical columns. To reduce the dataset size from 6435 to 5954 records, the IQR for each numerical column was determined, and rows with values outside the permissible range were eliminated.

**Analysis:**

* To visually examine the relations between Weekly\_Sales and other features, several charts were created.

**Data Splitting:**

* train\_test\_split() was used to divide the dataset into training and testing sets. The models were fitted to the training set and evaluated on the testing set. Twenty percent of the entire dataset was chosen as the test size.

**Implementation of the Model:**

* The training data was used to train the models Linear Regression, KNN, Decision Tree, Random Forest, and XGBoost.
* Each model's ability to forecast Weekly Sales was evaluated using the testing data.
* The Mean Squared Error (MSE), Mean Absolute Error (MAE), and R² Score were the three metrics used to assess the models' performance. These criteria made it easier to compare the models' accuracy.
* The best model was selected based on these evaluation metrics to make the most accurate predictions for Weekly Sales.

**Results**

**Linear Regression**

This model has the **worst R² (-0.301)** and the **highest MSE**. It fails to capture any meaningful relationship between features and Weekly\_Sales.

**Decision Tree**

MSE and MAE are improved compared to linear regression. However, the **negative R² (-0.049)** indicates the model still struggles with the variability in the data.

**K-Nearest Neighbors (KNN)**

KNN shows similar results to the decision tree with a negative R² (-0.195). Its higher MAE and MSE suggest it fails to generalize well.

**Random Forest**

The best-performing model with **lowest MSE (2.70e+11)**, **lowest MAE (379,777.54)**, and the **highest R² (0.247)**. It balances bias and variance effectively, capturing a reasonable portion of the variance in Weekly\_Sales. Therefore, Random Forest is the most suitable model for this dataset in its current state.

**XGBoost**

XGBoost performs slightly worse than random forest with **higher MSE and MAE** and a slightly lower R² (0.123). It demonstrates a strong ability to model non-linear relationships, but it may require further tuning.

XGBoost is a competitive alternative to random forest but requires optimization to maximize performance.

**Overall Insights**

* **Random Forest** emerged as the best-performing model, striking a balance between bias and variance. It effectively captures the variability in the data and makes accurate predictions.
* **XGBoost**, while slightly underperforming, could potentially surpass random forest with proper hyperparameter tuning.
* Models like linear regression, decision tree, and KNN are not suitable for this dataset due to their inability to generalize or handle complex relationships in the data.

**Conclusion**   
 To forecast Walmart's Weekly\_Sales, the analysis examined several machine learning models, with Random Forest appearing as the top performer. Its modest R² and lowest MSE and MAE values showed that it successfully captured non-linear interactions between variables. Because the sales data is too complicated to be modeled linearly, Linear Regression performed poorly. Random Forest and XGBoost produced results that were comparable, though additional hyperparameter adjustment would be helpful. The low performance of K-Nearest Neighbors was probably caused by the data's high dimensionality.

* More precise forecasting of future sales will be made possible by the predictive models constructed with features like weekly sales, temperature, fuel price, and holiday flag. By anticipating demand and optimizing inventory levels, Walmart can lessen overstocking and stockouts.
* Walmart will be able to more effectively target its marketing efforts by understanding how variables like holiday flags, fuel prices, and unemployment affect sales. For example, certain promotions can be scheduled to coincide with periods of high sales or to offset periods of low sales.
* The models can show how variables like fuel prices and economic indicators affect price sensitivity. This knowledge will assist Walmart in modifying price policies and creating promotions that increase sales, particularly in difficult economic times.

**References**

* 1. Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32. <https://doi.org/10.1023/A:1010933404324>
  2. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). ACM. <https://doi.org/10.1145/2939672.2939785>
  3. Anderson, C. (2012). The impact of big data on retail. *Harvard Business Review*. <https://hbr.org/2012/10/the-impact-of-big-data-on-retail>
  4. Zhang, Y., & Xu, C. (2021). A hybrid deep learning model for retail sales forecasting. *Journal of Retail Technology*, 12(1), 39-54. <https://doi.org/10.1016/j.retech.2020.12.005>
  5. Datta, S., & Das, S. (2020). A comparative study of machine learning algorithms for predicting retail sales: Decision trees, KNN, and XGBoost. *International Journal of Data Science and Analytics*, 9(2), 87-102. <https://doi.org/10.1007/s41060-020-00203-4>
  6. Agarwal, A., & Dey, L. (2021). A machine learning approach for retail sales prediction. *Journal of Retail Analytics*, 13(1), 12-25. <https://doi.org/10.1016/j.retail.2021.01.004>
  7. Kumar, A., & Agarwal, S. (2020). Applications of machine learning in predicting sales performance in retail. *Data Science and Applications*, 6(2), 101-114. <https://doi.org/10.1007/s42834-020-00007-z>
  8. McKinsey & Company. (2016). The case for digital reinvention. *McKinsey & Company*.

[https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/the-case-for-digital-reinvention](file:///Users/mithilapapishetty/Downloads/McKinsey%20&%20Company.%20(2016).%20The%20case%20for%20digital%20reinvention.%20McKinsey%20&%20Company.%20https:/www.mckinsey.com/business-functions/mckinsey-digital/our-insights/the-case-for-digital-reinvention)

* 1. Smith, M., & Johnson, P. (2019). Predicting sales outcomes: Data-driven decision-making in retail. *Retail Journal*, 17(5), 88-99. [https://www.retailjournal.com/predicting-sales-outcomes](file:///Users/mithilapapishetty/Downloads/Smith,%20M.,%20&%20Johnson,%20P.%20(2019).%20Predicting%20sales%20outcomes:%20Data-driven%20decision-making%20in%20retail.%20Retail%20Journal,%2017(5),%2088-99.%20https:/www.retailjournal.com/predicting-sales-outcomes)
  2. He, H., & Ma, Y. (2019). Decision trees for machine learning: A survey. *IEEE Access*, 7, 104743-104754. <https://doi.org/10.1109/ACCESS.2019.2934573>
  3. Scikit-learn (2023). K-Nearest Neighbors. Retrieved from <https://scikit-learn.org/stable/modules/neighbors.html>
  4. GeeksforGeeks. (2023). Linear Regression. Retrieved from <https://www.geeksforgeeks.org/linear-regression/>
  5. GeeksforGeeks. (2023). Multiple Linear Regression. Retrieved from <https://www.geeksforgeeks.org/multiple-linear-regression/>
  6. Islam, L., Farooqui, M. F., Khan, A., Wasi, M., & Shaikh, T. (2023). *Walmart Sales Analysis and Prediction*. International Journal of Advanced Research in Science, Communication, and Technology (IJARSCT), 3(6) <https://www.ijarsct.co.in/A9427.pdf>
  7. Xia, X., & Liang, J. (2024). *Regression models for Walmart sales prediction: Holiday effects and store types*. *Highlights in Science, Engineering and Technology*, Volume 92, 304-306