Group Members:
Tulsi Patel, cjw6jz, Chetu Barot, zrz9ry, Esha Sharma, exy8eb, Celestine Nguyen, haw3wk, Jennifer Vo, phb8pt, Shruti Bala, aen6ju

Github: https://github.com/CelestineNguyen/DS3001_SalesProject/tree/main

Final Paper

For our project, we decided to analyze what products affect sales in retail markets using data we sourced from a Kaggle competition for store sales time series forecasting. The data contains daily sales records for different products across multiple stores along with translation counts, oil prices, and store specific details. We used K-nearest neighbors (KNN) and Linear regression to develop and analyze sales patterns.

The KNN model outperformed the linear regression model with a relatively high R2 value of 0.89. It captured non-linear relationships by making accurate predictions for lower sales values and helped us look into some of the more irregular sales patterns. Residual analysis showed small systematic errors but the model struggled where we had fewer data observations for higher sales ranges due to fewer relevant neighbors. The linear regression model, which had a lower R2 score of 0.56, showed trends for lower sales but couldn't predict with accuracy the complexity of the higher sales. Recurring residual patterns showed missing features or unaccounted interactions, limiting its ability to generalize across more diverse sales ranges.

The analysis confirms that product type significantly impacts sales but is also dependent on other variables like store locations, local promotions, and seasonality of products. While KNN provided a more comprehensive framework for predicting sales using product families and understanding these influencing variables for the more complex higher sales, linear regression offers more insights into the linear relationships within the data.

Understanding the factors that drive sales in retail environments is critical for businesses aiming to optimize inventory, tailor marketing strategies, and maximize profitability. Grocery stores, in particular, face unique challenges due to the diversity of products, seasonal demand, and varying consumer preferences. This study addresses the central question: *"*Which products affect sales rates at stores?" By analyzing patterns in sales data, this research aims to identify product families that significantly influence overall sales, providing actionable insights for data-driven decision-making in store operations. The primary objective of this study is to identify which product families impact sales the most. Understanding these dynamics allows store managers to optimize shelf space allocation, develop targeted pricing and promotional strategies, and enhance customer satisfaction while maximizing revenue. Furthermore, the study evaluates the effectiveness of predictive models in capturing the complexities of sales behavior.

Two models, K-Nearest Neighbors (KNN) and Linear Regression, were employed to predict sales and analyze the relationship between product families and sales rates. The findings highlight significant insights into sales behavior. The KNN model achieved a high R2 value which indicates a strong ability to explain variance in sales. It effectively captured non-linear relationships and predicted lower sales values accurately. However, it struggled with sparsity in higher sales ranges because of limited data in these regions. On the other hand, Linear Regression with a moderate R2 score provided baseline insights into linear relationships but failed to capture the complexity of interactions affecting higher sales.

Product family emerged as a significant factor influencing sales rates, particularly for commonly bought, high-frequency items. Higher sales values revealed interactions with other variables, like store location, promotional events, and seasonality, suggesting the need for additional modeling sophistication. The residuals from KNN were more evenly distributed,

demonstrating minimal systematic errors. In contrast, Linear Regression did show patterns in residuals which reflected missing features or unaccounted interactions. This study provides a comprehensive analysis of sales drivers in retail and demonstrates the importance of advanced modeling techniques for accurate sales prediction. The findings underscore the advantages of non-linear models like KNN for capturing complex patterns in sales data. However, they also highlight the limitations of current approaches in modeling higher sales values, emphasizing the need for further feature engineering or advanced techniques such as ensemble methods.

The Stores Sales Kaggle competition came with 7 csv files:

samplesubmission.csv: includes two vars (id, sales). The id represents the product id and the sales represents the sales that the product will have in the future (in this competition it will be the next 15 days after the last recorded date). This csv is just a csv that shows the format for the kaggle competition submission. Both of these are key variables.

train.csv: This csv contains training data with store_nbr, family, on promotion, and sales. The key variables here are the sales and family variables as we can categorize the item and also see the total sales for that product on a certain date.

test.csv: Simply contains the 15 dates that we will be creating target sales numbers for.

stores.csv: A Csv that contains all the different Favorita Grocery Stores. It has the following variables: city, state, type, and cluster. Cluster allows us to see which stores are in relatively similar locations. The city could be a key variable as heavily populated cities could be an indicator of more sales for a particular product.

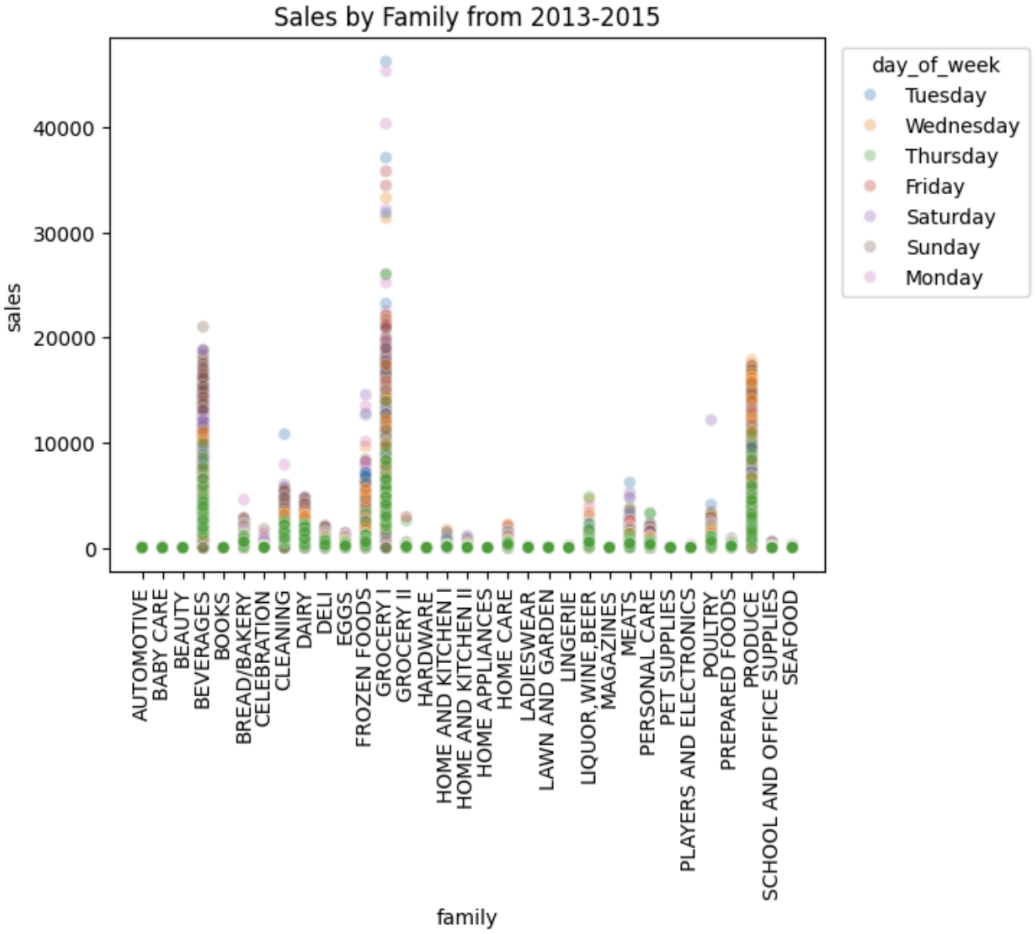
oil.csv: Includes dates and the daily oil price: Due to Ecuador being oil-dependent, the economy (therefore prices and hence sales) may also get affected.

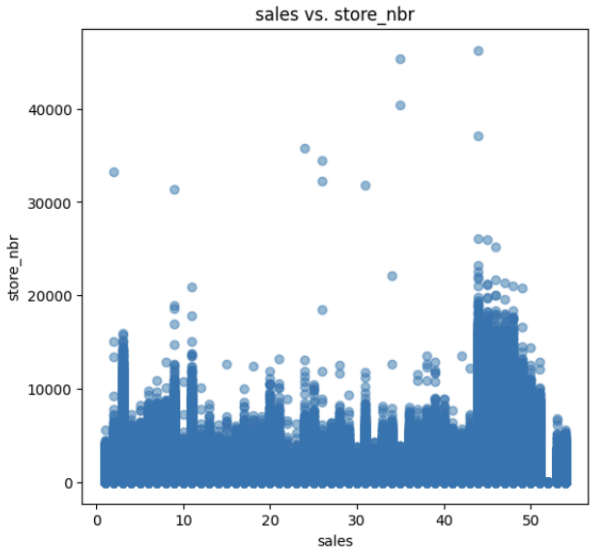
holidays_events.csv: Includes date, type, locale, locale_name, description

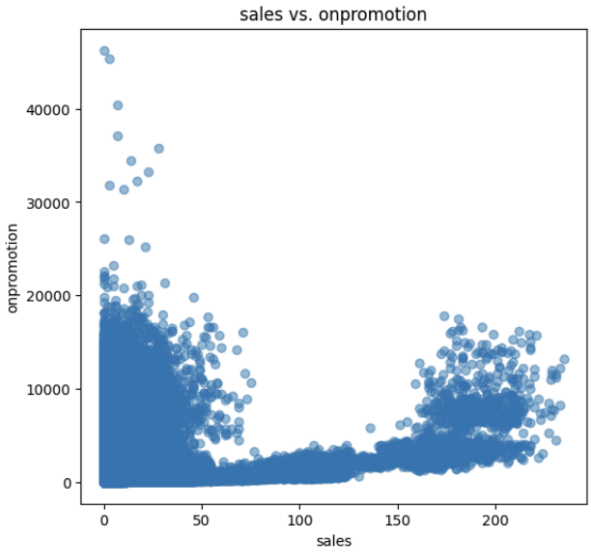
transactions.csv: Includes the date, store_nbr, and transactions. All of these variables are key as we must know the date, which store and total number of transactions for that store.

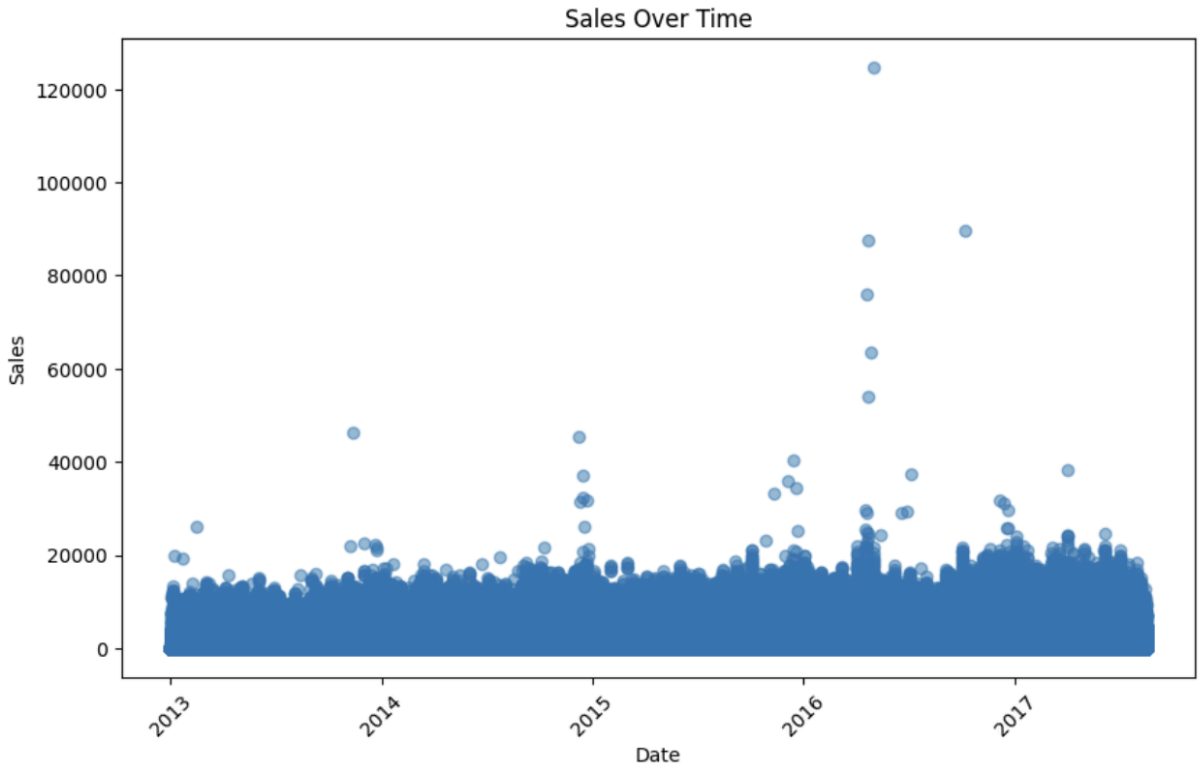
This data describes the sales of the product families sold at Favorita stores located in Ecuador. The train.csv dataset includes an id, date the item was sold, store number which is the store the item was sold at, family which is what type of item it was, # sales for that item, and on promotion which indicates if the item was on sale or not. All of the data is numeric except for the family column. There are 1048574 rows of data in the dataset. We then downloaded each data set and created pandas data frames. We then cleaned the data to convert dates to date-time format, add derived columns for the date, and filter data by the date range. After importing necessary libraries, we created the following scatter plots and bar graphs.

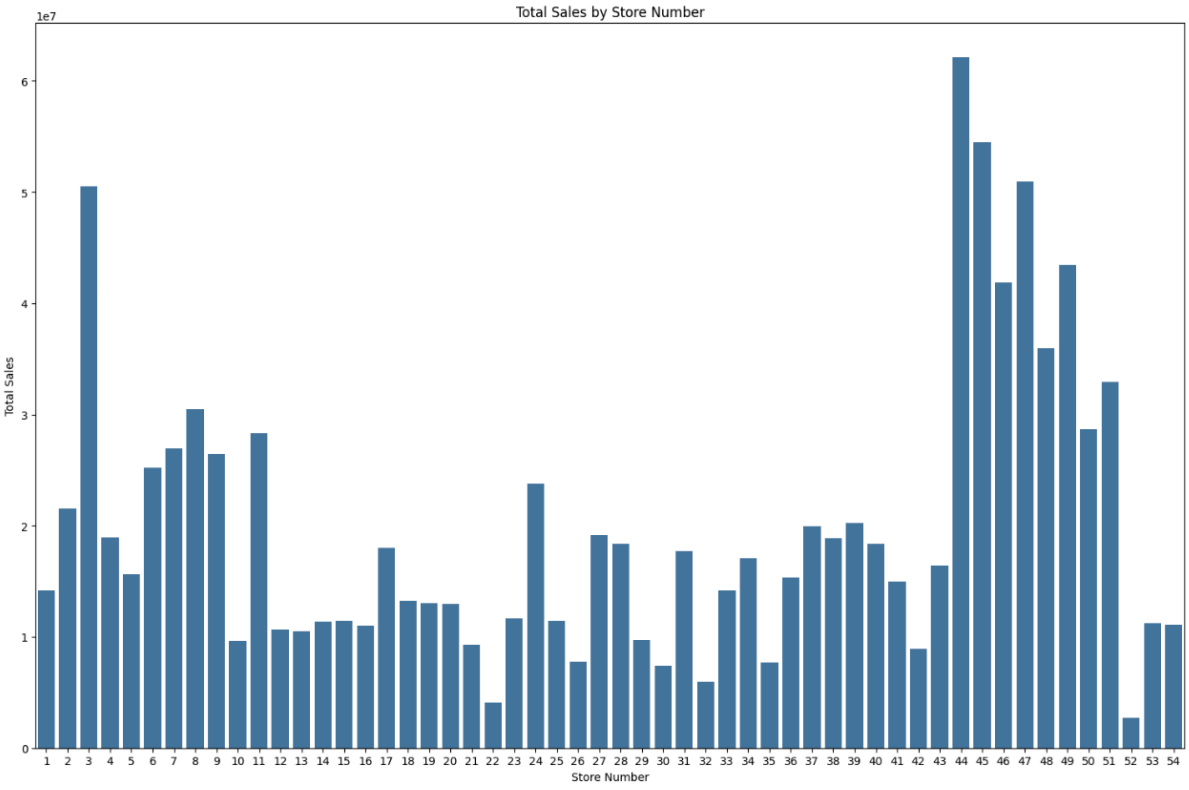
[ ]

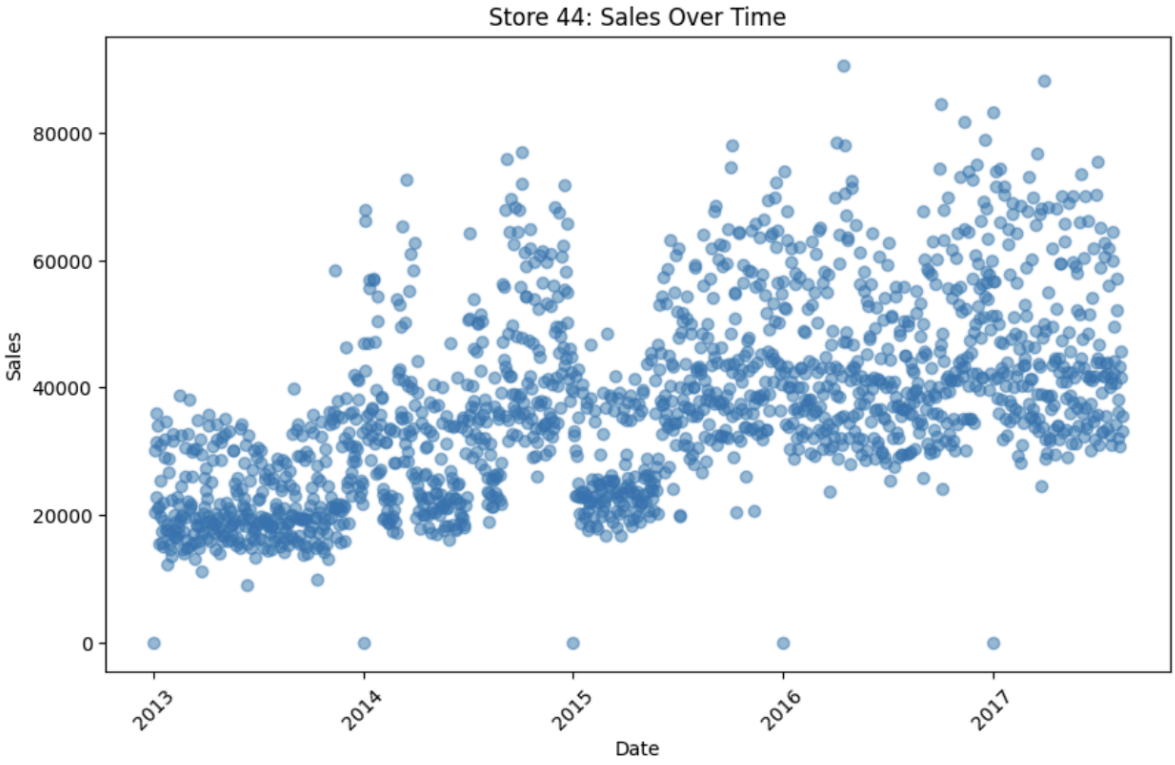

Sales by Family from 2013-2015
day_of_week
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Monday
sales
0
10000
20000
30000
40000
AUTOMOTIVE
BABY CARE
BEAUTY
BEVERAGES
BOOKS
BREAD/BAKERY
CELEBRATION
CLEANING
DAIRY
DELI
EGGS
FROZEN FOODS
GROCERY I
GROCERY II
HARDWARE
HOME AND KITCHEN I
HOME AND KITCHEN II
HOME APPLIANCES
HOME CARE
LADIESWEAR
LAWN AND GARDEN
LINGERIE
LIQUOR,WINE,BEER
MAGAZINES
MEATS
PERSONAL CARE
PET SUPPLIES
PLAYERS AND ELECTRONICS
POULTRY
PREPARED FOODS
PRODUCE
SCHOOL AND OFFICE SUPPLIES
SEAFOOD
family

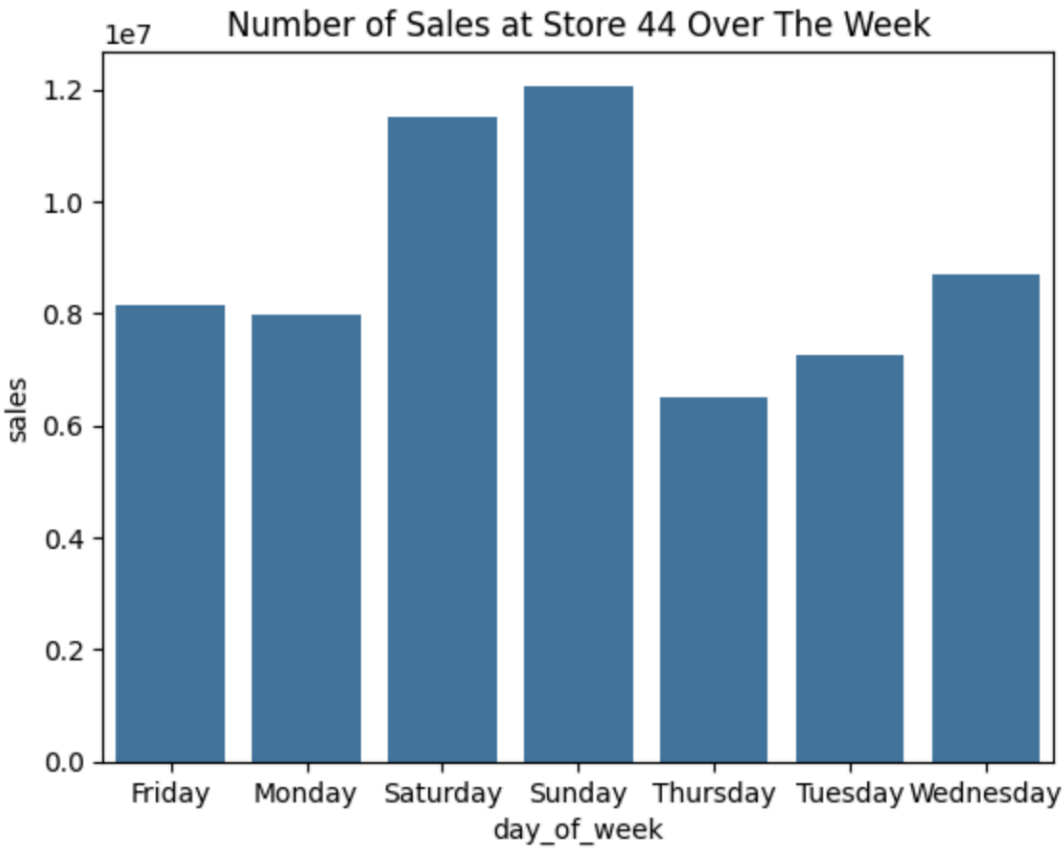

sales vs. store_nbr
store_nbr
0
10000
20000
30000
40000
0
10
20
30
40
50
sales

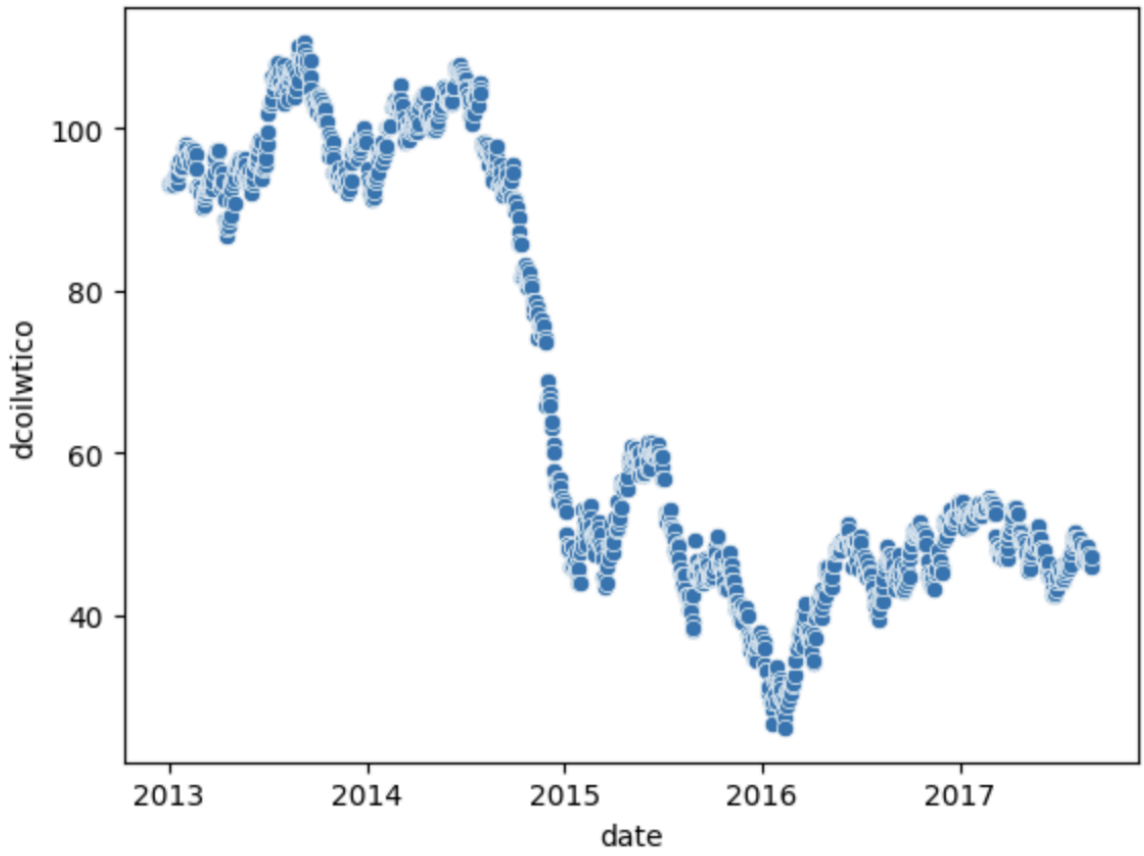

sales vs. onpromotion
onpromotion
0
10000
20000
30000
40000
0
50
100
150
200
sales


Sales Over Time
Sales
120000
100000
80000
60000
40000
20000
0
2013
2014
2015
2016
2017
Date


Total Sales by Store Number
1e7
Total Sales
Store Number


Store 44: Sales Over Time
Sales
80000
60000
40000
20000
0
2013
2014
2015
2016
2017
Date


Number of Sales at Store 44 Over The Week
1e7
1.2
1.0
0.8
0.6
0.4
0.2
0.0
sales
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday
day_of_week



We then calculated the total transactions per store

	store_nbr	transactions
0	1	2553963
1	2	3219901
2	3	5366350
3	4	2519007
4	5	2347877
5	6	3065896
6	7	2995993
7	8	4637971
8	9	3516162
9	10	1652493
10	11	3972488

Then we grouped and summed transactions using the transactions.csv. This code provides a summary of customer activity, highlighting the total number of transactions per store. This information can be useful for comparing store performance or identifying trends.

cluster	count
3	7
6	6
10	6
15	5
13	4
14	4

We then processed the oil.csv dataset. This snippet provides an initial look at the oil dataset, including its size and structure. Oil prices are likely used as a feature in the sales analysis, as they can influence economic conditions and purchasing behavior.

	date	dcoilwtico
0	2013-01-01	NaN
1	2013-01-02	93.14
2	2013-01-03	92.97
3	2013-01-04	93.12
4	2013-01-07	93.20

Methods:

The objective of this pre-analysis plan, or methods, is to outline the methodology and strategies for analyzing how different types of products sold affect store sale rates. The plan also addresses additional questions regarding the prediction of future sales, the impact of promotions,

and the effects of holidays and events on sales trends. To begin, the dataset includes critical observations such as the types of products sold, their associated product families, the date of data collection, the store identifier, and the number of sales predicted for specific product families. These observations form the foundation of the analysis.

This study employs supervised learning, specifically regression techniques, to predict the factors influencing sales. The primary target variable is the sales column, and the goal is to analyze how changes in various predictors impact this value. The analysis involves developing multiple regression models, including K-Nearest Neighbors (KNN) and Linear Regression, to compare their performance and identify the most influential factors on sales. By handling both linear and non-linear relationships, these models aim to provide comprehensive insights into sales trends.

The evaluation metrics for this project include R-squared (R^2) values and Root Mean Squared Error (RMSE). The best-performing model will exhibit the highest R^2 value and the lowest RMSE, ensuring strong predictive accuracy. Residual analysis will further help assess systematic errors and refine feature selection or model design. Success will be defined by the ability to predict sales accurately and to understand the effects of variables such as product type, store location, and seasonal trends.

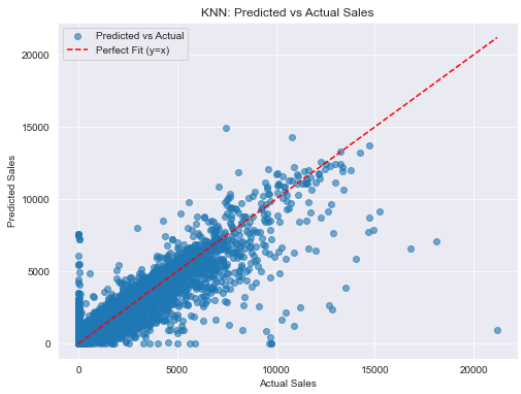
Several challenges are anticipated in the analysis. High dimensionality and noise in the dataset may necessitate careful feature selection or dimensionality reduction techniques like Principal Component Analysis (PCA). Additionally, poor model performance may require iterative refinement of the feature matrix (X) to focus on impactful variables. Sparse data in

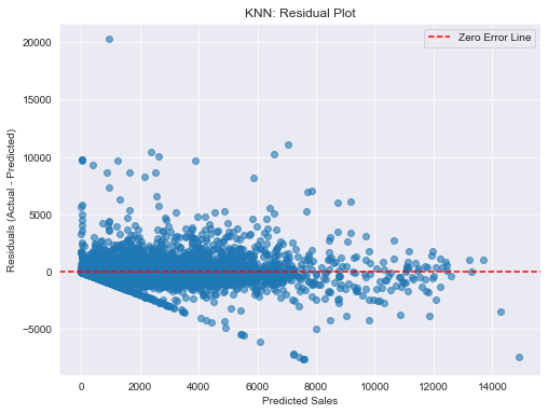
higher sales ranges could also lead to inaccuracies, which might be addressed through stratified sampling or data weighting strategies.

Feature engineering plays a critical role in preparing the dataset for analysis. Categorical variables such as the ‘family’ column will be one-hot encoded, while the ‘date’ column will be transformed into numerical features like day of the week, month, and year to capture temporal effects. Integration of additional datasets, such as store-specific details and promotion information, will be conducted by merging data based on store number and date.

The results of this analysis will provide insights into the effectiveness of the chosen models and their ability to capture complex relationships in the data, setting the stage for a detailed evaluation of their performance.

Results:

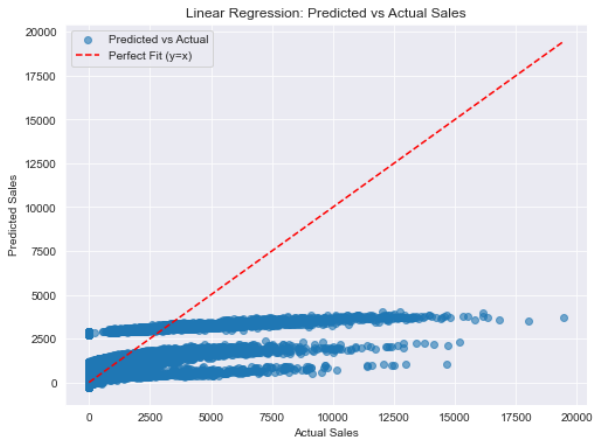


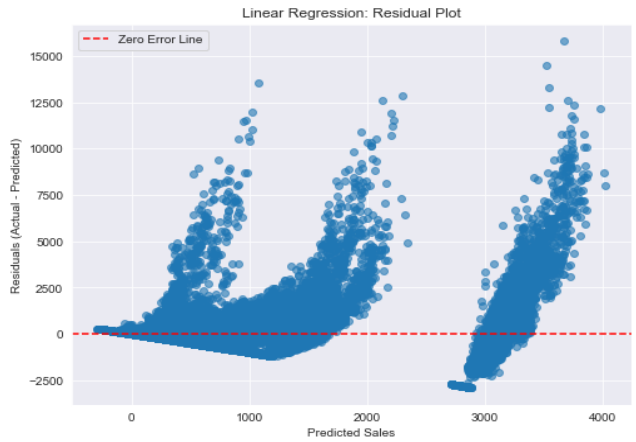


The KNN model demonstrates a strong ability to predict sales rates, particularly for lower sales ranges. In the scatter plot comparing predicted vs. actual sales, the points are closely aligned with the red "perfect prediction" line, suggesting that KNN predictions are generally accurate. However, as sales increase, predictions begin to deviate slightly, reflecting the model's

limitations in handling extreme values. This behavior indicates that while KNN captures non-linear relationships in the data better than linear regression, it still struggles with sparsity in higher sales regions, likely due to fewer relevant neighbors in these ranges.

The residual plot for KNN shows well-distributed errors with fewer discernible patterns compared to Linear Regression. The residuals remain closer to zero overall, indicating that the model does not systematically over- or under-predict sales across most ranges. However, there is some widening of residuals for higher sales, suggesting that KNN underestimates sales at these values. The high R^2 score of 0.89 underscores the model's effectiveness in explaining the variance in the data and its ability to generalize well for most sales predictions.





The regression model offers moderate predictive capabilities but falls short when handling higher sales values. The scatter plot comparing predicted vs. actual sales reveals that most predictions cluster near the correct values for lower sales. However, as actual sales increase, the predictions drift further from the red "perfect prediction" line, indicating that the model struggles to capture the complexity of the data in higher ranges. This issue is further highlighted by the residual plot, which exhibits a systematic curve and increasing variance in errors as predicted sales grow. Ideally, residuals should scatter randomly around zero with no

visible structure, but the observed patterns suggest the model cannot generalize well for higher sales.

The R^2 score for Linear Regression, at 0.56, reflects a moderate ability to explain the variance in the data. While the model captures some trends in the data, significant gaps in the scatter and residual plots confirm its limitations. This aligns with the observation that Linear Regression predicts lower sales reasonably well but struggles with higher values.

The graphs provide valuable insights into the predictive question, “Which products affect sales rates at stores?” The models demonstrate that product type is a significant factor influencing sales rates, as indicated by the variations captured in both the KNN and regression results. The scatter plots from both models show that lower sales values are predicted more accurately, which aligns with the idea that sales patterns for commonly sold products are easier to model due to their frequent occurrence in the dataset. The divergence from the perfect prediction line at higher sales values suggests that the influence of product type becomes more complex or interacts with other variables, such as store location or promotional events.

The KNN model, with its high R^2 score of 0.89, captures non-linear relationships more effectively, indicating that product types with irregular sales patterns might benefit from considering nearest-neighbor relationships based on sales trends and store-specific factors. On the other hand, the regression model, with a moderate R^2 of 0.56, highlights the limitations of linear assumptions in explaining the variance in sales rates. The residual analysis further underscores that neither model fully accounts for the complexity in higher sales, potentially pointing to interactions between product type and other features like promotions or seasonality.

Overall, the analysis confirms that products significantly impact sales rates, but their effects are intertwined with other variables. While KNN provides a more detailed and adaptable framework for predicting sales across diverse products, both models suggest the need for further feature engineering or advanced modeling to fully capture the dynamics of product influence on sales rates.

Conclusion:

The analysis conducted on the Kaggle "Store Sales Time Series Forecasting" dataset offers actionable insights into the relationship between product families and sales rates, underscoring key trends and areas for improvement. By utilizing K-Nearest Neighbors (KNN) and Linear Regression models, we gained a better understanding of how different products influence overall store performance. While the study demonstrates the use of machine learning models in retail analytics, it also reveals several challenges and opportunities for future exploration.

Our results confirm that product families significantly impact sales rates. The KNN model outperformed the Linear Regression model, achieving a high R^2 value of 0.89, indicating its superior ability to explain the variance in sales data and capture non-linear relationships. This highlights that sales patterns for certain product types, particularly those with irregular trends, benefit from models that account for nearest-neighbor relationships in sales trends and store-specific factors. The Linear Regression model with an R^2 of 0.56 demonstrated reasonable predictive accuracy for lower sales values but struggled to generalize for higher sales, which highlights the limitations of linear assumptions in complex retail environments.

Our analysis suggests that product types are easier to predict when they exhibit frequent and consistent sales patterns. The accuracy of both models at lower sales values aligns with the

observation that commonly sold products dominate the dataset. Although higher sales values presented more complexity as predictions deviated significantly from actual sales, this divergence suggests that interactions between product type and other features, such as store location, promotions, and seasonality, play a critical role in influencing sales rates.

While KNN demonstrated strong performance in predicting the sales rates for lower sales ranges, its limitations in higher sales ranges underscore involving advanced machine learning techniques such as Gradient Boosting Machines or Neural Networks for future work. These models could better capture non-linear relationships between variables, allowing improved generalization for higher sales values.

Additionally, feature engineering and selection could enhance the predictive accuracy of the model. For example, time series decomposition or interaction terms could be demonstrated with models that use time series decomposition or interaction terms [1]. The presence of consumer demographics and competitive pricing data in the model could further improve the analysis and results. Although we corrected missing values in our cleaned data set by filling gaps in the transactions and oil price datasets, imputation strategies or robust models, such as matrix factorization, could provide more accurate estimates for missing data points. For limitations, both models faced challenges in accurately predicting higher sales values. Sparse data for these cases likely caused the limitation, which requires extending our analysis to targeted sampling or bootstrapping techniques to address underrepresented scenarios.

References

[1] A. A. Khan, O. Chaudhari, and R. Chandra, “A review of Ensemble Learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation,” Expert Systems with Applications, https://www.sciencedirect.com/science/article/pii/S0957417423032803 (accessed Dec. 14, 2024).

[2] Alexis Cook, DanB, inversion, and Ryan Holbrook. Store Sales - Time Series Forecasting. https://kaggle.com/competitions/store-sales-time-series-forecasting, 2021. Kaggle.