Project title: Food demand forecasting for a meal delivery service

Project overview

Description

This project focuses on demand forecasting for a meal delivery company operating across multiple cities, which relies on fulfillment centers for dispatching meal orders. The goal is to predict meal demand for the next 10 weeks, helping centers plan their stock of perishable raw materials and staffing needs more efficiently.

The dataset is sourced from Kaggle provides details on cities where the meal delivery service operates, category of meal, pricing, and promotions.

Key steps and analysis

1. Data reading and cleaning

- Loaded the dataset using Pandas and explored initial data characteristics.
- Addressed inconsistencies, such as swapping `base_price` and `checkout_price` values when needed to ensure logical pricing.

2. Exploratory data analysis (EDA)

- Price analysis: Analyzed distributions of base and checkout prices, uncovering pricing trends and the impact of discounts.
- Demand patterns: Visualized demand fluctuations across weeks and explored correlations with pricing and promotions.
- Category & Cuisine Analysis: Identified the most popular categories (e.g., Beverages, Rice Bowl, and Sandwich) and cuisines (e.g., Italian and Thai).

3. Feature engineering

- Created features such as month, quarter, discount ratios, and interaction terms for promotions.
- One-hot encoded categorical variables to prepare data for machine learning models.

4. Modeling approach

Baseline models

- Linear regression: Initial R^2 score of 0.3, improved to 0.5 with feature engineering.

Advanced techniques

- Ridge and lasso regression: Implemented for regularization, achieving an R^2 of around 0.59.
- Random forest: Achieved the best performance with an R^2 of 0.79 on train data.
- Gradient boosting: Moderate performance with an R^2 of 0.61, requiring further tuning.

Final model

- Random forest: Chosen for its ability to capture complex relationships, achieving an R^2 of 0.809 on the test set with a Mean Absolute Error (MAE) of 75.53.

Feature importance

- Discount effects: Discounts were found to positively correlate with increased demand.
- Promotional strategies: Analyzed the impact of email and homepage promotions, both individually and combined, on order volumes.

1. Data Preparation

1.1 Importing Necessary Libraries

To start, I imported essential Python libraries to handle and analyze the dataset:

- **Pandas** and **NumPy**: For data manipulation and numerical operations.
- **Seaborn** and **Matplotlib**: For creating insightful visualizations to explore and understand the data.

1.2 Loading the Datasets

The dataset was sourced from Kaggle and consists of three files:

1. **Train Dataset**: Contains historical data for demand forecasting, including week, center_id, meal_id, checkout_price, base_price, emailer_for_promotion, homepage_featured, and num_orders.
2. **Test Dataset**: Similar to the train dataset but missing the num_orders column, which is the target variable to be predicted.
3. **Meal Information Dataset**: Provides additional details about meals, including their category and cuisine.

I used the read_csv() function from Pandas to load the datasets and performed the following initial checks:

- Used .head() to preview the first few rows of the data for a quick overview.
- Utilized .info() to confirm the data types of each column and verify that there were no missing values in any of the datasets.
- Leveraged .describe() to analyze the numerical features (e.g., maximums, minimums, mean, and quartiles) for consistency and to identify any anomalies.

1.3 Joining Datasets

To enrich the analysis and modeling, I merged the **Meal Information** dataset with both the **Train** and **Test** datasets. This allowed me to incorporate additional context, such as the category and cuisine of each meal, into the demand forecasting process.

1.4 Data Quality Check and Corrections

During the data exploration phase, I identified a data inconsistency:

- The base_price was occasionally smaller than the checkout_price. This is logically incorrect since the base price represents the standard cost, while the checkout price reflects potential discounts or promotions applied.

To resolve this issue:

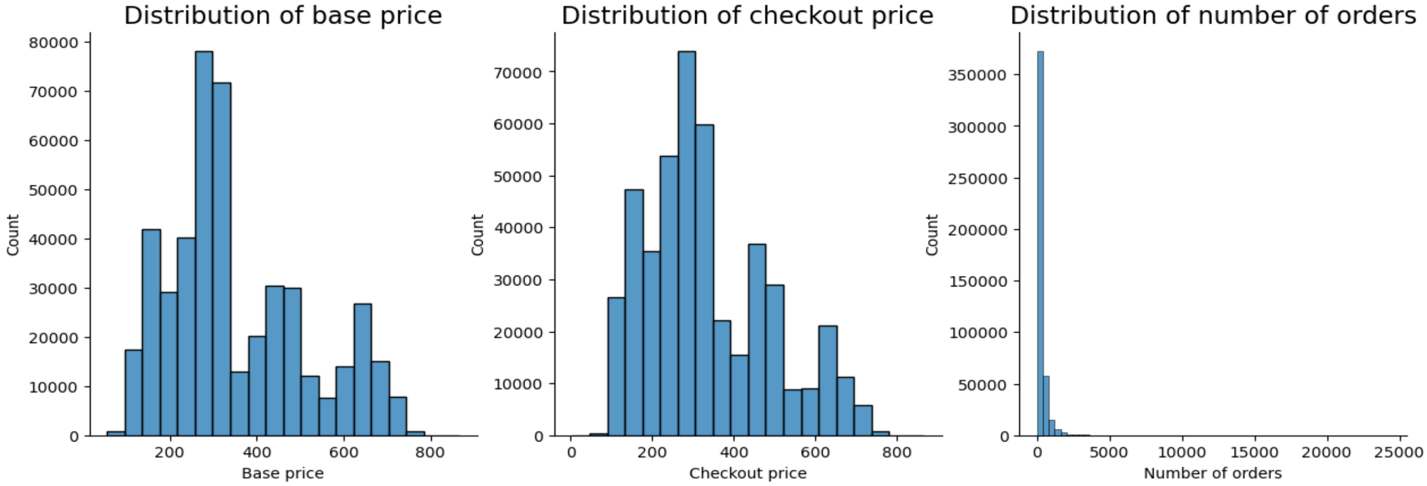
- I identified the affected rows where base_price < checkout_price.

- Corrected the base_price to match the checkout_price for these instances to ensure data accuracy and maintain the integrity of the dataset.

2.2 Exploratory Data Analysis (EDA)

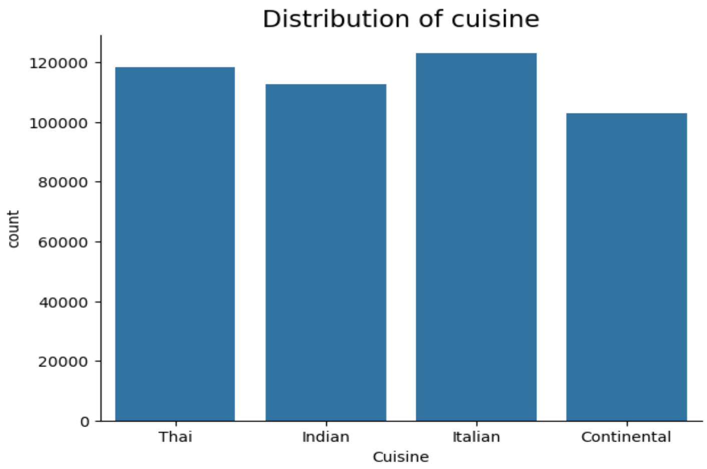
2.2.1 Summary Statistics and Distribution Analysis

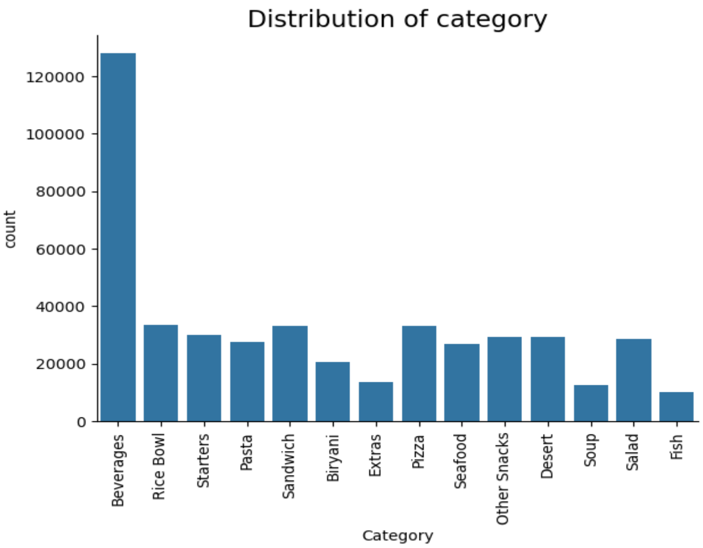
- **Base Price Distribution**: The base price follows a generally normal distribution with slight skewness and two notable outliers (additional peaks), which suggest some irregularities that may require closer inspection. This distribution is essential as the base price forms the foundation for the pricing strategy.
- **Checkout Price Distribution**: Similarly, the distribution of checkout prices exhibits a pattern comparable to the base price, confirming that the pricing strategies are aligned, though anomalies may be present due to outliers or errors, such as base prices being higher than checkout prices (which was corrected earlier).
- **Number of Orders Distribution**: The distribution of the number of orders is highly skewed, which is typical for demand forecasting tasks, as we often observe lower frequency of high-demand periods. This pattern reflects the cyclical nature of the business and may need special handling in modeling (e.g., log transformations).



2.2.2 Categorical Variables Analysis

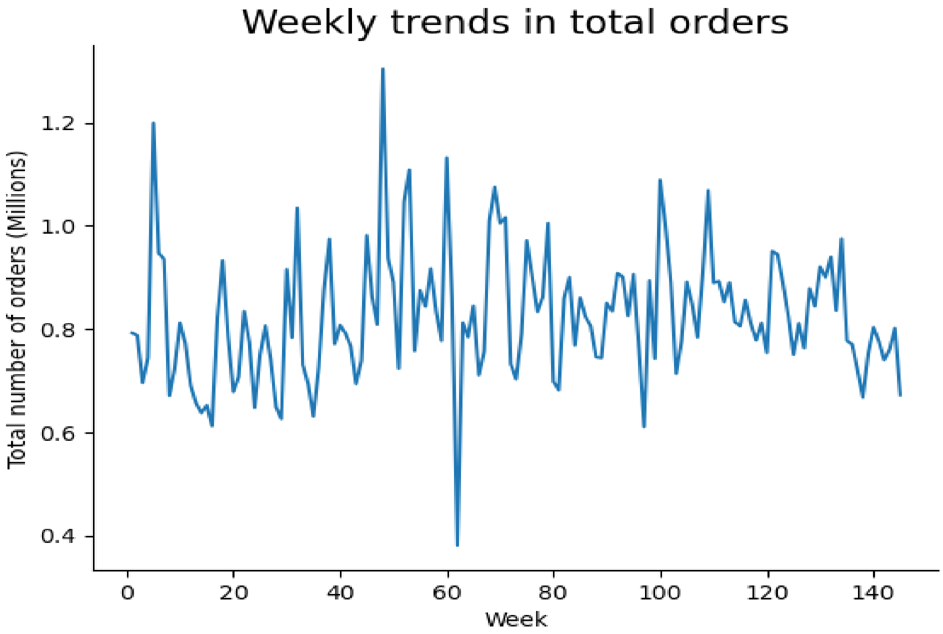
- **Cuisine Distribution**: The cuisines are fairly evenly distributed, with no extreme outliers. This indicates that, overall, there is no overwhelming preference for one cuisine, but there may still be subtle differences that are interesting when examined at a deeper level.
- **Meal Category Distribution**: The meal category distribution revealed that **Beverages** had the highest number of orders (over 120,000), while categories such as **Extras**, **Soup**, and **Fish** had the least, with only around 15,000 orders each. The remaining categories had orders around 30,000, suggesting varying customer preferences and demand levels.



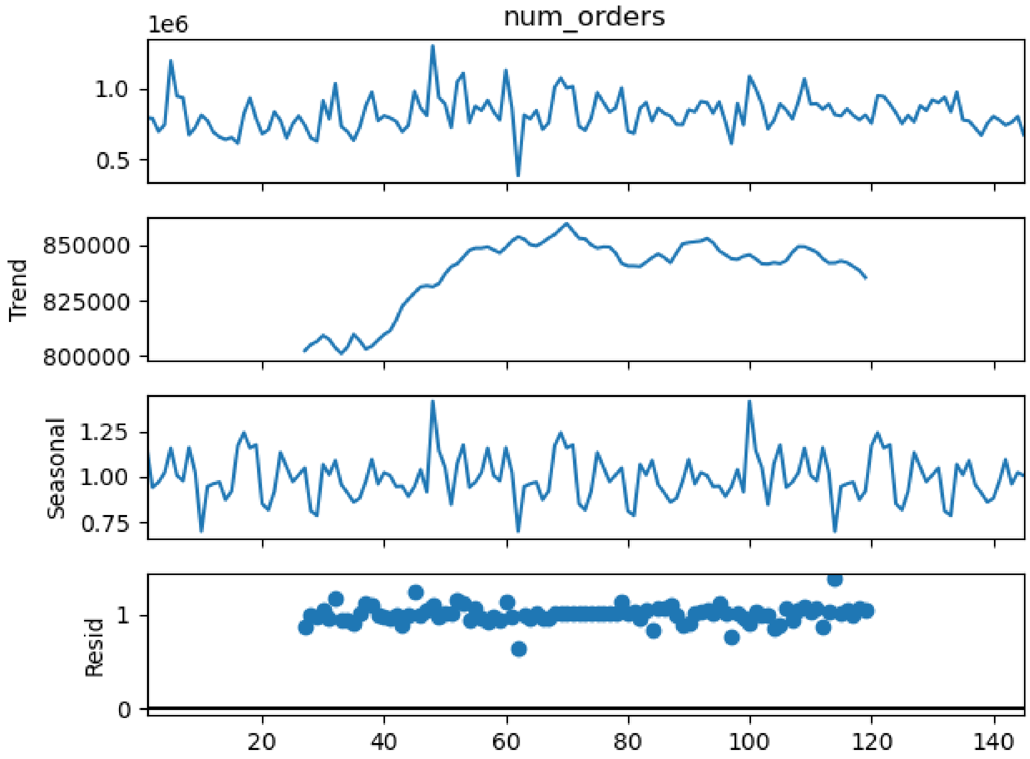


2.2.3 Time Series and Demand Analysis

- **Weekly Trend**: The weekly trend analysis shows periodic spikes in demand around **weeks 10**, **60**, and **100**, likely driven by external factors like promotions, holidays, or seasonal demand. This cyclical pattern highlights the need to forecast future spikes to optimize stock and staffing.

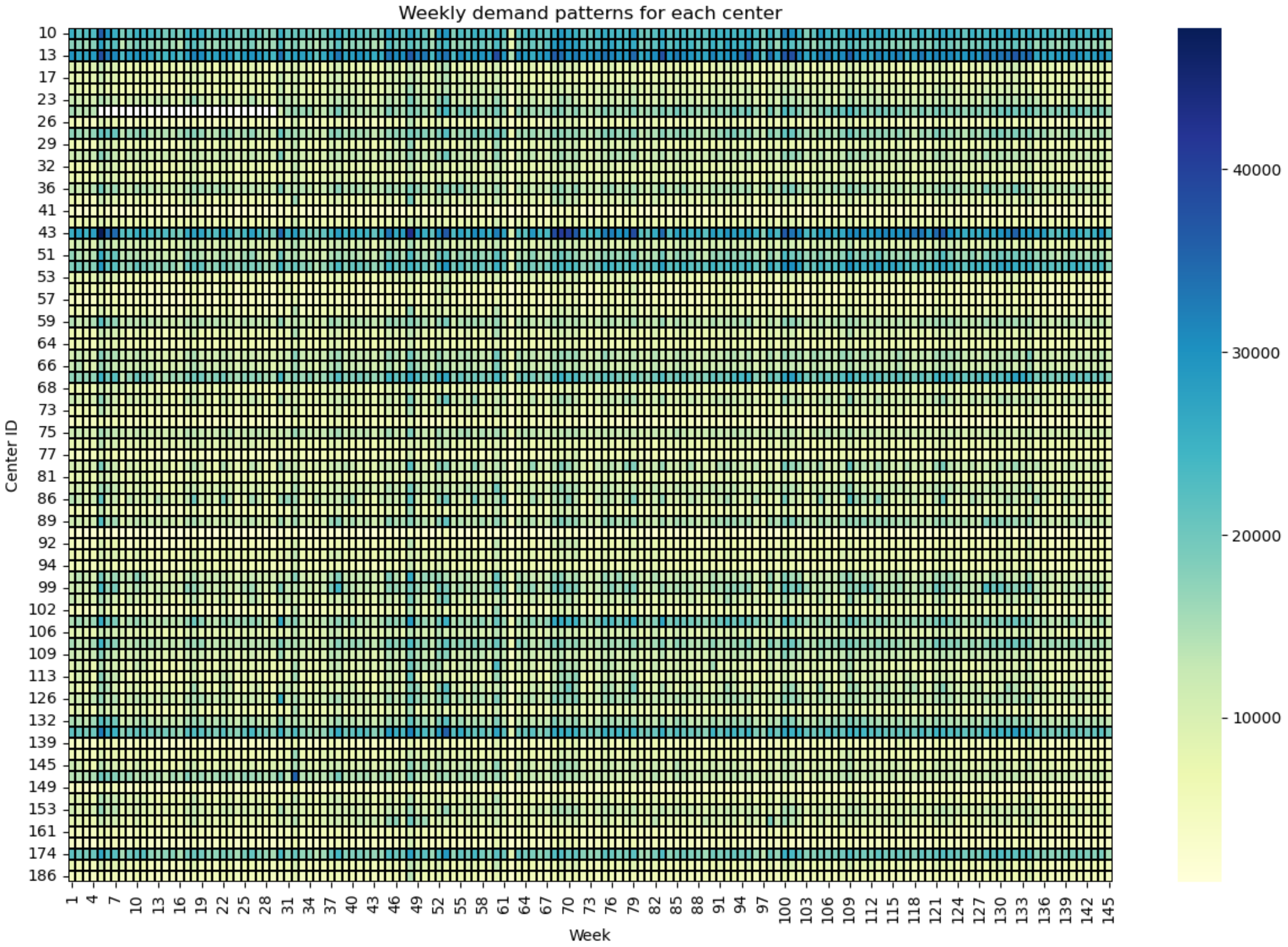


- **Seasonal Decomposition**: The decomposition of **num_orders** revealed an overall steady upward trend until around **week 80**, followed by a slight decline. The seasonal component exhibits regular fluctuations, suggesting predictable patterns, while the residuals (noise) appear random, indicating the decomposition model is well-suited for capturing these fluctuations.



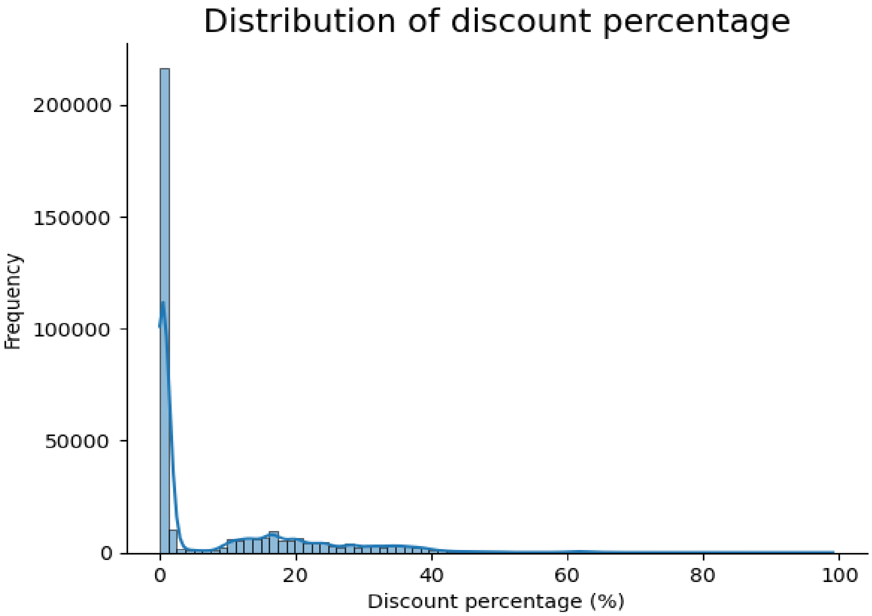
2.2.4 Demand Patterns and Heatmap Analysis

- **Heatmap of Demand Across Centers**: The heatmap analysis shows that certain centers, such as **Center 10**, **Center 11**, **Center 13**, and **Center 43**, consistently exhibit high demand. Meanwhile, centers like **51**, **52**, **67**, **137**, and **174** show medium demand, and others demonstrate lower demand. This heatmap provides insight into operational requirements and areas where demand optimization could be beneficial.

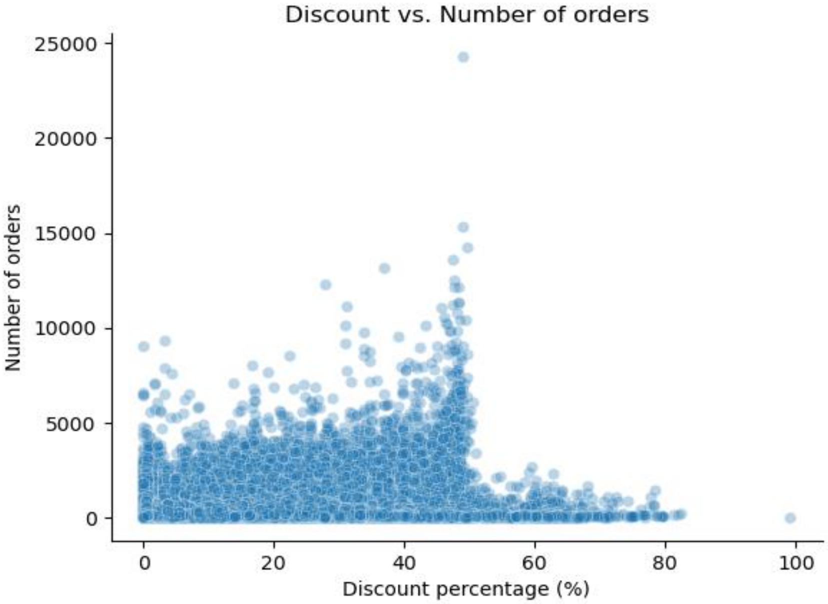


2.2.5 Price and Discount Analysis

- **Checkout Price vs. Base Price (Discount Analysis)**: The discount distribution revealed that most discounts are small, with the majority concentrated in the **0–10%** range. Discounts exceeding **20%** are rare, suggesting that large discounts are not a common strategy, which is typical for businesses aiming to maintain profit margins.



- **Discount Threshold Analysis**: After categorizing discounts into five bins (**0-10%, 10-20%, 20-30%, 30-50%, 50%+**), the average number of orders varied as follows:
  - **No discount**: Lowest average at 210 orders.
  - **30-50% discount**: Highest average at 550 orders.
  - **Other discount ranges** (10-20%, 20-30%, and 50%+): Around 350 orders on average. This analysis demonstrates how discounts influence sales volume, with moderate discounts yielding the highest demand.
- **Scatterplot between Number of Orders and Discount Percentage**: The scatter plot shows that most orders occur at lower discount percentages (**0–20%**), with occasional spikes in orders around **40%** discounts. This suggests that while discounts increase demand, there is no clear linear relationship, and higher discounts may not always translate into proportional sales growth.

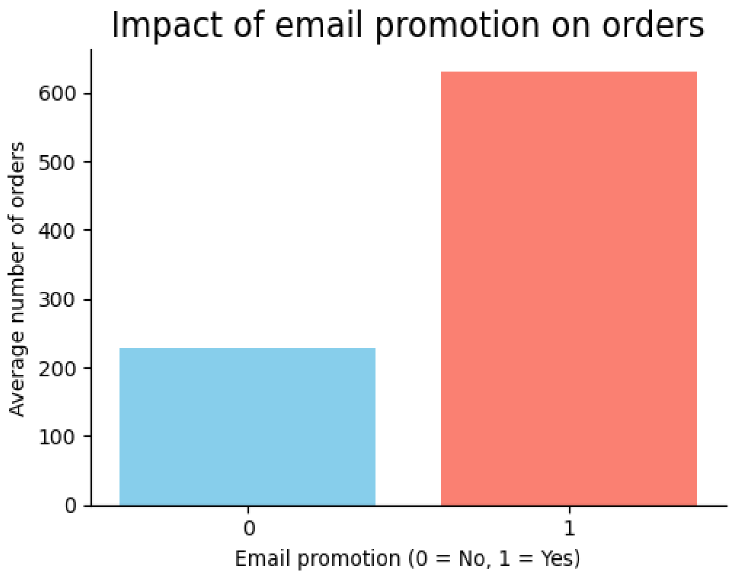


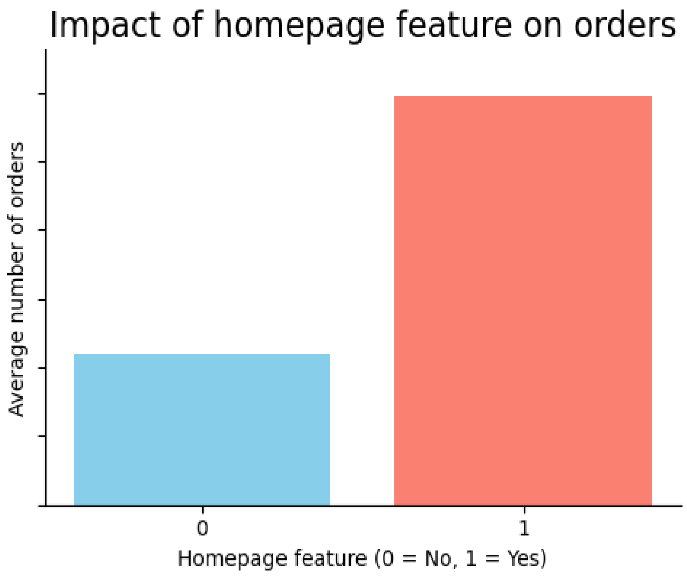
- **Correlation Between Discount and Price Variability**: The correlation matrix revealed weak relationships:
  - A slight positive correlation (**0.214**) exists between **discount** and **num_orders**, indicating that higher discounts are weakly associated with increased orders.
  - A weak negative correlation (**-0.175**) was found between **price_std** (price variability) and **num_orders**, suggesting that greater price variability might slightly decrease orders.
  - Minimal correlation (**0.131**) was observed between **discount** and **price_std**, implying that discounts are not strongly linked to price variability.

2.2.6 Promotion Effectiveness Analysis

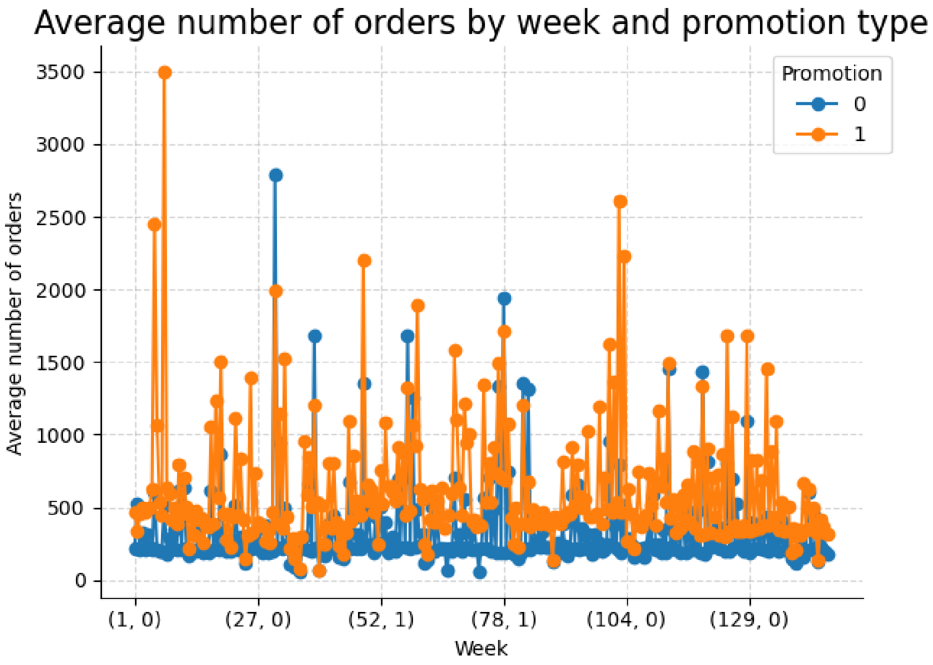
- **Average Orders with/without Promotions**:
  - Orders are significantly higher when **emailer_for_promotion** is active (average of 631 orders) compared to when it is not (average of 229 orders).
  - Featuring products on the **homepage** also leads to higher orders (average of 595 with homepage featured vs. 221 without).

- Combining both promotional strategies leads to the highest orders (average of **816 orders**), followed by using **emailer** (431 orders) and **homepage feature** (456 orders) independently. This highlights the powerful synergy of combining email promotions and homepage features.



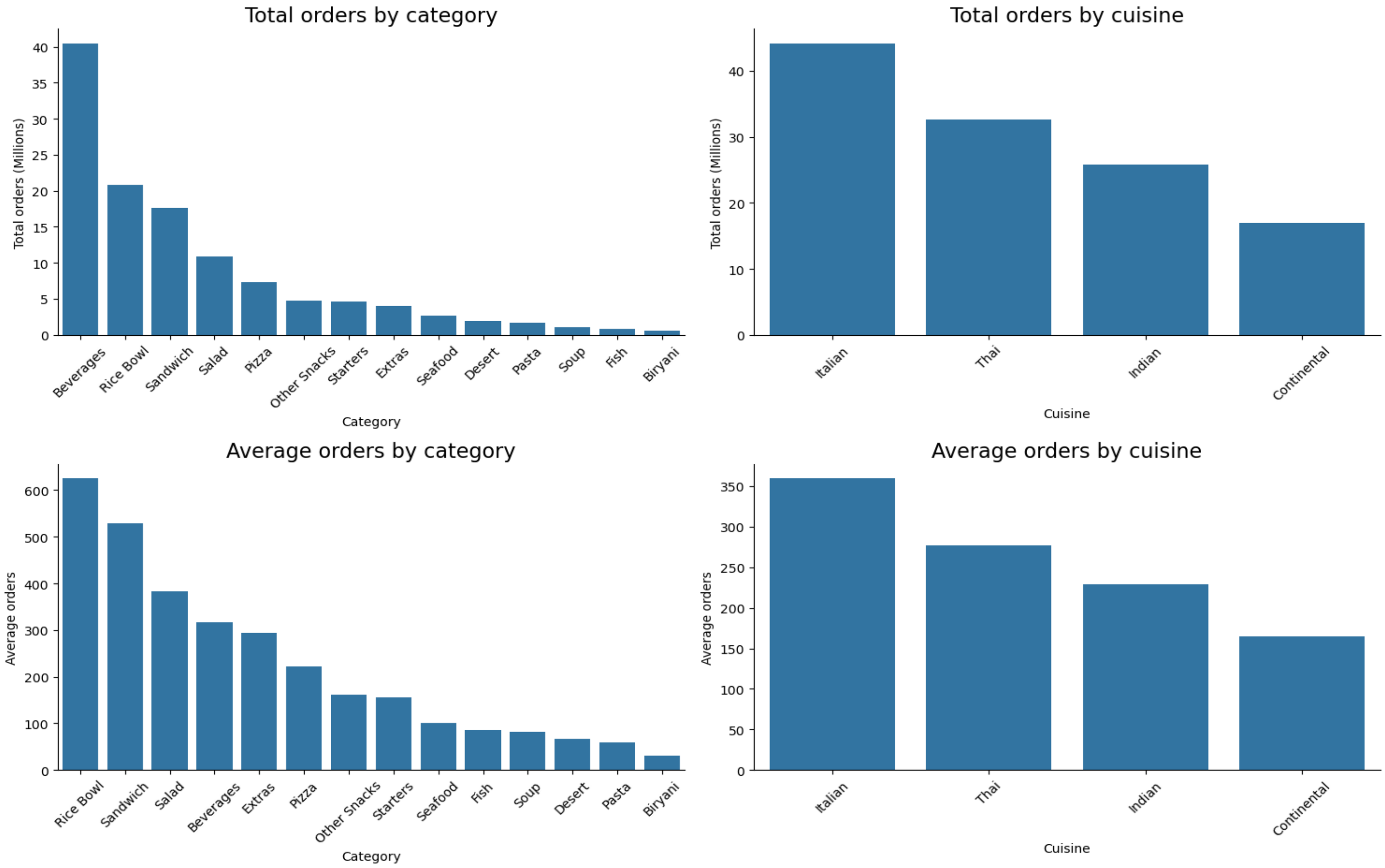


- **Promotion Impact Analysis**: Across all discount bins, the combination of **emailer** and **homepage feature** significantly boosted orders. Interestingly, the **30-50% discount** range generally resulted in the highest average orders, with diminishing returns observed beyond **50% discounts**.
- **Promotion Effectiveness by Week**: The weekly analysis confirmed that orders were higher during weeks with both promotions active. The **orange line** (both promotions) consistently exceeded the **blue line** (one or neither promotion), indicating the substantial impact of combined promotions on sales. The week-to-week fluctuations suggest varying levels of promotion success, with some weeks showing a more significant response than others.



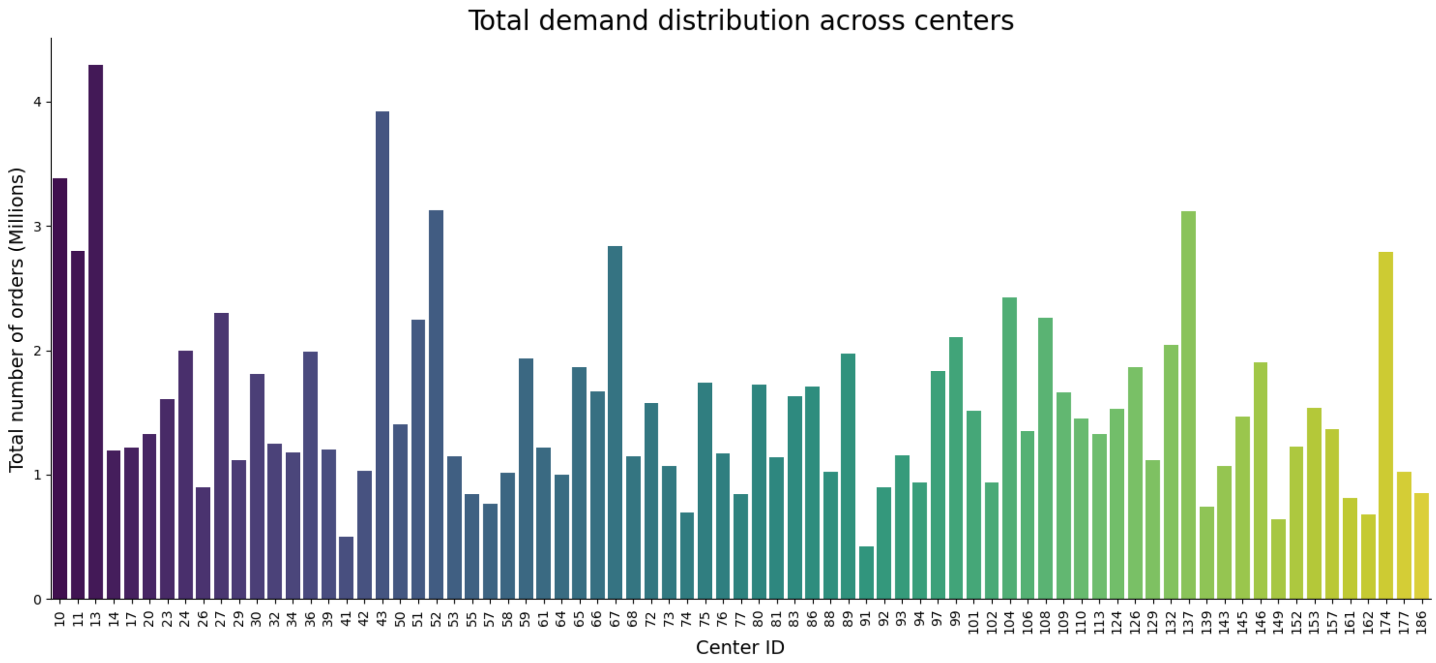
2.2.7 Category and Cuisine Analysis

- **Total Orders by Category**: **Beverages**, **Rice Bowls**, and **Sandwiches** dominated the orders, followed by **Salad** and **Pizza**. Categories such as **Biryani**, **Soup**, and **Fish** had much lower order volumes, indicating either limited popularity or less frequent promotions.
- **Total Orders by Cuisine**: **Italian** and **Thai** cuisines accounted for the highest number of orders, with **Italian** leading by a significant margin. **Indian**, **Continental**, and other cuisines had comparatively fewer orders, suggesting a preference for the top two cuisines in the dataset.
- **Average Orders by Category and Cuisine**: The patterns for average orders were similar to total orders, with **Rice Bowls** and **Sandwiches** leading in average orders per customer. Similarly, **Italian** and **Thai** cuisines also showed higher average orders, reinforcing the strong customer preference for these meal types.

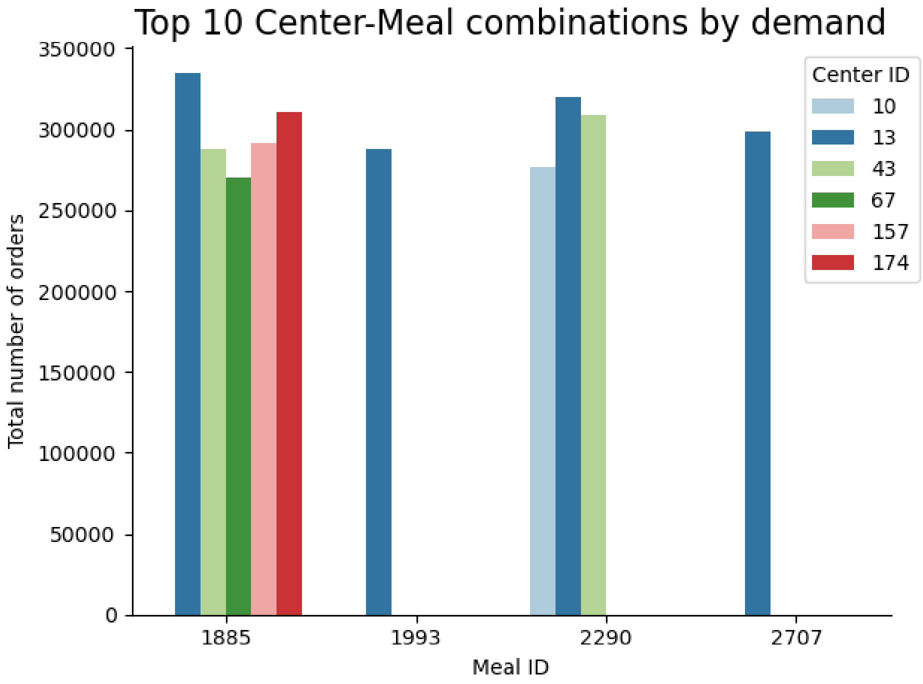


2.2.8 Center-Level Analysis

- **Demand Distribution Across Centers**: The demand distribution reveals significant variation across centers. Centers such as **Center 10**, **Center 43**, and **Center 132** exhibit high demand, while others like **Center 186** have lower demand. This disparity may reflect factors such as geographic location, service quality, or promotional efforts.



- **Center-Meal Combination Analysis**: Specific meal combinations show higher demand in certain centers. For instance, **Meal IDs 1885** (Beverages, Thai) and **1993** (Beverages, Thai) show the highest demand in **Center 10** and **Center 43**, while **Meal ID 2290** (Rice Bowl, Indian) performs well in **Center 67**. These insights can help identify which meal-center combinations need additional focus to optimize stock and staffing.



2.2.9 Correlation Matrix

- **Key Findings**:
    - Positive correlations between **num_orders** and promotional strategies (**emailer** and **homepage_featured**), as well as **discount**.

  - Negative correlations between **num_orders** and **checkout price** and **base price**, indicating that lower prices are associated with higher order volumes.

The correlation analysis provides insight into how promotions, pricing, and other factors interact, which will guide our modeling approach to accurately predict demand.

4. Modeling and Results

For predicting the number of orders, I implemented a regression approach, utilizing multiple models to ensure the most accurate prediction. I divided the dataset into training and testing sets, applying cross-validation on the training set and using the test set only for final model evaluation.

Feature Engineering

I created several new features, including:

- **Month** and **Quarter** features to capture seasonal trends.
- **Discount ratio** and **price difference** to quantify the effect of pricing.
- Interaction features such as:
  - **Price and discount interaction** to understand the relationship between pricing and discounts.
  - **Email promotion discount interaction** and **Homepage feature discount interaction** to explore how promotions and discounts impact order volume.
- **Log transformation** of the target variable, **number of orders**, to handle its skewed distribution and improve model performance.
- One-hot encoding was applied to all categorical features to enable their use in regression models.

Modeling Process

I first examined all available features and assessed their significance. Through feature selection, I removed the insignificant features, and used **Lasso Regression** to further shrink irrelevant features to zero. With a refined set of features, I proceeded with the modeling using the following approaches:

Model	R-squared	Mean absolute error
Multiple linear regression	0.5986	136.072
Ridge regression	0.599	136.05
Lasso regression	0.598	136.3
Random forest regressor	**0.786**	**76.5**
Gradient boosting regressor	0.6107	133.2

Multiple Linear Regression performed reasonably well in the initial tests, but it struggled with complex relationships in the data, making it less suitable for this problem where interactions and non-linearities are significant.

Ridge and Lasso Regression both showed similar results, offering a slight improvement over the linear regression model, but they were unable to handle the data's non-linear patterns as effectively as tree-based models.

Random Forest Regressor was by far the most successful model, with a significant improvement in performance, as indicated by its higher R² scores and its ability to capture complex interactions in the data. This model should be the primary choice for this type of predictive modeling task.

Gradient Boosting Regressor showed lower performance compared to Random Forest. It was less able to handle the complexity of the data, making Random Forest a more reliable choice.

Model Selection

After evaluating all models, I selected **Random Forest Regressor** as the best performer due to its ability to handle non-linearities and interactions effectively.

Final Model Evaluation:

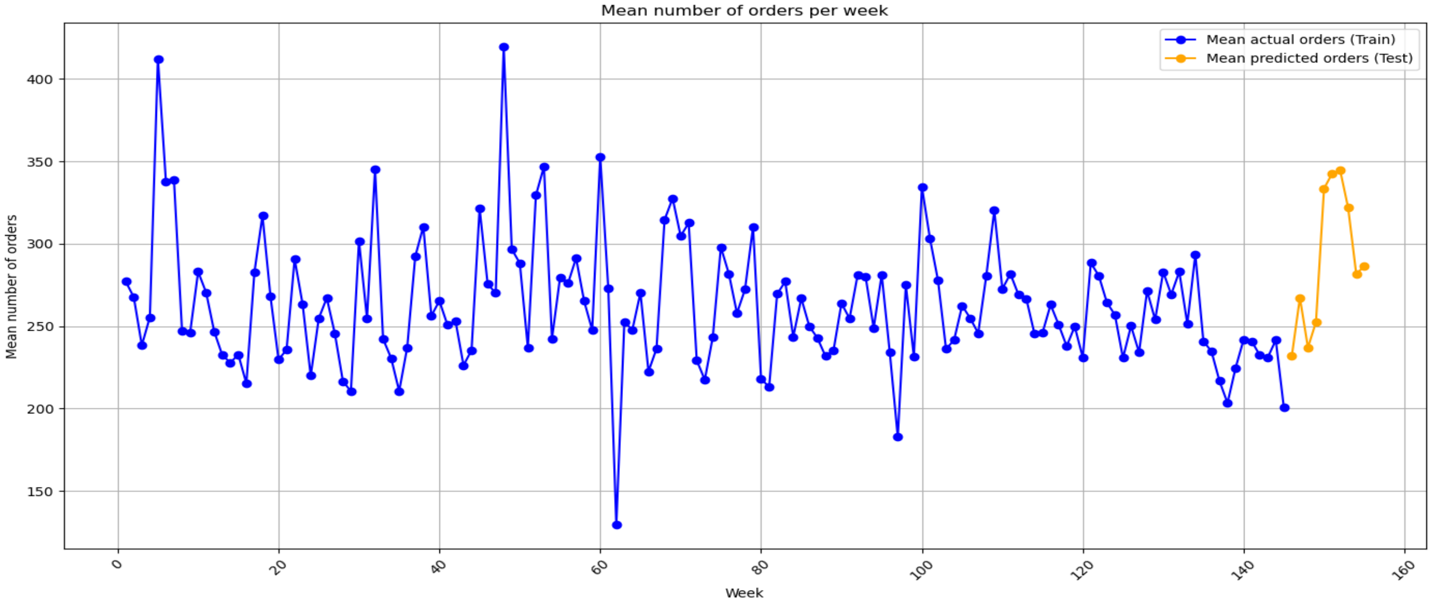
I validated the selected Random Forest model using the **test set**, achieving the following results:

- **Test Set R²**: 0.809
- **Test Set MAE**: 75.53

These results indicate that the model generalizes well to unseen data and provides reliable predictions for the number of orders.

Predictions for the Next 10 Weeks:

Using the **Random Forest** model, I predicted the **number of orders for the next 10 weeks**, which will assist in optimizing raw material stock and staffing plans for the meal delivery company.



Conclusion

The final model demonstrates strong performance, with an **R² of 0.809** and a **MAE of 75.53** on the test set. The feature engineering steps, including the creation of interaction features and the log transformation of the target variable, contributed significantly to the model's success. The **Random Forest Regressor** was the most effective model, capturing complex relationships in the data and providing accurate predictions. This model will be valuable for making data-driven decisions in inventory and staffing planning, ensuring that the company is prepared for future demand fluctuations.