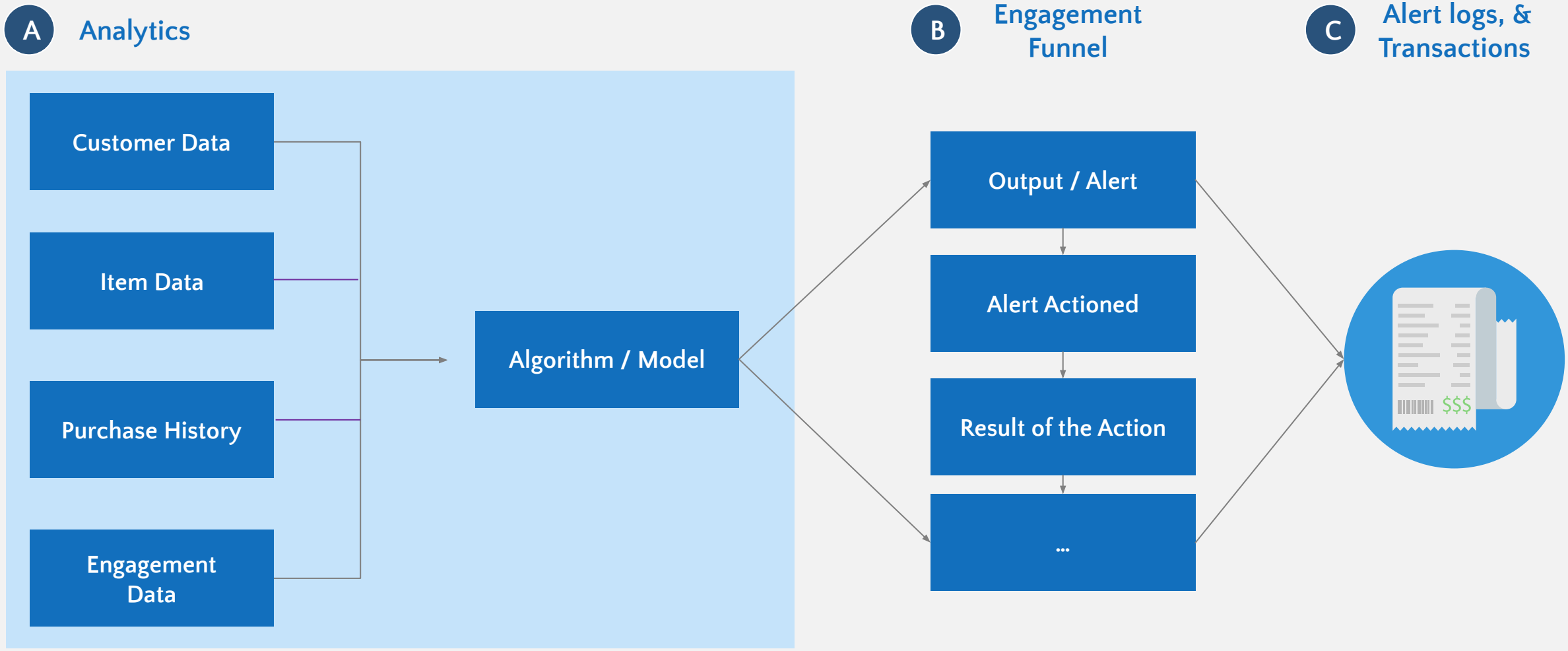
Advanced Predictive Modelling for Market Trend Analysis and Logistics Improvement

Business Case

High Level Project Framework

Below is an overview of the general steps we follow in a project. The first step is to develop the algorithm/model; those results are used to create alerts/outputs which are shared with the respective business unit and ultimately track engagement/benefit.



For the project we are going to focus on the Analytics part.

Basic statistics

- **Week:** The data spans from week 202027 to 202052, 202100 to 202152 and 202200 to 202216.
- **Customer Count:** Ranges from 20 to 536 with an average of about 176 customers per data point.
- **Cases:** Varies widely from about 177 to 36,498 cases, indicating significant variability in order volume.
- **Revenue:** Revenue ranges from a loss of approximately $298,467 to a gain of about $1,588,505. The mean revenue is around $271,453.
- **Cost:** Costs also show a wide range from about -$363,871 to $1,363,835, with an average cost around $224,027.

Given this overview, I have further analyzed trends, such as:

- How revenue and costs have evolved over time.
- Differences in performance between cities or types of cuisine.
- Analysis of product types and their impact on sales and profitability.

About the data:

Data Loading: Loaded data from the Excel file located on Google Drive using Pandas and checked the structure and summarized the information as below,

- City: Location of the customer data.
- Cuisine: Type of cuisine ordered.
- Product Type: Type of product ordered.
- Week: Specific week of data.
- Customer Count: Number of customers.
- Cases: Some measure, likely related to the amount of orders.
- Revenue: Revenue generated.
- Cost: Costs associated with the orders.

To get started, I have analyzed this data for insights such as:

- Total revenue and cost per city and cuisine.
- Trends over weeks in terms of customer count, revenue, and cost.
- Most popular products or cuisines.

Data Analysis:

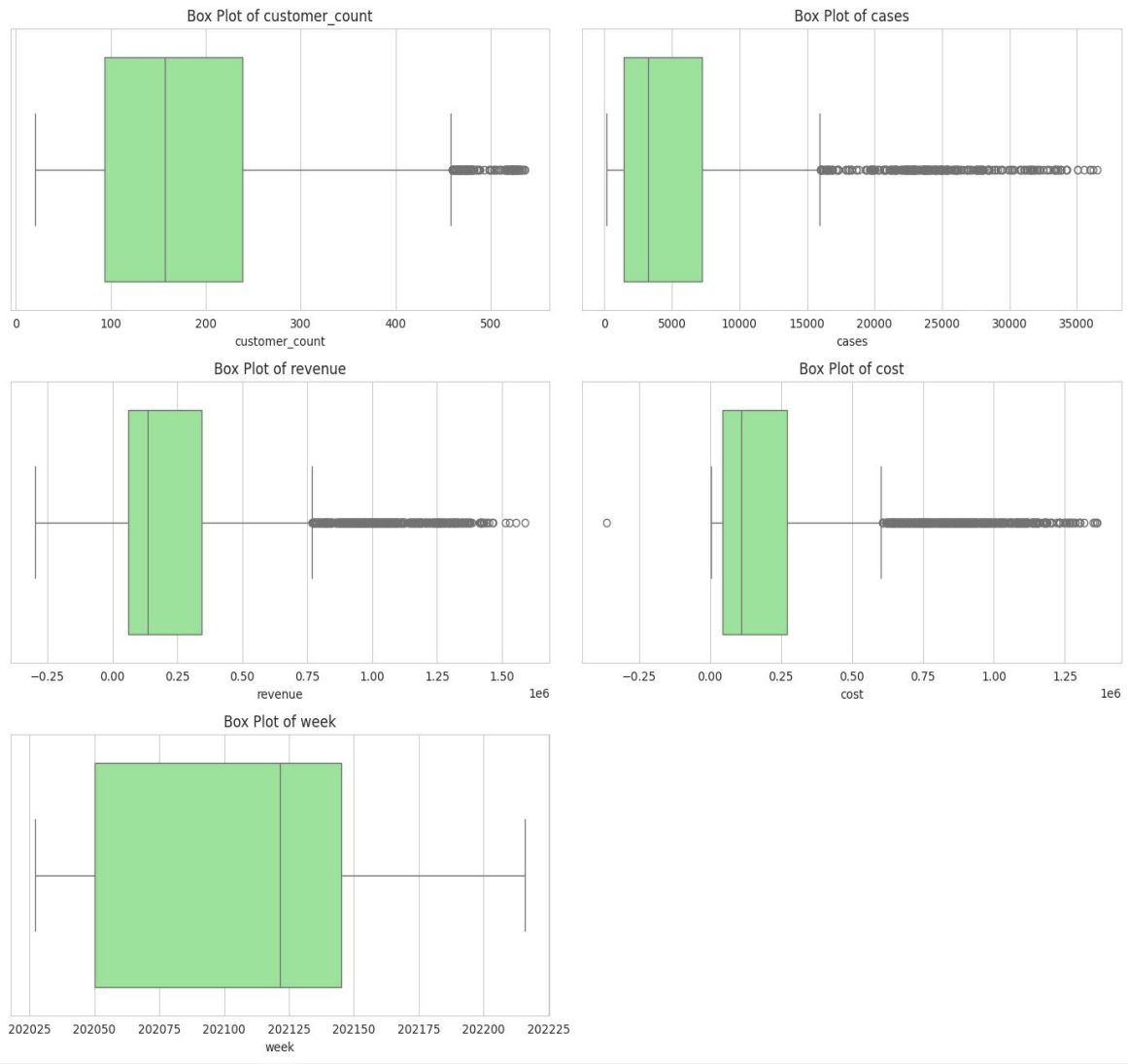
Outliers: Used visual methods like box plots to identify outliers in revenue, costs, and customer counts across different dimensions such as city and cuisine.

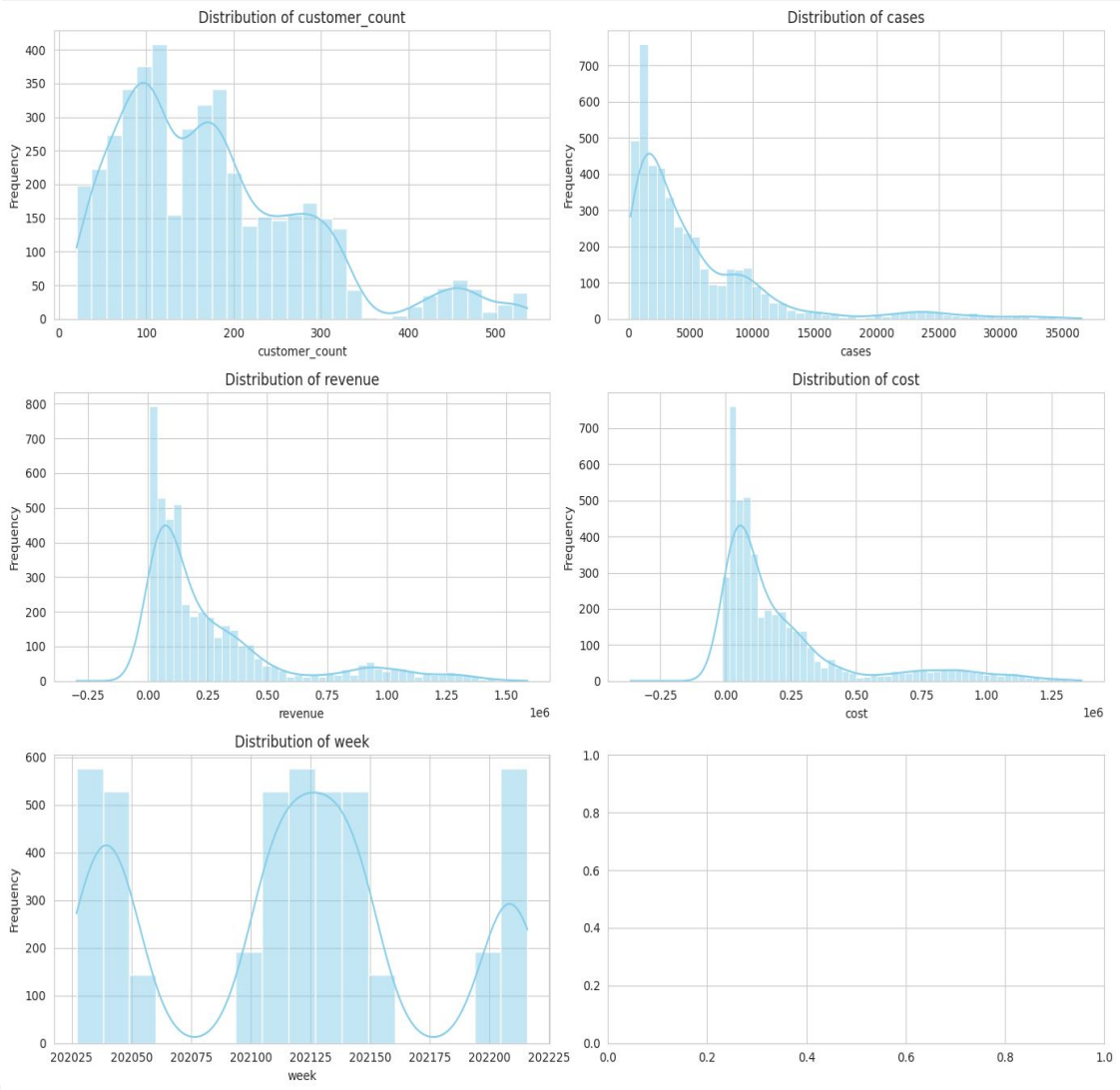
seasonality: Analyzed patterns across the weeks to see how customer behavior and sales metrics change over the year. This might reveal certain peaks or troughs that correspond to specific events or seasons.

Regression Analysis: Used statistical methods to predict the future outcomes based on historical data. This is particularly useful for forecasting sales, customer counts or revenue based on trends.

Profitability Analysis: Analyzed profit margins by looking at the difference between revenue and costs. This can be further detailed by segment to determine which products or locations offer the best returns on investment.

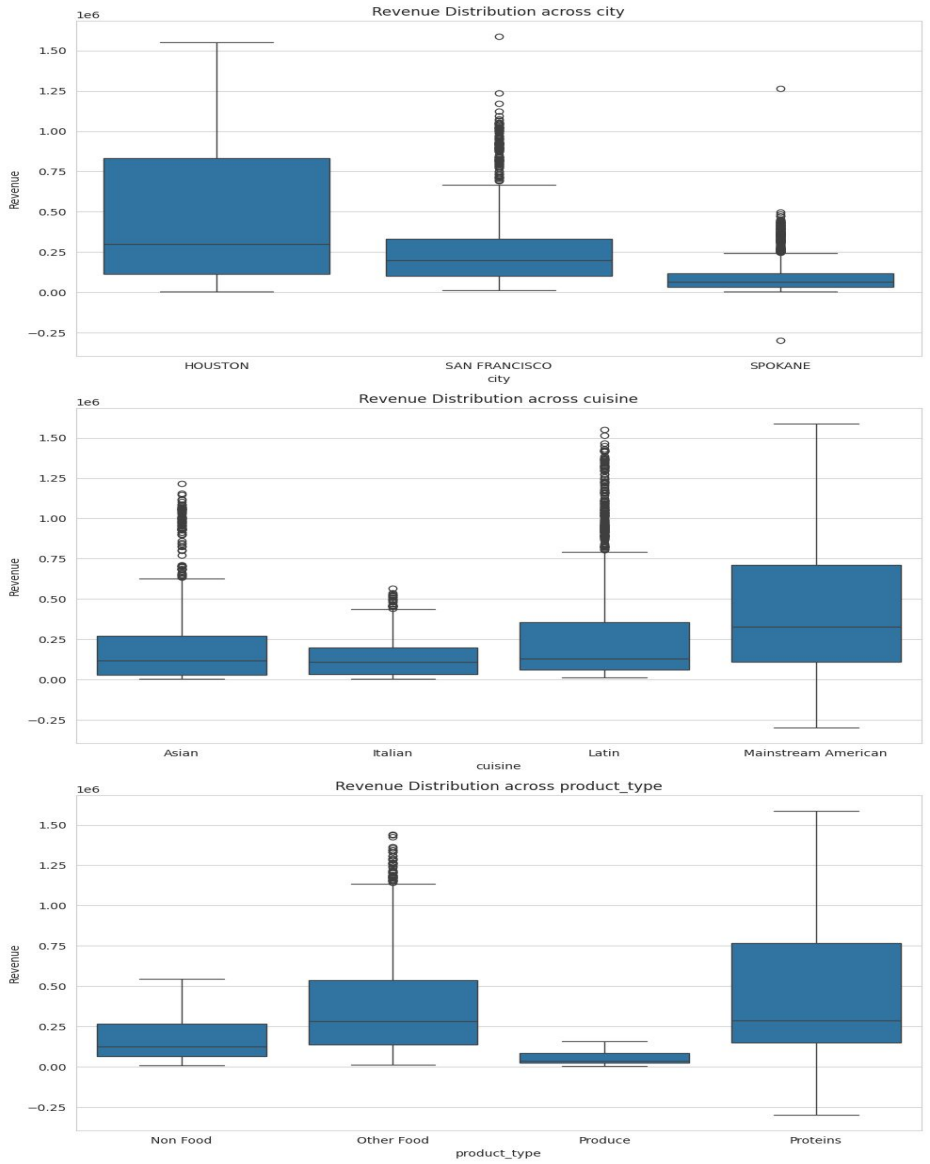
Results:


Box Plot of customer_count
customer_count
Box Plot of cases
cases
Box Plot of revenue
revenue
1e6
Box Plot of cost
cost
1e6
Box Plot of week
week


Distribution of customer_count
Frequency
customer_count
Distribution of cases
Frequency
cases
Distribution of revenue
Frequency
revenue
1e6
Distribution of cost
Frequency
cost
1e6
Distribution of week
Frequency
week

Results:


Weekly Revenue Trends
Average Revenue
350000
300000
250000
200000
150000
202025
202050
202075
202100
202125
202150
202175
202200
202225
Week


Revenue Distribution across city
1e6
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
Revenue
HOUSTON
SAN FRANCISCO
SPOKANE
city
Revenue Distribution across cuisine
1e6
Asian
Italian
Latin
Mainstream American
cuisine
Revenue Distribution across product_type
1e6
Non Food
Other Food
Produce
Proteins
product_type

Data Preprocessing and cleaning

Data Preprocessing

- Categorical Features Processing:
  - Identified categorical features: `city`, `cuisine`, `product_type`.
  - Applied `OneHotEncoder()` to transform these categorical variables into a format suitable for model input, creating binary columns for each category.
- Numerical Features Processing:
  - Integrated a custom `OutlierClipper` class utilizing the interquartile range to manage and reduce the impact of extreme outliers in the numerical features: week, customer_count, cases, and cost. This helps in maintaining data consistency and robustness.
  - Standardized the identified numerical features using `StandardScaler()` to normalize the data, ensuring equal contribution to the model's performance.

Data Cleaning

- This step is implicitly handled through the outlier clipping and the encoding of categorical variables.

Feature selection

- Selection Basis:
  a. Features are chosen based on domain knowledge, reflecting an understanding of factors believed to impact revenue.
  b. Categorical features include `city`, `cuisine`, and `product_type`.
  c. Numerical features include `week`, `customer_count`, `cases`, `cost`.
- Manual Inclusion:
  a. All selected features are manually included in the data preprocessing and modeling steps without prior statistical testing or evaluation for their predictive power.
- Integration into Models:
  a. Features are directly integrated into machine learning models through a preprocessing pipeline that handles scaling for numerical features and encoding for categorical features.

Model Selection, Evaluation and Performance analysis

Model Considered

- Three models are trained: Linear Regression, Decision Tree, and Random Forest Regressors.
- Models are evaluated using RMSE (Root mean Squared Error) to determine accuracy.
- Compared model performance with RMSE values.
- Random forest model has the low RMSE, i.e., Random Forest outperforms other models due to its ability to handle non-linear relationships and feature interactions effectively.

```
RMSE Linear Regression 94269.20014581826
RMSE Decision Tree Regression 49928.30513082444
RMSE Random Forest Regression 40869.29951601236
```

Predicting Future Revenue

Prediction Approach:

- The model predicts revenue for the next 12 weeks using data extrapolated from the last known data points in the test dataset, not by generating random values.

Data Utilization:

- The last 12 weeks of data from the test set are replicated to simulate future data points. This approach assumes that these weeks represent a relevant sample of near-future conditions.

Model Application:

- The trained Random Forest model, which performed best on historical data, is applied to these simulated future data points to generate revenue predictions.

Underlying Assumption:

- It's assumed that the conditions reflected in the last part of the test data continue similarly into the future, allowing the model to extrapolate based on these trends.

Practical Note:

- In real-world applications, this method would typically be enhanced by dynamically updating the dataset with actual ongoing data and potentially incorporating more sophisticated time-series forecasting techniques.

Assumptions and Open Questions

Exploratory Data Analysis (EDA)

- Assumptions: The distribution of numerical variables like 'customer_count', 'cases', and 'cost' directly impact revenue prediction. Outliers in data are present and managed by clipping to avoid skewing the model.
- Open Questions: Which additional variables could significantly influence revenue? How do outliers and data distribution nuances affect model accuracy?

Algorithm / Model Design

- Assumptions: Random Forest, Decision Tree, and Linear Regression can encapsulate the relationship between features and revenue. Historical data patterns will persist, allowing future revenue prediction.
- Open Questions: How to determine the most suitable algorithm for this dataset? Is the current model complexity optimal for the prediction task?

Model Performance

- Assumptions: RMSE is the chosen metric for performance, with the Random Forest model presumed to generalize best for future predictions.
- Open Questions: How might model performance vary with different data splits? What impact would feature selection and engineering have on the model?

Predicting Future Revenue

- Assumptions: The test set's last 12 weeks can be used to project the next 12 weeks' revenue. The model's extrapolation of unseen data is presumed accurate.
- Open Questions: How would unforeseen future events affect predictions? Should additional factors be incorporated for more accurate forecasting?