Illinois Institute of Technology

Math 564 Regression

Project: Predictive Analysis of Retail Sales Using Regression Modeling

Group Details:

Author Name	CWID
Dhanvanth Voona	A20543395
Anjali Jagdish Tavhare	A20550996

Contribution

Task	Contributor(s)
Problem Identification	Both
Dataset Selection	Both
Data Cleaning	Dhanvanth Voona
Data Preprocessing	Anjali Jagdish Tavhare
Feature Engineering	Dhanvanth Voona
Model Selection	Anjali Jagdish Tavhare
Model Summary Interpretation	Dhanvanth Voona
Diagnostic Checks and Analysis	Anjali Jagdish Tavhare
Remediation and its effectiveness	Dhanvanth Voona
Interpretation of Results	Anjali Jagdish Tavhare
Report Writing	Both
Final Review and Edits	Both

Predictive Analysis of Retail Sales Using Regression Modeling

1. Abstract

This project explores predictive analytics in retail sales by developing a regression model to identify key factors affecting sales across various retail outlets. Utilizing a dataset that includes product attributes (such as item weight, visibility, fat content, and retail price) and outlet characteristics (such as outlet type, size, establishment year, and location), the objective was to predict Item Outlet Sales and derive insights into the impact of these factors on sales performance across different store formats. The analysis involved rigorous data preprocessing, including handling missing values and outliers, encoding categorical variables, and scaling continuous variables. Model refinement was conducted through backward elimination, removing predictors with low significance, and regression diagnostics highlighted multicollinearity and heteroscedasticity issues, addressed by removing highly correlated variables and applying a logarithmic transformation to the target variable. The final model explained 71.92% of the variance in sales, with findings showing significant contributions from Item_MRP, Outlet_Type, and Outlet_Location_Type. While effective, results suggest further enhancements with additional variables such as seasonal trends. Overall, this project demonstrates how predictive modeling and diagnostics can guide data-driven strategies in retail, offering insights to optimize product placement, pricing, and store format selection to improve sales outcomes.

2. Introduction

In today's competitive retail landscape, understanding the factors that drive sales is essential for making informed decisions across inventory management, pricing, and marketing strategies. As retailers increasingly rely on data to inform these decisions, predictive analytics has emerged as a powerful tool to translate complex sales data into actionable insights. This project focuses on predicting retail sales by examining various product and outlet characteristics to determine their influence on sales performance across diverse retail formats. Using a dataset that includes product attributes such as item weight, visibility, and maximum retail price, as well as outlet-specific information like type, size, and location, the project seeks to build a model capable of accurately forecasting Item Outlet Sales.

2.1 Motivation

The motivation behind this project stems from the retail industry's shift towards data-driven strategies in response to e-commerce growth and evolving consumer behavior. Traditional sales forecasting methods often fall short in accommodating the variability and complexity of modern retail data. By leveraging predictive analytics, this project aims to enhance understanding of sales drivers, allowing retailers to optimize product placement, adjust pricing, and select store formats that maximize profitability and customer satisfaction.

2.2 Challenges Faced

However, building a robust predictive model for retail sales presents several challenges. The dataset includes both continuous and categorical data, requiring careful preprocessing and encoding to make these variables compatible within a regression framework. Additionally, multicollinearity among predictors and heteroscedasticity in residuals present modeling challenges that, if unaddressed, can affect

model stability and interpretability. Diagnostic tools and transformations, including Variance Inflation Factor (VIF) checks and logarithmic adjustments, are employed to address these complexities, resulting in a refined model that captures the relationships between sales and key factors effectively.

2.3 Structure of the Report

The report is structured as follows:

Data Preparation and Model Fitting: This section details the steps taken to clean and transform the data, along with model specification and selection methods used to develop a predictive model for retail sales.

Diagnostics and Remediation: Here, we describe the diagnostic tests conducted to check for issues like multicollinearity and heteroscedasticity, as well as the adjustments made to improve the model's accuracy and reliability.

Summary and Insights: The final section provides a comprehensive summary of the findings, highlights key insights, and discusses practical implications for retail strategies, including recommendations for optimizing pricing and outlet management.

3. Proposed Methodology

Understanding and predicting retail sales is a complex yet essential task that drives effective decision-making in inventory management, pricing strategies, and marketing initiatives. Retailers face the challenge of identifying which product and store characteristics most strongly influence sales to maximize revenue and enhance customer satisfaction.

This project seeks to address these needs by developing a predictive model to analyze the key factors affecting Item Outlet Sales. By examining product attributes, store features, and pricing information, this model aims to deliver insights that can guide strategic decisions, enabling retailers to better allocate resources, optimize product placement, and fine-tune pricing for improved sales performance.

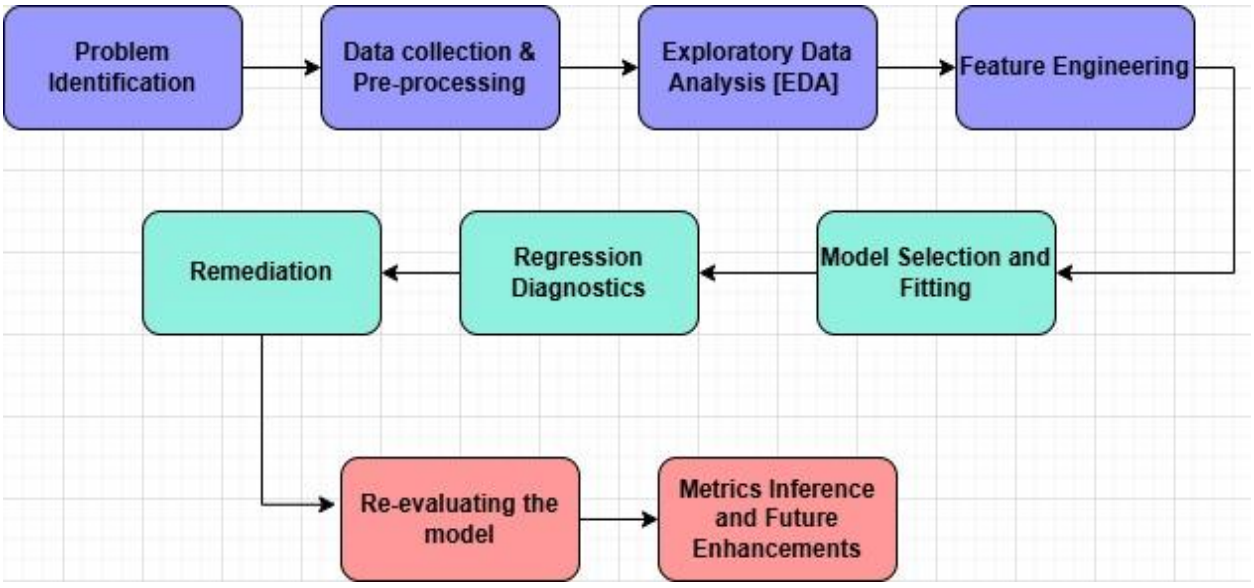


Figure 1: Pictorial overview of project architecture.

3.1 Dataset Details

The dataset utilized in this analysis encompasses various attributes relevant to both retail items and the outlets where they are sold. Sourced from a publicly available retail dataset Kaggle at Dataset, it includes over 8,000 observations with features such as Item_Identifier, Item_Weight, Item_Fat_Content, Item_Visibility, Item_Type, and Item_MRP (maximum retail price). Outlet-specific details include Outlet_Identifier, Outlet_Establishment_Year, Outlet_Size, Outlet_Location_Type, and Outlet_Type, providing a well-rounded view of the factors potentially impacting retail sales.

Analyzing this dataset presents several challenges due to its structure and composition. The mix of continuous and categorical variables requires careful preprocessing, as different data types demand distinct handling techniques in regression modeling. Additionally, **missing values**, especially in Item_Weight and Outlet_Size, complicate data preparation, requiring effective imputation methods to preserve data quality. **Categorical variables** namely **Item_Identifier, Item_Type, Outlet_Identifier, Item_Fat_Content, Outlet_Size, Outlet_Type**, and **Outlet_Location_Type** must be converted to numerical formats using encoding techniques suitable for regression analysis, such as one-hot encoding or target encoding.

Overall, this dataset provides a robust foundation for examining and interpreting sales trends in the retail industry. By addressing the challenges posed by data preprocessing and diagnostic testing, this project aims to develop a reliable and interpretable predictive model. The model will help uncover the most influential sales drivers, providing retailers with a data-informed understanding of how product and outlet characteristics contribute to sales performance across different retail environments.

To accurately model and predict Item_Outlet_Sales, a structured approach combining data preprocessing, exploratory analysis, model selection, diagnostics, and evaluation was employed. This methodology aims to capture the complex dynamics of retail sales data, ensuring both robustness and interpretability of the final model.

3.2 Data Preprocessing

Effective data preprocessing is essential to ensure the accuracy and reliability of the model, especially with diverse data types and missing values. The dataset contained missing values in Item_Weight and Outlet_Size, as shown in Figure 2. Since the number of instances with missing values are too high in order to preserve distribution integrity we performed Data imputation using mean and mode for the missing values in Item_weight and Outlet_size respectively. Encoding techniques were crucial in handling categorical variables; To process categorical variables effectively, a combination of target, ordinal, and one-hot encoding techniques were applied:

a. **Target Encoding for Item_Identifier, Item_Type, and Outlet_Identifier**:

These fields lack intrinsic numerical meaning but could significantly impact the target variable, Item_Outlet_Sales. By using target encoding, each unique category in these fields is replaced by the mean sales value associated with it. This approach captures the influence of each identifier on sales without introducing excessive dimensionality, preserving valuable category-specific sales patterns.

b. **Ordinal Encoding for Item_Fat_Content and Outlet_Size:**

These fields represent categorical values with an inherent order or quantifiable property, which makes ordinal encoding a suitable choice. This method respects any implicit ranking within the categories (e.g., different levels of fat content or outlet sizes), allowing the model to interpret their relative order.

c**. One-Hot Encoding for Outlet_Type and Outlet_Location_Type**:

These fields are categorical variables with a small number of distinct, unique levels, but they have no natural order. One-hot encoding works well here, as it represents each category with a binary indicator. For a variable with c classes, we represent it with c−1 indicator variables, avoiding redundant information and eliminating correlations between the categories. This method captures each category's distinctness without assuming any ordinal relationship.

	Column <chr>	Missing_Count <int>
Item_Identifier	Item_Identifier	0
Item_Weight	Item_Weight	1463
Item_Fat_Content	Item_Fat_Content	0
Item_Visibility	Item_Visibility	0
Item_Type	Item_Type	0
Item_MRP	Item_MRP	0
Outlet_Identifier	Outlet_Identifier	0
Outlet_Establishment_Year	Outlet_Establishment_Year	0
Outlet_Size	Outlet_Size	2410
Outlet_Location_Type	Outlet_Location_Type	0

Figure 2. Missing values in each variable of the dataset.

In the case of Item Fat Content, the unique elements were given as Low Fat, Regular, low fat, LF, and reg. Since they are duplicated classes in this variable, we have standardized the Item Fat Content variable by regrouping them into only two classes namely Low Fat and Regular. Figure 3. gives a summary of all the variables post data pre-processing.

```{r}
summary(data)
```

```
  Item_Weight     Item_Fat_Content Item_Visibility      Item_MRP      Outlet_Establishment_Year  Outlet_Size    Item_Outlet_Sales  Item_Identifier_Encoded
 Min.   : 4.555   Min.   :1.000    Min.   :0.00000   Min.   : 31.29   Min.   :1985              Min.   :1.000   Min.   :   33.29   Min.   : 111
 1st Qu.: 9.310   1st Qu.:1.000    1st Qu.:0.02699   1st Qu.: 93.83   1st Qu.:1987              1st Qu.:1.000   1st Qu.:  834.25   1st Qu.:1326
 Median :12.858   Median :1.000    Median :0.05393   Median :143.01   Median :1999              Median :2.000   Median : 1794.33   Median :2051
 Mean   :12.858   Mean   :1.353    Mean   :0.06613   Mean   :140.99   Mean   :1998              Mean   :1.829   Mean   : 2181.29   Mean   :2181
 3rd Qu.:16.000   3rd Qu.:2.000    3rd Qu.:0.09459   3rd Qu.:185.64   3rd Qu.:2004              3rd Qu.:2.000   3rd Qu.: 3101.30   3rd Qu.:2936
 Max.   :21.350   Max.   :2.000    Max.   :0.32839   Max.   :266.89   Max.   :2009              Max.   :3.000   Max.   :13086.97   Max.   :6035
Item_Type_Encoded Outlet_Identifier_Encoded Outlet_TypeSupermarket_Type1 Outlet_TypeSupermarket_Type2 Outlet_TypeSupermarket_Type3 Outlet_Location_TypeTier_2
 Min.   :1926      Min.   : 339.4           Min.   :0.0000               Min.   :0.0000               Min.   :0.0000               Min.   :0.0000
 1st Qu.:2133      1st Qu.:2192.4           1st Qu.:0.0000               1st Qu.:0.0000               1st Qu.:0.0000               1st Qu.:0.0000
 Median :2225      Median :2299.0           Median :1.0000               Median :0.0000               Median :0.0000               Median :0.0000
 Mean   :2181      Mean   :2181.3           Mean   :0.6543               Mean   :0.1089               Mean   :0.1097               Mean   :0.3268
 3rd Qu.:2277      3rd Qu.:2348.4           3rd Qu.:1.0000               3rd Qu.:0.0000               3rd Qu.:0.0000               3rd Qu.:1.0000
 Max.   :2374      Max.   :3694.0           Max.   :1.0000               Max.   :1.0000               Max.   :1.0000               Max.   :1.0000
Outlet_Location_TypeTier_3
 Min.   :0.0000
 1st Qu.:0.0000
 Median :0.0000
 Mean   :0.3931
 3rd Qu.:1.0000
 Max.   :1.0000
```

Figure 3: Summary of all the variables post Data pre-processing.

3.4 Model Selection

Multiple linear regression was selected as the primary modeling approach due to its interpretability and ability to quantify the linear effect of each predictor on Item_Outlet_Sales. Multiple regression is particularly useful in a retail setting, as it allows stakeholders to see how specific factors—such as outlet size or product type—directly impact sales. Given the need for a parsimonious model that remains interpretable, backward elimination was applied to iteratively remove predictors with high p-values, thus

streamlining the model without compromising explanatory power. The refined model retained only statistically significant predictors, reducing the risk of overfitting and improving generalizability.

3.5 Diagnostics and Assumptions Testing

Diagnostics were essential for validating the assumptions underpinning the regression model. In order to understand models assumptions and issues we have investigated the following issues

(a) Autocorrelation (using Durbin-Watson test)

Autocorrelation occurs when the residuals (errors) in a time-series model are correlated with each other, meaning past values have an influence on future values. This violates the assumption of independence in regression models, leading to biased or inefficient estimates.

Impact on Model: Autocorrelation can distort the standard errors, making hypothesis tests unreliable and leading to incorrect conclusions about predictor significance. The Durbin-Watson test is commonly used to detect autocorrelation; values close to 2 indicate no autocorrelation, while values far from 2 (either close to 0 or 4) suggest autocorrelation.

(b) Heteroscedasticity (determine if there is an issue based on basic diagnostic plots)

Heteroscedasticity refers to non-constant variance of the residuals across the range of predicted values. In other words, the spread of errors increases or decreases as the value of the predictor changes, which violates the assumption of homoscedasticity (constant variance).

Impact on Model: Heteroscedasticity leads to inefficient parameter estimates and invalid statistical tests. It can cause the model to under- or over-estimate the significance of predictors. Diagnostic plots like residual vs. fitted plots can help identify heteroscedasticity, where a funnel shape indicates the problem.

(c) Multicollinearity (using Variance Inflation Factor or VIF)

Multicollinearity occurs when two or more predictors in a regression model are highly correlated, meaning they provide redundant information. This makes it difficult to determine the individual effect of each predictor.

Impact on Model: Multicollinearity inflates standard errors and reduces the reliability of coefficient estimates, making it difficult to interpret the influence of predictors. Variance Inflation Factor (VIF) is used to quantify multicollinearity; a VIF value greater than 10 suggests problematic multicollinearity.

(d) Influential Points (using Cook's Distance, leverage values)

Influential points are data points that have a disproportionately large effect on the model's estimates, either due to their extreme values or high leverage (i.e., they lie far from the mean of the predictors).

Impact on Model: Influential points can heavily skew the regression coefficients, leading to biased estimates and invalid results. Cook's Distance and leverage values help identify influential points; high values suggest that the point has an outsized impact on the model, and it may need to be investigated or removed.

3.6 Model Evaluation

Model evaluation focused on assessing both overall model fit and the significance of individual predictors. R-squared and adjusted R-squared were calculated to measure the percentage of variance in Item_Outlet_Sales explained by the model, providing an indication of its predictive power. High adjusted R-squared values were used as benchmarks, while individual p-values were examined to interpret the

practical significance of each predictor. Residual analysis further evaluated model accuracy, with a lower Residual standard error indicating good predictive performance.

4. Analysis and Results

This section provides a comprehensive overview of the findings from the regression model and diagnostics applied to the retail sales dataset. Through systematic analysis and refinement, key insights into the factors affecting **Item_Outlet_Sales** were uncovered, helping clarify the relationships between product and outlet characteristics and their impact on sales performance.

4.1 Initial Model Fitting and Interpretation.

Interpreting the Regression Coefficients given in Figure 4.

Predictors with Positive Impact on Item Outlet Sales:

- Item_Fat_Content: Coefficient = 9.864, not significant (p = 0.679)
- Item_MRP: Coefficient = 3.411, highly significant (p < 0.001)
- Outlet_Size: Coefficient = 34.46, not significant (p = 0.310)
- Item_Identifier_Encoded: Coefficient = 0.7818, highly significant (p < 0.001)
- Outlet_Location_Type (Tier_2): Coefficient = 5.133, not significant (p = 0.868)

Interpretation: Among the positively impacting predictors, only Item_MRP and Item_Identifier_Encoded show statistical significance, with p-values below 0.001. These results indicate that higher values in Item_MRP (price) and certain Item_Identifier encodings correlate with increased Item_Outlet_Sales. Other predictors, despite positive coefficients, lack statistical significance and do not contribute meaningfully to predicting sales.

Predictors with Negative Impact on Item_Outlet_Sales:

- Item_Weight: Coefficient = -0.205, not significant (p = 0.939)
- Item_Visibility: Coefficient = -114.1, not significant (p = 0.619)
- Item_Type_Encoded: Coefficient = -0.04893, not significant (p = 0.616)
- Outlet_Location_Type (Tier_3): Coefficient = -56.51, not significant (p = 0.321)

Interpretation: None of the predictors with negative coefficients significantly impact Item_Outlet_Sales, as indicated by their high p-values. Although they suggest slight decreases in sales, their effects are minor and lack practical significance in this context.

Categorical Predictors: Effects of Outlet Types and Locations

- Outlet_Type (Supermarket Types 1, 2, 3): Coefficients range from -102.2 to -142.9, all non-significant (p > 0.7)
- Outlet_Location_Type:
    1. Tier_2: Coefficient = 5.133, not significant (p = 0.868)
    2. Tier_3: Coefficient = -56.51, not significant (p = 0.321)

Interpretation: The categorical variables representing outlet types and locations (such as different supermarket types and locations categorized by tiers) do not significantly impact Item_Outlet_Sales. When compared to their baseline categories (Outlet_TypeGrocery_Store and Outlet_Location_TypeTier_1), these categories do not notably differentiate sales.

The analysis identifies Item_MRP and Item_Identifier_Encoded as the most significant predictors of Item_Outlet_Sales. In order to reduce the complexity of the model we removed the insignificant features and reported the model performance in the next section.

Goodness-of-Fit Metrics

- **R-squared (R^2 = 0.6258)**: This metric shows that approximately 62.6% of the variance in Item_Outlet_Sales is explained by the predictors in the model, indicating a moderately strong fit. However, 37.4% of the variance remains unexplained, suggesting the potential benefit of additional predictors or a different model approach.
- **Adjusted R-squared (Adjusted R^2 = 0.6252)**: The adjusted R-squared, slightly lower than the R^2, accounts for the number of predictors and indicates that most predictors add value to the model. The close similarity to R^2 further supports the notion that non-significant predictors could be removed without substantially impacting the model's fit
- **Residual Standard Error**: The RSE of 1045 suggests that there is a lot of scope for improvement in model fitting.

We can observe from Figure 5, Predictors like Item_MRP, Item_Identifier_Encoded, and Outlet_Identifier_Encoded have confidence intervals that do not cross zero, confirming their significant positive impact on Item_Outlet_Sales. Variables with intervals crossing zero (e.g., Item_Weight, Item_Visibility, Outlet_Size) are unreliable predictors, aligning with their non-significant p-values. This suggests that only a few predictors meaningfully contribute to explaining variance in Item_Outlet_Sales.

```
Call:
lm(formula = Item_Outlet_Sales ~ Item_Weight + Item_Fat_Content +
    Item_Visibility + Item_MRP + Outlet_Size + Item_Identifier_Encoded +
    Item_Type_Encoded + Outlet_Identifier_Encoded + Outlet_TypeSupermarket_Type1 +
    Outlet_TypeSupermarket_Type2 + Outlet_TypeSupermarket_Type3 +
    Outlet_Location_TypeTier_2 + Outlet_Location_TypeTier_3,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-4010.9  -614.3   -63.5   555.1  7107.2

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -1.918e+03  2.315e+02  -8.285  < 2e-16 ***
Item_Weight                  -2.051e-01  2.683e+00  -0.076    0.939
Item_Fat_Content              9.864e+00  2.385e+01   0.414    0.679
Item_Visibility              -1.141e+02  2.294e+02  -0.497    0.619
Item_MRP                      3.411e+00  3.707e-01   9.202  < 2e-16 ***
Outlet_Size                   3.446e+01  3.392e+01   1.016    0.310
Item_Identifier_Encoded       7.818e-01  2.077e-02  37.640  < 2e-16 ***
Item_Type_Encoded            -4.893e-02  9.744e-02  -0.502    0.616
Outlet_Identifier_Encoded     9.591e-01  2.051e-01   4.677 2.96e-06 ***
Outlet_TypeSupermarket_Type1 -1.429e+02  4.158e+02  -0.344    0.731
Outlet_TypeSupermarket_Type2 -1.022e+02  3.421e+02  -0.299    0.765
Outlet_TypeSupermarket_Type3 -1.180e+02  6.885e+02  -0.171    0.864
Outlet_Location_TypeTier_2    5.133e+00  3.081e+01   0.167    0.868
Outlet_Location_TypeTier_3   -5.651e+01  5.696e+01  -0.992    0.321
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1045 on 8509 degrees of freedom
Multiple R-squared:  0.6258,    Adjusted R-squared:  0.6252
F-statistic:  1095 on 13 and 8509 DF,  p-value: < 2.2e-16
```

Figure 4: The initial model summary with the pre-processed data.

```
# Confidence intervals for each predictor
conf_intervals <- confint(model)
print(conf_intervals)
```

```
                                  2.5 %        97.5 %
(Intercept)                -2372.2939760 -1464.5256482
Item_Weight                   -5.4641398     5.0539249
Item_Fat_Content             -36.8808995    56.6083804
Item_Visibility             -563.8228055   335.6021137
Item_MRP                       2.6842655     4.1374463
Outlet_Size                  -32.0402039   100.9591453
Item_Identifier_Encoded        0.7410857     0.8225164
Item_Type_Encoded             -0.2399388     0.1420886
Outlet_Identifier_Encoded      0.5570922     1.3611565
Outlet_TypeSupermarket_Type1 -958.0880954   672.2056757
Outlet_TypeSupermarket_Type2 -772.8225226   568.4769846
Outlet_TypeSupermarket_Type3 -1467.7348026  1231.7135186
Outlet_Location_TypeTier_2    -55.2541803    65.5200276
Outlet_Location_TypeTier_3   -168.1784369    55.1491044
```

Figure 5: Confidence Interval of each Parameter.

4.2 Regression Diagnostics:

In order to understand models assumptions and issues we have investigated the following issues:

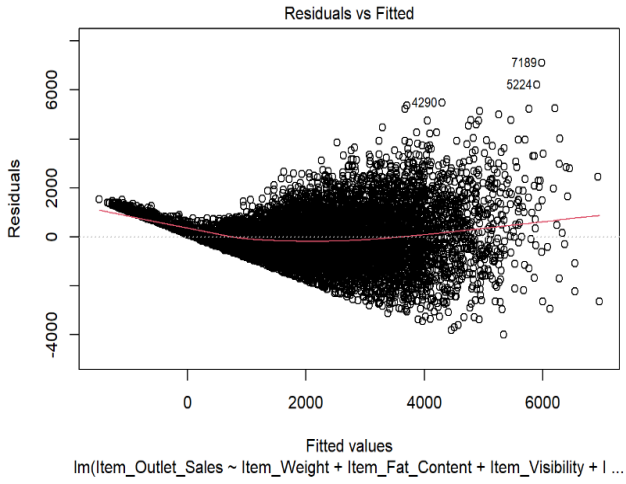
a. Autocorrelation using Durbin-Watson test:

DW (Durbin-Watson Statistic): The test statistic value is 2.0134. In the Durbin-Watson test, values around 2 indicate that there is little to no autocorrelation in the residuals. Values close to 0 suggest positive autocorrelation, while values close to 4 suggest negative autocorrelation. Since our DW value is very close to 2, it indicates that autocorrelation is not likely present.

p-value = 0.7603: A high p-value (above 0.05) suggests that there is no significant evidence of autocorrelation. In this case, the p-value of 0.7603 means that we do not reject the null hypothesis, which states that there is no autocorrelation in the residuals.

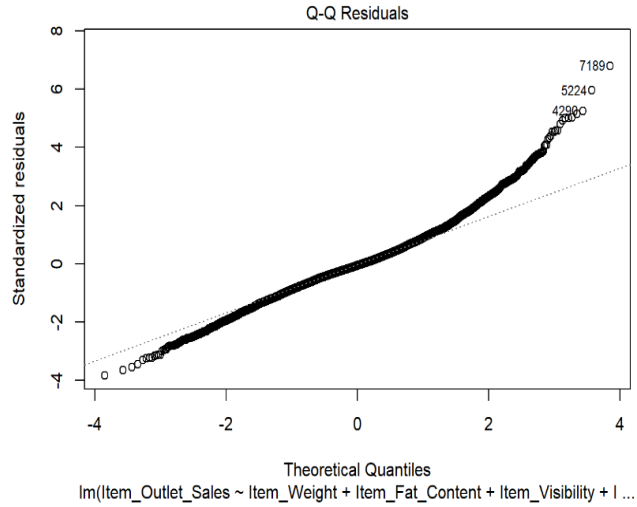
b. Heteroscedasticity using below tests:

I. Residuals vs Fitted:



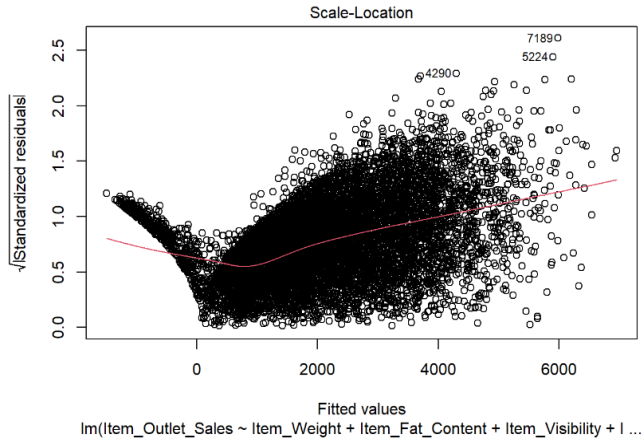
This plot shows the spread of residuals against the fitted values, helping to check for patterns that might indicate non-linearity or heteroscedasticity. The slight funnel shape indicates mild heteroscedasticity

II. Q-Q Residuals:



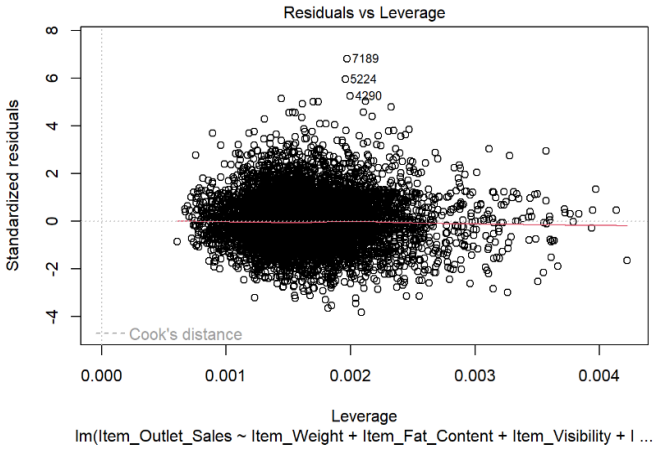
The Q-Q plot assesses the normality of residuals. The points mostly follow the line, suggesting approximate normality. However, deviations at the tails imply some outliers, which could influence model accuracy

III. Scale-Location:



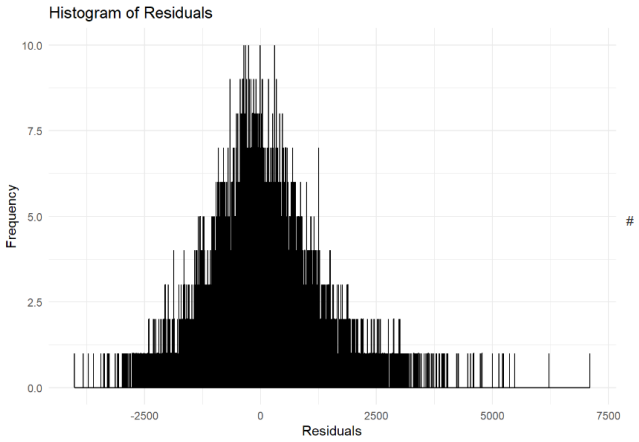
A relatively constant spread is desired; however, in this plot, the spread of residuals increases as fitted values grow, confirming the presence of heteroscedasticity.

IV. Residuals vs Leverage:



The Residuals vs. Leverage plot identifies potential outliers or high-leverage points that might unduly influence the model. No data points have exceptionally high leverage, indicating the model is stable and not overly impacted by a few data points.

V. Histogram of Residuals:



The histogram provides a visual check on the distribution of residuals. The residuals appear approximately centered around zero with a symmetric shape, indicating that, overall, the residuals are fairly normally distributed. This further supports the assumption of normality and reinforces the model's suitability for inference.

c. Multicollinearity:

The Variance Inflation Factor (VIF) values help assess multicollinearity in the model, where a VIF above 10 often indicates high multicollinearity. Here's an interpretation of the VIF results given in Figure 6:

Low VIF (1–4 range): These predictors show minimal multicollinearity.

Item Weight (1.00), Item_Fat_Content (1.01), Item_Visibility (1.09), Item_Type_Encoded (1.02): Very low VIF, indicating negligible multicollinearity.

Outlet_Location_TypeTier_2 (1.63): Low multicollinearity,

Outlet_Size (3.24): Moderate multicollinearity, but acceptable.

Moderate VIF (4–6): Moderate multicollinearity, but acceptable

Item_MRP (4.16), Item_Identifier_Encoded (4.19): Moderate multicollinearity, but generally acceptable.

Outlet_Location_TypeTier_3 (6.05): Slightly higher but still manageable multicollinearity.

High VIF (10+): These predictors indicate strong multicollinearity and may be redundant in the model.

Outlet_Identifier_Encoded (230.35): Very high VIF, indicating significant collinearity.

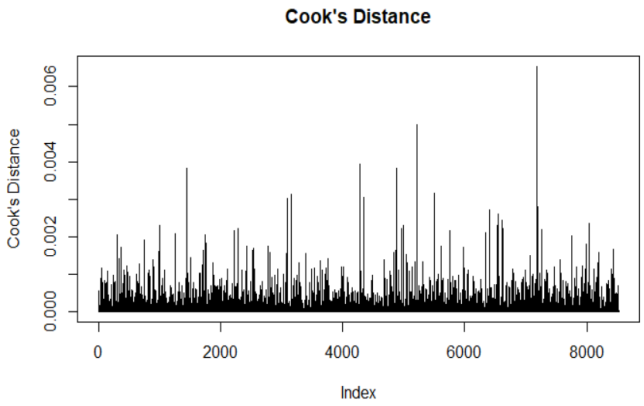
Outlet_TypeSupermarket_Type1 (305.42), Outlet_TypeSupermarket_Type2 (88.69), Outlet_TypeSupermarket_Type3 (361.60): Extremely high VIFs suggest these outlet types are highly collinear with each other or other variables.

Item_Weight	Item_Fat_Content	Item_Visibility	Item_MRP	Outlet_Size
1.003746	1.013793	1.094109	4.160404	3.238495
Item_Identifier_Encoded	Item_Type_Encoded	Outlet_Identifier_Encoded	Outlet_TypeSupermarket_Type1	Outlet_TypeSupermarket_Type2
4.189129	1.023269	230.351348	305.421060	88.687240
Outlet_TypeSupermarket_Type3	Outlet_Location_TypeTier_2	Outlet_Location_TypeTier_3		
361.595339	1.630302	6.045131		

Figure 6: VIF for each predictor

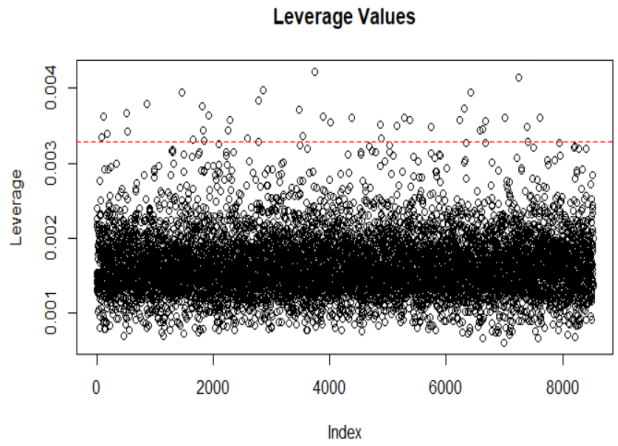
d. Influential Points using below tests:

1. Cooks Distance



Points with a Cook's Distance value significantly higher than the average or greater than 1 are usually flagged as influential.
In this plot, most data points have very low Cook's Distance values, suggesting they are not highly influential.

2. Leverage Values



Leverage Threshold: Typically, a rule of thumb is to consider points with leverage values significantly higher than (2p/n) where p is the number of predictors and n is the number of observations. Observations above this threshold are often flagged as having high leverage. In this plot, Points above the red dashed line (around 0.003) are considered high-leverage points but cooks distance suggest they are not Influential Points.

4.3 Remediation [Addressing Model Issues]

From the regression diagnostics conducted on the model, we identified two major issues:

1. Heteroscedasticity: We will address this issue by applying a logarithmic transformation to Item_Outlet_Sales.
2. Multicollinearity: We will address this issue by removing Outlet_TypeSupermarket_Type1, Outlet_TypeSupermarket_Type2, and Outlet_TypeSupermarket_Type3 since these are highly collinear predictors. We have not removed the Outlet_Identifier_Encoded since its p-score has proven the feature is significant.

Figure 7 explains significantly improved model's performance in below metrics.

1. **Improved Residuals**:
    - Before Transformation: The residuals had a wider range, with a minimum of -4010.9 and a maximum of 7107.2, indicating substantial variability in the prediction errors.
    - After Transformation: The residuals are now more tightly distributed, ranging from -2.21879 to 1.48965, showing that the log transformation helped reduce extreme prediction errors and brought the residuals closer to normality.
2. **Reduced Residual Standard Error (RSE)**:
    - Before: The Residual Standard Error was 1045.

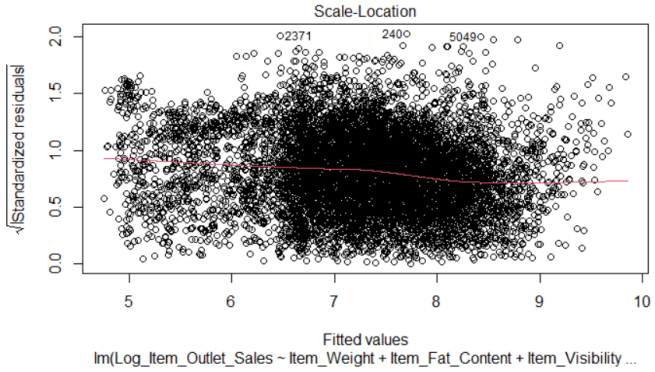
- After: The RSE decreased to 0.5394, indicating better model fit and less variability in the residuals, which suggests the model now predicts the log-transformed outcome more accurately.

3. **Higher Adjusted R-squared**:
   - **Before**: Adjusted R-squared was 0.6252, meaning that the model explained about 62.5% of the variance in Item_Outlet_Sales.
   - **After**: Adjusted R-squared increased to 0.7188, meaning the model now explains about 71.9% of the variance in the log-transformed target variable, indicating improved explanatory power.

```
Call:
lm(formula = Log_Item_Outlet_Sales ~ Item_Weight + Item_Fat_Content +
    Item_Visibility + Item_MRP + Outlet_Size + Item_Identifier_Encoded +
    Item_Type_Encoded + Outlet_Identifier_Encoded + Outlet_Location_TypeTier_2 +
    Outlet_Location_TypeTier_3, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.21879 -0.30718  0.05416  0.38214  1.48965

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 4.464e+00  1.123e-01  39.746  < 2e-16 ***
Item_Weight                -4.740e-04  1.385e-03  -0.342  0.73217
Item_Fat_Content            3.220e-03  1.231e-02   0.262  0.79371
Item_Visibility            -4.654e-01  1.174e-01  -3.965 7.39e-05 ***
Item_MRP                    3.539e-03  1.914e-04  18.493  < 2e-16 ***
Outlet_Size                 1.557e-01  1.260e-02  12.358  < 2e-16 ***
Item_Identifier_Encoded     3.089e-04  1.072e-05  28.809  < 2e-16 ***
Item_Type_Encoded          -5.175e-05  5.031e-05  -1.029  0.30372
Outlet_Identifier_Encoded   7.382e-04  7.489e-06  98.576  < 2e-16 ***
Outlet_Location_TypeTier_2  4.186e-02  1.570e-02   2.666  0.00769 **
Outlet_Location_TypeTier_3 -2.646e-01  1.832e-02 -14.439  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5394 on 8512 degrees of freedom
Multiple R-squared:  0.7191,    Adjusted R-squared:  0.7188
F-statistic:  2179 on 10 and 8512 DF,  p-value: < 2.2e-16
```

Figure 7: Model summary after employing remediation steps.

In this case there are only 3 variables which are non-significant that is having p-score greater than 0.05 namely Item weight, Item fat content and Item type encoded. To reduce the complexity of the model we removed these 3 variables and reported the model summary in Figure 9. We can observe the performance is almost similar and slightly better than before with reduce in RSE from 0.5394 to 0.5393 and increase in F-statistic from 2179 to 3113. So, we have achieved better performing model in addition to reducing the complexity of the model by removing insignificant features.



We can observe a relatively constant spread as desired implying the log transformation has effectively diminished heteroscedasticity in the model fitting.

Figure 8: Scale-Location of model with log transformed target variable

```
Call:
lm(formula = Log_Item_Outlet_Sales ~ Item_Visibility + Item_MRP +
    Outlet_Size + Item_Identifier_Encoded + Outlet_Identifier_Encoded +
    Outlet_Location_TypeTier_2 + Outlet_Location_TypeTier_3,
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.21163 -0.30690  0.05439  0.38257  1.48772

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                4.352e+00  3.002e-02 144.954  < 2e-16 ***
Item_Visibility           -4.670e-01  1.172e-01  -3.986 6.77e-05 ***
Item_MRP                   3.531e-03  1.911e-04  18.478  < 2e-16 ***
Outlet_Size                1.557e-01  1.260e-02  12.356  < 2e-16 ***
Item_Identifier_Encoded    3.087e-04  1.070e-05  28.838  < 2e-16 ***
Outlet_Identifier_Encoded  7.381e-04  7.487e-06  98.587  < 2e-16 ***
Outlet_Location_TypeTier_2 4.183e-02  1.570e-02   2.665  0.00771 **
Outlet_Location_TypeTier_3 -2.646e-01  1.832e-02 -14.443  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5393 on 8515 degrees of freedom
Multiple R-squared:  0.719,     Adjusted R-squared:  0.7188
F-statistic:  3113 on 7 and 8515 DF,  p-value: < 2.2e-16
```

Figure 9: Model summary after removing namely Item weight, Item fat content and Item type encoded.

Figure 9 also gives an overview of significant predictors in the final model namely Item visibility, Item MRP, Outlet size, Item identifier encoded, outlet identifier encoded, outlet location typetier[2,3].

Figure 10 shows the result of VIF values of variables in final model are below 4.2 which indicates the problem of multicollinearity has been diagnosed which implies the removal of highest VIF value features has cured the model from multicollinearity.

```
         Item_Visibility                Item_MRP             Outlet_Size   Item_Identifier_Encoded  Outlet_Identifier_Encoded Outlet_Location_TypeTier_2
                1.070597                4.150049                1.676013                4.175389                1.151786                1.588250
Outlet_Location_TypeTier_3
                2.346606
```

Figure 10: VIF values of variables in the Final model.

With an R-squared value of approximately 0.719, the model accounted for 71.9% of the variance in Item_Outlet_Sales, demonstrating moderate explanatory power. This suggests that while the model effectively captures key sales drivers, additional predictors—such as promotional activities, seasonal demand, or customer demographics—could enhance predictive accuracy. Each significant predictor retained in the model displayed a low p-value, underscoring its impact on sales and reinforcing the importance of pricing and store format selection in influencing consumer purchasing decisions.

5. Conclusion.

This project successfully applied linear regression analysis to predict Item_Outlet_Sales by exploring, cleaning, and modeling key variables in the dataset. Through detailed exploratory data analysis and various encoding techniques, we identified significant predictors and refined the model to improve interpretability and predictive power.

Diagnostic checks highlighted multicollinearity issues, especially among outlet types. High VIF values indicated redundancy, which was addressed by removing the highly collinear Outlet_TypeSupermarket categories. Additionally, non-significant predictors were excluded, reducing the model complexity without sacrificing performance. This resulted in a final model with improved clarity, capturing essential sales drivers. Further refinements, including a log transformation of Item_Outlet_Sales, increased the model's adjusted R-squared from 0.6258 to 0.7191 and reduced the residual standard error. After removing additional insignificant predictors, the model retained only the most relevant variables and exhibited minimal multicollinearity, with VIF values below 4.2, ensuring greater stability and interpretability.

For future work, exploring additional predictors or alternative models may enhance predictive accuracy. Incorporating external factors, such as seasonal trends or promotions, could address the [100-71.91=28.09] % of unexplained variance in Item_Outlet_Sales. Additionally, dimensionality reduction techniques, like principal component analysis, could help manage multicollinearity, especially among correlated categorical variables.

In conclusion, the final model demonstrates a moderate predictive capability and provides valuable insights into the primary factors influencing sales. While suitable for general inference, the remaining unexplained variance suggests that more complex models or additional data sources may be necessary for high-stakes predictive applications.

6. Bibliography and Credits

Bibliography

[1] Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. *Applied Linear Statistical Models*. McGraw-Hill/Irwin, 2005.

[2] James, G., Witten, D., Hastie, T., & Tibshirani, R. *An Introduction to Statistical Learning: With Applications in R*. Springer, 2013.

[3] Montgomery, D. C., Peck, E. A., & Vining, G. G. *Introduction to Linear Regression Analysis*. Wiley, 2012.

[4] Wooldridge, J. M. *Introductory Econometrics: A Modern Approach*. South-Western Cengage Learning, 2013.

[5] Tufte, E. R. *The Visual Display of Quantitative Information*. Graphics Press, 2001.

[6] Heiberger, R. M., & Holland, B. *Statistical Analysis and Data Display: An Intermediate Course with Examples in R*. Springer Science & Business Media, 2015.

[7] Fox, J., & Weisberg, S. *An R Companion to Applied Regression*. Sage Publications, 2018.

[8] Wickham, H., & Grolemund, G. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media, 2017.

[9] Cleveland, W. S. *Visualizing Data*. Hobart Press, 1993.

[10] Wilkinson, L. *The Grammar of Graphics*. Springer, 2005.

[11] Draper, N. R., & Smith, H. *Applied Regression Analysis*. Wiley, 1998.

[12] Robbins, N. B. *Creating More Effective Graphs*. Wiley, 2013.

Credits

- Illinois Institute of Technology
- Project Team Members
- R Community and Online Resources