Blinkit Grocery Data Analysis

BY

SANDRA GANESHAN

Introduction

The **Blinkit** dataset provides information about various items sold in retail outlets, including details on item fat content, sales, ratings, and outlet characteristics. The dataset is valuable for understanding sales trends, outlet performance, and product categorization. The analysis aims to uncover patterns and insights related to product sales, outlet types, and factors influencing the overall retail environment. This data is critical for businesses to make informed decisions on inventory, marketing, and sales strategies.

Objective:

- **Outlet Demographics Analysis**: Understanding the distribution of outlets by location type, size, and establishment year.
- **Item Type Insights**: Analyzing different item types and their sales performance to understand customer preferences.
- **Fat Content Impact**: Investigating how different fat content categories (e.g., Low Fat vs. Regular) influence sales and customer purchasing patterns.
- **Sales and Item Visibility**: Studying the relationship between item visibility and sales to identify potential marketing opportunities.
- **Outlet Performance**: Analyzing the sales performance by outlet and establishment year to identify trends and growth opportunities.

Aim

- **Optimize Sales Strategies**: By identifying trends in item types, fat content categories, and outlet performance, Blinkit can tailor its product offerings and marketing strategies to enhance sales.
- **Enhance Customer Understanding**: Analyzing item visibility and outlet characteristics helps Blinkit understand customer preferences, enabling better product placement and promotional strategies.
- **Assess Outlet Performance**: By examining sales trends across different outlet types and locations, Blinkit can assess the performance of each outlet and optimize its expansion and location strategies.
- **Inform Product and Pricing Decisions**: Insights from sales data, including variations by fat content and item type, can guide pricing strategies and inventory management to align with customer demand.

Data Cleaning & Preparation

Standardization:

- The 'Item Fat Content' column was standardized to combine similar categories: "low fat" and "LF" were replaced with "Low Fat," and "reg" was replaced with "Regular."

Null Value Handling:

- Checked for null values in the dataset. Columns like 'Item Weight' were dropped due to over 1000 missing values, which were deemed irrelevant for analysis.

Space Removal:

- Leading and trailing spaces were removed from string columns, including 'Item Identifier', 'Item Type', and 'Outlet Type.'

Duplicate Checking:

- Duplicate rows were identified, with the number of duplicates counted to ensure data uniqueness.

Feature Engineering:

- A new feature 'Year_count' was created to calculate the number of years since the outlet was established.
- 'sales_by_outlet' was created to calculate the total sales per outlet, while 'Avg Rating by Outlet' and 'Avg Sales by Item Type' were calculated using groupby operations.

Outlier Detection:

- Using the Interquartile Range (IQR) method, outliers in the 'Sales' column were identified and listed.

Univariate Analysis

Statistical Summary

```
for column in num_l:
    print(f"Statistical Functions for '{column}':")
    print("-------------------------------------")
    print(f"Minimum: {df[column].min()}")
    print(f"Maximum: {df[column].max()}")
    print(f"Mean: {df[column].mean():.2f}")
    print(f"Mode: {df[column].mode()[0]}")  # Mode returns a series;
take the first value
    print(f"Median: {df[column].median()}")
    print(f"Standard Deviation: {df[column].std():.2f}")
    print(f"Variance: {df[column].var():.2f}")
    print("_____________________________\n")
```

- **Sales Variability**: There is significant variability in sales, indicating the need for tailored marketing or operational strategies per outlet.
- **Item Visibility**: A large portion of items has low visibility, suggesting a possible opportunity to improve product placement or promotion.
- **Customer Satisfaction**: High ratings across the board reflect overall customer satisfaction, but minor improvements in visibility could enhance it further.

Boxplot for Numerical columns

```
for i in num_l:
  plt.boxplot(df[i])
  plt.title(f"box plot of {i}")
  plt.show()
```

- **Outliers**: Outliers in **Item Visibility**, **Rating**, **Sales by Outlet**, and **Avg Sales by Item Type** indicate areas for further investigation.
- **Consistency**: Some variables, like **Outlet Establishment Year** and **Avg Rating by Outlet**, show relatively consistent data with little variation.
- **Sales and Visibility**: The boxplots indicate that sales and visibility vary significantly across outlets and items, with specific products or outlets standing out as outliers.

Histplot for Numerical Columns

```
for i in num_l:
  plt.figure(figsize=(5,4))
  plt.hist(df[i], bins=7,color='blue')
```

```
  plt.title(f"hist plot of {i}")
  plt.xlabel(f"{i}")
  plt.ylabel('Count')
  plt.show()
```

Categorical Column Value count

```
for col in obj_l:
    print("univariate analysis of categorical column:")
    print(df[col].value_counts())
    print(f'Number of unique categories:{df[col].nunique()}')
    print("_____________________________________________________")
    print()
```

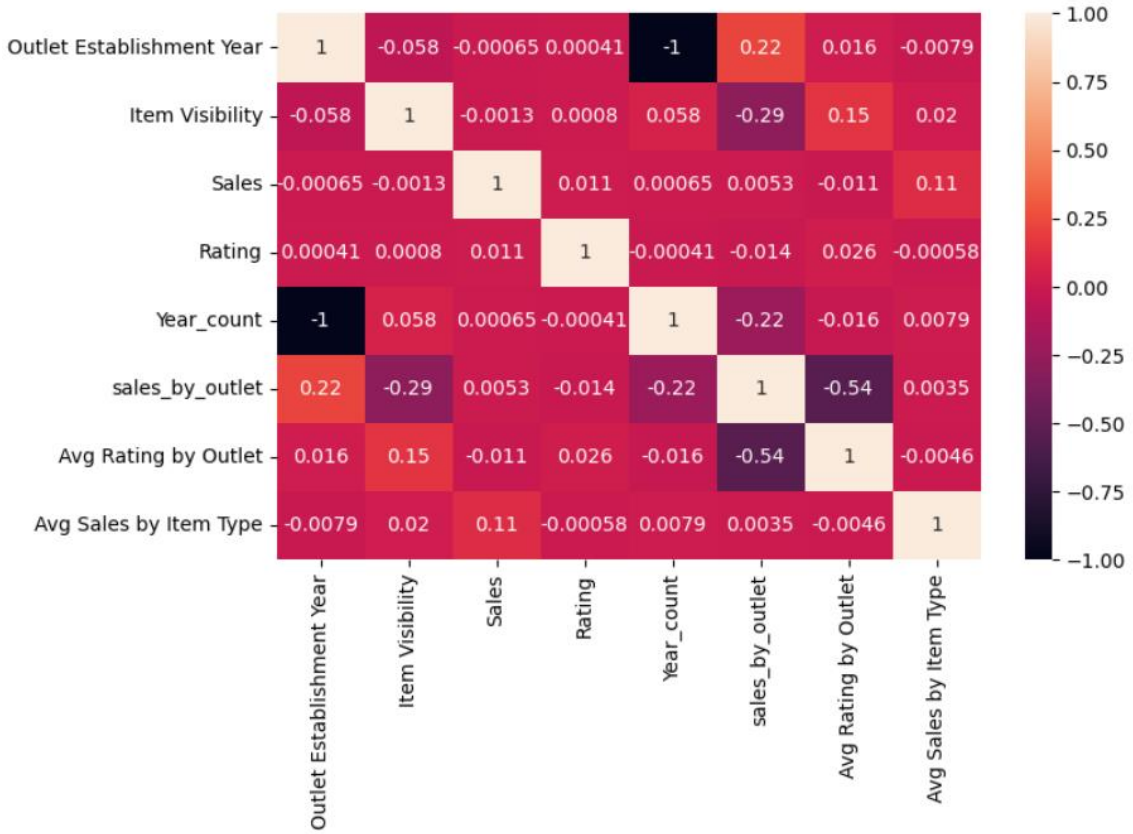
- Item Fat Content: Low Fat (5,517) is more common than Regular (3,006).
- Item Identifier: 1,559 unique identifiers, with frequent items appearing 9-10 times.
- Item Type: Fruits & Vegetables (1,232) and Snack Foods (1,200) are most common.
- Outlet Identifier: 10 outlets, with OUT027, OUT013, and OUT049 being the most frequent.
- Outlet Location Type: Tier 3 (3,350) has the most outlets.
- Outlet Size: Medium (3,631) is the most common.
- Outlet Type: Supermarket Type1 (5,577) is the most common.

Barplot for categorical column(Also known as Count plot)

```
for i in obj_l:
  if df[i].nunique()<=17:  #Unique categories
    plt.figure(figsize=(5,3))
    plt.bar(x=df[i].value_counts().index,height=df[i].value_counts())
    plt.title(f'Count plot of {i} Category')
    plt.xlabel(i)
    plt.ylabel("Count")
    plt.xticks(rotation=90)
    plt.show()
```

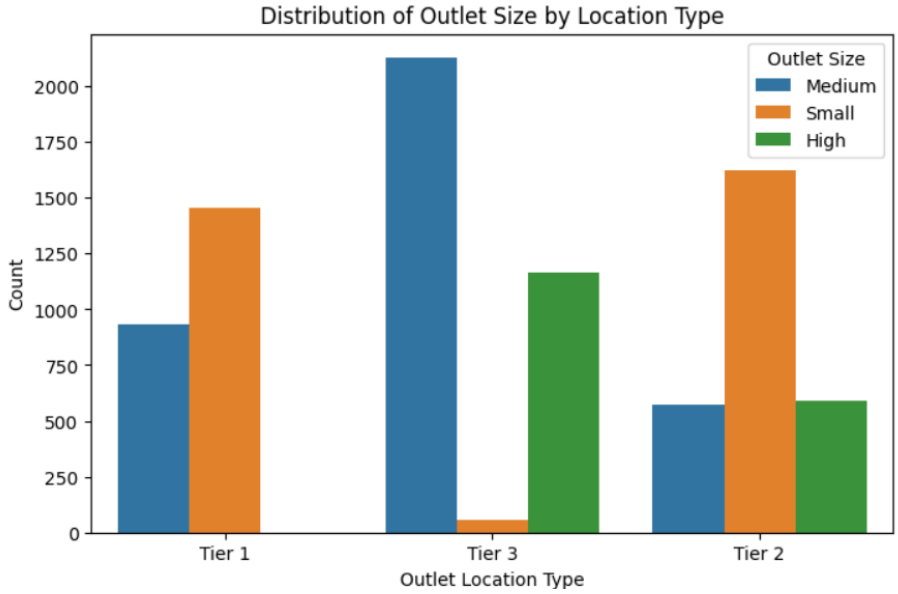
Bivariate Analysis

Heatmap



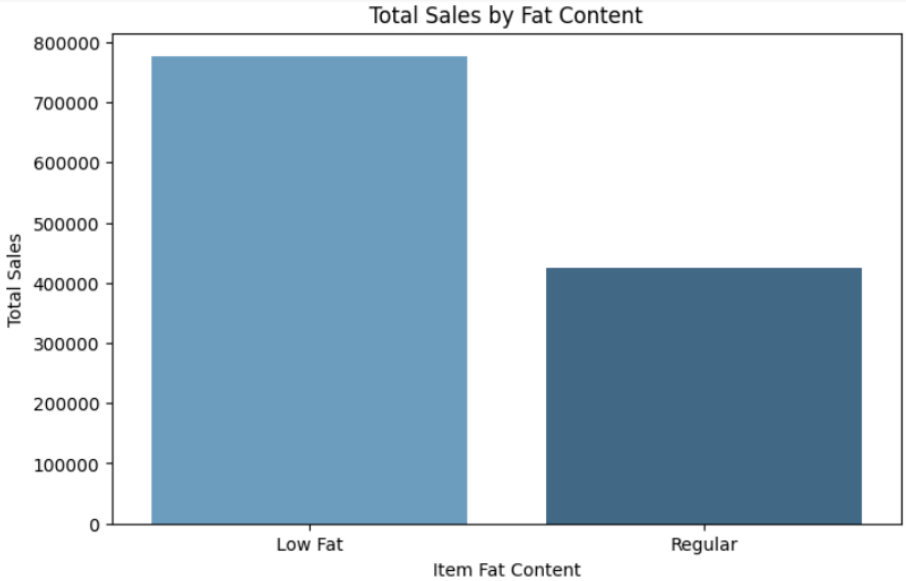
- Year Count & Outlet Establishment Year(-1): This is because "Year Count" is derived from "Outlet Establishment Year".
- Item Visibility shows weak correlations with most variables, suggesting it has minimal direct linear relationships with other numerical features.
- As the outlet gets older (Year Count increases), sales tend to slightly decrease, but the weak correlation suggests other factors likely have a stronger impact on sales.

Relationship between Outlet Location Type and Outlet Size(Count Plot)



- Tier 1 has more Small outlets than Medium outlets.
- Tier 2 has a more balanced distribution across High, Medium, and Small outlets.
- Tier 3 has a higher number of Medium and High outlets, but very few Small outlets.

Impact of Fat Content on Total Sales(Barplot)

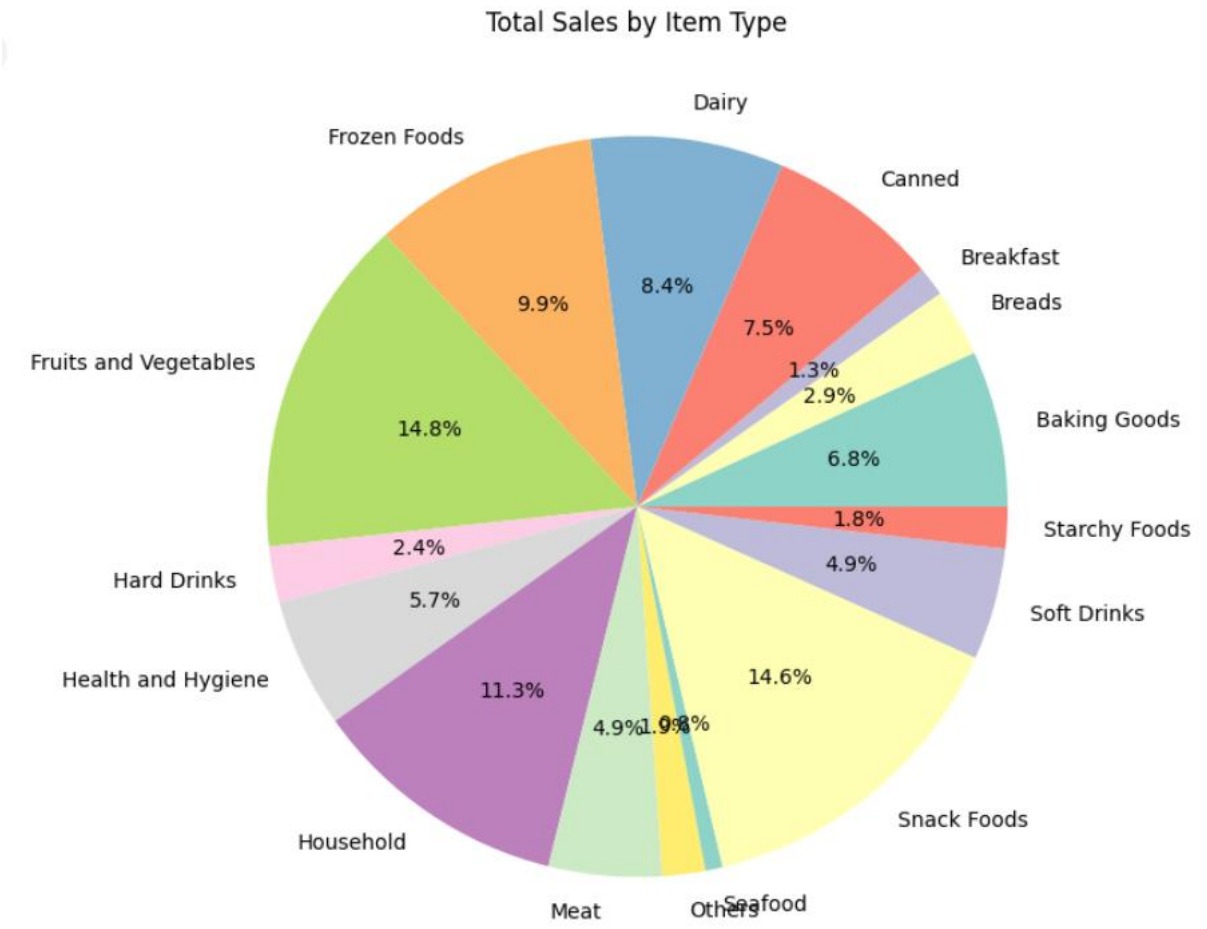


Item Fat Content	count	mean	std	min	25%	50%	\
Low Fat	5517.0	140.71410	62.311756	31.2900	94.4778	142.0812	
Regular	3006.0	141.50426	62.214786	32.0558	92.2212	144.0444	

Item Fat Content	75%	max
Low Fat	185.0898	266.8884
Regular	186.5214	266.8884

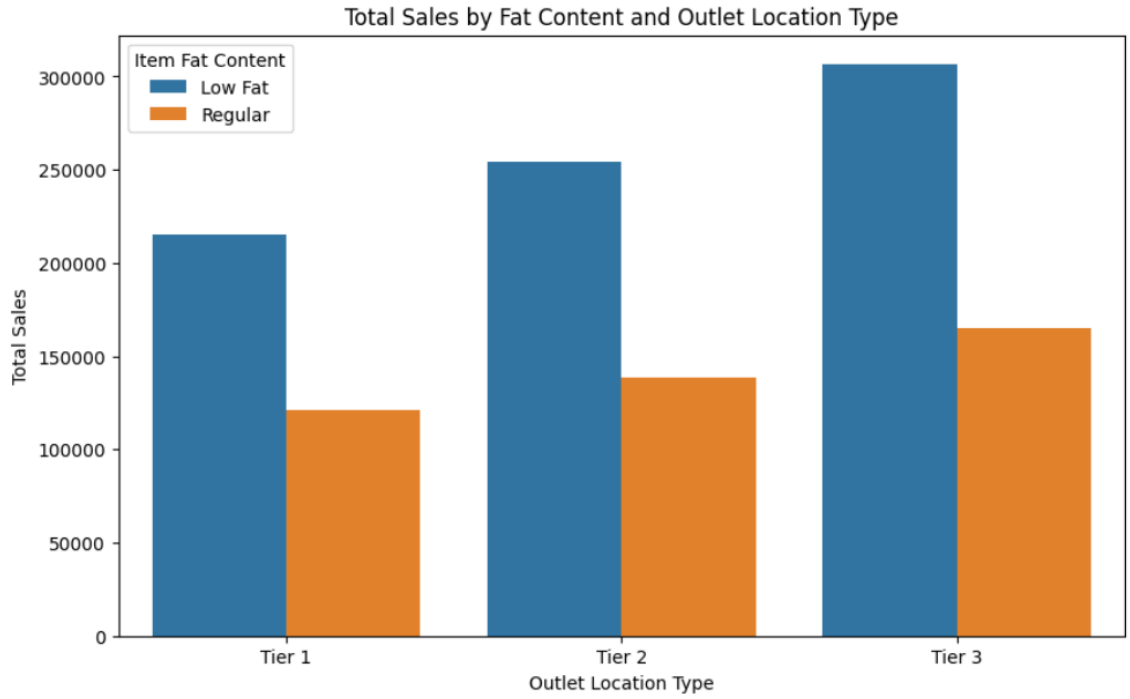
- **Total Sales**: Low Fat items have higher total sales, indicating greater consumer demand.
- **Mean Sales**: Average sales are similar for both categories (140.71 vs 141.50).
- **Sales Distribution**: Both categories show similar variability with a standard deviation around 62.
- **Range and Median**: Sales distributions are comparable, with slight differences in the interquartile range and median.

Total sales by item type



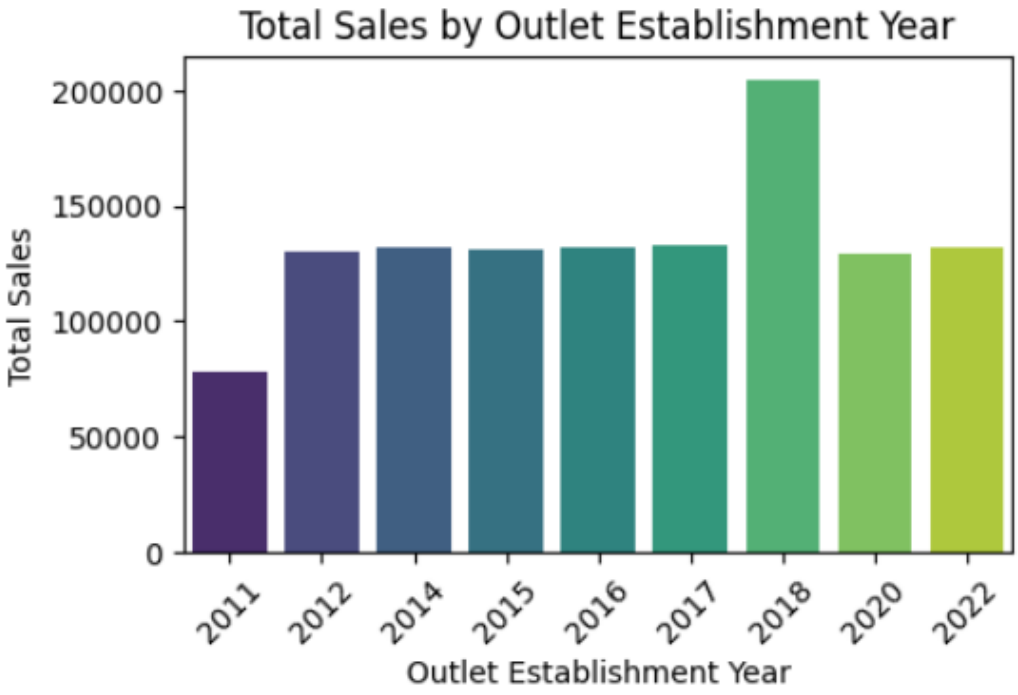
- Top Performers: Fruits and Vegetables lead with 178,124.08 in sales, followed by Frozen Foods (118,558.88) and Dairy (101,276.46).
- Lowest Sales: Seafood has the least sales at 9,077.87, with Starchy Foods and Others also underperforming.
- Popular Categories: Snack Foods (175,433.92), Household (135,976.53), and Baking Goods (81,894.74) show strong demand.
- Moderate Performance: Canned, Health and Hygiene, Meat, and Breads have steady but average sales.
- Growth Potential:Seafood and Starchy Foods offer opportunities for growth.

Fat content by outlet for total sales



- **Low Fat** items have a significant advantage in terms of total sales across all outlet types.
- **Tier 3** outlets have the highest sales, likely contributing to better overall performance in the sales data.

Total Sales by Outlet Establishment Year



1. **Increasing Sales**: Newer outlets, especially from 2017 onward, show higher total sales, with 2018 having the peak sales.
2. **Stable Early Years**: Sales for outlets established between 2011-2016 remain stable around 130,000.
3. **Lower Sales in Older Outlets**: Outlets from 2011 and 2012 show comparatively lower sales.
4. Overall, newer outlets tend to have higher sales, with the 2018 outlet leading.

Conclusion

The Blinkit Grocery Data Analysis provides valuable insights into various aspects of retail operations, sales trends, and product performance:

1.Outlet Demographics: The analysis shows that most outlets are concentrated in Tier 3 locations with medium-sized establishments. Newer outlets (established after 2017) tend to have higher total sales, indicating a trend of growing sales performance in newer stores.

2.Item Type Performance: Product categories like Fruits & Vegetables, Snack Foods, and Frozen Foods show the highest sales, while Seafood and Starchy Foods offer potential growth opportunities. Understanding customer preferences through item types can guide inventory and marketing strategies.

3. Fat Content Impact: Low Fat items outperform Regular items in terms of total sales, indicating a stronger consumer demand for healthier products. This trend could influence product selection and pricing strategies.

4. Outlet Type and Size: Supermarket Type 1 outlets, especially those in Tier 3, have the highest sales performance. Medium-sized outlets dominate in terms of count, suggesting a preference for medium-sized retail spaces in certain locations.

5.Sales Trends: The analysis highlights the potential to optimize sales strategies by focusing on the relationship between outlet size, location type, and sales performance. Furthermore, understanding the impact of item visibility and establishment year can help improve sales outcomes across various outlets.

6.Growth Opportunities: Certain product categories, like Seafood and Starchy Foods, show underperformance, presenting opportunities for targeted marketing and product placement. Similarly, Tier 3 outlets with Low Fat items are driving significant sales, suggesting a strategy to expand in these areas.

Overall, this analysis can assist Blinkit in refining its sales strategies, optimizing product offerings, and enhancing customer satisfaction through better inventory and outlet management.