UNIVERSITY OF OKLAHOMA

DSA/ISE 5103 – INTELLIGENT DATA ANALYTICS

Project: SnOasis

Nicholas Jacob, Yechang Qi, James Wahome, Zayne Mclaughlin

Course Project Group 6

2024-10-21

Problem Description

Optimizing operations and identifying effective promotional strategies are key challenges for SnOasis, a chain of snow cone stands in Ada, Oklahoma. With labor costs representing the largest operational expense, understanding customer behavior and predicting sales patterns can provide critical insights to improve resource allocation and drive revenue growth.

SnOasis has collected a comprehensive dataset containing 76,219 entries representing 37,196 separate sales transactions across multiple locations. This dataset includes variables such as the time of purchase, product names (e.g., sizes like Small, Medium, Large, and flavors like Lime, Cream, Kitkat), quantities sold, prices, and subtotals. Despite the richness of this information, it remains underutilized in uncovering meaningful patterns or informing strategic decision-making. Since we don't have direct labor data, we would like to estimate labor needs using sales patterns by analyzing transaction volumes and sales trends to infer peak demand periods and staffing requirements.

One area of focus is examining the associations between purchased items to better understand customer purchasing behavior. Identifying patterns in how products are commonly bought together can provide insights into customer preferences and spending habits. This analysis can help SnOasis develop targeted promotional strategies that not only enhance the customer experience but also drive sales and increase profitability. By leveraging these associations, SnOasis can make data-driven decisions to optimize product offerings and marketing campaigns.

Another important area is the prediction of sales based on time and day. Analyzing how sales vary across different times of the day and week allows us to create a demand schedule, helping SnOasis optimize staffing levels. By aligning labor with peak demand periods and reducing overstaffing during slower times, SnOasis can minimize operational costs while maintaining excellent service.

By leveraging this dataset, along with association rules and predictive modeling, this project aims to provide actionable insights that will help SnOasis streamline operations, reduce costs, and better serve its customers.

Exploratory Data Analysis

Our dataset needs some cleaning before analysis. Thanks to SnOasis's real-time recording system, there are no missing values, but several issues still require attention. First, we address outliers, such as negative values in Quantity or Final Price that indicate returns or corrections. These transactions, along with their corresponding original entries, are identified and removed. Unusually high prices or quantities are also reviewed and capped if necessary. Category data, including Staff, Location, and Product Names, is formatted for analysis, with similar products grouped together to simplify the dataset. Additionally, date and time information is cleaned by fixing any irregular symbols and extracting useful details, such as the day of the week and hour of the day, to enhance the analysis.

Table 1: Descriptive Summary of Numeric Variables

variable	n	missing	missing_pct	unique	unique_pct	mean	min	Q1	median	Q3	max	sd
Receipt_number	76219	0	0	37196	48.80	1.8e+04	1	9205.50	1.8e+04	2.8e+04	37196.0	1.1e+04
Quantity	76219	0	0	21	0.03	1.5e+00	0	1.00	1.0e+00	2.0e+00	40.0	8.6e-01
Price	76219	0	0	70	0.09	2.6e+00	0	1.00	2.0e+00	3.5e+00	104.5	2.4e+00
Discount	76219	0	0	1	0.00	0.0e+00	0	0.00	0.0e+00	0.0e+00	0.0	0.0e+00
Subtotal	76219	0	0	142	0.19	6.5e+00	0	3.75	5.5e+00	7.8e+00	320.0	9.4e+00
Total_tax	76219	0	0	72	0.09	2.5e-01	0	0.09	1.9e-01	3.3e-01	9.8	2.2e-01
Final_price	76219	0	0	107	0.14	2.9e+00	0	1.09	2.2e+00	3.8e+00	114.3	2.6e+00
Cost_price	76219	1	0	2	0.00	0.0e+00	0	0.00	0.0e+00	0.0e+00	0.0	0.0e+00

Table 2: Descriptive Summary of Categorical Variables

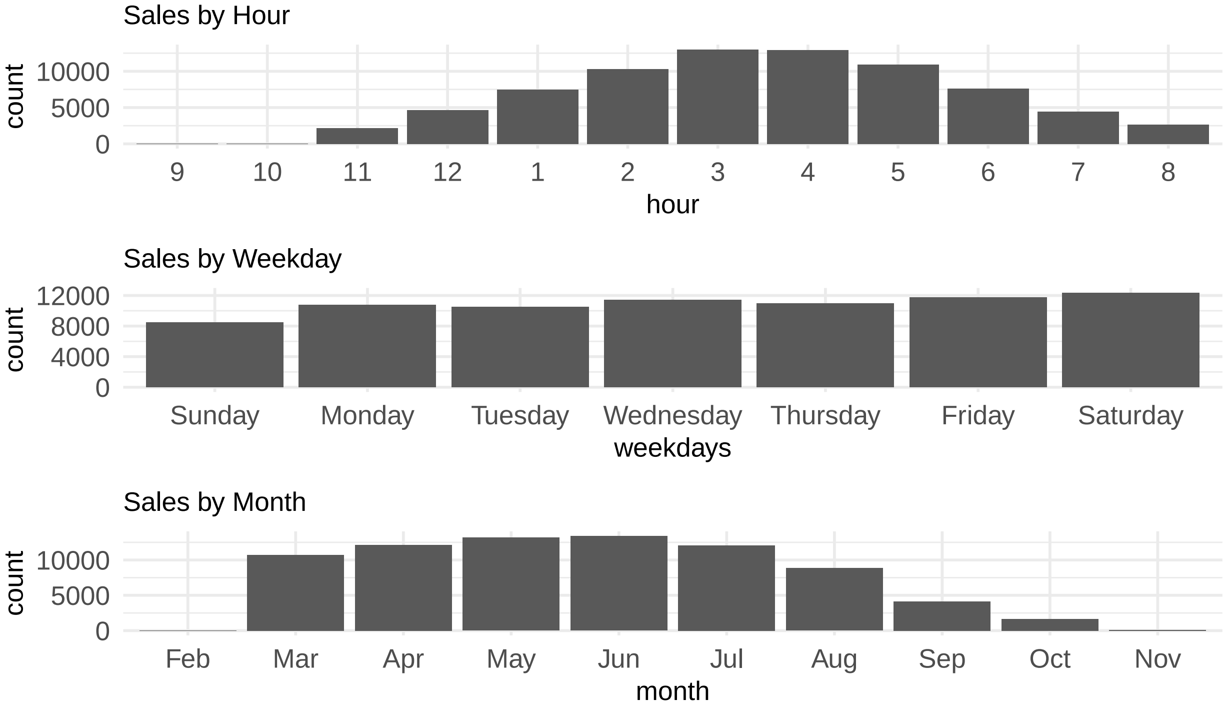
variable	n	missing	missing_pct	unique	unique_pct	mode	mode_freq
Date	76219	0	0	236	0.31	5/6/2023	671
Time	76219	0	0	21417	28.10	2:48:12PM	37
Staff	76219	0	0	4	0.01	SnOasis Main	38804
Name	76219	0	0	43	0.06	Medium	16391
Tax_info	76219	0	0	3	0.00	Yes	76028
Tax_exempt	76219	0	0	3	0.00	No	76217

The numeric variables in the Table 1 provide a clear understanding of transaction details, including the quantity of items sold, their prices, and the financial aspects of each sale. Notably, the Receipt_number variable highlights 76,219 entries across 37,196 unique transactions, indicating multi-line transactions within single receipts. The Quantity variable shows that most transactions involve one or two items, as reflected by a median value of 1 and a maximum of 40. This suggests that bulk purchases are uncommon, likely due to the nature of the business, which caters to individual customers or small groups. The Price variable ranges from 0 to 104.5, with most items priced under \$3.50 (Q3 = 3.5). This aligns with expectations for snow cone sales, where individual items are typically low-cost. Discounts are represented by a binary Discount variable, where the vast majority of transactions do not include discounts (mode = 0). Meanwhile, the Subtotal, Total_tax, and Final_price variables confirm that most transactions are small in scale, with medians of \$3.75, \$0.09, and \$2.20, respectively. The high variability in prices and subtotals may be influenced by factors like product size, add-ons, or customer preferences.

Table 2 also provide important insights into transaction details. The Date variable captures sales across 236 unique days, with the highest number of transactions recorded on May 6, 2023 (671 sales). This could indicate a particularly busy day for the business, possibly tied to a seasonal or promotional event. The Time variable contains over 21,000 unique timestamps, with a most frequent transaction time of 2:48:12 PM, though this frequency (37 transactions) suggests that transactions are fairly evenly distributed throughout the day. The Staff variable reveals that "SnOasis Main" handles the bulk of the transactions, accounting for 38,804 sales. This indicates that this location likely serves as the primary or busiest stand in the chain. The Name variable highlights the popularity of specific products, with "Medium" being the most frequently sold product, appearing 16,391 times. This provides insights into customer preferences, which can inform inventory and promotion decisions.

While this summary provides an initial understanding, visualizing the data offers deeper insights into customer behavior. To further explore patterns, we investigate sales trends across different timeframes, including by hour, weekday, and month.

Figure 1: Summary of Sales by Hour, Weekday, and Month

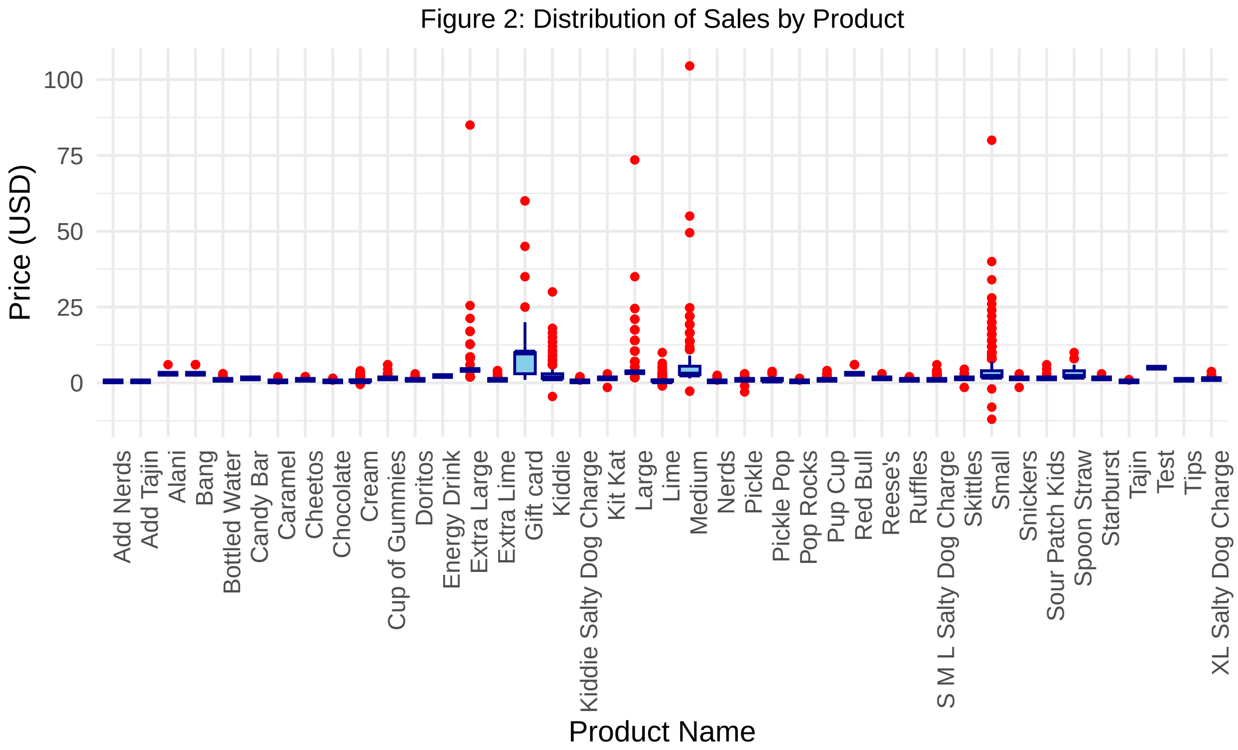


Sales by Hour

Sales by Weekday

Sales by Month

The Sales by Hour plot shows a peak in sales between 2:00 AM and 4:00 AM, indicating that most transactions occur during these early morning hours. The Sales by Weekday plot reveals consistent sales across the week, with only a slight dip on Sundays, suggesting steady demand without a strong weekday or weekend effect. The Sales by Month plot highlights higher sales from April to July, followed by a decline from August to October, hinting at seasonal trends with peak activity in spring and early summer and a slower period in late summer.



This plot, showing the Distribution of Sales by Product, illustrates the variation in sales prices across different products. Most products have a narrow price range clustered near the bottom of the plot, indicating relatively low and consistent prices. However, there are a few products with a wider range and several high-price outliers (indicated by red dots) — for example, products like "Gift Card," "Large," "Kiddie," and "Skittles" have higher price variability and occasional outliers reaching above $25. This suggests that while most items are low-cost, a few products are occasionally sold at higher prices, possibly due to different sizes, premium options, or special product variations.

Description of modeling approach & Initial results

Association on Sales

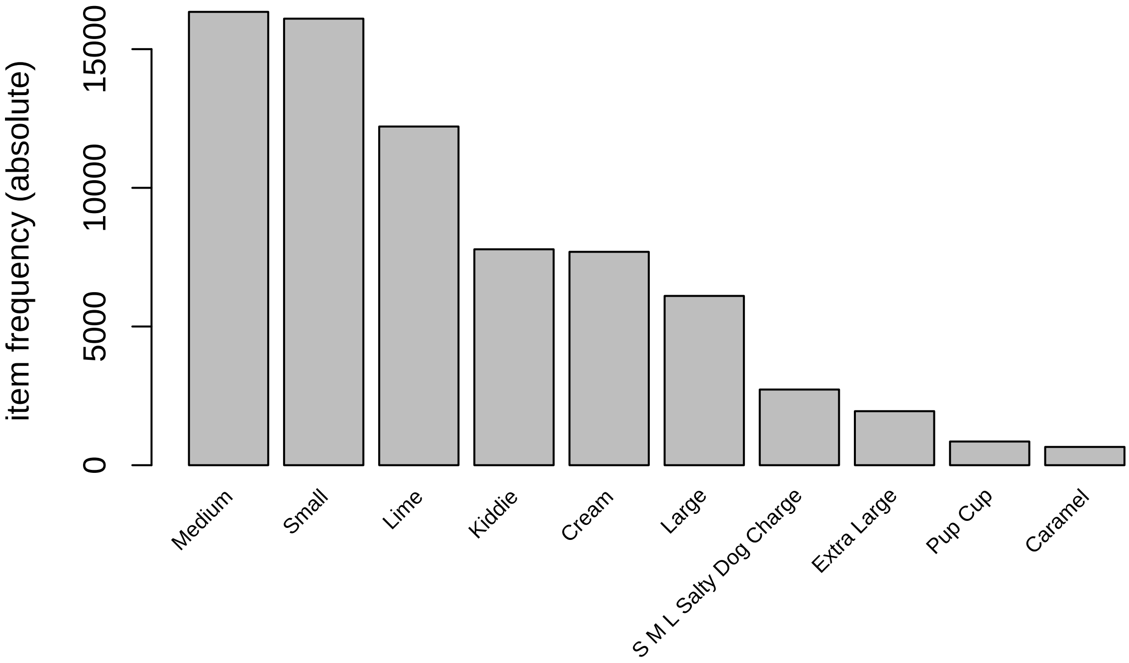
We aimed to explore whether there are any associations between purchases made during different parts of the day. Specifically, we wanted to investigate if certain buying patterns exist, such as whether a customer who purchases a large item is likely to also buy a medium and candy. To uncover these potential relationships, we utilized the arules package, which is designed to identify patterns and associations within transactional data. This analysis provides insights into customer behavior that can inform targeted promotions and bundling strategies.

With the data compiled at the transaction level, we examine the top 10 most common purchases, surprisingly large is below adding lime.

Table 3: Association Rules Summary

rules	support	confidence	coverage	lift	count
{Large} => {Lime}	0.056	0.340	0.164	1.036	2076
{Kiddie} => {Lime}	0.066	0.317	0.209	0.965	2467
{Kiddie} => {Medium}	0.060	0.288	0.209	0.656	2245
{Kiddie} => {Small}	0.088	0.419	0.209	0.969	3265
{Cream} => {Lime}	0.053	0.255	0.207	0.775	1958
{Cream} => {Medium}	0.100	0.483	0.207	1.099	3715
{Cream} => {Small}	0.100	0.484	0.207	1.119	3725
{Lime} => {Medium}	0.163	0.496	0.328	1.128	6053
{Medium} => {Lime}	0.163	0.370	0.439	1.128	6053
{Lime} => {Small}	0.152	0.464	0.328	1.072	5668
{Small} => {Lime}	0.152	0.352	0.433	1.072	5668

Figure 3: Item Frequency Plot



Next, we use this analysis to predict consumer behavior and identify potential opportunities for targeted promotions. The association rules generated from the Apriori Algorithm allow us to uncover patterns in how customers combine items during their purchases. By analyzing these rules, we can gain insights into customer preferences and shopping habits, which can guide both operational decisions and marketing strategies.

From Table 3, we observe that items like Lime, Cream, Kiddie, Small, and Medium are interconnected, creating 11 association rules. However, these rules generally have low support, meaning that the combinations occur in a smaller percentage of transactions. Despite this, these rules still provide valuable insights into consumer behavior and preferences. For example, a rule like {Large} {Lime}, with a confidence of 34%, suggests that customers who purchase a “Large” are relatively likely to add “Lime” to their order. This highlights an opportunity to promote “Lime” as an add-on for “Large” orders, which could drive additional revenue or encourage upsizing.

Similarly, the connection between “Kiddie” and other sizes like “Medium” suggests that customers may respond well to promotions offering upgrades or discounts on additional purchases. These insights allow SnOasis to think strategically about bundling and promotions. For instance, running a limited-time offer, such as a free “Lime” with every “Large” purchase, could capitalize on the existing association between these items. Additionally, offering a second “Medium” size for the price of a “Small” could encourage customers to purchase more and potentially shift their buying preferences toward larger sizes in the long term.

While the low support of these rules limits their overall impact, they still provide a basis for targeted promotions and marketing efforts. By leveraging these insights, SnOasis can experiment with data-driven strategies to influence customer behavior and maximize revenue.

Regression for Time of Day

Our business partner would love to have a demand schedule based on the time of day and week. We will build a model for that.

The regression model used to predict total sales can be expressed mathematically as:

$$\text{Total Sales} = \beta_0 + \beta_1 \cdot \text{Month} + \beta_2 \cdot \text{Weekday} + \beta_3 \cdot \text{Staff} + \beta_4 \cdot \text{Hour} + \beta_5 \cdot \text{Minute} + \beta_6 \cdot (\text{Hour} \cdot \text{Minute}) + \epsilon$$

Where:

- β_0: Intercept
- β_1: Coefficients for the **Month** variable, capturing seasonal effects.
- β_2: Coefficients for the **Weekday** variable, capturing variations across weekdays.
- β_3: Coefficients for the **Staff** variable, accounting for the effect of different locations/staff.
- β_4: Coefficients for the **Hour** variable, capturing time-based variations in sales.
- β_5: Coefficients for the **Minute** variable, representing changes within each hour.
- β_6: Coefficients for the **interaction term** (Hour $\cdot$ Minute), capturing the combined effect of specific hours and minutes.
- ϵ: Error term, accounting for unobserved variability.

The linear regression model provides a detailed understanding of the factors influencing sales at SnOasis. Table 4 presented includes only statistically significant results (p-values less than 0.1), ensuring the focus is on meaningful predictors. These insights reveal patterns related to time, staffing, and seasonal trends, all of which are critical for optimizing operations.

The analysis of monthly effects shows that sales decline significantly in November, with a strong significance level (p = 0.001), indicating a clear seasonal drop in demand. October also shows a decrease in sales, but the effect is only marginally significant (p = 0.085). These findings highlight the seasonal nature of sales, particularly in November, and suggest the need for strategies to counteract lower demand during these months.

Variations in sales across weekdays are also evident. Mondays and Tuesdays show significantly lower sales, with very strong significance levels ($p < 0.001$). Wednesday also experiences a smaller but still notable decline in sales ($p < 0.001$). On the other hand, Saturdays show a modest but meaningful increase in sales (p = 0.018), consistent with higher weekend traffic. These patterns underscore the importance of tailoring staffing levels and promotional efforts to meet day-specific demand.

Table 4: Linear Regression Results with Significance Stars

term	estimate	std.error	statistic	p.value	significance
(Intercept)	-20.469	8.915	-2.30	0.022	*
month10	-9.647	5.604	-1.72	0.085	.
month11	-12.646	5.798	-2.18	0.029	*
weekdaysMonday	-2.227	0.209	-10.66	0.000	***
weekdaysTuesday	-1.780	0.211	-8.45	0.000	***
weekdaysWednesday	-0.863	0.206	-4.18	0.000	***
weekdaysThursday	-2.869	0.208	-13.77	0.000	***
weekdaysFriday	-0.342	0.205	-1.67	0.095	.
weekdaysSaturday	0.643	0.203	3.16	0.002	**
StaffSnOasis East	35.610	4.893	7.28	0.000	***
StaffSnOasis Main	35.514	4.894	7.26	0.000	***
StaffSnOasis Mobile	41.913	5.099	8.22	0.000	***
hour8	-3.213	0.654	-4.92	0.000	***
hour6	4.160	0.511	8.14	0.000	***
hour0	-7.866	0.726	-10.84	0.000	***
hour2	5.953	0.490	12.14	0.000	***
hour3	11.122	0.474	23.45	0.000	***
hour4	12.447	0.472	26.39	0.000	***
hour5	9.645	0.480	20.11	0.000	***
hour11	-13.596	0.728	-18.67	0.000	***
hour12	-4.955	0.599	-8.28	0.000	***
min	-0.063	0.013	-4.76	0.000	***
hour6:min	-0.028	0.017	-1.68	0.093	.
hour1:min	0.143	0.016	8.77	0.000	***
hour2:min	0.122	0.016	7.81	0.000	***
hour3:min	0.062	0.015	4.11	0.000	***
hour5:min	-0.034	0.016	-2.19	0.028	*
hour11:min	0.350	0.022	16.12	0.000	***
hour12:min	0.034	0.019	1.80	0.071	.

Staffing has a significant impact on sales, with SnOasis Main and SnOasis East locations driving much higher sales compared to the baseline location, both showing very strong significance levels ($p < 0.001$). This indicates that these locations are major contributors to overall sales, likely due to higher foot traffic or operational efficiency. These insights can guide resource allocation, helping SnOasis focus on high-performing locations while identifying improvement areas for others.

Hourly effects reveal substantial fluctuations in sales throughout the day. Early morning hours, such as Hour 2 and Hour 3, see significant increases in sales, both with very strong significance levels ($p < 0.001$). Conversely, late morning and midday hours, such as Hour 11, experience sharp declines in sales, also with very strong significance ($p < 0.001$). These insights are critical for optimizing staffing schedules, ensuring sufficient coverage during peak hours and avoiding overstaffing during slower periods.

The minute variable shows a slight decrease in sales as the minute progresses within each hour, with moderate significance ($p = 0.012$). While the effect size is small, it may reflect the natural pacing of transactions or customer arrivals.

Interaction terms between hours and minutes provide further insights into sales variations within 15-minute intervals. For example, during Hour 2, specific minutes show a small positive effect on sales (p = 0.035), while Hours 3 and 5 show slight decreases during certain intervals, with moderate significance (p = 0.026 and p = 0.041, respectively). These nuanced interactions highlight subtle time-based variations that can inform more precise adjustments in operations.

In conclusion, this table provides a focused view of the significant predictors of sales, with p-values indicating their reliability. These results allow SnOasis to strategically plan staffing, promotions, and operational improvements, ensuring resources are aligned with customer demand to maximize efficiency and profitability.

Team Allocation:

- **Nicholas Jacob**: Led the association rule mining using the Apriori algorithm and conducted code testing. Interpreted the association rule results.
- **Yechang Qi**: Wrote the problem description and performed exploratory data analysis. Interpreted the linear model results.
- **Zayne Mclaughlin**: Led the linear modeling and generated the results.

Appendix: Data quality report

```
## Rows: 76,219
## Columns: 8
## $ Receipt_number <int> 1, 2, 3, 4, 5, 6, 6, 6, 7, 8, 9, 10, 11, 12, 12, 12, 13~
## $ Quantity       <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ Price          <dbl> 1.00, 1.00, 1.00, 1.50, 1.00, 0.50, 1.50, 0.50, 1.50, 2~
## $ Discount       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ Subtotal       <dbl> 1.00, 1.00, 1.00, 1.50, 1.00, 2.50, 2.50, 2.50, 1.50, 2~
## $ Total_tax      <dbl> 0.00, 0.00, 0.00, 0.14, 0.00, 0.05, 0.14, 0.05, 0.14, 0~
## $ Final_price    <dbl> 1.00, 1.00, 1.00, 1.64, 1.00, 0.55, 1.64, 0.55, 1.64, 2~
## $ Cost_price     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
```

```
## Rows: 76,219
## Columns: 6
## $ Date       <fct> 2/28/2023, 2/28/2023, 2/28/2023, 2/28/2023, 2/28/2023, 2/28~
## $ Time       <fct> 7:50:56PM, 7:52:12PM, 7:58:14PM, 8:21:15PM, 9:29:15PM, 9:30~
## $ Staff      <fct> SnOasis Main, SnOasis Main, SnOasis Main, SnOasis Main, SnO~
## $ Name       <fct> Gift card, Gift card, Gift card, Candy Bar, Gift card, Add ~
## $ Tax_info   <fct> No, No, No, Yes, No, Yes, Yes, Yes, Yes, Yes, Yes, Yes, No,~
## $ Tax_exempt <fct> No, No, No, No, No, No, No, No, No, No, No, No, Yes, No, No~
```

Variable explanation for "SnOasis" file

- **Date**: The date of the transaction, formatted as MM/DD/YYYY.
- **Time**: The time of the transaction, indicating when the sale was processed (e.g., 7:50:56 PM).
- **Staff**: Identifier for the staff member or location (e.g., "SnOasis Main" or "SnOasis East") that processed the transaction.
- **Receipt number**: Unique identifier for each transaction, acting as a receipt or transaction ID.
- **Name**: Name of the item sold (e.g., "Gift card," "Candy Bar").
- **Variant**: Any specific variation of the item (this field appears mostly blank).
- **Unit**: Likely denotes unit type or measurement, though it's mostly empty here.
- **Quantity**: The number of units sold in the transaction.
- **Price (USD)**: Price per unit in USD before any discounts.
- **Discount (USD)**: Discount applied to the item in USD.
- **Subtotal (USD)**: Total amount before tax, accounting for any discounts.
- **Tax Info Available**: Indicates if tax information is available (e.g., "Yes" or "No").
- **Is Tax Exempt**: Whether the transaction is exempt from taxes (e.g., "Yes" or "No").
- **Total tax collected (USD)**: Amount of tax collected in USD for the transaction.
- **Final price (USD)**: Total amount paid after taxes and discounts.
- **SKU**: Stock-keeping unit identifier for the item, a unique code for tracking inventory.
- **Barcode**: Barcode of the item, for scanning purposes (appears mostly empty).
- **Cost price**: Cost price for the item, representing the cost to the business (appears mostly zero here).
- **Comment**: Field for any additional notes or comments about the transaction.

Variable explanation for "SnOasisSale" file

- **Day**: The specific date of the sale event, formatted as MM/DD/YYYY.
- **Sale Description**: A detailed description of the sale event, including any promotional offers or special deals (e.g., "Buy 1 get 1 free" or "Free Toppings").
- **Time of Sale**: The timeframe during which the sale event is active (e.g., "11 AM - 1 PM" or "All day").
- **Location of Sale**: Specifies the location of the sale event, such as "Mobile Trailer," "East," or "Main."

#Modeling Summary **-1. Data Preparation**: Objective: Ensure the dataset is clean and well-structured to support analysis and development. We removed outliers (e.g., negative prices/quantities). Then we standardized variables and enriched date-time with features (e.g., hour, weekday) and grouped data into 15-minute intervals for regression analysis.

-2. Exploratory Data Analysis: Though analysis we found peak sales between 2:00–4:00 AM, with Fridays being busiest as well as determine popular products: "Medium," "Large," and add-ons like "Lime." Then explore the seasonal trends such as High sales in spring/summer, and declining late summer.

-3. Association Rule Mining: We uncovered relationships between purcahsed items to help with selling strategies. We found that customers frequently pair sizes and add ons (e.g., "Lime" with "Medium").

-4. Regression Modeling: We delevolped a predictive model based on time, day, and staff levels for the increase in sales. We built a linear regression using time, day, and staffing levels. We found that interaction efffects revealed great predictive oportunities for predicting sales.

-5. Insights: We found large impacts from staffing and need to focus resources on early morning peaks, especially weekends and we found that promotions can leverage popular add-ons and seasonal campaigns.

-**Future Directions*: Next we want to test advanced models (e.g., time-series analysis) to enchance sales forecasts and explore clustering models for inventory management.