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

Initial Data Analysis

Our dataset needs some cleaning before analysis. Although we don't have missing values thanks to our real-time recording system, we still need to handle a few things. First, we'll deal with outliers, specifically negative values in Quantity or Final Price that show returns or corrections. We'll find and remove these transactions along with their original entries. We'll also check if any unusually high prices or quantities need to be capped. Next, we'll properly format our category data like Staff, Location, and Product Names for analysis, possibly grouping similar products together to keep things simple. Finally, we'll clean up our date and time information, fixing any weird symbols and adding useful details like day of the week and hour of day.

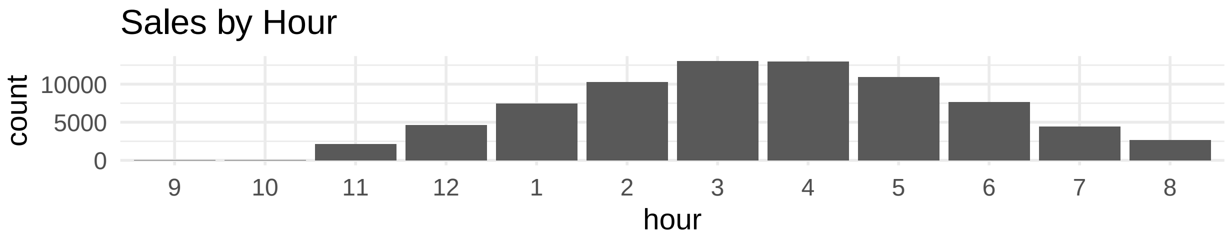
Table 1: Descriptive Summary of Numeric Variables

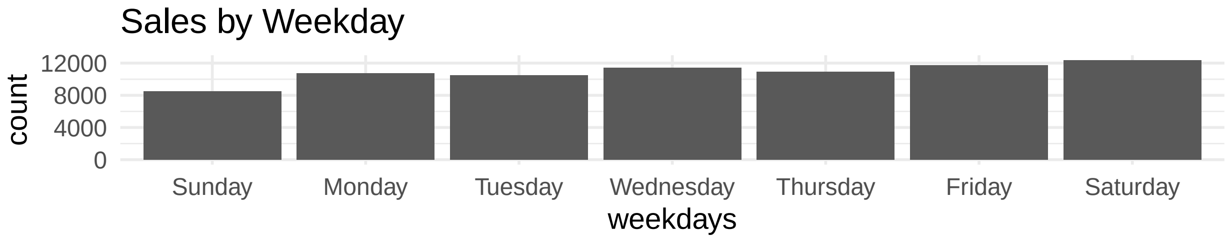
variable	n	missing	missing_pct	unique	unique_pct	mean	min	Q1	median	Q3	max	sd
Receipt_number	76219	0	0	37196	48.80	1.8e+04	1.0	9205.50	1.8e+04	2.8e+04	37196.0	1.1e+04
Quantity	76219	0	0	25	0.03	1.5e+00	-6.0	1.00	1.0e+00	2.0e+00	40.0	8.7e-01
Price	76219	0	0	79	0.10	2.6e+00	-12.0	1.00	2.0e+00	3.5e+00	104.5	2.4e+00
Discount	76219	0	0	34	0.04	0.0e+00	-8.2	0.00	0.0e+00	0.0e+00	0.0	6.0e-02
Subtotal	76219	0	0	149	0.20	6.5e+00	-12.5	3.75	5.5e+00	7.8e+00	320.0	9.4e+00
Total_tax	76219	0	0	81	0.11	2.5e-01	-1.1	0.09	1.9e-01	3.3e-01	9.8	2.2e-01
Final_price	76219	0	0	117	0.15	2.9e+00	-13.1	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	0.00	0.0e+00	0.0e+00	0.0	0.0e+00

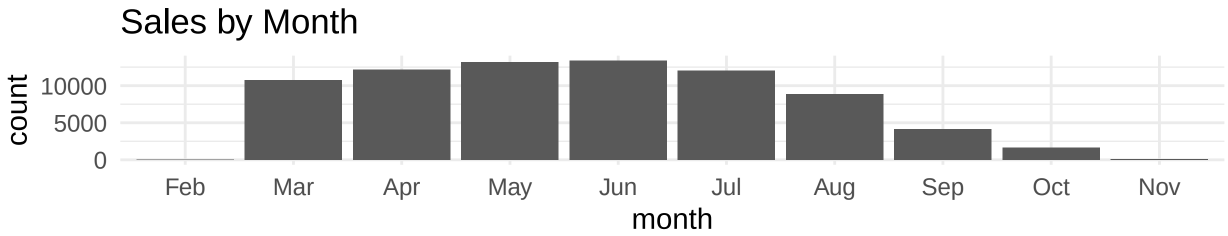
Table 2: Descriptive Summary of Categorical Variables

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

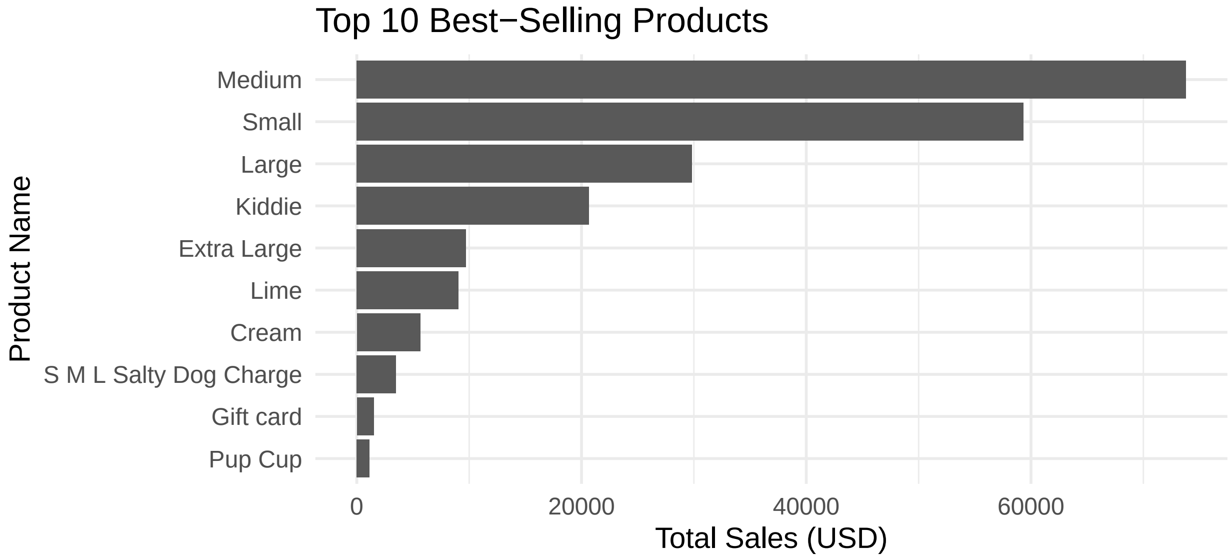
Visualizations



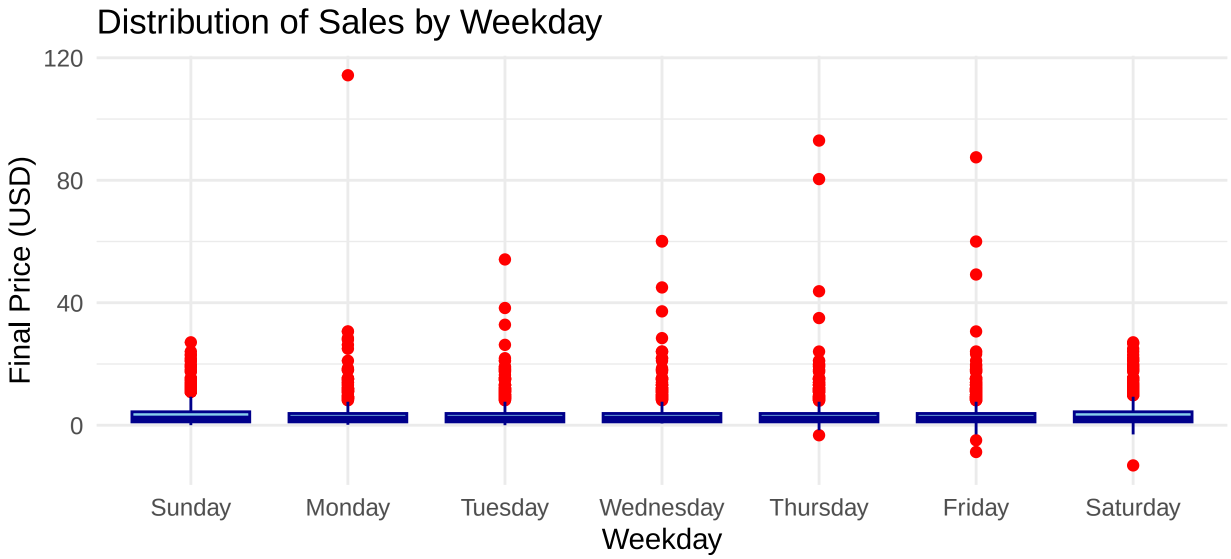




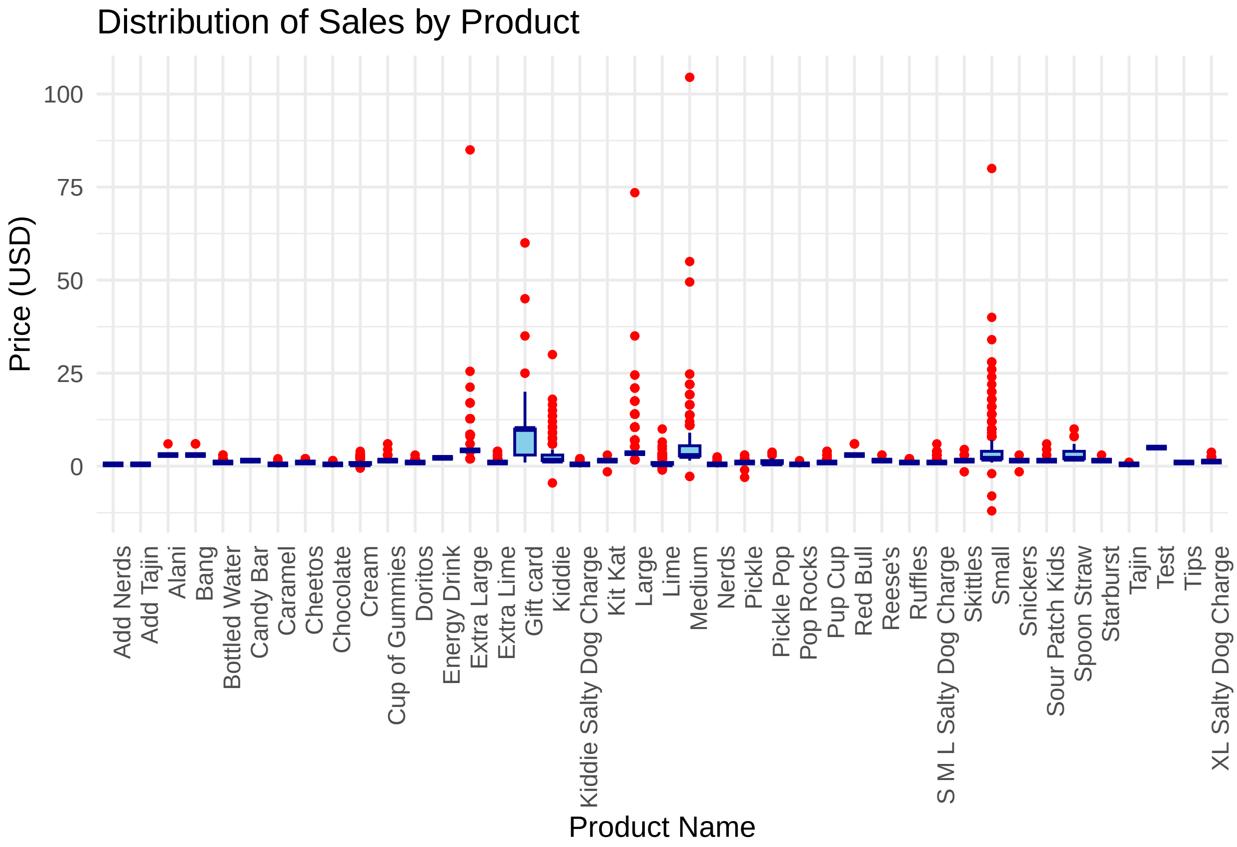
The Sales by Hour plot shows a peak in sales between 2:00 AM and 4:00 AM, indicating that most transactions occur during these early morning hours. The Sales by Weekday plot reveals consistent sales across the week, with only a slight dip on Sundays, suggesting steady demand without a strong weekday 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.



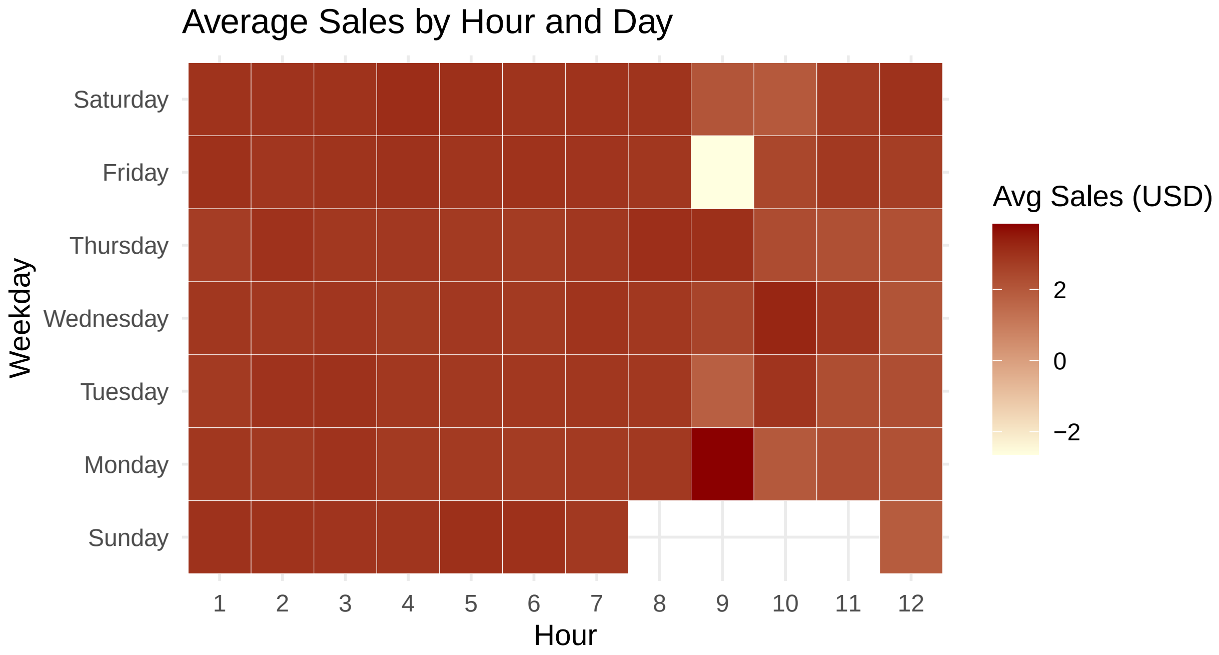
This chart shows that the highest-grossing products at SnOasis are primarily size options, with “Medium,” “Small,” and “Large” leading in total sales. Size variations dominate the top 10, indicating they are the most popular choices among customers. Additionally, add-ons like “Lime” and “Cream” and specialized items like “S M L Salty Dog Charge” and “Pup Cup” contribute significantly to sales. This suggests that offering a variety of sizes and add-ons is key to driving revenue at SnOasis.



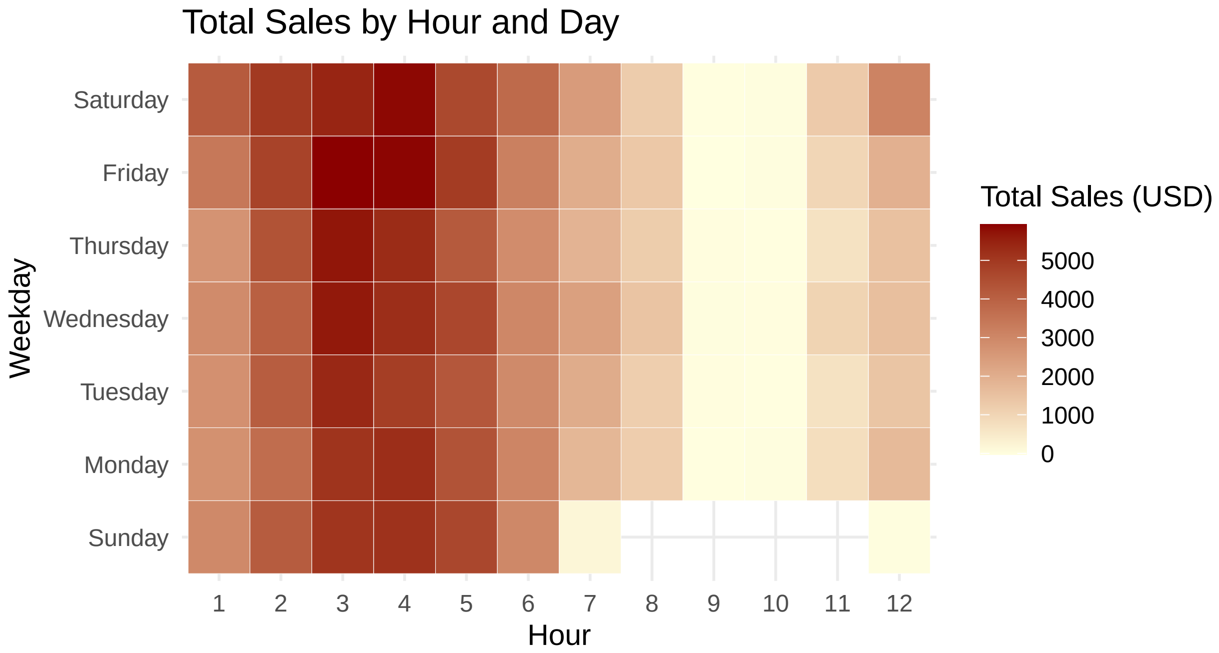
The Distribution of Sales by Weekday plot shows that the majority of sales prices are relatively low and consistent across all days, as indicated by the compact boxplots near the bottom. However, there are some high-price outliers (marked by red dots) each day, with the highest outlier reaching around $120. This suggests that while typical sales amounts are stable throughout the week, occasional high-value transactions occur randomly across all days.



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.



The heatmap shows that average sales are generally consistent but low across most days and hours, with one noticeable peak on Friday around 9:00 AM. Sunday has the lowest average sales, especially in the early morning hours. This suggests that SnOasis could focus promotions or staffing on the Friday morning peak and consider adjusting resources for lower-demand times, particularly on Sundays.

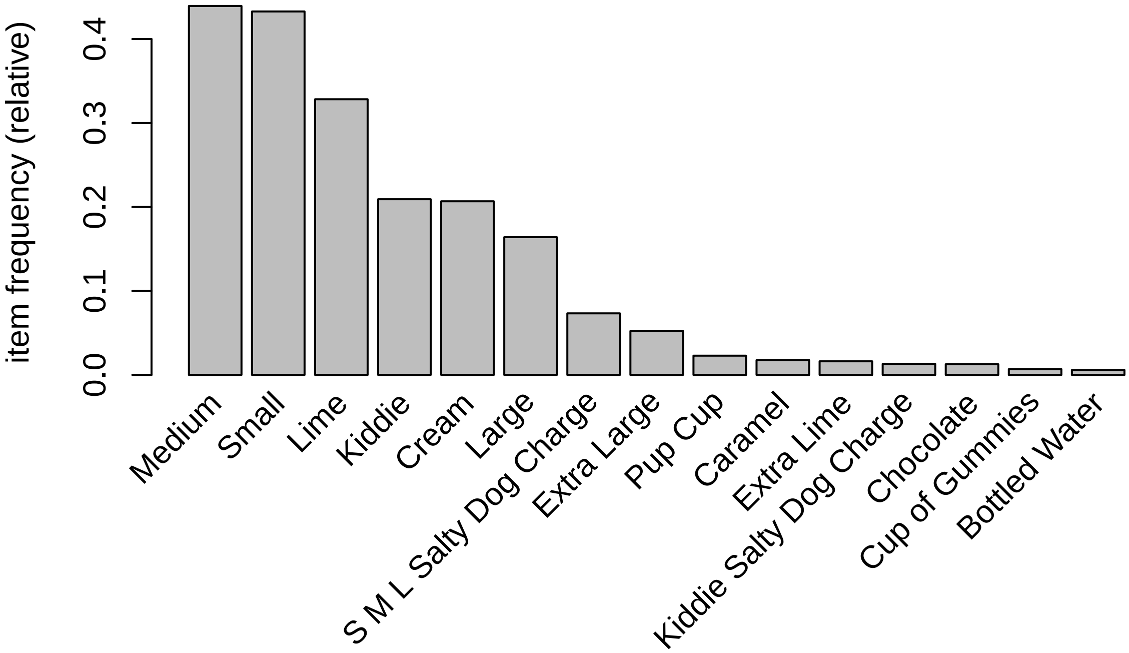


The pattern indicates that sales are concentrated in the early morning hours, peaking around 2:00 to 5:00 AM across most days, particularly from Friday through Sunday. Sales decrease significantly after 7:00 AM and remain low throughout the late morning. This suggests that SnOasis experiences its busiest periods in the early morning hours on weekends, which could inform staffing and inventory decisions for these peak times.

Association on Sales

I was curious to see if we could identify any association between the chunks of the day. Could we predict that if someone bought a large that they would also get a medium and candy? We utilize the `arules` package to examine this.

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



Next we use this to predict consumer behavior.

```
##      lhs         rhs      support confidence coverage lift  count
## [1]  {Large}  => {Lime}   0.0558  0.340      0.164    1.036 2076
## [2]  {Kiddie} => {Lime}   0.0663  0.317      0.209    0.965 2467
## [3]  {Kiddie} => {Medium} 0.0604  0.288      0.209    0.656 2245
## [4]  {Kiddie} => {Small}  0.0878  0.419      0.209    0.969 3265
## [5]  {Cream}  => {Lime}   0.0526  0.255      0.207    0.775 1958
## [6]  {Cream}  => {Medium} 0.0999  0.483      0.207    1.099 3715
## [7]  {Cream}  => {Small}  0.1001  0.484      0.207    1.119 3725
## [8]  {Lime}   => {Medium} 0.1627  0.496      0.328    1.128 6053
## [9]  {Medium} => {Lime}   0.1627  0.370      0.439    1.128 6053
## [10] {Lime}   => {Small}  0.1524  0.464      0.328    1.072 5668
## [11] {Small}  => {Lime}   0.1524  0.352      0.433    1.072 5668
```

Here we see that Lime, cream, Kiddie, Small and Medium are all tied together creating 11 rules that don't have very high support. While this doesn't give us much insight about these items, it does reveal something about consumer behavior. We might consider that a promotion for free lime with purchase would be very popular, many of our customers are getting this. Instead we might consider that giving free lime to a large might encourage customers to upsize which could be profitable to our partner. With the connection of kiddie and other sizes in these rules, we might consider running an early season promotion where the customer gets a second medium for the price of a small. Run this promotion for a limited time to see if you can influence long term those customers to change from both small and medium to only mediums.

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. We utilized time of day in 15 minute, month, and day with location to predict the total sales. We believe this will be a boon to our corporate sponsor.

```
## 
## Call:
## lm(formula = totalSales ~ month + weekdays + Staff + hour + min + 
##     hour:min, data = dfBy10min)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -31.47  -7.66  -1.23   6.14 307.07 
## 
## Coefficients: (1 not defined because of singularities)
##                     Estimate Std. Error t value             Pr(>|t|)    
## (Intercept)         -20.4693     8.9151   -2.30               0.0217 *  
## month3                5.6190     5.5955    1.00               0.3153    
## month4                6.7472     5.5943    1.21               0.2278    
## month5                7.1462     5.5940    1.28               0.2014    
## month6                9.0474     5.5934    1.62               0.1058    
## month7                4.1682     5.5943    0.75               0.4562    
## month8                1.7135     5.5949    0.31               0.7594    
## month9               -5.3392     5.5978   -0.95               0.3402    
## month10              -9.6472     5.6041   -1.72               0.0852 .  
## month11             -12.6458     5.7981   -2.18               0.0292 *  
## weekdaysMonday       -2.2270     0.2089  -10.66 < 0.0000000000000002 ***
## weekdaysTuesday      -1.7797     0.2106   -8.45 < 0.0000000000000002 ***
## weekdaysWednesday    -0.8632     0.2064   -4.18  0.00002896972063908 ***
## weekdaysThursday     -2.8686     0.2083  -13.77 < 0.0000000000000002 ***
## weekdaysFriday       -0.3419     0.2049   -1.67               0.0951 .  
## weekdaysSaturday      0.6428     0.2031    3.17               0.0015 ** 
## StaffSnOasis  East   35.6096     4.8933    7.28  0.00000000000034417 ***
## StaffSnOasis Main    35.5141     4.8940    7.26  0.00000000000040056 ***
## StaffSnOasis Mobile  41.9131     5.0987    8.22 < 0.0000000000000002 ***
## hour8                -3.2132     0.6536   -4.92  0.00000088467058842 ***
## hour9               -16.5675    13.9503   -1.19               0.2350    
## hour6                 4.1599     0.5109    8.14  0.00000000000000039 ***
## hour0                -7.8655     0.7256  -10.84 < 0.0000000000000002 ***
## hour1                -0.4614     0.5175   -0.89               0.3726    
## hour2                 5.9525     0.4903   12.14 < 0.0000000000000002 ***
## hour3                11.1219     0.4743   23.45 < 0.0000000000000002 ***
## hour4                12.4469     0.4716   26.39 < 0.0000000000000002 ***
## hour5                 9.6446     0.4797   20.11 < 0.0000000000000002 ***
## hour10              -28.5402    27.4460   -1.04               0.2984    
## hour11              -13.5964     0.7282  -18.67 < 0.0000000000000002 ***
## hour12               -4.9552     0.5988   -8.28 < 0.0000000000000002 ***
## min                  -0.0629     0.0132   -4.76  0.00000192740212685 ***
## hour8:min            -0.0165     0.0237   -0.69               0.4880    
## hour9:min             0.5482     0.6952    0.79               0.4303    
## hour6:min            -0.0278     0.0166   -1.68               0.0934 .  
## hour0:min                 NA         NA      NA                   NA    
## hour1:min             0.1433     0.0163    8.77 < 0.0000000000000002 ***
## hour2:min             0.1220     0.0156    7.81  0.00000000000000562 ***
## hour3:min             0.0624     0.0152    4.11  0.00003992602819586 ***
## hour4:min            -0.0147     0.0152   -0.97               0.3337    
## hour5:min            -0.0342     0.0156   -2.19               0.0282 *  
## hour10:min            0.3664     0.5511    0.66               0.5062    
## hour11:min            0.3505     0.0217   16.12 < 0.0000000000000002 ***
## hour12:min            0.0344     0.0191    1.81               0.0710 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 14.2 on 76176 degrees of freedom
## Multiple R-squared:  0.191,  Adjusted R-squared:  0.191 
## F-statistic:  428 on 42 and 76176 DF,  p-value: <0.0000000000000002
```

This is a model, I'll let someone else interpret…

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       <int> 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.

```
## Rows: 76,219
## Columns: 19
## $ Date                       <chr> "2/28/2023", "2/28/2023", "2/28/2023", "2/28~
## $ Time                       <chr> "7:50:56?PM", "7:52:12?PM", "7:58:14?PM", "8~
## $ Staff                      <chr> "SnOasis Main", "SnOasis Main", "SnOasis Mai~
## $ Receipt.number             <int> 1, 2, 3, 4, 5, 6, 6, 6, 7, 8, 9, 10, 11, 12,~
## $ Name                       <chr> "Gift card", "Gift card", "Gift card", "Cand~
## $ Variant                    <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ Unit                       <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ Quantity                   <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ Price..USD.                <dbl> 1.00, 1.00, 1.00, 1.50, 1.00, 0.50, 1.50, 0.~
## $ Discount..USD.             <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ Subtotal..USD.             <dbl> 1.00, 1.00, 1.00, 1.50, 1.00, 2.50, 2.50, 2.~
## $ Tax.Info.Available         <chr> "No", "No", "No", "Yes", "No", "Yes", "Yes",~
## $ Is.Tax.Exempt              <chr> "No", "No", "No", "No", "No", "No", "No", "N~
## $ Total.tax.collected..USD.  <dbl> 0.00, 0.00, 0.00, 0.14, 0.00, 0.05, 0.14, 0.~
## $ Final.price..USD.          <dbl> 1.00, 1.00, 1.00, 1.64, 1.00, 0.55, 1.64, 0.~
## $ SKU                        <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ Barcode                    <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ Cost.price                 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ Comment                    <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
```

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 purchased 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 effects revealed great predictive opportunities 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.