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.

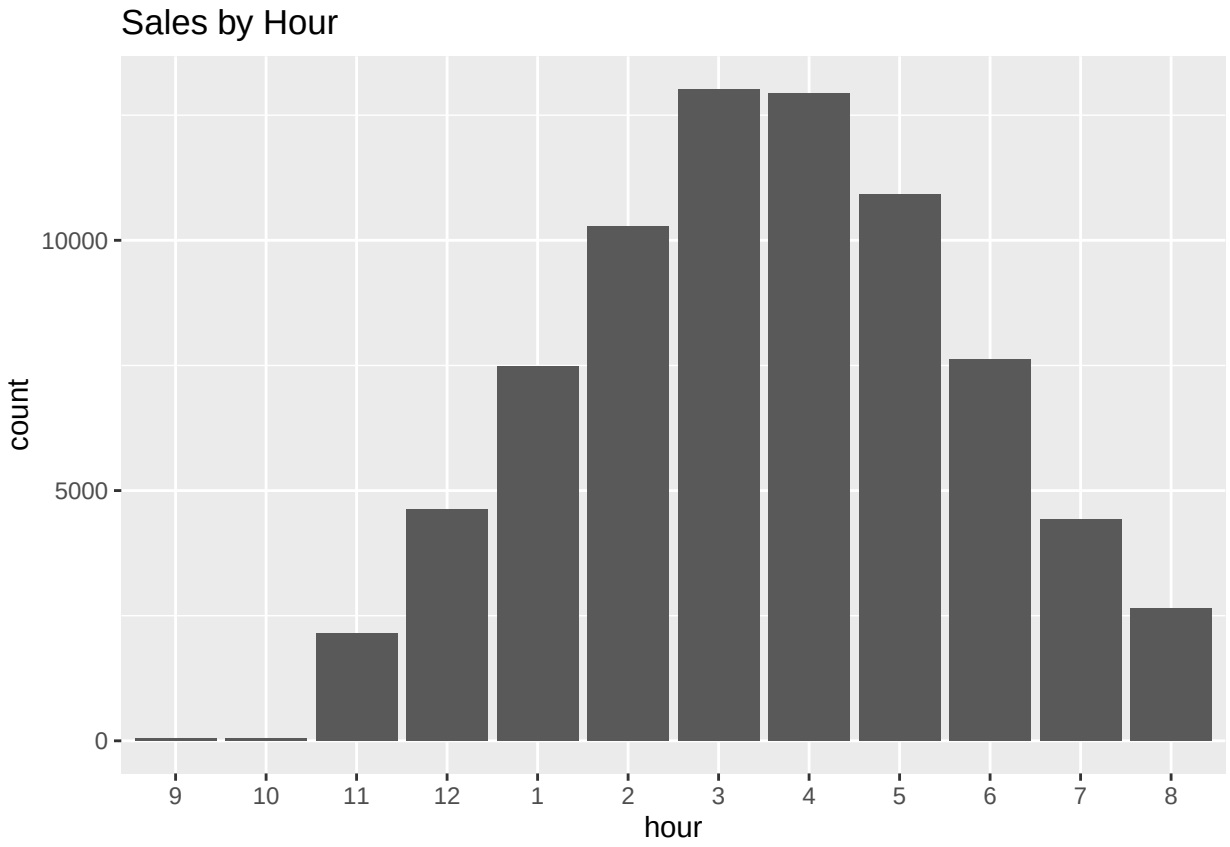
Table 1: Descriptive Summary of Numeric Variables

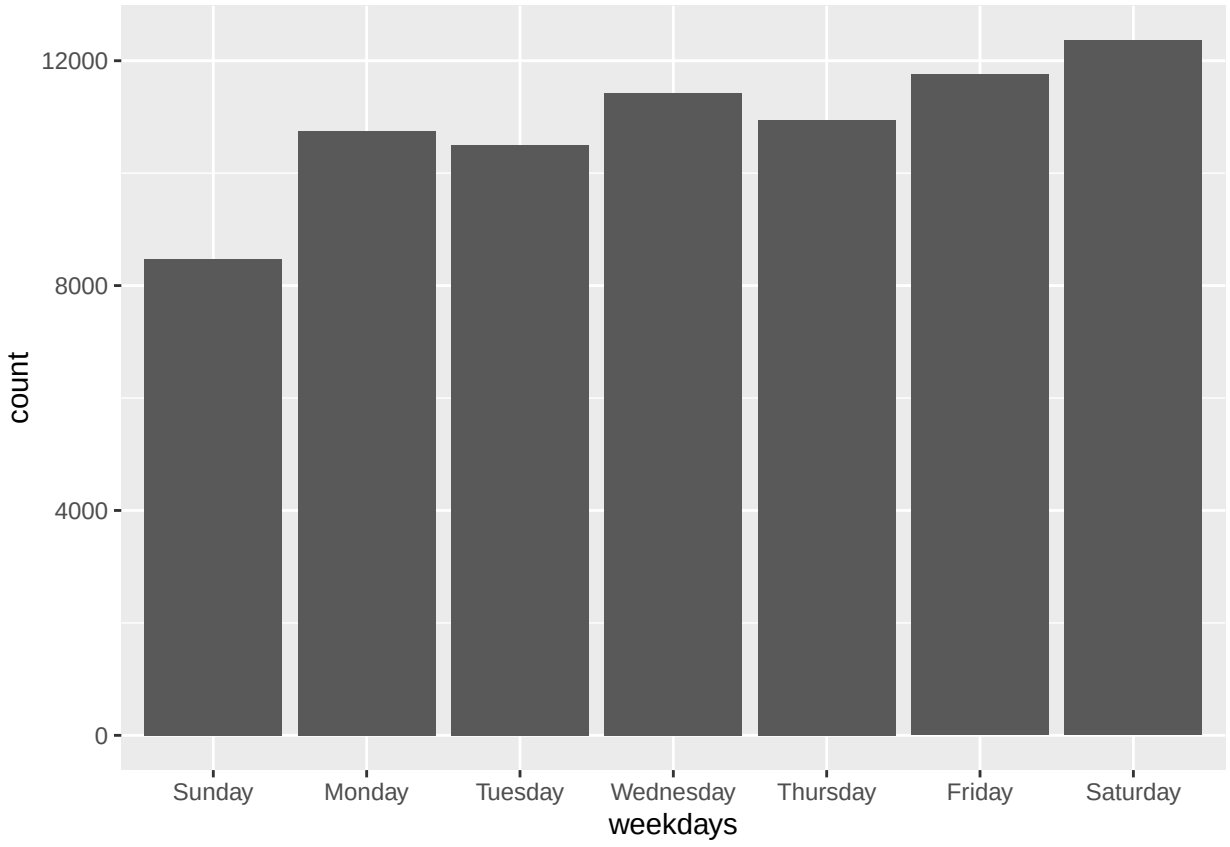
variable	n	missing	missing_pct	unique	unique_pct	mean	min	Q1	median	Q3	max	sd
Receipt_number	76219	0	0	37196	48.80	1.8e+04	1.0	9205.50	1.8e+04	2.8e+04	37196.0	1.1e+04
Quantity	76219	0	0	25	0.03	1.5e+00	-6.0	1.00	1.0e+00	2.0e+00	40.0	8.7e-01
Price	76219	0	0	79	0.10	2.6e+00	-12.0	1.00	2.0e+00	3.5e+00	104.5	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

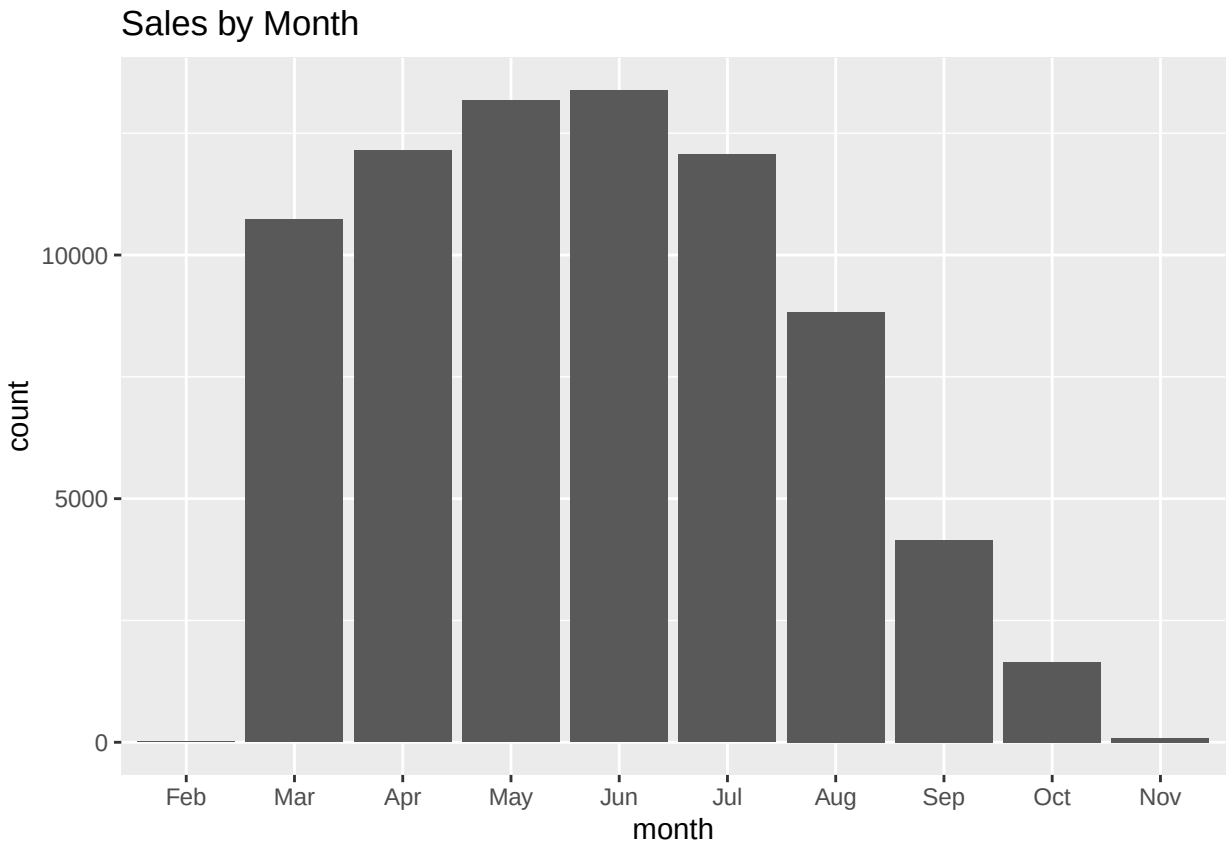
Table 2: Descriptive Summary of Categorical Variables

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

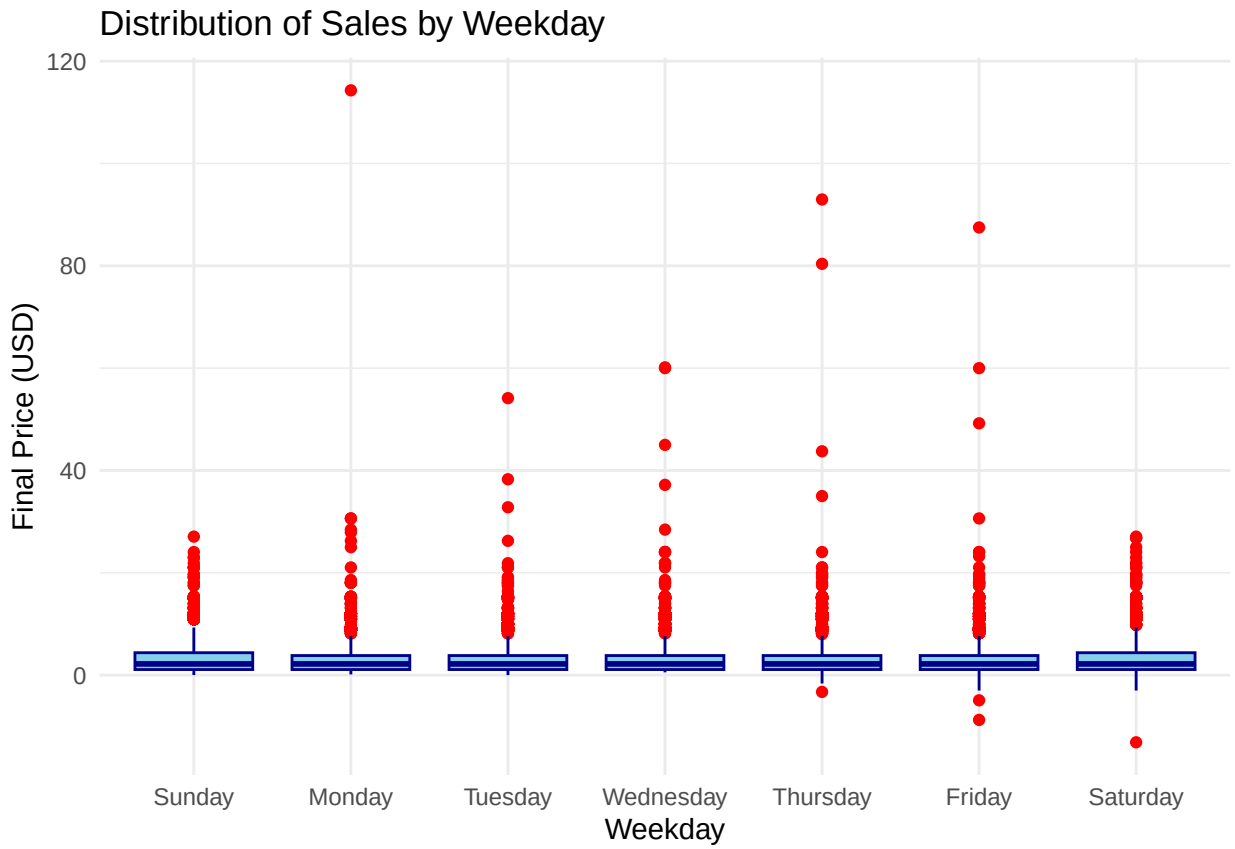
Visualizations



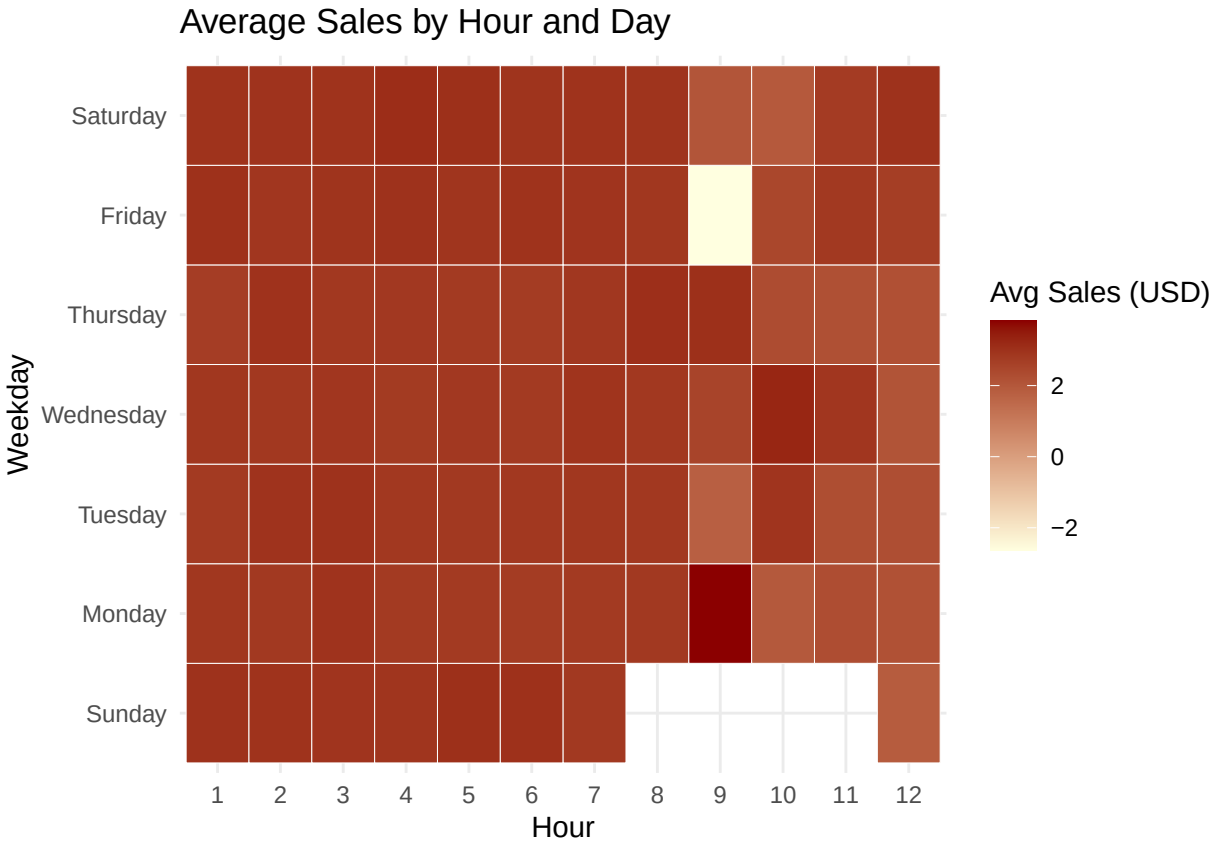




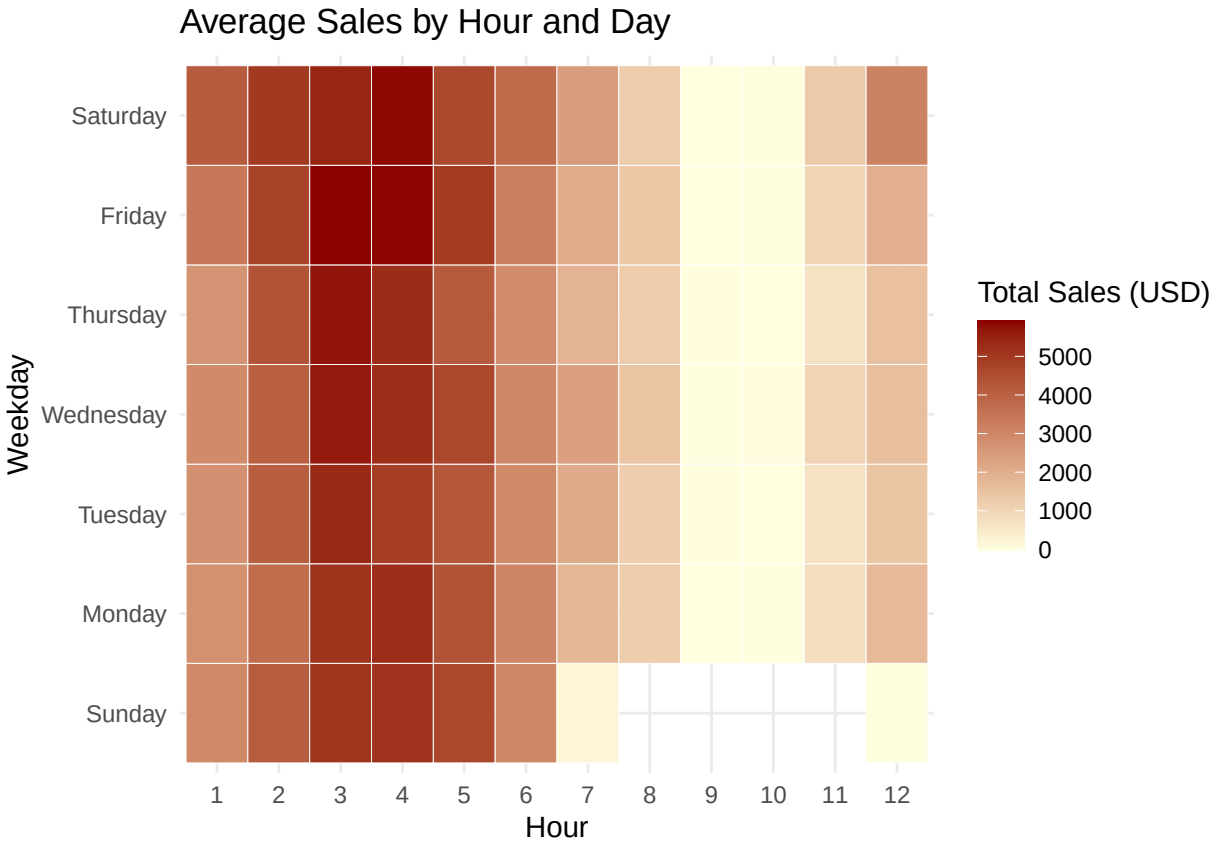
Create a box plot to show the distribution of sales by weekday This visualization helps identify variations in sales across days, highlighting any specific patterns It also shows outliers and spread, which can be useful for analyzing peak days and sales consistency



Create a heatmap to show average sales by hour and day of the week This visualization is useful for identifying peak sales times throughout the week, aiding in decisions around staffing or promotions for specific times



We were curious when the most sales occured. Instead of doing the heat map with the average, we recreated it with the total of all sales for those days and hours.


Average Sales by Hour and Day
Weekday
Saturday
Friday
Thursday
Wednesday
Tuesday
Monday
Sunday
1
2
3
4
5
6
7
8
9
10
11
12
Hour
Total Sales (USD)
5000
4000
3000
2000
1000
0

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~
```