Source - <https://www.kaggle.com/datasets/ahmedmohamed2003/cafe-sales-dirty-data-for-cleaning-training>

**Cafe Sales**

The purpose of this project is to use a spreadsheet to take dirty data, clean it, and find actionable insights for a theoretical cafe business.

Table Metadata:

| **Column Name** | **Description** | **Example Values** |
| --- | --- | --- |
| Transaction ID | A unique identifier for each transaction. Always present and unique. | TXN\_1234567 |
| Item | The name of the item purchased. May contain missing or invalid values (e.g., "ERROR"). | Coffee, Sandwich |
| Quantity | The quantity of the item purchased. May contain missing or invalid values. | 1, 3, UNKNOWN |
| Price Per Unit | The price of a single unit of the item. May contain missing or invalid values. | 2.00, 4.00 |
| Total Spent | The total amount spent on the transaction. Calculated as Quantity \* Price Per Unit. | 8.00, 12.00 |
| Payment Method | The method of payment used. May contain missing or invalid values (e.g., None, "UNKNOWN"). | Cash, Credit Card |
| Location | The location where the transaction occurred. May contain missing or invalid values. | In-store, Takeaway |
| Transaction Date | The date of the transaction. May contain missing or incorrect values. | 2023-01-01 |

Item Pricing Table:

| **Item** | **Price($)** |
| --- | --- |
| Coffee | 2 |
| Tea | 1.5 |
| Sandwich | 4 |
| Salad | 5 |
| Cake | 3 |
| Cookie | 1 |
| Smoothie | 4 |
| Juice | 3 |

**Prep**

***Goal:*** “We’re looking to find out what some of our best sellers are that produce the most revenue so we can focus on marketing them more to customers. We have sales data for the year of 2023. Could you look through it and see what you can find?”

***SMART Questions:***

* S - Is there a significant percentage difference in total revenue between the top selling item versus the second and third?
* M - How many times was the top selling item sold within the year? How does this compare to the other products?
* A - How could the lesser sold items possibly be changed to encourage customers to buy them?
* R - Is the top selling item cheaper or more expensive compared to the other items sold in the store? Does this contribute to why it sells more than the other items?
* T - Within each season, what products sold the most? What are the best and worst sales months of the year?

***Defining Limitations, Restrictions, and Guidelines:***

* As I don’t have actual communication with a cafe business to correct missing values, here is how the following columns with missing data will be treated:
  + Transaction ID, Payment Method, Location - Since it isn’t strictly needed to answer the problem, can leave ERROR and NULL values
  + Item, Transaction Date - Strictly needed to answer the problem. ERROR and NULL valued records will be deleted.
  + Quantity, Total Spent - Using separate columns, may be able to calculate ERROR and NULL values. If unable, the record will be deleted.
  + Price Per Unit - ERROR and NULL values can be corrected using the given pricing table
* Seasonal data will be determined by the following time periods:
  + December to February - Winter
  + March to May - Spring
  + June to August - Summer
  + September to November - Fall

***Cafe Sales Spreadsheet Cleaning / Analyzing Process Changelog:***

1. Created a copy of the sheet (‘dirty’) before making any changes.
2. Filtered for ERROR, UNKNOWN, and empty (NULL) values in the Item and Transaction Date columns. Deleted any records with these values.
3. Fixed Price Per Unit column

* Confirmed in the filter menu that only 8 options were available in the Item column.
* Filtered Item column for each item value along with Price Per Unit filter to find NULL and incorrect values.
* Updated all items to have correct Price Per Unit according to the prices listed in the above pricing table.
* Once all of the values were changed, checked the filter menu for the Price Per Unit column and confirmed no outliers remained.

1. Fixed Quantity and Total Spent columns

* Created a new column “Total Spent Formula” in which the price is calculated by multiplying Quantity and Price Per Item.
* Filtered the Total Spent Formula columns for 0s and errors along with filters on Quantity and Total Spent columns for any NULL values.
* Deleted any columns where Quantity and Total Spent columns both had NULL values.
* Created a new column “Quantity Formula” in which the quantity is calculated by dividing Total Spend Formula by Price Per Item.
* Filtered Quantity Formula column for any NULL or 0 values. Replaced formula to be Total Spend divided by Price Per Item.
* Once changed, checked the filter menu for the Quantity Formula column to confirm no outliers remained.
* Filtered Total Spent Formula once again for NULL or 0 values. Replaced formula to be Quantity Formula multiplied by Price Per Unit

1. Added conditional formatting to validate if Total Spent Formula was equal to Total Spent or was equal to Quantity Formula multiplied by Price Per Unit
2. In Payment Method and Location columns, changed ERROR, UNKNOWN, and empty values to NULL
3. Validated Transaction ID column

* Created a new column Transaction ID Len and determined that the length of a Transaction ID is 11 characters long
* Looked at the filter menu for the Len column and determined there weren’t any outliers

1. Cleaned up helper columns

* Duplicated and created sheet ‘helper\_columns’
* Pasted values of Total Spent Formula and Quantity Formula columns into their respective original columns
* Deleted all helper columns

1. Formatted data within columns

* Price Per Unit, Total Spent - Currency
* Transaction Date - Custom Date (YYYY-MM-DD)

1. Used ‘Remove duplicates’ and Trim whitespace’ Data tools. Both returned no results
2. Created helper column “Month” to assist in creating seasonal pivot tables
3. Created pivot tables for total and quarterly in a new sheet ‘cafe\_sales\_pt’ to create pivot and summary tables for the data
4. Created summary tables to compare the top three best and worst selling items annually.
5. Created summary tables to compare the top and bottom three selling items to each other.
6. Created pie graphs comparing annual total Quantity and Total Spent metics
7. Created a new sheet for monthly pivot tables and created a summary table for monthly units sold and monthly sales.
8. Created graphs on monthly sales pivot tables sheet

***Conclusions:***

For the entire year of 2023, the annual sales totaled $76,831 with a combined total of 25,988 units sold for every item.

Best selling

Salads were the best selling item with annual sales of $16,575 with 3,315 units sold. That means that Salads accounted for 21.57% of the total sales for the year while being 12.76% of total annual units sold.

The second best selling item was Sandwiches with annual sales of $13,068 with 3,267 units sold. That means that Sandwiches accounted for 17.01% of total sales for the year while being 12.57% of total annual units sold.

The third best selling item was Smoothies with annual sales of $12,776 with 3,194 units sold. That means that Smoothies accounted for 16.63% of total sales for the year while being 12.29% of total annual units sold.

Comparing the sales of the top three selling items, Salads sold $3,507 (26.84%) more than Sandwiches and 48 (1.47%) more units. Salads sold $3,799 (29.74%) more than Smoothies and 121 (3.79%) more units.

Worst Selling

Cookies were the worst selling item with annual sales of $3,076 with 3,076 units sold. That means Cookies accounted for 4.00% of the total sales for the year while being 11.84% of total annual units sold.

The second worst selling item was Tea with annual sales of $4,665 with 3,110 units sold. That means that Tea accounted for 6.07% of total sales for the year while being 11.97% of total annual units sold.

The third worst selling item was Coffee with annual sales of $6,814 with 3,407 units sold. That means that Coffee accounted for 8.87% of total sales for the year while being 13.11% of total annual units sold.

Comparing the sales of the bottom three selling items, Tea sold $1,589 (51.66%) more than Cookies and 34 (1.11%) more units. Coffee sold $3,738 (121.52%) more than Cookies and 331 (10.76%) more units.

Seasonal Sales

Comparatively speaking, the total sales per season showed no significant difference. Winter had $19,079.50 in sales with 6,457 units sold, Spring had $19,211.00 in sales with 6,468 units sold, Summer had $19,139.50 in sales with 6,474 units sold, and Fall had $19,401.00 in sales with 6,589 units sold.

Each season matched the annual trends exactly. Salads, Sandwiches, and Smoothies were the top three selling items while Cookies, Teas, and Coffee were bottom three.

Monthly Sales

However, there were some slight variations in the monthly sales figures. The top three months for sales were October sales of $6,755 with 2,326 units sold, January sales of $6,690 with 2,233 units sold, and June sales of $6,616 with 2,216 units sold. The bottom three months for sales were February sales of $5,953.50 with 2,033 units sold, July sales of $6,175.50 with 2,099 units sold, and September sales of $6,208 with 2,107 units sold.

Best Monthly Sales

October’s top three items were Salad, Soothie, and Sandwich while the bottom three items were Cookie, Tea, and Coffee. This mirrors the annual sales trend other than Smoothies being the second best selling item and Sandwiches being the third best selling item.

January’s top three items were Salad, Sandwich, and Juice while the bottom three items were Cookie, Tea, and Coffee. This mirrors the annual sales trend except for Juice replacing Smoothies as the third best selling item. A hypothesis for why this could be is that less customers may buy Smoothies in a cold month like January. However, looking at the total sales figures for the season of Winter shows that the season as a whole mirrors the annual sales trend.

June’s top three items were Salad, Smoothie, and Sandwich while the bottom three items were Cookie, Tea, and Coffee. Again, mirrors the annual trend with the exception of Smoothies and Sandwiches switching each other’s place.

Worst Monthly Sales

February’s top three items were Salad, Sandwich, and Smoothie while the bottom three items were Cookie, Tea, and Coffee. This matches the annual trend exactly.

July’s top three items were Salad, Smoothie, and Sandwich while the bottom three items were Cookie, Tea, and Coffee. This matches annual trends besides Smoothie and Sandwich places.

September’s top three items were Salad, Smoothie, and Sandwich while the bottom three items were Cookie, Tea, and Coffee. This matches the annual trend besides Smoothie and Sandwich places.

Monthly Conclusions

For the most part, monthly trends matched annual trends. For both the top three and bottom three months, Cookies, Tea, and Coffee sold the least for the month. While the top three items matched annual trends most of the time, sometimes Smoothies would be second best selling and Sandwiches would be third best selling. The only outlier in this regard is January where Juice was the third best selling item. A hypothesis for why this could be is that less customers may buy Smoothies in a cold month like January. However, looking at the total sales figures for the season of Winter shows that the season as a whole mirrors the annual sales trend.

Project Conclusion

Consistently, Salads, Sandwiches, and Smoothies are top selling items while Cookies, Tea, and Coffee are the least sold. If the company wishes to market their top selling items, they should therefore focus on marketing Salads, Sandwiches, and Smoothies. All three of the least selling items are the cheapest items on the menu. Maybe some sort of marketing campaign could be created by combining an expensive item with a cheaper item such as:

* “Buy a Salad, get a Tea or Coffee free”
* “Buy a Sandwich, get a Cookie free”
* “Buy a Smoothie, get a Cookie free”

Or could combine two expensive items with one at a discount such as:

* “Buy a Salad, get a Smoothie half off”

These would not only interest current customers who frequent the shop but could convince new customers to come and try some cafe items based on the deals.

As for technical conclusions, I realized once I was finished that, alternatively, I could have created a separate sheet with the pricing table and created a formula column in cafe\_sales that pulled the item price based on item name. This would have added a layer of validation to the item prices instead of manually changing prices with filters.