

# FitHack DA Internship assessment

## Analysis report

Initial look at the dataset reveals that the dataset consisted of sales data or invoices of various products categorized by certain broad categories. The dataset consisted of 1000 entries and had 17 features starting from Invoice ID to Rating of customers. The dataset didn't have any missing values or null values. The date and time columns were converted to Datetime objects and the analysis was set to motion. The goal of the analysis was to determine the factors that contribute towards a better gross margin percentage but as it turned out the gross margin percentage was constant throughout the dataset, hence other methods were adopted for noticing the influence of various factors.

The influence of these factors on gross income was taken into account. The dependencies of gross income was determined by using a correlation heatmap ( using correlation matrix ). It was noticed that the dataset was **augmented** with various calculated metrics rather than observed metrics like Tax 5%, Total, etc. Overlooking the metrics redundancies, it was found out the gross income is significantly positively correlated with **Unit Price and Quantity**. The contribution of customer ratings on gross income is negligible. Let's look at the analysis deeply.

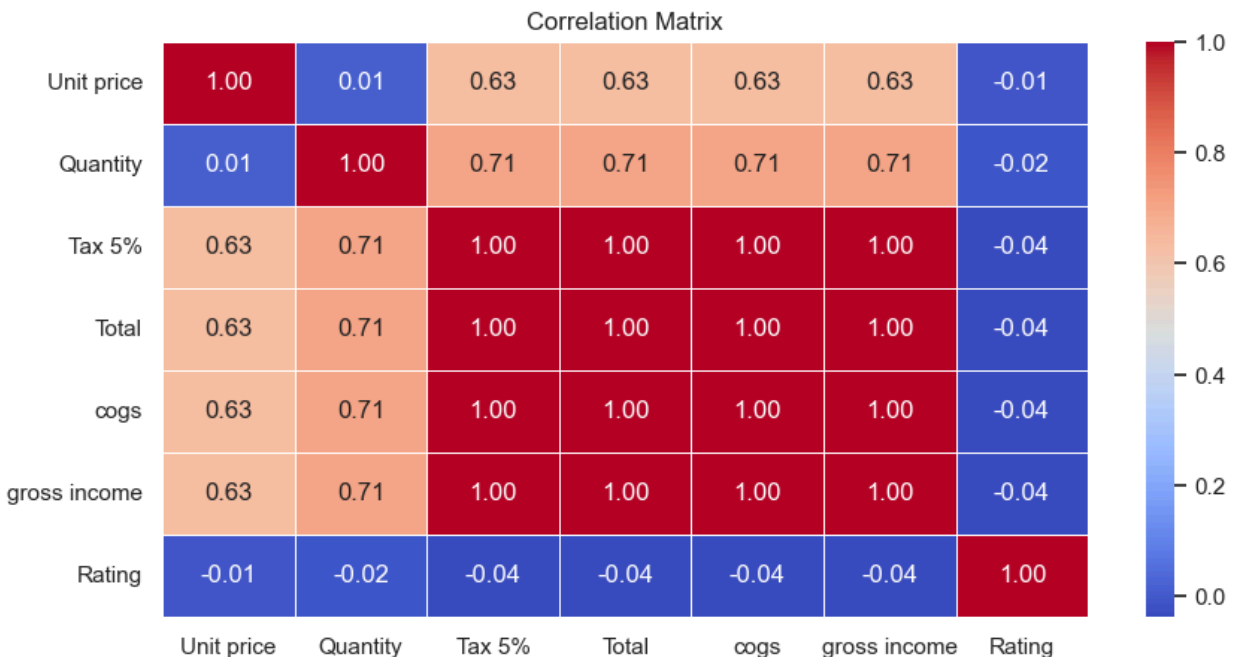
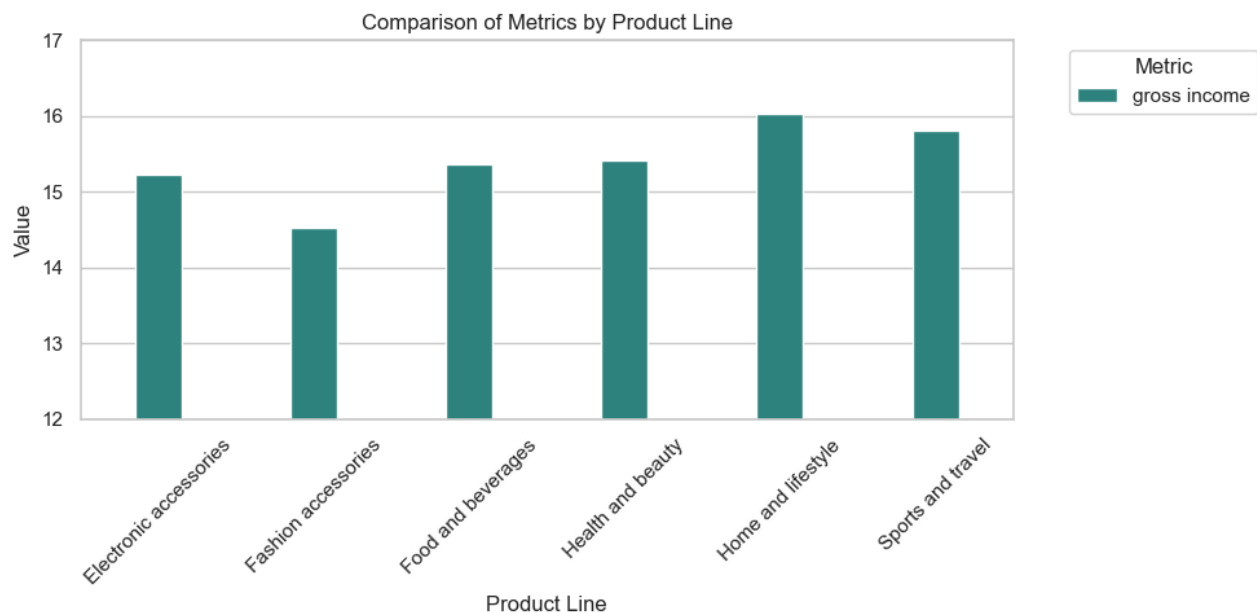


Figure 1 : Correlation heatmap

**Feature engineering :** A ‘Day’ column was added to conduct an analysis of which days contribute significantly to the gross income. Final dataset was cleaned and didn’t contain redundant calculated metrics.

## Approach and analysis

**Product based analysis** - The dataset was grouped by ‘Product Line’ and the corresponding average values of gross income, total, unit price of each product category was observed. It was found that ‘**Home and Lifestyle**’ came out on top in terms of capturing maximum average gross income, indicating strong profitability. ‘**Fashion Accessories**’ have the lowest average gross income, suggesting potential for improvement in this category. ‘**Sports and travel**’ is also turning in a significant amount of profitability.



*Figure 2 : Product line comparison*

**Payment method analysis** - The dataset was grouped by ‘Payment Method’ and the corresponding average values of similar metrics of each payment method was observed. Slightly unusual trend was noticed given the digitisation of payment methods nowadays. ‘**Cash**’ based transactions turned in the maximum amount of gross income, followed by ‘**Credit Card**’ and surprisingly ‘**Ewallet**’ in the last.

**Gender based analysis** - The dataset was grouped by ‘Gender’ and the corresponding average values of similar metrics of each gender was observed. ‘**Female**’ customers were fairly higher in frequency and the amount spent while buying products.



Figure 3 : Gender based distribution consisting of member and normal information

**Daily Sales Analysis** - The dataset was grouped by augmented 'Day' and the corresponding average values of similar metrics of each day was observed. On 'Saturday's the sales were higher and that resulted in generating greater profitability than any other day. The second significant performer in this analysis was 'Sunday'.

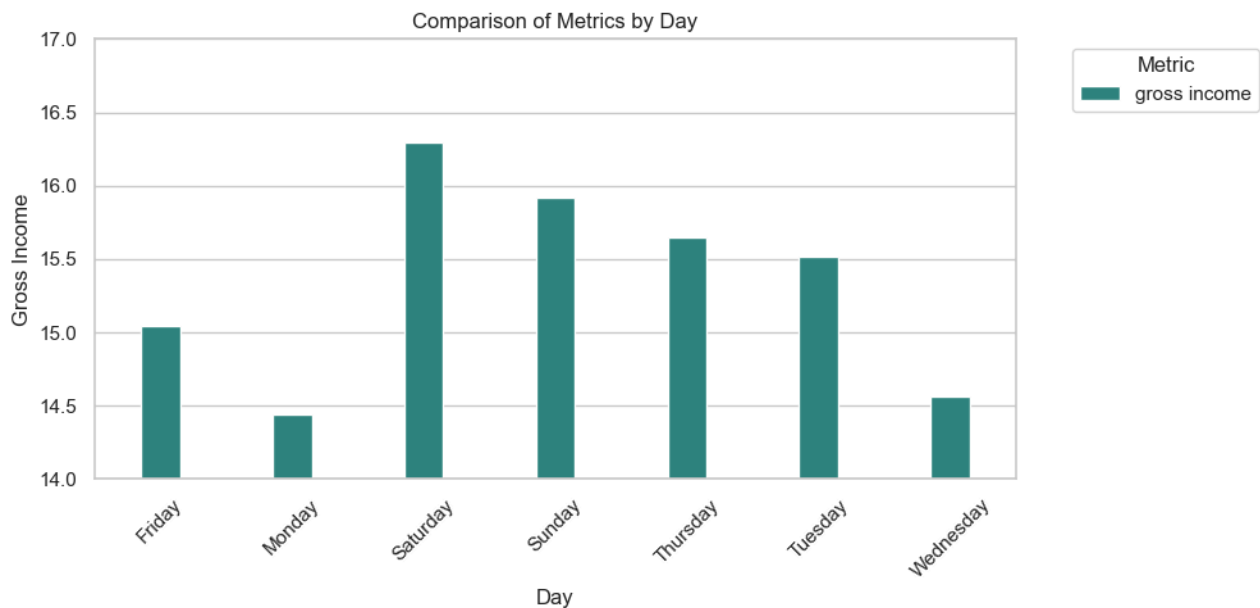


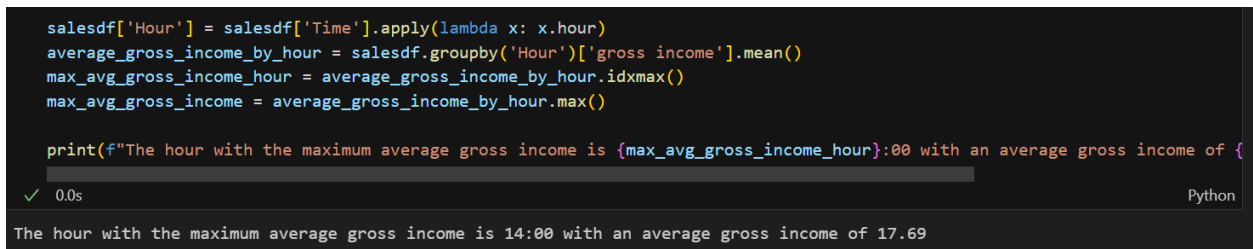
Figure 4: Day based comparison

**City based analysis** - The dataset was grouped by 'City' and the corresponding average values of similar metrics of each city present in the dataset was observed. It was observed that 'Naypyitaw' was the hottest city in terms of generating profitability.

**Member based analysis** - The dataset was grouped by 'Member' and the corresponding average values of similar metrics of each category was observed. It was observed that there was no significant difference between generated profitability by group of 'Member' customers and normal customers

**Branch based analysis** - The dataset was grouped by 'Branch' and the corresponding average values of similar metrics of each branch was observed. Branch 'C' was able to turn in maximum average gross income followed by branch 'B'

**Time based analysis** - For this the hour from the time column is pulled out using a lambda function and the dataset is grouped by this new 'hour' column. Similar average of the metrics like 'gross income' was calculated. It was found that the hottest time was **14:00** with average gross income reaching the maximum of **17.69**.



```
salesdf['Hour'] = salesdf['Time'].apply(lambda x: x.hour)
average_gross_income_by_hour = salesdf.groupby('Hour')['gross income'].mean()
max_avg_gross_income_hour = average_gross_income_by_hour.idxmax()
max_avg_gross_income = average_gross_income_by_hour.max()

print(f"The hour with the maximum average gross income is {max_avg_gross_income_hour}:00 with an average gross income of {max_avg_gross_income}")
```

✓ 0.0s Python

The hour with the maximum average gross income is 14:00 with an average gross income of 17.69

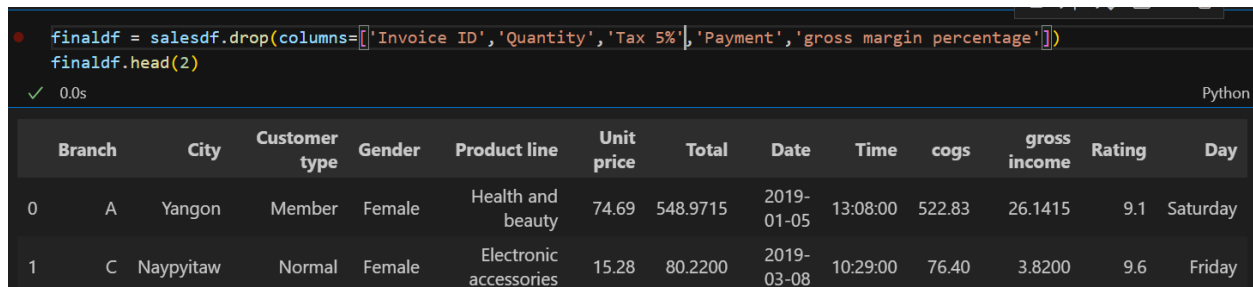
*Figure 5 : Time based result*

## Results and recommendations

1. Given that 'Home and Lifestyle' has the highest average gross income, it is worth considering expanding this product line by introducing new products, upselling, and cross-selling related items. Marketing efforts could be intensified for this category.
2. Focusing on 'Sports and Travel' is recommended. Since this category is also profitable, it may be worthwhile to explore opportunities to further enhance its product line, such as seasonal promotions or partnerships with sports and travel-related brands.
3. For 'Fashion accessories', analyzing whether certain products are underperforming due to pricing, quality, or market demand. Adjustments could include re-pricing, introducing more trendy items, or creating bundle deals.
4. It is recommended to encourage usage of digital payments through targeted promotions, discounts, loyalty rewards etc. This could lead to operational efficiency and reduce cash-handling costs. It is also recommended to consider customer surveys or market research to understand why people are having such problems with adoption.

5. Since female customers are contributing more to sales, they could be subject to targeted marketing campaigns that resonate with female shoppers, offering promotions on products that appeal to them. It is also recommended to explore male market reach.
6. Capitalizing on the higher sales during weekends (Saturday and Sunday) by running special weekend promotions or events that could further boost sales and profitability.
7. It is also recommended to introduce incentives for weekday shopping, such as mid-week discounts or loyalty points, to balance sales distribution throughout the week and increase overall profitability.
8. Since *Naypyitaw* is the most profitable city, new branches could be opened which would result in marketing efforts. It is recommended to reinforce the branches by increasing product availability.
9. Since Branch C is the top performer, analyzing what contributes to its success, be it location, customer demographics, product assortment, or staff performance and trying to replicate these factors across other branches would do the trick.
10. Since the hour of 14:00 sees the highest gross income, we can consider introducing time-limited offers or happy hours around this time to further maximize sales during peak hours.

Generally, continuing data driven decision making and reviewing the data to stay responsive to changing market trends is something very vital for any business. After removing the calculated metrics and removing some feature redundancies I found that this could be the perfect dataset recorded for further analysis.



```
finaldf = salesdf.drop(columns=['Invoice ID', 'Quantity', 'Tax 5%', 'Payment', 'gross margin percentage'])
finaldf.head(2)
```

	Branch	City	Customer type	Gender	Product line	Unit price	Total	Date	Time	cogs	gross income	Rating	Day
0	A	Yangon	Member	Female	Health and beauty	74.69	548.9715	2019-01-05	13:08:00	522.83	26.1415	9.1	Saturday
1	C	Naypyitaw	Normal	Female	Electronic accessories	15.28	80.2200	2019-03-08	10:29:00	76.40	3.8200	9.6	Friday

*Figure 6 : Final dataframe*

## Some interesting speculations about the analysis

Some speculations from the analysis suggest that 'Fashion Accessories' might be struggling because the products aren't resonating with customers, or prices are too high. The fact that cash transactions are outperforming digital payments is surprising—maybe people in this market just prefer using cash, or perhaps digital payments aren't as convenient. The low performance of 'Ewallet' could mean customers aren't fully comfortable with it yet. The similar profitability

between members and non-members hints that the membership program might need better perks. Lastly, Naypyitaw's strong performance might be because of less competition or a more prosperous local economy.