

# **Artificial Intelligence (CSEN933)**

## **Project 1**

### **Submitted by:**

Ali Abdrabou 52-1551

Karim Ossama 52-2402

Mohammed Abdelmoneam 52-4804

Omar Hegazy 52-9494

Yahia Shalaby 52-0556

**Date:** 20/10/2023

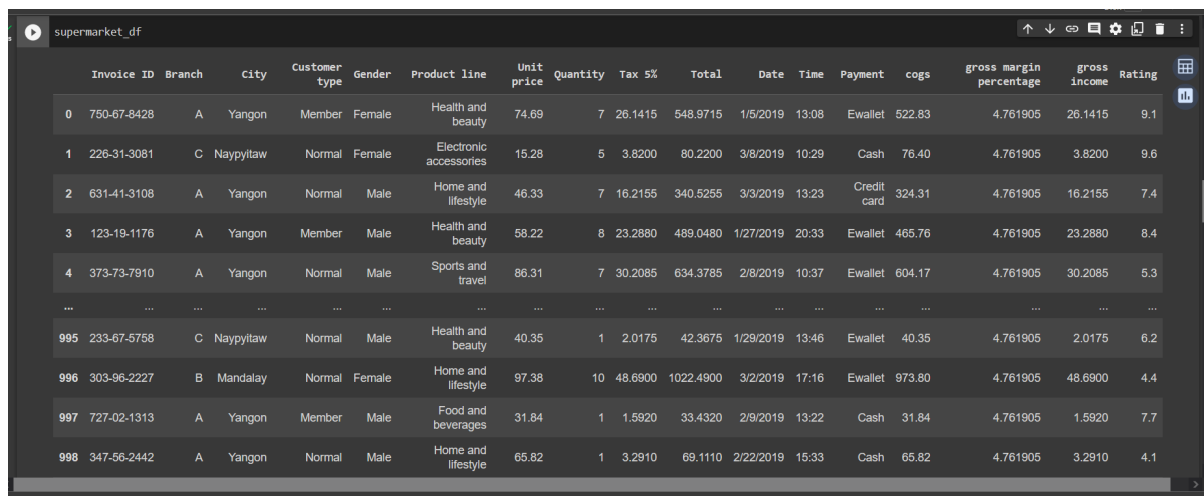
**1) Provide a discussion about the problem of choice, The challenges why this problem is important, and the methodology used in assessing such a problem.**

We attempted to tackle the question of whether relationships exist between different product types for example: the probability of a transaction purchasing product type A being followed by one of product type B and whether B will be followed by C, D, etc. Having this knowledge is important because it allows for the optimization of product displays in supermarkets and producing more enticing bundles therefore maximizing profit. We went about this using a Markov model wherein each transaction of a given product type was considered a state, computing the probability of one transaction following another (one state transitioning to another) after plugging in our data into a matrix.

**2) You should provide a discussion about the reasoning behind the choices made for the dataset of Kaggle, in order for it to be taken as the dataset for this case study.**

The data set used provided the transactional data of three different branches however we opted to use the data of a single branch (branch A) for the sake of brevity, we then picked out the product line attribute as it was the only relevant attribute when it comes to the needed calculations. The data set was properly formatted from the get-go which meant it was fit to use directly without the need for heavy data preparations or re-format.

**Screenshot of the chosen data set:**



	Invoice ID	Branch	City	Customer type	Gender	Product line	Unit price	Quantity	Tax 5%	Total	Date	Time	Payment	cogs	gross margin percentage	gross income	Rating
0	750-67-8428	A	Yangon	Member	Female	Health and beauty	74.69	7	26.1415	548.9715	1/5/2019	13:08	Ewallet	522.83	4.761905	26.1415	9.1
1	226-31-3081	C	Naypyitaw	Normal	Female	Electronic accessories	15.28	5	3.8200	80.2200	3/8/2019	10:29	Cash	76.40	4.761905	3.8200	9.6
2	631-41-3108	A	Yangon	Normal	Male	Home and lifestyle	46.33	7	16.2155	340.5255	3/3/2019	13:23	Credit card	324.31	4.761905	16.2155	7.4
3	123-19-1176	A	Yangon	Member	Male	Health and beauty	58.22	8	23.2880	489.0480	1/27/2019	20:33	Ewallet	465.76	4.761905	23.2880	8.4
4	373-73-7910	A	Yangon	Normal	Male	Sports and travel	86.31	7	30.2085	634.3785	2/8/2019	10:37	Ewallet	604.17	4.761905	30.2085	5.3
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
995	233-67-5758	C	Naypyitaw	Normal	Male	Health and beauty	40.35	1	2.0175	42.3675	1/29/2019	13:46	Ewallet	40.35	4.761905	2.0175	6.2
996	303-96-2227	B	Mandalay	Normal	Female	Home and lifestyle	97.38	10	48.6900	1022.4900	3/2/2019	17:16	Ewallet	973.80	4.761905	48.6900	4.4
997	727-02-1313	A	Yangon	Member	Male	Food and beverages	31.84	1	1.5920	33.4320	2/9/2019	13:22	Cash	31.84	4.761905	1.5920	7.7
998	347-56-2442	A	Yangon	Normal	Male	Home and lifestyle	65.82	1	3.2910	69.1110	2/22/2019	15:33	Cash	65.82	4.761905	3.2910	4.1

### Screenshot of the chosen data set after reducing it to branch A:

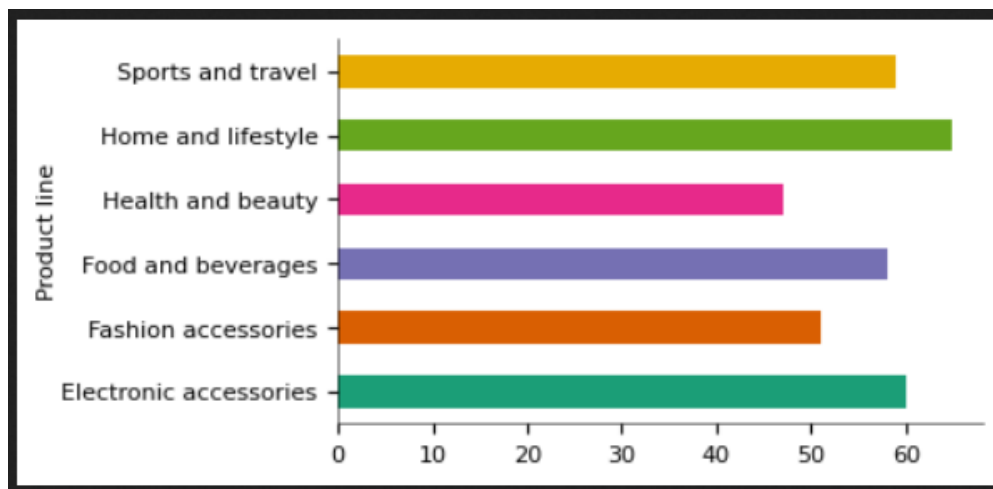
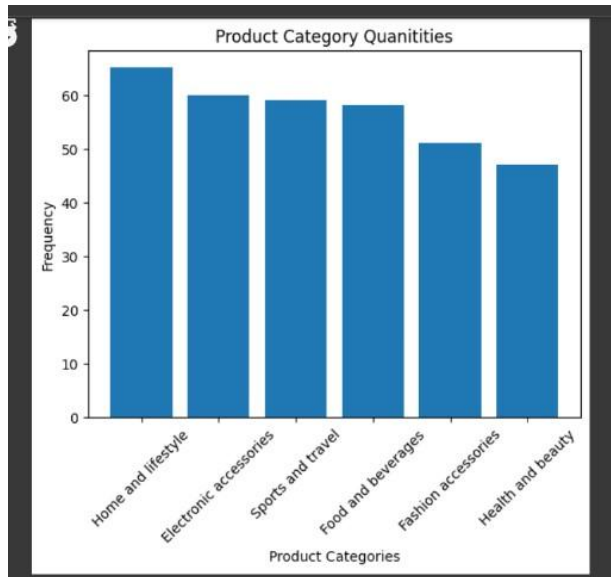
	Invoice ID	Branch	City	Customer type	Gender	Product line	Unit price	Quantity	Tax 5%	Total	Date	Time	Payment	cogs	gross margin percentage	gross income	Rating
0	750-67-8428	A	Yangon	Member	Female	Health and beauty	74.69	7	26.1415	548.9715	1/5/2019	13:08	Ewallet	522.83	4.761905	26.1415	9.1
2	631-41-3108	A	Yangon	Normal	Male	Home and lifestyle	46.33	7	16.2155	340.5255	3/3/2019	13:23	Credit card	324.31	4.761905	16.2155	7.4
3	123-19-1176	A	Yangon	Member	Male	Health and beauty	58.22	8	23.2880	489.0480	1/27/2019	20:33	Ewallet	465.76	4.761905	23.2880	8.4
4	373-73-7910	A	Yangon	Normal	Male	Sports and travel	86.31	7	30.2085	634.3785	2/8/2019	10:37	Ewallet	604.17	4.761905	30.2085	5.3
6	355-53-5943	A	Yangon	Member	Female	Electronic accessories	68.84	6	20.6520	433.6920	2/25/2019	14:36	Ewallet	413.04	4.761905	20.6520	5.8
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
990	886-18-2897	A	Yangon	Normal	Female	Food and beverages	56.56	5	14.1400	296.9400	3/22/2019	19:06	Credit card	282.80	4.761905	14.1400	4.5
992	745-74-0715	A	Yangon	Normal	Male	Electronic accessories	58.03	2	5.8030	121.8630	3/10/2019	20:46	Ewallet	116.06	4.761905	5.8030	8.8
997	727-02-1313	A	Yangon	Member	Male	Food and beverages	31.84	1	1.5920	33.4320	2/9/2019	13:22	Cash	31.84	4.761905	1.5920	7.7
998	347-56-2442	A	Yangon	Normal	Male	Home and lifestyle	65.82	1	3.2910	69.1110	2/22/2019	15:33	Cash	65.82	4.761905	3.2910	4.1
999	849-09-3807	A	Yangon	Member	Female	Fashion accessories	88.34	7	30.9190	649.2990	2/18/2019	13:28	Cash	618.38	4.761905	30.9190	6.6

### Screenshot of the final data we'll use after choosing the "Product Line" attribute:

```
0      Health and beauty
2      Home and lifestyle
3      Health and beauty
4      Sports and travel
6      Electronic accessories
...
990    Food and beverages
992    Electronic accessories
997    Food and beverages
998    Home and lifestyle
999    Fashion accessories
Name: Product line, Length: 340, dtype: object
```

**3) You should provide a discussion about the importance of the attributes which are chosen as the attributes upon which the predictive model is based, alongside the attributes to be predicted.**

The attribute used clusters differing products into set product groups or types that can then have relations drawn between them in a more uniform less sporadic manner (in contrary to, for instance, attempting to draw relations between every single product which would significantly increase complexity/reduce efficiency), we are trying to reach customer buying trends so for instance the probability between fashion accessory transactions being followed by a sports and travel related purchase or health and beauty leading to home and lifestyle purchases, etc. However, the attribute used is one of a categorical nature and so there are no minimum, maximum and mean values and so only mode and median operations were performed followed by a histogram being graphed.



```
[5] def get_ordinal_column_median(df, column_name):
    df_sorted = df.sort_values(by=column_name)
    n = len(df_sorted)
    if n % 2 == 0:
        return df_sorted.iloc[n // 2][column_name]
    else:
        return df_sorted.iloc[n+1// 2][column_name]

    ## MEDIAN AND MODE ARE PRINTED HERE!!!!!!!!!!!!!!
    median = get_ordinal_column_median(supermarket_A_df, 'Product line')
    print("The median is: "+median)
    mode = supermarket_A_df['Product line'].mode()[0]
    print("The mode is: "+mode)
```

```
The median is: Health and beauty
The mode is: Home and lifestyle
```

```
[6] freqTable = supermarket_A_df['Product line'].value_counts()
```

**4) You should provide a discussion about the conclusions documenting the end results and probabilities produced by the constructed Markov model showcasing whether the achieved results are satisfactory or not**

The model proved satisfactory, as it utilizes the probability matrix created on the chosen attribute, along with the number of purchases that we would like to view which product will most likely be purchased after (number of steps), and finally a starting product or position. After testing the model with several different starting positions and steps we were able to figure out which products usually end up being in the same purchase pattern or sequence, which ultimately allows us to better understand our customers. However, the model goes into an equilibrium after around 5-6 purchases, which indicates that the model can be better utilized for short-term predictions rather than long-term ones, which still suits our application as knowing the next 2-3 purchases is far more important than predicting the 20th purchase in the case of a supermarket run. In conclusion, we can proudly say the model worked adequately.

```
startingCategory = categoryList[np.where(vector == 1)[0][0]]
print("Starting with a purchase under "+startingCategory+",")
nextPurchase(result,n)
```

```
[0.16      0.18965517 0.18461538 0.18644068 0.18333333 0.25531915]
```

```
Starting with a purchase under Home and lifestyle,
```

```
After 1 purchase(s), it is most likely that a product under the, Health and beauty product line(s) will be chosen
with probability 0.2553191489361702
```