Methodology

1) Purchase Data -Finding Items Costs

- I started with loading the Purchase data sheet of the Excel file.
- I have chosen Candies, Mangoes and Milk Packets as the input features (X) and Payment as the output (y).
- Then I computed X rank to verify the information of the matrix.
- The pseudo-inverse method was then applied to estimate the unit cost of each item (Candies, Mangoes, Milk Packets).

2) Customer Classification RICH or POOR

- I have added a new column that is called Class.
- In case the payment exceeds 200 rupees, the customer is declared as RICH.
- Otherwise, the customer is rated as POOR.

3) IRCTC Stock Data Mean, Variance, Probability

- I loaded the Stock Price sheet of the IRCTC.
- I took out the Price column and did the following calculations:

1. Mean
2. Variance

- To find mean and variance, I computed it in 2 ways:

1. Using NumPy
2. Using my own functions

- Then I compared their time which was taken with 10 runs.
- I also found:

1. Average price on Wednesdays
2. Average price in April
3. With the Chg% column, I determined:
4. Probability of loss (Chg% < 0)
5. Wednesday (Chg% > 0 on Wed) probability of making profits.

4) Visualization

- I plotted a scatter graph:
- X-axis: Day

- Y-axis: Chg%
- This assists in the knowledge of the stock change variation throughout the week days.

5) Thyroid Data set: Data Checking

- I loaded thyroid0387 UCI sheet.
- I checked:
- Types of data (categorical and numerical).
- The minimum and maximum values of numeric column.
- Absent values in the columns.
- Outliers using IQR method
- Mean and variance of all numeric features were also computed by me.

6) Similarity Measures

- The two initial records I chose are of the dataset on thyroid.
- On binary (0/1) columns, I calculated:
- Jaccard similarity
- Simple Matching Coefficient (SMC)
- On numeric columns, I worked out:
- Cosine similarity

7) Heatmap for Similarity

- I selected the first 20 records.
- I developed a Jaccard similarity matrix on them.
- I then used Seaborn to plot the similarity as a heatmap so that it could be easily visualized.

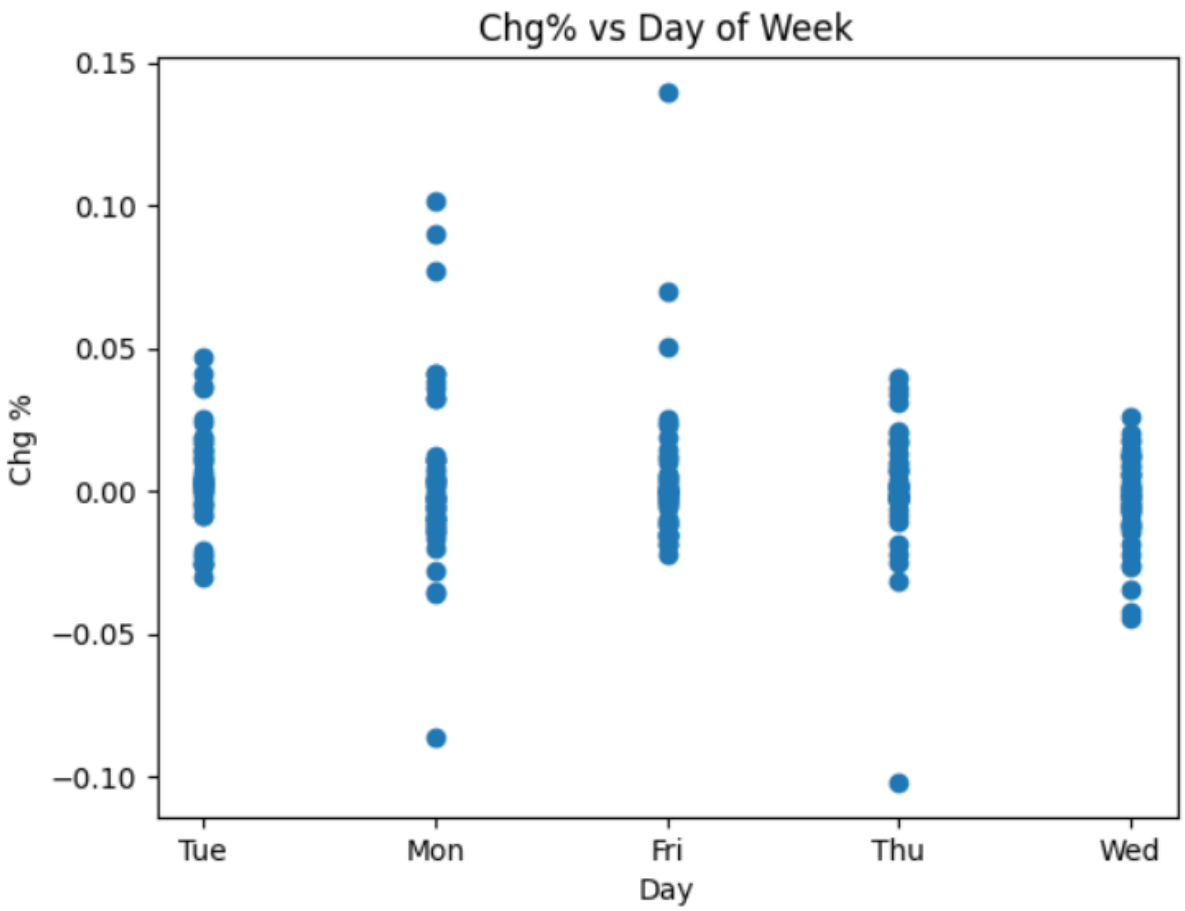
```
...    Customer  Candies (#)  Mangoes (Kg)  Milk Packets (#)  Payment (Rs)  \
0          C_1           20             6                 2           386
1          C_2           16             3                 6           289
2          C_3           27             6                 2           393
3          C_4           19             1                 2           110
4          C_5           24             4                 2           280
5          C_6           22             1                 5           167
6          C_7           15             4                 2           271
7          C_8           18             4                 2           274
8          C_9           21             1                 4           148
9         C_10           16             2                 4           198

   Unnamed: 5  Unnamed: 6  Unnamed: 7  Unnamed: 8  Unnamed: 9  ...  \
0         NaN         NaN         NaN         NaN         NaN  ...
1         NaN         NaN         NaN         NaN         NaN  ...
2         NaN         NaN         NaN         NaN         NaN  ...
3         NaN         NaN         NaN         NaN         NaN  ...
4         NaN         NaN         NaN         NaN         NaN  ...
5         NaN         NaN         NaN         NaN         NaN  ...
6         NaN         NaN         NaN         NaN         NaN  ...
7         NaN         NaN         NaN         NaN         NaN  ...
8         NaN         NaN         NaN         NaN         NaN  ...
9         NaN         NaN         NaN         NaN         NaN  ...
```

```
...   Unnamed: 12  Unnamed: 13  Unnamed: 14  Unnamed: 15  Unnamed: 16  \
0           NaN          NaN          NaN          NaN          NaN
1           NaN          NaN          NaN          NaN          NaN
2           NaN          NaN          NaN          NaN          NaN
3           NaN          NaN          NaN          NaN          NaN
4           NaN          NaN          NaN          NaN          NaN
5           NaN          NaN          NaN          NaN          NaN
6           NaN          NaN          NaN          NaN          NaN
7           NaN          NaN          NaN          NaN          NaN
8           NaN          NaN          NaN          NaN          NaN
9           NaN          NaN          NaN          NaN          NaN

   Unnamed: 17  Unnamed: 18  Candy  Mango  Milk
0          NaN          NaN    1.0   55.0  18.0
1          NaN          NaN    NaN    NaN   NaN
2          NaN          NaN    NaN    NaN   NaN
3          NaN          NaN    NaN    NaN   NaN
4          NaN          NaN    NaN    NaN   NaN
5          NaN          NaN    NaN    NaN   NaN
6          NaN          NaN    NaN    NaN   NaN
7          NaN          NaN    NaN    NaN   NaN
8          NaN          NaN    NaN    NaN   NaN
9          NaN          NaN    NaN    NaN   NaN
```

```
[10 rows x 22 columns]
Features (X):
[[20  6  2]
 [16  3  6]
 [27  6  2]
 [19  1  2]
 [24  4  2]
 [22  1  5]
 [15  4  2]
 [18  4  2]
 [21  1  4]
 [16  2  4]]
Payment (y):
[386 289 393 110 280 167 271 274 148 198]
Rank of Feature Matrix: 3
Cost of Candies     : 0.9999999999999989
Cost of Mangoes (Kg): 54.99999999999999
Cost of Milk Packets: 18.0
```

	Candies (#)	Mangoes (Kg)	Milk Packets (#)	Payment (Rs)	Class
0	20	6	2	386	RICH
1	16	3	6	289	RICH
2	27	6	2	393	RICH
3	19	1	2	110	POOR
4	24	4	2	280	RICH
5	22	1	5	167	POOR
6	15	4	2	271	RICH
7	18	4	2	274	RICH
8	21	1	4	148	POOR
9	16	2	4	198	POOR

```
Population Mean: 1560.6634538152612
Population Variance: 58496.49239931618
Mean (Own): 1560.6634538152612
Variance (Own): 58496.492399316136
Avg NumPy Time: 2.6917457580566408e-05
Avg Own Time : 1.1038780212402343e-05
Wednesday Mean: 1550.7060000000001
April Mean: 1698.9526315789474
Probability of Loss: 0.4979919678714859
Probability of Profit on Wednesday: 0.42
```

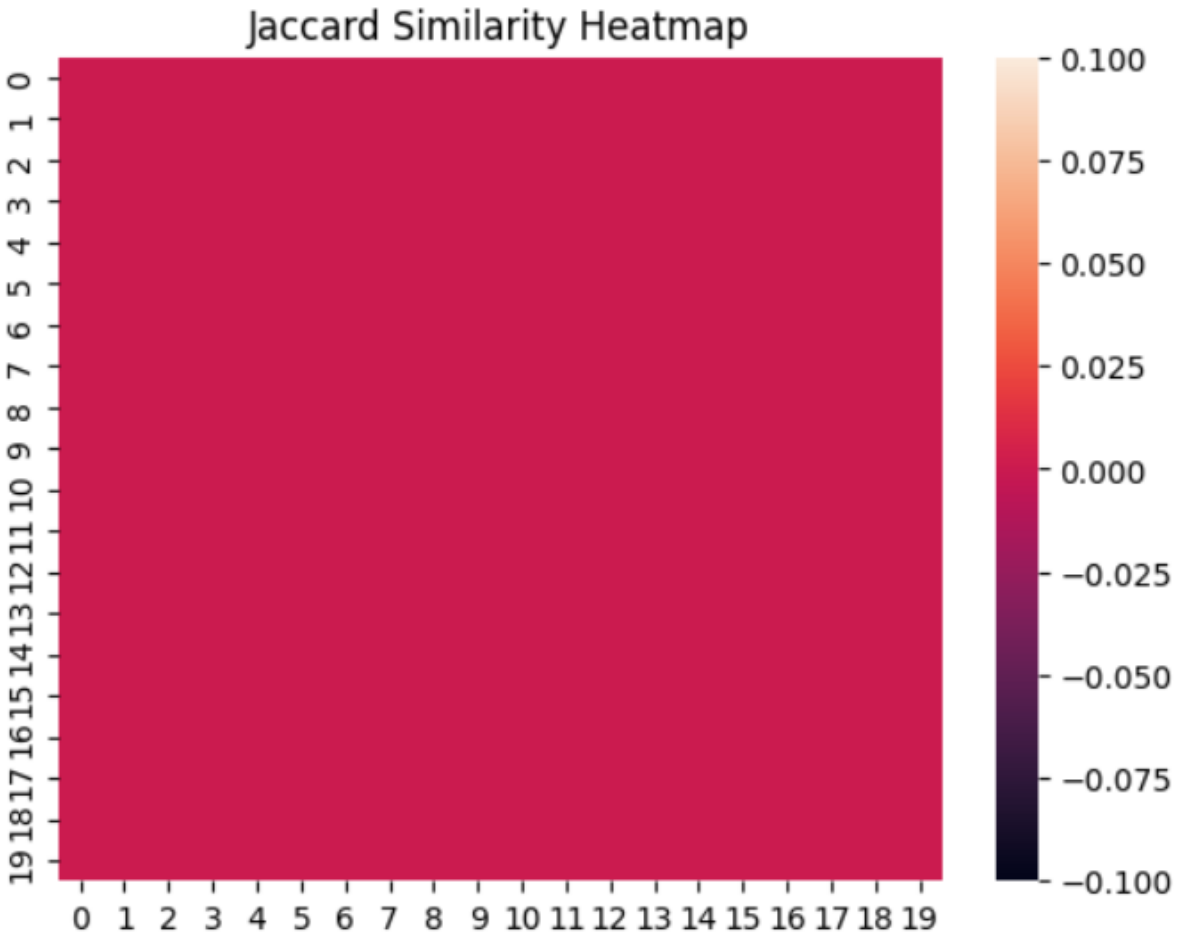

Chg% vs Day of Week
0.15
0.10
0.05
0.00
−0.05
−0.10
Chg %
Tue
Mon
Fri
Thu
Wed
Day

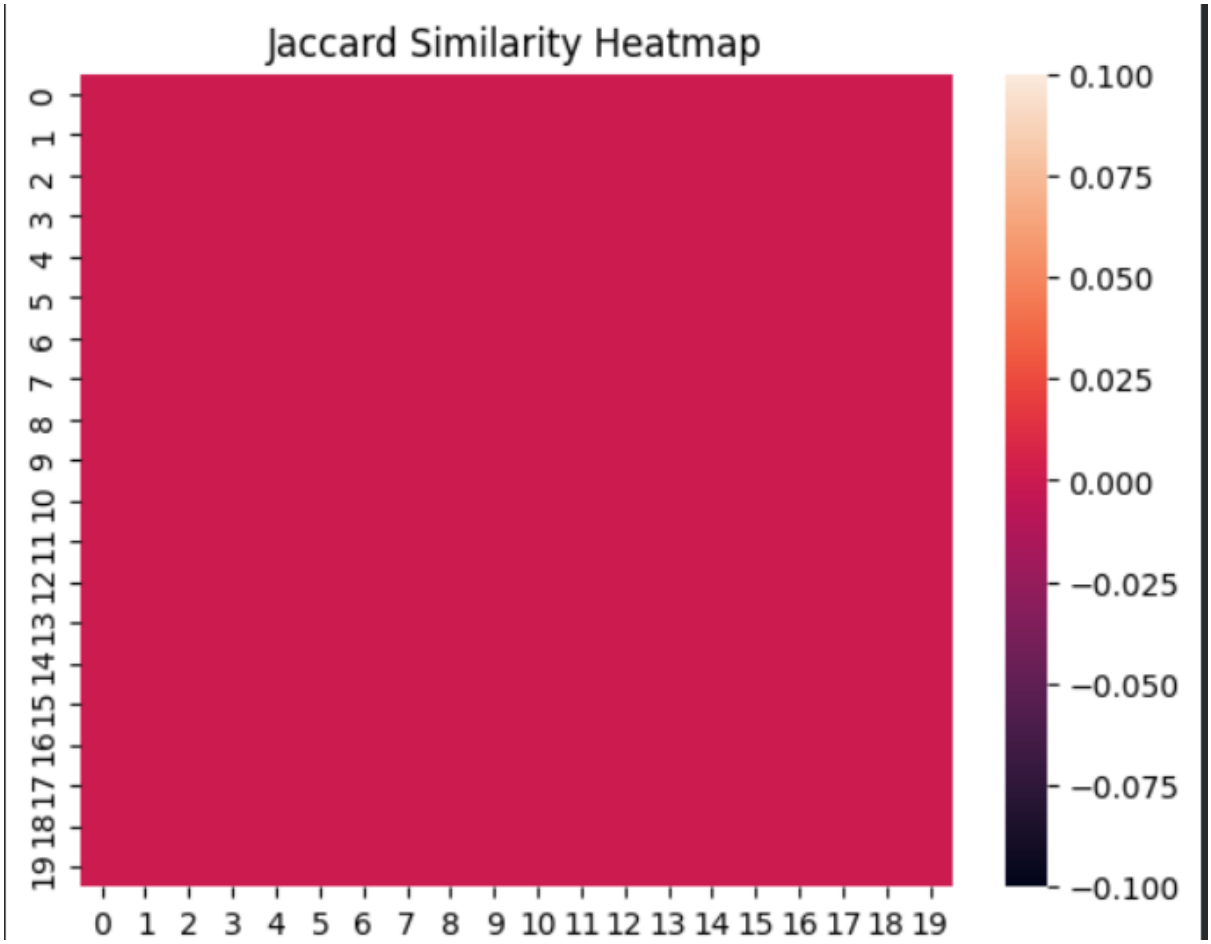
```
Attribute Datatypes:
Record ID                          int64
age                                int64
sex                               object
on thyroxine                      object
query on thyroxine                object
on antithyroid medication         object
sick                              object
pregnant                          object
thyroid surgery                   object
I131 treatment                    object
query hypothyroid                 object
query hyperthyroid                object
lithium                           object
goitre                            object
tumor                             object
hypopituitary                     object
psych                             object
TSH measured                      object
TSH                               object
T3 measured                       object
T3                                object
TT4 measured                      object
TT4                               object
T4U measured                      object
T4U                               object
FTI measured                      object
FTI                               object
TBG measured                      object
TBG                               object
referral source                   object
Condition                         object
dtype: object
```

```
Categorical Attributes: Index(['sex', 'on thyroxine', 'query on thyroxine',
       'on antithyroid medication', 'sick', 'pregnant', 'thyroid surgery',
       'I131 treatment', 'query hypothyroid', 'query hyperthyroid', 'lithium',
       'goitre', 'tumor', 'hypopituitary', 'psych', 'TSH measured', 'TSH',
       'T3 measured', 'T3', 'TT4 measured', 'TT4', 'T4U measured', 'T4U',
       'FTI measured', 'FTI', 'TBG measured', 'TBG', 'referral source',
       'Condition'],
      dtype='object')
Numerical Attributes: Index(['Record ID', 'age'], dtype='object')

Data Range (Numerical):
Record ID Min: 840801013 Max: 870119035
age Min: 1 Max: 65526
```

```
Missing Values:
Record ID                       0
age                             0
sex                             0
on thyroxine                    0
query on thyroxine              0
on antithyroid medication       0
sick                            0
pregnant                        0
thyroid surgery                 0
I131 treatment                  0
query hypothyroid               0
query hyperthyroid              0
lithium                         0
goitre                          0
tumor                           0
hypopituitary                   0
psych                           0
TSH measured                    0
TSH                             0
T3 measured                     0
T3                              0
TT4 measured                    0
TT4                             0
T4U measured                    0
T4U                             0
FTI measured                    0
FTI                             0
TBG measured                    0
TBG                             0
referral source                 0
Condition                       0
dtype: int64
```

```
Outlier Detection:
Record ID Outliers: 0
age Outliers: 4

Mean and Variance:
Record ID Mean: 852947346.6122983 Variance: 57486250586150.34
age Mean: 73.55582206716092 Variance: 1401800.8688713463
Jaccard Coefficient: 0
Simple Matching Coefficient: 0
Cosine Similarity: 0.9999999999999997
```


Jaccard Similarity Heatmap
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
−0.100


Jaccard Similarity Heatmap
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
−0.100