ExpendiTalks: Voicing the Spending Habits of NCR Households

Preliminary Period Project Deliverable for Data Mining (CS ELEC 4C)
Presented to the
Department of Computer Science
College of Information and Computing Sciences
University of Santo Tomas

In Partial Fulfillment
of the Requirements for the Degree
Bachelor of Science in Computer Science

By

Abrigo, Nathanael Chris
Cruz, Dwight Kenneth
Entrata, Joshua Kyle K.
Payumo, Edjin Jerney H.

Faculty:

Mr. Ahdrian Camilo C. Gernale

March 22, 2025

Table of Contents

I. Introduction

The National Capital Region (NCR) stands as the economic hub of the Philippines, contributing 31.5% of the country's GDP. Despite being the smallest region, it is the most densely populated, home to over 13.4 million Filipinos (DTI, 2020 National Census). As the nation's economic powerhouse, NCR households face higher living costs, influencing how they allocate income across necessities and discretionary spending. Urban financial behavior here is shaped by greater access to goods and services, convenience-based spending, and economic disparities. Understanding these household-level economic conditions is crucial for policymakers and businesses to develop targeted programs, address financial challenges, and identify market demands.

A. Background of the Dataset

The **Family Income and Expenditure Survey (FIES) 2023** is a nationwide survey conducted by the Philippine Statistics Authority (PSA) to collect comprehensive data on household income and expenditures. As the primary source of information on income distribution, consumption patterns, and economic disparities, the FIES plays a crucial role in shaping social and economic policies in the Philippines.

The 2023 FIES is the twenty-first iteration of the survey since its inception in 1957. It serves as a vital tool for government planners, policymakers, and researchers in designing targeted development programs. The dataset provides insights into:

- **Household consumption patterns** by expenditure category and income source.
- **Demographic and geographic data**, including household size, regional and provincial distribution, and urban or rural classification.

The dataset consists of 90 columns and approximately 163,268 records, offering a comprehensive view of the financial and socio-economic conditions of Filipino households. It captures both cash and in-kind income sources, various spending categories, employment details, and housing conditions, making it an essential resource for evidence-based decision-making and policy formulation.

B. Data Dictionary

Index	Features	Descriptions	Data Type	Feature Type
0	W_REGN	Region	Integer	Categorical
1	W_PROV	Province	Integer	Categorical
2	SEQ_NO	Household ID	Integer	Numerical
3	RPROV	Province Recode (Province with Highly Urbanized City (HUC) code)	Integer	Categorical
4	FSIZE	Average Family Size	Float	Numerical
5	REG_SAL	Salaries/Wages from Regular Employment	Integer	Numerical
6	SEASON_SAL	Salaries/Wages from Seasonal Employment	Integer	Numerical
7	WAGES	Salaries/Wages	Integer	Numerical
8	NETSHARE	Net Share of Crops, Fruits, etc. (Tot. Net Value of Share)	Integer	Numerical
9	CASH_ABROAD	Cash Receipts, Support, etc. from Abroad	Integer	Numerical
10	CASH_DOMESTIC	Cash Receipts, Support, etc. from Domestic Source	Integer	Numerical

11	RENTALS_REC	Rentals Received from Non-Agri Lands, etc.	Integer	Numerical
12	INTEREST	Interest	Integer	Numerical
13	PENSION	Pension and Retirement Benefits	Integer	Numerical
14	DIVIDENDS	Dividends from Investment	Integer	Numerical
15	OTHER_SOURCE	Other Sources of Income Not Elsewhere Classified (NEC)	Integer	Numerical
16	NET_RECEIPT	Family Sustenance Activities	Integer	Numerical
17	REGFT	Total Received as Gifts	Float	Numerical
18	NET_CFG	Crop Farming and Gardening	Integer	Numerical
19	NET_LPR	Livestock and Poultry Raising	Integer	Numerical
20	NET_FISH	Fishing	Integer	Numerical
21	NET_FOR	Forestry and Hunting	Integer	Numerical
22	NET_RET	Wholesale and Retail	Integer	Numerical
23	NET_MFG	Manufacturing	Integer	Numerical
24	NET_TRANS	Transportation, Storage Services	Integer	Numerical
25	NET_NEC_A8	Entrepreneurial Activities NEC	Integer	Numerical
26	NET_NEC_A9	Entrepreneurial Activities NEC	Integer	Numerical
27	NET_NEC_A10	Entrepreneurial Activities NEC	Integer	Numerical
28	EAINC	Total Income from Entrepreneurial Acitivites	Integer	Numerical
29	LOSSES	Losses from Entrepreneurial Activities	Integer	Numerical
30	BREAD	Bread and Cereals	Float	Numerical
31	MEAT	Meat	Float	Numerical
32	FISH	Fish and Seafood	Float	Numerical
33	MILK	Milk, Cheese, and Eggs	Float	Numerical
34	OIL	Oils and Fats	Float	Numerical
35	FRUIT	Fruit	Float	Numerical
36	VEG	Vegetables	Float	Numerical
37	SUGAR	Sugar, Jam, Honey, Chocolate, and Confetionery	Float	Numerical
38	FOOD_NEC	Food Products Not Elsewhere Classified	Float	Numerical
39	FRUIT_VEG	Fruit and Vegetable Juices	Float	Numerical
40	COFFEE	Coffee	Float	Numerical
41	TEA	Tea	Float	Numerical
42	COCOA	Cocoa drinks	Float	Numerical
43	WATER	Mineral Water	Float	Numerical
44	SOFTDRINKS	Softdrinks	Float	Numerical
45	OTHER_NON_ALCOHOL	Other Non Alcoholic Beverages	Float	Numerical
46	ALCOHOL	Alcoholic Beverages	Float	Numerical
47	TOBACCO	Tobacco	Float	Numerical
48	OTHER_VEG	Other Vegetable-Based Products	Float	Numerical
49	SERVICES_PRIMARY_GOODS	Services Primary Goods	Integer	Numerical
50	ALCOHOL_PROCDUCTION_SERVICES	Alcohol Procduction Services	Integer	Numerical
51	FOOD_HOME	Total Food Consumed at Home (Total)	Float	Numerical
52	FOOD_OUTSIDE	Food Regularly Consumed Outside The	Float	Numerical

		Home (Total)		
53	FOOD	Total Food Expeditures	Float	Numerical
54	CLOTH	Clothing and Footwear	Integer	Numerical
55	HOUSING_WATER	Housing, Water, Electricity, Gas and Other Fuels	Integer	Numerical
56	ACTRENT	Actual House Rent	Integer	Numerical
57	IMPUTED_RENT	Imputed House Rental Value	Integer	Numerical
58	BIMPUTED_RENT	Imputed Housing Benefit Rental Value	Integer	Numerical
59	RENTVAL	House Rent/Rental Value	Integer	Numerical
60	FURNISHING	Furnishings and Routine Household Maintenance	Integer	Numerical
61	HEALTH	Health	Integer	Numerical
62	TRANSPORT	Transport	Integer	Numerical
63	COMMUNICATION	Communication	Integer	Numerical
64	RECREATION	Recreation and Culture	Integer	Numerical
65	EDUCATION	Education	Integer	Numerical
66	INSURANCE	Insurance	Integer	Numerical
67	MISCELLANEOUS	Miscellaneous Goods and Services	Integer	Numerical
68	DURABLE	Durable Furniture and Equipment	Integer	Numerical
69	OCCASION	Special Family Occasion	Integer	Numerical
70	OTHER_EXPENDITURE	Other Expenditure (including Value Consumed and Losses)	Integer	Numerical
71	OTHER_DISBURSEMENT	Other Disbursement	Integer	Numerical
72	FOOD_ACCOM_SRVC	Food Regularly Consumed Outside The Home - Accomodation Services	Integer	Numerical
73	NFOOD	Total Non-Food Expenditure	Float	Numerical
74	TOINC	Total Income	Float	Numerical
75	TOTEX	Total Expenditure	Float	Numerical
76	TOTDIS	Total Disbursements	Float	Numerical
77	OTHREC	Total Other Receipts	Integer	Numerical
78	TOREC	Total Receipts	Float	Numerical
79	RPSU	PSU (Recode)	Integer	Numerical
80	RFACT	Family / Household Weight	Float	Numerical
81	MEM_RFACT	Population Weight	Float	Numerical
82	URB	Urban / Rural	Integer	Categorical
83	PERCAPITA	Per Capita Income	Float	Numerical
84	NPCINC	Per Capita Income Decile (National)	Integer	Numerical
85	RPCINC	Per Capita Income Decile (Region)	Integer	Numerical
86	PRPCINC	Per Capita Income Decile (Province)	Integer	Numerical
87	PPCINC	Per Capita Income Decile (Province and HUC)	Integer	Numerical
88	RPCINC_NIR	Per Capita Income Decile (Region with Negros Island Region (NIR))	Integer	Numerical
89	W_REGN_NIR	Region (with NIR)	Integer	Categorical

C. Statement of the Problem

The Family Income and Expenditure Survey (FIES) 2023 provides extensive data on household income and expenditures. However, it lacks a structured analysis of spending patterns across different income groups, making it difficult to derive actionable insights. Without a clear segmentation of households based on their financial behavior, policymakers struggle to design targeted assistance programs, and businesses lack the necessary insights to optimize their product offerings.

This study seeks to address the following key questions:

1. What are the distinct spending patterns among NCR households based on income levels?
2. How do different socioeconomic groups allocate their expenditures between essential needs (e.g., food, housing, healthcare, education) and discretionary expenses (e.g., entertainment, luxury goods)?
3. Can households be grouped into meaningful financial segments based on their income-expenditure behavior?

D. Objectives

This study aims to analyze household spending patterns in NCR using clustering techniques and association rule mining. Specifically, it seeks to:

1. **Identify Distinct Spending Patterns** – Examine how NCR households allocate their income across various expenditure categories based on income levels.
2. **Analyze Socioeconomic Expenditure Allocation** – Compare how different income groups prioritize essential needs (e.g., food, housing, healthcare) versus discretionary expenses (e.g., entertainment, luxury goods).
3. **Segment Households into Financial Groups** – Classify households into meaningful financial segments based on their income-expenditure behavior to uncover trends and disparities.

By applying data mining techniques, this study will uncover hidden patterns in household spending, providing valuable insights for:

1. **Government agencies** – To design more effective financial aid and subsidy programs.
2. **Businesses & marketers** – To align product offerings with consumer demand.
3. **Urban planners & policymakers** – To identify financial stress points and develop inclusive economic strategies.

By transforming raw survey data into actionable insights, this research will help policymakers, businesses, and urban planners better understand consumer behavior, design targeted programs, and address financial challenges in NCR, ultimately improving economic planning and consumer welfare in the region.

II. Methodology

A. Data Mining Pipeline

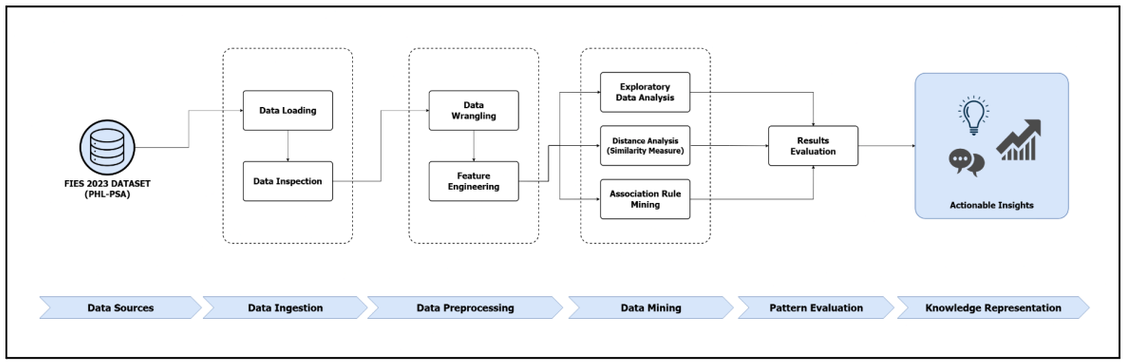
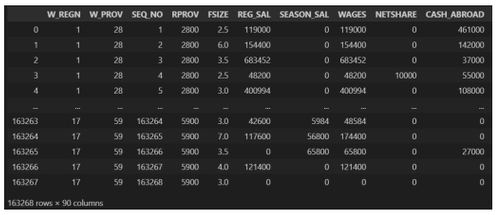


Figure 2.1: Data Mining Pipeline

Figure 2.1 showcases the Data Mining Pipeline, which follows a structured workflow designed to extract meaningful insights from the FIES 2023 dataset, specifically focusing on NCR households. The process consists of six key stages: Data Sources, Data Ingestion, Data Preprocessing, Data Mining, Pattern Evaluation, and Knowledge Representation.

B. Machine Problem Task Table

TASKS (with explanation)	PYTHON CODE	OUTPUT (Screenshot/Result)
1. Data Import & Exploration		
[PAYUMO] Load the large dataset	fies_23 = pd.read_csv('../data/raw/FIES PUF 2023 Volume2 Household Summary.csv')	loaded the PSA dataset with a total of 90 features and 163268 instances
2. Identifying Data and Attributes (10 pts)		
[PAYUMO] List all column names and data types	fies_23.info()	<class 'pandas.core.frame.DataFrame'> RangeIndex: 163268 entries, 0 to 163267 Data columns (total 90 columns): # Column Non-Null Count Dtype --- ------ -------------- ----- 0 W_REGN 163268 non-null int64 1 W_PROV 163268 non-null int64 2 SEQ_NO 163268 non-null int64 3 RPROV 163268 non-null int64 4 FSIZE 163268 non-null float64 5 REG_SAL 163268 non-null int64 6 SEASON_SAL 163268 non-null int64 7 WAGES 163268 non-null int64 8 NETSHARE 163268 non-null int64 9 CASH_ABROAD 163268 non-null int64 10 CASH_DOMESTIC 163268 non-null int64 11 RENTALS_REC 163268 non-null int64 12 INTEREST 163268 non-null int64 13 PENSION 163268 non-null int64 14 DIVIDENDS 163268 non-null int64 15 OTHER_SOURCE 163268 non-null int64 16 NET_RECEIPT 163268 non-null int64 17 REGFT 163268 non-null float64 18 NET_CFG 163268 non-null int64 19 NET_LPR 163268 non-null int64 ... 88 RPCINC_NIR 163268 non-null int64 89 W_REGN_NIR 163268 non-null int64 dtypes: float64(32), int64(58) 
3. Determining the Type of Dataset (10 pts)		
[PAYUMO] Check if columns are numerical, categorical, or mixed.	def classify_columns(df): numerical_cols = df.select_dtypes(include=['int64', 'float64']).columns.tolist() categorical_cols = df.select_dtypes(include=['object',	Numerical Columns: ['W_REGN', 'W_PROV', 'SEQ_NO', 'RPROV', 'FSIZE', 'REG_SAL', 'SEASON_SAL', 'WAGES', 'NETSHARE', 'CASH_ABROAD', 'CASH_DOMESTIC',

<table>
<tr>
<td>This will help classify data attributes, ensuring appropriate preprocessing for analysis.</td>
<td>'category']).columns.tolist()

    mixed_cols = []

    # Check if any categorical columns contain numerical values or vice versa
    for col in df.columns:
        unique_types = df[col].apply(type).nunique()
        if unique_types > 1:
            mixed_cols.append(col)

    # Remove mixed columns from numerical and categorical lists
    numerical_cols = [col for col in numerical_cols if col not in mixed_cols]
    categorical_cols = [col for col in categorical_cols if col not in mixed_cols]

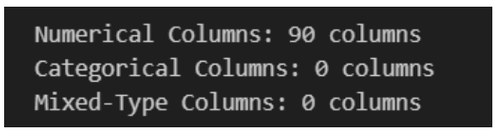
    return {
        "Numerical Columns": numerical_cols,
        "Categorical Columns": categorical_cols,
        "Mixed-Type Columns": mixed_cols
    }

column_types = classify_columns(fies_23)

# Print results
for col_type, cols in column_types.items():
    print(f"{col_type}: {cols}")

# Get column classifications
column_types = classify_columns(fies_23)

# Print counts for each category
for col_type, cols in column_types.items():
    print(f"{col_type}: {len(cols)} columns")</td>
<td>'RENTALS_REC', 'INTEREST', 'PENSION', 'DIVIDENDS', 'OTHER_SOURCE', 'NET_RECEIPT', 'REGFT', 'NET_CFG', 'NET_LPR', 'NET_FISH', 'NET_FOR', 'NET_RET', 'NET_MFG', 'NET_TRANS', 'NET_NEC_A8', 'NET_NEC_A9', 'NET_NEC_A10', 'EAINC', 'LOSSES', 'BREAD', 'MEAT', 'FISH', 'MILK', 'OIL', 'FRUIT', 'VEG', 'SUGAR', 'FOOD_NEC', 'FRUIT_VEG', 'COFFEE', 'TEA', 'COCOA', 'WATER', 'SOFTDRINKS', 'OTHER_NON_ALCOHOL', 'ALCOHOL', 'TOBACCO', 'OTHER_VEG', 'SERVICES_PRIMARY_GOODS', 'ALCOHOL_PROCDUCTION_SERVICES', 'FOOD_HOME', 'FOOD_OUTSIDE', 'FOOD', 'CLOTH', 'HOUSING_WATER', 'ACTRENT', 'IMPUTED_RENT', 'BIMPUTED_RENT', 'RENTVAL', 'FURNISHING', 'HEALTH', 'TRANSPORT', 'COMMUNICATION', 'RECREATION', 'EDUCATION', 'INSURANCE', 'MISCELLANEOUS', 'DURABLE', 'OCCASION', 'OTHER_EXPENDITURE', 'OTHER_DISBURSEMENT', 'FOOD_ACCOM_SRVC', 'NFOOD', 'TOINC', 'TOTEX', 'TOTDIS', 'OTHREC', 'TOREC', 'RPSU', 'RFACT', 'MEM_RFACT', 'URB', 'PERCAPITA', 'NPCINC', 'RPCINC', 'PRPCINC', 'PPCINC', 'RPCINC_NIR', 'W_REGN_NIR']
Categorical Columns: []
Mixed-Type Columns: []

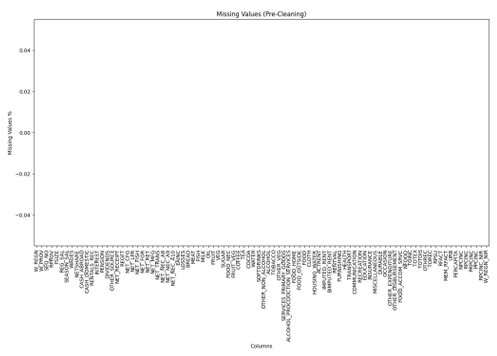
Upon inspecting the dataset, all columns are stored as numerical data types. However, some columns represent categorical variables despite being encoded as numbers. These include:

- W_REGN (Region Code): Encodes the administrative region of the household.
- W_PROV (Province Code): Encodes the province of the household.
- RPROV (Province Recode): A modified province code that includes Highly Urbanized Cities (HUCs).
- W_REGN_NIR (Region with NIR): A variation of the region code that includes the now-defunct Negros Island Region (NIR).
- URB (Urban/Rural Indicator): A binary variable indicating whether the household is in an urban (1) or rural (2) area.</td>
</tr>
<tr>
<td>4. Data Quality Assessment (20 pts)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[PAYUMO]
Check for missing values, duplicates, and wrong data.

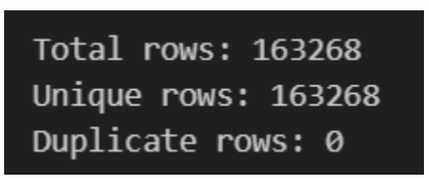
This will help identify data quality issues that may affect the accuracy of the analysis.</td>
<td># Calculating the Missing Values % contribution in DF
df_null = round(100*(fies_23.isnull().sum())/len(fies_23), 2)

# Plotting the df_null
plt.figure(figsize=(16,8))
sns.barplot(x=df_null.index, y=df_null.values, alpha=0.8)
plt.title('Missing Values (Pre-Cleaning)')
plt.ylabel('Missing Values %')
plt.xlabel('Columns')
plt.xticks(rotation=90)
plt.show()

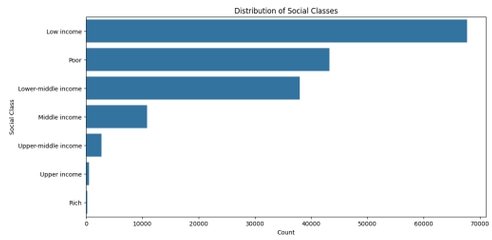
# Identifying duplicate rows using `duplicated()` method
duplicate_rows = fies_23[fies_23.duplicated()]
print(f"Number of duplicate rows: {duplicate_rows.shape[0]}")

# Identifying duplicate rows by comparing total rows with unique rows
total_rows = fies_23.shape[0]</td>
<td> 

 
</td>
</tr>
</table>

	unique_rows = fies_23.drop_duplicates().shape[0] duplicate_rows = total_rows - unique_rows print(f"Total rows: {total_rows}") print(f"Unique rows: {unique_rows}") print(f"Duplicate rows: {duplicate_rows}")	Based on the inspection results, no missing and duplicate values were identified

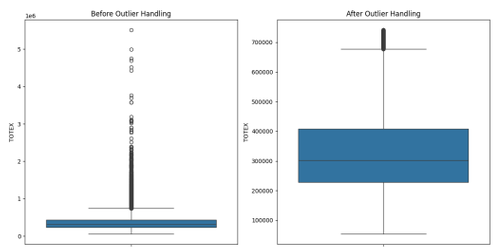
[PAYUMO]
Feature Engineering: SOCIAL_CLASS
 	
This will categorize households into income groups to analyze spending patterns across different economic levels.
	poverty_threshold = 13873 * 12 # Official poverty threshold (PSA, 2023)
	
	def classify_income(row):
	  total_income = row['TOINC']
	  if total_income < poverty_threshold:
	  return 'Poor'
	  elif total_income < 2 * poverty_threshold:
	  return 'Low income'
	  elif total_income < 4 * poverty_threshold:
	  return 'Lower-middle income'
	  elif total_income < 7 * poverty_threshold:
	  return 'Middle income'
	  elif total_income < 12 * poverty_threshold:
	  return 'Upper-middle income'
	  elif total_income < 20 * poverty_threshold:
	  return 'Upper income'
	  else:
	  return 'Rich'
	
	fies_feature['SOCIAL_CLASS'] = fies_feature.apply(classify_income, axis=1)
	
	fies_feature["SOCIAL_CLASS"].value_counts()
	
	### Visualize
	plt.figure(figsize=(12, 6))
	sns.countplot(
	  y="SOCIAL_CLASS",
	  data=fies_feature,
	
	order=fies_feature["SOCIAL_CLASS"].value_counts().index,
	)
	plt.title("Distribution of Social Classes")
	plt.xlabel("Count")
	plt.ylabel("Social Class")
	plt.show()
		SOCIAL_CLASS
		Low income 67700
		Poor 43265
		Lower-middle income 37991
		Middle income 10858
		Upper-middle income 2715
		Upper income 548
		Rich 191
		Name: count, dtype: int64
		
		 
		
		- To allow meaningful segmentation and comparative analysis of households, a new feature `SOCIAL_CLASS` was created by categorizing households based on their `TOINC` (Total Income). This allows for easier interpretation of income groups and supports targeted socioeconomic analysis.

[ENTRATA]
Data Wrangling: Filter NCR data
 	
This will extract and retain only the relevant data for the National Capital Region (NCR) analysis.
	fies_ncr = fies_feature[cleaned_fies['W_REGN'] == 13]
	
	# print how many rows remained after filtering
	print("Number of rows in NCR: ", fies_ncr.shape[0])
		Number of rows in NCR: 20690
		
		> The dataset was filtered to include only households from the NCR (National Capital Region) to ensure a more focused analysis within a highly urbanized and economically distinct area.
		
		This allows the study to capture spending and income patterns specific to NCR's urban households, where cost of living, income distribution, and consumption behaviors differ significantly from rural or other regional contexts.

[ENTRATA]
Handling Outliers for Total Income and Total Expenses

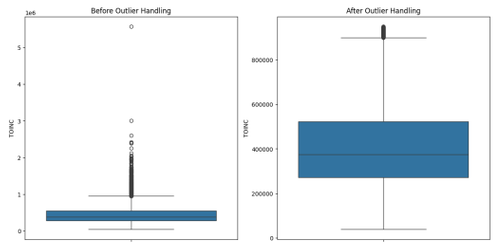
This will help minimize the impact of extreme values and improve the reliability of the analysis.

```
### Outliers in TOTEX (Total Expenditure)
fe_totex = fies_ncr['TOTEX']
Q1 = fe_totex.quantile(0.25)
Q3 =fe_totex.quantile(0.75)
IQR = Q3 - Q1
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
print(f'IQR (fe_totex): {IQR}')
print(f'Lower Bound (fe_totex): {lower_bound}')
print(f'Upper Bound (fe_totex) {upper_bound}')
outliers = (fies_ncr["TOTEX"] < lower_bound) \
    | (fies_ncr["TOTEX"] > upper_bound)

fe_totex_outlier_handled = fies_ncr[~outliers]

print(f'Number of rows before outlier handling for
TOTEX: {len(fies_ncr)}')
print(f'Number of rows after outlier handling for TOTEX:
{len(fe_totex_outlier_handled)}')
fig, axes = plt.subplots(1, 2, figsize=(12, 6))

sns.boxplot(data=fies_ncr.TOTEX, ax=axes[0])
axes[0].set_title("Before Outlier Handling")

sns.boxplot(data=fe_totex_outlier_handled.TOTEX,
ax=axes[1])
axes[1].set_title("After Outlier Handling")

plt.tight_layout()
plt.show()
```

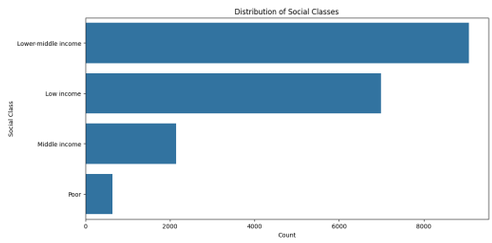
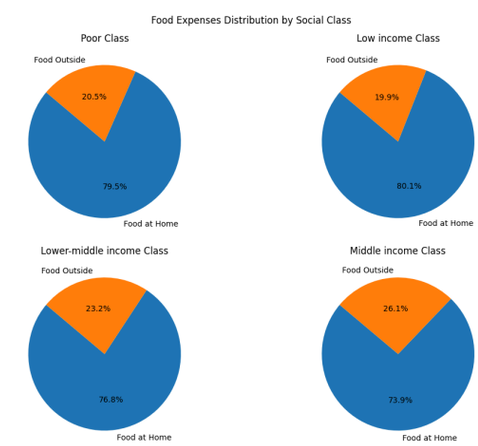


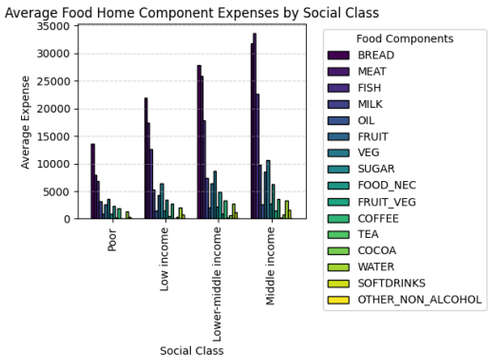

```
IQR (fe_totex): 203853.375
Lower Bound (fe_totex): -73435.3125
Upper Bound (fe_totex) 741978.1875
Number of rows before outlier handling for TOTEX: 20690
Number of rows after outlier handling for TOTEX: 19492
```

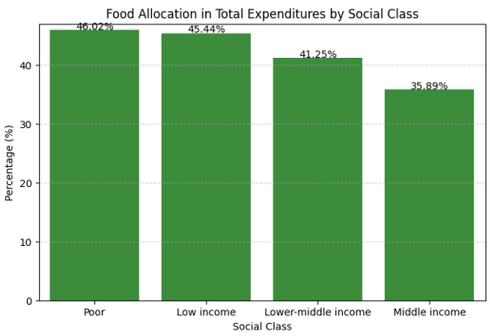
```
### Outliers in TOINC (Total Income)
fe_toinc = fe_totex_outlier_handled['TOINC']
Q1 = fe_toinc.quantile(0.25)
Q3 =fe_toinc.quantile(0.75)
IQR = Q3 - Q1
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
print(f'IQR (fe_toinc): {IQR}')
print(f'Lower Bound (fe_toinc): {lower_bound}')
print(f'Upper Bound (fe_toinc) {upper_bound}')
outliers = (fe_totex_outlier_handled["TOINC"] <
lower_bound) \
    | (fe_totex_outlier_handled["TOINC"] > upper_bound)

fe_toinc_outlier_handled =
fe_totex_outlier_handled[~outliers]

print(f'Number of rows before outlier handling for
TOINC: {len(fe_totex_outlier_handled)}')
print(f'Number of rows after outlier handling for TOINC:
{len(fe_toinc_outlier_handled)}')
fig, axes = plt.subplots(1, 2, figsize=(12, 6))

sns.boxplot(data=fe_totex_outlier_handled.TOINC,
ax=axes[0])
axes[0].set_title("Before Outlier Handling")

sns.boxplot(data=fe_toinc_outlier_handled.TOINC,
ax=axes[1])
axes[1].set_title("After Outlier Handling")

plt.tight_layout()
plt.show()
```

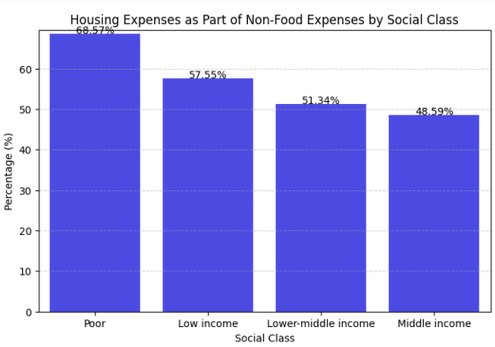


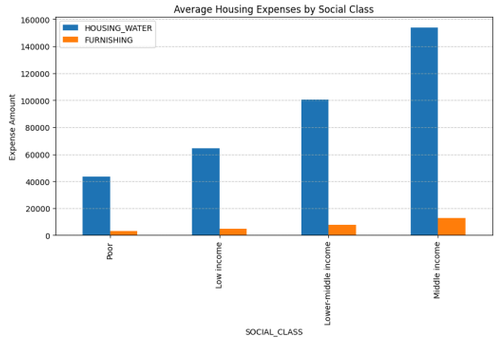
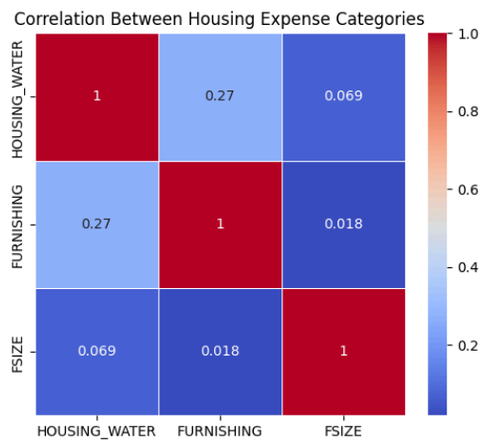

```
IQR (fe_toinc): 269893.5
Lower Bound (fe_toinc): -131035.0
Upper Bound (fe_toinc) 948539.0
Number of rows before outlier handling for TOINC: 19492
Number of rows after outlier handling for TOINC: 18848
```

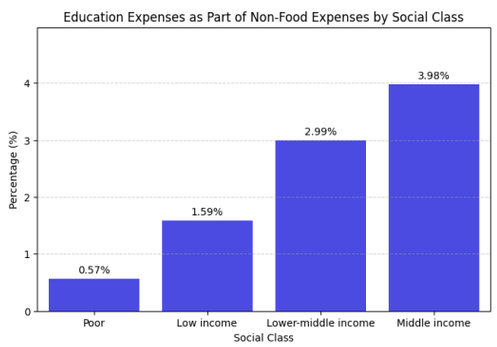
[PAYUMO] Review Social Class Distribution for NCR	final_fies["SOCIAL_CLASS"].value_counts() plt.figure(figsize=(12, 6)) sns.countplot(y="SOCIAL_CLASS", data=final_fies, order=final_fies["SOCIAL_CLASS"].value_counts().index,) plt.title("Distribution of Social Classes") plt.xlabel("Count") plt.ylabel("Social Class") plt.show()	SOCIAL_CLASS Lower-middle income 9071 Low income 6992 Middle income 2144 Poor 641 Name: count, dtype: int64  After filtering the dataset to include only NCR households and removing outliers based on Total Income (TOINC) and Total Expenses (TOTEX) using the IQR method, the resulting social class distribution remains imbalanced. Notably, the 'Upper-middle' and 'Rich' social classes were excluded due to being classified as outliers. This suggests that households in these income brackets exhibit significantly higher income and expenditure patterns compared to the majority, making them statistical anomalies rather than representative of typical spending behavior in NCR.
5. Quantitative Statistics (20 pts)		
[ABRIGO] Food expenditure analysis and visualization This will help identify spending patterns on food across different income groups and provide insights through visual representations.	final_fies['SOCIAL_CLASS'] = pd.Categorical(final_fies['SOCIAL_CLASS'], categories=['Poor', 'Low income', 'Lower-middle income', 'Middle income'], ordered=True) food_data = final_fies.groupby('SOCIAL_CLASS', observed=False)[['FOOD_HOME', 'FOOD_OUTSIDE']].sum() plt.figure(figsize=(12, 8)) for i, social_class in enumerate(food_data.index): plt.subplot(2, 2, i + 1) # Adjust grid size based on total classes plt.pie(food_data.loc[social_class], labels=['Food at Home', 'Food Outside'], autopct='%.1f%%', colors=['#1f77b4', '#ff7f0e'], startangle=140) plt.title(f'{social_class} Class') plt.suptitle('Food Expenses Distribution by Social Class') plt.tight_layout() plt.show()	 This visualization shows the proportion of food expenses by social class. Both middle income and low-middle income social group have higher percentage for food outside which can mean these groups may prefer buying processed food or buy ready to eat meals outside.

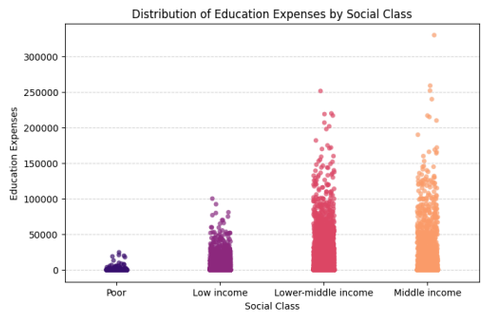
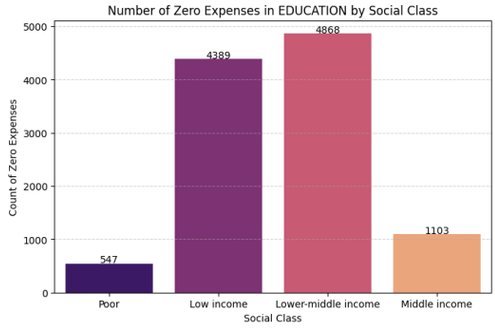
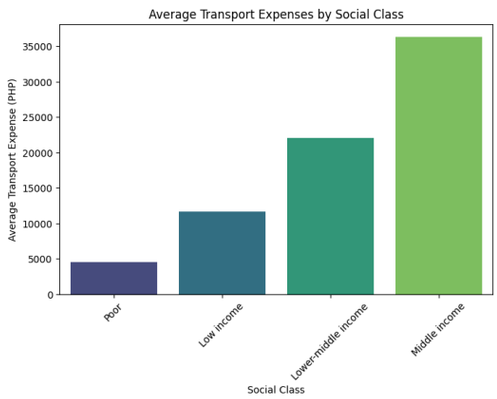
[ABRIGO]
Food expenditure analysis and visualization
 	
This will provide an overview of the average food and home component expenses by social class.
	components = [
	'BREAD', 'MEAT', 'FISH', 'MILK', 'OIL', 'FRUIT', 'VEG', 'SUGAR', 'FOOD_NEC',
	'FRUIT_VEG', 'COFFEE', 'TEA', 'COCOA', 'WATER', 'SOFTDRINKS', 'OTHER_NON_ALCOHOL'
	]
	
	# Group data by SOCIAL_CLASS and sum the component values
	component_data = final_fies.groupby('SOCIAL_CLASS', observed=False)[components].mean()
	
	# Plotting
	plt.figure(figsize=(12, 6))
	component_data.plot(kind='bar', width=0.8, colormap='viridis', edgecolor='black')
	
	plt.title('Average Food Home Component Expenses by Social Class')
	plt.xlabel('Social Class')
	plt.ylabel('Average Expense')
	plt.legend(title='Food Components', bbox_to_anchor=(1.05, 1), loc='upper left')
	plt.grid(axis='y', linestyle='--', alpha=0.5)
	plt.tight_layout()
	plt.show()
		
		This visualization shows the averages for all food home components expenses by social class. This shows how important food products such as bread, meat and fish are for all social classes. It is also noticeable that drinks like tea and cocoa only have enjoyers for low, low middle, and middle income social groups.

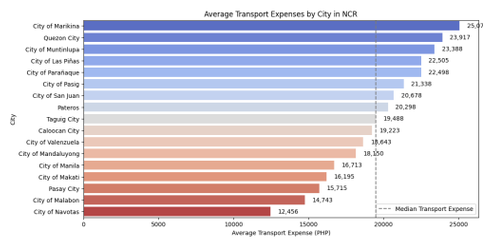
[ENTRATA]
Food expenditure analysis and visualization
 	
This will illustrate the proportion of total expenditure allocated to food across different social classes.
	# Group by social class and mean percentage of housing expenses
	housing_percentages = final_fies.groupby('SOCIAL_CLASS', observed=True).apply(
	lambda x: (x['FOOD'] / x['TOTEX']).mean() * 100,
	include_groups=False
	)
	
	plt.figure(figsize=(8, 5))
	sns.barplot(x=housing_percentages.index, y=housing_percentages.values, color='green', alpha=0.8)
	
	# percentage labels
	for i, v in enumerate(housing_percentages.values):
	plt.text(i, v + 0.1, f'{v:.2f}%', ha='center')
	
	plt.title('Food Allocation in Total Expenditures by Social Class')
	plt.xlabel('Social Class')
	plt.ylabel('Percentage (%)')
	plt.grid(axis='y', linestyle='--', alpha=0.5)
	plt.ylim(0, max(housing_percentages.values) + 1)
	plt.show()
		
		The visualization shows the food allocation in total expenditure by social class. It shows that the higher income groups allocate a lesser percentage of food allocation since they may have more financial flexibility.

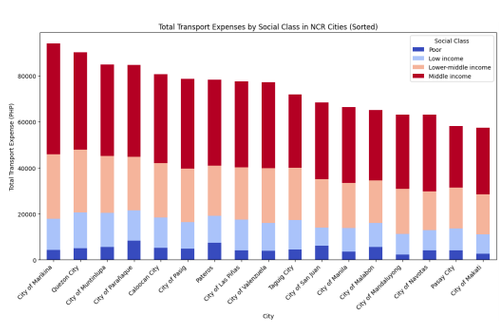
[CRUZ]
Housing expenditure analysis and visualization
 	
This will reveal spending patterns on housing across income groups and provide insights through visual representations.
	# Group by social class and mean percentage of housing expenses
	housing_percentages = final_fies.groupby('SOCIAL_CLASS', observed=True).apply(
	lambda x: ((x['HOUSING_WATER'] + x['FURNISHING']) / x['NFOOD']).mean() * 100,
	include_groups=False
	)
	
	plt.figure(figsize=(8, 5))
	sns.barplot(x=housing_percentages.index, y=housing_percentages.values, color='blue', alpha=0.8)
	
	for i, v in enumerate(housing_percentages.values):
	plt.text(i, v + 0.1, f'{v:.2f}%', ha='center')
	
	plt.title('Housing Expenses as Part of Non-Food Expenses by Social Class')
		
		This visualization shows that poor households spend the highest share on housing (68.57%) due to limited income, while higher-income groups allocate less as they can afford other expenses. As income rises, spending shifts toward education, transportation, and discretionary items, reducing housing's proportion of total expenses.

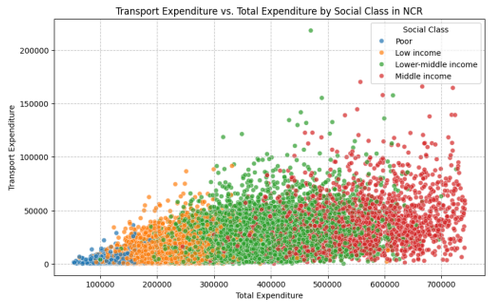
	plt.xlabel('Social Class') plt.ylabel('Percentage (%)') plt.grid(axis='y', linestyle='--', alpha=0.5) plt.ylim(0, max(housing_percentages.values) + 1) plt.show()	
[CRUZ] Housing expenditure analysis and visualization This will help in identifying the average of housing and furnishing for each income groups.	final_fies.groupby('SOCIAL_CLASS')[['HOUSING_WATER', 'FURNISHING']].mean().plot(kind='bar', figsize=(10, 5)) plt.title("Average Housing Expenses by Social Class") plt.ylabel("Expense Amount") plt.grid(axis='y', linestyle='--', alpha=0.7) plt.show()	 This visualization shows that housing and water expenses increase with income, as higher-income groups tend to afford better housing and utilities. Furnishing expenses also rise, but at a slower rate, reflecting increased discretionary spending among wealthier households.
[CRUZ] Housing expenditure analysis and visualization This will provide insight into the relationships between house expenses, furnishing, and family size.	plt.figure(figsize=(6, 5)) sns.heatmap(final_fies[['HOUSING_WATER', 'FURNISHING', 'FSIZE']].corr(), annot=True, cmap="coolwarm", linewidths=0.5) plt.title("Correlation Between Housing Expense Categories") plt.show()	 This heatmap shows the correlation between housing expenses and family size (FSIZE). Housing and water expenses have a weak positive correlation (0.27) with furnishing expenses, while both categories have very weak correlations with family size (0.069 and 0.018, respectively). This suggests that family size has little influence on housing and furnishing costs.

[ABRIGO]
Education expenditure analysis and visualization
 	
This will highlight spending patterns on education across income groups and provide insights through visual representations.
	percentages = final_fies.groupby('SOCIAL_CLASS', observed=True).apply(
	  lambda x: ((x['EDUCATION']) / x['NFOOD']).mean() * 100, include_groups=False
	)
	
	plt.figure(figsize=(8, 5))
	sns.barplot(x=percentages.index, y=percentages.values, color='blue', alpha=0.8)
	
	for i, v in enumerate(percentages.values):
	  plt.text(i, v + 0.1, f'{v:.2f}%', ha='center')
	
	plt.title('Education Expenses as Part of Non-Food Expenses by Social Class')
	plt.xlabel('Social Class')
	plt.ylabel('Percentage (%)')
	plt.grid(axis='y', linestyle='--', alpha=0.5)
	plt.ylim(0, max(percentages.values) + 1)
	plt.show()
		
		This visualization shows the proportion of education expenses in non food expenditures by social class. The 'Poor' social group have a very low proportion for education which can mean this group have a more priority expenses rather than education.

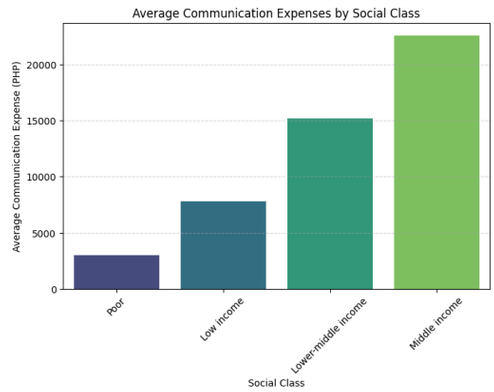
[ABRIGO] Education expenditure analysis and visualization This visualization will highlight the allocation of education expenses across various social classes.	plt.figure(figsize=(8, 5)) sns.stripplot(x='SOCIAL_CLASS', y='EDUCATION', data=final_fies, palette='magma', hue='SOCIAL_CLASS', jitter=True, alpha=0.7, size=5) plt.title('Distribution of Education Expenses by Social Class') plt.xlabel('Social Class') plt.ylabel('Education Expenses') plt.grid(axis='y', linestyle='--', alpha=0.5) plt.show()	 This visualization shows the distribution of education expenses by social class. It is noticeable here that both lower-middle income and middle income social groups have more money to spare for education. These groups also likely can afford more expensive education institutes like private schools.
[ABRIGO] Education expenditure analysis and visualization This will depict the prevalence of zero education expenses across social classes.	# zero expenses zero_education_counts = (final_fies.groupby('SOCIAL_CLASS', observed=False)['EDUCATION'] .apply(lambda x: (x == 0).sum()) .reset_index(name='Zero_Education_Count')) plt.figure(figsize=(8, 5)) sns.barplot(x='SOCIAL_CLASS', y='Zero_Education_Count', data=zero_education_counts, hue='SOCIAL_CLASS', palette='magma', legend=False) for i, v in enumerate(zero_education_counts['Zero_Education_Count']): plt.text(i, v + 10, str(v), ha='center') plt.title('Number of Zero Expenses in EDUCATION by Social Class') plt.xlabel('Social Class') plt.ylabel('Count of Zero Expenses') plt.grid(axis='y', linestyle='--', alpha=0.5) plt.show()	 This visualization shows the number of zero expenses in education expenses by social class. It shows a high count to both low income and lower-middle income followed by poor and middle income. This can mean that this count may have availed scholarships or decided not to spend anymore for education.
[ENTRATA] Transportation expenditure analysis and visualization This will identify spending patterns on transportation across income groups and provide insights through visual representations.	avg_transportation_expense = final_fies.groupby('SOCIAL_CLASS')['TRANSPORT'].mean().reset_index() order = ['Poor', 'Low income', 'Lower-middle income', 'Middle income'] plt.figure(figsize=(8,5)) sns.barplot(x='SOCIAL_CLASS', y='TRANSPORT', data=avg_transportation_expense, palette='viridis', hue='SOCIAL_CLASS', order=order) plt.title('Average Transport Expenses by Social Class') plt.xlabel('Social Class') plt.ylabel('Average Transport Expense (PHP)') plt.xticks(rotation=45) plt.show()	 - As social class increases (in terms of income), transport expenses also increase. - Lower-income groups most likely rely on cheaper public transportation, while higher-income individuals have more flexibility with private vehicles or premium transport options.

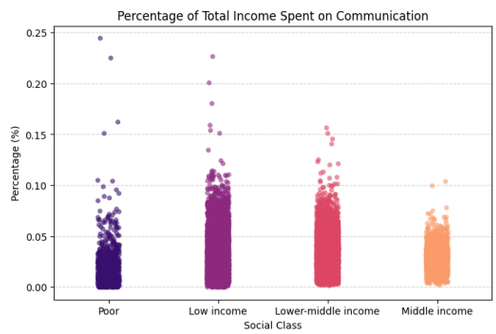
[ENTRATA]
Transportation expenditure analysis and visualization
 	
This will identify the average transport expenses per city in NCR.
	ncr_city_mapping = {
	  3900: "City of Manila",
	  7401: "City of Mandaluyong",
	  7402: "City of Marikina",
	  7403: "City of Pasig",
	  7404: "Quezon City",
	  7405: "City of San Juan",
	  7501: "Caloocan City",
	  7502: "City of Malabon",
	  7503: "City of Navotas",
	  7504: "City of Valenzuela",
	  7601: "City of Las Piñas",
	  7602: "City of Makati",
	  7603: "City of Muntinlupa",
	  7604: "City of Parañaque",
	  7605: "Pasay City",
	  7600: "Pateros",
	  7607: "Taguig City"
	}
	
	final_fies_ncr =
	final_fies[final_fies['RPROV'].isin(ncr_city_mapping.keys())].copy()
	final_fies_ncr['City'] =
	final_fies_ncr['RPROV'].map(ncr_city_mapping)
	
	avg_transport_ncr =
	final_fies_ncr.groupby('City')['TRANSPORT'].mean().reset_index()
	avg_transport_ncr =
	avg_transport_ncr.sort_values(by='TRANSPORT', ascending=False)
	
	plt.figure(figsize=(12, 6))
	ax = sns.barplot(x='TRANSPORT', y='City', data=avg_transport_ncr, palette='coolwarm')
	
	for i, v in enumerate(avg_transport_ncr['TRANSPORT']):
	  ax.text(v + 500, i, f"{int(v):,}", va='center', fontsize=10)
	
	# median reference line
	plt.axvline(x=avg_transport_ncr['TRANSPORT'].median(), color='gray', linestyle='--', label="Median Transport Expense")
	
	plt.title("Average Transport Expenses by City in NCR")
	plt.xlabel("Average Transport Expense (PHP)")
	plt.ylabel("City")
	plt.legend()
	plt.show()
		
		- Cities with higher transport expenses (Marikina, Quezon City, and Las Pinas) have the highest transport spending, possibly due to longer commutes or higher private vehicle usage
		
		- Cities like Navotas, Malabon, and Pasay have lower transport costs, maybe due to good public transport, shorter travel distances, or the distribution of lower income classes are dominant here.

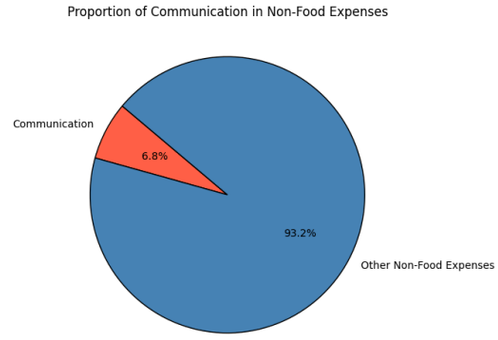
[ENTRATA]
Transportation expenditure analysis and visualization
 	
This will depict the distribution of each social class in the average transportation spending of people in each city in NCR.
	total_transport_social_ncr =
	final_fies_ncr.groupby(['City', 'SOCIAL_CLASS']) \
	  ['TRANSPORT'].mean().reset_index()
	
	df_pivot = total_transport_social_ncr.pivot(
	  index="City",
	  columns="SOCIAL_CLASS",
	  values="TRANSPORT"
	)
	
	df_pivot['Total'] = df_pivot.sum(axis=1)
	df_pivot = df_pivot.sort_values(by='Total', ascending=False).drop(columns=['Total'])
	
	plt.figure(figsize=(14, 7))
	df_pivot.plot(kind='bar', stacked=True, colormap='coolwarm', figsize=(14, 7))
	
	plt.title("Total Transport Expenses by Social Class in NCR Cities (Sorted)")
	plt.xlabel("City")
		
		- In all cities, the middle-income group dominates the total transport expenses
		
		- Lower-income and poor groups have minimal transport spending across all cities

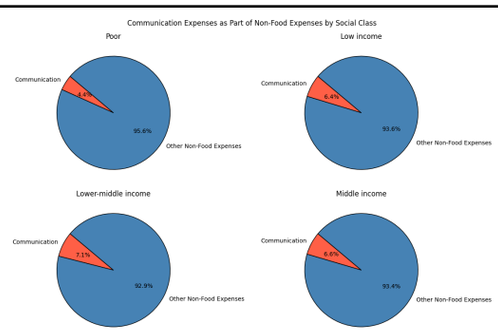
	plt.ylabel("Total Transport Expense (PHP)") plt.xticks(rotation=45, ha='right') plt.legend(title="Social Class", loc='upper right') plt.show()	
[ENTRATA] Transportation expenditure analysis and visualization This will illustrate the relationship between transport expenditure and total expenditure across different social classes in NCR.	plt.figure(figsize=(10, 6)) sns.scatterplot(data=final_fies, x="TOTEX", y="TRANSPORT", hue="SOCIAL_CLASS", alpha=0.7) plt.xlabel("Total Expenditure") plt.ylabel("Transport Expenditure") plt.title("Transport Expenditure vs. Total Expenditure by Social Class in NCR") plt.legend(title="Social Class", bbox_to_anchor=(1, 1)) plt.grid(True, linestyle="--", alpha=0.7) plt.show()	 - Higher-income groups allocate higher amounts to transportation as total expenditure increases. - Poor and low-income households allocate less to transport spending, likely due to public transport reliance and/or budget constraints. - There are some outliers in transport spending, most likely due to private vehicle ownership and its associated expenses.

[ENTRATA]
Transportation expenditure analysis and visualization
 	
This will illustrate the proportion of transport expenditure across various social classes.
	agg = final_fies.groupby('SOCIAL_CLASS').agg({
	'TRANSPORT': 'sum',
	'NFOOD': 'sum',
	'TOTEX': 'sum'
	}).reindex(order)
	
	agg['prop_nfood'] = (agg['TRANSPORT'] / agg['NFOOD']) * 100
	agg['prop_totex'] = (agg['TRANSPORT'] / agg['TOTEX']) * 100
	
	x = np.arange(len(agg))
	width = 0.35
	
	fig, ax = plt.subplots(figsize=(10, 6))
	bar1 = ax.bar(x - width/2, agg['prop_nfood'], width, label='Transport / Non-Food (%)', color='#66b3ff')
	bar2 = ax.bar(x + width/2, agg['prop_totex'], width, label='Transport / Total (%)', color='#ff9999')
	
	ax.set_xlabel('Social Class')
	ax.set_ylabel('Percentage (%)')
	ax.set_title('Proportion of Transport Expenditure by Social Class')
	ax.set_xticks(x)
	ax.set_xticklabels(agg.index, rotation=45)
	ax.legend()
	
	def autolabel(rects):
	for rect in rects:
	height = rect.get_height()
	ax.annotate(
	f'{height:.1f}%',
	xy=(rect.get_x() + rect.get_width() / 2, height),
	ha='center', va='bottom', fontsize=10
	)
	
	autolabel(bar1)
	autolabel(bar2)
	
	plt.tight_layout()
	plt.show()
		
		- Transport expenses only make up a small fraction of total expenditures
		
		- Transport has a slightly higher share in non-food expenses within non-food expenditures

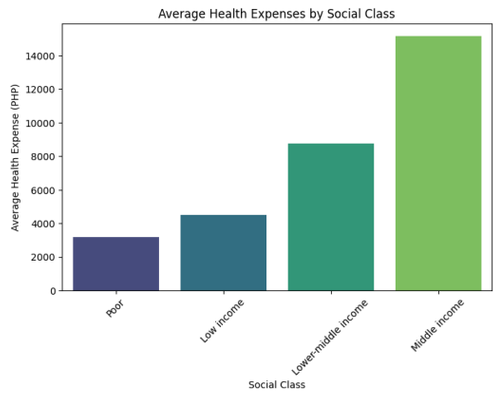
[ABRIGO]
Communication expenditure analysis and visualization
 	
This will identify spending patterns on communication services across income groups and provide insights through visual representations.
	avg_comm_expense = (
	  final_fies.groupby('SOCIAL_CLASS', observed=False)['COMMUNICATION']
	  .mean()
	  .reset_index()
	)
	
	plt.figure(figsize=(8, 5))
	sns.barplot(x='SOCIAL_CLASS', y='COMMUNICATION', data=avg_comm_expense, palette='viridis', hue='SOCIAL_CLASS')
	
	plt.title('Average Communication Expenses by Social Class')
	plt.xlabel('Social Class')
	plt.ylabel('Average Communication Expense (PHP)')
	plt.xticks(rotation=45)
	plt.grid(axis='y', linestyle='--', alpha=0.5)
	plt.show()
		
		This visualization shows the average expenditure for the communications category by the different social classes. The graph shows that the 'Middle income' social class have the highest average communication expense which can mean people in this group tend to utilize communication services for work or social engagement.

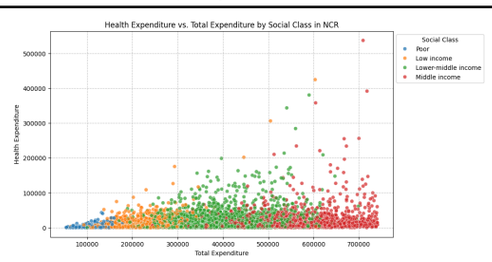
[ABRIGO]
Communication expenditure analysis and visualization
 	
This will show the percentage of total income allocated to communication services across various social groups.
	plt.figure(figsize=(8, 5))
	sns.stripplot(
	  x='SOCIAL_CLASS',
	  y=final_fies['COMMUNICATION'] / final_fies['TOINC'],
	  data=final_fies,
	  jitter=True,
	  alpha=0.6,
	  size=5,
	  palette='magma',
	  hue='SOCIAL_CLASS'
	)
	
	plt.title('Percentage of Total Income Spent on Communication')
	plt.xlabel('Social Class')
	plt.ylabel('Percentage (%)')
	plt.grid(axis='y', linestyle='--', alpha=0.5)
	plt.show()
		
		This visualization shows the percentage of total income spent on communication services per social group. The Middle income social group have the least or none 0.0% for communication expenses which can mean people of this group have a sure financial capability or may be required to have communication services.

[ABRIGO] Communication expenditure analysis and visualization This will show the proportion of communication expenses within non-food expenses.	nfood_total = final_fies['NFOOD'].sum() communication_total = final_fies['COMMUNICATION'].sum() remaining_nfood = nfood_total - communication_total labels = ['Communication', 'Other Non-Food Expenses'] sizes = [communication_total, remaining_nfood] colors = ['#FF6347', '#4682B4'] plt.figure(figsize=(6, 6)) plt.pie(sizes, labels=labels, colors=colors, autopct='%.1f%%', startangle=140, wedgeprops={'linewidth': 1, 'edgecolor': 'black'}) plt.title('Proportion of Communication in Non-Food Expenses') plt.show()	 This visualization shows the percentage of communication expenses for all non-food expenses. It shows a 6.8% which can mean that communication is considered an important expense for daily life in NCR

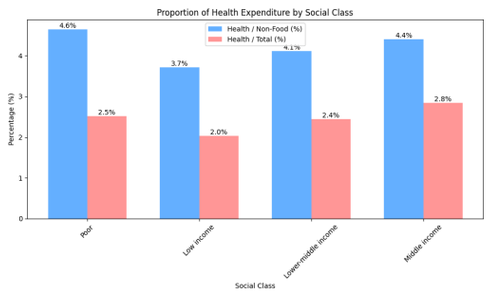
[ABRIGO]
Communication expenditure analysis and visualization
 	
This will illustrate the proportion of communication expenses within non-food expenses across all social classes.
	grouped_data = final_fies.groupby('SOCIAL_CLASS', observed=False).agg({
	  'COMMUNICATION': 'sum',
	  'NFOOD': 'sum'
	})
	
	num_classes = len(grouped_data)
	rows = (num_classes // 2) + (num_classes % 2 > 0) # For row layout
	
	fig, axes = plt.subplots(rows, 2, figsize=(12, rows * 4))
	axes = axes.flatten()
		
		This visualization shows the proportion of

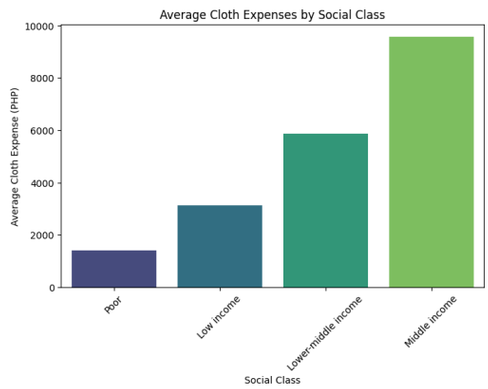
	for idx, (class_name, row) in enumerate(grouped_data.iterrows()):
	  communication_total = row['COMMUNICATION']
	  remaining_nfood = row['NFOOD'] - communication_total
	
	  sizes = [communication_total, remaining_nfood]
	  labels = ['Communication', 'Other Non-Food Expenses']
	
	  axes[idx].pie(sizes, labels=labels, autopct='%.1f%%',
	  colors=['#FF6347', '#4682B4'],
	  startangle=140, wedgeprops={'linewidth': 1, 'edgecolor': 'black'})
	
	  axes[idx].set_title(f'{class_name}')
	
	for idx in range(len(grouped_data), len(axes)):
	  axes[idx].axis('off')
	
	plt.suptitle('Communication Expenses as Part of Non-Food Expenses by Social Class')
	plt.tight_layout()
	plt.show()
		Communication in Non-Food Expenses by Social Classes. It shows low and lower middle income social groups have the highest percentage spent for communications expenditure which can mean that the minimum price for communications is expensive for these groups or a result of budget allocations.

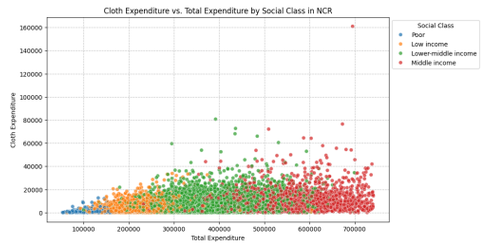
[ENTRATA]
Health expenditure analysis and visualization
 	
This will identify spending patterns on healthcare across income groups and provide insights through visual representations.
	avg_health_expense = final_fies.groupby('SOCIAL_CLASS')['HEALTH'].mean().reset_index()
	order = ['Poor', 'Low income', 'Lower-middle income', 'Middle income']
	
	plt.figure(figsize=(8,5))
	sns.barplot(
	  x='SOCIAL_CLASS',
	  y='HEALTH',
	  data=avg_health_expense,
	  palette='viridis',
	  hue='SOCIAL_CLASS',
	  order=order
	)
	
	plt.title('Average Health Expenses by Social Class')
	plt.xlabel('Social Class')
	plt.ylabel('Average Health Expense (PHP)')
	plt.xticks(rotation=45)
	plt.show()
		
		- Wealthier groups have greater access to healthcare. As seen in the visualizations, middle-income individuals likely prioritize healthcare and may afford private health services, insurance, and/or treatments. Unfortunately, lower-income groups may have limited access to private healthcare services due to financial constraints, that is why they allocate the least amount of money for health expenses.
		
		- The difference in health spending across social classes shows inequality in healthcare access.

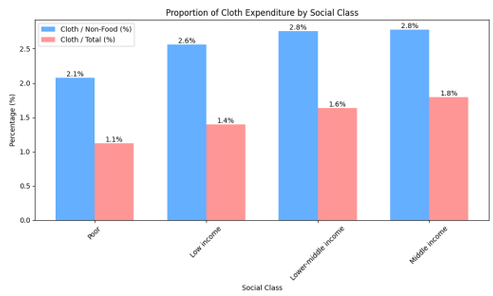
[ENTRATA] Health expenditure analysis and visualization This will illustrate the relationship between health expenditure and total expenditure and check out the outliers.	plt.figure(figsize=(10, 6)) sns.scatterplot(data=final_fies, x="TOTEX", y="HEALTH", hue="SOCIAL_CLASS", alpha=0.7) plt.xlabel("Total Expenditure") plt.ylabel("Health Expenditure") plt.title("Health Expenditure vs. Total Expenditure by Social Class in NCR") plt.legend(title="Social Class", bbox_to_anchor=(1, 1)) plt.grid(True, linestyle="--", alpha=0.7) plt.show()	 - As total expenditure increases, health expenditure also rises. However, the rate of increase is not uniform across all social classes. - The health expenditure for the "Poor" and "Low-income" groups is relatively low. This might be because they prioritize other necessities over healthcare. - A few red data points (middle-income households) indicate an extremely high allocation of income to health expenditures, possibly due to major medical

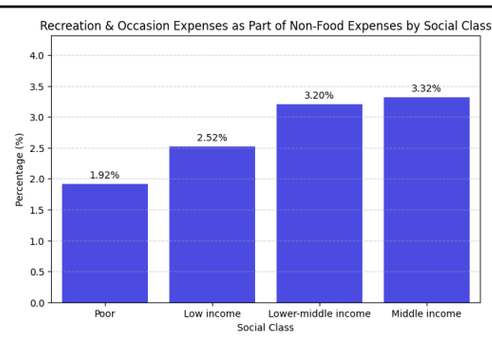
		procedures or chronic illnesses.

[ENTRATA]
Health expenditure analysis and visualization
 	
This will compare the proportion of health spending in non-food expenditure and total expenditure across all social classes.
	agg = final_fies.groupby('SOCIAL_CLASS').agg({
	'HEALTH': 'sum',
	'NFOOD': 'sum',
	'TOTEX': 'sum'
	}).reindex(order)
	
	agg['prop_nfood'] = (agg['HEALTH'] / agg['NFOOD']) * 100
	agg['prop_totex'] = (agg['HEALTH'] / agg['TOTEX']) * 100
	
	x = np.arange(len(agg))
	width = 0.35
	
	fig, ax = plt.subplots(figsize=(10, 6))
	bar1 = ax.bar(x - width/2, agg['prop_nfood'], width, label='Health / Non-Food (%)', color='#66b3ff')
	bar2 = ax.bar(x + width/2, agg['prop_totex'], width, label='Health / Total (%)', color='#ff9999')
	
	ax.set_xlabel('Social Class')
	ax.set_ylabel('Percentage (%)')
	ax.set_title('Proportion of Health Expenditure by Social Class')
	ax.set_xticks(x)
	ax.set_xticklabels(agg.index, rotation=45)
	ax.legend()
	
	def autolabel(rects):
	for rect in rects:
	height = rect.get_height()
	ax.annotate(
	f'{height:.1f}%',
	xy=(rect.get_x() + rect.get_width() / 2, height),
	ha='center', va='bottom', fontsize=10
	)
	
	autolabel(bar1)
	autolabel(bar2)
	
	plt.tight_layout()
	plt.show()
		
		- Middle-income households allocate the highest proportion of their budget to healthcare, while poor and low-income groups spend less
		
		- Compared to other expenses, all the social classes allocate a smaller percentage to health, likely because of poor healthcare programs in the country or they do not care/know the importance of taking care of their health.

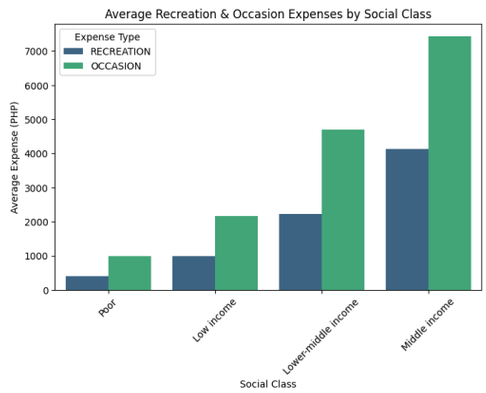
[ENTRATA] Clothing expenditure analysis and visualization This will identify spending patterns on clothing across income groups and provide insights through visual representations.	avg_cloth_expense = final_fies.groupby('SOCIAL_CLASS')['CLOTH'].mean().reset_index() order = ['Poor', 'Low income', 'Lower-middle income', 'Middle income'] plt.figure(figsize=(8,5)) sns.barplot(x='SOCIAL_CLASS', y='CLOTH', data=avg_cloth_expense, palette='viridis', hue='SOCIAL_CLASS', order=order) plt.title('Average Cloth Expenses by Social Class') plt.xlabel('Social Class') plt.ylabel('Average Cloth Expense (PHP)') plt.xticks(rotation=45) plt.show()	 - Common observation in different components of expenditures is that as income increases, the spending on clothes also rises. - The poor spend the least on clothing, likely prioritizing basic necessities over optional purchases. - Clothing expenditure may not just be based on necessity, but also lifestyle preferences and societal expectations.

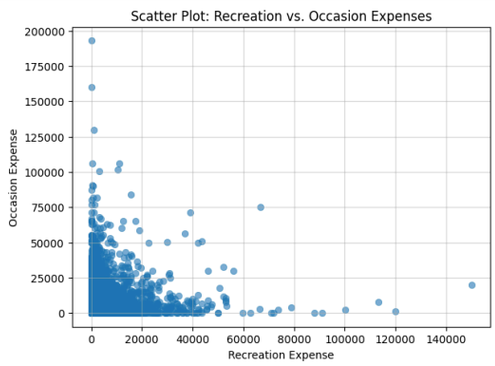
[ENTRATA] Clothing expenditure analysis and visualization This will illustrate how spending on clothing correlates with total expenditures across different social classes.	plt.figure(figsize=(10, 6)) sns.scatterplot(data=final_fies, x="TOTEX", y="CLOTH", hue="SOCIAL_CLASS", alpha=0.7) plt.xlabel("Total Expenditure") plt.ylabel("Cloth Expenditure") plt.title("Cloth Expenditure vs. Total Expenditure by Social Class in NCR") plt.legend(title="Social Class", bbox_to_anchor=(1, 1)) plt.grid(True, linestyle="--", alpha=0.7) plt.show()	 - Higher-income groups spend more on clothing, suggesting that they have more disposable income allocated for fashion, branded clothing, or luxury apparel. - There is a red-dot outlier (middle income), which suggests that a certain individual allocated a huge amount to clothing. This may be an influencer who needs luxury clothing or simply someone who enjoys spending on luxury brands.

[ENTRATA]
Clothing expenditure analysis and visualization
 	
This will show percentages relative to both non-food and total expenditures across social groups
	agg = final_fies.groupby('SOCIAL_CLASS').agg({
	'CLOTH': 'sum',
	'NFOOD': 'sum',
	'TOTEX': 'sum'
	}).reindex(order)
	
	agg['prop_nfood'] = (agg['CLOTH'] / agg['NFOOD']) * 100
	agg['prop_totex'] = (agg['CLOTH'] / agg['TOTEX']) * 100
	
	x = np.arange(len(agg))
	width = 0.35
	
	fig, ax = plt.subplots(figsize=(10, 6))
	bar1 = ax.bar(x - width/2, agg['prop_nfood'], width, label='Cloth / Non-Food (%)', color='#66b3ff')
	bar2 = ax.bar(x + width/2, agg['prop_totex'], width, label='Cloth / Total (%)', color='#ff9999')
	
	ax.set_xlabel('Social Class')
	ax.set_ylabel('Percentage (%)')
	ax.set_title('Proportion of Cloth Expenditure by Social Class')
	ax.set_xticks(x)
	ax.set_xticklabels(agg.index, rotation=45)
	ax.legend()
	
	def autolabel(rects):
	for rect in rects:
	height = rect.get_height()
	ax.annotate(
	f'{height:.1f}%',
	xy=(rect.get_x() + rect.get_width() / 2, height),
	ha='center', va='bottom', fontsize=10
	)
	
	autolabel(bar1)
	autolabel(bar2)
	
	plt.tight_layout()
	plt.show()
		
		- Surprisingly, poor households allocate 2.1% of their non-food budget to clothing, which is close to the 2.8% in middle-income households.
		
		- The difference between social classes is not as huge, suggesting that while middle-income earners spend more in absolute terms, they do not allocate an extreme portion of their budget to clothing.

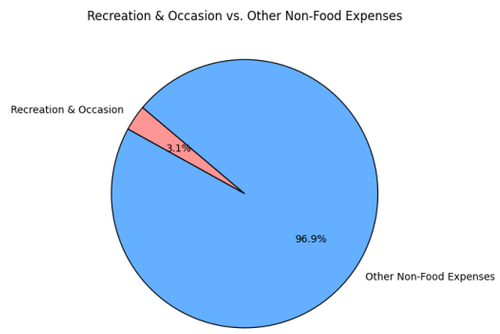
[CRUZ]
Recreation & Leisure expenditure analysis and visualization
 	
This will reveal spending patterns on recreation and leisure across income groups and provide insights through visual representations.
	# Group by social class and calculate mean percentage of recreation expenses
	recreation_percentages = final_fies.groupby('SOCIAL_CLASS', observed=True).apply(
	lambda x: ((x['RECREATION'] + x['OCCASION']) / x['NFOOD']).mean() * 100,
	include_groups=False
	)
	
	# Bar plot for better visualization of small percentages
	plt.figure(figsize=(8, 5))
		

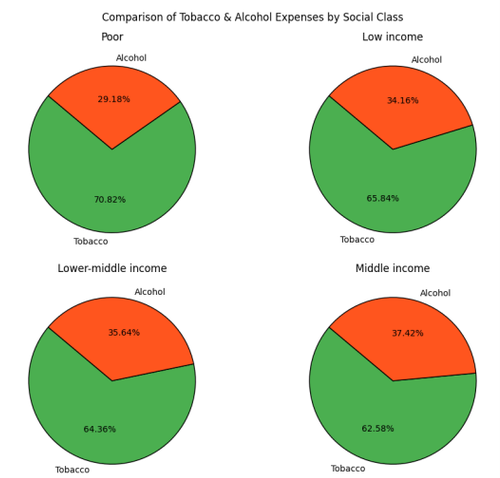
	sns.barplot(x=recreation_percentages.index, y=recreation_percentages.values, color='blue', alpha=0.8)
	
	# Add percentage labels for clarity
	for i, v in enumerate(recreation_percentages.values):
	  plt.text(i, v + 0.1, f'{v:.2f}%', ha='center')
	
	plt.title('Recreation & Occasion Expenses as Part of Non-Food Expenses by Social Class')
	plt.xlabel('Social Class')
	plt.ylabel('Percentage (%)')
	plt.grid(axis='y', linestyle='--', alpha=0.5)
	plt.ylim(0, max(recreation_percentages.values) + 1) # Adjust y-axis for better visibility
	plt.show()
		-The bar chart illustrates the proportion of Recreation & Occasion expenses relative to total Non-Food (NFOOD) expenses across different social classes.
		
		-Poor households allocate the smallest percentage (1.92%) of their non-food budget to recreation and special occasions.
		
		-As income increases, the proportion spent on these activities rises, reaching 3.32% for the Middle-income group.
		
		-This suggests that higher-income groups have more flexibility to spend on leisure and celebrations, whereas lower-income groups prioritize essential non-food needs.

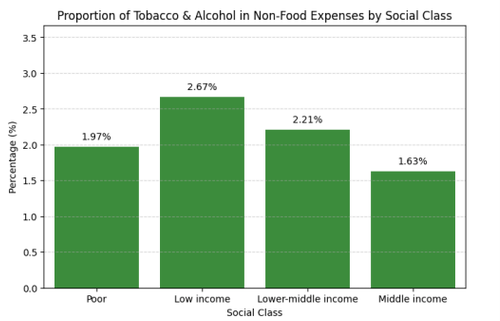
[CRUZ]
Recreation & Leisure expenditure analysis and visualization
 	
This will provide insights into the average recreation and occasion expenses across various social classes.
	# Compute average recreation & occasion expenses by social class
	avg_recreation_expense = final_fies.groupby('SOCIAL_CLASS')[['RECREATION', 'OCCASION']].mean().reset_index()
	order = ['Poor', 'Low income', 'Lower-middle income', 'Middle income']
	
	# Melt the dataframe for grouped bar plot
	avg_recreation_expense = avg_recreation_expense.melt(id_vars='SOCIAL_CLASS', var_name='Expense Type', value_name='Average Expense')
	
	# Bar plot
	plt.figure(figsize=(8,5))
	sns.barplot(
	  x='SOCIAL_CLASS',
	  y='Average Expense',
	  hue='Expense Type',
	  data=avg_recreation_expense,
	  palette='viridis',
	  order=order
	)
	
	plt.title('Average Recreation & Occasion Expenses by Social Class')
	plt.xlabel('Social Class')
	plt.ylabel('Average Expense (PHP)')
	plt.xticks(rotation=45)
	plt.legend(title='Expense Type')
	plt.show()
		
		-This grouped bar chart displays the average recreation and occasion expenses across different social classes.
		
		-Both expense categories (Recreation & Occasion) increase as income level rises.
		
		-The Middle-income group spends the most on both categories, with Occasion expenses surpassing Recreation.
		
		-The Poor social class has the lowest spending, reflecting limited discretionary income for leisure and celebrations.
		
		-The trend suggests that higher-income groups prioritize social and recreational activities more than lower-income groups.

[CRUZ]
Recreation & Leisure expenditure analysis and visualization
 	
This will reveal the spending dynamics between Recreation and Occasion expenses.
	plt.figure(figsize=(7, 5))
	sns.scatterplot(
	  x=final_fies['RECREATION'],
	  y=final_fies['OCCASION'],
	  alpha=0.6,
	  edgecolor=None
	)
	
	plt.title('Scatter Plot: Recreation vs. Occasion Expenses')
	plt.xlabel('Recreation Expense')
	plt.ylabel('Occasion Expense')
	plt.grid(alpha=0.5)
	plt.show()
		
		-This scatter plot visualizes the relationship between Recreation and Occasion expenses.
		
		-The majority of points cluster near the lower-left corner, meaning most households have low spending on both categories.
		
		-A few outliers represent households with very high spending on either Recreation or Occasion.

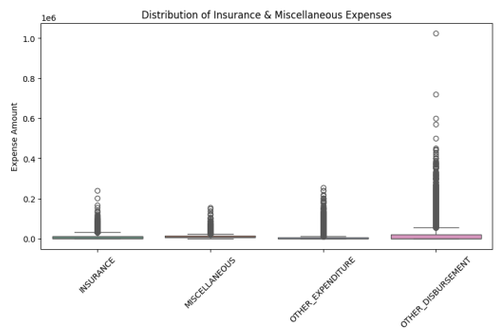
		-There appears to be a weak positive correlation, suggesting that those who spend more on Recreation also tend to spend more on Occasion expenses, but the relationship is not very strong.

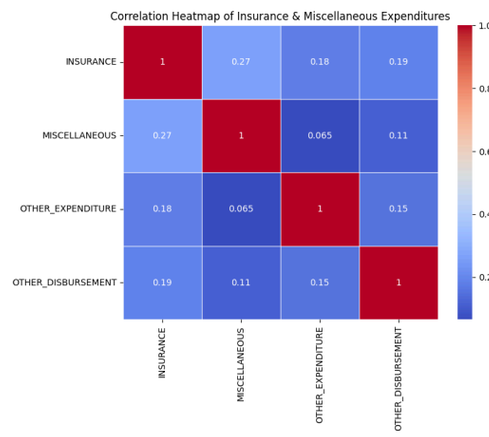
[CRUZ]
Recreation & Leisure expenditure analysis and visualization
 	
This will highlight the allocation of household budgets toward Recreation and Occasion expenses compared to other non-food categories.
	# Calculate total expenses for Recreation + Occasion and Other Non-Food
	total_recreation_occasion = final_fies[['RECREATION', 'OCCASION']].sum().sum()
	total_nfood = final_fies['NFOOD'].sum()
	
	# Calculate Other Non-Food Expenses
	other_nfood_expense = total_nfood - total_recreation_occasion
	
	# Data for the pie chart
	labels = ['Recreation & Occasion', 'Other Non-Food Expenses']
	sizes = [total_recreation_occasion, other_nfood_expense]
	colors = ['#ff9999', '#66b3ff']
	
	# Create pie chart
	plt.figure(figsize=(6, 6))
	plt.pie(sizes, labels=labels, autopct='%1.1f%%', colors=colors, startangle=140, wedgeprops={'edgecolor': 'black'})
	
	plt.title('Recreation & Occasion vs. Other Non-Food Expenses')
	plt.show()
		
		-The pie chart compares the share of Recreation & Occasion expenses against total non-food expenses.
		
		-Recreation & Occasion expenses make up only 3.1%, while Other Non-Food Expenses account for 96.9%.
		
		-This suggests that households allocate a very small portion of their budget to recreation and special occasions compared to other non-food necessities.

[ABRIGO]
Alcohol & Tobacco expenditure analysis and visualization
 	
This will identify spending patterns on alcohol and tobacco across income groups and provide insights through visual representations.
	percentage_data = final_fies.groupby('SOCIAL_CLASS', observed=False).apply(
	  lambda x: [
	  ((x['TOBACCO']) / x['NFOOD']).mean() * 100,
	  ((x['ALCOHOL']) / x['NFOOD']).mean() * 100
	  ],
	  include_groups=False)
	
	social_classes = percentage_data.index
	values = percentage_data.values
	
	plt.figure(figsize=(10, 8))
	
	for i, (label, value) in enumerate(zip(social_classes, values), 1):
	  plt.subplot(2, 2, i)
	  plt.pie(
	  value,
	  labels=['Tobacco', 'Alcohol'],
	  autopct='%.2f%%',
	  colors=['#4CAF50', '#FF5722'],
	  startangle=140,
	  wedgeprops={'edgecolor': 'black'}
	  )
	  plt.title(f'{label}')
	
	plt.suptitle('Comparison of Tobacco & Alcohol Expenses by Social Class')
	plt.tight_layout()
	plt.show()
		
		This visualization shows the comparison between tobacco and alcohol expenses by social class groups. All social groups spend more on tobacco than alcohol.

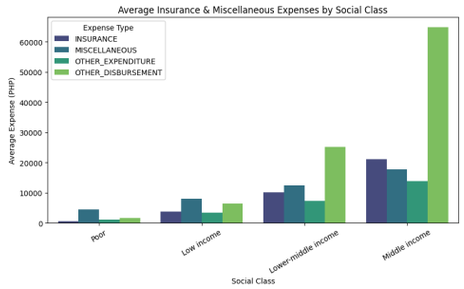
[ABRIGO]
Alcohol & Tobacco expenditure analysis and visualization
 	
This will examine the proportion of tobacco and alcohol expenses within non-food budgets across different social classes.
	percentages = final_fies.groupby('SOCIAL_CLASS', observed=False).apply(
	  lambda x: ((x['TOBACCO'] + x['ALCOHOL']) / x['NFOOD']).mean() * 100,
	  include_groups=False
	)
	
	plt.figure(figsize=(8, 5))
	sns.barplot(x=percentages.index, y=percentages.values, color='green', alpha=0.8)
	
	for i, v in enumerate(percentages.values):
	  plt.text(i, v + 0.1, f'{v:.2f}%', ha='center')
		
		This visualization shows the proportion of tobacco and

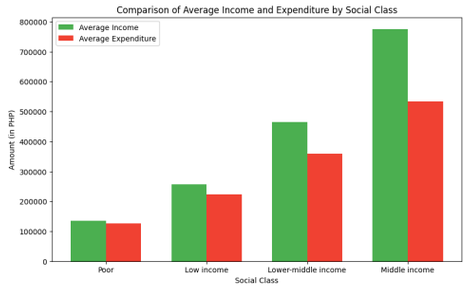
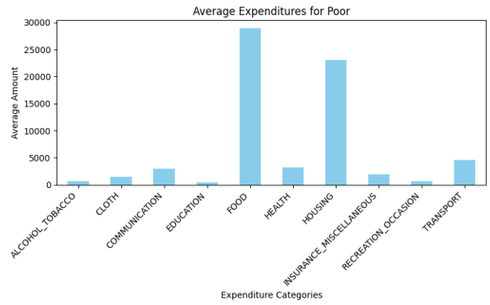
	plt.title('Proportion of Tobacco & Alcohol in Non-Food Expenses by Social Class') plt.xlabel('Social Class') plt.ylabel('Percentage (%)') plt.grid(axis='y', linestyle='--', alpha=0.5) plt.ylim(0, max(percentages.values) + 1) plt.show()	alcohol in non food expenses by social class. It shows lower-middle income, low income and poor social groups tend to allocate more budget for both products as compared to middle-income social groups. This may be influenced by events like celebrations or sad events where there is a culture to drink during those occasions.

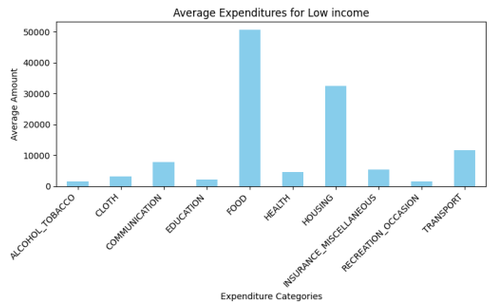
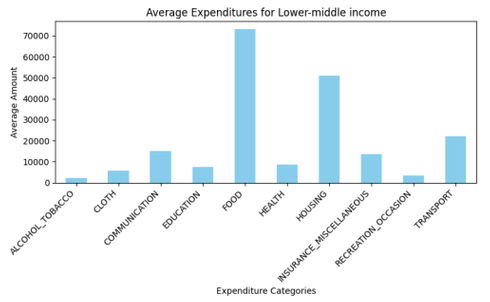
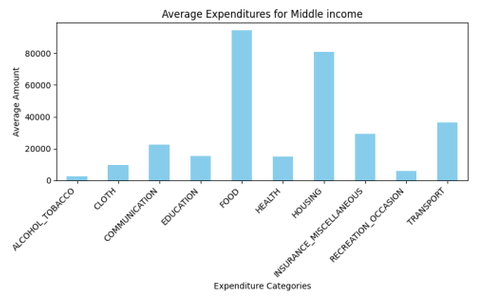
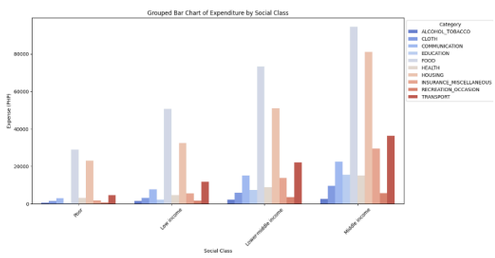
[CRUZ]
Insurance & Miscellaneous expenditure analysis and visualization
 	
This will identify spending patterns on insurance and other miscellaneous expenses across income groups and provide insights through visual representations.
	# Columns for analysis
	columns = ['INSURANCE', 'MISCELLANEOUS', 'OTHER_EXPENDITURE', 'OTHER_DISBURSEMENT']
	
	# Boxplot to check for outliers
	plt.figure(figsize=(10, 5))
	sns.boxplot(data=final_fies[columns], palette="Set2")
	plt.title("Distribution of Insurance & Miscellaneous Expenses")
	plt.ylabel("Expense Amount")
	plt.xticks(rotation=45)
	plt.show()
	
	# Summary statistics
	print(final_fies[columns].describe())
		
		INSURANCE MISCELLANEOUS OTHER_EXPENDITURE OTHER_DISBURSEMENT
		count 18848.000000 18848.000000 18848.000000 1.884800e+04
		mean 8666.942434 11158.858553 6387.040429 2.197597e+04
		std 12040.608436 7949.249707 16191.771556 4.771644e+04
		min 0.000000 558.000000 0.000000 0.000000e+00
		25% 0.000000 6366.000000 0.000000 0.000000e+00
		50% 5160.000000 9439.500000 300.000000 0.000000e+00
		75% 12480.000000 13725.500000 5000.000000 2.220000e+04
		max 239630.000000 156164.000000 253584.000000 1.024000e+06
		
		-Presence of Outliers: Circles above each box indicate outliers, especially in OTHER_DISBURSEMENT, exceeding 2,000,000 PHP.
		
		-Expense Concentration: Most expenses are low, with the median close to the bottom of each box. Interquartile Range (IQR): The small boxes indicate most values fall within a narrow range.
		
		-OTHER_DISBURSEMENT Variability: This category has the widest spread and extreme outliers.
		
		-Comparison Across Categories: INSURANCE, MISCELLANEOUS, and OTHER_EXPENDITURE have similar distributions, while OTHER_DISBURSEMENT has significantly higher values.
		
		-Potential Influences: Large outliers in OTHER_DISBURSEMENT may be due to one-time large expenses, business disbursements, or data errors.

[CRUZ]
Insurance & Miscellaneous expenditure analysis and visualization
 	
This will explore the relationships between insurance and miscellaneous expenditures, as visualized in the correlation heatmap.
	# Compute correlation
	correlation_matrix = final_fies[columns].corr()
	
	# Heatmap visualization
	plt.figure(figsize=(8, 6))
	sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', linewidths=0.5)
	plt.title("Correlation Heatmap of Insurance & Miscellaneous Expenditures")
	plt.show()
		
		-Correlation Strength: INSURANCE and MISCELLANEOUS have the highest correlation (0.27).
		Weak Relationships: Most correlations are below 0.3, indicating weak associations.
		
		-Self-Correlation: Each variable has a perfect correlation (1) with itself.

		-OTHER_EXPENDITURE & MISCELLANEOUS: Lowest correlation (0.065) suggests they are almost independent. -INSURANCE vs. OTHER_DISBURSEMENT: Moderate correlation (0.27) implies a weak link. -Color Interpretation: Darker red means higher correlation, while blue indicates weaker relationships.

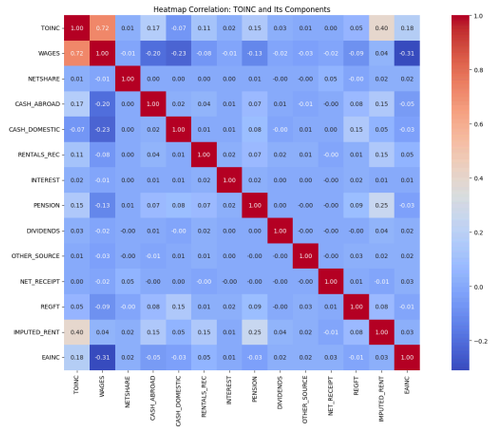
[CRUZ]
Insurance & Miscellaneous expenditure analysis and visualization
 	
This will compare the distribution of expenses across various social classes, specifically focusing on insurance, miscellaneous expenses, and other disbursement categories.
	# Group by Social Class and compute the mean
	avg_spending = final_fies.groupby("SOCIAL_CLASS")[columns].mean().reset_index()
	
	# Melt dataframe for easier plotting
	avg_spending_melted = avg_spending.melt(id_vars=["SOCIAL_CLASS"], var_name="Expense Type", value_name="Average Expense")
	
	# Barplot
	plt.figure(figsize=(10, 5))
	sns.barplot(x="SOCIAL_CLASS", y="Average Expense", hue="Expense Type", data=avg_spending_melted, palette="viridis")
	plt.title("Average Insurance & Miscellaneous Expenses by Social Class")
	plt.xlabel("Social Class")
	plt.ylabel("Average Expense (PHP)")
	plt.xticks(rotation=30)
	plt.legend(title="Expense Type")
	plt.show()
		
		-Expense Distribution: Middle-income groups have the highest insurance & miscellaneous expenses.
		-Income Effect: Higher social classes spend significantly more on OTHER_DISBURSEMENT.
		
		-Poor & Low Income: Lower expenses across all categories, except OTHER_DISBURSEMENT, which is still noticeable.
		
		-Spending Growth: OTHER_DISBURSEMENT increases exponentially as social class rises.
		
		-Comparison: INSURANCE, MISCELLANEOUS, and OTHER_EXPENDITURE increase steadily but at a lower rate.

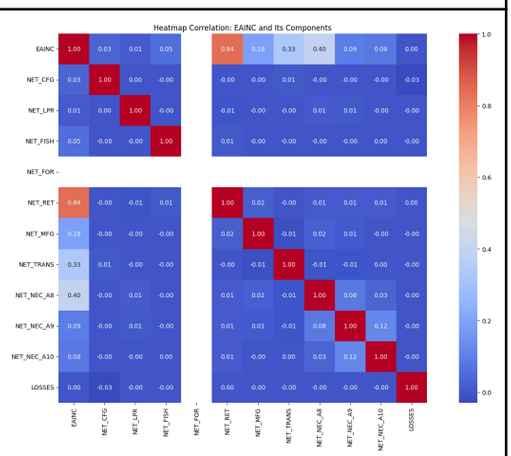
[ABRIGO] Generate statistics and provide EDA. Provide illustration This will summarize key dataset features, identify patterns and trends, and provide visual illustrations for better insights.	data = final_fies.groupby('SOCIAL_CLASS', observed=False)[['TOINC', 'TOTEX']].mean().reset_index() bar_width = 0.35 x = np.arange(len(data['SOCIAL_CLASS'])) plt.figure(figsize=(10, 6)) plt.bar(x - bar_width/2, data['TOINC'], width=bar_width, label='Average Income', color='#4CAF50') plt.bar(x + bar_width/2, data['TOTEX'], width=bar_width, label='Average Expenditure', color='#F44336') plt.xticks(x, data['SOCIAL_CLASS']) plt.title('Comparison of Average Income and Expenditure by Social Class') plt.xlabel('Social Class') plt.ylabel('Amount (in PHP)') plt.legend() plt.show()	 This visualization shows the comparison between average income and expenses by social group. It shows a higher average for income in all social classes which can mean people in NCR can budget their money wisely.
[ABRIGO] Generate statistics and provide EDA. Provide illustration This will illustrate the average expenditures by social class, emphasizing that food, housing, and transport dominate spending across all social classes.	avg_expenditure = final_fies.groupby('SOCIAL_CLASS', observed=False).agg({ 'ALCOHOL': 'mean', 'TOBACCO': 'mean', 'CLOTH': 'mean', 'COMMUNICATION': 'mean', 'EDUCATION': 'mean', 'FOOD_HOME': 'mean', 'FOOD_OUTSIDE': 'mean', 'HEALTH': 'mean', 'HOUSING_WATER': 'mean', 'FURNISHING': 'mean', 'IMPUTED_RENT': 'mean', 'INSURANCE': 'mean', 'MISCELLANEOUS': 'mean', 'OTHER_EXPENDITURE': 'mean', 'OTHER_DISBURSEMENT': 'mean', 'RECREATION': 'mean', 'OCCASION': 'mean', 'TRANSPORT': 'mean' }) # Combine related categories avg_expenditure['ALCOHOL_TOBACCO'] = avg_expenditure[['ALCOHOL',	

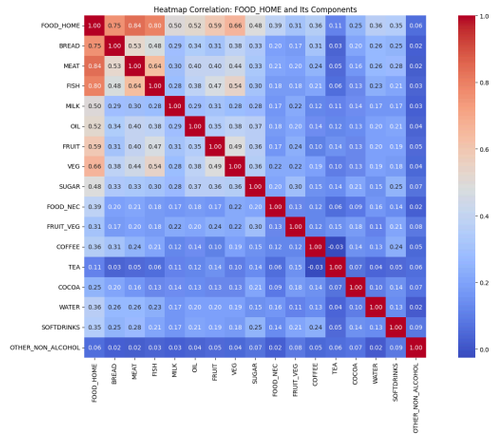
	'TOBACCO']].mean(axis=1) avg_expenditure['FOOD'] = avg_expenditure[['FOOD_HOME', 'FOOD_OUTSIDE']].mean(axis=1) avg_expenditure['HOUSING'] = avg_expenditure[['HOUSING_WATER', 'FURNISHING', 'IMPUTED_RENT']].mean(axis=1) avg_expenditure['RECREATION_OCCASION'] = avg_expenditure[['RECREATION', 'OCCASION']].mean(axis=1) avg_expenditure['INSURANCE_MISCELLANEOUS'] = avg_expenditure[['INSURANCE','MISCELLANEOUS', 'OTHER_EXPENDITURE', 'OTHER_DISBURSEMENT']].mean(axis=1) avg_expenditure = avg_expenditure[['ALCOHOL_TOBACCO', 'CLOTH', 'COMMUNICATION', 'EDUCATION', 'FOOD', 'HEALTH', 'HOUSING', 'INSURANCE_MISCELLANEOUS', 'RECREATION_OCCASION', 'TRANSPORT']] for social_class in avg_expenditure.index: plt.figure(figsize=(8, 5)) avg_expenditure.loc[social_class].plot(kind='bar', color='skyblue') plt.title(f'Average Expenditures for {social_class}') plt.ylabel('Average Amount') plt.xlabel('Expenditure Categories') plt.xticks(rotation=45, ha='right') plt.tight_layout() plt.show() #========================== avg_expenditure = avg_expenditure.reset_index().rename(columns={"inde x": "SOCIAL_CLASS"}) data_melted = avg_expenditure.melt(id_vars=["SOCIAL_CLASS"], var_name="Category", value_name="Expense") palette = sns.color_palette("coolwarm", len(avg_expenditure.columns) - 1) plt.figure(figsize=(14, 7)) sns.barplot(data=data_melted, x="SOCIAL_CLASS", y="Expense", hue="Category", palette=palette) plt.title("Grouped Bar Chart of Expenditure by Social Class") plt.xticks(rotation=45) plt.xlabel("Social Class") plt.ylabel("Expense (PHP)") plt.legend(title="Category", bbox_to_anchor=(1, 1)) plt.show()	    This visualization shows average expenditures by social class. It shows for all social classes that FOOD, HOUSING and TRANSPORT are the main expenditures. There is also a trend of higher average values for other details such as RECREATION_OCCASION which can mean the better social class standing, the more money can be allocated for other necessities like clothes and budget for relaxation and occasions.
6. Application of Proximity (Distance Analysis) (20 pts)		

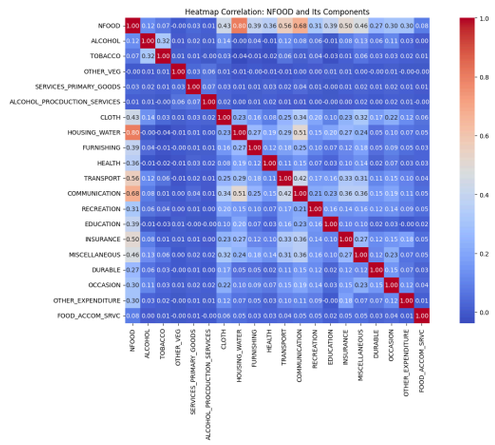
[PAYUMO]
Defined a **`plot_correlation_heatmap()`** function that generates a heatmap to visualize the correlation between a specified aggregated column and its component columns within a given dataframe.
	def plot_correlation_heatmap(dataframe, aggregated_column, components, title):
	# Subset the dataframe
	subset = dataframe[[aggregated_column] + components]
	
	# Calculate the correlation matrix
	correlation_matrix = subset.corr()
	
	# Plot heatmap
	plt.figure(figsize=(15, 10))
	sns.heatmap(correlation_matrix, annot=True, fmt=".2f", cmap="coolwarm", cbar=True, square=True)
	plt.title(title)
	plt.tight_layout()
		This task is only meant to create/define a function that will be used for the proceeding correlation analysis.

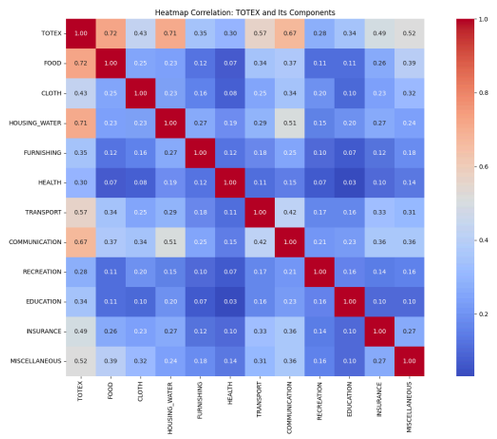
	plt.show()	

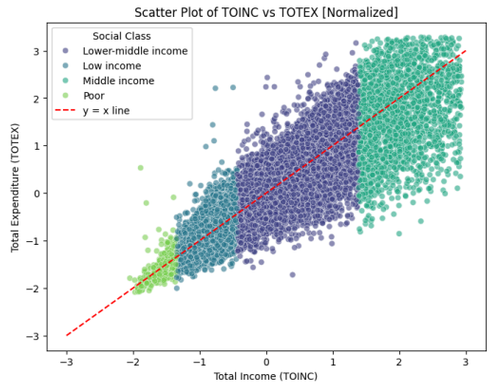
[PAYUMO]
Create a correlation heatmap for Total Income 'TOINC' and its components
 	
Visualize the correlation between Total Income (TOINC) and its components to identify key income drivers.
	# Define the aggregated column and its components
	aggregated_col = 'TOINC'
	components_list = ['WAGES', 'NETSHARE', 'CASH_ABROAD', 'CASH_DOMESTIC', 'RENTALS_REC', 'INTEREST', 'PENSION', 'DIVIDENDS', 'OTHER_SOURCE', 'NET_RECEIPT', 'REGFT', 'IMPUTED_RENT', 'EAINC']
	title = f'Heatmap Correlation: {aggregated_col} and Its Components'
	
	# Call the function
	plot_correlation_heatmap(final_fies, aggregated_col, components_list, title=title)
		
		- The `WAGES` component shows the highest correlation with `TOINC` at 0.72, emphasizing the strong dependence of households in NCR on employment income, which aligns with the region's highly urbanized and service-driven economy.
		- `IMPUTED_RENT` follows with a correlation of 0.40, reflecting the significant role of homeownership among households in NCR. This value represents the monetary benefit households gain by living in their own property instead of paying rent, particularly relevant in NCR's high property value context.
		- `EAINC` (Entrepreneurial Activities Income) and `CASH_ABROAD` show moderate correlations of 0.18 and 0.17, respectively. These values suggest that while some households engage in small businesses or receive overseas remittances, self-employment and remittance income play a smaller role compared to wage dependency in shaping total household income in NCR.

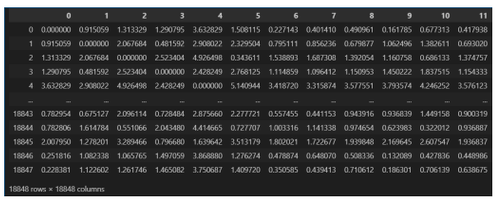
[PAYUMO]
Create a correlation heatmap for Income from Entrepreneurial Activities 'EAINC' and its components
 	
Visualize the correlation between Income from Entrepreneurial Activities (EAINC) and its components to identify key income contributors.
	# Define the aggregated column and its components
	aggregated_col = 'EAINC'
	components_list = ['NET_CFG', 'NET_LPR', 'NET_FISH', 'NET_FOR', 'NET_RET', 'NET_MFG', 'NET_TRANS', 'NET_NEC_A8', 'NET_NEC_A9', 'NET_NEC_A10', 'LOSSES']
	title = f'Heatmap Correlation: {aggregated_col} and Its Components'
	
	# Call the function
	plot_correlation_heatmap(final_fies, aggregated_col, components_list, title=title)
		
		- `NET_RET` shows the highest correlation with EAINC at 0.84, indicating that retail and wholesale activities are the primary drivers of entrepreneurial income among Filipino households.
		- No correlation was computed for `NET_FOR` due to all values being zero, which is expected given NCR's highly urbanized setting and lacking of forestry-related activities.

[PAYUMO]
Create a correlation heatmap for Indoor Food Expenses ‘FOOD_HOME’ and its components
 	
Visualize the correlation between Indoor Food Expenses (FOOD_HOME) and its components to understand spending patterns on home-consumed food.
	# Define the aggregated column and its components
	aggregated_col = 'FOOD_HOME'
	components_list = ['BREAD', 'MEAT', 'FISH', 'MILK', 'OIL', 'FRUIT', 'VEG', 'SUGAR', 'FOOD_NEC', 'FRUIT_VEG', 'COFFEE', 'TEA', 'COCOA', 'WATER', 'SOFTDRINKS', 'OTHER_NON_ALCOHOL']
	title = f'Heatmap Correlation: {aggregated_col} and Its Components'
	
	# Call the function
	plot_correlation_heatmap(final_fies, aggregated_col, components_list, title=title)
		
		- Based on the heatmap correlation of `FOOD_HOME` and its components, the top three contributors to household food expenses are `MEAT` (0.84), `FISH` (0.80), and `BREAD` (0.75), highlighting the significant share of protein sources and staple foods in NCR households' consumption.
		- Other notable contributors include `VEG` (0.66), `FRUIT` (0.59), `OIL` (0.52), `MILK` (0.50), and `SUGAR` (0.48), reflecting spending patterns focused on basic and versatile food items essential for daily meals.
		- These correlations suggest that households in NCR prioritize essential and nutrient-rich food items over non-essential products such as `COFFEE`, `TEA`, or `COCOA`, which show lower correlations with total food-at-home expenditure.

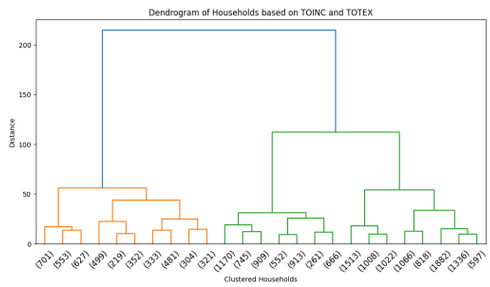
[PAYUMO]
Create a correlation heatmap for Non-Food Expenses ‘NFOOD’ and its components
 	
Visualize the correlation between Non-Food Expenses (NFOOD) and its components to analyze spending patterns beyond food consumption.
	# Define the aggregated column and its components
	aggregated_col = 'NFOOD'
	components_list = ['ALCOHOL', 'TOBACCO', 'OTHER_VEG', 'SERVICES_PRIMARY_GOODS', 'ALCOHOL_PROCDUCTION_SERVICES', 'CLOTH', 'HOUSING_WATER', 'FURNISHING', 'HEALTH', 'TRANSPORT', 'COMMUNICATION', 'RECREATION', 'EDUCATION', 'INSURANCE', 'MISCELLANEOUS', 'DURABLE', 'OCCASION', 'OTHER_EXPENDITURE', 'FOOD_ACCOM_SRVC']
	title = f'Heatmap Correlation: {aggregated_col} and Its Components'
	
	# Call the function
	plot_correlation_heatmap(final_fies, aggregated_col, components_list, title=title)
		
		- Based on the heatmap correlation of `NFOOD` and its components, the top contributors to non-food household expenses are `HOUSING_WATER` (0.80), `COMMUNICATION` (0.68), and `TRANSPORT` (0.56), highlighting the significant share of housing-related costs, connectivity, and mobility in NCR households' spending.
		- Other notable components include `INSURANCE` (0.50), `MISCELLANEOUS` (0.46), `CLOTH` (0.43), `FURNISHING` (0.39), `EDUCATION` (0.39), and `HEALTH` (0.36), reflecting additional necessary expenditures related to financial security, personal needs, and social services.
		- These correlations indicate that NCR households allocate most of their non-food spending to *essential living expenses*, while still maintaining moderate spending on education, health, and other personal and household-related needs.

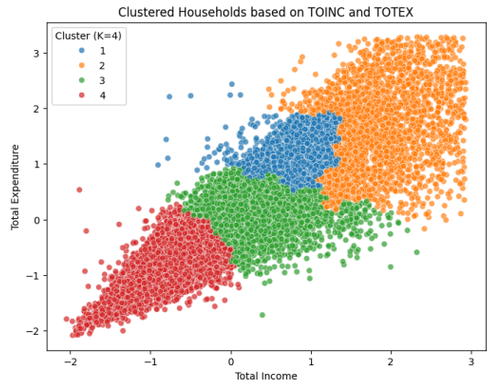
[PAYUMO]
Create a correlation heatmap for Total Expenses 'TOTEX' and its components
 	
Visualize the correlation between Total Expenses (TOTEX) and its components to identify major expenditure drivers.
	# Define the aggregated column and its components
	aggregated_col = 'TOTEX'
	components_list = ['FOOD', 'CLOTH', 'HOUSING_WATER', 'FURNISHING', 'HEALTH', 'TRANSPORT',
	'COMMUNICATION', 'RECREATION', 'EDUCATION', 'INSURANCE',
	'MISCELLANEOUS']
	title = f'Heatmap Correlation: {aggregated_col} and Its Components'
	
	# Call the function
	plot_correlation_heatmap(final_fies, aggregated_col, components_list, title=title)
		
		- Based on the heatmap correlation of `TOTEX` and its components, `FOOD` (0.72) and `HOUSING_WATER` (0.71) emerge as the most significant contributors, underscoring the substantial share of basic needs—daily consumption and housing-related expenses—in total household expenditures.
		- `HOUSING_WATER` captures spending on *housing, water, electricity, gas, and other fuels*, reflecting the essential cost of maintaining a household, especially in highly urbanized areas like NCR where utility expenses are substantial.
		- Other notable contributors include `COMMUNICATION` (0.67), `TRANSPORT` (0.57), `MISCELLANEOUS` (0.52), `INSURANCE` (0.49), and `CLOTH` (0.43), indicating that beyond food and housing, households also allocate significant portions of their expenses toward connectivity, mobility, financial security, and personal needs.

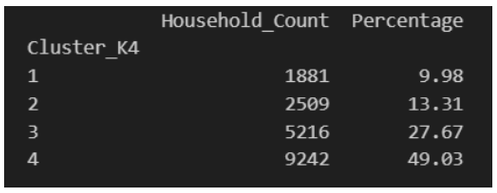
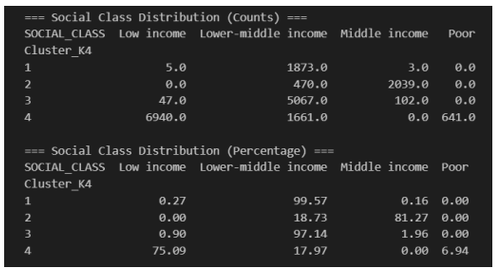
[PAYUMO]
Create a Scatter Plot showing the normalized data points in relation to the Total Income (TOINC) and Total Expenses (TOTEX). Also showed the distribution of `SOCIAL_CLASS` by setting it as the hue.
 	
Create a scatter plot to visualize the relationship between Total Income (TOINC) and Total Expenses (TOTEX) using normalized data points. Set SOCIAL_CLASS as the hue to highlight income-expenditure patterns across different social classes.
	# Prepare and scale the data while keeping the social class
	fies_normalized = final_fies[['TOINC', 'TOTEX', 'SOCIAL_CLASS']].copy()
	scaler = StandardScaler()
	fies_normalized[['TOINC_scaled', 'TOTEX_scaled']] = scaler.fit_transform(fies_normalized[['TOINC', 'TOTEX']])
	
	# Plot with hue based on SOCIAL_CLASS
	plt.figure(figsize=(8, 6))
	sns.scatterplot(
	data=fies_normalized,
	x='TOINC_scaled',
	y='TOTEX_scaled',
	hue='SOCIAL_CLASS',
	alpha=0.6,
	palette='viridis'
	)
	
	plt.title('Scatter Plot of TOINC vs TOTEX [Normalized]')
	plt.xlabel('Total Income (TOINC)')
	plt.ylabel('Total Expenditure (TOTEX)')
	
	plt.plot([-3, 3], [-3, 3], color='red', linestyle='--', label='y = x line')
	plt.legend(title='Social Class')
	plt.show()
		
		- The normalized scatter plot reveals a strong positive linear trend between Total Income (`TOINC`) and Total Expenditure (`TOTEX`), where there are generally more households that spend less than they earn, as seen by points below the red y = x line
		- The plot also shows that social class segmentation is distinct, with income groups clustering along the trend — lower-income households dominate the lower-left, while middle-income groups shift towards the upper-right.
		- Outliers are present, particularly households above the y = x line, suggesting some spend beyond their income, which may indicate borrowing, debt, or reliance on other non-reported resources.

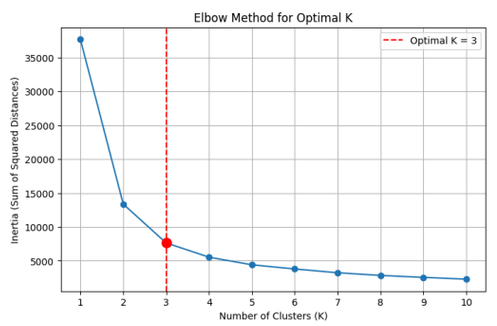
[PAYUMO]
Calculate the distance matrix **(Euclidean)** between the normalized TOINC and TOTEX columns. Generate a squareform to create a better display.
 	
To measure the similarity
	# Compute the pairwise Euclidean distances (condensed form)
	distance_matrix = pdist(fies_normalized[['TOINC_scaled', 'TOTEX_scaled']], metric='euclidean')
	
	# Display the distance matrix in a DataFrame
	distance_df =
		

between household income and expenditure patterns.	pd.DataFrame(squareform(distance_matrix), columns=fies_normalized.index, index=fies_normalized.index) distance_df	> After computing the Euclidean distances between households based on their Total Income (TOINC) and Total Expenditure (TOTEX), hierarchical clustering will be applied to systematically group households exhibiting similar economic behaviors. This method allows for uncovering underlying patterns and relationships within the data, providing a structured way to segment and profile households based on their income and spending characteristics.

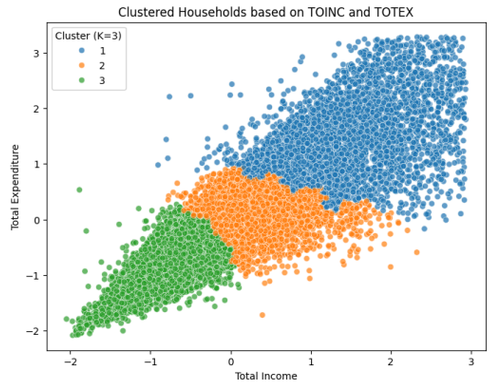
[PAYUMO]
Performed Ward's method to generate linkages in preparation of Hierarchical Clustering.
 	
Uses Ward's method to generate linkages for Hierarchical Clustering, ensuring compact and well-separated income-expenditure groups for better analysis.
	# Perform hierarchical clustering using Ward's method
	linkage_matrix = linkage(distance_matrix, method='ward')
	
	# Visualize the dendrogram
	plt.figure(figsize=(12, 6))
	dendrogram(linkage_matrix, truncate_mode='lastp', p=25, leaf_rotation=45, leaf_font_size=12)
	plt.title('Dendrogram of Households based on TOINC and TOTEX')
	plt.xlabel('Clustered Households')
	plt.ylabel('Distance')
	plt.show()
		
		- The dendrogram visualizes the hierarchical clustering of households based on normalized Total Income (`TOINC_scaled`) and Total Expenditure (`TOTEX_scaled`) using Ward's method, which minimizes within-cluster variance.
		- Two primary clusters emerge at higher linkage distances, representing distinct household groups with similar income-expenditure profiles, which are recursively partitioned into more homogeneous sub-clusters.
		- To interpret the cluster assignments, the dendrogram will be cut at a specified k number of clusters, and the resulting groups will be visualized through a scatter plot of `TOINC_scaled` vs. `TOTEX_scaled` with cluster-based color coding.

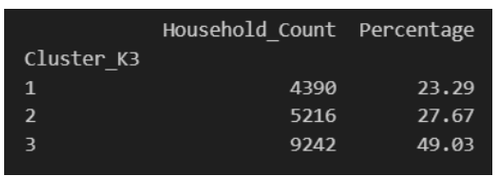
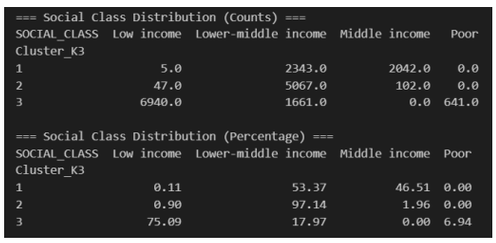
[PAYUMO] Performed Hierarchical Clustering using an arbitrary number of clusters (k=4) Helps to identify distinct financial behavior patterns across different socioeconomic segments in NCR.	fies_normalized['Cluster_K4'] = fcluster(linkage_matrix, t=4, criterion='maxclust') plt.figure(figsize=(8, 6)) sns.scatterplot(data=fies_normalized, x='TOINC_scaled', y='TOTEX_scaled', hue='Cluster_K4', palette='tab10', alpha=0.7) plt.title('Clustered Households based on TOINC and TOTEX') plt.xlabel('Total Income') plt.ylabel('Total Expenditure') plt.legend(title='Cluster (K=4)') plt.show()	 - The clusters align along a diagonal trend, reinforcing the positive relationship between income and expenditure—households with higher income generally spend more. The clusters effectively capture the economic progression across households in the dataset. - Cluster 4 (Red): Represents the households with the lowest income and expenditure, likely belonging to the poor and low-income classes facing financial constraints and limited spending capacity. - Cluster 3 (Green): Composed of households with moderate income but controlled spending, likely to reflect frugal low- to lower-middle-income families prioritizing efficient expense management. - Cluster 1 (Blue): Includes lower to middle-income households with balanced spending patterns, likely supported by stable employment or consistent income sources. - Cluster 2 (Orange): Captures the generally wealthier households with both high income and high expenditure, indicating greater financial flexibility and diversified consumption.

[PAYUMO]
Inspect the clusters (k=4) summary statistics
 	
This task helps segment households based on financial behavior
	cluster_counts = fies_normalized['Cluster_K4'].value_counts().sort_index()
	cluster_percent = (cluster_counts / cluster_counts.sum()) * 100
	result = pd.DataFrame({'Household_Count': cluster_counts, 'Percentage': cluster_percent.round(2)})
	print(result)
	
	profile = final_fies.merge(fies_normalized[['Cluster_K4']], left_index=True, right_index=True)
	cluster_profile = profile.groupby('Cluster_K4').agg({
	'TOINC': ['mean', 'median'],
	'TOTEX': ['mean', 'median'],
	'FSIZE': 'mean',
	'WAGES': 'mean',
	'EAINC': 'mean',
	'IMPUTED_RENT': 'mean',
	'FOOD': 'mean',
	'NFOOD': 'mean',
	'SOCIAL_CLASS': lambda x: x.value_counts(normalize=True).to_dict()
	})
	
	cluster_profile
	
	# Group by Cluster and SOCIAL_CLASS to get counts
	social_class_dist = profile.groupby(['Cluster_K4', 'SOCIAL_CLASS']).size().reset_index(name='Count')
	
	# Pivot the table for a cleaner view (clusters as rows, social classes as columns)
	social_class_pivot = social_class_dist.pivot(index='Cluster_K4', columns='SOCIAL_CLASS', values='Count').fillna(0)
	
	# Add percentage distribution per cluster
	social_class_pct = social_class_pivot.div(social_class_pivot.sum(axis=1), axis=0) * 100
	
	print("=== Social Class Distribution (Counts) ===")
	print(social_class_pivot)
	
	print("\n=== Social Class Distribution (Percentage) ===")
	print(social_class_pct.round(2))
		
		Household_Count Percentage
		Cluster_K4
		1 1881 9.98
		2 2509 13.31
		3 5216 27.67
		4 9242 49.03
		
		 
		
		 
		=== Social Class Distribution (Counts) ===
		SOCIAL_CLASS Low income Lower-middle income Middle income Poor
		Cluster_K4
		1 5.0 1873.0 3.0 0.0
		2 0.0 470.0 2039.0 0.0
		3 47.0 5067.0 102.0 0.0
		4 6940.0 1661.0 0.0 641.0
		
		=== Social Class Distribution (Percentage) ===
		SOCIAL_CLASS Low income Lower-middle income Middle income Poor
		Cluster_K4
		1 0.27 99.57 0.16 0.00
		2 0.00 18.73 81.27 0.00
		3 0.90 97.14 1.96 0.00
		4 75.09 17.97 0.00 6.94
		
		- The summary statistics provide deeper insights into each cluster's socio economic profile:
		- Cluster 4 is dominated by Low-Income households and is the only cluster containing the Poor class. Notably, food and non-food expenses are nearly equal, reflecting limited financial flexibility.
		- Clusters 3 and 1 share a similar Lower-Middle Income composition, but Cluster 1 shows higher income and expenditure, indicating better economic stability compared to the more frugal behavior of Cluster 3.
		- Cluster 2 represents the wealthier segment, primarily composed of Middle and some Lower-Middle Income households, with the highest average income and spending, suggesting greater financial capacity and diverse expenditures.

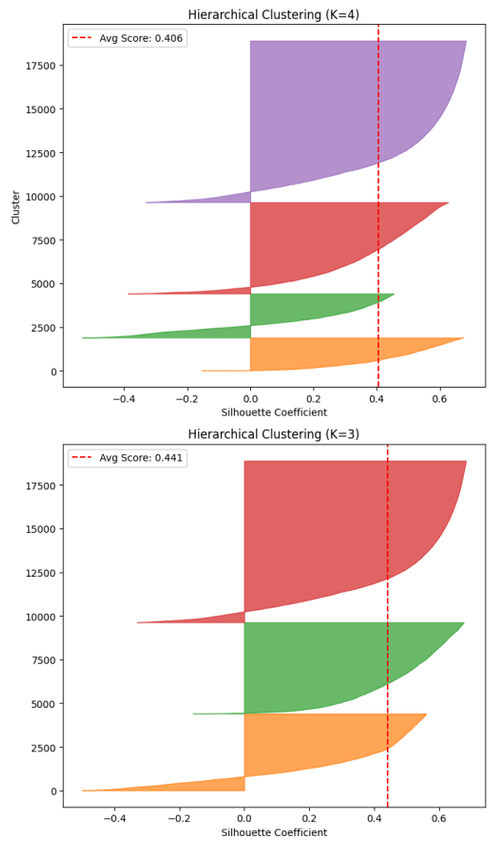
[PAYUMO]
Performed the Elbow Method as a validation step for the clusters
 	
To determine the optimal number of clusters for the dataset.
	X = fies_normalized[['TOINC_scaled', 'TOTEX_scaled']]
	inertia = []
	K_range = range(1, 11) # Test K from 1 to 10
	
	for k in K_range:
	kmeans = KMeans(n_clusters=k, random_state=42, n_init=10)
	kmeans.fit(X)
	inertia.append(kmeans.inertia_)
	# Detect the elbow (optimal k)
	knee = KneeLocator(K_range, inertia, curve='convex', direction='decreasing')
	optimal_k = knee.knee
	
	# Plot with annotation
	plt.figure(figsize=(8, 5))
	plt.plot(K_range, inertia, marker='o')
	plt.title('Elbow Method for Optimal K')
	plt.xlabel('Number of Clusters (K)')
	plt.ylabel('Inertia (Sum of Squared Distances)')
	plt.xticks(K_range)
	plt.grid(True)
		
		
		- Based on the Elbow Method plot, the optimal number of clusters appears to be K = 3. However, the initial K = 4 clustering remains valid, as it provided meaningful socioeconomic insights.
		- To further validate and compare the clustering results, hierarchical clustering will be reperformed with K = 3 to explore potential differences in household segmentation.

	# Annotate the detected optimal K
	if optimal_k:
	  plt.axvline(x=optimal_k, color='red', linestyle='--', label=f'Optimal K = {optimal_k}')
	  plt.scatter(optimal_k, inertia[optimal_k - min(K_range)], color='red', s=100, zorder=5)
	
	plt.legend()
	plt.show()
	
	print(f'Detected Optimal K: {optimal_k}')	

[PAYUMO]
Re-performed the Hierarchical Clustering using the optimal number of clusters (k=3)
 	
Refines Hierarchical Clustering with the optimal k=3, ensuring more precise income-expenditure group segmentation.
	# Perform Hierarchical Clustering with the optimal K
	fies_normalized['Cluster_K3'] = fcluster(linkage_matrix, t=optimal_k, criterion='maxclust')
	
	plt.figure(figsize=(8, 6))
	sns.scatterplot(
	  data=fies_normalized,
	  x='TOINC_scaled',
	  y='TOTEX_scaled',
	  hue='Cluster_K3',
	  palette='tab10',
	  alpha=0.7
	)
	plt.title('Clustered Households based on TOINC and TOTEX')
	plt.xlabel('Total Income')
	plt.ylabel('Total Expenditure')
	plt.legend(title='Cluster (K=3)')
	plt.show()
		
		- Based on the cluster plot, the data points are more well-distributed compared to the K=4 result, while still following a clear linear trend — higher income generally correlates with higher expenditure.
		  - Cluster 1 (Blue): Represents the majority of lower-middle to middle-income households, reflecting wealthier groups with generally higher income and higher expenditures.
		  - Cluster 2 (Orange): Captures the lower-middle income households, showing balanced income and expenditure levels.
		  - Cluster 3 (Green): Represents poor to lower-middle income households, reflecting financially constrained groups with low income and low expenditure.
		- Overall, the clusters demonstrate minimal overlap and few outliers, indicating a clear separation of socioeconomic groups based on income and spending patterns.

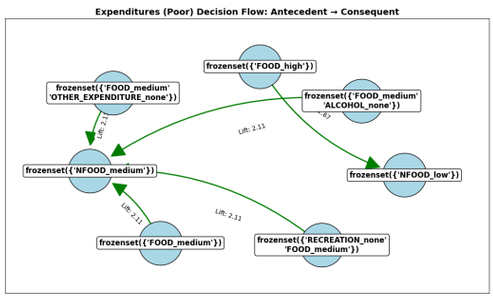
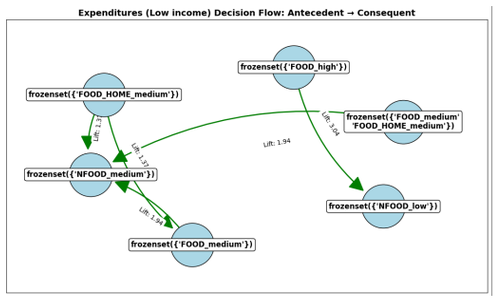
[PAYUMO]
Inspect the clusters (k=3) summary statistics
 	
Analyze the summary statistics of the k=3 clusters to understand their income-expenditure characteristics.
	cluster_counts = fies_normalized['Cluster_K3'].value_counts().sort_index()
	cluster_percent = (cluster_counts / cluster_counts.sum()) * 100
	result = pd.DataFrame({'Household_Count': cluster_counts, 'Percentage': cluster_percent.round(2)})
	print(result)
	
	profile = final_fies.merge(fies_normalized[['Cluster_K3']], left_index=True, right_index=True)
	cluster_profile = profile.groupby('Cluster_K3').agg({
	  'TOINC': ['mean', 'median'],
	  'TOTEX': ['mean', 'median'],
	  'FSIZE': 'mean',
	  'WAGES': 'mean',
	  'EAINC': 'mean',
	  'IMPUTED_RENT': 'mean',
	  'FOOD': 'mean',
	  'NFOOD': 'mean',
	  'SOCIAL_CLASS': lambda x: x.value_counts(normalize=True).to_dict()
	})
	
	cluster_profile
		
		 
		 
		- The summary statistics provide clearer insights into each cluster's socioeconomic characteristics:
		  - Cluster 1 (Blue 🔵) comprises the wealthier segment of the sample, a relatively balanced distribution of Middle and Lower-Middle Income households. This group has the highest average income and spending, with significant allocation toward non-food expenditures, indicating stronger financial capacity, asset ownership, and diverse

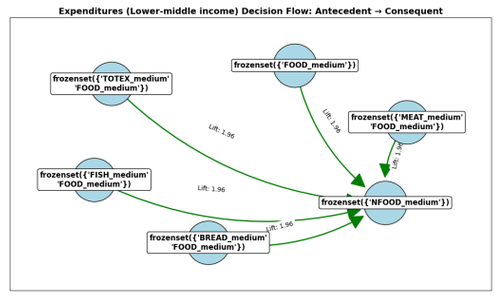
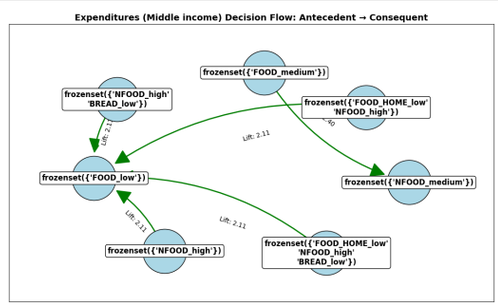
	# Group by Cluster and SOCIAL_CLASS to get counts
	social_class_dist = profile.groupby(['Cluster_K3', 'SOCIAL_CLASS']).size().reset_index(name='Count')
	
	# Pivot the table for a cleaner view (clusters as rows, social classes as columns)
	social_class_pivot = social_class_dist.pivot(index='Cluster_K3', columns='SOCIAL_CLASS', values='Count').fillna(0)
	
	# Add percentage distribution per cluster
	social_class_pct = social_class_pivot.div(social_class_pivot.sum(axis=1), axis=0) * 100
	
	print("=== Social Class Distribution (Counts) ===")
	print(social_class_pivot)
	
	print("\n=== Social Class Distribution (Percentage) ===")
	print(social_class_pct.round(2))
		consumption behavior.
		- Cluster 2 (Orange 🟠) captures the stable lower-middle income group with balanced income and expenses. While primarily composed of the Lower-Middle Income, this group demonstrates financial stability—able to cover necessities and some non-essential spending, signaling moderate economic comfort.
		- Cluster 3 (Green 🟢) captures the most financially constrained households, predominantly Low-Income, with some Lower-Middle Income and the only cluster containing the Poor class. Expenditure is focused heavily on essentials, with food and non-food spending nearly balanced, signaling limited financial flexibility and survival-level consumption behavior.

[PAYUMO]
Calculated and visualized the Silhouette scores to compare the cluster effectiveness between the two generated clustering groups (k=4 and k=3)
 	
Evaluate and visualize Silhouette scores to compare clustering effectiveness between k=4 and k=3 groups.
	def compute_silhouette(fies_data, cluster_col):
	  # Compute average silhouette score and sample scores
	  score = silhouette_score(fies_data[['TOINC_scaled', 'TOTEX_scaled']], fies_data[cluster_col])
	  silhouette_vals = silhouette_samples(fies_data[['TOINC_scaled', 'TOTEX_scaled']], fies_data[cluster_col])
	  return score, silhouette_vals
	
	fig, axes = plt.subplots(1, 2, figsize=(16, 6), sharey=True)
	
	for idx, (cluster_col, title) in enumerate([
	  ('Cluster_K4', 'Hierarchical Clustering (K=4)'),
	  ('Cluster_K3', 'Hierarchical Clustering (K=3)')
	]):
	  score, silhouette_vals = compute_silhouette(fies_normalized, cluster_col)
	  print(f'Silhouette Score for {cluster_col}: {score:.3f}')
	
	  y_lower = 10
	  ax = axes[idx]
	  for i in np.unique(fies_normalized[cluster_col]):
	  ith_cluster_vals = silhouette_vals[fies_normalized[cluster_col] == i]
	  ith_cluster_vals.sort()
	  size_cluster_i = ith_cluster_vals.shape[0]
	  y_upper = y_lower + size_cluster_i
	
	  color = plt.cm.tab10(i / 10)
	  ax.fill_betweenx(np.arange(y_lower, y_upper), 0, ith_cluster_vals,
	  facecolor=color, edgecolor=color, alpha=0.7)
	
	  y_lower = y_upper + 10 # Spacing between clusters
	
	  ax.axvline(x=score, color="red", linestyle="--", label=f"Avg Score: {score:.3f}")
	  ax.set_title(title)
	  ax.set_xlabel('Silhouette Coefficient')
	  if idx == 0:
	  ax.set_ylabel('Cluster')
	  ax.legend()
	
	plt.tight_layout()
	plt.show()
		
		- To further validate the clustering quality, a silhouette analysis was performed. The silhouette score measures how well each household fits within its assigned cluster relative to neighboring clusters, with higher scores indicating better-defined clusters.
		- Results show that the hierarchical clustering with K=4 yields an average silhouette score of 0.406, while K=3 achieves a slightly higher score of 0.441. This suggests that the K=3 clustering structure provides better overall cohesion and separation compared to K=4.
		- Both clustering solutions are acceptable, especially considering the complexity and natural overlaps inherent in socioeconomic data. Some data points in both scenarios have negative silhouette scores, indicating they are potentially misclassified or lie near cluster boundaries. However, these cases are minimal, and the majority of households have positive silhouette values.

		- The silhouette plots also reveal imbalanced cluster sizes, with certain clusters dominating in both K=3 and K=4 solutions. This is expected in socioeconomic datasets where wealth distributions are typically skewed. - Overall, K=3 appears slightly more effective based on the silhouette score and plot shape, offering more compact and well-separated clusters. However, K=4 may still be useful if the added segmentation aligns better with analytical or policy goals (e.g., distinguishing a smaller middle-income segment).
7. Data Mining: Association Rule Mining (20 pts)		
[ENTRATA] If needed, transform the dataset (one-hot encoding) and apply the Apriori algorithm to extract association rules.	poor_encoded = pd.get_dummies(df_poor) low_encoded = pd.get_dummies(df_low) lower_middle_encoded = pd.get_dummies(df_lower_middle) middle_encoded = pd.get_dummies(df_middle) MIN_SUPPORT = 0.3 MIN_THRESHOLD = 0.7	

[ENTRATA]
 	
This will create a plot to visualize directed acyclic graphs to process association rule data.
	def plot_dag(rules, title=None):
	G = nx.DiGraph()
	
	for _, row in rules.iterrows():
	G.add_edge(row["antecedents"], row["consequents"], weight=row["confidence"], lift=row["lift"])
	
	pos = nx.spring_layout(G, seed=42, k=3.0)
	
	plt.figure(figsize=(14, 8))
	
	edges, colors = [], []
	for u, v, d in G.edges(data=True):
	edges.append((u, v))
	lift = d["lift"]
	
	if lift > 1:
	colors.append("green")
	elif lift == 1:
	colors.append("gray")
	else:
	colors.append("red")
	
	nx.draw_networkx_nodes(
	G, pos, node_color="lightblue", node_size=5000, edgecolors="black"
	)
	
	nx.draw_networkx_edges(
	G, pos, edgelist=edges, edge_color=colors, width=2,
	arrows=True, arrowsize=50, connectionstyle="arc3,rad=0.2",
	min_target_margin=40
	)
	
	edge_labels = {(u, v): f"Lift: {d['lift']:.2f}" for u, v, d in G.edges(data=True)}
	nx.draw_networkx_edge_labels(
	G, pos, edge_labels=edge_labels, font_size=10, font_color="black", label_pos=0.4
	)
	
	labels = {node: str(node).replace(", ", "\n") for node in G.nodes()}
	nx.draw_networkx_labels(
	G, pos, labels, font_size=12, font_weight="bold",
	bbox=dict(facecolor="white", edgecolor="black", boxstyle="round,pad=0.3")
	)
		# Method for Visualization

	plt.title(f"{title} Decision Flow: Antecedent → Consequent", fontsize=14, fontweight="bold") plt.margins(0.2) plt.show()	
[ENTRATA] Association Rule Mining for Poor Households This will map the expenditure priorities of poor households, focusing on food and non-food spending.	plot_dag(no_dups_poor.nlargest(5, 'lift'), 'Expenditures (Poor)')	 **Insights** - Poor households allocate a moderate budget to both food and non-food expenses. This suggests that while food remains a necessity, some households still attempt to balance spending across essentials like utilities, transportation, and basic services. - Households that spend highly on food, often do not allocate spending to other discretionary expenditures like insurance or additional expenditures. This indicates that food remains the highest priority for poor households, potentially limiting spending on other essentials. - Households with low total expenditure tend to still prioritize food. However, this is often at the cost of other necessities, leaving little room for other essential expenses (non-food). **Recommendations** - There should be food security programs for poor households in NCR, so they do not have to compromise essential non-food expenses.
[ENTRATA] Association Rule Mining for Low-income Households This will analyze the spending behavior of low-income households with a focus on food and non-food expenditures.	plot_dag(no_dups_low.nlargest(5, 'lift'), 'Expenditures (Low income)')	 **Insights** - Low-income households tend to have moderate spending on food. This differs from poor-income households, which allocated a high percentage of their budget in food expenditures. - Low-income households exhibit a mix of both food and non-food expenditures. This suggests that low-income households may have slightly more flexibility in their spending but still prioritize food. - Households that spend high on food often decrease their non-food expenditures. That is why when needed, non-food spending is the first to be reduced in their budget.

[ENTRATA] Association Rule Mining for Lower-middle-income Households This will analyze the spending behavior of lower-middle-income households with a focus on food and non-food expenditures.	plot_dag(no_dups_lower_middle.nlargest(5, 'lift'), 'Expenditures (Lower-middle income)')	 **Insights** - Households with moderate spending on bread, meat, and fish are linked to moderate total food and non-food spending. This indicates that staples like bread and protein sources, such as meat and fish, remain core parts of their diet. - The presence of fish as a significant expenditure suggests a more diversified diet compared to lower-income households, which prioritize more affordable food items. This suggests that lower-middle income families are able to allocate more towards varied nutrition. - Unlike poor and low-income households, lower-middle-income families allocate a moderate amount to non-food expenses. This suggests a more balanced budget where essential non-food expenses like utilities, education, and transportation are not sacrificed in favor of food.
[ENTRATA] Association Rule Mining for Middle Income Households This will analyze the spending behavior of middle-income households with a focus on food and non-food expenditures.	plot_dag(no_dups_middle.nlargest(5, 'lift'), 'Expenditures (Middle income)')	 **Insights** - Households with low bread consumption are associated with low total food spending. This suggests that bread may not be a staple food for some middle-income families, possibly due to dietary preferences or a shift towards other food categories. - Households with low food spending tend to have high non-food expenses. This indicates that some middle-income families prioritize non-food necessities such as transportation, healthcare, or housing over food. - Households that spend less on food at home tend to allocate more to non-food necessities. This could imply a preference for dining out or convenience food, as well as a shift in budgeting where food is sacrificed for other essential expenses.

III. Data Analysis & Insights

A. Expenditure Analysis

a. Overall Spending Priorities by Income

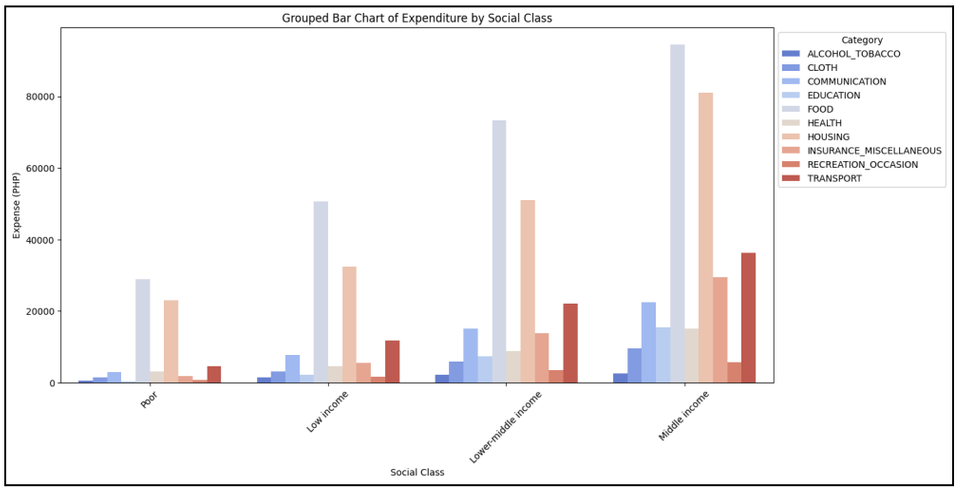


Figure 3.1: Overall Spending Priorities by Income

The household spending patterns in NCR vary depending on their social class. Lower-income families allocate a huge portion of their resources to basic necessities like food and housing, as shown in *Figure 3.1*, whereas higher-income families allocate a smaller share to essentials and more to education, health, insurance and other expenditures. For example, poor households spend a huge portion of their non-food budget on housing and utilities (about 68.57%), as shown in *Figure 3.2*, and a very high share of their total expenses on food (about 46.02%), as shown in *Figure 3.3*, leaving little room for other necessities. In contrast, middle-income households spend a lower percentage on housing as they have more disposable income for other expenses, such as insurance, education, health, and recreation.

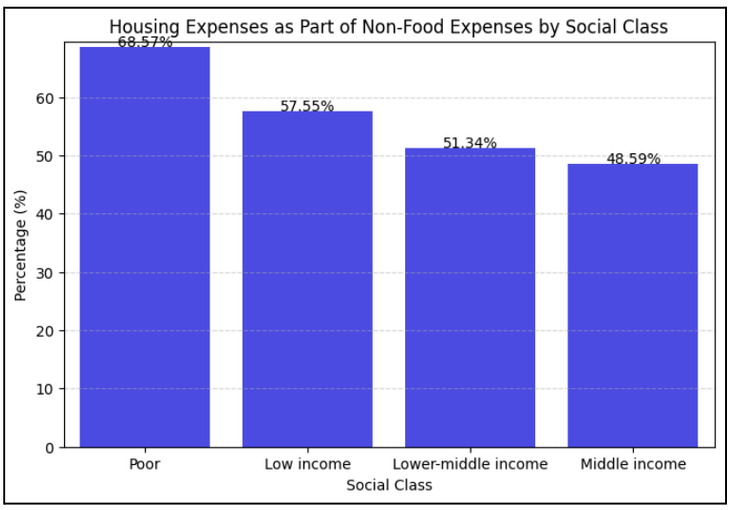


Figure 3.2: Housing Expenses as Part of Non-Food Expenses by Social Class

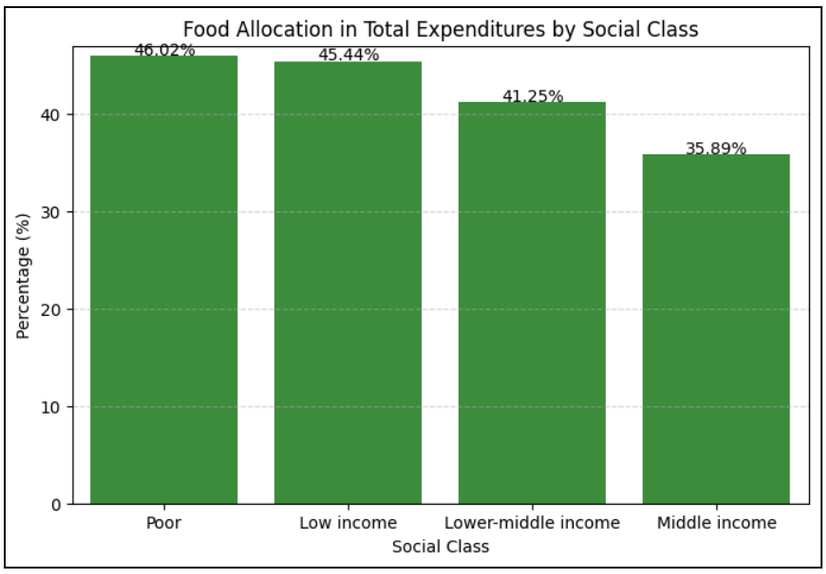


Figure 3.3: Food Allocation in Total Expenditures by Social Class

b. Food Expenditure

All social classes prioritize food, but they have different compositions of food spending. Poor and low-income households spend around 20% of their food budget on outside meals, while lower-middle and middle-income groups allocate a relatively higher percentage of food spending (about 25%) to dining out, as shown in *Figure 3.4*. Staples like bread, meat, and fish are the top food products consumed by all classes, but only the higher three groups report spending on non-essential beverages like tea, cocoa, and other non-alcoholic beverages, as shown in *Figure 3.5*. This suggests that as income rises, households seek more variety and preferences in their diet.

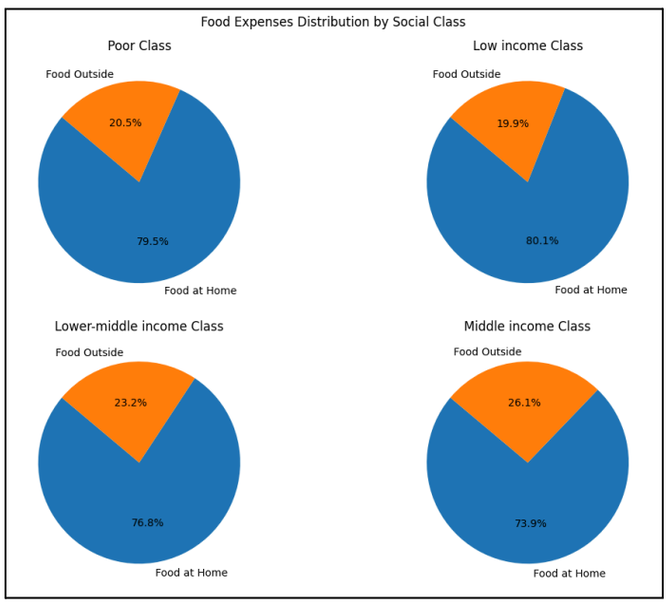


Figure 3.4: Food Expenses Distribution by Social Class

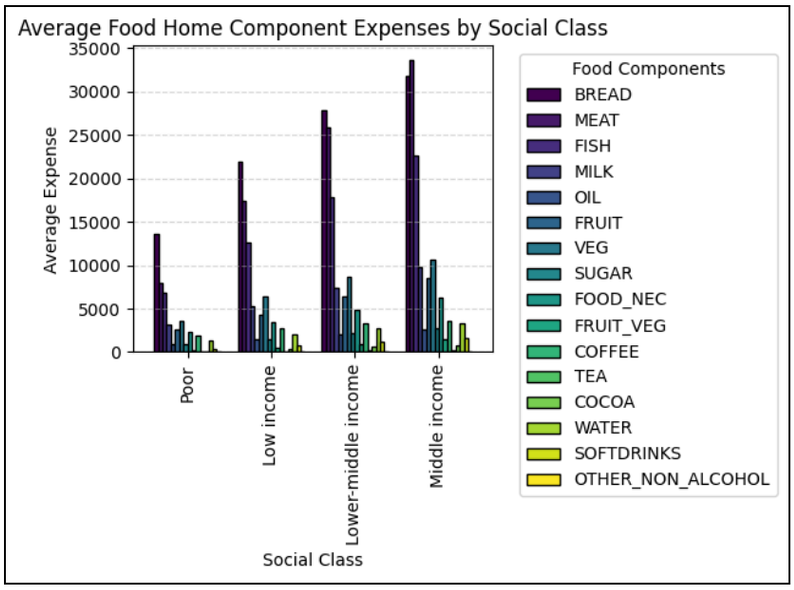


Figure 3.5: Average Food Home Component Expenses by Social Class

c. Education Expenditure

There is an alarming disparity in education spending. The poorest families allocate almost nothing of their expenses to education, as shown in *Figure 3.6*, and many low-income and lower-middle households report zero education spending, as shown in *Figure 3.7*. This non–spending may likely be due to relying on free public schooling or foregoing higher education. In contrast, middle-income and many lower-middle families are able to spend significantly on education, as shown in *Figure 3.8*, even affording private schools or college tuition.

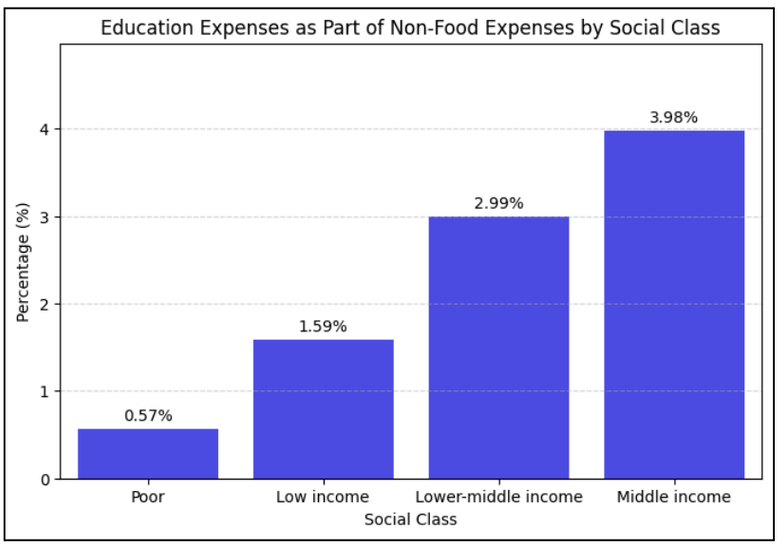


Figure 3.6: Education Expenses as Part of Non-Food Expenses by Social Class

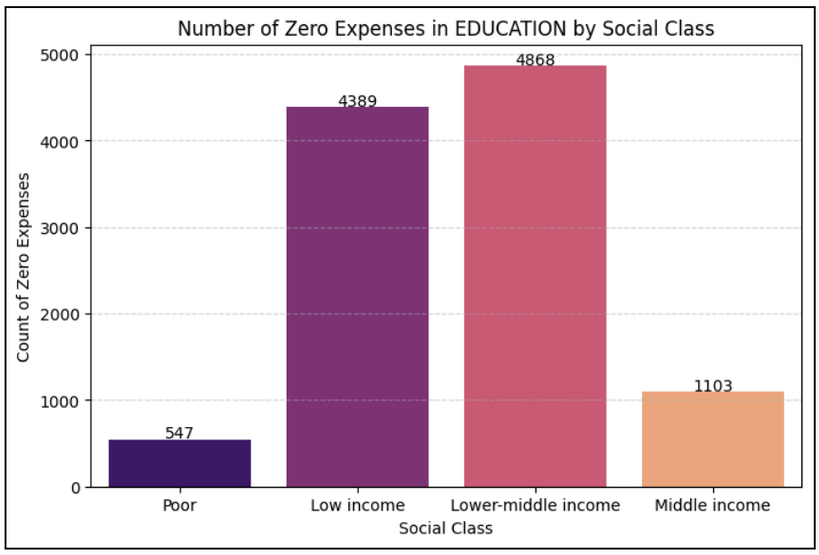


Figure 3.7: Number of Zero Expenses in Education by Social Class

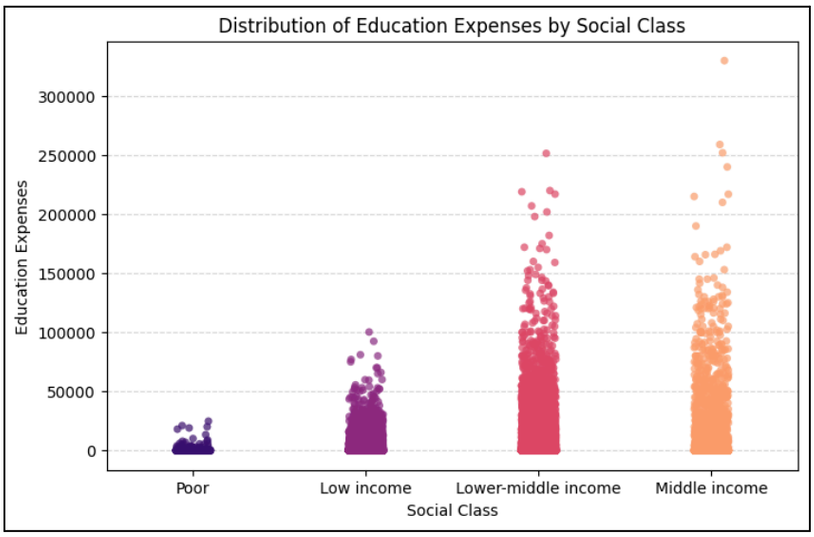


Figure 3.8: Distribution of Education Expenses by Social Class

d. Health expenditure

Health budgets rise with income, wherein wealthier households spend far more on health care, like clinic visits and medicine, than poorer households, as shown in *Figure 3.9*. Middle-income families appear to prioritize healthcare and can afford private services, while poor and low-income families allocate less to health due to financial constraints. The data suggest an inequality in healthcare access. Lower-income groups likely postpone medical care or rely on overcrowded public health services. In *Figure 3.10*, we can see outliers in health expenditure allocation, possibly due to major medical procedures or chronic illnesses. Unfortunately, the poor households are not capable of getting their needed medical attention.

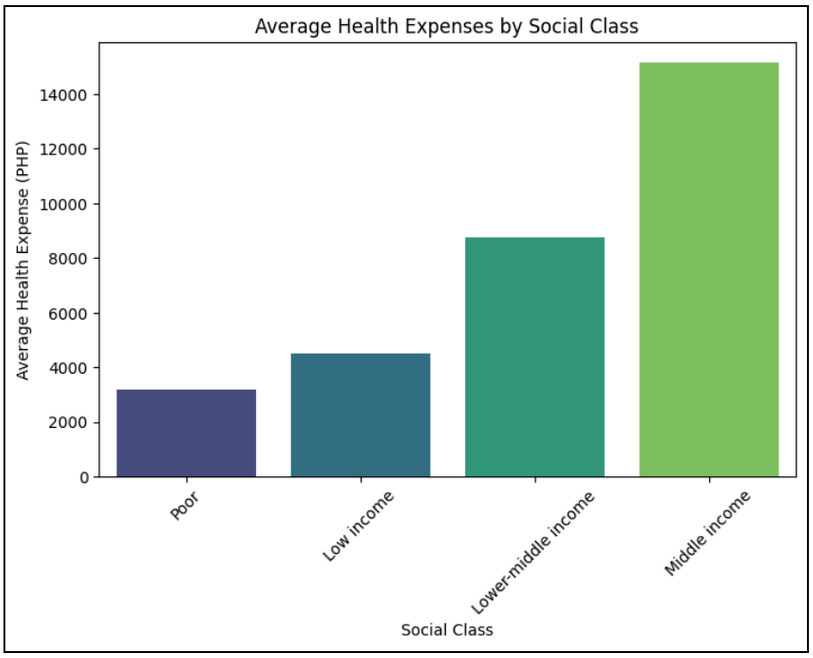


Figure 3.9: Average Health Expenses

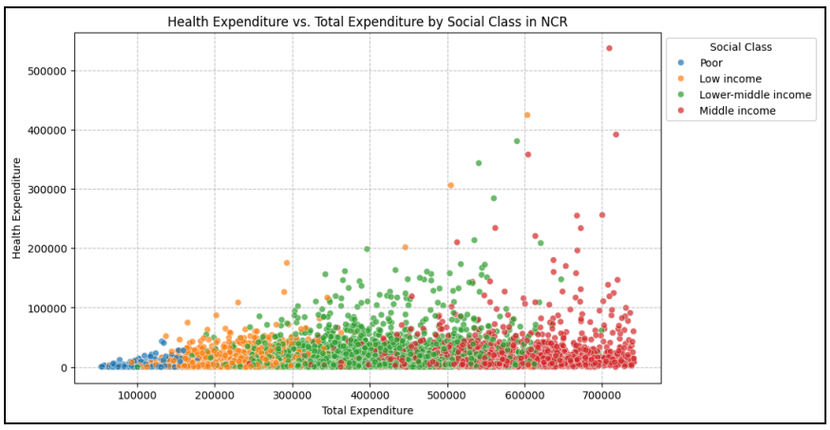


Figure 3.10: Health Expenditure by Social Class in NCR

e. Transportation Expenditure

Transportation expenses increase with income level. As shown in *Figure 3.11*, middle-income households allocate a significantly larger portion of their expenses to transportation, suggesting ownership of private vehicles, longer commutes, and/or use of ride-hailing applications. Meanwhile, poor and low-income households have a smaller transportation budget wherein they rely more on affordable public transportation options or working. However, even with minimal transport expenses, these lower-income groups still face a burden when commuting costs consume their limited income and also their time and resources. Transport costs also correlate with employment opportunities and access to services.

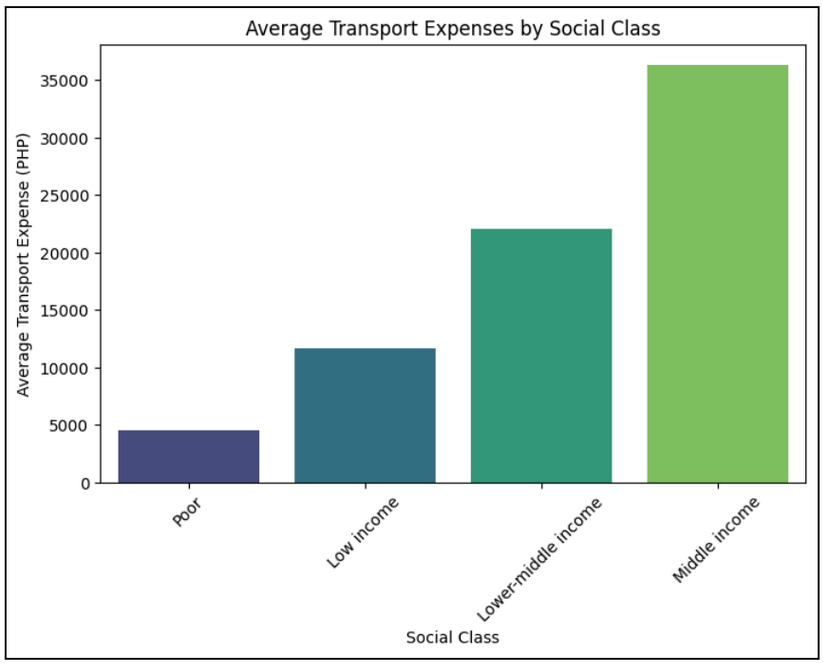


Figure 3.11: Average Transport Expenses by Social Class

f. Recreation and Discretionary Expenditure

As shown in *Figure 3.12*, recreation and occasion spending rise with income. Poor households spend only about 1.9% of their non-food budget on recreation, while middle-income households allocate about 3.3%. This suggests that discretionary speeding remains a luxury for lower classes. Still, all classes spend something on social and cultural activities, even though it is minimal. Recreation improves mental well-being and social participation, and even lower-income families seek ways to celebrate events.

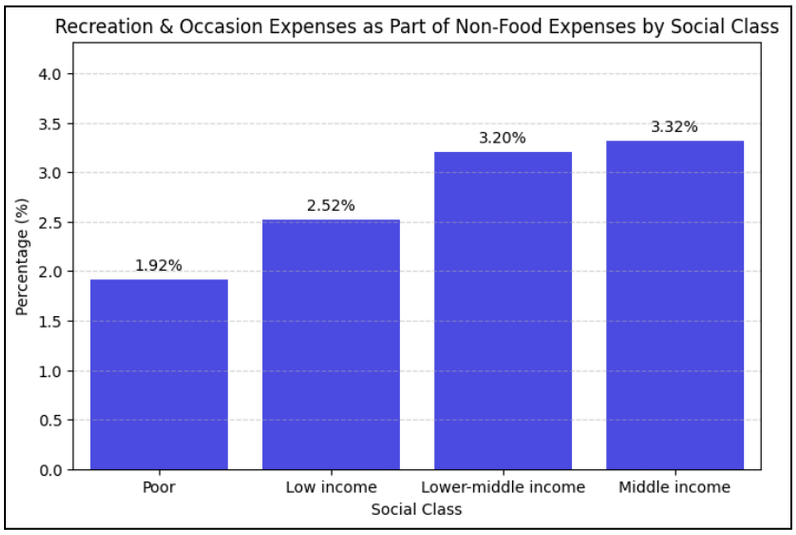


Figure 3.12: Recreation & Occasion Expenses as Part of Non-Food Expenses by Social Class

g. Vices Spending

Figure 3.13 shows that poorer households allocate a higher proportion of their budget to vices like tobacco and alcohol compared to middle-income households. This spending, even though it is relatively small compared to other expenditures, takes up a

larger share of tight budgets in poor and low income families. This reflects existing problems in our community such as potential addiction, coping behaviors, or cultural norms. This reduces funds available for more essential needs like food and education.

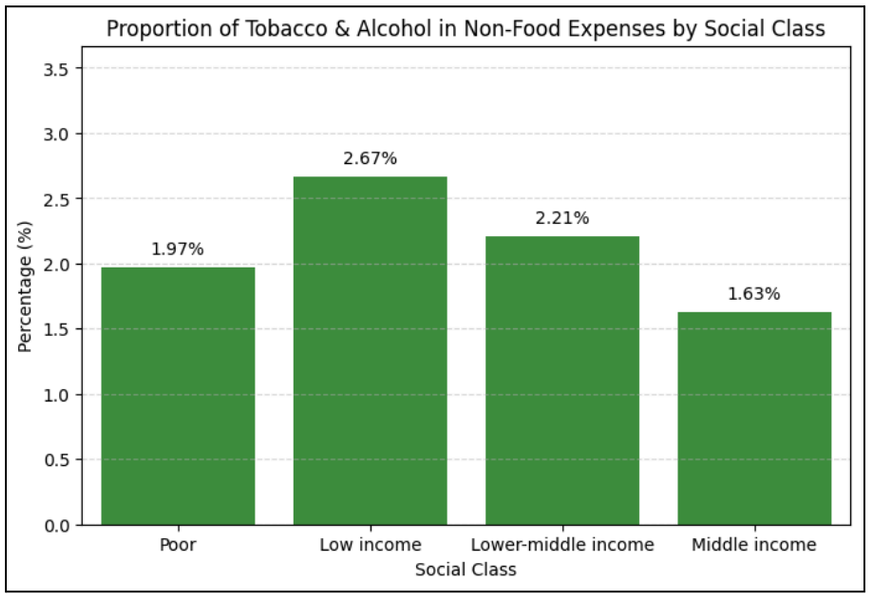


Figure 3.13: Proportion of Tobacco & Alcohol in Non-Food Expenses by Social Class

B. Distance Analysis

This section provides a comprehensive analysis of the financial landscape of NCR households by exploring the relationships between income, expenditures, and key demographic factors. Correlation analysis identifies the strongest contributors to aggregate financial metrics, highlighting the most influential spending drivers. Additionally, a distance matrix is computed to facilitate hierarchical clustering, uncovering hidden patterns among households based on their income and expenditure behaviors.

i. Correlation Analysis

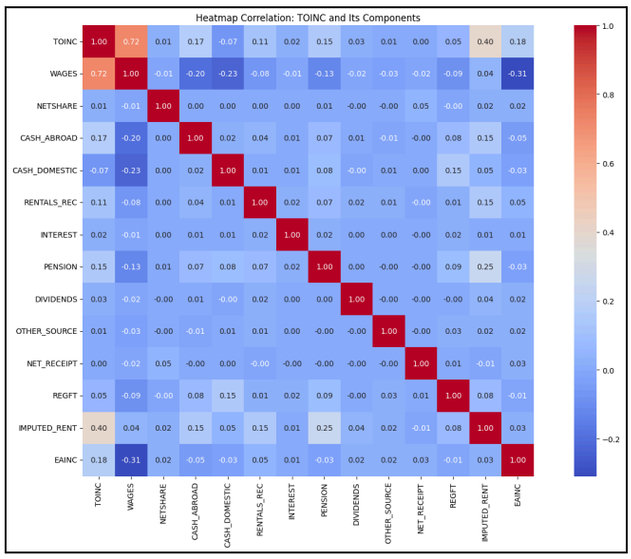


Figure 3.14: Heatmap Correlation between TONIC and its Components

Figure 3.14 illustrates the correlation between Total Income (TOINC) and its components, highlighting employment income (WAGES) as the primary driver (0.72

correlation) for NCR households, reflecting the region's service-driven economy. Homeownership (IMPUTED_RENT) also plays a significant role (0.40 correlation), indicating the financial benefits of owning property in NCR's high-value real estate market. In contrast, entrepreneurial income (EAINC, 0.18) and remittances (CASH_ABROAD, 0.17) have weaker correlations, suggesting that while small businesses and overseas transfers contribute to household earnings, they are secondary compared to wages. These findings underscore NCR's heavy reliance on formal employment, with homeownership providing additional financial stability.

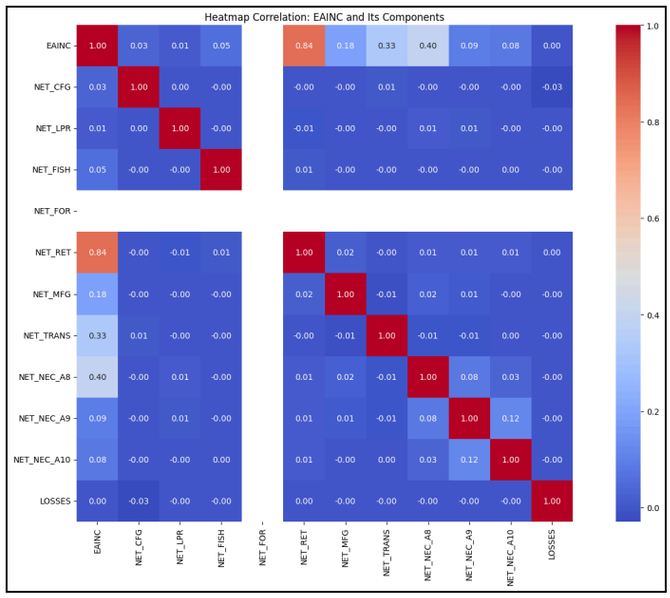


Figure 3.15: Heatmap Correlation between EAINC and its Components

Figure 3.15 presents the correlation between Entrepreneurial Income (EAINC) and its components, revealing that retail and wholesale trade (NET_RET) is the dominant source (0.84 correlation) of entrepreneurial earnings among NCR households. In contrast, forestry-related income (NET_FOR) shows no correlation, as all values are zero—an expected outcome given NCR's highly urbanized environment, where forestry activities are nonexistent. These findings highlight the strong reliance of entrepreneurs in NCR on commercial trade rather than agricultural or natural resource-based ventures.

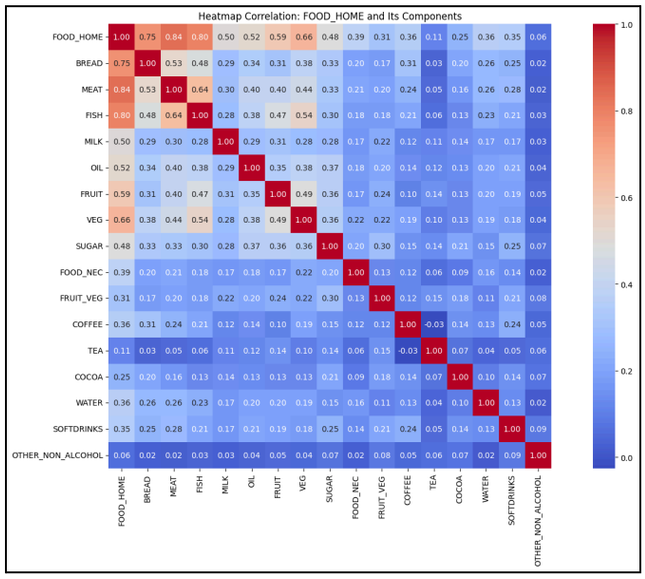


Figure 3.16: Heatmap Correlation between FOOD_HOME and its Components

Figure 3.16 illustrates the correlation between food-at-home expenses (FOOD_HOME) and its components, revealing that meat (0.84), fish (0.80), and bread

(0.75) are the top contributors to household food spending in NCR. Other key items such as vegetables (0.66), fruits (0.59), oil (0.52), milk (0.50), and sugar (0.48) also play a significant role, reflecting a focus on staple and nutrient-rich foods. In contrast, non-essential items like coffee, tea, and cocoa exhibit lower correlations, suggesting that NCR households allocate their food budget primarily toward essential and versatile food products.

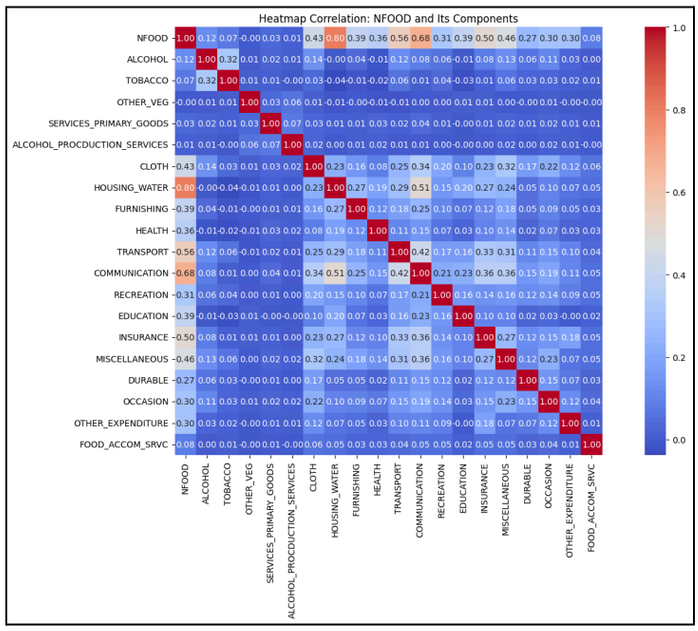


Figure 3.17: Heatmap Correlation between NFOOD and its Components

Figure 3.17 presents the correlation between non-food expenses (NFOOD) and its components, showing that housing & utilities (0.80), communication (0.68), and transport (0.56) are the primary drivers of non-food expenditures in NCR households. Other key contributors include insurance (0.50), miscellaneous expenses (0.46), clothing (0.43), furnishing (0.39), education (0.39), and health (0.36), indicating notable spending on financial security, personal needs, and essential services. These findings suggest that NCR households prioritize housing, connectivity, and mobility in their budgets while maintaining moderate allocations for education, health, and household necessities.

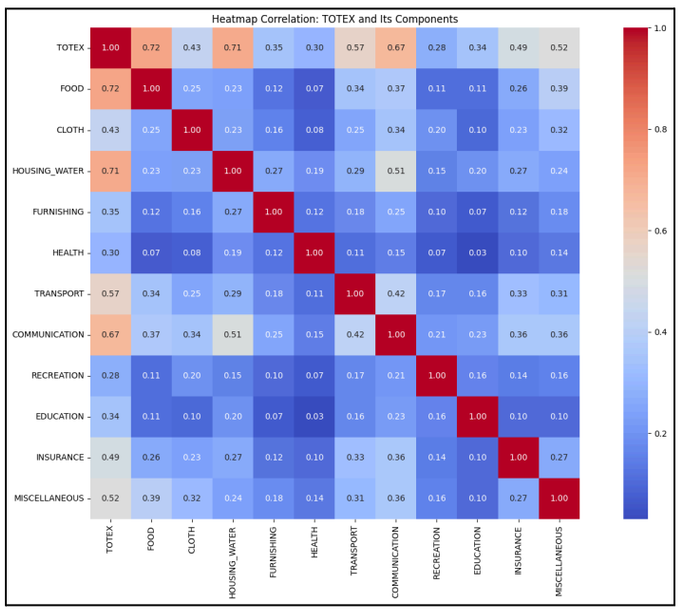


Figure 3.18: Heatmap Correlation between TOTEX and its Components

Figure 3.18 illustrates the correlation between total household expenditures (TOTEX) and its components, revealing that food (0.72) and housing & utilities (0.71) are the dominant spending categories, emphasizing the significant share of basic necessities in household budgets. Housing & utilities encompass essential costs like rent, electricity, water, and fuel, which are particularly high in urban areas like NCR. Other key contributors include communication (0.67), transport (0.57), miscellaneous expenses (0.52), insurance (0.49), and clothing (0.43), indicating that, beyond essentials, households allocate a considerable portion of their budget to connectivity, mobility, financial security, and personal needs.

ii. Distance Matrix Analysis

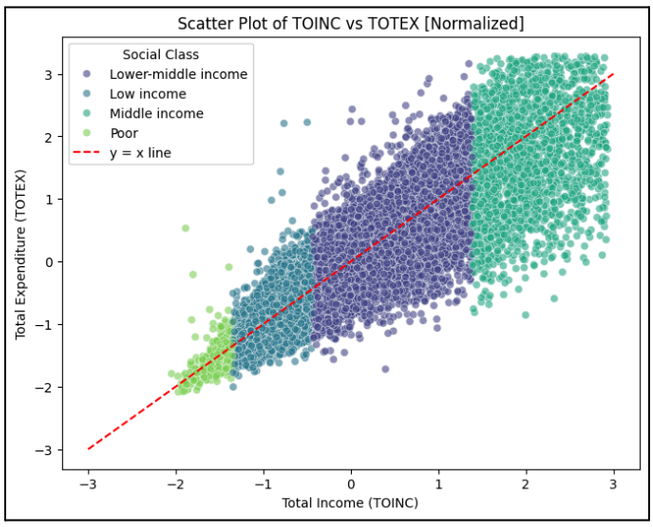


Figure 3.19: Scatter Plot of TOINC vs TOTEX

Figure 3.19 presents a scatter plot of Total Income (TOINC) vs. Total Expenditure (TOTEX), revealing a strong positive linear relationship between income and spending. More households are positioned below the red y = x line, indicating that they generally spend less than they earn. The plot also highlights distinct social class segmentation, with lower-income households clustering in the lower-left region and middle-income groups shifting towards the upper-right. Additionally, outliers above the y = x line suggest that some households spend beyond their income, possibly due to borrowing, debt, or unreported financial support.

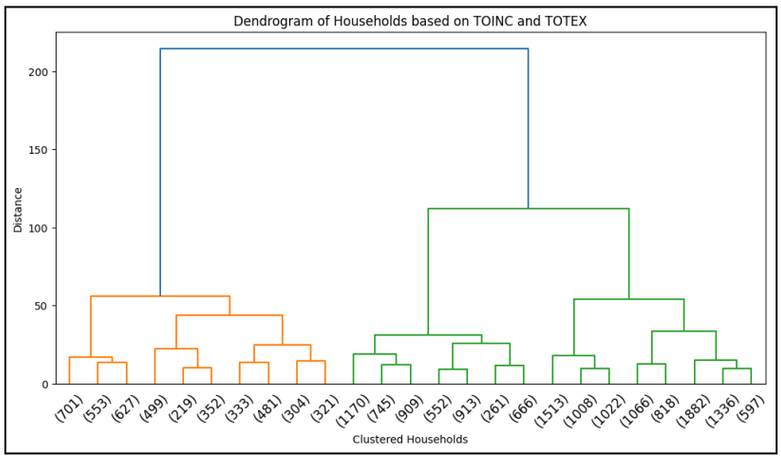


Figure 3.20: Dendogram of Households Based on TOINC vs TOTEX

Figure 3.20 presents a dendrogram of households based on Total Income (TOINC_scaled) and Total Expenditure (TOTEX_scaled), utilizing Ward's method to

minimize within-cluster variance. The hierarchical clustering process reveals two primary clusters at higher linkage distances, representing distinct household groups with similar income-expenditure profiles. These clusters are further subdivided into more homogeneous sub-clusters as the hierarchy progresses. To derive meaningful insights, the dendrogram will be cut at a predefined k number of clusters, with the resulting groups visualized in a scatter plot of TOINC_scaled vs. TOTEX_scaled, color-coded by cluster assignment.

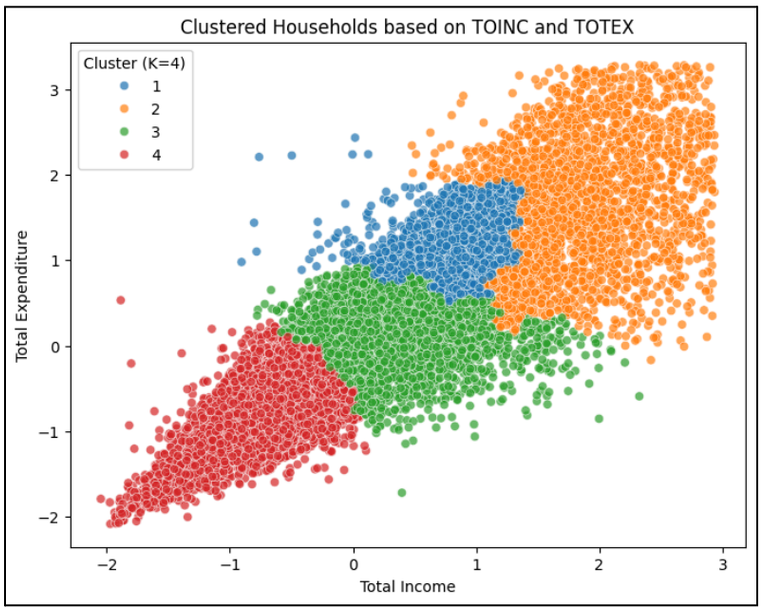


Figure 3.21: Clustered Household based on TOINC and TOTEX (K=4)

Figure 3.21 illustrates the clustered households based on Total Income (TOINC) and Total Expenditure (TOTEX) using K=4 clusters. The clusters align along a diagonal trend, reinforcing the strong positive relationship between income and expenditure, where higher-income households tend to spend more. The clustering effectively captures the economic progression of households:

- **Cluster 4 (Red):** Represents the lowest-income households with minimal expenditure, likely from the poor and low-income classes facing financial constraints.
- **Cluster 3 (Green):** Consists of households with moderate income but controlled spending, likely frugal low- to lower-middle-income families managing expenses efficiently.
- **Cluster 1 (Blue):** Encompasses lower- to middle-income households with balanced spending patterns, possibly supported by stable employment.
- **Cluster 2 (Orange):** Comprises higher-income households with high expenditure, reflecting greater financial flexibility and diversified consumption habits.

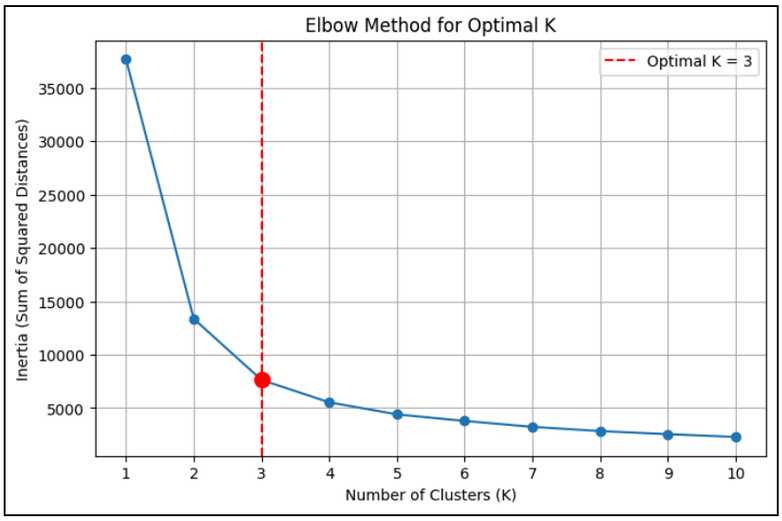


Figure 3.22: Elbow Method for Optimal K

Figure 3.22 presents the Elbow Method plot used to determine the optimal number of clusters (K). The plot suggests that K = 3 is the ideal choice, as it marks the point where the rate of decrease in within-cluster variance slows down. However, the initially chosen K = 4 clustering remains valid, as it provides meaningful socioeconomic insights into household segmentation. To further assess the robustness of the clustering results, hierarchical clustering will be reperformed with K = 3, allowing for a comparison of household groupings and potential refinements in the segmentation approach.

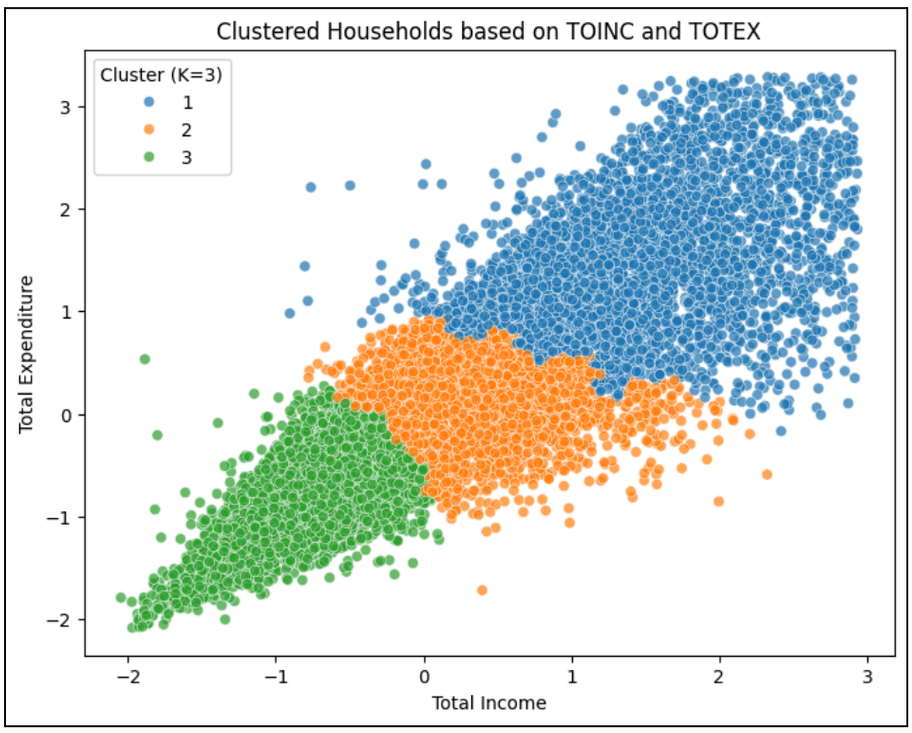


Figure 3.23: Clustered Household based on TOINC and TOTEX (K=3)

Figure 3.23 presents the K = 3 clustering of households based on Total Income (TOINC) and Total Expenditure (TOTEX). Compared to the K = 4 result, the clusters in this plot are more evenly distributed, while still following a clear positive trend—households with higher income generally spend more.

- Cluster 1 (Blue): Represents higher-income households, primarily from the lower-middle to middle-income groups, who exhibit greater financial flexibility and higher expenditures.

- Cluster 2 (Orange): Captures lower-middle income households, characterized by a more balanced income-expenditure pattern, indicating a controlled spending approach.
- Cluster 3 (Green): Represents poor to lower-middle income households, reflecting financially constrained groups with limited income and spending capacity.

The clustering result shows minimal overlap and few outliers, suggesting a clear segmentation of socioeconomic groups based on income and expenditure behavior.

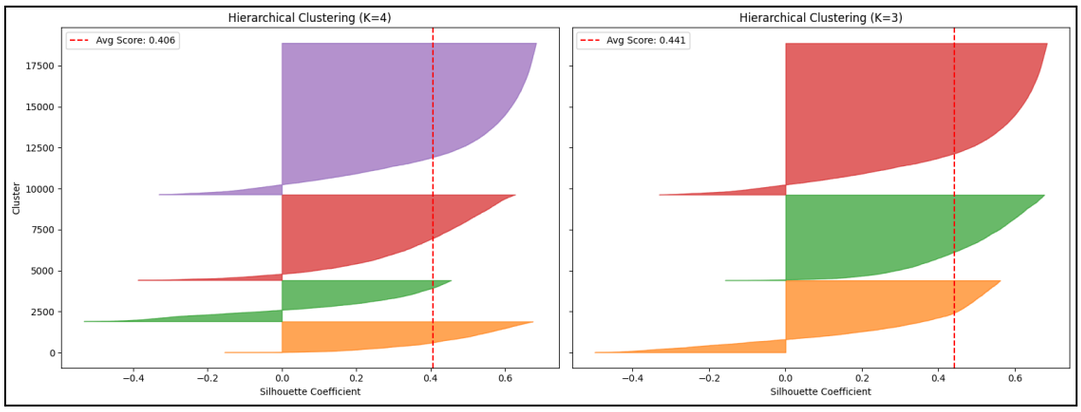


Figure 3.24: Comparison of Silhouette Scores for the Clustering Methods (K=3, K=4)

Figure 3.24 presents the silhouette score comparison for the K = 3 and K = 4 clustering solutions to assess cluster quality. The silhouette score, which measures how well households fit within their assigned clusters, indicates that K = 3 achieves a slightly higher score (0.441) compared to K = 4 (0.406), suggesting better-defined clusters with improved cohesion and separation.

While both clustering solutions are valid, some data points exhibit negative silhouette scores, indicating possible misclassification or proximity to cluster boundaries—a common occurrence in socioeconomic data due to natural overlaps. Additionally, cluster sizes are imbalanced, reflecting the skewed distribution of wealth across households.

Overall, K = 3 appears to provide more compact and distinct clusters, but K = 4 may still be relevant depending on analytical objectives, particularly if a finer distinction between socioeconomic groups is needed.

C. Association Rule Analysis

Association rule mining reveals critical expenditure behaviors and interdependencies across different income groups in NCR. By analyzing decision flows and relationships between expense categories, key financial behaviors were observed that drive household spending. Some normalization is done in this process, such as dropping columns where more than 50% of the rows have a value of zero. Without this normalization, the majority of the results in association rule mining would associate two itemsets with null values, which is not helpful for finding patterns.

i. Poor Households

Food spending dominates in financial decision-making, as evident in the previous figures. Rule mining analysis reveals strong dependencies between

food and non-food allocations, where moderate food spending is highly associated with zero spending on other non-essential categories. Additionally, higher food expenses directly reduce non-food expenditures, due to limited financial flexibility. A household that allocates a high budget to food is 2.11 times more likely to reduce non-food spending.

Interestingly, while earlier findings suggested that poorer households allocate some of their budget to vices, the rule mining results indicate that households that do not spend on alcohol tend to have even lower non-food spending overall.

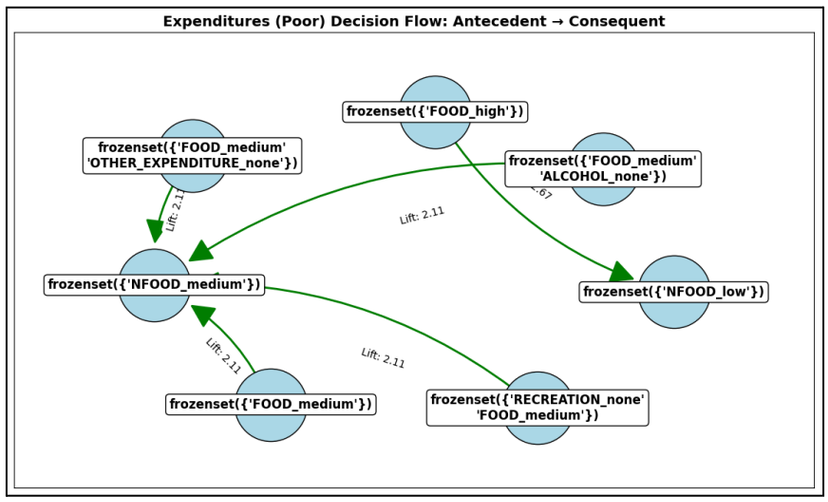


Figure 3.25: DAG Plot for Poor Households

ii. Low-income Households

Low-income households exhibit a more structured budget allocation compared to the poorest group, but food remains their top priority in overall expenditures. Rule mining analysis reveals strong similarities to poor households, particularly in how food spending dictates their financial decisions. However, a key distinction emerges wherein moderate spending on home-prepared meals is closely associated with moderate total food consumption. This indicates a shift towards home-cooked meals, allowing low-income households to maintain food security while managing other essential expenses.

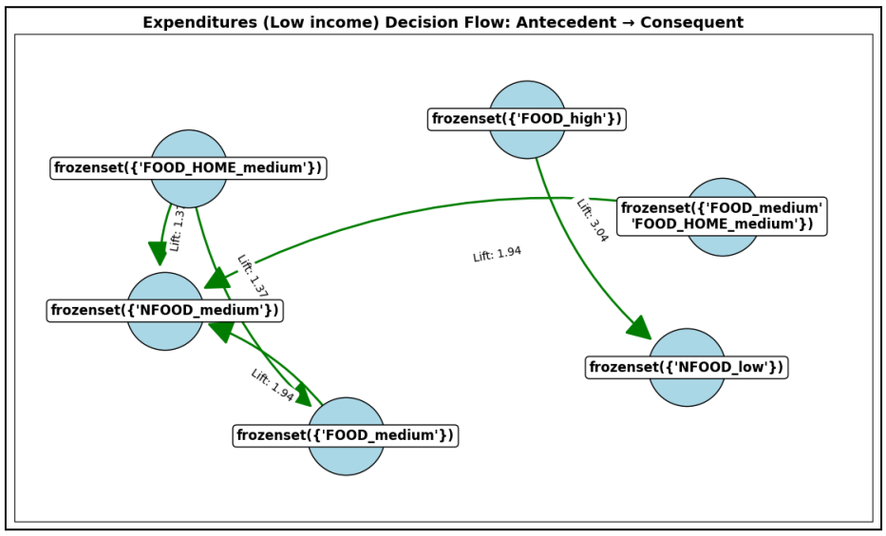


Figure 3.26: DAG Plot for Low Income Households

iii. Lower-Middle Income Households

A noticeable shift in spending patterns rises among lower-middle income households, where food diversity increases and allocation to specific food items like fish meat, and bread. Rule mining analysis reveals that households with medium food expenditures are strongly associated with purchasing fish, meat, and bread and they are 1.96x more likely to have a moderate spending on non-food essentials.

Additionally, non-food expenditures align more closely with total spending capacity, meaning that as household income increases, there is a balanced allocation toward both food and essential non-food categories. This financial balance suggests an improvement in overall economic stability, allowing them to invest in other essentials.

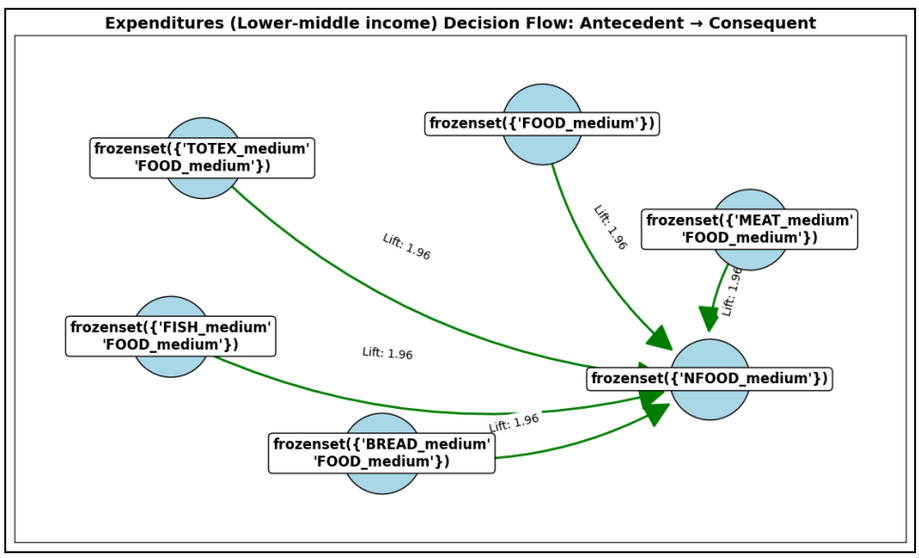


Figure 3.27: DAG Plot for Lower Middle Income Households

iv. Middle Income Households

Middle-income households exhibit a more flexible and discretionary spending pattern, where food is still a significant expense. However, home-prepared meals are deprioritized in favor of dining out. Rule mining analysis indicates that households with low home food spending are associated with higher non-food expenditures, confirming a lifestyle shift towards convenience and luxury to try out other things. As income rises, food takes up a smaller fraction of the budget, freeing funds for entertainment, travel, and leisure spending.

This trend highlights the economic confidence and a shift toward experience-based spending, where middle-income households allocate resources toward recreational activities, services, and other non-food essentials rather than increasing basic food expenses. Unlike lower-income groups, they have greater financial flexibility.

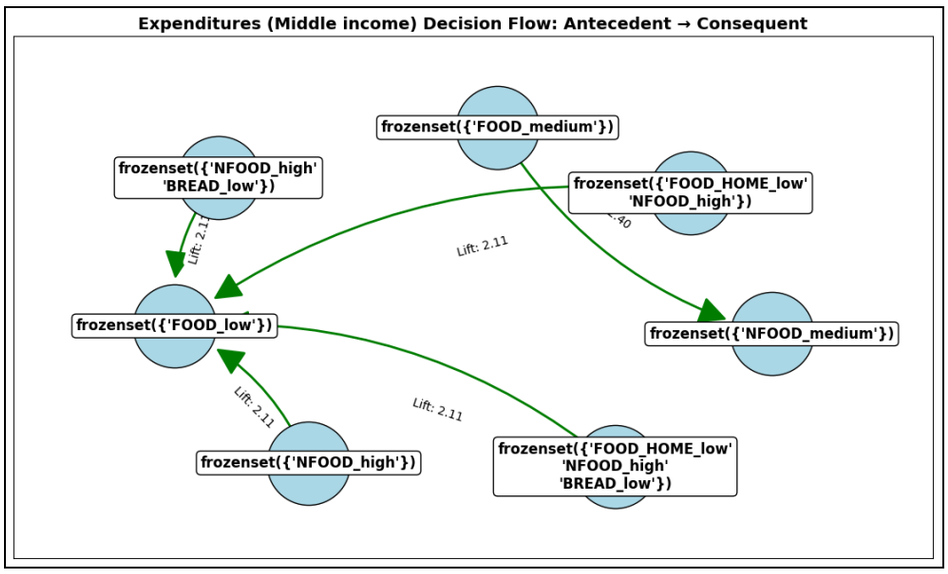


Figure 3.28: DAG Plot for Middle Income Households

D. Actionable Insights & Recommendations

Food remains the primary expenditure for all income groups, but lower-income households allocate the highest portion of their budget to it, often sacrificing other essential needs such as education and healthcare. Expanding food assistance programs, such as subsidized staple foods and feeding programs in urban poor areas, will help relieve financial pressure on these families. Additionally, ensuring access to affordable and nutritious food can reduce long-term health risks, as lower-income groups often struggle to allocate a budget for healthcare expenses.

While existing programs like the Pantawid Pamilyang Pilipino Program (4Ps) and the National Food Authority (NFA) Rice Program aim to provide financial and food assistance, challenges remain. One critical issue is food waste, as the National Capital Region (NCR) generates approximately 2,000 tons of food waste daily (Cos, 2022). The Food Donation Act of 2009 (RA 9803) encourages food donations but remains voluntary, limiting its impact.

To address both food insecurity and food waste, we propose the implementation of a **government-led surplus food redistribution system**. This initiative would require restaurants, hotels, and supermarkets to donate surplus but safe food to food banks and feeding programs managed by the government and accredited non-profit organizations. By creating a structured partnership between businesses and social welfare agencies, this system would ensure that excess food is efficiently redirected to communities in need instead of being discarded. Similar policies have been successfully implemented in countries like France, where the 2016 Food Waste Law (Loi Garot) mandates large supermarkets to donate unsold but still consumable food. Adopting a similar approach in the Philippines would not only strengthen food security for vulnerable households but also promote sustainable food management practices across the country.

IV. Conclusion

The analysis of NCR household expenditures reveals distinct financial segmentation, with spending behaviors closely tied to income levels. Lower-income households allocate the highest portion of their budget to food, often at the expense of other essential needs such as education and healthcare, while middle-income groups exhibit greater financial flexibility, allocating more towards discretionary and long-term investments.

Key findings emphasize the strong dependence on wages as the primary income source, the significant impact of food expenses on financial decisions, and the disparities in education and healthcare spending. Hierarchical clustering and rule mining further highlight how financial constraints dictate spending behavior—lower-income households remain focused on necessities, whereas middle-income households diversify their expenditures. Additionally, the prevalence of food insecurity alongside high food waste underscores inefficiencies in resource distribution.

To address these challenges, policy recommendations focus on expanding food assistance programs, improving access to education and healthcare, and enhancing public transportation infrastructure. A major proposal is the implementation of a government-led surplus food redistribution system, which would mandate businesses to donate surplus yet safe food to food banks and feeding programs, reducing both food insecurity and waste. Strengthening financial literacy and public health initiatives can further promote sustainable spending habits among households.

By aligning policies with the specific needs of each income group, a more inclusive and resilient economic environment can be cultivated—one where financial constraints are alleviated, essential services are accessible, and sustainable practices support long-term economic stability in NCR.

V. References

Cos, W. (2024, October 29). *Food wasted by the tons while millions of Filipinos go hungry*. ABS-CBN News. https://www.abs-cbn.com/news/10/04/22/tons-of-food-wasted-as-millions-of-filipinos-go-hungry

Department of Trade and Industry. (2022, October 12). *NCR - Regional Profile | Department of Trade and Industry Philippines*. Department of Trade and Industry Philippines. https://www.dti.gov.ph/regions/ncr/profile/

Mapa, C. (2024, August 15). *11 out of 18 Regions Recorded Significant Decreases in Poverty Incidence in 2023*. Philippine Statistics Authority. https://psa.gov.ph/statistics/poverty

Philippine Statistics Authority. (n.d.). *Family Income and Expenditure Survey*. PSADA. https://psada.psa.gov.ph/catalog/FIES/about