**NAME:-** SHIWANI KUMARI

**REGISTRATION:-** R24081010000075

**AIM:-** EXPENSE PATTERN| INCOME VS EXPENSE|FAMILY-AGE-AREA

**MINIMUM REQUIREMENTS:-**

**HARDWARE:-**

* RAM-8GB
* Processor- Intel Core i5 (10th Gen or later) / AMD Ryzen 5
* Storage- 512GB SSD or more
* Stable internet connection (for downloading and setting up required software)

**SOFTWARE:-**

PROGRAMMING LANGUAGE & IDEs

* Either VSC with Jupyter extension or Jupyter Notebook(Anaconda Distribution)
* Python 3.x(10,11 etc)
* Python’s libraries(Numpy, pandas, seaborn, Matplotlib etc.)

OPERATING SYSTEM

* Windows 10
* For MacBook - macOS

**STEPS:-**

1. **Download the Dataset**

Download the dataset from Kaggle:[**https://www.kaggle.com/datasets/aradhanahirapara/expense-pattern-income-vs-expensefamilyagearea**](https://www.kaggle.com/datasets/aradhanahirapara/expense-pattern-income-vs-expensefamilyagearea)

1. **Install Required Libraries**

Before running the code, install the necessary Python libraries. Open Command Prompt (Windows) / Terminal (Mac/Linux) and type:

pip install pandas/ pip install numpy / pip install matplotlib/ pip install seaborn/ pip install jupyter

1. **Open Jupyter Notebook**

To open Jupyter Notebook, run the following command in the terminal or command prompt:

1. **Create a New Notebook**
2. Click on "File" → “New File"→ “Jupyter Notebook.ipynb support" to create a new notebook.
3. Save the notebook by clicking ctrl+s (e.g., "Expense\_Pattern\_Analysis").
4. **Import Necessary Modules**

Now, import the required libraries for data handling, visualization, and analysis.

import pandas as pd # For data manipulation

import numpy as np # For numerical computations

import matplotlib.pyplot as plt # For basic visualization

1. **Load the Dataset**

If your dataset is in CSV format, load it using pandas.

# Load dataset (Ensure the dataset is in the same directory or provide full path)

df = pd.read\_csv("income\_expense\_data.csv")

# Display first 5 rows

df.head()

**After completing these steps, write the code according to the question and execute it.**

**Q.1: What is the total household income per year?**

**Ans**: import pandas as pd

# Load dataset

df = pd.read\_csv("income\_expense\_data.csv")

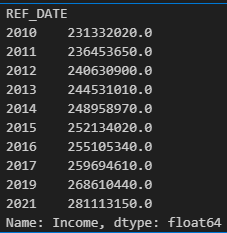
# Group by year and sum income

total\_income\_per\_year = df.groupby("REF\_DATE")["Income"].sum()

# Display results

print(total\_income\_per\_year)

**OUTPUT: This will give you a trend of how total household income has changed from 2010 to 2021**

****

**Q.2: What is the average expense per province?**

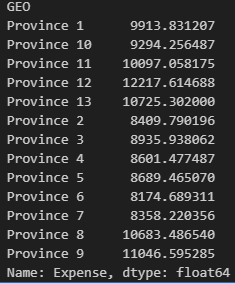
**Ans**: # Group by province and calculate the average expense

avg\_expense\_per\_province = df.groupby("GEO")["Expense"].mean()

# Display results

print(avg\_expense\_per\_province)

**OUTPUT:- This tells us which provinces have higher or lower average household expenses.**

****

**Q.3: What is the trend of income vs. expenses over the years?**

**Ans:** import matplotlib.pyplot as plt

# Group by year and sum income & expense

yearly\_data = df.groupby("REF\_DATE")[["Income","Expense"]].sum()

# Plot the data

plt.figure(figsize=(10,5))

plt.plot(yearly\_data.index, yearly\_data["Income"], marker='o', linestyle='-', label="Total Income",color='green')

plt.plot(yearly\_data.index, yearly\_data["Expense"], marker='o',linestyle='-', label="Total Expense", color='red')

# Labels and title

plt.xlabel("Year")

plt.ylabel("Amount(in units)")

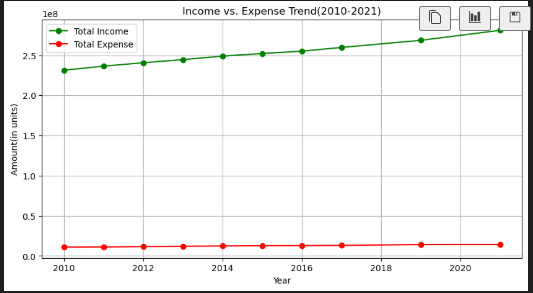
plt.title("Income vs. Expense Trend(2010-2021)")

plt.legend()

plt.grid(True)

plt.show()

**OUTPUT:- This chart helps identify whether expenses are growing faster than income over time.**

****

**Q.4: Which family type has the highest average income?**

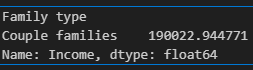
**Ans:** # Group by family type and calculate the average income

avg\_income\_family = df.groupby("Family type")["Income"].mean()

# Display results

print(avg\_income\_family.sort\_values(ascending=False))

**OUTPUT:- This helps in understanding how income distribution varies among different family structures.**

****

**Q.5: How does age affect household expenditure?**

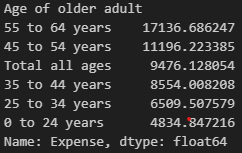
**Ans:** # Group by Age of older adult and calculate the average expense

avg\_expense\_by\_age = df.groupby("Age of older adult")["Expense"].mean()

# Display result

print(avg\_expense\_by\_age.sort\_values(ascending=False))

**OUTPUT:- This will show which age groups tend to spend more or less.**

****

**Q.6: What is the correlation between income and expenses?**

**Ans:** import numpy as np

# Compute correlation

correlation = np.corrcoef(df["Income"], df["Expense"])[0,1]

# Display result

print("Correlation between Income and Expense:",correlation)

**OUTPUT:- • If correlation ≈ 1, higher income leads to higher expenses.**

**• If correlation ≈ 0, there’s no direct relationship.**

**• If correlation ≈ -1, higher income leads to lower expenses (unlikely).**

****

**Q.7: Which income quintile spends the most?**

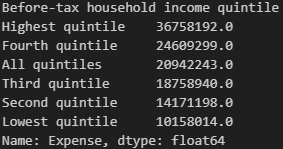
**Ans:** # Group by income quintile and sum expenses

expense\_by\_quintile = df.groupby("Before-tax household income quintile")["Expense"].sum()

# Display results

print(expense\_by\_quintile.sort\_values(ascending=False))

**OUTPUT:- Higher-income groups tend to spend more, but the lowest income groups may also spend a high proportion of their income.**

****

**Q8. Which provinces have the highest income-to-expense ratio?**

**Ans:** # Group by province and sum income and expenses

income\_expense\_ratio = df.groupby("GEO")[["Income","Expense"]].sum()

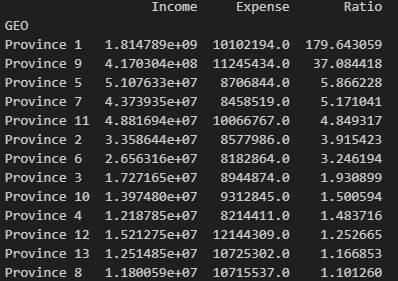
# Calculate the ratio

income\_expense\_ratio["Ratio"] = income\_expense\_ratio["Income"]/ income\_expense\_ratio["Expense"]

# Display results

print(income\_expense\_ratio.sort\_values(by="Ratio",ascending=False))

**OUTPUT:- • Provinces with a high ratio = More savings • Provinces with a low ratio = More financial stress**



**Q.** **9: How does expense distribution look?**

**Ans:** # Plot histogram

plt.figure(figsize=(8,5))

plt.hist(df["Expense"], bins=20, color='blue',alpha=0.7,edgecolor='black')

# Labels and title

plt.xlabel("Household Expense")

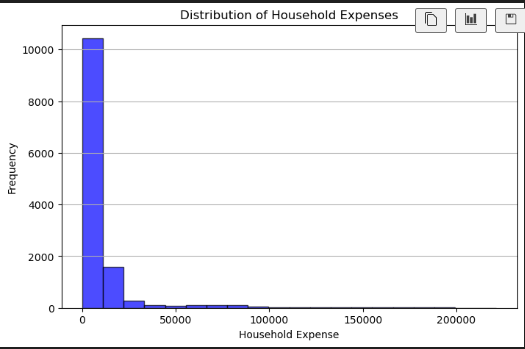
plt.ylabel("Frequency")

plt.title("Distribution of Household Expenses")

plt.grid(axis='y')

plt.show()

**OUTPUT:- This helps us understand if expenses are normally distributed or skewed.**

****

**Q.** **10: What is the trend of income inequality?**

**Ans:** # Filter data for highest & lowest income quintiles

high\_income = df[df["Before-tax household income quintile"] == "Highest"]

low\_income = df[df["Before-tax household income quintile"] == "Lowest"]

# Group by year and sum income

high\_income\_trend = high\_income.groupby("REF\_DATE")["Income"].sum()

low\_income\_trend = low\_income.groupby("REF\_DATE")["Income"].sum()

# Plot the data

plt.figure(figsize=(10,5))

plt.plot(high\_income\_trend.index, high\_income\_trend, marker='o',linestyle='-', label="Highest Income Quintile",color='blue')

plt.plot(low\_income\_trend.index, low\_income\_trend, marker='o',linestyle='-', label="Lowest Income Quintile",color='red')

# Labels and title

plt.xlabel("Year")

plt.ylabel("Total Income")

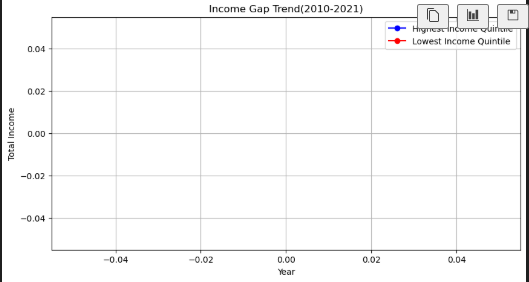
plt.title("Income Gap Trend(2010-2021)")

plt.legend()

plt.grid(True)

plt.show()

**OUTPUT:- This shows if income inequality is increasing or decreasing over time.**

****

**Q.11: What is the median expense by province?**

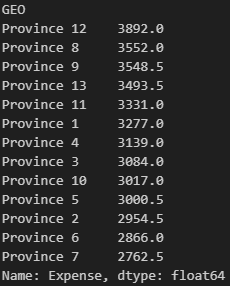
**Ans:** # Group by province and calculate median expense

median\_expense\_province = df.groupby("GEO")["Expense"].median()

# Display results

print(median\_expense\_province.sort\_values(ascending=False))

**OUTPUT:- The median is often a better measure than the average when there are extreme values.**

****

**Q.12: Which provinces have the highest number of records?**

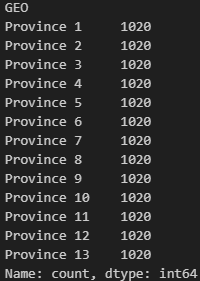
**Ans:** # Count number of records per province

record\_count = df["GEO"].value\_counts()

# Display results

print(record\_count)

**OUTPUT:- This helps us understand if some provinces have more data points than others.**

****

**Q. 13: How does expense distribution vary among income quintiles?**

**Ans:** import seaborn as sns

# Plot boxplot

plt.figure(figsize=(10,5))

sns.boxplot(x=df["Before-tax household income quintile"], y= df["Expense"])

# Labels and title

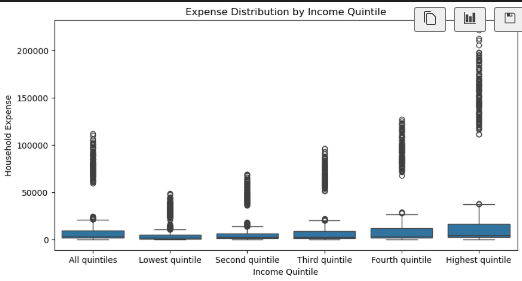
plt.xlabel("Income Quintile")

plt.ylabel("Household Expense")

plt.title("Expense Distribution by Income Quintile")

plt.show()

**OUTPUT:- This shows how expenses vary within each income level and highlights outliers.**

****

**Q.** **14: Which age group has the highest savings?**

**Ans:** # Group by age and sum income & expense

age\_savings = df.groupby("Age of older adult")[["Income","Expense"]].sum()

# Calculate savings ratio

age\_savings["Saving Ratio"] = age\_savings["Income"]/age\_savings["Expense"]

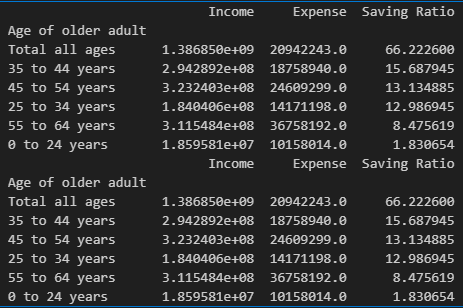
# Display results

print(age\_savings.sort\_values(by="Saving Ratio", ascending=False))

# Display results

print(age\_savings.sort\_values(by="Saving Ratio",ascending=False))

**OUTPUT:- Higher ratios mean more savings; lower ratios mean higher spending relative to income.**

****

**Q.15:** **What is the year-over-year income growth rate?**

**Ans**: # Group by year and sum income

yearly\_income = df.groupby("REF\_DATE")["Income"].sum()

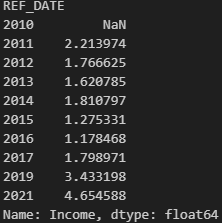
# Calculate percentage growth

income\_growth = yearly\_income.pct\_change()\*100

# Display results

print(income\_growth)

**OUTPUT:- This shows how income is growing (or declining) over time.**

****

**Q.** **16: What percentage of income is spent on expenses per year?**

**Ans:** # Group by year and sum income & expense

yearly\_data = df.groupby("REF\_DATE")[["Income","Expense"]].sum()

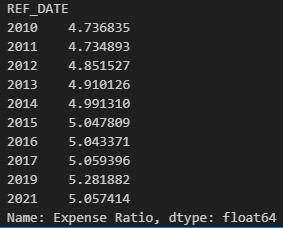
# Calculate expense-to-income ratio

yearly\_data["Expense Ratio"] = (yearly\_data["Expense"]/yearly\_data["Income"])\*100

# Display results

print(yearly\_data["Expense Ratio"])

**OUTPUT:- A higher percentage means people are spending more relative to their income.**

****

**Q.** **17: How do expenses vary over time for each income group?**

**Ans:** # Group by year and income quintile, then sum expenses

expense\_trend = df.groupby(["REF\_DATE","Before-tax household income quintile"])["Expense"].sum().unstack()

# Plot the data

expense\_trend.plot(figsize=(12,6),marker='o')

# Labels and title

plt.xlabel("Year")

plt.ylabel("Total Expenses")

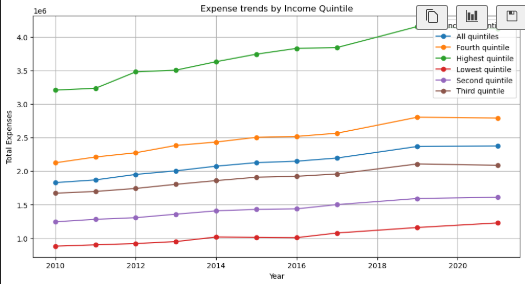
plt.title("Expense trends by Income Quintile")

plt.legend(title="Income Quintile")

plt.grid(True)

plt.show()

**OUTPUT:- : This shows how spending habits change over time for different income groups.**

****

**Q.** **18: How do different provinces compare in income and expense distribution?**

**Ans:** # Group by province and sum income & expense

province\_data = df.groupby("GEO")[["Income","Expense"]].sum()

# Scatter plot

plt.figure(figsize=(8,5))

sns.scatterplot(x=province\_data["Income"],y=province\_data["Expense"],hue=province\_data.index,s=100)

# Labels and title

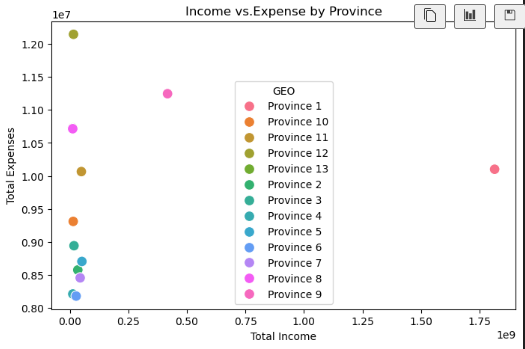
plt.xlabel("Total Income")

plt.ylabel("Total Expenses")

plt.title("Income vs.Expense by Province")

plt.show()

**OUTPUT:- This helps identify which provinces earn and spend the most**

****

**Q.** **19: What is the most common expense category?**

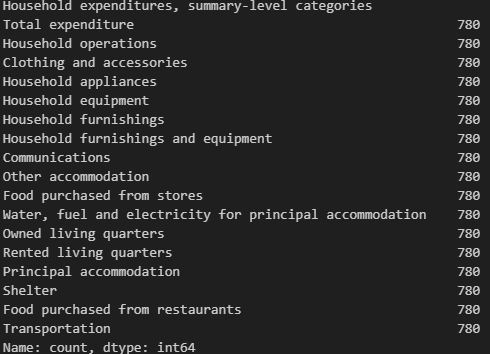
**Ans:** # Count frequency of each expense category

expense\_category\_counts = df["Household expenditures, summary-level categories"].value\_counts()

# Display results

print(expense\_category\_counts)

**OUTPUT:- This tells us where households spend the most money.**

****