income vs expenses

April 22, 2025

[ ]:

```
# Income vs Expense Affordability Analysis

import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Load the dataset
df = pd.read_csv('income_data/Inc_Exp_Data.csv')

# --- Basic Overview ---
print("Dataset Overview:")
print(df.head())
print("\nColumns:", df.columns.tolist())

# --- Descriptive Statistics ---
print("\nDescriptive Stats:")
print(df[['Mthly_HH_Income', 'Mthly_HH_Expense']].describe())

# --- Correlation Analysis ---
correlation = df['Mthly_HH_Income'].corr(df['Mthly_HH_Expense'])
print(f"\nCorrelation between income and expenses: {correlation:.2f}")

# --- Affordability Metric ---
df['Expense_%_of_Income'] = (df['Mthly_HH_Expense'] / df['Mthly_HH_Income']) *␣
 ↪100
print("\nExpense as % of Income (first 5 rows):")
print(df[['Mthly_HH_Income', 'Mthly_HH_Expense', 'Expense_%_of_Income']].head())

# --- Visualization: Income vs Expense ---
sns.scatterplot(data=df, x='Mthly_HH_Income', y='Mthly_HH_Expense')
plt.title("Monthly Income vs Monthly Expense")
plt.xlabel("Monthly Household Income")
plt.ylabel("Monthly Household Expense")
plt.grid(True)
plt.show()

# --- Visualization: Expense % of Income ---
```

```python
sns.histplot(df['Expense_%_of_Income'], kde=True)
plt.title("Proportion of Income Spent on Expenses")
plt.xlabel("Expenses as % of Income")
plt.grid(True)
plt.show()

# --- Visualization: Expense vs Earners, Colored by Family Size ---
sns.lmplot(data=df, x='No_of_Earning_Members', y='Mthly_HH_Expense', hue='No_of_Fly_Members', aspect=1.5)
plt.title("Expense vs Earners, Colored by Family Size")
plt.xlabel("Number of Earning Members")
plt.ylabel("Monthly Expense")
plt.grid(True)
plt.show()

# --- Affordability Thresholds ---
print("\nAffordability Threshold Guidelines:")
print("- < 40% of income: Generally affordable")
print("- 40% to 60%: Caution zone")
print("- > 60%: High expense burden (possibly unaffordable)")

# Categorize households by affordability
conditions = [
    df['Expense_%_of_Income'] < 40,
    df['Expense_%_of_Income'].between(40, 60),
    df['Expense_%_of_Income'] > 60
]
labels = ['Affordable', 'Moderate', 'High Burden']
df['Affordability_Level'] = pd.cut(df['Expense_%_of_Income'], bins=[0, 40, 60, 100], labels=labels)

# Affordability distribution plot
sns.countplot(x='Affordability_Level', data=df)
plt.title("Household Affordability Levels")
plt.xlabel("Affordability Category")
plt.ylabel("Number of Households")
plt.grid(True)
plt.show()

# Summary distribution
print("\nAffordability Level Distribution:")
print(df['Affordability_Level'].value_counts())

# --- Additional Insights ---
# Income and Expense by Education Level
plt.figure(figsize=(10, 5))
sns.boxplot(data=df, x='Highest_Qualified_Member', y='Mthly_HH_Income')
```

```python
plt.title("Monthly Income by Education Level")
plt.xticks(rotation=45)
plt.grid(True)
plt.show()

sns.boxplot(data=df, x='Highest_Qualified_Member', y='Mthly_HH_Expense')
plt.title("Monthly Expense by Education Level")
plt.xticks(rotation=45)
plt.grid(True)
plt.show()

# Affordability by Number of Earners
sns.boxplot(data=df, x='No_of_Earning_Members', y='Expense_%_of_Income')
plt.title("Expense % of Income by Number of Earners")
plt.xlabel("Number of Earning Members")
plt.ylabel("Expenses as % of Income")
plt.grid(True)
plt.show()

print("\nAdditional Observations:")
print("- Higher education levels generally correlate with higher income and slightly more controlled expenses.")
print("- Households with more earning members tend to have lower expense burdens proportionally.")
```