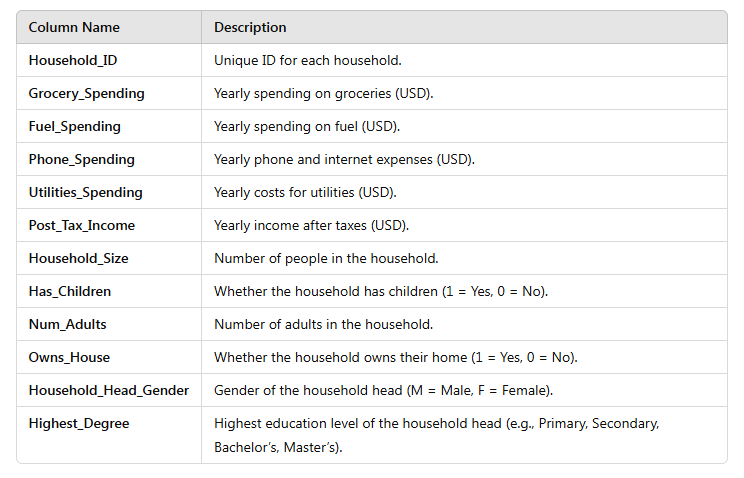
**Household Expenditure Analysis and Demographic Insights**

Aim

The case study aims to help you understand household spending patterns and how they connect to income and demographic factors such as family size and education. As you work through it, you’ll develop essential data analysis skills and gain valuable insights into consumer behavior.

#### **Dataset Information**

This dataset contains information expenditure in different categories, demographic characteristics, and other relevant variables. The goal of this analysis is to help you, the data analyst, understand household spending behavior and its potential relationship with demographic factors like income, household composition, and the head's education level.

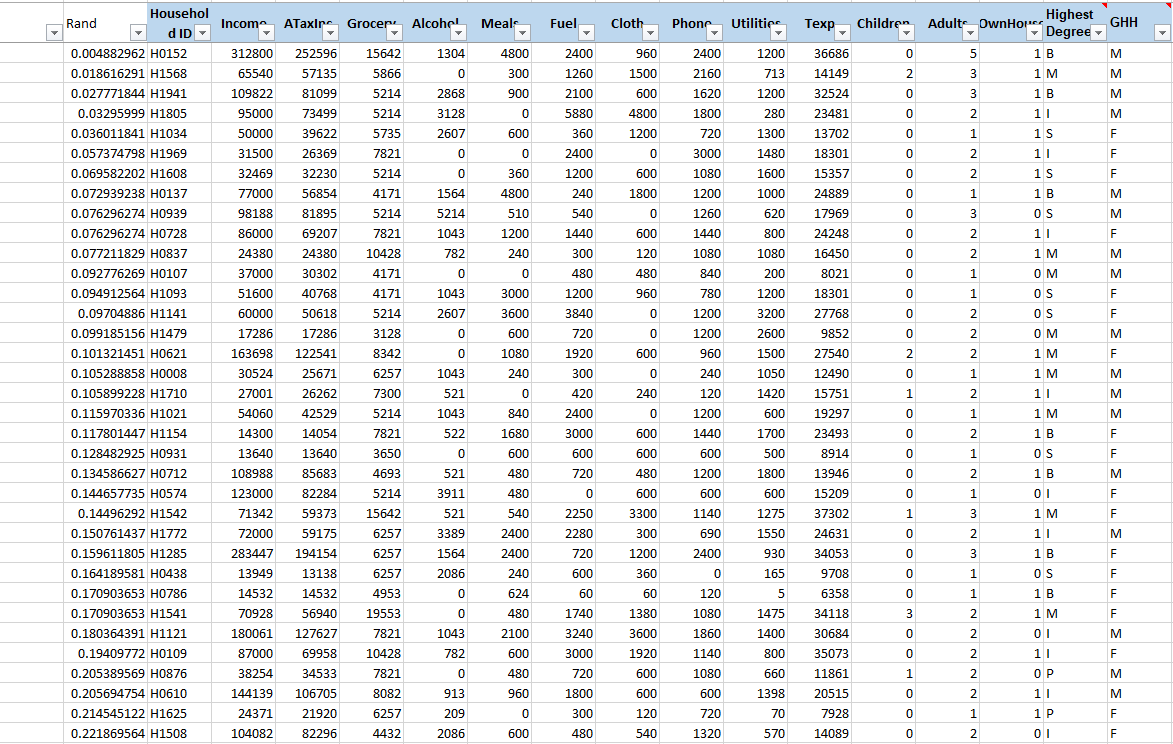


Following are the different tasks that you have to complete based on the dataset that is been provided to you:

**Task 1:**

1. Draw a unique random sample of two hundred and fifty (250) households following the instruction sheet available with the unprocessed data set.

Solution:

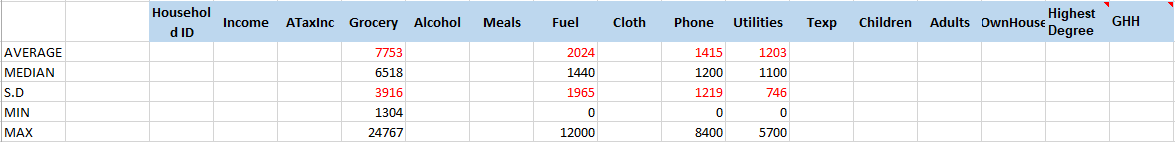


We selected the Random Sampling Method because it ensures every household has an equal chance of being chosen, minimizing bias and making the sample representative of the entire population. It’s simple to implement, statistically valid, and reproducible.

B. Compute the descriptive statistics for the following four expenditure variables.

1. Grocery
2. Fuel
3. Phone
4. Utilities

Solution:



C. Use information from the descriptive statistics in part (B) above to present a summary of your findings by contrasting different features of the distributions of the four variables.

Solution:

1. **Grocery Spending**:

* **Mean vs. Median**: The mean ($7,753) is higher because a few households with high grocery bills skew the average.
* **High Variability**: The large standard deviation ($3,916) reflects significant differences in spending, influenced by family size or premium items.
* **Range**: The wide range ($1,304 to $24,767) shows a mix of budget-conscious and high-end spenders.

2. **Fuel Spending**:

* **Mean vs. Median**: The mean ($2,024) is slightly higher, indicating a few households with high fuel costs.
* **High Variability**: The standard deviation ($1,965) shows that fuel spending varies due to factors like vehicle type and commuting habits.
* **Range**: The large range ($0 to $12,000) reflects significant differences in spending based on driving frequency and vehicle size.

3. **Phone Spending**:

* **Mean vs. Median**: The mean ($1,415) exceeds the median ($1,200) because some households spend much more on premium plans or phones.
* **High Variability**: The high standard deviation ($1,219) indicates varied spending, with some households spending very little and others a lot.
* **Range**: The range ($0 to $8,400) shows wide variation, influenced by plan choices and phone preferences.

4. **Utilities Spending**:

* **Mean vs. Median**: The mean ($1,203) is slightly higher than the median ($1,100), suggesting more uniform utility spending.
* **Lower Variability**: The standard deviation ($746) is lower, showing less variation in utility bills.
* **Range**: The range ($0 to $5,700) indicates some households have very low bills, while others have higher costs due to home size or energy use.

D. Draw a scatter plot of the natural log of total expenditures against the natural log of post-tax income, that is, ln(Texp) against ln(ATaxInc). Compute the coefficient of correlation between them. What relationship does the plot and correlation coefficient imply between the two variables?

Solution:

Scatter Plot

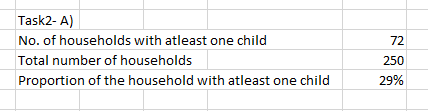
A **correlation coefficient of 0.66** means that there is **a moderate to strong positive correlation** between **the natural log of Total Expenditures (ln(Texp))** and **the natural log of Post-Tax Income (ln(ATaxInc)).**

As **Post-Tax Income** increases, **Total Expenditures** also tend to increase. This is a **positive relationship**, meaning that higher income is generally associated with higher spending, but it's not a perfect relationship.

**Task 2:**

A) Find the Proportion of Households with At Least One Child.

Solution:



* 1. For calculating no. of households with atleast one child.

=COUNTIF(Table1[Children], ">0")

* 1. For calculating the total number of households.

=COUNTA(Table1[Household ID])

* 1. To find out the proportion,

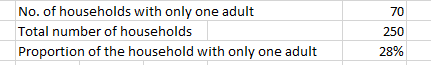
Proportion of Households with At Least One Child

= No. of households with atleast one child/ Total number of households.

According to the data **29% of households** in the dataset have at least one child.

B) Find the Proportion of Households with Exactly One Adult.

Solution:



* 1. For calculating no. of households with only one adult.

=COUNTIF(Table1[Adults], "=1")

b) For calculating the total number of households.

=COUNTA(Table1[Household ID])

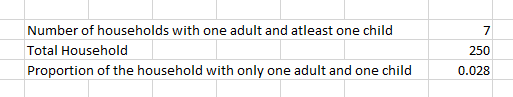
c) To find out the proportion,

Proportion of Households with At Least One Child

= No. of households with only one adult/ Total number of households

According to the data **28% of households** in the dataset have only one adult.

C) What is the proportion of the households where there are only one adult and at least one child?



* 1. For calculating the number of households with one adult and at least one child we have made use of COUNTIFS()

=COUNTIFS(Table3[Adults], 1, Table3[Children], ">0")

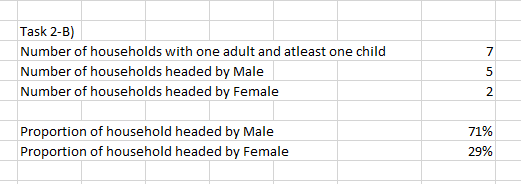
* 1. For calculating the total household in the dataset we have used COUNTA().

=COUNTA(Table3[Household ID])

* 1. Proportion is calculated using the formula:

Proportion = Household with only one adult and at least one child/ Total Household

D) Among the households consisting of only one adult and at least one child (the ones you looked at in part C above), find the proportion of male-headed households. Is a single father household more common than a single mother household?



1. For calculating the number of households with one adult and at least one child.

=COUNTIFS(Table3[Adults], 1, Table3[Children], ">0")

1. For calculating the number of households headed by Male.

=COUNTIFS(Table3[Adults], 1, Table3[Children], ">0", Table3[GHH], "M")

1. For calculating the number of households headed by Female.

=COUNTIFS(Table3[Adults], 1, Table3[Children], ">0", Table3[GHH], "F")

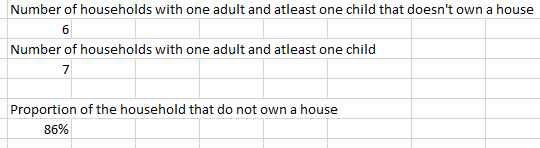
1. Proportion of household headed by Male,

Proportion= Number of households headed by Male/ Number of households with one adult and at least one child

1. Proportion of household headed by Female,
2. Proportion= Number of households headed by Female/ Number of households with one adult and at least one child

Based on the data, the proportion of male-headed households among those with only one adult and at least one child is 71% which is higher than the female.

E) Among the households consisting of only one adult and at least one child (the ones you looked at in part C above), find the proportion of households that do not own a house. Do the majority of single-parent households own a house?



* 1. For calculating the number of household with one adult and one child that doesn’t own a house.

= COUNTIFS(Table3[Adults], 1, Table3[Children], ">0", Table3[OwnHouse], 0**)**

* 1. For calculating the number of households with one adult and at least one child

=COUNTIFS(Table3[Adults], 1, Table3[Children], ">0")

* 1. For calculating the proportion of the household that do not own a house.

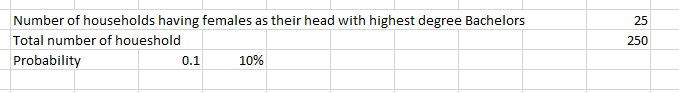
Proportion= Number of household with one adult and one child that doesn’t own a house/ Number of households with one adult and at least one child

Task 3: Household head's education

* 1. Construct a contingency table between the gender and the highest educational degree of the household head.

|  |  |
| --- | --- |
| **Row Labels** | **Count of Highest Degree** |
| **F** | **113** |
| Bachelors(B) | 25 |
| Intermediate(I) | 26 |
| Masters(M) | 19 |
| Primary(P) | 23 |
| Secondary(S) | 20 |
| **M** | **137** |
| Bachelors(B) | 37 |
| Intermediate(I) | 28 |
| Masters(M) | 25 |
| Primary(P) | 15 |
| Secondary(S) | 32 |
| **Grand Total** | **250** |

B. What is the probability that a randomly chosen head of a household is female and her highest degree is Bachelors?



* 1. Number of households having females as their head with highest degree Bachelors

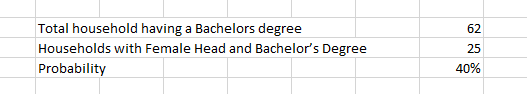
=COUNTIFS(Q2:Q251, "F", P2:P251, "B")

* 1. Total number of household.

=COUNTA(B2:B251)

* 1. Probability = Total Number of HouseholdsNumber of Female Heads with Bachelor’s Degree/ Total Number of Households

C. Among the households whose heads have a Bachelor’s degree, what is the probability that the household head is female?



1. Total household having a Bachelor’s degree.

=COUNTIF(P2:P251, "B")

1. Households with Female Head and Bachelor’s Degree.

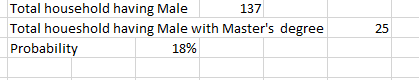
=COUNTIFS(Q2:Q251, "F", P2:P251, "B")

1. For calculating the probability,

Probability= Households with Female Head and Bachelor’s Degree/ Total household having a Bachelor’s degree.

The data shows that there is a 40% probability that the household head is female among households where the head has a Bachelor's degree.

D. Among the households whose heads are male, what is the probability that the household head has a Master’s degree as the highest educational degree?



1. Total household having Male.

=COUNTIF(Q2:Q251, "M")

1. Total household having Male with Master's degree.

=COUNTIFS(Q2:Q251, "M", P2:P251, "M")

1. Probability= Total household having Male with Master’s degree/ Total household having Male.

According to the data, we understand there is 18% probability that the household having Male having Master’s Degree.

The data shows that there is an 18% probability of a household having a male with a Master's degree.

E. Do you think that the two events "the gender of the household head is female" and "the highest educational degree of the household head is primary school certificate" are independent of each other? Justify your answer briefly.

X= “Gender of the household head is female”

Y= “"The highest educational degree of the household head is primary school certificate"

According to the data, we understand that the x and y are independent variable. Since gender has nothing to do with the highest degree that one perceive. Data also represents that there are male candidates who have perceived their highest degree as Primary.

The number of female candidate with the highest degree as the Primary are comparatively higher than the male candidate.