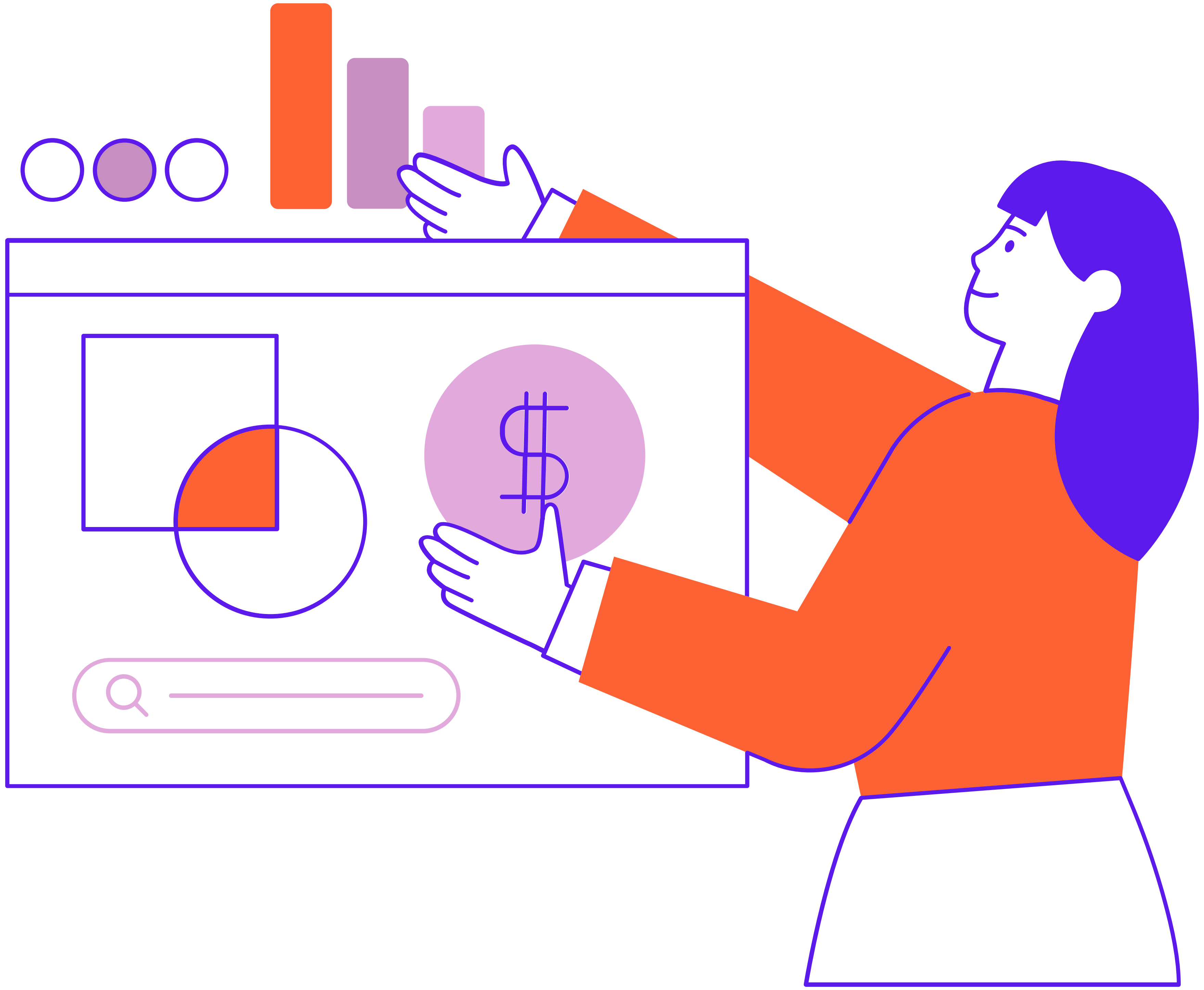
Income Prediction

By Group 2



SINTA BELLA (2802392330)
FARREL GILLAND WIJAYA (2802392381)
RICHIE RIZAWARDANA (2802403100)
ALEXANDER ABEL MAHA (2802420984)
ROBBEN WIJANATHAN (2802461681)

Introduction to Data Science

TABLE OF CONTENTS

01 Background

02 Purpose

03 Focus

04 Descriptive

05 Predictive

06 Conclusion

BACKGROUND

WHY INCOME?

Primary Indicator

Income is a primary indicator of well-being in our daily lives, helping to meet our essential needs.

Various Factors

Income levels are shaped by various factors that influence their distribution among individuals.

Impact

Researching income-influencing factors deepens our understanding of their impact and effectiveness.

OUR PURPOSE

The importance of understanding the factors influencing income distribution and how this knowledge can address disparities, create opportunities, and promote economic equity.

Identify Key Factors

Identify the key factors that influence income levels across different individuals and groups.

Examine Demographic Factors

Examine the role of demographic factors, such as age, education, gender housetype, and etc in shaping income disparities.

Offer Perspectives

Offer insightful and well-informed perspectives to enhance understanding and guide future research or policy decisions.

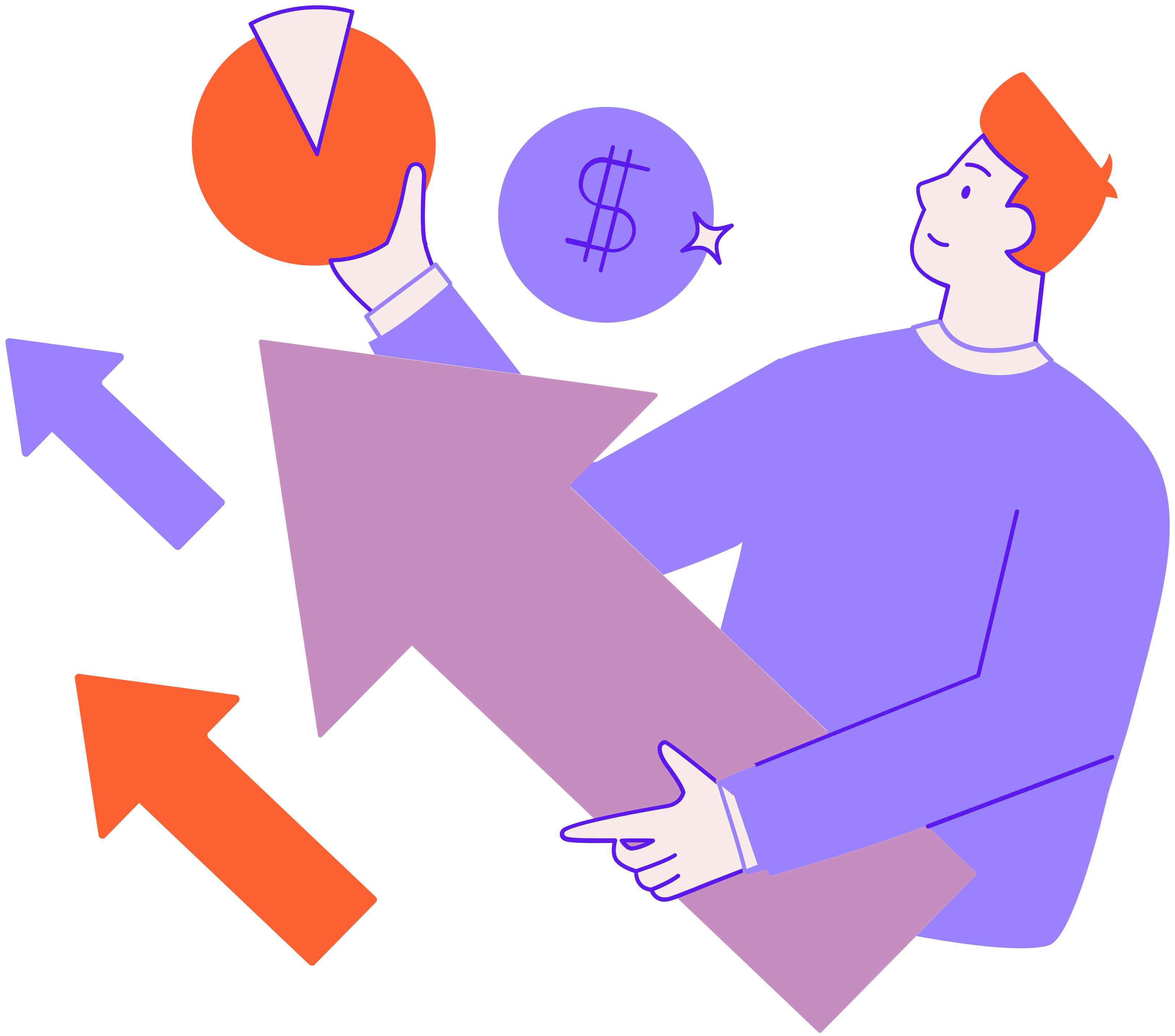
FOCUS OF OUR ANALYSIS

Descriptive

Examining how various factors have influenced income distribution among individuals in the past, helping to reveal key insights into income levels and disparities.

Predictive

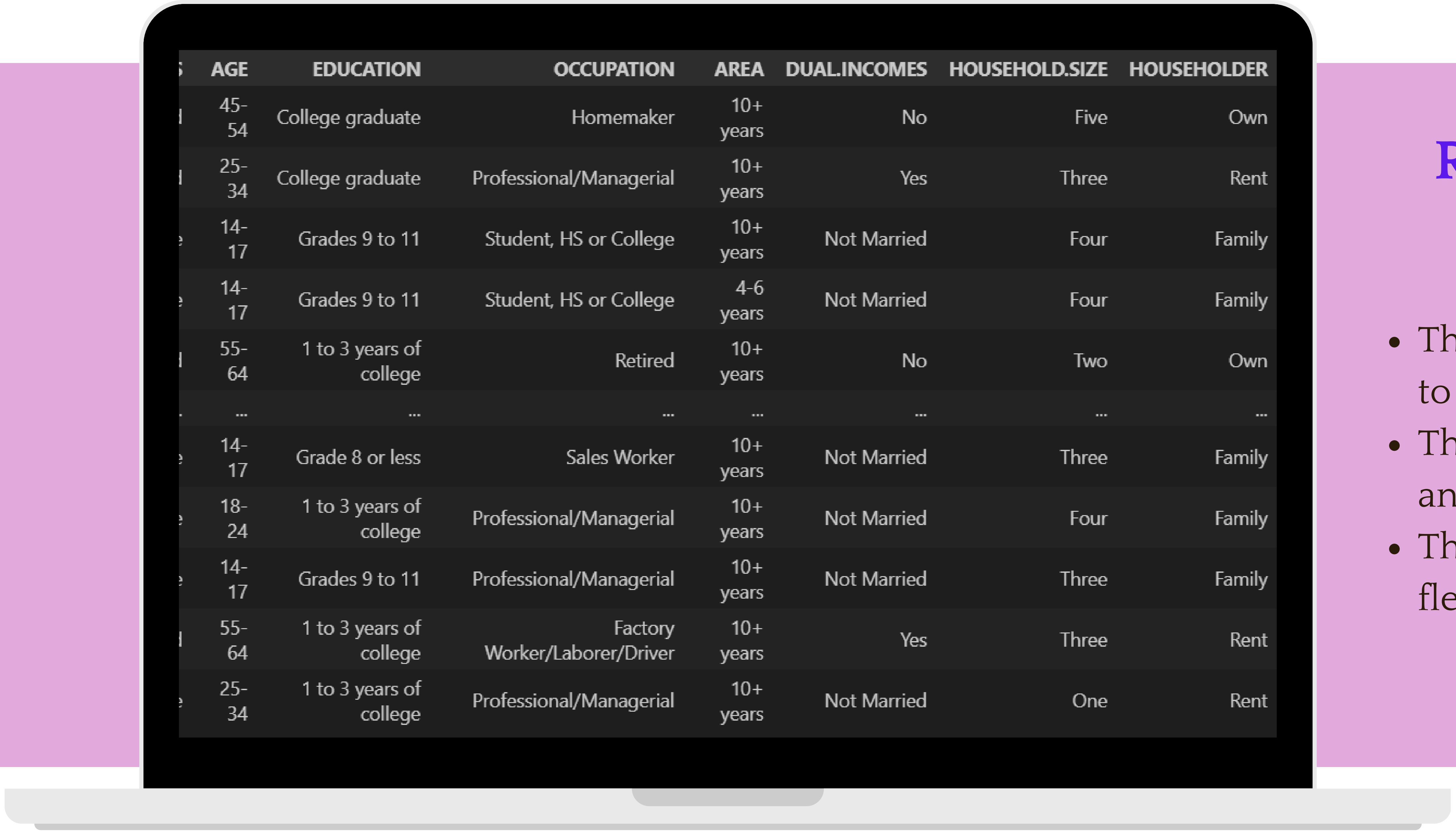
Identify which factors are likely to have the most significant impact on income levels in the future, providing valuable insights into potential income changes and inequalities.



OUR DATASET

Source

https://github.com/selva86/datasets/blob/master/income.csv

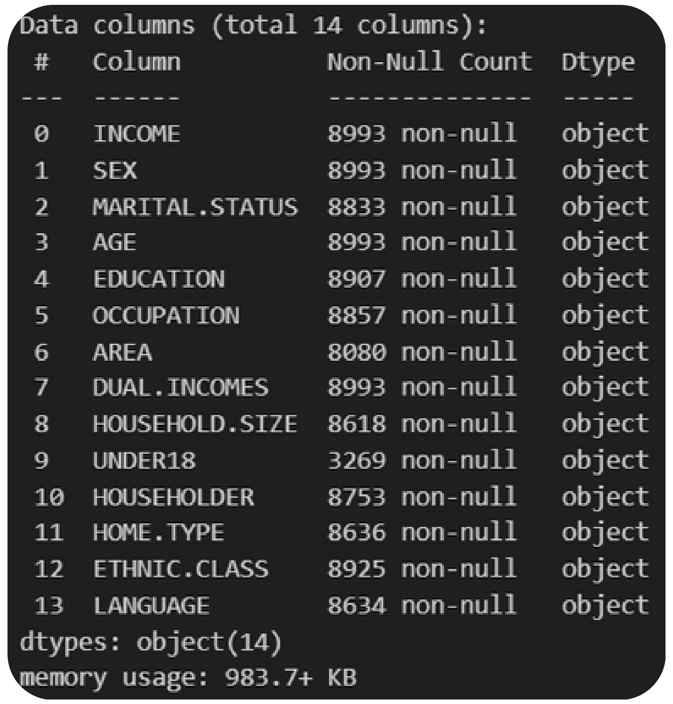


AGE	EDUCATION	OCCUPATION	AREA	DUAL.INCOMES	HOUSEHOLD.SIZE	HOUSEHOLDER
45-54	College graduate	Homemaker	10+ years	No	Five	Own
25-34	College graduate	Professional/Managerial	10+ years	Yes	Three	Rent
14-17	Grades 9 to 11	Student, HS or College	10+ years	Not Married	Four	Family
14-17	Grades 9 to 11	Student, HS or College	4-6 years	Not Married	Four	Family
55-64	1 to 3 years of college	Retired	10+ years	No	Two	Own
...	...	...	...	...	...	...
14-17	Grade 8 or less	Sales Worker	10+ years	Not Married	Three	Family
18-24	1 to 3 years of college	Professional/Managerial	10+ years	Not Married	Four	Family
14-17	Grades 9 to 11	Professional/Managerial	10+ years	Not Married	Three	Family
55-64	1 to 3 years of college	Factory Worker/Laborer/Driver	10+ years	Yes	Three	Rent
25-34	1 to 3 years of college	Professional/Managerial	10+ years	Not Married	One	Rent

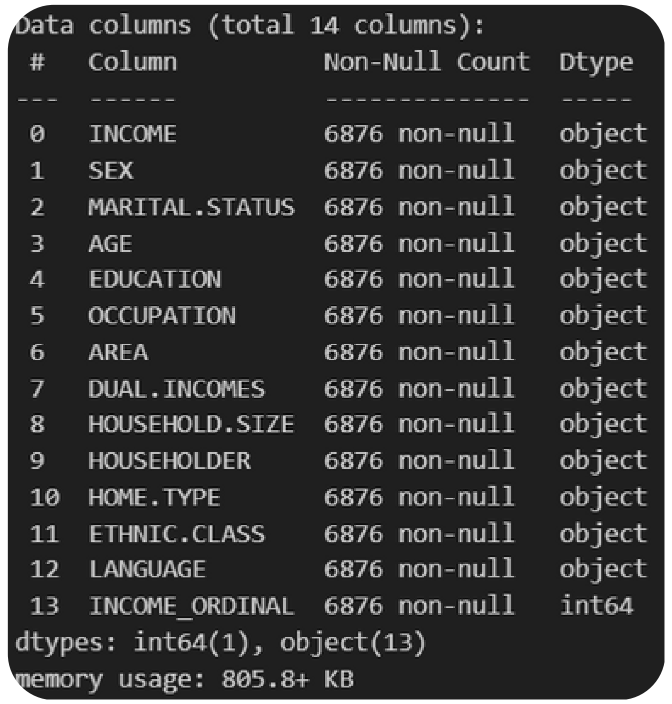
Reason for Using the Dataset:

- The complexity of data relevant to the research topic.
- The completeness of the data and its easy accessibility online.
- The diverse variables, making it flexible for analysis.

DATA CLEANSING & MAPPING



```
Data columns (total 14 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   INCOME          8993 non-null   object
 1   SEX             8993 non-null   object
 2   MARITAL.STATUS  8833 non-null   object
 3   AGE             8993 non-null   object
 4   EDUCATION       8907 non-null   object
 5   OCCUPATION      8857 non-null   object
 6   AREA            8080 non-null   object
 7   DUAL.INCOMES    8993 non-null   object
 8   HOUSEHOLD.SIZE  8618 non-null   object
 9   UNDER18         3269 non-null   object
 10  HOUSEHOLDER     8753 non-null   object
 11  HOME.TYPE       8636 non-null   object
 12  ETHNIC.CLASS    8925 non-null   object
 13  LANGUAGE        8634 non-null   object
dtypes: object(14)
memory usage: 983.7+ KB
```



```
Data columns (total 14 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   INCOME          6876 non-null   object
 1   SEX             6876 non-null   object
 2   MARITAL.STATUS  6876 non-null   object
 3   AGE             6876 non-null   object
 4   EDUCATION       6876 non-null   object
 5   OCCUPATION      6876 non-null   object
 6   AREA            6876 non-null   object
 7   DUAL.INCOMES    6876 non-null   object
 8   HOUSEHOLD.SIZE  6876 non-null   object
 9   HOUSEHOLDER     6876 non-null   object
 10  HOME.TYPE       6876 non-null   object
 11  ETHNIC.CLASS    6876 non-null   object
 12  LANGUAGE        6876 non-null   object
 13  INCOME_ORDINAL  6876 non-null   int64 
dtypes: int64(1), object(13)
memory usage: 805.8+ KB
```

Dropping Rows

We drop rows where any of the following critical columns have missing values: 'MARITAL.STATUS', 'EDUCATION', 'OCCUPATION', 'AREA', 'HOUSEHOLD.SIZE', 'HOUSEHOLDER', 'HOME.TYPE', 'ETHNIC.CLASS', 'LANGUAGE'.

Dropping Columns

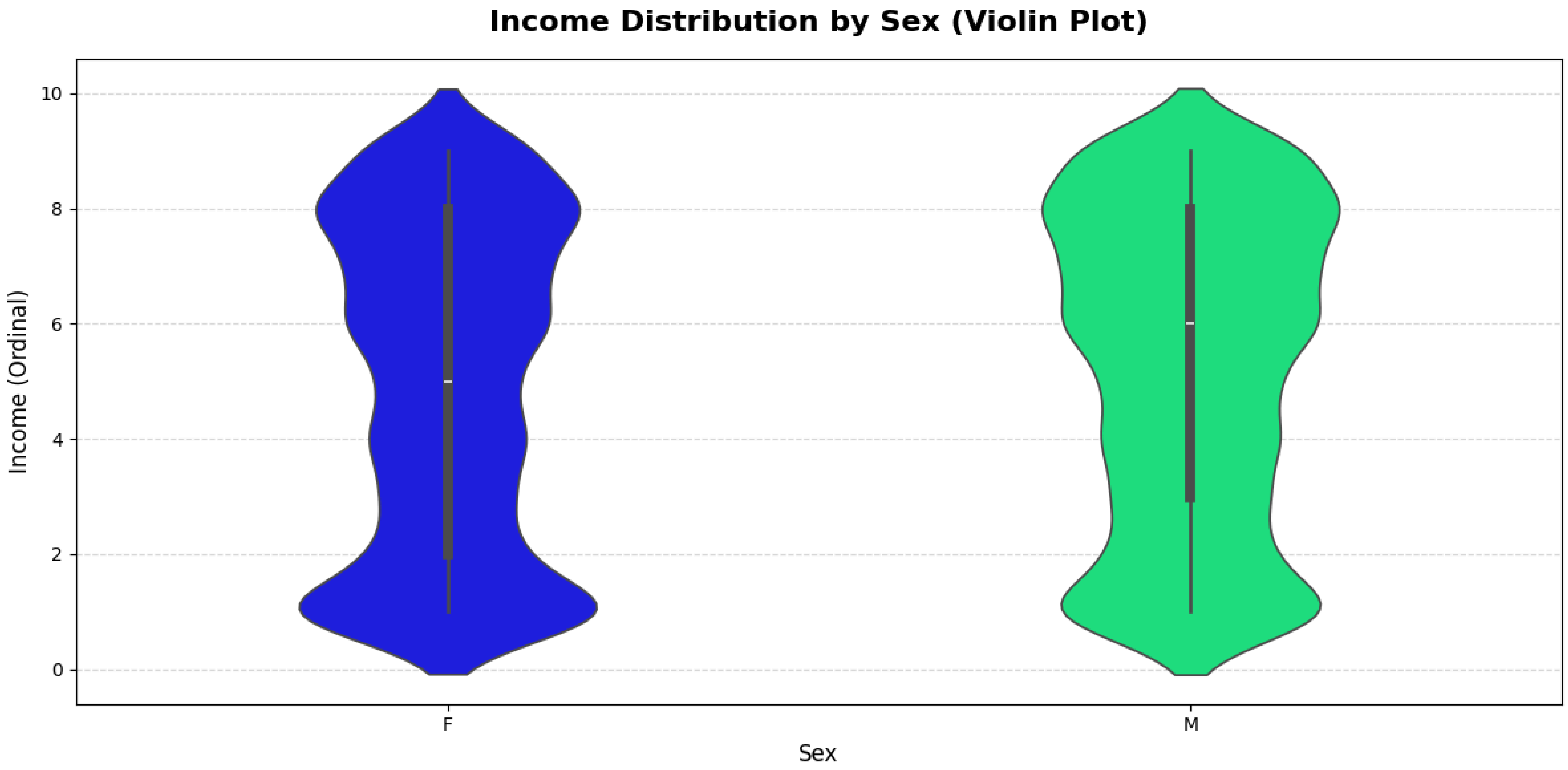
The column 'UNDER18' is removed because it is not relevant to the current analysis.

Mapping Ordinal Values

The INCOME column is an ordinal categorical variable, representing income ranges with the following mappings:

1: Less than $10,000
2: $10,000 – $15,000
3: $15,000 – $20,000
4: $20,000 – $25,000
5: $25,000 – $30,000
6: $30,000 – $40,000
7: $40,000 – $50,000
8: $50,000 – $75,000
9: More than $75,000

INCOME DISTRIBUTION BY BY SEX



Point 01

The graph shows similar income densities for men and women, but men earn slightly more, possibly due to differences in job roles, hours worked, or industry representation.

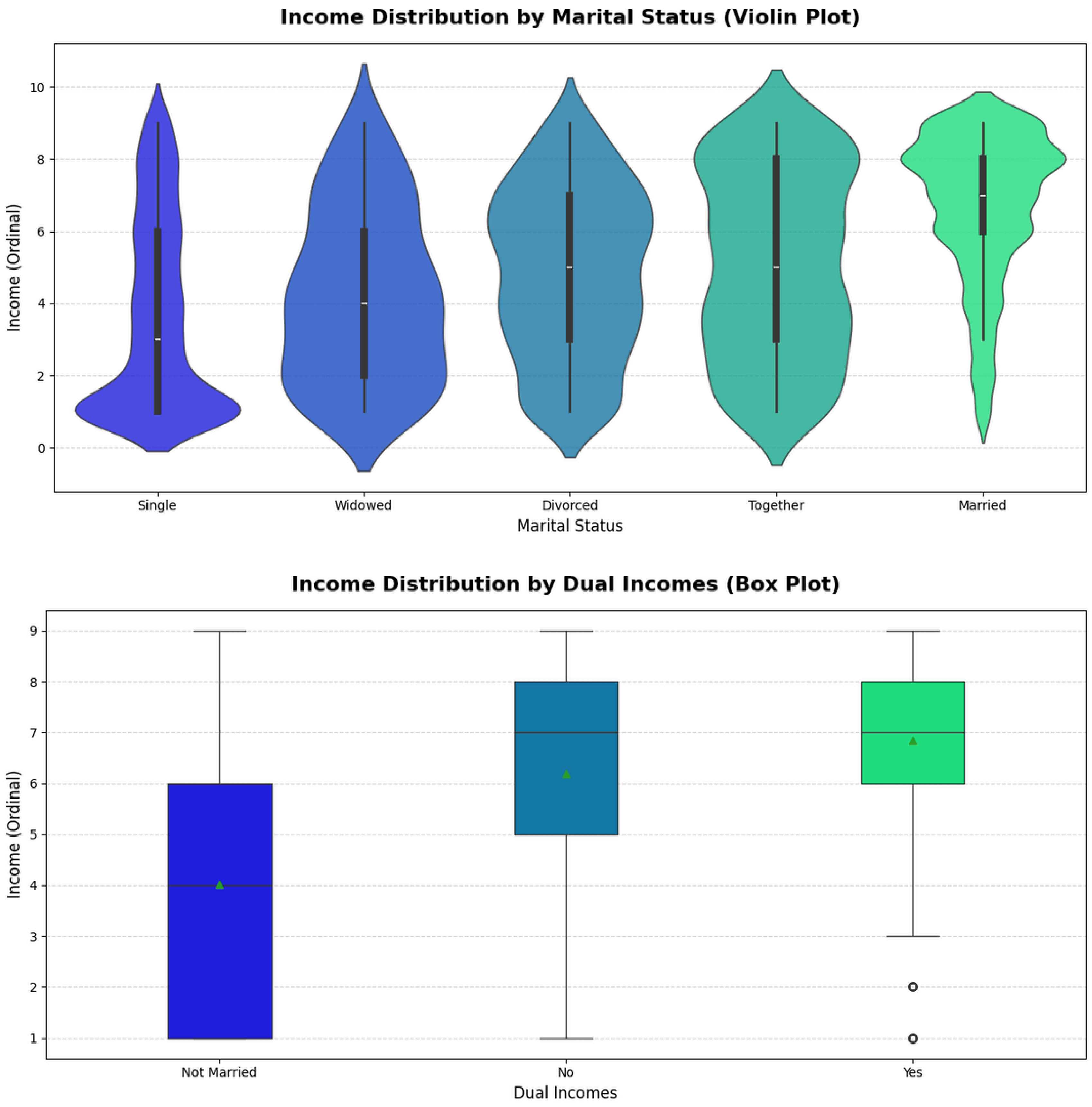
INCOME DISTRIBUTION BY MARITAL STATUS & DUAL INCOMES

Marital Status

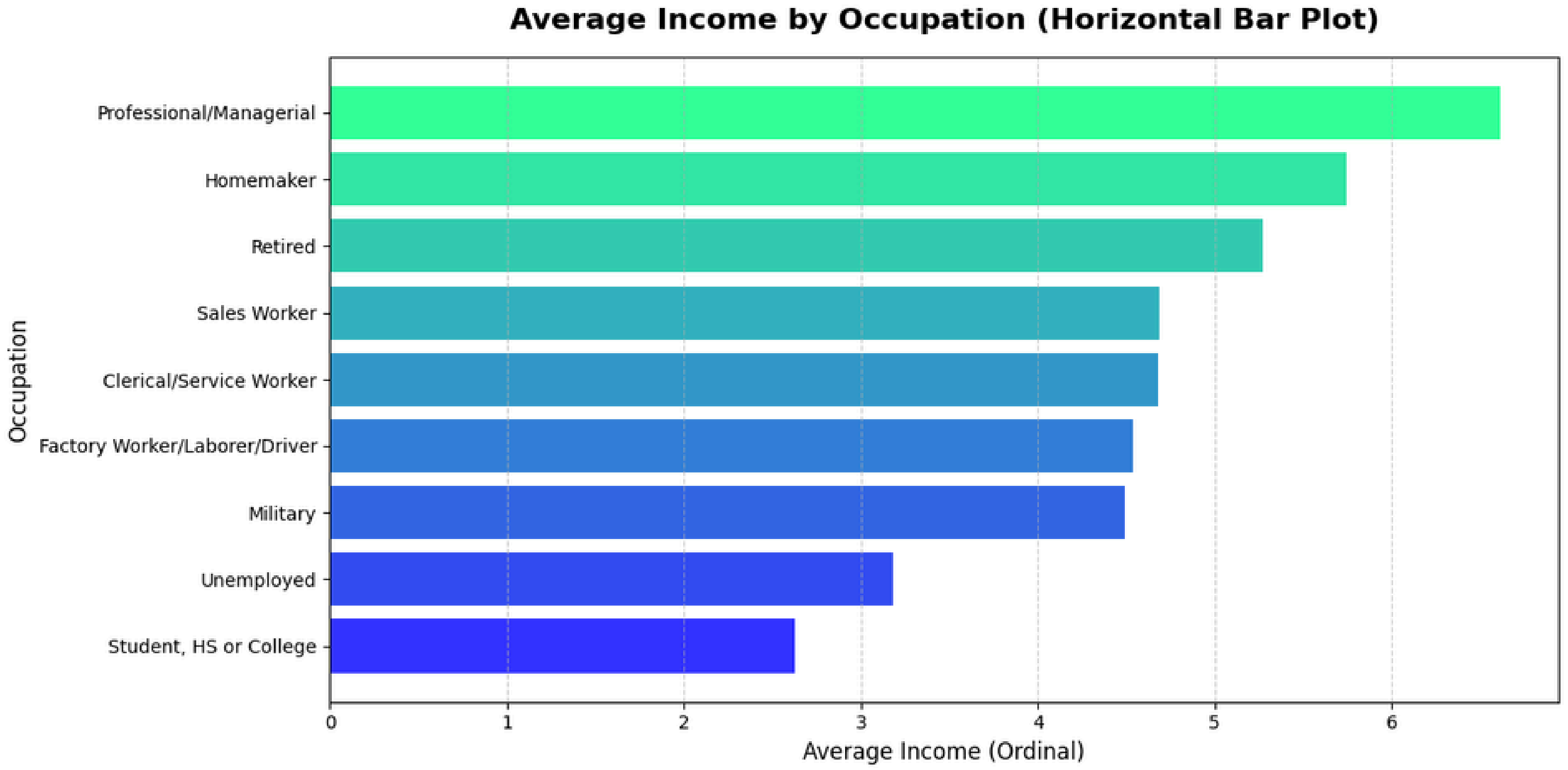
First graph varies by marital status, married or cohabiting individuals tend to have higher, more consistent incomes, while singles and widows have lower or more limited distributions.

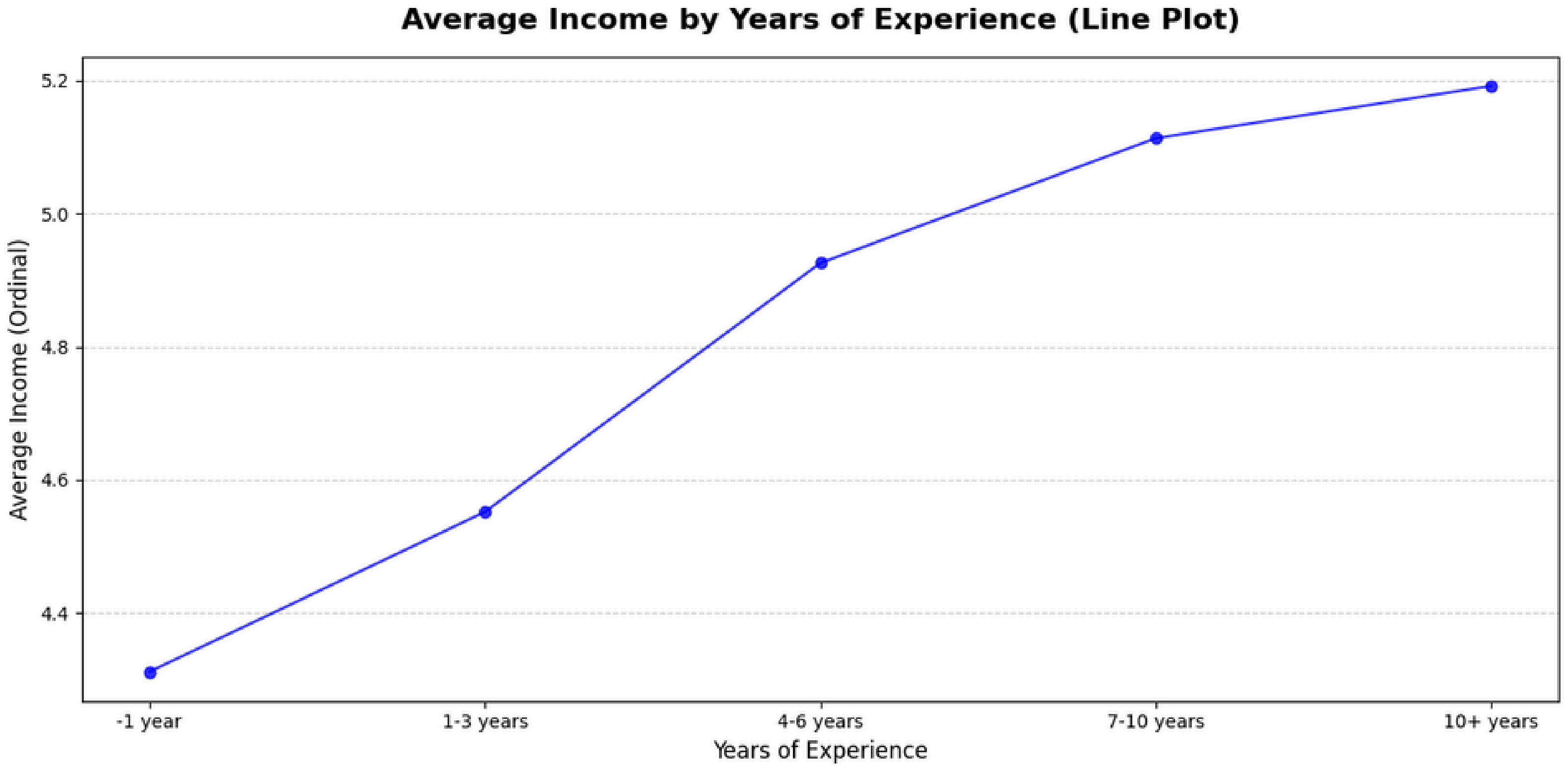
Dual Incomes

Second graph shows that among married individuals, households with dual incomes tend to have higher and more consistent average earnings.



AVERAGE INCOME BY OCCUPATION & YEARS OF EXPERIENCE





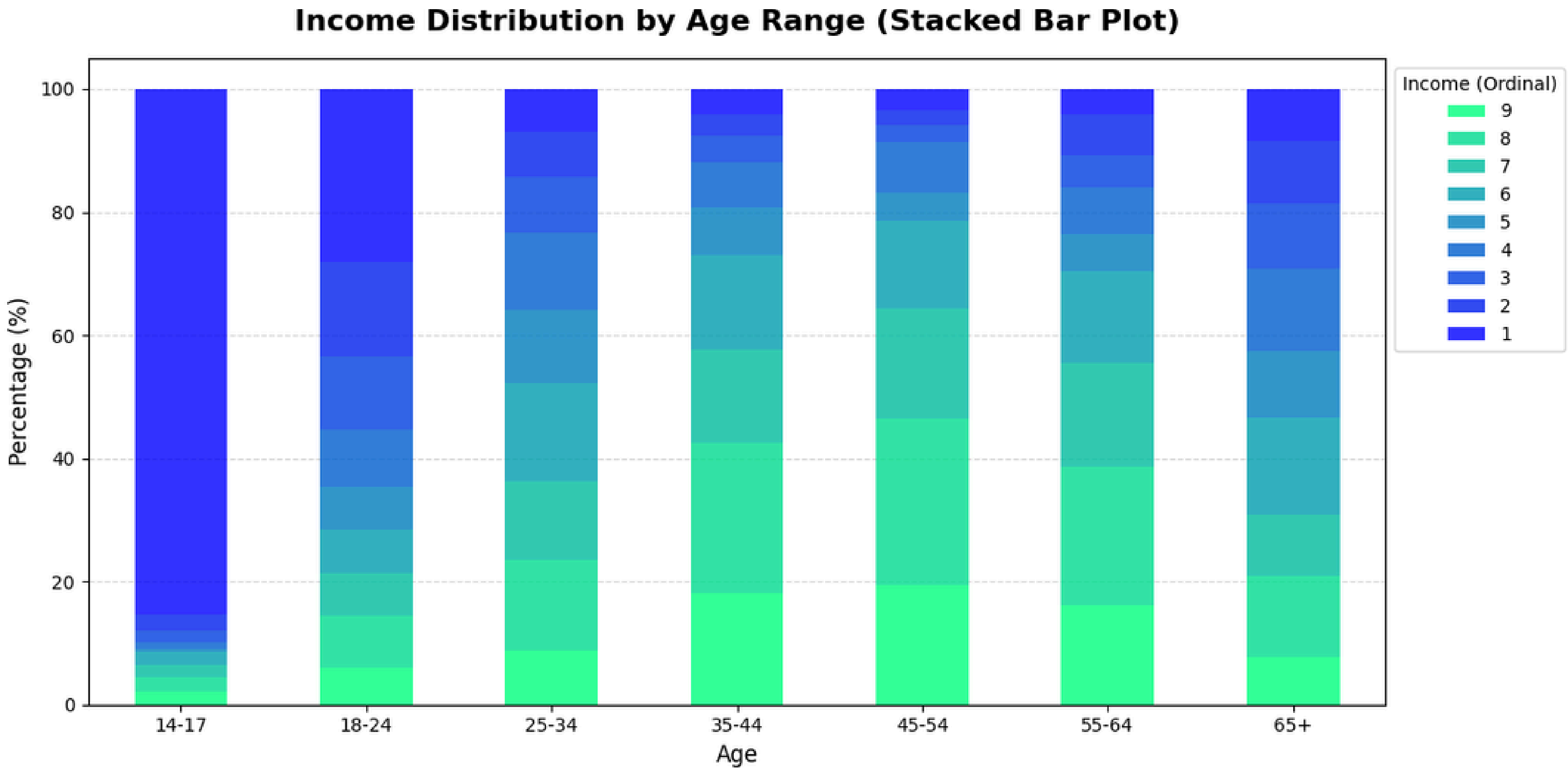
Occupation

The graph shows that individuals in professional roles have the highest income levels. Retirees also fall into a high-income category, likely due to accumulated savings or pensions. This indicates a correlation between one's occupation.

Years of Experience

The graph illustrates a steady increase in income with years of experience. As individuals gain experience and develop advanced skills, they become more capable of handling complex and challenging tasks. This progression explains why greater experience often leads to higher earning potential.

INCOME DISTRIBUTION BY AGE



Point 01

The 14-17 age group has the lowest income, with about 90% earning similarly (1), likely due to limited experience as they are often studying or job hunting.

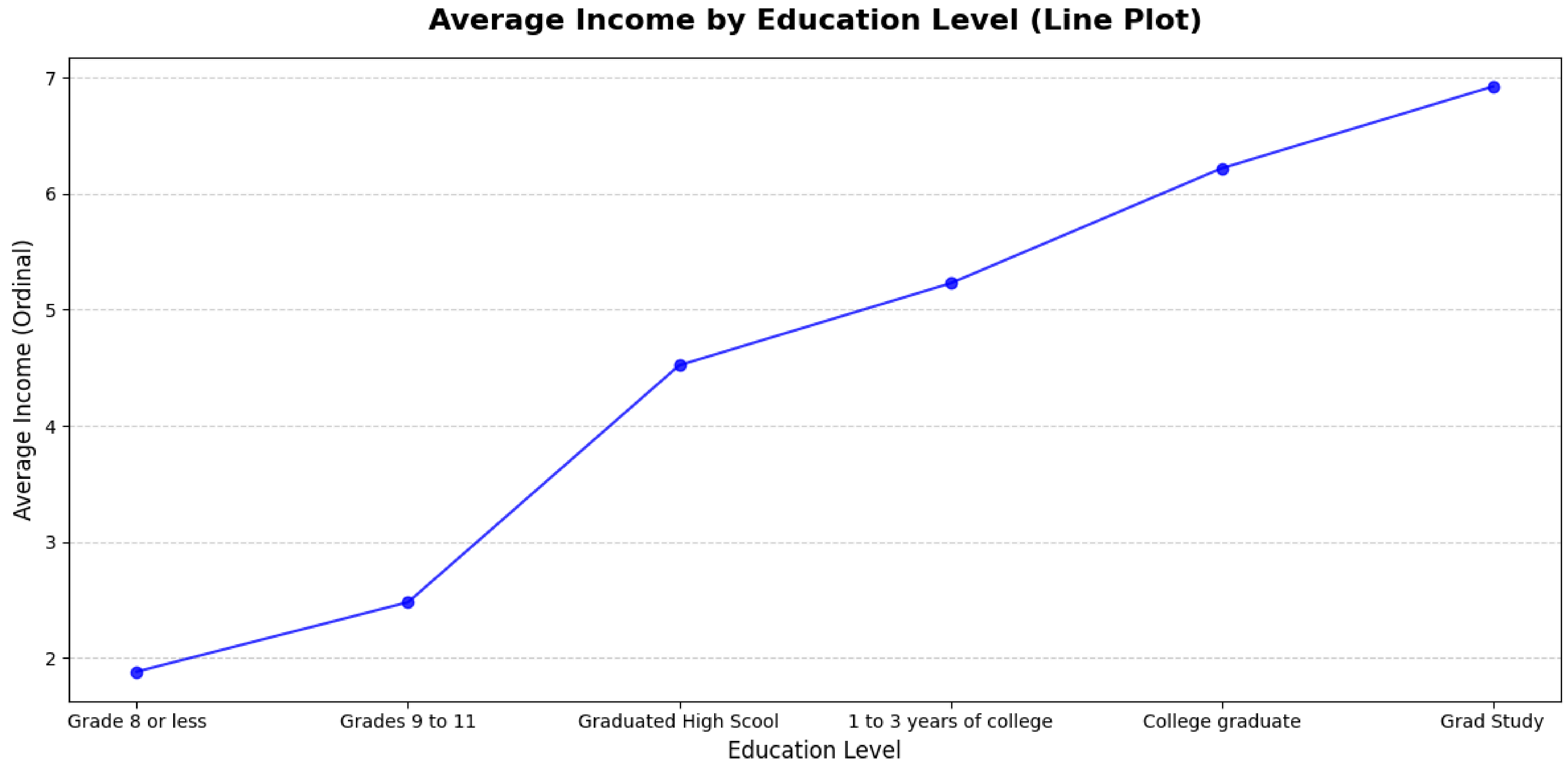
Point 02

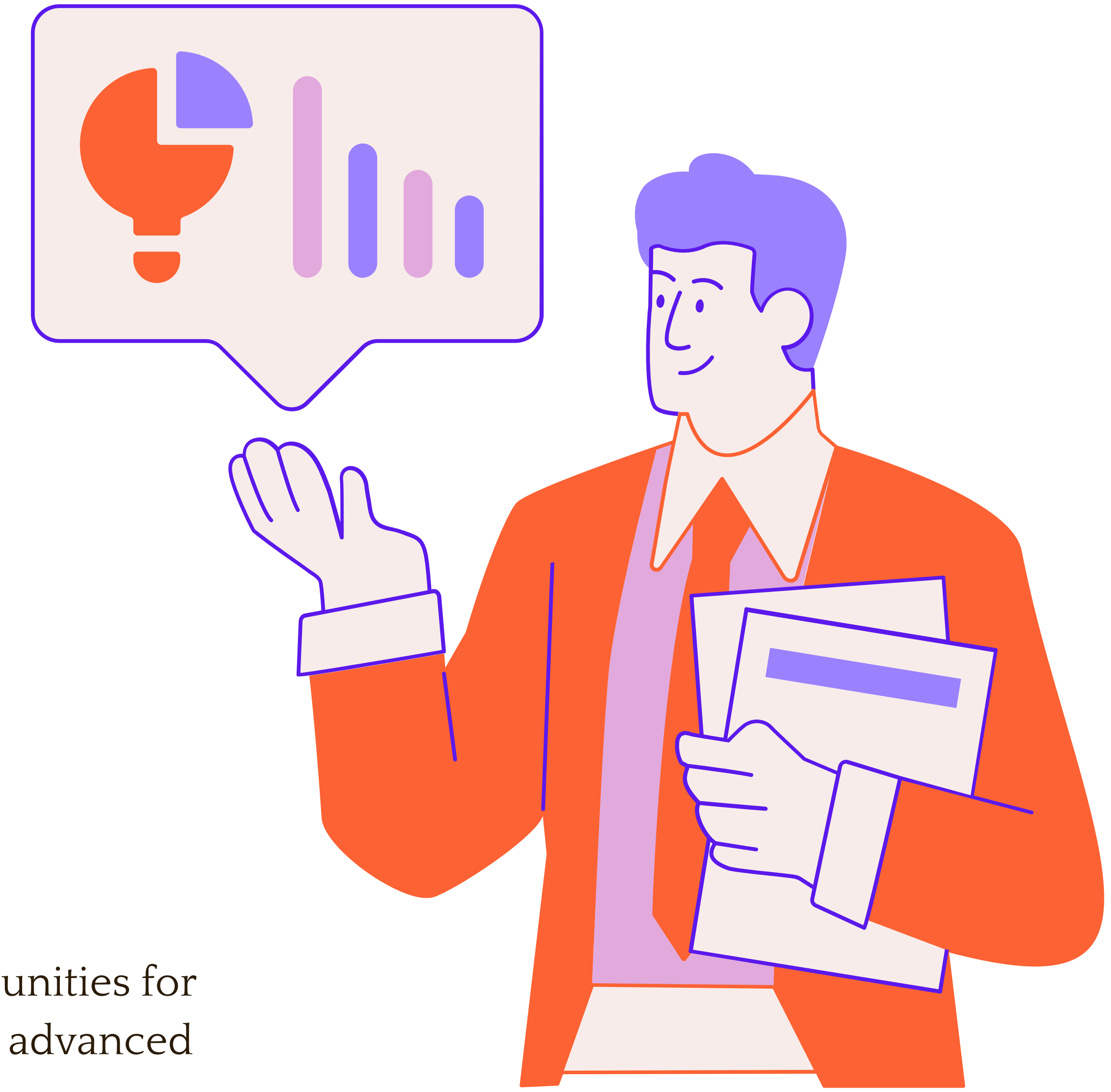
The 45-54 age group has the highest income, likely due to experience and seniority.

Point 03

Older age groups (55-65 & 65+) tend to earn less due to declining productivity with age.

AVERAGE INCOME BY EDUCATION LEVEL





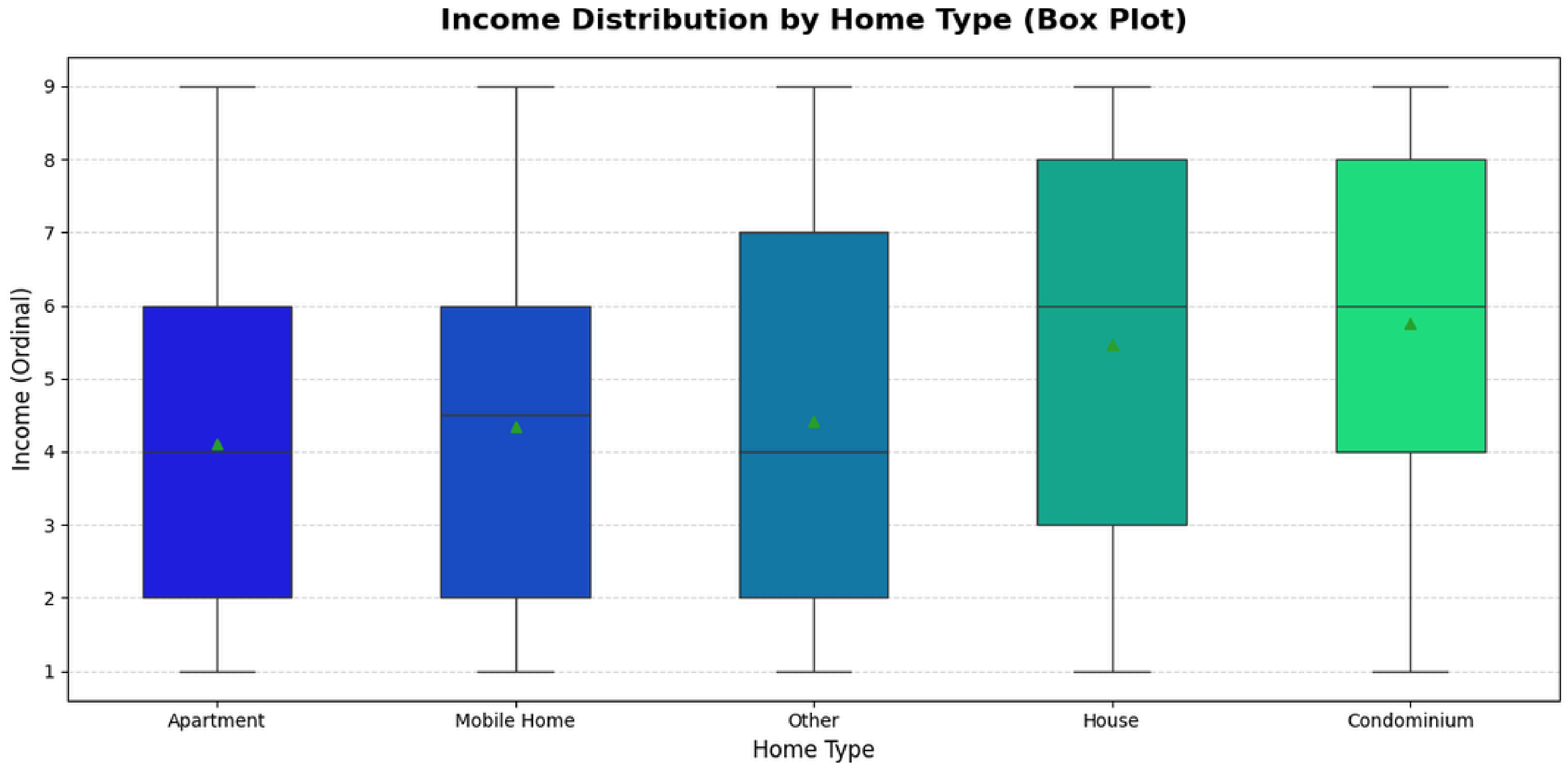
Point 01

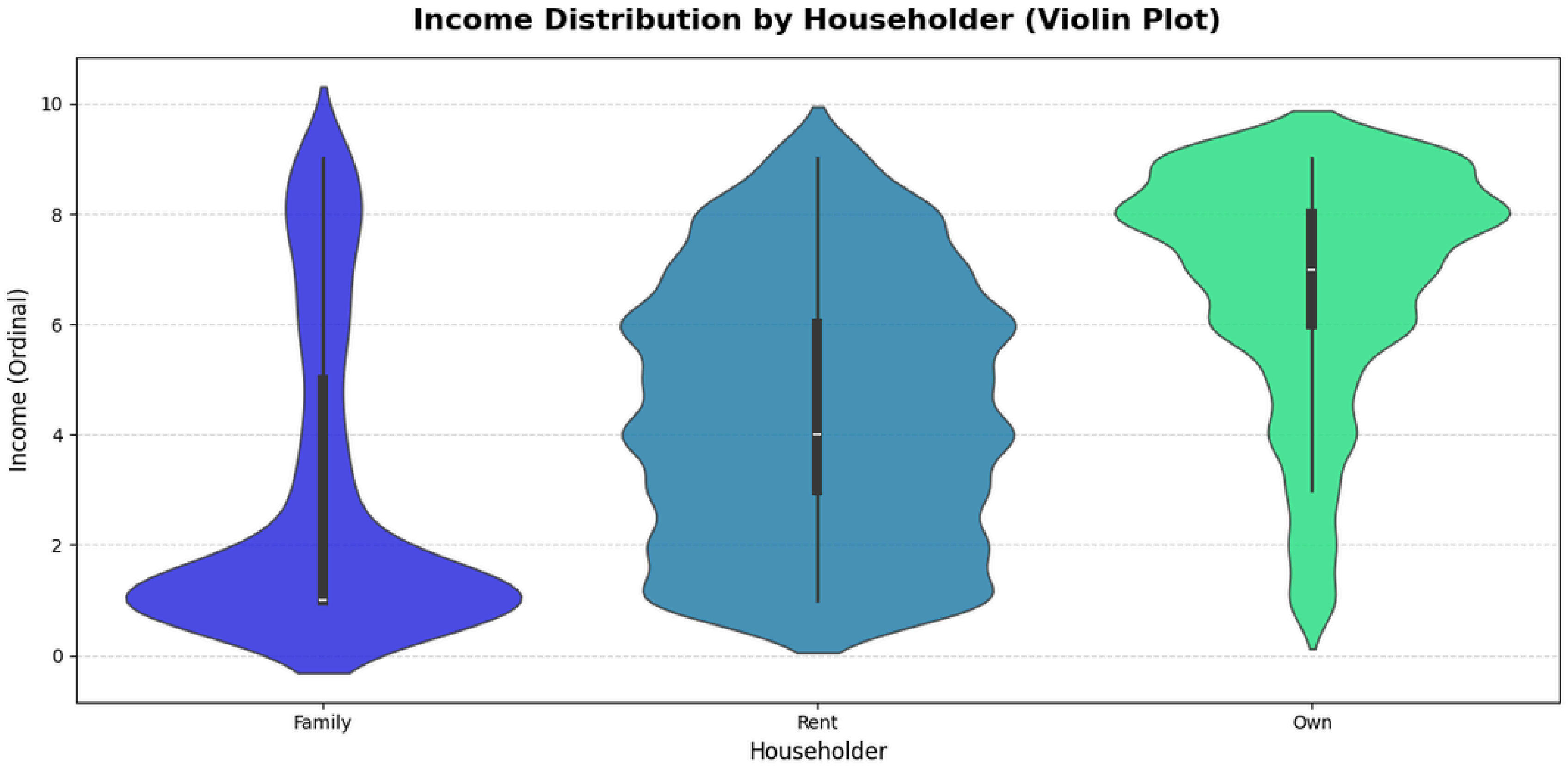
The graph illustrates a clear trend where individuals with higher education levels tend to have higher incomes.

Point 02

As education level increases, it opens up more opportunities for better-paying jobs, reflecting the correlation between advanced qualifications and increased earning potential.

INCOME DISTRIBUTION BY HOME TYPE AND HOUSEHOLDER





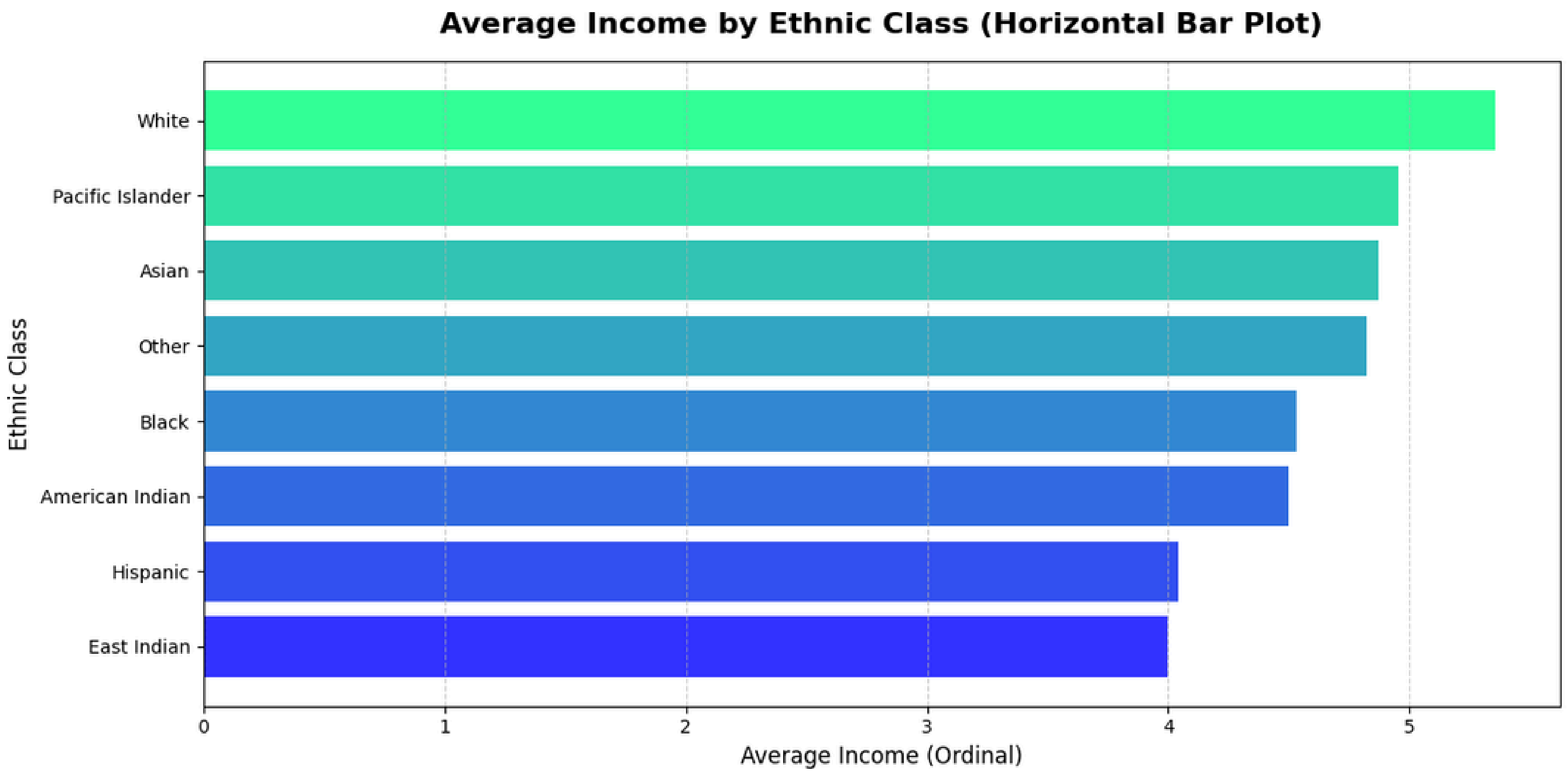
Home Type

The graph shows similar median incomes for apartments and mobile homes, with mobile homes slightly higher. Condominium and house owners have higher average incomes, linking income to housing type.

Householder

The graph shows those living with family have lower average incomes, renters have a wide income range, and homeowners typically have higher, stable incomes.

AVERAGE INCOME BY ETHNICITY



Point 01

The data reveals that White individuals form the majority, followed by Pacific Islanders and Asians.

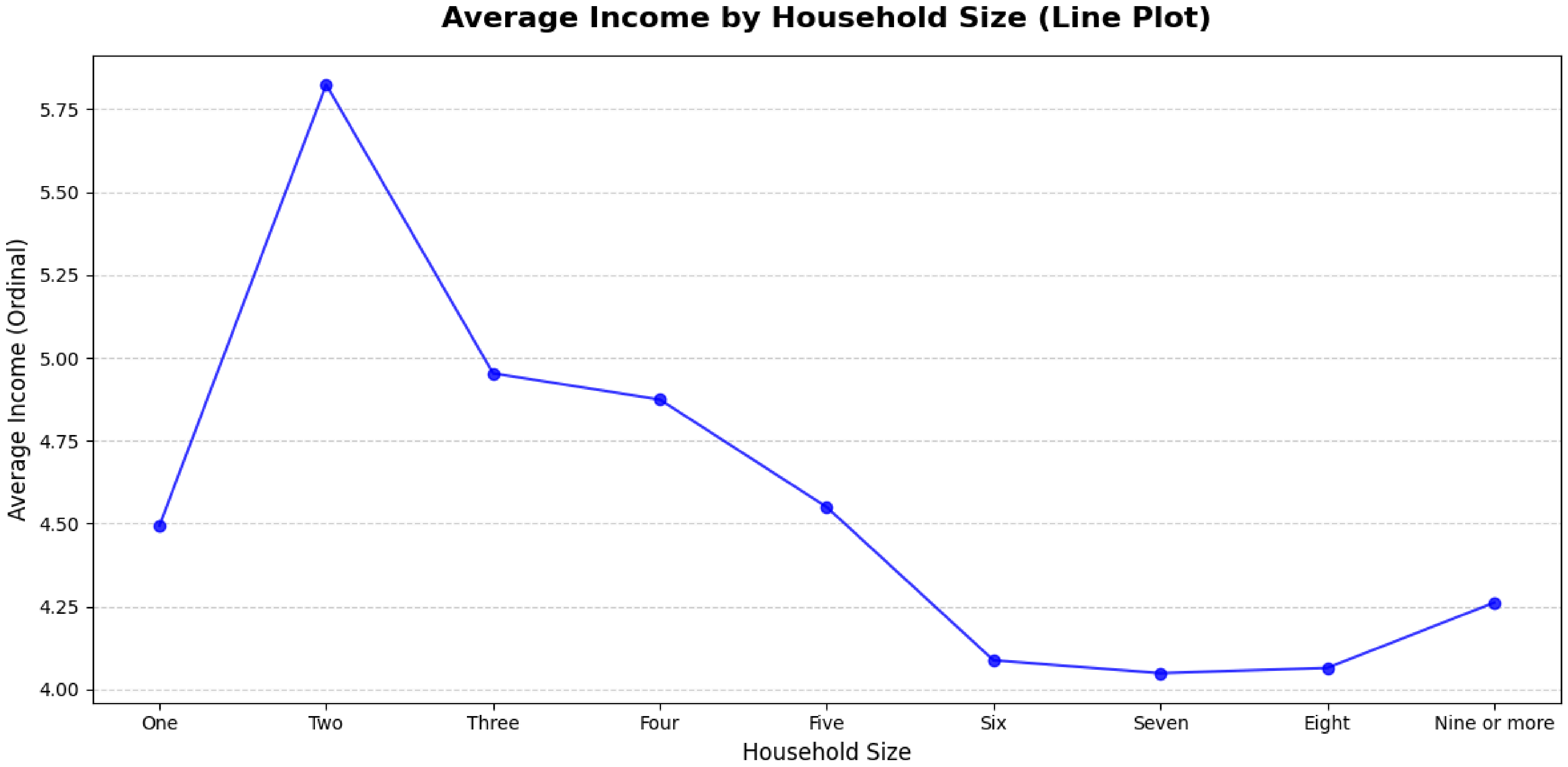
Point 02

Hispanic, American Indian, and Black individuals have lower representation, with West Indians being the least represented.

Point 03

This disparity may stem from socioeconomic factors like economic background and education.

AVERAGE INCOME BY HOUSEHOLD SIZE

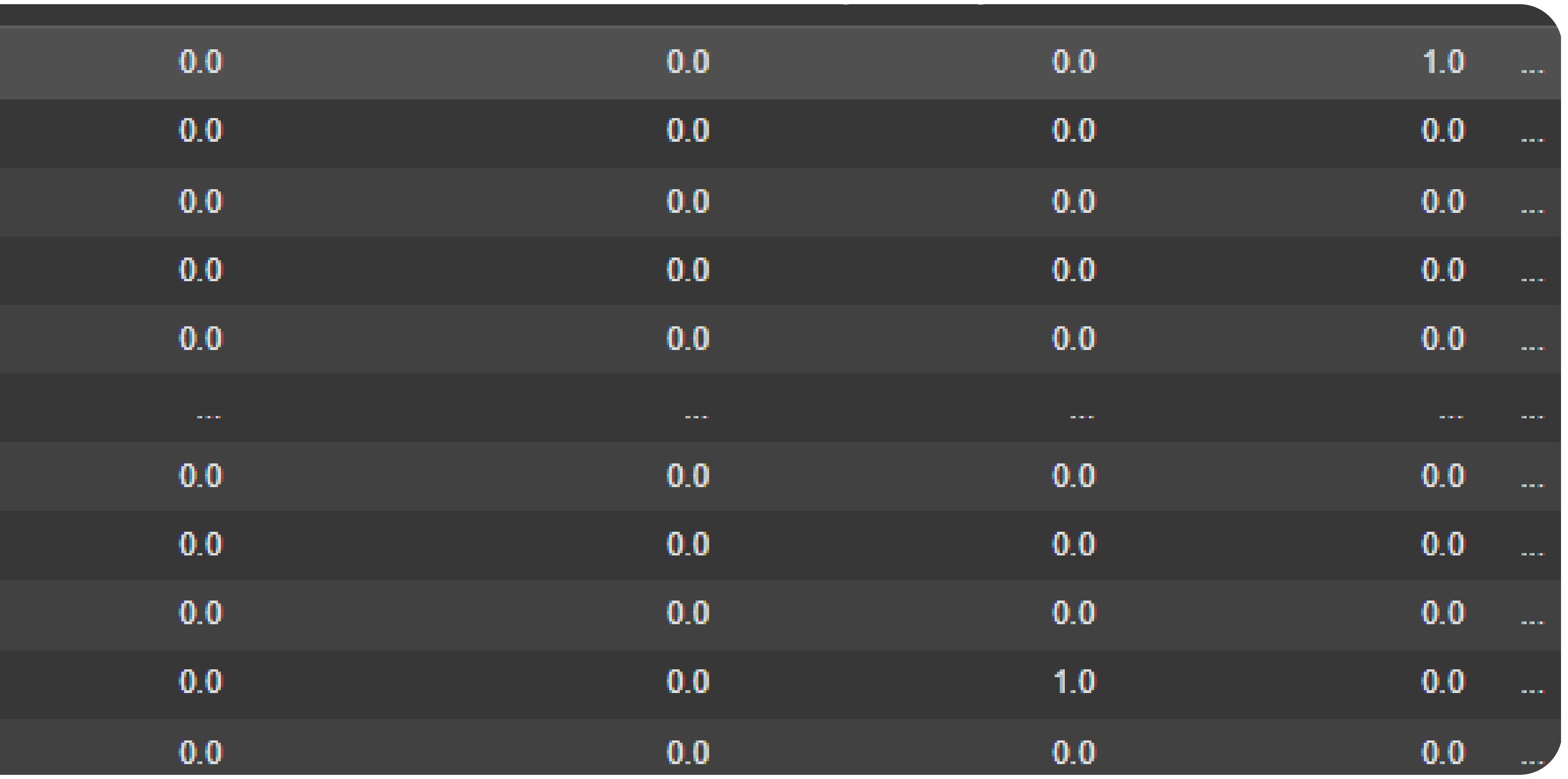


Point 01

Households with 2 members have the highest average income, while income decreases as household size increases, reaching the lowest in 7-member households.

Point 02

Larger households may experience lower incomes due to greater economic burdens. The dataset also reveals a significant drop in data representation starting from households with 3 members.



0.0	0.0	0.0	1.0	...
0.0	0.0	0.0	0.0	...
0.0	0.0	0.0	0.0	...
0.0	0.0	0.0	0.0	...
0.0	0.0	0.0	0.0	...
...	...	...	...	...
0.0	0.0	0.0	0.0	...
0.0	0.0	0.0	0.0	...
0.0	0.0	0.0	0.0	...
0.0	0.0	1.0	0.0	...
0.0	0.0	0.0	0.0	...

ENCODING

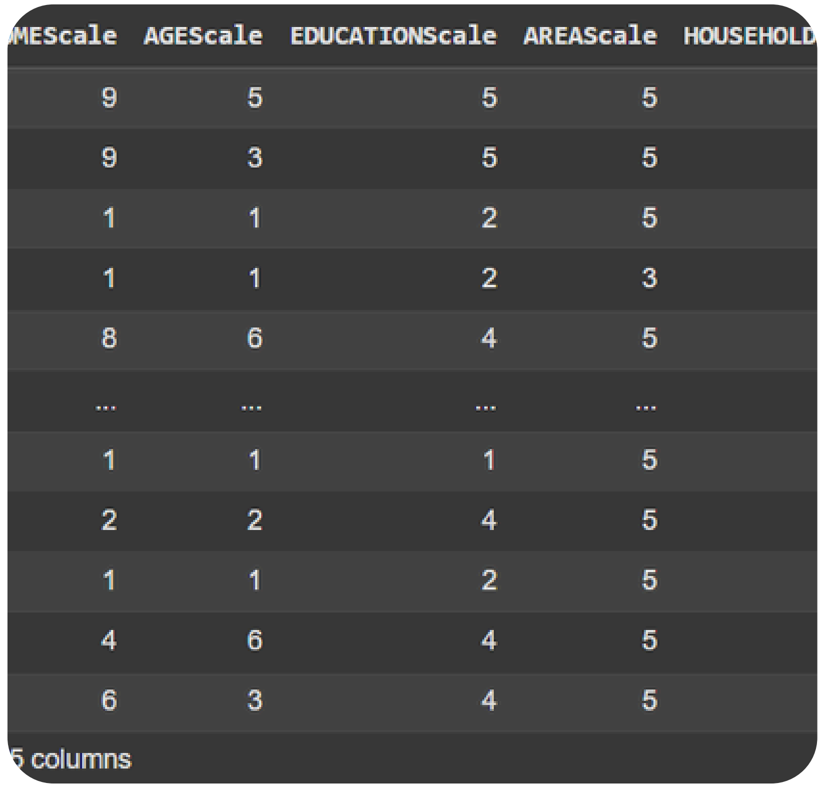
Encoding in machine learning converts categorical data into numerical formats, making it usable for algorithms. Common methods include one-hot encoding, which creates binary columns for each category, and ordinal encoding, which assigns ordered numerical values.

Ordinal

The income, age, education, area (experience years), and household size columns were converted from object to integer types by assigning integers to each category, enabling sorting.

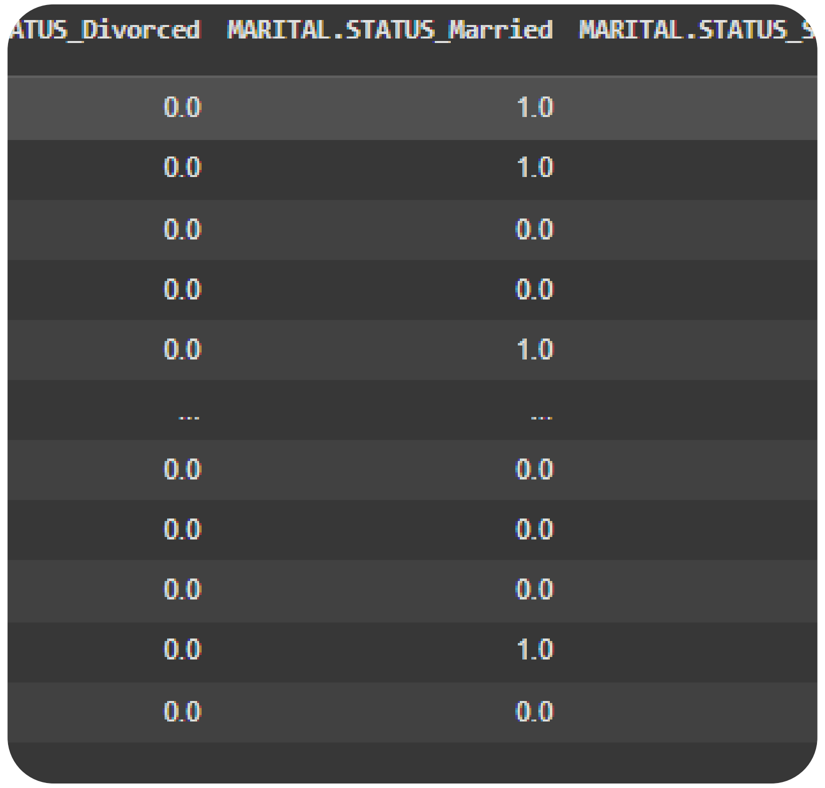
Onehot

The columns for gender, marital status, occupation, dual income, householder, house type, ethnic class, and language were converted from object to integer types, as their data cannot be sorted.



MEScale	AGEScale	EDUCATIONScale	AREAScale	HOUSEHOLD
9	5	5	5	
9	3	5	5	
1	1	2	5	
1	1	2	3	
8	6	4	5	
...	...	...	...	
1	1	1	5	
2	2	4	5	
1	1	2	5	
4	6	4	5	
6	3	4	5	

5 columns



ATUS_Divorced	MARITAL.STATUS_Married	MARITAL.STATUS_S
0.0	1.0	
0.0	1.0	
0.0	0.0	
0.0	0.0	
0.0	1.0	
...	...	
0.0	0.0	
0.0	0.0	
0.0	0.0	
0.0	1.0	
0.0	0.0	

SEHOLDER	...	ETHNIC.CLASS_Asian	ET
Own	...	0.0	
Rent	...	0.0	
Family	...	0.0	
Family	...	0.0	
Own	...	0.0	
...	...	...	
Family	...	0.0	
Family	...	0.0	
Family	...	0.0	
Rent	...	0.0	
Rent	...	0.0	

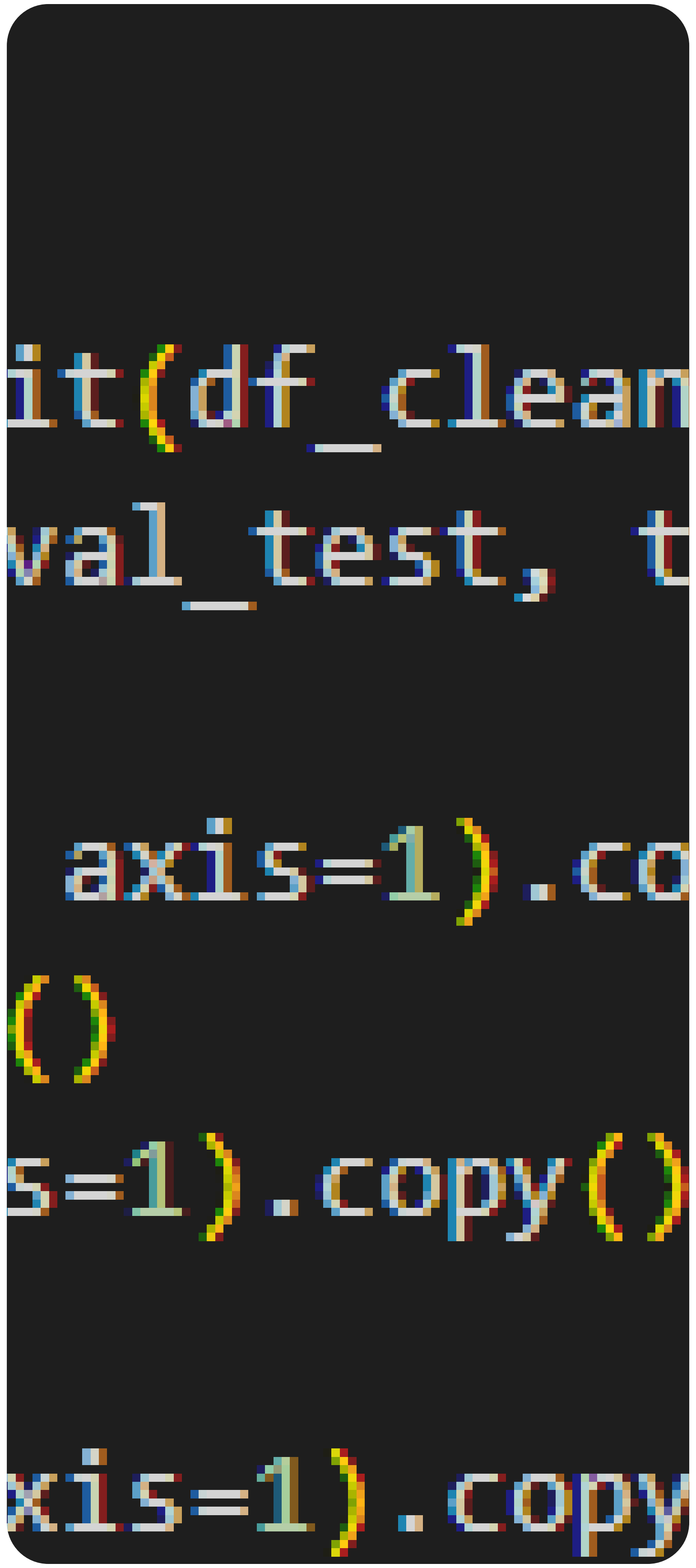
MERGE & SPLIT DATA

Merging

Combine the cleaned and encoded data, then remove columns with object data types, such as income, sex, marital status, age, education, occupation, experience, dual income, household size, householder, home type, ethnic class, and language.

Splitting

Divide the combined data into three parts, namely training data, validation data and testing data.



```
it(df_clean
val_test, t

 axis=1).co
()
s=1).copy()

xis=1).copy
```

K-FOLD VALIDATION

Grid search finds the best hyperparameter combination based on the highest evaluation score to optimize model performance. Results may vary depending on the dataset used.

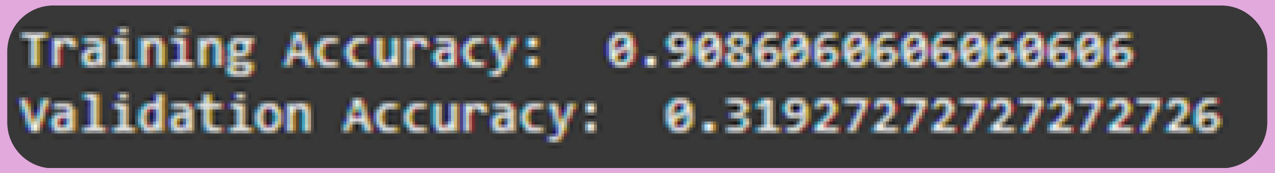
Grid Search Result

```
{'min_samples_leaf': 6, 'min_samples_split': 2, 'n_estimators': 100}
```

Before K-Fold

Validation Score

32%

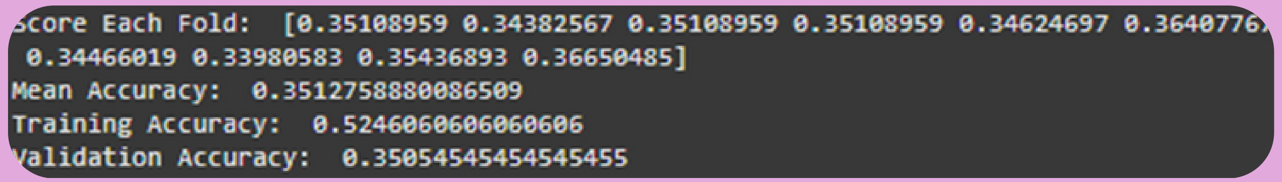


```
Training Accuracy:   0.9086060606060606
Validation Accuracy:   0.3192727272727272 6
```

After K-Fold

Validation Score

35%



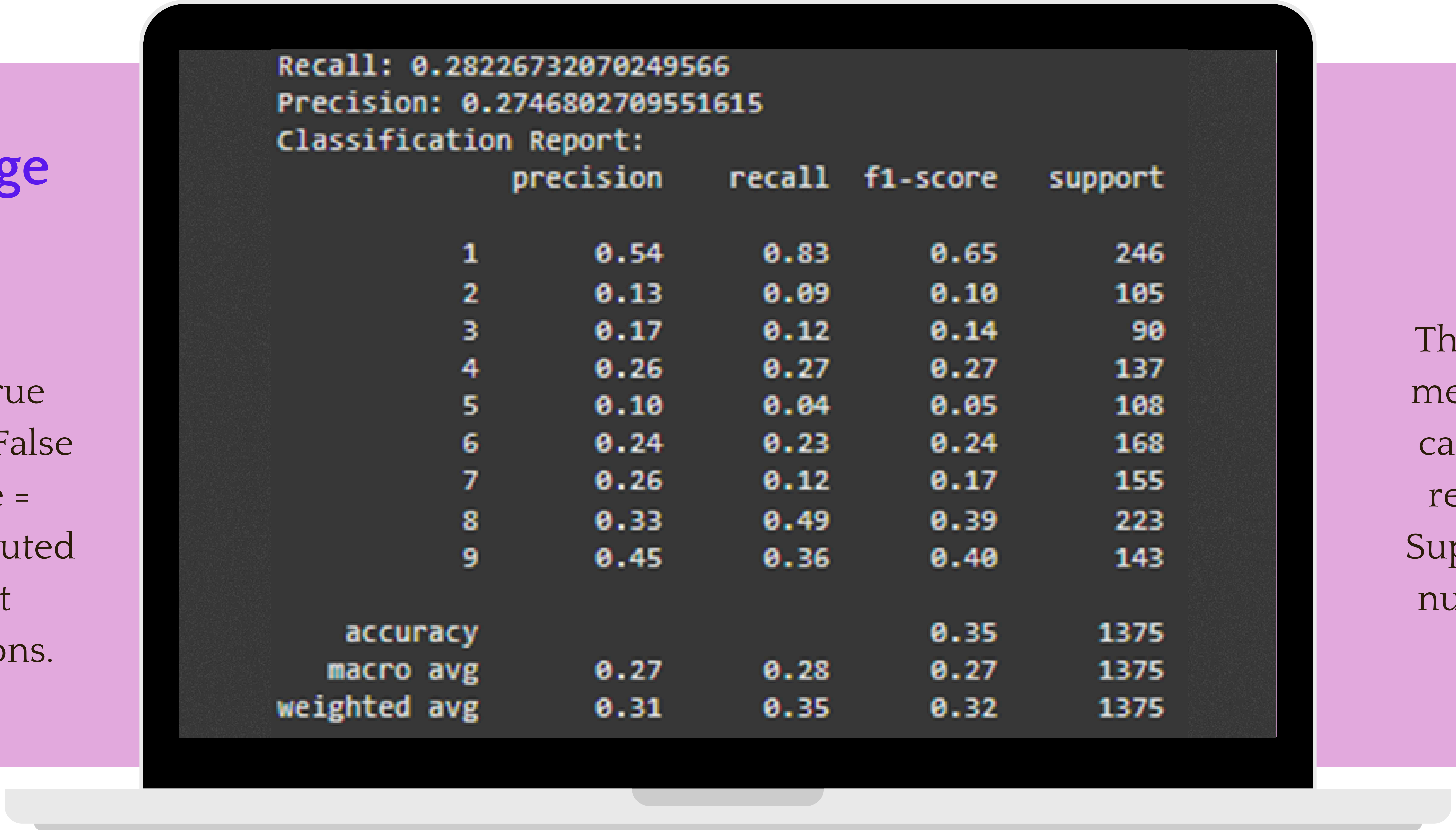
```
Score Each Fold:  [0.35108959 0.34382567 0.35108959 0.35108959 0.34624697 0.3640776
 0.34466019 0.33980583 0.35436893 0.36650485]
Mean Accuracy:  0.3512758880086509
Training Accuracy:  0.5246060606060606
Validation Accuracy:  0.3505454545454545 5
```

In K-Fold Cross Validation, hyperparameters optimized through Grid Search are used. The score of each fold represents the accuracy of training data split into 10 folds, while average accuracy shows the mean accuracy across all folds. Training and validation accuracy indicate how well the model learns and performs, with the possibility of overfitting.

RECALL AND PRECISION

Recall and Average Calculation

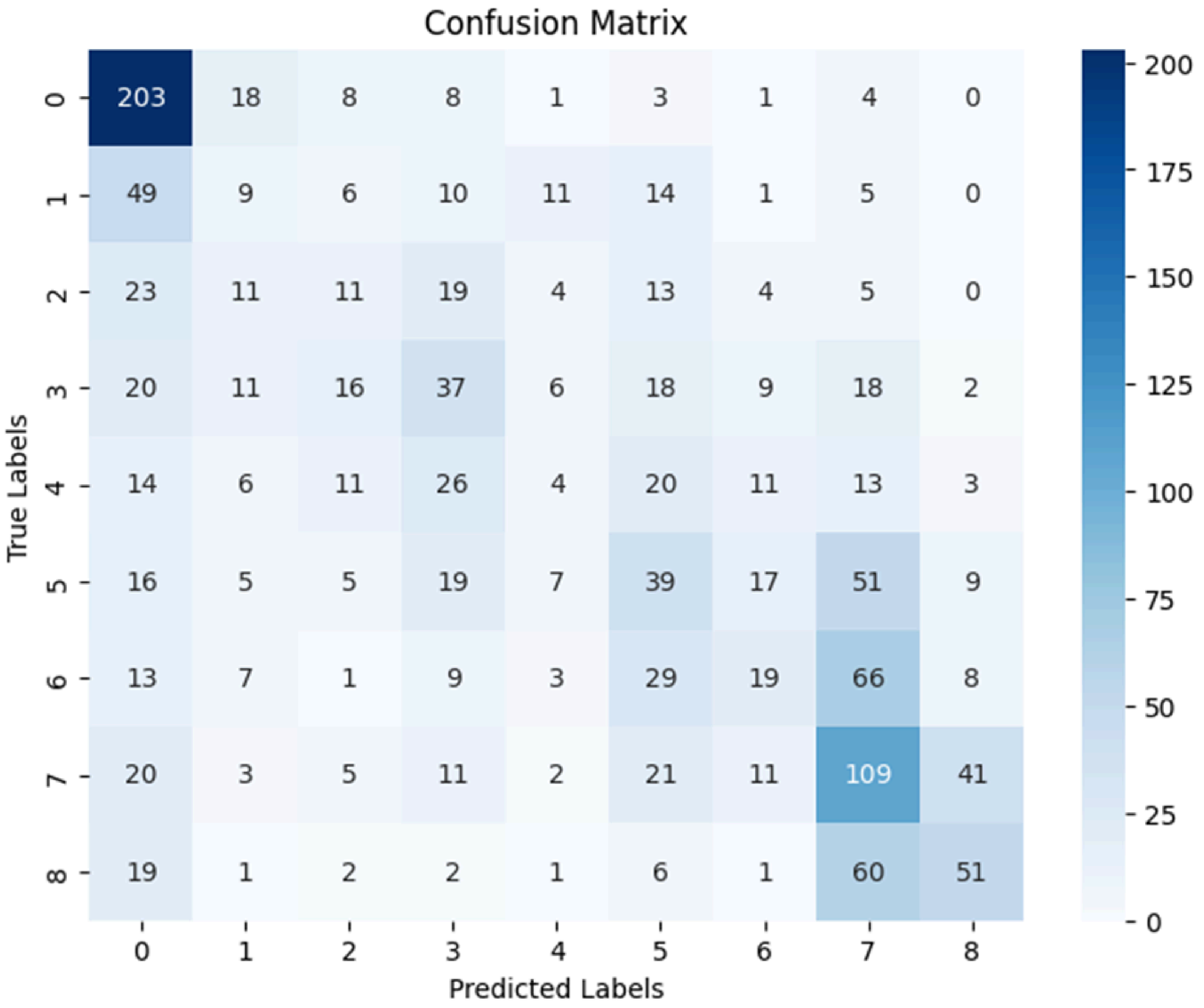
Recall measures model sensitivity, calculated as True Positives / (True Positives + False Negatives). When average = 'macro,' the average is computed across all classes without considering their proportions.



```
Recall: 0.28226732070249566
Precision: 0.2746802709551615
Classification Report:
              precision    recall  f1-score   support

           1       0.54      0.83      0.65       246
           2       0.13      0.09      0.10       105
           3       0.17      0.12      0.14        90
           4       0.26      0.27      0.27       137
           5       0.10      0.04      0.05       108
           6       0.24      0.23      0.24       168
           7       0.26      0.12      0.17       155
           8       0.33      0.49      0.39       223
           9       0.45      0.36      0.40       143

    accuracy                           0.35      1375
   macro avg       0.27      0.28      0.27      1375
weighted avg       0.31      0.35      0.32      1375
```

F1 Score and Support

The F1 score is the harmonic mean of precision and recall, calculated as 2 * (precision * recall) / (precision + recall). Support represents the actual number of samples for each class.



	0	1	2	3	4	5	6	7	8
0	203	18	8	8	1	3	1	4	0
1	49	9	6	10	11	14	1	5	0
2	23	11	11	19	4	13	4	5	0
3	20	11	16	37	6	18	9	18	2
4	14	6	11	26	4	20	11	13	3
5	16	5	5	19	7	39	17	51	9
6	13	7	1	9	3	29	19	66	8
7	20	3	5	11	2	21	11	109	41
8	19	1	2	2	1	6	1	60	51

CONFUSION MATRIX

Point 01

The main diagonal of the matrix (203, 9, 11, 37, 4, 39, 19, 109, 51) represents the model's accuracy for each class. For example, 203 in class 0 indicates strong performance, while 4 in class 4 shows difficulty in prediction.

Point 02

In class 7, the number 109 indicates relatively good performance, suggesting the model can predict this class with moderate accuracy.

PREDICTION TEST

Dataset Information

- Income: Shows the income per person.
- Sex: Shows the gender of the person.
- Marital status: Shows the person's marital status.
- Age: Shows the age group of the person.
- Education: The person's highest level of education.
- Occupation: A person's job or profession.
- Area: Shows the duration of time lived in each person's area.
- Dual Income: Shows when people do more than one source of income.
- Household Size: Shows the number of people living in the household.
- Household: Indicates whether the individual lives in their family’s, owns or rents a home.
- Home Type: Shows the person's home type.
- Ethnic Class: Shows each person's racial ethnicity.
- Language: Shows the person’s daily-used language.

TEST CASES

Test 01

Age: 25-34
Education: College Graduate
Experience: 1-3 years
Household Size: Two
Gender: Male
Marital Status: Together
Occupation: Homemaker
Dual Income: No
Householder Category: Renter
Home Type: Apartment
Ethnicity: White
Language: English

Predicted Income: $30,000 - $40,000

Test 02

Age: 14-17
Education: Grade 9 to 11
Experience: Less than 1 year
Household Size: Four
Gender: Female
Marital Status: Single
Occupation: Student (High School or College)
Dual Income: No
Householder Category: Family
Home Type: House
Ethnicity: Asian
Language: Other

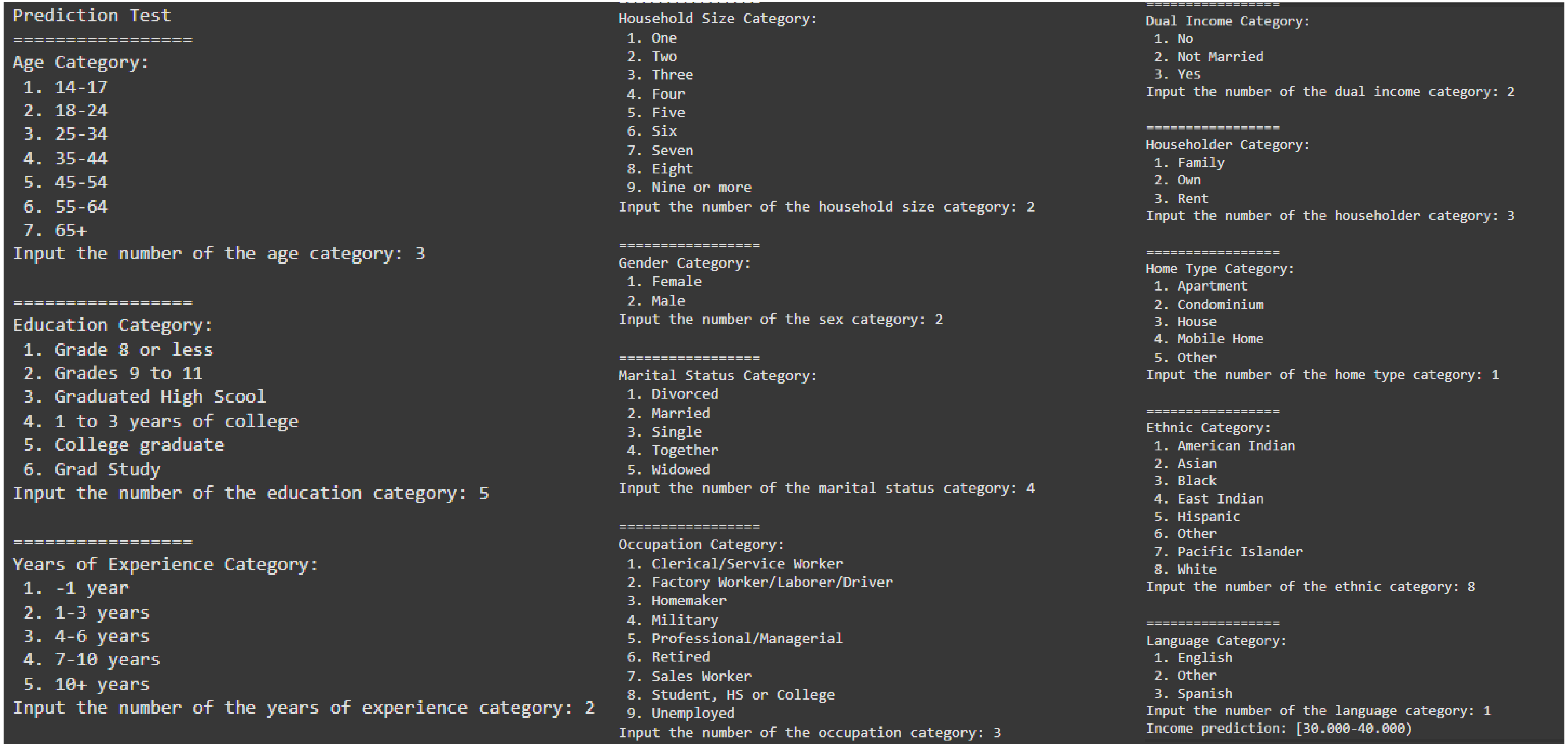
Predicted Income: Below $10,000

Test 03

Age: 55-64
Education: Graduate Studies
Experience: 10+ years
Household Size: Six
Gender: Male
Marital Status: Widowed
Occupation: Professional/Managerial
Dual Income: No
Householder Category: Homeowner
Home Type: Condominium
Ethnicity: Hispanic
Language: Spanish

Predicted Income: $75,000+

TEST 1

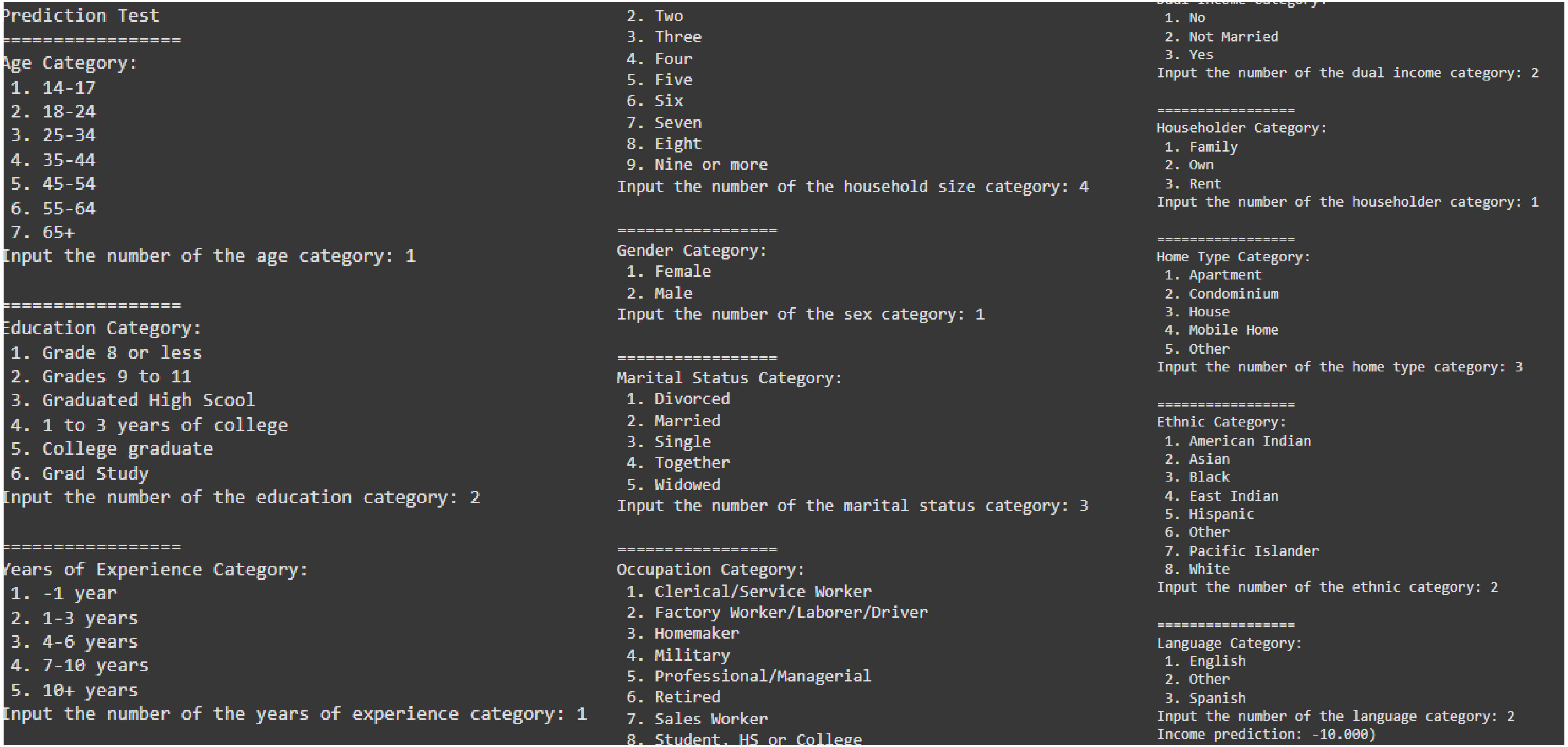


```
Prediction Test
=================
Age Category:
 1. 14-17
 2. 18-24
 3. 25-34
 4. 35-44
 5. 45-54
 6. 55-64
 7. 65+
Input the number of the age category: 3

=================
Education Category:
 1. Grade 8 or less
 2. Grades 9 to 11
 3. Graduated High Scool
 4. 1 to 3 years of college
 5. College graduate
 6. Grad Study
Input the number of the education category: 5

=================
Years of Experience Category:
 1. -1 year
 2. 1-3 years
 3. 4-6 years
 4. 7-10 years
 5. 10+ years
Input the number of the years of experience category: 2
```

```
=================
Household Size Category:
 1. One
 2. Two
 3. Three
 4. Four
 5. Five
 6. Six
 7. Seven
 8. Eight
 9. Nine or more
Input the number of the household size category: 2

=================
Gender Category:
 1. Female
 2. Male
Input the number of the sex category: 2

=================
Marital Status Category:
 1. Divorced
 2. Married
 3. Single
 4. Together
 5. Widowed
Input the number of the marital status category: 4

=================
Occupation Category:
 1. Clerical/Service Worker
 2. Factory Worker/Laborer/Driver
 3. Homemaker
 4. Military
 5. Professional/Managerial
 6. Retired
 7. Sales Worker
 8. Student, HS or College
 9. Unemployed
Input the number of the occupation category: 3
```

```
=================
Dual Income Category:
 1. No
 2. Not Married
 3. Yes
Input the number of the dual income category: 2

=================
Householder Category:
 1. Family
 2. Own
 3. Rent
Input the number of the householder category: 3

=================
Home Type Category:
 1. Apartment
 2. Condominium
 3. House
 4. Mobile Home
 5. Other
Input the number of the home type category: 1

=================
Ethnic Category:
 1. American Indian
 2. Asian
 3. Black
 4. East Indian
 5. Hispanic
 6. Other
 7. Pacific Islander
 8. White
Input the number of the ethnic category: 8

=================
Language Category:
 1. English
 2. Other
 3. Spanish
Input the number of the language category: 1
Income prediction: [30.000-40.000)
```

TEST 2

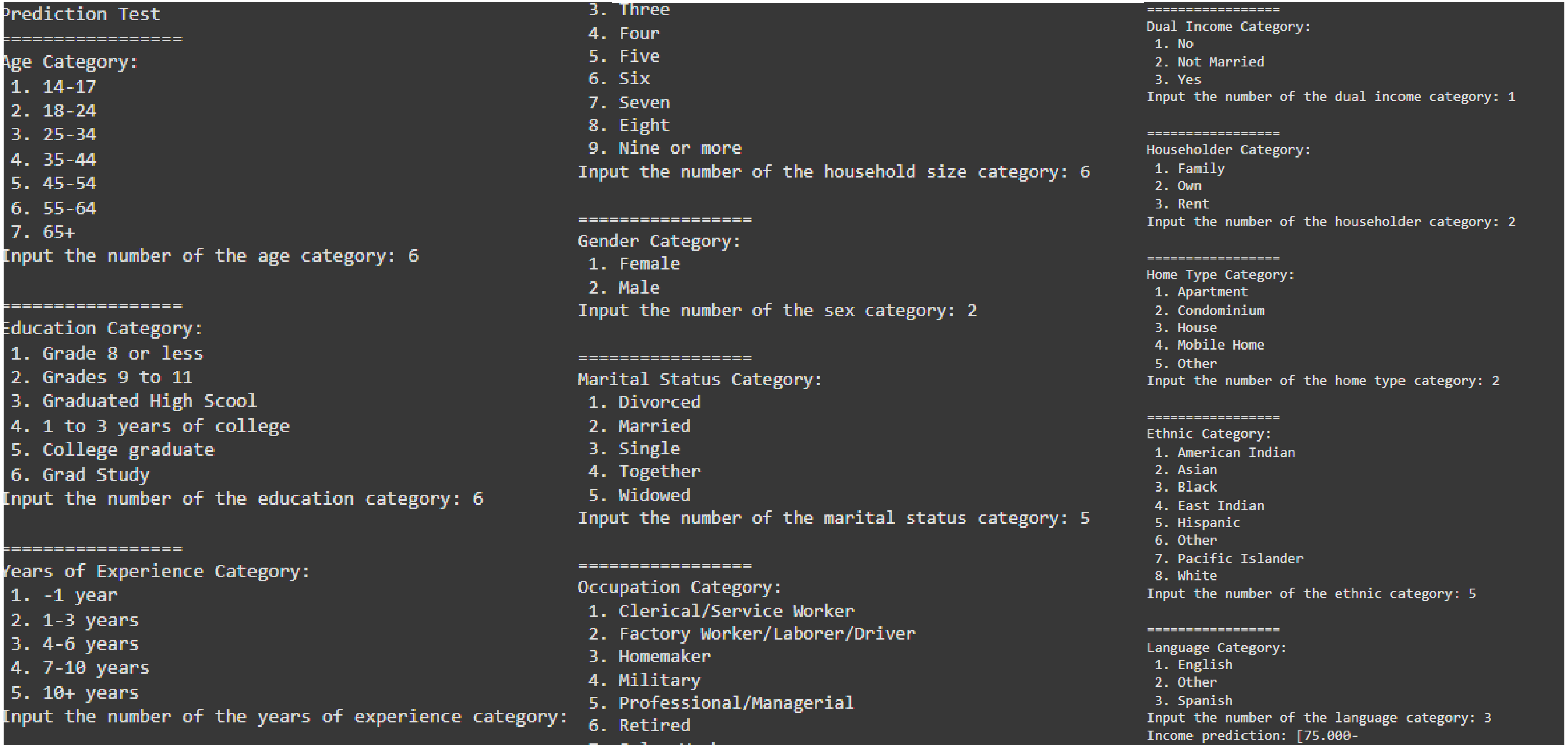


```
Prediction Test
=================
Age Category:
 1. 14-17
 2. 18-24
 3. 25-34
 4. 35-44
 5. 45-54
 6. 55-64
 7. 65+
Input the number of the age category: 1

=================
Education Category:
 1. Grade 8 or less
 2. Grades 9 to 11
 3. Graduated High Scool
 4. 1 to 3 years of college
 5. College graduate
 6. Grad Study
Input the number of the education category: 2

=================
Years of Experience Category:
 1. -1 year
 2. 1-3 years
 3. 4-6 years
 4. 7-10 years
 5. 10+ years
Input the number of the years of experience category: 1
```

```
 2. Two
 3. Three
 4. Four
 5. Five
 6. Six
 7. Seven
 8. Eight
 9. Nine or more
Input the number of the household size category: 4

=================
Gender Category:
 1. Female
 2. Male
Input the number of the sex category: 1

=================
Marital Status Category:
 1. Divorced
 2. Married
 3. Single
 4. Together
 5. Widowed
Input the number of the marital status category: 3

=================
Occupation Category:
 1. Clerical/Service Worker
 2. Factory Worker/Laborer/Driver
 3. Homemaker
 4. Military
 5. Professional/Managerial
 6. Retired
 7. Sales Worker
 8. Student, HS or College
```

```
Dual Income Category:
 1. No
 2. Not Married
 3. Yes
Input the number of the dual income category: 2

=================
Householder Category:
 1. Family
 2. Own
 3. Rent
Input the number of the householder category: 1

=================
Home Type Category:
 1. Apartment
 2. Condominium
 3. House
 4. Mobile Home
 5. Other
Input the number of the home type category: 3

=================
Ethnic Category:
 1. American Indian
 2. Asian
 3. Black
 4. East Indian
 5. Hispanic
 6. Other
 7. Pacific Islander
 8. White
Input the number of the ethnic category: 2

=================
Language Category:
 1. English
 2. Other
 3. Spanish
Input the number of the language category: 2
Income prediction: -10.000)
```

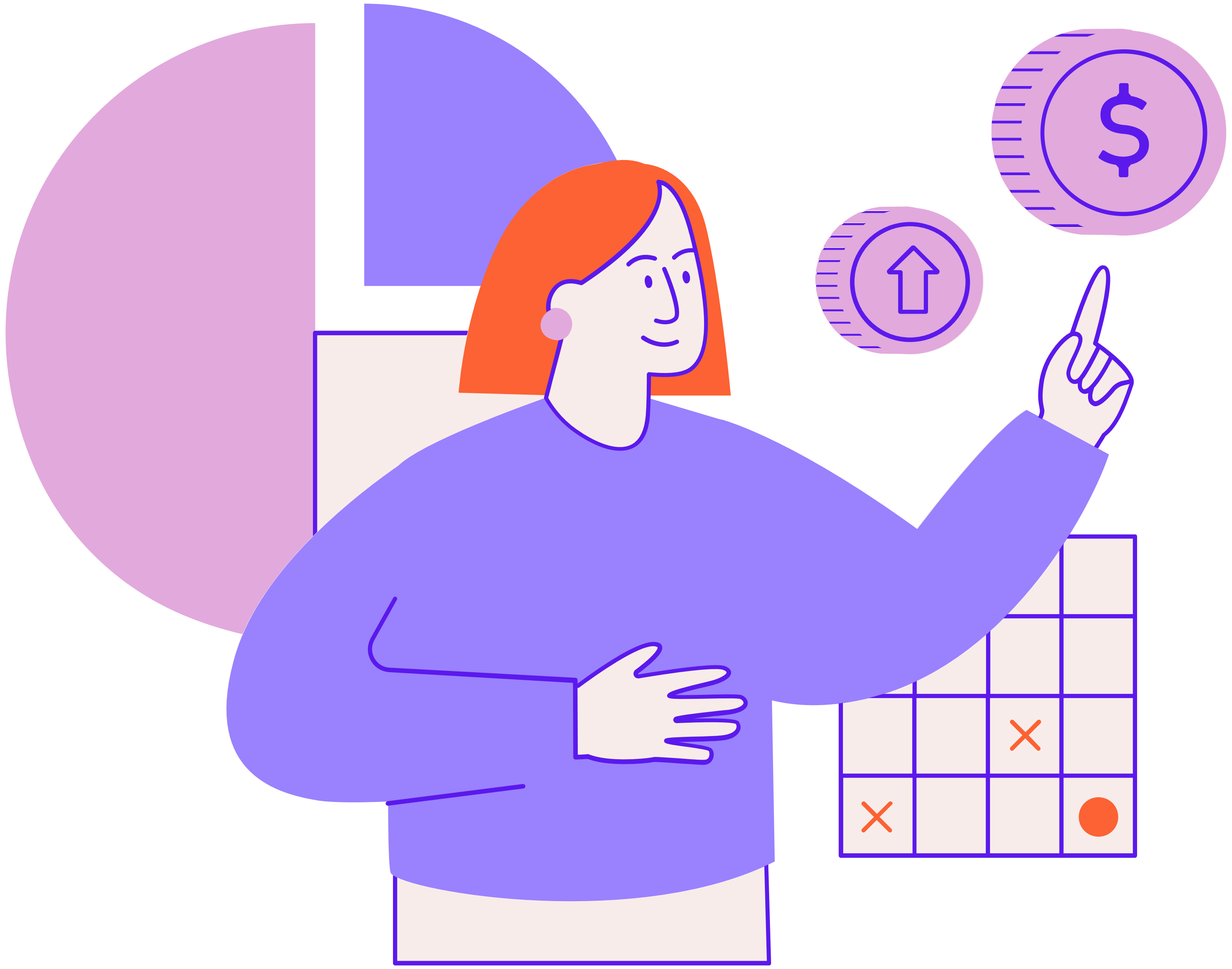
TEST 3



```
Prediction Test
=================
Age Category:
 1. 14-17
 2. 18-24
 3. 25-34
 4. 35-44
 5. 45-54
 6. 55-64
 7. 65+
Input the number of the age category: 6

=================
Education Category:
 1. Grade 8 or less
 2. Grades 9 to 11
 3. Graduated High Scool
 4. 1 to 3 years of college
 5. College graduate
 6. Grad Study
Input the number of the education category: 6

=================
Years of Experience Category:
 1. -1 year
 2. 1-3 years
 3. 4-6 years
 4. 7-10 years
 5. 10+ years
Input the number of the years of experience category:
```

```
 3. Three
 4. Four
 5. Five
 6. Six
 7. Seven
 8. Eight
 9. Nine or more
Input the number of the household size category: 6

=================
Gender Category:
 1. Female
 2. Male
Input the number of the sex category: 2

=================
Marital Status Category:
 1. Divorced
 2. Married
 3. Single
 4. Together
 5. Widowed
Input the number of the marital status category: 5

=================
Occupation Category:
 1. Clerical/Service Worker
 2. Factory Worker/Laborer/Driver
 3. Homemaker
 4. Military
 5. Professional/Managerial
 6. Retired
```

```
=================
Dual Income Category:
 1. No
 2. Not Married
 3. Yes
Input the number of the dual income category: 1

=================
Householder Category:
 1. Family
 2. Own
 3. Rent
Input the number of the householder category: 2

=================
Home Type Category:
 1. Apartment
 2. Condominium
 3. House
 4. Mobile Home
 5. Other
Input the number of the home type category: 2

=================
Ethnic Category:
 1. American Indian
 2. Asian
 3. Black
 4. East Indian
 5. Hispanic
 6. Other
 7. Pacific Islander
 8. White
Input the number of the ethnic category: 5

=================
Language Category:
 1. English
 2. Other
 3. Spanish
Input the number of the language category: 3
Income prediction: [75.000-
```



CONCLUSION

Point 01

Income levels are influenced by multiple factors, with age being a key determinant, as it reflects the accumulation of work experience.

Point 02

Education also plays a significant role, as higher education is often associated with greater job proficiency and the ability to take on professional roles.

Point 03

Married individuals tend to have higher incomes, likely due to the combined financial resources within the household.

Point 04

Mastering these factors, age, education, and marital status can greatly enhance one's economic prospects.

THANK YOU.