INT375: DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING

PROJECT REPORT

(Project Semester January-April 2025)

Exploratory Data Analysis of American Community Survey 2023 1-Year PUMS

Submitted by

Savinay Singh

Registration No. 12308126

Programme and Section: B.Tech CSE K23DW

Course Code: INT375

Under the Guidance of

Vikash Mangotra

UID: 31488

Assistant Professor

Discipline of CSE/IT

Lovely School of Computer Science and Engineering

Lovely Professional University, Phagwara

DECLARATION

I, Savinay Singh, student of B.Tech CSE under the CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report for the course INT375 (DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING) is based on my own intensive work and is genuine.

Date: 12 April 2025

Registration No. 12308126

Signature

Savinay Singh

CERTIFICATE

This is to certify that Savinay Singh bearing Registration No. 12308126 has completed the INT375 project titled, “Exploratory Data Analysis of American Community Survey 2023 1-Year PUMS” under my guidance and supervision. To the best of my knowledge, the present work is the result of his original development, effort and study.

Signature and Name of the Supervisor
Vikash Mangotra
Designation of the Supervisor: Faculty Coordinator
School of Computer Science and Engineering
Lovely Professional University
Phagwara, Punjab.

Date: 12 April 2025

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my supervisor, Mr. Vikash Mangotra, for the invaluable guidance, support, and encouragement throughout this minor project. His insights and feedback were instrumental in the exploration process and the successful completion of this work.

I am also thankful to Lovely Professional University and the School of Computer Science and Engineering for providing the necessary resources and academic environment conducive to this project.

Table of Contents

1. Introduction

This report presents a comprehensive Exploratory Data Analysis (EDA) performed on the **American Community Survey (ACS) 2023 1-Year Public Use Microdata Sample (PUMS)**. The dataset, originating from the U.S. Census Bureau, represents a vital resource for understanding the characteristics of the U.S. population. It provides a rich source of anonymized individual-level data covering a wide array of demographic, socio-economic, and housing characteristics across the United States, forming the basis for numerous research and policy decisions.

The primary objective of this project, undertaken as part of the **INT375 Data Science Toolbox: Python Programming** course, was to apply foundational data science techniques using Python for thorough data exploration. Exploratory Data Analysis is a critical initial step in any data-driven project. Its purpose is to understand the data's underlying structure, identify potential quality issues, uncover initial patterns and relationships, detect anomalies or outliers, and check assumptions before more formal modeling or hypothesis testing. This project specifically involved understanding the ACS PUMS structure, identifying and systematically addressing data quality issues (such as missing values and special codes), reducing dimensionality through strategic feature selection, and ultimately, uncovering preliminary patterns within the data using descriptive statistics, informative visualizations, and basic inferential tests.

Through this structured EDA, we aim to gain preliminary insights into the characteristics of the individuals represented in this specific sample. The analysis focuses on key variables including age, income, education level, gender, employment status, geographic region, and health insurance coverage, exploring their distributions and potential interconnections. The findings presented herein document the analytical journey, detail the data transformations performed, and highlight key observations. This work serves as a foundation for understanding the dataset's potential and limitations, paving the way for subsequent, more focused research or modeling tasks. This report details the methodology employed, the source and nature of the data, the step-by-step EDA process,

the analytical results derived from the cleaned subset, key visualizations, and finally, offers conclusions and potential avenues for future work.

2. Source of Dataset

The data utilized in this analysis is the **American Community Survey (ACS) 2023 1-Year Public Use Microdata Sample (PUMS)**. It was accessed as a CSV file named census_data.csv. The ACS PUMS is meticulously compiled and distributed annually by the **U.S. Census Bureau**.

PUMS files differ significantly from the pre-tabulated summary tables often released from the ACS. They contain records for a sample of individual housing units and the people within them, allowing data users immense flexibility to create custom tabulations and analyses tailored to specific research questions. These microdata records have undergone rigorous processing by the Census Bureau, including anonymization techniques and data swapping, to protect the confidentiality of survey respondents while maintaining statistical validity.

It is fundamentally important to recognize that the ACS is a survey, not a complete census count. Consequently, PUMS data represents a **sample** of the U.S. population and housing units. To generate statistically sound estimates that represent the characteristics of the *entire* population (e.g., average income for a state), users must apply the appropriate survey weights provided within the PUMS files. For analyses focused on individuals, the **Person Weight (PWGTP)** is used, while analyses of housing units utilize the **Housing Weight (WGTP)**. These weights account for the complex sampling design and non-response adjustments. **This particular EDA project, however, focused specifically on exploring the characteristics, distributions, and relationships present within the *sample data itself*. Therefore, survey weights were not applied in this analysis.** The results and conclusions pertain directly to the 6868 individuals in the cleaned subset, and direct generalization to the broader U.S. population requires the aforementioned weighting procedures.

Further details regarding the ACS PUMS, including methodologies, variable definitions (data dictionaries), data access options (like the FTP site or the online Microdata Analysis Tool - MDAT), user guides, and crucially, information on calculating margins of error to assess estimate reliability, can be found on the official ACS PUMS webpage: https://www.census.gov/programs-surveys/acs/microdata.html and its associated documentation pages.

3. EDA Process

The Exploratory Data Analysis followed a structured workflow executed within a Jupyter Notebook environment using Python and standard data science libraries (Pandas, NumPy, Matplotlib, Seaborn, SciPy, Statsmodels). The rationale and details of each step are outlined below:

Data Loading and Initial Overview:

The `census_data.csv` file was loaded into a Pandas DataFrame. Initial exploratory steps included examining the DataFrame's dimensions (`.shape`), reviewing data types and non-null counts (`.info()`), observing the first few rows (`.head()`), and getting a preliminary count of missing values per column (`.isnull().sum()`).

The initial .info() summary showed a mix of float, integer, and object data types, immediately highlighting the need for type review and potential cleaning. The initial missing value check revealed numerous columns with missing data, requiring a systematic approach.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6868 entries, 0 to 6867
Columns: 206 entries, record_type to immigration_year_flag
dtypes: float64(84), int64(118), object(4)
memory usage: 10.8+ MB
None
```

```
  record_type   household_id  census_division  person_id  puma_area  \
0           P  2023GQ0000108                9          1        101
1           P  2023GQ0000169                9          1        102
2           P  2023GQ0000878                9          1        300
3           P  2023GQ0001081                9          1        300
4           P  2023GQ0001269                9          1        400

   census_region  state_code  income_adjustment  age_years  \
0              4           2            1019518         71
1              4           2            1019518         90
2              4           2            1019518         90
3              4           2            1019518         22
4              4           2            1019518         49

   citizenship_status  ...  self-employment_income_flag  sex_flag  \
0                   1  ...                            0         0
1                   1  ...                            0         0
2                   1  ...                            0         0
3                   1  ...                            0         0
4                   5  ...                            0         0

   supplementary_security_income_flag  social_security_income_flag  \
0                                   0                            0
1                                   0                            1
2                                   0                            0
...
3                     0
4                     0

[5 rows x 206 columns]
```

```
Missing Values per Column:
 naturalization_year         6643
employment_class             2744
self_care_issue               401
independent_living_issue     1300
mobility_issue                401
                              ...
soc_occupation               2744
veteran_service_period       6150
medicare_coverage_given      5646
medicaid_coverage_given      4698
tricare_coverage_given       6231
Length: 86, dtype: int64
```

Initial Cleaning:

Columns containing replicate weights (identified by names containing 'weight_replicate') were programmatically identified and dropped from the DataFrame, as they are typically used for variance estimation, which was beyond the scope of this basic EDA.

Dropping these columns streamlined the dataset for this EDA's focus on point estimates and relationships within the sample, rather than complex variance calculations. The shape changed from (6868, 206) initially, reducing the column count.

Identification and Handling of Special Codes:

Key numerical variables, such as `poverty_ratio` and `migration_puma_area`, were inspected for high-frequency values that might represent special codes (e.g., 'Not Applicable', 'Topcoded'). Value counts confirmed specific codes (501 in `poverty_ratio`, 70000 in `migration_puma_area`) that were subsequently replaced with `np.nan` to treat them as missing or non-applicable data.

This step is crucial as treating these codes as valid numerical data would severely distort descriptive statistics (like mean, median) and visualizations. For example, the prevalence of code 501 in poverty_ratio required its conversion to NaN to avoid artificially inflating or deflating calculated poverty metrics based on this sample. Replacing with np.nan allows standard missing value handling functions to recognize and process these instances appropriately.

```
poverty_ratio Value Counts:
poverty_ratio
501.0    1878
0.0        61
149.0      35
151.0      33
58.0       32
134.0      31
128.0      30
276.0      30
243.0      30
233.0      27
Name: count, dtype: int64

migration_puma_area Value Counts:
migration_puma_area
400.0      229
100.0      192
200.0      141
300.0      112
1.0         29
2900.0      14
8626.0      13
7300.0      12
...
1110.0     21
4500.0     17
1260.0     12
Name: count, dtype: int64
```

```
Missing Values in Modified Columns:
poverty_ratio          2283
migration_puma_area    5911
dtype: int64
```

Column-Level Missing Value Assessment and Reduction:

The percentage of missing values for every column was calculated. A threshold was set (90%), and columns exceeding this threshold were removed. This step significantly reduced the dimensionality by dropping variables with insufficient data for reliable analysis or imputation (e.g., `naturalization_year`, various veteran service period columns, certain specific income/insurance flags).

The 90% threshold was chosen pragmatically; imputing columns with such a high proportion of missing data (e.g., naturalization_year at ~97% missing) is generally considered unreliable. This step significantly reduced the feature space from over 200 columns to 196, focusing subsequent efforts on variables with more complete information.

```
Columns with >50% Missing Values:
 naturalization_year            96.723937
veteran_disability_rate_2      97.408270
veteran_disability_rate_2.1    88.147932
english_proficiency            82.993593
recent_birth                   79.470006
grandcare_duration             98.281887
grandcare_responsible          96.840419
commute_time                   60.774607
vehicle_occupants              71.782178
transport_mode                 56.974374
post_9_11_service              89.545719
gulf_war_service               89.545719
service_1975_1990              89.545719
vietnam_era_service            89.545719
service_1955_1964              89.545719
korean_war_service             89.545719
pre_1950_service               89.545719
wwii_service                   89.545719
grade_level                    76.878276
immigration_year               92.516016
entry_decade                   92.516016
vehicle_count                  71.782178
parent_employment_status       77.999418
first_degree_field             81.523005
...
medicaid_coverage_given        68.404193
tricare_coverage_given         90.725102
dtype: float64
Shape after dropping high-missing columns: (6868, 196)
```

Feature Selection for Focused Analysis:

A subset of 15 variables deemed central for a preliminary socio-economic profile was selected. These included: `age_years`, `gender`, `education_level`, `total_income_12m`, `employment_status_code`, `census_region`, `race_detail_2_3`, `citizenship_status`, `disability_status`, `marital_status`, `wage_income_12m`, `commute_time`, `health_ins_coverage`, `state_code`, and `puma_area`.

The rationale for selecting these 15 variables was to create a manageable subset covering core demographic (age_years, gender, race_detail_2_3, citizenship_status), socio-economic, education_level, total_income_12m, wage_income_12m, employment_status_code), health (disability_status, health_ins_coverage), and context variables (marital_status, commute_time, census_region, state_code, puma_area). This allows for initial exploration of key relationships without the complexity of the full dataset.

Final Missing Value Imputation:

The remaining missing values (`NaN`s) within the selected 15-variable subset were imputed. Median imputation was chosen for numerical or ordinal variables (`age_years`, `total_income_12m`, `wage_income_12m`, `commute_time`, `education_level`) to minimize the influence of potential outliers. Mode imputation was used for categorical variables. This resulted in a complete dataset for the selected features.

Using the median for numerical variables like total_income_12m is particularly important given the observed skewness and outliers, as the median is less sensitive to extreme values than the mean. Mode imputation is standard practice for categorical features, replacing missing entries with the most frequent category. This resulted in a 6868x15 DataFrame ready for analysis.

```
Rows with >50% missing values: 0
Subset Shape after row filtering: (6868, 15)
```

```
Final Dataset Shape: (6868, 15)
Final Missing Values:
 age_years                 0
gender                     0
education_level            0
total_income_12m           0
employment_status_code     0
census_region              0
race_detail_2_3            0
citizenship_status         0
disability_status          0
marital_status             0
wage_income_12m            0
commute_time               0
health_ins_coverage        0
state_code                 0
puma_area                  0
dtype: int64
```

Cleaned Data Export:

The final, cleaned, and imputed 15-variable dataset was saved as `census_data_cleaned.csv`.

4. Analysis on Dataset

4.1 Introduction

This section elaborates on the analysis performed specifically on the cleaned data subset containing 6868 observations and 15 selected variables. The focus was on understanding the characteristics of the sample represented in this subset and identifying potential relationships between key demographic and socio-economic indicators through descriptive analysis, visualizations, and basic statistical tests.

4.2 General Description

The analysed subset represents 6868 individuals with complete data across the 15 selected variables. The population within this subset exhibits diverse characteristics:

- **Age (`age_years`)**: The age distribution is wide, spanning from infants (0) to seniors (90), with a mean age of approximately 39 years. The standard deviation is considerable (around 23 years), indicating significant age diversity. Normality tests suggest the distribution deviates from a perfect normal curve.

- **Income (`total_income_12m`):** This variable shows substantial positive skewness. The median income of $30,000 (after imputation) is considerably lower than the mean ($47,384), pulled up by high-income earners. Outlier analysis confirmed the presence of numerous high values (up to $787,000), suggesting that median is a more robust measure of central tendency for this variable in its original form. Wage income exhibits similar characteristics.

- **Education Level (`education_level`):** The median education level (code 16) signifies a Bachelor's degree, pointing towards a reasonably educated sample. However, the data includes individuals across the full spectrum of educational attainment.

- **Categorical Variables:** Gender distribution is nearly balanced (Mode = 2, Female). 'Employed' (code 1) is the most frequent employment status. Other categorical variables

like `census_region`, `race_detail_2_3`, `citizenship_status`, `disability_status`, `marital_status`, and `health_ins_coverage` provide further dimensions for potential stratified analysis.

4.3 Specific Requirements, Functions and Formulas

The analytical toolkit relied on several key Python libraries and statistical concepts:

Libraries Used:

- Primarily Pandas for data structures and manipulation, NumPy for numerical computation, Matplotlib and Seaborn for plotting, SciPy.Stats for hypothesis testing, and Statsmodels for VIF calculation.

Descriptive Statistics:

- Standard measures like mean, median, standard deviation, quantiles (for IQR), min, and max were computed using Pandas (`.describe()`, `.median()`, `.quantile()`) and NumPy (`np.mean`, `np.std`, etc.).

```
NumPy Stats for age_years:
Mean: 38.87, Std: 23.27, Range: 90
```

Data Handling:

- Missing values imputed via `.fillna()` with `.median()` or `.mode()[0]`. Special codes handled by replacing with `np.nan`.

Transformation:

- Log transformation (`np.log1p`) applied to income for visualization to mitigate skewness. A constant was added to handle non-positive values: log(Income + 6501).

Outlier Identification (IQR Method):

- IQR Method: Explicitly write the formulas: Lower Bound = Q1 - 1.5 * IQR; Upper Bound = Q3 + 1.5 * IQR. Mention that values outside these bounds are potential outliers requiring context.

```
IQR Outlier Bounds for total_income_12m: [-44687.50, 112612.50]
Number of Outliers: 521
Top 5 Outlier Values:
 4623    787000.0
969     586150.0
4981    541000.0
4899    536900.0
1773    532900.0
Name: total_income_12m, dtype: float64
```

Correlation:

- Pearson correlation (`.corr()`) used to measure linear association between numerical variables.

T-Test:

- `scipy.stats.ttest_ind(..., equal_var=False)` performed Welch's t-test to compare means of two independent groups (e.g., male vs. female income), assessing if the observed difference is statistically significant (p-value < 0.05 is typical threshold).

```
T-Test (Male vs. Female Income):
T-statistic: 9.794, P-value: 0.000
```

Chi-Squared Test:

- `scipy.stats.chi2_contingency` used to test for independence between two categorical variables (e.g., employment status and gender). A significant result (low p-value) suggests a relationship exists.

```
Chi-Squared Test (Employment Status vs. Gender):
Chi2: 123.537, P-value: 0.000, Degrees of Freedom: 5
```

Variance Inflation Factor (VIF):

- VIF: Explain the interpretation: VIF = 1 means no correlation, 1-5 is moderate, >5 suggests high correlation that might warrant attention, >10 indicates problematic multicollinearity.

```
VIF for Numerical Features:
            Feature       VIF
0         age_years  1.958517
1  total_income_12m  5.109350
2   wage_income_12m  4.170771
3      commute_time  1.558374
```

Normality Test:

- `scipy.stats.shapiro` used to check if data significantly deviates from a normal distribution (low p-value suggests non-normality).

```
NumPy Stats for age_years:
Mean: 38.87, Std: 23.27, Range: 90

First 5 Normalized Ages: [ 1.38098238  2.19761508  2.19761508 -0.72507035  0.43540768]
```

Hypothesis Tests:

- Briefly state the null hypothesis for each test. E.g., T-test: H0 = The means of the two groups are equal. Chi-Squared: H0 = The two categorical variables are independent. Shapiro-Wilk: H0 = The data is normally distributed.

4.4 Analysis Results

Distributional Characteristics:

The sample displays wide age variability, not conforming to a normal distribution (Shapiro-Wilk $p<0.001$). Income variables (`total_income_12m`, `wage_income_12m`) are highly right-skewed, with the mean ($47,384) significantly exceeding the median ($30,000), heavily influenced by high-income outliers (max $787,000). Log

transformation substantially improves income distribution symmetry for visualization. The median education level corresponds to a Bachelor's degree (code 16).

```
Shapiro-Wilk Test for age_years:
Statistic: 0.965, P-value: 0.000
```

Relationships and Statistical Tests:

Weak linear correlations were observed between age, income, and commute time via heatmap analysis. However, box plots revealed a clear positive trend between higher education levels and total income. Statistical testing confirmed significant differences: mean income varied significantly between genders (Welch's $t=9.79$, $p<0.001$), and there was a significant association between gender and employment status ($\chi^2=123.5$, $p<0.001$). Regional income comparisons hinted at differences, though the attempted A/B test between Northeast and West was inconclusive due to data limitations.

```
A/B Test (Region 1 vs. Region 4 Income):
T-statistic: nan, P-value: nan
```

Data Diagnostics:

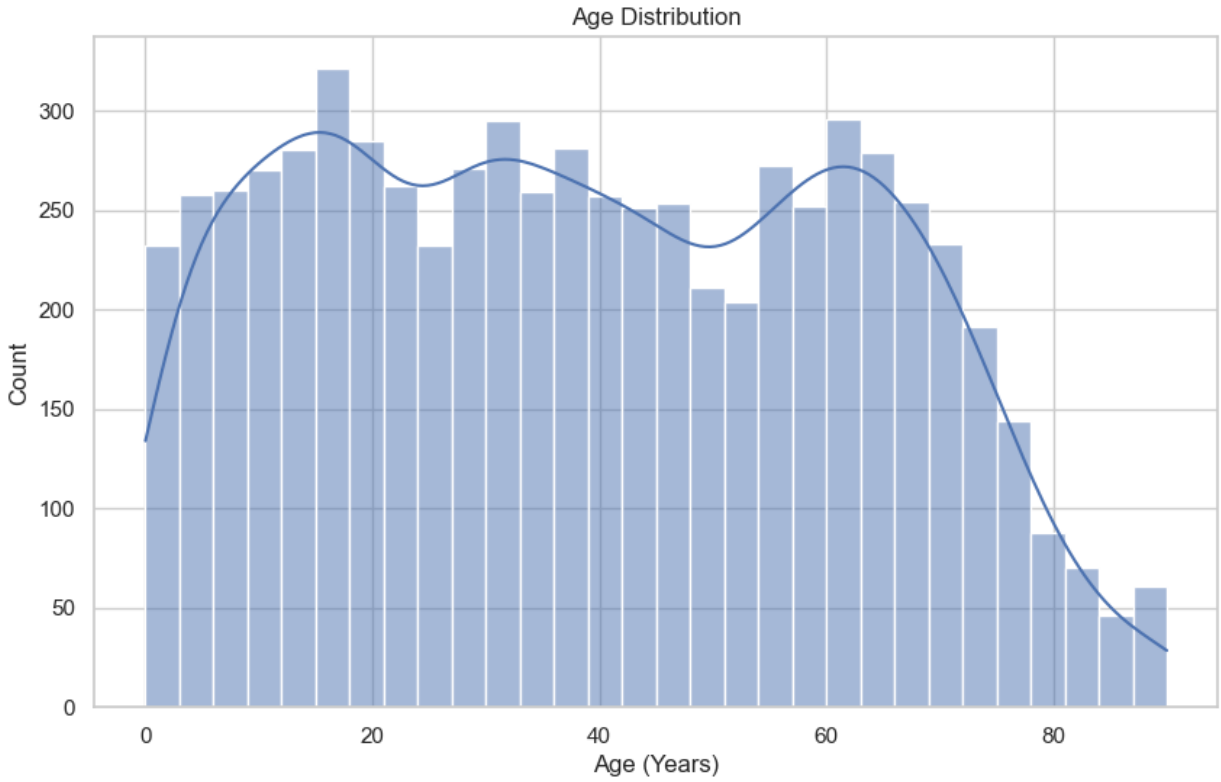
IQR analysis identified 521 high-income outliers in `total_income_12m`. VIF analysis on numerical predictors indicated low multicollinearity (max VIF ≈ 5.1), suggesting they could potentially be used together in regression models without severe collinearity issues.

```
VIF for Numerical Features:
            Feature       VIF
0         age_years  1.958517
1  total_income_12m  5.109350
2   wage_income_12m  4.170771
3      commute_time  1.558374
```

4.5 Visualization

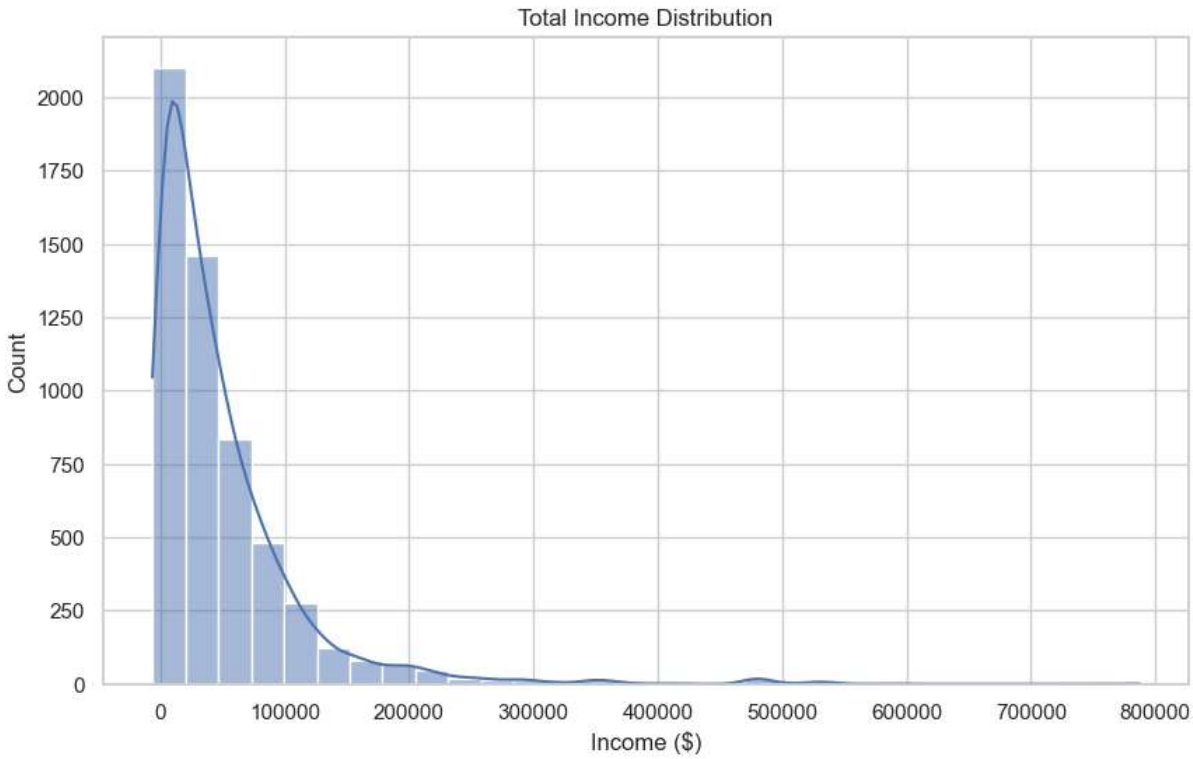
Visualizations were integral to this EDA, allowing for the identification of distributions, trends, correlations, and potential data quality issues. The following key plots were generated (images inserted below their respective captions):

Figure 1: Distribution of Age (Years) in the sample subset.



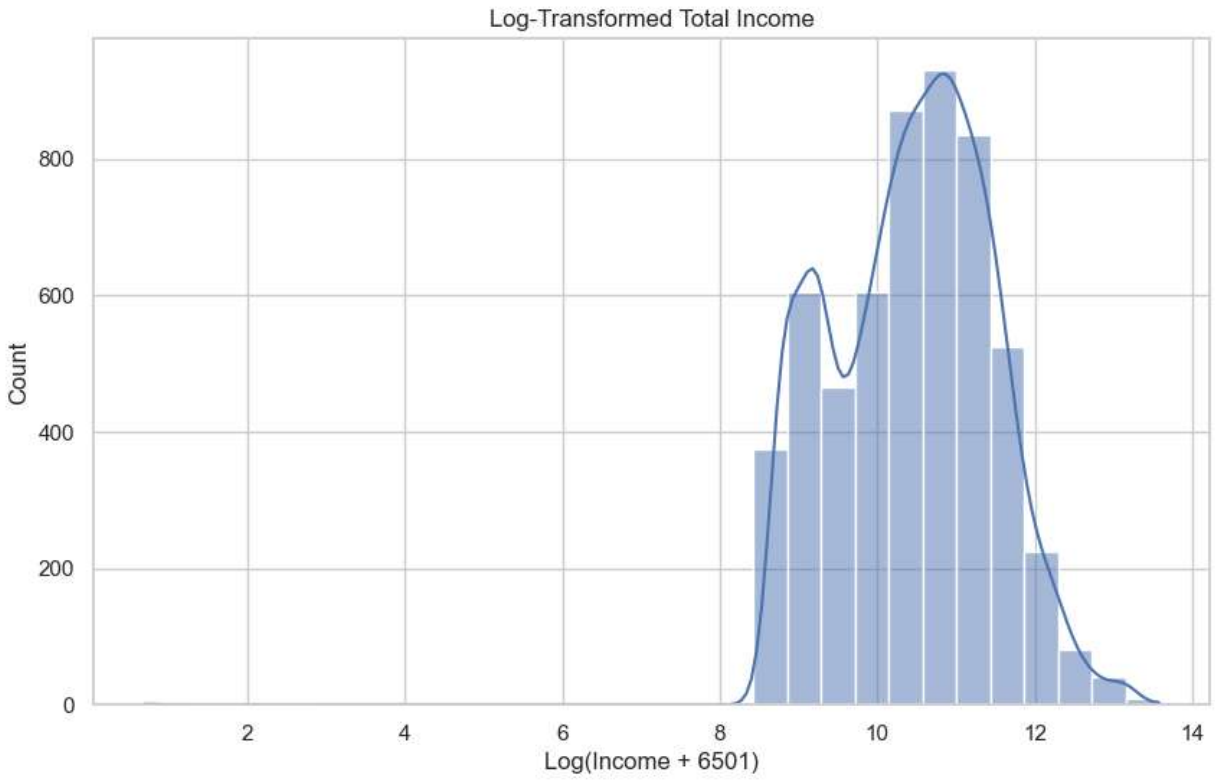
This histogram with a Kernel Density Estimate (KDE) overlay shows the age range from 0 to 90, indicating a diverse age representation in the sample.

Figure 2: Distribution of Total Income (Past 12 Months).



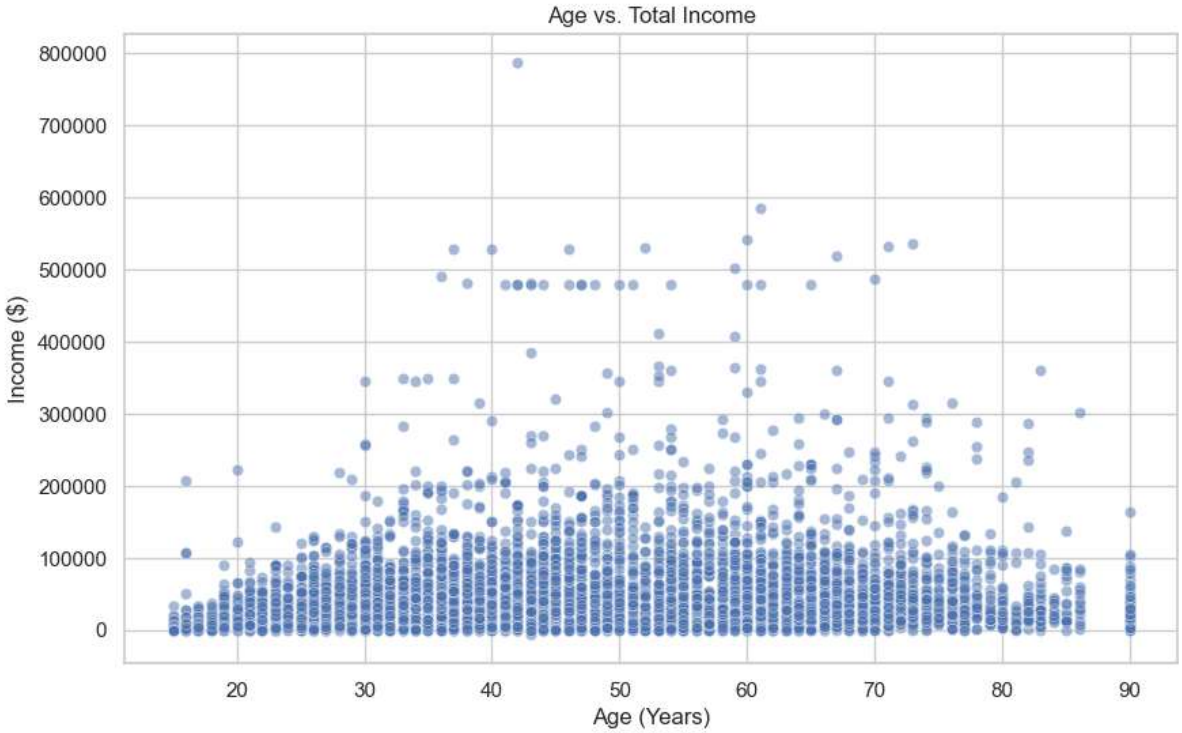
The heavy right-skewness in the income distribution is clearly visible, with most individuals earning lower to moderate incomes.

Figure 3: Distribution of Log-Transformed Total Income.



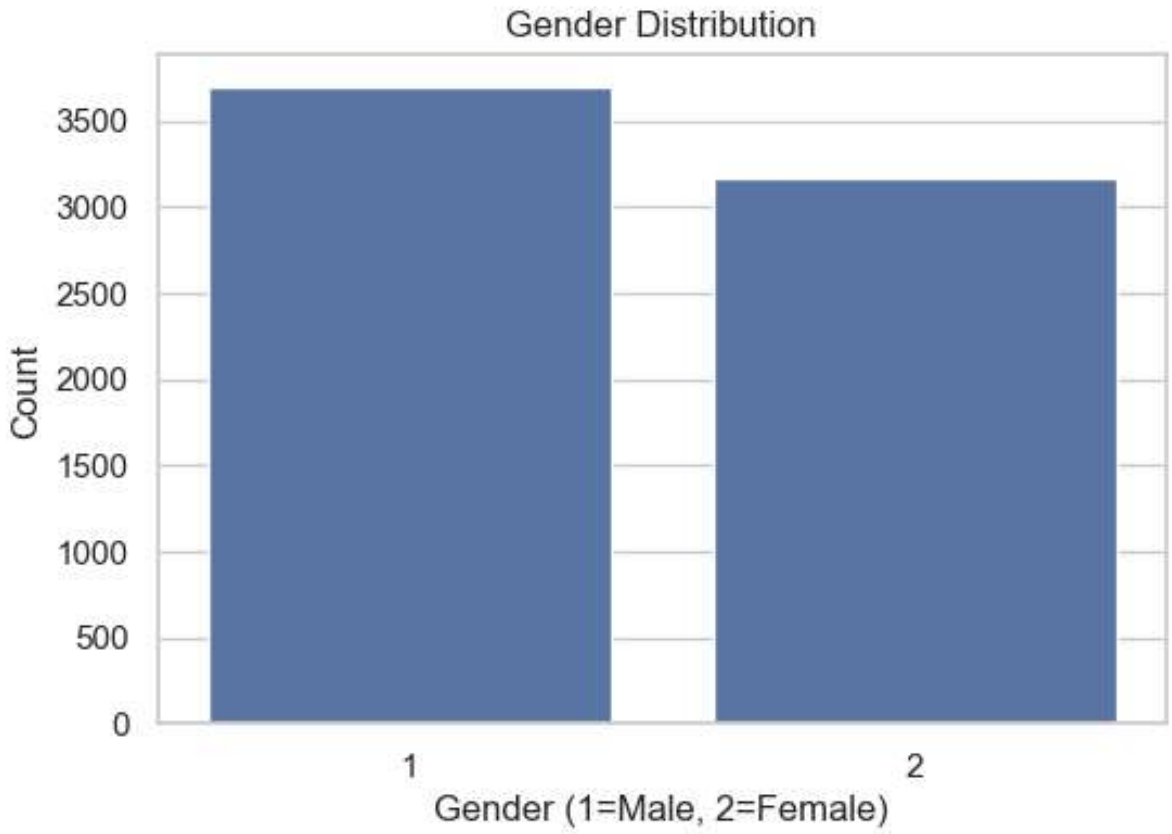
Applying a log transformation (log1p(x + 6501)) makes the income distribution more symmetrical, potentially suitable for modeling.

Figure 4: Relationship between Age and Total Income.



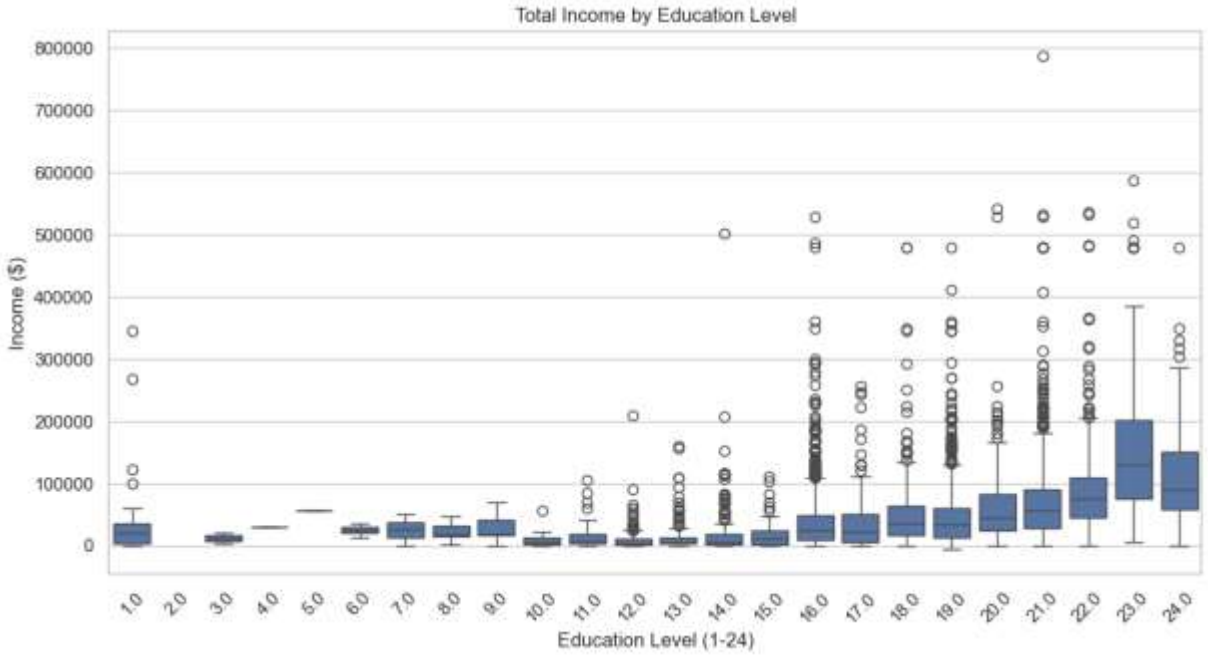
The scatter plot reveals no strong linear trend between age and income, although income spread appears wider for middle-aged groups.

Figure 5: Count of Individuals by Gender in the sample subset.



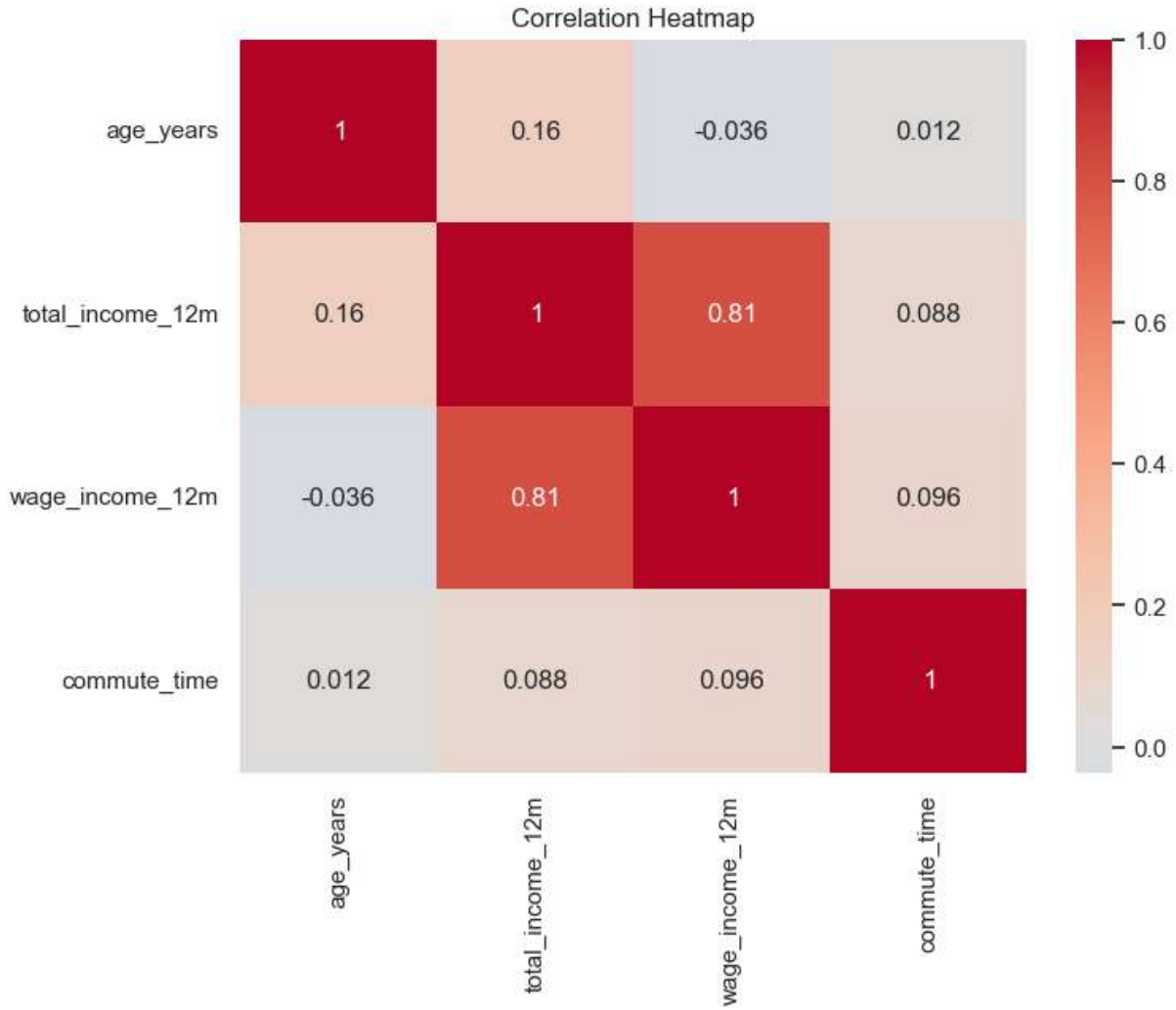
Shows the counts for gender categories (1=Male, 2=Female), indicating a fairly balanced distribution after imputation.

Figure 6: Total Income across different Education Levels.



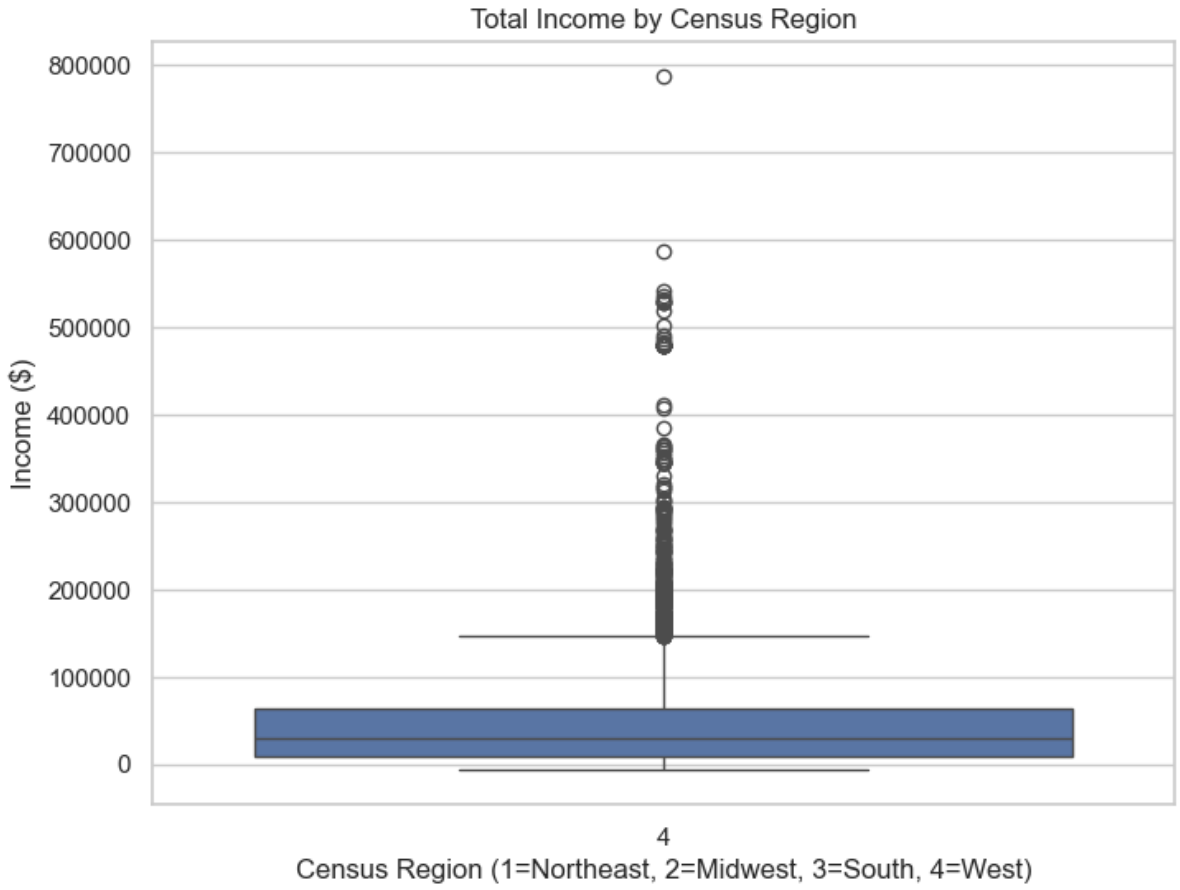
This box plot shows a positive association between education level and median total income, along with numerous high-income outliers at higher education levels.

Figure 7: Correlation Heatmap of Numerical Variables.



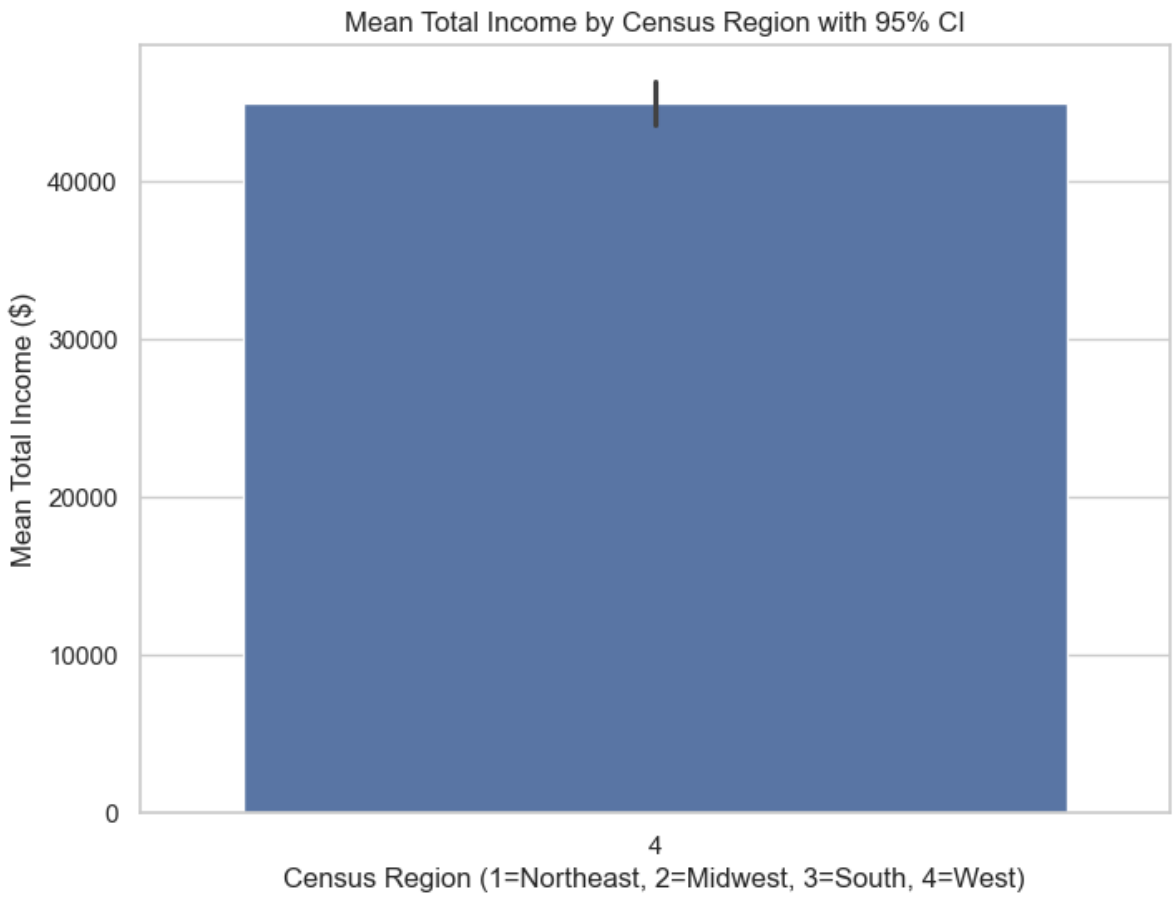
Highlights the Pearson correlation coefficients. Notable is the strong correlation between total income and wage income, and weaker correlations involving age and commute time.

Figure 8: Box plot illustrating the distribution and outliers in Total Income.



Clearly depicts the skewness and the extent of high-income outliers.

Figure 9: Mean Total Income by Census Region with 95% Confidence Intervals.



Compares average incomes across regions, suggesting potential differences (Note: based on limited data after cleaning).

5. Conclusion

The Exploratory Data Analysis of the ACS 2023 1-Year PUMS dataset, focusing on a cleaned 15-variable subset, provided valuable preliminary insights into the selected sample. The process involved careful data handling, including cleaning, imputation, and feature selection.

Key findings underscored the diversity within the sample, particularly in age and income. The pronounced right-skewness in income variables and the presence of high-income outliers were confirmed visually and through IQR analysis. A positive relationship between education level and income was observed. Statistically significant differences based on gender were detected for both mean income and employment status distributions. While regional differences in income were visually suggested, data limitations prevented conclusive statistical testing between all regions in this subset. Multicollinearity was assessed as low among the core numerical variables examined.

This EDA effectively highlights the data's structure, quality considerations, and potential areas for further investigation. It forms a basis for future analysis, keeping in mind the limitations of using unweighted sample data and the selected subset.

6. Future Scope

Building upon this initial EDA, future work could explore several productive avenues:

- **Population-Level Analysis:** Incorporate PUMS survey weights (PWGTP) to derive estimates representative of the entire U.S. population and calculate appropriate margins of error for statistical validity. This would allow for generalization of findings and comparison with official Census Bureau reports.

- **Predictive Modeling:** Develop regression models to predict income based on the analysed factors (education, age, gender, region, etc.), considering transformations or robust methods due to income skewness. Classification models could target employment status or other categorical outcomes. Exploring techniques like quantile regression might be beneficial for handling income skewness and predicting across the distribution, not just the mean.

- **Subgroup Analysis:** Perform deeper dives into how observed relationships differ across various demographic subgroups (e.g., race, citizenship status) by conducting stratified analyses. For example, comparing the income-education relationship across different race_detail_2_3 categories or between native-born citizens and naturalized citizens (citizenship_status).
- **Outlier Impact Assessment:** Systematically investigate the influence of high-income outliers on statistical results and model performance, potentially employing techniques like winsorization or robust statistical models. Techniques could include sensitivity analysis by removing outliers, using robust statistical methods less sensitive to extremes, or modeling outliers separately.

- **Expanded Feature Analysis:** Re-introduce relevant variables dropped during initial cleaning or selection to build a more comprehensive understanding of socio-economic dynamics (e.g., industry, occupation, detailed household characteristics) such as incorporating occupation and industry codes could provide much richer context for employment and income patterns

- **Geospatial Exploration:** Leverage PUMA and state codes for more granular geographic comparisons, potentially linking with external geographic datasets where feasible potentially revealing regional disparities in finer detail than census regions, although PUMA boundaries may change and require careful handling.

7. References

[1] U.S. Census Bureau, “American Community Survey (ACS): Public Use Microdata Sample (PUMS),” Accessed: Nov 2024. [Online]. Available: https://www.census.gov/programs-surveys/acs/microdata.html

[2] U.S. Census Bureau, “PUMS Documentation: User Guides, Data Dictionaries, Accuracy,” Accessed: Nov 2024. [Online]. Available: https://www.census.gov/programs-surveys/acs/microdata/documentation.html

[3] Pandas Development Team, “pandas: powerful Python data analysis toolkit,” Zenodo, 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3509134

[4] C. R. Harris et al., “Array programming with NumPy,” Nature, vol. 585, pp. 357–362, 2020.

[5] J. D. Hunter, “Matplotlib: A 2D graphics environment,” Comput. Sci. Eng., vol. 9, no. 3, pp. 90–95, 2007.

[6] M. L. Waskom, “Seaborn: statistical data visualization,” J. Open Source Softw., vol. 6, no. 60, p. 3021, 2021.

[7] P. Virtanen et al., “SciPy 1.0: Fundamental algorithms for scientific computing in Python,” Nat. Methods, vol. 17, pp. 261–272, 2020.

[8] S. Seabold and J. Perktold, “statsmodels: Econometric and statistical modeling with python,” presented at the 9th Python in Science Conference, 2010.