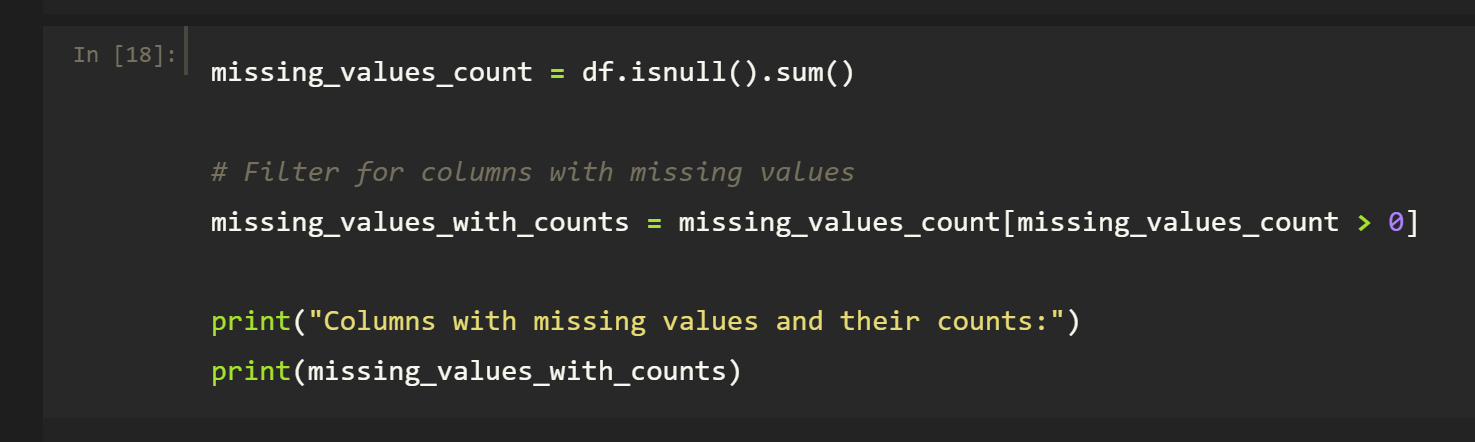
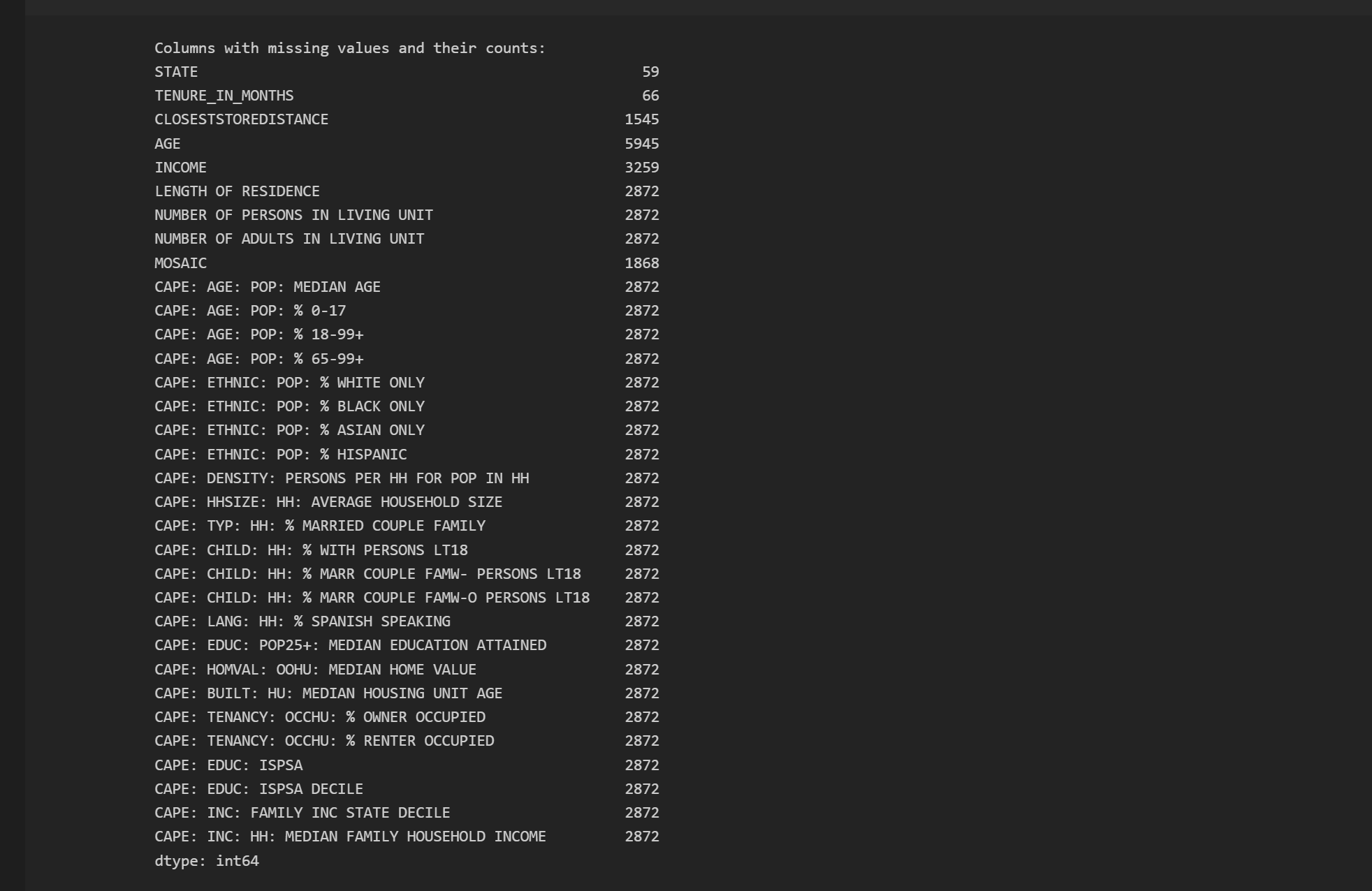
ACADIA DATA SCIENCE INTERN ASSIGNMENT

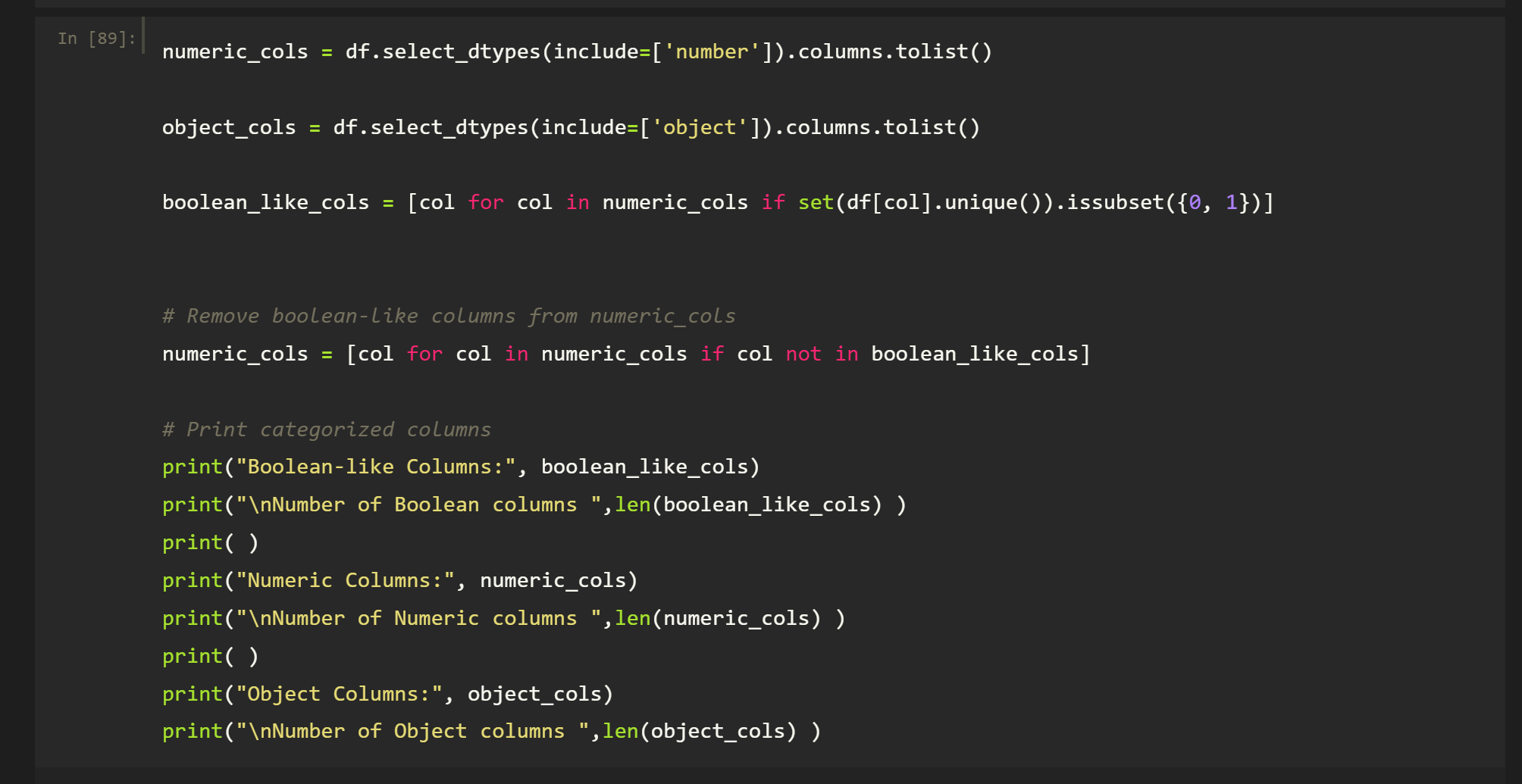
# **List down all the columns with missing values.**





I identified all columns with missing data along with the corresponding number of missing values in each column. This analysis is crucial for understanding the extent of missing data and its potential impact on subsequent analysis.

1. **Categorize the columns based on their data type and print, for ex: print all the numeric variables and other data types as well**





During EDA, I identified a PRIVATELABELTENDERFLAG column containing values 'Y' and 'N,' which I inferred to represent 'Yes' and 'No.' I transformed the values to align it with the Boolean data type, mapping 'Y' to 1 and 'N' to 0. This preprocessing step facilitated its inclusion under the Boolean datatype category, simplifying subsequent analysis.

The dataset contains three primary data types: Boolean, numeric, and object.

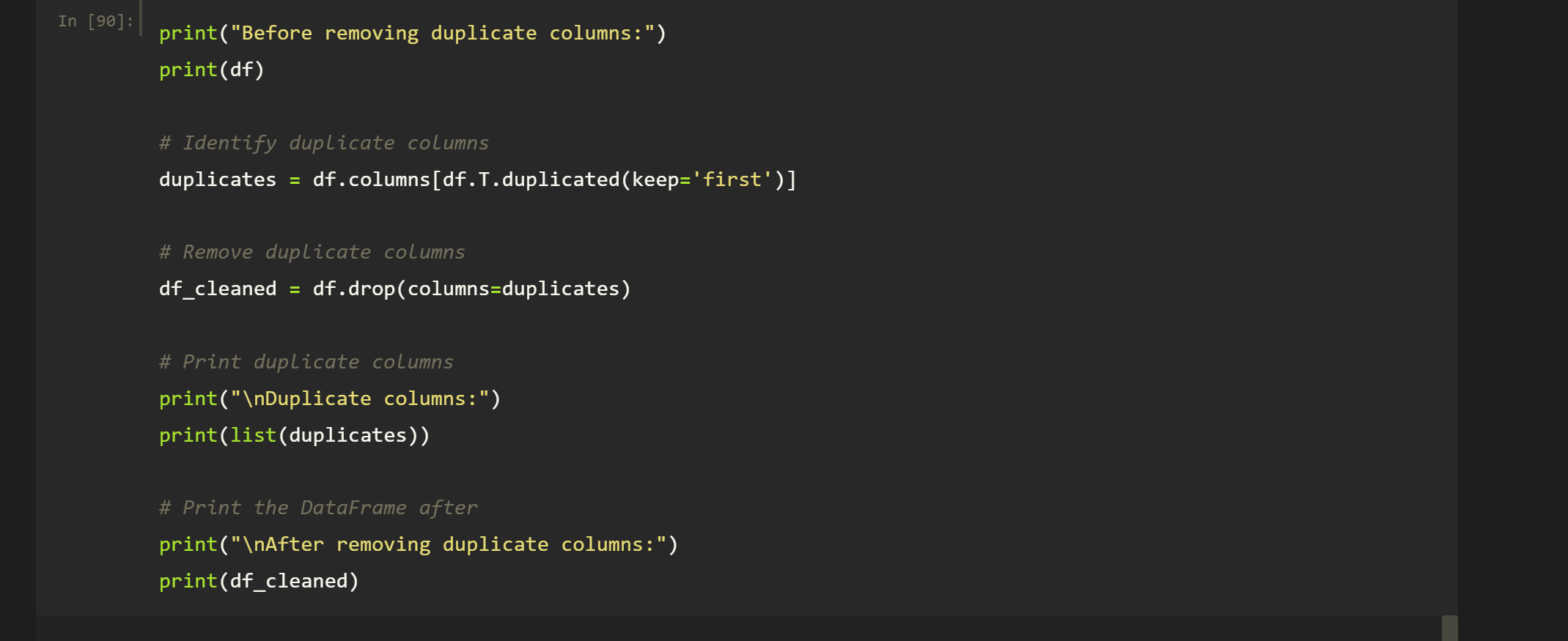
* **Boolean columns**: 28 columns, including fields like PRIVATELABELTENDERFLAG, were preprocessed to fit the Boolean datatype constraint.
* **Numeric columns**: 87 columns, representing continuous and discrete values such as AGE, INCOME, and various customer spending and demographic metrics.
* **Object columns**: 2 columns (STATE and MOSAIC), typically representing categorical data.

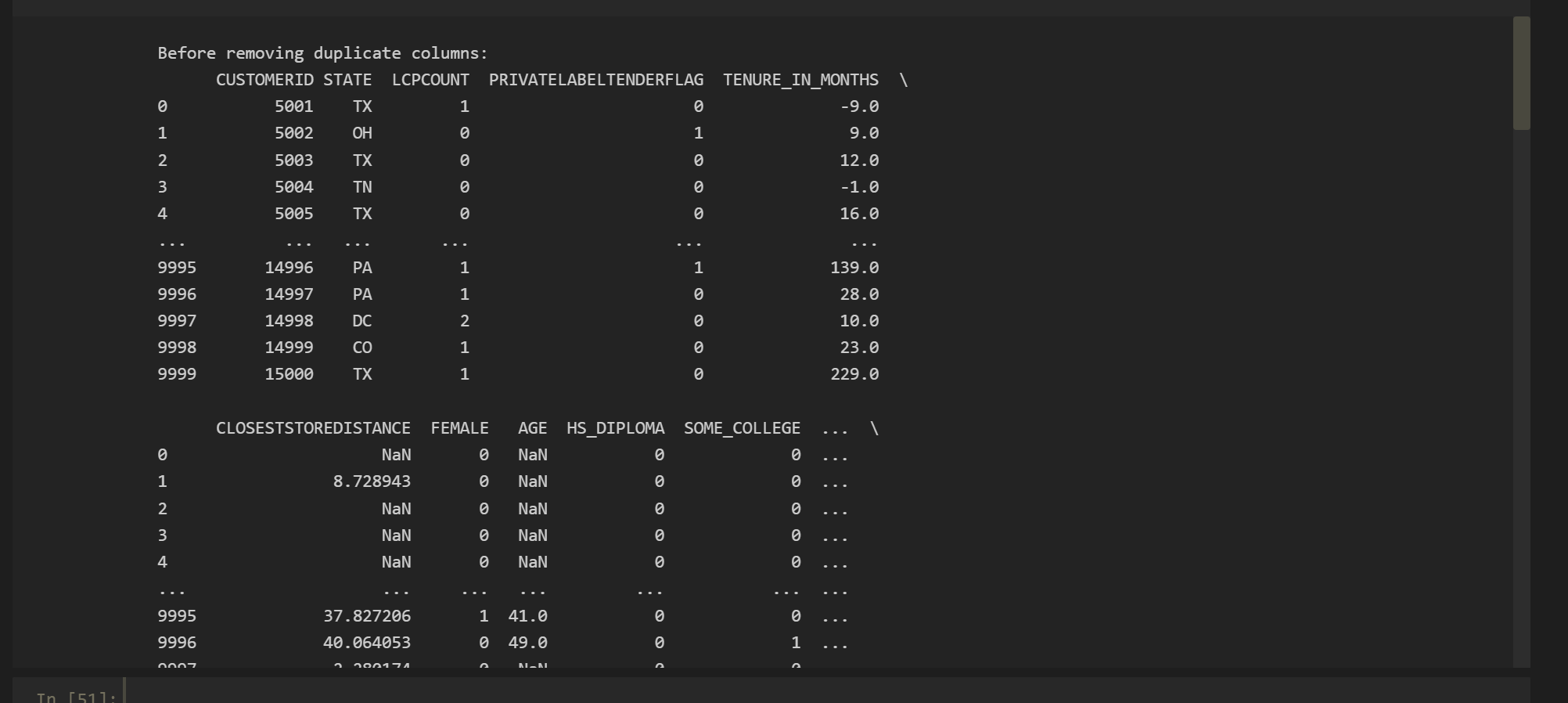
This categorization provides a comprehensive understanding of the dataset's structure and ensures appropriate handling of each type during analysis. For instance, numeric data will be used for statistical analysis and feature engineering, Boolean data for binary classification or filtering, and object data for encoding or segmentation.

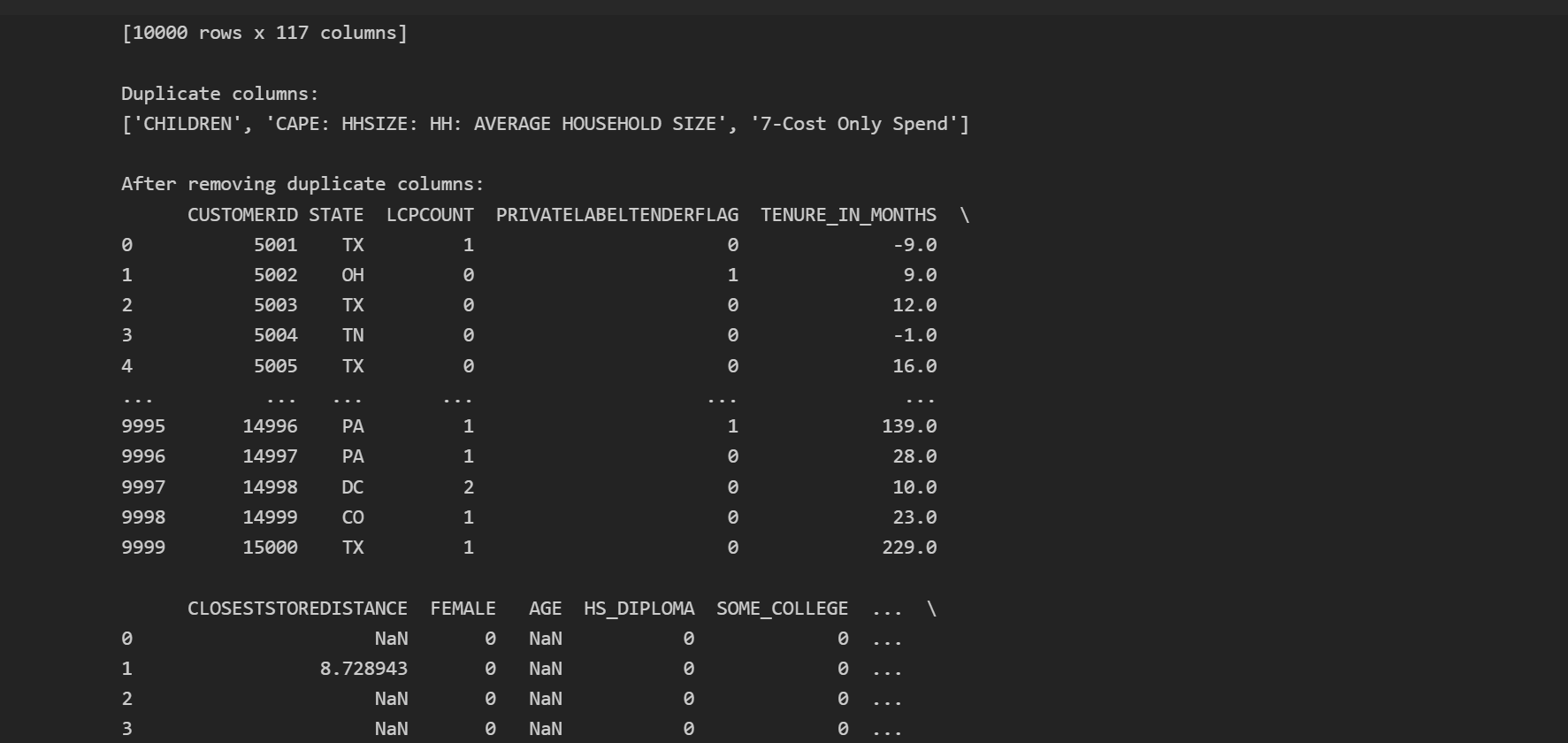
1. **List the columns with duplicates**

**a) Remove them**

**b) Print before and after**



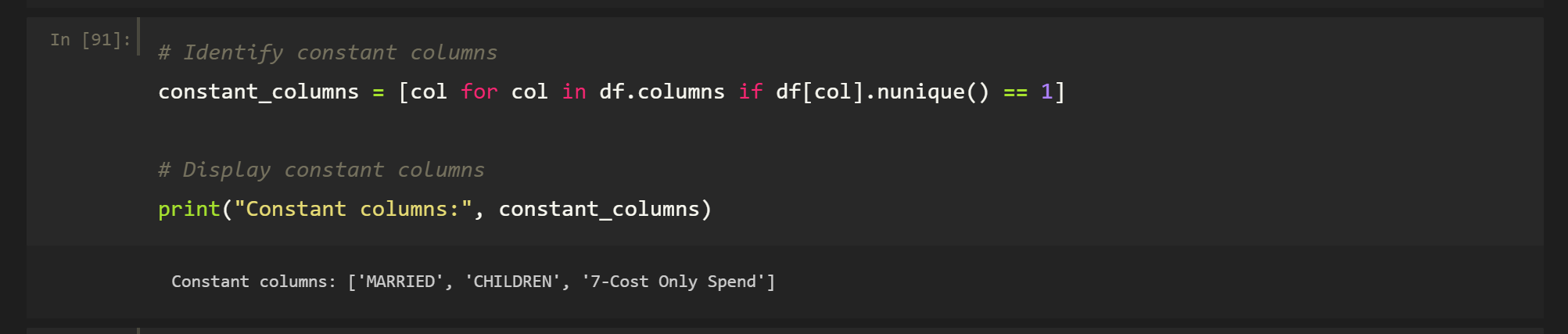


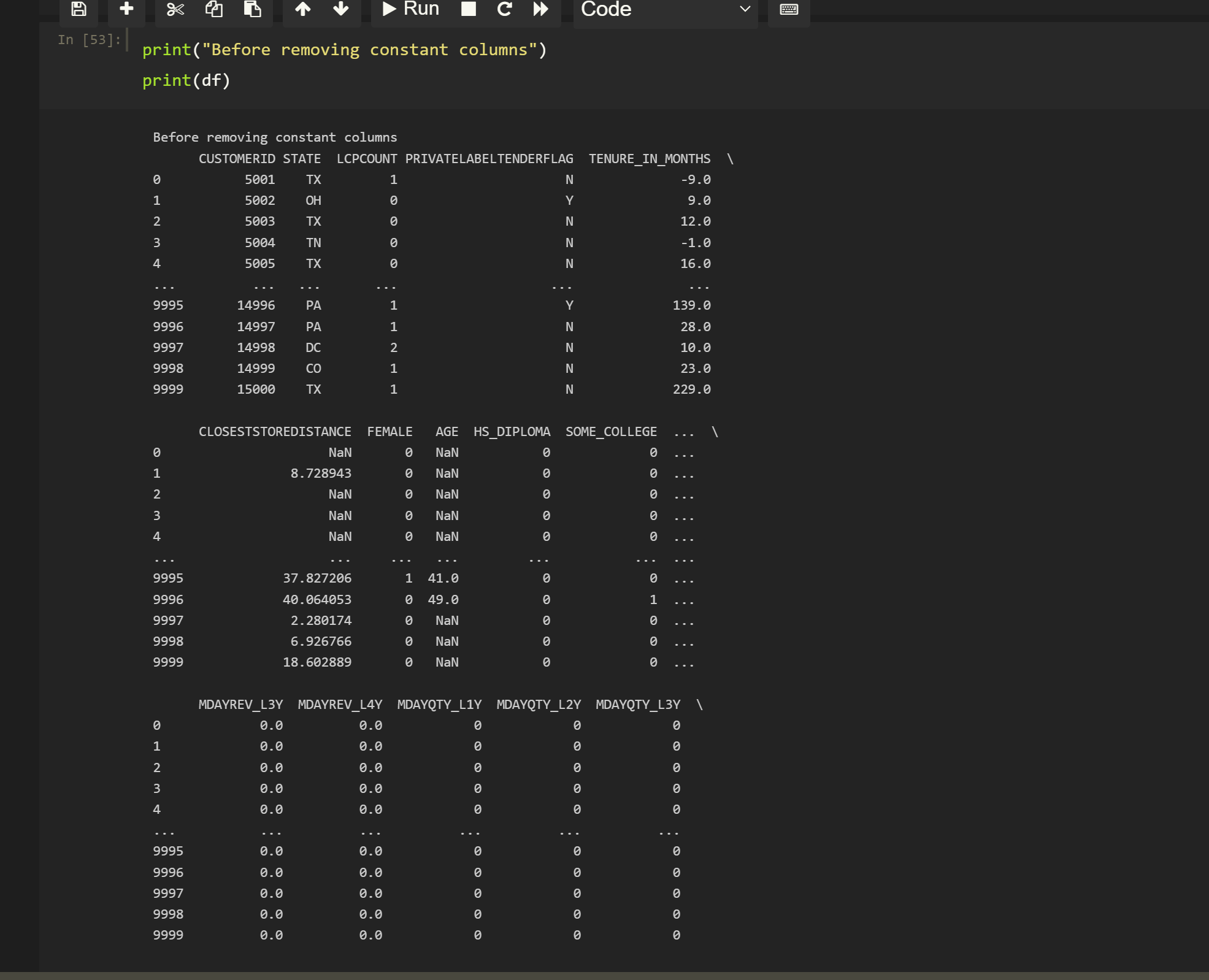


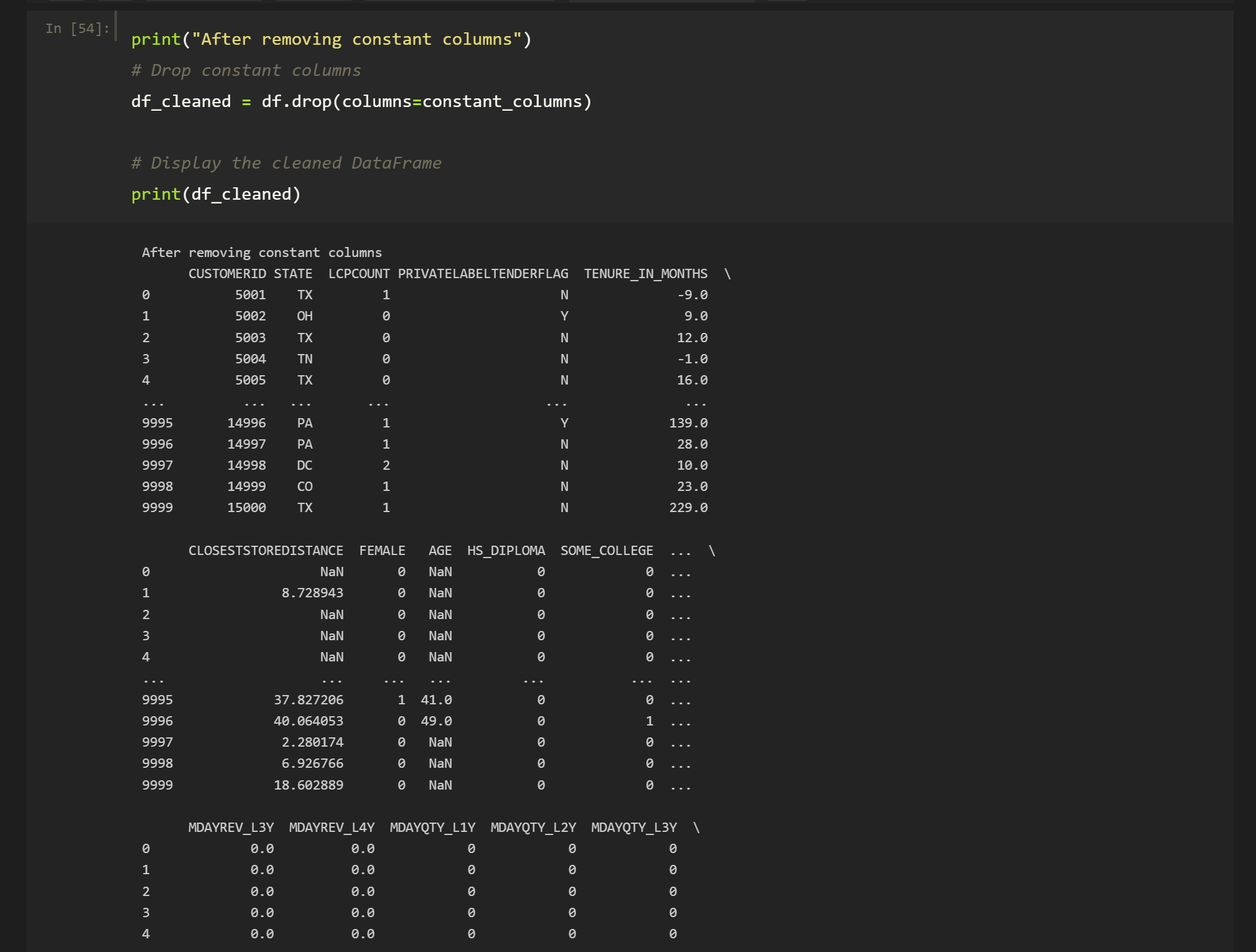
1. **List the constant columns**

**a) Remove them**

**b) Print before and after**



****

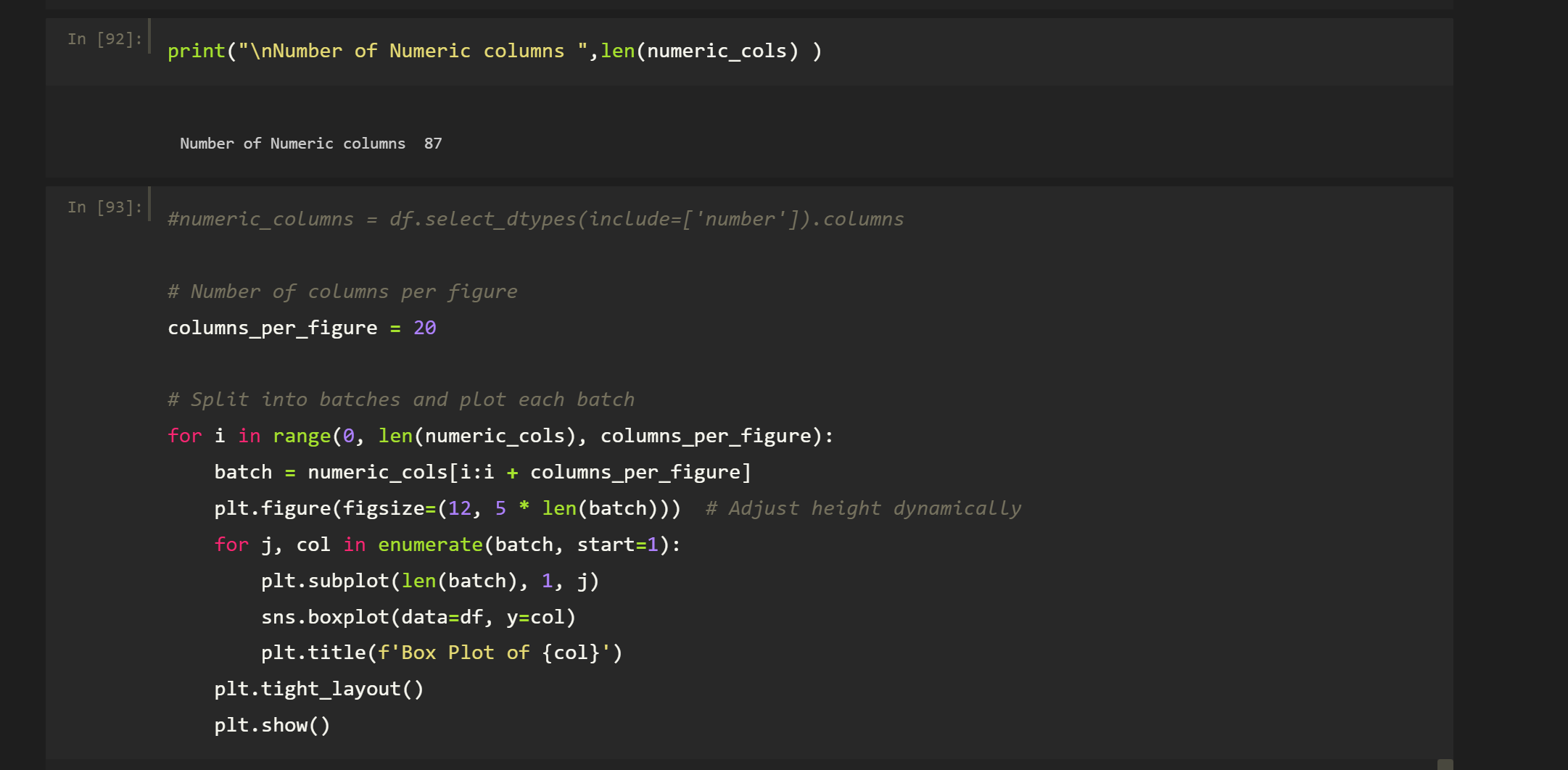
****

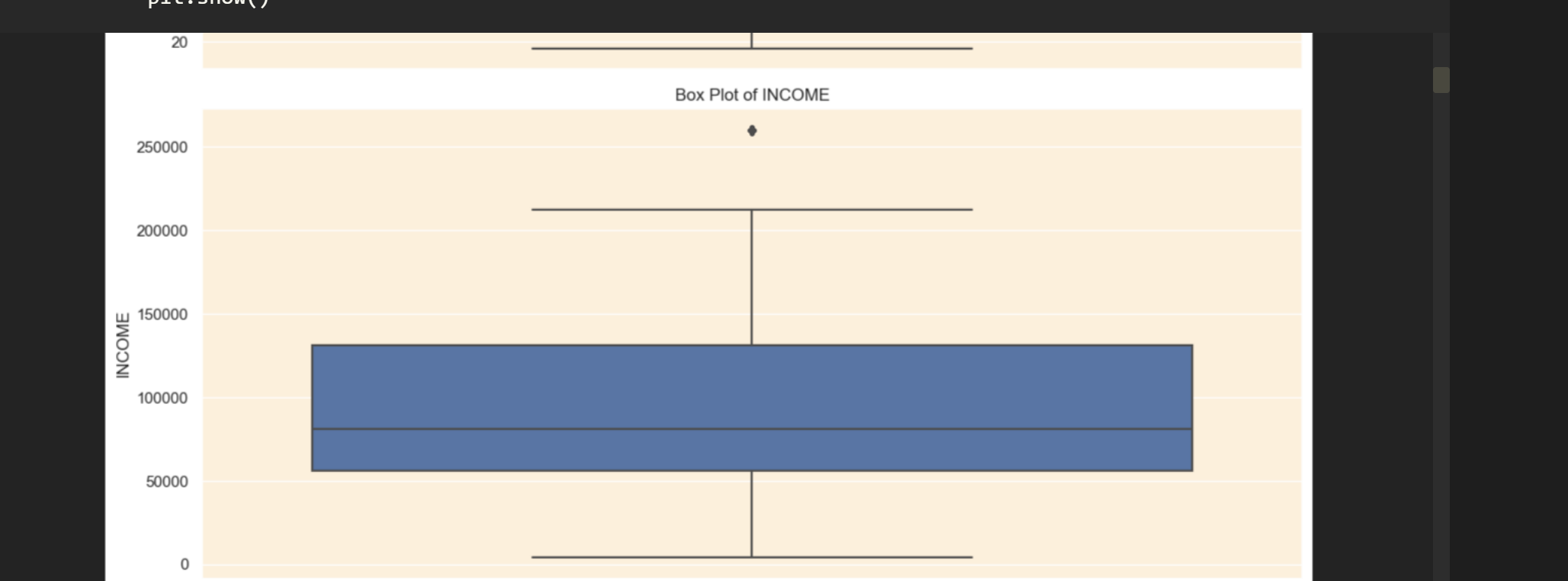
**It is common for datasets to include columns with constant values across all rows. Such columns lack variability and provide no meaningful information, making them safe to remove without impacting the analysis. Removing constant columns is an important preprocessing step that:**

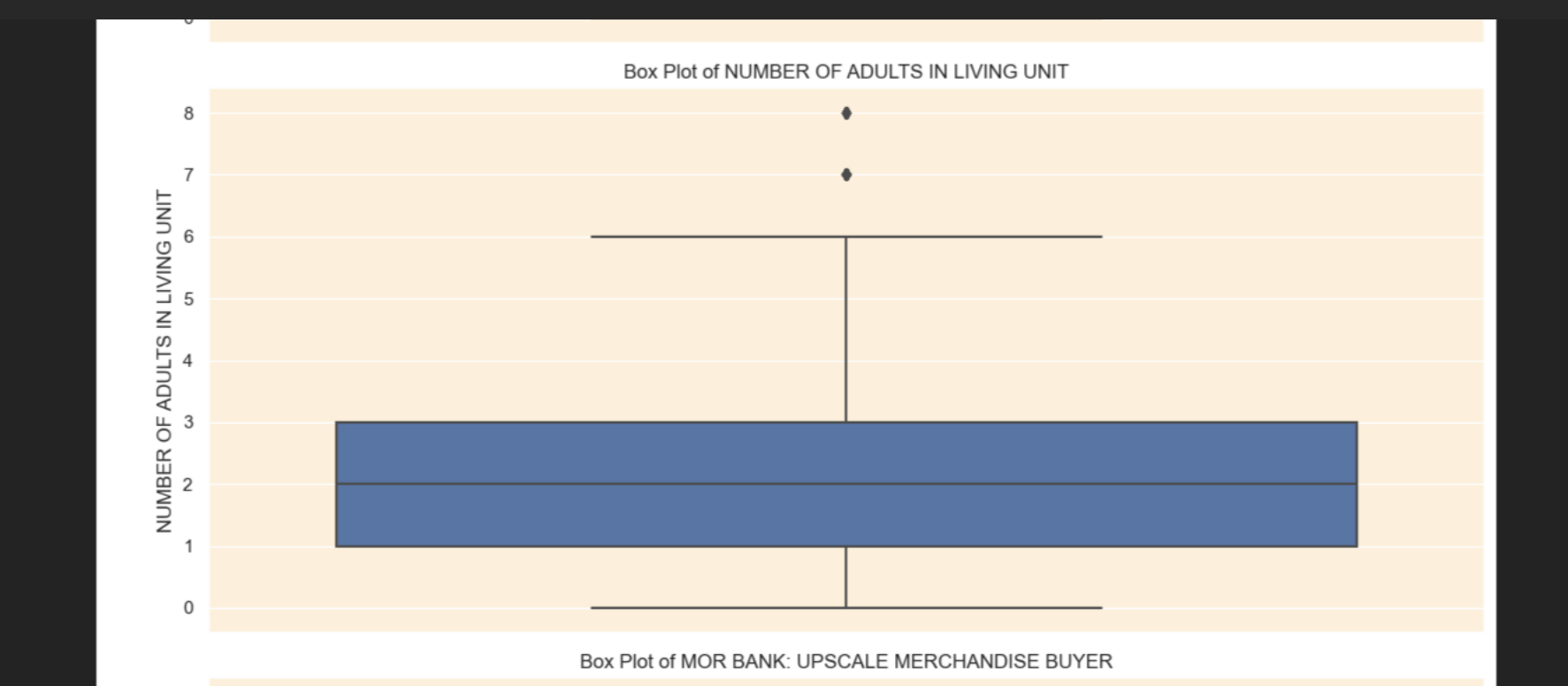
1. **Reduces dimensionality: Simplifies the dataset by removing redundant features.**
2. **Improves computational efficiency: Decreases processing time during model training and evaluation.**
3. **Enhances model interpretability: Eliminates irrelevant data, ensuring the model focuses on meaningful features.**

**This step ensures a cleaner dataset and aids in building more efficient and interpretable machine learning models.**

1. **Create a box plot to visualize the outliers of all the numeric columns**



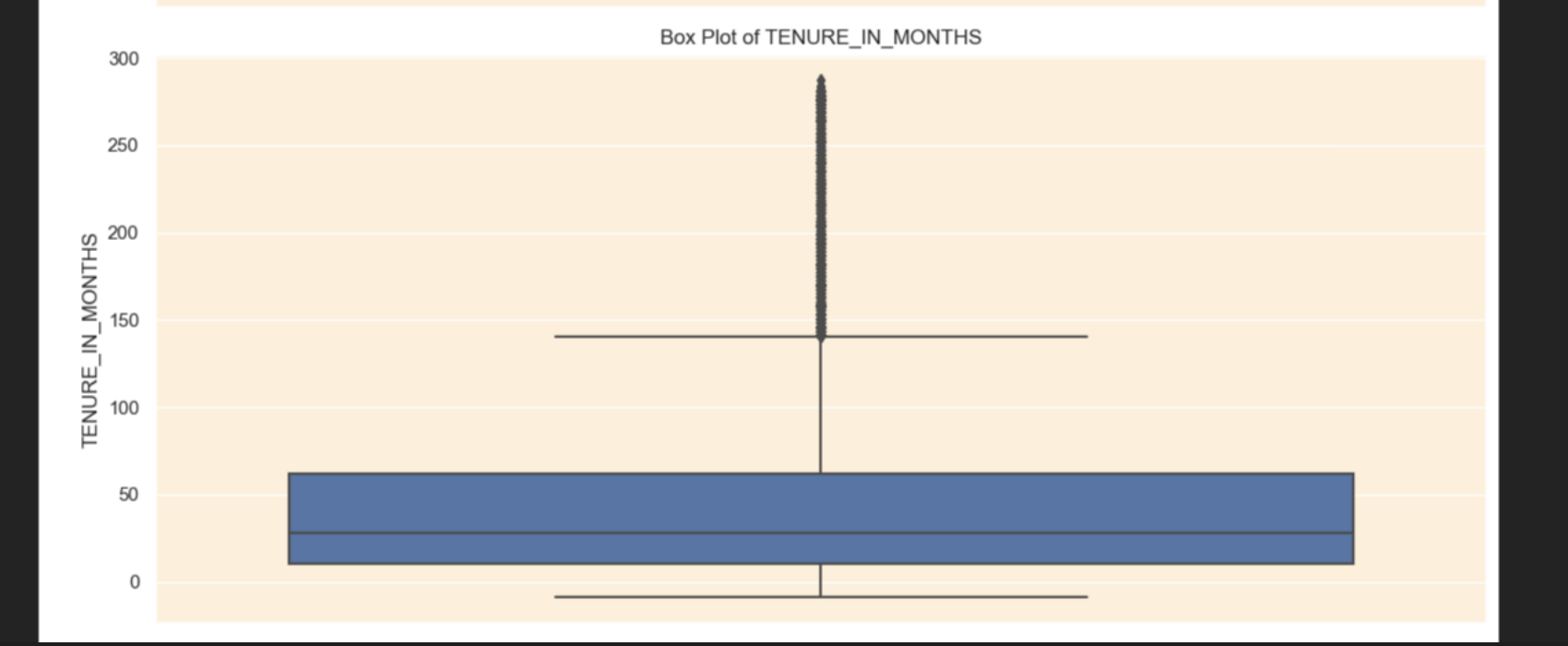
****

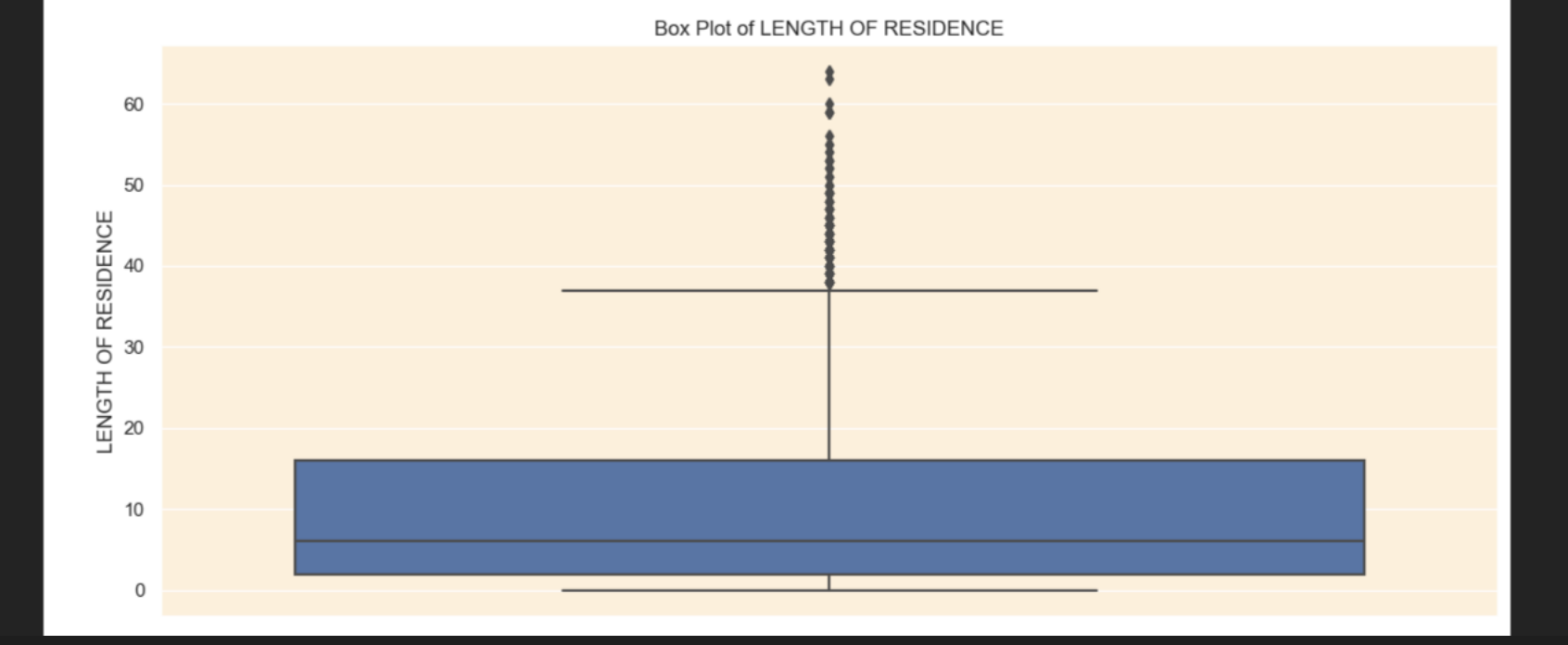
****

Generated box plots for all 87 numeric columns to visualize the distribution and detect outliers. While many plots provided limited insights, some columns revealed meaningful patterns. For instance:

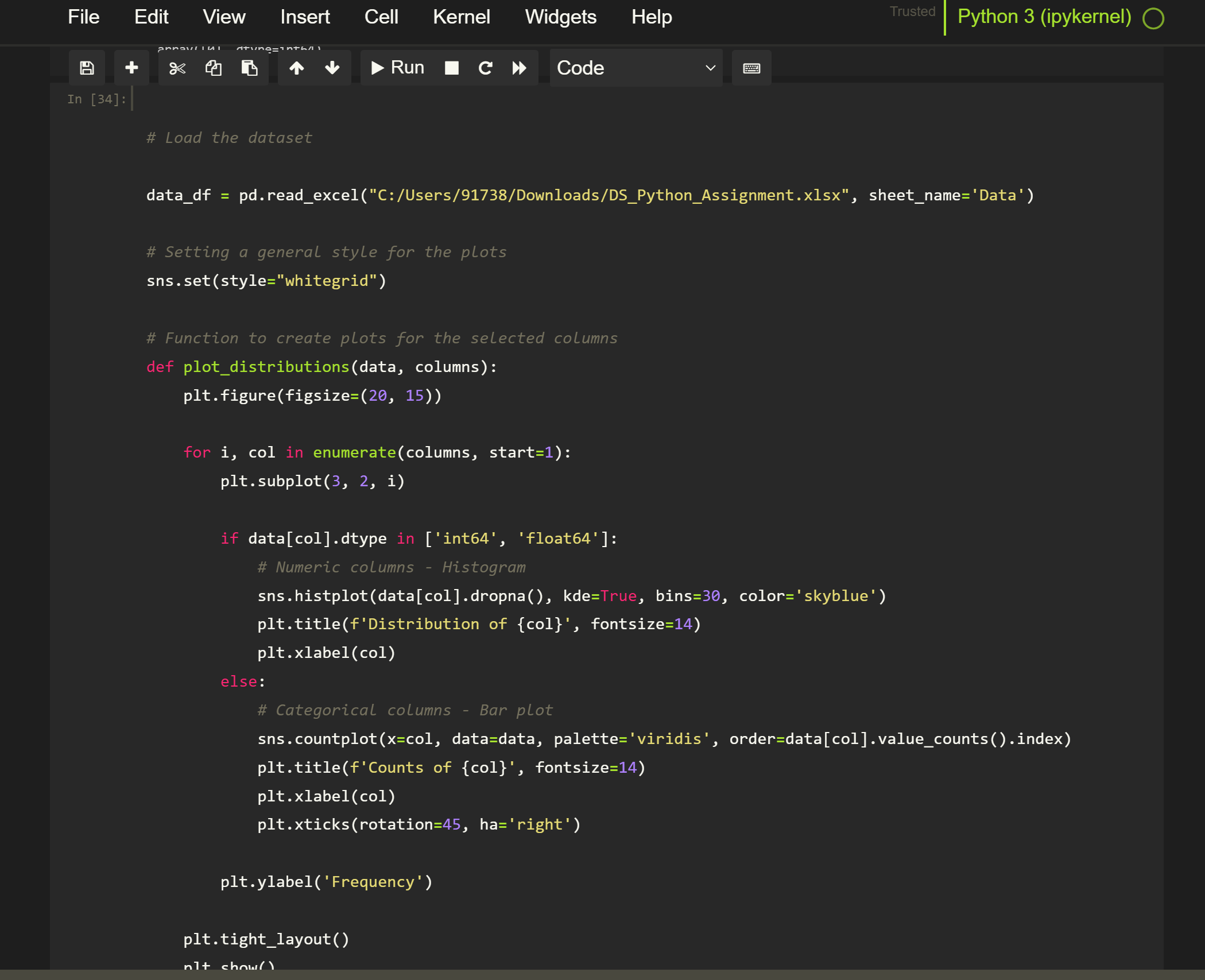
* **INCOME column**: The box plot indicates a single outlier.
* **NUMBER OF ADULTS IN LIVING UNIT column**: Two outliers are apparent in the plot.

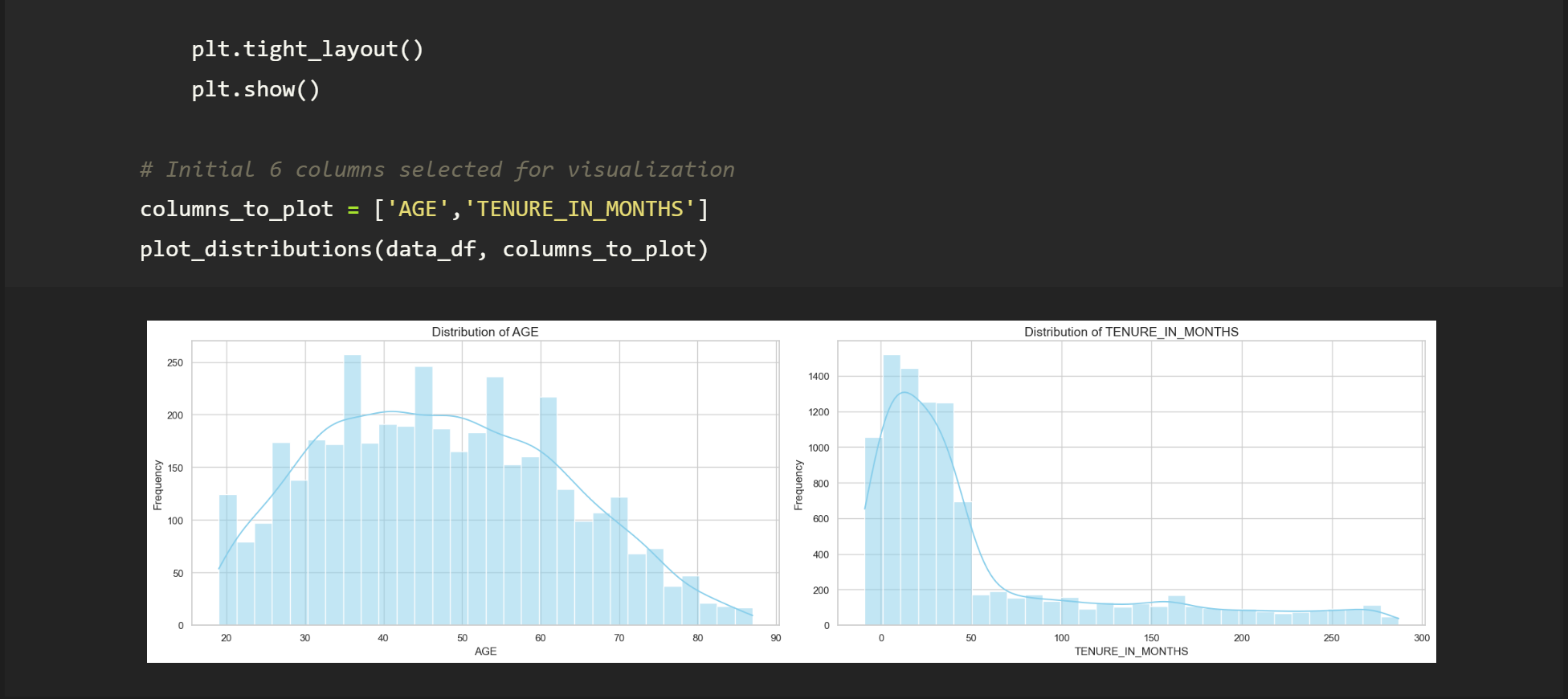
**Some of the other outliers that we are able to detect from the plotting box plots for all the numeric columns are as follows:**

****

****

1. **Create charts for any 6 columns and show their distribution**





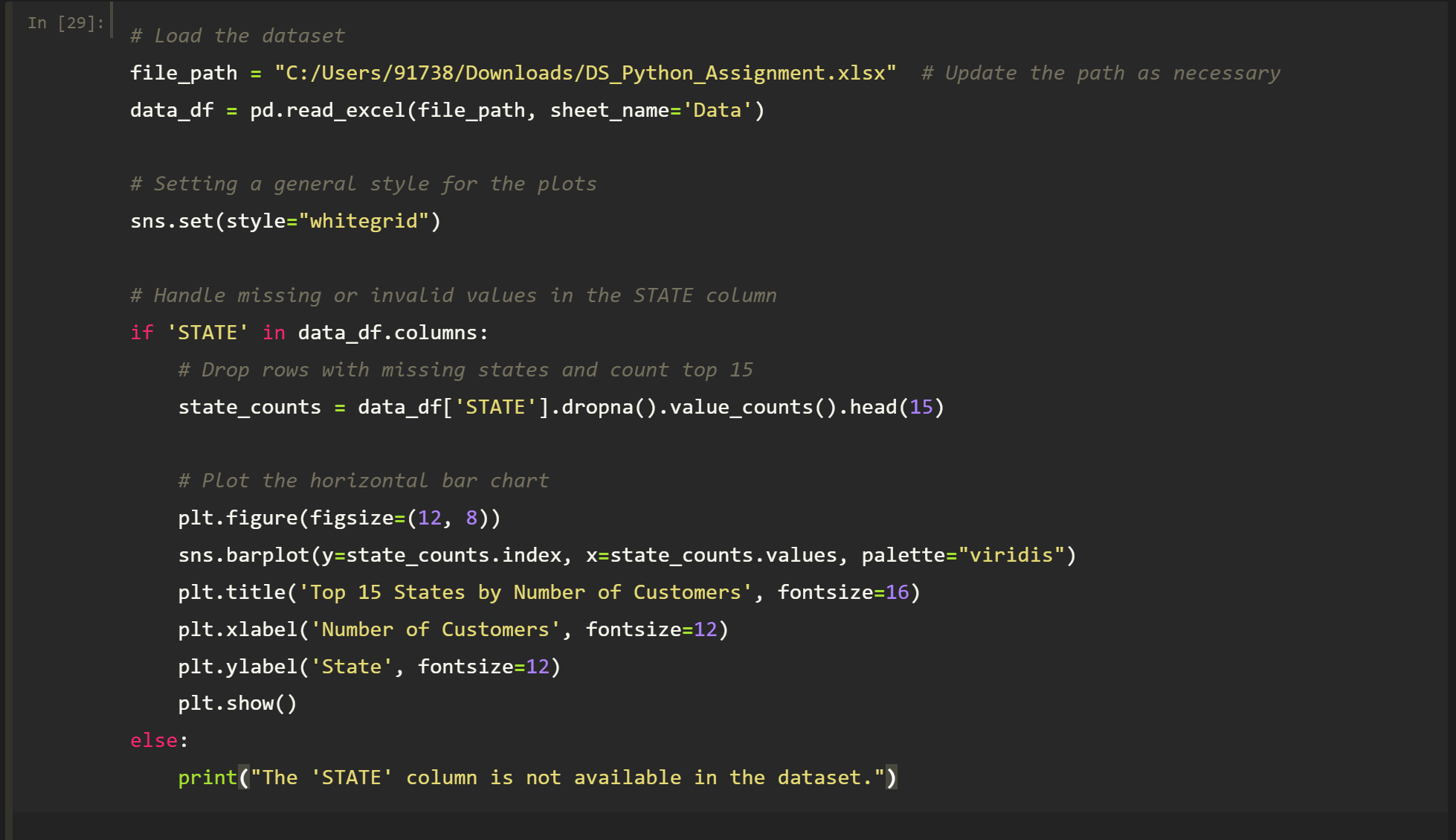
**Left Plot: Distribution of Age**

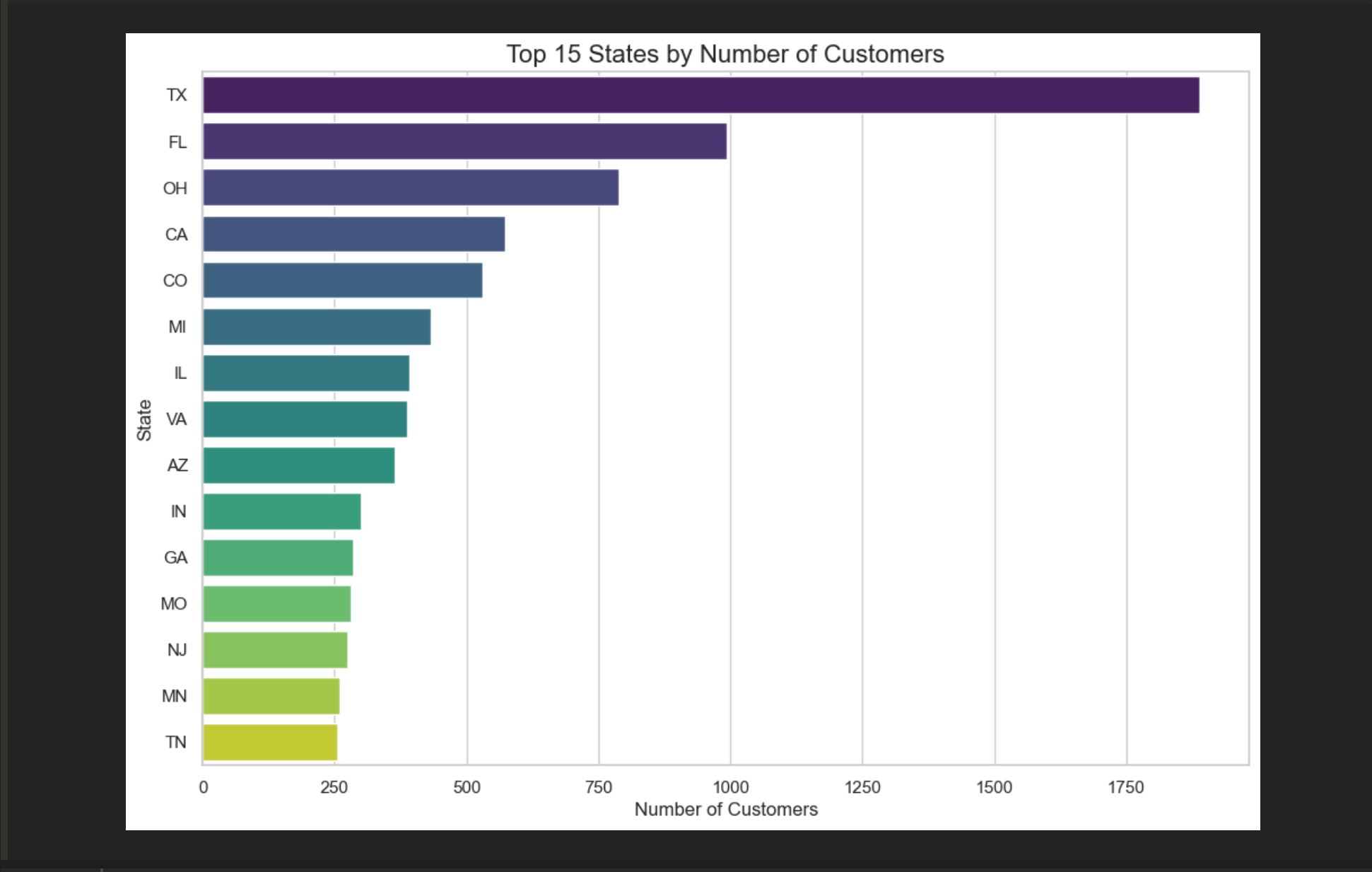
* **X-Axis (AGE)**: Ranges from 20 to 90 years.
* **Y-Axis (Frequency)**: Represents the count of individuals in each age group.
* **Histogram Bars**: The bars are light blue and indicate the frequency of individuals within different age bins. There’s a noticeable peak around the ages of 30-40.
* **Curve**: A smoothed line (density curve) overlays the histogram, suggesting that age distribution is approximately normal, with a slight skew reflecting a higher count in younger adults.

**Right Plot: Distribution of Tenure in Months**

* **X-Axis (TENURE\_IN\_MONTHS)**: Ranges from 0 to 300 months (or 0 to 25 years).
* **Y-Axis (Frequency)**: Represents the number of individuals based on their tenure.
* **Histogram Bars**: These are also light blue and show a clear right skew. Most individuals have lower tenure, peaking around 0-50 months, indicating a tendency towards shorter employment durations.
* **Curve**: A density curve follows the histogram, highlighting that the distribution is heavily concentrated at low tenure values, tapering off as tenure increases.

Overall, the left plot indicates a fairly normal distribution of age among individuals, while the right plot clearly shows that most individuals have shorter tenures, reflecting a right-skewed distribution.

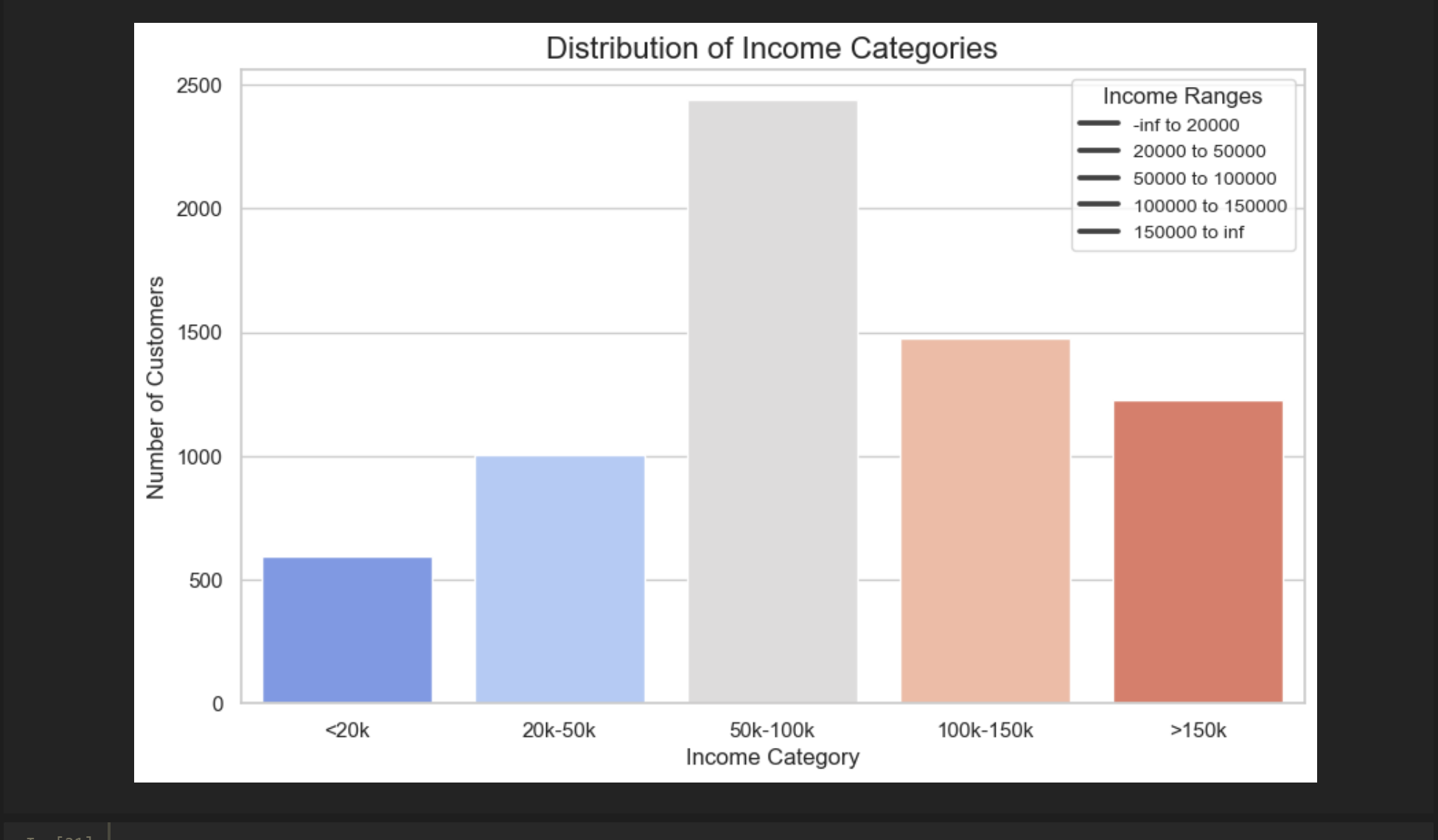




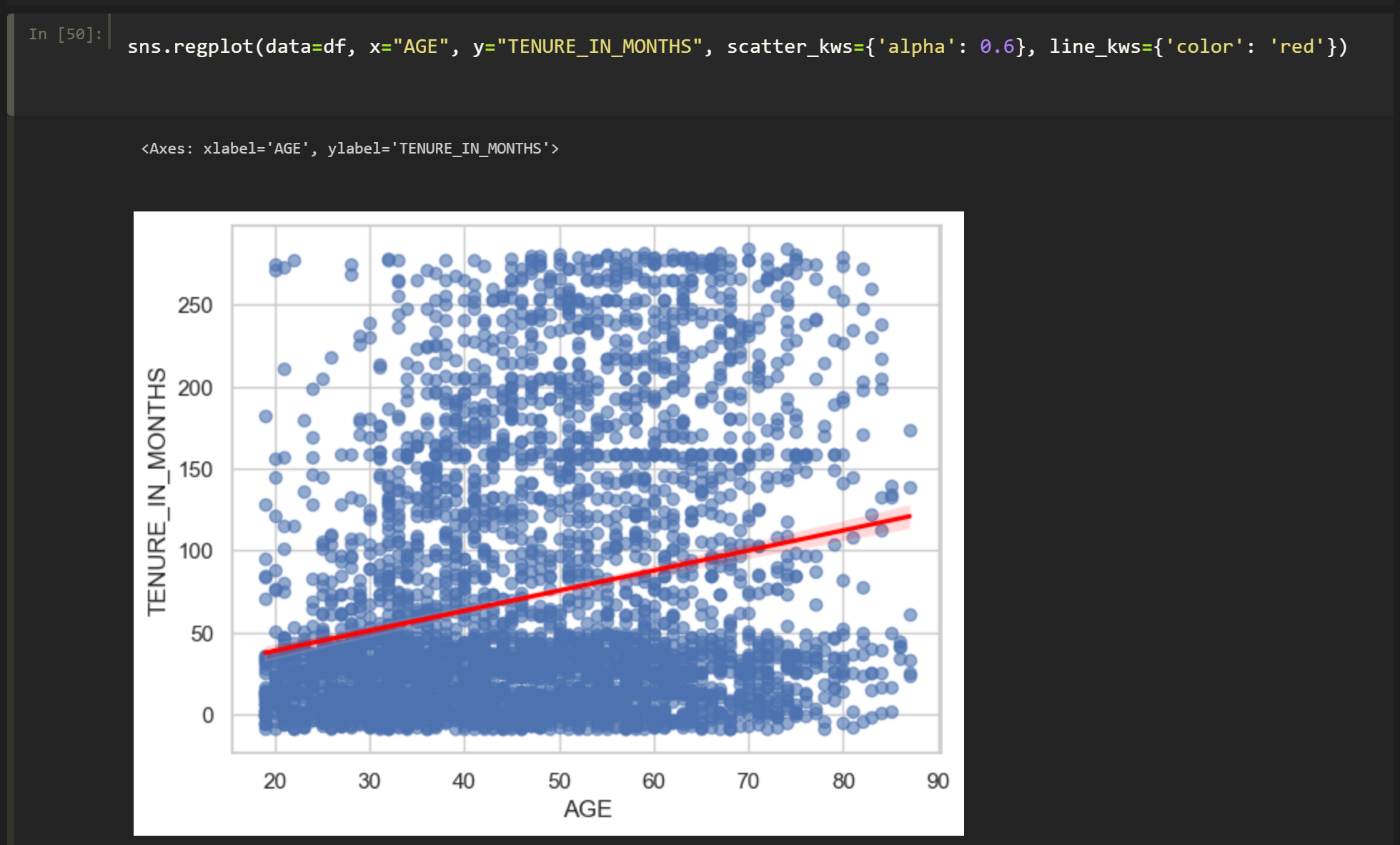
1. **Leading State:**
   * **Texas (TX)** has the highest number of customers by a significant margin, suggesting a strong presence or popularity in this state. It is almost leading by 1.75 times the second highest which is considerably a lot.
2. **Other Prominent States:**
   * **Florida (FL)** follows in second place, but the gap between TX and FL indicates that while FL has a substantial number of customers, it is much lower than that of TX.
   * **Ohio (OH)** and **California (CA)** are also among the top four, suggesting these states have large customer bases as well.
3. **Mid-tier States:**
   * States like **Colorado (CO)**, **Michigan (MI)**, **Illinois (IL)**, and **Virginia (VA)** show moderate customer counts, clustering between 500–1000 customers.
4. **Lower-end States:**
   * The states towards the bottom of the list, such as **Tennessee (TN)** and **Minnesota (MN)**, have customer counts that are significantly lower, indicating either lower market penetration or possibly a smaller population base in comparison to the leading states.

This chart provides a clear visual representation of customer distribution across states. Companies may use this data to tailor their marketing strategies, product offerings, and service expansions by focusing on states with high customer densities like Texas and Florida while also considering opportunities in lower-performing states.





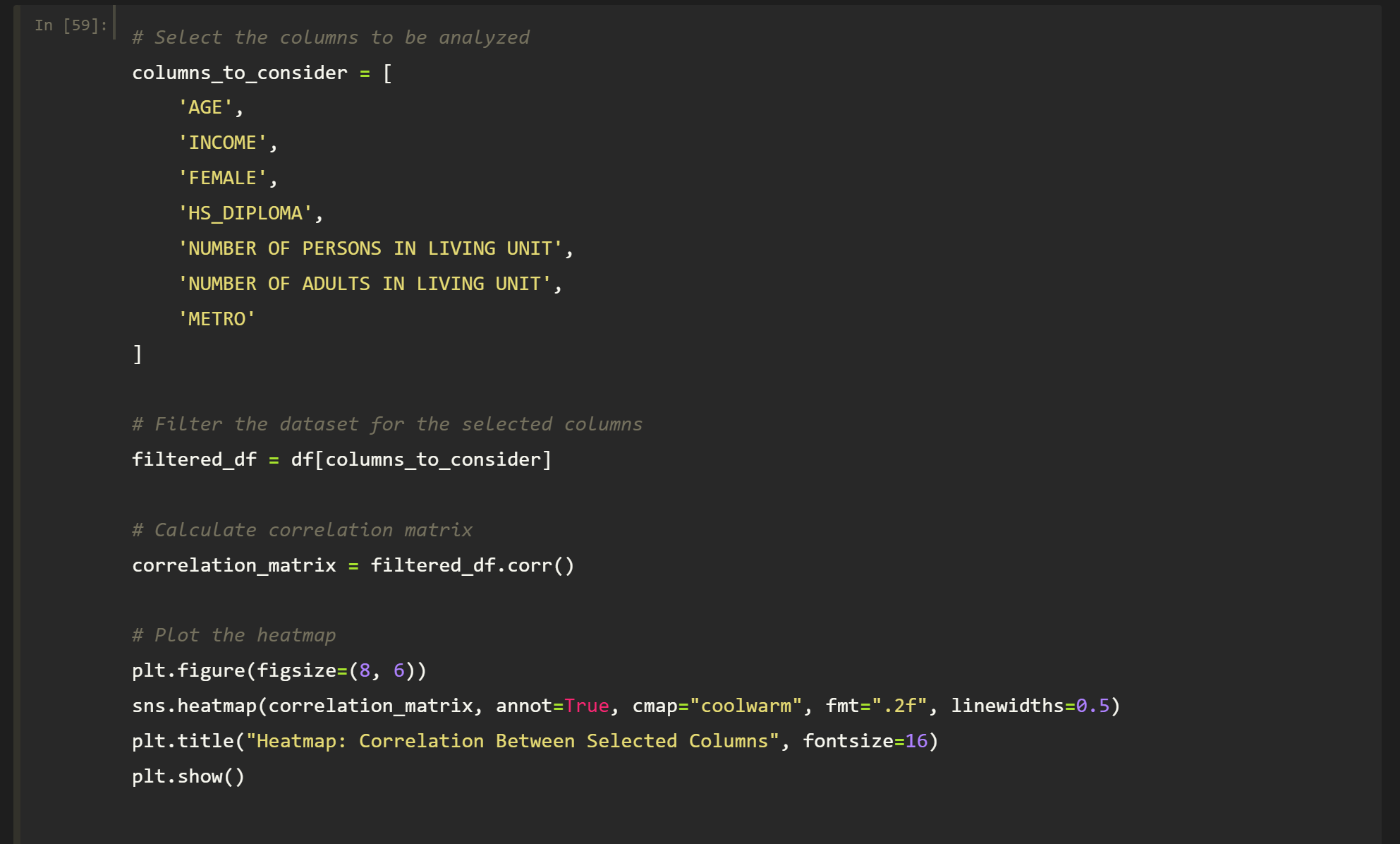
The bar chart shows customer distribution across income categories. Most customers (over 2,500) belong to the $50,000 - $100,000 income range, indicating a concentration in middle income, while the < $20,000 category has the fewest customers. This suggests a skew towards middle-income earners among the customer base.

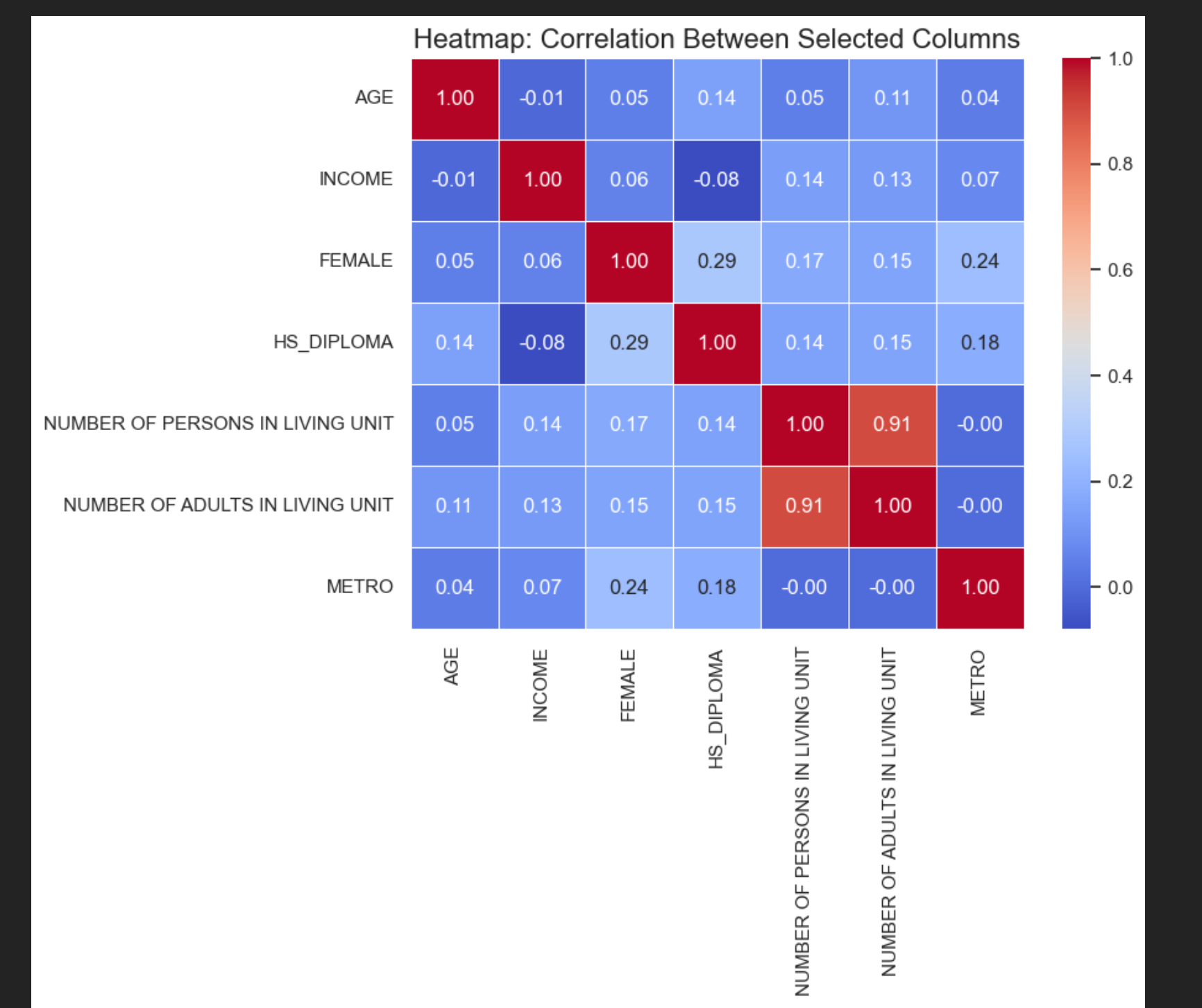


The scatter plot displays the relationship between two variables: age (on the x-axis) and tenure in months (on the y-axis). Here's a breakdown of its meaning:

1. **Trends:** The red line indicates a trend, suggesting that, generally, as age increases, tenure also tends to increase but the relationship appears to be weak. This could imply that older individuals might have longer tenure in their positions.
2. **Variability:** The wide distribution of points indicates substantial variability in tenure across all ages, with many younger individuals showing similar or low tenure levels.

Overall, the plot provides insights into how age might relate to tenure, but the correlation appears to be weak, indicating other factors may also play significant roles.



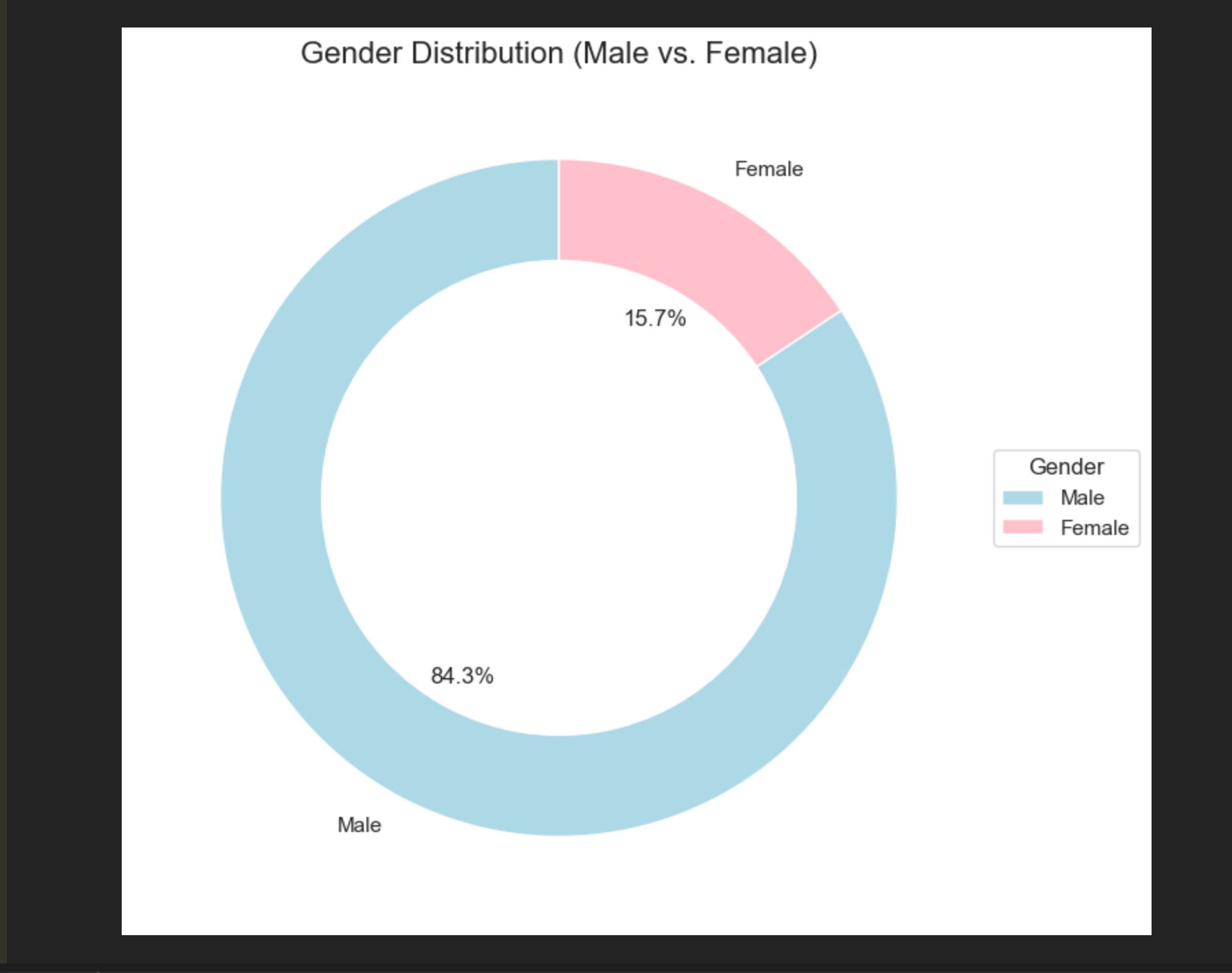


This heatmap shows the correlation between various demographic factors. Here are the key points:

1. **Strongest Correlations**:
   * **Number of Persons in Living Unit & Number of Adults in Living Unit**: 0.91 (very high correlation)
   * **HS Diploma & Female**: 0.29 (moderate correlation)
2. **Other Correlations**:
   * **Age & HS Diploma**: 0.14
   * **Metro & Female**: 0.24
   * **Income & Female**: 0.06 (weak correlation)
3. **Weak or No Correlations**:
   * Most correlations with **Income** and **Metro** are low, indicating little to no relationship with other factors.

Overall, the number of individuals in a living unit is closely related to the number of adults present, while gender and education have some moderate ties





* **Male:** 84.3% (represented by the blue portion of the circle)
* **Female:** 15.7% (represented by the pink portion of the circle)

The chart visually emphasizes the larger segment that corresponds to males compared to females, indicating a significant male majority in this dataset. The percentage labels aid in understanding the exact proportions of each gender in the context of the total population represented.