US Household Income Data Analysis

This document provides a detailed overview of the data cleaning process and the subsequent exploration and analysis of US Household Income data. The aim of the analysis is to identify trends, regional disparities, and the relationship between geographic factors and income levels across the US.

Data Sources

- **Tables Used**:
    - us_project.us_household_income
    - us_project.us_household_income_statistics

Data Cleaning Process

The data cleaning process focused on identifying and correcting issues such as duplicate records, inconsistencies in field values, and missing data to ensure the dataset was in a clean and consistent format for analysis.

1. Identifying and Removing Duplicates

- **Household Income Data**: Duplicates were identified based on the id column in the us_household_income table. The ROW_NUMBER() function was used to flag and remove duplicate records, retaining only the first occurrence of each id.
- **Outcome**: Duplicates were removed, and the dataset is now free of redundant records.
- **Household Income Statistics Data**: A check for duplicates in the us_household_income_statistics table showed no duplicates, ensuring data consistency.

2. Correcting State Names

- Errors in the state_name field were found (e.g., "georia" instead of "Georgia" and "alabama" written in lowercase).
- **Action Taken**: The state names were corrected to ensure proper spelling and consistent case formatting.

3. Correcting Missing Place Information

- Missing place data was identified for "Autauga County," where the place value was absent.
- **Action Taken**: The missing place value was updated for the specified county and city.

4. Correcting the Type Field

- An inconsistency was found where "Boroughs" was used instead of "Borough."
- **Action Taken**: The type field was updated to standardize the value.

5. Handling Missing or Invalid Values

- A check for missing or invalid values in the ALand and AWater fields was performed.

- **Outcome**: No records had both values missing or set to zero at the same time, so no changes were required.

Summary of Data Cleaning Actions

1. **State Name Corrections**: Fixed spelling and case issues in state names.
2. **Duplicate Removal**: Removed duplicate records from the us_household_income table.
3. **Missing Place Data**: Updated missing place information for specific counties.
4. **Type Field Correction**: Standardized the values in the type field.
5. **Validation of Numeric Fields**: Ensured no invalid zero or missing values in ALand and AWater.

Data Exploration and Analysis

1. Land and Water Area Analysis by State

- **Largest States by Land Area**: The top 10 states with the largest land areas were identified.
- **Largest States by Water Area**: States were ranked by water area, highlighting those with substantial water coverage.

2. Income Statistics Merging and Filtering

- The data was enriched by joining the us_household_income and us_household_income_statistics tables on the unique id field, adding additional metrics such as mean and median income levels.
- Rows with missing or zero income values were filtered to improve data quality.

3. State-Level Income Averages

- **States with the Highest Average Income**: The top 10 states with the highest average household income were identified.
- **States with the Lowest Average Income**: The bottom 5 states with the lowest average income were highlighted.
- **Highest and Lowest Median Incomes**: States with the highest and lowest median household incomes were ranked to illustrate regional income disparities.

4. Income Analysis by Area Type

- **Highest Income Areas by Type**: Some area types, such as municipalities, showed significantly higher average and median income levels. However, some of these areas had limited sample sizes (e.g., single entries).

- **Type Count Filtering**: Area types with fewer than 100 instances were excluded from the analysis to focus on more representative categories.

5. City-Level Income Averages

- The average and median household income was calculated for each city.
- **Top Cities by Household Income**: Cities with the highest income levels were identified, shedding light on regions with particularly high income.

Conclusion

This analysis provides a comprehensive view of household income distribution across the United States, segmented by state, area type, and city. The insights reveal significant disparities in income levels across different regions and geographic categories. By merging household income data with additional income statistics, this analysis enriches our understanding of income distribution and regional disparities, offering valuable context for policymakers, researchers, and businesses interested in the US income landscape.