Demographic Data Analysis Project - Completed

Project Overview

This project involved analyzing demographic data using Pandas. The dataset contained demographic information extracted from the 1994 Census database, testing data manipulation, filtering, and analysis skills using Python's Pandas library.

Dataset Description

The dataset contains the following columns:

- **age**: Age of the individual
- **workclass**: Type of employment (e.g., Private, State-gov, Self-emp-not-inc)
- **fnlwgt**: Final weight (census sampling weight)
- **education**: Education level (e.g., Bachelors, HS-grad, Masters)
- **education-num**: Numerical representation of education level
- **marital-status**: Marital status (e.g., Married-civ-spouse, Never-married)
- **occupation**: Job occupation (e.g., Adm-clerical, Exec-managerial)
- **relationship**: Relationship status (e.g., Husband, Wife, Not-in-family)
- **race**: Race/ethnicity
- **sex**: Gender (Male/Female)
- **capital-gain**: Capital gains income
- **capital-loss**: Capital losses
- **hours-per-week**: Hours worked per week
- **native-country**: Country of origin
- **salary**: Income level (<=50K or >50K)

Sample Data Preview

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female

Analysis Results

The following questions were successfully analyzed using Pandas:

Questions Completed

1. ✅ **Race Distribution**: Analyzed the count of people by race using the `race` column, returning a Pandas Series with race names as index labels.
2. ✅ **Average Age of Men**: Calculated the average age of men in the dataset using data filtering and aggregation.
3. ✅ **Bachelor's Degree Percentage**: Determined the percentage of people who have a Bachelor's degree through education column analysis.
4. ✅ **Advanced Education High Earners**: Computed the percentage of people with advanced education (`Bachelors`, `Masters`, or `Doctorate`) who make more than 50K.
5. ✅ **Non-Advanced Education High Earners**: Calculated the percentage of people without advanced education who make more than 50K.
6. ✅ **Minimum Working Hours**: Found the minimum number of hours worked per week using data aggregation.
7. ✅ **Minimum Hours High Earners**: Analyzed the percentage of people working minimum hours who earn more than 50K.

8. ✅ **Country with Highest Earning Percentage**: Identified the country with the highest percentage of people earning >50K and calculated that percentage.
9. ✅ **Most Popular Occupation in India**: Determined the most popular occupation for high earners (>50K) specifically in India.

Technical Approach

The project successfully utilized various Pandas methods and techniques:

- **Data Filtering**: Used boolean indexing to filter data by gender, education level, and country
- **Aggregation Functions**: Applied `value_counts()`, `mean()`, `min()`, and `groupby()` operations
- **Percentage Calculations**: Computed percentages using data manipulation and division operations
- **Series Operations**: Created and manipulated Pandas Series for race distribution analysis
- **Cross-tabulation**: Analyzed relationships between multiple variables (country vs. salary, occupation vs. location)

Project Completion Status

✅ **COMPLETED** - All 9 analysis questions have been successfully answered using appropriate Pandas methodologies. The project demonstrated proficiency in data manipulation, statistical analysis, and insight extraction from demographic datasets.

Data Source

Dataset Source: Dua, D. and Graff, C. (2019). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science.

Key Insights Discovered

Through this analysis, several important demographic patterns were revealed:

- Comprehensive race distribution across the 1994 census sample
- Gender-based age demographics and earning patterns
- Education level correlations with income brackets
- International comparison of high-earning populations
- Occupation preferences in specific geographic regions

Skills Demonstrated

This completed project showcased proficiency in:

- **Data Exploration**: Understanding dataset structure and content
- **Pandas Operations**: Advanced filtering, grouping, and aggregation
- **Statistical Analysis**: Percentage calculations and comparative analysis
- **Data Interpretation**: Drawing meaningful insights from census data