Midterm Task: Analysis of Expenditures

4780/6780 Fundamentals of Data Science

Kiril Kuzmin

Submit by 11pm, March 9, 2025

1 Introduction

Most states in the USA provide services and support to individuals with developmental disabilities (e.g., intellectual disability, cerebral palsy, autism, etc.) and their families. Imagine you are working for the Georgia Department of Behavioral Health and Developmental Disabilities. Your task is to analyze the provided dataset and determine **whether there is evidence of discrimination in the allocation of expenditures to Hispanic and White non-Hispanic consumers**.

2 Data Preprocessing

Familiarize yourself with the dataset `dat.csv` before proceeding with the analysis. Then start with introducing a new categorical variable, “`Age Cohort`”, using the following six groups:

- 0–5 years
- 6–12 years
- 13–17 years
- 18–21 years
- 22–50 years
- 51+ years

3 Analysis and Research Questions

Your analysis should address the following key questions:

A. Data Quality:

- Are there any missing or erroneous values in the dataset?
  Action: Remove all rows with missing or erroneous values. How many row did your delete?

- Are there any outliers in the *expenditure* data? (Assess using the interquartile range).
  Action: Exclude all such rows from further consideration.
  How many row did your delete?

B. Expenditure Analysis:

- Compute the average expenditure for the following groups:
  - All males
  - All Hispanics
  - All White non-Hispanic males
- Determine the *median* annual expenditure for each age cohort.
- Identify which gender has the highest average annual expenditure and calculate the expenditure difference between males and females.
- Calculate the total annual expenditures for each ethnic group.

C. Demographic Analysis:

- Identify the ethnic group with the highest number of consumers in the 22–50-year age cohort.
- Determine the most populous age cohort for each gender.

D. Main Question: Is There Discrimination/Bias?

- What percentage of all customers are *Hispanic* and *White non-Hispanic*?
- Compare expenditures between *Hispanic* and *White non-Hispanic* consumers *in total* and *across different age cohorts.*
- Assess if there is evidence of racial bias in the allocation of expenditures.

4 Submission Requirements

Submit your work on iCollege **by 11pm, March 9**:

1. A **1–2 page PDF report** discussing only the main question: *if there is evidence of racial bias in the allocation of expenditures between Hispanic and White non-Hispanic consumers.* The report should include necessary figures to effectively illustrate your findings.
2. An **IPython Notebook (ipynb file)** containing your data exploration, which answers all the questions listed above.