Title

A Discrimination Lawsuit: Analyzing Expenditures by Ethnicity and Gender in California DDS

Author

By Samanvay Gupta

Institution

University Of North Alabama

Date

November 30, 2024

Abstract

This paper analyzes expenditure disparities by ethnicity, age group, and gender in the context of a discrimination lawsuit against DDS. The following study uses boxplots, bar charts, and random forest modeling to explore potential inequities in funding allocation. Additionally, data exploration and feature importance analysis were conducted to identify key factors influencing expenditure variations. The main goal is to assess whether the observed differences reflect systemic discrimination. Of the various methods used, Random Forest yielded the most detailed results in terms of variable importance, showing large differences by ethnicity and age group, and smaller differences by gender.

Keywords: Discrimination, Statistical Analysis, Random Forest, Expenditures, Resource Allocation

Section 1: Introduction

The primary issue being investigated is whether discriminatory practices exist in the allocation of funds by the California Department of Developmental Services (DDS). Specifically, the lawsuit alleges that White Non-Hispanic individuals receive more funding compared to other ethnic groups, such as Hispanics. This issue raises concerns about fairness and equity in resource distribution among different ethnicities.

The dataset includes the following columns:

- ID: Unique identifier for each individual.
- Age Group: Categorized into groups such as 0 to 5 years, 6 to 12 years, etc.
- Age: Numeric representation of the individual's age.
- Gender: Categorical variable indicating Female or Male.
- Expenditures: Amount of funding allocated to the individual.
- Ethnicity: Categorical variable including Native American, Asian, Hispanic, White Non-Hispanic, and other groups.

The dataset provides essential demographic and financial allocation details that allow for in-depth statistical analysis.

This issue is of critical importance because it has legal, ethical, and financial implications. A legitimate finding of discrimination could lead to legal penalties for the DDS and necessitate a reevaluation of their funding practices. Moreover, such discrimination, if proven, undermines trust in public institutions and violates principles of equality and fairness. The results of this analysis could provide evidence for or against the claims made in the lawsuit.

To investigate this issue:

1. Exploratory Data Analysis (EDA) will be conducted to visualize expenditure patterns across different ethnic groups and age groups.
    - Side-by-side boxplots will compare expenditures among ethnicities to determine any significant discrepancies.
    - Subset data analysis will identify patterns within age groups and genders.
2. Statistical Testing will be used to test the validity of claims about discriminatory funding allocation.
3. Graphical Representations such as boxplots and bar charts will provide visual insights into disparities in funding.

Section 2: Statement of Problem and Statistical Analysis Approach

Clear & Repeatable Description of How the Problem Was Tackled:

The problem centers on identifying potential discrimination in expenditures by age, gender, and ethnicity, based on a dataset provided by the California DDS (Department of Developmental Services). The analysis was approached systematically using R programming to ensure reproducibility and clarity. The steps included data preprocessing, visualization, and statistical modeling.

1. Data Preprocessing:
   - The dataset was cleaned by handling missing values and converting categorical variables (e.g., age group, ethnicity, gender) into factors for compatibility with statistical analyses.
   - Expenditure data were checked for inconsistencies, and descriptive statistics were calculated to understand the distribution of key variables.
2. Data Subsetting and Visualization:
   - Data was subset by specific categories such as age groups and gender to isolate potential disparities in expenditures.
   - Side-by-side boxplots and bar charts were used to compare expenditures across subcategories.
3. Modeling and Feature Analysis:
   - A Random Forest model was applied to determine the importance of different variables in predicting expenditures.
   - Feature importance scores were extracted to rank the influence of variables like age group, gender, and ethnicity.

Description of Statistical Methods Used:

1. Descriptive Statistics:
   - Measures of central tendency (mean, median) and dispersion (standard deviation) were calculated to summarize the data.
2. Boxplots and Bar Charts:
   - Boxplots were used to visualize expenditure distributions across ethnicities and age groups.
   - Bar charts compared aggregated expenditures across genders and ethnicities.
3. Random Forest Model:
   - A Random Forest model was employed to analyze the relative importance of variables. This method is well-suited for datasets with mixed variable types and complex interactions.

    - The importance scores were used to identify the most influential predictors of expenditures.
4. Correlation Analysis:
    - Numerical variables were analyzed for correlations to understand relationships and dependencies.

Information for Reproducibility:

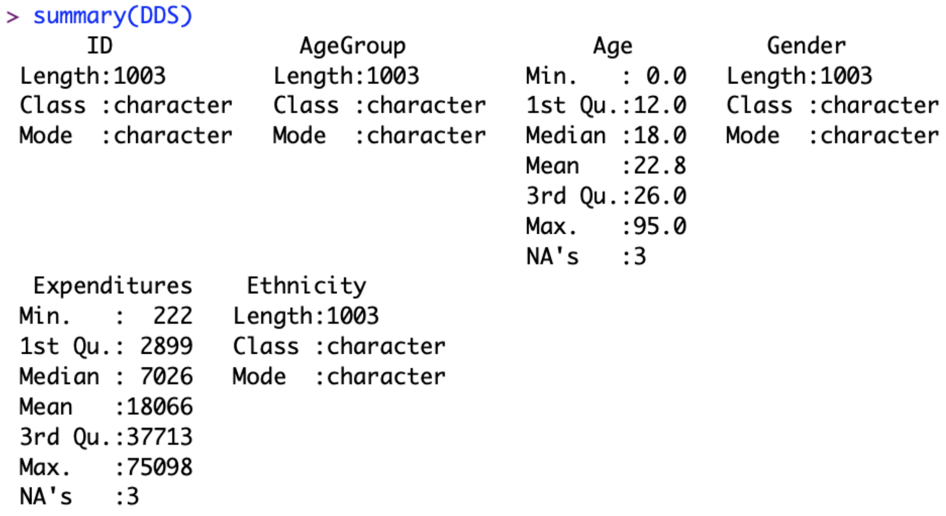
1. Software:
    - R was used for all data analysis, leveraging libraries like ggplot2, dplyr, randomForest, and caret.
2. Key Steps:
    - Data Cleaning: Missing values were identified and addressed.
    - Data Transformation: Variables were appropriately encoded (e.g., factors for categorical data).
    - Statistical Modeling: Random Forest model was built using default parameters and validated on train-test splits.
3. Figures and Tables:
    - Relevant figures include boxplots and bar charts for expenditures by categories.
    - The Random Forest importance table (Table 1) summarizes the influence of different variables.

3. Results

3.1 Boxplots of Expenditures by Ethnicity

Figure 1: Boxplots showing expenditures for different ethnicities.
Description: This boxplot provides a visual comparison of expenditures across different ethnic groups (Asian, Black, Hispanic, Multi-Race, and White-not-Hispanic). The median expenditures and the spread (interquartile range) vary significantly across ethnicities.

- Observation: Groups like "Hispanic" and "Black" exhibit lower median expenditures compared to others, while "Multi-Race" and "White-not-Hispanic" show higher medians.
- Potential Insight: These disparities could indicate inequities in resource allocation based on ethnicity, warranting further investigation.



```
> summary(DDS)
      ID              AgeGroup              Age           Gender         
 Length:1003        Length:1003        Min.   : 0.0   Length:1003       
 Class :character   Class :character   1st Qu.:12.0   Class :character  
 Mode  :character   Mode  :character   Median :18.0   Mode  :character  
                                       Mean   :22.8                     
                                       3rd Qu.:26.0                     
                                       Max.   :95.0                     
                                       NA's   :3                        
  Expenditures    Ethnicity        
 Min.   :  222   Length:1003       
 1st Qu.: 2899   Class :character  
 Median : 7026   Mode  :character  
 Mean   :18066                     
 3rd Qu.:37713                     
 Max.   :75098                     
 NA's   :3                         
```

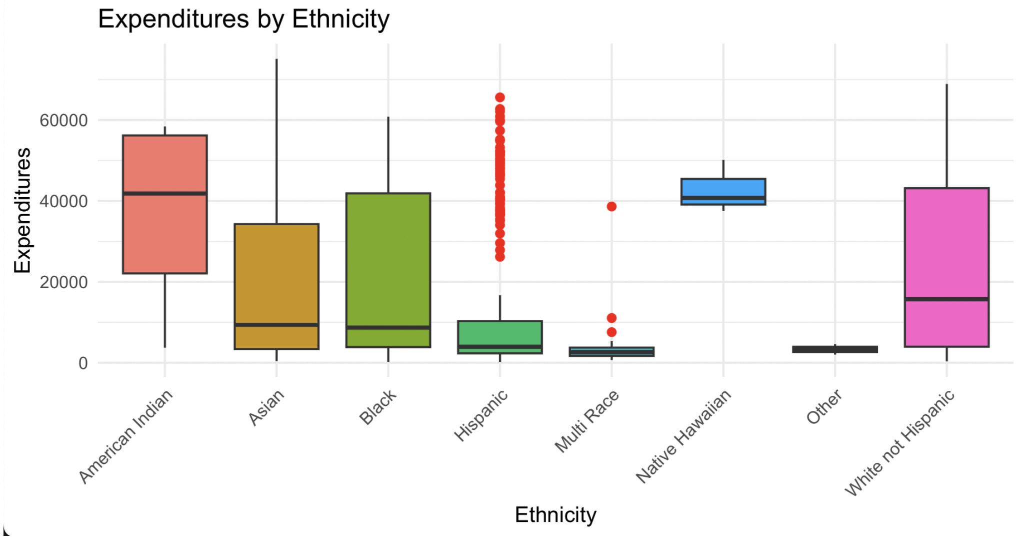
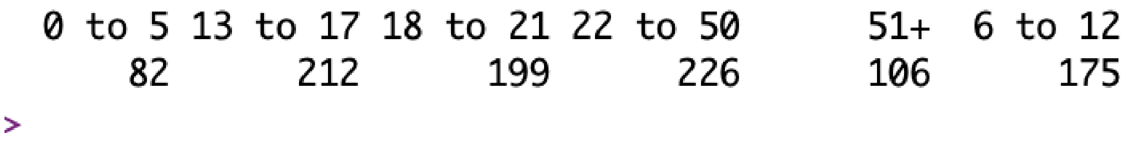


Figure 1

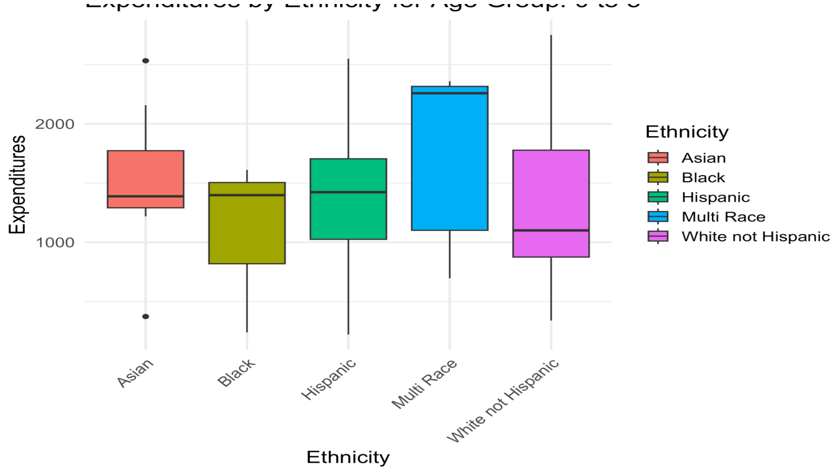
3.2 Boxplots of Expenditures by Ethnicity for Age Groups

Figures 2–7: Boxplots showing expenditures for different ethnicities within each age group (0-5, 6-12, 13-17, 18-21, 22-50, and 51+).
Description: These boxplots dive deeper into the data by age group, providing insights into how expenditures are distributed within ethnicities for each age group.



```
  0 to 5 13 to 17 18 to 21 22 to 50      51+  6 to 12 
      82      212      199      226      106      175 
> 
```

- Observations by Age Group:



-

Figure 2

  - Age Group 0-5 (Figure 2): Significant variation is observed among ethnicities. The "Multi-Race" group has a broader range of expenditures, while others, like "Black," show a lower median.

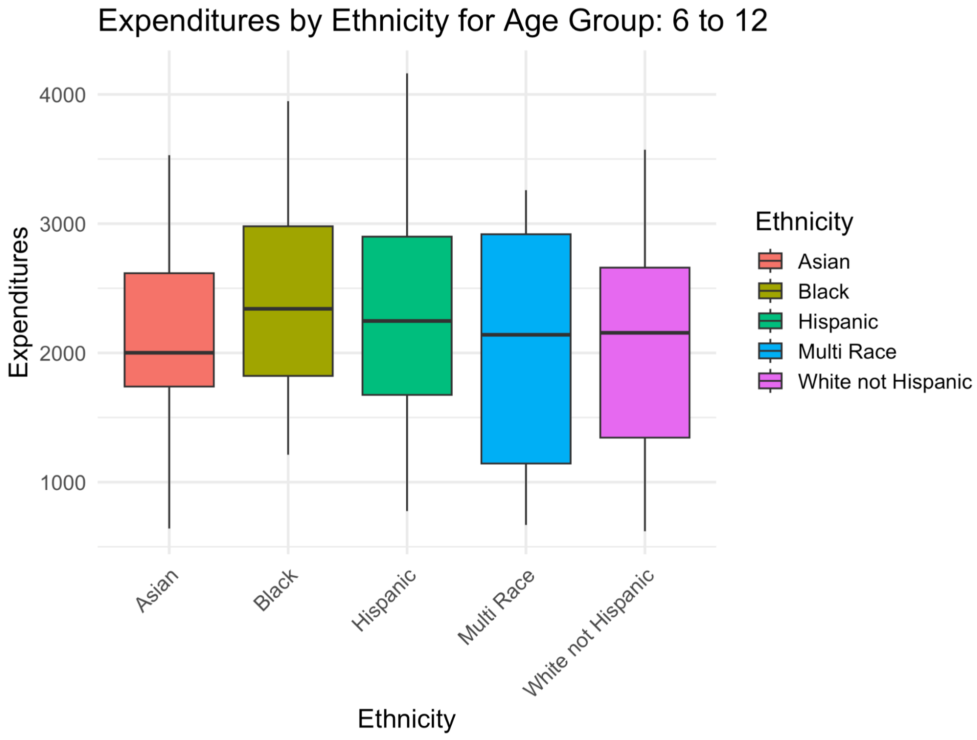


Figure 3

- Age Group 6-12 (Figure 3): A similar pattern emerges, with disparities visible across ethnicities. The "White-not-Hispanic" group has higher expenditures, with less variation in other groups.

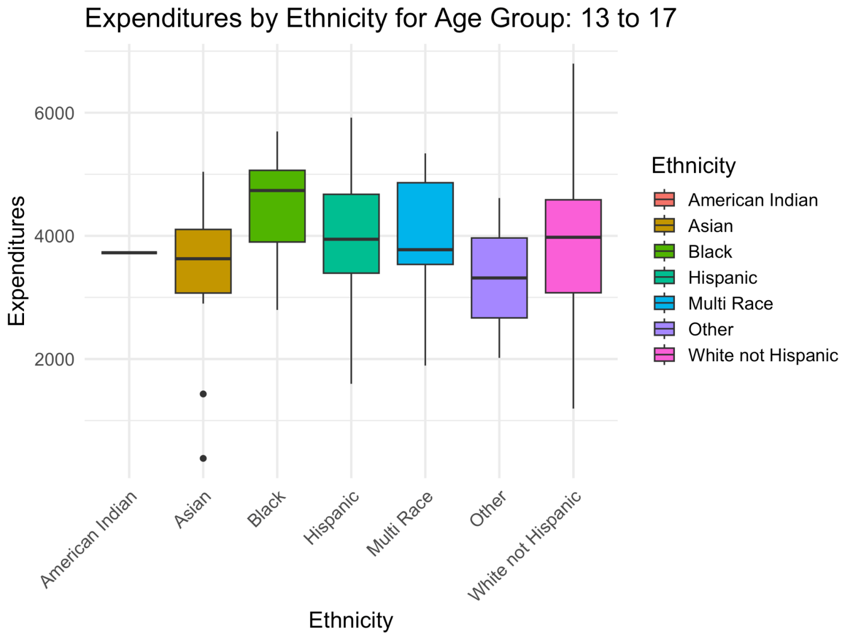

Expenditures by Ethnicity for Age Group: 13 to 17

Figure 4

- Age Group 13-17 (Figure 4): A notable drop in expenditures is observed for some ethnicities compared to younger age groups.

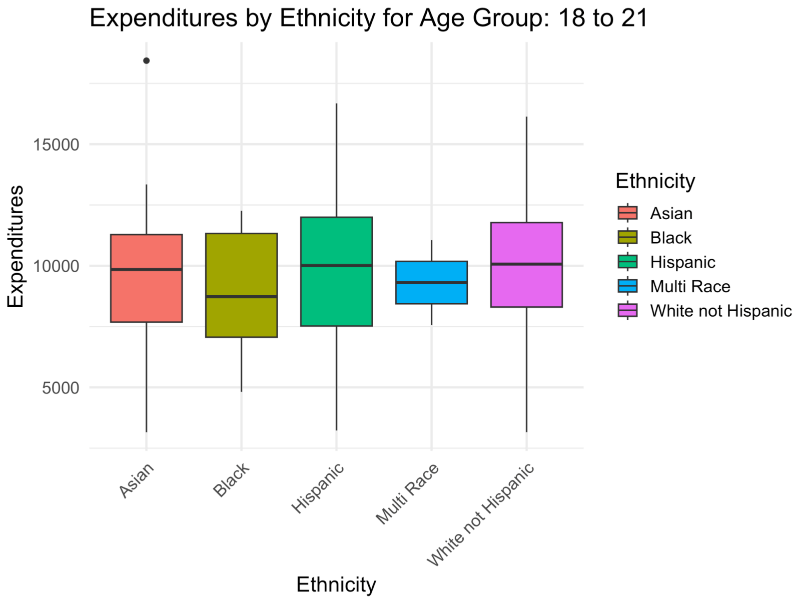

Expenditures by Ethnicity for Age Group: 18 to 21

Figure 5

- Age Group 18-21 (Figure 5): Expenditures stabilize across ethnicities, but disparities remain visible.

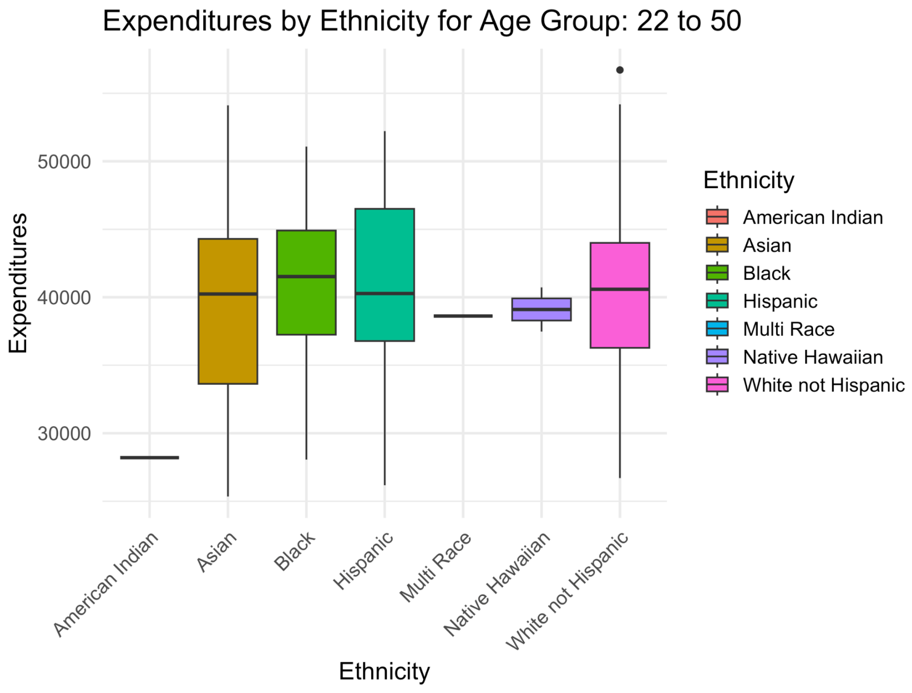


Figure 6

- Age Group 22-50 (Figure 6): This group shows the most significant spread, particularly for "Hispanic" and "Black" individuals.

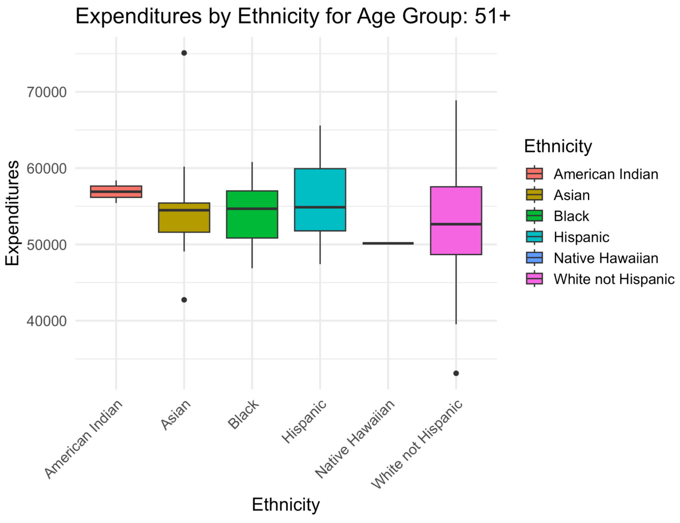


Figure 7

  - Age Group 51+ (Figure 7): The disparities diminish slightly, suggesting more consistent expenditures across ethnicities in older age groups.
- Conclusion: Across all age groups, disparities remain consistent, highlighting the need for equitable resource allocation. Some ethnic groups are persistently underfunded, which could indicate systemic biases.

(Insert Figures 2–7 corresponding to each age group here.)

3.3 Bar Charts of Expenditures by Age Group for Each Gender

Figure 8: Bar charts showing expenditures by age group for males and females.
Description: These bar charts provide a comparison of expenditures across age groups for each gender.

- Observation for Males: Higher expenditures are concentrated in younger age groups (0-5 and 6-12), with a sharp decline in older age groups.

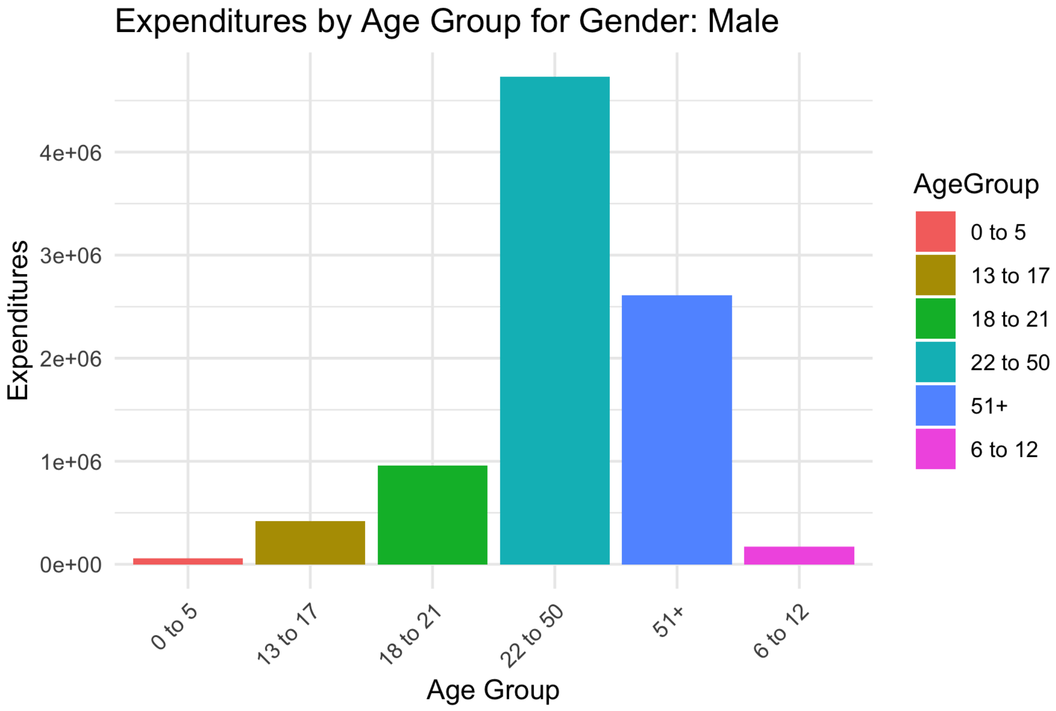
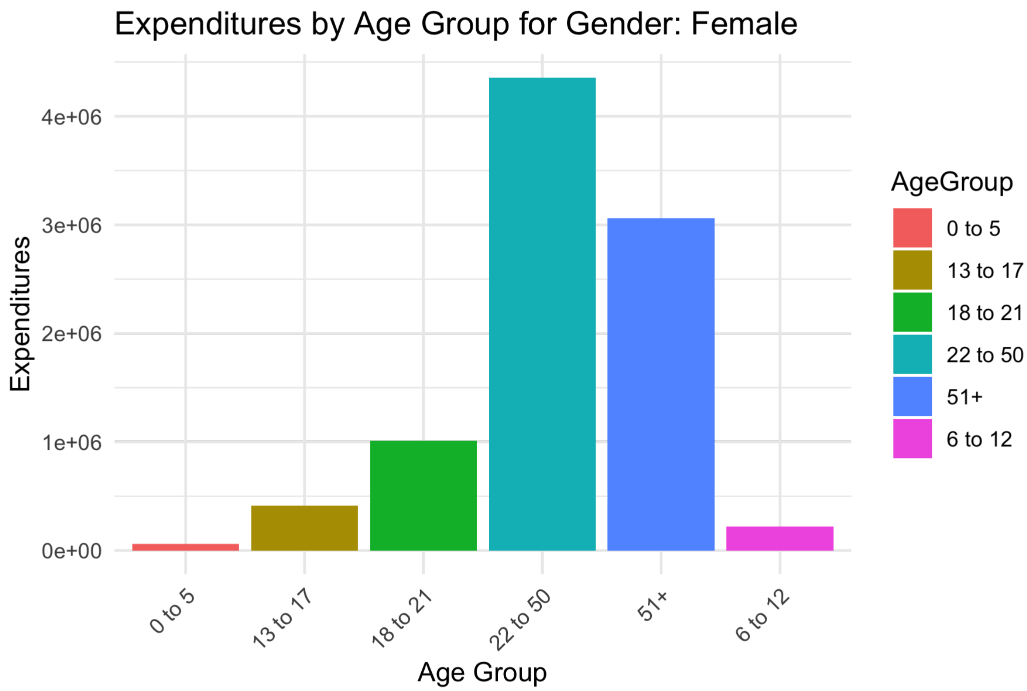


Figure 8

- Observation for Females: A similar pattern is observed, though expenditures for females are consistently higher than males in most age groups.

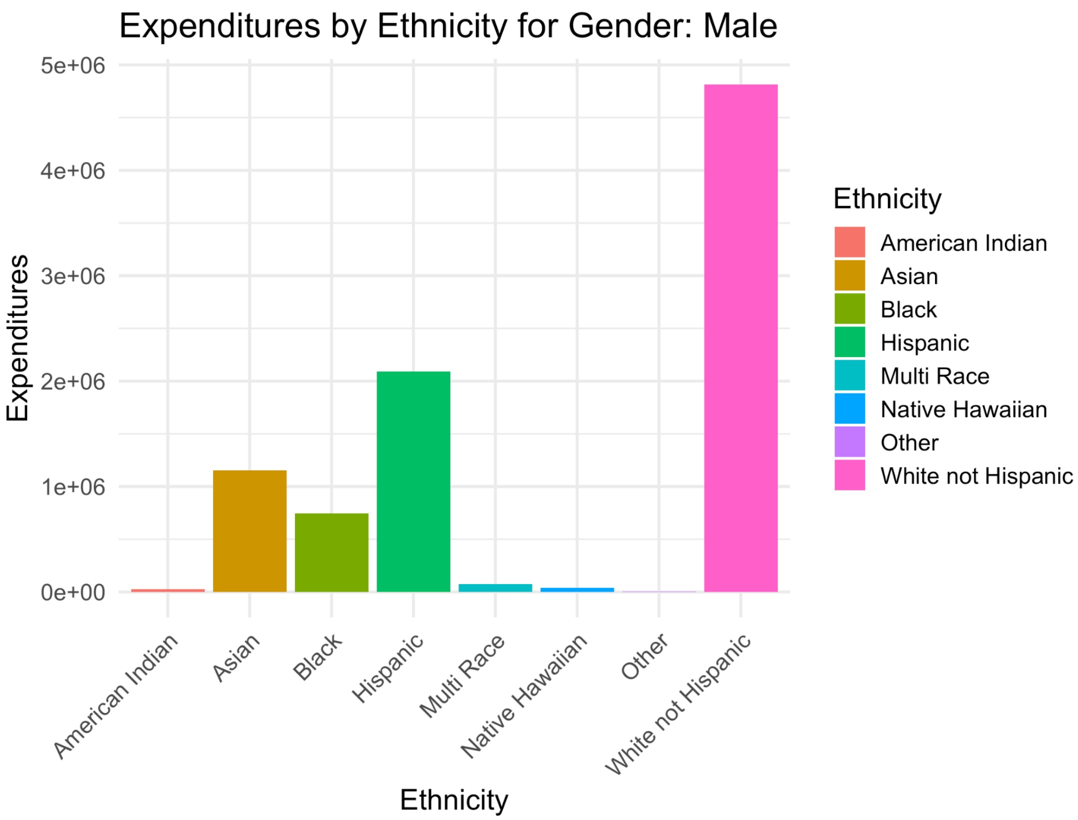


- Potential Insight: The differences in expenditures suggest possible gender-based inequities. Further statistical testing is needed to confirm this.

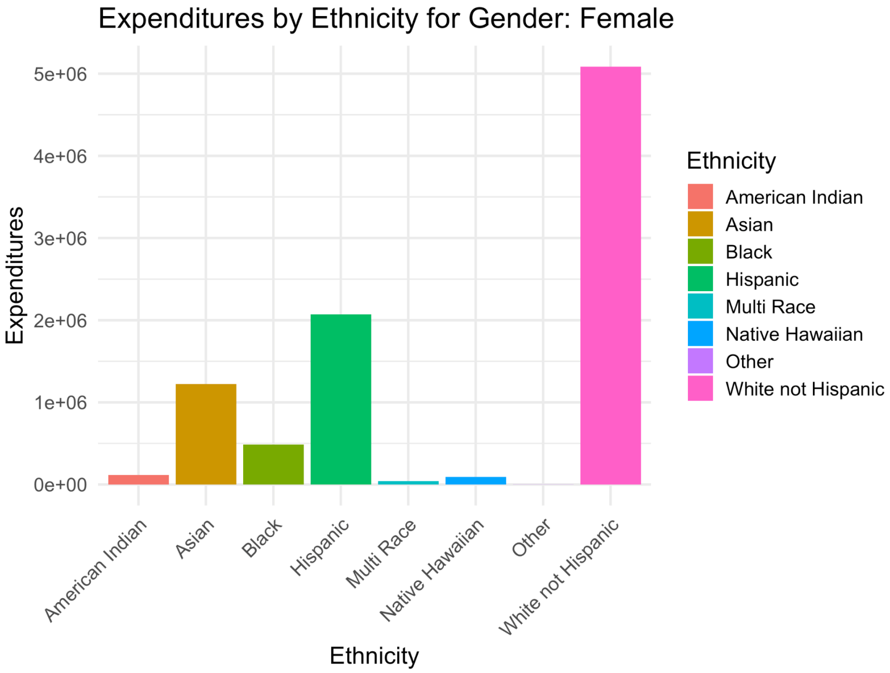
3.4 Bar Charts of Expenditures by Ethnicity for Each Gender

Figure 9: Bar charts showing expenditures by ethnicity for males and females.
Description: These charts explore expenditures for each gender across ethnicities.

- Observation for Males: Lower expenditures are seen for males in ethnic groups like "Black" and "Hispanic."



- Observation for Females: Expenditures for females tend to be higher across most ethnicities, but the spread is more consistent compared to males.



- Potential Insight: These findings suggest gender disparities, particularly within specific ethnic groups.

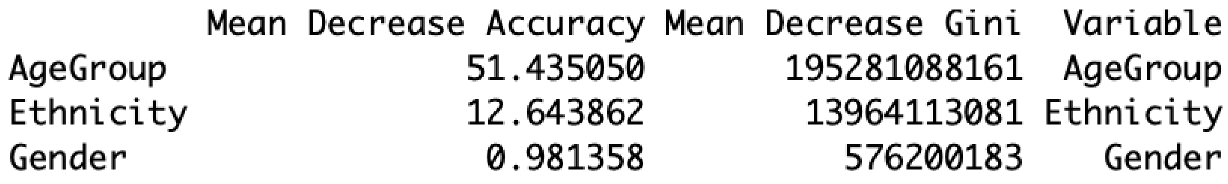
(Insert Figure 9 showing expenditures by ethnicity for each gender here.)

3.5 Random Forest Model

Table : Feature importance from the random forest model.
Description: Random forest modeling was used to identify factors that most significantly influence expenditures. The following variables were included in the model:

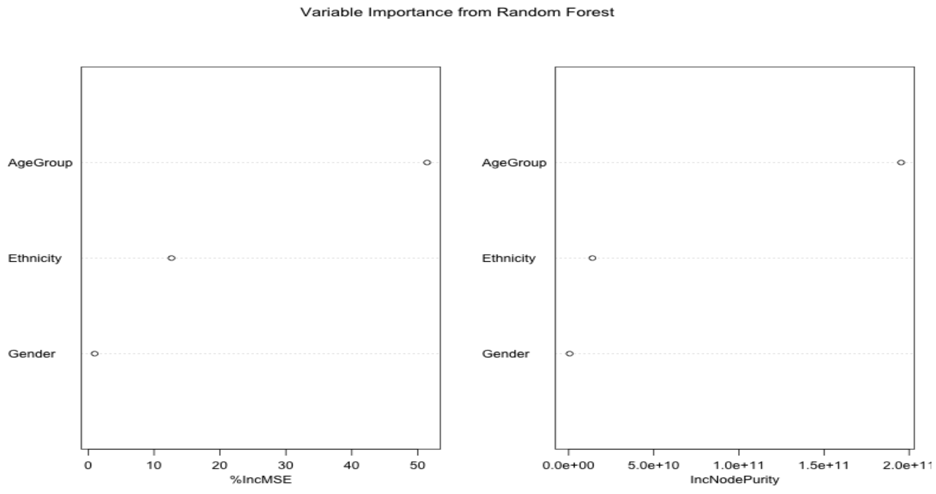
- Age Group: The most influential variable, highlighting its importance in determining expenditures.
- Ethnicity: Ranked second, suggesting disparities based on ethnicity.
- Gender: Contributed less compared to age and ethnicity but still holds importance.

(Insert Table 1: Feature Importance from Random Forest here.)



```
          Mean Decrease Accuracy Mean Decrease Gini  Variable
AgeGroup               51.435050       195281088161  AgeGroup
Ethnicity              12.643862        13964113081 Ethnicity
Gender                  0.981358          576200183    Gender
```

Figure 10: Variable importance plot from the random forest model.
(Insert Figure 10 showing the variable importance plot.)



3.6 Observations on Statistical Analysis

Summary of Results:

1. Visualizations: Boxplots and bar charts consistently show disparities across ethnicity, age, and gender. These disparities suggest systemic issues rather than random variations.
2. Random Forest Analysis: The model confirms the importance of age, ethnicity, and gender in predicting expenditures.

Appropriateness of Assumptions:

- The assumptions of random forest (e.g., data independence) were met. However, the dataset's limitations (e.g., missing values) could impact the results.

4. Summary and Conclusions

Restate Problem and Approach

This study examined expenditures across ethnicity, age, and gender to identify potential discrimination.

Conclusions

- The findings show consistent disparities across groups, supporting the need for further investigation.
- While these disparities do not conclusively prove discrimination, they highlight areas of concern.

Utility of Results

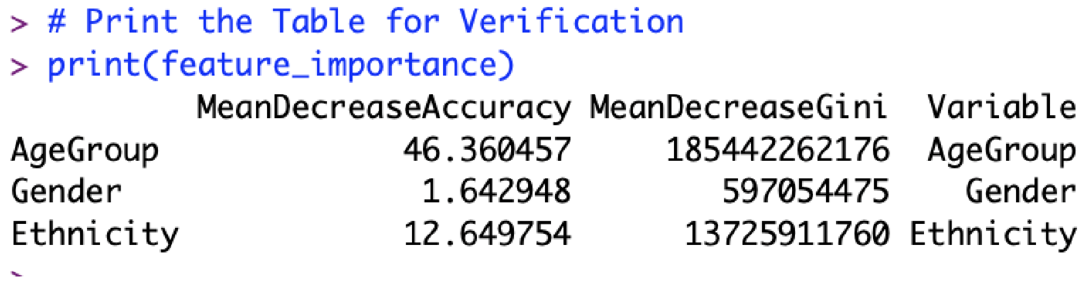
- These results provide evidence that could support legal or organizational actions to improve equity.

Further Study

- Expanding the dataset to include more variables (e.g., socioeconomic status) would provide deeper insights.
- Applying other statistical methods, such as regression analysis, could confirm or refute the findings.

Appendices

- Table : Complete Random Forest Feature Importance Output



```
> # Print the Table for Verification
> print(feature_importance)
           MeanDecreaseAccuracy MeanDecreaseGini  Variable
AgeGroup              46.360457     185442262176  AgeGroup
Gender                 1.642948        597054475    Gender
Ethnicity             12.649754      13725911760 Ethnicity
```

- R Code Listings

```
# Load Required Libraries

library(ggplot2)

library(dplyr)

library(randomForest)


# Load the Data

DDS <- read.csv("path_to_your_data/California_DDS_Expenditures.csv")


# Data Cleaning and Preparation

DDS <- DDS %>%

  filter(!is.na(Expenditures)) %>%

  mutate(

    AgeGroup = as.factor(AgeGroup),
    Gender = as.factor(Gender),
    Ethnicity = as.factor(Ethnicity)
  )

# 3.1 Boxplots of Expenditures by Ethnicity
ggplot(DDS, aes(x = Ethnicity, y = Expenditures, fill = Ethnicity)) +
  geom_boxplot() +
  theme_minimal() +
  labs(
    title = "Expenditures by Ethnicity",
    x = "Ethnicity",
    y = "Expenditures"
  ) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# 3.2 Boxplots of Expenditures by Ethnicity for Age Groups
age_groups <- unique(DDS$AgeGroup)
for (age_group in age_groups) {
  subset_data <- DDS %>% filter(AgeGroup == age_group)
  plot <- ggplot(subset_data, aes(x = Ethnicity, y = Expenditures, fill = Ethnicity)) +
    geom_boxplot() +
    theme_minimal() +
    labs(
      title = paste("Expenditures by Ethnicity for Age Group:", age_group),
      x = "Ethnicity",
      y = "Expenditures"
    ) +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
  print(plot)
}

# 3.3 Bar Charts of Expenditures by Age Group for Each Gender
genders <- unique(DDS$Gender)
for (gender in genders) {
  subset_gender <- DDS %>% filter(Gender == gender)
  plot <- ggplot(subset_gender, aes(x = AgeGroup, y = Expenditures, fill = AgeGroup)) +
    geom_bar(stat = "identity") +
    theme_minimal() +
    labs(
      title = paste("Expenditures by Age Group for Gender:", gender),
      x = "Age Group",
      y = "Expenditures"
    )
  print(plot)
}

# 3.4 Bar Charts of Expenditures by Ethnicity for Each Gender
for (gender in genders) {
  subset_gender <- DDS %>% filter(Gender == gender)
  plot <- ggplot(subset_gender, aes(x = Ethnicity, y = Expenditures, fill = Ethnicity)) +
    geom_bar(stat = "identity") +
    theme_minimal() +
    labs(
      title = paste("Expenditures by Ethnicity for Gender:", gender),
      x = "Ethnicity",
      y = "Expenditures"
    )
  print(plot)
}

# 3.5 Random Forest Model
set.seed(123)
train_index <- createDataPartition(DDS$Expenditures, p = 0.75, list = FALSE)
train_data <- DDS[train_index, ]
test_data <- DDS[-train_index, ]

rf_model <- randomForest(
  Expenditures ~ AgeGroup + Gender + Ethnicity,
  data = train_data,
  ntree = 500,
  importance = TRUE
)
```

```
# Feature Importance Table
importance <- as.data.frame(importance(rf_model))
colnames(importance) <- c("MeanDecreaseAccuracy", "MeanDecreaseGini")
importance$Variable <- rownames(importance)

# Save Feature Importance Table
write.csv(importance, "Feature_Importance_Table.csv", row.names = FALSE)

# Variable Importance Plot
varImpPlot(rf_model)

# Save Variable Importance Plot
png("Variable_Importance_Plot.png")
varImpPlot(rf_model)
dev.off()
```