A study on the alleged case of discrimination

NoSQL/MongoDB

02

The dataset involves a claim of discrimination based on ethnicity. The situation involved an alleged case of discrimination favoring White non-Hispanics over Hispanics in the allocation of funds to over 250,000 developmentally-disabled California Residents.

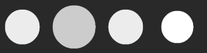
Background

Step 1: Design Mongo Collection

In [11]:
```
db= client['project']
collection= db['first']

df= pd.read_csv('./californiaDataSet.csv')
data= df.to_dict('records')
collection.insert_many(data)
```

Out[11]:
```
InsertManyResult([ObjectId('6619d8b0eb86fc47966602fb'), ObjectId('6619d8b0eb86fc47966602fc'), ObjectId('6619d8b0eb8
6fc47966602fd'), ObjectId('6619d8b0eb86fc47966602fe'), ObjectId('6619d8b0eb86fc47966602ff'), ObjectId('6619d8b0eb86
fc4796660300'), ObjectId('6619d8b0eb86fc4796660301'), ObjectId('6619d8b0eb86fc4796660302'), ObjectId('6619d8b0eb86f
c4796660303'), ObjectId('6619d8b0eb86fc4796660304'), ObjectId('6619d8b0eb86fc4796660305'), ObjectId('6619d8b0eb86fc
4796660306'), ObjectId('6619d8b0eb86fc4796660307'), ObjectId('6619d8b0eb86fc4796660308'), ObjectId('6619d8b0eb86fc4
796660309'), ObjectId('6619d8b0eb86fc479666030a'), ObjectId('6619d8b0eb86fc479666030b'), ObjectId('6619d8b0eb86fc47
9666030c'), ObjectId('6619d8b0eb86fc479666030d'), ObjectId('6619d8b0eb86fc479666030e'), ObjectId('6619d8b0eb86fc479
666030f'), ObjectId('6619d8b0eb86fc4796660310'), ObjectId('6619d8b0eb86fc4796660311'), ObjectId('6619d8b0eb86fc4796
660312'), ObjectId('6619d8b0eb86fc4796660313'), ObjectId('6619d8b0eb86fc4796660314'), ObjectId('6619d8b0eb86fc47966
60315'), ObjectId('6619d8b0eb86fc4796660316'), ObjectId('6619d8b0eb86fc4796660317'), ObjectId('6619d8b0eb86fc479666
0318'), ObjectId('6619d8b0eb86fc4796660319'), ObjectId('6619d8b0eb86fc479666031a'), ObjectId('6619d8b0eb86fc4796660
31b'), ObjectId('6619d8b0eb86fc479666031c'), ObjectId('6619d8b0eb86fc479666031d'), ObjectId('6619d8b0eb86fc47966603
1e'), ObjectId('6619d8b0eb86fc479666031f'), ObjectId('6619d8b0eb86fc4796660320'), ObjectId('6619d8b0eb86fc479666032
1'), ObjectId('6619d8b0eb86fc4796660322'), ObjectId('6619d8b0eb86fc4796660323'), ObjectId('6619d8b0eb86fc479666032
4'), ObjectId('6619d8b0eb86fc4796660325'), ObjectId('6619d8b0eb86fc4796660326'), ObjectId('6619d8b0eb86fc479666032
7'), ObjectId('6619d8b0eb86fc4796660328'), ObjectId('6619d8b0eb86fc4796660329'), ObjectId('6619d8b0eb86fc479666032
a'), ObjectId('6619d8b0eb86fc479666032b'), ObjectId('6619d8b0eb86fc479666032c'), ObjectId('6619d8b0eb86fc479666032
d'), ObjectId('6619d8b0eb86fc479666032e'), ObjectId('6619d8b0eb86fc479666032f'), ObjectId('6619d8b0eb86fc479666033
0'), ObjectId('6619d8b0eb86fc4796660331'), ObjectId('6619d8b0eb86fc4796660332'), ObjectId('6619d8b0eb86fc479666033
```



Step 2: Use Mongo Aggregation Pipeline

In [21]:

```
pipeline = [
    {# check if the person's ethinicity is white not hispanic or hispanic
        '$match': {
            'Ethnicity': {
                '$in': ['White not Hispanic', 'Hispanic']
            }
        }
    },
    {# group data by ethnicity then find descriptive statistics for both groups
        '$group': {
            '_id': '$Ethnicity',
            'average_expenditures': {
                '$avg': '$Expenditures'
            },
            'std_expenditures': {
                '$stdDevSamp': '$Expenditures'
            },
            'max_expenditures': {
                '$max': '$Expenditures'
            },
            'min_expenditures': {
                '$min': '$Expenditures'
            }
        }
    }
]
result = collection.aggregate(pipeline)
```

Step 2: Use Mongo Aggregation Pipeline

In [26]:
```
plotPipe = [
    {
        # Match documents where the Ethnicity field is either 'White not Hispanic' or 'Hispanic'
        '$match': {
            'Ethnicity': {
                '$in': ['White not Hispanic', 'Hispanic']
            }
        }
    }
]
result= list(collection.aggregate(plotPipe))
plotData= pd.DataFrame(result)
plotData.head()
```

Out[26]:

	_id	Id	Age Cohort	Age	Gender	Expenditures	Ethnicity
0	6619a038776ae5c590f19134	10210	13 to 17	17	Female	2113	White not Hispanic
1	6619a038776ae5c590f19135	10409	22 to 50	37	Male	41924	White not Hispanic
2	6619a038776ae5c590f19136	10486	0 to 5	3	Male	1454	Hispanic
3	6619a038776ae5c590f19137	10538	18 to 21	19	Female	6400	Hispanic
4	6619a038776ae5c590f19138	10568	13 to 17	13	Male	4412	White not Hispanic

05

Data Analysis

- Descriptive Statistics
- Graph Visualizations
- T-test

Descriptive Statistics

White people receive $13,632 more than hispanics on average

```
result = collection.aggregate(pipeline)

# empty dict to means
ethAvg= {}

# iterate through docs and print out information
for document in result:
    print(f"Ethnicity: {document['_id']}")
    print(f"Mean: {document['average_expenditures']:.2f}")
    print(f"Standard Deviation: {document['std_expenditures']:.2f}")
    print(f"Min: {document['min_expenditures']:.2f}")
    print(f"Max: {document['max_expenditures']:.2f}")
    print()
    ethnicity= document['_id']
    avgExp= document['average_expenditures']
    ethAvg[ethnicity]= avgExp
diff= ethAvg['White not Hispanic'] - ethAvg['Hispanic']
print(f'White people receive ~${diff:0.2f} more than hispanics on average.')
```

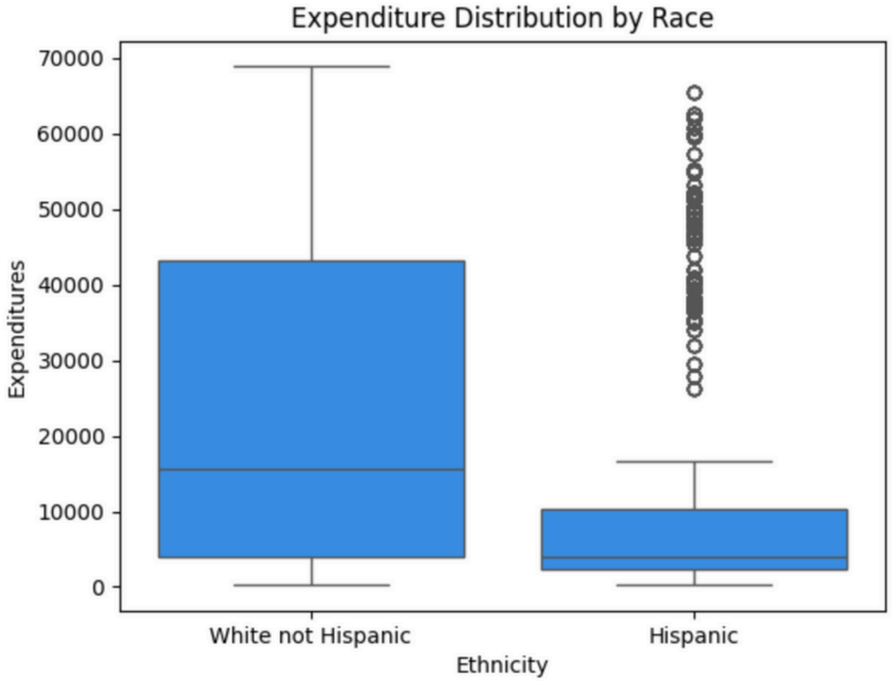
White not Hispanic

- Mean: 24697.55
- Standard Deviation: 20582.34
- Min: 340.00
- Max: 68890.00

Hispanic

- Mean: 11065.57
- Standard Deviation: 15612.01
- Min: 222.00
- Max: 65581.00

Expenditures by Ethnicity

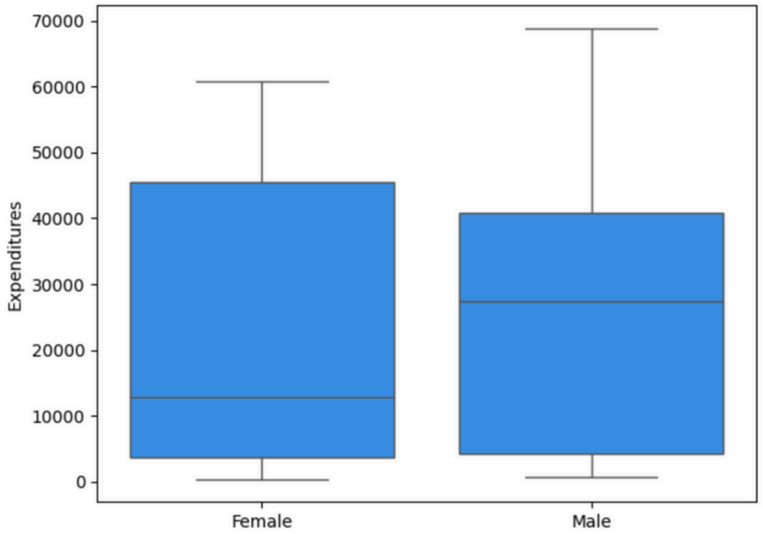


Significant disparity in the allocation of funds.

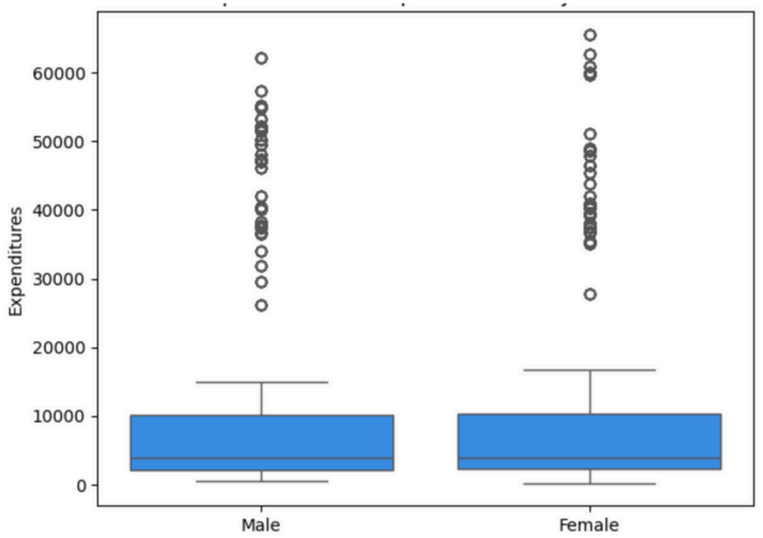
On average, **White non-Hispanic recipients receiving substantially higher** amounts compared to Hispanic recipients.

Expenditures by Gender

Expenditures on White, non-Hispanic



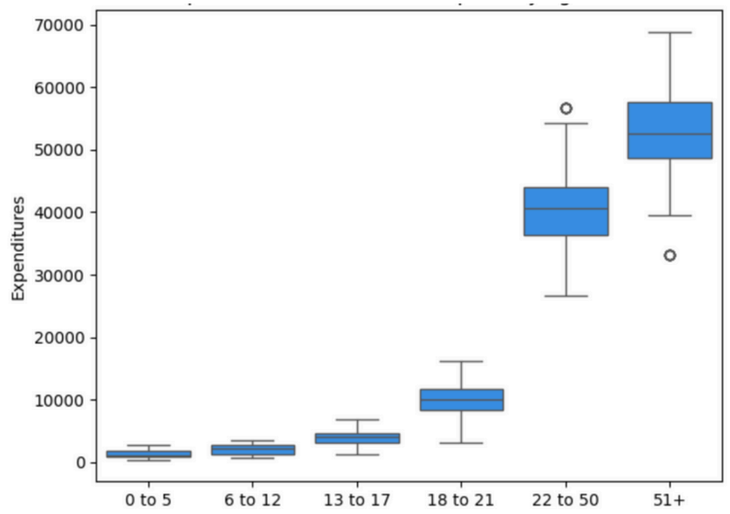
Expenditures on Hispanic



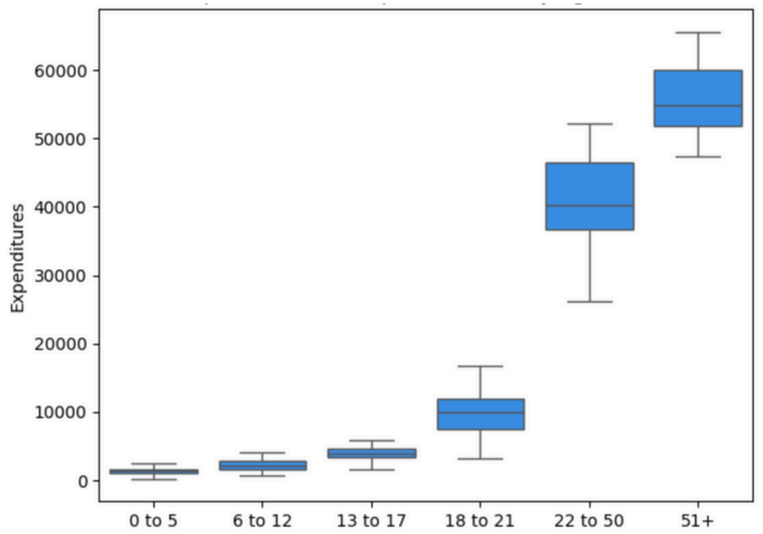
- The median expenditures for both White, non-Hispanic females and males are higher than Hispanic.
- White, non-Hispanic males tend to receive higher funds than the females
- Hispanic females and males tend to show less disparity
- Hispanic expenditures contain more outliers, suggesting variability among the highest spenders in the category
- White, non-Hispanic data appears to be more centrally distributed with fewer extreme values, suggesting more consistent approach to fund allocation among such demographic

Expenditures by Age Cohort

Expenditures on White, non-Hispanic



Expenditures on Hispanic



- Both groups show increasing median expenditures with age, peaking at the 51+ cohort.
- In both groups, the 18 to 21 and 22 to 50 cohorts show a marked increase in variability and median expenditures compared to younger age groups, which may reflect increased financial independence and responsibilities.
- The 51+ cohorts for both groups exhibit the highest variability and expenditures, which might be due to healthcare costs, retirement planning, or other age-related expenses.
- The White, non-Hispanic cohort shows outliers in the older age groups (18 to 21 and 51+), suggesting there are individuals in the White, non-Hispanic group with exceptionally high expenditures that are not as common in the Hispanic group.

T-test

T-Tests

- Comparing both whole populations
- Comparing by age and gender

In [47]:
```
# Query to retrieve age cohort, gender, expenditures, and ethnicity for "White not Hispanic" individuals
white_filter = {'Ethnicity': 'White not Hispanic'}
white_query_result = collection.find(white_filter, {'Age Cohort': 1, 'Gender': 1, 'Expenditures': 1, 'Ethnicity': 1,

# Convert the query results to a pandas DataFrame
white_df = pd.DataFrame(list(white_query_result))

# Query to retrieve age cohort, gender, expenditures, and ethnicity for "Hispanic" individuals
hispanic_filter = {'Ethnicity': 'Hispanic'}
hispanic_query_result = collection.find(hispanic_filter, {'Age Cohort': 1, 'Gender': 1, 'Expenditures': 1, 'Ethnicit

# Convert the query results to a pandas DataFrame
hispanic_df = pd.DataFrame(list(hispanic_query_result))
```

In [23]:
```
whiteExp = collection.find({'Ethnicity': 'White not Hispanic'}, {'Expenditures': 1, '_id': 0})

hispExp = collection.find({'Ethnicity': 'Hispanic'}, {'Expenditures': 1, '_id': 0})

whiteList = [doc['Expenditures'] for doc in whiteExp]
hispList = [doc['Expenditures'] for doc in hispExp]

whiteSeries = pd.Series(whiteList)
hispSeries = pd.Series(hispList)

from scipy import stats

tStat, pVal= stats.ttest_ind(whiteSeries, hispSeries, alternative='greater', equal_var=False)
if pVal < 0.05:
    print(f"White people received more money on average.")
    print(f'P-Value: {pVal}')
```

```
White people received more money on average.
P-Value: 3.525267812150838e-157
```

Conclusion

12 ×

1. Overall, **Hispanic individuals face discrimination.**

2. When gender is strictly controlled for, **both Hispanic males and females experience discrimination** compared to their white counterparts.

3. When age is strictly controlled for, none of the age groups face discrimination.

4. When both age and gender are controlled for, **discrimination is observed only within the 51+ Male category.**

5. Even after removing the 51+ age category, discrimination against the Hispanic population as a whole persists.