Disability Benefits Analysis

Author: Chris Chiang

Abstract and Introduction

In the United States, most states extend services and support to individuals (as well as their families) who are dealing with developmental disabilities, intellectual disabilities, autism, cerebral palsy, Alzheimer's, and many others. In California, the California Department of Developmental Services (DDS) handles these processes. They play a crucial role in distributing funds and aid over a quarter-million residents who have any of the aforementioned disabilities. A recent allegation of discrimination came to light, backed by analysts who took data from average annual expenditures on consumers that are categorized by race. The analysis claimed significant disparity within this: that the yearly expenditures on Latino the population were roughly 30% of that of the White population. This discovery triggered a deeper inquiry, which prompted state legislators and department leads to seek further statistical analysis. Within this report, I will look into the DDS data referenced in this article to assess whether there is enough evidence of ethnic, and even gender-based discrimination in the allocation of DDS funds.

Objectives and Background

I have 2 main goals in the project. My first aim to provide grouped summaries for categorical variables, visualization techniques for categorical variables and hypothesis generation based on EDA. The second part of the project is aimed at providing categorical variable analysis, model fitting and reporting of its parameters, and model-based visualizations. The population of interest that is being studied is developmentally-disabled residents in California. The type of sample being evaluated is a simple random sample based on population figures and data extracted from the state.

In [1]:

```
import numpy as np
import pandas as pd
import altair as alt
import sklearn.linear_model as lm
from sklearn.preprocessing import add_dummy_feature
```

Variable summaries

Name	Variable description	Type	Units of measurement
ID	ID	Numeric	ID
Age Cohort	Predefined Age Groups (0-5, 6-12, 13-17, 18-21, 22-50, and 51+)	Categorical	Category
Age	Age	Numeric	Years
Gender	Male/Female	Categorical	N/A
Expenditures	Annual Expenditures per Member	Numeric	USD
Ethnicity	Race	Categorical	None

Loading Example Data

In [2]:

```
dds = pd.read_csv('data/california-dds.csv')
dds.head(10)
```

Out[2]:

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

5	10690	13 to 17	15	Female	4566	Hispanic
6	10711	13 to 17	13	Female	3915	White not Hispanic
7	10778	13 to 17	17	Male	3873	Black
8	10820	13 to 17	14	Female	5021	White not Hispanic
9	10823	13 to 17	13	Male	2887	Hispanic

In [3]:

```
median_expend_by_eth = dds[['Ethnicity', 'Expenditures']].groupby('Ethnicity').median()
ethnicity_n = dds.Ethnicity.value_counts().rename('n')
table1 = pd.concat([median_expend_by_eth,ethnicity_n], axis=1, sort=True)
table1
```

Out[3]:

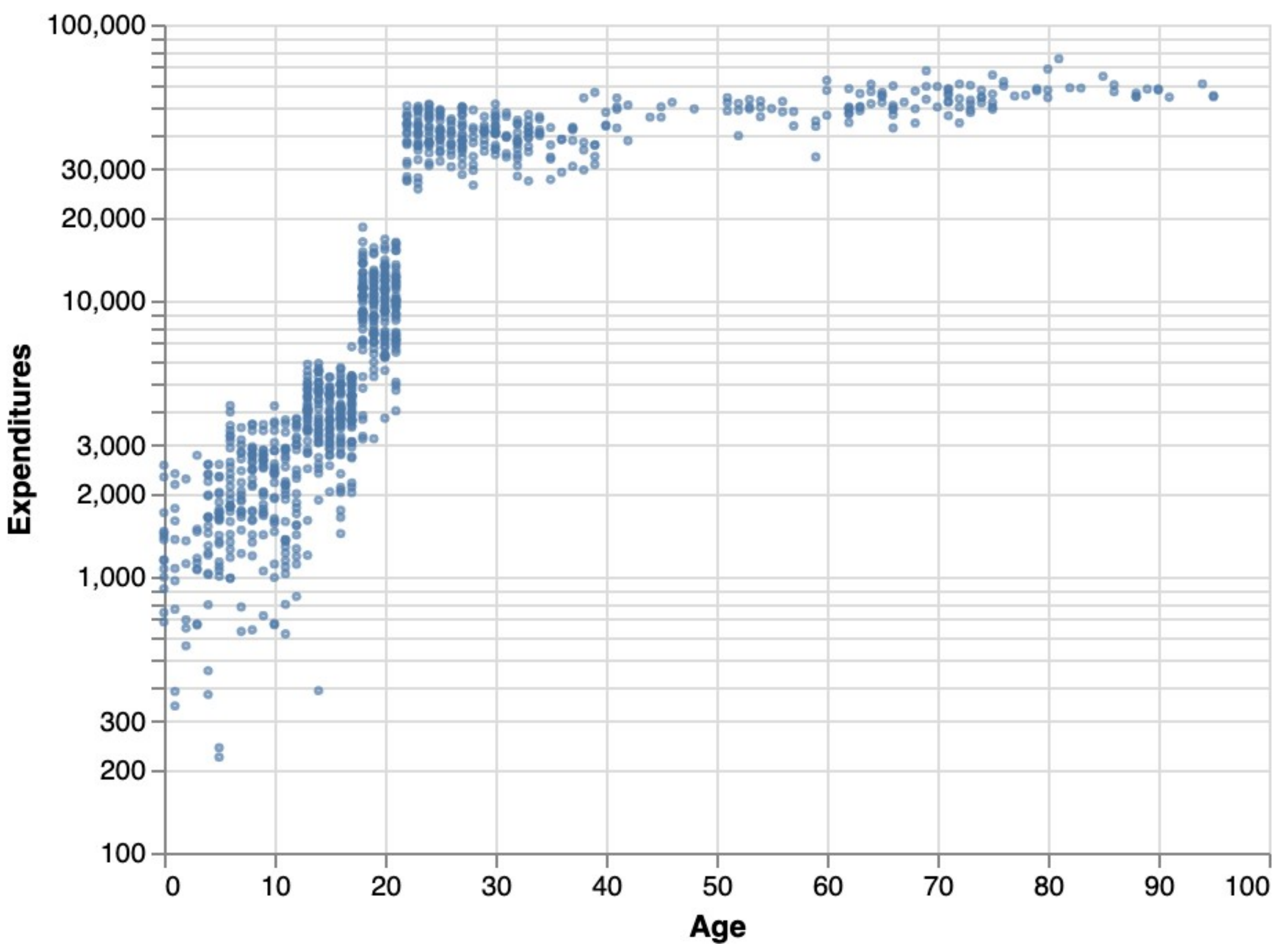
	Expenditures	Count
American Indian	41817.5	4
Asian	9369.0	129
Black	8687.0	59
Hispanic	3952.0	376
Multi Race	2622.0	26
Native Hawaiian	40727.0	3
Other	3316.5	2
White not Hispanic	15718.0	401

The main criteria used in this court case alleging discrimination is from ethnicity, so I did a calculation of the median expenditure for each group. From the table above, there are groups with significantly different median payments. Native Americans have a median funding of $41,817.5 while for mixed race people, it was $2,622. These two groups have a small sample size, and the results may be heavily skewed. The values for these races could be abnormally high or low because the limited sample misleads and skews the data significantly. For the races with a decent sample size, whites received more than asians, hispanics and blacks.

In [4]:

```
# solution
fig_1 = alt.Chart(dds).mark_point(size=2).encode(
        x =alt.X('Age'),
        y = alt.Y('Expenditures', scale = alt.Scale(type = 'log')))
fig_1
```

Out[4]:



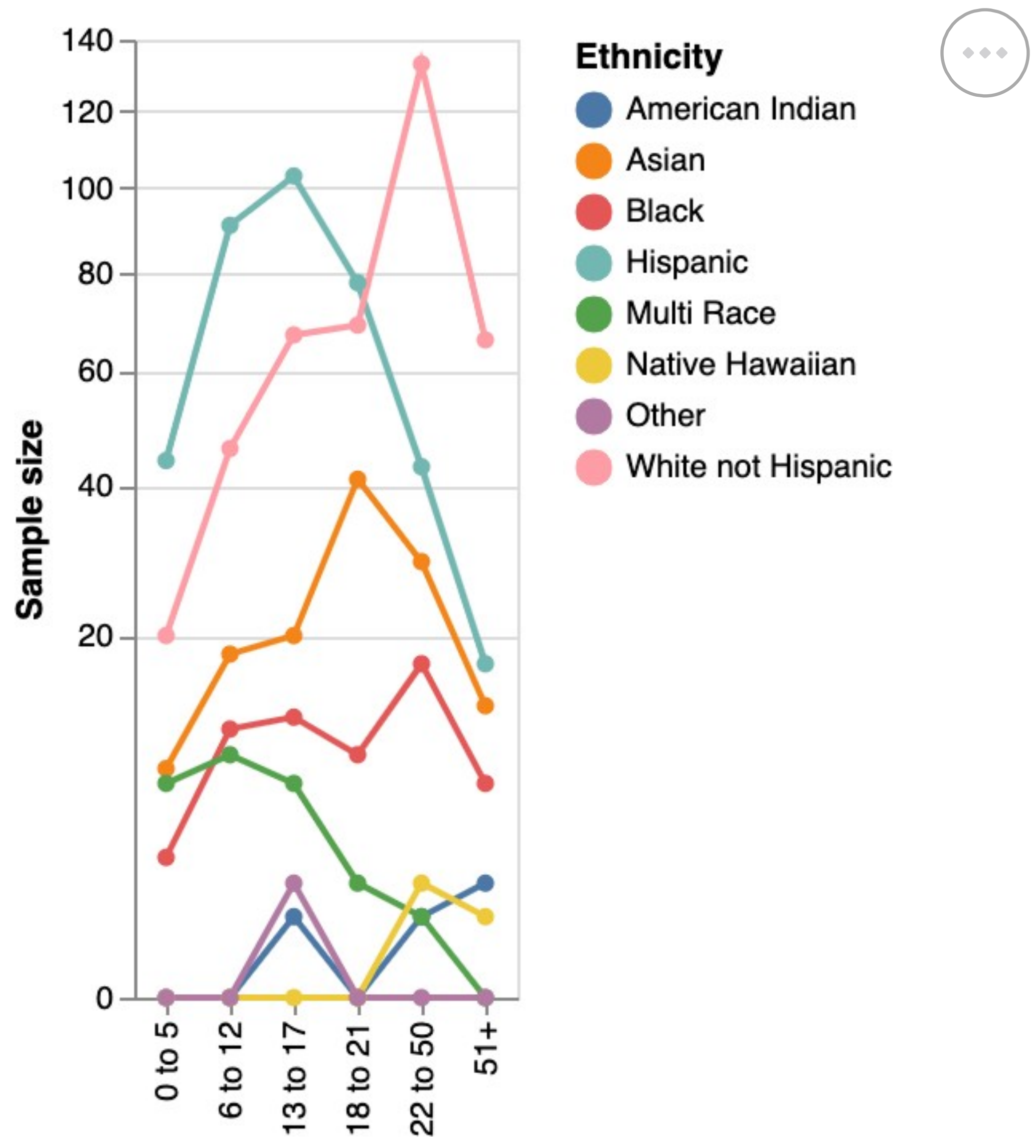
I plotted a scatterplot of age data vs expenditures. I used a log transformation because it is better depicts and captures all of the data visually. The data is not linear but there is a somewhat positive relationship between age and expenditures, especially for ages 0-20. After about the age of 30, the pattern stays consistent and doesn't change. Overall, it tends to increase with age. I deduce that after the age 20, a person's disability would require more care and medical attention (and thus cost) because symptoms and conditions tend to worsen with age. Additionally, before the age 20, most of this population are likely to be living at home, so the expenditure would be less, with additional people and resources that they may have access to. Moving forward, I will conduct analysis based on age cohorts instead of age, in order to be able to model the relationships categorically.

These are the cohorts I have chosen to break down age: 0-5: fewest needs and requires the least amount of attention. 6-12 and 13-17: many resources are provided by schools, thus lowering state expenditures directly. 18-21: typically the age people start moving out or living on their own. 22-50: typically does not live with parents but may still receive some care. 51+: has the most needs and require more funding because of age

In [5]:

```
samp_sizes= dds_cat.groupby(['Age Cohort','Ethnicity'])[['Id']].count()
samp_sizes['cohort_order'] = samp_sizes['Age Cohort']
fig_2 =
        alt.Chart(samp_sizes).mark_line(point=True).en
        code( x = alt.X('Age Cohort'),
        y = alt.Y('n', title = 'Sample size',
                  scale = alt.Scale(type =
        'sqrt')), color = 'Ethnicity')
fig_2
```

Out[5]:

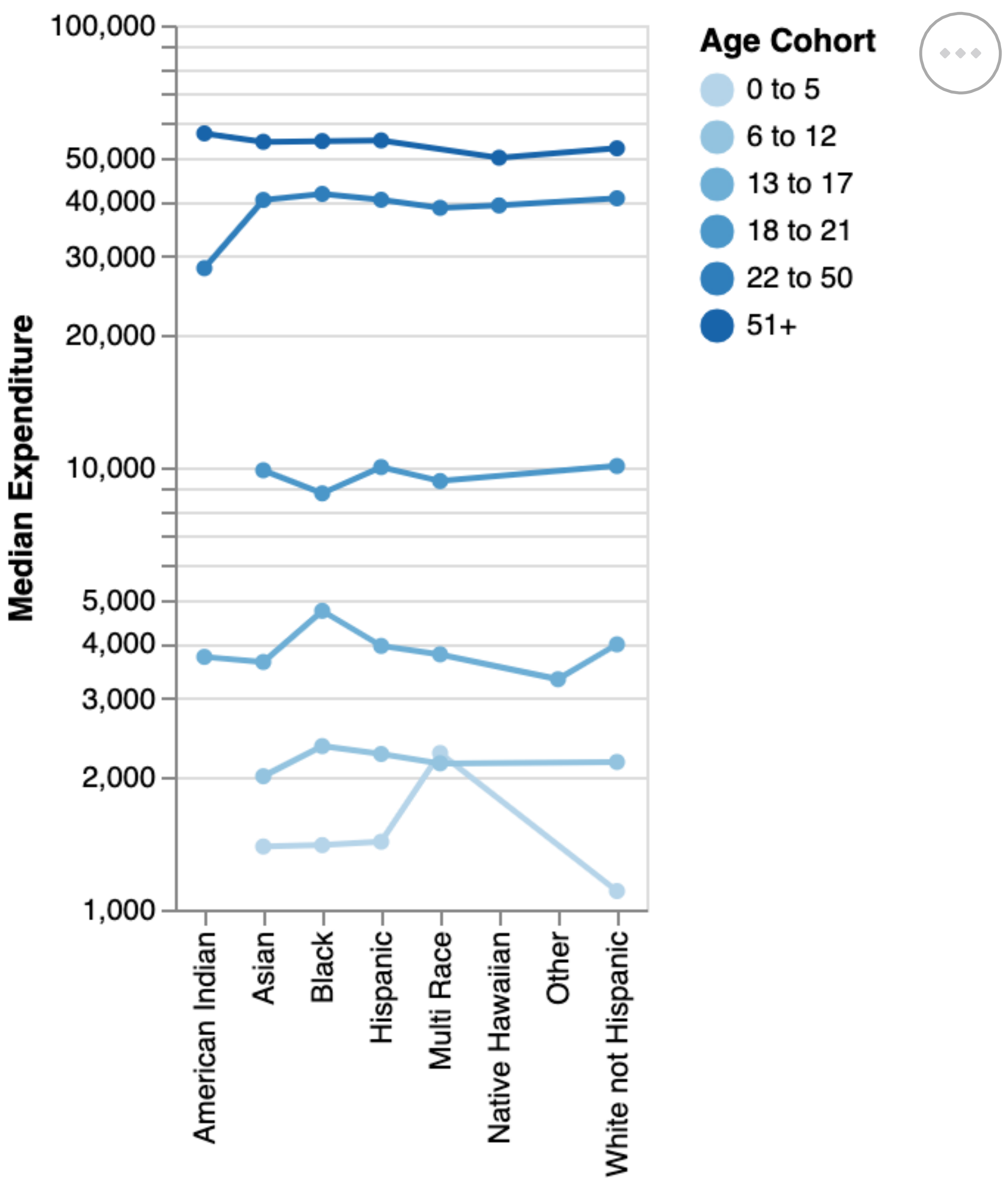


For most races (at least the ones with a decent size sample), the number of disabled people tends to drop with an increase in age but for the white population, it is the opposite effect: that number skyrockets. Most of the white population in this dataset who are disabled tend to be older (22-51+ years old). Most of the hispanic population in the sample are a much younger age (6-17 years old). Most of the multiracial people in the sample are under 5 years old and make up a much larger proportion of that age group than normal. This would make sense because most multiracial people are younger (generally speaking) and have low populations that are older. Many of the visually significant differences suggests that younger people tend to require less expenditures, and because the distributions for racial groups are not streamlined, it should be addressed.

In [6]:

```
dds_cat['cohort_order'] = dds_cat['Age Cohort'].cat.codes
fig_3 =
        alt.Chart(dds_cat).mark_line(point=True).encode(
        x = alt.X('Ethnicity',),
        y = alt.Y('median(Expenditures):Q', title = 'Median Expenditure',
                  scale = alt.Scale(type = 'log')),
        color = alt.Color('Age Cohort:O')
fig_3
```

Out[6]:



The last few prompts showed that the apparent discrimination could simply be an artifact of a difference of age structure and distributions based on a racial variable. I investigated this further by plotting median expenditure against ethnicity, as in the previous, but now also correcting for age cohort.

Regression Analysis

I am modeling the log of expenditures (response variable) as a function of gender, age cohort, and ethnicity:

$$\log\left(\text{expend}_i\right) = \beta_0 + \beta_1(\text{6-12})_i + + \beta_5(\text{51+})_i + \beta_6\text{male}_i + \beta_7\text{hispanic}_i + + \beta_{13}\text{other}_i + \epsilon_i$$

In this model, *all* of the explanatory variables are categorical and fitted using indicators, so linear model coefficients better capture means for each group. The response variable is log-transformed and all explanatory variables are categorical. For this model, I would like to add an error band to account for confidence figures, using a lower bound calculation of `expenditure` $- 2 \times$ `expenditure_s` and an upper bound calculation of `expenditure` $+ 2 \times$ `expenditure_s`.

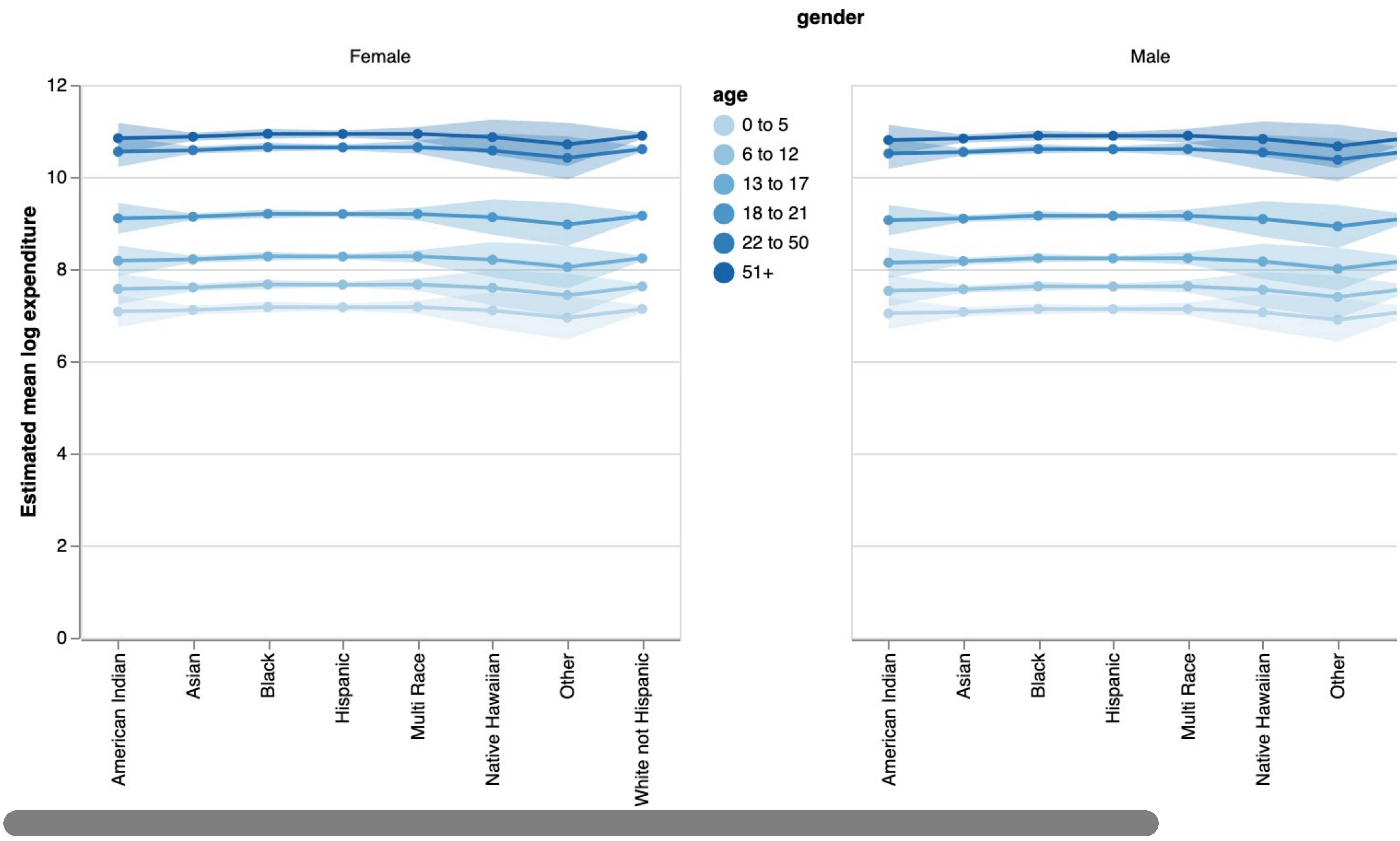
In [7]:

```
lines =alt.Chart(grid_df).mark_line(point=True).encode(
            x = alt.X('ethnicity',title=''),
            y = alt.Y('expenditure',title= 'Estimated mean log expenditure'),
            color = alt.Color('age:O',))

bands = lines.transform_calculate(
            lwr = 'datum.expenditure - 2*datum.expenditure_se',
            upr = 'datum.expenditure + 2*datum.expenditure_se'
    ).mark_errorband(opacity = 0.3).encode(
        x = alt.X('ethnicity',title=''),
        y = alt.Y('lwr:Q',title= 'Estimated mean log expenditure'),
        y2 = 'upr:Q',
        color = alt.Color('age:O', sort =
                        alt.EncodingSortField( field
                        ='cohort_order'), legend = None))

# layer and facet
fig_4 = alt.layer(lines,bands).resolve_scale().properties( width = 325, height
    = 300).facet(facet = 'gender:N')
fig_4
```

Out[7]:



Conclusion

The data is a random sample of Developmentally Disabled Residents in California with federally funded aid and is grouped by variables of ethnicity, age and gender. I initially saw that the amount of funding for hispanics, asians and blacks were significantly less than for Whites. However, a clear initial observation was that there is between expenditures and age, with higher age requiring more expenditures. I was able to visualize many relationships between the expenditures, ages, gender and ethnicities. I found out that race and gender had no effect on expenditures and the differences are simply because of a difference in distribution of age population of the racial groups. There is a positive correlated relationship between age and median expenditures and the data does NOT provide evidence of ethnic or gender discrimination in allocation of DDS funds.