PSTAT 100 Homework 4

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
from sklearn.linear_model import LinearRegression
```

Background: California Department of Developmental Services

From Taylor, S. A., & Mickel, A. E. (2014). Simpson's Paradox: A Data Set and Discrimination Case Study Exercise. Journal of Statistics Education, 22(1):

> 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. The agency through which the State of California serves the developmentally-disabled population is the California Department of Developmental Services (DDS) ... One of the responsibilities of DDS is to allocate funds that support over 250,000 developmentally-disabled residents. A number of years ago, an allegation of discrimination was made and supported by a univariate analysis that examined average annual expenditures on consumers by ethnicity. The analysis revealed that the average annual expenditures on Hispanic consumers was approximately one-third of the average expenditures on White non-Hispanic consumers. This finding was the catalyst for further investigation; subsequently, state legislators and department managers sought consulting services from a statistician.

In this assignment, you'll analyze the deidentified DDS data published with this article to answer the question: *is there evidence of ethnic or gender discrimination in allocation of DDS funds?*

Aside: The JSE article focuses on what's known as Simpson's paradox, an arithmetic phenomenon in which aggregate trends across multiple groups show the *opposite* of within-group trends. We won't emphasize this topic, though the data does provide a nice illustration -- if you're interested in learning more, you can follow the embedded link to the Wikipedia entry on the subject.

Assignment objectives

You'll answer the question of interest employing exploratory and regression analysis techniques from class. In particular, you'll practice the following skills.

Exploratory analysis:

- grouped summaries for categorical variables;
- visualization techniques for categorical variables;
- hypothesis generation based on EDA.

Regression analysis:

- categorical variable encodings;
- model fitting and fit reporting;
- parameter interpretation;
- model-based visualizations.

In addition, in **communicating results** at the end of the assignment, you'll practice a few soft skills that may be helpful in thinking about how to report results for your independent class project:

- composing a concise summary (similar to an abstract) of background and key findings; and
- determining which results (figures/tables) to reproduce in a presentation context.

---

0. Getting acquainted with the DDS data

The data for this assignment are already tidy, so in this section you'll just familiarize yourself with basic characteristics. The first few rows of the data are shown below:

In [2]:
```
dds = pd.read_csv('data/california-dds.csv')
dds.head()
```

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

Take a moment to open and read the data documentation (*data > california-dds-documentation.md*).

Question 0 (a). Sample characteristics

Answer the following questions based on the data documentation.

(i) Identify the observational units.

Answer: The observational units are each unique consumer IDs.

(ii) Identify the population of interest.

Answer: The population of interest is all individuals that receive DDS services.

(iii) What type of sample is this (*e.g.*, census, convenience, etc.)?

Answer: This sample is an administrative sample because the sample covers entire sampling frame, ethinicities recorded, but the frame does not cover the population, all ethnicities.

(iv) Is it possible to make inferences about the population based on this data?

Answer: Yes, because the sample is a random sample, we can make inferences about the popluation.

Question 0 (b). Variable summaries

Fill in the table below for each variable in the dataset.

Name	Variable description	Type	Units of measurement
ID	Unique consumer identifier	Numeric	None
Age Cohort	Age range that the unique consumer fits into	Categorical	Years
Age	Age of unique consumer	Numeric	Years
Gender	Gender of unique consumer	Categorical	None
Expenditures	Amount of money spent on support and services for the unique consumer	Numeric	Dollar
Ethnicity	Ethnicity of unique consumer	Categorical	None

---

1. Exploratory analysis

Question 1 (a). Alleged discrimination

These data were used in a court case alleging discrimination in funding allocation by ethnicity. The basis for this claim was a calculation of the median expenditure for each group. Here you'll replicate this finding.

(i) Median expenditures by ethnicity

Construct a table of median expenditures by ethnicity.

1. Slice the ethnicity and expenditure variables from `dds`, group by ethnicity, and calculate the median expenditure. Store the result as `median_expend_by_eth`.
2. Compute the sample sizes for each ethnicity using `.value_counts()`: obtain a Series object indexed by ethnicity with a single column named `n`. You'll need to use `.rename(...)` to avoid having the column named `Ethnicity`. Store this result as `ethnicity_n`.
3. Use `pd.concat(...)` to append the sample sizes in `ethnicity_n` to the median expenditures in `median_expend_by_eth`. Store the result as `tbl_1`.

Print `tbl_1`.

In [3]:
```
# compute median expenditures
median_expend_by_eth = dds.loc[:,['Ethnicity', 'Expenditures']].groupby('Ethnicity').median()
median_expend_by_eth

# compute sample sizes
ethnicity_n = pd.Series(dds['Ethnicity']).value_counts().rename('n')
print(ethnicity_n)

# concatenate
tbl_1 = pd.concat([median_expend_by_eth, ethnicity_n], axis = 1, sort = True)

# print
tbl_1
```

```
White not Hispanic    401
Hispanic              376
Asian                 129
Black                  59
Multi Race             26
American Indian         4
Native Hawaiian         3
Other                   2
Name: n, dtype: int64
```

Out[3]:

	Expenditures	n
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

(ii) Do there appear to be significant differences in funding allocation by ethnicity?

If so, give an example of two groups receiving significantly different median payments.

Answer

Yes, there appears to be significant differences in funding allocation by ethnicity. The two groups with significanty different median payments are American Indian , $41,817.5$, and Multi Race, $2,622$.

(iii) Which groups have small sample sizes? How could this affect the median expenditure in those groups?

Answer

The groups with small sample sizes are American Indian, Native Hawaiian, and Other ethnicities. These could affect the median expenditure in these groups by having insufficient data, making the median expenditure numbers to be inaccurate. Outliers in the data could have a larger impact since the size is small.

(iv) Display `tbl_1` visually.

Construct a point-and-line plot of median expenditure (y) against ethnicity (x), with:

- ethnicities sorted by descending median expenditure;
- the median expenditure axis shown on the log scale;
- the y-axis labeled 'Median expenditure'; and
- no x-axis label (since the ethnicity group names are used to label the axis ticks, the label 'Ethnicity' is redundant).

Store the result as `fig_1` and display the plot.

Hints:

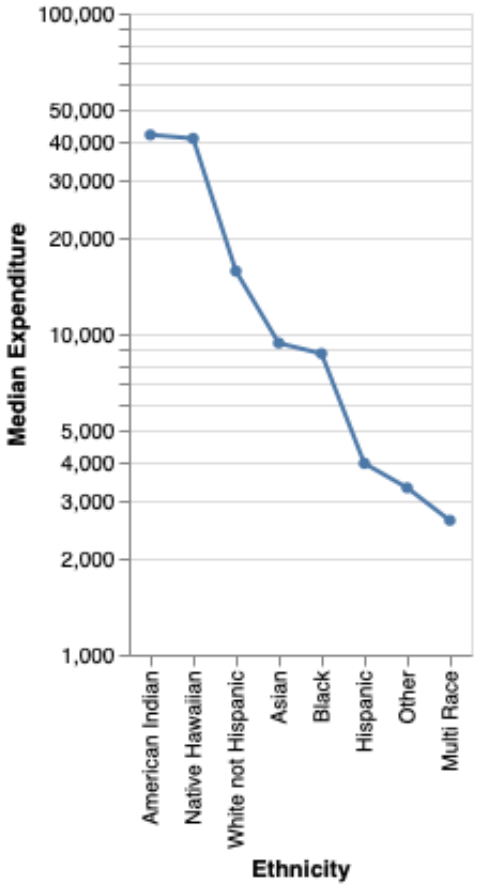
- you'll need to use `tbl_1.reset_index()` to obtain the ethnicity group as a variable;
- recall that `.mark_line(point = True)` will add points to a line plot;
- sorting can be done using `alt.X(..., sort = alt.EncodingSortField(field = ..., order = ...))`

In [4]:
```
# solution
fig_1 = alt.Chart(tbl_1.reset_index()).mark_line(point=True).encode(
    x = alt.X('index',sort = alt.EncodingSortField(field = 'Expenditures', order = 'descending'),
              title = 'Ethnicity'),
    y = alt.Y('Expenditures', title = 'Median Expenditure',scale = alt.Scale(type = 'log'))

)

fig_1
```

Out[4]:



Question 1 (b). Age and expenditure

Here you'll explore how expenditure differs by age.

(i) Construct a scatterplot of expenditure (y) versus age (x).

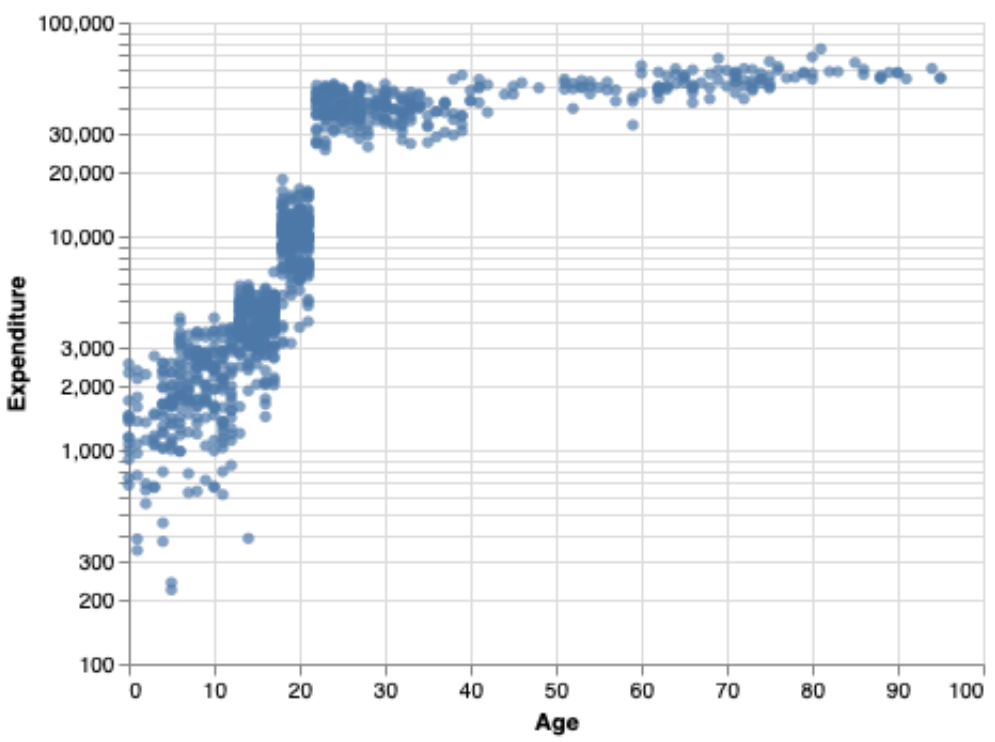
Use the quantitative age variable (not age cohort). Display expenditure on the y axis on the log scale, and age on the x axis on the usual (linear) scale.

Store the plot as `fig_2` and display the graphic.

In [5]:
```
# solution
fig_2 = alt.Chart(dds).mark_circle().encode(
    x = alt.X('Age', title = 'Age'),
    y = alt.Y('Expenditures', title = 'Expenditure',scale = alt.Scale(type = 'log'))

)
fig_2
```

Out[5]:



(ii) Does the relationship seem linear?

If so, describe the direction (positive/negative) and approximate strength (steep/slight) of relationship. If not, describe the pattern of relationship, if any, in 1-2 sentences.

Answer

This relationship could possibly be considered linear even though there seems to be some discontinuity, but I think it also could be non-linear beacuse it resembles a log function, which has some curvature to it. If we were to say it leans more towards the linear side, we could say that there is a positive direction and has a slightly steep relationship.

(iii) Overall, how does expenditure tend to change as age increases?

Answer

As age increases, expenditure for support and services increase.

(iv) What might explain the sudden increase in expenditure after age 20?

Answer

As individuals with developmental disabilities hit age 20, there could possibly be a need for more expenditure due to moving out of their guardians' home and getting a job.

Precisely because recipients have different needs at different ages that translate to jumps in expenditure, age has been discretized into age cohorts defined based on need level. Going forward, we'll work with these age cohorts -- by treating age as discrete, we won't need to attempt to model the discontinuities in the relationship between age and expenditure.

The cohort labels are stored as `Age Cohort` in the dataset. There are six cohorts; the cell below coerces the labels to an ordered category and prints the category levels.

In [6]:
```
# convert data types
dds_cat = dds.astype({'Age Cohort': 'category', 'Ethnicity': 'category', 'Gender': 'category'}).copy()

dds_cat['Age Cohort'] = dds_cat['Age Cohort'].cat.as_ordered().cat.reorder_categories(
    dds_cat['Age Cohort'].cat.categories[[0, 5, 1, 2, 3, 4]]
)

# age cohorts
dds_cat['Age Cohort'].cat.categories
```

Out[6]:
```
Index(['0 to 5', '6 to 12', '13 to 17', '18 to 21', '22 to 50', '51+'], dtype='object')
```

Here is an explanation of how the cohort age boundaries were chosen:

> The 0-5 cohort (preschool age) has the fewest needs and requires the least amount of funding. For the 6-12 cohort (elementary school age) and 13-17 (high school age), a number of needed services are provided by schools. The 18-21 cohort is typically in a transition phase as the consumers begin moving out from their parents' homes into community centers or living on their own. The majority of those in the 22-50 cohort no longer live with their parents but may still receive some support from their family. Those in the 51+ cohort have the most needs and require the most amount of funding because they are living on their own or in community centers and often have no living parents.

Question 1 (c). Age and ethnicity

Here you'll explore the age structure of each ethnic group in the sample.

(i) Group the data by ethnic group and tabulate the sample sizes for each group.

Use `dds_cat` so that the order of age cohorts is preserved. Write a chain that does the following.

1. Group by age cohort and ethnicity.
2. Slice the `Id` variable, which is unique to recipient in the sample.
3. Count the number of recipients in each group using `.count()`.
4. Reset the index so that age cohort and ethnicity are dataframe columns.
5. Rename the column of ID counts 'n'.

Store the result as `samp_sizes` and print the first four rows.

In [7]:
```
# solution
samp_sizes = dds_cat.groupby(['Age Cohort', 'Ethnicity']).Id.count().reset_index().rename(columns = {'Id' : 'n'})

# print
samp_sizes.head(4)
```

Out[7]:

	Age Cohort	Ethnicity	n
0	0 to 5	Asian	8
1	0 to 5	Black	3
2	0 to 5	Hispanic	44
3	0 to 5	Multi Race	7

(ii) Visualize the age structure of each ethnic group in the sample.

Construct a point-and-line plot of the sample size against age cohort by ethnicity.

1. To preserve the ordering of age cohorts, create a new column in `samp_sizes` called `cohort_order` that contains an integer encoding of the cohort labels in order. To obtain the integer encoding, slice the age cohort variable as a series and use `series.cat.codes`.
2. Construct an Altair chart based on `samp_sizes` with:
   - sample size (`n`) on the y axis;
   - the y axis titled 'Sample size' and displayed on a square root scale;
   - age cohort on the x axis, ordered by the cohort variable you created;
   - the x axis unlabeled; and
   - ethnic group mapped to color.

Store the plot as `fig_3` and display the graphic.

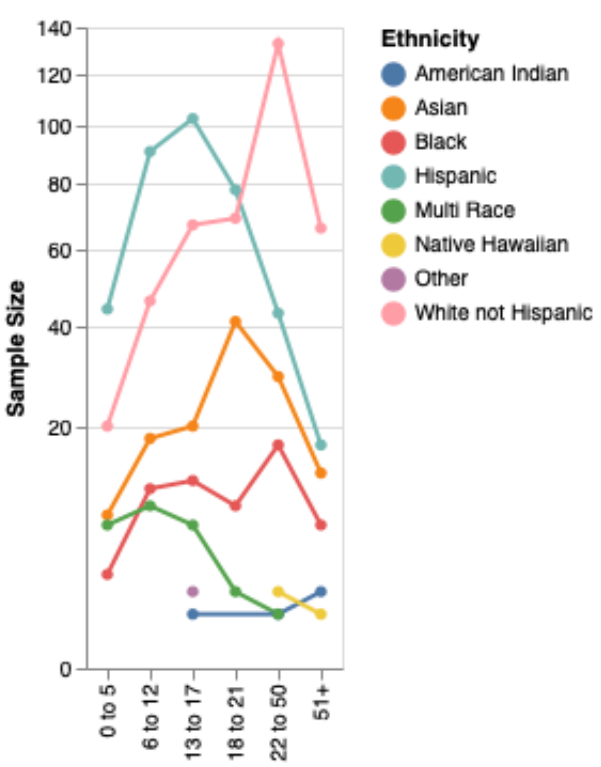
(*Hint*: sorting can be done using `alt.X(..., sort = alt.EncodingSortField(field = ..., order = ...))`.)

In [8]:
```
# add column with category codes
samp_sizes['cohort_order'] = dds_cat['Age Cohort'].cat.codes

# construct plot
fig_3 = alt.Chart(samp_sizes).mark_line(point=True).encode(
    x = alt.X('Age Cohort', sort = alt.EncodingSortField(field = 'cohort_order:O', order = 'descending'),
              title = '', scale = alt.Scale(zero=False)),
    y = alt.Y('n', title = 'Sample Size', scale = alt.Scale(type = 'sqrt')),
    color = alt.Color('Ethnicity')
)

# display
fig_3
```

Out[8]:



(iii) Are there differences in age structure?

If so, identify one specific example of two ethnic groups with different age structures and describe how the age structures differ.

Answer

Two ethnic groups with different age structures are White not Hispanic and American Indian. For example, we can see that 140 of the people in the White not Hispanic group are 22-50 years old while only about 5 people in the American Indian group are 22-50. Moreover, it seems that overall the American Indian group has the lowest number of people in each age range, making it very different from the other ethnic groups.

Question 1 (d). Correcting for age

Here you'll consider how the age structure among ethnic groups might be related to the observed differences in median expenditure.

(i) Combine your figures 1, 2, and 3.

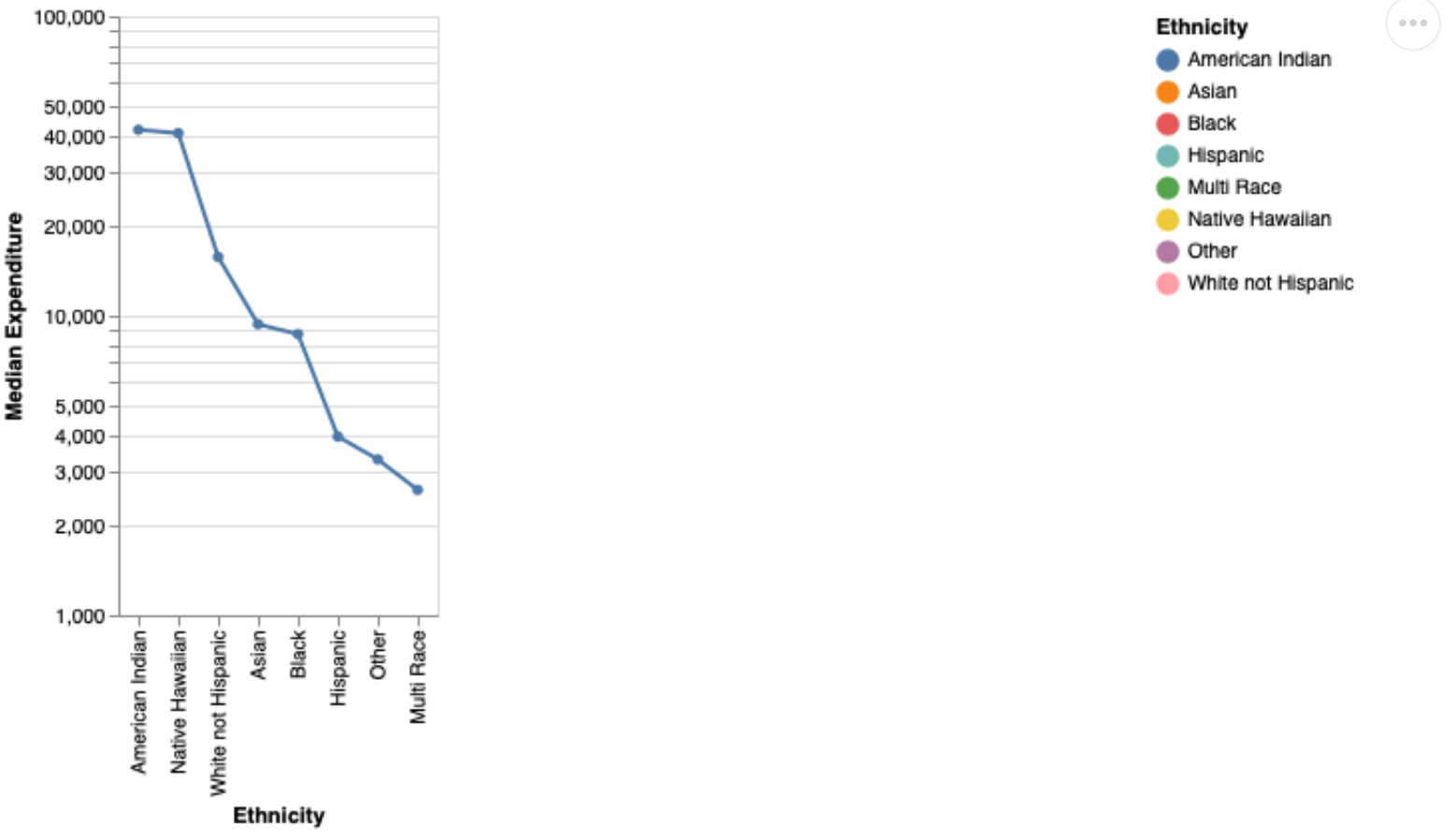
Place your figures together in three panels on a single row. You'll need to adjust the width of figure 2 (expenditure vs. age) so that the panels don't get cut off in rendering your notebook.

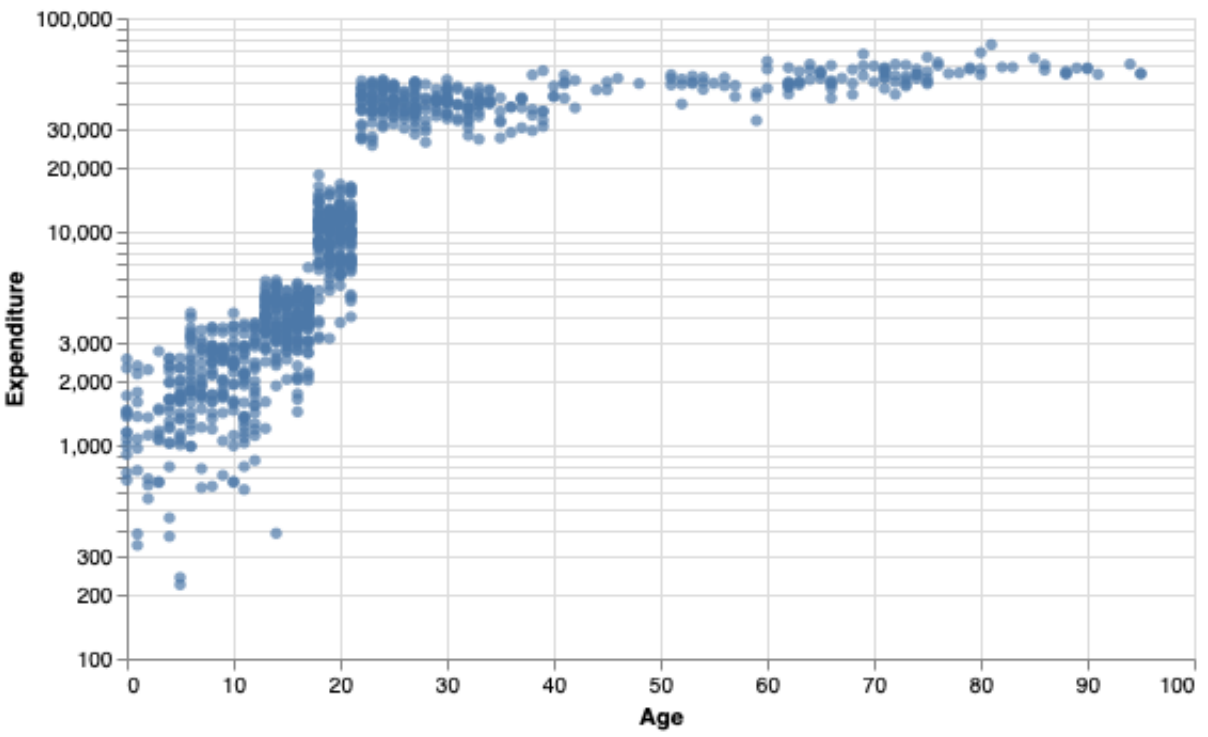
Store the figure panel as `eda_figures` and display the graphic.

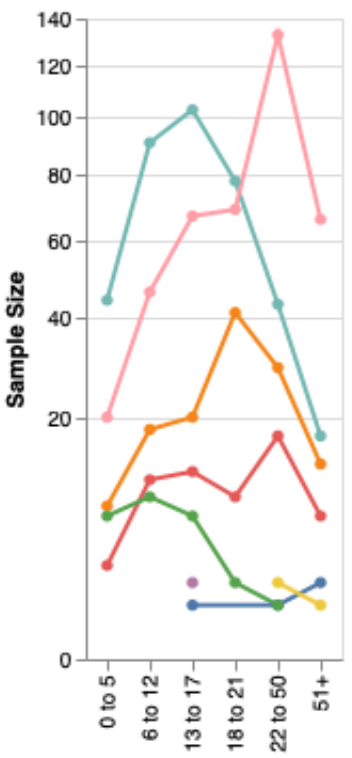
In [9]:
```
# solution
eda_figures = fig_1 & fig_2.properties(width=500,height=300) & fig_3

eda_figures
```

Out[9]:







(ii) Why is the median expenditure for the multiracial group so low?

Look at the age distribution for `Multi Race` and consider the age-expenditure relationship. Can you explain why the median expenditure for this group might be lower than the others? Answer in 1-2 sentences.

Answer

It seems that age distribution for Multi Race is around 13-17. Moreover, the people in this range seem to have an expenditure amount of 2000-3000. I think the median expenditure for this group is lower than the others because people this age are in high school, so a number of needed services are provided by schools, therefore they have less median expenditure.

(iii) Why is the median expenditure for the American Indian group so high?

Print the rows of `dds_cat` for this group (there aren't very many) and answer the question based on inspecting the rows.

In [10]:
```
# solution
dds_cat[dds_cat.Ethnicity == 'American Indian']
```

Out[10]:

	Id	Age Cohort	Age	Gender	Expenditures	Ethnicity
231	30234	51+	78	Female	55430	American Indian
575	61498	13 to 17	13	Female	3726	American Indian
730	74721	51+	90	Female	58392	American Indian
788	79645	22 to 50	32	Male	28205	American Indian

Answer

The median expenditure for the American Indian group is so high becuase most people in this group are 51+. Moreover, those in the 51+ cohort have the most needs and require the most amount of funding because they are living on their own or in community centers and often have no living parents, making their median expenditure much higher.

(iv) Plot expenditure against ethnicity by age.

Hopefully, the last few prompts convinced you that the apparent discrimination *could* simply be an artefact of differing age structure. You can investigate this by plotting median expenditure against ethnicity, as in figure 1, but now also correcting for age cohort.

1. To preserve the ordering of age cohorts, create a new column in `dds_cat` called `cohort_order` that contains an integer encoding of the cohort labels in order. To obtain the integer encoding, slice the age cohort variable as a series and use `series.cat.codes`.
2. Construct an Altair point-and-line chart based on `dds_cat` with:
   - ethnicity on the x axis;
   - no x axis label;
   - median expenditure on the y axis (*hint*: altair can parse `median(variablename)` within an axis specification);
   - the y axis displayed on the log scale;
   - age cohort mapped to color as an ordinal variable (meaning, use `:O` in the variable specification) and sorted in order of the `cohort_order` variable you created; and
   - lines connecting points that display the median expenditure for each ethnicity and cohort, with one line per age cohort.

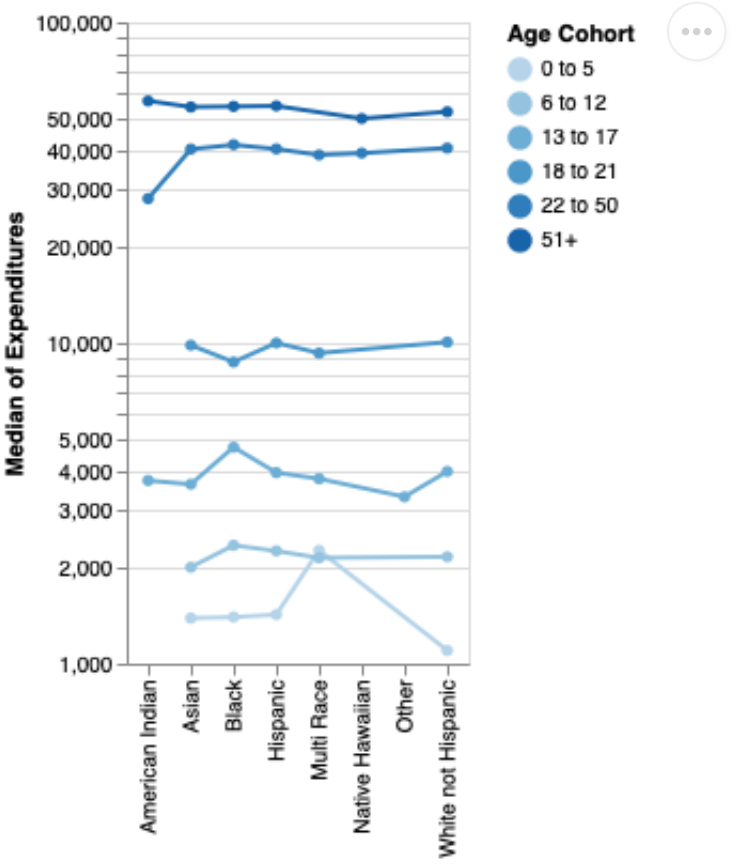
Store the result as `fig_4` and display the graphic.

In [11]:
```
# add column with category codes
dds_cat['cohort_order'] = dds_cat['Age Cohort'].cat.codes

# construct plot
fig_4 = alt.Chart(dds_cat).mark_line(point=True).encode(
    x = alt.X('Ethnicity', title = '', scale = alt.Scale(zero=False)),
    y = alt.Y('Expenditures', aggregate = 'median', scale = alt.Scale(type = 'log')),
    color = alt.Color('Age Cohort:O', sort = alt.EncodingSortField(field = 'cohort_order'))
)

# display
fig_4
```

Out[11]:



(v) Do the data reflect a difference in median expenditure by ethnicity after accounting for age?

Answer based on figure 4 in 1-2 sentences.

Answer

I think that overall, the data does not reflect a large difference in median expenditure by ethnicity after accounting for age. We can see that younger pople tend to have less median expenditure overall, while older people tend to have higher median expenditure overall, which follows the results we have seen thus far, but ethnicity does not seem to be such a large factor.

2. Regression analysis

Now that you've thoroughly explored the data, you'll use a linear model in this part to estimate the differences in median expenditure that you observed graphically in part 1.

More specifically, you'll model the log of expenditures (response variable) as a function of gender, age cohort, and ethnicity:

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

In this model, *all* of the explanatory variables are categorical and encoded using indicators; in this case, the linear model coefficients capture means for each group.

Because this model is a little different than the examples you've seen so far in two respects -- the response variable is log-transformed and all explanatory variables are categorical -- some comments are provided below on these features. You can review or skip the comments, depending on your level of interest in understanding the model better mathematically.

Commments about parameter interpretation

In particular, each coefficient represents a difference in means from the 'baseline' group. All indicators are zero for a white male recipient between ages 0 and 5, so this is the baseline group and:

$$\mathbb{E}\left(\log(\text{expend}) \mid \text{male, white, 0-5}\right) = \beta_1$$

Then, the expected log expenditure for a hispanic male recipient between ages 0 and 5 is:

$$\mathbb{E}\left(\log(\text{expend}) \mid \text{male, hispanic, 0-5}\right) = \beta_0 + \beta_7$$

So β_7 is *the difference in mean log expenditure between hispanic and white recipients after accounting for gender and age*. The other parameters have similar interpretations.

While the calculation shown above may seem a little foreign, you should know that the parameters represent marginal differences in means between genders (holding age and ethnicity fixed), between ages (holding gender and ethnicity fixed), and between ethnicities (holding age and gender fixed).

Comments about the log transformation

The response in this model is the *log* of expenditures (this gives a better model for a variety of reasons). The statistical assumption then becomes that:

$$\log(\text{expend})_i \sim N\left(\mathbf{x}_i'\beta, \sigma^2\right)$$

If the log of a random variable Y is normal, then Y is known as a *lognormal* random variable; it can be shown mathematically that the exponentiated mean of $\log Y$ is the median of Y. As a consequence, according to our model:

$$\text{median}(\text{expend}_i) = \exp\{\mathbf{x}_i'\beta\}$$

You'll work on the log scale throughout to avoid complicating matters, but know that this model for the log of expenditures is *equivalently* a model of the median expenditures.

Reordering categories

The cell below reorders the category levels to match the model written above. To ensure the parameters appear in the proper order, this reordering is done for you.

In [12]:
```
# remove ID and quantitative age
reg_data = dds_cat.copy().drop(columns = ['Id', 'Age'])

# reorder ethnicity
reg_data['Ethnicity'] = reg_data.Ethnicity.cat.as_ordered().cat.reorder_categories(
    reg_data.Ethnicity.cat.categories[[7, 3, 2, 1, 5, 0, 4, 6]]
)

# reorder gender
reg_data['Gender'] = reg_data.Gender.cat.as_ordered().cat.reorder_categories(['Male', 'Female'])
```

Question 2 (a). Data preprocessing

Here you'll extract the quantities -- explanatory variable matrix and response vector -- needed to fit the linear model.

(i) Categorical variable encoding.

Use `pd.get_dummies(...)` to encode the variables in `reg_data` as indicators. Be sure to set `drop_first = True`. Store the encoded categorical variables as `x_df` and print the first three rows and six columns. (There should be 13 columns in total.)

(*Hint*: `reg_data` can be passed directly to `get_dummies(...)`, and quantitative variables will be unaffected; a quick way to find `x_df` is to pass `reg_data` to this function and then drop the quantitative variables.)

In [13]:
```
# solution
x_df = pd.get_dummies(reg_data, columns = ['Age Cohort', 'Gender', 'Ethnicity'],
                      drop_first = True).drop(columns = ['Expenditures', 'cohort_order'])

x_df.iloc[0:3, 0:6]
```

Out[13]:

	Age Cohort_6 to 12	Age Cohort_13 to 17	Age Cohort_18 to 21	Age Cohort_22 to 50	Age Cohort_51+	Gender_Female
0	0	1	0	0	0	1
1	0	0	0	1	0	0
2	0	0	0	0	0	0

(ii) Add intercept.

Add an intercept column -- a column of ones -- to `x_df` using `add_dummy_feature(...)`. Store the result (an array) as `x_mx` and print the first three rows and six columns.

In [14]:
```
# solution
x_mx = add_dummy_feature(x_df, value = 1)
print(x_mx[0:3, 0:6])
```

```
[[1. 0. 1. 0. 0. 0.]
 [1. 0. 0. 0. 1. 0.]
 [1. 0. 0. 0. 0. 0.]]
```

(iii) Response variable.

Log-transform the expenditures column of `reg_data` and store the result in array format as `y`. Print the first ten entries of `y`.

In [15]:
```
# solution
y = np.array(np.log(reg_data['Expenditures']))
print(y[0:10])
```

```
[ 7.65586402 10.64361373  7.28207366  8.76405327  8.39208338  8.42639283
  8.27257061  8.26178468  8.5213844   7.96797318]
```

Question 2 (b). Model fitting

In this part you'll fit the linear model and summarize the results. You may find it helpful to have lab 6 open as an example to follow througout.

(i) Compute the estimates.

Configure a linear regression module and store the result as `mlr`; fit the model to `x_mx` and `y`. Be sure **not** to fit an intercept separately, since there's already an intercept column in `x_mx`.

(You do not need to show any output for this part.)

In [16]:
```
# solution

# configure module
mlr = LinearRegression(fit_intercept = False)

# fit model
mlr.fit(x_mx, y)
```

Out[16]:
```
LinearRegression(copy_X=True, fit_intercept=False, n_jobs=None, normalize=False)
```

(ii) Parameter estimate table.

Construct a table of the estimates and standard errors for each coefficient, and the estimate for the error variance parameter. The table should have two columns, 'estimate' and 'standard error', and rows should be indexed by parameter name. Follow the steps below.

1. Store the dimensions of `x_mx` as `n` and `p`.
2. Compute $(\mathbf{X}'\mathbf{X})$; store the result as `xtx`.
3. Compute $(\mathbf{X}'\mathbf{X})^{-1}$; store the result as `xtx_inv`.
4. Compute the residuals (as an array); store the result as `resid`.
   - (You can compute the fitted values as a separate step, or not, depending on your preference.)
5. Compute the error variance estimate, $\text{var}(\text{resids}) \times \frac{n-1}{n-p}$; store the result as `sigmasqhat`.
6. Compute the variance-covariance matrix of the coefficient estimates $\hat{\mathbf{V}} = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$; store the result as `v_hat`.
7. Compute the coefficient standard errors, $\sqrt{\hat{v}_{ii}}$; store the result (an array) as `coef_se`.
   - Append an `NaN` (`float('nan')`) to the array (for the error variance estimate).
8. Create an array of coefficient labels by appending 'intercept' to the column names of `x_df`, followed by 'error_variance'; store the result as `coef_labels`.
9. Create an array of estimates by appending the fitted coefficients with `sigmasqhat`; store the result as `coef_estimates`.
10. Create a dataframe with `coef_estimates` as one column, `coef_se` as another column, and indexed by `coef_labels`. Store the result as `coef_table`.

Print `coef_table`.

In [17]:

```
# store dimensions
n, p = x_mx.shape

# compute x'x
xtx = x_mx.transpose().dot(x_mx)

# compute x'x inverse
xtx_inv = np.linalg.inv(xtx)

# compute residuals
fitted_mlr = mlr.predict(x_mx)
resid = y - fitted_mlr

# compute error variance estimate
sigmasqhat = ((n - 1)/(n - p)) * resid.var()

# compute variance-covariance matrix
v_hat = xtx_inv * sigmasqhat

# compute standard errors
coef_se = np.sqrt(v_hat.diagonal())
coef_se = np.append(coef_se, float('nan'))

# coefficient labels
x_df.insert(loc=0, column='intercept', value=0)
x_df['error variance'] = 0
coef_labels = x_df.columns

# estimates
coef_estimates = np.append(mlr.coef_, sigmasqhat)

# summary table
coef_table = pd.DataFrame(
    data = {'coefficient estimate': coef_estimates, 'standard error': coef_se},
    index = coef_labels
)

# print
coef_table
```

Out[17]:

	coefficient estimate	standard error
intercept	7.092439	0.041640
Age Cohort_6 to 12	0.490276	0.043833
Age Cohort_13 to 17	1.101010	0.042761
Age Cohort_18 to 21	2.023844	0.043435
Age Cohort_22 to 50	3.470836	0.043500
Age Cohort_51+	3.762393	0.049536
Gender_Female	0.039784	0.020739
Ethnicity_Hispanic	0.038594	0.024881
Ethnicity_Black	0.041713	0.045702
Ethnicity_Asian	-0.021103	0.033454
Ethnicity_Native Hawaiian	-0.030725	0.189872
Ethnicity_American Indian	-0.054396	0.164828
Ethnicity_Multi Race	0.041024	0.067646
Ethnicity_Other	-0.189877	0.232793
error variance	0.106898	NaN

Now look at both the estimates and standard errors for each level of each categorical variable; if some estimates are large for at least one level and the standard errors aren't too big, then estimated mean log expenditures differ according to the value of that variable when the other variables are held constant.

For example: the estimate for `Gender_Female` is 0.04; that means that, if age and ethnicity are held fixed, the estimated difference in mean log expenditure between female and male recipients is 0.04. If $\log(a) - \log(b) = 0.04$, then $\frac{a}{b} = e^{0.04} \approx 1.041$; so the estimated expenditures (not on the log scale) differ by a factor of about 1. Further, the standard error is 0.02, so the estimate is within 2SE of 0; the difference could well be zero. So the model suggests there is no difference in expenditure by gender.

(iii) Do the parameter estimates suggest differences in expenditure by age or ethnicity?

First consider the estimates and standard errors for each level of age, and state whether any differences in mean log expenditure between levels appear significant; if so, cite one example. Then do the same for the levels of ethnicity. Answer in 2-4 sentences.

(*Hint*: it may be helpful scratch work to exponentiate the coefficient estimates and consider whether they differ by much from 1.)

In [18]:
```
# exponentiate age (not required)
np.exp(coef_table.iloc[1:6, 0])
```

Out[18]:
```
Age Cohort_6 to 12      1.632767
Age Cohort_13 to 17     3.007203
Age Cohort_18 to 21     7.567356
Age Cohort_22 to 50    32.163632
Age Cohort_51+         43.051330
Name: coefficient estimate, dtype: float64
```

In [19]:
```
# exponentiate ethnicity (not requried)
np.exp(coef_table.iloc[7:14, 0])
```

Out[19]:
```
Ethnicity_Hispanic           1.039348
Ethnicity_Black              1.042595
Ethnicity_Asian              0.979118
Ethnicity_Native Hawaiian    0.969742
Ethnicity_American Indian    0.947057
Ethnicity_Multi Race         1.041877
Ethnicity_Other              0.827061
Name: coefficient estimate, dtype: float64
```

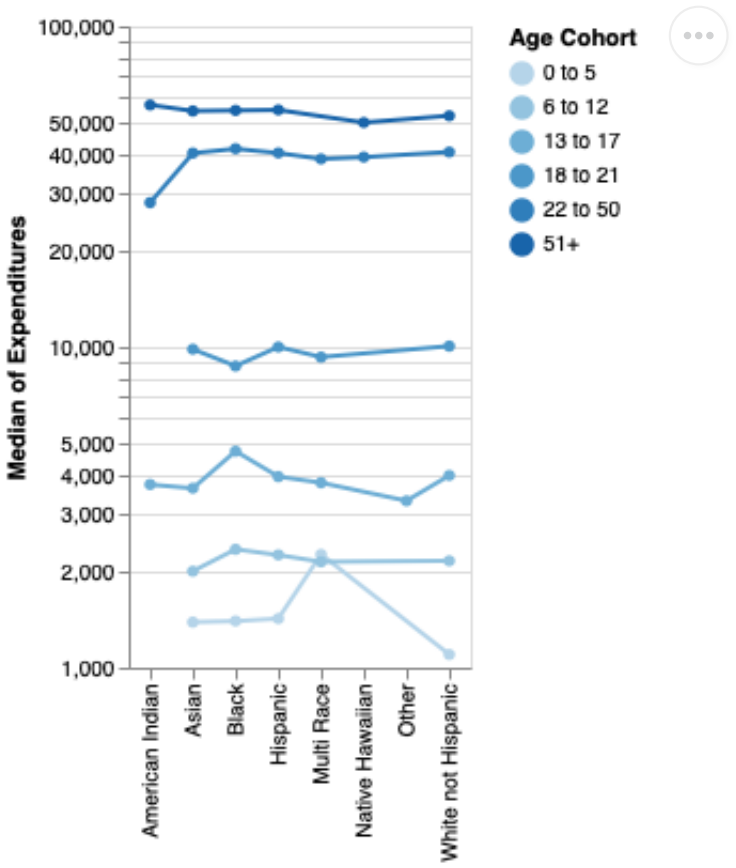
Answer

The exponentiated coefficient estimates for each level of age being larger than 1 seem to indicate that the differences in expenditure between levels are significant. The most obvious example of this is in the Age Cohort 51+ that has an exponentiated coefficient estimate of 43.05133, which is much larger than an exponentiated coefficient estimate of 1. On the other hand, the exponentiated coefficient estimates for each level of ethnicity being close to a factor of 1 seem to suggest that there is no difference in expenditure by ethnicity, and Ethnicity_Other, the least convincing ethnicity level, has a coefficient estimate of -0.189877 and a standard error of 0.232793, indicating that the estimate is well within 1SE of 0.

Now as a final step in the analysis, you'll visualize your results. The idea is simple: plot the estimated mean log expenditures for each group. Essentially you'll make a version of your figure 4 from part 1 in which the points are estimated rather than observed. So the model visualization graphic will look similar to this:

In [20]:
```
fig_4
```

Out[20]:



In order to construct a 'model version' of this plot, however, you'll need to generate estimated mean log expenditures for each unique combination of categorical variable levels. The cell below generates a 'grid' of every such combination.

In [21]:

```
# store unique levels of each categorical variable
genders = reg_data.Gender.unique()
ethnicities = reg_data.Ethnicity.unique()
ages = reg_data['Age Cohort'].unique()

# generate grid of each unique combination of variable levels
gx, ex, ax = np.meshgrid(genders, ethnicities, ages)
ngrid = len(genders)*len(ethnicities)*len(ages)
grid_mx = np.vstack([ax.reshape(ngrid), gx.reshape(ngrid), ex.reshape(ngrid)]).transpose()
grid_df = pd.DataFrame(grid_mx, columns = ['age', 'gender', 'ethnicity']).astype(
    {'gender': 'category', 'ethnicity': 'category', 'age': 'category'}
)

# reorder category levels so consistent with input data
grid_df['ethnicity'] = grid_df.ethnicity.cat.as_ordered().cat.reorder_categories(
    grid_df.ethnicity.cat.categories[[7, 3, 2, 1, 5, 0, 4, 6]]
)
grid_df['gender'] = grid_df.gender.cat.as_ordered().cat.reorder_categories(['Male', 'Female'])
grid_df['age'] = grid_df.age.cat.as_ordered().cat.reorder_categories(
    grid_df.age.cat.categories[[0, 5, 1, 2, 3, 4]]
)
grid_df['cohort_order'] = grid_df.age.cat.codes

# preview
grid_df.head()
```

Out[21]:

	age	gender	ethnicity	cohort_order
0	13 to 17	Female	White not Hispanic	2
1	22 to 50	Female	White not Hispanic	4
2	0 to 5	Female	White not Hispanic	0
3	18 to 21	Female	White not Hispanic	3
4	51+	Female	White not Hispanic	5

Question 2 (c). Model visualization

Your task in this question will be to add fitted values and standard errors to the grid above and then plot it.

(i) Create an explanatory variable matrix from the grid.

Pretend for a moment that you're going to treat `grid_df` as if it were the data. Create a new `x_mx` based on `grid_df`:

1. Use `pd.get_dummies(...)` to obtain the indicator variable encoding of `grid_df`; store the result as `pred_df`.
2. Add an intercept column to `pred_df` using `add_dummy_feature(...)`; store the result (an array) as `pred_mx`.

Print the first three rows and six columns of `pred_mx`.

In [22]:
```
# variable encodings
pred_df = pd.get_dummies(grid_df, columns = ['age', 'gender', 'ethnicity'],
                         drop_first = True).drop(columns = ['cohort_order'])

# add intercept
pred_mx = add_dummy_feature(pred_df, value = 1)

# preview
pred_mx[0:3, 0:6]
```

Out[22]:
```
array([[1., 0., 1., 0., 0., 0.],
       [1., 0., 0., 0., 1., 0.],
       [1., 0., 0., 0., 0., 0.]])
```

(ii) Compute fitted values and standard errors on the grid.

Now add a new column to `grid_df` called `expenditure` that contains the estimated log expenditure (*hint*: use `mlr_predict(...)` with your result from (i) immediately above).

In [23]:
```
# solution
grid_df['expenditure'] = mlr.predict(pred_mx)
```

The cell below adds the standard errors for estimated log expenditure.

In [24]:
```
# add standard errors
grid_df['expenditure_se'] = np.sqrt(pred_mx.dot(xtx_inv).dot(pred_mx.transpose()).diagonal() * sigmasqhat)
```

(iii) Plot the estimated means and standard errors.

Construct a model visualization matching figure 4 in the following steps.

1. Construct a point-and-line plot called `lines` based on `grid_df` with:
   - ethnicity on the x axis;
   - no x axis title;
   - log expenditure on the y axis;
   - the y axis title 'Estimated mean log expenditure';
   - age cohort mapped to the color encoding channel as an *ordinal* variable and shown in ascending cohort order (refer back to your codes for figure 4).
2. Construct an error band plot called `bands` based on `grid_df` with:
   - a `.transform_calculate(...)` step computing lower and upper band boundaries
     - $\texttt{lwr} = \texttt{expenditure} - 2 \times \texttt{expenditure\_se}$ and
     - $\texttt{upr} = \texttt{expenditure} + 2 \times \texttt{expenditure\_se}$;
   - ethnicity on the x axis;
   - no x axis title;
   - `lwr` and `upr` passed to the `y` and `y2` encoding channels;
   - the `y` channel titled 'Estimated mean log expenditure';
   - age cohort mapped to the color channel exactly as in `lines`.
3. Layer `lines` and `bands` and facet the layered chart into columns according to gender. Store the result as `fig_5`.

Display `fig_5`.

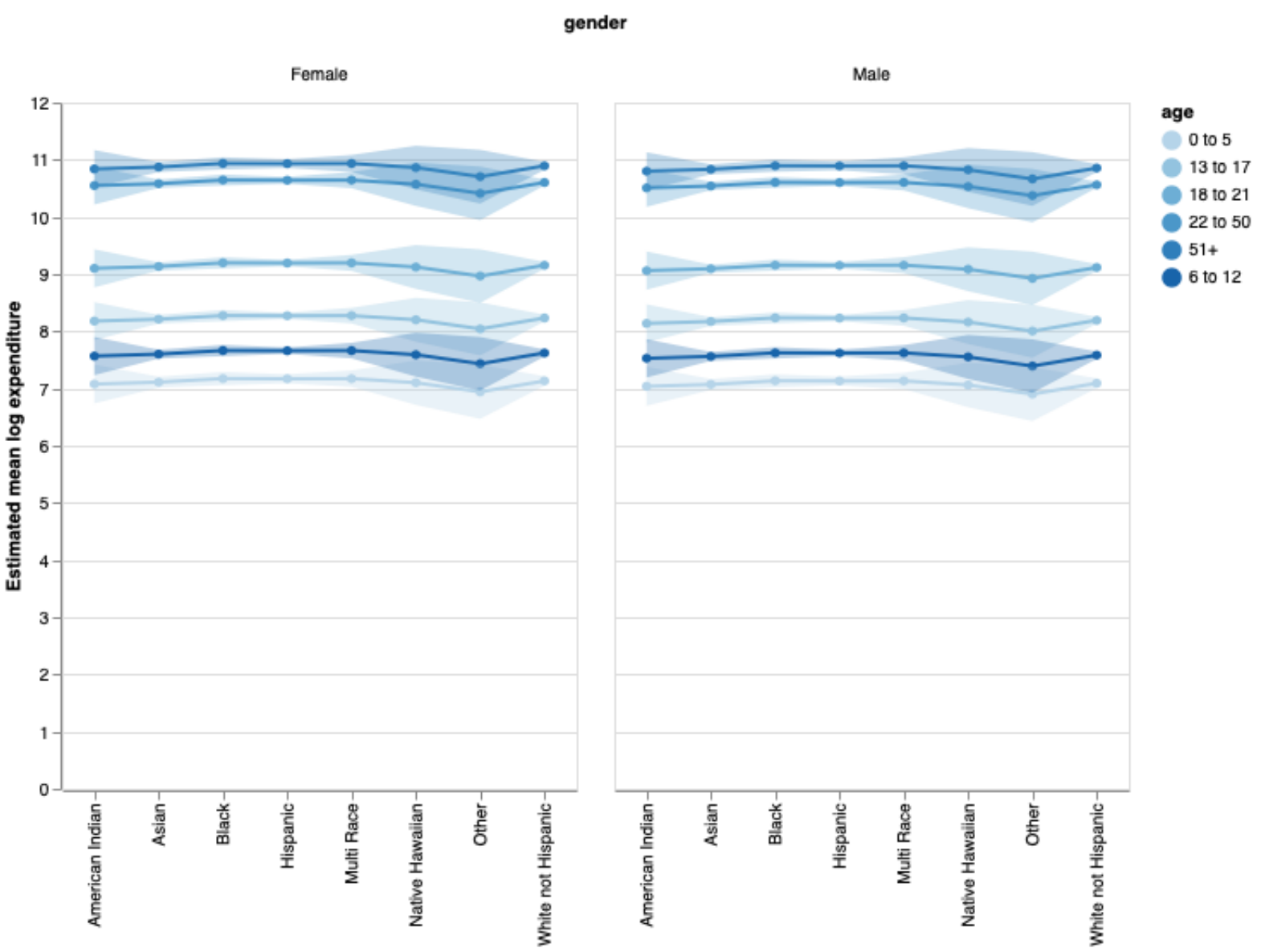
In [25]:
```
# point and line plot
lines = alt.Chart(grid_df).mark_line(point=True).encode(
    x = alt.X('ethnicity', title = '', scale = alt.Scale(zero=False)),
    y = alt.Y('expenditure', title = 'Estimated mean log expenditure'),
    color = alt.Color('age:O', sort = alt.EncodingSortField(field = 'cohort_order', order = 'ascending'))
)

# error bands
bands = alt.Chart(grid_df).transform_calculate(
    lwr = 'datum.expenditure - (2 * datum.expenditure_se)',
    upr = 'datum.expenditure + (2 * datum.expenditure_se)'
).mark_errorband(extent='ci').encode(
    x = alt.X('ethnicity', title = '', scale = alt.Scale(zero=False)),
    y = alt.Y('lwr:Q', title = 'Estimated mean log expenditure'),
    y2 = 'upr:Q',
    color = alt.Color('age:O', sort = alt.EncodingSortField(field = 'cohort_order', order = 'ascending'))
)

# layer and facet
fig_5 = (lines + bands).properties(height=400, width=300).facet(column = 'gender')

# display
fig_5
```

Out[25]:



(iv) Sanity check.

Does the model visualization seem to accurately reflect the pattern in your exploratory plots? Answer in 1 sentence.

Answer

The model visualization does seem to accurately reflect the pattern in our exploratory plots due to the fact that there are significant differences in estimated mean log expenditure between the age levels (with older age cohorts having larger estimated mean log expenditure) but not too much variation based on the ethnicity levels.

(v) Which estimates have greater uncertainty and why?

Identify the ethnic groups for which the uncertainty band is relatively wide in the plot. Why might uncertainty be higher for these groups? Answer in 2 sentences.

(*Hint*: it may help to refer to figure 3.)

Answer

The ethnic groups that have relatively wide uncertainty bands in the plot are American Indian, Native Hawaiian, and Other. Uncertainty seems to be higher for these groups due to the sample sizes for these groups being significantly smaller than the rest of the ethnic groups.

---

3. Communicating results

Review your exploratory and regression analyses above, and then answer the following questions.

Question 3 (a). Summary

Write a one-paragraph summary of your analysis. Focus on answering the question, 'do the data provide evidence of ethnic or gender discrimination in allocation of DDS funds?'

Your summary should include the following:

- a one-sentence description of the data indicating observations, variables, and whether they are a random sample;
- one to two sentences describing any important exploratory findings;
- a one-sentence description of the method you used to analyze the data (don't worry about capturing every detail);
- one sentence desribing findings of the analysis;
- an answer to the question.

Answer

A random sample was conducted (since the total sample size was so neatly 1000) to gather data about DDS expenditure information for different ethnic groups, along with additional information such as ID, age, and gender. After making different exploratory plots, we found that the data provided evidence of a difference in the allocation of DDS funds depending on the different age cohorts. We were able to come to this conclusion by using visualizations to analyze the different relationships between the covariates and in order to further support our claims, we performed linear regression to ensure our conclusions from earlier were consistent with our regression findings. In conclusion, we found that the data do not provide evidence of ethnic or gender discrimination in the allocation of DDS funds.

Question 3 (b). Supporting information

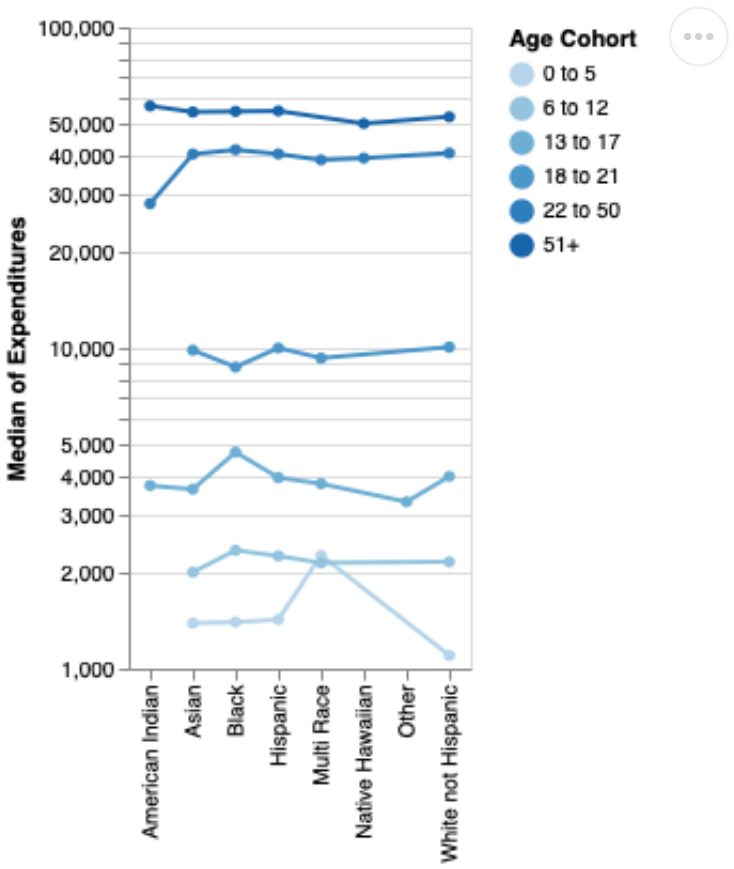
Choose one table or figure from part 1 and one table and figure from part 2 that support your summary of results. Write a caption for each of your choices.

(i) First figure/table.

Figure 4 shows the median expenditures of each ethnicity, separated by age cohort. The points are connected by age cohort and generally form a level line, indicating that across all ethnicities, the median expenditure is similar for each age cohort.

In [26]:
```
# show figure/table
fig_4
```

Out[26]:



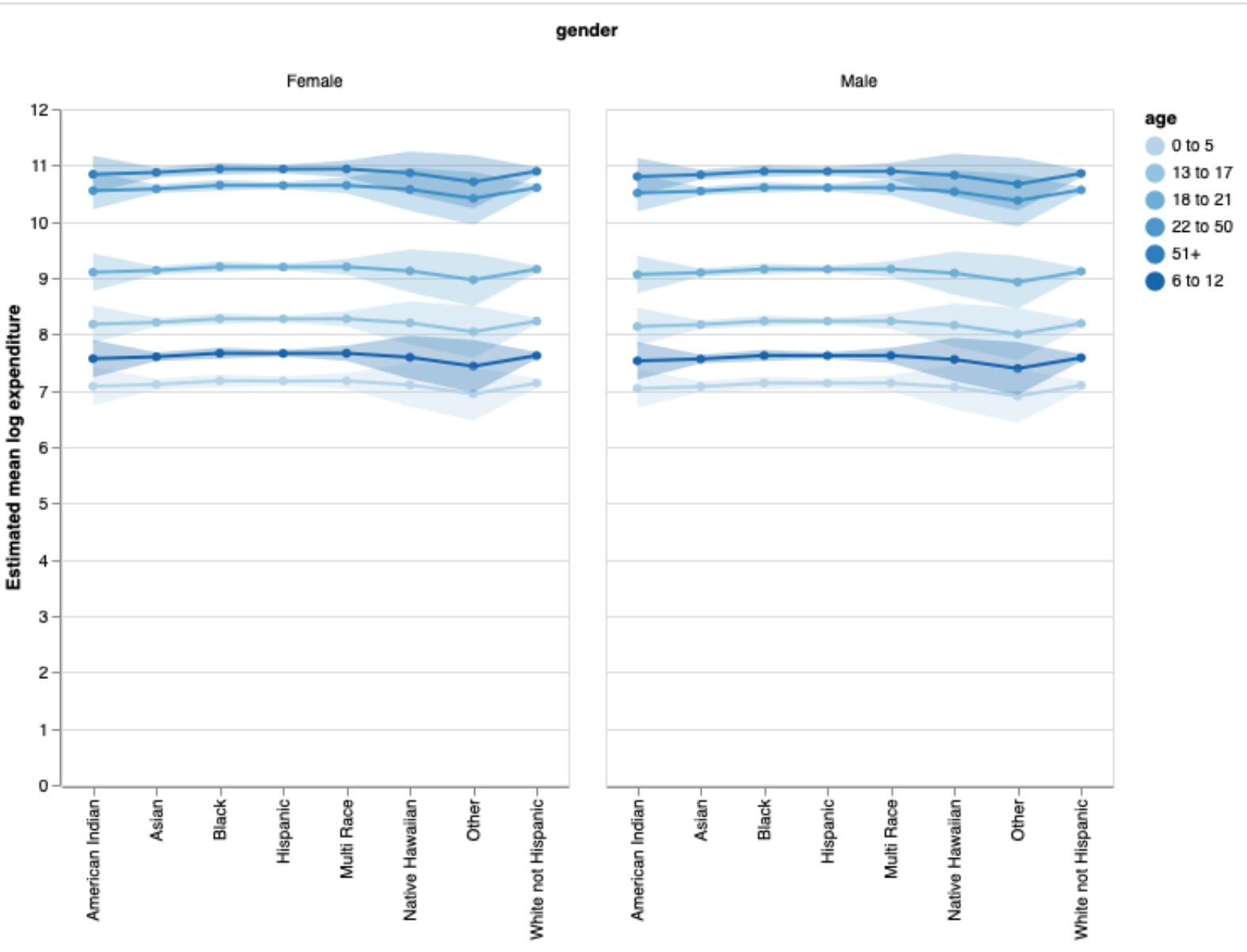
(ii) Second figure/table.

Figure 5 shows two mean log expenditurea for each ethnicity, separated by age cohort, and split by sex. The points are connected by age cohort and generally form a level line, indicating that across all ethnicities, the mean log expenditure is similar for each age cohort. Furthermore, the lines seem identical between sex, indicating that there is no difference in mean log expenditure based on sex.

In [27]:

```
# show figure/table
fig_5
```

Out[27]:



Submission Checklist

1. Save file to confirm all changes are on disk
2. Run *Kernel > Restart & Run All* to execute all code from top to bottom
3. Save file again to write any new output to disk
4. Select *File > Download as > HTML*.
5. Open in Google Chrome and print to PDF on A3 paper in portrait orientation.
6. Submit to Gradescope