PSTAT 100 Homework 3

In [ ]:
```
# Initialize Otter
import otter
grader = otter.Notebook("hw3-dds.ipynb")
```

In [ ]:
```
import numpy as np
import pandas as pd
import altair as alt
import sklearn.linear_model #as lm
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import add_dummy_feature

import warnings
warnings.simplefilter(action='ignore', category=FutureWarning)
```

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

Out[ ]:

	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 consumers (developmentally-disabled residents)*.

(ii) Identify the population of interest.

Answer: *The population of interest are the 250,000 developmentally-disabled residents of California*.

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

Answer: *This is a random sample*.

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

Answer: *Yes, it is possible to make inferences about the population based on this data, since the statistical properties of the sample are expected to match those of the population*.

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 of consumer	Categorical	Years
Age	Exact Age of consumer	Numeric	Years
Gender	Gender of consumer	Categorical	None
Expenditures	Amount spent per yr. on consumer	Numeric	USD ($)
Ethnicity	Ethnic group of 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 [ ]:

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

# compute sample sizes
ethnicity_n = dds.loc[:,'Ethnicity'].value_counts().rename('n')
#type(ethnicity_n)
```

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

# print
tbl_1
```

Out[ ]:

	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. 'White not Hispanic' has a large sample size and a 15k+ median whereas 'Hispanic' and 'Asian' have only 3k - 9k for their median.

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

Answer

'American Indian', 'Native Hawaiian', 'Multi Race', and 'Other' all have small sample sizes. Because of this, it can very easily skew the median to extremely large or small values for those groups and is not actually an acurate picture.

(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:

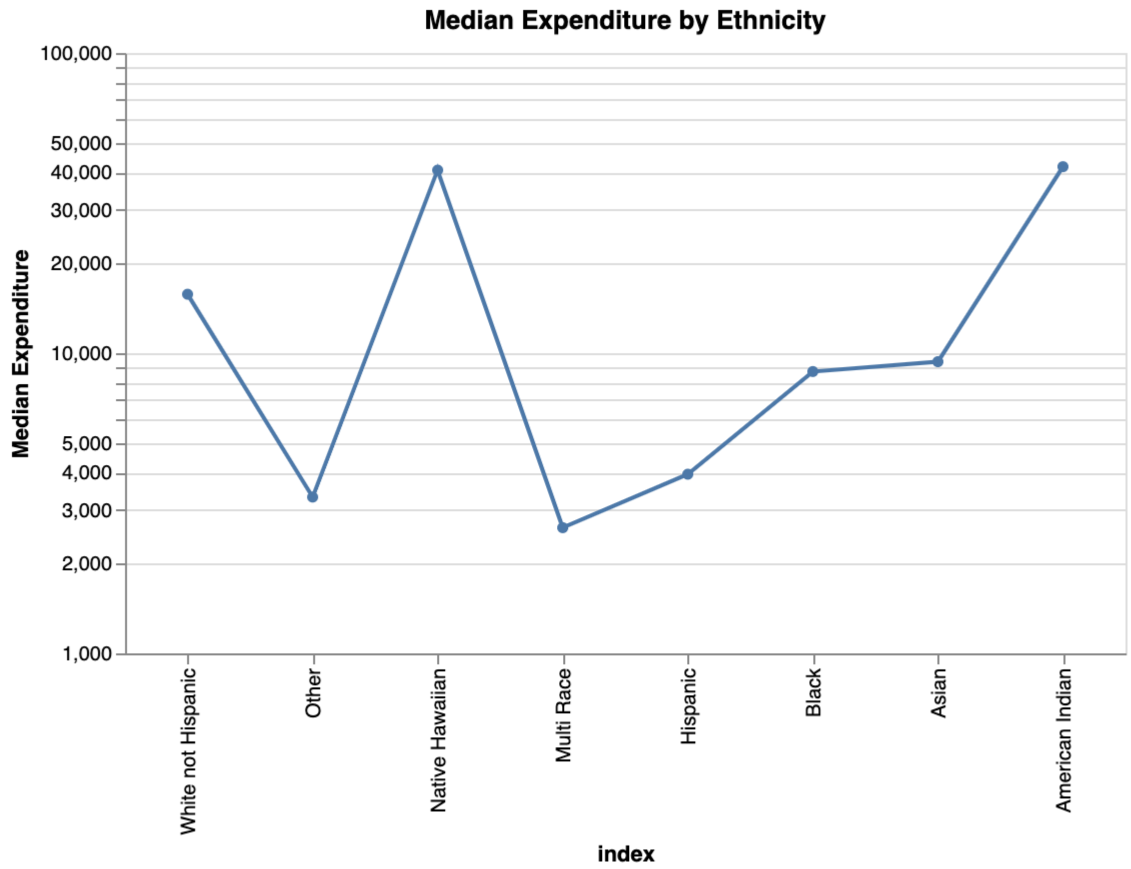
- you'll need to use `tbl_1.reset_index()` to obtain the ethnicity group as a variable;
- recall that `.mark_line(point = True)` will add points to a line plot;
- sorting can be done using `alt.X(..., sort = alt.EncodingSortField(field = ..., order = ...))`

In [ ]:

```
# solution
fig_1 = alt.Chart(tbl_1.reset_index()).mark_line(point=True).encode(
    x = alt.X('index:N', scale = alt.Scale(zero = False),
              sort = alt.EncodingSortField(field='index', order='descending')),
    y = alt.Y('Expenditures', title = 'Median Expenditure', scale = alt.Scale(type = 'log'))
).properties(width = 500,
             title = 'Median Expenditure by Ethnicity')

fig_1
```

Out[ ]:



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.

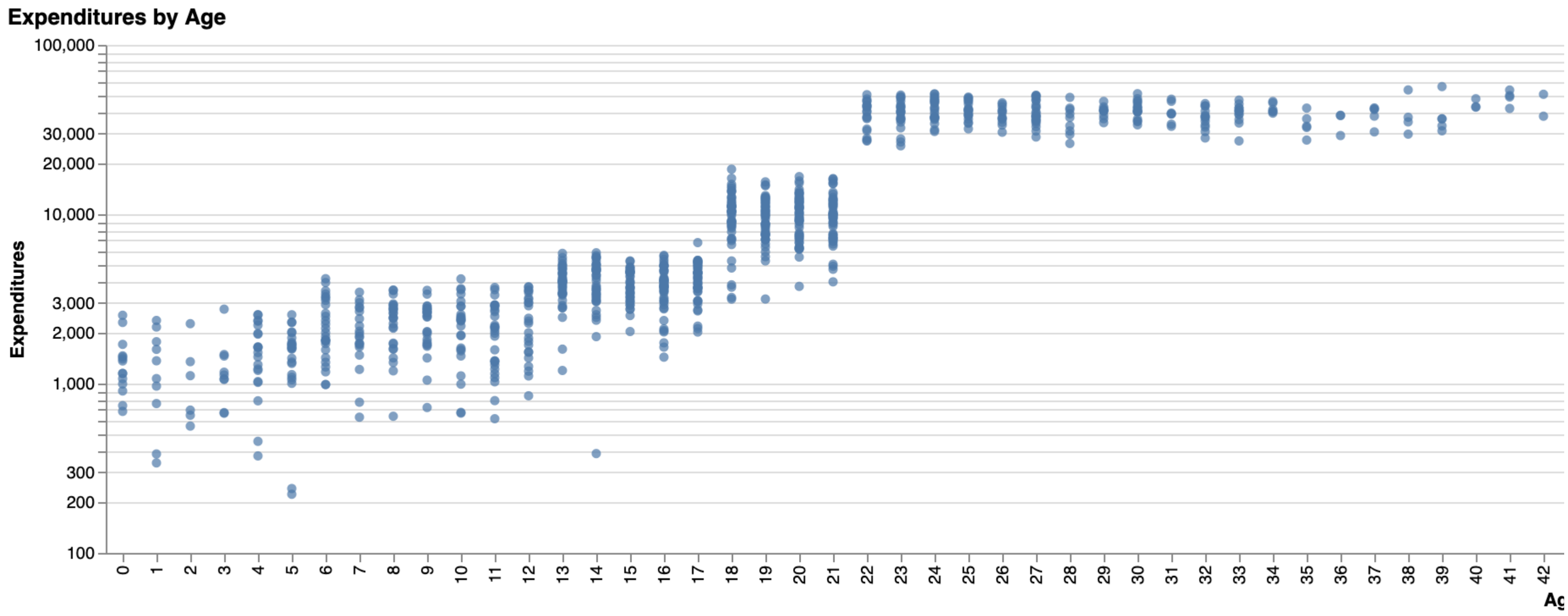
Store the plot as `fig_2` and display the graphic.

In [ ]:

```
# solution
fig_2 = alt.Chart(dds).mark_circle().encode(
    x = alt.X('Age:N', scale = alt.Scale(type = 'linear')),
    y = alt.Y('Expenditures', scale = alt.Scale(type = 'log'))
).properties(title = 'Expenditures by Age')

fig_2.configure_title(
    anchor='start',
)
```

Out[ ]:



(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

The relationship is somewhat linear in the beginning until around age 21. Past this age, the graph makes a jump and flattens out and we don not observe a linear relationship anymore. The overall direction is always positive. In the beginning this is a steep or strong relationship, and later becomes slight and less strong.

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

Answer

As age increases, up until around age 21, there is a positive relationship and trend with expenditures. An increase in age corresponds with an increase in spending. After this age, expenditure amount flatlines and is about the same for the rest of the ages.

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

Answer

After age 20, most people are at the age when they move away from home or are graduating from school and begining to do things on their own. This means that they need a bit more money since their situations are changing and are moving towards a lifestyle that just costs more money. They might also have less help from family since they have moved out, and need to get that help from somewhere else.

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

# print
samp_sizes
```

Out[ ]:

	Age Cohort	Ethnicity	n
0	0 to 5	American Indian	0
1	0 to 5	Asian	8
2	0 to 5	Black	3
3	0 to 5	Hispanic	44
4	0 to 5	Multi Race	7
5	0 to 5	Native Hawaiian	0
6	0 to 5	Other	0
7	0 to 5	White not Hispanic	20
8	6 to 12	American Indian	0
9	6 to 12	Asian	18
10	6 to 12	Black	11
11	6 to 12	Hispanic	91
12	6 to 12	Multi Race	9
13	6 to 12	Native Hawaiian	0
14	6 to 12	Other	0
15	6 to 12	White not Hispanic	46
16	13 to 17	American Indian	1
17	13 to 17	Asian	20
18	13 to 17	Black	12
19	13 to 17	Hispanic	103
20	13 to 17	Multi Race	7
21	13 to 17	Native Hawaiian	0
22	13 to 17	Other	2
23	13 to 17	White not Hispanic	67
24	18 to 21	American Indian	0
25	18 to 21	Asian	41
26	18 to 21	Black	9
27	18 to 21	Hispanic	78
28	18 to 21	Multi Race	2
29	18 to 21	Native Hawaiian	0
30	18 to 21	Other	0
31	18 to 21	White not Hispanic	69
32	22 to 50	American Indian	1
33	22 to 50	Asian	29
34	22 to 50	Black	17
35	22 to 50	Hispanic	43
36	22 to 50	Multi Race	1
37	22 to 50	Native Hawaiian	2
38	22 to 50	Other	0
39	22 to 50	White not Hispanic	133
40	51+	American Indian	2
41	51+	Asian	13
42	51+	Black	7
43	51+	Hispanic	17
44	51+	Multi Race	0
45	51+	Native Hawaiian	1
46	51+	Other	0
47	51+	White not Hispanic	66

(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.

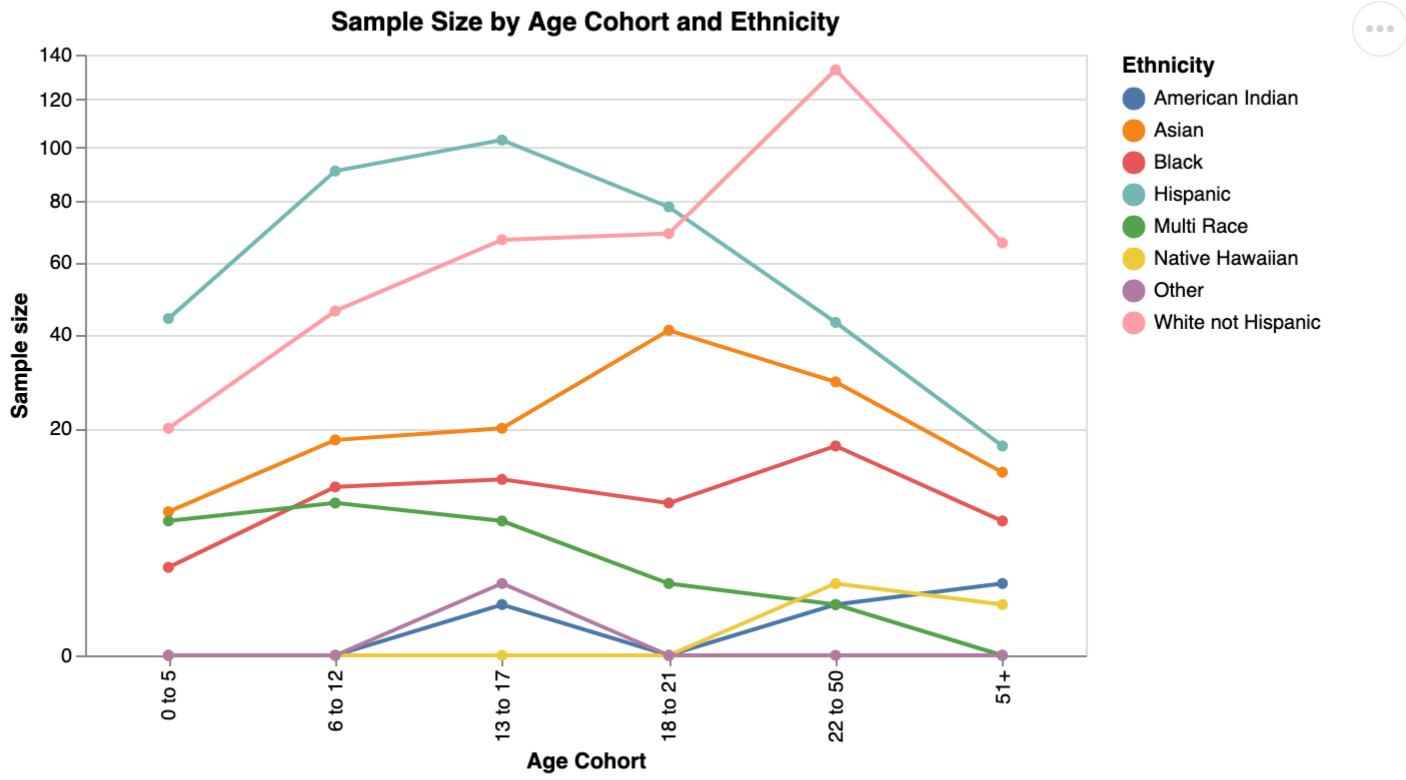
(*Hint*: sorting can be done using `alt.X(..., sort = alt.EncodingSortField(field = ..., order = ...))`.)

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

# construct plot
fig_3 = alt.Chart(samp_sizes).mark_line(point=True).encode(
    x = alt.X('Age Cohort:N', sort=alt.EncodingSortField(field='cohort_order', order='ascending')),
    y = alt.Y('n', title = 'Sample size', scale = alt.Scale(type = 'sqrt')),
    color = 'Ethnicity'
).properties(width = 500,
             title = 'Sample Size by Age Cohort and Ethnicity')

# display
fig_3
```

Out[ ]:



(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

There are differences in age structure, according to the chart created above. We can see that the red line is relatively flat and there are about the same number of people in each age group. However this is different from the pink line corresponding to "White not Hispanic" has an overall larger number for each group and the blue line for Hispanic is decreasing.

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.

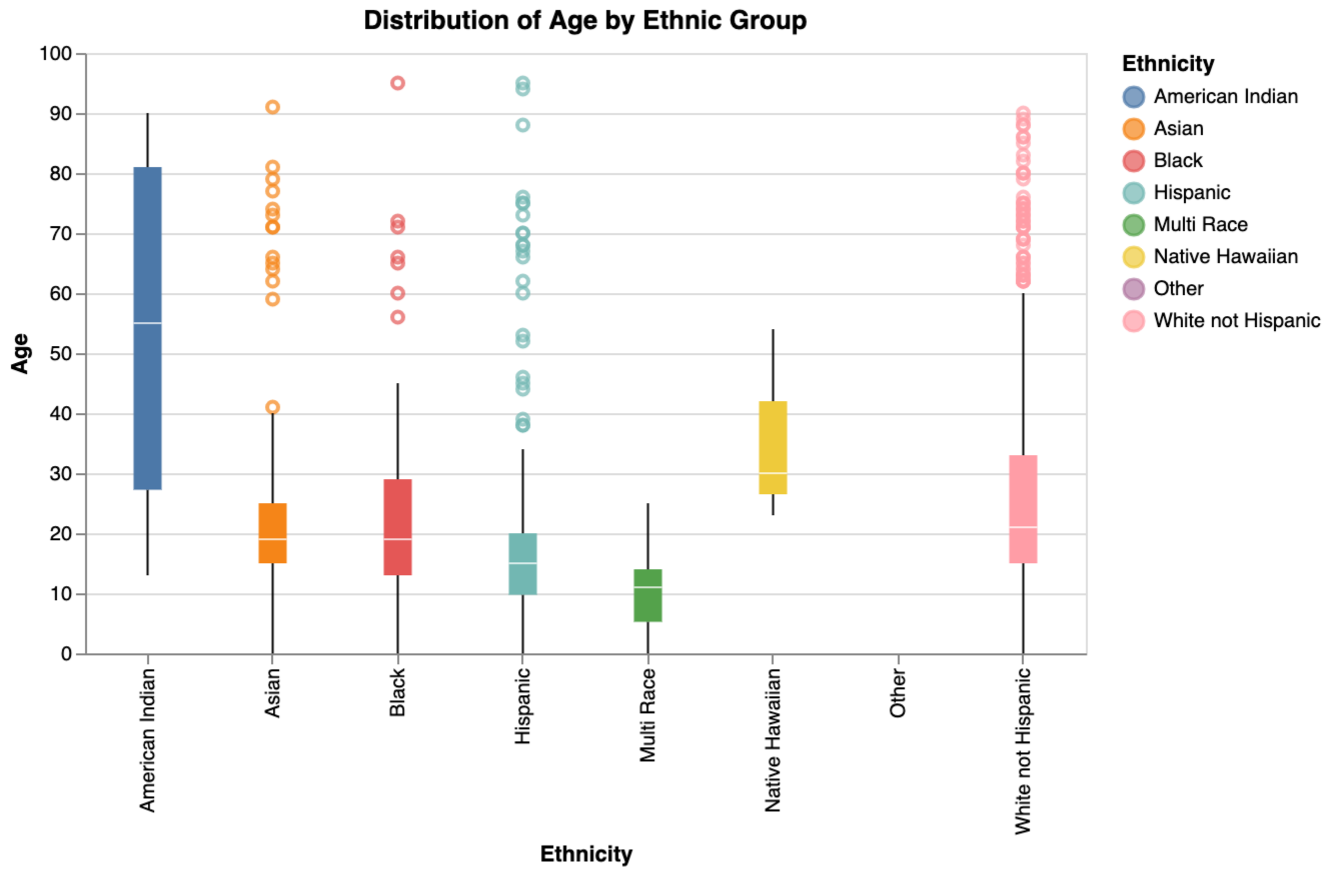
(i) Distribution of Age by ethnic group

Construct the boxplots of the distribution of age by ethnic group.

1. Construct an Altair chart based on `dds_cat` with:
   - Ethnicity on the x axis;
   - Age on the y axis;
   - ethnic group mapped to color.

In [ ]:
```
# solution
alt.Chart(dds_cat).mark_boxplot(outliers=True).encode(
    x = 'Ethnicity',
    y = 'Age',
    color = 'Ethnicity'
).properties(
    width = 500,
    title = 'Distribution of Age by Ethnic Group')
```

Out[ ]:



(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.

In [ ]:

```
dds_cat[dds_cat['Ethnicity'] == 'Multi Race']
```

Out[ ]:

	Id	Age Cohort	Age	Gender	Expenditures	Ethnicity	cohort_order
13	11189	13 to 17	17	Male	5340	Multi Race	2
30	12850	13 to 17	13	Male	3775	Multi Race	2
84	18383	0 to 5	0	Male	1149	Multi Race	0
145	22988	13 to 17	16	Male	4664	Multi Race	2
191	26437	0 to 5	0	Male	2296	Multi Race	0
243	31168	6 to 12	11	Female	2918	Multi Race	1
288	35360	6 to 12	10	Female	1622	Multi Race	1
330	39942	13 to 17	14	Male	3399	Multi Race	2
362	43291	6 to 12	11	Male	2140	Multi Race	1
393	45755	6 to 12	11	Male	1144	Multi Race	1
410	47043	22 to 50	25	Male	38619	Multi Race	4
443	50222	18 to 21	19	Female	7564	Multi Race	3
517	56736	18 to 21	18	Female	11054	Multi Race	3
569	61120	6 to 12	7	Male	3000	Multi Race	1
570	61187	6 to 12	11	Male	2885	Multi Race	1
668	69542	0 to 5	5	Female	1053	Multi Race	0
686	71073	13 to 17	14	Female	5062	Multi Race	2
839	84388	0 to 5	2	Female	697	Multi Race	0
871	87444	13 to 17	14	Female	1893	Multi Race	2
906	90953	6 to 12	10	Female	669	Multi Race	1
934	93628	6 to 12	6	Male	3259	Multi Race	1
948	94595	0 to 5	4	Female	2335	Multi Race	0
977	97426	0 to 5	1	Female	2359	Multi Race	0
978	97793	6 to 12	9	Female	1048	Multi Race	1
994	99529	0 to 5	2	Male	2258	Multi Race	0
997	99718	13 to 17	17	Female	3673	Multi Race	2

Answer

The median expenditure for the Multi Race group is lower than others because if we look at the dataframe corresponding to only Multi Race, we can see that the maximum age is only 25. Previously, we saw that the expenditures is much higher after around age 20, so when this group has predominantly young ages, they require lower spend and therefore this group has a lower overall median.

(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 [ ]:
```
# solution
dds_cat[dds_cat['Ethnicity'] == 'American Indian']
```

Out[ ]:

	Id	Age Cohort	Age	Gender	Expenditures	Ethnicity	cohort_order
231	30234	51+	78	Female	55430	American Indian	5
575	61498	13 to 17	13	Female	3726	American Indian	2
730	74721	51+	90	Female	58392	American Indian	5
788	79645	22 to 50	32	Male	28205	American Indian	4

Answer

The median expenditure for the American Indian group is so high because there are only 4 observations and 2 of those are very large. This will skew the median and is not a good measure of the middle for this group.

(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.

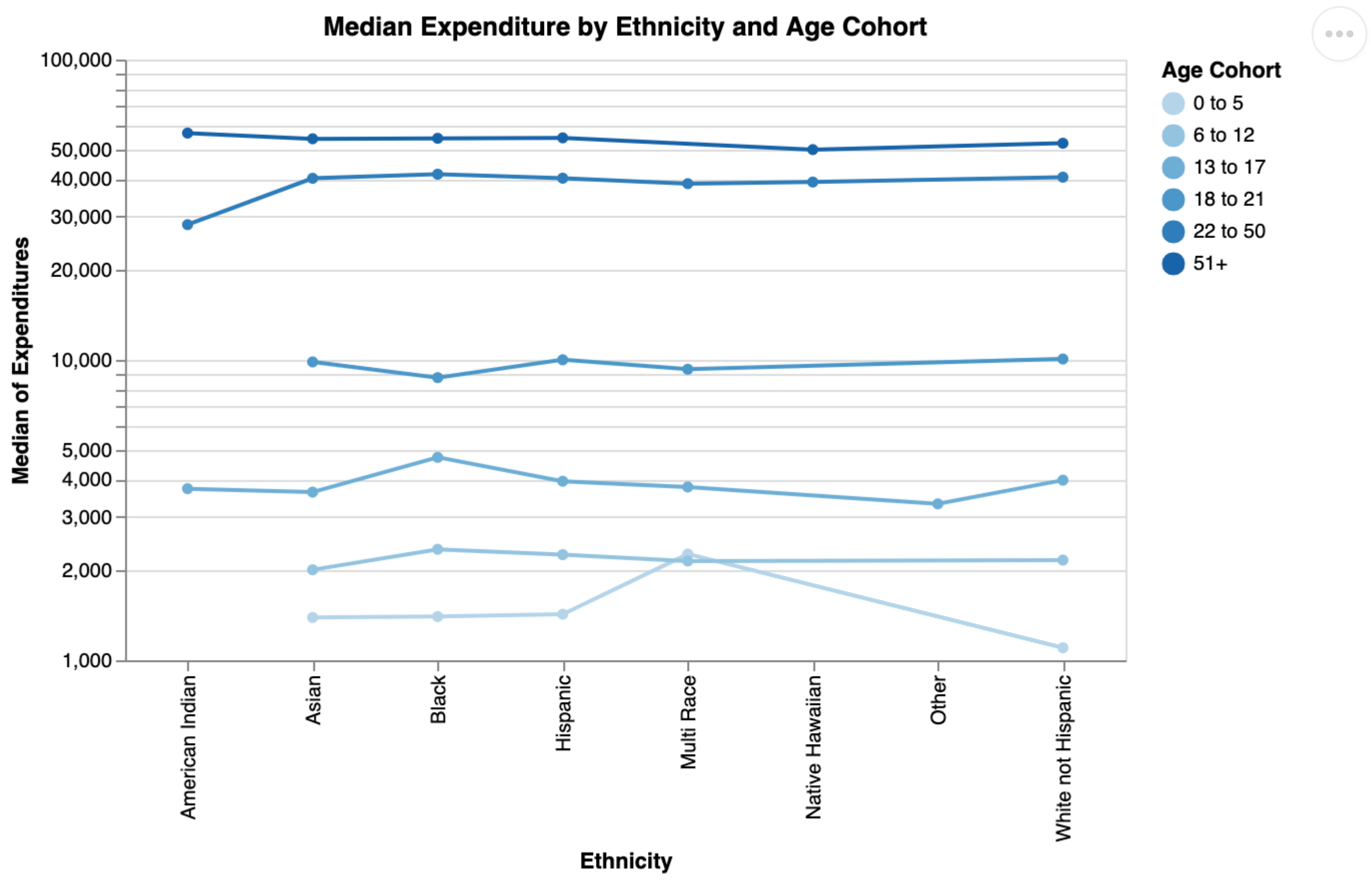
Store the result as `fig_4` and display the graphic.

In [ ]:
```
# 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:N'),
    y = alt.Y('median(Expenditures)', scale = alt.Scale(type = 'log')),
    color = alt.Color('Age Cohort:O', sort=alt.EncodingSortField(field='cohort_order', order='ascending'))
).properties(width = 500,
             title = 'Median Expenditure by Ethnicity and Age Cohort')

# display
fig_4
```

Out[ ]:



(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

No, the data does not reflect a difference in median expenditure by ethnicity. The increase in spending is related to the different age cohorts and not their respective ethnic groups. Across all ethnicities, the spending is about the same.

---

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:

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

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:

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

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 [ ]:
```
# 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'])
reg_data
```

Out[ ]:

	Age Cohort	Gender	Expenditures	Ethnicity	cohort_order
0	13 to 17	Female	2113	White not Hispanic	2
1	22 to 50	Male	41924	White not Hispanic	4
2	0 to 5	Male	1454	Hispanic	0
3	18 to 21	Female	6400	Hispanic	3
4	13 to 17	Male	4412	White not Hispanic	2
...	...	...	...	...	...
995	51+	Female	57055	White not Hispanic	5
996	18 to 21	Male	7494	Hispanic	3
997	13 to 17	Female	3673	Multi Race	2
998	6 to 12	Male	3638	Hispanic	1
999	22 to 50	Male	26702	White not Hispanic	4

1000 rows × 5 columns

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 [ ]:

```
# solution
x_df = pd.get_dummies(reg_data, drop_first=True).drop(columns = ['Expenditures', 'cohort_order'])
x_df.iloc[0:3, 0:6]
```

Out[ ]:

	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 [ ]:

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

Out[ ]:

```
array([[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 [ ]:

```
# solution
y = np.log(reg_data.Expenditures)
y.iloc[0:11, ]
```

Out[ ]:

```
0      7.655864
1     10.643614
2      7.282074
3      8.764053
4      8.392083
5      8.426393
6      8.272571
7      8.261785
8      8.521384
9      7.967973
10     8.331827
Name: Expenditures, dtype: float64
```

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 [ ]:
```
# solution
mlr = LinearRegression(fit_intercept = False)
mlr.fit(x_mx, y)
```

Out[ ]:
```
LinearRegression(fit_intercept=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 [ ]:
```
# 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
se = np.sqrt(v_hat.diagonal())
coef_se = np.append(se, float('nan'))

# coefficient labels
coef_labels = np.append("intercept", list(x_df.columns.values))
coef_labels = np.append(coef_labels, 'error_variance')

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

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

# print
coef_table
```

Out[ ]:

	coefficent estimate	coefficient standard errors
intercept	7.092439	0.041661
Age Cohort_6 to 12	0.490276	0.043855
Age Cohort_13 to 17	1.101010	0.042783
Age Cohort_18 to 21	2.023844	0.043456
Age Cohort_22 to 50	3.470836	0.043521
Age Cohort_51+	3.762393	0.049561
Gender_Female	0.039784	0.020749
Ethnicity_Hispanic	0.038594	0.024893
Ethnicity_Black	0.041713	0.045725
Ethnicity_Asian	-0.021103	0.033470
Ethnicity_Native Hawaiian	-0.030725	0.189967
Ethnicity_American Indian	-0.054396	0.164910
Ethnicity_Multi Race	0.041024	0.067680
Ethnicity_Other	-0.189877	0.232910
error_variance	0.107005	NaN

In [ ]:

```
grader.check("q2_b_ii")
```

Out[ ]:

q2_b_ii passed! 🚀

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 [ ]:

```
# exponentiate age (not required)
np.exp(coef_estimates)

# 1.63276654e+00
# 3.00720315e+00
# 7.56735586e+00
# 3.21636320e+01
# 4.30513297e+01
```

Out[ ]:

```
array([1.20283826e+03, 1.63276654e+00, 3.00720315e+00, 7.56735586e+00,
       3.21636320e+01, 4.30513297e+01, 1.04058576e+00, 1.03934836e+00,
       1.04259499e+00, 9.79118208e-01, 9.69742449e-01, 9.47056543e-01,
       1.04187723e+00, 8.27060911e-01, 1.11293969e+00])
```

In [ ]:

```
# exponentiate ethnicity (not requried)
np.exp(coef_estimates)

# 1.03934836e+00
# 1.04259499e+00
# 9.79118208e-01
# 9.69742449e-01
# 9.47056543e-01
# 1.04187723e+00
# 8.27060911e-01

# Most of these are close to 1
```

Out[ ]:

```
array([1.20283826e+03, 1.63276654e+00, 3.00720315e+00, 7.56735586e+00,
       3.21636320e+01, 4.30513297e+01, 1.04058576e+00, 1.03934836e+00,
       1.04259499e+00, 9.79118208e-01, 9.69742449e-01, 9.47056543e-01,
       1.04187723e+00, 8.27060911e-01, 1.11293969e+00])
```

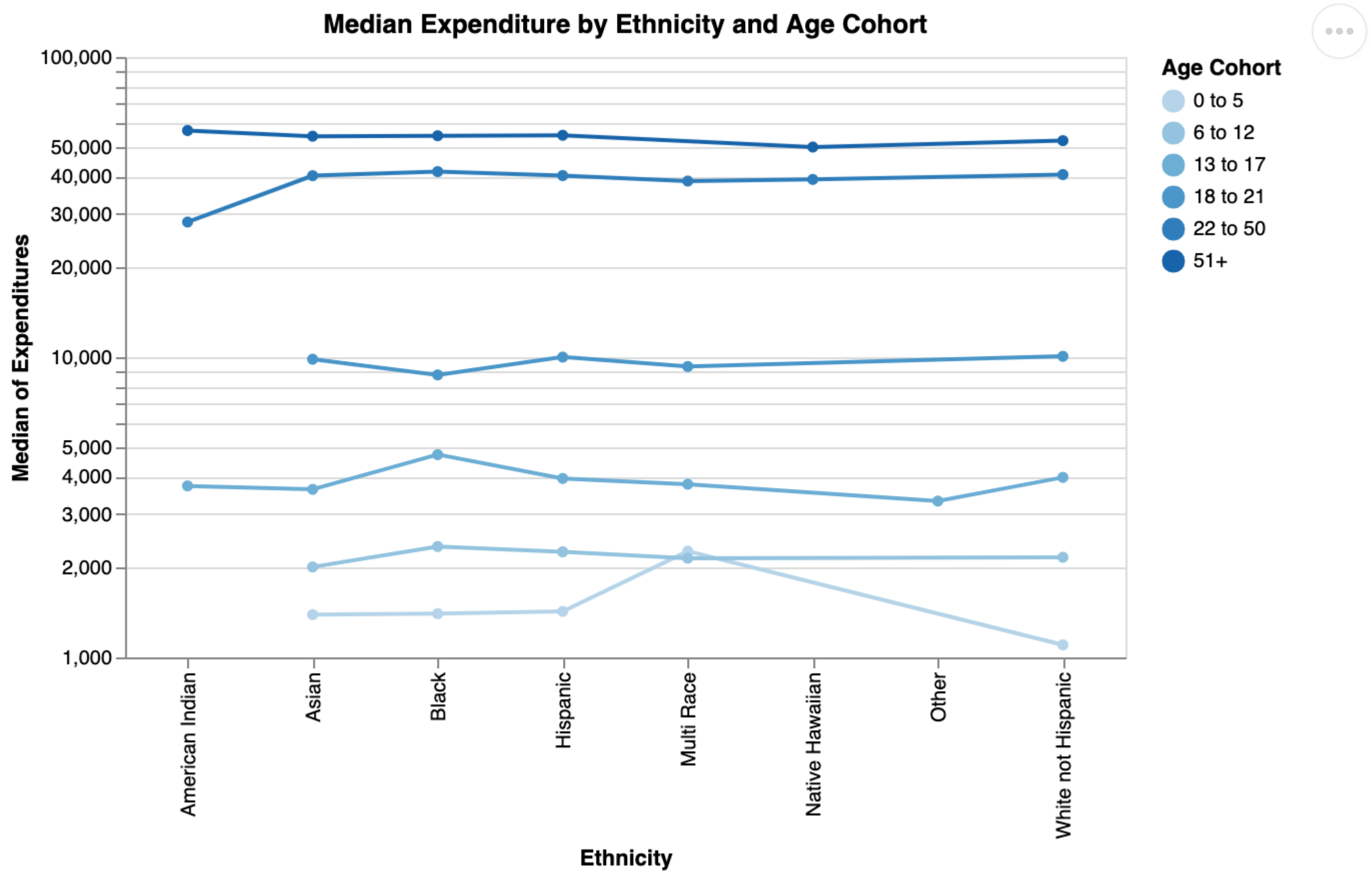
Answer

The parameter estimates differences suggest differences by age, and not by ethnicity.

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 [ ]:
```
fig_4
```

Out[ ]:



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 [ ]:
```
# 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[ ]:

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

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

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

Out[ ]:

```
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 [ ]:

```python
# solution
grid_df['expenditure'] = mlr.predict(pred_mx)
grid_df
```

Out[ ]:

	age	gender	ethnicity	cohort_order	expenditure
0	13 to 17	Female	White not Hispanic	2	8.233233
1	22 to 50	Female	White not Hispanic	4	10.603059
2	0 to 5	Female	White not Hispanic	0	7.132223
3	18 to 21	Female	White not Hispanic	3	9.156067
4	51+	Female	White not Hispanic	5	10.894616
...	...	...	...	...	...
91	22 to 50	Male	Native Hawaiian	4	10.532551
92	0 to 5	Male	Native Hawaiian	0	7.061714
93	18 to 21	Male	Native Hawaiian	3	9.085558
94	51+	Male	Native Hawaiian	5	10.824108
95	6 to 12	Male	Native Hawaiian	1	7.551990

96 rows × 5 columns

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

In [ ]:

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

Out[ ]:

	age	gender	ethnicity	cohort_order	expenditure	expenditure_se
0	13 to 17	Female	White not Hispanic	2	8.233233	0.029081
1	22 to 50	Female	White not Hispanic	4	10.603059	0.026060
2	0 to 5	Female	White not Hispanic	0	7.132223	0.041358
3	18 to 21	Female	White not Hispanic	3	9.156067	0.029215
4	51+	Female	White not Hispanic	5	10.894616	0.034409
...	...	...	...	...	...	...
91	22 to 50	Male	Native Hawaiian	4	10.532551	0.189774
92	0 to 5	Male	Native Hawaiian	0	7.061714	0.193982
93	18 to 21	Male	Native Hawaiian	3	9.085558	0.191777
94	51+	Male	Native Hawaiian	5	10.824108	0.191167
95	6 to 12	Male	Native Hawaiian	1	7.551990	0.192156

96 rows × 6 columns

(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
    - lwr=expenditure-2xexpenditure_se
    - upr=expenditure+2xexpenditure_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`.

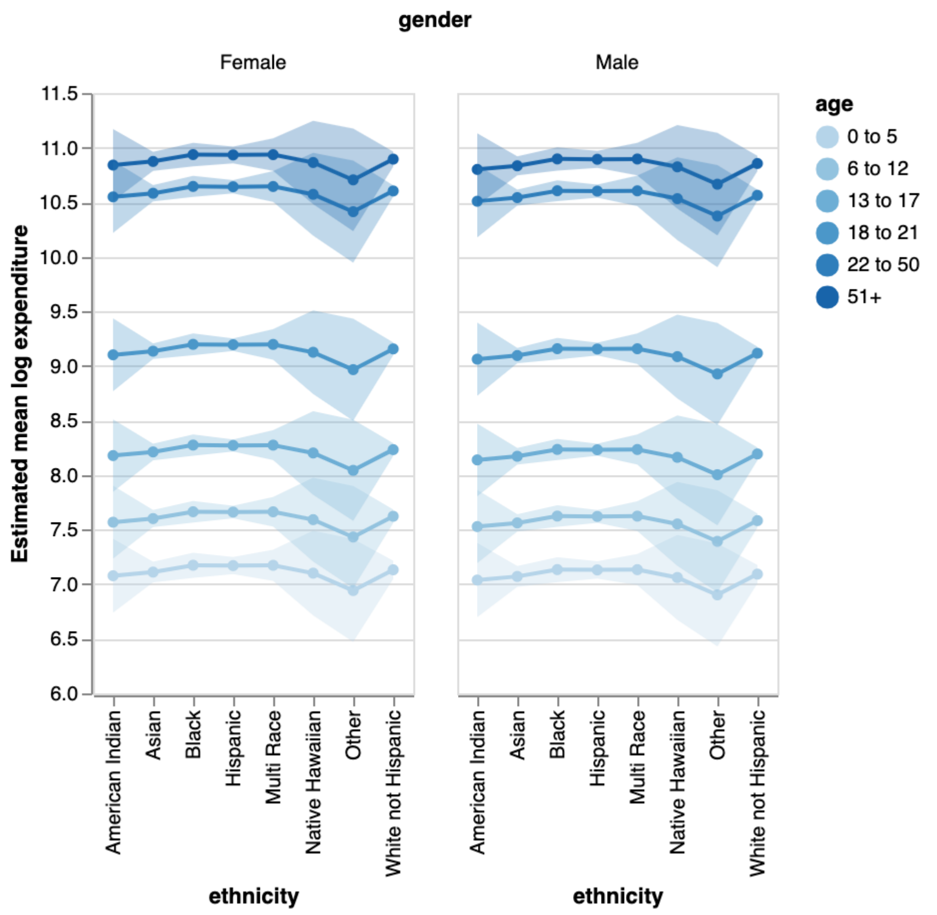
In [ ]:
```python
# point and line plot
lines = alt.Chart(grid_df).mark_line(point=True).encode(
    x = alt.X('ethnicity:N'),
    y = alt.Y('expenditure', title='Estimated mean log expenditure', scale = alt.Scale(zero = False)),
    color = alt.Color('age:O', sort=alt.EncodingSortField(field='cohort_order',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

# error bands
bands = lines.transform_calculate(lwr = 'datum.expenditure - 2*datum.expenditure_se',
                                  upr = 'datum.expenditure + 2*datum.expenditure_se').mark_errorband().encode(
    alt.X('ethnicity:N'), alt.Y('lwr:Q', title = 'Estimated mean log expenditure'), alt.Y2('upr:Q'),
    color = alt.Color('age:O', sort=alt.EncodingSortField(field='cohort_order', order='ascending')))

# layer and facet
fig5 = lines + bands

# display
fig5.facet('gender')
```

Out[ ]:



(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 that we observed in the intial exploratory plots. There is hardly any difference in spending when accounting for gender or ethnicity. The only reason for an increase in spend is because of age, and this is what we expected to see.

(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 uncertainty band is relatively wide for the 'American Indian', 'Native Hawaiian' and 'Other' groups. This is because the sample sizes for each of these are very low when compared to others. In figure 3, we can barely see the lines since the number in each age cohort is so small, less than 5 for each cohort, and this will surely create a larger uncertainty band.

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

The DDS data in this homework contains observations of individuals of all different ages, genders, and ethnic backgrounds, and we were tasked with investigating the available data to see if there was evidence discrimination with how funds were allocated. The variables in this dataset include those mentioned previously, as well as expenditure amount, an ID varable,and 'Age Cohort' which is a range of ages that a person will fall into. This is also a random sample, because of course we are not given data on all 250,000+ people that funds are given to, and there is nothing specific as to the individuals in the dataset that we have for being included. As seen in the exploratory plots, spending tends to increase from age 0 - 22, where after that it stays relatively contstant. We also see that there are age differences across the age groups, but the median spending is about the same for each ethnic group, when accounting for the ones that have a sizeable amount of data. To analyze the data, multiple linear regression techniques were used as well as charting and data manipulation. Overall, after inspecting the data and creating several plots, we can clearly see that the data does not provide any evidence of ethnic or gender discrimination in the allocation of DDS funds The amount recieved is soley dependent upon individuals age.

Question 3 (b). Supporting information

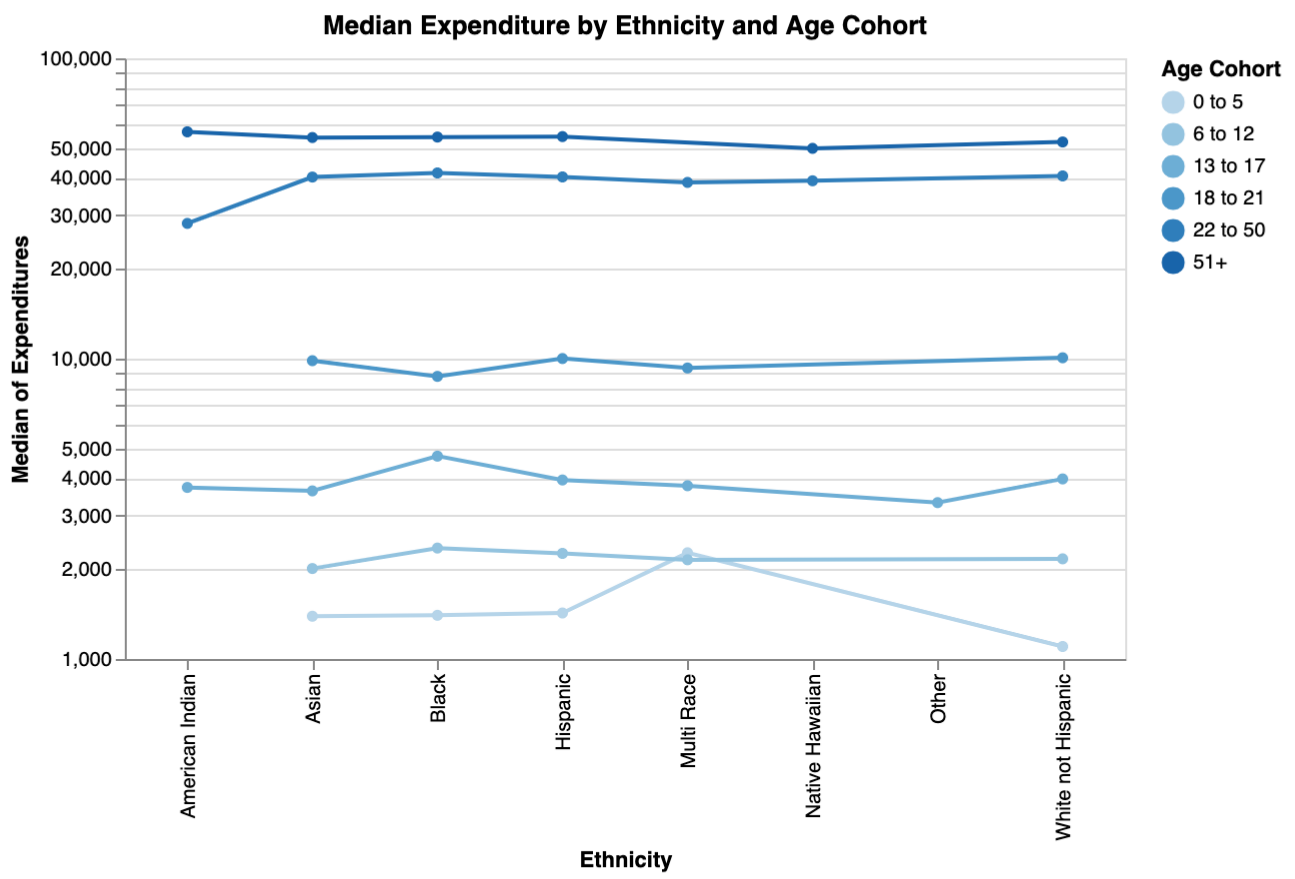
Choose one table or figure from part 1 and one table and figure from part 2 that support your summary of results. Write a caption for each of your choices.

(i) First figure/table.

Plot of median expenditure by Ethnicity by Age Cohort.

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

Out[ ]:

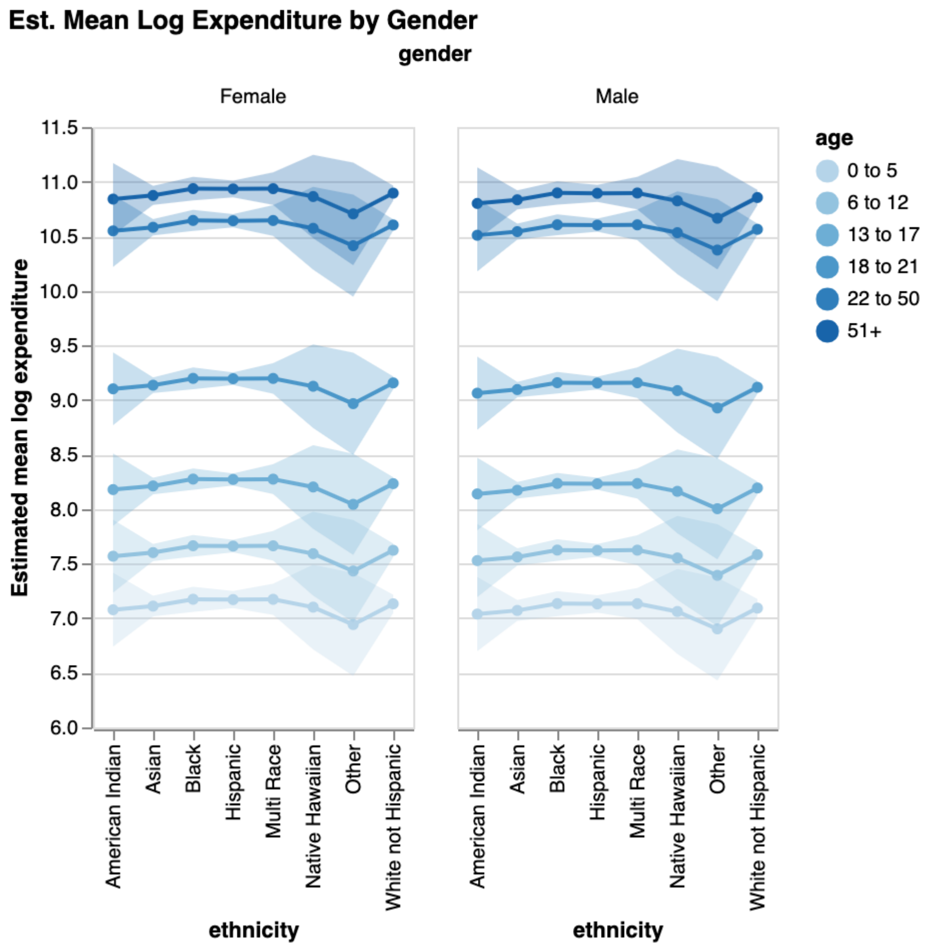


(ii) Second figure/table.

Plot of esitmated means and standard errors (using MLR) of log expendiutre by Ethnicity, Age Cohort, and separated by Gender.

In [ ]:
```
# show figure/table
fig5.facet('gender').properties(title = 'Est. Mean Log Expenditure by Gender')
```

Out[ ]:



In [ ]:

```
grader.check_all()
```

Out[ ]:

```
q2_b_ii results: All test cases passed!
```