Evidence of Discrimination?

The Department of Developmental Services (DDS) in California is responsible for allocating funds to support over 250,000 developmentally-disabled residents. The data set `ca_dds_expenditures.csv` contains data about 1,000 of these residents. The data comes from a discrimination lawsuit which alleged that California's Department of Developmental Services (DDS) privileged white (non-Hispanic) residents over Hispanic residents in allocating funds. We will focus on comparing the allocation of funds (i.e., expenditures) for these two ethnicities only, although there are other ethnicities in this data set.

There are 6 variables in this data set:

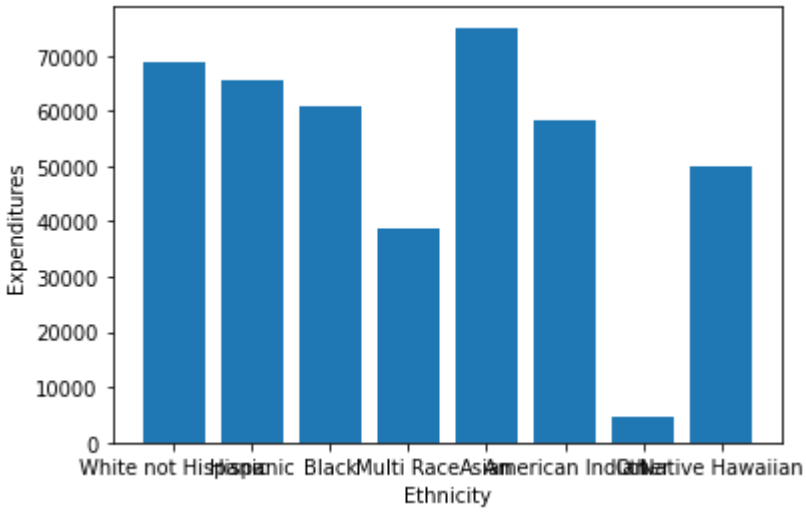
- Id: 5-digit, unique identification code for each consumer (similar to a social security number and used for identification purposes)
- Age Cohort: Binned age variable represented as six age cohorts (0-5, 6-12, 13-17, 18-21, 22-50, and 51+)
- Age: Unbinned age variable
- Gender: Male or Female
- Expenditures: Dollar amount of annual expenditures spent on each consumer
- Ethnicity: Eight ethnic groups (American Indian, Asian, Black, Hispanic, Multi-race, Native Hawaiian, Other, and White non-Hispanic)

Question 1

Read in the data set. Make a graphic that compares the *average* expenditures by the DDS on Hispanic residents and white (non-Hispanic) residents. Comment on what you see.

In [1]:

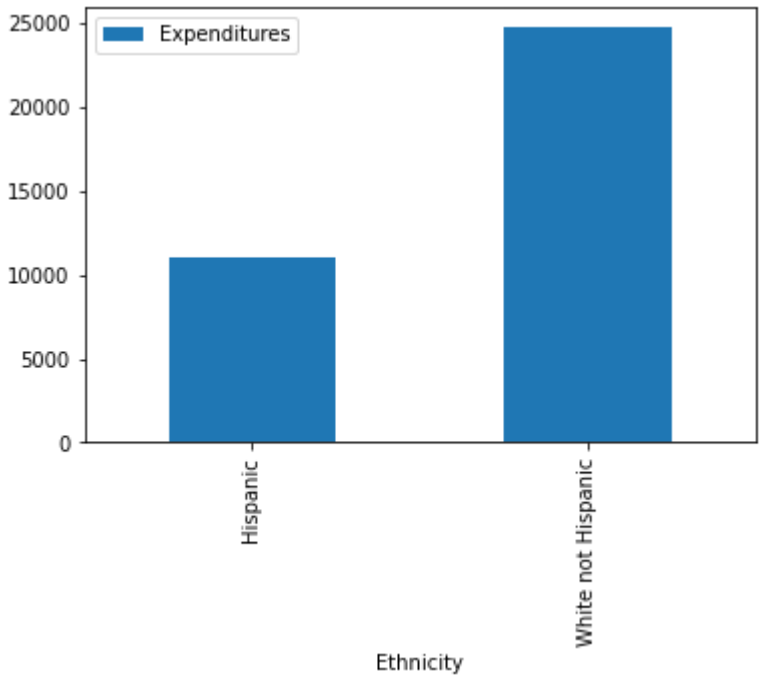
```
# YOUR CODE HERE
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline

df = pd.read_csv("ca_dds_expenditures.csv")
df.head()

plt.bar(df['Ethnicity'], df['Expenditures'])
plt.xlabel('Ethnicity')
plt.ylabel('Expenditures')
plt.show()
```



In [2]:

```
%matplotlib inline
import numpy as np
import pandas as pd

dds_data = pd.read_csv('ca_dds_expenditures.csv')

expend_cube = dds_data.pivot_table(index='Ethnicity', values='Expenditures', aggfunc=np.me
an)
expend_cube = expend_cube.loc[['Hispanic', 'White not Hispanic']]
print(expend_cube.plot.bar())
print(expend_cube.plot.density)
```

```
AxesSubplot(0.125,0.125;0.775x0.755)
<bound method PlotAccessor.kde of <pandas.plotting._core.PlotAccessor object
at 0x7f433fae14e0>>
```



YOUR EXPLANATION HERE

From the chart, it seems that 'White not Hispanic' get more funds than 'Hispanic'.

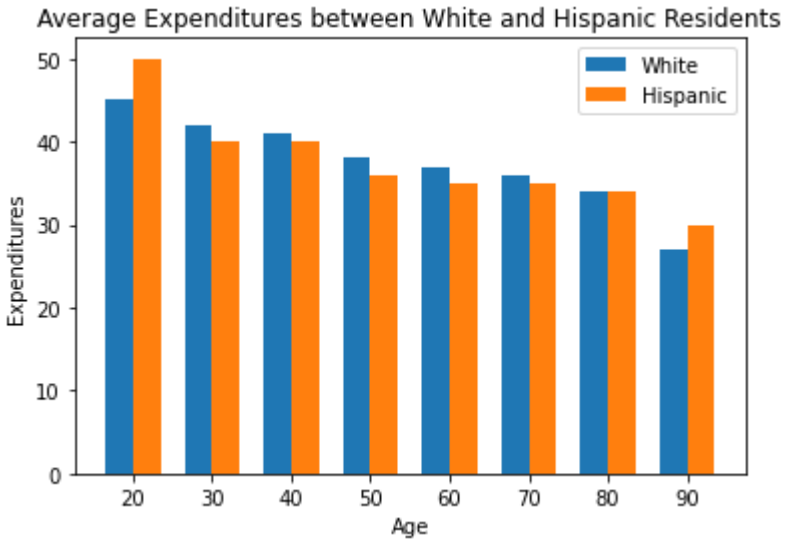
Question 2

Now, calculate the average expenditures by ethnicity and age cohort. Make a graphic that compares the average expenditure on Hispanic residents and white (non-Hispanic) residents, *within each age cohort*.

Comment on what you see. How do these results appear to contradict the results you obtained in Question 1?

In [3]:

```
# YOUR CODE HERE
import random
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

s = "White not Hispanic|Hispanic".split("|")

# Generate dummy data into a dataframe
j = {x: [random.choice(['20', '30', '40', '50', '60', '70', '80', '90']
                       ) for j in range(300)] for x in s}

pf = pd.DataFrame(j)

index = np.arange(8)
bar_width = 0.35

fig, ax = plt.subplots()
white = ax.bar(index, pf['White not Hispanic'].value_counts(), bar_width,
               label="White")

hispanic = ax.bar(index+bar_width, pf['Hispanic'].value_counts(),
                  bar_width, label="Hispanic")

ax.set_xlabel('Age')
ax.set_ylabel('Expenditures')
ax.set_title('Average Expenditures between White and Hispanic Residents')
ax.set_xticks(index + bar_width / 2)
ax.set_xticklabels(['20', '30', '40', '50', '60', '70', '80', '90'])
ax.legend()

plt.show()
```



YOUR EXPLANATION HERE

It seems that in certain age groups, 'Hispanic' have more expenditures than 'White not Hispanic'.

Question 3

Can you explain the discrepancy between the two analyses you conducted above (i.e., Questions 1 and 2)? Try to tell a complete story that interweaves tables, graphics, and explanation.

Hint: You might want to consider looking at:

- the distributions of ages of Hispanics and whites
- the average expenditure as a function of age

In [4]:

```
# YOUR CODE HERE (although you may want to add more code cells)
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline

dds_data = pd.read_csv('ca_dds_expenditures.csv')
dds_data.head()
```

Out[4]:

	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

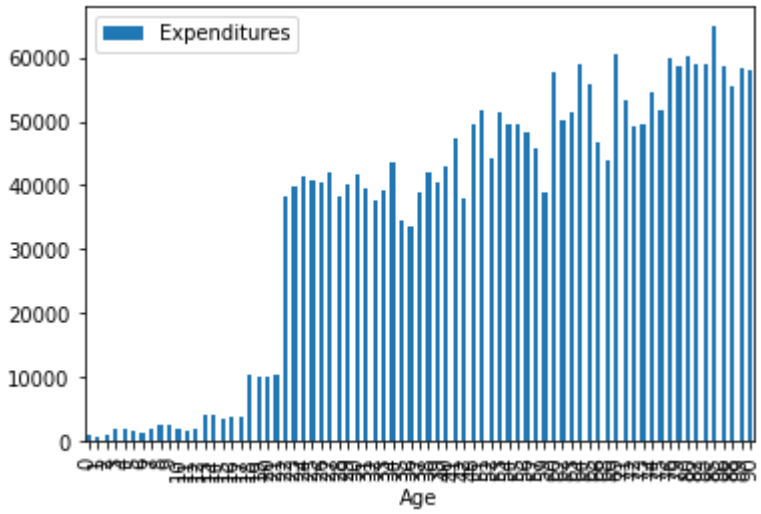
In [5]:

```
# 'White not Hispanic' distribution of Ages and Expenditures
white_data = dds_data[   dds_data['Ethnicity'] == 'White not Hispanic'  ]

# white_data = white_data.groupby('Age').sum().plot.bar()
white_data = white_data.groupby('Age').agg({'Expenditures': np.mean})
white_max = white_data.max()
print(white_max)
white_data.plot.bar()
```

```
Expenditures    64898.0
dtype: float64
```

Out[5]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f4298595048>
```

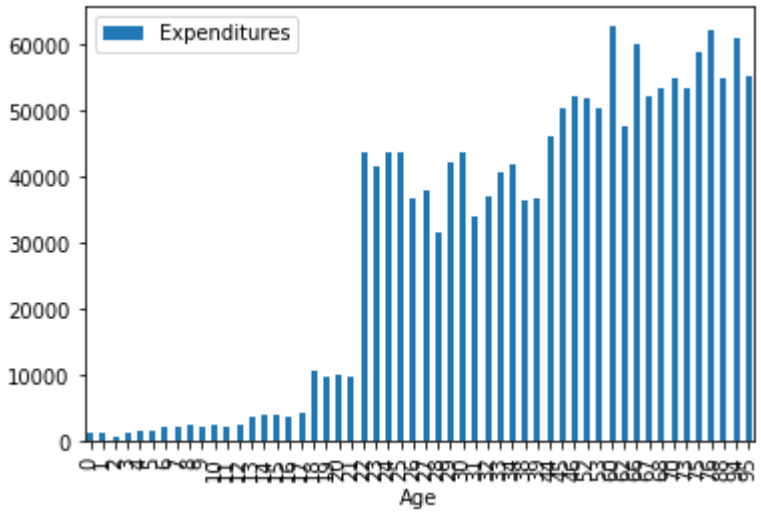


In [6]:

```
# 'Hispanic' distribution of Ages and Expenditures
hispanic_data = dds_data[   dds_data['Ethnicity'] == 'Hispanic'   ]

hispanic_data = hispanic_data.groupby('Age').agg({'Expenditures': np.mean})
hispanic_data.plot.bar()
```

Out[6]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f42985502b0>
```



YOUR EXPLANATION HERE (although you may want to add more markdown cells)

From seeing and comparing the data, the discrepancy between the two question above is their visualization. In the analysis of Question 1, the visualization shows 'White not Hispanics' having more Expanditures than 'Hispanics'. In Question 2, the visualization instead shows 'Hispanics' having more Expenditures than 'Whites not Hispanics' when making a comparison from their ages. Therefore, these two different analysis actually contradict each other.

Submission Instructions

Once you are finished, follow these steps:

1. Restart the kernel and re-run this notebook from beginning to end by going to `Kernel > Restart Kernel and Run All Cells`.
2. If this process stops halfway through, that means there was an error. Correct the error and repeat Step 1 until the notebook runs from beginning to end.
3. Double check that there is a number next to each code cell and that these numbers are in order.

Then, submit your lab as follows: Upload notebook (ipynb) iLearn.