Table of Contents

DSO 553 - Final Group Project - Group 4

Load Data into Notebook

In [27]:
```
import pymongo
import pandas as pd
import numpy as np
import json # Convert CSV to JSON
```

In [6]:
```
client = pymongo.MongoClient(
    'mongodb+srv://nellytian:nellytian@cluster0.z4dp2r1.mongodb.net/?retryWrites=true&w=majority&appName=Cluster0',
    socketTimeoutMS=60000,
    connectTimeoutMS=60000
)
```

In [7]:
```
db = client["grouproject"]

db
```

Out[7]:
```
Database(MongoClient(host=['ac-w1marnp-shard-00-02.z4dp2r1.mongodb.net:27017', 'ac-w1marnp-shard-00-01.z4dp2r1.mongodb.net:27017', 'ac-w1marnp-shard-00-00.z4dp2r1.mongodb.net:27017'], document_class=dict, tz_aware=False, connect=True, retrywrites=True, w='majority', appname='Cluster0', authsource='admin', replicaset='atlas-v1v2az-shard-0', tls=True, sockettimeoutms=60000, connecttimeoutms=60000), 'grouproject')
```

In [8]:
```
collection = db["Discrimination"]
```

In [9]:
```
data = collection.find()
```

In [10]:
```
print(data)
```

```
<pymongo.cursor.Cursor object at 0x7f86b27a3fa0>
```

In [39]:
```
# for row in data:
    # print(row)
```

In [11]:
```
data = collection.find()
```

In [12]:
```
print(data)
```

```
<pymongo.cursor.Cursor object at 0x7f86b2799c40>
```

In [13]:
```
list_cursor = list(data)
```

In [40]:
```
# print(list_cursor)
```

In [15]:
```
# Convert into DataFrame

df = pd.DataFrame(list_cursor)
```

In [16]:
```
df.head()
```

Out[16]:

	_id	Id	Age Cohort	Age	Gender	Expenditures	Ethnicity
0	6612aeb1d21067eea8048fe3	10210	13 to 17	17	Female	2113	White not Hispanic
1	6612aeb1d21067eea8048fe4	10409	22 to 50	37	Male	41924	White not Hispanic
2	6612aeb1d21067eea8048fe5	10486	0 to 5	3	Male	1454	Hispanic
3	6612aeb1d21067eea8048fe6	10538	18 to 21	19	Female	6400	Hispanic
4	6612aeb1d21067eea8048fe7	10568	13 to 17	13	Male	4412	White not Hispanic

In [17]:
```
df.tail()
```

Out[17]:

	_id	Id	Age Cohort	Age	Gender	Expenditures	Ethnicity
995	6612aeb1d21067eea80493c6	99622	51+	86	Female	57055	White not Hispanic
996	6612aeb1d21067eea80493c7	99715	18 to 21	20	Male	7494	Hispanic
997	6612aeb1d21067eea80493c8	99718	13 to 17	17	Female	3673	Multi Race
998	6612aeb1d21067eea80493c9	99791	6 to 12	10	Male	3638	Hispanic
999	6612aeb1d21067eea80493ca	99898	22 to 50	23	Male	26702	White not Hispanic

In [18]:
```
df.shape
```

Out[18]:
```
(1000, 7)
```

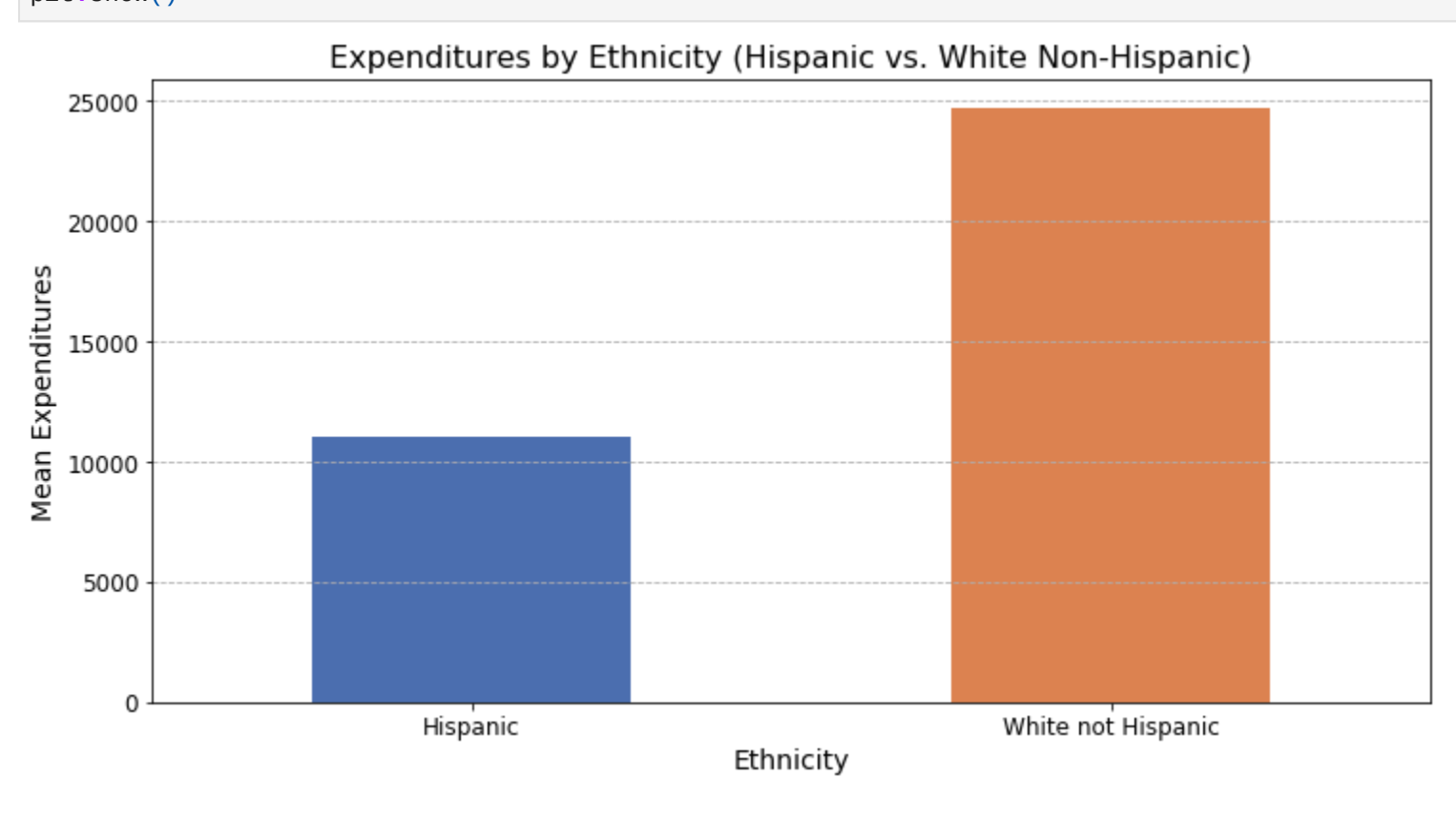
Primary Analysis

In [19]:
```
import matplotlib.pyplot as plt
```

In [20]:
```
# Filter the data to only include Hispanic and White Non-Hispanic
filtered_df = df[df['Ethnicity'].isin(['Hispanic', 'White not Hispanic'])]
```

Average Expenditures by Ethnicity

In [21]:
```
# It would make sense to see the average Expenditures across these two categories
# Group the data by Ethnicity and calculate the mean Expenditures
ethnicity_expenditures = filtered_df.groupby('Ethnicity')['Expenditures'].mean()

# Visualize using a Bar Chart
plt.figure(figsize=(12, 6))
ethnicity_expenditures.plot(kind='bar', color=['#4C72B0', '#DD8452'])
plt.xlabel('Ethnicity', fontsize=14)
plt.ylabel('Mean Expenditures', fontsize=14)
plt.title('Expenditures by Ethnicity (Hispanic vs. White Non-Hispanic)', fontsize=16)
plt.xticks(rotation=0, fontsize=12)
plt.yticks(fontsize=12)
plt.grid(axis='y', linestyle='--')
plt.show()
```



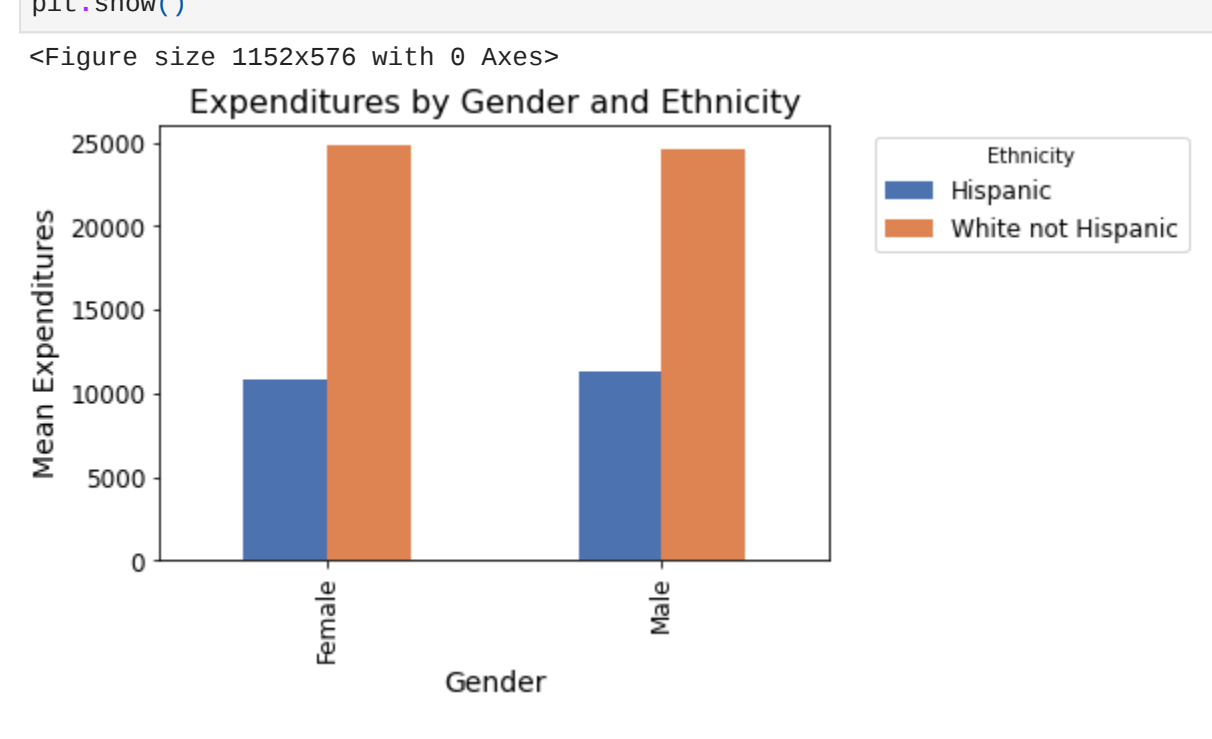
We use the **average expenditures by ethnicity** to compare the differences in allocation of funds between Hispanics and White non Hispanics.

The above chart shows that White not Hispanics receive higher average expenditures (more funding), but we must continue this analysis by looking at the other variables available to us (eg. Gender, Age Cohort).

Average Expenditures by Ethnicity and Gender

In [22]:
```
# Check how Mean Expenditures vary on other Variables

# Female vs. Male, Depending on Ethnicity

# Group the data by Gender and Ethnicity, and calculate the mean Expenditures
expenditures_by_gender_ethnicity = filtered_df.groupby(['Gender', 'Ethnicity'])['Expenditures'].mean().reset_index()

# Create a pivot table
pivot_table = expenditures_by_gender_ethnicity.pivot(index='Gender', columns='Ethnicity', values='Expenditures')

plt.figure(figsize=(16, 8))  # Adjust the figure size as needed
pivot_table.plot(kind='bar', color=['#4C72B0', '#DD8452'])
plt.xlabel('Gender', fontsize=14)
plt.ylabel('Mean Expenditures', fontsize=14)
plt.title('Expenditures by Gender and Ethnicity', fontsize=16)
plt.legend(title='Ethnicity', fontsize=12, bbox_to_anchor=(1.05, 1), loc='upper left')
plt.xticks(fontsize=12)
plt.yticks(fontsize=12)
plt.show()
```

```
<Figure size 1152x576 with 0 Axes>
```

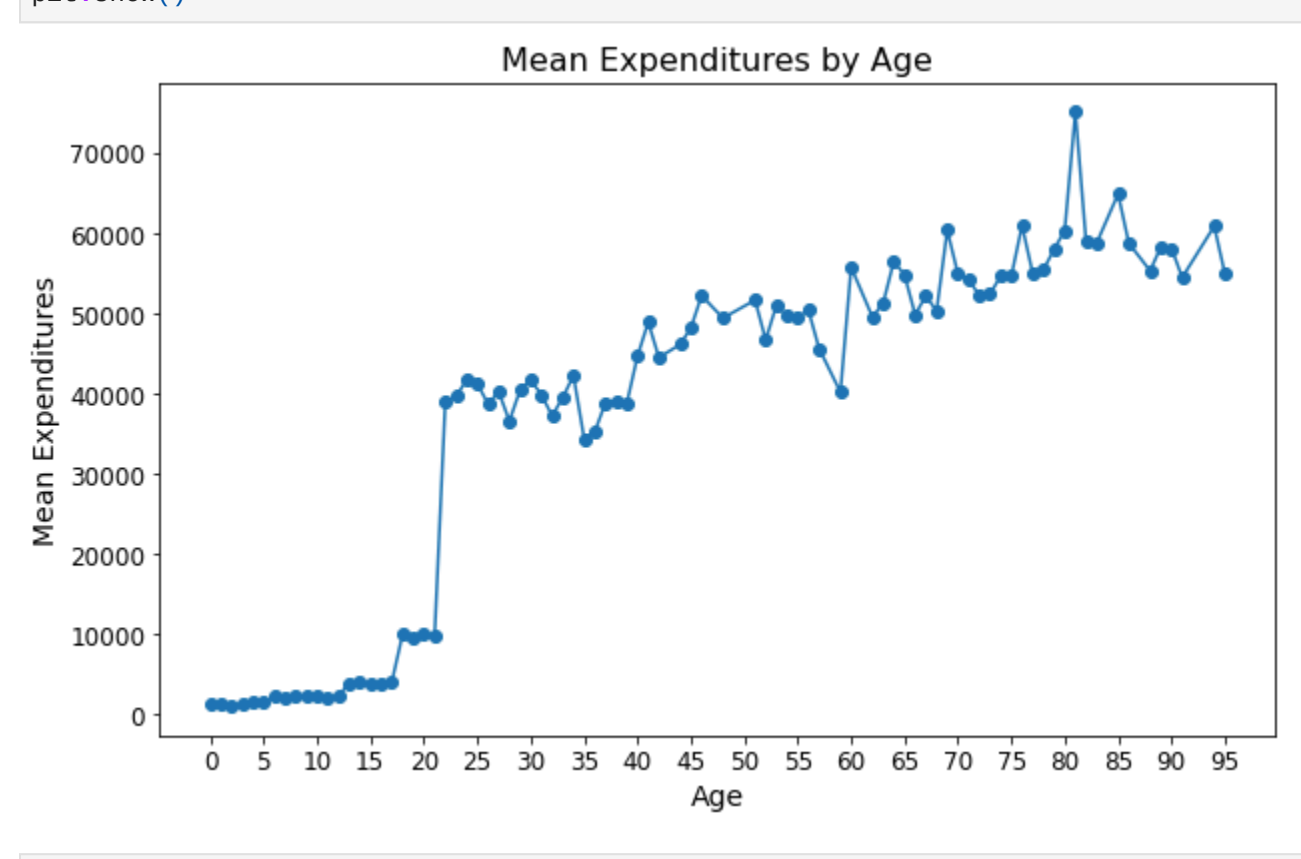


When looking at the average expenditures by gender, it seems that White not Hispanics have higher mean expenditures than do Hispanics, for both female and male.

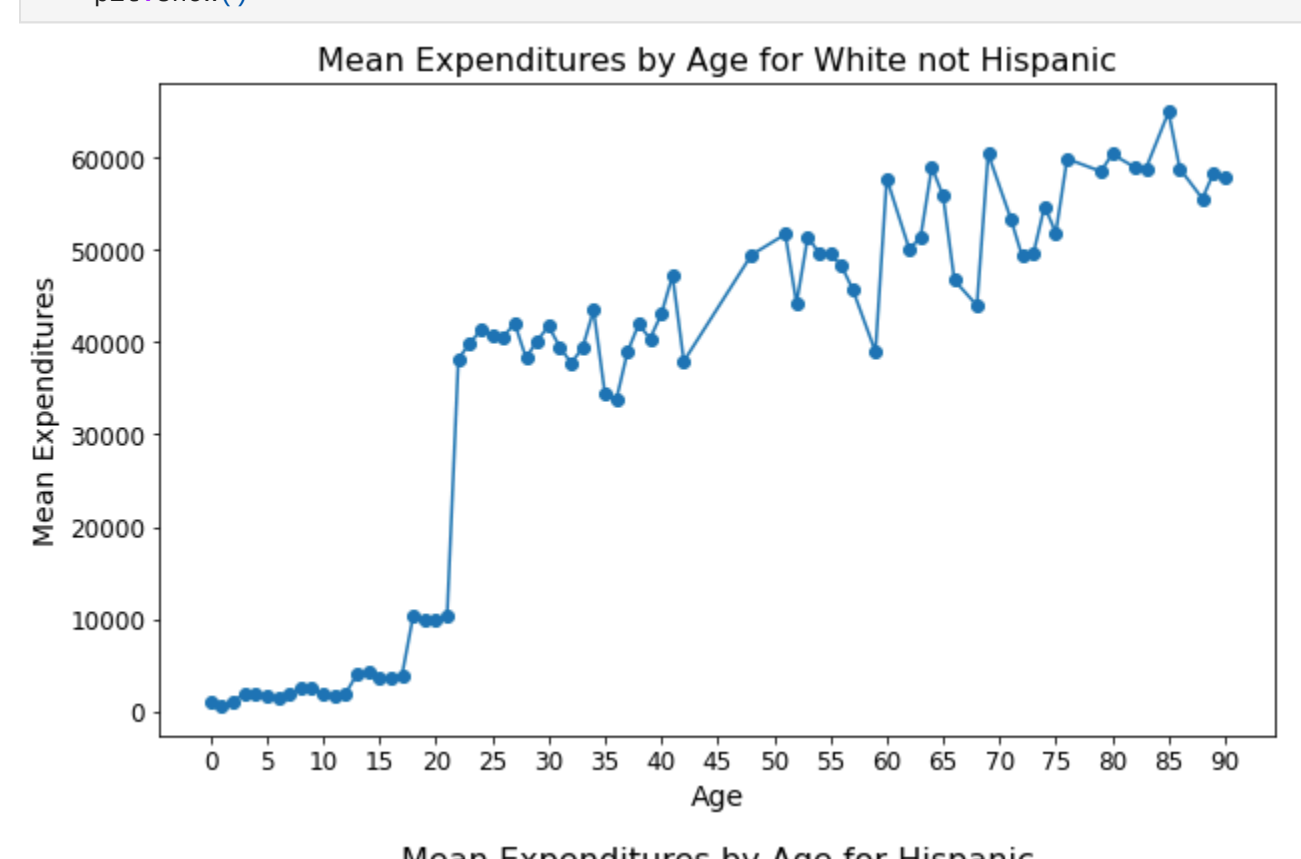
Average Expenditures by Ethnicity and Age Groups

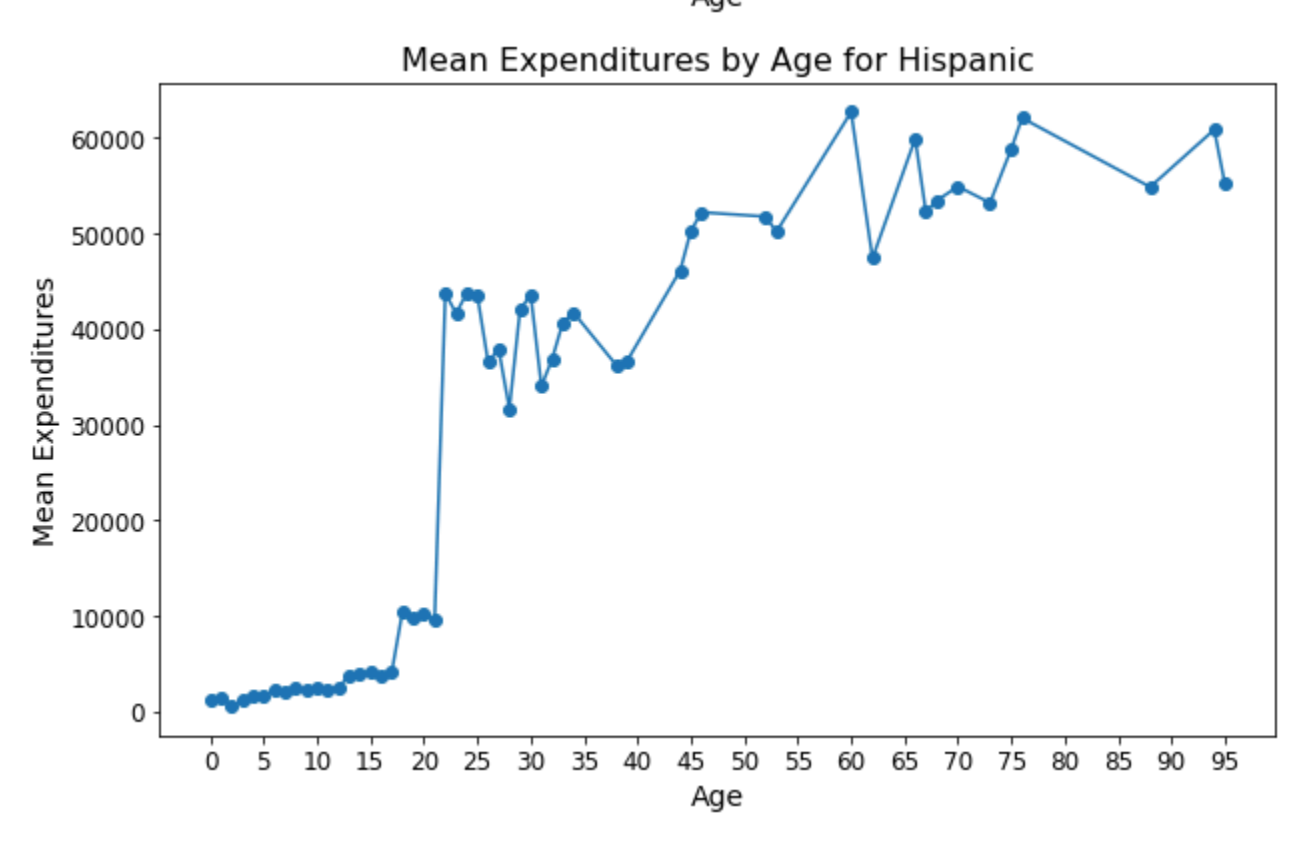
We want to see whether there is the mean expenditures vary depending on the age of an individual, and then also whether this is different across the two ethnic groups in analysis.

In [28]:
```
# Group the data by 'Age' and calculate mean 'Expenditures'
expenditures_by_age = df.groupby('Age')['Expenditures'].mean().reset_index()

# Sort the data by 'Age'
expenditures_by_age = expenditures_by_age.sort_values('Age')

# Create a line chart
plt.figure(figsize=(10, 6))
plt.plot(expenditures_by_age['Age'], expenditures_by_age['Expenditures'], marker='o')
plt.xlabel('Age', fontsize=14)
plt.ylabel('Mean Expenditures', fontsize=14)
plt.title('Mean Expenditures by Age', fontsize=16)

# Set x-axis ticks for every 5 years
min_age = expenditures_by_age['Age'].min()
max_age = expenditures_by_age['Age'].max()
age_ticks = np.arange(min_age, max_age + 1, 5)
plt.xticks(age_ticks, age_ticks, fontsize=12)

plt.yticks(fontsize=12)
plt.show()
```



In [29]:
```
ethnicities = ["White not Hispanic", "Hispanic"]

for ethnicity in ethnicities:
    ethnic_df = df[df['Ethnicity'] == ethnicity]
    expenditures_by_age = ethnic_df.groupby('Age')['Expenditures'].mean().reset_index()
    expenditures_by_age = expenditures_by_age.sort_values('Age')

    plt.figure(figsize=(10, 6))
    plt.plot(expenditures_by_age['Age'], expenditures_by_age['Expenditures'], marker='o')
    plt.xlabel('Age', fontsize=14)
    plt.ylabel('Mean Expenditures', fontsize=14)
    plt.title(f'Mean Expenditures by Age for {ethnicity}', fontsize=16)

    min_age = expenditures_by_age['Age'].min()
    max_age = expenditures_by_age['Age'].max()
    age_ticks = np.arange(min_age, max_age + 1, 5)
    plt.xticks(age_ticks, age_ticks, fontsize=12)

    plt.yticks(fontsize=12)
    plt.show()
```





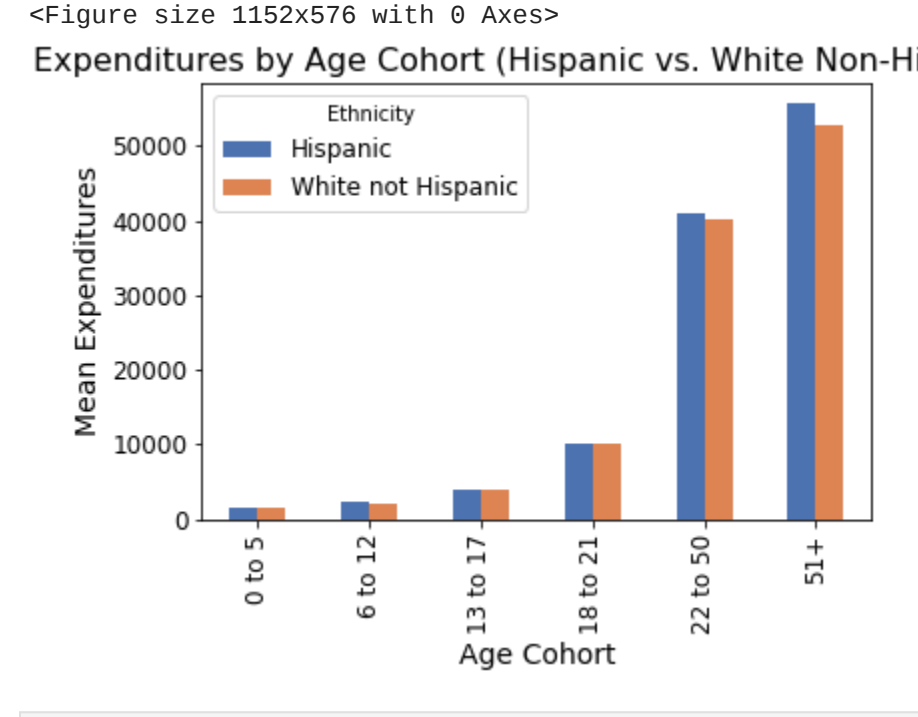
The above graphs show us that for both "White not Hispanic" and "Hispanic" ethnicities, the mean expenditures go up significantly depending on the age of the individual in question. This is evident by the surge in mean expenditures that occurs at around age 20 for both considerations, and increases even more thereafter.

In [26]:
```
# Check the Age Cohorts

distinct_age_cohorts = df['Age Cohort'].unique().tolist()
distinct_age_cohorts
```

Out[26]:
```
['13 to 17', '22 to 50', '0 to 5', '18 to 21', '51+', '6 to 12']
```

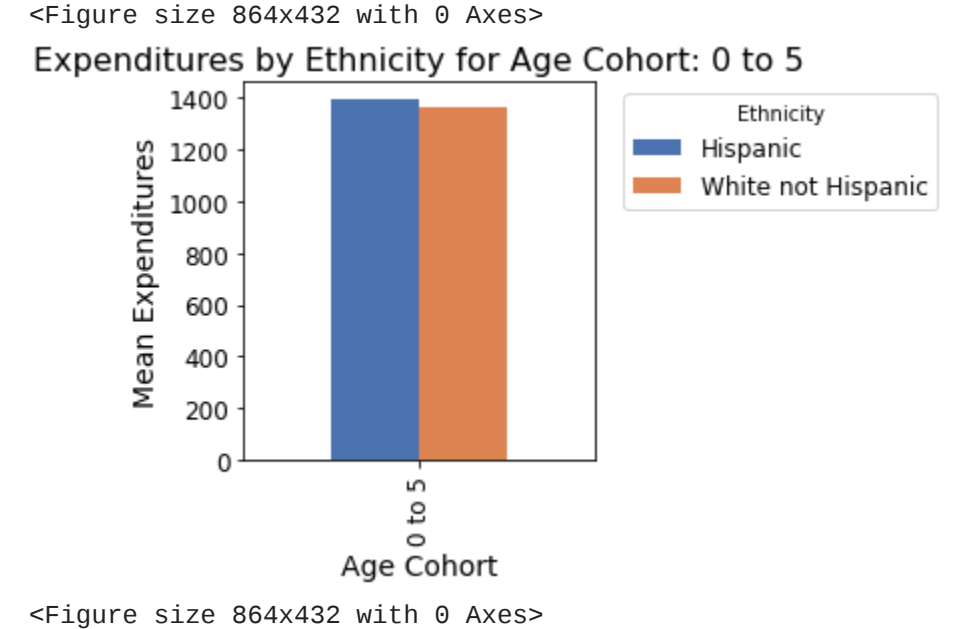
In [27]:
```
# Group the data by 'Age Cohort' and 'Ethnicity'
expenditures_by_age_ethnicity = filtered_df.groupby(['Age Cohort', 'Ethnicity'])['Expenditures'].mean().reset_index()

# Set the order of the 'Age Cohort' values
age_cohort_order = ['0 to 5', '6 to 12', '13 to 17', '18 to 21', '22 to 50', '51+']
expenditures_by_age_ethnicity['Age Cohort'] = pd.Categorical(expenditures_by_age_ethnicity['Age Cohort'], categories=age_cohort_order, ordered=True)

# Visualize Using a Bar Chart
pivot_table = expenditures_by_age_ethnicity.pivot(index='Age Cohort', columns='Ethnicity', values='Expenditures')

plt.figure(figsize=(16, 8))  # Adjust the figure size as needed
pivot_table.plot(kind='bar', color=['#4C72B0', '#DD8452'])
plt.xlabel('Age Cohort', fontsize=14)
plt.ylabel('Mean Expenditures', fontsize=14)
plt.title('Expenditures by Age Cohort (Hispanic vs. White Non-Hispanic)', fontsize=16)
plt.legend(title='Ethnicity', fontsize=12)
plt.xticks(fontsize=12)
plt.yticks(fontsize=12)
plt.show()
```

```
<Figure size 1152x576 with 0 Axes>
```

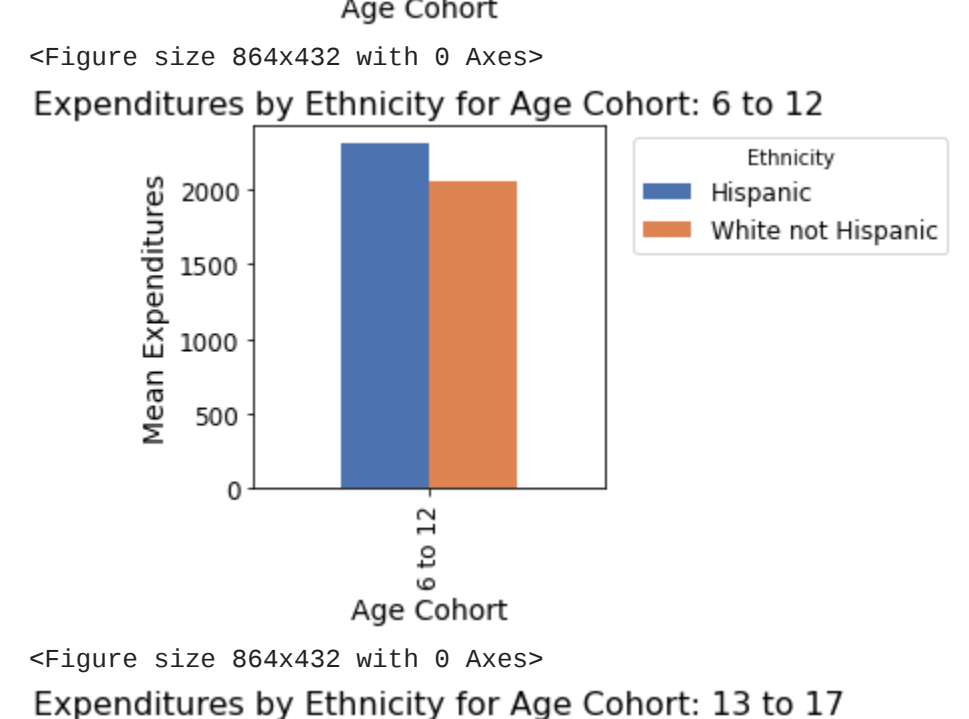


In [120…
```
# Create a separate chart for each Age Cohort
for age_cohort in age_cohort_order:
    age_cohort_data = expenditures_by_age_ethnicity[expenditures_by_age_ethnicity['Age Cohort'] == age_cohort]
    pivot_table = age_cohort_data.pivot(index='Age Cohort', columns='Ethnicity', values='Expenditures')

    plt.figure(figsize=(12, 6))  # Adjust the figure size as needed
    ax = pivot_table.plot(kind='bar', color=['#4C72B0', '#DD8452'])
    plt.xlabel('Age Cohort', fontsize=14)
    plt.ylabel('Mean Expenditures', fontsize=14)
    plt.title(f'Expenditures by Ethnicity for Age Cohort: {age_cohort}', fontsize=16)
    plt.legend(title='Ethnicity', bbox_to_anchor=(1.05, 1), loc='upper left', fontsize=12)
    plt.xticks(fontsize=12)
    plt.yticks(fontsize=12)
    plt.tight_layout()
    plt.show()
```

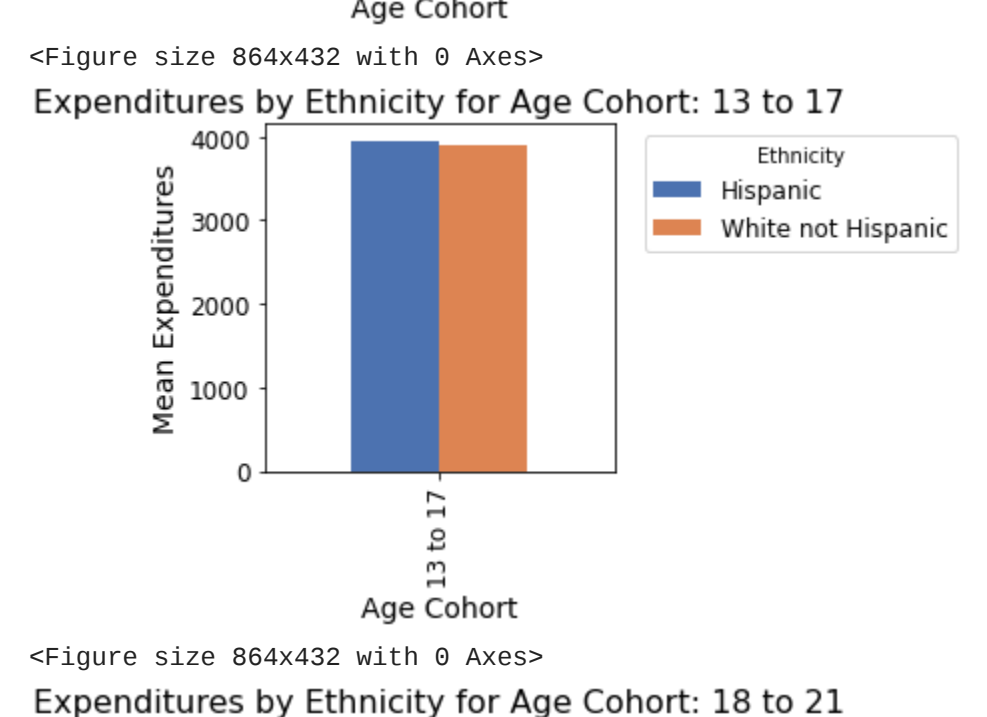
```
<Figure size 864x432 with 0 Axes>
```



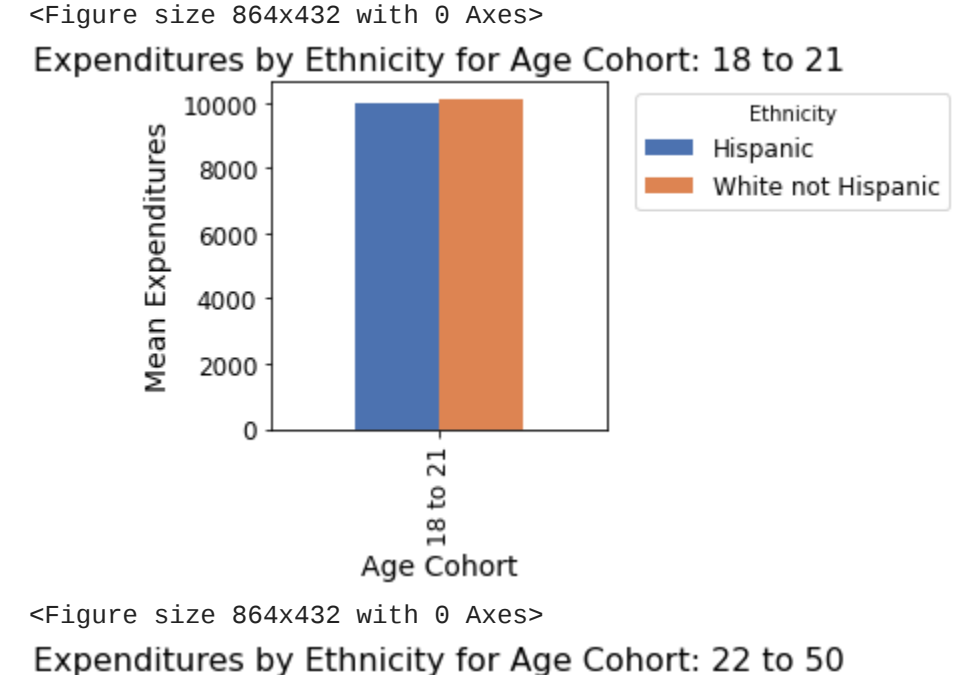

```
<Figure size 864x432 with 0 Axes>
```



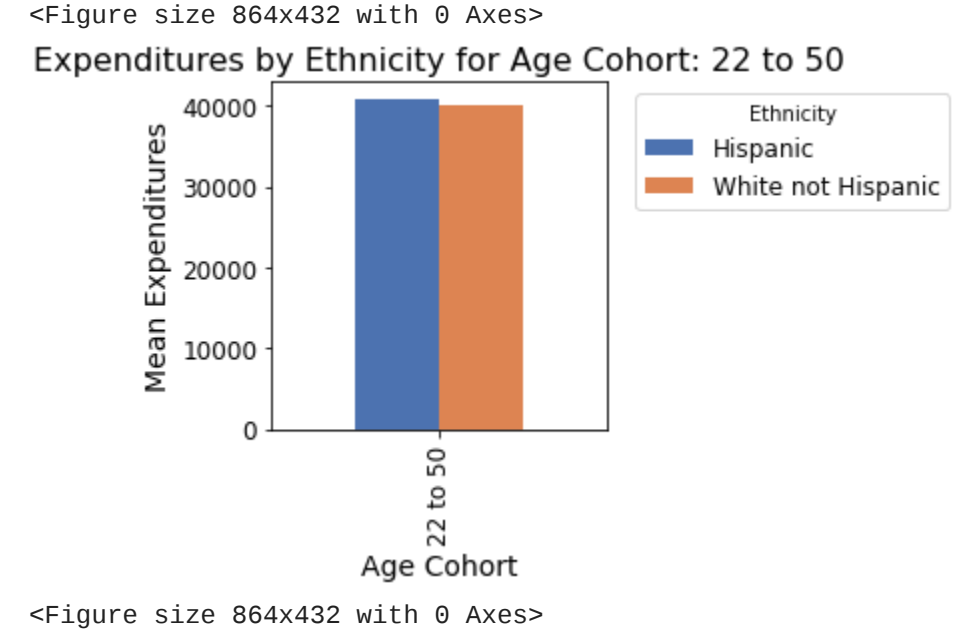

```
<Figure size 864x432 with 0 Axes>
```



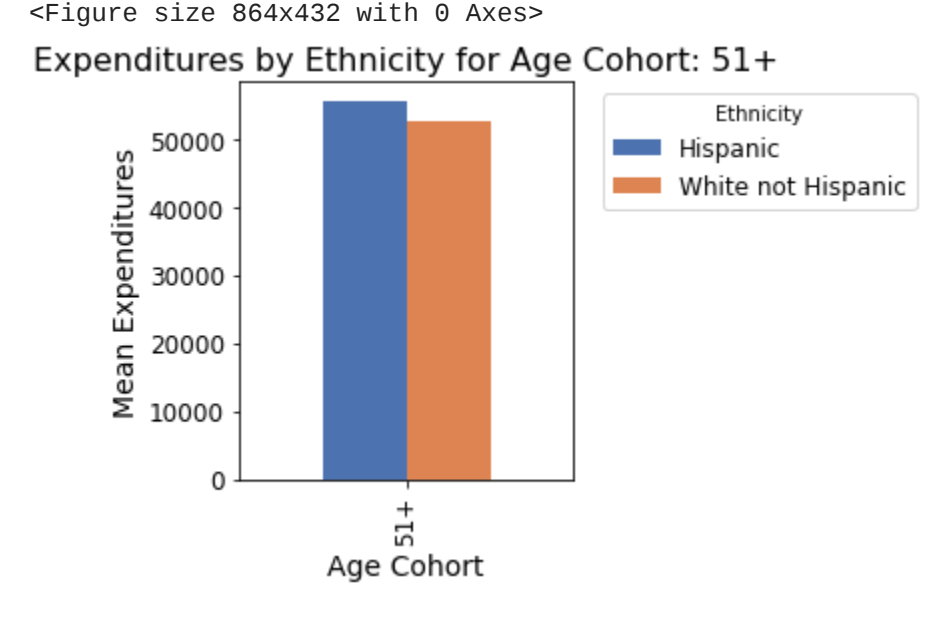

```
<Figure size 864x432 with 0 Axes>
```




```
<Figure size 864x432 with 0 Axes>
```



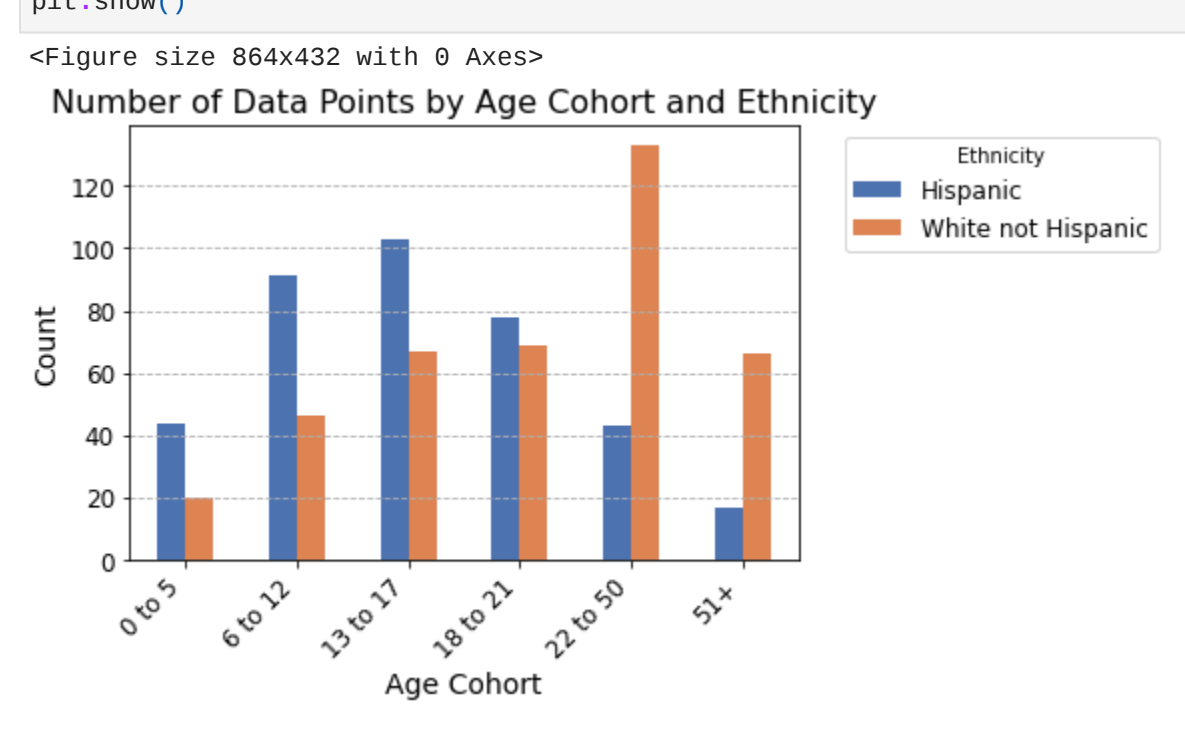

```
<Figure size 864x432 with 0 Axes>
```



The above graphs show that when looking at the average expenditures for each age cohorts, Hispanics actually have *higher* expenditures than do White not Hispanics. This contradicts the initial claim.

In [36]:
```
# Count the number of data points in each age cohort for each ethnicity
age_cohort_counts = filtered_df.groupby(['Ethnicity', 'Age Cohort']).size().reset_index(name='Count')

# Set the order of the 'Age Cohort' values
age_cohort_order = ['0 to 5', '6 to 12', '13 to 17', '18 to 21', '22 to 50', '51+']
age_cohort_counts['Age Cohort'] = pd.Categorical(age_cohort_counts['Age Cohort'], categories=age_cohort_order, ordered=True)

# Create a pivot table
pivot_table = age_cohort_counts.pivot(index='Age Cohort', columns='Ethnicity', values='Count')

# Create the plot
plt.figure(figsize=(12, 6))
pivot_table.plot(kind='bar', color=['#4C72B0', '#DD8452'])
plt.xlabel('Age Cohort', fontsize=14)
plt.ylabel('Count', fontsize=14)
plt.title('Number of Data Points by Age Cohort and Ethnicity', fontsize=16)
plt.legend(title='Ethnicity', fontsize=12, bbox_to_anchor=(1.05, 1), loc='upper left')
plt.xticks(rotation=45, ha='right', fontsize=12)
plt.yticks(fontsize=12)
plt.grid(axis='y', linestyle='--')
plt.show()
```

```
<Figure size 864x432 with 0 Axes>
```



From the above, we see that there are a lot more Hispanics who fall in younger age cohorts, which have much lower mean expenditures than older age groups.

This explains why at first glance, there seems to be a higher average expenditure for White not Hispanic populations. In reality, this is only due to the greater number of White not Hispanics who are in older age categories, thus who incur higher expenditures in the first place.

Mongo Aggregation Pipeline Output: Comparative Analysis

In [126…
```
pipeline = [
    {
        '$match': {
            'Ethnicity': {'$in': ['White not Hispanic', 'Hispanic']}
        }
    },
    {
        '$group': {
            '_id': {
                'Ethnicity': '$Ethnicity',
                'Age Cohort': '$Age Cohort'
            },
            'mean_expenditures': {
                '$avg': '$Expenditures'
            }
        }
    },
    {
        '$project': {
            'Ethnicity': '$_id.Ethnicity',
            'Age Cohort': '$_id.Age Cohort',
            'mean_expenditures': 1,
            '_id': 0
        }
    },
    {
        '$sort': {
            'Age Cohort': 1
        }
    }
]

# Execute the aggregation pipeline and retrieve the results
results = list(collection.aggregate(pipeline))

# Print the results
for result in results:
    print(result)
```

```
{'mean_expenditures': 1393.2045454545455, 'Ethnicity': 'Hispanic', 'Age Cohort': '0 to 5'}
{'mean_expenditures': 1366.9, 'Ethnicity': 'White not Hispanic', 'Age Cohort': '0 to 5'}
{'mean_expenditures': 3904.3582089552237, 'Ethnicity': 'White not Hispanic', 'Age Cohort': '13 to 17'}
{'mean_expenditures': 3955.2815533980583, 'Ethnicity': 'Hispanic', 'Age Cohort': '13 to 17'}
{'mean_expenditures': 10133.057971014492, 'Ethnicity': 'White not Hispanic', 'Age Cohort': '18 to 21'}
{'mean_expenditures': 9959.846153846154, 'Ethnicity': 'Hispanic', 'Age Cohort': '18 to 21'}
{'mean_expenditures': 40924.11627906977, 'Ethnicity': 'Hispanic', 'Age Cohort': '22 to 50'}
{'mean_expenditures': 40187.624060150374, 'Ethnicity': 'White not Hispanic', 'Age Cohort': '22 to 50'}
{'mean_expenditures': 52670.42424242424, 'Ethnicity': 'White not Hispanic', 'Age Cohort': '51+'}
{'mean_expenditures': 55585.0, 'Ethnicity': 'Hispanic', 'Age Cohort': '51+'}
{'mean_expenditures': 2052.2608695652175, 'Ethnicity': 'White not Hispanic', 'Age Cohort': '6 to 12'}
{'mean_expenditures': 2312.186813186813, 'Ethnicity': 'Hispanic', 'Age Cohort': '6 to 12'}
```

In [37]:
```
pipeline = [
    {
        '$match': {
            'Ethnicity': {'$in': ['White not Hispanic', 'Hispanic']}
        }
    },
    {
        '$group': {
            '_id': {
                'Ethnicity': '$Ethnicity',
                'Age Cohort': '$Age Cohort'
            },
            'mean_expenditures': {
                '$avg': '$Expenditures'
            },
            'count': {
                '$count': {}
            }
        }
    },
    {
        '$project': {
            'Ethnicity': '$_id.Ethnicity',
            'Age Cohort': '$_id.Age Cohort',
            'mean_expenditures': 1,
            'count': 1,
            '_id': 0
        }
    },
    {
        '$sort': {
            'Age Cohort': 1
        }
    }
]

# Execute the aggregation pipeline and retrieve the results
results = list(collection.aggregate(pipeline))

# Print the results
for result in results:
    print(result)
```

```
{'mean_expenditures': 1366.9, 'count': 20, 'Ethnicity': 'White not Hispanic', 'Age Cohort': '0 to 5'}
{'mean_expenditures': 1393.2045454545455, 'count': 44, 'Ethnicity': 'Hispanic', 'Age Cohort': '0 to 5'}
{'mean_expenditures': 3955.2815533980583, 'count': 103, 'Ethnicity': 'Hispanic', 'Age Cohort': '13 to 17'}
{'mean_expenditures': 3904.3582089552237, 'count': 67, 'Ethnicity': 'White not Hispanic', 'Age Cohort': '13 to 17'}
{'mean_expenditures': 9959.846153846154, 'count': 78, 'Ethnicity': 'Hispanic', 'Age Cohort': '18 to 21'}
{'mean_expenditures': 10133.057971014492, 'count': 69, 'Ethnicity': 'White not Hispanic', 'Age Cohort': '18 to 21'}
{'mean_expenditures': 40187.624060150374, 'count': 133, 'Ethnicity': 'White not Hispanic', 'Age Cohort': '22 to 50'}
{'mean_expenditures': 40924.11627906977, 'count': 43, 'Ethnicity': 'Hispanic', 'Age Cohort': '22 to 50'}
{'mean_expenditures': 52670.42424242424, 'count': 66, 'Ethnicity': 'White not Hispanic', 'Age Cohort': '51+'}
{'mean_expenditures': 55585.0, 'count': 17, 'Ethnicity': 'Hispanic', 'Age Cohort': '51+'}
{'mean_expenditures': 2052.2608695652175, 'count': 46, 'Ethnicity': 'White not Hispanic', 'Age Cohort': '6 to 12'}
{'mean_expenditures': 2312.186813186813, 'count': 91, 'Ethnicity': 'Hispanic', 'Age Cohort': '6 to 12'}
```

Key Insights

- When we simply look at the average expenditures across the Hispanic vs. White not Hispanic ethnicities, without conduting any bivariate analysis, we see that the average funding provided to White not Hispanic is much greater.
- The same results hold when we add in 'Gender' as a variable into the analysis and compare across the following four groups:

1. Female Hispanic
2. Female White not Hispanic
3. Male Hispanic
4. Male White not Hispanic

- However, when we add **'Age Cohort'** into the mix, we find that **there are no significant differences in average expenditures across the Hispanic vs. White not Hispanic ethnicities**. *In fact, across 5 of the 6 cohorts, Hispanics actually seem to have a higher average funding than their White not Hispanic counterparts.*
- We also realize that as the age of the individual in consideration plays a role in their mean expenditures. As a person ages, their expenditures increase.
- To provide further explanation to this, we explore the number of data points that are collected for each age cohort, across Hispanic vs. White not Hispanic ethnicities. We see that there is a much higher number of White not Hispanic data points for the older Age Cohorts, which usually incur higher average expenditures in the first place (as explained by the prior bullet).
- As such, we conclude that **there is no significance difference between the allocation of funds between Hispanics and White non Hispanics**, as per our analysis across the different age cohorts. Specifically, when expenditures are categorized by **age**, the initial assumption of discrimination fails to hold.