Assignment 3.2:

Question 1 and Question 4:

In this homework Assignment I kept the first 6 variables that I was originally using. These were Total, white, black, Asian, bachelor population and then the median household income based on tracts in San Diego. Ideally I was hoping to study the Hispanic population, but was having trouble finding this in the census data. This would have been more insightful I feel in comparison to other races because there are larger percents of Hispanics in San Diego.

Since this was the case, I decided to take a new approach. I wanted to test if the rich really got richer. To do so I had to pull data from 2017 and 2019 and then also calculate their differences. I then used these to create data frames with percents, and these are what I used to cluster.

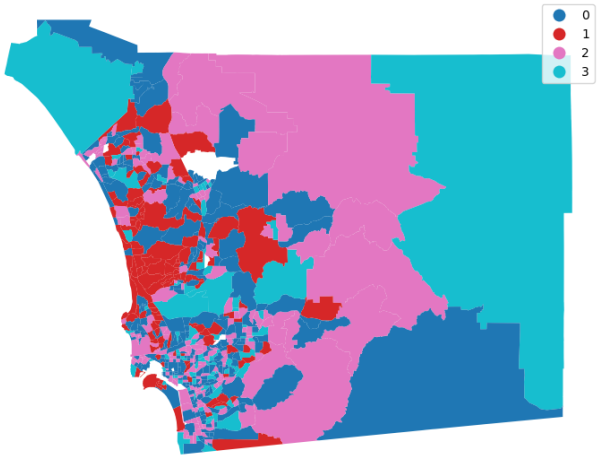
My question was not as answered by the clustering as much, but I discuss how my question was answered. The clustering did indeed tell me that the clusters were mostly weighted on the median household income change over the two years (i.e. tracts with similar changes in household income were grouped together in clusters). To find the answer to my question I instead looked at the values of change as binary, either 0 it decreased or 1 it increased. Then I compared those that increased and those that decreased with respect to their median household income. I noticed that 97 of 614 tracts increased in median household income. Of those 97, 91 live in the top half (307) of the tracts ordered from highest household income to lowest household. Also, in the top 88 tracts for median household income, we notice that half of those that showed an increase in household income. This is very disproportionately distributed over the tracts. This leads us to believe that it is indeed true that the rich get richer.

I used cluster sizes of 3 and 4.

Things to note from the pdf distributions are that in the clusters of 3 variables is that the tract 1 with the largest decrease in household income, had the most dense population with the peak with the greatest increase in population with bachelors degree. This is maybe unexpected, but also I could look more into this.

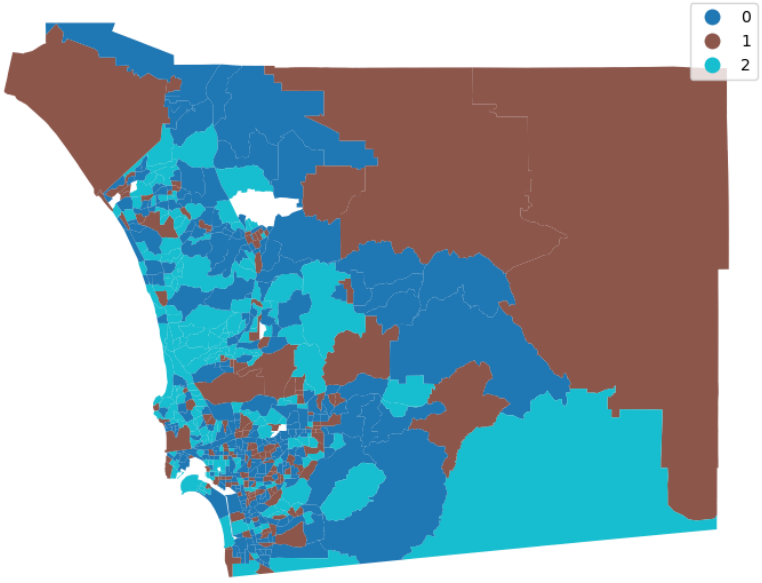
```
]:
f, ax = plt.subplots(1, figsize=(9,9))
df_for_clustering.plot('labels4', categorical = True, legend = True, linewidth = 0, ax = ax)
plt.title('KMeans with 4 clusters')
ax.set_axis_off()
plt.show()
```

KMeans with 4 clusters

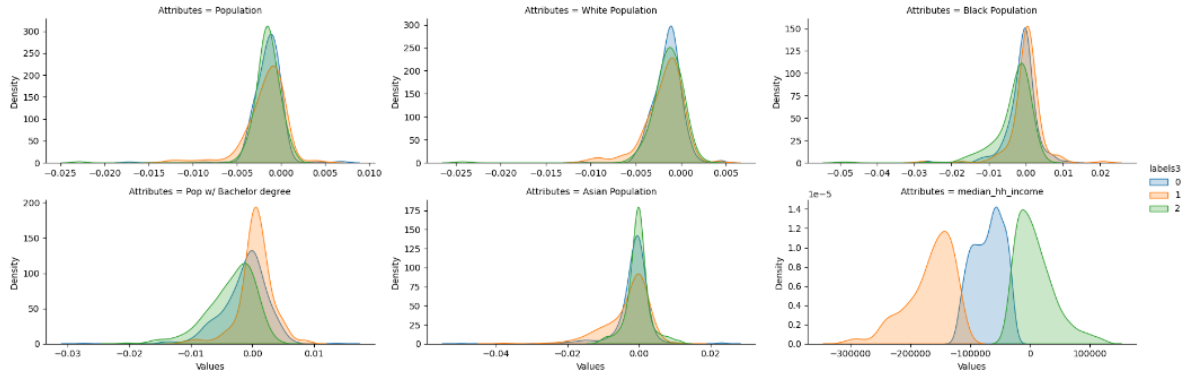



```
f, ax = plt.subplots(1, figsize=(9,9))
df_for_clustering.plot('labels3', categorical = True, legend = True, linewidth = 0, ax = ax)
plt.title('KMeans with 3 clusters')
ax.set_axis_off()
plt.show()
```

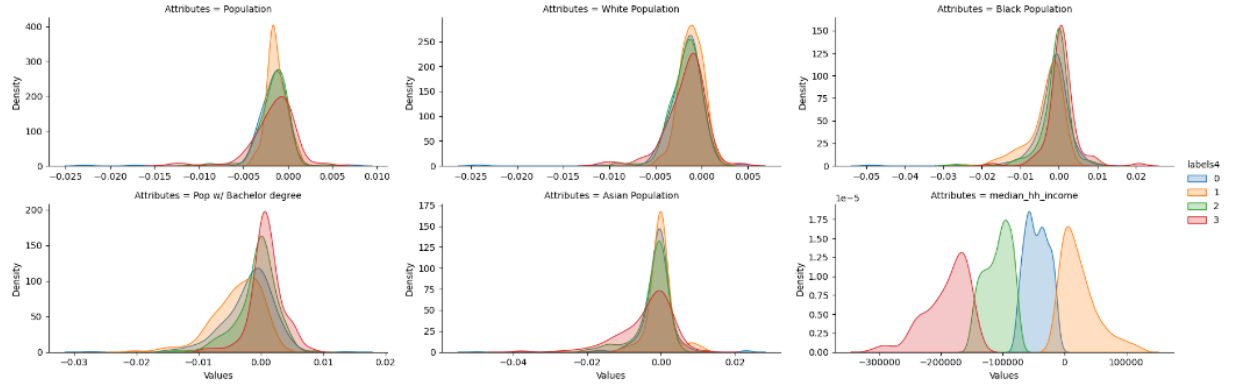
KMeans with 3 clusters



```
facets = sns.FacetGrid(data = tidy3, col = 'Attributes', hue = 'labels3', sharey=False, sharex=False, aspect=2, col_wrap=
_ = facets.map(sns.kdeplot, 'Values', shade = True).add_legend()
plt.show()
```




```
facets = sns.FacetGrid(data = tidy4, col = 'Attributes', hue = 'labels4', sharey=False, sharex=False, aspect=2, col_wrap=
_ = facets.map(sns.kdeplot, 'Values', shade = True).add_legend()
plt.show()
```



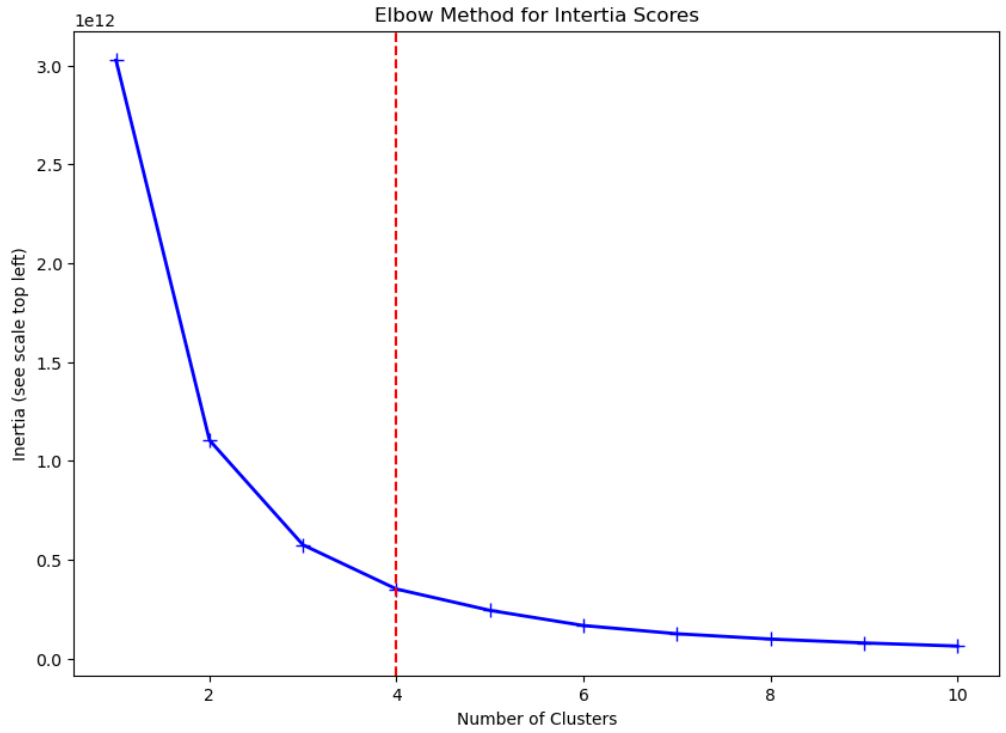
Question 2

In question two we were asked to use the elbow method to understand what number of clusters will be the best. Attached below is the elbow curve and we see that 4 is likely the optimal number of clusters. I ended up choosing 3 and 4, however upon further consideration I think choosing a value such at 6 could have created value

```
8]:
_ = plt.figure(figsize = (10,7))
_ = plt.plot(range(1,11), inertia_scores, linewidth = 2, color = 'blue', marker = '+', markersize = 8)
_ = plt.title('Elbow Method for Intertia Scores')
_ = plt.xlabel('Number of Clusters')
_ = plt.ylabel('Inertia (see scale top left)')

num_clusters = 4

_ = plt.axvline(x = num_clusters, color = 'red', linestyle = '--')
_ = plt.show()
```



Question 3:

I will display the report of both the clusters I chose to do. These are the tables with the mean values of each variable based on the cluster. Remember that these will be changes in percents so they are very small. Therefor it is more meaningful to just look at the sign associated with them. The negative sign means that the percentage in the tract fell over the 2 year period from 2017-2019. I will also show the histograms for the areas.

```
k3means = df_for_clustering.groupby('labels3').mean()
k3means.T
```

labels3	0	1	2
Population	-0.001521	-0.001806	-0.001597
White Population	-0.001542	-0.001916	-0.001479
Black Population	-0.001173	0.000040	-0.003417
Asian Population	-0.001643	-0.003354	-0.000340
Pop w/ Bachelor degree	-0.001321	0.000520	-0.003541
median_hh_income	-70253.310345	-167483.014815	8801.280423
area_sqkm	10.758022	40.412079	12.582975

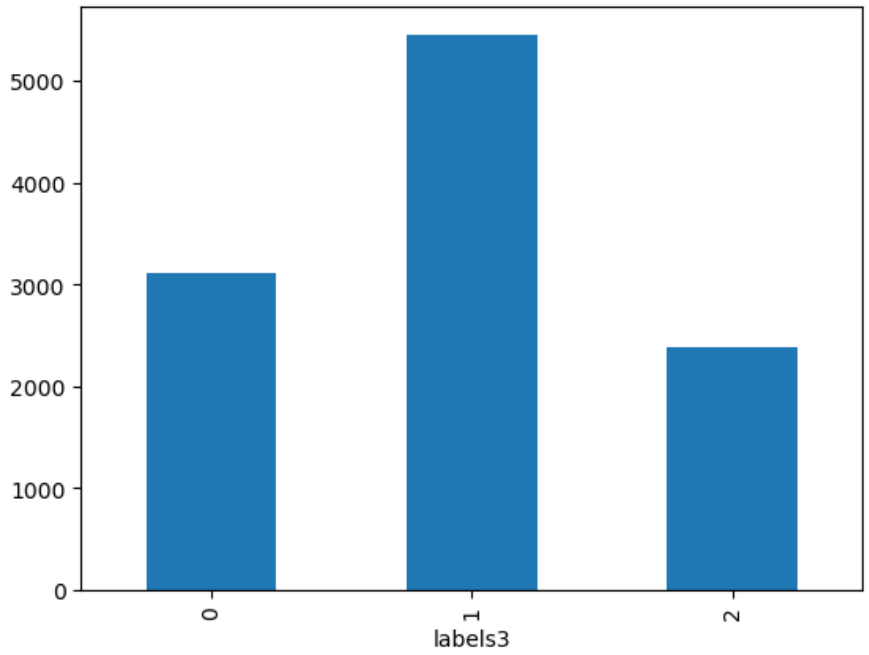
```
k4means = df_for_clustering.groupby('labels4').mean()
k4means.T
```

labels4	0	1	2	3
Population	-0.001679	-0.001311	-0.001687	-0.001697
White Population	-0.001735	-0.001099	-0.001732	-0.001751
Black Population	-0.001769	-0.003532	-0.000981	0.000687
Asian Population	-0.001046	-0.000124	-0.002402	-0.004023
Pop w/ Bachelor degree	-0.001907	-0.003870	-0.000625	0.000847
median_hh_income	-44652.093220	23584.157480	-108569.447674	-191184.481013
area_sqkm	11.608473	7.140872	20.590738	47.665089

```
areas3 = df_for_clustering.dissolve(by = 'labels3', aggfunc = 'sum')['area_sqkm']
areas3
```

```
labels3
0    3119.826470
1    5455.630702
2    2378.182202
Name: area_sqkm, dtype: float64
```

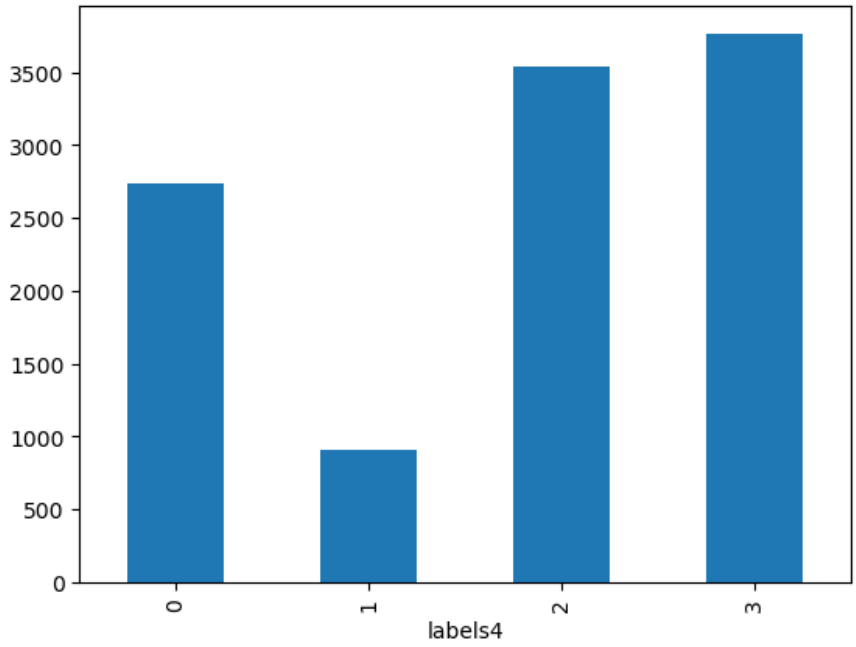
```
areas3.plot.bar()
plt.show()
```



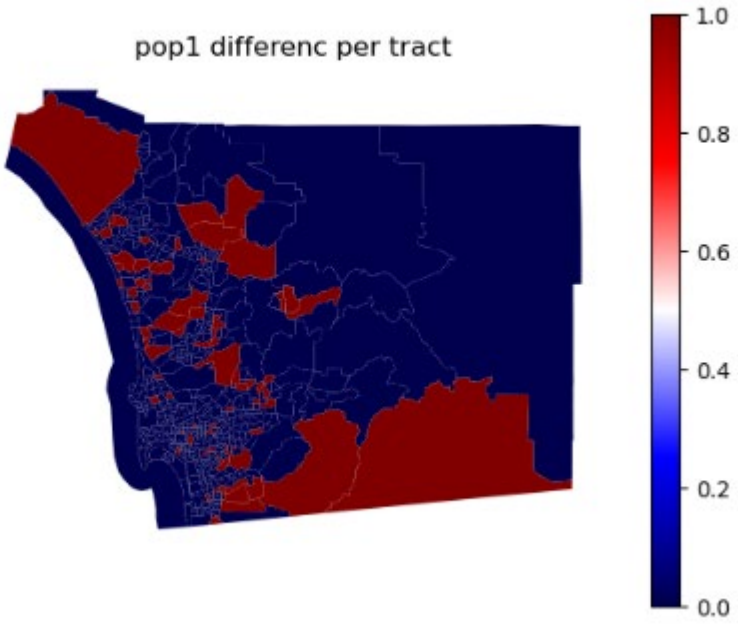
```
areas4 = df_for_clustering.dissolve(by = 'labels4', aggfunc = 'sum')['area_sqkm']
areas4
```

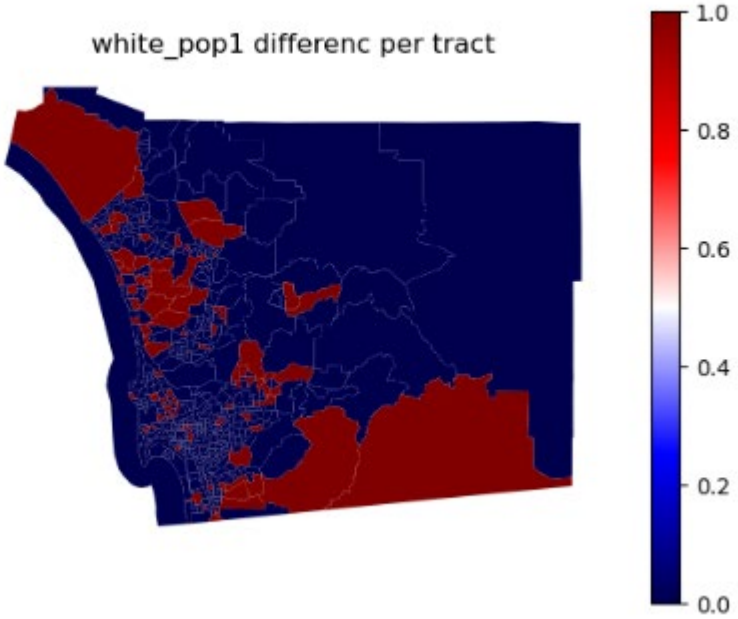
```
labels4
0    2739.599551
1     906.890779
2    3541.607021
3    3765.542023
Name: area_sqkm, dtype: float64
```

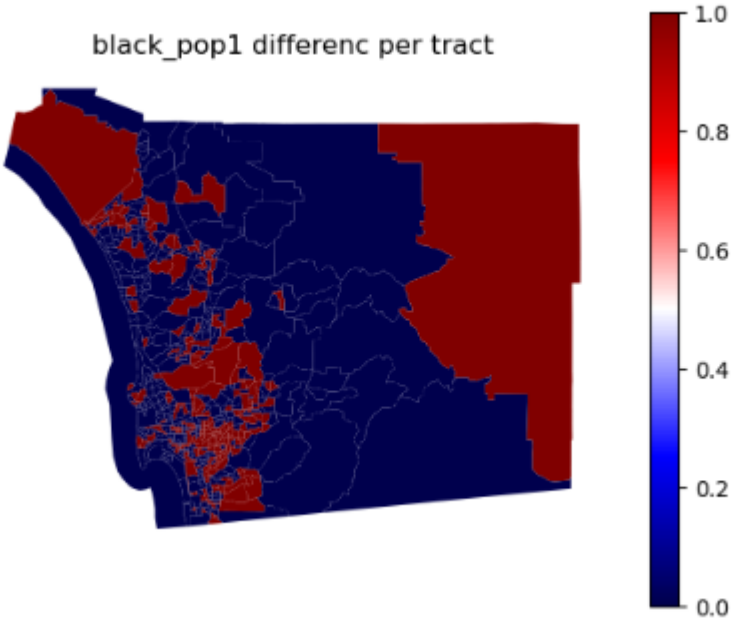
```
areas4.plot.bar()
plt.show()
```

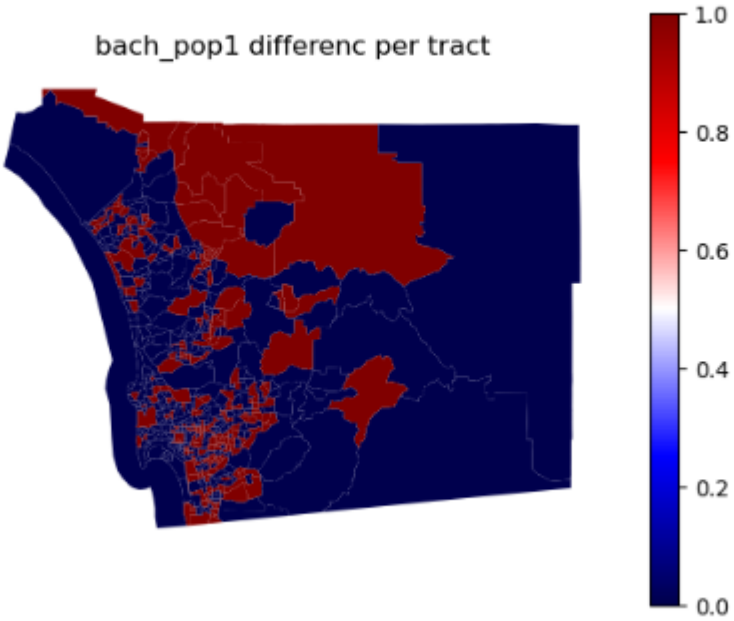


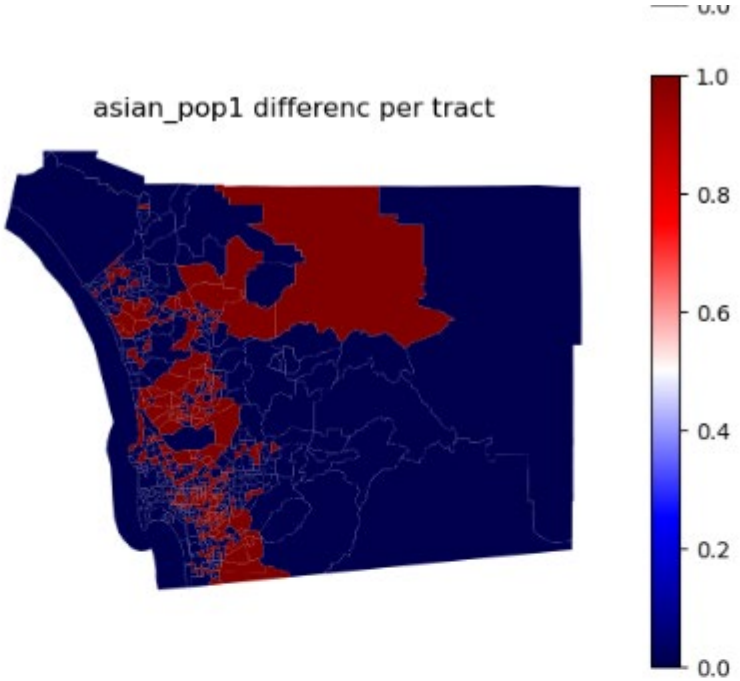
For Fun I will include the plots of the tracts that increased vs decreased. Note red is an increase in the percentage, and blue is a decrease. I chose these colors because there is a clear distinction between them.

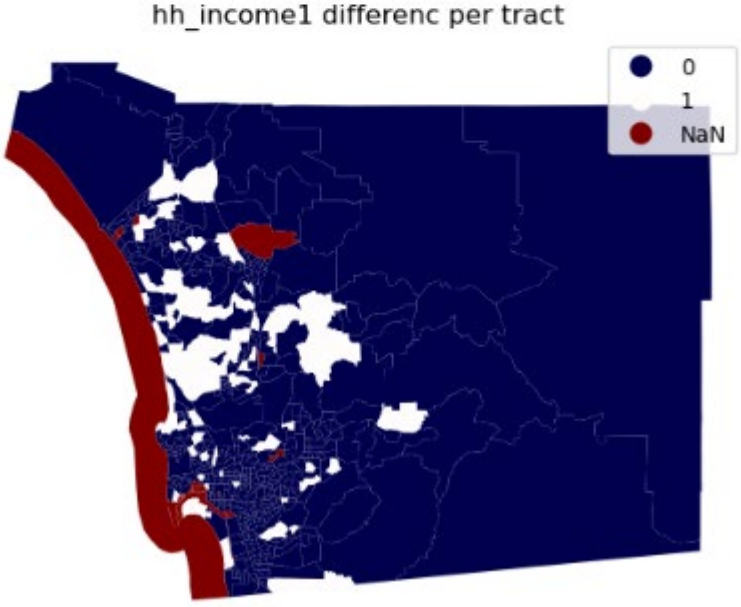

pop1 differenc per tract
1.0
0.8
0.6
0.4
0.2
0.0


white_pop1 differenc per tract
1.0
0.8
0.6
0.4
0.2
0.0


black_pop1 differenc per tract
1.0
0.8
0.6
0.4
0.2
0.0


bach_pop1 differenc per tract
1.0
0.8
0.6
0.4
0.2
0.0


asian_pop1 differenc per tract
1.0
0.8
0.6
0.4
0.2
0.0


hh_income1 differenc per tract
0
1
NaN