In [1]:

```python
# Reporting version of Capstone project
```

In [2]:

```python
# Code to import packages - (learn how to hide)
import descartes

import folium # map rendering library

import geopandas as gpd
from geopy.geocoders import Nominatim # convert an address into latitude and longitude

import json

import matplotlib.pyplot as plt


import numpy as np
from numpy.polynomial.polynomial import polyfit

import pandas as pd
from pandas.plotting import scatter_matrix

import requests # library to handle requests

from scipy.stats import chi2_contingency

import seaborn as sns

from shapely import wkt
from shapely.geometry import MultiPolygon, Polygon

from sklearn.cluster import KMeans # KMeans clustering
from sklearn.decomposition import PCA
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_absolute_error
from sklearn.metrics import r2_score
from sklearn.model_selection import train_test_split
from sklearn import preprocessing
```

Introduction

With the national debt being at an all-time high for the US, becoming more efficent at managing funds at the local-level could be a great opportunity to also maximize spending at a federal level (and nationwide). So this study is going to explore a dataset of various cities within a specific area that includes various socioeconomic factors, by grouping them into different groups and creating different benchmarks for each group. By using unsupervised machine learning models that help us cluster cities into groups, this would hopefully help us provide us with enough data to make informed decisions on how to use supervised machine learning techiques to create potential benchmarks and (logisical) models that cities can use to measure the effectiveness of current resources and predict future success. Once the data has been group accordingly, then we are going to grabbing location data from popular areas in each city to see if there is an indirect relationship that we can identify (for future studies).

Data

Understanding that Los Angeles county ranks #1 for largest county population in the US (~10 Million - larger than US 41 states), the goal of this study is to use statisical analysis and machine learning techiques on this dataset to classify cities within LA County into groups of clusters that help indentify population averages, benchmarks and indicators of success for each group - based on a variety of socioeconomic factors (i.e., income, school enrollment, life expectancy, etc.). This will be helpful for city planning and future research purposes by building off the initial research (www.measureofamerica.org/los-angeles-county/ (http://www.measureofamerica.org/los-angeles-county/)). This framework will also be useful for inserting other Los Angeles datasets for classification purposes.

Leveraging data made available by the County of Los Angeles at (www.data.lacounty.gov/ (http://www.data.lacounty.gov/)), we will be using 'A Portrait of Los Angeles County using the Human Development Index: GIS Data' at (www.data.lacounty.gov/Community/A-Portrait-of-Los-Angeles-County-using-the-Human-D/j7aj-mn8v (http://www.data.lacounty.gov/Community/A-Portrait-of-Los-Angeles-County-using-the-Human-D/j7aj-mn8v)). HD Index explaination - (https://ssrc-static.s3.amazonaws.com/moa/PoLA%20Methodological%20Note.pdf (https://ssrc-static.s3.amazonaws.com/moa/PoLA%20Methodological%20Note.pdf))

Once the cities have been grouped into clusters, we will be grabbing population locations for each city and grouping the location data by cluster for further analysis.

Original data:

In [3]:
```
LA_HPI_CSV='A_Portrait_of_Los_Angeles_County_using_the_Human_Development_Index__GIS_Dat
LA_HPI=pd.read_csv(LA_HPI_CSV) # Read in csv data into a pandas dataframe
LA_HPI.head() # Dataframe preview
```

Out[3]:

	the_geom	GEO_NAME	GEO_ID	GEO_TYPE	HD_INDEX	LIFE_EXPEC	LESS_HS	BACHEL(
0	MULTIPOLYGON (((-118.22611962104467 34.0621774...	Northeast Los Angeles	1010	City of Los Angeles Community Plan Area	4.85	83.3	30.9	;
1	MULTIPOLYGON (((-118.37014808865722 34.1963466...	North Hollywood - Valley Village	2130	City of Los Angeles Community Plan Area	4.92	81.6	19.9	:
2	MULTIPOLYGON (((-118.22539176891415 34.0719216...	Central City North	1110	City of Los Angeles Community Plan Area	3.50	82.3	39.0	;
3	MULTIPOLYGON (((-118.62899162601589 34.1472726...	Canoga Park - Winnetka - Woodland Hills - West...	2200	City of Los Angeles Community Plan Area	6.02	82.8	14.8	:
4	MULTIPOLYGON (((-118.37114153268259 34.2598174...	Sun Valley - La Tuna Canyon	2170	City of Los Angeles Community Plan Area	4.19	82.1	33.5	'

Cleaning up data up by reformatting columns, dropping irrelevant columns, and converting coordinates into polygon objects for mapping

In [4]:
```
LA_HPI.drop(columns=['GEO_TYPE','GEO_ID'],inplace=True) # Drop irrelevant columns
LA_HPI_columns=['Polygon','City','Human Development Index', 'Life Expectancy', 'No HS D
       'School Enrollment', 'Earnings', 'Health Index', 'Education Index', 'Income Inde
LA_HPI.columns=LA_HPI_columns # Replace column names
LA_HPI["Polygon"]=LA_HPI["Polygon"].apply(wkt.loads) # Create polygon object for graphi
LA_HPI.head() # Dataframe preview
```

Out[4]:

	Polygon	City	Human Development Index	Life Expectancy	No HS Diplomas	Bachelors Degrees	Graduate Degrees	School Enrollment
0	(POLYGON ((-118.2261196210447 34.0621774102914...	Northeast Los Angeles	4.85	83.3	30.9	25.4	8.0	80.3
1	(POLYGON ((-118.3701480886572 34.1963466238140...	North Hollywood - Valley Village	4.92	81.6	19.9	32.8	8.4	74.1
2	(POLYGON ((-118.2253917689142 34.0719216988227...	Central City North	3.50	82.3	39.0	22.2	6.9	54.4
3	(POLYGON ((-118.6289916260159 34.1472726007404...	Canoga Park - Winnetka - Woodland Hills - West...	6.02	82.8	14.8	37.2	12.9	79.9
4	(POLYGON ((-118.3711415326826 34.2598174361240...	Sun Valley - La Tuna Canyon	4.19	82.1	33.5	17.4	4.1	77.6

The dataset contains 140 rows and 12 columns. One row for each city; along with various columns for factors that pertain to health, education, and living standards, along with name and geographic information.

After downloading and formatting the dataset into a pandas dataframe (to make it easy to manipulate, plot, map and analyze the data), we now create another dataframe that we can use for calculations by transforming our cleaned up 140x12 dataset into a 140x10 dataset by setting 'City' as the index and removing the 'Polygon' column.

In [5]:

```
LA_HPI_Table=LA_HPI # Create table dataframe
LA_HPI_Table=LA_HPI_Table.drop(columns='Polygon') # Drop city column from table datafra
LA_HPI_Table.set_index('City',inplace=True) # Set city names as index
LA_HPI_Table.head() # Dataframe preview
```

Out[5]:

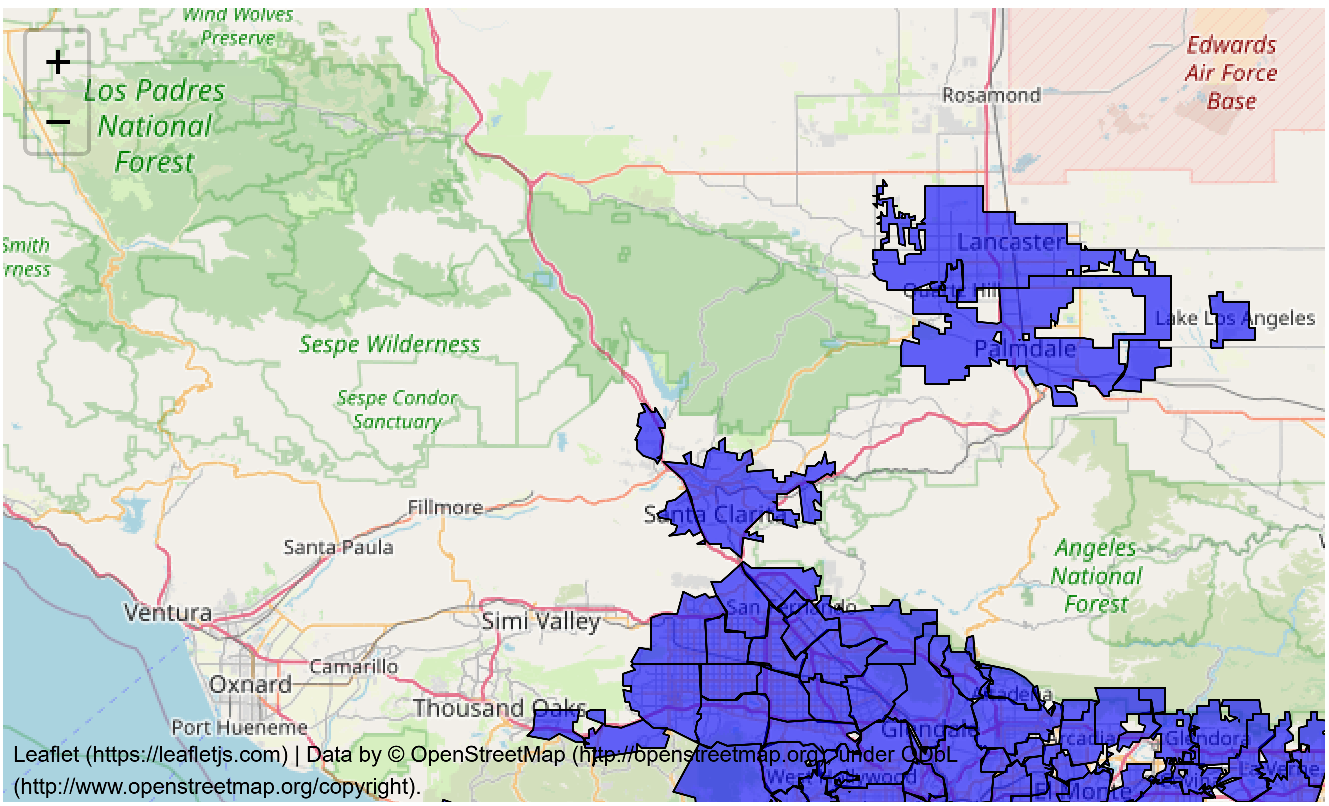
City	Human Development Index	Life Expectancy	No HS Diplomas	Bachelors Degrees	Graduate Degrees	School Enrollment	Earnings	Health Index	Educ I
Northeast Los Angeles	4.85	83.3	30.9	25.4	8.0	80.3	24503	7.22	
North Hollywood - Valley Village	4.92	81.6	19.9	32.8	8.4	74.1	27157	6.48	
Central City North	3.50	82.3	39.0	22.2	6.9	54.4	20909	6.77	
Canoga Park - Winnetka - Woodland Hills - West Hills	6.02	82.8	14.8	37.2	12.9	79.9	34243	7.00	
Sun Valley - La Tuna Canyon	4.19	82.1	33.5	17.4	4.1	77.6	22596	6.72	

Now we take a look at how these different areas differ from city to city using maps:

In [6]:

```python
# Get coordinates (latitude, longtitude) for Los Angeles County
address='Los Angeles County, US'
geolocator = Nominatim(user_agent="CA_explorer")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude

# Converting 'Polygon' column from dataframe into geodataframe for plotting
LA_HPI_gdf=gpd.GeoDataFrame(LA_HPI,geometry='Polygon')
LA_HPI_gdf_json=LA_HPI_gdf.to_json() # Convert from geodataframe to json for choropleth

# Create map of Los Angeles County using latitude and longitude values
map_LA_County = folium.Map(location=[latitude, longitude], zoom_start=9)

# Map features
LA_HPI_gdf_Points = folium.features.Choropleth(LA_HPI_gdf_json)
map_LA_County.add_child(LA_HPI_gdf_Points)

# Exporting the map to a HTML Image File
map_LA_County.save('LA County Map.html')
map_LA_County.save('LA County Map.PNG')

# Display Map
map_LA_County
```

Out[6]:



Map of Cities within LA County (above)

Map of cities by category density

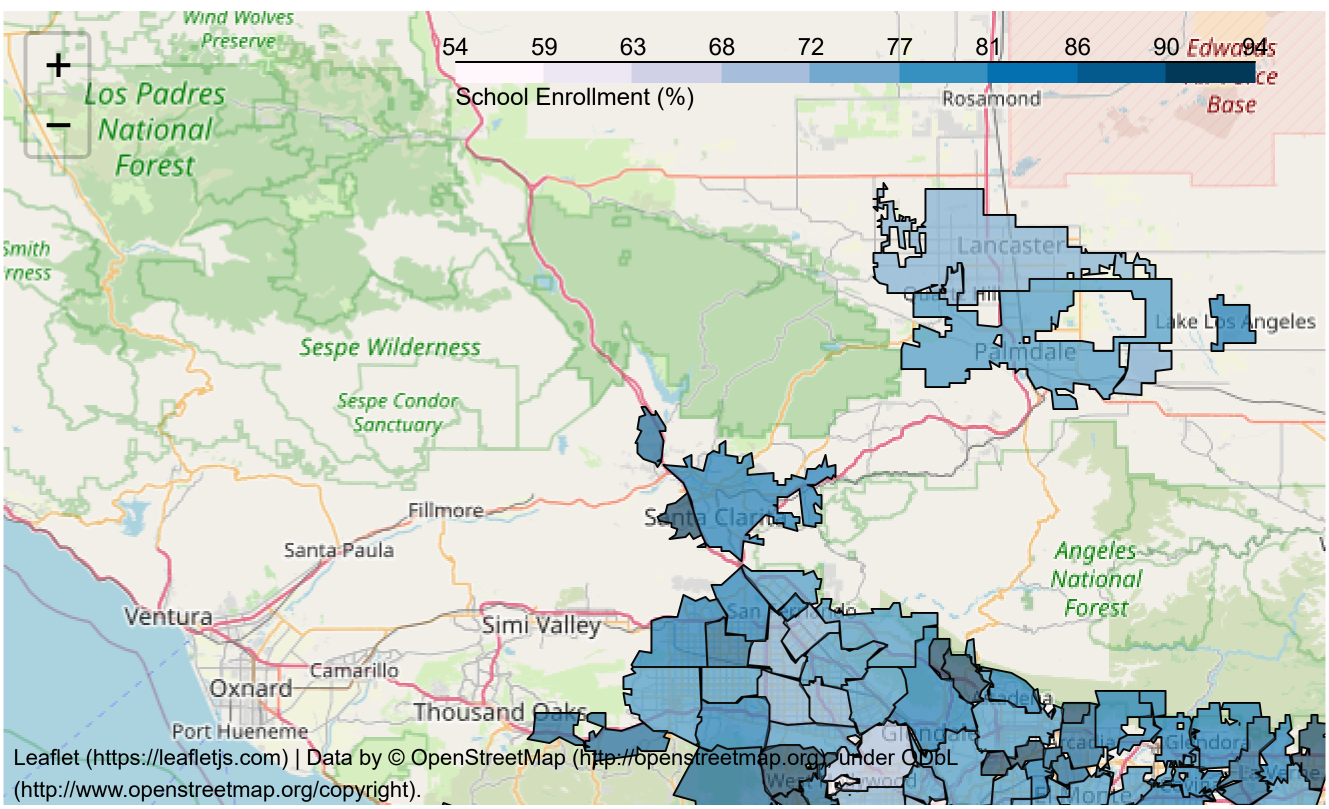
In [7]:

```
# For plotting features on map
style_function = lambda x: {'fillColor': '#ffffff',
                            'color':'#000000',
                            'fillOpacity': 0.1,
                            'weight': 0.1}
highlight_function = lambda x: {'fillColor': '#000000',
                                'color':'#000000',
                                'fillOpacity': 0.50,
                                'weight': 0.1}
```

In [8]:

```python
Enrollment_Geo=['City','School Enrollment']

# Initialize the map:
map_LA_County = folium.Map([latitude, longitude], zoom_start=9)

choropleth=folium.Choropleth(
    geo_data=LA_HPI_gdf_json,
    name='choropleth',
    data=LA_HPI[Enrollment_Geo],
    columns=Enrollment_Geo,
    key_on='feature.properties.City',
    bins=9,
    fill_color='PuBu',
    fill_opacity=0.7,
    line_opacity=1.2,
    legend_name='School Enrollment (%)',
    highlight=True
).add_to(map_LA_County)
choropleth.geojson.add_child(
    folium.features.GeoJsonTooltip(['City'],labels=False)
)

choropleth=folium.features.GeoJson(
    LA_HPI_gdf_json,
    style_function=style_function,
    control=False,
    highlight_function=highlight_function,
    tooltip=folium.features.GeoJsonTooltip(
        fields=Enrollment_Geo,
        aliases=['City: ','School Enrollment in population %: '],
        style=("background-color: white; color: #333333; font-family: arial; font-size:
    )
)
map_LA_County.add_child(choropleth)

map_LA_County
```

Out[8]:

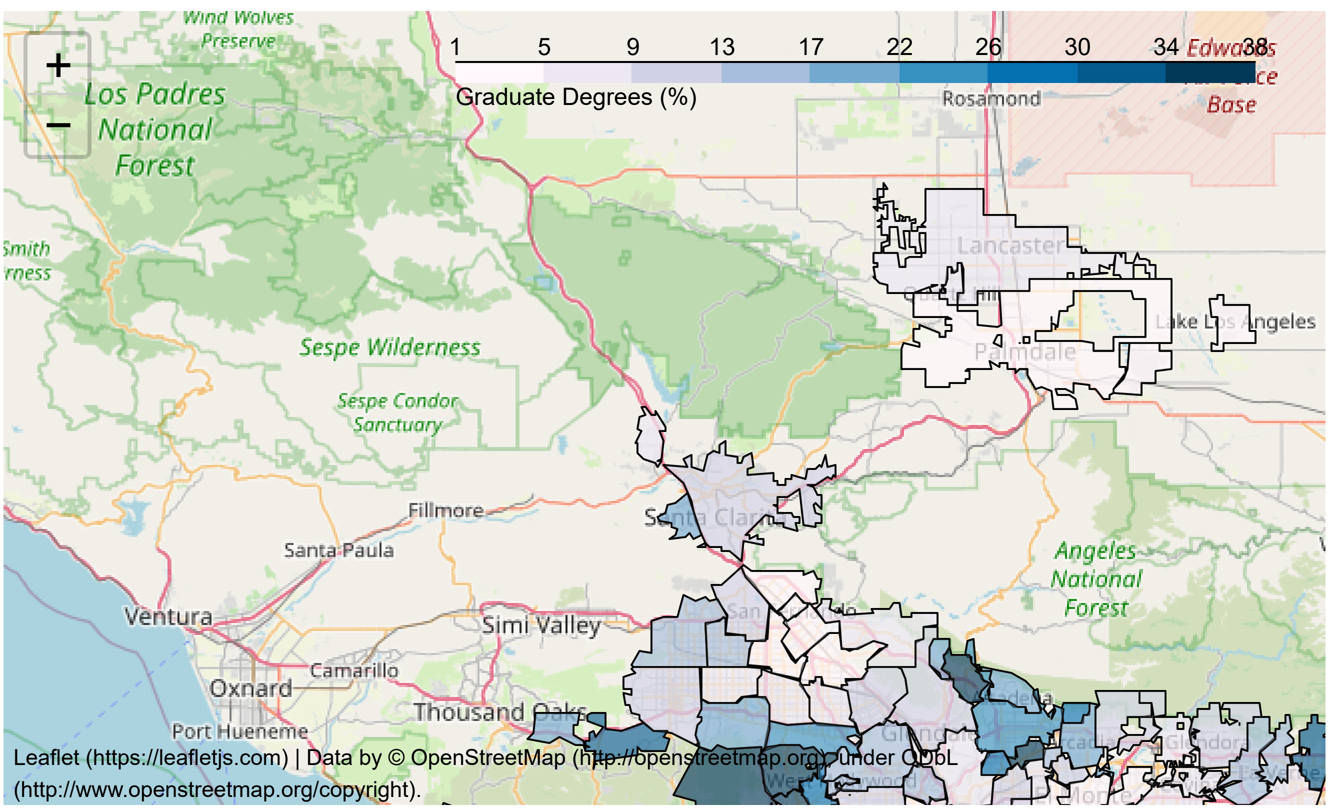


Map of Cities within LA County by School Enrollment (above)

In [9]:

```python
Graduate_Geo=['City','Graduate Degrees']

# Initialize the map:
map_LA_County = folium.Map([latitude, longitude], zoom_start=9)

choropleth=folium.Choropleth(
    geo_data=LA_HPI_gdf_json,
    name='choropleth',
    data=LA_HPI[Graduate_Geo],
    columns=Graduate_Geo,
    key_on='feature.properties.City',
    bins=9,
    fill_color='PuBu',
    fill_opacity=0.7,
    line_opacity=1.2,
    legend_name='Graduate Degrees (%)',
    highlight=True
).add_to(map_LA_County)
choropleth.geojson.add_child(
    folium.features.GeoJsonTooltip(['City'],labels=False)
)

choropleth=folium.features.GeoJson(
    LA_HPI_gdf_json,
    style_function=style_function,
    control=False,
    highlight_function=highlight_function,
    tooltip=folium.features.GeoJsonTooltip(
        fields=Graduate_Geo,
        aliases=['City: ','Graduate degrees in population %: '],
        style=("background-color: white; color: #333333; font-family: arial; font-size:
    )
)
map_LA_County.add_child(choropleth)

map_LA_County
```

Out[9]:

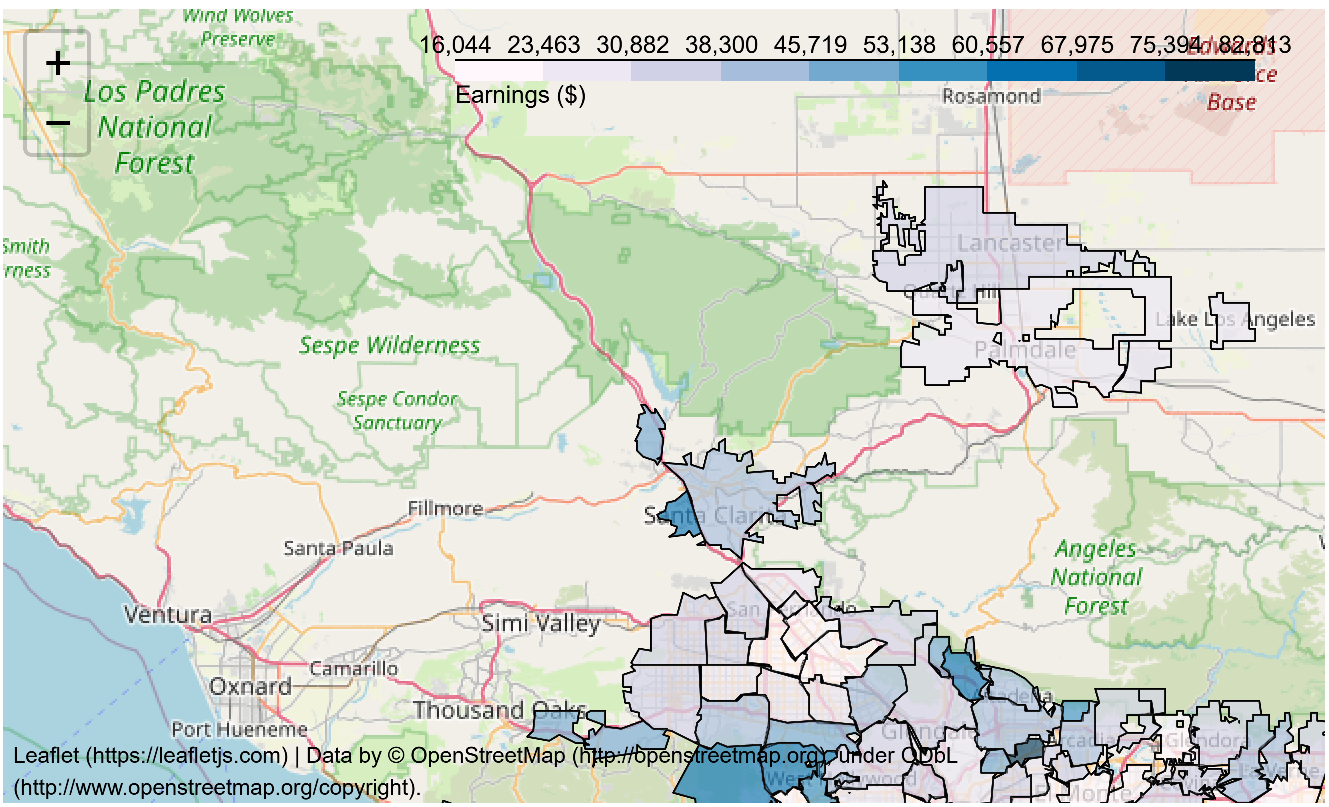


Map of Cities within LA County by Graduate Degrees (above)

In [10]:

```python
Earnings_Geo=['City','Earnings']

# Initialize the map:
map_LA_County = folium.Map([latitude, longitude], zoom_start=9)

choropleth=folium.Choropleth(
    geo_data=LA_HPI_gdf_json,
    name='choropleth',
    data=LA_HPI[Earnings_Geo],
    columns=Earnings_Geo,
    key_on='feature.properties.City',
    bins=9,
    fill_color='PuBu',
    fill_opacity=0.7,
    line_opacity=1.2,
    legend_name='Earnings ($)',
    highlight=True
).add_to(map_LA_County)
choropleth.geojson.add_child(
    folium.features.GeoJsonTooltip(['City'],labels=False)
)

choropleth=folium.features.GeoJson(
    LA_HPI_gdf_json,
    style_function=style_function,
    control=False,
    highlight_function=highlight_function,
    tooltip=folium.features.GeoJsonTooltip(
        fields=Earnings_Geo,
        aliases=['City: ','Earnings in population $: '],
        style=("background-color: white; color: #333333; font-family: arial; font-size:
    )
)
map_LA_County.add_child(choropleth)

map_LA_County
```

Out[10]:

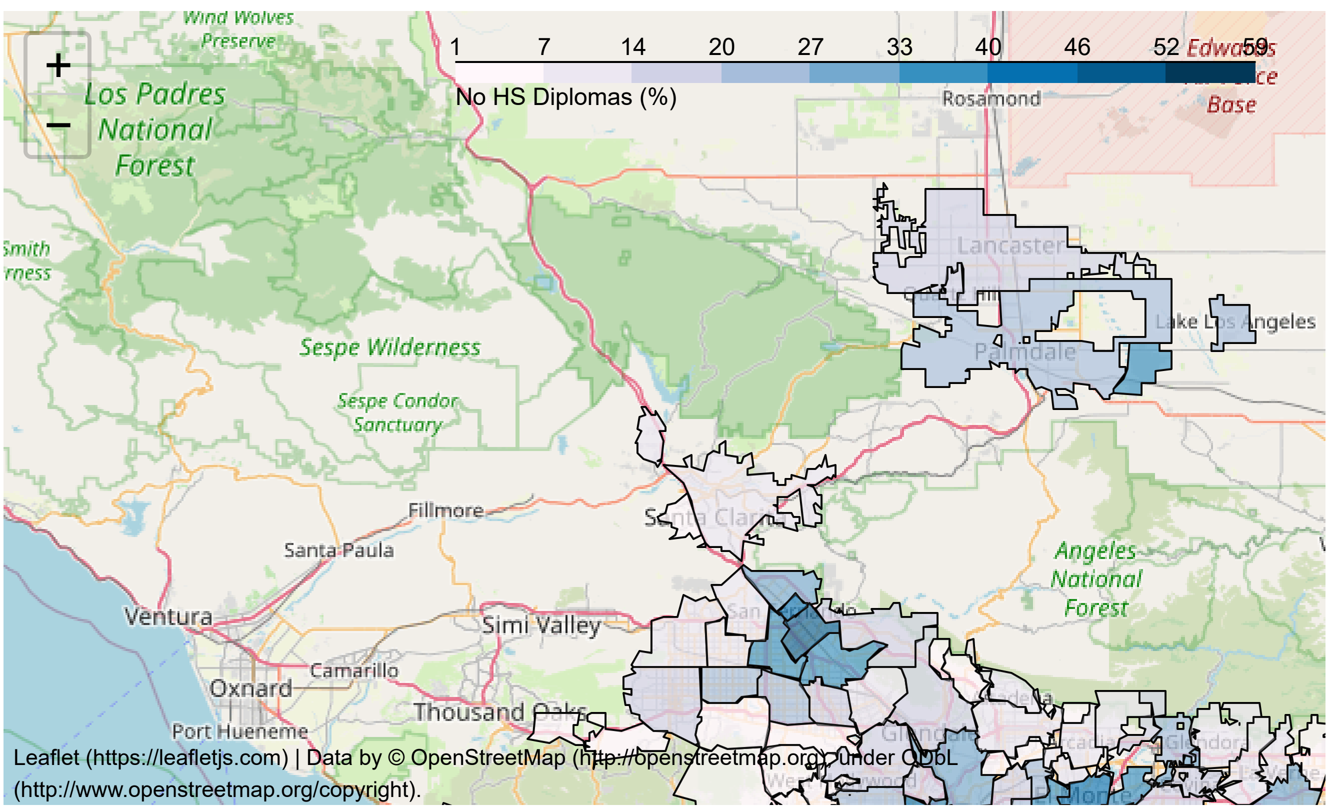


Map of Cities within LA County by Earnings (above)

In [11]:

```python
No_HS_Geo=['City','No HS Diplomas']

# Initialize the map:
map_LA_County = folium.Map([latitude, longitude], zoom_start=9)

choropleth=folium.Choropleth(
    geo_data=LA_HPI_gdf_json,
    name='choropleth',
    data=LA_HPI[No_HS_Geo],
    columns=No_HS_Geo,
    key_on='feature.properties.City',
    bins=9,
    fill_color='PuBu',
    fill_opacity=0.7,
    line_opacity=1.2,
    legend_name='No HS Diplomas (%)',
    highlight=True
).add_to(map_LA_County)
choropleth.geojson.add_child(
    folium.features.GeoJsonTooltip(['City'],labels=False)
)

choropleth=folium.features.GeoJson(
    LA_HPI_gdf_json,
    style_function=style_function,
    control=False,
    highlight_function=highlight_function,
    tooltip=folium.features.GeoJsonTooltip(
        fields=No_HS_Geo,
        aliases=['City: ','No HS Diplomas in population %: '],
        style=("background-color: white; color: #333333; font-family: arial; font-size:
    )
)
map_LA_County.add_child(choropleth)

map_LA_County
```

Out[11]:

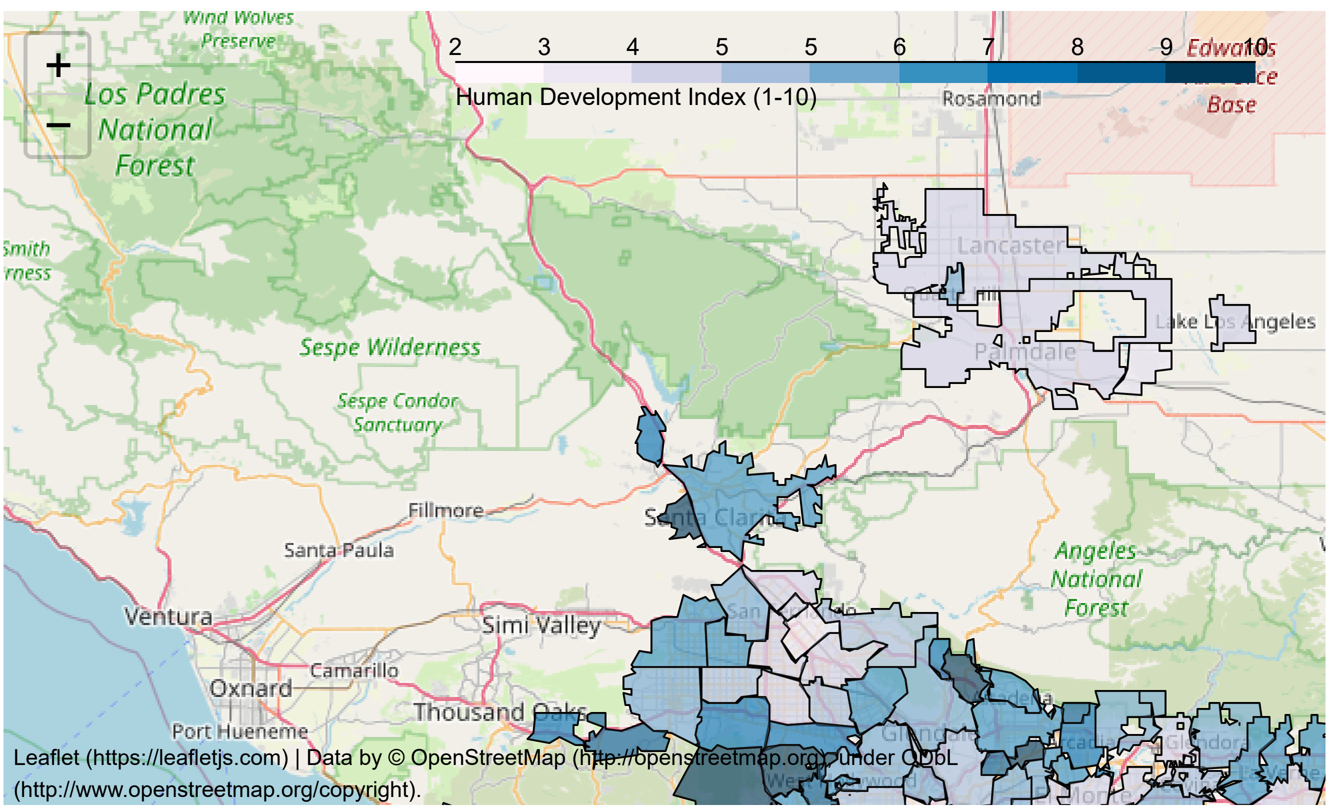


Map of Cities within LA County by No HS Diplomas (above)

In [12]:

```python
HDI_Geo=['City','Human Development Index']

# Initialize the map:
map_LA_County = folium.Map([latitude, longitude], zoom_start=9)

choropleth=folium.Choropleth(
    geo_data=LA_HPI_gdf_json,
    name='choropleth',
    data=LA_HPI[HDI_Geo],
    columns=HDI_Geo,
    key_on='feature.properties.City',
    bins=9,
    fill_color='PuBu',
    fill_opacity=0.7,
    line_opacity=1.2,
    legend_name='Human Development Index (1-10)',
    highlight=True
).add_to(map_LA_County)
choropleth.geojson.add_child(
    folium.features.GeoJsonTooltip(['City'],labels=False)
)

choropleth=folium.features.GeoJson(
    LA_HPI_gdf_json,
    style_function=style_function,
    control=False,
    highlight_function=highlight_function,
    tooltip=folium.features.GeoJsonTooltip(
        fields=HDI_Geo,
        aliases=['City: ','Human Development Index in population (1-10): '],
        style=("background-color: white; color: #333333; font-family: arial; font-size:
    )
)
map_LA_County.add_child(choropleth)

map_LA_County
```

Out[12]:

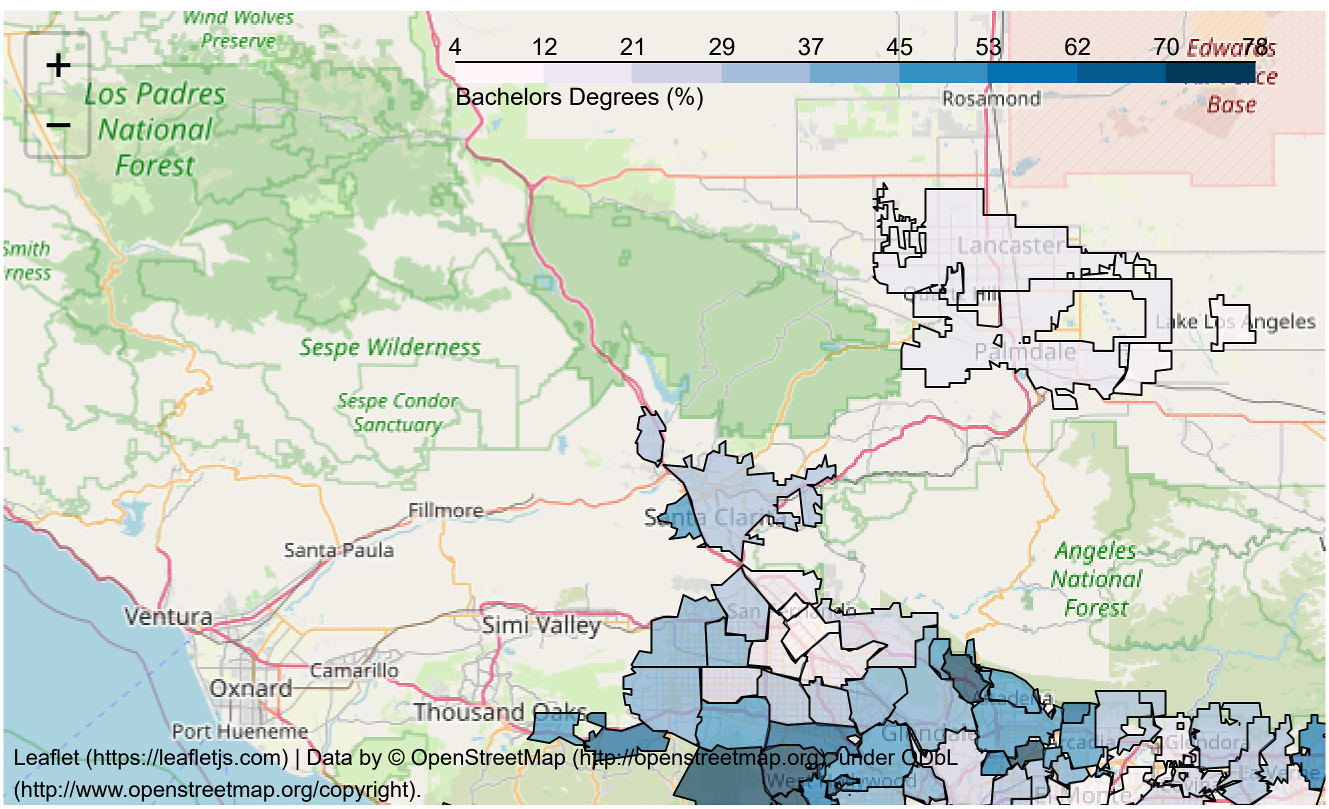


Map of Cities within LA County by Human Development Index (above)

In [13]:

```python
Bachelors_Geo=['City','Bachelors Degrees']

# Initialize the map:
map_LA_County = folium.Map([latitude, longitude], zoom_start=9)

choropleth=folium.Choropleth(
    geo_data=LA_HPI_gdf_json,
    name='choropleth',
    data=LA_HPI[Bachelors_Geo],
    columns=Bachelors_Geo,
    key_on='feature.properties.City',
    bins=9,
    fill_color='PuBu',
    fill_opacity=0.7,
    line_opacity=1.2,
    legend_name='Bachelors Degrees (%)',
    highlight=True
).add_to(map_LA_County)
choropleth.geojson.add_child(
    folium.features.GeoJsonTooltip(['City'],labels=False)
)

choropleth=folium.features.GeoJson(
    LA_HPI_gdf_json,
    style_function=style_function,
    control=False,
    highlight_function=highlight_function,
    tooltip=folium.features.GeoJsonTooltip(
        fields=Bachelors_Geo,
        aliases=['City: ','Bachelors Degrees in population (%): '],
        style=("background-color: white; color: #333333; font-family: arial; font-size:
    )
)
map_LA_County.add_child(choropleth)

map_LA_County
```

Out[13]:



Map of Cities within LA County by Bachelors Degrees (above)

From looking at the different maps, we see a clear correlation between higher performing cities and their proximity to the ocean which is not surprising. But their also seems to be a line of high performing cities that run from the ocean through LA all the way up into the Angeles forest. Would be interesting the analyze the ages in these populations to see if this is predictive of general migration patterns as people progress throughout their careers.

After the data was clean-up and formatted, we then do a quick visual analysis of the data to get a better understanding of the overall distribution for the different categories. Using histograms:

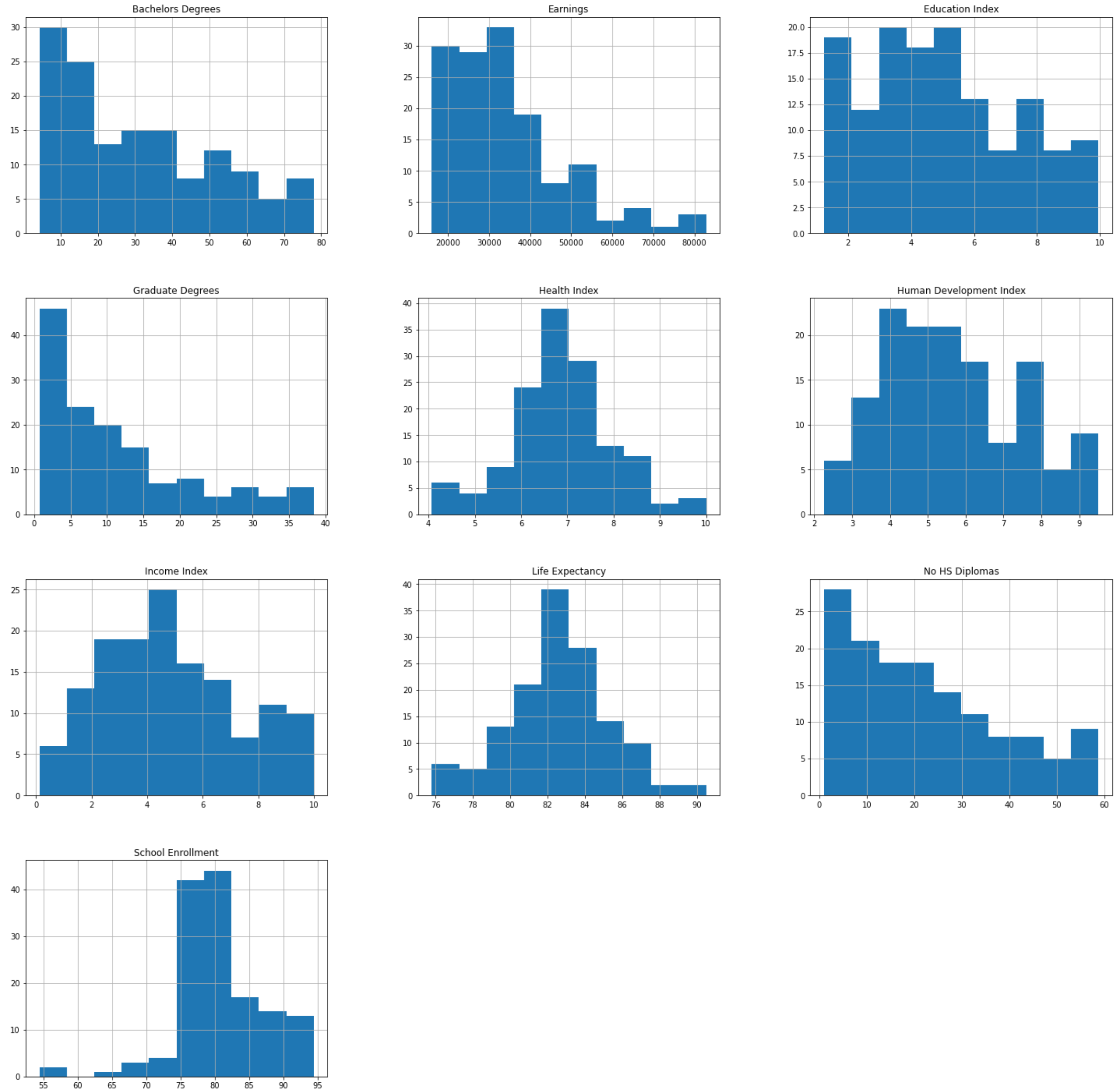
In [14]:
```
LA_HPI_Table
```

Out[14]:

City	Human Development Index	Life Expectancy	No HS Diplomas	Bachelors Degrees	Graduate Degrees	School Enrollment	Earnings	Health Index	Educ I
Northeast Los Angeles	4.85	83.3	30.9	25.4	8.0	80.3	24503	7.22	
North Hollywood - Valley Village	4.92	81.6	19.9	32.8	8.4	74.1	27157	6.48	
Central City North	3.50	82.3	39.0	22.2	6.9	54.4	20909	6.77	
Canoga Park - Winnetka - Woodland Hills - West Hills	6.02	82.8	14.8	37.2	12.9	79.9	34243	7.00	
Sun Valley - La Tuna Canyon	4.19	82.1	33.5	17.4	4.1	77.6	22596	6.72	
...	...	...	...	...	...	...	...	...	
Diamond Bar	7.38	85.4	7.9	50.9	17.2	86.7	41012	8.08	
Downey	5.12	81.4	23.9	21.4	6.1	79.1	31152	6.43	
Redondo Beach	7.99	82.3	4.2	56.9	21.9	85.4	59819	6.78	
San Dimas	6.62	81.9	7.9	35.7	13.3	87.2	40843	6.62	
Harbor Gateway	3.91	78.7	27.4	19.7	4.0	78.1	23106	5.29	

140 rows × 10 columns

In [15]:
```
LA_HPI_Table.hist(figsize=(25,25)) # Create hisogram table
plt.show() # Plot histogram (remove pre-plot messages)
```

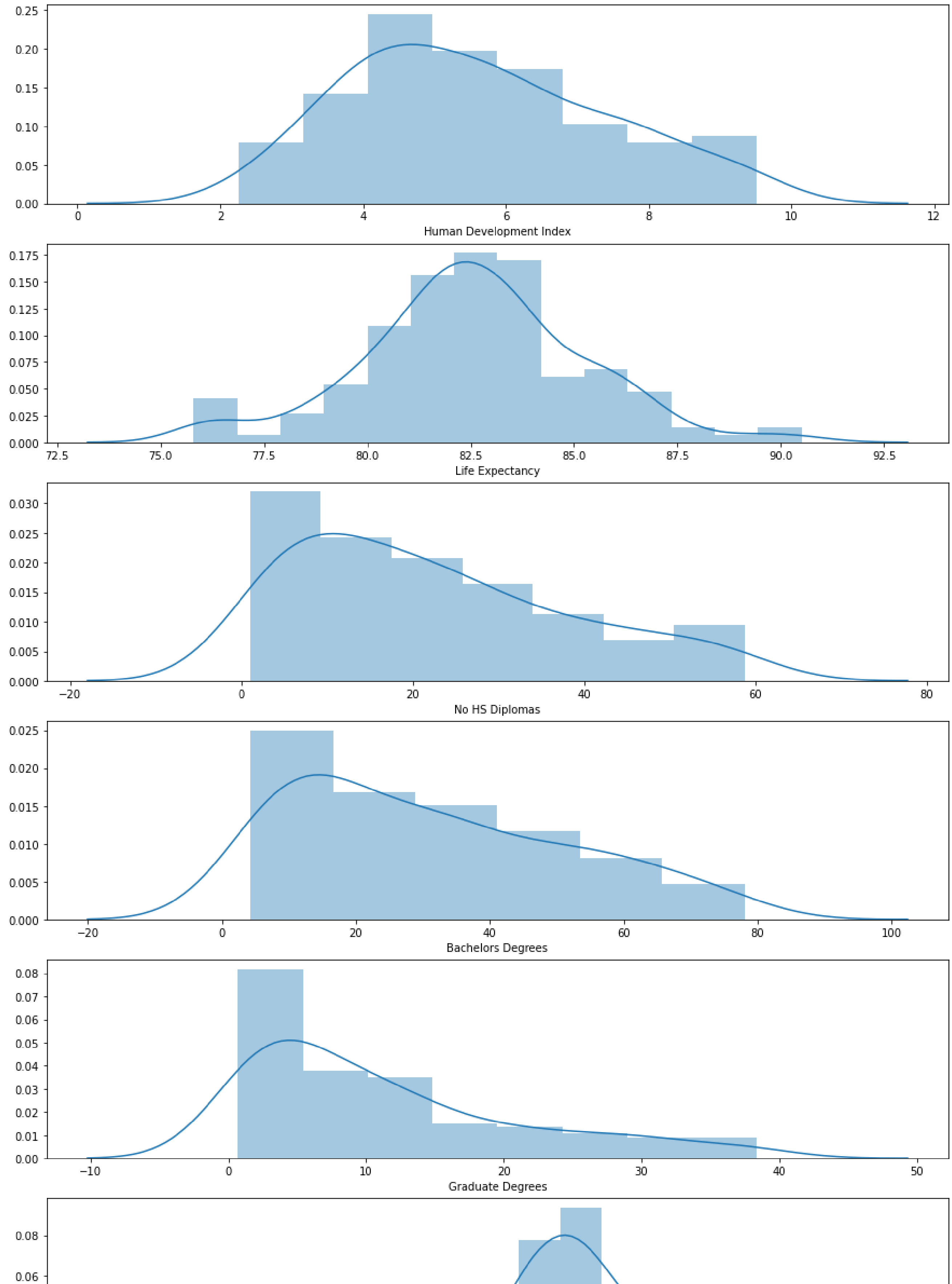


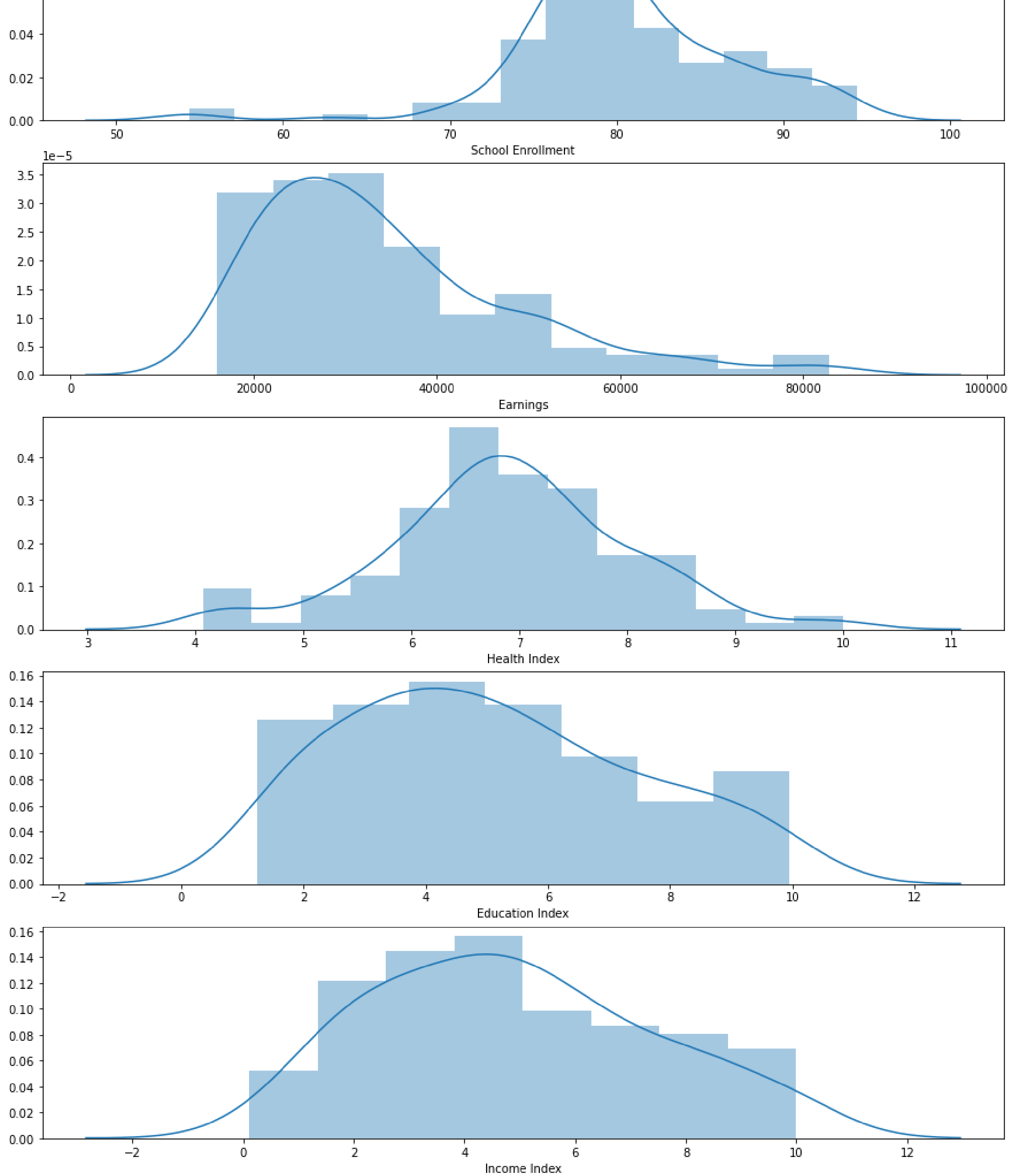
In [16]:
```
LA_HPI_Table.columns
```

Out[16]:
```
Index(['Human Development Index', 'Life Expectancy', 'No HS Diplomas',
       'Bachelors Degrees', 'Graduate Degrees', 'School Enrollment',
       'Earnings', 'Health Index', 'Education Index', 'Income Index'],
      dtype='object')
```

In [17]:

```
f, axes = plt.subplots(10, 1, figsize = [15,40])
order = 0

for column in LA_HPI_Table.columns:
    sns.distplot(a = LA_HPI_Table[column], ax=axes[order])
    order = order + 1

plt.savefig('LA County Histograms.png')
```





Looking at the histogram shows a a couple of different features.

School Enrollment:

Compared to the other charts, there doesn't seem to be the least amount of disparity between cities in this area, so seeing how this doesn't directly transfer to the greater disparity that we see with bachelor's degrees and earning, this could be worth investigating to see if these communities are doing a poor job of educating their residents or doing a poor job of retaining their residents once they are educated and higher-income earners. \

Graduate Degrees vs Bachelors Degrees:

Seems like graduate degrees are a lot more concentrated then how Bachelors degrees are distributed around LA county.

Redundent Indexes:

Their doesn't seem to be any significant relationships between the indexes and their corresponding values, so we will be dropping these later to increase the predictive power of our clustering model.

Methodology

Given our general understanding of how different area of cities within LA County are performing in different areas, now we look to explore the strength of the relationships between the different variables by looking at the correlations, to help us determine what is important for our calculations that will help us classify the cities.

After our intial exploratory data analysis, we now move onto the data cleaning phase by using machine learning to help determine which factors would be relevant for building our dimensions, clusters, and for further analysis.

Since the goal of the study is to understand how the cities within Los Angeles county group together and differ, we will be using unsupervised machine learning methods in the form of PCA and k-means clustering -- to find out how many dimensions and clusters our data should be grouped together to give us the best results.

First, we start off by standardizing our data in order to get a better understanding of the relationships within the variables. Then we create a heatmap and scatterplots to explore the relationships.

In [18]:

```
LA_HPI_fit=preprocessing.StandardScaler().fit(LA_HPI_Table).transform(LA_HPI_Table) # S
LA_HPI_fit=pd.DataFrame(LA_HPI_fit, columns=LA_HPI_Table.columns) # Converting into dat
LA_HPI_corr=LA_HPI_fit.corr() # Create correlation analysis object
LA_County_Heatmap = sns.heatmap(LA_HPI_corr) # Map correlation analysis as heatmap
LA_County_Heatmap;
```

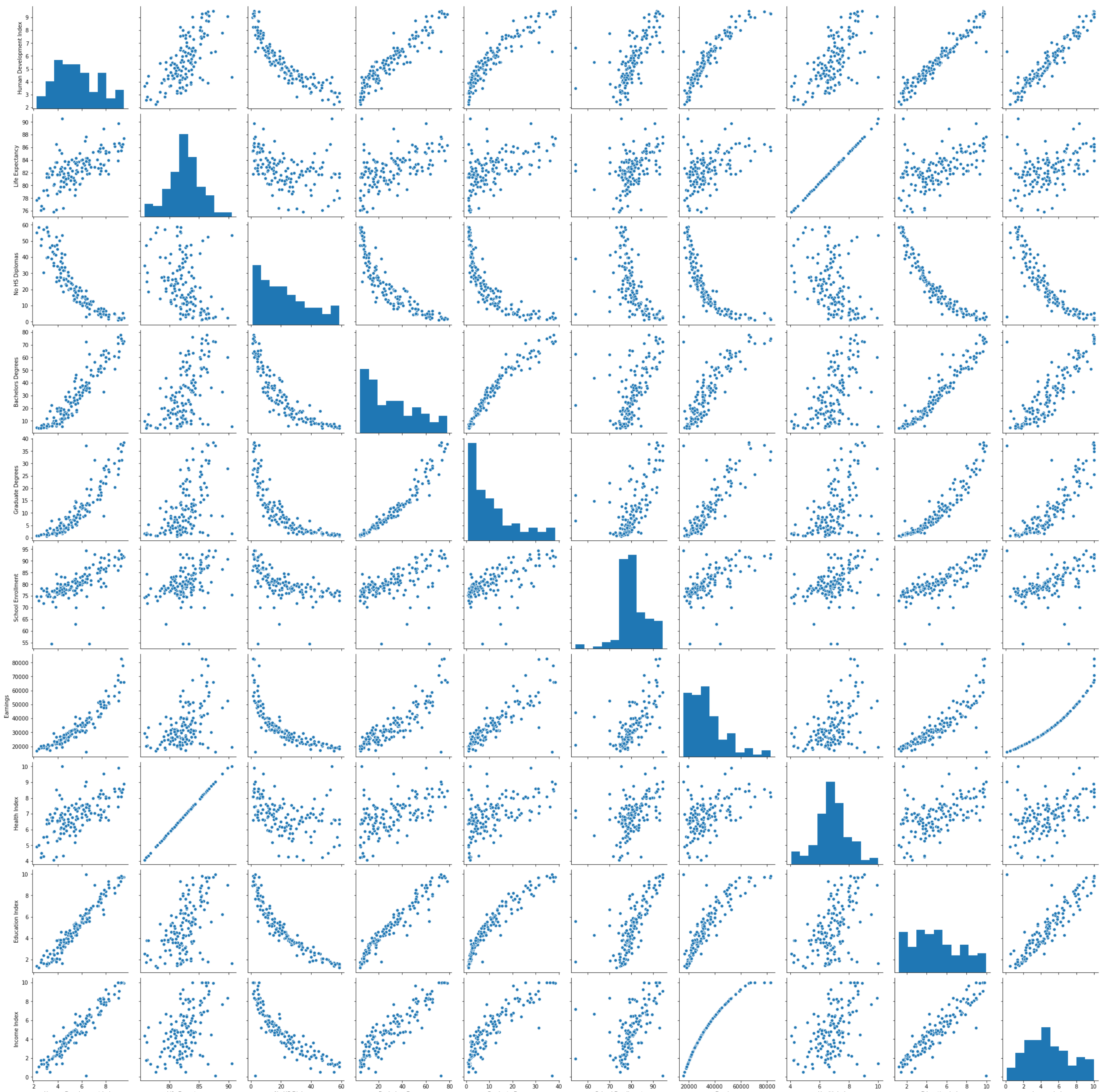
In [19]:

```
LA_County_Heatmap.figure.savefig('LA County Heatmap.PNG')
```

Correlation matrix of our dataset (above)

In [20]:
```
sns.pairplot(LA_HPI, diag_kind='hist',size=2.85) # Create scatterplot of all the variab
plt.show() # Plot
```

```
C:\Users\marky\anaconda3\envs\geo_env\lib\site-packages\seaborn\axisgrid.py:2079: UserWarning: The `size` parameter has been renamed to `height`; please update your code.
  warnings.warn(msg, UserWarning)
```

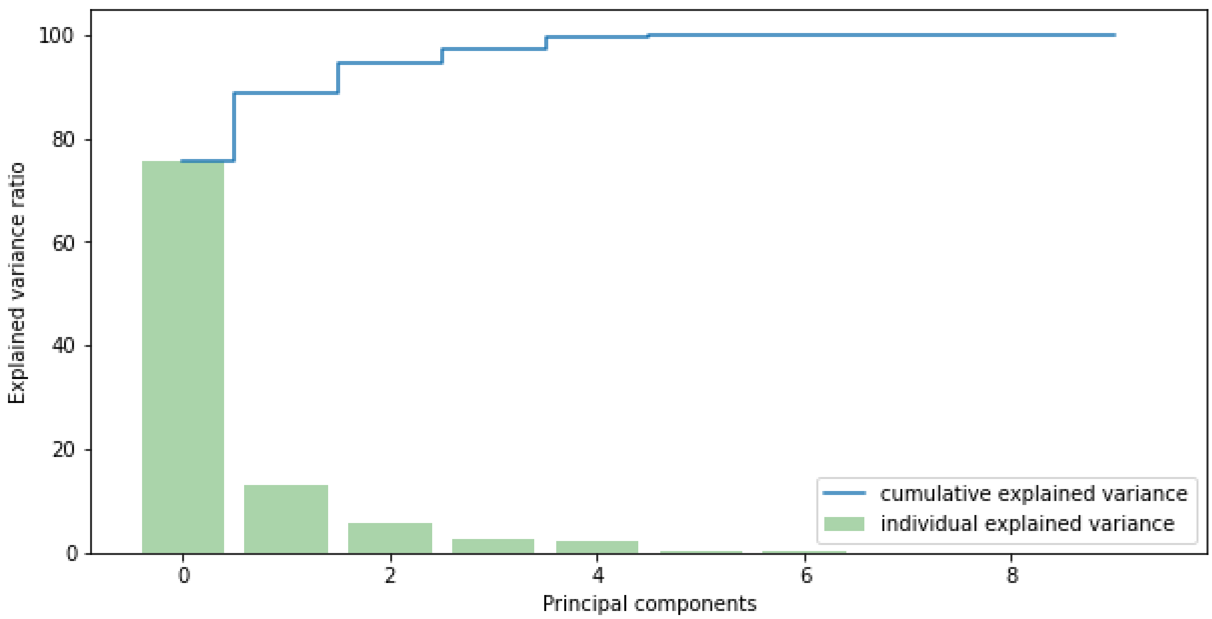


Our inital look at the strenghs of the different relationships from the correlation charts also shows us that there are clear redundancies between indexes and their corresponding values (i.e., life expectancy and health index). So we do a principal componenets analysis to make sure our dataset has enough predicitve power in it's first few columns, so that we can get rid of redundant columns.

In [21]:
```
#Calculating Eigenvecors and eigenvalues of Covariance matrix
mean_vec = np.mean(LA_HPI_fit, axis=0)
cov_mat = np.cov(LA_HPI_fit.T)
eig_vals, eig_vecs = np.linalg.eig(cov_mat)
```

In [22]:
```
eig_pairs = [ (np.abs(eig_vals[i]),eig_vecs[:,i]) for i in range(len(eig_vals))] # Crea
eig_pairs.sort(key = lambda x: x[0], reverse= True) # Sort from high to low
# Calculation of Explained Variance from the eigenvalues
tot = sum(eig_vals)
var_exp = [(i/tot)*100 for i in sorted(eig_vals, reverse=True)] # Individual explained
cum_var_exp = np.cumsum(var_exp) # Cumulative explained variance
```

In [23]:
```
# PLOT OUT THE EXPLAINED VARIANCES SUPERIMPOSED
plt.figure(figsize=(10, 5))
plt.bar(range(len(var_exp)), var_exp, alpha=0.3333, align='center', label='individual e
plt.step(range(len(cum_var_exp)), cum_var_exp, where='mid',label='cumulative explained
plt.ylabel('Explained variance ratio')
plt.xlabel('Principal components')
plt.legend(loc='best')
plt.show()
print(cum_var_exp)
```




```
[ 75.6127039   88.84934737  94.61488324  97.38941368  99.60122855
  99.80392209  99.98870637  99.9983045   99.9999758  100.        ]
```

Here we see that 3 components can account for 94.62% of variance in our dataset. So we remove redundent columns (to give our data greater predictive power) and then re-analyze the relationships between the variables.

In [24]:

```
LA_HPI_fit_V2=LA_HPI_fit # Storing information onto new dataframe
LA_HPI_fit_V2=LA_HPI_fit_V2.drop(columns=['Human Development Index','Health Index','Edu
LA_HPI_corr_V2=LA_HPI_fit_V2.corr() # Build correlation object
LA_County_Heatmap_Model = sns.heatmap(LA_HPI_corr_V2) # Create heatmap of correlation c
LA_County_Heatmap_Model;
```

In [25]:

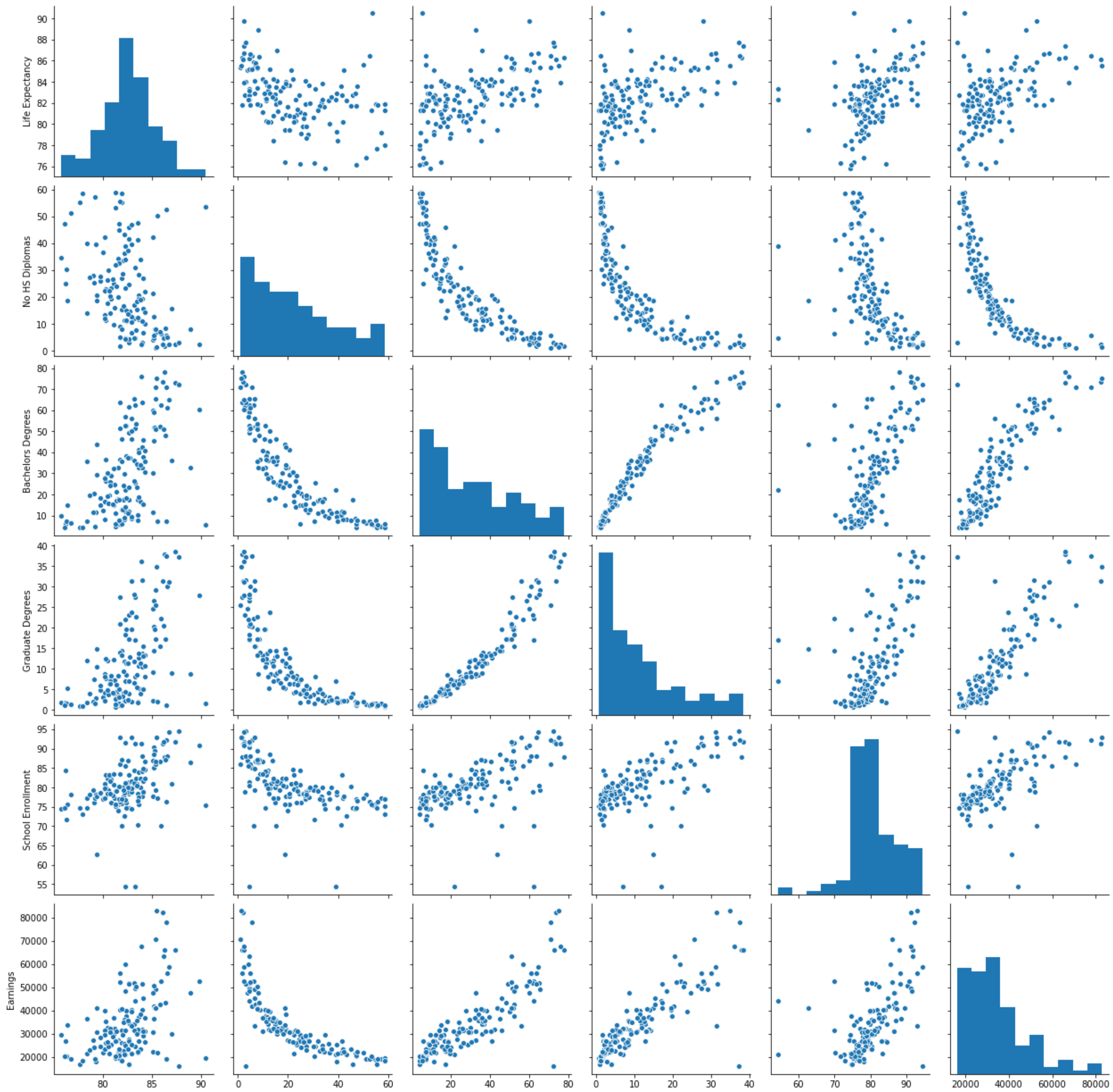
```
LA_County_Heatmap_Model.figure.savefig('LA County Heatmap Model.PNG')
```

Correlation matrix of refined dataset (above)

In [26]:

```
LA_HPI_V2=LA_HPI # Create dataframe for scatterplots
LA_HPI_V2=LA_HPI_V2.drop(columns=['Human Development Index','Health Index','Education I
LA_County_Pairplot = sns.pairplot(LA_HPI_V2, diag_kind='hist',size=2.85) # Create scatt
LA_County_Pairplot;
```

```
C:\Users\marky\anaconda3\envs\geo_env\lib\site-packages\seaborn\axisgrid.py:2079: UserWarning: The `size` parameter has been renamed to `height`; please update your code.
  warnings.warn(msg, UserWarning)
```



In [27]:

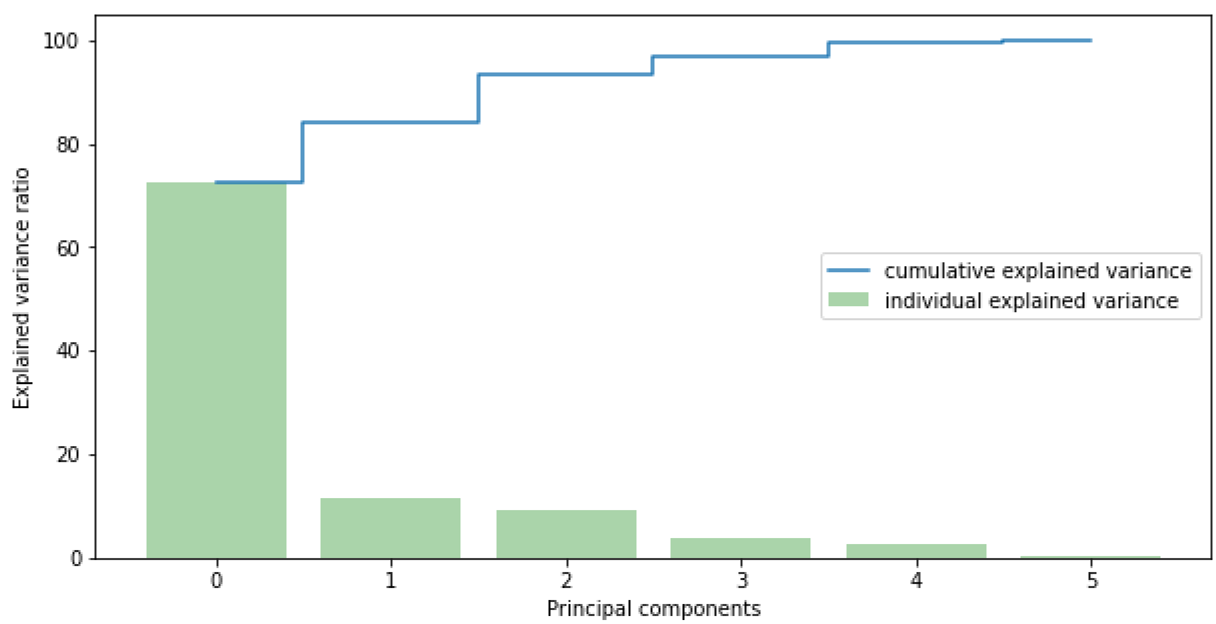
```
LA_County_Pairplot.savefig('LA County Pairplot Model.PNG')
```

Scatterplot matrix of refined dataset (above)

Now that we are happy with our dataset we then re-do a principal component analysis to see how many dimensions we should split our data into, in order to give us the most predictive power per dimension

In [28]:

```
#Calculating Eigenvecors and eigenvalues of Covariance matrix
mean_vec = np.mean(LA_HPI_fit_V2, axis=0)
cov_mat = np.cov(LA_HPI_fit_V2.T)
eig_vals, eig_vecs = np.linalg.eig(cov_mat)

# Create a list of (eigenvalue, eigenvector) tuples
eig_pairs = [ (np.abs(eig_vals[i]),eig_vecs[:,i]) for i in range(len(eig_vals))]

# Sort from high to low
eig_pairs.sort(key = lambda x: x[0], reverse= True)

# Calculation of Explained Variance from the eigenvalues
tot = sum(eig_vals)
var_exp = [(i/tot)*100 for i in sorted(eig_vals, reverse=True)] # Individual explained
cum_var_exp = np.cumsum(var_exp) # Cumulative explained variance

# PLOT OUT THE EXPLAINED VARIANCES SUPERIMPOSED
LA_County_PCA = plt.figure(figsize=(10, 5))
plt.bar(range(len(var_exp)), var_exp, alpha=0.3333, align='center', label='individual e
plt.step(range(len(cum_var_exp)), cum_var_exp, where='mid',label='cumulative explained
plt.ylabel('Explained variance ratio')
plt.xlabel('Principal components')
plt.legend(loc='best')
plt.show()
```



In [29]:

```
LA_County_PCA.savefig('LA County PCA.PNG')
```

In [30]:
```
cum_var_exp
```

Out[30]:
```
array([ 72.57163142,  84.21590562,  93.28990288,  96.8841107 ,
        99.68416985, 100.        ])
```

In [31]:
```
pca=PCA()
pca.fit(LA_HPI_fit_V2)
pca.explained_variance_ratio_
```

Out[31]:
```
array([0.72571631, 0.11644274, 0.09073997, 0.03594208, 0.02800059,
       0.0031583 ])
```

With our consolidated correlation matrix, our top 3 variables still account for 93.29% for variability for a dataset, so we will move forward with this 140 x 6 table.

In [32]:
```
LA_HPI_V2.head()
```

Out[32]:

	Polygon	City	Life Expectancy	No HS Diplomas	Bachelors Degrees	Graduate Degrees	School Enrollment	Earnings
0	MULTIPOLYGON (((-118.22612 34.06218, -118.2260...	Northeast Los Angeles	83.3	30.9	25.4	8.0	80.3	24503
1	MULTIPOLYGON (((-118.37015 34.19635, -118.3659...	North Hollywood - Valley Village	81.6	19.9	32.8	8.4	74.1	27157
2	MULTIPOLYGON (((-118.22539 34.07192, -118.2253...	Central City North	82.3	39.0	22.2	6.9	54.4	20909
3	MULTIPOLYGON (((-118.62899 34.14727, -118.6289...	Canoga Park - Winnetka - Woodland Hills - West...	82.8	14.8	37.2	12.9	79.9	34243
4	MULTIPOLYGON (((-118.37114 34.25982, -118.3705...	Sun Valley - La Tuna Canyon	82.1	33.5	17.4	4.1	77.6	22596

Given how most of the variance in the LA County datset can be explained through 3 'principal component' variables (from the analysis above), we use Prinicipal Component Analysis (PCA) to reduce the number of features from our dataset into 3.

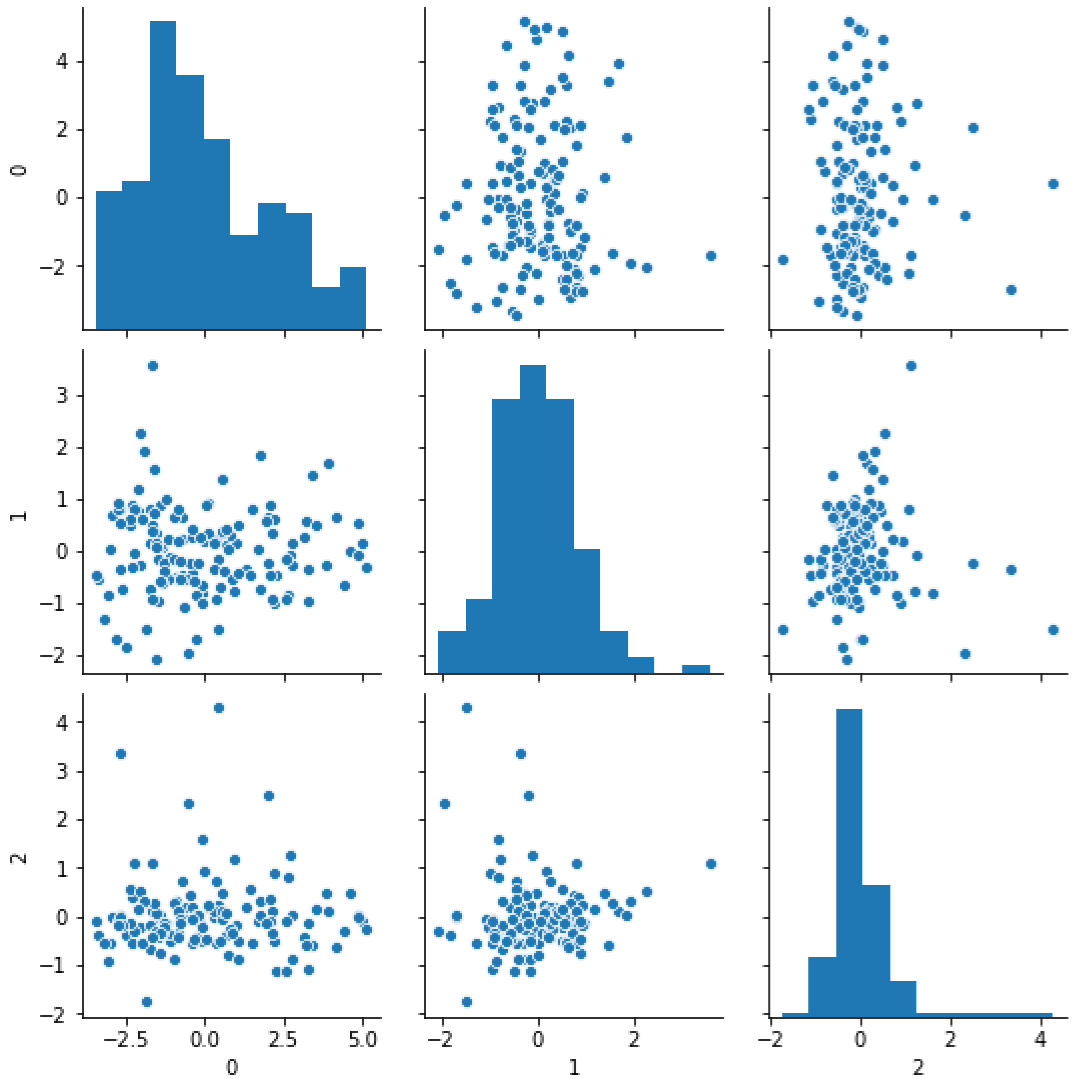
In [33]:
```
pca3 = PCA(n_components=3) # PCA object for grouping dataset into three dimensions, by
x_3d = pca3.fit_transform(LA_HPI_fit_V2) # Fit to our dataset, then transform it based
```

In [34]:
```
x_3d[:5,:] # Preview of our 3 dimensional dataset
```

Out[34]:
```
array([[-0.74351387,  0.63911838, -0.06540718],
       [-0.71017222, -0.45970881,  0.73013331],
       [-2.69684519, -0.36407416,  3.36475871],
       [ 0.39989   , -0.14889716,  0.21648856],
       [-1.5119643 ,  0.33249197,  0.01259393]])
```

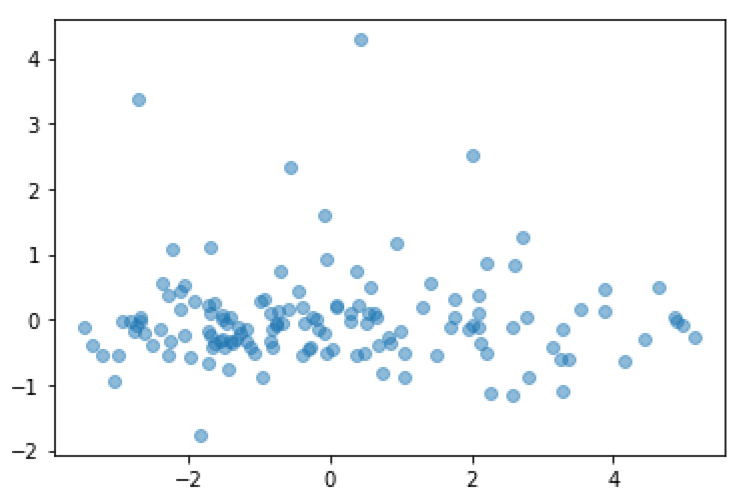
In [35]:
```python
df_pca3=pd.DataFrame(x_3d) # Dataframe from principal component analysis of 3
sns.pairplot(df_pca3) # Plot dataframe
```

Out[35]: <seaborn.axisgrid.PairGrid at 0x270b381b2b0>



In [36]:
```python
plt.scatter(x_3d[:,0],x_3d[:,2], alpha=0.5)
```

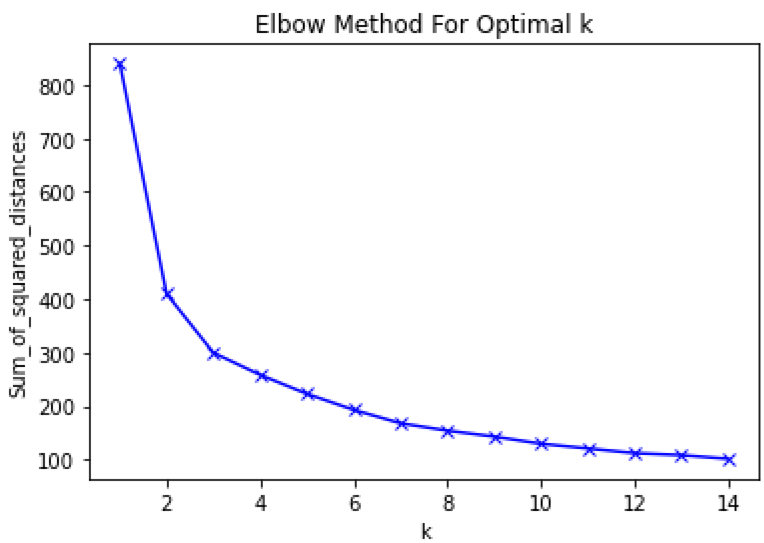
Out[36]: <matplotlib.collections.PathCollection at 0x270b372c430>



After transforming our data into 3 dimension (above), now we find out what would be our optimal k for using k-means to cluster the data.

In [37]:
```
# For Loop to collect 'sum of squared distances' for k-means clustering ranging from 1
Sum_of_squared_distances = []
K = range(1,15)
for k in K:
    km = KMeans(n_clusters=k)
    km = km.fit(LA_HPI_fit_V2)
    Sum_of_squared_distances.append(km.inertia_)
```

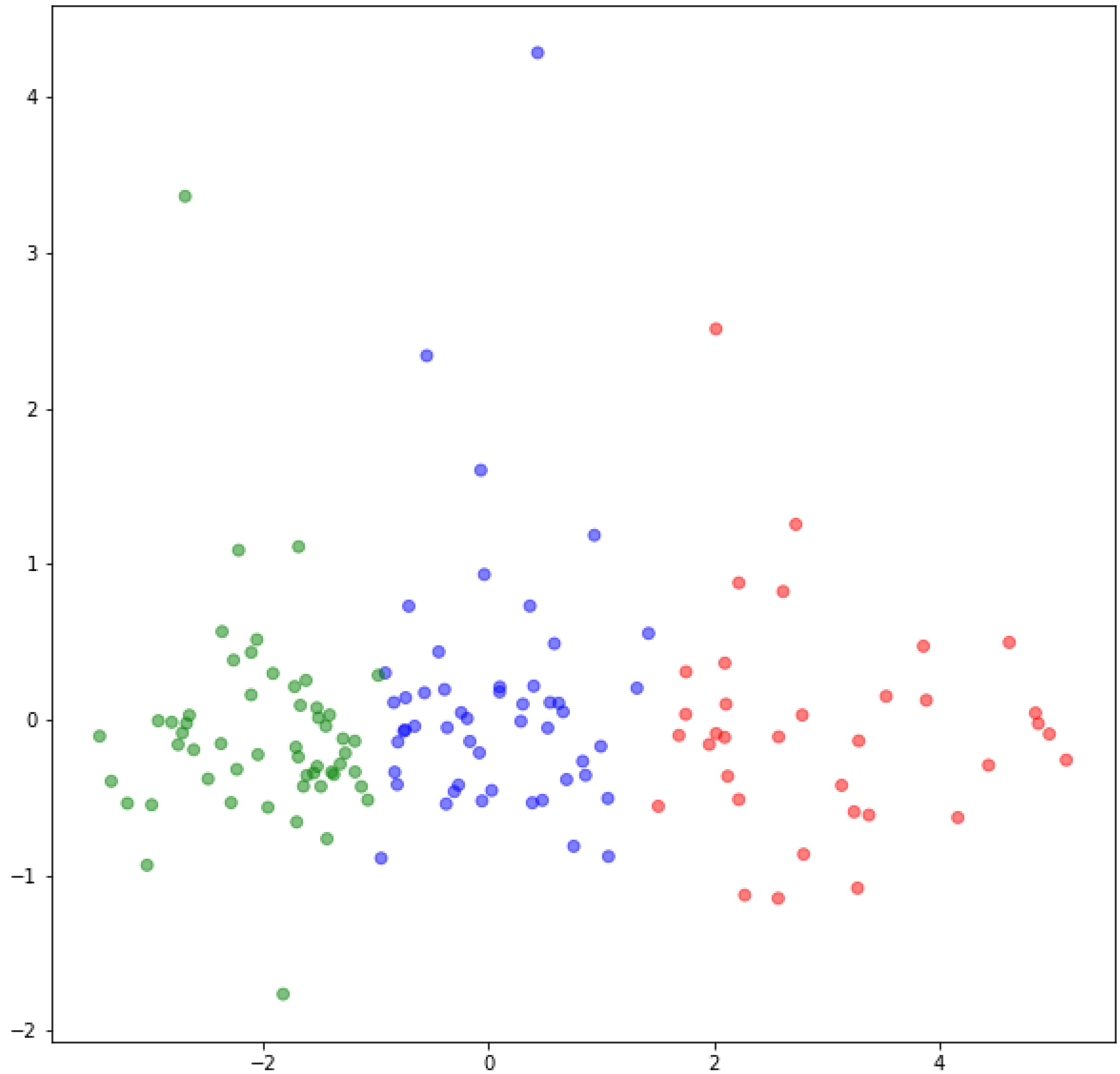
In [38]:
```
plt.plot(K, Sum_of_squared_distances, 'bx-')
plt.xlabel('k')
plt.ylabel('Sum_of_squared_distances')
plt.title('Elbow Method For Optimal k')
plt.show()
```



We see that our biggest drop off in accuracy comes where K is equal to 3, so we will use that for our K-means clustering of the PCA below.

In [39]:

```
kmeans=KMeans(n_clusters=3) #Set a 3 KMeans clustering

X_clustered=kmeans.fit_predict(LA_HPI_fit_V2) #Compute cluster centers and predict clus

LABEL_COLOR_MAP = {0:'r', 1: 'g', 2: 'b'} #Define our own color map
label_color = [LABEL_COLOR_MAP[l] for l in X_clustered]

# Plot the scatter digram
plt.figure(figsize = (10,10))
plt.scatter(x_3d[:,0],x_3d[:,2], c=label_color, alpha=0.5)
plt.show()
```



3 Clusters formed from (3-Dimension) PCA data (above)

We also visualiza how these groups cluster together based on the different dimensions that were created from PCA, along with mapping how the clusters form on a map

In [69]:
```
# Create a temp dataframe from our PCA projection data "x_10d"
df=pd.DataFrame(x_3d)
df['X_cluster']=X_clustered
LA_HPI['Cluster']=X_clustered
```

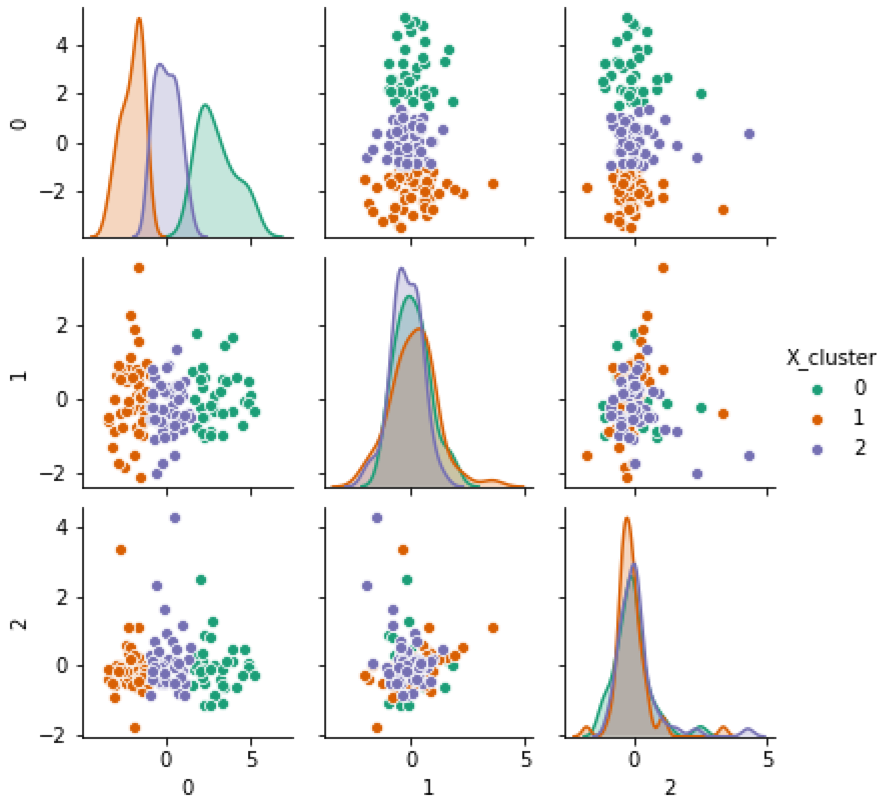
In [70]:
```
X_clustered # Our array of clusters that were formed
```

Out[70]:
```
array([2, 2, 1, 2, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 1, 0, 1, 1, 1, 2, 0, 1,
       2, 2, 0, 2, 2, 2, 1, 0, 0, 0, 0, 1, 1, 2, 2, 0, 1, 0, 0, 0, 0, 1,
       1, 1, 1, 0, 2, 2, 2, 2, 2, 2, 0, 0, 1, 1, 1, 2, 1, 2, 2, 1, 1, 1,
       1, 2, 1, 2, 1, 0, 1, 1, 0, 1, 1, 0, 2, 0, 1, 1, 0, 2, 2, 1, 0, 1,
       0, 2, 1, 0, 1, 1, 2, 1, 2, 1, 0, 2, 2, 0, 2, 1, 2, 2, 2, 1, 0, 2,
       2, 1, 2, 2, 1, 0, 2, 2, 0, 1, 2, 1, 1, 0, 2, 0, 1, 2, 2, 0, 0, 1,
       2, 0, 1, 0, 2, 0, 2, 1])
```

Our array of clusters that were formed (above)

In [42]:
```
# Call Seaborn's pairplot to visualize our feature interactions based on clusters
LA_County_PCA_plot = sns.pairplot(df, hue='X_cluster', palette= 'Dark2', diag_kind='kde
LA_County_PCA_plot;
```

```
C:\Users\marky\anaconda3\envs\geo_env\lib\site-packages\seaborn\axisgrid.py:2079: Use
rWarning: The `size` parameter has been renamed to `height`; please update your code.
  warnings.warn(msg, UserWarning)
```



In [43]:
```
LA_County_PCA_plot.savefig('LA County Pairplot PCA.PNG')
```

Map of our PCA data based on the clusters that were formed using k-means (above)

Results

After our clusters of groups have been created, then we place the cluster data into our earlier graphs to get a better understanding of how LA County is broken down.

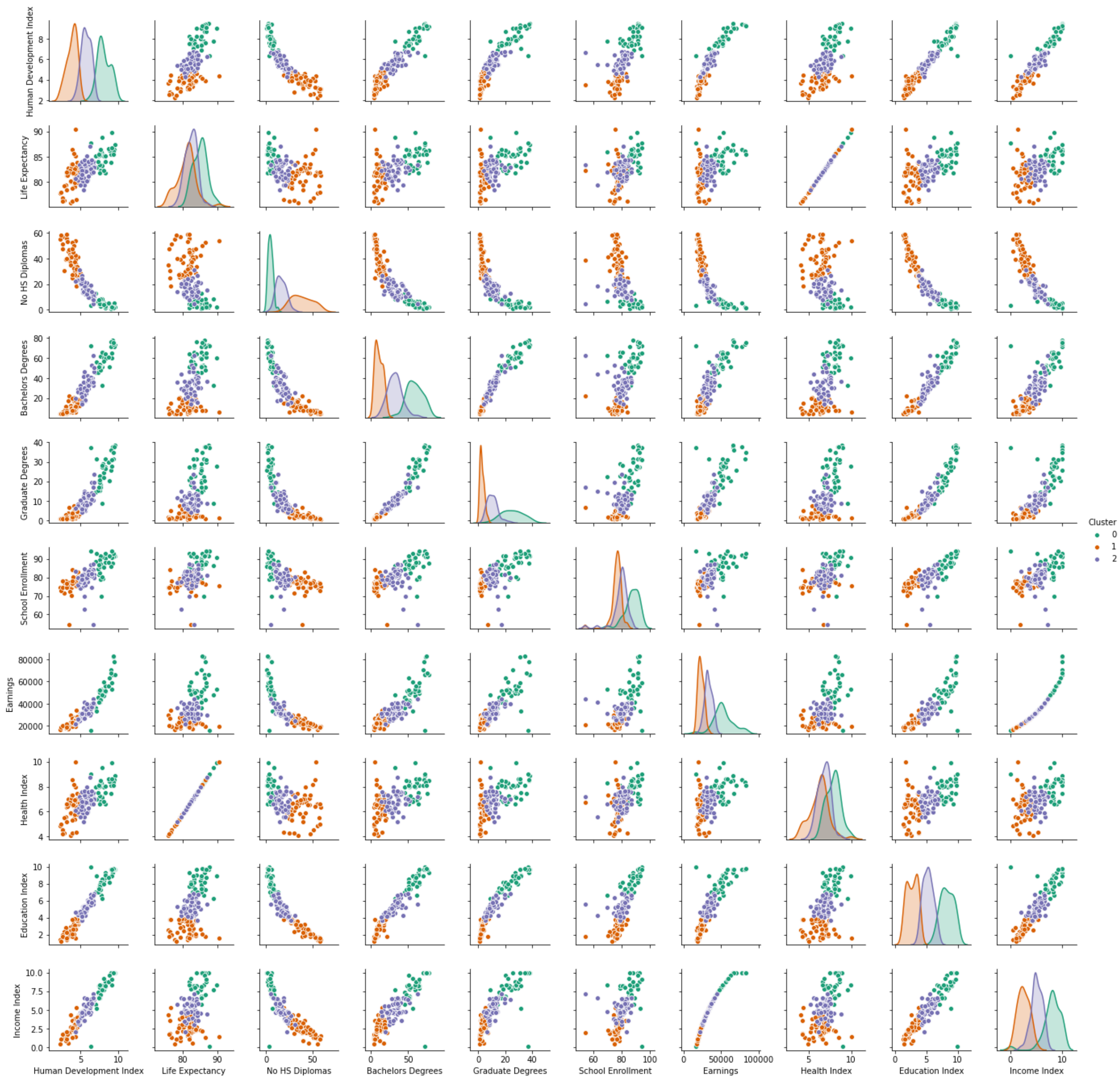
In [44]:

```python
# Call Seaborn's pairplot to visualize our KMeans clustering on the PCA projected data
sns.pairplot(LA_HPI, hue='Cluster', palette= 'Dark2', diag_kind='kde',size=1.85)
```

```
C:\Users\marky\anaconda3\envs\geo_env\lib\site-packages\seaborn\axisgrid.py:2079: UserWarning: The `size` parameter has been renamed to `height`; please update your code.
  warnings.warn(msg, UserWarning)
```

Out[44]:

```
<seaborn.axisgrid.PairGrid at 0x270b3ee2700>
```

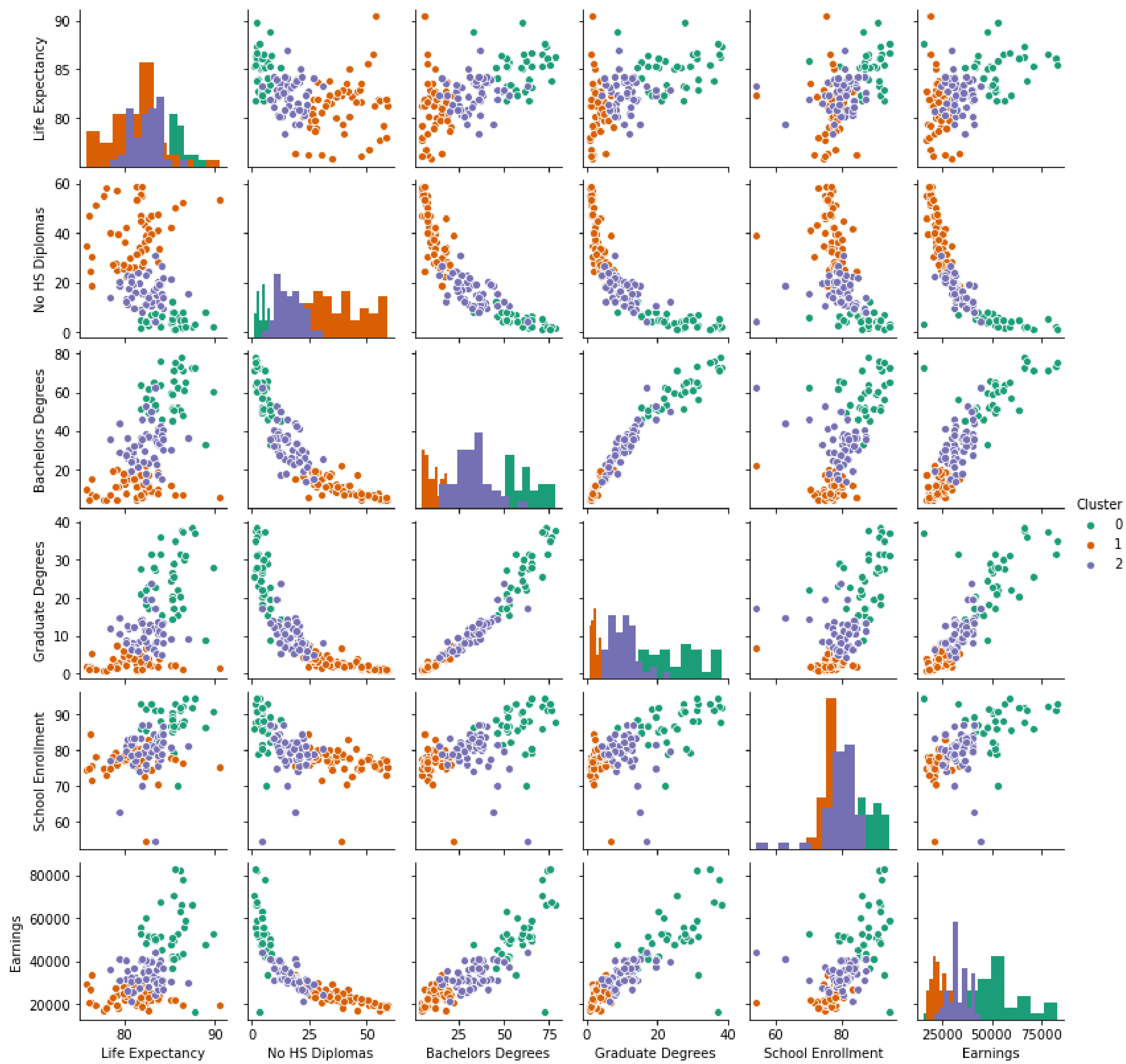


Our original dataframe grouped by clusters (above)

In [45]:
```
LA_HPI_V2['Cluster']=X_clustered
```

In [46]:
```
# Call Seaborn's pairplot to visualize our KMeans clustering on the PCA projected data
LA_County_PCA_model_plot = sns.pairplot(LA_HPI_V2, hue='Cluster', palette= 'Dark2', dia
LA_County_PCA_model_plot;
```

```
C:\Users\marky\anaconda3\envs\geo_env\lib\site-packages\seaborn\axisgrid.py:2079: Use
rWarning: The `size` parameter has been renamed to `height`; please update your code.
  warnings.warn(msg, UserWarning)
```



In [47]:
```
LA_County_PCA_model_plot.savefig('LA County Pairplot Model PCA.PNG')
```

Our refined dataframe grouped by clusters (above)

Upon our initial research for how the factors correlated to each other, we discovered an interesting relationship between 'school enrollment', 'earnings' and 'bachelors degrees' that could warrant further analysis.

To help faciliate further research, we grabbed location data from the top 3 popular places in each city using foursquare, and segmented by cluster below.

In [48]:

```
# Credentials and Parameters
# Private
```

In [49]:

```
VENUE_List=[]
# for loop for column rows
for i in range(len(LA_HPI)):
    CITY=LA_HPI['City'][i]
    CLUSTER=LA_HPI['Cluster'][i]
    CITIES=LA_HPI['City'][i].split(" - ")

# for loop for column items
    for j in range(len(CITIES)):
        NEAR=CITIES[j] +', CA'

        # create the API request URL
        url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret
            CLIENT_ID,
            CLIENT_SECRET,
            VERSION,
            NEAR,
            LIMIT,
            INTENT)

        # make the GET request
        results = requests.get(url).json()

        if results['meta']['code']==200:
            for k in range(LIMIT):
                # Save relevant field from results into a dataframe
                NAME=results['response']['groups'][0]['items'][k]['venue']['name']
                CATEGORY=results['response']['groups'][0]['items'][k]['venue']['categor
                LOCATION=results['response']['geocode']['where']
                AREA=CITY
                GROUP=CLUSTER+1
                VENUE=(NAME,CATEGORY,LOCATION,AREA,GROUP)
                VENUE_List.append(VENUE)
#                print(VENUE)
        else:
                j=j+1
```

In [50]:

```
## Select columns for dataframe to download results
Venue_Columns=('Name','Category','City','Area','Group')
# Convert list to dataframe, add columns
df_VENUE_List=pd.DataFrame(VENUE_List,columns=Venue_Columns)
# Formate 'city' column dataframe within the dataframe by Capitalizing it and removing
df_VENUE_List['City']=df_VENUE_List['City'].str.title().str.rstrip(' Ca')
# Save results into a csv
df_VENUE_List.to_csv('LA_County_Venue_List.csv')
```

Venue Location Dataframe

In [51]:

```python
# Read from csv
df_VENUE_List_File=pd.read_csv('LA_County_Venue_List.csv')
df_VENUE_List_File.drop(columns='Unnamed: 0')
```

Out[51]:

	Name	Category	City	Area	Group
0	Moby's Coffee & Tea Company	Coffee Shop	North Hollywood	North Hollywood - Valley Village	3
1	Movement Lifestyle Studio	Dance Studio	North Hollywood	North Hollywood - Valley Village	3
2	Trader Joe's	Grocery Store	North Hollywood	North Hollywood - Valley Village	3
3	Frends Beauty	Cosmetics Shop	Valley Village	North Hollywood - Valley Village	3
4	Gelson's	Grocery Store	Valley Village	North Hollywood - Valley Village	3
5	Miya Sushi	Sushi Restaurant	Valley Village	North Hollywood - Valley Village	3
6	Aquarium City	Pet Store	Canoga Park	Canoga Park - Winnetka - Woodland Hills - West...	3
7	Pastries By Edie	Café	Canoga Park	Canoga Park - Winnetka - Woodland Hills - West...	3
8	Pho 21	Asian Restaurant	Canoga Park	Canoga Park - Winnetka - Woodland Hills - West...	3
9	Brent's Deli	Deli / Bodega	Winnetk	Canoga Park - Winnetka - Woodland Hills - West...	3
10	Costco Food Court	Food Court	Winnetk	Canoga Park - Winnetka - Woodland Hills - West...	3
11	Panini Cafe	Sandwich Place	Winnetk	Canoga Park - Winnetka - Woodland Hills - West...	3
12	Topanga Canyon Hills	Scenic Lookout	Woodland Hills	Canoga Park - Winnetka - Woodland Hills - West...	3
13	Calabasas Farmer's Market	Farmers Market	Woodland Hills	Canoga Park - Winnetka - Woodland Hills - West...	3
14	Health Nut	Salad Place	Woodland Hills	Canoga Park - Winnetka - Woodland Hills - West...	3
15	El Pollo Amigo	Mexican Restaurant	West Hills	Canoga Park - Winnetka - Woodland Hills - West...	3
16	Yozen Frogurt	Ice Cream Shop	West Hills	Canoga Park - Winnetka - Woodland Hills - West...	3
17	Starbucks	Coffee Shop	West Hills	Canoga Park - Winnetka - Woodland Hills - West...	3
18	786 Degrees Wood Fired Pizza Co.	Pizza Place	Sun Valley	Sun Valley - La Tuna Canyon	2
19	Softline Solutions	Office	Sun Valley	Sun Valley - La Tuna Canyon	2
20	In-N-Out Burger	Fast Food Restaurant	Sun Valley	Sun Valley - La Tuna Canyon	2
21	Old Time Drive In	American Restaurant	La Tuna Canyon	Sun Valley - La Tuna Canyon	2
22	Rise N Shine Café	Breakfast Spot	La Tuna Canyon	Sun Valley - La Tuna Canyon	2
23	Grocery Outlet	Grocery Store	La Tuna Canyon	Sun Valley - La Tuna Canyon	2
24	Long Beach Creamery	Ice Cream Shop	Wilmington	Wilmington - Harbor City	2

	Name	Category	City	Area	Group
25	555 East American Steakhouse	Steakhouse	Wilmington	Wilmington - Harbor City	2
26	The Lions Lighthouse for Sight	Lighthouse	Wilmington	Wilmington - Harbor City	2
27	Cliffs of Palos Verdes	Scenic Lookout	Harbor City	Wilmington - Harbor City	2
28	La Española Meats	Spanish Restaurant	Harbor City	Wilmington - Harbor City	2
29	Pacific Coast Hobbies	Hobby Shop	Harbor City	Wilmington - Harbor City	2
30	La Michoacana	Snack Place	Mission Hills	Mission Hills - Panorama City - North Hills	2
31	In-N-Out Burger	Fast Food Restaurant	Mission Hills	Mission Hills - Panorama City - North Hills	2
32	Ay Papa Que Rico	Cuban Restaurant	Mission Hills	Mission Hills - Panorama City - North Hills	2
33	In-N-Out Burger	Fast Food Restaurant	Panorama City	Mission Hills - Panorama City - North Hills	2
34	La Sirenita Restaurant	Seafood Restaurant	Panorama City	Mission Hills - Panorama City - North Hills	2
35	Chipotle Mexican Grill	Mexican Restaurant	Panorama City	Mission Hills - Panorama City - North Hills	2
36	La Sirenita Restaurant	Seafood Restaurant	North Hills	Mission Hills - Panorama City - North Hills	2
37	Stinking Crawfish	Seafood Restaurant	North Hills	Mission Hills - Panorama City - North Hills	2
38	Rincon Taurino	Mexican Restaurant	North Hills	Mission Hills - Panorama City - North Hills	2
39	Vista Hermosa Park	Park	Westlake	Westlake	2
40	The Theatre at the Ace	Theater	Westlake	Westlake	2
41	Whole Foods Market	Grocery Store	Westlake	Westlake	2

Cluster 1

In [52]:

```
LA_Cluster_Data_1=LA_HPI[LA_HPI['Cluster']==0].mean()
df_VENUE_List_File.loc[df_VENUE_List_File['Group'] == 1]
```

Out[52]:

Unnamed: 0	Name	Category	City	Area	Group

Cluster 2

In [53]:
```
LA_Cluster_Data_2=LA_HPI[LA_HPI['Cluster']==1].mean()
df_VENUE_List_File.loc[df_VENUE_List_File['Group'] == 2]
```

Out[53]:

	Unnamed: 0	Name	Category	City	Area	Group
18	18	786 Degrees Wood Fired Pizza Co.	Pizza Place	Sun Valley	Sun Valley - La Tuna Canyon	2
19	19	Softline Solutions	Office	Sun Valley	Sun Valley - La Tuna Canyon	2
20	20	In-N-Out Burger	Fast Food Restaurant	Sun Valley	Sun Valley - La Tuna Canyon	2
21	21	Old Time Drive In	American Restaurant	La Tuna Canyon	Sun Valley - La Tuna Canyon	2
22	22	Rise N Shine Café	Breakfast Spot	La Tuna Canyon	Sun Valley - La Tuna Canyon	2
23	23	Grocery Outlet	Grocery Store	La Tuna Canyon	Sun Valley - La Tuna Canyon	2
24	24	Long Beach Creamery	Ice Cream Shop	Wilmington	Wilmington - Harbor City	2
25	25	555 East American Steakhouse	Steakhouse	Wilmington	Wilmington - Harbor City	2
26	26	The Lions Lighthouse for Sight	Lighthouse	Wilmington	Wilmington - Harbor City	2
27	27	Cliffs of Palos Verdes	Scenic Lookout	Harbor City	Wilmington - Harbor City	2
28	28	La Española Meats	Spanish Restaurant	Harbor City	Wilmington - Harbor City	2
29	29	Pacific Coast Hobbies	Hobby Shop	Harbor City	Wilmington - Harbor City	2
30	30	La Michoacana	Snack Place	Mission Hills	Mission Hills - Panorama City - North Hills	2
31	31	In-N-Out Burger	Fast Food Restaurant	Mission Hills	Mission Hills - Panorama City - North Hills	2
32	32	Ay Papa Que Rico	Cuban Restaurant	Mission Hills	Mission Hills - Panorama City - North Hills	2
33	33	In-N-Out Burger	Fast Food Restaurant	Panorama City	Mission Hills - Panorama City - North Hills	2
34	34	La Sirenita Restaurant	Seafood Restaurant	Panorama City	Mission Hills - Panorama City - North Hills	2
35	35	Chipotle Mexican Grill	Mexican Restaurant	Panorama City	Mission Hills - Panorama City - North Hills	2
36	36	La Sirenita Restaurant	Seafood Restaurant	North Hills	Mission Hills - Panorama City - North Hills	2
37	37	Stinking Crawfish	Seafood Restaurant	North Hills	Mission Hills - Panorama City - North Hills	2
38	38	Rincon Taurino	Mexican Restaurant	North Hills	Mission Hills - Panorama City - North Hills	2
39	39	Vista Hermosa Park	Park	Westlake	Westlake	2
40	40	The Theatre at the Ace	Theater	Westlake	Westlake	2
41	41	Whole Foods Market	Grocery Store	Westlake	Westlake	2

Cluster 3

In [54]:
```
LA_Cluster_Data_3=LA_HPI[LA_HPI['Cluster']==2].mean()
df_VENUE_List_File.loc[df_VENUE_List_File['Group'] == 3]
```

Out[54]:

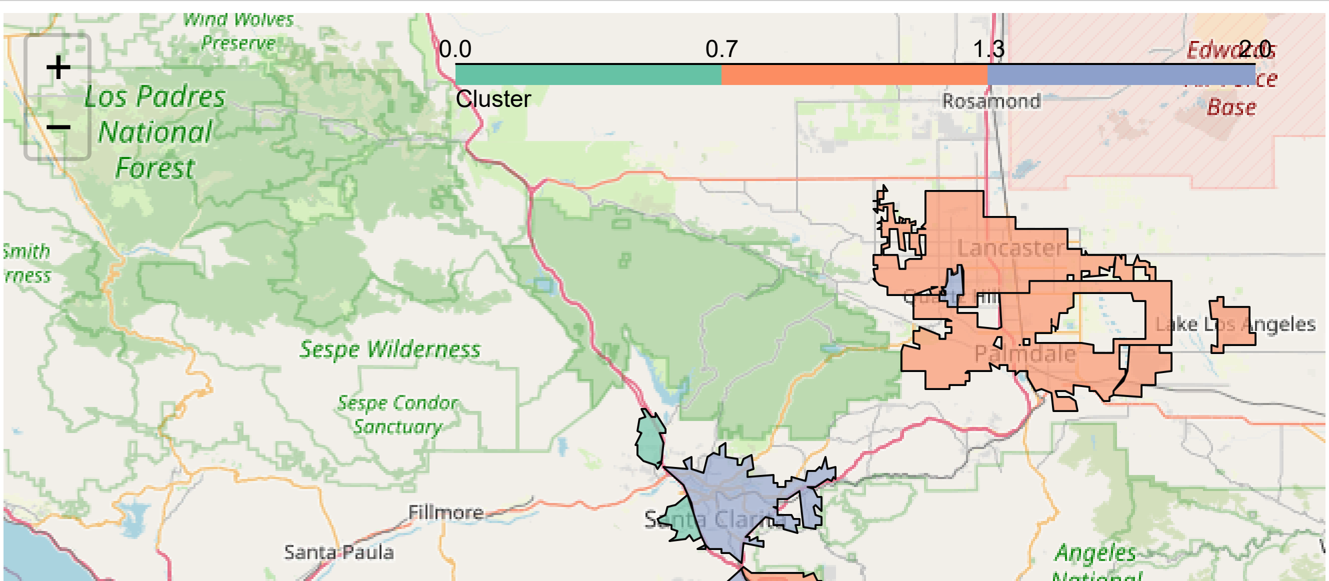
	Unnamed: 0	Name	Category	City	Area	Group
0	0	Moby's Coffee & Tea Company	Coffee Shop	North Hollywood	North Hollywood - Valley Village	3
1	1	Movement Lifestyle Studio	Dance Studio	North Hollywood	North Hollywood - Valley Village	3
2	2	Trader Joe's	Grocery Store	North Hollywood	North Hollywood - Valley Village	3
3	3	Frends Beauty	Cosmetics Shop	Valley Village	North Hollywood - Valley Village	3
4	4	Gelson's	Grocery Store	Valley Village	North Hollywood - Valley Village	3
5	5	Miya Sushi	Sushi Restaurant	Valley Village	North Hollywood - Valley Village	3
6	6	Aquarium City	Pet Store	Canoga Park	Canoga Park - Winnetka - Woodland Hills - West...	3
7	7	Pastries By Edie	Café	Canoga Park	Canoga Park - Winnetka - Woodland Hills - West...	3
8	8	Pho 21	Asian Restaurant	Canoga Park	Canoga Park - Winnetka - Woodland Hills - West...	3
9	9	Brent's Deli	Deli / Bodega	Winnetk	Canoga Park - Winnetka - Woodland Hills - West...	3
10	10	Costco Food Court	Food Court	Winnetk	Canoga Park - Winnetka - Woodland Hills - West...	3
11	11	Panini Cafe	Sandwich Place	Winnetk	Canoga Park - Winnetka - Woodland Hills - West...	3
12	12	Topanga Canyon Hills	Scenic Lookout	Woodland Hills	Canoga Park - Winnetka - Woodland Hills - West...	3
13	13	Calabasas Farmer's Market	Farmers Market	Woodland Hills	Canoga Park - Winnetka - Woodland Hills - West...	3
14	14	Health Nut	Salad Place	Woodland Hills	Canoga Park - Winnetka - Woodland Hills - West...	3
15	15	El Pollo Amigo	Mexican Restaurant	West Hills	Canoga Park - Winnetka - Woodland Hills - West...	3
16	16	Yozen Frogurt	Ice Cream Shop	West Hills	Canoga Park - Winnetka - Woodland Hills - West...	3
17	17	Starbucks	Coffee Shop	West Hills	Canoga Park - Winnetka - Woodland Hills - West...	3

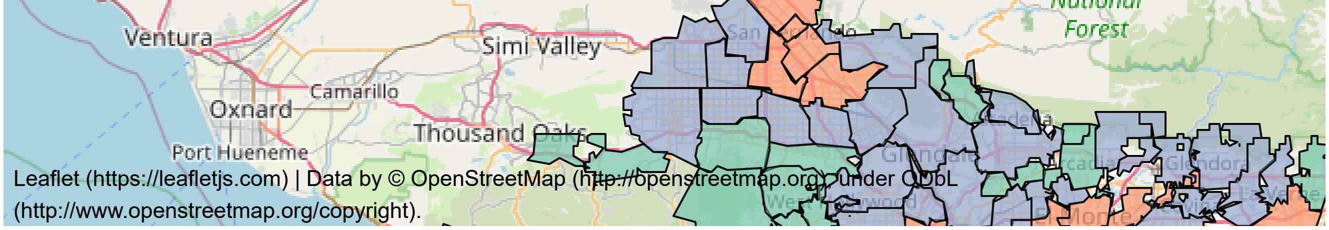
Cluster Map

In [72]:

```
# Converting 'Polygon' column from dataframe into geodataframe for plotting
LA_HPI_gdf=gpd.GeoDataFrame(LA_HPI,geometry='Polygon')
LA_HPI_gdf_json=LA_HPI_gdf.to_json() # Convert from geodataframe to json for choropleth

Cluster_Geo=['City','Cluster']

# Initialize the map:
map_LA_County = folium.Map([latitude, longitude], zoom_start=9)

choropleth=folium.Choropleth(
    geo_data=LA_HPI_gdf_json,
    name='choropleth',
    data=LA_HPI[Cluster_Geo],
    columns=Cluster_Geo,
    key_on='feature.properties.City',
    bins=3,
    fill_color='Set2',
    fill_opacity=0.7,
    line_opacity=1.2,
    legend_name='Cluster',
    highlight=True
).add_to(map_LA_County)
choropleth.geojson.add_child(
    folium.features.GeoJsonTooltip(['City'],labels=False)
)

choropleth=folium.features.GeoJson(
    LA_HPI_gdf_json,
    style_function=style_function,
    control=False,
    highlight_function=highlight_function,
    tooltip=folium.features.GeoJsonTooltip(
        fields=['City','Cluster','Human Development Index', 'Life Expectancy', 'No HS D
       'School Enrollment', 'Earnings', 'Health Index', 'Education Index', 'Income Inde
                ],
        style=("background-color: white; color: #333333; font-family: arial; font-size:
    )
)
map_LA_County.add_child(choropleth)


map_LA_County
```

Out[72]:





Looking at the map, we see a clear positive relationship between higher performing cities and their proximiting to the ocean. We also see that inner cities regions with Los Angeles and the San Fernando Valley are the worst performers, to go along with the Lancaster region. There also seem to be pockets of higher former cities in pockers of more mountain areas as well.

In [68]:

```
LA_Clusters=[]
LA_Clusters=pd.concat([LA_Cluster_Data_1,LA_Cluster_Data_2,LA_Cluster_Data_3],axis=1)
LA_Clusters.sort_values(by='Human Development Index',axis=1,inplace=True)
LA_Clusters=LA_Clusters.transpose().reset_index(drop=True)
LA_Clusters=LA_Clusters.drop(columns = {'Cluster'})
LA_Clusters = round(LA_Clusters, 2)
# LA_Clusters = LA_Clusters.sort_index()
LA_Clusters
```

Out[68]:

	Human Development Index	Life Expectancy	No HS Diplomas	Bachelors Degrees	Graduate Degrees	School Enrollment	Earnings	Health Index	Education Index	In
0	3.87	81.09	39.00	11.34	2.87	76.75	23251.83	6.29	2.73	
1	5.74	82.37	16.33	32.56	10.60	79.57	33352.08	6.82	5.29	
2	8.09	85.04	4.81	59.40	25.39	87.53	53037.40	7.93	8.33	

From the breakdown of the averages for the different groups above, we see the least disparity in life expectancy and school enrollment, while we the highest disparity is seen in no HS diplomas, graduate degrees and earnings.

Discussion

Now that we grouped the cities within LA counties into clusters and have seen how they are plotted out on a map, it is very interesting to see how the different clusters seemed to be grouped throughout the area. There seems to be an obvious association between highest performing cities and their proximity to the ocean, but we also see highest performing cities among mountain regions which would be interesting to explore from an age perspective to see if this is representative of migration patterns within LA County. It's also worth noting how close the different city clusters are in their school enrollment levels, while there is a fair amount of discrepancy in other categories. This could also be worth further explaination in the form of creating a logistical regression model, and also seeing if this is a result of the quality of education in various regions or if it is a results of cities not retaining their citizens once they have become educated and involved in the workforce.

Conclusion

From the results of our studies, it seems like there could be a lot of good information to further explore education effective and migration patterns within LA to see how they effect earnings and graduation rates. An imporant question to ask is are higher performing areas offering better education and/or are higher earning individuals moving to these areas once they've reached a certain level of income. While this dataset was limited to factors that related to health, income and education -- we are fortunate that LA County has a great amount of dataset available that can evaluated under a similar model to help with other classification tasks. Once we understand the different clusters and where their greatest opportunities for improvements are, we can use these clusters to develop benchmarks and allocate resources where they will 'move the needle' the most.