1. *Changes Over Time:* I hypothesize that there will be significant differences in my educational data between the years 2021 and 2022, possibly reflecting changes in school demographics, performance, or other factors.

```
In [57]: df21 = data[data['School\nYear'] == 2021]
         df22 = data[data['School\nYear'] == 2022]

In [58]: df21
```

Out[58]:

	School\nYear	State	Region 13	District	African\nAmerican	Hispanic	White	American\nIndian	Asian	Pacific\nIslander	Two or\nMore\nRaces	Special\nEd\n(Curre
2	2021.0	0.67	0.70	0.87	89%	81%	80%	*	92%	-	100%	6
4	2021.0	0.39	0.44	0.53	78%	31%	47%	*	65%	-	43%	2
6	2021.0	0.19	0.24	0.26	11%	8%	20%	*	42%	-	14%	
9	2021.0	0.62	0.63	0.69	44%	54%	59%	*	90%	-	50%	3
11	2021.0	0.31	0.32	0.35	11%	25%	16%	*	63%	-	0%	1
...	...	...	...	...	...	...	...	...	...	...	...	
236	2021.0	0.43	0.44	0.51	40%	61%	36%	-	63%	-	83%	3
239	2021.0	0.26	0.29	0.46	26%	30%	38%	*	64%	*	61%	1
241	2021.0	0.28	0.30	0.46	29%	35%	37%	*	65%	*	63%	1
243	2021.0	0.41	0.44	0.65	63%	53%	59%	*	76%	*	83%	3
245	2021.0	0.37	0.39	0.54	34%	44%	45%	*	73%	*	64%	3

103 rows × 17 columns

```
In [59]: df22
```

Out[59]:

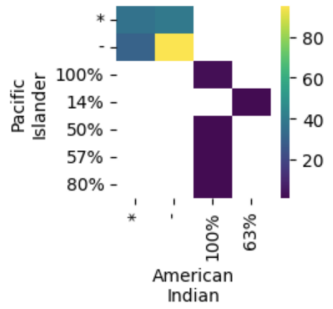
	School\nYear	State	Region 13	District	African\nAmerican	Hispanic	White	American\nIndian	Asian	Pacific\nIslander	Two or\nMore\nRaces	Special\nEd\n(Curre
1	2022.0	0.76	0.78	0.97	91%	91%	97%	*	100%	*	88%	9
3	2022.0	0.51	0.56	0.81	64%	64%	82%	*	89%	*	88%	5
5	2022.0	0.30	0.35	0.60	55%	36%	55%	*	73%	*	50%	2
8	2022.0	0.71	0.72	0.91	82%	73%	92%	*	98%	*	88%	7
10	2022.0	0.43	0.45	0.68	64%	45%	53%	*	89%	*	25%	5
...	...	...	...	...	...	...	...	...	...	...	...	
235	2022.0	0.48	0.50	0.81	87%	69%	75%	-	91%	-	86%	4
238	2022.0	0.34	0.37	0.57	36%	36%	49%	*	80%	*	42%	2
240	2022.0	0.36	0.39	0.61	44%	40%	54%	*	81%	*	51%	2
242	2022.0	0.53	0.57	0.78	64%	65%	78%	*	87%	*	75%	4
244	2022.0	0.43	0.45	0.66	52%	47%	56%	*	87%	*	57%	3

2. *Demographic Differences:* I anticipate that there will be notable differences in some variables (potentially related to school performance or enrollment) between the 'American Indian' and 'Pacific Islander groups in my dataset. I attempted to visualize this through a heatmap.

```
def heatmap(df, x_colname, y_colname, figsize=(2, 2), mpl_palette_name='viridis'):
  from matplotlib import pyplot as plt
  import seaborn as sns
  import pandas as pd
  plt.subplots(figsize=figsize)
  df_2dhist = pd.DataFrame({
      x_label: grp[y_colname].value_counts()
      for x_label, grp in df.groupby(x_colname)
  })
  sns.heatmap(df_2dhist, cmap=mpl_palette_name)
  plt.xlabel(x_colname)
  plt.ylabel(y_colname)
  return autoviz.MplChart.from_current_mpl_state()

chart = heatmap(df_6752476028165918961, *['American\nIndian', 'Pacific\nIslander'], **{})
chart
```

Out[69]:

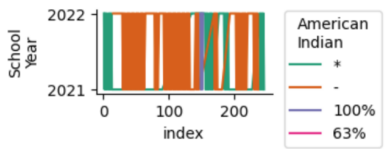



```
    ys = counted['counts']
  else:
    xs = series[timelike_colname]
    ys = series[value_colname]
  plt.plot(xs, ys, label=series_name, color=palette[series_index % len(palette)])

fig, ax = plt.subplots(figsize=figsize, layout='constrained')
df = df.sort_values(timelike_colname, ascending=True)
if series_colname:
  for i, (series_name, series) in enumerate(df.groupby(series_colname)):
    _plot_series(series, series_name, i)
  fig.legend(title=series_colname, bbox_to_anchor=(1, 1), loc='upper left')
else:
  _plot_series(df, '')
sns.despine(fig=fig, ax=ax)
plt.xlabel(timelike_colname)
plt.ylabel(value_colname)
return autoviz.MplChart.from_current_mpl_state()

chart = time_series_multiline(df_5016765184939975526, *['index', 'School\nYear', 'American\nIndian'], **{})
chart
```

Out[68]:



3. *Geographical Variations:* Since 'State', 'Region 13', and 'District' appear to be geographical identifiers, I hypothesize that there will be meaningful differences in the data based on these geographical divisions. I attempted to explore the relationships between these variables and another numerical variable using a 3D visualization.

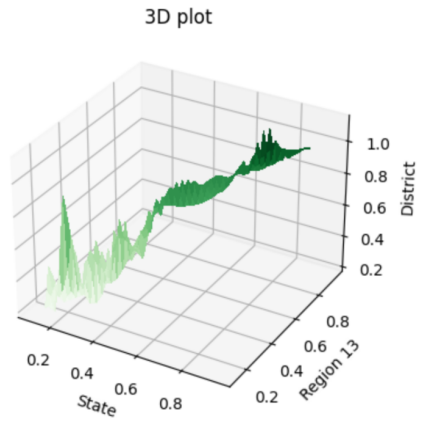
In [50]:
```
data['State'] = data['State'].str.strip('%').astype(float)/100
data['Region 13'] = data['Region 13'].str.strip('%').astype(float)/100
data['District'] = data['District'].str.strip('%').astype(float)/100
data
```

Out[50]:

	School\nYear	State	Region 13	District	African\nAmerican	Hispanic	White	American\nIndian	Asian	Pacific\nIslander	Two or\nMore\nRaces	Special\nEd\n(Curre
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
1	2022.0	0.76	0.78	0.97	91%	91%	97%	*	100%	*	88%	9
2	2021.0	0.67	0.70	0.87	89%	81%	80%	*	92%	-	100%	6
3	2022.0	0.51	0.56	0.81	64%	64%	82%	*	89%	*	88%	5
4	2021.0	0.39	0.44	0.53	78%	31%	47%	*	65%	-	43%	2
...	...	...	...	...	...	...	...	...	...	...	...	
241	2021.0	0.28	0.30	0.46	29%	35%	37%	*	65%	*	63%	1
242	2022.0	0.53	0.57	0.78	64%	65%	78%	*	87%	*	75%	4
243	2021.0	0.41	0.44	0.65	63%	53%	59%	*	76%	*	83%	3
244	2022.0	0.43	0.45	0.66	52%	47%	56%	*	87%	*	57%	3
245	2021.0	0.37	0.39	0.54	34%	44%	45%	*	73%	*	64%	3

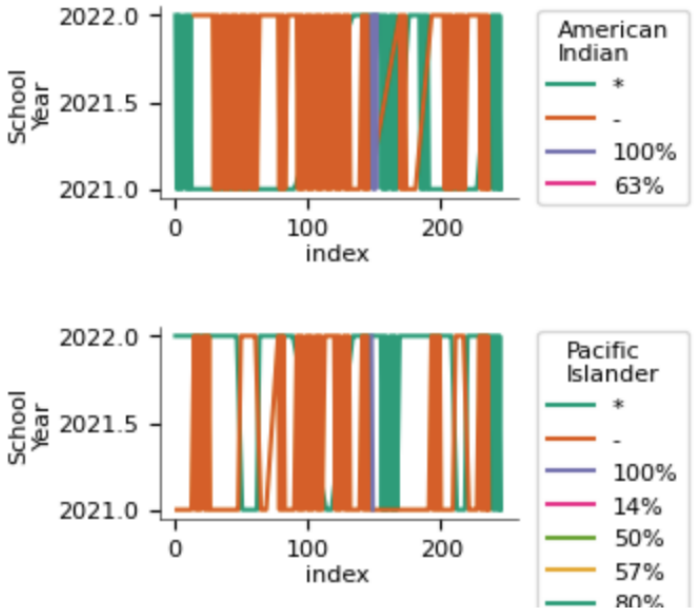
240 rows × 17 columns

```
fig=plt.figure()
ax = fig.add_axes(Axes3D(fig))
surf=ax.plot_surface(xi,yi,zi,cmap='Greens',linewidth=0,antialiased=False)
fig.colorbar(surf, location = 'bottom')
ax.set_title('3D plot')
ax.set_xlabel('State')
ax.set_ylabel('Region 13')
ax.set_zlabel('District')
plt.show()
```



4. *Time-Series Trends:* Given the function I created for multi-line time series visualization, I expect to find that trends over time differ based on some categorizing variable in my dataset (possibly the different demographic groups).

Time series


2022.0
2021.5
2021.0
School Year
0
100
200
index
American Indian
*
-
100%
63%

2022.0
2021.5
2021.0
School Year
0
100
200
index
Pacific Islander
*
-
100%
14%
50%
57%
80%