Mini project 2: primary productivity in coastal waters

Amy Nguyen

In this project you're again given a dataset and some questions. The data for this project come from the EPA's National Aquatic Resource Surveys, and in particular the National Coastal Condition Assessment (NCCA); broadly, you'll do an exploratory analysis of primary productivity in coastal waters.

By way of background, chlorophyll A is often used as a proxy for primary productivity in marine ecosystems; primary producers are important because they are at the base of the food web. Nitrogen and phosphorus are key nutrients that stimulate primary production.

In the data folder you'll find water chemistry data, site information, and metadata files. It might be helpful to keep the metadata files open when tidying up the data for analysis. It might also be helpful to keep in mind that these datasets contain a considerable amount of information, not all of which is relevant to answering the questions of interest. Notice that the questions pertain somewhat narrowly to just a few variables. It's recommended that you determine which variables might be useful and drop the rest.

As in the first mini project, there are accurate answers to each question that are mutually consistent with the data, but there aren't uniquely correct answers. You will likely notice that you have even more latitude in this project than in the first, as the questions are slightly broader. Since we've been emphasizing visual and exploratory techniques in class, you are encouraged (but not required) to support your answers with graphics.

The broader goal of these mini projects is to cultivate your problem-solving ability in an unstructured setting. Your work will be evaluated based on the following:

- choice of method(s) used to answer questions;
- clarity of presentation;
- code style and documentation.

Please write up your results separately from your codes; codes should be included at the end of the notebook.

Part 1: dataset

Merge the site information with the chemistry data and tidy it up. Determine which columns to keep based on what you use in answering the questions in part 2; then, print the first few rows here (but *do not include your codes used in tidying the data*) and write a brief description (1-2 paragraphs) of the dataset conveying what you take to be the key attributes. Direct your description to a reader unfamiliar with the data; ensure that in your data preview the columns are named intelligibly.

Suggestion: export your cleaned data as a separate `.csv` file and read that directly in below, as in: `pd.read_csv('YOUR DATA FILE').head()`.

In [3]:
```
# show a few rows of clean data
ncca.head()
```

Out[3]:

	Unique Site ID	Waterbody Name	State	Region	Date Collected	Chlorophyll A (ug/L)	Nitrogen (mg N/L)	Phosphorus (mg P/L)
0	59	Mission Bay	CA	West Coast	1-Jul-10	3.34	0.40750	0.061254
1	60	San Diego Bay	CA	West Coast	1-Jul-10	2.45	0.23000	0.037379
2	61	Mission Bay	CA	West Coast	1-Jul-10	3.82	0.33625	0.048100
3	62	San Diego Bay	CA	West Coast	1-Jul-10	6.13	0.23875	0.044251
4	63	White Oak River	NC	East Coast	9-Jun-10	9.79	0.63250	0.090636

For my tidy dataset `ncca`, I dediced to only keep the variables `Unique Site ID`, `Waterbody Name`, `State`, `Region`, `Date Collected`, `Chlorophyll A (ug/L)`, `Nitrogen (mg N/L)`, and `Phosphorus (mg P/L)` for a total of 1092 observations and 8 variables. Each row in the dataframe represents a single record for one Unique Site ID where the water chemistry recordings were taken. Chlorophyll A will be the variable measuring primary productivity, while nitrogen and phosphorus corresponding to each Site ID represent the nutrient concentrations for each Site ID.

Part 2: exploratory analysis

Answer each question below and provide a visualization supporting your answer. A description and interpretation of the visualization should be offered.

Comment: you can either designate your plots in the codes section with clear names and reference them in your answers; or you can export your plots as image files and display them in markdown cells.

What is the apparent relationship between nutrient availability and productivity?

Comment: it's fine to examine each nutrient -- nitrogen and phosphorus -- separately, but do consider whether they might be related to each other.

There appears to be a positively linear relationship between nutrient availability and productivity. After grouping by waterbody names and calculating the average concentration levels for Chlorophyll A, nitrogen, and phosphorus, we see in Fig1a that as the average nutrient concentrations increase, so does Chlorophyll A production. Fig1a also indicates that nitrogen is more widely available than phosphorus as the average phosphorus concentration does not exceed 0.5mg/L, unlike nitrogen which goes up to an average concentration of about 2.4mg/L.

To better understand the relationship between nitrogen and phosphorus availability, Fig1b displays average phosphorus concentration against average nitrogen concentration. An increase in nitrogen levels corresponds to an increase in phosphorus levels, indicating a positively linear relationship between the two nutrients.

Are there any notable differences in available nutrients among U.S. coastal regions?

Fig2b displays a bar chart of the average nutrient availability by coastal region. Based on the plot, the Gulf Coast has the richest availability of nutrients for both nitrogen and phosphorus. The Great Lakes has the least amount of available phosphorus and the west coast has the least amount of available nitrogen. The spread of each of the coastal regions' available nutrients can be visualized in Fig2a.

Based on the 2010 data, does productivity seem to vary geographically in some way?

If so, explain how; If not, explain what options you considered and ruled out.

Fig2a shows that productivity does vary geographically. The East coast has the largest spread and also the highest levels of Chlorophyll A and also has ample concentrations of nutrient concentrations. In contrast, the west coast has lower levels of nutrient concentrations which in turn leads to lower levels of Chlorophyll A.

How does primary productivity in California coastal waters change seasonally in 2010, if at all?

Does your result make intuitive sense?

For this dataset, the collection dates range from June till September which technically all falls in the summer season. However, when plotting productivity for each month in June, July, August, and September, (fig3) we see that July and August have the most Chlorophyll A productivity which makes is intuitive since these are the warmest months.

What is the distribution of productivity for each coastal region?

The density estimates of productivity shown in fig4 display the distributions of Chlorophyll A for each coastal region. The Great Lakes is skewed right and most of the Unique ID sites have low Chlorophyll A production with nearly 35% of The Great Lakes observations having a Chlorophyll measure below 1. The west coast is also right skewed and its density estimate tapers off at a Chlorophyll level around 14ug/L, but its most common Chlorophyll measurement is about 3ug/L. The east coast is has the 3rd most common readings for productivity and while the Gult Coast has the lowest density estimates, it is because this region has the widest range of values for Chlorophyll which are subsequently also higher measures of Chlorophyll than the other regions.

Codes

In [2]:

```
import pandas as pd
import numpy as np
import altair as alt

ncca_raw = pd.read_csv('assessed_ncca2010_waterchem.csv')
ncca_sites = pd.read_csv('assessed_ncca2010_siteinfo.csv')

# merge ncca_raw and ncca_sites
merged_data = pd.merge(
    ncca_sites, ncca_raw, how = 'right',
    on = ['UID', 'STATE']
)
# create dict of new column names
new_names = {'UID': 'Unique Site ID', 'WTBDY_NM': 'Waterbody Name',
             'STATE': 'State',
             'NCA_REGION': 'Region', 'DATE_COL_x': 'Date Collected',
             'Chlorophyll A': 'Chlorophyll A (ug/L)',
             'Total Nitrogen': 'Nitrogen (mg N/L)',
             'Total Phosphorus': 'Phosphorus (mg P/L)'}

# specify column order
name_order = ['Unique Site ID', 'Waterbody Name','State',
              'Region', 'Date Collected', 'Chlorophyll A (ug/L)',
              'Nitrogen (mg N/L)', 'Phosphorus (mg P/L)']

# pivot, rename, and
ncca = merged_data.pivot(
    index = ['UID', 'WTBDY_NM', 'STATE', 'NCA_REGION', 'DATE_COL_x'],
    columns = 'PARAMETER_NAME',
    values = 'RESULT'
).reset_index().rename_axis(
    columns=None
).rename(
    columns = new_names
).loc[:, name_order]
```

In [4]:

```
# group by Waterbody Name and aggregate mean
ncca_agg = ncca.groupby(
    ['Waterbody Name', 'Region']
).mean().reset_index()

ncca_agg.head()
```

Out[4]:

	Waterbody Name	Region	Unique Site ID	Chlorophyll A (ug/L)	Nitrogen (mg N/L)	Phosphorus (mg P/L)
0	Alazan Bay	Gulf Coast	1542.0	12.760000	0.882500	0.143675
1	Albermarle Sound	East Coast	3390.5	24.461667	0.597187	0.032193
2	Alligator River	East Coast	339.0	4.040000	0.793500	0.024905
3	Alsea Bay	West Coast	618.0	6.640000	0.501250	0.072810
4	Anclote Anchorage	Gulf Coast	249.0	1.270000	0.372500	0.008185

In [5]:

```
# melt for plotting
agg_plot_df = ncca_agg.melt(
    id_vars = ['Waterbody Name', 'Region', 'Chlorophyll A (ug/L)'],
    value_vars = ['Nitrogen (mg N/L)', 'Phosphorus (mg P/L)'],
    var_name = 'Nutrient',
    value_name = 'Average Nutrient Concentration'
)

agg_plot_df.head()
```

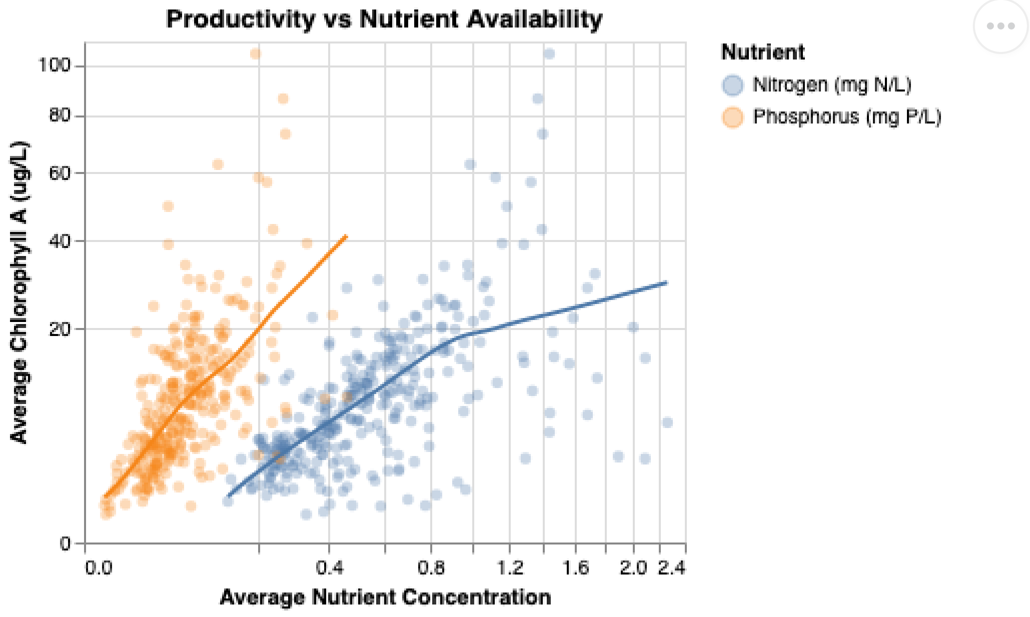
Out[5]:

	Waterbody Name	Region	Chlorophyll A (ug/L)	Nutrient	Average Nutrient Concentration
0	Alazan Bay	Gulf Coast	12.760000	Nitrogen (mg N/L)	0.882500
1	Albermarle Sound	East Coast	24.461667	Nitrogen (mg N/L)	0.597187
2	Alligator River	East Coast	4.040000	Nitrogen (mg N/L)	0.793500
3	Alsea Bay	West Coast	6.640000	Nitrogen (mg N/L)	0.501250
4	Anclote Anchorage	Gulf Coast	1.270000	Nitrogen (mg N/L)	0.372500

In [6]:

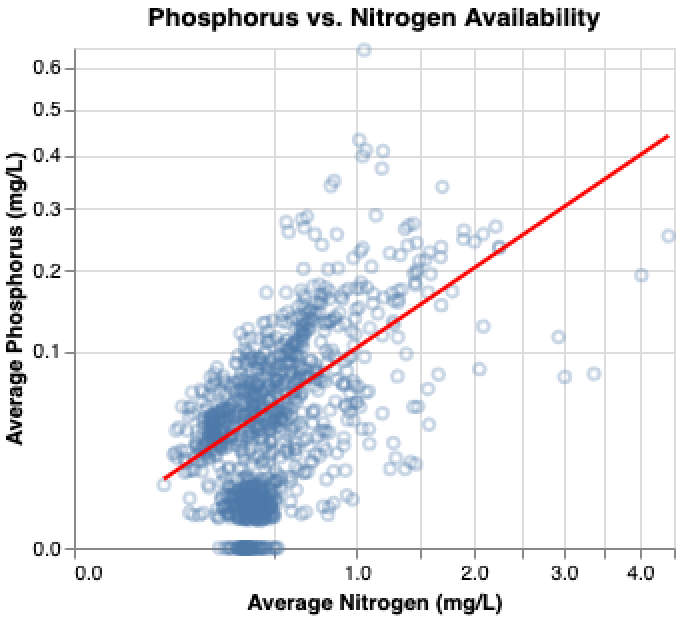
```
# visualizing relationship between nutrient availability and productivity
scatter = alt.Chart(agg_plot_df).mark_circle(opacity = 0.3).encode(
    x = alt.X('Average Nutrient Concentration:Q', scale = alt.Scale(type = 'sqrt')),
    y = alt.Y('Chlorophyll A (ug/L):Q',
              title = 'Average Chlorophyll A (ug/L)',
              scale = alt.Scale(type = 'sqrt')),
    color = alt.Color('Nutrient')
).properties(
    width = 300,
    height = 250,
    title = 'Productivity vs Nutrient Availability'
)

# smooth line
smooth = scatter.transform_loess(
    groupby = ['Nutrient'],
    on = 'Average Nutrient Concentration',
    loess = 'Chlorophyll A (ug/L)',
    bandwidth = 0.8
).mark_line(color = 'black')

fig1a = scatter + smooth
fig1a
```

Out[6]:



In [7]:
```
# visualize relationship between nitrogen and phosphorus availability
scatter1 = alt.Chart(ncca).mark_point(opacity=0.3).encode(
    x = alt.X('Nitrogen (mg N/L):Q',
              title = 'Average Nitrogen (mg/L)',
              scale = alt.Scale(type = 'sqrt')),
    y = alt.Y('Phosphorus (mg P/L):Q',
              title = 'Average Phosphorus (mg/L)',
              scale = alt.Scale(type = 'sqrt'))
).properties(
    width = 300,
    height = 250,
    title = 'Phosphorus vs. Nitrogen Availability'
)
# compute trend line
trend1 = scatter1.transform_regression(
    on = 'Nitrogen (mg N/L)',
    regression = 'Phosphorus (mg P/L)',
).mark_line(color = 'red')

fig1b = scatter1 + trend1
fig1b
```

Out[7]:



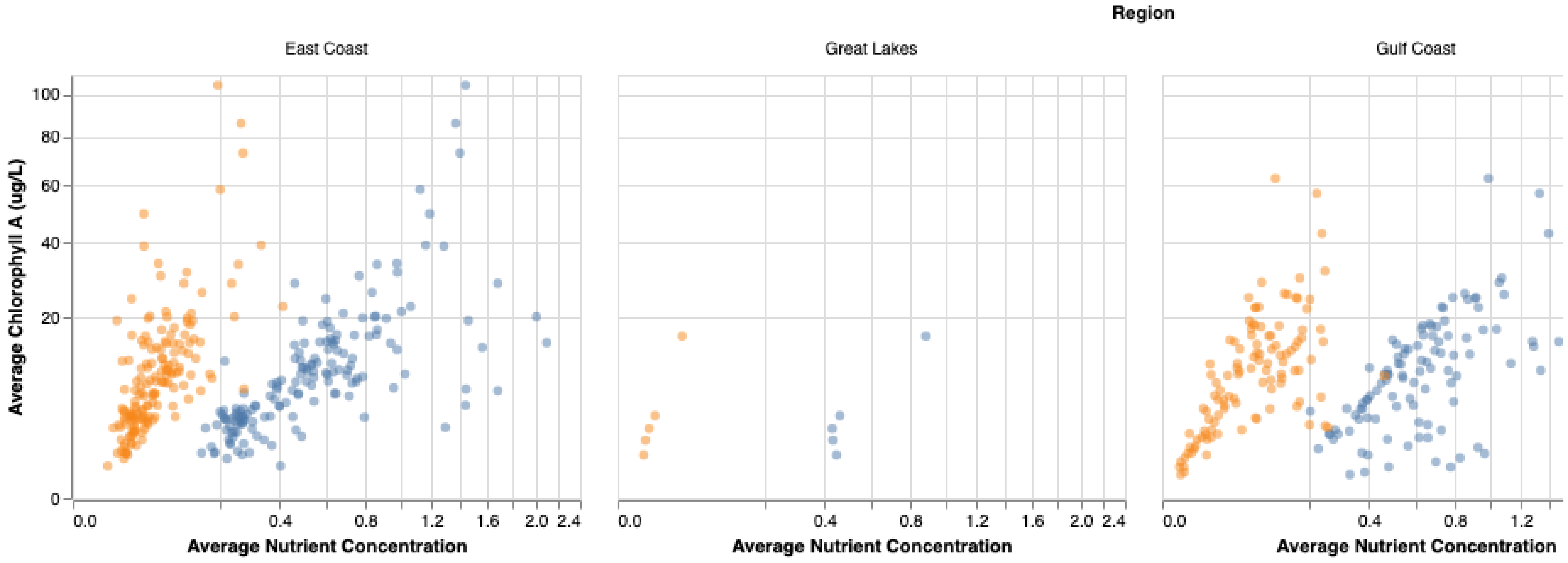
In [8]:
```
agg_region = ncca.groupby(
    ['Region', 'Waterbody Name']
).mean().drop(
    columns='Unique Site ID'
).reset_index().melt(
    id_vars = ['Region', 'Waterbody Name', 'Chlorophyll A (ug/L)'],
    value_vars = ['Nitrogen (mg N/L)', 'Phosphorus (mg P/L)'],
    var_name = 'Nutrient',
    value_name = 'Average Nutrient Concentration'
)

agg_region.head()
```

Out[8]:

	Region	Waterbody Name	Chlorophyll A (ug/L)	Nutrient	Average Nutrient Concentration
0	East Coast	Albermarle Sound	24.461667	Nitrogen (mg N/L)	0.597187
1	East Coast	Alligator River	4.040000	Nitrogen (mg N/L)	0.793500
2	East Coast	Assawoman Bay	10.160000	Nitrogen (mg N/L)	0.660000
3	East Coast	Back Bay	19.400000	Nitrogen (mg N/L)	1.459000
4	East Coast	Back Sound	3.040000	Nitrogen (mg N/L)	0.164375

In [9]:
```
fig2a = alt.Chart(agg_region).mark_circle(opacity = 0.5).encode(
    x = alt.X('Average Nutrient Concentration:Q', scale = alt.Scale(type = 'sqrt')),
    y = alt.Y('Chlorophyll A (ug/L):Q',
              title = 'Average Chlorophyll A (ug/L)',
              scale = alt.Scale(type = 'sqrt')),
    color = 'Nutrient'
).properties(
    width = 300,
    height = 250,
    title = 'Productivity vs Nutrient Availability'
).facet('Region')

fig2a
```

Out[9]:



In [10]:

```
# group by coastal region and aggregate by mean
ncca_region = ncca.drop(
    columns=['Unique Site ID']
).groupby(['Region']).mean().reset_index().melt(
    id_vars = ['Region', 'Chlorophyll A (ug/L)'],
    value_vars = ['Nitrogen (mg N/L)', 'Phosphorus (mg P/L)'],
    var_name = 'Nutrient',
    value_name = 'Average Nutrient Concentration'
)

ncca_region.sort_values('Average Nutrient Concentration', ascending=False)
```

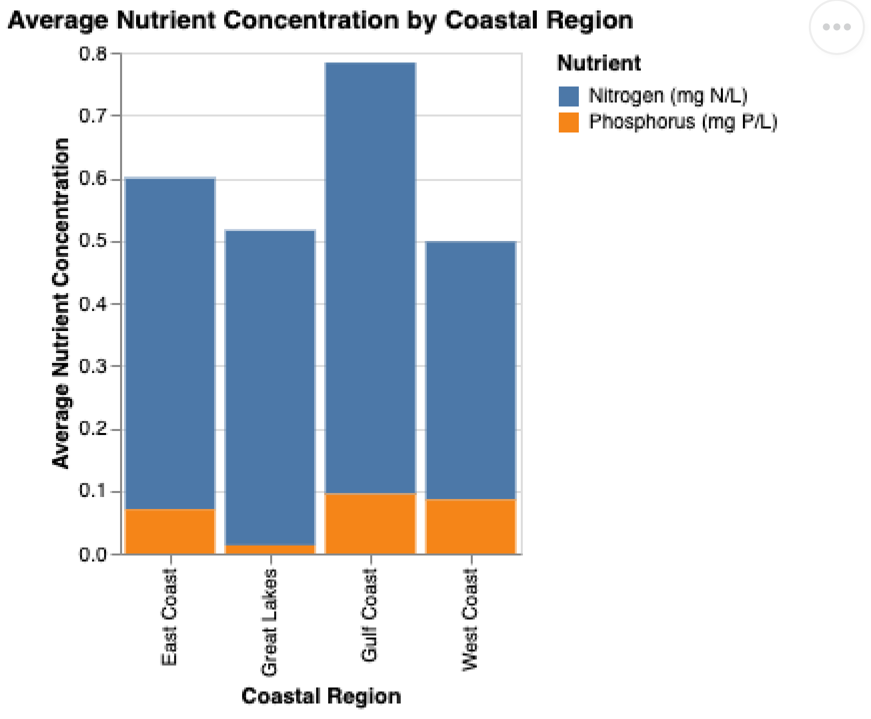
Out[10]:

	Region	Chlorophyll A (ug/L)	Nutrient	Average Nutrient Concentration
2	Gulf Coast	11.603818	Nitrogen (mg N/L)	0.689250
0	East Coast	10.617785	Nitrogen (mg N/L)	0.529884
1	Great Lakes	4.475248	Nitrogen (mg N/L)	0.503336
3	West Coast	5.545900	Nitrogen (mg N/L)	0.412794
6	Gulf Coast	11.603818	Phosphorus (mg P/L)	0.094474
7	West Coast	5.545900	Phosphorus (mg P/L)	0.085498
4	East Coast	10.617785	Phosphorus (mg P/L)	0.069867
5	Great Lakes	4.475248	Phosphorus (mg P/L)	0.013019

In [11]:

```
fig2b = alt.Chart(ncca_region).mark_bar().encode(
    x = alt.X('Region', title = 'Coastal Region'),
    y = 'Average Nutrient Concentration',
    color = 'Nutrient'
).properties(
    width = 200,
    height = 250,
    title = 'Average Nutrient Concentration by Coastal Region')

fig2b
```

Out[11]:

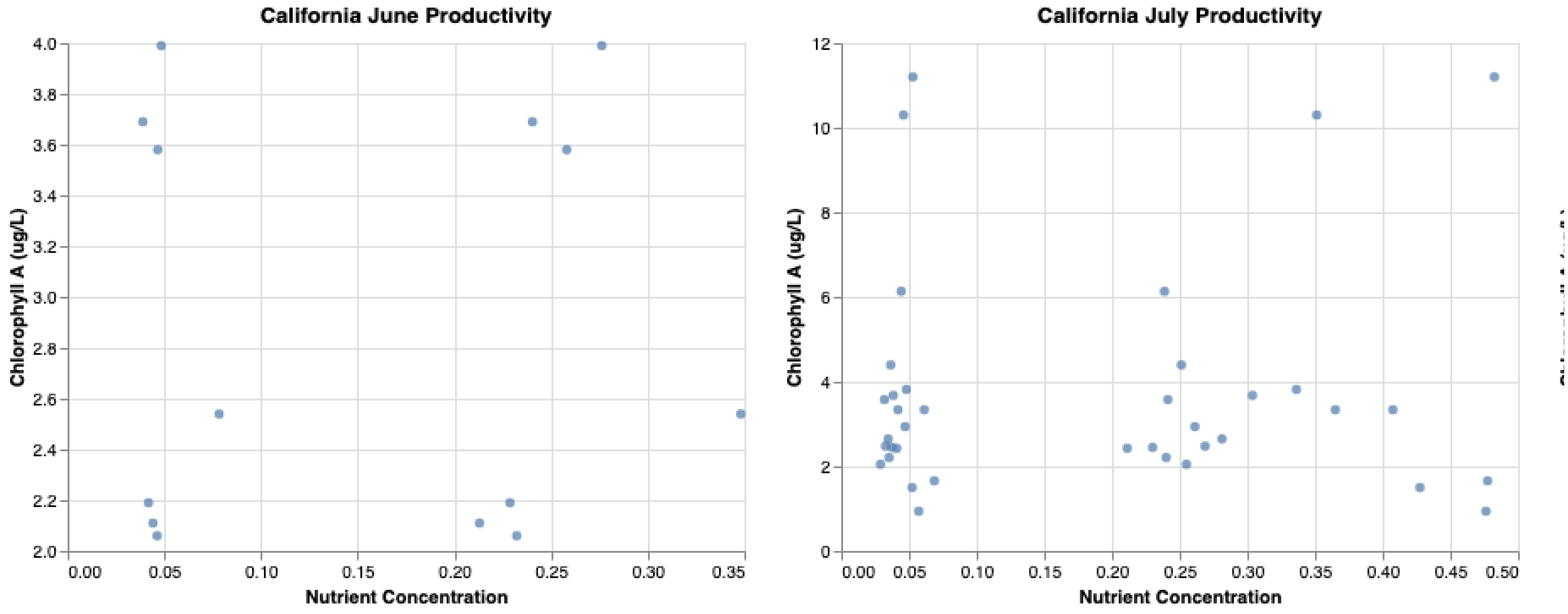


In [12]:

```
# melt for plotting
season = ncca.loc[ncca['State']=='CA'].melt(
    id_vars = ['State', 'Date Collected', 'Chlorophyll A (ug/L)'],
    value_vars = ['Nitrogen (mg N/L)', 'Phosphorus (mg P/L)'],
    var_name = 'Nutrient',
    value_name = 'Nutrient Concentration'
)
```

```
# extract data for each month and plot
june = season[season['Date Collected'].str.contains('Jun')]
july = season[(season['Date Collected'].str.contains('Jul'))]
aug = season[(season['Date Collected'].str.contains('Aug'))]
sep = season[(season['Date Collected'].str.contains('Sep'))]

june_fig = alt.Chart(june).mark_circle().encode(
    x = alt.X('Nutrient Concentration:Q'),
    y = alt.Y('Chlorophyll A (ug/L):Q', scale = alt.Scale(zero=False)),
).properties(
    title = 'California June Productivity'
)

jul_fig = alt.Chart(july).mark_circle().encode(
    x = alt.X('Nutrient Concentration:Q'),
    y = alt.Y('Chlorophyll A (ug/L):Q', scale = alt.Scale(zero=False)),
).properties(
    title = 'California July Productivity'
)

aug_fig = alt.Chart(aug).mark_circle().encode(
    x = alt.X('Nutrient Concentration:Q'),
    y = alt.Y('Chlorophyll A (ug/L):Q', scale = alt.Scale(zero=False)),
).properties(
    title = 'California August Productivity'
)

sep_fig = alt.Chart(sep).mark_circle().encode(
    x = alt.X('Nutrient Concentration:Q'),
    y = alt.Y('Chlorophyll A (ug/L):Q', scale = alt.Scale(zero=False)),
).properties(
    title = 'California September Productivity'
)

fig3 = june_fig | jul_fig | aug_fig | sep_fig
fig3
```

Out[12]:



In [21]:

```
plot_df = ncca.melt(
    id_vars = ['Chlorophyll A (ug/L)', 'Region'],
    value_vars = ['Nitrogen (mg N/L)', 'Phosphorus (mg P/L)'],
    var_name = 'Nutrient',
    value_name = 'Nutrient Concentration'
)
plot_df.head()
```

Out[21]:

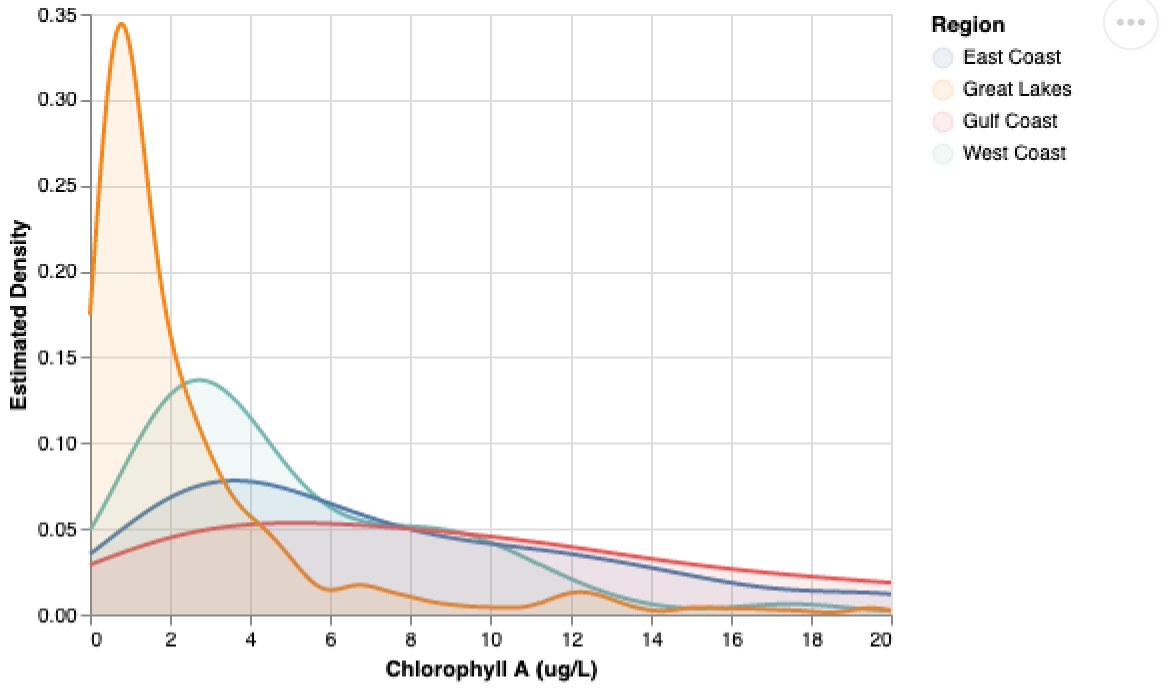
	Chlorophyll A (ug/L)	Region	Nutrient	Nutrient Concentration
0	3.34	West Coast	Nitrogen (mg N/L)	0.40750
1	2.45	West Coast	Nitrogen (mg N/L)	0.23000
2	3.82	West Coast	Nitrogen (mg N/L)	0.33625
3	6.13	West Coast	Nitrogen (mg N/L)	0.23875
4	9.79	East Coast	Nitrogen (mg N/L)	0.63250

In [38]:

```
# density estimates of productivity for each region
p = alt.Chart(plot_df).transform_density(
    density = 'Chlorophyll A (ug/L)',
    groupby = ['Region'],
    as_ = ['Chlorophyll A (ug/L)', 'Estimated Density'],
    extent = [0, 20],
    steps = 1000
).mark_line().encode(
    x = 'Chlorophyll A (ug/L):Q',
    y = 'Estimated Density:Q',
    color = 'Region:N'
)
```

```
fig4 = p + p.mark_area(opacity = 0.1)
fig4
```

Out[38]:



In [ ]: