Primary Productivity in Coastal Waters

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:

- approach 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: data description

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. You do not need to describe preprocessing steps. 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 [13]:
```
# show a few rows of clean data

ncca.head()
```

Out[13]:

	Waterbody Location	Region	Ammonia	Chlorophyll A	Nitrate/Nitrite	Nitrogen	Phosphorus
0	Alazan Bay	Gulf Coast	0.042	12.760000	0.107	0.882500	0.143675
1	Albermarle Sound	East Coast	0.003	24.461667	0.000	0.597187	0.032193
2	Alligator River	East Coast	0.031	4.040000	0.093	0.793500	0.024905
3	Alsea Bay	West Coast	0.020	6.640000	0.320	0.501250	0.072810
4	Anclote Anchorage	Gulf Coast	0.015	1.270000	0.037	0.372500	0.008185

My tidy dataset includes the Waterbody Location, Region, Ammonia, Chlorophyll A, Nitrate/Nitrite, Nitrogen, and Phosphorus variables. I chose these attributes because it is helpful to find the correlation between these nutrients and the regions where they reside, espeically the Chlorophyll A variable that allows one to see its productivity. I also picked these specific nutrients because all the other related variables have no values. In this dataset, each row shows the location with its nutrient concentration.

Part 2: exploratory analysis

Answer each question below and provide a graphic or other quantitative evidence supporting your answer. A description and interpretation of the graphic/evidence should be offered.

- (i) 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.
- (ii) Are there any notable differences in available nutrients among U.S. coastal regions?
- (iii) 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 why you ruled them out.

- (iv) How does primary productivity in California coastal waters change seasonally in 2010, if at all? Does your result make intuitive sense?
- (v) Pose and answer one additional question.

(i) To answer this question, I made a scatterplot graph comparing chlorophyll A with the nutrients. As shown, there is a positive relationship between nutrient availability and productivity among nutrients like nitrogen and phosphorus while ammonia and nitrate/nitrite peaks and later decreases. This shows that an increase in nitrogen and phosphorus increases chlorophyll production.

(ii) I created bar graphs featuring the average amount of nutrient concentration among different U.S. coastal regions. Based off of each chart, the Gulf Coast has the highest amount of nitrate/nitrite and nitrogen. On the other hand, the Great lakes has the least amount of all types of nutrients that are displayed, and the East Coast has the highest amount of ammonia. Overall, except for the Great Lakes, the coastal regions have around the same amount of phosphorus.

(iii) Productivity does seem to vary geographically in some way. I figured this out by using the same melted ncca data set that I used for the last question to create into scatterplot graphs. Analyzing these, there seems to be a different kind of spread among the different geographic regions. However, this result may be biased because one can tell that there are not an even amount of data points for each area, especially for the Great Lakes graph; therefore, it can be hard to accurately compare the Great Lakes to other coastal regions

(iv) Looking at the data, the only dates the data is collected in California is only among four consecutive months (June, July, August, September), so it is difficult to tell whether the waters change seasonally in 2010. However, there are some differences when I checked each month to see its primary productivity. August has the most amount of spread compared to the other months, although we may face the same problem of not having the same or enough data points to make a full comparison.

(v) One question I made was which state had the highest amount of total chlorophyll A concentration. By creating a bar graph that features 28 states, Ohio seems to have the most amount of chlorophyll A concentration.

Code appendix

In [5]:
```
import pandas as pd
import numpy as np
import altair as alt

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

In [6]:
```
merge = pd.merge(ncca_sites, ncca_raw, how = 'right', on = ['UID', 'STATE'])

ncca_merge = merge.pivot(
    index = ['UID', 'WTBDY_NM', 'STATE', 'NCA_REGION', 'DATE_COL_x'],
    columns = 'PARAMETER_NAME', values = 'RESULT').reset_index(
    ).rename_axis(columns=None).rename(columns = {'WTBDY_NM': 'Waterbody Location', 'NCA_REGION': 'Region'})

ncca_merge.head()
```

Out[6]:

	UID	Waterbody Location	STATE	Region	DATE_COL_x	Ammonia	Chlorophyll A	Dissolved Inorganic Nitrogen	Dissolved Inorganic Phosphate	Dissolved Silica	Nitrate	Nitrate/Nitrite	Nitrite	I
0	59	Mission Bay	CA	West Coast	1-Jul-10	0.000	3.34	0.014	0.028	NaN	NaN	0.014	NaN	
1	60	San Diego Bay	CA	West Coast	1-Jul-10	0.010	2.45	0.020	0.026	NaN	NaN	0.010	NaN	
2	61	Mission Bay	CA	West Coast	1-Jul-10	0.000	3.82	0.009	0.030	NaN	NaN	0.009	NaN	
3	62	San Diego Bay	CA	West Coast	1-Jul-10	0.000	6.13	0.010	0.028	NaN	NaN	0.010	NaN	
4	63	White Oak River	NC	East Coast	9-Jun-10	0.002	9.79	0.030	0.043	NaN	NaN	0.028	NaN	

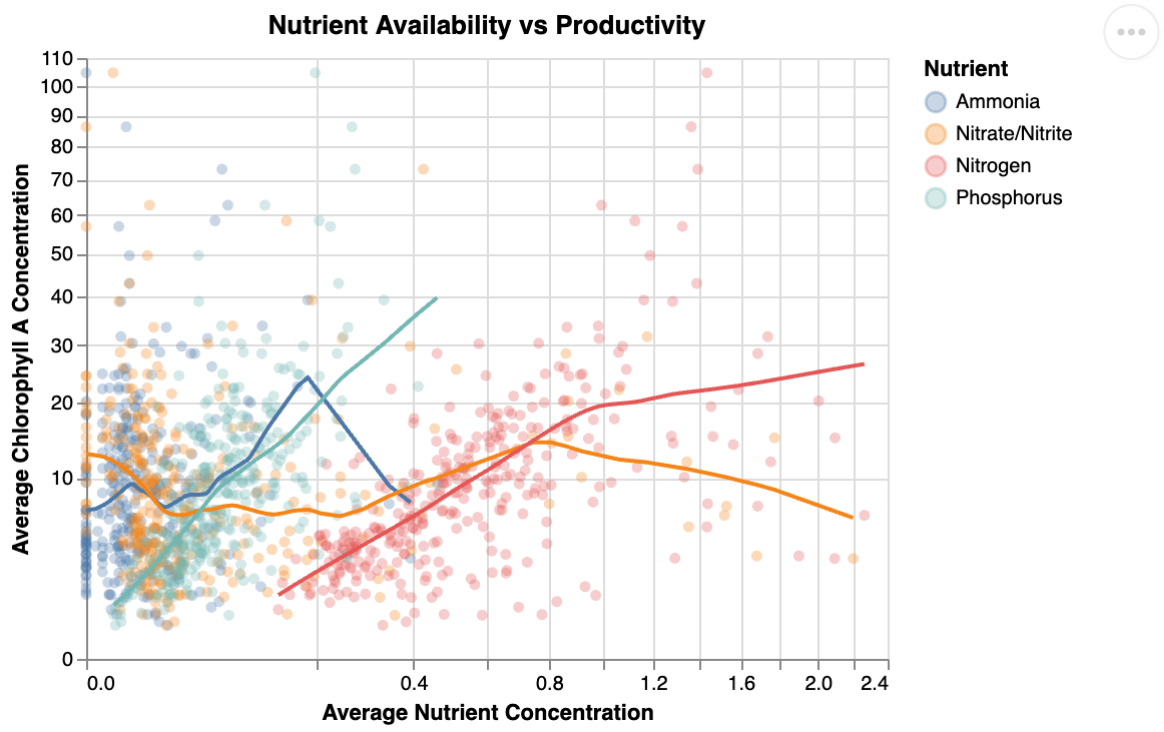
In [7]:
```
drop = ['UID', 'Dissolved Silica', 'Nitrate', 'Nitrite',
        'Nitrogen Particulate', 'Phosphorus Particulate',
        'Total Dissolved Nitrogen', 'Total Dissolved Phosphorus',
        'Total Kjeldahl Nitrogen', 'Dissolved Inorganic Nitrogen',
        'Dissolved Inorganic Phosphate']

ncca = ncca_merge.drop(columns = drop).drop(columns = ['DATE_COL_x', 'STATE']).groupby(['Waterbody Location', 'Regi
               ).reset_index().rename(columns = {'Total Nitrogen':'Nitrogen',
                                                 'Total Phosphorus':'Phosphorus'})

ncca.head()
```

Out[7]:

	Waterbody Location	Region	Ammonia	Chlorophyll A	Nitrate/Nitrite	Nitrogen	Phosphorus
0	Alazan Bay	Gulf Coast	0.042	12.760000	0.107	0.882500	0.143675
1	Albermarle Sound	East Coast	0.003	24.461667	0.000	0.597187	0.032193
2	Alligator River	East Coast	0.031	4.040000	0.093	0.793500	0.024905
3	Alsea Bay	West Coast	0.020	6.640000	0.320	0.501250	0.072810
4	Anclote Anchorage	Gulf Coast	0.015	1.270000	0.037	0.372500	0.008185

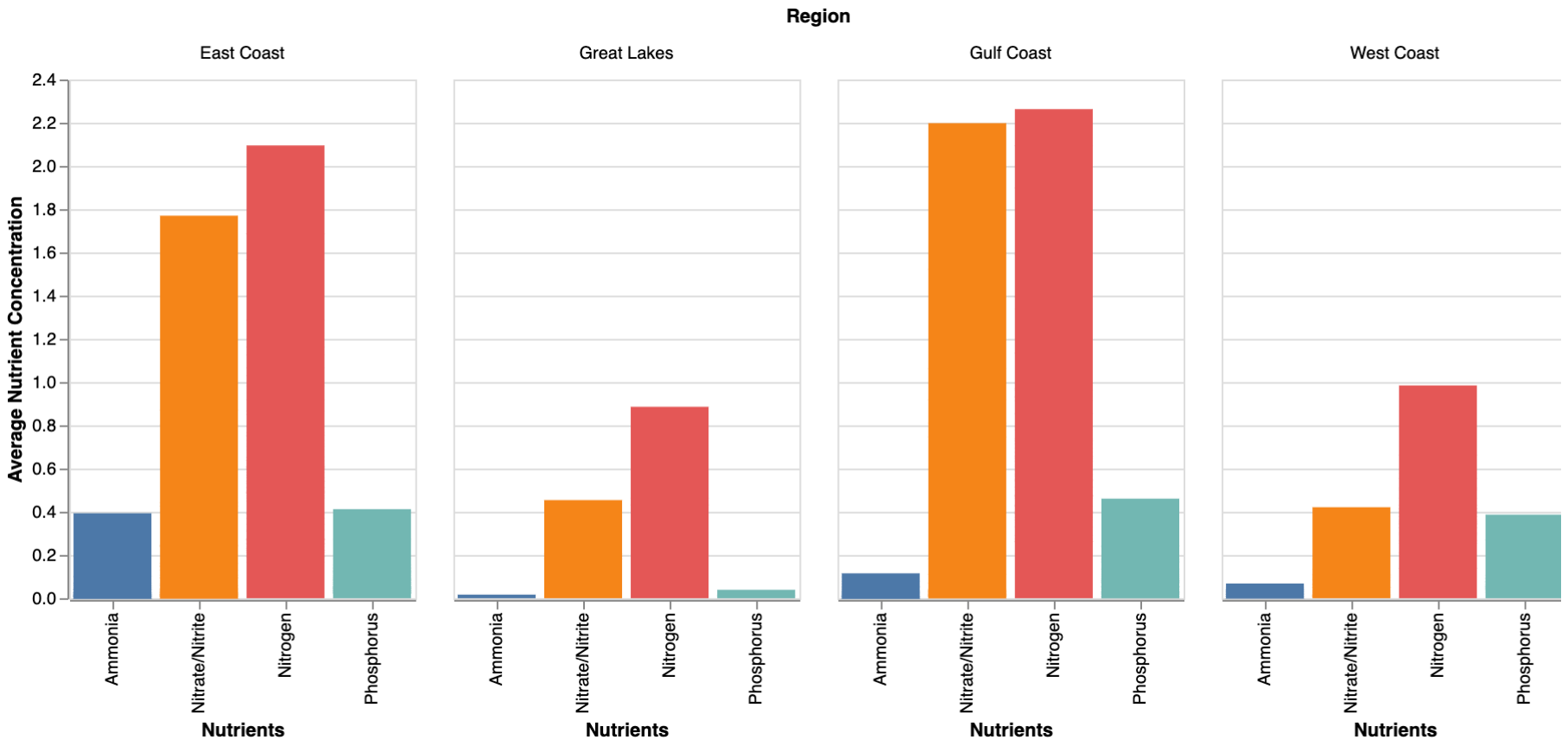
In [31]:
```
i_melt = ncca.melt(
    id_vars = ['Waterbody Location', 'Region', 'Chlorophyll A'],
    value_vars = ['Ammonia', 'Nitrate/Nitrite', 'Nitrogen', 'Phosphorus'],
    var_name = 'Nutrient',
    value_name = 'Average Nutrient Concentration')

i_scatter = alt.Chart(i_melt).mark_circle(opacity = 0.3).encode(
    x = alt.X('Average Nutrient Concentration:Q', scale = alt.Scale(type = 'sqrt')),
    y = alt.Y('Chlorophyll A:Q',
              title = 'Average Chlorophyll A Concentration',
              scale = alt.Scale(type = 'sqrt')),
    color = alt.Color('Nutrient')
    ).properties(title = 'Nutrient Availability vs Productivity')

i_smooth = i_scatter.transform_loess(
    groupby = ['Nutrient'],
    on = 'Average Nutrient Concentration',
    loess = 'Chlorophyll A',
    bandwidth = 0.7
).mark_line(color = 'black')

i_scatter + i_smooth
```

Out[31]:



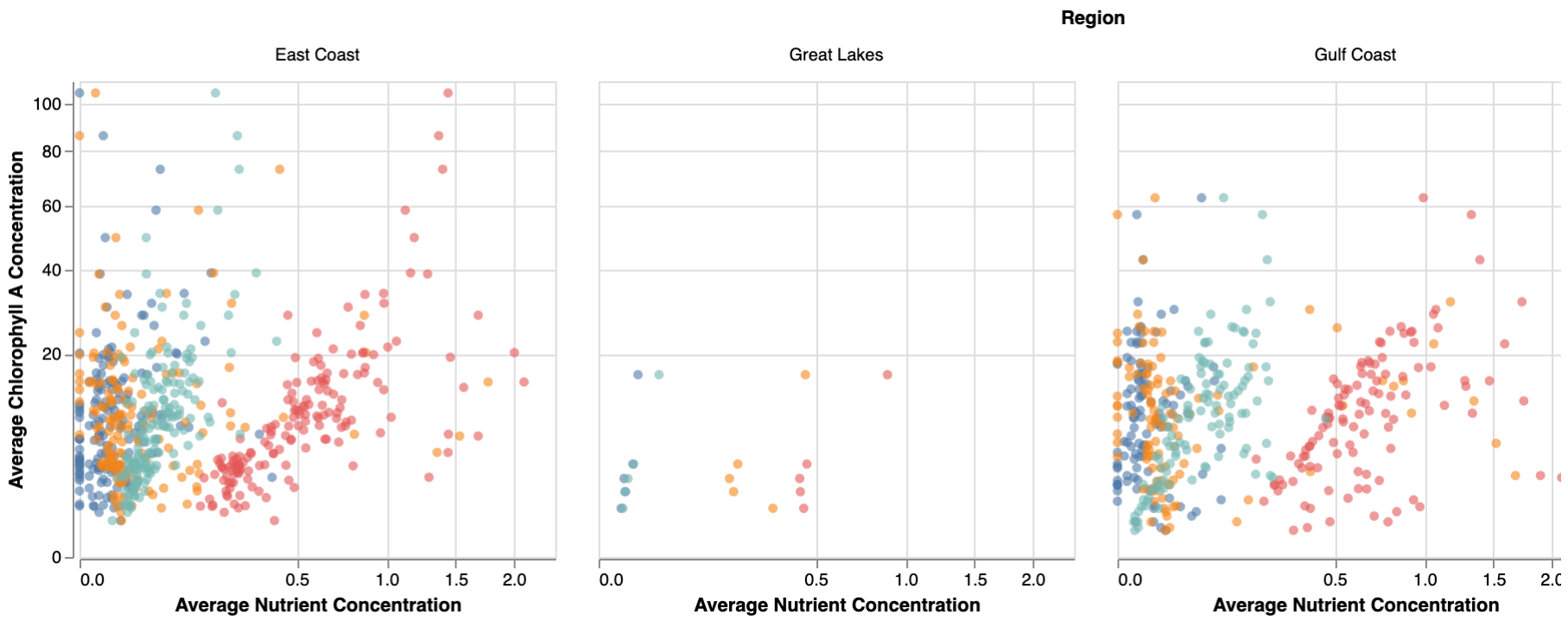
In [15]:
```
ii_melt = ncca.melt(
    id_vars = ['Region', 'Chlorophyll A'],
    value_vars = ['Ammonia', 'Nitrate/Nitrite', 'Nitrogen', 'Phosphorus'],
    var_name = 'Nutrient',
    value_name = 'Average Nutrient Concentration')

ii_bar = alt.Chart(ii_melt).mark_bar().encode(
    x = alt.X('Nutrient', title = 'Nutrients'),
    y = 'Average Nutrient Concentration',
    color = 'Nutrient').properties(width = 200).facet('Region')

ii_bar
```

Out[15]:



In [16]:
```
iii_scatter = alt.Chart(ii_melt).mark_circle(opacity = 0.6).encode(
    x = alt.X('Average Nutrient Concentration:Q', scale = alt.Scale(type = 'sqrt')),
    y = alt.Y('Chlorophyll A:Q',
              title = 'Average Chlorophyll A Concentration',
              scale = alt.Scale(type = 'sqrt')),
    color = alt.Color('Nutrient')
    ).properties(width = 275, height = 275,
                 title = 'Nutrient Availability vs Productivity').facet('Region')

iii_scatter
```

Out[16]:

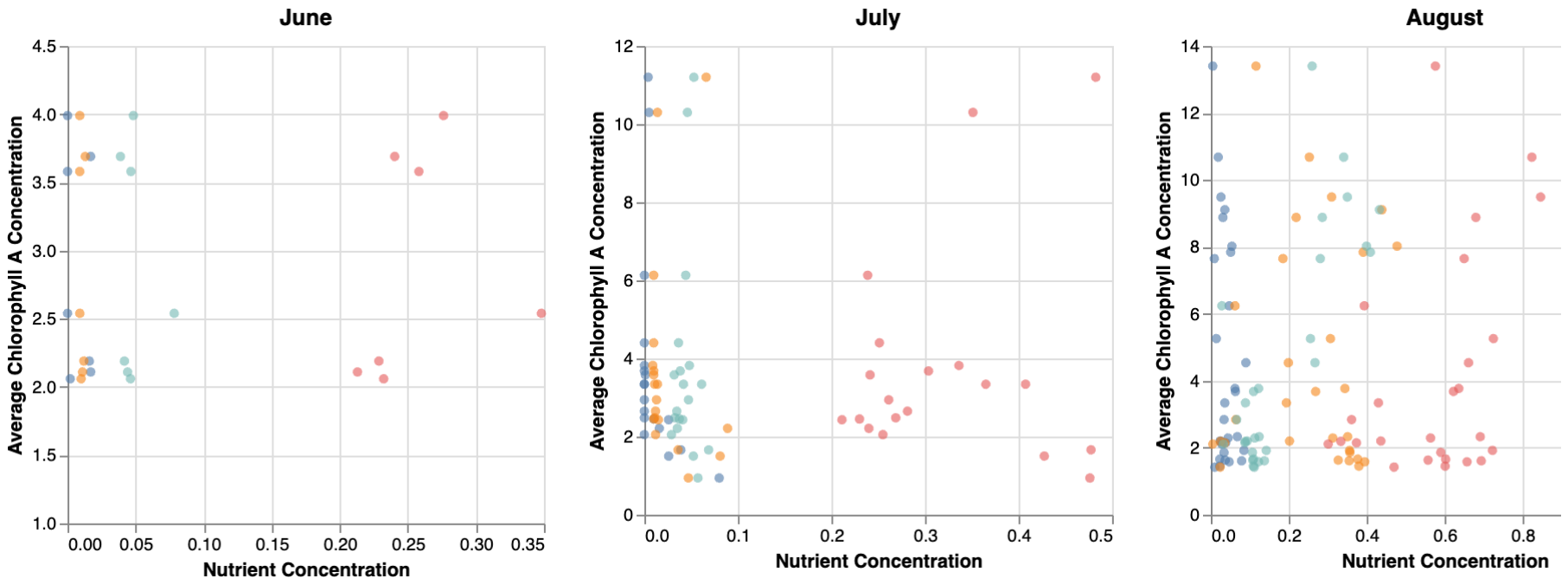


In [34]:
```
iv_ncca = ncca_merge.drop(columns = drop
                          ).rename(columns = {'WTBDY_NM': 'Waterbody Location', 'NCA_REGION': 'Region',
                                              'DATE_COL_x': 'Date Collected', 'STATE': 'State',
                                              'Total Nitrogen':'Nitrogen', 'Total Phosphorus':'Phosphorus'})

iv_ca = iv_ncca[iv_ncca['State'] == 'CA'].drop(columns = ['Waterbody Location', 'Region', 'State'])

iv_june = iv_ca[iv_ca['Date Collected'].str.contains('Jun')].drop(columns = 'Date Collected').melt(
    id_vars = 'Chlorophyll A', value_vars = ['Ammonia', 'Nitrate/Nitrite', 'Nitrogen', 'Phosphorus'],
    var_name = 'Nutrient', value_name = 'Nutrient Concentration')

iv_july = iv_ca[iv_ca['Date Collected'].str.contains('Jul')].drop(columns = 'Date Collected').melt(
    id_vars = 'Chlorophyll A', value_vars = ['Ammonia', 'Nitrate/Nitrite', 'Nitrogen', 'Phosphorus'],
    var_name = 'Nutrient', value_name = 'Nutrient Concentration')

iv_august = iv_ca[iv_ca['Date Collected'].str.contains('Aug')].drop(columns = 'Date Collected').melt(
    id_vars = 'Chlorophyll A', value_vars = ['Ammonia', 'Nitrate/Nitrite', 'Nitrogen', 'Phosphorus'],
    var_name = 'Nutrient', value_name = 'Nutrient Concentration')

iv_september = iv_ca[iv_ca['Date Collected'].str.contains('Sep')].drop(columns = 'Date Collected').melt(
    id_vars = 'Chlorophyll A', value_vars = ['Ammonia', 'Nitrate/Nitrite', 'Nitrogen', 'Phosphorus'],
    var_name = 'Nutrient', value_name = 'Nutrient Concentration')
```

In [33]:
```
iv_june_scatter = alt.Chart(iv_june).mark_circle(opacity = 0.6).encode(
    x = alt.X('Nutrient Concentration:Q', scale = alt.Scale()),
    y = alt.Y('Chlorophyll A:Q', scale = alt.Scale(domain = [1, 4.5]),
              title = 'Average Chlorophyll A Concentration'),
```

```
    color = alt.Color('Nutrient')).properties(width = 275, height = 275,
                title = 'June')

iv_july_scatter = alt.Chart(iv_july).mark_circle(opacity = 0.6).encode(
    x = alt.X('Nutrient Concentration:Q', scale = alt.Scale()),
    y = alt.Y('Chlorophyll A:Q', scale = alt.Scale(),
              title = 'Average Chlorophyll A Concentration'),
    color = alt.Color('Nutrient')).properties(width = 270, height = 270,
                title = 'July')

iv_august_scatter = alt.Chart(iv_august).mark_circle(opacity = 0.6).encode(
    x = alt.X('Nutrient Concentration:Q', scale = alt.Scale()),
    y = alt.Y('Chlorophyll A:Q', scale = alt.Scale(),
              title = 'Average Chlorophyll A Concentration'),
    color = alt.Color('Nutrient')).properties(width = 270, height = 270,
                title = 'August')

iv_september_scatter = alt.Chart(iv_september).mark_circle(opacity = 0.6).encode(
    x = alt.X('Nutrient Concentration:Q', scale = alt.Scale()),
    y = alt.Y('Chlorophyll A:Q', scale = alt.Scale(),
              title = 'Average Chlorophyll A Concentration'),
    color = alt.Color('Nutrient')).properties(width = 270, height = 270,
                title = 'September')

iv_june_scatter | iv_july_scatter | iv_august_scatter | iv_september_scatter
```

Out[33]:



In [32]:

```
v_bar_cha = alt.Chart(iv_ncca).mark_bar().encode(
    x = alt.X('State', title = 'States'),
    y = alt.Y("Chlorophyll A", title = 'Total Chlorophyll A Concentration')
    ).properties(title = 'Total Chlorophyll A Concentration in Each State')

v_bar_cha
```

Out[32]: