Mini project 2: 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:

- 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 [26]:

```
# show a few rows of clean data
pd.read_csv('data.csv').head(5)
```

Out[26]:

	Unique_ID	Site_ID	State	Water_Body	Region	Coastal_Region	Date_Collected	Ammonia mg N/L	Ch
0	59	NCCA10-1111	CA	Mission Bay	West	West Coast	7/1/2010	0.000	
1	60	NCCA10-1119	CA	San Diego Bay	West	West Coast	7/1/2010	0.010	
2	61	NCCA10-1123	CA	Mission Bay	West	West Coast	7/1/2010	0.000	
3	62	NCCA10-1127	CA	San Diego Bay	West	West Coast	7/1/2010	0.000	
4	63	NCCA10-1133	NC	White Oak River	Southeast	East Coast	6/9/2010	0.002	

The tidy data has had many columns dropped for brevity and relevance. All of the data comes from waterbodies located within or around the USA. The columns that are left relate to which site the samples were collected from, where the site is located, when the samples were collected, and which samples were collected along with their measurements. Each sample was collected from a specfic water body. The samples are the levels of Chlorophyll A per ug/L which is known as the productivity of a waterbody and the various nutrients in the water. The plants and living organisms that produce Chlorophyll A rely on these nutrients to do so.

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.

Examining fig1-fig3 it can be seen that as nutrient availbility increases so does productivity. This pattern of increase seems consistent for both nitrogen and phosphorus.

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

It can be seen from hist2-hist4 that there are notable difference in the available nutrients depending on the coastal region. Phosphorous is much higher in the gulf coast than anywhere else and there is significantly more nitrogen in the great lakes than any of the coastal regions.

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.

Yes, examining hist2 and hist4 productivity seems to be highest where there is more available nitrogen.

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

Does your result make intuitive sense?

Yes, looking at the histograms from hist5 it can be seen that there spikes of primary productivity throughout various parts of the year in the different regions. The spikes seem to be relative to the time in which that region is expierencing spring/summer.

Pose and answer one additional question.

Which region has the most consistent amount of productivity? Why might that be?

From hist5 It appears that the east coast has the highest level of productivity. The levels are the most consistent out of the other regions. Looking at hist1 it can be seen that the east coast has the largest amount of samples in the 0-20 Chlorophyll A ug/L range so maybe the waterbodies are just not as affected by changes in seasonality like other regions.

Codes

In [3]:
```
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')
```

In [4]:
```
dcols = [4,5,9,10,11,12,13,14,15,16,17]

# Drop hopefully unimportant columns
ncca_raw_mod = ncca_raw.drop(ncca_raw.columns[dcols], axis = 1)

# Want to be able to pivot the data so combining the parameter name with units to make

ncca_raw_mod["PARAMETER_NAME"] = ncca_raw_mod["PARAMETER_NAME"] + " "+ ncca_raw_mod["U
ncca_raw_mod = ncca_raw_mod.drop("UNITS", axis = 1)

ncca_raw_mod.head()
```

Out[4]:

	UID	SITE_ID	STATE	DATE_COL	PARAMETER_NAME	RESULT
0	59	NCCA10-1111	CA	7/1/2010	Total Nitrogen mg N/L	0.407500
1	59	NCCA10-1111	CA	7/1/2010	Nitrate/Nitrite mg N/L	0.014000
2	59	NCCA10-1111	CA	7/1/2010	Dissolved Inorganic Phosphate mg P/L	0.028000
3	59	NCCA10-1111	CA	7/1/2010	Dissolved Inorganic Nitrogen mg N/L	0.014000
4	59	NCCA10-1111	CA	7/1/2010	Total Phosphorus mg P/L	0.061254

In [5]:
```
dcols = [3,4,6,7,8,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30]

# Drop hopefully unimportant columns
ncca_sites_mod = ncca_sites.drop(ncca_sites.columns[dcols], axis = 1)
pd.set_option('display.max_columns', None)

ncca_sites_mod.head()
```

Out[5]:

	UID	SITE_ID	STATE	WTBDY_NM	NCCR_REG	NCA_REGION
0	59	NCCA10-1111	CA	Mission Bay	West	West Coast
1	60	NCCA10-1119	CA	San Diego Bay	West	West Coast
2	61	NCCA10-1123	CA	Mission Bay	West	West Coast
3	62	NCCA10-1127	CA	San Diego Bay	West	West Coast
4	63	NCCA10-1133	NC	White Oak River	Southeast	East Coast

In [6]:
```
rawdata = pd.merge(ncca_raw_mod,ncca_sites_mod, how = 'left',on = ["SITE_ID", "UID","S

pd.set_option('display.max_columns', None)
rawdata.head(10)
# Looks good just need to pivot
```

Out[6]:

	UID	SITE_ID	STATE	DATE_COL	PARAMETER_NAME	RESULT	WTBDY_NM	NCCR_REG	NCA_RE
0	59	NCCA10-1111	CA	7/1/2010	Total Nitrogen mg N/L	0.407500	Mission Bay	West	West
1	59	NCCA10-1111	CA	7/1/2010	Nitrate/Nitrite mg N/L	0.014000	Mission Bay	West	West
2	59	NCCA10-1111	CA	7/1/2010	Dissolved Inorganic Phosphate mg P/L	0.028000	Mission Bay	West	West
3	59	NCCA10-1111	CA	7/1/2010	Dissolved Inorganic Nitrogen mg N/L	0.014000	Mission Bay	West	West
4	59	NCCA10-1111	CA	7/1/2010	Total Phosphorus mg P/L	0.061254	Mission Bay	West	West
5	59	NCCA10-1111	CA	7/1/2010	Ammonia mg N/L	0.000000	Mission Bay	West	West
6	59	NCCA10-1111	CA	7/1/2010	Chlorophyll A ug/L	3.340000	Mission Bay	West	West
7	60	NCCA10-1119	CA	7/1/2010	Total Nitrogen mg N/L	0.230000	San Diego Bay	West	West
8	60	NCCA10-1119	CA	7/1/2010	Nitrate/Nitrite mg N/L	0.010000	San Diego Bay	West	West
9	60	NCCA10-1119	CA	7/1/2010	Ammonia mg N/L	0.010000	San Diego Bay	West	West

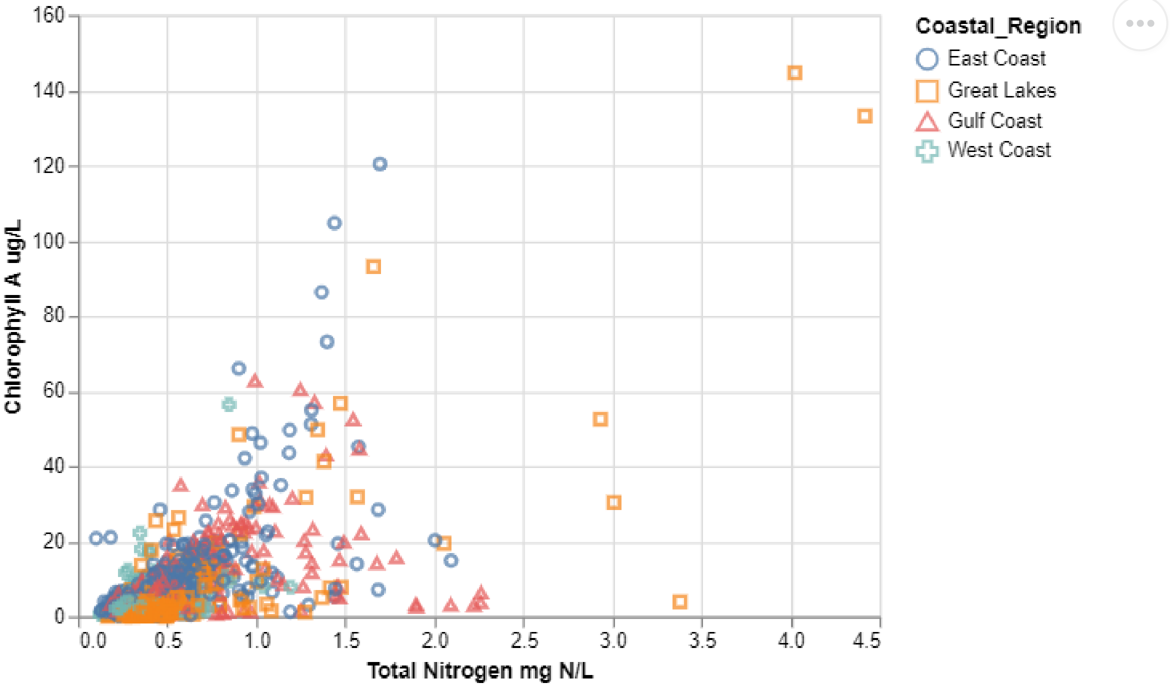
In [25]:

```
raw_data_mod1 = rawdata.pivot(index = np.append(ncca_sites_mod.columns,"DATE_COL"),col
                             ).reset_index().rename_axis(None, axis=1)
raw_data_mod2 = raw_data_mod1.rename(columns = {"UID":"Unique_ID" , "SITE_ID": "Site_I
                                                "WTBDY_NM":"Water_Body", "NCCR_REG": '
                                                "Coastal_Region", "DATE_COL":"Date_Col

data = raw_data_mod2
#pd.set_option('display.max_columns', None)

# The data is now tidy and ready to be used for plots

#data.head()
#data.to_csv("data.csv",index = False)
```

In [8]:

```
fig1 = alt.Chart(data).mark_point().encode(
    y = 'Chlorophyll A ug/L:Q',
    x = 'Total Nitrogen mg N/L:Q',
    color = "Coastal_Region",
    shape = "Coastal_Region",
    #size = alt.Size('NCCR_REG',scale = alt.Scale(type = 'log'))
)

fig1
```

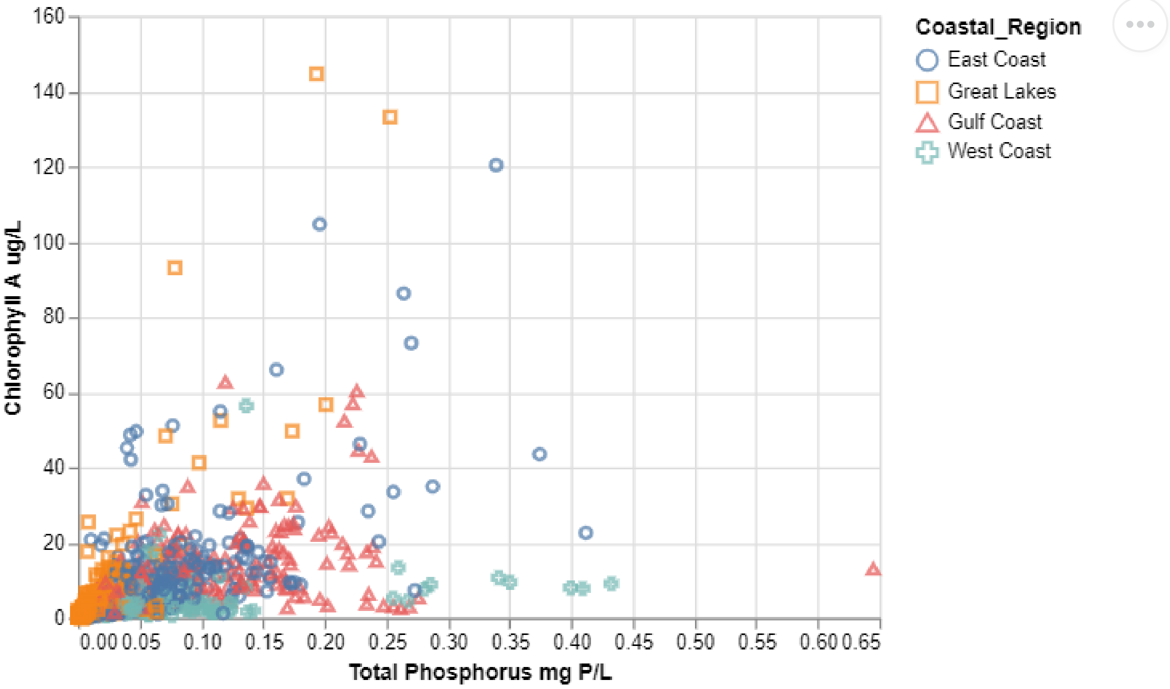
Out[8]:



In [9]:

```
fig2 = alt.Chart(data).mark_point().encode(
    y = 'Chlorophyll A ug/L:Q',
    x = 'Total Phosphorus mg P/L:Q',
    color = "Coastal_Region",
    shape = "Coastal_Region",
)
fig2
```

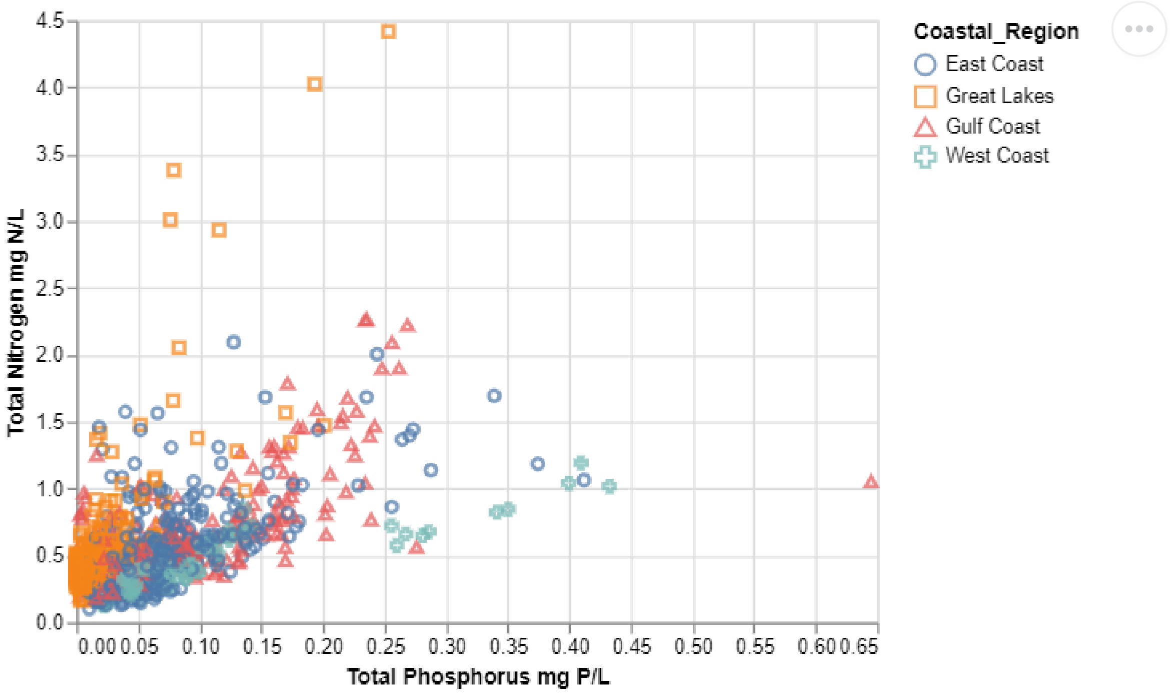
Out[9]:



In [10]:

```
fig3 = alt.Chart(data).mark_point().encode(
    y = 'Total Nitrogen mg N/L:Q',
    x = 'Total Phosphorus mg P/L:Q',
    color = "Coastal_Region",
    shape = "Coastal_Region",
)
fig3
```

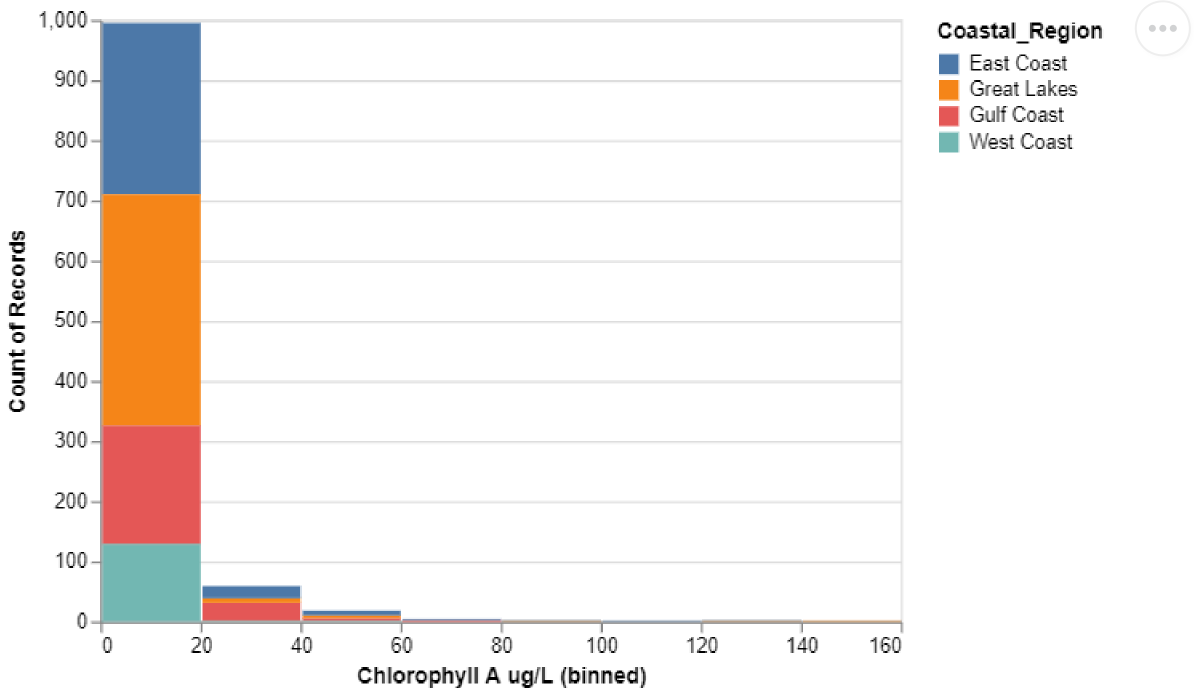
Out[10]:



In [11]:

```
hist1 = alt.Chart(data).mark_bar().encode(
    #alt.X('GDP per capita', scale = alt.Scale(type = 'log')),
    alt.X('Chlorophyll A ug/L:Q',bin = True),
    #y = 'Chlorophyll A ug/L:Q',
    y='count()',
    color = "Coastal_Region"
)
hist1
```

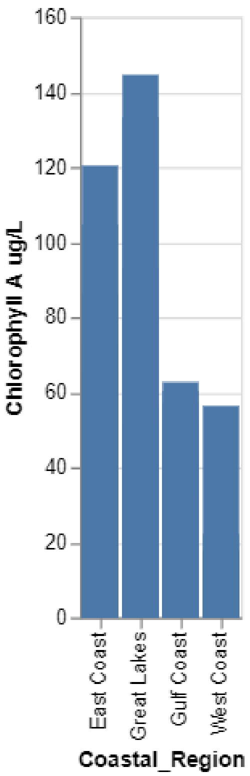
Out[11]:



In [12]:

```
hist2 = alt.Chart(data).mark_bar().encode(
    x = "Coastal_Region:N",
    y='Chlorophyll A ug/L:Q'

)
hist2
```

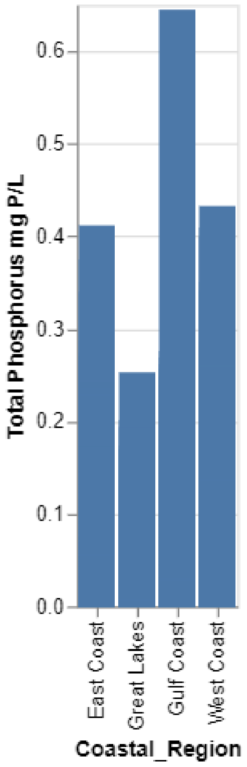
Out[12]:



In [13]:

```
hist3 = alt.Chart(data).mark_bar().encode(
    x = "Coastal_Region:N",
    y='Total Phosphorus mg P/L:Q'

)
hist3
```

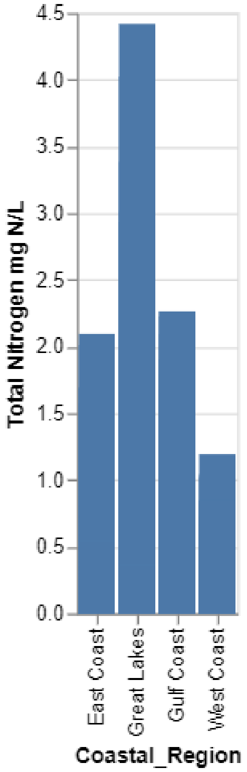
Out[13]:



In [14]:

```
hist4 = alt.Chart(data).mark_bar().encode(
    x = "Coastal_Region:N",
    y='Total Nitrogen mg N/L:Q'
)
hist4
```

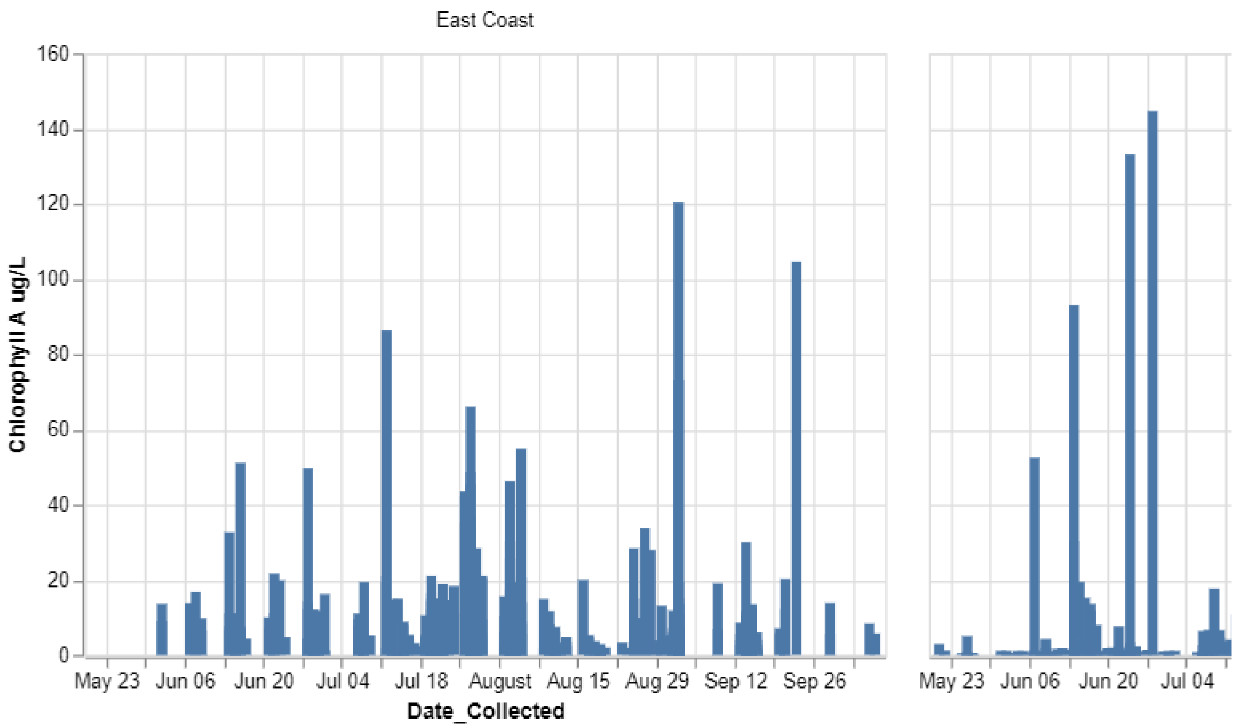
Out[14]:



In [15]:

```
hist5 = alt.Chart(data).mark_bar().encode(
    x = "Date_Collected:T",
    y='Chlorophyll A ug/L:Q',
    #color = "Coastal_Region"
).facet(
column = "Coastal_Region")
hist5
```

Out[15]:



In [ ]: