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:

- 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.

In [1]:
```python
import pandas as pd
import numpy as np
import altair as alt
```

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 [2]: #show a few rows of clean data
        pd.read_csv('tidy_data.csv').head()
```

Out[2]:

	UID	Site ID	State	Date	NCCR Region	Chlorophyll A (ug/L)	Total Nitrogen (mg N/L)	Total Phosphorus (mg P/L)
0	59	NCCA10-1111	CA	7/1/2010	West	3.34	0.40750	0.061254
1	60	NCCA10-1119	CA	7/1/2010	West	2.45	0.23000	0.037379
2	61	NCCA10-1123	CA	7/1/2010	West	3.82	0.33625	0.048100
3	62	NCCA10-1127	CA	7/1/2010	West	6.13	0.23875	0.044251
4	63	NCCA10-1133	NC	6/9/2010	Southeast	9.79	0.63250	0.090636

The dataset contains information on nutrient concentrations in coastal regions of the United States, specifically focusing on the Great Lakes, Gulf, Northeast, Southeast, and West regions. The key attributes of the dataset include measurements of three important nutrients: Total Nitrogen, Total Phosphorus, and Chlorophyll A. These nutrients play a crucial role in the health and quality of coastal ecosystems, and can be used to analyzed primary productivity in these coastal regions.

The dataset provides data at various sites within each coastal region, identified by the "Site ID" and "UID" columns (both unique identifiers). Each data entry is also associated with a specific date, captured in the "Date" column. Additionally, the dataset includes information about the state where the data was collected and the respective NCCR (National Coastal Condition Report) region. By analyzing this dataset, we can gain valuable insights into the variability of nutrient concentrations among different coastal regions, helping us understand the environmental conditions and potential impacts on the marine ecosystems.

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 observe the relationship between nutrient availability and productivity, we can examine the correlation between nutrient levels and Chlorophyll A levels which are a measure of productivity. The correlation between Total Nitrogen and Chlorophyll A is about 0.64, and the correlation between Total Phosphorus and Chlorophyll A is about 0.51. These correlation coefficients indicate a moderate positive correlation between Total Nitrogen and Chlorophyll A, and a slightly weaker positive correlation between Total Phosphorus and Chlorophyll A. The higher correlation coefficient for Total Nitrogen and Chlorophyll A suggests a relatively stronger relationship compared to Total Phosphorus and Chlorophyll A, suggesting that Nitrogen may have a more significant influence on the productivity of coastal waters, as measured by Chlorophyll A levels, when compared to Phosphorus. Scatter plots below graphically confirm these results about the relationship between Nitrogen, Phosphorous, and Chlorophyll A. According to the direction and slopes of the circles on the charts, both relationships are clearly on the positive side (but not too extreme), with Nitrogen being slightly stronger. We must also consider whether they might be related to each other. The correlation coefficient between Total Nitrogen and Total Phosphorus is about 0.57. This indicates another moderate positive relationship, this time between the two nutrients. The positive correlation suggests that as the levels of Total Nitrogen increase, the levels of Total Phosphorus also tend to increase, and vice versa.

(ii) The scatter plot highlighting nutrient values among U.S. coastal regions reveals interesting insights regarding the variations in nutrient availability. The plot shows a positive trend between Total Nitrogen and Total Phosphorus, indicating a potential relationship between these two nutrients. Additionally, distinct clusters of points are visible, indicating notable differences in nutrient levels among the different coastal regions. Notably, the cluster of blue points at the bottom of the plot represents the Great Lakes region, suggesting lower nutrient levels in that area. Conversely, the green points located at the top of the line represent the West region, indicating relatively higher nutrient levels there. This visual representation demonstrates the presence of some regional disparities in nutrient availability along the U.S. coast. The scatter plot provides valuable evidence of the diverse nutrient profiles across U.S. coastal regions.

(iii) To determine if productivity varies geographically based on the 2010 data, we can examine the relationship between productivity (Chlorophyll A) and the geographic location represented by the 'NCCR_REG' column. By visualizing the distribution of productivity across different waterbodies or coastal regions, we can gain insights into any potential geographic variations. One approach is to create a box plot that shows the distribution of Chlorophyll A values for each NCCA region. This allows us to compare the median and variability of productivity across different locations. Based on the information from the box and whisker plots, it appears that there are differences in the productivity levels (represented by Chlorophyll A) among the regions. The Gulf region generally has higher productivity levels compared to other regions, indicated by its higher median and larger interquartile range. The Great Lakes region has the lowest productivity levels overall, with a smaller interquartile range and a lower maximum value. The presence of outliers, indicated by the individual data points beyond the whiskers, can also provide insights. For example, the Northeast region has an outlier with a value of 120.37, suggesting a potentially exceptional productivity event in that specific location at a certain point in time. While the range of the values of Chlorophyll A may not seem that spread out, each region has distinct differences in the makeup of its box plots. Overall, these differences in the plots suggest that productivity levels vary geographically among the regions.

(iv) Yes, based on the line plot created below showing the seasonal changes in primary productivity in California coastal waters for 2010, there appears to be variation in Chlorophyll A levels throughout the season. There is a noticeable variation in primary productivity throughout the summer season that can be seen visually through the plot and by studying the mean Chlorophyll A levels. The Chlorophyll A levels exhibit both peaks and drops, indicating changes in primary productivity over time. The highest peak in Chlorophyll A levels occurs on July 27th, with a value of 11.2. This suggests a period of potentially increased primary productivity during that day. Additionally, on July 29th, the Chlorophyll A level still remains high at 10.3. There are also other fluctuations and transitions in primary productivity throughout the summer months. For example, there is a large drop in Chlorophyll A levels on August 10th (2.15), followed by an immediate increase on August 11th (6.24). This indicates a shift in primary productivity during that small period. These observed seasonal changes in primary productivity align with the expected patterns in coastal ecosystems and do make intuitive sense. Some factors such as variations in sunlight, temperature, and nutrient availability can influence primary productivity. The higher Chlorophyll A levels during certain times suggest increased photosynthetic activity and primary production, while drops in levels may indicate environmental changes or shifts in the community dynamics. In conclusion, this analysis reveals seasonal changes in primary productivity in California coastal waters during 2010.

(v) Question: How does the variability of each nutrient concentrations differ between coastal regions in the United States? Answer: To visualize the differences in nutrient variability among coastal regions, we can create box plots for each of the nutrient parameters. These plots allow us to compare the distribution of nutrient concentrations across regions and identify any notable differences in variability. In analyzing the variability of nutrient concentrations across coastal regions in the United States, we examined three key nutrients: Total Nitrogen, Total Phosphorus, and Chlorophyll A. The results reveal interesting patterns in variability among the regions. Total Nitrogen and Total Phosphorus concentrations exhibited relatively consistent levels across all regions, with moderate spreads and some outliers. The Gulf region displayed slightly wider spreads compared to the Great Lakes, Northeast, and West regions. Chlorophyll A showed more pronounced variability. The Great Lakes and Northeast regions had narrower spreads, while the Gulf, Southeast, and West regions displayed wider spreads. Notably, the Gulf region demonstrated higher variability in Chlorophyll A concentrations. Overall, this analysis suggests that while nutrient concentrations remain relatively consistent across coastal regions, differences in Chlorophyll A variability may indicate variations primary productivity, particularly in the Gulf region. Even further investigation into the underlying factors driving these variations would provide valuable insights for coastal management and conservation efforts if those are the goals of this data collection.

Code appendix

In [3]:
```
ncca_raw = pd.read_csv('data/assessed_ncca2010_waterchem.csv')
ncca_sites = pd.read_csv('data/assessed_ncca2010_siteinfo.csv')
```

In [4]:
```
ncca_raw.drop(columns = ['BATCH_ID', 'MDL', 'MRL', 'PQL', 'HOLDING_TIME', 'C
```

In [5]:
```
ncca_sites = ncca_sites[['UID', 'SITE_ID', 'STATE', 'NCCR_REG']]
```

In [6]:
```
#merge the dataframes based on the common columns
merged_data = pd.merge(ncca_raw, ncca_sites, on=['UID', 'SITE_ID', 'STATE'])
```

In [7]:
```
#convert merged_data to csv so easier to read
merged_data.to_csv('merged_data.csv', index=False)
```

In [8]:
```
#pivot data so parameters are variables with results filling in and drop unr
pivot_data = merged_data.pivot(
    index=['UID', 'SITE_ID', 'STATE', 'DATE_COL', 'NCCR_REG'],
    columns = 'PARAMETER_NAME', #want parameters to be columns
    values = 'RESULT',  #want result values to fill
).reset_index().rename_axis(columns={'PARAMETER_NAME': ''})
```

```
pivot_data=pivot_data[['UID', 'SITE_ID', 'STATE', 'DATE_COL', 'NCCR_REG', 'C

tidy_data = pivot_data.rename(columns={
    'Chlorophyll A': 'Chlorophyll A (ug/L)',
    'Total Nitrogen': 'Total Nitrogen (mg N/L)',
    'Total Phosphorus': 'Total Phosphorus (mg P/L)',
    'SITE_ID': 'Site ID',
    'DATE_COL': 'Date',
    'NCCR_REG': 'NCCR Region',
    'STATE':'State'
})
```

In [9]:

```
#convert tidy_data to csv so easier to read
tidy_data.to_csv('tidy_data.csv', index=False)
```

In [10]:

```
#code for part (i)
#continue working with pivot_data which has raw column names
```

In [11]:

```
#study nitrogen productivity based on correlation with Chlorophyll A levels
corr_nitrogen = pivot_data['Total Nitrogen'].corr(pivot_data['Chlorophyll A'
print("Correlation between Total Nitrogen and Chlorophyll A:", corr_nitroger
```

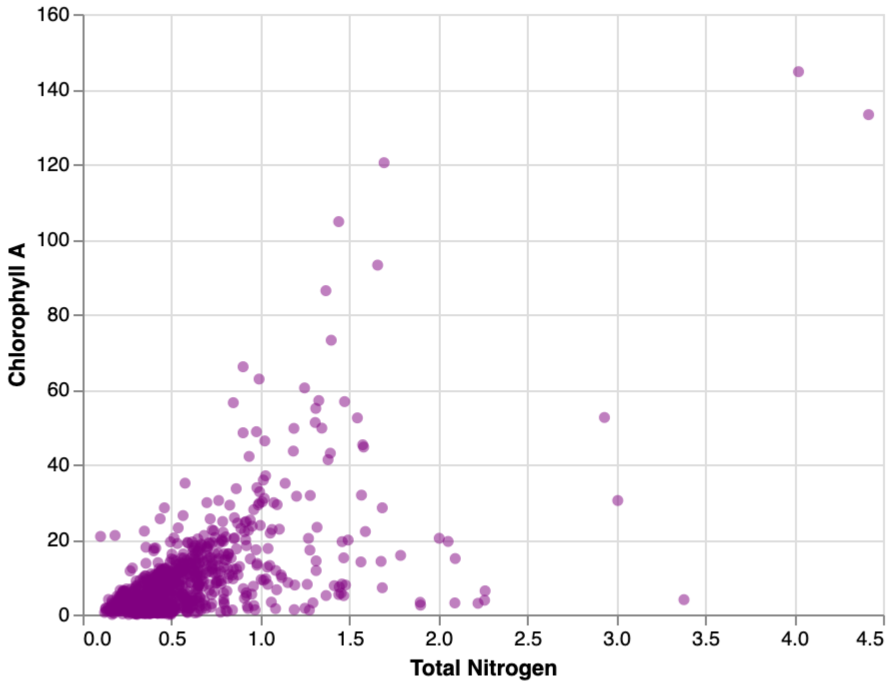
```
Correlation between Total Nitrogen and Chlorophyll A: 0.6411651592236679
```

In [12]:

```
#create visual representation of N productivity
n_scatter = alt.Chart(pivot_data).mark_circle(opacity=0.5, color='purple').e
    x=alt.X('Total Nitrogen', scale = alt.Scale(zero = False)),
    y=alt.Y('Chlorophyll A', scale = alt.Scale(zero = False))
)

n_scatter
```

Out[12]:



In [13]:

```
#study phosphorous productivity
corr_phosphorus = pivot_data['Total Phosphorus'].corr(pivot_data['Chlorophyl
print("Correlation between Total Phosphorus and Chlorophyll A:", corr_phosph
```

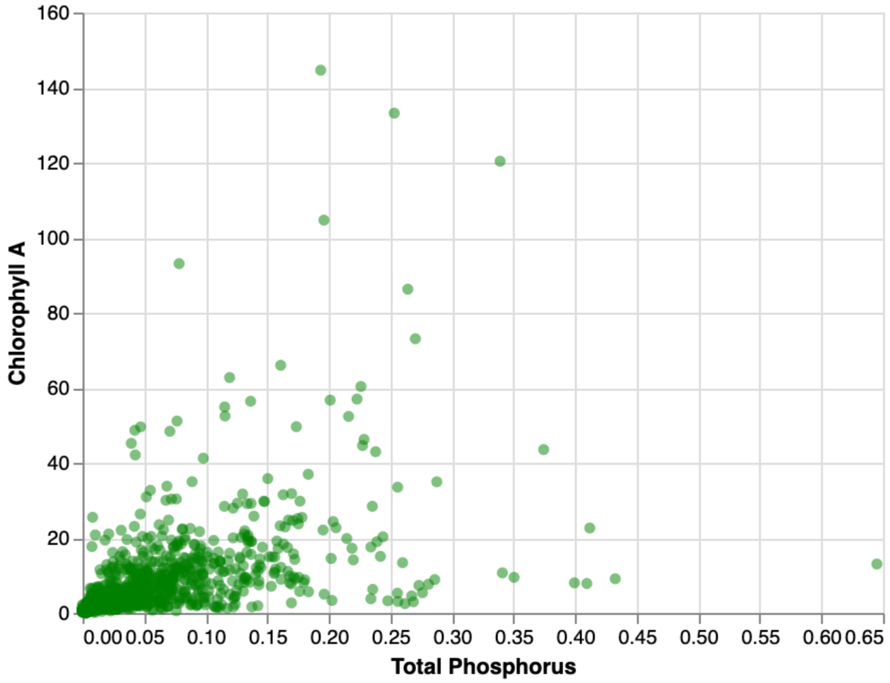
```
Correlation between Total Phosphorus and Chlorophyll A: 0.512930505369619
```

In [14]:

```
#create visual representation of P productivity
p_scatter = alt.Chart(pivot_data).mark_circle(opacity=0.5, color='green').en
    x=alt.X('Total Phosphorus', scale = alt.Scale(zero = False)),
    y=alt.Y('Chlorophyll A', scale = alt.Scale(zero = False))
)

p_scatter
```

Out[14]:



In [15]:

```
#now check relationship among the two nutrients
corr_nutrients = pivot_data['Total Nitrogen'].corr(pivot_data['Total Phospho
print("Correlation between Total Nitrogen and Total Phosphorus:", corr_nutri
```

```
Correlation between Total Nitrogen and Total Phosphorus: 0.5660930496417647
```

In [16]:

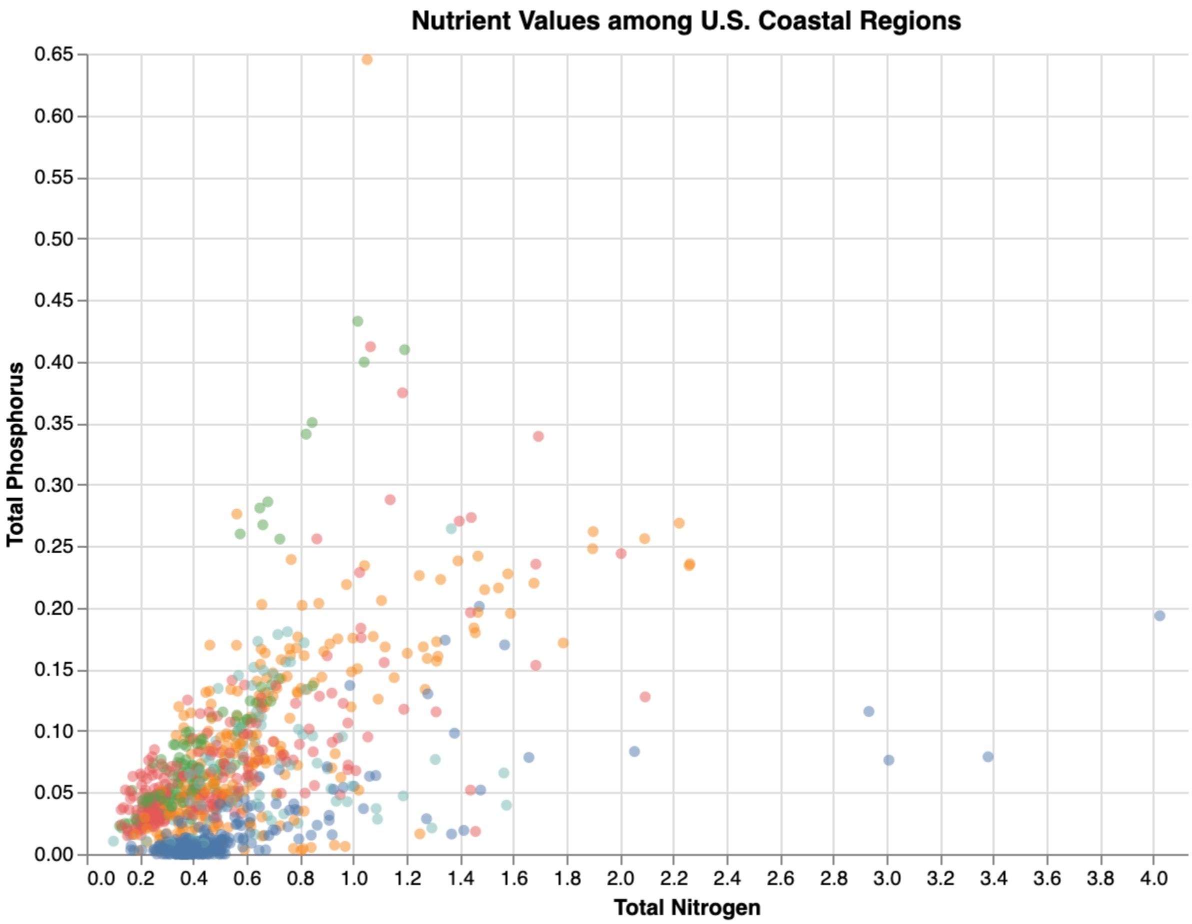
```
#code for part (ii)
```

In [17]:

```
#plot nutrient values colored by region to see regional differences
scatter_nutrient = alt.Chart(pivot_data).mark_circle(opacity=0.5).encode(
    x=alt.X('Total Nitrogen', title='Total Nitrogen', scale = alt.Scale(zero
    y=alt.Y('Total Phosphorus', title='Total Phosphorus', scale = alt.Scale(
    color='NCCR_REG:N'
).properties(
    width=600,
    height=400,
    title='Nutrient Values among U.S. Coastal Regions'
)

scatter_nutrient
```

Out[17]:

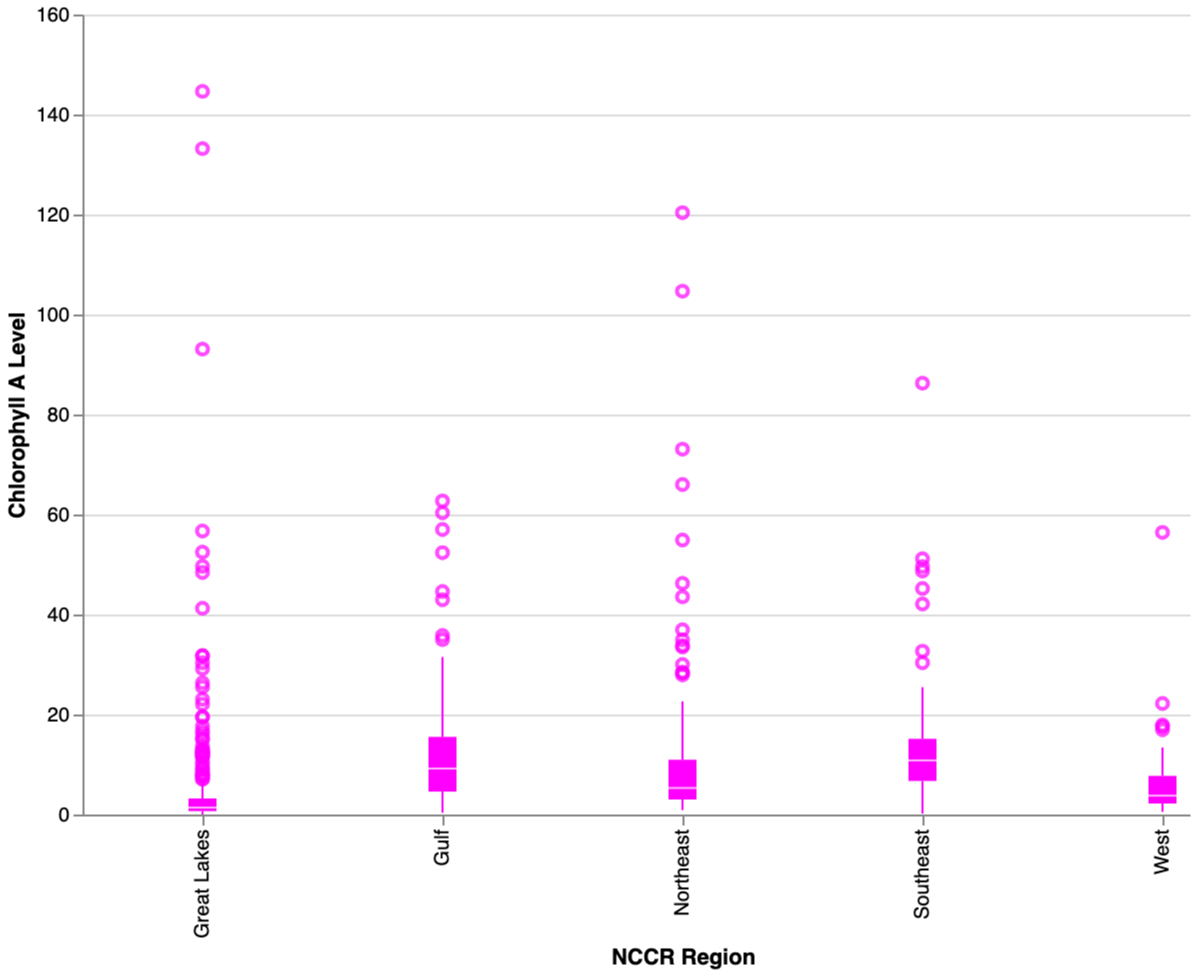


In [18]:

```
#check chlorophyll A levels in each NCCA region
box_plot = alt.Chart(pivot_data).mark_boxplot(color='magenta').encode(
    x=alt.X('NCCR_REG', title='NCCR Region'),
    y=alt.Y('Chlorophyll A:Q', title='Chlorophyll A Level')
).properties(
    width=600,
    height=400
)

box_plot
```

Out[18]:



In [19]:

```
#code for part (iv)
```

In [20]:

```
#make mini dataframe for just california observations
ca = pivot_data[pivot_data['STATE'] == 'CA']
```

In [21]:

```
#group data by date and calculate average Chlorophyll A
seasonal_productivity = ca.groupby('DATE_COL')['Chlorophyll A'].mean()
```

In [22]:

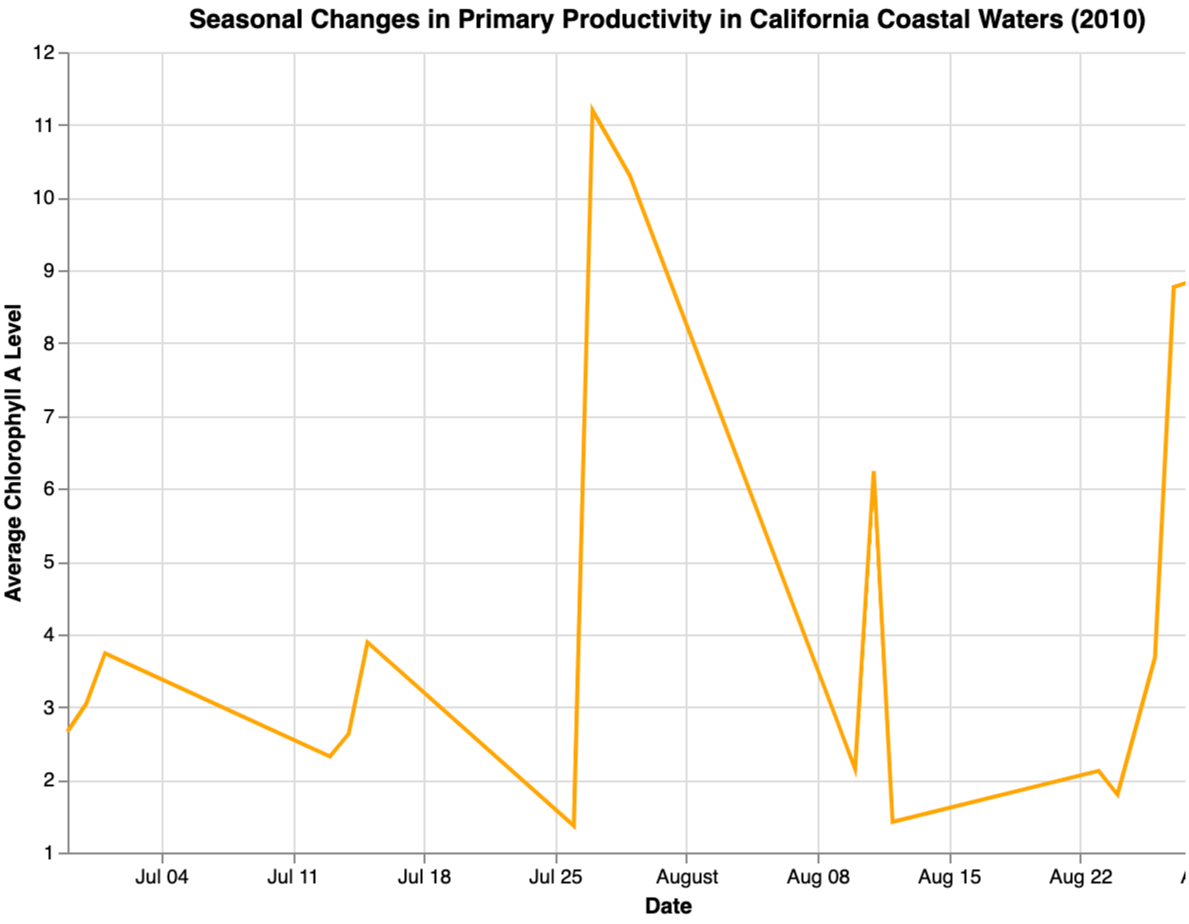
```
#convert seasonal productivity data into a dataframe object to use for chart
ca_data = pd.DataFrame({
    'DATE_COL': ['6/29/2010', '6/30/2010', '7/1/2010', '7/13/2010', '7/14/20
                 '7/26/2010', '7/27/2010', '7/29/2010', '8/9/2010', '8/10/20
                 '8/23/2010', '8/24/2010', '8/26/2010', '8/27/2010', '8/30/2
    'Chlorophyll A': [2.663333, 3.042500, 3.736000, 2.320000, 2.630000, 3.88
                      11.200000, 10.300000, 2.840000, 2.149333, 6.240000, 1.
                      3.680000, 8.770000, 9.028000, 4.870000, 4.752500]
})
```

In [23]:

```
#line plot to show mean chlorophyll A levels through time to study seasonal
seasonal_changes = alt.Chart(ca_data).mark_line().encode(
    x=alt.X('DATE_COL:T', title='Date', scale = alt.Scale(zero = False)),
    y=alt.Y('Chlorophyll A:Q', title='Average Chlorophyll A Level', scale =
    color=alt.value('orange')
).properties(
    width=600,
    height=400,
    title='Seasonal Changes in Primary Productivity in California Coastal Wa
)

seasonal_changes
```

Out[23]:


Seasonal Changes in Primary Productivity in California Coastal Waters (2010)

In [24]:

```
#code for part (v)
```

In [25]:

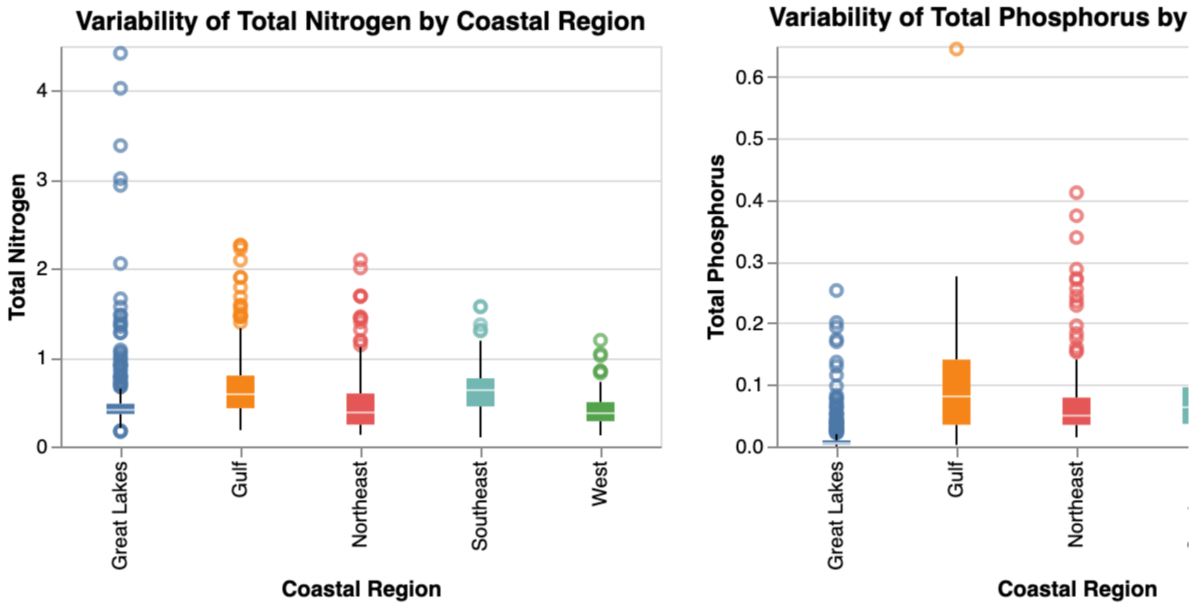
```
#box plot to visualize total nitrogen variability by region
boxplot_n = alt.Chart(pivot_data).mark_boxplot().encode(
    x=alt.X('NCCR_REG:N', title='Coastal Region'),
    y=alt.Y('Total Nitrogen:Q', title='Total Nitrogen', scale=alt.Scale(zerc
    color=alt.Color('NCCR_REG:N', title='Coastal Region')
).properties(
    width=300,
    height=200,
    title='Variability of Total Nitrogen by Coastal Region'
)


#box plot to visualize total phosphorous variability by region
boxplot_p = alt.Chart(pivot_data).mark_boxplot().encode(
    x=alt.X('NCCR_REG:N', title='Coastal Region'),
    y=alt.Y('Total Phosphorus:Q', title='Total Phosphorus', scale=alt.Scale(
    color=alt.Color('NCCR_REG:N', title='Coastal Region')
).properties(
    width=300,
    height=200,
    title='Variability of Total Phosphorus by Coastal Region'
)

#box plot to visualize total chlorophyll a variability by region
boxplot_c = alt.Chart(pivot_data).mark_boxplot().encode(
    x=alt.X('NCCR_REG:N', title='Coastal Region'),
    y=alt.Y('Chlorophyll A:Q', title='Chlorophyll A', scale=alt.Scale(zero=F
    color=alt.Color('NCCR_REG:N', title='Coastal Region')
).properties(
    width=300,
    height=200,
    title='Variability of Chlorophyll A by Coastal Region'
)

#plot side-by-side
(boxplot_n | boxplot_p | boxplot_c)
```

Out[25]:


Variability of Total Nitrogen by Coastal Region
Total Nitrogen
Great Lakes
Gulf
Northeast
Southeast
West
Coastal Region
Variability of Total Phosphorus by
Total Phosphorus
Great Lakes
Gulf
Northeast
Coastal Region