mp2-ncca

June 2, 2023

1 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.

1.1 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()`.

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

[3]:
```
Item  UID Date collected      Region  Latitude  Longitude  Chlorophyll A
0      59       7/1/2010  West Coast  32.77361 -117.21471           3.34  \
1      60       7/1/2010  West Coast  32.71424 -117.23527           2.45
2      61       7/1/2010  West Coast  32.78372 -117.22132           3.82
3      62       7/1/2010  West Coast  32.72245 -117.20443           6.13
4      63       6/9/2010  East Coast  34.75098  -77.12117           9.79

Item  Total Nitrogen  Total Phosphorus
0            0.40750          0.061254
1            0.23000          0.037379
2            0.33625          0.048100
3            0.23875          0.044251
4            0.63250          0.090636
```

In this dataset, I find the followings to be key attributes: "UID," "Date collected," 'Result,' 'Units,' 'Region,' 'Latitude,' and 'Longitude', 'Chlorophyll A,' 'Total Nitrogen,' and 'Total Phosphorus.' (I have modifed the namses). "UID" is a unique identifier for the site/visit, which can be beneficial in terms of distinguishing the data. "Date collected" refers to the date the site visit occurred, which can be useful for analyzing the data in a specific period in 2010. "Region," "Latitude," and "Longitude" are useful when dealing with geographical analysis. "Chlorophyll A," "Total Nitrogen," and "Total Phosphorus" are the nutrient that we are gonna analysize.

1.2 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) Based on the scatter plots, the sqrt of Chlorophyll A and Nitrogen have a distinct postive correlation. The sqrt of Chlorophyll A and Phosphorus also have a distinct postive correlation. In

addition, the correlation matrix highlights the relationships between sqrt-transformed Chlorophyll A, Total Nitrogen, and Total Phosphorus. The positive correlations indicate that higher levels of Chlorophyll A are associated with higher levels of both Nitrogen and Phosphorus. Additionally, there is a moderate positive correlation between Nitrogen and Phosphorus. The correlations are shown in the correlation matrix and heat map. These findings support the understanding that nutrient availability, specifically Nitrogen and Phosphorus, can influence the productivity of aquatic ecosystems, as reflected by the Chlorophyll A levels.

ii) Based on the bar-charts, there are noteable differences in available nutrients among U.S. coastal regions. The Gulf Coast have the most abundant average Nitrogen and Phosphorus. For Nitrogen, the values for other three are quite close, with West Coast being the lowest. However, for Phosphorus the West Coast has the second highest average value, with Great Lakes significantly being lower than other regions.

iii) Based on the 2010 data, the productivity seem to vary geographically. As Longitude increases, the value of Chlorophyll A also increases slightly. In contrast, as Latitude increases, the value of Chlorophyll A decreases to a large extent. In addition, there are clusters and gradients in different longitude and latitude, which suggest that productivity may vary geographically.

iv) Based on the line chart, Chlorophyll A is monotonically increasing as the Month moves from May to October. In particular, the increases from May to June and from September to October are dramatic, while the increases from Jun to September are more moderate. The observed seasonal increase in Chlorophyll A, with higher concentrations from May to October in California coastal waters, aligns with the expected pattern of higher primary productivity during warmer months. This result is intuitive, as it reflects the influence of environmental factors and seasonal variations on algal growth and photosynthesis.

v) Additional question: How does the chlorophyll concentration vary between different coastal regions in the United States? Based on the bar chart, the average Chlorophyll A is most abundant in Gulf Coast and East Coast. In contrast, the average is quite low for West Coast and Great Lakes, which are merely half as the other two. One interesting finding is that Gulf Coast has the highest value for all the nutrients we have analyzed in this dataset.

2 Code appendix

[1]:
```
import pandas as pd
import numpy as np
import altair as alt
import statsmodels.api as sm
alt.data_transformers.disable_max_rows()
alt.renderers.enable('mimetype')

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

[2]:
```
merged_data = pd.merge(ncca_raw, ncca_sites, on=['UID', 'SITE_ID','STATE'])

# selecting the columns of interest
columns_to_keep = [
```

```
    'UID',
    'DATE_COL_x',
    'PARAMETER',
    'PARAMETER_NAME',
    'RESULT',
    'UNITS',
    'NCA_REGION',
    'ALAT_DD',
    'ALON_DD'
]
merged_data = merged_data[columns_to_keep]
merged_data = merged_data.rename(columns={
    'DATE_COL_x': 'Date collected',
    'PARAMETER_NAME': 'Item',
    'PARAMETER' : 'Parameter',
    'RESULT': 'Result',
    'UNITS': 'Units',
    'NCA_REGION': 'Region',
    'ALAT_DD': 'Latitude',
    'ALON_DD': 'Longitude'
})

selected_parameters = ['Chlorophyll A', 'Total Nitrogen', 'Total Phosphorus']
filtered_data = merged_data[merged_data['Item'].isin(selected_parameters)]

# pivot the data
merged_data = filtered_data.pivot_table(index=['UID', 'Date collected',␣
 ↪'Region', 'Latitude', 'Longitude'],
                                        columns='Item', values='Result').
 ↪reset_index()
merged_data
```

[2]:

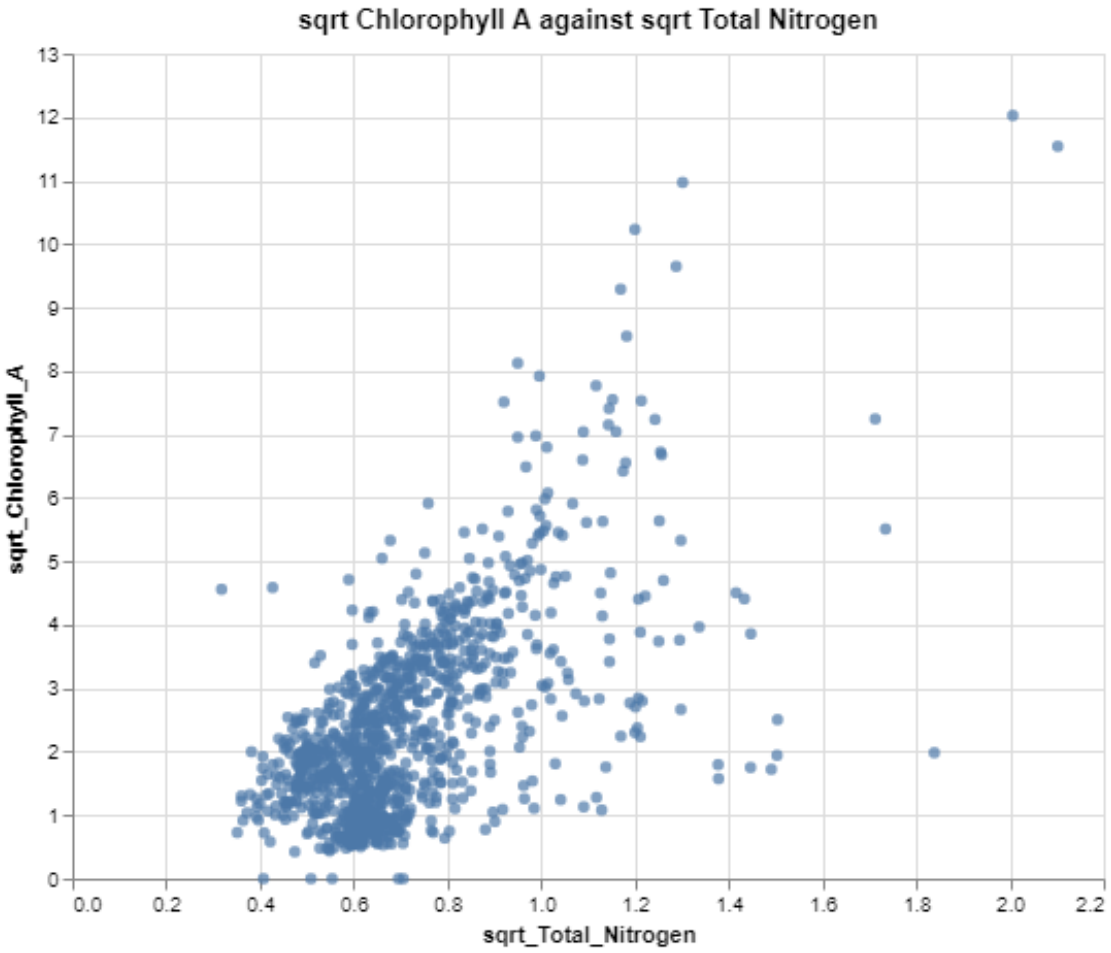
Item	UID	Date collected	Region	Latitude	Longitude	Chlorophyll A	\
0	59	7/1/2010	West Coast	32.77361	-117.21471	3.34	\
1	60	7/1/2010	West Coast	32.71424	-117.23527	2.45	
2	61	7/1/2010	West Coast	32.78372	-117.22132	3.82	
3	62	7/1/2010	West Coast	32.72245	-117.20443	6.13	
4	63	6/9/2010	East Coast	34.75098	-77.12117	9.79	
…	…	…	…	…	…	…	
1087	16727	6/18/2010	Great Lakes	44.98607	-85.64046	0.75	
1088	16728	6/25/2010	Great Lakes	44.94789	-85.94790	2.27	
1089	16729	6/16/2010	Great Lakes	44.83721	-85.52862	1.11	
1090	16730	6/29/2010	West Coast	32.66443	-117.13879	2.11	
1091	16731	6/29/2010	West Coast	32.66243	-117.12712	2.19	

Item	Total Nitrogen	Total Phosphorus
0	0.407500	0.061254

```
1           0.230000          0.037379
2           0.336250          0.048100
3           0.238750          0.044251
4           0.632500          0.090636
…                 …                 …
1087        0.380000          0.000000
1088        0.437625          0.006249
1089        0.361250          0.000000
1090        0.213000          0.044127
1091        0.228750          0.041821

[1092 rows x 8 columns]
```

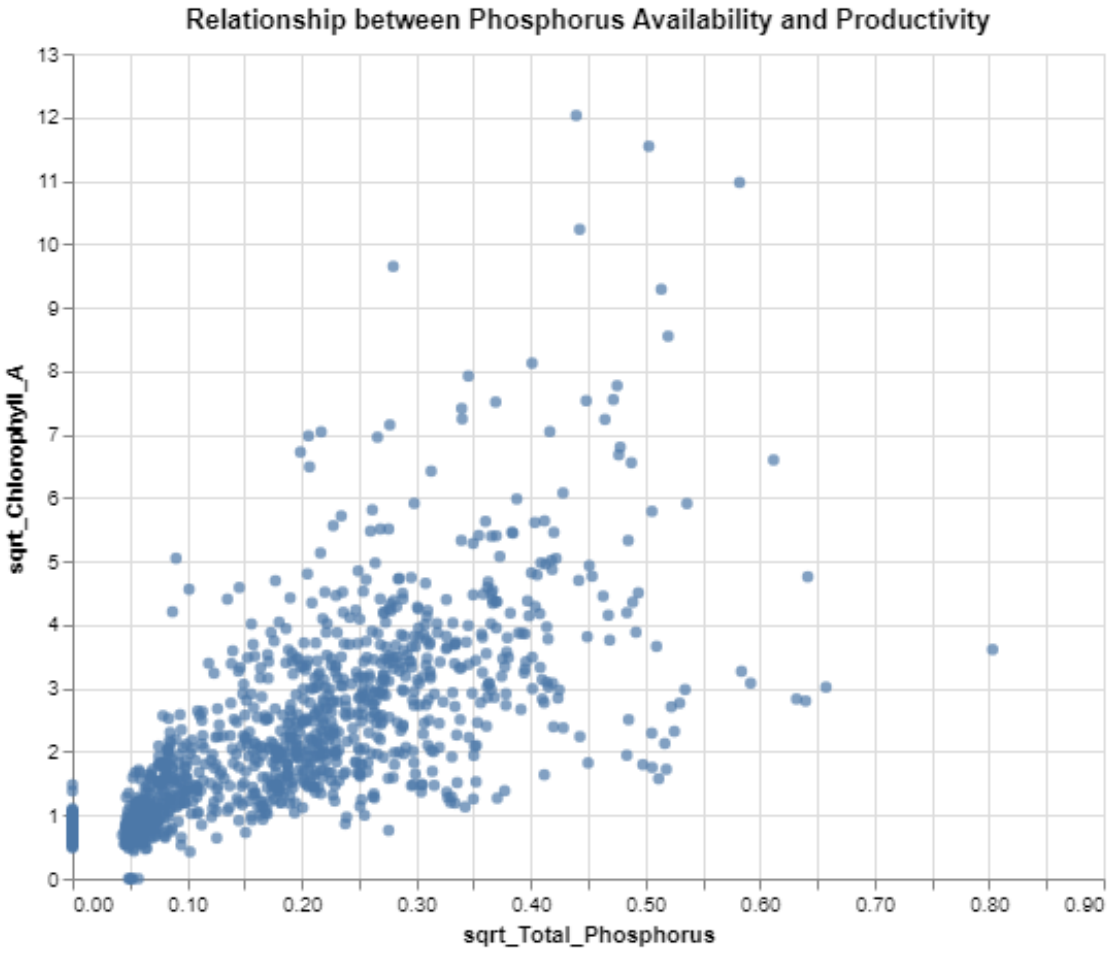
[4]:
```
# i
merged_data['sqrt_Total_Nitrogen'] = np.sqrt(merged_data['Total Nitrogen'])
merged_data['sqrt_Chlorophyll_A'] = np.sqrt(merged_data['Chlorophyll A'])

N2_chart = alt.Chart(merged_data).mark_circle().encode(
    x='sqrt_Total_Nitrogen',
    y='sqrt_Chlorophyll_A',
    tooltip=['Total Nitrogen', 'Chlorophyll A']
).properties(
    width = 500,
    height = 400,
    title = "sqrt Chlorophyll A against sqrt Total Nitrogen"
)

N2_chart
# There's clearly a postive trend between sqrt(Chlorophyll A) and sqrt(Total␣
 ↪Nitrogen)
```

[4]:



[5]:
```
merged_data['sqrt_Total_Phosphorus'] = np.sqrt(merged_data['Total Phosphorus'])

chart_phosphorus = alt.Chart(merged_data).mark_circle().encode(
    x='sqrt_Total_Phosphorus',
    y='sqrt_Chlorophyll_A',
    tooltip=['Total Phosphorus', 'Chlorophyll A']
).properties(
    width=500,
    height=400,
    title='Relationship between Phosphorus Availability and Productivity'
)

chart_phosphorus

# Here we also observe a postive trend between sqrt(ChlorophyllA) and␣
 ↪sqrt(Total Phosphorus)
```

[5]:

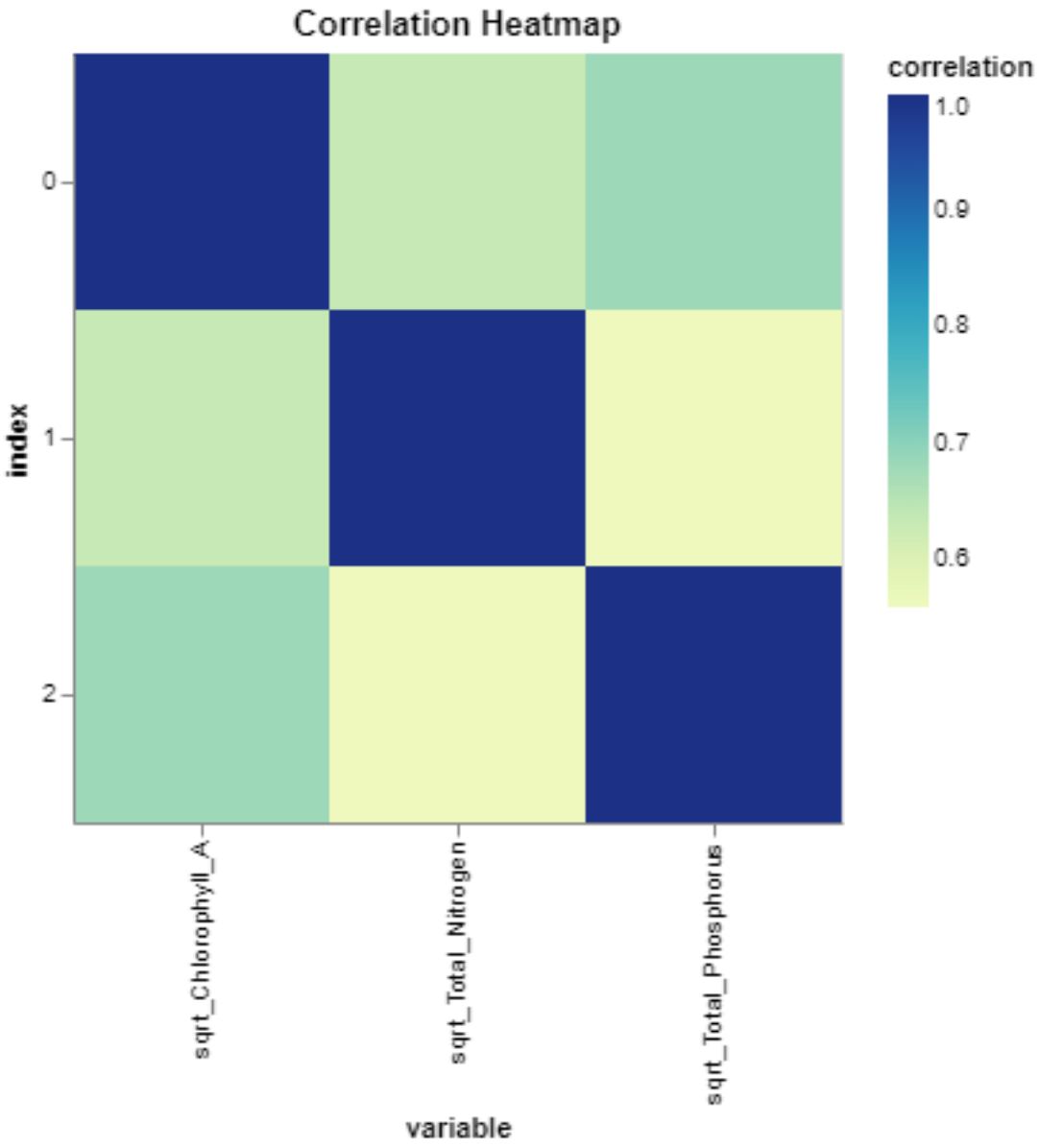


[6]:
```
correlation_matrix = merged_data[['sqrt_Chlorophyll_A', 'sqrt_Total_Nitrogen',␣
↪'sqrt_Total_Phosphorus']].corr()
correlation_matrix

# these nutrients are correlated based on the correlation matrix.
```

[6]:
```
Item                  sqrt_Chlorophyll_A  sqrt_Total_Nitrogen
Item
sqrt_Chlorophyll_A              1.000000            0.630860  \
sqrt_Total_Nitrogen             0.630860            1.000000
sqrt_Total_Phosphorus           0.680913            0.557689

Item                  sqrt_Total_Phosphorus
Item
sqrt_Chlorophyll_A                 0.680913
sqrt_Total_Nitrogen                0.557689
sqrt_Total_Phosphorus              1.000000
```

[7]:
```
# Create the correlation matrix
correlation_matrix = merged_data[['sqrt_Chlorophyll_A', 'sqrt_Total_Nitrogen',␣
 ↪'sqrt_Total_Phosphorus']].corr()

# Assign sequential indices to the correlation matrix
correlation_matrix['index'] = range(len(correlation_matrix))

# Melt the dataframe to convert it to long format
correlation_melted = correlation_matrix.melt(id_vars='index',␣
 ↪var_name='variable', value_name='correlation')

# Create the heatmap
heatmap = alt.Chart(correlation_melted).mark_rect().encode(
    x='variable:O',
    y='index:O',
    color='correlation:Q',
    tooltip=['index', 'variable', 'correlation:Q']
).properties(
    width=300,
    height=300,
    title='Correlation Heatmap'
)

heatmap
# This heat map confirms that the correlation between these nutrients are␣
 ↪positive
```

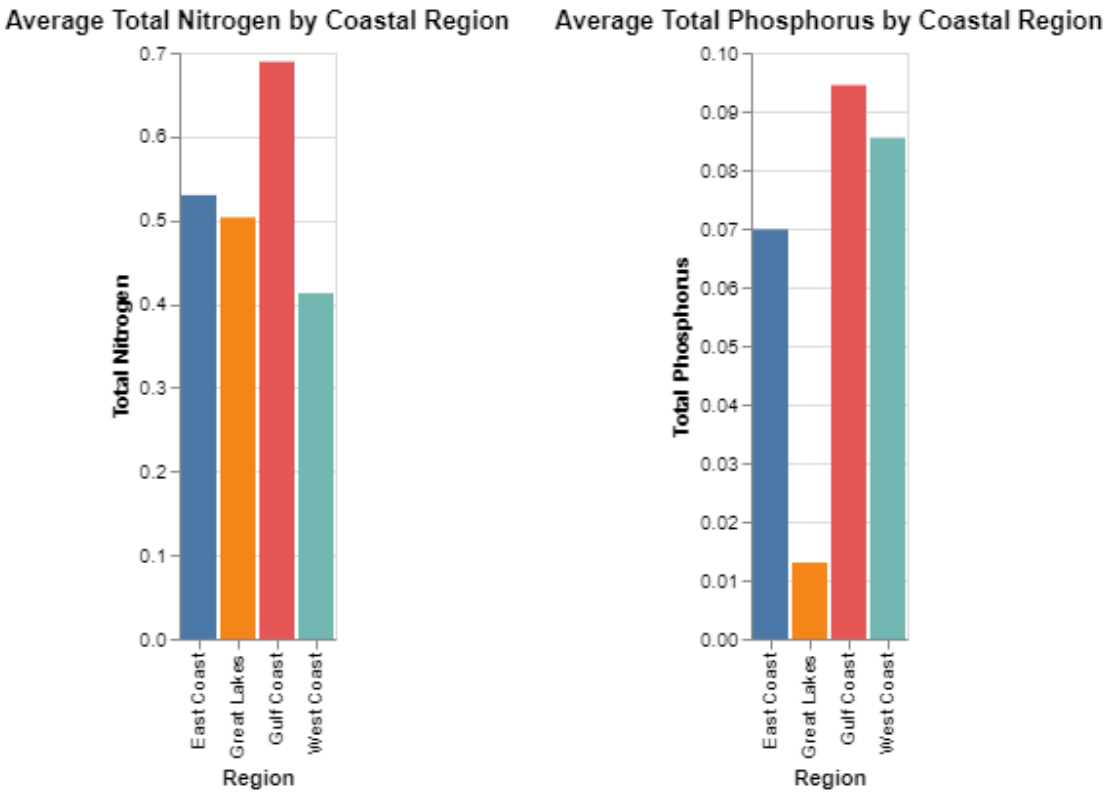
[7]:



[8]:
```
# ii)
region_summary = merged_data.groupby('Region').agg({
    'Total Nitrogen': 'mean',
    'Total Phosphorus': 'mean'
}).reset_index()

# create a bar plot for Total Nitrogen
bar_nitrogen = alt.Chart(region_summary).mark_bar().encode(
    x='Region',
    y='Total Nitrogen',
    color=alt.Color('Region', legend=None),
```

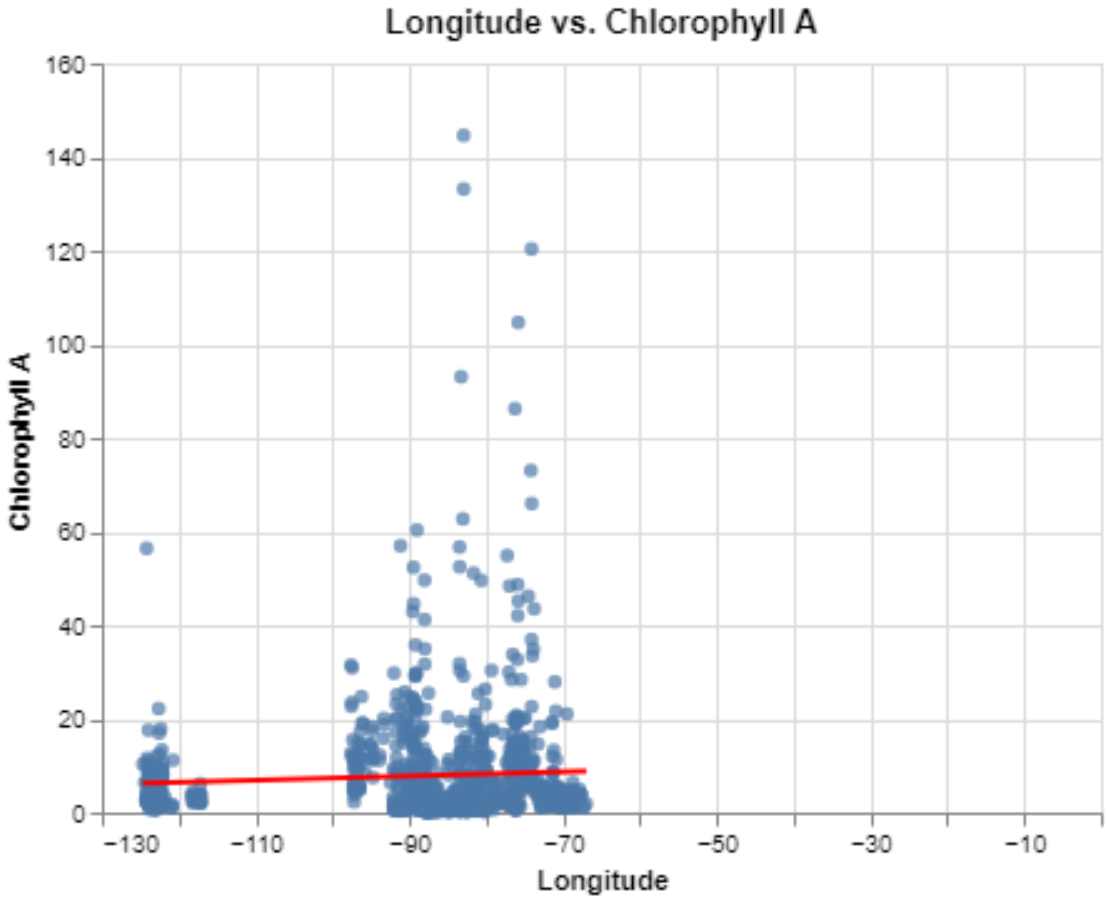
```
    tooltip=['Region', 'Total Nitrogen']
).properties(
    title='Average Total Nitrogen by Coastal Region'
)

# create a bar plot for Total Phosphorus
bar_phosphorus = alt.Chart(region_summary).mark_bar().encode(
    x='Region',
    y='Total Phosphorus',
    color=alt.Color('Region', legend=None),
    tooltip=['Region', 'Total Phosphorus']
).properties(
    title='Average Total Phosphorus by Coastal Region'
)

# combine the plots
combined_plot = (bar_nitrogen | bar_phosphorus).
 ↪resolve_scale(color='independent')

# display the combined plot
combined_plot
```

[8]:



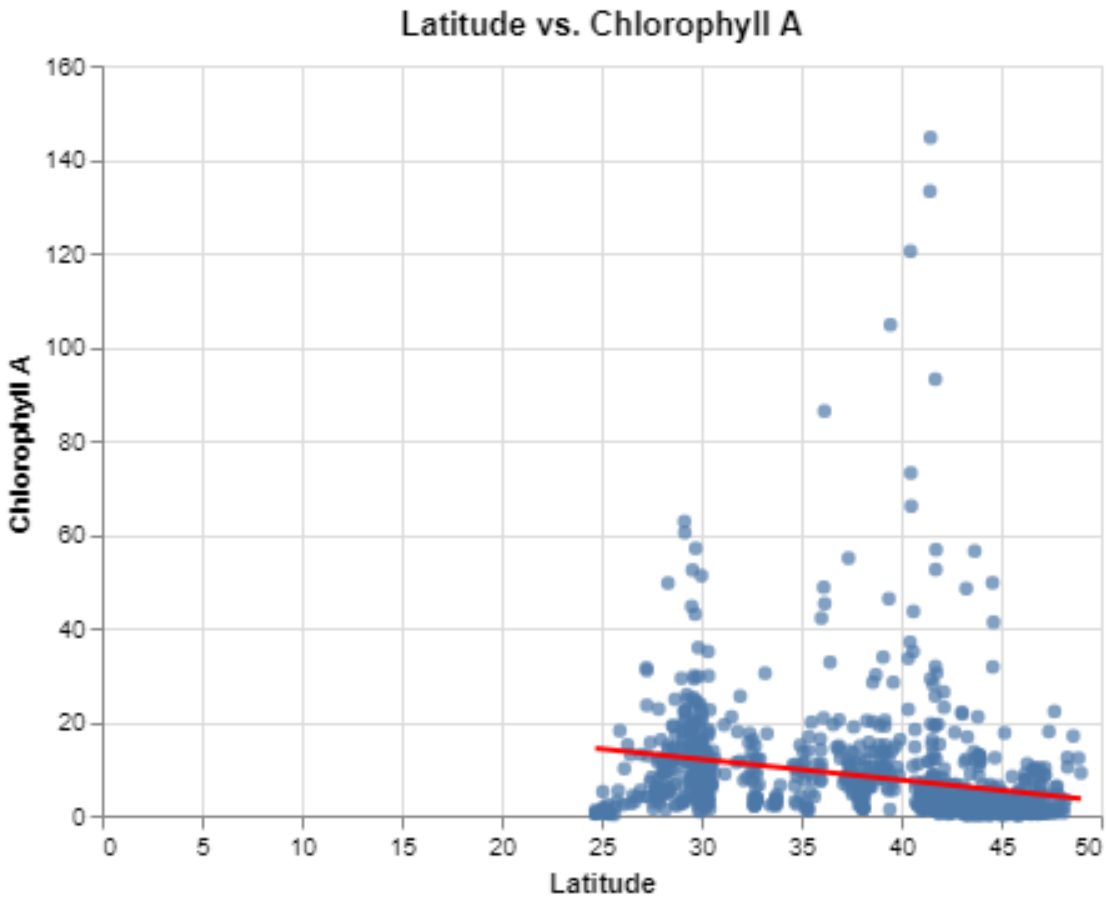
[9]:
```
# iii)
# scatter plot of Longitude vs. Chlorophyll A
scatter_longitude = alt.Chart(merged_data).mark_circle().encode(
    x='Longitude',
    y='Chlorophyll A',
    tooltip=['Longitude', 'Chlorophyll A']
).properties(
    title='Longitude vs. Chlorophyll A'
)

# add a regression line
reg_line_longitude = scatter_longitude.transform_regression(
    'Longitude', 'Chlorophyll A'
).mark_line(color='red')

# combine scatter plot and regression line
scatter_with_regression_longitude = (scatter_longitude + reg_line_longitude)

scatter_with_regression_longitude

# There's a slightly positive trend between Chlorophyll A and Longtitude.
```

[9]:



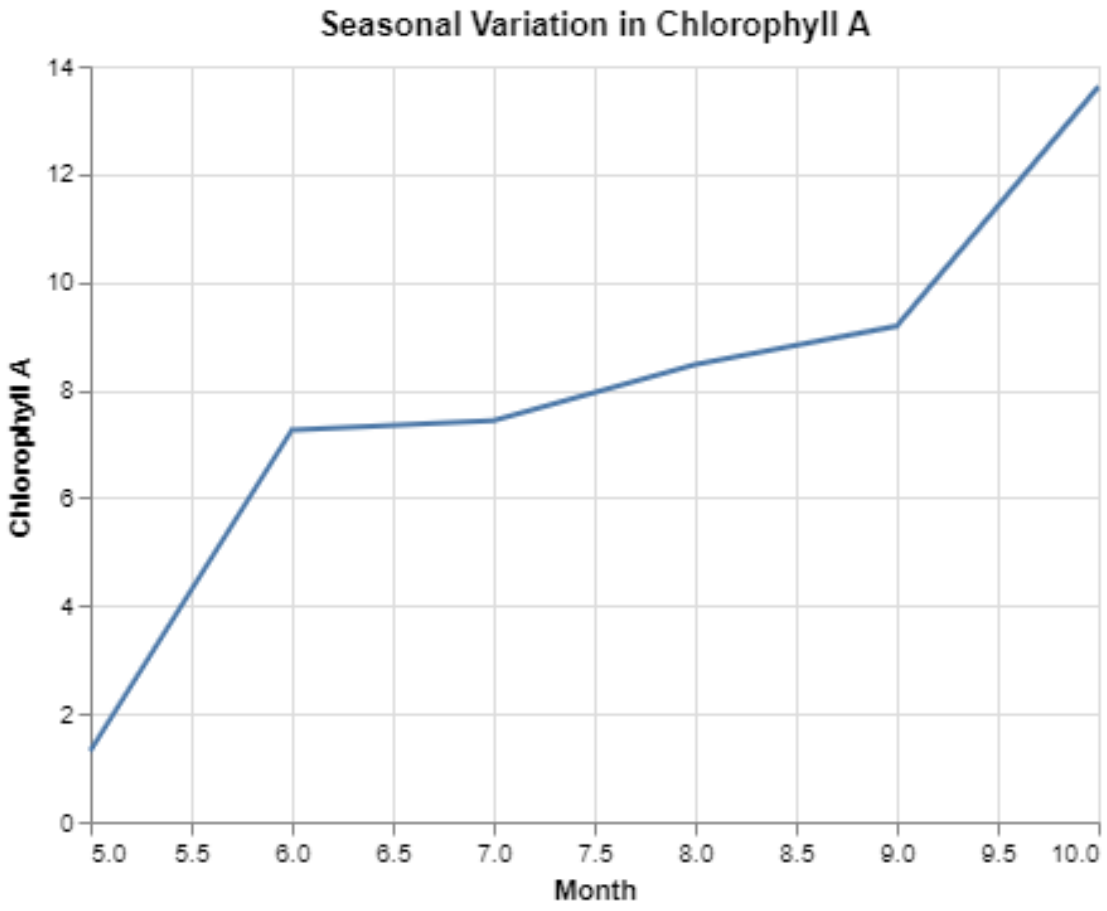
[10]:
```
# scatter plot of Latitude vs. Chlorophyll A
scatter_latitude = alt.Chart(merged_data).mark_circle().encode(
    x='Latitude',
    y='Chlorophyll A',
    tooltip=['Latitude', 'Chlorophyll A']
).properties(
    title='Latitude vs. Chlorophyll A'
)

# add a regression line
reg_line_latitude = scatter_latitude.transform_regression(
    'Latitude', 'Chlorophyll A'
).mark_line(color='red')

# combine scatter plot and regression line
scatter_with_regression_latitude = (scatter_latitude + reg_line_latitude)
scatter_with_regression_latitude

# There's a clear negative trend between Chlorophyll A and latitude.
```

[10]:



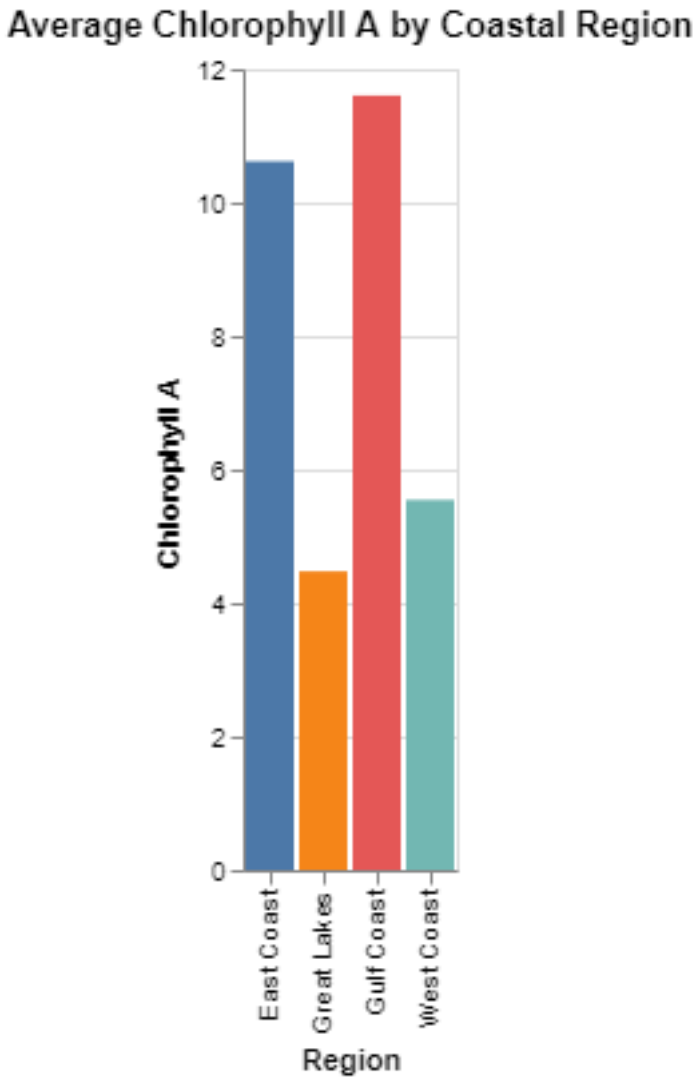
[11]:
```
# iv)
# extract month from the 'Date collected' column
merged_data['Month'] = pd.to_datetime(merged_data['Date collected']).dt.month

# calculate the monthly mean of Chlorophyll A
monthly_mean = merged_data.groupby('Month')['Chlorophyll A'].mean().
 ↪reset_index()

# create a line plot of monthly mean Chlorophyll A
Seasonal_plot = alt.Chart(monthly_mean).mark_line().encode(
    x='Month',
    y='Chlorophyll A',
    tooltip=['Month', 'Chlorophyll A']
).properties(
    title='Seasonal Variation in Chlorophyll A'
)

Seasonal_plot
# increasing trend
```

[11]:



[12]:
```
# iv) How does the chlorophyll concentration vary between different coastal␣
 ↪regions in the United States?
# calculate average chlorophyll concentration by region
region_summary = merged_data.groupby('Region')['Chlorophyll A'].mean().
 ↪reset_index()

# create a bar plot
bar_chart = alt.Chart(region_summary).mark_bar().encode(
    x='Region',
    y='Chlorophyll A',
    color=alt.Color('Region', legend=None),
    tooltip=['Region', 'Chlorophyll A']
).properties(
    title='Average Chlorophyll A by Coastal Region'
)

bar_chart
```

[12]:


Average Chlorophyll A by Coastal Region
12
10
8
6
4
2
0
Chlorophyll A
East Coast
Great Lakes
Gulf Coast
West Coast
Region

[ ]: