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.

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()`.

This ncca data is a dataset containing information about chemicals in multiple sites in the National Aquatic Resource Surveys. Each row has a specific site and the information about it (this includes things like the site id, state, date, chlorophyl, nh3, nitrogen, phosphorous, etc.) The columns take measured concentrations of the chemical measured and we can use this to make inferences about the dataset. The nitrogen and phophorous are the nutrients present in each site and the cholorphyl is the productivity measurement in the ecosystem.

Cleaning up the data, starts with mergining the two datasets, subsetted to the specific columns and then pivoting to reshape the data. The data is then converted to lower case, and then we create new columns of dates. Then we subset the dates farther where we only keep areas with the percent of missing values less than 1%.

In [20]:
```
ncca.head()
```

Out[20]:

	uid	site_id	state	date_col	chla	din	nh3	no3no2	ntl	ptl	srp	wt
0	59	NCCA10-1111	CA	7/1/2010	3.34	0.014	0.000	0.014	0.40750	0.061254	0.028	
1	60	NCCA10-1119	CA	7/1/2010	2.45	0.020	0.010	0.010	0.23000	0.037379	0.026	S
2	61	NCCA10-1123	CA	7/1/2010	3.82	0.009	0.000	0.009	0.33625	0.048100	0.030	
3	62	NCCA10-1127	CA	7/1/2010	6.13	0.010	0.000	0.010	0.23875	0.044251	0.028	S
4	63	NCCA10-1133	NC	6/9/2010	9.79	0.030	0.002	0.028	0.63250	0.090636	0.043	W

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.

1) Based on all the graphs(below), we can see there is a positive relationship between the cholorphyl and phosphorous, chlorophyl and nitrogen, and nitrogen and phosphorous. Based on the correlation coefficients we calculated, the strongest positive relationship is between the chloropyl and nitrogen.

2) There is a higher concentration of nitrogen in all regions, more than phosphorous. The amounts of nitrogen and phophorous are different based on the region. Gulf regions have the most nutrients. There is a bit of discrepancy with the second highest as west has the second higherest phosphorus but the lowest nitrogen. The great lakes is lowest for phosphorus but not for nitrogen.

3) There is a higher productivity of chlorophyl in different areas of the United States. The Southeast and Gulf regions have some of the highest concentrations of chlorophyl. This is more present on the eastern side of the United States, which is where there is less desert and more fertile area. Therefore we can say there is a geographical difference.

4) The primary productivity in california does change with the seasons. As we go into the later months (summer to spring), we can see that the average chlorophyl production increases, and almost doubles. We can probably associate this with the longer days of the summer and the higher presence of sun.

5) *Question: How does altitude change the primary productivity?*

Based on the graph it looks like the primary productivity of chlorphyl stays mostly the same throughout the graph, while there are a lot of outliers through the middle the net graph seems to stay around and below 20. However when we print out the average in ranges, it actually shows a negative trend that increased altitude has less chlorophyl.

Code appendix

In [2]:

```python
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')
ncca_sites.head()
```

Out[2]:

	UID	SITE_ID	STATE	VISIT_NO	DATE_COL	WTBDY_NM	SITESAMP	INDEX_VISIT	EPA_
0	59	NCCA10-1111	CA	1.0	1-Jul-10	Mission Bay	Y	Y	
1	60	NCCA10-1119	CA	1.0	1-Jul-10	San Diego Bay	Y	Y	
2	61	NCCA10-1123	CA	1.0	1-Jul-10	Mission Bay	Y	Y	
3	62	NCCA10-1127	CA	1.0	1-Jul-10	San Diego Bay	Y	Y	
4	63	NCCA10-1133	NC	1.0	9-Jun-10	White Oak River	Y	Y	

5 rows × 31 columns

In [3]:

```
#data clean up
ncca = pd.merge(
    ncca_raw.iloc[:, [0, 1, 2, 3, 5, 7]].pivot(
        index=['UID', 'SITE_ID', 'STATE', 'DATE_COL'],
        columns='PARAMETER',
        values='RESULT'
    ).reset_index(),
    ncca_sites.loc[:, ['UID', 'WTBDY_NM', 'NCCR_REG', 'ALAT_DD', 'ALON_DD']]
    how='left',
    on='UID'
)

ncca.columns = ncca.columns.str.lower()

ncca_dates = pd.concat([
    ncca,
    ncca.date_col.str.split(
        pat='/',
        n=3,
        expand=True
    ).rename(columns={0: 'month', 1: 'day', 2: 'year'})
], axis=1)

ncca = ncca_dates.loc[:, ncca_dates.isna().mean() < 0.01]
ncca.head()
```

Out[3]:

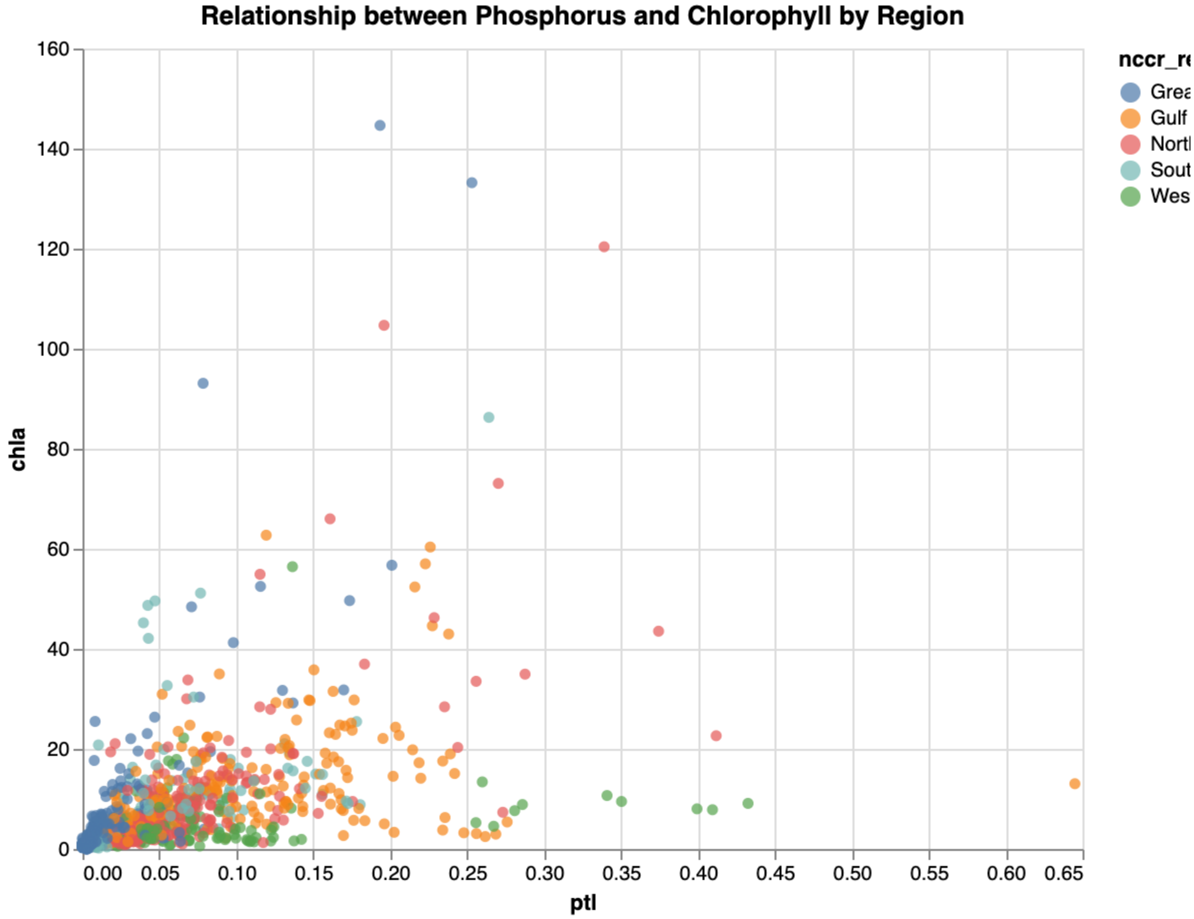
	uid	site_id	state	date_col	chla	din	nh3	no3no2	ntl	ptl	srp	wt
0	59	NCCA10-1111	CA	7/1/2010	3.34	0.014	0.000	0.014	0.40750	0.061254	0.028	
1	60	NCCA10-1119	CA	7/1/2010	2.45	0.020	0.010	0.010	0.23000	0.037379	0.026	S
2	61	NCCA10-1123	CA	7/1/2010	3.82	0.009	0.000	0.009	0.33625	0.048100	0.030	
3	62	NCCA10-1127	CA	7/1/2010	6.13	0.010	0.000	0.010	0.23875	0.044251	0.028	S
4	63	NCCA10-1133	NC	6/9/2010	9.79	0.030	0.002	0.028	0.63250	0.090636	0.043	W

In [4]:

```
#relationship between phosphorous and chlorphyl
fig1 = alt.Chart(ncca).mark_circle().encode(
    x='ptl',
    y='chla',
    color='nccr_reg',
    tooltip=['ptl', 'chla', 'nccr_reg']
).properties(
    width=500,
    height=400,
    title='Relationship between Phosphorus and Chlorophyll by Region'
)

fig1
```

Out[4]:



In [5]:

```
correlation = ncca['ptl'].corr(ncca['chla'])
print("Correlation coefficient between PTL and CHLA:", correlation)
```

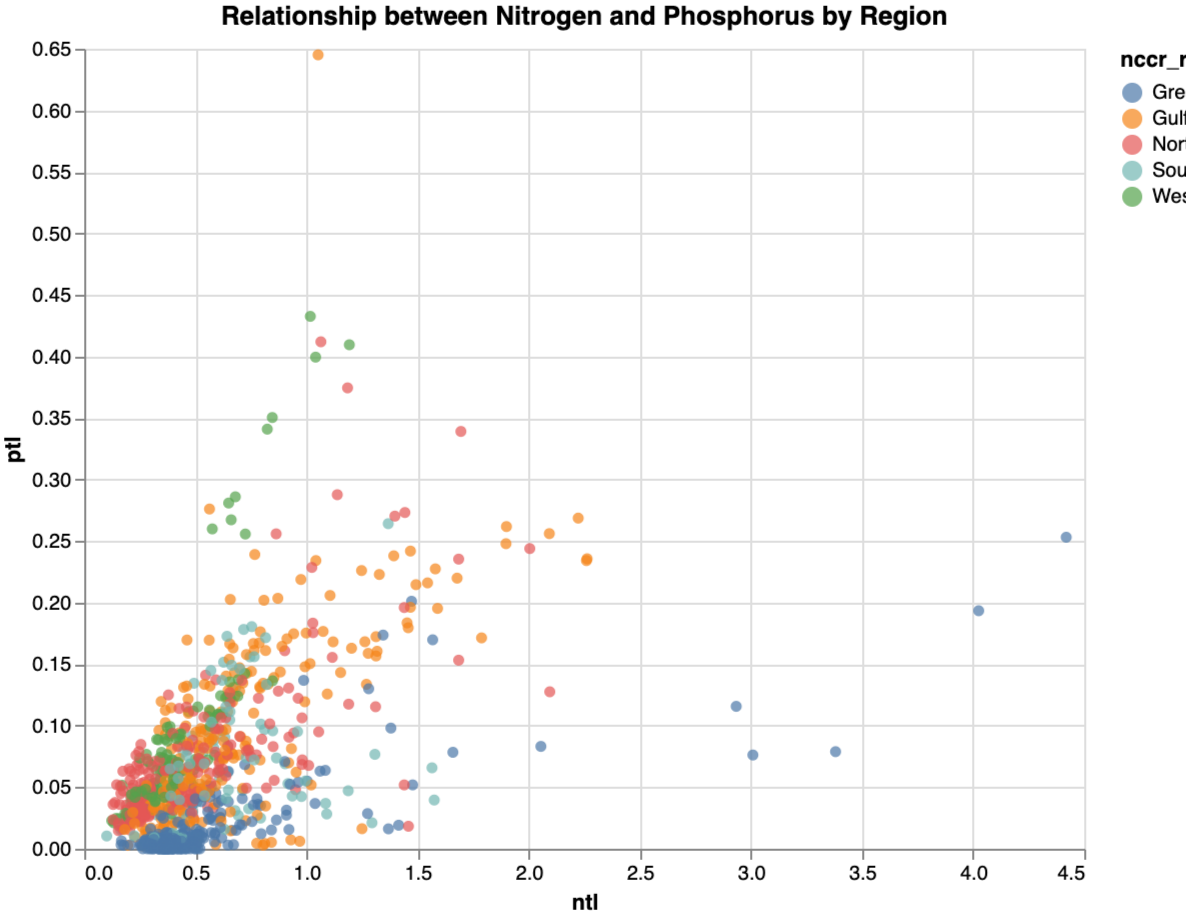
```
Correlation coefficient between PTL and CHLA: 0.512930505369619
```

In [6]:

```
# relationship between nitrogen and phopshorous
fig2 = alt.Chart(ncca).mark_circle().encode(
    x='ntl',
    y='ptl',
    color='nccr_reg',
    tooltip=['ntl', 'ptl', 'nccr_reg']
).properties(
    width=500,
    height=400,
    title='Relationship between Nitrogen and Phosphorus by Region'
)

fig2
```

Out[6]:



In [7]:

```
correlation = ncca['ptl'].corr(ncca['ntl'])
print("Correlation coefficient between PTL and NTL:", correlation)
```

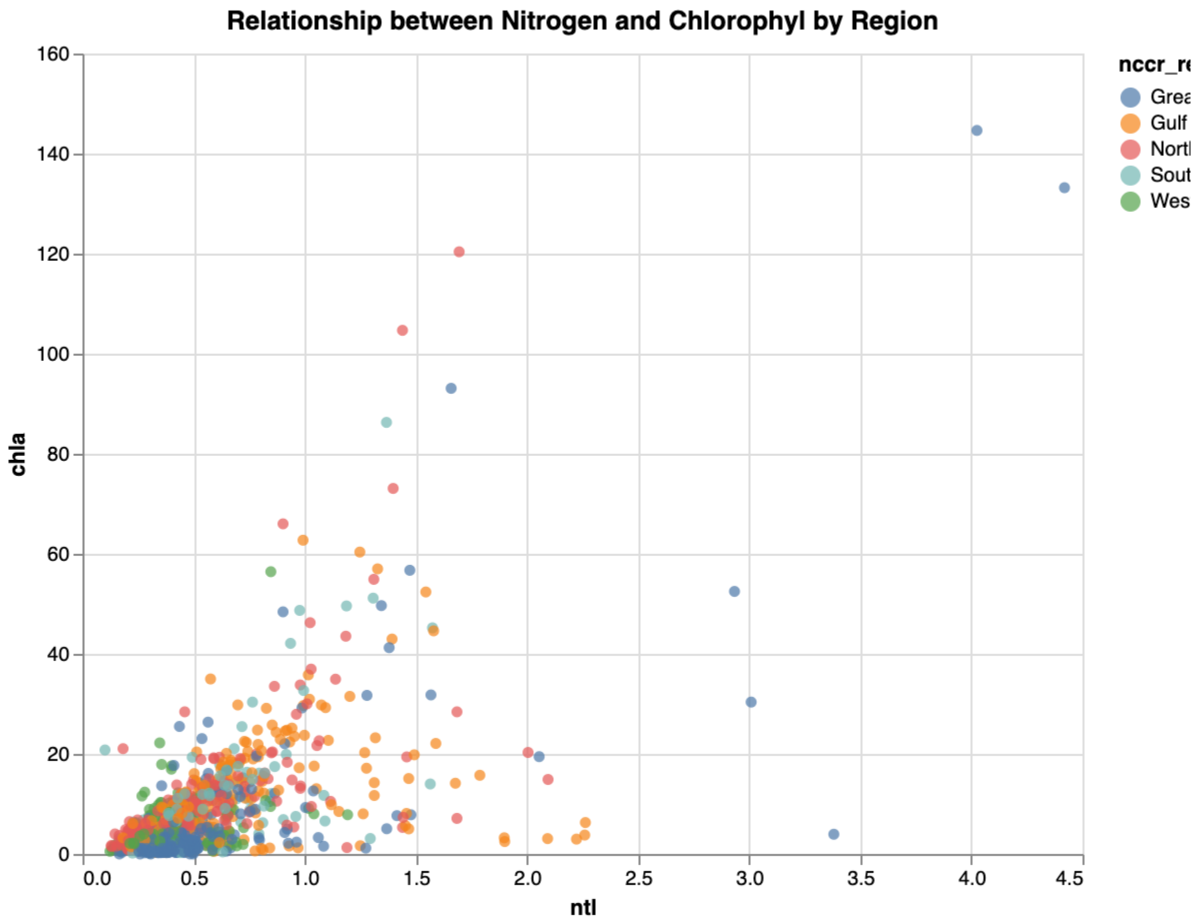
```
Correlation coefficient between PTL and NTL: 0.5660930496417647
```

In [8]:

```
# relationship between nitrogen and chlorophyll
fig3 = alt.Chart(ncca).mark_circle().encode(
    x='ntl',
    y='chla',
    color='nccr_reg',
    tooltip=['ntl', 'chla', 'nccr_reg']
).properties(
    width=500,
    height=400,
    title='Relationship between Nitrogen and Chlorophyl by Region'
)

fig3
```

Out[8]:



In [9]:

```
correlation = ncca['chla'].corr(ncca['ntl'])
print("Correlation coefficient between CHLA and NTL:", correlation)
```

```
Correlation coefficient between CHLA and NTL: 0.6411651592236679
```

Part 2

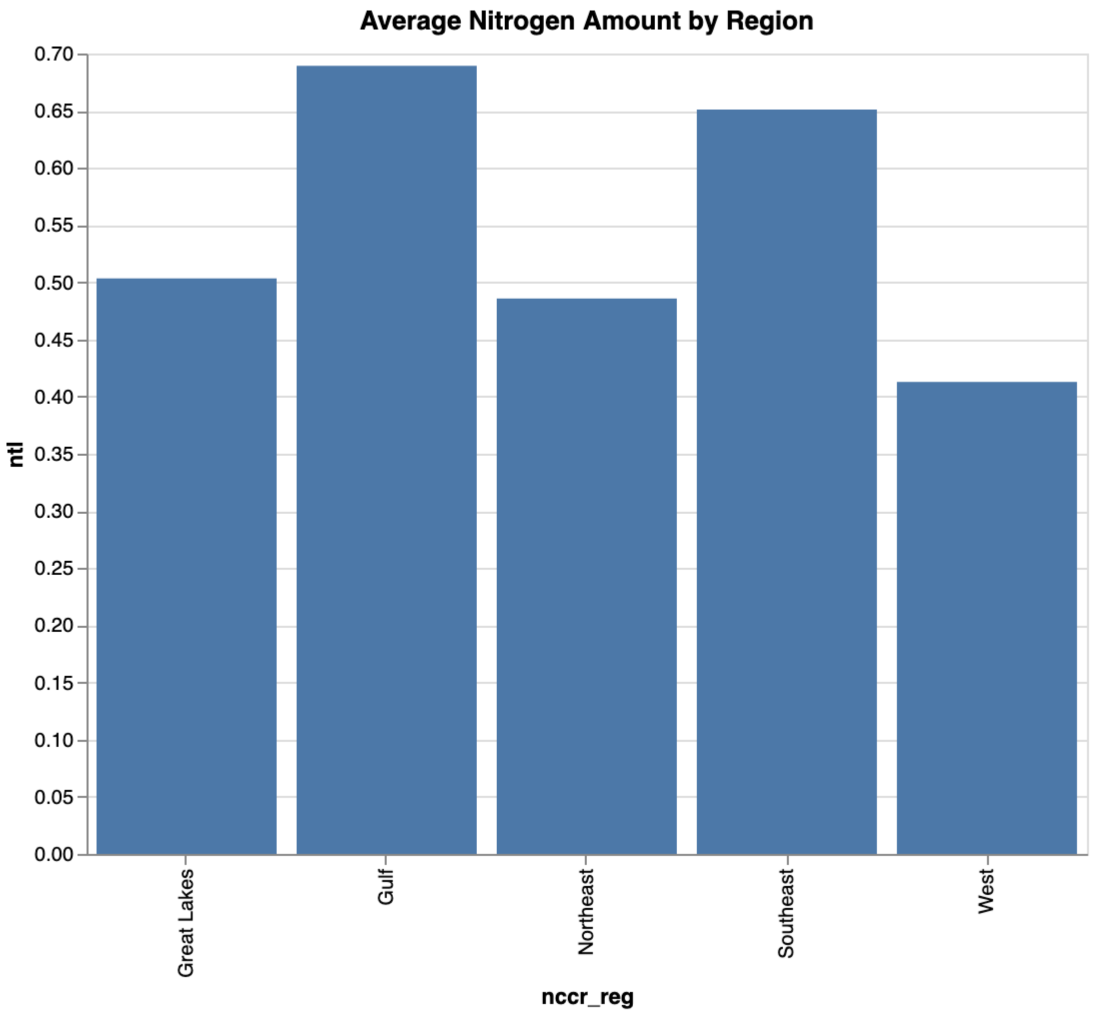
In [10]:

```
average_nitrogen_region = ncca.groupby('nccr_reg')['ntl'].mean().reset_index

fig4 = alt.Chart(average_nitrogen_region).mark_bar().encode(
    x='nccr_reg',
    y='ntl',
    tooltip=['nccr_reg', 'ntl']
).properties(
    width=500,
    height=400,
    title='Average Nitrogen Amount by Region'
)

fig4
```

Out[10]:



In [11]:
```
average_nitrogen_region
```

Out[11]:

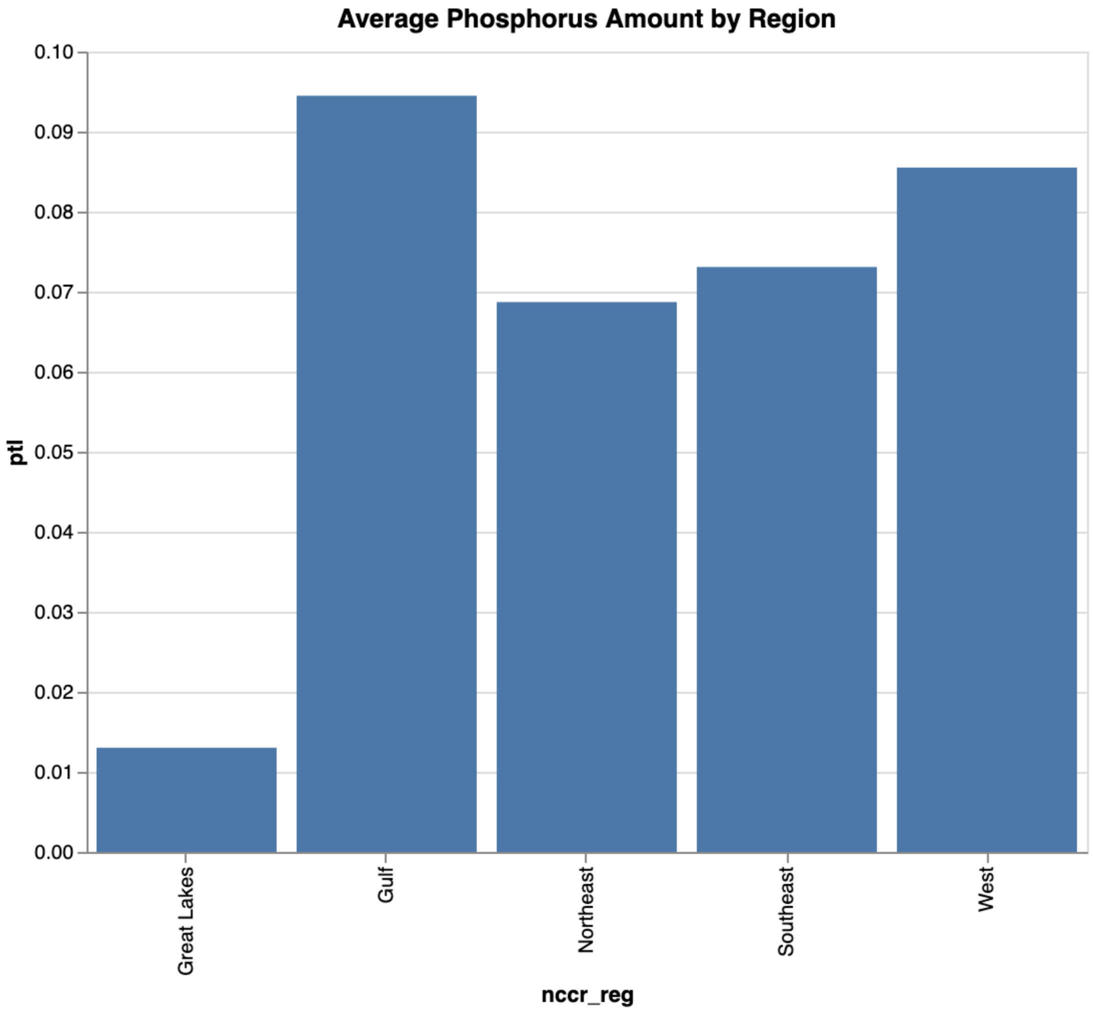
	nccr_reg	ntl
0	Great Lakes	0.503336
1	Gulf	0.689250
2	Northeast	0.485748
3	Southeast	0.650999
4	West	0.412794

In [12]:
```
average_phosphorus_region = ncca.groupby('nccr_reg')['ptl'].mean().reset_inc

fig5 = alt.Chart(average_phosphorus_region).mark_bar().encode(
    x='nccr_reg',
    y='ptl',
    tooltip=['nccr_reg', 'ptl']
).properties(
    width=500,
    height=400,
    title='Average Phosphorus Amount by Region'
)
```

```
fig5
```

Out[12]:



In [13]:

```
average_phosphorus_region
```

Out[13]:

	nccr_reg	ptl
0	Great Lakes	0.013019
1	Gulf	0.094474
2	Northeast	0.068696
3	Southeast	0.073080
4	West	0.085498

Part 3

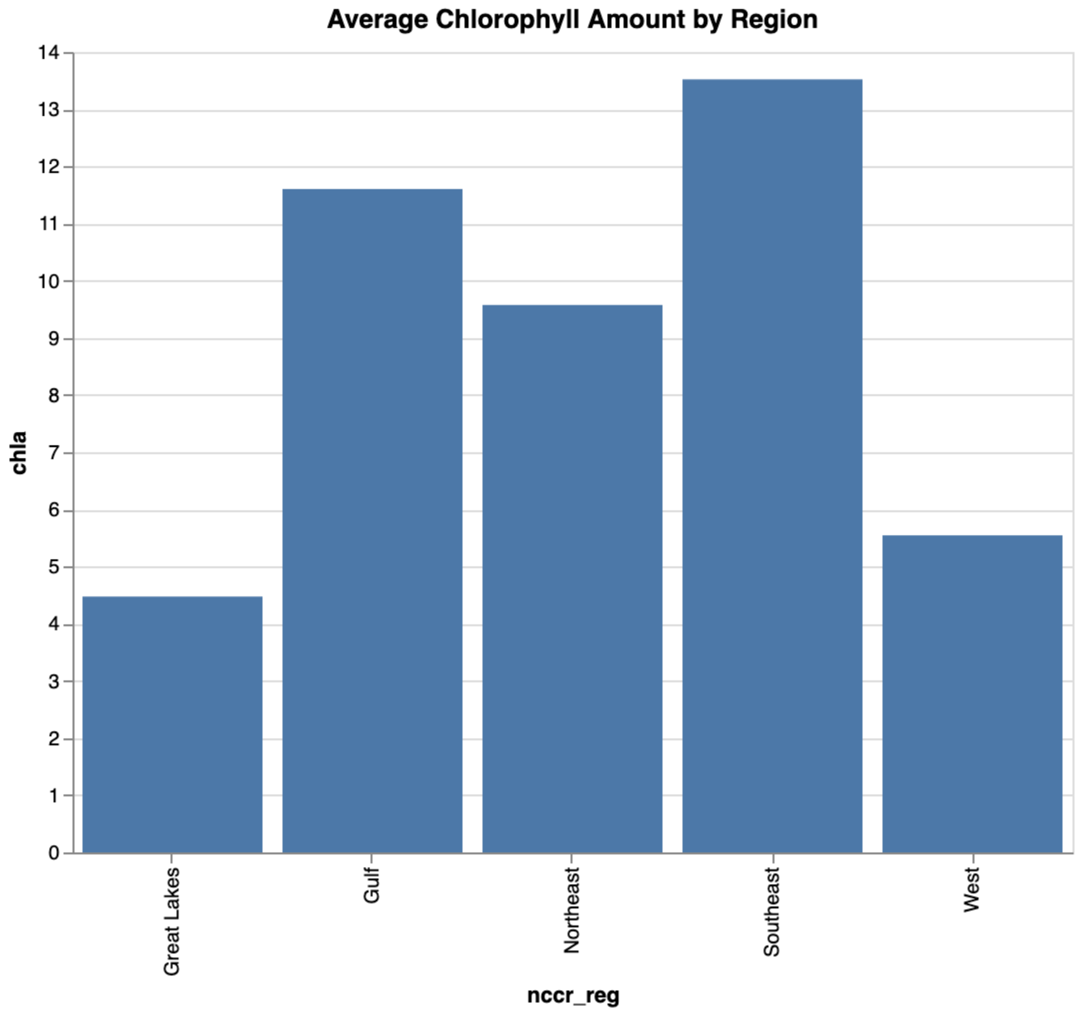
In [14]:

```
average_chlorophyll_region = ncca.groupby('nccr_reg')['chla'].mean().reset_i

fig6 = alt.Chart(average_chlorophyll_region).mark_bar().encode(
    x='nccr_reg',
    y='chla',
    tooltip=['nccr_reg', 'chla']
).properties(
    width=500,
    height=400,
    title='Average Chlorophyll Amount by Region'
)

fig6
```

Out[14]:



In [15]:

```
average_chlorophyll_region
```

Out[15]:

	nccr_reg	chla
0	Great Lakes	4.475248
1	Gulf	11.603818
2	Northeast	9.575352
3	Southeast	13.521707
4	West	5.545900

Part 4

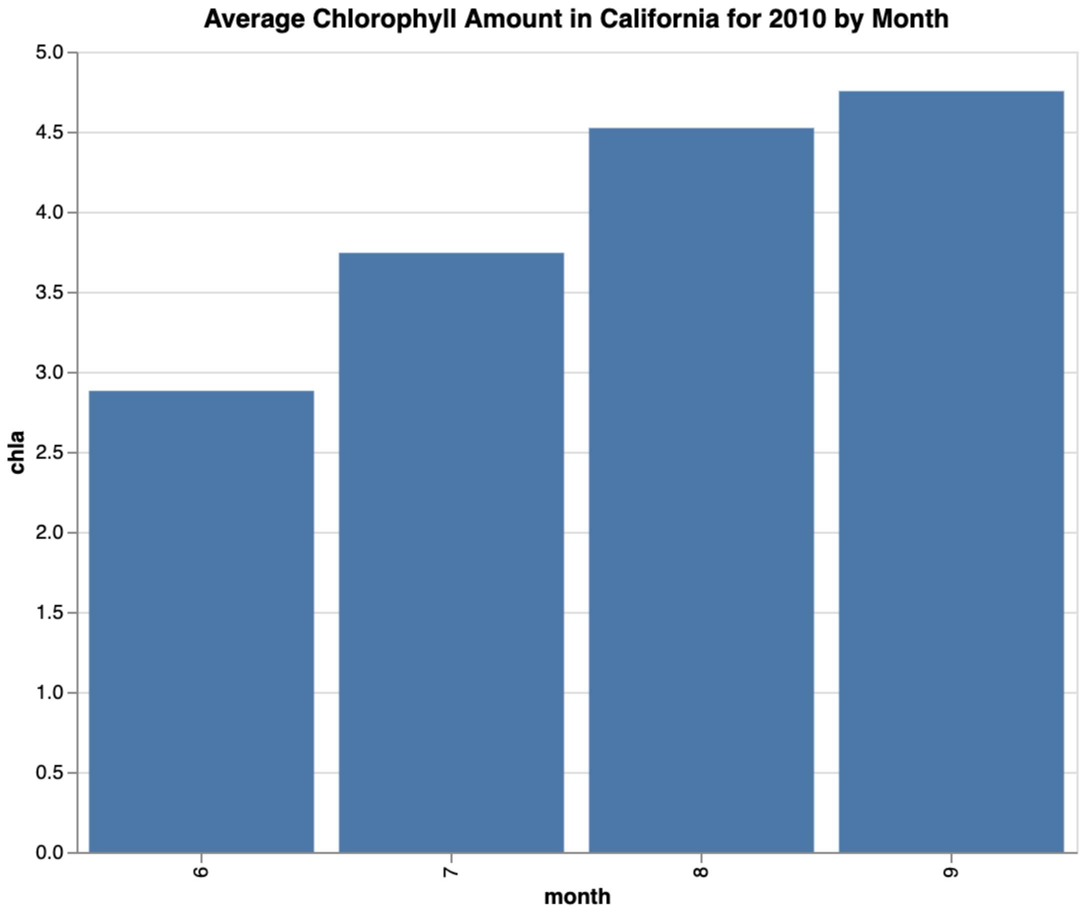
In [16]:

```
ca_average_chlorophyll_2010 = ncca[(ncca['state'] == 'CA') & (ncca['year'] =

fig7 = alt.Chart(ca_average_chlorophyll_2010).mark_bar().encode(
    x='month',
    y='chla',
    tooltip=['month', 'chla']
).properties(
    width=500,
    height=400,
    title='Average Chlorophyll Amount in California for 2010 by Month'
)

fig7
```

Out[16]:



In [17]:

```
ca_average_chlorophyll_2010
```

Out[17]:

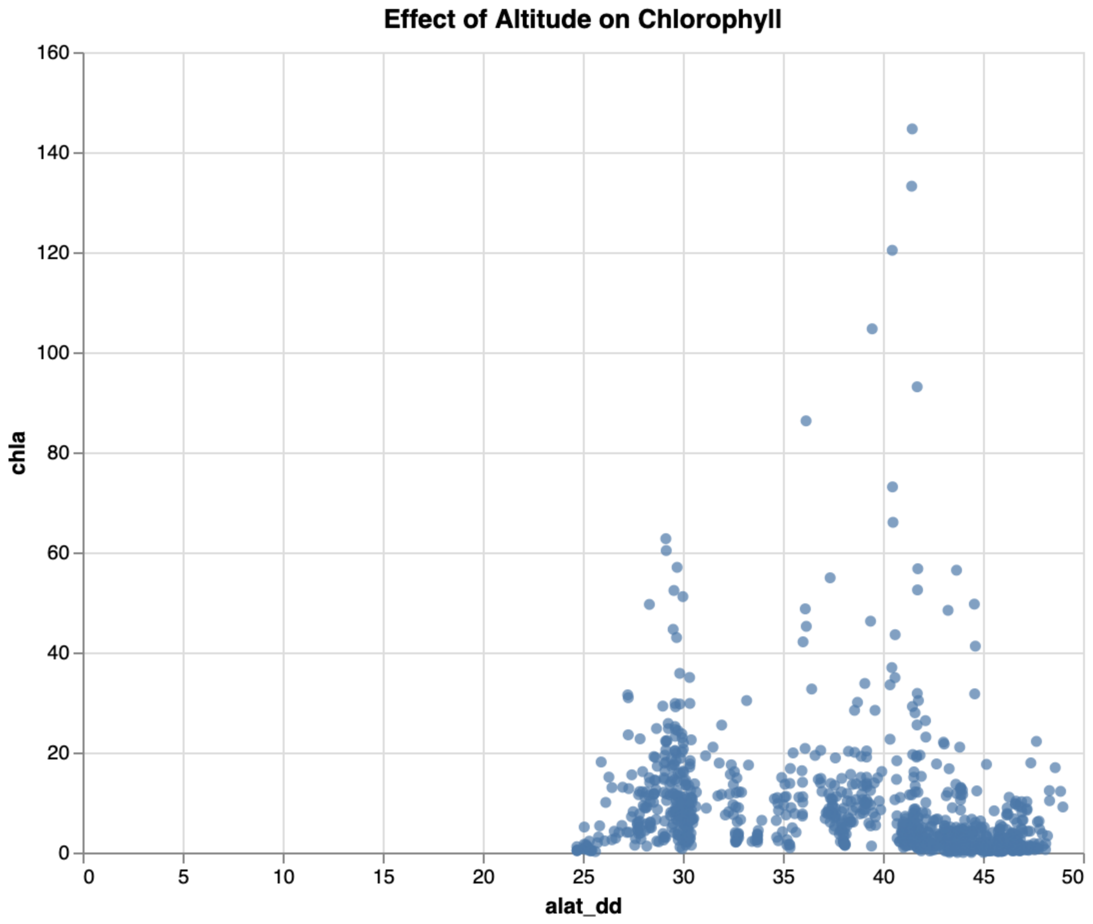
	month	chla
0	6	2.880000
1	7	3.742105
2	8	4.522000
3	9	4.752500

In [18]:

```
fig8 = alt.Chart(ncca).mark_circle().encode(
    x='alat_dd',
    y='chla',
    tooltip=['alat_dd', 'chla']
).properties(
    width=500,
    height=400,
    title='Effect of Altitude on Chlorophyll'
)

fig8
```

Out[18]:



In [19]:

```
altitude_ranges = [(25, 30), (31, 40), (41, 50)]

for start, end in altitude_ranges:
    average_chlorophyll = ncca[(ncca['alat_dd'] >= start) & (ncca['alat_dd']
    print(f"Altitude Range: {start} - {end} meters")
    print(f"Average Chlorophyll: {average_chlorophyll}\n")
```

```
Altitude Range: 25 - 30 meters
Average Chlorophyll: 12.325145013076925

Altitude Range: 31 - 40 meters
Average Chlorophyll: 11.408424676616916

Altitude Range: 41 - 50 meters
Average Chlorophyll: 4.7047784341594685
```