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 [51]:

```
# show a few rows of clean data
data.head()
```

Out[51]:

	UID	SITE_ID	STATE	Date collected	Nutrient formula	Nutrient	Nutrient Amount	UNITS	Date Anlyzed	Day he
0	59	NCCA10-1111	CA	7/1/2010	NTL	Total Nitrogen	0.407500	mg N/L	7/14/2010	13
1	59	NCCA10-1111	CA	7/1/2010	NO3NO2	Nitrate/Nitrite	0.014000	mg N/L	7/8/2010	7
2	59	NCCA10-1111	CA	7/1/2010	SRP	Dissolved Inorganic Phosphate	0.028000	mg P/L	7/8/2010	7
3	59	NCCA10-1111	CA	7/1/2010	DIN	Dissolved Inorganic Nitrogen	0.014000	mg N/L	NaN	Na
4	59	NCCA10-1111	CA	7/1/2010	PTL	Total Phosphorus	0.061254	mg P/L	7/14/2010	13

I would say that significant columns from both datasets were the UID (Unique Idenifier), site_ID, and state as those there catagories correspond with each other. Secondly, the Nutrient Formula, Nutrient namek, and Nutrient amount are also key attributes in the ncca_raw dataset as it provides numerical calulations of each nutrient detectected for each UID. Lastly, the water sampled, region, and province, are key attributes in the ncca_sites dataset as it provides specifical georgraphic locations for their corresponding UID numbers.

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.

Of the 28 states thar provided information about nutrient availabiltiy, only 50% of all nutrients in our dataset can be found in all these states and in very small amounts. These are: Ammonia, Dissolved inorganic Nitrogen, Dissolved inorganic Phosphate, Nitrate, Total Nitrogen and Total Phosphorus. The other 50% can only be found in specific states with Virginia(VA) having amounts of almost all of the remaining 50% of nutirents not found in other states; dissolved silica, nitrate, nitrite, nitrogen particulate, phosporus particulate, Total dissolved Nitrogen and Total dissolved Phosphorus. Chlorophyll A is the only nutrient with large amounts in each state which makes sense since it serves as a proxy for primary productionin marine ecosystems.

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

After filtering our 'Region' column and pulling out all the rows that contain the name "coast" we found given data on three coastal regions: East Coast, West Coast, and the Gulf Coast (under texas). Notable differences among these coasts are that the East Coast is the only coast to contain Dissolved Silica, Nitrate, Nitrite, Nitrogen Particulate, Phosphorus Particulate, Total Dissolved Nitrogen, and Total Dissolved Phosphorus. The East Coast also has larger amounts if Chlorophyll which may contribute to why all the other nutrients are found in the East Coast.

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.

Given that we found the East coast to produce a majority of our nutrients we found that Virginia has almost all of available nutrients within one state which may contribute to why nutrients are detected in our coastal graph. This may contribute to the extra amount of chlorophyll produced in the east coast.

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

Does your result make intuitive sense?

We sort the dataset by pulling all rows with state CA. We plot by Water source to see where exactly nutrients are coming from. We see that primary productivity only slightly varys seasonally with Chlorophyll seeing the biggest variation compared to other nutrients seeing barely any change.

Pose and answer one additional question.

Why are my TA's so cool?

- Its because they have mad swag and have great smiles. :)

P.S. (I ran out of time lolz)

Codes

In [1]:
```
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 [2]:
```
##Part 1 - cleaning data
#ncca_raw #7876 values
ncca_raw.isna().sum()
```

Out[2]:
```
UID                  0
SITE_ID              0
STATE                0
DATE_COL             0
BATCH_ID             0
PARAMETER            0
PARAMETER_NAME       0
RESULT               0
UNITS                0
MDL               1092
MRL               4088
PQL               7722
DATE_ANALYZED     1144
HOLDING_TIME      1414
QACODE            4660
LAB_SAMPLE_ID     1354
SAMPLE_ID         1191
METHOD            7700
dtype: int64
```

In [ ]:
```
#ncca_sites #1104
ncca_sites.isna().sum()
```

Out[ ]:
```
UID                       0
SITE_ID                   0
STATE                     0
VISIT_NO                 10
DATE_COL                 10
WTBDY_NM                  0
SITESAMP                  0
INDEX_VISIT               0
EPA_REG                   0
NCCR_REG                  0
NCA_REGION                0
COUNTRY                   0
PROVINCE                  0
STATION_DEPTH            12
STATION_DEPTH_UNITS      12
ALAT_DD                   0
ALON_DD                   0
MAP_DATUM                11
DSNTYPE                   0
MDCATY                    0
NEP_NM                  730
NPSPARK                1076
PANEL                     0
STATUS10                  0
STRATUM                   0
TNT                       0
WGT_CAT                   0
WGT_NCCA10                0
RSRC_CLASS                0
QA_CODES               1091
COMMENT                1091
dtype: int64
```

In [ ]:
```
ncca_raw.columns
```

Out[ ]:
```
Index(['UID', 'SITE_ID', 'STATE', 'DATE_COL', 'BATCH_ID', 'PARAMETER',
       'PARAMETER_NAME', 'RESULT', 'UNITS', 'MDL', 'MRL', 'PQL',
       'DATE_ANALYZED', 'HOLDING_TIME', 'QACODE', 'LAB_SAMPLE_ID', 'SAMPLE_I
D',
       'METHOD'],
      dtype='object')
```

In [5]:
```
ncca_raw.head()
```

Out[5]:

	UID	SITE_ID	STATE	DATE_COL	BATCH_ID	PARAMETER	PARAMETER_NAME	RE
0	59	NCCA10-1111	CA	7/1/2010	100714.1	NTL	Total Nitrogen	0.40
1	59	NCCA10-1111	CA	7/1/2010	100708.1	NO3NO2	Nitrate/Nitrite	0.01
2	59	NCCA10-1111	CA	7/1/2010	100708.1	SRP	Dissolved Inorganic Phosphate	0.02
3	59	NCCA10-1111	CA	7/1/2010	IM_CALCULATED	DIN	Dissolved Inorganic Nitrogen	0.01
4	59	NCCA10-1111	CA	7/1/2010	100714.1	PTL	Total Phosphorus	0.06

In [7]:
```
ncca_sites.head()
```

Out[7]:

	UID	SITE_ID	STATE	VISIT_NO	DATE_COL	WTBDY_NM	SITESAMP	INDEX_VISIT	EPA_R
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 [10]:
```
ncca_raw1 = ncca_raw.rename(
    columns = {'DATE_COL':'Date collected',
               'PARAMETER': 'Nutrient formula',
               'PARAMETER_NAME':'Nutrient',
               'RESULT':'Nutrient Amount',
               'HOLDING_TIME': 'Days held',
               'DATE_ANALYZED':'Date Anlyzed'}).drop(
    columns = ['METHOD','LAB_SAMPLE_ID','SAMPLE_ID','BATCH_ID','MDL','MRL','
```

In [ ]:
```
ncca_raw1
```

Out[ ]:

	UID	SITE_ID	STATE	Date collected	Nutrient formula	Nutrient	Nutrient Amount	UNITS	Dat Anlyze
0	59	NCCA10-1111	CA	7/1/2010	NTL	Total Nitrogen	0.407500	mg N/L	7/14/201
1	59	NCCA10-1111	CA	7/1/2010	NO3NO2	Nitrate/Nitrite	0.014000	mg N/L	7/8/201
2	59	NCCA10-1111	CA	7/1/2010	SRP	Dissolved Inorganic Phosphate	0.028000	mg P/L	7/8/201
3	59	NCCA10-1111	CA	7/1/2010	DIN	Dissolved Inorganic Nitrogen	0.014000	mg N/L	Nal
4	59	NCCA10-1111	CA	7/1/2010	PTL	Total Phosphorus	0.061254	mg P/L	7/14/201
...	...	...	...	...	...	...	...	...	.
7871	16731	NCCA10-1108	CA	6/29/2010	NTL	Total Nitrogen	0.228750	mg N/L	7/7/201
7872	16731	NCCA10-1108	CA	6/29/2010	PTL	Total Phosphorus	0.041821	mg P/L	7/7/201
7873	16731	NCCA10-1108	CA	6/29/2010	SRP	Dissolved Inorganic Phosphate	0.033000	mg P/L	7/2/201
7874	16731	NCCA10-1108	CA	6/29/2010	NH3	Ammonia	0.016000	mg N/L	7/1/201
7875	16731	NCCA10-1108	CA	6/29/2010	NO3NO2	Nitrate/Nitrite	0.012000	mg N/L	7/2/201

7876 rows × 11 columns

In [52]:

```
ncca_sites1 = ncca_sites.rename(
    columns={'WTBDY_NM':'Water collected',
             'SITESAMP':'Site sampled',
             'INDEX_VISIT':'Visit used',
             'WGT_NCCA10':'Adjusted site weight in square miles',
             'NCA_REGION':'Region',
             'STATION_DEPTH':'Water depth in Meters'
            }).drop(columns=['QA_CODES','COMMENT','VISIT_NO','EPA_REG','COUN
                             'ALON_DD','DSNTYPE','MDCATY','NEP_NM','NPSPARK'
                             'STRATUM', 'TNT','WGT_CAT','NCCR_REG','MAP_DATU
ncca_sites1
```

Out[52]:

	UID	SITE_ID	STATE	Water collected	Site sampled	Visit used	Region	PROVINCE	Water depth in Meters
0	59	NCCA10-1111	CA	Mission Bay	Y	Y	West Coast	Californian Province	2.5
1	60	NCCA10-1119	CA	San Diego Bay	Y	Y	West Coast	Californian Province	3.5
2	61	NCCA10-1123	CA	Mission Bay	Y	Y	West Coast	Californian Province	2.2
3	62	NCCA10-1127	CA	San Diego Bay	Y	Y	West Coast	Californian Province	9.5
4	63	NCCA10-1133	NC	White Oak River	Y	Y	East Coast	Carolinian Province	1.0
...	...	...	...	...	...	...	...	...	...
1099	2010099	NCCAGL10-GLBA10-174	MI	Lake Michigan	N	Y	Great Lakes	Great Lakes Province	NaN
1100	2010110	NCCAGL10-GLBA10-183	MI	Lake Michigan	N	Y	Great Lakes	Great Lakes Province	NaN
1101	2010113	NCCA10-2326	LA	Fourleague Bay	N	Y	Gulf Coast	Louisianian Province	NaN
1102	2010135	NCCA10-2328	LA	Hackberry Lake	N	Y	Gulf Coast	Louisianian Province	NaN
1103	2010141	NCCAGL10-GLBA10-179	MI	Lake Michigan	N	Y	Great Lakes	Great Lakes Province	NaN

1104 rows × 10 columns

In [63]:

```
##Part 1 - merging data
data = pd.merge(ncca_raw1, ncca_sites1, how = 'right', on = ['UID','SITE_ID'
data
```

Out[63]:

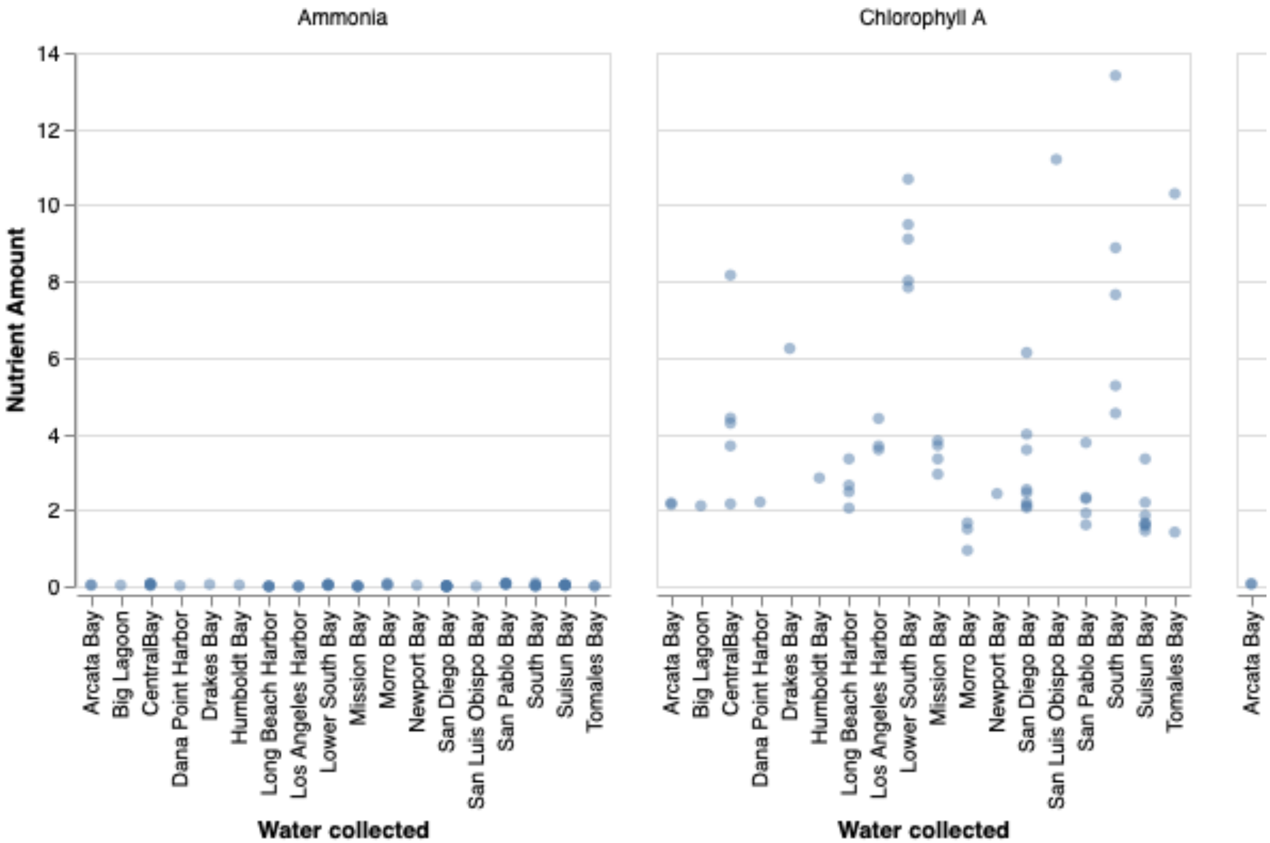
	UID	SITE_ID	STATE	Date collected	Nutrient formula	Nutrient	Nutrient Amount	UNITS	Dat Anlyze
0	59	NCCA10-1111	CA	7/1/2010	NTL	Total Nitrogen	0.407500	mg N/L	7/14/201
1	59	NCCA10-1111	CA	7/1/2010	NO3NO2	Nitrate/Nitrite	0.014000	mg N/L	7/8/201
2	59	NCCA10-1111	CA	7/1/2010	SRP	Dissolved Inorganic Phosphate	0.028000	mg P/L	7/8/201
3	59	NCCA10-1111	CA	7/1/2010	DIN	Dissolved Inorganic Nitrogen	0.014000	mg N/L	Nal
4	59	NCCA10-1111	CA	7/1/2010	PTL	Total Phosphorus	0.061254	mg P/L	7/14/201
...	...	...	...	...	...	...	...	...	.
7873	16731	NCCA10-1108	CA	6/29/2010	NTL	Total Nitrogen	0.228750	mg N/L	7/7/201
7874	16731	NCCA10-1108	CA	6/29/2010	PTL	Total Phosphorus	0.041821	mg P/L	7/7/201
7875	16731	NCCA10-1108	CA	6/29/2010	SRP	Dissolved Inorganic Phosphate	0.033000	mg P/L	7/2/201
7876	16731	NCCA10-1108	CA	6/29/2010	NH3	Ammonia	0.016000	mg N/L	7/1/201
7877	16731	NCCA10-1108	CA	6/29/2010	NO3NO2	Nitrate/Nitrite	0.012000	mg N/L	7/2/201

7876 rows × 18 columns

In [71]:

```
##Part 2; question 1
#relationship between  nutrient availability and productivity

alt.data_transformers.enable('default', max_rows=None)
# facet by nutrient to see each nutrient amount
alt.Chart(data).mark_circle(opacity = 0.5).encode(
    x = alt.X('STATE'),
    y = alt.Y('Nutrient Amount', scale = alt.Scale(zero = False)),).properti
    width = 250, height = 250
).facet(
    column = 'Nutrient'
)
#gives us the values for each nutrient in each state.

##Part 2; question 2
#any notable differences in available nutrients among U.S. coastal regions
#we filter our 'Region' column by pulling out all the rows that contain the
#facet by nutrient to see each nutrient amount
coast = data[data['Region'].str.contains("Coast")]

alt.Chart(coast).mark_circle(opacity = 0.5).encode(
    x = alt.X('Region'),
    y = alt.Y('Nutrient Amount', scale = alt.Scale(zero = False)),).properti
    width = 250, height = 250
).facet(
    column = 'Nutrient'
)

##Part 2; question 3
#sort into a cali datset by pulling all rows with state CA. We plot by Water
cali = data[data.STATE == 'CA'].dropna(thresh=10)
cali.sort_values(by = 'Water collected')
cali

alt.Chart(cali).mark_circle(opacity = 0.5).encode(
    x = alt.X('Water collected'),
    y = alt.Y('Nutrient Amount', scale = alt.Scale(zero = False)),).properti
    width = 250, height = 250
).facet(
    column = 'Nutrient'
)
```

Out[71]:



Ammonia
Chlorophyll A
Nutrient Amount
14
12
10
8
6
4
2
0
Arcata Bay
Big Lagoon
CentralBay
Dana Point Harbor
Drakes Bay
Humboldt Bay
Long Beach Harbor
Los Angeles Harbor
Lower South Bay
Mission Bay
Morro Bay
Newport Bay
San Diego Bay
San Luis Obispo Bay
San Pablo Bay
South Bay
Suisun Bay
Tomales Bay
Water collected