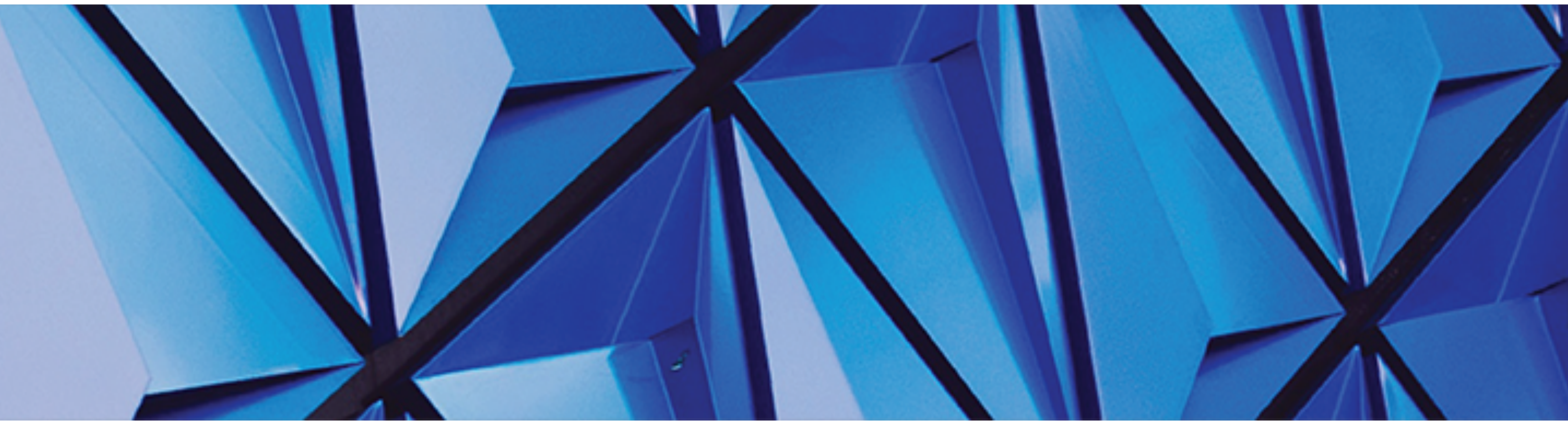


FIT5147 - Data Exploration and Visualisation S1 2019

Masters in Data Science

Assignment Report

Analysis on Coral Bleaching

Roopak Tj
Student ID : 29567467
Tutor: Joy Zhao

17 March 2019

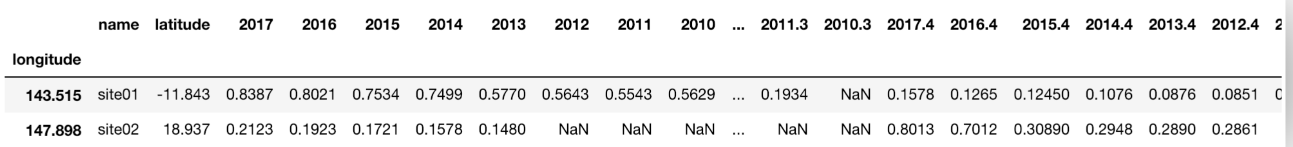
Programming Exercise : Tableau Public

Overview: Due to human impact, Coral Reefs are depleting rapidly. Damaging activities include coral mining, pollution (organic and non-organic), overfishing, blast fishing, the digging of canals and access into islands and bays. Data is about Coral bleaching of different coral types in 8 different sites spanned over 7 years timeframe (2010-2017).

Data Format: Excel sheet (xlsx file format)

Data exploration: Initial data understanding done in Python using pandas. Some of the observations:

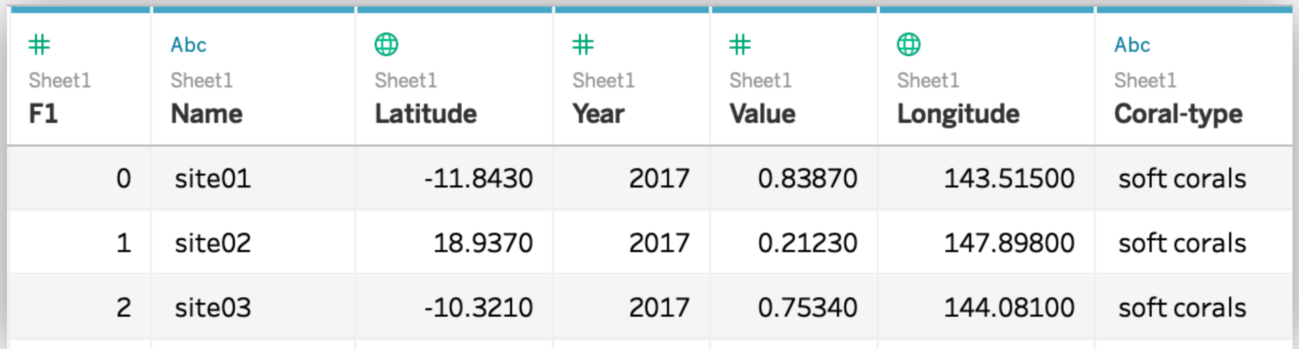
- 8 rows and 42 columns
- Some of the years have partial missing values where in some site in some type of corals, data was completely missing for the timeframe.
- Column names are not proper - names like 2011.3, 2010.3 (After importing into pandas)
- Column names in different hierarchy (longitude in different level as seen below)



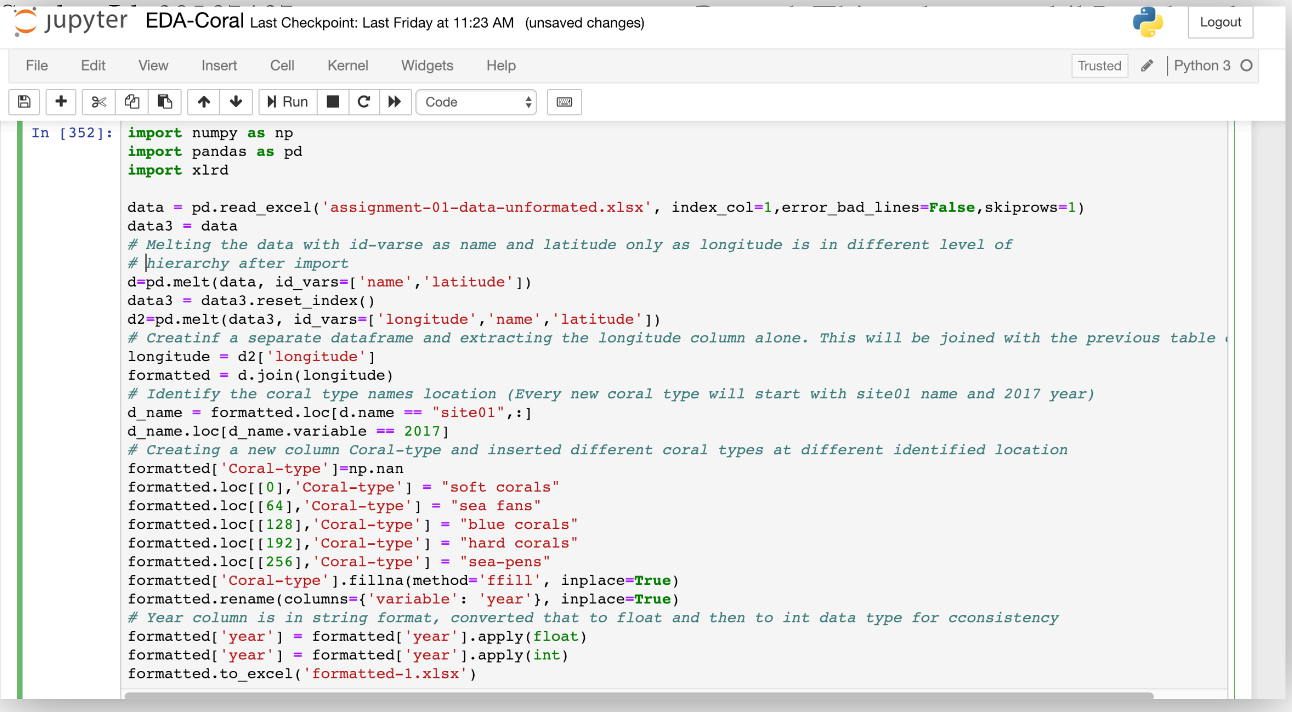
	name	latitude	2017	2016	2015	2014	2013	2012	2011	2010	...	2011.3	2010.3	2017.4	2016.4	2015.4	2014.4	2013.4	2012.4	2
longitude																				
143.515	site01	-11.843	0.8387	0.8021	0.7534	0.7499	0.5770	0.5643	0.5543	0.5629	...	0.1934	NaN	0.1578	0.1265	0.12450	0.1076	0.0876	0.0851	0
147.898	site02	18.937	0.2123	0.1923	0.1721	0.1578	0.1480	NaN	NaN	NaN	...	NaN	NaN	0.8013	0.7012	0.30890	0.2948	0.2890	0.2861	

Data format correction using Pandas (Python):

- Used melt data frame function to create new columns year and value, keeping longitude, latitude and name as id-varse.
- In this data frame, coral type information was missing. Created a new empty column 'Coral Types' and found the location at which there was a type change. Inserted Coral type values in that location and padded the column to fill NaN values with the previous filled value till the end.
- Year column was having datatype string with values like 2017.4 etc. Converted the datatype of the column to float and then to int to retain consistency of value throughout the column.
- Formatted file : Rows - 320, Columns - 6



#	Abc	🌐	#	#	🌐	Abc
Sheet1	Sheet1	Sheet1	Sheet1	Sheet1	Sheet1	Sheet1
F1	**Name**	**Latitude**	**Year**	**Value**	**Longitude**	**Coral-type**
0	site01	-11.8430	2017	0.83870	143.51500	soft corals
1	site02	18.9370	2017	0.21230	147.89800	soft corals
2	site03	-10.3210	2017	0.75340	144.08100	soft corals



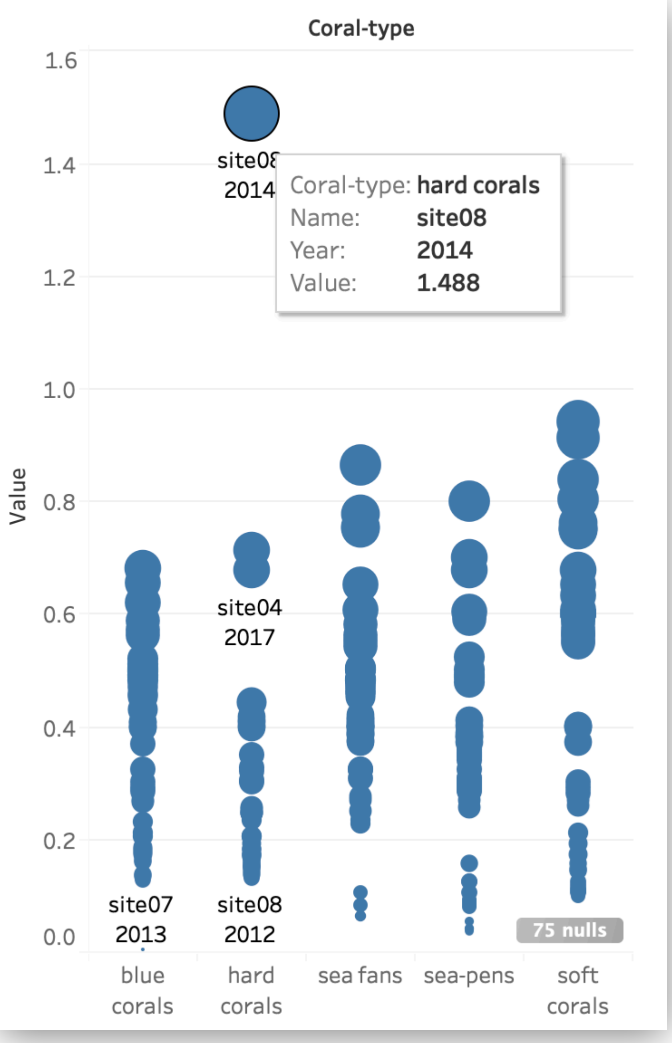
```
import numpy as np
import pandas as pd
import xlrd

data = pd.read_excel('assignment-01-data-unformated.xlsx', index_col=1,error_bad_lines=False,skiprows=1)
data3 = data
# Melting the data with id-varse as name and latitude only as longitude is in different level of
# hierarchy after import
d=pd.melt(data, id_vars=['name','latitude'])
data3 = data3.reset_index()
d2=pd.melt(data3, id_vars=['longitude','name','latitude'])
# Creatinf a separate dataframe and extracting the longitude column alone. This will be joined with the previous table
longitude = d2['longitude']
formatted = d.join(longitude)
# Identify the coral type names location (Every new coral type will start with site01 name and 2017 year)
d_name = formatted.loc[d.name == "site01",:]
d_name.loc[d_name.variable == 2017]
# Creating a new column Coral-type and inserted different coral types at different identified location
formatted['Coral-type']=np.nan
formatted.loc[[0],'Coral-type'] = "soft corals"
formatted.loc[[64],'Coral-type'] = "sea fans"
formatted.loc[[128],'Coral-type'] = "blue corals"
formatted.loc[[192],'Coral-type'] = "hard corals"
formatted.loc[[256],'Coral-type'] = "sea-pens"
formatted['Coral-type'].fillna(method='ffill', inplace=True)
formatted.rename(columns={'variable': 'year'}, inplace=True)
# Year column is in string format, converted that to float and then to int data type for cconsistency
formatted['year'] = formatted['year'].apply(float)
formatted['year'] = formatted['year'].apply(int)
formatted.to_excel('formatted-1.xlsx')
```

Outlier-Identification:

Imported the .xlsx file to Tableau public for exploration and Outlier Identification. Outliers can be find in different the below visualisations.

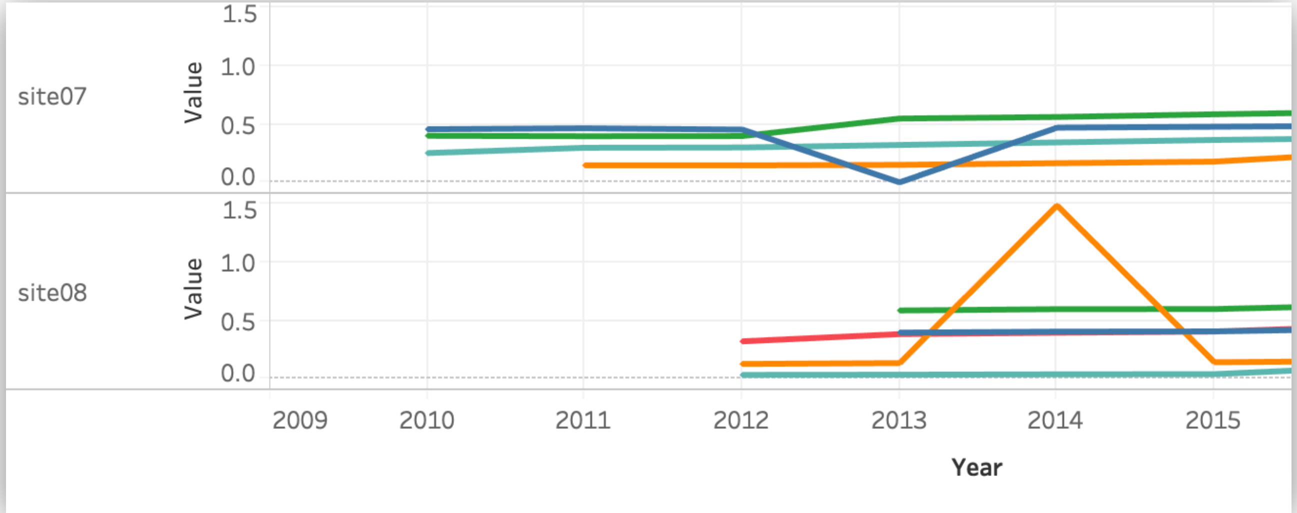
Bubble chart with Coral type in Column and Value(Percentage) in rows with Site name and Year in Text(Mark) and Value in bubble size.

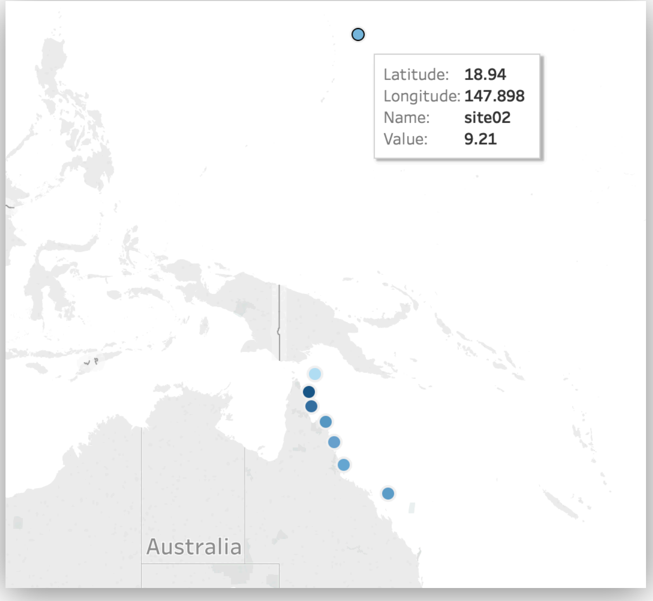
Observation: Coral bleach at site08 in 2014 has percentage of 148 which is unusual considering the previous years percentage for the same coral type in different years.



Same outlier can be verified using continuous line chart(Year in column and Name/Value in Rows with coral-type indicating colour). You can see that there is a sudden rise in the bleach percentage for site08 in 2014 which goes beyond 100% and then goes back to the same range of values.

An anomaly can also be identified for site07 in 2013 where the coral type goes to very low percentage(almost zero) comparing to its previous and next value.





Latitude: 18.94
Longitude: 147.898
Name: site02
Value: 9.21

World map with Latitude in columns and Longitude in rows.

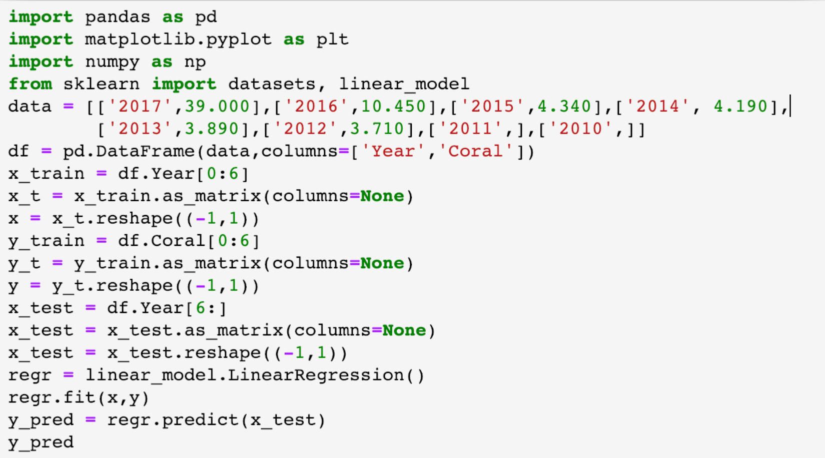
Another Outlier can be identified with the value of latitude for site02. Comparing to other Coral sites, site02 locates itself in the middle of the ocean. On checking the latitude value of other sites, it is clear that negative sign is missing from the latitude value of site02 which can be an entry error.

Outlier Fix:

Latitude outlier fix: For site02, latitude value is positive and on comparison with latitude values of other sites, it should be negative. So manually changing the latitude value os site02 from 18.94 to -18.94.

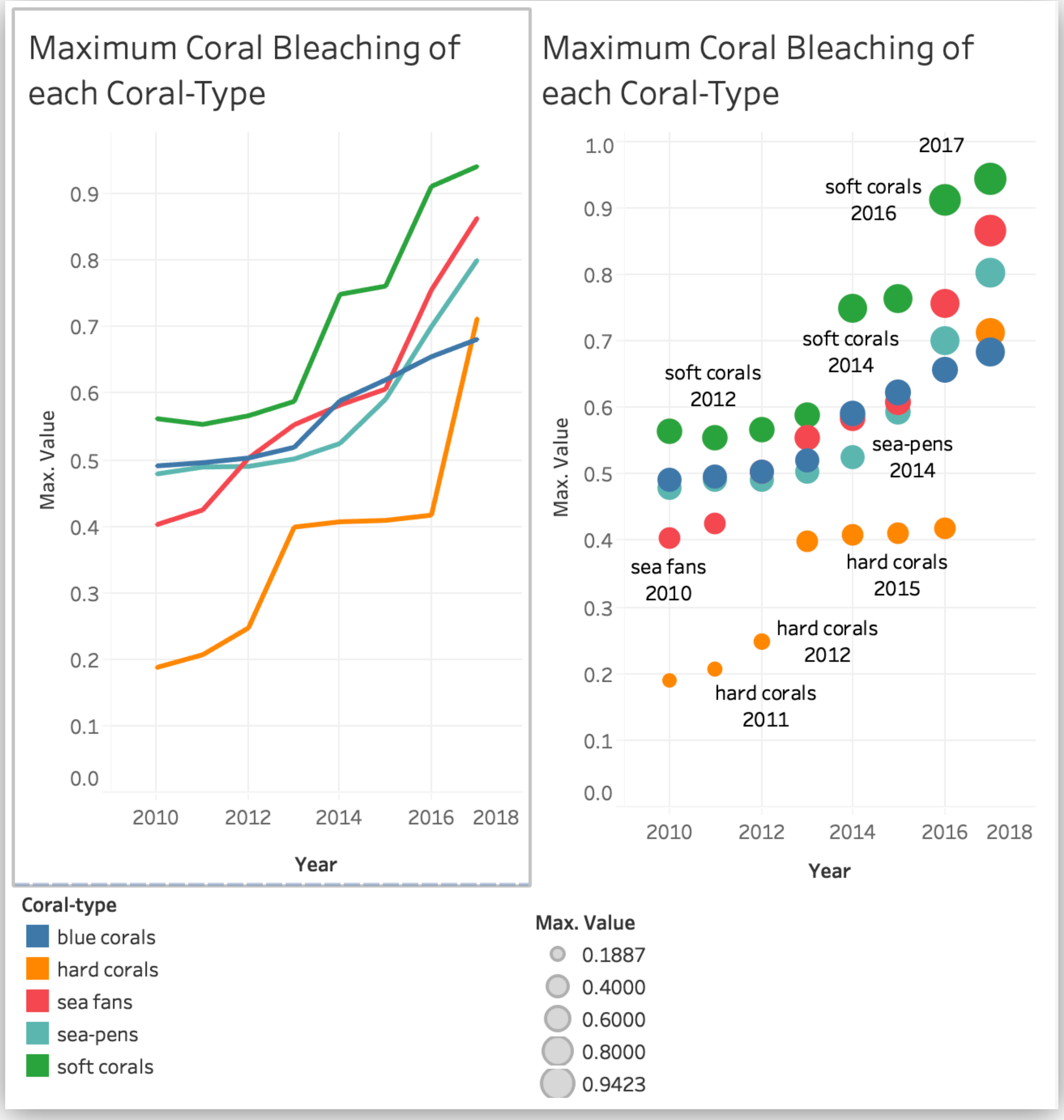
Percentage fix: Percentage for site07 and site08 has manual entry errors. It is corrected using comparing the coral percentage of same type in previous years

Missing Value fix : In some rows, there are partially missing data and in others complete data is missing. Partial missing data has been fixed using estimating the value using regression technique and imputing the value to the sheet. Please find the code in the image.



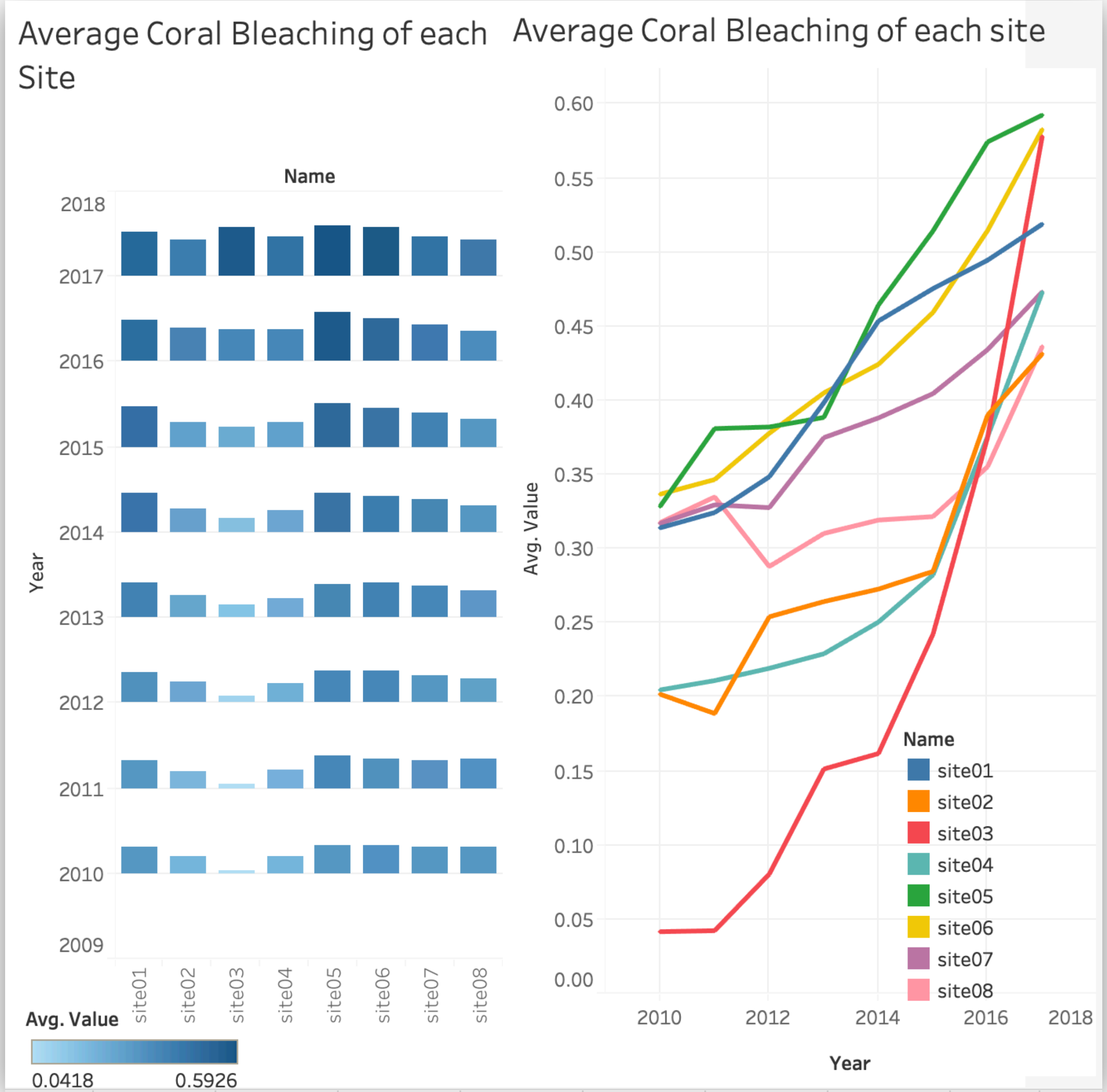
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
from sklearn import datasets, linear_model
data = [['2017',39.000],['2016',10.450],['2015',4.340],['2014', 4.190],
        ['2013',3.890],['2012',3.710],['2011',],['2010',]]
df = pd.DataFrame(data,columns=['Year','Coral'])
x_train = df.Year[0:6]
x_t = x_train.as_matrix(columns=None)
x = x_t.reshape((-1,1))
y_train = df.Coral[0:6]
y_t = y_train.as_matrix(columns=None)
y = y_t.reshape((-1,1))
x_test = df.Year[6:]
x_test = x_test.as_matrix(columns=None)
x_test = x_test.reshape((-1,1))
regr = linear_model.LinearRegression()
regr.fit(x,y)
y_pred = regr.predict(x_test)
y_pred
```

Q1) In which years and for which kinds of coral bleaching is the worst



The information can be represented using a continuous line chart and a bubble chart(for better understanding)
Representation clearly compares the maximum Coral Bleaching for each type in each year. As seen in the graph, for all the years soft corals has seen the maximum Coral bleaching

Q2) How the location of the site affects bleaching on the different kinds of coral.



The visualisation represents how bleaching (average) happens on each site. The above two types of visualisation can be used to represent the same.

On site01 and site05 Coral bleaching is high which also increase in a steady rate. Highest rate of increase of Coral Bleaching is observed in Site03 which is almost exponential growth.

Site08 observed a small decrease in bleaching but it gradually increased from 2012.