Exam 1

CS 4731

Michael Montgomery

September 23, 2019

1 Problem definition

Water salinity and temperature is important for a number of things. One of which is its impact of on water density. Ocean temperatures and salinity levels can affect how ocean currents form and move, and it can also affect marine life. One example is the Dead Sea, located between Israel and Jordan. The Dead Sea has extremely high salinity levels which makes harboring marine life extremely difficult if not borderline impossible. Other reasons for concern is due the rise of global temperatures the oceans could become inherently saltier. Warmer temperatures causes the ocean's water to evaporate leaving behind salt when the water rises into the atmosphere. This cycle causes the salinity levels of the ocean to rise, and this could become increasingly worrisome due to the fact that the salinity levels of the world's oceanic bodies affect living organisms. We have already observed this through the case of the Dead Sea and its inability to support life other than microorganisms.[2] [6]

This project's goals is to draw a correlation between water temperature and its salinity levels. So, in essence, how does water temperature affect its salinity? Can we draw a correlation between the two variables? What does this mean for rising global temperatures and its effect on the oceanic ecosystems, etc.

2 How linear regression can help model/predict

Linear regression can help model how the ocean's salinity is affected by its temperature. With salinity being the dependent variable we can potentially achieve a model that predicts how rising temperature could affect the ocean's salinity.

3 Data Set

The data set used for this project was collected from the California Cooperative Oceanic Fisheries Investigation (CalCOFI). CalCOFI was established in 1949, and its focus today is the study of the marine environment off the coast of California. Their collected data is used to study climate change, etc. They collect oceanic data down to the depth of 500 meters, and their data extends back to the 1940s. In their data set, named "bottle.csv", the measure a multitude of things, but this project focuses on their collection of water temperature, measured in degrees Celsius, and salinity, measured in grams per kilograms (g/kg). This data set includes approximately 850,000 data points, and this should be more than enough to draw some sort of conclusion between the two variables.[5]

Data Set: https://www.kaggle.com/sohier/calcofi

4 Proposed Solution

To find out the relationship between water temperature and its salinity this project uses various python libraries and statistical methods to evaluate the line of best fit. The data file will be read in using Pandas.[4] Pandas is a good library for reading and analyzing files. Numpy[1] is used for this project's mathematical operations, and Matplotlib[3] is used for this project's data visualization.

5 Solution

Reading in the CSV file

```
# read in csv file
data = pd.read_csv('bottle.csv',low_memory=False)
```

Extracting data and casting it from a Pandas Series to a Numpy Array

```
# slice the two rows
temp = data.iloc[:,5]
salt = data.iloc[:,6]

# cast Panda series to Numpy arrays
temp = temp.to_numpy()
salt = salt.to_numpy()
```

Calculate line of best fit - Least Squares

```
#     https://stackoverflow.com/questions/13643363/linear-regression-of-arrays-conta
#     this is used to filter out all the non-values
mask = ~np.isnan(temp) & ~np.isnan(salt)

#     https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.linreg
slope, intercept, r_value, p_value, std_err = stats.linregress(temp[mask],salt[mask]
```

Function to evaluate y = b1 + b0x

```
f = lambda b0, b1, x: b0 + b1*x


domain = np.linspace(1,np.nanmax(temp),1000)

values = f(intercept,slope,domain)
```

Data Visualization

```
data.plot(x='T_degC', y='Salnty', style='rx' )

plt.title('Temperature (C) vs Salinity (g/kg)')

plt.xlabel('Temperature (C)')

plt.ylabel('Salinity')

plt.savefig('Graph_of_total_points.pdf', bbox_inches='tight'

# see Appendix for graph



ax.plot(temp,salt,'rx')

ax.plot(domain, values,'g-')

ax.set_title('Data with regression line')

ax.set_xlabel('Temperature')

ax.set_ylabel('Salinity')

plt.savefig('regression.png')

# see second graph in Appendix


plt.hist(temp)
plt.title('Distrobution of Temperature variables')
plt.savefig('Hist_temp.png', bbox_inches='tight')
#see third graph in Appendix


plt.hist(salt)
plt.title('Distrobution of Salinity variables')
plt.savefig('Hist_Salinity.png', bbox_inches='tight')
# see fourth graph in Appendix
```

6 Evaluation

The data analysis yield the line $y = 34.4409 - 0.0552x$. This means that the temperature of the water negatively affects the salinity of the water according to this data set. Higher water temperatures yield lower water salinity levels. The distribution of the data across the temperature variables peak around ten degrees Celsius, and the salinity variables peak around thirty-four grams per kilograms.

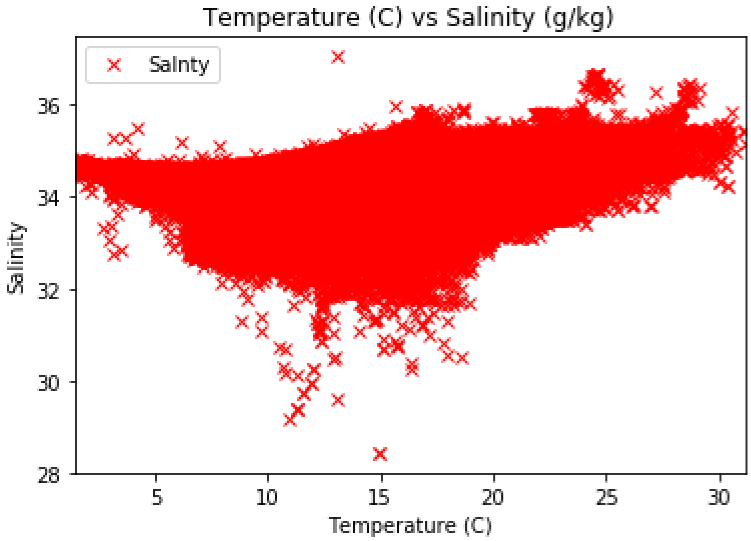
7 Discussion

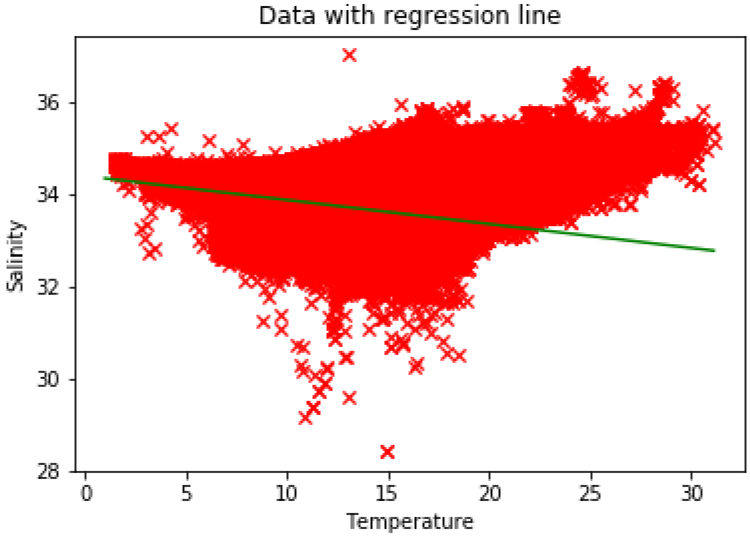
There could be several contributing factors behind the data set that would give a negative correlation between the temperature and salinity variables. One of which includes that saltier water is more dense than its less salty counterpart. Therefor the saltier water sinks to the bottom and cools down. So, despite the contributing factors of the global temperature rising inherently speeding up the evaporation process, this data set yielded an opposite correlation of the two values. However, due to rising temperatures the salinity levels in the ocean could begin to grow larger as a whole, and its effect on oceanic currents could cause problems to the oceanic ecosystem.

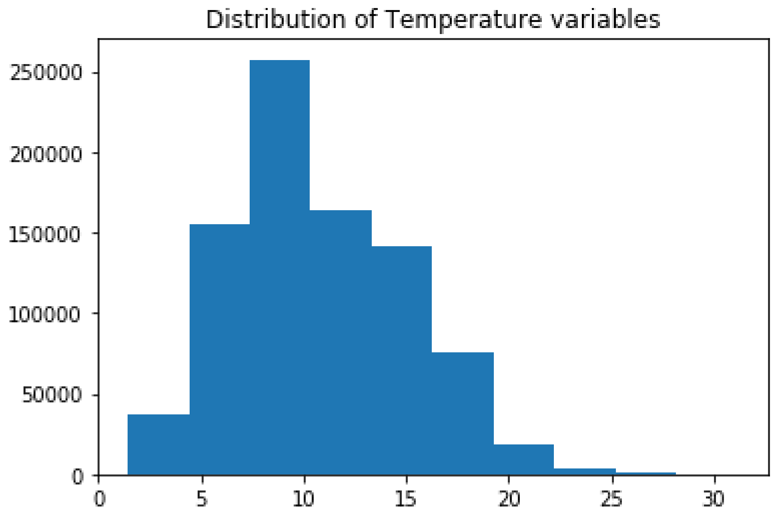
8 Future work

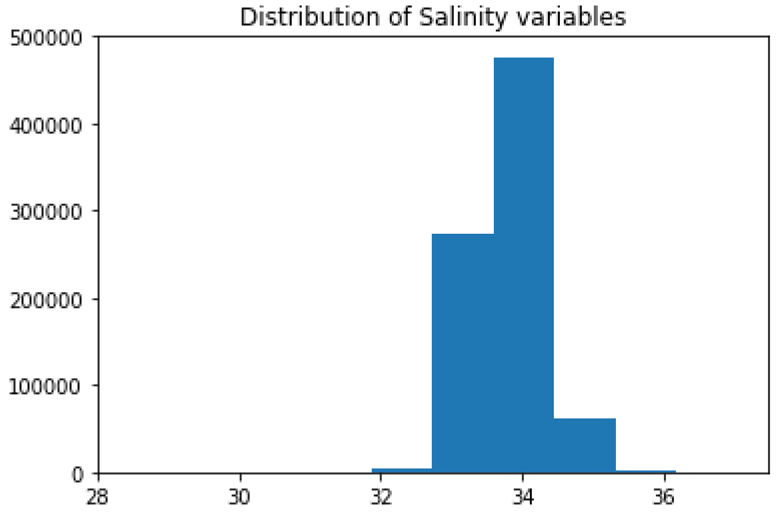
Future work on this project might include: creating more data visualizations, pulling in more data points for analysis, and cleaning up and commenting on the code to make it easier to read and understand.

A Appendix









References

[1] Numpy and scipy documentation¶.

[2] Ocean salinity.

[3] Overview¶.

[4] powerful python data analysis toolkit¶.

[5] Webmaster. Program overview.

[6] B Wilkansky. Life in the dead sea. *Nature*, 138(3489):467, 1936.