**Live Green | Live Happy**

**Problem and Goal:**

Looking to promote sustainable practices, Wells Fargo has taken the initiative to transition themselves and individuals into a low-carbon economy. As such, they have invited students with an interest in sustainable practices to utilize machine learning and data science to discover ways to minimize the carbon footprint of individuals. The difficulty in this is maintaining the quality of life for the individuals involved.

 Here at the College of Charleston, we have adopted Sustainability Literacy as a bridge to addressing 21st-century problems. Indeed, our Quality Enhancement Plan for the next 10 years is focused on this idea of sustainability in the environmental, economic, and social spheres. As such, this was the perfect opportunity to represent the college and Wells Fargo in finding a solution to the carbon-minimization problem.

**General Solution and Results:**

The primary difficulty in this dataset was working with missing values, whether they be in conversion rates, or data collected from the 1002 individuals. However, I was able to create several primary subsets of the data after a significant amount of feature engineering.

The first is the carbon dataframe, which shows the individual number, activity, quality of life importance, and carbon usage. This showed the overall data for every individual. However, the goal of the project was to identify solutions on an individual basis, so I developed the indRows set. This subsets the data to the individual, while also rating every activity based on a regression line.

One final note is that there were clear outliers in the solution sets. Car trips, garbage disposal, and air conditioning often skewed the regression line significantly. However, two major points stood out:

1. Composting greatly decreased the carbon footprint of an individual.
2. This project should be called: Stop Flying on Planes! The carbon emissions from air travel was just overwhelming.

**Mathematics and Equations:**

On Conversion Rates:

I first attempted to find actual values for these rates online, but they were often too general to be helpful. As such, I used the average of known ratios to calculate any null values in the second datasheet.

The equations often varied depending on what values were known and which needed to be calculated, so it was far more efficient for me to do the operations by excel or hand. As I did them in short periods of time, I opted for entering the equations in python by hand, but the process could certainly be automated if more ratios would be needed.

I’ve included some of the equations in this package, but the method I used should be sufficient in understanding my process.

On Calculating Carbon Use:

While a significant amount of this was coupled with feature engineering, the general principle for my calculations can be explained independently.

Every source of energy creation, such as solar or gas, was marked as used or unused. Then, the ratio between every energy source and activity was found based on the selected value. Unfortunately, the exact proportions weren’t included, so the chosen ratios were then averaged, and multiplied by the consumption of carbon (whether it be by activity, hour, etc.). These mathematics were completed by a looping mechanism in python.

**Feature Engineering:**

Feature Engineering was likely one of the most difficult parts of the process, due to the vast amount of missing values and table manipulations that were required:

Many of the conversion rates were not required, as there was no precedent for their use in the dataset. As such, they were left as NaN or 0.

In addition, any rows with NaN for every energy source were subsetted to create an index. This index was later used to subset the data and replace the NaN rows with binary digits. The binary selection was based off a precedent for activity and energy source.

When only a single conversion rate was given for an activity, all NaN were replaced with 1 for that conversion rate and multiplied by the consumption.

If consumption or quality of life values weren’t included, I left them out of the analysis. They could have been replaced by average values but would have only distracted from the insights gained from the collected data.

Many tables were revised and edited for easier use, comparison, merging, or subsetting, as noted throughout the entirety of the code.

Finally, several dataframes were composed for a variety of purposes. Some are prepared for individual analysis, while others can be used to identify clusters among an entire sample or population.

**Machine Learning Analysis:**

The primary machine learning algorithm that I used was a simple linear regression model on an individual basis.

With this dataset, there isn’t a clear minimization solution, as the equation would be dependent on two variables:

Min(ax + by)  
Where a and b are inversely related

As a result, instead of assigning arbitrary values for a, b, and a group of relation-based rates, I opted for a primarily informative method.

A regression line plots a persons’ carbon emissions vs quality of life per activity (excluding outliers), and the distance is calculated accordingly. A large positive value shows that the activity is largely positive, whereas a large negative value shows it has significant emissions and minimal effect on one’s quality of life. The regression line was made to be positive, as the ratio between quality and quantity was most important. However, it was interesting that some individuals found high carbon activities to be less enjoyable or important.

This was done at an individual level, as directed by the prompt from Wells Fargo. I also attempted a few clustering algorithms, but they either didn’t provide any significant insights or were less effective (as the data wasn’t binary, but rather gradual) than linear regression and were subsequently removed from the final code. I also attempted some clustering on the entire dataset, but it was also relatively inconclusive (no real insights or correlations). The other machine learning algorithms I experimented with were k-means, logistic regression, and SVMs.