Data compilation is a major issue within the ESG (environmental, Social, Governance) industry. With data presented by multiple different metrics across industries and companies, it is difficult to conclude a company's environmental impact. Despite this, with careful application, algorithms can synthesize and benchmark this data for us. In this project, I've compiled the resource consumption, resource waste prevention, and/or greenhouse gas prevention of two Russell 2000 companies: Green Plains Inc (GPRE) and SunRun (RUN). Seeing that Buena Ola has a large focus on the planet and profit in their triple-bottom-line. I decided to create a program that would create a linear regression model that can predict/ analyze the correlation between the companies' sustainable behavior to their ability to profit while maintaining sustainable behavior. This way we more clearly visualize whether these companies' business model and sustainability model can coexist

GPRE is an agri-tech company that specializes in creating low-carbon biofuel creation. The company makes it clear that one of the defining characteristics is its ability to maintain sustainable profits. SunRun on the other hand is a company that focuses on providing residential homes with access to solar energy by equipping them with solar panels and batteries to an on-grid system. One reason I chose these companies was to use their expected financial outcome, as it would make sense that a company that provides sustainable resources such as biofuel and solar energy would reduce its environmental impact while increasing earnings over time. In this case, I wanted to see if this would be the case for my model.

For my program, I decided to create a function that allows a user to input which company they would like to analyze(GPRE, RUN). From there the user is prompted by

the program to input which environmental factors (i.e., total greenhouse gasses emitted, greenhouse gases prevented, water usage, electricity usage, etc.) they would like to analyze with contrast to the companies earnings. The program makes use of the environmental data provided by these companies as well as the Yahoo Finance library (yfinance) to parse the data into pandas dataframes. The program then converts the data frames into NumPy arrays which get put into a regression algorithm to calculate the linear equation which would model future sustainability to earnings correlation. The program then returns this information in the form of a graph that plots the data points as well as the regression line (using Matplotlib and sklearn libraries) for the user to see.

One drawback to this approach is that both GPRE and RUN only provided approximately 3-5 points of data per environmental factor, which means that the program is not very accurate when predicting the future. However, given the opportunity to access more data the program would increase in accuracy assuming the relations are linear (in that case we would need to use a different regression model). Another drawback of the program is that the graph which compares all the environmental factors is hard to visualize (this is partially due to the poor sample size of data) because of the scale of the data relative to each other, as well-being measured in different units.

With more time and resources (Data) I would have given each individual data comparison pair their own regression model to best predict the outcomes. In addition, I would have implemented better user interface features that would make navigating and testing the program easier.