**STAT 412 – Interim Report**  
**Author**: Sanada Naco  
**Date**: May 10, 2025  
**Project Title**: ESG and Financial Indicators Analysis

**1. Aim of the Project**

The aim of this project is to investigate the relationship between companies’ ESG (Environmental, Social, and Governance) performance and their financial and environmental indicators over the years 2015–2025. It also evaluates how ESG scores vary by industry and region, and explores the predictive power of financial variables on ESG scores.

**2. Data Source and Variables**

The dataset used in this project was obtained from Kaggle: [ESG and Financial Performance Dataset](https://www.kaggle.com/datasets/shriyashjagtap/esg-and-financial-performance-dataset/data). It contains company-level ESG and financial performance indicators from 2015 to 2025, totaling 11,000 observations. The key variables include:

* **Dependent Variable**: ESG\_Overall – the overall ESG score of each company.
* **Independent Variables**:
  + Revenue (log-transformed)
  + MarketCap (log-transformed)
  + CarbonEmissions (log-transformed)
  + ProfitMargin
  + GrowthRate
  + Industry
  + Region

**3. Data Cleaning and Tidying & EDA**

* Checked the structure of the dataset and variable types
* Converted CompanyID to character and Year to integer
* Summarized the data to understand distributions and possible data entry issues

The dataset consists of 11,000 observations across 16 variables, capturing company-level information such as industry, region, year, financial performance, environmental impact, and ESG scores from 2015 to 2025. Categorical fields like Industry and Region describe each company's classification, while numeric variables represent financial indicators and ESG metrics. A summary of the data reveals substantial variation in company size and performance. For instance, Revenue ranges from as low as 36 to over 180,000, with a mean of 4,691 but a much lower median of 1,890, indicating the presence of a few very large firms that skew the average. Similar patterns are seen in MarketCap, CarbonEmissions, and EnergyConsumption, all of which show extreme upper values and strong right-skewness, justifying the need for log transformation during modeling. ESG-related variables (ESG\_Overall, ESG\_Environmental, ESG\_Social, ESG\_Governance) are on a standardized 0–100 scale and appear relatively balanced, with medians around 50–55. However, nearly all key variables contain missing values—typically around 1,000 observations per variable—highlighting the importance of careful imputation.

A screenshot of a computer screen

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect.

* **Histograms**

Several financial and environmental variables such as Revenue, MarketCap, CarbonEmissions, and EnergyConsumptionwere highly right-skewed, with a few companies showing extremely large values. To correct this and make the data more suitable for linear modeling, we applied log transformations. The resulting distributions were much closer to normal, improving model assumptions like linearity and constant variance. ESG scores and variables like ProfitMargin and GrowthRate already had roughly symmetric distributions and did not require transformation. These steps ensured the dataset was well-prepared for regression analysis.

A graph of different types of data

AI-generated content may be incorrect.A group of green and white graphs

AI-generated content may be incorrect.

**Correlation Plots**

The correlation analysis showed that ESG\_Overall has moderate positive relationships with its component scores—Environmental, Social, and Governance—which confirms internal consistency. However, its correlation with financial variables like log\_Revenue and ProfitMargin was positive but weak, suggesting that larger or more profitable companies tend to score slightly higher on ESG, but the relationship is limited. Environmental impact variables such as CarbonEmissions, WaterUsage, and EnergyConsumption were highly correlated with each other, indicating that they reflect similar underlying company behaviors. These high inter-correlations justify the decision to include only one of them (e.g., log\_CarbonEmissions) in the final regression model to avoid multicollinearity. Overall, the results align with expectations and support the variable selections made during model building.

A screenshot of a graph

AI-generated content may be incorrect.A screenshot of a graph

AI-generated content may be incorrect.

* **Correlation analysis**

The correlation matrix highlights strong positive relationships among size-related variables such as Revenue, MarketCap, and their log-transformed counterparts, with coefficients above 0.80–0.90. Environmental impact measures like CarbonEmissions, EnergyConsumption, and WaterUsage also show high correlations, reflecting their common underlying drivers. ESG\_Overall is moderately correlated with its components—Environmental, Social, and Governance—as expected. However, its correlation with financial variables like Profit and Revenue remains low to moderate (around 0.25–0.30), indicating that while financial size may play a role, it is not a dominant factor in ESG performance.

A graph with numbers and a bar chart

AI-generated content may be incorrect.

**Research Questions:**

How has ESG performance changed over time (2015–2025)?

To examine how ESG performance has evolved, we plotted the **mean ESG\_Overall score** for each year between 2015 and 2025. The line chart reveals a **clear upward trend**: the average ESG score increased from approximately **51.5 in 2015** to about **58.2 in 2025**. Although there are minor fluctuations in a few years—particularly a slight dip around 2020—the overall pattern is consistently positive. This suggests that companies have, on average, made steady improvements in their environmental, social, and governance practices over the past decade. The rise may reflect growing global pressure on corporate sustainability and increased ESG reporting standards.

A graph with orange lines and black dots

AI-generated content may be incorrect.

Is there a relationship between profitability and ESG performance? & Do companies with higher revenue or market capitalization tend to have better ESG scores?

To explore how financial performance relates to ESG scores, we plotted ESG\_Overall against two key variables: log\_Revenue and ProfitMargin. Both scatterplots show a large spread of points with a slight upward trend, as captured by the fitted regression lines. This indicates that companies with higher revenue and better profit margins tend to have slightly higher ESG scores. However, the relationship is weak and not tightly clustered around the trend line, suggesting that financial performance alone does not strongly predict ESG behavior. These visual insights are consistent with the earlier correlation analysis and regression results, where the coefficients were significant but small in magnitude.

A green line in the middle of a graph

AI-generated content may be incorrect.A blue line in the middle of a graph

AI-generated content may be incorrect.

Are there significant differences in ESG\_Overall scores across different industries?

The boxplot shows the distribution of ESG\_Overall scores across different industries. While most industries have median ESG scores clustered between 50 and 70, there are clear differences in central tendency and spread. For example, the **Energy** and **Finance** sectors show relatively high medians and tighter interquartile ranges, suggesting more consistent ESG performance. In contrast, **Transportation** displays a wider spread with a lower median, indicating more variability and generally lower ESG scores within the sector. Outliers are present in nearly all industries, reflecting exceptional performers or underperformers.

A graph showing a number of blue squares

AI-generated content may be incorrect.

**4. Exploration of Missing Data Mechanism**

To explore whether missing values in ESG\_Overall occur randomly, we plotted the distribution of financial variables (ProfitMargin, GrowthRate, MarketCap, and Revenue) split by missingness of ESG\_Overall. In each density plot, the distribution for records with missing ESG scores (in blue) closely overlaps with the distribution for non-missing records (in red). This suggests that the probability of ESG\_Overall being missing does not depend on the values of these variables. Since the shapes, centers, and spreads of the distributions are nearly identical, this provides **visual evidence in support of the MCAR (Missing Completely At Random)** assumption for ESG\_Overall. These findings support the use of unbiased imputation methods such as multiple imputation without requiring further adjustment for non-random missingness.

A graph with a line

AI-generated content may be incorrect. A graph of a normal distribution

AI-generated content may be incorrect.A graph with numbers and lines

AI-generated content may be incorrect.A graph with lines and numbers

AI-generated content may be incorrect.

Since earlier visual tests suggested that missingness was likely Missing Completely At Random (MCAR), we chose the mice (Multiple Imputation by Chained Equations) method. This technique allowed us to impute missing values based on predictive models that respect the multivariate structure of the data, ensuring a statistically sound foundation for later analysis.To evaluate the quality of the imputed values for missing ESG-related data, we examined both strip plots and density plots after multiple imputation. The strip plots show that the imputed values (in red) are evenly spread across the full range of each variable and blend well with the observed values (in blue), indicating that the imputation process preserved the natural distribution and variability. The density plot comparing the distribution of imputed ESG\_Overall values with the original missing cases shows a smooth, bell-shaped curve, confirming that the imputed values align with the expected range and pattern. Together, these visual checks suggest that the imputation process was successful, and the generated values are plausible, unbiased, and appropriate for downstream modeling.

A screenshot of a graph

AI-generated content may be incorrect.A graph with a line drawn on it

AI-generated content may be incorrect.

**5. Feature Engineering**

The following derived variables were created:

* **Profit**: Calculated using revenue and profit margin
* **ESG Spread**: Difference between governance and environmental scores
* **Emissions per Revenue** and **Emissions per Profit**: Adjusted carbon emission metrics
* **ESG High**: Binary variable based on whether a company's ESG score is above the median

**6. Confirmatory Data Analysis**

The regression analysis confirms that ESG\_Overall scores have significantly increased over the 2015–2025 period, with an average rise of about 0.67 points per year. This result is statistically highly significant, as shown by the very low p-value and large t-statistic. However, the model’s **R² value of only 1.8%** indicates that the year alone explains just a small fraction of the variability in ESG scores. In other words, while there is a clear trend of ESG improvement over time, **most of the variation across companies is driven by other factors**. These likely include industry-specific norms, geographic regulatory environments, and firm-level financial or operational characteristics.

A screenshot of a computer

AI-generated content may be incorrect.

These regression models support the earlier scatterplot findings, showing **weak but statistically significant positive relationships** between ESG scores and financial performance. In the model predicting revenue, a one-point increase in ESG\_Overall was associated with a 98.29-unit increase in revenue, yet the model's **R-squared value is only 0.0229**, meaning it explains just **2.29%** of the variation in revenue. Similarly, in the model predicting profit margin, a one-point ESG increase corresponded to a 0.059 percentage point rise in profit margin, with an **R-squared of 0.0115** (just **1.15%** of variance explained). These results suggest that while ESG scores are positively associated with financial metrics, their **predictive power is limited**. ESG is likely one of many contributing factors, and more complex, multivariable models are needed to capture the full picture of financial performance.

A screenshot of a computer program

AI-generated content may be incorrect.A computer screen shot of a number

AI-generated content may be incorrect.

The ANOVA results provide strong statistical evidence that ESG scores vary significantly across industries. With an F-value of 138.2 and a p-value less than 2e–16, we reject the null hypothesis that all industries have the same mean ESG\_Overall score. This supports what was visually observed in the boxplot: certain industries—such as Energy and Finance—tend to have higher ESG performance, while others, like Transportation, show lower average scores. These findings highlight the importance of including industry as a categorical predictor in the final regression model, as it clearly plays a significant role in explaining ESG variability.

A black background with white text

AI-generated content may be incorrect.

**8. Model Building and Evaluation**

* **Final Model**:  
  ESG\_Overall ~ log\_Revenue + log\_CarbonEmissions + Industry + Region
* **Assumptions tested**:
  + Residuals: Q-Q plots
  + Homoscedasticity: Breusch-Pagan test (bptest)
* **Train-Test Split** (80/20 using caret::createDataPartition)
* **Model Evaluation**:
  + Used postResample to evaluate RMSE and R² on test data
  + Compared predicted vs actual values with scatter plot

The final linear regression model predicting ESG\_Overall used four key predictors: log\_Revenue, log\_CarbonEmissions, Industry, and Region. All variables were statistically significant (*p* < 2e–16), and the model fit was strong, with an **adjusted R² of 0.6287** on the training set.

* **Financial Size**: The coefficient for log\_Revenue was **+40.89**, meaning a one-unit increase in log-revenue (roughly a 2.7× increase in actual revenue) was associated with a **40-point increase in ESG score**, all else equal. This confirms that larger firms are typically rated more positively in ESG metrics.
* **Environmental Impact**: The coefficient for log\_CarbonEmissions was **–40.49**, showing that companies with higher emissions tend to receive lower ESG scores — reinforcing the ESG framework's environmental accountability.
* **Industry Effects**: Compared to the reference industry, sectors like **Energy (+63.12)**, **Transportation (+43.02)**, and **Utilities (+55.20)** were associated with higher ESG scores. In contrast, **Finance (–98.29)**, **Healthcare (–31.64)**, and **Retail (–28.64)** had significantly lower ESG scores, aligning with the earlier boxplot and ANOVA results.
* **Regional Differences**: Firms headquartered in **Europe (+22.38)** and **North America (+17.21)** scored substantially higher than those in the reference region, while companies in the **Middle East (–2.38)** showed slightly lower ESG scores on average.

A screenshot of a computer

AI-generated content may be incorrect.

The small drop in R² (from 62.9% to 59.9%) and RMSE indicates **strong generalization**. The predicted vs. actual plot shows points closely following the 45-degree reference line, confirming that the model makes accurate and stable predictions.

> train\_metrics

RMSE Rsquared MAE

9.6522426 0.6294768 7.8063834

> test\_metrics

RMSE Rsquared MAE

9.8327625 0.5993319 7.9280077

A graph with blue dots

AI-generated content may be incorrect.

**9. Conclusion**

This project set out to understand how company-level ESG (Environmental, Social, and Governance) performance relates to financial and environmental indicators across a diverse global dataset covering the years 2015 to 2025. Through a structured approach that combined data cleaning, exploratory analysis, statistical testing, and predictive modeling, we arrived at several important insights.

We first observed a steady increase in average ESG scores over time, suggesting growing emphasis on sustainable practices across industries. Initial scatterplots and correlation tests revealed weak but statistically significant relationships between ESG scores and financial indicators like revenue and profit margin. However, it became clear that financial metrics alone could not explain most of the variation in ESG performance.

More meaningful differences emerged when we examined ESG scores across industries and regions. ANOVA and post-hoc tests confirmed that these structural factors played a significant role in shaping ESG outcomes, justifying their inclusion in our final model. Missing values in ESG and financial variables were carefully imputed using the micemethod, supported by visual and statistical evidence suggesting that missingness was at least partially random.

The final regression model, which incorporated log-transformed revenue and emissions along with categorical indicators for industry and region, performed strongly—explaining approximately 63% of the variation in ESG scores in the training set and 60% in the test set. Coefficients aligned with expectations: higher revenue predicted higher ESG scores, while greater carbon emissions predicted lower ones. Industry and regional effects were also substantial and highly significant.

Although the model is both interpretable and predictive, there is room for refinement. Future work could explore non-linear methods such as random forests or gradient boosting, introduce interaction terms, or include time-lagged variables to better reflect real-world dynamics. Additional data sources—like qualitative disclosures or ESG controversies—could also enhance the model's depth.

In conclusion, this project provided a solid statistical foundation for understanding ESG performance and demonstrated that company structure, geographic context, and environmental footprint matter just as much—if not more—than pure financial strength. The approach is scalable, interpretable, and grounded in rigorous analysis, offering valuable insights for both researchers and decision-makers interested in sustainability.

**10. GitHub Repository**

* *Link to GitHub repository with full code and data: [Insert your link here]*