FIN41910: Green Data Science

Group Data Project

June 23, 2023

I/We declare that all material included in this project is the end result of my/our own work and that due acknowledgement has been given in the bibliography and references to all sources be they printed, electronic or personal.

	Name & Surname	Student Number
1.	Gavin Connolly	18308483
2.	Sofia Emelianova	22205187
3.	Ajit Nambiyar	22200172
4.	Fanis-Filippos Papanikolaou	22200286

1 Executive summary

1.1 Research Purpose

The growing global concern about climate change's effect on our environment has grown in recent years. Companies from varying sectors, contributing to carbon emissions, are now subjected to environmental impact inspections. The factors contributing to a company's carbon emissions are multifaceted and interconnected, influenced by diverse and complicated aspects of their operations. By identifying and addressing these factors, businesses can take proactive steps towards mitigating their carbon footprint and fostering sustainability.

This paper aims to identify and explore the various factors that potentially explain companies' carbon emissions from a finance perspective, with a focus on examining the effect a company's efficiency of capital allocation, as measured by Return On Equity, has on the level of emissions exhibited by the company. The study was conducted on randomly selected 100 companies listed on Race To Zero (RTZ) which vary in both size and industry. Understanding these factors can help companies identify areas where emissions reductions can be made, develop effective strategies to minimize their environmental impact, and contribute to the global effort to combat climate change.

1.2 Research Design

Using emissions data from the Race To Zero data explorer, and corresponding financial data sourced from Refinitiv Eikon, this report seeks to model Scope 1, Scope 2 (Market Based), & Scope 2 (Location Based) emissions using companies financial information.

1.3 Findings

The model showed a non-significant parameter for Scope 1 emissions, with Scope 2 emissions having a stronger relationship with ROE. This implies that while ROE is not a great predictor for direct emissions from a company, it can be useful in predicting the indirect emissions.

1.4 Research Limitations

The primary limitation of this report is the voluntary and non-standardized nature of emissions reporting. The emissions data was sourced from the RaceToZero data explorer, which is a voluntary campaign seeking to halve emissions by the year 2030. As a result, only companies which are actively seeking to lessen their carbon footprint in a meaningful way are included in the study, which may introduce sampling bias into the data, harming the generalizability of findings to the overall market.

In addition to this, due to the lack of availability of financial data for non-listed companies, the study has been limited to publicly-listed companies which are subject to disclosure requirements.

1.5 Research Implications

The report found return on equity to be useful for estimating direct emissions, Scope 2 Market-Based & Location-Based, but not indirect Scope 1 emissions. Investors can align their portfolios with sustainability goals using ROE, but should be cautious and consider other relevant factors. These findings would allow pursue their financial objectives, while also contributing to a sustainable investment landscape.

2 Literature review and hypothesis development

A number of studies have been conducted throughout the years regarding this topic. The main takeaway from reading relevant literature is that there are many factors that can potentially explain companies' carbon emissions. By using statistical analysis, this paper will focus on 7 carefully selected factors in an attempt to answer the given problem.

2.1 Return-on-Equity

Return-on-Equity demonstrates the extent of a the company's capabilities in generating profits therefore, high Return-on-Equity represents successful utilisation of capital. The rationale behind the choice of this variable is that high Return-on-Equity can hint that there might be some investments in emission-intensive activities. Additionally, companies with higher Return-on-Equity tend to be more profitable and will experience faster growth rates. Expansions may increase the carbon footprint and so high Return-on-Equity might also be associated with increased emissions. A study by *Hendler & Hunter (2022)* did not find a statistically significant relationship between Return-on-Equity and total Greenhouse gas emissions. However, *Choi et al. (2013)* mentioned that financially strong companies, may afford to invest in extra personnel or capital required for improved prevention of carbon emissions.

H_0: *A company's efficiency in their allocation of capital has no relationship with emissions levels.*
H_1: *A company's efficiency in their allocation of capital has a relationship with emissions levels.*

3 Sample and Data

To avoid industry, country of origin/operation, and company size bias, 100 Race to Zero members were selected at random. The final randomised sample of industries yielded the following results:

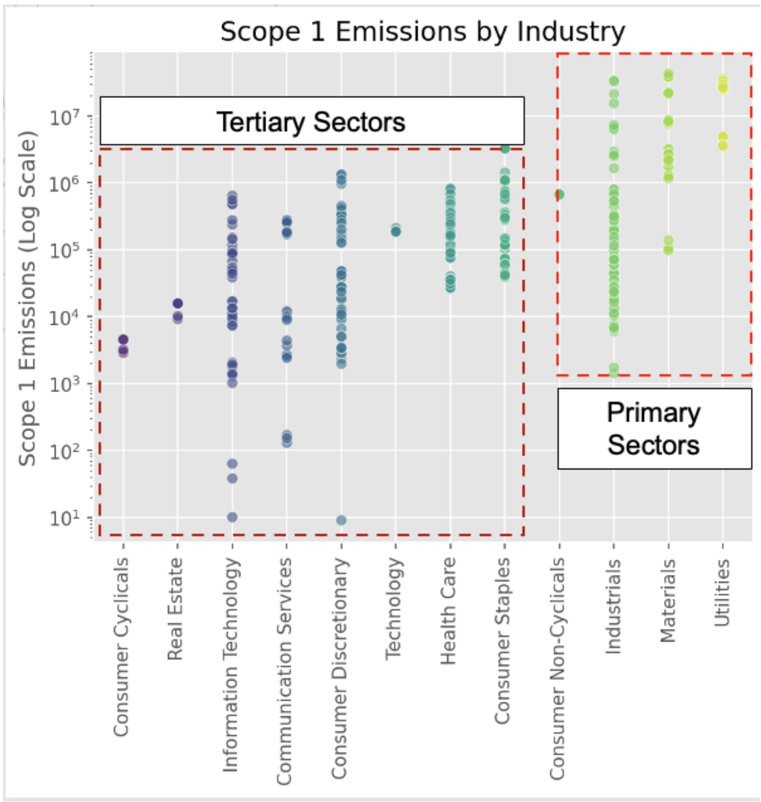
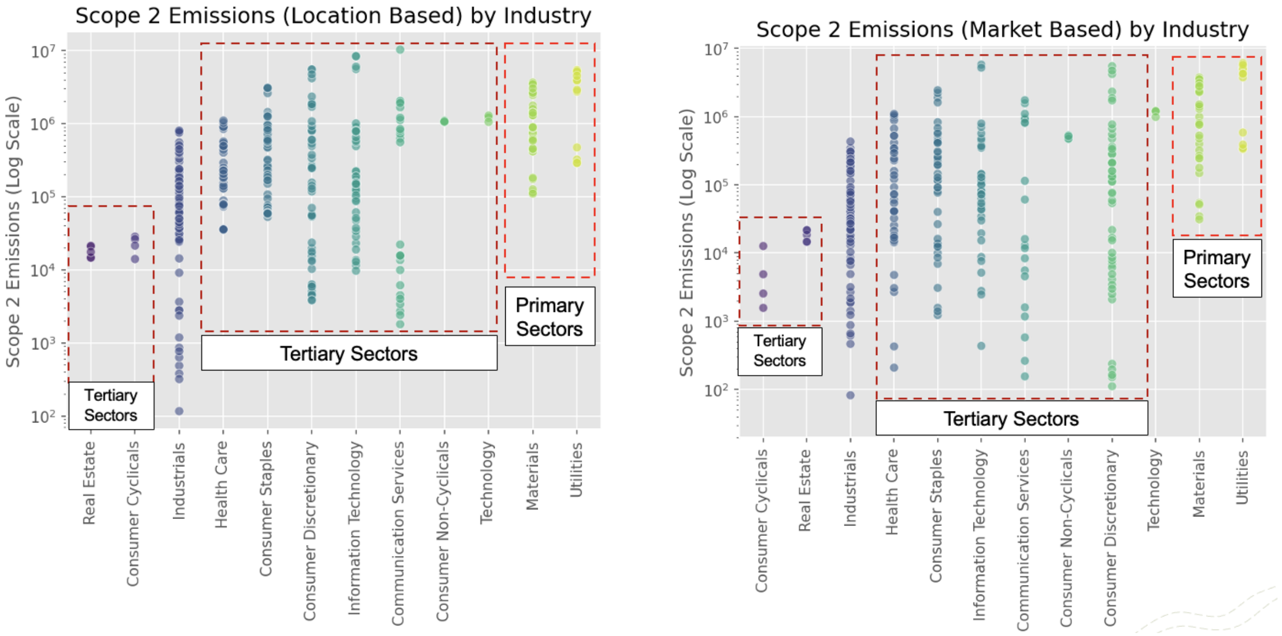


Figure 1: Scope 1 Emissions by Industry



(a) Location based

(b) Market based

Figure 2: Scope 2 Emissions by Industry

Scope 1 (direct emission) and scope 2 (indirect emission) were the dependent variables of interest in this study and as such, were sourced accordingly for each company.

For both dependent and independent variables, annual four-year emissions data for the period of 2018-2021 was used. To account for accuracy and consistency, annual financial data was sourced for each company for the month of December throughout the four-year period of study. Since December is a common month for end-of-year company reports, this eliminated any potential missing data for the independent variables.

Figure 1 illustrates the logarithmic representation of Scope 1 emissions to investigate the emission distribution among companies across various industries. The analysis reveals that industries associated with primary sectors, such as Industrials, Materials, and Utilities, exhibit higher Scope 1 emissions compared to tertiary sectors like Technology, Healthcare, and Information Technology. This observation indicates that emissions tend to be more significant in the lower segment of the economic value chain and gradually decrease as we ascend the value chain.

Figure 2 and Figure 3 depict the relationship between Location-based and Market-based Scope 2 emissions, respectively. Similar to the findings in Figure 1, both figures demonstrate a comparable pattern. However, it should be noted that certain companies in the tertiary sector exhibit higher emissions when compared to those in the primary sector. Nonetheless, the emissions within these tertiary industries exhibit a wide dispersion, suggesting a higher degree of variability. These instances can be considered as outliers since the average emissions of these tertiary industries still remain lower than those of primary sector-based industries.

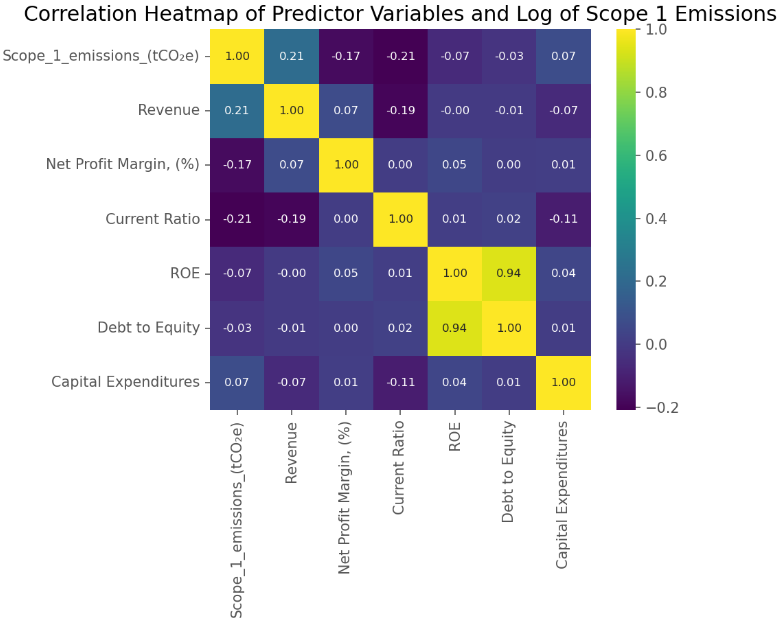
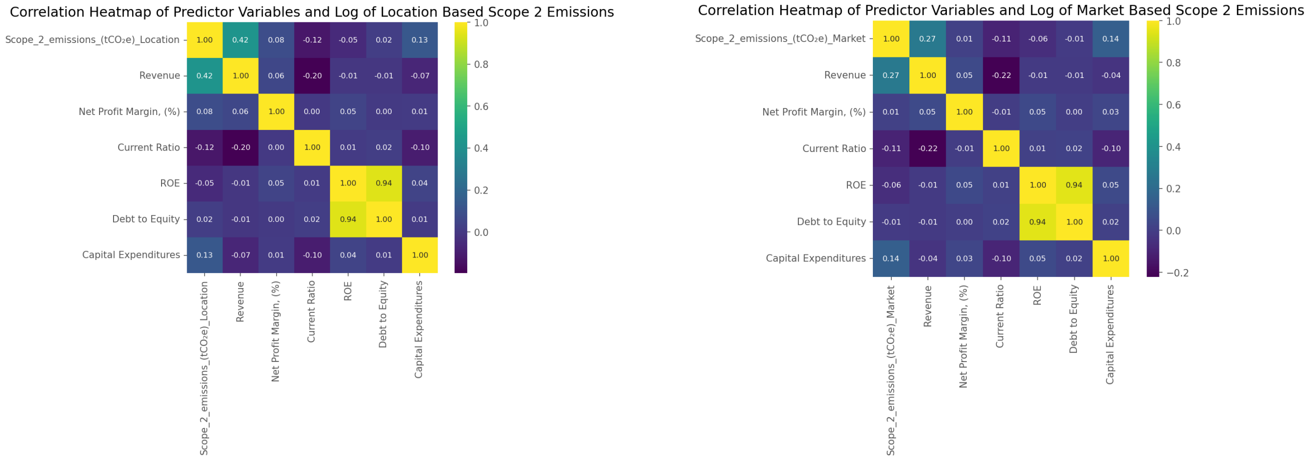


Figure 3: Scope 1 Emissions Correlation Heatmap of Predictor Variables



(a) Location based

(b) Market based

Figure 4: Scope 2 Emissions Correlation Heatmap of Predictor Variables

In figure 3 revenue correlates the highest with Scope 1 emissions out of all. This is because revenue was used as a proxy for company size. Both Return-on-Equity and Debt-to-Equity have the highest correlation among one another which is due to the fact that both ratios include total equity in the denominator. A positive correlation indicates a direct relationship between two variables. In statistical terms, it means that as one variable increases, the other variable tends to increase as well, and vice versa. The remainder of the variables do not significantly correlate with one another. This means that this study will not suffer from multicollinearity.

Similar results are apparent for Scope 2 emissions in figure 4. There appear to be a few variables that negatively correlate with one another which means that there is an inverse relationship between two variables i.e., as one variable increases, the other variable tends to decrease, and vice versa. Some variables have a correlation of zero throughout the heatmaps. This means that there is no linear relationship between two variables i.e., there is no consistent pattern or trend between the values of the two variables. This implies that changes in one variable do not correspond to predictable changes in the other variable. The data points may appear scattered or randomly distributed on the correlation heatmap.

4 Methodology

Emissions data was scraped from the RTZ data explorer using the 'selenium' package. Each cell in the table was iterated through and scraped for the four years of emissions data of each of the 500 companies. The industry associated with each company was similarly gathered from the RTZ data explorer, by iterating through each sector in the table's filter, storing the companies contain in each sector, and then outputting these to a list. These observations could then be merged together based on company names to give the finalized set of emissions data. After screening out non-listed companies as well as one company (Interpublic Group of Companies, Inc.) which was included twice in the RTZ data explorer, 100 companies were sampled from this dataset.

Financial data associated with each of the 100 company was then gathered using the Refinitiv Eikon Python API. 6 financial measurements, Revenue, Net Profit Margin, Current Ratio, ROE, Debt to Equity, & Capital Expenditure, were chosen in order to capture the size, profitability, liquidity, efficiency, leverage and rate of infrastructure development respectively. In order to avoid diluting the size measurement, the capital expenditures of each company were regularized with respect to their revenues.

In order to address the huge variance in the scale in the emissions levels as well as the revenues of the 100 companies in the sample, a log-transform was performed on each of these variables. To capture the varying sector-based emission levels discussed previously, an indicator variable was created to account for whether a company was operating in a tertiary industry, with Communication Services, Consumer Discretionary, Information Technology, Health Care, Consumer Staples, Real Estate, Consumer Cyclicals, Consumer Non-Cyclicals & Technology being classified as tertiary industries and Industrials, Materials & Utilities being classified as non-tertiary industries. Initially, a panel OLS model with time effects was implemented for each of Scope 1, Scope 2 Market-Based & Scope 2 Location-Based emissions, the parameters & model summary of which are given in the Appendix. However, as the F-test on the hypothesis that the time effects were jointly different from zero, failed to show significance at a 5% level (p-values of 0.6521, 0.3178 & 0.0939 respectively), there was not sufficient statistical evidence to suggest a panel regression was appropriate and so a pooled model was considered instead.

Thus, the each of the final models, using the relevant emissions classification can be expressed as:

$$\begin{aligned} log\left(Emissions_i\right) = \quad & \beta_0 + \beta_1 \cdot log\left(Revenue_i\right) + \beta_2 \cdot NetProfitMargin_i + \beta_3 \cdot CurrentRatio_i \\ & + \beta_4 \cdot ROE_i + \beta_5 \frac{Debt_i}{Equity_i} + \beta_6 \cdot \frac{CapitalExpenditure_i}{Revenue_i} + \beta_7 \cdot \mathbb{1}_{Tertiary} \end{aligned} \tag{1}$$

The output of these models was then analysed, with resulting and findings presented in the following sections.

5 Empirical results

A model summary for each of the models is provided as follows:

Dep. Variable:	Scope_1_emissions_(tCO$_2$e)	**R-squared:**	0.3699
Estimator:	PooledOLS	**R-squared (Between):**	0.3910
No. Observations:	354	**R-squared (Within):**	-0.5740
Log-likelihood	-757.69	**R-squared (Overall):**	0.3699
Dep. Variable:	Scope_2_emissions_(tCO$_2$e)_Location	**R-squared:**	0.4591
Estimator:	PooledOLS	**R-squared (Between):**	0.5047
No. Observations:	349	**R-squared (Within):**	-3.4502
Log-likelihood	-635.47	**R-squared (Overall):**	0.4591
Dep. Variable:	Scope_2_emissions_(tCO$_2$e)_Market	**R-squared:**	0.2876
Estimator:	PooledOLS	**R-squared (Between):**	0.3087
No. Observations:	322	**R-squared (Within):**	-0.1451
Log-likelihood	-685.97	**R-squared (Overall):**	0.2876

A summary of the ROE coefficients & significance levels for each model are as follows:

	Parameter	**Std. Err.**	**T-stat**	**P-value**	**Lower CI**	**Upper CI**
Scope 1 Emissions	-0.0760	0.1255	-0.6053	0.5454	-0.3229	0.1709
Scope 2 Emissions Location-Based	-0.2572	0.1312	-1.9607	0.0508	-0.5152	0.0009
Scope 2 Emissions Market-Based	-0.3204	0.0926	-3.4612	0.0006	-0.5024	-0.1383

A full model summary & parameter estimates for each model can be found in the Appendix.

From these results, it can seen that ROE was useful in estimating Scope 2 Market-Based emissions, with a marginal p-value for Location-Based Scope 2 emission. The parameter estimate for Scope 1 emissions was not statistically significantly different from 0, indicating that ROE is not useful in predicting this class of emissions.

The actual vs. predicted values were then plotted, with ROE being represented by the colour of the points in the scatterplot, with the following results.

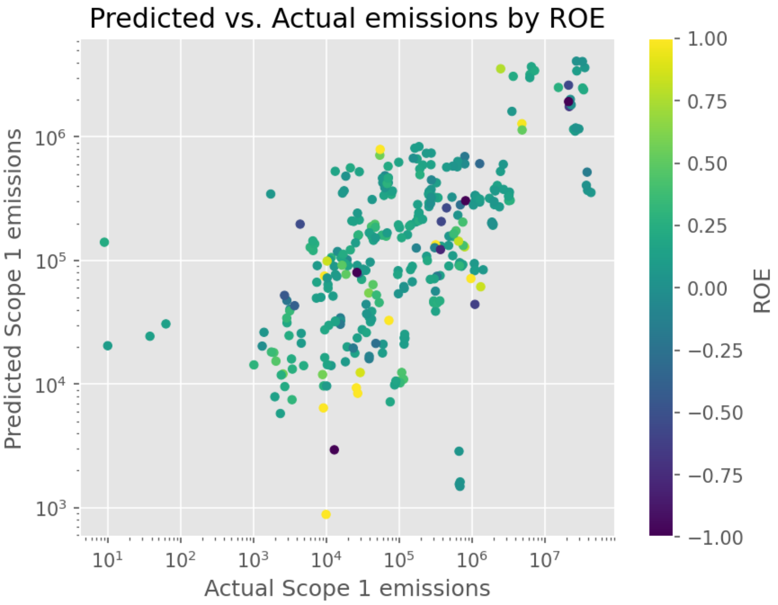
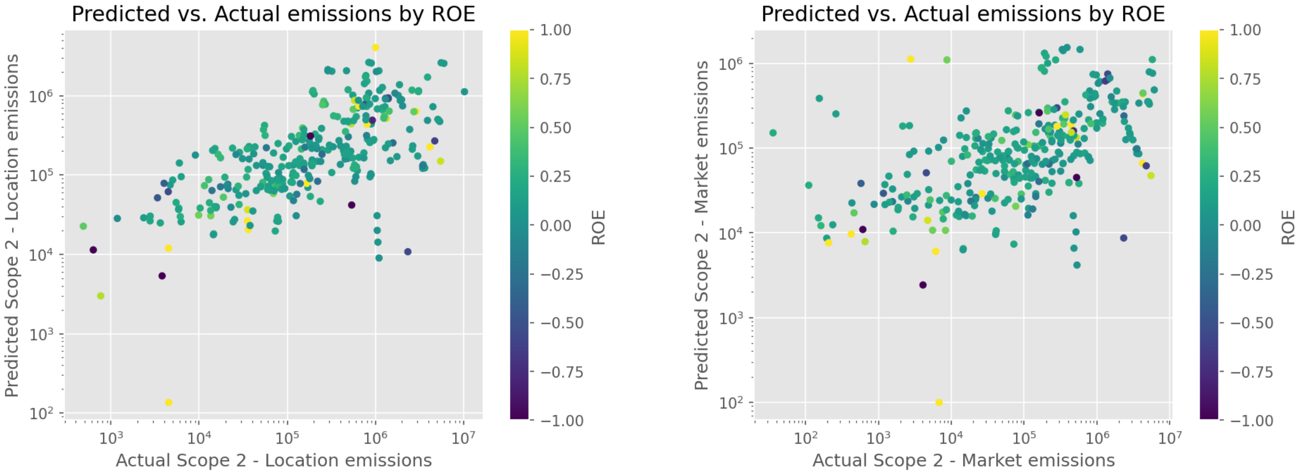


Figure 5: Scope 1 Emissions - Predicted vs. Fitted



(a) Location based

(b) Market based

Figure 6: Scope 2 Emissions - Predicted vs. Fitted

These graphs demonstrate that the model used has some explanatory power in predicting emissions levels, particularly in the case of Scope 2 emissions. The observed general trend reveals that companies with lower ROE tend to exhibit higher emissions, while those with higher ROE tend to have lower emissions. Although this trend holds true in many instances, it is essential to acknowledge the presence of outliers at either end, indicating that the relationship between ROE and emissions is not absolute.

6 Conclusion

In conclusion, the findings of this study provide valuable insights for investors who prioritize environmental, social, and governance (ESG) considerations and are concerned about emissions in their investment portfolios. Specifically, two key findings shed light on the relationship between return on equity (ROE) and emissions.

Firstly, the results indicate that ROE is a useful metric for estimating Scope 2 Market-Based emissions. The marginal p-value for Location-Based Scope 2 emissions suggests a moderate relationship between ROE and emissions. This finding implies that investors who focus on ROE as a key performance indicator can use it as a valuable tool in estimating and assessing the environmental impact of companies' operations related to Scope 2 Market-Based emissions. However, it is important to note that the parameter estimate for Scope 1 emissions was not statistically significant, indicating that ROE may not be a reliable predictor for this category of emissions.

For investors concerned about emissions and committed to ESG principles, these findings offer valuable guidance. By incorporating ROE as a factor in their investment decision-making process, investors can gain insights into the potential environmental impact of companies' operations, particularly in relation to Scope 2 Market-Based emissions. This knowledge allows investors to align their portfolios with their sustainability objectives and make informed investment choices.

Nevertheless, it is important to exercise caution when interpreting the results and making investment decisions solely based on ROE. While the relationship between ROE and emissions is evident to some extent, other factors, such as industry-specific characteristics, regulatory environments, and company-specific practices, may also influence emissions levels. Therefore, investors should adopt a comprehensive approach that considers a range of ESG indicators, alongside ROE, to ensure a well-rounded assessment of a company's sustainability performance.

In summary, investors focusing on ESG and emissions concerns can leverage the findings of this study by incorporating ROE as a metric to estimate and assess Scope 2 Market-Based emissions. This approach enhances their ability to align their investment portfolios with sustainability objectives. However, it is crucial to recognize the limitations of ROE as a standalone indicator and to consider other relevant factors when making investment decisions. By doing so, investors can contribute to a more sustainable and responsible investment landscape while pursuing their financial goals.

A Appendix

A.1 Panel OLS Model Summaries

A.1.1 Scope 1 Emissions

Dep. Variable:	Scope_1_emissions_(tCO$_2$e)	**R-squared:**	0.3701
Estimator:	PanelOLS	**R-squared (Between):**	0.3911
No. Observations:	354	**R-squared (Within):**	-0.5910
Log-likelihood	-756.85	**R-squared (Overall):**	0.3698

Scope 1 Emissions (tCO$_2$e)	**Parameter**	**Std. Err.**	**T-stat**	**P-value**	**Lower CI**	**Upper CI**
Constant	-10.459	2.4256	-4.3118	0.0000	-15.229	-5.6877
Revenue	0.9852	0.1002	9.8352	0.0000	0.7882	1.1822
Net Profit Margin, (%)	-2.5369	0.6669	-3.8042	0.0002	-3.8486	-1.2252
Current Ratio	0.1518	0.1168	1.3000	0.1945	-0.0779	0.3814
ROE	-0.0901	0.1268	-0.7106	0.4778	-0.3395	0.1593
Debt to Equity	0.0123	0.0224	0.5500	0.5827	-0.0318	0.0564
Capital Expenditures	13.192	4.2749	3.0859	0.0022	4.7837	21.600
Industry_Tertiary	-2.0380	0.2389	-8.5325	0.0000	-2.5078	-1.5682

A.1.2 Scope 2 Emissions Location

Dep. Variable:	Scope_2_emissions_(tCO$_2$e)_Location	**R-squared:**	0.4612
Estimator:	PanelOLS	**R-squared (Between):**	0.5042
No. Observations:	349	**R-squared (Within):**	-3.4483
Log-likelihood	-633.66	**R-squared (Overall):**	0.4589

Scope 2 Emissions (tCO$_2$e) Location	**Parameter**	**Std. Err.**	**T-stat**	**P-value**	**Lower CI**	**Upper CI**
const	-14.838	1.7674	-8.3951	0.0000	-18.314	-11.361
Revenue	1.1117	0.0730	15.238	0.0000	0.9682	1.2552
Net Profit Margin, (%)	0.2756	0.4838	0.5696	0.5693	-0.6761	1.2272
Current Ratio	0.4087	0.0847	4.8280	0.0000	0.2422	0.5752
ROE	-0.3311	0.0931	-3.5559	0.0004	-0.5142	-0.1479
Debt to Equity	0.0611	0.0165	3.7010	0.0003	0.0286	0.0936
Capital Expenditures	15.216	3.0849	4.9324	0.0000	9.1479	21.284
Industry_Tertiary	-0.1880	0.1744	-1.0779	0.2818	-0.5310	0.1551

A.1.3 Scope 2 Emissions Market

Dep. Variable:	Scope_2_emissions_(tCO_2e)_Market	**R-squared:**	0.2896
Estimator:	PanelOLS	**R-squared (Between):**	0.3133
No. Observations:	322	**R-squared (Within):**	-0.1625
Log-likelihood	-682.67	**R-squared (Overall):**	0.2867

Scope 2 Emissions (tCO_2e) Market	**Parameter**	**Std. Err.**	**T-stat**	**P-value**	**Lower CI**	**Upper CI**
Constant	-14.004	2.6671	-5.2507	0.0000	-19.252	-8.7565
Revenue	1.0606	0.1103	9.6117	0.0000	0.8435	1.2777
Net Profit Margin, (%)	-0.4458	0.6668	-0.6686	0.5043	-1.7578	0.8662
Current Ratio	0.4344	0.1198	3.6278	0.0003	0.1988	0.6701
ROE	-0.2962	0.1314	-2.2548	0.0248	-0.5547	-0.0377
Debt to Equity	0.0514	0.0233	2.2054	0.0282	0.0055	0.0973
Capital Expenditures	16.657	4.6987	3.5449	0.0005	7.4114	25.902
Industry_Tertiary	-0.9065	0.2437	-3.7190	0.0002	-1.3860	-0.4269

A.1.4 Estimated Time Effects

Year	**Scope 1**	**Scope 2 Location**	**Scope 2 Market**
2018	0.26	0.28	0.32
2019	-0.03	-0.03	0.30
2020	-0.10	-0.13	-0.14
2021	-0.10	-0.08	-0.38
F-test for Poolability	0.5445	1.1786	2.1506
p-value	0.6521	0.3178	0.0939

Back to Report

A.2 Pooled OLS Model Summaries

A.2.1 Scope 1 Emissions

Dep. Variable:	Scope_1_emissions_(tCO_2e)	**R-squared:**	0.3699
Estimator:	PooledOLS	**R-squared (Between):**	0.3910
No. Observations:	354	**R-squared (Within):**	-0.5740
Log-likelihood	-757.69	**R-squared (Overall):**	0.3699

Scope 1 (tCO$_2$e)	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
Constant	-10.550	2.4197	-4.3599	0.0000	-15.309	-5.7904
Revenue	0.9887	0.0999	9.8938	0.0000	0.7921	1.1852
Net Profit Margin, (%)	-2.4641	0.6600	-3.7332	0.0002	-3.7623	-1.1659
Current Ratio	0.1498	0.1165	1.2857	0.1994	-0.0794	0.3789
ROE	-0.0760	0.1255	-0.6053	0.5454	-0.3229	0.1709
Debt to Equity	0.0090	0.0222	0.4048	0.6859	-0.0346	0.0526
Capital Expenditures	13.597	4.2274	3.2163	0.0014	5.2819	21.911
Industry_Tertiary	-2.0487	0.2381	-8.6029	0.0000	-2.5171	-1.5803

A.2.2 Scope 2 Emissions Location

Dep. Variable:	Scope_2_emissions_(tCO$_2$e)_Location	**R-squared:**	0.4591
Estimator:	PooledOLS	**R-squared (Between):**	0.5047
No. Observations:	349	**R-squared (Within):**	-3.4502
Log-likelihood	-635.47	**R-squared (Overall):**	0.4591

Scope 2 (tCO$_2$e) Location	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	-14.856	1.7688	-8.3989	0.0000	-18.335	-11.376
Revenue	1.1120	0.0730	15.230	0.0000	0.9684	1.2556
Net Profit Margin, (%)	0.3357	0.4804	0.6988	0.4851	-0.6092	1.2807
Current Ratio	0.4054	0.0847	4.7862	0.0000	0.2388	0.5720
ROE	-0.3204	0.0926	-3.4612	0.0006	-0.5024	-0.1383
Debt to Equity	0.0583	0.0164	3.5557	0.0004	0.0261	0.0906
Capital Expenditures	15.639	3.0601	5.1105	0.0000	9.6196	21.658
Industry_Tertiary	-0.1929	0.1745	-1.1058	0.2696	-0.5361	0.1503

Scope 2 Emissions Market

Dep. Variable:	Scope_2_emissions_(tCO$_2$e)_Market	**R-squared:**	0.2876
Estimator:	PooledOLS	**R-squared (Between):**	0.3087
No. Observations:	322	**R-squared (Within):**	-0.1451
Log-likelihood	-685.97	**R-squared (Overall):**	0.2876

Scope 2 (tCO_2e) Market	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	-14.043	2.6756	-5.2486	0.0000	-19.308	-8.7789
Revenue	1.0607	0.1107	9.5815	0.0000	0.8429	1.2785
Net Profit Margin, (%)	-0.2317	0.6635	-0.3492	0.7272	-1.5372	1.0738
Current Ratio	0.4304	0.1203	3.5781	0.0004	0.1937	0.6671
ROE	-0.2572	0.1312	-1.9607	0.0508	-0.5152	0.0009
Debt to Equity	0.0439	0.0232	1.8891	0.0598	-0.0018	0.0896
Capital Expenditures	18.129	4.6845	3.8701	0.0001	8.9125	27.346
Industry_Tertiary	-0.9426	0.2446	-3.8542	0.0001	-1.4238	-0.4614

Back to Report

A.3 Python Code

A.3.1 Scraping the Dependent Variables

```
import pandas as pd
from selenium import webdriver
from selenium.webdriver import FirefoxOptions
from selenium.webdriver.common.by import By
from selenium.webdriver.support.wait import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from selenium.webdriver.common.keys import Keys
import time
from sqlalchemy import create_engine
from sqlalchemy_utils import database_exists, create_database

url = 'https://racetozerodataexplorer.org/proceed/'

scope1_lst = []
scope2_local_lst = []
scope2_market_lst = []
name_lst = []

master_lst = [scope1_lst, scope2_local_lst, scope2_market_lst, name_lst]

options = FirefoxOptions()
options.add_argument("--headless") # comment out to see how it works

with webdriver.Firefox(options=options) as driver:
    driver.get(url)
    headers = driver.find_elements(By.XPATH,
    "//*[@class= 'styles_table__jWPKo']/thead/tr/th")
    headers = [header.text for header in headers]
    table = driver.find_element(By.TAG_NAME,
    "table").find_element(By.TAG_NAME, "tbody").find_elements(By.TAG_NAME, 'tr')
    for row in table:
        name = row.find_element(By.TAG_NAME, 'th')
        # print(name.text)
        master_lst[3].append(name.text)
        for i, cell in enumerate(row.find_elements(By.TAG_NAME, 'td')):
            if i in [0,1,2]: # only interested in first 3 columns
                try:
                    cell.find_element(By.CLASS_NAME, 'styles_button__IS06I').click()
                    WebDriverWait(driver, 10).until( # Wait for pop-up to load
                    EC.presence_of_element_located((By.CLASS_NAME, 'styles_axisValue__ip0q1')))
                    values = driver.find_elements(By.CLASS_NAME, 'styles_axisValue__ip0q1')
                    master_lst[i].append([value.text.replace(',','') for value in values])
                    driver.find_element(By.CLASS_NAME, 'styles_close__cpVjT').click()
                except:
                    master_lst[i].append(['', '', '', ''])

industries = [
    'Basic Materials',
    'Communication Services',
    'Consumer Cyclicals',
    'Consumer Discretionary',
    'Consumer Non-Cyclicals',
    'Consumer Staples',
    'Energy',
    'Health Care',
    'Industrials',
    'Information Technology',
    'Materials',
    'Real Estate',
    'Technology',
    'Utilities',
]

sectors = []

with webdriver.Firefox(options=options) as driver:
    driver.get(url)
    for industry in industries:
        filter = driver.find_element(By.ID, 'react-select-q-input')
        filter.click()
        filter.send_keys(industry)
        time.sleep(0.2)
        if industry in ['Materials', 'Technology']:
            filter.send_keys(Keys.DOWN)
        filter.send_keys(Keys.ENTER)
        time.sleep(0.2)
        table = driver.find_element(By.TAG_NAME, "table").find_element(By.TAG_NAME, 'tbody')
        names = table.find_elements(By.TAG_NAME, 'th')
        for name in names:
            sectors.append([industry, name.text])
        driver.find_element(By.CLASS_NAME, 'css-g010ao').click()
```

```
headers = [header.strip().replace('\n', ' ').replace(
    '4 years trend (ending in reporting year)', '') for header in headers]
years = ['2018', '2019', '2020', '2021']

# Need to merge all these
df_scope1 = pd.DataFrame(master_lst[0], index=master_lst[3],
columns=years).reset_index(drop=False)
df_scope2_loc = pd.DataFrame(master_lst[1], index=master_lst[3],
columns=years).reset_index(drop=False)
df_scope2_mkt = pd.DataFrame(master_lst[2], index=master_lst[3],
columns=years).reset_index(drop=False)

df_sectors = pd.DataFrame(master_lst[4], columns=['Industry'], index=master_lst[3])

engine = create_engine("sqlite:///GDS.db")
if not database_exists(engine.url):
    create_database(engine.url)

# Add dataframes to database to avoid having to rerun webscraping each time
for df, header in [(df_scope1, headers[1]), (df_scope2_loc, headers[2]),
(df_scope2_mkt, headers[3])]:
    df.rename(columns={'index': 'Company Name'}, inplace=True)
    df.to_sql(header.replace(' ', '_'), engine, if_exists='replace', index=False)

df_sectors.to_sql('Sectors', engine, if_exists="replace", index=False)

df_scope1 = pd.read_sql_table('Scope_1_emissions_(tCO2e)', engine)
df_scope2_loc = pd.read_sql_table('Scope_2_emissions_(tCO2e)_Location', engine)
df_scope2_mkt = pd.read_sql_table('Scope_2_emissions_(tCO2e)_Market', engine)

df_sectors = pd.read_sql_table('Sectors', engine)

# Drop Interpublic Group of Companies, Inc. which is duplicated in data source
df_scope1 = df_scope1.loc[~df_scope1['Company Name'].duplicated()]
df_scope2_loc = df_scope2_loc.loc[~df_scope2_loc['Company Name'].duplicated()]
df_scope2_mkt = df_scope2_mkt.loc[~df_scope2_mkt['Company Name'].duplicated()]

# Filter out companies with no data
df_scope1 = df_scope1.loc[~df_scope1.iloc[:,1:].duplicated()]
df_scope2_loc = df_scope2_loc.loc[~df_scope2_loc.iloc[:,1:].duplicated()]
df_scope2_mkt = df_scope2_mkt.loc[~df_scope2_mkt.iloc[:,1:].duplicated()]

df_scope1 = df_scope1.melt(id_vars='Company Name',
value_name='Scope_1_emissions_(tCO2e)', var_name='Year')
df_scope2_loc = df_scope2_loc.melt(id_vars='Company Name',
value_name='Scope_2_emissions_(tCO2e)_Location', var_name='Year')
df_scope2_mkt = df_scope2_mkt.melt(id_vars='Company Name',
value_name='Scope_2_emissions_(tCO2e)_Market', var_name='Year')
companies_sample

df = pd.read_sql_table('Emissions_Data_Joined', engine)
companies_sample = pd.read_csv("CompaniesSample.csv")
```

```
df = df.merge(companies_sample['Company Name'], on='Company Name', how='inner')
df.replace('', '-', inplace=True)
df.to_sql('Emissions_Data_Joined', engine, if_exists='replace', index=False)

table = df[['Year', 'Scope_1_emissions_(tCO2e)']].loc[df
['Scope_1_emissions_(tCO2e)']!='-'].astype(int).groupby('Year').describe(percentiles=[])
table.columns = table.columns.set_levels(table.columns.levels[1].str.replace('50%',
'median'), level=1)
table

table = df[['Year', 'Scope_2_emissions_(tCO2e)_Location']].loc[df
['Scope_2_emissions_(tCO2e)_Location']!='-'].astype(int).groupby('Year').describe(percentiles=[])
table.columns = table.columns.set_levels(table.columns.levels[1].str.replace('50%',
'median'), level=1)
table

table = df[['Year', 'Scope_2_emissions_(tCO2e)_Market']].loc[df
['Scope_2_emissions_(tCO2e)_Market']!='-'].astype(int).groupby('Year').describe(percentiles=[])
table.columns = table.columns.set_levels(table.columns.levels[1].str.replace('50%',
'median'), level=1)
table
```

A.3.2 Gathering the Financial Data

```
import eikon as ek
import pandas as pd
from sqlalchemy import create_engine
ek.set_app_key('150b92bce26746bbbee1258f4ac8453b6cb6a1ae')

df_tickers = pd.read_csv("CompaniesSample.csv")

tickers_lst = df_tickers['Ticker'].to_list()

fin_vars_lst = [
    'TR.TotalReturn.date',
    'TR.PCTotAsset',
    'TR.TotalEquity',
    'TR.Revenue',
    'TR.NetProfitMargin',
    'TR.TotalDebtOutstanding',
    'TR.CapitalExpenditures',
    'TR.CurrentRatio'
    ]

df, err = ek.get_data(tickers_lst, fin_vars_lst, {'SDate': '2017-12-31',
'EDate': '2022-12-31', 'FRQ': 'FY', 'Curn': 'USD'})
df

df_fin = df_tickers[['Company Name', 'Ticker']].merge(df, right_on='Instrument',
left_on='Ticker').drop(columns='Instrument')
```

```
df_fin['Date'] = df_fin['Date'].str[:-10] # Get rid of TZ info
df_fin = df_fin.loc[df_fin['Date'].str[:4].isin(['2017','2018',
'2019','2020','2021'])].reset_index(drop=True)

engine = create_engine("sqlite:///GDS.db")
df_fin.to_sql('FinancialsTableRaw', engine, index=False, if_exists="replace")

df_fin = pd.read_sql_table('FinancialsTableRaw', engine)

df_fin['Net Profit Margin, (%)'] = df_fin['Net Profit Margin, (%)']/100
df_fin['Net Profit'] = df_fin['Net Profit Margin, (%)']*df_fin['Revenue']
df_fin['ROE'] = df_fin['Net Profit']/df_fin['Total Equity']

df_fin['Debt to Equity'] = df_fin['Total Debt']/df_fin['Total Equity']

df_fin['Capital Expenditures'] = df_fin['Capital Expenditures,
Cumulative']/df_fin['Total Assets (Pvt)']*-1

df_fin.drop(columns=['Ticker', 'Total Equity', 'Total Assets (Pvt)', 'Total Debt',
'Capital Expenditures, Cumulative', 'Net Profit'], inplace=True)
df_fin

df_fin.to_sql('FinancialsTableFinal', engine, index=False, if_exists="replace")
```

A.3.3 Basic plots and correlation heatmaps

```
import pandas as pd
from sqlalchemy import create_engine
import matplotlib.pyplot as plt
import matplotlib as mpl
import seaborn as sns
import numpy as np
plt.style.use('ggplot')
mpl.rcParams['figure.dpi'] = 150

engine = create_engine("sqlite:///GDS.db")
df = pd.read_sql_table('Emissions_Data_Joined', engine)
df_sectors = pd.read_sql_table('Sectors', engine)

df_plot = df[['Company Name', 'Scope_1_emissions_(tCO2e)']].merge(
    df_sectors, on='Company Name').loc[df['Scope_1_emissions_(tCO2e)']!='-']

df_plot['Scope_1_emissions_(tCO2e)'] = df_plot['Scope_1_emissions_(tCO2e)'].astype(int)

df_plot = df[['Company Name', 'Scope_1_emissions_(tCO2e)']].merge(
    df_sectors, on='Company Name').loc[df['Scope_1_emissions_(tCO2e)']!='-']

df_plot['Scope_1_emissions_(tCO2e)'] = df_plot['Scope_1_emissions_(tCO2e)'].astype(int)

industries_sorted = df_plot.groupby('Industry').mean().sort_values(by=
'Scope_1_emissions_(tCO2e)').index
```

```
df_plot['Industry'] = pd.Categorical(df_plot['Industry'], industries_sorted)

sns.scatterplot(data=df_plot.sort_values('Industry'), alpha=0.6, x='Industry',
                y='Scope_1_emissions_(tCO2e)',
                hue='Industry', palette='viridis', legend=False)
plt.yscale("log")
plt.ylabel('Scope 1 Emissions (Log Scale)')
plt.xlabel('')
plt.xticks(rotation=90)
plt.title('Scope 1 Emissions by Industry')
plt.show()

df_plot = df[['Company Name', 'Scope_2_emissions_(tCO2e)_Location']].merge(
    df_sectors, on='Company Name').loc[df['Scope_2_emissions_(tCO2e)_Location']!='-']

df_plot['Scope_2_emissions_(tCO2e)_Location'] = df_plot[
'Scope_2_emissions_(tCO2e)_Location'].astype(int)

df_plot['Industry'] = pd.Categorical(df_plot['Industry'], industries_sorted)

sns.scatterplot(data=df_plot.sort_values('Industry'), alpha=0.6,
                x='Industry', y='Scope_2_emissions_(tCO2e)_Location',
                hue='Industry', palette='viridis', legend=False)
plt.yscale("log")
plt.ylabel('Scope 2 Emissions (Log Scale)')
plt.xlabel('')
plt.xticks(rotation=90)
plt.title('Scope 2 Emissions (Location Based) by Industry')
plt.show()

df_plot = df[['Company Name', 'Scope_2_emissions_(tCO2e)_Market']].merge(
    df_sectors, on='Company Name').loc[df['Scope_2_emissions_(tCO2e)_Market']!='-']

df_plot['Scope_2_emissions_(tCO2e)_Market'] = df_plot[
'Scope_2_emissions_(tCO2e)_Market'].astype(int)

df_plot['Industry'] = pd.Categorical(df_plot['Industry'], industries_sorted)

sns.scatterplot(data=df_plot.sort_values('Industry'), alpha=0.6,
                x='Industry', y='Scope_2_emissions_(tCO2e)_Market',
                hue='Industry', palette='viridis', legend=False)
plt.yscale("log")
plt.ylabel('Scope 2 Emissions (Log Scale)')
plt.xlabel('')
plt.xticks(rotation=90)
plt.title('Scope 2 Emissions (Market Based) by Industry')
plt.show()

df_fin = pd.read_sql_table('FinancialsTableFinal', engine)
df_fin['Year'] = df_fin['Date'].str[:4].astype(int)+1
df_fin.drop(columns='Date', inplace=True)
```

```
df['Year'] = df['Year'].astype(int)
df_merged = df.merge(df_fin, on=['Company Name', 'Year']).replace('-', None)

df_scope1 =  df_merged.drop(columns=['Scope_2_emissions_(tCO2e)_Location',
             'Scope_2_emissions_(tCO2e)_Market']).dropna().drop(columns='Company Name')
df_scope1['Scope_1_emissions_(tCO2e)'] =
np.log10(df_scope1['Scope_1_emissions_(tCO2e)'].astype(float))

corr_matrix = df_scope1.drop(columns='Year').corr()
cols = corr_matrix.columns
hm = sns.heatmap(corr_matrix, cbar=True, annot=True, square=True, fmt='.2f',
                 annot_kws={'size': 8}, yticklabels=cols, xticklabels=cols, cmap='viridis')
hm.set_title("Correlation Heatmap of Predictor Variables and Log of Scope 1 Emissions")
plt.show()

df_scope2 =  df_merged.drop(columns=['Scope_1_emissions_(tCO2e)',
             'Scope_2_emissions_(tCO2e)_Market']).dropna().drop(columns='Company Name')
df_scope2['Scope_2_emissions_(tCO2e)_Location'] = np.log10(df_scope2
['Scope_2_emissions_(tCO2e)_Location'].astype(float))

corr_matrix = df_scope2.drop(columns='Year').corr()
cols = corr_matrix.columns
hm = sns.heatmap(corr_matrix, cbar=True, annot=True, square=True, fmt='.2f',
                 annot_kws={'size': 8}, yticklabels=cols, xticklabels=cols, cmap='viridis')
hm.set_title("Correlation Heatmap of Predictor Variables and
Log of Location Based Scope 2 Emissions")
plt.show()

df_scope2 =  df_merged.drop(columns=['Scope_1_emissions_(tCO2e)',
             'Scope_2_emissions_(tCO2e)_Location']).dropna().drop(columns='Company Name')
df_scope2['Scope_2_emissions_(tCO2e)_Market'] = np.log10(df_scope2
['Scope_2_emissions_(tCO2e)_Market'].astype(float))

corr_matrix = df_scope2.drop(columns='Year').corr()
cols = corr_matrix.columns
hm = sns.heatmap(corr_matrix, cbar=True, annot=True, square=True, fmt='.2f',
                 annot_kws={'size': 8}, yticklabels=cols, xticklabels=cols, cmap='viridis')
hm.set_title("Correlation Heatmap of Predictor Variables and
Log of Market Based Scope 2 Emissions")
plt.show()
```

A.3.4 Model Building & Evaluation

```
import pandas as pd
from sqlalchemy import create_engine
import statsmodels.api as sm
from linearmodels.panel import PanelOLS
from linearmodels.panel import PooledOLS
import numpy as np
import matplotlib.pyplot as plt
import matplotlib as mpl
```

```
plt.style.use('ggplot')
mpl.rcParams['figure.dpi'] = 150

engine = create_engine("sqlite:///GDS.db")
df_emissions = pd.read_sql_table('Emissions_Data_Joined', engine)
df_emissions['Year'] = df_emissions['Year'].astype(int)

df_sectors = pd.read_sql_table('Sectors', engine)
df_fin = pd.read_sql_table('FinancialsTableFinal', engine)

df_fin['Year'] = df_fin['Date'].str[:4].astype(int)+1
df_fin.drop(columns='Date', inplace=True)
df_fin['Revenue'] = np.log(df_fin['Revenue']) # log transform revenue
df_panel = df_emissions.merge(df_sectors, on='Company Name').merge(
    df_fin, on=['Company Name', 'Year']).set_index(['Company Name', 'Year'])

df_panel = df_panel.loc[~df_panel['Capital Expenditures'].isna()] # Drop rows with missing values

primary = ['Industrials', 'Materials', 'Utilities']
df_panel.loc[~df_panel['Industry'].isin(primary), 'Industry'] = 'Tertiary'
df_panel.loc[df_panel['Industry'].isin(primary), 'Industry'] = 'Primary'
df_panel = pd.get_dummies(df_panel, columns=['Industry'], drop_first=True)

x_vars = df_panel.columns[3:]

df_scope1 = df_panel[x_vars].merge(df_panel['Scope_1_emissions_(tCO2e)'],
                                    left_index=True, right_index=True)
df_scope2_loc = df_panel[x_vars].merge(df_panel['Scope_2_emissions_(tCO2e)_Location'],
                                        left_index=True, right_index=True)
df_scope2_mkt = df_panel[x_vars].merge(df_panel['Scope_2_emissions_(tCO2e)_Market'],
                                        left_index=True, right_index=True)

# Drop observations without emissions data
df_scope1 = df_scope1.loc[df_scope1['Scope_1_emissions_(tCO2e)'] != '-'].astype(float)
df_scope2_loc = df_scope2_loc.loc[df_scope2_loc['Scope_2_emissions_(tCO2e)_Location'] != '-'].ast
df_scope2_mkt = df_scope2_mkt.loc[df_scope2_mkt['Scope_2_emissions_(tCO2e)_Market'] != '-'].astyp

x = sm.add_constant(df_scope1[x_vars])
y = np.log(df_scope1[['Scope_1_emissions_(tCO2e)']])

model1 = PanelOLS(dependent=y, exog=x, time_effects=True)
res1 = model1.fit()
res1.summary

x = sm.add_constant(df_scope2_loc[x_vars])
y = np.log(df_scope2_loc[['Scope_2_emissions_(tCO2e)_Location']])

model2_loc = PanelOLS(dependent=y, exog=x, time_effects=True)
res2_loc = model2_loc.fit()
print(res2_loc.summary.as_latex())

x = sm.add_constant(df_scope2_mkt[x_vars])
```

```
y = np.log(df_scope2_mkt[['Scope_2_emissions_(tCO2e)_Market']])

model2_mkt = PanelOLS(dependent=y, exog=x, time_effects=True)
res2_mkt = model2_mkt.fit()
print(res2_mkt.summary.as_latex())

x = sm.add_constant(df_scope1[x_vars])
y = np.log(df_scope1[['Scope_1_emissions_(tCO2e)']])

model1 = PooledOLS(dependent=y, exog=x)
res1 = model1.fit()
print(res1.summary.as_latex())

x = sm.add_constant(df_scope2_loc[x_vars])
y = np.log(df_scope2_loc[['Scope_2_emissions_(tCO2e)_Location']])

model2_loc = PooledOLS(dependent=y, exog=x)
res2_loc = model2_loc.fit()
print(res2_loc.summary.as_latex())

x = sm.add_constant(df_scope2_mkt[x_vars])
y = np.log(df_scope2_mkt[['Scope_2_emissions_(tCO2e)_Market']])

model2_mkt = PooledOLS(dependent=y, exog=x)
res2_mkt = model2_mkt.fit()
print(res2_mkt.summary.as_latex())

norm = mpl.colors.Normalize(vmin=-1, vmax=1)

np.exp(model1.predict(params=res1.params, exog=x)).merge(
    df_scope1[['Scope_1_emissions_(tCO2e)', 'ROE']], left_index=True,
        right_index=True).plot.scatter(x='Scope_1_emissions_(tCO2e)',
            y='predictions', c='ROE', colormap='viridis', norm=norm)
plt.xscale('log')
plt.yscale('log')
plt.ylabel('Predicted Scope 1 emissions')
plt.xlabel('Actual Scope 1 emissions')
plt.title('Predicted vs. Actual emissions by ROE')
plt.show()

np.exp(model2_loc.predict(params=res2_loc.params, exog=x)).merge(
    df_scope2_loc[['Scope_2_emissions_(tCO2e)_Location', 'ROE']], left_index=True,
        right_index=True).plot.scatter(x='Scope_2_emissions_(tCO2e)_Location',
            y='predictions', c='ROE', colormap='viridis', norm=norm)
plt.xscale('log')
plt.yscale('log')
plt.ylabel('Predicted Scope 2 - Location emissions')
plt.xlabel('Actual Scope 2 - Location emissions')
plt.title('Predicted vs. Actual emissions by ROE')
plt.show()

np.exp(model2_mkt.predict(params=res2_mkt.params, exog=x)).merge(
    df_scope2_mkt[['Scope_2_emissions_(tCO2e)_Market', 'ROE']], left_index=True,
    right_index=True).plot.scatter(x='Scope_2_emissions_(tCO2e)_Market',
        y='predictions', c='ROE', colormap='viridis', norm=norm)
plt.xscale('log')
plt.yscale('log')
plt.ylabel('Predicted Scope 2 - Market emissions')
plt.xlabel('Actual Scope 2 - Market emissions')
plt.title('Predicted vs. Actual emissions by ROE')
plt.show()
```

Back to Report

B Bibliography

- Xia, M. and Cai, H.H. (2023) 'The driving factors of corporate carbon emissions: An application of the Lasso model with survey data', *Environmental Science and Pollution Research*, 30(19), pp. 56484–56512. doi:10.1007/s11356-023-26081-7.

- Drempetic, S., Klein, C. and Zwergel, B. (2019) 'The influence of firm size on the ESG score: Corporate Sustainability Ratings under review', *Journal of Business Ethics*, 167(2), pp. 333–360. doi:10.1007/s10551-019-04164-1.

- I Made Narsa, A.N. (2021) 'Factors that can be predictors of carbon emissions disclosure', *Jurnal Akuntansi*, 25(1), p. 70. doi:10.24912/ja.v25i1.725.

- Serafeim, G. and Velez Caicedo, G. (2022) 'Machine learning models for prediction of scope 3 carbon emissions', *SSRN Electronic Journal* [Preprint]. doi:10.2139/ssrn.4149874.

- Oestreich, A.M. and Tsiakas, I. (2015) 'Carbon Emissions and Stock Returns: Evidence from the EU emissions trading scheme', *Journal of Banking & Finance*, 58, pp. 294–308. doi:10.1016/j.jbankfin.2015.05.005.

- *Discover how the 500 largest members of race to zero are taking climate action* (2021) *Race to Zero*. Available at: https://racetozerodataexplorer.org/proceed/ (Accessed: 12 June 2023).

- Chen, Y. et al. (2022) 'Does the Carbon Emission Trading Scheme Boost Corporate Environmental and financial performance in China?', *Journal of Cleaner Production*, 368, p. 133151. doi:10.1016/j.jclepro.2022.133151.

- Hendler, A. and Hunter, E. (2022) 'Investing Green to Become More Green: An Analysis of Whether S&P 100 Companies are Decreasing their Carbon Footprint Proportional to their Liquidity', *Senior Honors Papers / Undergraduate Theses*, 44.

- Deniswara, K., Wijaya, F.K. and Handitya, E.J. (2023) 'Determinant of carbon emission disclosure: Empirical studies on energy companies in Indonesia', *Proceedings of the 2023 14th International Conference on E-Education, E-Business, E-Management and E-Learning* [Preprint]. doi:10.1145/3588243.3588283.

- Luo, L., Tang, Q. and Lan, Y. (2013) 'Comparison of propensity for carbon disclosure between developing and developed countries', *Accounting Research Journal*, 26(1), pp. 6–34. doi:10.1108/arj-04-2012-0024.

- Karim, A.E., Albitar, K. and Elmarzouky, M. (2021) 'A novel measure of corporate carbon emission disclosure, the effect of capital expenditures and corporate governance', *Journal of Environmental Management*, 290. doi:10.1016/j.jenvman.2021.112581.

- Esty, D.C. and Porter, M.E. (1998) 'Industrial ecology and competitiveness.', *Journal of Industrial Ecology*, 2(1), pp. 35–43. doi:10.1162/jiec.1998.2.1.35.

- Reinhardt, F.L. (1999) 'Bringing the Environment Down to Earth', *Harvard Business Review*, 77(4), pp. 149–157. PMID: 10539206.

- Walley, N. and Whitehead, B. (1994) 'It's not easy being green', *Harvard Business Review*, 72(3), pp. 46–52.

- Qu, S. and Ma, H. (2022) 'The impact of carbon policy on carbon emissions in various industrial sectors based on a hybrid approach', *Environment, Development and Sustainability* [Preprint]. doi:10.1007/s10668-022-02673-0.

- Ali, M. U., Gong, Z. M., and Yao, C. (2021). 'Fossil energy consumption, economic development, inward FDI impact on CO(2) emissions in Pakistan: Testing EKC hypothesis through ARDL model', *International Journal of Finance & Economics*, 26(3), 3210–3221.

- Fan, D. L., Huang, Y. X., Yong-Jian, P. U., et al. (2017). 'CO2 emission from fossil energy consumption in chongqing and prediction of its peak', *Journal of Southwest University (natural Science Edition)*, 39, 180–185.