

A FINANCIAL COMPARISON OF ENVIRONMENTALLY-FRIENDLY AND NON-FRIENDLY INDUSTRIES

BUAN 6356.001

PROFESSOR ZHE ZHANG

Group 2

JORDAN DICK

YOOHYUN CHEON

MEGDELAWIT WOLDESELASSIE

RABIN SUBEDI

**Index**

Executive Summary…………………………………………………………………………………... 3

Motivation…………………………………………………………………………………………….... 5

Data Description………………………………….…………………………………………………... 7

BI Models………………………………………….…………………………………………………... 11

Conclusion………………………………………..…………………………………………………… 16

Supporting Tables………………………………..…………………………………………………… 20

References……………………………………….……………………………………………………. 37

**Executive Summary**

At a high level, this report aims to determine whether there is a relationship between the environmental-friendliness of certain industries and the financial performance of those industries. More precisely, the goal is to discover any explanatory relationships between industry-specific environmental stewardship data gathered from the EPA on a variety of industries and the return from hand-picked ETFs intended to accurately track the financial performance of those industries. This environmental-friendliness data includes factors denoted as RSEI (Risk Screening Environmental Indicators) and TRI (Toxic Release Inventory). The RSEI Score is further broken down into more sub-components. RSEI, in a general sense, captures the overall significance of polluting activities based on various factors like the level of danger of toxic chemicals and compounds, the amount of pollutants emitted, and the proximity of those emissions to large population centers. The TRI is also a multivariate factor that accounts for the amount of waste emitted into either the water, land, or air by certain industries. Even though all of this data is collected at a granular facility-specific level, only the higher-level, industry-specific data will be necessary for the purposes of this report. Any conclusions about explanatory relationships in this data can only be based on the relationship between the multivariate factors, RSEI Score and TRI data, and the ETFs which track industries associated with the factors, not necessarily the industries themselves. Furthermore, any conclusions about the effectiveness of environmental data in explaining the performance of these ETFs would be inappropriate since not all possibly available environmental data is analyzed. Only the relationship between various ETFs annual returns, RSEI and TRI data can be investigated, and all conclusions can only be relative to that context. Furthermore, since this is an observational study, causal relationships cannot be proven as a matter of fact, only inferred.

Upon completing various forms of supervised machine learning to investigate the relationship between the environmental data and the dependent variable, or annual return, it seemed as though there was no significant relationship between the environmental data collected from the EPA and the annual return of the hand-picked ETFs. Supervised machine learning techniques employed include a fixed-effects linear regression model, a neural network, and a decision tree. The fixed-effects model made theoretical sense, but upon performing iterative f-tests, none of the data proved valuable in explaining the dependent variable. The same was true for both the neural network and the decision tree which both failed to even predict a simple majority of binary outcomes in the validation dataset. Those binary outcomes were based on whether that ETF would or would not have beaten the market index in any given year from 2007-2018. The market index used as a benchmark was SPY. The neural network and the decision tree results were worse than simply flipping a coin since they were less than 50% accurate. The only conclusion that could be drawn from the initial supervised machine learning analysis was simply that no explanatory relationship exists between industry-specific RSEI or TRI factors and the return of ETFs associated with those industries. Had analysis been performed on a more granular level, then it is possible that a more significant relationship could have been found between RSEI or TRI data and the annual return of specific companies. However, it was not feasible to collect this type of data from the EPAs website as the facilities were not always tied to specific companies in the available datasets.

Since this environmental data is collected from the US government, there may be other implications from this analysis pertinent to public policy. It could be indicative of a general failure of tax policies to align financial incentives with environmental ones. Since the government collects this data, it could create a proportional tax system that penalizes companies who are poor stewards of the environment. The availability of the data makes this possible. The large scale financial effects of such a tax policy cannot be understood unless they are implemented, but the data shows that, clearly, there is no such tax policy in place which effectively aligns financial and environmental incentives. The lack of sufficient evidence to support an explanatory relationship between environmental stewardship data and financial performance of high-carbon, high-pollution industries at a national scale serves as evidence in and of itself - incentives are misaligned for these industries to prioritize the environment.

Unsupervised machine learning was then employed to determine if there were hidden relationships in the data. A single linkage hierarchical system was applied to aggregated data by industry. This method was robust against outliers. The findings resulted in multiple clusters where some of the members within those clusters had easily observable, logical similarities between the operations and end markets of those industries. These sub-clusters were regressed together as the relationships between those similar industries was likely to be more information rich. This assumption proved true because multiple significant variables were discovered for two different clusters - Cluster 2 (Electrical Equipment and Hazardous Waste) and Cluster 3 (Machinery and Miscellaneous Manufacturing).

Electrical Equipment and Hazardous Waste both share costs in the disposal of highly toxic, and even deadly, chemicals and compounds which are used in manufacturing of a variety of electronic products, like batteries. The by-products from these manufacturing processes can be extremely expensive to dispose of. Hazardous waste companies typically share these disposal costs with electrical equipment manufacturers, but they also must deal with this type of waste produced by society at large. The relationship between these two industries, and the resulting regression, were therefore unsurprising given this premise. The costs associated with those industries are naturally correlated to the financial performance of those types of companies, which bodes well for the environment. For Cluster 3, the business operations for companies engaged in either the Machinery or the Miscellaneous Manufacturing industries tend to overlap extensively. For these types of companies, the total amount of bulk waste created from their business processes seems to be a relevant factor in explaining annual return. Traditionally, the waste, or by-product, from the business operations of these two industries has been an overlooked expense, but that may be a luxury modern business managers cannot afford. There is a very large variety of wastes generated by these two industries, and an ever-increasing volume. The costs associated with disposal of this myriad of wastes are also increasing per pound year over year faster than inflation. Again, the relationship between TRI modeled pounds and the return of ETFs for both the Machinery and the Miscellaneous Manufacturing Industries makes sense. A more detailed explanation of the similarities in the cost structures of these industries in each cluster can be found in the BI Model section of this report. Additionally, each of these different clusters carry different implications from their respective regression results, and these implications are also discussed in further detail in the BI Model section of this report.

**Project Motivation/Background**

For our project, we had different options to choose from. Fields such as healthcare, technology, environment pollution and finance are relevant and interesting subjects to base our project on. However, our group wanted to work on a topic for a field that is relevant in today’s society but normally neglected by scholars to research on. Thus, we came upon the topic of environment, social and corporate governance (ESG). ESG is a measure of performance of a publicly traded company on standards that are other-than profit and business. As we know, the main goal of any business is to make profit and provide the best value to its investors whether they are private investors (such as in a private company) or shareholders (in public companies). However, as people have become wiser over the years, they have gradually begun to understand that a public company has additional responsibilities other than providing the best return to its shareholders. More specifically, a public company should be environmentally responsible and perform its business operations in such a way that it causes the least possible harm and depletion to the natural environment and its resources. Similarly, a public company should respect and honor the different social norms and values that are practised in whatever natural regions, country or state it operates in. Furthermore, a public company should be alert, ardent and resolute in discovering and removing any kind of improper corporate governance practices within its entities, thus providing more transparency and trust to its shareholders and the general public. This is the notion behind ESG.

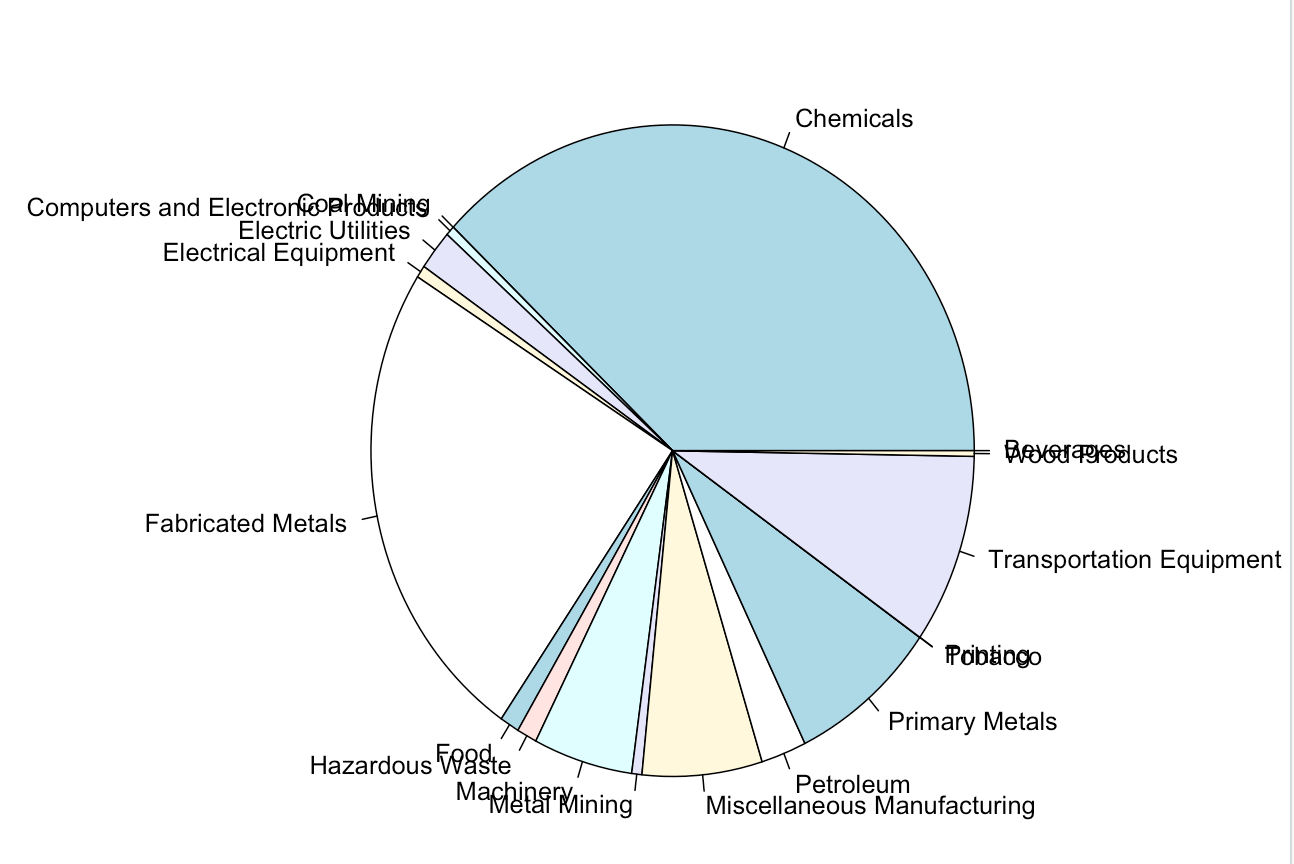
It is not uncommon, nowadays, to notice that public companies spend a great deal of time, capital and effort to improve their ESG standards. In its annual report, every public company explains in detail what different steps it took in the past year to become more environmentally responsible, to respect the social norms and to practise healthy corporate governance. Thus, we see that public companies care about how the general public perceives them in terms of these ESG standards.

After understanding the growing popularity of ESG measures for public companies and the inclusion of these measures on the popular financial websites such as YahooFinance and quarterly reports of the public companies themselves, we decided that working on ESG measures of public companies would make up an interesting and useful project for us. To make the project more useful, we tried to relate the financial and profit aspect of a public company with its ESG aspect. More specifically, we tried to find whether a public company with good ESG measures performs well financially or not. However, we soon realized that ESG is a broad topic and it is difficult and not feasible for us to work on all the aspects of ESG. This is because a company could be socially responsible but have unhealthy corporate governance practises going on. Similarly, a company could be less harmful to the environment but may not be socially active and care less about the society it operates in. Therefore, we decided to limit our project to only the environmental aspect of the ESG. That is to say, we will compare different companies solely on how much or how less they pollute the natural environment. This decision was also based on the fact that high-quality data in the environmental category was more easily available compared to social and governance data. Finally, we felt as though one of the most important issues for modern civilization is to maintain a sustainable footprint on Earth both for ourselves, and for future generations.

We quickly realized that to compare a company from one industry with another company from a totally different industry would be an unfair comparison and will not yield useful insight because, by nature, some industries are heavy polluters of the environment and some industries are less polluters. A company such as a chemical plant will naturally cause more pollution and harm to the natural environment than a company such as Google which is heavily technology-oriented and mostly operaties in the digital dimension. The difference in their contribution of pollution to the environment is not caused by their company-specific characteristics; rather, it is caused by the different industries they operate in. Thus, to make more sensible comparisons, we decided to compare industries themselves, rather than individual companies. We also decided to only compare industries which contribute the most to pollution relative to the total amount of pollution emitted. Industries whose overall contribution to pollution is a miniscule and insignificant proportion of the entire amount of pollution being added to our water, land and air every year are unlikely to have financial performance tied to their degree of pollution in any statistically significant way. So, only industries which contribute significant proportions of overall pollution were analyzed.

After comparing different industries on how heavily or lightly they pollute the environment, we will then compare the financial aspects of these industries such as return on investment. Since we are comparing industries and not individual companies, we decided to take Exchange Traded Funds (ETFs) from different industries and compare them. An ETF is a basket of securities (stocks, bonds and other financial instruments) from different holding companies that are held and tracked by that specific ETF. Thus, it is a form of an index that tracks the performance of the different holding companies that fall under that ETF. It is important, however, that we select only those ETFs from industries that best represent that industry alone and not other industries. In other words, we will select an ETF for an industry if a large percentage of that ETF contains the securities from holding companies from that industry alone. After we have selected ETFs for different industries, we will then compare these ETFs and test our hypothesis that industries which pollute the environment less are generally more lucrative to invest. Finally, we will form our conclusion and provide any additional insight we will obtain during the course of research.

**Data Description**

The selection process behind determining which industries to compare began with determining which industries proportionally contribute the most to overall RSEI pollution criteria. As you can see in the below pie chart, all of the industries that were selected for analysis constituted a significant proportion of the overall RSEI Scores from all industries that report to the EPA. After determining which industries to analyze, ETFs were hand-picked to track each industry, and the returns of those ETFs were used as our output variables. So, annual return from the ETFs is the dependent variable.

Exchange Traded Funds, commonly known as ETFs, is a type of investment that is similar to a mutual fund. ETFs are pooled investment vehicles that offer diversified exposure to a particular area of the market. It allows investors to receive interest from the pooled investment they contributed to after the funds are used to invest in stocks, bonds, or other assets.

The data we collected is focused on specific industries. After we narrowed down our research to selected industries, we picked ETFs for each industry that we believed best represented that industry. The list of the industries and their corresponding ETFs we selected include: chemical (FSCHX), metal fabrication (PICK), mining (XME), primary metals (JJMTF), transportation Equipment (XTN), manufacturing (FIDU), utilities (XLU), petroleum (XOP), machinery (XLI), computers and electronic equipment (XLK), electrical equipment (VPU), miscellaneous manufacturing (VIS), food & beverage (PBJ), wood products (CUT), coal mining (KOL), tobacco (BTI), printing (IP), hazardous waste (EVX).

SPY is a very well-known and recognized ETF in the US. It is also one of the oldest and top ranking for largest assets under management (AUM) and greatest trading volume. It is seen as the benchmark for trillions in dollars of investment. SPY was used as the overall market benchmark in this report. FSCHX, which stands for the Fidelity Select Chemicals Portfolio, is one of the highest rated Fidelity funds. FSCHX is known for its best lifetime annualized returns. Since its inception in 1985, its annualized return has averaged above 13 percent.

For the mining industry, we used the XME ETF. We picked XME because it is the only fund that tracks the US metals and mining segment. It gives equal weight to the different firms it holds which can be a little problematic since this could lead to underweight of integrated mining firms and higher volatility. However, since it tracks a well-known, annually rebalancing index with highly liquid securities, it is considered an important representative ETF.

JJMTF is used to track the primary metals industry. It is designed to focus more heavily on the medal processing industry than on the mining industry, but it is not necessarily a pure play on primary metals. Unfortunately, some ETFs are not perfectly indexed representations of the underlying market, but are sufficient for analysis, and are the best that we could find.

XTN is the best ETF we could find for the transportation equipment manufacturing industry. It is known for its efficient fund, index tracking, and provides a best solution for the industry's high concentration by equal weighting its holdings.

Fidelity MSCI industrial index ETF (FIDU) is an ETF, without the fees and expenses, seeks to provide investment returns that directly relate to the general performance of the industrial index. At minimum, 80% of assets are invested in securities included in the fund's underlying index.

XLU is among the cheapest and most liquid options available. XLU is a massive, highly liquid fund that only invests in utilities companies that are included in the S&P 500. This ETF consists of select few companies that do not fully represent the market capitalization. However, the companies held by this ETF are massive and as such, they have big assets and volume. These features make XLU a popular choice among ETFs that deal with utilities companies.

XLK is an ETF that is concentrated on the technology sector. This ETF avoids small size and mid-size companies which makes it more stable and less volatile. It includes securities from only a few selected technology companies.

PBJ is an ETF designed to solely track the food and beverage industries. No other ETF that tracked the food and beverage industries had existed as long as PBJ or was as concentrated as PBJ was in just those industries. We felt that not only did PBJ do a superior job of only tracking those industries, but it also had a significant amount of historical data for the entire period of interest - 2007-2018. PBJ seemed like the best ETF for both industries, so it was used for. Other ETFs that tracked just food were more commodity focused instead of focused on various companies' performance and food processing/manufacturing. Also, no ETF could be found that only tracked the beverages industry. So, PBJ was used for both the food and the beverage industry, which the EPA categorized into two, separate industries.

CUT is an ETF that comprises the 25 largest forest and timberland firms around the world. However, the loose definitions and criteria for a forest ETF allows it to spread out into other industries that would normally not be considered as forest and timberland industry. It has also greater geographic inclusion than usual ETFs. Nevertheless, the inclusion of the 25 largest forestry firms makes it a rich representation of forest and timberland industry.

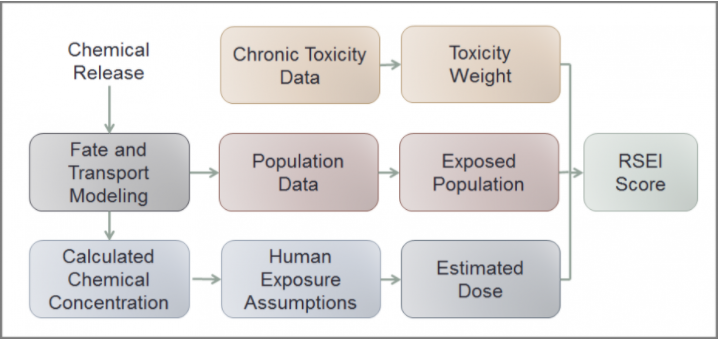
KOL is an ETF that is based on coal mining, transportation, equipment and trade. However, a large percentage of revenues for this ETF comes from companies that are not purely based on coal mining but rather on transportation and coal equipment. Although it has broader exposure, the total holding companies within it is few. This ETF uses modified market capitalization to give weights to securities from different holdings. Nevertheless, it is still one of the best representations for the global coal industry.

ACT is an ETF that is based on companies that derive a large portion of their revenues from tobacco, alcoholic beverages and marijuana. This ETF is relatively new in the market but has become quickly popular because of the heavy concentration of its portfolio on products that are normally considered vices. Boston Beer and Constellation Brands are two of its popular holdings. Upon further analysis, ACT has only existed for a brief amount of time, so there was insufficient historical data to render it useful in our analysis. This was also the only tobacco ETF we could find, and it was not a pure play tobacco industry index either. So, instead of using this ETF, we decided to select BTI which is the British American Tobacco Company. This company is one of the largest tobacco companies on earth, and it has many facilities in the US. It is a very good representation of the overall tobacco industry, especially compared to ACT.

IP (International Paper Co.) is a big company that produces renewable fiber based paper and pulps. However, this is not an ETF, but a company itself that is owned by many other ETFs. The reason we chose a company instead of an ETF for the paper industry is because there is no such ETF that can represent the paper industry. This company operates in North America, Latin America, Europe, North Africa, India and Russia. Its main segments are Industrial Packaging, Global Cellulose Fibers and Printing Papers.

EVX is an ETF that is heavily concentrated on firms that deal with waste management. It has 3 different tiers and within each tier, the companies are given equal weights. In the first tier, the big four companies receive 10% weight of the total portfolio. The 5 smallest companies receive 2% weight of the portfolio and the remaining 50% weight is distributed among the remaining firms. However, this ETF is concentrated on US securities which differs it from the market as the market has broader global coverage and inclusion.

Understanding RSEI Score results is very crucial in this paper. The Risk-Screening Environmental Indicators (RSEI) is a model that provides us with data on releases of toxic substances from industrial facilities. RSEI scores do not have units of measurement and, due to this reason, the scores only have meaning in comparison to other RSEI Scores. The scores are assigned values based on the size of the chemical they release, the toxicity of the chemical, where that chemical ends up, and the people it affects. To calculate the RSEI Score, we take the toxicity weight multiplied by the exposed population multiplied by the estimated dose.



To rank the industries from high RSEI score to low RSEI score we need to compare them with each other. If we observe the RESI score 2 times higher than another, we can say the potential for risk is 2 times higher. Toxicity weight plays a big role in the RSEI score, as the toxicity level increases so does the RSEI score even if the release of this chemical is low. The estimated exposure to population is also a key factor in determining the RSEI score. The higher the population exposure the higher the RSEI score, however, the effect on human health needs to be determined.

On the other hand, a low RSEI Score indicates there is less environmental concern but we need to be aware of the environmental risks we have not considered and the different pollution caused by the facilities which they fail to report. Due to the different aspects of RSEI score, we need to look at the RSEI score and further investigate. RSEI score does not explicitly tell us the exact level of risk nor help us draw conclusions.

In order to gain a good understanding of the RSEI scores we need to look at the different types of RSEI score. In this paper we further collected data of the RSEI Modeled Hazardous, RSEI Modeled Pounds,TRI Pounds, RSEI Score-Cancer and the RSEI Score-Noncancer. RSEI Hazard unlike RSEI Score does not include freight and transport modeling or adjustments for population exposure. It tells us the size of the release and the chemical’s toxicity or also referred to as toxicity-weighted pounds. When we calculate the RSEI Hazard over the set of modeled releases we refer to it as RSEI Modeled Hazard because we are not considering all the possible releases. RSEI Modeled Pounds refers to the pounds for those releases and transfers.

RSEI Score-Cancer and the RSEI Score-Noncancer allow us to clearly see the RSEI score that has a direct impact on humans such as cancer. The RSEI Score cancer indicated the toxic chemicals released are directly linked or are identified as cancer-causing toxins. The Toxics Release Inventory (TRI) is used to track the industrial management of toxic chemicals that may cause harm to human health and the environment. TRI data are reported by certain industrial and federal facilities.

RSEI Hazard is often higher than RSEI Modeled Hazard for the same facility because more releases and transfers are included. RSEI Hazard can be calculated for any TRI release or transfer. Generally, cases where RSEI Score is relatively low and RSEI Modeled Hazard is relatively high indicate that something in the environmental modeling is mitigating the impact of the release on human health—for instance, a small exposed population around a facility in a remote area, a low calculated concentration due to a fast rate of chemical decay, or treatment in an off-site incinerator or water treatment plant.

**BI Models**

Multiple models were employed to assess whether the environmental data gathered from the EPA would be effective in explaining the annual returns of the given industries. Each of those models will be defined, but the more significant findings will receive more detailed explanations. Those models will be mentioned in the order in which they were developed, and the corresponding analysis will be enumerated. The BI models utilized in this project include: Fixed Effects Linear Regression, Neural Network, Decision Tree, and Sub-Regressions of Correlated Industries. The Sub-Regressions of Correlated Industries in the only method employed which combines both supervised and unsupervised analysis, and it also produces the most useful results.

**Fixed Effects Linear Regression**

The first model was a fixed effects linear regression which defines dummy, or indicator, variables for each industry. The theory behind this model was that each industry has many different, unique factors affecting it which can be best explained by giving every industry its own intercept using indicator variables. Some of those unique, fixed effects for a given industry may be a result from lobbying groups for that industry, the reputation of that industry, the managerial culture common to that industry, regulations unique to that industry, and much more. Most of these factors are not available for us to analyze using a BI model. Therefore, fixed effects models are useful as they account for outside factors which have constant effects. When processing and preparing the data for analysis, an important consideration in creating the indicator variables was that one of the indicator variables had to be removed before performing the regression. The purpose of this is to avoid exact collinearity in our data. To avoid exact collinearity (and ruining the model) the indicator variable for Beverages was removed. Therefore, beverages will serve as the base indicator variable, and its intercept will simply be the intercept for the model, with no added fixed effects. Normalized data was not used in the fixed effects regression model because normalized data is not necessary in linear regressions.

The results from the fixed effects regression can be seen in Table 1.1. Per this regression, the only significant environmental variable at a 90% significance level was TRI Pounds. Also, only some of the fixed effects indicator variables were significant at a 90% level of significance. This initial model is insufficient for multiple reasons. Since it has a very small R-squared value, the model fails to fit the data well. Also, a negative Adjusted R-squared indicates that not only does the model do a poor job of fitting the data, but the variables actually detract from its predictive value. Additionally, an F-test performed on all the variables in the full model only yields an f-statistic of 0.39, which means that none of the variables in the model have predictive value. Clearly, this model is not sufficient, so other models were devised in order to find valuable relationships in the data. However, before trying different models, some investigation was performed to determine why the results are insignificant, and if there may be some insight to extract from the data using this fixed effects regression method.

The reduced model only contains the environmental variables, not the fixed effects indicator variables. The reduced model’s regression results can be seen in Table 1.2. Another f-test was performed comparing the full model to the reduced model which removed all of the fixed effects variables per Table 1.3. This f-test illustrated that none of the fixed effects variables effectively explain the dependent variable - annual return. Another F-test, Table 1.5, was performed against all of the environmental variables where only the indicator variables were left in the reduced model. The regression results from this second reduced model can be seen in Table 1.4. The results were similar - none of the variables seem to be significant predictors of annual return. So, the current conclusion is that the environmental data gathered from the EPAs website on RSEI and TRI has no significant predictive value on the return of the ETFs that were hand-picked to track these various industries. However, this is not to say that no environmental data can predict returns, but the data gathered on RSEI and TRI cannot predict returns for the selected ETFs in this analysis.

**Neural Network**

With the Neural Network, the first step was to separate the data into both a training and a validation dataset. Since the environmental data is relatively limited, 60% of the entire dataset was used for training and 40% for validation. The data was randomly selected, and not all industries may be equally represented in both datasets. This should not have serious implications on our results, however, since there are a sufficient number of records in both datasets. With a sufficient number of records, there is enough information, or degrees of freedom, to extract a significant amount of information from the data. The same datasets were also used for training and testing/validating the decision tree.

In spite of the fact that neural networks are a black box, they can sometimes have almost as much, or more, predictive value than other BI Models like linear or logistic regressions. The Neural network in Table 2.1 was designed to only have a single, boolean output, while the inputs were all of the environmental variables from the data. To construct the boolean result variable column, a sequential vector was created which compared the annual return of each ETF with the corresponding annual return of the market. If the ETF returned more that year than the market, then it was assigned a 1, and otherwise it was assigned a 0. So, this neural network is designed to predict whether a certain industry, or ETF, will or will not beat the market based on environmental data.

Unfortunately, the neural network created to describe the dataset containing RSEI and TRI environmental factors fell short of being useful. The actual diagram of the neural network can be seen in Table 2.1. The confusion matrix of the neural network can be seen in Table 2.2. Since the accuracy rate of this neural network is only 43.7%, it would be better if no neural network were used at all. With an accuracy rate below 50%, this model would actually lose a company money if they made decisions based off of it. It would be better to flip a coin to determine as to whether the underlying ETF would beat the market. So, this neural network, based on the environmental input variables, has no explanatory value for the output, or whether or not the underlying ETFs will beat the market.

**Decision Tree**

The same training and validation data that was employed for the neural network was also used for the Decision Tree, and similar results about the effectiveness of the model were retrieved. The decision tree is aiming to predict the same thing as the neural network - can factors about the environmental friendliness of certain industries be used to accurately determine whether or not certain ETF that track those same industries will beat the market? The general purpose of applying this method after obtaining insignificant results from the neural network was to determine if there were any simple information gaining partitions of data that could enable some basic decisions about the data. The initial decision tree is very large, and is not useful for predictions, but it can be viewed in Table 3.1. The accuracy of the initial tree using the testing, or validation, dataset is only 44.8%, which is insignificantly better than the neural network. The confusion matrix used to calculate the accuracy of the initial tree is in Table 3.2.

The decision tree was pruned, per Table 3.3, and is more easily interpreted, but results are still insignificant. Although the pruned tree has only 7 partitions, compared to 44 for the original tree, it still has 45.9% accuracy according to the confusion matrix in Table 3.4. Even though the decision tree was decreased to 1/6th the size, it actually gained in interpretive value! So, the pruned tree is a far better tool, but only relative to the original tree. Neither provide significant predictive value. In conclusion, the decision trees failed to provide valuable insight into whether the annual return of the ETFs that represent their underlying industries will or will not beat the return of the market based on the EPAs environmental factors regarding those same industries.

**Hierarchical Clustering and Regression Analysis**

Given that the data as a whole has little predictive value, an unsupervised machine learning technique was employed to assess whether some valuable information was being obscured when analyzed all together. Since there are only 18 different industries being analyzed, Hierarchical Clustering is most appropriate. In order to prepare the data for analysis using this unsupervised method, the data had to be normalized so that everything was on the same scale. This was important because some naturally larger measurements may obscure smaller ones but may not necessarily be more information rich. Next, to create these clusters, all of the environmental data was aggregated together by industry. Each variable was summed up by industry so that each record reflected the total RSEI and TRI emissions for that industry from 2007-2018 as well as the total return for that given industry over that same period. The resulting data frame only had 18 records with the total returns and total volume of emissions from each industry. This data was then further processed and prepared for clustering by removing the return column so that only the environmental factors are considered when creating the clusters.

After normalization and aggregation, a distance matrix was created from which the different industries could be compared, and clustered, based on minimizing the single, or minimum, distance. Since the single linkage method is sensitive to noise, or outliers, the results from clustering using the single linkage method were compared against the results for the average linkage method, and results were nearly the same for the purposes of this analysis. So, our data is robust against outliers when performing hierarchical cluster analysis when the data is normalized, and, therefore, using the single linkage method will yield useful results. To compare the Hierarchical clusters using the single linkage and average linkage methods, see Table 4.1 and 4.2 below, respectively. Additionally, the ‘sqldf’ package was used in R to aggregate the data together grouping it all by industry. The aggregated data was simply summed together. The theory behind this was that there could be a relationship between the total amount of pollution any given industry has emitted over a certain time period and the total return from investments in that industry. This would seem logical if financial and environmental incentives were effectively aligned. Based on this premise, the sum total for each variable relative to each industry was calculated over the period 2007-2018, and then this data was used to perform unsupervised clustering analysis.

After obtaining the clusters from the unsupervised hierarchical clustering single linkage technique, the heatmap, Table 4.3, can be utilized to explain why those industries were grouped together the way they were. This heatmap will help provide insight into why certain relationships were made and will have implications on the interpretation of the results from the regressions of the clusters of industries. The hierarchical cluster was rather arbitrarily cut at two heights, 5 and 10. Then, each cluster was regressed as a standalone subset of data. The results were both intriguing and conceptually difficult to interpret. First, at the cut height of 5, three clusters were of interest that were regressed as their own, standalone sub-clusters:

1) Printing, Beverages, Coal Mining, Tobacco, Computers and Electronics, Wood Products (Table 4.4);

2) Electrical Equipment, Hazardous Waste (Table 4.5);

3) Machinery, Miscellaneous Manufacturing (Table 4.6)

According to the heatmap (Table 4.3), Cluster 1 Industries are all grouped together because of the similar amount of RSEI Model Pounds that they emit, Cluster 2 Industries are grouped together because of similar TRI emissions, and Cluster 3 Industries are grouped together because of similar RSEI and Carcinogen emissions. Regressions were performed using both the normalized data and the non-Transformed data, but the significance of results did not vary regardless. Therefore, the sub-cluster regressions utilize non-transformed data instead of normalized data for enhanced interpretability of results.

The regression results for the first sub-cluster regression was insignificant (Table 4.4). The f-test for the first regression result was not indicative of any predictive value in the environmental data for the ETFs of the industries in that sub-cluster. However, interestingly, the regression results for Cluster 2 and Cluster 3 did yield significant variables. Specifically, the sub-cluster regression for the Electrical Equipment and Hazardous Waste industries (Table 4.5) Indicated that the RSEI Score for Noncancerous emissions was a significant indicator of the financial performance of the ETFs that represented those two industries - VPU (Electrical Equipment) and Hazardous Waste (EVX). Additionally, the F-test for this regression was relatively strong (.114) compared to all of the other F-statistics obtained from all of the other analytical methods employed on this dataset. Furthermore, a 95% confidence interval for this RSEI Score for Non-Cancerous emissions yields an interval of -3.46475386 and -0.3929518, which are both below zero. This means that this variable can be reliably understood to be negative. In conclusion for the sub-cluster regression for Cluster 2, as the RSEI Score for non-cancerous emissions increases by 1, the annual return of VPU (Electric Equipment) and EVX (Hazardous Waste) decreases by -1.895e-06%. This relationship between these two seemingly unrelated industries is likely due to the fact that the Electric Equipment Industry has to dispose of a significant amount of highly toxic chemicals, metals, and other by-products from semiconductor, analog chip, wire, transformers, and other electrically conductive materials. It can be very expensive to dispose of or recycle (in few cases) the by-products from Electrical Equipment Manufacturing. The more of this material an Electrical Equipment manufacturer must dispose of, the more costs they will incur. So, this relationship between Non-Cancerous RSEI Scores and financial performance makes sense for this industry. Similarly, the Hazardous Waste Disposal Industry typically signs contracts with developers of local governments to collect and dispose of garbage and waste of all kinds. These contracts are less valuable when the Waste Disposal industry must dispose of extremely hazardous materials that must be handled in specialized, complex ways. So, again, it makes sense that the financial performance of the Hazardous Waste Disposal industry, represented by EVX, has a negative relationship with the Non-Cancerous RSEI Score that measures the amount of frequently specialized, hazardous materials that are costly to dispose, and incurring a major fine by the EPA is not a valid alternative for gross negligence.

Cluster 3 also yielded significant results (Table 4.6). Both the Machinery and the Miscellaneous Manufacturing Industries, represented by the ETFs XLI and VIS, respectively, tend to perform worse in the stock market when the amount of RSEI Modeled Pounds increases. Specifically, as the factor RSEI Modeled Pounds increases 1 pound, the return from investments in these industries, or really in their corresponding ETFs (XLI and VIS), will decrease by -9.559e-08%. The effects from this result are significant even at a 95% confidence level, just like the result for Non-Cancerous RSEI Scores for the regression of Cluster 2. A 95% confidence interval for the RSEI Modeled Pounds variable from the regression of Cluster 3 is between -1.779244e-07 and -1.326146e-08, which indicates this factor is reliably below zero. This has many implications on both the Miscellaneous Manufacturing and the Machinery Industries, or at least all of the companies involved in those activities who are represented by the ETFs XLI and VIS.

Practically speaking, many Miscellaneous Manufacturing companies use a tremendous amount of machinery in their business operations. Both industries, Miscellaneous Manufacturing and Machinery, are closely aligned, and frequently overlap. As the trend of vertical integration in the US business landscape has accelerated, this line has been further blurred. So, since these two industries are involved in most of the same manufacturing and construction processes, these two industries tend to perform very similarly. That said, one of the most overlooked costs for these industries is that of waste disposal. Every year, waste disposal costs increase for a variety of reasons, some not included. First, the number of available, or legal, disposal venues is constantly being restricted due to there being more and more waste every year. Also, waste disposal is constantly being more regulated. Not only is solid waste in land and water being more regulated, but so is air waste. RSEI Modeled Pounds measures the weight of all of the waste and is definitely correlated to the cost of disposal of that waste. Waste Disposal Companies will typically charge more for contracts when they must dispose of heavier, higher volume waste. Every manufacturing process, and most construction processes, utilize heavy and light machinery which produce a considerable volume and variety of waste, or by-products (See Table 5). Oftentimes, these costs are overlooked in manufacturing and machining processes, but good managers may not want to make that mistake anymore. Not only is reducing waste beneficial to the environment, but it may be a significant avenue for cost reduction if companies can find new ways to reduce that waste, or to economically reuse or recycle it. Per the results in the sub-regression of Cluster 3, the total amount of weight of waste that must be disposed of every year by the Miscellaneous Manufacturing and the Machinery Industries can significantly describe returns.

**Managerial Implications and Conclusion**

A general lack of correlation between environmental stewardship factors and the return of ETFs associated with those industries may imply that the regulatory environment does a poor job of aligning environmental stewardship with fiscal incentives. Although one-time fines are imposed on companies that are grossly negligent in a single case, there is not currently a proportional tax on polluters or on carbon emitters, and that is most likely why there is not a significant relationship between financial performance of these polluting industries, measured by return of ETFs that track those industries, and the degree to which those same industries pollute, per RSEI and TRI emission data from the EPA. Since larger volumes of pollution for certain industries does not entail worse performance, that can serve as evidence that current tax policies enforced by the government do a poor job of aligning tax incentives with environmental ones. A transition to clean operations for any given industry may require such an alignment. If the government enforced a laddered tax policy which forced polluters to pay more in taxes, then there would likely be a relationship between the performance of some of these ETFs and the amount of polluting chemicals those industries emit. The fact that no significant relationship exists between the financial performance of the ETFs that track all 18 of these industries and the amount of pollution that these industries emit is evidence that current tax incentives for reducing pollution are insufficient for bolstering better environmental stewardship. In other words, the lack of the existence of a relationship between the industry-specific environmental stewardship data and financial performance return data for these industries indicates that financial incentives are not sufficiently aligned with environmental stewardship incentives. While this assertion may be logical, it is speculative and cannot be concluded as a matter of fact from the analysis on this data given that the scope of this data does not provide sufficient insight into environmental tax law implications. The only conclusion that can be drawn from this data is that RSEI and TRI factors collected by the Environmental Protection Agency about various industries cannot accurately determine the financial performance of the ETFs that are closely associated with the same industries. On a national scale, there is no significant predictive relationship between the returns of ETFs that track various industries and the industry specific RSEI or TRI data associated with those exact same industries.

Upon further analysis, more granular conclusions and insights can be drawn about a handful of related industries. Those industries include those mentioned in Cluster 2 and Cluster 3 in the BI Model Analysis Section. These clusters were found using an unsupervised method of machine learning called single linkage hierarchical clustering. Each sub-cluster is defined for different similar characteristics that makes those industries similar in their environmental impact, and, in some cases, related in their business operations, too. Interestingly, the environmental impact data collected by the EPA can be insightful in understanding the nature of the business operations that various industries engage in.

In Cluster 2, both the Electrical Equipment and the Hazardous Waste Industries bear the costs of disposal of many hazardous and toxic materials from the manufacture of all kinds of electronics and electrical equipment. This equipment is known for being very toxic because it uses a variety of chemicals and metals for products like batteries, semiconductors, electrical wire, transformers, cell phones, and so much more that must be disposed of in extremely specific, expensive ways. Furthermore, the by-products of these processes are expensive to dispose of. Electrical Equipment manufacturers pay Hazardous Waste Disposal companies to dispose of this toxic waste via contracts. Sometimes, those Hazardous Waste Disposal companies lose money on those contracts because of how difficult it can be to dispose of some of these toxic chemicals and metals like lithium, mercury, lead, and many others. Since both the Electrical Equipment Industry and the Hazardous Material industry share these costs, their businesses can be materially affected if the volume of waste they must dispose of increases. It is likely that because of this relationship, these two industries were clustered together in the Hierarchical Clustering Analysis in Tables 4.1 and 4.3. The similarity of these industries entails that the regression of the sub-cluster of these two industries could provide insight into whether the effects of disposal of hazardous waste in the Electrical Equipment Industry and the Hazardous Waste industry affects the financial performance of ETFs which closely track those industries. The results were positive. The regression of the 24 records consisting of the annual returns of the ETFs representing Electrical Equipment and Hazardous Waste, VPU and EVX, respectively, and the RSEI and TRI factors from the EPAs website on those same industries yielded a significant factor - RSEI Score Non-Cancer. The annual return of both of the ETFs which represent the Electrical Equipment and the Hazardous Waste Industries (VPU and EVX) is significantly described by the RSEI Score of Non-Cancerous Emissions from those industries in that while their RSEI level of Non-Cancerous emissions rise by 1 unit, their annual return decreases by -1.9289% from the year 2007-2018 (12 year period). These results are reliable at a 95% significance level. What’s most important from these results is that the relationship is significantly negative.

Another granular conclusion to be drawn from this analysis is from Cluster 3 which included both Machinery and Miscellaneous manufacturing. According to the Heatmap in Table 4.3, the Machinery and Miscellaneous Manufacturing industries are very similar in their RSEI and RSEI-Cancer Causing Scores, which is why they are clustered together and regressed as a standalone dataset per Table 4.6. This relationship is likely due to the fact that both of these industries overlap in many areas pertinent to the RSEI Score like the type of wastes, the amount of those wastes emitted, and the proximity of those emissions to large population centers. Not only are the actual business operations of these businesses similar, but so are the end-markets they rely on. Furthermore, many of the Miscellaneous Manufacturers use the same equipment employed by the Machinery Industry. So, again, there is a tremendous amount of overlap in these industries. So, it makes sense that they would be clustered together. Upon performing the regression, the results for one of the variables were significant at a 95% significance level. The return for the XLI and VIS ETFs from the year 2007-2018 can be significantly explained by the RSEI Modeled Pounds variable at a 95% significance level. For each additional modeled pound for the Machinery and Miscellaneous Manufacturing Industries, the return for these ETFs decreases by -9.559e-08%.

It is also important to note that the above conclusions may not necessarily be causal. It is possible that the relationships discovered are simply corollary. For example, it is possible that the value of the underlying ETFs increase due to widespread efficient and effective capital allocation as well as effective company management, but that the amount of emissions decreases for multiple, unrelated reasons. Emissions could have been reduced overall because of exceptional company leadership decisions while similar decisions resulted in better stock performance on an industry-wide scale. This could account for the negative relationship between aforementioned, industry-specific environmental stewardship data and returns from the associated ETFs. Although Ultimately, a causal relationship cannot be concluded from this data since this is an observational study, not a randomized, controlled experiment. It would be nearly impossible to conduct a randomized, controlled experiment at such a grand scale. However, one thing is clear, in the absence of widespread, comprehensive, and proportional environmental tax policies that align both operational and financial incentives with environmental stewardship incentives, there is unlikely to ever be a significant relationship between financial performance and level of pollution for any industry.

**Tables and Diagrams**

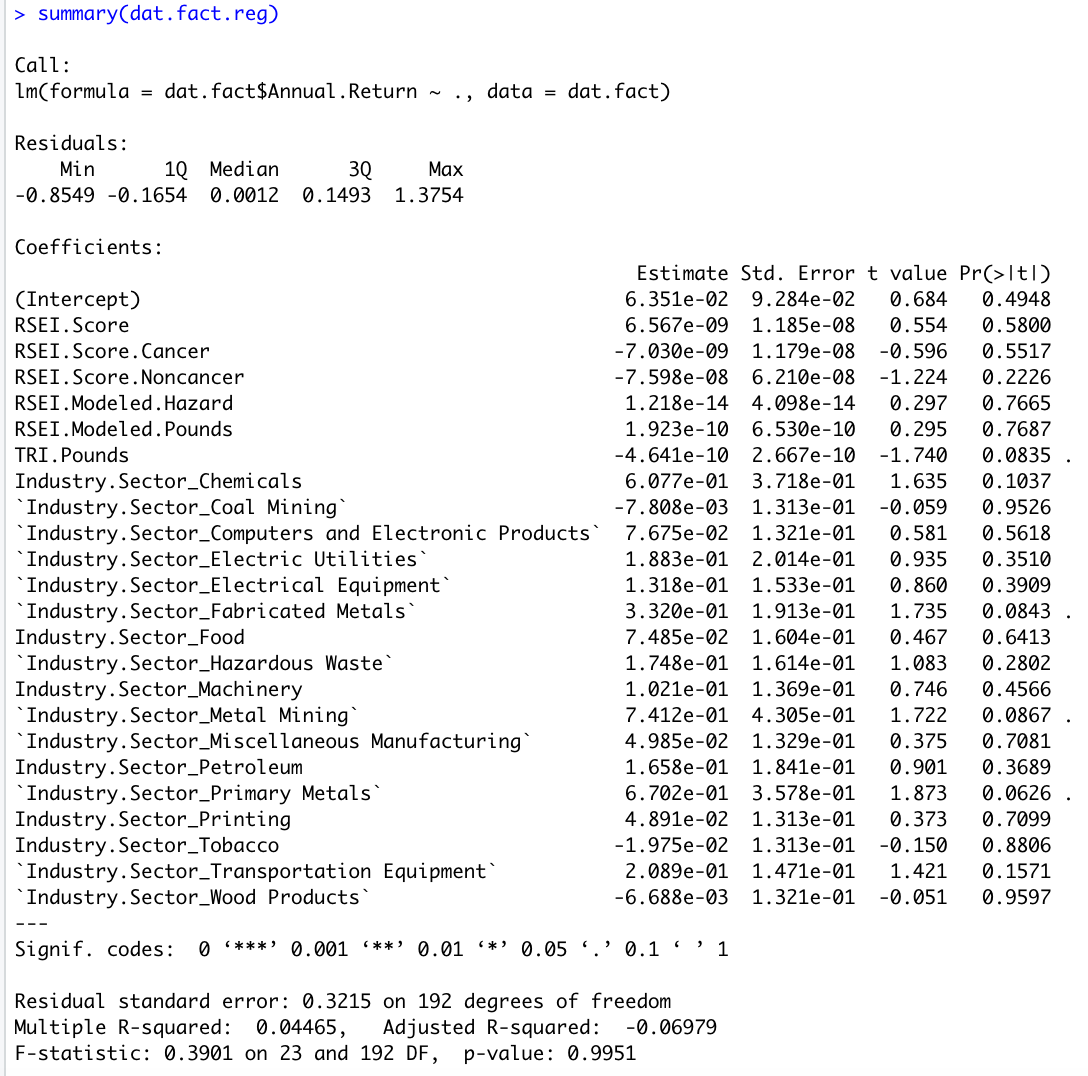
Table 1.1 - Fixed Effects Model: Full Model

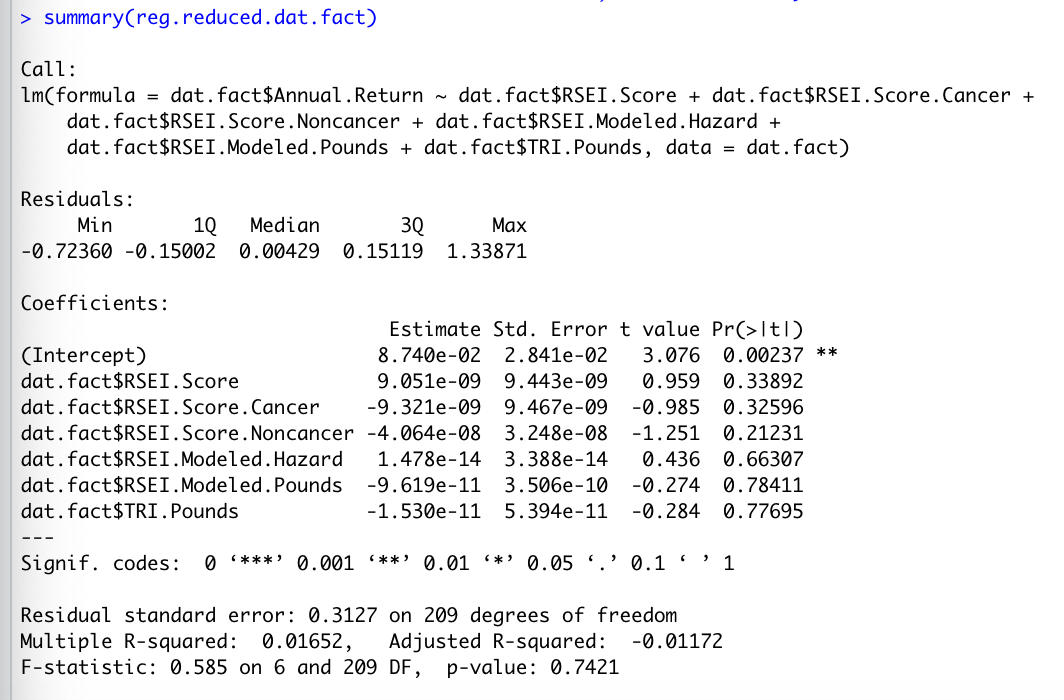
Table 1.2 - Reduced Model With Environmental Variables

Table 1.3 - ANOVA or F-Test of Model in Table 1.2 against the Full Model

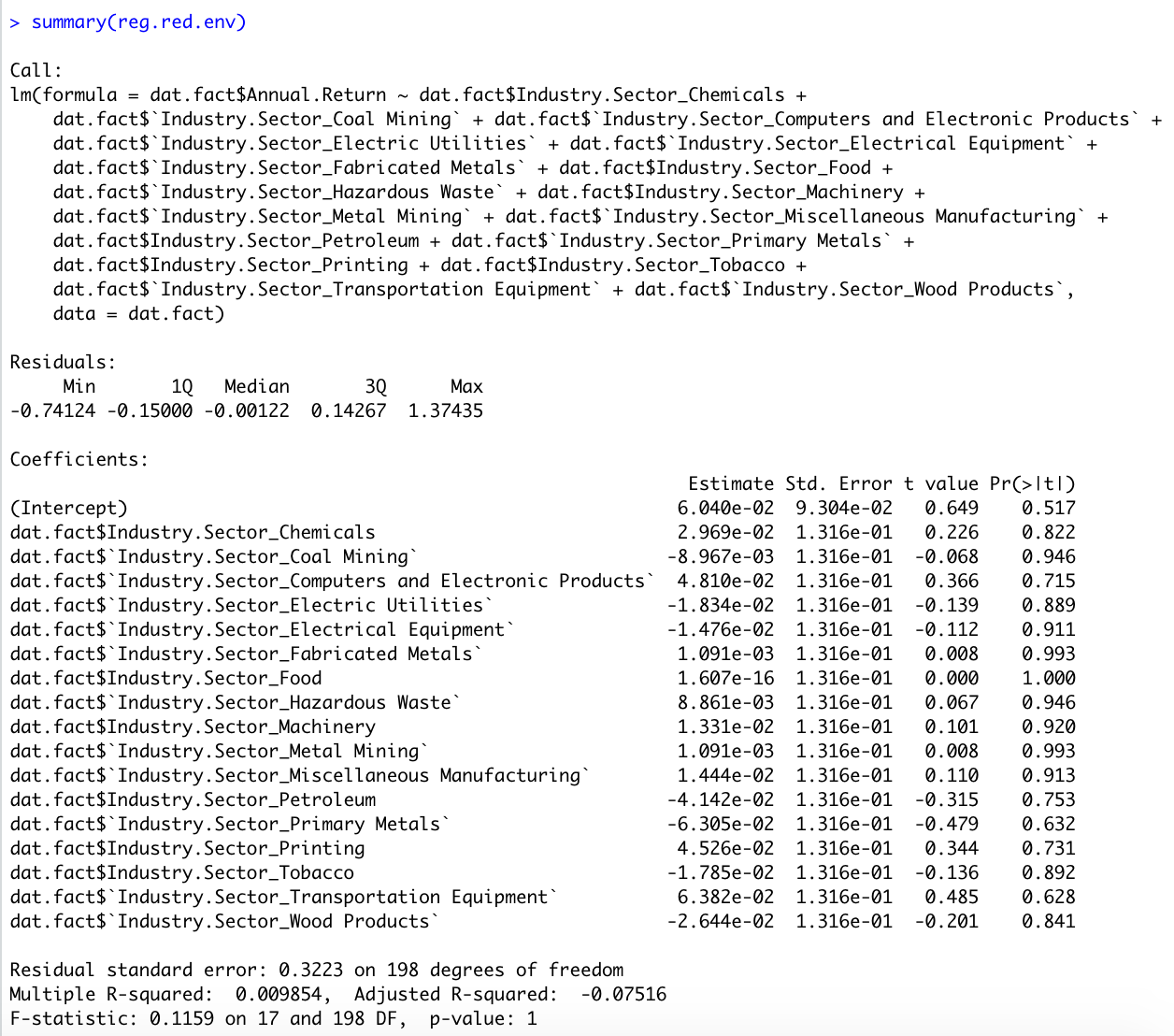
Table 1.4 - Reduced Model With Fixed Effects Variables

Table 1.5 - ANOVA or F-Test of the Model in Table 1.4 compared to the Full Model

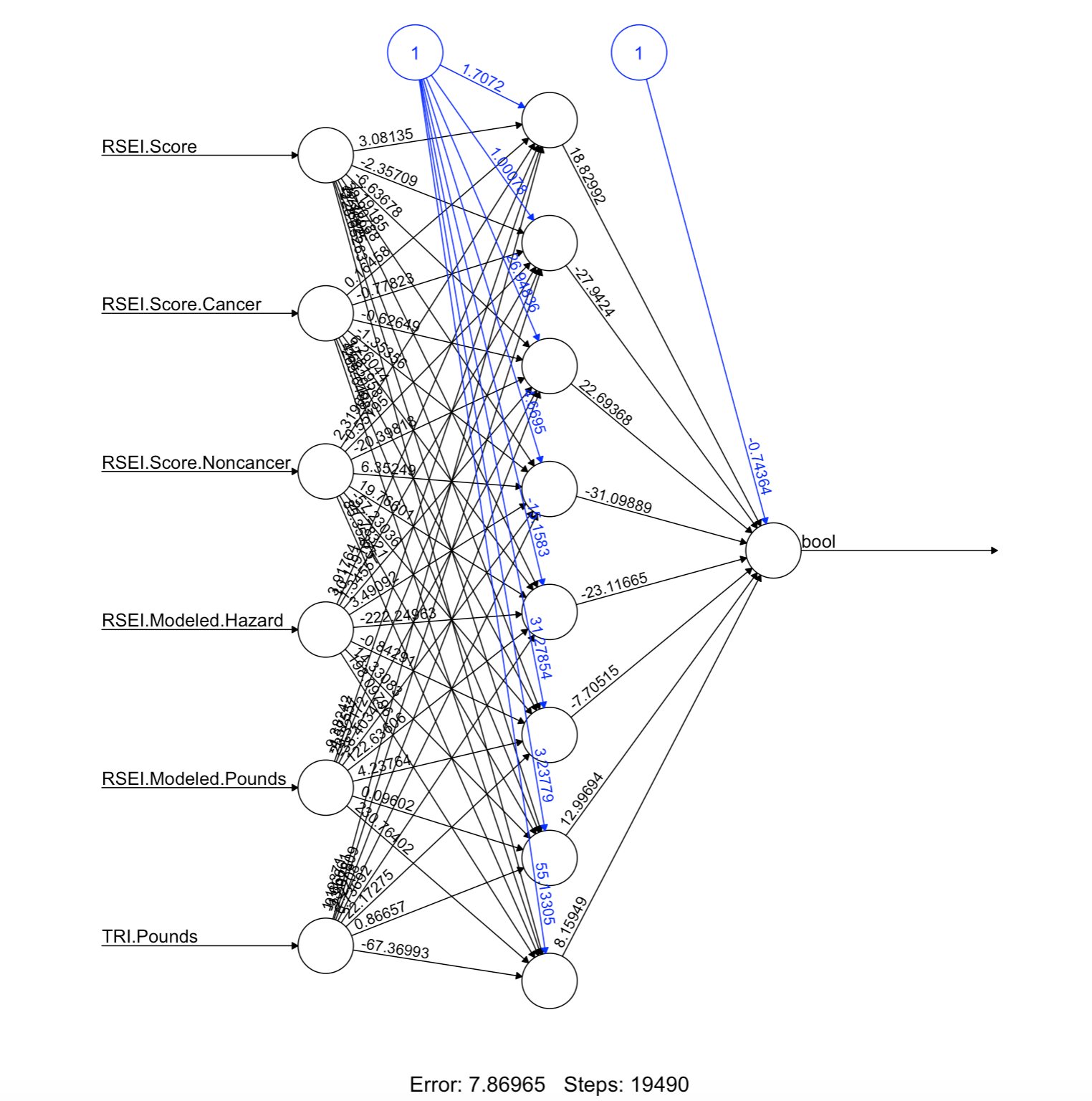
Table 2.1 - Neural Network

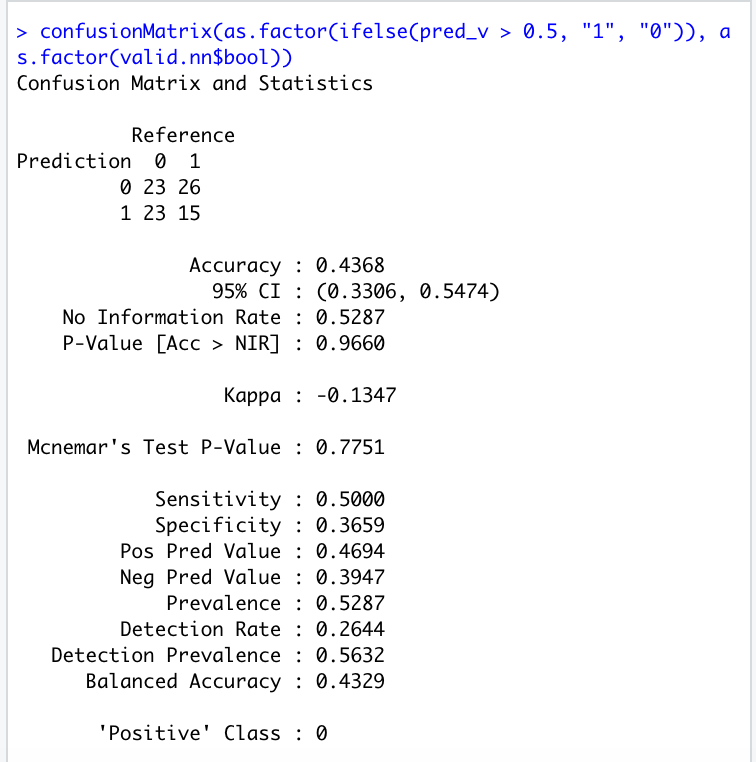
Table 2.2 - Confusion Matrix for Neural Network in Table 2.1

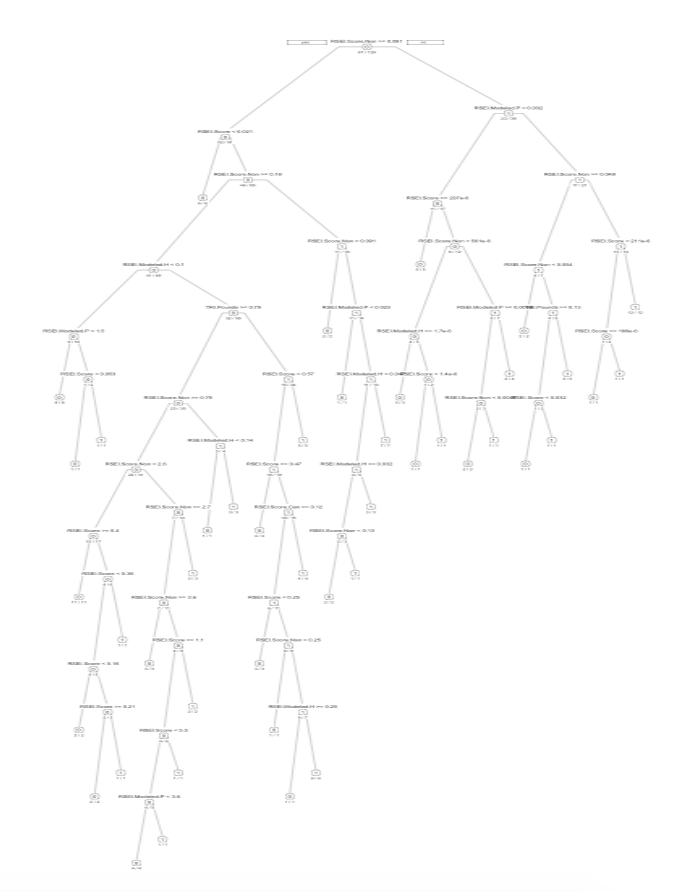
Table 3.1 - Un-Pruned Decision Tree

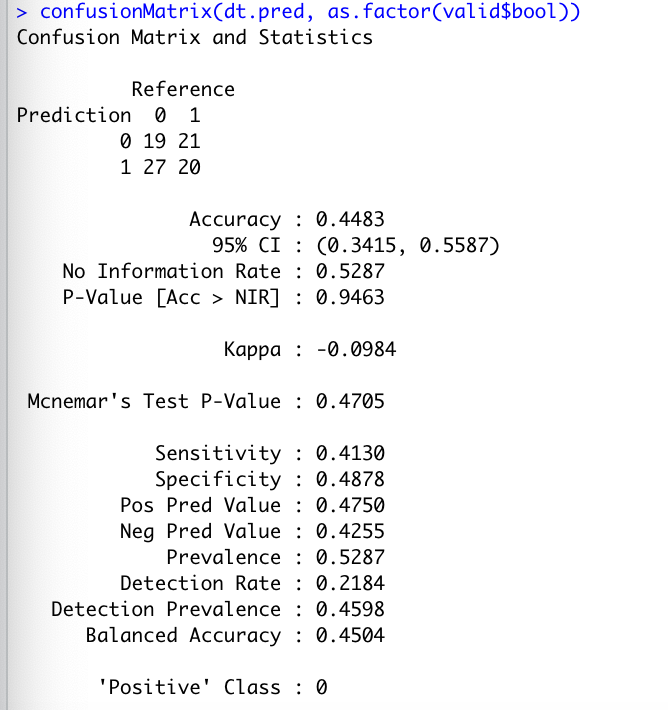
Table 3.2 - Confusion Matrix for Decision Tree in Table 3.1

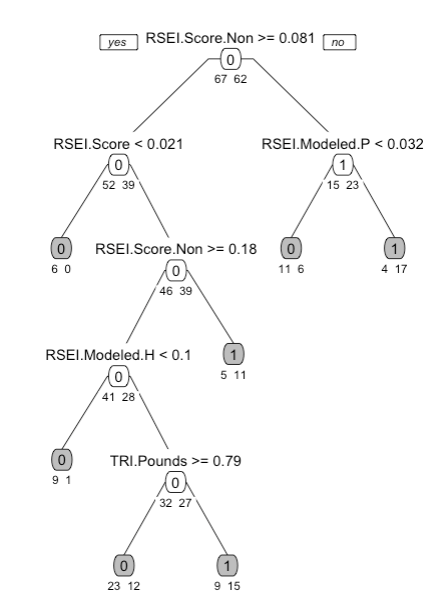
Table 3.3 - Pruned Decision Tree

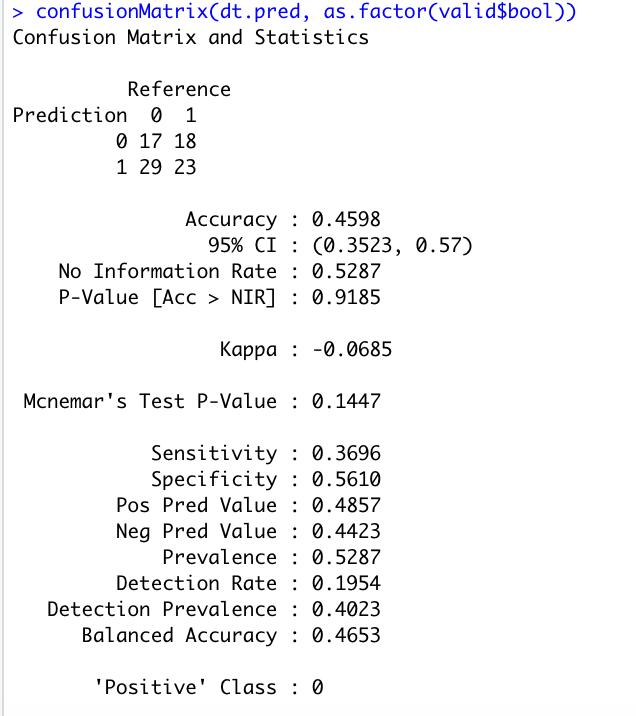
Table 3.4 - Confusion Matrix for Pruned Decision Tree in Table 3.3

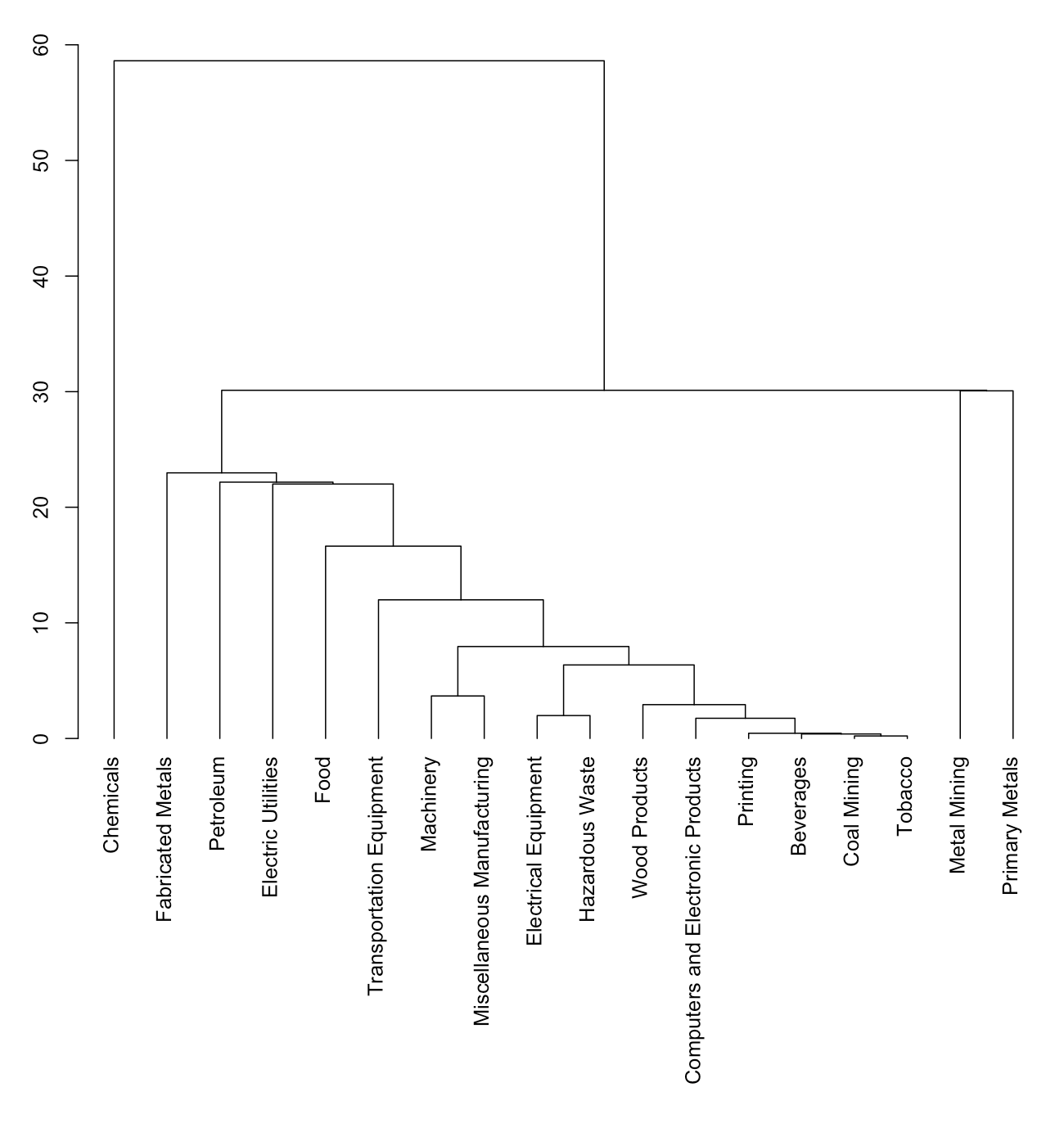
Table 4.1 - Hierarchical Cluster Single Linkage Method

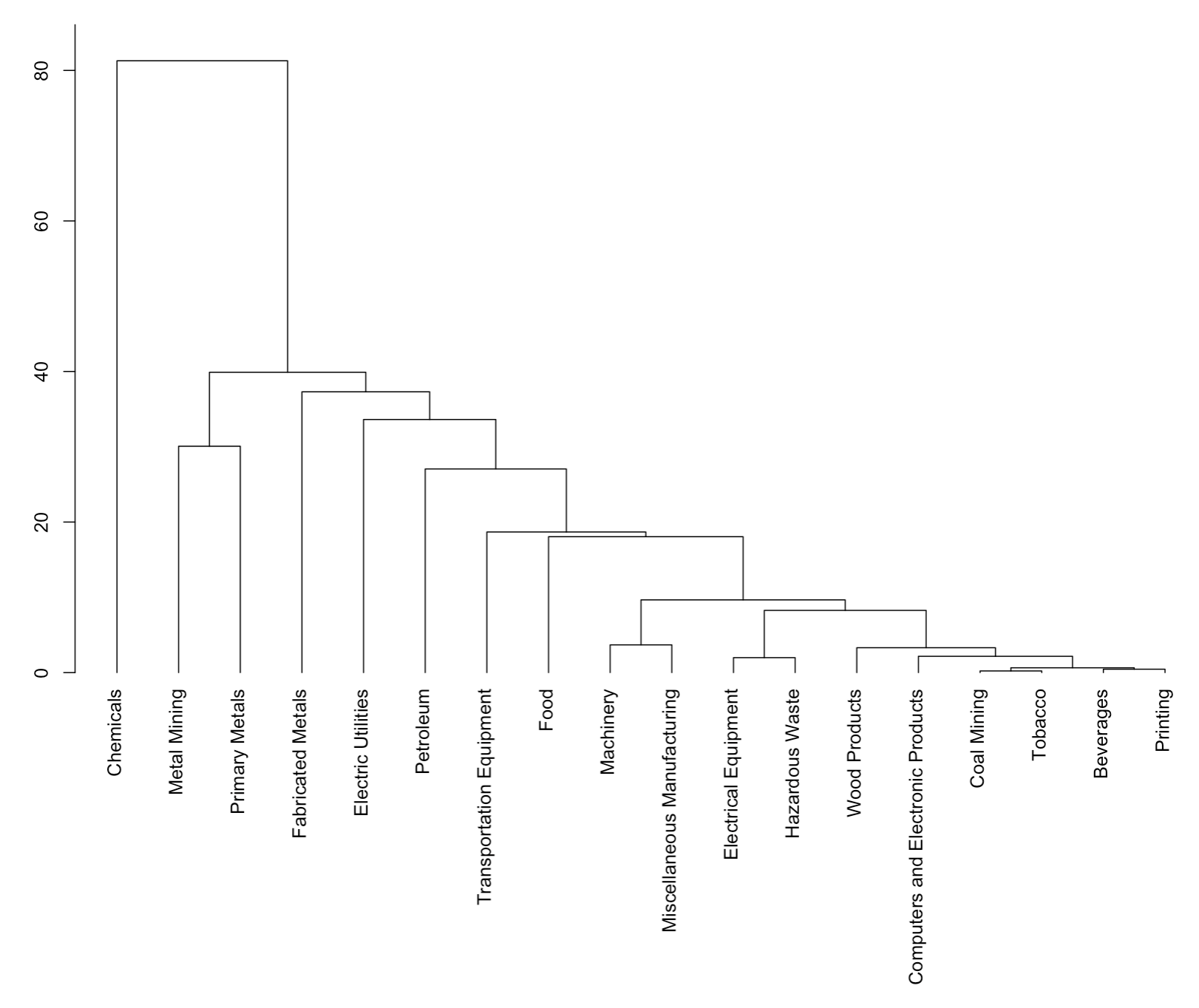
Table 4.2 - Hierarchical Cluster Average Linkage Method

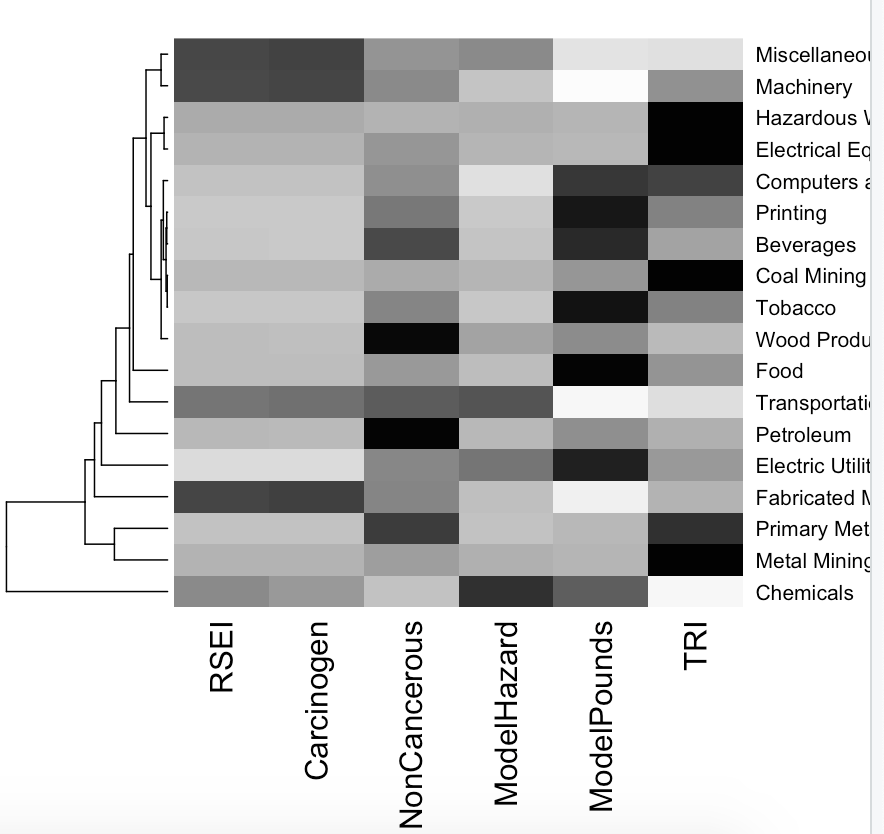
Table 4.3 - Heatmap Explaining Hierarchical Clustering Relationships

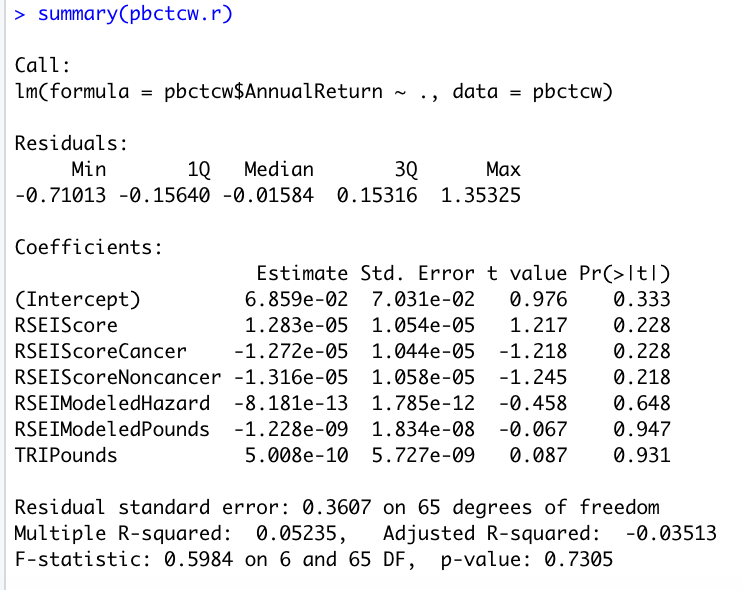
Table 4.4 - Regression for Printing, Beverages, Coal Mining, Tobacco, Computers and Electronic Equipment, and Wood Products Industries

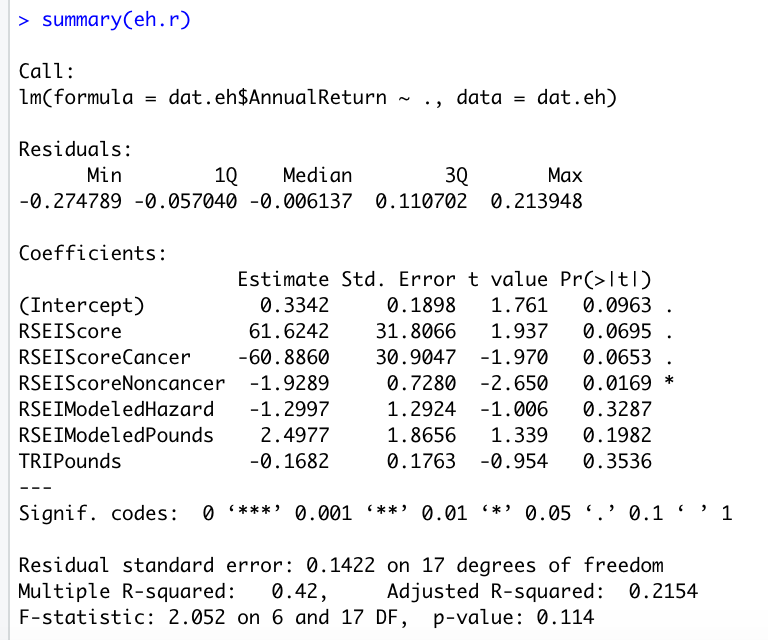
Table 4.5 - Regression for Electrical Equipment and Hazardous Waste Industries

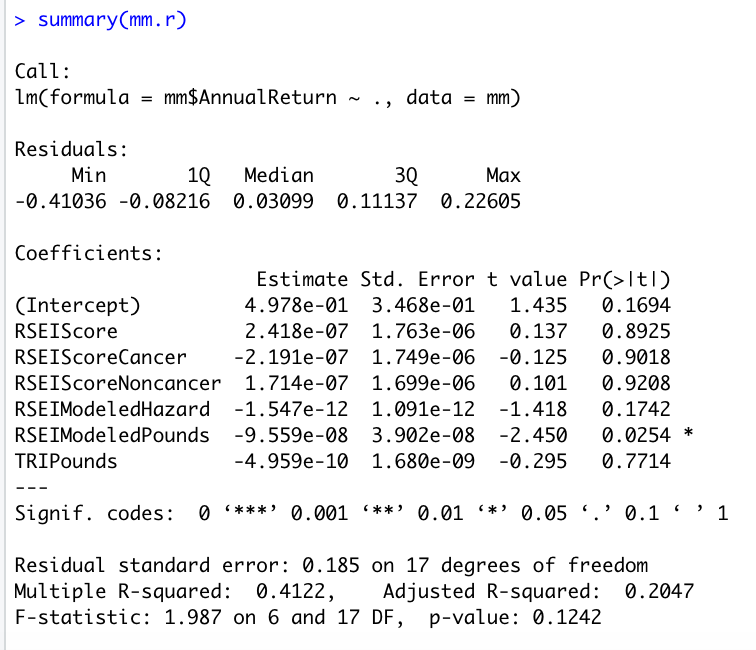
Table 4.6 - Regression for Machinery and Miscellaneous Manufacturing Industries

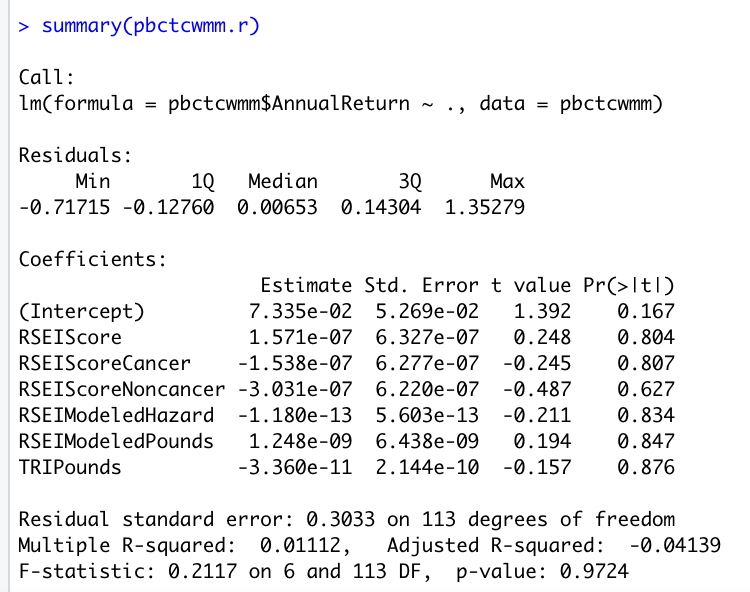
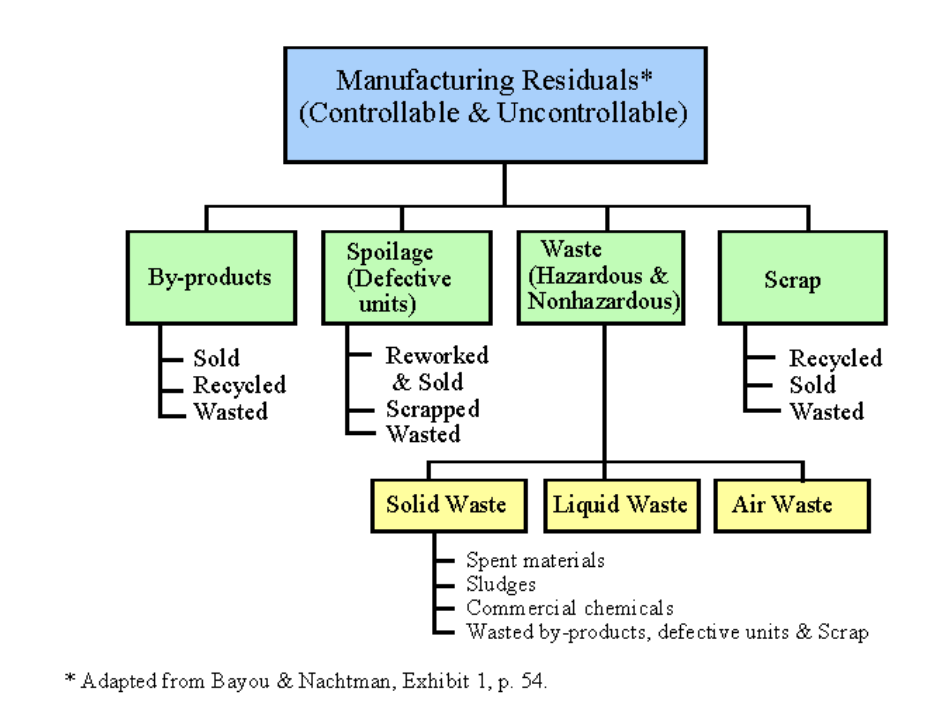
Table 4.7 - Sub-Cluster Regression for Hierarchical Cluster Cut at Height=10

Table 5.1 - Manufacturing Residuals Diagram

**References:**

ETF.com. (2020, July 01). What Is An ETF? Retrieved November 13, 2020, from <https://www.etf.com/etf-education-center/etf-basics/what-is-an-etf>

Martin, J. (n.d.). Bayou, M. E. and J. B. Nachtman. 1992. Costing for manufacturing wastes. Journal of Cost Management (Summer): 53-62. Retrieved November 13, 2020, from <https://maaw.info/ArticleSummaries/ArtSumBayouNachtman92.htm>

Thune, K. (n.d.). These Are the Top 10 Sector Funds Available at Fidelity. Retrieved November 13, 2020, from <https://www.thebalance.com/sector-funds-to-buy-at-fidelity-investments-4126394>

Tsydenova, O., & Bengtsson, M. (2010, September 23). Chemical hazards associated with treatment of waste electrical and electronic equipment. Retrieved November 13, 2020, from <https://www.sciencedirect.com/science/article/abs/pii/S0956053X10004393>

Understanding RSEI Results. (2018, December 23). Retrieved November 13, 2020, from <https://www.epa.gov/rsei/understanding-rsei-results>

Www.epa.gov. (2020, November 11). Retrieved November 12, 2020, from <https://edap.epa.gov/public/extensions/EasyRSEI/EasyRSEI.html>

Yahoo Finance - Stock Market Live, Quotes, Business & Finance News. (n.d.). Retrieved November 13, 2020, from <https://finance.yahoo.com/>

ETFs Holding IP: International Paper Co. (n.d.). Retrieved November 13, 2020, from <https://www.etfchannel.com/symbol/ip/>