Final Project

Group 29

March 20 2021

1 Introduction

For many years, humans concentrate on developing ourselves in an attempt to build up a better world. Technologies have been largely improved and the efficiency of production has reached new level. However, it has to be admitted the existence of sacrificing environment and some social problems. As people will always keep chasing for the better lives, it is crucial and urgent for us to solve the induced environmental and social problems just as improve the technology in the past. Obviously, ensuring the main variables that triggering such environment related and social problems are basic and important. Hence, the main aim for accomplishing this project is to develop a methodology for calculating the key performance indicators (KPIs) that related to environmental and social problems discussed in the given datasets. Noticing the potential existence of relationships between environmental problems and social problems, the big question in this project will be exploring the relationships of problems related to environment and society based on the found key performance indicators. After exploring the offered datasets and noticing the big question is complex, we decide to separate the main question into three main parts.

The first part will concentrate on the exploration of climate change from 2018 to 2020 reflected in the "corporation" folder. In this part, we further separate this part into several questions. The first will be holding the subset selection in order to ““. The second will focus on exploring whether the government and corporation strategies can trigger the environmental issues by finishing the research on how the engagement of corporations (how companies manage the whole value chain save energy) influence low GHG service (GHG: low greenhouse gas based on specific CO2 gram/mile maximum emission) and how the potential changing low GHG service influence scope 2 emission (indirect emissions from source that are owned by agencies) in the next year. Based on the subset selection, we aim at selecting the key variables that needed to be considered and in the second part, we want to further ensure a more obvious causal relationship. Furthermore, in the third part, the main theme will be exploring how scope 1 emission (direct emissions from sources owned by agencies) and scope 2 emissions in 2018 influence the set emission target in 2019 and finally influence the public policies held in 2020. This part explores the potential relationships between environmental issues and public polices under the main theme of climate change in two directions which can show stronger relationships between the two aspects "environmental issues" and "social problems" in the big question.

The second part will focus on exploring potential relationships between social and environmental issues under the theme of water security. Likewise, we further separate this part into several small questions. The first will be how the environmental influence the social aspect. In this part, we are willing to explore whether there exist inherent water-related risks which will impose the economic influence on the your business that will influence the strategic decision and whether there exists linkages between water and the operations of the organizations. The second will be how the social aspect influence the environmental issues. In this part, we focus on evaluating the importance of considering environmental aspects such as water quality in the social activities such as in the operation of organizations. The regular social activities can actually impose influence on environmental issues as monitoring means potential change on people's activities related to environment, resulting in trigger final influence on environmental issues. The last part will aim at accomplishing research on potential relationship between environmental and social issues under the theme of city. In this part, we switch from focusing on relationships between environmental and organizations to the relationship between the environmental issues and government level. Similar to the previous two parts, we still try to test the relationships in two directions. The first is to prove that the environment will influence the social decisions made by the government. The second will be test whether the social aspects from government level will influence the environmental issues.

Different from previous parts, we hold a test of accuracy in this part to further strengthen the credibility of the results of our project.

By separating these three main parts, we show the relationships between environmental issues and the social issues which is the big question of this project in three ways, including testing relationships between climate and management rules of organizations, relationships between water security and operations of social aspect in business, and the relationships between environment and decisions made in government-status. This, the comprehensiveness can be in some extent promised and even though further researchers are needed, the credibility of current project can be somewhat ensured.

2 Data

The datasets we used for investigating the relationship between water securities and social issues are "2018_Full_Water_Security_Dataset.csv", "2019_Full_Water_Security_Dataset.csv", and "2020_Full_Water _Security_Dataset.csv" provided by Kaggle. We select the specific questions and their corresponding responses for each years and classified them as either environmental issues or the social issues. For the environmental issues, we use the column 1 of W1.1 (the direct importance rating of water quality and water quantity to the success of the participants' business), the column 1 of W1.2 (the proportion of the water aspects being regularly measured), column 1 of W1.2b(the total volumes of water withdrawn, discharged, and consumed across all participants' operations), both columns of W1.2h (total water withdrawal data by source), both columns of W1.2i (total water discharge data by destination). For the social issues, we select W4.1 (identification of water-related risks with the potential to have a substantive financial or strategic impact on participants' business), the third column of W4.1c (the number and proportion of facilities exposed to water risks that have a substantive impact on participants' business), W9.1(identification of linkages or tradeoffs between water and other environmental issues in their organizations' operations), and W9.1a (the linkages or trade-offs). W9.1 and W9.1 are only available for the datasets in 2019 and 2020. The continuous numerical variables includes the column 1 of W1.2b, column 2 of W1.2h and W1.2i, and column 3 (column 2 in 2020) of 4.1c, while all other variables are categorical variables. For the categorical variables with importance rating (both columns of W1.1), we replace the observations with 6 as the most important and 0 as the least important. For the categorical variables with relevance (column 1 of W1.2h and W1.2i), the observations are replaced with 1 as "relevant", 0 as "relevant but volumes unknown", and -1 as "not relevant". For all the categorical variables with "yes" or "no" answers (W4.1 and W9.2), "yes" is replaced with 1 and "no" is replaced with -1. For the column 1 of W9.1a, "tradeoff" is replaced with -1, and "linkage" is replaced with 1. For W1.2, the observations are replaced with 1 as the smallest range and 6 as the biggest range.

Variables in Climate change dataset emis_percentage:a numerical variable of emission in percentage based on corporation emission target. scope_2_market: a numerical variable of Scope 2 Emission based on market data Scope_2_loc: a numerical variable of Scope 2 Emission based on local data Scope 1: a numerical variable of Scope 1 Emission low_GHG_Service: a categorical variable indicating whether corporation produce/provide low carbon product or service emission_target: a numerical variable indicating the type and amount of emission target a company has public_policy: a categorical variable indicating whether government implement new public policy regarding environmental issue this year engagement: a categorical variable indicating type of engagements the corporation implement to take care of climate-related issue in its value chain. energy_percent: a numerical variable indicating the percentage of corporation-wide spending on energy cost financial_opportunity: a categorical variable indicating the type of financial opportunity a company is facing regarding environmental issue The data is also categorized into different years(2018, 2019, 2020).If the variable is from year 2018, nothing will be added to the suffix; if 2019, then "_19" will be added to the end; if 2020, "_20" will be added. For example, energy_percent in year 2018, 2019, and 2020 will be the following: energy_percent, energy_percent_19, energy_percent_20.

3 Background

Over many years, humans have a reach a comparably high status of both technology and living standard. Unlike past harder time, compared to mainly focus on developing economics, more and more people turn their eyes on their influences which trigger environmental problems and explore the potential relationships between the environmental problems and social changes including different kinds of decisions made by governments and organizations such as public or business polices. Before continuing showing the reports

of some related researches on this area, the first step is to highlight some terms which will be used in the basic introduction and further personal analysis parts: low GHG refers to low greenhouse gas based on specific CO2 gram/mile maximum emission; scope 2 emission is indirect emissions from source that are owned by agencies; scope 1 emission equals to direct emissions from sources owned by agencies. According to Angeloantonio Russo who is in the department of management in Bocconi University, corporate environmentalism was correlated with improved financial performance in the short term, but corporate growth correlated with increased emissions. In his research, the professor admits the ensured relationships between environmental issues such as GHG emission and social issues such as the profits made by companies, and further explore the performance of such relationships on a specific company. Meanwhile, according to Zineb Moumen, etc., water security can impose potential influence on humans' activities including social activities such as different kinds of development related policies made by organizations and governments. Such crucial aspect in environmental issues should also be taken into consideration. Therefore, in this project, not just concentrate on those environmental problems, we would like to show a more specific progress to prove not only the existence of the relationships between environmental and social issues, but also discover the key performance indicators which play important roles and how they are related to social aspects.

4 Investigation

4.1 Corporations - Climate Change

In this section, we will focus on the climate change data set for corporation, and we will explore potential factors that might relate or contribute to the climate change.The data set for climate change covers 3 different years- 2018, 2019, 2020, and we will combine three together for this analysis.

4.1.1 Multivariate Regression

In the climate change data set, the variables that directly indicate the climate change can not be found. However, we are sure about the following facts: the emission of the carbon dioxide and other green house gases (GHG) is the major cause of the global warming. Therefore, we will consider GHG emission as the measurement of the environmental factors, and all of the remaining as the social factors. As a result, we have four variables for the environmental factors ("scope_1", "scope_2_market", "scope_2_loc", "emis_percentage"), and four environmental variables ("low_GHG_service", "emission_target", "public_policy", "engagement","energy_percent", "financial_opportunity").

To do some initial data exploration and to capture variables that might potentially related to the GHG emission, we will do a multivariate regression for each environmental variables against all of the social variables.

First, we will perform a multivariate regression for the scope1 emission of GHG for the year 2020.

```
lm(formula = scope_1_20 ~ ., data = climate_dt[, -c("account_number",
    "scope_2_market_20", "scope_2_loc_20", "emis_percentage_20",
    "scope_1_19", "scope_2_market_19", "scope_2_loc_19", "emis_percentage_19",
    "scope_1", "scope_2_market", "scope_2_loc", "emis_percentage")])

Residuals:
      Min        1Q    Median        3Q       Max
-21879874  -1763099   -606597    634103  64956468

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              -5000331    1726607  -2.896  0.00399 **
low_GHG_service            554907    1712250   0.324  0.74605
emission_target           -611001     697808  -0.876  0.38179
public_policy             1301043    2157368   0.603  0.54682
engagement                1793758     684562   2.620  0.00913 **
energy_percent             773609     385291   2.008  0.04536 *
financial_opportunity      389242     959240   0.406  0.68513
low_GHG_service_19        1319306    2405837   0.548  0.58375
emission_target_19         543884     821086   0.662  0.50811
public_policy_19         -1231413    3022884  -0.407  0.68397
engagement_19            -1073030     915542  -1.172  0.24192
energy_percent_19         1168113     441119   2.648  0.00843 **
financial_opportunity_19  -269865    1242846  -0.217  0.82822
low_GHG_service_20       -1450840    1930667  -0.751  0.45283
emission_target_20         388571     682310   0.569  0.56935
public_policy_20           835999    2395142   0.349  0.72725
engagement_20               56983     827693   0.069  0.94515
energy_percent_20         1008987     244881   4.120 4.64e-05 ***
financial_opportunity_20  -363393    1057548  -0.344  0.73132
---
```

Figure 1. The multivariate regression of scope 1 emission of 2020 against the social factors. We find the variables engagement, energy$_p ercent, energy_p ercent_1 9, energy_p ercent_2 0 a re s i g n i f i c a n t l y r e l a t e d t o s c o p e 1 e m i s s i o n$

It's not hard to find that the percent of operational cost spent on the energy(energy_percent) has a significant relationship with scope 1 emission of year 2020.

Then we repeat the multivariate regression for another environmental variable - "scope_2_market_20".

```
lm(formula = scope_2_market_20 ~ ., data = climate_dt[, -c("account_number",
    "scope_1_20", "scope_2_loc_20", "emis_percentage_20", "scope_1_19",
    "scope_2_market_19", "scope_2_loc_19", "emis_percentage_19",
    "scope_1", "scope_2_market", "scope_2_loc", "emis_percentage")])

Residuals:
     Min       1Q   Median       3Q      Max
-3971180  -623988  -273109   248418 17226817

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                 130564     597361   0.219  0.82716
low_GHG_service             936760     565894   1.655  0.09906 .
emission_target             134200     208618   0.643  0.52061
public_policy               156994     623182   0.252  0.80130
engagement                 -191194     199749  -0.957  0.33937
energy_percent              289576     114733   2.524  0.01220 *
financial_opportunity      -261805     343858  -0.761  0.44712
low_GHG_service_19         -548732     821325  -0.668  0.50466
emission_target_19          469359     263431   1.782  0.07596 .
public_policy_19            106641     894972   0.119  0.90524
engagement_19               -59195     271116  -0.218  0.82734
energy_percent_19            51061     127781   0.400  0.68978
financial_opportunity_19    323984     417580   0.776  0.43854
low_GHG_service_20          123613     632697   0.195  0.84525
emission_target_20         -615016     212235  -2.898  0.00408 **
public_policy_20            -70976     734631  -0.097  0.92311
engagement_20               319493     250057   1.278  0.20250
energy_percent_20            17717      76138   0.233  0.81618
financial_opportunity_20  -454209     317576  -1.430  0.15385
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 2. The multivariate regression of scope 2 of market emission of 2020 against the social factors. We find the variable energy$_p ercent, and emission_t arget_2 0 i s h i g h l y r e l a t e d t o t h e s c o p e_{2m} a r k e t_2 0$

Then we will repeat the regression for the other two environment variables of year 2020, and find the variable "energy_percent" is related to the variable "scope_2_loc_20", and variable "engagement_20" and "emission_target" are related to the "emission_percent".

To get a comprehensive picture on the potential relationship between variables, we also repeat the multivariate regression on environmental variables for year 2018 and 2019. Also, we did multivariate regression for each social variables against all of other variables. Since it might be too lengthy to mention everything in the report, I will only include observations that's key to the following analysis.

First, we have observed a significant correlation between variables the scope 1 emission of 2018, emission target of 2019, and public policy of 2020, which will be discussed in the later sections.

Another key observation is that the variable percent of operational cost spent on energy (energy_percent) seems to have a strong relation with the GHG emission, and we will further examine this is next section.

4.1.2 Examine variable energy_percent

As we have observed earlier, percentage of the operational cost spent on the energy is a factor that contributes greatly to emission.

```
Call:
lm(formula = climate_dt$scope_1 ~ climate_dt$energy_percent_19)

Residuals:
      Min        1Q    Median        3Q       Max
-30482466   -351079   -334976   -156171  74385357

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                -2574112     539746  -4.769  2.5e-06 ***
climate_dt$energy_percent_19  2925269     203653  14.364  < 2e-16 ***
---
```

Figure 3. This is the linear regression of scope 1 emission of year 2020 against the percentage of operational cost spent on energy in 2019.

From the above graph, we realize that the variable energy_percent is highly related to the green house gas emission, which makes sense since the consumption of the energy is the major source of the green house gas emission.

Again, we performed a multivariate regression on percentage of operational cost spent on energy in 2019 (energy_percent_19) against all of the other variables.However, no significant correlation is found except with energy_percent variables that are from different years. Therefore, at least in the variables we examined, the percent of the operational cost spent on energy is not a mediator between social variables and the climate variables, but the variable itself is highly related to climate variables.

4.1.3 Governmental/Corporation Strategies ⇒ Environment Issues I

Based on the results in multivariate regression, we look for variables that possibly have an effect on each other. To ensure a more likely causal relationship, we only seek for statistical relationships between a vector recording events in a base year and vectors in the following year since some implemented strategies take time to have an effect, and at least there will not be a reversal causal inference. Following the multivariate regression results, we find that the variable "engagement" in 2018, a variable indicating how much the company engage in climate-related issues in her value chain (how many sectors of chain interacted with), possibly relates to "low_GHG_Service" in 2019, a variable indicating whether product or service provided by this company expiates at a low carbon level. In the multivariate regression section, this relation has a p-value of 0.0327, which is lower than 0.05, indicating a high possibility of correlation.

```
        Pearson's Chi-squared test

data:  climate_dt$engagement and climate_dt$low_GHG_service_19
X-squared = 54.258, df = 3, p-value = 9.887e-12
```

Figure 4. Chi-square test for independence of climate-related corporation engagement and low GHG service

The null hypothesis of this chi-square test is that there is no relationship between the variables "engagement" in 2018 and "low GHG service" in 2019, that they are independent of each other. By testing independence of these two variables, we get that the p-value of this chi-square test is 9.887e-12. Hence, we can conclude that we can successfully reject the null hypothesis at a 99.9% confidence level that these two variables are dependent on each other. That means if a company engages more actively in taking climate-related issues in her value chain into consideration in 2018, she will probably provide more low GHG services in the following year. This is an internal relationship within companies implying whether and how business administration can affect its own supply from an environmental perspective. This has two possible implications. First, if we fly higher and observe companies from a social vision, we can manage to define engagement level as a key performance indicator about whether company takes actions to face climate issues. Second, engagement possibly affects future company-wide emissions. To prove these two implications, we should test whether low GHG service level changes emission level.

In the following test, we want to seek for statistical proof of the relationship between low GHG service level in 2019 and Scope 2 emission level. From multivariate regression, we observe that the p-value between these two variables is 0.092073, which is still pretty low.

```
Call:
lm(formula = climate_dt$scope_2_market_20 ~ climate_dt$low_GHG_service_

Residuals:
     Min       1Q   Median       3Q      Max
 -754158  -701117  -284181   -23777 20257842

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                    263828     172413   1.530   0.1269
climate_dt$low_GHG_service_19  490330     208289   2.354   0.0192 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1741000 on 322 degrees of freedom
  (206 observations deleted due to missingness)
Multiple R-squared:  0.01692,   Adjusted R-squared:  0.01387
F-statistic: 5.542 on 1 and 322 DF,  p-value: 0.01917
```

Figure 5. linear regression between low GHG service in 2019 and Scope 2 emission in 2020

The null hypothesis is that low GHG service provided by companies 2019 has no statistical correlation with Scope 2 emission level in 2020. In this test, we observe that the p value is 0.0192. Therefore, we can conclude that we are 95% confident that there is a correlation between these two variables, statistically. Unfortunately, we observe that the slope is 490330, a positive, big number. This is the opposite of what we expected, since we usually expect a more environmental implementation results in a reduce of emission. We then notice that we are testing on emission scope 2, which is the indirect type of emission. Therefore, we guess that it is possible that this service reduces scope 1 mission. Hence there is the following test:

```
Call:
lm(formula = climate_dt$scope_1_20 ~ climate_dt$low_GHG_service_19)

Residuals:
     Min       1Q   Median       3Q      Max
-3497718 -3442152  -984586  -866354 84715785

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                    984586     629739   1.563  0.11859
climate_dt$low_GHG_service_19 2513193     797379   3.152  0.00172 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8542000 on 487 degrees of freedom
  (41 observations deleted due to missingness)
Multiple R-squared:  0.01999,   Adjusted R-squared:  0.01798
F-statistic: 9.934 on 1 and 487 DF,  p-value: 0.001723
```

Figure 6. linear regression between low GHG service in 2019 and Scope 1 emission in 2020

The p-value this time is 0.00172. Hence, we clearly reject the null hypothesis. We are very confident about the existence of correlation between these two variables. So unfortunately, we observe the slope is yet positive. This is not following common sense, but shows the strength of data analysis right away. Statistically, we can conclude that low GHG services may increase emissions of both types in the following year. Subsequently, we should reject our implications made above, at least at this level. We cannot conclude that the fact that implementations of more climate-related corporation engagement can increase low GHG service will result in less emission in following two years, and should not conclude that we should set up government engagement as a performance indicator of climate consideration. However, this provides us an opportunity for further studies. We can talk about whether such engagement is right, how to appropriately address and implement it, and what exhausts its effect, etc.

4.1.4 Governmental/Corporation Strategies ⇒ Environment Issues II

In this section, we again reference to the multivariate regression result we get. We want to test whether companies take emission target into account, that the target will influence its emission outcome in the following year. Researching on the regression list, we observe that the predictive variable referencing to 2019 emission target has a relatively small p-value when the explanatory variable is emission percentage 2020. Since the variable emission target is categorical that it records

4.1.5 Environmental issues ⇒ Governmental/Corporation policies

In this section, we perform a mediation analysis on how scope 1 emissions will affect corporation's emission target, and such targets will affect administrative policies regarding environmental issues implemented by government in the following year.

To perform such analysis, we first test whether Scope 1 emission in 2018 will affect governmental policies implemented in 2020. This time difference is reasonable if we take the time to discuss and enact a new systematic policy. And we choose to analyze Scope 1 emission total amount because we don't rely on multivariate regression this time and this is the emissions that is directly expiated by the things owned or controlled by this corporation. It is a more direct measure of environmental impact comparing to Scope 2's.

```
Call:
lm(formula = climate_dt$public_policy_20 ~ climate_dt$scope_1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.89848  0.09676  0.10228  0.10253  0.10257

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        8.974e-01  1.370e-02  65.500   <2e-16 ***
climate_dt$scope_1 2.709e-09  1.373e-09   1.974    0.049 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2931 on 491 degrees of freedom
  (37 observations deleted due to missingness)
Multiple R-squared:  0.007872,  Adjusted R-squared:  0.005851
F-statistic: 3.896 on 1 and 491 DF,  p-value: 0.04897
```

Figure 7. linear regression between Scope 1 Emission in 2018 and Public policy implemented in 2020

We observe that in this test, the p-value is 0.049. That means we are 95% that increasing Scope 1 emission by 1 unit will result in a higher possibility of implementing various type of environmental public policy. As a result, we can conclude from a statistical sense that if the government observes more emission, it will try to negotiate on the issue and make new related policies.

Then, we perform a regression test to see whether there exists a relationship between corporation's emission target in 2019 and the public policies in 2020.

```
Call:
lm(formula = climate_dt$emission_target_19 ~ climate_dt$scope_1)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6327 -0.5054  0.3696  0.5086  1.5087

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        1.491e+00  4.381e-02   34.04  < 2e-16 ***
climate_dt$scope_1 1.251e-08  4.499e-09    2.78  0.00563 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9625 on 516 degrees of freedom
  (12 observations deleted due to missingness)
Multiple R-squared:  0.01476,   Adjusted R-squared:  0.01285
F-statistic:  7.73 on 1 and 516 DF,  p-value: 0.00563
```

Figure 8. linear regression between Emission target in 2019 and Scope 1 Emission in 2018

From this test, we observe that the p-value of this test is 0.00563. This is a very low value that we are 99.9% confident that there exists a statistical correlation between Emission target set in 2019 and Emission amount (Scope 1) in 2018. Since there is a time constraint, we may achieve more values from the correlation's perspective. We observe that for each 1 unit of amount of emission increased, a small amount of increase of unit of emission target will be achieved.

Then, we perform linear regression of public policy in 2020 on emission in 2018 and emission target in 2019 to answer the question

```
Call:
lm(formula = climate_dt$public_policy_20 ~ climate_dt$scope_1 +
    climate_dt$emission_target_19)

Residuals:
     Min       1Q   Median       3Q      Max
-0.96759  0.03241  0.07974  0.12688  0.17443

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  8.256e-01  2.475e-02  33.362  < 2e-16 ***
climate_dt$scope_1           2.151e-09  1.367e-09   1.574 0.116178
climate_dt$emission_target_19 4.733e-02  1.364e-02   3.470 0.000566 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2899 on 490 degrees of freedom
  (37 observations deleted due to missingness)
Multiple R-squared:  0.03167,   Adjusted R-squared:  0.02772
F-statistic: 8.013 on 2 and 490 DF,  p-value: 0.0003764
```

Figure 9. linear regression of public policy in 2020 on Emission target in 2019, Scope 1 Emission in 2018

From the regression of multiple predictive variables, the p-values for Emission target and Emission amount are 0.000566, 0.116178, respectively. Then we are 99.9% confident that even if we take two variables into consideration, emission target in 2019 still affects public policy in 2020. This is reasonable because as companies increase number of targets, governments will notice the potential threat of environmental issues and start to make policies and implement them in the following years. Since there is a statistical correlation between emission and its target, along with this test, we can conclude that emission target in 2019 at least partially mediate the relationship. While at the same time, we cannot provide statistical evidence about whether emission amount in 2018 affect policy in 2020 through this relationship. We then calculate the amount of effects. The total effect of emission in 2018 on public policy in 2020 is 2.709e-09, the direct effect of this relationship is 2.151e-09, while the indirect effect is 12.51e-09*4.733e-02 = 5.92e-10. Last but not least, we need to perform Sobel test to prove the significance of this mediation analysis. We calculate that T equals 2.169514, which is greater than 1.96. Then we are confident at a 95% level that the mediation analysis is statistically true. Last but not least, the 95% confidence interval is (-4.2522, 4.2522). Though slope is positive, we can't observe the bound absolute value's difference because indirect effect mean is extremely small relative to T value. Emission target set by companies in 2019 will partially mediate the correlation between Emission amount in 2018 and Public environmental policy made in 2020. That means when government made public policy, governors indeed look at and possibly catch the most recent news, information to make relevant policies (though there is a 1-2 years lag). We can say companies take total Scope 1 emission into account and make target accordingly. This is showing a sense of social responsibility on their side. In the mean time, government takes environmental factor (total emission) and social factor (corporation action) into account when they make their relevant policies.

4.2 Corporations - Water Securities

4.2.1 Environmental Issues ⇒ Social Issues

By taking a close look at the response dataset of the water security questionnaire 2018, several attributes associated with the environmental or social issues are selected manually and re-combined into a new data frame. The selected question numbers are Question1.1, Question1.2b, Question1.2h, Question1.2i, Question4.1, Question4.1c, Question9.1 and Question9.2. In order to have a better understanding regarding the impact of environmental issue on corporations' social development, social-impact related question 4.1– have you identified any inherent water-related risks with the potential to have a substantive financial or strategic impact on your business—is picked to be the dependent variable.

The primary goal of this section is to find which environmental factor is most important in determine Question 4.1. Since the question is "Yes/No" question, we re-code the answers "Yes" to "1" and "No" to "-1". Then we apply the random forest algorithm for feature selection on this binary classification. For the random forest parameters, we set number of trees to 100 and get he error estimations.

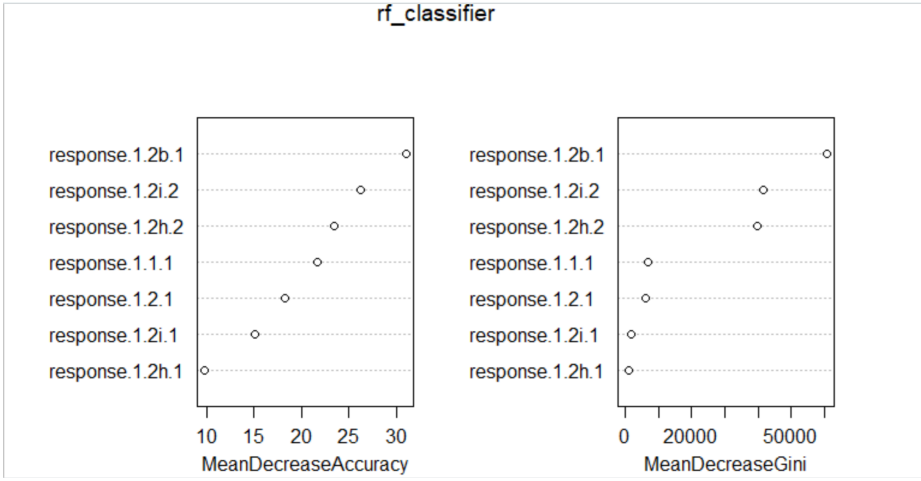

Figure 10. Mean Decrease Accuracy and Mean Decrease Gini for feature selection

```
Call:
 randomForest(formula = response.4.1 ~ ., data = training, importance = TRUE,      ntree = 30,
mtry = 2, na.action = na.omit)
               Type of random forest: classification
                     Number of trees: 30
No. of variables tried at each split: 2

        OOB estimate of  error rate: 0.09%
Confusion matrix:
       -1      1 class.error
-1 116912    382 0.003256774
1      55 353118 0.000155731
```

Figure 11. Evaluation metrics of random forest

As the data table shown above, the OOB error rate is only 0.09%, which indicates that the model is a perfect match for the prediction of question 4.1. Moreover, we found out that response 1.2b.1, the continuous variable -total volumes of water withdrawn, discharged, and consumed across the operations is the most important attributes to identify the inherent water-related risks because it has the highest mean decease accuracy score. Thus, its coefficient correlation is more vital to other variables'.

4.2.2 Social Issues ⇒ Environmental Issues

The same procedure of feature selection as the one in last section is performed here. This time, we from a new data frame with social issue questions, Question 4.1, Question 4.1c, Question 9.1, Question 9.1a as the independent variable of the random forest algorithm, and one environmental issue Question 1.1.1 as the dependent variable.

Since the Question 1.1.1 is the importance of water rating, we re-code the answers to ordinal scale from 1 to 6. Then we apply the random forest algorithm for feature selection on this multi-variate classification. For the random forest parameters, we set number of trees to 100 and get the error estimations.

```
Call:
 randomForest(formula = response.1.1.1 ~ ., data = training2,      importance = TRUE, ntree
= 30, mtry = 2, na.action = na.omit)
               Type of random forest: classification
                     Number of trees: 30
No. of variables tried at each split: 2

        OOB estimate of  error rate: 46.49%
Confusion matrix:
   -2 0      1 2      3 class.error
-2  0 0   8517 0   5544   1.0000000
0   0 0 125137 5  88852   1.0000000
1   0 0 865213 3 420469   0.3270412
2   0 1  91164 4 130159   0.9999819
3   0 0 362640 0 553132   0.3959938
```

Figure 12. Mean Decrease Accuracy and Mean Decrease Gini for feature selection

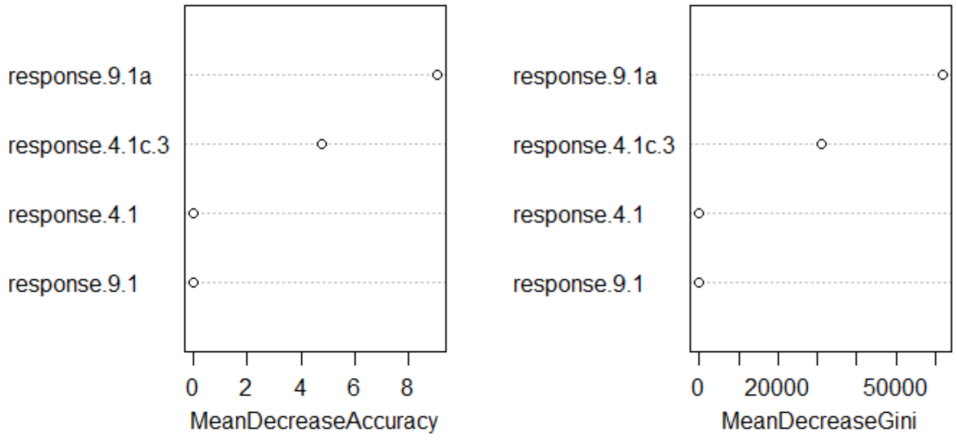

Figure 13. Evaluation metrics of random forest

As seen in the table above, we got an error of 46.49%, which means that this model is not a good model to use for predicting the environmental problem from the social aspects. The bigger mean decrease accuracy, the less redundant attribute. Response 1.2b.1 is the most important variable here, however, it is not that significant to the environmental variable, the importance of water, because of the error rate of the model. Thus, its coefficient correlation cannot be determined as more important than others'.

4.2.3 Influence of Years

To investigate if the environmental issues and social issues in the past has increasing influence in recent years, we do the causal inference by regression to decide if their relationship is positive. First, we select the column 1 of question W4,1 as the reference. W4.1 asks the participants if they find the water securities may have the impact on their business. We assume that as the years increase, the problem shouldn't be mitigate since the social issues and the environmental issues promote each others.

```
Call:
lm(formula = response.2020 ~ response.2018 + response.2019, data = test4.1.2)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8596 -0.1598  0.1404  0.1404  1.8402

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.009698   0.042755   0.227    0.821
response.2018 0.122059   0.123532   0.988    0.325
response.2019 0.727877   0.122652   5.935  1.7e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5389 on 164 degrees of freedom
Multiple R-squared:  0.7071,    Adjusted R-squared:  0.7036
F-statistic:   198 on 2 and 164 DF,  p-value: < 2.2e-16
```

Figure 14. Regression Outcomes of W4.1 for 2020 respect to 2018 and 2019

According to the outcomes of the multivariate regression, the adjusted R-squared is 0.7036, which is closed to 1, so this regression is significant. Both of the estimates for the responses in 2018 and 2019 are positive, and the response in 2019 respect to 2020 has bigger slope and lower p-value, our assumption of causal inference has been showed. Similarly, we select the W1.1 as the reference. The question W1.1 asks the participants to rate the importance of water quality and quantity to their business, so it should have the same outcomes as W4.1.

```
Call:
lm(formula = response.2020 ~ response.2018 + response.2019, data = test1.1.2)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3454 -0.7080  0.0010  0.6546  3.3836

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.70802    0.05283   13.40   <2e-16 ***
response.2018  0.25480    0.02584    9.86   <2e-16 ***
response.2019  0.29100    0.02637   11.03   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.186 on 1327 degrees of freedom
Multiple R-squared:  0.2281,    Adjusted R-squared:  0.2269
F-statistic:   196 on 2 and 1327 DF,  p-value: < 2.2e-16
```

Figure 15. Regression Outcomes of W1.1 for 2020 respect to 2018 and 2019

Similar to the outcomes of W4.1, both of the estimates are positive, and the slope for 2019 is bigger. So for the same company with the same account number, the water's importance is constantly increasing, which also shows that there might be causal relations among years.

4.3 Cities

In the researching process on the cities data, we mainly focused on finding the connections between cities administrative actions and past data and the emission of the following year, mainly through cities' responses to the questionnaire. We focus on climate issues instead of water security due to the fact that more information was given in the data sets related to climate issues rather than water security issues in the city level. In studying the relationship between corporations and climate issues, we found that the corporation policies and strategies towards environmental issues successfully addressed the issues with statistical proves. In cities level, we would also be interested in finding the relation between cities governance strategies and the outcome of dealing with climate issues. The outcome was measured by the amount of emission regarding scope1 and scope2.

In the cities data, we picked Question 2.0, 4.0, 5.0, 5.5, 6.1, 7.6, 7.6, 7.6, 8.4, 11.0, 14.2. Those are answers to the questionnaire that were easy to be interpreted by statistical tools. It didn't mean that other answers from the data sets were not valuable in our research. Similarly to the corporation investigation, we would consider green house gases (GHG) as our criterion of cites tackling climate issues since GHG is the major cause to global warming. In the city-level point of view, we mainly focused on two variables: "Scope 1 emission" and "Scope 2 emission". We used mainly the data from 2019 and 2020, since the questionnaire of 2018 differed significantly to that of the following two years, which made it difficult for comparison.

We performed multivariate tests for both scope 1 and scope 2 emission for the year of 2020. We found that most variables we recorded in the cities level did not have strong connections to the emission of GHG in the same year. Considering the chronological lag of city policies and strategies, we would compare the factor of 2019 to the emission recorded in 2020 to get a better understanding of the effectiveness of those policies and strategies by repeating the multivariate regression.

4.3.1 City-wide Emission Inventory to Environmental Issues

We want to seek statistical proof that whether the cities having a city-wide emission inventory would have an impact on the GHG emission of the particular city in the next year. From linear regression we can find that the p-value is 1.76e-06 which is low considering a 95 percent confidence interval. As a result, we can conclude that we are 95 percent confident that there is a correlation between having a city-wide emission inventory and scope 2 emission the following year. However, after doing the same test by applying linear regression on scope 1 emission, we find a p-value of 0.632, which does not support us to override the null hypothesis that having a city-wide emission inventory does not have statistical correlation with scope 1 emission the following year.

```
Call:
lm(formula = scope2_emission_19 ~ citywide_emmision_inventory_19,
    data = cities_dt, na.action = na.omit)

Residuals:
     Min       1Q   Median       3Q      Max
-1443587  -143579  -133136   -48942  7758471

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                     1564744     281256   5.563 7.21e-08 ***
citywide_emmision_inventory_19  -472774      96422  -4.903 1.76e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 672700 on 234 degrees of freedom
  (350 observations deleted due to missingness)
Multiple R-squared:  0.09317,   Adjusted R-squared:  0.08929
F-statistic: 24.04 on 1 and 234 DF,  p-value: 1.764e-06
```

Figure 16. linear regression of scope 2 emission in 2020 against having a city-wide emission inventory in 2019.

4.3.2 Size of Park Space to Environmental Issues

By similar approaches through linear regression, we found that scope 2 emission the following year is also significantly correlated with the size of park space in cities with a p-value of 0.00589. With 95 percent confidence we can conclude that there is a correlation between park space and the scope 2 emission the following year. Interestingly, when applying the same test on scope 1 emission the following year and park space, we fail to prove that there is a correlation between the two, with a p-value of 0.071.

```
Call:
lm(formula = scope2_emission_20 ~ size_park_space_km2_19, data = cities_dt,
    na.action = na.omit)

Residuals:
     Min       1Q   Median       3Q      Max
-1153880  -260855    -7877   570516   570516

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            -33170     330337  -0.100  0.92222
size_park_space_km2_19   15466       4314   3.585  0.00589 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 615400 on 9 degrees of freedom
  (575 observations deleted due to missingness)
Multiple R-squared:  0.5881,    Adjusted R-squared:  0.5424
F-statistic: 12.85 on 1 and 9 DF,  p-value: 0.005886
```

Figure 17. linear regression of scope 2 emission in 2020 against the city-wide park space in 2019.

5 Conclusion

In this project, our group aims to find how environmental issues and social issue affect each other. The main environmental issues we concern about is the ones that closely relate to climate change and water security. We investigate closely that when companies face or create environmental problems themselves, how do they make changes in respond to that. Subsequently, how do the government recognize environmental changes and accordingly makes policy and governmental administrations on the issues. Under this sense, we also look for evidence to support how these entities strive for finding and recognizing the issues, and whether they take these factors into account. On the other hand, the relationship also goes the other way. We want to investigate how social issues related to government and companies will affect our environments regarding climate and water bodies.

Since there is a huge package of data sets, finding an appropriate way to manage and use it is crucial. Depending on the feature that it includes data from the year of 2018, 2019, and 2020, we transform it into three big data sets, each concerning one aspect of climate change, water security, and governmental administration, which have many columns listing factors of records from a specific year. With such data, we can manage to use overarching testing method to see if there are some general relationships among the factors. Though this is not all time accurate, it is obvious that there are some clear relationship between the environmental and social factors listed. In addition, we try to manipulate data from different years to force a more probable correlation and make more inferences.

Here is the analysis results of this report. We first investigate social issues along with climate-related issues (risk and opportunity). Among the social issue, the variable percent of the operational cost on energy (energy_percent)shows strong correlation with climate-related issues, but it's not a mediator between climate issues and other social issues. Based on the overarching test, we also find that companies' engagement level in 2018, the variable indicating whether the company cares about climate-related issues in each sector of her value chains, affects the company's low green house gas product/service in the following year. This is an implication saying if a company takes care of climate-related issue by supervising its sectors, it will secure a reduction of green house gas emission indirectly later. This means it is possible that company's action of this will indeed have an influence on the market. Such product can release the company's advertising force, educating customers from a perspective of environmental friendliness, and so on. At the very least, such action is effective and will indirectly contribute to solve environmental problems in society, the complex system with so many contributors acting. Hence, engagement level can become a KPI measuring company's contribution to society regarding climate issues. To extend this result, we test whether low green house gas product contributes to a direct reduction of emission. Surprisingly, the result is not only positive, but in the opposite way we expect. The more low GHG service we produce and supply, the more emission we get. This is counter-intuitive but worth discerning. If low GHG service rate changes, what is another moving factor that contribute to the increase? Fake low GHG standard? Market dynamic? Extension of chain results? We don't know it yet. But it provides us a perfect start for further investigation. Moreover, this is a qualification for the above test. The following result is more interesting. We investigate how Scope 1 emission will affect corporation and government's actions. Scope 1 emission is the more measurable and direct emission contributed by an entity. With an increase of Scope 1 emission total amount in 2018, we find that company's emission target will increase (which means company will assign new targets, regarding intensity or absolute numbers) in the following year. This means clearly that company takes its previous emission into account and makes her emission target accordingly. Also, we find that such emission has a positive relationship with public policies made in 2 years later. Then it is reasonable to deduce that government will recognize emission level in the base year, making plans and revisions, then finally reach a final policy in years. That is, government is making policies accordingly, not in a random manner in the least saying. More interestingly, we find that there is a mediation between these three variables, that policy made in 2020 relates to both Scope 1 emission in 2018 and emission target in 2019. We also observe that the mediation tells that the government recognizes the climate influence both through direct observation of the emission numbers and through corporations' floating targets: the government looks at both environmental and social aspects from a high vision. This sounds like a reasonable, rigorous approach, and requires further studies on this issue to confirm.

After testing on climate-related issues, we also look at water security-related issues.In this section, several attributes are selected manually to investigate social and environmental problems correlation. Thus, we are looking for the most significant environment variable that correlates to the social issue and the most important social variable that correlates to the environmental variables. The results are that total volumes of water withdraws, discharged and consumed correlates to the social issues - inner risk of corporations - the most among all of the environmental factors. However, no strong social factors show obvious correlation to the water importance of the company which related to the environmental issue.

Our research might be restricted by the number of observations in the cities datasets. Indeed, the cities datasets contains an enough number of observations, yet many of which contains NA values. It means that many fields of cities' questionnaire are left to blank.

By interpreting the data we find that some strategies of cities would affect the result of scope 2 emission of green house gases, yet many other strategies might not have a statistical impact. The impacting factors on scope 2 emission were whether cities have city-wide emission inventories and the amount of park space in km^2. However, it is possible that there is a confounding factor that lead to both decrease in scope 2 emission and having an city-wide emission inventory. More research should be done on this topic in the future.

After all, we start to have a structured picture of how environmental issues and social issues relate and interact with each other, or among themselves. On an overarching sense, we can say that there are so many interacting factors, while they have so many further study required to perform to finish the test, that they influence each other. We start to understand that environment is not an independent issue; it is not that "we should take care of our environment", but we should take care of everything around as a whole. These interactions not only tells that environment is well involved in our daily life: corporation, governmental administration, market... and they also imply to us that we should both protect the environment directly and through a business aspect because at least corporation is inter-correlated with environment. Maybe other factors as well. Meanwhile, it is crucial to see how environmental issues alter how people act in a social manner. Conclusively, theses issues are seperately-considered issues, but in fact they are on entity. We should care about both, perhaps in a sense of a social equivalence, human-nature dynamic equilibrium, in order to finally achieve an improvement in both aspects.

6 Theory

6.1 Mean

Mean equals to the average value of data in your sample or population. Such value can always be basically categorized sample mean and population mean, and the sample mean is always used to estimate the situation of the whole population. The sample mean can be an unbaised estimator as it can be the least squares estimator according to the Gauss Markov theorem.

6.2 Standard Error

The estimator for standard error are defined as the following:

$$\hat{SE}(\bar{x}) = \frac{\sqrt{\bar{x}(1-\bar{x})}}{\sqrt{n-1}} \frac{\sqrt{N-n}}{\sqrt{N}} \tag{1}$$

6.3 (Multivariate) Linear Regression

Multivariate linear regression is used to estimate the association between multiple explanatory variables and a response variable. The model comes in the following format:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \tag{2}$$

Several important assumptions are made regarding to this model:

- The response variable is linearly dependent on all of covariates $x_1, ..., x_p$.
- All of the observations $y_1, ..., y_n$ are independent.
- x_j is uncorrelated to the error ϵ.
- The errorϵ is normally distributed.
- The covariates $x_1, ..., x_p$ are not highly correlated.
- The mean and variance of ϵ_i are the same for i.

6.4 Chi-squared Test

A chi-squared test is a hypothesis test used to determine whether a sample follows the chi-squared distribution. Researchers can use the chi-squared test to determine whether two groups of categorical data are statistically independent or not, and whether the data fall in bins are following the desired distribution. In this study, the chi-square test is used to determine if two variables are statistically independent or not. Then

H_0 : The two groups of data are independent from each other, that there is no statistical relationship between the variables,

H_1 : The two groups of data are dependent on each other, that there is some statistical relationship between the variables

6.5 Confidence Interval

This is a theory used to cover a range of possible values varying from the estimated mean. Using this theory, we can uncover an interval that we are highly confident that the population mean is within it if we repeatedly draw random samples. This is an estimate approach to predict true mean from samples.

$$samplemean + / - standarderror * t - score \tag{3}$$

6.6 Sobel Test

$H_0 : ab = 0$ (where a,b are indirect effects)

$$T = \frac{\hat{a}\hat{b}}{\sqrt{(\hat{b} * exp(2))(SE(\hat{a}) * exp(x) + (\hat{a} * exp(2))(SE(\hat{b}) * exp(x))}} \tag{4}$$

In the Sobel test, the null hypothesis is that the indirect effect of the mediation analysis yield 0. When the test statistic's absolute value is larger than 1.96, the null hypothesis will be rejected. If so, at a alpha level = 0.05, we say that we are 95% confident that the indirect effect is not equal to zero. In addition, this is a very conservative test.

6.7 Random Forest

A random forest predictor is an ensemble of individual decision tree predictors and is commonly used in unsupervised data. As part of the construction, random forest predictors naturally lead to a dissimilarity measure between the observations. One can also define an random forest dissimilarity measure between unlabeled data: the idea is to construct an random forest classifier that distinguishes the "observed" data from suitably generated synthetic data.

6.8 Mean Decreased Accuracy

In the feature selection by using the random forest, the mean decrease accuracy is one of the methods to order the importance of the variables. By derived variable importance index, it is possible to determine correlation coefficient related features are more significant than others.

7 Appendix

7.1 Contribution

1. Junqian Liu: Did question selection and data cleaning of water security data sets; wrote part of 4.2.2, Data, and Theory.
2. Shuyang Zhang: Did data cleaning of water security data sets, did random forest algorithms on water security data sets; wrote social and environmental correlation part of water security; did code merging; did part of the theory.
3. Yibei Cai: Did data cleaning of climate change data set, and did part 4.1.1 Multivariate Regression and 4.1.2 Examine Variable Energy Percent
4. Yiteng Lu: Did part 4.1.3, 4.1.4, 4.1.5, and wrote the conclusion
5. Yuan Lin: Did data cleaning of cities data sets. Wrote part 4.3 Cities and part of conclusion.
6. Zetong Lai: Wrote introduction, background and part of 4.2.2.

7.2 Citation

1. EPA Center for Corporate Climate Leadership. `https://www.epa.gov/climateleadership/scope-1-and-scope-2-inventory-guidance`
2. CDP - Unlocking Climate Solution. `https://www.kaggle.com/c/cdp-unlocking-climate-solutions/overview`

3. Pogutz, Stefano and Russo, Angeloantonio, Eco-Efficiency vs Eco-Effectiveness: Exploring the Link between GHG Emissions and Firm Performance (September 3, 2009). Academy of Management Annual Conference Best Paper Proceedings, Available at SSRN: https://ssrn.com/abstract=1467790 or http://dx.doi.org/10.2139/ssrn.1467790

4. Shi,Tao,and Steve Horvath. "Unsupervised learning with random forest predictors." nal of Computational and Graphical Statistics 15.1 (2006): 118-138.

5. Zineb Moumen,Najiba El Amrani El Idrissi,Manuela Tvaronaviienė, Abderrahim Lahrach.Water security and sustainable development.Insights into Regional Development,Entrepreneurship and Sustainability Center,2019,1(4),pp.301-317. 10.9770/ird.2019.1.4(2). hal-02342701

6. Lee, Hansoo, et al. "Case Dependent Feature Selection using Mean Decrease Accuracy for Convective Storm Identification." 2019 International Conference on Fuzzy Theory and Its Applications (iFUZZY). IEEE, 2019.