Title of Project:

Modelling Greenhouse Gas Emissions in the Canadian Industrial Sector using Economic out and Energy usage.

Full names

1. Frank Kwaku Kuukyee
2. Olayinka Soremekun

1. INTRODUCTION

1.1. MOTIVATION

1.1.1. Context

The environmental impact of industrial activities has become a focal point of global concern as nations strive to achieve sustainable development and mitigate climate change. The Canadian industrial sector holds a significant position due to its diverse and expansive industrial landscape. As industries play a crucial role in contributing to the nation's economic growth, there is an increasing need to comprehend and model the relationship between economic indicators, and environmental sustainability. Through this project, we endeavor to provide a comprehensive understanding of the two pivotal factors—Gross Domestic Product (GDP) and Energy Usage. By harnessing data-driven modeling techniques, we aim to unravel the complex interplay between economic activities, energy consumption, and their collective impact on greenhouse gas emissions. The insights derived from this analysis hold the potential to guide policy decisions, foster sustainable industrial practices, and pave the way for a more environmentally conscious industrial landscape in Canada.

1.1.2. Problem

The Canadian industrial sector plays a significant role in national economic growth, but it also contributes to greenhouse gas emissions. This study seeks to address the challenge of quantifying and predicting these emissions based on economic activities and energy consumption.

1.2. OBJECTIVES

1.2.1. Overview

This project work sought to evaluate the interaction between the economic output and energy consumption, and their collective impact on Greenhouse gas emissions. And also, develop the best regression model to predict Greenhouse gas emissions using Gross Domestic product (GDP) and Energy usage.

1.2.2 Goals & Research Questions

Goals:

The goal of the project work is to determine the statistically significant main effect, higher order and interactive terms to develop a reliable Multiple Linear Regression model for predicting greenhouse gas emissions.

Research Questions:

The project work addressed the questions presented below;

1. What is the statistical significance of the individual predictors?
2. What is the statistical significance of the interaction between GDP and energy consumption for the prediction of Greenhouse gas emissions?
3. What is the best model for predicting Greenhouse gas emissions?

2. METHODOLOGY

2.1 Data

Our data was obtained from the website of Natural Resources Canada. Data was collected on Green House Gas emission measured in Metric Tonnes of CO2 equivalent [Mt.CO2e], Energy consumption in Peta Joules[PJ] and Activity in Gross Domestic Product (GDP) for the major industries in Canada between 2017 and 2020.The Industries covered based on the classification by Natural Resources Canada are: Iron Mines, Gold and Silver Mines, Potash Mines, Upstream Mining, Pulp Mills, Paper and newsprint, Petroleum Refining/petrochemical industry, All other basic inorganic chemical manufacturing, Chemical fertilizer (except potash) manufacturing, Other Chemical Manufacturing, Cement Industry, Iron and Steel, Primary Production of Alumina and

Aluminum, Other Non-Ferrous Smelting and Refining, Motor Vehicle Industry, Other Manufacturing, Construction, and Forestry.

Averages values were calculated for the Green House Gas emission, Energy consumption and GDP between 2017 and 2020. The Greenhouse gas emissions [Mt.CO2e] were converted to CO2 equivalent [CO2e] to increase the significant figures. The period was chosen because of data availability and consistent data points. The final data used for the analysis is in the file GHGas3.csv attached as appendix 1. A snippet of the data set is shown below:

	Industry <chr>	Emissions <dbl>	Energy <dbl>	GDP <dbl>
1	Iron Mines	2735.994	36.84	6683.25
2	Gold and Silver Mines	3521.629	48.43	11615.50
3	Potash Mines	2522.204	33.86	7933.75
4	Upstream Mining	73228.501	851.20	121019.67
5	Pulp Mills	4418.469	310.37	1185.75
6	Paper and newsprint	3567.012	240.18	2166.00

2.1.1 Model Specification and Variable Explanations

All variables are reported annually with units shown in square brackets. The full model was specified using the following variables;

- Dependent variable:(Y): Emissions [CO2e] (continuous variable)
- Independent variables:
    - Energy [PJ] (continuous variable)
    - GDP [million $2012] (continuous variable)

2.2 Approach

We planned to approach this project using the methods we have learned in Data 603. We have chosen the Multiple Linear Regression approach as it allows us to model the relationships between multiple predictors and the response variable (greenhouse gas emissions). We expect this approach to provide insights into the joint influence of economic output and energy usage on emissions.

2.3 Workflow

1. Data Preprocessing: Handle missing values
2. Exploratory Data Analysis (EDA): Understand data distributions, identify outliers
3. Model Selection: Identify the most relevant predictors for inclusion in the regression model.
4. Model Evaluation: Assess model performance using relevant metrics such as R-squared, Mean Squared Error, individual T-test and partial F-test.
5. Interpretation of Results: Analyze coefficients and make inferences about the impact of economic output and energy usage on greenhouse gas emissions

2.4 Workload Distribution

Frank Kuukyee: Data preprocessing, model selection and interpretation of results.

Olayinka Soremekun: Exploratory Data Analysis, evaluation, and conclusion.

3 MAIN RESULTS OF THE ANALYSIS

A first-order model was built which consisted of every explanatory variable as a base comparison as shown below. This was helpful in selecting different variables through various selection procedures.

3.1 Results

3.1.1 Full model

```
Call:
lm(formula = Emissions ~ Energy + GDP, data = GHGemmisions)

Residuals:
   Min     1Q Median     3Q    Max
-12393  -3641   2303   3633   8954

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.446e+03  1.852e+03  -2.401   0.0298 *
Energy       6.815e+01  7.951e+00   8.571 3.64e-07 ***
GDP          8.849e-02  3.832e-02   2.309   0.0356 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5855 on 15 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.8899,    Adjusted R-squared:  0.8752
F-statistic: 60.61 on 2 and 15 DF,  p-value: 6.517e-08
```

3.1.2 Estimated model

$$\widehat{Emissions} = (-4.446e+03) + (6.815e+0)Energy_i + (8.849e-02)GDP_i$$

Hypothesis Statement for Individual T-tests:

H (0): βi = 0

H(A): βi ≠ 0, i = Energy, and GDP

Main Effects Individual T-tests:

Energy: t= 8.571, p= 3.64e-07

GDP: t= 2.309, p= 0.0356

Individual T-tests were used in our variable selection to determine the best predictors based on a significance level of α = 0.05. From the results of these tests, we reject the null hypothesis in favor of the alternative. This suggests that Energy consumption and GDP are significant predictors of Greenhouse Gas Emissions.

3.1.3 Higher order model

```
Call:
lm(formula = Emissions ~ Energy + I(Energy^2), data = GHGemmisions)

Residuals:
    Min      1Q  Median      3Q     Max
-6499.8  -931.0  -509.7   758.8  8055.5

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 3716.60084 1420.33761   2.617   0.0194 *
Energy       -10.16138   12.31975  -0.825   0.4224
I(Energy^2)    0.10750    0.01454   7.395 2.24e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3162 on 15 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.9679,    Adjusted R-squared:  0.9636
F-statistic: 225.9 on 2 and 15 DF,  p-value: 6.334e-12
```

Hypothesis Statement for Individual T-tests (Higher Order Terms):

H(0) : βi = 0

H(A) : βi ≠ 0, i = Energy^2,

Higher order Individual T-tests:

(Energy^2) : t= 7.395, p= 2.24e-06

From the p value we will reject the null and add (Energy^2) to our model. Our higher order model is below:

$$\widehat{Emissions} = (3716.60084) + (-10.16138)Energy_i + (0.10750)Energy_i^2$$

```
Analysis of Variance Table

Model 1: Emissions ~ Energy + I(Energy^2)
Model 2: Emissions ~ Energy + GDP
  Res.Df       RSS Df  Sum of Sq F Pr(>F)
1     15 150014738
2     15 514179436  0 -364164697
```

Hypothesis Statement for ANOVA Test:

H (0): $\beta p-q+1 = \beta p-q+2 = ... = \beta p = 0$: Higher order terms are not significant

H(A): at least one $\beta p \neq 0$:

There was no p-value and F-values from the ANOVA so we could not conclude on the significance of the Higher order model.

3.1.4 Interactive model

```
Call:
lm(formula = Emissions ~ (Energy + GDP)^2, data = GHGemmisions)

Residuals:
    Min      1Q  Median      3Q     Max
-5496.5  -625.7   -72.1   852.4  7286.0

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.209e+03  1.285e+03   0.941  0.36285
Energy       2.753e+01  7.432e+00   3.704  0.00236 **
GDP         -1.661e-02  2.541e-02  -0.654  0.52381
Energy:GDP   4.916e-04  7.514e-05   6.543 1.31e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3008 on 14 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.9729,    Adjusted R-squared:  0.967
F-statistic: 167.3 on 3 and 14 DF,  p-value: 3.365e-11
```

Hypothesis Statement for Individual T-tests (Interaction Terms):

H (0): $\beta i = 0$

H(A): $\beta i \neq 0$

i = *Energy: GDP*

Interaction Term T-tests:

Energy*GDP: t = 6.543, p = 1.31e-05

From the results of these tests, we would reject the null hypothesis in favor of the alternative. This suggests that Energy*GDP is significant predictor of Greenhouse Gas Emissions. The interaction model is shown below.

$$\widehat{Emissions} = 1209 + 27.53Energy_i - 0.01661GDP_i + 0.0004916Energy * GDP_i$$

```
Analysis of Variance Table

Model 1: Emissions ~ (Energy + GDP)^2
Model 2: Emissions ~ Energy + GDP
  Res.Df       RSS Df  Sum of Sq      F    Pr(>F)
1     14 126709714
2     15 514179436 -1 -387469722 42.811 1.306e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hypothesis Statement for ANOVA Test:

H(0) : βp−q+1 = βp−q+2 = ... = βp = 0 : Interactive terms are not significant

H(A) : at least one βp ≠ 0 : Interactive terms are significant

F=42.811, p-value is very small (1.306e-05) compared to significance level of 0.05.

Therefore, we reject the null hypothesis and conclude that Interactive terms are significant. The Interactive model does contribute significantly to explaining the variance in the response variable (Emissions) compared to the Full Model. Our best model for the prediction Greenhouse emissions is shown below.

$$\widehat{Emissions} = 1209 + 27.53Energy_i - 0.01661GDP_i + 0.0004916Energy * GDP_i$$

3.1.5 Interpretations of coefficients

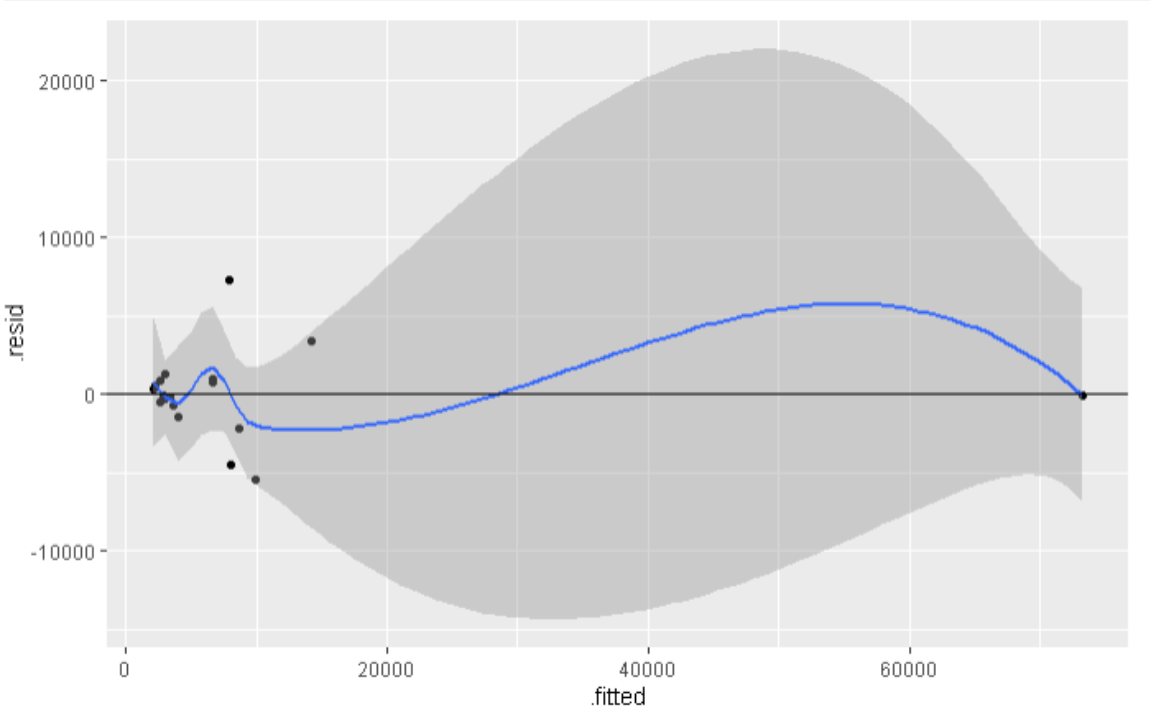
- Intercept (β0): The estimated intercept is 1209. However, its p-value is 0.36285, which is greater than the significance level of 0.05. This suggests that the intercept is

not significantly different from zero, and it might not be a meaningful parameter in explaining the variance in emissions.

- Energy (β1): The estimated coefficient for Energy is 27.53. This implies that, holding GDP constant, a one PJ increase in Energy consumption is associated with an estimated increase of 27.53 CO2e in Emissions. The p-value (0.00236) is less than 0.05, indicating that the coefficient for Energy is statistically significant.
- GDP (β2): The estimated coefficient for GDP is r -0.01661. This suggests that, holding Energy constant, a million $2012 increase in GDP is associated with an estimated decrease of 0.01661 CO2e in Emissions. However, the p-value (0.52381) is greater than 0.05, indicating that the coefficient for GDP is not statistically significant. Therefore, GDP might not be a significant predictor in the model

3.1.6 Testing for the Assumptions

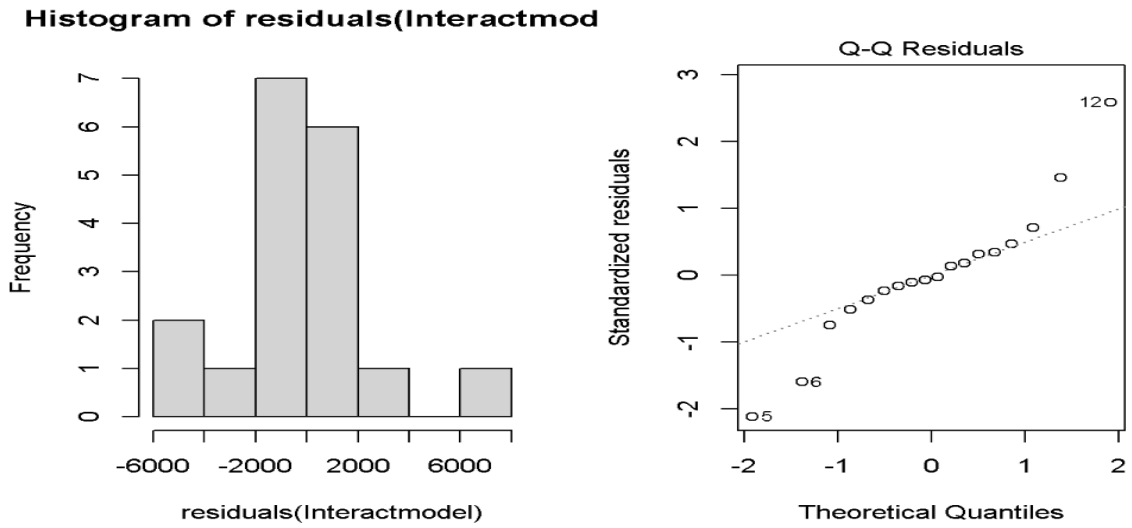
1. Linearity and Independence



Using residual plots as shown above we check to see if there are any discernible patterns that are non-linear. From the plot, we see that there are no prominent patterns showing in the trend of our data, suggesting that it passes the linearity and independence assumptions.

2. Normality

We can see that the distribution of the residuals in the histogram follows a fairly normal trend with some data points occurring near the tail ends. Additionally, a normal probability plot of residuals is provided. Again, we see that most of the data points approximate the normal line, however, there are a few points flaring outwards near the tails indicating the presence of possible outliers.



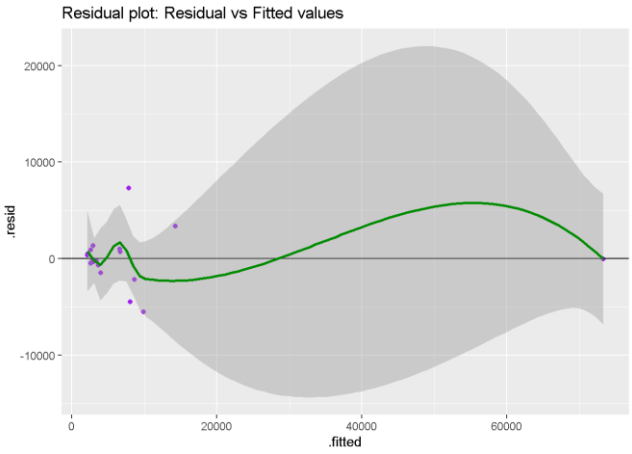
Testing for Normality using Shapiro-Wilk test:

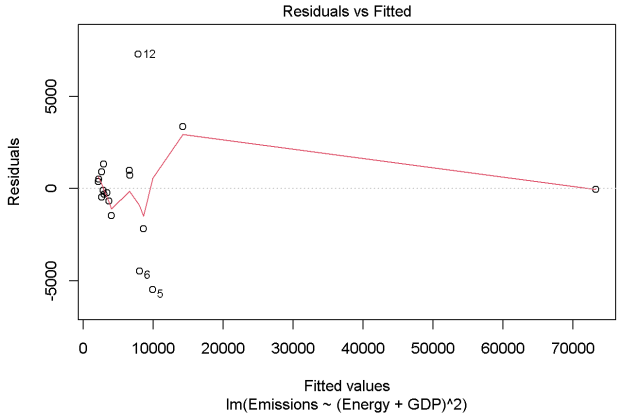
Null Hypothesis H(0): The sample data is normally distributed

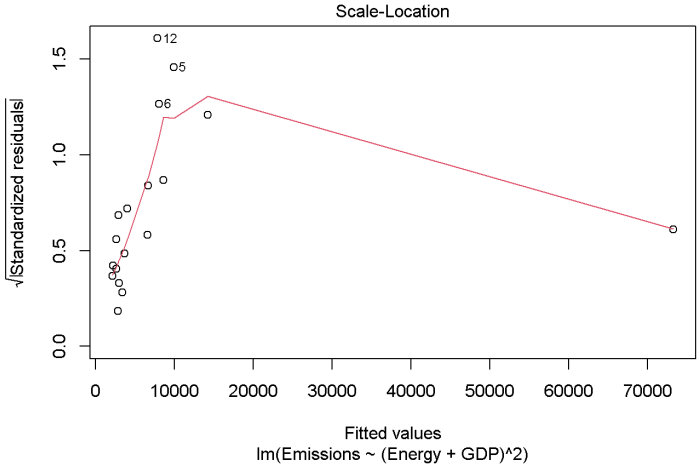
Alt. Hypothesis H(A): The sample data is not normally distributed

Based on $\alpha = 0.05$, the results of the Shapiro-Wilks test ($W = 0.89914$, $p = 0.05552$. We fail to reject the null hypothesis. Data meets the normality condition

3. Equal Variance Assumption







Test for Heteroscedascity using the studentized Breusch-Pagan test:

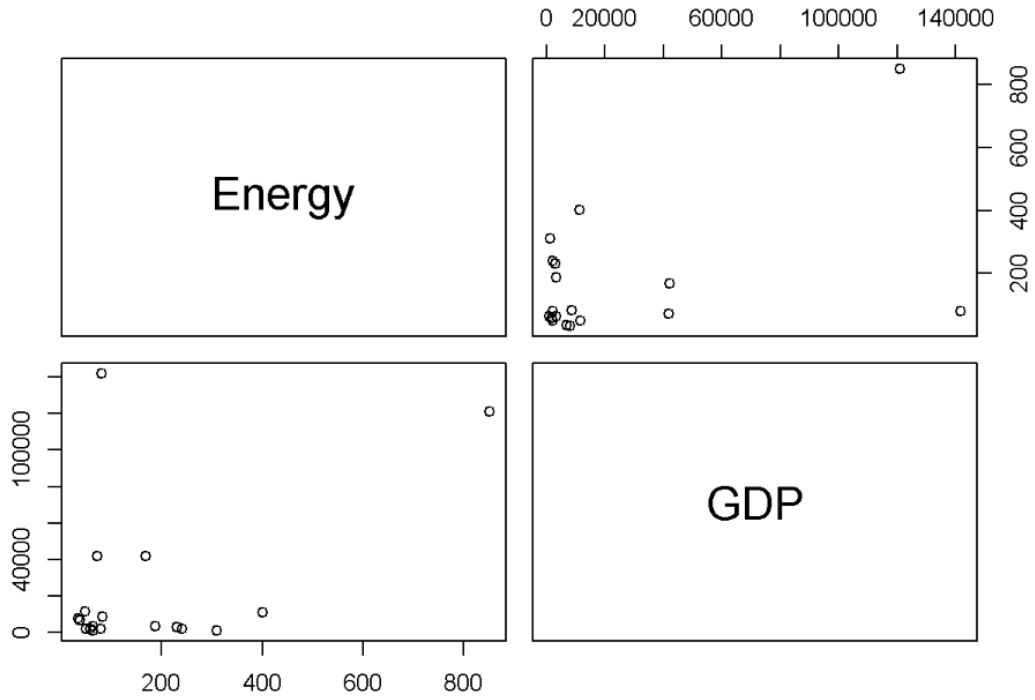
Null Hypothesis H (0): Heteroscedascity is not present

Alt. Hypothesis H(A): Heteroscedascity is present

We tested to see if our data is homoscedastic through a plot of fits to residuals as well as the studentized Breusch-Pagan test. Looking at the plot of fits to residuals, we see that there is no pattern to the plotted data as the fitted values increase. This is an indicator that the data have common variance. From the results of the Breusch-Pagan test (BP =8.0486, p = 0.04502), we fail to reject the null hypothesis in favor of the alternative, suggesting that our model is homoscedastic.

4. Multi-colinearity

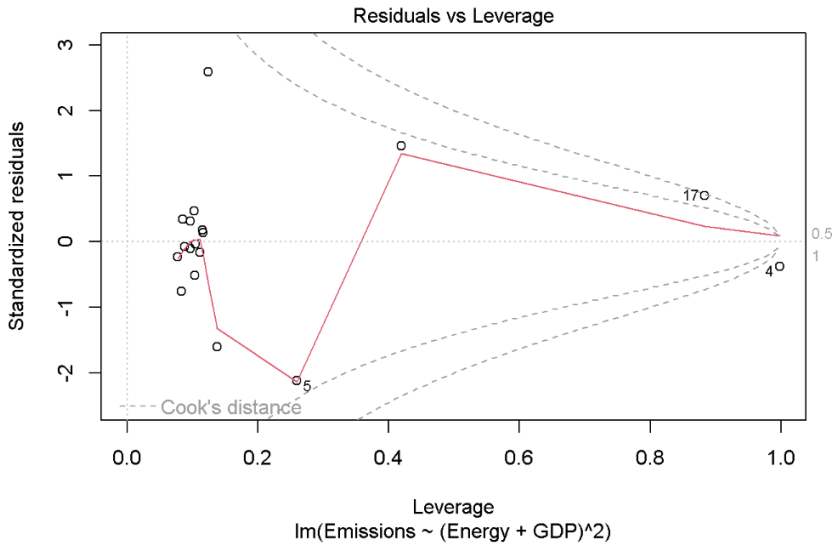
To test for multicollinearity in our models, we looked at pair plots and multiple variance inflation



Testing for Multicolinearity

```
##
## Call:
## imcdiag(mod = Interactmodel, method = "VIF")
##
##
##  VIF Multicollinearity Diagnostics
##
##                VIF detection
## Energy     4.1510         0
## GDP        2.0893         0
## Energy:GDP 6.1322         0
##
## NOTE:  VIF Method Failed to detect multicollinearity
##
##
## 0 --> COLLINEARITY is not detected by the test
##
```

5. Influential Points and Outliers



From the residuals vs leverage and the other plots, we see that there are no points beyond Cook's distance. There however appears to be one influential point that that represented the Upstream mining sector which is the highest contributor in Greenhouse gas emissions and Energy consumption.

4 CONCLUSION AND DISCUSSION

4.1 Approach

The Multiple Linear Regression model, built on the relationship between greenhouse gas emissions (Emissions), energy consumption (Energy), and economic output (GDP), yielded insightful results. The model's output, demonstrates the significant impact of both Energy and the interaction term Energy: GDP on greenhouse gas emissions.

The coefficients indicate that both Energy and the interaction term Energy: GDP have statistically significant effects on greenhouse gas emissions. The Adjusted R-squared of 0.967 demonstrates the model's high explanatory power, capturing approximately 96.7% of the variance in emissions.

4.2 Future Work

The success of the model opens avenues for future work. While the present study focused on the main predictors, additional variables such as industrial sector specifics and technological advancements could further enhance the model's accuracy. Continuous monitoring and updating of the model with new data will ensure its relevance and applicability in a dynamic industrial landscape.

Moreover, the model's adherence to various assumptions, including linearity, independence, normality, equal variance, multicollinearity, and outliers, provides confidence in its robustness. However, further exploration and sensitivity analyses can strengthen the model's resilience and uncover potential refinements.

The implications of this research extend beyond academic interest, offering valuable insights for policymakers, industry leaders, and environmentalists. By understanding the intricate relationships between economic activities, energy usage, and greenhouse gas emissions, stakeholders can formulate informed strategies to balance industrial growth with environmental sustainability. The model provides a practical tool for scenario analysis and decision-making, contributing to a greener and more sustainable Canadian industrial sector.

5 REFERENCES

Canada, E. A. C. C. (2023). *Greenhouse gas emissions*. Canada.ca. https://www.canada.ca/en/environment-climate-change/services/environmental-indicators/greenhouse-gas-emissions.html

Division, N. R. C.-. O. O. E. E.-. D. P. a. A. (2023). *Industrial Sector – Disaggregated Industries Canada Table 1: Secondary energy use by energy source*. Natural Resources Canada. https://oee.nrcan.gc.ca/corporate/statistics/neud/dpa/showTable.cfm?type=CP§or=id&juris=ca&year=2020&rn=1&page=0

6 APPENDIX

R-codes used

```
GHGemmisions=read.csv('C:/Users/HP/Desktop/Frank Kuukyee 602 Assigment 3/GHGas3.csv')

head(GHGemmisions)

 #fitting the full model

Fullmodel<-lm(Emissions~Energy+GDP, data = GHGemmisions)

summary(Fullmodel)

#fitting the full model

Fullmodel<-lm(Emissions~Energy+GDP, data = GHGemmisions)

summary(Fullmodel)

#fitting the higher order model

Higherordermodel<-lm(Emissions~Energy+GDP+I(Energy^2)+I(GDP^2), data = GHGemmisions)

summary(Higherordermodel)

#fitting the higher order model

Higherordermodel2<-lm(Emissions~Energy+I(Energy^2)+I(Energy^3), data = GHGemmisions)

summary(Higherordermodel2)

# fitting the higher order model

Higherordermodel3<-lm(Emissions~Energy+I(Energy^2), data = GHGemmisions)

summary(Higherordermodel3)

#A partial F-test for the Higher order model against the Full model

anova(Higherordermodel3,Fullmodel)

#fitting the interactive model

Interactmodel<-lm(Emissions~(Energy+GDP)^2, data = GHGemmisions)

summary(Interactmodel)

#A partial F-test for the Interactive model against the Full model

anova(Interactmodel,Fullmodel)

#Residual plts foe checking linearity

ggplot(Interactmodel, aes(x=.fitted, y=.resid)) +

  geom_point() +geom_smooth()+

  geom_hline(yintercept = 0)

#Residual plts foe checking linearity
```

```
ggplot(Interactmodel, aes(x=.fitted, y=.resid)) +
  geom_point() +geom_smooth()+
  geom_hline(yintercept = 0)
#Residual plts foe checking linearity
ggplot(Interactmodel, aes(x=.fitted, y=.resid)) +
  geom_point() +geom_smooth()+
  geom_hline(yintercept = 0)
# (histogram)
par(mfrow=c(1,2))
hist(residuals(Interactmodel))
plot(Interactmodel, which=2) #a Normal plot
#Testing for Normality
shapiro.test(residuals(Interactmodel))
#Residual plot
ggplot(Interactmodel, aes(x=.fitted, y=.resid)) +
  geom_point(colour = "purple") +
  geom_hline(yintercept = 0) +
  geom_smooth(colour = "green4")+
  ggtitle("Residual plot: Residual vs Fitted values")
#scale location
ggplot(Interactmodel, aes(x=.fitted, y=sqrt(abs(.stdresid)))) +
geom_point(colour = "purple") +
geom_hline(yintercept = 0) +
geom_smooth( colour = "green4")+
ggtitle("Scale-Location plot : Standardized Residual vs Fitted values")

#residuals plot
plot(Interactmodel, which=3)
#
bptest(Interactmodel)
#pair plot
pairs(~Energy+GDP+Energy*GDP, data=GHGemmisions)
imcdiag(Interactmodel, method="VIF")
```

```
plot(Interactmodel,which=5)
```