Big Data Project

For my Big Data project, I decided to focus on the topic of climate change. This has been a topic I have been fascinated with since I was a kid and one I continue to be passionate about to this day. My motivation for this project was to see for myself what I can learn from the data and to see how my conclusions compare to the experts. If a topic is important enough to you, you should try and do your own research on it and that was my goal.

I started my analysis by looking into the temperature trends over the past 50 years. After deploying a seasonal ARIMA forecast on the temperature trends, I was able to quickly predict and visualize the increasing temperature trend into the next 20 years. The goal of this plot was to help me introduce the topic/issue with a simple visual I created (https://imgur.com/a/E4lzPYu). This model was more for the introduction for my presentation, for my analysis I wanted to explore the relationship between different emission statistics and temperature change.

I used two datasets for this analysis, one on temperature change and one on emission statistics. The temperature change dataset was sourced from the National Aeronautics and Space Administration Goddard Institute for Space Studies (NASA-GISS) and distributed by the Food and Agriculture Organization Corporate Statistical Database (FAOSTAT). This dataset has average temperature change data year over year from 1961 to 2019 in 190 countries. The emissions dataset was sourced from the Global Carbon Project, an international scientific project aiming to quantify and understand the global carbon cycle and its interactions with Earth's climate. This dataset had data from 1750 to 2020 tracking country-specific emissions statistics by year. It had data for coal, oil, gas, cement, flaring, and a column for per capita C02 emissions. Before I began modeling, I first had to prepare the data. I merged these two datasets by the country and year columns that they shared. I had to cut off the data from 1750-1961 in the emission dataset in order for it to match, a lot of the data from this cutoff period was missing anyways. After joining the data I had to figure out how I would deal with the missing data. After graphing the percent of missing data from each column I found that each of the relevant columns had about a quarter missing. I also graphed the percent of missing data over the time component to see if there was anything significant there. I found that it hovered around 24 to 28 percent until 1991 when it somewhat drastically dipped to about 18 percent, where it hovered around until another big dip for the last 2 years of data. I considered dropping the earlier portion but ultimately decided against it, I decided the dip was not drastic enough to offset the amount of

data I would be losing. My goal with this combined dataset was to create various models to see what emissions statistics had better predictive power for temperature change.

Initially, my models were based on the notion that each data point would be like it was in the merged dataset, a country in a year. Through this initial stage, I implemented eleven models with different methods. Some of the methods/strategies I explored were: multiple linear regression, using interaction terms, introducing and withholding 'Year' variable, using different variations of lagged components, Lasso Regression, and Random Forests. The multiple linear regression models had little success, with less than a 10% explained variance. Ultimately the best models ended up being the Random Forest models and the models taking the lagged 5, 10, and 25-year averages.

Next, I decided to switch the types of models I was doing to ones that did not incorporate the country variable but instead just look at global yearly averages across all the statistics. I made this switch because emissions from one region don't just affect the temperature in that region but the whole globe. I ended up deploying basically the same models as I did in the last part. These models performed, expectedly, much better than the ones from before, with all of them being able to explain at least 85% of the variance. As far as interpretability goes for these models, looking at the emissions statistics, the coal and cement predictors were pretty consistently more significant than the rest. But, the coal predictor often had a negative coefficient, which intuitively does not make sense.

There are definitely some issues with the analysis I attempted to do. First off, the first part of my analysis did not work as I intended it to. As I said before, this was because of the lack of geographic impact of the emissions on temperature. This was something I should have thought about in more depth prior to my analysis. Secondly, my lack of experience with time series analysis made it difficult for me to navigate model selection. This semester we had primarily focused on different types of problems more suited to problems based on prediction that did not involve this type of time series component. I knew this going into it, but I underestimated the challenge I would have working through this problem. Thirdly, additional data could have been introduced to help me better capture the true relationship between the temperature and the emission variables. Variables like population, solar radiation, and GDP could have potentially been helpful to include.

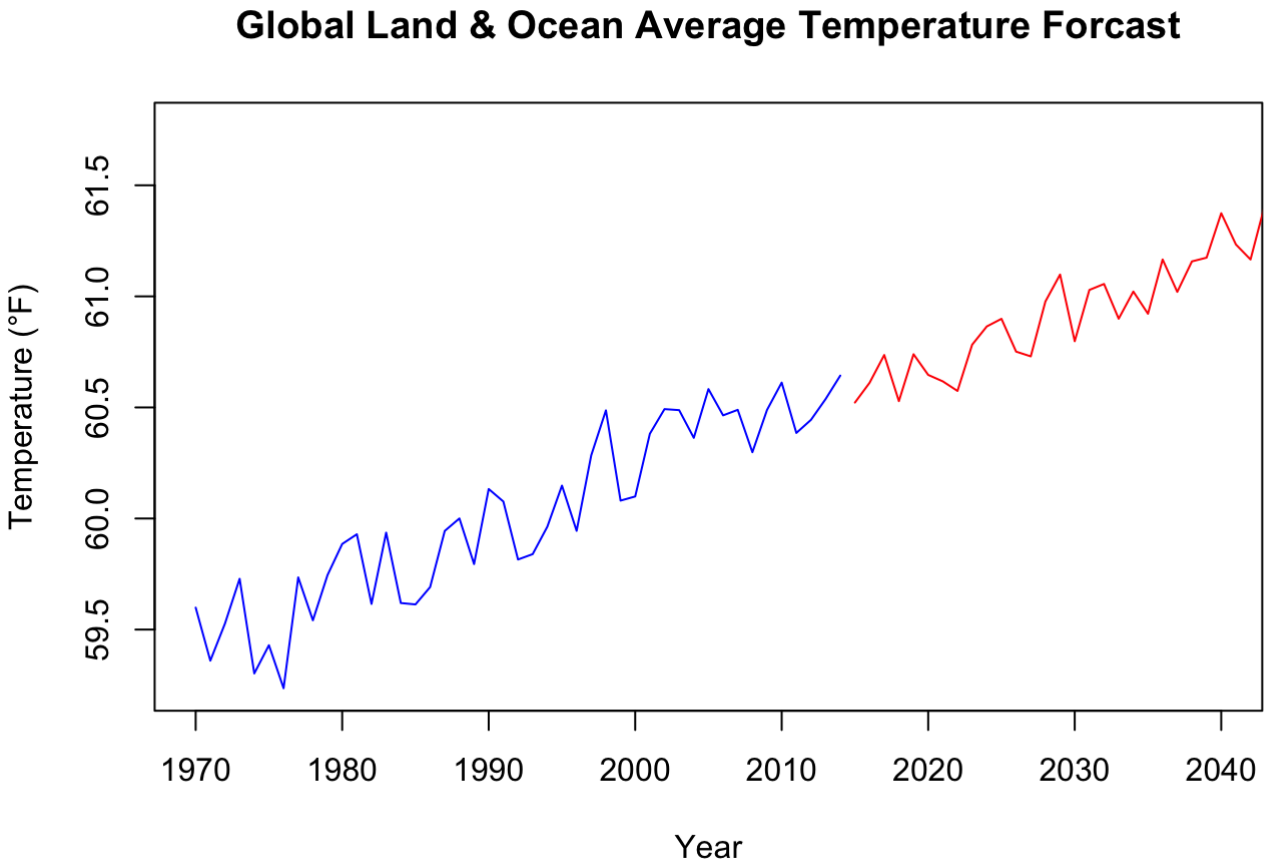
After facing these shortcomings in my analysis I decided to do some additional analysis on some climate change-related questions so I could get more out of this project. First, I gathered a dataset on hurricanes and typhoons in the past couple of hundred years to see if I could find a relationship there with temperature change. I saw that they did indeed follow a similar positive trend and I deployed a simple linear regression model that expressed some significance in that relationship. But, that obviously does not imply causation. Secondly, I took a look at some of the impactful climate change events and movements to see what effect they had on emissions and temperature data. There ended up not being a visible relationship in those spots. That is not to say that the events had no impact, it just can be tough for things like those to stand out against a very macro-level variable.

Going into this analysis my goal was to explore a topic that is important to me. While I did not end up getting the results I was hoping for, I still gained some valuable insights about the issue and learned more about the challenge of real-world data analysis. Looking back, I probably should have planned everything out better from the data selection stage. I was a bit over my head with the analysis I was attempting to do. But, I still was able to gain insights about climate trends that I would not have reached without this project. Ultimately, this was a great learning experience and I learned a lot about problem-solving, statistics, and planet Earth.

Climate Change Analysis (Models and Plots)

Benjamin Coleman

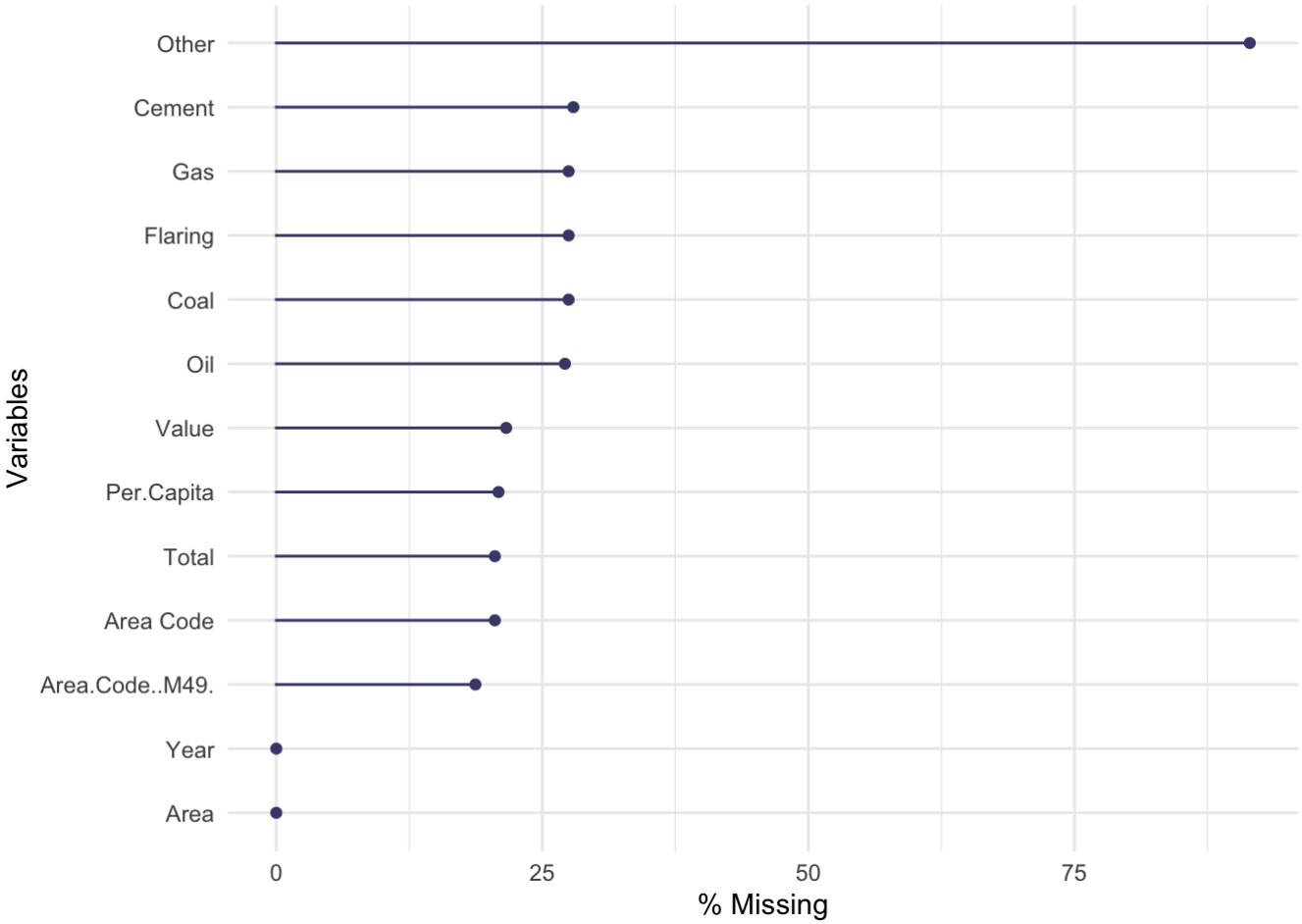
2023

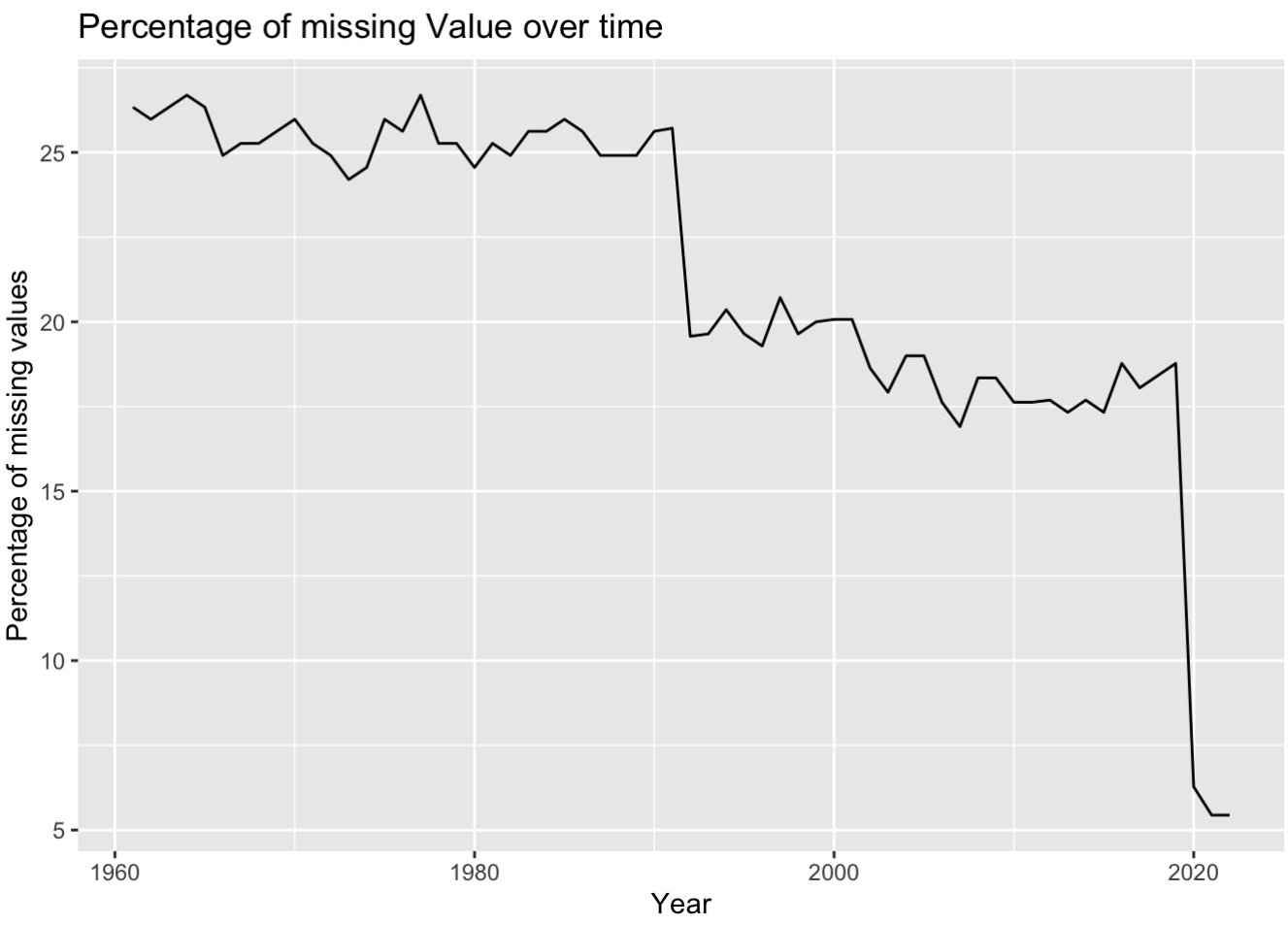


This graph was mostly for the presentation. The goal was to use it to illustrate the fact that following historical temperature trends, the temperature will continue to rise. The red line is the forecast.

Statistical methods used: Seasonal ARIMA model trained on annual average temperature data from 1970-2014.

Visualizing missing data


Other
Cement
Gas
Flaring
Coal
Oil
Value
Per.Capita
Total
Area Code
Area.Code..M49.
Year
Area
Variables
0
25
50
75
% Missing


Percentage of missing Value over time
25
20
15
10
5
Percentage of missing values
1960
1980
2000
2020
Year

- These helped me view the trend of missing data and evaluate the best way to handle it.

Statistical Modeling - learn more about emissions impact on temperature change

Multiple Linear Regression Model, Predictors: Cement, Gas, Flaring, Coal, Oil

```
##
## Call:
## lm(formula = Value ~ Cement + Gas + Flaring + Coal + Oil, data = data_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5343 -0.4242 -0.0498  0.3781  2.5862
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.728e-01  6.721e-03  70.346  < 2e-16 ***
## Cement       2.446e-03  9.454e-04   2.587  0.00969 **
## Gas          6.847e-03  3.230e-04  21.203  < 2e-16 ***
## Flaring     -2.075e-03  1.538e-03  -1.349  0.17752
## Coal        -1.730e-04  9.136e-05  -1.893  0.05838 .
## Oil         -1.419e-03  1.396e-04 -10.171  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6054 on 9527 degrees of freedom
## Multiple R-squared:  0.05383,    Adjusted R-squared:  0.05333
## F-statistic: 108.4 on 5 and 9527 DF,  p-value: < 2.2e-16
```

- These predictors (Cement, Gas, Flaring, Coal, Oil) were able to explain about 5.3% of the variance in annual temperature.
- Cement, Gas, and Oil were statistically significant predictors ($p<0.05$).

Multiple Linear Regression Model with Interactions, Predictors: Cement, Gas, Flaring, Coal, Oil

```
##
## Call:
## lm(formula = Value ~ Cement * Gas * Flaring * Coal * Oil, data = data_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.52468 -0.41761 -0.04749  0.37288  2.59636
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  4.639e-01  7.175e-03  64.654  < 2e-16 ***
## Cement                       2.975e-02  5.272e-03   5.643 1.71e-08 ***
## Gas                          1.020e-02  6.627e-04  15.386  < 2e-16 ***
## Flaring                     -1.408e-02  2.512e-03  -5.605 2.15e-08 ***
## Coal                        -8.379e-04  2.303e-04  -3.638 0.000277 ***
## Oil                         -3.274e-03  3.493e-04  -9.374  < 2e-16 ***
## Cement:Gas                  -4.348e-04  9.625e-05  -4.517 6.34e-06 ***
## Cement:Flaring              -3.420e-03  1.642e-03  -2.084 0.037221 *
## Gas:Flaring                 -3.790e-05  1.712e-04  -0.221 0.824861
## Cement:Coal                 -9.977e-06  3.005e-06  -3.320 0.000902 ***
## Gas:Coal                     2.132e-06  5.946e-06   0.359 0.719893
## Flaring:Coal                 3.372e-04  8.082e-05   4.172 3.04e-05 ***
## Cement:Oil                  -1.342e-05  1.160e-05  -1.157 0.247460
## Gas:Oil                      5.400e-06  3.602e-06   1.499 0.133814
## Flaring:Oil                  3.354e-04  6.723e-05   4.989 6.16e-07 ***
## Coal:Oil                     3.369e-06  1.243e-06   2.711 0.006719 **
## Cement:Gas:Flaring           6.607e-05  2.390e-05   2.764 0.005718 **
## Cement:Gas:Coal              1.193e-07  2.352e-08   5.073 3.99e-07 ***
## Cement:Flaring:Coal          4.149e-06  1.530e-06   2.711 0.006728 **
## Gas:Flaring:Coal            -1.014e-06  1.782e-06  -0.569 0.569163
## Cement:Gas:Oil               6.375e-07  1.641e-07   3.884 0.000103 ***
## Cement:Flaring:Oil           7.787e-06  5.806e-06   1.341 0.179901
## Gas:Flaring:Oil             -2.943e-06  7.939e-07  -3.706 0.000212 ***
## Cement:Coal:Oil              1.599e-09  2.460e-09   0.650 0.515677
## Gas:Coal:Oil                -3.079e-08  1.636e-08  -1.882 0.059842 .
## Flaring:Coal:Oil            -1.980e-06  5.494e-07  -3.605 0.000314 ***
## Cement:Gas:Flaring:Coal     -2.710e-08  6.360e-09  -4.261 2.06e-05 ***
## Cement:Gas:Flaring:Oil      -1.357e-07  6.359e-08  -2.134 0.032874 *
## Cement:Gas:Coal:Oil         -9.253e-11  1.866e-11  -4.959 7.22e-07 ***
## Cement:Flaring:Coal:Oil     -1.634e-09  1.174e-09  -1.392 0.164045
## Gas:Flaring:Coal:Oil         1.248e-08  5.059e-09   2.466 0.013678 *
## Cement:Gas:Flaring:Coal:Oil  1.750e-11  5.344e-12   3.275 0.001061 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5995 on 9501 degrees of freedom
## Multiple R-squared:  0.07476,    Adjusted R-squared:  0.07174
## F-statistic: 24.76 on 31 and 9501 DF,  p-value: < 2.2e-16
```

- Introducing these interaction terms increased the explained variance to about 7.5%.
- Many of the interaction terms were found to be statistically significant.

Multiple Linear Regression, introducing time component with the 'Year' variable

```
##
## Call:
## lm(formula = Value ~ Cement + Gas + Flaring + Coal + Oil + Year,
##     data = data_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.34229 -0.26716 -0.02171  0.24211  2.16598
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.939e+01  5.503e-01 -89.745  < 2e-16 ***
## Cement      -8.313e-04  6.939e-04  -1.198    0.231
## Gas          1.869e-03  2.430e-04   7.691 1.60e-14 ***
## Flaring      3.019e-04  1.128e-03   0.268    0.789
## Coal         9.440e-05  6.703e-05   1.408    0.159
## Oil         -4.288e-04  1.029e-04  -4.169 3.09e-05 ***
## Year         2.505e-02  2.764e-04  90.607  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4437 on 9526 degrees of freedom
## Multiple R-squared:  0.4918, Adjusted R-squared:  0.4915
## F-statistic:  1536 on 6 and 9526 DF,  p-value: < 2.2e-16
```

- The addition of the time variable significantly increased the explained variance to about 49.2%.
- Gas, Oil, and Year were statistically significant predictors.

Multiple Linear Regression Model with Lagged Effects

```
##
## Call:
## lm(formula = Value ~ Cement + Gas + Flaring + Coal + Oil + Cement_lag +
##     Gas_lag + Flaring_lag + Coal_lag + Oil_lag, data = data_clean_lag)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.64209 -0.41728 -0.04629  0.37782  2.57981
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.4793710  0.0067906  70.593  < 2e-16 ***
## Cement       0.0194316  0.0057936   3.354 0.000800 ***
## Gas         -0.0047689  0.0024385  -1.956 0.050535 .
## Flaring     -0.0017076  0.0053079  -0.322 0.747674
## Coal        -0.0001595  0.0005598  -0.285 0.775757
## Oil         -0.0041861  0.0012466  -3.358 0.000788 ***
## Cement_lag  -0.0161974  0.0058856  -2.752 0.005934 **
## Gas_lag      0.0119270  0.0024714   4.826 1.41e-06 ***
## Flaring_lag -0.0002617  0.0053503  -0.049 0.960985
## Coal_lag    -0.0001050  0.0005692  -0.185 0.853616
## Oil_lag      0.0027931  0.0012700   2.199 0.027874 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6042 on 9329 degrees of freedom
## Multiple R-squared:  0.05876,    Adjusted R-squared:  0.05775
## F-statistic: 58.24 on 10 and 9329 DF,  p-value: < 2.2e-16
```

- The multiple linear regression model with lagged emissions predictors explained about 5.9% of the variance.
- Cement, Oil, Cement_lag, and Gas_lag were statistically significant predictors.

Random Forest Model, Predictors: Cement, Gas, Flaring, Coal, Oil

```
##
## Call:
##  randomForest(formula = Value ~ Cement + Gas + Flaring + Coal +      Oil, data = data
_clean)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 1
##
##           Mean of squared residuals: 0.2523093
##                     % Var explained: 34.82
```

- The Random Forest model explained 34.89% of the variance in Value, which is a good improvement over the linear models without the time component, but less than the linear model with the time component.
- The Mean Squared Error is about 0.252.

Lasso Regression Model, Predictors: Cement, Gas, Flaring, Coal, Oil

```
##
## Call:  cv.glmnet(x = x, y = y, alpha = 1)
##
## Measure: Mean-Squared Error
##
##      Lambda Index Measure       SE Nonzero
## min 0.00007    82  0.3669 0.008873       5
## 1se 0.05866     9  0.3754 0.008837       1
```

```
## 6 x 1 sparse Matrix of class "dgCMatrix"
##                        s1
## (Intercept)  0.4726899238
## Cement       0.0022566566
## Gas          0.0068413064
## Flaring     -0.0020492448
## Coal        -0.0001555264
## Oil         -0.0014154996
```

- The Lasso model selected 5 non-zero coefficients, with a Mean Squared Error of about 0.3669 at the best lambda (regularization parameter).
- At a slightly larger lambda (1se rule), the model selected only one predictor.

Combining Random Forests Model with the Year variable

```
rf_model_year <- randomForest(Value ~ Cement + Gas + Flaring + Coal + Oil + Year, data =
data_clean)
print(rf_model_year)
```

```
##
## Call:
##  randomForest(formula = Value ~ Cement + Gas + Flaring + Coal +      Oil + Year, data
= data_clean)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##           Mean of squared residuals: 0.1391273
##                     % Var explained: 64.06
```

- This model has a good fit of the data with a 64.06% of explained variability.

Different variations of the lagged model

MLR model with 5 year lag

```
##
## Call:
## lm(formula = Value ~ Cement + Gas + Flaring + Coal + Oil + Cement_lag +
##     Gas_lag + Flaring_lag + Coal_lag + Oil_lag, data = data_clean_lag_5)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.58305 -0.40207 -0.04238  0.37237  2.53744
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.5215184  0.0070269  74.217  < 2e-16 ***
## Cement       0.0208763  0.0031807   6.564 5.56e-11 ***
## Gas         -0.0010858  0.0011114  -0.977 0.328608
## Flaring     -0.0020306  0.0022498  -0.903 0.366797
## Coal        -0.0012724  0.0002690  -4.730 2.28e-06 ***
## Oil         -0.0023057  0.0004514  -5.108 3.32e-07 ***
## Cement_lag  -0.0183106  0.0034891  -5.248 1.57e-07 ***
## Gas_lag      0.0090647  0.0011830   7.662 2.02e-14 ***
## Flaring_lag -0.0015944  0.0023426  -0.681 0.496141
## Coal_lag     0.0010189  0.0002777   3.669 0.000245 ***
## Oil_lag      0.0009369  0.0004920   1.904 0.056920 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5952 on 8557 degrees of freedom
## Multiple R-squared:  0.06729,    Adjusted R-squared:  0.0662
## F-statistic: 61.74 on 10 and 8557 DF,  p-value: < 2.2e-16
```

- About a 1.5% increase in explained variance using the statistics from 5 years ago instead of that current year, makes logical sense.

MLR model with lagged value of average of the past 5 years

```
##
## Call:
## lm(formula = Value ~ Cement_avg + Gas_avg + Flaring_avg + Coal_avg +
##     Oil_avg + Year, data = data_clean_5yr_avg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.30817 -0.26694 -0.01956  0.24501  2.08727
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.297e+01  6.075e-01 -87.202  < 2e-16 ***
## Cement_avg  -1.464e-03  7.351e-04  -1.992 0.046412 *
## Gas_avg      2.023e-03  2.609e-04   7.752 1.01e-14 ***
## Flaring_avg  1.241e-04  1.182e-03   0.105 0.916376
## Coal_avg     1.514e-04  6.966e-05   2.174 0.029740 *
## Oil_avg     -4.055e-04  1.060e-04  -3.827 0.000131 ***
## Year         2.684e-02  3.048e-04  88.039  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4386 on 8754 degrees of freedom
## Multiple R-squared:    0.5,  Adjusted R-squared:  0.4997
## F-statistic:  1459 on 6 and 8754 DF,  p-value: < 2.2e-16
```

MLR model with lagged value of average of the past 10 years

```
##
## Call:
## lm(formula = Value ~ Cement_avg + Gas_avg + Flaring_avg + Coal_avg +
##     Oil_avg + Year, data = data_clean_10yr_avg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2755 -0.2721 -0.0186  0.2477  2.0800
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.594e+01  7.076e-01 -79.056  < 2e-16 ***
## Cement_avg  -1.733e-03  8.012e-04  -2.163 0.030579 *
## Gas_avg      2.328e-03  2.906e-04   8.011 1.31e-15 ***
## Flaring_avg -5.521e-04  1.285e-03  -0.430 0.667491
## Coal_avg     1.764e-04  7.405e-05   2.382 0.017238 *
## Oil_avg     -4.195e-04  1.130e-04  -3.712 0.000207 ***
## Year         2.832e-02  3.546e-04  79.849  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4392 on 7790 degrees of freedom
## Multiple R-squared:  0.4835, Adjusted R-squared:  0.4831
## F-statistic:  1215 on 6 and 7790 DF,  p-value: < 2.2e-16
```

MLR model with lagged value of average of the past 25 years

```
##
## Call:
## lm(formula = Value ~ Cement_avg + Gas_avg + Flaring_avg + Coal_avg +
##     Oil_avg + Year, data = data_clean_25yr_avg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.90203 -0.27089 -0.02787  0.23468  2.08301
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.220e+01  1.257e+00 -49.487  < 2e-16 ***
## Cement_avg  -2.402e-03  1.150e-03  -2.089   0.0368 *
## Gas_avg      2.569e-03  4.408e-04   5.827 5.99e-09 ***
## Flaring_avg -1.743e-03  1.945e-03  -0.896   0.3704
## Coal_avg     1.918e-04  9.452e-05   2.029   0.0425 *
## Oil_avg     -1.669e-04  1.543e-04  -1.082   0.2792
## Year         3.143e-02  6.277e-04  50.072  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4406 on 4996 degrees of freedom
## Multiple R-squared:  0.3667, Adjusted R-squared:  0.3659
## F-statistic: 482.1 on 6 and 4996 DF,  p-value: < 2.2e-16
```

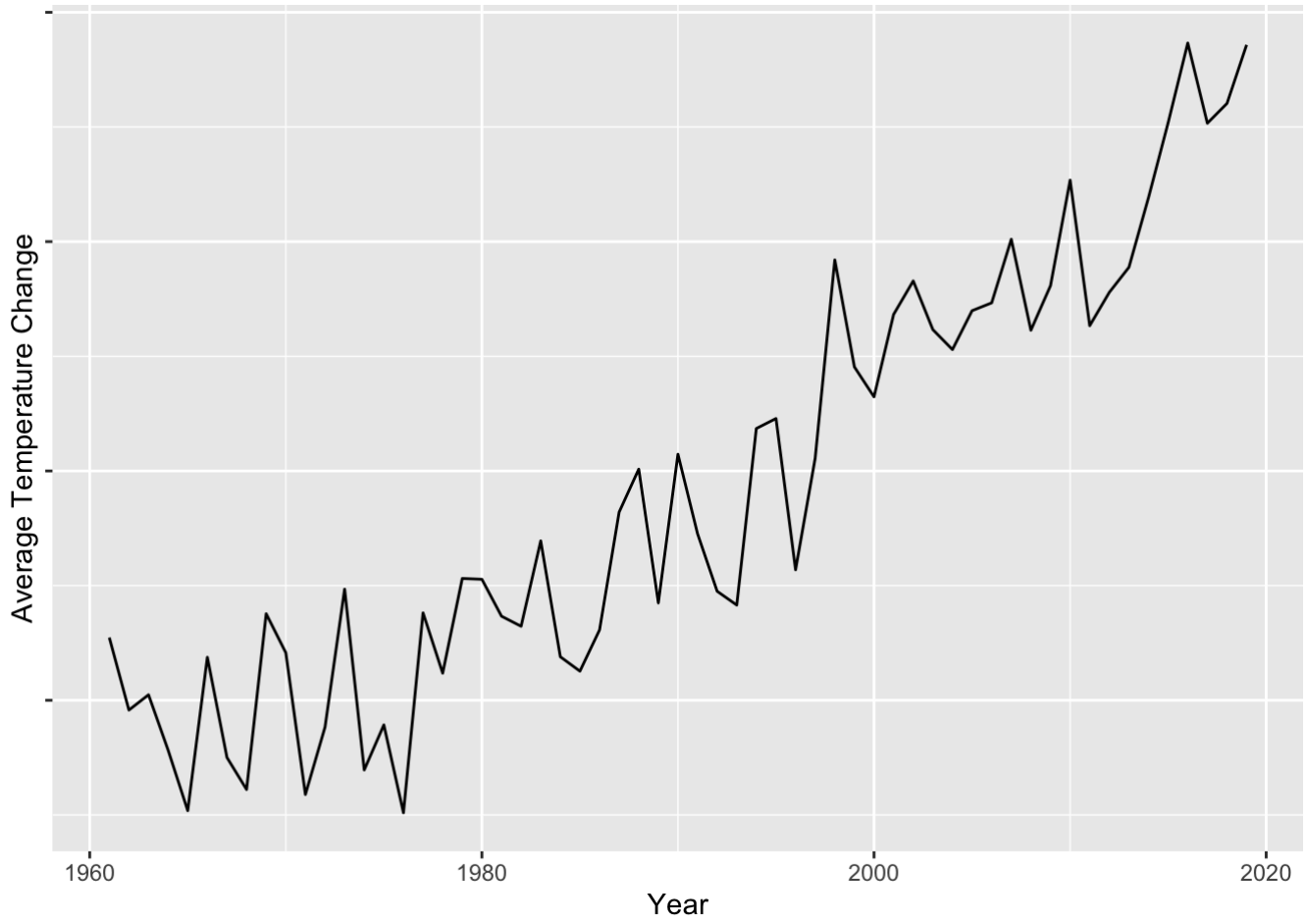
- The 10 year average lagged model preformed better then the 5 year one, but the 25 year average one preformed worse than both.
- All of these are preforming better than models in the past.

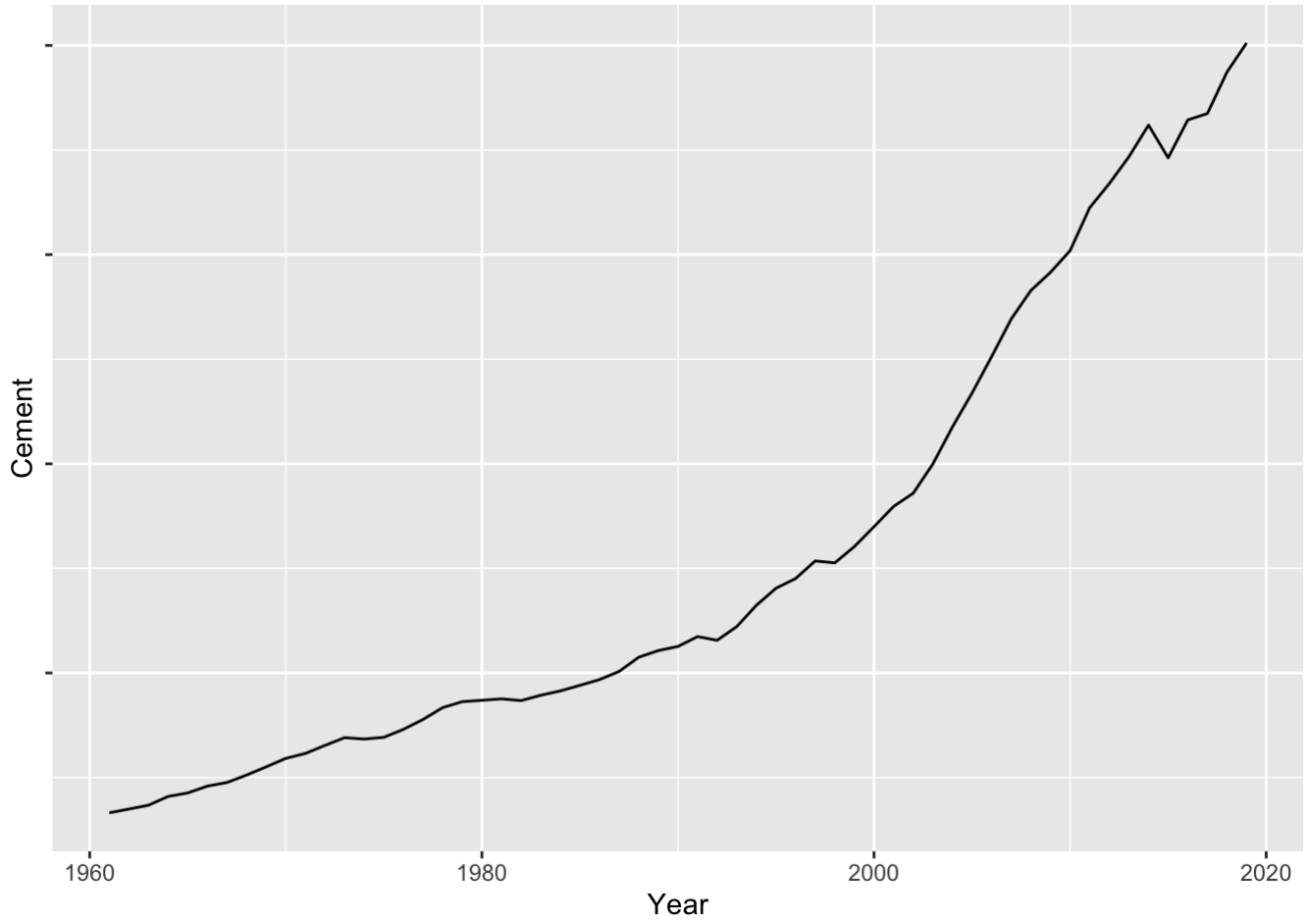
What have I learned from these initial models about which variables are more indicative of climate change?

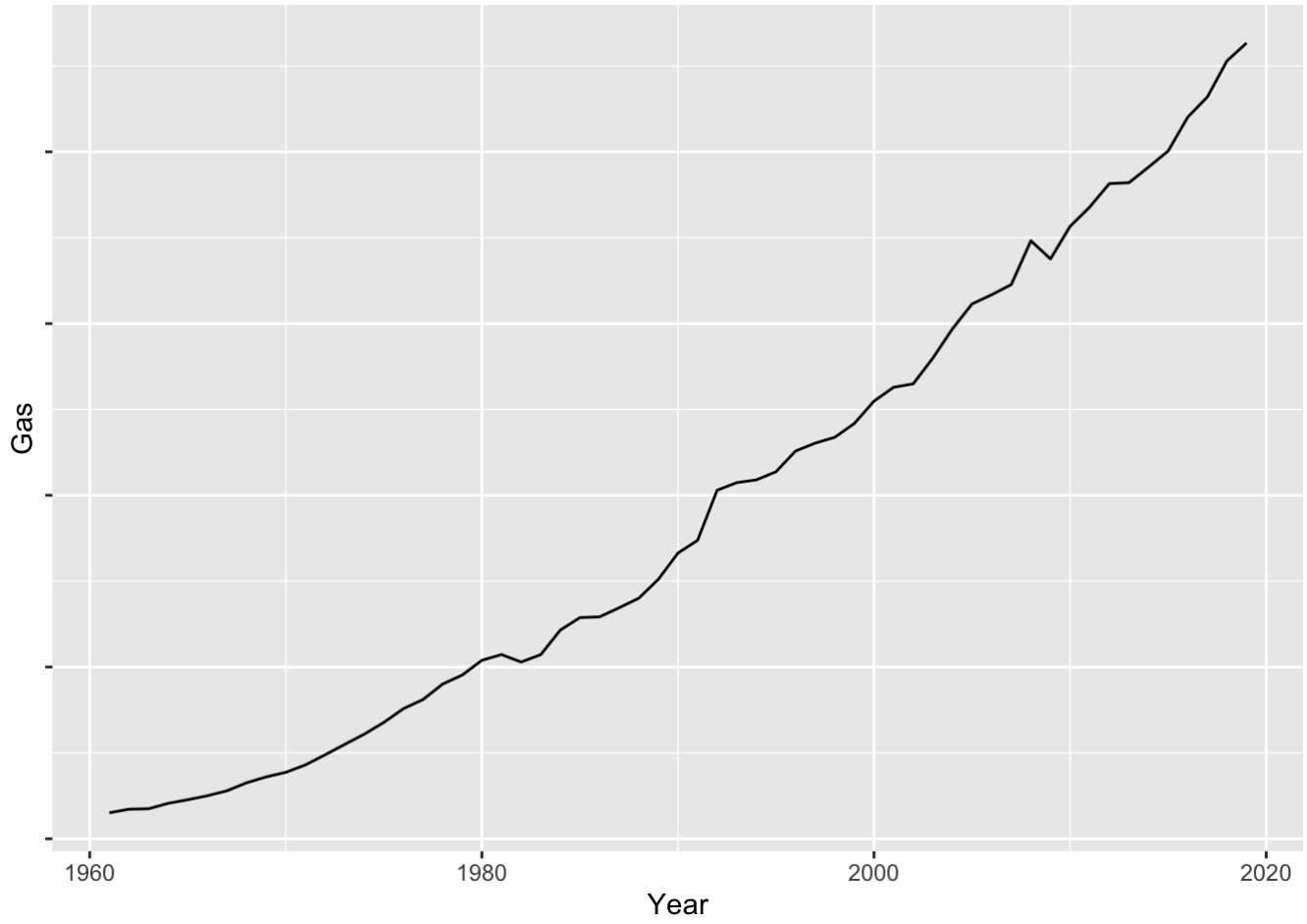
- It seems like the variables Gas, Oil, and Cement have shown consistent significance in most of the models. This suggests that these variables might have a strong relationship with the dependent variable, which I assume is a proxy for climate change.
- The ‘Year’ variable added a significant improvement to the model, suggesting that temporal factors have a strong role in climate change.
- The lagged variables of Gas and Cement were significant in the lagged effects model, suggesting that past values of these variables may influence the dependent variable.

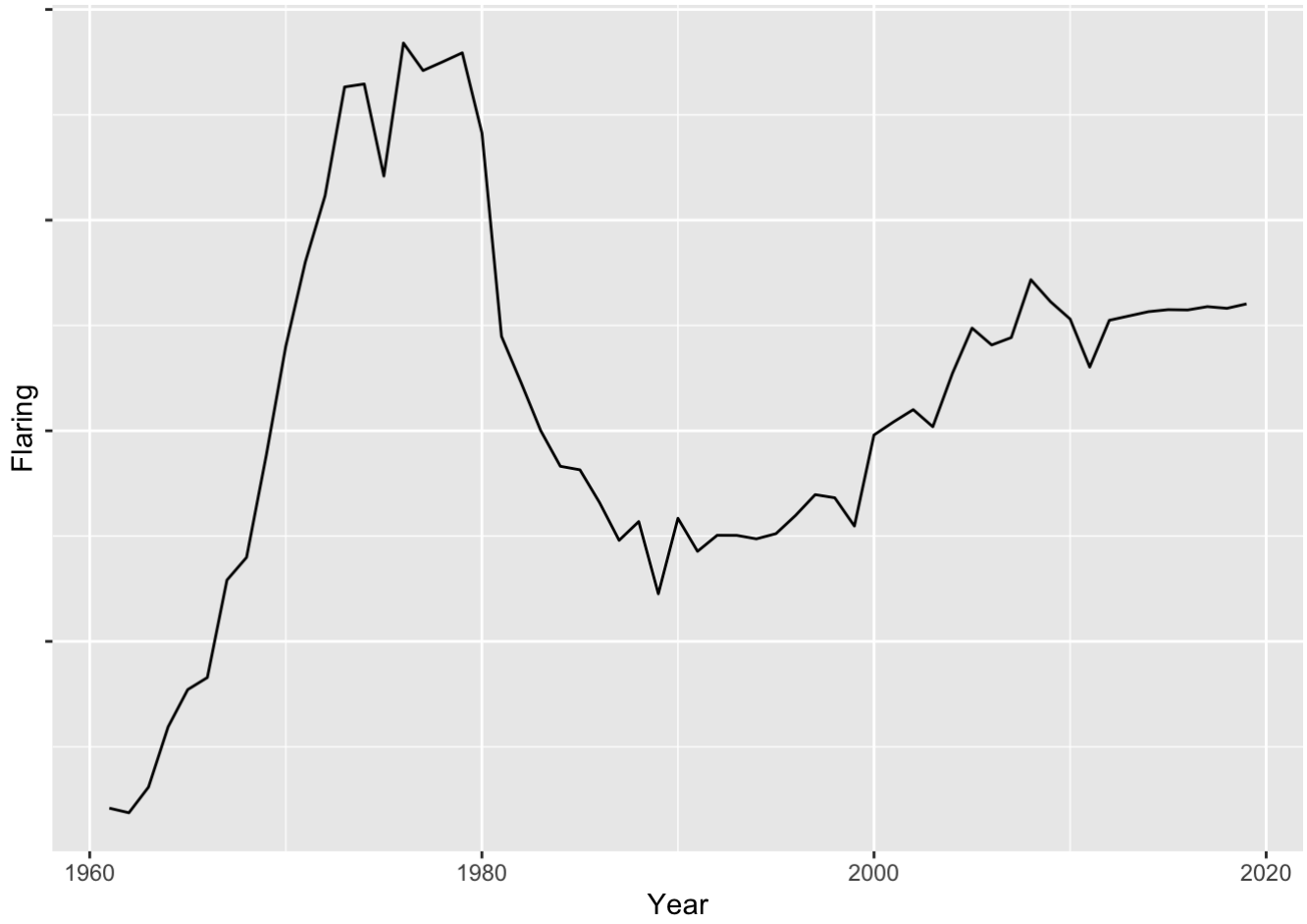
Pivoting Analysis - removing geographical component

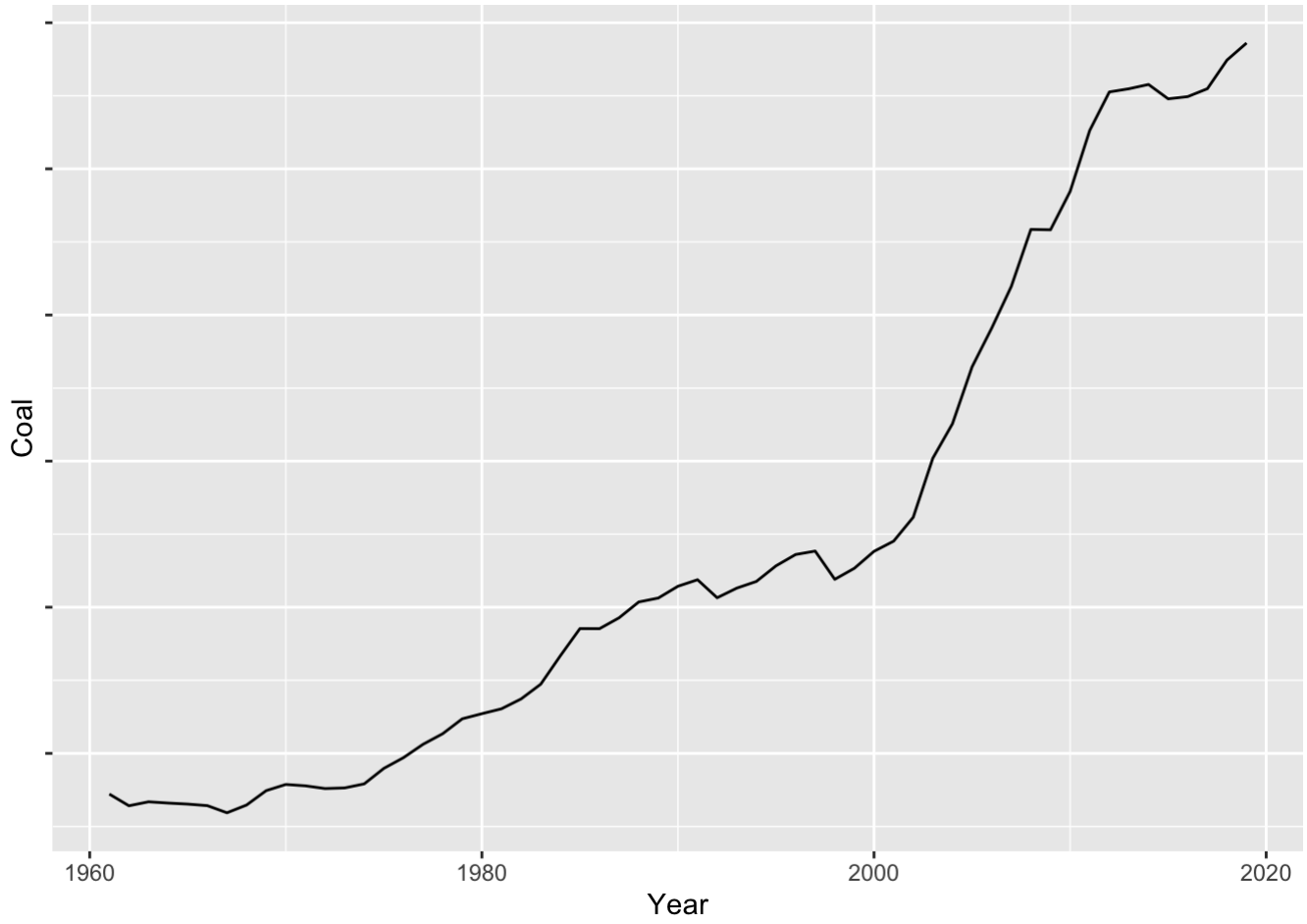
EDA: Visualizing dataset components over time

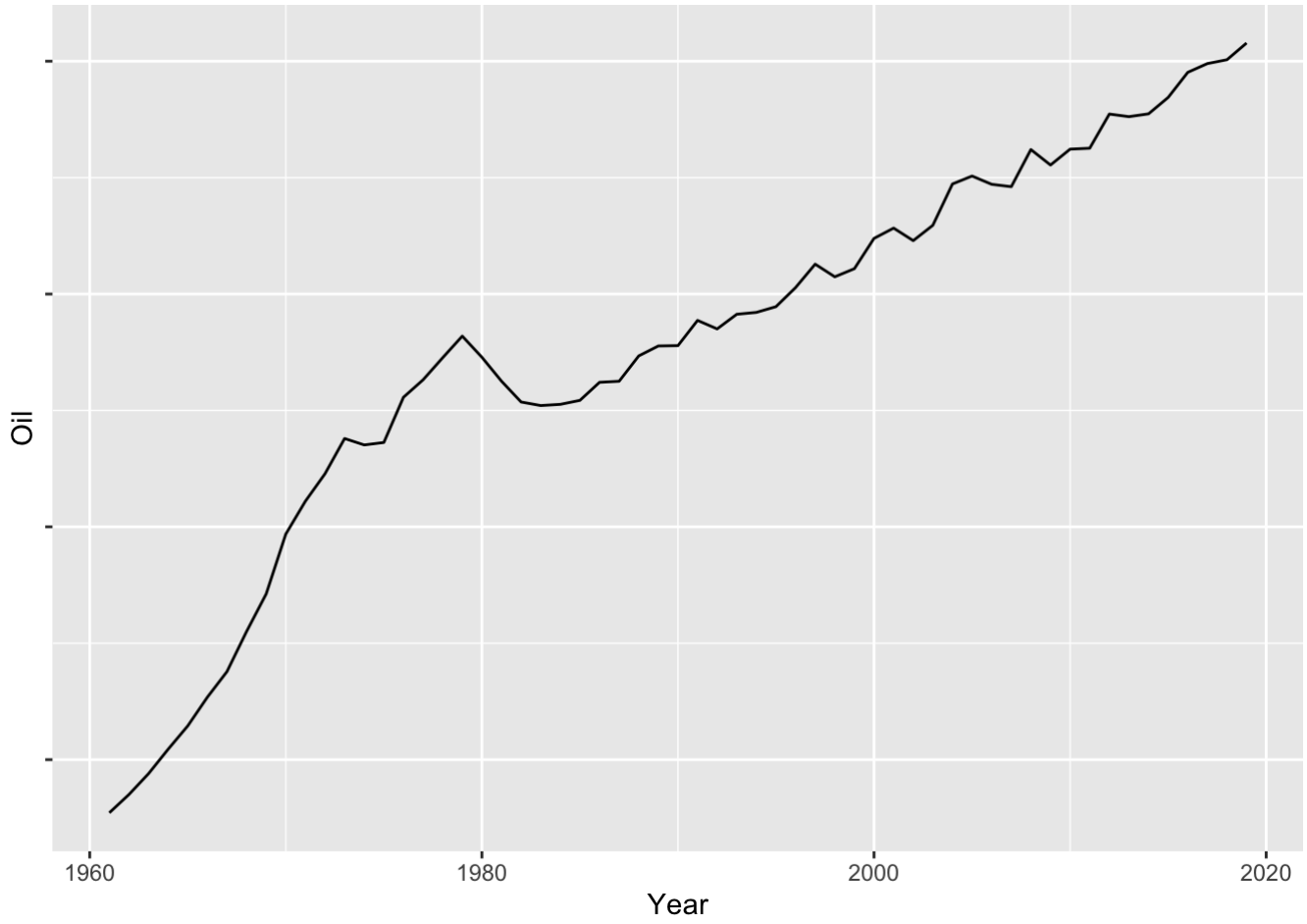

Average Temperature Change
1960
1980
2000
2020
Year

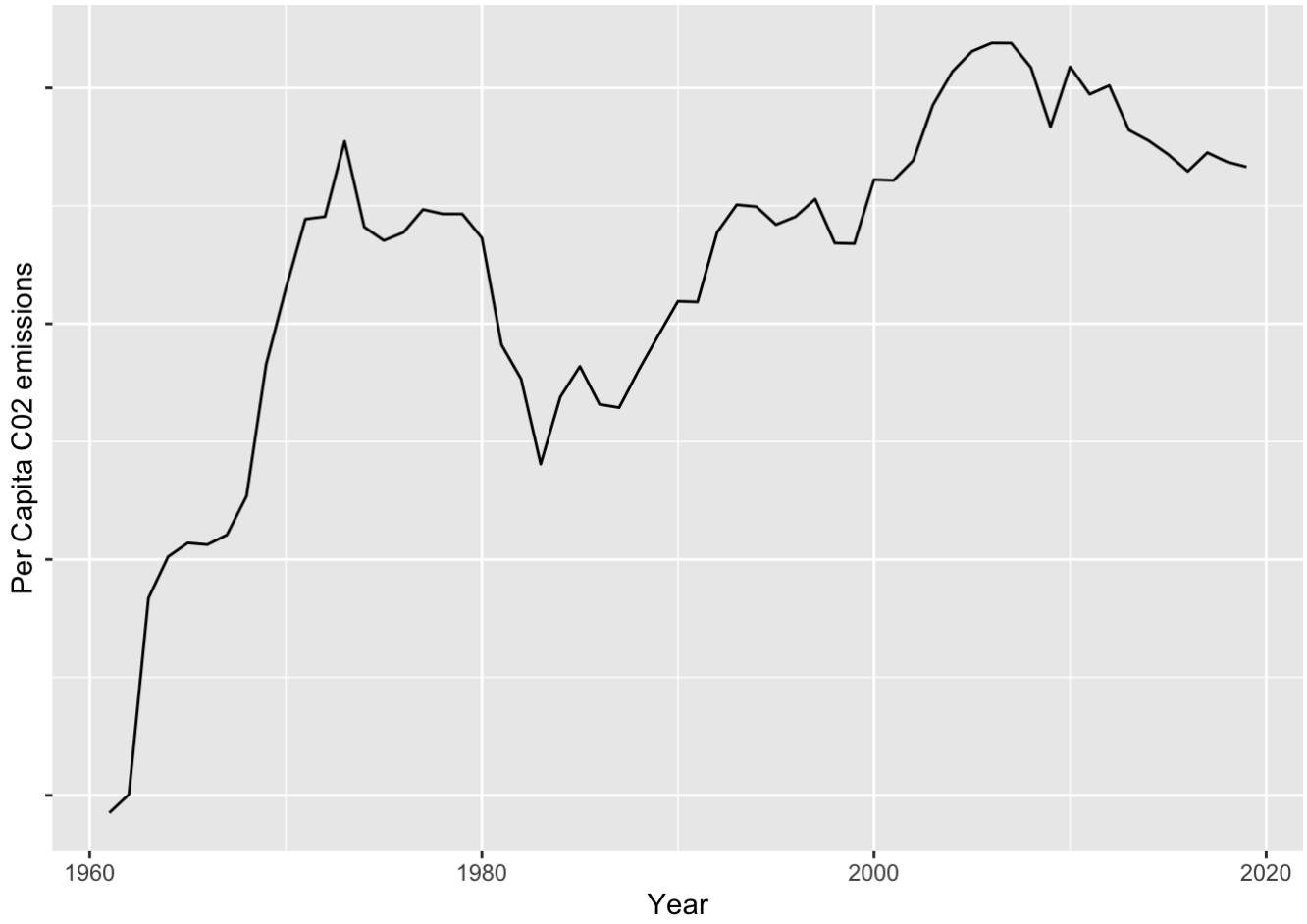

Cement
1960
1980
2000
2020
Year


Gas
1960
1980
2000
2020
Year


Flaring
1960
1980
2000
2020
Year


Coal
1960
1980
2000
2020
Year


Oil
1960
1980
2000
2020
Year



Multiple Linear Regression, using yearly averages of emissions/temperature change

```
##
## Call:
## lm(formula = Value ~ Cement + Gas + Flaring + Coal + Oil, data = yearly_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.27429 -0.09425 -0.01747  0.12982  0.28134
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.052879   0.167033   0.317  0.75281
## Cement       0.387547   0.141196   2.745  0.00825 **
## Gas          0.005854   0.038365   0.153  0.87931
## Flaring     -0.304336   0.136516  -2.229  0.03006 *
## Coal        -0.035830   0.012271  -2.920  0.00513 **
## Oil          0.025168   0.016220   1.552  0.12669
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1499 on 53 degrees of freedom
## Multiple R-squared:  0.9057, Adjusted R-squared:  0.8968
## F-statistic: 101.8 on 5 and 53 DF,  p-value: < 2.2e-16
```

- Wow this model is much stronger than the past ones

- Cement, Coal, and Flaring are significant. Only Cement has a positive coefficent though, potential issues here.

Multiple Linear Regression with interactions, using yearly averages of emissions/temperature change

```
##
## Call:
## lm(formula = Value ~ Cement * Gas * Flaring * Coal * Oil, data = yearly_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.20167 -0.08617 -0.01217  0.06062  0.22365
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -3.063e+01  2.480e+01  -1.235   0.2274
## Cement                       3.640e+01  4.535e+01   0.803   0.4292
## Gas                         -1.077e+01  7.583e+00  -1.420   0.1669
## Flaring                      1.205e+01  1.760e+01   0.685   0.4995
## Coal                         1.581e+00  1.471e+00   1.075   0.2920
## Oil                          1.816e+00  1.020e+00   1.780   0.0864 .
## Cement:Gas                   9.114e-01  3.746e+00   0.243   0.8096
## Cement:Flaring              -1.482e+00  3.313e+01  -0.045   0.9647
## Gas:Flaring                  8.146e+00  5.728e+00   1.422   0.1665
## Cement:Coal                 -1.579e+00  2.747e+00  -0.575   0.5703
## Gas:Coal                     4.531e-01  4.316e-01   1.050   0.3032
## Flaring:Coal                -1.598e+00  1.296e+00  -1.233   0.2281
## Cement:Oil                  -1.833e+00  1.216e+00  -1.507   0.1433
## Gas:Oil                      3.579e-01  2.173e-01   1.647   0.1111
## Flaring:Oil                 -1.361e+00  7.816e-01  -1.742   0.0929 .
## Coal:Oil                    -9.066e-02  5.227e-02  -1.734   0.0943 .
## Cement:Gas:Flaring          -1.022e+00  2.932e+00  -0.349   0.7301
## Cement:Gas:Coal             -3.974e-02  9.599e-02  -0.414   0.6821
## Cement:Flaring:Coal          8.802e-01  2.102e+00   0.419   0.6786
## Gas:Flaring:Coal            -4.060e-01  3.318e-01  -1.224   0.2317
## Cement:Gas:Oil               3.882e-03  8.903e-02   0.044   0.9655
## Cement:Flaring:Oil           5.328e-01  8.519e-01   0.625   0.5369
## Gas:Flaring:Oil             -2.537e-01  1.712e-01  -1.482   0.1500
## Cement:Coal:Oil              7.703e-02  7.328e-02   1.051   0.3025
## Gas:Coal:Oil                -1.411e-02  1.292e-02  -1.093   0.2842
## Flaring:Coal:Oil             1.005e-01  5.601e-02   1.795   0.0839 .
## Cement:Gas:Flaring:Coal      4.431e-02  7.209e-02   0.615   0.5439
## Cement:Gas:Flaring:Oil       1.520e-02  6.810e-02   0.223   0.8250
## Cement:Gas:Coal:Oil         -9.513e-05  2.162e-03  -0.044   0.9652
## Cement:Flaring:Coal:Oil     -5.017e-02  5.564e-02  -0.902   0.3752
## Gas:Flaring:Coal:Oil         1.122e-02  9.953e-03   1.127   0.2697
## Cement:Gas:Flaring:Coal:Oil -3.060e-04  1.603e-03  -0.191   0.8500
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1483 on 27 degrees of freedom
## Multiple R-squared:  0.953,  Adjusted R-squared:  0.899
## F-statistic: 17.65 on 31 and 27 DF,  p-value: 2.463e-11
```

- This model explains slightly more variance than the last
- None of the interactions are significant here.

Random Forest Model, using yearly averages of emissions/temperature change

```
##
## Call:
##  randomForest(formula = Value ~ Cement + Gas + Flaring + Coal +      Oil, data = year
ly_data)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 1
##
##           Mean of squared residuals: 0.03040743
##                     % Var explained: 85.79
```

- This is a strong model as well with a % explained variance at 85.79

Lasso Regression Model, using yearly averages of emissions/temperature change

```
##
## Call:  glmnet(x = predictors, y = response, alpha = 1, lambda = optimal_lambda)
##
##   Df  %Dev    Lambda
## 1  5 90.56 9.176e-05
```

```
## 6 x 1 sparse Matrix of class "dgCMatrix"
##                       s0
## (Intercept)  0.041088982
## Cement       0.366190374
## Gas          0.008571105
## Flaring     -0.290296032
## Coal        -0.033666574
## Oil          0.023846055
```

```
## [1] "Mean Squared Error:  0.02020657061986"
```

```
## [1] "R-squared:  0.905601641334736"
```

```
## [1] "Adjusted R-squared:  0.896696135800277"
```

MLR model with lagged value of average of the past 5 years

```
##
## Call:
## lm(formula = Value ~ Cement_avg + Gas_avg + Flaring_avg + Coal_avg +
##     Oil_avg + Year, data = data_clean_5yr_avg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28465 -0.09906 -0.02239  0.12187  0.31755
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.898328  84.509581  -0.105   0.9166
## Cement_avg   0.503454   0.238463   2.111   0.0400 *
## Gas_avg     -0.035771   0.084682  -0.422   0.6746
## Flaring_avg -0.402326   0.312283  -1.288   0.2038
## Coal_avg    -0.043467   0.018828  -2.309   0.0253 *
## Oil_avg      0.042827   0.053845   0.795   0.4303
## Year         0.004471   0.043227   0.103   0.9180
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.152 on 48 degrees of freedom
## Multiple R-squared:  0.9051, Adjusted R-squared:  0.8933
## F-statistic: 76.33 on 6 and 48 DF,  p-value: < 2.2e-16
```

MLR model with lagged value of average of the past 10 years

```
##
## Call:
## lm(formula = Value ~ Cement_avg + Gas_avg + Flaring_avg + Coal_avg +
##     Oil_avg + Year, data = data_clean_10yr_avg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.29171 -0.11291 -0.02582  0.12732  0.30496
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 197.37208  151.32113   1.304    0.199
## Cement_avg    0.63885    0.52670   1.213    0.232
## Gas_avg       0.07684    0.16337   0.470    0.640
## Flaring_avg  -1.03072    0.62650  -1.645    0.107
## Coal_avg     -0.05320    0.03910  -1.361    0.181
## Oil_avg       0.17358    0.10420   1.666    0.103
## Year         -0.10098    0.07730  -1.306    0.198
##
## Residual standard error: 0.1516 on 43 degrees of freedom
## Multiple R-squared:    0.9,  Adjusted R-squared:  0.886
## F-statistic: 64.49 on 6 and 43 DF,  p-value: < 2.2e-16
```

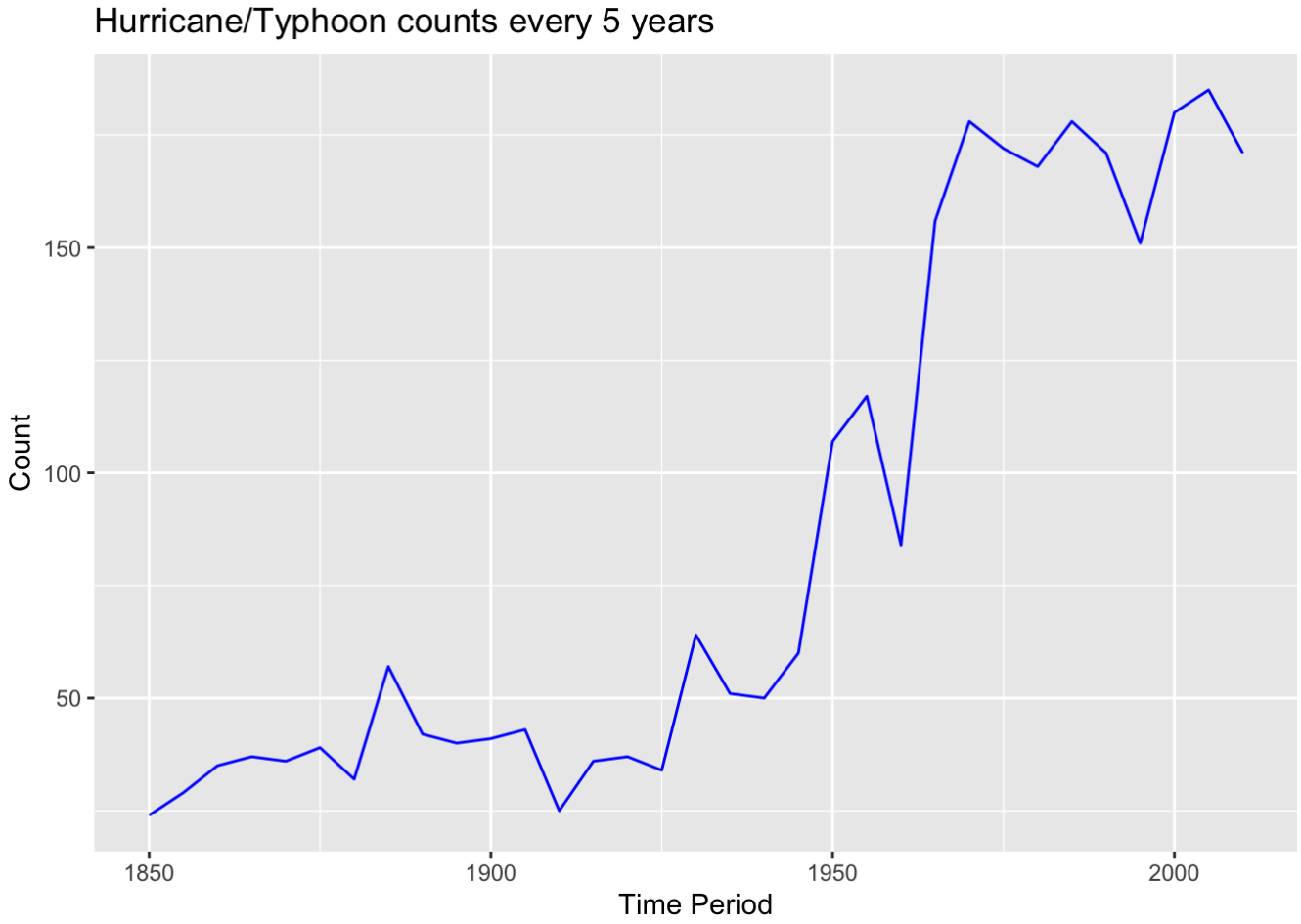
MLR model with lagged value of average of the past 25 years

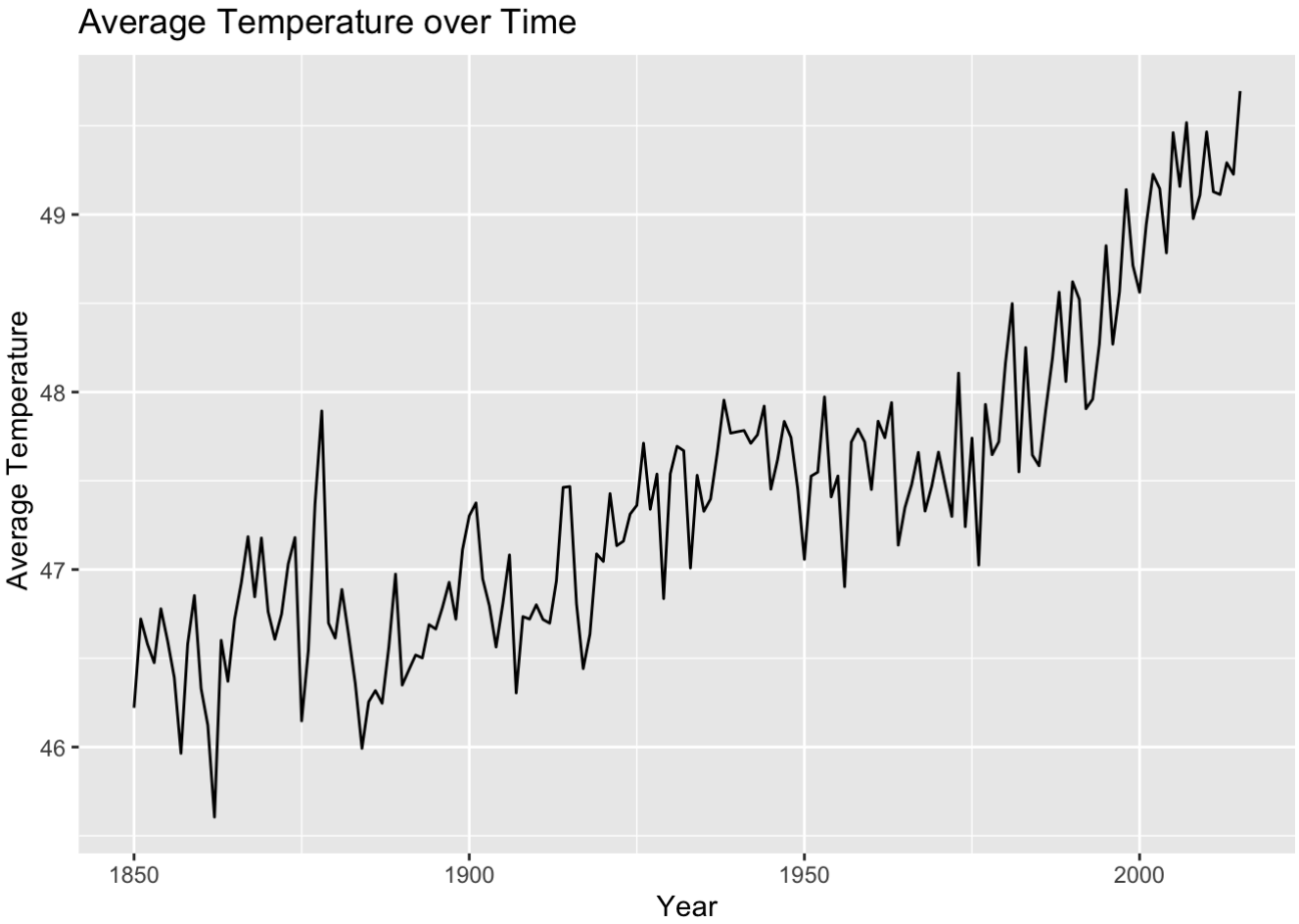
```
##
## Call:
## lm(formula = Value ~ Cement_avg + Gas_avg + Flaring_avg + Coal_avg +
##     Oil_avg + Year, data = data_clean_25yr_avg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.29097 -0.07152 -0.00738  0.08695  0.39499
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -801.6958   938.6640  -0.854   0.4003
## Cement_avg     3.7428     1.6629   2.251   0.0324 *
## Gas_avg        1.0096     0.9743   1.036   0.3090
## Flaring_avg    6.3014     5.3083   1.187   0.2452
## Coal_avg      -1.0688     0.5014  -2.132   0.0419 *
## Oil_avg       -0.5450     0.6558  -0.831   0.4130
## Year           0.4125     0.4789   0.861   0.3963
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.147 on 28 degrees of freedom
## Multiple R-squared:  0.8696, Adjusted R-squared:  0.8417
## F-statistic: 31.13 on 6 and 28 DF,  p-value: 3.793e-11
```

- Lagged average models are preforming well but not as well as I expected compared to the other ones in this section.

Additional Analysis

Temperature Change Effect on Hurricanes/Typhoons



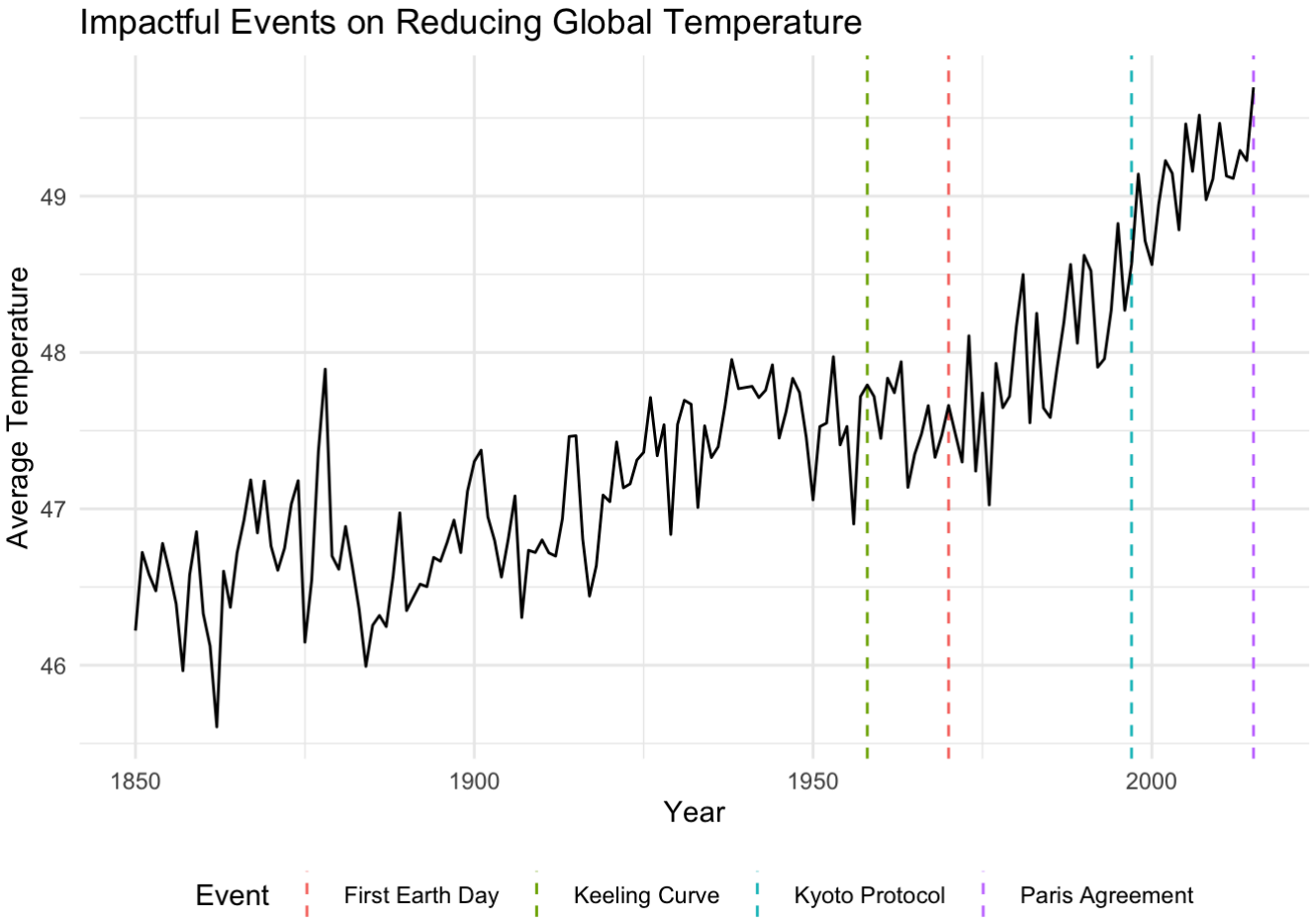


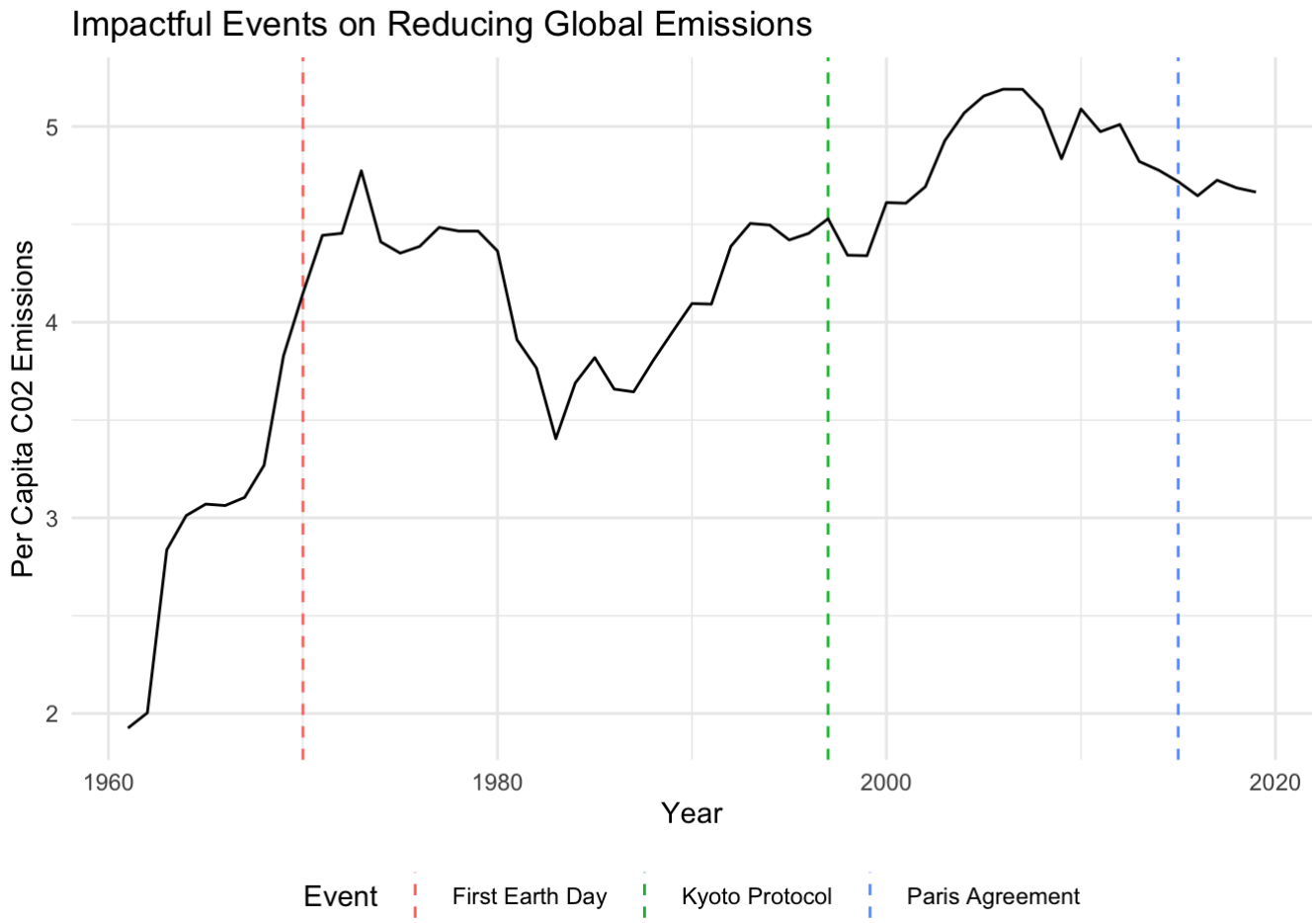
Simple Linear Regression Model - Storm Count ~ Average Global Land Temperature (Yearly)

```
##
## Call:
## lm(formula = Count ~ LandAverageTemperature, data = storm_counts_temp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.700  -5.530  -1.369   3.602  23.281
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -502.8449    38.7412  -12.98   <2e-16 ***
## LandAverageTemperature   10.9674     0.8168   13.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.696 on 162 degrees of freedom
##   (102 observations deleted due to missingness)
## Multiple R-squared:  0.5267, Adjusted R-squared:  0.5238
## F-statistic: 180.3 on 1 and 162 DF,  p-value: < 2.2e-16
```

- The average temperature and the storm counts follow a similar trend, could potentially be a causal relationship.

Analyzing impactful Climate Change laws/movements





- I wouldn't say these events have had much impact yet on global temperature or emissions.

Raw Code

Benjamin Coleman

2023

```
global_temps = read.csv("data/GlobalTemperatures.csv")

global_temps$LandAverageTemperature <- (global_temps$LandAverageTemperature * 9/5) + 32
global_temps$LandMaxTemperature <- (global_temps$LandMaxTemperature * 9/5) + 32
global_temps$LandMinTemperature <- (global_temps$LandMinTemperature * 9/5) + 32
global_temps$LandAndOceanAverageTemperature <- (global_temps$LandAndOceanAverageTemperat
ure * 9/5) + 32



global_temps$dt = as.Date(global_temps$dt, format = "%Y-%m-%d")

global_temps$Year = year(global_temps$dt)
global_temps$Month = month(global_temps$dt)

annual_data <- global_temps %>%
  group_by(Year) %>%
  summarise(
    LandAverageTemperature = mean(LandAverageTemperature, na.rm = TRUE),
    LandMaxTemperature = mean(LandMaxTemperature, na.rm = TRUE),
    LandMinTemperature = mean(LandMinTemperature, na.rm = TRUE),
    LandAndOceanAverageTemperature = mean(LandAndOceanAverageTemperature, na.rm = TRUE)
  )
annual_data_complete = na.omit(annual_data)


annual_data_1970 <- annual_data_complete %>% filter(Year >= 1970)

temps_ts_1970 <- ts(annual_data_1970$LandAndOceanAverageTemperature, start = 1970, frequ
ency = 1)

train_ts_1970 <- window(temps_ts_1970, end = c(2014))

temps_sarima_1970 <- auto.arima(train_ts_1970, seasonal = TRUE)

temps_forecast_sarima_1970 <- forecast(temps_sarima_1970, h = 50)

x_hist <- 1970:(2014)
x_forecast <- 2015:2064

combined_data_sarima <- c(train_ts_1970, temps_forecast_sarima_1970$mean)

wiggles <- runif(length(x_forecast), -0.2, 0.2)

temps_forecast_sarima_1970_wiggles <- temps_forecast_sarima_1970$mean + wiggles


plot(x = x_hist, y = train_ts_1970, main = "Global Land & Ocean Average Temperature Forc
ast",
     xlab = "Year", ylab = "Temperature (°F)", type = "l", col = "blue", ylim = range(co
```

```
mbined_data_sarima), xlim = c(1970, 2040))
lines(x = x_forecast, y = temps_forecast_sarima_1970_wiggles, col = "red")
```

This graph was mostly for the presentation. The goal was to use it to illustrate the fact that following historical temperature trends, the temperature will continue to rise. The red line is the forecast.

Statistical methods used: Seasonal ARIMA model trained on annual average temperature data from 1970-2014.

```
temps = read.csv("data/FAOSTAT_data_en_5-11-2023.csv")
temps = temps[, -7]

emissions = read.csv("data/GCB2022v27_MtCO2_flat.csv")
emissions <- emissions %>%
  rename(
    Area = Country,
    `Area Code` = ISO.3166.1.alpha.3
  ) %>% filter(Year >= 1961 & Year <= 2019)

data = merge(x = temps, y = emissions, by = c("Area", "Year"), all = "True")
data =  select(data, -c("Flag.Description", "Flag", "Element.Code", "Year.Code", "Unit",
"Element", "Months", "Domain.Code", "Domain"))
```

Visualizing missing data

```
gg_miss_var(data, show_pct = TRUE)

data <- data %>%
  mutate(is_missing = is.na(Value))

missing_counts <- data %>%
  group_by(Year) %>%
  summarise(missing_count = sum(is_missing), total_count = n())

missing_counts <- missing_counts %>%
  mutate(missing_percentage = missing_count / total_count * 100)

ggplot(missing_counts, aes(x = Year, y = missing_percentage)) +
  geom_line() +
  labs(title = "Percentage of missing Value over time",
       x = "Year",
       y = "Percentage of missing values")
```

- These helped me view the trend of missing data and evaluate the best way to handle it.

Statistical Modeling - learn more about emissions impact on temperature change

Multiple Linear Regression Model, Predictors: Cement, Gas, Flaring, Coal, Oil

```
data <- data[complete.cases(data$Value), ]
data_clean <- na.omit(data[, c("Area", "Year", "Value", "Cement", "Gas", "Flaring", "Coa
l", "Oil", "Per.Capita")])

mlr_model <- lm(Value ~ Cement + Gas + Flaring + Coal + Oil, data = data_clean)
summary(mlr_model)
```

- These predictors (Cement, Gas, Flaring, Coal, Oil) were able to explain about 5.3% of the variance in annual temperature.
- Cement, Gas, and Oil were statistically significant predictors ($p<0.05$).

Multiple Linear Regression Model with Interactions, Predictors: Cement, Gas, Flaring, Coal, Oil

```
mlr_model_interactions <- lm(Value ~ Cement * Gas * Flaring * Coal * Oil, data = data_cl
ean)
summary(mlr_model_interactions)
```

- Introducing these interaction terms increased the explained variance to about 7.5%.
- Many of the interaction terms were found to be statistically significant.

Multiple Linear Regression, introducing time component with the 'Year' variable

```
mlr_model_time <- lm(Value ~ Cement + Gas + Flaring + Coal + Oil + Year, data = data_cle
an)
summary(mlr_model_time)
```

- The addition of the time variable significantly increased the explained variance to about 49.2%.
- Gas, Oil, and Year were statistically significant predictors.

Multiple Linear Regression Model with Lagged Effects

```
data_clean_lag <- data_clean %>%
  dplyr::group_by(Area) %>%
  dplyr::mutate_at(vars(Cement, Gas, Flaring, Coal, Oil), list(lag = dplyr::lag)) %>%
  dplyr::ungroup()

data_clean_lag <- na.omit(data_clean_lag)

mlr_model_lag <- lm(Value ~ Cement + Gas + Flaring + Coal + Oil + Cement_lag + Gas_lag +
Flaring_lag + Coal_lag + Oil_lag, data = data_clean_lag)
summary(mlr_model_lag)
```

- The multiple linear regression model with lagged emissions predictors explained about 5.9% of the variance.
- Cement, Oil, Cement_lag, and Gas_lag were statistically significant predictors.

Random Forest Model, Predictors: Cement, Gas, Flaring, Coal, Oil

```
rf_model <- randomForest(Value ~ Cement + Gas + Flaring + Coal + Oil, data = data_clean)
print(rf_model)
```

- The Random Forest model explained 34.89% of the variance in Value, which is a good improvement over the linear models without the time component, but less than the linear model with the time component.
- The Mean Squared Error is about 0.252.

Lasso Regression Model, Predictors: Cement, Gas, Flaring, Coal, Oil

```
x <- model.matrix(Value ~ Cement + Gas + Flaring + Coal + Oil, data = data_clean)[,-1]
y <- data_clean$Value

lasso_model <- glmnet(x, y, alpha = 1)
set.seed(123)  # For reproducibility
cv.lasso_model <- cv.glmnet(x, y, alpha = 1)
print(cv.lasso_model)
coef(cv.lasso_model, s = cv.lasso_model$lambda.min)
```

- The Lasso model selected 5 non-zero coefficients, with a Mean Squared Error of about 0.3669 at the best lambda (regularization parameter).
- At a slightly larger lambda (1se rule), the model selected only one predictor.

Combining Random Forests Model with the Year variable

```
rf_model_year <- randomForest(Value ~ Cement + Gas + Flaring + Coal + Oil + Year, data =
data_clean)
print(rf_model_year)
```

- This model has a good fit of the data with a 64.06% of explained variability.

Different variations of the lagged model

MLR model with 5 year lag

```
data_clean_lag_5 <- data_clean %>%
  dplyr::group_by(Area) %>%
  dplyr::mutate_at(vars(Cement, Gas, Flaring, Coal, Oil, Per.Capita), list(lag = ~dply
r::lag(., 5))) %>%
  dplyr::ungroup()

data_clean_lag_5 <- na.omit(data_clean_lag_5)

mlr_model_lag_5 <- lm(Value ~ Cement + Gas + Flaring + Coal + Oil + Cement_lag + Gas_lag
+ Flaring_lag + Coal_lag + Oil_lag, data = data_clean_lag_5)
summary(mlr_model_lag_5)
```

- About a 1.5% increase in explained variance using the statistics from 5 years ago instead of that current year, makes logical sense.

MLR model with lagged value of average of the past 5 years

```
data_clean_5yr_avg <- data_clean %>%
  dplyr::group_by(Area) %>%
  dplyr::mutate_at(vars(Cement, Gas, Flaring, Coal, Oil), list(avg = ~zoo::rollmean(.,
5, fill = NA, align = "right"))) %>%
  dplyr::ungroup()

data_clean_5yr_avg <- na.omit(data_clean_5yr_avg)

mlr_model_5yr_avg <- lm(Value ~ Cement_avg + Gas_avg + Flaring_avg + Coal_avg + Oil_avg
+ Year, data = data_clean_5yr_avg)
summary(mlr_model_5yr_avg)
```

MLR model with lagged value of average of the past 10 years

```
data_clean_10yr_avg <- data_clean %>%
  dplyr::group_by(Area) %>%
  dplyr::mutate_at(vars(Cement, Gas, Flaring, Coal, Oil), list(avg = ~zoo::rollmean(., 1
0, fill = NA, align = "right"))) %>%
  dplyr::ungroup()

data_clean_10yr_avg <- na.omit(data_clean_10yr_avg)

mlr_model_10yr_avg <- lm(Value ~ Cement_avg + Gas_avg + Flaring_avg + Coal_avg + Oil_avg
+ Year, data = data_clean_10yr_avg)
summary(mlr_model_10yr_avg)
```

MLR model with lagged value of average of the past 25 years

```
data_clean_25yr_avg <- data_clean %>%
  dplyr::group_by(Area) %>%
  dplyr::mutate_at(vars(Cement, Gas, Flaring, Coal, Oil), list(avg = ~zoo::rollmean(., 2
5, fill = NA, align = "right"))) %>%
  dplyr::ungroup()

data_clean_25yr_avg <- na.omit(data_clean_25yr_avg)

mlr_model_25yr_avg <- lm(Value ~ Cement_avg + Gas_avg + Flaring_avg + Coal_avg + Oil_avg
+ Year, data = data_clean_25yr_avg)
summary(mlr_model_25yr_avg)
```

- The 10 year average lagged model preformed better then the 5 year one, but the 25 year average one preformed worse than both.
- All of these are preforming better than models in the past.

What have I learned from these initial models about which variables are more indicative of climate change?

- It seems like the variables Gas, Oil, and Cement have shown consistent significance in most of the models. This suggests that these variables might have a strong relationship with the dependent variable, which I assume is a proxy for climate change.

- The ‘Year’ variable added a significant improvement to the model, suggesting that temporal factors have a strong role in climate change.
- The lagged variables of Gas and Cement were significant in the lagged effects model, suggesting that past values of these variables may influence the dependent variable.

Pivoting Analysis - removing geographical component

EDA: Visualizing dataset components over time

```
yearly_data = data_clean %>%
  group_by(Year) %>%
  summarize(
    Value = mean(Value, na.rm = TRUE),
    Cement = mean(Cement, na.rm = TRUE),
    Gas = mean(Gas, na.rm = TRUE),
    Flaring = mean(Flaring, na.rm = TRUE),
    Coal = mean(Coal, na.rm = TRUE),
    Oil = mean(Oil, na.rm = TRUE),
    Per.Capita = mean(Per.Capita, na.rm = TRUE)
  )

ggplot(data = yearly_data, aes(x = Year, y = Value)) + geom_line() + ylab("Average Tempe
rature Change") + theme(axis.text.y = element_blank())
ggplot(data = yearly_data, aes(x = Year, y = Cement)) + geom_line() + theme(axis.text.y
= element_blank())
ggplot(data = yearly_data, aes(x = Year, y = Gas)) + geom_line() + theme(axis.text.y = e
lement_blank())
ggplot(data = yearly_data, aes(x = Year, y = Flaring)) + geom_line() + theme(axis.text.y
= element_blank())
ggplot(data = yearly_data, aes(x = Year, y = Coal)) + geom_line()+ theme(axis.text.y = e
lement_blank())
ggplot(data = yearly_data, aes(x = Year, y = Oil)) + geom_line()+ theme(axis.text.y = el
ement_blank())
ggplot(data = yearly_data, aes(x = Year, y = Per.Capita)) + geom_line() + ylab("Per Capi
ta C02 emissions")+ theme(axis.text.y = element_blank())
```

Multiple Linear Regression, using yearly averages of emissions/temperature change

```
mlr_model_yearly <- lm(Value ~ Cement + Gas + Flaring + Coal + Oil, data = yearly_data)
summary(mlr_model_yearly)
```

- Wow this model is much stronger than the past ones
- Cement, Coal, and Flaring are significant. Only Cement has a positive coefficent though, potential issues here.

Multiple Linear Regression with interactions, using yearly averages of emissions/temperature change

```
mlr_model_yearly_interactions <- lm(Value ~ Cement * Gas * Flaring * Coal * Oil, data =
yearly_data)
summary(mlr_model_yearly_interactions)
```

- This model explains slightly more variance than the last
- None of the interactions are significant here.

Random Forest Model, using yearly averages of emissions/temperature change

```
rf_model_yearly = randomForest(Value ~ Cement + Gas + Flaring + Coal + Oil, data = yearl
y_data)
print(rf_model_yearly)
```

- This is a strong model as well with a % explained variance at 85.79

Lasso Regression Model, using yearly averages of emissions/temperature change

```
predictors <- as.matrix(yearly_data[,c("Cement", "Gas", "Flaring", "Coal", "Oil")])
response <- yearly_data$Value

lasso_model <- glmnet(predictors, response, alpha = 1)

cv.lasso <- cv.glmnet(predictors, response, alpha = 1)
optimal_lambda <- cv.lasso$lambda.min

lasso_model <- glmnet(predictors, response, alpha = 1, lambda = optimal_lambda)

print(lasso_model)
print(coef(lasso_model))

lasso_predictions <- predict(lasso_model, newx = predictors)

MSE <- mean((response - lasso_predictions)^2)
R2 <- 1 - sum((response - lasso_predictions)^2) / sum((response - mean(response))^2)
n <- length(response)
p <- ncol(predictors)
adj_R2 <- 1 - (1-R2)*(n-1)/(n-p-1)

print(paste("Mean Squared Error: ", MSE))
print(paste("R-squared: ", R2))
print(paste("Adjusted R-squared: ", adj_R2))
```

MLR model with lagged value of average of the past 5 years

```
data_clean_5yr_avg <- yearly_data %>%
  dplyr::mutate_at(vars(Cement, Gas, Flaring, Coal, Oil), list(avg = ~zoo::rollmean(.,
5, fill = NA, align = "right"))) %>%
  dplyr::ungroup()

data_clean_5yr_avg <- na.omit(data_clean_5yr_avg)

mlr_model_5yr_avg <- lm(Value ~ Cement_avg + Gas_avg + Flaring_avg + Coal_avg + Oil_avg
+ Year, data = data_clean_5yr_avg)
summary(mlr_model_5yr_avg)
```

MLR model with lagged value of average of the past 10 years

```
data_clean_10yr_avg <- yearly_data %>%
  dplyr::mutate_at(vars(Cement, Gas, Flaring, Coal, Oil), list(avg = ~zoo::rollmean(., 1
0, fill = NA, align = "right"))) %>%
  dplyr::ungroup()

data_clean_10yr_avg <- na.omit(data_clean_10yr_avg)

mlr_model_10yr_avg <- lm(Value ~ Cement_avg + Gas_avg + Flaring_avg + Coal_avg + Oil_avg
+ Year, data = data_clean_10yr_avg)
summary(mlr_model_10yr_avg)
```

MLR model with lagged value of average of the past 25 years

```
data_clean_25yr_avg <- yearly_data %>%
  dplyr::mutate_at(vars(Cement, Gas, Flaring, Coal, Oil), list(avg = ~zoo::rollmean(., 2
5, fill = NA, align = "right"))) %>%
  dplyr::ungroup()

data_clean_25yr_avg <- na.omit(data_clean_25yr_avg)

mlr_model_25yr_avg <- lm(Value ~ Cement_avg + Gas_avg + Flaring_avg + Coal_avg + Oil_avg
+ Year, data = data_clean_25yr_avg)
summary(mlr_model_25yr_avg)
```

- Lagged average models are preforming well but not as well as I expected compared to the other ones in this section.

Additional Analysis

Temperature Change Effect on Hurricanes/Typhoons

```
atlantic = read.csv("data/atlantic.csv")
pacific = read.csv("data/pacific.csv")
atlantic = atlantic %>% select("ID", "Date")
pacific = pacific %>% select("ID", "Date")
atlantic = atlantic[!duplicated(atlantic$ID), ]
pacific = pacific[!duplicated(pacific$ID), ]
storm = rbind(atlantic, pacific)
storm$Date <- as.Date(as.character(storm$Date), format = "%Y%m%d")

storm_counts_5 <- storm %>%
  mutate(YearGroup = floor(as.numeric(format(Date, "%Y")) / 5) * 5) %>%
  group_by(YearGroup) %>%
  summarize(Count = n())
storm_counts_5 <- storm_counts_5[1:(nrow(storm_counts_5) - 1), ]
ggplot(storm_counts_5, aes(x = YearGroup, y = Count, group = 1)) +
  geom_line(color = "blue") +
  labs(x = "Time Period", y = "Count", title = "Hurricane/Typhoon counts every 5 years")
```

```
annual_data_1850 = annual_data %>% filter(Year > 1849)
ggplot(data = annual_data_1850, aes(x = Year, y = LandAverageTemperature)) + geom_line()
+ ylab(" Average Temperature") + ggtitle("Average Temperature over Time")
```

Simple Linear Regression Model - Storm Count ~ Average Global Land Temperature (Yearly)

```
storm_counts <- storm %>%
  mutate(YearGroup = floor(as.numeric(format(Date, "%Y")))) %>%
  group_by(YearGroup) %>%
  summarize(Count = n())
storm_counts <- storm_counts[1:(nrow(storm_counts) - 1), ]

storm_counts_temp <- merge(annual_data, storm_counts, by.x = "Year", by.y = "YearGroup",
all.x = TRUE, all.y = FALSE)


mlr_model <- lm(Count ~ LandAverageTemperature, data = storm_counts_temp)
summary(mlr_model)
```

- The average temperature and the storm counts follow a similar trend, could potentially be a causal relationship.

Analyzing impactful Climate Change laws/movements

```
events_df <- data.frame(
  Year = c(2015, 1997, 1970, 1958),  # Add the relevant years for the impactful events
  Event = c("Paris Agreement", "Kyoto Protocol", "First Earth Day", "Keeling Curve")
)

ggplot() +
  geom_vline(data = events_df, aes(xintercept = Year, color = Event), linetype = "dashe
d") +
  geom_line(data = annual_data_1850, aes(x = Year, y = LandAverageTemperature), color =
"black") +
  labs(x = "Year", y = "Average Temperature", title = "Impactful Events on Reducing Glob
al Temperature") +
  theme_minimal() +
  theme(legend.position = "bottom")
```

```
events_df <- data.frame(
  Year = c(2015, 1997, 1970),  # Add the relevant years for the impactful events
  Event = c("Paris Agreement", "Kyoto Protocol", "First Earth Day")
)
ggplot() +
  geom_vline(data = events_df, aes(xintercept = Year, color = Event), linetype = "dashe
d") +
  geom_line(data = yearly_data, aes(x = Year, y = Per.Capita), color = "black") +
  labs(x = "Year", y = "Per Capita C02 Emissions", title = "Impactful Events on Reducing
Global Emissions") +
  theme_minimal() +
  theme(legend.position = "bottom")
```

- I wouldn’t say these events have had much impact yet on global temperature or emissions.