Project Draft

Due: October 30, 11:59pm

```
library(tidyverse)
library(broom)
library(caret)
climate <- read_csv("data/climate.data.processed.csv")
```

INTRODUCTION AND DATA

As the world struggles to convince so many people about the urgency of climate change it needs to be known that the threat we face is not just that of rising sea levels or CO2 levels—it is that of losing our homes. It is that of entire cities having to uproot and move elsewhere because they can no longer sustain themselves. Far from just a small increase in temperature, but a disruption of our lives as we know it.

Each year, tens of millions of people are driven from their homes by floods, storms, and droughts. The Ecological Threat Register, conducted by The Sydney-based Institute for Economics and Peace (IEP), measures ecological threats over 157 independent states and territories. The report projects that as many as 1.2 billion people around the world could be displaced by 2050 (Institute for Economics and Peace, 2020). The adverse effects of global climate change will induce more extreme weather, growing food and water insecurity, and rising sea levels which will cause the number of displaced people to rise (UNHCR, 2019). The IEP report additionally identifies three clusters of ecological hotspots: the Sahel-Horn belt of Africa, from Mauritania to Somalia; the Southern African belt, from Angola to Madagascar, and the Middle East and Central Asian belt, from Syria to Pakistan.

The intersection of climate change and migration requires comprehensive data analysis and solutions to the multidimensional challenges it creates (Podesta, 2019). Therefore, analyzing the dynamics between climate change indicators and displaced people not due to conflict can reveal opportunities for interventions.

Our primary goal in this project is to understand the correlation between climate change indicators and the refugee flow from at-risk countries. In order to make the analysis more manageable, we will focus on the climate change and refugee data of the Middle East region.

According to the WBD (World Bank Database), most of the data from the data set comes directly from each country in the World Bank Group's national statistical systems. The data itself contains many variables, and the series name tells us the metric for which we are getting data. Within each series, the data is broken down into the data for each nation, and each row below the country name corresponds with that country's data for that series name for each year between 1980-2029. This data set contains all the markers that the WDB has tracked in association with climate change in almost every country on Earth. This includes variables such as CO2 emissions levels of every country and agricultural output of each country. Unfortunately, the WDB does not have very complete data for some of the variables. However, we will focus mainly on variables that have sufficient data points, unless the variable is unlikely to change much over time, in which case we have decided to turn these into binary variables.

We will start by examining variables in the WDB data that scientific literature identifies as climate change predictors such as CO2 emissions (as measured in kiltons), other greenhouse emissions (in kilotons), and percentage of arable land as well as overall refugee flow out of the countries. Then, we will analyze the correlation between displacement and climate indicators. The predictors will be used to build a linear model that attempts to explain some of the variance in refugee flow. While refugee flow from a country can be influenced by an almost innumerable amount of variables, the hope for the project is not predict all of that

variance. A model that is able to significantly predict 20% of the variance in refugee flow would be a useful result. We hypothesize that CO2 emissions (as measured in kilotons), other greenhouse emissions (in kilotons), and percentage of arable land will be able to explain, with statistical significance, at least 15% of the variance in refugee flow from countries in the Middle East and Central Asian Belt, an at-risk region identified by The Ecological Threat Register.

METHODOLOGY

Visualizations and Exploratory Data Analysis

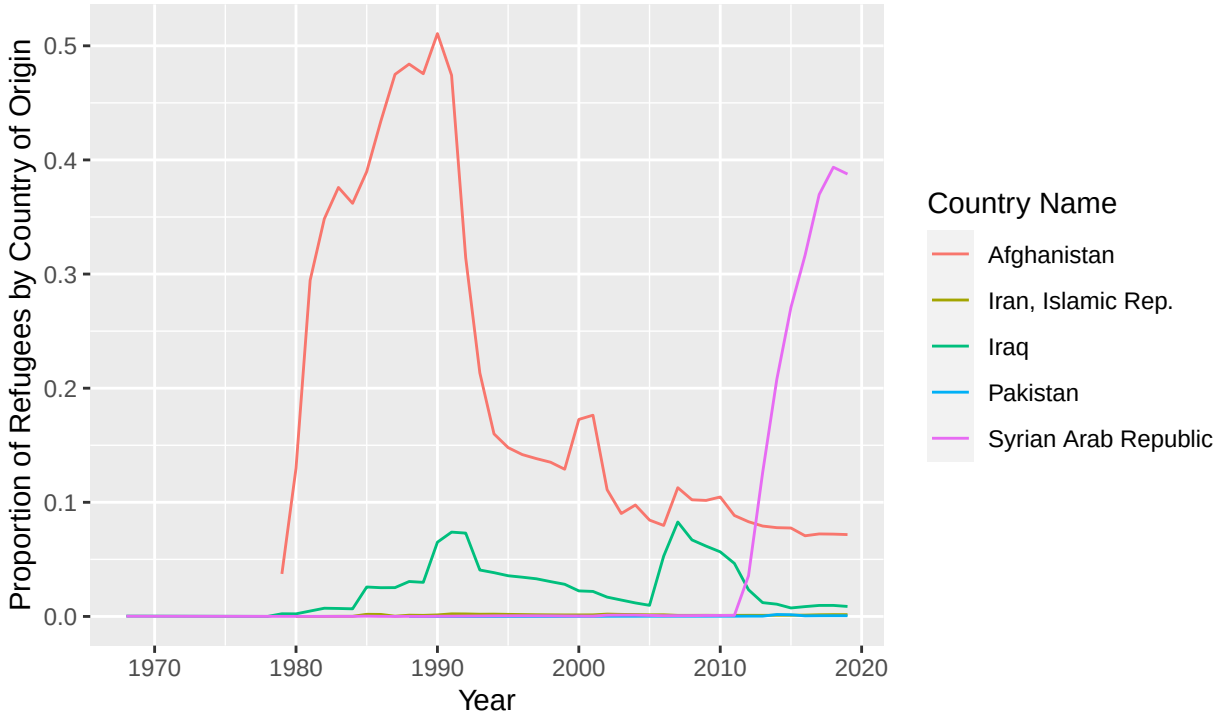
To visualize climate change predictors within each Middle Eastern country over time, we plotted the yearly proportion of population leaving as refugees, CO2 emissions, N2O emissions, methane emissions, percent of country's land that was arable, hectares of arable land used for cereal cultivation.

```
# refugee -----------

climate_refugee <- climate_me %>%
  select(`Time`, `Country Name`, refugee, prop_refugee)%>%
  filter(prop_refugee != is.na(prop_refugee))

ggplot(data = climate_refugee, mapping = aes(x = `Time`,
                                             y= prop_refugee,
                                      color = `Country Name`,
                                      fill = `Country Name`)) +
  #facet_wrap( ~ `Country Name`) +
  geom_line() +
  labs(x = "Year", y = "Proportion of Refugees by Country of Origin",
       title = "Graph 1: Number of refugees in Iran and Pakistan remained
       relatively similar while that in Syria increased sharply after 2011
       civil war")
```

Graph 1: Number of refugees in Iran and Pakistan remained relatively similar while that in Syria increased sharply after 2011 civil war



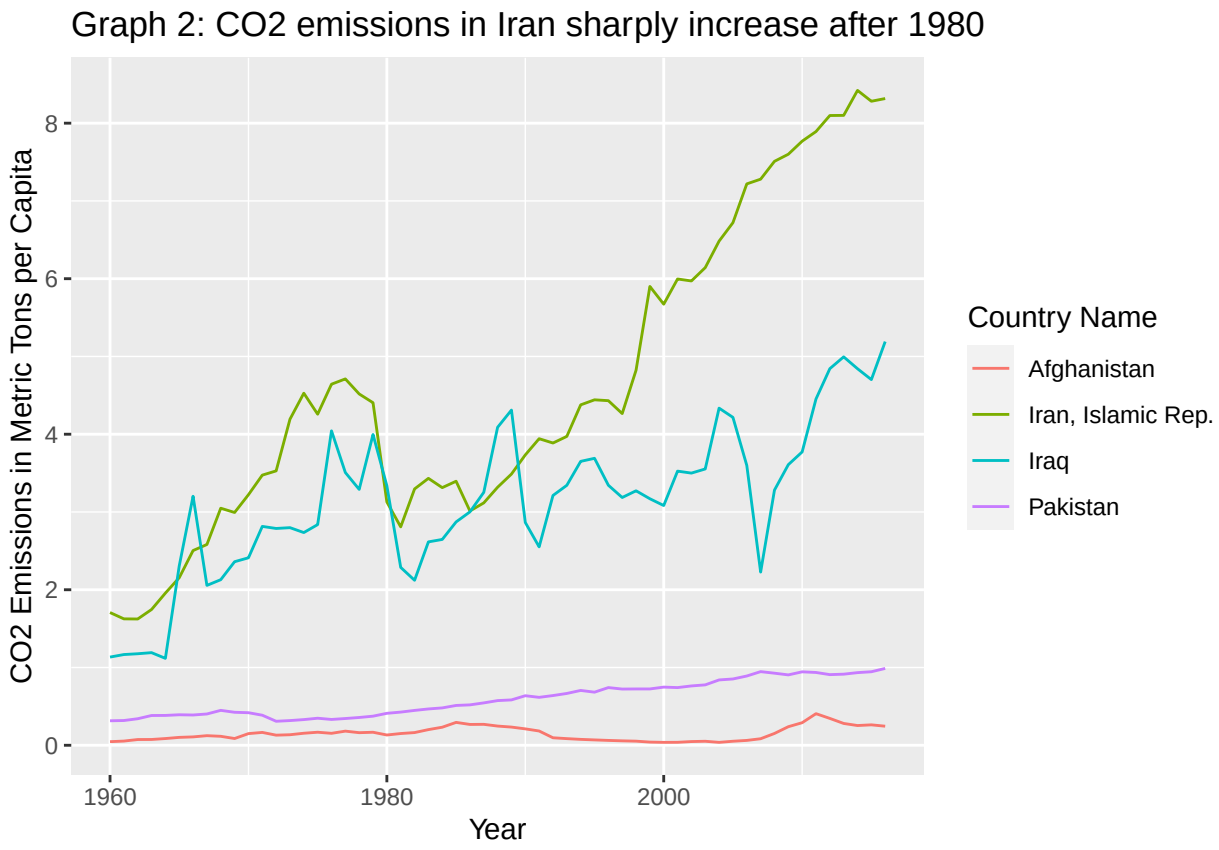
The region as generally defined includes 5 countries, Syria, Iran, Iraq, Afghanistan, and Pakistan. All of these were included in the modeling for our analysis except for Syria. Syria had barely any refugees leaving their country until the start of the Syrian Civil War in 2011. It is clear in the exploratory data analysis, specifically the line graph of refugee population (as proportion of overall population) overtime, that there is a massive spike. That massive increase in refugee population is clearly due to little more than the conflict. Thus data from that country was excluded. However, one can also see significant spikes in refugee population fleeing Afghanistan. Despite this, Afghanistan was included in the model. This is because the refugee population spikes do not correlate as well with periods of war/conflict. The number of refugees fleeing Afghanistan actually decreased dramatically during civil wars in the 1990's. While it is undeniable that Afghanistan did experience periods of civil wars and attacks by foreign powers, that alone is not enough to disqualify them. The refugee population fleeing Afghanistan has existed at significant levels regardless of period of conflict or not. This is also the case with Iraq. While the country experienced conflict during the period of study, the refugee trends do not correlate well with that conflict. In fact, the refugee population from Iraq was still slightly decreasing at the start of the American invasion. This was the justification for including Iraq in the data. More broadly, the mere existence of conflict does not justify the exclusion of a country from our analysis. It is unlikely that climate factors alone would push someone to leave their country of residence. Moreover, the goal of this project is not to entirely explain the variance in refugee population with only these factors. That said, we wanted to observe the how influential climate factors are when there are other factors (i.e. conflict) that would increase a person's desire to leave their country of residence. Thus countries that have had significant refugee populations both with and without conflict present are being included.

```
# co2 -----------

climate_co2 <-climate_data %>%
  select(`Time`, `Country Name`, co2_tons)%>%
  filter(co2_tons != is.na(co2_tons))

ggplot(data = climate_co2, mapping = aes(x = `Time`,
```

```
                                        y= co2_tons,
                                    color = `Country Name`,
                                    fill = `Country Name`)) +
  geom_line() +
  labs(x = "Year", y = "CO2 Emissions in Metric Tons per Capita",
       title = "Graph 2: CO2 emissions in Iran sharply increase after 1980")
```

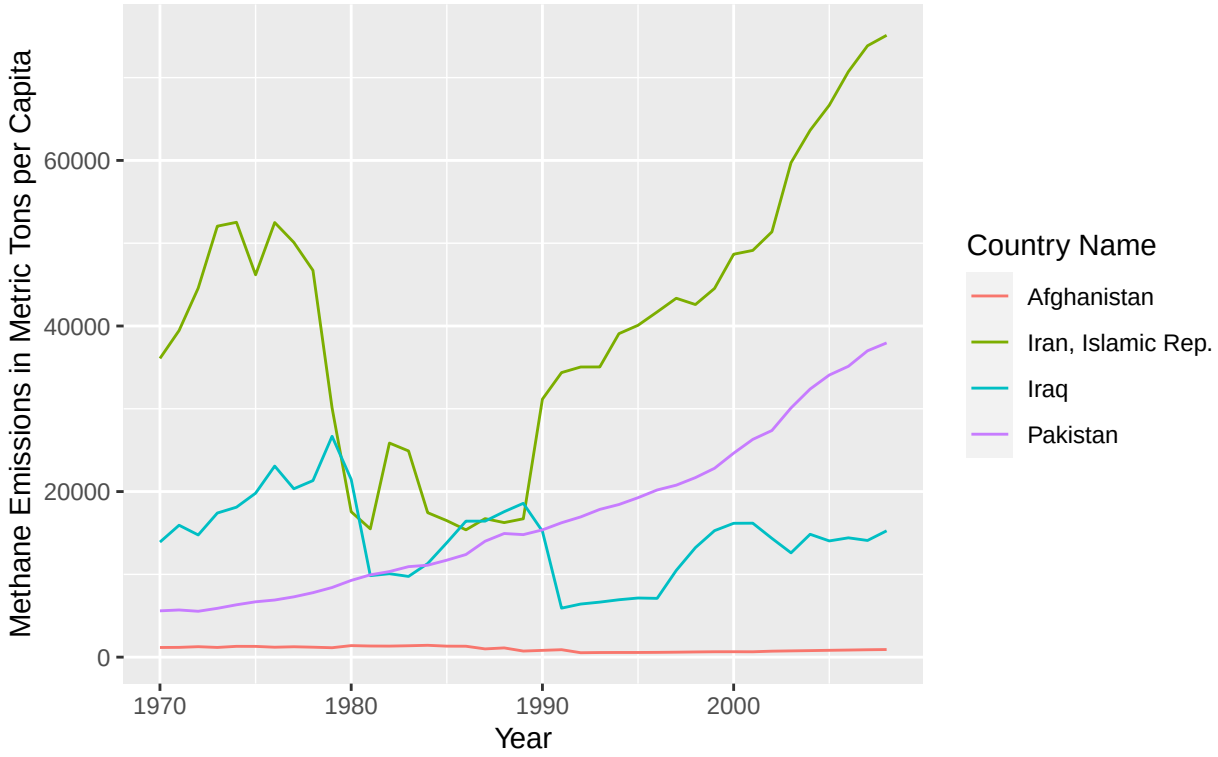



```
# methane -----------

climate_methane <- climate_data %>%
  select(`Time`, `Country Name`, methane_tons)%>%
  filter(methane_tons != is.na(methane_tons))

ggplot(data = climate_methane, mapping = aes(x = `Time`,
                                        y= methane_tons,
                                    color = `Country Name`,
                                    fill = `Country Name`)) +
  geom_line() +
  labs(x = "Year", y = "Methane Emissions in Metric Tons per Capita",
       title = "Graph 3: Methane emissions in Afghanistan remaine relatively
       constant")
```

Graph 3: Methane emissions in Afghanistan remaine relatively constant

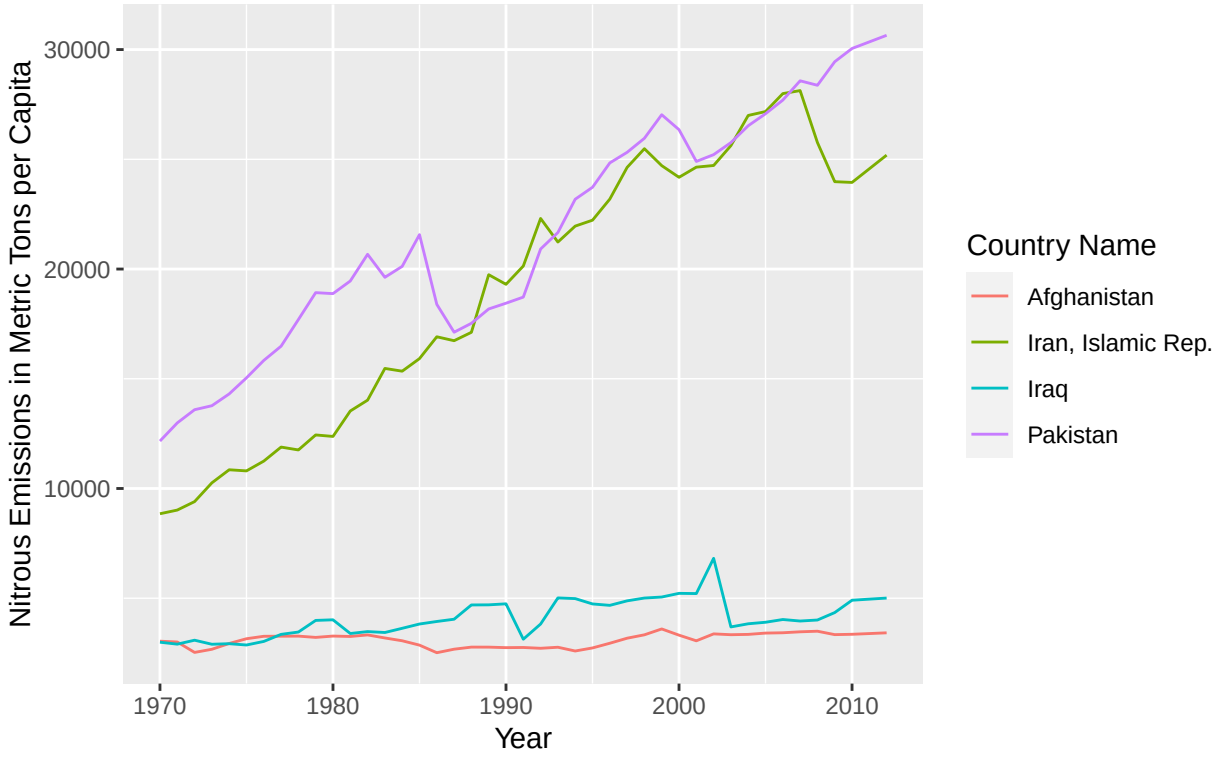



```
# nitrous ----------

climate_methane <- climate_data %>%
  select(`Time`, `Country Name`, nitrous_tons) %>%
  filter(nitrous_tons != is.na(nitrous_tons))

ggplot(data = climate_methane, mapping = aes(x = `Time`, y = nitrous_tons,
                                             color = `Country Name`,
                                             fill = `Country Name`)) +
  geom_line() +
  labs(x = "Year", y = "Nitrous Emissions in Metric Tons per Capita",
       title = "Graph 4: Nitrous emissions increase at similar rates for Iran
       and Pakistan ")
```

Graph 4: Nitrous emissions increase at similar rates for Iran and Pakistan

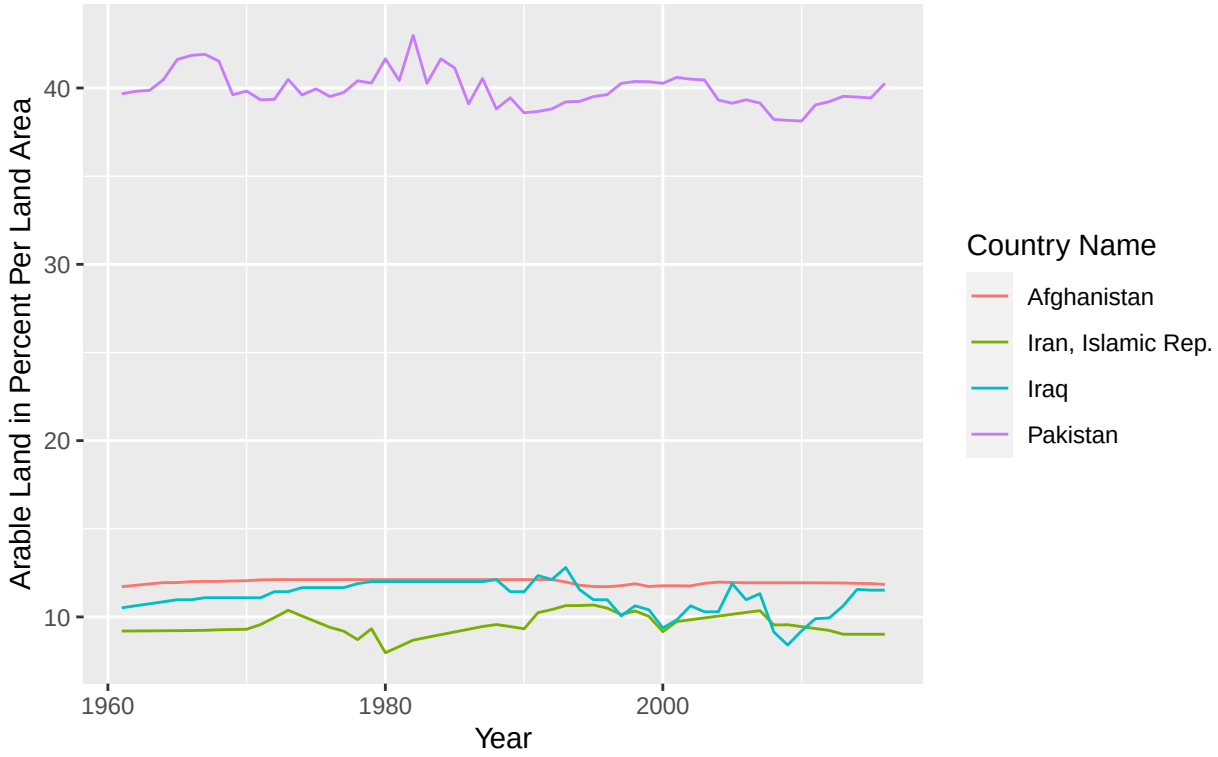



```
# arable_land ----------

climate_arable <- climate_data %>%
  select(`Time`, `Country Name`, arable_land) %>%
  filter(arable_land != is.na(arable_land))

ggplot(data = climate_arable, mapping = aes(x = `Time`, y = arable_land,
                                            color = `Country Name`,
                                            fill = `Country Name`)) +
  geom_line() +
  labs(x = "Year", y = "Arable Land in Percent Per Land Area",
       title = "Graph 5: Percentage of arable land remained relatively constant
       for all four countries ")
```

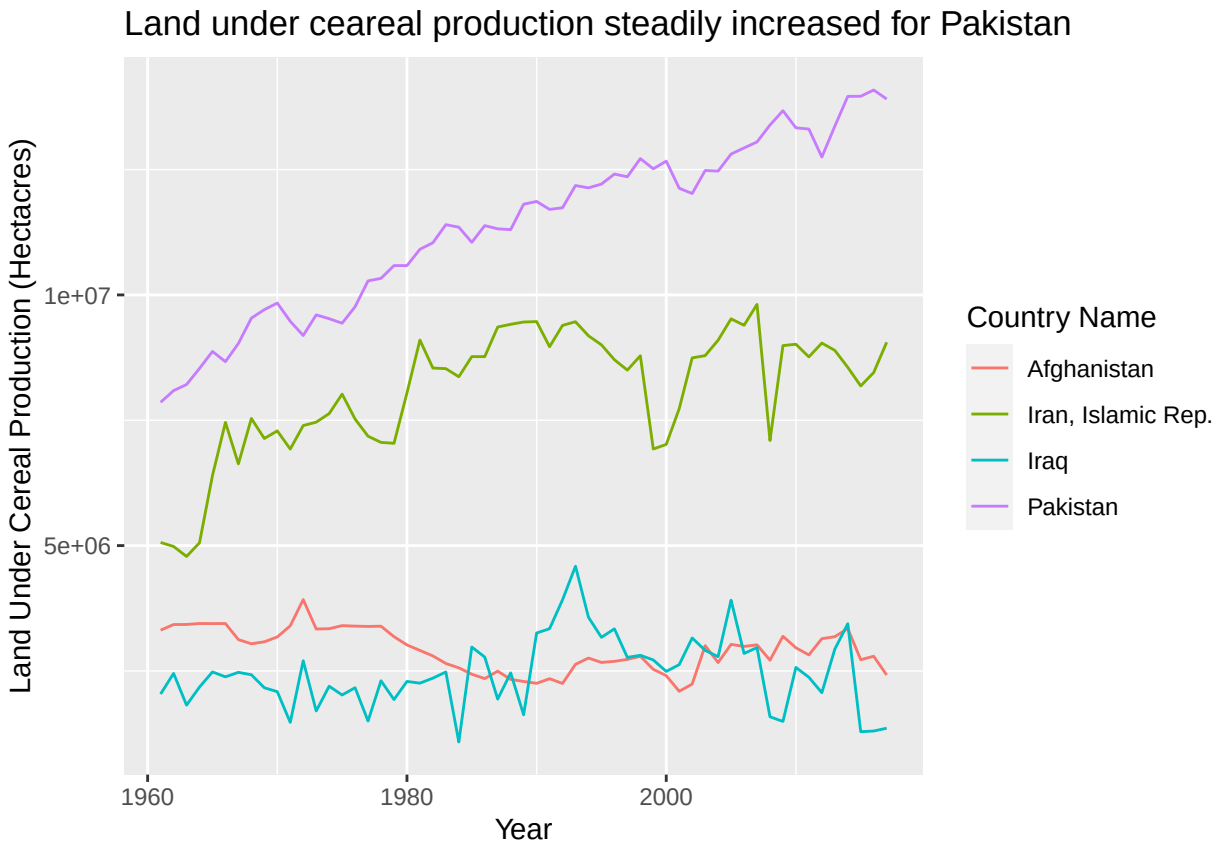
Graph 5: Percentage of arable land remained relatively constant for all four countries




```
# land_cereal ----------

climate_cereal <- climate_data %>%
  select(`Time`, `Country Name`, land_cereal) %>%
  filter(land_cereal != is.na(land_cereal))

ggplot(data = climate_cereal, mapping = aes(x = `Time`, y = land_cereal,
                                            color = `Country Name`,
                                            fill = `Country Name`)) +
  geom_line() +
  labs(x = "Year", y = "Land Under Cereal Production (Hectacres)",
       title = "Land under ceareal production steadily increased for Pakistan")
```



Linear Regression Model with Log Transformation and Assessing Quality of Fit

In order to analyze our data, we will run multiple linear regressions on internally displaced people associated with disasters and other variables that have been associated with displacement. We will look at regions of countries in the world in order to make the data analysis easier for us. We would like to see which variables are predictors of internal displacement. Seeing these predictors will allow us to easily see which variables are associated with high amounts of displacement due to disaster. Some of the variables that we will focus on include annual rainfall, agricultural yield, and the percentage of land that is low-lying in the country. For some of the variables with fewer data points but are unlikely to change over the past 20 years, such as the percentage of land that is low-lying, we will turn these variables into binary categorical variables. This will allow us to use these variables in our data analysis.

Once we identify some of the better predictors of displacement, we would like to run two sample hypothesis tests to see if there are differences in certain variables in countries that are prone to displacement. Specifically, we could run two sample hypothesis tests between high- and low-displacement regions as mentioned in the introduction with relation to certain predictor variables. This way, we could see the extent to how different these variables are for these two classifications. Additionally, we would like to run a Chi-Squared test on the displacement due to disasters and displacement due to conflict. We would like to run this test because it can help us elucidate whether these displacements are related. We would run this test because, if there is a high amount of displacement due to conflict, this could affect the displacement due to disaster, so we would want to see if these variables are independent.

```
climate_model <- lm(log(prop_refugee) ~ arable_land + land_cereal + co2_tons +
                      nitrous_tons + methane_tons,
                    data = climate_data)

tidy(climate_model, conf.int = TRUE, conf.level = 0.95)
```

```
climate_model_aug <- augment(climate_model)
climate_model_aug

glance(climate_model)%>%
  pull(r.squared)

climate_model %>%
  tidy() %>%
  select(term, estimate) %>%
  mutate(estimate = round(exp(estimate), 3))

#Worth conducting model with interactions; Look at p-values for interactions

#FURTHER MODELS/TESTING
#Apply equation to different country; could try to predict a country outside the region and see how wel
#Compare refugee numbers between countries from other regions with different climate but that have simi
#Conducting 2 sample t.test for refugees; filter out periods of war
```

Diagnostic Plots for Linear Regression Model

Since the dataset includes many different predictor variables, it was first necessary to understand which would be the most powerful holding the others constant. In order to make inferences in regression, certain conditions must be met. In order to simplify the process, a simple model with one predictor and one response was created. The first two are linearity which requires that the relationship between the variables and the predictor be linear and equal variance which means that residuals have relatively constant variance. These conditions were checked by creating a basic linear model and plotting the residuals such that we could examine their variance and linearity.

```
ggplot(climate_model_aug, mapping = aes(x = .fitted, y = .std.resid)) +
  geom_point() +
  geom_hline(yintercept = 0, lwd = 2, col = "red", lty = 2) +
  labs(x = "Value", y = "Residuals")
```

The next is independence which requires that the residuals be independent. This is checked very easily by simply plotting all the residuals on a graph.

```
ggplot(data = climate_model_aug,
       aes(x = 1:nrow(climate_data),
           y = .std.resid)) +
  geom_point() +
  labs(x = "Index", y = "Residual")
```

The next requirement is that of normality. The checks for this involved both a plotting of the residuals on a histogram as well as the creation of a Q-Q plot. The histogram was examined for a potential relationship between the residuals and the fit of the model to the qq line was checked too.

```
ggplot(climate_model_aug, mapping = aes(sample = .std.resid)) +
  stat_qq() +
  stat_qq_line()

ggplot(climate_model_aug, mapping = aes(x = .std.resid)) +
  geom_histogram() +
  labs(x = "Residuals")
```

Given that the data passed these conditions checks we decided to perform them with a much more complicated model in order to observe the power of our predictors in the presence of each other.

Applying the Model to Another Country

The other region to which our model can be tested is the Sahel Belt of Africa. This is another region identified by the WORLD REPORT to be an at risk region. It includes Senegal, Mauritania, Mali, Burkina Faso, Niger, Chad, Sudan, Eritrea. These countries have similar conflict levels and similar economic situations thus lending it to another application of our model.

```
climate_sahel <- climate %>%
  filter(`Country Name` == "Burkina Faso" | `Country Name` == "Senegal" |
           `Country Name` == "Mauritania" | `Country Name` == "Niger" |
           `Country Name` == "Mali")

climate_mod_sahel <- lm(log(prop_refugee) ~ arable_land + land_cereal + co2_tons +
                      nitrous_tons + methane_tons,
                    data = climate_sahel)

tidy(climate_mod_sahel, conf.int = TRUE, conf.level = 0.95)
```

```
## # A tibble: 6 x 7
##   term           estimate   std.error statistic    p.value   conf.low  conf.high
##   <chr>             <dbl>       <dbl>     <dbl>      <dbl>      <dbl>      <dbl>
## 1 (Intercep~     -5.93e+0 1.01           -5.87     6.97e-8   -7.94e+0   -3.92e+0
## 2 arable_la~     -6.25e-2 0.0582         -1.07     2.86e-1   -1.78e-1    5.32e-2
## 3 land_cere~      6.78e-8 0.000000129     0.526    6.00e-1   -1.88e-7    3.24e-7
## 4 co2_tons        5.53e+0 1.42            3.90     1.85e-4    2.71e+0    8.34e+0
## 5 nitrous_t~     -2.04e-4 0.000115       -1.78     7.88e-2   -4.33e-4    2.40e-5
## 6 methane_t~     -1.93e-3 0.000833       -2.32     2.25e-2   -3.59e-3   -2.80e-4
```

```
climate_model_aug_sahel <- augment(climate_mod_sahel)
climate_model_aug_sahel
```

```
## # A tibble: 98 x 12
##    .rownames `log(prop_refug~ arable_land land_cereal co2_tons nitrous_tons
##    <chr>                <dbl>       <dbl>       <dbl>    <dbl>        <dbl>
##  1 122                 -11.9       10.7       1846796   0.0618        3043.
##  2 127                 -11.4       10.9       2233598   0.0617        3192.
##  3 132                 -11.2       11.1       2701884   0.0606        3333.
##  4 137                 -12.0       11.2       2551189   0.0635        3480.
##  5 145                 -14.2       16.1       1217422   0.387         2450.
##  6 146                  -3.72       0.369      193452   1.44          1420.
##  7 150                  -6.19      16.1       1259076   0.506         2613.
##  8 151                  -3.43       0.388      118921   0.445         1491.
##  9 155                  -4.83      16.1       1229004   0.423         2968.
## 10 156                  -3.26       0.398      156051   0.441         1550.
## # ... with 88 more rows, and 6 more variables: methane_tons <dbl>,
## #   .fitted <dbl>, .std.resid <dbl>, .hat <dbl>, .sigma <dbl>, .cooksd <dbl>
```

```
glance(climate_mod_sahel)%>%
  pull(r.squared)
```

```
## [1] 0.6053138
```

```
climate_mod_sahel %>%
  tidy() %>%
  select(term, estimate) %>%
  mutate(estimate = round(exp(estimate), 3))
```

```
## # A tibble: 6 x 2
##   term         estimate
##   <chr>           <dbl>
## 1 (Intercept)     0.003
## 2 arable_land     0.939
## 3 land_cereal     1
## 4 co2_tons      251.
## 5 nitrous_tons    1
## 6 methane_tons    0.998
```

Testing for Multicollinearity

```
climate_model <- lm(log(prop_refugee) ~ arable_land + land_cereal + co2_tons +
                      nitrous_tons + methane_tons,
                    data = climate_data)

car::vif(climate_model)
```

```
##  arable_land  land_cereal     co2_tons nitrous_tons methane_tons
##     4.262434    24.421321     6.380504    42.350822    16.966068
```

RESULTS

DISCUSSION

Given the rapidly accelerating nature of climate change, it is imperative that more data analysis be done on the massive impacts it has. Our goal was to see what climate change factors like CO2 emissions and agricultural outputs would influence refugee flow from a country. Refugee flow was picked as a terminal measure of climate change impacts on a country. With so many people denying that climate change impacts their lives because it just makes some days hotter, we decided to show a very concrete and serious impact on the lives of millions of people. Using a linear model allowed us to see not only the direct effects of several different variables all at play at once, but also their interactions. In all of the countries we analyzed, climate change was not the only factor that influenced refugee flow. However, the ability to explain even 20% of the variance in something as intensely complicated as refugee flow with a linear model would be quite the feat. In this experiment we were actually able to obtain and adjusted R squared of ADJ R SQUARED HERE. GOOD result: This was surprising and could be due to LIMITATIONS, but the conditions for inference were mostly satisfied leading us to believe we had constructed a robust model. BAD result: While we were not able to construct a model that explained a significant amount of the variance in refugee flow, this was still a good exercise in trying to uncover relationships that could help the general understanding of migration.

Our variables were X X X X X. Some like X and X were chosen because they are very obvious markers of human impact on the climate. Additionally X and X were chose because they are downstream markers of anthropogenic climate change that are closer to affecting human lives. There was also an interaction effect between X and X measured for THIS REASON.

The significance of the main effects for X X X were Y Y Y respectively. These are great results indicating there likely is a relationship between these variables and refugee flow.

Overall, this not only served as an excellent exercise for us in applying linear regression to the real world, but also can act as the basis for future work trying to develop a quantitative prediction of refugee flow in the future.

Limitations and Concerns

The main concern with the validity of our models here was the lack of data. The World Bank data which was used was missing for several years and we had to develop our regional groupings due to lack of collation on the part of the World Bank. Moreover, robust, comprehensive data on refugee populations from each country

doesn't begin until the late 1980s and early 1990s whereas other metrics such as carbon dioxide emissions begin in the 1960s. In addition to the lack of data, the data on refugee populations was a raw sum and didn't disaggregate the data based on reason for refugee status. Further studies on the relationship between climate predictors and displaced populations should look at refugees and displaced people due to climate change. Though the World Bank data includes data for internally displaced people due to disasters, this metric lacks enough data for analysis.