# Project 1

#Project 1 ###Audrey Malcolm ###aem3886

Chosen Datasets

The first three datasets includes the anuual emissions of methane CO2, and total greenhouse gasses from various countries since the 1960s.The datasets are from the World Bank’s statistics. CO2 emissions are recorded in kilotons (kt). Moreover, methane emissions are recorded as kts of a CO2 equivalent. This variable takes into account the warming potential of the greenhouse gasses in order to scale for each gas’s contribution to global warming. Because one methane atom releases 25 times more heat than a CO2 atom, kts of a CO2 equivalent is the methane emissions in kts times 25. The total greenhouse gas emissions for each country was also measured in kts of CO2 equivalent. The fourth and fifth datasets include the annual number and rate of serious assaults and robberies in various countries since 2003. The crime data was sourced from the United Nations statistics The datasets were chosen to give insigt to how global warming may affect psychology of populations, especially in regard to violent crimes. I expect there to be some correlation between crimes and emissions, but I can’t for sure say the emissions are what cause the increase in crimes. It is more likely a number of factors that influence the amount of violent crime in a region including income, education, and mental health resources.

Datasets: https://data.worldbank.org/indicator/EN.ATM.CO2E.KT (https://data.worldbank.org/indicator/EN.ATM.CO2E.KT) https://data.worldbank.org/indicator/EN.ATM.CO2E.KT (https://data.worldbank.org/indicator/EN.ATM.CO2E.KT) https://data.worldbank.org/indicator/EN.ATM.GHGT.KT.CE (https://data.worldbank.org/indicator/EN.ATM.GHGT.KT.CE) https://dataunodc.un.org/crime/serious_assault (https://dataunodc.un.org/crime/serious_assault) https://dataunodc.un.org/crime/robbery (https://dataunodc.un.org/crime/robbery)

Tidy Data into ‘Emissions’ and ‘Crime’

The datasets were tidied using a variety of functions in order to place year and emissions in their own column. Then, the CO2, methane, and total emissions data were joined into one dataset called ‘Emissions’ using a left join so the country name, region, and year variables remained common in the new dataset. The robbery and assault data were already tidy, so they were simply joined into one dataset called ‘crime.’

```
library(tidyverse)
```

```
## -- Attaching packages ------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts --------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readr)

robbery<-robbery%>%select(-7)
robbery[, -c(1:3)] <- sapply(robbery[, -c(1:3)], as.numeric)
crime<-left_join(robbery, serious_assault, by=c("Year","Country","Sub Region", "Region"))
head(crime)
```

```
## # A tibble: 6 x 8
##   Region `Sub Region` Country  Year Robbery_Count Robbery_Rate
##   <chr>  <chr>        <chr>   <dbl>         <dbl>        <dbl>
## 1 Africa Eastern Afr~ Burundi  2008          4009       0.0401
## 2 Africa Eastern Afr~ Burundi  2009          4231       0.0423
## 3 Africa Eastern Afr~ Burundi  2010          3039       0.0304
## 4 Africa Eastern Afr~ Burundi  2011          4266       0.0427
## 5 Africa Eastern Afr~ Burundi  2012          4246       0.0425
## 6 Africa Eastern Afr~ Burundi  2013          4108       0.0411
## # ... with 2 more variables: Assault_Count <dbl>, Assault_Rate <dbl>
```

```
Co2[, -c(1:4)] <- sapply(Co2[, -c(1:4)], as.numeric)
Co2<-Co2%>%pivot_longer(cols=-c(1:4), names_to = "Year", values_to = "CO2_Emissions")%>%select(-
c(3:4))

methane[, -c(1:4)] <- sapply(methane[, -c(1:4)], as.numeric)
methane<-methane%>%pivot_longer(cols = -c(1:4), names_to = "Year", values_to = "Methane_Emission
s")%>%select(-c(3:4))

total_emission[, -c(1:4)] <- sapply(total_emission[, -c(1:4)], as.numeric)
total_emission<-total_emission%>%pivot_longer(cols = -c(1:4), names_to = "Year", values_to = "To
tal_Emissions")%>%select(-c(3:4))

Emissions<-Co2%>%left_join(methane, by=c("Country Code", "Year", "Country Name"))
Emissions<-Emissions%>%left_join(total_emission, by=c("Country Code", "Year","Country Name"))
head(Emissions)
```

```
## # A tibble: 6 x 6
##   `Country Name` `Country Code` Year  CO2_Emissions Methane_Emissio~
##   <chr>          <chr>          <chr>         <dbl>            <dbl>
## 1 Aruba          ABW            1960             NA               NA
## 2 Aruba          ABW            1961             NA               NA
## 3 Aruba          ABW            1962             NA               NA
## 4 Aruba          ABW            1963             NA               NA
## 5 Aruba          ABW            1964             NA               NA
## 6 Aruba          ABW            1965             NA               NA
## # ... with 1 more variable: Total_Emissions <dbl>
```

Joining the Datasets into ‘Join’

The crime and the emissions datasets were joined using a left join because there are several common variables shared throughout the datasets including year and country name. NA values were ommited so that only countries with both crime data and emissions data were considered in the following analysis. Moreover, reduntant values were removed including robbery rate and assault rate because they represented the same statistics as the robbery and assault count. Similarly, country code was removed because it is the same value as country name. As a result of the removal of NAs, 1,482 cases were reduced to 716. Moreover, many countries were excluded from the analysis because they either didn't have emissions data or crime data for that given year. Some of these countries like the United States and China might have been interesting cases to analyze, but won't be considered because they didn't have overlap in the datasets. This also means that of the data analyzed, the country with the highest emissions might not be the actual highest emissions, but rather highest emissions that also had crime data available.

```
Emissions[, -c(1:2)] <- sapply(Emissions[,-c(1:2)], as.numeric)
crime[, -c(1:3)] <- sapply(crime[,-c(1:3)], as.numeric)

join<-crime%>%left_join(Emissions, by=c("Year","Country"="Country Name"))%>%na.omit%>%select(-6,
-8,-9)
```

##Summary Statistics

Looking at the summary of the numeric data, the robbery count and the assault count range are large, which makes sense when looking at the entire world. All the emissions statistics appear to be heavily right skewed, which indicates most countries are not producing a significant ammount of the world's greenhouse gasses. Rather, a few countries are producing most of the world's greenhouse gasses. The countries with the highest average greenhouse gas emissions are Russia, India, Brazil, and Japan. Similarly, there appears to be a significant right skew in the crime data, so a few countries have high rates of robberies and assaults while the other countries have similar levels of crime. The countries with the highest average crime rates were Brazil and Mexico.

```
join%>%select_if(is.numeric)%>%summary()
```

```
##       Year      Robbery_Count       Assault_Count    CO2_Emissions    
##  Min.   :2003   Min.   :      0.0   Min.   :     0   Min.   : 0.02271  
##  1st Qu.:2006   1st Qu.:    939.5   1st Qu.:  1109   1st Qu.: 1.66652  
##  Median :2008   Median :   3886.0   Median :  6197   Median : 4.63046  
##  Mean   :2008   Mean   :  36471.9   Mean   : 37891   Mean   : 6.14900  
##  3rd Qu.:2010   3rd Qu.:  16680.5   3rd Qu.: 32279   3rd Qu.: 8.54823  
##  Max.   :2012   Max.   :1087059.0   Max.   :732913   Max.   :60.90081  
##  Methane_Emissions  Total_Emissions  
##  Min.   :    27.6   Min.   :    155.3  
##  1st Qu.:  3898.2   1st Qu.:  23282.6  
##  Median : 11760.9   Median :  68847.7  
##  Mean   : 41411.5   Mean   : 240752.8  
##  3rd Qu.: 27487.0   3rd Qu.: 152890.0  
##  Max.   :636395.8   Max.   :3002894.9
```

```
join%>%group_by(Country)%>%summarize(mean_Emissions=mean(Total_Emissions),sd(Total_Emissions))%
>%arrange(desc(mean_Emissions))
```

```
## # A tibble: 109 x 3
##    Country            mean_Emissions `sd(Total_Emissions)`
##    <chr>                       <dbl>                 <dbl>
##  1 Russian Federation       2670707.               176026.
##  2 India                    2495458.               325338.
##  3 Brazil                   2273444.               780553.
##  4 Japan                    1406225.                58505.
##  5 Germany                   975483.                30184.
##  6 Canada                    911177.               104711.
##  7 Indonesia                 846752.               169662.
##  8 Australia                 776528.                12985.
##  9 Mexico                    638916.                27548.
## 10 France                    537658.                22672.
## # ... with 99 more rows
```

```
join%>%group_by(Country)%>%mutate(crime=Robbery_Count+Assault_Count)%>%summarize(mean=mean(crim
e),sd(crime),n(), var(crime), median(crime), mode(crime), min(crime), max(crime), range=max(crim
e)-min(crime), cor(crime,Total_Emissions))%>%arrange(desc(mean))
```

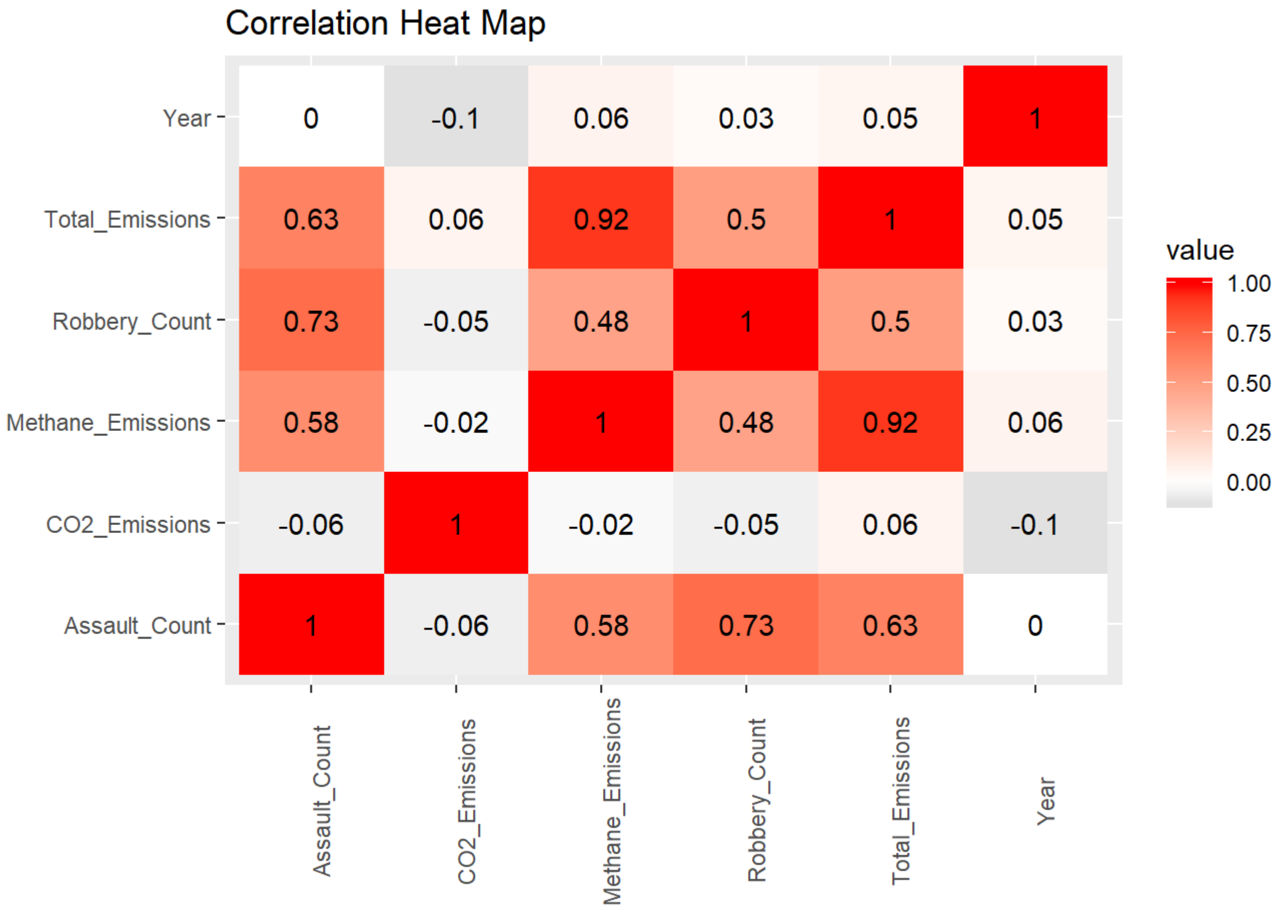
```
## # A tibble: 109 x 11
##    Country   mean `sd(crime)` `n()` `var(crime)` `median(crime)`
##    <chr>    <dbl>       <dbl> <int>        <dbl>           <dbl>
##  1 Brazil  1.66e6     113146.     7 12801989373.        1584544
##  2 Mexico  8.78e5      87792.     9  7707479372.         907809
##  3 Argent~ 4.99e5      23903.     4   571366230.         492881
##  4 Germany 4.06e5     186602.    10 34820150009.         519908
##  5 India   3.08e5      24965.     9   623255144.         301623
##  6 France  3.06e5      39003.    10  1521249679.         293060.
##  7 Russia~ 2.77e5      99601.     9  9920305526.         289393
##  8 Belgium 2.20e5      81893.    10  6706532374.         264849
##  9 Turkey  1.76e5      60078.    10  3609350894.         167388
## 10 Spain   1.24e5      19565.     5   382787536.         113960
## # ... with 99 more rows, and 5 more variables: `mode(crime)` <chr>,
## #   `min(crime)` <dbl>, `max(crime)` <dbl>, range <dbl>, `cor(crime,
## #   Total_Emissions)` <dbl>
```

As for how a country’s emissions and crime rates are related, the correlation matrix shows a high correlation of 0.7 between the number of robberies and the number of assaults. This indicates that within countries, crime seems to increase together. There is also a correlation between the total emissions and the number of assaults with 0.63. The correlation between total emissions and robberies of 0.5. Looking at the countries with the highest total greenhouse gases released, Brazil, Germany and India have the highest rates of crimes. This supports the correlation of these variables based on the fact these countries were among the highest in average crime and emissions.

```
cor<-join_nums<-join%>%select_if(is.numeric)%>%scale()%>%cov

join%>%filter(Total_Emissions>15000)%>%arrange(desc(Assault_Count), desc(Robbery_Count))
```

```
## # A tibble: 573 x 9
##    Region `Sub Region` Country  Year Robbery_Count Assault_Count
##    <chr>  <chr>        <chr>   <dbl>         <dbl>         <dbl>
##  1 Ameri~ South Ameri~ Brazil   2012        979571        732913
##  2 Ameri~ South Ameri~ Brazil   2011       1087059        717185
##  3 Ameri~ South Ameri~ Brazil   2010       1081041        715702
##  4 Ameri~ South Ameri~ Brazil   2006        894978        652778
##  5 Ameri~ South Ameri~ Brazil   2009        910679        651879
##  6 Ameri~ South Ameri~ Brazil   2008        934548        649996
##  7 Ameri~ South Ameri~ Brazil   2007        927667        649027
##  8 Europe Western Eur~ Germany  2007         52949        523283
##  9 Europe Western Eur~ Germany  2008         49913        518499
## 10 Europe Western Eur~ Germany  2006         53696        510775
## # ... with 563 more rows, and 3 more variables: CO2_Emissions <dbl>,
## #   Methane_Emissions <dbl>, Total_Emissions <dbl>
```

```
join%>%select_if(is.numeric)%>%cor%>%as.data.frame%>%  rownames_to_column%>%pivot_longer(-1)%>%
ggplot(aes(rowname,name,fill=value))+geom_tile()+  geom_text(aes(label=round(value,2)))+  xlab(
"")+ylab("") +scale_fill_gradient2(low="black", high = "red") + theme(axis.text.x=element_text(a
ngle=90))+ggtitle("Correlation Heat Map")
```

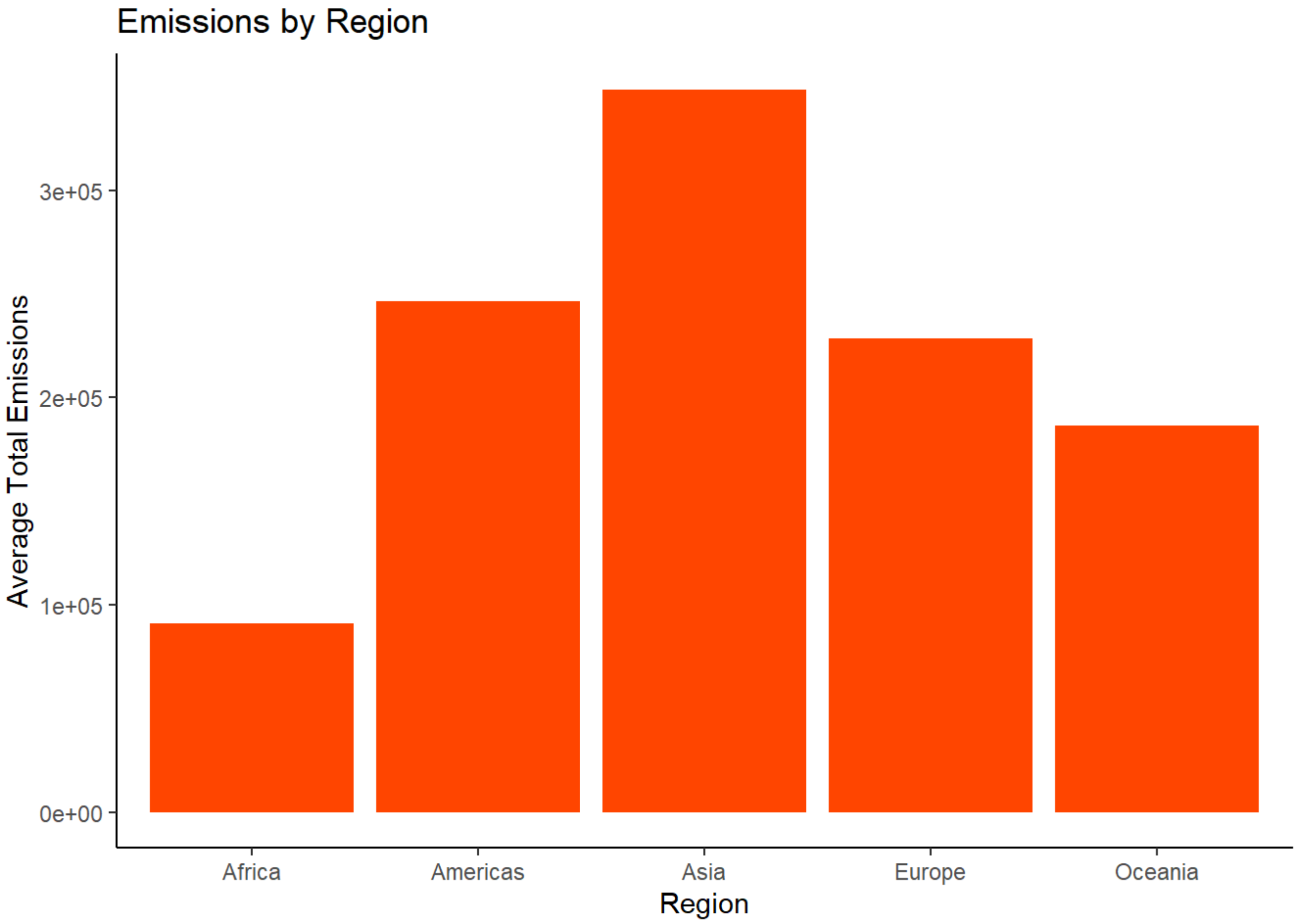


Visualizations

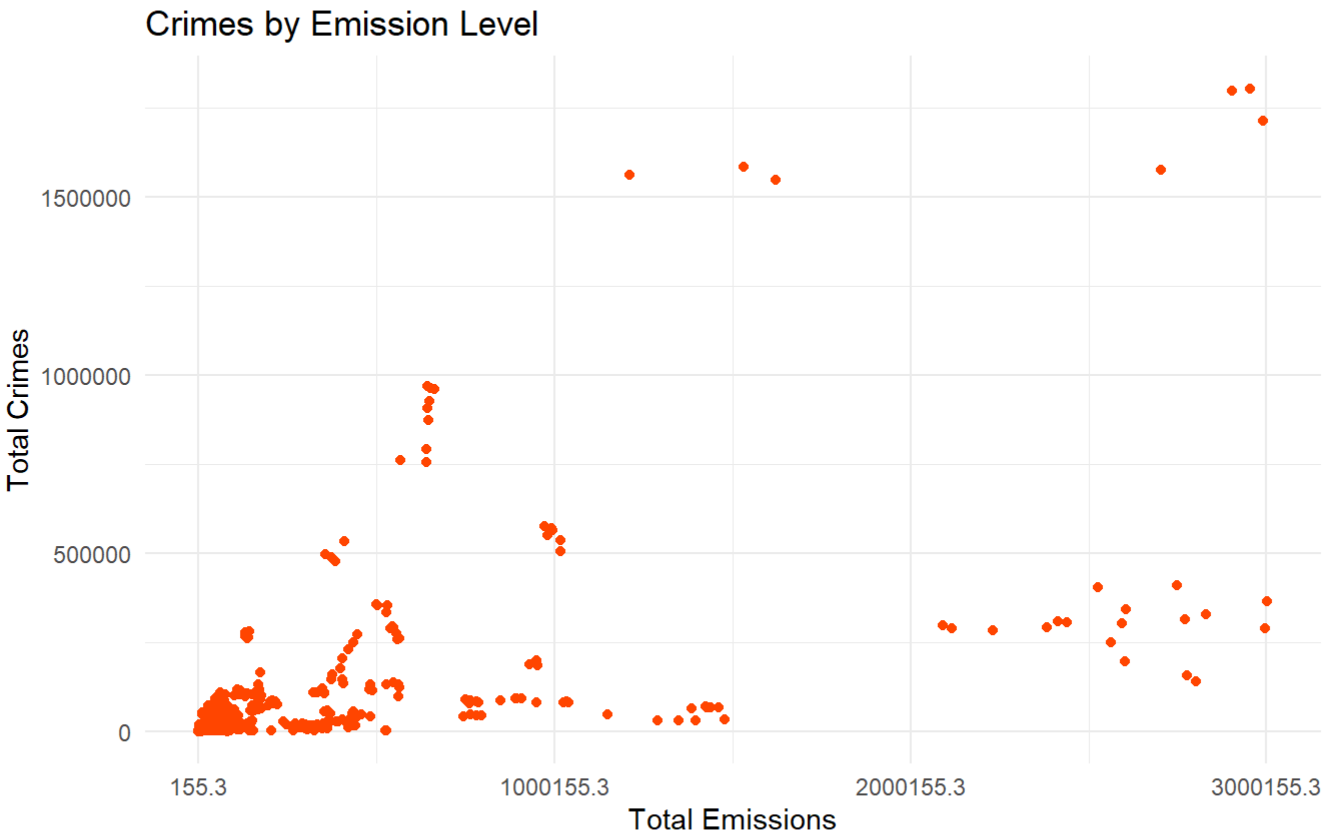
A summary graph of average emissions by region was created to demonstrate where in the world the most emissions are. Asia has the highest overall emissions while Europe, the Americas, and Oceania have about the same emissions. Africa has the lowest average emissions. The amount of emissions A new variable called crime was created as an indicator of a country's overall crime. Crime was the sum of the number of assaults and number of roberies per country. The new crime variable was graphed against the total number of emissions. Most of the points appeared to cluster near 0 in a somewhat linear pattern with some outliers that were either high crime and low emissions, or high emissions and low crime. However, the overall trend appears to be positively correlated.

```
library(ggplot2)

join%>%ggplot(aes(Region, Total_Emissions))+geom_bar(stat="summary", fill="orangered")+labs(titl
e="Emissions by Region", y="Average Total Emissions")+theme_classic()
```

```
## No summary function supplied, defaulting to `mean_se()
```



```
join2<-join%>%mutate(crime=Robbery_Count+Assault_Count)

ggplot(join2)+geom_point(aes(Total_Emissions, crime), color='orangered')+coord_fixed()+labs(titl
e="Crimes by Emission Level", x="Total Emissions", y="Total Crimes")+theme_minimal()+scale_x_con
tinuous(breaks = round(seq(min(join2$Total_Emissions), max(join2$Total_Emissions), by = 1000000
),1)) +scale_y_continuous(breaks = round(seq(min(join2$crime), max(join2$crime), by = 500000),1
))
```



##PCs

After creating a graph for the proportion of variance explained by each principle component, I decided to keep two principal components for the analysis because that seems to be where the elbow is in the graph. PC1 and PC2 are also the only two components with eigenvalues greater than 1.

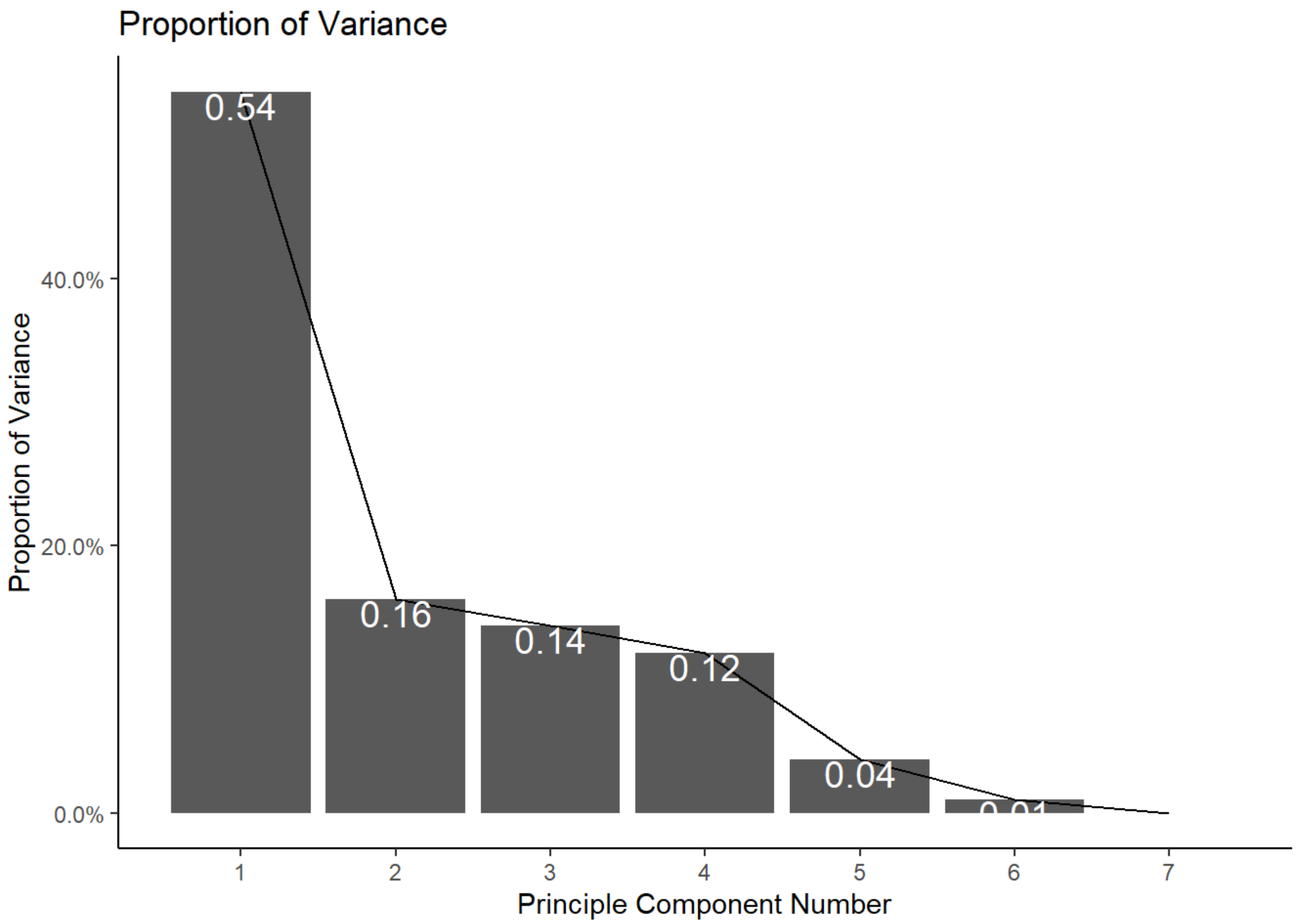
```
nums<-join2%>%select_if(is.numeric)%>%scale
rownames(nums)<-join2$Total_Emissions
pca<-princomp(nums)
summary(pca, loadings=T)
```

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
## Standard deviation     1.9360753 1.0510854 0.9940243 0.9052387 0.5053819
## Proportion of Variance 0.5362328 0.1580465 0.1413523 0.1172290 0.0365383
## Cumulative Proportion  0.5362328 0.6942794 0.8356317 0.9528607 0.9893990
##                           Comp.6       Comp.7
## Standard deviation     0.2722193 3.152771e-08
## Proportion of Variance 0.0106010 1.421981e-16
## Cumulative Proportion  1.0000000 1.000000e+00
## 
## Loadings:
##                   Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
## Year                      0.657  0.571  0.489                     
## Robbery_Count      0.440        -0.305  0.320 -0.606         0.489
## Assault_Count      0.462        -0.182  0.115  0.774         0.363
## CO2_Emissions           -0.740  0.381  0.549                     
## Methane_Emissions  0.417         0.406 -0.413 -0.174  0.676       
## Total_Emissions    0.430 -0.114  0.406 -0.331        -0.726       
## crime              0.483        -0.271  0.250               -0.793
```

```
eigval<-pca$sdev^2
eigval
```

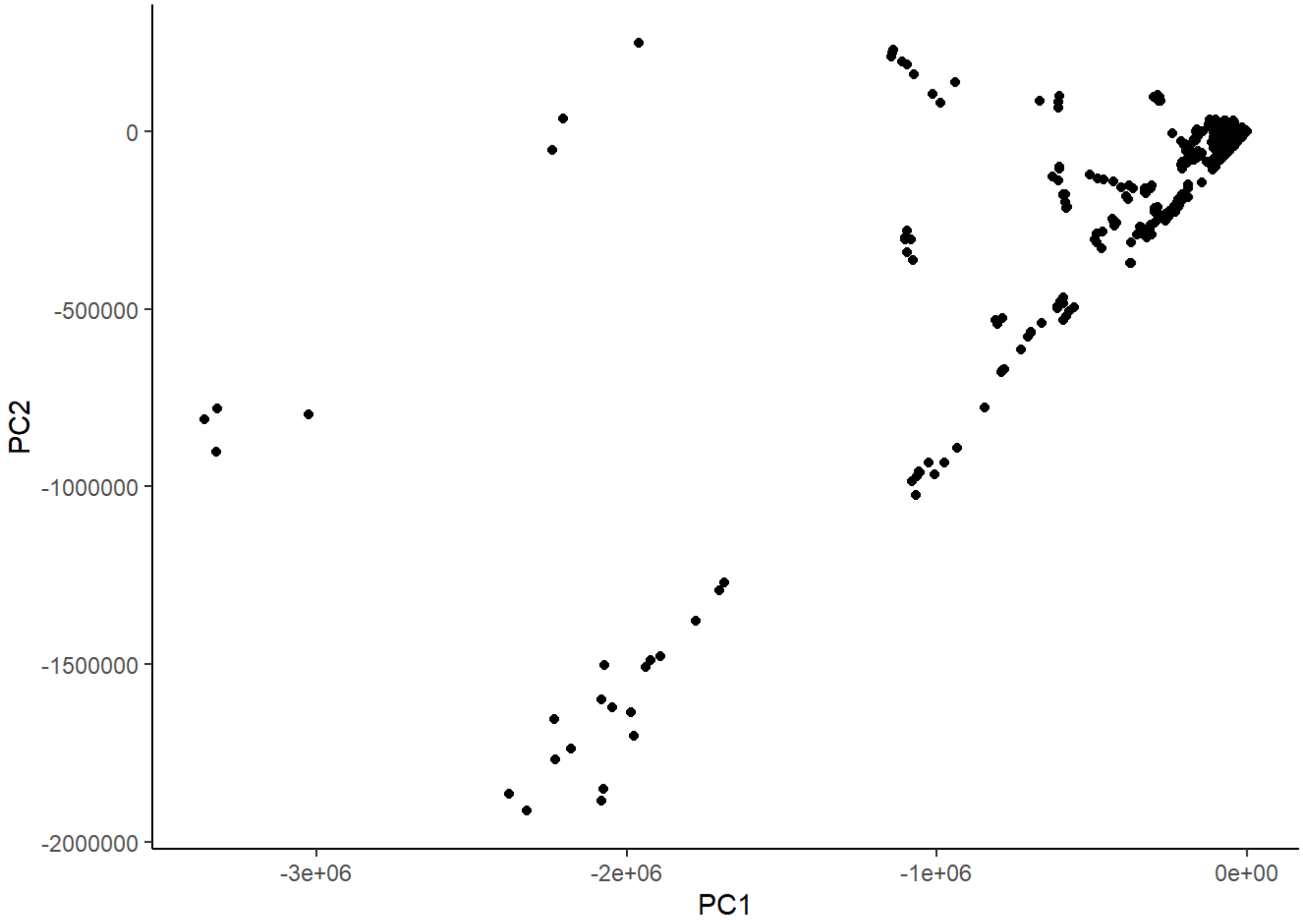
```
##       Comp.1       Comp.2       Comp.3       Comp.4       Comp.5 
## 3.748387e+00 1.104781e+00 9.880843e-01 8.194570e-01 2.554109e-01 
##       Comp.6       Comp.7 
## 7.410333e-02 9.939966e-16
```

```
varprop=round(eigval/sum(eigval),2)
ggplot()+geom_bar(aes(y=varprop,x=1:7),stat="identity")+xlab("")+geom_path(aes(y=varprop,x=1:7))
+  geom_text(aes(x=1:7,y=varprop,label=round(varprop,2)),vjust=1,col="white",size=5)+  scale_y_c
ontinuous(breaks=seq(0,.6,.2),labels = scales::percent)+  scale_x_continuous(breaks=1:10)+theme_
classic()+labs(y="Proportion of Variance", x="Principle Component Number", title="Proportion of
 Variance")
```

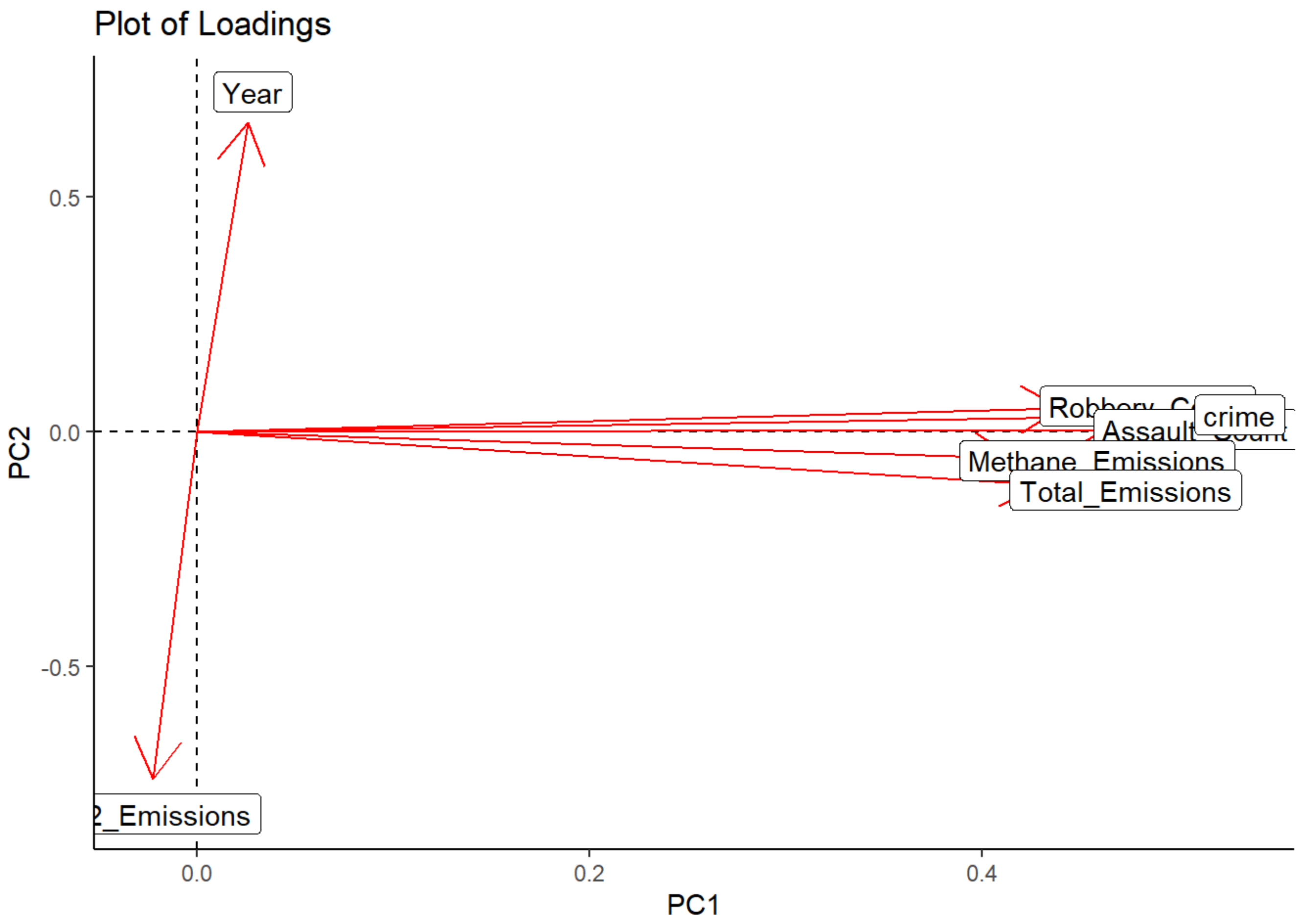


From the summary of PCAs, it is shown that PC1 explains the variation between total emissions, methan emissions, robberies, and assaults. It seems as the total emissions increases, so does the number of crimes. PC2 explains how emissions change over time. The emissions of methane have increased over time, but the emissions of CO2 have decreased.

```
eig<-join2%>%select(Total_Emissions,crime)%>%cor%>%eigen
xy<-join2%>%select(crime, Total_Emissions)%>%as.matrix
scores<--t(t(eig$vectors)%*%t(xy))%>%as.data.frame()
ggplot(scores, aes(V1,V2))+geom_point()+xlab("PC1")+ylab("PC2") + theme_classic()
```



```
pca$loadings[1:7,1:2]%>%as.data.frame%>%rownames_to_column%>%ggplot()+geom_hline(aes(yintercept=
0),lty=2)+  geom_vline(aes(xintercept=0),lty=2)+ylab("PC2")+xlab("PC1")+  geom_segment(aes(x=0,y
=0,xend=Comp.1,yend=Comp.2),arrow=arrow(),col="red")+  geom_label(aes(x=Comp.1*1.1,y=Comp.2*1.1,
label=rowname))+ggtitle("Plot of Loadings")+theme_classic()
```



```
Country<-join2$Country
pca$scores%>%as.data.frame%>%cbind(Country,.)%>%top_n(15,Comp.1)
```

```
##                Country    Comp.1      Comp.2    Comp.3    Comp.4
## 1               Mexico  6.212697  1.03782068 -2.314229  2.757682
## 2               Mexico  6.175897  1.26990499 -2.076579  2.946872
## 3               Mexico  6.165860  1.50339912 -1.833110  3.104607
## 4               Brazil 12.248014  0.09483711 -3.287240  1.617256
## 5               Brazil 13.366051  0.10197786 -2.290853  1.185459
## 6               Brazil 12.462697  0.59630146 -3.006463  2.135977
## 7               Brazil 12.073686  0.91369445 -2.975604  2.431518
## 8               Brazil 15.079921  0.84210366 -2.152488  2.233891
## 9               Brazil 15.288485  1.05426539 -1.815771  2.306011
## 10              Brazil 14.844640  1.21488269 -1.183587  2.075737
## 11               India  6.147832 -0.01778286  3.485173 -3.612155
## 12               India  6.387969  0.19227912  3.803385 -3.524182
## 13               India  6.586261  0.41345446  4.046052 -3.365981
## 14               India  7.031463  0.60983435  4.312236 -3.221170
## 15 Russian Federation  5.986757 -1.72404011  2.540471 -2.174981
##         Comp.5      Comp.6        Comp.7
## 1  -1.99370529 -0.06907615 -8.881784e-15
## 2  -2.19767379 -0.10004570 -1.021405e-14
## 3  -2.16079467 -0.10854165 -9.769963e-15
## 4   0.06789444  0.89393027  5.329071e-15
## 5  -0.16749757 -0.63686124  2.664535e-15
## 6  -0.13807389  1.10996758  3.552714e-15
## 7   0.02503011  1.58242862  4.440892e-15
## 8  -0.39039612 -0.67416800  0.000000e+00
## 9  -0.43999010 -0.56576462  8.881784e-16
## 10  0.21895455 -0.51613404  3.552714e-15
## 11  0.92656585  0.78474221  6.550316e-15
## 12  1.00865264  0.54661139  6.883383e-15
## 13  1.11586032  0.53268739  7.438494e-15
## 14  1.31821308  0.36674265  8.215650e-15
## 15 -2.39482977 -0.53982772 -1.088019e-14
```

```
pca$scores%>%as.data.frame%>%cbind(Country,.)%>%top_n(3,Comp.2)
```

```
##   Country     Comp.1   Comp.2    Comp.3    Comp.4       Comp.5
## 1 Burundi -0.7650898 1.740049 0.4152912 0.3388170 -0.020691893
## 2  Rwanda -0.7535612 1.735054 0.4160087 0.3413993  0.009325632
## 3  Uganda -0.4962320 1.706434 0.4980804 0.2632403  0.088038776
##         Comp.6        Comp.7
## 1 -0.007576093 -4.996004e-16
## 2 -0.002908708 -3.330669e-16
## 3  0.030867171  1.110223e-16
```

A plot of loadings graph was generated that showed crime, methane emissions, and total emissions were highly correlated and contributing to PC1. CO2 emissions and year were not correlated with the other factors and contributed more to PC2. The reason CO2 emissions are highly correlated with crime may be because countries that produce a high amount of CO2 are generally more developed nations with economies involved in manufacturing. Wealth can contribute to decreased crimes, so CO2 emissions may be indicative of the countries overall wealth. Methane is a greenhouse gas generally produced in agriculture, which is a more important contributor to the economy in less developed nations. As such, PC1 can be considered the relative wealth axis, while PC2 is the emissions over time axis. The largest contributing countries to PC1 are Mexico, Brazil, and India. The countries contributing the most to PC2 are Burundi, Rwanda, and Uganda