Writing Assignment 3 - EDA

Melton Team - Dana Kenney, Jacob Voyles, Sydney Yeargers, Tatim Laughery, and William Holbert

3/8/2022

# Data Cleaning

All of our data to this point has been manually entered into Excel spreadsheets. We have been continuously working to source more data and fill in any gaps that remain in the information. A positive of manually entering all our data is that it eliminates the need to undertake any cleaning steps, as we can standardize the data as we enter it.

We are limited by the sources we can pull from. Microsoft’s data reporting practices have seemingly improved over time. While the most recent data is very complete, looking back 8-12 years ago results in data with more NA values (especially in years 2013 and 2014). Because our data is completely manually entered, we are being very diligent to ensure that no data entry errors are made. If we see a data point that appears to be a major outlier, we check the input source for an explanation. Thus far we’ve noticed that outliers seem to occur more when data reporting was spottier (i.e., more NA values) or reporting practices appeared to change.

# Data Reshaping

Currently, all of our data is coming from the Carbon Disclosure Project (CDP). The only data reshaping that needs to be done is accounting for any changes in data reporting practices over time. For example, some years had data for North America, and some had that data split into Canada and United Sates, and had to be aggregated after being entered. We are currently also putting together a data set on Data Centers (with information including location, power usage, size, etc.), but getting that data has been slow. Once we have that completed data we have plans to correlate it with our existing CDP data and create our model.

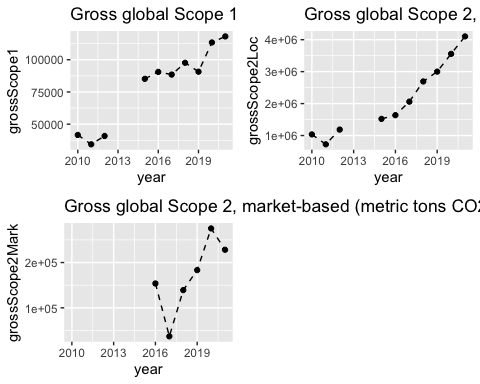
# Key Metrics and Goals

We are trying to track carbon consumption over time. So metrics comparing carbon levels at various points in time, and breaking down how that carbon is being consumed is critical for our success. Other important metrics that we will be considering is how that carbon is being offset. How much and what type of offsets are being purchased can be correlated with our predictions for future carbon emissions and help us to predict the future cost of offsets.

# Visualizations Over Time

## Global Emissions

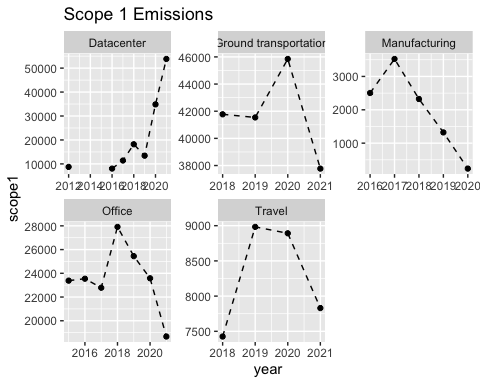
library(ggplot2)  
library(dplyr)  
library(tidyr)  
library(ggpubr)  
ccGlobal <- read.csv('Data/ClimateChangeGlobal.csv', header=FALSE)  
ccGlobal = setNames(data.frame(t(ccGlobal[,-1])), ccGlobal[,1])  
ccGlobal = ccGlobal[!is.na(ccGlobal[,1]),0:4]  
ccGlobal = ccGlobal %>% rename(year = `Year (C6.1 - 6.5)`,  
 grossScope1 = `Gross global Scope 1 emissions (metric tons CO2e)`,  
 grossScope2Loc = `\_Gross global Scope 2, location-based\_ (metric tons CO2e)`,  
 grossScope2Mark = `Gross global Scope 2, market-based (metric tons CO2e)`)   
  
globalScope1 <- ggplot(ccGlobal, aes(x=year, y=grossScope1))+  
 geom\_point()+geom\_line(linetype='dashed')+  
 ggtitle("Gross global Scope 1 emissions (metric tons CO2e)")  
  
globalScope2Loc <- ggplot(ccGlobal, aes(x=year, y=grossScope2Loc))+  
 geom\_point()+geom\_line(linetype='dashed')+  
 ggtitle("Gross global Scope 2, location-based (metric tons CO2e)")  
  
globalScope2Mark <- ggplot(ccGlobal, aes(x=year, y=grossScope2Mark))+  
 geom\_point()+geom\_line(linetype='dashed')+  
 ggtitle("Gross global Scope 2, market-based (metric tons CO2e)")  
  
  
ggarrange(globalScope1, globalScope2Loc, globalScope2Mark, nrow=2, ncol=2)



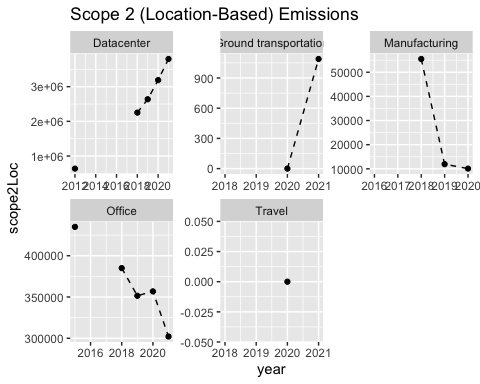
These graphs show critical data for our project: total emissions over time, which is what we are trying to predict with our model. What we can see is that location-based emissions are going up consistently, and Microsoft is opting to utilize buybacks to ‘reduce’ their emissions.

## By Business Activity

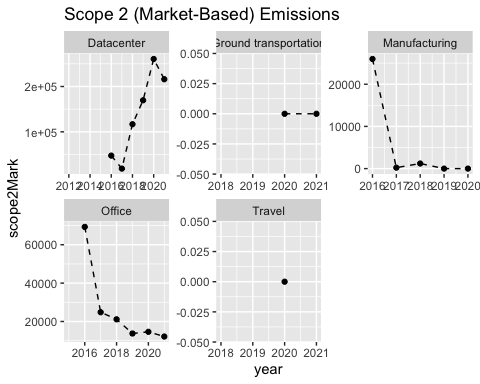
ccBusAct <- read.csv('Data/ClimateChangeBusinessActivity.csv')  
ccBusAct = ccBusAct %>% rename(year = Year..C7.3.,  
 scope1 = Scope.1.emissions..metric.tons.CO2e.,  
 scope2Loc = Scope.2..location.based..metric.tons.CO2e.,  
 scope2Mark = Scope.2..market.based..metric.tons.CO2e.) %>%  
 select(year,Activity,scope1,scope2Loc,scope2Mark) %>%   
 filter(!is.na(year),  
 !is.na(Activity))  
  
BAScope1Plot <- ggplot(ccBusAct, aes(x=year,y=scope1))+  
 geom\_point()+ facet\_wrap(~Activity, scales='free')+  
 ggtitle("Scope 1 Emissions")+  
 geom\_line(linetype='dashed')  
  
BAScope2LocPlot <- ggplot(ccBusAct, aes(x=year,y=scope2Loc))+  
 geom\_point()+ facet\_wrap(~Activity, scales='free')+  
 ggtitle("Scope 2 (Location-Based) Emissions")+  
 geom\_line(linetype='dashed')  
  
BAScope2MarkPlot <- ggplot(ccBusAct, aes(x=year,y=scope2Mark))+  
 geom\_point()+ facet\_wrap(~Activity, scales='free')+  
 ggtitle("Scope 2 (Market-Based) Emissions")+  
 geom\_line(linetype='dashed')  
  
BAScope1Plot



BAScope2LocPlot



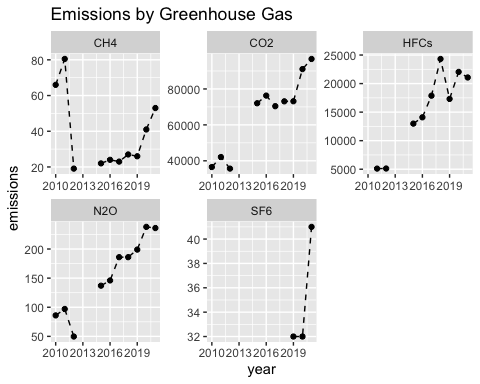
BAScope2MarkPlot



This data is a large part of the reason that we decided to focus on data centers. As can be seen in the plots above, emissions are generally going down for almost every business activity over time, but data center emissions are growing exponentially. Tracking data center emissions could be a decent representation of all emissions.

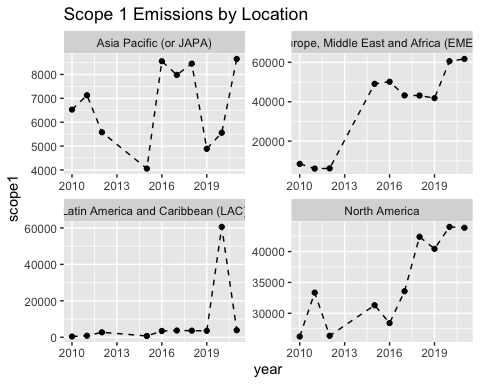
## By Greenhouse Gas

ccGas <- read.csv('Data/CCGas.csv')  
ccGas = ccGas %>% rename(year = Year..C7.1a.,  
 emissions = Scope.1.Emissions..metric.tons.CO2e.)  
  
GasPlot <- ggplot(ccGas, aes(x=year, y=emissions))+  
 geom\_point()+ facet\_wrap(~Gas, scales='free')+  
 ggtitle("Emissions by Greenhouse Gas")+  
 geom\_line(linetype='dashed')  
GasPlot

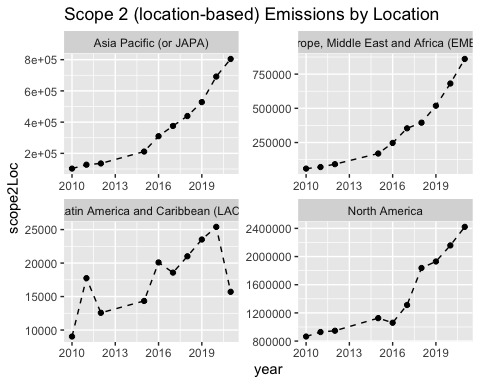


## By Area

ccArea <- read.csv('Data/CCArea.csv')  
ccArea = ccArea %>% rename(year = Year..C7.2.7.5.,  
 scope1 = Scope.1.Emissions..metric.tons.CO2e.,  
 scope2Loc = Scope.2..location.based..metric.tons.CO2e.,  
 scope2Mark = Scope.2..market.based..metric.tons.CO2e.,  
 ElHeStCool = Purchased.and.consumed.electricity..heat..steam.or.cooling..MWh.,  
 lowCarbElHeStCool = Purchased.and.consumed.low.carbon.electricity..heat..steam.or.cooling.accounted.for.in.Scope.2.market.based.approach..MWh.)%>%   
 filter(!is.na(Area))  
  
ccArea$Area[ccArea$Area=="Latin America (LATAM)"]="Latin America and Caribbean (LAC)"  
ccArea = ccArea[!(ccArea$Area=="Canada"|ccArea$Area=="United States of America"),]  
AreaScope1Plot <- ggplot(ccArea, aes(x=year, y=scope1))+  
 geom\_point()+ facet\_wrap(~Area, scales='free')+  
 ggtitle("Scope 1 Emissions by Location")+  
 geom\_line(linetype='dashed')  
AreaScope1Plot

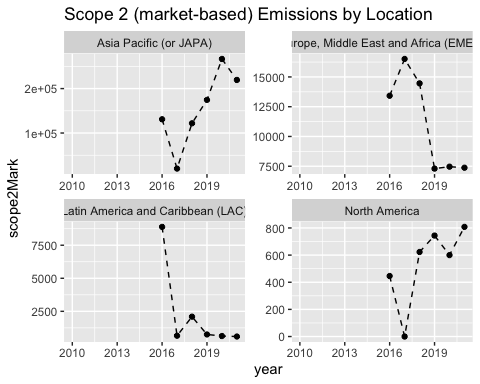


AreaScope2LocPlot <- ggplot(ccArea, aes(x=year, y=scope2Loc))+  
 geom\_point()+ facet\_wrap(~Area, scales='free')+  
 ggtitle("Scope 2 (location-based) Emissions by Location")+  
 geom\_line(linetype='dashed')  
AreaScope2LocPlot



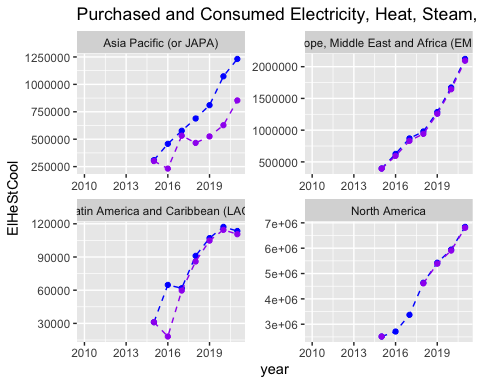
This graph is showing location-based carbon emissions within scope 2 . We see a steady increase across the board in all major regions. One important thing to note in these plots is that the scale varies for each one. While they are decent representations of the change in emissions within their area, they shouldn’t be used to compare between regions. For example, while the fluctuations in Latin America and Caribbean (LAC) appear dramatic comparatively, they are actually fairly small.

AreaScope2MarkPlot <- ggplot(ccArea, aes(x=year, y=scope2Mark))+  
 geom\_point()+ facet\_wrap(~Area, scales='free')+  
 ggtitle("Scope 2 (market-based) Emissions by Location")+  
 geom\_line(linetype='dashed')  
AreaScope2MarkPlot



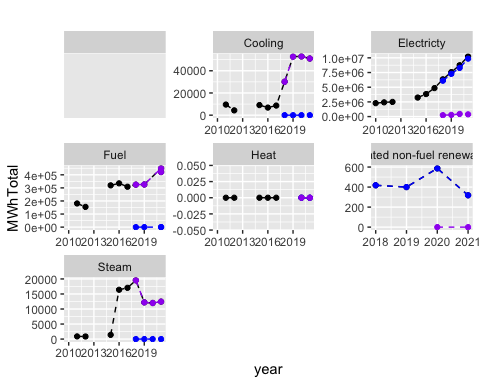
These graphs don’t paint a complete picture of total emissions emitted by region because these are graphs of market-based emissions. Market-based emissions take into account buybacks, which largely started in 2016. What it does clearly show us is that most buybacks are happening in Europe.

ggplot(ccArea, aes(x=year))+  
 geom\_point(aes(y=ElHeStCool), color="blue")+  
 geom\_line(aes(y=ElHeStCool), color="blue",linetype="dashed")+  
 geom\_point(aes(y=lowCarbElHeStCool), color="purple")+  
 geom\_line(aes(y=lowCarbElHeStCool), color="purple",linetype="dashed")+  
 facet\_wrap(~Area, scales='free')+  
 ggtitle("Purchased and Consumed Electricity, Heat, Steam, or Cooling by Location")



## By Consumption Type

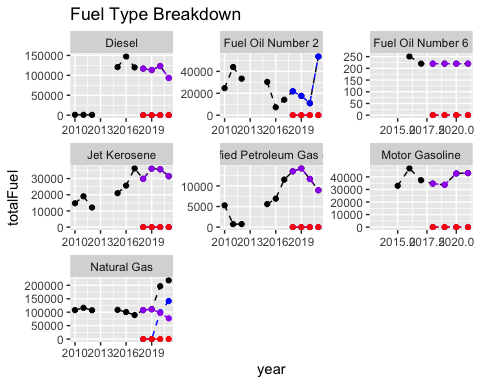
ccConType <- read.csv('Data/EnergyConsumptionOverview.csv')  
ccConType = ccConType %>% rename(year = Year..C8.2.,  
 conType = Consumption.Type..purchased.or.aquired.,  
 MWhRenew = MWh.from.renewable.sources,  
 MWhNonRenew = MWh.from.non.renewable.sources,  
 MWhTotal = Total..renewable.and.non.renewable..MWh) %>%  
 select(year,conType,MWhRenew,MWhNonRenew,MWhTotal) %>%  
 filter(!is.na(conType))  
  
ggplot(ccConType, aes(x=year))+  
 geom\_point(aes(y=MWhTotal), color="black")+  
 geom\_line(aes(y=MWhTotal), color="black",linetype="dashed")+  
 geom\_point(aes(y=MWhRenew), color="blue")+  
 geom\_line(aes(y=MWhRenew), color="blue",linetype="dashed")+  
 geom\_point(aes(y=MWhNonRenew), color="purple")+  
 geom\_line(aes(y=MWhNonRenew), color="purple",linetype="dashed")+  
 facet\_wrap(~conType, scales='free')+  
 ggtitle("")



In these plots we are looking at the breakdown of uses of energy consuming carbon. Black is the total, Blue is renewable, and purple is non-renewable. We see a start in the color change in 2018 because they changed how they were reporting MWh Totals that year. We also see that electricity and fuel are the largest sources of carbon consumption. We also see an effort by Microsoft to make all of their electricity renewable around the same time they made a net zero pledge.

## By Fuel Breakdown

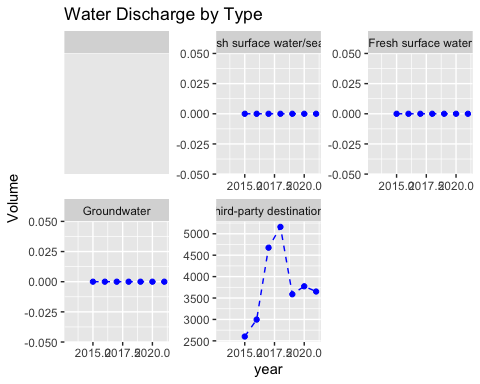
FuelBreakdownType <- read.csv('Data/FuelUseBreakdown.csv')  
FuelBreakdownType = FuelBreakdownType %>% rename(year = Year..C8.2c.,  
 fuelType = Fuel.Type,  
 totalFuel = Total.fuel.MWh.consumed.by.the.organization,  
 MWhElectric = MWh.fuel.consumed.for.self.generation.of.electricity,  
 MWhHeat = MWh.fuel.consumed.for.self.generation.of.heat,  
 EmissFactor = Emission.factor..metric.tons.CO2e.per.MWh.) %>%  
 select(year,fuelType,totalFuel,MWhElectric,MWhHeat,EmissFactor) %>%  
 filter(!is.na(fuelType))  
  
  
  
ggplot(FuelBreakdownType, aes(x=year))+  
 geom\_point(aes(y=totalFuel), color="black")+  
 geom\_line(aes(y=totalFuel), color="black",linetype="dashed")+  
 geom\_point(aes(y=MWhElectric), color="blue")+  
 geom\_line(aes(y=MWhElectric), color="blue",linetype="dashed")+  
 geom\_point(aes(y=MWhHeat), color="purple")+  
 geom\_line(aes(y=MWhHeat), color="purple",linetype="dashed")+  
 geom\_point(aes(y=EmissFactor), color="red")+  
 geom\_line(aes(y=EmissFactor), color="red",linetype="dashed")+  
 facet\_wrap(~fuelType, scales='free')+  
 ggtitle("Fuel Type Breakdown")



This graph helps to break down the uses of non-renewable fuel sources between heat and electricity. Heat is represented by blue and electricity is represented by purple with black being the total. Most of the fuel breakdown goes towards electricity, except for Fuel Oil Number 2 and Natural Gas. These plots also show an effort by Microsoft to reduce their usage of non-renewable sources over time in most categories.

## By Water Discharge

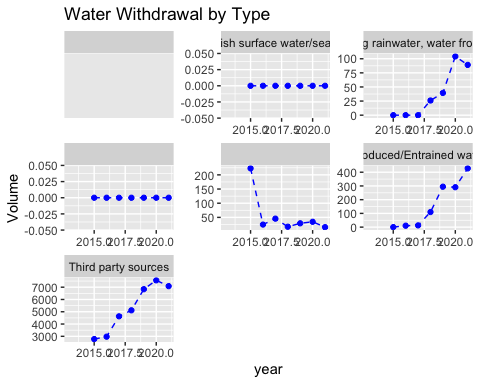
WaterDischarge <- read.csv('Data/WaterDischarge.csv')  
WaterDischarge = WaterDischarge %>% rename(year = Year..W1.2i.,  
 waterType = Type,  
 Volume = Volume..megaliters.year.,) %>%  
 select(year,waterType,Volume) %>%  
 filter(!is.na(waterType))  
  
   
  
ggplot(WaterDischarge, aes(x=year))+  
 geom\_point(aes(y=Volume), color="blue")+  
 geom\_line(aes(y=Volume), color="blue",linetype="dashed")+  
 facet\_wrap(~waterType, scales='free')+  
 ggtitle("Water Discharge by Type")



Water discharged by Microsoft can effect local habitats and have an indirect effect on climate change. From this data it appears that discharge has been mostly stable or near 0. I don’t see us using this as a factor for future carbon emissions.

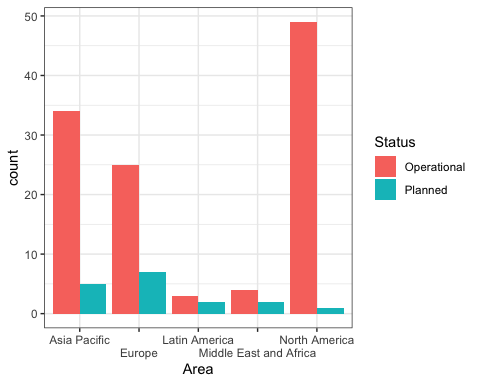
## By Water Withdrawal

WaterWithdrawal <- read.csv('Data/WaterWithdrawal.csv')  
WaterWithdrawal = WaterWithdrawal %>% rename(year = Year..W1.2h.,  
 waterType = Water.Source,  
 Volume = Volume..megaliters.year.,) %>%  
 select(year,waterType,Volume) %>%  
 filter(!is.na(waterType))  
  
   
  
ggplot(WaterWithdrawal, aes(x=year))+  
 geom\_point(aes(y=Volume), color="blue")+  
 geom\_line(aes(y=Volume), color="blue",linetype="dashed")+  
 facet\_wrap(~waterType, scales='free')+  
 ggtitle("Water Withdrawal by Type")



### By Data Center Location

library(ggplot2)  
library(dplyr)  
  
#reading in csv  
center <- read.csv("Data/datacenter\_data.csv")   
  
#editing name of attribute  
names(center)[4] = "YearBuilt"  
center\_temp1 = center  
centers0 = center\_temp1 %>% select(Name, Area, Status)  
  
#Plotting count of data centers  
dc\_plot = ggplot(centers0) + theme\_bw() + geom\_bar(aes(x = Area, fill = Status),position = 'dodge')+ scale\_x\_discrete(guide = guide\_axis(n.dodge = 2))  
dc\_plot



We wanted to work with a different data set than CDP to see what we could find. When we started to create some basic graphs, we noticed that there was a “Noth America” along with a “North America” entry. After cleaning that data, we were able to get the above visual.

# Conclusions So Far

From the data we have, it is clear that Microsoft’s electricity usage and carbon emissions are on the rise. With Microsoft planning and operating 50-100 new data centers every year, (whether though building, buying, or leasing) it’s no wonder that carbon emissions for data centers are rising the fastest. By utilizing predictive models concerning data centers, we can give a good estimate of how carbon consumption is likely to rise at Microsoft, and how that is going to translate to future buybacks.