Project Assignment 2

Akwesi Ntim Duodu

2024-03-18

```
##Research Question: **Are carbon emissions from industry associated wi
th carbon emissions from transport?**

##explanatory variable: Carbon emission from industry
##response variable: Carbon emission from transport
```

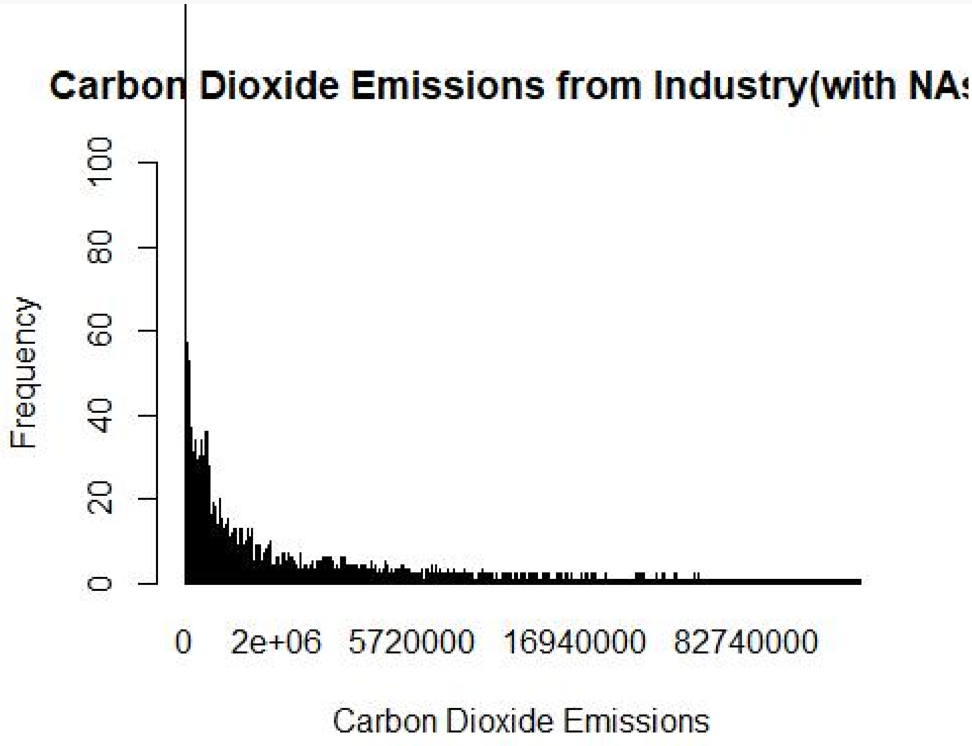
#load data set and libraries

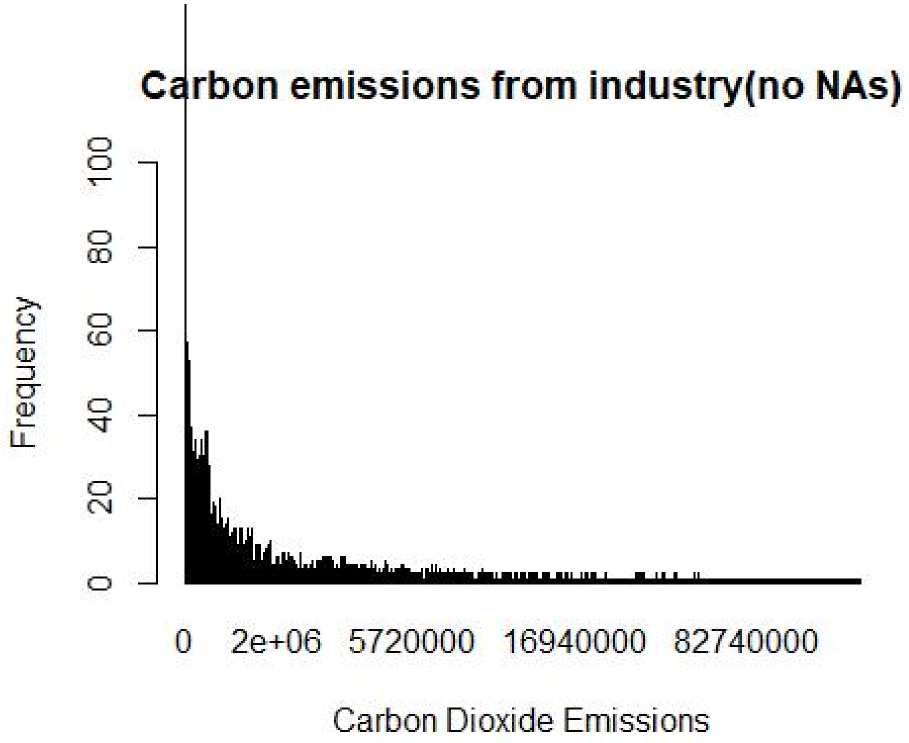
create variable subset

```
variables_to_investigate <- co_emission_by_sector[, c("Entity", "Year",
 "Carbon.dioxide.emissions.from.industry", "Carbon.dioxide.emissions.fr
om.transport")]
save(variables_to_investigate,file="akwesi_ntim_duodu_projectassignment
2")
```

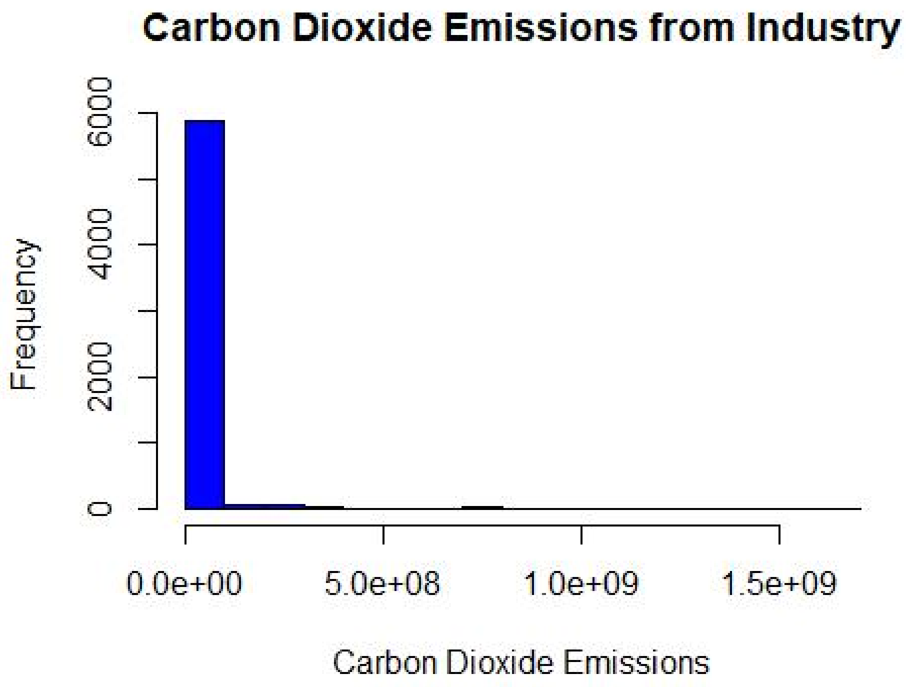
##Data management I

```
{
freq(variables_to_investigate$Carbon.dioxide.emissions.from.industry,
     main = "Carbon Dioxide Emissions from Industry(with NAs)",
     xlab = "Carbon Dioxide Emissions",
     ylab = "Frequency",
     ylim = c(0, 100))

#Identify error codes of NA and assign them AS NA's
# Replace NA values in the Carbon.dioxide.emissions.from.industry varia
ble with NA
variables_to_investigate$Carbon.dioxide.emissions.from.industry[is.na(v
ariables_to_investigate$Carbon.dioxide.emissions.from.industry)] <- NA

# Check the frequency of values in the Carbon.dioxide.emissions.from.in
dustry variable
freq(variables_to_investigate$Carbon.dioxide.emissions.from.industry,
     main= "Carbon emissions from industry(no NAs)",
      xlab = "Carbon Dioxide Emissions",
     ylab = "Frequency",
     ylim = c(0, 100))

#EXCLUDE NA VALUES FOR INDUSTRY
co_emission_by_sector$Carbon.dioxide.emissions.from.industry[is.na(co_e
mission_by_sector$Carbon.dioxide.emissions.from.industry)] <- NA
#EXCLUDE NA VALUES FOR TRANSPORT
co_emission_by_sector$Carbon.dioxide.emissions.from.transport[is.na(co_
emission_by_sector$Carbon.dioxide.emissions.from.transport)] <- NA
}
```

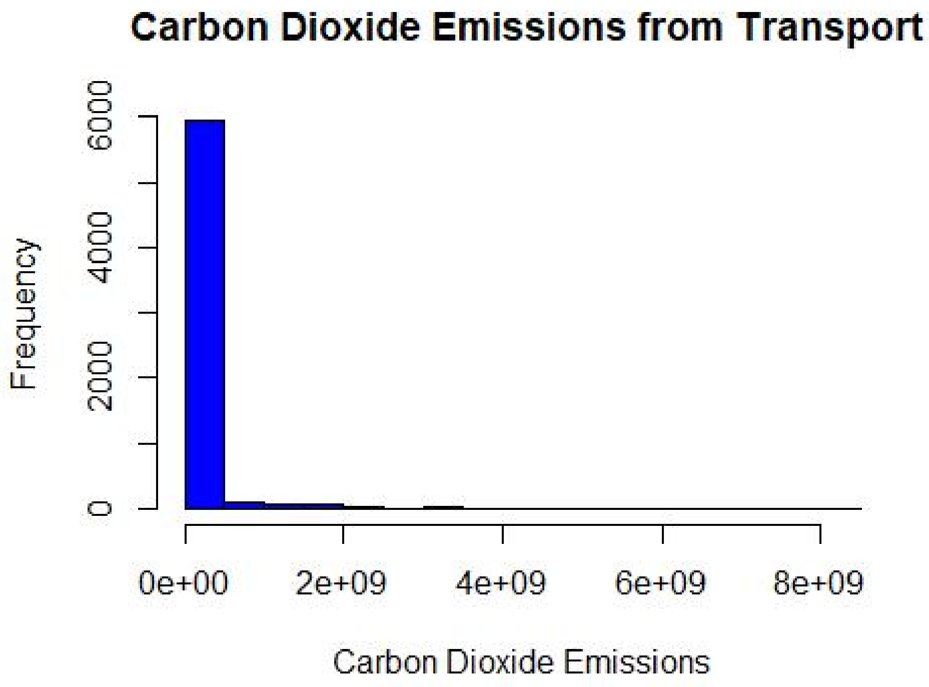




```r
# Create histogram for Carbon dioxide emissions from industry
hist(co_emission_by_sector$Carbon.dioxide.emissions.from.industry,
     main = "Carbon Dioxide Emissions from Industry",
     xlab = "Carbon Dioxide Emissions",
     ylab = "Frequency",
     col = "blue")
```



```r
#create histogram for carbon emissions from transport
hist(co_emission_by_sector$Carbon.dioxide.emissions.from.transport,
     main = " Carbon Dioxide Emissions from Transport",
     xlab = "Carbon Dioxide Emissions",
     ylab = "Frequency",
     col = "blue")
```



#Data Management II: Subsetting data(for possible further needs)

```
{
# Filter data for years above 2010
subset_data <- variables_to_investigate[variables_to_investigate$Year >
 2010, ]


# Group emissions into 5-year intervals from 1990
subset_data$Year_Group <- cut(subset_data$Year, breaks = seq(1990, 2025,
 by = 5), labels = paste(seq(1990, 2020, by = 5), "-", seq(1995, 2025,
by = 5)))

# Check the structure of the new data
str(subset_data)


secondary_subset <- variables_to_investigate[variables_to_investigate$Y
ear >= 2015, ]
str(secondary_subset)
}

## 'data.frame':    2050 obs. of  5 variables:
##  $ Entity                                    : chr  "Afghanistan" "Afgh
anistan" "Afghanistan" "Afghanistan" ...
##  $ Year                                      : int  2011 2012 2013 2014
```

```
 2015 2016 2017 2018 2019 2020 ...
##  $ Carbon.dioxide.emissions.from.industry : num  10000 30000 40000 3
0000 40000 80000 40000 60000 40000 60000 ...
##  $ Carbon.dioxide.emissions.from.transport: num  6710000 5850000 433
0000 3530000 4290000 3310000 3940000 4410000 4530000 3260000 ...
##  $ Year_Group                             : Factor w/ 7 levels "1990
 - 1995",..: 5 5 5 5 5 6 6 6 6 6 ...
## 'data.frame':    1230 obs. of  4 variables:
##  $ Entity                                 : chr  "Afghanistan" "Afgh
anistan" "Afghanistan" "Afghanistan" ...
##  $ Year                                   : int  2015 2016 2017 2018
 2019 2020 2015 2016 2017 2018 ...
##  $ Carbon.dioxide.emissions.from.industry : num  40000 80000 40000 6
0000 40000 ...
##  $ Carbon.dioxide.emissions.from.transport: num  4290000 3310000 394
0000 4410000 4530000 ...
```

##Descriptive Statistics for emissions from industry and transport

```
# Summary statistics for carbon emissions from industry
summary(secondary_subset$Carbon.dioxide.emissions.from.industry)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA'
s 
## 0.000e+00 3.000e+04 7.900e+05 3.078e+07 3.495e+06 1.633e+09       1
8
```

```
mean(secondary_subset$Carbon.dioxide.emissions.from.industry, na.rm = T
RUE)
```

```
## [1] 30778441
```

```
sd(secondary_subset$Carbon.dioxide.emissions.from.industry, na.rm = TRU
E)
```

```
## [1] 164723844
```

```
# Summary statistics for carbon emissions from transport
summary(secondary_subset$Carbon.dioxide.emissions.from.transport)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA'
s 
## 0.000e+00 1.055e+06 5.170e+06 1.406e+08 2.316e+07 8.269e+09       1
2
```

```
mean(secondary_subset$Carbon.dioxide.emissions.from.transport, na.rm =
TRUE)
```

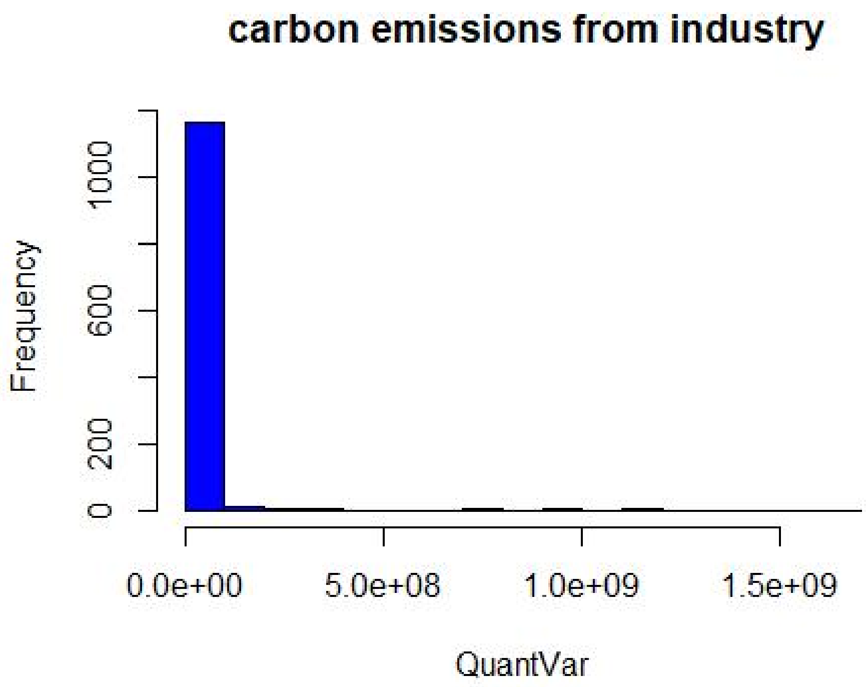
```
## [1] 140616871
```

```
sd(secondary_subset$Carbon.dioxide.emissions.from.transport, na.rm = TR
UE)
```

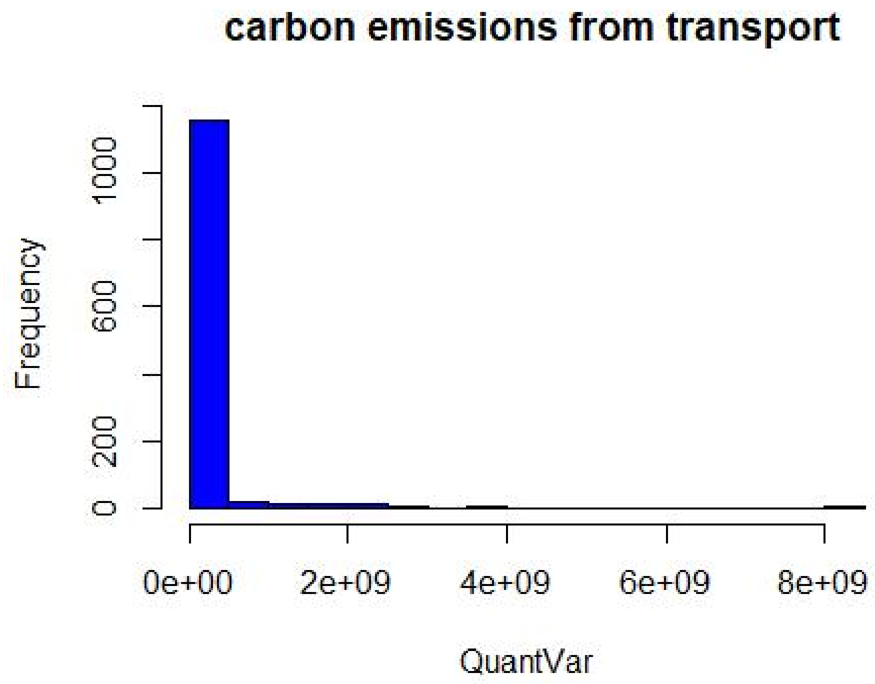
```
## [1] 676455968
```

#Univariate analysis of two quantitative variables

```
# Histogram for carbon emissions from industry
hist(secondary_subset$Carbon.dioxide.emissions.from.industry,
     main = " carbon emissions from industry",
     xlab = "QuantVar",
     ylab = "Frequency",
     col = "blue")
```



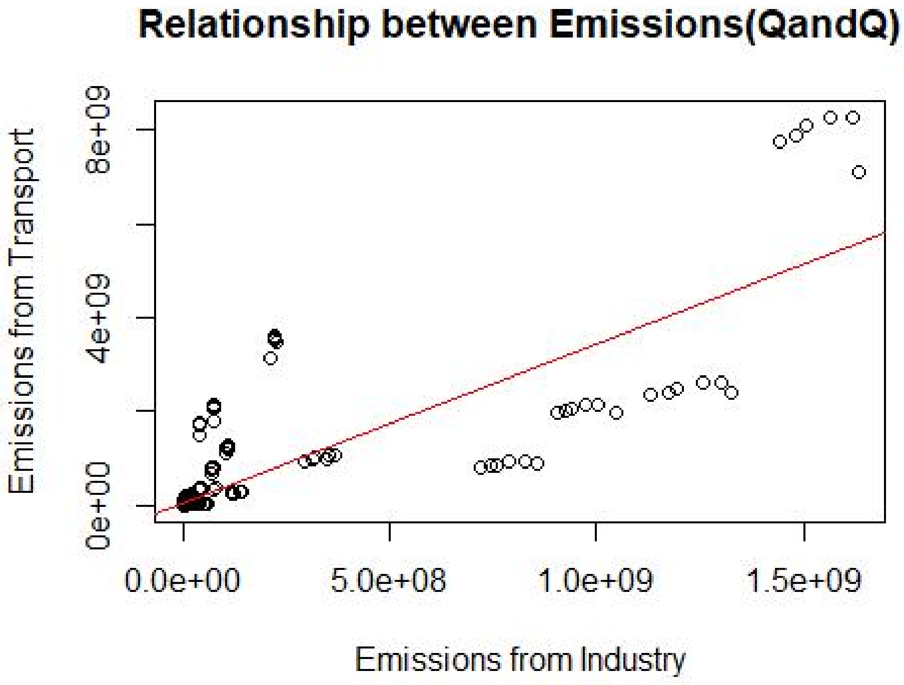
```
# Histogram for carbon emissions from transport
hist(secondary_subset$Carbon.dioxide.emissions.from.transport,
     main = " carbon emissions from transport",
     xlab = "QuantVar",
     ylab = "Frequency",
     col = "blue")
```



#Bivariate graphing

```
# Summary statistics
summary_industry <- summary(secondary_subset$Carbon.dioxide.emissions.f
rom.industry)
summary_transport <- summary(secondary_subset$Carbon.dioxide.emissions.
from.transport)

# Print summary statistics
print("Summary statistics for emissions from industry:")

## [1] "Summary statistics for emissions from industry:"

print(summary_industry)

##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA'
s
## 0.000e+00 3.000e+04 7.900e+05 3.078e+07 3.495e+06 1.633e+09        1
8

print("Summary statistics for emissions from transport:")

## [1] "Summary statistics for emissions from transport:"

print(summary_transport)

##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA'
s
```

```
## 0.000e+00 1.055e+06 5.170e+06 1.406e+08 2.316e+07 8.269e+09        1
2

#Create a formula for the scatterplot
formula <- formula(Carbon.dioxide.emissions.from.transport ~ Carbon.dio
xide.emissions.from.industry)

#Graph for emissions from industry vs. emissions from transport
plot(formula,
     data = secondary_subset,
     main = "Relationship between Emissions(QandQ)",
     xlab = "Emissions from Industry",
     ylab = "Emissions from Transport")

#line of best fit
abline(lm(formula, data = secondary_subset), col = "red")
```



```
##Summary

#Transport and industry have significant ecological effects. On the pre
mise of the supply chain, the two variables bear a relationship. Indust
ry relies on transport to move raw materials, components, and finished 
goods between suppliers, manufacturers, distributors, and customers. In
 the economy, transport plays a vital role in distributing goods from i
ndustrial facilities to retail stores or directly to consumers. Choosin
g to analyse the relationship between two quantitative variables has re
inforced learning of the fact that there could exist a variety of relat
```

```
ionships between variables, principal among them being associations. For example, the relationship can follow a linear pattern or a non-linear pattern. For some observations,it can be possible to observe a cluster of data points and periodic patterns. Choosing to use summaries of statistics such as the mean and the standard deviation can showcase leading trends for further investigation. I have learnt with this project that summaries of statistics are not "ends" in statistical investigation but tools used for further inferential analysis.

#Analysing two quantitative variables also proves that correlation will not necessarily imply causation. It can be possible for the explanatory variable to correlate with output from the response variable but this relationship can also owe to an effect from other variables. Moreover, the researcher needs to test further to arrive at fairly calculated confidence intervals. By this rationale, this project keeps in the secondary subset, other variables like the years associated with the emissions and the countries. Given this, it can be possible to determine whether larger or smaller emissions owe to the countries and or the years other than the two quantitative variables under test. On this front, building inferential analysis can be worked at if the hypothetical testing for the effect of the explanatory quantitative variable on the response quantitative variable proves null in the ensuing tests in the next project. From this project, I observed that there were no significant differences between the dataset with the presence of NAs and when NAs were absent. Excluding these values were not actions that significantly alter the data since the recording of an NA as per the rationale of those who undertook the study, was only done when there was no available data after field test.
```