Coffee

Michael Porco

2023-02-06

ASK

I will be performing an analysis for the fictional company "Michael's Coffee Company." I am tasked with analyzing exports of coffee. The company sources its coffee from all over the world, and wants to preemptively stay ahead of any supply chain shortages by analyzing coffee exports per country. The data of these countries over the past five years will be what I base my analysis off of. I obtained my data from Kaggle, where a JupyterLab Python Notebook took data from the International Coffee Organization's website and created csv files from it. I will be presenting my findings to the board of "Michael's Coffee Company."

PREPARE

The coffee dataset is about the countries that produce, export,or import coffee.

The data was made available by the ico. https://www.ico.org/new_historical.asp

This data was made usable by a JupyterLab Python notebook.

To learn more about the python notebook, look here. https://github.com/MSI17819/Coffee_data_analysis/blob/codeimpro/Coffee_codeimpro.ipynb

```
library(lubridate)
library(tidyverse)
library(janitor)
```

From the data's source, I have downloaded the files onto my pc. Here, I download the data into dataframes.

While downloading, I do the first basic tasks of cleaning the data. I remove any NA pieces of the data, and I make sure that the data does not have any unsupported characters.

PROCESS

```
coffee_export <- drop_na(clean_names(read.csv("Coffee_export.csv")))
```

Next, I take a look at the column names of the data frames to see what the data stores.

```
colnames(coffee_export)
```

```
##  [1] "country"      "x1990"        "x1991"        "x1992"        "x1993"
##  [6] "x1994"        "x1995"        "x1996"        "x1997"        "x1998"
## [11] "x1999"        "x2000"        "x2001"        "x2002"        "x2003"
## [16] "x2004"        "x2005"        "x2006"        "x2007"        "x2008"
## [21] "x2009"        "x2010"        "x2011"        "x2012"        "x2013"
## [26] "x2014"        "x2015"        "x2016"        "x2017"        "x2018"
## [31] "x2019"        "total_export"
```

The stakeholders are only looking for data trends over the past five years. I will be filtering out every year except for 2015 through 2019. I will also have to make a new column for totals because they start with the total of all the years in the original dataset.

```
coffee_export <- select(coffee_export, country, x2015, x2016, x2017, x2018, x2019)
```

Next I will use the glimpse function to see the number of rows and columns while also seeing what data type they are, and the first few items in the columns.

It is important to note that the unit of measurement here is in kilograms of exported coffee.

```
glimpse(coffee_export)
```

```
## Rows: 55
## Columns: 6
## $ country <chr> "Angola", "Bolivia (Plurinational State of)", "Brazil", "Burun~
## $ x2015   <dbl> 660000, 1800000, -2147483648, 13800000, 23400000, 2580000, 762~
## $ x2016   <int> 660000, 1320000, 2056140000, 12240000, 16860000, 4800000, 7698~
## $ x2017   <int> 540000, 1560000, 1855500000, 10140000, 14700000, 1080000, 7791~
## $ x2018   <int> 540000, 1320000, 2138220000, 12120000, 17220000, 2340000, 7684~
## $ x2019   <dbl> 1380000, 1200000, -2147483648, 17580000, 15000000, 1140000, 82~
```

CLEAN

First, I need to match data types. I will be making all of numbers into the double numerical data type so that they can all match. Countries and coffee type will stay as character strings.

During processing of the original dataset in python, some negative values were created when they should be positive. To remedy this I will be using the absolute value of all numerical values.

```
coffee_export <- mutate(coffee_export, x2016 = as.double(x2016))
coffee_export <- mutate(coffee_export, x2017 = as.double(x2017))
coffee_export <- mutate(coffee_export, x2018 = as.double(x2018))
coffee_export <- mutate(coffee_export, x2015 = abs(x2015))
coffee_export <- mutate(coffee_export, x2016 = abs(x2016))
coffee_export <- mutate(coffee_export, x2017 = abs(x2017))
coffee_export <- mutate(coffee_export, x2018 = abs(x2018))
coffee_export <- mutate(coffee_export, x2019 = abs(x2019))
```

Now that I have the completed dataset, I will make sure it is still clean and that no errors were created when manipulating the data.

```
coffee_export <- clean_names(drop_na(coffee_export))
```

ANALYZE

With the data checked and cleaned, I will begin analyzing it. First I will create a new column for totals of just the five years I am looking at.

```
coffee_export <- mutate(coffee_export, total = rowSums(coffee_export[,2:6]))
coffee_export <- mutate(coffee_export, change = x2019 - x2015)
```

These countries had the most exports over the past five years.

```
head(arrange(coffee_export, desc(total)))
```

```
##     country      x2015      x2016      x2017      x2018      x2019       total
## 1    Brazil 2147483648 2056140000 1855500000 2138220000 2147483648 10344827296
## 2  Viet Nam 1316640000 1783260000 1505520000 1883100000 1644000000  8132520000
## 3  Colombia  762960000  769860000  779100000  768480000  820320000  3900720000
## 4 Indonesia  502740000  392700000  491880000  272340000  380040000  2039700000
## 5  Honduras  301800000  318360000  440460000  428640000  405900000  1895160000
## 6     India  315720000  365160000  392520000  358020000  361680000  1793100000
##       change
## 1          0
## 2  327360000
## 3   57360000
## 4 -122700000
## 5  104100000
## 6   45960000
```

These countries exported the least coffee over the last five years.

```
head(arrange(coffee_export, total))
```

```
##             country  x2015  x2016  x2017  x2018 x2019  total change
## 1             Congo      0      0      0      0     0      0      0
## 2 Equatorial Guinea      0      0      0      0     0      0      0
## 3             Gabon      0      0      0      0     0      0      0
## 4          Paraguay      0      0      0      0     0      0      0
## 5         Sri Lanka      0  60000 120000 120000 60000 360000  60000
## 6            Guyana 120000 120000  60000  60000 60000 420000 -60000
```

I will compare the exports of 2015 and 2019 to see which country has grown the most in exports.

```
select(filter(coffee_export, change == max(change)), country, change)
```

```
##    country    change
## 1 Viet Nam 327360000
```

Here is the country the lowered its exports most.

```
select(filter(coffee_export, change == min(change)), country, change)
```

```
##     country     change
## 1 Indonesia -122700000
```

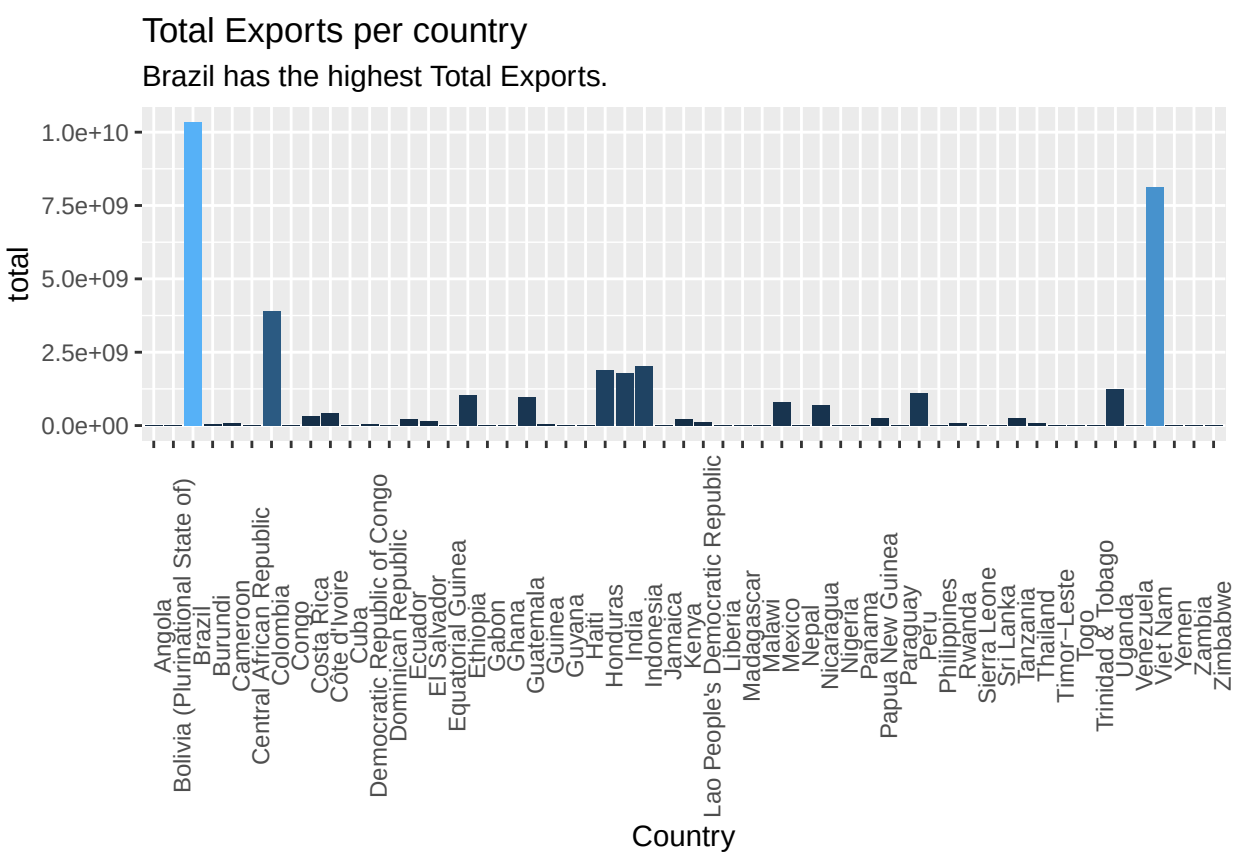
Here, I will create new dataframes for the top three exporters, and the country that shrank the most, to be visualized later.

```
vietnam <- filter(coffee_export, country == "Viet Nam")
vietnam <- pivot_longer(vietnam, cols = c("x2015", "x2016", "x2017", "x2018", "x2019"), names_to = "yea
brazil <- filter(coffee_export, country == "Brazil")
brazil <- pivot_longer(brazil, cols = c("x2015", "x2016", "x2017", "x2018", "x2019"), names_to = "year"
colombia <- filter(coffee_export, country == "Colombia")
colombia <- pivot_longer(colombia, cols = c("x2015", "x2016", "x2017", "x2018", "x2019"), names_to = "y
indonesia <- filter(coffee_export, country == "Indonesia")
indonesia <- pivot_longer(indonesia, cols = c("x2015", "x2016", "x2017", "x2018", "x2019"), names_to =
```

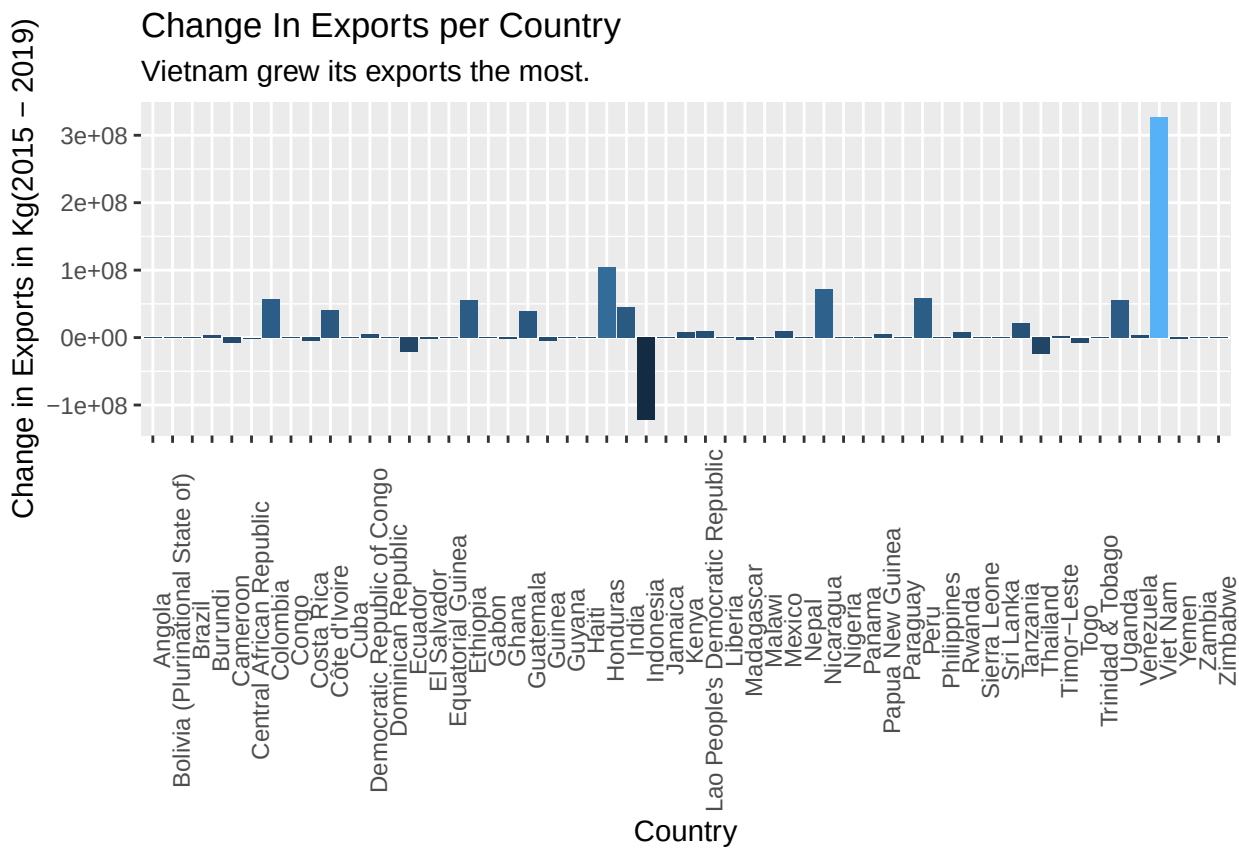
SHARE

Now I will create some vizualizations to show what the coffee dataset can tell us. I have graphs to show the total exports of all the countries, changes in exports, and the change in exports of some notable countries.

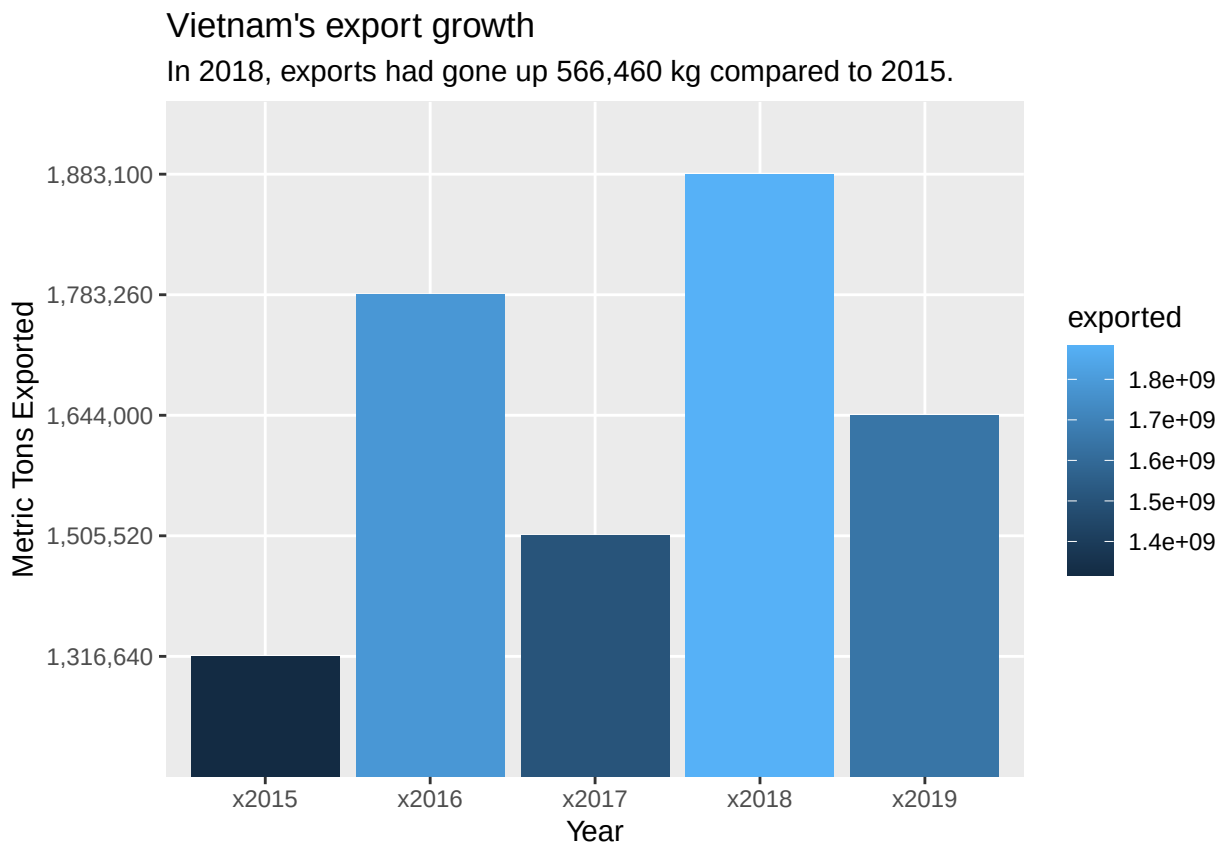
```
ggplot(data = coffee_export) + geom_col(mapping = aes(x = country, y = total, fill = total)) + theme(ax
```



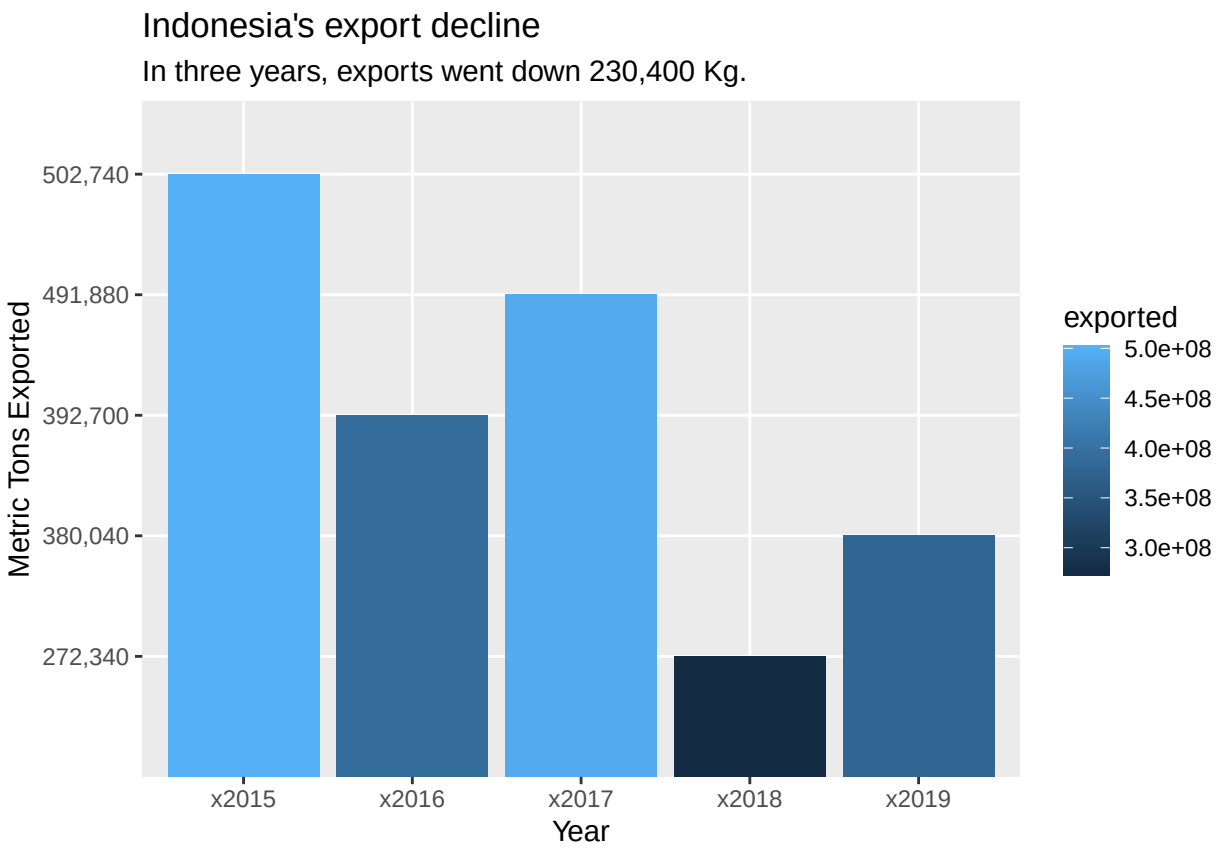
```
ggplot(data = coffee_export) + geom_col(mapping = aes(x = country, y = change, fill = change)) + theme(a
```



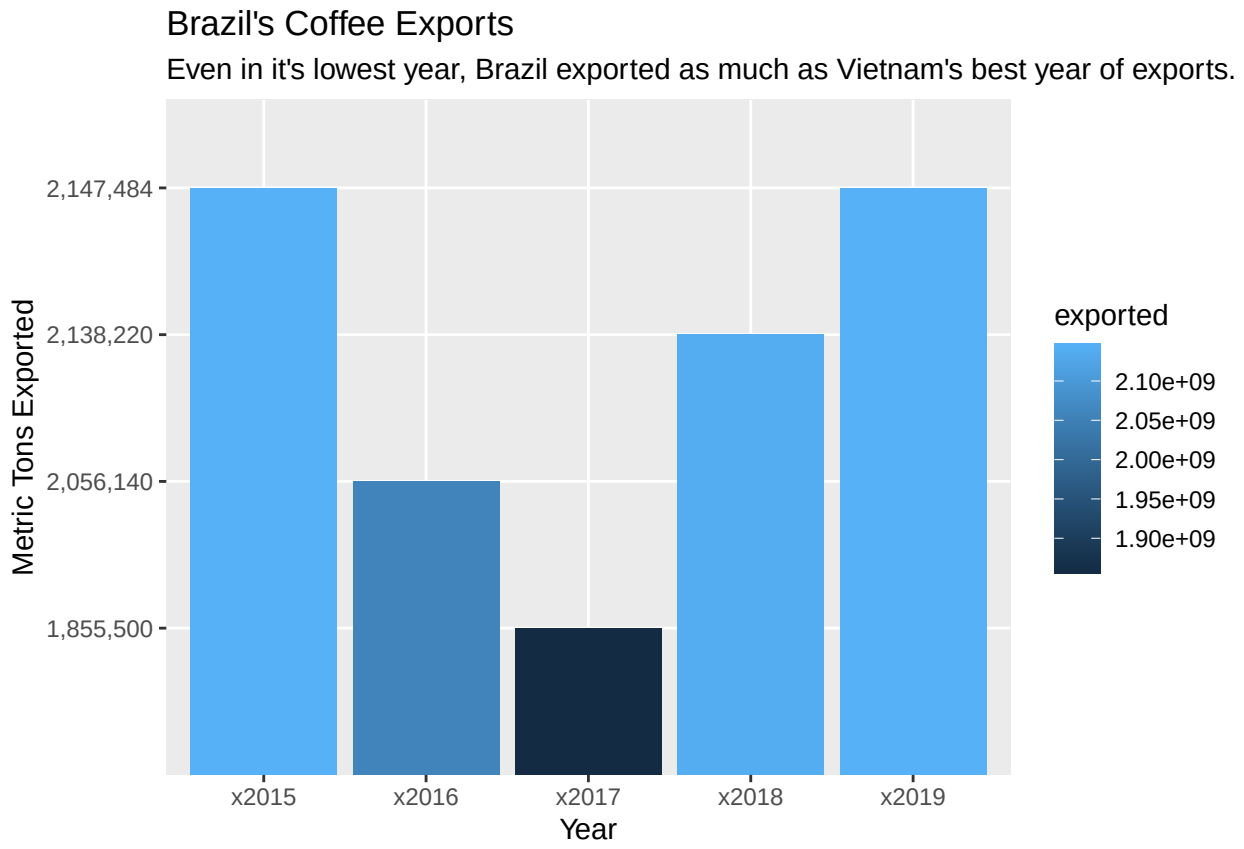

```
ggplot(data = vietnam) + geom_col(mapping = aes(x = year, y = format(exported/1000, big.mark = ",", sci
```



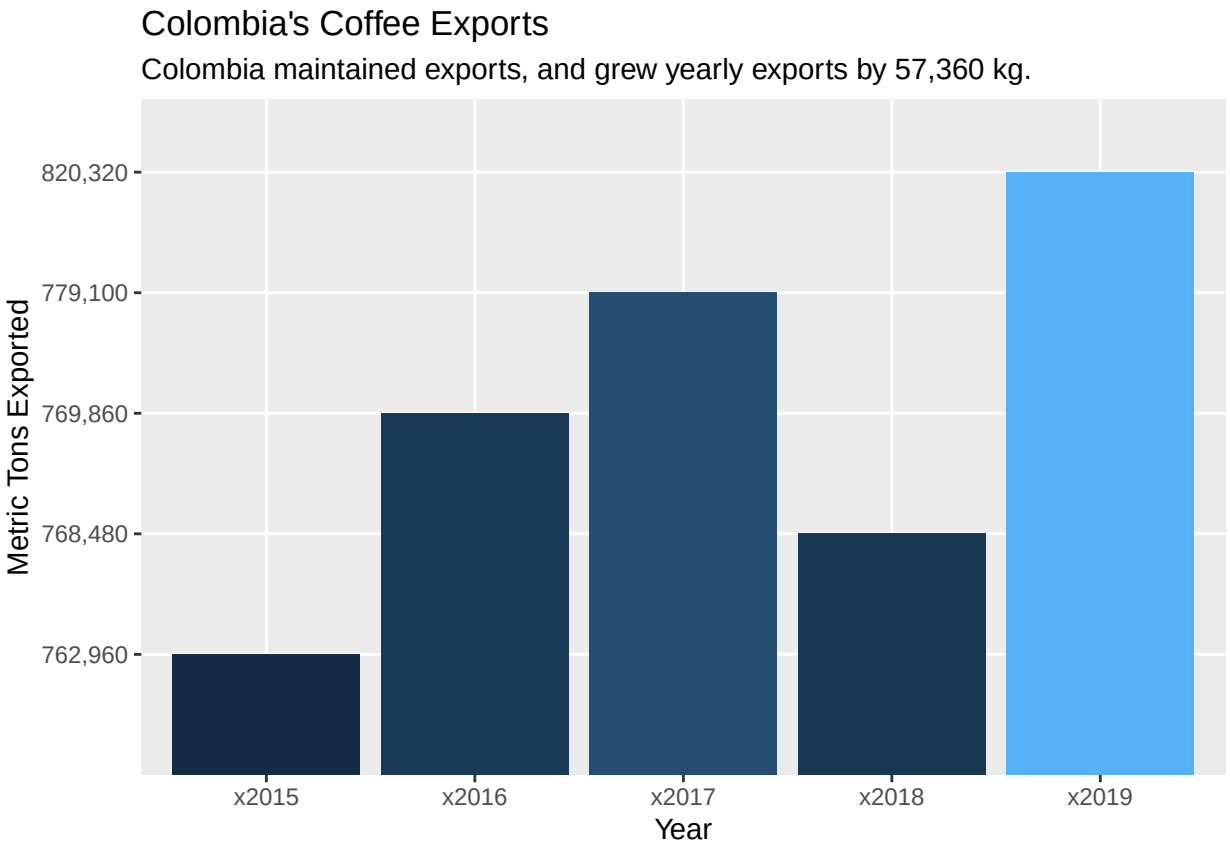

```
ggplot(data = indonesia) + geom_col(mapping = aes(x = year, y = format(exported/1000, big.mark = ",", s
```




```
ggplot(data = brazil) + geom_col(mapping = aes(x = year, y = format(exported/1000, big.mark = ",", scie
```




```
ggplot(data = colombia) + geom_col(mapping = aes(x = year, y = format(exported/1000, big.mark = ",", sc
```



ACT

Now, I present my findings to Michael's Coffee Company. Some key takeaways are:

Brazil maintains high levels of exports.

Indonesia is wavering in its yearly exportation of coffee.

Instead of Indonesia, Vietnam could be a good source, as it is growing consistently.

Colombia is also increasing its exports and could be a good source to replace countries with lowering exports.