```r
## -------- Assignment of analyzing coal consumption---by Kiran Kumar -chakk1k

# Read the CSV file manually
coal <- read.csv(file.choose())  # Manually reading the CSV file
head(coal)
```

```
##                           X    X1980    X1981    X1982    X1983   X1984
## 1             North America 16.45179 16.98772 16.47546 17.12407 18.4267
## 2                   Bermuda        0        0        0        0       0
## 3                    Canada  0.96156  0.99047  1.05584  1.11653 1.23682
## 4                 Greenland  0.00005  0.00005  0.00003  0.00003 0.00003
## 5                    Mexico  0.10239  0.10562  0.11967  0.12869 0.13071
## 6 Saint Pierre and Miquelon        0        0        0        0       0
##      X1985    X1986    X1987    X1988    X1989   X1990   X1991    X1992
## 1 18.81819 18.52559 19.43781 20.40363 20.62571 20.5602 20.4251 20.64672
## 2        0        0        0        0        0       0       0        0
## 3  1.20679  1.12583  1.25072  1.35809  1.35196 1.21338 1.26457  1.32379
## 4        0        0        0        0        0       0       0        0
## 5  0.14646  0.15609  0.17001  0.15967  0.17359  0.1694 0.15916  0.16584
## 6        0        0        0        0        0       0       0        0
##      X1993    X1994    X1995    X1996    X1997   X1998   X1999    X2000
## 1 21.28219 21.39631 21.64225 22.57572 23.20491 23.5002 23.4747 24.55583
## 2        0        0        0        0        0       0       0        0
## 3  1.22875  1.24492  1.28479  1.30032  1.44933 1.50985   1.505  1.61651
## 4        0        0        0        0        0       0       0        0
## 5  0.19118   0.1836  0.20768  0.25067  0.26373 0.26753 0.28947  0.29444
## 6        0        0        0        0        0       0       0        0
##      X2001    X2002    X2003    X2004   X2005    X2006    X2007    X2008
## 1 23.62705 23.69876 24.17788 24.36024 24.6876 24.32174 24.54746 24.11993
## 2        0        0        0        0       0        0        0        0
## 3  1.35444  1.36876  1.38766  1.43684 1.44948  1.42135  1.38369  1.37388
## 4        0        0        0        0       0        0        0        0
## 5  0.32908  0.36525  0.41878  0.31944 0.39739  0.39244  0.38911  0.32008
## 6        0        0        0        0       0        0        0        0
##      X2009
## 1 21.14803
## 2        0
## 3  1.14314
## 4        0
## 5   0.3365
## 6        0
```

```r
# Rename the first column to "Region"
colnames(coal)[1] <- "Region"
colnames(coal)
```

```
##  [1] "Region" "X1980"  "X1981"  "X1982"  "X1983"  "X1984"  "X1985"  "X1986"
##  [9] "X1987"  "X1988"  "X1989"  "X1990"  "X1991"  "X1992"  "X1993"  "X1994"
## [17] "X1995"  "X1996"  "X1997"  "X1998"  "X1999"  "X2000"  "X2001"  "X2002"
## [25] "X2003"  "X2004"  "X2005"  "X2006"  "X2007"  "X2008"  "X2009"
```

```r
# Display summary statistics of the dataset
summary(coal)
```

```
##     Region             X1980             X1981             X1982
##  Length:232         Length:232         Length:232         Length:232
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##     X1983             X1984             X1985             X1986
##  Length:232         Length:232         Length:232         Length:232
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##     X1987             X1988             X1989             X1990
##  Length:232         Length:232         Length:232         Length:232
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##     X1991             X1992             X1993             X1994
##  Length:232         Length:232         Length:232         Length:232
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##     X1995             X1996             X1997             X1998
##  Length:232         Length:232         Length:232         Length:232
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##     X1999             X2000             X2001             X2002
##  Length:232         Length:232         Length:232         Length:232
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##     X2003             X2004             X2005             X2006
##  Length:232         Length:232         Length:232         Length:232
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##     X2007             X2008             X2009
##  Length:232         Length:232         Length:232
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
```

```r
# Load necessary libraries
library(tidyverse)  # Load the library for pivot_longer and pivot_wider functions.
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```
## Warning: package 'dplyr' was built under R version 4.3.2
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(dplyr)

# Pivot the dataset to long format using pivot_longer
# Pivot the dataset to long format using pivot_longer
long_coal_dataset <- pivot_longer(coal, cols = -Region, names_to = "Year", values_to = "Consumption")
head(long_coal_dataset)
```

```
## # A tibble: 6 x 3
##   Region        Year  Consumption
##   <chr>         <chr> <chr>
## 1 North America X1980 16.45179
## 2 North America X1981 16.98772
## 3 North America X1982 16.47546
## 4 North America X1983 17.12407
## 5 North America X1984 18.4267
## 6 North America X1985 18.81819
```

```r
## When we check the long_coal_dataset in the year, we see that multiple variables are stored in one co
####--------####
## We observe that in the year column multiple variables are stored in one column

# Using transform function and gsub to replace "X" with an empty string
long_coal_dataset <- transform(long_coal_dataset, Year = gsub("X", "", Year))

head(long_coal_dataset)
```

```
##          Region Year Consumption
## 1 North America 1980    16.45179
## 2 North America 1981    16.98772
## 3 North America 1982    16.47546
## 4 North America 1983    17.12407
## 5 North America 1984     18.4267
## 6 North America 1985    18.81819
```

```r
## Let's change the year into numeric
long_coal_dataset$Year <- as.numeric(as.character(long_coal_dataset$Year))
is.numeric(long_coal_dataset$Year)  # Check if it changed to numeric or not
```

```
## [1] TRUE
```

```r
## Let's change the other column, Consumption, from character to numeric

## Let's check whether it is numeric or not
is.numeric(long_coal_dataset$Consumption)
```

```
## [1] FALSE
```

```r
## Let's check whether it's character
is.character(long_coal_dataset$Consumption)
```

```
## [1] TRUE
```

```
### It is character, so let's change it into numeric
long_coal_dataset$Consumption <- as.numeric(long_coal_dataset$Consumption)
```

```
## Warning: NAs introduced by coercion
```

```
is.numeric(long_coal_dataset$Consumption)
```

```
## [1] TRUE
```

```
summary(long_coal_dataset)
```

```
##     Region               Year       Consumption
##  Length:6960        Min.   :1980   Min.   : -0.0003
##  Class :character   1st Qu.:1987   1st Qu.:  0.0000
##  Mode  :character   Median :1994   Median :  0.0002
##                     Mean   :1994   Mean   :  1.3256
##                     3rd Qu.:2002   3rd Qu.:  0.0773
##                     Max.   :2009   Max.   :138.8298
##                                    NA's   :517
```

```
head(long_coal_dataset)
```

```
##          Region Year Consumption
## 1 North America 1980    16.45179
## 2 North America 1981    16.98772
## 3 North America 1982    16.47546
## 4 North America 1983    17.12407
## 5 North America 1984    18.42670
## 6 North America 1985    18.81819
```

```
#### Using sapply
## Let's check the class for all the columns together
sapply(long_coal_dataset, class)
```

```
##      Region        Year Consumption
## "character"   "numeric"   "numeric"
```

```
summary(long_coal_dataset)
```

```
##     Region               Year       Consumption
##  Length:6960        Min.   :1980   Min.   : -0.0003
##  Class :character   1st Qu.:1987   1st Qu.:  0.0000
##  Mode  :character   Median :1994   Median :  0.0002
##                     Mean   :1994   Mean   :  1.3256
##                     3rd Qu.:2002   3rd Qu.:  0.0773
##                     Max.   :2009   Max.   :138.8298
##                                    NA's   :517
```

```
### Let's remove the NA's from here

library(janitor)
```

```
## Warning: package 'janitor' was built under R version 4.3.2
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
head(long_coal_dataset)
```

```
##          Region Year Consumption
## 1 North America 1980    16.45179
## 2 North America 1981    16.98772
## 3 North America 1982    16.47546
## 4 North America 1983    17.12407
## 5 North America 1984    18.42670
## 6 North America 1985    18.81819
```

```
library(tidyr)

tidy_coaldata <- long_coal_dataset
tidy_coaldata <- tidy_coaldata %>%
  drop_na()

head(tidy_coaldata)
```

```
##          Region Year Consumption
## 1 North America 1980    16.45179
## 2 North America 1981    16.98772
## 3 North America 1982    16.47546
## 4 North America 1983    17.12407
## 5 North America 1984    18.42670
## 6 North America 1985    18.81819
```

```
sum(is.na(tidy_coaldata))
```

```
## [1] 0
```

```
summary(tidy_coaldata)
```

```
##     Region               Year       Consumption
##  Length:6443        Min.   :1980   Min.   : -0.00025
##  Class :character   1st Qu.:1987   1st Qu.:  0.00000
##  Mode  :character   Median :1995   Median :  0.00016
##                     Mean   :1995   Mean   :  1.32555
##                     3rd Qu.:2002   3rd Qu.:  0.07728
##                     Max.   :2009   Max.   :138.82977
```

```r
final_coal <- tidy_coaldata

### Separate continents, regions, and countries for grammar of graphic visualization

# Define a vector of continent names
continent_names <- c("Africa", "Asia", "Europe", "North America", "Central & South America",
                     "Former U.S.S.R.", "Middle East", "Central African Republic",
                     "Asia & Oceania", "Antarctica")

# Filter rows in 'final_coal' where the 'Region' column matches any continent name
continent <- filter(final_coal, Region %in% continent_names)
head(continent)
```

```
##          Region Year Consumption
## 1 North America 1980    16.45179
## 2 North America 1981    16.98772
## 3 North America 1982    16.47546
## 4 North America 1983    17.12407
## 5 North America 1984    18.42670
## 6 North America 1985    18.81819
```

```r
# Filter rows in 'final_coal' where the 'Region' column does not match any continent name
non_continents <- final_coal %>%
  filter(!(Region %in% continent_names))
head(non_continents)
```

```
##    Region Year Consumption
## 1 Bermuda 1980           0
## 2 Bermuda 1981           0
## 3 Bermuda 1982           0
## 4 Bermuda 1983           0
## 5 Bermuda 1984           0
## 6 Bermuda 1985           0
```

```r
# Filter out the 'World' region from 'non_continents'
countries <- non_continents[!(non_continents$Region %in% "World"),]

# View the resulting data frames
head(countries,n=15)  # Display data for individual countries
```

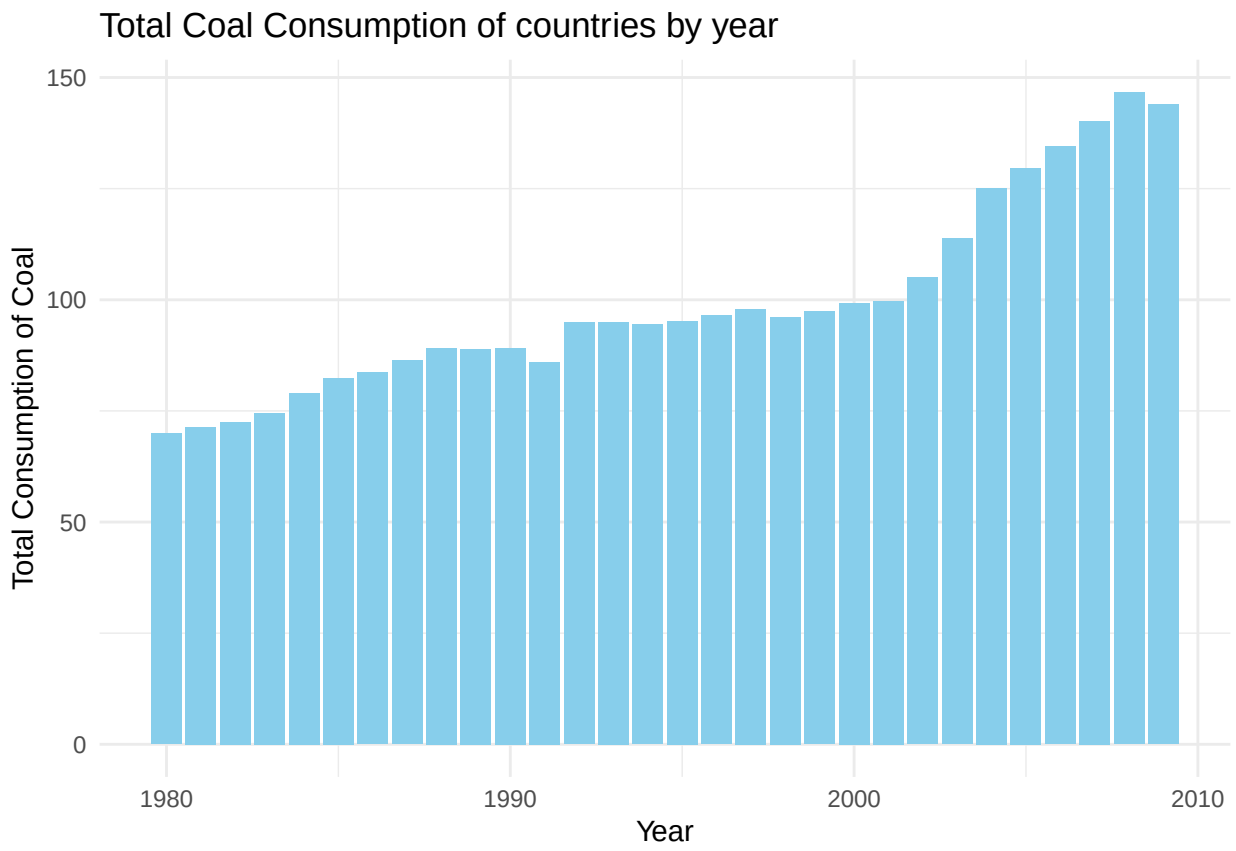
```
##     Region Year Consumption
## 1  Bermuda 1980           0
## 2  Bermuda 1981           0
## 3  Bermuda 1982           0
## 4  Bermuda 1983           0
## 5  Bermuda 1984           0
## 6  Bermuda 1985           0
## 7  Bermuda 1986           0
## 8  Bermuda 1987           0
## 9  Bermuda 1988           0
## 10 Bermuda 1989           0
## 11 Bermuda 1990           0
```

```
## 12 Bermuda 1991           0
## 13 Bermuda 1992           0
## 14 Bermuda 1993           0
## 15 Bermuda 1994           0
```

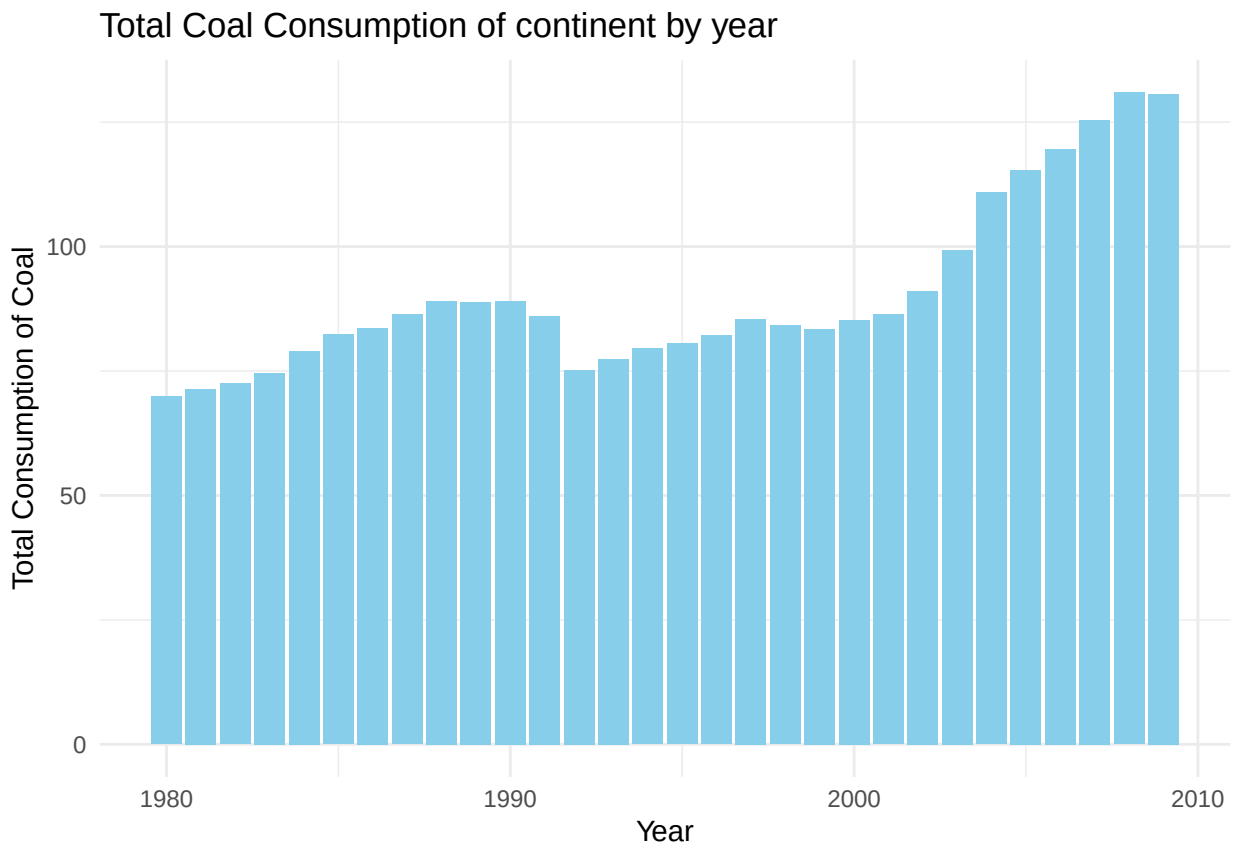
```r
head(continent,n=15)  # Display data for continents
```

```
##           Region Year Consumption
## 1  North America 1980    16.45179
## 2  North America 1981    16.98772
## 3  North America 1982    16.47546
## 4  North America 1983    17.12407
## 5  North America 1984    18.42670
## 6  North America 1985    18.81819
## 7  North America 1986    18.52559
## 8  North America 1987    19.43781
## 9  North America 1988    20.40363
## 10 North America 1989    20.62571
## 11 North America 1990    20.56020
## 12 North America 1991    20.42510
## 13 North America 1992    20.64672
## 14 North America 1993    21.28219
## 15 North America 1994    21.39631
```

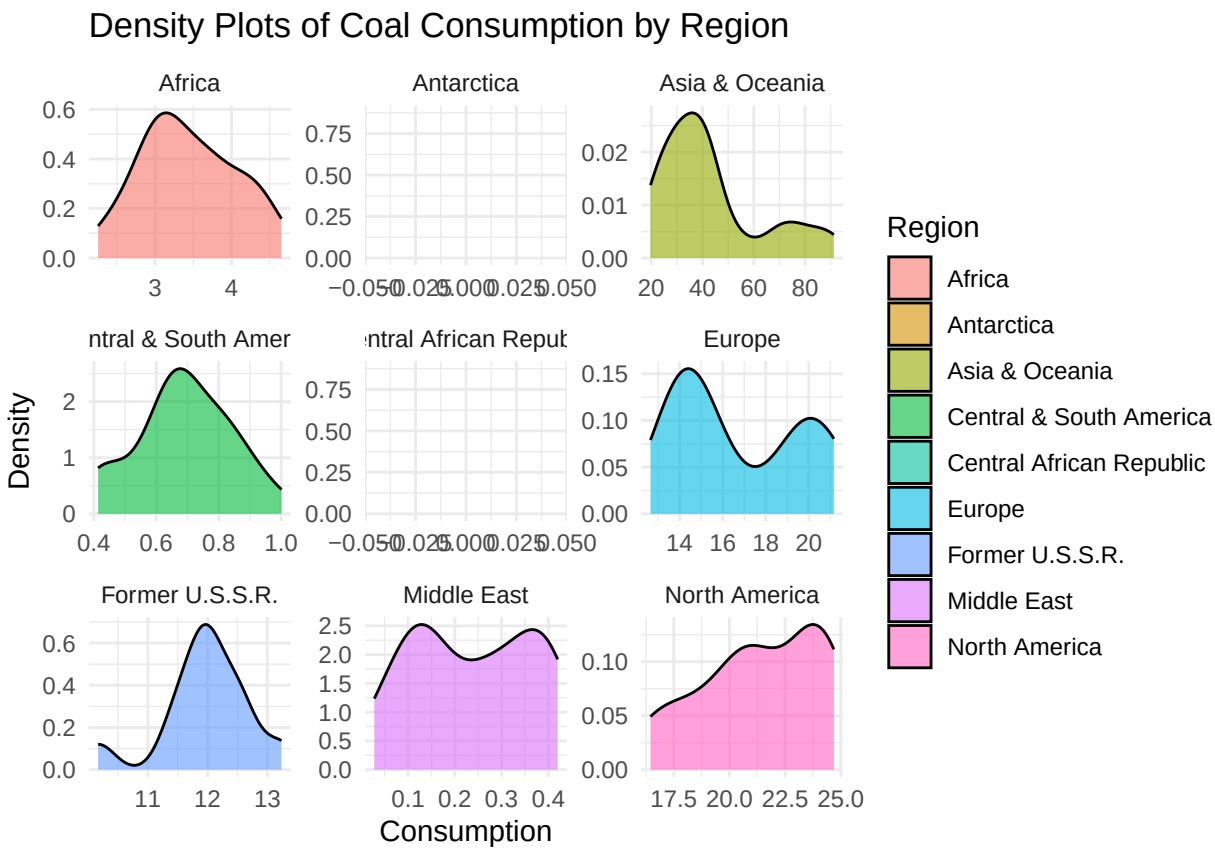
```r
##### Visualization using Grammer of graphix (ggplot2)
library(ggplot2)


### ggplot for countries

# Calculate total consumption by year of countries
total_consumption_countries <- aggregate(Consumption ~ Year, data = countries, FUN = sum)

# Bar plot for total consumption by year of countries
barplot_countries<-ggplot(total_consumption_countries, aes(x = Year, y = Consumption)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(
    title = "Total Coal Consumption of countries by year",
    x = "Year",
    y = "Total Consumption of Coal"
  ) +
  theme_minimal()
barplot_countries
```



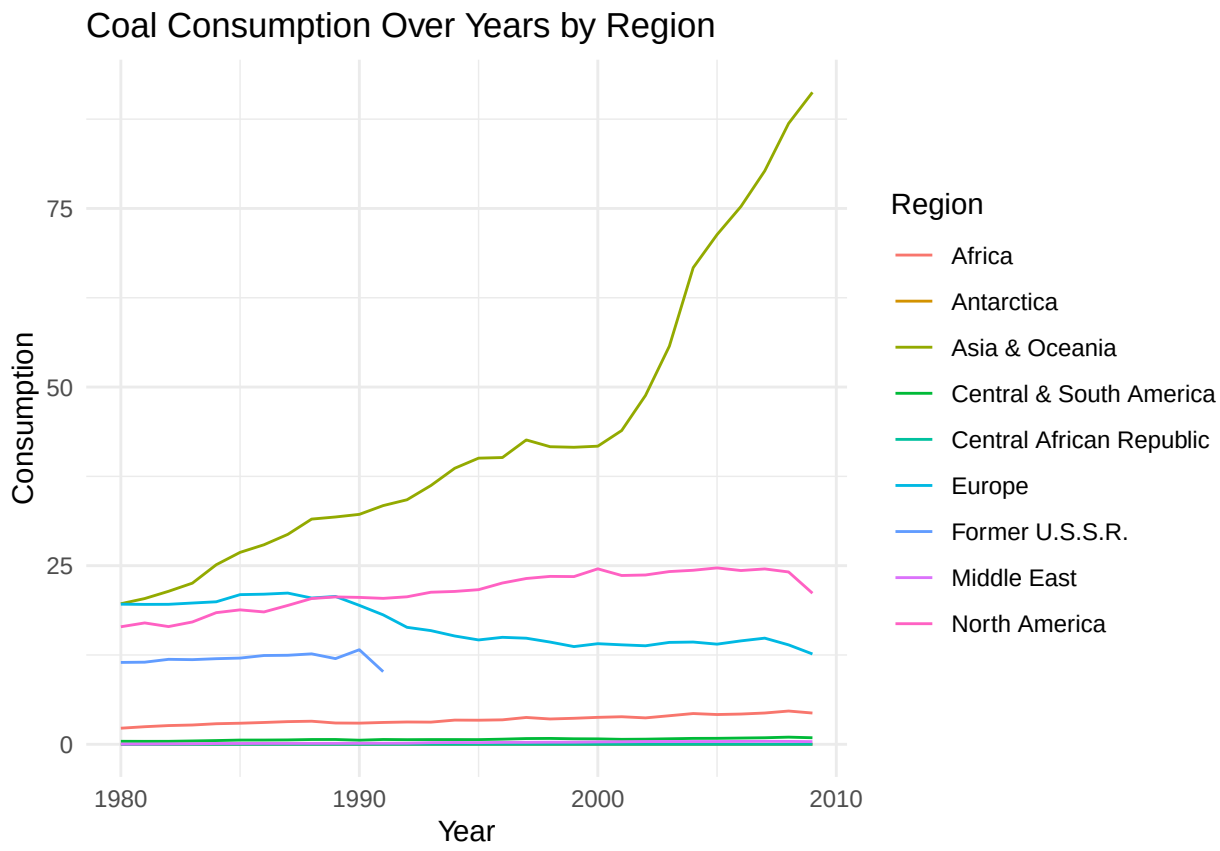

```
#Similarly lets calculate total consumption of coal by year of Continent

total_consumption_continent <- aggregate(Consumption ~ Year, data = continent, FUN = sum)

# Bar plot for total consumption by year of countries
barplot_continents<-ggplot(total_consumption_continent, aes(x = Year, y = Consumption)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(
    title = "Total Coal Consumption of continent by year",
    x = "Year",
    y = "Total Consumption of Coal"
  ) +
  theme_minimal()
barplot_continents
```



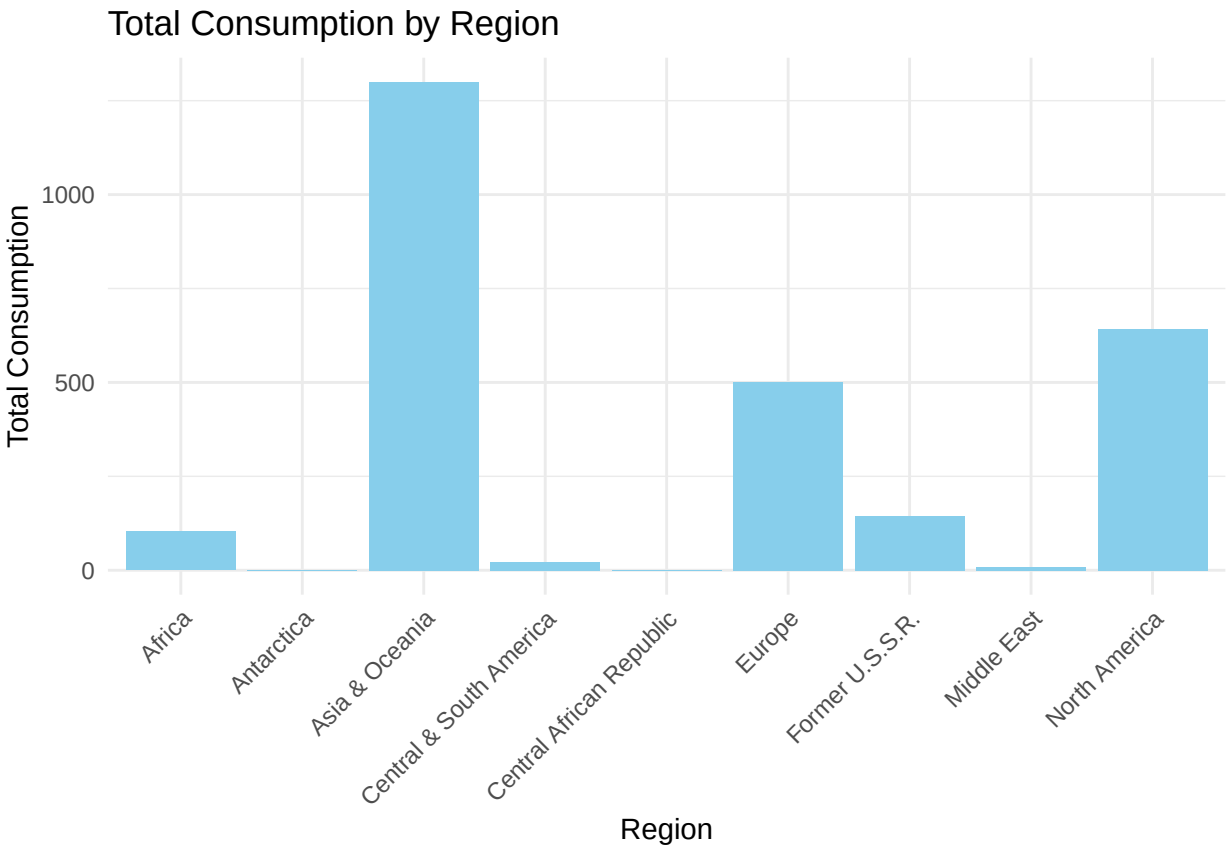

```
# Density plots of consumption by region
density_region<-ggplot(continent, aes(x = Consumption, fill = Region)) +
  geom_density(alpha = 0.6) +
  facet_wrap(~Region, scales = "free") +
  labs(
    title = "Density Plots of Coal Consumption by Region",
    x = "Consumption",
    y = "Density"
  ) +
  theme_minimal()

density_region
```




```r
# Aggregate data for consumption over regions and years

regionYearConsumption <- aggregate(Consumption ~ Region + Year, data =continent, FUN = sum)

time_overyears<-ggplot(regionYearConsumption, aes(x = as.numeric(Year), y = Consumption, color = Region)
  geom_line() +
  labs(
    title = "Coal Consumption Over Years by Region",
    x = "Year",
    y = "Consumption"
  ) +
  theme_minimal()
time_overyears
```




```r
# Create bar chart
bar_continent_region<-ggplot(continent, aes(x = Region, y = Consumption)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(
    title = "Total Consumption by Region",
    x = "Region",
    y = "Total Consumption"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

bar_continent_region
```


Total Consumption by Region
Total Consumption
1000
500
0
Africa
Antarctica
Asia & Oceania
Central & South America
Central African Republic
Europe
Former U.S.S.R.
Middle East
North America
Region