Data Wrangling on Coal Consumption

I have retrieved data for coal consumption from http://594442.youcanlearnit.net/coal.csv

The main purpose is to prepare the dataset for analysis.

Business Question

Identify trends in coal consumption from 1980 to 2010.

Analysis

I am going to use tidyverse library to perform data wrangling operations

```
#install and update tidyverse
install.packages("tidyverse")
library(tidyverse)
```

Now let's import data and have a look at the variables

```
coal <- read_csv('http://594442.youcanlearnit.net/coal.csv' , skip = 2)
glimpse(coal)
```

18:1 (Untitled) R Script

Console Terminal Jobs

~/Data Wrangling/

```
> glimpse(coal)
Observations: 232
Variables: 31
$ region <chr> "North America", "Bermuda", "Canada", "Greenland", "Mexico", "S...
$ `1980` <chr> "16.45179", "0", "0.96156", "0.00005", "0.10239", "0", "15.3877...
$ `1981` <chr> "16.98772", "0", "0.99047", "0.00005", "0.10562", "0", "15.8915...
$ `1982` <chr> "16.47546", "0", "1.05584", "0.00003", "0.11967", "0", "15.2999...
$ `1983` <chr> "17.12407", "0", "1.11653", "0.00003", "0.12869", "0", "15.8788...
$ `1984` <chr> "18.4267", "0", "1.23682", "0.00003", "0.13071", "0", "17.05914...
$ `1985` <chr> "18.81819", "0", "1.20679", "0", "0.14646", "0", "17.46494", "0...
$ `1986` <chr> "18.52559", "0", "1.12583", "0", "0.15609", "0", "17.24367", "0...
```

Summary of dataset:

```
Console   Terminal ×   Jobs ×
~/Data Wrangling/
> summary(coal)
    region              1980               1981               1982
 Length:232         Length:232         Length:232         Length:232
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
     1983               1984               1985               1986
 Length:232         Length:232         Length:232         Length:232
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
     1987               1988               1989               1990
 Length:232         Length:232         Length:232         Length:232
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
     1991               1992               1993               1994
 Length:232         Length:232         Length:232         Length:232
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
     1995               1996               1997               1998
 Length:232         Length:232         Length:232         Length:232
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
     1999               2000               2001               2002
 Length:232         Length:232         Length:232         Length:232
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
     2003               2004               2005               2006
 Length:232         Length:232         Length:232         Length:232
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
     2007               2008               2009
 Length:232         Length:232         Length:232
 Class :character   Class :character   Class :character
```

The dataset is wide. A good rule of thumb is to create wide dataset to long. I am going to use [gather] function.

```
> coal_long <- gather(coal, Year, Coal_consumption , -region )
> glimpse(coal_long)
Observations: 6,960
Variables: 3
$ region           <chr> "North America", "Bermuda", "Canada", "Greenland", "M...
$ Year             <chr> "1980", "1980", "1980", "1980", "1980", "1980", "1980...
$ Coal_consumption <chr> "16.45179", "0", "0.96156", "0.00005", "0.10239", "0"...
>
```

All the variables are of Character datatype, I will convert year as integer datatype and coal_consumption = numeric.

```
> coal_long$Year <- as.integer(coal_long$Year)
> summary(coal_long)
    region              Year      Coal_consumption
 Length:6960        Min.   :1980   Length:6960
 Class :character   1st Qu.:1987   Class :character
 Mode  :character   Median :1994   Mode  :character
                    Mean   :1994
                    3rd Qu.:2002
                    Max.   :2009
> coal_long$Coal_consumption <- as.numeric(coal_long$Coal_consumption)
```

Now, I can see that the regions variables has to be classified as it consists of many values which are countries, world and continents. For our analysis, we will simplify it to continent levels and call them regions. I can also see world as a row. We will keep this record for now and will not delete it.

```
> unique(coal_long$region)
  [1] "North America"                        "Bermuda"
  [3] "Canada"                               "Greenland"
  [5] "Mexico"                               "Saint Pierre and Miquelon"
  [7] "United States"                        "Central & South America"
  [9] "Antarctica"                           "Antigua and Barbuda"
 [11] "Argentina"                            "Aruba"
 [13] "Bahamas, The"                         "Barbados"
 [15] "Belize"                               "Bolivia"
 [17] "Brazil"                               "Cayman Islands"
 [19] "Chile"                                "Colombia"
 [21] "Costa Rica"                           "Cuba"
 [23] "Dominica"                             "Dominican Republic"
 [25] "Ecuador"                              "El Salvador"
 [27] "Falkland Islands (Islas Malvinas)"    "French Guiana"
 [29] "Grenada"                              "Guadeloupe"
 [31] "Guatemala"                            "Guyana"
 [33] "Haiti"                                "Honduras"
 [35] "Jamaica"                              "Martinique"
 [37] "Montserrat"                           "Netherlands Antilles"
 [39] "Nicaragua"                            "Panama"
 [41] "Paraguay"                             "Peru"
 [43] "Puerto Rico"                          "Saint Kitts and Nevis"
 [45] "Saint Lucia"                          "Saint Vincent/Grenadines"
 [47] "Suriname"                             "Trinidad and Tobago"
 [49] "Turks and Caicos Islands"             "Uruguay"
 [51] "Venezuela"                            "Virgin Islands,  U.S."
 [53] "Virgin Islands, British"              "Europe"
 [55] "Albania"                              "Austria"
 [57] "Belgium"                              "Bosnia and Herzegovina"
 [59] "Bulgaria"                             "Croatia"
 [61] "Cyprus"                               "Czech Republic"
```

```
> noncountries <- c("North America" ,"Central & South America" ,"Antarctica",
+                   "Europe" , "Eurasia","Middle East","Africa","Asia & Oceania",
+                   "World" )
```

Data preparation

```
> matches <- which(!is.na(match(coal_long$region, noncountries))) #list of rows with noncountry values
> summary(coal_long)
    region               Year      Coal_consumption
 Length:6960        Min.   :1980   Min.   : -0.0002
 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
> coal_country <- coal_long[-matches,]
> coal_region <- coal_long[matches,]
> unique(coal_country)
# A tibble: 6,690 x 3
   region                     Year Coal_consumption
   <chr>                     <int>            <dbl>
 1 Bermuda                    1980          0
 2 Canada                     1980          0.962
 3 Greenland                  1980          0.00005
 4 Mexico                     1980          0.102
 5 Saint Pierre and Miquelon  1980          0
 6 United States              1980         15.4
 7 Antigua and Barbuda        1980          0
 8 Argentina                  1980          0.0348
 9 Aruba                      1980         NA
10 Bahamas, The               1980          0
# ... with 6,680 more rows
```

Now we have prepared data for analysis.

```
> unique(coal_region)
# A tibble: 270 x 3
   region                    Year Coal_consumption
   <chr>                    <int>            <dbl>
 1 North America            1980          16.5
 2 Central & South America  1980           0.420
 3 Antarctica               1980           0
 4 Europe                   1980          19.6
 5 Eurasia                  1980          11.5
 6 Middle East              1980           0.0279
 7 Africa                   1980           2.25
 8 Asia & Oceania           1980          19.7
 9 World                    1980          69.9
10 North America            1981          17.0
# ... with 260 more rows
```

Let's use ggplot2 library to visualize

```
> library(ggplot2)
> ggplot(data = coal_region , mapping=aes(x = Year , y = Coal_consumption))+
+   geom_line(mapping = aes(color=region))
```

Insights

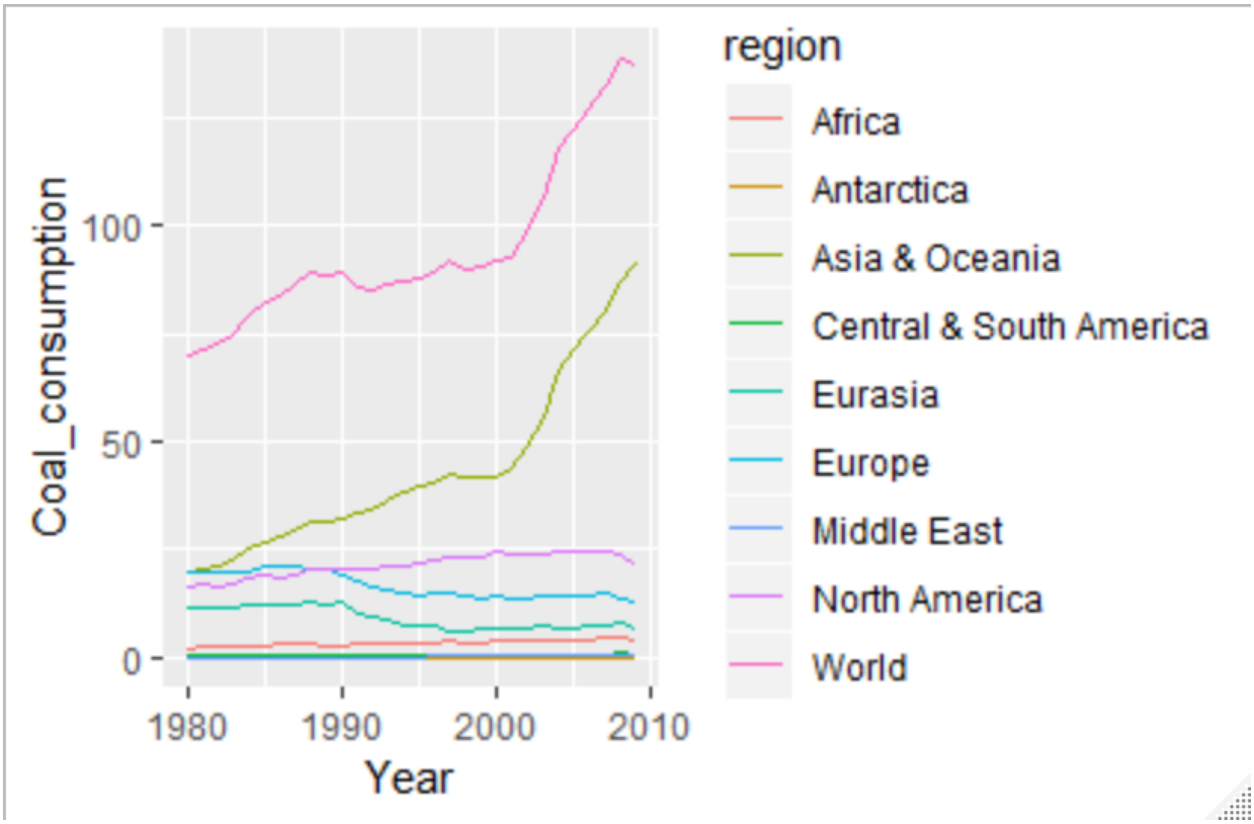

Figure 1: Plot for coal consumption trends across regions

From the line plot, we can see the trends in coal consumption in different regions of the world. From figure 1, we can see that the overall coal consumption has increased over the years.

Asia & Oceania has highest coal consumption and thus the trend line for coal consumption has peaked despite other regions having low/ consumption.