R Data Code

Kevin Babb
October 28, 2018

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document.

Loading of packages required for the data analysis

```
## ── Attaching packages
──────────────────────────────────────────────────────────
tidyverse 1.2.1 ──

## ✔ ggplot2 3.1.0     ✔ purrr   0.2.5
## ✔ tibble  1.4.2     ✔ dplyr   0.7.7
## ✔ tidyr   0.8.1     ✔ stringr 1.3.1
## ✔ readr   1.1.1     ✔ forcats 0.3.0

## ── Conflicts
─────────────────────────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

Loading of data into R

```
raw_stats <-
read.csv("~/Documents/Class/CKME-136/Workshop/CKME136_Capstone/Da
ta/all_energy_statistics.csv")
```

We now look at the data loaded

```
View(raw_stats)
```

Looking further:

```
summary(raw_stats)

##       country_or_area
##  Germany      :  20422
##  United States:  19847
##  Poland       :  19802
##  Austria      :  17440
##  Romania      :  17357
##  France       :  17236
##  (Other)      :1077378
##
commodity_transaction
##  From combustible fuels – Main activity                   :
6601
##  Electricity - Gross demand                               :
5532
##  Electricity - Gross production                           :
5523
##  Electricity - net production                             :
5523
##  Electricity - Own use by electricity, heat and CHP plants:
5523
##  Electricity - total production, main activity            :
5523
##
(Other)                                                  :1155257

##       year                          unit           quantity

##  Min.   :1990   Cubic metres, thousand : 52032   Min.   :   -
864348
##  1st Qu.:1997   Kilowatt-hours, million:147741   1st Qu.:
14
##  Median :2003   Kilowatts,  thousand   : 50229   Median :
189
##  Mean   :2003   Metric Tons            :   684   Mean   :
184265
##  3rd Qu.:2009   Metric tons,  thousand :759859   3rd Qu.:
2265
##  Max.   :2014   Terajoules             :178937
Max.   :6680329000
##

##  quantity_footnotes                      category
##  Min.   :1          total_electricity        :133916
##  1st Qu.:1          gas_oil_diesel_oil       : 97645
##  Median :1          fuel_oil                 : 75132
##  Mean   :1          natural_gas_including_lng: 64161
##  3rd Qu.:1          liquified_petroleum_gas  : 62156
##  Max.   :1          motor_gasoline           : 53198
##  NA's   :1025536    (Other)                  :703274
```

```
str(raw_stats)
```

```
## 'data.frame':    1189482 obs. of  7 variables:
##  $ country_or_area      : Factor w/ 243 levels
```

```
"Afghanistan",..: 14 14 21 21 21 21 21 21 58 58 ...
##  $ commodity_transaction: Factor w/ 2452 levels "Additives and
Oxygenates - Exports",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ year                 : int  1996 1995 2014 2013 2012 2011
2010 2009 1998 1995 ...
##  $ unit                 : Factor w/ 6 levels "Cubic metres,
thousand",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ quantity             : num  5 17 0 0 35 25 22 45 1 7 ...
##  $ quantity_footnotes   : int  NA NA NA NA NA NA NA NA NA
NA ...
##  $ category             : Factor w/ 71 levels
"additives_and_oxygenates",..: 1 1 1 1 1 1 1 1 1 1 ...

anyNA(raw_stats$quantity_footnotes)

## [1] TRUE

sum(is.na(raw_stats$quantity_footnotes))

## [1] 1025536

ncol(raw_stats)

## [1] 7

nrow(raw_stats)

## [1] 1189482
```

Dataset is 7 columns x 1,189,482 rows. Lots of N/A's in "quantity footnotes variable". Check to see how many.

```
(sum(is.na(raw_stats$quantity_footnotes))/nrow(raw_stats))*100

## [1] 86.21703
```

86% N/As! We will need to drop this column. For now, we need some descriptive statistics of the individual columns. First country_or_area

```
country_detail <- raw_stats %>% group_by(country_or_area) %>%
summarise(occurences = length(country_or_area)) %>%
arrange(desc(occurences))

head(country_detail, n=10)

## # A tibble: 10 x 2
##    country_or_area occurences
##    <fct>                <int>
##  1 Germany              20422
##  2 United States        19847
##  3 Poland               19802
##  4 Austria              17440
##  5 Romania              17357
##  6 France               17236
##  7 Japan                17037
##  8 Czechia              16588
##  9 Italy                16312
## 10 Netherlands          15955
```

```
tail(country_detail, n=10)
```

```
## # A tibble: 10 x 2
##    country_or_area                          occurences
##    <fct>                                         <int>
##  1 South Sudan                                     305
##  2 Germany, Fed. R. (former)                       293
##  3 Bonaire, St Eustatius, Saba                     224
##  4 Sint Maarten (Dutch part)                       219
##  5 German Dem. R. (former)                         106
##  6 Antarctic Fisheries                              90
##  7 Pacific Islands (former)                         68
##  8 Yemen, Dem. (former)                             61
##  9 Yemen Arab Rep. (former)                         45
## 10 Commonwealth of Independent States (CIS)         16
```

```
anyNA(country_detail)
```

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

```
str(country_detail)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    243 obs. of  2
variables:
##  $ country_or_area: Factor w/ 243 levels "Afghanistan",..: 84
229 172 14 178 77 111 58 109 153 ...
##  $ occurences     : int  20422 19847 19802 17440 17357 17236
17037 16588 16312 15955 ...
```

```
summary(country_detail)
```

```
##        country_or_area   occurences
##  Afghanistan   :  1    Min.   :   16
##  Albania       :  1    1st Qu.: 1914
##  Algeria       :  1    Median : 3406
##  American Samoa:  1    Mean   : 4895
##  Andorra       :  1    3rd Qu.: 5890
##  Angola        :  1    Max.   :20422
##  (Other)       :237
```

Commodity transaction stats:

```
commodity_detail <- raw_stats %>% group_by(commodity_transaction)
%>% summarise(occurences = length(commodity_transaction)) %>%
arrange(desc(occurences))
```

```r
head(commodity_detail, n=10)
```

```
## # A tibble: 10 x 2
##    commodity_transaction
occurences
##    <fct>
<int>
##  1 From combustible fuels – Main activity
6601
##  2 Electricity - Gross demand
5532
##  3 Electricity - Gross production
5523
##  4 Electricity - net production
5523
##  5 Electricity - Own use by electricity, heat and CHP plants
5523
##  6 Electricity - total production, main activity
5523
##  7 Electricity - total net installed capacity of electric
powe…       5521
##  8 Electricity - total net installed capacity of electric
powe…       5521
##  9 Electricity - Final energy consumption
5499
## 10 Electricity - Consumption by other
5491
```

```r
tail(commodity_detail, n=10)
```

```
## # A tibble: 10 x 2
##    commodity_transaction
occurences
##    <fct>
<int>
##  1 Refinery gas - Transformation in coke ovens
1
##  2 "Vegetal waste - Consumption by construction "
1
##  3 "Vegetal waste - Consumption by mining and quarrying "
1
##  4 "White spirit and special boiling point industrial spirits
…          1
##  5 "White spirit and special boiling point industrial spirits
…          1
##  6 "White spirit and special boiling point industrial spirits
…          1
##  7 White spirit and special boiling point industrial spirits -
…          1
##  8 "White spirit and special boiling point industrial spirits
```

```
…          1
##  9 "White spirit and special boiling point industrial spirits
…          1
## 10 "White spirit and special boiling point industrial spirits
…          1
```

```r
anyNA(commodity_detail)
```

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

```r
str(commodity_detail)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    2452 obs. of  2
variables:
##  $ commodity_transaction: Factor w/ 2452 levels "Additives and
Oxygenates - Exports",..: 832 719 720 737 744 766 758 759 718 702
...
##  $ occurences           : int  6601 5532 5523 5523 5523 5523
5521 5521 5499 5491 ...
```

```r
summary(commodity_detail)
```

```
##
commodity_transaction
##  Additives and Oxygenates - Exports                   :   1

##  Additives and Oxygenates - Imports                   :   1

##  Additives and Oxygenates - Production                :   1

##  Additives and Oxygenates - Receipts from other sources:   1

##  Additives and Oxygenates - Stock changes             :   1

##  Additives and Oxygenates - Total energy supply       :   1

##  (Other)                                              :2446

##   occurences
##  Min.   :   1.0
##  1st Qu.:  23.0
##  Median :  99.0
##  Mean   : 485.1
##  3rd Qu.: 476.0
##  Max.   :6601.0
##
```

Year is pretty straightforward.

```r
year_detail <- raw_stats %>% group_by(year) %>%
summarise(occurences = length(year)) %>%
arrange(desc(occurences))

year_detail

## # A tibble: 25 x 2
##     year occurences
##    <int>      <int>
##  1  2014      56264
##  2  2013      56109
##  3  2012      55838
##  4  2011      55214
##  5  2010      54544
##  6  2008      53852
##  7  2009      53769
##  8  2007      52248
##  9  2006      49397
## 10  2005      49203
## # ... with 15 more rows

anyNA(year_detail)

## [1] FALSE

str(year_detail)

## Classes 'tbl_df', 'tbl' and 'data.frame':    25 obs. of  2
variables:
##  $ year      : int  2014 2013 2012 2011 2010 2008 2009 2007
2006 2005 ...
##  $ occurences: int  56264 56109 55838 55214 54544 53852 53769
52248 49397 49203 ...

summary(year_detail)

##       year        occurences
##  Min.   :1990   Min.   :36280
##  1st Qu.:1996   1st Qu.:43550
##  Median :2002   Median :46520
##  Mean   :2002   Mean   :47579
##  3rd Qu.:2008   3rd Qu.:53769
##  Max.   :2014   Max.   :56264
```

Unit column:

```
unit_detail <- raw_stats %>% group_by(unit) %>%
summarise(occurences = length(unit)) %>%
arrange(desc(occurences))

unit_detail

## # A tibble: 6 x 2
##   unit                    occurences
##   <fct>                        <int>
## 1 Metric tons,  thousand      759859
## 2 Terajoules                  178937
## 3 Kilowatt-hours, million     147741
## 4 Cubic metres, thousand       52032
## 5 Kilowatts,  thousand         50229
## 6 Metric Tons                    684
```

```
anyNA(unit_detail)
```

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

```
str(unit_detail)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    6 obs. of  2
variables:
##  $ unit      : Factor w/ 6 levels "Cubic metres, thousand",..:
5 6 2 1 3 4
##  $ occurences: int  759859 178937 147741 52032 50229 684
```

```
summary(unit_detail)
```

```
##                         unit     occurences
##  Cubic metres, thousand :1   Min.   :   684
##  Kilowatt-hours, million:1   1st Qu.: 50680
##  Kilowatts,  thousand   :1   Median : 99886
##  Metric Tons            :1   Mean   :198247
##  Metric tons,  thousand :1   3rd Qu.:171138
##  Terajoules             :1   Max.   :759859
```

Quantity column:

```
anyNA(raw_stats$quantity)
```

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

```
str(raw_stats$quantity)
```

```
##  num [1:1189482] 5 17 0 0 35 25 22 45 1 7 ...
```

```
summary(raw_stats$quantity)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.
Max.
##    -864348         14        189     184265       2265
6680329000
```

We already know about quantity_footnotes so next up is the category column:

```
category_detail <- raw_stats %>% group_by(category) %>%
summarise(occurences = length(category)) %>%
arrange(desc(occurences))
```

```
head(category_detail, n=10)

## # A tibble: 10 x 2
##    category
occurences
##    <fct>
<int>
##  1 total_electricity
133916
##  2 gas_oil_diesel_oil
97645
##  3 fuel_oil
75132
##  4 natural_gas_including_lng
64161
##  5 liquified_petroleum_gas
62156
##  6 motor_gasoline
53198
##  7 fuelwood
52032
##  8 electricity_net_installed_capacity_of_electric_power_plants
50229
##  9 other_kerosene
43466
## 10 hard_coal
42307

tail(category_detail, n=10)

## # A tibble: 10 x 2
##    category                               occurences
##    <fct>                                       <int>
##  1 gasoline_type_jet_fuel                       1293
##  2 falling_water                                 962
##  3 solar_electricity                             953
##  4 nuclear_electricity                           756
##  5 oil_shale_oil_sands                           756
##  6 uranium                                       684
##  7 geothermal                                    496
##  8 gas_coke                                      365
##  9 other_coal_products                           105
## 10 tide_wave_and_ocean_electricity                58

anyNA(category_detail)

## [1] FALSE

str(category_detail)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    71 obs. of  2
variables:
##  $ category  : Factor w/ 71 levels
"additives_and_oxygenates",..: 67 27 24 42 37 39 25 21 51 31 ...
##  $ occurences: int  133916 97645 75132 64161 62156 53198 52032
50229 43466 42307 ...
```

```
summary(category_detail)
```

```
##                        category    occurences
##  additives_and_oxygenates: 1   Min.   :    58
##  animal_waste            : 1   1st Qu.:  2208
##  anthracite              : 1   Median :  6470
##  aviation_gasoline       : 1   Mean   : 16753
##  bagasse                 : 1   3rd Qu.: 20236
##  biodiesel               : 1   Max.   :133916
##  (Other)                 :65
```

We do some cleanup.

```
rm(category_detail)

rm(commodity_detail)

rm(country_detail)

rm(unit_detail)

rm(year_detail)
```

Lastly we drop the quantity footnotes column and use the raw statistics as a tibble dataframe going forward.

```
test_data <- as_tibble(raw_stats)

class(test_data)

## [1] "tbl_df"     "tbl"        "data.frame"

test_data <- test_data %>% select(-quantity_footnotes)
```

Part I: Hard Coal

We filter the categories of interest, beginning with ‘Hard coal’. We drop columns we don’t need, group the countries together, and sort the results in ascending order by country followed by year. Lastly we nest the result by the grouped country.

```
hard_coal <- test_data %>% filter(commodity_transaction == "Hard
coal - transformation in electricity, CHP and heat plants") %>%
select(-commodity_transaction, -category) %>%
group_by(country_or_area) %>% arrange(country_or_area, year) %>%
nest()

head(hard_coal)
```

```
## # A tibble: 6 x 2
##   country_or_area data
##   <fct>           <list>
## 1 Afghanistan     <tibble [16 × 3]>
## 2 Argentina       <tibble [25 × 3]>
## 3 Australia       <tibble [25 × 3]>
## 4 Austria         <tibble [25 × 3]>
## 5 Bangladesh      <tibble [19 × 3]>
## 6 Belarus         <tibble [9 × 3]>
```

```
# Check to see the structure of the 'data' tibble - say
Afghanistan
pluck(hard_coal, "data") %>% pluck(1) %>% head()
```

```
## # A tibble: 6 x 3
##    year unit                   quantity
##   <int> <fct>                     <dbl>
## 1  1990 Metric tons,  thousand       40
## 2  1991 Metric tons,  thousand       40
## 3  2001 Metric tons,  thousand       20
## 4  2002 Metric tons,  thousand       20
## 5  2003 Metric tons,  thousand       30
## 6  2004 Metric tons,  thousand       30
```

We create new data columns using the 'mutate' and 'map' commands. From the data we extract the following information: - initial_year: (first recorded year of transforming this resource), initial_transformation (recorded units of transformation in first recorded year) - linear model: (derived linear model of transformation units as described by year) - slope: (slope of linear model: +ve/-ve) - r_squared: (statistical measure of how close the model data is to the fitted regression line)

```
hard_coal <- test_data %>% filter(commodity_transaction == "Hard
coal - transformation in electricity, CHP and heat plants") %>%
select(-commodity_transaction, -category) %>%
group_by(country_or_area) %>% arrange(country_or_area, year) %>%
nest() %>% mutate(initial_year = map_int((map(data, "year")), 1),
initial_transformation = map_dbl((map(data, "quantity")), 1),
model = map(data, ~lm(quantity ~ year, data =  .)), slope =
map_dbl(model, ~pluck(coef(.), "year")), r_squared =
map_dbl(model, ~pluck(glance(.), "r.squared")) )

head(hard_coal)
```

```
## # A tibble: 6 x 7
##   country_or_area data  initial_year initial_transfo… model
slope
##   <fct>           <lis>        <int>            <dbl> <lis>
<dbl>
## 1 Afghanistan     <tib…         1990               40 <S3:…
0.707
## 2 Argentina       <tib…         1990              205 <S3:…
23.3
## 3 Australia       <tib…         1990            23913 <S3:… -
139.
## 4 Austria         <tib…         1990             1421 <S3:…
19.1
## 5 Bangladesh      <tib…         1990                0 <S3:…
26.6
## 6 Belarus         <tib…         2006               73 <S3:…
-7.12
## # ... with 1 more variable: r_squared <dbl>
```

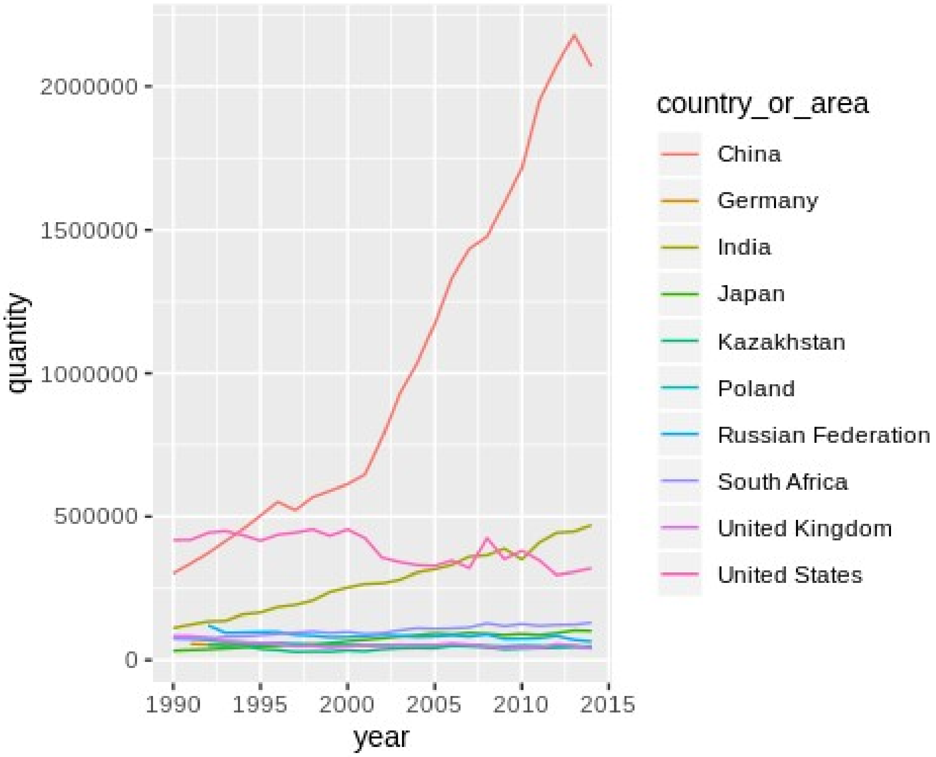
We can now begin our analysis on this data. We obtain the a list of the top 10 countries that began with the highest transformtion of coal into electricity.

```
hard_coal %>% arrange(desc(initial_transformation)) %>% head(10)
```

```
## # A tibble: 10 x 7
##    country_or_area data  initial_year initial_transfo… model
slope
##    <fct>           <lis>        <int>            <dbl> <lis>
<dbl>
##  1 United States   <tib…         1990           418513 <S3:…
-5766.
##  2 China           <tib…         1990           301998 <S3:…
81557.
##  3 Russian Federa… <tib…         1992           121629 <S3:…
-1343.
##  4 India           <tib…         1990           111940 <S3:…
14854.
##  5 United Kingdom  <tib…         1990            84014 <S3:…
-1218.
##  6 Poland          <tib…         1990            77554 <S3:…
-1010.
##  7 South Africa    <tib…         1990            74186 <S3:…
2371.
##  8 Germany         <tib…         1991            55723 <S3:…
-622.
##  9 Kazakhstan      <tib…         1992            52140 <S3:…
197.
## 10 Japan           <tib…         1990            31785 <S3:…
3103.
## # ... with 1 more variable: r_squared <dbl>
```

At this point we can generate a chart to see how these countries hard coal transformation into electricity change over time.

```
hard_coal %>% arrange(desc(initial_transformation)) %>% head(10)
%>% unnest(data) %>% ggplot(country_or_area, mapping = aes(x =
year, y = quantity)) + geom_line(mapping = aes(color =
country_or_area))
```



```
# We may need to tease this out or do a logarithmic chart to
better represent this data.
```

Part II: Brown Coal

Same code as before but different variable.

```
brown_coal <- test_data %>% filter(commodity_transaction ==
"Brown coal - Transformation in electricity, CHP and heat
plants") %>% select(-commodity_transaction, -category) %>%
group_by(country_or_area) %>% arrange(country_or_area, year) %>%
nest()

head(brown_coal)

## # A tibble: 6 x 2
##   country_or_area        data
##   <fct>                  <list>
## 1 Australia              <tibble [25 × 3]>
## 2 Austria                <tibble [17 × 3]>
## 3 Belgium                <tibble [15 × 3]>
## 4 Bosnia and Herzegovina <tibble [23 × 3]>
## 5 Bulgaria               <tibble [25 × 3]>
## 6 Cambodia               <tibble [7 × 3]>
```

```
pluck(brown_coal, "data") %>% pluck(1) %>% head()
```

```
## # A tibble: 6 x 3
##    year unit                   quantity
##   <int> <fct>                     <dbl>
## 1  1990 Metric tons,  thousand    58421
## 2  1991 Metric tons,  thousand    62332
## 3  1992 Metric tons,  thousand    64012
## 4  1993 Metric tons,  thousand    61619
## 5  1994 Metric tons,  thousand    64849
## 6  1995 Metric tons,  thousand    66407
```

```
brown_coal <- test_data %>% filter(commodity_transaction ==
"Brown coal - Transformation in electricity, CHP and heat
plants") %>% select(-commodity_transaction, -category) %>%
group_by(country_or_area) %>% arrange(country_or_area, year) %>%
nest() %>% mutate(initial_year = map_int((map(data, "year")), 1),
initial_transformation = map_dbl((map(data, "quantity")), 1),
model = map(data, ~lm(quantity ~ year, data =  .)), slope =
map_dbl(model, ~pluck(coef(.), "year")), r_squared =
map_dbl(model, ~pluck(glance(.), "r.squared")) )
```

```
head(brown_coal)
```

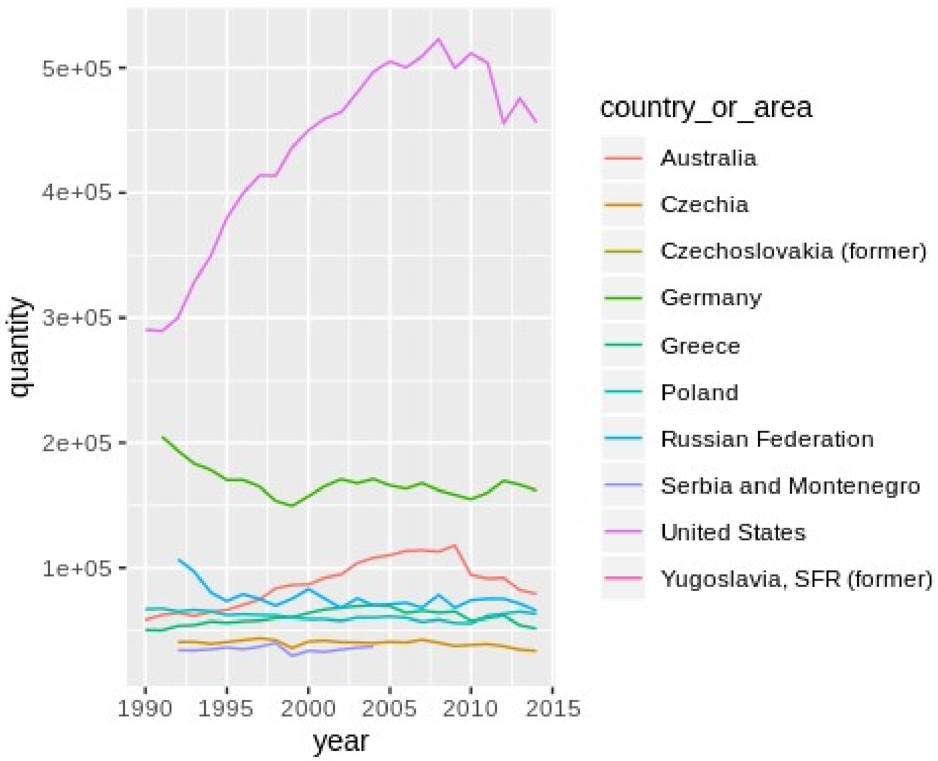
```
## # A tibble: 6 x 7
##   country_or_area data  initial_year initial_transfo… model
slope
##   <fct>           <lis>        <int>            <dbl> <lis>
<dbl>
## 1 Australia       <tib…         1990            58421 <S3:…
1780.
## 2 Austria         <tib…         1990             2133 <S3:…
-43.7
## 3 Belgium         <tib…         1990              936 <S3:…
-56.3
## 4 Bosnia and Her… <tib…         1992             7317 <S3:…
389.
## 5 Bulgaria        <tib…         1990            26211 <S3:…
213.
## 6 Cambodia        <tib…         2008                0 <S3:…
58.4
## # ... with 1 more variable: r_squared <dbl>
```

Analysis and charts

```
brown_coal %>% arrange(desc(initial_transformation)) %>% head(10)
```

```
## # A tibble: 10 x 7
##    country_or_area data  initial_year initial_transfo… model
slope
##    <fct>           <lis>        <int>             <dbl> <lis>
<dbl>
##  1 United States   <tib…         1990            290523 <S3:…
8599.
##  2 Germany         <tib…         1991            204903 <S3:…
-986.
##  3 Russian Federa… <tib…         1992            106834 <S3:…
-830.
##  4 Poland          <tib…         1990             66915 <S3:…
-234.
##  5 Czechoslovakia… <tib…         1990             63000 <S3:…
NA
##  6 Yugoslavia, SF… <tib…         1990             60458 <S3:…
NA
##  7 Australia       <tib…         1990             58421 <S3:…
1780.
##  8 Greece          <tib…         1990             50531 <S3:…
302.
##  9 Czechia         <tib…         1992             40889 <S3:…
-224.
## 10 Serbia and Mon… <tib…         1992             34158 <S3:…
41.7
## # ... with 1 more variable: r_squared <dbl>
```

```
brown_coal %>% arrange(desc(initial_transformation)) %>% head(10)
%>% unnest(data) %>% ggplot(country_or_area, mapping = aes(x =
year, y = quantity)) + geom_line(mapping = aes(color =
country_or_area))
```



Part III: Fuel Oil

```
fuel_oil <- test_data %>% filter(commodity_transaction == "Fuel
oil - Transformation in electricity, CHP and heat plants") %>%
select(-commodity_transaction, -category) %>%
group_by(country_or_area) %>% arrange(country_or_area, year) %>%
nest()

head(fuel_oil)

## # A tibble: 6 x 2
##   country_or_area     data
##   <fct>               <list>
## 1 Afghanistan         <tibble [24 × 3]>
## 2 Albania             <tibble [18 × 3]>
## 3 Algeria             <tibble [8 × 3]>
## 4 Angola              <tibble [25 × 3]>
## 5 Antigua and Barbuda <tibble [25 × 3]>
## 6 Argentina           <tibble [25 × 3]>

pluck(fuel_oil, "data") %>% pluck(1) %>% head()

## # A tibble: 6 x 3
##    year unit                      quantity
##   <int> <fct>                        <dbl>
## 1  1990 Metric tons,  thousand           4
## 2  1991 Metric tons,  thousand           3
## 3  1992 Metric tons,  thousand           2
## 4  1993 Metric tons,  thousand        2
## 5  1994 Metric tons,  thousand        2
## 6  1995 Metric tons,  thousand        2
```

```
fuel_oil <- test_data %>% filter(commodity_transaction == "Fuel
oil - Transformation in electricity, CHP and heat plants") %>%
select(-commodity_transaction, -category) %>%
group_by(country_or_area) %>% arrange(country_or_area, year) %>%
nest() %>% mutate(initial_year = map_int((map(data, "year")), 1),
initial_transformation = map_dbl((map(data, "quantity")), 1),
model = map(data, ~lm(quantity ~ year, data =  .)), slope =
map_dbl(model, ~pluck(coef(.), "year")), r_squared =
map_dbl(model, ~pluck(glance(.), "r.squared")) )
```

```
## Warning in stats::summary.lm(x): essentially perfect fit:
summary may be
## unreliable
```

```
head(fuel_oil)
```

```
## # A tibble: 6 x 7
##   country_or_area data  initial_year initial_transfo… model
slope
##   <fct>           <lis>        <int>            <dbl> <lis>
<dbl>
## 1 Afghanistan     <tib…         1990                4 <S3:…  -
0.0818
## 2 Albania         <tib…         1990              169 <S3:…  -
6.77
## 3 Algeria         <tib…         1990                0 <S3:…  -
0.0357
## 4 Angola          <tib…         1990               40 <S3:…
6.96
## 5 Antigua and Ba… <tib…         1990                9 <S3:…
1.26
## 6 Argentina       <tib…         1990             1800 <S3:…
67.1
## # ... with 1 more variable: r_squared <dbl>
```

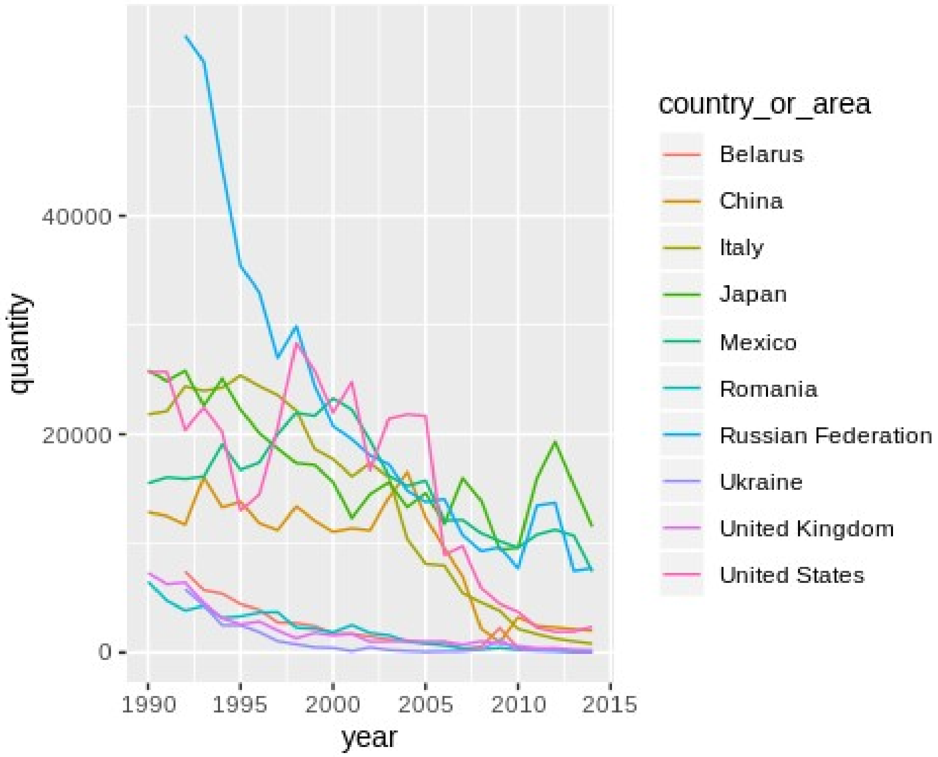
Analysis and charts

```
fuel_oil %>% arrange(desc(initial_transformation)) %>% head(10)
```

```
## # A tibble: 10 x 7
##    country_or_area data  initial_year initial_transfo… model
slope
##    <fct>           <lis>        <int>            <dbl> <lis>
<dbl>
##  1 Russian Federa… <tib…         1992            56504 <S3:… -
1905.
##  2 Japan           <tib…         1990            25834 <S3:…
-536.
```

```
##  3 United States   <tib…         1990            25666 <S3:…
-999.
##  4 Italy           <tib…         1990            21798 <S3:… -
1197.
##  5 Mexico          <tib…         1990            15508 <S3:…
-407.
##  6 China           <tib…         1990            12856 <S3:…
-547.
##  7 Belarus         <tib…         1992             7434 <S3:…
-264.
##  8 United Kingdom  <tib…         1990             7313 <S3:…
-235.
##  9 Romania         <tib…         1990             6492 <S3:…
-229.
## 10 Ukraine         <tib…         1992             5800 <S3:…
-159.
## # ... with 1 more variable: r_squared <dbl>
```

```
fuel_oil %>% arrange(desc(initial_transformation)) %>% head(10)
%>% unnest(data) %>% ggplot(country_or_area, mapping = aes(x =
year, y = quantity)) + geom_line(mapping = aes(color =
country_or_area))
```



Part IV: Gas Oil/Diesel Oil

```
gasdiesel_oil <- test_data %>% filter(commodity_transaction ==
"Gas Oil/ Diesel Oil - Transformation in electricity, CHP and
heat plants") %>% select(-commodity_transaction, -category) %>%
group_by(country_or_area) %>% arrange(country_or_area, year) %>%
nest()

head(gasdiesel_oil)

## # A tibble: 6 x 2
##   country_or_area     data
##   <fct>               <list>
## 1 Afghanistan         <tibble [25 × 3]>
## 2 Albania             <tibble [3 × 3]>
## 3 Algeria             <tibble [25 × 3]>
## 4 Angola              <tibble [18 × 3]>
## 5 Anguilla            <tibble [25 × 3]>
## 6 Antigua and Barbuda <tibble [25 × 3]>

pluck(gasdiesel_oil, "data") %>% pluck(1) %>% head()

## # A tibble: 6 x 3
##    year unit                   quantity
##   <int> <fct>                     <dbl>
## 1  1990 Metric tons,  thousand       50
## 2  1991 Metric tons,  thousand       50
## 3  1992 Metric tons,  thousand       50
## 4  1993 Metric tons,  thousand       50
## 5  1994 Metric tons,  thousand       50
## 6  1995 Metric tons,  thousand       50

gasdiesel_oil <- test_data %>% filter(commodity_transaction ==
"Gas Oil/ Diesel Oil - Transformation in electricity, CHP and
heat plants") %>% select(-commodity_transaction, -category) %>%
group_by(country_or_area) %>% arrange(country_or_area, year) %>%
nest() %>% mutate(initial_year = map_int((map(data, "year")), 1),
initial_transformation = map_dbl((map(data, "quantity")), 1),
model = map(data, ~lm(quantity ~ year, data =  .)), slope =
map_dbl(model, ~pluck(coef(.), "year")), r_squared =
map_dbl(model, ~pluck(glance(.), "r.squared")) )

## Warning in stats::summary.lm(x): essentially perfect fit:
summary may be
## unreliable

head(gasdiesel_oil)

## # A tibble: 6 x 7
##   country_or_area data  initial_year initial_transfo… model
slope
##   <fct>           <lis>        <int>            <dbl> <lis>
<dbl>
## 1 Afghanistan     <tib…         1990               50 <S3:…  -
1.58
## 2 Albania         <tib…         2000               21 <S3:…  -
7.5
```

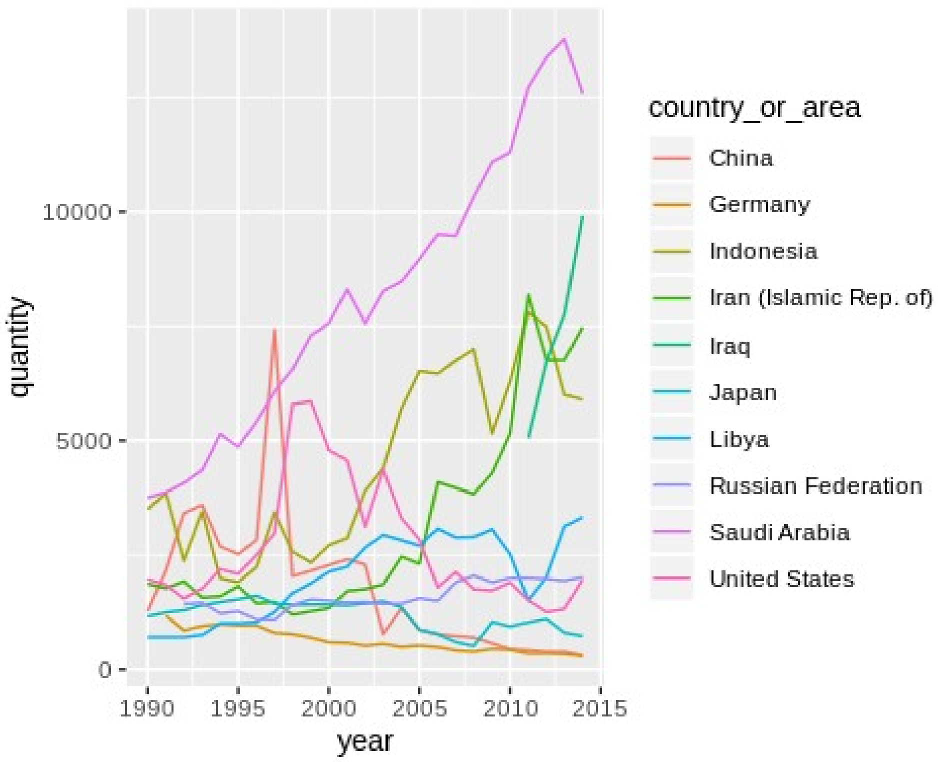
```
## 3 Algeria         <tib…         1990              125 <S3:…
25.2
## 4 Angola          <tib…         1997               51 <S3:…
42.1
## 5 Anguilla        <tib…         1990                4 <S3:…
0.807
## 6 Antigua and Ba… <tib…         1990               24 <S3:…
1.68
## # ... with 1 more variable: r_squared <dbl>
```

Analysis and charts

```
gasdiesel_oil %>% arrange(desc(initial_transformation)) %>%
head(10)
```

```
## # A tibble: 10 x 7
##    country_or_area data  initial_year initial_transfo… model
slope
##    <fct>           <lis>        <int>            <dbl> <lis>
<dbl>
##  1 Iraq            <tib…         2011             5061 <S3:…
1559.
##  2 Saudi Arabia    <tib…         1990             3752 <S3:…
417.
##  3 Indonesia       <tib…         1990             3500 <S3:…
216.
##  4 United States   <tib…         1990             1969 <S3:…
-40.7
##  5 Iran (Islamic … <tib…         1990             1868 <S3:…
246.
##  6 Russian Federa… <tib…         1992             1430 <S3:…
39.3
##  7 China           <tib…         1990             1269 <S3:…
-139.
##  8 Germany         <tib…         1991             1172 <S3:…
-33.6
##  9 Japan           <tib…         1990             1163 <S3:…
-29.7
## 10 Libya           <tib…         1990              700 <S3:…
103.
## # ... with 1 more variable: r_squared <dbl>
```

```
gasdiesel_oil %>% arrange(desc(initial_transformation)) %>%
head(10) %>% unnest(data) %>% ggplot(country_or_area, mapping =
aes(x = year, y = quantity)) + geom_line(mapping = aes(color =
country_or_area))
```



Part V: Natural Gas (including LNG)

```
natural_gas <- test_data %>% filter(commodity_transaction ==
"Natural gas (including LNG) - transformation in electricity, CHP
and heat plants") %>% select(-commodity_transaction, -category)
%>% group_by(country_or_area) %>% arrange(country_or_area, year)
%>% nest()

head(natural_gas)
```

```
## # A tibble: 6 x 2
##   country_or_area data
##   <fct>           <list>
## 1 Algeria         <tibble [25 × 3]>
## 2 Argentina       <tibble [25 × 3]>
## 3 Armenia         <tibble [23 × 3]>
## 4 Australia       <tibble [25 × 3]>
## 5 Austria         <tibble [25 × 3]>
## 6 Azerbaijan      <tibble [23 × 3]>
```

```
pluck(natural_gas, "data") %>% pluck(1) %>% head()
```

```
## # A tibble: 6 x 3
##    year unit        quantity
##   <int> <fct>          <dbl>
## 1  1990 Terajoules    179712
## 2  1991 Terajoules    192337
## 3  1992 Terajoules    200313
## 4  1993 Terajoules   237719
## 5  1994 Terajoules   252618
## 6  1995 Terajoules   259020
```

```
natural_gas <- test_data %>% filter(commodity_transaction ==
"Natural gas (including LNG) - transformation in electricity, CHP
and heat plants") %>% select(-commodity_transaction, -category)
%>% group_by(country_or_area) %>% arrange(country_or_area, year)
%>% nest() %>% mutate(initial_year = map_int((map(data, "year")),
1), initial_transformation = map_dbl((map(data, "quantity")), 1),
model = map(data, ~lm(quantity ~ year, data =  .)), slope =
map_dbl(model, ~pluck(coef(.), "year")), r_squared =
map_dbl(model, ~pluck(glance(.), "r.squared")) )

head(natural_gas)
```

```
## # A tibble: 6 x 7
##   country_or_area data  initial_year initial_transfo… model
slope
##   <fct>           <lis>        <int>            <dbl> <lis>
<dbl>
## 1 Algeria         <tib…         1990           179712 <S3:…
1.64e4
## 2 Argentina       <tib…         1990           243136 <S3:…
1.99e4
## 3 Armenia         <tib…         1992            22800 <S3:… -
3.06e1
## 4 Australia       <tib…         1990           161478 <S3:…
1.76e4
## 5 Austria         <tib…         1990            82181 <S3:…
3.44e2
## 6 Azerbaijan      <tib…         1992           117775 <S3:…
7.82e3
## # ... with 1 more variable: r_squared <dbl>
```

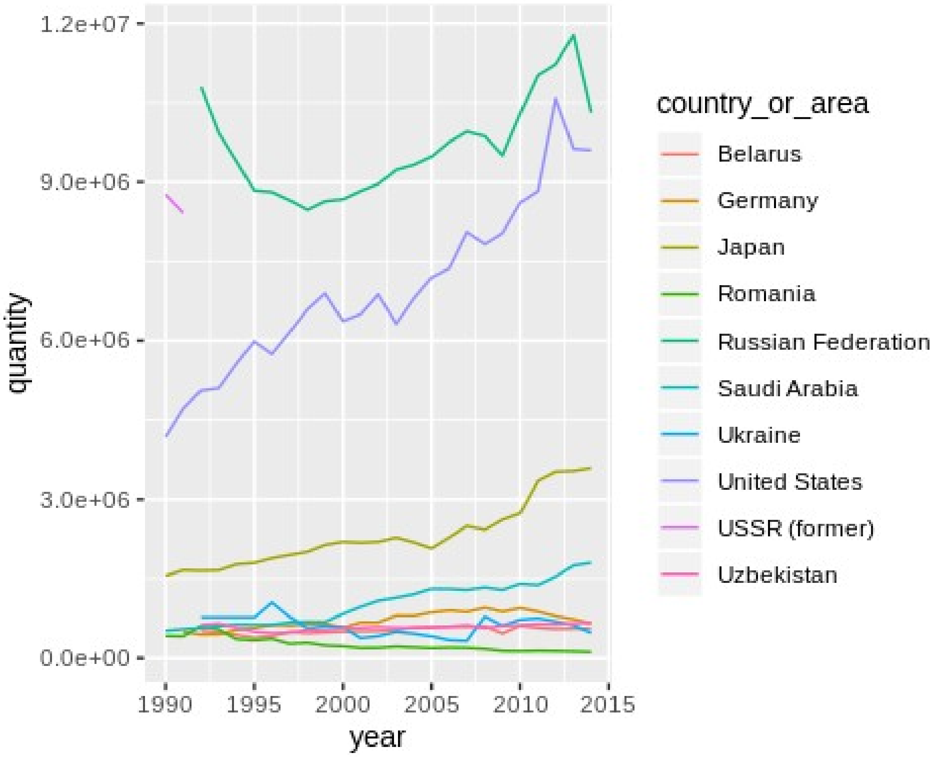
Analysis and charts

```
natural_gas %>% arrange(desc(initial_transformation)) %>%
head(10)
```

```
## # A tibble: 10 x 7
##    country_or_area data  initial_year initial_transfo… model
slope
##    <fct>           <lis>        <int>            <dbl> <lis>
<dbl>
##  1 Russian Federa… <tib…         1992         10794027 <S3:…
7.88e4
##  2 USSR (former)   <tib…         1990          8765937 <S3:… -
3.51e5
##  3 United States   <tib…         1990          4175718 <S3:…
2.10e5
```

```
##   4 Japan            <tib…          1990          1555133 <S3:…
7.63e4
##   5 Ukraine          <tib…          1992           765500 <S3:… -
9.35e3
##   6 Uzbekistan       <tib…          1992           622140 <S3:…
4.58e3
##   7 Saudi Arabia     <tib…          1990           516377 <S3:…
5.30e4
##   8 Belarus          <tib…          1992           511257 <S3:…
6.30e3
##   9 Germany          <tib…          1991           496505 <S3:…
1.77e4
## 10 Romania          <tib…          1990           417957 <S3:… -
1.58e4
## # ... with 1 more variable: r_squared <dbl>
```

```
natural_gas %>% arrange(desc(initial_transformation)) %>%
head(10) %>% unnest(data) %>% ggplot(country_or_area, mapping =
aes(x = year, y = quantity)) + geom_line(mapping = aes(color =
country_or_area))
```



We may want to export this data for some work in Hive.

```
brown_coal %>% arrange(desc(initial_transformation)) %>% head(10)
%>% select(-initial_year, -initial_transformation) %>%
unnest(data) %>% write_csv('brown_coal.csv')
```

```r
fuel_oil %>% arrange(desc(initial_transformation)) %>% head(10)
%>% select(-initial_year, -initial_transformation) %>%
unnest(data) %>% write_csv('fuel_oil.csv')

gasdiesel_oil %>% arrange(desc(initial_transformation)) %>%
head(10) %>% select(-initial_year, -initial_transformation) %>%
unnest(data) %>% write_csv('gasdiesel_oil.csv')

hard_coal %>% arrange(desc(initial_transformation)) %>% head(10)
%>% select(-initial_year, -initial_transformation) %>%
unnest(data) %>% write_csv('hard_coal.csv')

natural_gas %>% arrange(desc(initial_transformation)) %>%
head(10) %>% select(-initial_year, -initial_transformation) %>%
unnest(data) %>% write_csv('natural_gas.csv')
```