COS 30045 Data Visualisation
Process Book

Co2 Emissions from Fuel Combustion

Peter Micevski
ID: 100651602

Contents

1 Introduction

1.1 Background and Motivation

Proposal

Witnessing the beginning of the 21st century is unofficially believed by many people as one of the greatest moments of our time. We are not only exposed but also beneficiaries to the most advanced scientific developments of our time, with one of the next major advancements currently taking place in artificial intelligence, which may also be a technological development beyond our imagination.

However, as exciting as this may sound its not a moment in time where we can just stop working and go on permanent holiday. The current level of scientific advancement also requires equal effort in responsibility, the first place this responsibility needs to be applied is getting our house in order and this house is our planet and its health.

Due to the level of discussion in our society regarding co2 emissions produced by fossil fuels I would like a more detailed understanding of co2 emissions, data used will be from a report published by the International Energy Agency titled "CO2 emissions from fuel combustion 2018 highlights".

1.2 Project Objectives

Proposal

The main objective will be investigating CO2 emissions made by coal, oil and natural gas. These fuels dominate and support our modern social ecosystem, with great discussion and protests had, regarding coal fired power stations and their contributions to CO2 emissions. I feel the conversation can be biased at times with too much focus on coal and not enough focus on oil and natural gas. The primary questions will be:

What is the contribution of CO2 emissions by fuel type by nation?

What is the contribution of CO2 emissions by sector by nation?

Do wealthier nations have higher emissions?

Has technological advancement over time helped curb CO2 emissions?

Benefits I wish to achieve, will be helping people develop a more balanced and clearer interpretation of CO2 emissions produced by our modern ecosystem.

1.3 Project Schedule

Table 1. Project Schedule

Semester Week	Work to do…
Week 4	Have all data cleaned and prepared for exploration
Week 5	Begin exploration
Week 6	Begin Implementation and Progress Report
Week 7-10	Final Implementation and Final Report
Week 11-12	Presentation and Submission

2 Data

2.1 Data Source

Proposal

Data will be sourced from The International Energy Agency (IEA), published In the report “CO2 emissions from fuel combustion highlights 2018”, https://webstore.iea.org/co2-emissions-from-fuel-combustion-2018-highlights

The data sets are Tables, table items are countries and categories of countries and attributes are Categorical quantitative and Temporal time-series. There will be a total of up to seven data sets used including one data set for population time-series.

Data not included from the sets will be attribute data of percentage change in time-series data as it only refers to a portion of the data, as well as a set of items whose nomenclature is not relevant for the analysis.

2.2 Data Processing

Proposal

Data will require cleaning as it will be imported from pdf files, conversion into .xlsx to .csv requires reformatting tables, fonts and removal of page titles. For the moment no quantities will be derived however the us of a derived will be used, the variable is CO2/population (the amount of CO2 produced per each person). The processing and exploration of data will be performed by RStudio.

Progress Report

Initial processing required transforming the PDF files to CSV, first step taken was using PDFsam Basic Software to extract all pages with tables from Co2 combustions highlights 2018 report.

Next, converting PDF to CSV was a process of trial and error, this began by using online open source vendors such as www.zamzar.com and www.pdftoexcel.com however these sources turned out unsuitable for as data extraction was not in its full form. The unsatisfactory data extraction process led me to using Adobe Software available on Swinburne campus computers.

Moving on, data cleaning was performed by a mix of manually copy and pasting between CSV files and R-Studio processing, R-Studio code for process CSV files is displayed in Appendix A1, Data Clean Up Process Code.

At this stage data is ready for processing.

Final Report

Data processing revealed to be challenging task in the sense that it's a process in constant fluctuation, what I thought I was needed ended up being not required.

I thought that attempting to visualise all 149 countries would be adequate to address the questions I was asking, however this ended up not being the case. Having the viewer visualise each county individually doesn't provide the viewer with a large enough scope regarding the issue i.e. the ability in randomly selecting five countries means that they may select countries with little relation between them, and being able to memorise which country had what emission for 149 countries is a mentally strenuous task.

Fortunately, the dataset contained OECD and Non-OECD data by regions, extracting this data was as simple as copy/paste from PDF to CSV with some R-studio processing to eliminate white spaces.

And it's here that I think data processing can be challenge, was it not for the OECD and Non-OECD data available I would have needed to collate this data myself, by manually grouping countries in there respective OECD grouping and then sum all groups, an extremely time consuming process. On the other hand, the initial data processing was not used as a simpler solution was suitable, so time consumption was still a factor due to initial data encoding oversight.

In total five dataset were produced

- WORLD_FUEL_COMBUSTION_SECTOR_2016_fix.csv
- WORLD_REGION_**NATURAL_GAS**_COMB_LONG.csv
- WORLD_REGION_**TOTAL_COAL**_COMB_LONG.csv
- WORLD_REGION_**TOTAL_FUEL**_COMB_LONG.csv
- WORLD_REGION_**TOTAL_OIL**_COMB_LONG.csv

Region items used were

- OECD Americas
- OECD Europe
- OECD Asia Oceania
- Non-OECD Europe and Eurasia
- Africa
- Non-OECD Asia (Excl, China)
- China
- Non-OECD Americas
- Middle East
- Australia

Sector attributes used were

- Electricity/Heat production
- Other energy Ind
- Manufacturing/construction
- Transport
- of which Road
- Residential
- Commercial/public services

Finally, one last process was required having to do with algorithm implementation, datasets needed to be transformed from wide to long for, natural gas, total coal, total fuel and total oil. This was done by grouping all years together in a “year” attribute column and with there respective values in a value column. Value measurements of all datasets was by Million tonnes of Co2.

3 Requirements

3.1 Must-Have Features

Proposal

Must have features will be bar and line charts as simple as it may sound, I feel that being able to present my vision would be difficult.

Final Report

Must have features have been delivered as part of the project, line chart and bar chart delivered.

3.2 Optional Features

Proposal

The ability of interactivity of toggling between variables and multi-lined and stacked area charts would be a fantastic contribution towards the visual aspects of presenting the data.

Final Report

Interactivity was delivered however not as imagined, buttons and not toggles where used to change between variables on the line chart between Total Fuel, Coal, Oil and Natural Gas, a multiline chart for the previous variables was also implemented. A staked area line chart was not delivered due to complexity of D3.js.

4 Visualisation Design

Proposal

Visualisation will be presented in two charts, line and bar chart.

Line Chart Features:
Y-axis: Values are “Million Tonnes of Co2”
X-axis: Year Time-Series in 5-year intervals
Will be able to plot up to 5 countries, with the added feature of population for each country controlled by toggle switch, displayed in Figure 2, Line Chart Design.

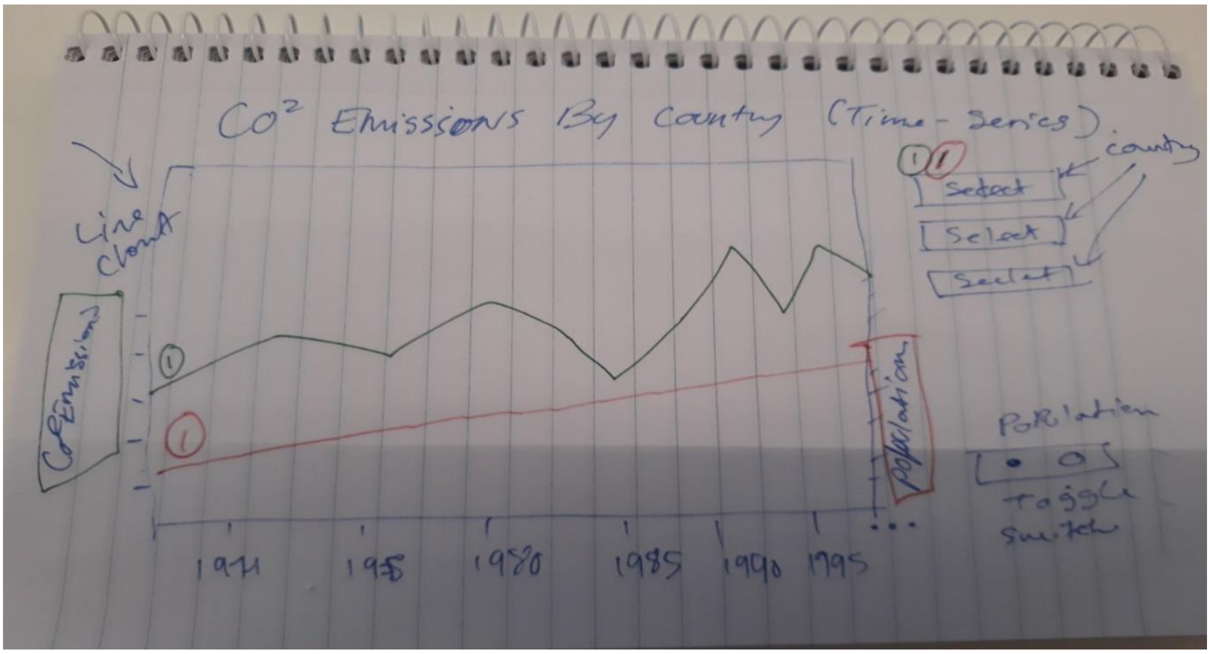

Figure 2, Line Chart Design

Bar Chart Features:
Y-axis: Values are “Million Tonnes of Co2”
X-axis: Categorical sector of each country, Electricity/Heat production, Other energy Ind, Manufacturing/construction, Transport, of which Road, Residential and Commercial/public services for 5 countries, displayed in Figure 3, Grouped Bar Chart.

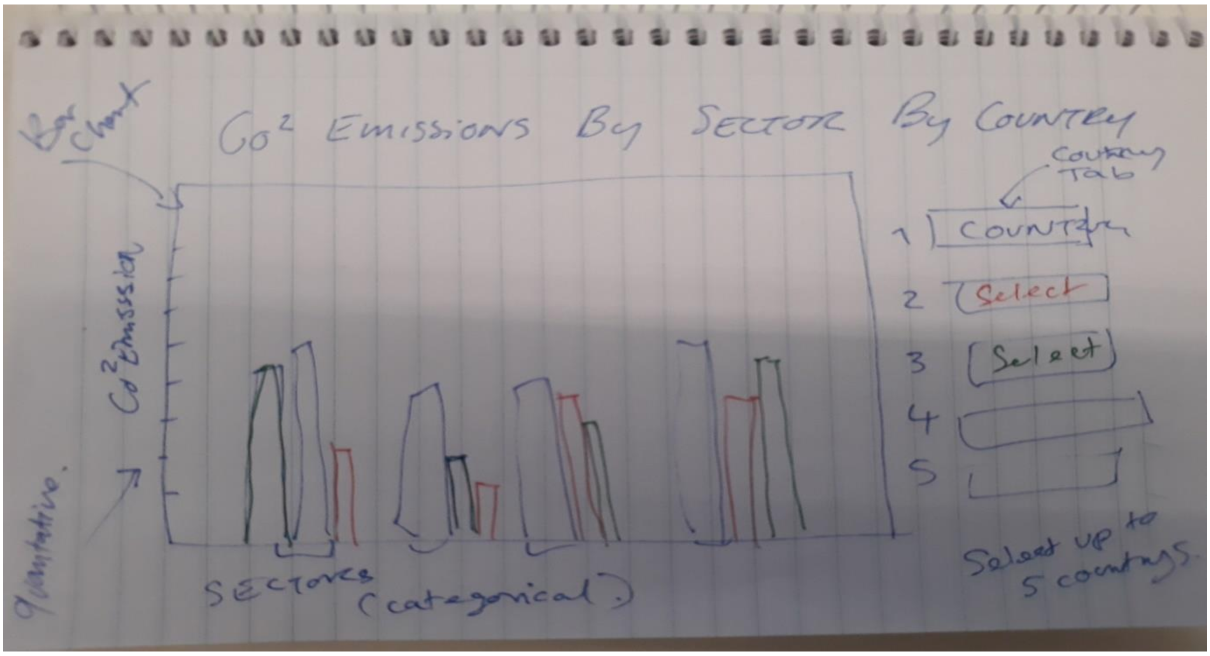

Figure 3, Grouped Bar Chart

Progress Report

First attempt at exploring visualisation was using R-studio gg-plot this was a line graph will all 149 counties, displayed in Figure 4, GG-Plot Line Chart by countries.

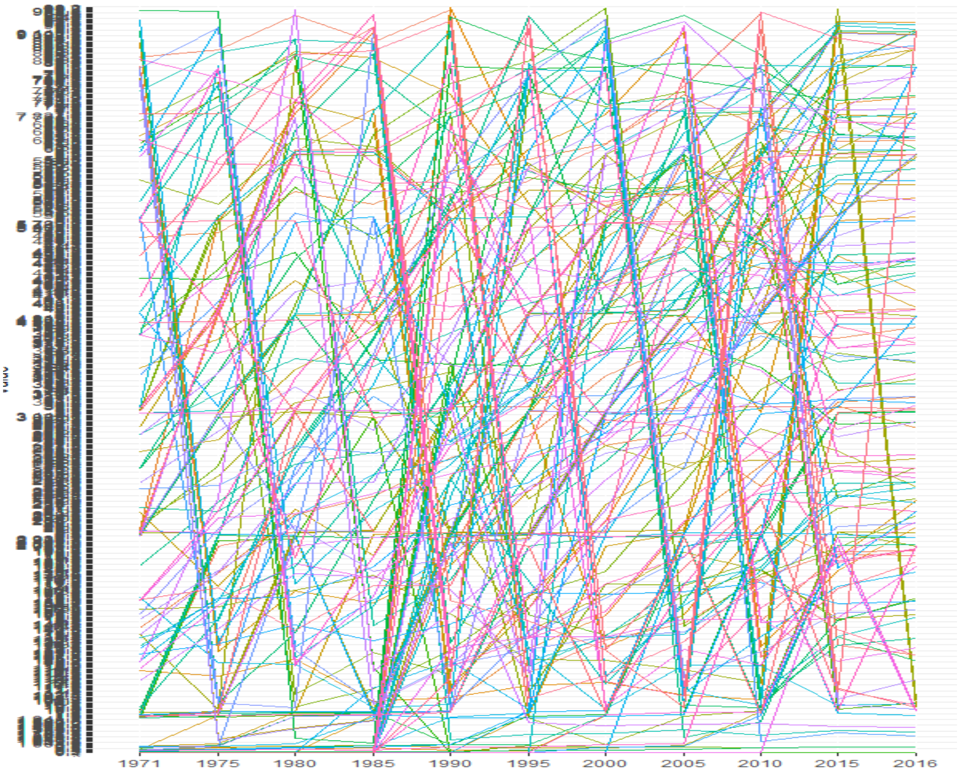

Figure 4, GG-Plot Line Chart by countries

As can be viewed making sense of the data in figure 4, GG-Plot Line Chart by countries is impossible, at this point I realised I needed to change my approach, I needed to group data in a regional format, fortunately OECD and Non-OECD data was also part of the dataset, data encoding was changed and produced the following output, displayed in Figure 6, GG-Plot Line Chart by Regions.

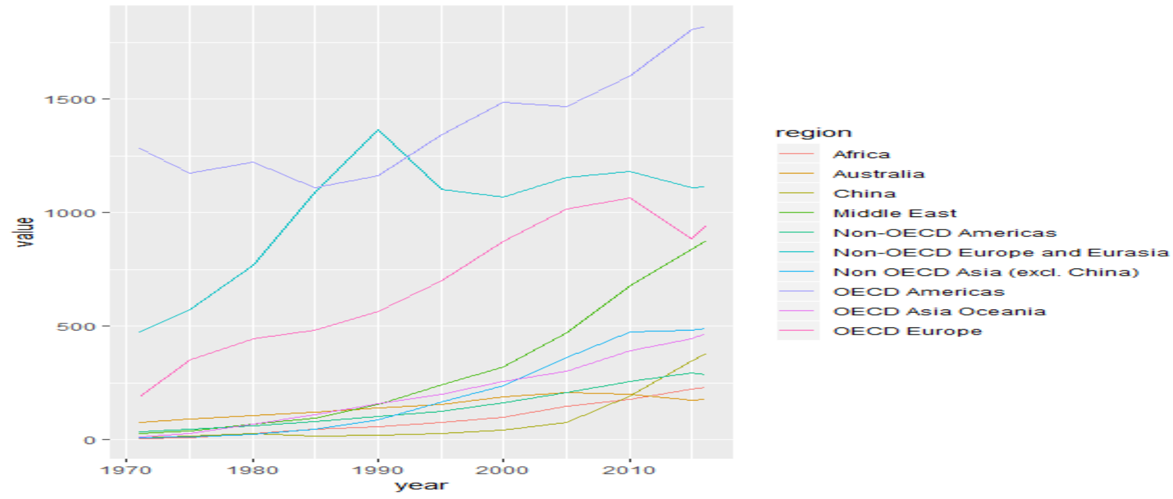

Figure 6, GG-Plot Line Chart by Regions

The above chart clearly indicates all regions and their contribution of Co2 emissions from 1971 to 2016, this validates data encoding is correct as the difference between variables is clearly visualised and provides the viewer with adequate information regarding the subject matter.

Moving on to exploring grouped bar chart displayed in Figure 7, GG-Plot Grouped Bar chart by Sector, reveals encoding is adequate, visualisation clearly displays data.

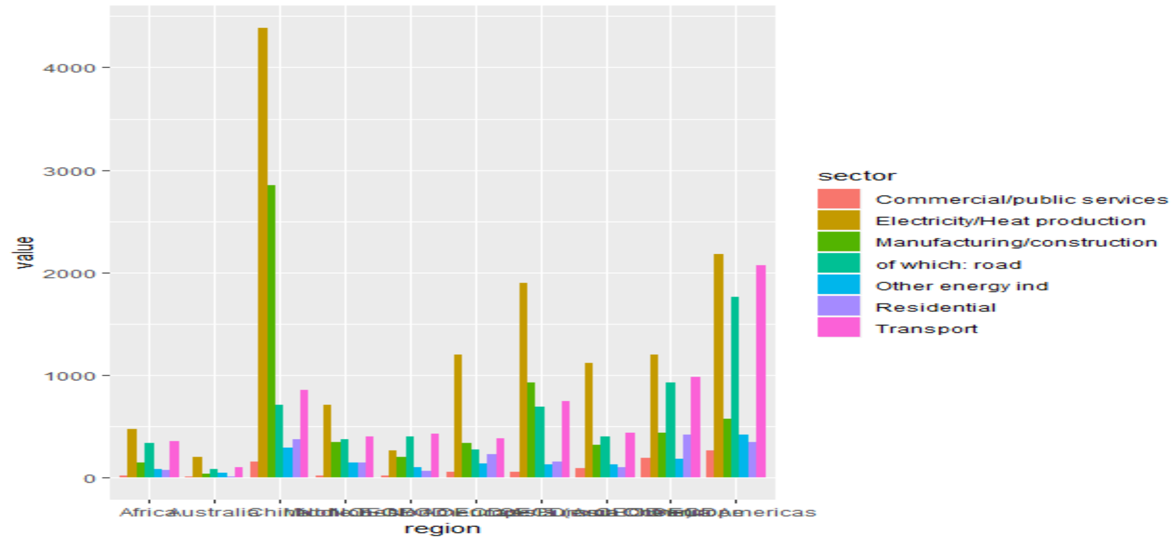

Figure 7, GG-Plot Grouped Bar chart by Sector

At this moment no visualisations are available in D3.

Final Report

Line Chart

The final visualisation of the multi-line chart envisioned in the progress report has been delivered, displayed in Figure 8, Co2 Emissions by Region.

Line chart background presents no grid lines, thus allowing a clean presentation of line data, I believe grid lines in this case would have delivered too much chart noise the 10 item variables and grid lines would have been in direct competition and present a challenge for the viewer especially in the lower bounds of the chart.

Both X and Y axis have clear labelling representing the measurements used “Million tonnes of Co2” for Y axis and “Emissions by Year” for each value attribute data point.

Colour used is Hue for categorical data colour schema used was the D3 schemeCategory20.

Legend on the top left corner clearly presents the respective colour to item “region” variable.

Button options allow the viewer the ability to select between fuel types the four buttons represent data for Total Fuel, Coal, Oil and Natural Gas.

Background colour used is WhiteSmoke I feel the subtlety of this colour is comforting for the viewer.

Lastly a contemporary approach was used in labelling the visualisation with bold heading above, Co2 Emissions by Fuel Type between OECD and Non-OECD Regions.

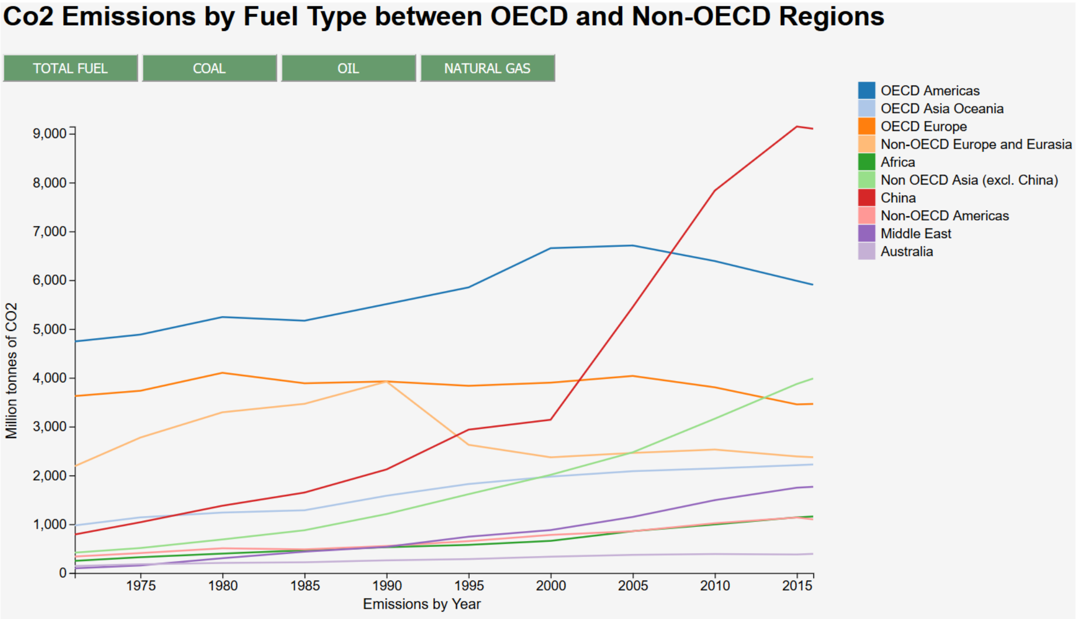

Figure 8, Co2 Emissions by Region

Grouped Bar Chart

The final visualisation of the grouped bar chart envisioned in the progress report has been delivered, displayed in Figure 9, Co2 Emissions by Sector.

Once again chart background presents no grid lines, to further add and valid for the above chart. Grid lines represent a finer detail of measurement, whilst completing this unit I have come to the belief what is most important in visualising data is about presenting the size and separation/distance between data attributes. Measurements need to be present and correct however not at the detriment of the visualisation, especially in my case as I'm presenting many variables a total of 70 in the grouped bar chart. Ten regions by Seven sectors.

All colour coding is same as the line chart retaining the same theme, X axis labelling is by its individual regions, I felt this was clear enough as to not need an additional label of "Region". Y axis label is in intervals of 500 and labelling "Million tonnes of Co2" was horizontal rather than vertical.

Legend was placed inside the chart compared to besides the chart in the line chart, I felt the filling of space was acceptable as the last three regions have an extremely low count. If on the other hand data present displayed more volatility, the legend would have been placed outside of the chart.

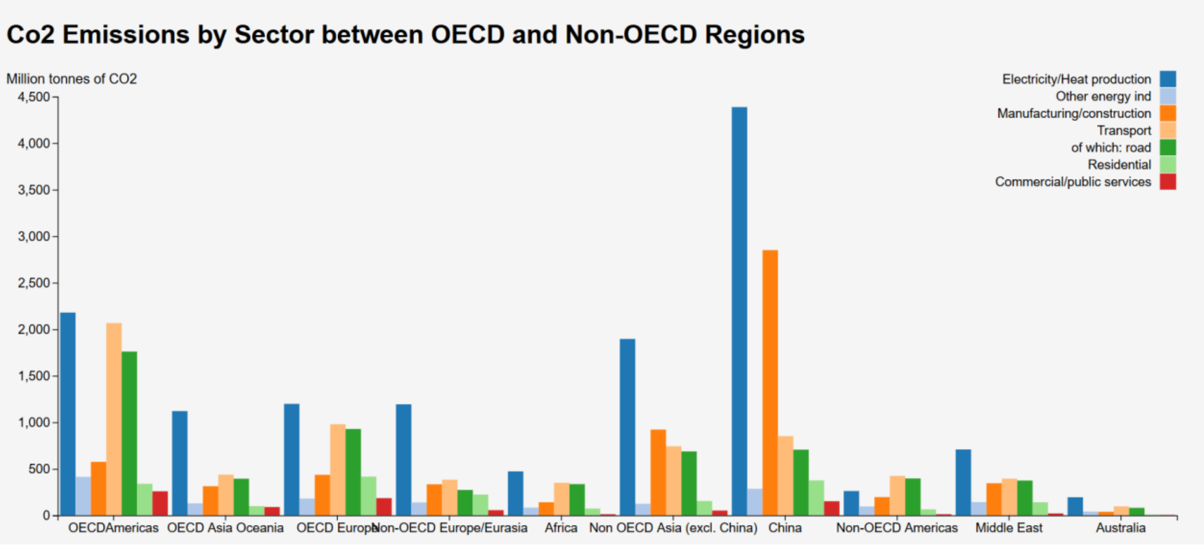

Figure 9, Co2 Emissions by Sector

5 Conclusion

Final

Project started with the simple task of selecting a dataset to visualise, however finding and selecting a dataset was not a simple process, it required many hours of thinking about what to visualise, which was then followed by viewing many datasets trying to establish whether they were adequate for visualising.

Having selected the dataset, the Domain had been established “Co2 Emissions form Fuel Combustion”. Data Abstraction was next, here is where my questions where established

- What is the contribution of Co2 by fuel type by nation?
- What is the contribution of Co2 by Sector by nation?
- Do wealthier Nations have higher emissions?
- Has Technology advancement over time help curb emissions?

The above questions ended up changing from “by nation” to “by region” due to complexity of visualising every nation.

Data encoding followed next, Items were, regions by OECD and Non-OECD. Attributes were Sectors – categorical/quantitative and Fuel Types – temporal/time-series.

Next step was transforming the dataset from PDF to CSV. The initial processes of open source web sites were fruitless tasks, data was not converted in full. This was followed by using Adobe on Swinburne Campus computers, data was successfully converted. Cleaning of data was performed on R-Studio and some copy/paste.

Data exploration was also performed on R-Studio, here is where I established the initial data abstraction was incorrect and need to change the analysis from nation to OECD and Non-OECD region.

Algorithm implementation of D3 graphs proved to be the greatest challenge of all, the first discovery was the need for transformation of datasets from wide to long for all fuel types. Year attributes were collated on a single attribute labelled year and their respective values were placed in a value attribute. This was achieved by R-Studio.

Next followed a slow process of trial and error when creating the visualisations, the low-level complexity of D3 required a detailed understanding of Java script in order to navigate D3. The delivery of these graphs was assisted by the examples provided at [1] for grouped bar charts, [2] d3-documantaion for learning how to handle csv files, colour examples, nests, scales. The multi-line graph[3], helping of creating the legend[4] and Interactive data visualisation[5].

The experience of delivering this visualisation was at times a testing exercise, the most testing of all was algorithm implementation which is only a matter of experience in java script and D3. Most important was the four nested levels of validation, understanding the domain and how to implement its encoding is most important, spending time in properly understanding how items and attributes can be best represented will assist immensely towards having these variables communicate their values. This was something I had discovered as part of this project, requiring me to revisit encoding a number of times throughout the process, it may not always be possible to get validation right the first time, however limiting the amount times you go back could be extremely time benefiting.

Chart design and colour selection is another important element, this part of visualisation is the difference of not only an appealing design, but the ability to communicate the data with integrity. Displaying information not relative to the data is labelled as chart junk [6] and takes away from the intended purpose.

Being aware of colour blindness is another important element of selecting colour schemas, having the right colour scheme could also make the difference of communicating the data correctly. This problem is solved by using colourbrewer.com.

In conclusion Data Visualisation has introduced me to the to an extremely broad discipline, from nested validation to algorithm deployment, humans visual perception and its psychology when using colours and shapes are incorrectly implemented.

The most amazing discovery for me is the awareness of the power the human visual channel has in absorbing large amounts of information when graphs are implemented correctly, is amazing.

References

[1]"Grouped Bar Chart", *Gist*, 2019. [Online]. Available: https://gist.github.com/mbostock/3887051. [Accessed: 02- Jun- 2019].

[2]"d3/d3", *GitHub*, 2019. [Online]. Available: https://github.com/d3/d3/wiki. [Accessed: 02- Jun- 2019].

[3]"Independent scale multi-line graph", *Bl.ocks.org*, 2019. [Online]. Available: https://bl.ocks.org/d3noob/6d50c569ee39ee5d6f26435e7586bc03. [Accessed: 02- Jun- 2019].

[4]"D3.js part 7 of 9 - Adding a legend to explain the data - Competa", *Competa*, 2019. [Online]. Available: https://www.competa.com/blog/d3-js-part-7-of-9-adding-a-legend-to-explain-the-data/. [Accessed: 02- Jun- 2019].

[5]S. Murray, *Interactive data visualization for the Web*

[6]E. Tufte, *The visual display of quantitative information*. Cheshire (Connecticut): Graphics Press, 2015.

Appendix A

A1, Data Clean Up Process Code

```
######### READING DATA
library(readr)
dv2Key_World_2018_2 <- read_csv("co2_combustion_71_16.csv")

######### COVERTING TO DATA FRAME
co2emdf <- as.data.frame(dv2Key_World_2018_2) #1

######## ELIMIMATING SPACES
co2emdf$X14 <- gsub(" ","",co2emdf$X14) #2

for(i in 1:length(co2emdf)){
  #print(i)
  co2emdf[[i]] <- gsub(" ","",co2emdf[[i]]) # working
}

######### CONVERTING DF TO NUMERIC
co2df_num <-as.data.frame(sapply(co2emdf, as.numeric))

rownames(co2df_num) <- make.names(co2emdf$`Selected indicators for 2016`, unique =
T) #### RENAMED DF ROW NAME AFTER BEING CHANGED TO NUMERIC

co2df_num <- co2df_num[,-c(1)]

saveRDS(co2df_num, file = "co2.rds")

co2df_num <- readRDS(file = "co2.rds")

library(plyr)

co2emdf <- rename(co2emdf, c("Selected indicators for 2016"="Country")) ### RENAMED
ROW NAME
```

A2, Transforming Wind to Long

```
########### D3 DATA TIDY WIDE TO LONG

library(readr)
library(tidyr)

Em_Fuel_Comb_Y <-
read_csv("WORLD_FUEL_COMBUSTION_SECTOR_2016_fix.csv")
```

```
Em_Fuel_Comb_Y <- gather(Em_Fuel_Comb_Y, `Electricity/Heat production`,
`Other energy ind`, `Manufacturing/construction`,
                    `Transport`, `of which: road`, `Residential`, `Commercial/public
services`, key=sector, value = "value")

Em_Fuel_Comb_Y <- gather(dv2Key_World_2018_2, `1971`, `1975`, `1980`,
                    `1985`, `1990`, `1995`, `2000`, `2005`,
                    `2010`, `2015`, `2016`, key=`year`, value = "value")
Em_Fuel_Comb_Y[3]

for(i in 3:length(Em_Fuel_Comb_Y)){
  #print(i)
  Em_Fuel_Comb_Y[[i]] <- gsub(" ","",Em_Fuel_Comb_Y[[i]]) # working
}

write_csv(Em_Fuel_Comb_Y, "ausNATtest.csv")
```