Visualising COVID-19

This Project will visualise the spread of COVID-19 in the early months of 2020 using R.

This was inspired by the DataCamp Project Visualizing COVID-19 by Richie Cotton.

First we will look into the confirmed covid cases found around the world in confirmed\_cases\_worldwide.csv data set.

#load the relevant packages  
library(readr)  
library(ggplot2)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

# Read datasets/confirmed\_cases\_worldwide.csv into confirmed\_cases  
confirmed\_cases <- read\_csv("Downloads/Visualizing COVID-19/datasets/confirmed\_cases\_worldwide.csv")

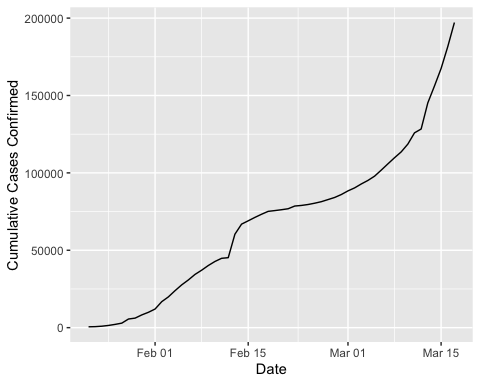
##   
## ── Column specification ───────────────────────────────────────────────  
## cols(  
## date = col\_date(format = ""),  
## cum\_cases = col\_double()  
## )

# Print out confirmed\_cases  
confirmed\_cases

## # A tibble: 56 x 2  
## date cum\_cases  
## <date> <dbl>  
## 1 2020-01-22 555  
## 2 2020-01-23 653  
## 3 2020-01-24 941  
## 4 2020-01-25 1434  
## 5 2020-01-26 2118  
## 6 2020-01-27 2927  
## 7 2020-01-28 5578  
## 8 2020-01-29 6166  
## 9 2020-01-30 8234  
## 10 2020-01-31 9927  
## # … with 46 more rows

The confirmed cases worldwide can be represented in a line plot below by visualising the cases against the date beginning in 22th January 2020.

# Draw a line plot of cumulative cases over each date  
  
# Label the y-axis  
ggplot(confirmed\_cases, aes(date, cum\_cases)) +  
 geom\_line() +  
 xlab("Date") +  
 ylab("Cumulative Cases Confirmed")



We can see a steep incline beginning towards the end of March with figures reaching almost 2000,000. We can also see a sharp increase around February 15th before returning to a steady incline that’s interesting. However it is interesting to focus on Chinas experience and compare to worldwide figures.

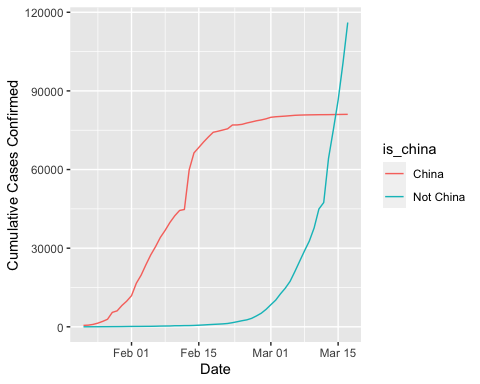
# Read in datasets/confirmed\_cases\_china\_vs\_world.csv  
confirmed\_cases\_with\_china <- read\_csv("Downloads/Visualizing COVID-19/datasets/confirmed\_cases\_china\_vs\_world.csv")

##   
## ── Column specification ───────────────────────────────────────────────  
## cols(  
## is\_china = col\_character(),  
## date = col\_date(format = ""),  
## cases = col\_double(),  
## cum\_cases = col\_double()  
## )

# See the result  
print(glimpse(confirmed\_cases\_with\_china))

## Rows: 112  
## Columns: 4  
## $ is\_china <chr> "China", "China", "China", "China", "China", "China", "Chin…  
## $ date <date> 2020-01-22, 2020-01-23, 2020-01-24, 2020-01-25, 2020-01-26…  
## $ cases <dbl> 548, 95, 277, 486, 669, 802, 2632, 578, 2054, 1661, 2089, 4…  
## $ cum\_cases <dbl> 548, 643, 920, 1406, 2075, 2877, 5509, 6087, 8141, 9802, 11…  
## # A tibble: 112 x 4  
## is\_china date cases cum\_cases  
## <chr> <date> <dbl> <dbl>  
## 1 China 2020-01-22 548 548  
## 2 China 2020-01-23 95 643  
## 3 China 2020-01-24 277 920  
## 4 China 2020-01-25 486 1406  
## 5 China 2020-01-26 669 2075  
## 6 China 2020-01-27 802 2877  
## 7 China 2020-01-28 2632 5509  
## 8 China 2020-01-29 578 6087  
## 9 China 2020-01-30 2054 8141  
## 10 China 2020-01-31 1661 9802  
## # … with 102 more rows

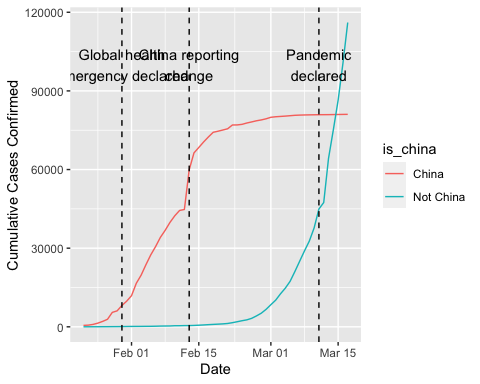
# Draw a line plot of cumulative cases vs. date, grouped and colored by is\_china  
  
# Define aesthetics within the line geom  
plt\_cum\_china\_vs\_world <- ggplot(confirmed\_cases\_with\_china) +  
 geom\_line(aes(date, cum\_cases, group = is\_china, color = is\_china)) +  
 xlab("Date") +  
 ylab("Cumulative Cases Confirmed")  
  
# See the plot  
plt\_cum\_china\_vs\_world



We can see that China contained a great volume of the casesIn February, the majority of cases were in China. To understand the significance of the time we will enter the landmark dates that indicate the escalation of COVID into a pandemic.

With the World Health Organisation declaring a global emergency on 30th January 2020. We have China reporting a change in the number of cases on the 13th February 2020 and then a pandemic declared on the 11th March 2020.

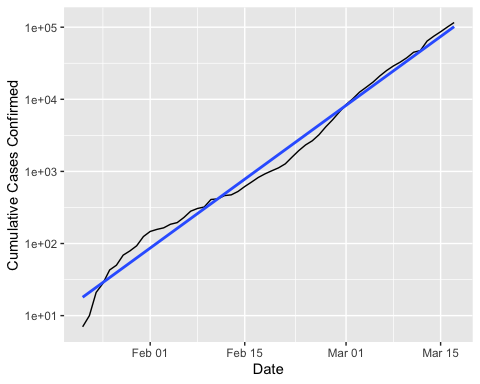
#add the event dates and the name of the event into the plot  
events <- tribble(  
 ~ date, ~ event,  
 "2020-01-30", "Global health\nemergency declared",  
 "2020-03-11", "Pandemic\ndeclared",  
 "2020-02-13", "China reporting\nchange"  
) %>%  
 mutate(date = as.Date(date))  
  
# Using events, add vertical dashed lines with an xintercept at date  
# and text at date, labeled by event, and at 100000 on the y-axis  
plt\_cum\_china\_vs\_world +  
 geom\_vline(aes(xintercept = date), data = events, linetype = "dashed") +  
 geom\_text(aes( x=date, label = event), data = events, y = 1e5)



A logarithmic scale will be introduced on the y axis to explore the changes in number of cases expected in the future as shown below with a trend line. We can use the trend line to predict how the number of cases worldwide, other than China, will change in the future.

# Modify the plot to use a logarithmic scale on the y-axis  
  
# Filter confirmed\_cases\_with\_china for not China  
no\_china <- confirmed\_cases\_with\_china %>%  
 filter(is\_china == "Not China")  
  
# Using not\_china, draw a line plot cum\_cases vs. date  
  
# Add a smooth trend line using linear regression, no error bars  
plt\_worldwide\_log\_trend<- ggplot(no\_china, aes(date, cum\_cases)) +  
 geom\_line() +  
 geom\_smooth(method = "lm", se = FALSE) +  
 xlab("Date") +  
 ylab("Cumulative Cases Confirmed") +  
 scale\_y\_log10()  
  
plt\_worldwide\_log\_trend

## `geom\_smooth()` using formula 'y ~ x'



We can see that the trend line is a good fit but indicates a increasing trend line and predict a future increase in cases worldwide.

We will now explore the top seven countries worst effected by COVID-19.

# read the data set for each country  
confirmed\_cases\_by\_country <- read\_csv("Downloads/Visualizing COVID-19/datasets/confirmed\_cases\_by\_country.csv")

##   
## ── Column specification ───────────────────────────────────────────────  
## cols(  
## country = col\_character(),  
## province = col\_character(),  
## date = col\_date(format = ""),  
## cases = col\_double(),  
## cum\_cases = col\_double()  
## )

# Group by country, summarize to calculate total cases, find the top 7  
top\_countries\_by\_cases <- confirmed\_cases\_by\_country %>%  
 group\_by(country) %>%  
 summarize(total\_cases = max(cum\_cases)) %>%  
top\_n(7, total\_cases)

## `summarise()` ungrouping output (override with `.groups` argument)

# See the result  
top\_countries\_by\_cases

## # A tibble: 7 x 2  
## country total\_cases  
## <chr> <dbl>  
## 1 France 7699  
## 2 Germany 9257  
## 3 Iran 16169  
## 4 Italy 31506  
## 5 Korea, South 8320  
## 6 Spain 11748  
## 7 US 6421

We can visualise the change in the number of cases for the countries that have been effected the most.

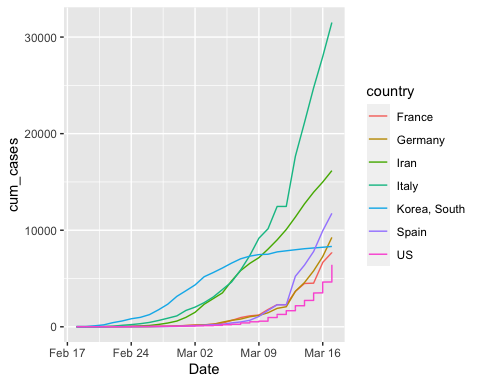
# Read in the dataset from datasets/confirmed\_cases\_top7\_outside\_china.csv  
top7\_countries\_outside\_china <- read\_csv("Downloads/Visualizing COVID-19/datasets/confirmed\_cases\_top7\_outside\_china.csv")

##   
## ── Column specification ───────────────────────────────────────────────  
## cols(  
## country = col\_character(),  
## date = col\_date(format = ""),  
## cum\_cases = col\_double()  
## )

# Glimpse the contents of confirmed\_cases\_top7\_outside\_china  
glimpse(top7\_countries\_outside\_china)

## Rows: 2,030  
## Columns: 3  
## $ country <chr> "Germany", "Iran", "Italy", "Korea, South", "Spain", "US", …  
## $ date <date> 2020-02-18, 2020-02-18, 2020-02-18, 2020-02-18, 2020-02-18…  
## $ cum\_cases <dbl> 16, 0, 3, 31, 2, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13…

# Using confirmed\_cases\_top7\_outside\_china, draw a line plot of  
# cum\_cases vs. date, grouped and colored by country  
ggplot(top7\_countries\_outside\_china, aes(date, cum\_cases, color = country, group = country)) +  
 geom\_line() +  
 xlab("Date")



ylab("Cumulative Cases Confirmed")

## $y  
## [1] "Cumulative Cases Confirmed"  
##   
## attr(,"class")  
## [1] "labels"

We can see that Italy was the country that was the most seriously effected by the virus aside from China by around 15,000 cases.