Covid19_VS_VisualAnalytics_Timeseries_Data

V.Srinivas

06/05/2020

```r
# Load the required packages
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
```

```r
library(tidyverse)
```

```
## -- Attaching packages ---------- tidyverse 1.3.0 --

## v tibble  2.1.3      v purrr   0.3.3
## v tidyr   1.0.2      v stringr 1.4.0
## v readr   1.3.1      v forcats 0.5.0

## -- Conflicts ------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(caret)
```

```
## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift
```

```r
library(scales)
```

```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:purrr':
##
##     discard
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

```r
# Read datasets/confirmed_cases_worldwide.csv into confirmed_cases_worldwide
covid_ds <- data.table::fread('https://raw.githubusercontent.com/RamiKrispin/coronavirus/master/csv/core
  mutate (date=as.Date(date)) %>%
  select(-lat, -long)
```

```r
# View info about the dataset
glimpse(covid_ds)
```

```
## Observations: 94,200
## Variables: 4
## $ date    <date> 2020-01-22, 2020-01-23, 2020-01-24, 2020-01-25, 2020-01-26...
## $ country <chr> "Afghanistan", "Afghanistan", "Afghanistan", "Afghanistan",...
## $ type    <chr> "confirmed", "confirmed", "confirmed", "confirmed", "confir...
## $ cases   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
```

```r
str(covid_ds)
```

```
## 'data.frame':    94200 obs. of  4 variables:
##  $ date   : Date, format: "2020-01-22" "2020-01-23" ...
##  $ country: chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ type   : chr  "confirmed" "confirmed" "confirmed" "confirmed" ...
##  $ cases  : int  0 0 0 0 0 0 0 0 0 0 ...
```

```r
tail(covid_ds,20)
```

```
##               date country       type cases
## 94181 2020-05-01    China recovered     1
## 94182 2020-05-02    China recovered     0
## 94183 2020-05-03    China recovered     0
## 94184 2020-05-04    China recovered     2
## 94185 2020-05-05    China recovered     0
## 94186 2020-05-06    China recovered     0
## 94187 2020-05-07    China recovered     0
## 94188 2020-05-08    China recovered     0
## 94189 2020-05-09    China recovered     0
## 94190 2020-05-10    China recovered     0
## 94191 2020-05-11    China recovered     0
```

```
## 94192 2020-05-12    China recovered     0
## 94193 2020-05-13    China recovered     0
## 94194 2020-05-14    China recovered     0
## 94195 2020-05-15    China recovered     0
## 94196 2020-05-16    China recovered     0
## 94197 2020-05-17    China recovered     0
## 94198 2020-05-18    China recovered     0
## 94199 2020-05-19    China recovered     0
## 94200 2020-05-20    China recovered     0
```

```r
dim(covid_ds)
```

```
## [1] 94200     4
```

```r
covid_ds$type[covid_ds$type == 'death'] <-'dead'

# Convert negative cases to positive assuming it is typo error
#covid_ds$cases = abs(covid_ds$cases)
#covid_ds %>% filter(cases < 0) %>% select(country,cases, type)
```

```r
# compute daily cases for all cuntries
daily_cases <- covid_ds %>%
  group_by(date, country, type) %>%
  summarize (daily_cases = sum(cases)) %>%
  ungroup()
```

```r
# Totals so far worldwide
totals <- covid_ds %>%
  group_by(type) %>%
  summarize(total=sum(cases))

totals %>%
  ggplot(aes(type, total, fill=type)) +
  geom_col() +
  scale_y_continuous(labels=comma)+
  geom_text(aes(label=total), vjust=-.5) +
  labs(subtitle="Total cases world wide")
```

Total cases world wide
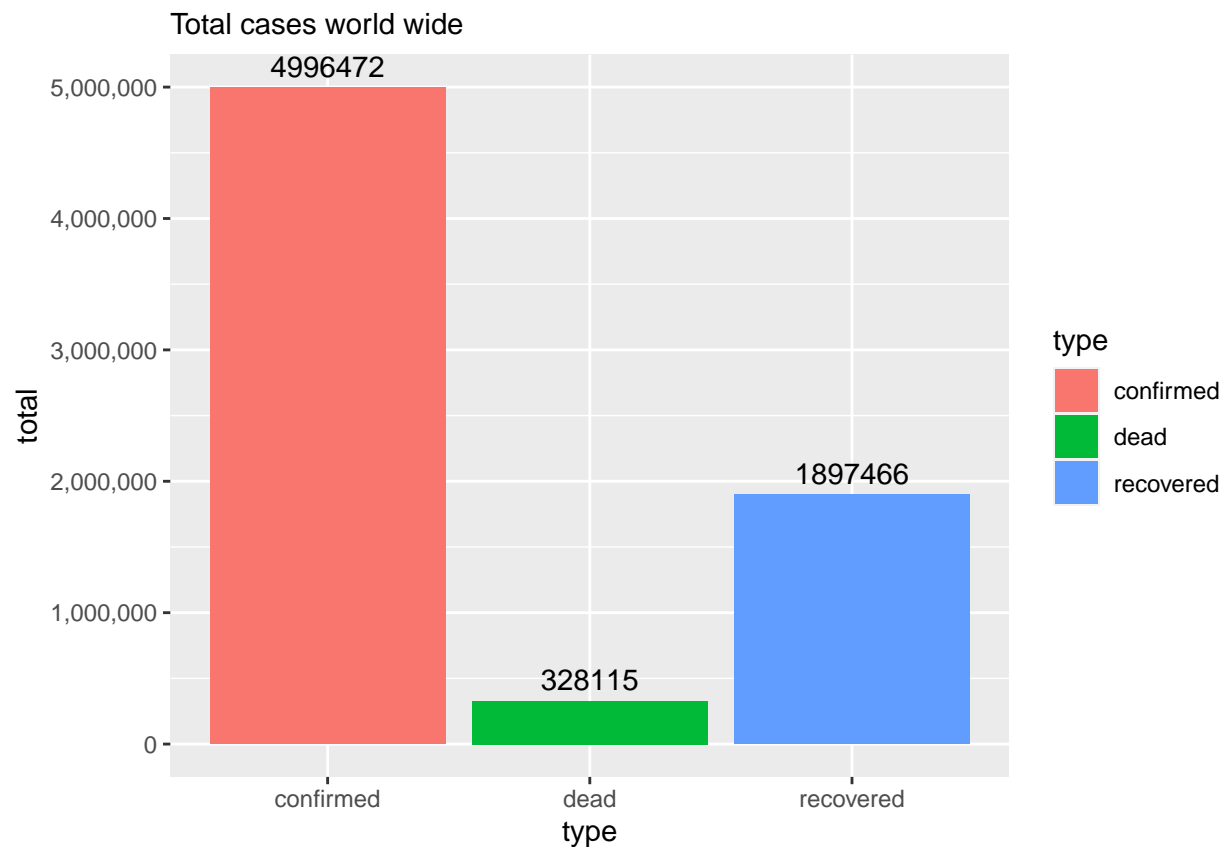
```r
# Total daily cases across the globe
tot_daily_all_countries <- covid_ds %>%
  group_by(date, type) %>%
  mutate (total_daily_cases = sum(cases)) %>%
  arrange(desc(total_daily_cases)) %>%
  select(date, type, total_daily_cases)

tot_daily_all_countries %>%
  filter(total_daily_cases == max(total_daily_cases)) %>%
  head(1)
```
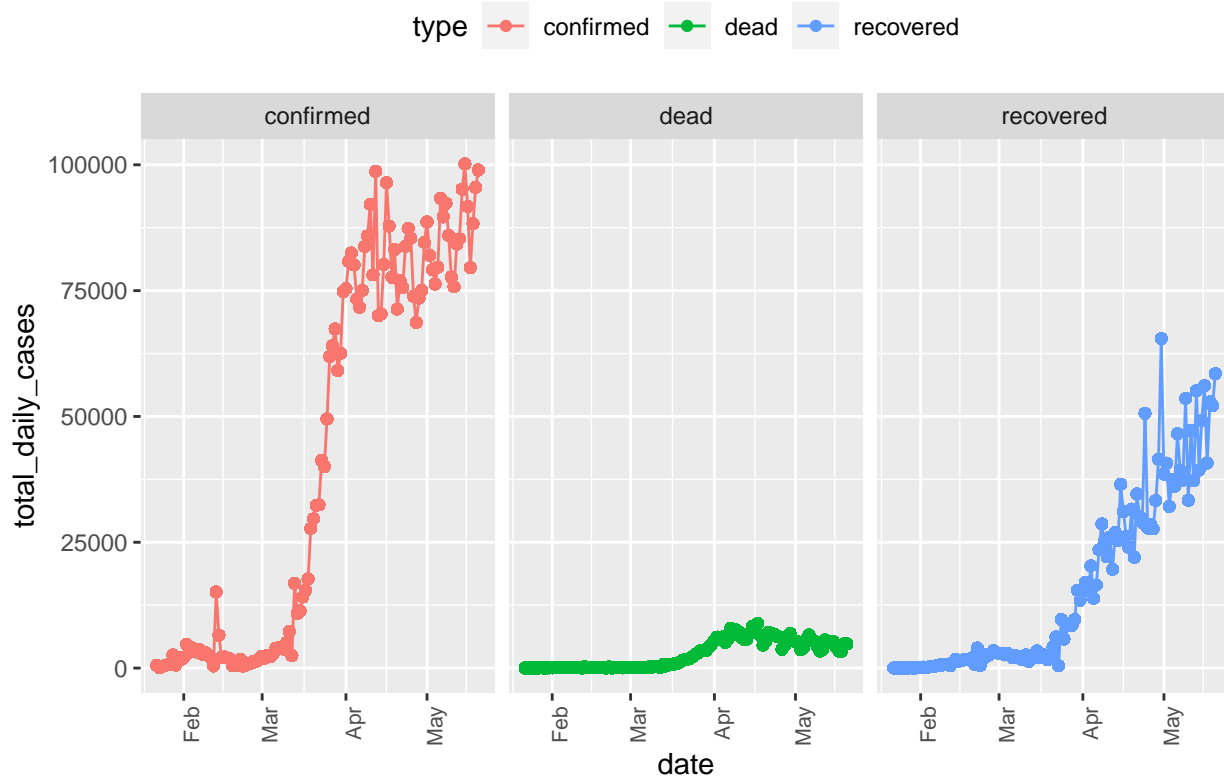
```
## # A tibble: 1 x 3
## # Groups:   date, type [1]
##   date       type       total_daily_cases
##   <date>     <chr>                  <int>
## 1 2020-05-15 confirmed             100184
```

```r
tot_daily_all_countries %>% ggplot(aes(date,total_daily_cases, col=type)) +
  geom_point()+
  geom_line() +
  facet_wrap(~type)+
  theme(axis.text.x = element_text(
    angle = 90,
    size = 8,
    hjust = 1
```

```
),
legend.position = "top") +
labs(subtitle=" Total daily cases across all nations")
```

Total daily cases across all nations


```
# Confirmed cumulative cases country_wise
confirmed <- daily_cases %>%
  filter(type=='confirmed') %>%
  select(-type) %>%
  group_by(country) %>%
  mutate(cum_c_cases= cumsum(daily_cases)) %>%
  select(-daily_cases) %>%
  ungroup()
```

```
# Recovered cumulative cases
recovered <- daily_cases %>%
  filter(type=='recovered') %>%
  select(-type) %>%
  group_by(country) %>%
  mutate(cum_r_cases= cumsum(daily_cases)) %>%
  select(-daily_cases) %>%
  ungroup()
```

```
# dead cumulative cases
dead <- daily_cases %>%
```

```r
  filter(type=='dead') %>%
  select(-type) %>%
  group_by(country) %>%
  mutate(cum_d_cases= cumsum(daily_cases)) %>%
  select(-daily_cases) %>%
  ungroup()
```

```r
# Combine cumulative cases of all 3 categories
all_cases <- confirmed %>%
  inner_join(recovered, by=c('date', 'country')) %>%
  inner_join(dead, by=c('date', 'country')) %>%
  mutate(country=as.factor(country))
```

```r
# All cumultaive cases, all countries on all dates with all 3 categories - Top 50
all_cases %>% arrange(desc(date,cum_c_cases)) %>% head(50)
```

```
## # A tibble: 50 x 5
##    date       country            cum_c_cases cum_r_cases cum_d_cases
##    <date>     <fct>                    <int>       <int>       <int>
##  1 2020-05-20 Afghanistan               8145         930         187
##  2 2020-05-20 Albania                    964         758          31
##  3 2020-05-20 Algeria                   7542        3968         568
##  4 2020-05-20 Andorra                    762         639          51
##  5 2020-05-20 Angola                      52          17           3
##  6 2020-05-20 Antigua and Barbuda         25          19           3
##  7 2020-05-20 Argentina                 9283        2933         403
##  8 2020-05-20 Armenia                   5271        2419          67
##  9 2020-05-20 Australia                 7081        6470         100
## 10 2020-05-20 Austria                  16353       14882         633
## # ... with 40 more rows
```

```r
# All cumulative cases of all countries on all dates
countries_cases <- all_cases %>%
  rename(confirmed=cum_c_cases, recovered=cum_r_cases, dead=cum_d_cases) %>%
  gather(type, cum_cases, confirmed:dead) %>%
  mutate(type=as.factor(type)) %>% arrange(desc(date))
```

```r
# Select, major hit nations (top 30)
top_20_countries <- covid_ds %>% group_by(country) %>% summarize(total_cases=sum(cases)) %>% arrange(des
select_countries <- top_20_countries$country
```
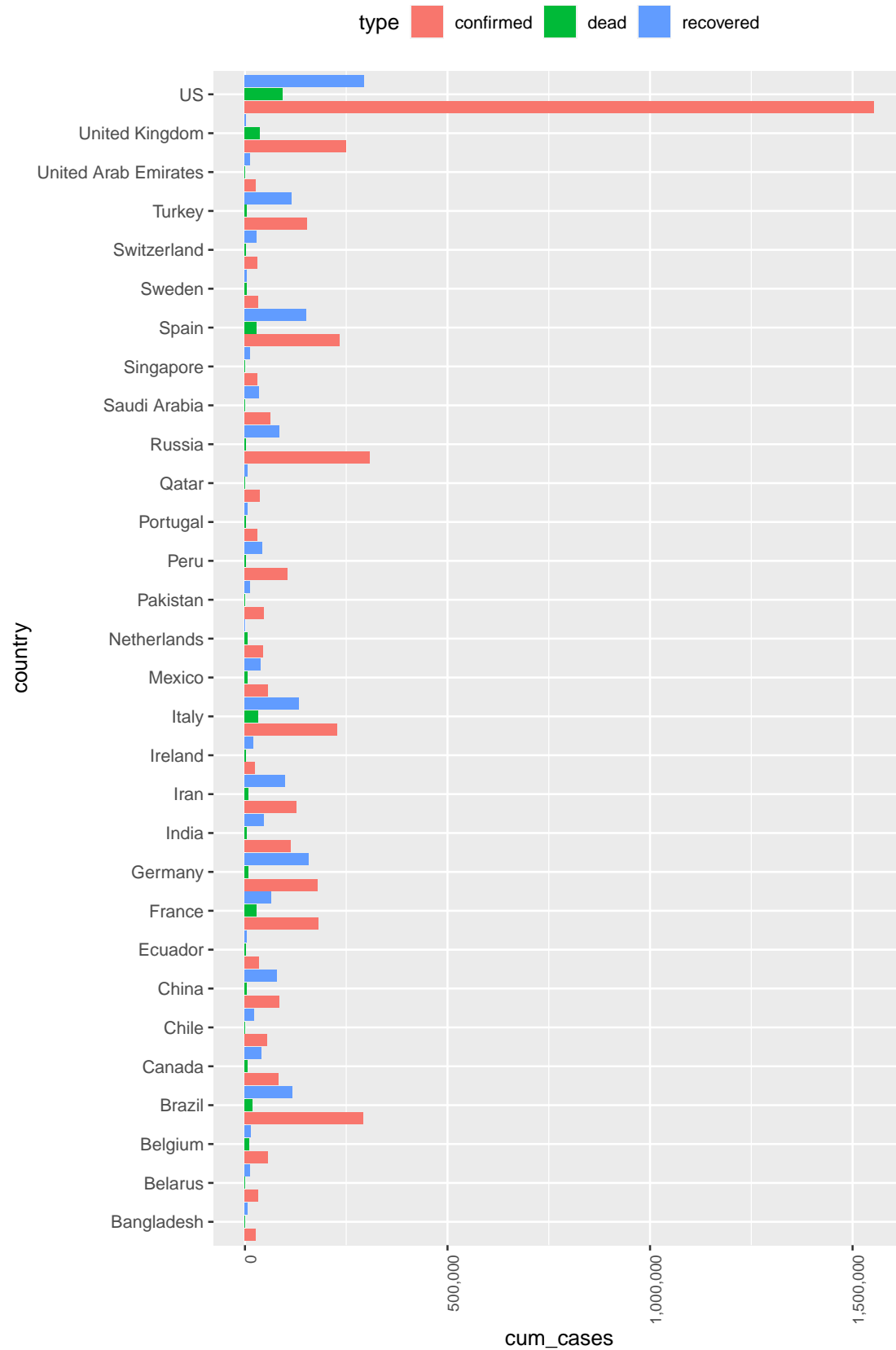
```r
countries_cases %>%
  filter(country %in% select_countries) %>%
  ggplot(aes(country, cum_cases, fill=type)) +
  geom_col( position = position_dodge(1)) +
  scale_y_continuous(labels = comma)+
  coord_flip() +
  theme(axis.text.x = element_text(
    angle = 90,
    size = 8,
    hjust = 1
  ),
```

```
    legend.position = "top") +
labs(subtitle="Total numbers across select nations")
```

Total numbers across select nations

```r
# Cumulative cases by date and type
countries_cases_all <- countries_cases %>%
  group_by(date, type) %>%
  mutate(tot_cum_cases = sum(cum_cases)) %>%
  arrange(desc(tot_cum_cases))

# Cumulative cases by date and type - Top 50
cum_cases_all <- countries_cases_all %>%
  group_by(date, type) %>%
  top_n(1, wt = cum_cases) %>%
  ungroup()
```

```r
cum_cases_all %>%
  select(date, type, tot_cum_cases) %>%
  arrange(desc(date, type)) %>% head(50)
```

```
## # A tibble: 50 x 3
##    date       type        tot_cum_cases
##    <date>     <fct>               <int>
##  1 2020-05-20 confirmed         4996472
##  2 2020-05-20 recovered         1897466
##  3 2020-05-20 dead               328115
##  4 2020-05-19 confirmed         4897492
##  5 2020-05-19 recovered         1838995
##  6 2020-05-19 dead               323285
##  7 2020-05-18 confirmed         4801943
##  8 2020-05-18 recovered         1786875
##  9 2020-05-18 dead               318481
## 10 2020-05-17 confirmed         4713620
## # ... with 40 more rows
```

```r
cum_cases_all %>%
  ggplot(aes(date, tot_cum_cases, col=type)) +
  geom_line(size=1) +
  scale_y_continuous(label=comma) +
  theme(axis.text.x = element_text(
    angle = 90,
    size = 8,
    hjust = 1
  ),
  legend.position = "top") +
  labs(subtitle=" Cumulative number progression across all nations")
```

Cumulative number progression across all nations
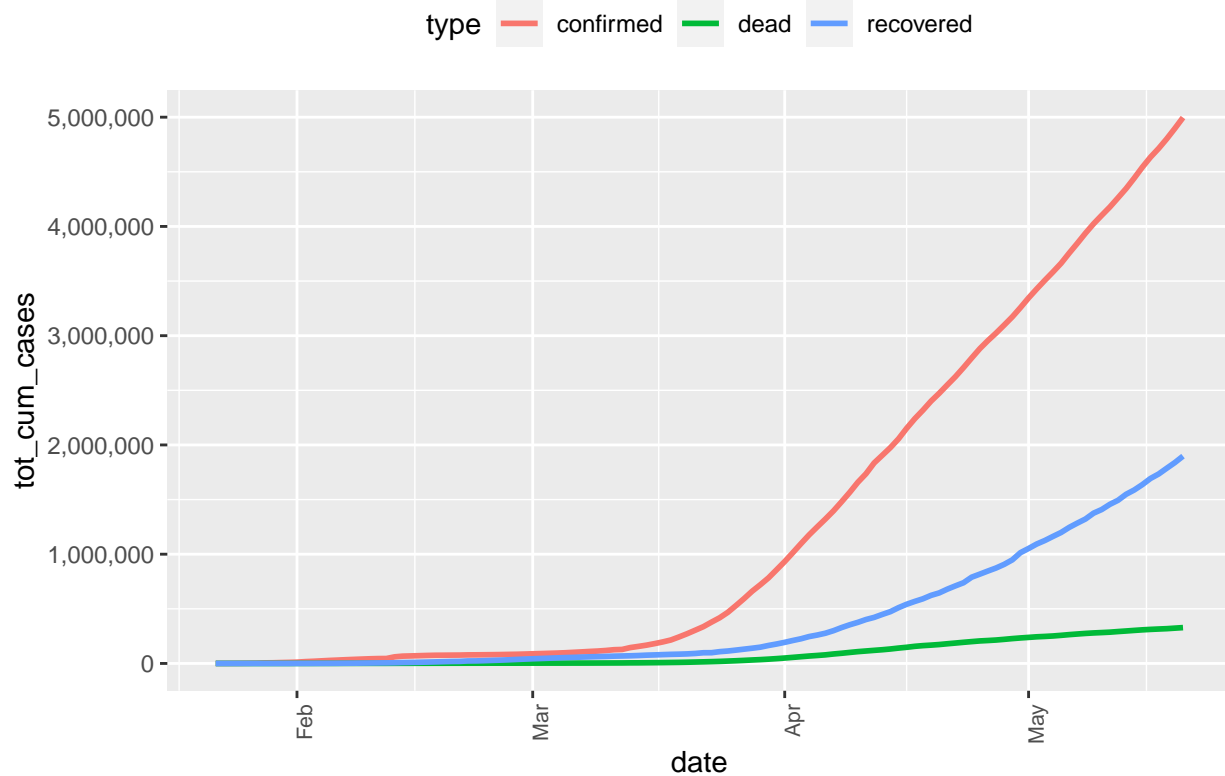

```r
# Cases in select, major hot nations
select_cases <- countries_cases %>%
  filter(country %in% select_countries)

select_cases %>%
  ggplot(aes(date, cum_cases, col=country)) +
  geom_line(size=1) +
  scale_y_continuous(label=comma) +
  facet_grid(~type) +
  theme(axis.text.x = element_text(
    angle = 90,
    size = 8,
    hjust = 1
  ),
  legend.position = "top") +
  labs(subtitle=" Cumulative number progression across major hit nations")
```

Cumulative number progression across major hit nations
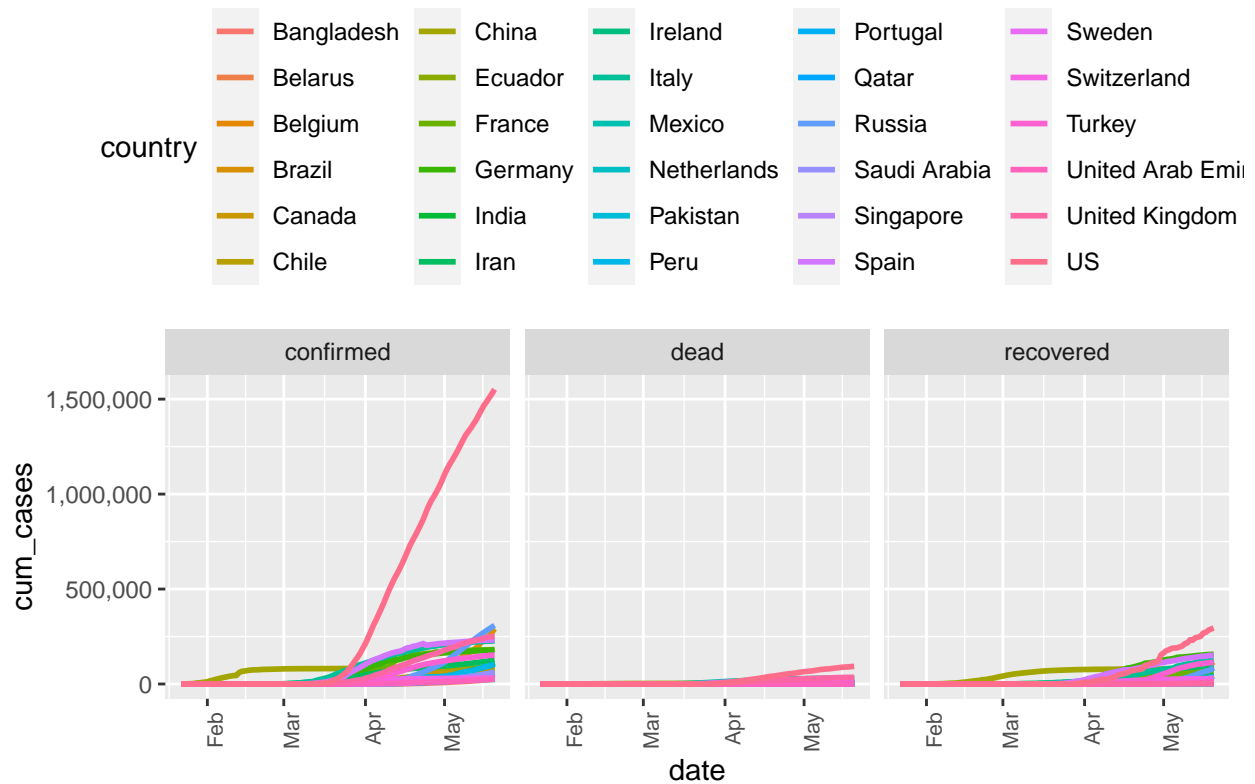

```r
# Compare China and USA
china_vs_usa <- countries_cases %>%
  filter(country %in% c('China', 'US'))

china_vs_usa %>%
  ggplot(aes(date, cum_cases, col=country)) +
  geom_line(size=1) +
  scale_y_continuous(label=comma) +
  facet_grid(~type)  +
  theme(axis.text.x = element_text(
    angle = 90,
    size = 8,
    hjust = 1
  ),
  legend.position = "top") +
  labs(subtitle=" China vs USA cumulative progression")
```

China vs USA cumulative progression
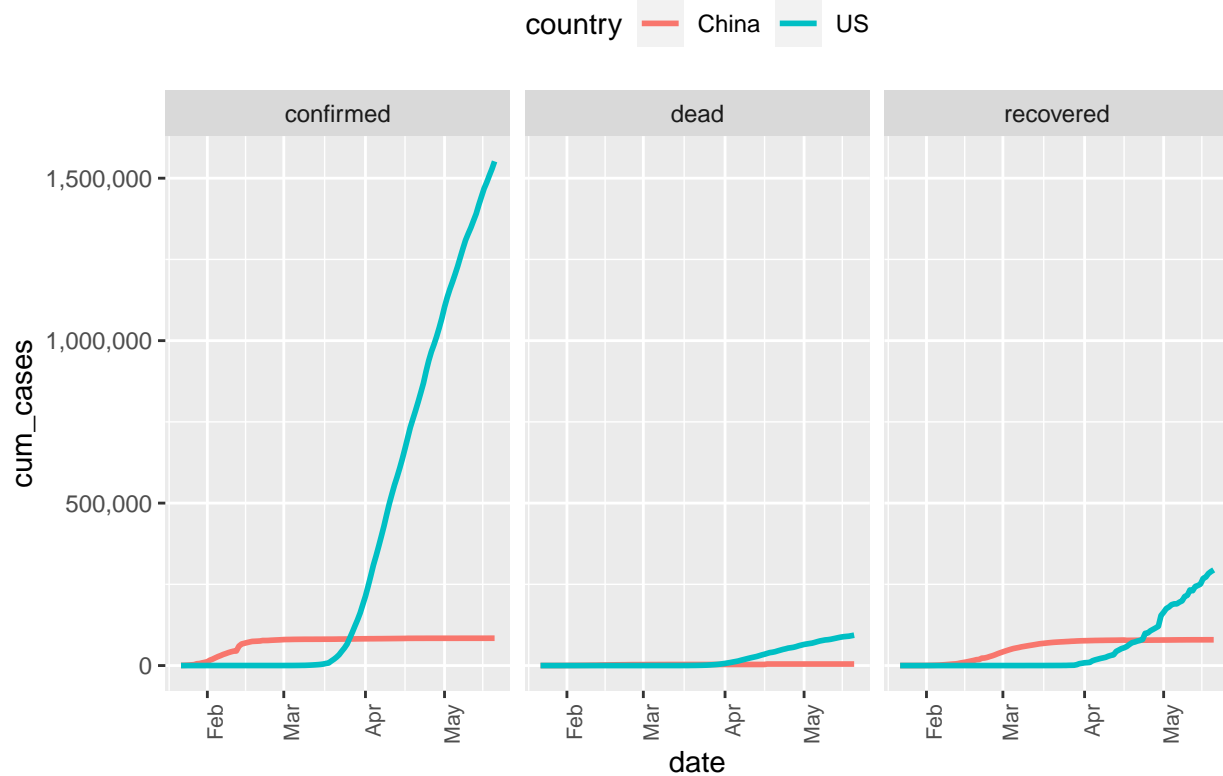

```r
# non-China and non_USA cases (non-extreme case countries)
non_china_usa_cases <- countries_cases %>%
  filter(country %in% select_countries & country !='China' & country != 'US')

non_china_usa_cases %>%
  ggplot(aes(date, cum_cases, col=country)) +
  geom_line(size=1) +
  facet_grid(~type)  +
  theme(axis.text.x = element_text(
    angle = 90,
    size = 8,
    hjust = 1
  ),
  legend.position = "top") +
  labs(subtitle="Non_china, Non_USA cumulative number progression")
```

Non_china, Non_USA cumulative number progression
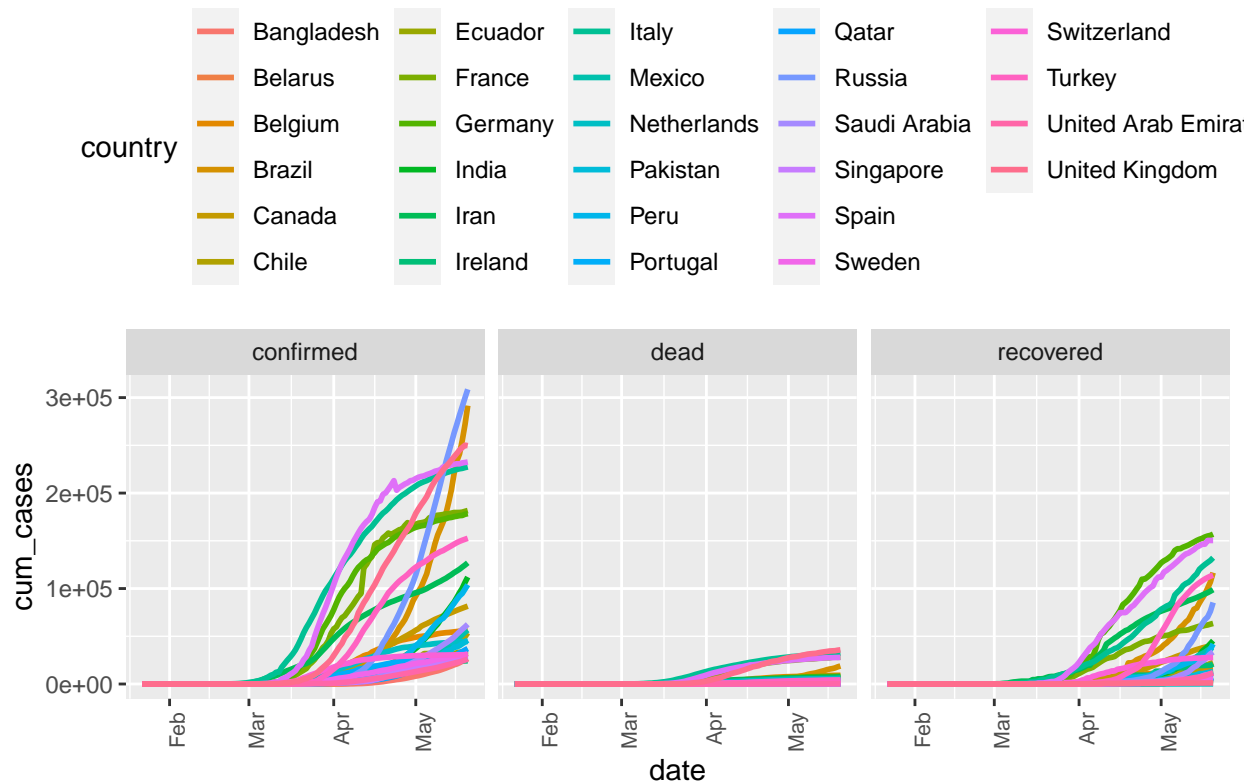

```r
# Cases in india
india_cases <- countries_cases %>%
  filter(country %in% c('India')) %>%
  select(-country)

india_cases %>% arrange(desc(date)) %>% head(50)
```

```
## # A tibble: 50 x 3
##     date       type       cum_cases
##     <date>     <fct>         <int>
##  1 2020-05-20 confirmed    112028
##  2 2020-05-20 recovered     45422
##  3 2020-05-20 dead           3434
##  4 2020-05-19 confirmed    106475
##  5 2020-05-19 recovered     42309
##  6 2020-05-19 dead           3302
##  7 2020-05-18 confirmed    100328
##  8 2020-05-18 recovered     39233
##  9 2020-05-18 dead           3156
## 10 2020-05-17 confirmed     95698
## # ... with 40 more rows
```

```r
india_cases %>% group_by(type) %>% top_n(1) %>% rename(total_cases = cum_cases) %>%
  ggplot(aes(type, total_cases, fill=type)) +
  geom_col() +
```
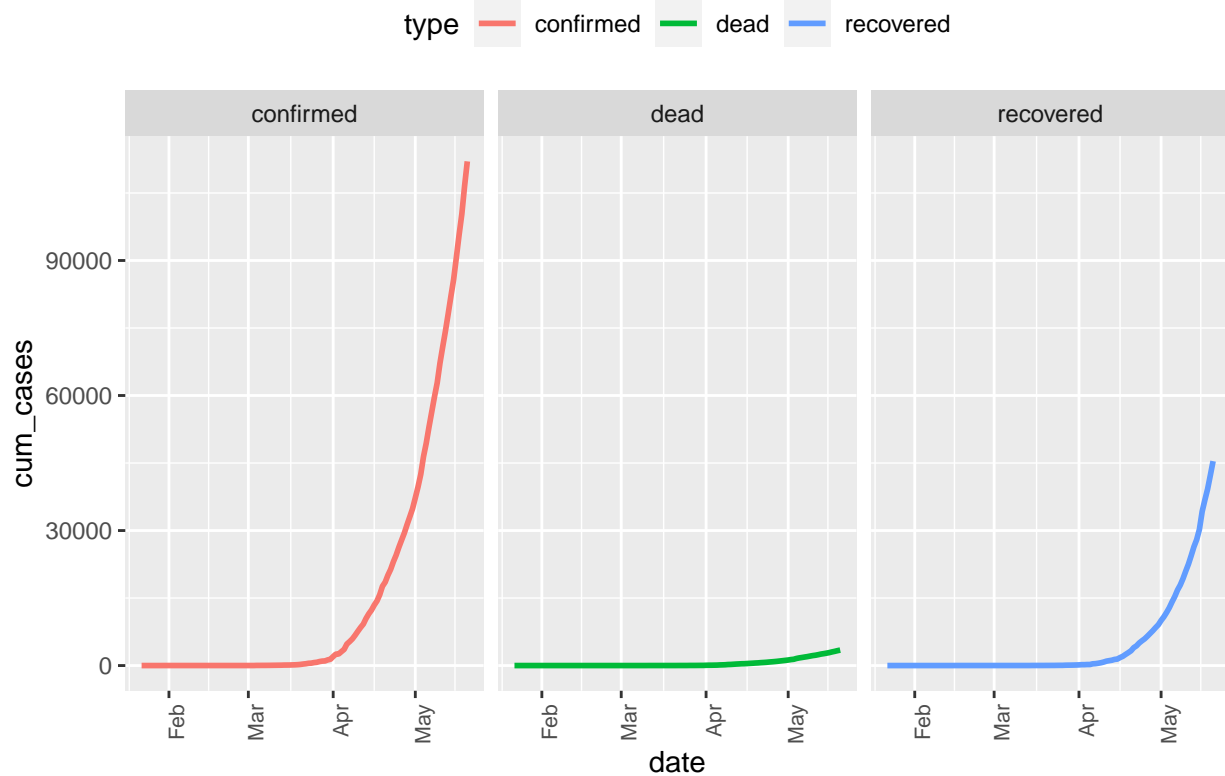
```
  geom_text(aes(label = total_cases), vjust = -0.5) +
  labs(subtitle=" Total cases in India")
```

Selecting by cum_cases
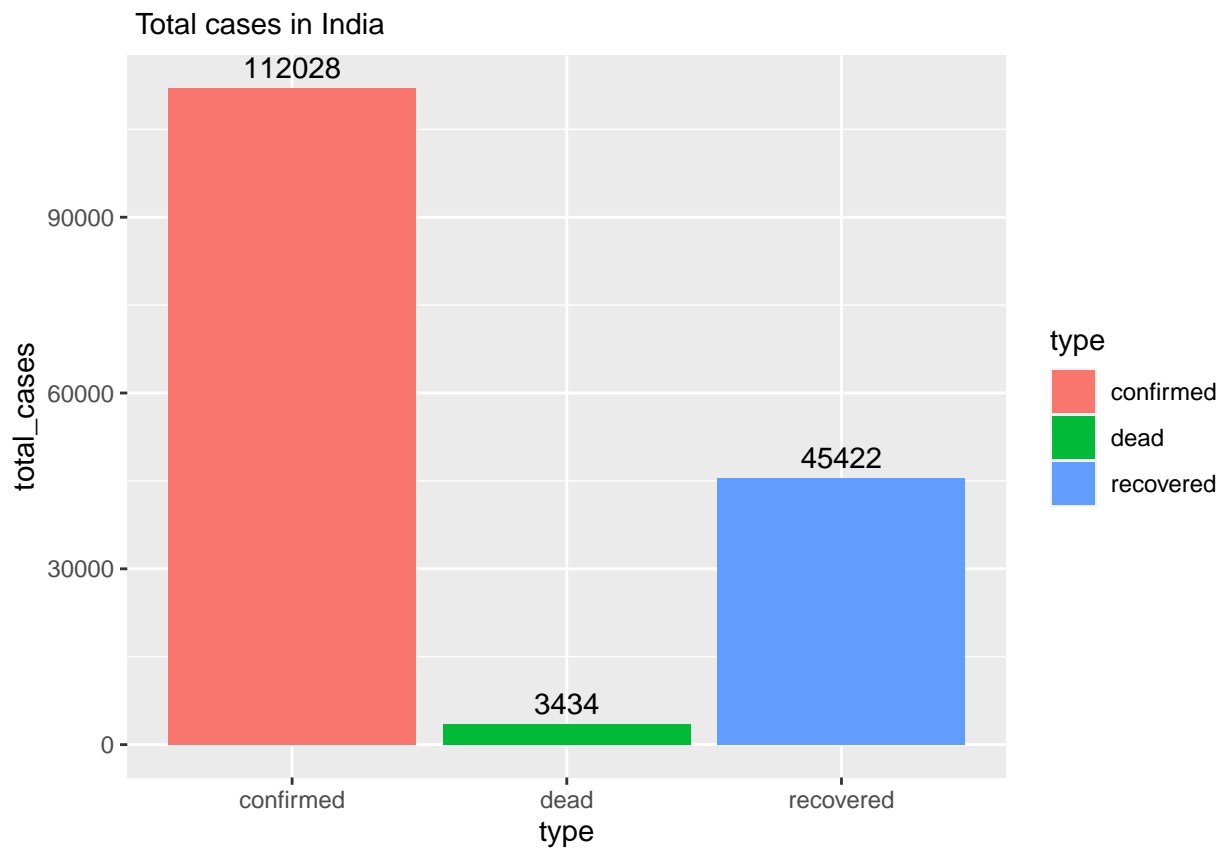

```
india_cases %>%
  ggplot(aes(date, cum_cases, col=type)) +
  geom_line(size=1) +
  facet_grid(~ type) +
  theme(axis.text.x = element_text(
    angle = 90,
    size = 8,
    hjust = 1
  ),
  legend.position = "top") +
  labs(subtitle=" Cumulative case progression in India")
```

Cumulative case progression in India


```r
# Total daily cases in India
tot_d_india_cases <- covid_ds %>%
  filter(country=='India') %>%
  group_by(date, type) %>% mutate (total_daily_cases = sum(cases)) %>%
  arrange(desc(total_daily_cases))%>%
  select(date, type, total_daily_cases)

tot_d_india_cases %>% ggplot(aes(date,total_daily_cases, col=type)) +
  geom_point()+
  geom_line() +
  facet_wrap(~type)+
  theme(axis.text.x = element_text(
    angle = 90,
    size = 8,
    hjust = 1
  ),
  legend.position = "top") +
  labs(subtitle=" Total daily cases in India")
```

Total daily cases in India
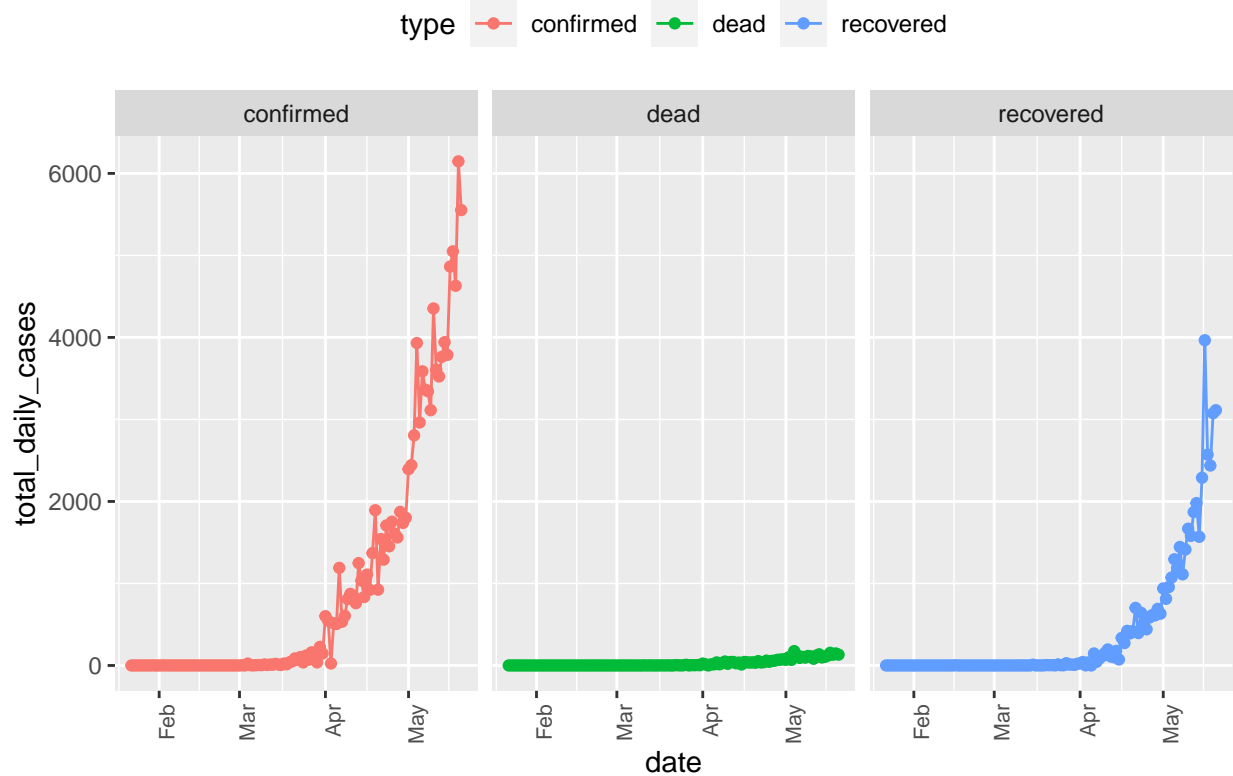

```
# Cases in Oman & UAE
oman_vs_uae <- countries_cases %>%
  filter(country %in% c('Oman', 'United Arab Emirates'))

oman_vs_uae %>% group_by(country,type) %>% top_n(1) %>% rename(total_cases = cum_cases) %>%
  ggplot(aes(country, total_cases, fill=type)) +
  geom_col(position = position_dodge(1)) +
  geom_text(aes(label = total_cases),position = position_dodge(1), vjust = -0.5) +
  labs(subtitle=" Total cases in Oman & UAE")
```
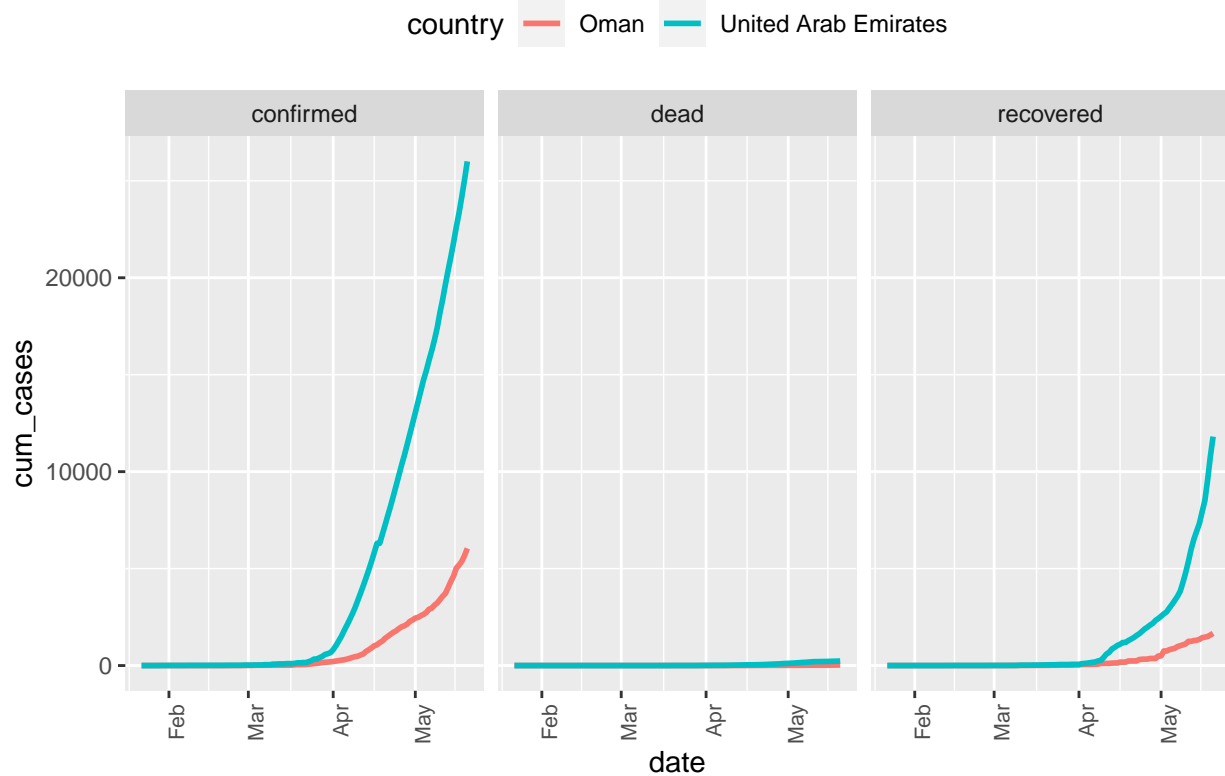
```
## Selecting by cum_cases
```

Total cases in Oman & UAE
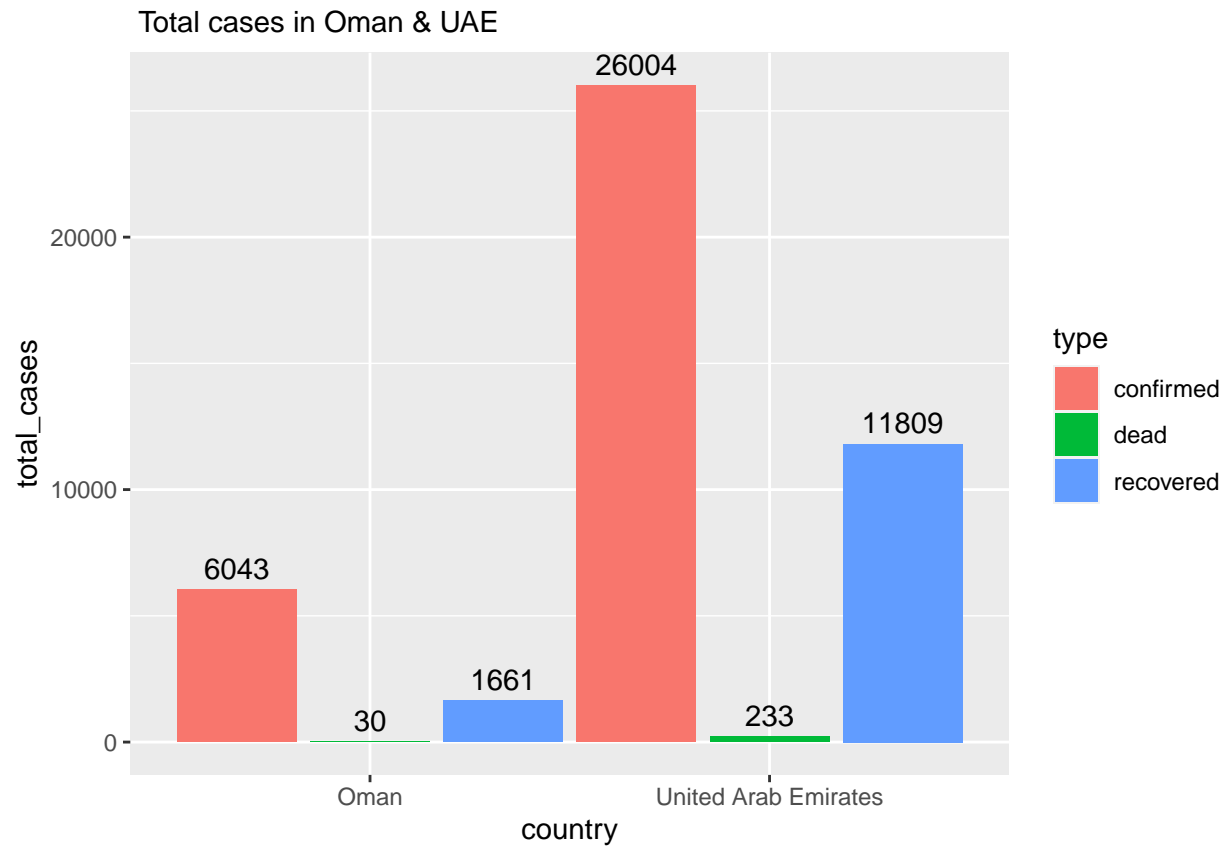

```r
oman_vs_uae %>% ggplot(aes(date, cum_cases, col=country)) +
  geom_line(size=1) +
  facet_grid(~type)  +
  theme(axis.text.x = element_text(
    angle = 90,
    size = 8,
    hjust = 1
  ),
  legend.position = "top") +
  labs(subtitle=" Oman vs UAE cumulative progression")
```

Oman vs UAE cumulative progression


```r
# Total daily cases across the globe
oman_vs_uae_daily <- covid_ds %>%
  filter(country=='Oman' | country == "United Arab Emirates") %>%
  group_by(date,country, type) %>% mutate (total_daily_cases = sum(cases)) %>%
  arrange(desc(total_daily_cases))%>%
  select(date, country, type, total_daily_cases)
```

```r
oman_vs_uae_daily %>% ggplot(aes(date,total_daily_cases, col=country)) +
  geom_point()+
  geom_line() +
  facet_wrap(~type)+
  theme(axis.text.x = element_text(
    angle = 90,
    size = 8,
    hjust = 1
  ),
  legend.position = "top") +
  labs(subtitle=" Total daily cases in Oman & UAE")
```

Total daily cases in Oman & UAE
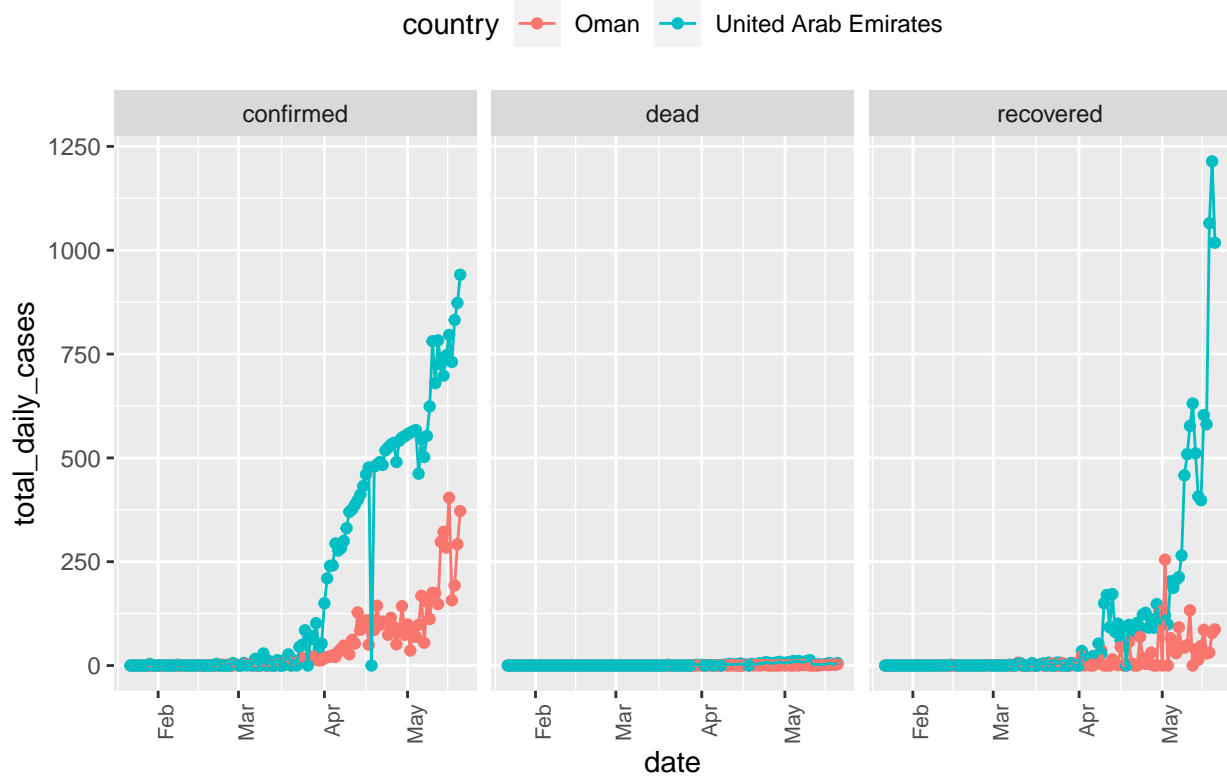

```r
# Number of people dead for every 100 people recovered in the top 30 hits countries
percents <- covid_ds %>%
  select(country, type, cases) %>%
  group_by(country, type) %>%
  summarize(tot_cases= sum(cases)) %>%
  spread(type, tot_cases) %>%
  summarize(recovery_percent= 100*sum(recovered)/sum(confirmed),
            dead_percent= 100*sum(dead)/sum(confirmed),
            dead_to_recovery= 100*sum(dead)/sum(recovered)) %>%
  gather(type, percentage, recovery_percent:dead_to_recovery)
```

```r
# Plot Number of people dead for every 100 people recovered in select countries
percents %>%
  filter(country %in% select_countries) %>%
  ggplot(aes(country, percentage, fill=country)) +
  geom_col()+
  facet_grid(type~.)+
  theme(axis.text.x = element_text(
    angle = 90,
    size = 8,
    hjust = 1
  ),
  legend.position = "none") +
  labs(subtitle="% of dead, recoveries and dead to recoveries")
```

% of dead, recoveries and dead to recoveries
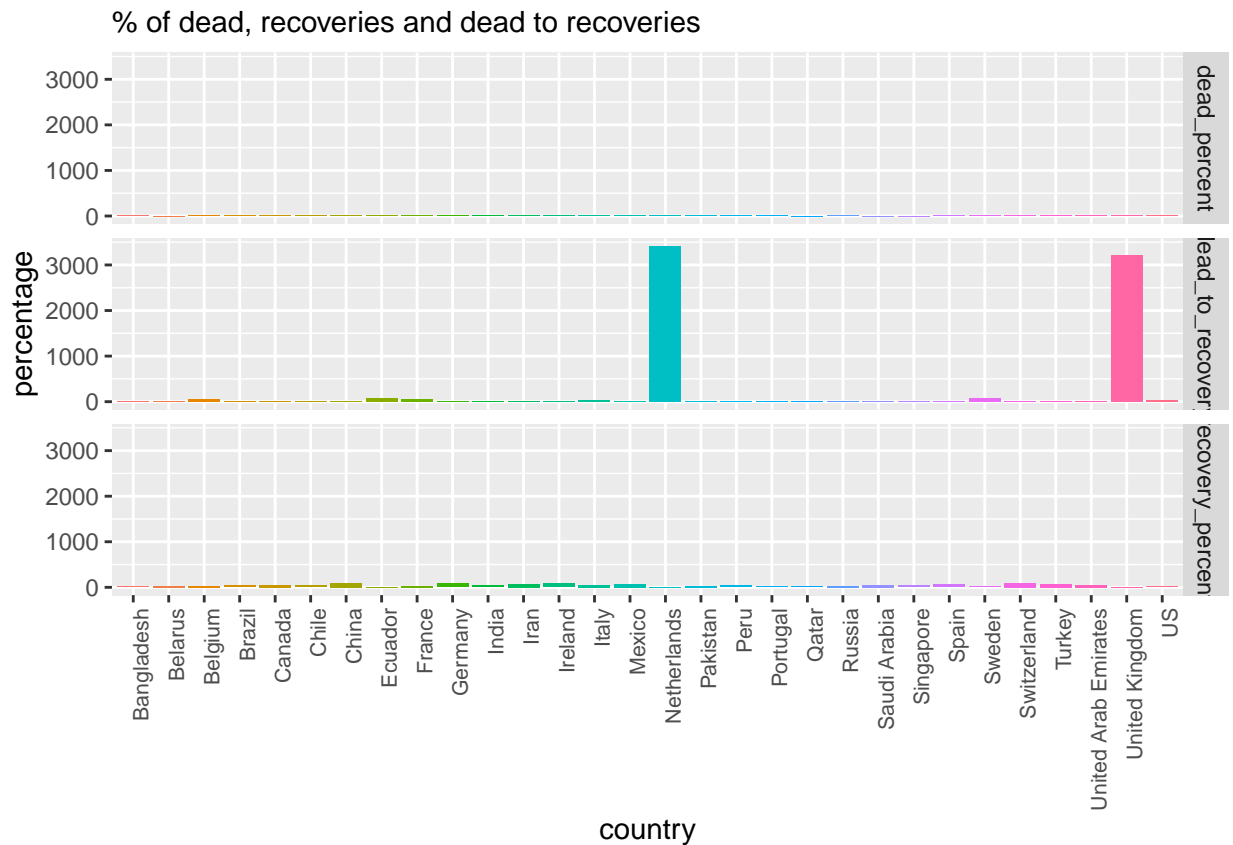

```r
# Plot Number of people dead for every 100 people recovered in select countries, excluding UK and Nethe
percents %>%
  filter(country %in% select_countries & country != 'United Kingdom' & country != 'Netherlands') %>%
  ggplot(aes(country, percentage, fill=country)) +
  geom_col()+
  facet_grid(type~.)+
  theme(axis.text.x = element_text(
    angle = 90,
    size = 8,
    hjust = 1
  ),
  legend.position = "none") +
  labs(subtitle="% of dead, recoveries and dead to recoveries")
```

% of dead, recoveries and dead to recoveries