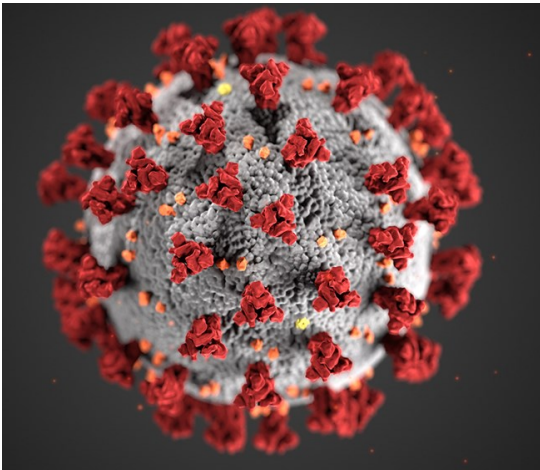


1. From epidemic to pandemic



In December 2019, COVID-19 coronavirus was first identified in the Wuhan region of China. By March 11, 2020, the World Health Organization (WHO) categorized the COVID-19 outbreak as a pandemic. A lot has happened in the months in between with major outbreaks in Iran, South Korea, and Italy.

We know that COVID-19 spreads through respiratory droplets, such as through coughing, sneezing, or speaking. But, how quickly did the virus spread across the globe? And, can we see any effect from country-wide policies, like shutdowns and quarantines?

Fortunately, organizations around the world have been collecting data so that governments can monitor and learn from this pandemic. Notably, the Johns Hopkins University Center for Systems Science and Engineering created a [publicly available data repository](https://github.com/RamiKrispin/coronavirus) (<https://github.com/RamiKrispin/coronavirus>) to consolidate this data from sources like the WHO, the Centers for Disease Control and Prevention (CDC), and the Ministry of Health from multiple countries.

In this notebook, you will visualize COVID-19 data from the first several weeks of the outbreak to see at what point this virus became a global pandemic.

Please note that information and data regarding COVID-19 is frequently being updated. The data used in this project was pulled on March 17, 2020, and should not be considered to be the most up to date data available.

In [3]:

```
# Load the readr, ggplot2, and dplyr packages
library(readr)
library(ggplot2)
library(dplyr)

# Read datasets/confirmed_cases_worldwide.csv into confirmed_cases_worldwide
confirmed_cases_worldwide <- read_csv("datasets/confirmed_cases_worldwide.csv")

# See the result
confirmed_cases_worldwide
```

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

— Column specification —

```
cols(  
  date = col_date(format = ""),  
  cum_cases = col_double()  
)
```

A spec_tbl_df: 56 × 2

date	cum_cases
<date>	<dbl>
2020-01-22	555
2020-01-23	653
2020-01-24	941
2020-01-25	1434
2020-01-26	2118
2020-01-27	2927
2020-01-28	5578
2020-01-29	6166
2020-01-30	8234
2020-01-31	9927
2020-02-01	12038
2020-02-02	16787
2020-02-03	19881
2020-02-04	23892
2020-02-05	27635
2020-02-06	30817
2020-02-07	34391
2020-02-08	37120
2020-02-09	40150
2020-02-10	42762
2020-02-11	44802
2020-02-12	45221
2020-02-13	60368
2020-02-14	66885
2020-02-15	69030
2020-02-16	71224
2020-02-17	73258
2020-02-18	75136
2020-02-19	75639
2020-02-20	76197
2020-02-21	76823
2020-02-22	78579
2020-02-23	78965
2020-02-24	79568
2020-02-25	80413
2020-02-26	81395

date	cum_cases
<date>	<dbl>
2020-02-27	82754
2020-02-28	84120
2020-02-29	86011
2020-03-01	88369
2020-03-02	90306
2020-03-03	92840
2020-03-04	95120
2020-03-05	97882
2020-03-06	101784
2020-03-07	105821
2020-03-08	109795
2020-03-09	113561
2020-03-10	118592
2020-03-11	125865
2020-03-12	128343
2020-03-13	145193
2020-03-14	156097
2020-03-15	167449
2020-03-16	181531
2020-03-17	197146

In [4]:

```

library(testthat)
library(IRkernel.testthat)

soln_confirmed_cases_worldwide <- read_csv("datasets/confirmed_cases_worldwide.csv")

run_tests({
  test_that("readr is loaded", {
    expect_true(
      "readr" %in% .packages(),
      info = "Did you load the `readr` package?"
    )
  })
  test_that("ggplot2 is loaded", {
    expect_true(
      "ggplot2" %in% .packages(),
      info = "Did you load the `ggplot2` package?"
    )
  })
  test_that("dplyr is loaded", {
    expect_true(
      "dplyr" %in% .packages(),
      info = "Did you load the `dplyr` package?"
    )
  })

  test_that("confirmed_cases_worldwide is a data.frame", {
    expect_s3_class(
      confirmed_cases_worldwide,
      "data.frame",
    )
  })
  test_that("confirmed_cases_worldwide has the correct column", {
    expect_identical(
      colnames(confirmed_cases_worldwide),
      colnames(soln_confirmed_cases_worldwide),
      info = "The column names of the `confirmed_cases_worldwide` data frame do not correspond with the ones in the CSV file: `\"datasets/confirmed_cases_worldwide.csv\"`."
    )
  })
  test_that("has the correct data", {
    expect_equal(
      confirmed_cases_worldwide,
      soln_confirmed_cases_worldwide,
      info = "The data of the `confirmed_cases_worldwide` data frame do not correspond with data in the CSV file: `\"datasets/confirmed_cases_worldwide.csv\"`."
    )
  })
})

```

Attaching package: 'testthat'

The following object is masked from 'package:dplyr':

matches

Column specification

```
cols(  
  date = col_date(format = ""),  
  cum_cases = col_double()  
)
```

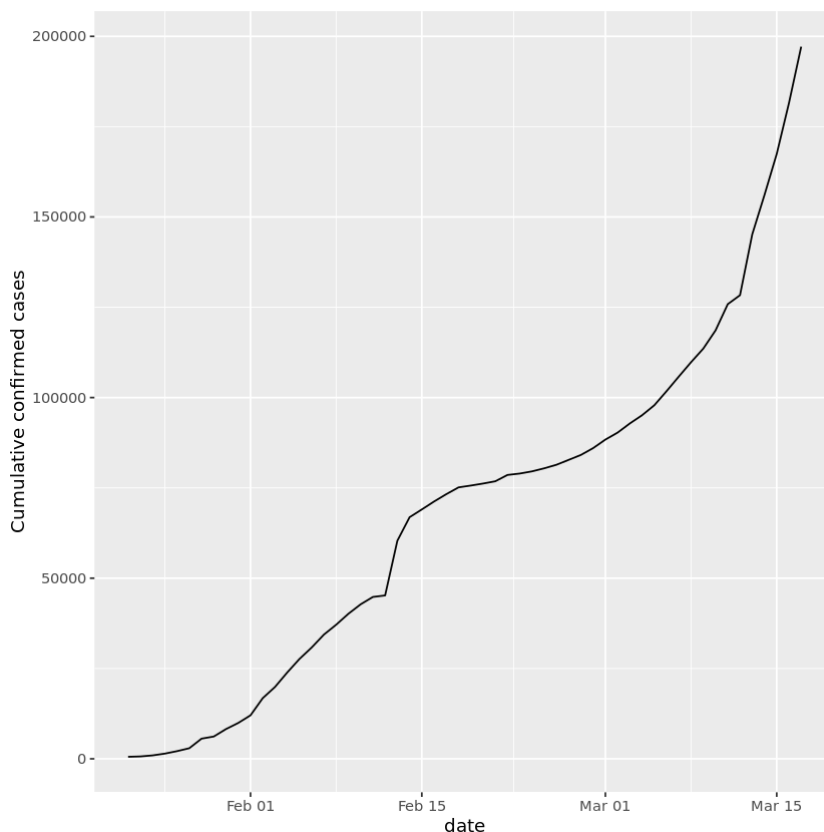
6/6 tests passed

2. Confirmed cases throughout the world

The table above shows the cumulative confirmed cases of COVID-19 worldwide by date. Just reading numbers in a table makes it hard to get a sense of the scale and growth of the outbreak. Let's draw a line plot to visualize the confirmed cases worldwide.

In [5]:

```
# Draw a line plot of cumulative cases vs. date  
# Label the y-axis  
ggplot(confirmed_cases_worldwide, aes(x=date, y=cum_cases)) +  
  geom_line() +  
  labs(y='Cumulative confirmed cases')
```



In [6]:

```

run_tests({
  plot <- last_plot()
  test_that("the plot is created", {
    expect_false(
      is.null(plot),
      info = "Could not find a plot created with `ggplot()``."
    )
  })
  test_that("the plot uses the correct data", {
    expect_equal(
      plot$data,
      confirmed_cases_worldwide,
      info = "The dataset used in the last plot is not `confirmed_cases_worldwide``."
    )
  })
  test_that("the plot uses the correct x aesthetic", {
    expect_equal(
      quo_name(plot$mapping$x),
      "date",
      info = "The x aesthetic used in the last plot is not `date``."
    )
  })
  test_that("the plot uses the correct y aesthetic", {
    expect_equal(
      quo_name(plot$mapping$y),
      "cum_cases",
      info = "The y aesthetic used in the last plot is not `cum_cases``."
    )
  })
  test_that("the plot uses the correct geom", {
    expect_true(
      "GeomLine" %in% class(plot$layers[[1]]$geom),
      info = "The geom used in the last plot is not `geom_line()``."
    )
  })
  test_that("the plot uses the correct y label", {
    expect_true(
      grepl("[Cc]umulative\\s+[Cc]onfirmed\\s+[Cc]ases", plot$labels$y),
      info = "The y label used in the last plot is not `\"Cumulative confirmed cases\"``."
    )
  })
})

```

6/6 tests passed

3. China compared to the rest of the world

The y-axis in that plot is pretty scary, with the total number of confirmed cases around the world approaching 200,000. Beyond that, some weird things are happening: there is an odd jump in mid February, then the rate of new cases slows down for a while, then speeds up again in March. We need to dig deeper to see what is happening.

Early on in the outbreak, the COVID-19 cases were primarily centered in China. Let's plot confirmed COVID-19 cases in China and the rest of the world separately to see if it gives us any insight.

We'll build on this plot in future tasks. One thing that will be important for the following tasks is that you add aesthetics within the line geometry of your ggplot, rather than making them global aesthetics.

In [7]:

```
# Read in datasets/confirmed_cases_china_vs_world.csv
confirmed_cases_china_vs_world <- read_csv("datasets/confirmed_cases_china_vs_world.csv")

# See the result
confirmed_cases_china_vs_world

# Draw a line plot of cumulative cases vs. date, colored by is_china
# Define aesthetics within the line geom
plt_cum_confirmed_cases_china_vs_world <- ggplot(confirmed_cases_china_vs_world) +
  geom_line(aes(x=date, y=cum_cases, color='is_china')) +
  ylab("Cumulative confirmed cases")

# See the plot
plt_cum_confirmed_cases_china_vs_world
```

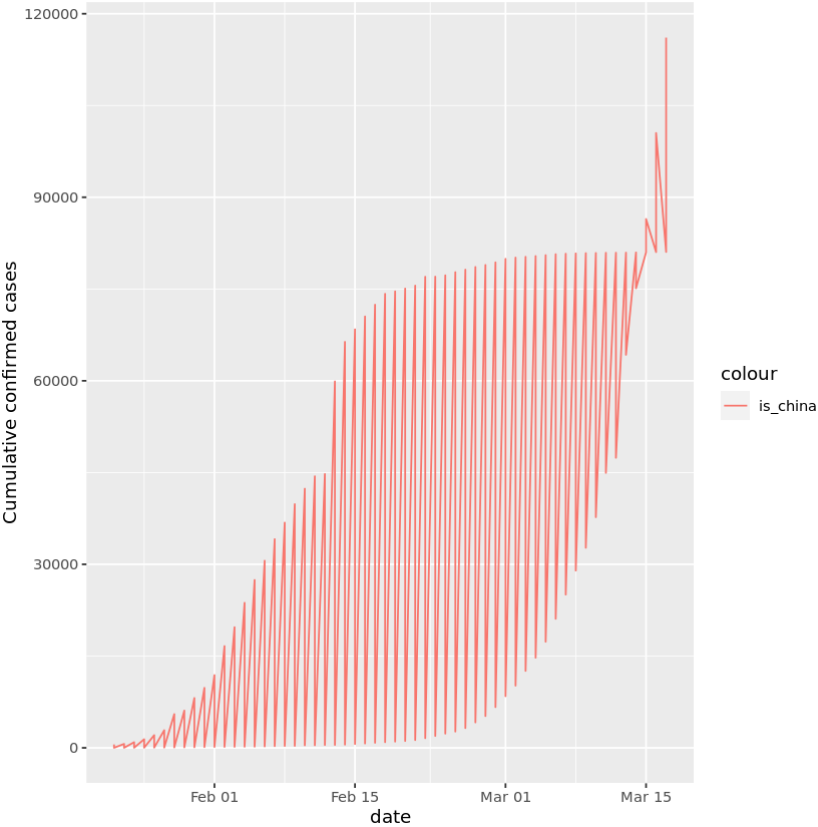
Column specification

```
cols(  
  is_china = col_character(),  
  date = col_date(format = ""),  
  cases = col_double(),  
  cum_cases = col_double()  
)
```

A spec_tbl_df: 112 × 4

is_china	date	cases	cum_cases
<chr>	<date>	<dbl>	<dbl>
China	2020-01-22	548	548
China	2020-01-23	95	643
China	2020-01-24	277	920
China	2020-01-25	486	1406
China	2020-01-26	669	2075
China	2020-01-27	802	2877
China	2020-01-28	2632	5509
China	2020-01-29	578	6087
China	2020-01-30	2054	8141
China	2020-01-31	1661	9802
China	2020-02-01	2089	11891
China	2020-02-02	4739	16630
China	2020-02-03	3086	19716
China	2020-02-04	3991	23707
China	2020-02-05	3733	27440
China	2020-02-06	3147	30587
China	2020-02-07	3523	34110
China	2020-02-08	2704	36814
China	2020-02-09	3015	39829
China	2020-02-10	2525	42354
China	2020-02-11	2032	44386
China	2020-02-12	373	44759
China	2020-02-13	15136	59895
China	2020-02-14	6463	66358
China	2020-02-15	2055	68413
China	2020-02-16	2100	70513
China	2020-02-17	1921	72434
China	2020-02-18	1777	74211
China	2020-02-19	408	74619
China	2020-02-20	458	75077
:	:	:	:
Not China	2020-02-17	113	824
Not China	2020-02-18	101	925
Not China	2020-02-19	95	1020
Not China	2020-02-20	100	1120
Not China	2020-02-21	153	1273

is_china	date	cases	cum_cases
<chr>	<date>	<dbl>	<dbl>
Not China	2020-02-22	305	1578
Not China	2020-02-23	365	1943
Not China	2020-02-24	384	2327
Not China	2020-02-25	332	2659
Not China	2020-02-26	570	3229
Not China	2020-02-27	925	4154
Not China	2020-02-28	1038	5192
Not China	2020-02-29	1463	6655
Not China	2020-03-01	1782	8437
Not China	2020-03-02	1733	10170
Not China	2020-03-03	2409	12579
Not China	2020-03-04	2155	14734
Not China	2020-03-05	2611	17345
Not China	2020-03-06	3749	21094
Not China	2020-03-07	3957	25051
Not China	2020-03-08	3921	28972
Not China	2020-03-09	3729	32701
Not China	2020-03-10	5004	37705
Not China	2020-03-11	7239	44944
Not China	2020-03-12	2467	47411
Not China	2020-03-13	16837	64248
Not China	2020-03-14	10872	75120
Not China	2020-03-15	11326	86446
Not China	2020-03-16	14052	100498
Not China	2020-03-17	15590	116088



In [8]:

```

soln_confirmed_cases_china_vs_world <- read_csv("datasets/confirmed_cases_china_vs_world.csv")

run_tests({
  test_that("confirmed_cases_china_vs_world is a data.frame", {
    expect_s3_class(
      confirmed_cases_china_vs_world,
      "data.frame"
    )
  })
  test_that("confirmed_cases_china_vs_world has the correct column names", {
    expect_identical(
      colnames(confirmed_cases_china_vs_world),
      colnames(soln_confirmed_cases_china_vs_world),
      info = "The column names of the `confirmed_cases_china_vs_world` data frame do not correspond with the ones in the CSV file: `\"datasets/confirmed_cases_china_vs_world.csv\"`."
    )
  })
  test_that("confirmed_cases_china_vs_world has the correct data", {
    expect_equal(
      confirmed_cases_china_vs_world,
      soln_confirmed_cases_china_vs_world,
      info = "The data of the `confirmed_cases_china_vs_world` data frame do not correspond with data in the CSV file: `\"datasets/confirmed_cases_china_vs_world.csv\"`."
    )
  })
  # NOTE: glimpse is not tested. Can this be done?
  test_that("plt_cum_confirmed_cases_china_vs_world is not NULL", {
    expect_false(
      is.null(plt_cum_confirmed_cases_china_vs_world),
      info = "`plt_cum_confirmed_cases_china_vs_world` is NULL."
    )
  })
  test_that("plt_cum_confirmed_cases_china_vs_world is a plot", {
    expect_true(
      "ggplot" %in% class(plt_cum_confirmed_cases_china_vs_world),
      info = "`plt_cum_confirmed_cases_china_vs_world` is not a `ggplot()` object."
    )
  })
  test_that("plt_cum_confirmed_cases_china_vs_world uses the correct data", {
    expect_equal(
      plt_cum_confirmed_cases_china_vs_world$data,
      confirmed_cases_china_vs_world,
      info = "The dataset used in `plt_cum_confirmed_cases_china_vs_world` is not `confirmed_cases_china_vs_world`."
    )
  })
  layer <- plt_cum_confirmed_cases_china_vs_world$layers[[1]]
  test_that("plt_cum_confirmed_cases_china_vs_world uses the correct geom", {
    expect_false(
      is.null(layer),
      info = "The geom used in `plt_cum_confirmed_cases_china_vs_world` is not `geom_line()`."
    )
  })
  test_that("plt_cum_confirmed_cases_china_vs_world uses the correct geom", {

```

```

    expect_true(
      "GeomLine" %in% class(layer$geom),
      info = "The geom used in `plt_cum_confirmed_cases_china_vs_world` is not `geom_line`."
    )
  })
  test_that("plt_cum_confirmed_cases_china_vs_world uses the correct x aesthetic", {
    expect_equal(
      quo_name(layer$mapping$x),
      "date",
      info = "The x aesthetic used in `plt_cum_confirmed_cases_china_vs_world` is not `date`."
    )
  })
  test_that("plt_cum_confirmed_cases_china_vs_world uses the correct y aesthetic", {
    expect_equal(
      quo_name(layer$mapping$y),
      "cum_cases",
      info = "The y aesthetic used in `plt_cum_confirmed_cases_china_vs_world` is not `cum_cases`."
    )
  })
  test_that("plt_cum_confirmed_cases_china_vs_world uses the correct color aesthetic", {
    expect_equal(
      quo_name(layer$mapping$colour),
      "is_china",
      info = "The color aesthetic used in `plt_cum_confirmed_cases_china_vs_world` is not `is_china`."
    )
  })
})

```

Column specification

```

cols(
  is_china = col_character(),
  date = col_date(format = ""),
  cases = col_double(),
  cum_cases = col_double()
)

```

11/11 tests passed

4. Let's annotate!

Wow! The two lines have very different shapes. In February, the majority of cases were in China. That changed in March when it really became a global outbreak: around March 14, the total number of cases outside China overtook the cases inside China. This was days after the WHO declared a pandemic.

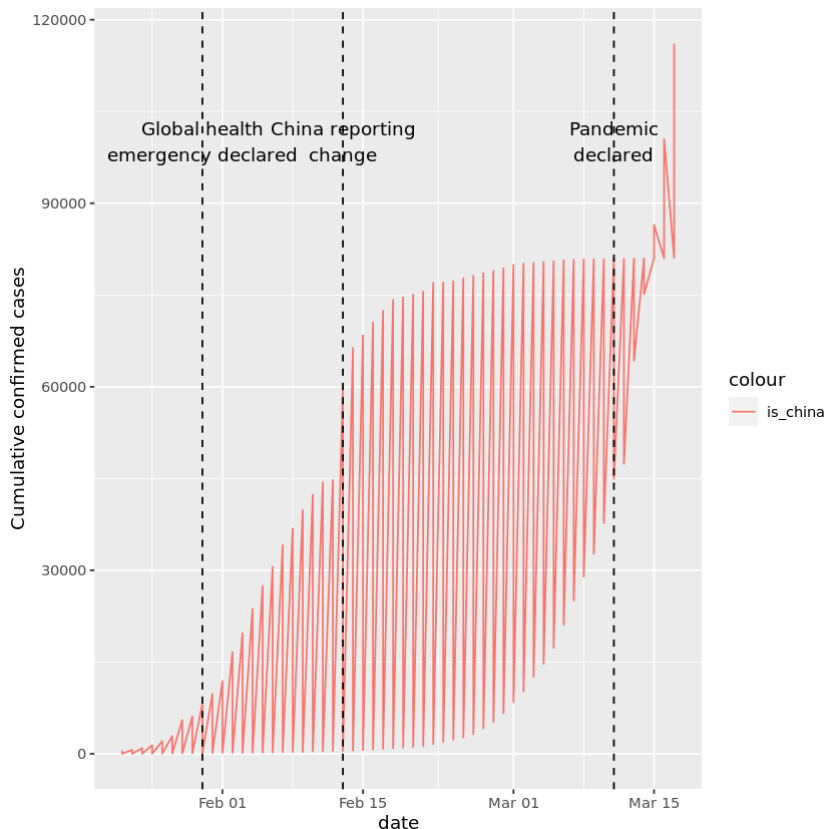
There were a couple of other landmark events that happened during the outbreak. For example, the huge jump in the China line on February 13, 2020 wasn't just a bad day regarding the outbreak; China changed the way it reported figures on that day (CT scans were accepted as evidence for COVID-19, rather than only lab tests).

By annotating events like this, we can better interpret changes in the plot.

In [9]:

```
who_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 who_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_confirmed_cases_china_vs_world +
  geom_vline(data=who_events, aes(xintercept=date), linetype="dashed") +
  geom_text(data=who_events, aes(x=date, label=event), y=1e5)
```



In [10]:

```

run_tests({
  plot <- last_plot()
  test_that("the plot got created", {
    expect_false(
      is.null(plot),
      info = "Could not find a plot created with `ggplot()`."
    )
  })
  layer1 <- plot$layers[[2]]
  layer2 <- plot$layers[[3]]
  test_that("the plot has both geoms", {
    expect_false(
      is.null(layer1) || is.null(layer2),
      info = "Could not find `geom_vline()` and `geom_text()` in your last plot."
    )
  })
  test_that("the plot has both geoms", {
    expect_true(
      "GeomVline" %in% class(layer1$geom) && "GeomText" %in% class(layer2$geom) |
      "GeomText" %in% class(layer1$geom) && "GeomVline" %in% class(layer2$geom),
      info = "Could not find `geom_vline()` and `geom_text()` in your last plot."
    )
  })
  if ("GeomVline" %in% class(layer1$geom)) {
    vline <- layer1
    text <- layer2
  } else {
    vline <- layer2
    text <- layer1
  }
  test_that("the plot uses the correct data", {
    expect_equal(
      vline$data,
      who_events,
      info = "The dataset used in the `geom_vline()` is not `who_events`."
    )
  })
  test_that("the geom uses the correct xintercept aesthetic", {
    expect_equal(
      quo_name(vline$mapping$xintercept),
      "date",
      info = "The xintercept aesthetic used in the `geom_vline()` is not `date`."
    )
  })
  test_that("the geom uses the correct linetype parameter", {
    expect_equal(
      vline$aes_params$linetype,
      "dashed",
      info = "The linetype parameter used in the `geom_vline()` is not `\"dashed\"`."
    )
  })
  test_that("the geom uses the correct data", {
    expect_equal(
      text$data,
      who_events,
      info = "The dataset used in the `geom_text()` is not `who_events`."
    )
  })
})

```

```
})
test_that("the geom uses the correct x aesthetic", {
  expect_equal(
    quo_name(text$mapping$x),
    "date",
    info = "The x aesthetic used in the `geom_text()` is not `date`."
  )
})
test_that("the geom uses the correct label aesthetic", {
  expect_equal(
    quo_name(text$mapping$label),
    "event",
    info = "The label aesthetic used in the `geom_text()` is not `event`."
  )
})
if(!is.null(text$aes_params$y)) {
  test_that("the geom uses the correct y parameter", {
    expect_equal(
      text$aes_params$y,
      100000
    )
  })
} else if (!is.null(quo_name(text$mapping$y))) {
  test_that("the geom uses the correct y parameter", {
    expect_equal(
      quo_name(text$mapping$y),
      '1e+05'
    )
  })
}
})
```

10/10 tests passed

5. Adding a trend line to China

When trying to assess how big future problems are going to be, we need a measure of how fast the number of cases is growing. A good starting point is to see if the cases are growing faster or slower than linearly.

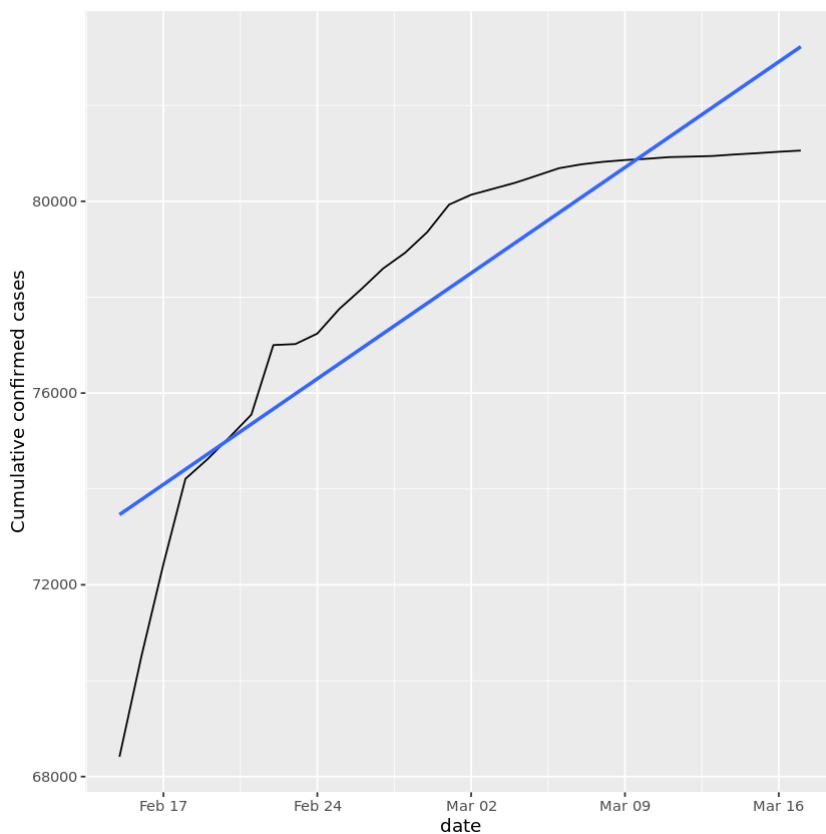
There is a clear surge of cases around February 13, 2020, with the reporting change in China. However, a couple of days after, the growth of cases in China slows down. How can we describe COVID-19's growth in China after February 15, 2020?

In [11]:

```
# Filter for China, from Feb 15
china_after_feb15 <- confirmed_cases_china_vs_world %>%
  filter(is_china=="China",date >= "2020-02-15")

# Using china_after_feb15, draw a line plot cum_cases vs. date
# Add a smooth trend line using linear regression, no error bars
ggplot(china_after_feb15,aes(x=date,y=cum_cases)) +
  geom_line() +
  geom_smooth(method="lm",se = FALSE) +
  ylab("Cumulative confirmed cases")
```

`geom_smooth()` using formula 'y ~ x'



In [12]:

```

run_tests({
  test_that("the data is filtered correctly", {
    soln_china_after_feb15 <- confirmed_cases_china_vs_world %>%
      filter(is_china == "China", date >= "2020-02-15")
    expect_equivalent(
      soln_china_after_feb15,
      china_after_feb15,
      info = "`china_after_feb15` has not been filtered correctly."
    )
  })
  plot <- last_plot()
  test_that("the plot is created", {
    expect_false(
      is.null(plot),
      info = "Could not find a plot created with `ggplot()`."
    )
  })
  test_that("the plot uses the correct data", {
    expect_equal(
      plot$data,
      china_after_feb15,
      info = "The dataset used in the last plot is not `soln_china_after_feb15`."
    )
  })
  test_that("the plot uses the correct x aesthetic", {
    expect_equal(
      quo_name(plot$mapping$x),
      "date",
      info = "The x aesthetic used in the last plot is not `date`."
    )
  })
  test_that("the plot uses the correct y aesthetic", {
    expect_equal(
      quo_name(plot$mapping$y),
      "cum_cases",
      info = "The y aesthetic used in the last plot is not `cum_cases`."
    )
  })
  layer1 <- plot$layers[[1]]
  layer2 <- plot$layers[[2]]
  test_that("the plot has the correct geoms", {
    expect_false(
      is.null(layer1) || is.null(layer2),
      info = "Could not find `geom_line()` and `geom_smooth()` in your last plot."
    )
  })
  test_that("the plot has the correct geoms", {
    expect_true(
      "GeomLine" %in% class(layer1$geom) && "GeomSmooth" %in% class(layer2$geom)
      ||
      "GeomSmooth" %in% class(layer1$geom) && "GeomLine" %in% class(layer2$geom),
      info = "Could not find `geom_line()` and `geom_smooth()` in your last plot."
    )
  })
  if ("GeomLine" %in% class(layer1$geom)) {
    line <- layer1
    smooth <- layer2
  } else {
    line <- layer2

```

```
    smooth <- layer1
  }
  test_that("the geom has the correct method parameter", {
    expect_equal(
      smooth$stat_params$method,
      "lm",
      info = "The method parameter used in the `geom_smooth()` is not `\"lm\"`."
    )
  })
  test_that("the geom has the correct se parameter", {
    expect_equal(
      smooth$stat_params$se,
      FALSE,
      info = "The se parameter used in the `geom_smooth()` is not `\"FALSE\"`."
    )
  })
})
```

9/9 tests passed

6. And the rest of the world?

From the plot above, the growth rate in China is slower than linear. That's great news because it indicates China has at least somewhat contained the virus in late February and early March.

How does the rest of the world compare to linear growth?

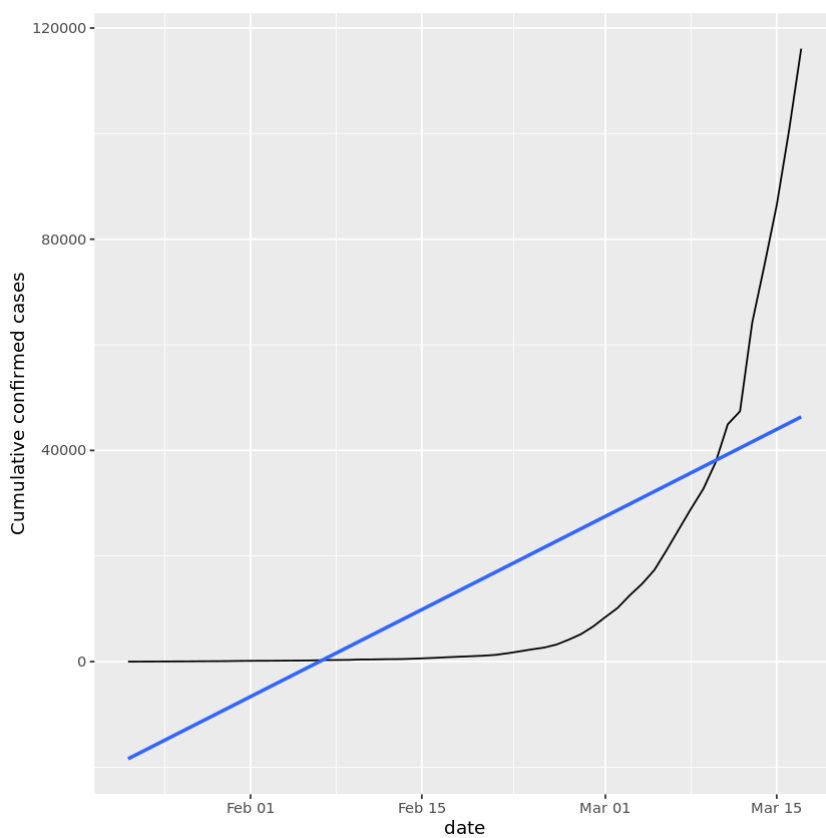
In [13]:

```
# Filter confirmed_cases_china_vs_world for not China
not_china <- confirmed_cases_china_vs_world %>%
  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_not_china_trend_lin <- ggplot(not_china, aes(x=date,y=cum_cases)) +
  geom_line() +
  geom_smooth(method="lm" , se= FALSE) +
  ylab("Cumulative confirmed cases")

# See the result
plt_not_china_trend_lin
```

`geom_smooth()` using formula 'y ~ x'



In [14]:

```

run_tests({
  test_that("the data is filtered correctly", {
    soln_not_china <- confirmed_cases_china_vs_world %>%
      filter(is_china == "Not China")
    expect_equal(
      soln_not_china,
      not_china,
      info = "`not_china` has not been filtered correctly."
    )
  })
  plot <- last_plot()
  test_that("the plot is created", {
    expect_false(
      is.null(plot),
      info = "Could not find a plot created with `ggplot()`. "
    )
  })
  test_that("the plot uses the correct data", {
    expect_equal(
      plot$data,
      not_china,
      info = "The dataset used in the last plot is not `not_china`."
    )
  })
  test_that("the plot uses the correct x aesthetic", {
    expect_equal(
      quo_name(plot$mapping$x),
      "date",
      info = "The x aesthetic used in the last plot is not `date`."
    )
  })
  test_that("the plot uses the correct y aesthetic", {
    expect_equal(
      quo_name(plot$mapping$y),
      "cum_cases",
      info = "The y aesthetic used in the last plot is not `cum_cases`."
    )
  })
  layer1 <- plot$layers[[1]]
  layer2 <- plot$layers[[2]]
  test_that("the plot uses the correct geoms", {
    expect_false(
      is.null(layer1) || is.null(layer2),
      info = "Could not find `geom_line()` and `geom_smooth()` in your last plot."
    )
  })
  test_that("the plot uses the correct geoms", {
    expect_true(
      "GeomLine" %in% class(layer1$geom) && "GeomSmooth" %in% class(layer2$geom)
      ||
      "GeomSmooth" %in% class(layer1$geom) && "GeomLine" %in% class(layer2$geom),
      info = "Could not find `geom_line()` and `geom_smooth()` in your last plot."
    )
  })
  if ("GeomLine" %in% class(layer1$geom)) {
    line <- layer1
    smooth <- layer2
  } else {
    line <- layer2

```



```
smooth <- layer1
}
test_that("the geom uses the correct method parameter", {
  expect_equal(
    smooth$stat_params$method,
    "lm",
    info = "The method parameter used in the `geom_smooth()` is not `\"lm\"`."
  )
})
test_that("the geom uses the correct se parameter", {
  expect_equal(
    smooth$stat_params$se,
    FALSE,
    info = "The se parameter used in the `geom_smooth()` is not `\"FALSE\"`."
  )
})
})
```

9/9 tests passed

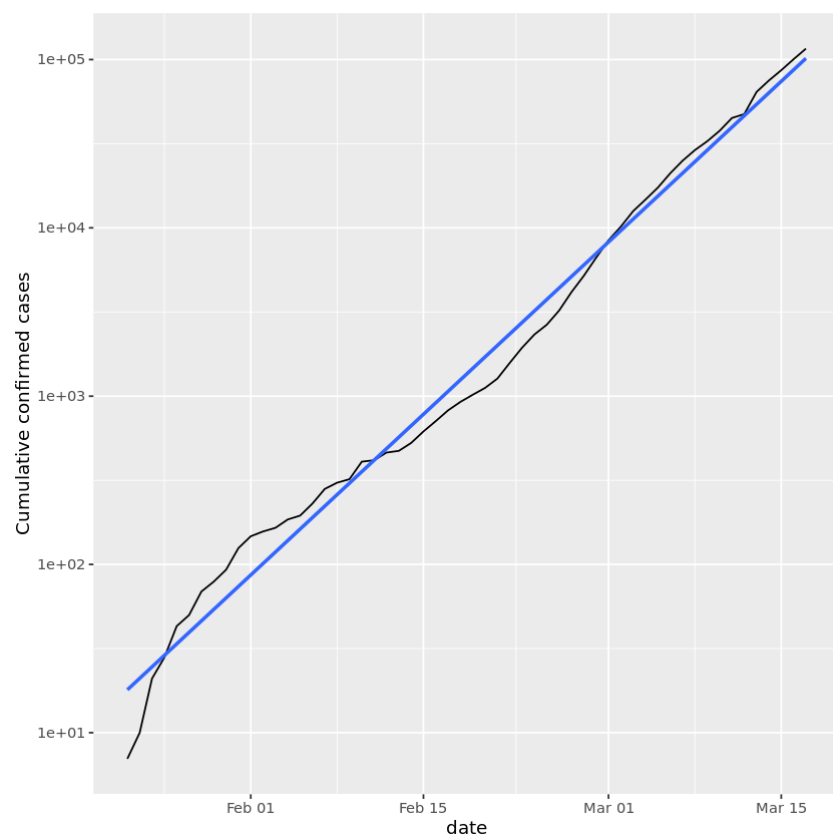
7. Adding a logarithmic scale

From the plot above, we can see a straight line does not fit well at all, and the rest of the world is growing much faster than linearly. What if we added a logarithmic scale to the y-axis?

In [15]:

```
# Modify the plot to use a logarithmic scale on the y-axis  
plt_not_china_trend_lin +  
  scale_y_log10()
```

`geom_smooth()` using formula 'y ~ x'



In [16]:

```
run_tests({
  plot <- last_plot()
  test_that("the plot is created", {
    expect_false(
      is.null(plot),
      info = "Could not find a plot created with `ggplot()`."
    )
  })
  scale <- plot$scales$get_scales(aes("y"))
  test_that("the plot has a scale", {
    expect_false(
      is.null(scale),
      info = "Could not find a scale in your last plot."
    )
  })
  test_that("the plot uses the correct scale", {
    expect_equal(
      scale$trans$name,
      "log-10",
      info = "Could not find a logarithmic y scale: `scale_y_log10()`."
    )
  })
})
```

3/3 tests passed

8. Which countries outside of China have been hit hardest?

With the logarithmic scale, we get a much closer fit to the data. From a data science point of view, a good fit is great news. Unfortunately, from a public health point of view, that means that cases of COVID-19 in the rest of the world are growing at an exponential rate, which is terrible news.

Not all countries are being affected by COVID-19 equally, and it would be helpful to know where in the world the problems are greatest. Let's find the countries outside of China with the most confirmed cases in our dataset.

In [17]:

```
# Run this to get the data for each country
confirmed_cases_by_country <- read_csv("datasets/confirmed_cases_by_country.csv")
glimpse(confirmed_cases_by_country)

# Group by country, summarize to calculate total cases, find the top 7
top_countries_by_total_cases <- confirmed_cases_by_country %>%
  group_by(country) %>%
  summarize(total_cases = max(cum_cases)) %>%
  top_n(7,total_cases)

# See the result
top_countries_by_total_cases
```

Column specification

```
cols(
  country = col_character(),
  province = col_character(),
  date = col_date(format = ""),
  cases = col_double(),
  cum_cases = col_double()
)
```

Rows: 13,272

Columns: 5

```
$ country <chr> "Afghanistan", "Albania", "Algeria", "Andorra", "Antigua
and...
$ province <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA, ...
$ date <date> 2020-01-22, 2020-01-22, 2020-01-22, 2020-01-22, 2020-01
-22,...
$ cases <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, ...
$ cum_cases <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, ...
```

A tibble: 7 × 2

country	total_cases
<chr>	<dbl>
France	7699
Germany	9257
Iran	16169
Italy	31506
Korea, South	8320
Spain	11748
US	6421

In [18]:

```
run_tests({  
  test_that("the data is manipulated correctly", {  
    soln_top_countries_by_total_cases <- confirmed_cases_by_country %>%  
      group_by(country) %>%  
      summarize(total_cases = max(cum_cases)) %>%  
      top_n(7, total_cases)  
    expect_equivalent(  
      soln_top_countries_by_total_cases,  
      top_countries_by_total_cases,  
      info = "`top_countries_by_total_cases` has not been filtered correctly."  
    )  
  })  
})
```

1/1 tests passed

9. Plotting hardest hit countries as of Mid-March 2020

Even though the outbreak was first identified in China, there is only one country from East Asia (South Korea) in the above table. Four of the listed countries (France, Germany, Italy, and Spain) are in Europe and share borders. To get more context, we can plot these countries' confirmed cases over time.

Finally, congratulations on getting to the last step! If you would like to continue making visualizations or find the hardest hit countries as of today, you can do your own analyses with the latest data available [here](https://github.com/RamiKrispin/coronavirus) (<https://github.com/RamiKrispin/coronavirus>).

In [19]:

```
# Read in the dataset from datasets/confirmed_cases_top7_outside_china.csv
confirmed_cases_top7_outside_china <- read_csv("datasets/confirmed_cases_top7_outside_c
hina.csv")

# Glimpse at the contents of confirmed_cases_top7_outside_china
glimpse(confirmed_cases_top7_outside_china)

# Using confirmed_cases_top7_outside_china, draw a line plot of
# cum_cases vs. date, colored by country
ggplot(confirmed_cases_top7_outside_china, aes(x=date, y=cum_cases, color=country))+
  geom_line() + ylab("Cumulative confirmed cases")
```

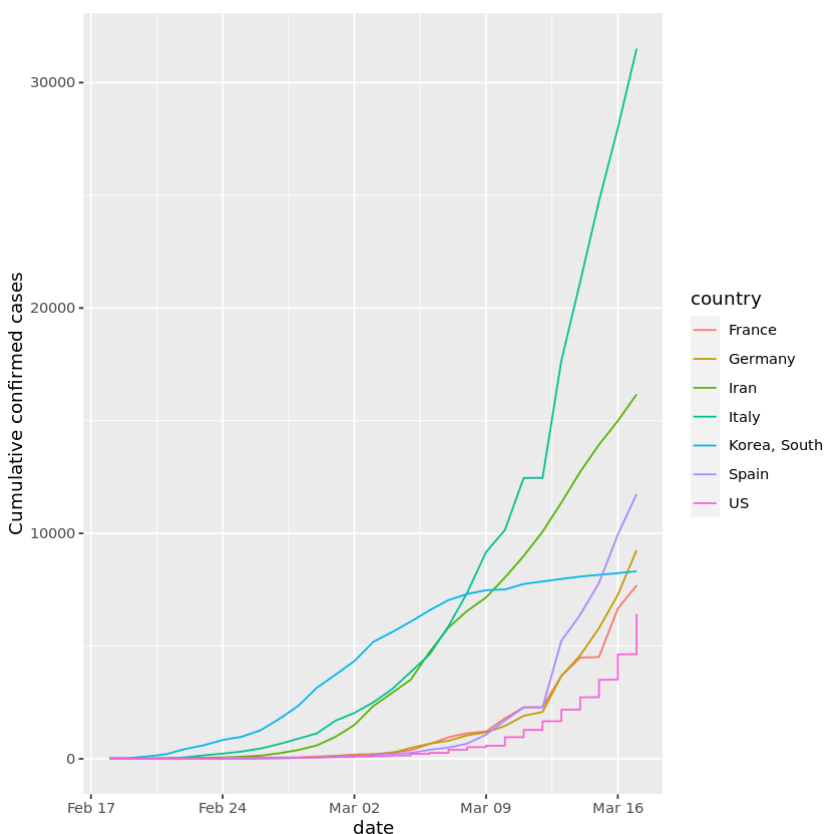
Column specification

```
cols(
  country = col_character(),
  date = col_date(format = ""),
  cum_cases = col_double()
)
```

Rows: 2,030

Columns: 3

```
$ country <chr> "Germany", "Iran", "Italy", "Korea, South", "Spain", "U
S", "...
$ 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,...
```



In [20]:

```
soln_confirmed_cases_top7_outside_china <- read_csv("datasets/confirmed_cases_top7_outside_china.csv")

run_tests({
  test_that('confirmed_cases_top7_outside_china is a data.frame', {
    expect_s3_class(
      confirmed_cases_top7_outside_china,
      'data.frame'
    )
  })
  test_that('confirmed_cases_top7_outside_china had the correct column names', {
    expect_identical(
      colnames(confirmed_cases_top7_outside_china),
      colnames(soln_confirmed_cases_top7_outside_china),
      info = "The column names of the `confirmed_cases_top7_outside_china` data frame do not correspond with the ones in the CSV file: `\"datasets/confirmed_cases_top7_outside_china.csv\"`."
    )
  })
  test_that('confirmed_cases_top7_outside_china had the correct data', {
    expect_equal(
      confirmed_cases_top7_outside_china,
      soln_confirmed_cases_top7_outside_china,
      info = "The data of the `confirmed_cases_top7_outside_china` data frame do not correspond with data in the CSV file: `\"datasets/confirmed_cases_top7_outside_china.csv\"`."
    )
  })
  # NOTE: glimpse is not tested. Can this be done?
  plot <- last_plot()
  test_that('the plot is created', {
    expect_false(
      is.null(plot),
      info = "Could not find a plot created with `ggplot()`. "
    )
  })
  test_that('the plot uses the correct data', {
    expect_equal(
      plot$data,
      confirmed_cases_top7_outside_china,
      info = "The dataset used in the last plot is not `not_china`."
    )
  })
  line <- plot$layers[[1]]
  test_that('the plot uses the correct geom', {
    expect_false(
      is.null(line),
      info = "Could not find `geom_line()` in your last plot."
    )
  })
  test_that('the plot uses the correct geom', {
    expect_true(
      'GeomLine' %in% class(line$geom),
      info = "Could not find `geom_line()` in your last plot."
    )
  })
  mapping <- plot$mapping
  geom_mapping <- line$mapping
  test_that('the plot uses the correct x aesthetic', {
```

```
    expect_true(
      !is.null(mapping$x) && quo_name(mapping$x) == "date" ||
      !is.null(geom_mapping$x) && quo_name(geom_mapping$x) == "date",
      info = "The x aesthetic used in the last plot is not `date`."
    )
  })
  test_that('the plot uses the correct y aesthetic', {
    expect_true(
      !is.null(mapping$y) && quo_name(mapping$y) == "cum_cases" ||
      !is.null(geom_mapping$y) && quo_name(geom_mapping$y) == "cum_cases",
      info = "The y aesthetic used in the last plot is not `cum_cases`."
    )
  })
  test_that('the plot uses the correct color aesthetic', {
    expect_true(
      !is.null(mapping$colour) && quo_name(mapping$colour) == "country" ||
      !is.null(geom_mapping$colour) && quo_name(geom_mapping$colour) == "country"
    ,
      info = "The color aesthetic used in the last plot is not `country`."
    )
  })
})
```

Column specification

```
cols(
  country = col_character(),
  date = col_date(format = ""),
  cum_cases = col_double()
)
```

10/10 tests passed