Data Analysis

Alexander Kopp

1/17/2022

Task

1. Select in Moodle one of the available datasets.
2. Download the data and put it into your project.
3. Read the data into an RMD file, perform data wrangling as required and useful, analyse the data, visualize the data and interpret the data.

Used library:

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

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

Data Wrangling:

First we retrieve the data as tibble using read_delim with the delimiter being ";". Due to the fact that there is a one-to-one relationship between `BundeslandID` and `Name` we will drop `BundeslandID`. Finally we rename the columns on the one hand by translating german into english, on the other to make the structure clearer with shorter names.

```
data <- read_delim("Corona-Tests.csv", delim = ";") |>
  select(-BundeslandID) |> # drops mentioned column
  rename("Date" = "Datum",
         "PharmacyTests" = "TestungenApotheken",
         "PharmacyPCR" = "TestungenApothekenPCR",
         "PharmacyAntigen" = "TestungenApothekenAntigen",
         "BusinessTests" = "TestungenBetriebe")
```

```
## Rows: 2530 Columns: 7

## -- Column specification --------------------------------------------------------
## Delimiter: ";"
## chr (2): Datum, Name
## dbl (5): BundeslandID, TestungenApotheken, TestungenApothekenPCR, TestungenA...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Finally we see the structure of the data: for every day we can view the total of tests made in pharmacies and businesses in each state of Austria and the country itself. In addition the tests in pharmacies are seperated into PCR and antigen tests. It is important to note that the data for PCR tests starts in september.

```
data |> head(20)
```

```
## # A tibble: 20 x 6
##    Date       Name             PharmacyTests PharmacyPCR PharmacyAntigen BusinessTests
##    <chr>      <chr>                    <dbl>       <dbl>           <dbl>         <dbl>
##  1 28.04.2021 Burgenland              236422          NA          236422         68761
##  2 28.04.2021 Kärnten                 176205          NA          176205         78015
##  3 28.04.2021 Niederösterreich        397169          NA          397169        510424
##  4 28.04.2021 Oberösterreich          347180          NA          347180        306641
##  5 28.04.2021 Salzburg                246114          NA          246114        101554
##  6 28.04.2021 Steiermark              542714          NA          542714        388944
##  7 28.04.2021 Tirol                   321005          NA          321005         67924
##  8 28.04.2021 Vorarlberg              204500          NA          204500         56577
##  9 28.04.2021 Wien                    987752          NA          987752        296169
## 10 28.04.2021 Österreich             3459061          NA         3459061       1875009
## 11 29.04.2021 Burgenland              242288          NA          242288         73433
## 12 29.04.2021 Kärnten                 181185          NA          181185         83225
## 13 29.04.2021 Niederösterreich        405068          NA          405068        523172
## 14 29.04.2021 Oberösterreich          357106          NA          357106        315155
## 15 29.04.2021 Salzburg                253543          NA          253543        105762
## 16 29.04.2021 Steiermark              557460          NA          557460        402836
## 17 29.04.2021 Tirol                   327972          NA          327972         72415
## 18 29.04.2021 Vorarlberg              214257          NA          214257         61870
## 19 29.04.2021 Wien                   1011712          NA         1011712        345276
## 20 29.04.2021 Österreich             3550591          NA         3550591       1983144
```

For easier handling we convert the Date column into the right type. This proves to be a more complex task than initially thought as the date format mysteriously changes beginning december 28th 2021 (which equals row 2441).

```
# split the data into two tibbles with different format
data_upper <- data[1:2440,1:6]
data_lower <- data[2441:nrow(data),1:6] # nrow computes the number of rows

# convert the tibbles to the same format
data_upper <- mutate(data_upper, Date = as.Date(Date, "%d.%m.%Y")) # day.month.year
data_lower <- mutate(data_lower, Date = as.Date(Date, "%Y-%m-%d")) # year-month-day
```

```r
# combine both tibbles into data again
data <- rbind(data_upper, data_lower)
```

Now we split the data into 10 different tibbles, each corresponding to a state respectively Austria.

```r
bgl <- filter(data, Name == "Burgenland")
car <- filter(data, Name == "Kärnten")
noe <- filter(data, Name == "Niederösterreich")
ooe <- filter(data, Name == "Oberösterreich")
sal <- filter(data, Name == "Salzburg")
stm <- filter(data, Name == "Steiermark")
tir <- filter(data, Name == "Tirol")
vbg <- filter(data, Name == "Vorarlberg")
vie <- filter(data, Name == "Wien")
aut <- filter(data, Name == "Österreich")
```

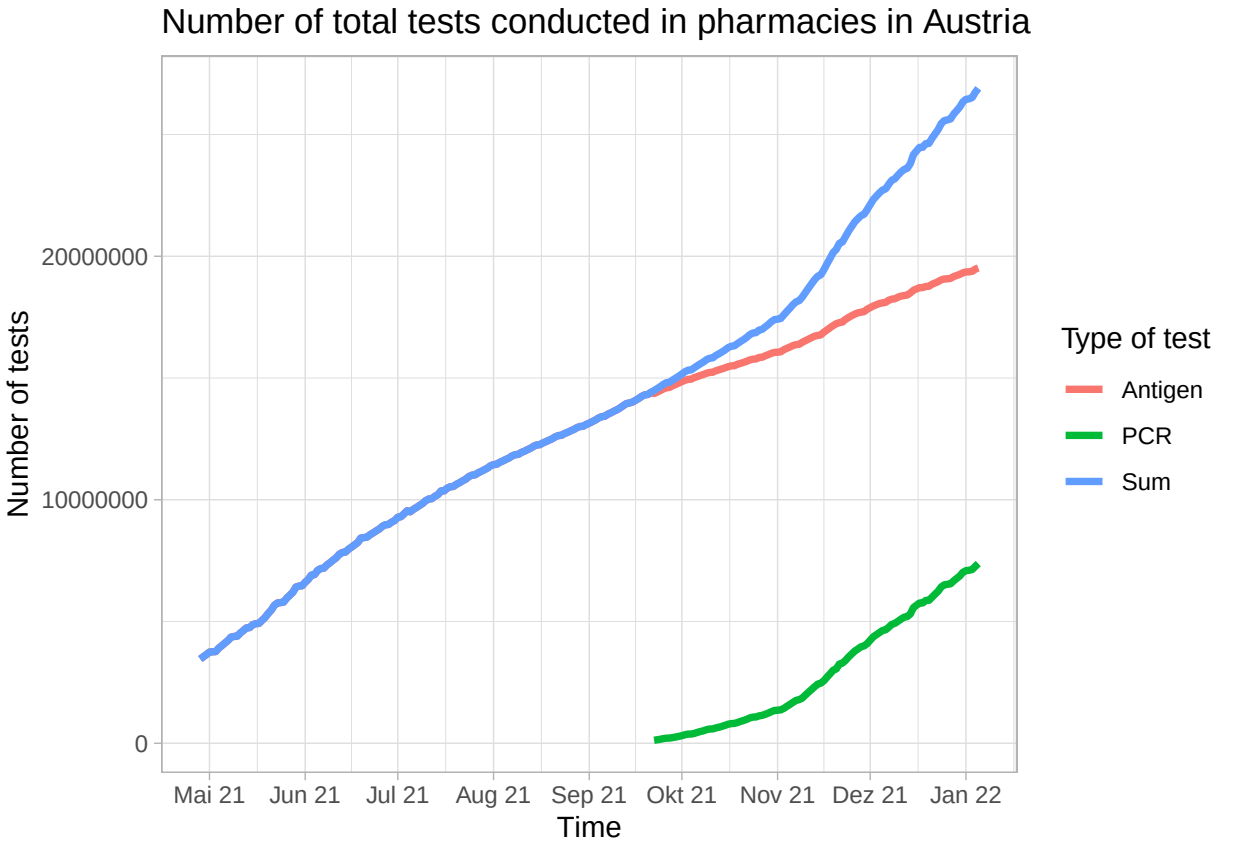
Data Visualisation & Analysis

Pharmacy test type Austria

At first we examine the total number of tests conducted in pharmacies comparing PCR to antigen.

```r
options(scipen = 1000) # opt out scientific notation

# use gather for key value pairs, this comes in very handy for plots with multiple graphs
longaut <- gather(aut,
                  key = "Type of test",
                  value = "Number of tests",
                  PharmacyPCR, PharmacyAntigen, PharmacyTests)

ggplot(longaut, aes(x = Date, y = `Number of tests`, col = `Type of test`)) +
  geom_line(size = 1.3) +
  scale_color_discrete(labels = c("Antigen", "PCR", "Sum")) +
  scale_x_date(date_breaks = "1 month", date_labels = "%b %y") +
  theme_light() +
  ggtitle("Number of total tests conducted in pharmacies in Austria") +
  xlab("Time") +
  ylab("Number of tests")
```

```
## Warning: Removed 147 row(s) containing missing values (geom_path).
```



We can see that there is a great demand for PCR tests at the pharmacy since november, whereas the demand for antigen tests increased just a little.

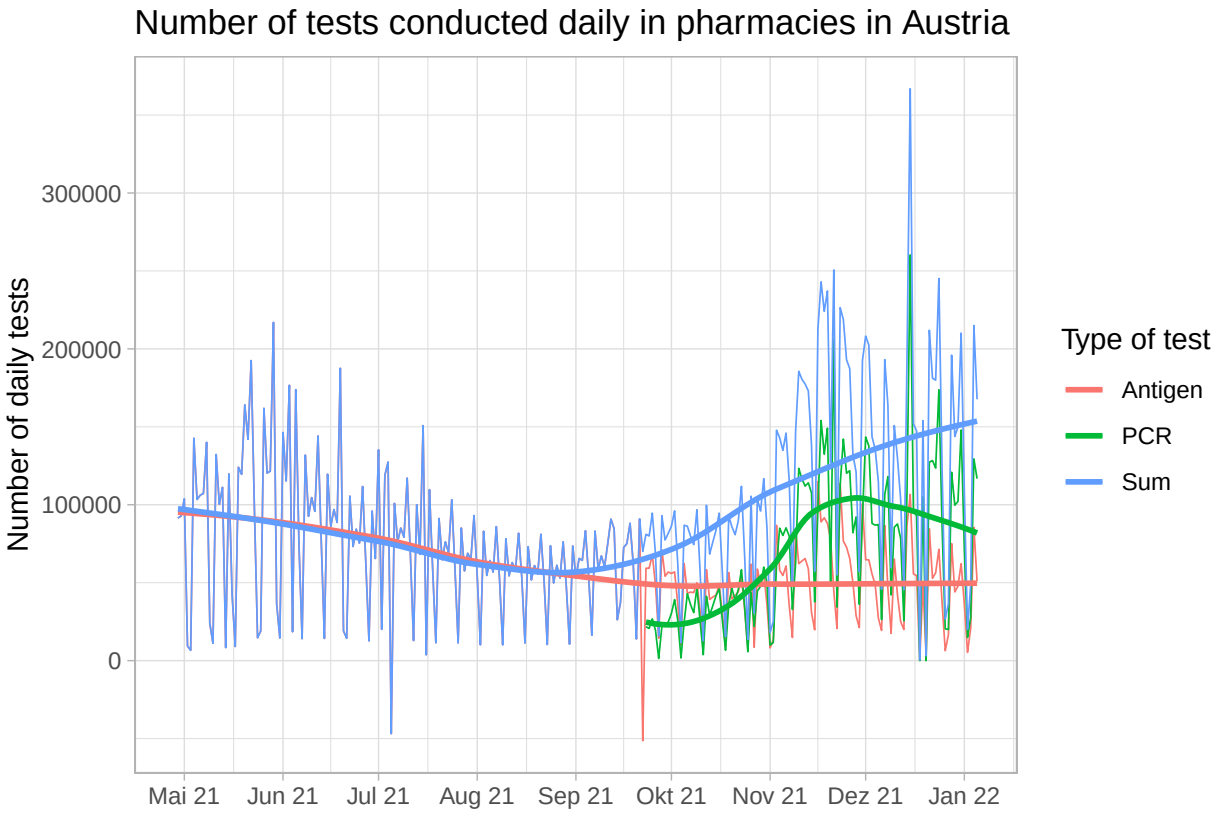
Of course we're not only interested for the total number of tests ever conducted, but also at the daily amount. For this reason we create new columns using the lag operator...

```
autdaily <- mutate(aut,
              PTdaily = PharmacyTests - lag(PharmacyTests),
              PCRdaily = PharmacyPCR - lag(PharmacyPCR),
              Antidaily = PharmacyAntigen - lag(PharmacyAntigen))
```

and view them as a lineplot. Due to the vivid changes per week, we additionally smooth the data.

```
longautdaily <- gather(autdaily,
                       key = "Type of test",
                       value = "Number of tests per day",
                       PCRdaily, Antidaily, PTdaily)
ggplot(longautdaily, aes(x = Date, y = `Number of tests per day`, col = `Type of test`)) +
  geom_line(size = 0.3) +
  geom_smooth(se = FALSE) +
  scale_color_discrete(labels = c("Antigen", "PCR", "Sum")) +
  scale_x_date(date_breaks = "1 month", date_labels = "%b %y") +
  theme_light() +
  ggtitle("Number of tests conducted daily in pharmacies in Austria") +
  xlab("") +
  ylab("Number of daily tests")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'

## Warning: Removed 150 rows containing non-finite values (stat_smooth).

## Warning: Removed 150 row(s) containing missing values (geom_path).
```



We can see many different properties:

- There is a day for the sum and two days for antigen where negative daily data was registered. This is probably due to the fact that on days before or after, to many tests were listed and the total number had been corrected.
- Here we actually see, that already during the end of october the demand for PCR tests was rising (still being the most dramatic during november). Since december the requests are very high, but declining a small amount.
- December 15th 2021 was a record breaking day for Tests at pharmacies, both PCR and antigen. Added together more than 350.000 tests were conducted this day. This probably correlates with the need for christmas shopping, as the lockdown had ended just two days earlier.

Similar to the whole country of Austria we can also examine single states. As an example we will investigate the data of Vienna and Salzburg.

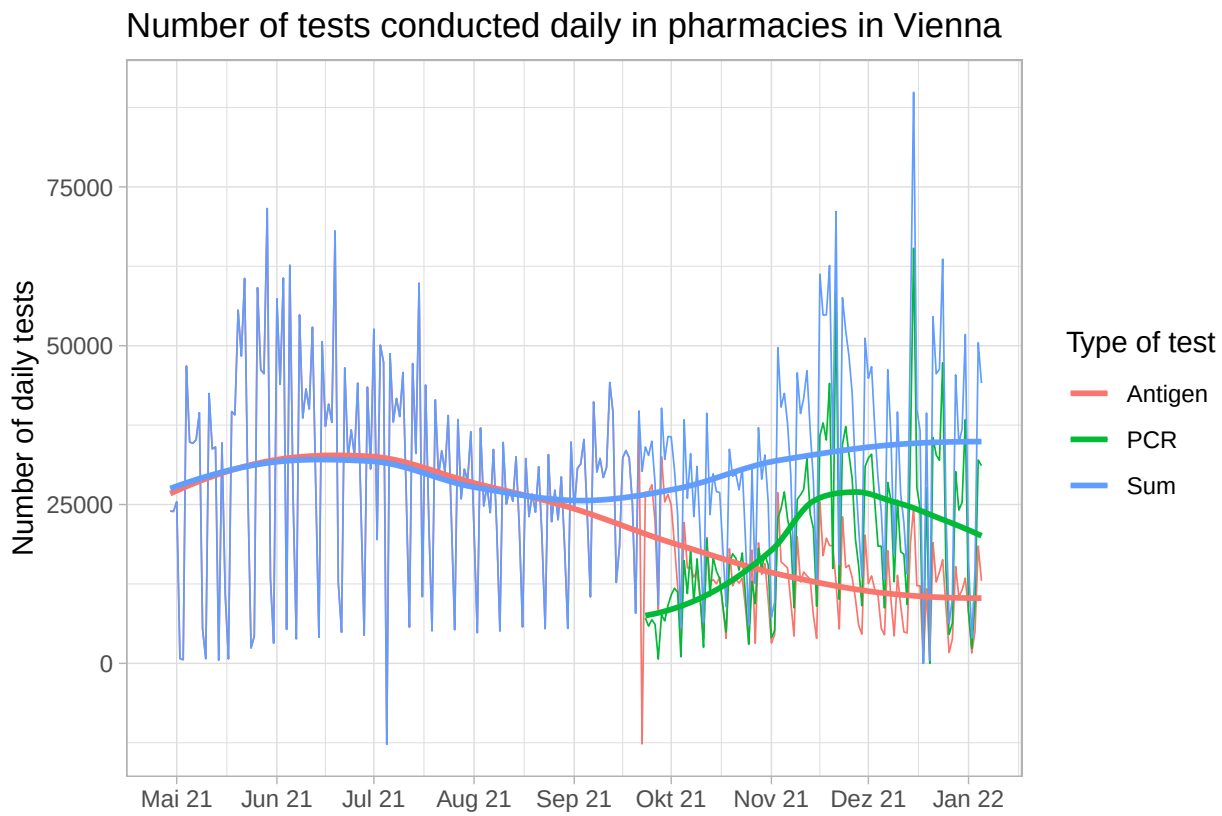
Vienna

We use the daily view because it is more meaningful.

```
viedaily <- mutate(vie,
              PTdaily = PharmacyTests - lag(PharmacyTests),
              PCRdaily = PharmacyPCR - lag(PharmacyPCR),
              Antidaily = PharmacyAntigen - lag(PharmacyAntigen))
```

```
longviedaily <- gather(viedaily,
                       key = "Type of test",
                       value = "Number of tests per day",
                       PCRdaily, Antidaily, PTdaily)

ggplot(longviedaily, aes(x = Date, y = `Number of tests per day`, col = `Type of test`)) +
  geom_line(size = 0.3) +
  geom_smooth(se = FALSE) +
  scale_color_discrete(labels = c("Antigen", "PCR", "Sum")) +
  scale_x_date(date_breaks = "1 month", date_labels = "%b %y") +
  theme_light() +
  ggtitle("Number of tests conducted daily in pharmacies in Vienna") +
  xlab("") +
  ylab("Number of daily tests")
```

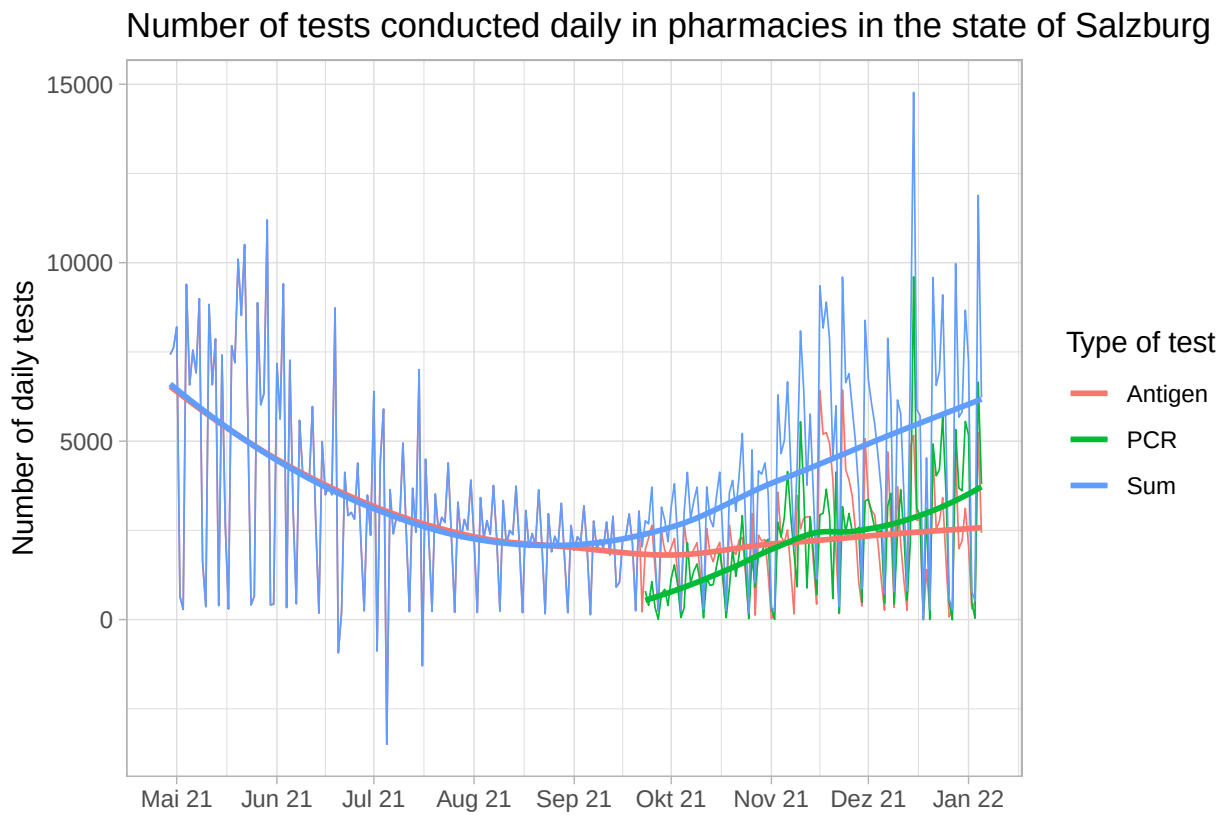
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'

## Warning: Removed 150 rows containing non-finite values (stat_smooth).

## Warning: Removed 150 row(s) containing missing values (geom_path).
```



Salzburg

```
saldaily <- mutate(sal,
              PTdaily = PharmacyTests - lag(PharmacyTests),
              PCRdaily = PharmacyPCR - lag(PharmacyPCR),
              Antidaily = PharmacyAntigen - lag(PharmacyAntigen))

longsaldaily <- gather(saldaily,
                       key = "Type of test",
                       value = "Number of tests per day",
                       PCRdaily, Antidaily, PTdaily)

ggplot(longsaldaily, aes(x = Date, y = `Number of tests per day`, col = `Type of test`)) +
  geom_line(size = 0.3) +
  geom_smooth(se = FALSE) +
  scale_color_discrete(labels = c("Antigen", "PCR", "Sum")) +
  scale_x_date(date_breaks = "1 month", date_labels = "%b %y") +
  theme_light() +
  ggtitle("Number of tests conducted daily in pharmacies in the state of Salzburg") +
  xlab("") +
  ylab("Number of daily tests")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'

## Warning: Removed 150 rows containing non-finite values (stat_smooth).

## Warning: Removed 150 row(s) containing missing values (geom_path).
```



Comparing Vienna with Salzburg we view some differences:

- In Salzburg the demand for tests fell strong during summer whereas in Vienna there was a just small and later decline.
- The number of PCR tests in Salzburg is just a little bit higher than antigen since november, whereas the number for PCR tests is almost the double of antigen in Vienna.

Tests pharmacies vs. business Austria

Finally we compare the total number of daily tests made in pharmacies with businesses in Austria.

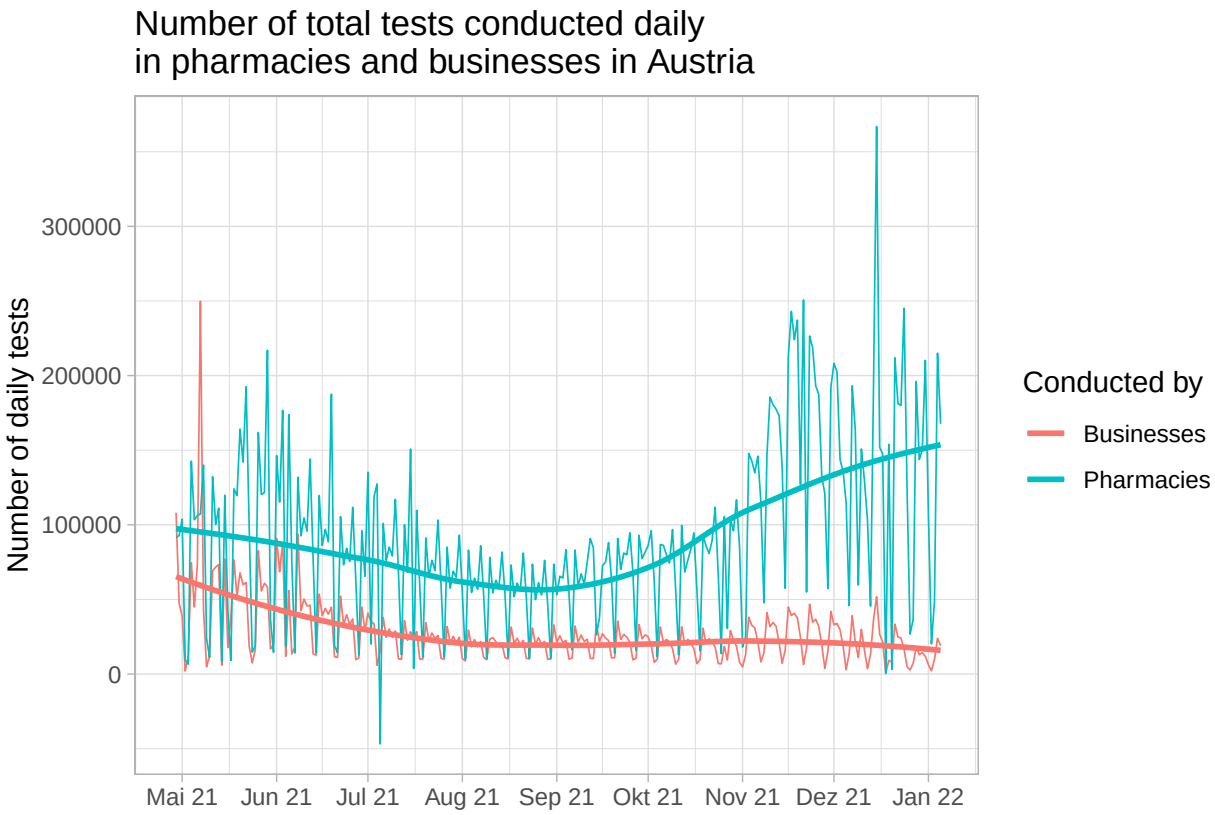
```
autvs <- mutate(aut,
                Pharmacies = PharmacyTests - lag(PharmacyTests),
                Businesses = BusinessTests - lag(BusinessTests))

longautvs <- gather(autvs,
                    key = "Conducted by",
                    value = "Number of tests per day",
                    Pharmacies, Businesses)

ggplot(longautvs, aes(x = Date, y = `Number of tests per day`, col = `Conducted by`)) +
  geom_line(size = 0.3) +
  geom_smooth(se = FALSE) +
  scale_x_date(date_breaks = "1 month", date_labels = "%b %y") +
  theme_light() +
```

```
  ggtitle("Number of total tests conducted daily \nin pharmacies and businesses in Austria") +
  xlab("") +
  ylab("Number of daily tests")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## Warning: Removed 2 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 2 row(s) containing missing values (geom_path).
```



As we see, far more tests are conducted in pharmacies compared to businesses (except one day in april and one in may). Especially since october the numbers for pharmacies were almost eight times as high than for businesses.