Data Wrangling Final Project: COVID19 Effect in Italy

Github repository link (https://github.com/narenbakshi97/COVID-19-effect-in-Italy-)

COVID 19:

The COVID-19 pandemic, also known as the coronavirus pandemic, is an ongoing pandemic of coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The outbreak was identified in Wuhan, China, in December 2019. The World Health Organization declared the outbreak a Public Health Emergency of International Concern on 30 January, and a pandemic on 11 March.As of 4 May 2020, more than 3.52 million cases of COVID-19 have been reported in 187 countries and territories, resulting in more than 248,000 deaths. More than 1.13 million people have recovered.

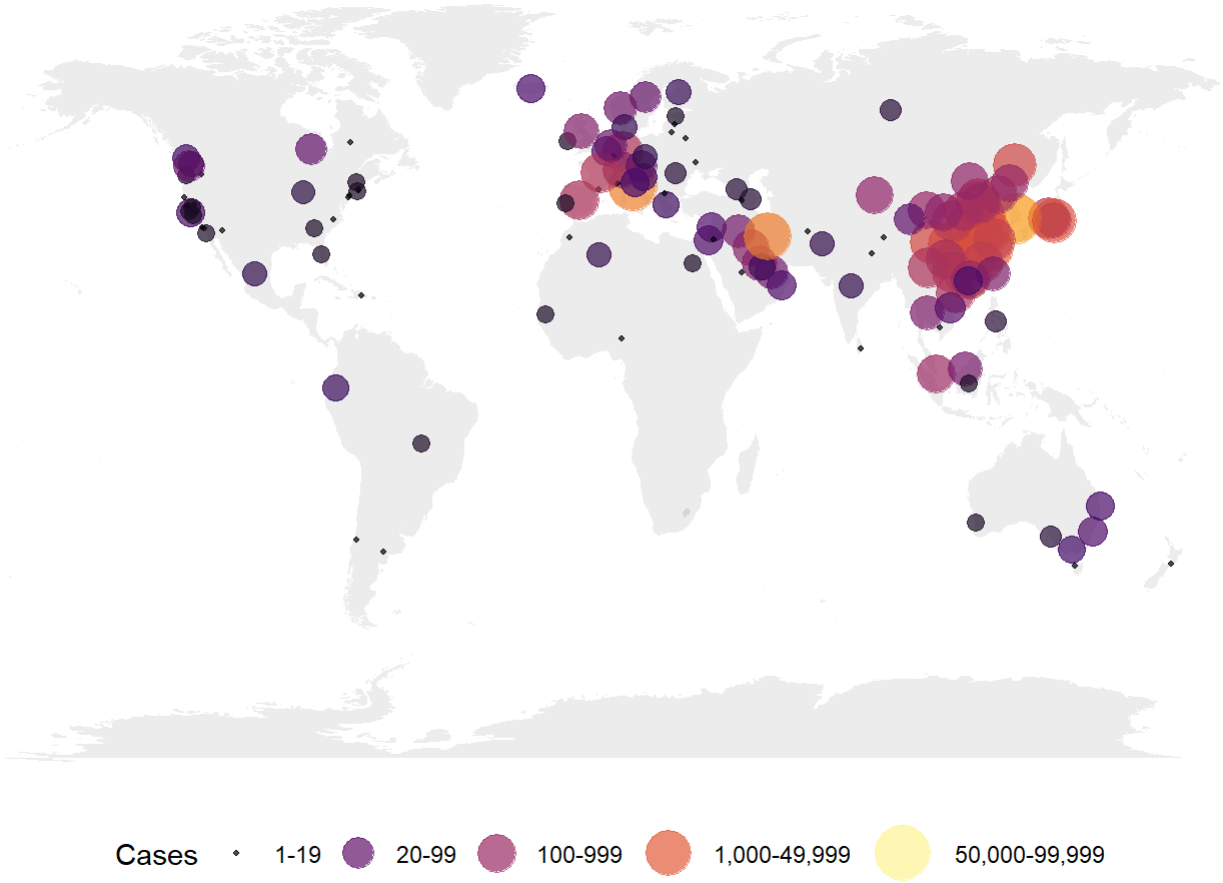
- Source: Wikipedia (https://en.wikipedia.org/wiki/COVID-19_pandemic)

```
## [1] "Using direct authentication"
```

1. Global Impact:

Let's first visualize the global impact of COVID-19. I used the package managed by Johns Hopkins University. The repository contains time series data of confirmed, recovered, and deaths cases world-wide and for US in seperate CSV files, I used the csv file containing information about the confirmed cases world wide.

- Github repository source (https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series)



Worldwide effect of COVID-19

This is an archived version of the data, which was updated 1 month ago. Here we can see that number of case are still building up in the US while the outbreak was serious and getting out of control for Italy and other Europian countries, along with China.

2.Italy

As we can see from the world data as well, we can see that Italy is one of the most affected countries by COVID-19. I have chose multiple datasets for analysing the effect based on different parameters. The datasets from Kaggle. Here are the sources.

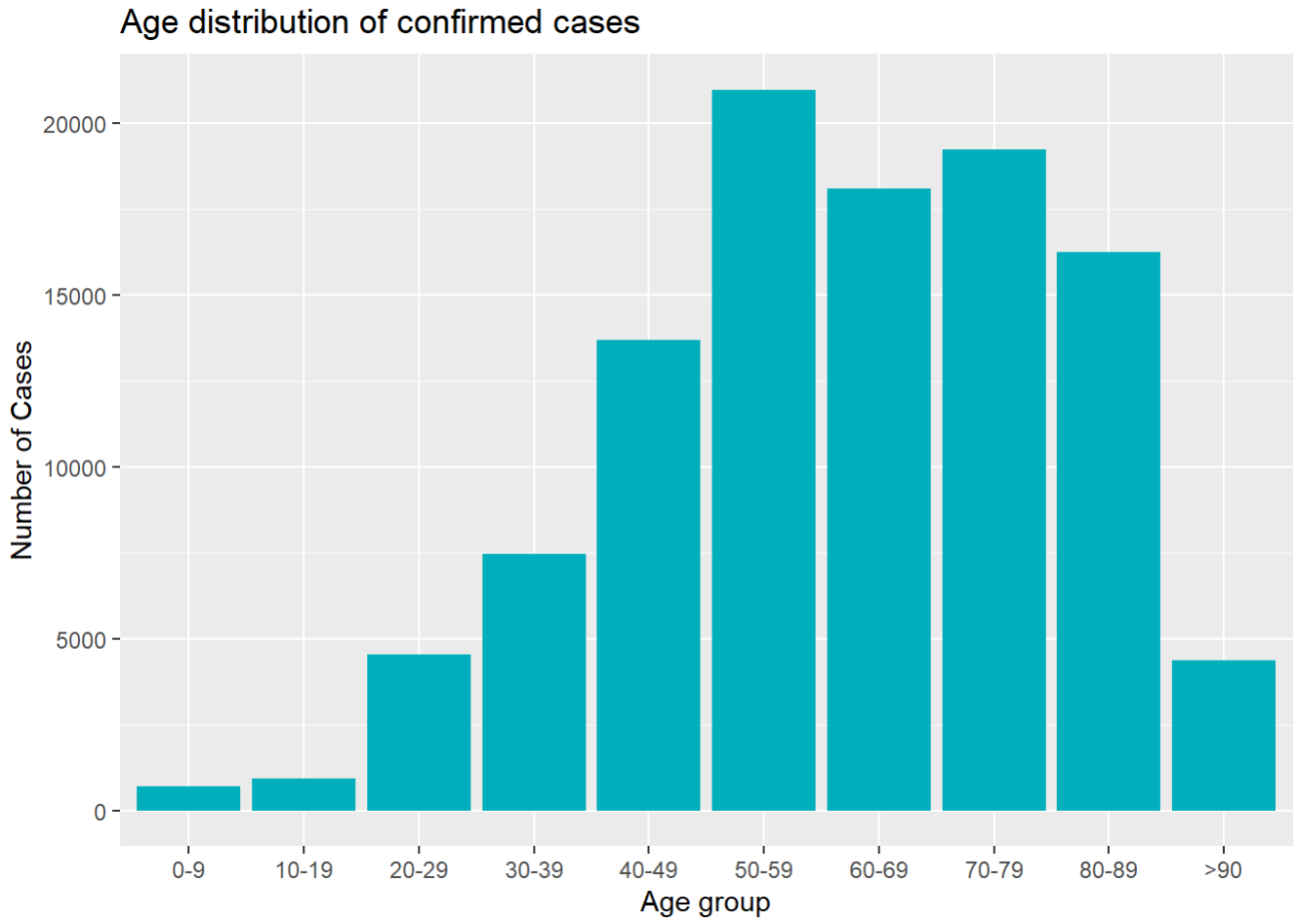
1. Novel Coronavirus COVID-19 Italy Dataset (https://www.kaggle.com/virosky/italy-covid19)
2. COVID-19 in Italy (https://www.kaggle.com/sudalairajkumar/covid19-in-italy)
3. Novel Corona Virus 2019 Dataset (https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset)

The analysis is divided into further parts based on the following factors:

1. Gender
2. Area
3. Age

3. Analysis by Age

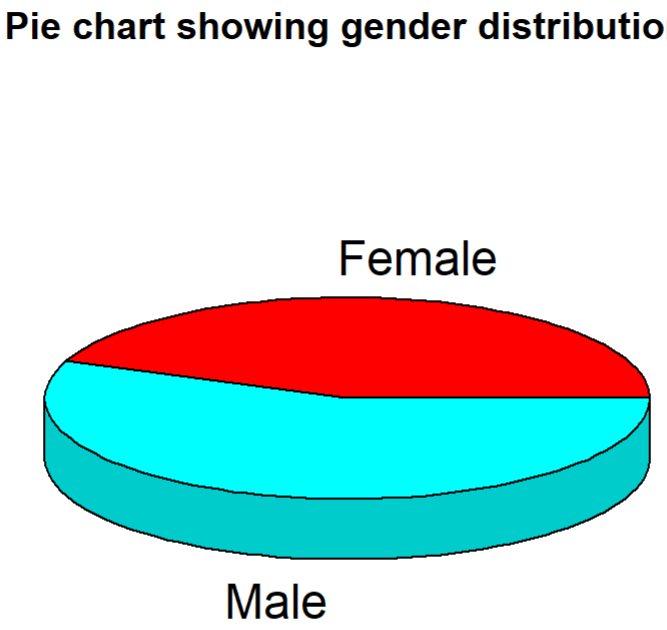
```
# Loading the data
covid_italy <- read.csv("data/covid-age.csv")

# exploring the data
#head(covid_italy)

#dealing with unknown values
covid_italy <- subset(covid_italy, covid_italy$age_classes != "Uknown")

covid_italy$age_classes <- factor(covid_italy$age_classes, levels = c("0-9", "10-19", "20-29",
"30-39", "40-49","50-59","60-69","70-79","80-89",">90"))
```



As we can see the most affected age group in Italy is 50 to 59 years, followed by 70 to 79 years. So we can clearly see that old and middle aged people are more likely to get the virus. The yongest and oldest age group has not been affected much as the other ones. This could be because in any situation both age groups are considred fragile and they have less immunity, so people care extra for them. There are less cases confirmed of younger groups. The other reason for having lesser cases in younger people might be due to the fact that Italy has the oldest population across globe by count. According to EU statistics Italy has the lowest percentage of young people.

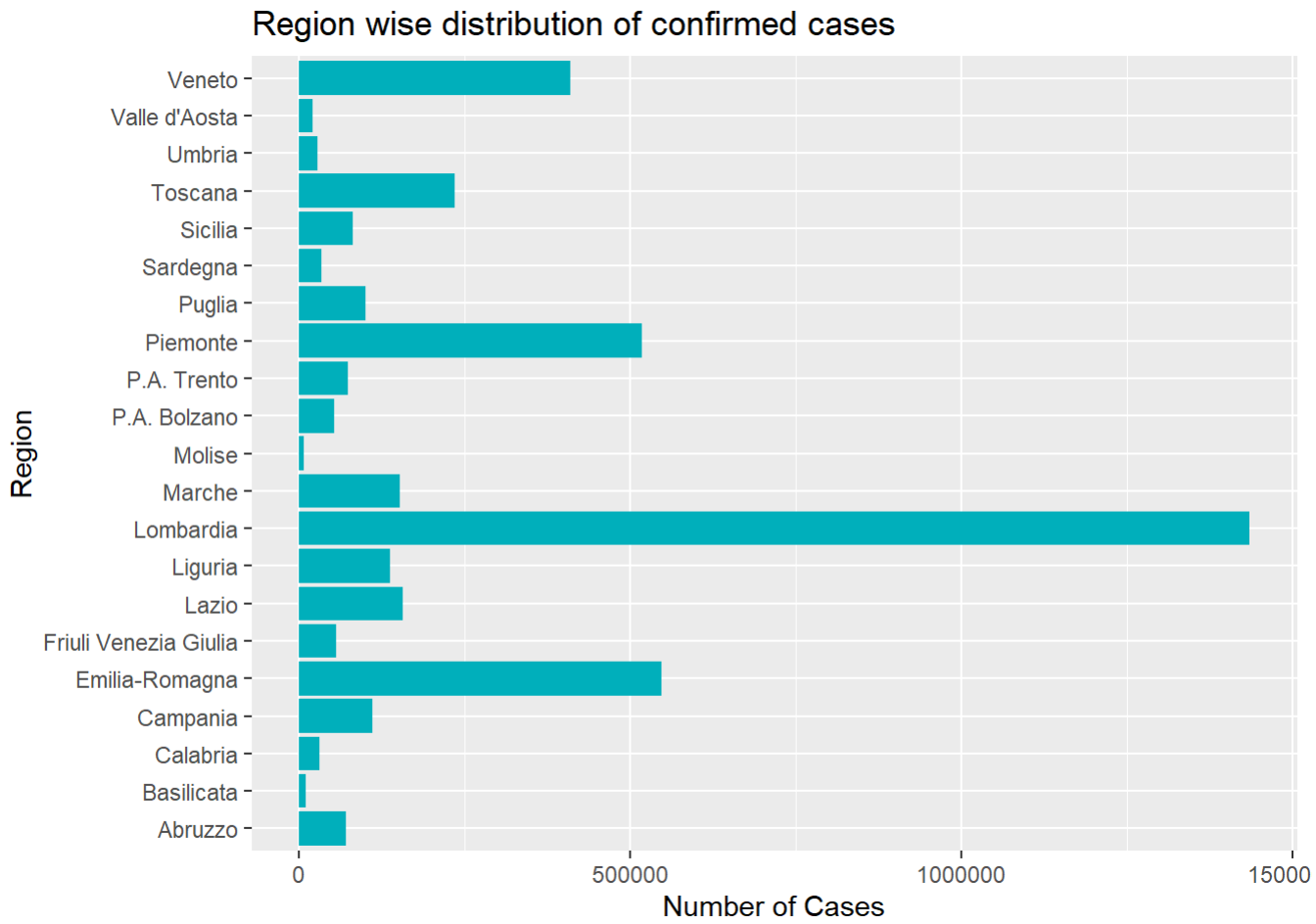
4. Analysis by Gender

```
# loading the data for Gender analysis
covid_gender <- read.csv("data/COVID19_open_line_list.csv")

# fetching the gender counts
gender <- c(table(covid_gender$sex))
female <- c(gender["female"] + gender["Female"])
male <- c(gender["male"] + gender["Male"])
gender <- c(female, male)
lbls <- c("Female", "Male")
pie3D(gender, labels = lbls, main="Pie chart showing gender distribution")
```

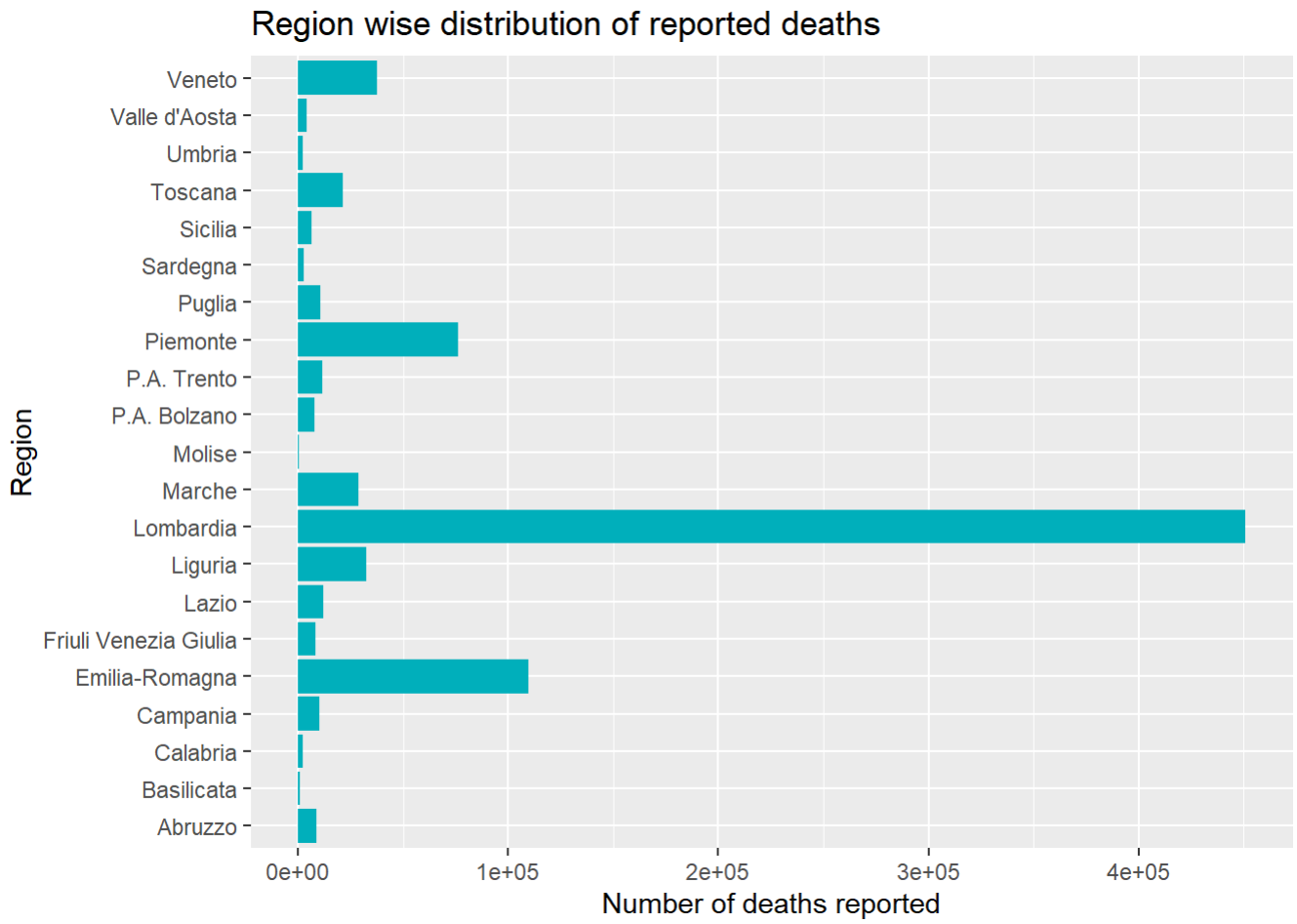


This pie chart suggests that man are more likely to die from coronavirus than women. Researchers had found that the infection rate among men and women is the same but the death rate among men is 2.8% as compared with 1.7% for women. As there are a few reasons men are more likely to die from the new coronavirus. As because Women have a heightened immunity system response. Scientists have explained the reason for more numbers of infected people from Coronavirus in men over women. As in china almost 50- 80% of men do smoking whereas only 2-3% of women do smoking which affects the respiratory system of men over women.

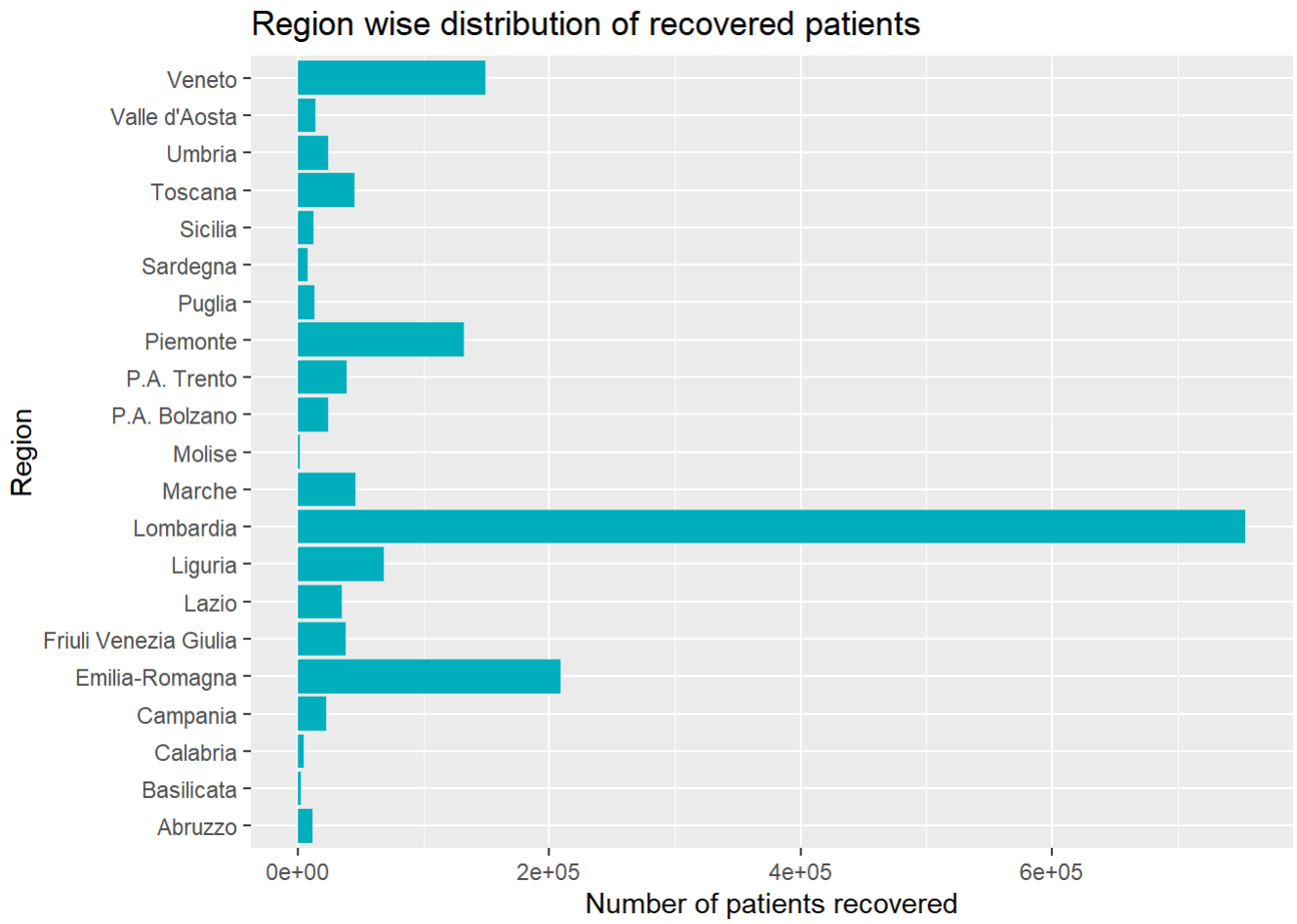
5. Analysis by Area



The plot supports what we have heard in the news and over all reality. Lombardia region has the most number of cases reported, followed by Emilia-Romagna. Now let's see number of deaths reported per area



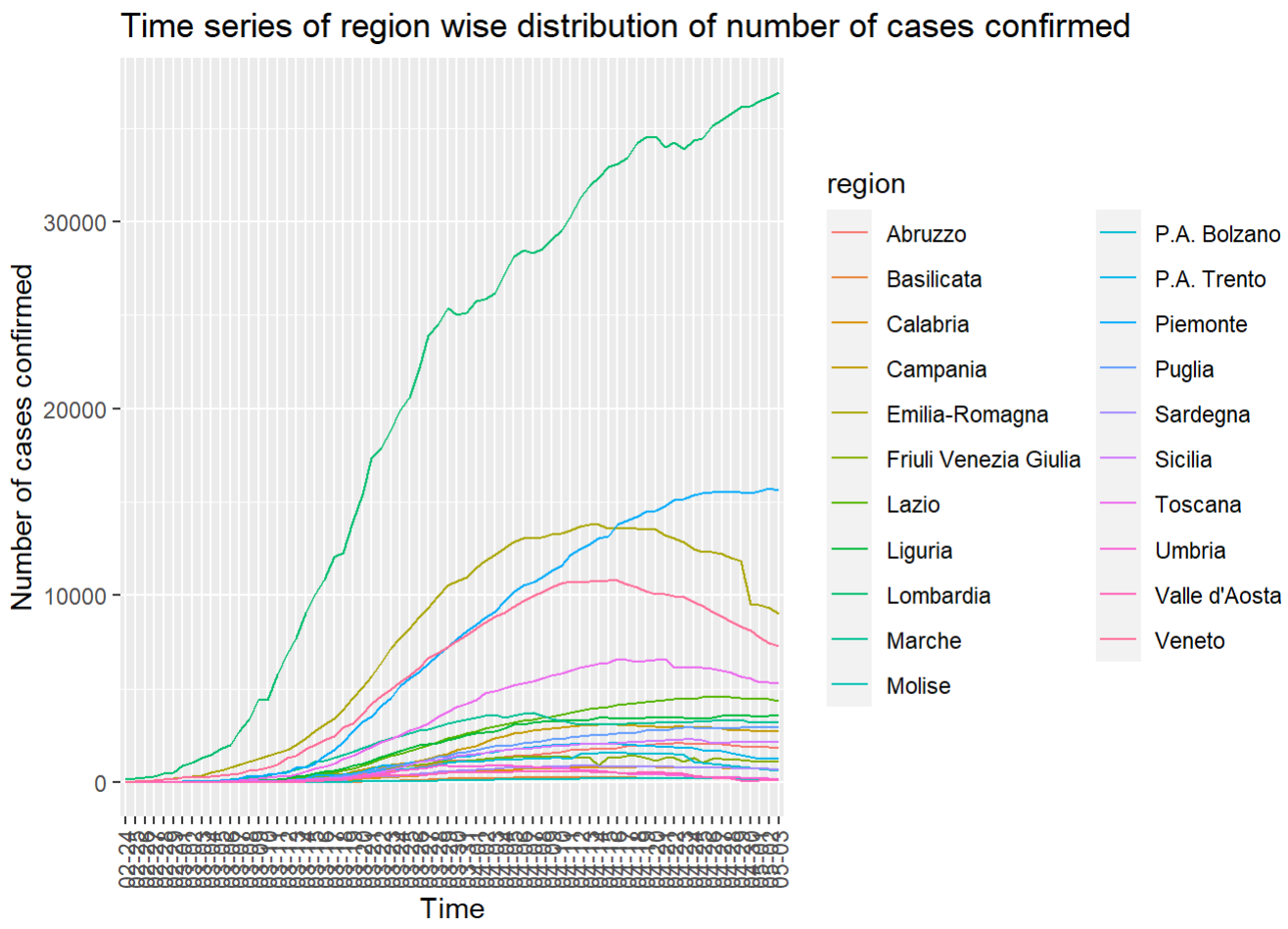
Number of deaths reported go linearly with the number of cases confirmed, as seen in the above plot



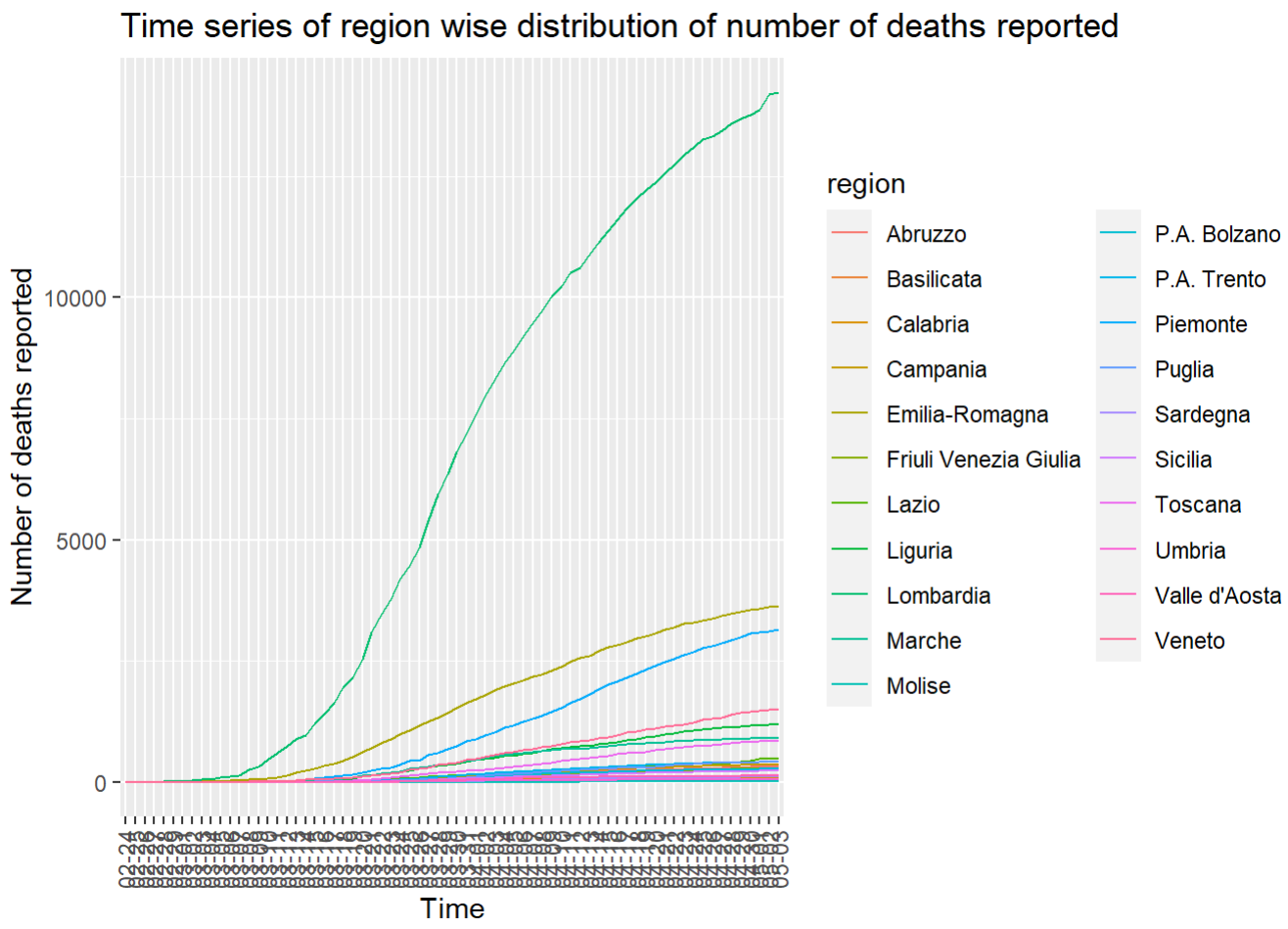
Same applies to the recovered patients, they are proportional to number of cases confirmed so it means that no other factors are affecting, fortunately.

6. Timeseries analysis of confirmed cases

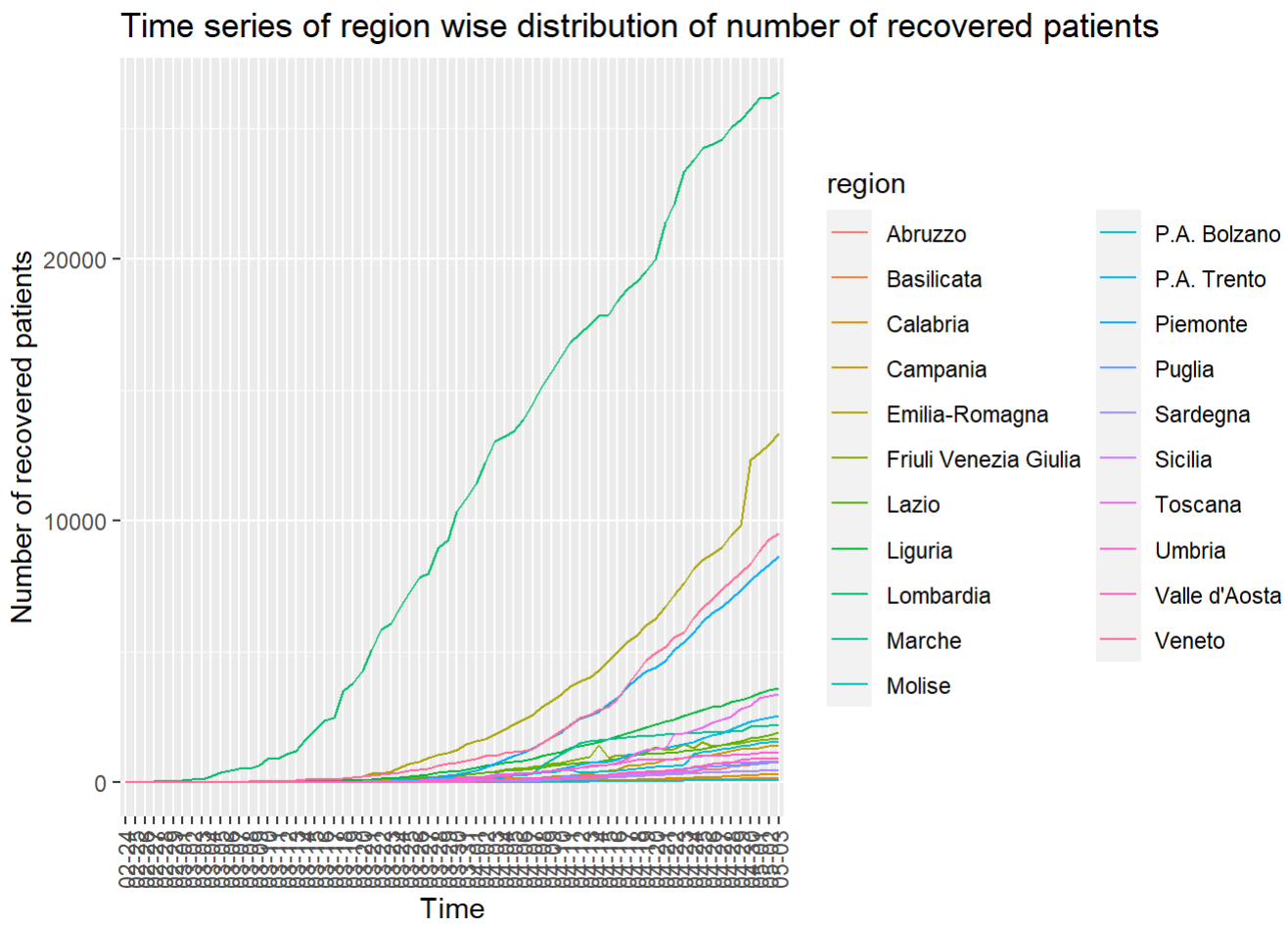
```
# extrating date and month from our time series
covid_region$dm <- factor(format(as.Date(covid_region$date), "%m-%d"))

# confirmed
ggplot(covid_region, aes(x = dm, y = total_confirmed_cases, group = region, col = region)) +
  geom_line() +
  ggtitle("Time series of region wise distribution of number of cases confirmed") +
  xlab("Time") +
  ylab("Number of cases confirmed") +
theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))
```



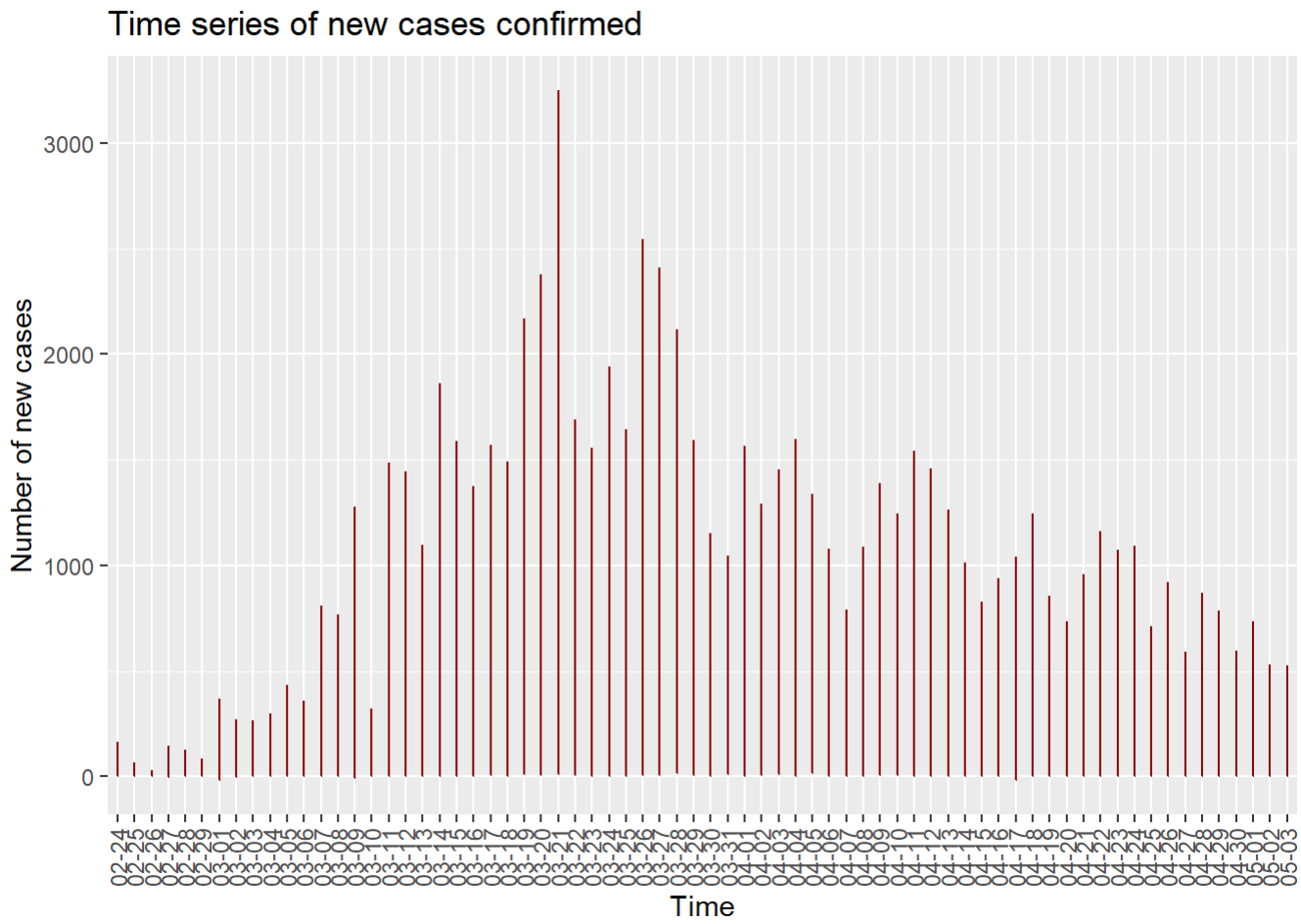

```
ggplot(covid_region, aes(x = dm, y = deaths, group = region, col = region)) +
  geom_line() +
  ggtitle("Time series of region wise distribution of number of deaths reported") +
  xlab("Time") +
  ylab("Number of deaths reported") +
theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))
```




```
ggplot(covid_region, aes(x = dm, y = recovered, group = region, col = region)) +
  geom_line() +
  ggtitle("Time series of region wise distribution of number of recovered patients") +
  xlab("Time") +
  ylab("Number of recovered patients") +
theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))
```




```
ggplot(covid_region, aes(x = dm)) +
  geom_line(aes(y = new_confirmed_cases), color = "darkred") +
  ggtitle("Time series of new cases confirmed") +
  xlab("Time") +
  ylab("Number of new cases") +
theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))
```



As we can see that the curve is getting flattened, which is a good sign.

Prevention:

To avoid the critical situation people are suggested to do following things

- Avoid contact with people who are sick.
- Avoid touching your eyes, nose, and mouth.
- Stay home when you are sick.
- Cover your cough or sneeze with a tissue, then throw the tissue in the trash.
- Clean and disinfect frequently touched objects and surfaces using a regular household
- Wash your hands often with soap and water, especially after going to the bathroom; before eating; and after blowing your nose, coughing, or sneezing.If soap and water are not readily available, use an alcohol-based hand sanitizer.