Fighting Tuberculosis

Start

Tuberculosis

- Tuberculosis (TB) remains a leading cause of death from infectious disease worldwide, with an estimated 10 million new cases in 2019.
- Current approaches to preventing, diagnosing, and treating TB are inadequate. Drug-resistant strains of TB have also emerged, creating a growing sense of urgency to control the spread of the disease.
- We are working to better understand the basic science behind the TB epidemic and to support the development of new tools for prevention, diagnosis, and treatment as well as optimal delivery of TB care.

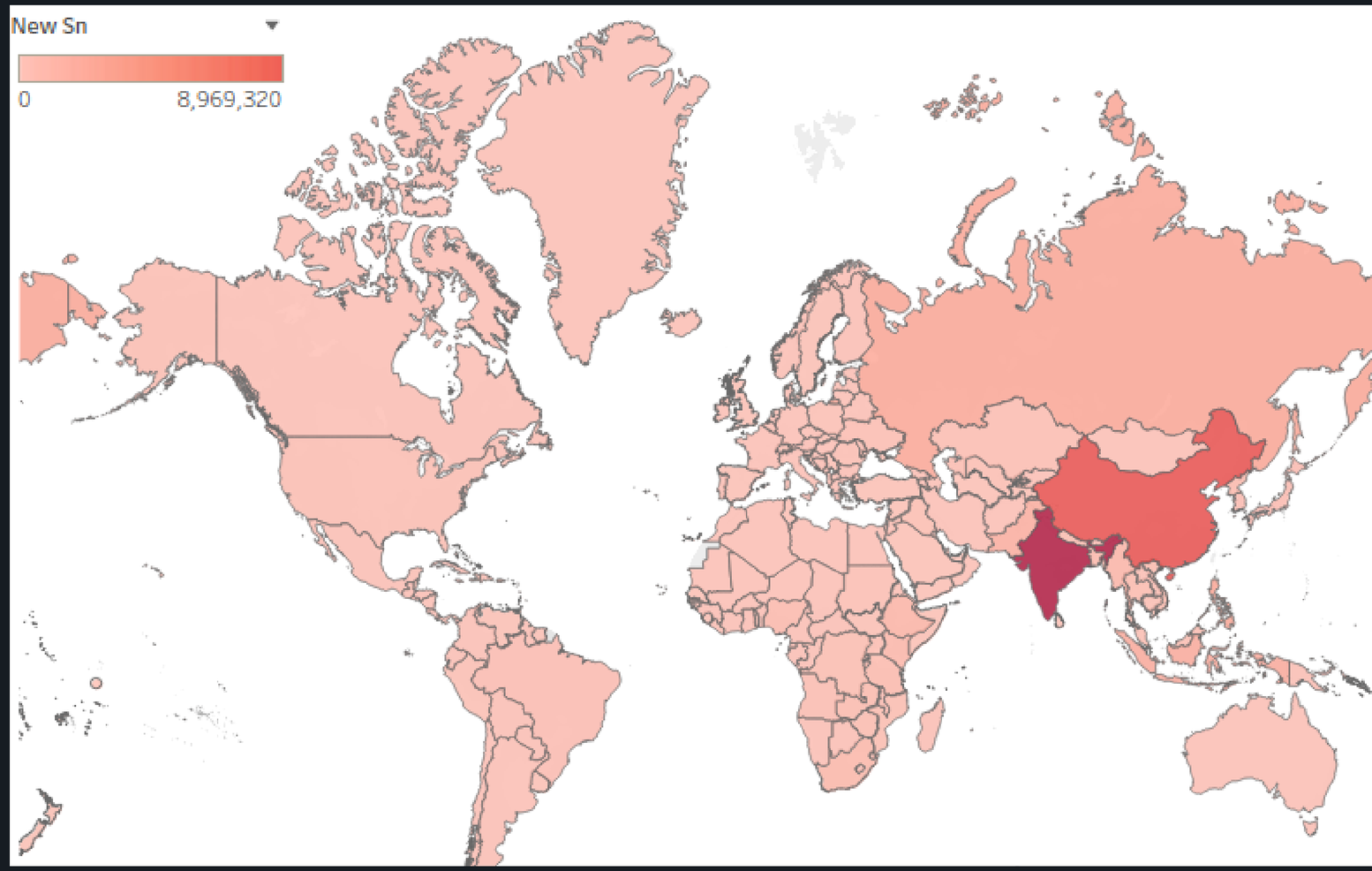
With this background given by Gates Foundation website, let's analyze data about tuberculosis at different countries.

New pulmonary smear-negative cases

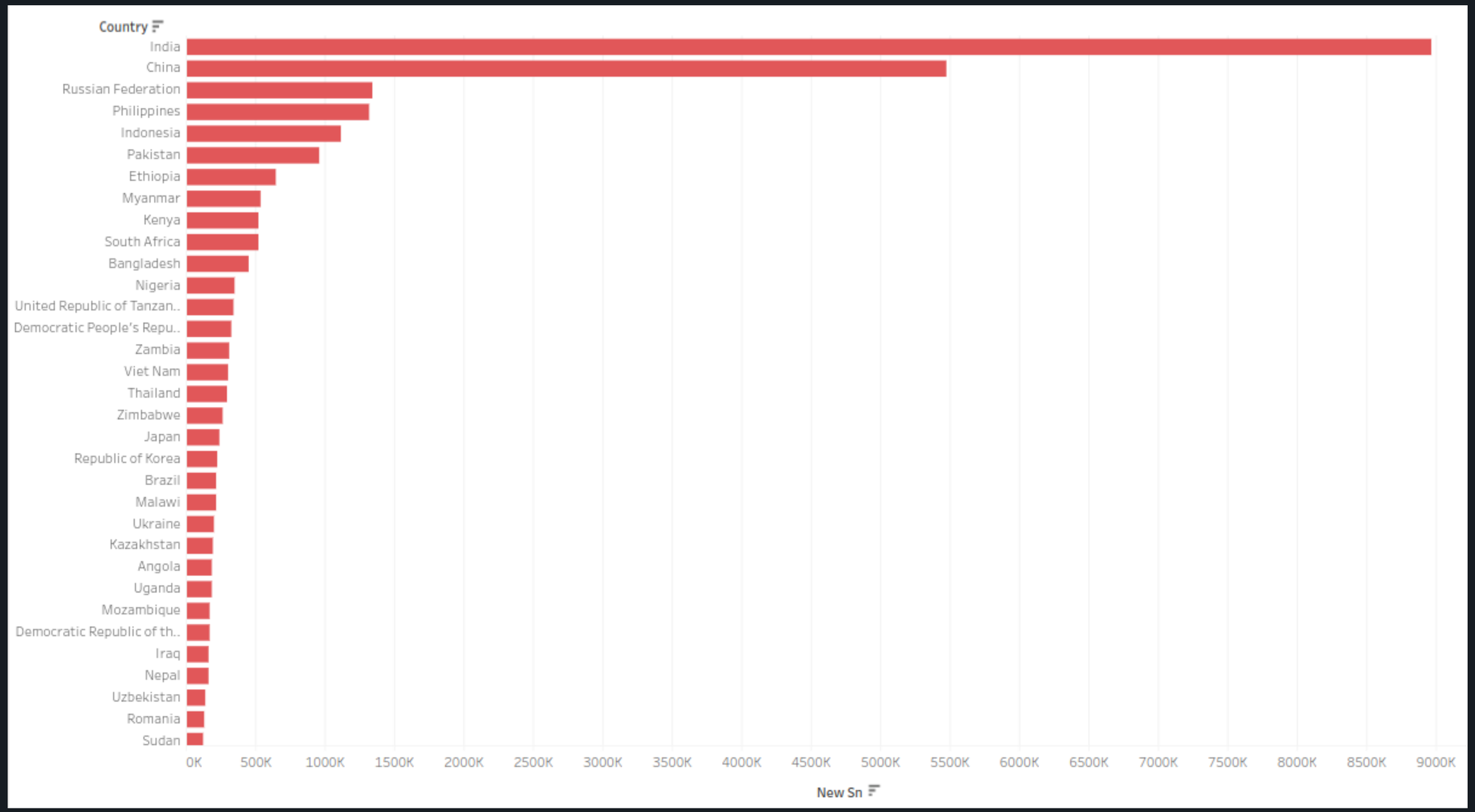
These are the most infectious and most likely to transmit the disease in their surroundings.

At this map we can see the countries that have had more cases during the last years.

These are the countries that require more attention.

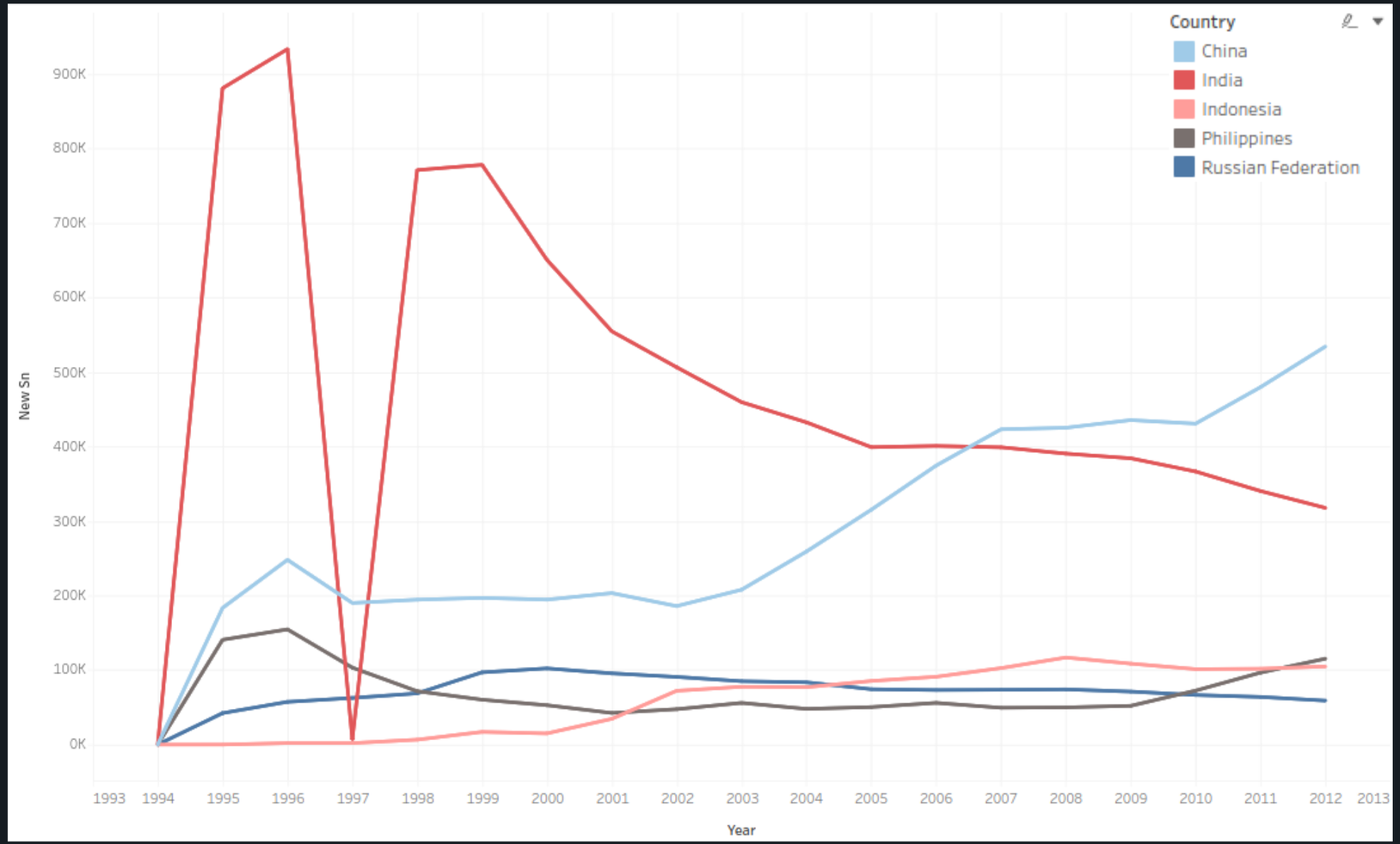


New pulmonary smear-negative cases



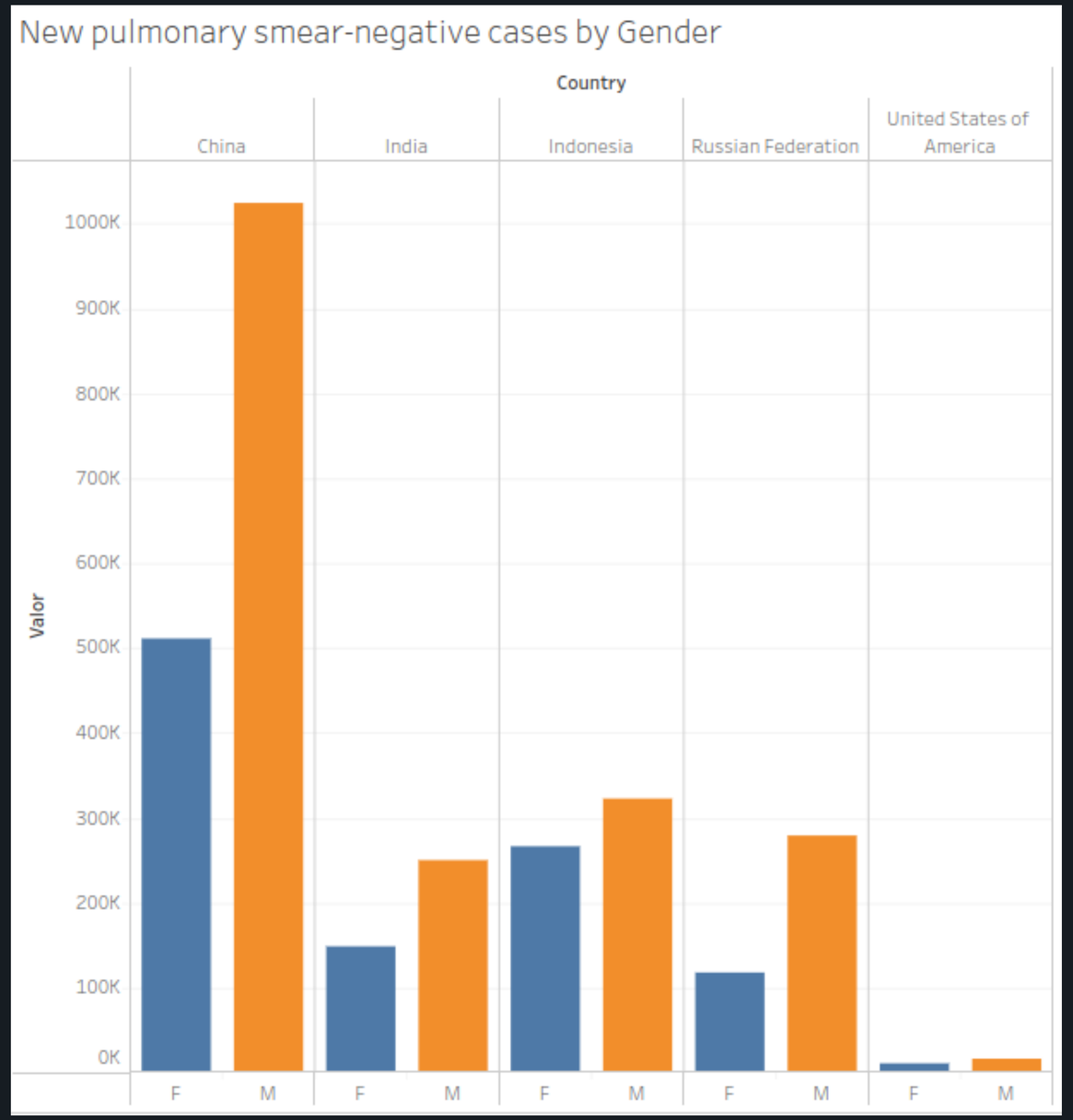
New pulmonary smear-negative cases

Despite of the decrease in the number of cases, almost every country have had a continuous increase in the number of cases each year. India seems to be making a great effort to fight it, but China's cases keep going up.

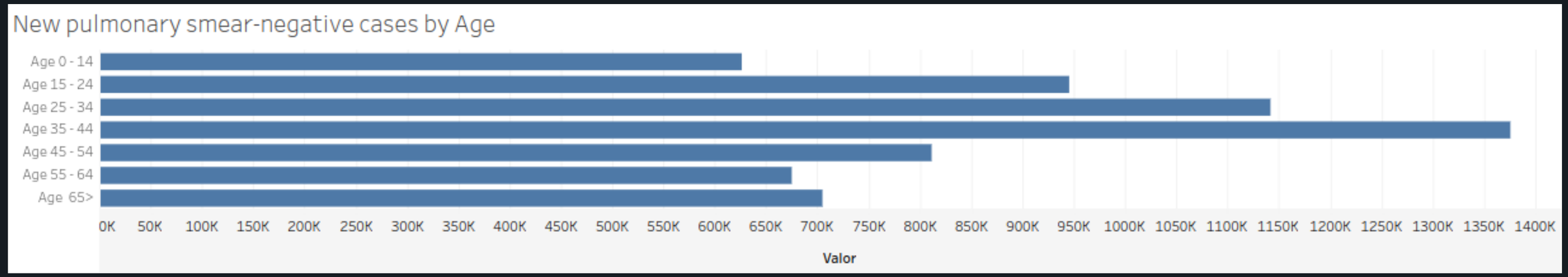


New pulmonary smear-negative cases by Gender

Men are more likely to get the disease compared with women, at least for most of the countries it is a similar situation. United States were included in this chart, and we can observe the same pattern is shown.



New pulmonary smear-negative cases by Age



We can see an increase in the inffected patients as they become older, reaching the highest probabilities to get the disease at ages 35 to 44, after that group of age the probabilities are lower, but at 65 years the probabilities increase again.

What data science programming language should we use moving forward?

I'm not very familiarized using R, so I would say using Python is the best option, first, because I've already used Python before, second, since there are many repetitive tasks, a Python's for loop will be a good option.

While trying to manipulate the data I used an open-source MS Excel equivalent, it stopped working when trying to change many values, Python wouldn't stop in a similar situation and will do the work correctly.

I haven't used R before so, maybe if I learn more about R and apply it to projects my opinion could change.
SQL is not a programming language I would like to use for these kinds of projects.

Appendix

When manipulating the data I had to:

- Change NA values to 0s (number zero)
- Created fields to sum totals of different columns
- Created age groups, since there where many fields about age groups I had to be careful about it
- Data was disorganized
- Columns name's were not very useful, I had to go to the definitions every time before manipulate it

Fighting Tuberculosis

End