Data exploration of the COVID-19 cased in the State of São Paulo, Brazil

Rodolph B Mendes   
School of Computing, Higher diploma in Data Science  
NCI. National College of IrelandDublin, Ireland  
x19213298@student.ncirl.ie

*Abstract*—The current COVID-19 crisis in Brazil has been increasing without much intervention of the State.  
Observations of the data from that area can bring a perspective to which point the cases can come when the interactions are not restricted.

Keywords—COVID-19, Brazil, São Paulo

# Introduction

This report hopes through analysing COVID-19 data from the states of São Paulo, Rio de Janeiro, Minas Gerais, Ceará. Bahia and Manaus to get a broader view and understanding of the current situation of the level of spread of the disease. We hope on this report to (1) Through the exploration of the data, have an overview of the level of infection in those said states and between them (2) Observe the variables and observe the relations between them (3) Using time series for not only visualization of the progress of the total daily deaths, but also creating a prediction tool (4) and finally through the use of linear regression, try to create a statistical prediction tool for future cases.

# General observation of the Brazilian states

## The dataset

The used information was extracted from Brasil.io [1], a website that have been aggregating all COVID-19 data from Brazil when the official data do not seem to be reliable as described by the media [2] and the creation of such websites outside of the government reach was deemed necessary.

## The sample size used

Due to the amount of data, we favoured a sample utilizing the states of the 5 biggest cities in Brazil. It was observed that the dataset for the Distrito Federal state, where the city of Brasíllia is, was somehow incomplete, not to mention that the state is similar to the Washington DC in the USA, so the use of the data would overcomplicate the results, that being said, it was decided to include two other states instead, the state of the 6th biggest city, Minas Gerais and the state of Amazonas, which was one of the most affected during the pandemic.

# Preparation of the data

The data obtained from the Brasil.io website had to be downloaded in different files, since a single csv file could not hold all the entries at once. For the preparation of the data, first it was needed to use an encoding, since Brazilian cities use a great variety of special characters and the CSV format cannot read them, it was used the UTF-8 encoding when uploading the files.

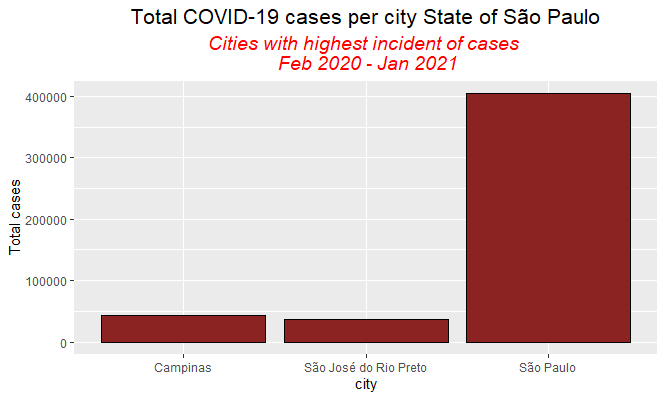
This was followed by a first subsetting from Minas Gerais and Bahia states, since the dataset that contained their data also contained the data of some other states and this first subsetting aimed to have unique variables for all the states.

Next, all data from NA values was excluded from the dataset. From observations on the datasets themselves, just a few cities had blank entries and a few cities altogether had no name, so removing those did not pass 5% of the total amount of entries, which was acceptable.  
To exclude the blank entries, those were converted into NA’s and then removed. Further subsetting was done after the exclusions to be sure all the data had only the required states.

# Exploring the states data

## Identifying the cities with the highest number of total cases

By aggregating the daily data, we could observe and compare the number of cases per state. We began with the São Paulo State with a simple observation regarding absolute total number of cases:



**Figure 1.** *total COVID-19 cases per city, State of São Paulo.*

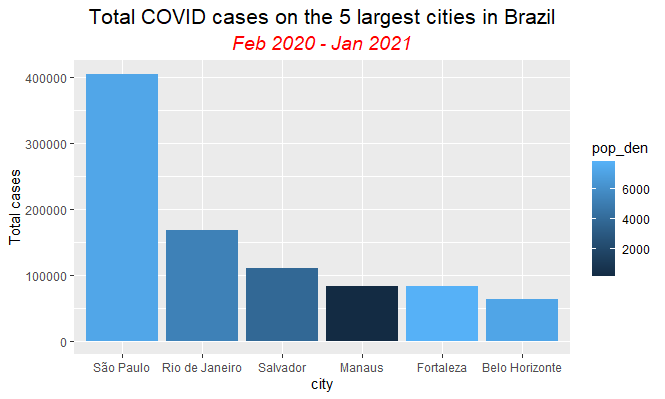
By observing the huge difference of the cities and the capital, it was chosen to the compare the total cases between the largest cities in the country as well as Manaus, who was heavily afflicted by the spread. Before we analyse graph, let us give some perspective regarding the cities actual size.

## As we can observe below, according the latest data from IBGE [3] the Brazilian institute of Geography and Statistics, we have the city of São Paulo as the biggest in population, but not as size, which goes to the Manaus city, which in turn have a small populational density due the fact that there is a great green area around the city.



**Figure 2.** *Comparison between cities: Population, Area and populational density. Source: IBGE(2021)*

## With that we can observe the below graph:

 **Figure 3.** *Comparison between cities: Total cases and populational density*

From the graph we can observe that due the total population, São Paulo was the most affected, but we can see that the city of Manaus, even being with the smallest population and populational density, was greatly affected. We can further observe this as below:



**Figure 4.** *Comparison between cities: Affected % population*

We can see here that from all the cities, the city of Manaus was the one most affected in term of total population, with a rate of infection in their population close to 4%.

## With those observations, it is clear that the total population is not a good measure when we talk about rate of infection in the cities, so we further explored the cases using by basis the cases by 100.00 inhabitants, which puts them all in the same level of comparison.

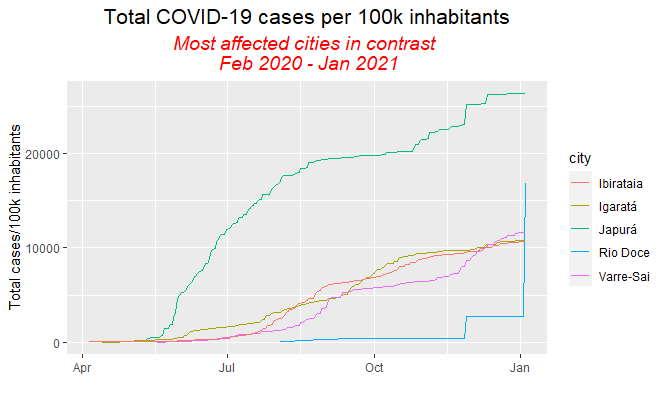
## Cases by 100.000 inhabitants

To observe the total amount of cases by 100.00 inhabitants, we used two different approaches from the data. First, was looking for the absolute maximum, which returned what city, independent of the data, had most cases for each of the studied states. To compare and see if there was any evolution, we also looked on the last day.

Based on that information, we gathered a list of the most affected cities per state, as we can see below:

 **Figure 5.** *Comparison between most affected cities per 100.000 inhabitants. Affected % of population*

It can be observed a worrisome percentage of the population of Japurá affected by the pandemic. The Amazonas state is a well known state not only for the Amazon rainforest, but for being poor, with former slave communities called “Quilombolas” and indigenous communities which have little to no access to formal education and basic healthcare and assist on inflating the number for the state.

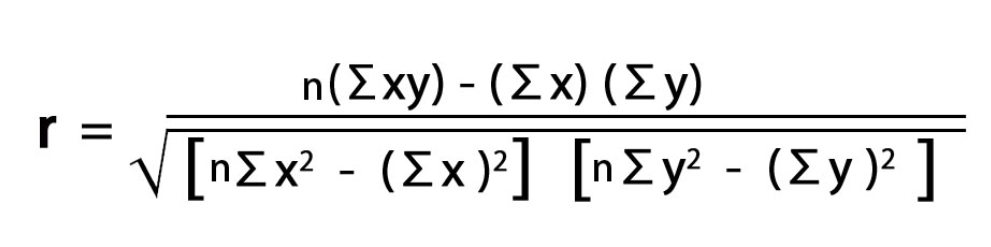
 **Figure 6.** *Infection curve, most affected cities per 100.000 inhabitants*

We can also observe the curve of the mentioned cities above. We can see that for Japurá, the curve is in fact very different from the other observed cities, that grew on a more stable curve that Japurá. For Rio Doce, we can see that cases were sub-notified until they were released all at once, causing the number to explode overnight.

# Observing relationships between the variables

## Hypothesis: Estimated population is a good predictor for confirmed cases per 100.000 inhabitants

In order to better understand our variables, we will test for the correlation between the Estimated population and the confirmed cases per 100.00 inhabitants. Since we have a large number of states to analyse, for the correlation we chose São Paulo, for being the biggest and Amazonas, due the level of the spread in the state.

 **Figure 7.** *Correlation formula. WallStreetMojo, 2021[4]*

The correlation coefficient (r) can be from -1 to +1, where they represent a perfect positive correlation at +1 (the two variables increase on a scale of 1 to 1) or perfectly negative (when one variable increase, the other decrease by the same amount), with a correlation at 0 showing no correlation between the variables.

First, to understand what method we would use, we have to test the two datasets and see if they follow a normal curve.

For that we have used two methods, using the Shapiro-Wilk test for normality and by visualizing the data.

After testing for Shapiro-Wilk, the datasets returned as below:

Estimated population São Paulo:

P-value: .00000000000000022

Estimated population Amazonas:

P-value: .00000000000000022

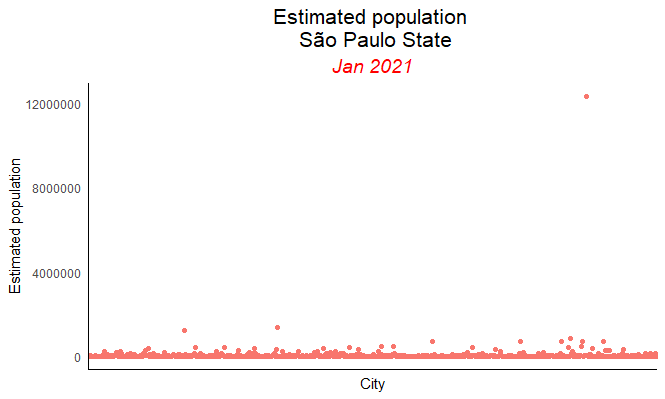
Estimated population São Paulo:

P-value: .00000000000000022

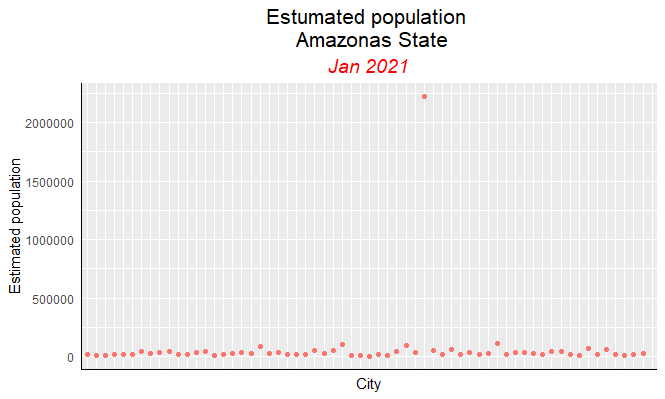
Estimated population Amazonas:

P-value: .00000000000000022

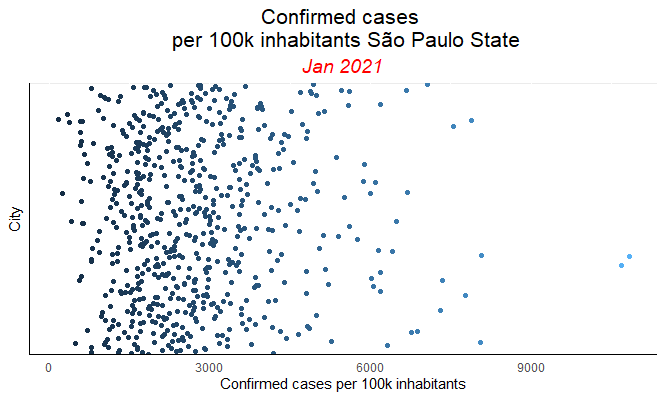
According to the Shapiro-Wilk normality test, we can only consider as normally distributed any values that have the p-value > 0.05, therefore, for all the tested datasets, we will use the Spearman method the correlations, since this method applied better to non-normal distributions. We can confirm this statement by visualizing the data below:



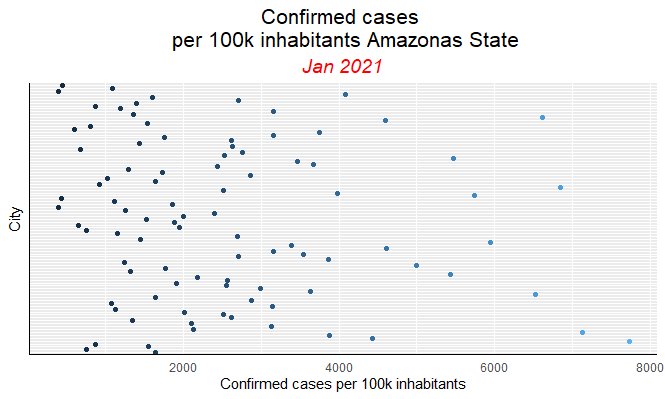
**Figure 8.** *Estimated population curve, São Paulo state*



**Figure 9** *Estimated population curve, Amazonas state*



**Figure 10.** *Confirmed cases per 100.000 inhabitants curve, São Paulo state*

 **Figure 11.** *Confirmed cases per 100.000 inhabitants curve, Amazonas state*

From this we performed the correlation test between the Estimated population x Confirmed cases per 100.000 inhabitants for both states for the below results:

São Paulo State: 0.1983428. With this, we can affirm that there is a weak positive correlation between the estimated population and the confirmed cases per 100.000 inhabitants on the State of São Paulo.

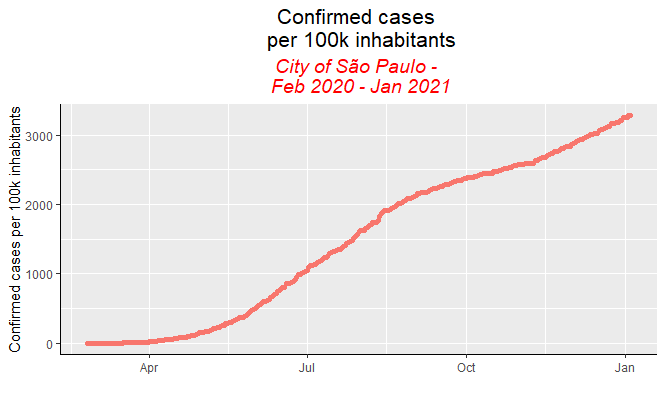
Amazonas State: -0.2219564. We can observe that there is a weak negative correlation between the variables in the Amazonas State.

Conclusion: In general, we reject the hypothesis that estimated population is a good predictor to the Confirmed cases per 100.000 inhabitants for both the States of São Paulo and Amazonas.

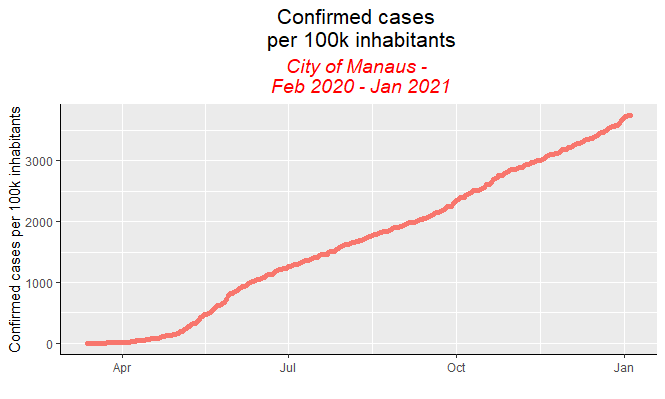
## Hypothesis: Confirmed cases per 100.000 inhabitants is a good predictor for total deaths

Following the same principles we observed on the previous Hypothesis, we will test the variables for normality. Due to limitations for Shapiro-Wilk test can on R, we will only utilize the visual method for this test, since the test can test variables no greater than 5.000 rows.

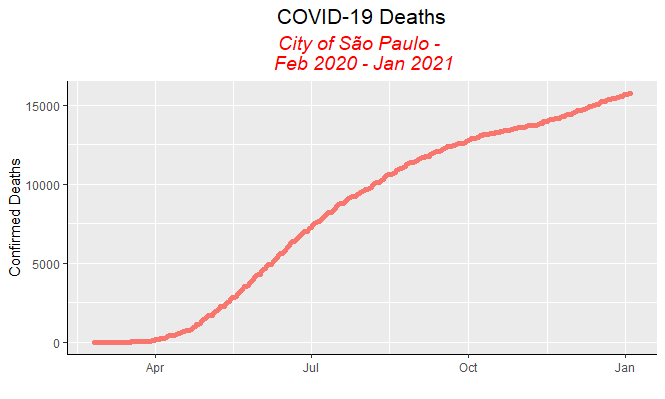
In order to observe a single curve instead of several, we will conduct the correlation with the cities of São Paulo and Manaus, instead of their states.



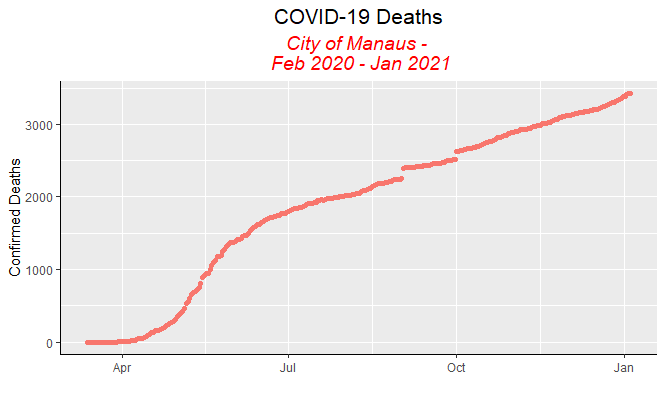
**Figure 12.** *Confirmed cases per 100.000 inhabitants curve, São Paulo city*



**Figure 13.** *Confirmed cases per 100.000 inhabitants curve, Manaus city*



**Figure 14.** *COVID-19 deaths curve, São Paulo city*



**Figure 15.** *COVID-19 deaths curve, Manaus city*

We can observe that even leaning toward normality in a near future, none of the curves are normally distributed, therefore, we will again apply the Spearman method to the correlation.

São Paulo city: 0.9998533. As expected, the relation between the number of cases possesses a strong positive correlation with the total number of deaths, being close to perfectly explaining it.

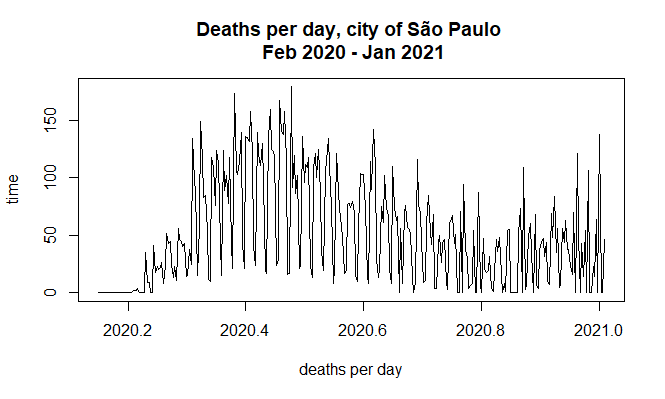
Manaus city 0.9999094. We can see here once again a very strong correlation between the number of confirmed cases per 100.000 inhabitants and the total number of deaths.

Conclusion: We fail to reject the hypothesis that the total number of cases per 100.000 inhabitants is a good predictor for the total number of deaths in both states, as the correlations show that the variables greatly explain each other.

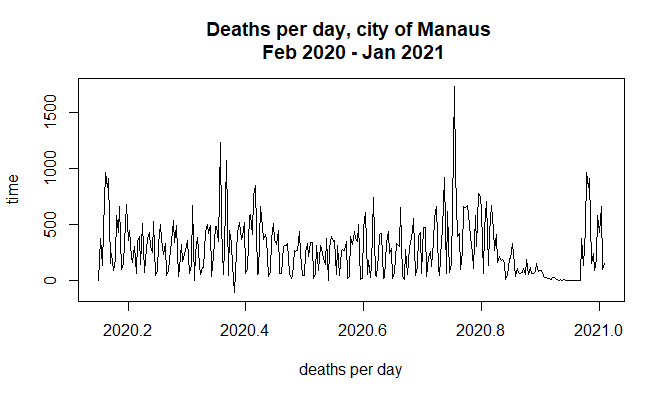
# identifying the sazonality

By utilizing time series, we can not only create graphs that assist us on observing the seasonality of COVID-19, but also act as a predictor for future cases.

Again, we will be exploring the data from the cities of São Paulo and Manaus

At a first moment we can observe the total number of deaths per day in São Paulo city. Even with number in a somewhat constant, we can identify the two waves that have affected the city so far, one in March, which lasted until about August and a second one that started close to November:

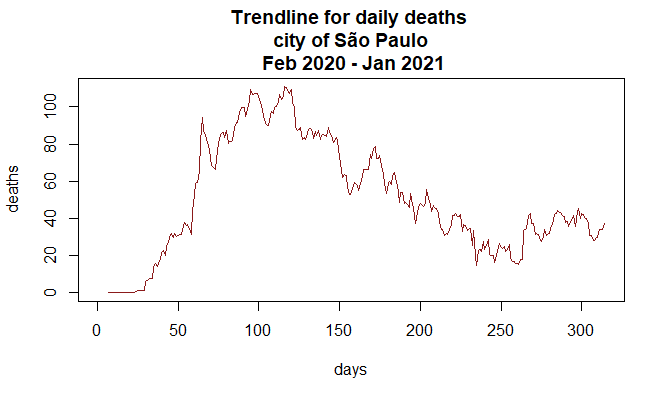
**Figure 16.** *COVID-19 deaths per day, Feb 2020-Jan 2021, City of São Paulo*



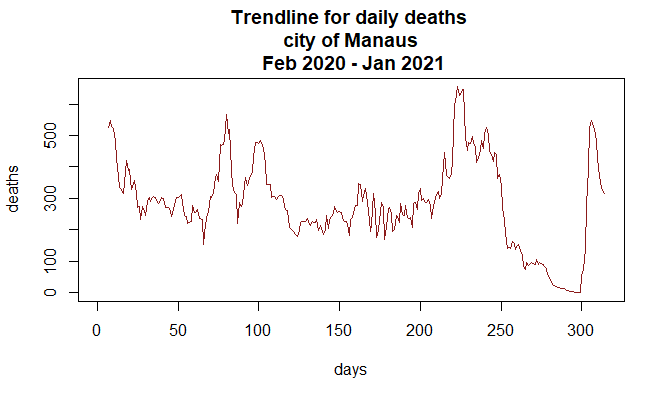
**Figure 17.** *COVID-19 deaths per day, Feb 2020-Jan 2021, City of Manaus*

Differently from São Paulo, we can observe four waves of spread happening in Manaus city, a spike in cases at the beginning of the pandemic, another close to March, a third by July and the last one close to 2021.

To better understand the evolution of the number of deaths, we created a trendline smoothed by the simple moving average, to better visualize the impact in the long term.



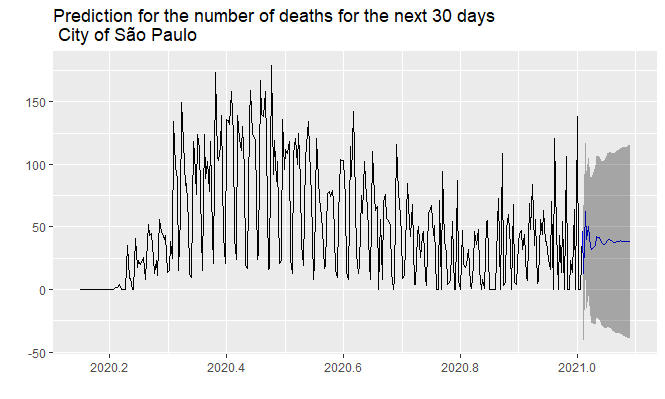
**Figure 18.** *Deaths trendline, Feb 2020-Jan 2021, City of São Paulo*

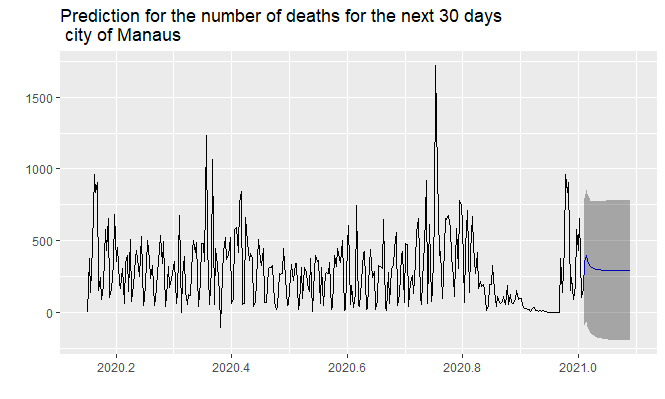


**Figure 19.** *Deaths trendline, Feb 2020-Jan 2021, City of Manaus*

Through the trendlines, the waves that affected each of the cities can be seen more easily.

Finally, by utilizing the exponential moving average, we were able to create a prediction line for each of the states. The graphs were plotted using a confidence level of 95% and are showing the expected cases for the next 30 days:

 **Figure 20.** *Prediction on the number of Deaths for the next 30 days, Feb 2020-Jan 2021, City of São Paulo*



**Figure 21.** *Prediction on the number of Deaths for the next 30 days, Feb 2020-Jan 2021, City of Manaus*

By the two predictions, we can unfortunately expect that the trend on the deaths to keep escalating for both States.

# Predicting the number of deaths São Paulo city

Another statistical tool we can use in conjunction with the correlation we have seen previously is the linear regression, through it, we can create a simple machine learning prediction tool that will allow us, based on the input of a first variable, predict where the second variable will be at that stage.

The regression equation is given by:

ŷ = bX + a,

Where ŷ is the number we are trying to predict, x is the input, b is the slope or inclination of the curve and a is the intercept, or where the curve touches the axis x.

For this prediction tool, we will use the previously observed correlation between total number of cases per 100.000 inhabitants x deaths in order to create this simple predictor.

After creating the linear regression on R, we can observe the following results:

*b = 4.875, a = 1037.526*

*R-Squared: expected to be higher than 0.7. Output: 0.982*

*Adj R-Squared: High value expected. Output: 0.982*

*F-Statistics: High value expected. Output: 1.71e+04*

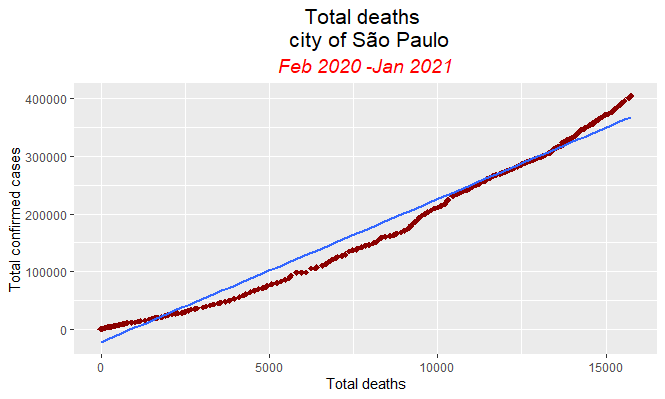
*Std Error: The closest to 0, the better. Output: 0.0003025*

*t-statistics: Need to be higher than 1.96, so p-value is smaller than 0.05; Output: 130.78*

*P-value: Expected to be smaller than 0.05. Output: < 0.00000000000000022*

According to the above results, we can consider that the model is significant and will be good to predict future cases.

We can observe the regression line created from the model below:

 **Figure 22.** *Total deaths prediction tool, Feb 2020-Jan 2021, São Paulo city*

We can apply the created model and from it, make a simple prediction on the total number of deaths when cases get to one million:

ŷ = X\* 4.875 + 1037.52613

ŷ = 1000000\* 4.875 + 1037.52613

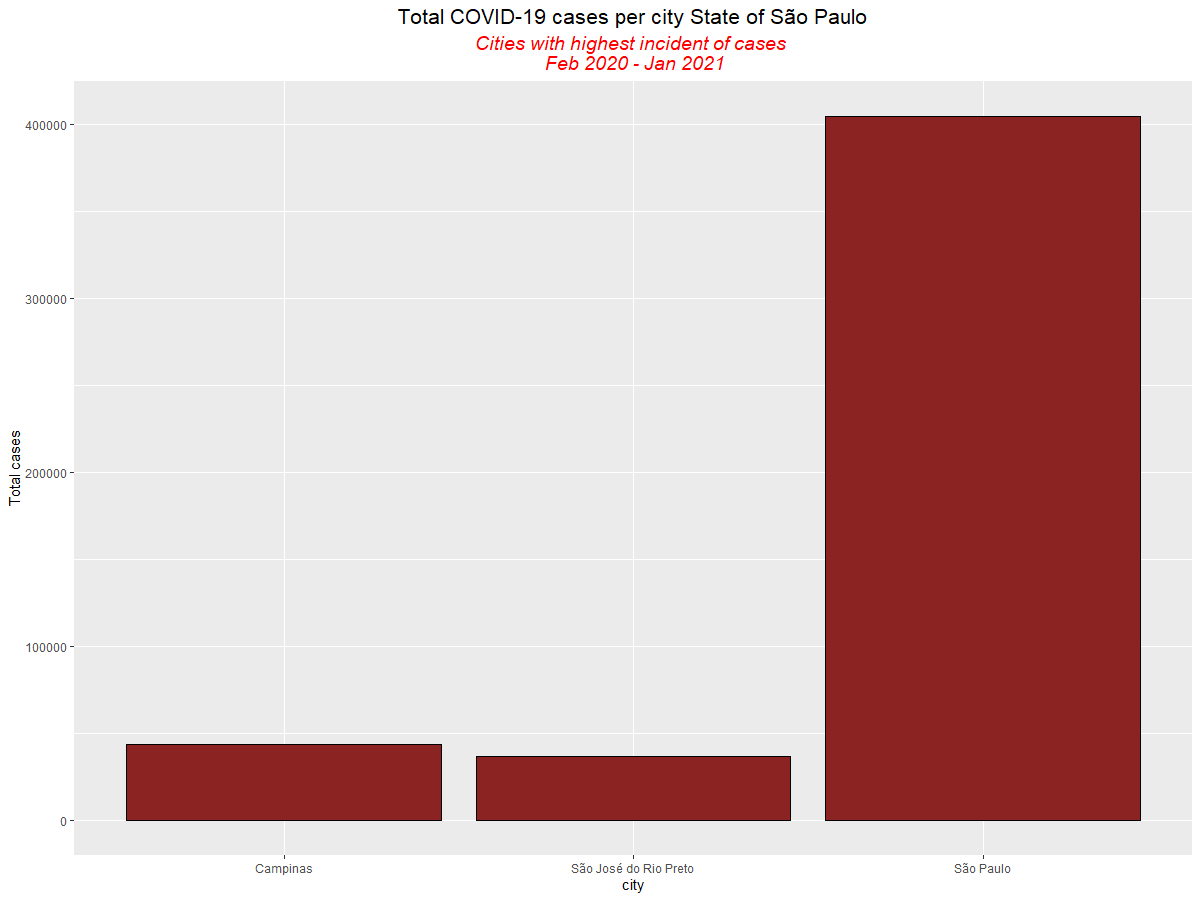
ŷ = 4876037.526

According to our prediction tool, we can expect at million cases, the deaths to come to 4876037.

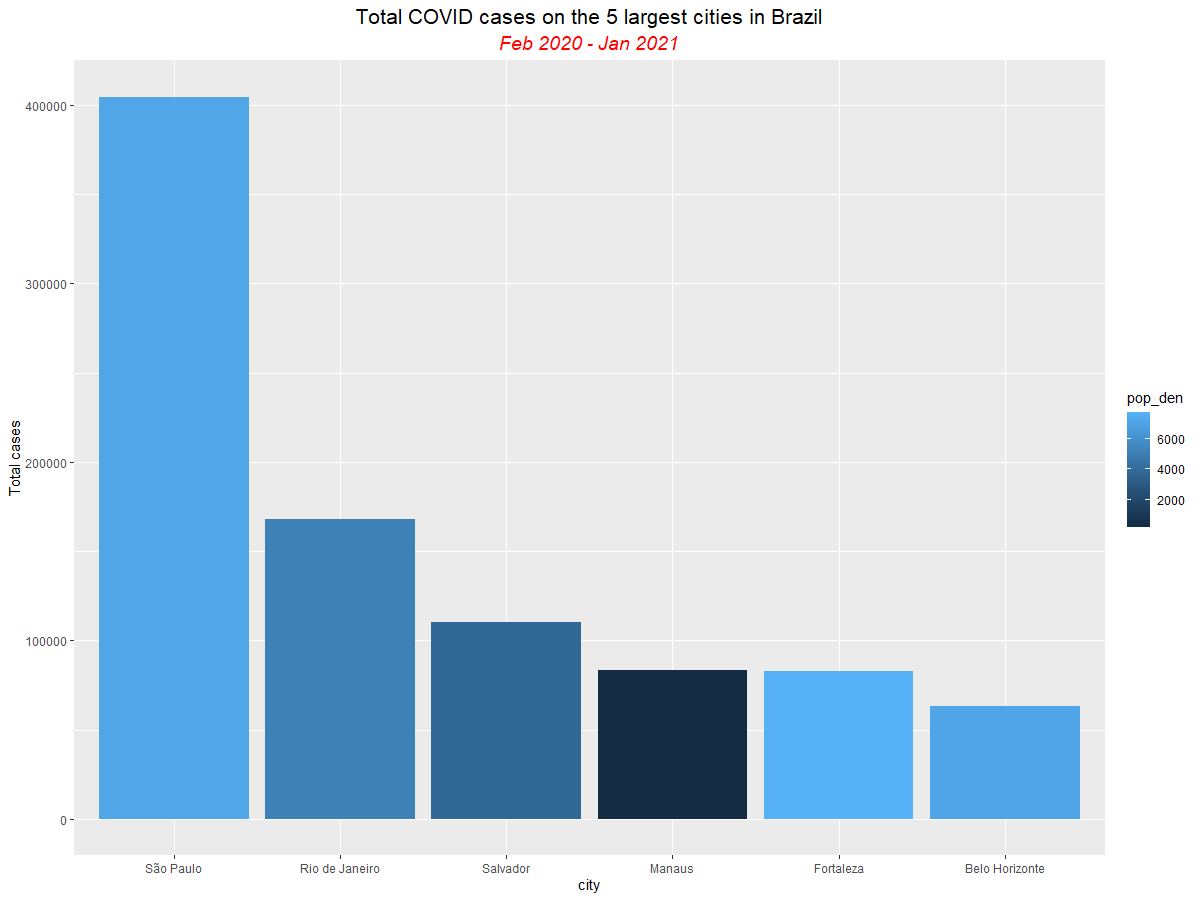
##### References

1. Brasil.io, Dataset: COVID-19, viewed on 31 Dec 2020, <https://brasil.io/dataset/covid19/files/>
2. Daniel Gallas, 2020, Coronavírus: por que números de casos e mortes por covid-19 no Brasil podem estar longe da realidade, BBC, viewed 10 Jan 2021, <https://www.bbc.com/portuguese/brasil-53287455J>
3. IBGE, Instituto Brasileiro de Geografia e Estatística, 2021, Cidades e Estados, viewed on Jan 10th, 2021, < https://www.ibge.gov.br/cidades-e-estados.html?view=municipio>
4. WallStreetMojo, 2021, Correlation Coefficient, viewed 10 Jan 2021 < https://www.wallstreetmojo.com/correlation-coefficient-formula/>
5. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

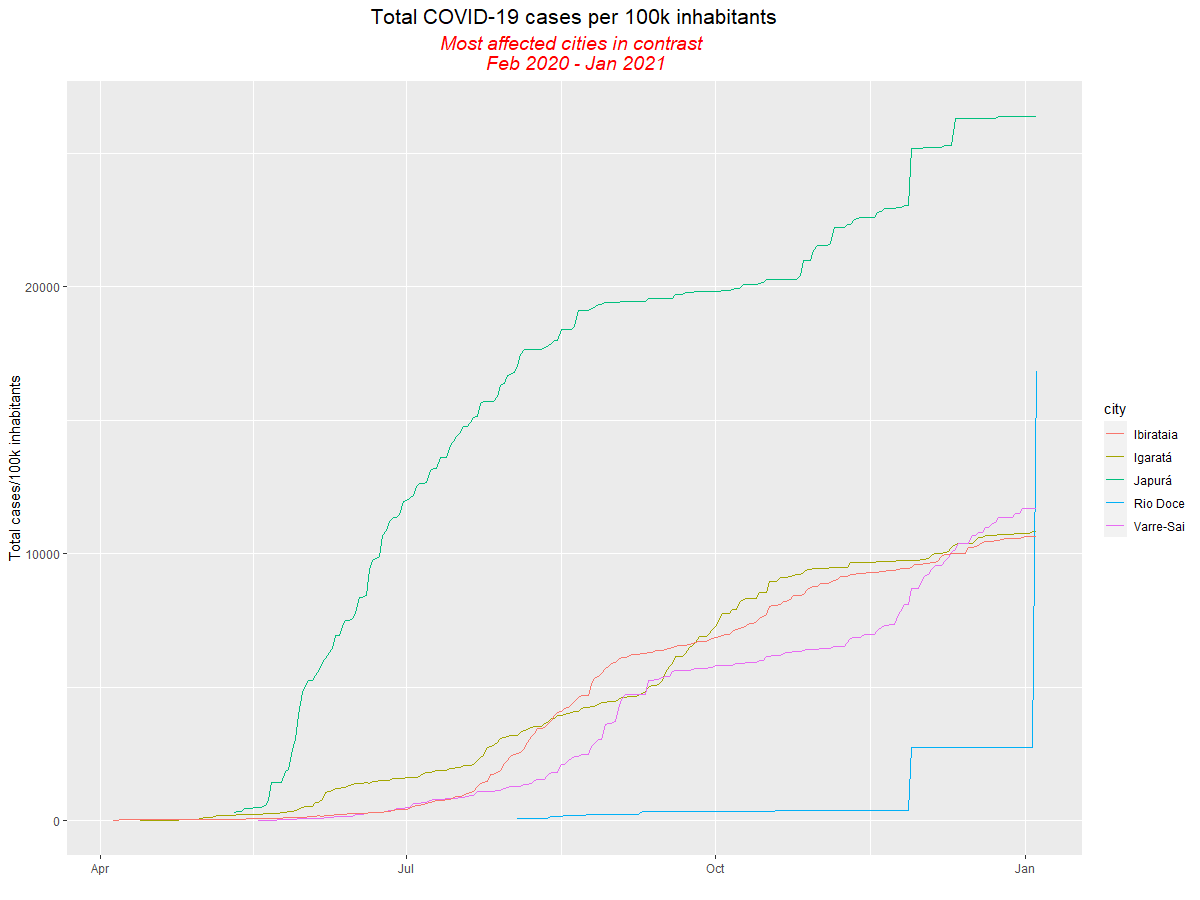
Annexes:



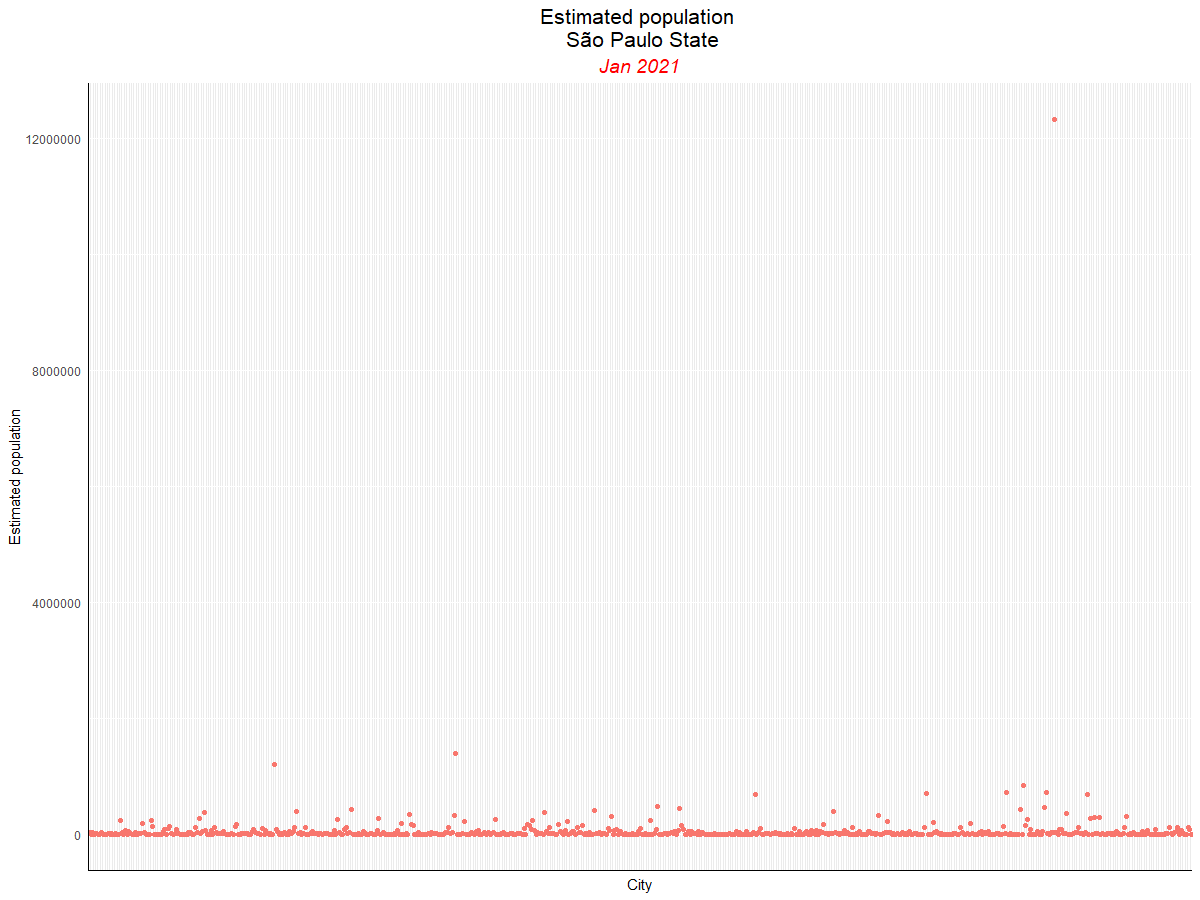
**Figure.23:** Higher definition graph of Figure 1, *total COVID-19 cases per city, State of São Paulo.*



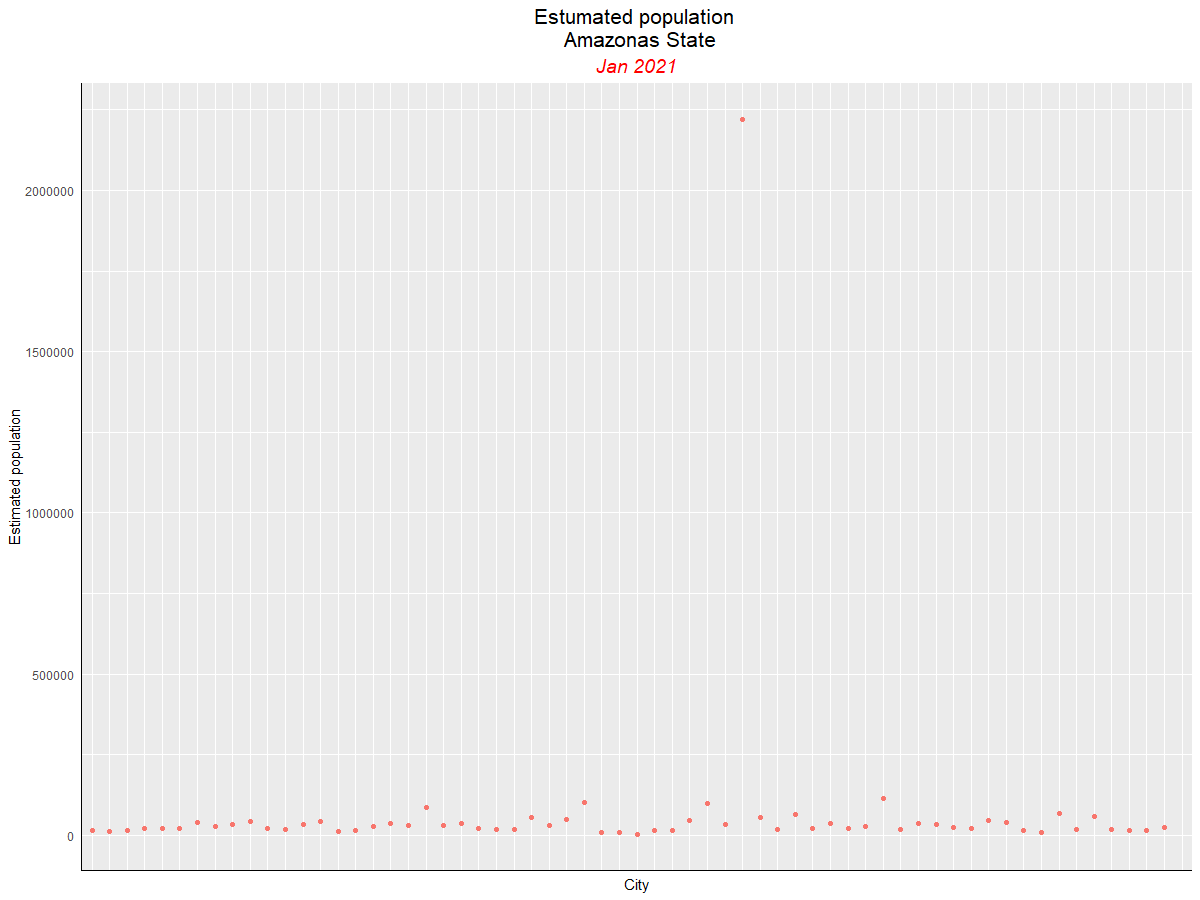
**Figure.24:** *Higher definition of Figure 3.* *Comparison between cities: Total cases and populational density*



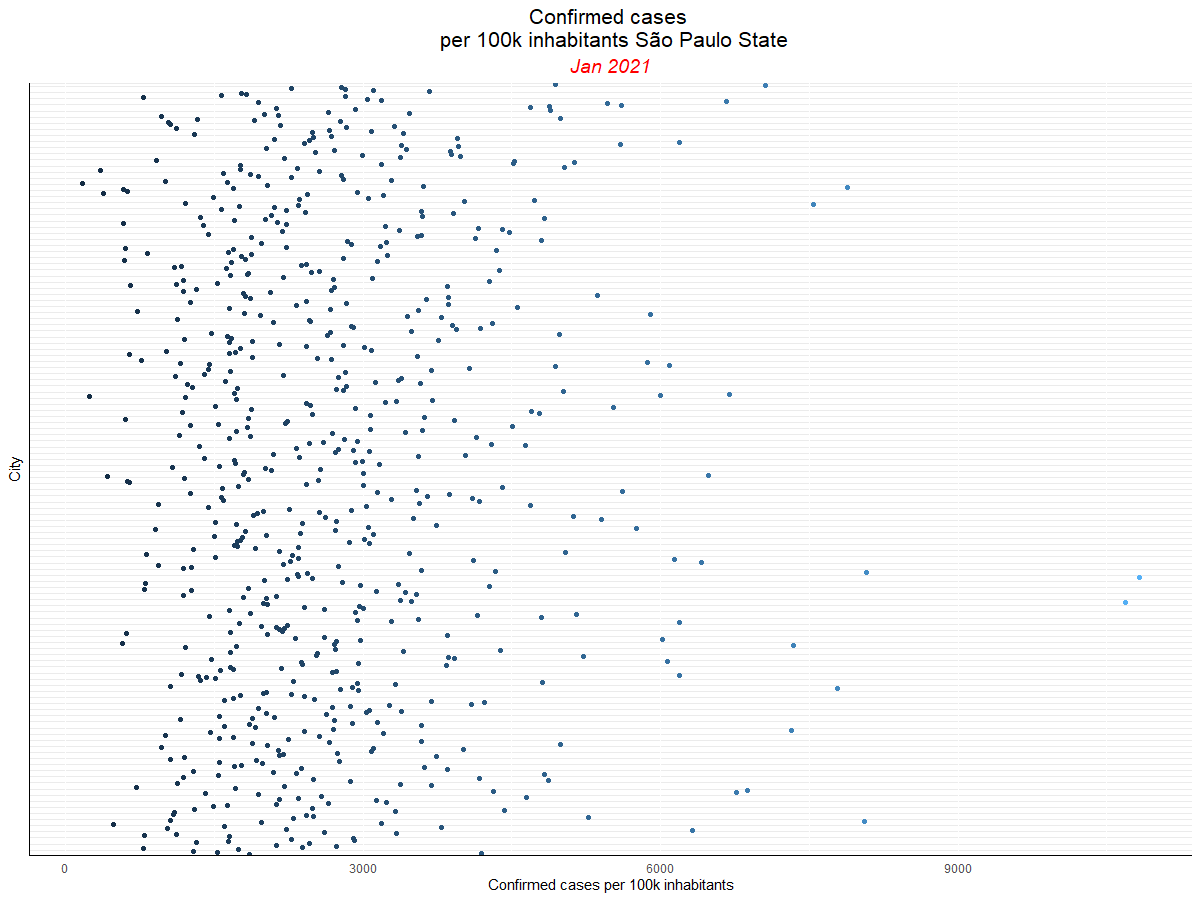
**Figure.25:** *Higher definition of Figure 6***.** *Infection curve, most affected cities per 100.000 inhabitants*



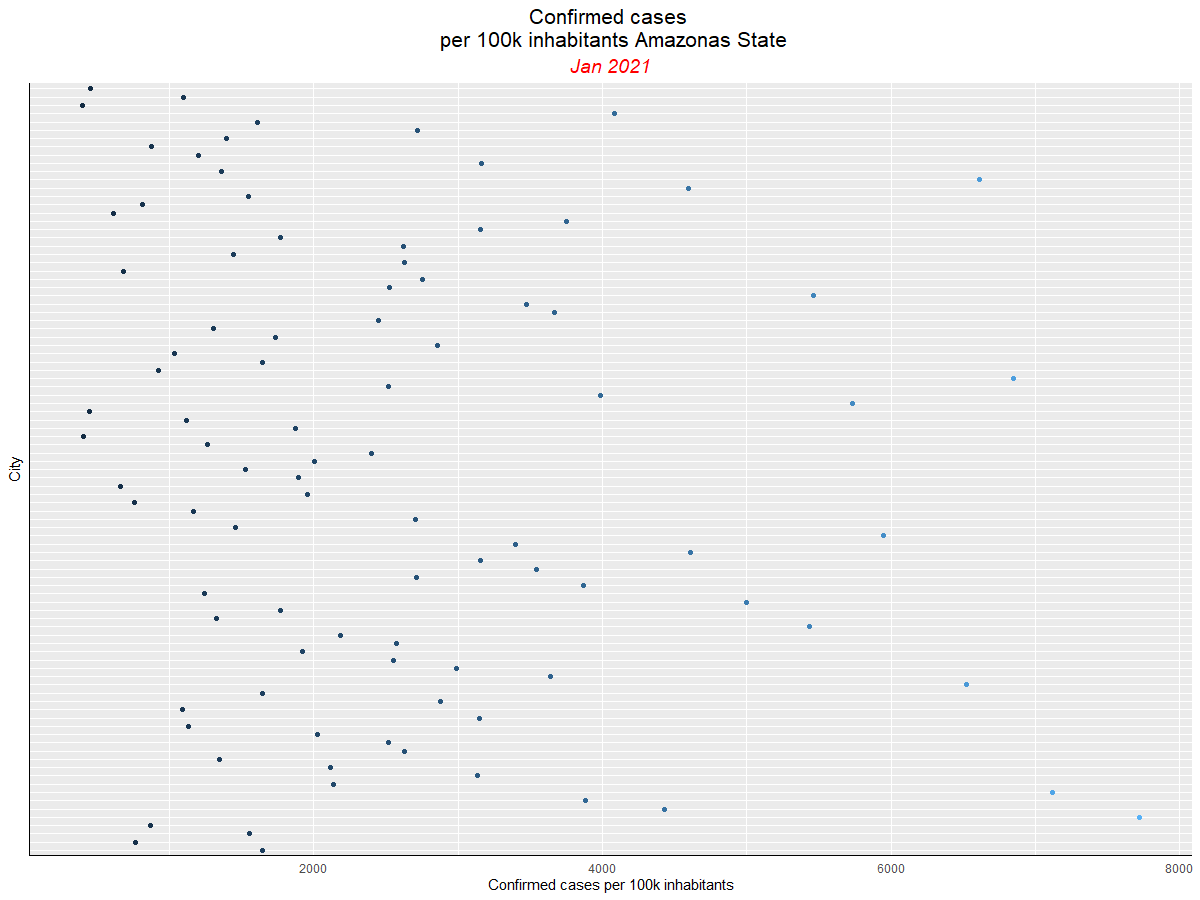
**Figure.26:** *Higher definition of Figure 8***.** *Estimated population curve, São Paulo state*



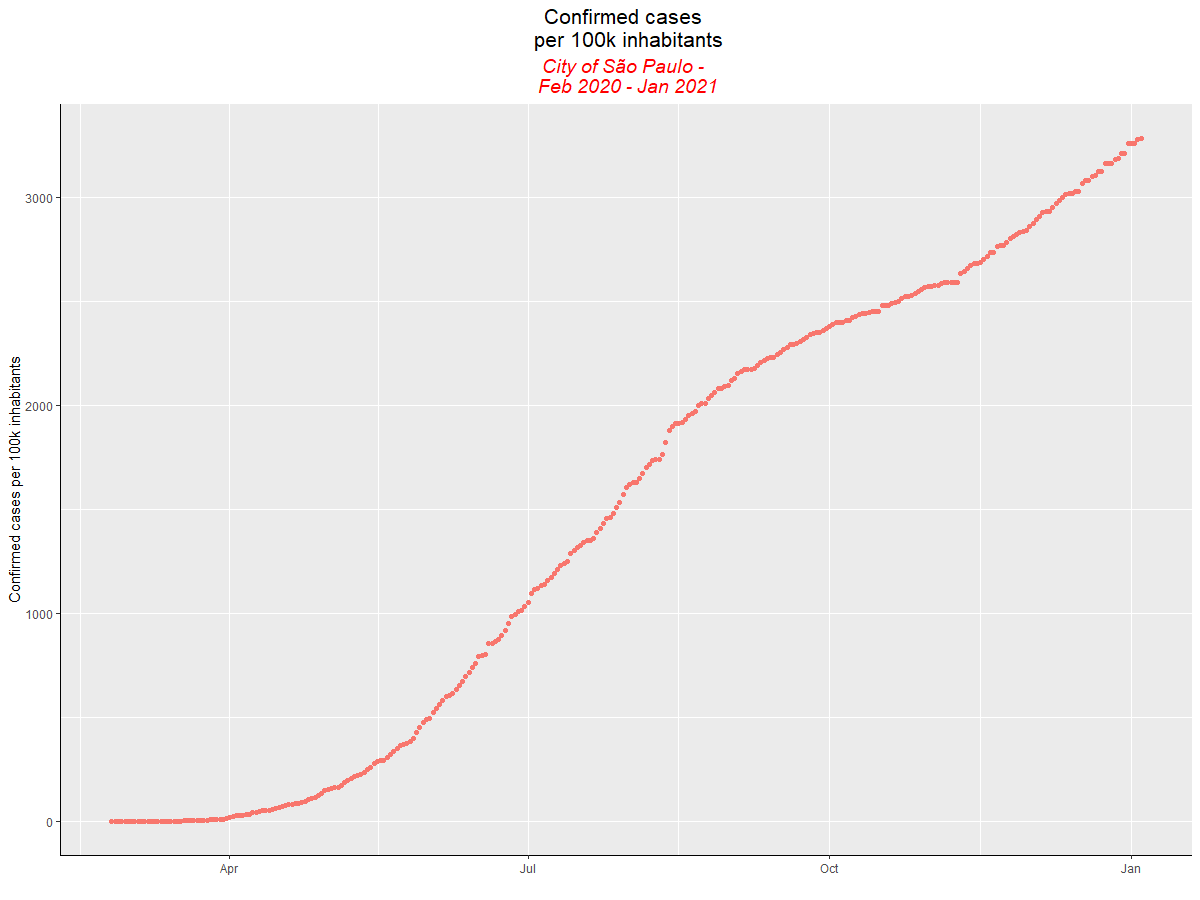
**Figure.27:** *Higher definition of Figure 9* *Estimated population curve, Amazonas state*



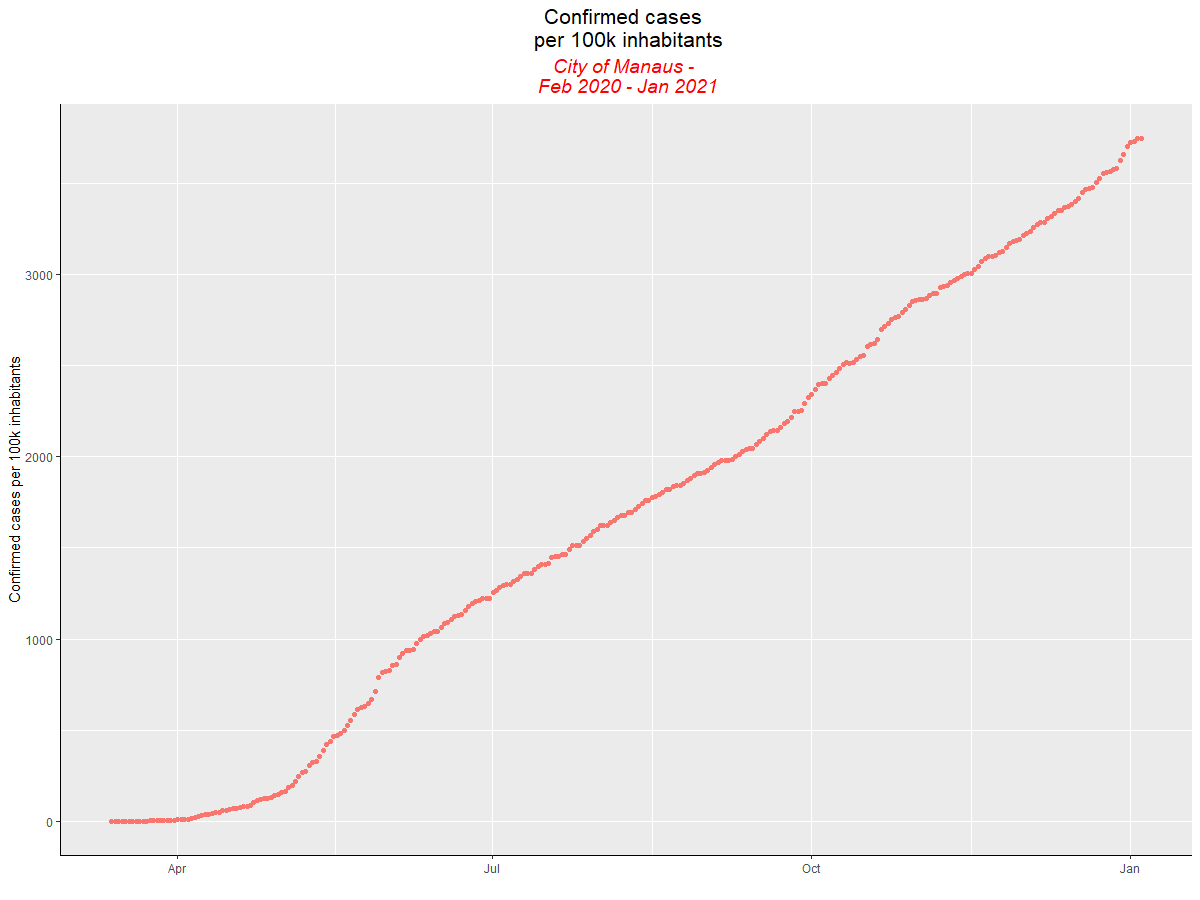
**Figure.28:** *Higher definition of Figure 10***.** *Confirmed cases per 100.000 inhabitants curve, São Paulo state*



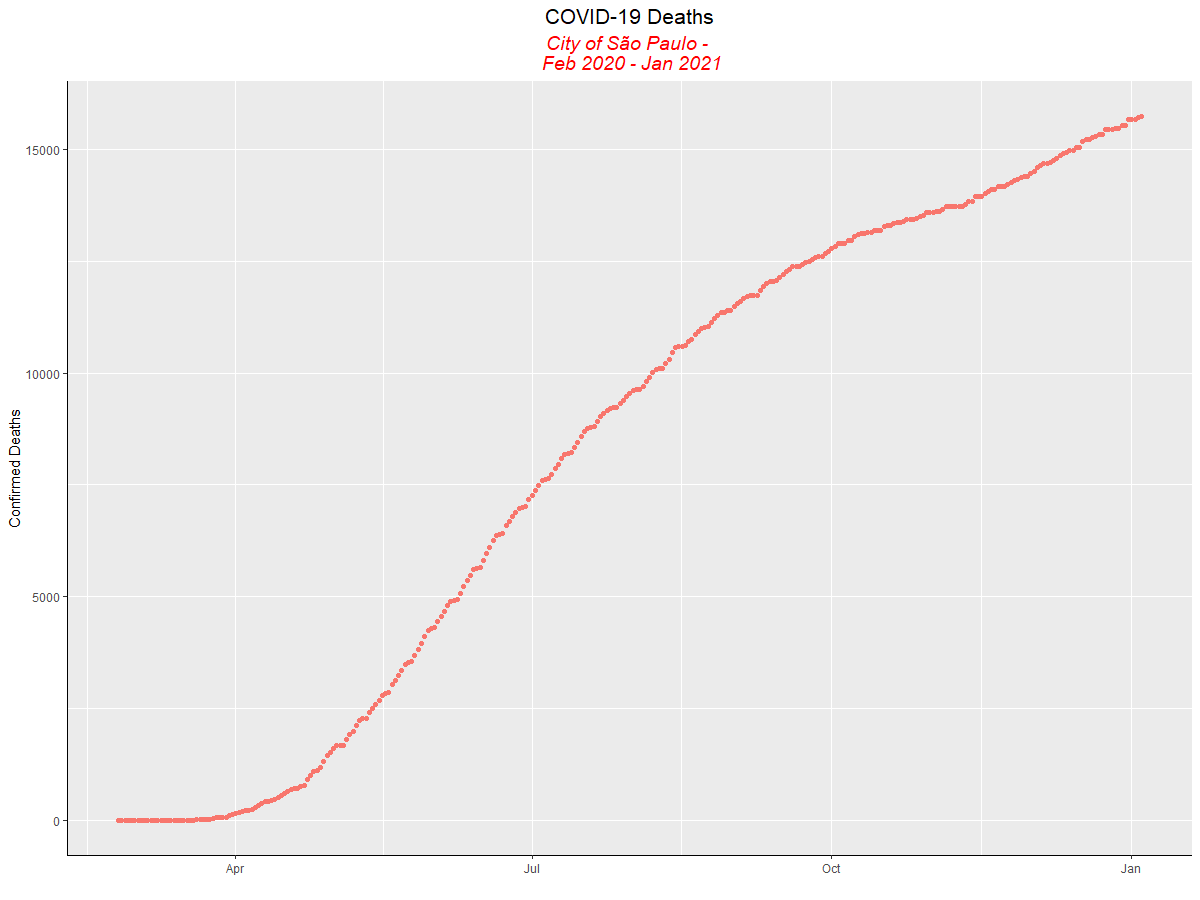
**Figure.29:** *Higher definition of Figure 11***.** *Confirmed cases per 100.000 inhabitants curve, Amazonas state*



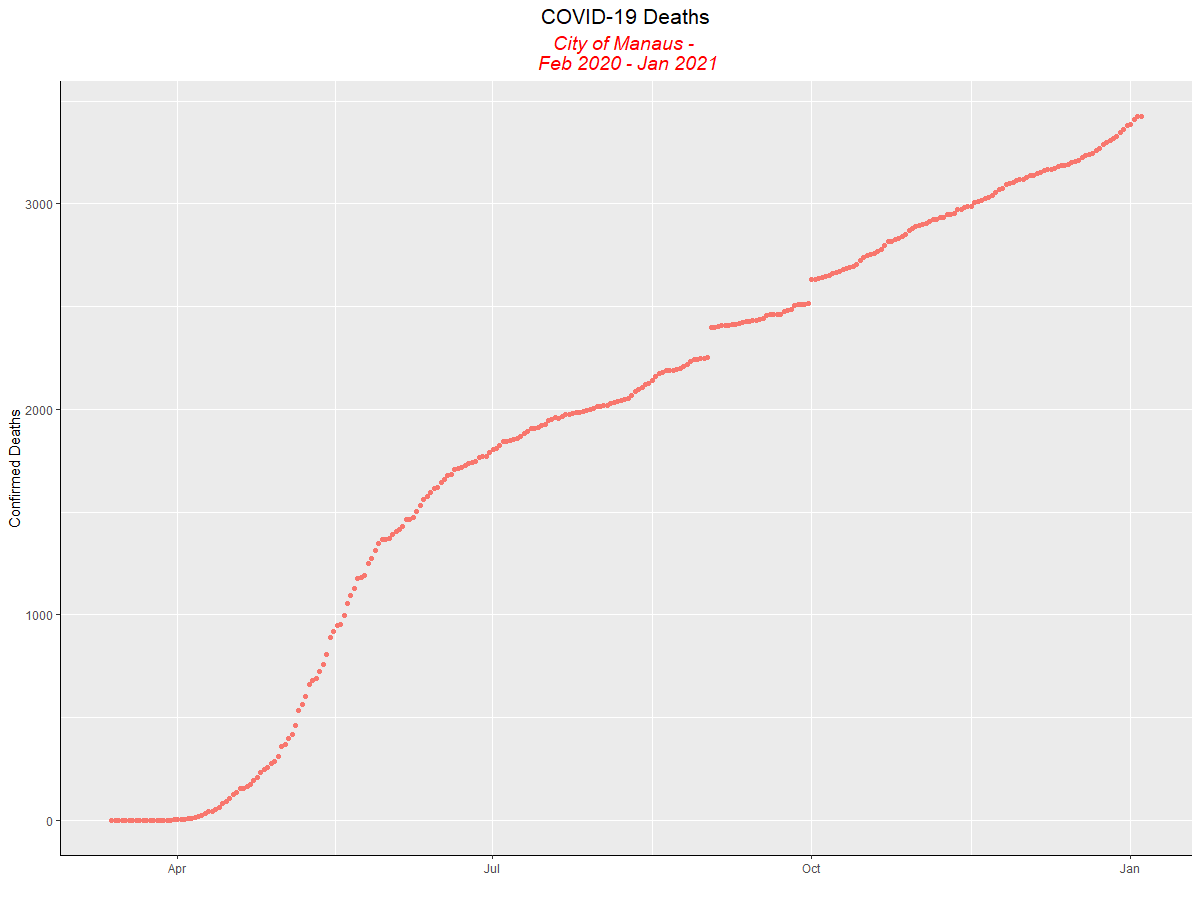
**Figure.30:** *Higher definition of Figure 12***.** *Confirmed cases per 100.000 inhabitants curve, São Paulo city*



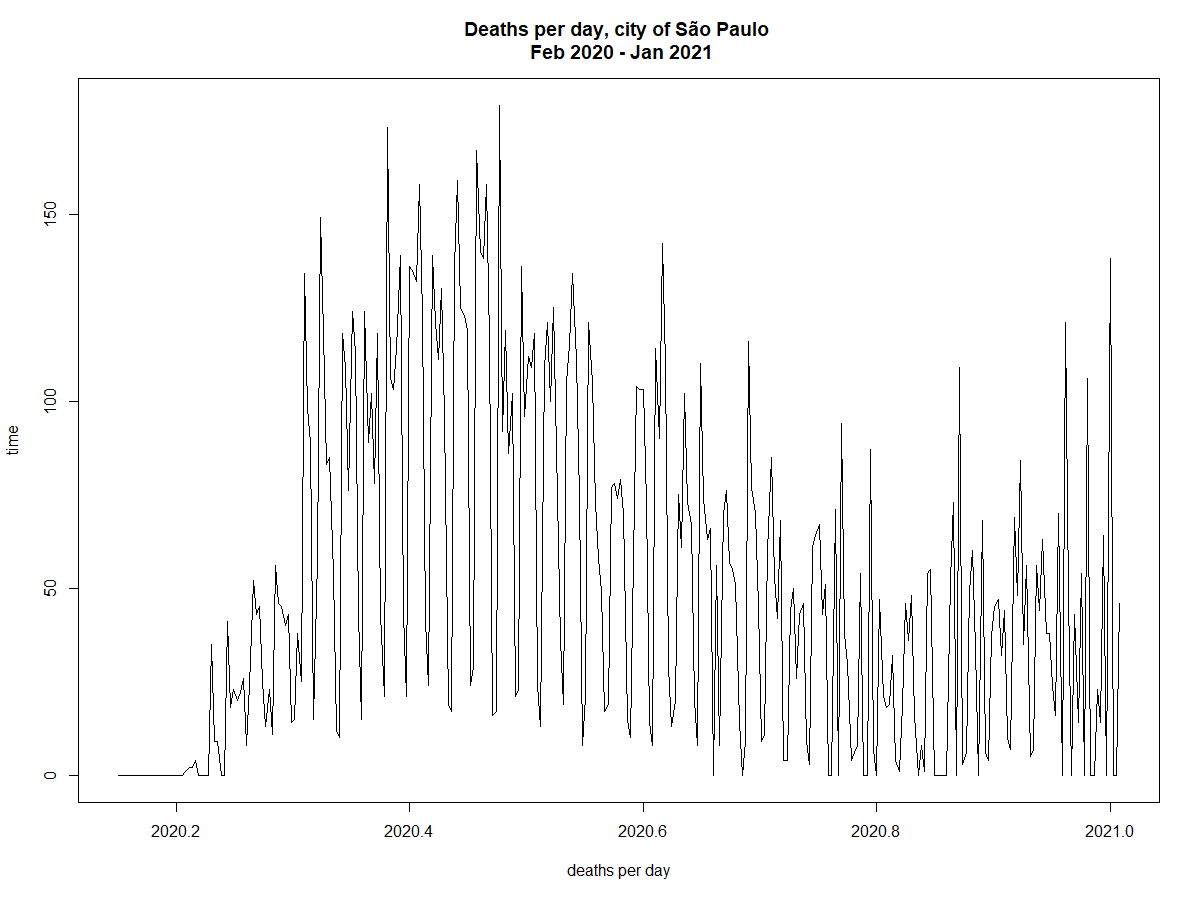
**Figure.31:** *Higher definition of Figure 13***.** *Confirmed cases per 100.000 inhabitants curve, Manaus city*



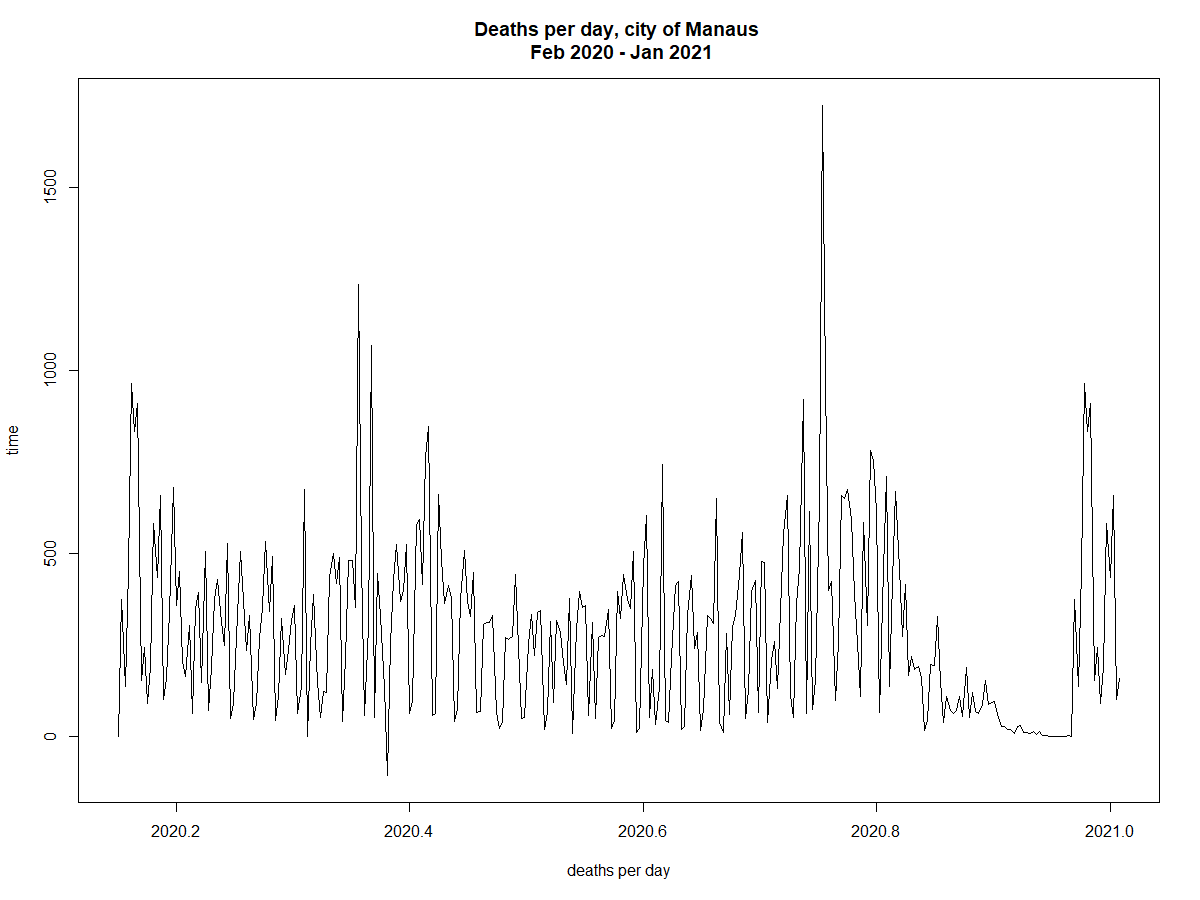
**Figure.32:** *Higher definition of Figure**14***.** *COVID-19 deaths curve, São Paulo city*

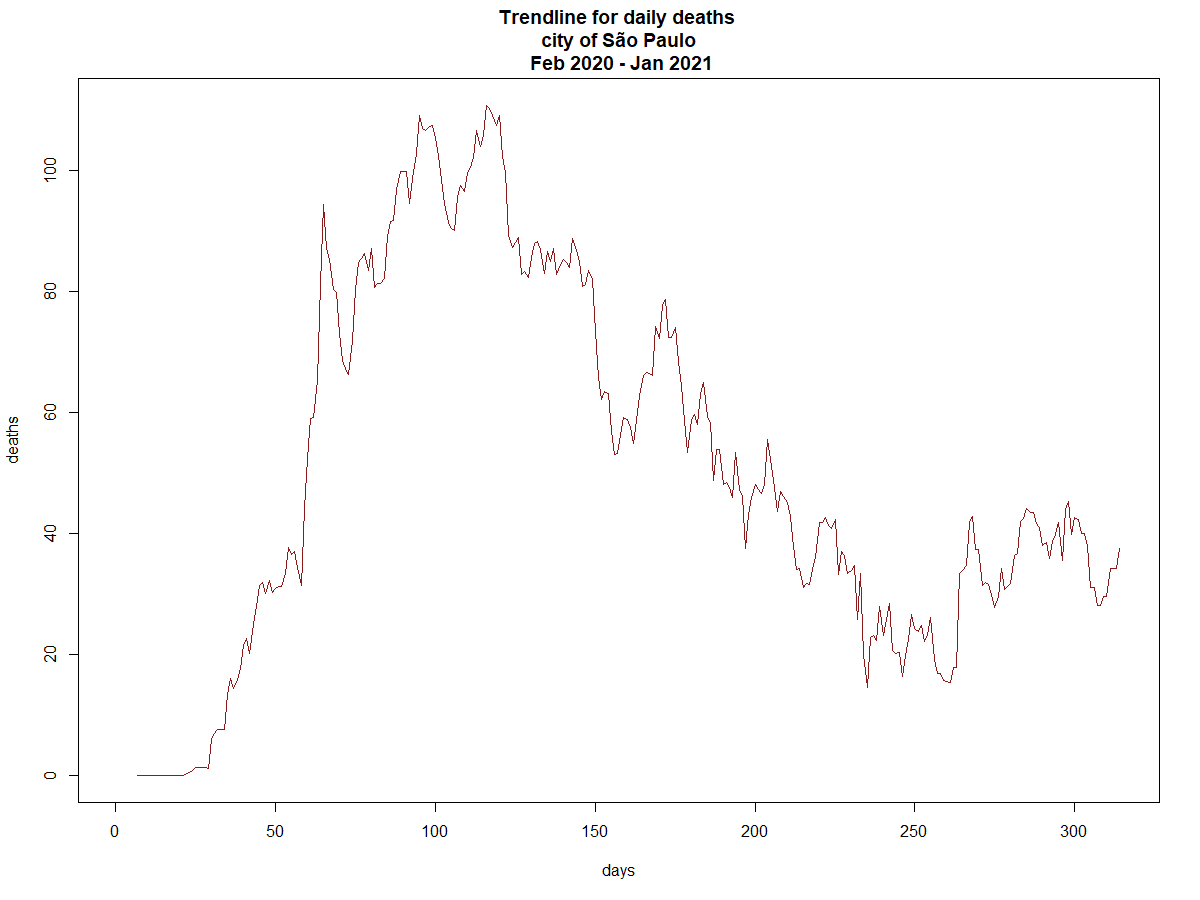


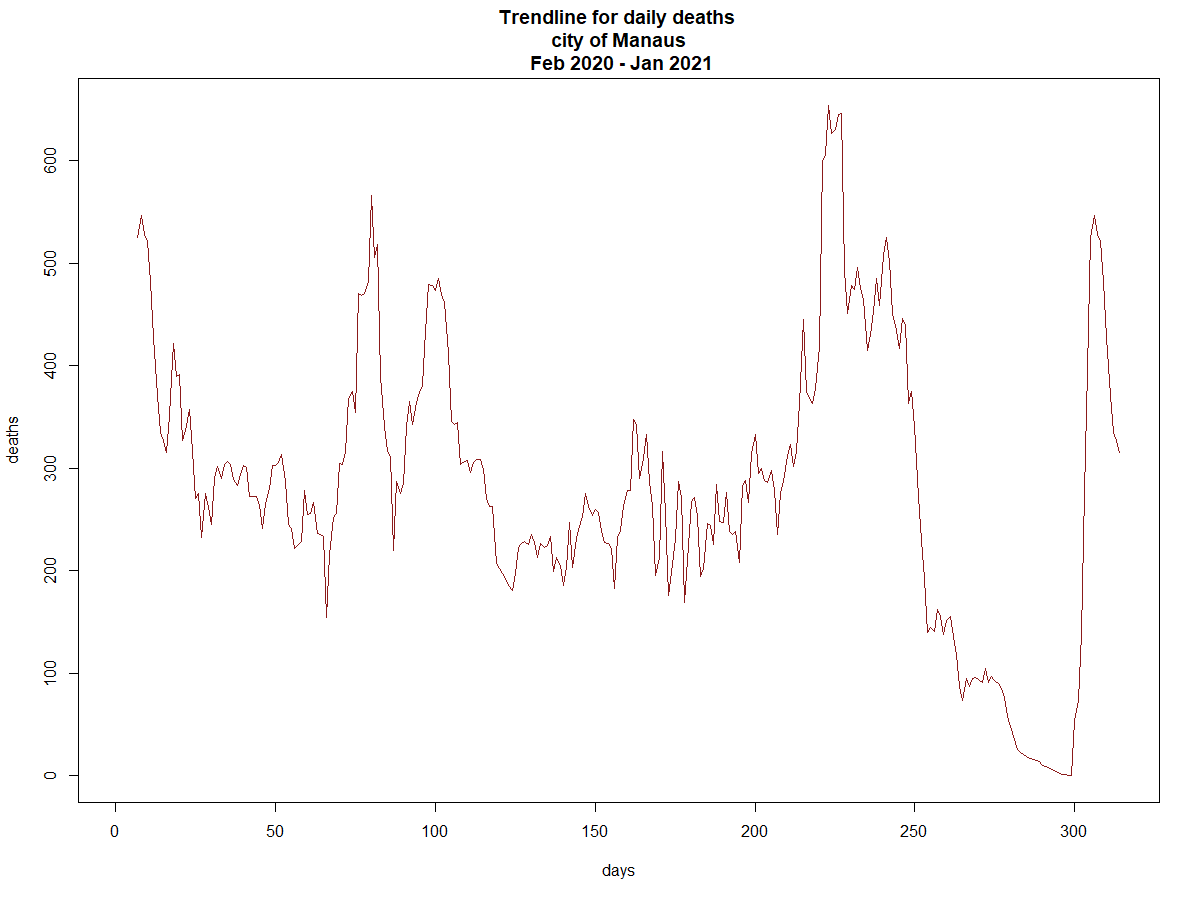
**Figure.33:** *Higher definition of Figure 15***.** *COVID-19 deaths curve, Manaus city*



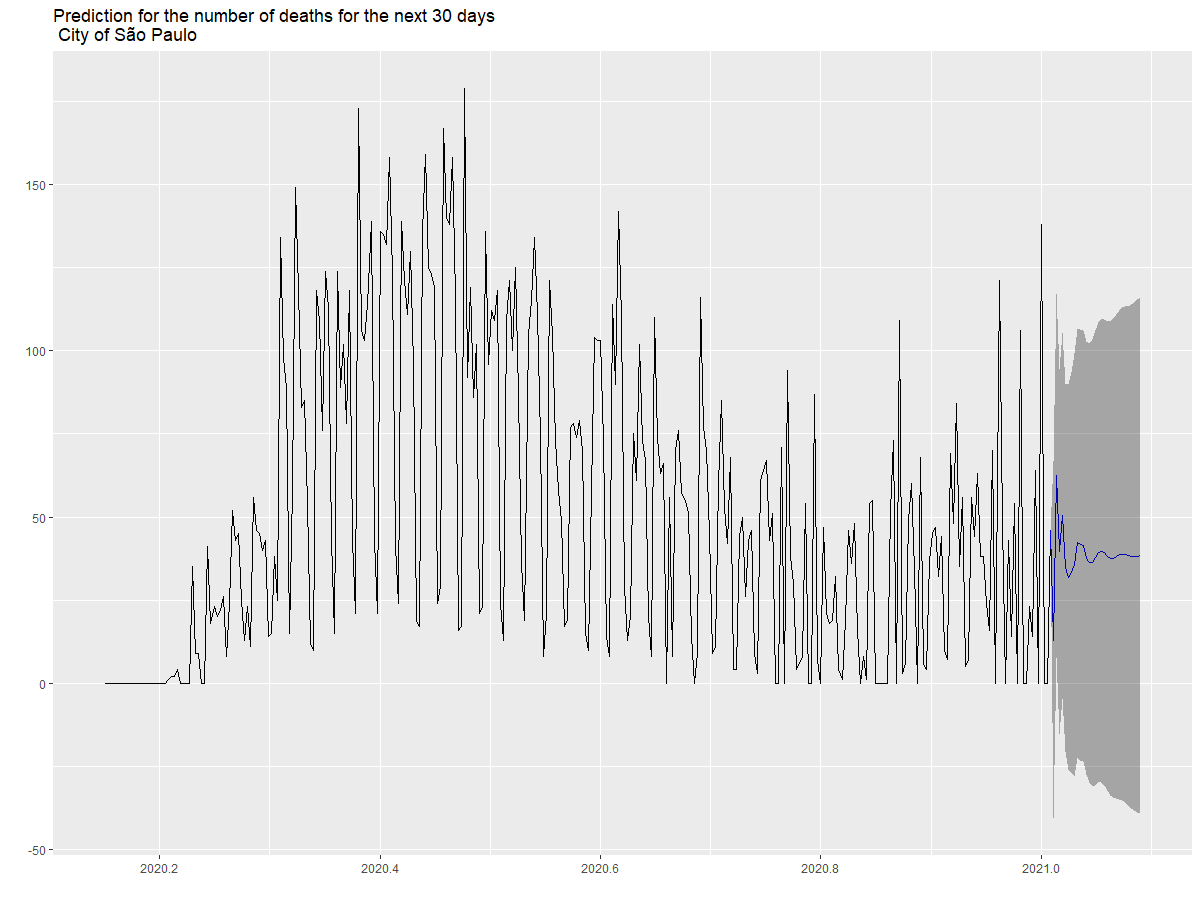
**Figure.34:** *Higher definition of Figure 16***.** *COVID-19 deaths per day, Feb 2020-Jan 2021, City of São Paulo*

 **Figure.35:** *Higher definition of Figure 17***.** *COVID-19 deaths per day, Feb 2020-Jan 2021, city of Manaus*

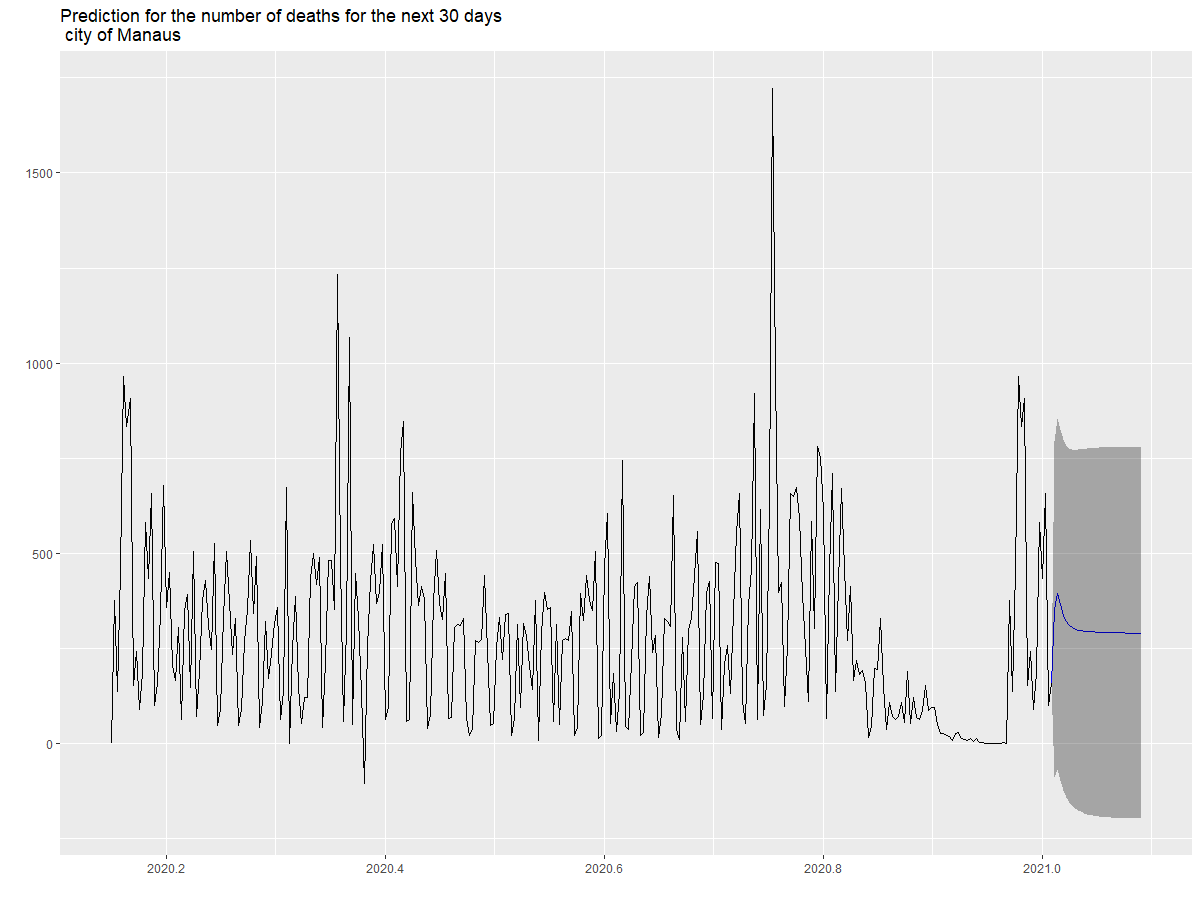
 **Figure.36:** *Higher definition of Figure 18***.** *Deaths trendline, Feb 2020-Jan 2021, City of São Paulo*



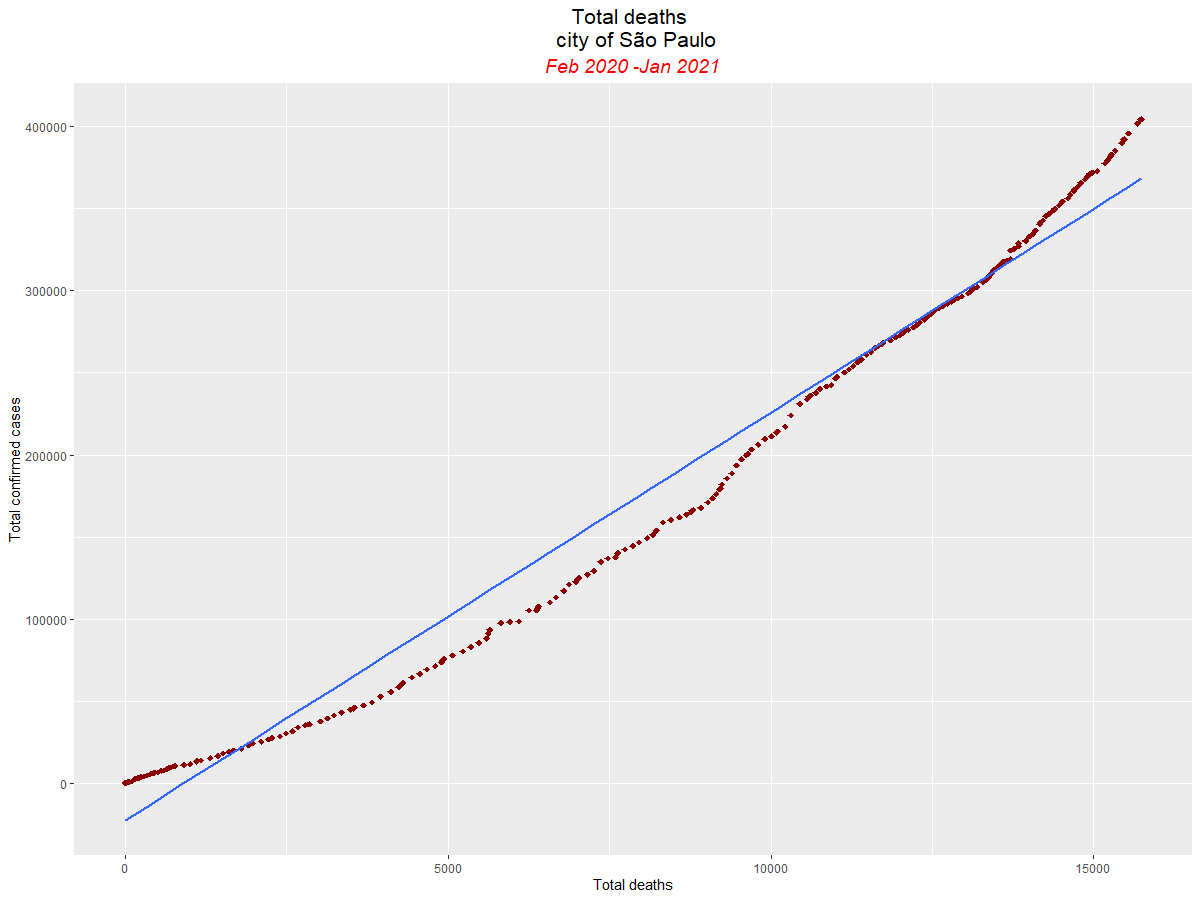
**Figure.37:** *Higher definition Figure 19***.** *Deaths trendline, Feb 2020-Jan 2021, City of Manaus*



**Figure.38:** *Higher definition Figure 20***.** *Prediction on the number of Deaths for the next 30 days, Feb 2020-Jan 2021, City of São Paulo*



**Figure.39:** *Higher definition Figure 21***.** *Prediction on the number of Deaths for the next 30 days, Feb 2020-Jan 2021, City of Manaus*



**Figure.40:** *Higher definition Figure 22***.** *Total deaths prediction tool, Feb 2020-Jan 2021, São Paulo city*