Analysis of the relationship between the diffusion of COVID-19 and the presence of metro, train and LRV in the neighborhoods of Rio de Janeiro

Gabriel Leuzinger Coutinho

May 2021

1 Introduction

1.1 Background

The COVID-19 has significantly impacted the whole world and Brazil is no exception. The country has already registered more than 14 million cases and 400 thousand deaths. Rio de Janeiro is one of the most impacted states of Brazil, especially the state's capital, the city of Rio de Janeiro. The city of 6.7 million habitants has already recorded 264 thousand cases of COVID-19 and 24 thousand deaths. Therefore, stopping the diffusion of COVID-19 in Rio is essential.

Many studies have showed a direct relationship between the use of public transportation, including metro, train, and light rail vehicles (LRV), and the diffusion of COVID-19 in urban areas. People are usually more exposed to the being infected during commutes in metro, train, and light rail vehicles (LRV), especially if the wagon is crowded. Although the use of these modes of transportation has significantly decreased in Rio since the beginning of the pandemic, they still carry more than 9 million passengers every month.

1.2 Problem

Metro, train and LRV might be vector for the diffusion of the COVID-19 through the city of Rio de Janeiro, as is the case in other cities, e.g., New York and Hong Kong. However, there are no studies on this topic in Rio. This project aims to test if the following hypothesis is true: metro, train, and LRV are contributing to the diffusion of COVID-19 in the City of Rio de Janeiro.

1.3 Interest

This information can inform public health officials in Rio about more effective ways of halting the diffusion of COVID-19 in the city. This is even more important now that new variants of the virus are starting to spread in Brazil. Besides, companies may use this result to decide which mode of transportation to recommend to their employees. Some companies in Brazil are already providing individual modes of transportation to their

employees, for them to avoid public transportation. If the proposed hypothesis is confirmed, this will give this business more elements to justify this kind of policy. Besides, it may encourage other companies to adopt similar policies.

2 Data acquisition and cleaning

2.1 Data sources

The first source of data used is the register of COVID-19 cases made available by Rio de Janeiro government on the website https://coronavirus.rio. Data on the modes of transportation analyzed were based on the information available at the website's of the metro company (https://www.metrorio.com.br), the train company (https://www.supervia.com.br), and the LRV company (https://www.vltrio.com.br/#/). The lines and stations for each of these modes of transportation were identified in the websites and the geographic coordinates of each of the stations were collected using Google maps.

Besides, data about Rio de Janeiro 163 neighborhoods were retrieved from the Wikipedia. Data on the population of each of the neighborhoods was collected on the Rio de Janeiro government website: https://www.data.rio/search?collection=Dataset&tags=população. Finally, it was also necessary to access the Rio de Janeiro government website to retrieve a geojson file of Rio neighborhoods, which is used to create maps.

2.2 Data cleaning and use

First, I use the dataset of COVID-19 cases to compile a list of cases per month per neighborhood for the period that is analyzed: from January 2020 to April 2021. Then, I combine this data with the data on the population of each neighborhood to calculate the cases of COVID-19 per population for each of these neighborhoods for each month in the period analyzed. However, the population data is from the last Brazilian census, in 2010. The problem is that three neighborhood, Jabour, Lapa, and Vila Kennedy were created after this year. Therefore, I must adjust the dataset of COVID-19 cases to combine these three neighborhoods with the ones they used to be part of: Senador Camará, Centro, and Bangu, respectively.

This same problem is detected in the dataset of Rio de Janeiro neighborhoods, retrieved from Wikipedia. Therefore, I do the same adjustment in this dataset. Combining these three datasets, I create a data frame of COVID-19 cases per month per neighborhood per

population of the neighborhood. Besides, I also create a second data frame to store the data of cumulative COVID-19 cases per neighborhood per population of the neighborhood.

The next step is to prepare the geojson file to be used to create the maps. It is necessary to adjust the neighborhood names in the file and then attribute a number to each neighborhood. The necessary information to create the maps are retrieved from the geojson file into a data frame and combined with the numbers attributed to the neighborhoods. This data frame is combined with the data frame with the data of cumulative COVID-19 cases per neighborhood per population of the neighborhood to create a choropleth map to display the data with a time slider, as well as other analysis.

Then, I divide the dataset on the modes of transport in three sets, one for each mode. These sets are used to create a map to displaying these stations' locations. Finally, this map is combined with the choropleth map to create a third map. This final map displays the evolution of COVID-19 cases over time in each neighborhood and the modes of transportation present there.

Finally, the data frame of COVID-19 cases per population for each of these neighborhoods by month is combined with the information of the modes of transportation present in each neighborhood. This final data frame is used to perform a series of statical analysis to determine if the project hypothesis is valid or not.

3 Methodology

There are two steps in analyzing the data. First, I explore the data by creating maps and visually inspecting the relationship between the COVID-19 dispersion in Rio and presence of metro, train, and LRV stations in the neighborhoods. Second, I make a series of statical analysis to determine if the project hypothesis is valid or not.

3.1 Exploring the data

As already discussed, I start exploring the data by creating maps to visually inspect it. The first map I create is a map displaying all the metro, train and LRV stations in Rio de Janeiro. This allowed me to check if all these stations are in the right place, which is essential to later compare the dispersion of COVID-19 though neighborhoods with the presence of these modes of transportation in these neighborhoods.

The next step is also the creation of a map. This map is a choropleth to display the cumulative number of cases of COVID-19 per total population in each neighborhood. It is important to normalize the data by dividing the number of COVID-19 cases by the total population of the neighborhood, because this population considerably vary. For example, the less populated neighborhood, Grumari, has only 167 habitants, while the most populated neighborhood, Campo Grande, has almost 330 thousand habitants.

To allow a properly visual inspection of the evolution of the dispersion of COVID-19 through the neighborhoods, I add a time slider to the map, which allow to see the cumulative number of COVID-19 cases per total population for each month of the period being analyzed. Therefore, we can slide from January 2020 up to April 2021, viewing how COVID-19 is evolving through Rio. This map also allows me to check if the data displayed is compatible with the dataset.

Then, I can finally combine the information of the two maps in one single map. Using this final map, we can detect any relationship between the presence of metro, train, and LRV in the neighborhoods and the dispersion of COVID-19.

The final step in data exploration is creating some line plots to display the data. The first one displays the number of COVID-19 cases per total neighborhood habitants per transport mode (metro, train, and LRV) present in the neighborhood over time. The second line plot is very similar to the first, but this time is displaying the cumulative number of COVID-19 cases instead of the monthly total.

3.2 Statistical analysis

The first step in the statistical analysis is deciding which type of statistical method should be used. Since we have as independent variable categorical values, i.e., the type of mode of public transport (metro, train, and LRV) present in the neighborhood and continuous values, i.e., the number of COVID-19 cases, as the dependent variable, a good method to analyze the data is the one-way analysis of variance (ANOVA).

ANOVA is used when you have a categorical independent variable (with two or more categories) and a normally distributed interval dependent variable and you wish to test for differences in the means of the dependent variable broken down by the levels of the independent variable. If the test is positive, it means that at least two of the categories analyzed are significantly different. Therefore, we can reject the null hypothesis: that the presence or not of different modes of public transportation (metro, train, LRV) in a

neighborhood in Rio de Janeiro have no impact in the number of COVID-19 cases in this neighborhood.

Before doing the one-way ANOVA test, I had to make sure that a few assumptions are met from the data:

- The samples are independent of one another.
- Data is normally distributed. The outcome variable must follow a normal distribution in each subpopulation. Normality is only needed for small sample sizes, say $n < 20$.
- The population standard deviations of the groups are homoscedastic. The variances within all subpopulations must be equal. Homogeneity is only needed if sample sizes are very unequal.

I start this checking by plotting some box plots and histograms to have a look at the distribution of the data in the categories defined:

(0) No metro, train, or LRV present in the neighborhood.
(1) Only LRV present in the neighborhood.
(2) Only train present in the neighborhood.
(3) Train and LRV present in the neighborhood.
(4) Only metro present in the neighborhood.
(5) Metro and LRV present in the neighborhood.
(6) Train and LRV present in the neighborhood.
(7) Metro, train, and LRV present in the neighborhood.

Categories (3) and (5) are discarded because there is no observation in these subgroups. Besides, there is only one observation in the subgroup (7), what invalidate using this sample in many of the tests. Therefore, the observation in group 7 was included in group (6). This division of the dataset guarantees that all the samples are independent from one another because no observation is part of more than one sample.

Then, I create quantile-quantile (QQ) plots for some visual representation of the data. The QQ plot is a much better visualization of the data, providing more certainty about the normality. QQ plot is a scatterplot created by plotting two sets of quantiles against one another. f the two distributions which we are comparing are exactly equal then the points on the Q-Q plot will perfectly lie on a straight-line $y = x$. Therefore, if they our data points are drawn from a normal distribution, the points should lie on the line in the QQ-plot.

Finally, I perform one last test to check the data follow a normal distribution: the Shapiro-Wilk test, which examines if a variable is normally distributed in some population. It tests the hypothesis that the distribution is normal. Therefore, if the p-value is greater than 0.05, we cannot reject the null hypothesis that a variable is normally distributed in some population.

Next, I check for homoscedasticity given that the sample sizes are very unequal. I use two tests: Bartlett and Leven. The null hypothesis for these tests is that the groups we're comparing all have equal population variances. If we reject the null hypothesis, we can also reject the assumption of homoscedasticity.

After all assumptions are checked, I finally perform the one-way ANOVA test. First, I use the one-way ANOVA test to analyze the data on the total number of COVID-19 cases per neighborhood. Then, I repeat the test for each month in the period between January 2020 and April 2021.

However, the one-way ANOVA test left an unanswered question: which precisely means are different? The one-way ANOVA test tells us if our results are significant or not but does not tell us where the results are significant. The most common solution to this problem is using Tukey's Honestly Significant Difference (HSD) procedure. It allows us to interpret the statistical significance of our ANOVA test and find out which specific groups' means (compared with each other) are different. Therefore, the final step of the statical analysis is performing the Tukey's HSD procedure for the datasets for which the one-way ANOVA test rejected the null hypothesis.

4 Results and discussion

The first map I created, displaying all the metro, train and LRV stations in Rio de Janeiro, is shown in Figures 1 and 2. As can be seen, all the stations were correctly displayed.

Figure 1 – Map of the metro, train and LRV stations of the city of Rio de Janeiro

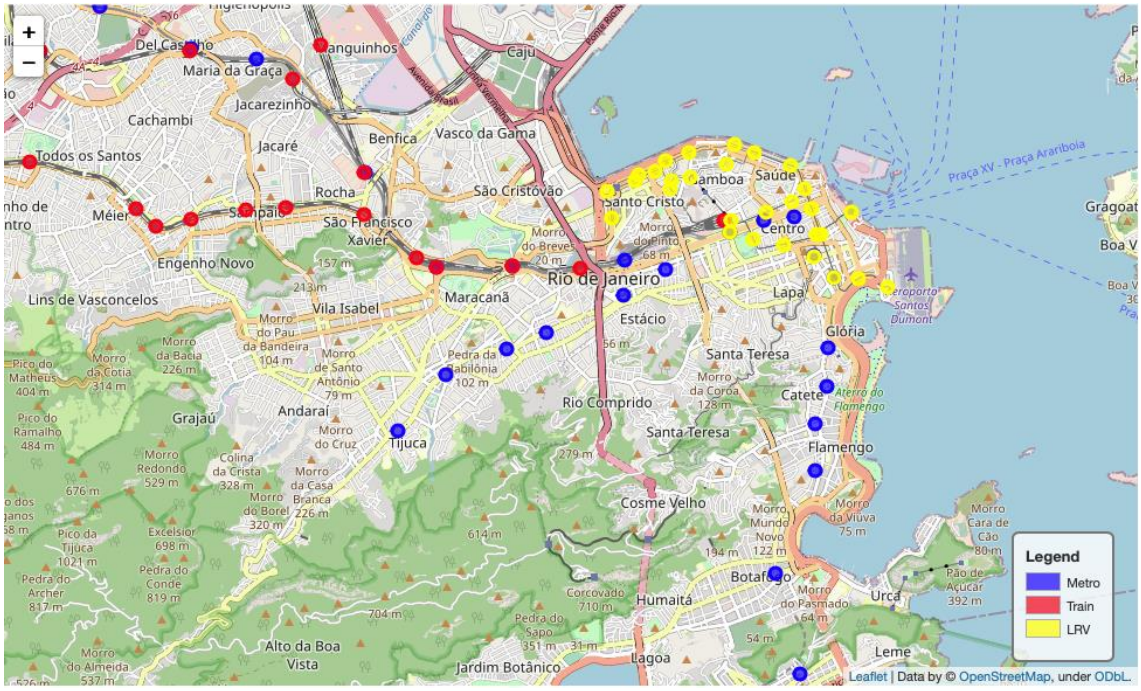

Figure 2 – Map of the metro, train and LRV stations of the central region of city of Rio de Janeiro

The second map created is the choropleth to display the cumulative number of cases of COVID-19 per total population in each neighborhood, which can be seen in Figures 3 and 4.

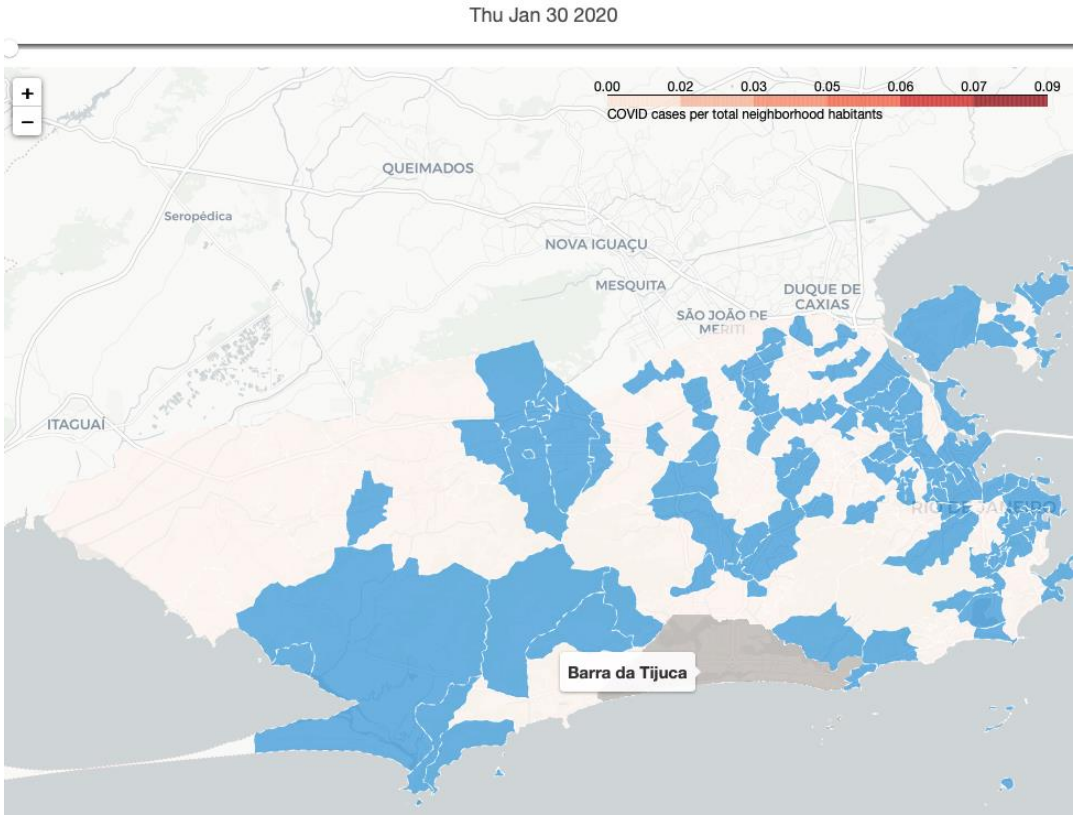

Figure 3 – Choropleth of the cumulative number of cases of COVID-19 per total population in each neighborhood for January 2020

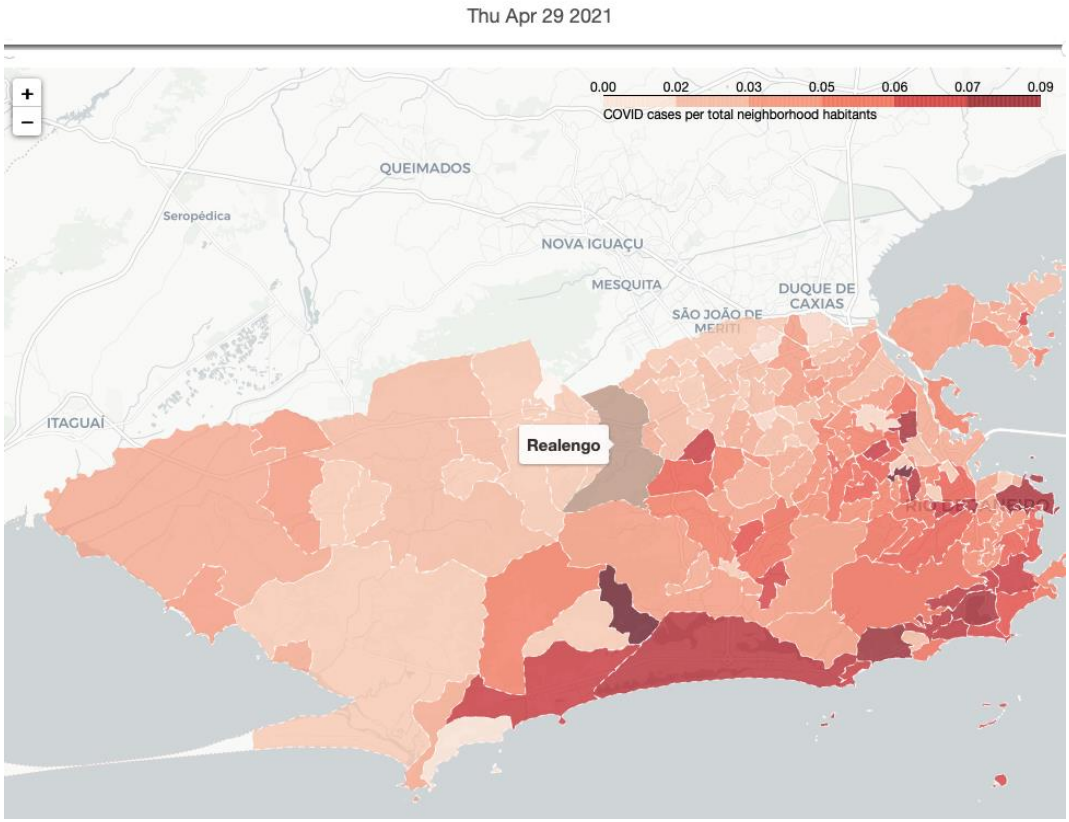

Figure 4 – Choropleth of the cumulative number of cases of COVID-19 per total population in each neighborhood for April 2021

Then, I combine the two previous maps in a single map. The analysis of Figures 5 to 10 clearly indicate that the cases of COVID-19 are increasing before in the central region of Rio de Janeiro and later spread to the other regions. Besides, some points seem to indicate that this spread accompanied the metro and train stations, especially in the first months of the pandemic. This pattern is not clear though and is not distinguishable in the later months.

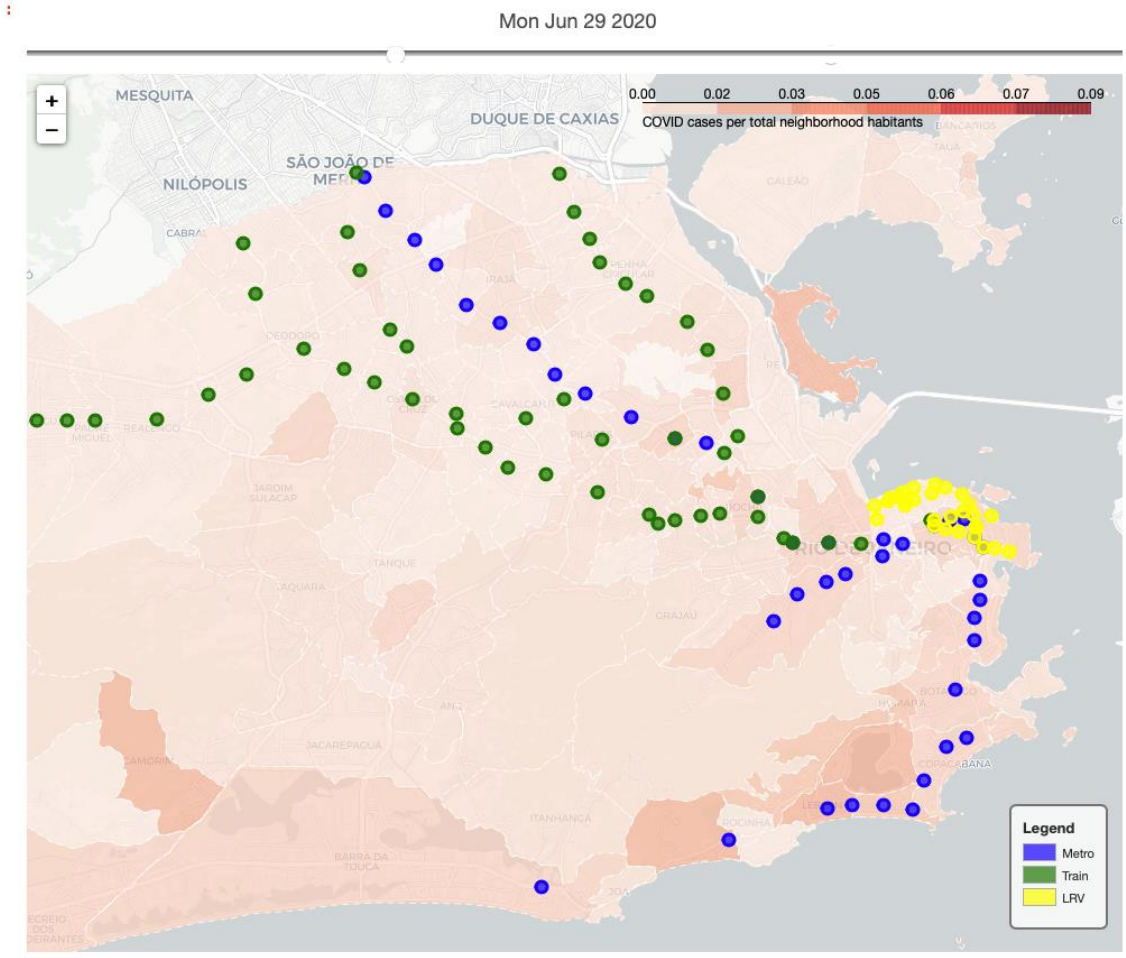

Figure 5 – Choropleth of the cumulative number of cases of COVID-19 per total population in each neighborhood for June 2020

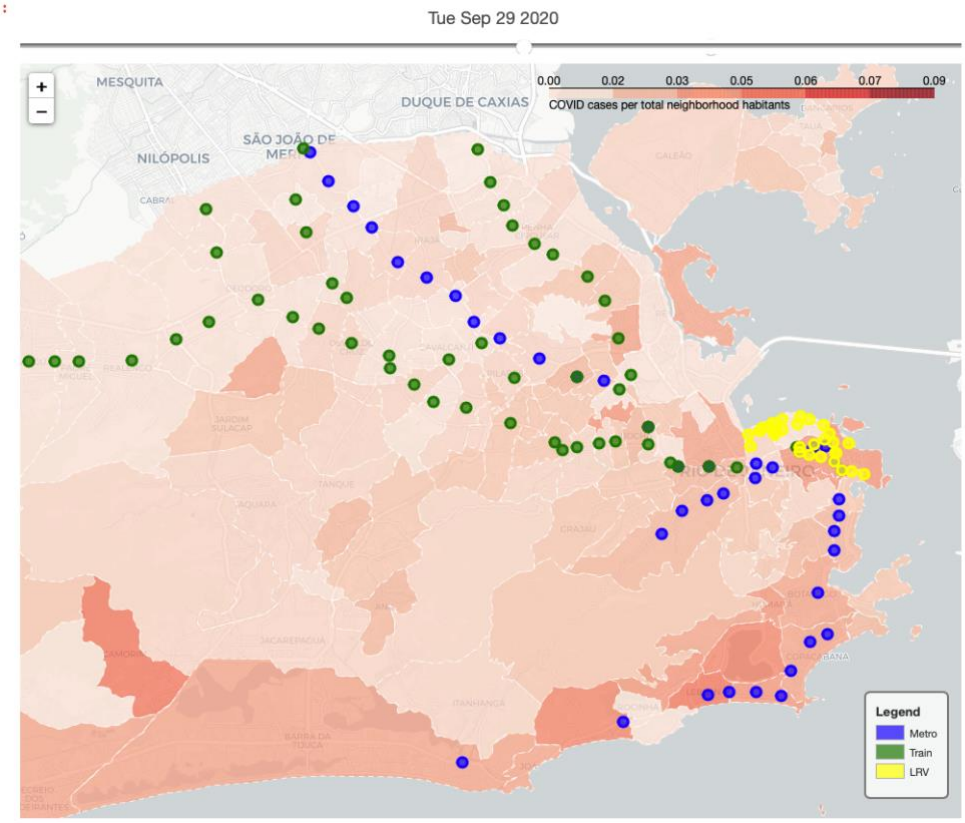

Figure 6 – Choropleth of the cumulative number of cases of COVID-19 per total population in each neighborhood for September 2020

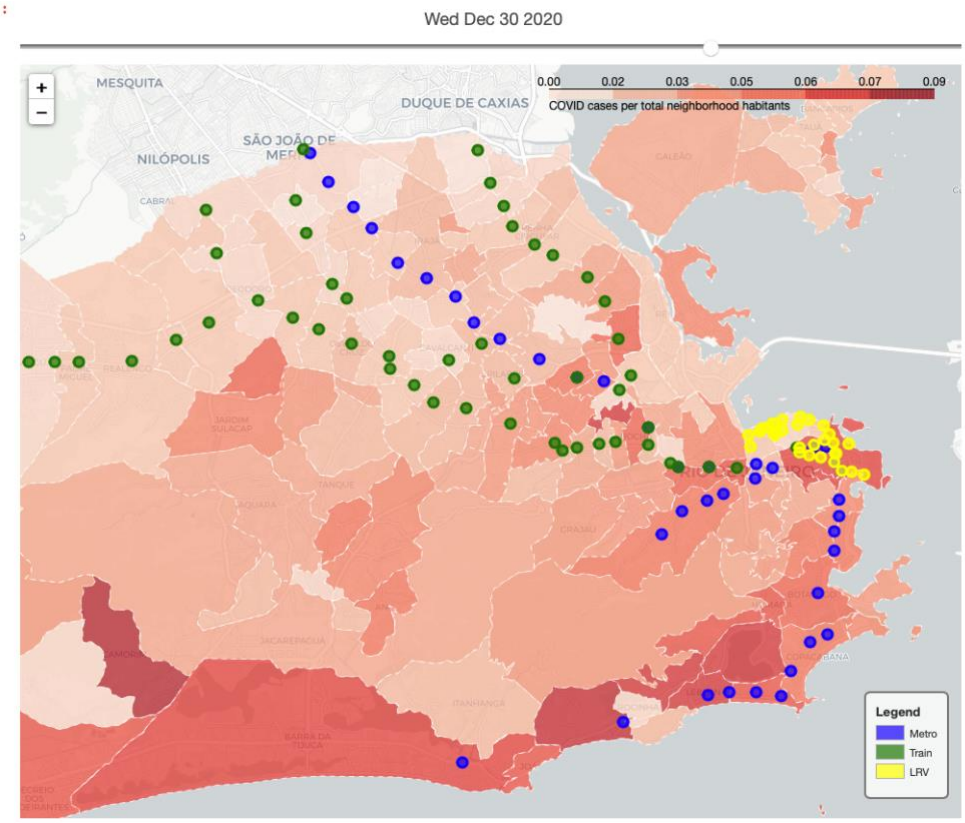

Figure 7 – Choropleth of the cumulative number of cases of COVID-19 per total population in each neighborhood for December 2020

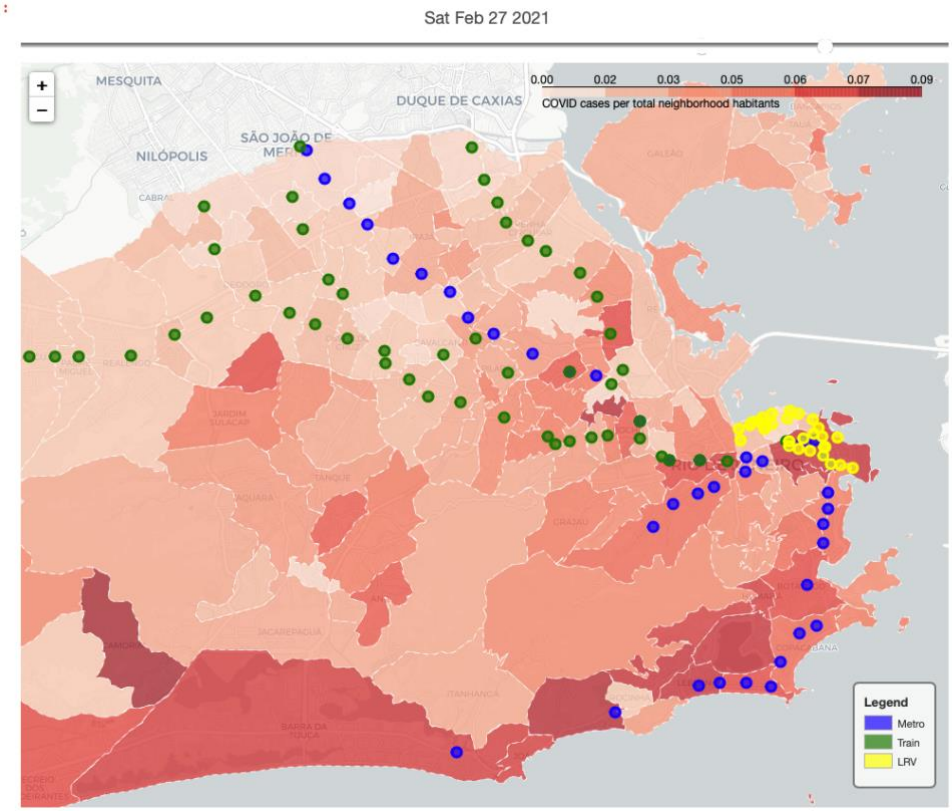

Figure 8 – Choropleth of the cumulative number of cases of COVID-19 per total population in each neighborhood for February 2021

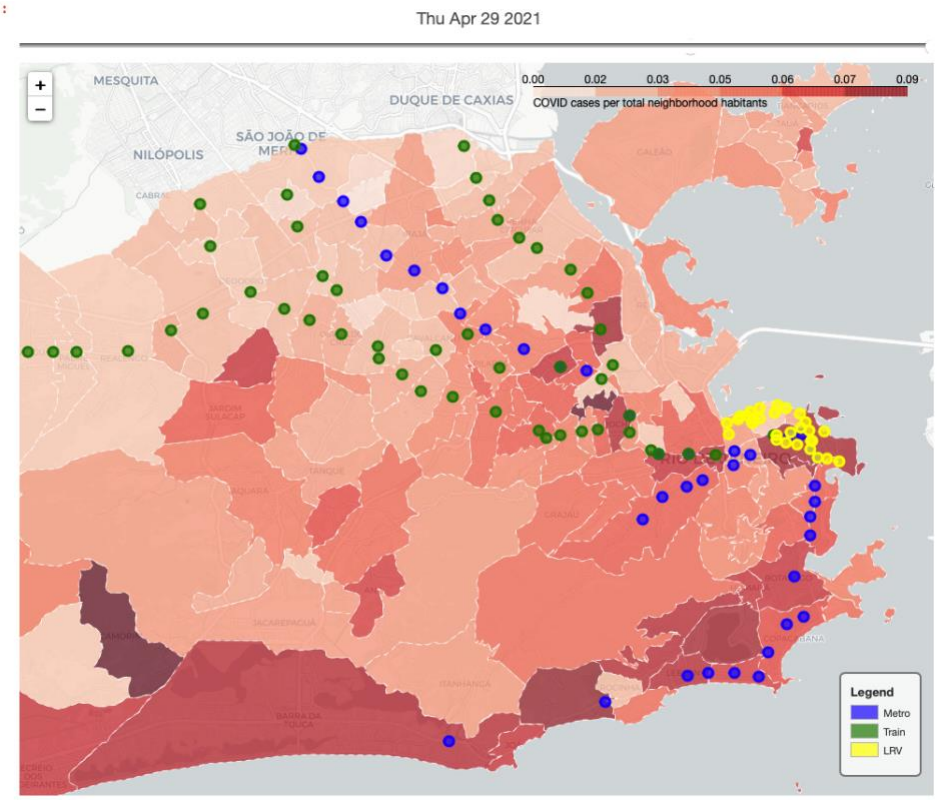

Figure 9 – Choropleth of the cumulative number of cases of COVID-19 per total population in each neighborhood for April 2021

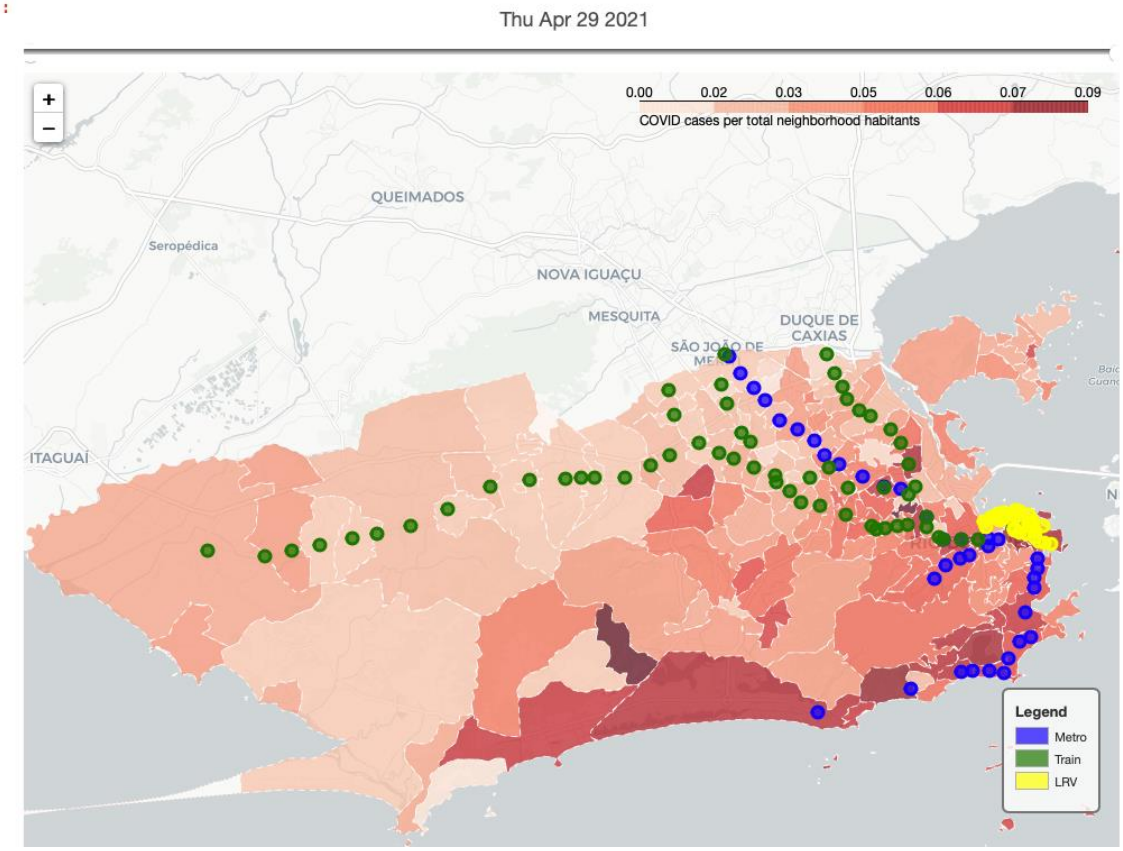

Figure 10 – Choropleth of the cumulative number of cases of COVID-19 per total population in each neighborhood for April 2021

The first line plot I created displays the number of COVID-19 cases per total neighborhood habitants per transport mode (metro, train, and LRV) present in the neighborhood over time (Figure 11). The graphic indicates that the COVID-19 cases follow similar patterns independent of the mode of transportation present in each neighborhood. Nonetheless, it also indicates that neighborhoods with metro stations and those with metro, train and LRV stations have more cases of COVID-19/total population of the neighborhood than the others.

The second line plot (Figure 12) shows cumulative number of COVID-19 cases per total neighborhood habitants per transport mode (metro, train, and LRV) present in the neighborhood over time. As expected, the graphic indicates that that neighborhoods with metro stations and those with metro, train and LRV stations have more cases of COVID-19 than the others.

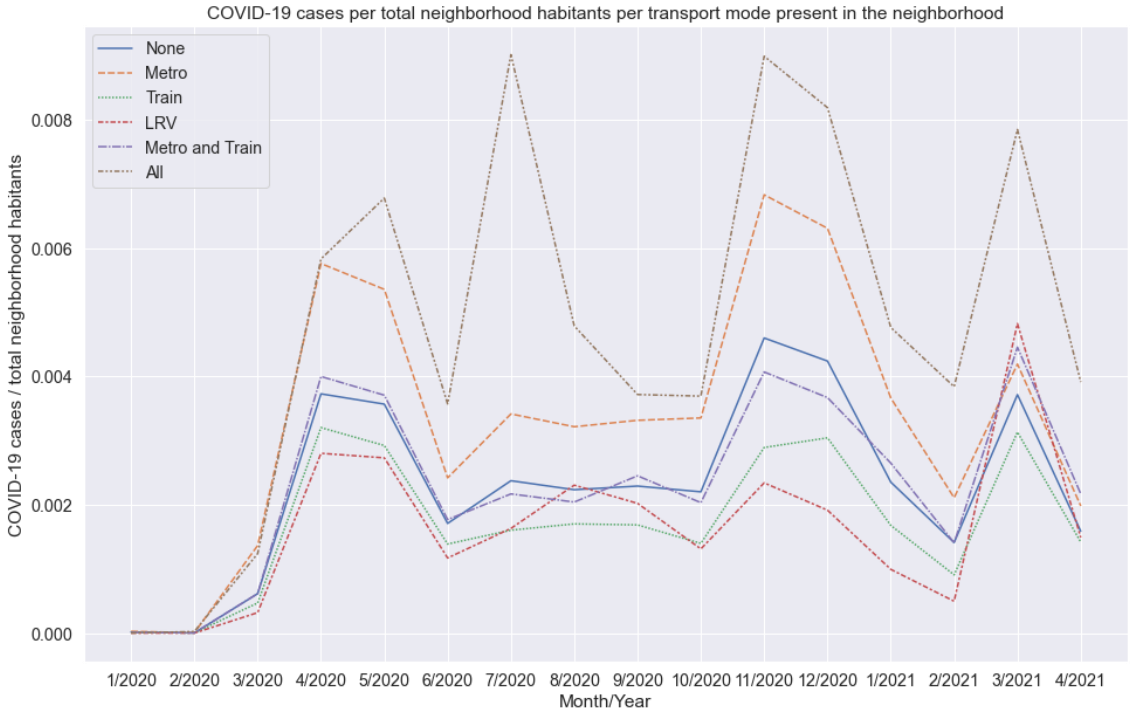

Figure 11 – Number of COVID-19 cases per total neighborhood habitants per transport mode (metro, train, and LRV) present in the neighborhood

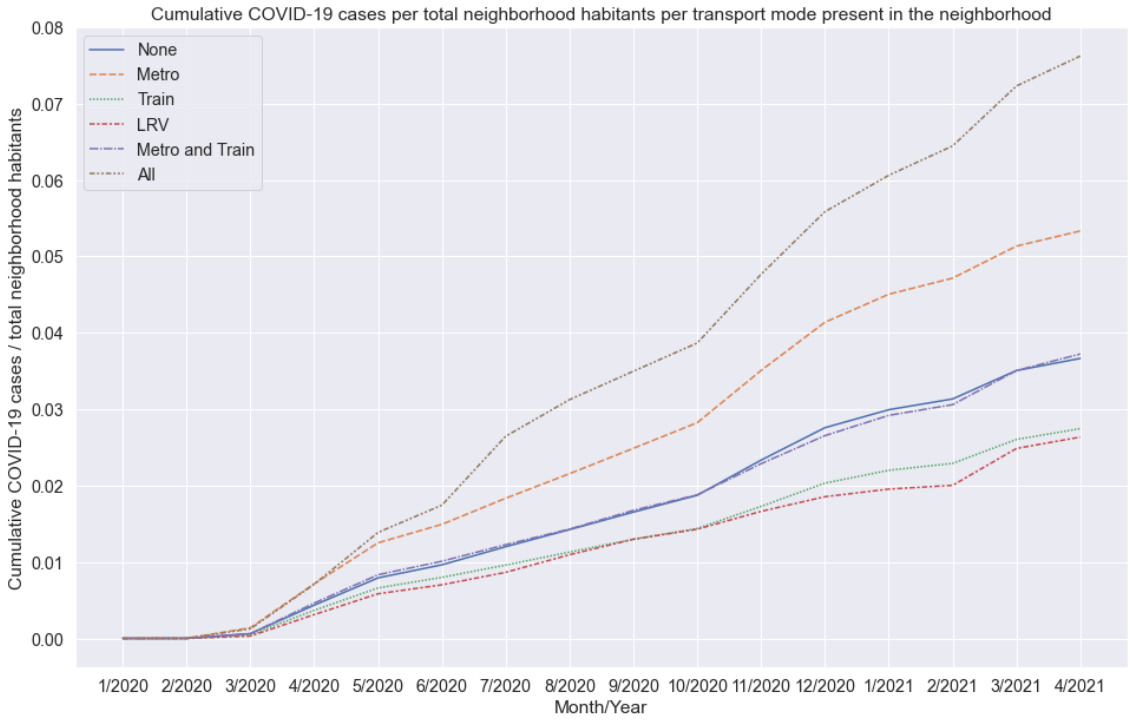

Figure 12 – Cumulative number of COVID-19 cases per total neighborhood habitants per transport mode (metro, train, and LRV) present in the neighborhood

Now we look at the results of the statistical analysis. First, I create boxplots to see the COVID-19 cases per neighborhood habitant's distribution by mode of transport present

in each neighborhood (Figure 13). Then, I also create histograms (Figure 14) for each of the classes that are being analyzed, as described in the Methodology.

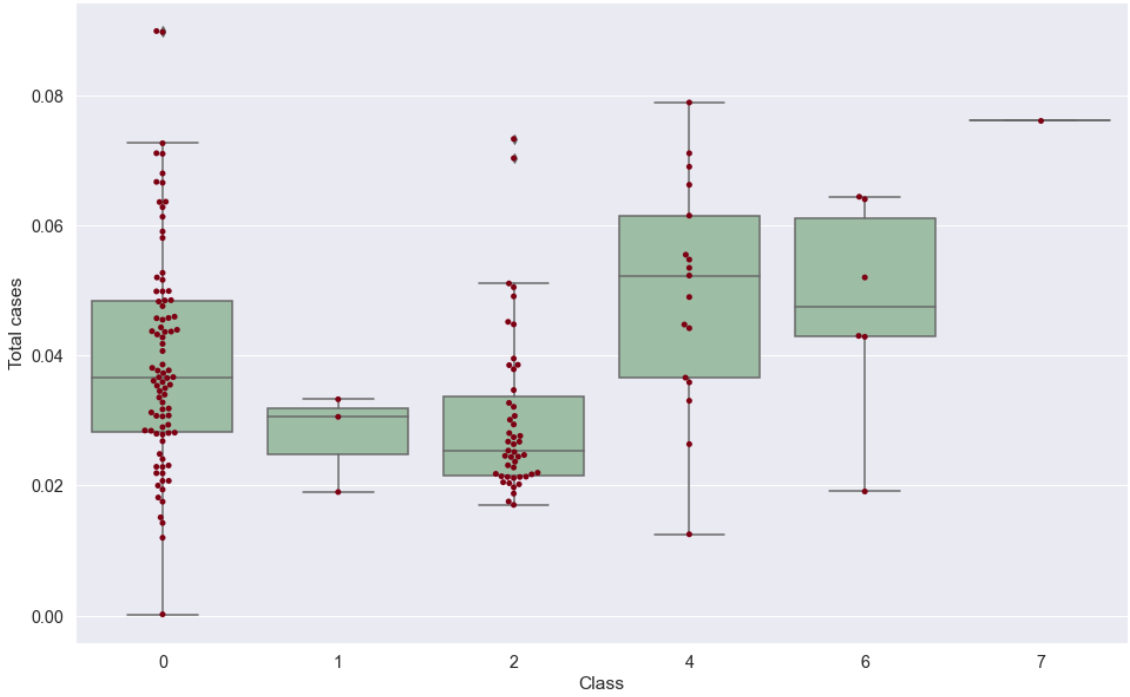

Figure 13 – Boxplots

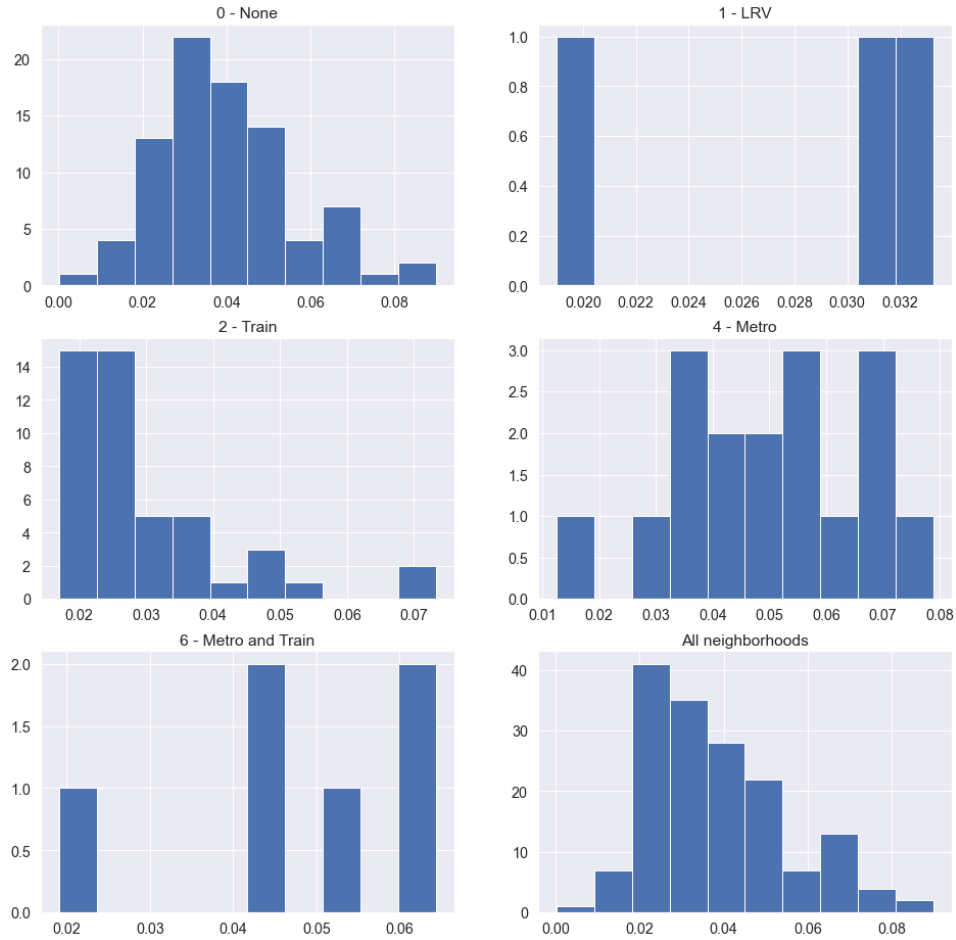

Figure 14 – Histograms

The boxplots and histograms indicate that data may not follow a normal distribution. Histograms for the whole data and for trains and no transport modes (none) are skewed to right. So, let's do a few more tests. First, I look at the quantile-quantile (QQ) plots for some visual representation of the data.

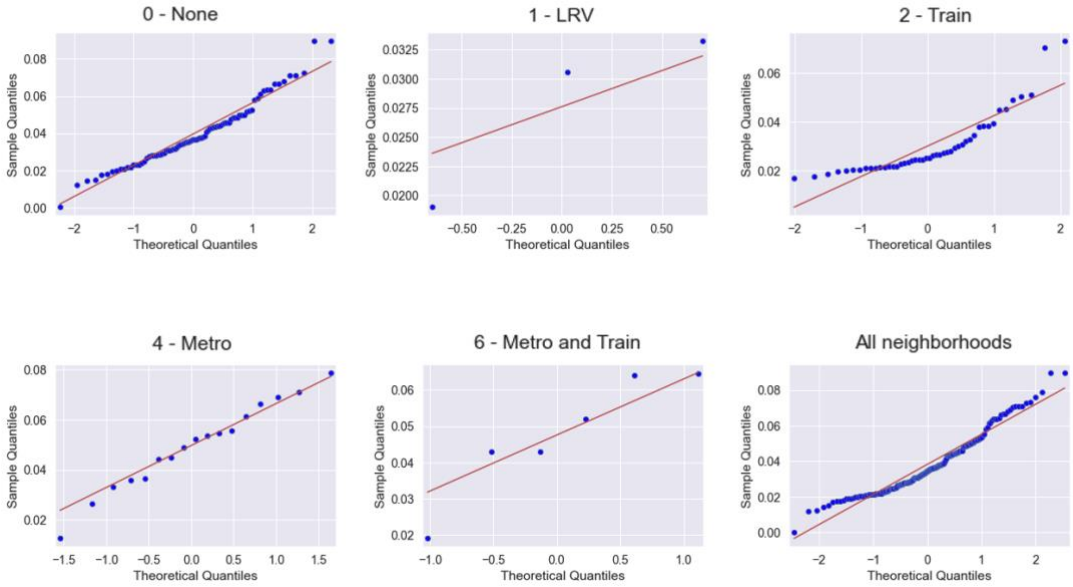

Figure 15 – QQ-plots

The QQ-plots also indicate that some of the data subgroups, as well as the whole dataset is skew to the right. I also use the Shapiro-Wilk to test normality. The results of the test are shown in Table 1.

Table 1 – Results of the Shapiro-Wilk

Subgroup / Class	F-statistic	P-value
No metro, train, or LRV	0.966	0.021
Only LRV	0.886	0.342
Only Train	0.795	1.251
Metro	0.984	0.985
Metro and Train	0.954	0.765
All neighborhoods	0.948	1.186

The results confirm my first observation. The whole data, the subgroup no transport modes (0), and the subgroup trains (2) are not normally distributed because their P-values are smaller than 0.05. However, even if the distribution of the individual observations is not normal, the distribution of the sample means will be normally distributed if your sample size is about 30 or larger. This is due to the "central limit theorem" that shows that even when a population is non-normally distributed, the distribution of the "sample means" will be normally distributed when the sample size is 30 or more. Therefore, given that subgroups (0) and (2) are both larger than 30, we can use the one-way ANOVA test in our data.

Finally, I check for homoscedasticity given that our sample sizes are very unequal. We use two tests: Bartlett and Leven. The results are shown in Table 2.

Table 2 – Results of the Leven and Bartlett tests

Test	F-statistic	P-value
Leven	1.887	0.116
Bartlett	6.696	0.153

The results show that both tests have failed to reject the null hypothesis. Therefore, the hypothesis of equal population variances is valid. Now, we can finally perform the one-way ANOVA test. The results are shown in Table 3.

Table 3 – Results of the one-way ANOVA test

F-statistic	P-value
7.095	2.849 e-05

The P-value obtained is significant ($P<0.05$). But to conclude the analysis, we need to calculate the F-critical. If the F-statistic is greater than the F-critical, we can reject the null hypothesis. The result of the calculation of the F-critical is shown in Figure 16.

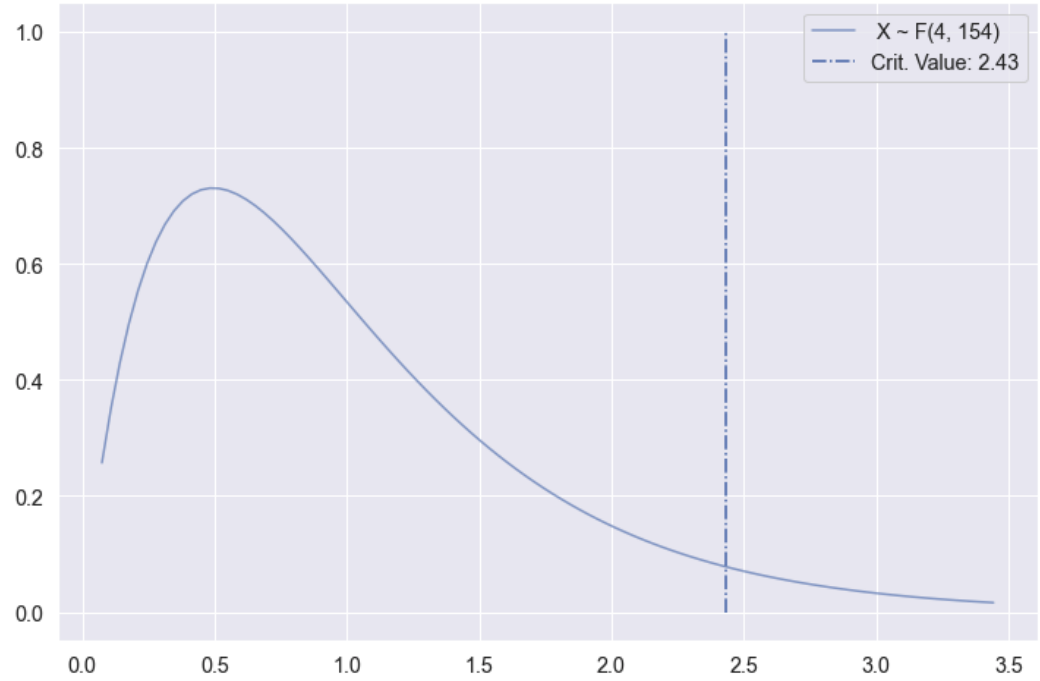

Figure 16 – One-way ANOVA test F-distribution

The F-statistic of 7.10 is greater than the F-critical of 2.43, thus we can conclude that there are significant differences among the number of COVID-19 cases in Rio de Janeiro neighborhoods according to the mode of transport (metro, train, LRV) present in the neighborhood.

As detailed in the Methodology, the one-way ANOVA test tells us if our results or significant or not but does not tell us where the results are significant. The most common solution to this problem is using Tukey's Honestly Significant Difference (HSD) procedure. The results of the Tukey's HSD procedure are present in Table 4.

Table 4 – Results of the Tukey's HSD procedure

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
0	1	-0.012	0.6729	-0.0376	0.0136	False
0	2	-0.0095	0.0102	-0.0174	-0.0016	True
0	4	0.0101	0.1188	-0.0015	0.0217	False
0	6	0.0121	0.3004	-0.0051	0.0292	False
1	2	0.0025	0.9	-0.0235	0.0285	False
1	4	0.0221	0.1735	-0.0052	0.0494	False
1	6	0.024	0.1832	-0.0061	0.0541	False
2	4	0.0196	0.001	0.0072	0.0319	True
2	6	0.0215	0.0085	0.0039	0.0392	True
4	6	0.002	0.9	-0.0176	0.0215	False

Tukey's HSD results indicates that the significant statistical difference is occurring between group (2) and groups (0), (4), and (6). Thus, the presence of train stations in the neighborhood is the main statistical different group. This is coherent with the line plot of the exploratory analysis displaying the cumulative number of COVID-19 cases that indicated that the number of COVID-19 cases in neighborhoods with access to train was lower than in most of the other neighborhoods. Besides, the lack of significant statistical difference between the neighborhoods with access to metro and those that do have access to it, may indicate that metro is not a significant vector in the dispersion of COVID-19.

I also use the one-way ANOVA test to check if the presence or not of different modes of public transportation (metro, train, LRV) in Rio de Janeiro neighborhoods have no impact in the number of COVID-19 cases in this neighborhood for each month in the period analyzed (01/2020 to 04/2021). The results are shown in Table 5.

Table 5 – Results of the one-way ANOVA test

Period	F-statistic	P-value	Reject
01/2020	1.718	0.149	False
02/2020	1.712	0.150	False
03/2020	7.168	2.539 e-05	True
04/2020	5.169	6.188 e-04	True
05/2020	5.664	2.789 e-04	True
06/2020	5.250	0.001	True
07/2020	8.158	5.371 e-06	True
08/2020	5.542	3.391 e-04	True
09/2020	4.659	0.001	True
10/2020	4.244	0.003	True
11/2020	5.999	1.630 e-04	True
12/2020	4.861	0.001	True
01/2021	6.971	3.466 e-05	True

02/2021	4.259	0.003	True
03/2021	2.231	0.068	False
04/2021	3.123	0.017	True

The presence or not of different modes of public transportation (metro, train, LRV) in Rio de Janeiro neighborhoods have an impact in the number of COVID-19 cases in this neighborhood for most months during the pandemic outbreak. The exceptions are the first two months, when the pandemic was still starting and March 2021. A possible explanation is that in March 2021 a strong lock-down was established in Rio.

Finally, we perform the Tukey Test months for which the null hypotheses were rejected. The results are shown in Tables 6 to 18.

Table 6 – Results of the Tukey's HSD procedure for 03/2020

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
0	1	-0.0002	0.9	-0.0011	0.0006	False
0	2	-0.0002	0.3221	-0.0005	0.0001	False
0	4	0.0006	0.001	0.0002	0.001	True
0	6	0.0002	0.8748	-0.0004	0.0008	False
1	2	0.0001	0.9	-0.0008	0.0009	False
1	4	0.0008	0.0947	-0.0001	0.0018	False
1	6	0.0004	0.7425	-0.0006	0.0014	False
2	4	0.0008	0.001	0.0004	0.0012	True
2	6	0.0004	0.4045	-0.0002	0.001	False
4	6	-0.0004	0.44	-0.0011	0.0003	False

Table 7 – Results of the Tukey's HSD procedure for 04/2020

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
0	1	-0.0015	0.5853	-0.0044	0.0014	False
0	2	-0.0007	0.1587	-0.0016	0.0002	False
0	4	0.0013	0.0418	0.0	0.0027	True
0	6	0.0007	0.8136	-0.0012	0.0027	False

1	2	0.0008	0.9	-0.0022	0.0037	False
1	4	0.0029	0.0841	-0.0002	0.006	False
1	6	0.0023	0.3633	-0.0012	0.0057	False
2	4	0.0021	0.001	0.0007	0.0035	True
2	6	0.0015	0.2568	-0.0005	0.0035	False
4	6	-0.0006	0.9	-0.0028	0.0016	False

Table 8 – Results of the Tukey's HSD procedure for 05/2020

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
0	1	-0.0012	0.6553	-0.0038	0.0013	False
0	2	-0.0008	0.0323	-0.0016	-0.0	True
0	4	0.001	0.1678	-0.0002	0.0021	False
0	6	0.001	0.524	-0.0008	0.0027	False
1	2	0.0004	0.9	-0.0022	0.003	False
1	4	0.0022	0.1874	-0.0006	0.005	False
1	6	0.0022	0.265	-0.0008	0.0053	False
2	4	0.0018	0.001	0.0006	0.003	True
2	6	0.0018	0.0429	0.0	0.0036	True
4	6	0.0	0.9	-0.002	0.002	False

Table 9 – Results of the Tukey's HSD procedure for 06/2020

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
0	1	-0.0005	0.7105	-0.0016	0.0006	False
0	2	-0.0004	0.0518	-0.0007	0.0	False
0	4	0.0004	0.2119	-0.0001	0.0009	False
0	6	0.0005	0.4389	-0.0003	0.0012	False
1	2	0.0002	0.9	-0.001	0.0013	False
1	4	0.0009	0.246	-0.0003	0.0021	False
1	6	0.001	0.2657	-0.0004	0.0023	False
2	4	0.0007	0.0022	0.0002	0.0013	True
2	6	0.0008	0.0363	0.0	0.0016	True
4	6	0.0001	0.9	-0.0008	0.0009	False

Table 10 – Results of the Tukey's HSD procedure for 07/2020

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
0	1	-0.0008	0.7446	-0.0026	0.0011	False
0	2	-0.0008	0.0033	-0.0013	-0.0002	True
0	4	0.0007	0.148	-0.0001	0.0015	False
0	6	0.0011	0.1363	-0.0002	0.0023	False
1	2	0.0	0.9	-0.0018	0.0019	False
1	4	0.0015	0.2381	-0.0005	0.0035	False
1	6	0.0018	0.1418	-0.0003	0.004	False
2	4	0.0015	0.001	0.0006	0.0023	True
2	6	0.0018	0.0013	0.0005	0.0031	True
4	6	0.0004	0.9	-0.0011	0.0018	False

Table 11 – Results of the Tukey's HSD procedure for 08/2020

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
0	1	-0.0004	0.9	-0.002	0.0012	False
0	2	-0.0006	0.0103	-0.0011	-0.0001	True
0	4	0.0006	0.1962	-0.0002	0.0013	False
0	6	0.0004	0.8667	-0.0007	0.0014	False
1	2	-0.0002	0.9	-0.0018	0.0014	False
1	4	0.001	0.5241	-0.0007	0.0027	False
1	6	0.0008	0.7715	-0.0011	0.0027	False
2	4	0.0012	0.001	0.0004	0.0019	True
2	6	0.001	0.1214	-0.0001	0.0021	False
4	6	-0.0002	0.9	-0.0014	0.001	False

Table 12 – Results of the Tukey's HSD procedure for 09/2020

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
0	1	-0.0006	0.9	-0.0027	0.0016	False
0	2	-0.0007	0.0628	-0.0013	0.0	False
0	4	0.0008	0.1467	-0.0002	0.0018	False
0	6	0.0006	0.7673	-0.0009	0.0021	False
1	2	-0.0001	0.9	-0.0023	0.0021	False
1	4	0.0014	0.4772	-0.001	0.0037	False
1	6	0.0012	0.7041	-0.0014	0.0037	False
2	4	0.0015	0.0014	0.0004	0.0025	True

2	6	0.0013	0.1544	-0.0003	0.0028	False
4	6	-0.0002	0.9	-0.0019	0.0014	False

Table 13 – Results of the Tukey's HSD procedure for 10/2020

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
0	1	-0.0008	0.8722	-0.0031	0.0015	False
0	2	-0.0008	0.0212	-0.0015	-0.0001	True
0	4	0.0006	0.452	-0.0004	0.0017	False
0	6	0.0003	0.9	-0.0012	0.0018	False
1	2	-0.0	0.9	-0.0023	0.0023	False
1	4	0.0014	0.5042	-0.001	0.0038	False
1	6	0.0011	0.773	-0.0016	0.0038	False
2	4	0.0014	0.0048	0.0003	0.0025	True
2	6	0.0011	0.3159	-0.0005	0.0027	False
4	6	-0.0003	0.9	-0.0021	0.0014	False

Table 14 – Results of the Tukey's HSD procedure for 11/2020

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
0	1	-0.0021	0.5745	-0.0061	0.0019	False
0	2	-0.0017	0.002	-0.0029	-0.0005	True
0	4	0.0009	0.5875	-0.0009	0.0027	False
0	6	0.001	0.8127	-0.0017	0.0037	False
1	2	0.0004	0.9	-0.0036	0.0045	False
1	4	0.0031	0.2772	-0.0012	0.0073	False
1	6	0.0031	0.3535	-0.0016	0.0078	False
2	4	0.0026	0.0021	0.0007	0.0045	True
2	6	0.0027	0.0575	-0.0001	0.0054	False
4	6	0.0001	0.9	-0.003	0.0031	False

Table 15 – Results of the Tukey's HSD procedure for 12/2020

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
0	1	-0.002	0.5357	-0.0057	0.0016	False
0	2	-0.0012	0.0381	-0.0023	-0.0	True

0	4	0.0012	0.2678	-0.0005	0.0029	False
0	6	0.0009	0.8579	-0.0016	0.0033	False
1	2	0.0009	0.9	-0.0028	0.0046	False
1	4	0.0032	0.1536	-0.0007	0.0071	False
1	6	0.0029	0.3461	-0.0014	0.0072	False
2	4	0.0024	0.0026	0.0006	0.0041	True
2	6	0.002	0.1792	-0.0005	0.0046	False
4	6	-0.0003	0.9	-0.0031	0.0025	False

Table 16 – Results of the Tukey's HSD procedure for 01/2021

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
0	1	-0.0013	0.4339	-0.0034	0.0008	False
0	2	-0.0006	0.0758	-0.0013	0.0	False
0	4	0.0009	0.0545	-0.0	0.0019	False
0	6	0.0012	0.1468	-0.0002	0.0026	False
1	2	0.0007	0.8978	-0.0014	0.0028	False
1	4	0.0022	0.0503	-0.0	0.0045	False
1	6	0.0025	0.0487	0.0	0.0049	True
2	4	0.0015	0.001	0.0005	0.0026	True
2	6	0.0018	0.0074	0.0003	0.0032	True
4	6	0.0002	0.9	-0.0014	0.0018	False

Table 17 – Results of the Tukey's HSD procedure for 02/2021

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
0	1	-0.001	0.433	-0.0025	0.0006	False
0	2	-0.0004	0.1951	-0.0009	0.0001	False
0	4	0.0005	0.3419	-0.0002	0.0012	False
0	6	0.0006	0.4589	-0.0004	0.0017	False
1	2	0.0006	0.8242	-0.001	0.0022	False
1	4	0.0014	0.123	-0.0002	0.0031	False
1	6	0.0016	0.1199	-0.0002	0.0034	False
2	4	0.0009	0.017	0.0001	0.0016	True
2	6	0.001	0.0773	-0.0001	0.0021	False
4	6	0.0002	0.9	-0.001	0.0014	False

Table 18 – Results of the Tukey's HSD procedure for 04/2021

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
0	1	-0.0002	0.9	-0.0022	0.0018	False
0	2	-0.0002	0.8587	-0.0008	0.0004	False
0	4	0.0003	0.8924	-0.0006	0.0012	False
0	6	0.0015	0.0222	0.0001	0.0028	True
1	2	0.0	0.9	-0.002	0.002	False
1	4	0.0005	0.9	-0.0016	0.0027	False
1	6	0.0017	0.2703	-0.0006	0.0041	False
2	4	0.0005	0.5764	-0.0005	0.0015	False
2	6	0.0017	0.0078	0.0003	0.0031	True
4	6	0.0012	0.2096	-0.0003	0.0027	False

Tukey's HSD results indicates a significant difference between neighborhood with access to metro and those without access to it only for March/2020 and April/2020. This may indicate that the metro was only significant in the diffusion of COVID-19 in the start of the pandemic in Rio. After that, the reduced use of this transport mode associated with sanitary measures by Rio de Janeiro government may have reduced this issue.

5 Conclusion

The results of the statistical analysis show that there is no significant difference between the number of COVID-19 per habitant for neighborhoods with and without access to metro, train, and LRV in Rio de Janeiro. The exceptions are the months of March and April 2020, right in the beginning of the pandemic, when the neighborhoods with access to metro were significantly more impacted than those without access to this mode of transportation. One possible explanation is that the sharp decrease in metro, train, and LRV users since March 2021 until now associated with the sanitary measures adopted by the Rio de Janeiro government have minimized the impact that public transportation could have in the diffusion of COVID-19.

Nonetheless, we may not affirm that public transportation was not a significant vector of diffusion of COVID-19 in Rio de Janeiro because the present analysis did not consider the bus, which is an important mode of public transportation in Rio de Janeiro. Therefore, a possible future path the advancing the present study is to include data from the bus system of Rio de Janeiro.