# Capstone Project 1 – Statistical Analysis

Performing a set of statistical tests on the four clusters of Grids we got from our clustering model helps us to understand the behavior and patterns of these clusters. The one-way analysis of variance (**ANOVA**) is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups. A t-test is used to determine if there is a significant difference between the means of two groups.

The null hypothesis and alternate hypothesis for both the tests are stated as,

Null Hypothesis: The mean is the same for all groups.

Alternate Hypothesis: The mean is not the same for all groups.

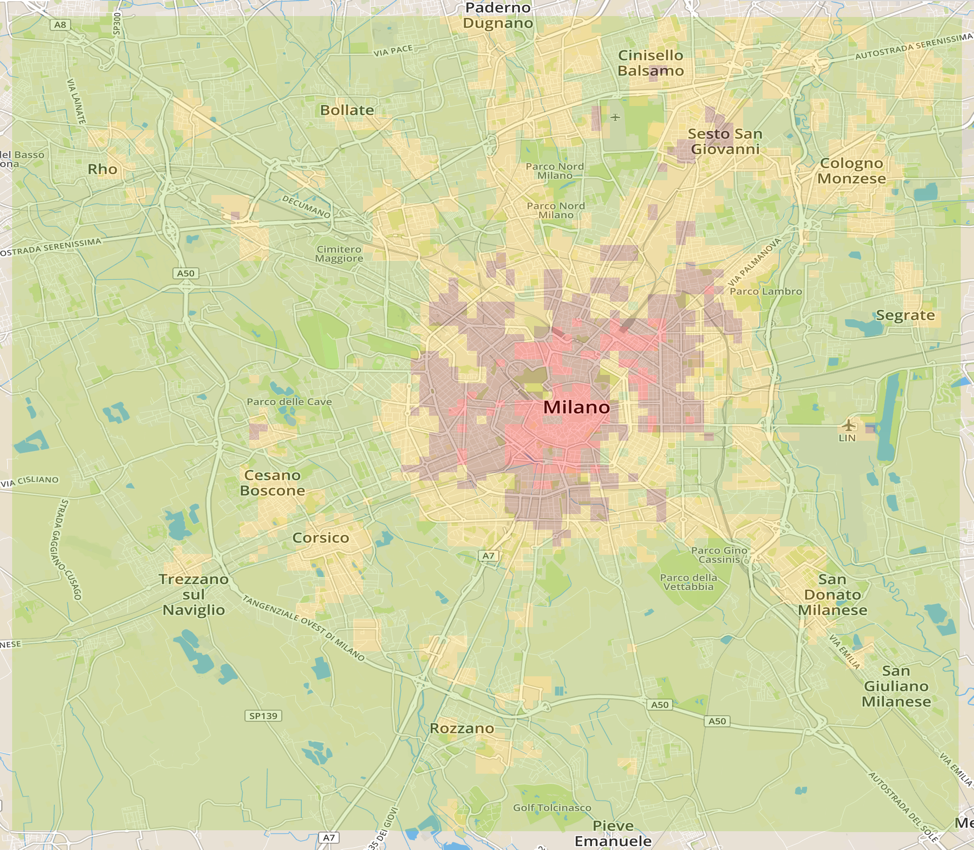
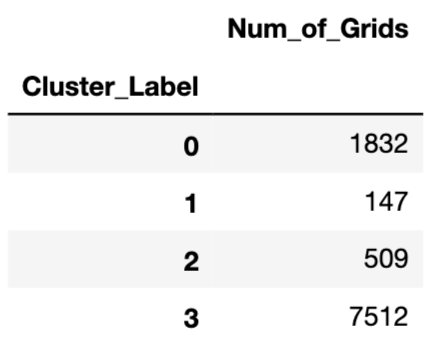
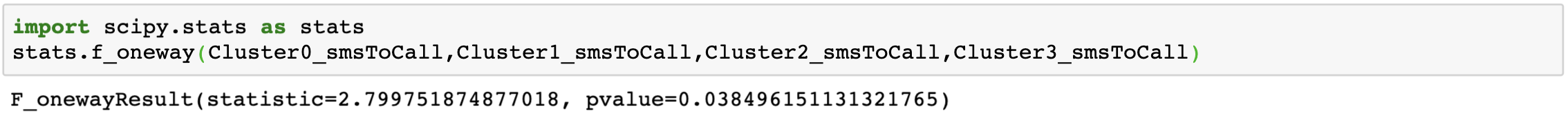
 

Fig: Four clusters from our clustering model, represented in geojson file

## ANOVA test 1:

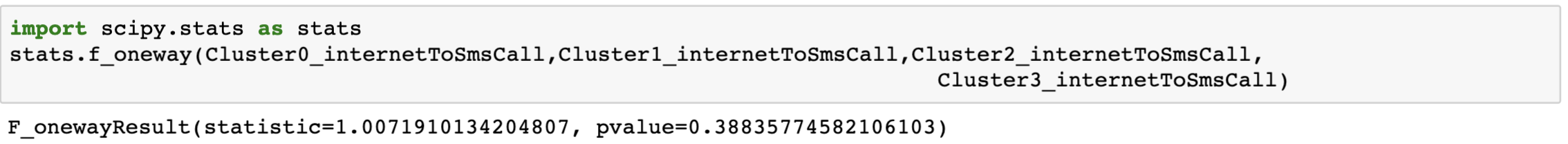
SMS to Call Ratio: Is the ratio of total number of SMS to total number of Calls handled by the Grids from 11/01/2013 – 01/01/2014.



Performing a one-way ANOVA test on the four cluster’s SMS to Call ratios results in a P-value < 0.05, hence we reject the Null hypothesis. Which means that the average of SMS to Call ratio is different for the 4 clusters.

### ANOVA test 2:

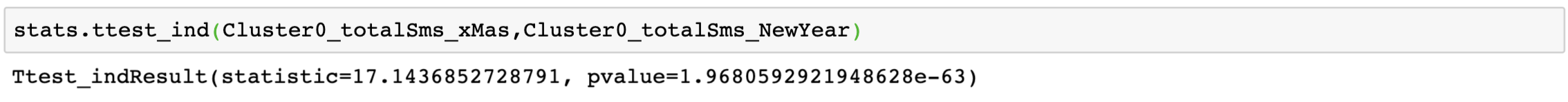
Internet to SMS & Call Ratio: Is the ratio of total number of Internet activity to combined total of SMS & Calls handled by the Grids from 11/01/2013 – 01/01/2014

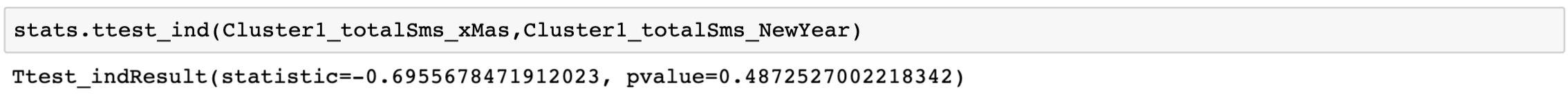


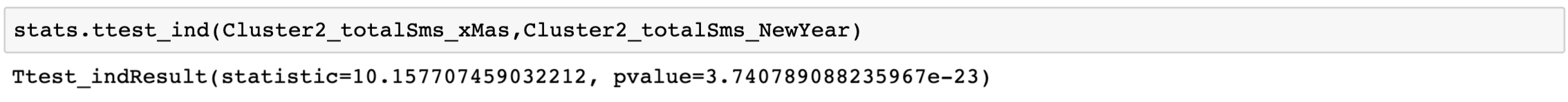
Performing a one-way ANOVA test on the four cluster’s Internet to SMS & Call ratios results in a P-value > 0.05, hence we conclude that there is enough evidence that the mean of Internet to SMS & Call Ratio is same for all the 4 clusters. This could be because Milan is a metropolitan city in Italy, with good internet connectivity throughout the city. We see that all 4 clusters have a mean Internet to SMS & Call Ratio ~ 5.0.

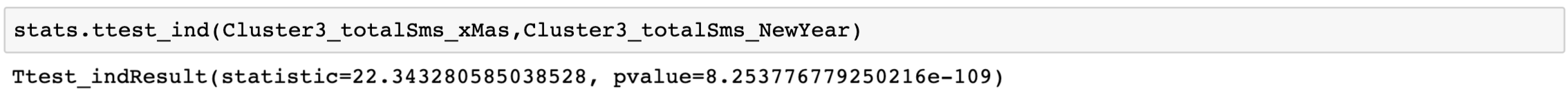
### T-test 1:

SMS: A comparison of the SMS volumes handled during Christmas & New Year holidays, by each cluster.



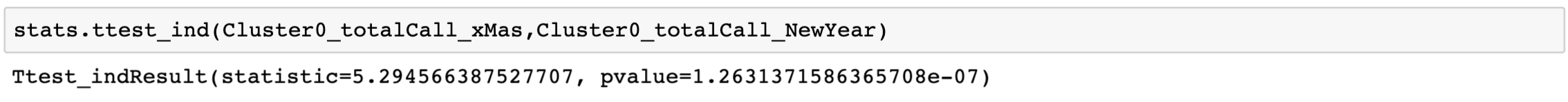


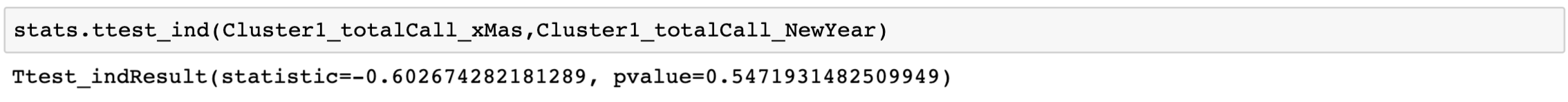


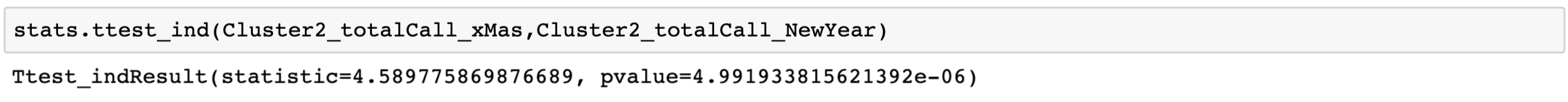


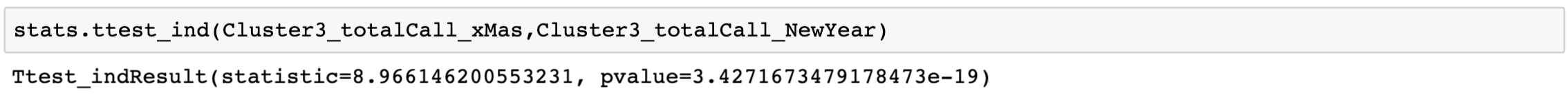
### T-test 2:

Call: A comparison of the Call volumes handled during Christmas & New Year holidays, by each cluster.



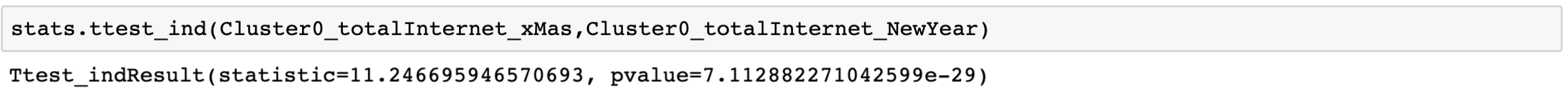


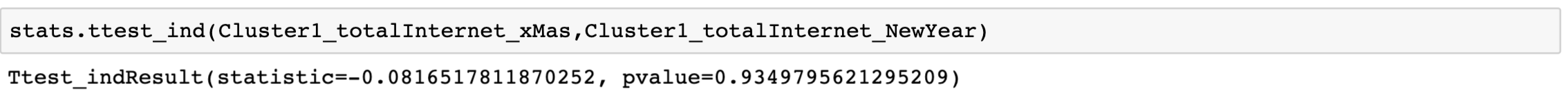


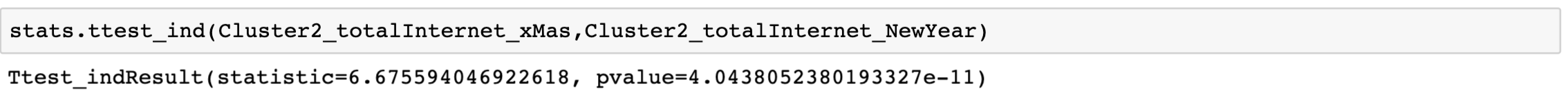


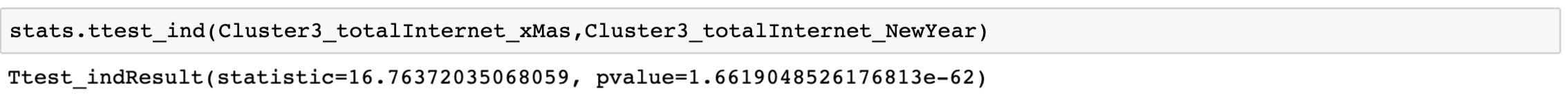
### T-test 3:

Internet: A comparison of the Internet volumes handled during Christmas & New Year holidays, by each cluster.









Cluster0, Cluster2, Cluster3 have P-value < 0.05 for all three tests. So, we reject the null hypothesis and conclude that the mean volumes of SMS, Calls and Internet activity on Christmas day and New Year day are different. Interestingly, it is not so for Cluster1. For all the three telecom activities, respective t-tests results in a P-value > 0.05. The average volumes of SMS, Calls and Internet activity on Christmas and New Year are the same. This could be because Cluster1 consists of Grids with high telecommunication activities in general, and a quick look at these Grids in the geojson file shows they are also mostly the transport hubs.