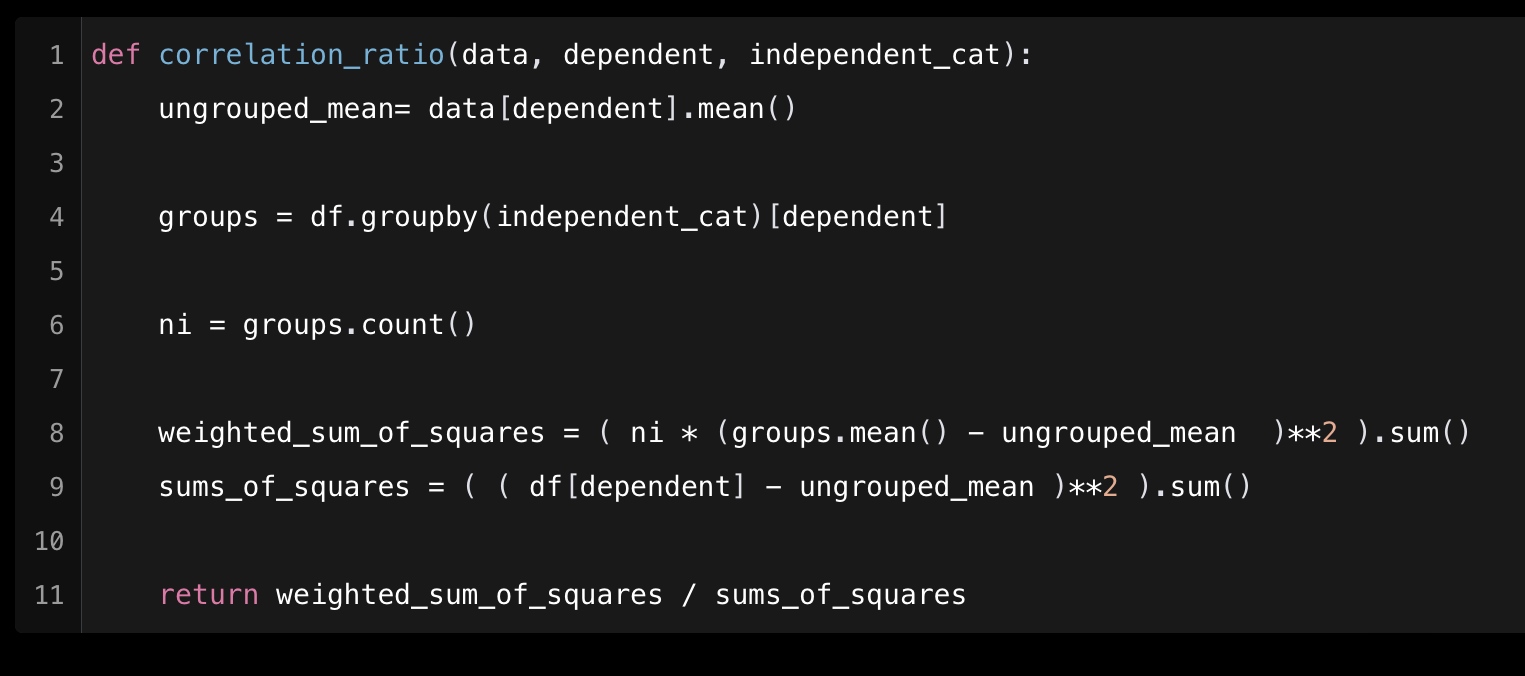
**Note:** This document forgoes the introduction and is intended as an extension of the previous deliveries.

**Measures of Relationships**

Firstly, we analyze if a correlation between city zones and the use of public transport exists, as we theorized on our visualizations.

Unfortunately, typical correlation measures such as Pearson's coefficient cannot apply here since one of our variables is categorical. To measure correlation, we will use the correlation ratio, or . We define the following function to calculate it. ([source](https://en.wikipedia.org/wiki/Correlation_ratio))

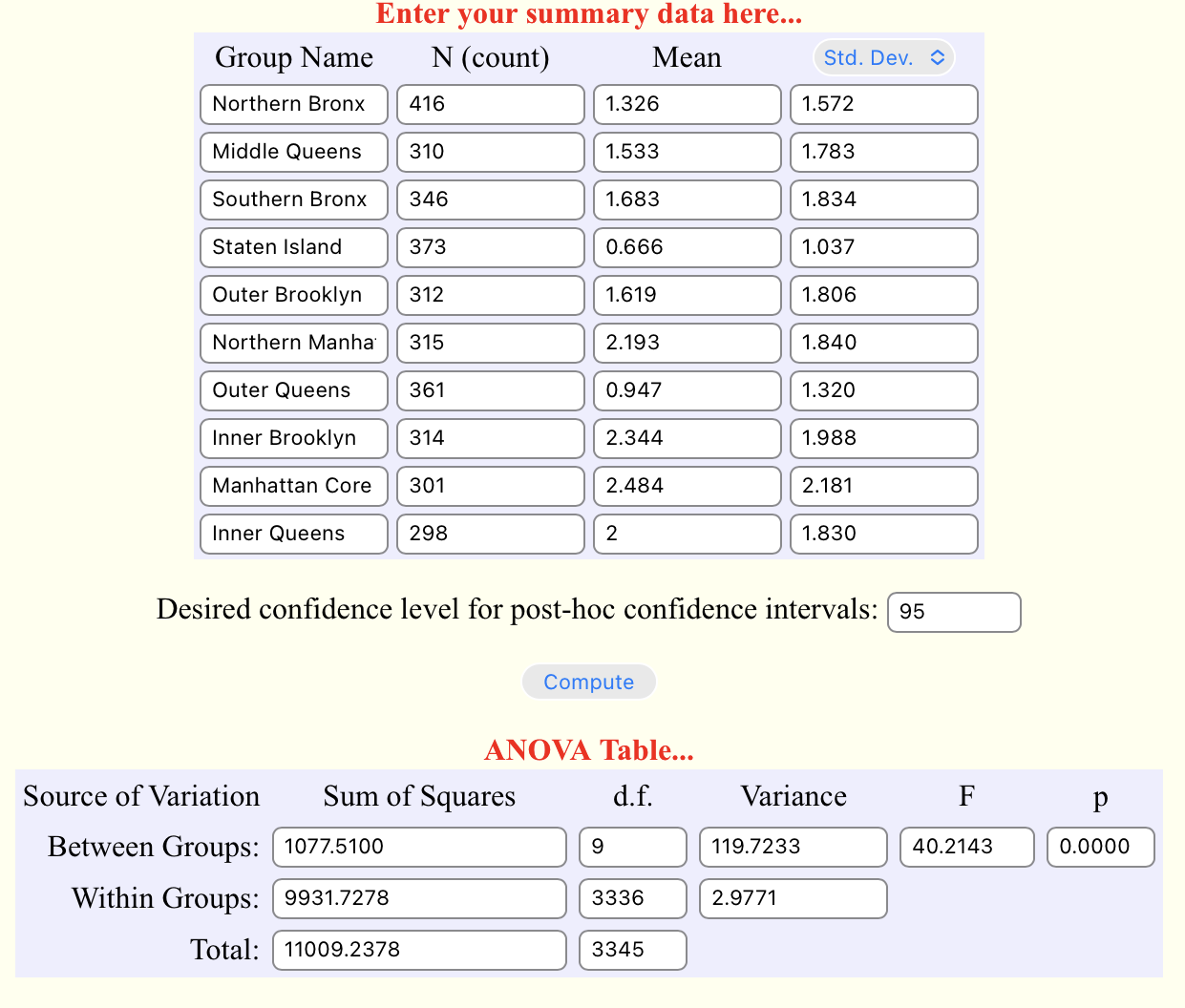


*Fig 1. Function to calculate the correlation ratio.*

With the function defined, we calculate the correlation ratio between ‘mean\_num\_walk\_trips\_per\_day’ and ‘cms\_zone’. Taking the square root of the value returned by the function we obtain 0.313 as our ratio.

We see a mid-to-low level of correlation, which could be due to the fact that if we recall the violin plot for ‘mean\_num\_walk\_trips\_per\_day’ and ‘cms\_zone’, only two of the groups clearly stood out (Staten Island and Outer Queens) while the others looked relatively simmilar. We can also explore this correlation ratio for ‘mean\_num\_car\_trips\_per\_day’, which showed no difference in visualization. The metric returned for that is 0.176, indicating a very weak correlation.

To ensure our first pair of variables are correlated, we perform hypothesis testing on them. We postulate as the null hypothesis, stating there is no difference in the means of the different groups, and or the alternative hypothesis that there is a difference depending on the group. We perform ANOVA testing to answer this question, since we are dealing with many different groups.

*Fig 2. ANOVA Calculation table results.*

Computing mean and standard deviation for all groups, as well as using the number of instances on each group we perform a One-Way ANOVA to test our hypothesis. The One-Way ANOVA reveals that, within a 99% confidence interval, there is a statistically signifficant difference between at least two of the 10 studied groups

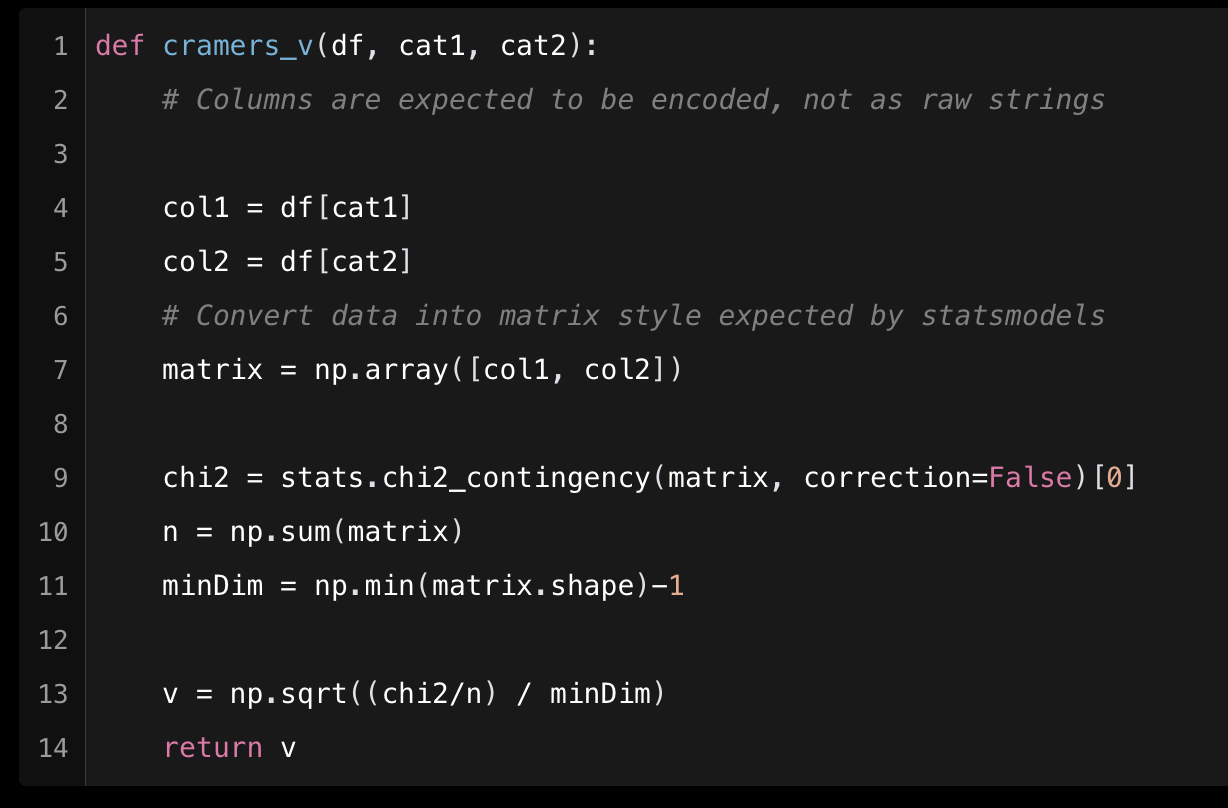
We therefore reject the Null Hypothesis that stated there was no statistically significant difference between the groups. Furthermore, since the p-value is 0 we can state there exists a very statisticall

If we are interested in only two of the groups, we perform Welch’s T-Test, since we are testing on populations with different number of samples and variance. For example, we will look at if inside the district of Queens, if a difference exists in means depending on if we select a ‘cms\_zone’ closer to the center (Inner Queens) or not (Outer Queens)

This reveals that there is a significant difference in ‘mean\_num\_walk\_trips\_per\_day’ between the two groups, ‘Inner Queens’ and ‘Outer Queens’ ;

The T-Test reveals that within a 99% confidence interval, there is a statistically signifficant difference between both groups. We therefore reject the Null Hypothesis and state there is a very statistically signifficant difference.

Continuing on this line we can also take a look at the correlation between ‘work\_mode’ and ‘employment’. Like before, classic correlation metrics do not work since both are really categorical nominal variables. To analyze if a correlation exists, we use [Cramér's V](https://en.wikipedia.org/wiki/Cramér%27s_V), also known as . We define the following function to calculate it. ([source](https://www.statology.org/cramers-v-in-python/)).

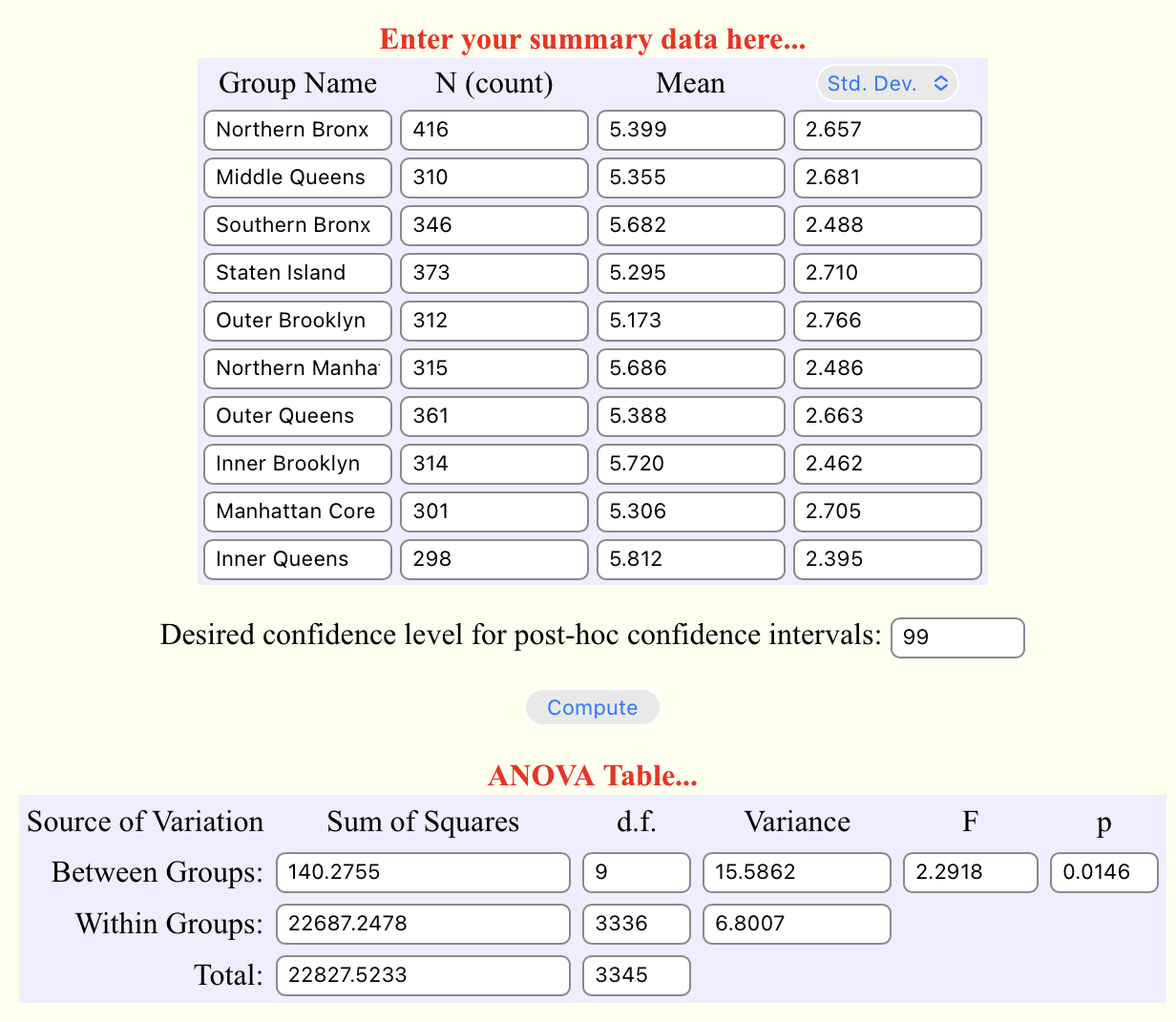


*Fig 3. Function to calculate Cramér’s V*

This correlation metric returns a value of 0.161, indicating that the two might not be as related as we thought.

Following that, we will look at if walking is preferred over the use of public transport city wide. To do this we perform a Student T-Test to study if there is a signifficant difference between the means.

The test returns a value of which with a 99% confidence we can say rejects the null hypothesis and suggests a very strong signifficance in the difference between means.

To close, we will look at the paricipation level of survey respondents by city zone. To do this, we do an ANOVA test of the number of days a subject participated between the different city zones, with as the null hypothesis stating there is no difference in the means of the groups, and as the alternative hypothesis stating there is at least a difference between two osf the groups.

The ANOVA test reveals that, within a 99% confidence interval, we fail to reject the null hypothesis and no statistically signifficant difference is found among the 10 groups.

This reveals there was no difference in participation levels among participants from different city zones.