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 η^2. We define the following function to calculate it. (source)

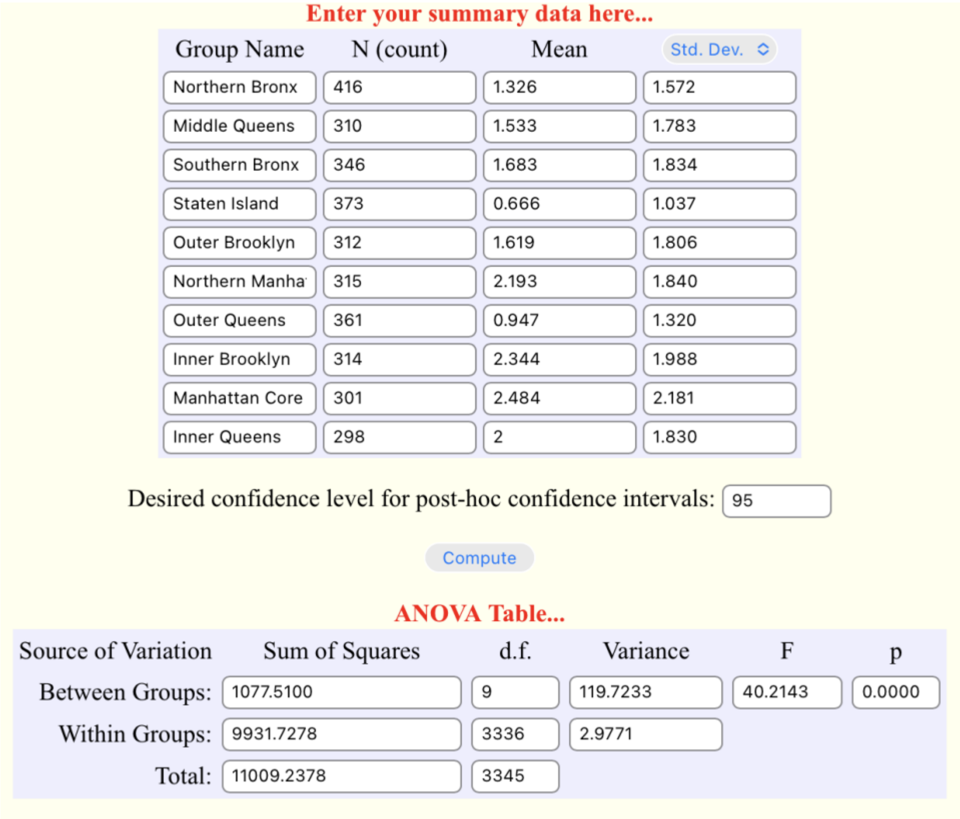
```
def correlation_ratio(data, dependent, independent_cat):
    ungrouped_mean= data[dependent].mean()

    groups = df.groupby(independent_cat)[dependent]

    ni = groups.count()

    weighted_sum_of_squares = ( ni * (groups.mean() - ungrouped_mean  )**2 ).sum()
    sums_of_squares = ( ( df[dependent] - ungrouped_mean )**2 ).sum()

    return weighted_sum_of_squares / sums_of_squares
```

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 H_0 as the null hypothesis, stating there is no difference in the means of the different groups, and H_1or 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.



Enter your summary data here...

Group Name	N (count)	Mean	Std. Dev.
Northern Bronx	416	1.326	1.572
Middle Queens	310	1.533	1.783
Southern Bronx	346	1.683	1.834
Staten Island	373	0.666	1.037
Outer Brooklyn	312	1.619	1.806
Northern Manha	315	2.193	1.840
Outer Queens	361	0.947	1.320
Inner Brooklyn	314	2.344	1.988
Manhattan Core	301	2.484	2.181
Inner Queens	298	2	1.830

Desired confidence level for post-hoc confidence intervals: 95

Compute

ANOVA Table...

Source of Variation	Sum of Squares	d.f.	Variance	F	p
Between Groups:	1077.5100	9	119.7233	40.2143	0.0000
Within Groups:	9931.7278	3336	2.9771		
Total:	11009.2378	3345			

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 significant difference between at least two of the 10 studied groups

$$F(1077.51, 9931.72) = 40.213, p = 0$$

We therefore reject the Null Hypothesis H_0 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' ($M = 2, SD = 1{,}823$) and 'Outer Queens' ($M = 0.947, SD = 1.32$);

$$t = 8.305, p = 8.455 * 10^{-16}$$

The T-Test reveals that within a 99% confidence interval, there is a statistically significant difference between both groups. We therefore reject the Null Hypothesis H_0 and state there is a very statistically significant 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, also known as φ_c. We define the following function to calculate it. (source).

```python
def cramers_v(df, cat1, cat2):
    # Columns are expected to be encoded, not as raw strings

    col1 = df[cat1]
    col2 = df[cat2]
    # Convert data into matrix style expected by statsmodels
    matrix = np.array([col1, col2])

    chi2 = stats.chi2_contingency(matrix, correction=False)[0]
    n = np.sum(matrix)
    minDim = np.min(matrix.shape)-1

    v = np.sqrt((chi2/n) / minDim)
    return v
```

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 $p = 2.004e - 65$, which with a 99% confidence we can say rejects the null hypothesis H_0 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 H_0 as the null hypothesis stating there is no difference in the means of the groups, and H_1 as the alternative hypothesis stating there is at least a difference between two osf the groups.

Enter your summary data here...

Group Name	N (count)	Mean	Std. Dev.
Northern Bronx	416	5.399	2.657
Middle Queens	310	5.355	2.681
Southern Bronx	346	5.682	2.488
Staten Island	373	5.295	2.710
Outer Brooklyn	312	5.173	2.766
Northern Manha	315	5.686	2.486
Outer Queens	361	5.388	2.663
Inner Brooklyn	314	5.720	2.462
Manhattan Core	301	5.306	2.705
Inner Queens	298	5.812	2.395

Desired confidence level for post-hoc confidence intervals: 99

Compute

ANOVA Table...

Source of Variation	Sum of Squares	d.f.	Variance	F	p
Between Groups:	140.2755	9	15.5862	2.2918	0.0146
Within Groups:	22687.2478	3336	6.8007		
Total:	22827.5233	3345			

The ANOVA test reveals that, within a 99% confidence interval, we fail to reject the null hypothesis H_0 and no statistically signifficant difference is found among the 10 groups.

$$F(140.276, 22687.248) = 2.292, p = 0.0146$$

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