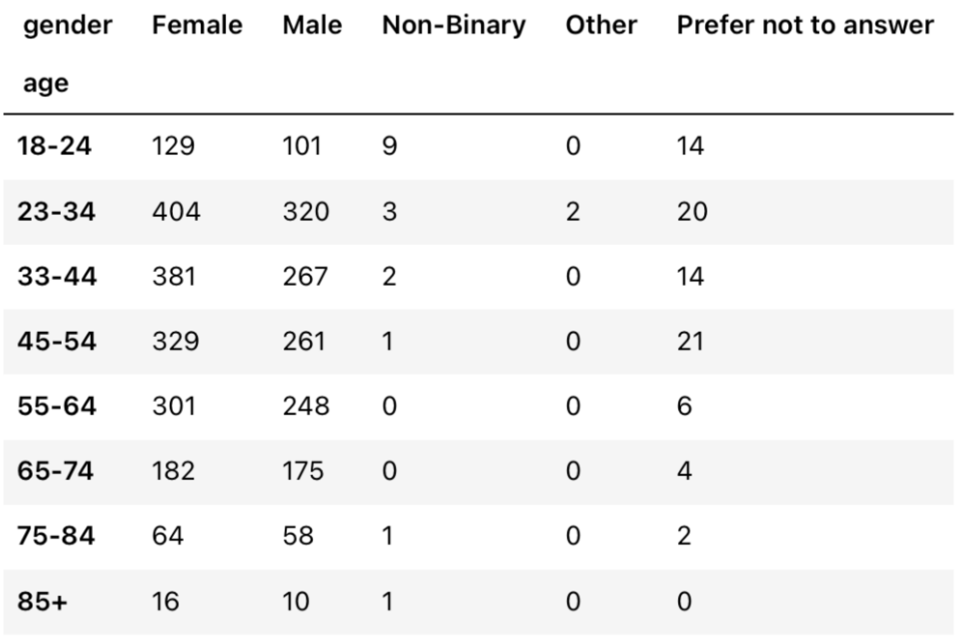
Note: This document forgoes the introduction and is intended as an extension of the Introductory document.

Exploratory Data Analysis

We look now at histogram views of a subset of trips per day using certain modes of transportation, which will help guide the hypothesis we want to test as well as gaining an understanding of the people who comprise our dataset.

The members of the dataset

In an effort to assert if the dataset sample is in fact balanced and representative of New York City's population, we will explore demographic data about the survey respondents as well as representation of each city zone.



gender	Female	Male	Non-Binary	Other	Prefer not to answer
age					
18-24	129	101	9	0	14
23-34	404	320	3	2	20
33-44	381	267	2	0	14
45-54	329	261	1	0	21
55-64	301	248	0	0	6
65-74	182	175	0	0	4
75-84	64	58	1	0	2
85+	16	10	1	0	0

Fig 1. Cross-Classification Table for the members of the dataset in terms of 'Age' and 'Gender'

We can see that the population of the dataset contains more Female than Male participants across all age groups, and that other genders have what would correspond as an equal representation on the dataset as in the general population.

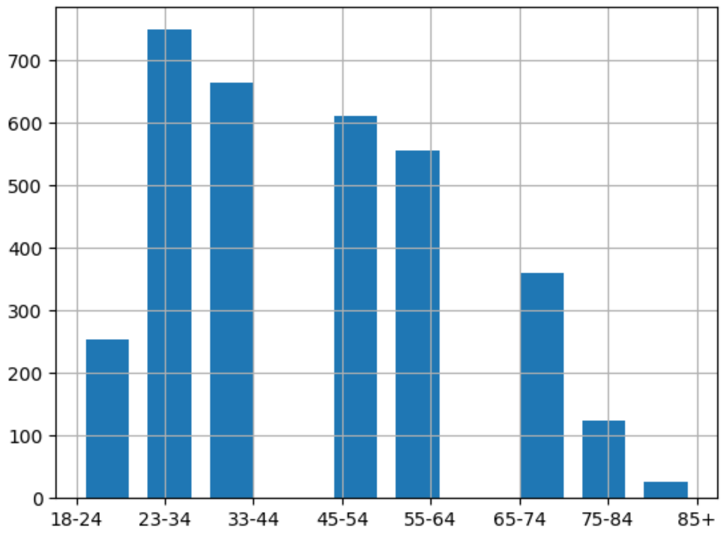


Fig 2. Age range histogram

This lines up with the typical age distribution for cities nowadays, and verifies the claim that the survey was representative of the population. Below is the breakdown of population in age ranges for 2021.

Note that the graph shown uses different age groups

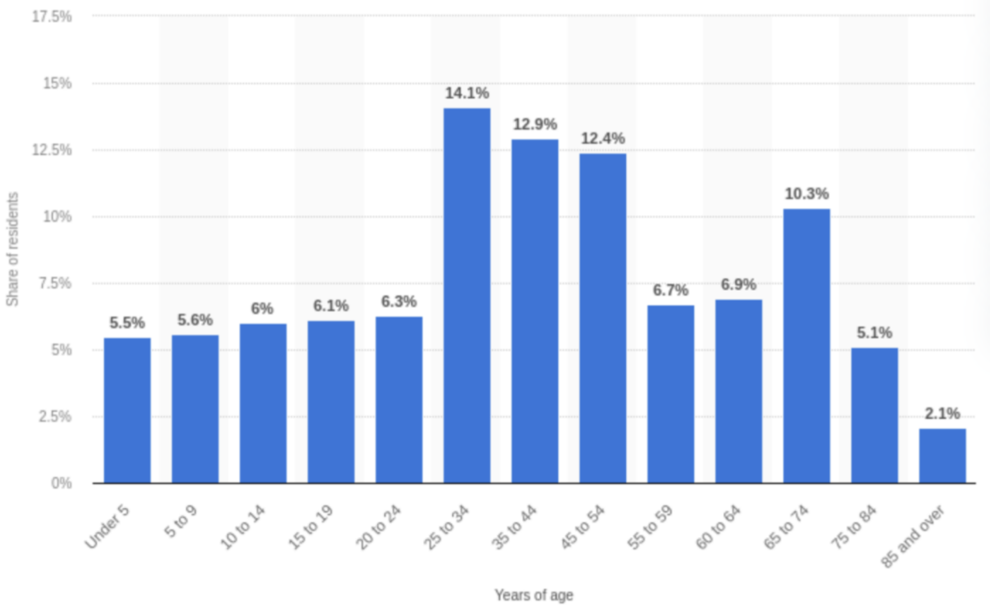
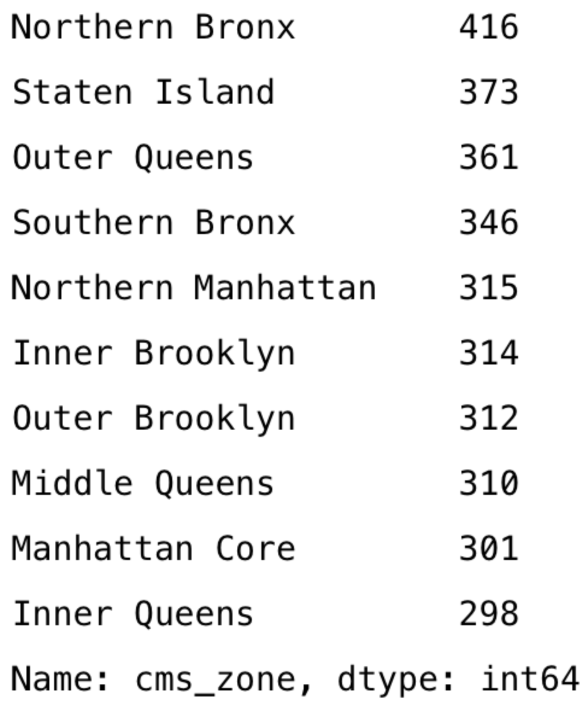


Fig 3. Age Histogram for New York City in 2021, gathered by Statista

Next, we explore the representation of each city area:



```
Northern Bronx        416
Staten Island         373
Outer Queens          361
Southern Bronx        346
Northern Manhattan    315
Inner Brooklyn        314
Outer Brooklyn        312
Middle Queens         310
Manhattan Core        301
Inner Queens          298
Name: cms_zone, dtype: int64
```

Fig 4. Instance counts for each city zone

Like before, the majority of zones have more or less the same representation. Northern Bronx, Staten Island, Outer Queens and Southern Bronx stand out as having more representation than the rest outside a reasonable margin. We will have to keep this overrepresentation in mind.

Stats per city zones

Next, having talked about the composition of our dataset, we will explore statistics regarding transport per city zones.

With the goal to guide our research question, we will look at the mean number of trips for certain modes of transportation per city zone.

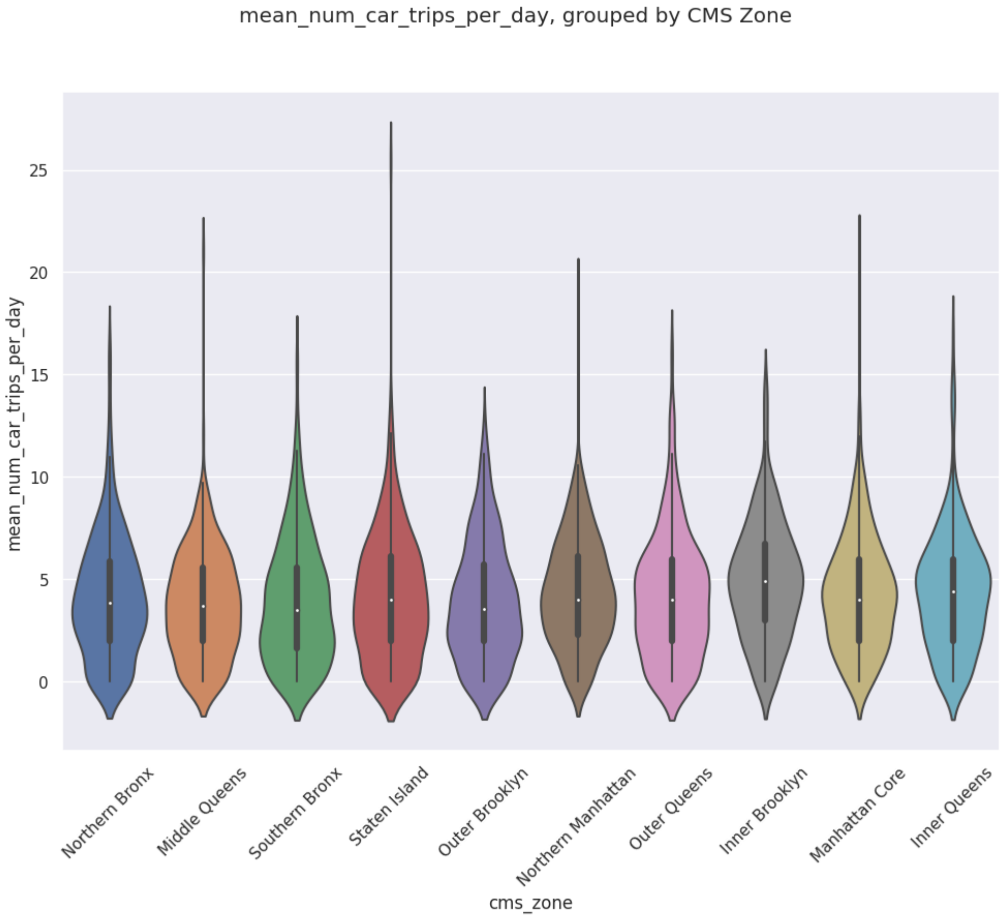


Fig 5. Violin plot of the mean number of car trips per day, per city zone

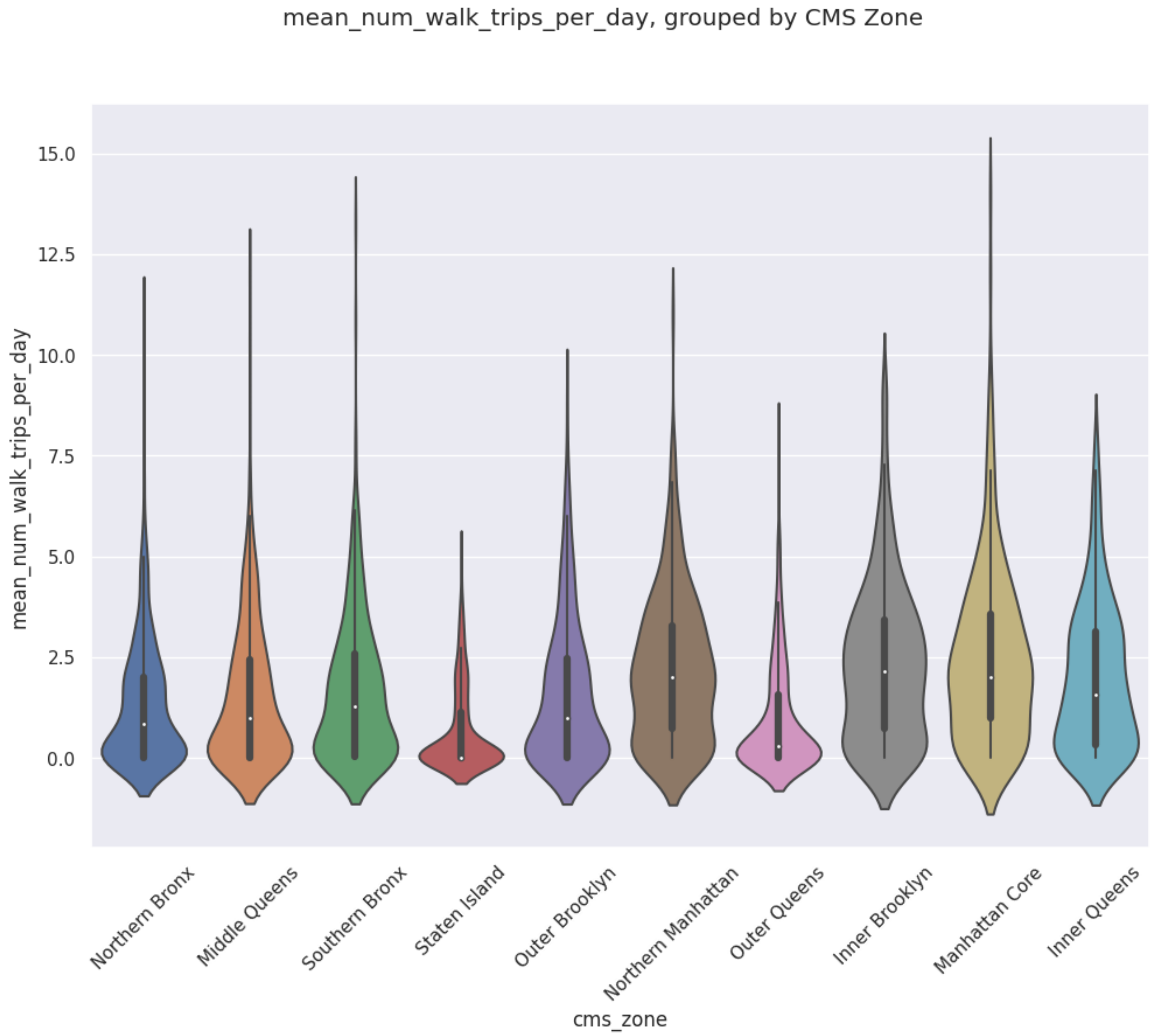


Fig 6. Violin plot of the mean number of walking trips per day, per city zone

Here we notice a difference. While the violin plots for mean car trips per day appeared the same across all city zones, it is not the case for walking trips. This suggests a correlation between city zones and the mean number of walking trips per day.

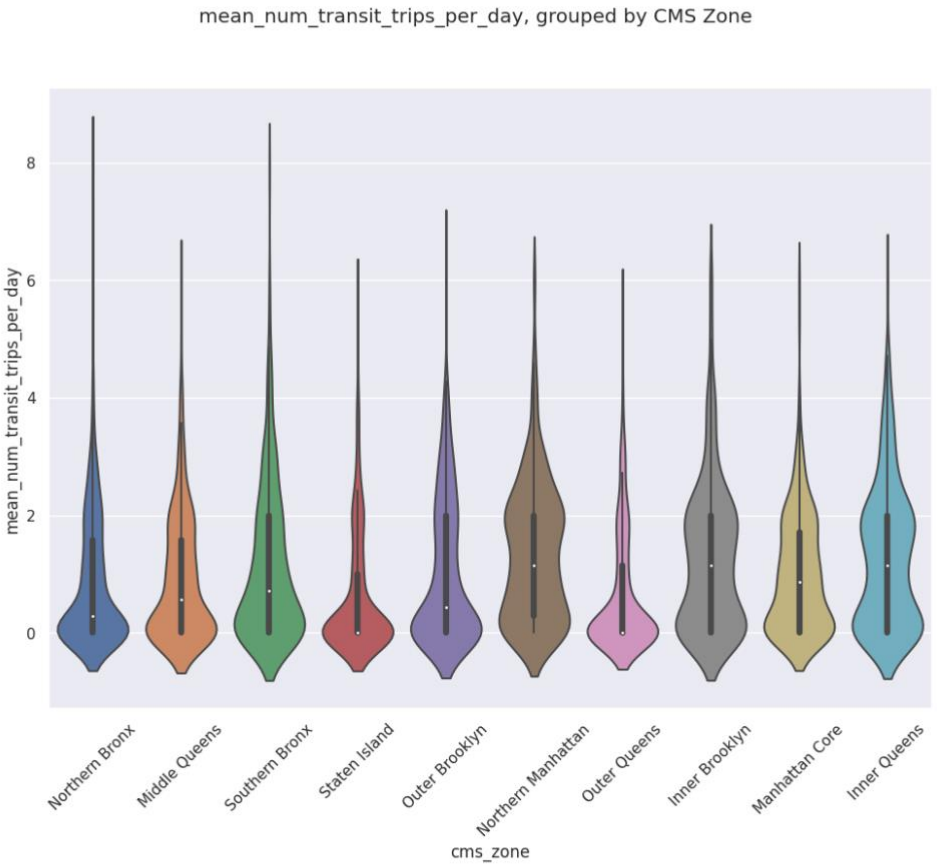
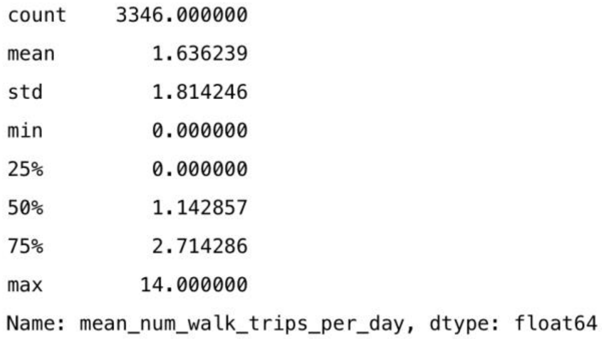


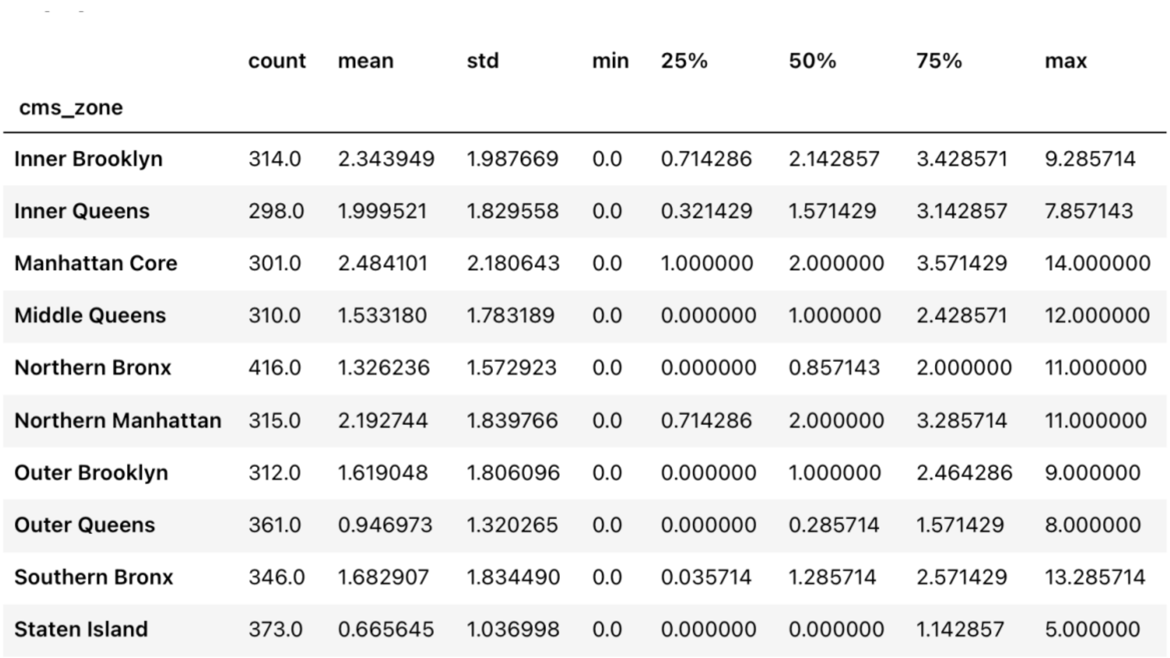
Fig 7. Violin plot of the mean number of public transit trips per day, per city zone

This idea is further reinforced by looking at the violin plot for public transit trips. Here we see what appears to be a pattern: Areas closer to the center of the city tend to have a higher number of public transit and walking trips than others. We verify this with numerical data for 'mean_num_walk_trips_per_day'.



```
count    3346.000000
mean        1.636239
std         1.814246
min         0.000000
25%         0.000000
50%         1.142857
75%         2.714286
max        14.000000
Name: mean_num_walk_trips_per_day, dtype: float64
```

Fig 8. Statistical information about 'mean_num_walk_trips_per_day'



cms_zone	count	mean	std	min	25%	50%	75%	max
Inner Brooklyn	314.0	2.343949	1.987669	0.0	0.714286	2.142857	3.428571	9.285714
Inner Queens	298.0	1.999521	1.829558	0.0	0.321429	1.571429	3.142857	7.857143
Manhattan Core	301.0	2.484101	2.180643	0.0	1.000000	2.000000	3.571429	14.000000
Middle Queens	310.0	1.533180	1.783189	0.0	0.000000	1.000000	2.428571	12.000000
Northern Bronx	416.0	1.326236	1.572923	0.0	0.000000	0.857143	2.000000	11.000000
Northern Manhattan	315.0	2.192744	1.839766	0.0	0.714286	2.000000	3.285714	11.000000
Outer Brooklyn	312.0	1.619048	1.806096	0.0	0.000000	1.000000	2.464286	9.000000
Outer Queens	361.0	0.946973	1.320265	0.0	0.000000	0.285714	1.571429	8.000000
Southern Bronx	346.0	1.682907	1.834490	0.0	0.035714	1.285714	2.571429	13.285714
Staten Island	373.0	0.665645	1.036998	0.0	0.000000	0.000000	1.142857	5.000000

Fig 9. Statistical information for 'mean_num_walk_trips_per_day', grouped by city zones

This further reinforces our idea and strongly suggests the existence of some sort of relationship between city zones and the mean number of daily trips using public transportation or by walking.