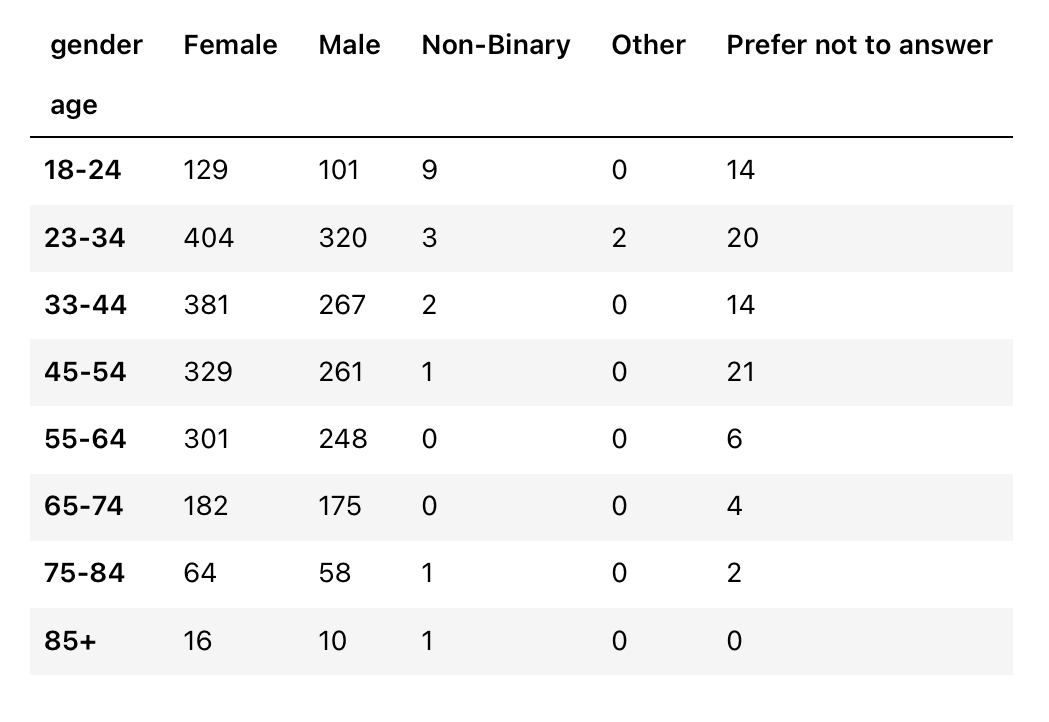
**Note:** This document forgoes the introduction and is intended as an extension of the Introductory document.

**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.

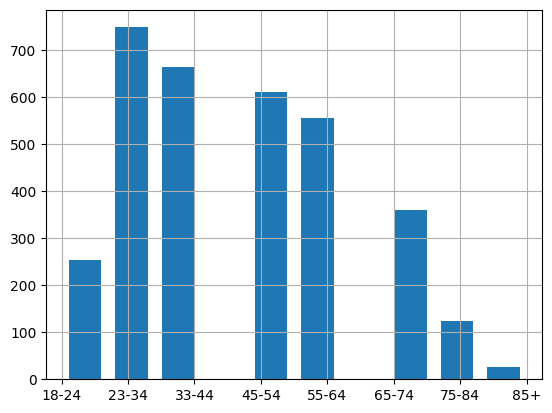
**The members of the dataset**

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.



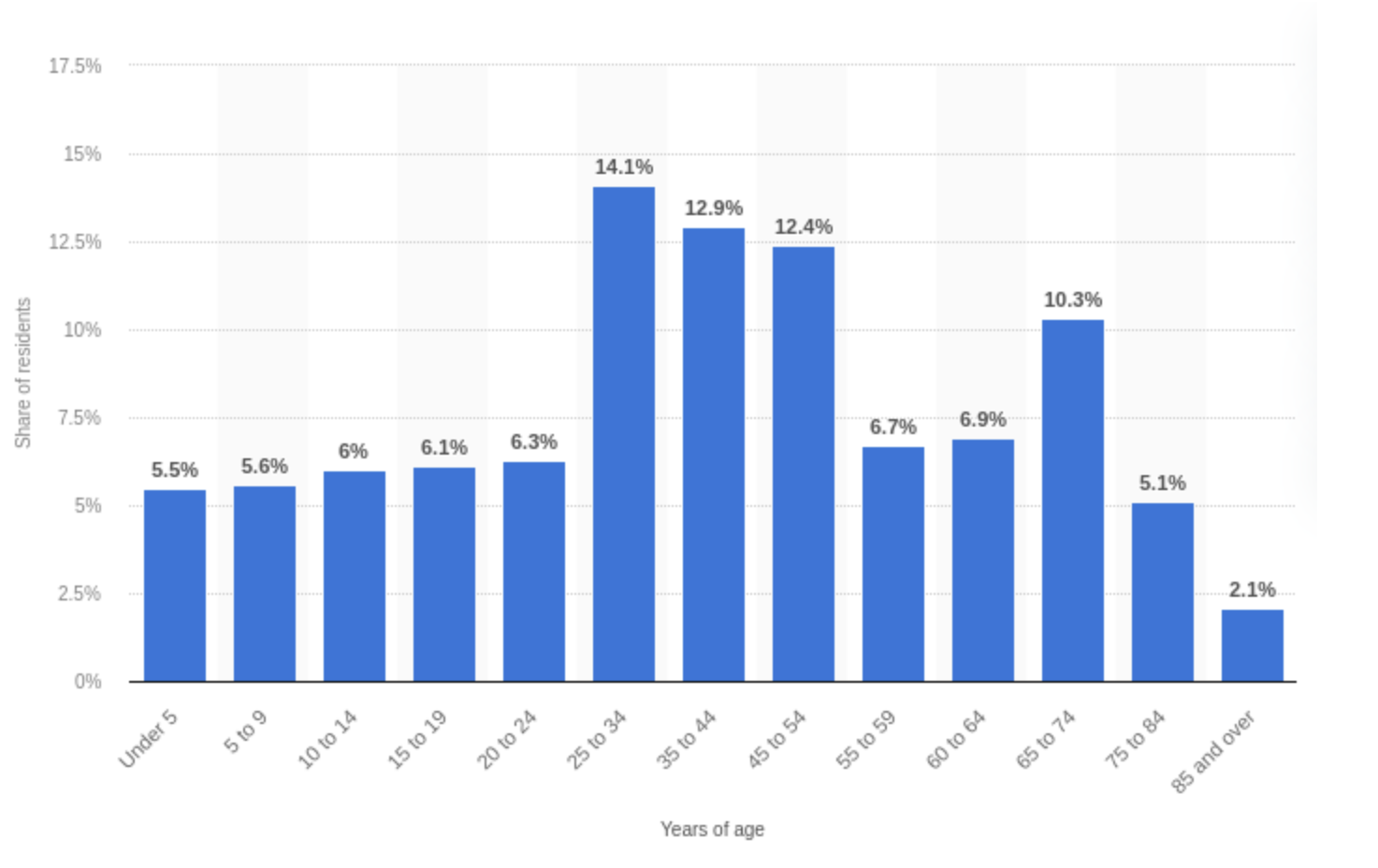
*Fig 1. Cross-Classification Table for the members of the dataset in terms of ‘Age’ and ‘Gender’*

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.



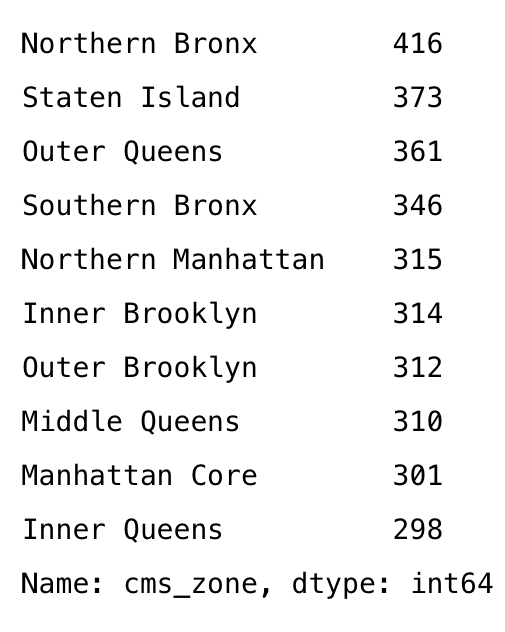
*Fig 2. Age range histogram*

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*](https://www.statista.com/statistics/911456/new-york-population-share-age-group/)

Next, we explore the representation of each city area:

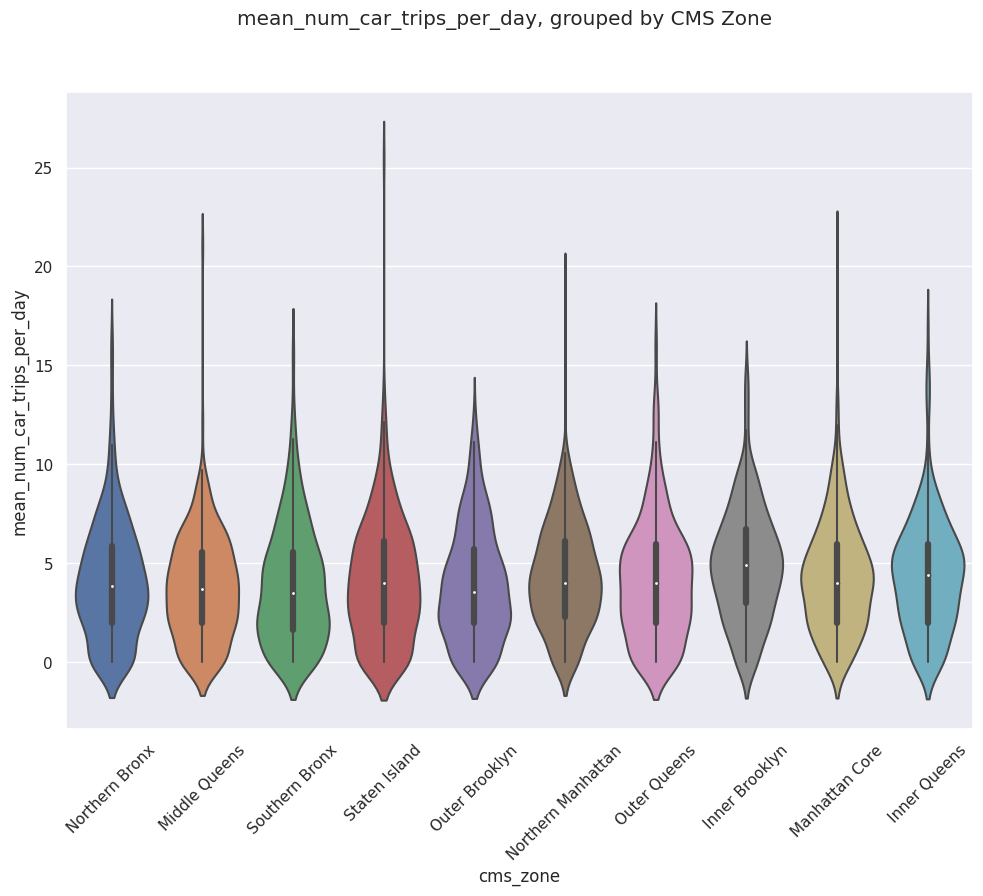
*Fig 4. Instance counts for each city zone*

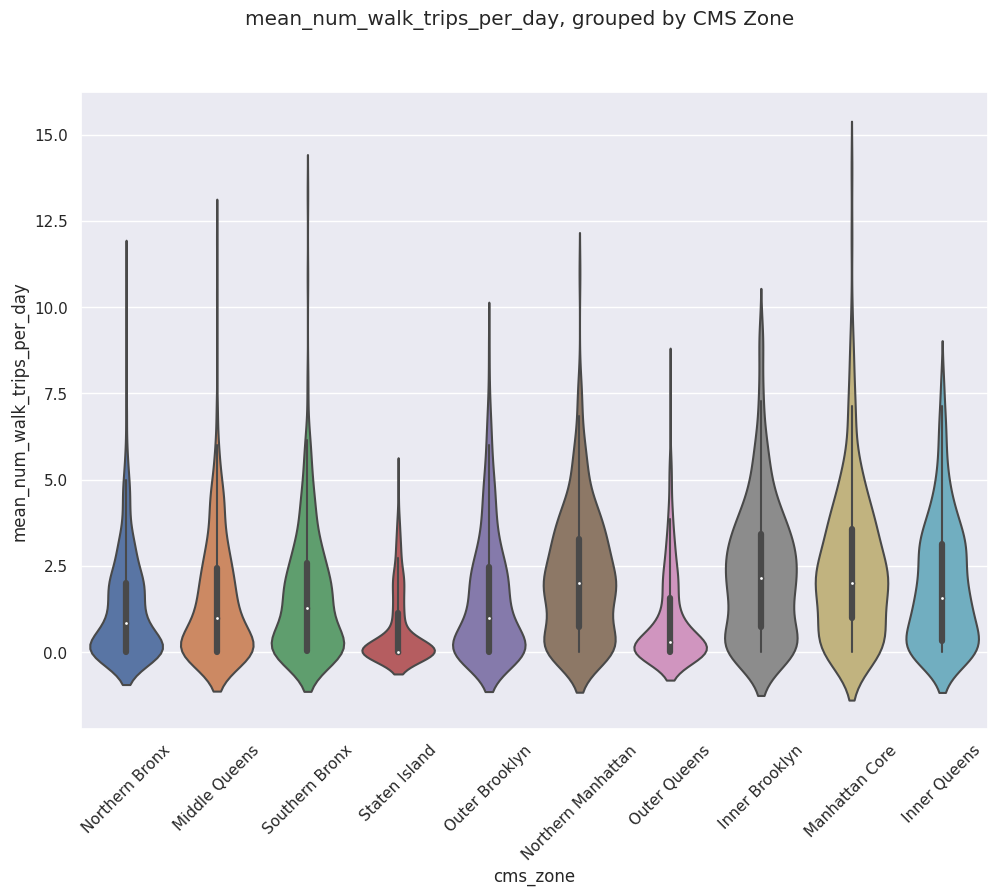
Like before, the majority of zones have more or less the same representation. Northern Bronx, Staten Island, Outer Queens and Southern Bronx stand out as having more representation than the rest outside a reasonable margin. We will have to keep this overrepresentation in mind.

**Stats per city zones**

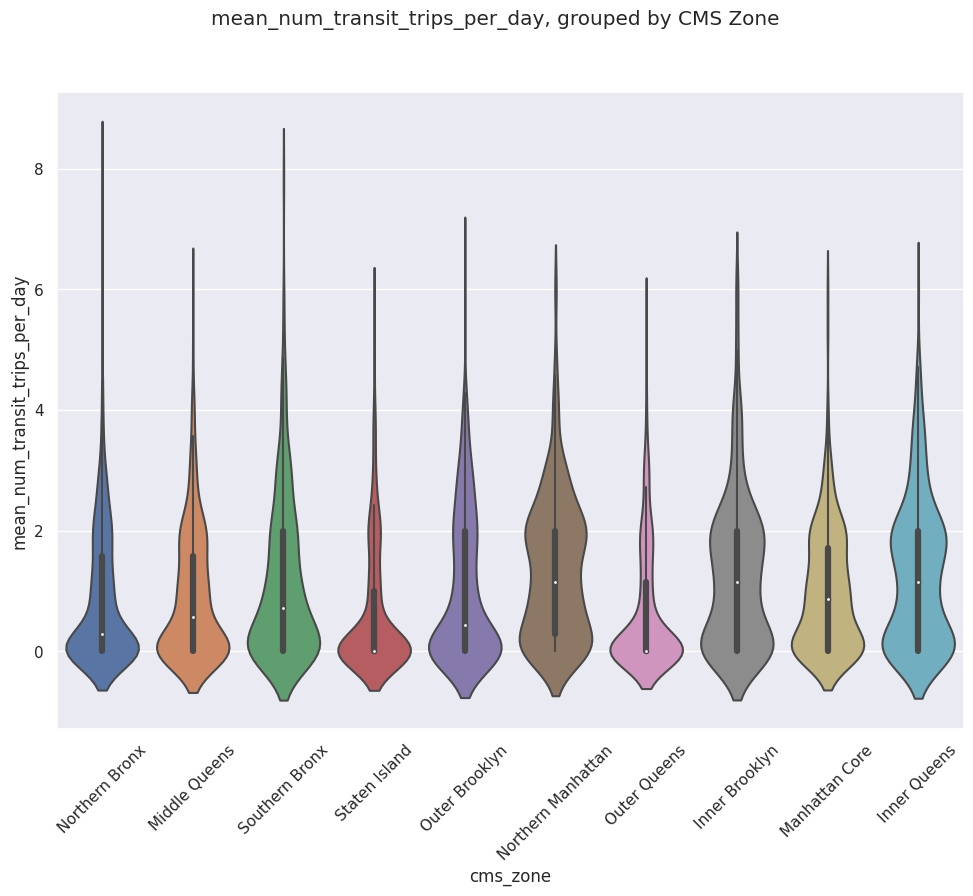
Next, having talked about the composition of our dataset, we will explore statistics regarding transport per city zones.

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.

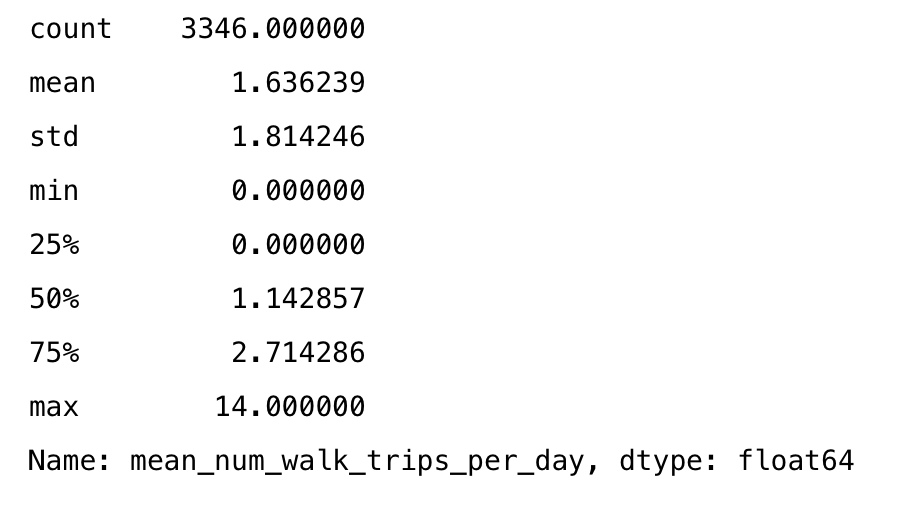
*Fig 5. Violin plot of the mean number of car trips per day, per city zone*

 *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’.



*Fig 8. Statistical information about ‘mean\_num\_walk\_trips\_per\_day’*

*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.