Boston Harbor Now

Waterfront Visitation and Equity Study: Project Feasibility Assessment Report

Author: Ford Fishman

Introduction

The Boston Harbor Now Waterfront Data Project aims to analyze Veraset mobility data from 2018 to 2022 to understand waterfront visitation more deeply. The conclusions from the analysis will be made public for interested parties to gain insight from the project. To serve as a point of comparison, BHN also acquired a data set from MassInc for August 2023 consisting of foot traffic, occupation, and survey information for five parks: Castle Island Park, Christopher Columbus Park, Martin's Park, Piers Park, and Pope John Paul II Park. This data set will allow us to understand how confident we can be in the Veraset data.

Tectonix Platform

Phase I of the data analysis consisted of data exploration and pre-processing. The Veraset data is hosted on a geospatial data platform called Tectonix. To access the Veraset data, I met with Tectonix team members for hands-on instruction on the platform. I created a dashboard consisting of polygons of the five parks corresponding to the MassInc data. Due to some abnormally high visitor and visit counts for Christopher Columbus Park for August 2022, a more natural analog for August 2023, I solely exported August 2021, as the next closest analog.

The records I exported from Tectonix included the number of visits and the number of visitors. Visits are a more direct comparison to the MassInc foot traffic data, while visitors are more comparable to the survey data, as it is unlikely survey participants were interviewed more than once. To compare to the survey data, I exported the following demographic visitor data for the five parks: gender, education, age, race, and income.

As I became more familiar with the Tectonix platform, I also provided feedback to the Tectonix team regarding the usability of various features of their platform. Various UX elements of Tectonix make accessing data and exporting it difficult. For instance, specific dates can only be picked by manually moving a slider imprecisely. Individual metrics have to be exported on their own to individual .csv files. I have spoken to the Tectonix team about their UX issues, and they plan to work on them, but the exact timeline for any fixes is unclear. Additionally, they may provide access to a Python-based API to the platform, which would allow me to build more precise and less time-consuming queries.

Data pre-processing

Both data sets were imported and pre-processed in Python using the pandas library. The Veraset data required less pre-processing, as I did not have direct access to individual records. Exported values for Asian and Pacific visitors were combined to match the MassInc race features. Additionally, all visitor counts broken down by demographic groups were normalized by the total visitor count for that park. This allowed me to compare proportions of demographic groups across parks, rather than raw visitor counts, which are unreliable, according to Tectonix team members.

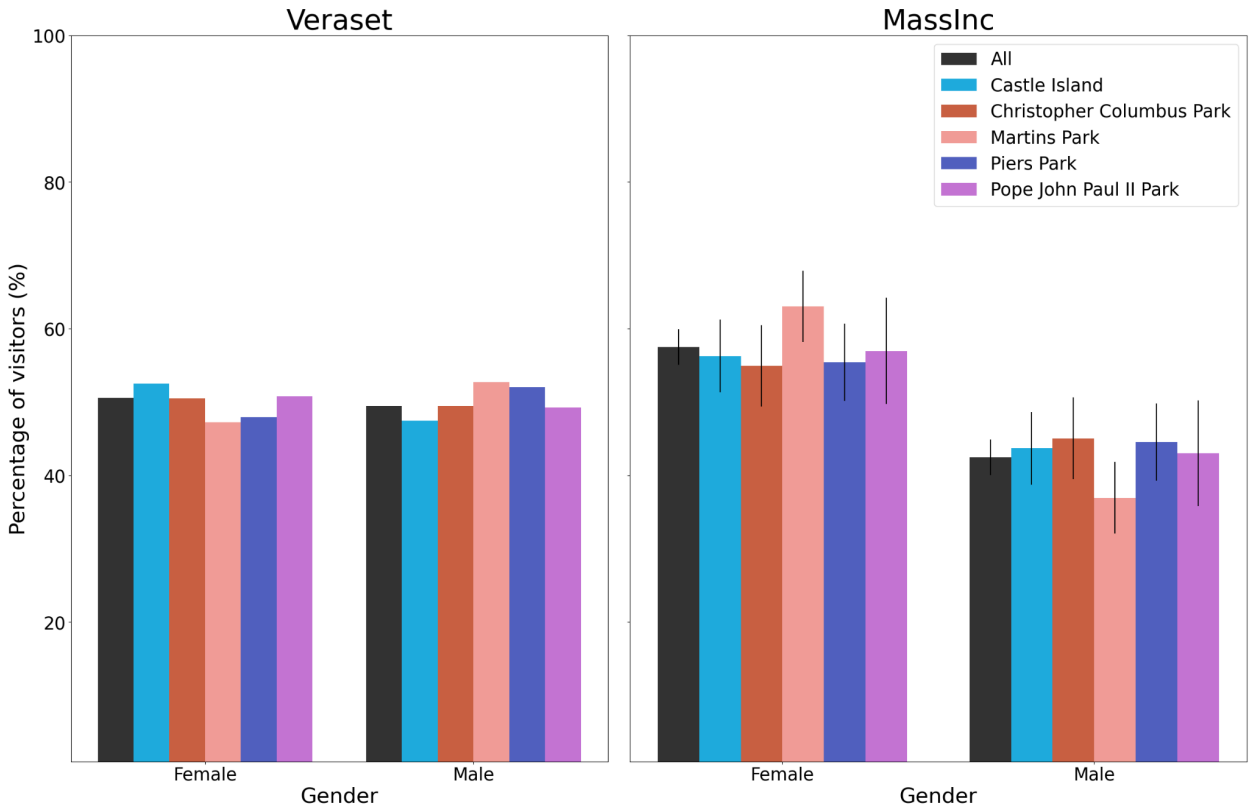
The MassInc data required more pre-processing to be comparable to the Veraset data. For instance, various survey questions allowed the participants to answer “Prefer not to say”, which has no analog in the Veraset data, so those responses were labeled as missing data and not taken into account in the visualizations below. Responses for gender of “Transgender” or “Non-binary” and for race of “Middle Eastern / Arab” were also removed from the analysis, as there is no Veraset analog for these responses. Certain responses were also binned together for comparison, namely the education responses of “High school graduate” and “11th grade or less” and the income responses of “Below 25,000” and “25 to 49 thousand”.

Data set comparison

To understand the Veraset data, we want to see how much it agrees with the MassInc data. We can make some key comparisons to see how strongly they agree on different points. Below, I have visualized these comparisons for various demographic groupings as bar plots. The MassInc survey data has error bars representing 95% confidence intervals. These bars represent the range of possible true values given the data. If two error bars overlap, we cannot say the true values are distinct. The Veraset data do not have error bars because we do not have access to individual records; all records have been pre-aggregated for us. Therefore, we have much less confidence in estimating values for the Veraset data, and we can only comment on what we observe, rather than making strong conclusions about the underlying populations behind the sample.

For several of the demographic comparisons, the Veraset and MassInc data are not binned analogously. For instance, the Veraset data include visitors under the age of 18, while the MassInc data do not. Some demographics align more closely than others, and the more the bins diverge, the less confident we can be in making conclusions from our comparisons. It is also important to note that these are values for two different years: 2023 for MassInc and 2021 for Veraset, so minor deviations between the two are not unexpected.

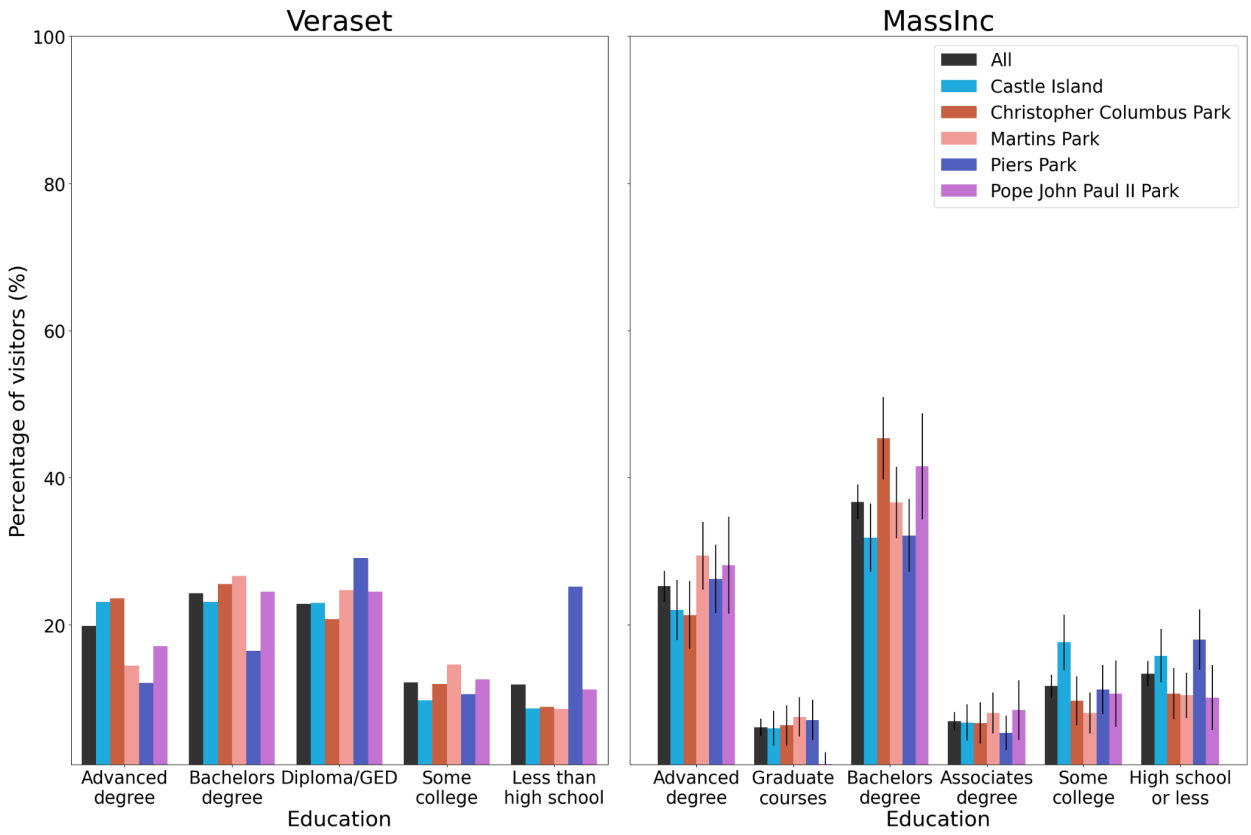
Gender



The ratio of male to female visitors was fairly similar across the parks within each data set. However, there was a relatively large difference in the ratio between data sets. With Veraset, the ratio was approximately 50:50, but with MassInc, the ratio of male to female visitors was approximately 40:60. There could be several reasons for this, including those providing the survey being more likely to approach female visitors than male visitors or male visitors being more likely to decline to take the survey.

Regardless of the reason, this can present a problem for future analysis. For example, if there is any relationship between gender and income, the imbalance in the gender ratio in the MassInc data could lead to incorrect conclusions. One approach could be to oversample male visitors in future analyses. This process would include randomly sampling the male visitors as duplicate observations in our data to make sure that they are not underrepresented in our analysis, thus evening the ratio.

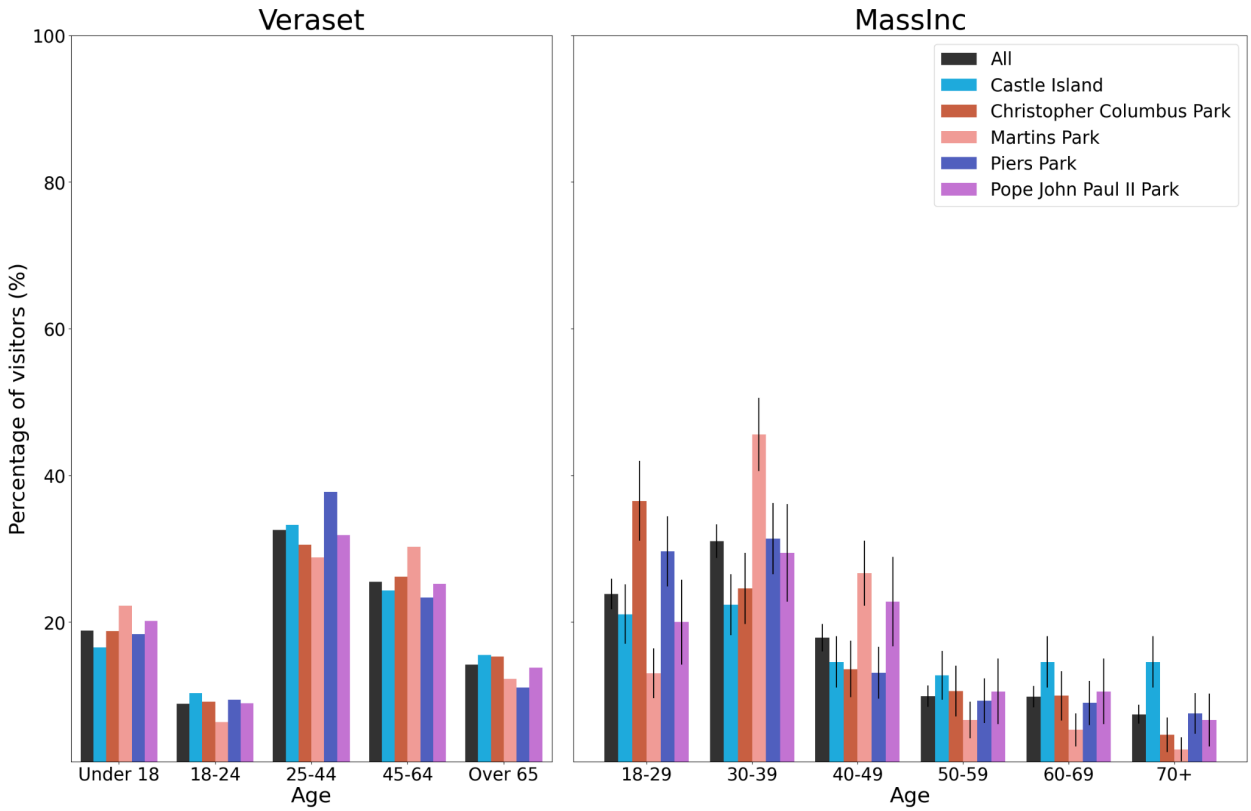
Education



The categories of education in the two data sets were not exactly the same. They both shared “Advanced Degree”, “Bachelors Degree”, and “Some College”. However, MassInc had “Graduate courses”, which had no Veraset analog. There was likely some overlap between Veraset’s “Diploma/GED” and MassInc’s “Associates degree”, but they were not quite the same. Finally, it is unclear which of Veraset’s categories would match visitors with a high school diploma but no college experience.

However, we can still make some general conclusions about education. The most noticeable observation was that visitors with Bachelor’s degrees in all parks made up a larger percentage of visitors in the MassInc data as opposed to the Veraset data. There was also a slight uptick in the percentage of advanced degrees in the MassInc data. Additionally, there were some differences among parks that occurred in one data set and not the other. For example, there were higher percentages of Bachelor’s degrees in Christopher Columbus Park and Pope John Paul II Park in MassInc compared to other parks, but not in the Veraset data. However, both showed a large proportion of visitors with less than a high school education at Piers Park, though this is less prominent in the MassInc data. These differences may be due to biases in self-reporting education levels in the survey data. That being said, there was only minimal agreement on education between the datasets.

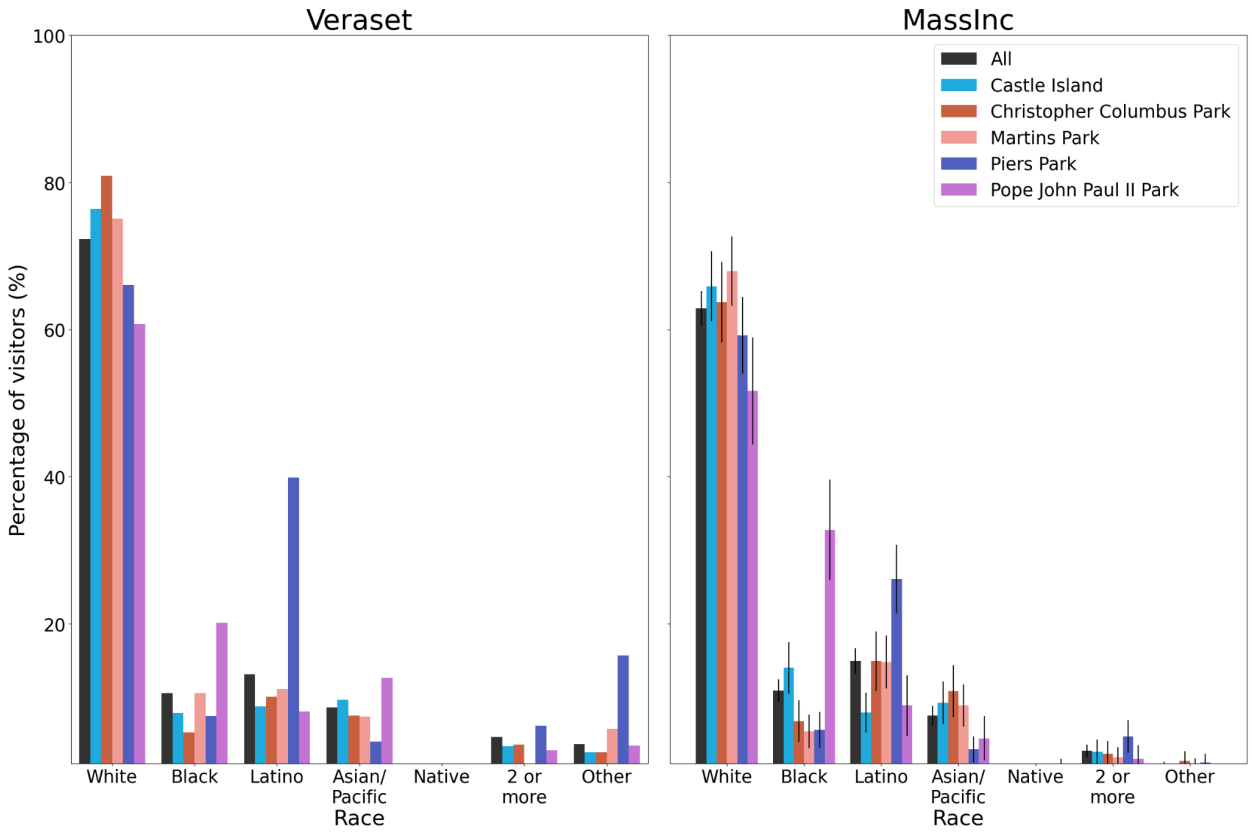
Age



The manner in which age was binned in the two data sets is quite divergent. As mentioned above, Veraset included individuals under 18, while MassInc did not. For all other ages, the bins did not align, as well. In general, both data sets show that the oldest adults made up a smaller proportion of the visitors than younger adults. Both also showed that the youngest bin of adults make up a smaller proportion of visitors than slightly older adults. However, this also may be due to how Veraset's group of youngest adults only consisted of 6 years of age (18-24), compared to MassInc's 11 years of age (18-29).

The most obvious difference between the data sets is that Veraset showed that with a given age bin, there is minimal difference among the parks, while MassInc showed that certain parks are more commonly visited by different age groups. For instance, MassInc showed that the oldest adults visited Castle Island Park more than other parks and that Martin's Park was heavily visited by visitors aged 30-50. Christopher Columbus Park had a large amount of visitors aged 18-29, as well. Veraset shows none of these patterns. This difference could be largely due to the different binning ranges of ages, or it could mean that Veraset's data may not accurately represent the age of park visitors.

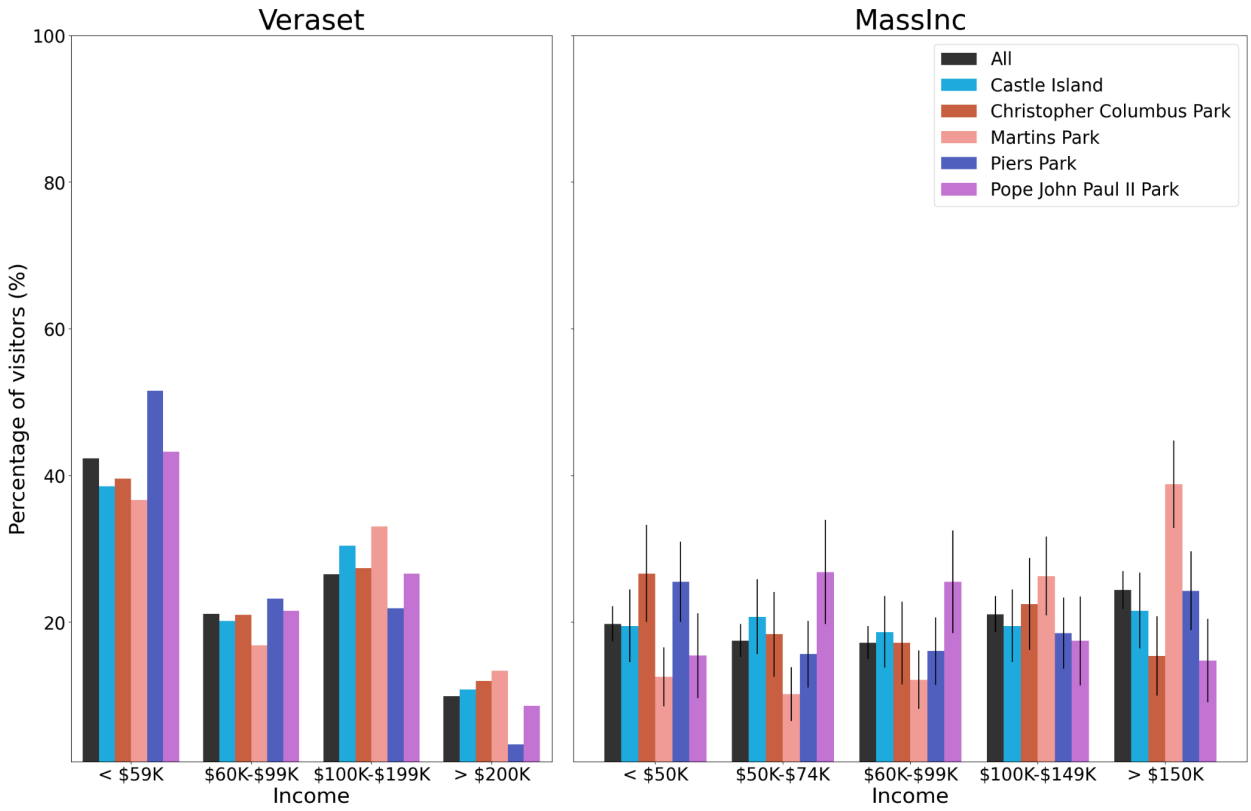
Race



Of all the demographic categories, the two data sets most strongly aligned on race. All race categories were the same between the two. For both, white visitors made up the majority of all visitors, with about 70% of all visitors being white in the Veraset data and about 60% in the MassInc data. Notably, both data sets also showed spikes of black visitors at Pope John Paul II Park and of Latino visitors in Piers Park. While the size of these spikes was not of the same magnitude in both data sets, their presence in both suggests strong agreement between the two them.

The Veraset data did include a spike in Asian/Pacific visitors at Pope John Paul II Park that was not present in the MassInc data, and there were considerably more visitors labeled as “Other” in the Veraset data, especially at Piers Park. There could be some biases in who was surveyed due to race, and there could also be differences in how “Other” was compiled for Veraset. Overall, there was a high degree of alignment between the two data sets with regard to race.

Income



Similar to the age categories, income was not binned in an easily comparable manner. However, we can make some general conclusions by seeing how the percentage of visitors changed with income bins. Here, we can notice some large differences between the data sets. Income was relatively evenly distributed among the income bins for MassInc. For Veraset, the bins of < $59K and $100K-$199K made a much larger proportion of visitors than the other two bins. This may be due to the range of the income bins themselves.

Additionally, there were many differences among the parks in the MassInc data that were not present in the Veraset data. For instance, Pope John Paul II Park had large proportions of middle-income visitors with MassInc, which was not the case with Veraset. MassInc also showed that the largest proportion of visitors came from the highest income bin, which was not the case with Veraset. Veraset also showed that Piers Park had few visitors from the highest income bin, which was not the case with MassInc.

Next steps

This analysis showed that while the MassInc and Veraset data sets do not perfectly align, they can be compared. Here, I focused on a qualitative comparison, but in the future, I will conduct an Inter-rater Reliability Analysis (IRR) to quantify the degree of agreement between the two. This will consist of creating hypotheses and questions to ask of both data sets. For instance,

one potential hypothesis could be “Castle Island Park has the lowest proportion of visitors with Bachelor’s degrees”. For each data set, I will assess if this is true or false. If both parks have the same answer, then they agree; otherwise, they disagree. I will do this for a number of hypotheses. Once all are tested, I can provide a value for percent agreement between the two data sets.

Once this step has been completed, I will move to a series of queries BHN has provided about the larger Boston waterfront. I will use Veraset data from 2019 to 2022 to address these queries, with data from across the waterfront. For each query, I will conduct and analysis and provide conclusions. I will provide a reliability grade for each query to convey my confidence in those conclusions.