Geocoding Truck Stops Update

William Co

2025-08-03

This report documents the geocoding validation process, highlighting discrepancies arising from integrating data sources like Yelp and Yellow Pages. Despite matching entries by phone number, these platforms introduce a high rate of false positives. We apply a variety of filtering to correct for this. We then provide a detailed account of the types of errors observed and offer a numerical estimate of the remaining discrepancies. These errors represent a small proportion of the entire dataset, affecting approximately 17 locations.

Table of contents

1 Introduction

This section details the iterative process of refining our geocoding validation. An initial update addressed a bug in latitude and longitude extraction from the "RVers and Travelers" website. Subsequent analysis focused on resolving discrepancies between different data sources.

Initially, we attempted to match locations based on properties such as primary and secondary phone numbers. However, this approach proved complex and prone to error. A more effective strategy was developed, which involves matching locations based on the minimum distance between them. The closest geographic match is considered the correct one. This method significantly reduced the number of discrepancies.

For instance, postal code input errors were identified as a source of discrepancies. The initial analysis, including outliers, showed a number of locations with a discrepancy of less than one mile (see Figure 1).

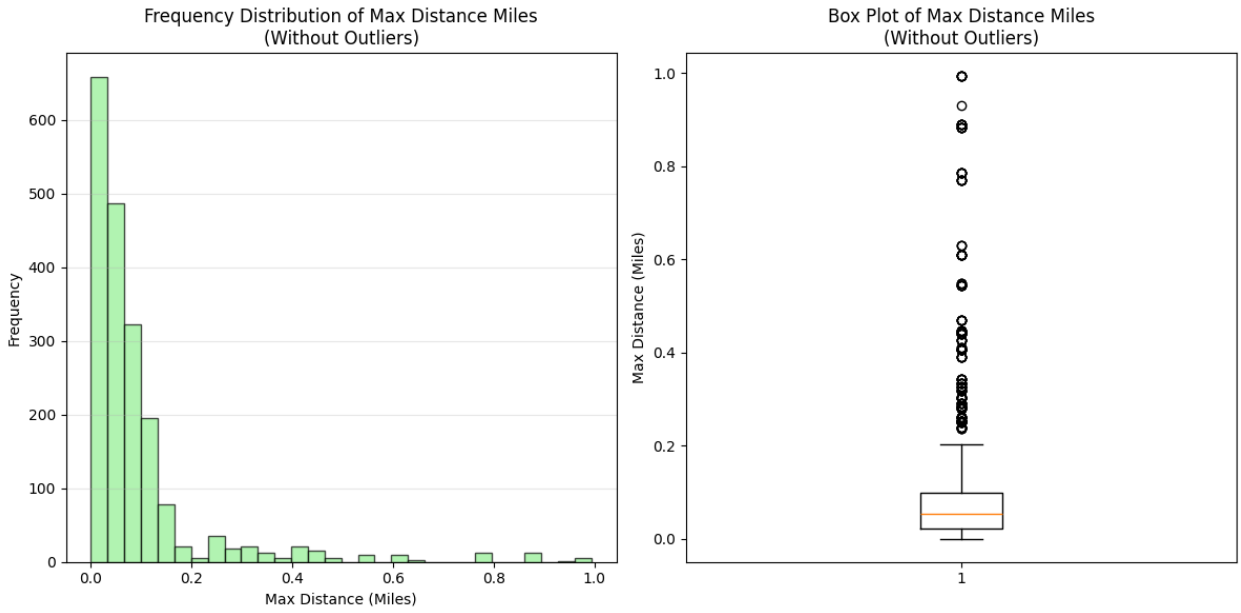

Figure 1: Analysis of discrepancies less than one mile.

Including all outliers, the dataset initially appeared as shown in Figure 2.

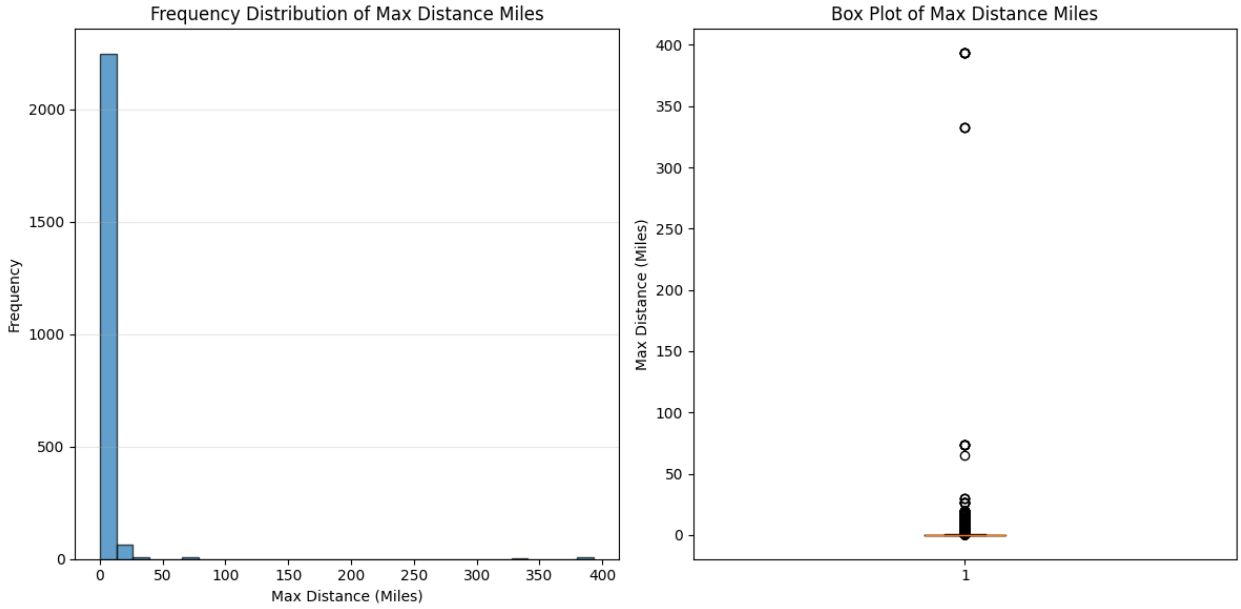

Figure 2: Dataset including all outliers.

After implementing the minimum distance comparison method, we were able to reduce the number of locations with a discrepancy greater than one mile from seven to five. The results of this improved approach are visualized in Figure 3.

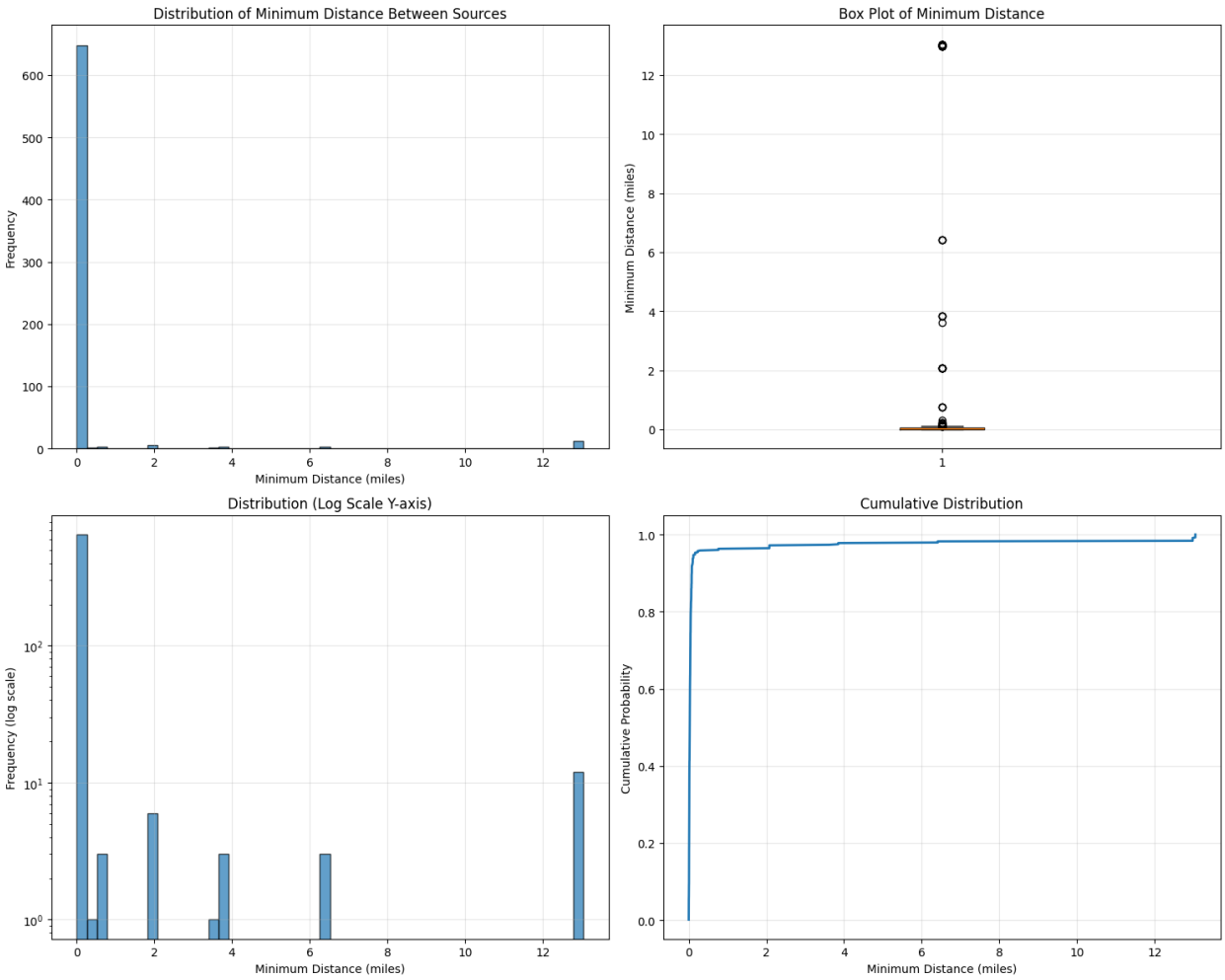

Figure 3: Results of the minimum distance comparison approach.

This refined approach has left us with only five locations having a discrepancy of more than one mile, a significant improvement in data accuracy.

2 Manual Verification of Remaining Discrepancies

The five remaining locations with significant discrepancies were manually verified. The analysis revealed no single, discernible pattern to the errors. Instead, they fall into two main categories:

1. **Geocoding Inaccuracy**: In some cases, the geocoding service returned coordinates with a spatial offset from the actual address. For a given address intended to map to location X, the service returned coordinates at $X + E$, where E represents the spatial error.
2. **Non-Physical Addresses**: Another source of error was the presence of Post Office (P.O.) boxes instead of physical street addresses in the source data. P.O. boxes do not represent a physical location and thus cannot be accurately geocoded for our purposes (see Figure 4).

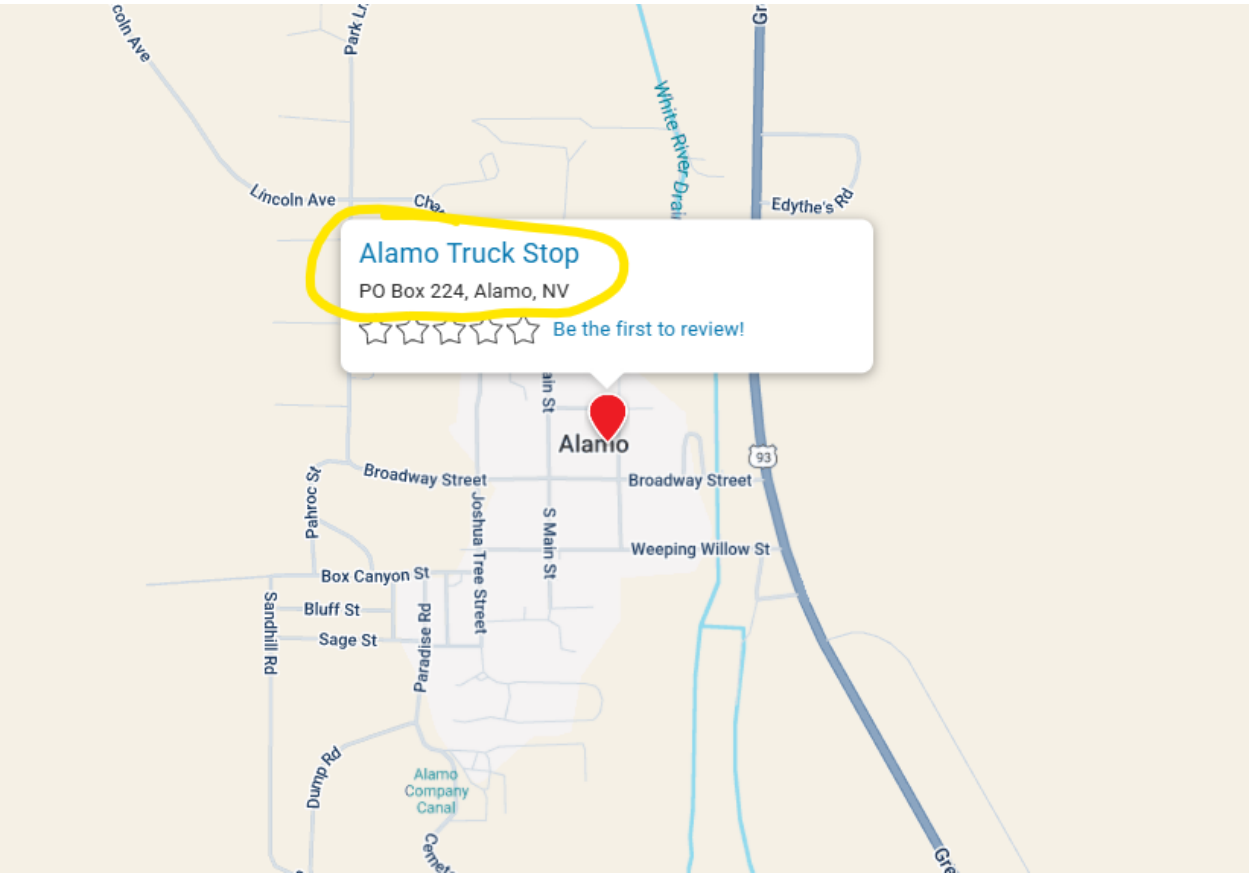

Figure 4: An example of a P.O. Box listed as an address.

explaining the "false" place name matches. The place name matches introduced a small bit of error because We made sure to match down to the road level. Which is fine but not as accurate as if we match down to the label level.

ive matched everything down to 200 meters.

next step is now to finalize the lat and long coordinates. I will average out the 300 meter discrepancies to the midpoint,