Quarterly Report for DOE on Mode Choice Estimation

## Overview

Problem: Using multinomial logit to estimate mode choice was problematic, as pointed out by reviewers. The model and estimated results varied widely across different model specifications, were very dependent on small tweaks to the model (i.e., it was not very robust to small changes). In addition, for a chosen model specification, the estimated results did not support heterogeneity across populations; it did not vary across user class or other characteristics that we would expect to change the results, and instead were only correlated with changes in cost / price. [ANNA YOU MIGHT NEED TO CHANGE THE ABOVE PARAGRAPH]

Possible solution: rather than using multinomial logit as the mechanism to estimate mode choice, we want to use fractional split logit. We will try several different specifications to test out the performance of these specifications relative to each other, and also relative to multinomial. Reviewers in the last round suggested this, and it may help solve some of these problems with the mode choice estimation.

## Goals:

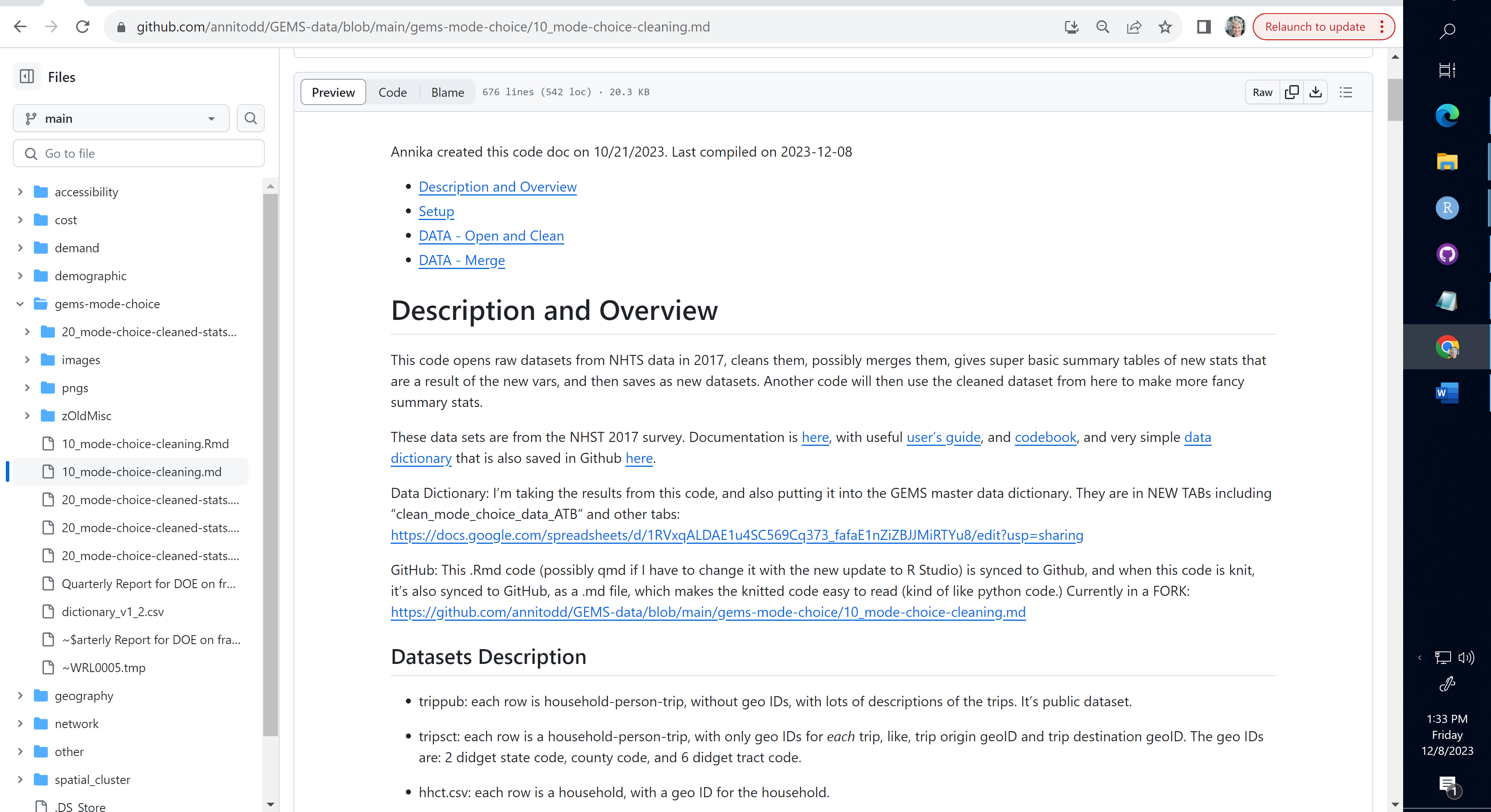
* Prepare a cleaned dataset with all of the variables needed for the estimation. This should be in a format that is clean, and is able to be easily manipulated to make it into various datasets that can be used for a variety of fractional logit specifications.
* The dataset should be rigorously cleaned, and should be well documented. The code and methods should be easily readable for those who do not know the programming language (i.e., the code should be commented in a way that it is obvious what is going on even to a non-coder).
* Examine the dataset to see what types of fractional logit specifications could be supported. Specifically, we need enough variation within each observation that is being defined in the fractional split to be able to estimate heterogeneous effects.
* Specify a variety of fractional split logits, and run these specifications.
* Compare the different fractional split logit specifications to each other.
* Re-run the model as a multinomial split logit (there may be differences because the data may have been cleaned in a more rigorous way than previously).
* Compare the fractional split logits to the multinomial.

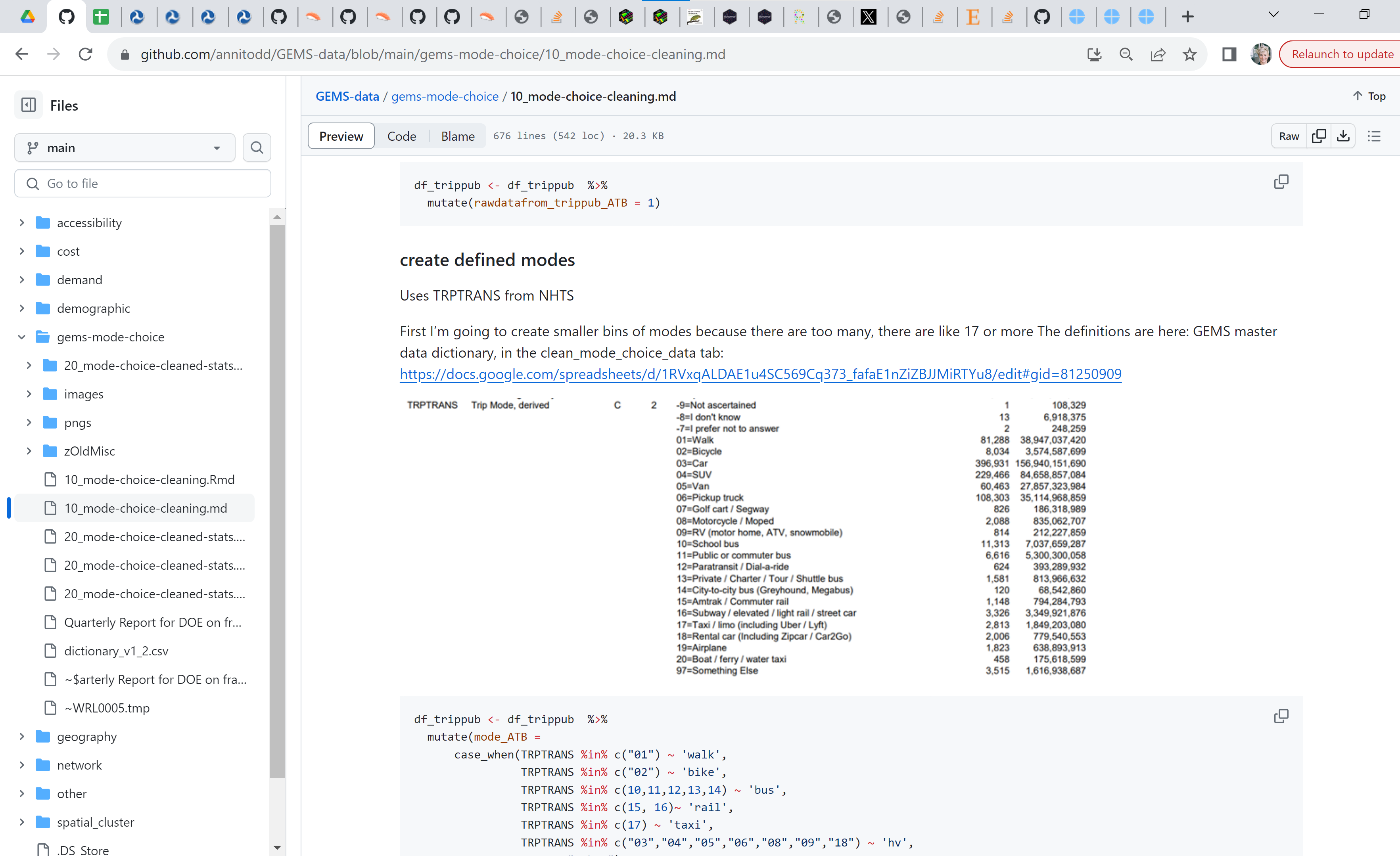
## Difficulties and barriers, and progress, reaching these goals:

* Barrier: It is unclear that the dataset used in the first round is rigorously cleaned; it is not well documented, and there are a few obvious errors that we want to correct. This means starting to create a new dataset from scratch.
  + Progress: we have a newly prepared dataset derived from the original raw data. We do not have all of the variables that we need prepared, but we have most of them prepared in a rigorous way.
* Barrier: rigorous data-cleaning means checking and re-checking and doing many gut checks on every data import, merging, manipulating, etc.
  + Progress: many checks have been run, that highlight and show, convincingly, that the data is being prepared in a rigorous way, without errors.
* Barrier: Thoroughly documenting code is best practices, but takes much longer!
  + Progress: we are using a markdown formatted R code that allows easily readable paragraphs of prose describing the purpose of the code, comments describing what each small code line / chunk is doing, and documenting decisions made with respect to cleaning. Links to external documentation are embedded in the code. See below for an example.
* Barrier: examining the dataset to ensure that enough observations are available to provide heterogeneity across user classes requires a thorough process (again with many checks).
  + Progress: we are examining the levels of heterogeneity that are likely supported. I.e., we want to make sure that there are enough observations within each user class we intend to look at such that the assumptions for the fractional split logit model are met.
* Barrier: a fractional split logit model requires a modified dataset format that is not the current format, and that takes a bit longer to make sure it’s specified in the correct way, with even more checks (many multiples of checks). Each different specification of the fractional split requires a dataset that is manipulated and created in a different way; this requires time and documentation.
  + Progress: one proof-of concept dataset has been constructed.
* Barrier: literature on the theory and practical application of fractional split logit is sparse in the academic literature. We have a handful of examples, but these examples do not thouroughly document the theory and what the constraints of the model are, why they chose their specification and the underpinning theory as to why. We are referring back to the seminal theory paper from the 90s to do much of our theoretical underpinnings.
  + Progress: we are talking to experts in the field of econometrics and statistics to make sure we understand the constraints under which fractional split logit operates, and the underlying theory and assumptions needed to produce believable results.

## Examples of the progress:

### Example of rigorous, readable, documentation of the code:



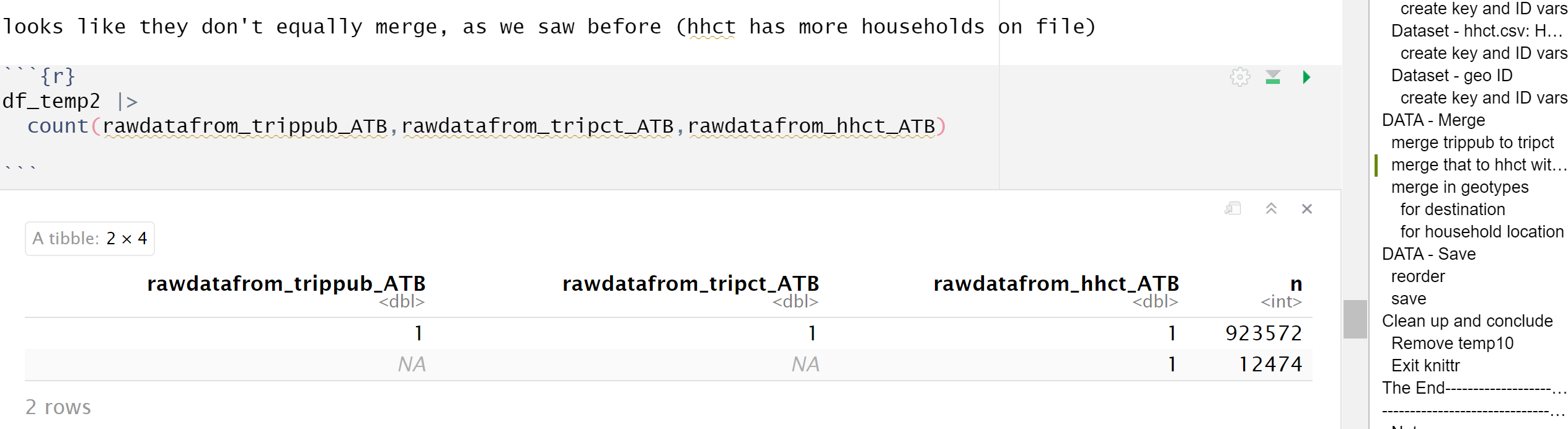


### Example of checks

#### Merging

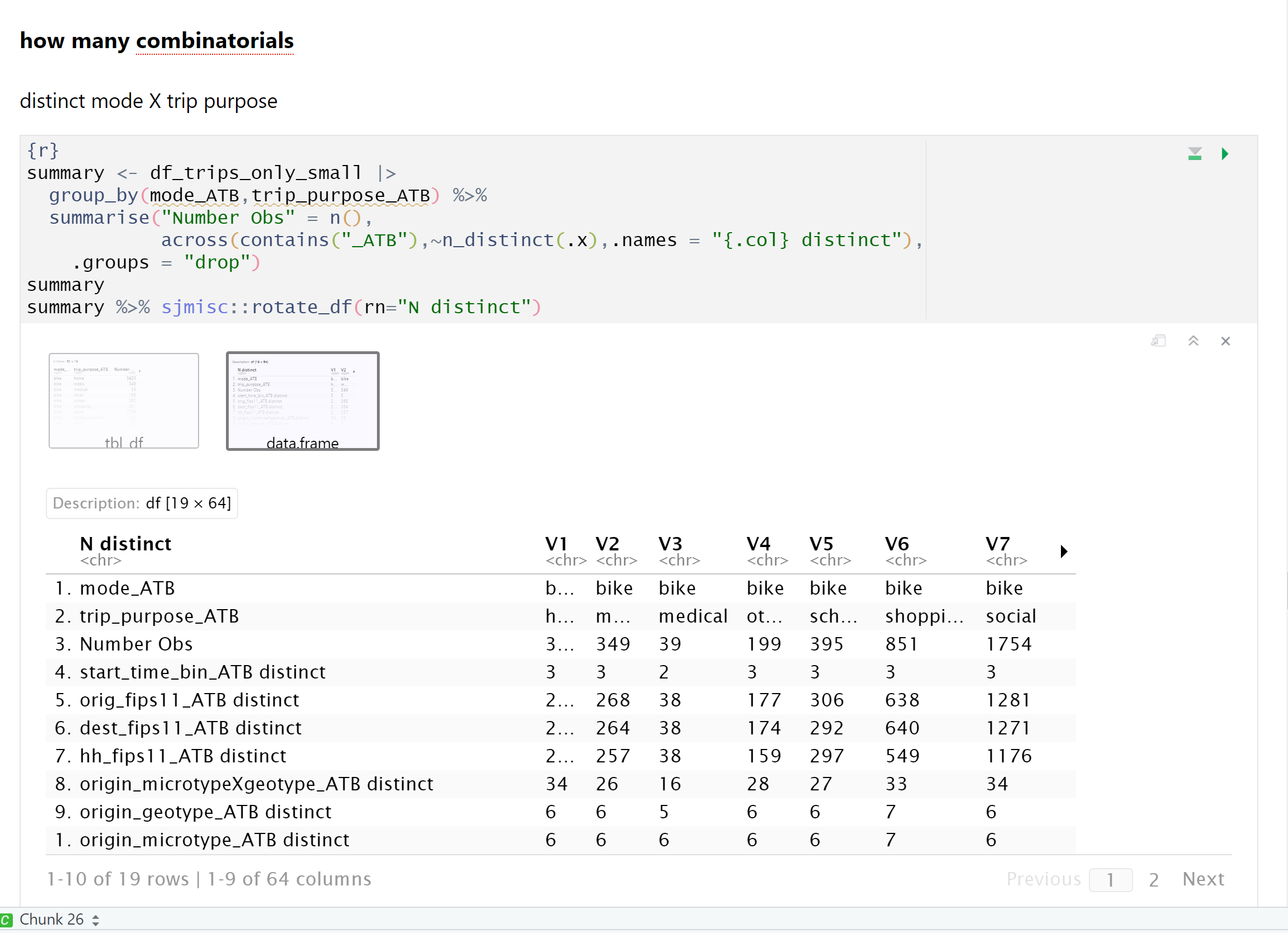


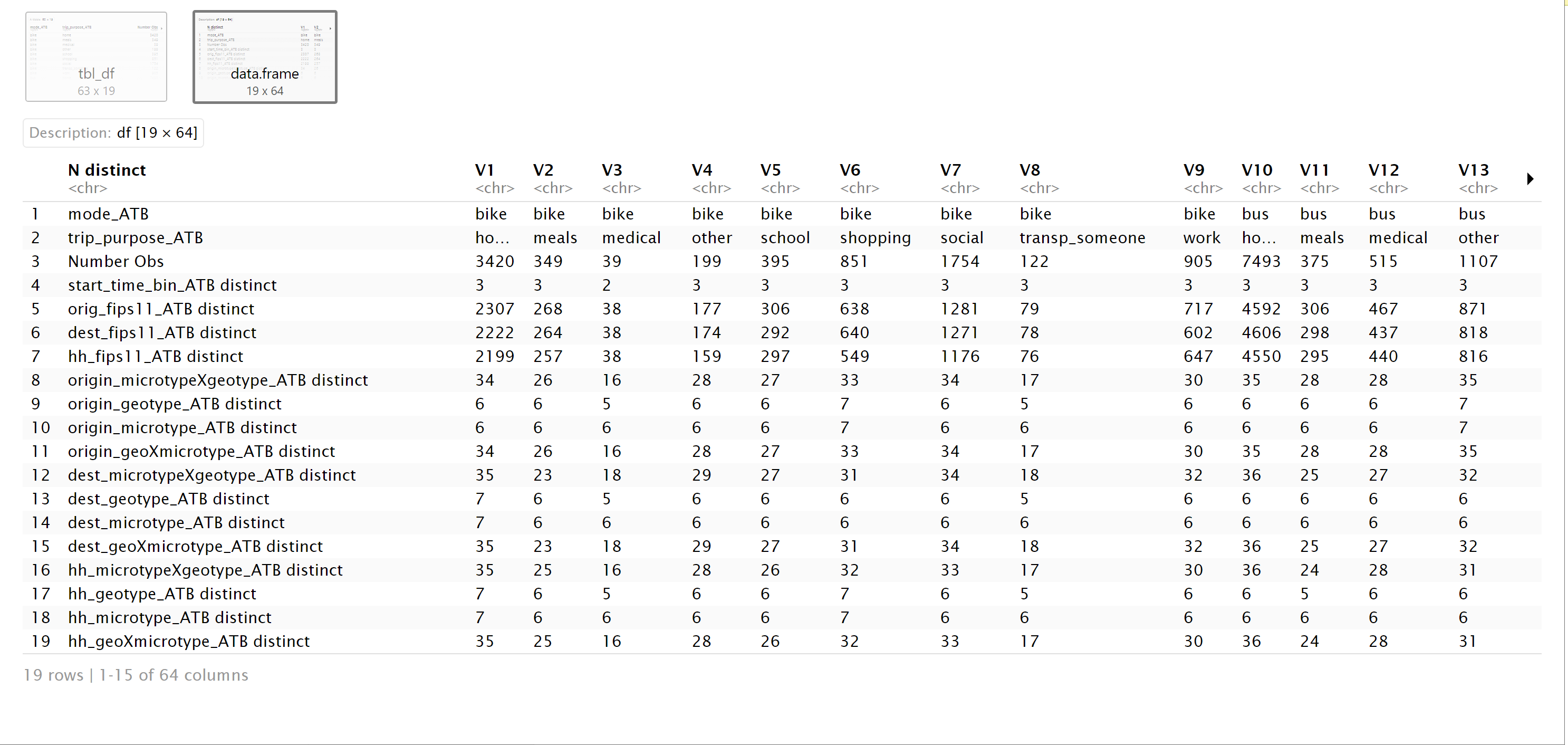
Merging



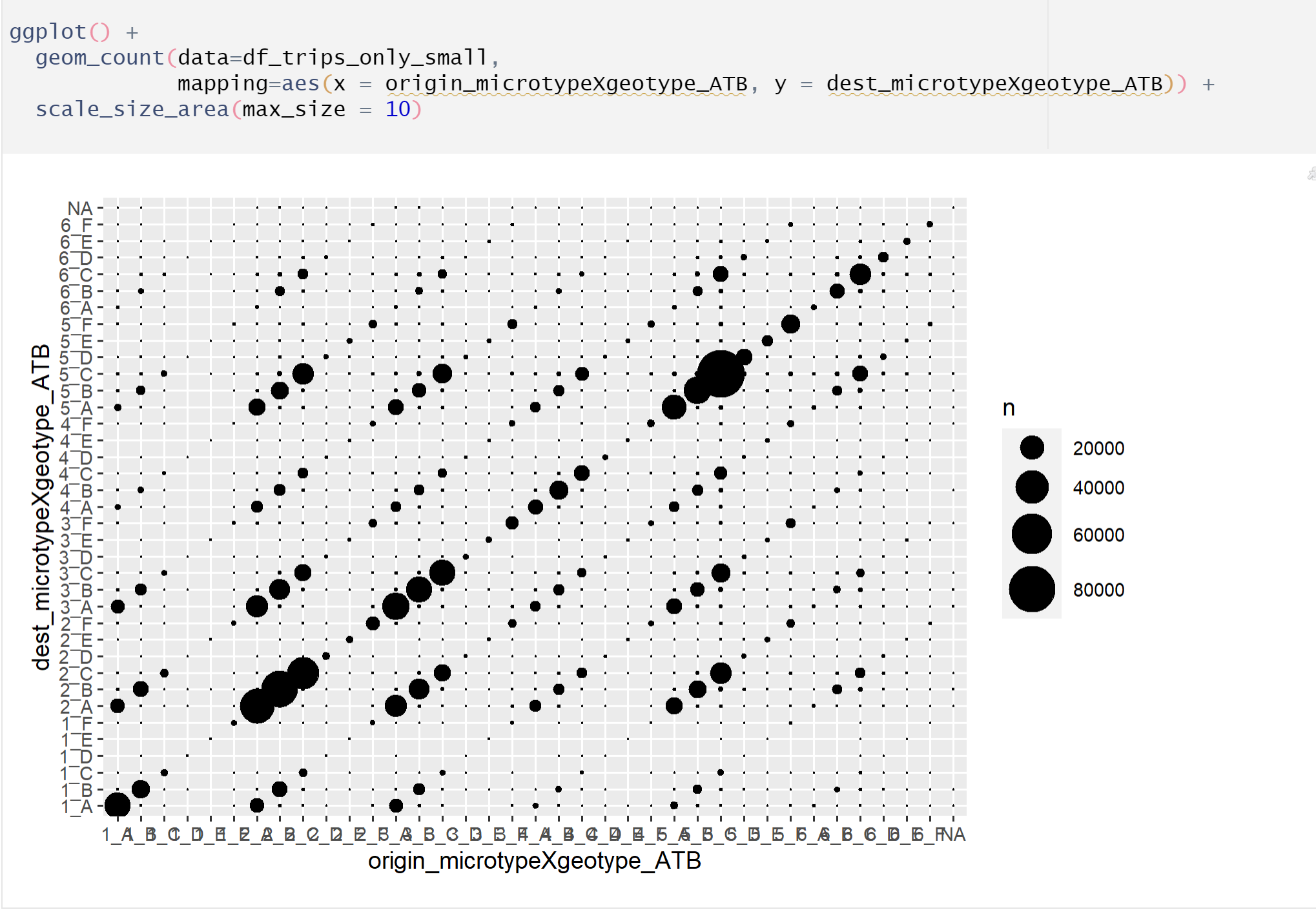
### Examining what the data will supportL

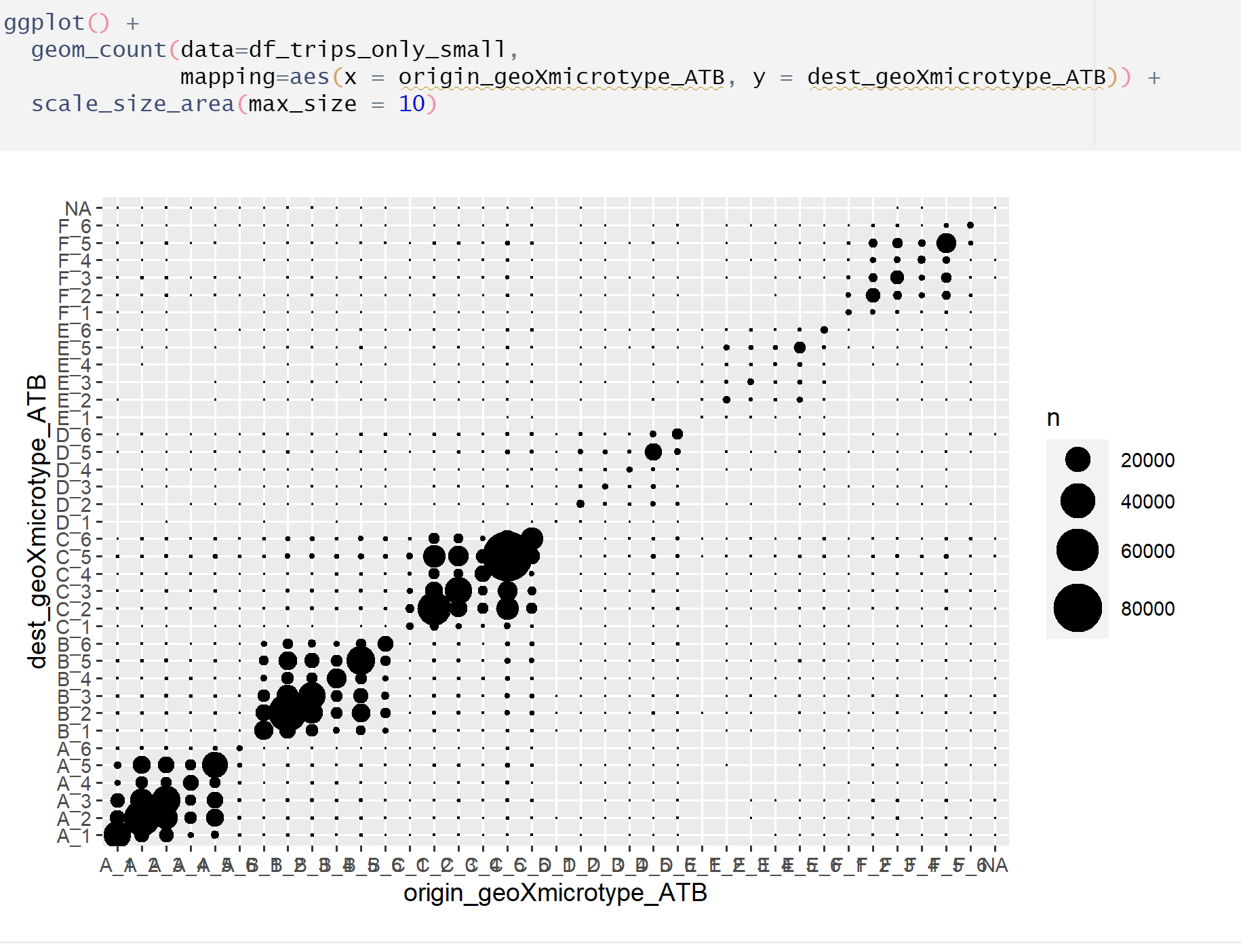
#### What level of heterogeneity is supported:



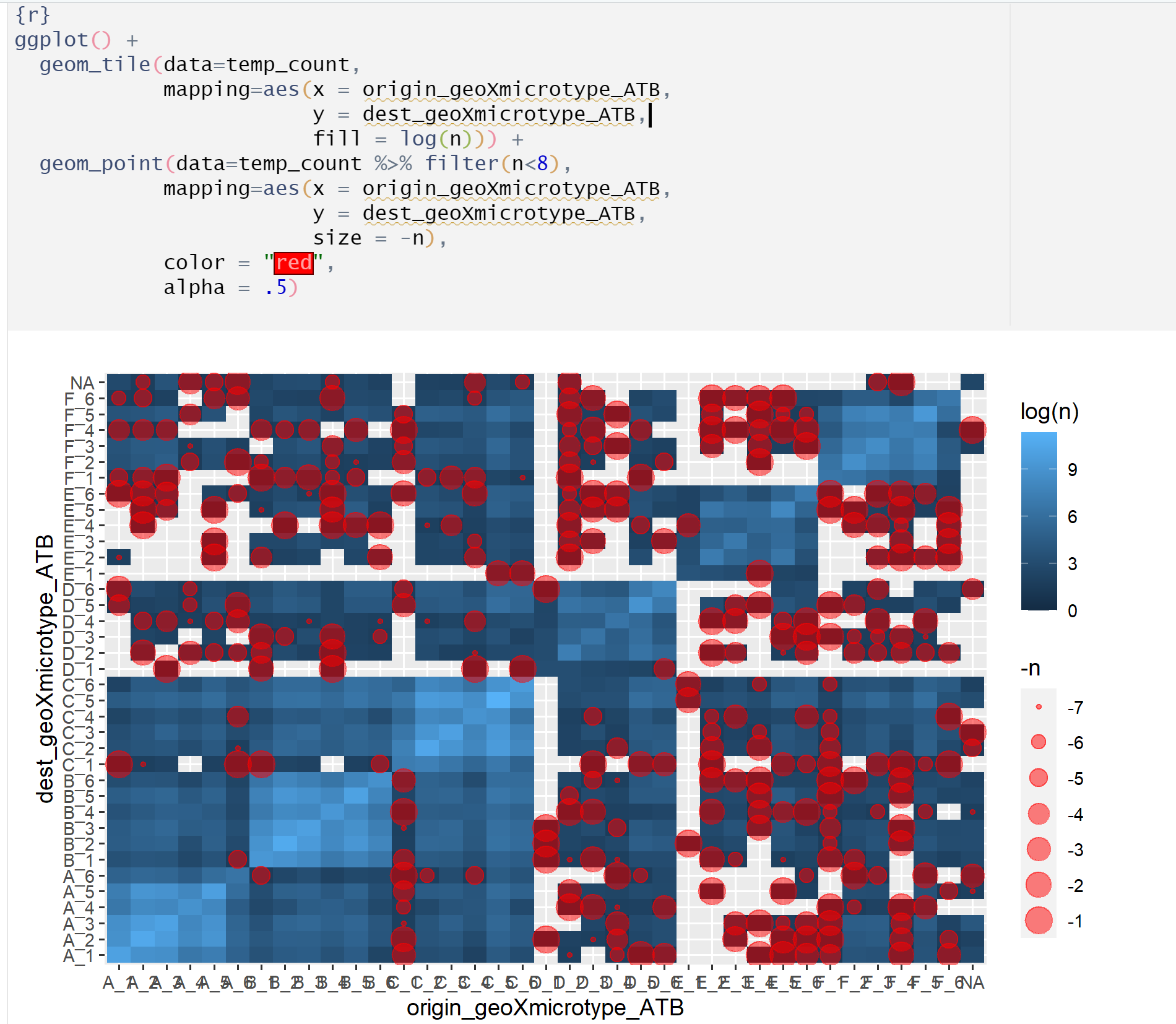


#### What level of heterogeneity is supported:

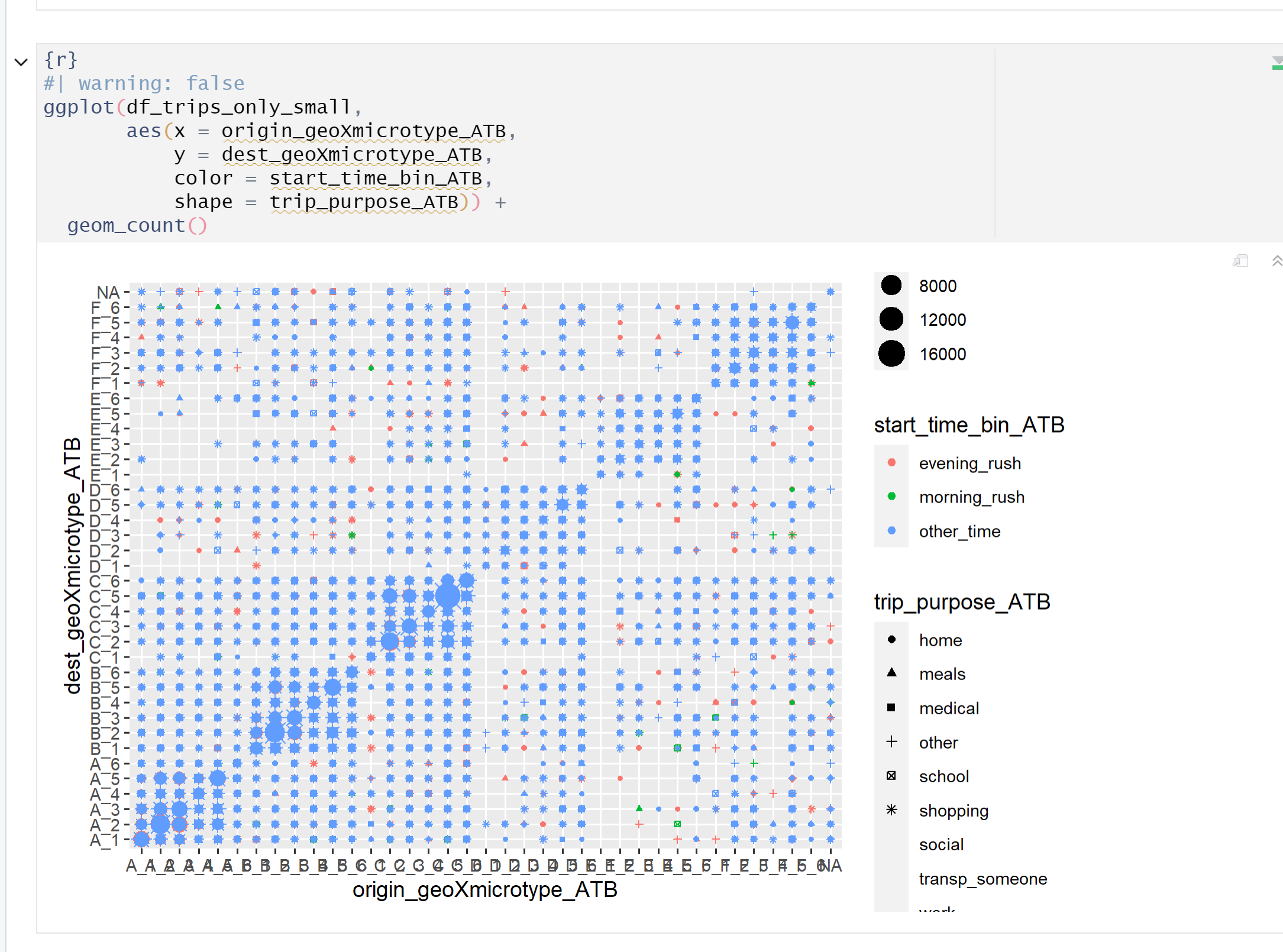




#### Identifying problematic geo types (less than 10 observations):



#### Showing where there may be gaps in trip purpose, start time bins (this could be displayed better; but for now, dots where the shapes are not overlaid means that some trip purposes are missing; a lack of all of the colors means that start times may be missing)



#### Examining how many geo X microtypes have all 7 modes:

