OpenStreetMap Data Case Study

1. Map Area

Location: Los Angeles , CA United States

<https://www.openstreetmap.org/history#map=14/34.0292/-118.2836>

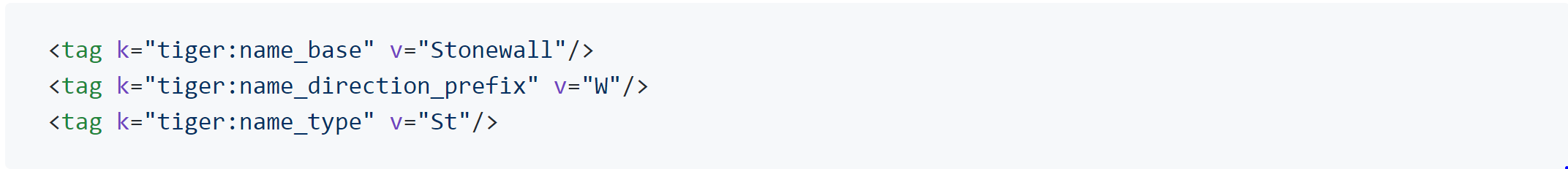
I am pursuing the master degree in Los Angeles and it is also my first city where I visited in United States. Thus, it is a good starting point to conduct my case study. However, the Los Angeles City is too large so that once I tried it for outputting csv file, It always takes me very long time and my computer clashed before end of compiling up the file. Thus, I restricted the map zone to south bound to exposition park and north bound to STAPLES center so that I can conduct the map researching much easily.

2. The Problems which I encountered in the Map.

After trial and chosen the region which my computer can perform map saving normally, the original file size is about 80MB. Excepted the data size, I also encounter some problem in the exploration the map especially in Way\_tags.csv. The problem and corresponding solution will be discussed individually.

Problem a. same source type of image in node\_tags.csv and way\_tags.csv: In node\_tags.csv, the tags “source” record how the image is uploaded. There are multiple sources and some of them are same but different order or abbreviation. For example, “Bing” is same meaning as “bing\_imagery”, and “survey;image;usgs\_imagery” is same as “usgs\_imagery; survey;image”. For this problem, In the beginning of scraping .osm file, I correct them to their respective mappings iteratively. Though the final file it still have some typo or misusing comma to semicolon, it solve the most these two type of duplicate words and integrated in to one phrase.

Problem b. align the “addr” type output to tiger GPS output in way\_tags.csv: Fortunately, most of way\_tags in this region are codify by tigers GPS. The xml format of tigers GPS are pasted below.



The tiger GPS segment the original full street name into name\_base, name\_direction\_prefix, and name\_type. The problem is all the type and prefix are abbreviation but in the “addr” type, the street name might be abbreviation or full name. In order to practice my coding skill and align with the street output of type “addr”, First, I change the format of name type by mapping so that all the street type is full name, and perform same procedire to name direction. Second, besides the original “street” key, I created the “name\_direction” and “name\_type” key for saving the original street data’s direction and type. Each of them are stored in the full name not abbreviation. I used regular expression to exclude house number then using the split function to locate the necessary data. In order to mapping correctly, I revise the previous quiz mapping directory and add new abbreviation street type and direction to make sure there are no missing street name in whole dataset.



Problem c. multiple values separated by semicolon exist in type “bld\_id” and “aid”: Though there are multiple values in one value column, after overviewing all dataset and each field, It is unnecessary to fix them in to each individual row due to meaningless filed and none of advance use in this project.

3. Data Overview and Additional Ideas

I used SQLite portable instead of executing sqlite3 on command line.

a. Datafile size

project\_map.osm: 82,850 KB

project\_db.db: 47,045 KB

nodes.csv: 32,806 KB

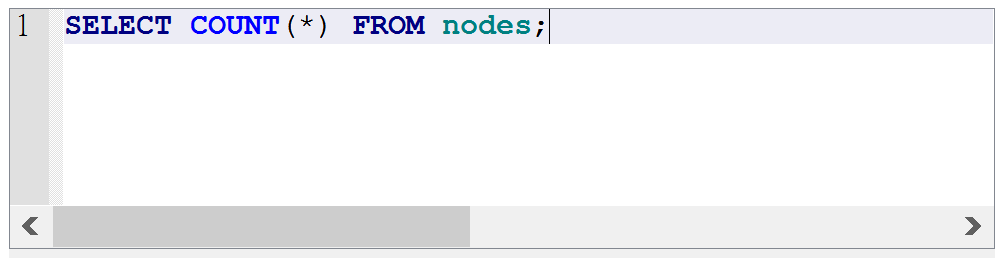
nodes\_tags.csv: 201 KB

ways.csv: 2,359 KB

ways\_nodes.csv: 9,943 KB

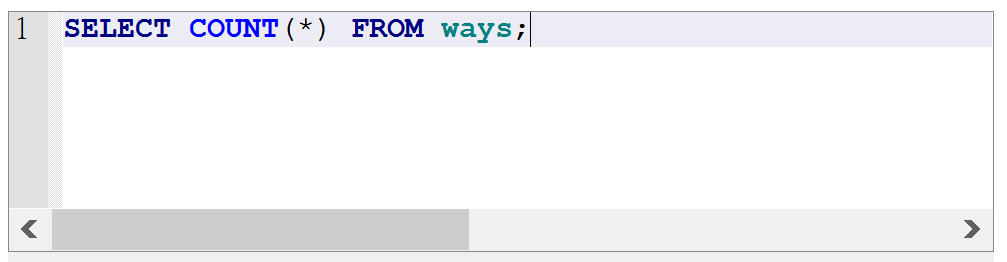
ways\_tags.csv: 7,908 KB

b. Number of Nodes



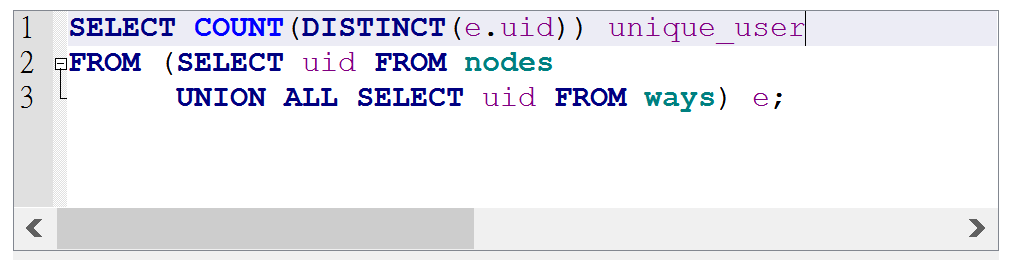
358097

c. Number of Ways



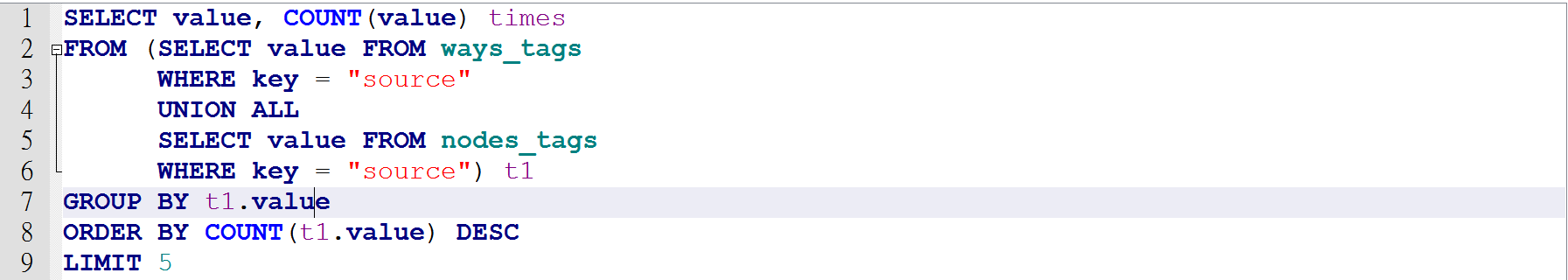
36010

d. Number of unique users

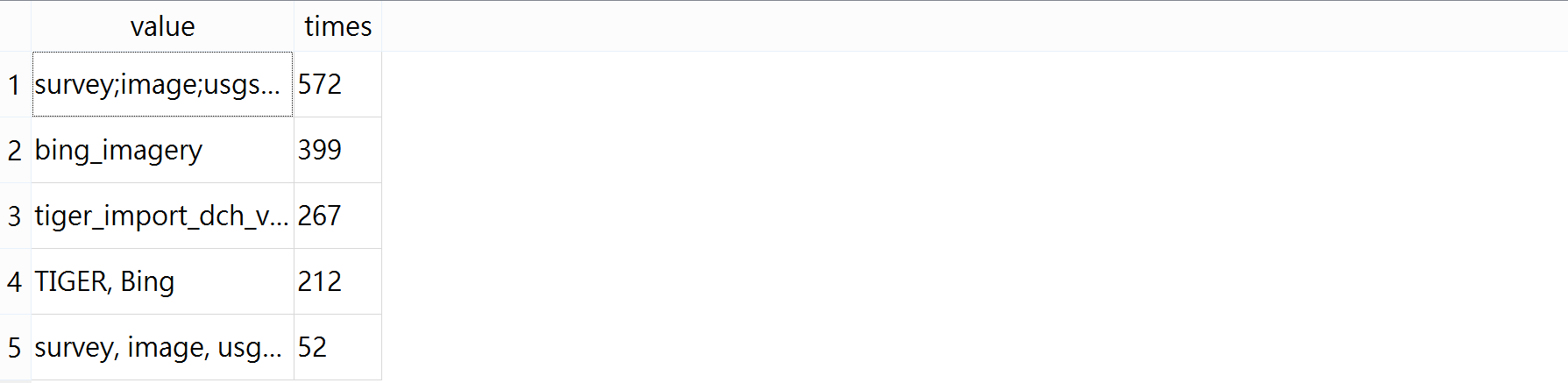


341

e. Top five of sources in ways

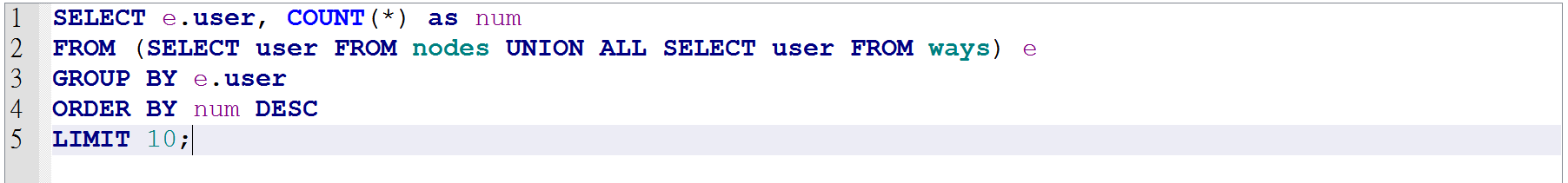


Result:

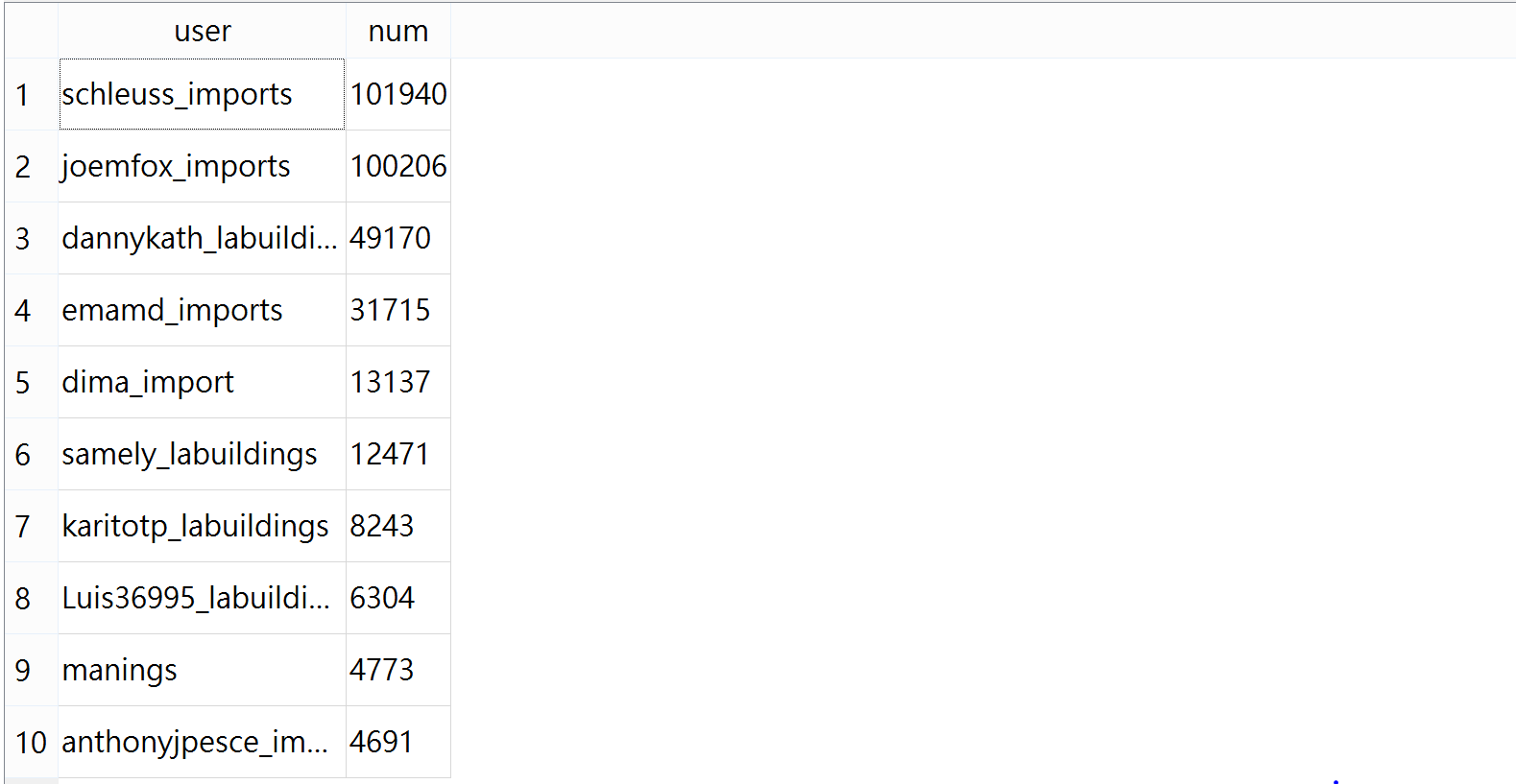


This shows that I still have some improvement on fixing the duplicate meaning value such as point 1 and point 5, but It doesn’t show up in the small range sample.

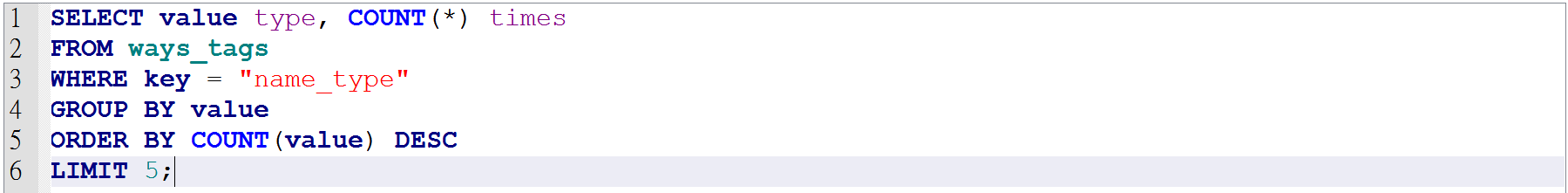
f. Top 10 contributing users



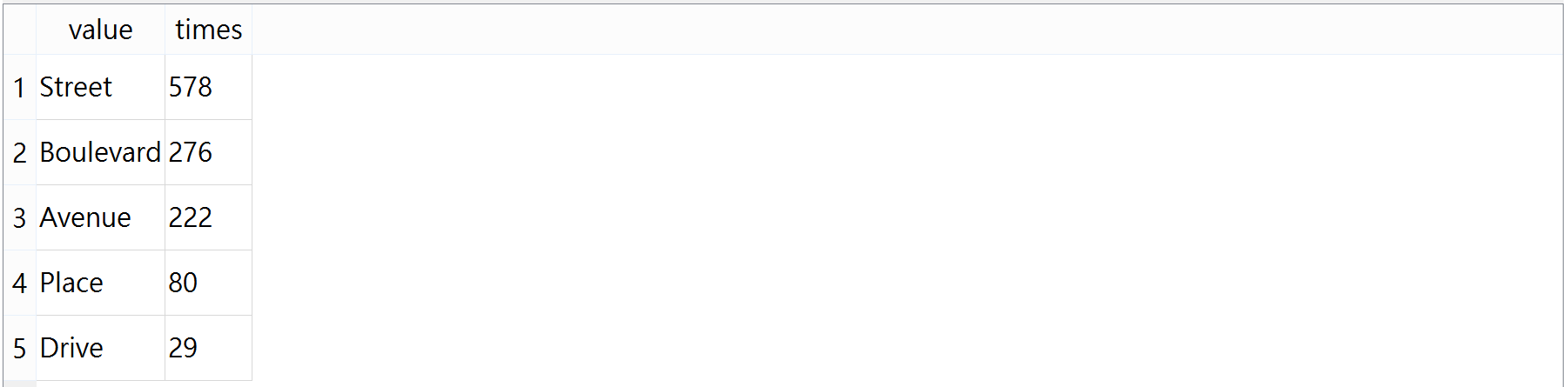
Result:



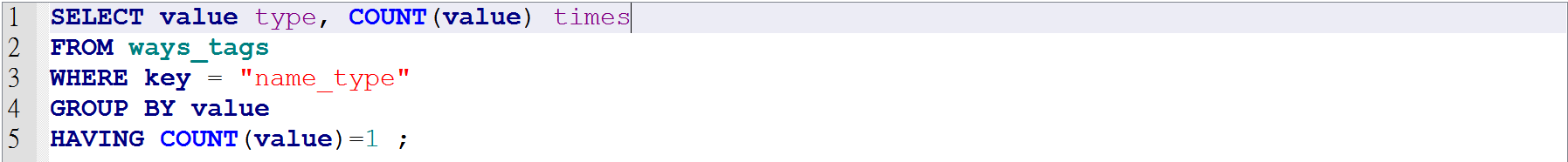
g. Top 5 street name type



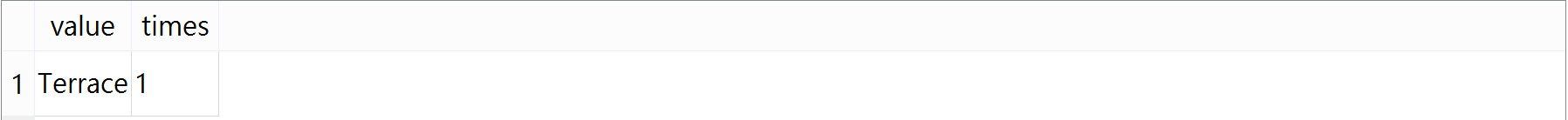
Result:



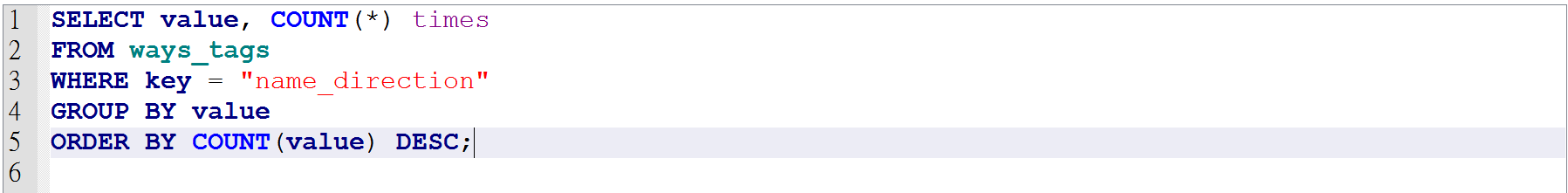
h. street name type only showed once



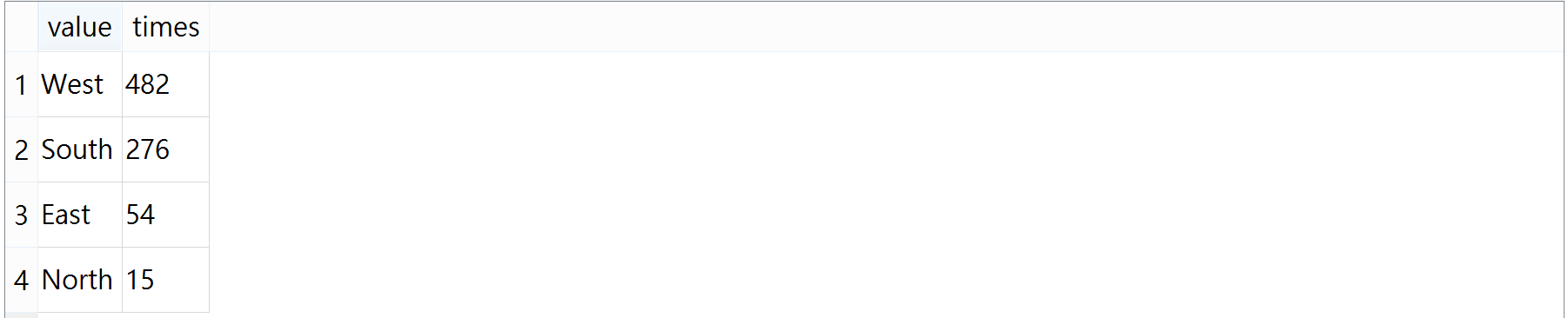
Result:



i. street name direction distribution

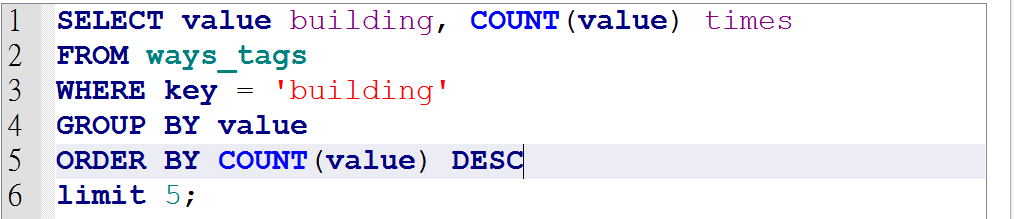


Result:

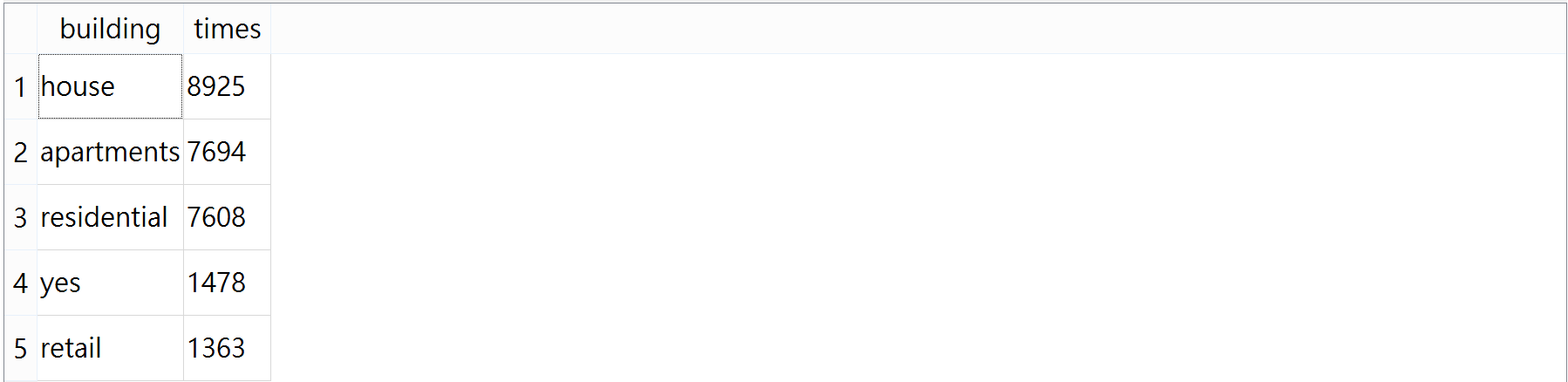


4. Further Exploration

For the further exploration, I try to find the top five of building style in the ways tags database.



However, the result show that the forth place of the sorting is called “yes”



Totally, “yes” indicate that there is exist a building but we cannot confirm what type of it is. Without re-entry or codify the Open Street Map, one of the solution is looking up for next id’s or previous id’s building type and modify to its value. For example, if id = 123 have building = retail, then give the id = 124, which have original building value = yes to the retail. This way can perform easier when the next or previous building type are already confirmed. However, it might have some biased-on misclassification because we only guess the value by its neighbor. Thus, the other way to improve this case is enlarge the range of we looked up. For example , for the guess of building type of id = 124, we look up more from 121 to 127, to suggest the value by more its neighbor. Same as the first method, this still have some probability on misclassification because most answers might be between house, apartment, and residential which are top three building type near the zone I chosen.