• Are any of the attributes useless for analysis? How can you tell?

I also think the bar charts displaying data for one attribute is useless because it’s only displaying data for one attribute, and there is nothing to compare it too.

• Why did we only consider some of the numeric attributes to identify potential outliers using the rule of thumb we discussed in class. Is there a potential problem with this approach for some attributes in this dataset?

We want to consider attributes that have outliers when there is a lot of data on a variety of attributes instead of just one attribute because then we can make correlations between things and use the rule of thumb on outliers and suspected outliers. The potential problem would be like if there isn’t enough data, i.e the bar charts that have data for only one attribute because then how will you make correlations with other data?? You can’t.

• Which attributes have suspected outliers or outliers, based on the Box and Whisker plots?

A lot of data has outliers and suspected outliers based on the rule of thumb.

• Looking at the bar charts for the categorical variables, do some values for attributes appear to be very unusual?

I think for example, the DST Host Count bar chart is weird because all the attributes have a low display of data, except for the last one, which is exponentially high compared to the other attributes.

• Identify attributes with a large number of zeros, based on the interquartile range from the table of numeric attributes. Looking at the histograms, what do you notice about these attributes?

These attributes have an abnormal amount of zeroes in other attributes such as in the srv count histogram, but on the ends of the barchart, the numbers are really high.. so it means there are really high spikes in data at certain points.

• We will be seeking to distinguish normal and attack behavior. Do you think that we should eliminate rows with unusual values?

I think we should eliminate rows with unusual values because it means nothing to us, it is just garbage values… but if those values were to mean something, then it could easily mean that those would be spikes in the data.

• Why are some of the correlation coefficients set to NaN?

Because there is too much correlation for there to even consider a certain number or there is no data on something.

• Which pairs of attributes have a correlation coefficient greater than or equal to 0.98? Why might these attributes be correlated?

The bar charts that have data on only one attribute has a correlation coefficient greater than 0.98 because it is the only attribute in that chart that has an overwhelming amount of data. I think they are correlated this way because of how often something got attacked.

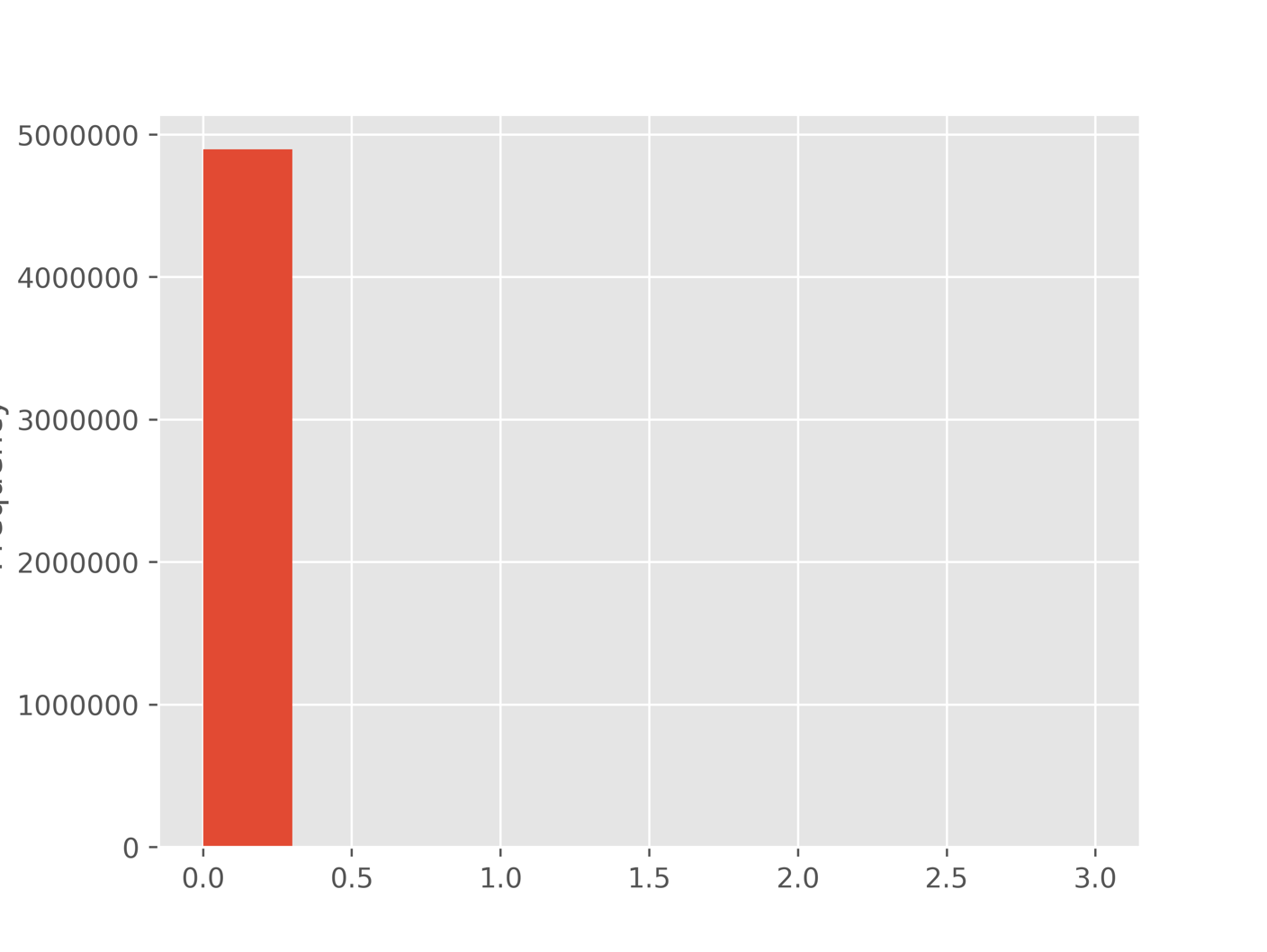
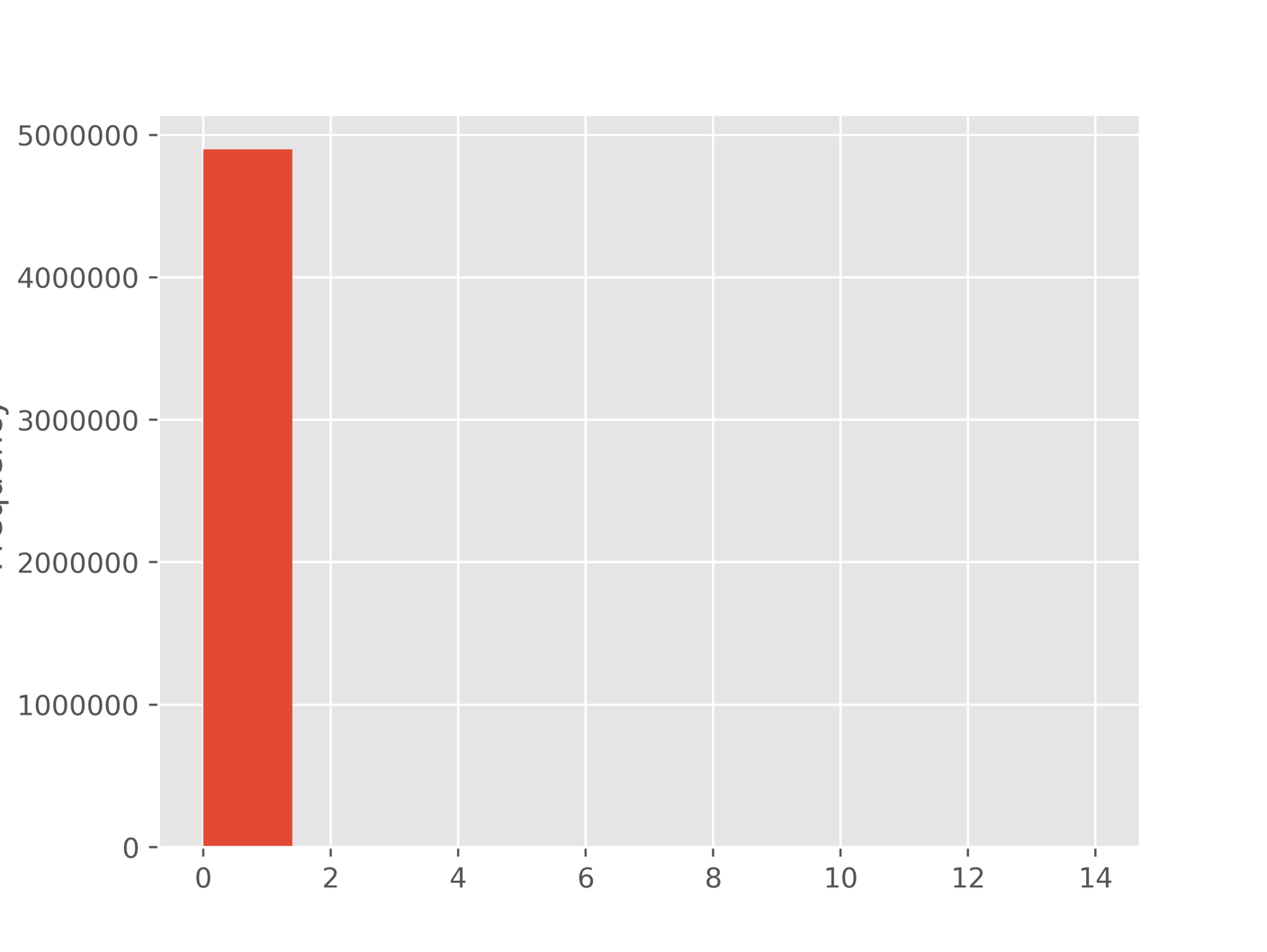
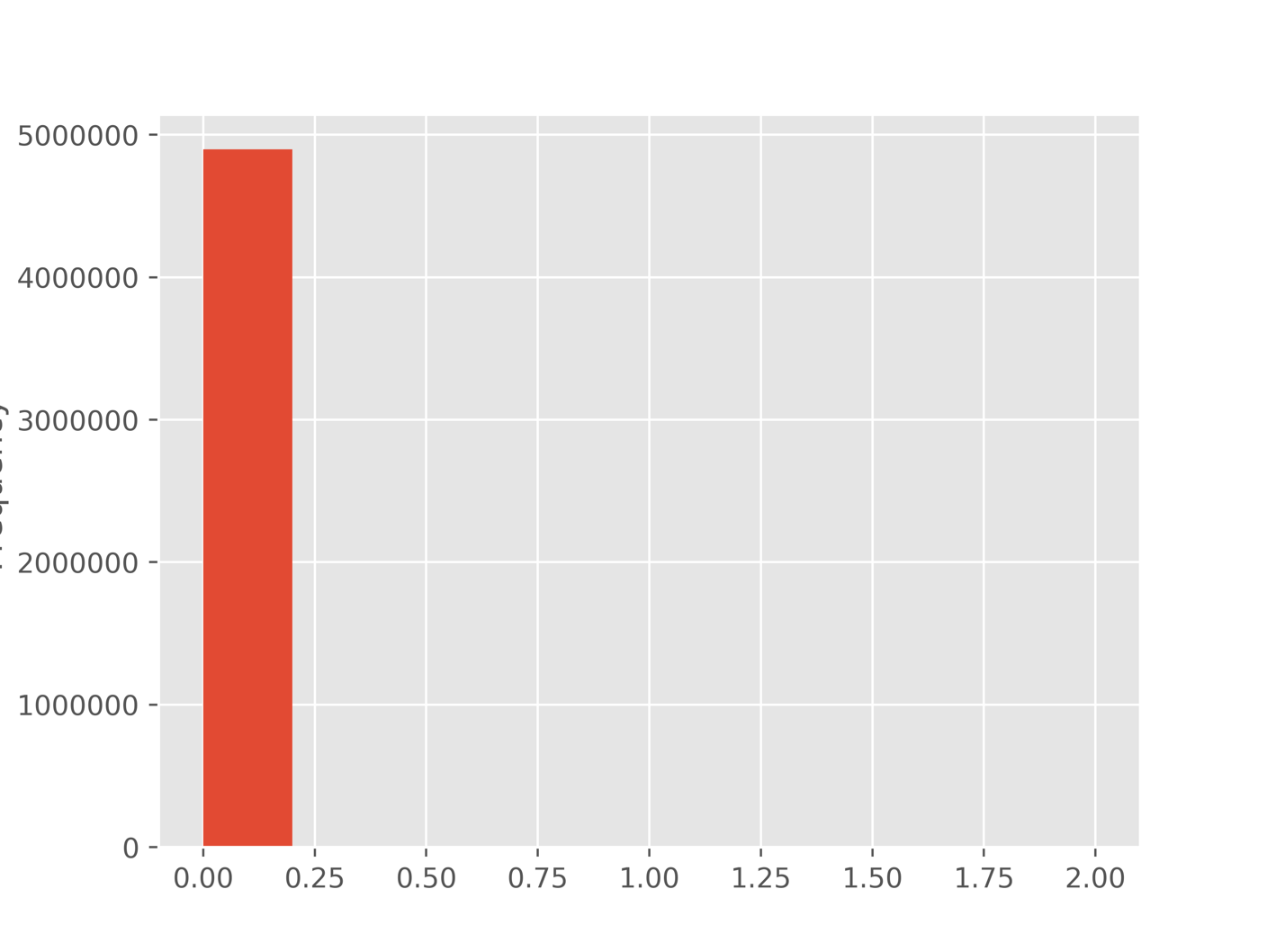
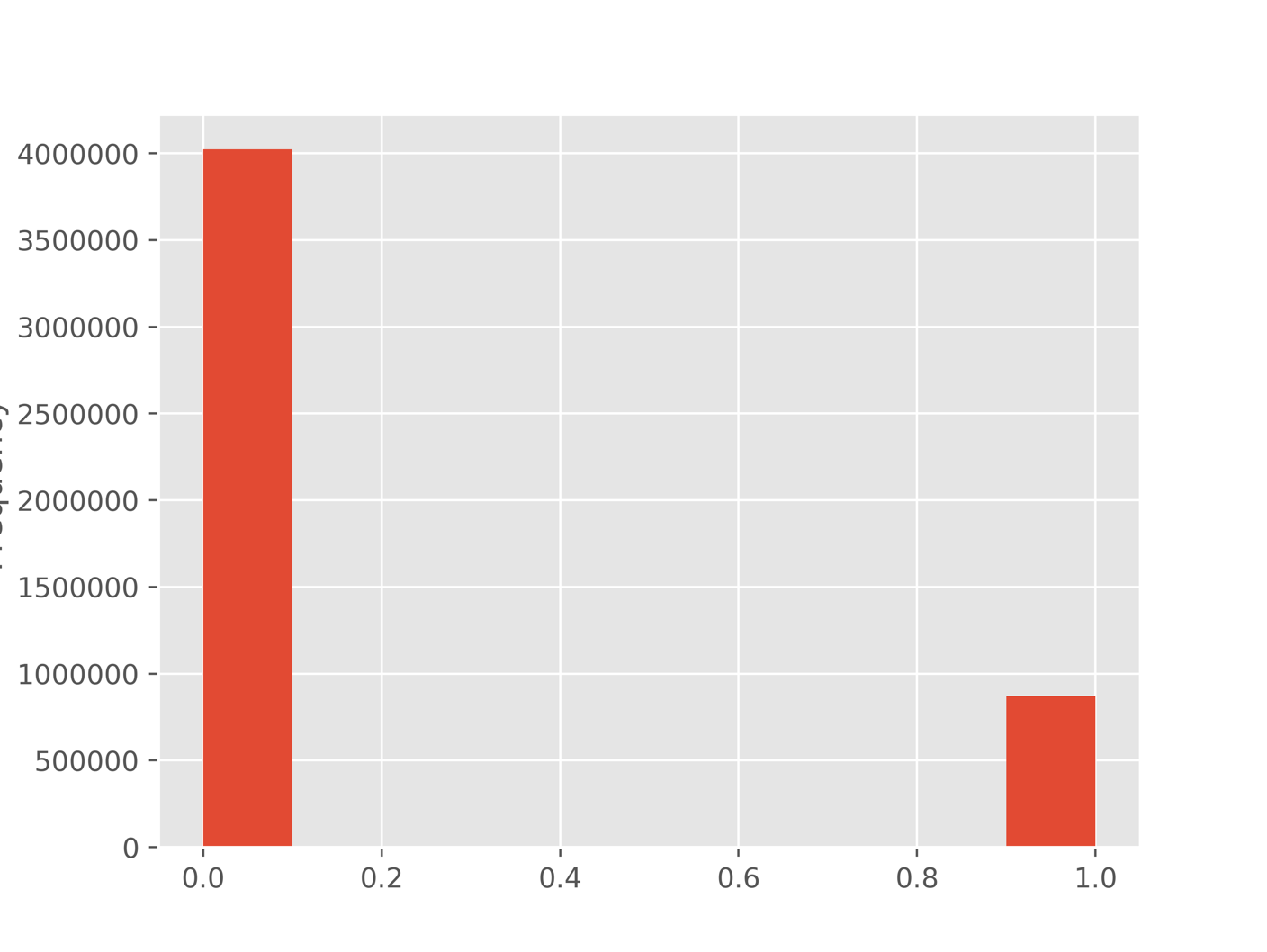
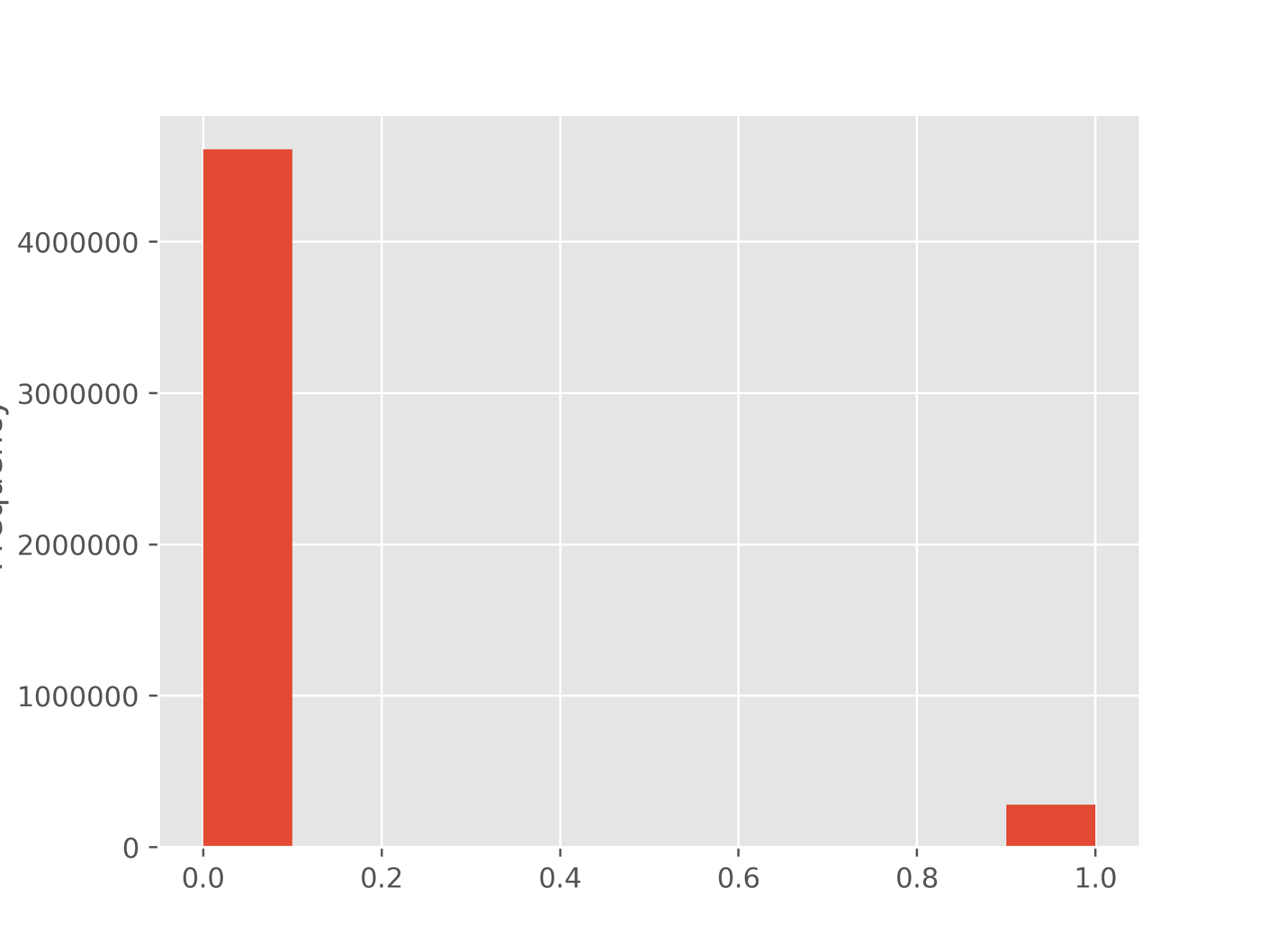
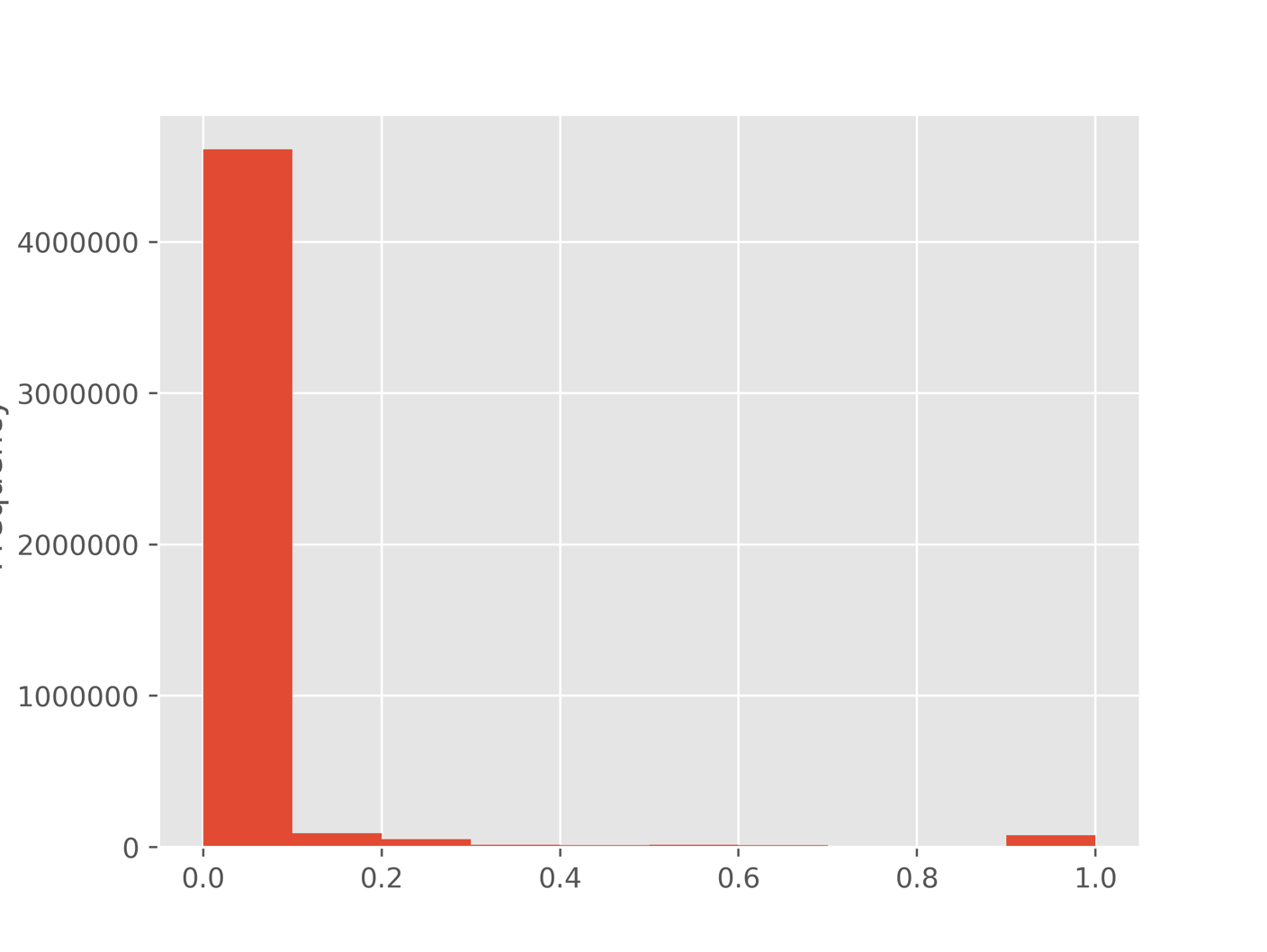
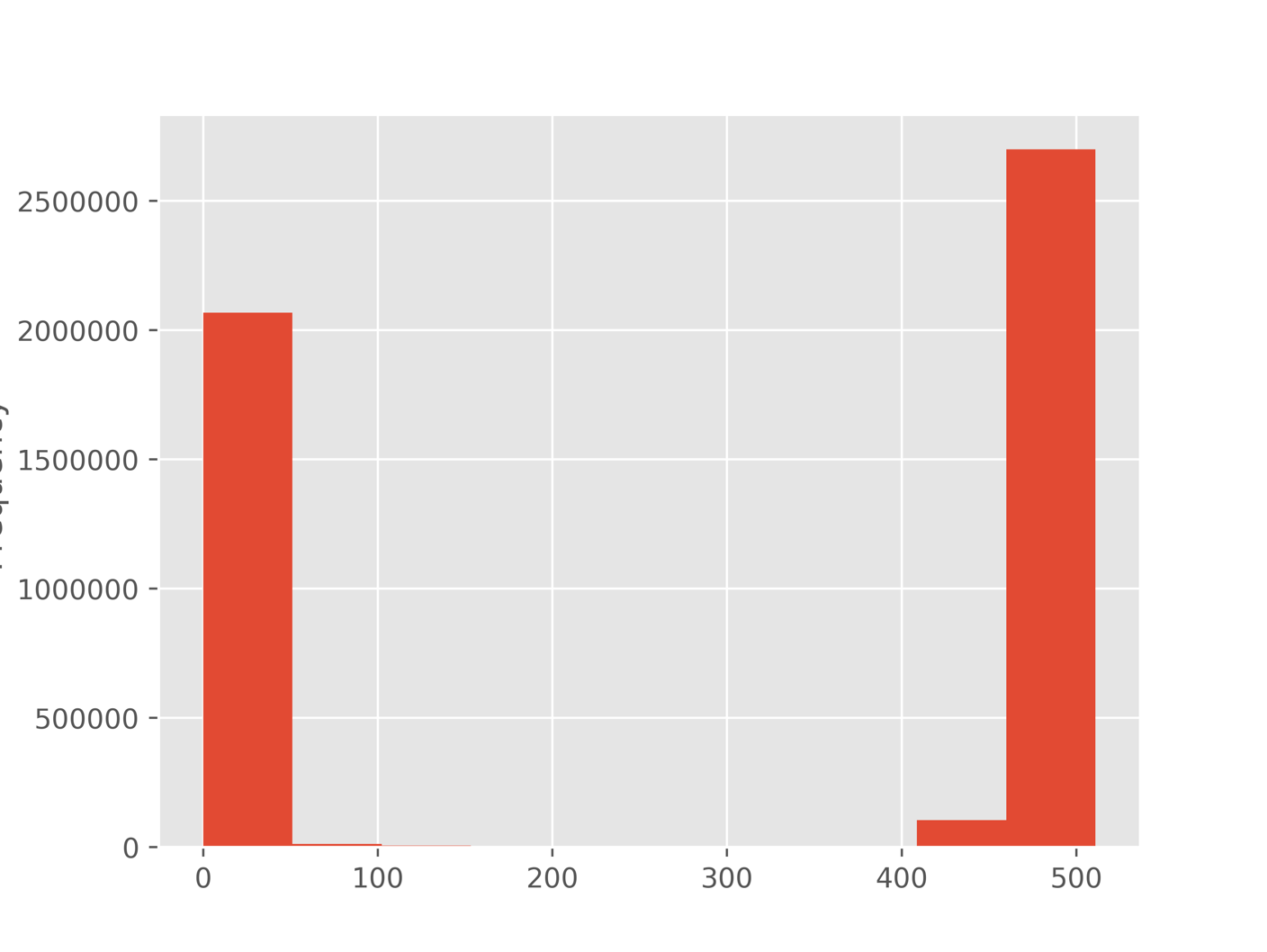
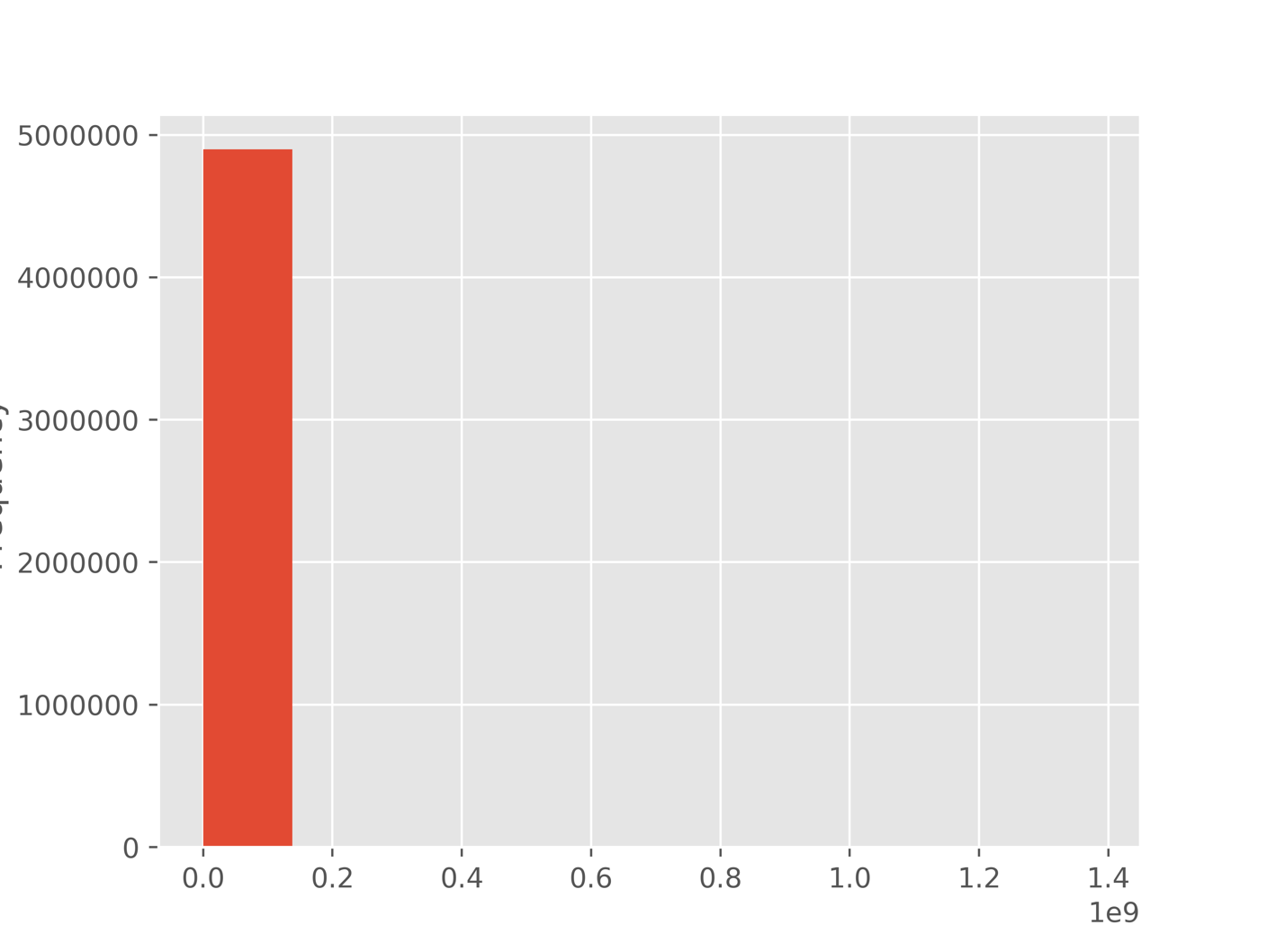
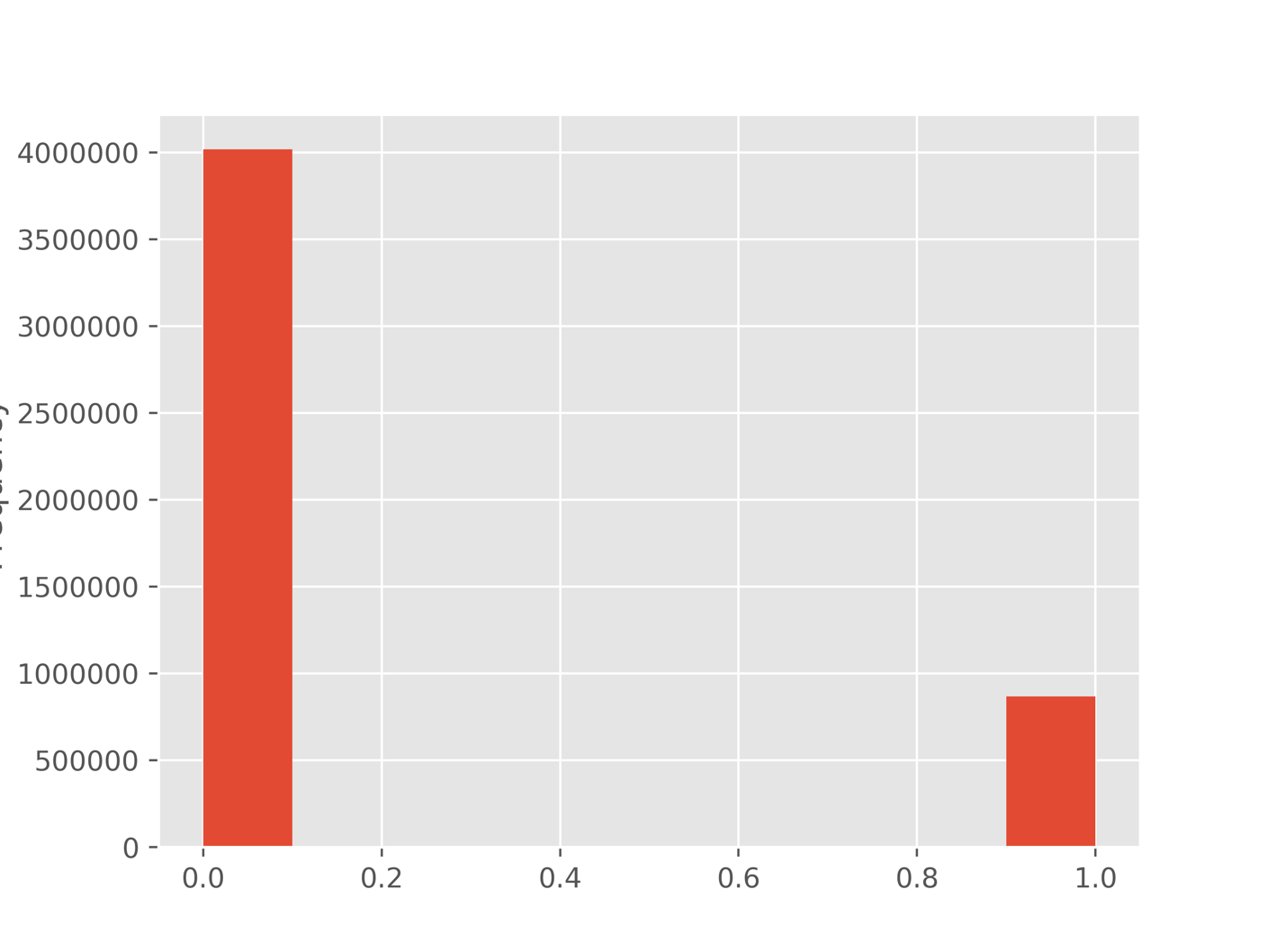
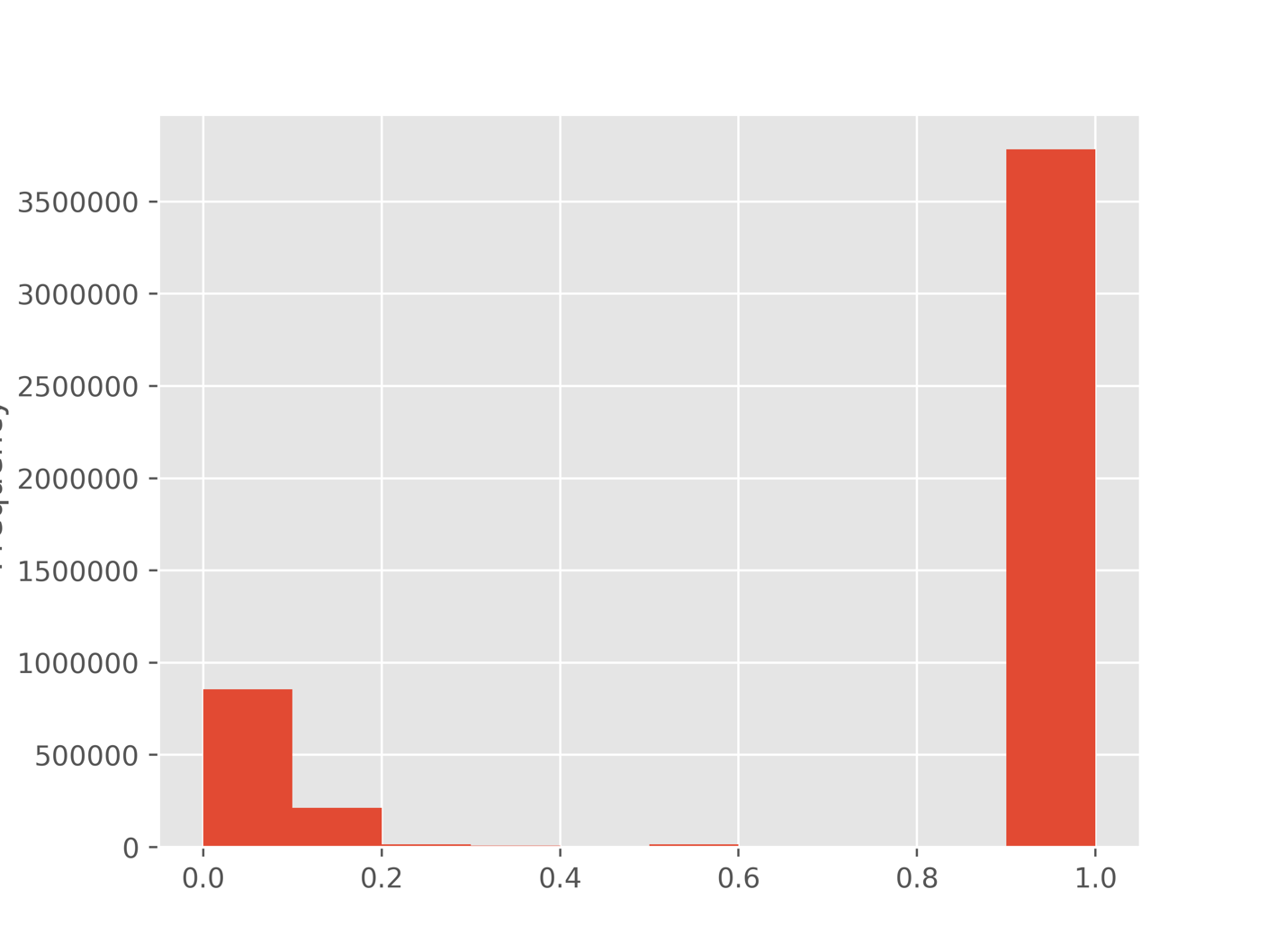
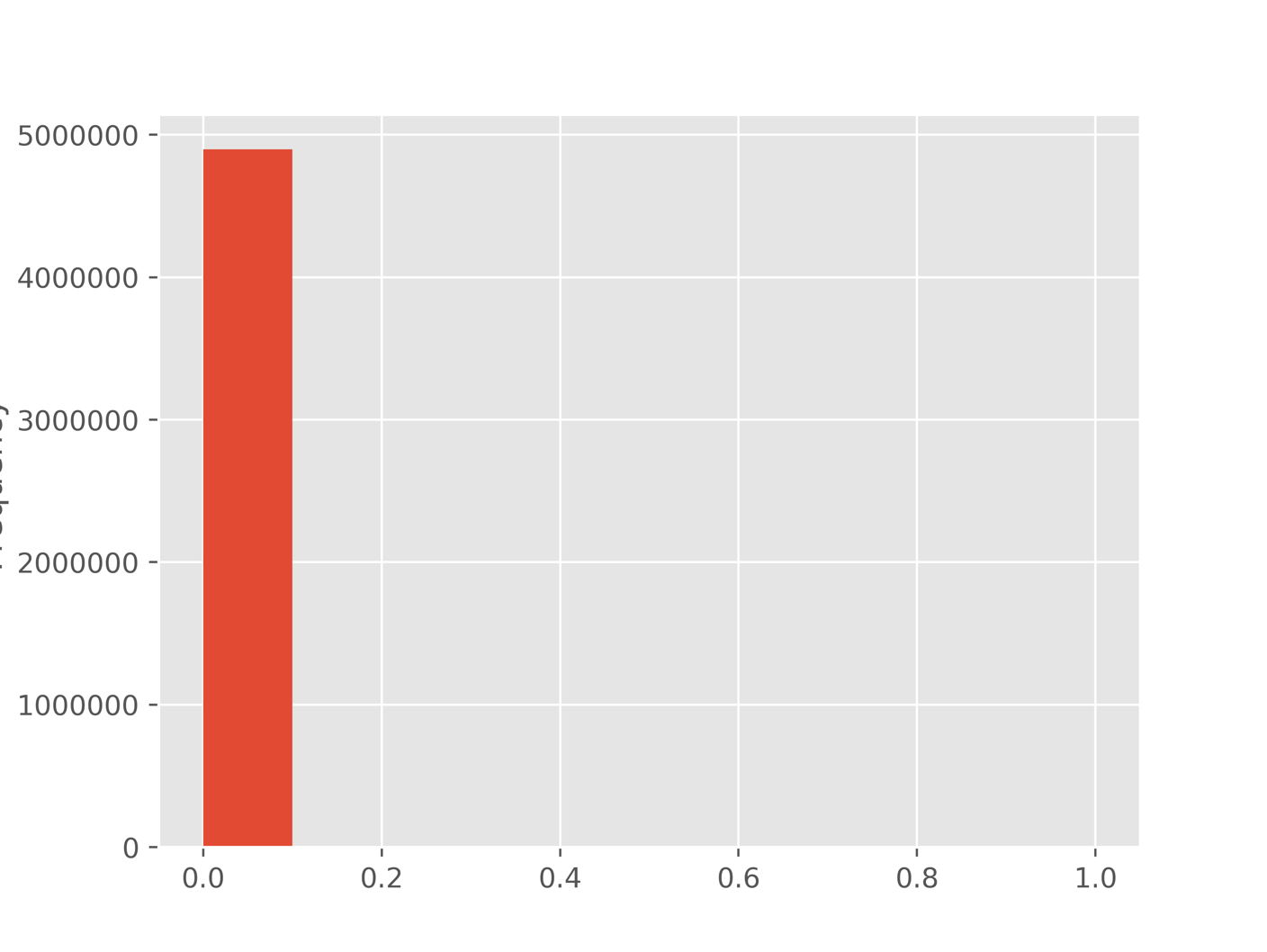
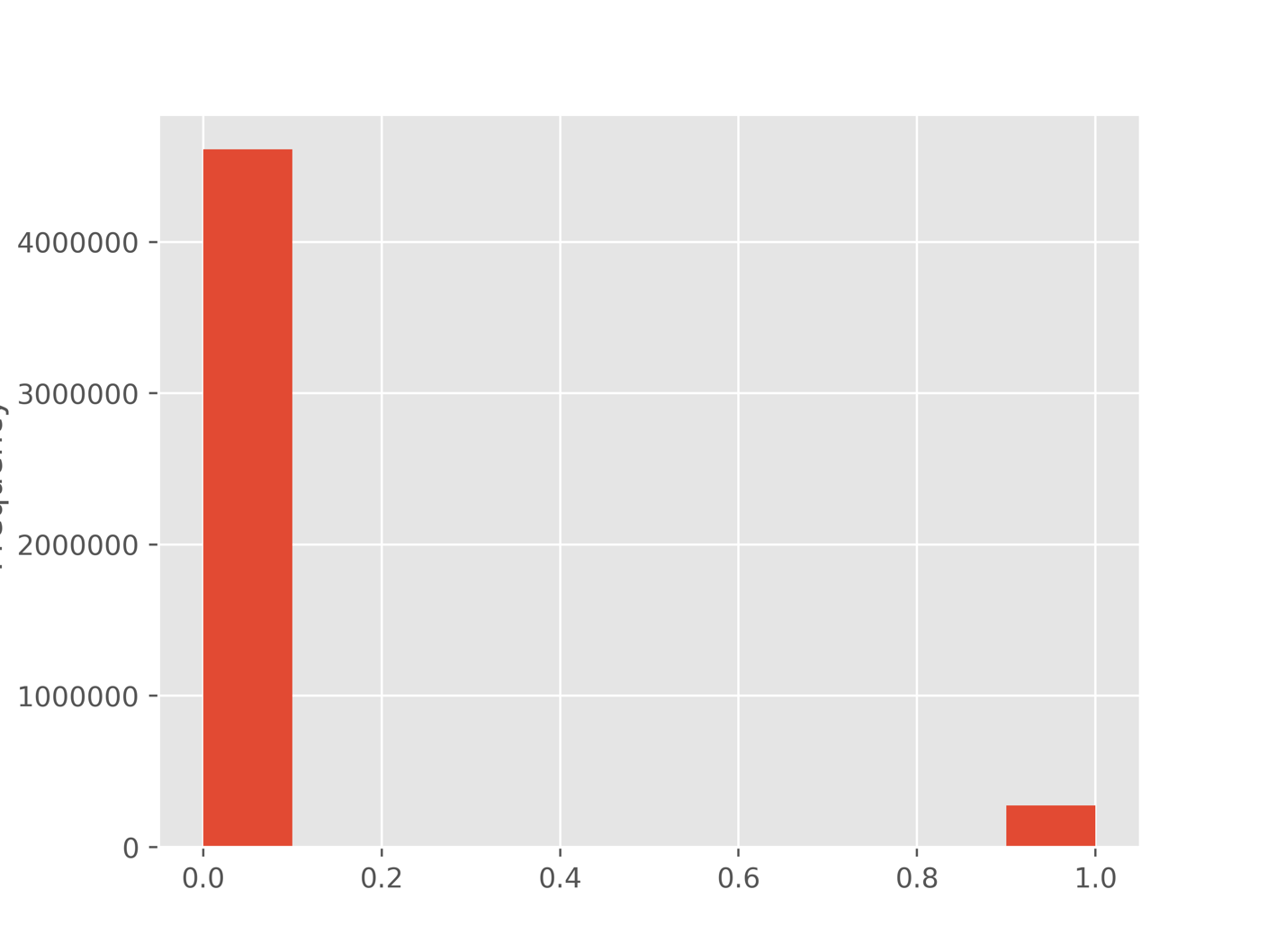
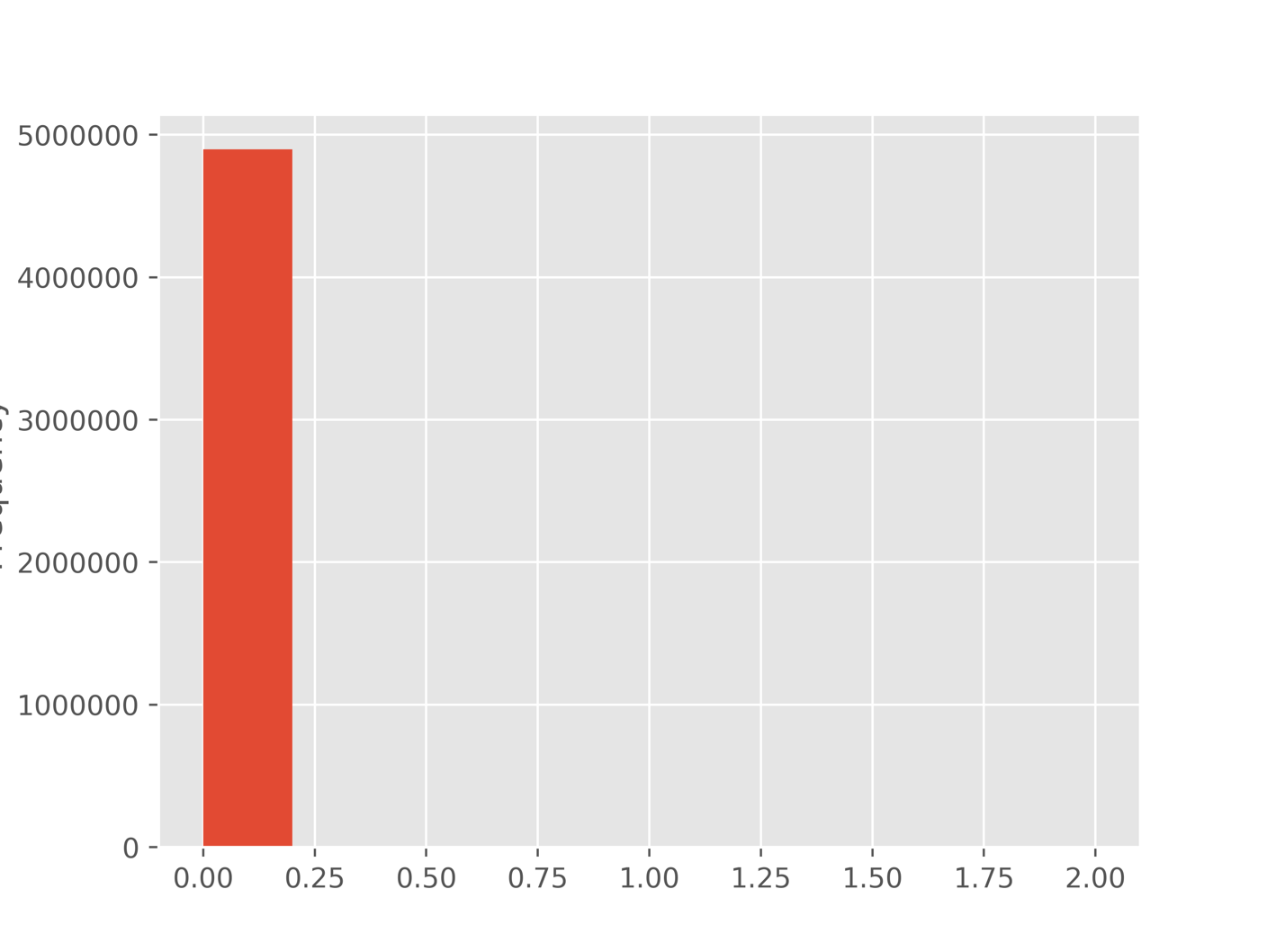
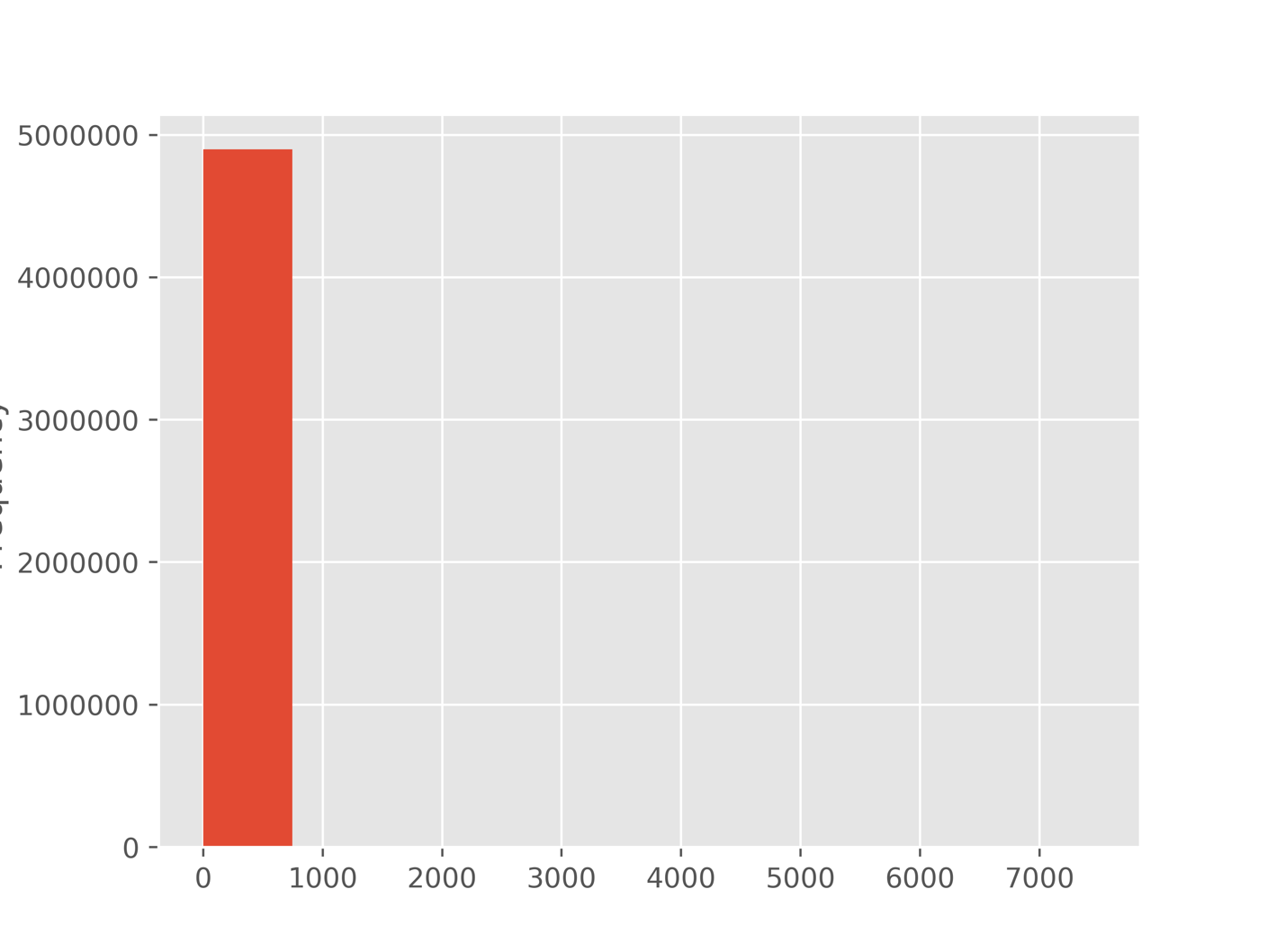
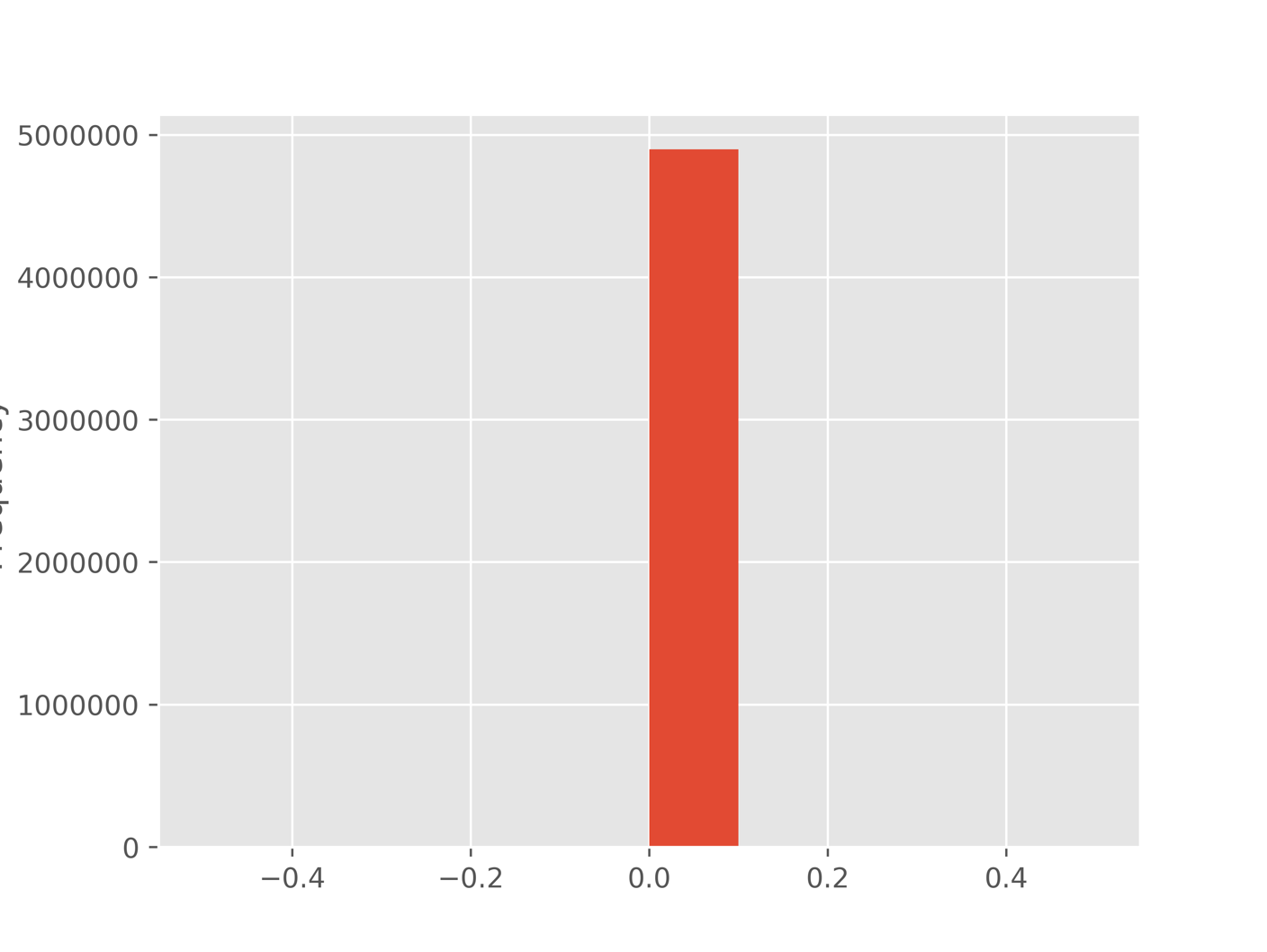
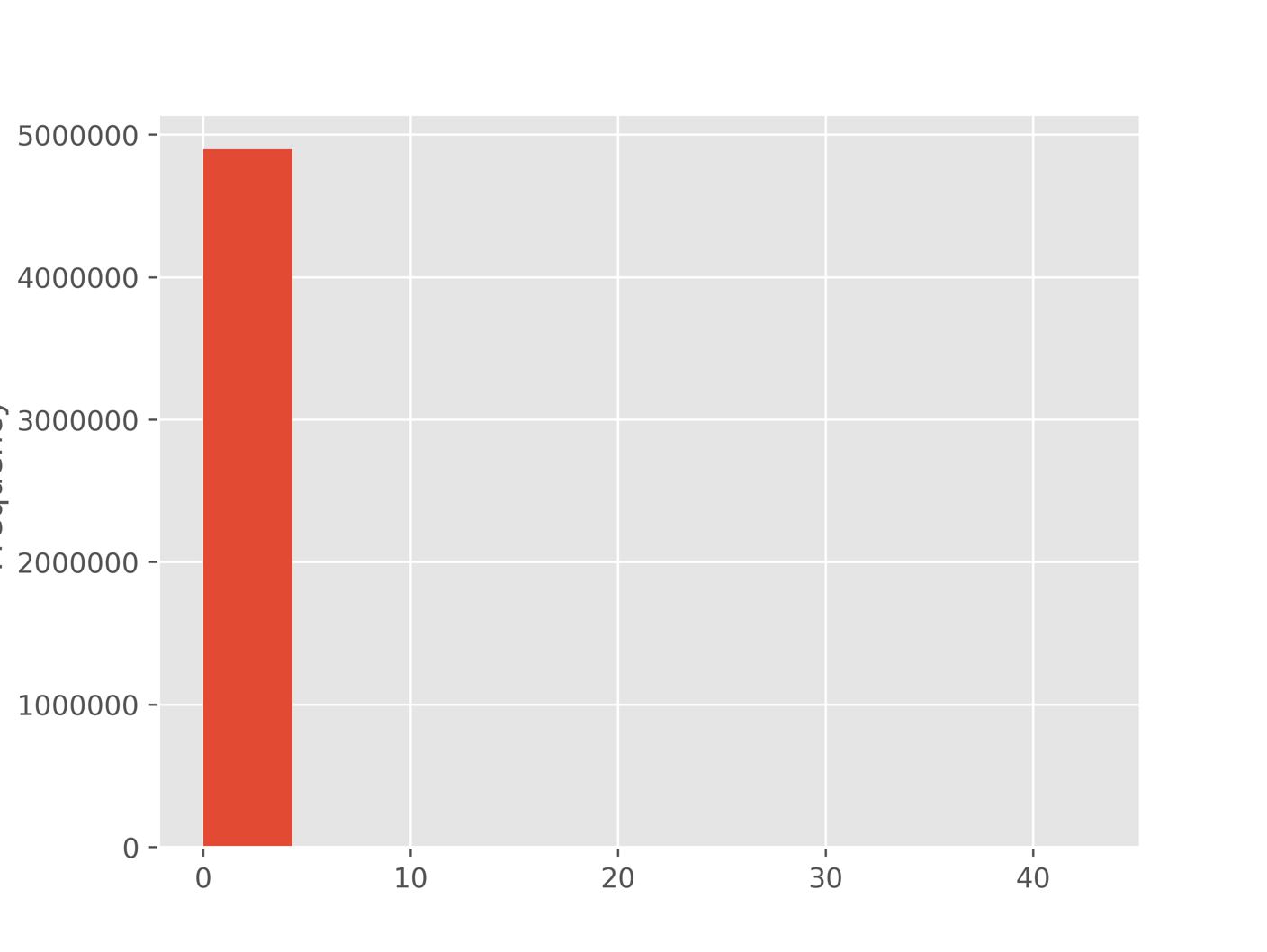
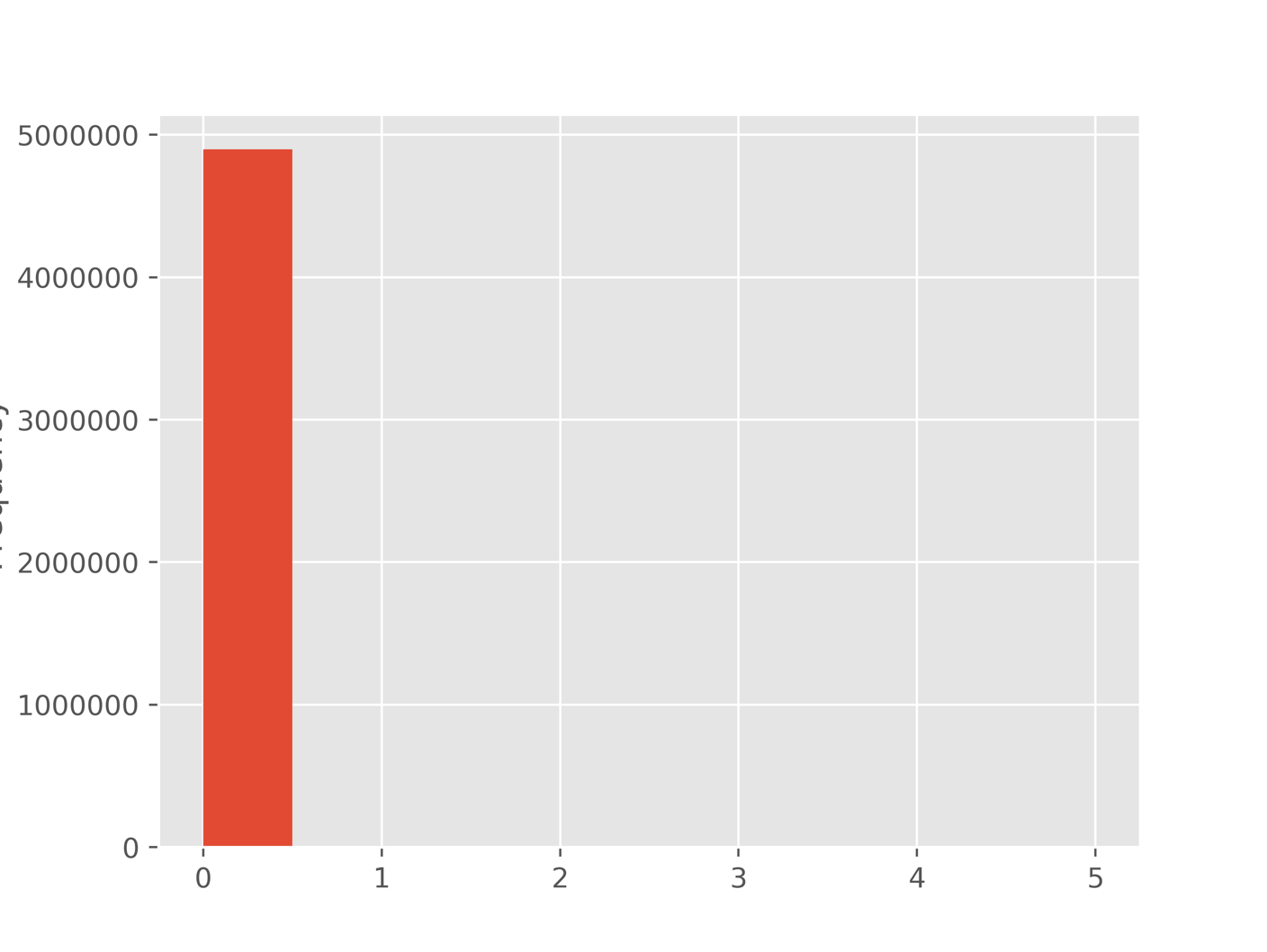
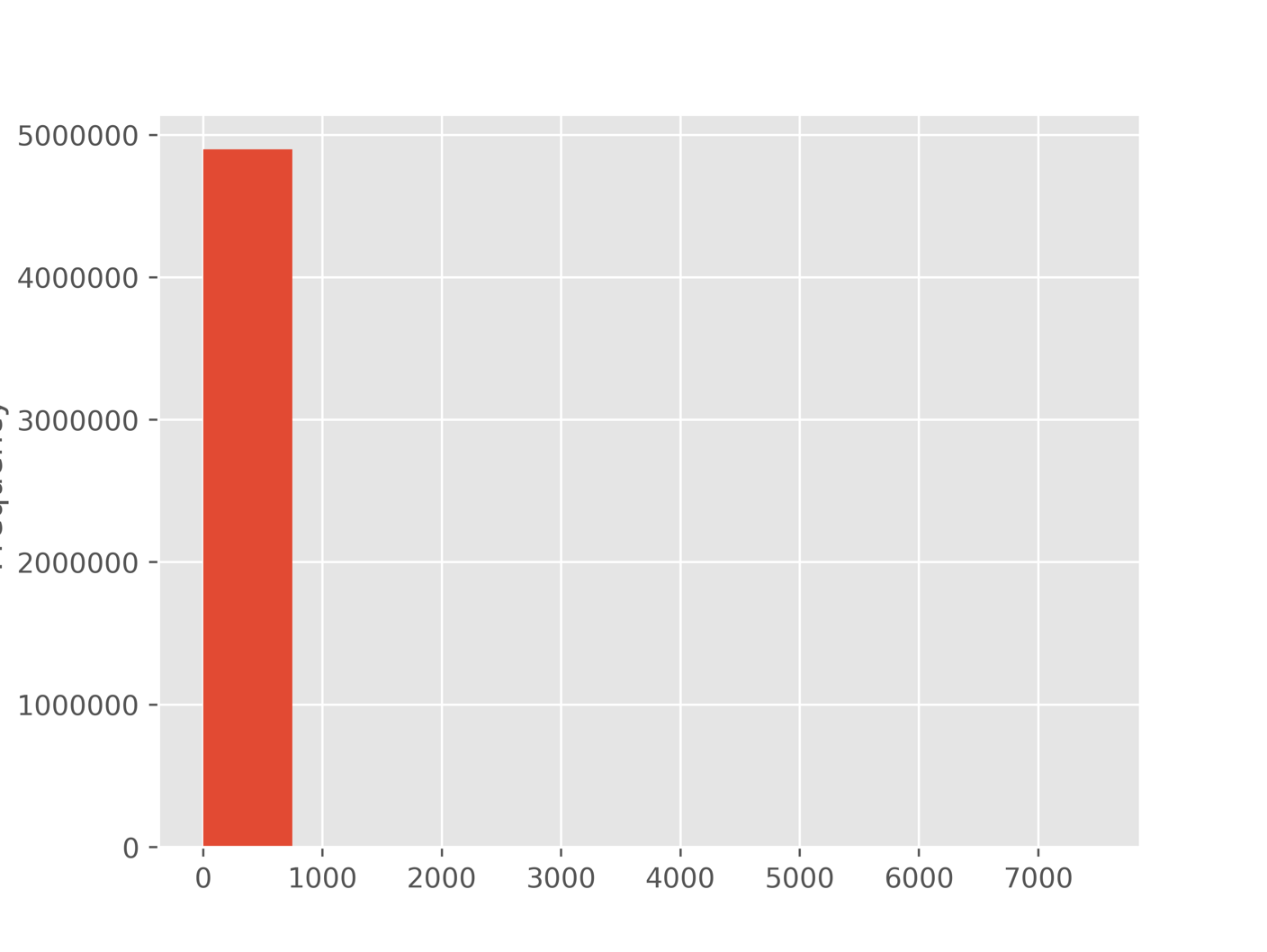
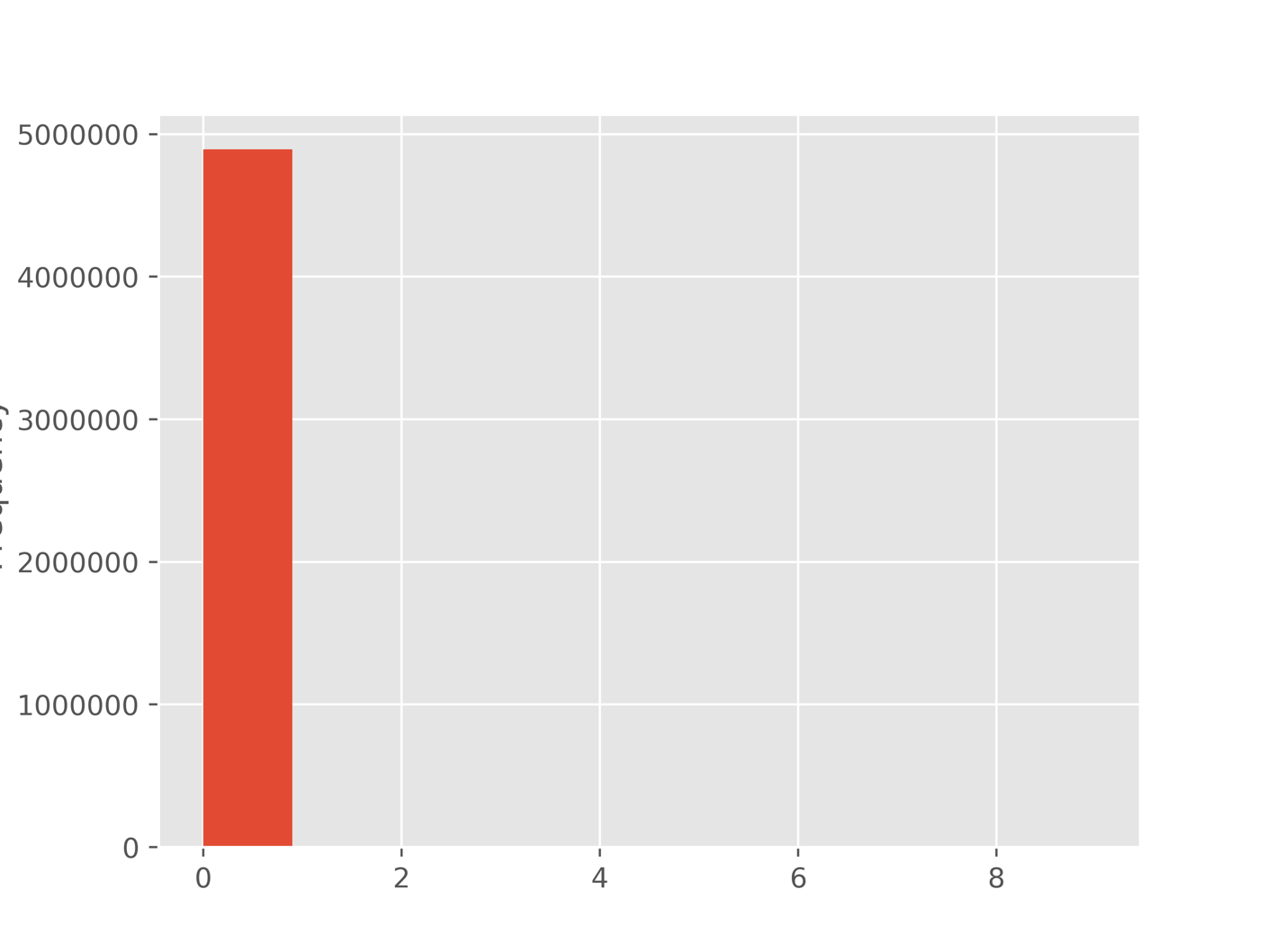
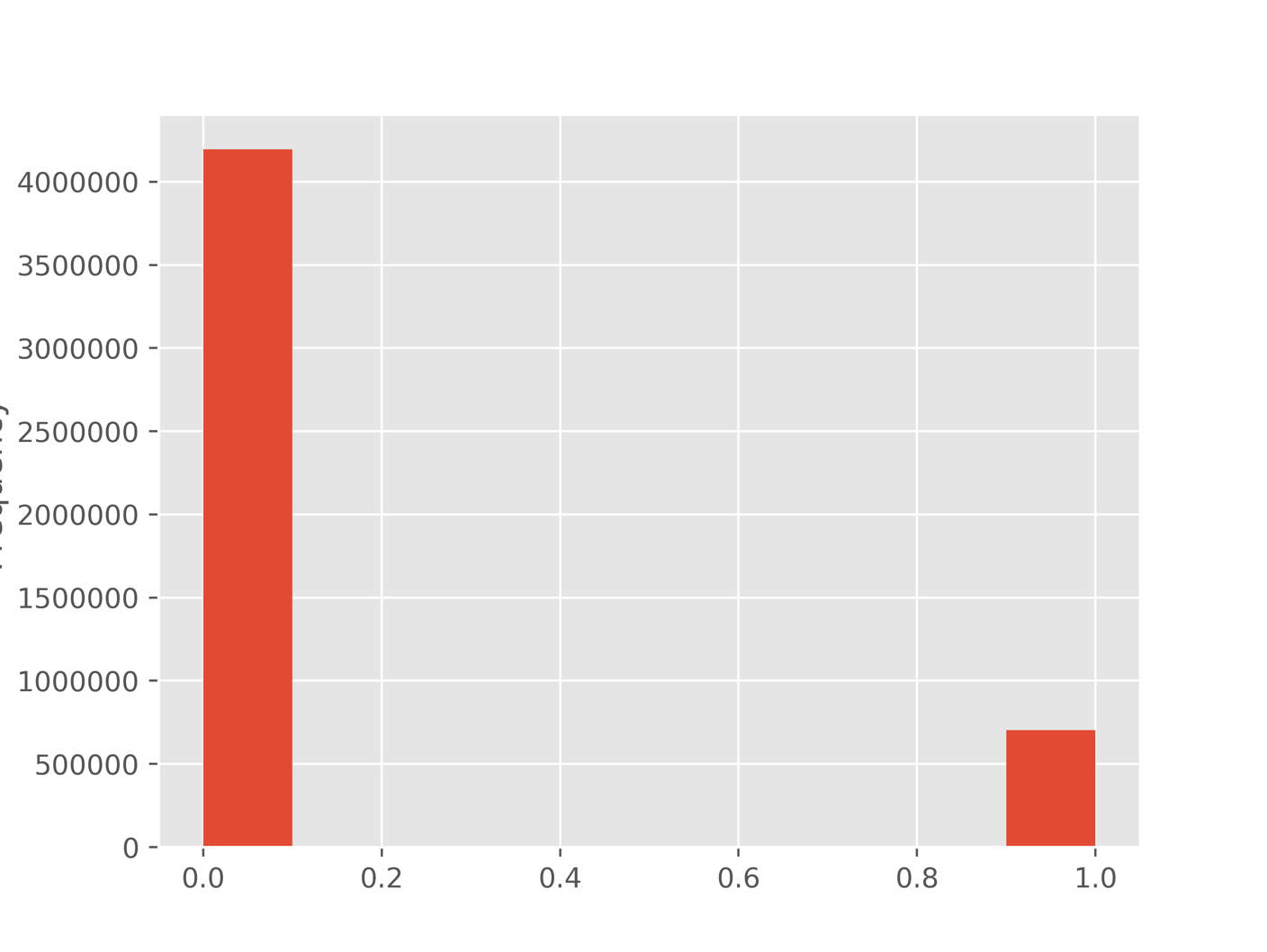
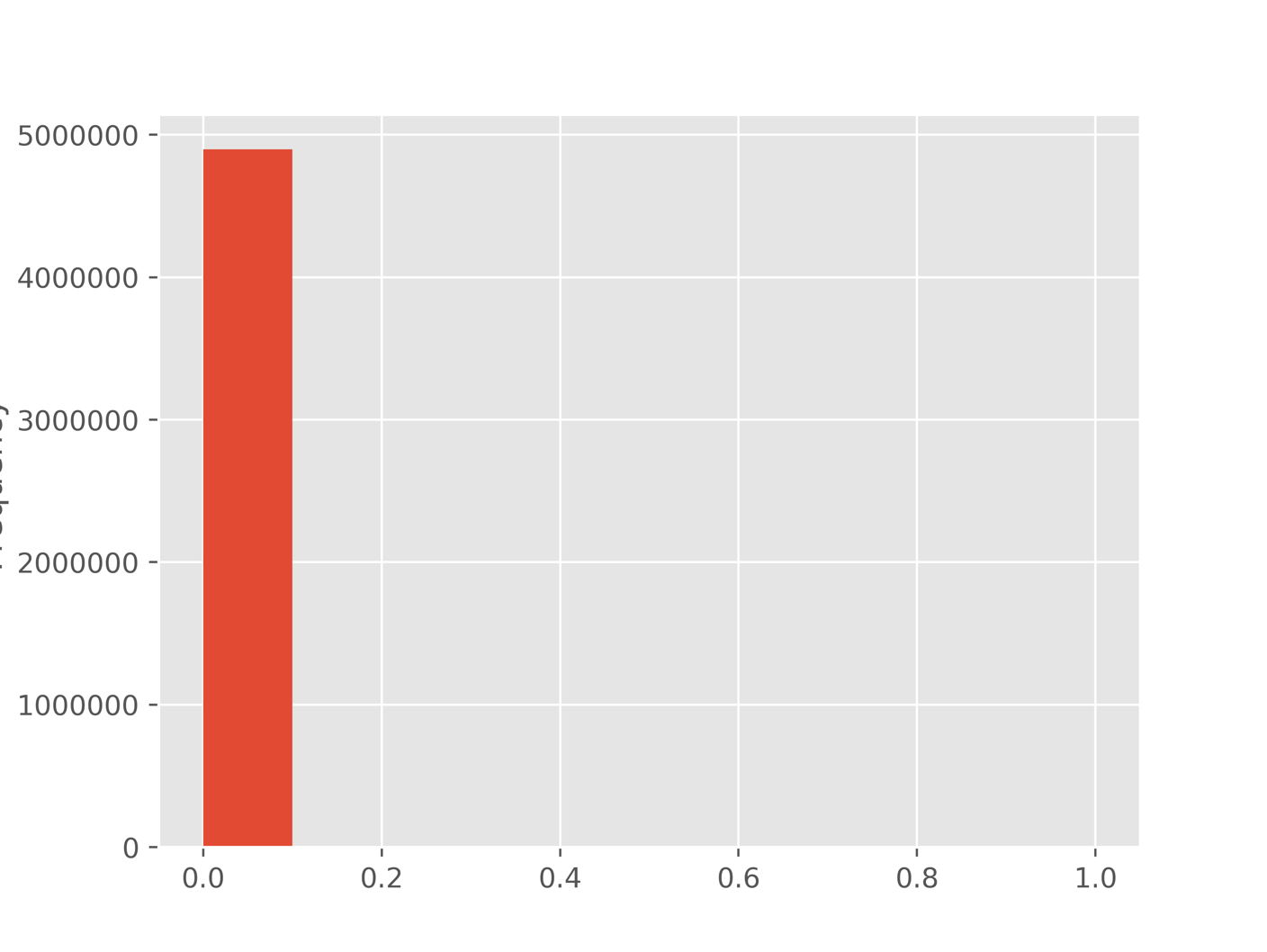
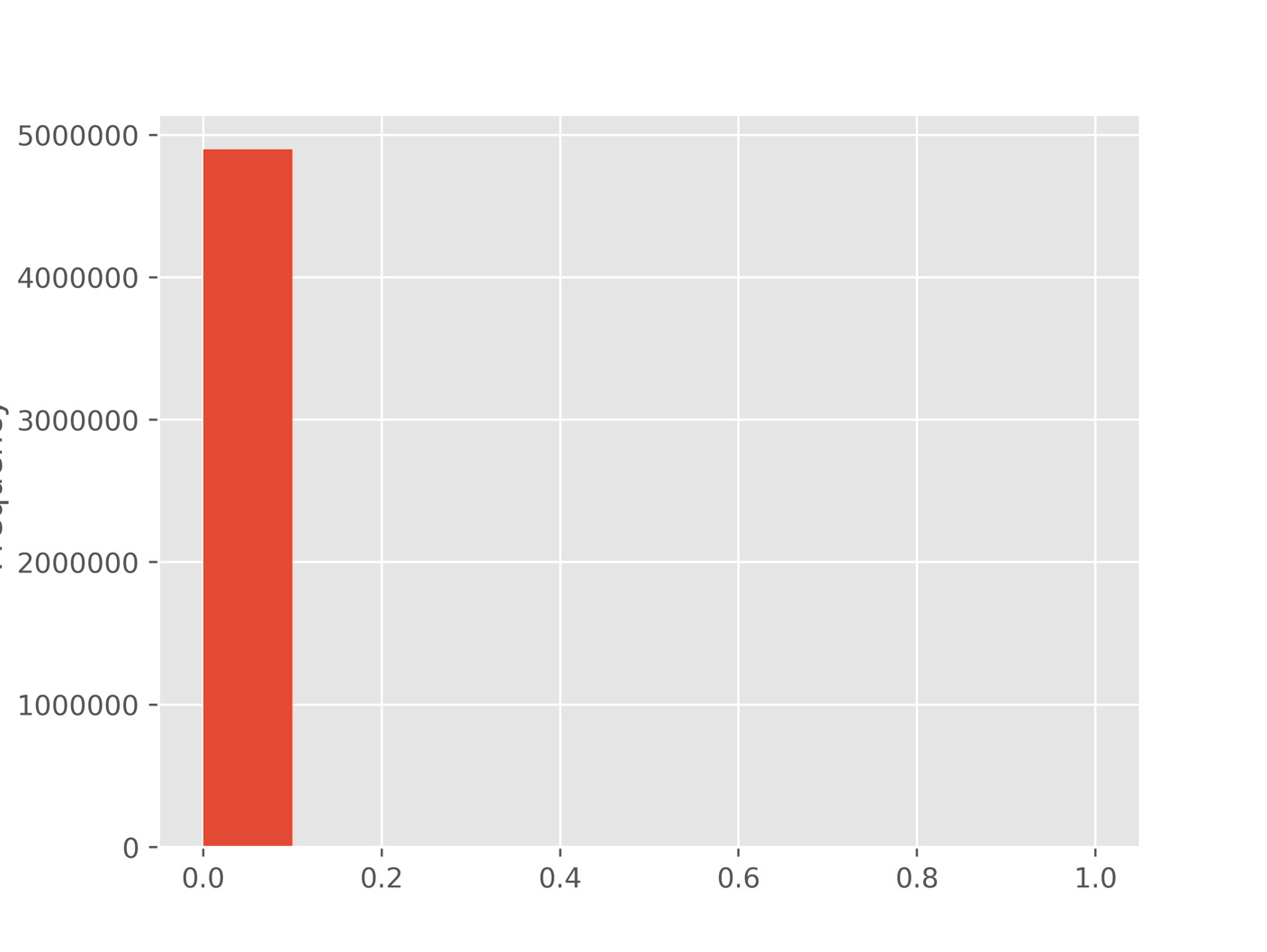
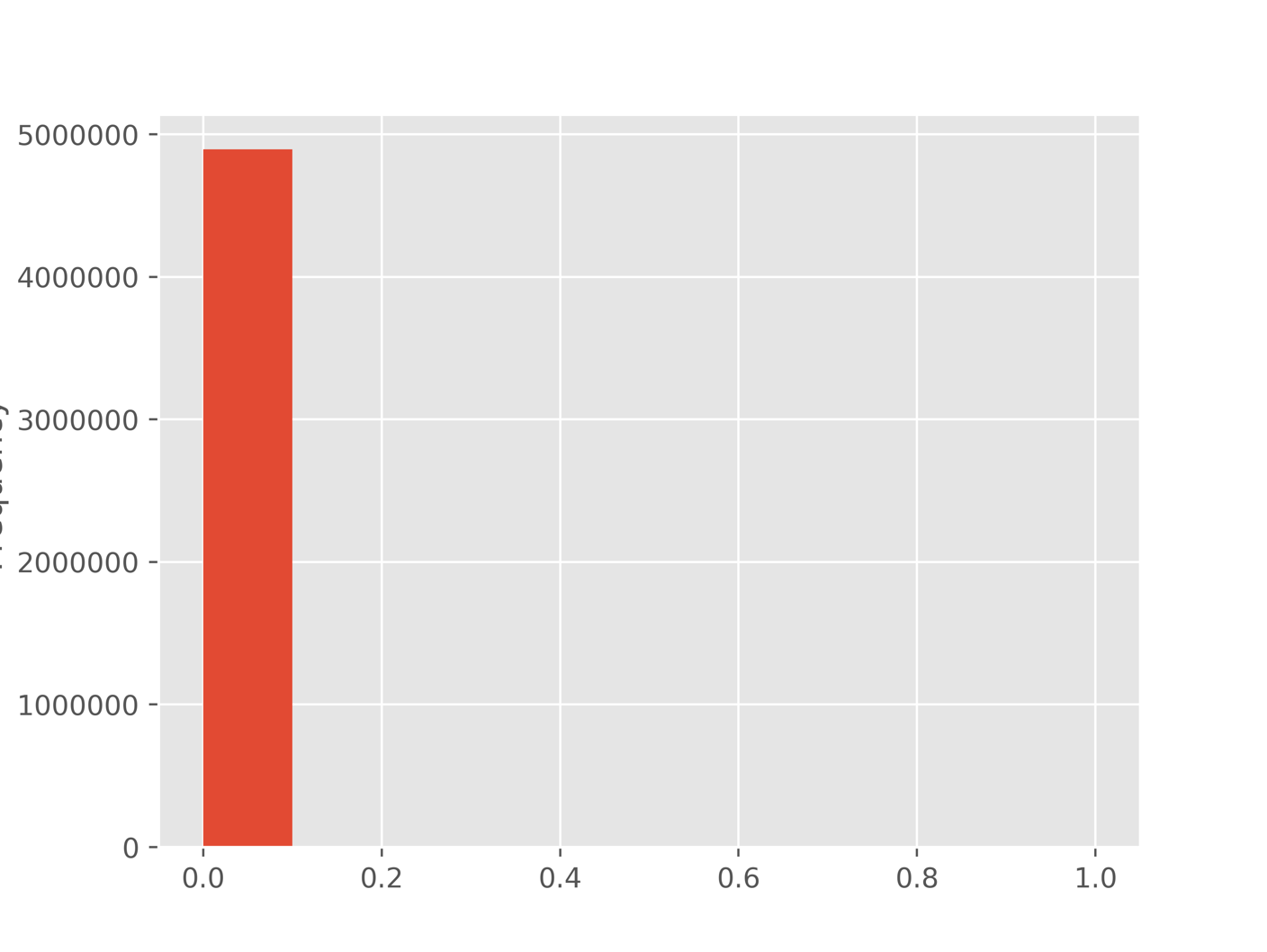
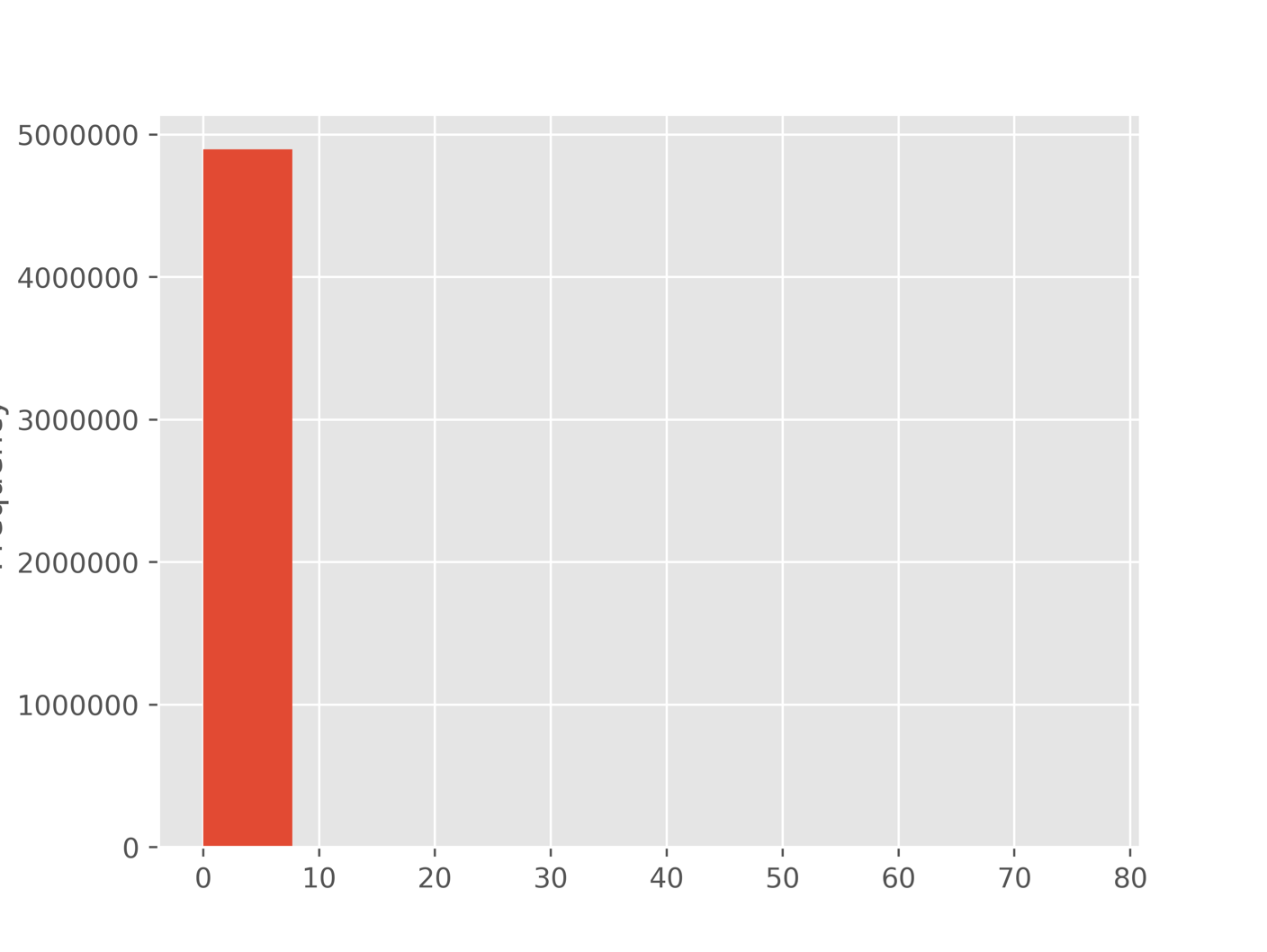
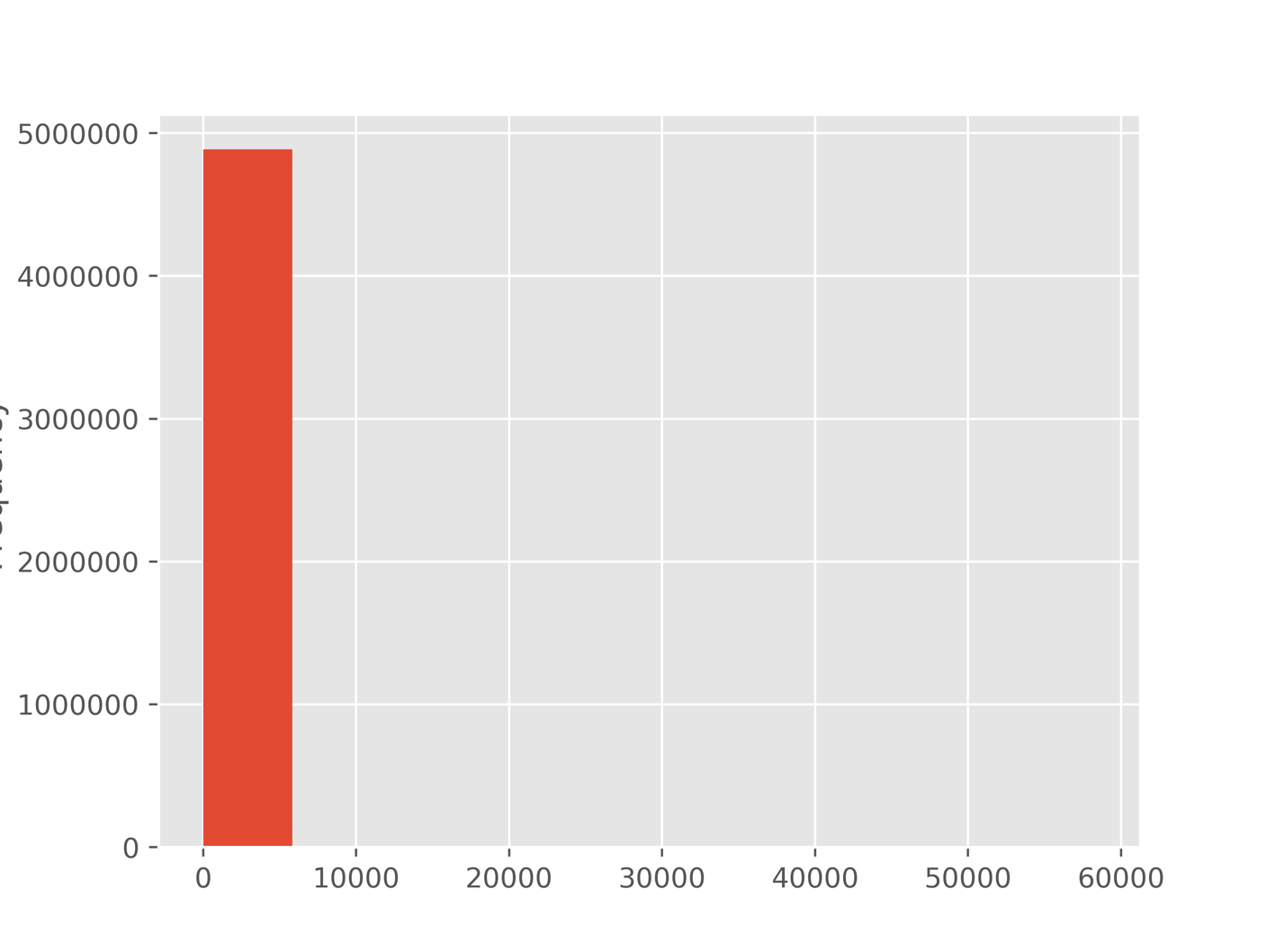
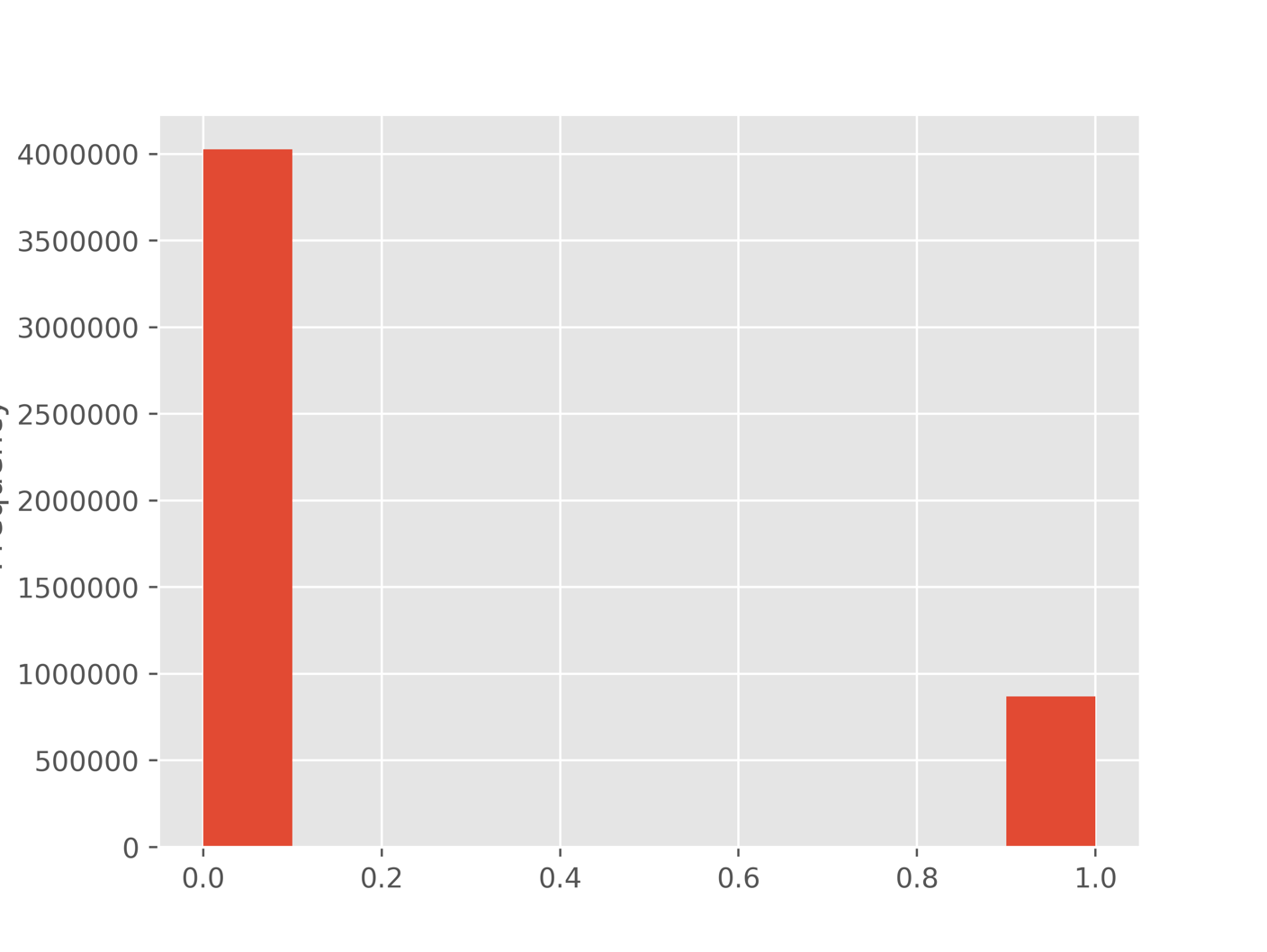
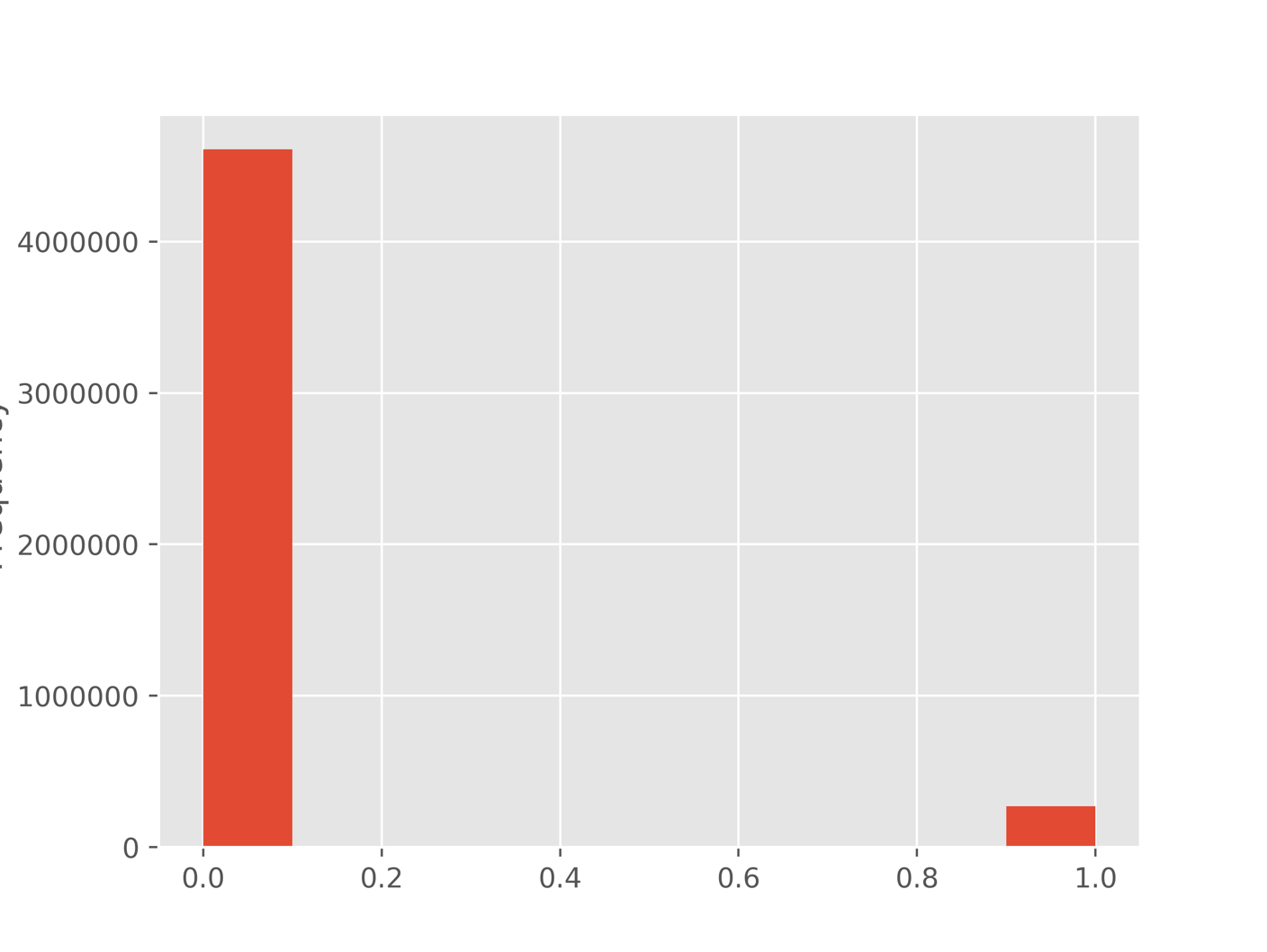
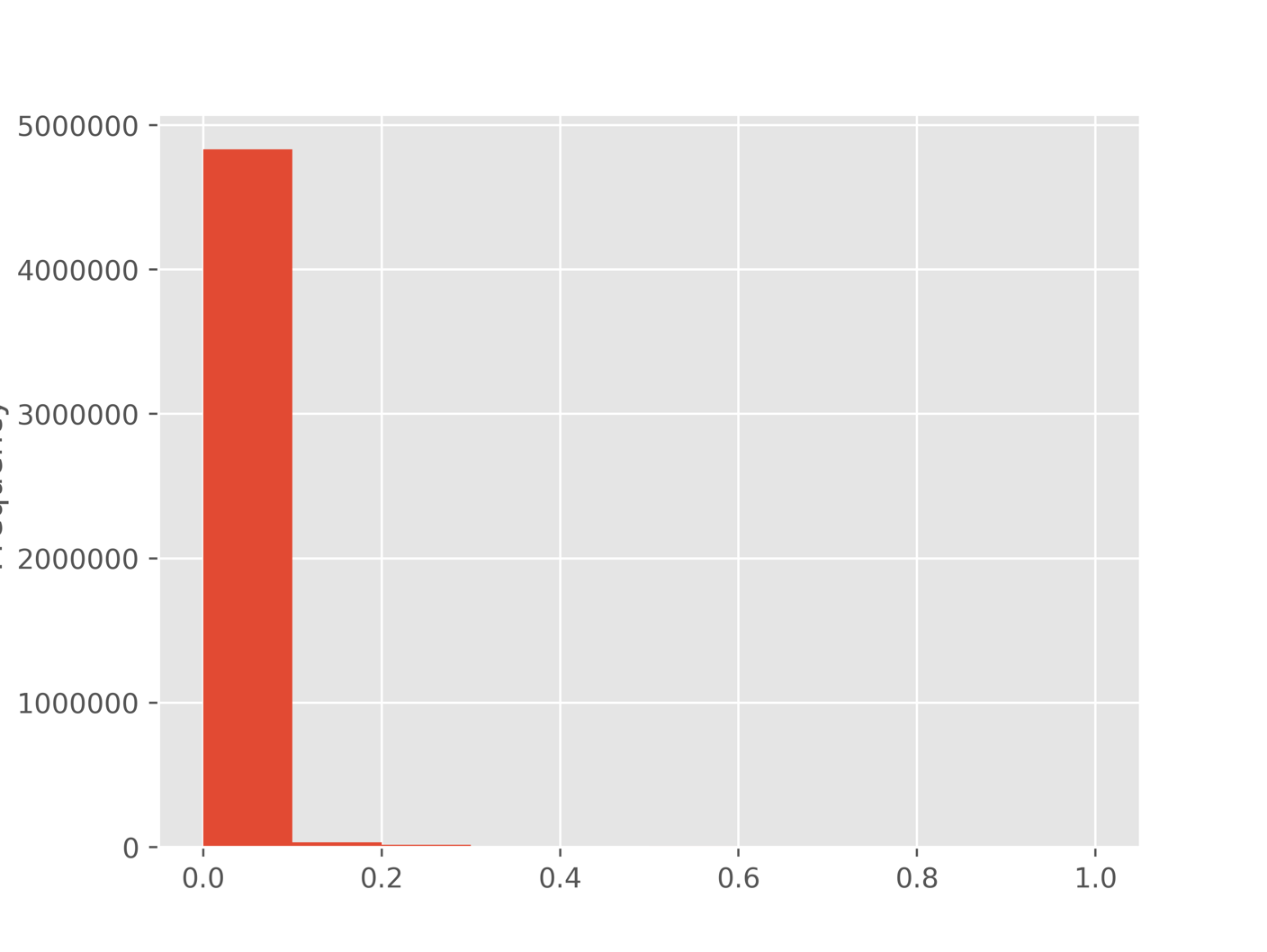
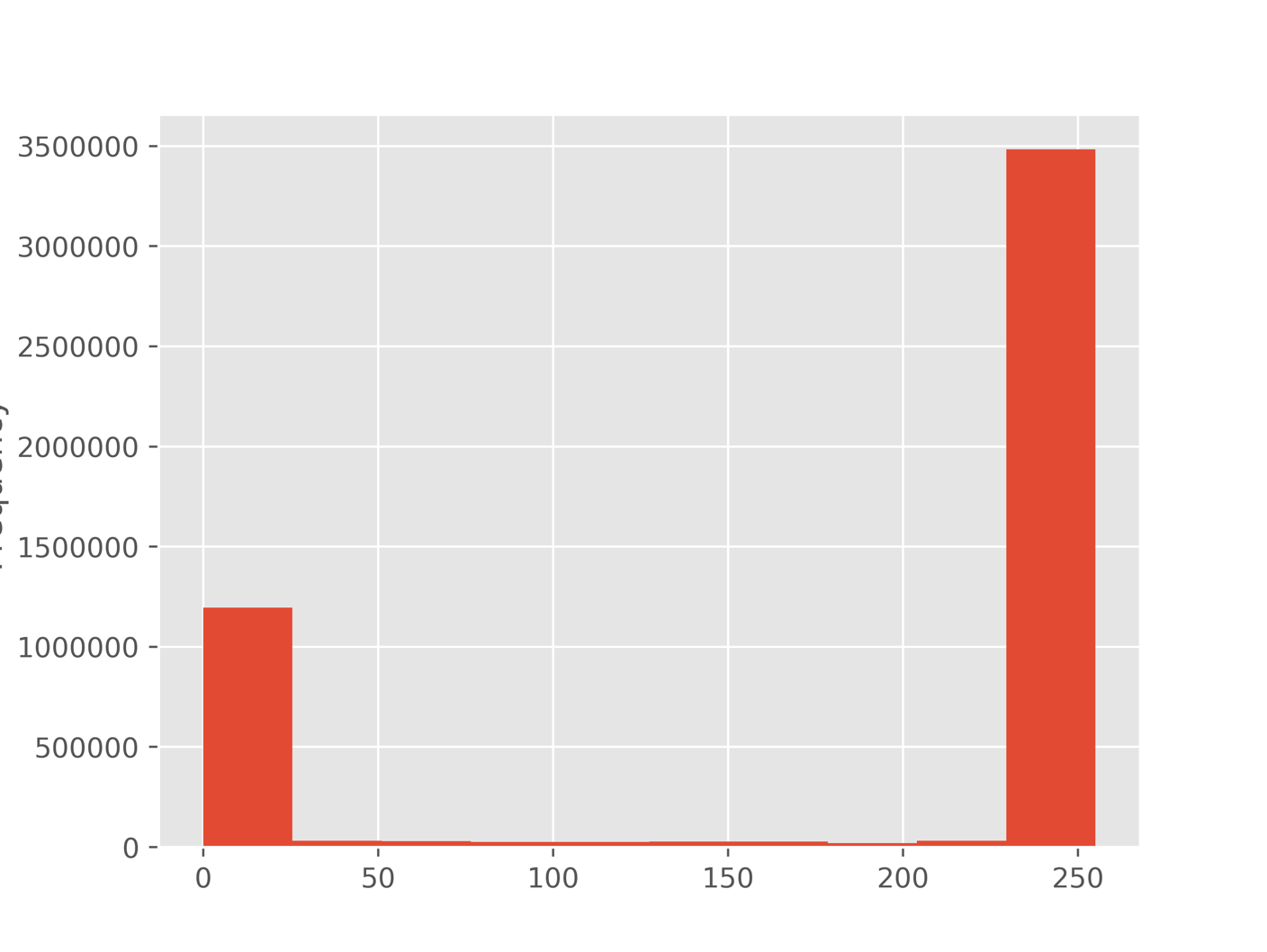
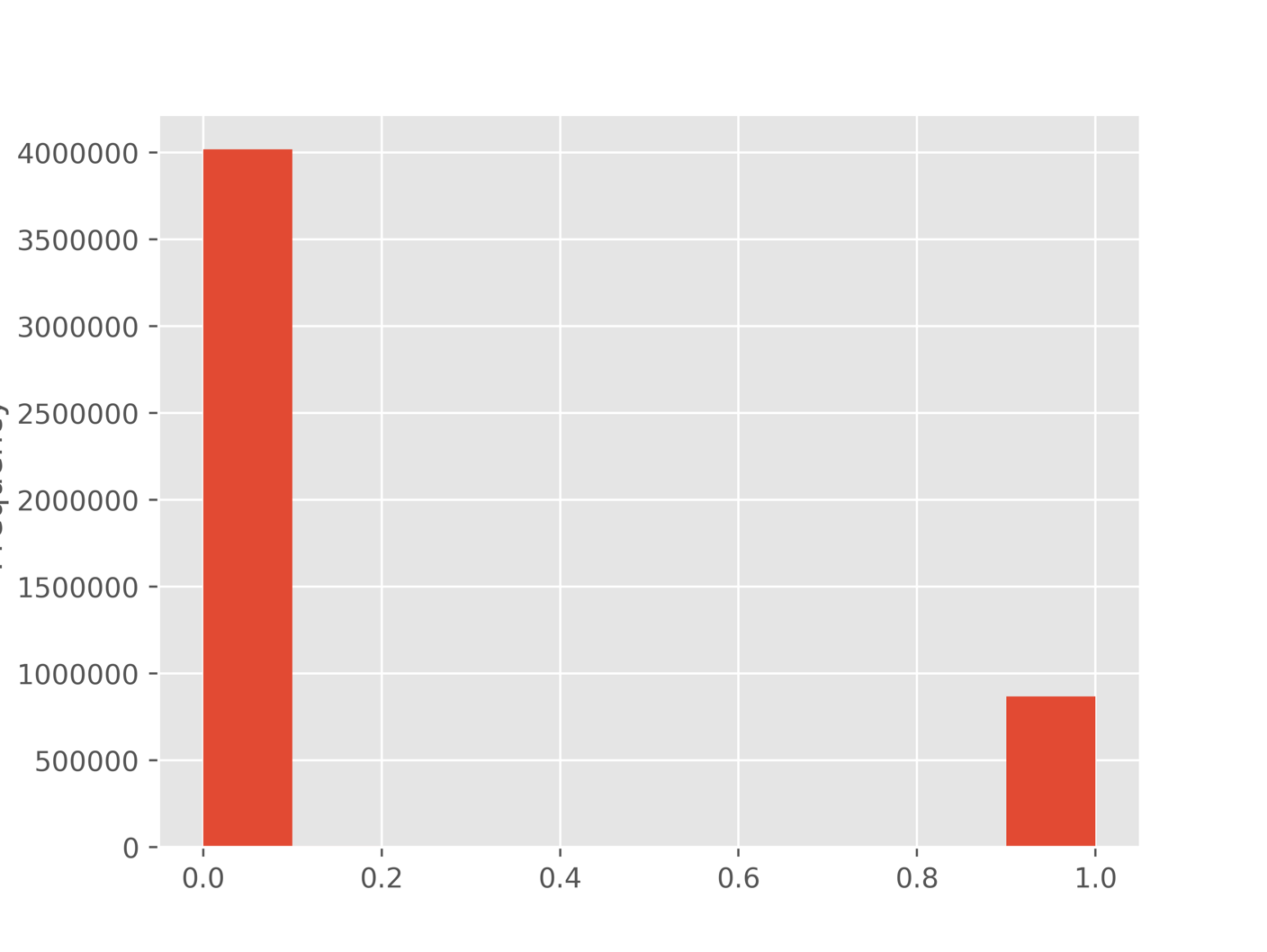
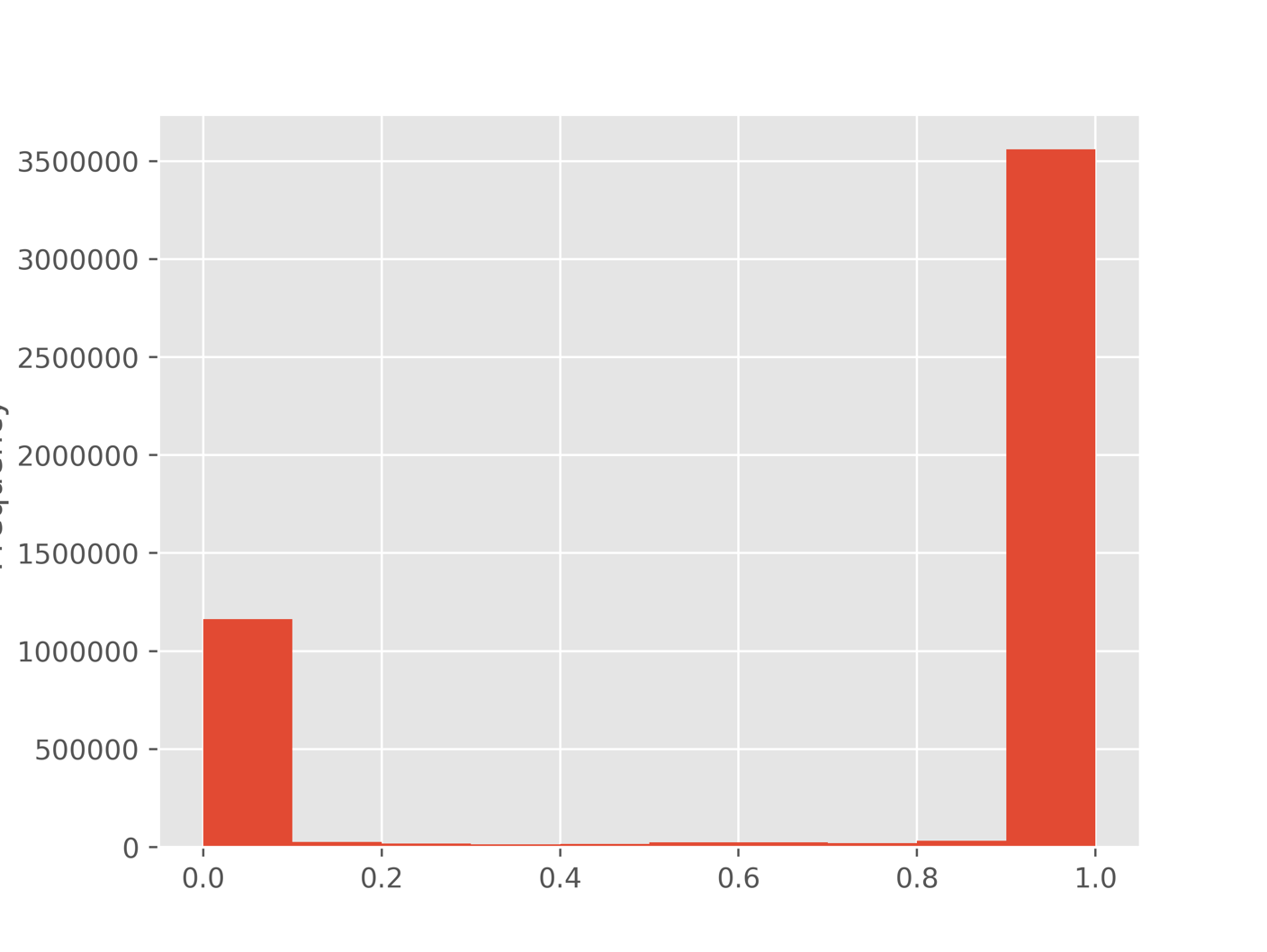
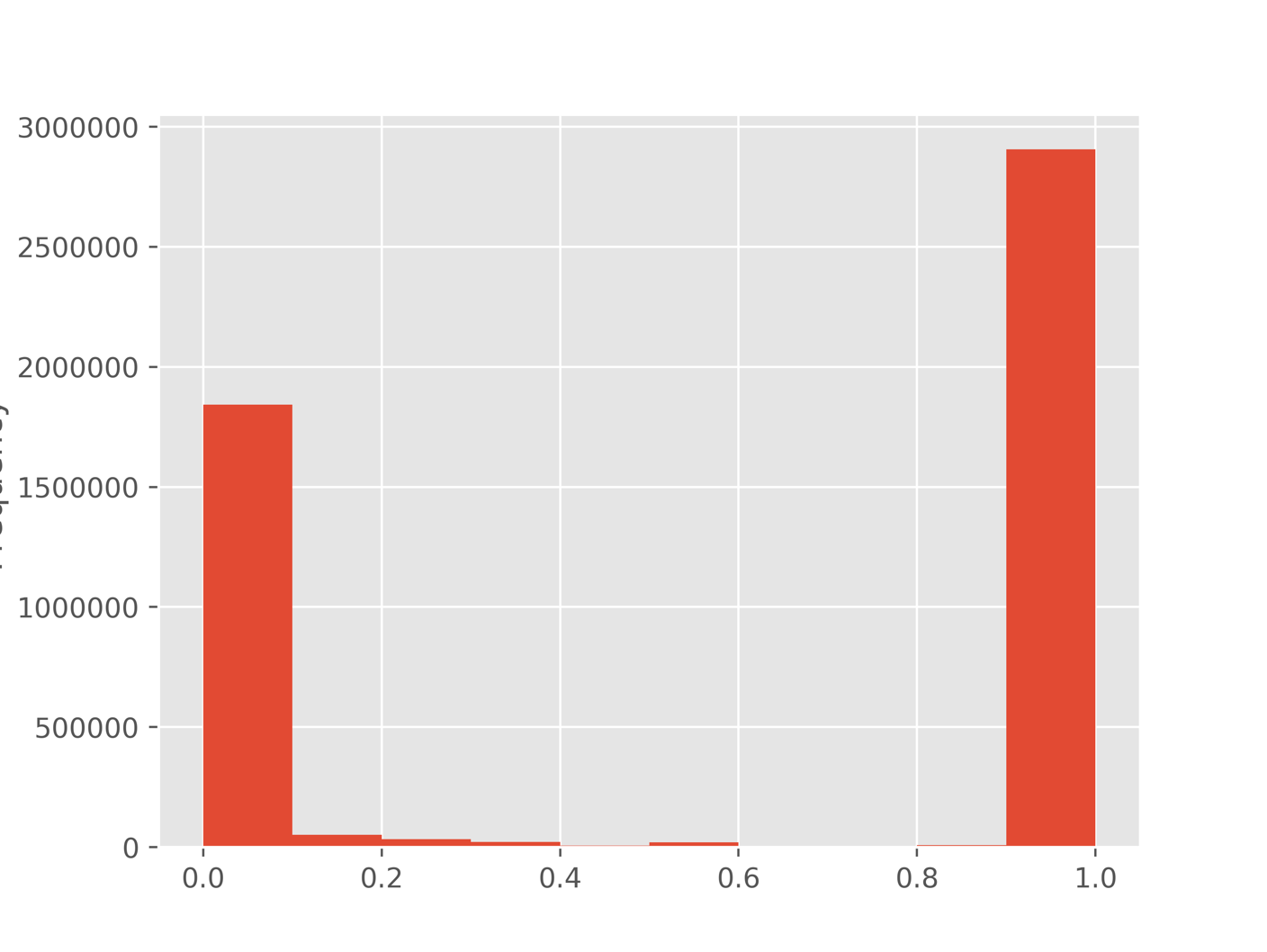
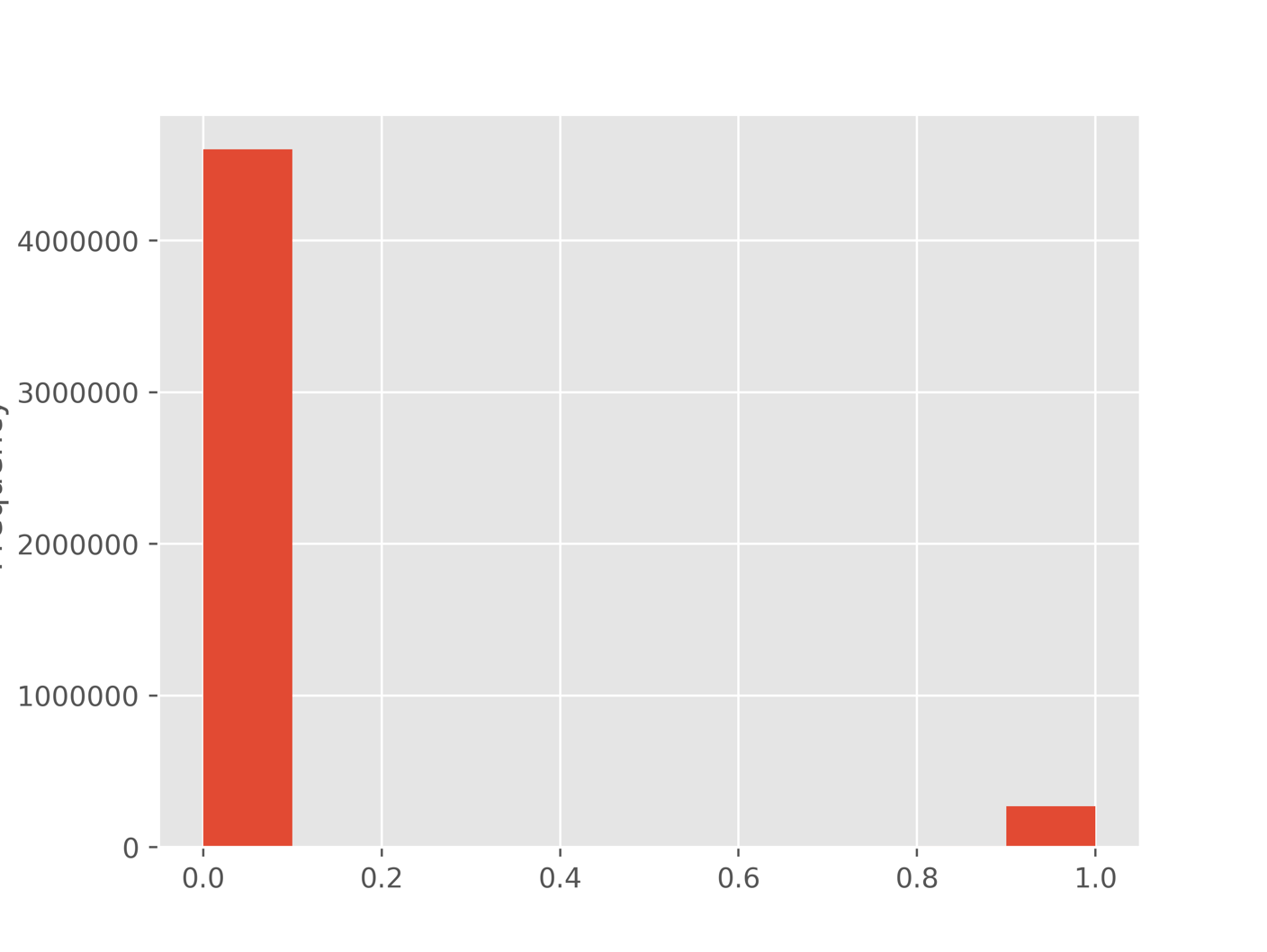
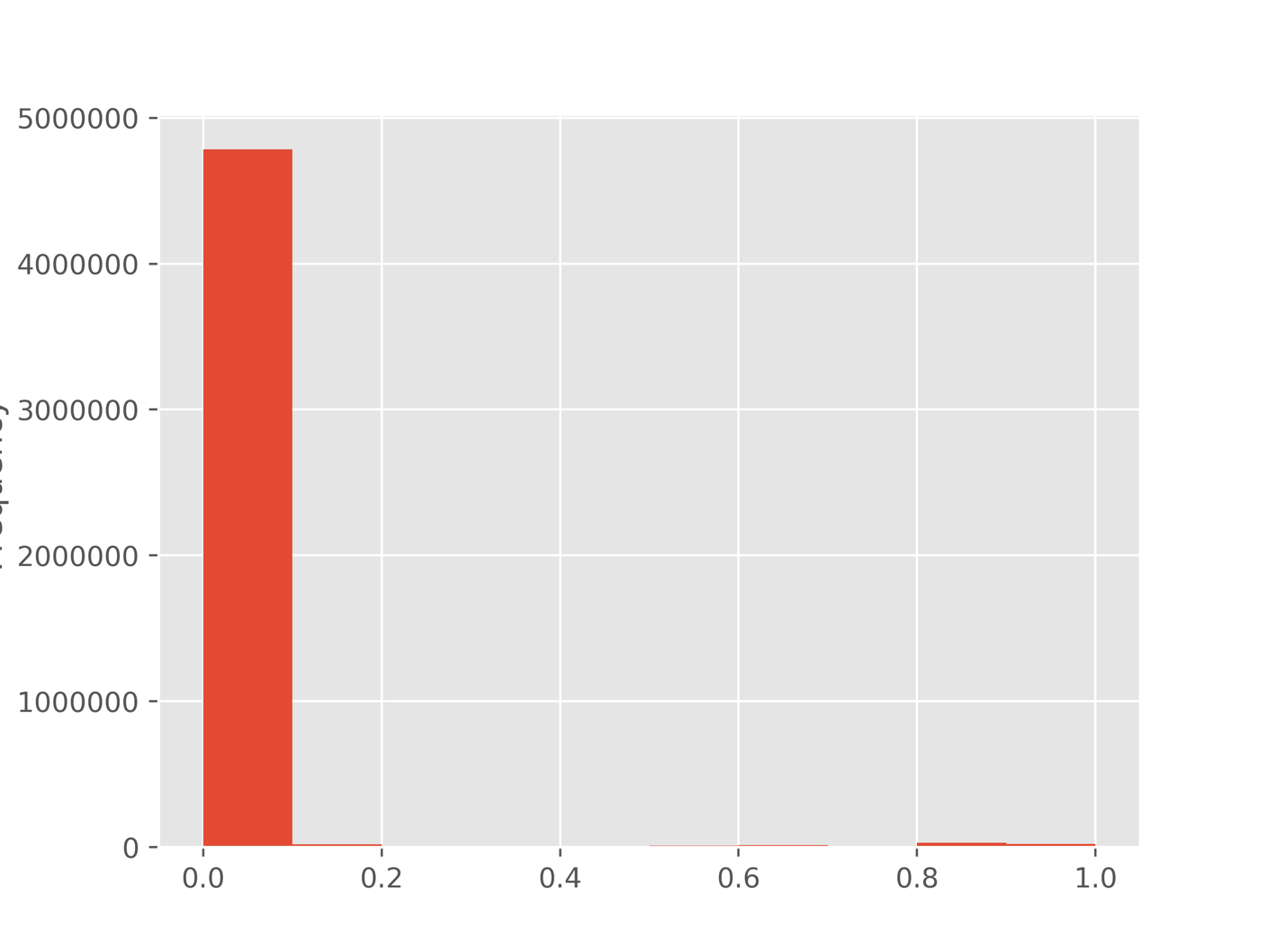
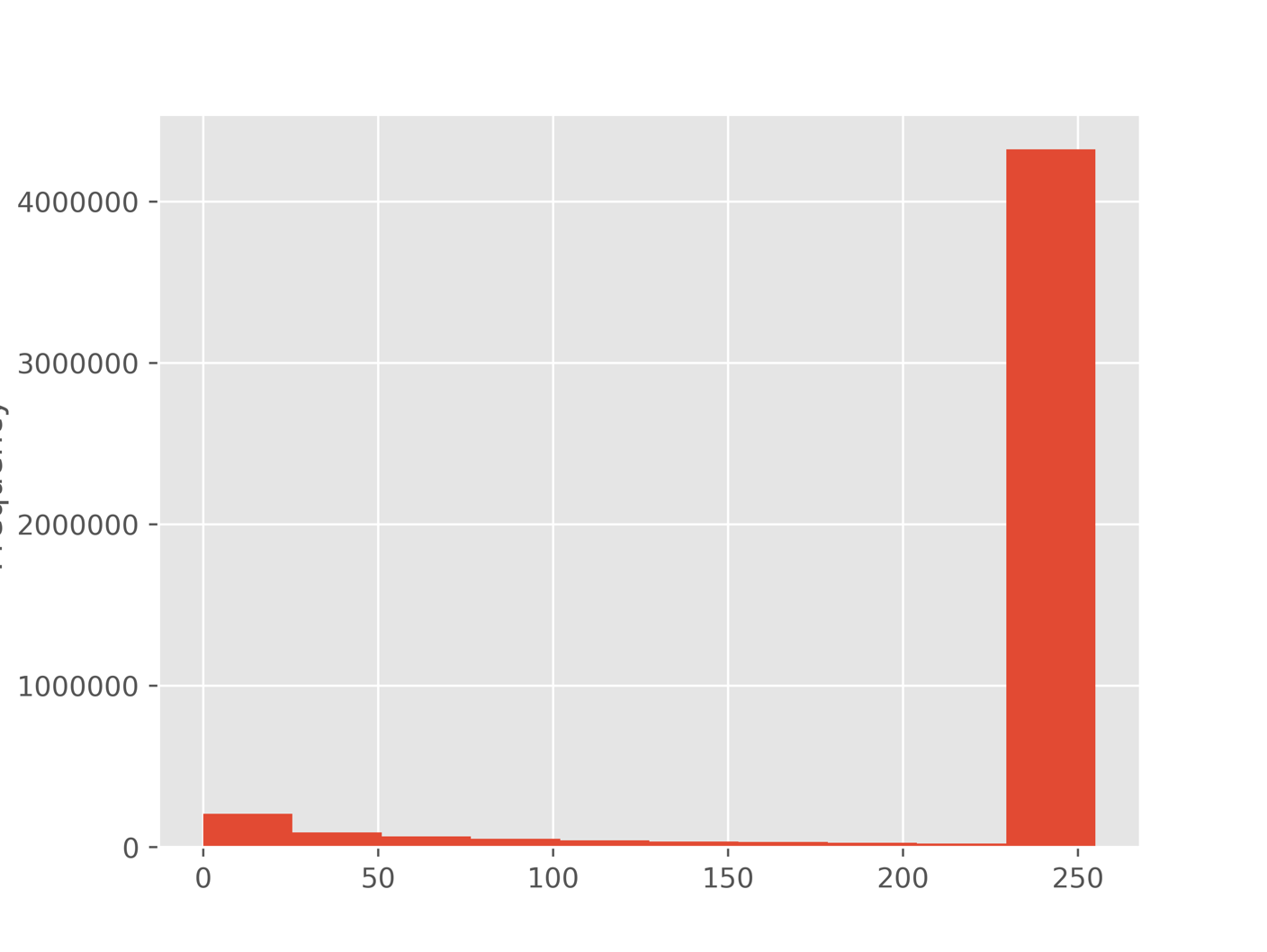
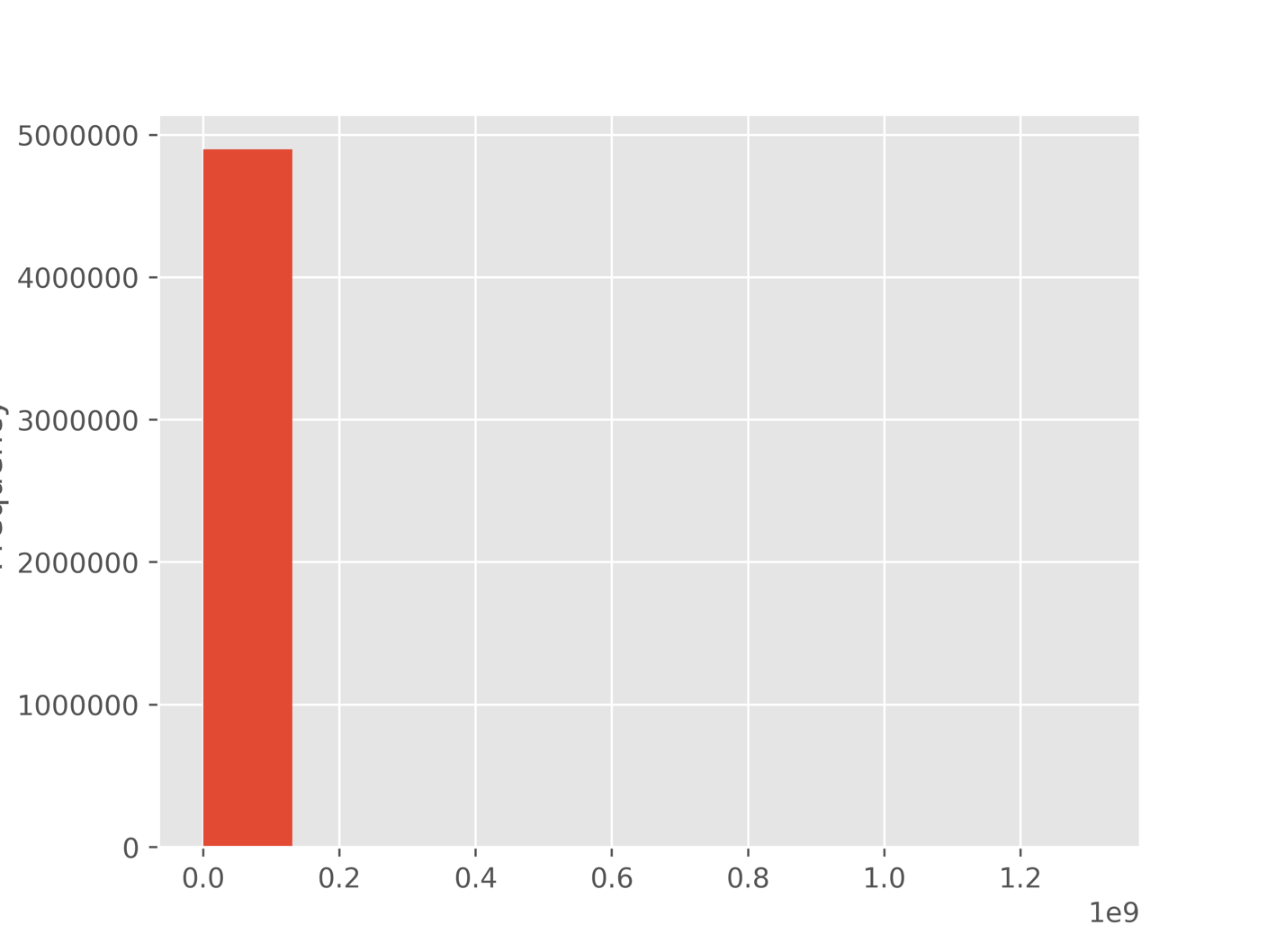
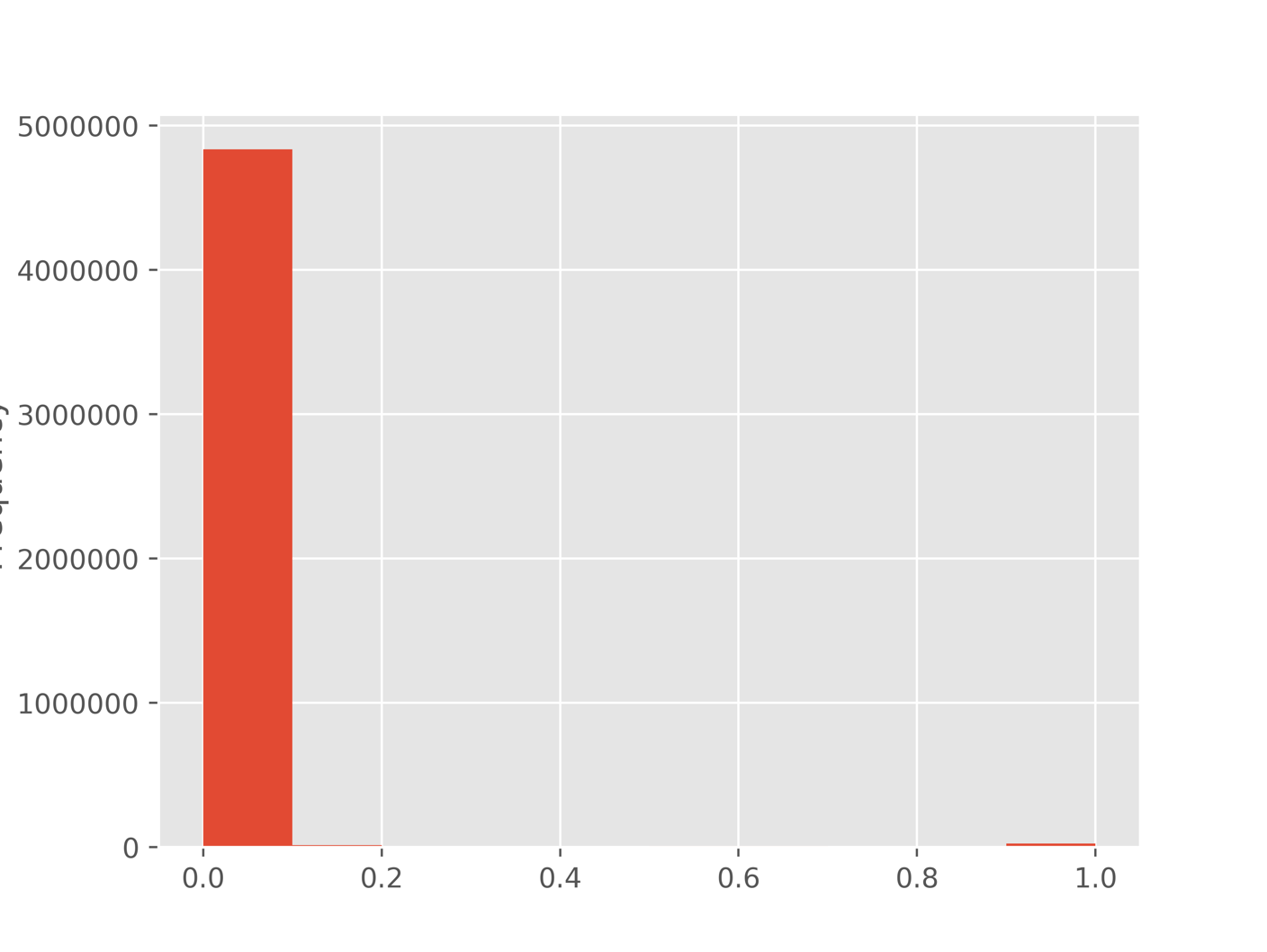
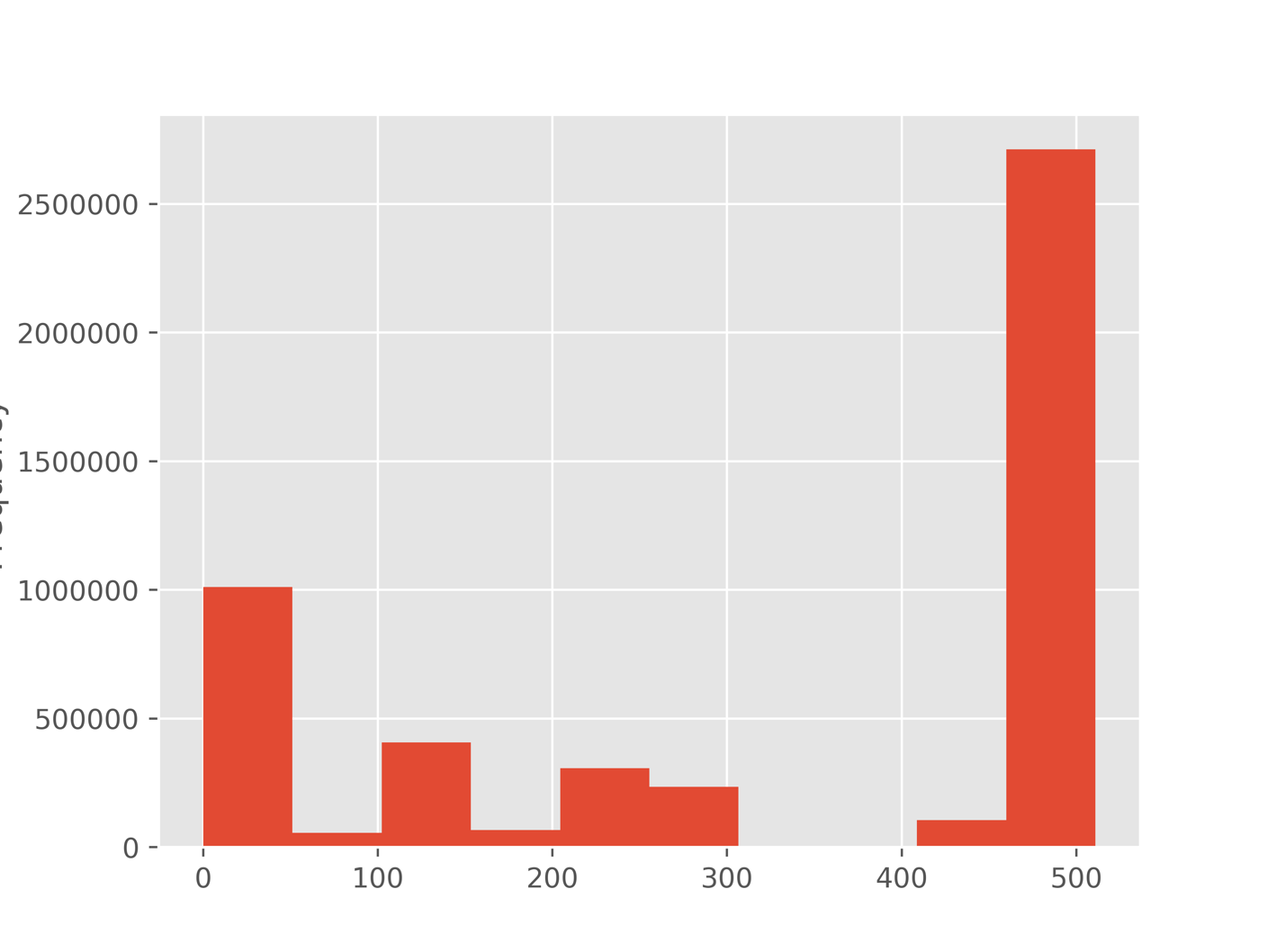
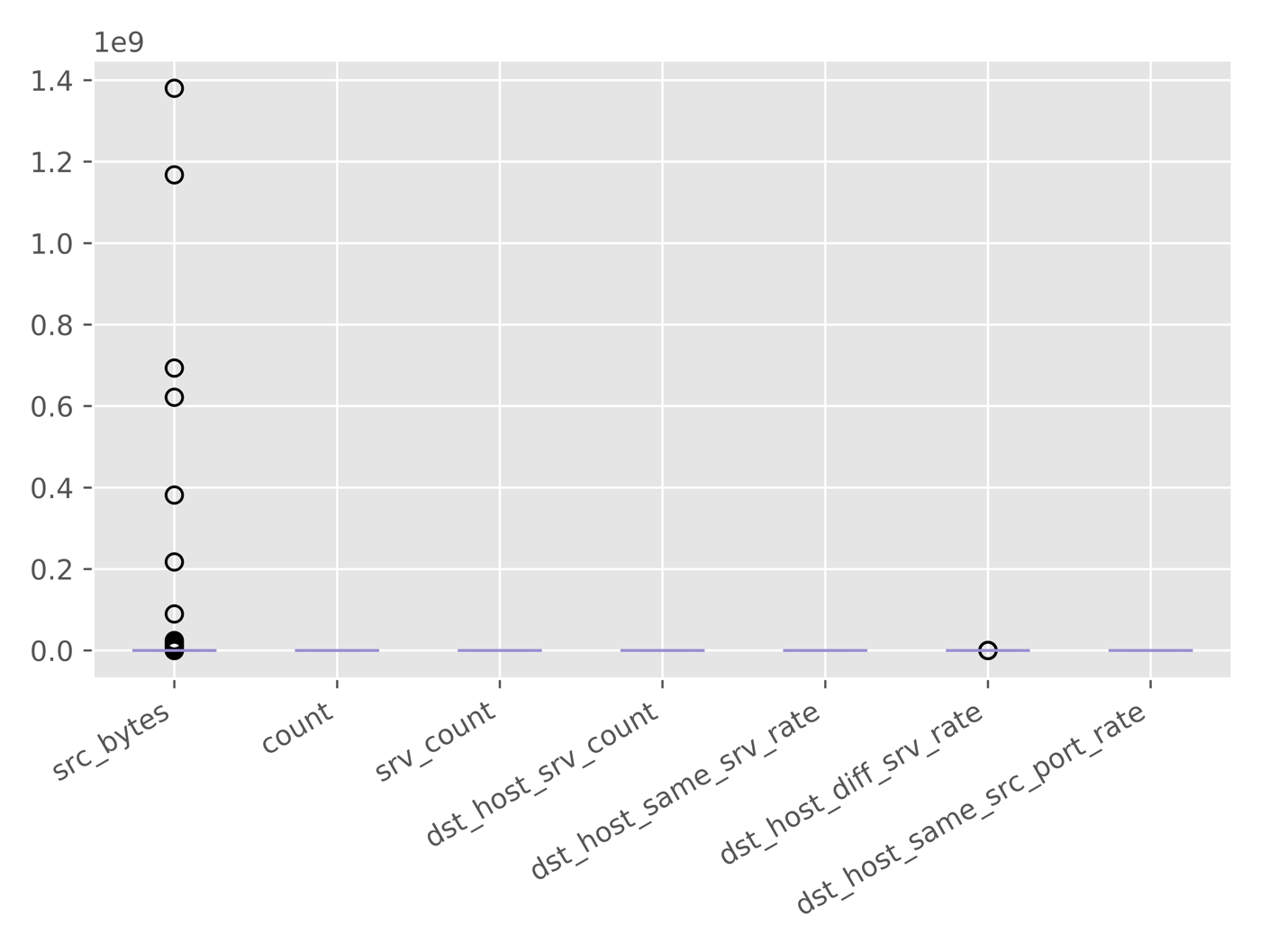
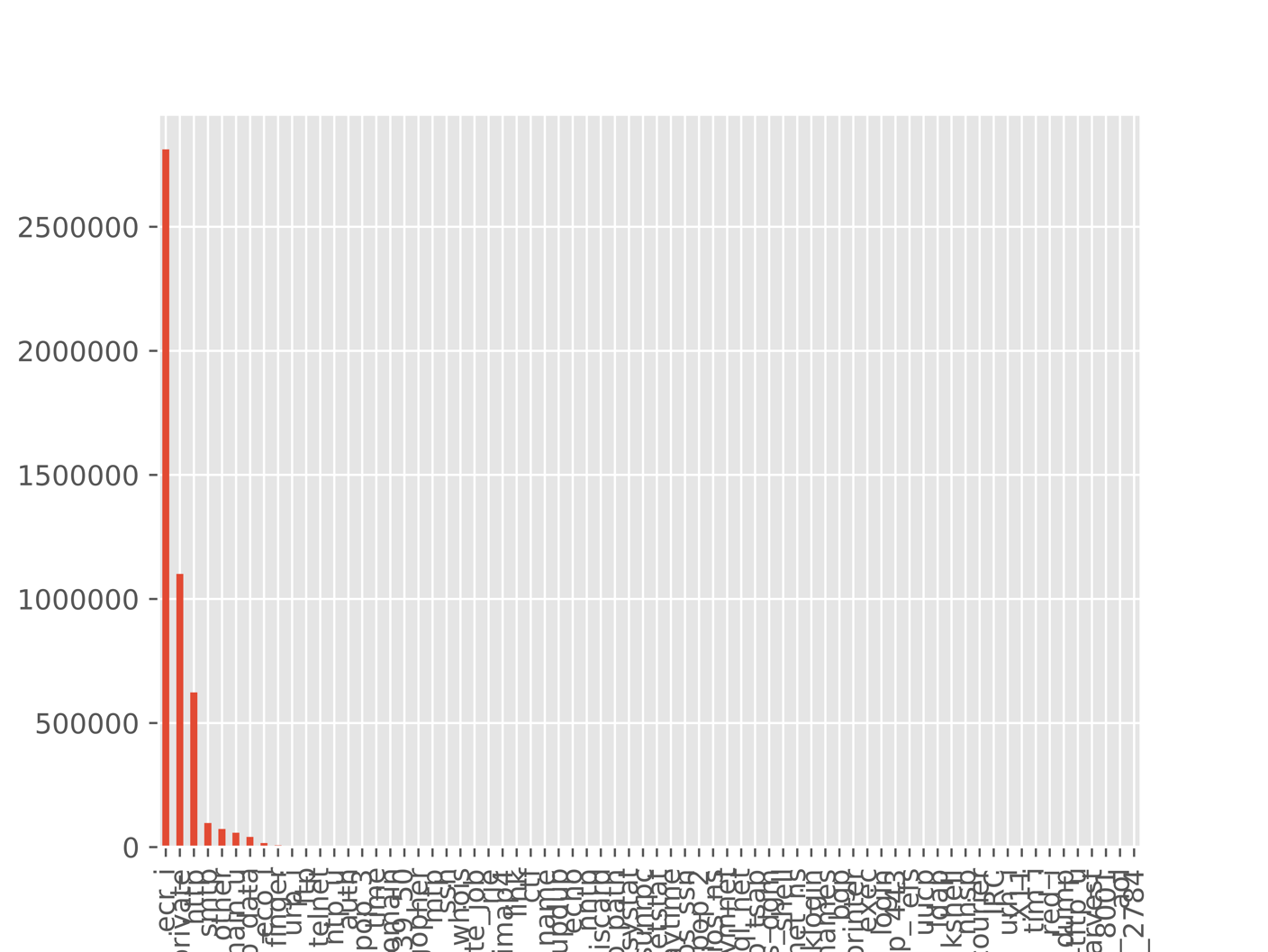
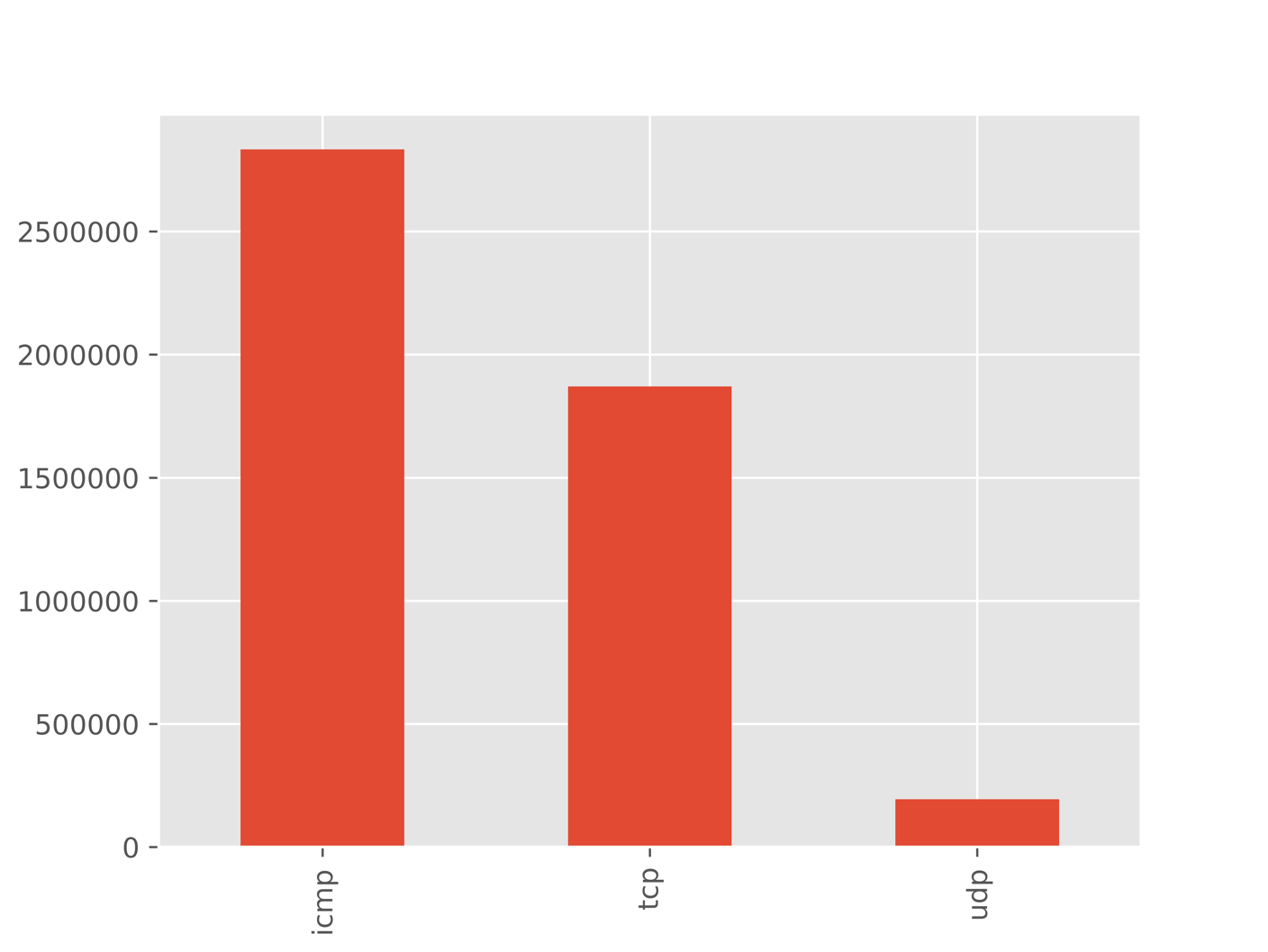
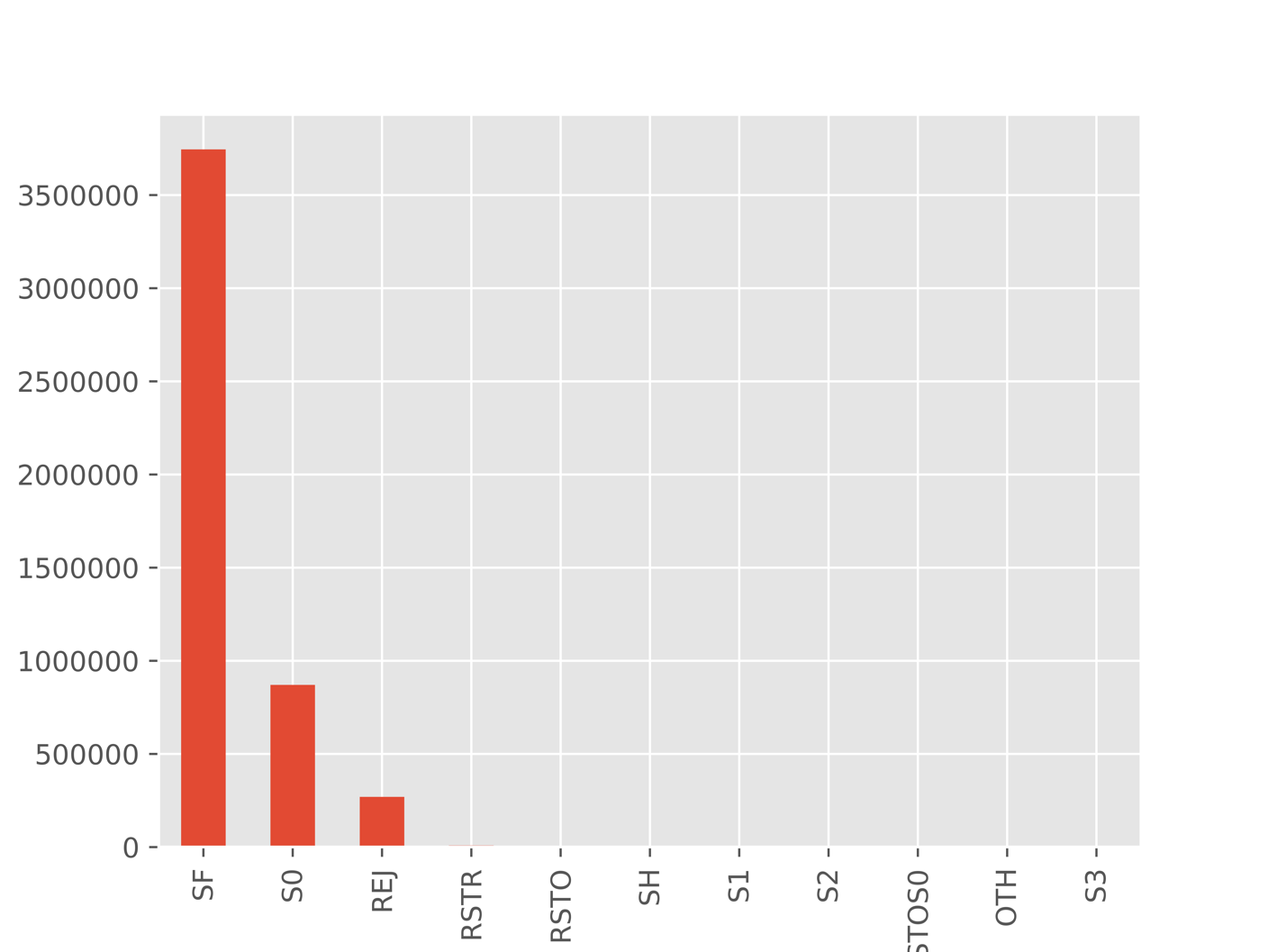
• Which pairs of attributes have a correlation coefficient less than -0.80? Why might these attributes be negatively correlated?

The bar charts that are more skewed tend to have coefficients less than -0.80.

• Are there any other interesting observations you can make about the data?

No, I think there aren’t any other interesting observations. But, I do think it is cool how a lot of data was skewed and one sided.





|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | count | unique values/cardinality | mode | mode frequency | % missing |
| protocol\_type | 4898431 | 3 | icmp | 2833545 | 0 |
| service | 4898431 | 70 | ecr\_i | 2811660 | 0 |
| flag | 4898431 | 11 | SF | 3744328 | 0 |



