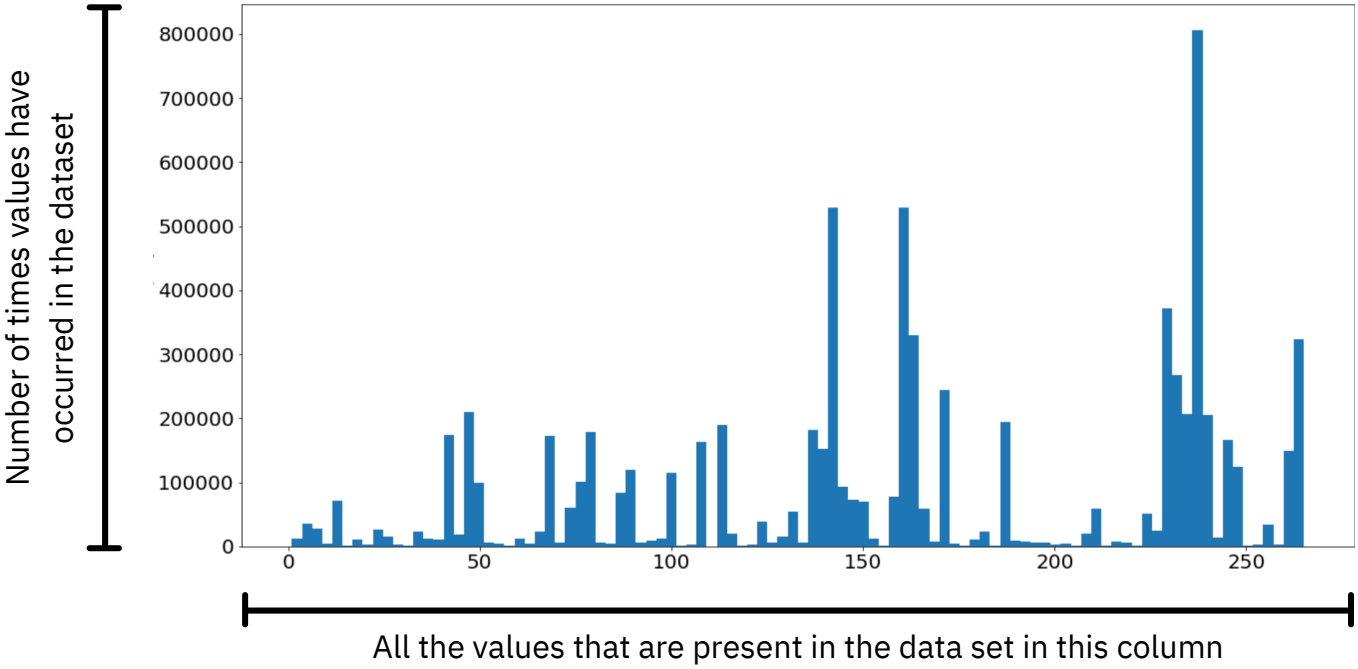
Reading histograms

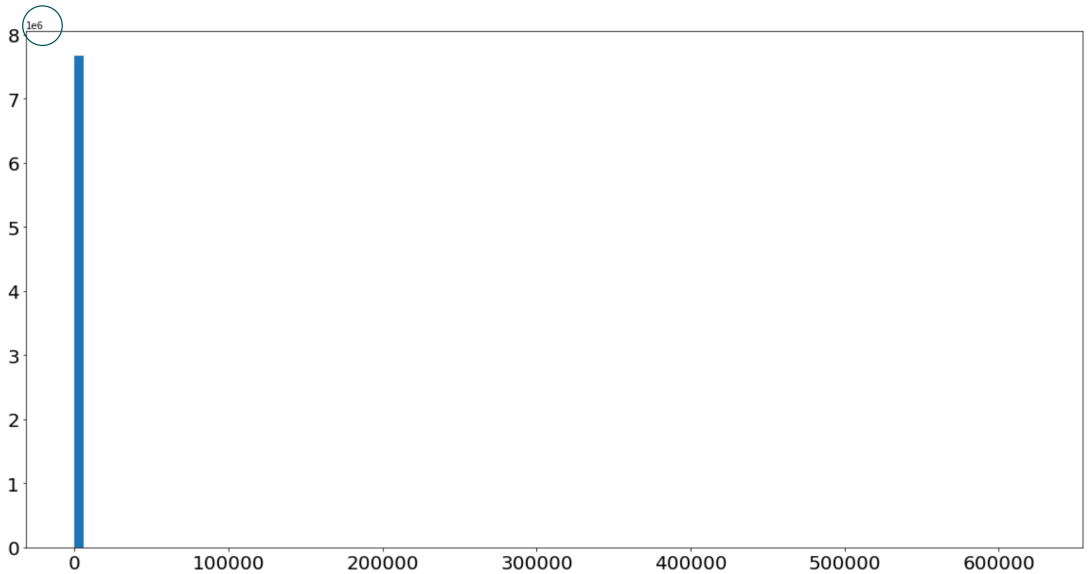
Histograms are extremely useful to understand your data. It can help you understand the distribution of the data but also alert you to potentials problems in the data. But first, you need to know what you're looking at.

Anatomy of a histogram



You can make a histogram for both categorical and numerical values. The histogram above shows us the distribution of a categorical feature. This specific feature is Location ID. So the numbers on the x-axis all correspond to a location.

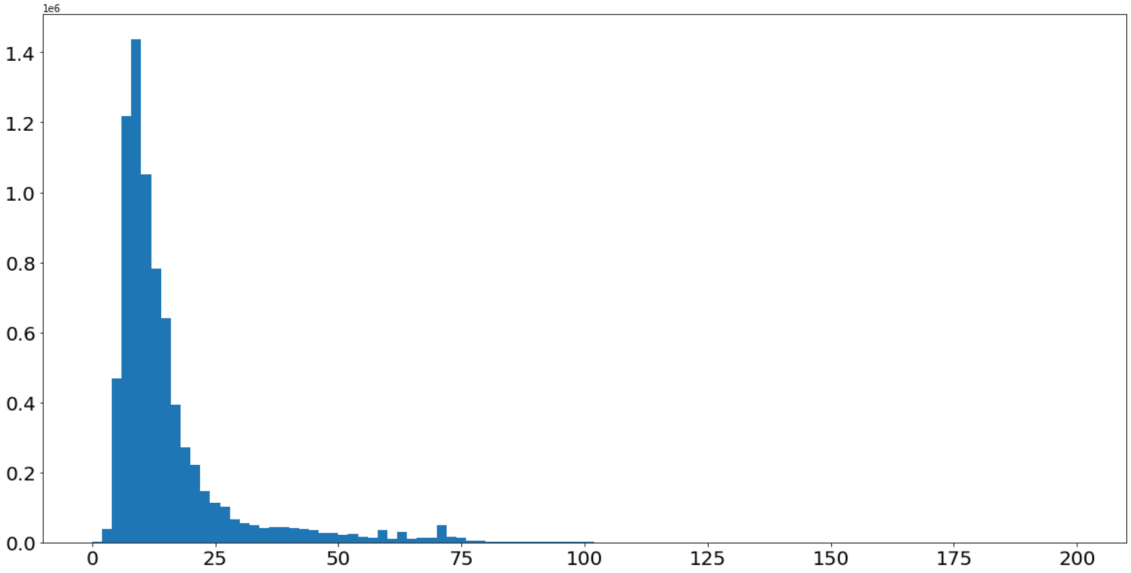
The histogram below illustrates the distribution of a numerical feature. When we make histograms for numerical features, they tend to look close to distribution plots. The feature we see here is the amount of money spent on a given taxi trip (total_amount).



Note: the y-axis numbers might look very low, but make sure you see that it is actually in the unit of 10 to the power of 6. So 1 is actually 1000000.

Finding problems using the histogram

The x-axis is not randomly generated. It spans from the lowest value in the dataset to the highest value. So this would mean, even though we can't really see it, in the previous plot, there is a total_amount value that is around 600000. This obviously was a mis-entered value. Or someone made a trip to the moon and back. Either way, we don't want that data point in our dataset as it is either mistaken or is an outlier. Here is how the histogram looks like after we get rid of the outliers/faulty values.



Make sure you realise the difference of the range on the x-axis. It now goes from 0 to 200.