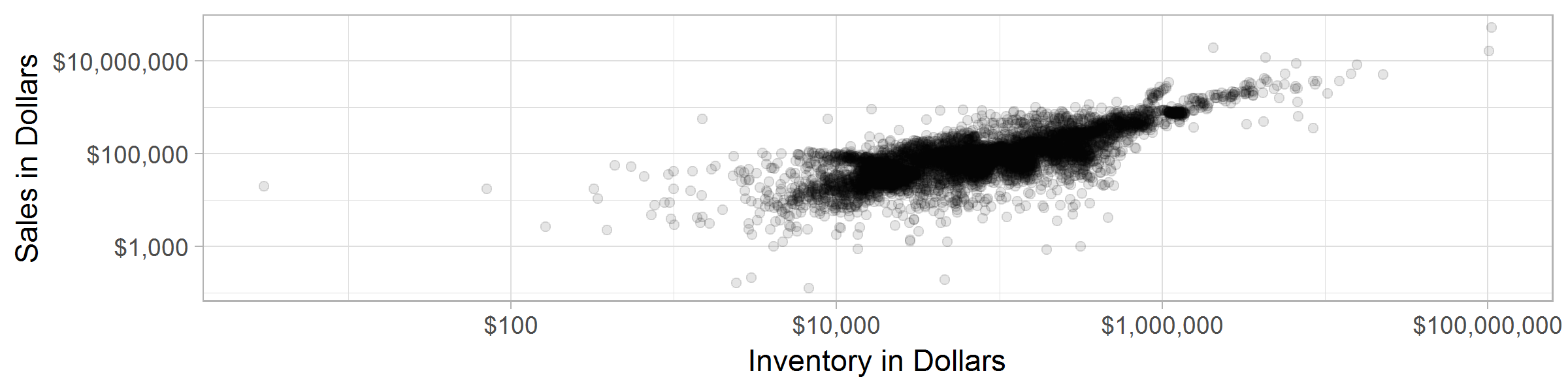
Description of Variables

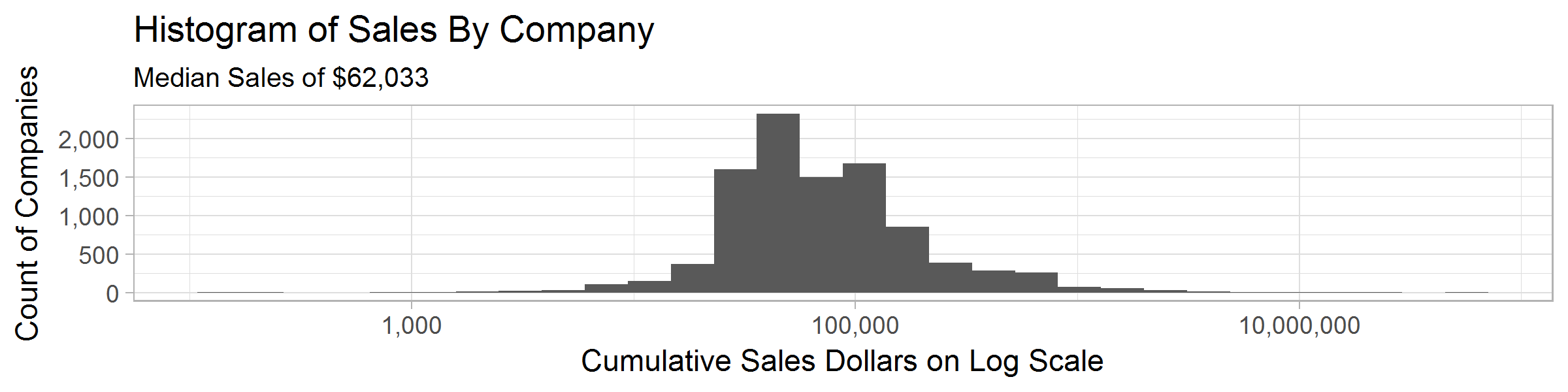
The data set contains the sales and inventory value for 9,762 companies. A scatterplot, and correlation analysis, both reveal a high level in interrelation between the two variables.



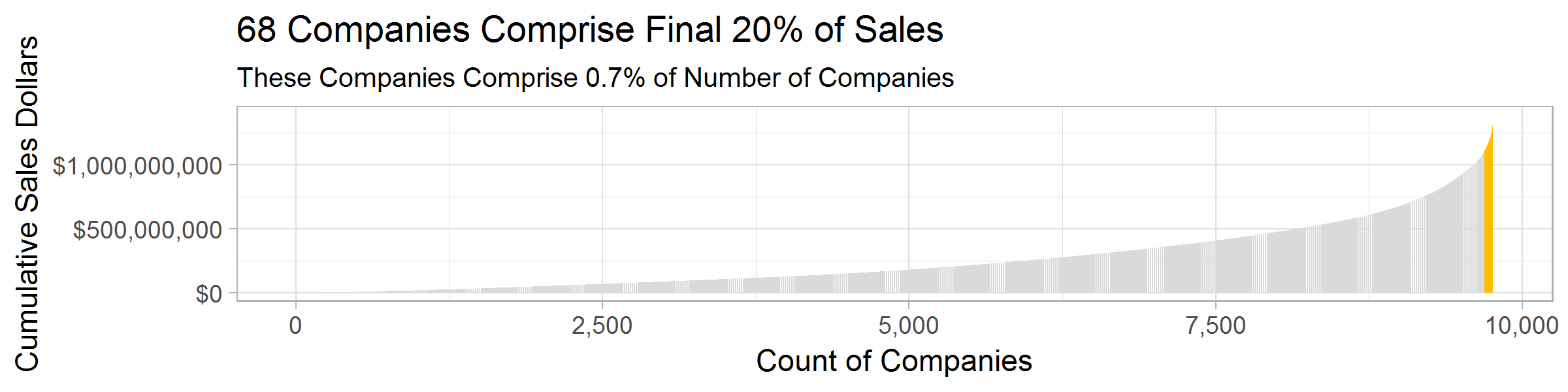
In fact, the correlation of the two variables is found to be 0.83. There is reason to believe that many of the observations would fall into similar strata no matter which variable is used, however, some of the examples will shift between strata.

Certainty Stratum – Methodology and Selection

Exploration of the data found useful information about the shape of the distribution of the data set for Sales by Company. Among the 9,762 companies included, the Mean Sales was $142,267, but with a much lower median - $62,033. This indicates a long tail distribution – with several outliers on the upper end of sales that are likely increasing the mean as well as the standard deviation of the sales value.



Considering this information, we investigated the possibility that a small number of the companies comprises a large share of the overall sales. In any scenario where the variables of interest – in this case, the total sales of the population – are of interest, a certainty stratum that comprises a large share of the overall value might allow a better insight into the final result. In a non-probability study, this could be done as a cut-off sample – but with the proposed survey design for this study, we propose that the companies providing the final 20% of cumulative sales to be taken as a certainty stratum.



This comprises 68 companies, but which sell 20% of the overall value of all companies in the population. This will allow for 432 of the 500 samples to be allocated to the remainder of the population, 9,693 companies. This gives an overall probability of selection for the remaining companies of 0.045. Overall probability of selection is 0.051 (500/9,762).

Stratum Creation and Break Points

With the certainty stratum determined, the next consideration is the creation of the stratum and the allocation of sampling units by stratum. The allocation method has already been determined: Neyman allocation, which will consider weight the stratum’s contribution to total variability of the sample, and then move more samples to the stratum with the largest contribution to variability. Based on our preliminary exploration, we would expect more of the allocation to be given to the upper end of the spectrum, as the large number of companies with sales less than the mean offer a smaller share of the overall variability of the population.

Considering the histogram of sales helps to determine some idea of the number of stratum required. It is clear that the long tail at the upper end of the sales spectrum is deserving of its own certainty stratum – and creating this certainty stratum provides for all companies with sales greater than $1.818MM to be included in the final sample. Considering the remainder of the companies in the histogram, there is a clear median value, but then three higher frequency bins to either side of the median. The lower tail and the space between the upper median range and the certainty stratum could be considered another group. In total, we might consider 8 stratum – including the certainty – but will test two different counts of stratum and determine the best method based on the expected variance produced.

Two methods of the determining the break points, and number of stratum, were considered. The first method considered was the Cumulative Method. This method considers the value of Sales, arranged from least to greatest, and then summed on a cumulative basis moving from least to greatest individual companies. Then this is broken into equal size strata.

Determination of Stratum

Selection of Variable