Exercise to explore, filter and summarize the distribution of numeric variables within a QC spreadsheet with graphical plots

We supply a datasheet, `data.csv`, that has some quality metrics from assemblies of bacterial genome sequencing. Please process this data to answer the questions below and upload the code and results to a public git repository such as github.com or gitlab.com.

Question 1

How many samples are there that have failed the contamination check (confindr.contam_status.check_result) and have contamination (confindr.percentage_contamination.metric_value) of over 5.0 (%)

Question 2

How many samples are there that have less than or equal to 50 contigs and a N50 value of greater than or equal to 750,000

Question 3

Select all numeric columns and rename them to remove the `.quast` prefix and `.metric_value` suffix, and rename `confindr.percentage_contamination` to `contamination_percent`

Question 4

Make a box plot of `Total length (>= 1000 bp)`

Question 5

Pivot the data so that it is 'tidy' with one observation per row and have final column headings of sample_name, metric, value

Question 6

Make a violin plot for each of the numeric variables in a single plot. Bonus: include jittered data points