1 newpage

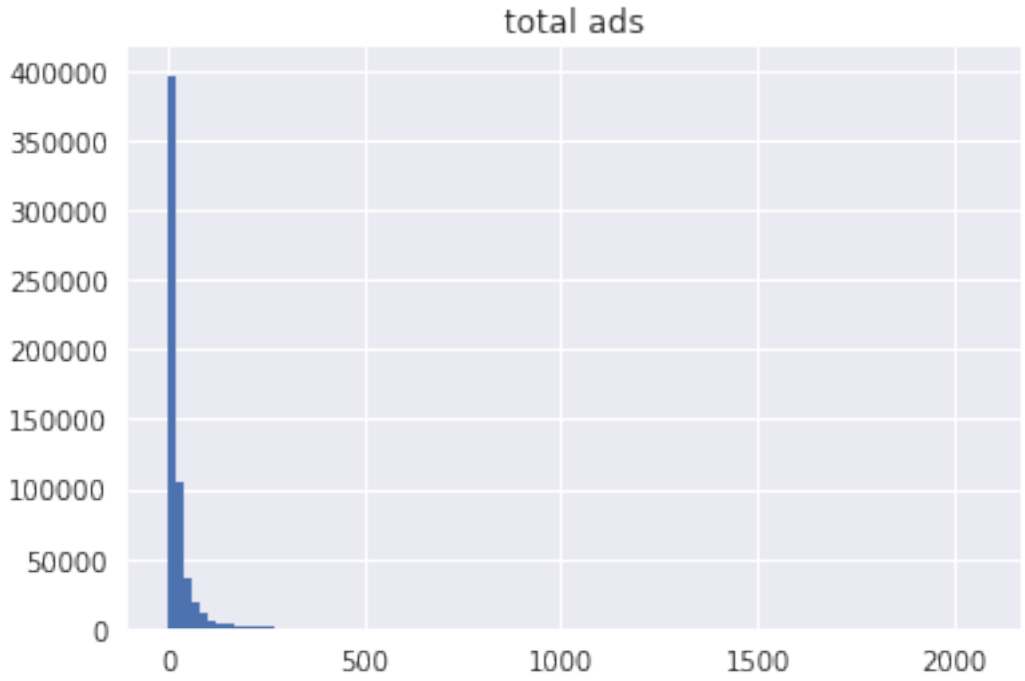
2 Data Science: Bridging Principles and Practice

2.1 Part 5: Numerical Data and Histograms [SOLUTIONS]

2.2 5a. Histograms

EXERCISE: Use `hist` to create the histogram for the "total ads" column in the `ads` table. Set the number of bins to 100.

```
In [4]: # create a histogram for the total ads column
        ads.hist("total ads", bins=100);
```



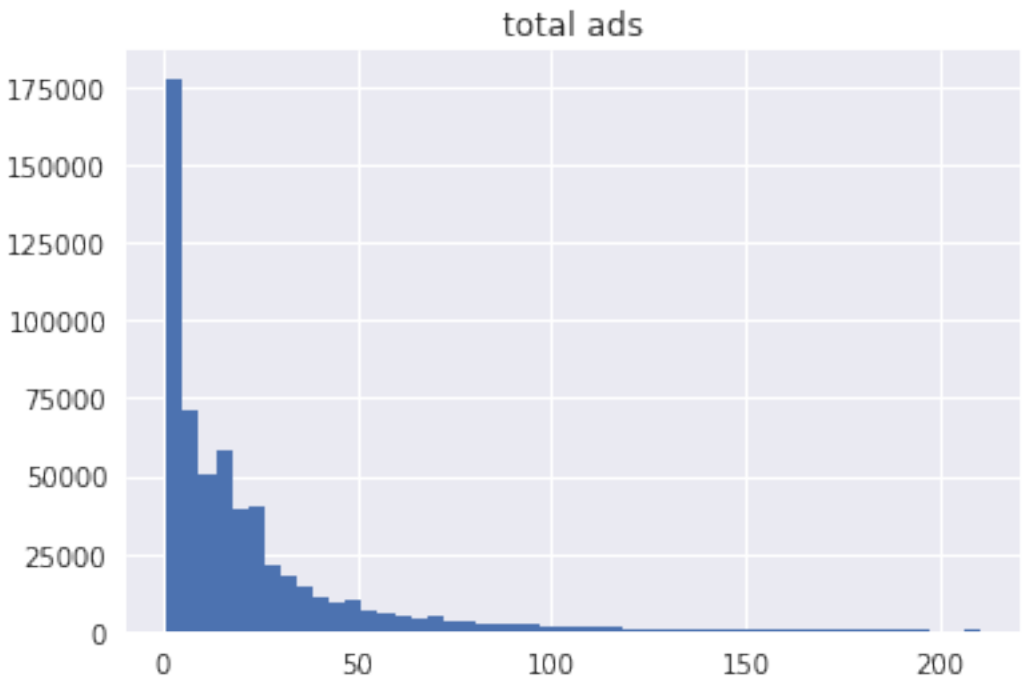
QUESTION: what does the histogram tell us about the distribution of "total ads"?

ANSWER: The distribution of "total ads" is *right-skewed*- a relatively small number of subjects saw between 250 and 2000 ads, which pulls the histogram out to the right. A large majority of subjects in the Rocket Fuel study saw less than 100 total ads.

2.3 5b. Consumer Response vs. Total Ads Seen

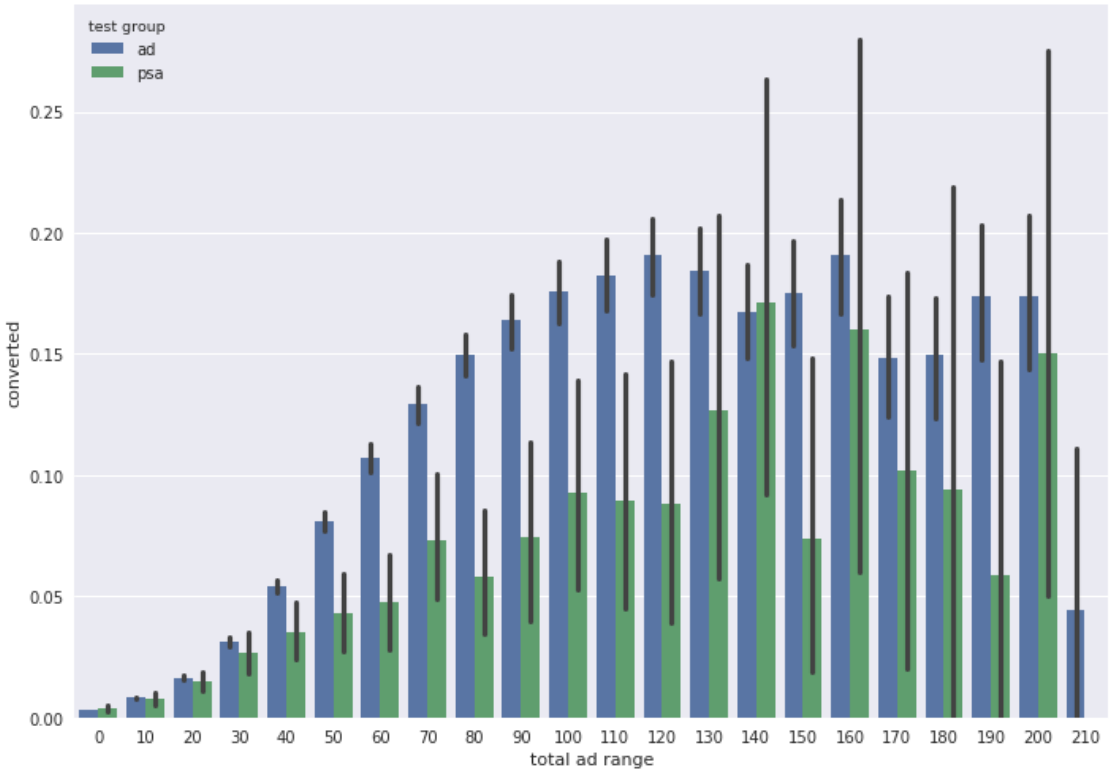
EXERCISE: Re-draw the histogram for the `total ads` column using the `ads_small` DataFrame. Set the number of bins equal to 50.

```
In [6]: # create a histogram for the distribution of total ads
        ads_small.hist("total ads", bins=50);
```



In [9]:
```
# make the plots bigger
sns.set(rc={'figure.figsize':(11.7,8.27)})

# make a bar plot of the conversion rates by ranges of ads seen
sns.barplot(x="total ad range", y="converted", hue="test group",
            data=ads_small);
```



QUESTION: What can you infer from the plot? For what range is advertising most effective?

ANSWER: As in the last notebook, we can define advertising efficacy as the difference in conversion rate between the control (PSA) and experimental (ad) groups. In this case, that difference is the difference in the heights of the adjacent green and blue bars.

The ad group had a greater rate than the PSA group if at least 10 total ads are seen, and we tend to see fairly similar, positive differences between about 80 and 120 total ads seen. For those ranges, the ad group always had a higher conversion rate than the PSA group, and the differences in conversion rates are all pretty close. After 130 total ads seen, the differences between the two groups vary a lot more (the conversion rate for the PSA group is actually a bit *better* than for the ad group at the 140 total ad range).

It's important here to keep our histogram in mind. A majority of people in the study saw 50 or fewer total ads. The small sample sizes for the larger total ad ranges mean that the differences we see may not be *statistically significant*. We'll talk more about determining statistical significance elsewhere in the course.

QUESTION: What do the above figures imply for the design of the next campaign assuming that consumer response would be similar?

ANSWER: This questions is up for some debate. Let's assume that the subjects in this campaign are *representative* of the target population of the next campaign (i.e. that they reacted and behaved similarly). If we're going by the first time we see the ad group rate exceed the PSA group rate, we would advise marketing to aim for at least 10 total ads seen, and that any additional ads they can show to that consumer (up to about 80 total) will result in an increased conversion rate. Any more ads than that, and the conversion rate difference compared to not seeing an ad at all either remains about the same or jumps up and down.

To fully answer this question, we need some more information: - how much does it cost to show a user more ads, and is that cost offset by the subsequent increase in conversion rate? - given the smaller sample sizes for groups above 25 total ads (see histogram), are the differences in rates statistically significant? That is, can we attribute the fact that consumers in the experimental group bought more handbags to the fact that they saw the ad, or was that difference due to random chance?

We'll explore these questions further in subsequent notebooks.