1. Determine if there are differences in NPS score by country.
    a. Are any differences statistically significant, and how did you decide?

There are significant differences in NPS scores by country. I ran an analysis of variance test (ANOVA), which gives us two numbers that describe the difference between groups (the F-statistic, and a p-value). Our F-statistic was large (14.69) and our p-value was very small ($p<2*10^{-16}$) which means it is extremely unlikely that the NPS scores would occur this way by chance. **The country has quite a lot to do with NPS scores.**

2. Is there a difference in NPS by number of jobs posted?
    a. Are any differences statistically significant, and how did you decide?

For this question I set up a simple linear regression. The idea is to calculate some numbers that describe how much the number of jobs a company has posted is related to their NPS. **The test indicated that there was essentially no relationship between the number of jobs posted and NPS** (R-squared = -0.0011, p = 0.9578).

3. Company leadership is interested in segmenting NPS data by company size, but we don't collect the size of the company. What segmentation strategy would you suggest based on the data available?
    a. Why did you choose that segmentation strategy?
    b. What would you suspect are 1-2 dangers from this approach, and how would you convey these to company leadership? Why is your strategy still viable despite those dangers?

I would anticipate that companies with more jobs listed would tend to be larger, so using the number of jobs posted as a natural proxy for company size. The issue with that is that as mentioned above, there is no real relationship between NPS and number of postings. This could potentially mean that NPS is unrelated to company size. It may also be true that larger companies are more likely to buy into certain visibility levels, though the two most popular visibility levels have nearly identical average NPS scores. **Combining visibility levels with the number of jobs posted is likely our best bet for finding a good predictor in these data.**

4. What strategies/methodologies would you use to extract the most common themes in verbatim responses from customers in the NPS survey?
    a. What are some themes in these responses? If you're able to execute some code to find some themes, great, but don't spend too much time here. It's more important to demonstrate your understanding of methodologies.

Sentiment analysis would be very useful for this kind of question. **Latent Dirichlet Allocation (LDA) can be used to automatically identify topics or themes in text.** This model tells us what words are statistically most likely to appear together in text. Using this information, we can get an idea of what kinds of comments are written most often. This model isn't perfect, and does still require human interpretation, but this process would be far superior to extracting themes from verbatim responses by hand. **Identifying themes could also be done by hand as a last resort**, depending on the size of the data set and time constraints. For a large data set a random sample could be hand coded to get an idea of some of the themes.

Some themes I see from reading a sample of the verbatim responses are: not getting enough applicants/candidates, feeling satisfies with service/praise of the service, candidates being underqualified, comments about the cost of the service