# EDA

Halid Kopanski

2022-11-21

## Exploratory Data Analysis

### Summary

For this study, we are presented with data from an 'Effervescent Experiment'. The data contains the dissolving times of two different brands of cold medicine tablets that were obtained under various conditions. Those conditions include varying water temperatures (6°, 23°, 406°) and the presence of stirring (magnetic stir bar at 350 rpm). This was a complete block design with stirring acting as the blocking effect. In all, the data contains 48 rows and 6 columns. The 6 columns include 3 explanatory variables (Brand, Temp, Stirred categorical factors), a single response variables (Time, a continuous variable), and 1 descriptor (sample order). Prior to starting any analysis, we will explore the data to gain an understanding of what to expect and to check for violations of any assumptions.

From the summary statistics table, we can see that each group has exactly 4 entries, eliminating concerns with respect to design imbalance.

| Brand | Temp | Stirred | Min | 25% | Mean | Median | 75% | Max | Range | Var | n |
|---|---|---|---|---|---|---|---|---|---|---|---|
| name | 6 | yes | 75.80973 | 75.83358 | 76.20241 | 75.89223 | 76.26107 | 77.21547 | 1.4057377 | 0.4593492 | 4 |
| name | 6 | no | 78.15246 | 78.79910 | 78.99061 | 79.04435 | 79.23586 | 79.72130 | 1.5688327 | 0.4146440 | 4 |
| name | 23 | yes | 69.08937 | 71.82180 | 72.69145 | 73.14894 | 74.01859 | 75.37855 | 6.2891789 | 6.9869087 | 4 |
| name | 23 | no | 76.06895 | 76.20492 | 76.36351 | 76.27622 | 76.43481 | 76.83265 | 0.7636940 | 0.1078134 | 4 |
| name | 40 | yes | 64.45156 | 64.87321 | 65.85343 | 65.43863 | 66.41886 | 68.08492 | 3.6333543 | 2.5499751 | 4 |
| name | 40 | no | 69.99943 | 70.28754 | 70.55511 | 70.50947 | 70.77705 | 71.20207 | 1.2026434 | 0.2544033 | 4 |
| store | 6 | yes | 76.24402 | 77.06561 | 77.33703 | 77.60659 | 77.87801 | 77.89089 | 1.6468708 | 0.5964884 | 4 |
| store | 6 | no | 77.78345 | 79.01994 | 79.49240 | 79.63219 | 80.10465 | 80.92176 | 3.1383169 | 1.6942517 | 4 |
| store | 23 | yes | 65.92809 | 66.08831 | 66.19126 | 66.22629 | 66.32923 | 66.38436 | 0.4562787 | 0.0411024 | 4 |
| store | 23 | no | 67.08353 | 67.14393 | 67.51552 | 67.52360 | 67.89520 | 67.93138 | 0.8478521 | 0.2060739 | 4 |
| store | 40 | yes | 58.24407 | 58.90895 | 59.12529 | 59.21659 | 59.43293 | 59.82388 | 1.5798100 | 0.4320148 | 4 |
| store | 40 | no | 58.53920 | 58.76884 | 58.96347 | 58.99050 | 59.18513 | 59.33370 | 0.7945066 | 0.1202191 | 4 |

There are insights to extract from this table. We can see that there does appear to be a disparity between the mean dissolving times of store brand cold medications when compared to name brand. Store brand cold medicines dissolve in a shorter amount of time as a whole over all effects than does the name brand medicine. This disparity becomes even more pronounced as the effect of temperature is introduced and as the temperature increases. When the brand type is store the means between each of the same temperature effects are much closer in similarity than the between temperature of name brand medicines. The differences between means store brand medicines of 6°, 23° and 40° are 2.15, 1.32 and 0.162 respectively. The differences between means name brand medicines of 6°, 23° and 40° are 2.79, 3.67, and 4.70 respectively. This may indicate an interaction effect on dissolving times by brand as the average difference in dissolving times is 1.21 for store brand and 3.72 for name brand.

Next, consider the range in variability at each factor and level. The range in values of the variance is 6.9458; an interesting result considering nine of the twelve observable variances fall within a range of 0.04 and 1.69. The variability between observations of name brand cold medicines are elevated, especially in cases when the observation was stirred at 23° and 40°. Variance of non-stirred observations, particularly within temperature values of 23° and 40°, are noticeably lower than their stirred counterparts within the same brand. The variance seems to jump by a large amount between the groups. Contrast analysis might be a concern due to the small sample size.

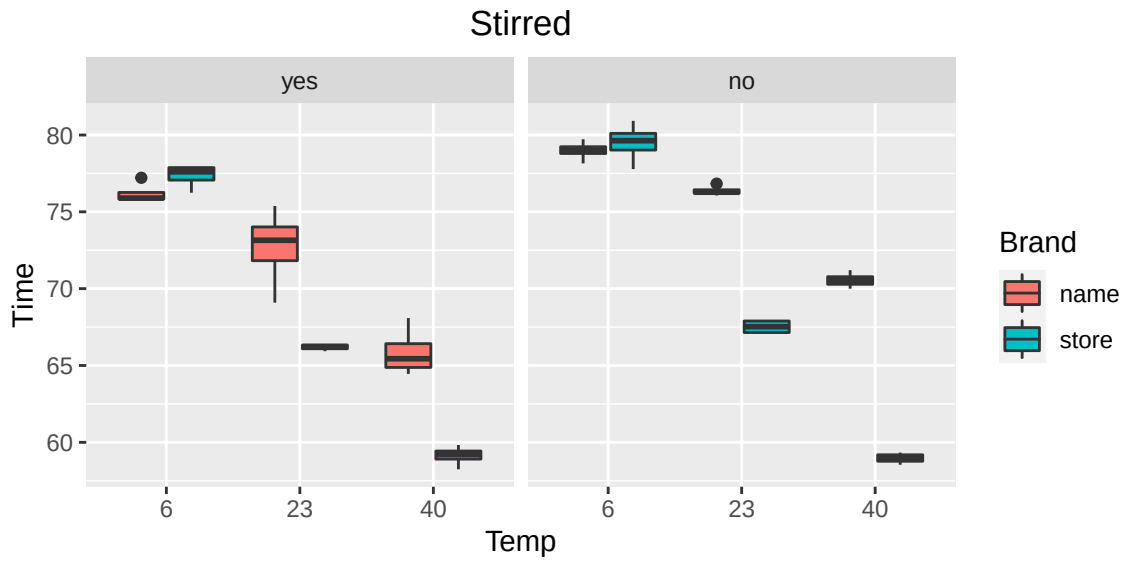
Further contextualizing central tendency and spread, we can see this illustrated in a different perspective by grouping effects together. Take note that temperature has the smallest set of ranges, both at each individual level and as a whole when compared to brand type and stirring effect. There is not a lot of varability among temperature compared to the other effects though they are different enought to identify an irregular increases as temperature increases. The range of values provides additional context to the data's story; data points when focused only on temperature groups tends to be within a smaller range of each other indicating less variability; meanwhile brand and stirring effects have a wider distribution and more variability–outliers withstanding.

**Interactions**

From the boxplots below, we can immediately see that stirring seems to increase the variance of the name-brand medicine–while also decreasing the mean differences of brand within each temperature grouping. An interaction effect between temperature and brand can be deduced if lines are drawn through the centers of the boxes. Earlier, we had introduce an insight from the summary statistics output indicating an inverse relationship between temperature and dissolve time–as temperature increases dissolve time decreases. The boxplot reinforces this idea. In fact, we can also make the claim that temperature has an inverse effect on dissolving times whether stirring is present or not–indications of temperature having a strong effect on dissolving time by itself. Stirring might have an additive effect regardless of temperature.

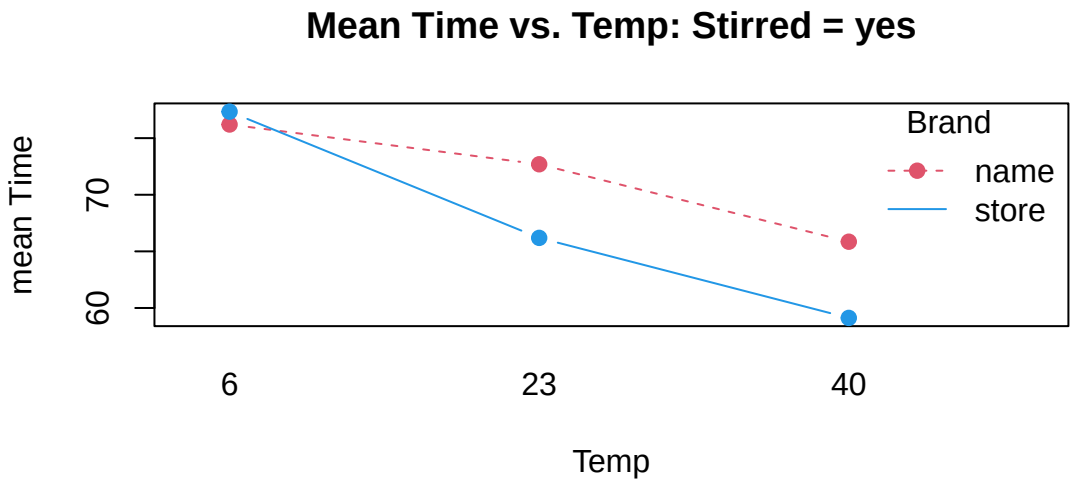
It is simply conjecture at this point, however we noticed that observations of dissolve time while stirring the water seems to have increased the name brand variability, while not stirring the water seems to have increased the store brand variability. Perhaps worth looking into the blocking effects of stirred on variability at various temperatures.

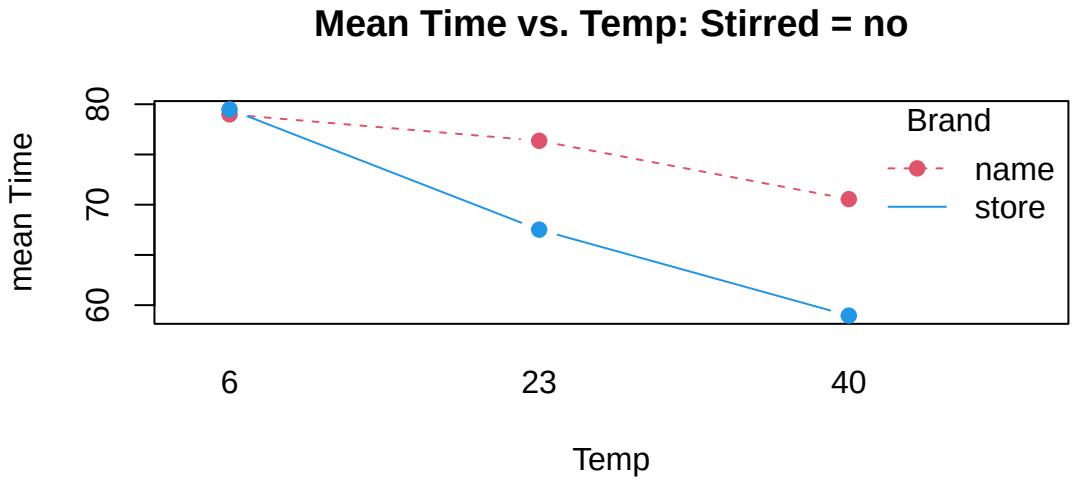
Outliers are present in our boxplots and we will address these data points when looking at assumptions.



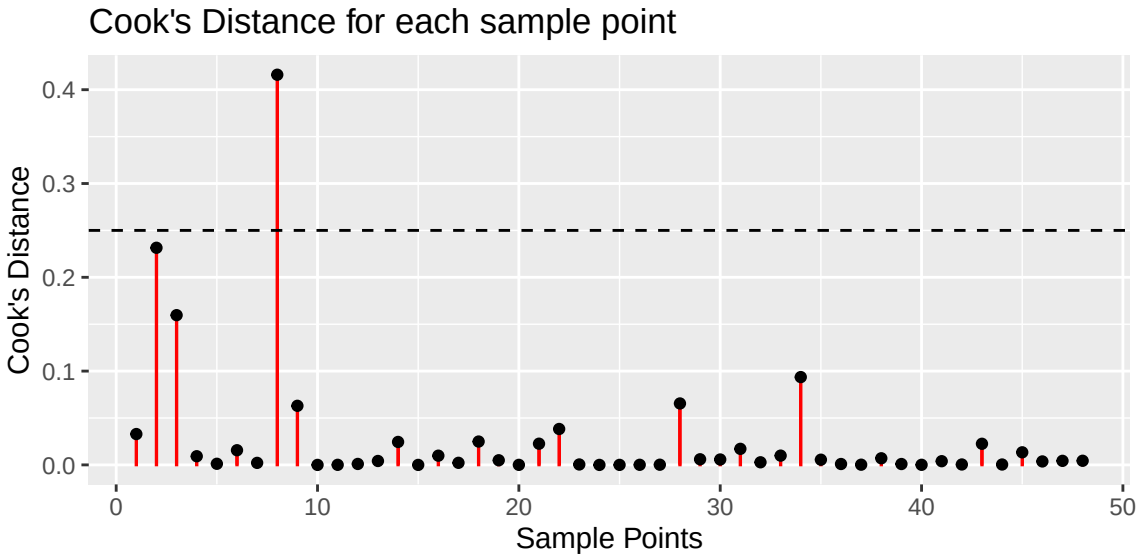

The possible interaction between brand and temperature becomes even more noticeable in the preceding three-factor interaction plots. Specifically, the brand and temperature interaction can be seen when the

temperature increases. The slope for the store brand has a more pronounced negative slope than the slope of the name brand. In addition, there might be a slight three-factor interaction between brand, temperature, and stirring as the name and store brand lines appear to be closer together in the stirred=yes plot than the stirred=no plot.



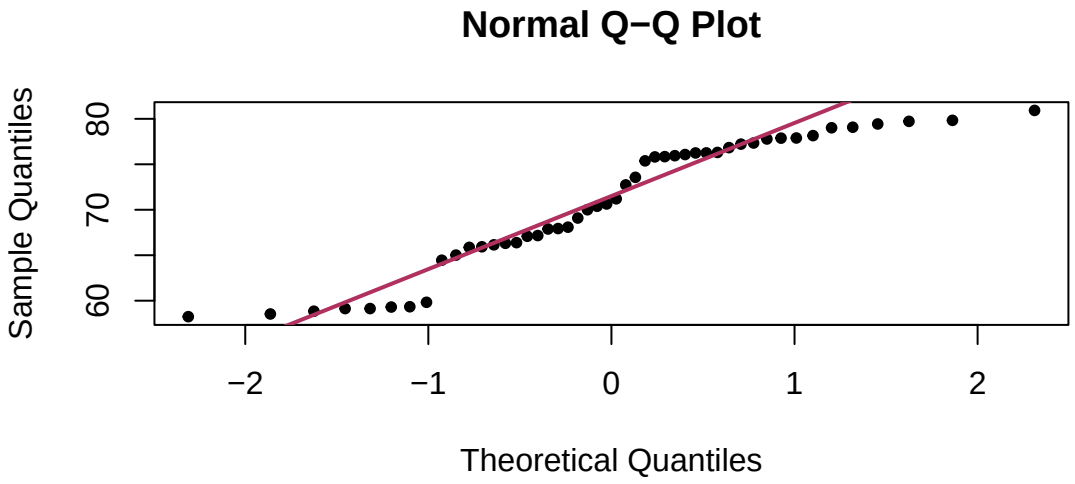



**Assumptions and Violations**




**Interaction**

From the boxplots, we were able to see a small number of outliers. To confirm if there is any concern we plotted the Cook's Distance for each point based on a full linear model. Point 8 has a higher Cook's distance than the rest of the points which may require removal for analysis if it is suspected of causing issues in the analysis. This would have to be weighed against the risks caused by introducing imbalances.




Finally, we check the normality of the data. Here a Q-Q plot is generated for the full model residuals. The data appears to suffer from heavy tails, multimodality and/or gaps in data between the left tail and the center. Since downstream analysis hinges on the assumption that our data is normally distributed, these issues may pose a problem.

**Recycling Bin**