EDA

Halid Kopanski

2022-11-21

Exploratory Data Analysis

For this study, we are presented with data from an 'Effervescent Experiment'. The data contained the dissolving times of two different brands of cold medicine tablets that were obtained under various conditions. Those conditions included varying water temperatures (6°, 23°, 40°) 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 contained 48 rows and 6 columns. The 6 columns include 3 explanatory variables (Brand, Temp, Stirred categorical factors), 2 response variables (Time and Org Time, both numerical), 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.

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

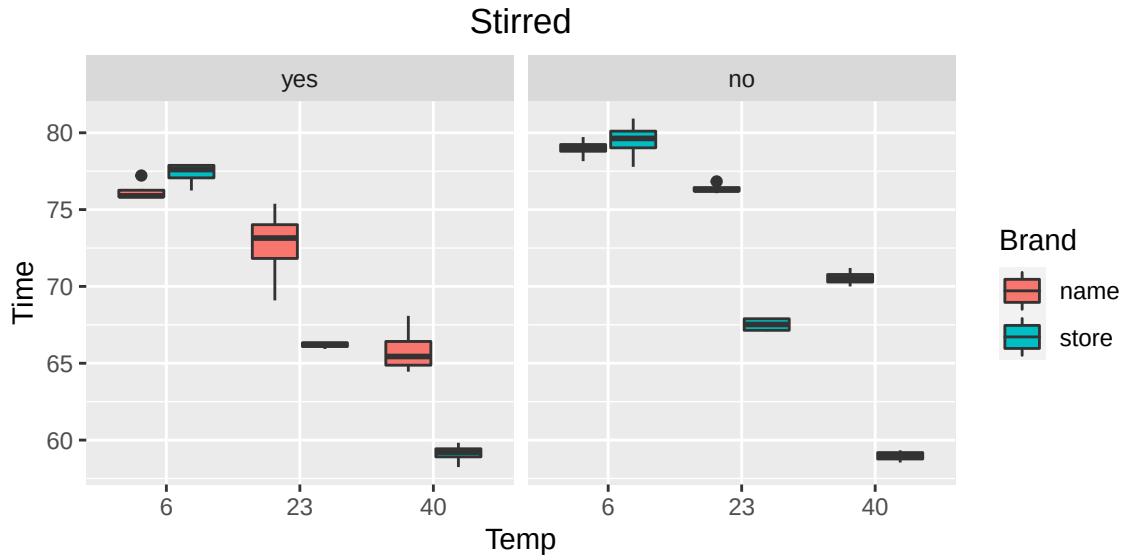
From the summary statistics table, we can see that each group has exactly 4 entries, eliminating concerns with respect to design imbalance. There appears to be a decrease in the mean dissolving times of store brand cold medicines compared to name brand. The disparity between means by brands is especially more prominent as the temperature increases.

```
##                     Min     Max   Range
## Store           58.2441 80.9218 22.6777
## Name            64.4516 79.7213 15.2697
## Temperature 6   75.8097 80.9218  5.1120
## Temperature 23 65.9281 76.8326 10.9046
## Temperature40   58.2441 71.2021 12.9580
## Stirred         58.2441 77.8909 19.6468
## Not Stirred     58.5392 80.9218 22.3826
```

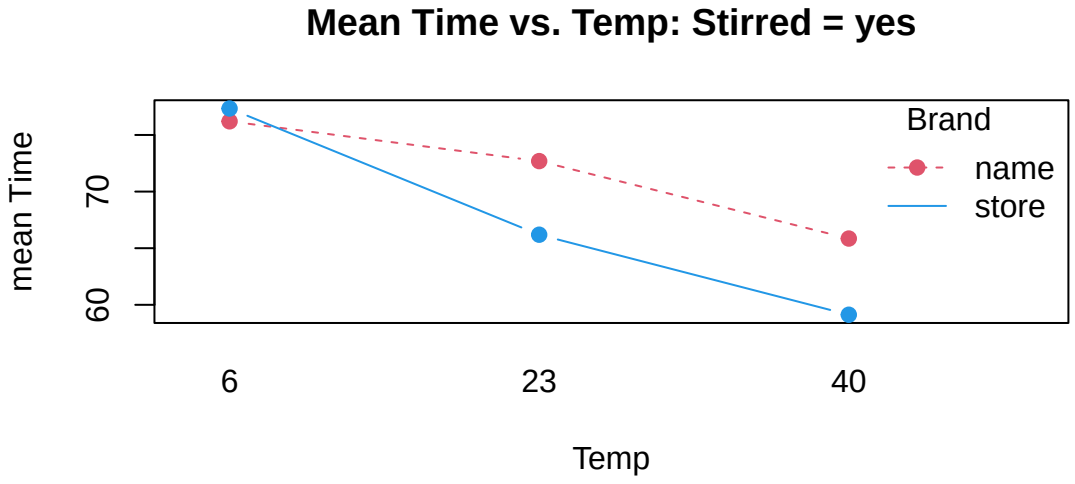
When effects are group together, we can consider the range of observations of dissolving time. Notice that out of the three effects, temperature has the smallest range both at each individual level and as a whole when compared to brand type and stirring effect. The range of values provides additional context to the

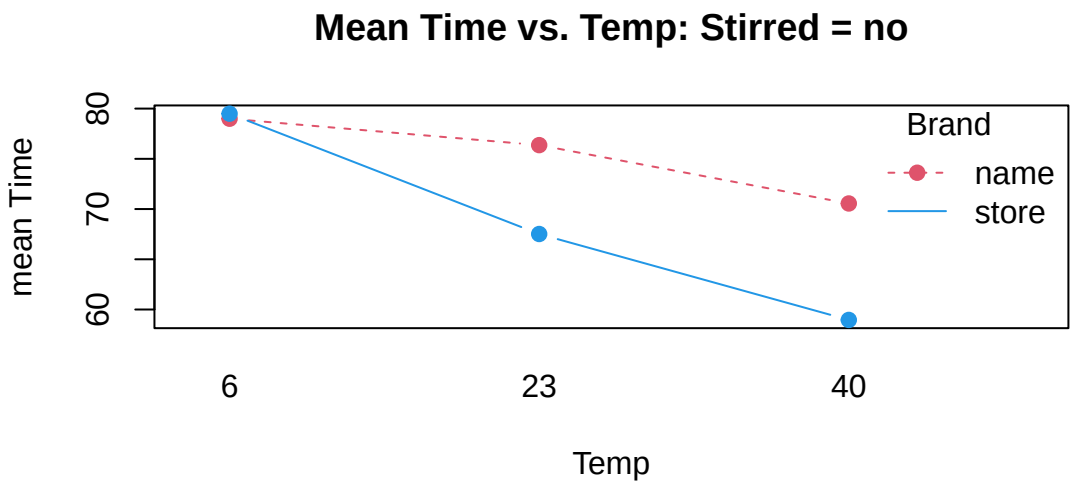
story the data; that temperature groups tend to be within a smaller range of each indicating less variability while brand and stirring effects have a wider distribution and more variability.

With range of observations in mind, let us consider the range in variability. The range in values of the variance is 6.9458, an interesting result considering 75% of all variance values fall between 0.1078 and 1.6943. We will want to look at the observations of name brand cold medicines stirred at 23° and 40° closely and highlight any interesting details. The variance seems to jump by quite a large amount between the groups, so contrast analysis might be a concern due to the small sample size.

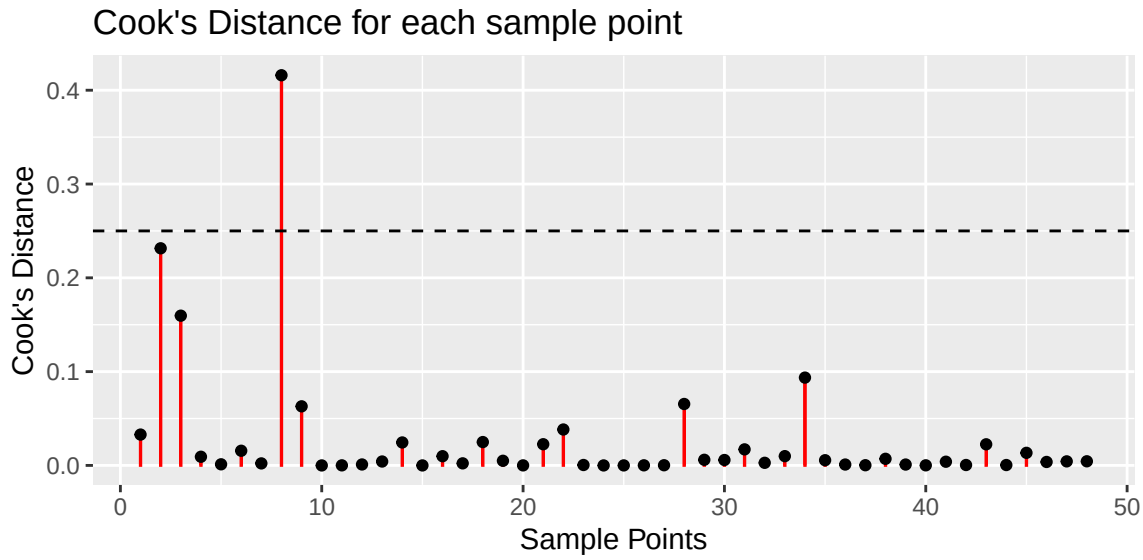


Immediately it can be seen that stirring seems to increase the variance of the name-brand medicine. Also, an interaction effect between temperature and brand can be deduced if lines are drawn through the centers of the boxes. We can also see that temperature has an inverse effect on dissolving times whether stirring is present or not. Stirring might have an additive effect regardless of temperature.

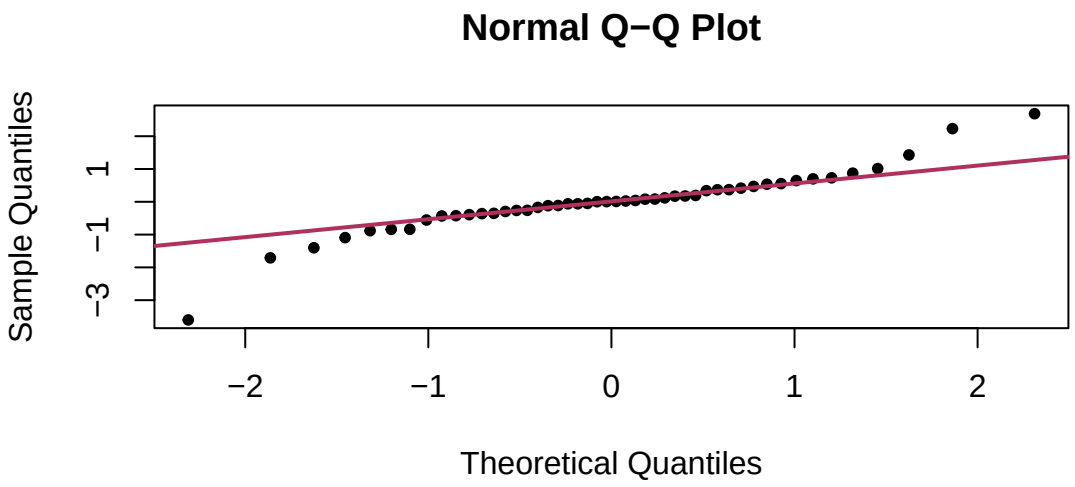




The possible interaction between brand and temperature becomes even more noticeable in the preceding three-factor interaction plots. Specifically, the brand and temperature interaction can be seen when the temperature increases. The slope for the store brand has a more pronounced negative slope than 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.



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 gap 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.