EDA

Halid Kopanski

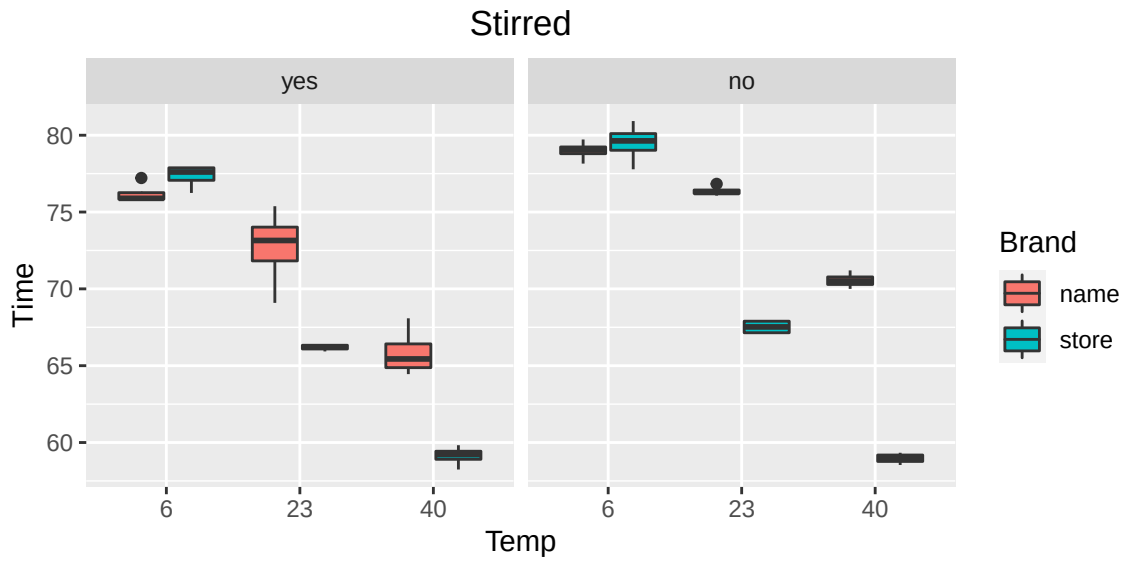
2022-11-14

Exploratory Data Analysis

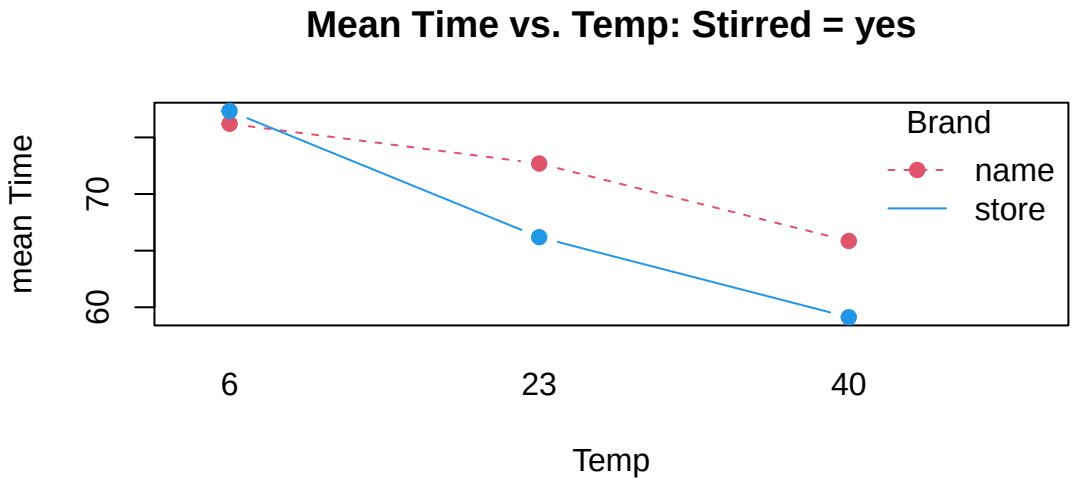
For this study we are presented with data from an 'Effervescent Experiment'. The data contained dissolving times of two different brands of cold medicine tablets that were obtained under various conditions. Those conditions included varying water temperatures (6°, 23°, 406°) and the presence of stirring (magnetic stir bar at 350 rpm). This was a complete block design with stirring acting as the blocking effect. In all, the data 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.

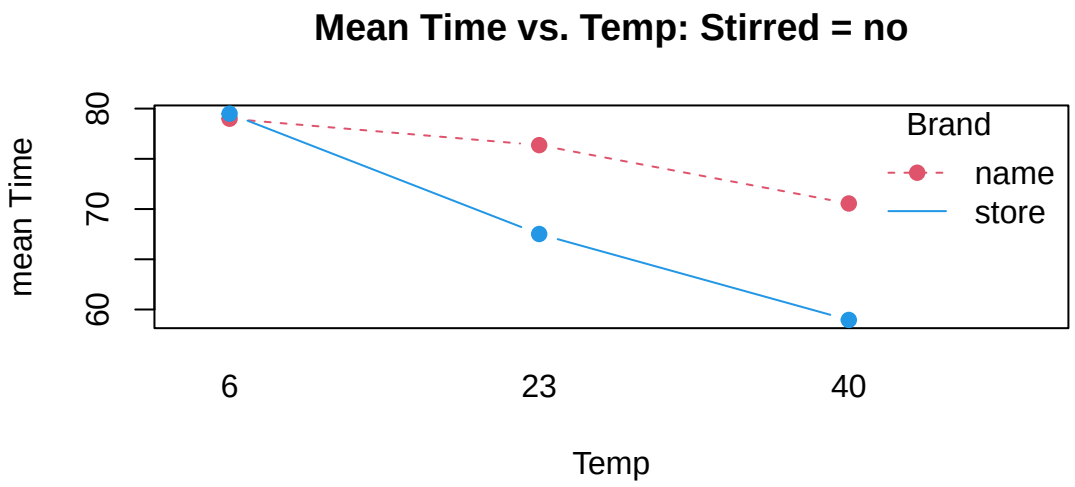
Brand	Temp	Stirred	25%	Mean	Median	75%	Var	n
name	6	yes	75.83358	76.20241	75.89223	76.26107	0.4593492	4
name	6	no	78.79910	78.99061	79.04435	79.23586	0.4146440	4
name	23	yes	71.82180	72.69145	73.14894	74.01859	6.9869087	4
name	23	no	76.20492	76.36351	76.27622	76.43481	0.1078134	4
name	40	yes	64.87321	65.85343	65.43863	66.41886	2.5499751	4
name	40	no	70.28754	70.55511	70.50947	70.77705	0.2544033	4
store	6	yes	77.06561	77.33703	77.60659	77.87801	0.5964884	4
store	6	no	79.01994	79.49240	79.63219	80.10465	1.6942517	4
store	23	yes	66.08831	66.19126	66.22629	66.32923	0.0411024	4
store	23	no	67.14393	67.51552	67.52360	67.89520	0.2060739	4
store	40	yes	58.90895	59.12529	59.21659	59.43293	0.4320148	4
store	40	no	58.76884	58.96347	58.99050	59.18513	0.1202191	4

From the summary statistics table, we can see that each group has exactly 4 entries, so no imbalance concerns. 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.

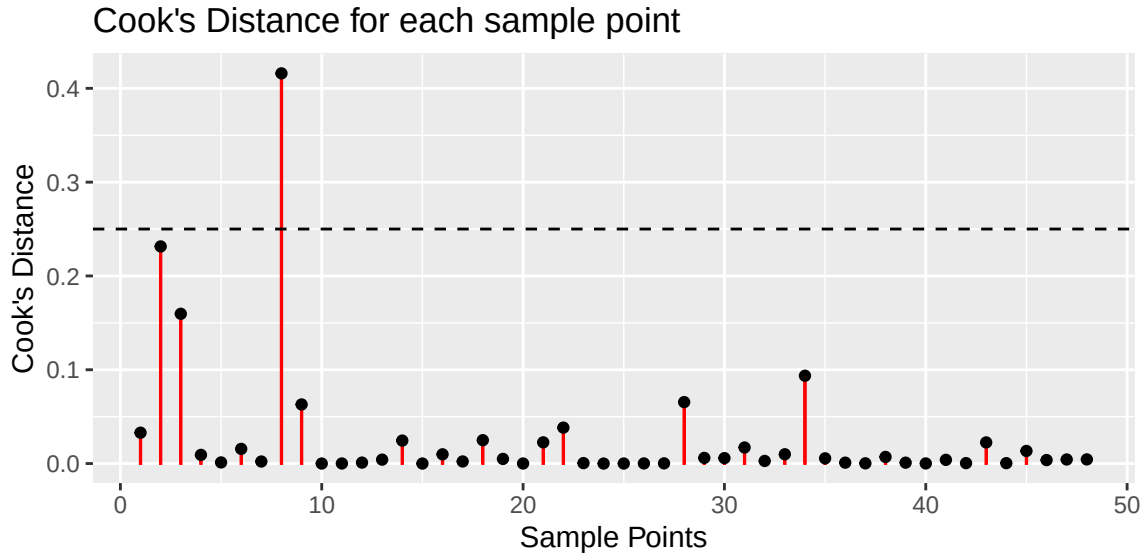


Immediately it can be seen that stirring seems to increase the variance of the name brand medicine. Also, an interaction effect between temperature and brand can be deduced if lines are drawn through the centers of the boxes. We can also see that temperature has an inverse effect on dissolving times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

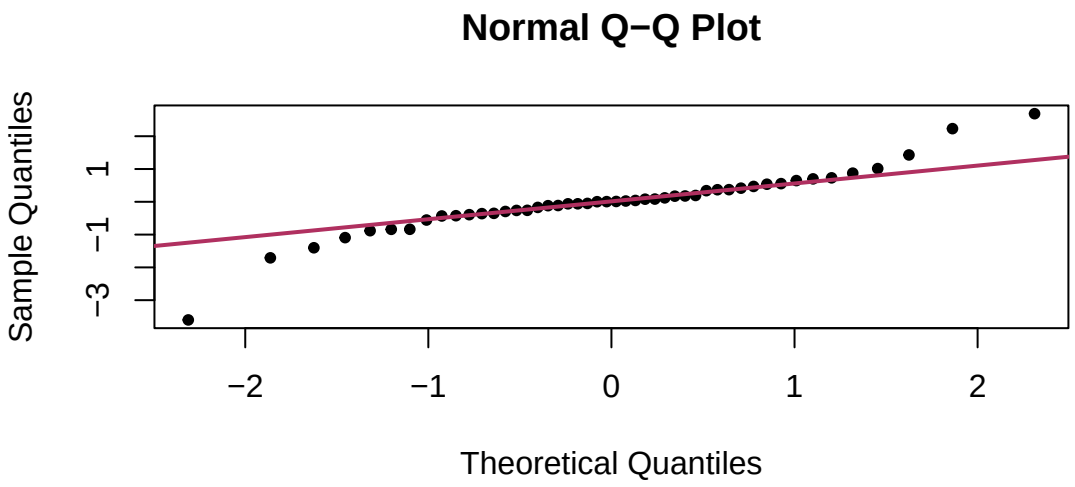




The possible interaction between brand and temperature becomes even more noticeable in the preceding three factor interaction plots. Specifically the brand and temperature interaction can be seen with increasing temperature the store brand line has a more negative slope than the name brand line. In addition, 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 amount of outliers. To confirm if there are any of 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 cause by introducing imbalances.



Finally, we check the normality of the data. Here a QQ plot is generated for the full model residuals. The data seems to be indicative of heavy tails. This might pose a problem for some of our analyses.