What we learned from the sponsor:

The sponsor gave the team 4 feedback points.

1. Sales outliers are systematically dealt with due to lag between live data flow and periodic audits. With this knowledge, we are creating a new variable "price" to identify mistakes in the data where the quantity or sales are calculated incorrectly.
2. Maverik advised us to impute the missing temperature data points and they confirmed they will be utilizing 10-day forecasted weather data in production. We learned our dataset only contains stores located in Utah County, which we could leverage if we choose to utilize a free historical weather API to impute the NAs.
3. They noted our point about singularities and commented that projects affect patterns significantly and suggested potentially creating two models, one for sites without projects and one for sites with projects.
4. Finally, Maverik suggested leveraging location id and limiting the model to sites that have been online for an extended period of time.

Results:
We have not generated any results so far.

Remains to be done:

We are moving forward with building the model in Python. We will use the statsmodel package (included in the approved list from Maverik) using an ARIMA algorithm adfuller(). We found this article about how to create an ARIMA sales forecasting model here - https://www.analyticsvidhya.com/blog/2020/10/how-to-create-an-arima-model-for-time-series-forecasting-in-python/. It is a good reference for us about the process of setting it up, checking for stationarity, and how to solve the model. This is the model we would recommend and seems to best fit our needs. https://github.com/Kydoimos97/CapstoneMSBA2020/issues/8

Problems we are encountering:

Some of the team is running into computing power problems because the dataset is large. We are looking into the possibility of using cloud platforms to help solve this.