**The Impact of Outliers: The Raisin River Canoe Race**

Downriver racing is a sport where competitors take their watercraft (usually a canoe, kayak, or stand-up paddle board) down a race course in competition with other paddlers, sometimes navigating rapids, in an attempt to get the fastest time either overall or in their class (classes are split up by boat and paddler type). The Raisin River Canoe race has been running for over 50 years on the Raisin River in Eastern Ontario. The usual race course is 30 km (about 18.6 miles) and contains rapids that fluctuate in difficulty with the water level. 

In race results, a DNF (did not finish) refers to a boat that did not make it to the finish line. Most of the time, these boats are paddlers who at some point decided to drop out of the race. This is most common in longer races (like Raisin River) or races with dangerous rapids. While the rapids on the Raisin River course are not too intimidating, as water level increases, the difficulty (or, more importantly, the fear factor) increases as well.

Race directors are responsible for keeping track of their competitors and making sure they get off the river safely, so it is important that race officials are aware of possible DNFs. Having a way to predict the proportion of participants that may DNF the race based on the water level (information that is known ahead of time) could be very useful. For this activity we will be using data containing DNF and water level measures from the Raisin River Race from 2005-2025 to look at how outliers can impact a model.

This data set contains flow measurements (indicative of water level) in ft3/sec and the proportion of DNFs (# of DNFs/total participants) from the Raisin River Race from 2005-2025 (excluding 2020 and 2021, when the race didn't run, and 2019, which has incomplete data).

Explore the relationship between flow rate and proportion of DNFs

1. Graph a scatterplot that shows flow rate as a predictor for proportion of DNF
2. Fit a model using flow rate as a predictor for proportion of DNF and meaningfully interpret the slope coefficient.
3. The model has a negative coefficient for flow's impact on DNF proportion. Does this make sense with what you saw on the graph from question 1?
4. Note the p-value for the flow coefficient. Is it significant?
5. Comment on the appropriateness of the model

Investigate influential observations

1. Do any observations have unusual leverage? standardized residual? influence?

Consider the outlier

1. We only want to remove outliers if we have reason to believe that outside factors may be influencing the data in a way that prevents data analysis. In this case, the outlying observation is from 2017. In 2017, the water level was dangerously high, causing the race officials to move the race start to Delaney Road (roughly 4 miles further down the river), leaving the finish in the normal place. The map below shows the race course. Do you think this change warrants removing the year from the data set? Why or why not?
2. Regardless of your answer to question 4, remove the influential observation from the data set.
3. Regraph the scatterplot without the influential observation and add a line of best fit.
4. Refit the model with the new version of the data set.
5. Provide a meaningfully interpret the slope coefficient.
6. If the flow level is 1100 ft3/sec and there are 200 competitors, how many DNFs can we expect?