Disney World Ride Wait Times - Proposal

Gus Lipkin

I started the Disney World Ride Wait Times project in Spring 2020 when I took Intro to Data Science with Dr Sanchez. I enjoyed working on the project and visually identifying seasonal trends that allowed us to make recommendations on when to visit the parks in order to avoid longer wait times. While I would consider that project a success, it definitely was a beginner project. The most advanced tool I used was `facet_wrap` to create calendar heatmaps. Our last slide talked about next steps for the project. I was mildly familiar with linear regression and trend lines from my time in high school physics, but I knew that more could be done. The one regression I did for that slide left a lot to be desired. My plan is to revisit the original dataset and instead of predicting the best time of year to visit, I want to identify which lines will become shorter before that happens.

I'll be using the Disney World Wait Time dataset from touringplans.com. I have wait time data from 2012 all the way through the end of 2021 for fourteen rides. In addition, I have metadata for each of those days. The metadata has area specific data such as weather and sunset time, park specific data such as opening and closing times, and other important data such as the amount of students in school in a given region such as the US northeast.

I don't want to perform a time-series analysis, but I will have to employ some time-series techniques to account for data seasonality. I will also have to make sure that I log transform the appropriate variables because wait times cannot be negative. I will also have to make other considerations since the dataset is missing lots of times in a random manner. On the other hand, since there is so much data, I'll need to be smart about my operations and making sure they're energy efficient. I'm not entirely sure that I will be able to get the answers I want, but I'm hoping for the best and accuracy rates over 50%. That said, there are a few ways I can approach this. I can do one big regression, and pretend that the rides are not in separate areas (a bad idea), do a multinomial logistic regression to predict which ride will have the shortest wait time (not entirely sure how to do this), or a bunch of regressions on each of the rides (probably my best bet). Of course, that's a lot of work :(.

Realistically, I'm just going to figure it out as I go, but I'm confident it'll at least be something I can be proud of.