Statistical Hypothesis:

By exploring a dataset of wait times for the Walt Disney World ride *Splash Mountain*, along with a metadata set that contains relevant information about the park and other factors, I expect to find:

- A positive relationship between average daily temperature and wait times.
- A negative relationship between daily precipitation and wait times.
- A positive relationship between holidays and wait times.
- A negative relationship between the percentage of schools in session and wait times
- No relationship between the number of hours the park is open and wait times.

Outcome:

Initially with two scatterplots, I was able to observe a positive relationship between daily temperatures and wait times. This supported my hypothesis. I also found a positive relationship between rainfall and wait times, which I did not expect to find. Upon performing an OLS regression, I found relationships at significant levels for daily temperatures, precipitation, schools in session and park hours. This regression model had an adjusted r^2 value of 0.303 indicating that the model can explain 30.3% of the variance in daily wait times. Adding the holiday factor to the model did not improve the model any and the column had a very large p-value, 0.36, and was not included in the final regression.

	coef	std err	t	P>|t|
const	-42.0235	4.412	-9.526	0.00
Rain	-11.3347	5.976	-1.897	0.06
Temp	0.6325	0.049	12.984	0.00
in_session	-10.2395	1.065	-9.617	0.00
park_hours	3.1204	0.177	17.636	0.00

Ultimately, I found positive relationships between temperature and park hours and wait times. The relationship between temperature and wait times was what I had expected to find, but I was surprised to see a relationship between park hours and wait times. I also found negative relationships between rainfall and wait times and the percentage of schools in session in wait times. These results were expected.

One of the challenges of this analysis is that I am working with third-party data. As a result, I do not have access to actual, logged wait times, just estimated wait times that were reported through Disney’s mobile application. Having actual wait times would lead to a more accurate model. Without these, I’m essentially trying to estimate an estimate.

I believe the rainfall estimates may be flawed for several reasons. For one, when exploring the dataset I found a minimum value for precipitation of 0.03. No day has less precipitation than this. I know it rains a lot in Florida but it’s not quite a daily occurrence. In the process of cleaning the dataset, I also had to remove times when the attraction was not operating, which would include thunderstorms in the area. For these reasons, I’m not sure the rainfall factor was fully properly captured.