Group Assignment #1: Restaurant Lunch Sales

A mom-and-pop restaurant in southern Brazil finally reopened for business after the pandemic. The owners hired you to predict customer traffic. They have given you two years of data on the number of lunch sales per day. The data also include information about the weather and the local tourist season. The complete list of variables is included at the end of this assignment description. The owners have asked you to generate insights about the relationship between some of these variables and lunch sales.

The sample period spans two pre-Covid years (February 2018 to March 2020). The owners understand that customer traffic may not follow the same patterns in the future, but these data are the best available, so you have been asked to analyze them anyway.

To answer the owners' questions (below), estimate a model of lunch sales regressed on the following variables: precipitation, temperature, humidity, whether it is high tourist season, number of open competitors, and day of the week.

Hint: Pay attention to which variables are categorical variables.

Use your regression results to answer the following questions:

1. Use your regression results to analyze the effects of competition on lunch sales.

   a. What is the relationship between the number of open competitors and lunch sales, all else equal? Does the sign of the coefficient estimate make sense?

   b. Re-run the regression using log lunch sales as the dependent variable. What is the effect of having one more open competitor on lunch sales?

   c. The owners are concerned that the estimated relationship above may be biased. Discuss one potential omitted variable and the direction of the bias.

You can use either model (i.e., lunch sales in level or log) for the remaining questions. Be clear about your choice and stick to it throughout.

2. Which day of the week has the highest average lunch sales, all else equal? Which day of the week has the lowest? Is the difference between the two statistically significant?

3. Implement a residual analysis of your model.

   a. Does the plot reveal some potential issue with the regression model? What might explain the patterns of the residuals?

   b. How would you fix the issue identified above? Briefly discuss how fixing the issue may change the estimated effects of competition on sales. (You do not need to implement the fix)

Dataset Description:

Your dataset, *brazilianrestaurant.csv*, is composed of observations at the day level and contains the following variables:

Variable	**Description**
Date	Date in yyyy-mm-dd format
DayInData	Observation number
DayOfWeek	Day of the week, e.g. Monday, Tuesday, etc. (string variable)
Month	Calendar month
NumberOfLunchSales	Number of lunch sales that day (each sale is a ticket that may contain multiple dishes)
Weekend	Dummy, =1 if Saturday or Sunday
HighSeason	Dummy, =1 during high tourist season in southern Brazil
NumberOfOpenCompetitors	Number of other restaurants open in the neighborhood that day
PrecipitationMm	Precipitation in millimeters
LunchtimeTemperatureCelsius	Temperature at noon, in degrees Celsius
HumidityPct	Relative humidity, on a scale of 0 to 100

Useful code:
You can use the “relevel” function to change the baseline category for a categorical variable.
For example, to make Monday the baseline category for DayOfWeek, you can write:
df_lunch$DayOfWeek <- relevel(factor(df_lunch$DayOfWeek), ref="Monday")