## Cleaning Data for Move Run-Over Prediction

**Setting Up**

First, I imported pandas so I could work with this data as a DataFrame and numpy for feature engineering. I read the csv into a pandas DataFrame, with column ‘order\_id’ as the index. (After each step, I either used the pandas head(), info(), or describe() method to check if the code had the intended effect.)

**Removing Extraneous Columns and Rows**

This whole dataset incorporates a few months of data for customers that booked both on the web and over the phone. I removed the columns beginning with ‘calculator\_’ because they are the result of a pre-move survey Bellhops only gave to some customers who booked over the phone over the past couple of weeks. The survey asks questions that would seem helpful in the move length estimation process like “is there a basement at the first location?” “is there an elevator at the second location?” etc. However, because about 98% of the responses were null, they were only from recent phone orders and not randomly distributed throughout the dataset, and there was no logical way to replace the values, I decided to drop the 10 columns. (Hopefully in a future analysis those can be incorporated.)

Furthermore, I dropped about 300 rows out of almost 9,000 where the ‘move\_length\_hours’ column was null or less than 0. (There is clearly an issue with the data if it says the start time was after the end time.) As the independent variable is based on this column, it made no sense to replace values.

**Feature Engineering**

With this project, I am trying to predict whether or not a move will run over 4 hours. Bellhops never intentionally books a move over four hours as it can cause a back-up throughout the rest of the day. I generated a column to represent the independent variable called ‘over’ that gives a '1' if the move length was over 4 hours and a '0' if not.

I transformed the ‘property\_type\_name’ column, which describes whether a property is a house, apartment, storage facility, etc. from an object type into a categorical variable and transformed the ‘booked\_at’ and ‘reservation\_start’ columns into datetimes.

From the ‘reservation\_start’ column, I added columns that represent the hour of day the move began, whether or not that was in the morning, the day of week the move began, and whether or not that was a weekend day.

Finally, I wrote the data to a csv, ready for exploratory data analysis.