* a half page write-up of any observations/visuals from the baseline model (I already have a little of this at the top of the notebook)

For our baseline model, we were to fit a linear regression model using the relevant features to predict price. However, the main hurdle in this task was massaging/manipulating each of the raw features so that we could fit our model while achieving interpretability and computational efficiency.

To incorporate the days of the week and holidays, we used our work on the “Average Difference from Listing’s Own Mean Price” from Milestone #3 in visualizing how prices changed throughout the year. Ultimately, we found that the real increase in per-night rental costs came on Friday, Saturday, and around the New Years’ holidays. Additionally, we found that non-holiday dates in January and February showed the lowest prices, which we deem as “slump” dates. Thus, we made categorical variables to denote the day of the week (weekend or no weekend), holiday (3 days around New Years’), and slump dates (January and February dates that aren’t around New Years’).

Because the categorical variables of neighborhood and zipcode have over 200 distinct values each, one-hot encoding would produce to far too many variables for a linear regression model – leading to long computational time and a small chance of over-fitting. To solve this, we create four categorical variables for each of these features that separate the neighborhoods and zipcodes by price into quartiles. Each quartile is its own categorical variables (e.g. most expensive 25% of neighborhoods, least expensive 25% of zipcodes). Thus, we move away from trying to account for individual neighborhoods such as “Tribeca” separately and instead choose to analyze the most expensive neighborhoods together. While we lose some degree of granularity, we believe that what is gained in computational efficiency and streamlined interpretability is well worth it.