The original raw data (ski\_resort\_data.csv) had 330 rows where each row represented a ski resort in USA. Among the 27 different information categories reported for each resort, some of the data were not available and the corresponding values were set to NaN in the data. However, all the 27 columns for the Big Mountain Resort had non-null values.

To obtain a pricing model for the ski resort, there are few pieces of information that are very critical, such as ticket price. Our further analysis revealed around 16% of the weekday and weekend ticket prices were missing. However, we ensure to check all the data before deleting the data for resorts where the ticket prices are not reported, because they might include useful information which can be used to clean the data. There were few things we tried to check in the data, such as:

* Make sure we did not have any duplicated results by checking the uniqueness of the resort name + region/state combination
* Distribution of feature values

Studying the distribution of each column’s information over all the resorts provided very useful information about the possible issues on the data. Based the abnormality in the results, we either deleted some rows, columns, or corrected the corresponding values. For instance, further investigations of very odd distribution of “yearsOpen” showed there is one resort that the corresponding value was set to 2019. We deleted the corresponding row. Or the further study of the weird shape for “SkiableTerrain\_ac” showed the value was reported wrong. Finally, investigating the “Fast eight” suggested it was only available for one resort. So, it did not carry useful information and we deleted the whole column.

Finally, there were a useful information missing from the data, such as the population and area of each state. We obtained this information from <https://simple.wikipedia.org/w/index.php?title=List_of_U.S._states&oldid=7168473> and merged it to our dataset.

Finally comparing the data for the ticket prices showed, the data missed more weekday prices than the weekend. So we dropped the information on the weekday prices and deleted the rows that the weekend prices were missing. Then the final data with 277 rows is saved into the file named “ski\_data\_cleaned.csv”.