During the data wrangling process the following data sets were analyzed for a specific purpose (respectively):

* ski\_resort\_data.csv – details on the various ski resorts across various regions and states
* <https://simple.wikipedia.org/w/index.php?title=List_of_U.S._states&oldid=7168473> – wiki page with state information (population, square mileage, etc.)

Initially when analyzing the data from ski resort data, it contained 330 rows and 27 columns.

These columns included ticket prices for adults during the weekday and weekend, total skiable

terrain (acres), chairs included throughout the resort as well as the types of lifts and so on.

Upon further analysis there were columns that included a significant amount of null values such

as the column *fastEight* and *NightSkiing\_ac* with null values accounting for 50% and 43%

respectively. Even for the ticket price columns we were saying no values show up roughly 14%

at the time. In addition to looking at these numerical values another observation was made into

categorical columns in the data such as *Name*, *Region* and *state*. The values of these columns

provided insight to the number of potential unique records there was in the data. However,

upon further investigation it appeared that 90% of the time the *Region* and *state* column were

the same value while 10% were different. The question that came about was what exactly is a

*Region* in the context of the data? Looking through the data and finding the *Regions* include

data that include states and locations of a smaller geographic area within a state (i.e. California:

Sierra Nevada and Northern California). When doing a comparison between both *Region* and

*state* regarding the number of resorts in both cases New York had the highest number of

resorts compared to the others. The question arose can Big Mountain’s pricing rates be based

on New York’s pricing premium due to it resort’s proximity to population? Or should the focus

be primarily on the state of Montana where this resort resides? One way to address this

question was to visualize the distribution of data across the various states using a boxplot. The

boxplot provided insight into the how the pricing on average for each state varied from average

its mean price in comparison to each other. After comparing price data for each state, the next

step was to look at the other columns in the data. The following columns were cause for

concern in the data:

* SkiableTerrain\_ac - many of the values clustered at the low end
* Snow\_Making\_ac – many of the values clustered at the low end as well
* fastEight – little variance with many of the values being 0
* fastSixes – more variance than fasEight but many of the values being 0
* trams – many of the values around zero with little variance
* yearsOpen – this has a vastly wide variance from 0 to 2019

One of the ways to verify some of the values that were seen for some of these columns was to double check the outlier information such as *SkiableTerrain\_ac* which had a maximum value of 26,819 acres. The name of the resort related to this value was Silverton Mountain. Upon further inspection it was noted that this data was incorrect and the actual acreage was 1819. Validating records in the provided data is a time-consuming process and the effort has to provide significant contribution to the overall objective. For example there was a record that had incorrect *Snow\_Making\_ac*, however, this record didn’t include any pricing information. Efforts rectifying this data would prove useless if pursued. This meant that some of the data would not be relevant to addressing the original objective of this project. As a result a filter was applied to the data to remove rows that were outliers to the general distribution of the data, including data that didn’t contain pricing information (~14% of the records). This approach was also taken with some of the columns in the data. An example of this was the *fastEight* column, more than half of its data was missing in the csv provided and as result it was decided to drop this column. Upon cleaning up these outliers, the distribution of data was now in a state to provide comparable summaries of the resort information for each state. To do this a new DataFrame was created that provided information on aggregation of all the relavant columns (*Name, SkiableTerrain\_ac, daysOpenLastYear, TerrainParks, NightSkiing\_ac*) for each state. The next step after grouping and summarizing was to supplement this data with information pertaining to the state itself. As mentioned earlier, could there be a correlation between resort price and population or pricing and state square miles, or a combination of both? For this external data set Wikipedia was used as a source providing the needed information. However, with this new dataset there were inconsistencies in the data such as extraneous characters that prevented a match within the DataFrame from the ski resort data. The resolution was to simply find and replace all instances of the special characters in the wiki data set and that resolved the issue of missing rows in the data. Next was to compare the columns *AdultWeekday* and *AdultWeekend* to find a relationship between the prices using a scatterplot diagram. At first glance there appears to be a linear relationship between the two as they both share the same price and the only difference between the two are at the lower priced resorts. However, when looking at the state of Montana it can was noted that the average price for both *AdultWeekday* and *AdultWeekend* at the various resorts are the same. With that being said it was concluded that the *AdultWeekday* column can be dropped, since both columns are the same price in the state of Montana. Another look through missing values was done and it appeared that every fourth column of data was removed manually, not sure what conclude but will need to do a deeper investigation into the data to decide whether these columns should be dropped as well. This concludes the initial data wrangling and the results of this analysis have been saved into the following files (ski\_data\_cleaned.csv and state\_summary.csv).