**Q: 3**

This exercise is focused on evaluating the relationship between different features of the resorts across United States to determine the target variable for modelling the ticket price for increasing the revenue of the Big Mountain Resort, a ski resort located in Montana. Dataset used, ski\_resort\_data.csv, contains information about different resorts across the country with 330 rows and 27 columns. Name column/feature has resort names including Big Mountain Resort (resort of interest) and doesn’t have duplicate resort names when region/state is considered.

AdultWeekend and AdultWeekday columns are weekend and weekday adult ticket price in dollars; so considered these columns as target variables. Information from these columns across states suggests, except for few states, most ticket prices were in range of 25 to over $100. Besides, average weekend ticket price was higher than weekday price for all the states except Colarado, NewJersey, NewMexico, Montana, South Dakota.

While processing data, droppedfastEight feature because all but one had value of 0, so it has very little variance, and half the values are missing. Also, fastSixes and trams features has more variability, but still mostly 0. Not deleted yet. Also dropped rows where missing values observed in both AdultWeekend and AdultWeekday columns. Inaddition, found issues with certain features and were able to fix them. For example:

1. SkiableTerrain\_ac feature had wide range of values and max is 26819 (could be an editing error, so corrected value by spot checking data), clustered down the low end. So, filtered the rows having SkiableTerrain\_ac >10000
2. yearsOpen because most values are low but it has a maximum of 2019, which strongly suggests someone recorded calendar year rather than number of years. So, filtered rows having yearsOpen <1000.

Also used another data frame containing population and area data for the US states. Comparing the states column between both data frames, used regular expression for text modifications (deleted square brackets and their contents) to make sure the consistency. Merged population and area dataframe to ski resorts one.

For determining the target variable for modelling the ticket price, plotted the relationship between weekend and weekday variables using matplotlib and found a clear line indicating those prices were equal as seen in box plot earlier. But weekend prices were higher than weekday for the prices up to $100. However, weekday and weekend prices were same for Montana resorts including Big Mountain resort. Since more missing data, weekday column is dropped from ski resorts data and saved for future analysis. Finally, left with 277 rows and 25 columns.

Left: relationships between the states.