\*\*Q: 1\*\* Write a summary of the exploratory data analysis above. What numerical or categorical features were in the data?

Aside from the state name, region and name of the resort, all the variables were numerical. However, we did create a categorical variable called ‘Quartile’ at one point to look at the quartiles of AdultWeekend resort prices. There was no clear pattern that emerged when we compared price quartile to the first two components of the primary components analysis.

Was there any pattern suggested of a relationship between state and ticket price?

There was no straightforward pattern between states and average ticket prices for ski resorts in that state. We looked at the distribution of prices. We also visualized how ticket price varied by principal components generated from state-level numerical data, resort-level numerical data using a heat map, and scatter plots between ticket price and numerical variables at the state level.

In terms of distribution, the average of weekend and weekday ticket prices in a state varied from $42 to $93 across the 34 states in the data frame; the distribution of this average ticket price was not normally distributed. Montana was in the lowest quartile and the 8th cheapest in terms of the average between its’ weekend and weekday ticket price.

Based on the numeric data included in our data set, PCA analysis showed that the first two components explained 77.2% of variation in ticket price. The same PCA showed that the ticket prices in Vermont and New Hampshire stood far above other states, presumably due in part to their relatively high number of resorts per 100 square miles and resorts per 100,000 people in those states. New York, Colorado, Michigan, and California prices also seemed stood apart seemingly because of the relatively large number total open days at resorts in the state the previous year and the average number of terrain parks at ski resorts in those state.

A heat map visualizing the strength and magnitude of two-way correlations between all numerical values in the full resort-level data set showed that the adult weekend price was strongly and positively correlated with vertical drop, the number of fast quads, the number of total chairs, the number of runs, the area covered by snow making machines. I think the relationship between resort ticket price and factors such as the ratios that reflect the relative skiable area covered in lights for night skiing in a given state, the relative number of square miles of skiable area in a state, and the relative number of ski resorts in a state are less clear and strong. In addition, some of these ratios cannot be influenced by an individual ski resort in Montana, such as our client, so they do not seem useful in terms of figuring out what the park could do to charge more per ticket but could be useful if the client was interested in further honing its pricing strategy based on a subset of competitors with similar characteristics.

What did this lead us to decide regarding which features to use in subsequent modeling?

Based on the results of the EDA summary above, I think the features worth including in subsequent modeling are the following:

* The ratio of the number of terrain parks at a ski resort to the total number of terrain parks at ski resorts in its state (e.g. it’s market share within the state)
* The vertical change in elevation from the summit to the base at a ski resort
* The number of fast quads (four-person chairlifts) at a ski resort
* The total number of chairs at a ski resort
* The number of runs at a ski resort
* The number of acres area covered by snow making machines at a ski resort

In addition, resort characteristics worth adjusting for at the state level which may influence resort prices but that our Montana ski resort cannot necessarily control are the following:

* The ratio of the number of resorts per 100 square miles within a state
* The ratio of the number of resorts per 100,000 people within a state
* The ratio of the number total open days at resorts in the state the previous year

What aspects of the data (e.g. relationships between features) should you remain wary of when you come to perform feature selection for modeling?

We should be weary of the covariance or collinearity between variables. Adding correlated variables to our model is redundant in that it doesn’t give our model new additional information to work off of. For example, the number of resorts per 100 square miles and the number of resorts per capita are strongly and positively correlated. Likewise, there is strong correlation between the number of chairs at a ski resort and the number of fast quads, because both are counting the number of chairlifts.

Two key points that must be addressed are the choice of target feature for your modelling and how, if at all, you're going to handle the states labels in the data.

In terms of handling states labels, the provided notebook argues that it probably makes most sense to keep all the states together in one model for the time being as no clear groupings emerged from the plot of states by principal component 1 and principal component 2. However, I would personally remove the six states that had eigen scores above 2 for the 2nd component and eigen scores over 2 for the first component.