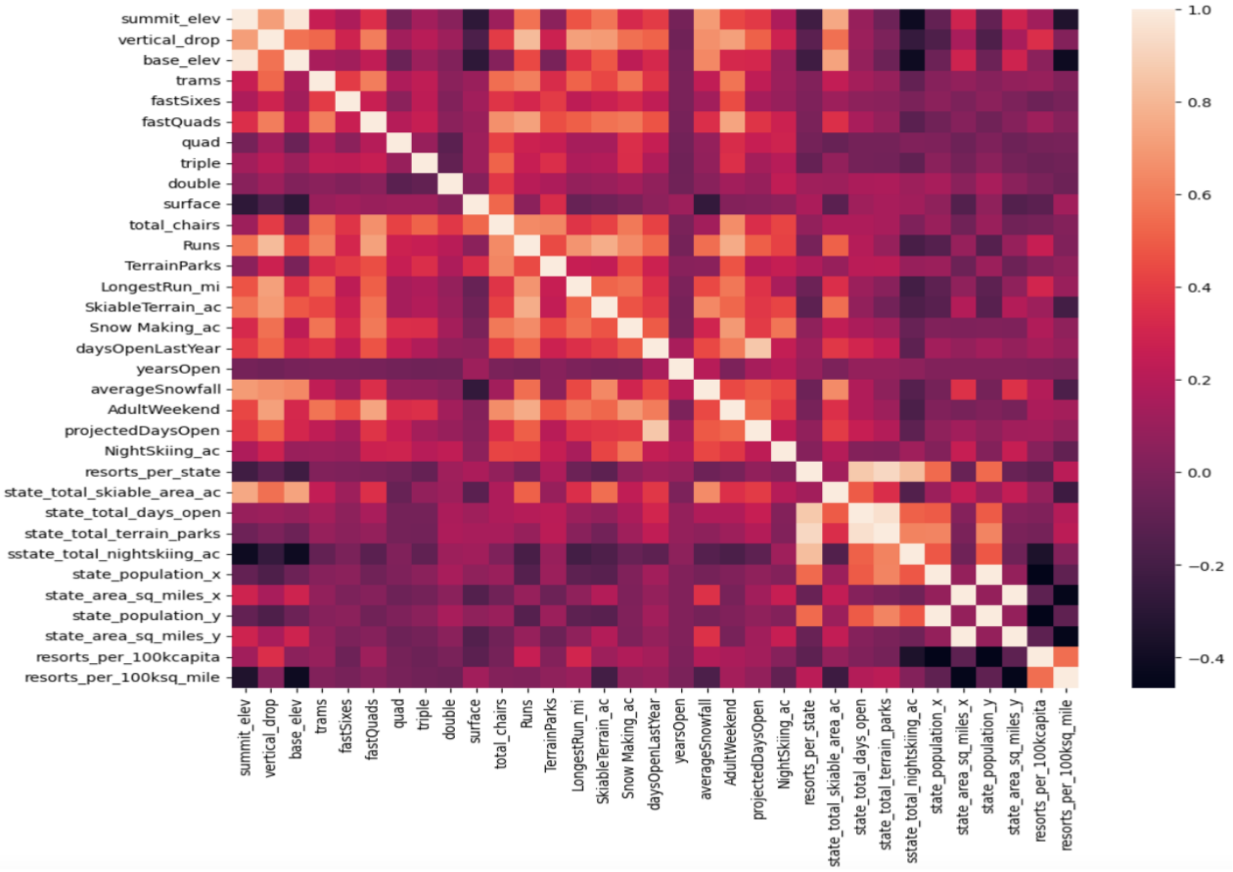
Big Mountain Resort just invested 1.54 M into another lift and needs to redistribute costs in a way that captures more value within the next year that is greater than market value. I have been tasked with performing a data analysis of the data provided.

Initial data wrangling showed that there were missing data within the data set as well as data that needed to be deleted because it just wasn't relevant to our analysis. Columns fastEight and AdultWeekday were dropped. Rows for null values for AdultWeekend ticket prices were also dropped. FastEight was dropped due to excessive values of 0. AdultWeekday ticket prices were dropped because of excessive null values. And all the null values for AdultWeekend were dropped to give a better data set of values. A lot of columns had values of zero or were skewed more to the zero value which indicated incomplete data. SkiableTerrain_ac because values are clustered down the low end, Snow Making_ac for the same reason, fastEight because all but one value is 0 so it has very little variance, and half the values are missing, fastSixes raises an amber flag; it has more variability, but still mostly 0, trams also may get an amber flag for the same reason, 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. The ski resort with year 2019 was dropped, the skiable_terrain_ac was corrected from 26819 to 1819, and Heavenly Mountain having a snow making of 3379 was corrected to 2880. We are currently finished with data wrangling. All data that is important has been tabulated and is ready to move on to the next step. I believe the target feature is the snow_making_ac. There were 277 rows left and 25 columns.

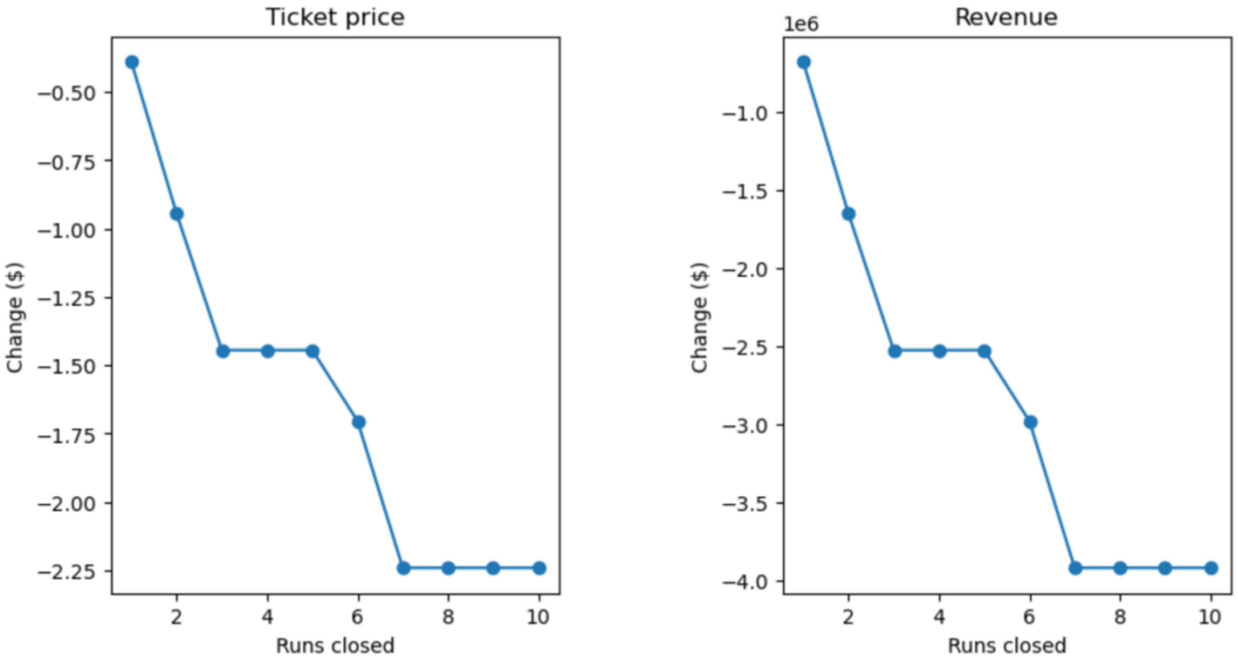
Next an exploratory data analysis was performed to see if there were any relationships between data sets. Yes, there was a relationship between state and ticket prices. We used PCA and also PCA with ticket prices to see if there were any correlations. Finally, we did a heat map to compare each category with each other and against itself to see if there were correlations, see figure 1 for details. Some valuable correlations were found. Remain wary of ticket prices with resorts, as well as the chairs a resort has to the number of runs.

Figure 1:



During our preprocessing and training we implemented cross validation as well as a random forest generator. Big Mountain Resorts modeled price was so much higher due to what all the resort has to offer. It came in on the top 5 of chairs, snow maker ac, runs, and skiable terrain. By taking the average price of the test data we were able to get the price within $9 using the mean absolute error. Going further and imputing missing values and performing further machine learning only caused our results to be worse. The use of cross validation results highlights that assessing model performance in inherently open to variability. You'll get different results depending on the quirks of which points are in which fold. The random forest had worse results than the cross validation. Cross validating on the rf_grid_cv.best_estimator_, X_train, y_train, scoring='neg_mean_absolute_error'. Its estimation for the absolute mean error was 8 compared to the 9 previously done with just ticket prices so gave a better overall result. I think using the random forest regressor is the best route to go moving forward. I think shutting down three to five runs is the best route to go, the company would only lose $1.50 per ticket as well as only losing $2,500,000 in revenue. I think it would come as a surprise to the executives because they probably already feel they are priced high but the data clearly shows they aren’t charging enough. They could make use of it by using the random forest regression and comparing their prices with other states and adjusting the price to fit the market.

Figure 2:


Ticket price
Change ($)
Runs closed
Revenue
1e6
Change ($)
Runs closed