**Problem Statement**

Develop an estimated multiple linear regression model by January 31st 2024, to obtain a price value, which is the response variable for the model. The model’s multiple coefficients will be sorted in descending order to determine the impact a specific facility feature has on price. The higher the coefficient, the greater the impact that facility has on price. Management can use the information to decide which facilities and features are good candidates for further or reduced investments. The response variable, the price, is calculated by the model. The calculated price capitalizes on each facility, not on a percentage added to average prices of all resorts in the segment, as currently is the common practice. To calculate the break-even duration for a projected investment, management uses the modeled price, the cost of investments and the forecast revenues. A minimum 70% coefficient of determination for the model, is considered acceptable. The p-value for the model must be less than 0.05 to be statistically significant. The model is a tool for management to develop a data driven investment strategy to control costs, increase net profits and optimize distribution of investments at Big Mountain ski resort, thus capitalizing on its facilities.

**Context**

The Big Mountain Resort is a ski resort in Montana, with a unique location near the Glacier National Park. Currently, the pricing strategy is to charge a premium above the average price charged by other resorts in same market segment. To overcome the limitations of such strategy, Management needs a new pricing strategy approach that capitalizes on its existing facilities. Information on facilities such as chairlifts, trams, ski runs, terrain parks, skiable areas day and night and their attributes such as cost, speed, length, count, elevation, area, will be used to determine which facility and which features, attributes are most common in the 330 resorts across the US considered in this project and what are the average prices charged per facility. The data analysis will produce a list of facilities based on popular features, price charged, and specific attributes. This will enable management to decide in which facilities to invest more, and which facilities justify a higher price.

**Criteria for success**

* Clean data by January 15th, 2024 and identify facilities attributes that are fixed and variable, i.e.:
  + elevation is fixed variable,
  + a run drop is fixed and cannot be changed,
  + a run length is variable, can be extended depending on local geography.
* Obtain a model with relevant predictor variables by Jan 20th, 2024
* Use model’s coefficients to rank facilities’ features based on effect on price from highest to lowest by Jan 25th, 2024
* Use the model to calculate price recommendations when each of the facility features are increased by a factor of 1%, 5%, 10% and 20% and plot the results using confidence intervals to visualize price increase vs facilities’ features by January 30th, 2024

**Scope of solution space**

Use descriptive statistical analysis to describe data set. Perform a multiple linear regression analysis and an ANOVA statistic to find the relationship between multiple predictor variables, facilities attributes, and the response variable, price. The data analysis to be completed by January 31st, 2024 and produce a model output, an estimated equation with multiple coefficients corresponding to attributes of facilities. The reliability of the multiple linear regression to be checked by verifying assumptions of:

* linear relationship between predictor variables and response variable
* independence of residuals by performing a Durbin-Watson Test
* homoscedasticity of residuals by performing a Breusch-Pagan Test
* normality of residuals, by visually examining a quantile-quantile plot
* multicollinearity doesn’t exist among predictor variables, by calculating the VIF value (variance inflation factor) of each predictor variable

Release the model for continue use as new data becomes available and recalculate the coefficients on a yearly basis to include all seasons once.

**Constraints within solution space**

* An R-Square value close to minimum acceptable 0.70 indicates a lower reliability of the coefficients and thus the response variable - price - may be too high or too low.
* Calculated coefficients may induce management bias to opt for greater investments in fewer number of facilities.
* The increase in price as calculated by the model does not take in consideration the economic considerations such as price elasticity of demand.
* The ratio of percentage change in the number of facilities, services or the features demanded by customers to the percentage change in price: A marketing research study of how price change affects the supply and demand would be highly recommended to put realistic limits on how much the price can be increased, realistically.
* Variables outside of this model and which cannot be controlled or only partially offset, can affect the predicted efficiency of the model, i.e. snowfall variability (can be partially offset by snow making machine), macro-economic conditions, accommodation shortfalls during high demand, risk of avalanches, bad weather, lack of support from local community for expansion projects.

**Stakeholders to provide key insight**

* The Big Mountain CEO
* Jimmy Blackburn, Director of Operations
* Alesha Eisen, the Database Manager
* Head of Data Science team

**Key Data sources**

Single CSV file provided by database manager