**Data Wrangling:**

I started with importing the ski\_data csv file, then proceeded to glance over it’s data using df.info and df.head ().

A decent amount of missing values appeared while checking the .head() table, so i naturally felt as if that needed more investigation. The original csv file had 330 rows and 27 columns of data, I later discovered that a good chunk of data was missing , especially from the fastEight column, which has  166 missing values according to isnull().sum(). Other columns also had a number of missing values , most significantly to our objective are Adultweekday and adultweekend columns  (‘Ticket prices’) , less significant but extremely informative NightSkiing(provides an insight if customers can ski at night ) with  almost half of its values missing .

The database had some confusing region and state columns that were easy to navigate.  Also, some duplicates, but with some sorting, it appeared that our subject operated in 38 regions across 35 states, with New York and Michigan having the highest concentration of resorts followed by Colorado and California.

With some visual data analysis, using bar and box plots to discover resort concentration, as well as pricing distribution. Montana pricing was bounded between $25 and $75. Other states saw higher divergence in pricing, but the Montana average prices could not be placed using them.

Just over 82 % of resorts have no missing ticket prices, while 3% are missing 1 value and 14% missing both.

**EDA:**

While performing Exploratory Data Analysis, we tried to patronize and justify pricing based on state, yet that yielded insignificant results. Through a heatmap we were able to find some correlations between prices and features:

Positive Correlations from the data:

 -summit height

 -base elevation

 -vertical drop

-snow making.

     Negative Correlations:

 -lower chair numbers coincide with higher ticket prices

**Pre-Processing & Training:**

After inspecting the data and setting the R2, ss error, ss total, and mean squared error . The basics of the model was ready to be tested against scikit learn's preset metrics, then filling the missing data.

the data was scaled, then was tested on a linear regression model that resulted in explaining over 80% of the variance on the train set and 70% on the test set, it also suggested that the data was overfitted, the simple linear regression model was able to predict within 9% of the real price.

we defined the first pipeline and selected some features using SelectKBest which made things worse, apparently selecting a subset of features had an impact on the model's performance. k=10 was worse than all features, which rendered the question: what is the best K value?

So far we have built a pipeline that : imputed missing values,scaled the data , select K best features , and trained the linear regression model.Yet we need to build a cross validation technique to estimate the Model's performance.

Using GridSearchCV we built a cross validation model to determine the best K value, which showed that at k=8 the model will perform best. implementing the GridSearchCSV model on the existing pipeline with a random forest regression enabled us to understand features importance in price determination:

- runs

- fast quads -

-snow making

- skiable terrain

 performed higher on the importance scale. After assessing the performance of both models (Random forest, linear regression) the random forest performed linear regression in mean error, standard deviation error, and absolute error. Hence, the random forest model was selected.

**Modeling:**

when the Big Mountain resort was modeled, the modeled price was 95.88 and the actual price was 81.00

Features that came up as important in the modeling:

    vertical\_drop

    Snow Making\_ac

    total\_chairs

    fastQuads

    Runs

    LongestRun\_mi

    trams

    SkiableTerrain\_ac

When compared to other resorts in Montana , Big Mountain scores on the high end area covered by snow , total chairs ,skiable terrain , and fast quads, middle in vertical drop, number of runs which places it at the medium to high end of ticket pricing .

The model suggested that closing 1 run makes no difference in price, and closing 2 to 5 runs would reduce support for ticket price , it's important to note that the price effect is identical between 3 and 5 ive runs closed. while 6 or more runs closed would produce less support or ticket price.

An increase o 1 run and 150 drop and installing additional lift will increase the price by $1.61, based on visitor projection amounts to $2815217 ... note\* adding more acres of snow making makes no difference to the price.

extending the longest run by 0.2 miles and adding 4 acres of snow making makes no difference on the price as well.

In order to make a better cost cutting and price prediction model, more data is needed. for Example: operation cost in terms of energy consumption and labor cost would be extremely helpful in determining better prices and cost reduction. visitors purchase information such as meals, stay, quality of service.

The current price seems to be based on competition rather than qualification, from a glance it seemed that Big Mountain undervalued its services to attract a larger sum of visitors a year. There could be a strategy built within to increase revenue based on consumer spending on other amenities, yet it is unclear with the current data the reason behind the undervaluation.

The Business leaders will find my findings to be of interest to them since we both aim to develop a just pricing for the services offered , but as stated before I'm unable to build an accurate price prediction without knowing the intricacies of the business and the actual cost incurred providing services, sufficient data should be collected and provided to accurately make the business run at full efficiency. The business at the current stage could experiment with pricing provided and test consumer's willingness to pay.

 I would hope of a recurring relationship between the Business and me to which business development would be a key component to it, any new data is welcomed to calibrate the model and produce more accurate pricing systems. this model will be available for business analyst to replicate and use at Github.