Guided Capstone Project Report

Abstract

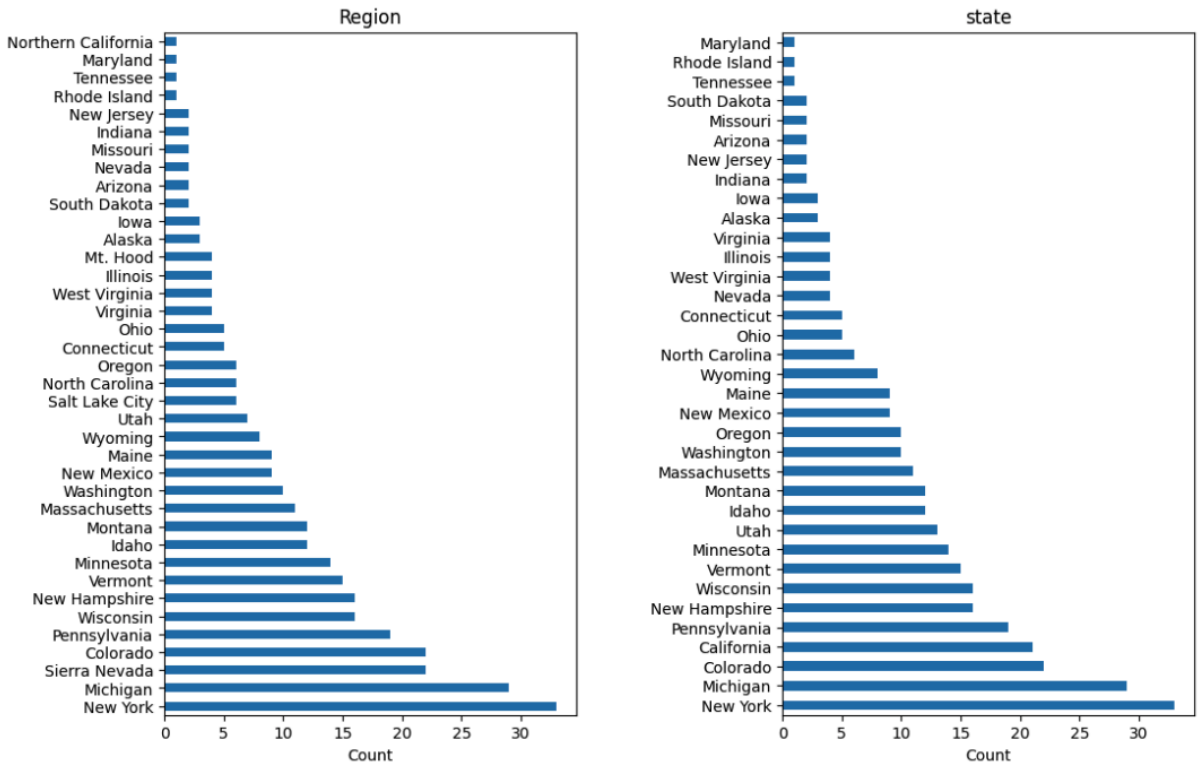
Big Mountain Resort recently installed an additional chair lift which increases operation costs by $1,540,000 this season. While Big Mountain Resort has been able to charge a premium above the average market price, it has limitations. The business wants to **maximize the capitalization on its facilities** by adjusting ticket prices and revising costs.

Focusing on the question, "*How can we maximize the capitalization on the facilities this season by setting the right price ticket and/or adjusting the operation cost?*", we conducted analysis on ski resort data, and would recommend increasing ticket price by **$1.99** while increasing the vertical drop by adding a run to a point 150 feet lower down and installing an additional chair lift to bring skiers back up. With this, we can expect **$3,474,638** revenue increase

Data Wrangling

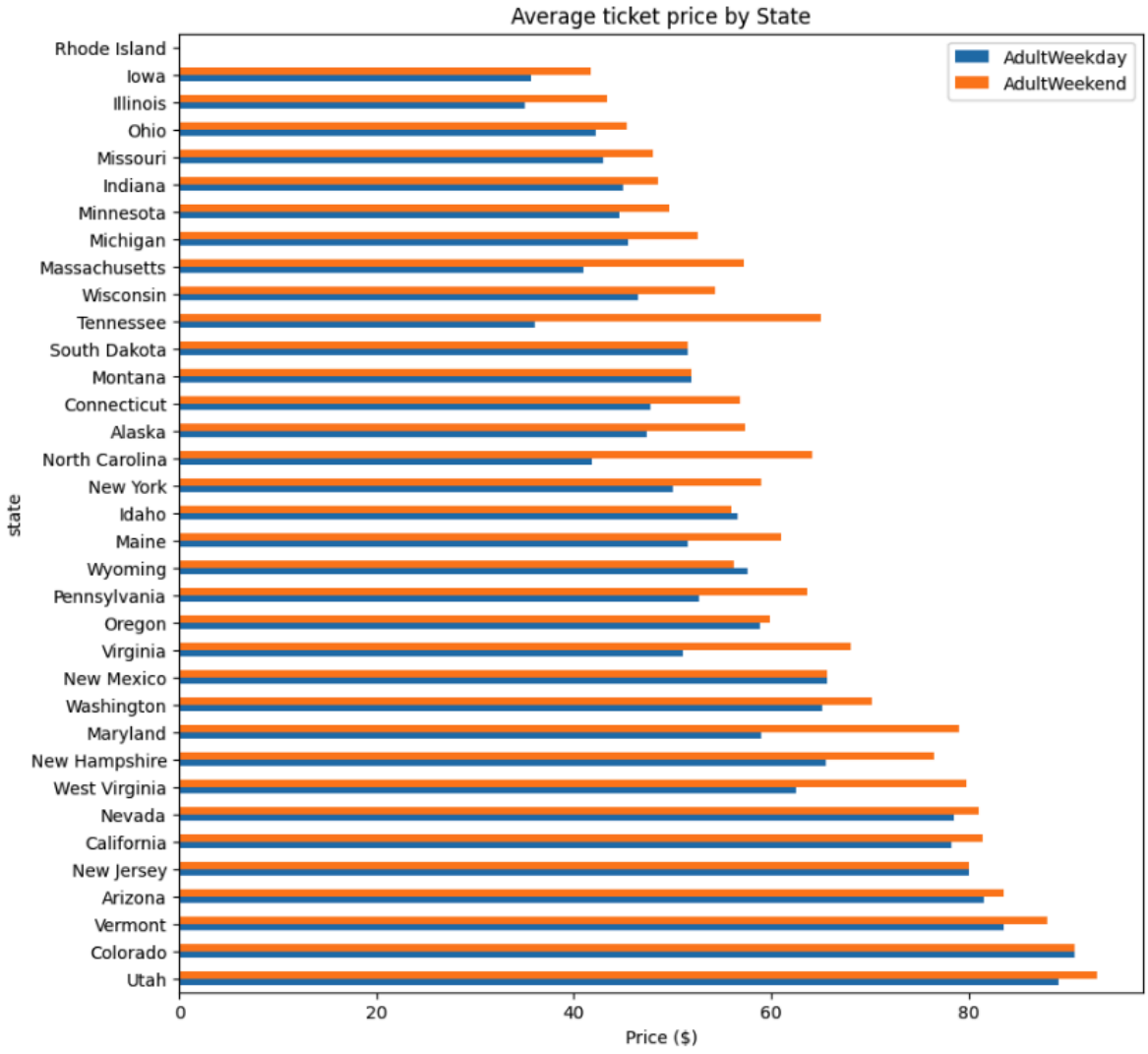
First thing first, we imported data from `ski_resort_data.csv`, and there we found two columns for prices `AdultWeekday` and `AdultWeekened`. We made sure that each row represents a unique resort by checking name+region combination, and also found multiple columns with missing values which includes `fastEight` where 50% of the data was missing and `NightSkiing_ac` where 43% of the data is missing.

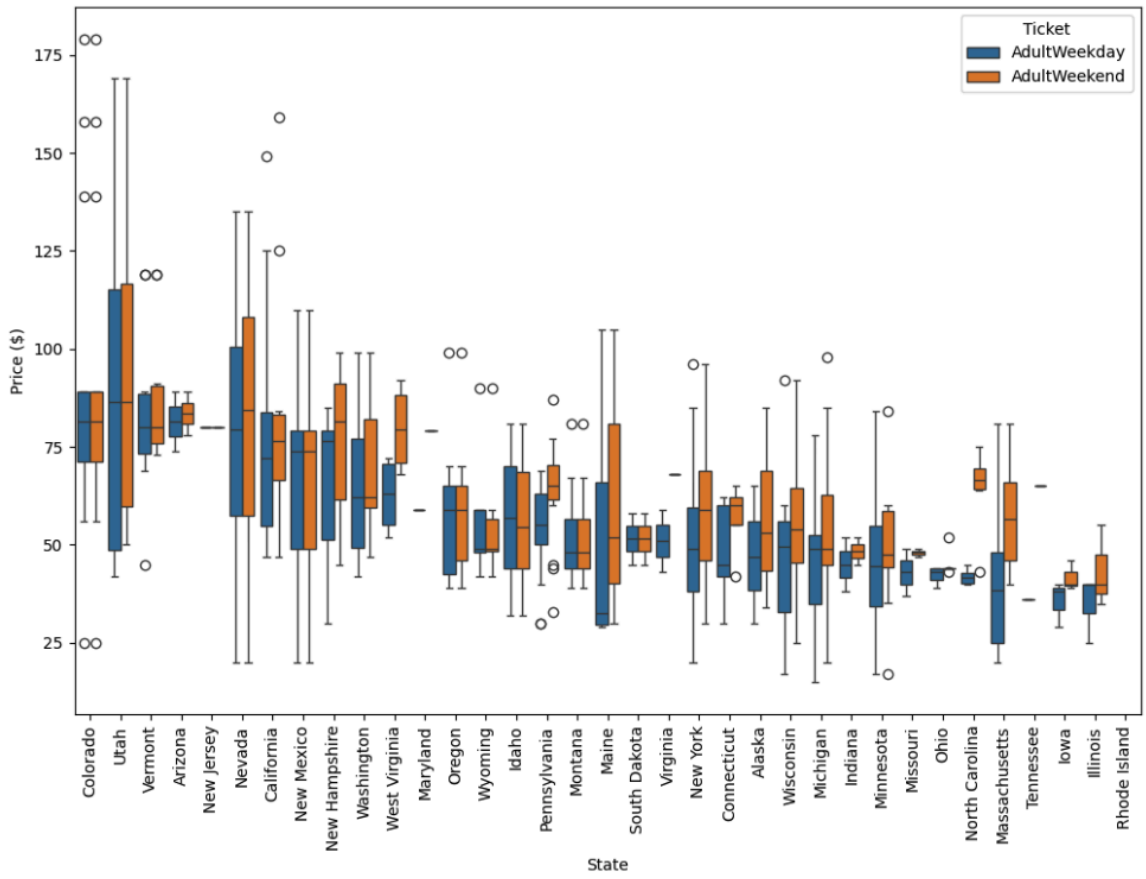
The resort data came with each resort's region and state. And this is how the distribution of resorts looks like.



Based on the chart, New York is accounting for the majority of resorts, and our target resort's location, Montana, comes in at 13th place.

We also looked into the ticket price data by State. Aside from some relatively expensive ticket prices in California, Colorado, and Utah, most prices appear to lie in a broad band from around 25 to over 100 dollars. Some States show more variability than others. Montana and South Dakota, for example, both show fairly small variability as well as matching weekend and weekday ticket prices. Nevada and Utah, on the other hand, show the most range in prices. Some States, notably North Carolina and Virginia, have weekend prices far higher than weekday prices.


Average ticket price by State
AdultWeekday
AdultWeekend
Rhode Island
Iowa
Illinois
Ohio
Missouri
Indiana
Minnesota
Michigan
Massachusetts
Wisconsin
Tennessee
South Dakota
Montana
Connecticut
Alaska
North Carolina
New York
Idaho
Maine
Wyoming
Pennsylvania
Oregon
Virginia
New Mexico
Washington
Maryland
New Hampshire
West Virginia
Nevada
California
New Jersey
Arizona
Vermont
Colorado
Utah
state
0
20
40
60
80
Price ($)



This leave us with 2 questions:

- What to do about the two types of ticket price?
- What to do about the state information?

Regarding the ticket price types, over 82% of resorts have no missing ticket price, 3% are missing one value, and 14% are missing both. We first dropped 14% of the row that had no price data. We also found that weekday and weekend prices in Montana are the same. While the weekend prices have the least missing values, we drop the weekday prices and then keep just the rows that have the weekend price.

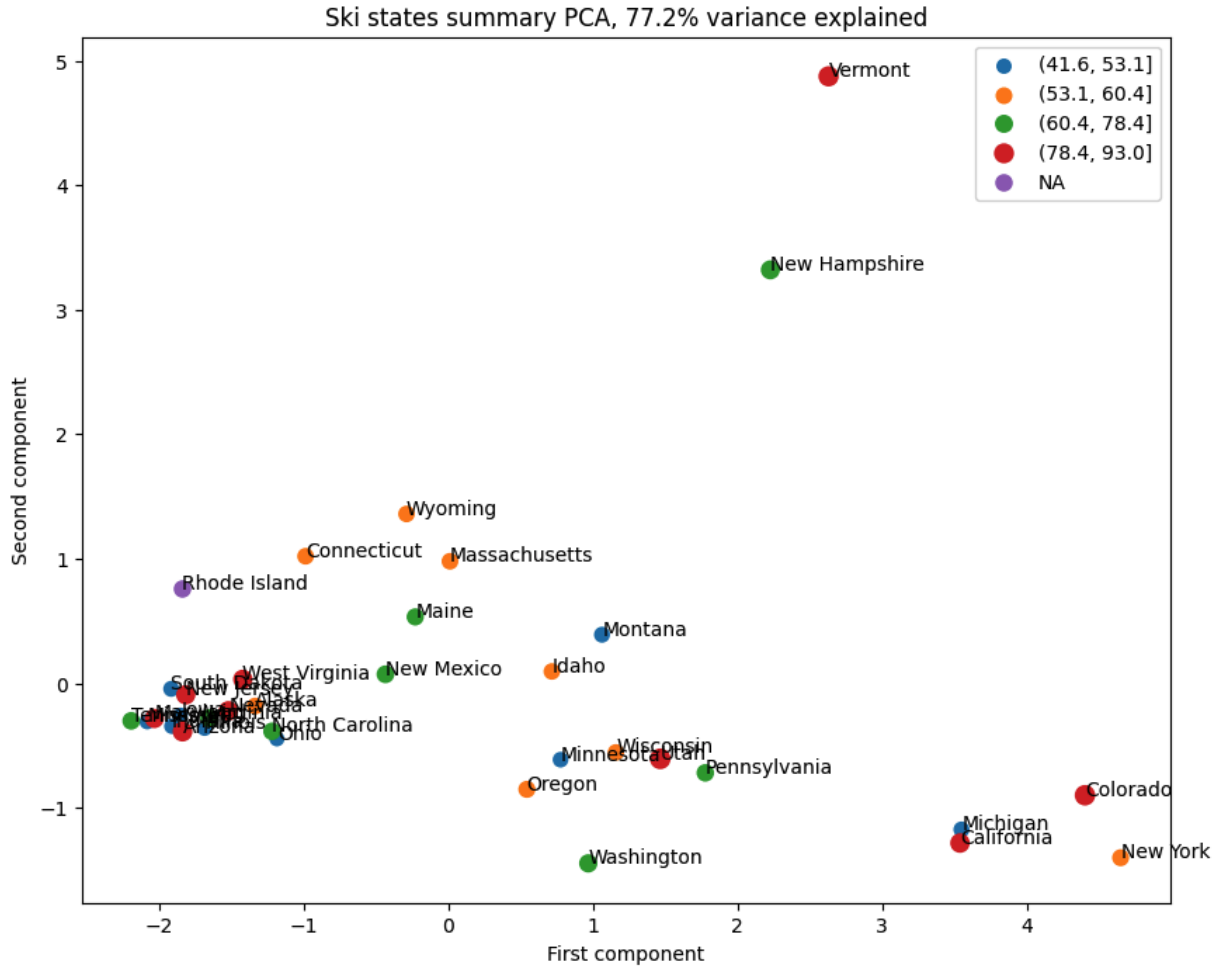
Regarding the state, in order to gain further insights, we first added some aggregated columns such as `resorts_per_state`, `state_total_skiable_area_ac`, etc. We also added the population and area data for the US states that we obtained from wikipedia.

Aside from the above, we also checked the distribution of features. There we fix the value on `SkiableTerrain_ac`, dropped the `fastEight` column (which was missing 50% of the values), and dropped the rows that had extreme values for `yearsOpen`.

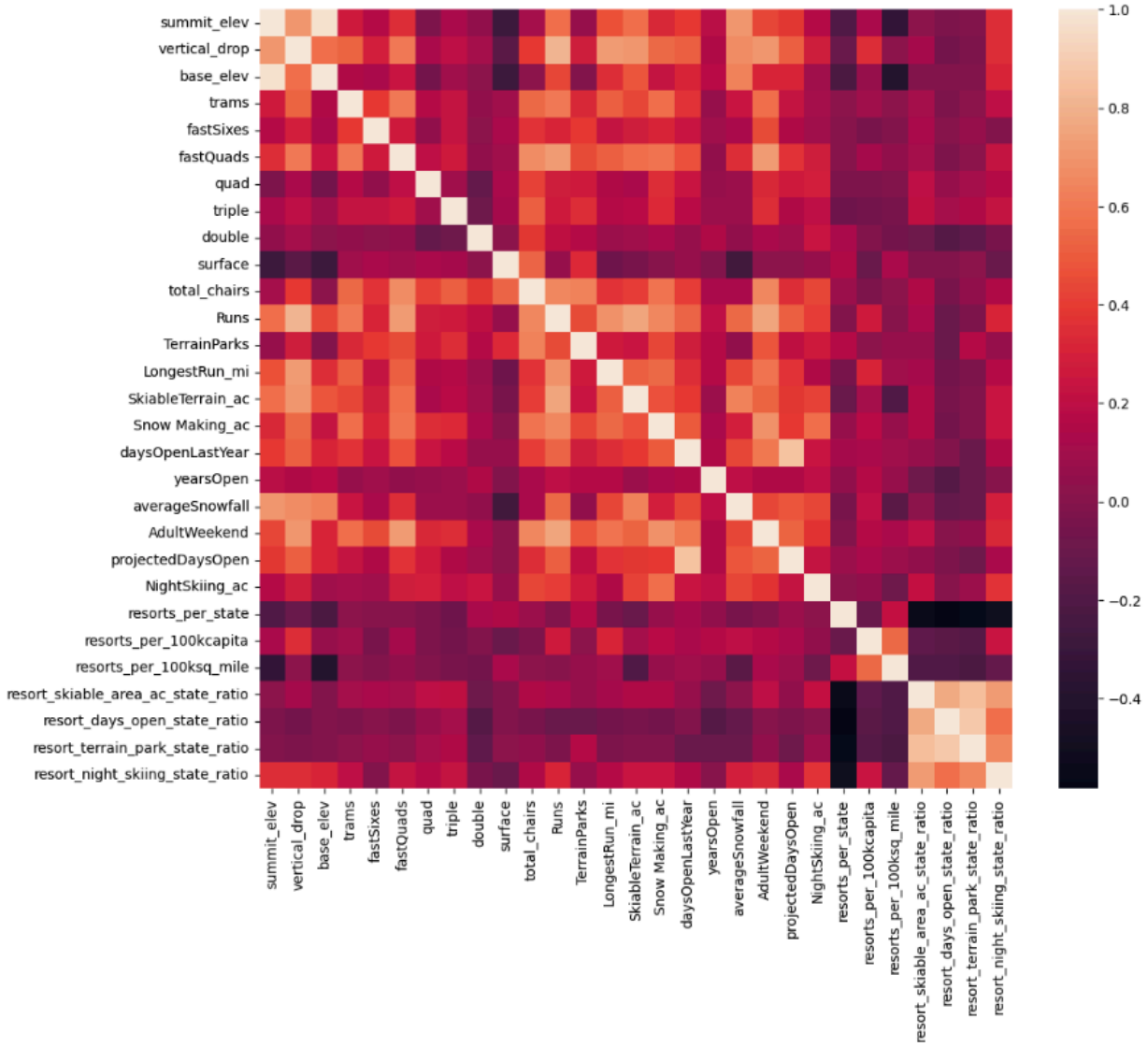
Finally, we saved the data in `ske_data_cleaned.cvs` and `state_summary.cvs`

Exploratory data Analysis

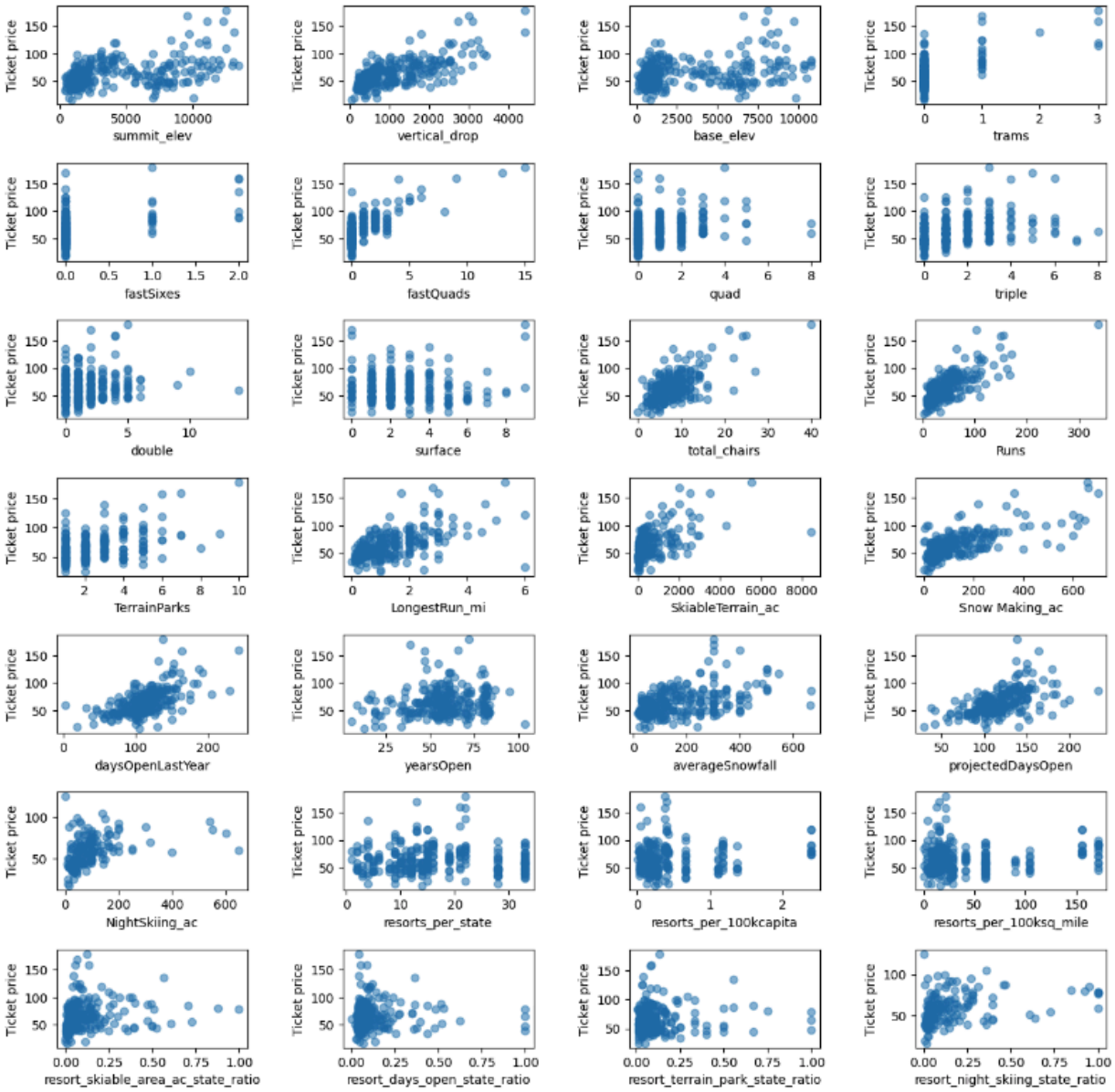
After some basic explorations, we conducted PCA. Where we first scaled the data, then fit the PA transformation, applied the transformation to the data to create the derived features, and finally, used the derived features to look for patterns in the data. We plotted the first two components from PCA and didn't see a pattern with price. Therefore we will be building a pricing model that considers all states together, without treating any one particularly specially.



We also merged the ski_data with the state_summary data, and created new features by putting each resort within the context of its state so we can understand better what share of states' skiing "assets" is accounted for by each resort. Then we generated a correlation heatmap. While focusing on the `AdultWeekend`, we see couple other features showed strong correlation, which includes `fastQuads`, `Runs`, `Snow Making_ac`, and `resort_night_skiing_state_ratio`.



We then create a series of scatterplots to really dive into how ticket price varies with other numeric features.



It seems that the more chairs a resort has to move people around, relative to the number of runs, ticket price rapidly plummets and stays low.

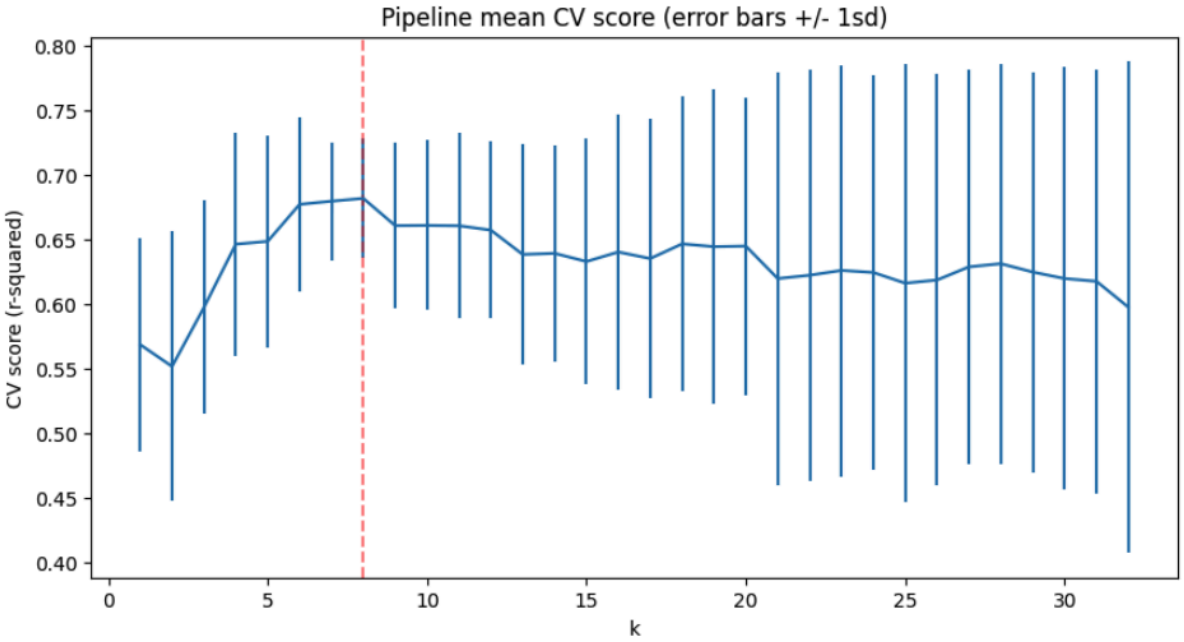
Model Preprocessing

Now after exploring the data, we want to build a model to predict the adult weekend ticket price.

First, we splitted the data into 70/30 train/test data and used mean as the baseline to compare the modal performance.

We tested with 2 modals, which were Linear Regression and Random Forest. For each modal, we evaluated with R-squared, Mean Absolute Error and Mean Square Error.

For Linear Regression, we preprocessed by imputed missing values, scaling feature data and selecting the best k features. We used cross-validation to get the best performing k, which was 8 features.



Another model we tried was the Random Forest Model. We preprocessed by imputing the missing values and scaling the feature data. Random Forest turned out to have a lower cross-validation mean absolute error by almost $1. It also exhibits less variability.

Modeling & Price recommendation

Given the Random Forest model we got from the preprocessing step, we refited the modal with the dataset.

We ran a couple scenario cases and the most practical one was to increase the vertical drop by adding a run to a point 150 feet lower down but requiring the installation of an additional chair lift to bring skiers back up. There, we can increase ticket prices by $1.99, and over the season, this could be expected to amount to $3,474,638.

We also predict that adding 2 acres of snow making cover to the above scenario makes no difference to the ticket price, so for the sake of cost saving, we suggest not adding it.

Also, increasing the longest run by .2 miles and guaranteeing its snow coverage by adding 4 acres of snow making capability does not make any difference to the ticket price.

Alternatively, we can close runs to save cost. We predict that closing one run makes no difference. Closing 2 and 3 reduces support for ticket price and can reduce revenue by $710,145 for 2 reduces and $1,166,666 for 3 reduces. Also, if we were to close down 3 runs, we may as well close down 4 or 5 as there's no further loss in ticket price. Increasing the closures down to 6 or more leads to a further large drop.

One call out here is that we were able to predict the ticket price based on feature changes, but since we don't have data for each of the facility maintenance cost data, we cannot accurately access the net revenue impact.

Future scope of work

Here are the following steps:

- Get data for operating cost of the facilities. Then combined with our price prediction senario, we can provide better expected revenue impact number
- Big Mountain prices are already higher compared to other resorts. Would like to understand what is the acceptable price range from the business context