SKI RESORT ANALYSIS

Cameron Hamilton
and
Sofia Scribner

December 11, 2024

Abstract

Determining the price of a single-day lift ticket at a ski resort involves looking at numerous factors, including geographical location, percentage of different run compositions (beginner, intermediate, advanced), elevation, total run length, and more. In this project, we aimed to predict the price of a single-day pass for a hypothetical ski resort. Using exploratory data analysis, variable selection, linear regression models, and inference, we identified several key factors that go into the of pricing of ski resorts and applied these insights to predict ticket costs for ski resorts across various countries. Our findings provide a data-driven framework for pricing strategies in the ski industry.

1 Introduction

Our company Peak Outdoors has been in the ski and rental shop business for quite some time. We've expanded around the globe becoming the one of the first global ski and rental shops that isn't tied to one brand. We have seen a lot of success come because of the increased popularity of skiing and snowboarding that social media has brought. With all of this increased capital we have decided to expand and create a new ski resort. Our two questions to help us create a new ski resort are: Can we predict the price of a day pass at a ski resort with certain specifications? And, what factors are the most important in determining day ticket prices? We want this new ski resort to be profitable. Our planning committe and general contractors have surveyed that we have enough capital to build a ski resort with the following specifications:

Elevation:

Base Elevation: 1,500 meters (4,921 feet)
Summit Elevation: 3,500 meters (11,483 feet)
Vertical Drop: 2,000 meters (6,562 feet)
Runs:
Total Runs: 150
Total Run Length: 114 km (71 miles)
Longest Run: 3.74 km (2.32 miles)
Lifts: Total number of lifts: 22 (total capacity: 30228)
Gondolas: 2 (Large capacity)
Chairlifts: 15 (Medium capacity)
Surface Lifts: 5 (Small capacity)

We are also going to determine which regions of the world would be best to place our ski resort. We have selected the most popular regions in the world for skiing (grouped by country) These regions are; The Alps, Asia, North America, and Scandinavia.

1.1 Data Description

We selected our data from a data database, Kaggle, and it can be found at this link:
https://www.kaggle.com/datasets/ulrikthygepedersen/ski-resorts.
In this particular data set there were 499 different resorts that had variables associated with them. There were a lot of extra variables in our data set, that after some EDA we determined were irrelevant to the questions we wanted to research, and wouldn't add much to the overall model. We initially created a model using these variables:

Country
Price (Originally in Euros, converted to Dollars)
Highest Point of elevation (meters)
Lowest Point of elevation (meters)
Number of Beginner Slopes
Number of Intermediate Slopes
Number of Difficult Slopes
Total Slopes
Longest Run (Kilometers)
Lift capacity (number of people per hour)

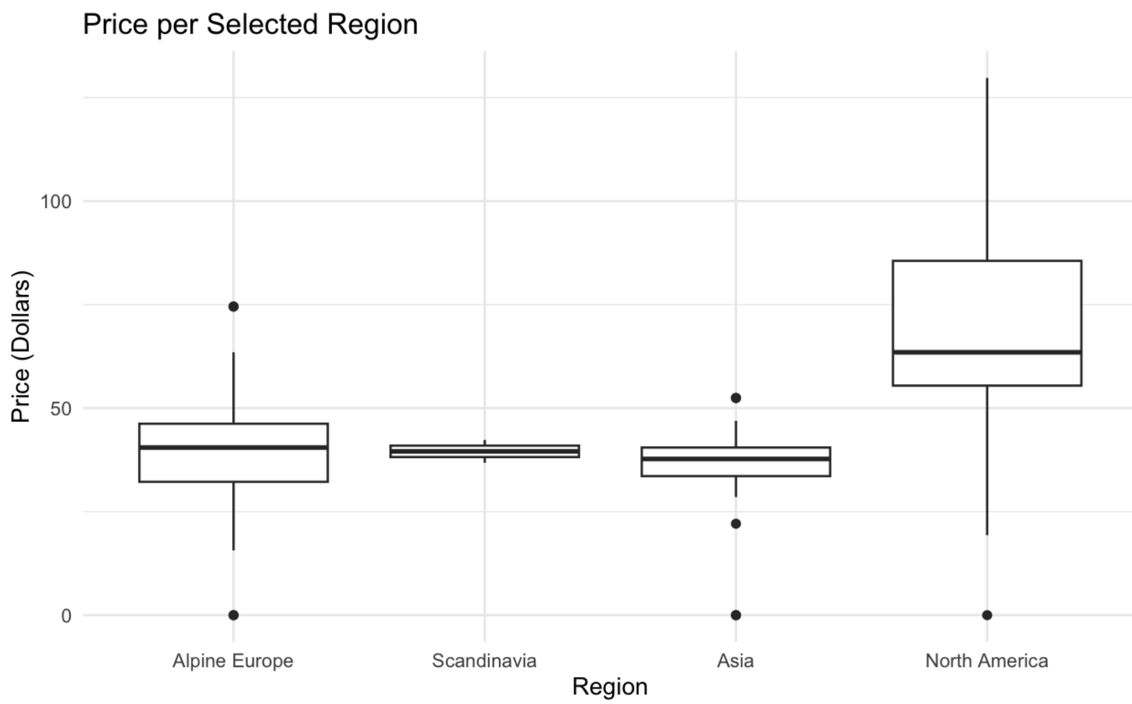

Figure 1: Boxplots of how price is different in each region that we have decided to study.

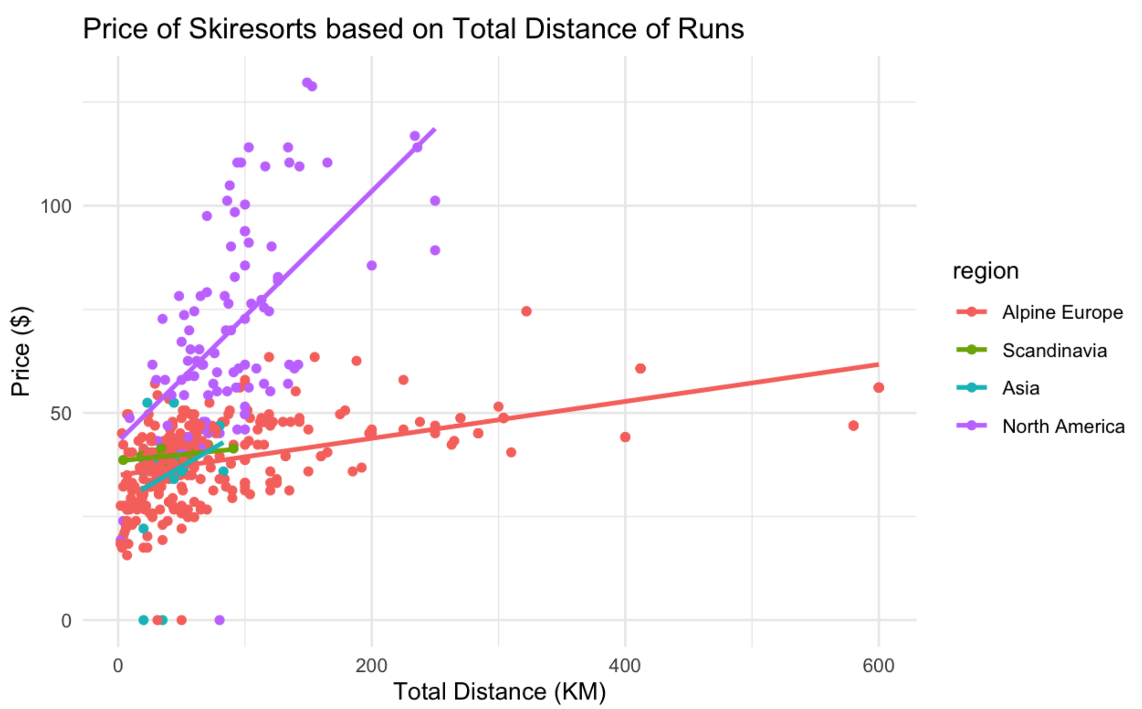

Figure 2: Scatterplot of Price of Ski Resort (in dollars) by total length of runs in ski resorts. The color corresponds to the region).

2 Model Selection and Validation

Because we are an American company, we transformed the Price variable using the 0.92 conversion rate for Euros to USD. We tested for multicollinearity, and realized that many of our original variables were nearly perfectly correlated with each other, such as the number of difficult slopes and the total number of slopes. To eliminate this multicollinearity issue, we transformed the slope counts by creating a new column recording the proportion of the slopes marked as difficult at each resort, rather than just the raw count data.

Most notably, rather than use every single country as a separate indicator variable, we focused on a few regions of interest: The Alps, Scandinavia, North America, and Asia. To do this, we created a new variable simply called "Region."

The final model we determined to use to answer our question is:

$$\begin{aligned} \text{PassPrice}_i = \beta_0 &+ \beta_1 \times \text{HighestPoint}_i \\ &+ \beta_2 \times \text{LongestRun}_i \\ &+ \beta_3 \times \text{LiftCapacity}_i \\ &+ \beta_4 \times \text{I}\left(\text{Region}_i = \text{NorthAmerica}\right) \\ &+ \beta_5 \times \text{I}\left(\text{Region}_i = \text{Asia}\right) \\ &+ \beta_6 \times \text{I}\left(\text{Region}_i = \text{Scandanavia}\right) \\ &+ \epsilon_i, \end{aligned} \tag{1}$$

where

$$\epsilon_i \sim N(0, \sigma^2).$$

This final model was selected using variable selection techniques. First, we removed the variables from the data that we felt were not relevant, and then we transformed the data as stated above.

After that, we tried a few different variable selection methods, including the stepwise method using AIC as a metric. We additionally tried using the LASSO model because it is an ideal method when a large number of explanatory variables are available. We used cross-validation to test and find the best LASSO model we could. The LASSO and stepwise methods produced very similar results. One major difference was that the effects for region were all reduced to 0 in the LASSO model, except for North America. We decided to go with the model including all the effects for region, because we are very interested in comparing different locations.

We also visually checked pairwise plots between region and all of the continuous variables in our model (see Appendix), in order to see if we could recognize an obvious need for an interaction term, which we did not.

3 Analyses, Results, and Interpretation

There were a lot of variables that were in the data that we found. In fact, there were a lot of data that were highly correlated with each other. So after we determined through variable selection which were most important. We fitted our model. Our fitted model as shown below in Table 1, helps us to better determine the biggest impacts to day pass ticket pricing.

Based on our model that we fit to this data set, our selected multiple linear regression model had 6 variables The highest point in elevation, the length of the longest run, total lift capacity in an hour. The other 3 were indicator variables that help us to better understand where to locate our ski resort.

Coefficient	Estimate	Confidence.Interval
(Intercept)	18.56761	(14.68223, 22.45298)
Highest.point	0.00622	(0.00446, 0.00798)
Longest.run	0.92165	(0.613, 1.23031)
Lift.capacity	0.00009	(6e-05, 0.00011)
regionScandinavia	10.37923	(4.29073, 16.46773)
regionAsia	2.59816	(-3.14882, 8.34514)
regionNorth America	32.34804	(29.43003, 35.26604)

Table 1: Coefficients and Confidence Intervals for Linear Model

Our analysis shows us that if there were no other factors influencing a ski pass, a ski pass would cost about \$18.57 (our intercept term) We can also say that we are 95% confident that this value is between \$14.68 and \$22.45, this highlights a solid baseline independent of specific features and regional factors.

All the other individual predictors; Highest point, longest run, and lift capacity all substantially impact the day price of a ski ticket. Elevation (Highest.point), tells us that for every additional meter in elevation, the price increases by .00622 cents. This isn't a very noticeable difference on a small scale (say by hundreds of feet), but most ski resorts (including ours) are thousands of feet above sea level. The Length of the run has a great impact on the price of a day ticket. For every kilometer that the longest run can be, there is a .92 cent increase. For our resort just for our longest run alone we can charge \$3.44 on each ticket. Finally, for each person on a ski lift at any given time we can charge \$2.72. All of these may seem relatively small but add up in the end to the overall ticket pricing.

The biggest impact on ticket costs though, is where our ski resort is located. Ski passes in Scandinavia are priced significantly higher than most other regions, as indicated by an estimate of \$10.38. This could be attributed to unique regional features or market conditions. Conversely, the Asia region, with an estimate of \$2.60, does not show a significant effect, implying that pricing in this region may be influenced by other unmeasured factors or that there is higher variability in prices within this region.

The most noticeable effect of ski ticket pricing however is our observed coefficient for North America. Just placing our ski resort in North America means that we would be able to charge \$32.35 more per ticket. indicating a substantial regional premium.

4 Conclusions

In our linear regression model presented above, we found the answers to our questions. If we wanted to price a Ski resort with the highest point of 3500 meters, a longest run of 3.74 kilometers, and a max lift capacity of 30228 riders it would depend where we wanted to put our ski resort. Here are the 95% prediction intervals for the price of a day pass in a resort with our specifications for each region.

Because of our analysis and the results above, we were able to answer our research questions. We learned that the most important factors in maximizing profit of a ski resort are: Highest point in elevation, the length of the longest run, how many people can travel on our lifts in 1 hour, and the region that the ski resort is in. If we were going solely off of profit we would obviously put our resort somewhere in North America. Since a resort

Region	Fit	Lower	Upper
Alpine Europe	40.14	15.74	64.54
Scandinavia	50.52	25.41	75.63
Asia	42.74	17.74	67.73
North America	72.49	48.01	96.96

Table 2: Prediction Intervals for Ski Resorts in Different Regions

with similar specs would have an average day ticket that costs $72.49. A Scandinavian resort would also help us to be more profitable with a ticket that costs $50.52. However, this leads us to one of our first drawbacks of this project. North America has the largest coefficient. Meaning, that no matter what specs we have for our ski resort, North America will always give us a higher day ticket cost. This could be cause for a future study where we would look at the different mountain regions in North America (Rockies, Appalachians, Sierra Nevada) to see if there is a price difference between those.

All in all, we as a company are going to move forward with building a ski resort in North America making sure that we make an emphasis on the elevation where we put our resort, the length of the longest run and the total lift capacity.

APPENDIX

Our LASSO regression results on how we selected our specific variables.

Coefficient	Estimate
(Intercept)	2.894592e+01
Highest.point	4.179358e-03
Lowest.point	.
Longest.run	4.173976e-01
Lift.capacity	2.755960e-05
NightskiingNo	.
NightskiingYes	.
Summer.skiingYes	.
regionScandinavia	.
regionAsia	.
regionNorth America	2.520433e+01
perc$_dif$	.

Table 3: Lasso Regression Resutls

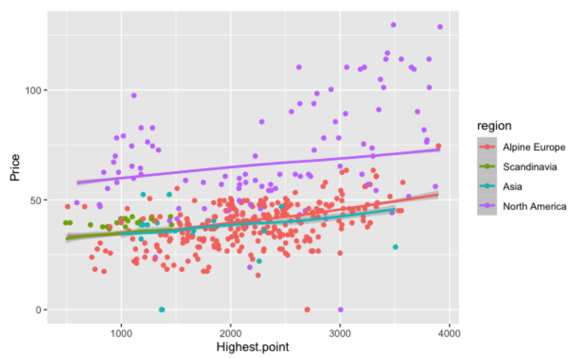

Figure 3: Interaction plot to better determine if we needed an interaction term between the variables. Since Alpine was so close to all the other regions, we determined that we would need one.