Estimating the Price of Skiing Hotels

Datasci 203: Statistical Analysis Lab 2

Atharva Mehendale, Leanna Chraghchian, Sally Fang

Contents

1 Introduction

For European ski hotels, the battle to attract guests has never been fiercer and more important. While pandemic lockdowns have ended, ski hotels in Europe are faced with the issue of an increased cost of fuel due to the conflicts between Russia and Ukraine. Additionally, a notable post-pandemic trend is a surge in short-term trips taken during the year without long-term planning. This “freestyle tourism” has brought new players, like Airbnb, to compete with ski hotels with their flexible options. The dilemma is whether ski hotels should reduce hours, charge more, or find better solutions. Where can hotels save money to stay competitive? The 14 major transformative projects taking place in the U.S. this winter also foreshadow a European wave to follow – as ski hotels compete on an international scale. The push for green technology and necessity of differentiation almost guarantees major future projects in European hotels.

For these reasons, we are analyzing data on 178 European ski hotels to focus on how price is affected by different factors. We acknowledge that these factors can be shared among hotels that are part of the same ski resort (share the same mountain). We will answer the following research question: How does the distance from the nearest lift affect the pricing of a hotel? The answer to this question will help guide future construction projects and allow for improvements to create new, more convenient ways to decrease the distance between the hotel and ski lifts (i.e. shuttles). Hotels can keep this study in mind when making decisions about future projects as it affects revenue and competitive position.

2 Data & Operationalization

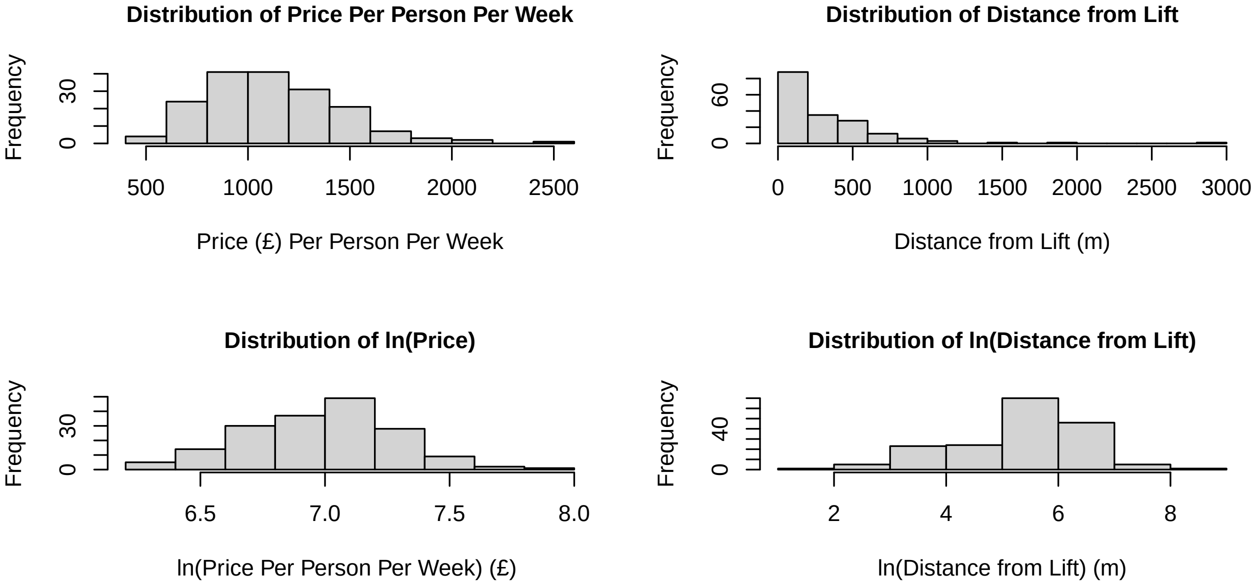
The data source for this study is an observational dataset of ski hotels within resorts across six countries in Europe, with columns describing the various amenities and offerings at each hotel. This data was compiled by Jack Lacey and posted as part of the public domain on Kaggle. Each row of the dataset represents a different ski hotel as part of a larger resort and details on its operations such as price per person per week, the number of lifts in the resort, and the total piste distance. We used 44% of the data given after filtering out rows that contained “unknown” and “Inf” in columns that we used for our model.

Our X and Y concepts were quite straightforwardly operationalized. The dataset included columns called distance_from_lift_(m) and price (£) that we used to operationalize X and Y, respectively. The first describes the smallest distance to a lift from a hotel as it is the distance to the nearest lift, which is in meters. The second describes the amount of money, in English pounds, a person would be spending per week. The “price of a hotel” is most precisely described by the price a person would pay for a length of time to stay at that hotel, so we felt that this variable makes the most sense for us to use.

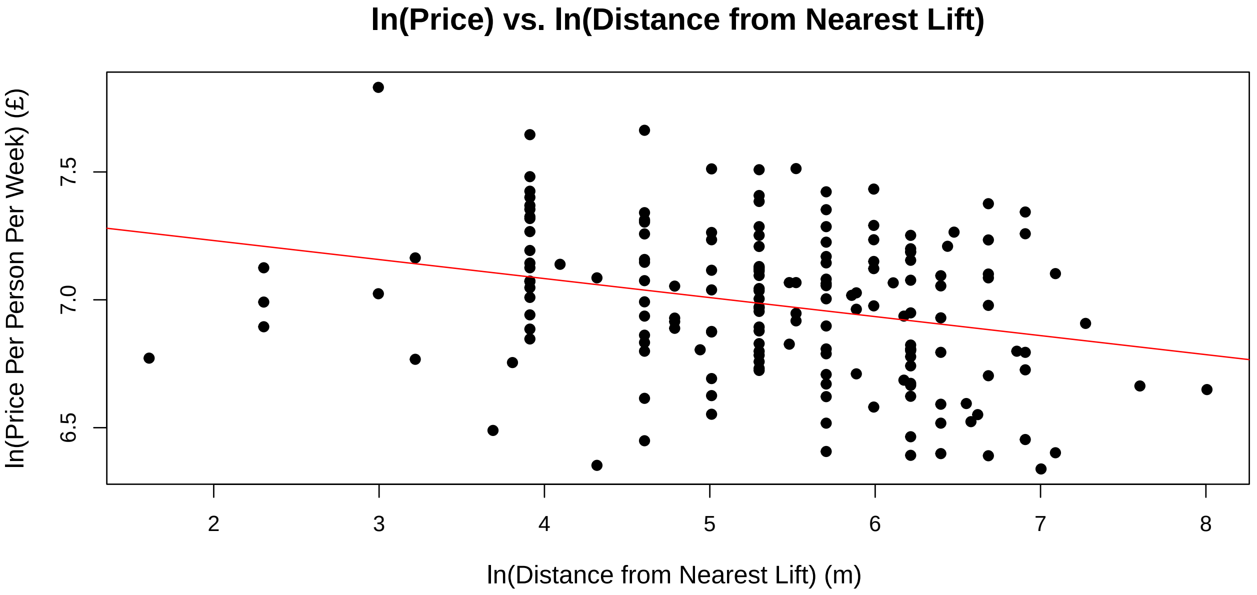
3 Explanation of Key Modeling Decisions

The following columns were excluded from our dataset: X (unique identifier), country, resort, hotel, link, and a series of columns on the snow recorded in 2020. X, country, resort, hotel, and link columns are removed from the data because the key questions we attempted to model do not require data at an individual level. We also dropped columns recording snowfall on slopes because we wanted to focus our analysis on features that hotel management can change. We also removed all null entries from our dataset. Specifically, any entry missing a key value from our model was excluded from our analysis. This leaves our dataset with a total of 13 columns and 175 rows after cleaning. After examining the distribution of our key variables of interest, log transformations were applied to the price of the hotel and the distance from the nearest lift.

As evident from Figure 1 below applying log transformations to both variables make the distributions more normally distributed. We believe a log-log transformation is appropriate in this case because both variables are always positive and span multiple orders of magnitude. We also believe that a log-log transformation yields meaningful percent changes in the topic of our analysis. We understand that the predicted price of a hotel is predicted to become negative after a certain point which introduces a limitation to our model. For this reason, we applied a log transformation to our model.



We also converted our X variable from factor-level data to a numeric type. This allows us to make an analysis without worrying about overfitting by interpreting a metric variable as a factor-level variable. Similar transformations from character type to numeric type were applied to other key variables of interest such as the column sleeps (number of people the hotel can sleep). We also created one additional variable, the ratio of total lifts to total slopes, computed using totalLifts and totalRuns, as a covariate to compare against our main variable of interest. Specifically, we hope this ratio describes how crowded a ski hotel can be.



We are interested in estimating the price of skiing hotels based on the distance the hotel is from the nearest lift. Our exploratory plot (Figure 2) suggests that this relationship is roughly linear and negative. Therefore, the model we regress on takes the form of

$$ln(price\ per\ person\ per\ week) = \beta_0 + \beta_1 \cdot ln(distance\ from\ nearest\ lift) + \mathbf{Z}\gamma$$

Where β_1 represents the percentage increase in price per person per week given 1 percentage decrease in distance from the nearest lift, $\mathbf{Z}$ represents additional covariates we are interested in exploring, and γ represents the coefficients of those covariates.

Several covariates were intentionally left out of our model, including the country the hotel is located in and records of snowfall in 2020. These covariates were excluded from our analysis because we wanted to focus on features the hotel management can change. Furthermore, including a categorical term like country reduces the precision of our dataset, and modeling snowfall (by taking an average of all records) gave us a statistically insignificant result. We believe that snowfall on slopes influences the price of a skiing hotel; however, it is likely that given the low precision of our dataset by including this variable, we are unable to evaluate the true relationship that this covariate casts on our outcome variable.

Table 1: Estimated Regressions

	Dependent variable:		
	ln(Price Per Person Per Week) (£)		
	(1)	(2)	(3)
ln(Distance from Lift) (m)	-0.074^{***} (0.019)	-0.070^{***} (0.020)	-0.061^{***} (0.020)
Total Lifts		0.002^{***} (0.001)	-0.004^{**} (0.002)
Altitude (m)		-0.00004 (0.00005)	-0.00002 (0.00005)
Lift to Run Ratio		-0.202^{***} (0.070)	-0.108 (0.084)
Total Piste Length (km)		-0.0002 (0.0002)	-0.0002 (0.0002)
Available Beds		-0.0001 (0.0002)	0.00003 (0.0001)
Black Diamond Slopes			0.003 (0.003)
Chairlifts			0.008^{**} (0.004)
Gondolas			0.015^{***} (0.004)
Constant	7.380^{***} (0.106)	7.487^{***} (0.160)	7.335^{***} (0.163)
Observations	175	175	175
R^2	0.078	0.173	0.256
Adjusted R^2	0.073	0.144	0.215
Residual Std. Error	0.281 (df = 173)	0.270 (df = 168)	0.258 (df = 165)
F Statistic	14.720^{***} (df = 1; 173)	5.869^{***} (df = 6; 168)	6.302^{***} (df = 9; 165)
Note:			*p<0.1; **p<0.05; ***p<0.01

4 Discussion of Results

According to the table above (Table 1), the coefficient of interest was highly statistically significant (at a significance level of 1%) across all three modeled regressions. The coefficient itself varied from -0.061 to -0.074. Considering model 3's result, this means a 1% decrease in distance from the nearest lift increases the price of the hotel by 0.06%. A more practical interpretation of this model is to consider a hypothetical scenario in which a hotel remodels and constructs an additional lift that decreases the distance by 25%, this would increase the price of the hotel by 1.75%.

This result indicated that if hotel management is considering remodeling the hotel, they should consider building additional lifts closer to the hotel site. In future constructions, hotel management should also consider having lifts as close to the hotel as possible. Several practical limitations we acknowledged include geographical constraints (can't move the hotel closer to an existing lift, there's a limit to how close the lift be constructed) and operational constraints.

One key observation we made is that all statistically significant results are related to lifts (distance to lift, type of lifts). It's also interesting to us that the number of beds available at the hotel did not cast a statistically significant result on the price of the hotel. The less important results we concluded are the results from the altitude and total piste length since they yield statistically insignificant results and are also features that hotels can't really change.

5 Discussion of Limitations

In our evaluation of the large sample model assumptions, we acknowledged that geographical clustering might pose problems for our analysis. We partially accounted for this by only looking at European countries, hoping to minimize potential variations across countries. Furthermore, multiple hotels do belong to the same resort so there's a potential issue with geographical clustering. Potential temporal dependency was addressed since all hotel prices are collected in 2020. We accounted for the assumption to have a unique BLP by applying log transformations on key variables. As evident in Figure 1, without the log transformation, the distance to the nearest lift variable posed a heavy-tailed distribution. After applying the log transformation, the distributions are visually more normally distributed. We also made sure that no perfect collinearity exists since no variable was dropped from our regressions.

An important omitted variable is a shuttle service variable noting if the hotel already has a shuttle service. This variable would have a positive correlation with the price variable since having shuttle services is a luxury and would thus increase the price of the hotel. It would have a negative correlation with the lift distance variable, since having shuttle services would decrease the perceived distance from the hotel to the lift. This would lead to a negative omitted variable bias that points away from zero, thus making the model overconfident. Another important omitted variable is a restrictions variable referring to the government or geographic restrictions the hotel needs to abide by. For example, a hotel cannot build a ski lift because it crosses into a residential zone. This variable would have a negative correlation with the price variable since more restrictions would limit the hotel and would thus decrease the price. It would have a positive correlation with the lift distance variable, since having more governmental restrictions could stop the construction of new lifts and geographic restrictions could increase the distance from the hotel to the lift. This leads to a negative omitted variable bias that points away from zero, thus making the model overconfident.

We believe that the possibility of reverse causality is slim. Because once a hotel is built somewhere the decision is sunk and over time they will find a price that attracts people to the hotel. Given the operational and physical constraints of moving a constructed hotel towards a lift, there's little chance of reverse causality of the price of the hotel affecting the distance to the nearest lift.

The number of gondolas could also be affected by the distance from the nearest lift because gondolas serve two purposes: 1. transporting people to the base of the mountain and 2. transporting people to the top of the mountain. This is potentially an outcome variable on the right-hand side of our model because a greater distance from the nearest ski lift can lead to more gondolas that transport people from the hotel to ski lifts at the base of the mountain. If we take away the gondolas variable, the distance from the lift variable will absorb the sign of the more positive gondolas coefficient, and thus move towards zero and making the model underconfident.

6 Conclusion

In this study, we quantified the relationship between the price that a hotel guest would have to pay on a weekly basis and the distance that the hotel is to the nearest lift. With a 1% decrease in distance from the nearest lift, the price of the hotel as paid by the average hotel customer would increase by somewhere between 0.061 and 0.074. In addition, we found that only lifts, as compared to other hotel features like occupant availability, altitude, and the number of slopes, had statistically significant results, meaning that our initial hypothesis was not only right in which variable we selected to have the greatest impact, but also right in that we need only that variable for statistically significant results.

The next steps for this research project would be to expand the dataset to not only include European nations but all locations that have ski hotels available. With this data available, we would be able to duplicate the modeling efforts and therein ensure that our results make sense on a global scale, not just on a subset of hotels in one geographical area. The significance of this work lies in that ski hotels will be able to better plan their layouts to maximize ski lifts near there, especially for hotels that are currently being designed. In addition, popular ski hotels can 1) make more lifts available for their customers and 2) make lifts more accessible in order to increase their prices. Customers will also benefit from having an easier time getting to the lifts which will ultimately allow them to maximize their time on the slopes - everybody wins.