Research on Factors that affect ski resort pricing

IS 507 Final Journal Article

Xinyuan Chen(xc44@illinois.edu), Yizhan Xue (netID: yizhanx2@illinois.edu)

A. Abstract

Skiing has always been a very popular winter entertainment in Europe and the United States. In recent years, many investors have begun to lock the market in Asia and invest in the construction of ski resorts in Asia, like northern China. Thanks to its unique climate and environmental advantages, northern China even successfully hosted the Winter Olympics this year, attracting a large number of tourists for ski entertainment. However, pricing for ski resorts in different regions and sizes varies widely. Tourists always want a high-quality skiing-vacation experience at the least cost. What factors affect resorts' ticket prices and tourists' choices most? Aside from pricing, what factors should investors be most aware of when managing a ski resort are the research questions we are focusing on. To explain the questions, three statistical models are introduced in the research: Multiple Linear Regression (MLR), Principal Component Analysis (PCA), and Logistic Regression (LR). Respectively, the MLR model shows that the slopes, lifting ability, and altitude are the factors that influence ticket prices most; the PCA model shows resort facilities and natural environment as the two most important dimensions for the research; the LR model identifies continent, slopes, nightskiing, a snowpark and lifting ability as the factors that influence the likelihood of affecting resort ticket prices. Finally, the study group will discuss the advantages and disadvantages of each model and suggest that operators need to follow up on the maintenance and renewal of ski facilities if they want to attract more tourists.

For tourists, when looking at resort tickets, the environment, climate, facilities, and service friendliness of the resort all need to be considered.

B. Introduction

According to the survey by Vanat, L. (2021), 67 countries around the world offering fully equipped outdoor ski resorts in 2021. If counting indoor ski resorts, there are about 50 indoor snow centers in 20 countries around the world. Approximately only 100 ski resorts have a mountaineering and other activities such as dry runs. An estimation of 2,000 ski resorts in the world is with more than 4 lifts. In addition to the major ski resorts with the highest number of skiers, there are other smaller ski resorts whose ski industries are still developing. The latest rising ski resorts are in Eastern Europe and China but some other countries, such as Cyprus, Greece, India, Iran, Israel, Lebanon, Lesotho, Morocco, New Zealand, Pakistan, South Africa, Turkey, etc. There are also ski resorts developing rapidly in the middle.

In the comparison of market share, the Alps are the biggest ski destination in the world, reports of 43% of skier visits worldwide reference the statistic from Berard-Chenu, L. (2021). The second most popular region is America (mostly North America), accounting for 21% of skier visits worldwide. Japan and China consist of Asia’s ski market, but gradually Japan’s market share is replaced by growing China. In Feb 2022, China held the Winter Olympic Games successfully, a series of ensuing policies motivated the development of the country's ski industry and Chinese tourists’ attention to skiing.

In the future, competition in the ski resort industry between Europe and America continues to be fierce, but the Asian market is gradually becoming a strong competitor. In the next few years, these three continents will hold the majority of the global ski market share.

C. Literature Review

There are many advantages to skiing sports, like the excitement to challenge humans' limits, boosting moods, and making the body more flexible. To reach their expectations on skiing, tourists are strict in the ski resort selection. Based on research about altitude sickness and skiing given by Hatzenbuehler, J. et al. (2009), over half of the respondents have altitude sickness concerns before skiing, and 40% of respondents will refer to the highest altitude when choosing a ski resort. Another concern for tourists is the risk of skiing injuries. According to the ski injury, risk assessment model created by Boris, D. et al. (2020), the injury risk is highly correlated with the region, due to the climate factor, and terrain that some slope angles are challenging for beginners. Hence, consumers also care about the potential risk of getting injured, like whether the climate is stable, whether the slopes are friendly for first learners, etc.

Some researchers also pick meteorology as their research target. In the research on environmental impact on skiing speed conducted by Carus, L. et al. (2021), in extreme weather, such as strong winds, poor visibility, and poor lighting, skiers generally underestimate their skiing speed, resulting in reduced reflexes and increased risk of injury. However, it just explores the impact of extreme weather on skiing and does not calculate a specific coefficient to explain how severe the impact of extreme weather on skiing is. Besides, Carus' study only takes into account the impact of meteorological factors on the safety of skiing, without other environmental

factors. Many scholars have researched investing in ski resorts and have given pertinent suggestions. Snow conditions have a positive effect on attracting investment. The higher the quality of the snow, the more tourists, and therefore more investment, can be attracted. Hence, they mind ski resort operators investing more in artificial snow. However, this conclusion is too one-sided, considering only improving the quality of artificial snow, and ignoring other safety factors. Therefore, in this research, more comprehensive factors, such as other facilities on the site and the maintenance mechanism of ski equipment, will be included in the impact of skiing safety for general considerations.

D. Methods

We collect data from the Ski-resort-stats website (2020), which provides data about resort information, lift information, slope information, and snow information of ski resorts all over the world. After data preprocessing, we get our dataset with 512 instances and 19 variables. Each instance is a line containing information about one ski resort. The 19 variables include Ski pass price for adults, different kinds of slopes number, different kinds of lifts number, and so on.

In the data preprocessing, we looked at the correlation matrix of the variables, identified two of them with highly aliased coefficients, then removed them. After that, we calculate the VIF based on the current variables, only four out of 19 variables have VIF greater than 5 and no variable has VIF greater than 10. Thus the multiple collinearities of variables are acceptable.

The models are chosen in the research. The first model is Multiple Linear Regression (MLR). Our goal of the task is to find a linear fitting function of the “Day Pass Price for Adult” variable

and find the most important independent variables that can affect the dependent variable. The basic MLR formula is shown as: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_i X_i$ referenced the Eberly, L. E. (2007).

The second model is Principal Component Analysis (PCA). In this project, PCA is used for identifying the most important dimensions among the factors that impact adult ticket prices. The basic PCA formula is shown in plot 1 referenced the Jolliffe, I. T., & Cadima, J. (2016).

The third model is Logistic Regression(LR). LR is a Machine Learning algorithm that is used for testing the hypothesis of relationships between a categorical dependent variable and multiple independent variables. Based on the correlation coefficients the model outputs, analysts can create a formula that calculates the likelihood of the researched condition happening. The basic LR formula is shown in plot 2 referenced the Peng, C. (2002).

E. Discussion and Results

1. Multiple Linear Regression (MLR)

Firstly we implement Multiple Linear Regression (MLR) to our dataset. We need to make the assumptions of the model:

- A linear relationship between the outcome variable and the independent variables.
- Multivariate Normality.
- No multicollinearity, which is always ensured by checking the VIF.
- Homoscedasticity.

We need to make sure all assumptions are satisfied, so we have the following steps to modify our model.

We draw the scatter plots for all independent variables with the dependent variable separately and list them below. The variables like Continent and Child Friendly etc. obviously do not have linearity with dependent variable Ski Pass Price Adult from scatterplots. We could remove them from multiple linear regression tasks to make sure of the linear relationship assumption. (Please review the scatterplots in *plot 3*.) We remove 5 variables without linear relationship with the dependent variable: Continent, Child.friendly, Snowparks, Nightskiing, Summer.skiing. Totally 13 independent variables left.

By checking the VIF, 3 out of 13 variables have VIF greater than 5 and no variable has VIF greater than 10. It means only 3 variables have weak multicollinearity and other variables have not multicollinearity. Therefore, we can assume that the data set satisfies the assumption of no multicollinearity (Please review the VIF results in *plot 4.)*

To check the assumptions, a diagnostic plots for manual regression is shown in *plot 5*:

- In the “residuals vs Fitted” plot, the blue line is not a horizontal line without patterns, so it means the model is not totally satisfied with the linear relationship even if we try to remove some of the variables.
- In the “Normal Q-Q” plot, the left bottom and right upper data points are not in the diagonal, so it means the model is not totally satisfied with normality (the residuals are normally distributed).

- In the “Scale-Location” plot, the blue line is not a horizontal line with equal spread points, so it means the model is not totally satisfied with homoscedasticity.

  With the linear relationship check, VIF check and assumption check, the dataset is not totally satisfied with the regression task. However, using multiple linear regression to predict a reasonable price for a ski resort based on the ski resort attribute is an important application for our custom. We decide to build a multiple linear regression model with stepwise equation following and show the model performance.

We build a multiple linear regression model with stepwise equations based on 10 independent variables and 1 dependent variable described above. The output of the regressions shown in the *plot 6*.

Because Multiple R-squared values and Adjusted R-squared values are different within one percent, there is no obvious overfit in the stepwise regression.

Suppose H0: all beta coefficients are equal to zero. H1: at least one beta coefficient is significantly different from zero. With p-value = 2.2e-16 < 0.05, reject H0 and accept H1. As the Pr(>|t|), the Difficult.slopes, Surface.lift.etc., Lowest.point, Longest.run, Chairlifts.etc. are significantly different from zero at the .05 level. We decide to select these five variables as independent variables to predict the Day Pass Price of Adult in our regression model.

Referring to results in *plot 7*, suppose H0: all beta coefficients are equal to zero. H1: at least one beta coefficient is significantly different from zero. With p-value = 2.2e-16 < 0.05, reject H0 and accept H1. As the Pr(>|t|), all variables are significantly different from zero at .05 level. Because the model p-value and the p-value of all individual predictor variables are less than 0.05, the linear model to be statistically significant with 0.05 level. With the statistical significance, the Residual Standard Error is 14.5, the Multiple R-squared is 0.5823, and the F-statistic is 46.37.

The dependent (response) variable that has not sufficiently been explained by this model as the R-squared < 0.7. The reason comes from the assumptions of the multiply linear regression model that are not totally satisfied. For example, there is not a strong linear relationship between the independent variables and the dependent variable, this nonlinear relationship performs not good enough when fitted with a linear model. However, our regression model can still be used as a baseline model to suggest a reliable adult daily ticket price for ski resort managers to adjust their pricing strategy based on the number of difficult slopes, number of surface lift, the lowest point, the longest run and the number of chairlifts in the ski resort.

Overall, the formula that determines the resorts ticket prices is:

$$Y_{ticket\text{-}price} = 30.231 + 0.558*\text{Difficult.slopes} - 0.585*\text{Surface.lift.etc} + 7.061*\text{Lowest.point} + 1.013*\text{Longest.run} + 0.378*\text{Charlifts.etc}$$

2. Principal Component Analysis (PCA)

We need to determine the number of components firstly. Two methods can be used to determine the number of components for the PCA model: the Kaiser-Meyer-Olkin (KMO) method and the Scree Plot Knee method.

– *The KMO method:*

Using the variance bar chart, the model puts a horizontal line at 1 in the picture, and any factors that the variance bar exceeds the horizontal line can be chosen as the determinant of components. Given the KMO visualized model below, 6 factors exceed/reach the horizontal line, hence the number of components is 6. (*plot 8*)

– *Knee of Scree Plot Method*:

A scree chart shows the variances (eigenvalues) on the y-axis and the number of dimensions on the x-axis. Variances are descending and sorted in the chart, with a broken link connecting them. The dimensions in front of the first elbow give the most information, hence the number of dimensions before the first elbow determines the number of components in the PCA model. In the chart below, there are three dimensions before the first elbow, hence the number of determinants is three. (*plot 9*)

Now that the numbers of components recommended by the KMO method and the Knee method are different. Given the number of total factors in this dataset is only 20, 3 components might be more suitable for the project.

As recommended above, the number of rotate components is 3, and the cutoff index is 0.4, the PCA model presents *plot 10*:

In the PCA model, the first rotate component consists of Beginner Slope, Intermediate Slope, Surface Lift, Chairlift, Gondola, Lift Capacity, and difficult slopes. Beginner, Intermediate and Difficult Slope belong to the category 'Slope', Surface Lift, Chairlift, Gondola, and Lift Capacity belongs to the category 'Lifting Ability'. The first rotate component gives the most information, it can be concluded that slope conditions and lifting capacity are the strongest factors that influence the tourists' skiing feeling. The factors can be characterized as 'Resort Facilities'. All indexes are positive, representing a positive influence on the research topic. The formula of RC1 is:

$$Y_{tourist\text{-}feeling} = 0.907*Beginner.slopes + 0.832*Intermediate.slopes + 0.885*Surface.lift.etc. + 0.852*Chairlifts.etc. + 0.840*Gondola.etc. + 0.552*Lift.capacity + 0.489*Difficult.slopes$$

The second rotate component consisted of Continent, Highest point, Lowest point, Adult Pass Ticket Price, and Difficult Slopes. The categories of these factors are duplicates, roughly they could be characterized as 'Natural Environmental Factor'. All indexes are positive but not as high as RC1, which represents a moderate positive impact on the skiing experience. The third rotation component consisted of Child Friendly, Summer Skiing, Avg snow in the last 5 seasons, and Best Weeks. They can be categorized as 'Resort Service'. Only summer skiing shows a negative relationship, all other factors have a positive relationship with the skiing experience.

3. Logistic Regression Model

In the original dataframe, the VIF of some factors is relatively large (but are below 10), and there is a problem of moderate collinearity among these factors. If all of them are used in the logistic regression model, the model will produce very exaggerated odds ratios, making the model overfit. Therefore, in the case where the vif of all factors does not exceed 10, factors in RC1 and RC2 are subset from the PCA model to form a new dataframe.

Moreover, The logistic regression model aims at searching relationships for binary dependent variables. Hence, the factor 'Ski pass prices adult' should be revalued in '0/1' format in the new dataframe. Prices no more than the median 70 are revalued as '0', representing a low price, while prices higher than '70' are revalued as '1', which represents a high price. The model is rendered in *plot 11*.

Given the significance level at 0.05, the factor of which p-value less than 0.05 is regarded as a significant factor that has a significant relationship with the dependent variable (ticket price). Factors Continent, Intermediate.slopes, Chairlifts.etc, have significant relationship with the resort ticket price. The bigger the odds ratio is from the number 1, the stronger impact the factor will have on the dependent variable. Among the significant factors, Continent has a strong positive impact on the resort ticket price, but its odd ratio has exceeded 10, which is unusual, hence this factor is overfitted to the model (*plot 12*).

Basing on the summary of Logistic Regression model reported through R language, a concrete function on calculating the probability of affecting resort ticket prices can be built (only significant factors of which the p-value less than 0.05 are taken into consideration):

In Machine Learning, apart from applying algorithms to build the model, another common task is to make predictions on given data. The next step is to split 70% of the data into test data, and the rest 30% of the data into train data. The prediction result is presented as a confusion matrix in *plot 13*.

The sensitivity rate is 0.986, which means that the proportion of actual positives is 98.6%. The specificity rate is 0.588, which means that the proportion of actual negatives is 58.8%. The positive predictive value is 0.951, which means that 95.1% of total positives are actual true positives. The negative predictive value is 0.833, which means that 83.3% of total negatives are actual true negatives. The accuracy is 0.942, which means that 94.2% data from the train dataset is predicted right. The F-1 score is 0.968, which means that the model is well in classification.

ROC curve is a plot to judge whether the data is well classified. The horizontal axis is false positive rate, the vertical axis is true positive rate. When the curve is closer to the upper left corner of the coordinate system, the better the dataset is classified. In the project's logistic regression model, the curve is very close to the upper left corner, indicating that the dataset is well classified (*plot 14*).

AUC stands for the area under the curve. AUC represents the area under the ROC curve. In this model, the AUC reaches 0.965, which helps to prove the dataset is well-classification (plot 15).

By comparison, logistic regression is the most suitable model for this dataset, with a classification accuracy of almost 95%. From the diagnostic plot, we can see that the data is not normally distributed because there are outliers in the data. The research overall has three limitations. One is the small sample size, approximately only 20 factors and 500 resorts' records are involved in the research. Secondly the characteristics of independent variables are not prominent in certain dimensions. Thirdly, there are over 50 missing data in the original dataset, but when process the dataset, all of them are omitted. The logistic regression model can just meet this defect of the dataset. Interestingly, all three models turn out that skiing slopes, and resorts' lifting ability are the most significant factors that influence tourists' reviews and resorts' ticket prices. Last but not least is the overfitting problem. Even though no factor's VIF index is more than 10, when building the logistic regression model, some of their odds ratios are extreme unusual, even reach 60, representing a warn sign that if all factors are involved, the model will be overfitted. To solve this problem, a new dataframe is subset from the original dataset.

Only three models are used in the research, logistic regression is just a relevant and suitable model, but it doesn't mean logistic regression can best answer the research question. In the further research steps, more statistical models need to be involved in the research, so that we can extract more useful information from it.

F. Conclusion

From the three models, we can conclude that resorts' facilities and resorts environmenta are the factors that tourists are concern most when choosing a ski resort as a destination. These are also the strongest factors that influence the resort's ticket prices. The more advanced the ski facilities, and the better the environment condition of the resort, the higher the ticket price will be. We

suggest that when choosing a resort, in addition to considering cost and tickets, tourists also need to consider safety factors, such as whether the resort's facilities are human-friendly, and whether the terrain in the ski area is steep. The operators of ski resorts should maintain and update ski resort facilities from time to time while paying attention to the quality of artificial snow and invest more medical emergency resources to ensure the safety of tourists. When investors invest in ski resorts, in addition to paying attention to income, they should also pay attention to the business strategies and service attitude of ski resorts. The significance of this research is to provide our research objects: tourists, investors, and ski resort operators with some references for choosing/operating ski resorts. Through research, readers can learn more about the ski market. For the ski industry, the research results focus on safety considerations, which can promote the industry to further improve various safety protection facilities in the ski resort.

Reference

Berard-Chenu, L., Cognard, J., François, H., Morin, S., & George, E. (2021). *Do changes in snow conditions have an impact on snowmaking investments in French Alps ski resorts?* International Journal of Biometeorology, 65(5), 659-675.

Boris, D., Makajić-Nikolić, D., Ćirović, M., Petrović, N., & Suknović, M. (2020). *A ski injury risk assessment model for ski resorts*. *Journal of Risk Research*, 23(12), 1590–1602. https://doi-org.proxy2.library.illinois.edu/10.1080/13669877.2020.1749113

Carus, L., & Castillo, I. (2021). *Managing risk in ski resorts: Environmental factors affecting actual and estimated speed on signposted groomed slopes in a cohort of adult recreational alpine skiers*. Plos one, 16(8), e0256349.

Eberly, L. E. (2007). Multiple linear regression. *Topics in Biostatistics*, 165-187.

Hatzenbuehler, J., Glazer, J., & Kuhn, C. (2009). *Awareness of Altitude Sickness Among Visitors to a North American Ski Resort*. Wilderness & Environmental Medicine, 20(3), 257–260. https://doi-org.proxy2.library.illinois.edu/10.1580/08-WEME-OR-191R.1

Jolliffe, I. T., & Cadima, J. (2016). *Principal component analysis: a review and recent developments*. Philosophical Transactions: Mathematical, Physical and Engineering Sciences, 374(2065), 1–16.

Peng, C.-Y. J., Lee, K. L., & Ingersoll, G. M. (2002). *An Introduction to Logistic Regression Analysis and Reporting*. The Journal of Educational Research, 96(1), 3–14.

Ski-resort-stats. (2020, April 12). *Find resort*. Ski resort statistics. Retrieved May 5, 2022, from https://ski-resort-stats.com/find-ski-resort/

Vanat , L. (2021, April). *2021 International Report on Snow & Mountain Tourism*. vanat.ch. Retrieved May 11, 2022, from https://www.vanat.ch/RM-world-report-2021.pdf.

Appendix 1. Factors Definition

Variable	Definition	Data Type
Resort	The name of the ski & snowboard resort.	chr(character)
Continent	The name of the continent in which the resort is located	chr
Country	The name of the country in which the resort is located	chr
Highest point	The highest mountain point at the ski resort	num(number)
Lowest point	The lowest possible point to ski at the ski resort	num
Child friendly	Is the ski resort child friendly or not?	chr
Ski pass prices adult	The price shows what it costs for 1 adult for 1 day in the main season in Euro €.	int
Season	Shows when the resort normally start and end the ski season.	chr
Beginner slopes	The total amount of “beginner” slopes in kilometer at the resort.	int

Intermediate slopes	The total amount of “intermediate” slopes in kilometer at the resort.	int
Difficult slopes	The total amount of “difficult” slopes in kilometer at the resort.	int
Total slopes	The sum of “beginner slopes” + “intermediate slopes” + “difficult slopes”	int
Longest run	The longest possible run at the ski resort, without using any lifts.	int
Snowparks	Does the resort have one or more snowparks, or not?	chr
Nightskiing	Does the resort offer skiing on illuminated slopes?	chr
Summer skiing	Does the resort offer summer skiing or not?	chr
Surface lift etc.	The amount of lifts in this category: T-bar, Sunkidslift, Rope lifts and people mower	int
Chairlift etc.	The total amount of chairlifts.	int
Gondola etc.	The amount of lifts in this category: Gondola, Train lifts, Funicular, Combined gondola and chairlifts, Helicopter lifts, Snowcats and	int

	Aerial tramways.	
Total lifts	The sum of “surface lifts etc” + “gondola etc” + “chairlifts etc”	int
Lift capacity	How many passengers can the lift system at the ski resort more in one hour?	num
Total lifts open	The amount of current open and runnings lifts at the ski resort.	chr
Snow cannons	The total amount of snow cannons at the ski resort.	num
Avg snow last 5 seasons	The average snow depth for the last 5 seasons.	chr

Appendix 2. Tables and Figures

$$X = t_1p_1^T + t_2p_2^T \dots + t_Rp_R^T + E$$

$$= TP^T + E$$

Where

X (*I x J*) is a data matrix,

T (*I x R*) are the scores,

P (*J x R*) are the loadings

E (*I x J*) are the residuals

R is the number of principal components used to describe X

Plot 1

$$\log\left(\frac{Pi}{1-Pi}\right) = \beta 0 + \beta 1X1 + \cdots \beta nXn$$

OR

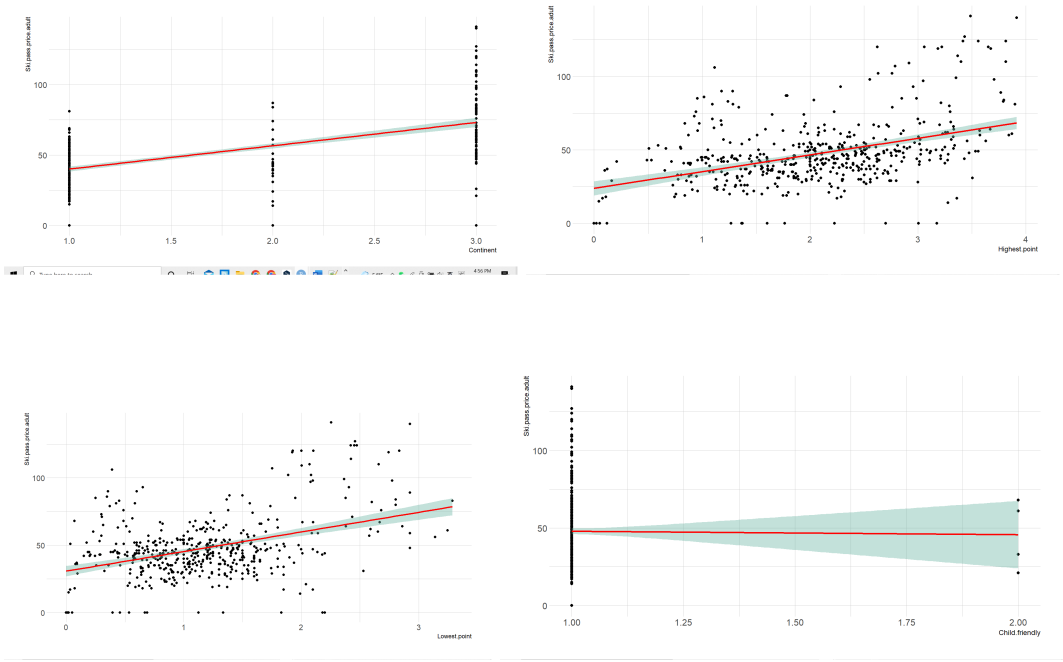
$$P = \frac{e^{\beta 0 + \beta 1X1 + \cdots \beta nXn}}{1 + e^{\beta 0 + \beta 1X1 + \cdots \beta nXn}}$$

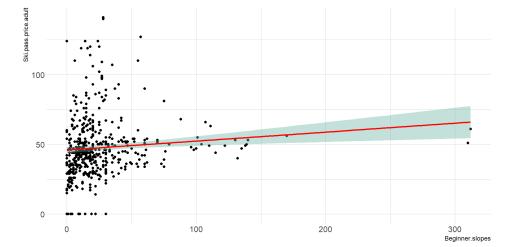
Where P is the target(dependent variable) likelihood

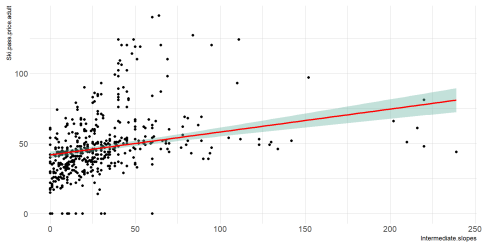
X is the independent variable

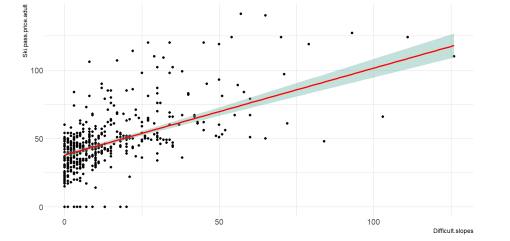
β is the beta coefficients

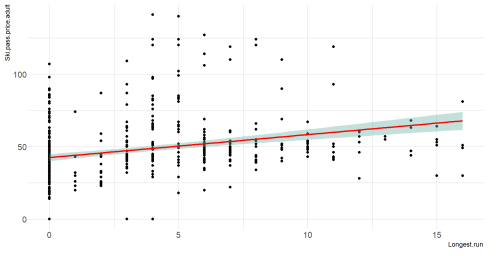
Plot 2

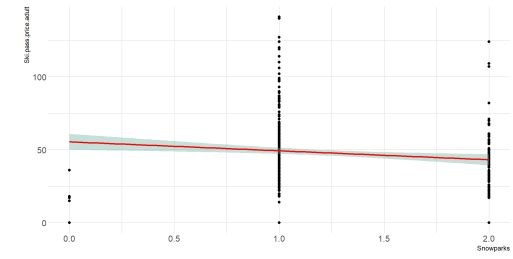


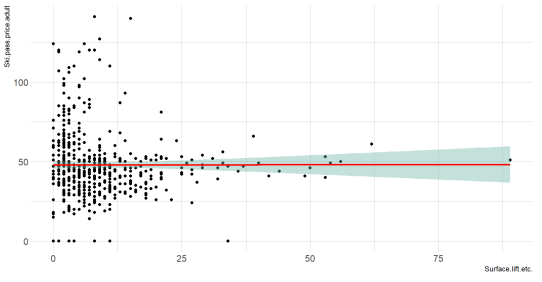

Ski.pass.price.adult
0
50
100
0
100
200
300
Beginner.slopes

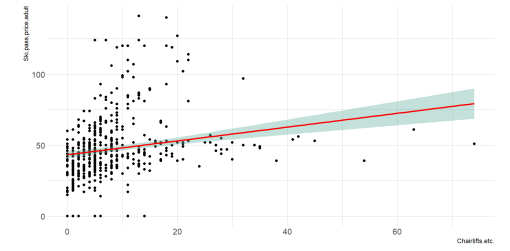

Ski.pass.price.adult
0
50
100
0
50
100
150
200
250
Intermediate.slopes

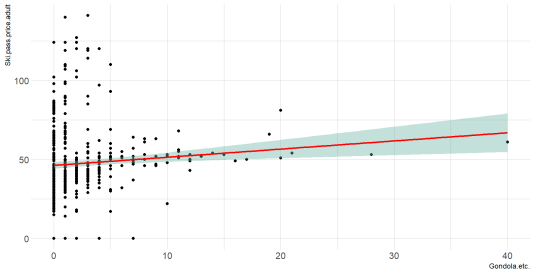

Ski.pass.price.adult
0
50
100
0
50
100
Difficult.slopes

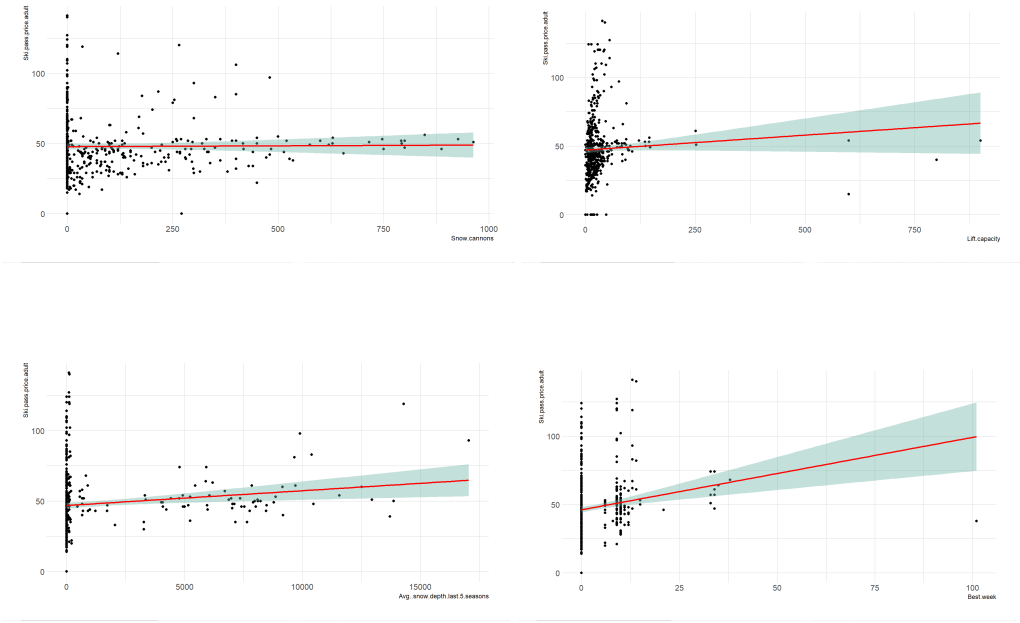

Ski.pass.price.adult
0
50
100
0
5
10
15
Longest.run


Ski.pass.price.adult
0
50
100
0.0
0.5
1.0
1.5
2.0
Snowparks


Ski.pass.price.adult
0
50
100
0
25
50
75
Surface.lift.etc.


Ski.pass.price.adult
0
50
100
0
20
40
60
Chairlifts.etc.

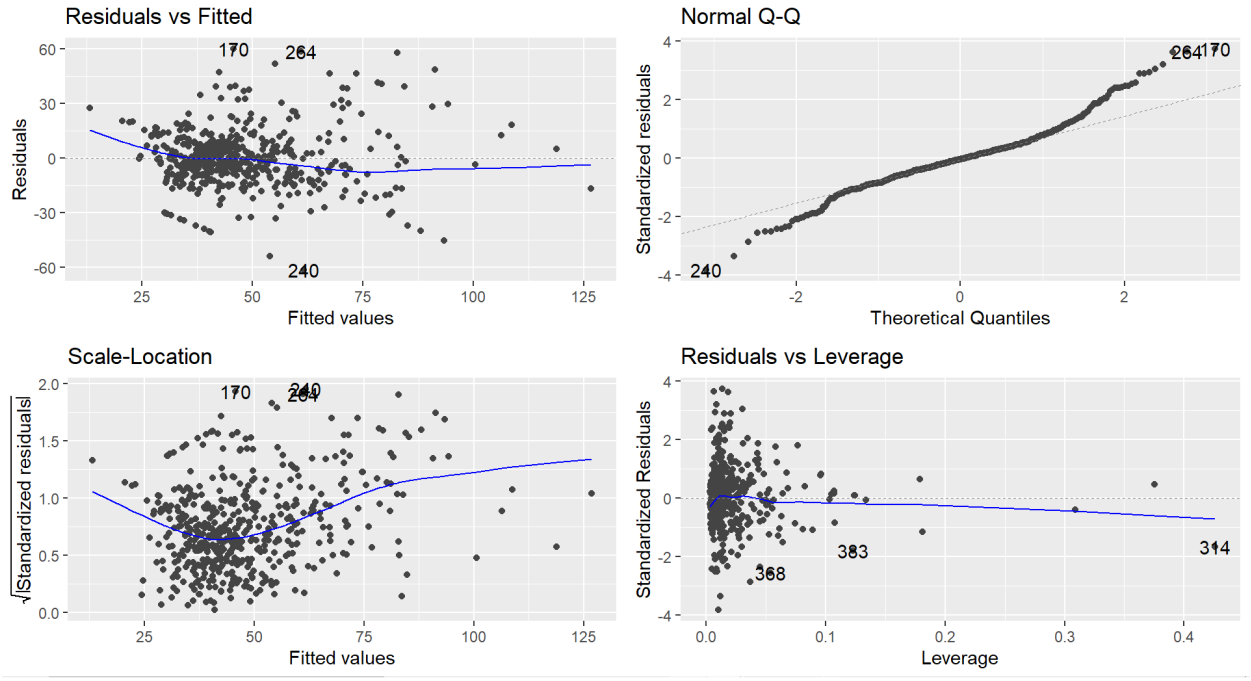

Ski.pass.price.adult
0
50
100
0
10
20
30
40
Gondola.etc.



Plot 3

```
          Highest.point                Lowest.point
               5.459112                    4.725760
         Beginner.slopes          Intermediate.slopes
               7.581188                    5.289118
       Difficult.slopes                 Longest.run
               2.666369                    1.435026
Surface.lift.etc.                Chairlifts.etc.
               3.946818                    4.595630
          Gondola.etc.                Lift.capacity
               3.693212                    1.380994
          Snow.cannons Avg..snow.depth.last.5.seasons
               1.238450                    1.376911
           Best.week
               1.237614
```

Plot 4



Plot 5

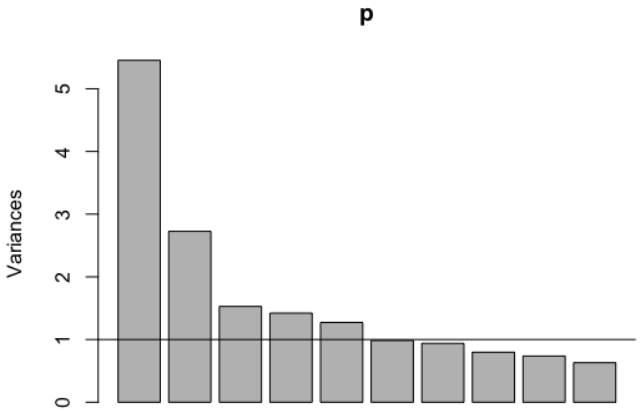
```
Residuals:
    Min      1Q  Median      3Q     Max
-62.120  -8.533  -0.888   6.889  61.533

Coefficients:
                                Estimate Std. Error t value
(Intercept)                   30.3047591  1.8249661  16.606
Difficult.slopes               0.5582813  0.0479682  11.639
Surface.lift.etc.             -0.6039301  0.0825822  -7.313
Lowest.point                   6.9590083  1.2803171   5.435
Longest.run                    0.9908855  0.1872728   5.291
Chairlifts.etc.                0.3560837  0.1055164   3.375
Avg..snow.depth.last.5.seasons 0.0005253  0.0002778   1.891
                              Pr(>|t|)
(Intercept)                    < 2e-16 ***
Difficult.slopes               < 2e-16 ***
Surface.lift.etc.             1.02e-12 ***
Lowest.point                  8.50e-08 ***
Longest.run                   1.81e-07 ***
Chairlifts.etc.               0.000796 ***
Avg..snow.depth.last.5.seasons 0.059187 .
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.18 on 508 degrees of freedom
Multiple R-squared:  0.4702,    Adjusted R-squared:  0.464
F-statistic: 75.15 on 6 and 508 DF,  p-value: < 2.2e-16
```

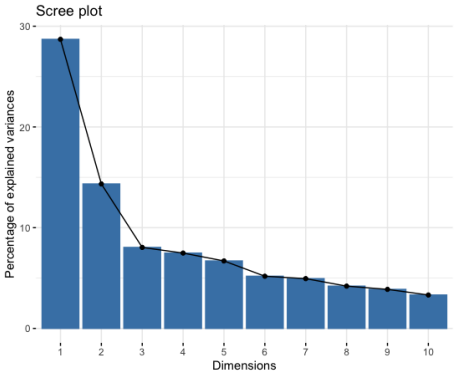
Plot 6

```
Residuals:
    Min      1Q  Median      3Q     Max
-62.658  -8.533  -0.906   7.034  61.209

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      30.23084    1.82916  16.527  < 2e-16 ***
Difficult.slopes  0.55756    0.04809  11.595  < 2e-16 ***
Surface.lift.etc. -0.58451    0.08215  -7.115 3.81e-12 ***
Lowest.point      7.06089    1.28242   5.506 5.83e-08 ***
Longest.run       1.01288    0.18738   5.405 9.95e-08 ***
Chairlifts.etc.   0.37781    0.10515   3.593 0.000359 ***


---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.5 on 499 degrees of freedom
Multiple R-squared:  0.5823,    Adjusted R-squared:  0.5697
F-statistic: 46.37 on 15 and 499 DF,  p-value: < 2.2e-16
```

Plot 7


p
Variances
0
1
2
3
4
5

Plot 8


Scree plot
Percentage of explained variances
0
10
20
30
1
2
3
4
5
6
7
8
9
10
Dimensions

Plot 9

```
Loadings:
                               RC1    RC2    RC3
Beginner.slopes                 0.907
Intermediate.slopes             0.832
Surface.lift.etc.               0.885
Chairlifts.etc.                 0.852
Gondola.etc.                    0.840
Lift.capacity                   0.552
Continent                              0.677
Highest.point                          0.768
Lowest.point                           0.771
Ski.pass.price.adult                   0.800
Difficult.slopes                0.489  0.696
Child.friendly                                0.504
Summer.skiing                                -0.597
Avg..snow.depth.last.5.seasons                0.542
Best.week                                     0.740
Longest.run
Snowparks
Nightskiing
Snow.cannons

                 RC1   RC2   RC3
SS loadings    4.979 3.168 1.559
Proportion Var 0.262 0.167 0.082
Cumulative Var 0.262 0.429 0.511
```

Plot 10

Characteristic	OR[1]	95% CI[1]	p-value
Beginner.slopes	0.98	0.94, 1.01	0.14
Intermediate.slopes	1.03	1.02, 1.05	<0.001
Surface.lift.etc.	0.91	0.81, 1.00	0.073
Chairlifts.etc.	1.08	1.02, 1.16	0.011
Gondola.etc.	1.11	0.91, 1.36	0.3
Lift.capacity	1.00		0.4
Difficult.slopes	1.00	0.98, 1.03	0.7
Continent	13.9	5.33, 50.1	<0.001
Highest.point	0.54	0.09, 3.13	0.5
Lowest.point	2.84	0.47, 18.3	0.3
Difficult.slopes.1			

[1] OR = Odds Ratio, CI = Confidence Interval

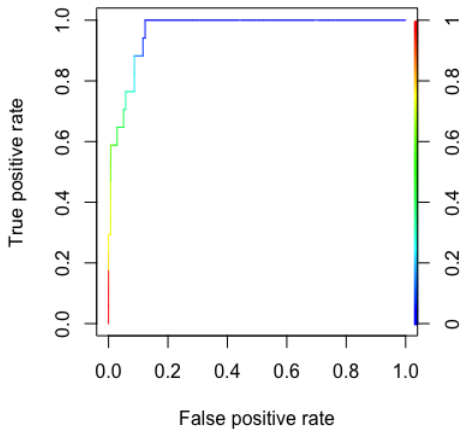
Plot 11

```
Call:
NULL

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.46420  -0.10358  -0.04496  -0.03253   2.84809

Coefficients: (1 not defined because of singularities)
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -9.769538   2.307628  -4.234  2.3e-05 ***
Beginner.slopes    -0.015303   0.020227  -0.757 0.449320
Intermediate.slopes  0.030719   0.011567   2.656 0.007914 **
Surface.lift.etc.  -0.103660   0.064124  -1.617 0.105977
Chairlifts.etc.     0.076114   0.036044   2.112 0.034712 *
Gondola.etc.        0.054567   0.133890   0.408 0.683604
Lift.capacity       0.003379   0.005549   0.609 0.542581
Difficult.slopes    0.008509   0.015132   0.562 0.573903
Continent           2.553214   0.689536   3.703 0.000213 ***
Highest.point      -0.451608   1.107497  -0.408 0.683440
Lowest.point        0.919360   1.122465   0.819 0.412755
Difficult.slopes.1         NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 242.73  on 359  degrees of freedom
Residual deviance: 113.72  on 349  degrees of freedom
AIC: 135.72

Number of Fisher Scoring iterations: 8
```

Plot 12

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 136   7
         1   2  10

               Accuracy : 0.9419
                 95% CI : (0.8926, 0.9731)
    No Information Rate : 0.8903
    P-Value [Acc > NIR] : 0.02037

                  Kappa : 0.6587

 Mcnemar's Test P-Value : 0.18242

            Sensitivity : 0.9855
            Specificity : 0.5882
         Pos Pred Value : 0.9510
         Neg Pred Value : 0.8333
              Precision : 0.9510
                 Recall : 0.9855
                     F1 : 0.9680
             Prevalence : 0.8903
         Detection Rate : 0.8774
   Detection Prevalence : 0.9226
      Balanced Accuracy : 0.7869

       'Positive' Class : 0
```

Plot 13



Plot 14

	modnames	dsids	curvetypes	aucs
1	m1	1	ROC	0.9654731
2	m1	1	PRC	0.7888287

Plot 15

Appendix 3. Syntax/Code (R Language)

```
library(plyr)
library(stringr)
library(RVAideMemoire)
library(gmodels)
library(tidyverse)
library(DescTools)

Import dataset
data = read.csv(file.choose(), header=T)
data <- na.omit(data) # omit missing data
df = subset(data, select = -c(Total.lifts.open, Resort.name)) # Remove obvious useless column

Revalue factors/pre-processing
df$Continent <- as.numeric(revalue(df$Continent, c(
  "Europe" = 1,
  "Rest of the world" = 2,
  "America" = 3
))) # Continent
x <- unique(df$Country)
for (i in 1:length(df[,1])) {
  for (j in 1:length(x)) {
    if(df$Country[i] == x[j]) {
      df$Country[i] <- as.numeric(j)
      break}}}
df$Country <- strtoi(df$Country) # Country
df$Child.friendly <- as.numeric(revalue(df$Child.friendly, c(
  "Yes" = 1,
  "No" = 2
))) # Child.friendly
x <- unique(df$Season)
for (i in 1:length(df[,1])) {
  for (j in 1:length(x)) {
    if(df$Season[i] == x[j]) {
      df$Season[i] <- as.numeric(j)
      break
    }
  }
}
df$Season <- strtoi(df$Season) # Season
df$Snowparks <- as.numeric(revalue(df$Snowparks, c(
  "Yes" = 1,
  "No" = 2,
  "no report" = 0
))) #Snowparks
```

```
df$Nightskiing <- as.numeric(revalue(df$Nightskiing, c(
  "Yes" = 1,
  "No" = 2
))) # Nightskiing
df$Summer.skiing <- as.numeric(revalue(df$Summer.skiing, c(
  "Yes" = 1,
  "No" = 2,
  "no report" = 0
))) # Summer.skiing
df$Avg..snow.depth.last.5.seasons <- revalue(df$Avg..snow.depth.last.5.seasons, c(
  "no report" = 0
))
df$Avg..snow.depth.last.5.seasons <- str_replace_all(df$Avg..snow.depth.last.5.seasons, ",", "")
df$Avg..snow.depth.last.5.seasons <- strtoi(df$Avg..snow.depth.last.5.seasons) # Avg…5seasons
# Remove variables with aliased coefficients
rdf <- subset(df, select = -c(Total.lifts, Total.slopes,Country,Season))
```

MLR model:

```
# Remove them to make sure the VIF < 5
tmp <- subset(rdf, select = -c(Beginner.slopes, Highest.point, Intermediate.slopes))
# Remove them to make sure linear relationship
tmp <- subset(tmp, select = -c(Continent, Child.friendly, Snowparks, Nightskiing,
Summer.skiing))
#Create Initial Linear Regression Model with Enter Method
model1 <- lm(Ski.pass.price.adult ~ ., data=tmp)
model1
#Check VIF
VIF(model1)
summary(model1)
#Diagnostic Plots for Model Fit
# approach 1: par(mar=c(1,1,1,1))
par(mfrow = c(2, 2))
plot(model1)
# approach 2: library(ggfortify)
autoplot(model1)
par(mfrow = c(1, 1))
#Using Stepwise Multiple Linear Regression
null = lm(Ski.pass.price.adult ~ 1, data=tmp)
null
full = lm(Ski.pass.price.adult ~ ., data=tmp)
full
#Stepwise Regression
train_Step = step(null, scope = list(upper=full), direction="both")
summary(train_Step)
# Keep the variables that stepwise regression output
```

```
tmp <- subset(tmp, select = c(Difficult.slopes, Surface.lift.etc., Lowest.point, Longest.run,
Chairlifts.etc., Ski.pass.price.adult))
# Build model
model2 <- lm(Ski.pass.price.adult ~ ., data=tmp)
summary(model2)
```

PCA model:
Determine the number of components:
Approach 1: The Knee Method

```
library(factoextra)
p<-prcomp(rdf,center=T, scale=T)
fviz_eig(p)
```

Approach 2: The KMO method

```
plot(p)
abline(1,0)
```

Create the PCA model:

```
library(psych)
p2<-psych::principal(rdf,rotate='varimax',nfactors=3,score=TRUE) # using the Knee method
p3<-print(p2$loadings, cutoff=.4, sort=T)
```

Logistic Regression Model:
Re-subset dataframe (choose factors from PCA's RC1 and RC2)

```
rdf1<-subset(rdf,select =
c(Ski.pass.price.adult,Beginner.slopes,Intermediate.slopes,Surface.lift.etc.,Chairlifts.etc.,Gondol
a.etc.,Lift.capacity,Difficult.slopes,Continent,Highest.point,Lowest.point,Difficult.slopes))
```

Revalue the dependent variable (Ski.pass.price.adult. price <=70: 0; price>70: 1)

```
rdf1$Ski.pass.price.adult<-as.numeric(rdf1$Ski.pass.price.adult)
rdf1$Ski.pass.price.adult_1<-cut(rdf1$Ski.pass.price.adult,c(-1,70,150))
table(rdf1$Ski.pass.price.adult_1)
rdf1$Ski.pass.price.adult_1<-0*(rdf1$Ski.pass.price.adult_1=='no')+0*(rdf1$Ski.pass.price.adult
_1=='(-1,70]')+1*(rdf1$Ski.pass.price.adult_1=='(70,150]')
rdf2<-subset(rdf1,select = -c(Ski.pass.price.adult))
names(rdf2)
```

Calculate Odds Ratios

```
rdf2$Ski.pass.price.adult_1<-as.factor(rdf2$Ski.pass.price.adult_1)
log_reg <- glm(Ski.pass.price.adult_1 ~ ., family = "binomial", data = rdf2)
library(gtsummary)
log_reg %>%
  gtsummary::tbl_regression(exp = TRUE)
```

Classification (70% training data, 30% test data)

```
set.seed(123)
split <- initial_split(rdf2, prop = .7, strata = "Ski.pass.price.adult_1")
train <- training(split)
test  <- testing(split)

library(caret)
```

```
train$Ski.pass.price.adult_1<- as.factor(train$Ski.pass.price.adult_1)
log_reg_1 = train(
  form = Ski.pass.price.adult_1 ~ .,
  data = train,
  method = "glm",
  family = "binomial"
)
Build Confusion Matrix
library(rsample)
pred <- predict(log_reg_1, test)
confusionMatrix(pred, as.factor(test$Ski.pass.price.adult_1), mode = 'everything')
ROC Curve
log_reg_train <- glm(Ski.pass.price.adult_1 ~ ., data=train, family=binomial)
library(ROCR)
log_reg_test_prob <- log_reg_train %>% predict(test, type = "response")
preds <- prediction(as.numeric(log_reg_test_prob), test$Ski.pass.price.adult_1)
perf <- performance(preds,"tpr","fpr")
plot(perf,colorize=TRUE)
AUC values
library(precrec)
precrec_obj <- evalmod(scores = log_reg_test_prob, labels = test$Ski.pass.price.adult_1)
sm_aucs <- auc(precrec_obj)
sm_aucs
```