

# **Skiing Hotels in Europe**

Bell, Kai & Movius, Luke

# Agenda

Introduction

Summary

Visualization

Regression Analysis

Classification






# Introduction



# Background & Description

- This dataset includes information about ski hotels across Europe
- Including: the country that each resort is located in, the name of the resort, the name of the specific hotel within the resort, the price per person per week, the distance from each hotel to the nearest lift, the altitude of each resort, the total piste distance, the number of lifts, and the number of gondolas
- We chose this dataset because we want to know what the objective best ski resort and hotel combination exists in Europe, since it can be very expensive to travel there



# **Manipulation and Summary**

# Cleaning

- Removed 14 columns, including data about how many of each type of lift there are at each hotel, how many slopes of each difficulty, the links to each hotel, and the snow data from each hotel
  - This resulted in a decrease of 24 columns to 9 columns
- Replaced original price column with the prices converted to USD, replaced original distance from lift column with new column converted from meters to feet, also replaced piste column with conversion to feet, as well as altitude converted to feet
  - Done since we chose the data set to discover the best value for our traveling, we will spend USD and not GBP
  - We also use the imperial system rather than metric, so we changed the distances and altitude from meters to feet and gave it a cleaner name
- Removed all hotels with NA values, mostly those missing the distance from nearest lift
- Created columns containing a normalized 1-10 scale for the price, distance from nearest lift, and piste length for each hotel
- Then, created the BestValue column which has the average score for the 3 previous values
  - The higher the number, the better value
  - Additional column “HiLoValue” to easily give viewer an indication of whether a hotel has a high or low value ( $\geq$  median is high)
- After Cleaning:
  - 14 columns: 4 categorical, 10 numerical

# Summary of Variables

Var	Min	1st Quartile	Median	Mean	3rd Quartile	Max
PriceInUSD	690.5	1048.6	1326.1	1359.5	1564.0	3070.7

Var	Min	1st Quartile	Median	Mean	3rd Quartile	Max
AltInFt	590.4	3202.9	4788.8	4604.1	5904.0	7544.0

Var	Min	1st Quartile	Median	Mean	3rd Quartile	Max
BestValue	3.0	5.5	6.0	5.997	6.5	8.9

Var	Min	1st Quartile	Median	Mean	3rd Quartile	Max
TotalPisteInMi	0.0	71.3	124.0	150.9	186.0	756.4

Country	Mean Price (USD)
Andorra	1049.67
Austria	1518.64
Bulgaria	828.01
Finland	1255.38
France	1494.13
Italy	1178.23

Country	Mean Altitude (Ft)
Andorra	5904
Austria	3479
Bulgaria	3749
Finland	623
France	5772
Italy	5103


Country	Mean Total Piste (Mi)
Andorra	101
Austria	146
Bulgaria	37
Finland	19
France	226
Italy	126

Country	Mean Best Value
Andorra	6.34
Austria	5.69
Bulgaria	5.69
Finland	5.53
France	6.32
Italy	6.10

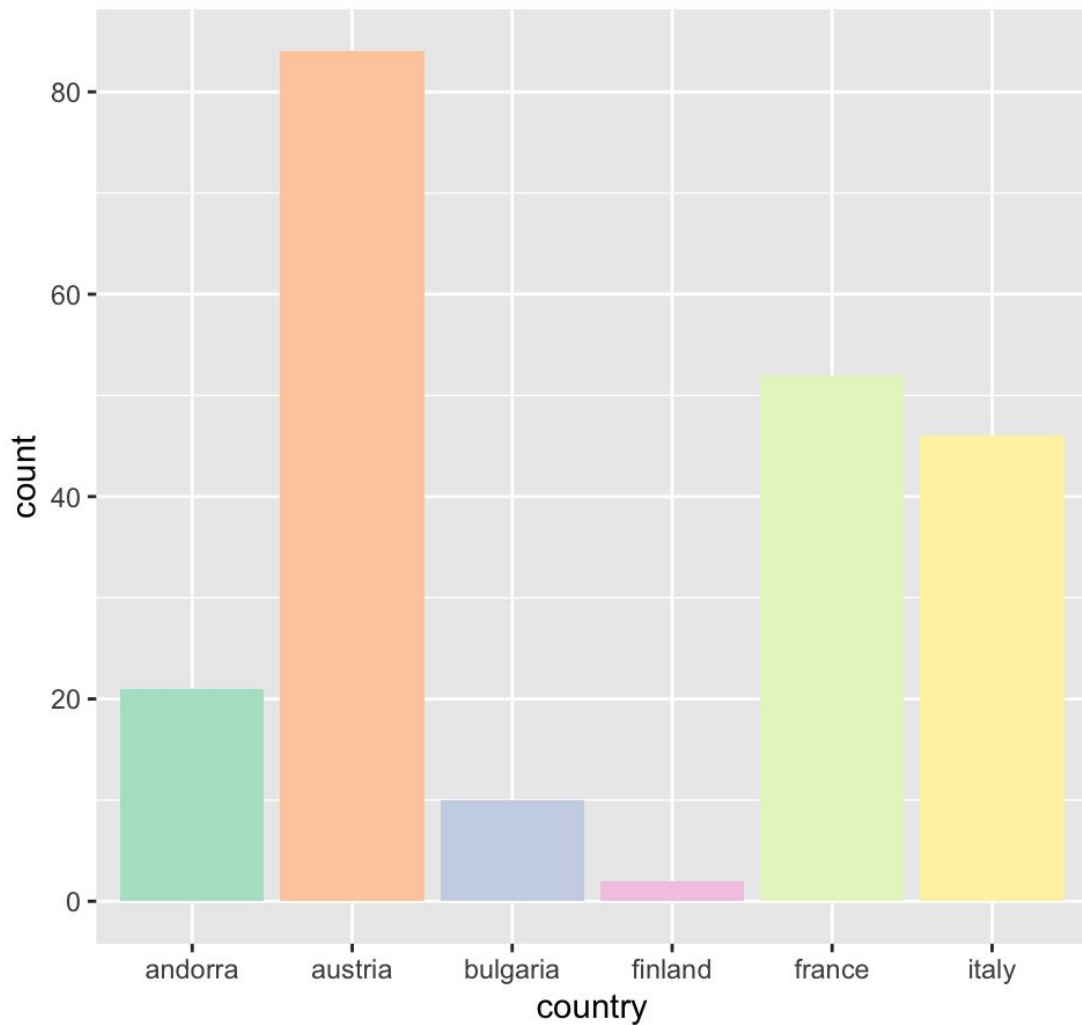


# Visualization

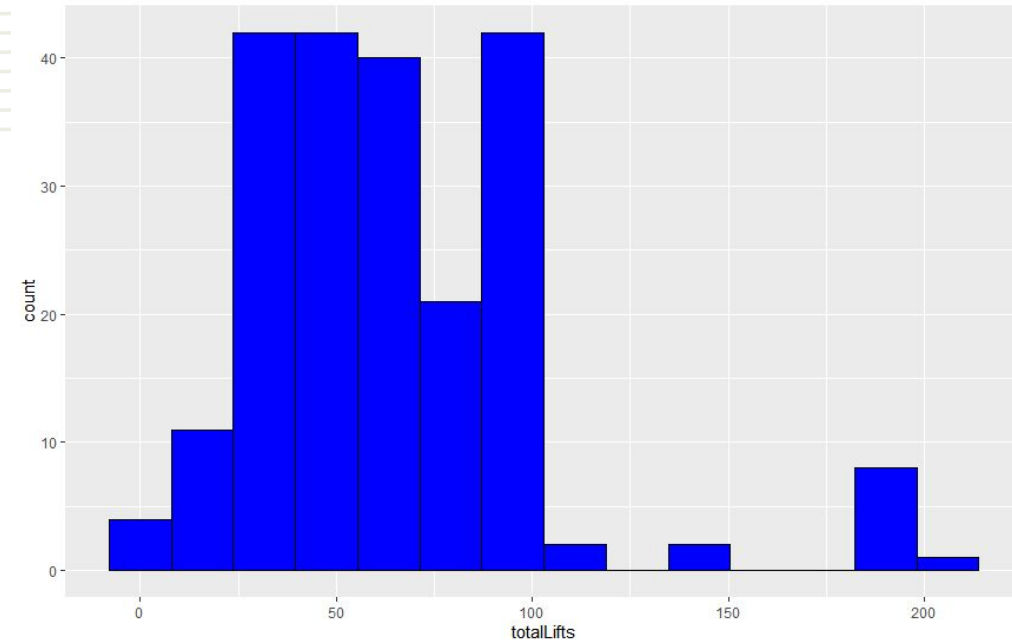





Starting with something basic, here is the frequency of each country present in the data set. As you can see, Finland only has two data points, while Austria has over 80. Knowing this information is important, as we can't make super broad statements about Finland, as we don't have a large sample size.



# How many lifts are at each resort?



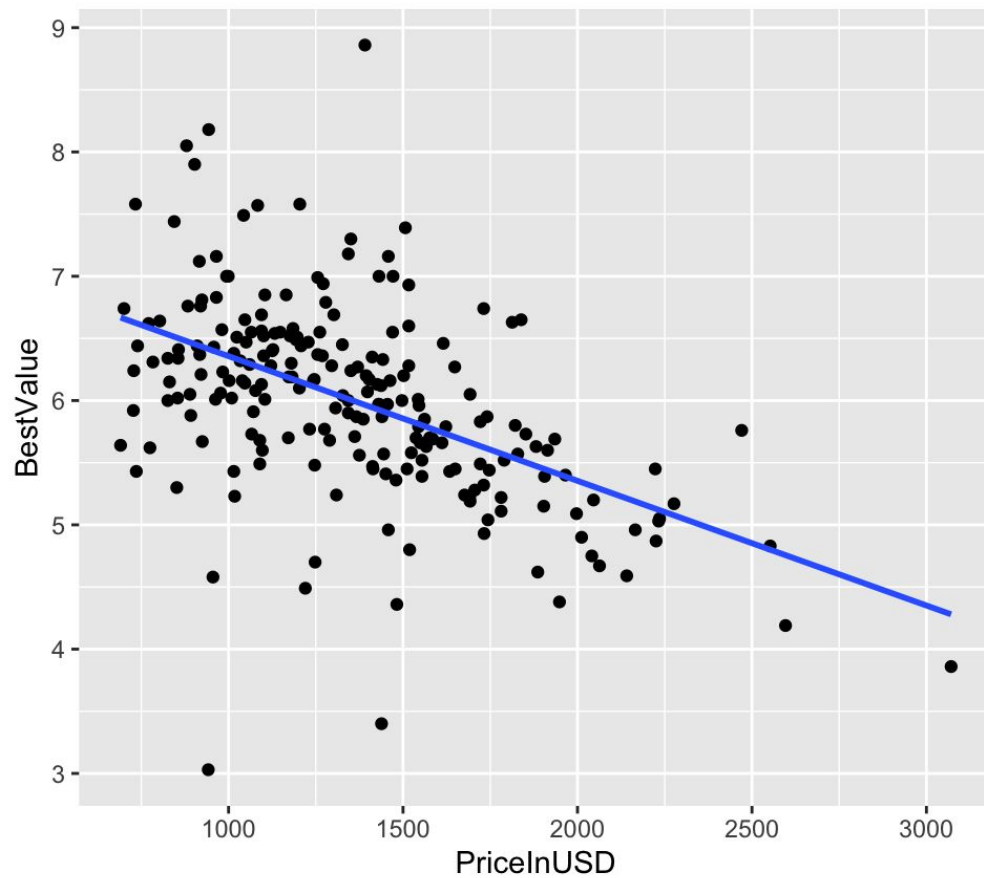
- Pictured left is a histogram of the count of totalLifts



# Are the most expensive resorts the best value?

We aim to answer this by using a scatter plot  
comparing the value of the resorts to their price.

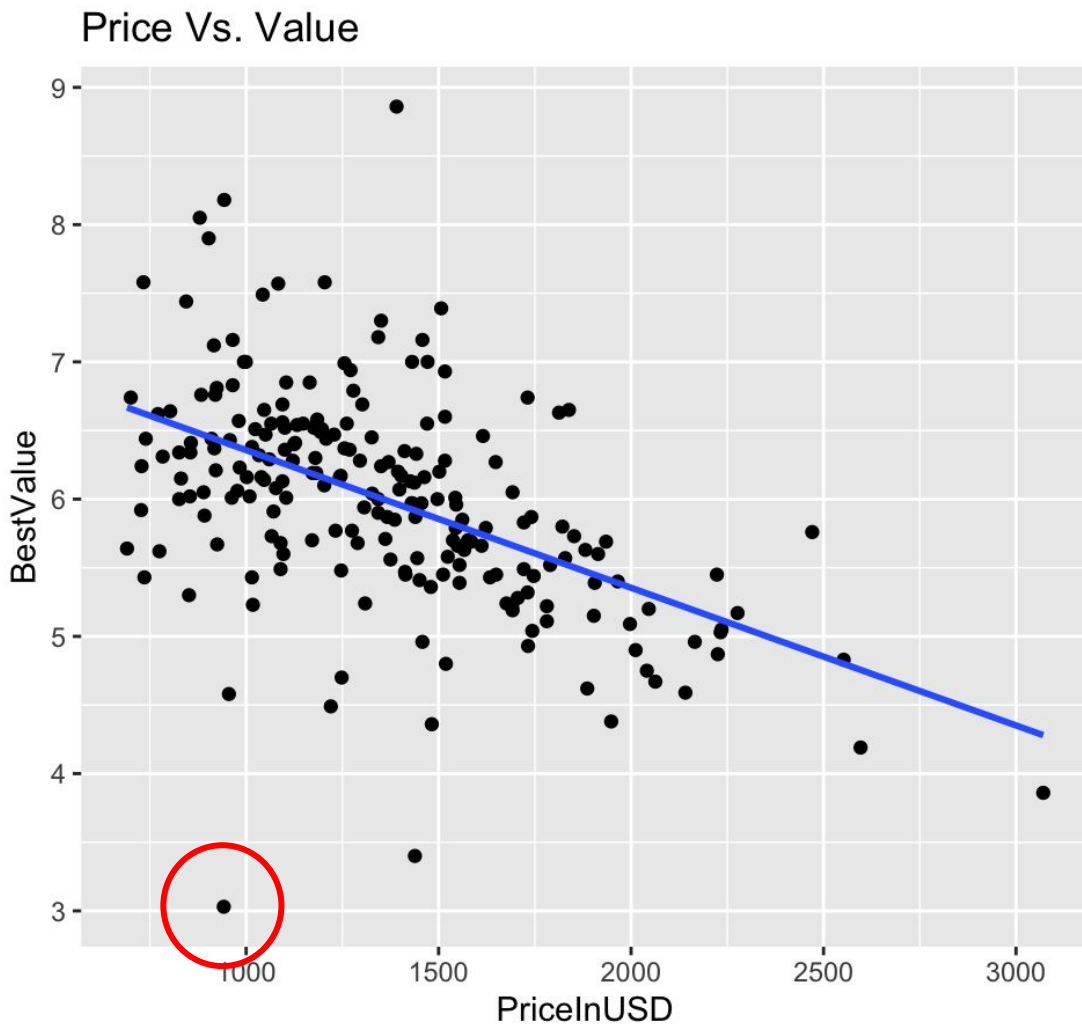
Price Vs. Value



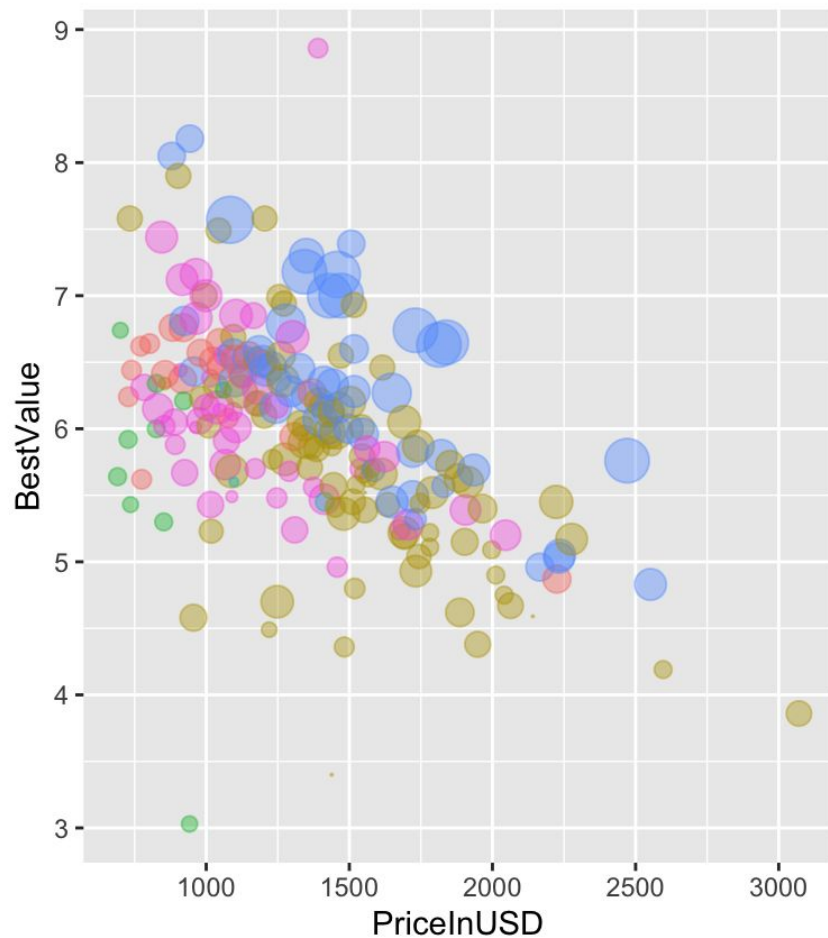
As the scatter plot reveals, there is a negative correlation between the price of a resort, and its value.

This suggests that if you are looking for a high-value resort, you are better off choosing one of the cheaper ones.

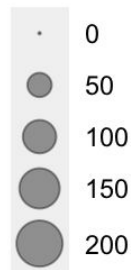
There were some outliers, though. For example, hotel-orlovets is relatively cheap, but also is low-value.



# Price Vs. Value



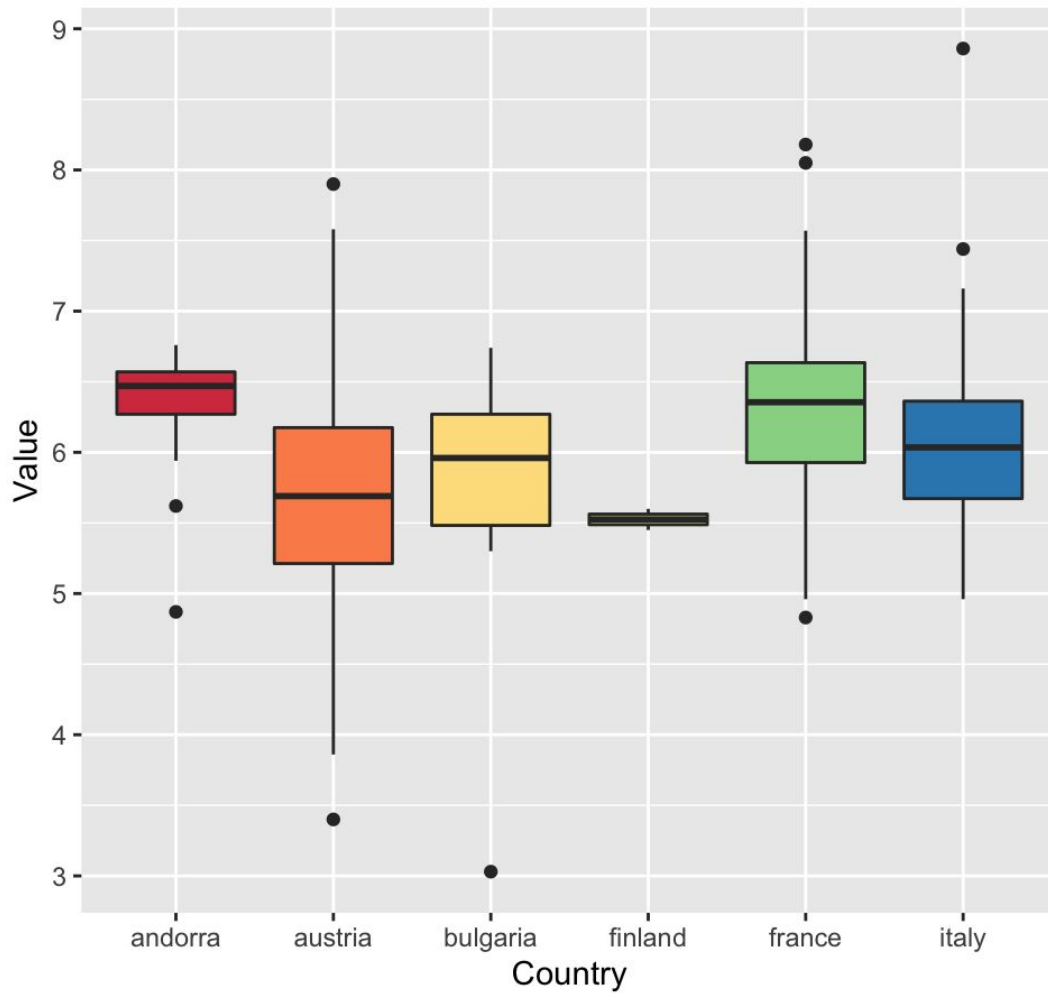
## totalLifts



## country



This is the same scatter plot but showing the resort's country and the amount of lifts it has. From this graph, it seems that either France or Andorra have the best value on average. Let's test this.


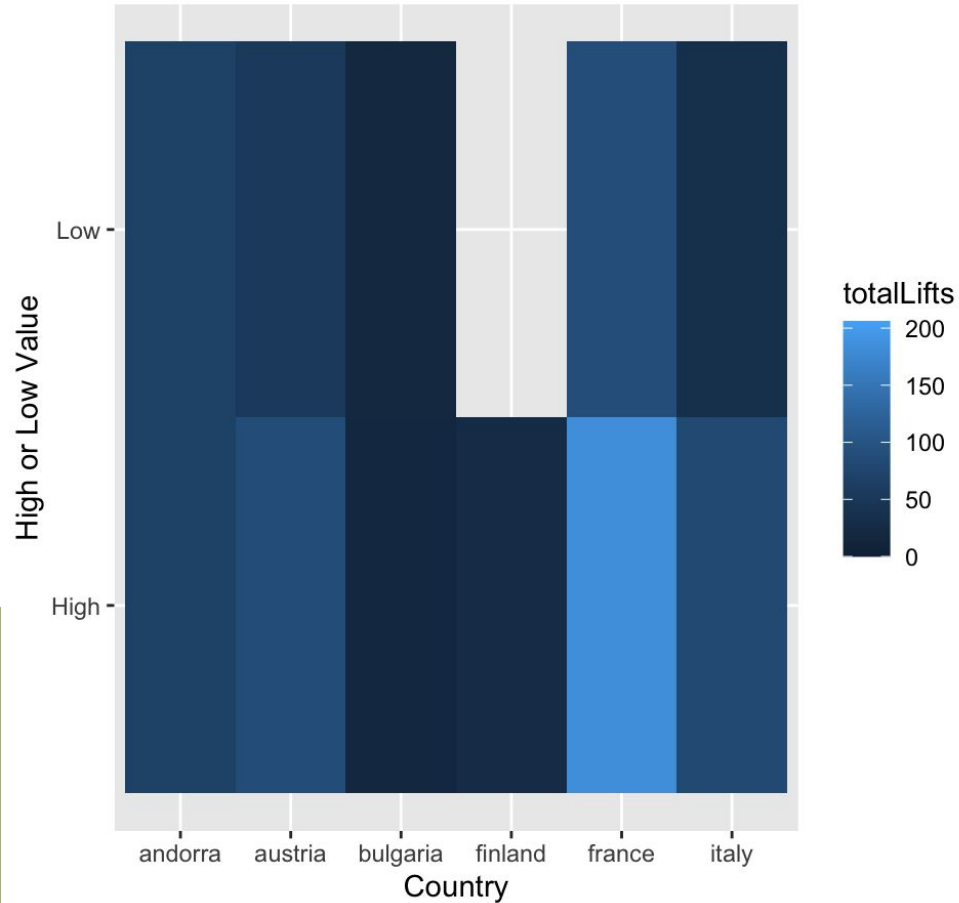


By a slim margin, Andorra has a higher **median** value than France, followed by Italy, Bulgaria, Austria, and Finland. But, do they also have the higher mean?

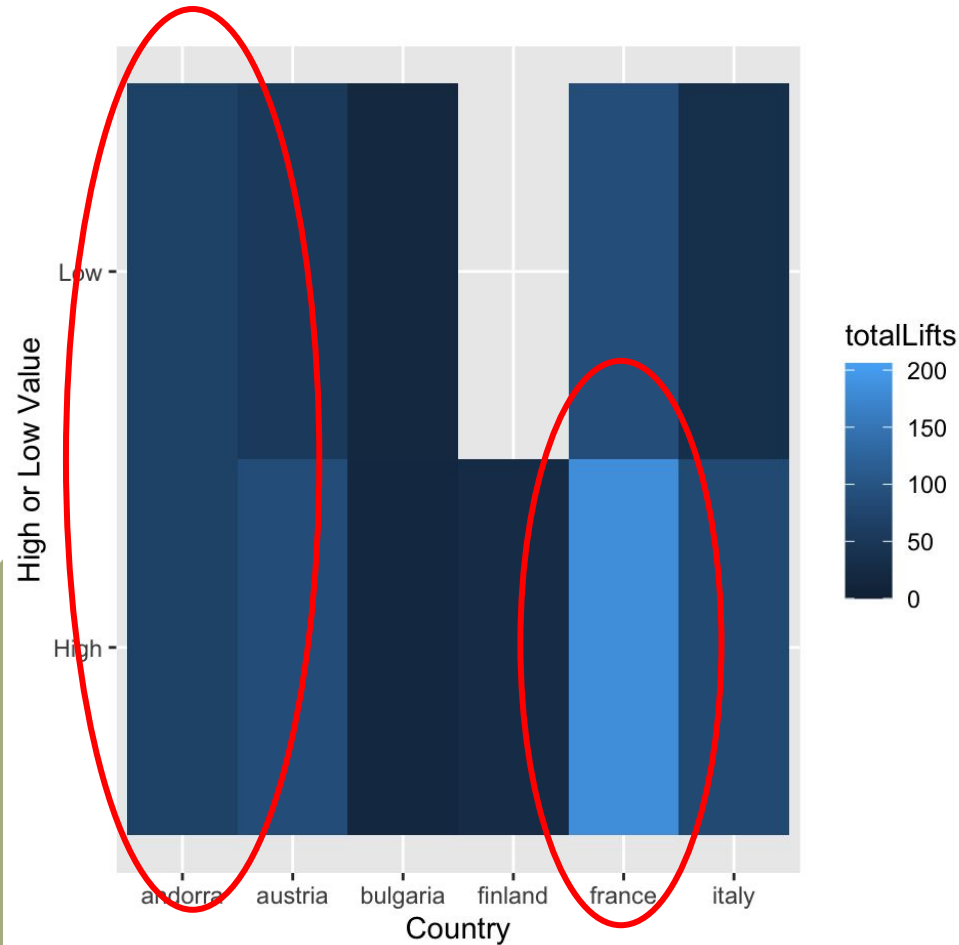
# What can we conclude from the boxplot?

It turns out that Andorra also has a higher mean value (6.342381) than France (6.323654) although it is very close. Though there are outliers of course, this suggests that Andorra is on average the best place to plan a trip if you want the most value for your money (and you may want to steer clear of the two resorts in Finland).





This is a heatmap comparing the resort's country, whether or not it is high or low value, and then how many ski lifts it contains. We can take away a few things from this graph. ----->



- France, the highest value country, also has the most lifts
- In some cases, like Bulgaria and Andorra, the amount of lifts tends to not be a great predictor of the value
- But, in Italy, Austria, and France, how many lifts a resort has does predict how much value you will get.



This wordcloud was created to visualize the average BestValue of each country



italy austria  
andorra france  
bulgaria finland



# Regression Analysis

# Multiple Regression

Call:

```
lm(formula = BestValue ~ NormalizedPisteScale + NormalizedPriceScale +  
    NormalizedDistScale, data = sdnew)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.0041334	-0.0028654	0.0000434	0.0030846	0.0041691

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.0042851	0.0017212	-2.49	0.0136 *
NormalizedPisteScale	0.3334878	0.0001246	2676.00	<2e-16 ***
NormalizedPriceScale	0.3334643	0.0001122	2972.53	<2e-16 ***
NormalizedDistScale	0.3336569	0.0001493	2235.29	<2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.002701 on 211 degrees of freedom

Multiple R-squared: 1, Adjusted R-squared: 1

F-statistic: 6.255e+06 on 3 and 211 DF, p-value: < 2.2e-16

- Since we made a scale that is directly correlated with the calculations for BestValue, the  $r^2$  value of this model is 1. This means that we can make near perfect predictions based on our interpretation of the data.
- Based on this, we can predict that scores of 8, 2, and 4 on the NormalizedPisteScale, NormalizedPriceScale, and NormalizedDistScale respectively will yield a BestValue score of 4.66

# Logistic Regression

For our logistic regression we subsetting our data, keeping four variables

We used our HiLoValue variable for the logistic regression, as it is a binomial categorical variable.

We based our model on three other variables: country, total lifts, and price in USD.

	country	totalLifts	PriceInUSD	HiLoValue
3	bulgaria	24	690.52	High
4	bulgaria	18	700.28	High
5	bulgaria	24	727.12	High
6	andorra	30	728.34	High
7	austria	52	733.22	High
8	bulgaria	18	735.66	High

# Building our model for logistic regression

We then built our model based on these variables

Here is what our summary revealed:

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.44253  -0.36610  -0.19831  -0.09857   2.97330

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.046e+00  1.796e+00  -3.366  0.000762 ***
countryaustria -1.236e-01  1.305e+00  -0.095  0.924571
countrybulgaria 1.306e+00  1.739e+00   0.751  0.452902
countryfinland -1.477e+01  1.687e+03  -0.009  0.993014
countryfrance  -1.191e+00  1.515e+00  -0.786  0.431868
countryitaly   -1.249e+00  1.613e+00  -0.775  0.438602
totallifts     -2.986e-02  1.202e-02  -2.483  0.013031 *
PriceInUSD      3.751e-03  8.796e-04   4.264  2.01e-05 ***
```

```
Null deviance: 133.075  on 214  degrees of freedom
Residual deviance:  87.863  on 207  degrees of freedom
AIC: 103.86
```

# Interpreting our Summary

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.44253  -0.36610  -0.19831  -0.09857   2.97330

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.046e+00  1.796e+00  -3.366  0.000762 ***
countryaustria -1.236e-01  1.305e+00  -0.095  0.924571
countrybulgaria  1.306e+00  1.739e+00   0.751  0.452902
countryfinland -1.477e+01  1.687e+03  -0.009  0.993014
countryfrance  -1.191e+00  1.515e+00  -0.786  0.431868
countryitaly   -1.249e+00  1.613e+00  -0.775  0.438602
totallifts     -2.986e-02  1.202e-02  -2.483  0.013031 *
PriceInUSD      3.751e-03  8.796e-04   4.264  2.01e-05 ***
```

Our Y intercept is -6.046

Then, we have our slope coefficients for each of our variables.

So, if we were to write an equation, here is what it would look like:

$$Y = -6.046 + 0.003751\text{PriceInUSD}$$

You can write that out for all the variables, of course.



# Interpreting our deviances

To test how good our model is at making predictions, we can find a p value using our deviances

First we subtract our Residual deviance from our Null deviance to find the chi squared value:

$$133.075 - 87.863 = 45.212$$

We can then put this into a calculator along with of DOF to find our p value.

Our p value is 0.00000, meaning that our model is **highly useful** in predicting values.

```
Null deviance: 133.075 on 214 degrees of freedom
Residual deviance: 87.863 on 207 degrees of freedom
AIC: 103.86
```



# Classification

# What do we want to predict?

- Our model will be attempting to predict which country a resort is located in
- First, we are going to make up a new data point
- We are going to use plausible values:

NormalizedPisteScale	NormalizedPriceScale	NormalizedDistScale	BestValue	HiLoValue
2.29	9.18	6.65	6.78	High

DistFromLiftInFt	altitude..m.	totalPiste..km.	totalLifts	PriceInUSD	TotalPisteInMi
1200	1529	220	22	1006.23	88.99

# How many neighbors?

Using, that new data point, we built a model to see how many neighbors we will base our prediction off of

Here is the output revealing that value:

```
9-nearest neighbor model
```

```
Training set outcome distribution:
```

andorra	austria	bulgaria	finland	france	italy
21	84	10	2	52	46

# What the output means

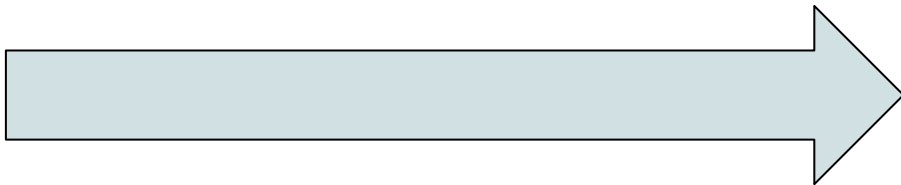
Our model will use **nine** neighbors to predict our new resort.

So, let's see what our model predicts!

9-nearest neighbor model

Training set outcome distribution:

andorra	austria	bulgaria	finland	france	italy
21	84	10	2	52	46



# Results

Our new resort was predicted to reside in Italy!

This means that based on the **nine** nearest neighbors, Italy, was the country that was overall “nearest” to our imaginary new resort.

```
[1] italy
```

```
Levels: andorra austria bulgaria finland france italy
```



# Link to Dataset

<https://www.kaggle.com/datasets/jacklacey/skiing-hotels>