Ski Resort Recommendation System Leveraging Data Mining Report

Juncheng Man
University of Colorado Boulder
Boulder, Colorado, USA
juma1543@colorado.edu
CSCI4502-001
StudentID:111448584

Audrey Ly
University of Colorado Boulder
Boulder, Colorado, USA
auly2364@colorado.edu
CSCI4502-001
StudentID:110435577

Grace Waida
University of Colorado Boulder
Boulder, Colorado, USA
grwa6713@colorado.edu
CSCI4502-001
StudentID:110068866

Justin Nguyen
University of Colorado Boulder
Boulder, Colorado, USA
jung1504@colorado.edu
CSCI4502-002
StudentID:110167485

1 ABSTRACT

This project aims to develop a ski resort recommendation system that addresses the diverse needs and preferences of skiers while accounting for variable factors like weather, location, and cost. Leveraging data mining methodologies, the system aggregates data from various sources, including resort characteristics, climate patterns, snowfall trends, and ticket prices, to deliver tailored recommendations. By integrating historical and real-time climate data, the system provides dynamic insights into resort conditions, enabling users to make well-informed decisions.

Key components of the project include data collection, cleaning, aggregation, and visualization. Advanced techniques, such as time-series forecasting and machine learning models like XGBoost, are employed to predict snowfall and uncover relationships among features such as ticket pricing, resort facilities, and geographical factors. Preliminary analysis highlights trends such as the influence of elevation, snowfall, and lift capacity on pricing. The system is designed with scalability in mind, paving the way for potential enhancements such as real-time data integration, user preference inputs, and personalized recommendations through an interactive platform.

2 INTRODUCTION

Skiing is one of the most popular winter sports in the United States, with a total of 60.4 million skiers reported for the 2023-24 season by The National Ski Areas Association (NSAA) [1]. As the demand for skiing continues to grow, there arises a pertinent problem to find the best ski resort that fits users needs and wants.

Making a choice on the exact resort to visit is difficult as it depends on personal preferences such as location, difficulty, climate, price, and snowfall level. This problem has been compounded by the impacts of climate change, with many ski resorts experiencing reduced snowfall and seasons cutting shorter [2].

To address these challenges, we propose a ski resort recommendation system, specifically tailored to help users choose what resort they prefer within the United States. Our system recommends these ski resorts based on user preferences like difficulty, location, price, and travel time, while also taking into account volatile factors like snowfall and weather. Data mining methodologies and tools will be utilized to gain knowledge and extract patterns from past ski resorts and climate data.

3 RELATED WORK

A number of previous works have focused on developing ski resort recommendation systems, predominantly in the e-commerce sector. An interactive dashboard was developed using data on ski resorts and snow coverage from the Mavens Data playground which was used for the Maven Slope Challenge [5]. The dataset were cleaned and pre-processed using Microsoft Excel, after which it would be stored in a database to be queried by users. The dashboard allows users to narrow down on their ideal ski resort using filter factors like location, season and snow, cost, slope length, and ski lift availability. However, the documentation did not provide evaluation metrics to determine the effectiveness of the solution. Another group developed a deep neural network as a recommendation system that predicts suitable ski resorts for customers of a web agency [8]. The deep neural network was trained on a dataset consisting of archived data of Valraiso, a e-Commerce Web Agency of France. It takes in the features: group, budget, duration between order week and stay week, stay week, altitude, ski difficulty, size of skiing area, and country of customer, and outputs a recommended ski resort. The evaluation of the neural network show a performance with a TOP-10 accuracy that reaches 98.95% and 99.30% for French and non-French customers, respectively, where TOP-n accuracy corresponds to the accuracy where the true class matches with any one of the n most probable classes predicted by the model. The NYC Data Science Academy has also published a ski resort recommendation system developed by a student, which helps users determine the optimal ski resort using data scraped from a popular skiing website (www.onthesnow.com) [11]. Two "web spiders" were built via the Scrapy Python framework, with one extracting ski resort data and the other extracting daily snowfall data for each resort. Average annual snow totals, variance in snow totals, cumulative snowfall and lift ticket prices were used to determine the optimal location and timing for skiing. Once again, no evaluation metrics were provided to assess the effectiveness of this solution.

A common pattern across these solutions is the mining of data from resorts and snowfall datasets, and the use of popular features like cost, snowfall, and location to provide recommendations for users. However, none of them has examined in detail the role of climate in shaping the desirability of ski resorts. While the impacts of climate change on snowfall may not have been prominent when the above solutions were developed, in today's world it can no longer be ignored. Gaining knowledge from historical temporal snowfall data is no longer sufficient, and snowfall forecasts have to be combined with climate patterns to provide a more accurate recommendation for ski resorts, as we will do in our project.

4 PROPOSED WORK

4.1 Data Collection

The first step of our analysis is data collection. Our study makes use of open-source datasets from Kaggle, GitHub and government agencies. They include the Ski Resorts and Snow Coverage datasets [7] to understand ski resort types as well as snowfall distribution, Ticket Prices [4] to determine affordability, and Past Weather Data [6], [3] for model construction to predict snowfall for ski locations using live climate data. These datasets are easily downloadable from the respective websites as csv files, which can be further processed using the Python3 programming language in a Jupyter notebook environment.

4.2 Data Aggregation

The second step of our analysis is to aggregate data from the different dataset sources to create a combined dataset where possible for model construction and visualization. The aggregation process involves merging data columns from the respective datasets, enabling us to gain insights from a comprehensive view of skiing and provide skiers with in-depth information related to ski resorts. For model construction, aggregation is performed for the following climate features: temperature, relative humidity, wind speed, wind direction, snow depth. While it is difficult to aggregate the columns of the snowfall and ski resort datasets due to the time-dependence of snowfall dataset and the time-independence of ski resort dataset, we can overlay the datasets visually to analyze them simultaneously.

By integrating our datasets, we aim to deliver high-quality recommendations that consider key factors such as affordability, resort difficulty, snowfall, location, travel time, and weather conditions. Our comprehensive aggregated datasets will be instrumental in uncovering correlations among the various data sources, enabling us to analyze a broad spectrum of information. Leveraging these insights, we can provide users with highly tailored and accurate recommendations based on the identified patterns and relationships.

4.3 Data Pre-processing

The third step of our analysis is data pre-processing, where we ensure the data is of high quality and ready for analysis. This step involves removing duplicates, filling in missing or incomplete information, and cleaning out unnecessary data. This is particularly important when analysing the snowfall and climate datasets, where NaN values or erroneous values appear in bursts likely due to sensor malfunctioning or sensor shutdown. Such data, if untreated, will introduce undesirable bias and noise in our model construction process and must be augmented or removed beforehand. Due to the large size of the snowfall dataset, removal of erroneous and NaN data entries is performed as the cleaned dataset still remains significantly large for model training purposes, totaling up to 23580 entries.

4.4 Knowledge Mining

With the necessary data preparations and cleaning complete, the final step of our analysis is to look at our datasets for valuable patterns and trends to gain insights into how variables pertaining to skiing are related to and influence each other. Understanding the nuanced interrelations in our data guides us to curate the best recommendation based on preferences we believe skiers would be looking for.

Key patterns to be analyzed include the correlation relationship between features, such as the correlation between cost, skiing difficulty, and location with the popularity of a ski resort, to gain knowledge on the most crucial factors that shape the ski resort choice of the average skier. Such correlation analysis will be performed across all sets of possible features in our dataset to extract the maximum value out of them. In addition, time-series analysis will also be performed to understand the temporal evolution of some features and how relationships between our data features could change over time. We aim to analyze the correlation between snowfall levels and lift ticket prices to determine whether snow conditions influence the cost of skiing at a specific resort. By exploring this relationship, we hope to understand how weather impacts pricing strategies within the ski industry. This can involve exploring the historical trend of snowfall and climate conditions over time, understanding the general seasons where extremities or fluctuations can occur, and using this knowledge to perform time-series forecasting for future predictions.

In addition, data visualization tools like Seaborn and Matplotlib in Python help us visualize the distribution of our data with respect to each other and time. This helps us better identify outliers and trends across complex and multifaceted data like our aggregated dataset. Such visualization may also be provided to the end user for recommendation purposes to help the end user better understand the decision-making process of our algorithm.

4.5 Tools

The tools which we will adopt for the project are as follows.

(1) Jupyter Notebook
(2) Python3
(3) Microsoft Excel
(4) GitHub
(5) Python Libraries: Pandas, Numpy, Matplotlib, Seaborn, Scipy, Scikit-Learn, xgboost
(6) Kaggle and other datasets

5 EVALUATION

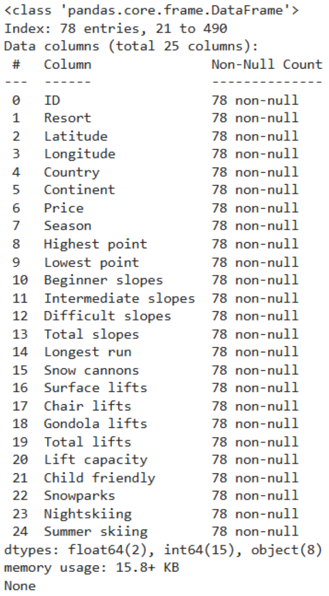
The evaluation of our ski resort recommendation system focuses on several key metrics that assess the system's ability to meet skier preferences and provide accurate, useful recommendations. These include the accuracy of recommendations, cost-effectiveness, and the alignment of ski run characteristics with skier skill levels.

(1) **Accuracy of Recommendations:** The effectiveness of the recommendation system is considered by examining how well the system matches skiers with resorts that align with their preferences. By analyzing correlations with the datasets, we aim to identify the most relevant factors that influence the skier's resort choice, such as location, difficulty, and snow coverage. Visual representations of these correlations, such as graphs and scatter plots, will help demonstrate the system's ability to provide accurate and targeted recommendations.

(2) **Cost-Effectiveness:** We also evaluate the system's ability to provide cost-effective recommendations by analyzing ticket pricing across different ski resorts. By examining the relationship between snow coverage and pricing, the system will help skiers make informed decisions based on both affordability and the value of snow conditions. This evaluation ensures that the system can provide recommendations that balance cost with the quality of skiing conditions.

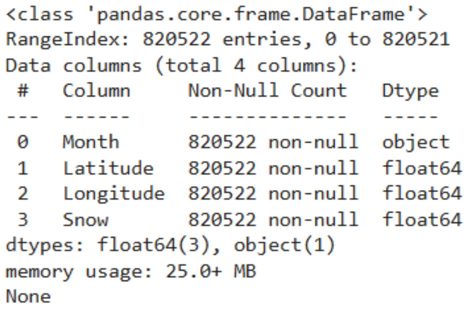
6 RESULTS

We have collected datasets from Kaggle and various government sources to ensure a robust data foundation. The datasets include the Ski Resorts and Snow Coverage datasets [7] in CSV format and Past Weather Data in both CSV[6] and netCDF[3] format. Data aggregation was performed for the weather datasets to combine the climate variables together with snowfall depth to facilitate model training. Finally, the datasets were cleaned through several pre-processing steps, including removing duplicate and erroneous entries, handling NaN values, discarding irrelevant columns, and filtering relevant rows. Figures 1, 2, 3 shows our datasets after aggregation and cleaning.



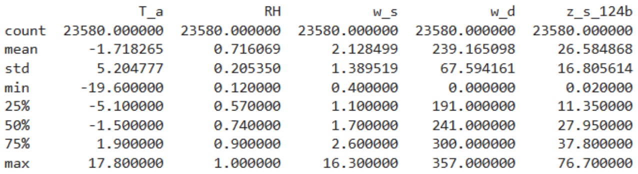
```
<class 'pandas.core.frame.DataFrame'>
Index: 78 entries, 21 to 490
Data columns (total 25 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   ID                   78 non-null     int64
 1   Resort               78 non-null     object
 2   Latitude             78 non-null     float64
 3   Longitude            78 non-null     float64
 4   Country              78 non-null     object
 5   Continent            78 non-null     object
 6   Price                78 non-null     int64
 7   Season               78 non-null     object
 8   Highest point        78 non-null     int64
 9   Lowest point         78 non-null     int64
 10  Beginner slopes      78 non-null     int64
 11  Intermediate slopes  78 non-null     int64
 12  Difficult slopes     78 non-null     int64
 13  Total slopes         78 non-null     int64
 14  Longest run          78 non-null     int64
 15  Snow cannons         78 non-null     int64
 16  Surface lifts        78 non-null     int64
 17  Chair lifts          78 non-null     int64
 18  Gondola lifts        78 non-null     int64
 19  Total lifts          78 non-null     int64
 20  Lift capacity        78 non-null     int64
 21  Child friendly       78 non-null     object
 22  Snowparks            78 non-null     object
 23  Nightskiing          78 non-null     object
 24  Summer skiing        78 non-null     object
dtypes: float64(2), int64(15), object(8)
memory usage: 15.8+ KB
None
```

Figure 1: Filtered US Ski Resorts Dataset Summary



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 820522 entries, 0 to 820521
Data columns (total 4 columns):
 #   Column     Non-Null Count   Dtype
---  ------     --------------   -----
 0   Month      820522 non-null  object
 1   Latitude   820522 non-null  float64
 2   Longitude  820522 non-null  float64
 3   Snow       820522 non-null  float64
dtypes: float64(3), object(1)
memory usage: 25.0+ MB
None
```

Figure 2: Snowfall Distribution Dataset Summary



```
               T_a           RH          w_s          w_d     z_s_124b
count  23580.000000  23580.000000  23580.000000  23580.000000  23580.000000
mean      -1.718265      0.716069      2.128499    239.165098     26.584868
std        5.204777      0.205350      1.389519     67.594161     16.805614
min      -19.600000      0.120000      0.400000      0.000000      0.020000
25%       -5.100000      0.570000      1.100000    191.000000     11.350000
50%       -1.500000      0.740000      1.700000    241.000000     27.950000
75%        1.900000      0.900000      2.600000    300.000000     37.800000
max       17.800000      1.000000     16.300000    357.000000     76.700000
```

Figure 3: Cleaned and Aggregated Weather and Snowfall Dataset Summary

For knowledge gaining, we first explored the features and their interrelations within each dataset to understand the distributions of the features better. For instance, since the price is likely to be a significant factor in recommending a suitable ski resort, we looked at the distribution of ticket pricing across all of the ski resorts in the United States, highlighted in figure 4.

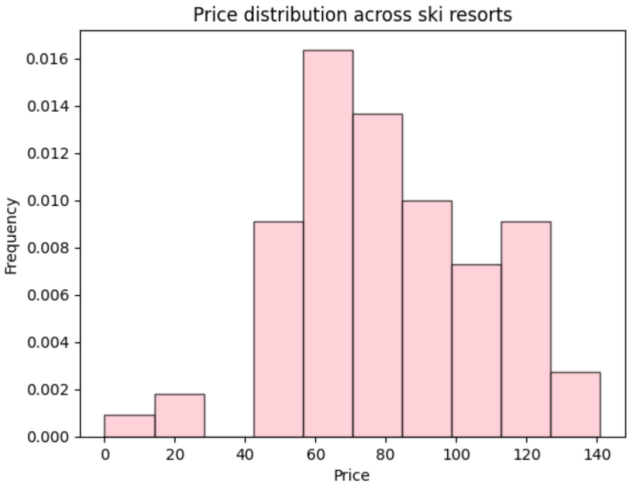


Figure 4: Price Distribution across Ski Resorts

Here, we see that most ticket pricing is clustered about the $50 - $100 range, with very few ski resorts falling below $40. This knowledge helps us to understand that $50 - $100 is what majority of skiers are expecting for ticket pricing when they decide to go skiing, and also that prices in the range below $40 are rare to come by, and definitely should be given higher emphasis by our recommendation for skiers who value affordable skiing destinations.

We wanted to see the distribution of pricing not only numerically but also geographically. Hence, we also investigated how geographical location and region across the United States may affect the pricing of ski resorts. Figure 5 shows a map of the United States with the various ski resort locations labeled and color-coded by pricing.

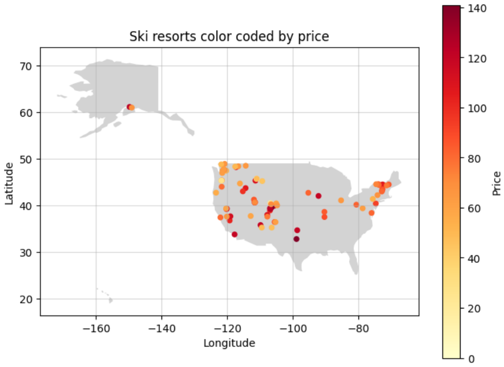


Figure 5: Geographical Distribution of Ski Resort Prices

This exploration reveals that some of the priciest ski resorts are located along the main mountain ranges of the United States, with good slopes that provide a suitable terrain for skiing. This knowledge indicates that the physical terrain of the ski resort is most certainly a strong contributor to the ticket pricing and must be accounted for by our recommendation. This is further backed up by our exploration of Figure 6, which reveals that some of the priciest resorts do occupy areas with the highest points for skiing. Hence, the recommendation must reflect that a skier who looks for the best slopes and highest points for skiing cannot expect low ticket pricing.

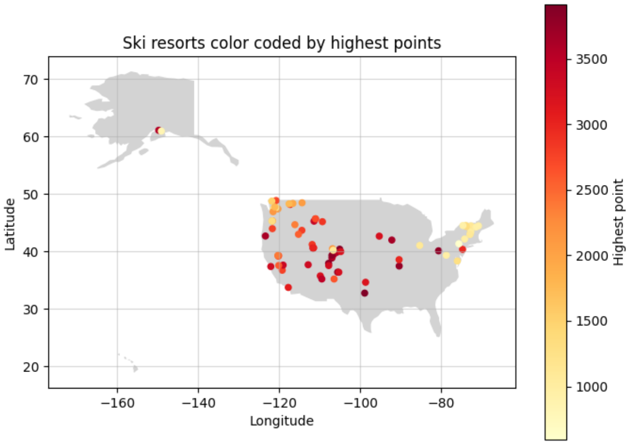


Figure 6: Geographical Distribution of Ski Resort Highest Points

In addition, we explored how pricing is affected by the snowfall conditions, the lift capacity, and other features at the ski resorts.

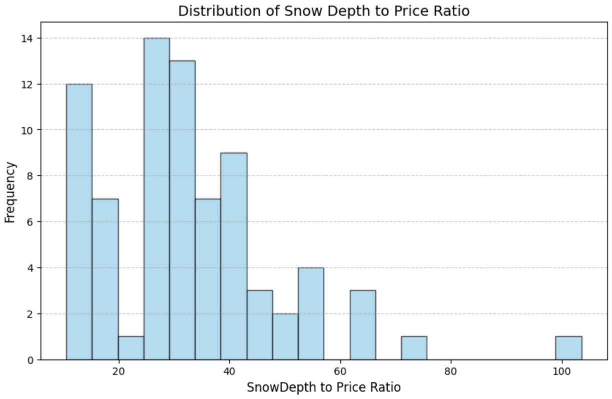


Figure 7: Snow Depth To Price Ratio

This histogram illustrates the distribution of the ratio between snow depth (highest point of the resort) and ticket price across different ski resorts. The x-axis represents the ratio of snow depth to ticket price, while the y-axis indicates the frequency of resorts within each ratio range. The exploration shows how resorts differ in their "value for snowfall"—that is, how much snow depth a skier gets for the price paid. A higher ratio indicates better value (more snow per dollar). The majority of resorts are clustered around certain ratio ranges, suggesting some standardization in pricing strategies based on snow depth.

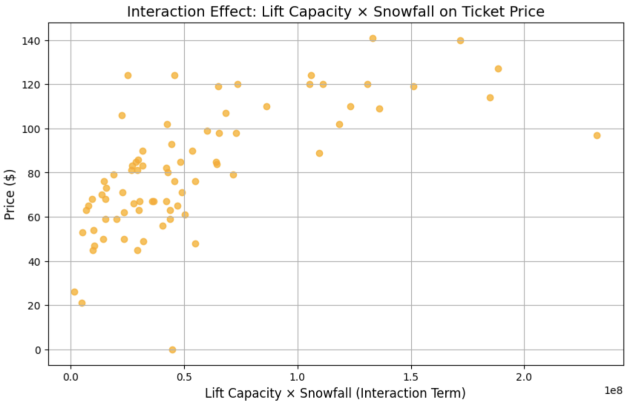


Figure 8: Lift Capacity × Snowfall on Ticket Price

This scatter plot visualizes the interaction between lift capacity and snowfall and its relationship with ticket prices at ski resorts. The x-axis represents the interaction term, which is a product of lift capacity and snowfall. Higher values suggest resorts with both high lift capacity and significant snowfall. The y-axis represents the ticket price in dollars. The exploration reveals a general trend where ticket prices increase as the interaction term grows, indicating that resorts with better infrastructure and more snowfall charge higher prices. They have a positive correlation.

We further explored how ticket pricing could be affected by other features, such as the services and amenities offered by the ski resort itself. To do this, we constructed a correlation heatmap showing the level of correlation between ticket pricing and different features of the ski resort, as shown in Figure **??**.

The heatmap reveals that ski resort pricing is most strongly correlated with both the lift capacities of the resort and the number of slopes, suggesting that the quality and quantity of facilities offered by the ski resort also influence the pricing of the resort on a level similar to that of the natural terrain of the ski resort. This will be factored into the recommendation for when skiers emphasize the availability of ample facilities in the resort.

Moving on, we also considered how to categorize the many ski resorts in the United States into a few main categories, which could help skiers to zone down on a particular category they may favor quickly, and greatly reduce the search space needed to find an optimal resort by narrowing the search to only that category. This can not only save time for skiers but also resources for the calculations. We approached this through a K-means clustering method to cluster the ski resorts into k clusters, leveraging unsupervised learning techniques. Without knowing the ideal cluster size K, we used the elbow method to first decide on a good value of K, which will be the value of K that gives the turning elbow point in Figure 10.

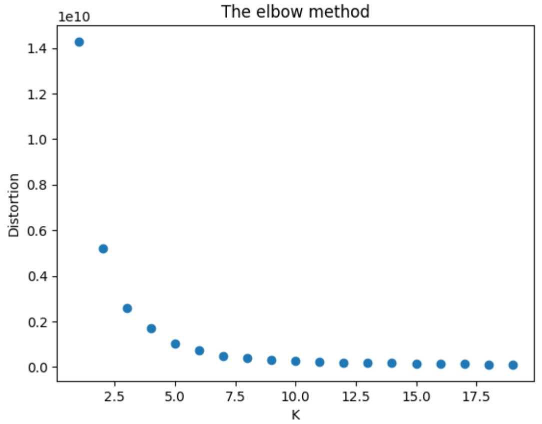


Figure 10: Elbow Method

The graph reveals that the ideal size for K lies around 3 to 4, so we performed K-means clustering using both values of K, shown in Figures 11 and 12.

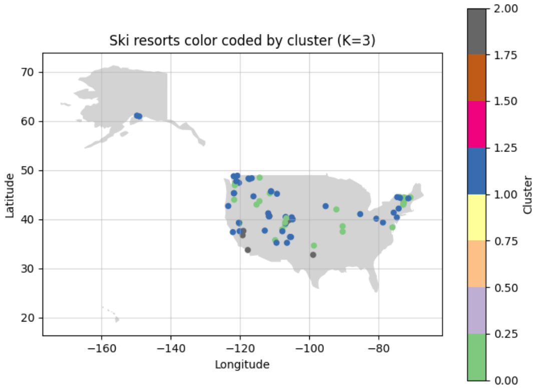


Figure 11: Clustering with K=3

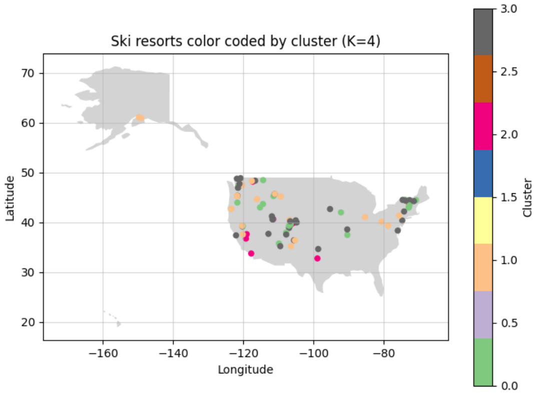


Figure 12: Clustering with K=4

The clustering results show a distinct cluster along the West Coast, with the remaining clusters distributed evenly across the country, suggesting that geography is not a primary factor in clustering. However, no resorts are in the southeast, indicating that because of flatter terrains, and warmer weather, these climates are not ideal as locations for resorts. To better understand the rationale behind the clustering results, we analyzed the characteristics and patterns across the clusters for the k=4 case. We looked at the box plots of the ticket prices, lift capacity, total lifts, and highest point across the resorts of the 4 clusters.

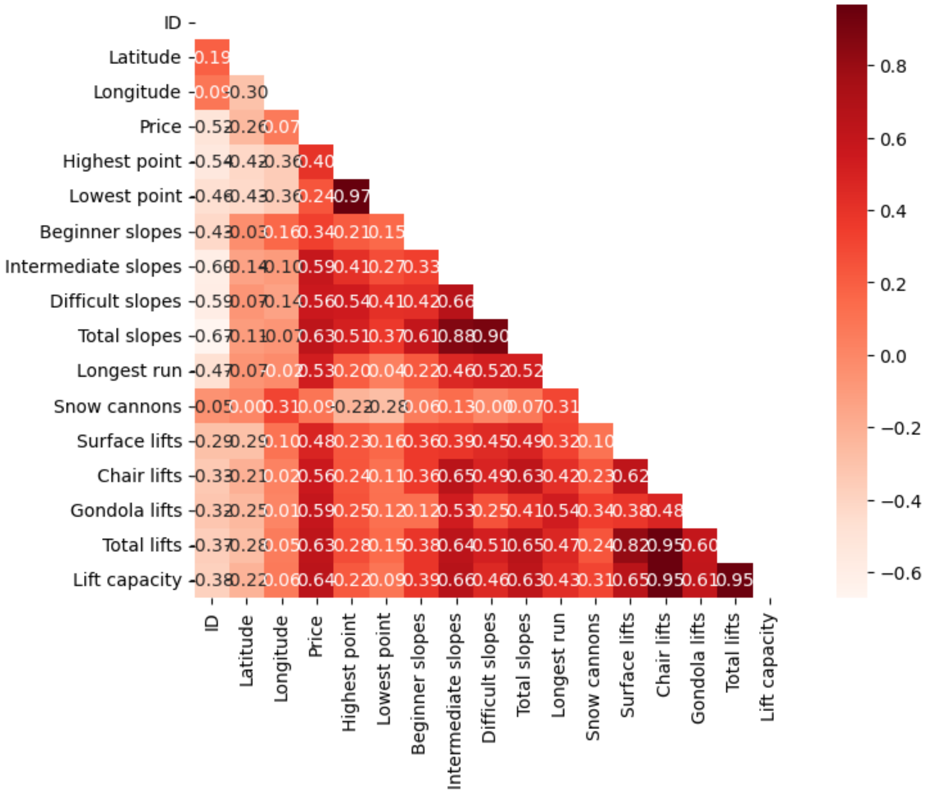


Figure 9: Heatmap with Resort Ratings

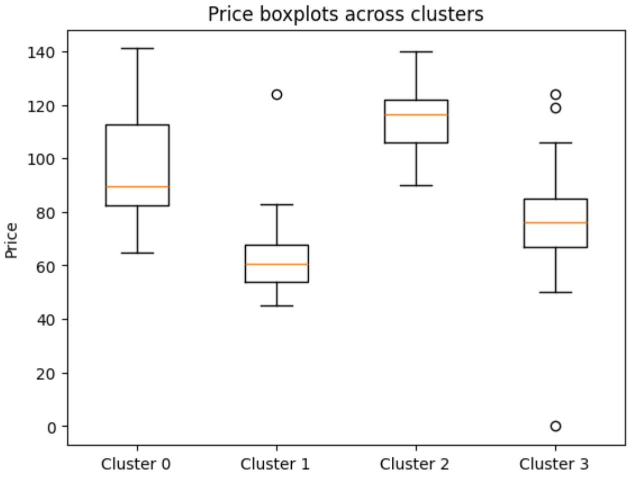


Figure 13: Boxplot of price across clusters

The price box plots show some level of distinction between the clusters, with cluster 2 having the highest price range and cluster 1 the lowest price range. However, there is significant overlap between the inter-quartile range of the clusters, suggesting that the pricing of the ski resorts is not a strong candidate for clustering.

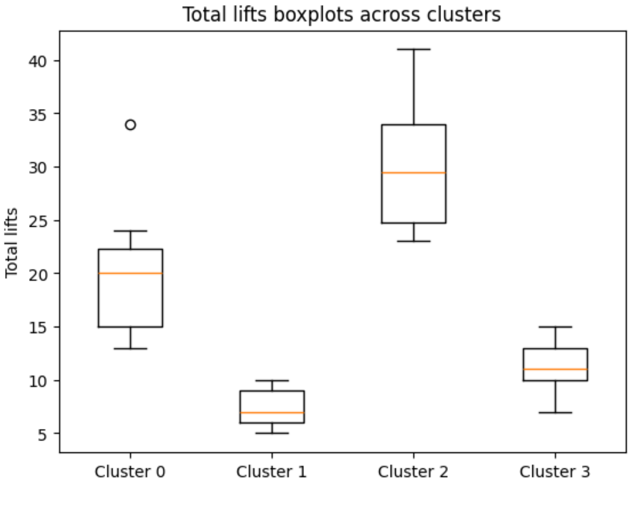


Figure 14: Boxplot of total lifts across clusters

The total lifts box plots show a good distinction between the clusters, with cluster 2 having the highest range of total number of lifts and cluster 1 the lowest. The overlap between the range of the clusters is minimal, suggesting that the total lifts of the ski resorts is a strong candidate for clustering.

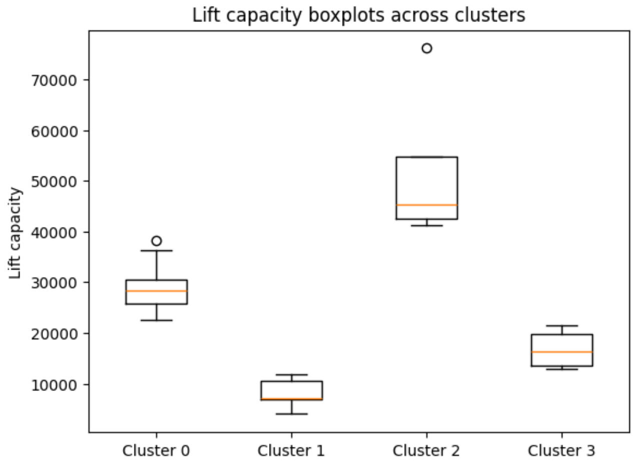


Figure 15: Boxplot of lift capacity across clusters

The lift capacity box plots show a clear distinction between the clusters, with cluster 2 having the highest range of lift capacity and cluster 1 the lowest. There is no overlap between the ranges of the clusters, suggesting that the lift capacity of the ski resorts is the best candidate for clustering.

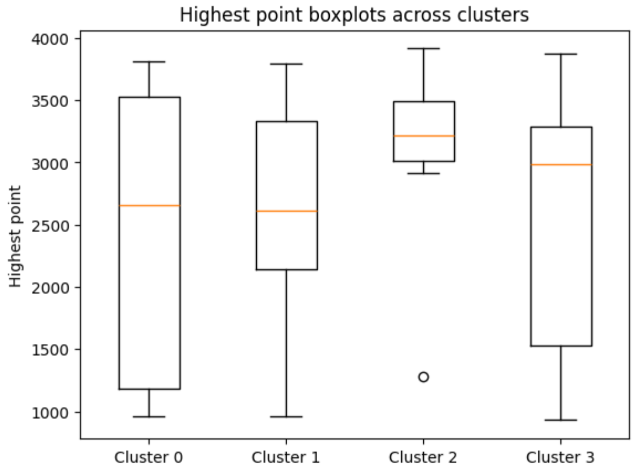


Figure 16: Boxplot of highest points across clusters

The highest point box plots show little distinction between the clusters, with no clear indication of which cluster has the greatest range of highest points. There is significant overlap between the inter-quartile ranges of the clusters, suggesting that the highest point of the ski resorts is a poor candidate for clustering.

Hence, from the analysis of the clustering results, it can be concluded that the clusters are split most strongly by lift capacity and least strongly by the highest points of the resorts. Since result is obtained from an unsupervised learning algorithm, it implies that the ski resorts across the United States have the most differentiation in terms of the lift capacity offered at the resorts and least differentiation in their highest points on average. When helping skiers narrow down their search for ski resorts by focusing on a specific cluster, it must be noted to skiers that such clustering would not be of significant value if highest point is a key consideration, but would significantly assist in the search if lift capacity is a key consideration.

We further explored the relation of ski resort pricing with the snowfall they receive, as snow is at the foundation of ski resorts. Thus, intuitively, the pricing should be positively correlated with snowfall, where more snow leads to more skiing activity opportunities. This is investigated by plotting snowfall received across the United States over the period of one year against the ski resorts with their pricing color-coded. The plots are split by month and documented in Table 1. Incorporating the temporal snowfall data reveals an interesting observation: some of the priciest ski resorts do not actually receive substantial snowfall through the year. Some of the ski resorts in the southwest and central regions of the United States have some of the highest prices, but yet the snowfall volume remains on the lower end of the bar. This could imply that such pricing are unwarranted, which must be incorporated into our recommendation to ensure that such ski resorts are penalized for high prices despite low snowfall levels which translates to less skiing opportunities.

At the same time, we aim to take into account the effects of climate change on snowfall by building a machine learning model capable of predicting based on current weather conditions. Such predictions will be helpful when combined with past snowfall trends to reflect a more accurate projection of snowfall. This is because with the effects of climate change, fluctuations in weather could mean that snowfall patterns deviate from the norm. To construct our model, we made use of the aggregated and cleaned past weather dataset that combines 11 years of snow and weather conditions, making a good spread over time to learn the relationship between weather parameters and snowfall. The model of choice is the xgboost regressor, where squared error is taken as the loss metric. The dataset is then randomly split into a training and testing subset where testing subset takes up 15% of the original dataset. The testing dataset is then further split such that the first 5000 entries are used as validation and the remaining for testing. In order to optimize the hyperparameters for the model, we adopted the GridSearchCV algorithm to iteratively fit and look for the best performing hyperparameters. After 32805 fits, the algorithm result is the following set of hyperparameters: {'alpha': 0.1, 'colsample_bytree': 0.8, 'lambda': 5, 'learning_rate': 0.05, 'max_depth': 6, 'min_child_weight': 5, 'n_estimators': 300, 'subsample': 0.8}. The finally errors for the model are a Mean Absolute Error (MAE) of 12.6818 and Root Mean Squared Error (RMSE) of 15.5756. Figure **??** shows the model's performance on the testing dataset and figure **??** shows the importance of the respective features used by the model, where w_d is wind direction, w_s is wind speed, RH is relative humidity, and T_a is temperature. It is rational that relative humidity has the strongest influence on snowfall since higher humidity indicates more water to form snow. It is however unexpected that temperature has a lower influence.

Another aspect we decided to explore is how actual popularity of the ski resorts are dependent on the previously discussed features and areas. All discussion on relationships to popularity have been

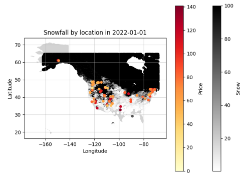
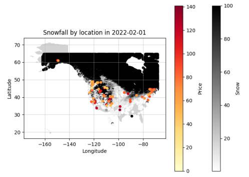
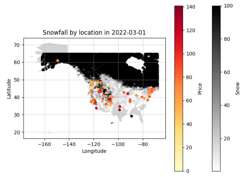
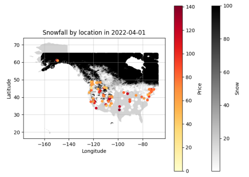
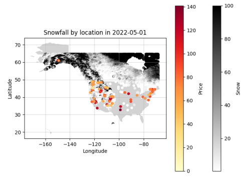
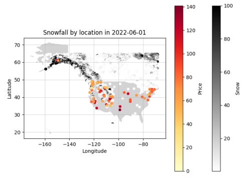
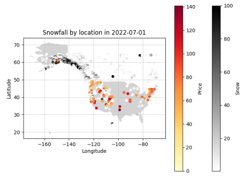
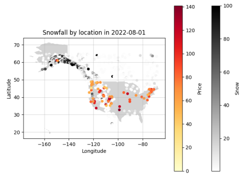
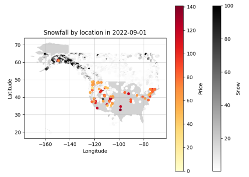
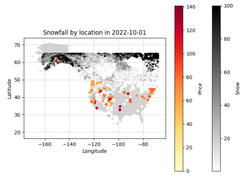
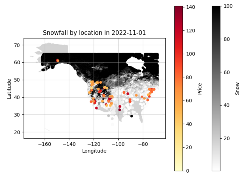
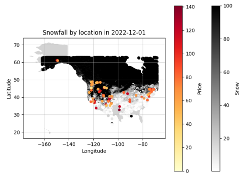
Nr.	1	2	3	4
1	 Snowfall by location in 2022-01-01 Latitude Longitude Price Snow	 Snowfall by location in 2022-02-01 Latitude Longitude Price Snow	 Snowfall by location in 2022-03-01 Latitude Longitude Price Snow	 Snowfall by location in 2022-04-01 Latitude Longitude Price Snow
2	 Snowfall by location in 2022-05-01 Latitude Longitude Price Snow	 Snowfall by location in 2022-06-01 Latitude Longitude Price Snow	 Snowfall by location in 2022-07-01 Latitude Longitude Price Snow	 Snowfall by location in 2022-08-01 Latitude Longitude Price Snow
3	 Snowfall by location in 2022-09-01 Latitude Longitude Price Snow	 Snowfall by location in 2022-10-01 Latitude Longitude Price Snow	 Snowfall by location in 2022-11-01 Latitude Longitude Price Snow	 Snowfall by location in 2022-12-01 Latitude Longitude Price Snow

Table 1: Snowfall against Ski Resort Price Distribution (1 year)

theoretical up to now. To understand their relationships in reality, we incorporated the online rating popularity of the ski resorts into our dataset. The rating data is extracted from the Global Ski Resort Rankings Dataset from OPENICPSR[10], and added as a column to our existing ski resort dataset. This data aggregation led to the reduction of the dataset as ski resorts without corresponding available ratings were dropped. Nonetheless, there remains 61 resorts from across all of United States, forming a good representation still.

Generating a new correlation heatmap including the resorts ratings reveals interesting relationships between the resort features and resort ratings. In Figure 17 below, the resort features with the strongest correlations with ski resort ratings are the price, total number of slopes and the number of intermediate and difficult slopes. While correlation does not imply causation, it is likely that the correlation with the price is actually a form of causation where higher rating of the resort causes higher prices. This is reasonable since a higher rating makes the demand for the ski resort more price inelastic, thereby calling for price hikes for profit maximization. This must be factored in when making recommendations to skiers who value the rating of resorts, as skiers must expect that a ski resort with high ratings is most likely more expensive as well. The strong correlation of ratings with the slopes of the ski resorts is another important observation which could be used to explain why some ski resorts which receive less snowfall remains popular with high prices, an observation made earlier. The strong correlation with slopes suggests that skiers look out for the variety of runs they can experience above all else. This implies that even if a resort does not receive the most snowfall, as long as there is sufficient snowfall to facilitate skiing, skiers would value the resort more if the resort offers a greater variety of slopes. Thus, it is likely that the previously identified resorts with high prices that do not receive ample snowfall offer a large number of slopes, providing them with a sound basis for maintaining high prices. Taking into account this observation, the recommendation system will need to assign the total slopes a high weight when generating recommendations, given that the slope variety is a feature that skiers value significantly. However, one more observation is that while the number of intermediate and difficult slopes have a strong correlation with resort rating, the number of beginner slopes have a very weak correlation. This abnormality in trend suggests that it is likely that the ratings for resorts are left by those who value the intermediate and difficult slopes. This is reasonable as those who leave ratings are more likely to be skiers truly passionate about the sport, as compared to individuals just seeking a regular vacation. Being passionate in the sport also suggests that they are well-trained and aiming to take on the higher difficulty slopes. However, this means that the ratings for the resorts may not reflect the true visitor-ship of the resorts, since it does not encompass information regarding those trying beginner slopes or seeking a casual vacation, whose numbers may actually surpass that of passionate skiers.

Thus, we decided to further investigate the relationship between the rating received by a resort and the total number of visitors of the resort. To do this, we compared the visitor numbers to the ratings received for 10 resorts, matching with the most-visited ski resorts dataset from Statista[9]. We created a new dataset through joining our current ski dataset with that from Statista on rows where the ski resorts name match, and thereafter investigated the level of correlation between the rating and visitor number of the resorts.

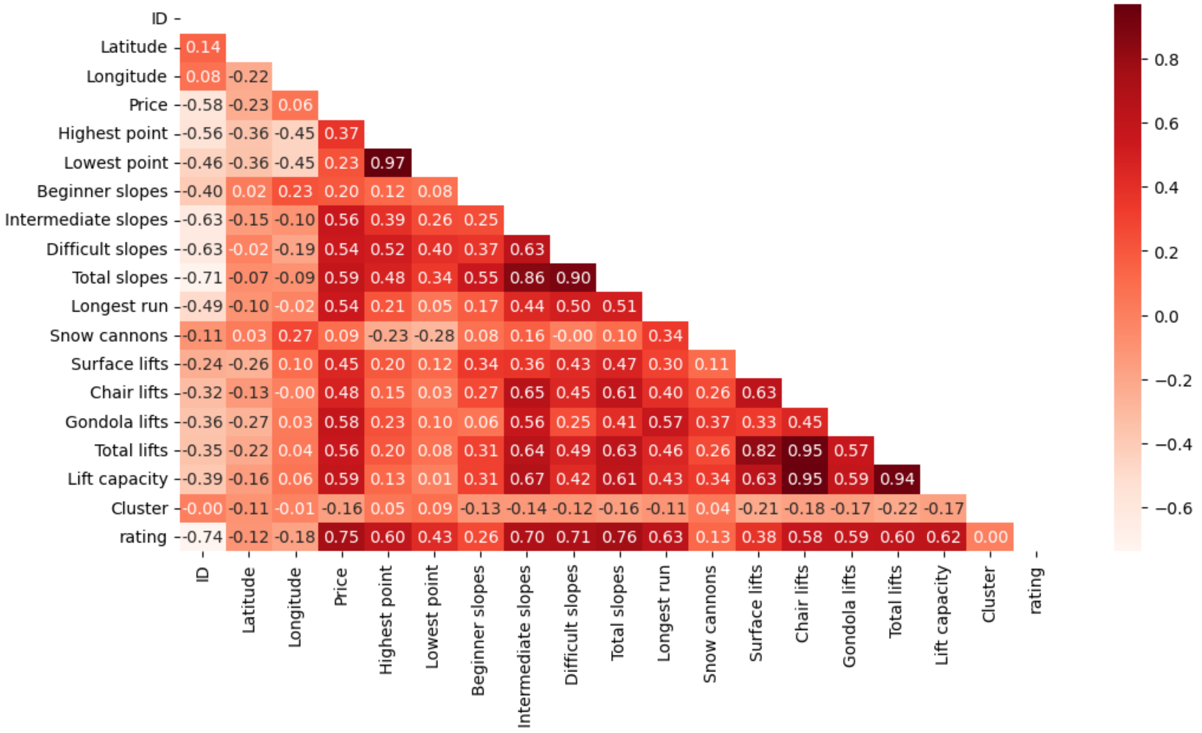


Figure 17: Heatmap with Resort Ratings

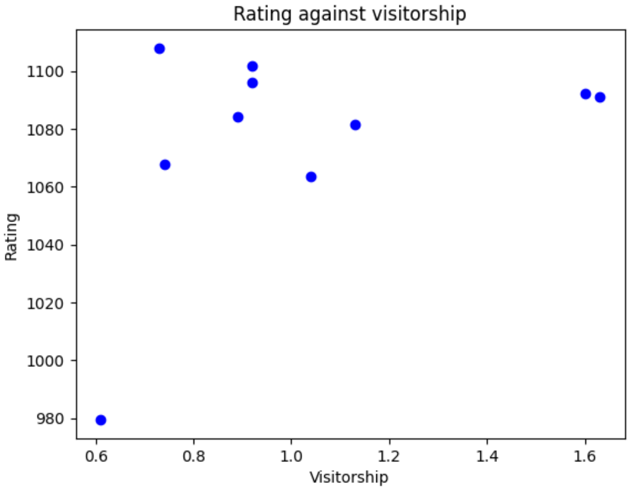


Figure 18: Ratings against Visitor Numbers

Plotting the ski resorts rating against visitor number yields the following results. Pearson Correlation coefficient: 0.39908858911983636, P-value: 0.25325521894960146. The small Pearson coefficient suggests that there is no clear linearity between the two, and the large p-value suggests the correlation is likely due to random chance. This establishes that the high ratings of a resort does not usually translate into high visitor numbers, due to the exclusion of casual vacationers in the rating statistics as aforementioned. As a result, the recommendation system developed using ratings will not be applicable to all skiers, but could come in more useful for the better-trained and prepared skiers. To create recommendations that are applicable to all skiers, data analysis using visitor numbers per resort must be used instead. However, such data is not readily and openly available for all resorts, as some resorts may choose to keep it confidential, unlike rating data that is publicly viewable. Hence, further improvements can be made for the recommendation system by performing extensive mining for visitor number data for the ski resorts and conducting further analysis using them.

7 DISCUSSION

In the 2023/2024, season there were 486 ski areas operating in 37 states. New York holds the lead for the most operating ski resorts, followed by with Michigan having 40 resorts. There is a widespread availability of resorts in the US, and considering different options and preferences.

This project was originated to create a recommendation system that caters to users preferences by identifying suitable ski resorts. We initially focused on ski resorts EPIC and IKON network. When we progressed, we realized that the availability for data was limited. We then addressed this by expanding the scope of our resorts to be all resorts across the United States. This significantly helped us to improve our resources but also introduced other challenges in making sure the datasets were consistent and relevant.

A key challenged faced was determining what factors to include in our recommendation system. Because ski resorts have so many different varying offerings like snowfall, amenities, cost, etc, we had to incorporate these elements effectively and carefully consider them. These challenges showed us how important it was to narrow down what factors that would be most impactful for users.

Collaboration was critical in overcoming many of the challenges we faced. Being able to pool our teams expertise, we were able to brainstorm and test strategies for data analysis. We identified gaps in our data and supplemented it with external resources if it was possible. Our team was able to greatly adapt and innovate when we faced challenges, which helped us make our project successful. allowed us to save time and resources for calculations and analysis.

7.1 Future Ideas

Our groups future plans for this project is to transform this into a fully developed application that provides users with a highly interactive and personalized experience. The app would allow users to input specific preferences—such as desired ski run difficulty, budget, and travel distance—and receive tailored ski resort recommendations that align perfectly with their needs. Additionally, we aim to enhance our datasets with real-time updates, particularly for snow conditions, ensuring the most accurate and up-to-date information is available.

An important consideration for future development is the size of ski resorts. Small resorts, which make up 59% of all operations but attract only 13% of total skiers, represent an untapped opportunity to tailor recommendations for a broader range of user preferences. Incorporating resort size as a factor could add depth to the recommendations, helping users discover hidden gems that might otherwise be overlooked.

Beyond generating recommendations, the envisioned app would offer advanced features to elevate the user experience. These would include user reviews, real-time weather updates, hourly snow reports, and even resort-specific services such as lift ticket booking. By integrating these features, we aim to create a comprehensive resource for ski enthusiasts. With its personalized recommendations, live updates, and community-driven feedback, this app has the potential to become the ultimate go-to platform for skiers and snowboarders, delivering convenience and expertise in one seamless package.

8 CONCLUSION

In this project, we developed a ski resort recommendation system using data mining methodologies to address the common needs and preferences of skiers and snowboarders when choosing the ideal resort. By integrating datasets from various sources, including resort characteristics, climate data, and ticket pricing, we developed a comprehensive framework to uncover and analyze correlations between key factors influencing ski resort selection. The project emphasized key steps, including data collection, aggregation, preprocessing, and knowledge mining, to uncover meaningful patterns and correlations.

Our analysis revealed key insights into the factors driving ski resort desirability, including the significant influence of lift capacity, snowfall, and pricing. By applying strategies learned in class, we utilized K-means clustering to categorize ski resorts based on critical attributes, demonstrating a practical method for narrowing down options and simplifying the recommendation process. Furthermore, we implemented advanced techniques, such as the XGBoost regressor, to predict snowfall trends. This approach integrated knowledge of climate variability, allowing us to account for dynamic weather conditions in our recommendations—a method reflecting the principles of data mining and predictive modeling emphasized in our coursework.

Despite challenges such as finding datasets that matched the specific information we were seeking and aligned with those we had already selected, the project successfully delivered actionable insights. Another significant hurdle was determining which data was most relevant to our objectives and filtering out information that offered little value. Through careful analysis and strategic decision-making, we identified key pricing trends, the role of resort features in determining popularity, and the relationship between snowfall and cost-effectiveness, demonstrating the effectiveness of the methodologies we applied.

Moving forward, there is significant potential to expand our extensive dataset into a fully functional system capable of analyzing customer preferences and providing personalized ski resort recommendations based on their inputs. The details of this vision are outlined in the Future Ideas section, which highlights the opportunities for further development and enhancements.

Our project demonstrates how data-driven methods can be effectively applied to provide meaningful insights, creating a strong foundation for future improvements in personalized ski resort recommendations.

9 APPENDIX

On my honor, as a University of Colorado Boulder student, I have neither given nor received unauthorized assistance.

Contributions:

Juncheng Man contributed by selecting and analyzing relevant datasets for the project, including datasets on resorts features, snowfall, weather, and resorts popularity. He performed the pre-processing and cleaning of the datasets and aggregated the datasets into combine datasets. He also investigated the inter-relationships of the various features in the combined dataset and rendered the results graphically to assist with analysis (all figures aside from figures 7 and 8). Finally, he was responsible for training a supervised machine learning model to predict snowfall based on live weather conditions, and contributed to the draft of the reports.

Justin Nguyen contributed by researching and identifying valuable datasets for the project. We cleaned and sorted through the data sets to find what is most relevant to the project. He assisted with data pre-processing and used these datasets to create useful graphs that advanced our analysis (figures 7 and 8). Additionally, he helped identify key features for the recommendation system, supporting the team in determining the most important factors in selecting the best ski resort for the user. He focused on. I also contributed in writing the project report and making all of the presentations.

Grace Waida contributed to the project by conducting research to identify the key attributes the group aimed to analyze when identifying datasets. She sourced relevant datasets that aligned with the project requirements and tested pre-processing ideas using Excel. Additionally, Grace provided support to the team by assisting with tasks as needed and helped develop both the presentation and the final report.

Audrey Ly contributed by searching for datasets that would be useful for the project. She researched what questions to ask using the datasets, and also helped to narrow down them down to focus on the most important factors in the selection of ski resorts. She also helped with data pre-processing and helped to analyze the relevant datasets. Additionally, she supported various tasks throughout the project and collaborated in the development of the final report.

REFERENCES

[1] Forbes. U.s. ski industry had 5th busiest season despite record warm temps in 2023-24, 2024.

[2] Guardian. Ski resorts' era of plentiful snow may be over due to climate crisis, study finds., 2024.

[3] Laboratory, O. R. N. Daymet: Daily surface weather data on a 1-km grid for north america, version 4 r1, 2022.

[4] Lai, S. Onthesnow, 2019.

[5] Malcom, O. C. Find a ski resort for your next vacation: Data analytics project, 2023.

[6] of Agriculture, D. Data from: Eleven years of mountain weather, snow, soil moisture and stream flow data from the rain-snow transition zone, 2019.

[7] Pedersen, U. T. Ski resorts, 2022.

[8] Serdouk, Y., C. T. C. E. . M. C. Ski resorts recommendation using deep neural networks. *Proceedings of the Workshop on Recommenders in Tourism Co-Located with the 15th ACM Conference on Recommender Systems (RecSys 2021) 2974* (2021), 85–89.

[9] Statista. Most-visited ski resorts in the united states between the 2007/2008 and the 2010/2011 seasons (in millions)*, 2012.

[10] Tank, T. Global ski resort rankings dataset, 2024.

[11] Toujas, A. Data web scraping to help plan a ski vacation, 2017.