SHOULD WE GO SKIING? ✕❄✕

Evaluating decision-making conditions and behaviors when choosing to participate in snowsports at a Norwegian ski resort

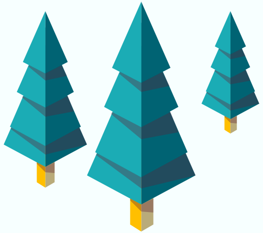
LISA GIRARD
Thinkful Data Science
Capstone III: Unsupervised Clustered Learning
January 10, 2023

Data Source & What Can We Learn?

Data originally collected in 2018 at Hafjell Ski Resort, Norway. Raw data collected via conjoint survey collection and analysis.
400 respondents, 7,200+ response rows, 87 features.

Research Question:

What conditions contribute to the decision to participate in snowsports (ski/snowboard) at a ski resort?



Suggested Audience

MARKETING

Brand management, advertising/marketing, business/customer development, media relations, internal/external communications

OPERATIONS

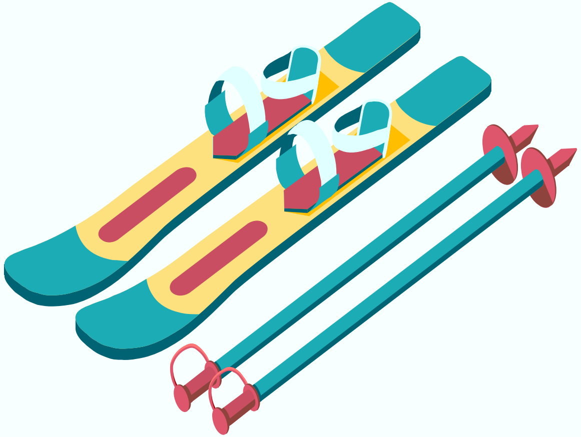
Lifts, food/hospitality, rentals/equipment, maintenance, snowmaking, facility and terrain management, transportation

C-LEVELS

C-Level leaders: Management/Executive, Finance, HR, Marketing/Communications, Operations, Technology, Risk Management, Legal

EXPLORING THE DATA

CONSIDERED FACTORS



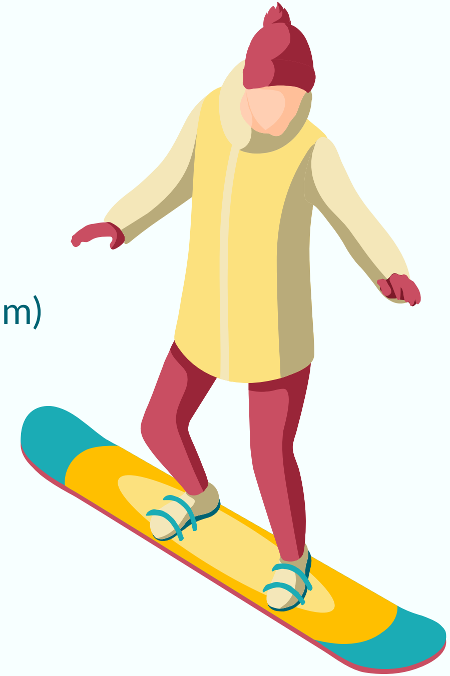
- Weather conditions
- Lift line wait
- Regular week/vacation
- Weekday / weekend
- Percentage of slopes/runs open

- Family status
- Distance from resort
- Interest in skiing/snowsports
- Age
- Gender (M/F)

The Process

90% random sample   n=4277

1. Created subset data with above conditions
2. Scaled data set
3. Applied 4 clustering models
    a. KMeans
    b. Agglomerative Clustering
    c. DBSCAN
    d. GMM
4. Used clustering suggestions per model (elbow, dendrogram)
5. Applied various dimensionality reductions
    a. PCA
    b. TSNE
    c. UMAP
6. Calculated Silhouette scores per model and per model dimensionality reduction

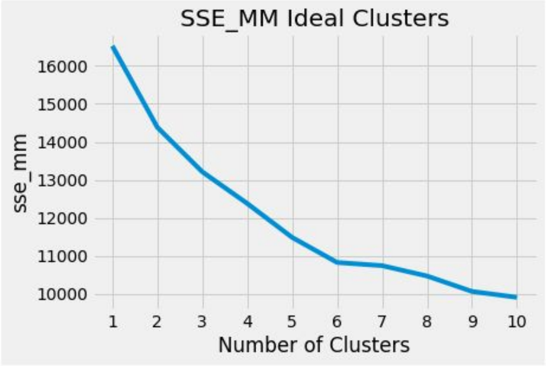


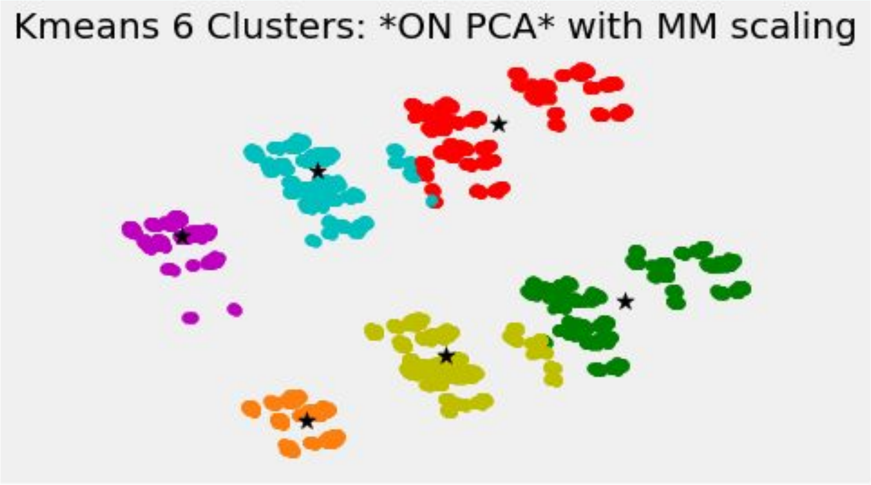
KMeans Clustering

90% random sample   n=4277

6 Defined Clusters

- 90% random sampling   n=4277
- MinMax scaled data (values 0 to 1)
- PCA Silhouette Score: 62%





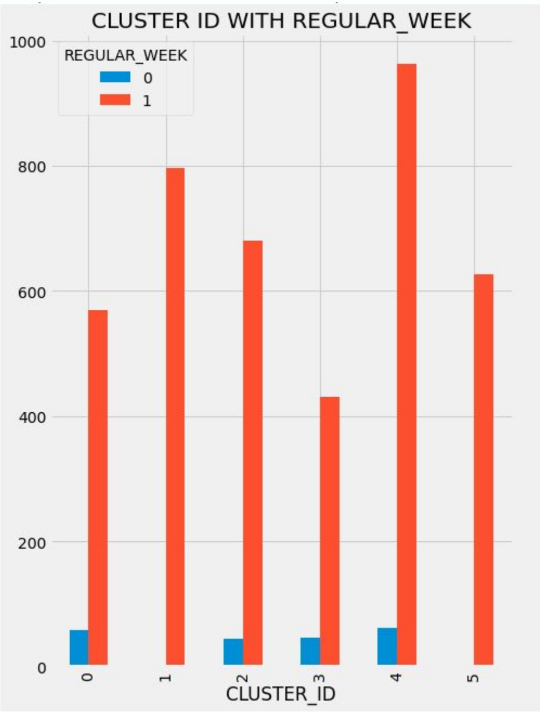
KMeans Clustering

90% random sample   n=4277

Cluster Characteristics

TIME FRAME:
REGULAR WEEK vs VACATION

- All clusters indicated high attendance during regular weeks
- Very few clusters indicated attendance during vacation periods



KMeans Clustering

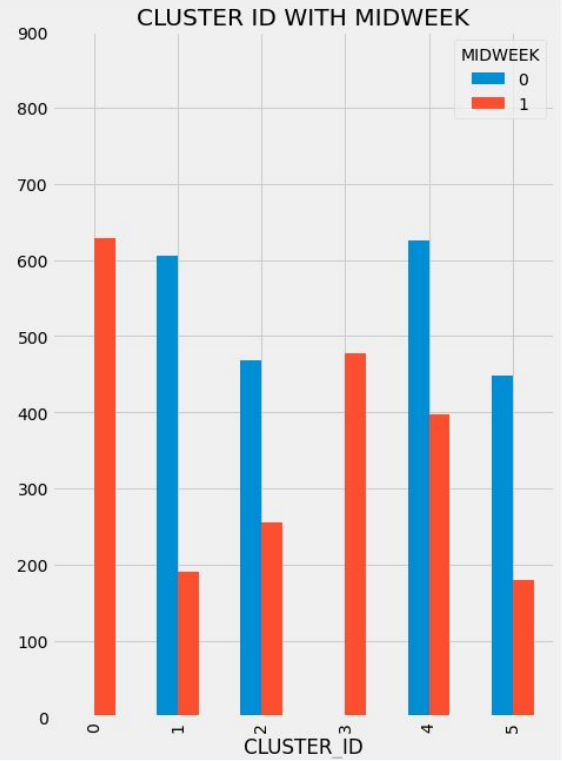
90% random sample   n=4277

Cluster Characteristics

TIME FRAME:
MIDWEEK vs WEEKEND

- Cluster0 and Cluster3 indicated only visiting resort during midweek, 0 weekend visits

- Clusters 1,2,4,5 indicated higher weekend visits



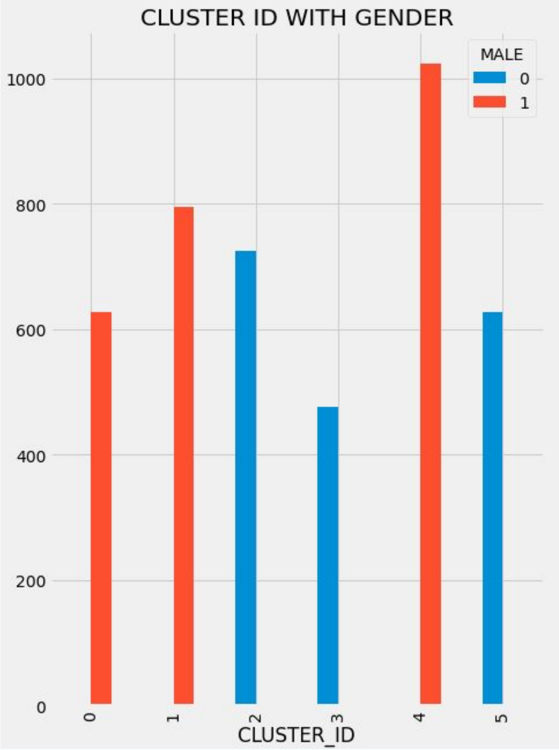
KMeans Clustering

90% random sample   n=4277

Cluster Characteristics

GENDER:
MALE vs FEMALE

- Clusters 0,1,4 are all Male
- Clusters 2,3,5 are all Female



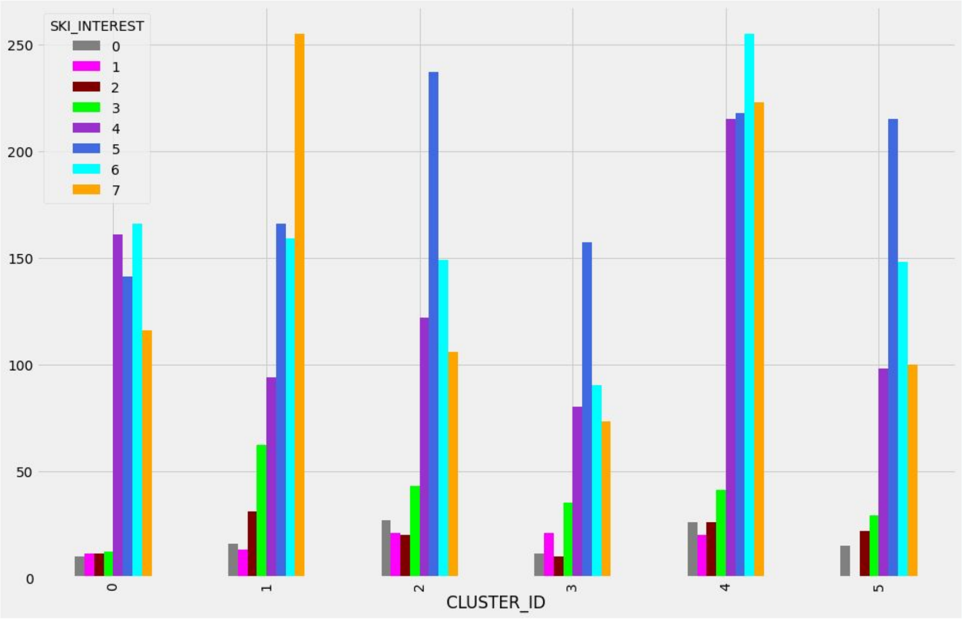
KMeans Clustering

90% random sample   n=4277

Cluster Characteristics

SKI INTEREST:
1-low interest to 7-high interest

- Cluster1 indicates highest interest, Cluster4 at 2nd highest interest
- Clusters 2,3,5 indicated mid interest
- Cluster0 shows most balanced interest with “4,5,6” indicated for mid-range interest



Interest seems to average about 5 for all clusters

KMeans Clustering

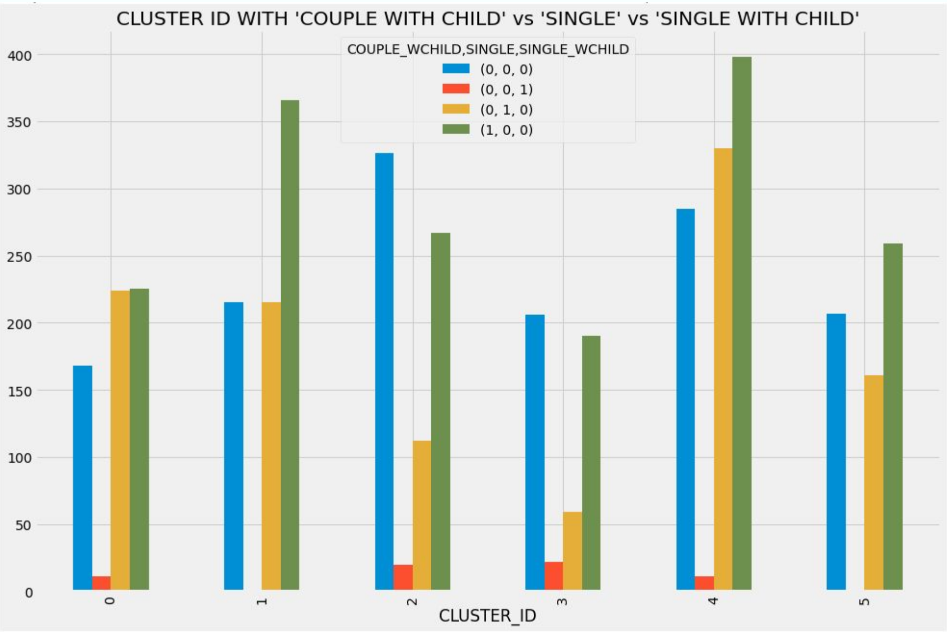
90% random sample   n=4277

Cluster Characteristics

**FAMILY STATUS:
COUPLE wCHILD vs SINGLE
vs SINGLE wCHILD**

- Low indication across all clusters of “Single With Child”
- Clusters 1,4,5 indicate more family visitors with “Couple With Child”

- SINGLE vs NOT-SINGLE:
    - *More group skiing taking place across clusters than solo single skiing (blue > yellow)*



KMeans Clustering

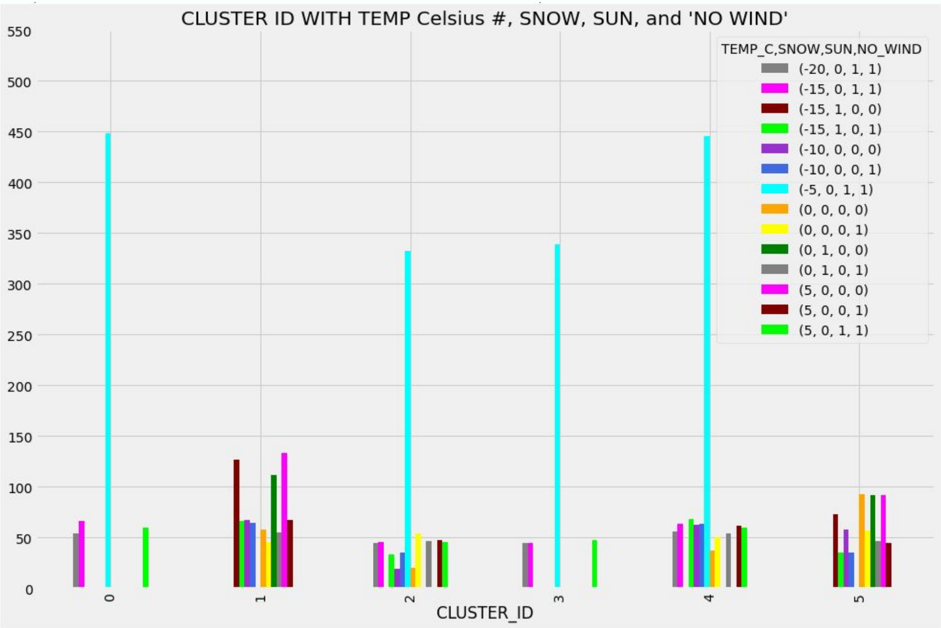
90% random sample   n=4277

Cluster Characteristics

WEATHER CONDITIONS:
TEMP vs SNOW vs SUN vs WIND

- Almost all clusters indicate visitors on "bluebird days" at
  - -5C (23F)
  - no snow
  - a sunny day
  - no wind

- Cluster5 - balance between 0-5C, no snow, no sun, windy and snowy



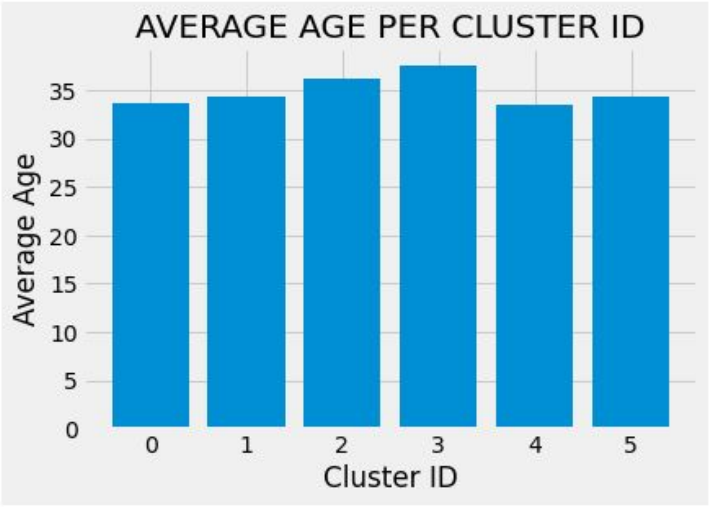
KMeans Clustering

90% random sample   n=4277

Cluster Characteristics

AGE:
34 across all clusters

- Cluster0: 34
- Cluster1: 34
- Cluster2: 36
- Cluster3: 38
- Cluster4: 34
- Cluster5: 34



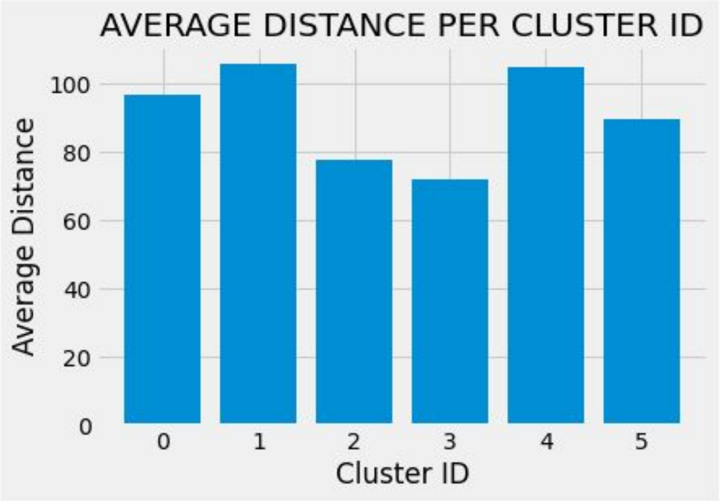
KMeans Clustering

90% random sample   n=4277

Cluster Characteristics

DISTANCE:
91 kilometers for all clusters

- Cluster0: 97km
- Cluster1: 106km
- Cluster2: 78km
- Cluster3: 72km
- Cluster4: 105km
- Cluster5: 90km



KMeans Clustering

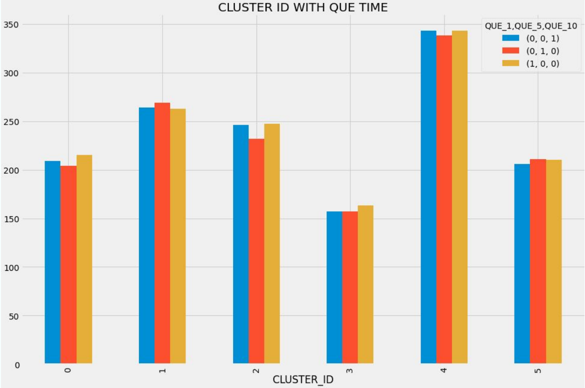
90% random sample   n=4277

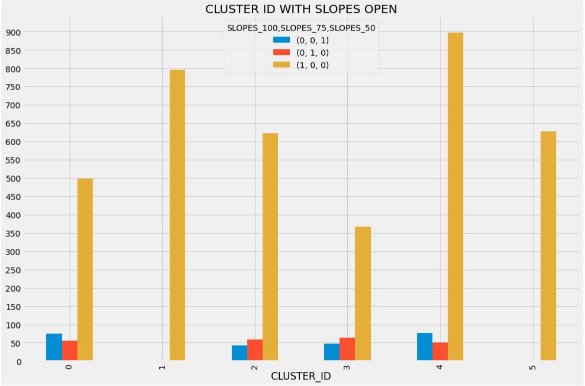
Cluster Characteristics

LIFT LINES & SLOPES OPEN:
Que time 1,5,10 mins
Slopes open 50%, 75%, 100%

- Lift line wait times were evenly distributed among the clusters
- Slopes 100% open predominantly dominated all clusters

With que times evenly distributed and slopes open percentage so high at 100%, there's not enough variance to consider these as determining cluster factors.





Recap: CLUSTERO

- Mostly regular week (not vacation)
- All midweek
- All male
- Mid-range interest
- Couple-with-child, single, not-single
- Bluebird day (weather)
- Average age: 34
- Average distance from resort: 97km

Recap: CLUSTER1

- All regular week (not vacation)
- Mostly weekend
- All male
- High interest
- Couple-with-child, single = not-single
- 0-5C, snowy/no snow, cloudy, windy days
- Average age: 34
- Average distance from resort: 106km

Recap: CLUSTER2

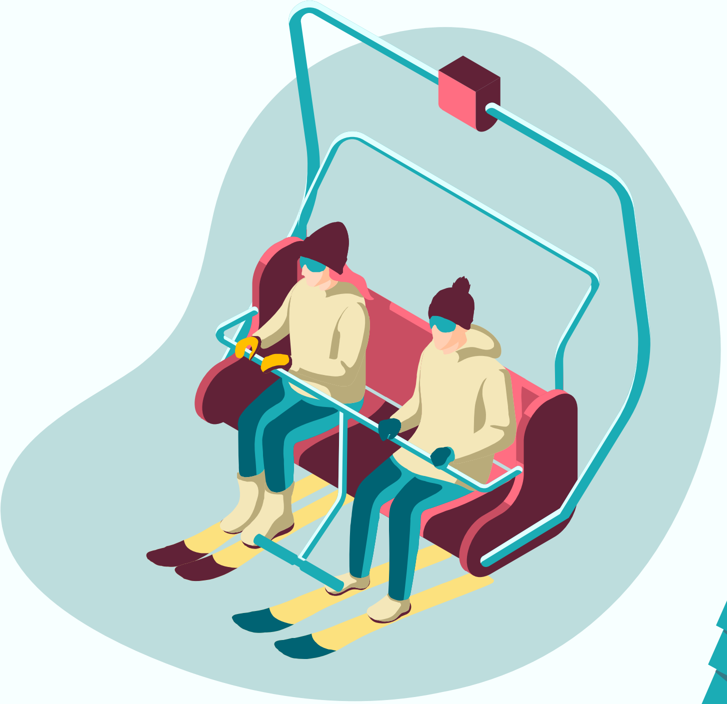
- Mostly regular week (not vacation)
- Mostly weekend
- All female
- Mid-range interest
- Not-single, couple-with-child, single
- Bluebird day (weather)
- Average age: 36
- Average distance from resort: 78km

Recap: CLUSTER3

- Mostly regular week (not vacation)
- All midweek
- All female
- Mid-range interest
- Single, couple with child
- Bluebird day (weather)
- Average age: 38
- Average distance from resort: 72km

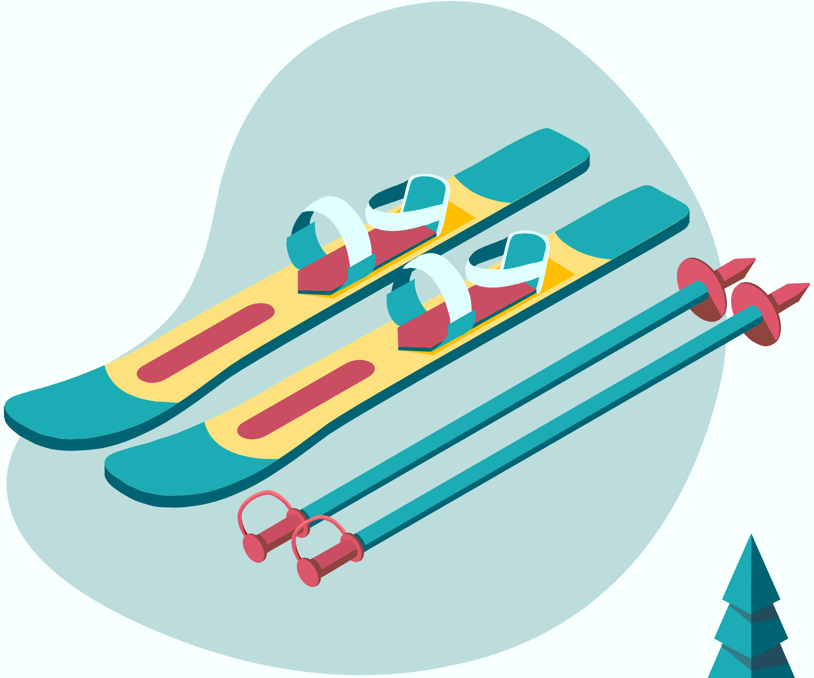
Recap: CLUSTER4

- Mostly regular week (not vacation)
- Mostly weekend
- All male
- High interest
- Couple-with-child, single, not-single
- Bluebird day (weather)
- Average age: 34
- Average distance from resort: 105km



Recap: CLUSTER5

- All regular week (not vacation)
- Mostly weekend
- All female
- Mid-range interest
- Couple with child, not-single, single
- 0-5C, no snow/snowy, cloudy, windy days
- Average age: 34
- Average distance from resort: 90km



ALTERNATIVE MODELS

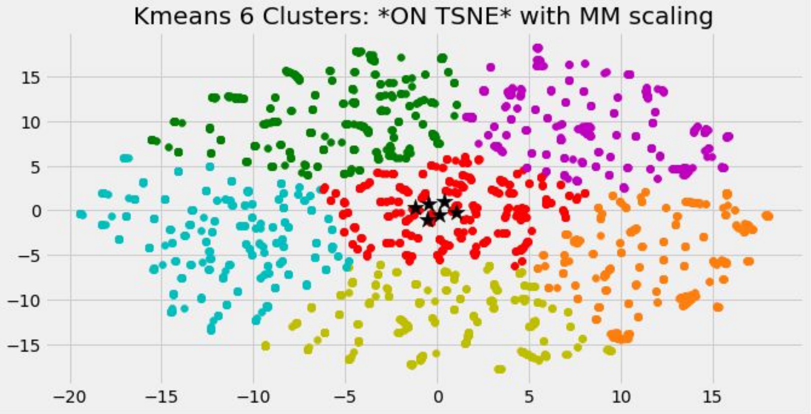
KMeans TSNE / UMAP

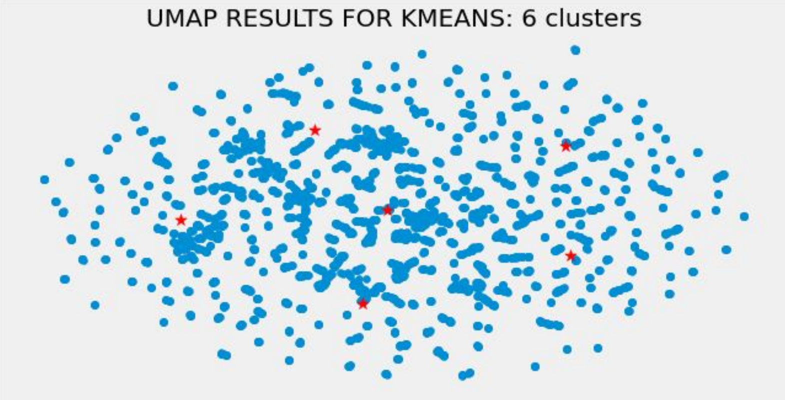
90% random sample n=4277

Dimensionality Reduction

TSNE & UMAP

- No defined clusters with either
- TSNE Silhouette Score: 37%
- UMAP Silhouette Score: 36%



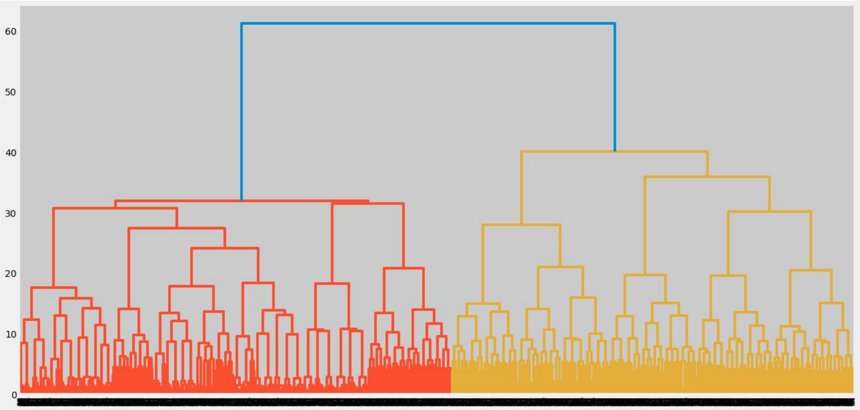


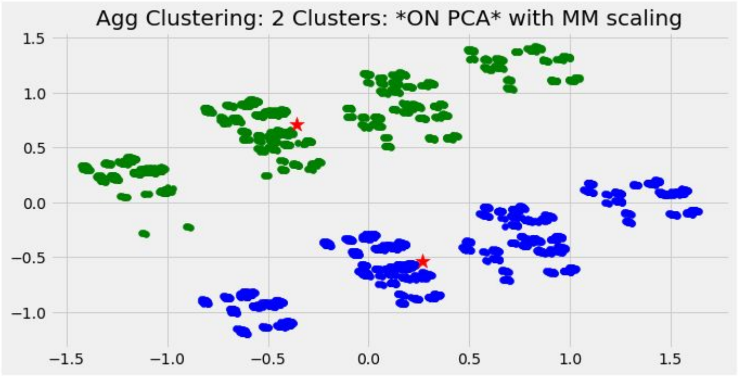
Agglomerative Clustering

90% random sample   n=4277

2 Defined Clusters

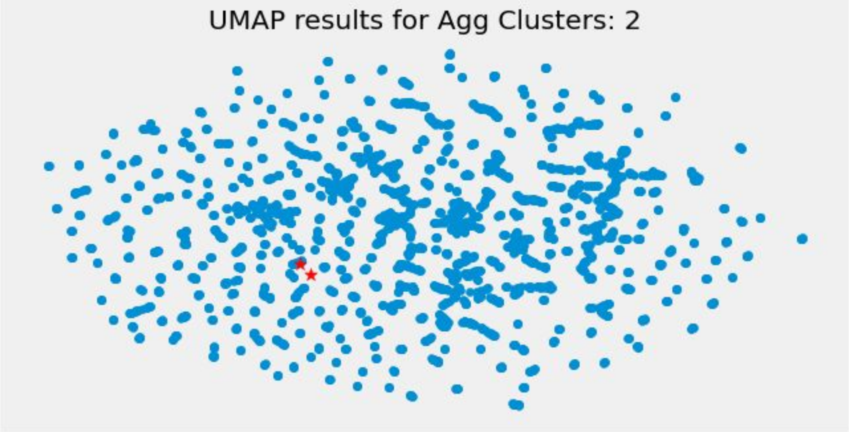
- Dendrogram indicates 2 optimal clusters
- PCA dimensionality reduction
- PCA Silhouette Score: 49%





Agg Cluster TSNE / UMAP

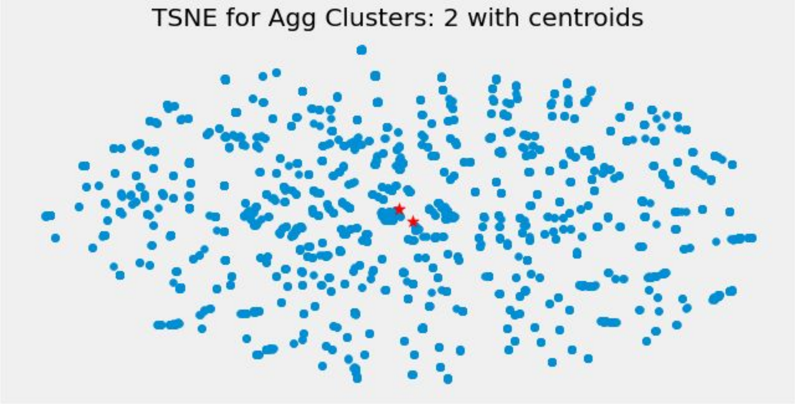
90% random sample   n=4277


UMAP results for Agg Clusters: 2

Dimensionality Reduction

TSNE & UMAP

- No defined clusters with either
- TSNE Silhouette Score: 32%
- UMAP Silhouette Score: 31%


TSNE for Agg Clusters: 2 with centroids

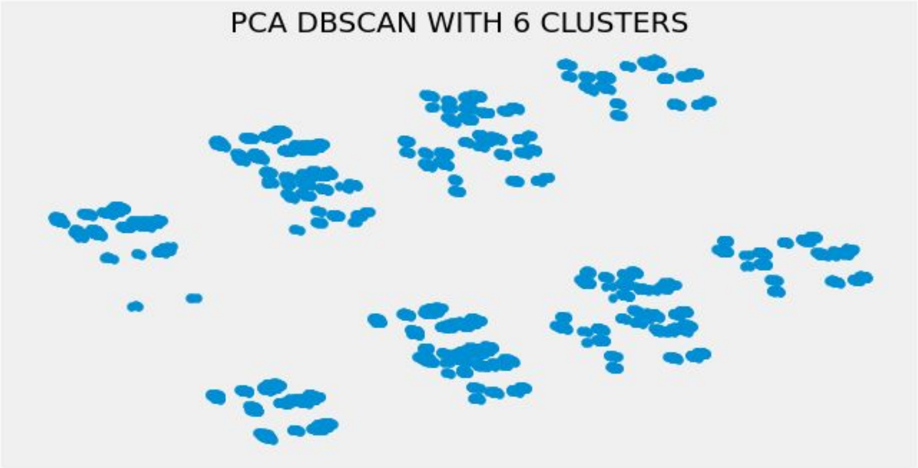
DBSCAN Clustering

90% random sample n=4277

6 Coded Clusters

8 clusters returned

- 90% random sampling n=4277
- PCA dimensionality reduction
- PCA Silhouette Score: 50%



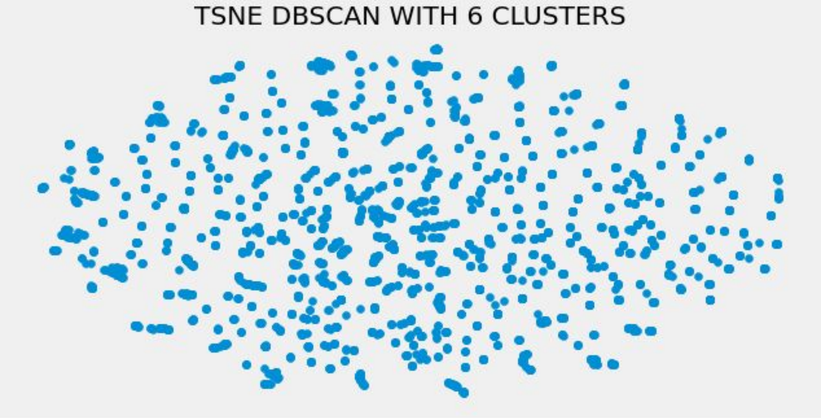
DBSCAN TSNE / UMAP

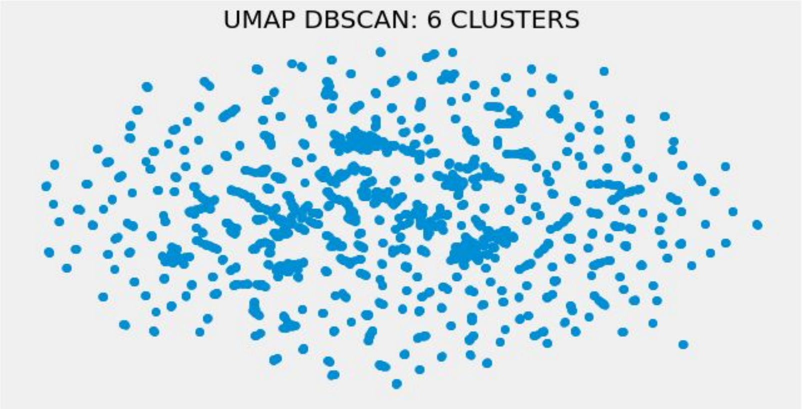
90% random sample   n=4277

Dimensionality Reduction

TSNE & UMAP

- No defined clusters with either
- TSNE Silhouette Score: 40%
- UMAP Silhouette Score: -12%



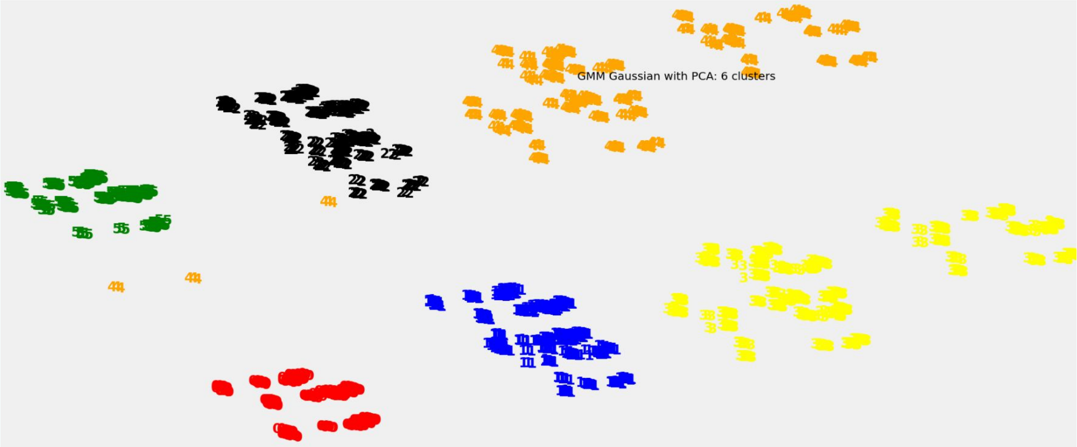


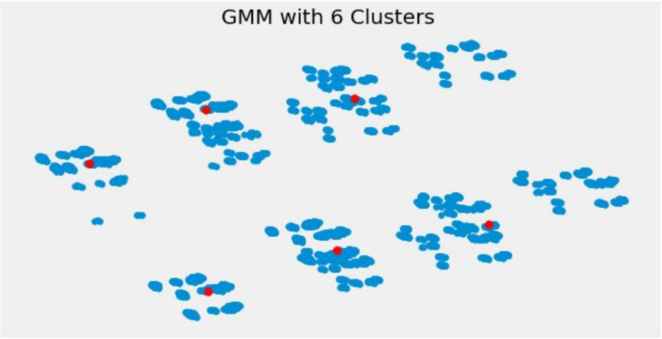
GMM Clustering

90% random sample   n=4277

6 Defined Clusters

- 90% random sampling   n=4277
- PCA dimensionality reduction
- PCA Silhouette Score: 62%





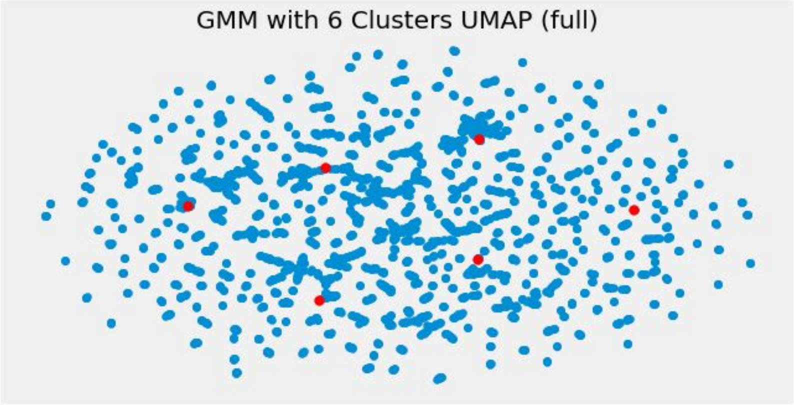
GMM TSNE / UMAP

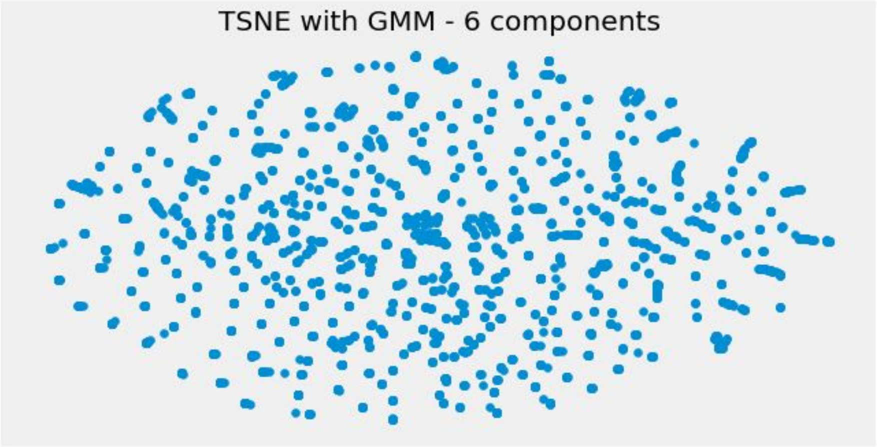
90% random sample   n=4277

Dimensionality Reduction

TSNE & UMAP

- No defined clusters with either
- TSNE Silhouette Score: 36%
- UMAP Silhouette Score: 32%

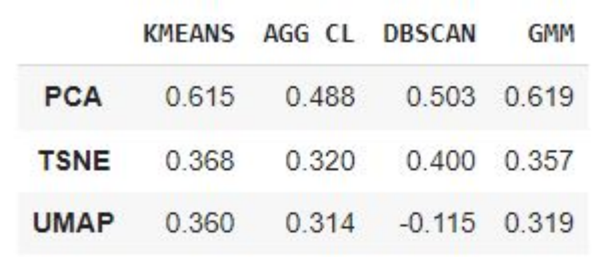

GMM with 6 Clusters UMAP (full)


TSNE with GMM - 6 components

SILHOUETTE SCORES

90% random sample   n=4277

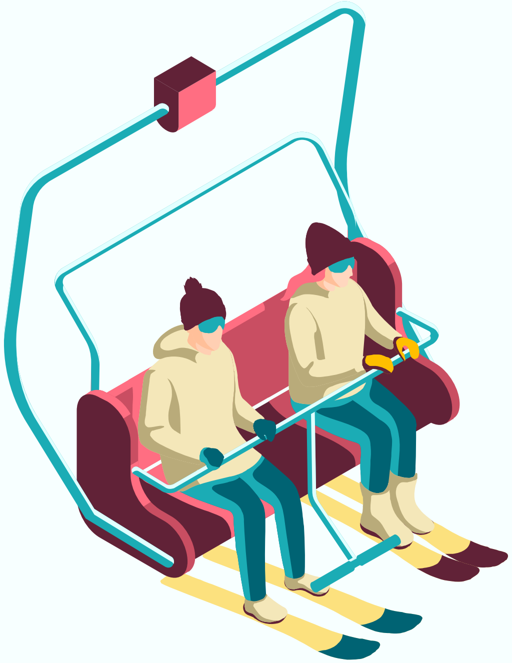
- Measures goodness of fit of clusters (cohesion within cluster and separation from other clusters)
- PCA clustering produced most distinct clusters and as coded (6 for most models)
- KMeans PCA and GMM PCA both produced Silhouette scores of 62%*
- ARI scoring was not considered since there was no defined y target to compare with



	KMEANS	AGG CL	DBSCAN	GMM
PCA	0.615	0.488	0.503	0.619
TSNE	0.368	0.320	0.400	0.357
UMAP	0.360	0.314	-0.115	0.319

*** I chose to go with cluster identification with KMeans due to established labels and better visual definition and centroid clustering.**

DATA-DRIVEN ACTION



MARKETING APPLICATION

MARKET SEGMENTATION

Clusters provide market segmentation for custom marketing and advertising campaigns

- family package deals
- rental deals
- email campaigns
- coupon/discount codes
- social media communications and campaigns
- ski/ride programs catered to age and interest levels

BRAND MANAGEMENT

Clusters allow for branding considerations around targeted ages and demographics (ie “the family mountain” or “where Gen Z goes to ski”)

OPERATIONS APPLICATION

MOUNTAIN OPERATIONS MANAGEMENT

Clusters provide insight into customer behavior, preferences, and potential. Operations can use clusters for:

- weather/crowd accommodations
- staffing and scheduling
- facility accommodations (ie family spaces, nursing rooms, janitorial management)
- lift load management and maintenance
- shuttle/transportation considerations

C-LEVEL APPLICATION

MARKET SEGMENTATION

Cluster segment awareness can contribute to optimal:

- staffing and training
- facility and resource management (ie hospitality, mechanical, terrain)
- marketing towards audience/customers
- HR staffing to accommodate employees, expected crowds, and demographics
- business development around defined cluster segments
- technology to meet expectations and demands of segments and business operations
- Risk Management/Legal management for crowds, customers, employees and overall safety/legal operations, practices, and policies

NEXT STEPS

EXPANSION

Create other sub datasets with additional features available (ie lift ticket costs, equipment rental, etc.)

CONDENSE

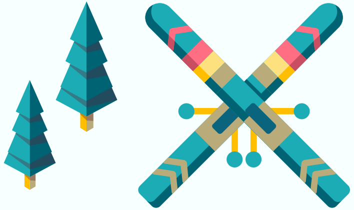
Create smaller, more defined clusters

TARGET

Model supervised learning on targets defined in clustering

MARKET/MANAGE

Use data modeling results to implement marketing and management changes



WEAKNESS CONSIDERATIONS

- The TSNE and UMAP models produced very scattered results, even after parameter tuning.

- Originally, I tried modeling on just weather conditions. These results were not conducive to clustering and only delivered about 15 actual data points due to duplicative data.

- The final cluster models delivered results with “all male” and “all female”. Some model runs did have clusters with both male/female and I would have preferred to have more gender diversity in my clusters.

FINAL CREDITS

CONTACT INFO

Lisa Girard
nonansensedata@gmail.com
lisaraygirard@gmail.com
lisagirard.com

Data processing with Jupyter Colab Notebook in Python
(click link to review)

Original Data Sources
https://data.mendeley.com/datasets/6w4tzrs3yw
https://www.tandfonline.com/doi/full/10.1080/23311886.2019.1681246

Images: from author and
https://www.flickr.com/photos/21085902@N08/16784653735

Presentation template by Slidesgo, icons by Flaticon, infographics and some images by Freepik.