BEHIND THE GOGGLES

unmasking ski resort customers

Using data to determine clustered
audience identification and segmentation

Lisa Girard

Capstone IV - Final
Thinkful Data Science
February 2, 2023

SKI SCHOOL

DATA SOURCE

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

RESEARCH PURPOSE

- An analysis of data pertaining to customer habits, intentions, behaviors, demographics, and characteristics of visitors to a Norwegian ski resort

- Analysis includes unsupervised clustered machine learning followed by supervised machine learning to predict and identify the resulting clusters
    - unsupervised data analysis
    - supervised data analysis

PROBLEMS TO SOLVE:

Customer identification is crucial to successful business operations, especially in a customer-focused, experience-based service industry like snowsports.

SKI SCHOOL

Applicable problems:

- **low utilization rates of ski school program offerings**
- **low facility utilization on bad weather days**
- **low weekday visitor counts**
- **long lift line wait times / less customer participation**
- **high male / low female ratio**
- **high singles / low family ratio**

Suggested Audiences

Marketing

- **Market Segmentation**
- **Brand Management**
- **Advertising/Marketing**
- **Business/Customer Development**
- **Media Relations**
- **Internal/External Communications**

Cluster identification is especially beneficial for:

- targeted marketing campaigns
- new customer segmentation
- re-engagement campaigns

Suggested Audiences

Operations

- **Lifts**
- **Food/Hospitality**
- **Rentals/Equipment**
- **Maintenance**
- **Snowmaking**
- **Facility and terrain management**
- **Transportation**

Cluster identification is especially beneficial for:

- facility management and utilization
- asset and resource management
- creating and managing on-mountain experience expectations

Suggested Audiences

C-Suite Executives

- **CEO/Board of Directors**
- **Finance**
- **Human Resources**
- **Marketing/Communications**
- **Operations**
- **Information Technology**
- **Risk Management**
- **Legal**

Cluster identification is especially beneficial for:

- audience/customer insight
- future business planning
- market and industry positioning

DOWNHILL DATA

THE PROCESS

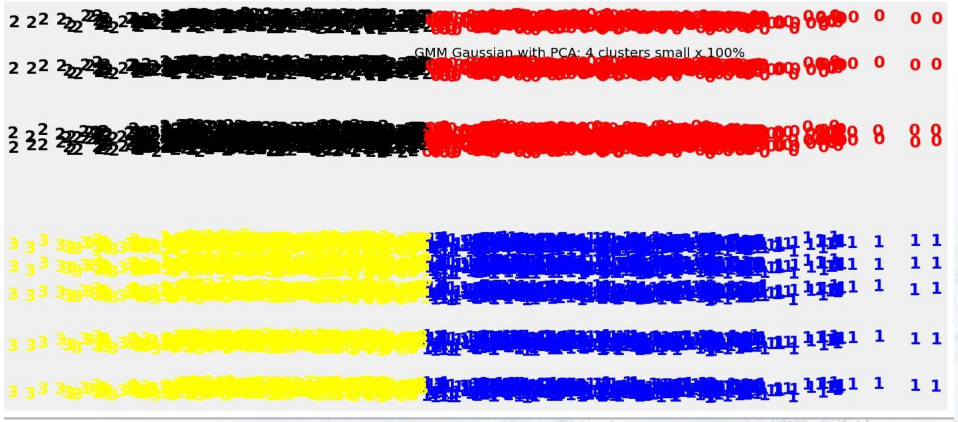
1. EDA / data cleaning of original data source
2. Feature engineering (subset of data with encoded values)
3. Applied 4 unsupervised clustering models with 3 dimensionality reducing techniques (PCA, TSNE, UMAP)
   a. KMeans
   b. Agglomerative Clustering *
   c. DBSCAN *
   d. GMM
4. Calculated silhouette scores per model and per model dimensionality reduction for model fit determination
5. Trained supervised learning models on cluster targets
6. Executed visual graphing plots to determine characteristics, patterns, demographics, and other identifiable features of the cluster observations (ski resort visitors)
7. Determination of business-case use

*Agglomerative clustering, DBSCAN models, and various dimensionality reducers not included in linked documentation as they were not productive. I settled on working with KMeans and GMM

CREATING CLUSTERS

Kmeans clustering

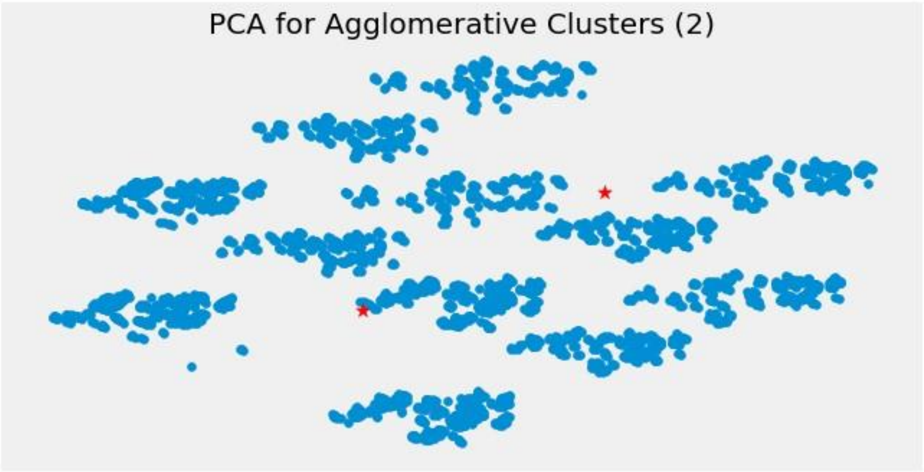
- identified 4 distinct clusters
- centered on defined centroids
- silhouette score: 47.26



GMM clustering

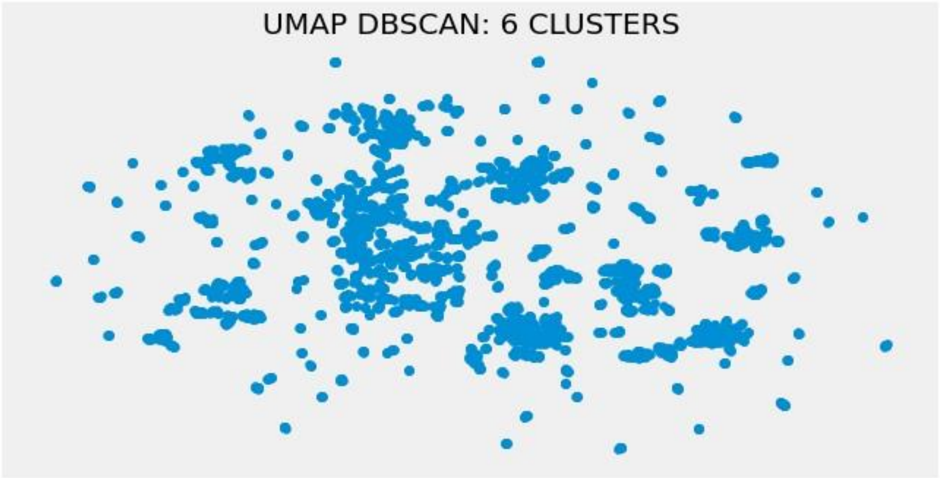
- identified 4 distinct clusters via text labels
- silhouette score: 47.25

OTHER CLUSTER MODELS



Agglomerative clustering w/PCA

- **did not identify 2 distinct clusters as coded**
- **silhouette score: 35.65**



GMM clustering w/UMAP

- **did not identify 6 distinct clusters as coded**
- **silhouette score: -3.48**

Who Are These Clusters?

Predicting Features

STANDARD LOGISTIC REGRESSION model

```
[ ] trainscore= lr.score(X_train,y_train)
    trainscore

0.9995648389904265

[ ] testscore = lr.score(X_test, y_test)
    testscore

0.9973913043478261

[ ] report = classification_report(y_test, y_preds)
    print(report)

              precision    recall  f1-score   support

           0       0.99      1.00      1.00       350
           1       1.00      1.00      1.00       253
           2       1.00      1.00      1.00       266
           3       1.00      0.99      1.00       281

    accuracy                           1.00      1150
   macro avg       1.00      1.00      1.00      1150
weighted avg       1.00      1.00      1.00      1150
```

- **Not a good model fit**
- **Too accurate**
- **Possible data leakage**

Who Are These Clusters?

Predicting Features

RANDOM FOREST model

```
[ ] acc_rf_score = classifier_rf.score(X_test, y_test)
    acc_rf_score

0.9191304347826087
```

```
[ ] # CLASSIFICATION FOR RF
    print(classification_report(y_test, preds_rf))
```

	precision	recall	f1-score	support
0	0.88	0.99	0.93	350
1	0.97	0.87	0.92	253
2	0.88	0.98	0.93	266
3	0.98	0.83	0.90	281
accuracy			0.92	1150
macro avg	0.93	0.91	0.92	1150
weighted avg	0.93	0.92	0.92	1150

```
kf=KFold(n_splits=5)
xv_rf = cross_val_score(classifier_rf,X,y,cv=kf) #
print("Cross Validation Scores are {}".format(xv_
print("Average Cross Validation score: {}".format(xv_rf.mean()))
xv_rf_score = xv_rf.mean()

Cross Validation Scores are [0.92434783 0.92254134 0.9329852  0.91470844 0.91906005]
Average Cross Validation score: 0.9227285730502895
```

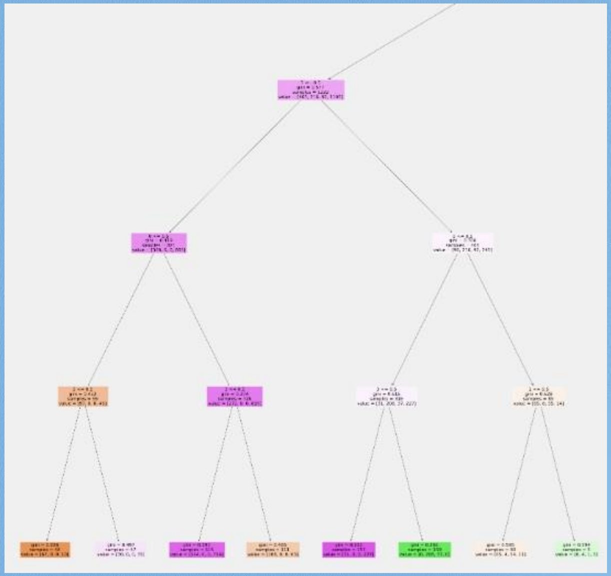
```
classifier_rf.oob_score_

0.918189730200174
```

- Great model fit
- Scores in high 80s, 90s
- OOB score: 92%
- Cross Validation score: 92%

Who Are These Clusters?

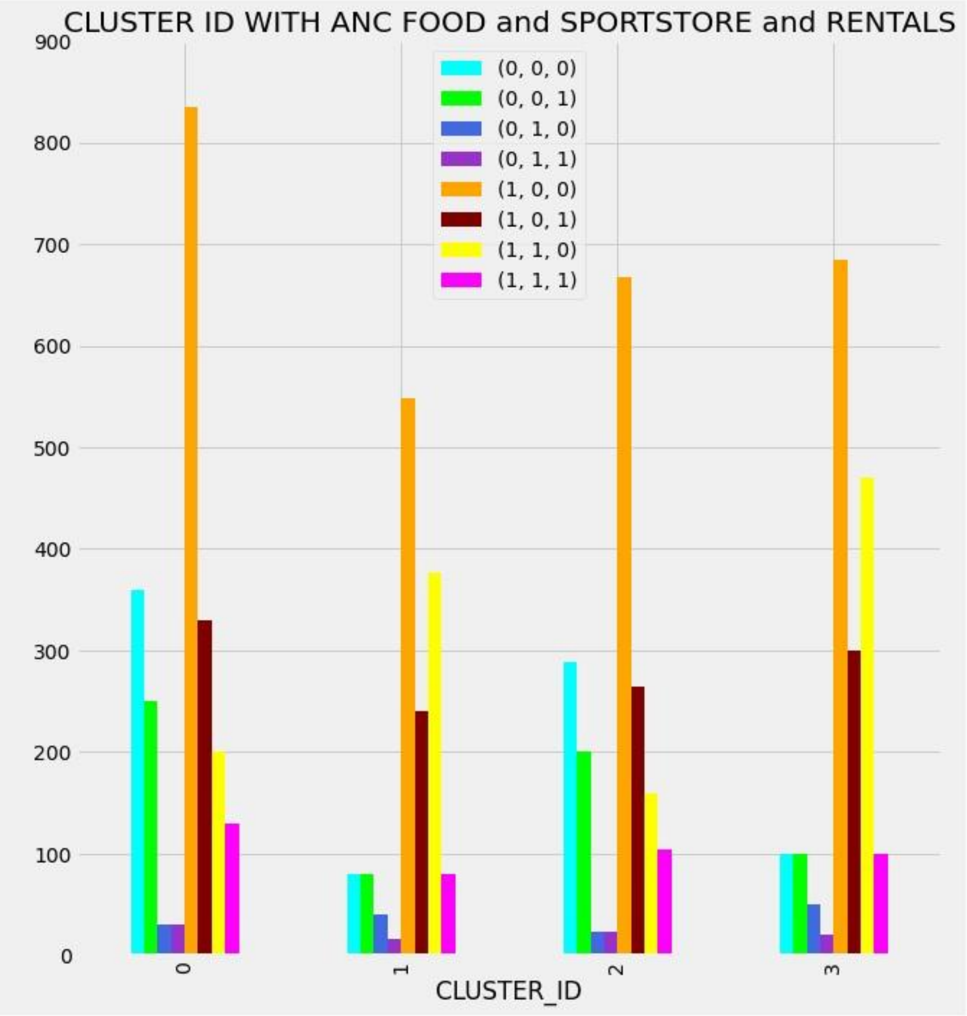
Predicting Features

RANDOM FOREST model



- RF handles complex variable relationships (ie possible data leakage)
- Includes *Feature Importance* measurement
- Aggregates results of several decision trees
- Handles both categorical and continuous data
- Good for high dimensional data (ie 87 columns)

Cluster Snapshots



(1,0,0) - yes for food, no for store, no for rentals
(1,0,1) - yes for food, no for store, yes for rentals
(1,1,0) - yes for food, yes for store, no for rentals

Ancillary spending on:

- restaurants/food
- sport store
- rentals

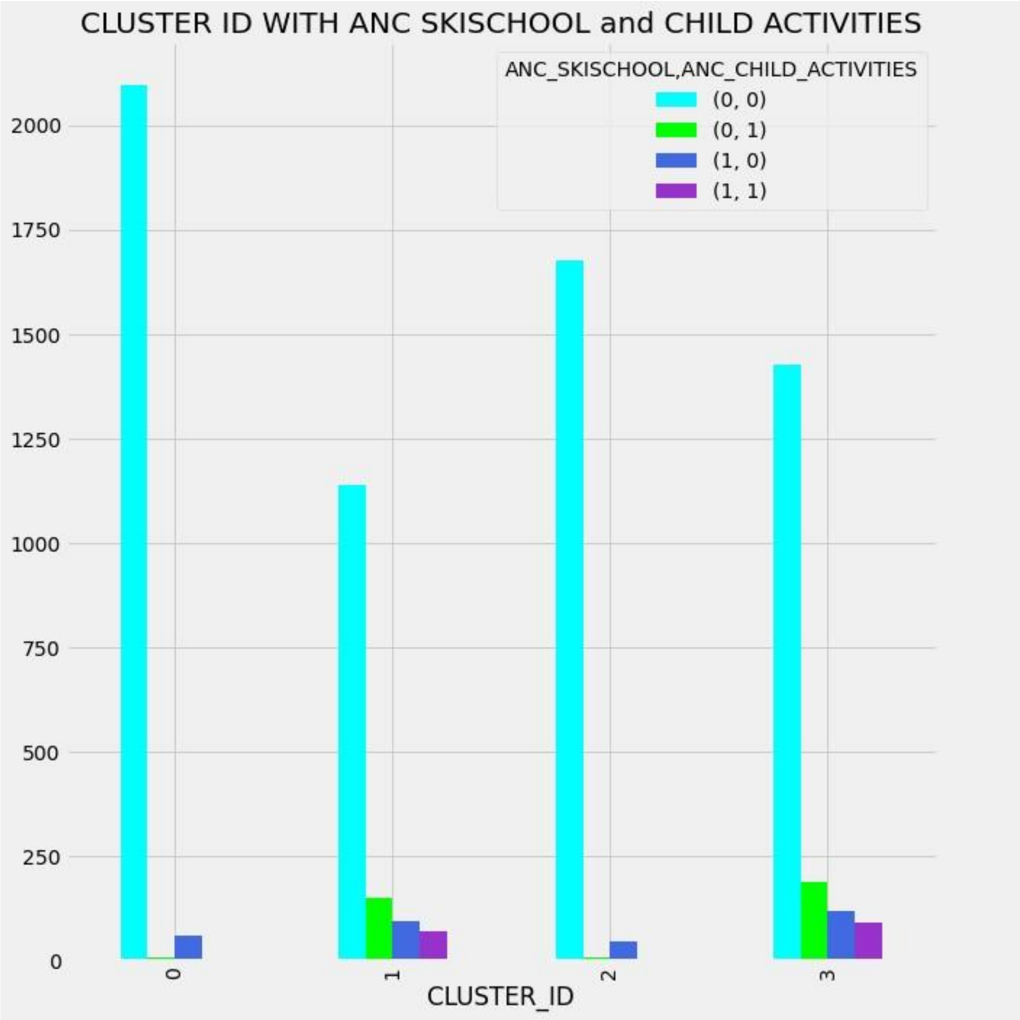
Interpretations:

- all clusters spend more money on food than other ancillary services
- food and sport store and food and rentals are 2nd most popular spending combinations

Business Case Use:

- special offers/coupons for food with rental/sport store purchase
- ensure excellent customer service in restaurant sector
- maintain food service quality/ quantity customer expectations

Cluster Snapshots



(0,0) - no for ski school, no for child activities
(0,1) - no for ski school, yes for child activities
(1,0) - yes for ski school, no for child activities
(1,1) - yes for ski school, yes for child activities

Ancillary spending on:

- ski school
- child activities

Interpretations:

- all clusters spend very little on ski school
- clusters 1 and 3 spend more money on child activities than ski school

Business Case Use:

- marketing campaigns geared toward increasing awareness of ski school program offerings
- special offer for ski school discount/trial with purchase of child activity
- market/customer research about ski school opinion/experience
- robust evaluation/development of ski school program

Cluster Snapshots

(0,0,0,0,1) - yes for price NOK650, no for all other prices
(0,0,0,1,0) - yes for price NOK550, no for all other price
(0,0,1,0,0) - yes for price NOK450, no for all other price
(0,1,0,0,0) - yes for price NOK350, no for all other price
(1,0,0,0,1) - yes for price NOK250, no for all other price

Lift Ticket Price

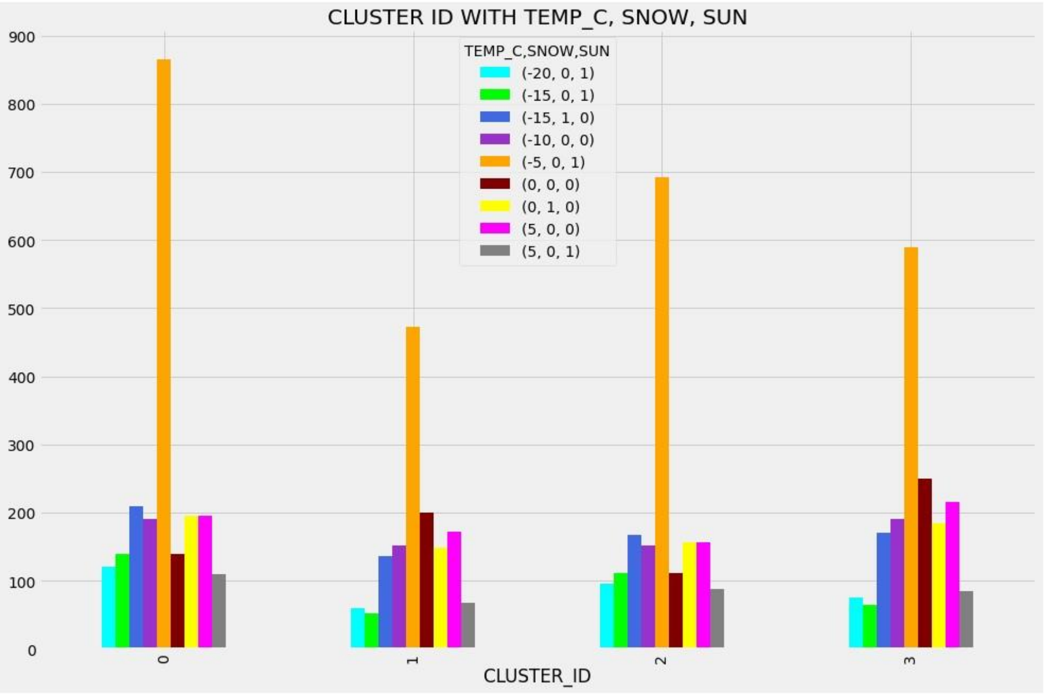
Interpretations:

- **lift ticket sweet spot prices:**
  - **NOK 450-550 (USD $45-55)**
  - **NOK 250 (USD $25)**
- **Highest priced ticket – less visitors**

Business Case Use:

- **consider dynamic lift ticket pricing (demand-driven)**
- **consider weather-related pricing**
- **consider multi-pack pricing (ie "Loveland 4pack")**
- **review relationship between day part/pricing(full-day/half-day), weather, visitor demographics (distance, family status,etc.)**

Cluster Snapshots



(-5,0,1): -5C°, no snow, sunny day
(0,0,0): 0C° , no snow, not sunny day
(5,0,0): 5C°, no snow, not sunny day
(-15,1,0): -15C°,snowy, not sunny day

Weather Conditions

Interpretations:

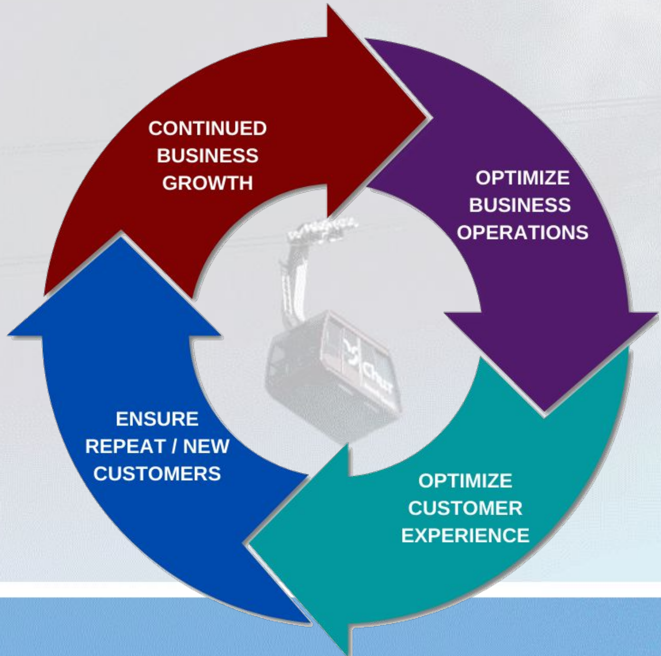
- all clusters with visitors on days with
  - temps at ~ -5C° (23F°)
  - no snow
  - sunny
- Other popular weather conditions:
  - cooler temps (~ 5-32F°), snowy/not-snowy, no sun
  - warmer temps (~ 41F°), no snow, no sun

Business Case Use:

- consider weather-related lift ticket pricing to increase business/utilization on non-optimal weather days
- facility management on good weather days (lifts, operations, transportation, staffing)
- ancillary services support (ie food, rentals) for good weather days

CONCLUSION

- Successfully clustered ~ 7200 observations into 4 customer groups
- Identified characteristics and features inside each of 4 customer groups
- Identified business case uses for some unpacked clusters regarding:
  - ancillary spending
  - lift ticket pricing
  - weather

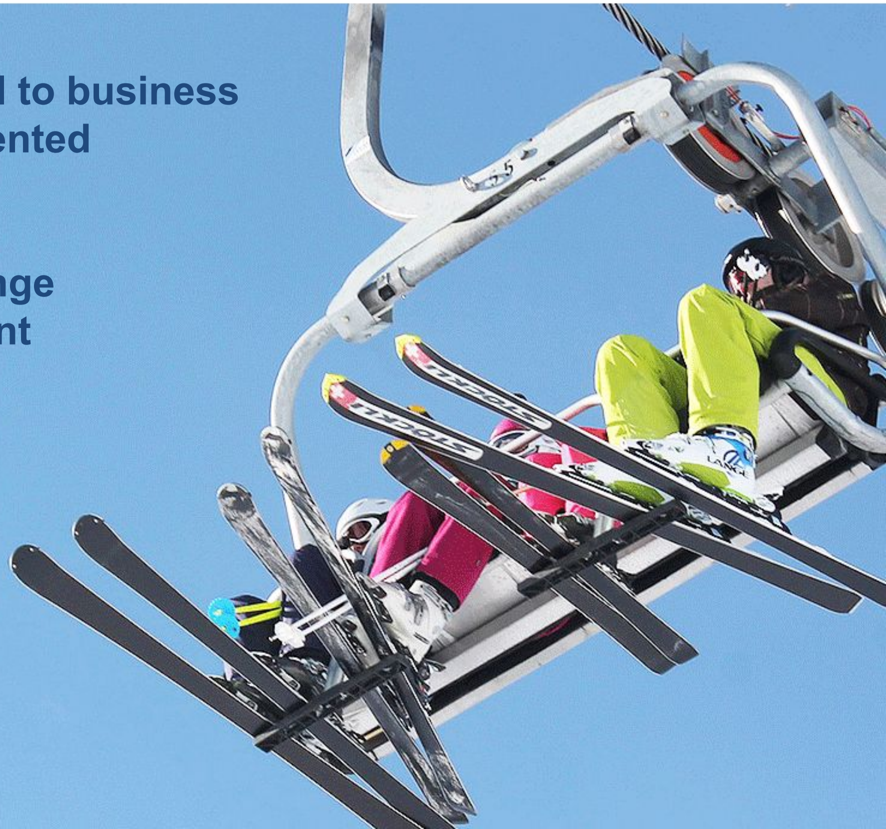


WHY DATA MATTERS

- Identifying customer characteristics is crucial to business success, especially in service/experience oriented industries

- Snowsports industry unique business challenge
  - controllable typical business environment
  - uncontrollable physical environment

- Extra layers of challenge
  - geo-physical location
  - required equipment
  - skills learning curve
  - inherent risk/danger
  - cost-prohibitive expense



RECOMMENDATIONS

- Continue unpacking clusters and developing business case uses from findings

IF

Cluster0 contains more “Couple with Child”

THEN

market ski school/child activities to Cluster0

IF

Cluster1 is “Work Full Time”, ski on good weather days

THEN

provide opportunities that cater towards their weekend availability and weather preference, and, provide mountain services to address volume/needs

RECOMMENDATIONS

- **Continue unpacking clusters and developing business case uses from findings**

IF

Cluster2 skews female with high interest

THEN

provide ladies-only programs, clinics, and other offerings/events that cater towards high interest, female.

IF

Cluster3 is “Price NOK250”

THEN

consider these “bargain skiers” a group that prefers lower costs and be sure to offer other services at lower price points

RECOMMENDATIONS

- **Cross-reference clusters for more analysis**
    - **“check Cluster1 against Cluster3” for more refinement, pattern recognition**

- **Continual cluster modeling to determine other patterns, refine to small/bigger clusters**

- **Share cluster finds and data analysis with entire team to:**
    - **build data literacy**
    - **ensure team engagement**
    - **solidify data strategies that will aid in data informed decisions**

FINAL CREDITS

CONTACT INFO

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

Thinkful Data Science
online bootcamp
Capstone IV - Final
February 2, 2023

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