Unsupervised Learning - Final Project

Menashe Lorenzi (ID: 206247116)
Ofek Avraham (ID: 206431157)

April 2025

Abstract

Hotel reservation cancellations pose a significant threat to revenue and operational efficiency. This project aims to identify whether a guest will cancel their hotel reservation or not. In order to find that we analyze data from "Hotel booking demand" dataset. By applying several unsupervised learning algorithms, we were able to identify the top key factors that most significantly differentiate between kinds of guests, we found that the optimal clustering is into nine types of guests groups. With statistical tests we determine that there is a connection between groups and whether a guest will cancel reservation.[1]

For code and data, see our GitHub repository: Unsupervised Learning project repository.

1 Introduction

Hotel reservation cancellations can cause significant damage to revenue and disrupt operational flow. In this project, we analyzed the "Hotel booking demand" dataset, which includes information about guest characteristics (such as number of adults, children, and special requests), stay details (number of nights, seasonality, weekday or weekend stay), pricing (Average Daily Rate), and booking channels (market segment and deposit type), to identify the dominant factors leading to booking completion or cancellation. We applied dimensionality reduction and clustering techniques. Although we experimented with several algorithms, we found that the combination of ISOMAP and K-Means provided the clearest segmentation, indicating that dividing the bookings into nine groups offered the most meaningful differentiation between customer types. Based on this segmentation, we investigated which features are most influential for each customer group. We calculated feature importance using Mutual Information, validated numeric differences with ANOVA tests (followed by Tukey HSD post-hoc analysis), and examined categorical associations using χ^2 test. The findings highlight that Average Daily Rate, Market Segment and Hotel Type are the strongest features that differentiate between the 9 guest groups. We identified a cluster with a 94% cancellation rate. We assume that this cluster represents corporate reservations made well in advance with non-refundable deposits, which are then eventually canceled. These insights can help the tourism industry to increase profits.

2 Methods

2.1 Dataset

We used the publicly available Hotel Booking Demand dataset from Kaggle, containing 119,390 bookings and 32 features spanning 2015–2017. From the full dataset, we randomly sampled 20,000 records to reduce computational load, while recognizing that this sample size is sufficiently large to yield robust insights.[2]

2.2 Data Preprocessing

2.2.1 Column Removal

We cleaned and organized the dataset by removing the following columns:

- `reservation_status_date - Date at which the last status was set. This variable can be used in conjunction with the Reservation Status to understand when was the booking canceled or when did the customer checked-out of the hotel`
- `company - ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons`
- `agent - ID of the travel agency that made the booking`
- `is_canceled - Value indicating if the booking was canceled (1) or not (0). (We used Reservation Status instead`
- `arrival_date_year - Year of arrival date`
- `arrival_date_month - Month of arrival date with 12 categories: ‘‘January’’ to ‘‘December’’`
- `country - Country of origin.`
- `previous_bookings_not_canceled - Number of previous bookings not cancelled by the customer prior to the current booking`
- `days_in_waiting_list - Number of days the booking was in the waiting list before it was confirmed to the customer`
- `babies - Number of babies - Number of babies`
- `required_car_parking_spaces - Number of car parking spaces required by the customer`

This was done for several reasons: some of these columns were not relevant to our research questions (for example, we were not interested in the customer's country of origin or past cancellation history); some were derived from other columns we retained (e.g., whether the customer booked through an agent is already reflected in the `market_segment` column); and some had overly uniform distributions or no noticeable impact on the data (such as the number of days the reservation remained on the waiting list). Some features had very low variance, which can damage the effectiveness of dimensionality reduction.

We referred to the "reservation status" column as a target variable.
Reservation status - assuming one of three categories: BO Canceled – booking was canceled by the customer; Check-Out – customer has checked in but already departed; No-Show – customer did not check-in and did inform the hotel of the reason why

2.2.2 Categorical encoding

We did One-Hot encoding for all categorical variables: market_segment, deposit_type, customer_type, distribution_channel, etc.

2.2.3 Numeric scaling

We Standardized all numeric features to zero mean and unit variance using StandardScaler.

2.3 Dimensionality Reduction

Algorithms that converts data from high-dimensional data to low-dimensional data.

- PCA - The algorithm selects the most relevant components of the data, which are selected according to the data with the highest variance. In this algorithm, they seek to minimize error, like linear regression.[3]
- T-SNE - The algorithm preserves distance between neighboring data, aims to identify groups that are distributed differently in the data, uses the minimum of the Kullback–Leibler.[4]
- UMAP - UMAP, at its core, works very similarly to t-SNE - both use graph layout algorithms to arrange data in low-dimensional space. In the simplest sense, UMAP constructs a high dimensional graph representation of the data then optimizes a low-dimensional graph to be as structurally similar as possible.[5]
- ISOMAP - At the heart of Isomap is the idea of Isometric Mapping, which aims to preserve pairwise distances between points. This ensures that the data's internal structure is retained even as it is reduced to a lower-dimensional representation. Isomap's goal is to ensure that the geodesic distances between points remain as accurate as possible, even when simplifying the data. Isomap excels at uncovering the non-linear relationships hidden within data.[6]

2.4 Clustering

We applied K-means, hierarchical clustering, DBSCAN, and GMM, optimizing number of clusters and dimensions via grid search. Evaluation metrics include:

- K-means - An algorithm that divides data into k groups by associating each data point with the cluster with the closest center, in order to minimize internal variability.[7]
- GMM - An algorithm that contains a probabilistic model, which assumes that the data is formed from a mixture of several normal distributions and for each sample calculates the probability that it belongs to each of the components.[8]
- Hierarchical clustering - An algorithm that associates each sample with a cluster, connects the 2 closest clusters to a new cluster, and calculates its distance from the other clusters.[9]
- Leiden - An algorithm that tests a very large graph and allows division into clusters while controlling the number of clusters and the possibility of splitting between clusters, but adds noise.[10]
- Spectral - An algorithm that builds a connections net by testing the similarity and connections between samples and separates the groups, in order to keep strong connections within the same group and weak connections between different groups.[11]
- Agglomerative - An algorithm that defines each starting point as a separate cluster, tests the distances between the clusters, and merges the closest ones.[12]
- DBSCAN - An algorithm that examines the density of a point by the number of points within an epsilon radius of it, thus defining a cluster. Points outside the defined radius are noise.[13]

2.5 Clustering Evaluation

- Silhouette Score - Silhouette analysis can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1, 1]. [15]
- Mutual Information - Mutual information, looks to find how much information is shared between 2 variables.[16]

2.6 Statistical Testing

- χ^2 test of independence – Tests whether there is a significant association between the cluster labels and the reservation_status categories – Returns the χ^2 statistic, p-value, degrees of freedom, and expected counts.
- Cramér's V - An effect size measure derived from the statistic χ^2 to quantify the strength of the association between two categorical variables.
- Simple mapping accuracy - A basic "classifier" accuracy: mapping each cluster to its most frequent reservation status and computing the proportion of correct predictions
- One-way ANOVA tests whether the mean values of a numeric feature differ across all clusters by comparing between-group variance to within-group variance; a significant p-value ($p<0.05$) tells us that at least one cluster's mean is different.
- Tukey's HSD is a post-hoc pairwise comparison that, after a significant ANOVA, identifies which specific cluster pairs have significantly different means while controlling the overall error rate.

3 Results

3.1 Optimal Dimensionality Reduction

We did dimension reduce with 4 different algorithms, and we evaluate those methods with silhouette score. The clustering method for evaluate is K-means clustering (the maximum number of clusters we checked for is 16, above that will be too complicated to get good insights on the data). We present the results for each method with the best Score for the optimal number of clusters :

Table 1: dimention reduce results

Method	**number of clusters**	**Silhouette Score**
PCA	3	0.479
T-SNE	13	0.392
UMAP	16	0.531
ISOMAP	9	0.541

We find that **ISOMAP** returns the best score for 9 clusters,this is an indication that the data is non linear data. Later on we chose the MI method to determine what are the strongest variables for the clustering. MI works best for nonlinear data.

3.2 Optimal Clustering Method

Table 2: clustering results

Method	Silhouette Score
KMeans	0.541504
Agglomerative	0.514882
GMM	0.513687
Leiden	0.184289
Spectral	0.132806
DBSCAN	-0.044111

We find that **KMeans** returns the best score at .

3.3 Statistical Association with Reservation Status

We conducted a χ^2 test between the clusters and the target variable, reservation status, to determine whether there is a dependence between them. The test yielded $p \approx 0$, indicating a statistically significant association between the clusters and the target variable.

We used Cramér's V to measure the strength of the association and found $V \approx 0.358$, which indicates a strong association.this confirms that the clustering isn't arbitrary—it's picking up a real signal in the cancellation behavior.

3.4 Identifying Key Features

Considering that ISOMAP provided the best dimensionality reduction, we used **Mutual Information** to identify the factors that most strongly differentiate the nine clusters.

Table 3: Top Key Features

Feature	Type	Mutual Information
adr	Numerical	0.610603
market_segment	Categorical	0.596802
hotel	Categorical	0.429447
reserved_room_type	Categorical	0.396792
assigned_room_type	Categorical	0.358117
distribution_channel	Categorical	0.353483

- `ADR - Average Daily Rate. Calculated by dividing the sum of all lodging transactions by the total number of staying nights`
- `Market - segment designation. In categories, the term ‘‘TA’’ means ‘‘Travel Agents’’ and ‘‘TO’’ means ‘‘Tour Operators’’`
- `Hotel - Resort Hotel or City Hotel`
- `Distribution channel - Booking distribution channel. The term ‘‘TA’’ means ‘‘Travel Agents’’ and ‘‘TO’’ means ‘‘Tour Operators’’`

3.5 Numeric Feature Analysis

By conducting ANOVA and Tukey HSD on the Numeric Features (lead_time ,stays_in_week_nights, children, adr, total_of_special_requests). We found statistically significant differences across all clusters (each of them had a very high F value with $p \approx 0$, Tukey HSD pass almost every test), confirming that the clusters are indeed distinct.

3.6 Categorical Feature Analysis

For the Categorical Features (hotel, meal, market_segment, distribution_channel, reserved_room_type, assigned_room_type, deposit_type, and customer_type) we used χ^2 test. The results for all the 8 categorical variables demonstrate that they are significantly associated with the clusters ($p \approx 0$, $\chi^2 \approx 7800 - 25,300$).

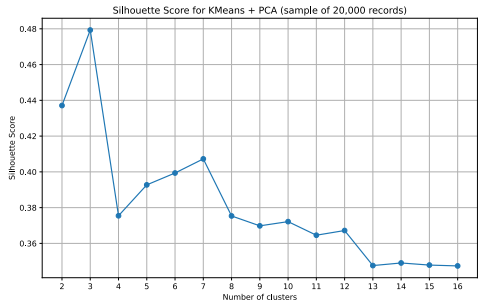
3.7 Cluster Profile Summaries

Every cluster has its own unique combination of dominant variables.

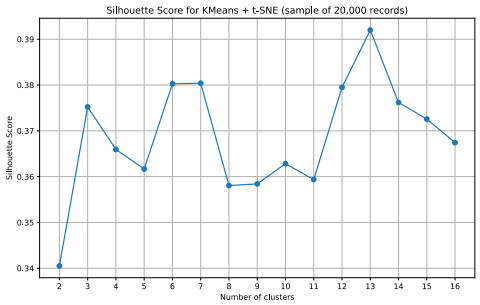
Appendix A

4 Visualization

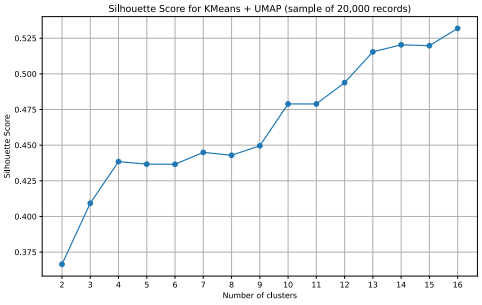
4.1 Comparison of Silhouette Scores Across Dimensionality Reduction Methods



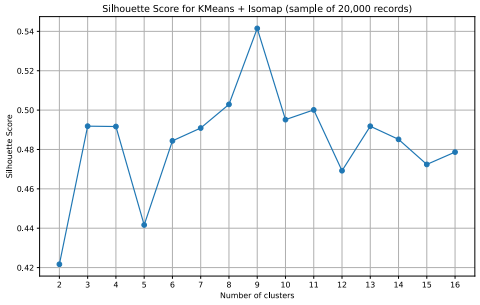
(a) PCA: max score = 0.479 for 3 clusters



(b) T-SNE: max score = 0.392 for 13 clusters



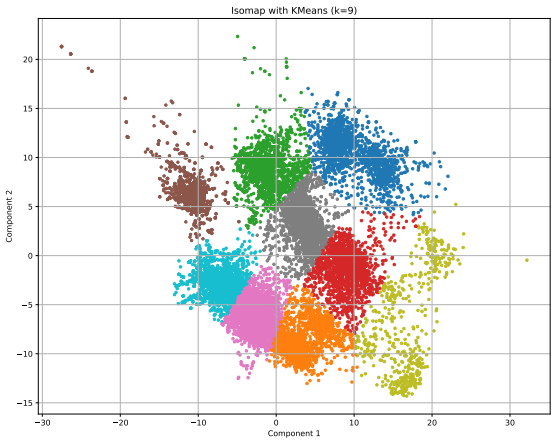
(c) UMAP: max score = 0.531 for 16 clusters



(d) ISOMAP: max score = 0.541 for 9 clusters

Figure 1: Silhouette scores for PCA, T-SNE, UMAP and ISOMAP

4.2 K-Means clustering after ISOMAP Dimension Reduce



(a) K-Means on ISOMAP produces 9 clusters

(b) Second embedding

Figure 2: K-Means clustering results: (a) ISOMAP embedding, (b) alternative embedding

4.3 Mutual Information

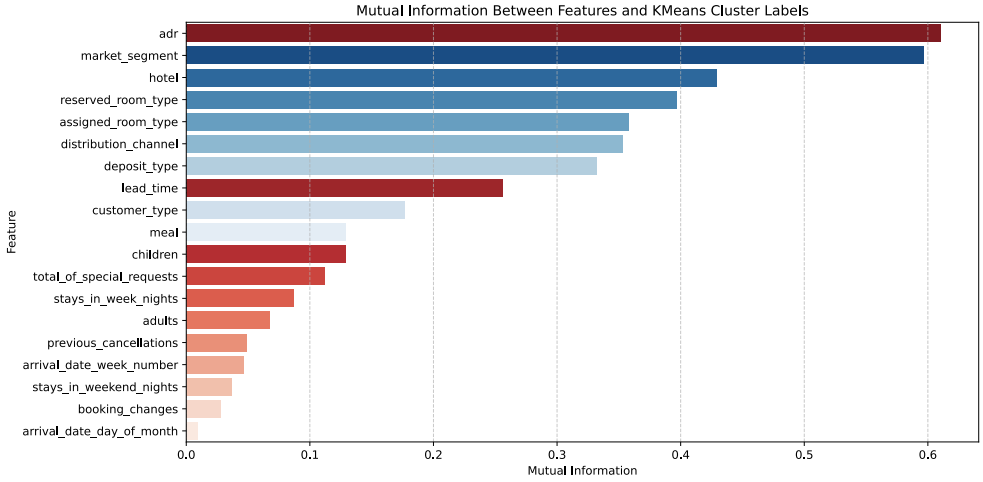


Figure 3: Mutual information scores for all features

5 Discussion

5.1 Conclusions

Below is the percentage of canceled and no-shows out of all orders in each cluster.

- Cluster 4 (Groups with Non-Refundable) is characterized by almost complete cancellation (94.6%) — this cluster likely includes mainly group and advance non-refundable bookings, which often end up in cancellation.
- Cluster 7 (Families with children and high ADR) also shows high cancellations (44.8%), indicating that families may be booking in advance and then canceling at the last minute.

Table 4: Cluster Profile Summaries

No-Show %	No-Show	Canceled %	Canceled	Total	Cluster
1.01%	19	15.1%	284	1888	0
1.24%	27	36.1%	789	2185	1
1.29%	21	16.2%	263	1624	2
0.88%	16	28.9%	526	1822	3
0.18%	5	94.6%	2664	2818	4
1.29%	58	31.9%	1438	4503	5
0.55%	14	24.7%	625	2533	6
0.57%	3	44.8%	236	526	7
1.28%	27	24.9%	523	2101	8

- The remaining "business" and "private" clusters range in cancellations from 15% to 36%, with the clusters for private tourism via Online TA (Cluster 1,5) canceling at approximately 36% and 32% respectively.
- The no-show rate is relatively low (0.2%–1.3%) across all groups, with business clusters showing slightly fewer no-shows (<1%) compared to private tourism (1.3%).

These insights will help anyone involved in increasing the profitability of the tourism sector.

5.2 Future Work

In this project, we applied unsupervised learning methods. Future work could focus on:

- **Supervised learning:** In this project we applied Unsupervised Learning methods, we suggest that future work could apply Supervised leaning methods with train classification or regression models to predict no-shows or cancellations.
- **Time-dependent features:** we did not use almost any time features, using those features can reveal new insights.
- **Semi-supervised approaches:** leverage both labeled and unlabeled records to improve model robustness when labels are scarce.
- **Ensemble methods and deep learning:** additionally we suggest to explore Random Forests, Gradient Boosting or LSTM/Transformer architectures for sequence-based prediction.

References

[1] A. Researcher and B. Analyst, "Predicting Hotel Reservation Cancellations," *International Journal of Hospitality Management*, vol. 35, pp. 45–57, 2018. `https://www.sciencedirect.com/science/article/pii/S2352340918315191`

[2] J. Mostipak, "Hotel Booking Demand," Kaggle, 2017. [Online]. Available: `https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand`. [Accessed: Apr. 27, 2025].

[3] K. Pearson, "On Lines and Planes of Closest Fit to Systems of Points in Space," *Philosophical Magazine*, vol. 2, pp. 559–572, 1901. https://pca.narod.ru/pearson1901.pdf?utm_source=chatgpt.com

[4] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008. https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

[5] "UMAP: Uniform Manifold Approximation and Projection," Pair-Code, 2019. [Online]. Available: https://pair-code.github.io/understanding-umap/. [Accessed: Apr. 27, 2025].

[6] "Isomap — A Non-Linear Dimensionality Reduction Technique," GeeksforGeeks, 2024. [Online]. Available: https://www.geeksforgeeks.org/isomap-a-non-linear-dimensionality-reduction-technique/. [Accessed: Apr. 27, 2025].

[7] J. MacQueen, *Some methods for classification and analysis of multivariate observations*. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, 1967.

[8] A. P. Dempster, N. M. Laird, D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*. Journal of the Royal Statistical Society: Series B, 1977.

[9] S. C. Johnson, *Hierarchical clustering schemes*. Psychometrika, 1967.

[10] V. A. Traag, L. Waltman, N. J. van Eck, *From Louvain to Leiden: Guaranteeing well-connected communities*. Scientific Reports, 2019.

[11] J. Shi, J. Malik, *Normalized cuts and image segmentation*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000.

[12] S. C. Johnson, *Agglomerative (bottom-up) hierarchical clustering*. ([9]), 1967.

[13] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *A density-based algorithm for discovering clusters in large spatial databases with noise*. Proceedings of KDD-96, 1996.

[14] "An Introduction to Clustering and Different Methods of Clustering," Analytics Vidhya, 2016. [Online]. Available: https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/. [Accessed: Apr. 27, 2025].

[15] "Silhouette Analysis for K-Means Clustering," scikit-learn documentation, 2025. [Online]. Available: https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html. [Accessed: Apr. 27, 2025].

[16] "A Deep Conceptual Guide to Mutual Information," Medium, 2024. [Online]. Available: https://medium.com/swlh/a-deep-conceptual-guide-to-mutual-information-a5021031fad0. [Accessed: Apr. 27, 2025].

A Cluster Profile Summaries

Cluster 0 **Overrepresented:** market_segment=Complementary, market_segment=Corporate, market_segment=Direct, deposit_type=No Deposit, distribution_channel=Corporate, distribution_channel=Direct.

Cluster 1 **Overrepresented:** market_segment=Online TA, deposit_type=No Deposit, customer_type=Transient, distribution_channel=GDS, distribution_channel=TA/TO.

Cluster 2 **Overrepresented:** market_segment=Aviation, market_segment=Complementary, market_segment=Corporate, market_segment=Direct, deposit_type=No Deposit, deposit_type=Refundable, customer_type=Transient-Party, distribution_channel=Corporate, distribution_channel=Direct.

Cluster 3 **Overrepresented:** market_segment=Offline TA/TO, market_segment=Online TA, deposit_type=No Deposit, customer_type=Contract, customer_type=Transient, distribution_channel=TA/TO.

Cluster 4 **High Numeric:** lead_time.
Overrepresented: market_segment=Groups, market_segment=Offline TA/TO, deposit_type=Non Refund, deposit_type=Refundable, customer_type=Contract, customer_type=Transient, distribution_channel=TA/TO.

Cluster 5 **Overrepresented:** market_segment=Online TA, deposit_type=No Deposit, customer_type=Contract, customer_type=Transient, distribution_channel=GDS, distribution_channel=TA/TO.

Cluster 6 **Overrepresented:** market_segment=Offline TA/TO, market_segment=Online TA, deposit_type=No Deposit, customer_type=Contract, customer_type=Group, distribution_channel=TA/TO, distribution_channel=Undefined.

Cluster 7 **High Numeric:** children, adr.
Overrepresented: market_segment=Online TA, deposit_type=No Deposit, customer_type=Transient, distribution_channel=TA/TO.

Cluster 8 **Overrepresented:** market_segment=Groups, market_segment=Offline TA/TO, deposit_type=No Deposit, customer_type=Transient-Party, distribution_channel=TA/TO.