ORIE 5250: Final Report

Finding Optimal Assortments of Expedia By Customer Clustering

Chenyang Yan cy477
Yutong Wang yw2364
Hongyi Nie hn327

Introduction

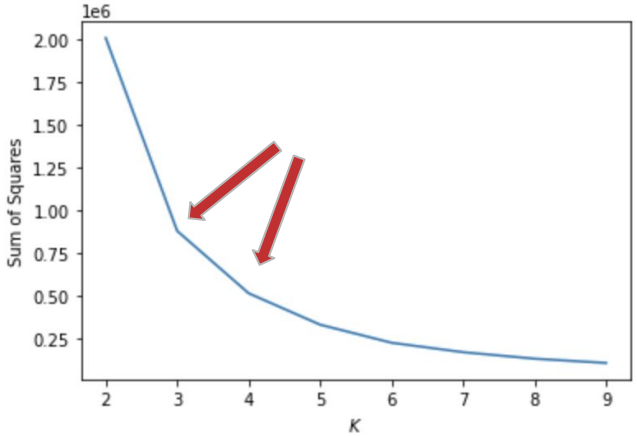
- For the second dataset “Expedia”, Divide the customer into more segments by using K means clustering and Spectral Clustering, to find the optimal assortment to maximize the expected revenue.
- Mixture of MNL models for multi clusters
- Single MNL model for unknown type of clusters
- MLE estimation
- Compare the mixture MNL and the single MNL model result

KMeans Clustering

We tried to better define customer types using the features:

'srch_booking_window', 'srch_adults_count',
'srch_children_count', 'srch_room_count'

To find the best **K** to predict type, we first evaluate the clusterings using **The Elbow Method**:

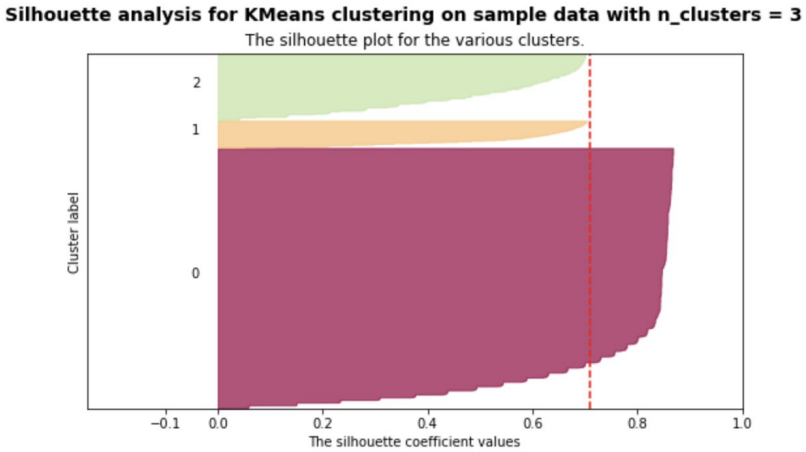


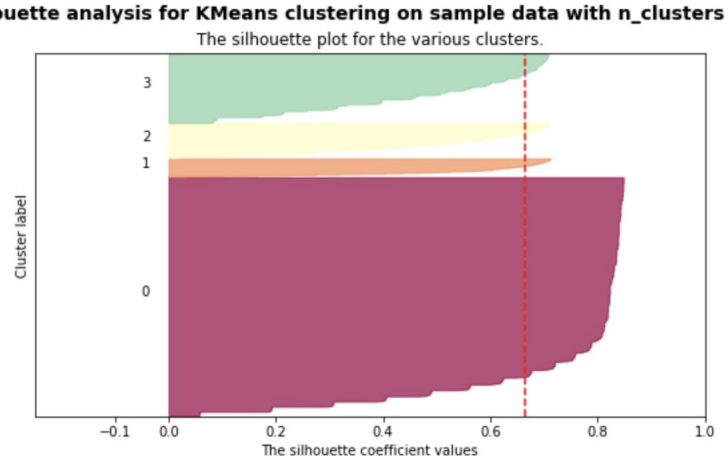
In this method, we look at the sum-of-squares error in each cluster against K. We compute the distance from each data point to the center of the cluster (centroid) to which the data point was assigned.

KMeans Clustering

To further determine the value for K, we then evaluate the clusterings using **The Silhouette Method**:

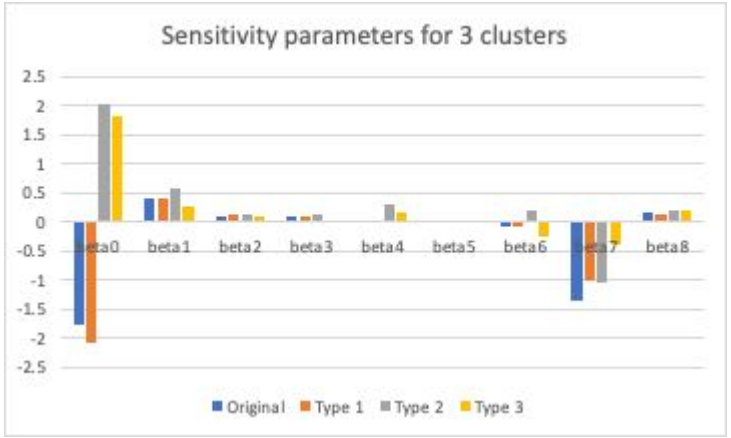
This method measures how well each datapoint "fits" its assigned cluster *and also* how poorly it fits into other clusters.



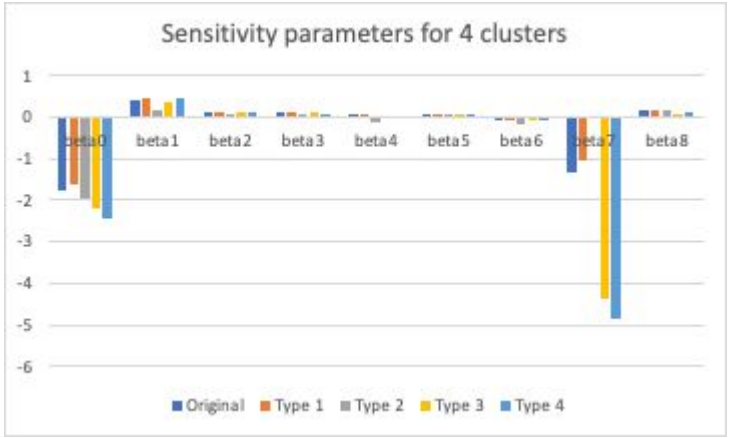


K means: N = 3 vs N = 4

Beta for N = 3



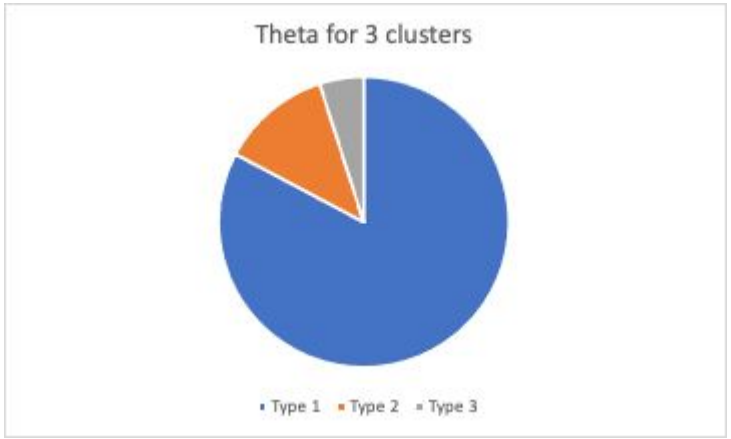
Beta for N = 4



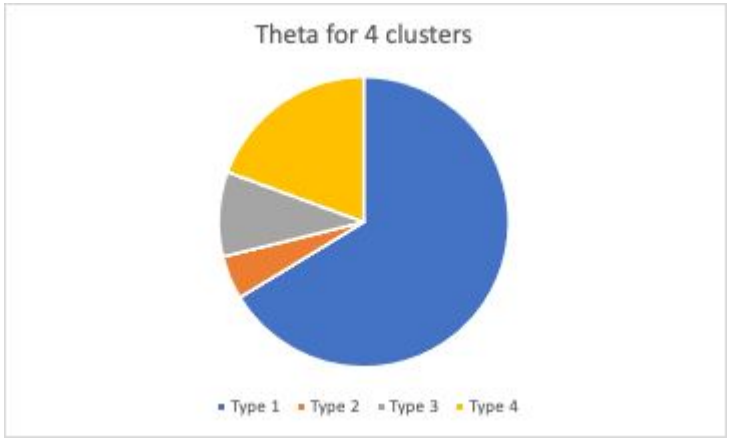
Beta1	Beta2	Beta3	Beta4
prop_starrating	**prop_review_score**	**prop_brand_bool**	**prop_location_score**
Beta5	Beta6	Beta7	Beta8
prop_accesibility_score	**prop_log_historica**	**price_usd**	**promotion_flag**

K means: N = 3 vs N = 4

Theta for N = 3

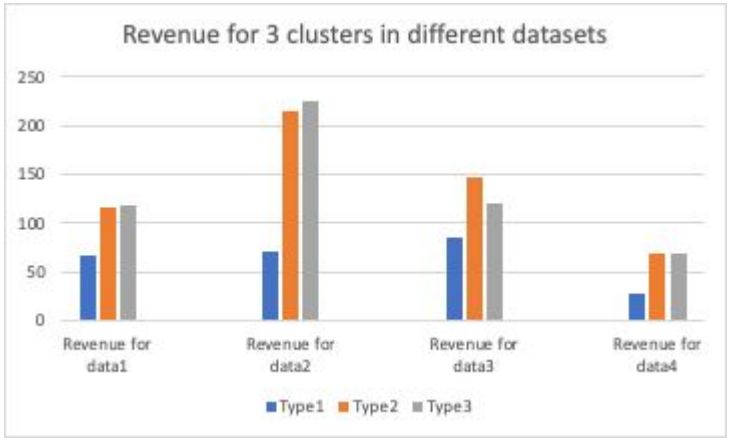


Theta for N = 4

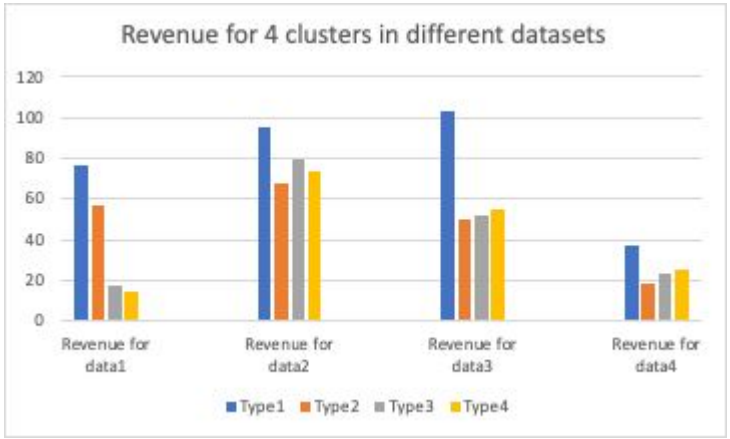


K means: N = 4 vs N = 3

Expected Revenue for N = 3

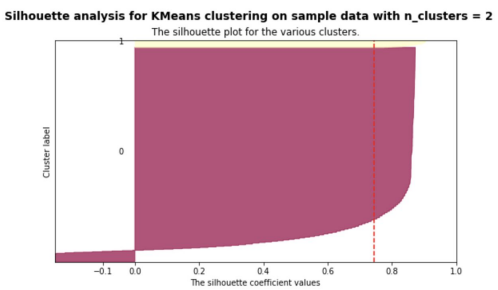


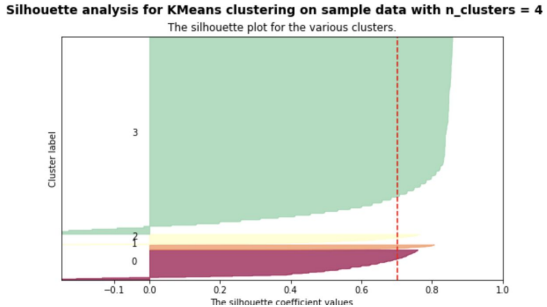
Expected Revenue for N = 4

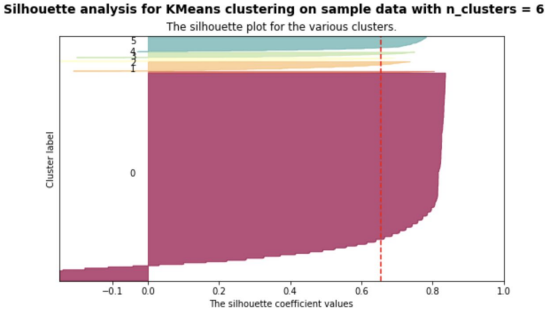


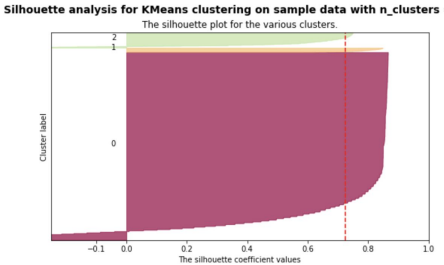
Spectral Clustering

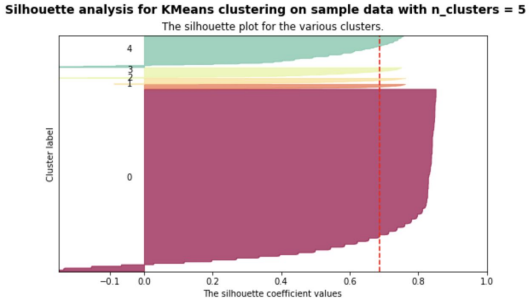
We also used Spectral Clustering to define customer types. Since this method does not compute any centers of clusters, we can only use **The Silhouette Method** to evaluate the performance of models with different number of clusters.

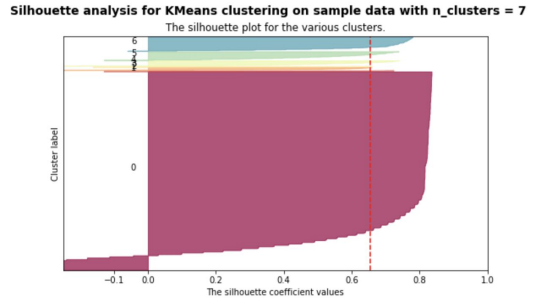






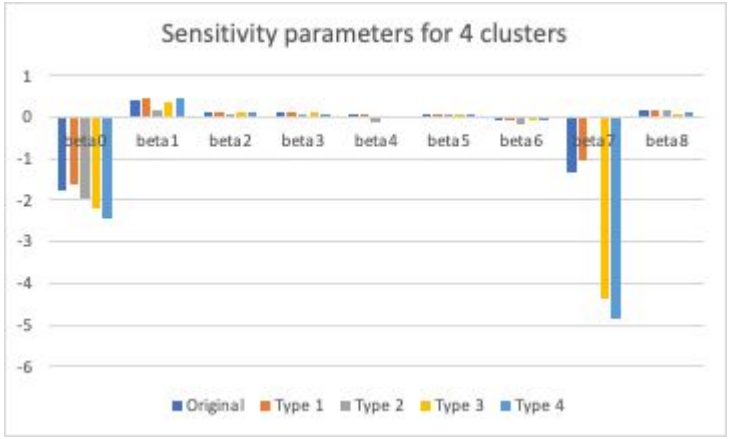




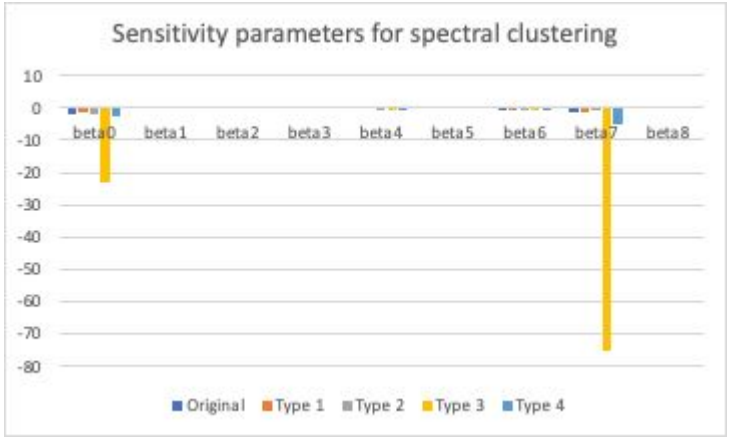


N = 4: K means vs Spectral Clustering

Beta for K means



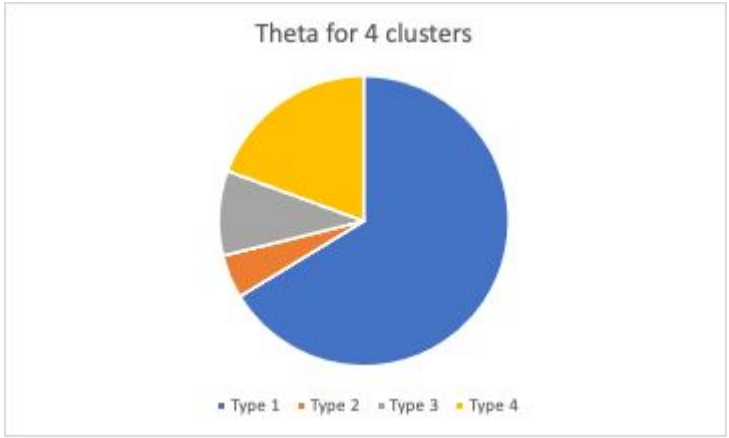
Beta for Spectral Clustering



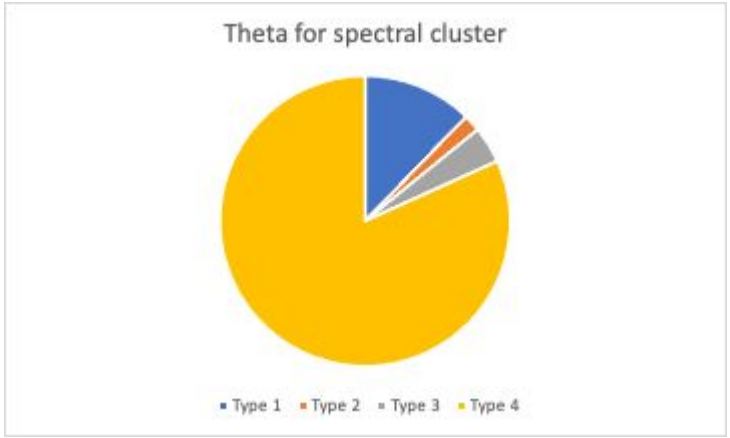
Beta1	Beta2	Beta3	Beta4
prop_starrating	**prop_review_score**	**prop_brand_bool**	**prop_location_score**
Beta5	Beta6	Beta7	Beta8
prop_accesibility_score	**prop_log_historica**	**price_usd**	**promotion_flag**

N = 4: K means vs Spectral Clustering

Theta for K means

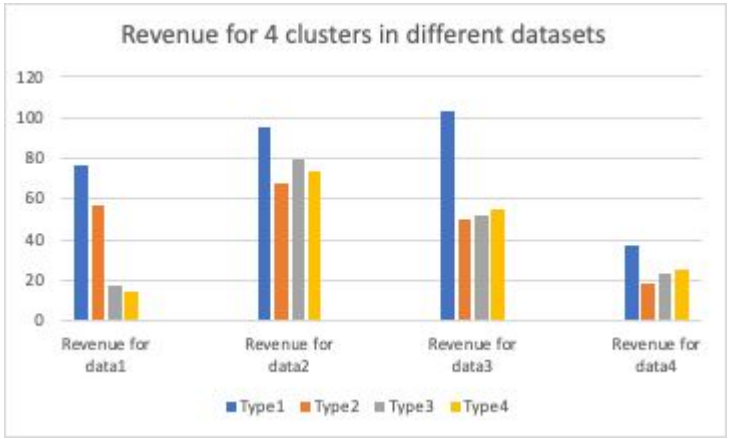


Theta for Spectral Clustering

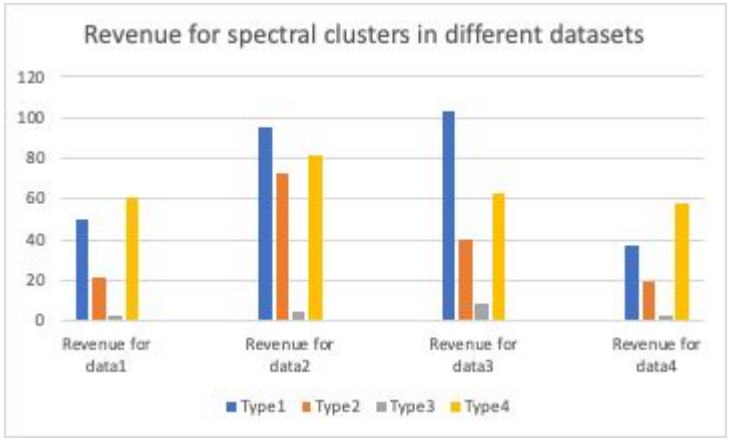


N = 4: K means vs Spectral Clustering

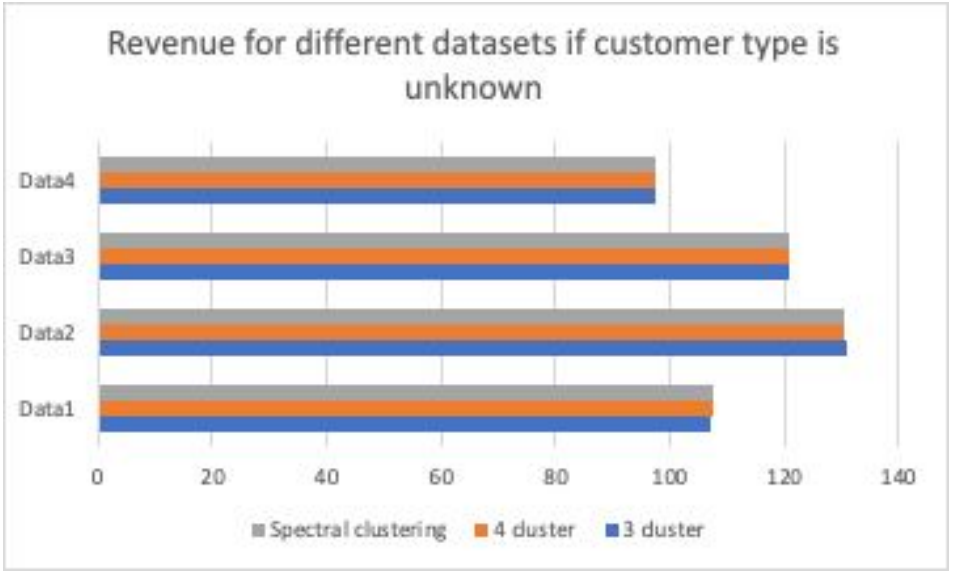
Expected Revenue for K means



Expected Revenue for Spectral Clustering



Expected revenue comparison for all three method



Conclusion

- K means N = 3 vs N = 4:
  - Revenue of k = 3 is higher than k = 4
  - Elbow Method find “best n value” based on sum-of-squares error, not based on revenue
  - Consider consumer habits

- K means vs Spectral Clustering when N = 4 : Revenue close, cost different
- we should cluster based on promotion flag, prob starting and booking window.

Thanks for Watching!