

|  |
| --- |
| Analytical Decision Modeling Tools I |
|  |
| Course Project Report  Team: B5  Project Sponsor: Room8 |

Team Members

Hongyu Lyu

Maanav Setlur

Sarah Pickel

Sowmiya Muruganandam

[Executive Summary 3](#_Toc57837798)

[Objective 4](#_Toc57837799)

[Business Problem 4](#_Toc57837800)

[Methodology 4](#_Toc57837801)

[Overview: 4](#_Toc57837802)

[Data Collection: 4](#_Toc57837803)

[Method 1: 5](#_Toc57837804)

[Method 2: 5](#_Toc57837805)

[Clustering Model 5](#_Toc57837806)

[Cluster Plots: 8](#_Toc57837807)

[Matching Model 10](#_Toc57837808)

[Caveats 12](#_Toc57837809)

[Conclusion 12](#_Toc57837810)

[Appendix 13](#_Toc57837811)

[Python code for Twitter Data Mining using Tweepy 13](#_Toc57837812)

[Python code for NLP based Attribute extraction 14](#_Toc57837813)

[Python CVX implementation for Linear Matching model 16](#_Toc57837814)

[Citation 17](#_Toc57837815)

Executive Summary

|  |
| --- |
| The rise in the number of Gen Z and Millennial population is booming job centers like San Francisco, New York, among many other cities. Their quest for affordable living has fueled the co-living market's growth in the United States.  ROOM8, a data-driven organization, caters to this demographic's needs by offering a platform to find the perfect roommate. The company is now expanding its market share by finding new users for their social media platform and targeting them with their marketing campaign for a better conversion rate.  In this project, we have helped ROOM8 identify potential users looking for roommates on social media, grouped them into market segments/clusters and provided a matching algorithm to find the appropriate match to the users using Optimization. |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Objective  This project explores the possibilities of using Optimization to cluster market segments on social media platforms and provide a stable matching algorithm for ROOM8.  Business Problem  In this project, we are trying to solve two business problems using Optimization,   1. To Identify Market Clusters   We can see a considerable increase in the number of users trying to find rooms/roommates using social media. People post statuses/tweets, join Facebook/Twitter groups, or even attend Facebook events to look for roommates and find a suitable place to rent. These users could be the prospective users of the ROOM8 platform. Identifying such users and targeting them would be the best expansion strategy for market growth for ROOM8. We tried to incorporate an Optimization model to determine the clusters/segments of such users on social media.   1. To Match Roommates   Finding a stable match between roommates based on their priorities is the goal of ROOM8. We tried to find a stable match between the set of identified users using an optimization model.  Methodology Overview: 1) We mined social media for obtaining the list of people looking for roommates and their attributes.  2) With the help of an evolutionary solver, we grouped the identified users into different market segments/clusters with the K Means Ideology  3) We performed Linear Optimization in each cluster to identify a stable match between the users. Data Collection: We decided to mine data from the Twitter platform to identify users looking for roommates on social media. We wanted to represent each of these users in five attributes with a Boolean value: clean, student, night person, smoker, pet lover. We focused on one city for the user search: Los Angeles, California. Method 1: We mined Twitter data using the Tweepy library. We performed a hashtag, and keyword search to retrieve user tweets to a CSV file, then used Natural language processing to extract the desired attribute values from the retrieved tweets. Method 2: Due to mining limitations on the twitter algorithm, we could not fetch data older than 30 days. Hence we decided to extract tweets manually.  We used Natural Language Processing to extract our desired five attributes from the tweets.  After data cleansing, we chose 50 users out of the total users from the above two methods to build our optimization model. Clustering Model On our selected 50 users and their five Boolean attributes, we performed K Means clustering using the evolutionary solver in Excel. Our clustering model's goal is to minimize the distance of the cluster members to the cluster center by choosing optimal cluster centers.  Firstly, we assigned integer indexes to each of the users; then calculated standardized data for the attributes using the excel standardize function. This standardized data was used for further calculations.  We randomly chose cluster centers among the users and calculated the Euclidean distance between the standardized user values and the first level cluster centers. We used the evolutionary solver to find optimal cluster centers by minimizing the standard error between the users and the cluster centers.  We performed clustering with different K sizes to identify the optimal number of clusters representing our user segments.  We used the four Step Model development strategy to build our clustering model, as below:  Step1: Identify relevant concepts   1. Inputs:   : index representing users, where  : index representing cluster, j  : cluster center  : user i  K: number of clusters   1. Objective:   To minimize the squared error   1. Decision Variable:   : cluster center   1. Constraints:       Step2: Setup Algebraic Model  The algebraic model to minimize distance is as below,  )    Step3: Execute  “Spreadsheet modeling” is used to execute the above said algebraic model and the following results were obtained  Excel Results:  SSE for each segment is calculated and the attributes with the highest average in each segment are marked in crimson in the below tables   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Output for K=1 | | | | | | |  | Student | Pet Owner | Clean | Smoker | Night owl | |  |  |  |  |  |  | | AVERAGE | 0.18 | 0.10 | 0.38 | 0.12 | 0.24 | | Segments | Number | % | SSE/Segment |  |  | | Segment 1 | 50 | 100.0% | 38.1 | SSE Total | 38.1 | | TOTAL | 50 | 100.0% |  |  |  | |  | | | | | | |  |  |  |  |  |  | | Output for K=2 | | | | | | |  | Student | Pet Owner | Clean | Smoker | Night owl | |  |  |  |  |  |  | | Segment 1 | 0.16 | 0.10 | 0.00 | 0.16 | 0.29 | | Segment 2 | 0.21 | 0.11 | 1.00 | 0.05 | 0.16 | |  |  |  |  |  |  | | Total | Number | % | SSE/Segment |  |  | | Segment 1 | 31 | 62.0% | 17.5 | SSE Total | 25.9 | | Segment 2 | 19 | 38.0% | 8.4 |  |  | | TOTAL | 50 | 100.0% |  |  |  | |  |  |  |  |  |  | | Output for K=3 | | | | | | |  | Student | Pet Owner | Clean | Smoker | Night owl | |  |  |  |  |  |  | | Segment 1 | 0.25 | 0.08 | 0.25 | 0.00 | 1.00 | | Segment 2 | 0.14 | 0.09 | 0.00 | 0.23 | 0.00 | | Segment 3 | 0.19 | 0.13 | 1.00 | 0.06 | 0.00 | |  |  |  |  |  |  | | Total | Number | % | SSE/Segment |  |  | | Segment 1 | 12 | 24.0% | 5.4 |  |  | | Segment 2 | 22 | 44.0% | 8.3 | SSE Total | 18.8 | | Segment 3 | 16 | 32.0% | 5.1 |  |  | | TOTAL | 50 | 100.0% |  |  |  | |  |  |  |  |  |  | | Output for K=4 | | | | | | |  | Student | Pet Owner | Clean | Smoker | Night owl | |  |  |  |  |  |  | | Segment 1 | 1.00 | 0.11 | 0.44 | 0.00 | 0.33 | | Segment 2 | 0.00 | 0.00 | 0.17 | 1.00 | 0.00 | | Segment 3 | 0.00 | 0.14 | 0.00 | 0.00 | 0.33 | | Segment 4 | 0.00 | 0.07 | 1.00 | 0.00 | 0.14 | |  |  |  |  |  |  | | Total | Number | % | SSE/Segment |  |  | | Segment 1 | 9 | 18.0% | 5.1 |  |  | | Segment 2 | 6 | 12.0% | 0.8 | SSE Total | 15.8 | | Segment 3 | 21 | 42.0% | 7.2 |  |  | | Segment 4 | 14 | 28.0% | 2.6 |  |  | | TOTAL | 50 | 100.0% |  |  |  | |  |  |  |  |  |  | | Output for K=5 | | | | | | |  | Student | Pet Owner | Clean | Smoker | Night owl | |  |  |  |  |  |  | | Segment 1 | 0.20 | 0.00 | 0.00 | 0.00 | 0.00 | | Segment 2 | 0.27 | 0.00 | 0.27 | 0.00 | 1.00 | | Segment 3 | 0.20 | 1.00 | 0.40 | 0.00 | 0.20 | | Segment 4 | 0.00 | 0.00 | 0.17 | 1.00 | 0.00 | | Segment 5 | 0.15 | 0.00 | 1.00 | 0.00 | 0.00 | |  |  |  |  |  |  | | Total | Number | % | SSE/Segment |  |  | | Segment 1 | 15 | 30.0% | 2.4 |  |  | | Segment 2 | 11 | 22.0% | 4.4 | SSE Total | 12.1 | | Segment 3 | 5 | 10.0% | 2.8 |  |  | | Segment 4 | 6 | 12.0% | 0.8 |  |  | | Segment 5 | 13 | 26.0% | 1.7 |  |  | | TOTAL | 50 | 100.0% |  |  |  |  Cluster Plots: We have plotted the segments for each K as below,  1) K=2    2) K=3    3)K=4    4)K=5        Step4: Evaluation  We performed clustering with different K values to identify the best model that represents each of the segments, and calculated segment-wise SSEs. Cluster with K=5 has the minimum SSE and also represents the attributes well. Matching Model The next step is to find a stable match for the identified roommate users using Optimization. A match is stable only if the user is matched with the best possible option. The model is inspired by Irving’s algorithm, except that we used Linear optimization to make the match instead of eliminating the unstable pairs.  The individual cluster user count was smaller. Hence, we decided to find the best possible match for all our 50 users. The attributes for the users are captured in binary. Therefore, we used Hamming distance as a metric to measure the distance/dissimilarity between the users. The distance between users was calculated using python.  Since we have the distance between the users based on their preference attributes, the problem can be easily solved with the Linear Integer Optimization Approach by assigning the match between the users as zero or one with the necessary constraints.  We used the four Step Model development strategy to build our matching model, as below:  Step1: Identify relevant concepts   1. Inputs:   : index representing users in set 1, where  : index representing users in set 2, j  : hamming distance between the users  : Users   1. Objective:   To minimize the distance between the users   1. Decision Variable:   : binary decision variable   1. Constraints:       Step2: Setup Algebraic Model  The algebraic model to minimize distance between users is as below,    Step3: Execute  We used Python cvxpy to execute our linear matching model and obtained the following output,    The final output is a matrix that has a match between the users. The output file attached with this project report has the users in the row and column headers and the 1’s in the matrix represent a match and the 0’s represent no match. Every user is matched with one potential user in the group.  Step 4: Evaluation  We randomly picked out matched users and checked out the similarity measure to conclude that the model has provided us with an optimal solution. Caveats For the clustering model, there are certain limitations as below,   1. Conventional solvers cannot handle the increase in user size, and processing time might be an issue going forward 2. To find the optimal K value, we had to run several iterations, which could be eliminated if we included optimal calculation of K Value as part of the model 3. Automatically fetching and clustering large scale data on a time-on-time basis will be a problem with the existing model 4. We could also try other approaches like convex clustering to compare the best results   For the Matching model,  1) Increase in user base will increase the processing time. Conclusion Therefore, we have created two optimization models, one to cluster market segments and one to find a stable match. AppendixPython code for Twitter Data Mining using Tweepy    Python code for NLP based Attribute extraction      Python CVX implementation for Linear Matching model  Citation “How Does the Template Calculation Work?” *Cluster Analysis 4 Marketing*, 12 July 2020, www.clusteranalysis4marketing.com/technical-aspects-cluster-analysis/how-does-the-template-calculation-work/.  “Stable Roommates Problem.” *Wikipedia*, Wikimedia Foundation, 28 Aug. 2020, en.wikipedia.org/wiki/Stable\_roommates\_problem.  “Scipy.spatial.distance.hamming¶.” *Scipy.spatial.distance.hamming - SciPy v1.5.4 Reference Guide*, docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.hamming.html.  IRVING, ROBERT W. “An Efficient Algorithm for the ‘Stable Roommates’ Problem .” *JOURNAL OF ALGORITHMS 6*, 1 May 1984, www.dcs.gla.ac.uk/~pat/jchoco/roommates/papers/Comp\_sdarticle.pdf. |