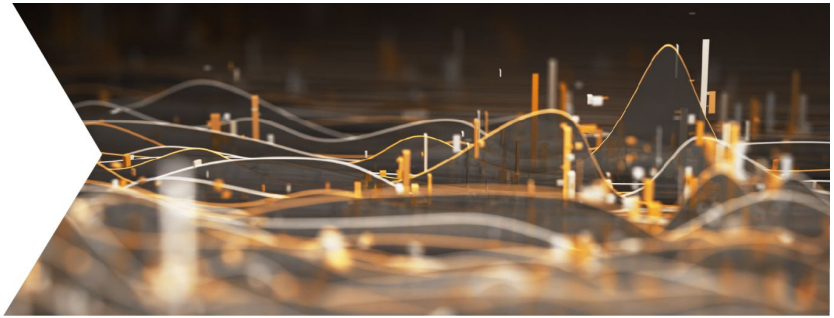
How to Use Data

EAS 5740



EAS 5470 Final Project

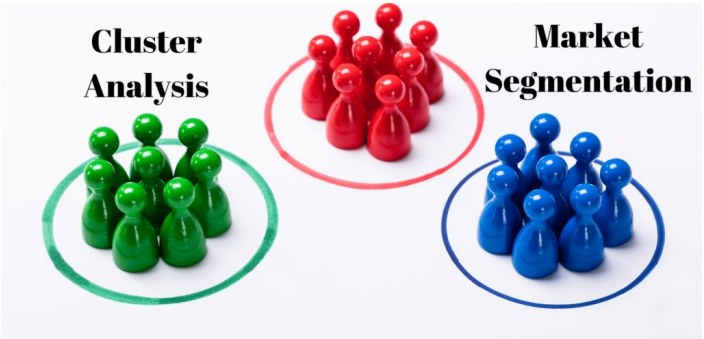
Contributors: Shihao Zhang (MCIT), Willis Yee (MCIT), Simon Alam (MCIT)

Defining the problem

Technical objective: Customer Segmentation (Track 1)

Business questions:

- Can you identify distinct user segments based on their reviews?
- What are the key characteristics of each segment?
- How can marketing strategies be tailored to each segment to promote businesses and increase their customer base?



Data collection and preparation

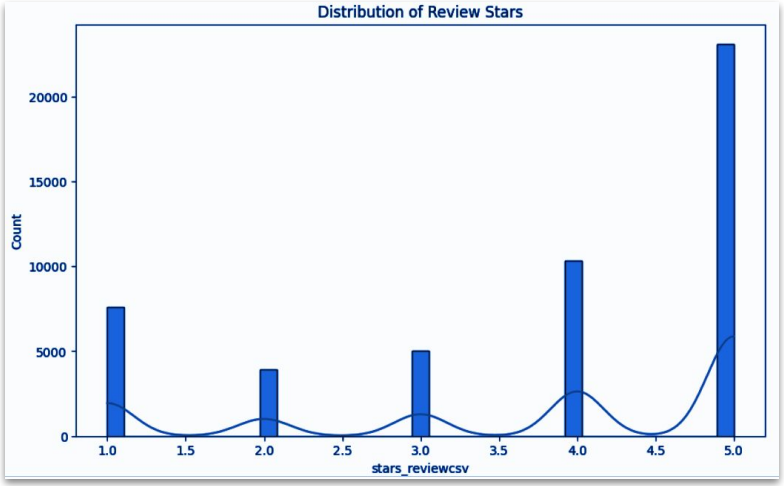
- Yelp businesses dataset - 29,527 rows, 60 columns
- Yelp reviews dataset - 50,000 rows, 9 columns
- Yelp users dataset - 45,019 rows, 22 columns
- Loaded and merged datasets in Google Colab
    - Merged “reviews” dataset with “businesses” dataset on “business_id” column, creating “merged_df”
    - Merged “merged_df” with “users” dataset on “user_id” column

Data Exploration

Identify characteristics of the columns from the merged df:

- .columns, .dtypes, .isnull(), .head(), .describe(), etc

Utilize plotting and matrices:



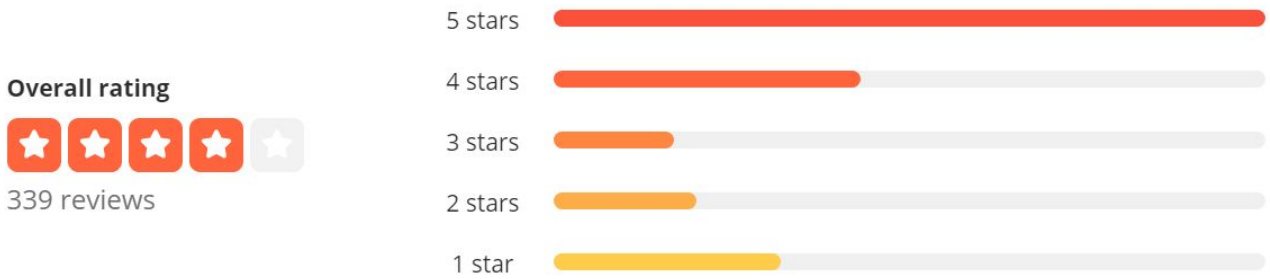
Correlation Matrix:

	stars_reviewcsv	useful_reviewcsv	funny_reviewcsv	\
stars_reviewcsv	1.00	-0.07	-0.04	
useful_reviewcsv	-0.07	1.00	0.74	
funny_reviewcsv	-0.04	0.74	1.00	
cool_reviewcsv	0.06	0.82	0.82	
review_count_usercsv	0.03	0.24	0.21	
useful_usercsv	0.02	0.41	0.41	
funny_usercsv	0.01	0.38	0.43	
cool_usercsv	0.02	0.41	0.42	
fans_usercsv	0.03	0.35	0.31	
average_stars_usercsv	0.58	-0.01	-0.00	
compliment_hot_usercsv	0.01	0.45	0.53	

Initial trends/patterns?

Insights:

- Reviews tend to skew toward 5.0 stars, and the average around 4 stars.
- Out of the numerical columns, the useful, cool, and funny columns are the most correlated.
- The top ten business categories predominantly revolve around food.



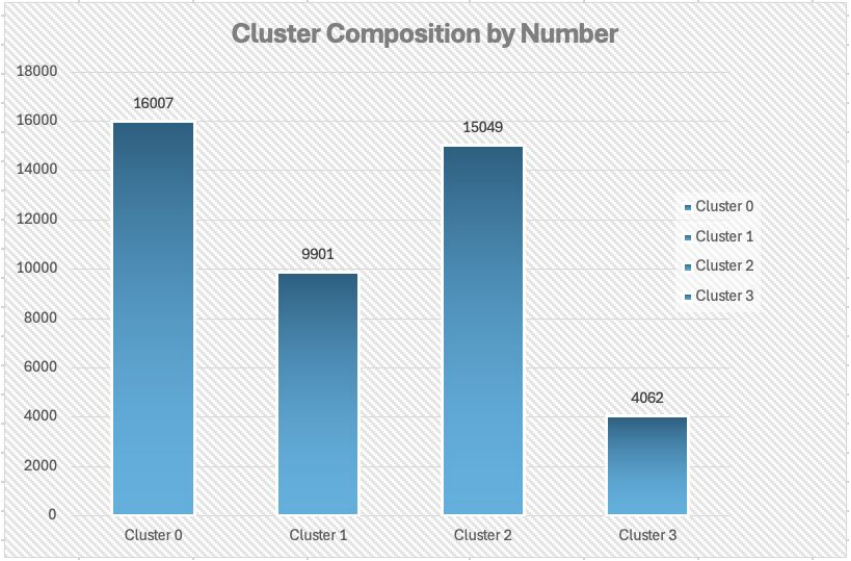
Modeling Approach(es) and Techniques

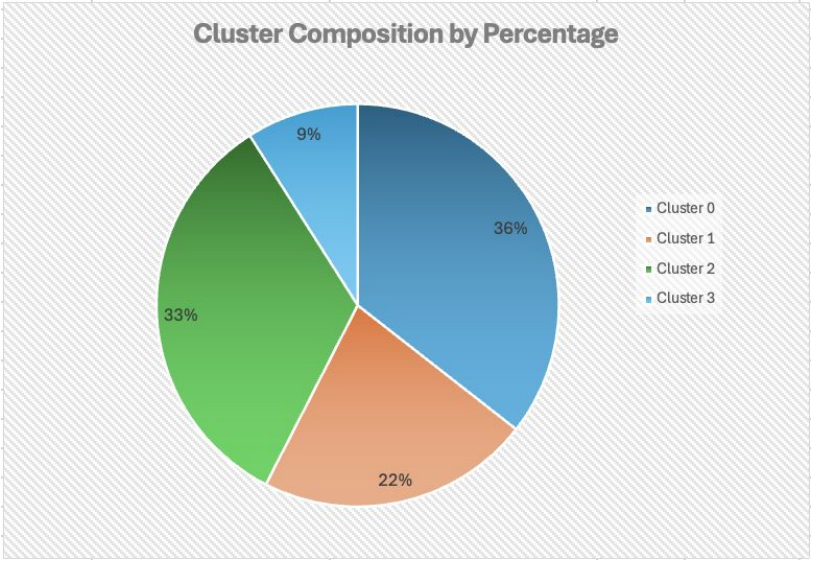
Techniques:

- Feature Engineering: create new data out of existing columns,
  - i.e. avg review length, engagement score (combine useful, cool, funny)
- Normalization of features
  - Min-max scaling
- Kmeans
  - Elbow method
- PCA
  - Reduce overfitting, allow visualization

Interpretation of Results

- Our segmentation has split users into four clusters, ranging from 4062 user in the smallest cluster and 16007 in the largest cluster.





Interpretation of Results

- Cluster 0 are 'enthusiastic users' who are enthusiastic about writing reviews and have many fans.
- Cluster 3 are 'invisible users' who normally won't give reviews, tend to give lower scores and don't have many fans.
- Cluster 1 are 'regular, picky users' who are not very active nor invisible, and tend to give relatively low score.
- Cluster 2 are 'regular, undemanding users' who are not very active nor invisible, and tend to give relatively high score.

By learning the characteristics of different clusters, we will be able to draft a customer profile for each cluster and predict customer behaviors based on customer profile.

Characteristics	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Average
avg_review_length	585.96	646.6	405.7	621.28	542.23
review_frequency	11.04	5.16	5.42	0.94	6.96
category_diversity	6.04	5.38	5.56	4.96	5.64
avg_rating_deviation	0.48	-1.35	0.76	-1.56	-0.02
engagement_score	1030.35	217.44	196.6	12.77	481.05
tenure_days	4965.37	3859.08	3148.6	3136.83	3949.77
average_stars_usercsv	3.91	3.35	4.36	1.67	3.73
compliment_hot_usercsv	23.16	3.52	2.58	0.01	9.88
fans_usercsv	15.43	3.23	3.33	0.12	7.32
review_count_usercsv	155.12	56.77	46.7	8.29	84

Green: Highest

Red: Lowest

Interpretation of Results

By calculating the TRUE ratio of different business attributes for each cluster, we will be able to get some senses of what business attributes that each cluster cares and to design the marketing strategies accordingly.

- Cluster 0: marketing bike parking and reservation.
- Cluster 1: marketing bike RestaurantsTableService and AcceptsCreditCards.
- Cluster 2: marketing DogsAllowed, Caters, Open24hours and OutdoorSeating.
- Cluster 3: marketing DriveThru, GoodForKids, HasTV and Delivery.
- Relatively, all three clusters much care about AcceptCreditCards, the TRUE ratio is above 96%.

TRUE for Business Attributes	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Average
GoodForKids	77.92%	82.15%	83.43%	84.31%	80.99%
DogsAllowed	25.74%	24.93%	27.32%	23.22%	26.01%
RestaurantsDelivery	62.92%	66.75%	69.26%	72.00%	66.43%
Caters	54.44%	55.44%	59.07%	58.14%	56.40%
RestaurantsTableService	75.22%	76.95%	75.99%	74.52%	75.82%
OutdoorSeating	64.49%	61.94%	67.34%	58.80%	64.47%
HasTV	70.29%	75.84%	73.30%	80.19%	73.11%
BusinessAcceptsCreditCards	96.02%	97.21%	97.21%	96.99%	96.73%
BikeParking	82.85%	80.21%	81.75%	75.96%	81.46%
RestaurantsReservations	47.68%	46.43%	45.61%	41.22%	46.32%
DriveThru	24.69%	30.33%	24.01%	43.93%	27.48%
Open24Hours	26.53%	10.00%	41.18%	28.57%	27.71%

Green: Highest

Red: Lowest