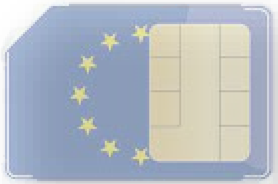
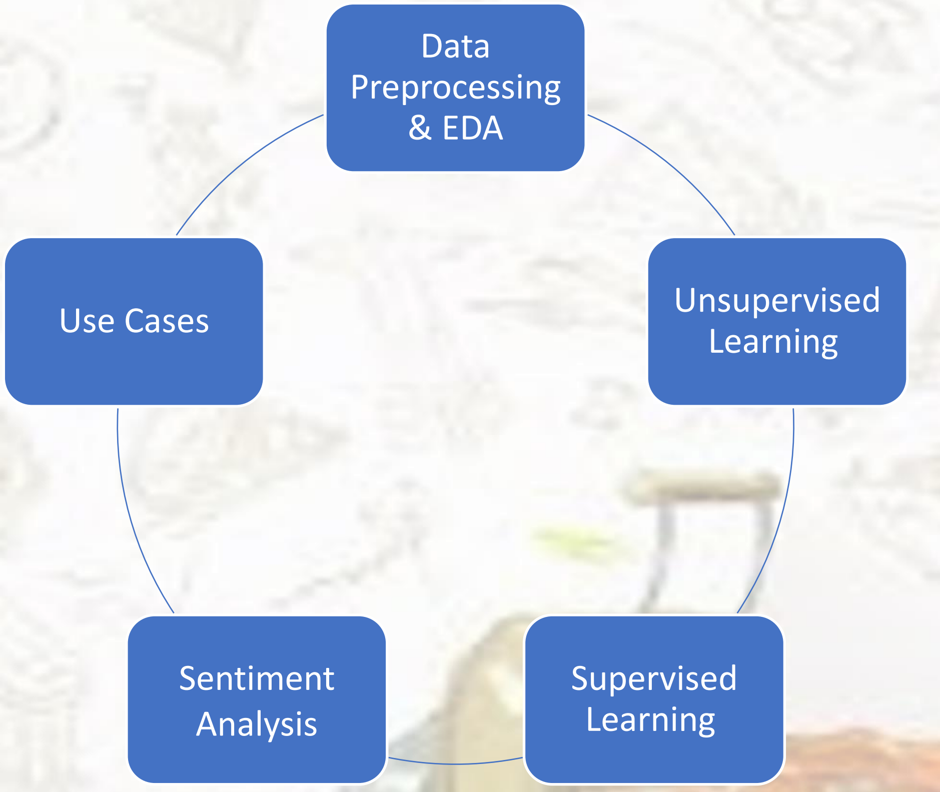
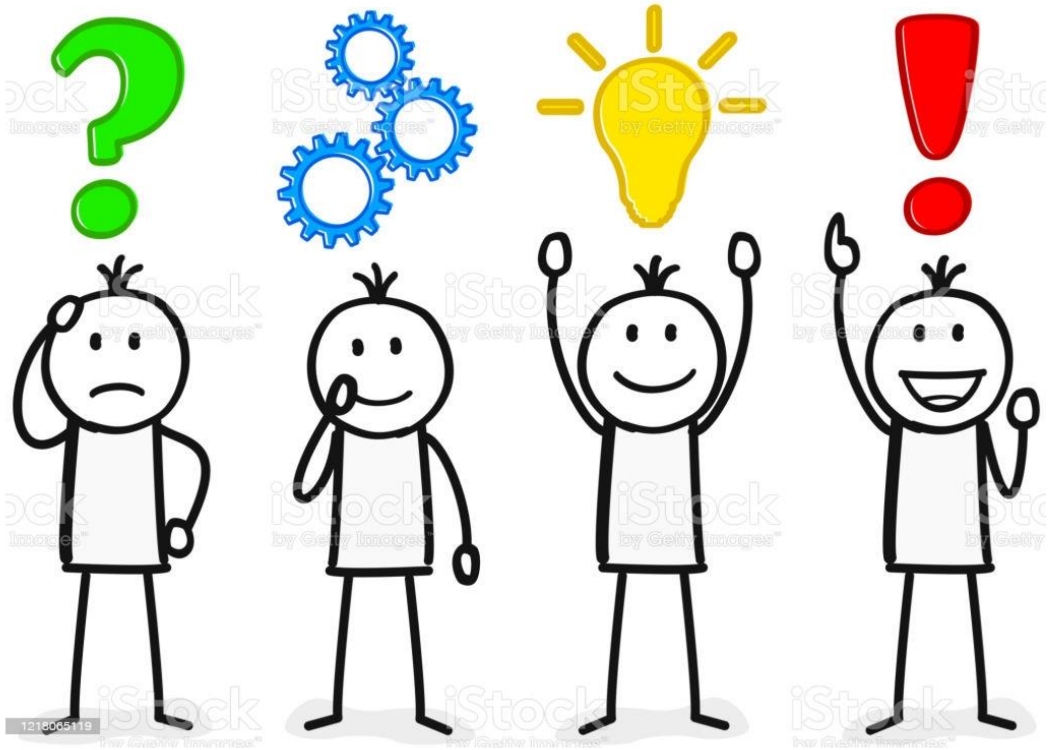
Agenda



1. Introduction and Problem Statement
2. Data Pre-processing & EDA
3. Unsupervised
4. Supervised
5. Sentiment analysis
6. Use case and deployment



1. Introduction & Problem Statement



Introduction

- Analyst Group presenting to Europe Tourism Board
- Data set is from Google Review & Trip Advisor
- 24 Attractions are being reviewed
- Costumer's feedback review on hotels
- Rating is an average from a scale from 0 to 5 (5 being the best)

Google

Tripadvisor

EUROPEAN
TRAVEL
COMMISSION

Introduction

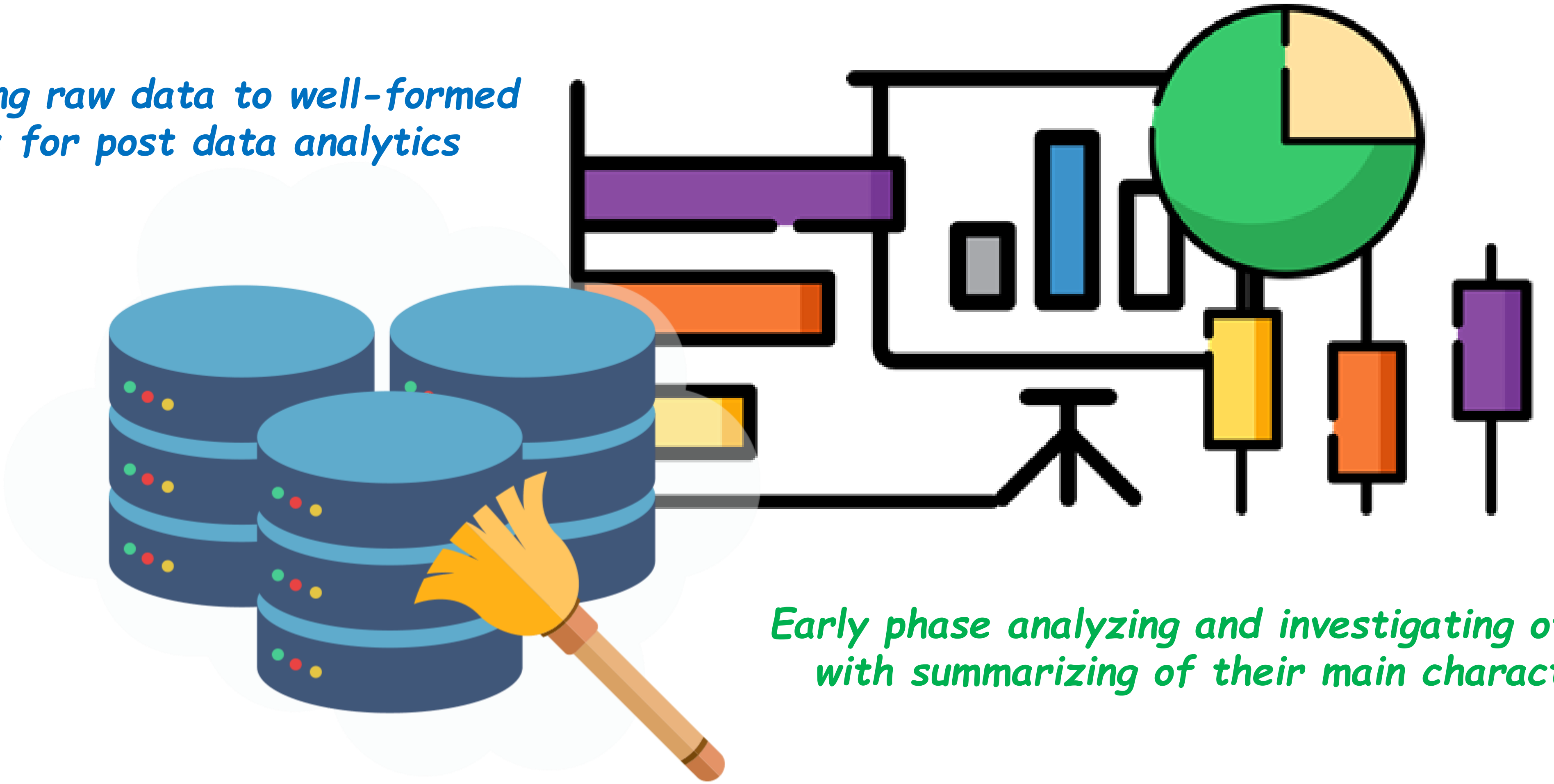
- **Goal is to boost tourism across Europe**
- **Provide content to advertise & attract different costumer profiles**
- **Analyze feedback to target improvement**

Problem Statement

- **Understanding & identifying costumer profile.**
- **Characteristic of each attraction.**
- **Correlation between customers & attraction.**
- **Improve service quality & hospitality experience.**

2. Data Preprocessing & EDA

Transforming raw data to well-formed datasets for post data analytics



Early phase analyzing and investigating of datasets with summarizing of their main characteristics

Raw Data Understanding

In [2]:
```
# read the raw datasets from Google Travel Ratings .csv file on Kaggle

travel_review = pd.read_csv("../mining/google_review_ratings.csv")
```

In [3]:
```
# preview the raw datasets from Google Travel Ratings .csv file on Kaggle

travel_review.head()
```

Out[3]:

	User	Category 1	Category 2	Category 3	Category 4	Category 5	Category 6	Category 7	Category 8	Category 9	...	Category 16	Category 17	Category 18	Category 19	Categor 2
0	User 1	0.0	0.0	3.63	3.65	5.0	2.92	5.0	2.35	2.33	...	0.59	0.5	0.0	0.5	0
1	User 2	0.0	0.0	3.63	3.65	5.0	2.92	5.0	2.64	2.33	...	0.59	0.5	0.0	0.5	0
2	User 3	0.0	0.0	3.63	3.63	5.0	2.92	5.0	2.64	2.33	...	0.59	0.5	0.0	0.5	0
3	User 4	0.0	0.5	3.63	3.63	5.0	2.92	5.0	2.35	2.33	...	0.59	0.5	0.0	0.5	0
4	User 5	0.0	0.0	3.63	3.63	5.0	2.92	5.0	2.64	2.33	...	0.59	0.5	0.0	0.5	0

5 rows × 26 columns

In [4]:
```
# understand the original shape of the datasets
# total of 5456 observations and 26 variables

travel_review.shape
```

Out[4]:
```
(5456, 26)
```

Data Cleansing Process

	User	Category 1	Category 2	Category 3	Category 4	Category 5	Category 6	Category 7	Category 8	Category 9	...	Category 16	Category 17	Category 18	Category 19
0	User 1	0.00	0.00	3.63	3.65	5.00	2.92	5.00	2.35	2.33	...	0.59	0.50	0.00	0.50
1	User 2	0.00	0.00	3.63	3.65	5.00	2.92	5.00	2.64	2.33	...	0.59	0.50	0.00	0.50
2	User 3	0.00	0.00	3.63	3.63	5.00	2.92	5.00	2.64	2.33	...	0.59	0.50	0.00	0.50
3	User 4	0.00	0.50	3.63	3.63	5.00	2.92	5.00	2.35	2.33	...	0.59	0.50	0.00	0.50
4	User 5	0.00	0.00	3.63	3.63	5.00	2.92	5.00	2.64	2.33	...	0.59	0.50	0.00	0.50
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5451	User 5452	0.91	5.00	4.00	2.79	2.77	2.57	2.43	1.09	1.77	...	0.66	0.65	0.66	0.69
5452	User 5453	0.93	5.00	4.02	2.79	2.78	2.57	1.77	1.07	1.76	...	0.65	0.64	0.65	1.59
5453	User 5454	0.94	5.00	4.03	2.80	2.78	2.57	1.75	1.05	1.75	...	0.65	0.63	0.64	0.74
5454	User 5455	0.95	4.05	4.05	2.81	2.79	2.44	1.76	1.03	1.74	...	0.64	0.63	0.64	0.75
5455	User 5456	0.95	4.07	5.00	2.82	2.80	2.57	2.42	1.02	1.74	...	0.64	0.62	0.63	0.78

5456 rows × 26 columns

In [15]:
```
# Data Pre-processing step 5 - Renaming the variables to description for easy representation
description = ['user_id', 'churches', 'resorts', 'beaches', 'parks', 'theatres', 'museums', 'malls', 'zoo', 'restaurants',
travel_review.columns = description
travel_review.head()
```

In [13]:
```
# Data Pre-processing step 4 - Perform conversion of the data-type to Category 11 into numerical
travel_review['Category 11'] = pd.to_numeric(travel_review['Category 11'], errors='coerce')
```

Missing Data

Unclear Variables

Incorrect Type

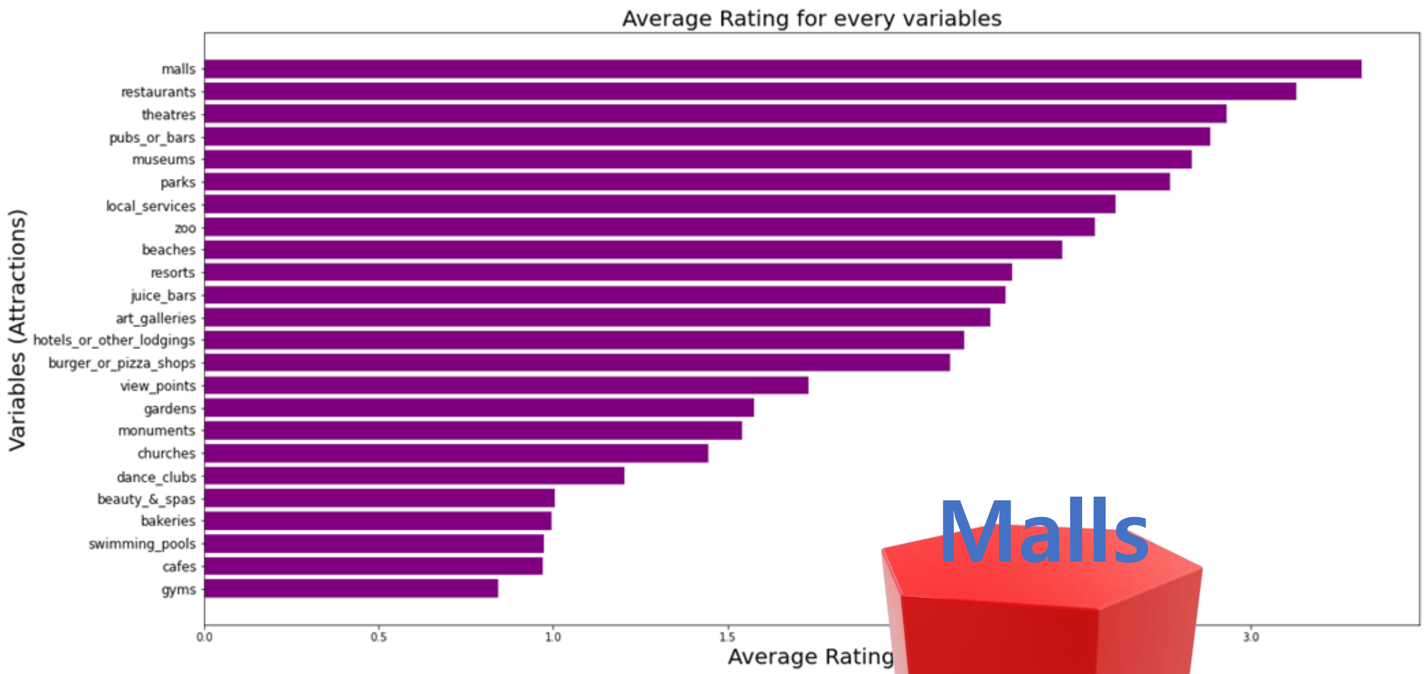
In [7]:
```
# Data Pre-processing step 2 - removal or replace missing data or invalid variables (columns)
# Based on observation, there is a unnamed column as "Unnamed: 25"
# Delete this invalid column
del travel_review['Unnamed: 25']
```

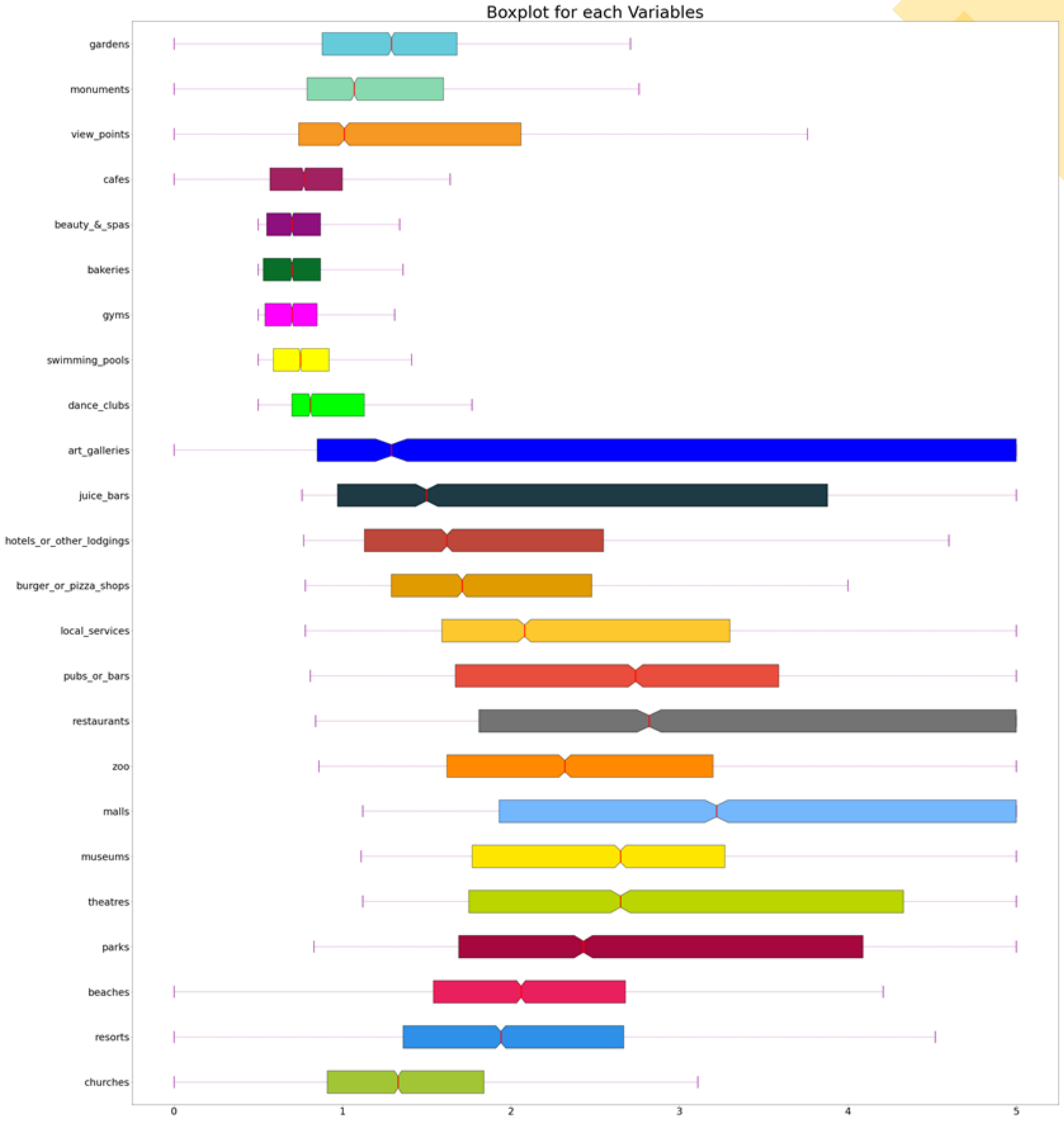
[9]:
```
# Data Pre-processing step 2 - removal or replace missing data or invalid variables (columns)
# Based on observation, categories 12 & 24 which is the rating score is less than 5% which can be removed.
travel_review = travel_review.dropna()
```

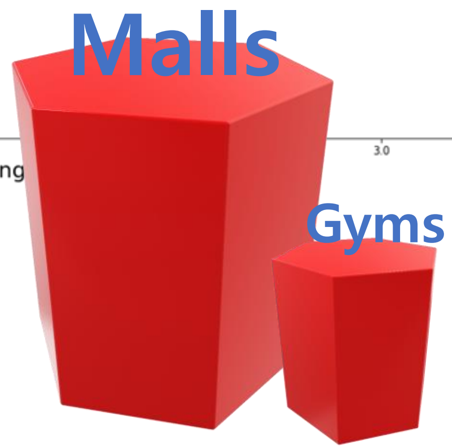
	user_id	churches	resorts	beaches	parks	theatres	museums	malls	zoo	restaurants	...	art_galleries	dance_clubs	swimming_pools	gyms	ba
0	User 1	0.00	0.00	3.63	3.65	5.00	2.92	5.00	2.35	2.33	...	1.74	0.59	0.50	0.00	
1	User 2	0.00	0.00	3.63	3.65	5.00	2.92	5.00	2.64	2.33	...	1.74	0.59	0.50	0.00	
2	User 3	0.00	0.00	3.63	3.63	5.00	2.92	5.00	2.64	2.33	...	1.74	0.59	0.50	0.00	
3	User 4	0.00	0.50	3.63	3.63	5.00	2.92	5.00	2.35	2.33	...	1.74	0.59	0.50	0.00	
4	User 5	0.00	0.00	3.63	3.63	5.00	2.92	5.00	2.64	2.33	...	1.74	0.59	0.50	0.00	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
5451	User 5452	0.91	5.00	4.00	2.79	2.77	2.57	2.43	1.09	1.77	...	5.00	0.66	0.65	0.66	
5452	User 5453	0.93	5.00	4.02	2.79	2.78	2.57	1.77	1.07	1.76	...	0.89	0.65	0.64	0.65	
5453	User 5454	0.94	5.00	4.03	2.80	2.78	2.57	1.75	1.05	1.75	...	0.87	0.65	0.63	0.64	
5454	User 5455	0.95	4.05	4.05	2.81	2.79	2.44	1.76	1.03	1.74	...	5.00	0.64	0.63	0.64	
5455	User 5456	0.95	4.07	5.00	2.82	2.80	2.57	2.42	1.02	1.74	...	0.85	0.64	0.62	0.63	

5454 rows × 25 columns

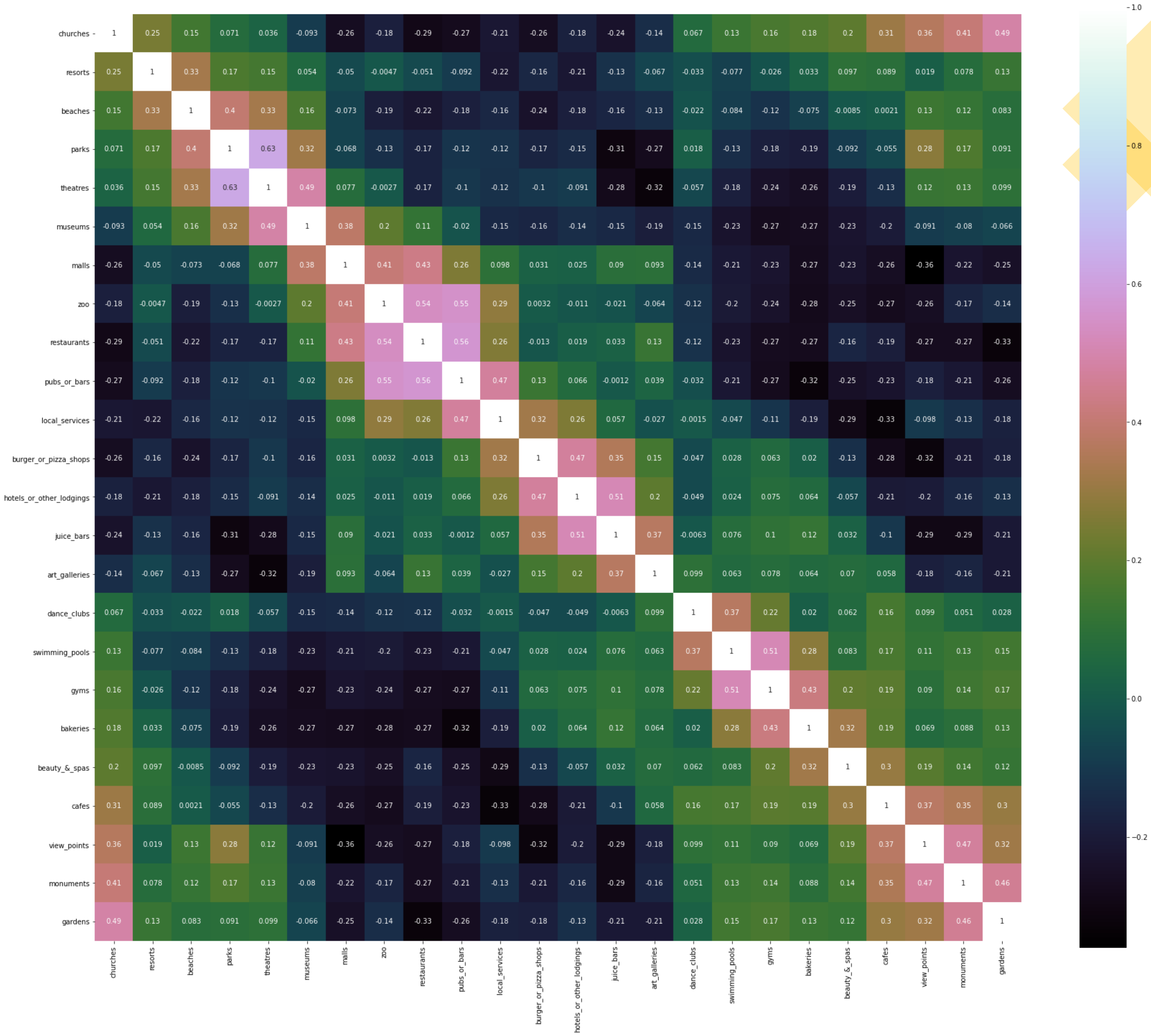
Exploratory Data Analysis







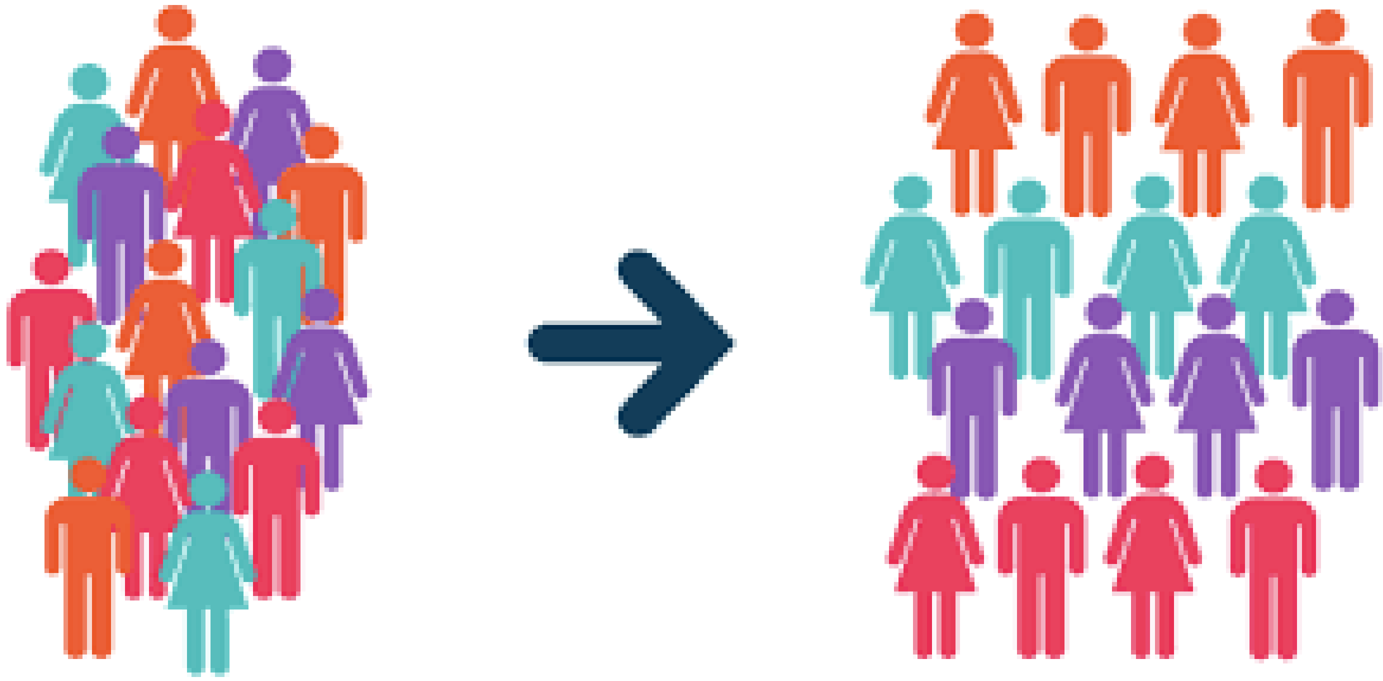
Exploratory Data Analysis



Representation of how each variable relates to others

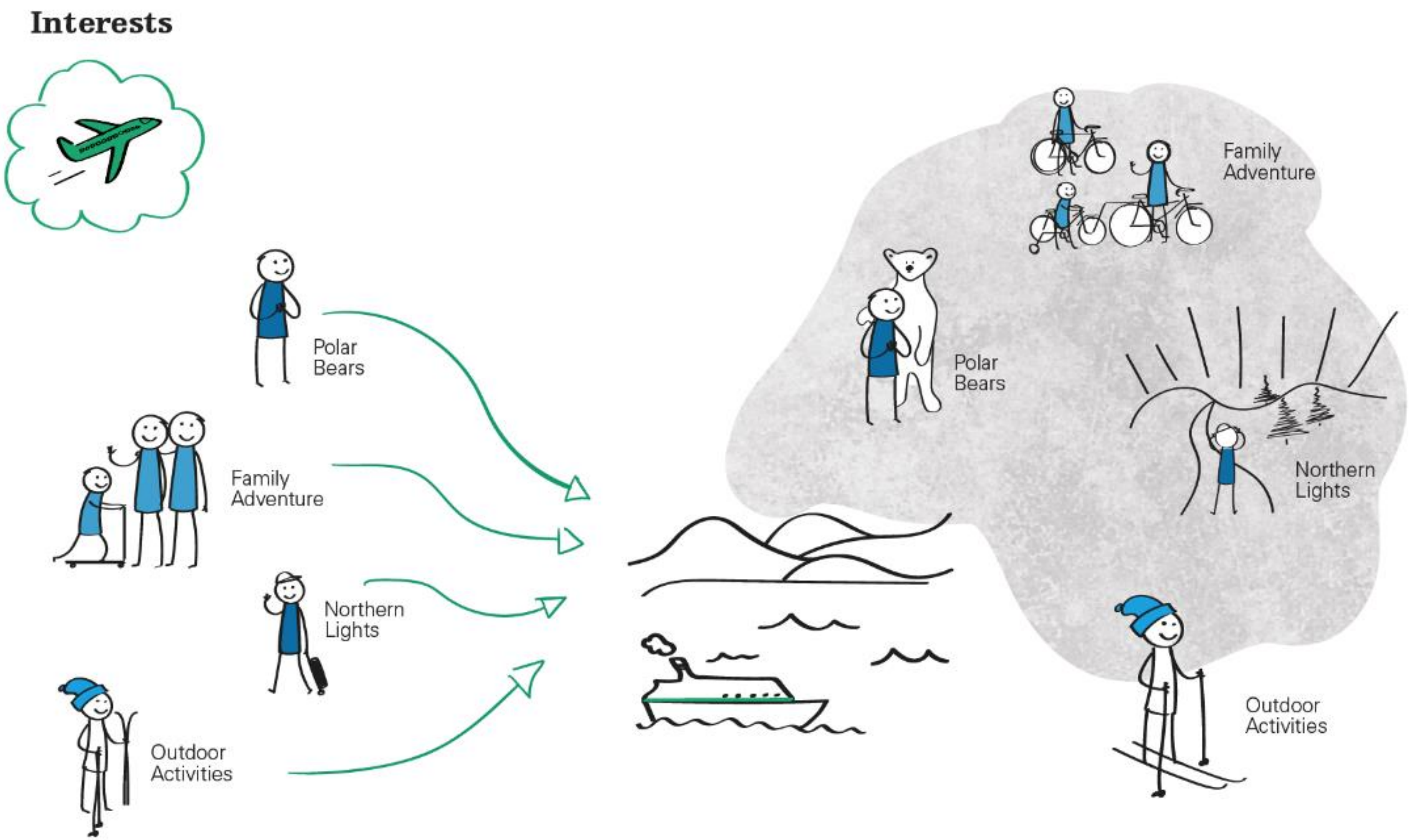
E.g. Theatres has a high correlation to Amusement Parks and moderately to Museums and Beaches

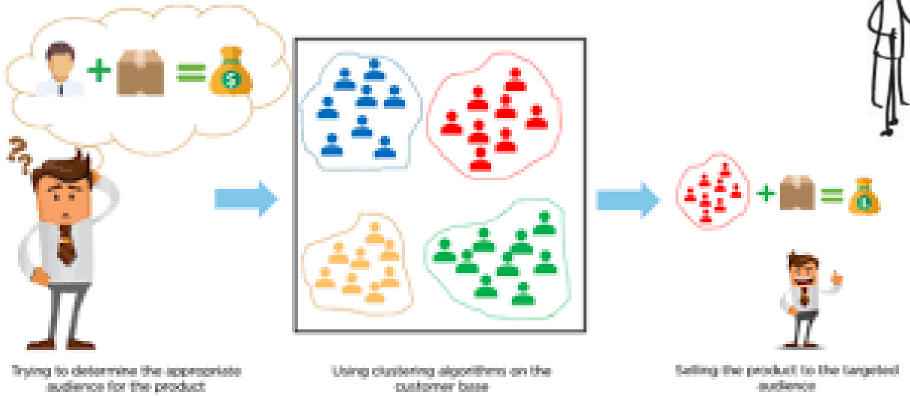
3. Unsupervised Learning



Purpose:

- Divide the Google review user into different clusters with similar interest.
- Allow tourists to get to preferred venues quickly.



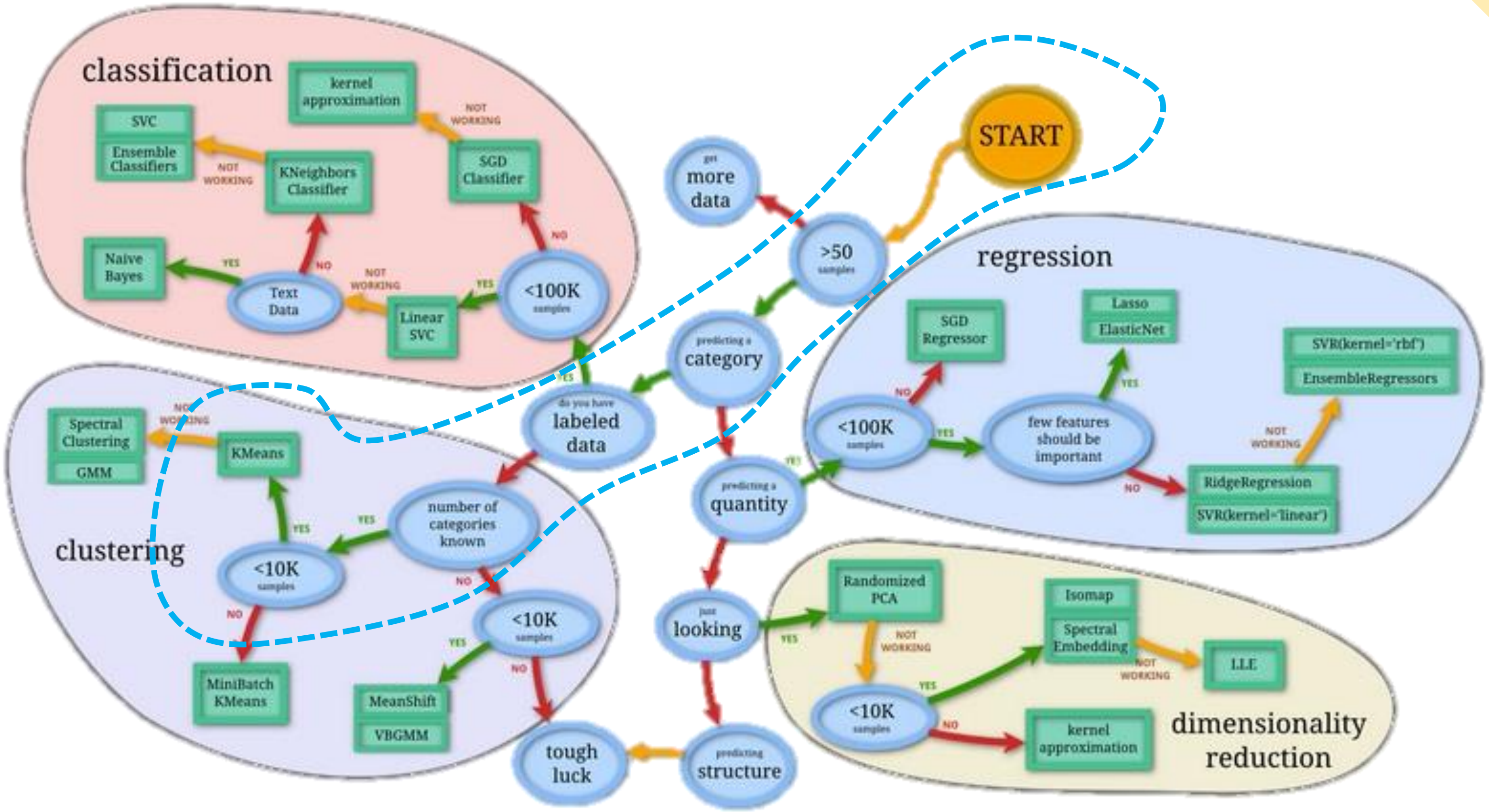


Model Selection----K-Means

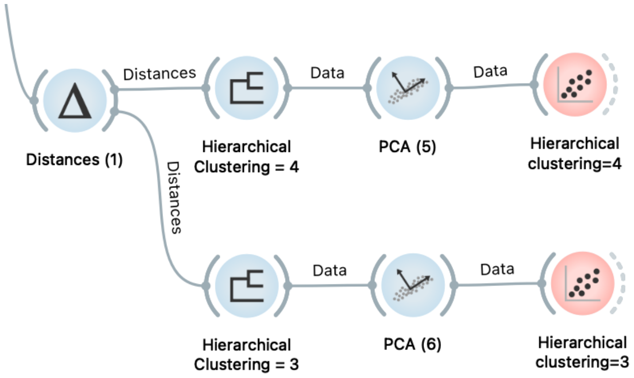
- **Algorithm:**

K-Means Clustering & Hierarchical Clustering

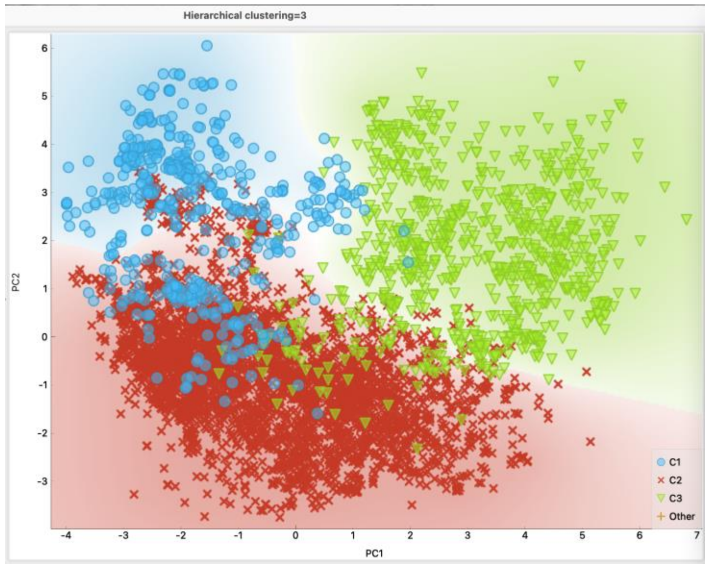
- **Try different algorithm on 2 scenarios and compared its result:**
  - ❑ **Hierarchical Clustering on original data (Cluster = 3/ Cluster = 4)**
  - ❑ **K-Means Clustering on scaled original data (Cluster = 3/ Cluster = 4)**



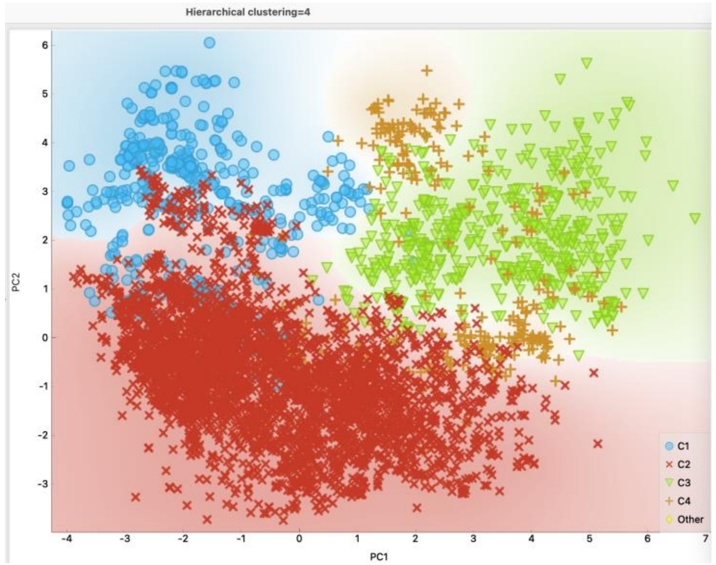
Hierarchical Clustering



Cluster = 3



How about
Cluster = 4 ?

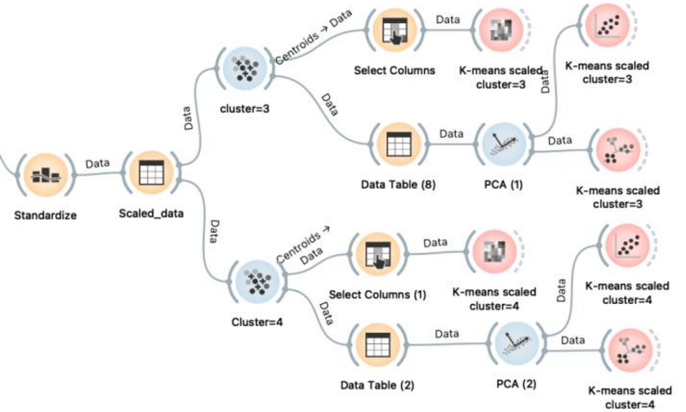


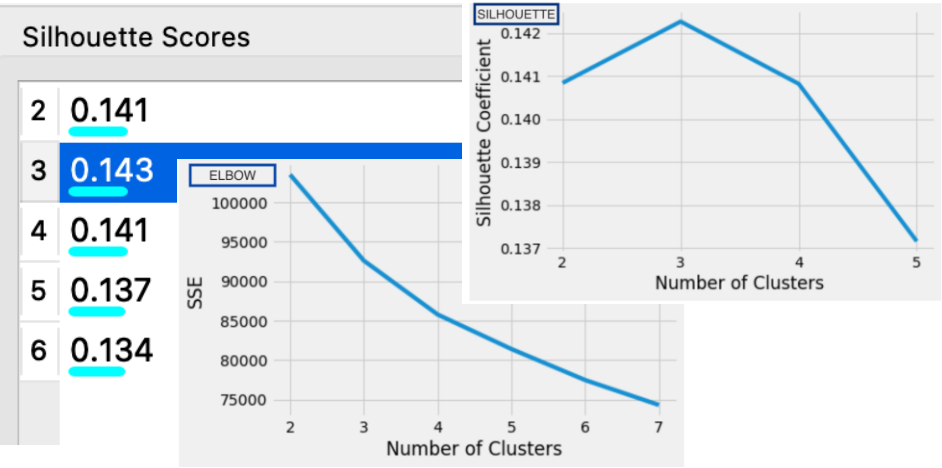
Overlapping heavily！！

Let's have a try on
k-means

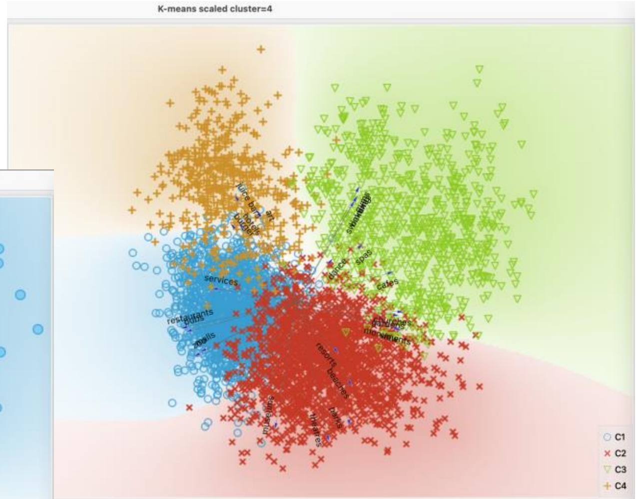
^_^

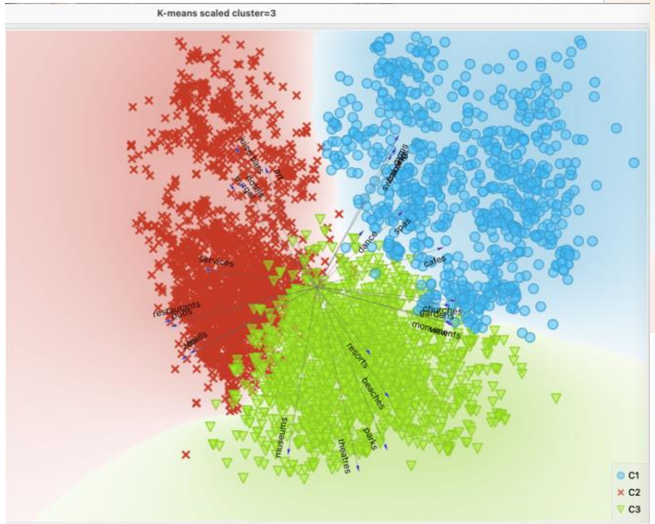
K-Means Clustering on scaled original data





The scores
are similar.
Cluster = 3 ? Or 4 ?



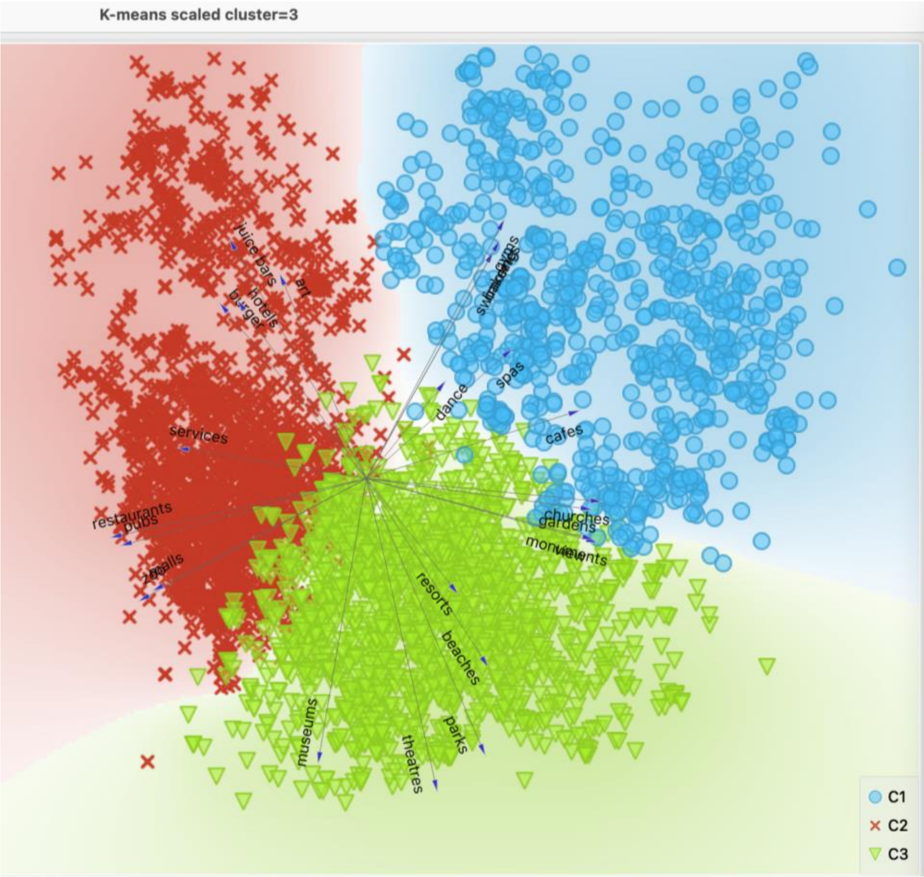


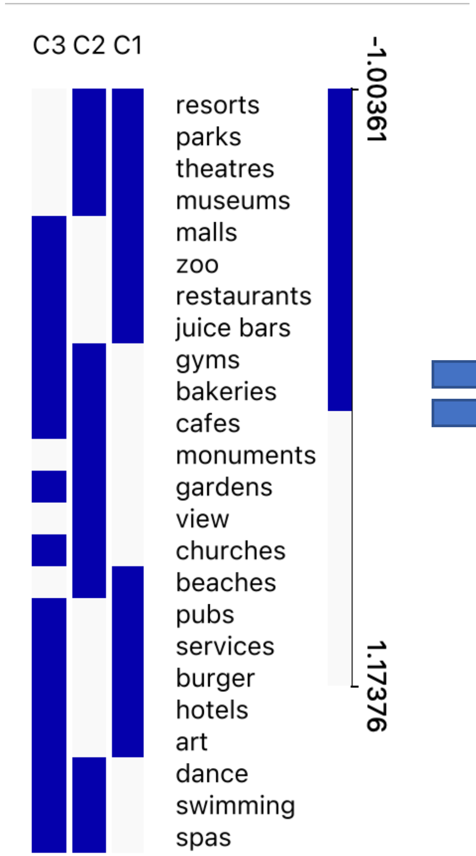
Similar >_<

It's hard to choose!

Result of K-Means Cluster=3

We can use it go for **Supervised learning** later! (cluster=3)

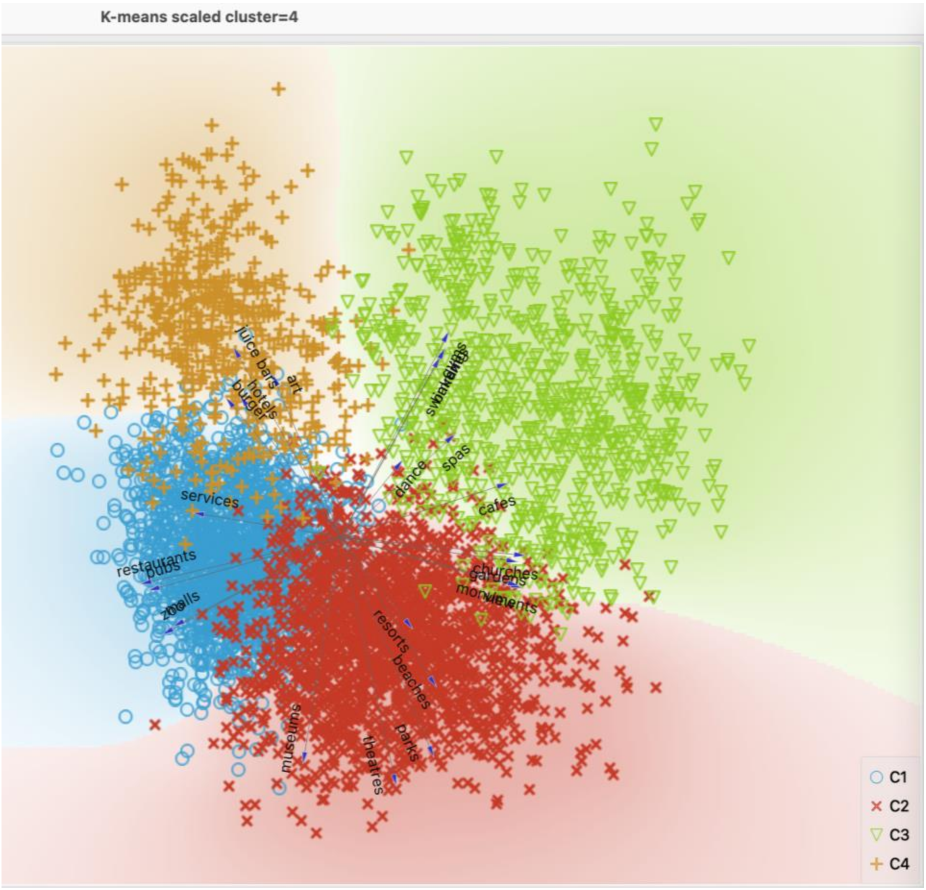


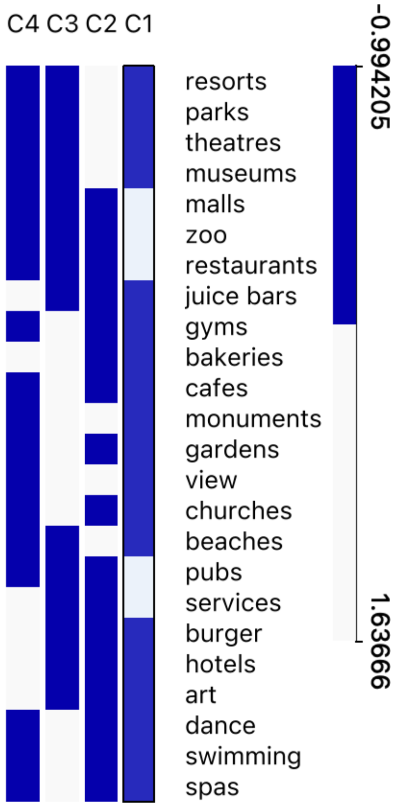


GROUP	CATEGORY
Cluster_1	Active and Healthy Leisure
Cluster_2	City Recreation
Cluster_3	Amusement Park

Result of K-Means Cluster=4

We can use it go for **Supervised learning** now! (cluster=4)



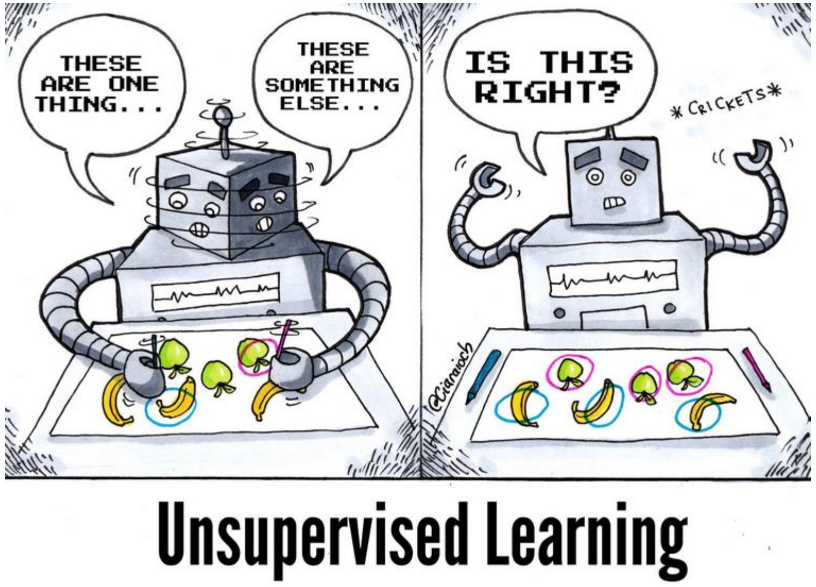


GROUP	CATEGORY
Cluster_1	City Recreation
Cluster_2	Amusement Park
Cluster_3	Active & Healthy Leisure
Cluster_4	Staycation

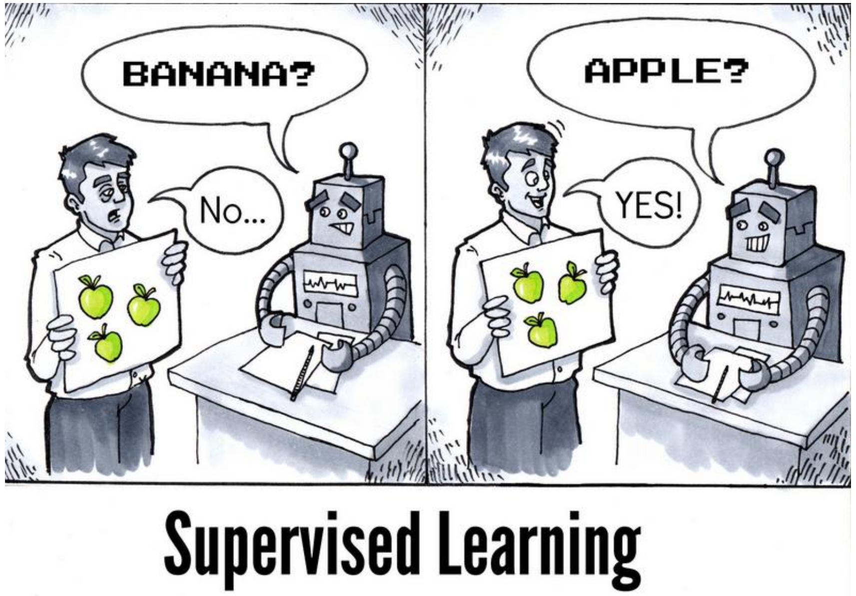
4. Supervised Learning

Predicting A Customer Travel Preference

- **Use the clusters from unsupervised learning as label for the supervised model.**
- **Classification models are used to determine the customer preference by classify the customer to the cluster characteristic.**



This group of object with similar features is Apple



Supervised Model Selection

- **Model selection based on the use case to :**

  i) understand the customer group profiling

  ii) use of the input features is important

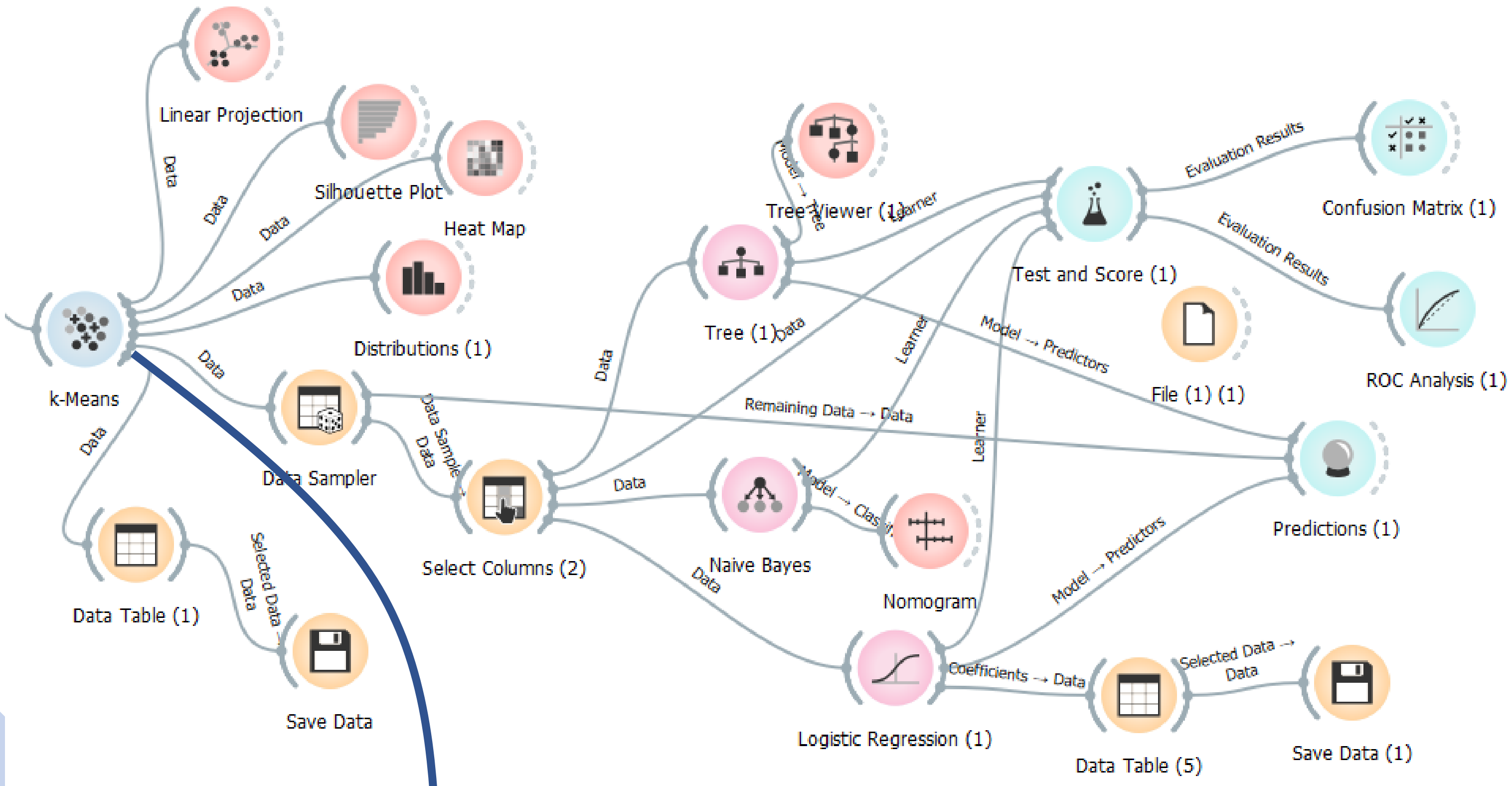
  iii) precision or type 2 error not vital

Hence high interpretability for decision classification models will be priority.

Classification Models	Interpretability	Performance
Decision Tree	High	Low
Random Forest	Low	High
Support Vector Machine	Low	High
K Nearest Neighbor	Low	High
Naive Bayes	High	Low
Logistic Regression	High	Moderate

Supervised Model Evaluation

- **Cluster from K-Means become the label/target for Supervised Model.**
- **As expected the Logistic Regression is a better model among the 3.**



Evaluation with 10-fold Cross Validation.

Test and Score (1)

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Tree	0.953	0.933	0.933	0.933	0.933
Naive Bayes	0.988	0.925	0.925	0.926	0.925
Logistic Regression	0.997	0.965	0.965	0.966	0.965

K-Means Clusters become target for Supervised Model training

	User	Cluster	Silhouette	resorts	parks	theatres	museums	malls	zoo	restaurants
1	User 1	C2	0.510679	0.00	3.65	5.00	2.92	5.00	2.35	2.
2	User 2	C2	0.508844	0.00	3.65	5.00	2.92	5.00	2.64	2.
3	User 3	C2	0.508721	0.00	3.63	5.00	2.92	5.00	2.64	2.
4	User 4	C2	0.512292	0.50	3.63	5.00	2.92	5.00	2.35	2.

Supervised Model Optimization

- **Accuracy Logistic Regression can easily improve with reducing the penalizes the sum of square weightage by using Ridge Regularisation instead of Lasso.**
- **Prediction use a hidden data proved that the Tree and Naïve Bayes models does not yield as high accuracy compared to Logistic Regression.**

Logistic Regression accuracy improved with Ridge regularisation.

Test and Score (1)

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Tree	0.953	0.933	0.933	0.933	0.933
Naive Bayes	0.988	0.925	0.925	0.926	0.925
Logistic Regression	0.997	0.965	0.965	0.966	0.965

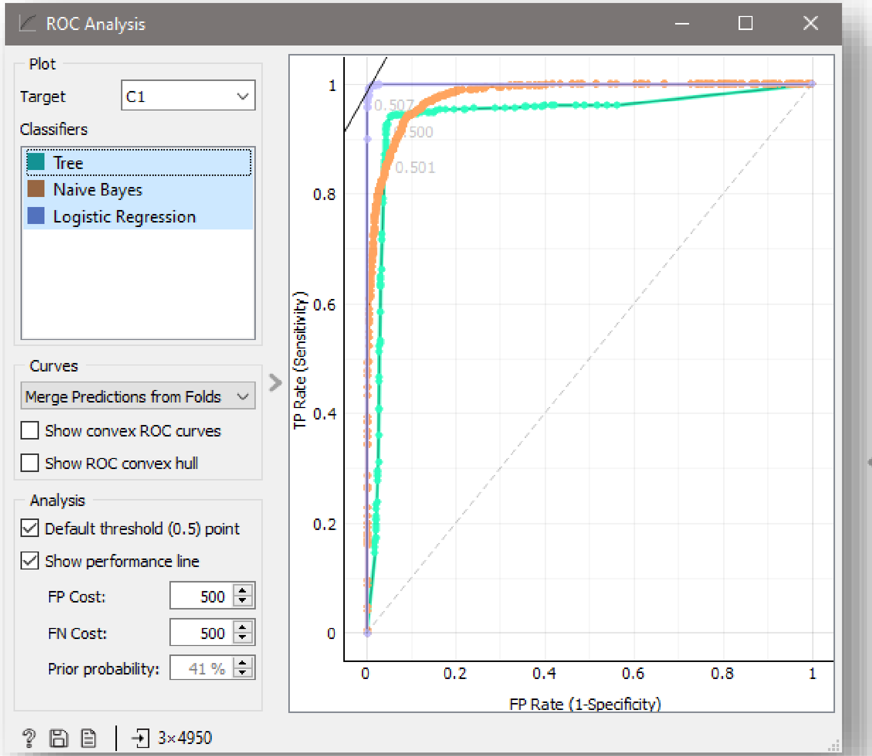
+3.6%

Test and Score (1)

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Tree	0.953	0.933	0.933	0.933	0.933
Naive Bayes	0.988	0.925	0.925	0.926	0.925
Logistic Regression	1.000	0.991	0.991	0.991	0.991

Logistic Regression has better accuracy



Predictions (1) (1)

	Logistic Regression	Tree	Naive Bayes	Cluster
30	C1	C1	C1	C1
31	C3	C3	C1	C3
32	C2	C2	C2	C2
33	C2	C2	C2	C2
34	C4	C4	C4	C4
35	C2	C1	C1	C2
36	C2	C2	C2	C2
37	C3	C3	C3	C3
38	C4	C4	C4	C4
39	C1	C1	C1	C1
40	C2	C2	C2	C2
41	C1	C1	C1	C1
42	C2	C2	C2	C2

Model	AUC	CA	F1	Precision	Recall
Logistic Regression	0.999	0.959	0.960	0.963	0.959
Tree	0.943	0.918	0.918	0.918	0.918
Naive Bayes	0.986	0.939	0.938	0.941	0.939

49 | 49 | 3×49

https://www.analyticsvidhya.com/blog/2016/01/ridge-lasso-regression-python-complete-tutorial/

Validation of Features Using Logistic Regression Odds Ratio

- **Using 95% confidence interval, odds ratio > 1.96.**
- **Select the features related to each cluster.**
- **Compare to K-Means clusters features. Result comparable with K-Means Clusters.**

Logistic Regression Model:

$$\ln(\text{odds ratio}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

Where k = number of independent variables in the model

ε_i = random error in observation i

3-Classes Logistic Regression

Odds Ratio

Feature Name	Cluster_1	Cluster_2	Cluster_3	
restaurants	11.258	0.699	0.127	City Recreation (Family, Young and energetic weekend shopping & party)
zoo	7.857	0.692	0.184	
pubs	5.647	0.554	0.320	
services	4.461	0.861	0.260	
malls	3.102	0.996	0.324	
juice bars	2.741	0.423	0.863	
burger	2.696	0.764	0.486	
art	2.481	0.495	0.815	
hotels	1.989	0.727	0.692	
museums	0.544	3.370	0.545	Amusement Park (Family high spending group)
resorts	0.458	1.486	1.469	
theatres	0.221	9.815	0.462	
beaches	0.195	4.149	1.238	
Amusement parks	0.168	7.959	0.748	
spas	0.534	0.511	3.660	Active and Healthy Leisure (Non family working group)
gyms	0.517	0.214	9.043	
bakeries	0.490	0.406	5.025	
dance	0.462	0.605	3.573	
view	0.337	1.403	2.113	
cafes	0.327	0.452	6.761	
monuments	0.242	1.671	2.474	
Botanic gardens	0.237	0.942	4.480	
swimming	0.198	0.436	11.571	
churches	0.104	0.964	9.993	

K-Means 3 Clusters

4-Classes Logistic Regression

Odds Ratio

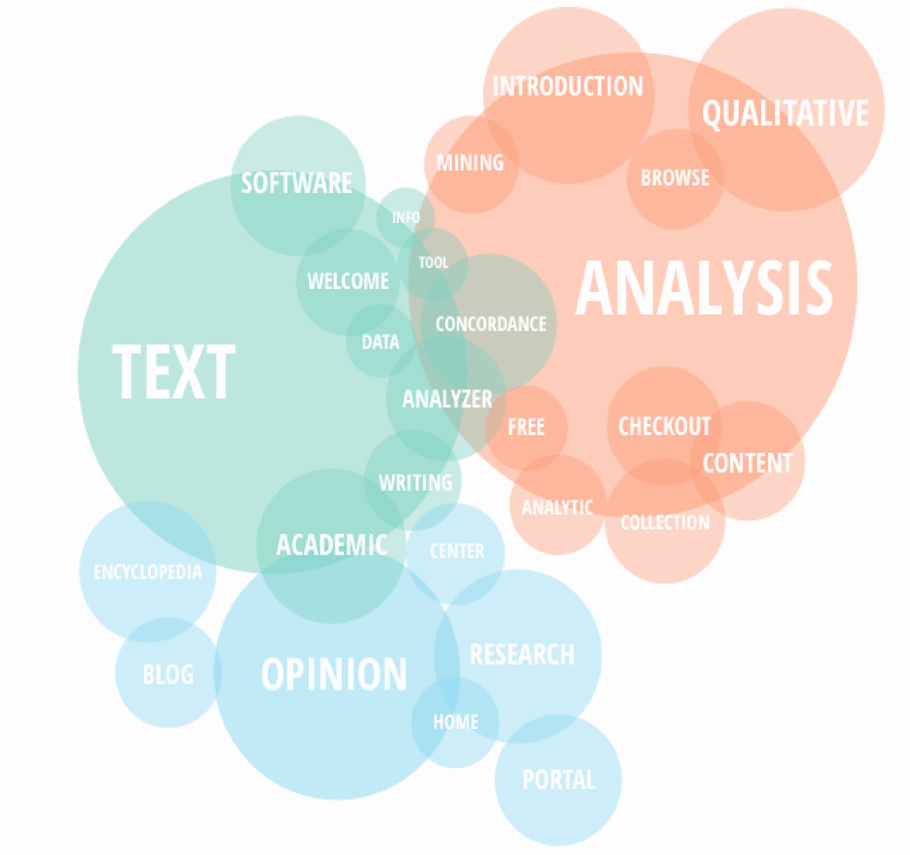
Feature Name	Cluster_1	Cluster_2	Cluster_3	Cluster_4	
monuments	2.642	0.511	0.167	1.488	Amusement Park
view	1.903	0.603	0.390	0.804	
resorts	1.784	1.043	0.320	0.520	
beaches	4.575	0.520	0.268	0.449	
Amusement parks	10.299	0.416	0.187	0.225	
theatres	9.994	0.535	0.234	-0.223	
museums	3.652	1.400	0.299	-0.424	
malls	1.025	4.080	0.708	-1.086	City Recreation
pubs	0.412	5.893	1.347	-1.185	
restaurants	0.629	10.274	0.987	-1.852	
zoo	0.902	15.066	0.518	-1.951	
bakeries	0.388	0.218	2.300	1.638	Staycation
art	0.379	0.998	3.786	-0.359	
juice bars	0.289	0.629	8.383	-0.422	
hotels	0.397	0.447	10.978	-0.669	
services	0.522	1.879	3.067	-1.101	
burger	0.350	0.515	23.062	-1.424	
churches	1.767	0.372	0.073	3.041	Active Leisure
swimming	0.476	0.259	0.480	2.828	
gyms	0.244	0.464	1.006	2.174	
cafes	0.687	0.558	0.322	2.092	
Botanic gardens	1.666	0.374	0.305	1.663	
spas	0.532	0.431	1.137	1.345	
dance	0.637	0.524	0.951	1.149	

- **As the 3 or 4 classes are possible groupings of tourists. Both will be proposed to the tourism board who has the expertise to decide which is better to improve sales and tourist experiences.**

https://www.ncbi.nlm.nih.gov/books/NBK431098/

5. Sentiment Analysis

Hotel Rating Review



Objective:

To improve customer's satisfaction on the overall travelling experiences and hence boost the tourism industry

How:

User reviews were extracted from website and the rating score is ranging from 1 to 5.

Rate 1 is the lowest score which gave us a hint that the chance of visitor not coming back again is high.

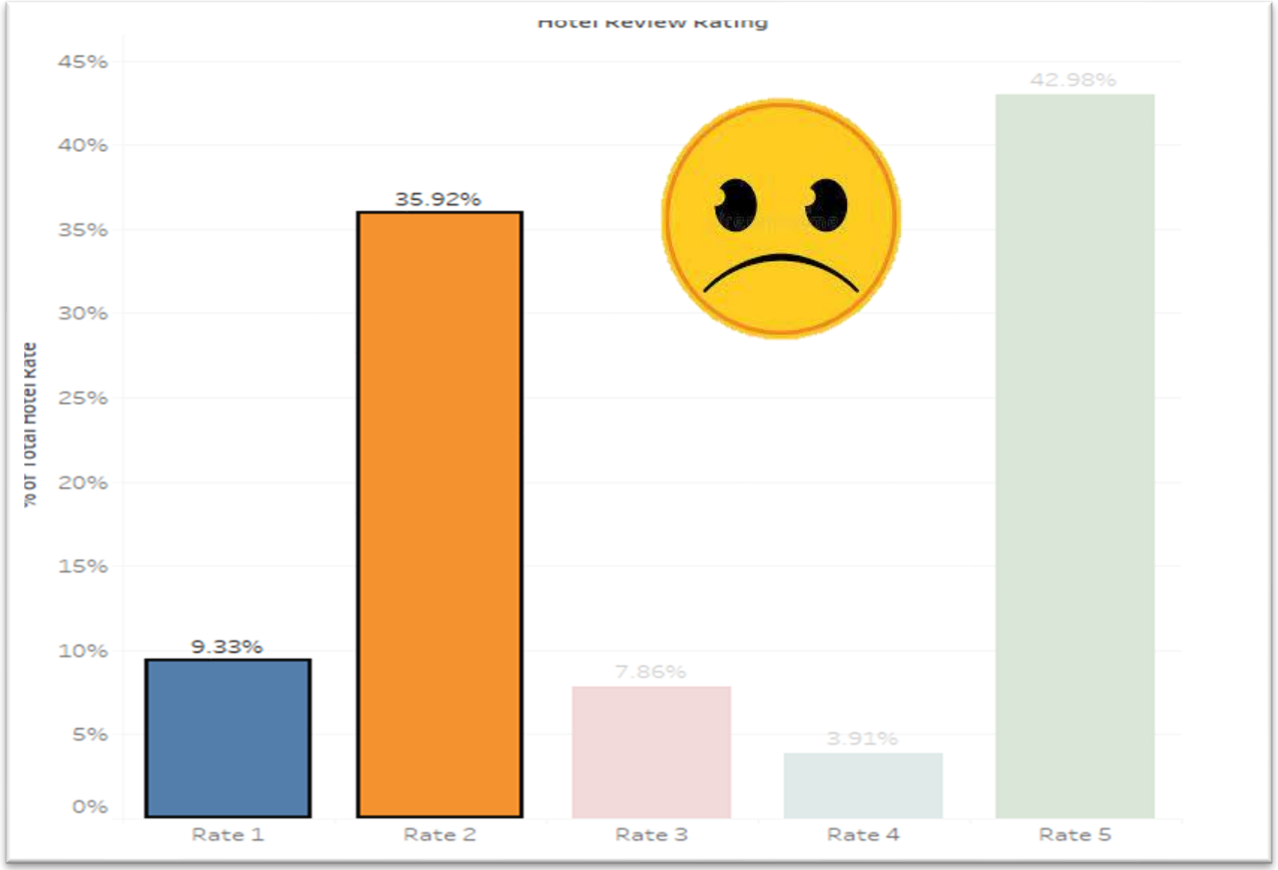
Hotel Rating Review

Total **45.25%** of hotel reviews were at Rate 1 & 2

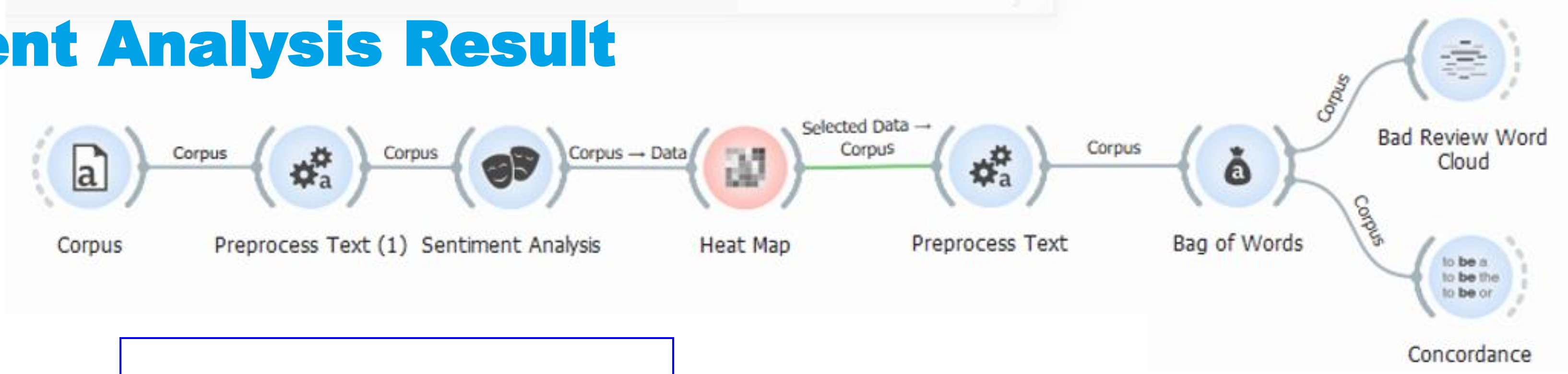
To improve customer's satisfaction, we would like to raise the rating from 1, 2 to at least Rate 3.

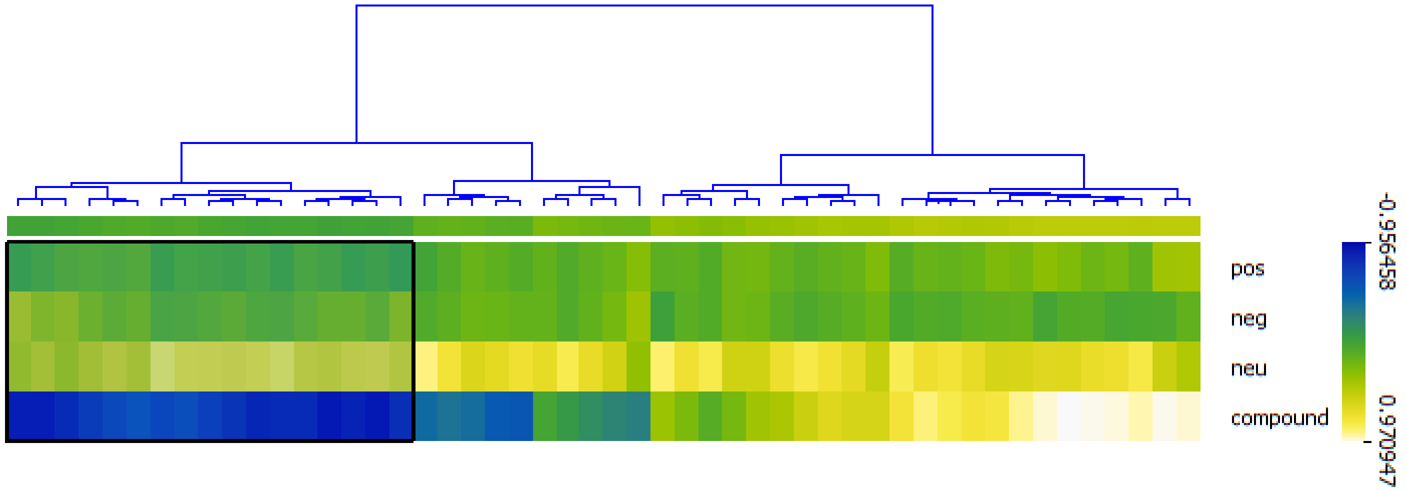
By using the Sentiment Analysis, we would like to understand what is the main most concern factor or expectation for visitor during their stay in hotel.

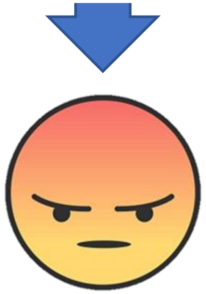
Data sampling for Sentiment Analysis will be only focus on rate 1 & rate 2 and total 500 data will be used for analysis (limitation on computing power)



Sentiment Analysis Result

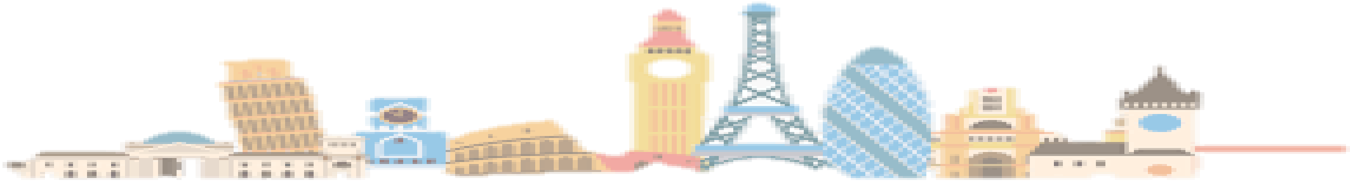






Widgets above shows the flow how we extract the user review using Sentiment Analysis.

Compound result <-0.5 will be fetched into Word Cloud to understand the negative feedback / criticism from users.



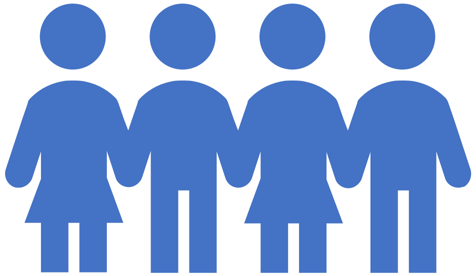
Sentiment Analysis Result

From the Word Cloud result, we can observe that most of the negative feedback are focus on *customer service* as well as the *cleanliness* of the room.

As a result, these should be the improvement areas in to order to delight customer and achieving better review score.

6. Use Case and Deployment

Mass Marketing



Niche Marketing

Use cases and deployment

Use Case 1: Improve marketing efficiency and effectiveness

- **Advertise places within clusters**
  - ➢ **E.g. Tourists like zoo also like pub**
- **Form business partnership to boost sales**
  - ➢ **E.g. 1 ticket for multiple attraction places, Zoo, pub.**

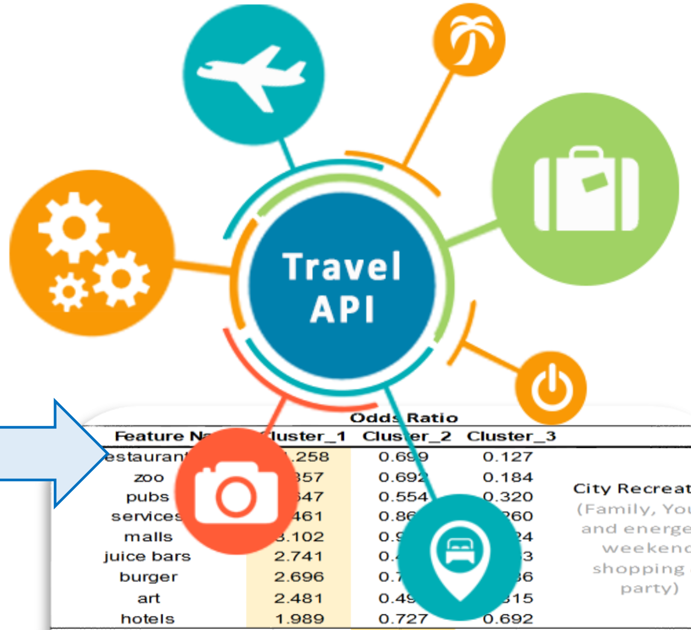
Multi-Park		Adult/Youth^	Child (Ages 3 to 12)
ParkHopper Plus Includes · Admission to Jurong Bird Park, Night Safari, River Wonders and Singapore Zoo · Tram rides at Jurong Bird Park, Night Safari and Singapore Zoo · Wild Animal Carousel ride at Singapore Zoo (child ticket only) · Amazon River Quest ride at River Wonders · Valid for 14 days from your first visit and for one-time entry to each park.		S$ 90	S$ 69
4-Park Admission - save $70!		S$ 83.60 (U.P. $88.00)	S$ 61.75 (U.P. $65.00)
2-Park Admission		S$ 50	S$ 40
2-Park Admission	Any 2 Parks	S$ 74.10 (U.P. $78.00)	S$ 52.25 (U.P. $55.00)

Use cases and deployment

Use Case 1: Improve marketing efficiency and effectiveness

- **Predict package that suit tourist's preference**
  - **Input to Model: Tourist's Profile (via Survey, spending trend)**
  - **Output from Model: Recommend Package**

Cafe

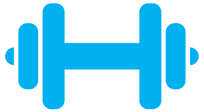


[illegible]	Odds Ratio			
Feature N[illegible]	[illegible]luster_1	Cluster_2	Cluster_3	
[illegible]estaurant[illegible]	[illegible]	0.699	0.127	City Recreation (Family, Young and energetic weekend shopping & party)
zoo	[illegible]	0.692	0.184	
pubs	[illegible]	0.554	0.320	
services	[illegible]	[illegible]	[illegible]	
malls	[illegible]	[illegible]	[illegible]	
juice bars	2.741	[illegible]	[illegible]	
burger	2.696	[illegible]	[illegible]	
art	2.481	[illegible]	[illegible]	
hotels	1.989	0.727	0.692	
museums	0.544	3.370	0.545	Amusement Park (Family high spending group)
resorts	0.458	1.486	1.469	
theatres	0.221	9.815	0.462	
beaches	0.195	4.149	1.238	
Amusement parks	0.168	7.959	0.748	
spas	0.534	0.511	3.660	Active and Healthy Leisure (Non family working group)
gyms	0.517	0.214	9.043	
bakeries	0.490	0.406	5.025	
dance	0.462	0.605	3.573	
view	0.337	1.403	2.113	
cafes	0.327	0.452	6.761	
monuments	0.242	1.671	2.474	
Botanic gardens	0.237	0.942	4.480	
swimming	0.198	0.436	11.571	
[illegible]	0.104	0.964	9.993	

Trained Model

Recommend Package

Collaborative Marketing



Tourist A in Cluster 3 already went to Dance, Bakeries and Garden.

Recommend Tourist A go to Gym.

Tourist in the same cluster also go to this attractions.

Use cases and deployment

Use Case 2: Improve hospitality service to enhance customer's experience

- **Process customers' feedback using text mining technique to identify problems**
  - **E.g. poor customer services and house keeping**
- **Develop plan to address the problems**
  - **Retrain employees to improve skills**
  - **Increase employee engagement**
  - **Hire professions with the right skillsets**
- **Implement plan and process new feedback**
- **Measure performance, learn and refine plan**


THANK YOU