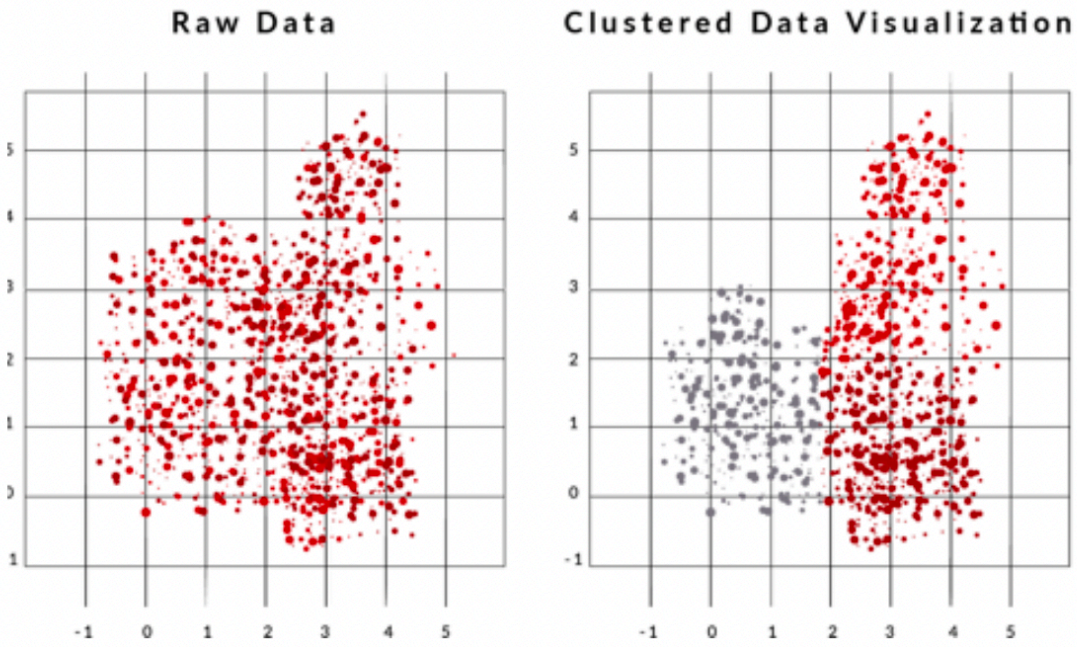
FEB 2021



CLOUD KITCHEN BRANDS CLUSTERING

DATA SCIENCE PROJECT

BRAND PERSONA

BY: RACHID ELKHAYAT

OVERVIEW:

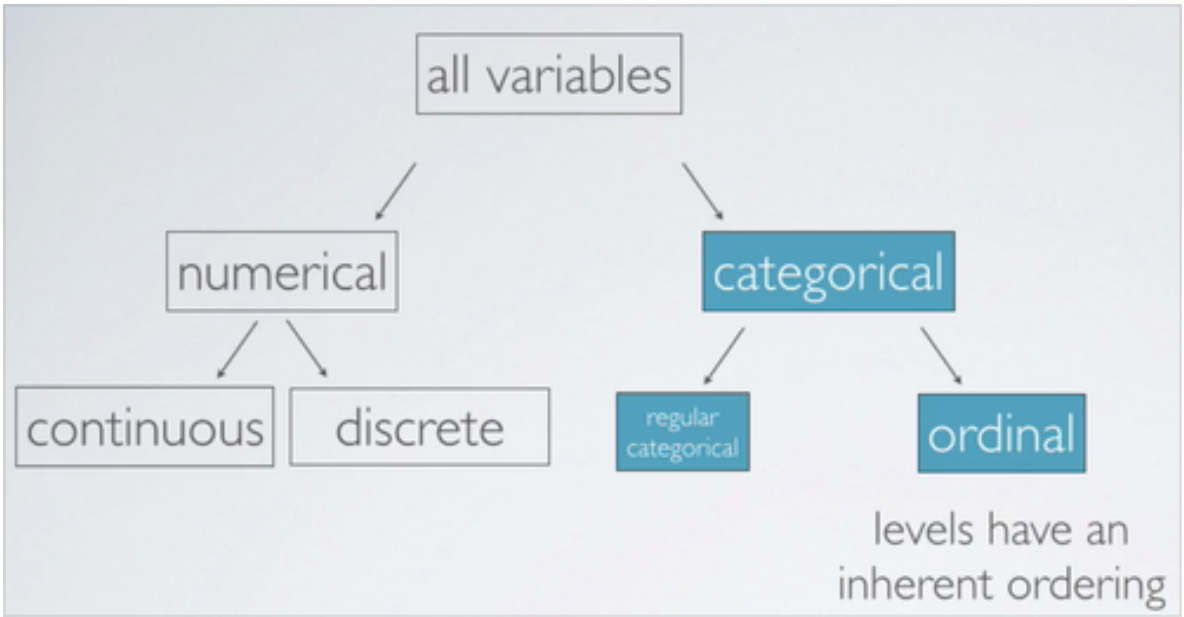
In this project we are creating groups of a set of brands that we call a cluster, that will help understand the similarities and the differences using unsupervised machine learning method of identifying and grouping similar data points in larger datasets without concern for the specific outcome.

Steps implemented:

- Finding the features
- Collecting and cleaning the data
- Feature engineering
- Encoding
- Selecting the number of Clusters(elbow method)
- Running Clustering algorithm(K-means)
- Visualization and evaluation

THE FEATURES:

In general, there are two types of features that we will be dealing with, and each one of them can be classified into two categories.



Since K-MEAN model input has to be numeric, we need to encode the data that we have using different encoding methods based on the type of feature we have.

```python
from sklearn.preprocessing import OneHotEncoder

# creating instance of one-hot-encoder
enc = OneHotEncoder(handle_unknown='ignore')

# label encoded values
enc_df = pd.DataFrame(enc.fit_transform(df1[['Cuisine']]).toarray())
# merge with main df on key values
df1 = df1.join(enc_df)
```

Features engineering is a key process in our project, as the feature will be the deciding factor of how the brands will be grouped together. In order to get the right features into place we use Feature-Engineering which is the Science of extracting more information from existing data, to help the Machine Learning algorithm to understand and work accordingly.

Features source:
The features has be gathered and put together from different sources in within the company, the product manager managed to fill features related to their experience with the brands. and on the other we leveraged the data that we have in our database warehouse to obtain useful features using SQL scripts as showing below:

```sql
-- Average price -----------------------------------------------
select
 POSITION_BRAND_NAME
,round(Median(median_price)) as median_price

from

(select
       POSITION_BRAND_NAME
      ,POSITION_NAME
      ,case
       when country_name ='Saudi Arabia'          then NVL(P_MENU_PRICE_B_TAX,0)/3.75
       when country_name ='Kuwait'                then NVL(P_MENU_PRICE_B_TAX,0)* 3.29
       when country_name ='United Kingdom'        then NVL(P_MENU_PRICE_B_TAX,0)/1.25
       when country_name ='United Arab Emirates'  then NVL(P_MENU_PRICE_B_TAX,0)/3.67
       End  as Median_price


from "KAFKA_DB"."KAFKA_BI_ANALYTICS_SCHEMA"."FACT_POSITION_DETAILS"
where POSITION_BRAND_NAME not like [redacted]
AND  date(ORDER_POSITION_CREATED_AT) >= '2020-11-01'
AND P_MENU_PRICE_B_TAX >5
group by 1,2,3
)
group by 1
```

Below is the list of features that were considered in this project, eight of which where taking into consideration as an input for the machine learning algorithm.

Features						
Brand age	New	young	mature	old		
Country Of Origin	Lebanon	GCC	India	Italy	China	
Health oriented	yes	No				
Social media engagement	Low	Moderate	High			
breakfast	Yes	No				
Occasional	yes	No				
Peak Time	Morning	AFTERNOON	EVENING			
Brand ownership	Standalone	Group				
International presence	Yes	No				
Cuisine	Arabic	Asian	American	Indian	International	Turkish
Main ingridient	Beef	Chicken	Seafood	Pastry	Plant Based	others
Kitopi regions (UKS)	UKS	UK	US	U	S	K
Number of Menu Items	< 20	20-40	40+			
Avg main-course price	$	$$	$$$			

Below are the set of features that we found it fit to describe the personality of the brands by helping to differentiate the clusters in the best way possible. Each Feature contains different categories where the dominance of which will identify the personalities of the brand as well as the clusters.

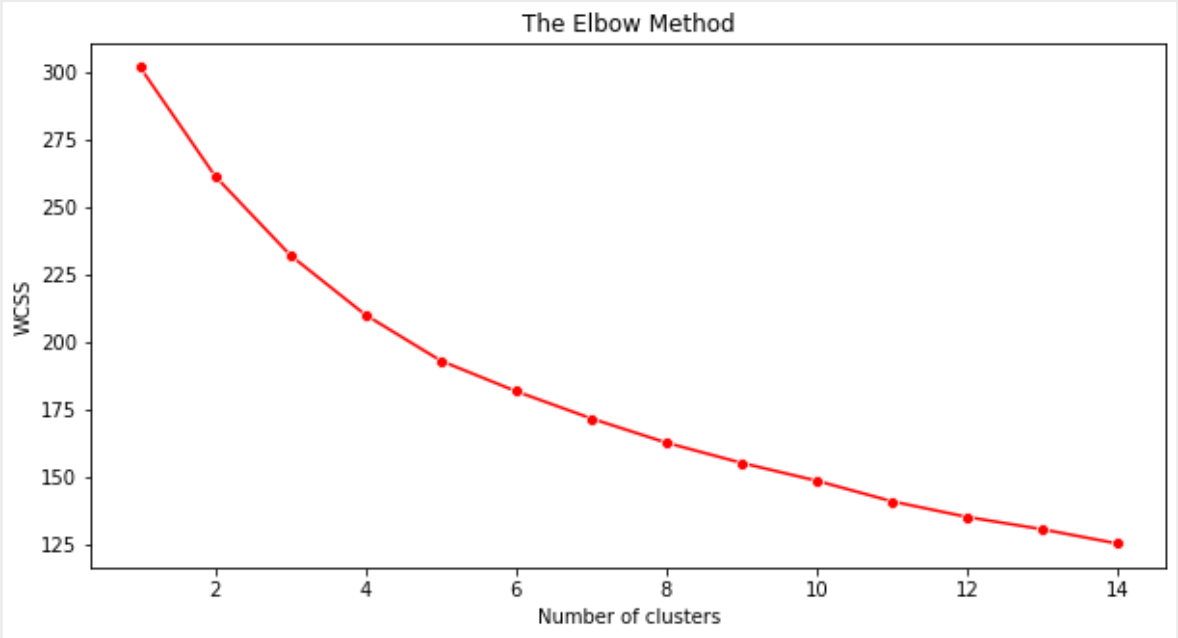
Brand Name	Avg item price	Number of Menu Items	Health oriented	International	breakfast	Occasional	Brand ownership	Cuisine	peak	M	AF	EV
ORG	HIGH	LOW	Y	N	N	N	Standalone	INTERNATIONAL	AF	0	1	0
SUS	HIGH	LOW	N	N	N	N	Group	ASIAN	EV	0	0	1
LOV	MEDIUM	LOW	Y	Y	N	N	Group	DESERT	EV	0	0	1
ALL	MEDIUM	LOW	Y	Y	Y	N	Group	DESERT	EV	0	0	1
POP	MEDIUM	LOW	Y	N	N	N	Group	DESERT	AFEV	0	1	1
FRY	MEDIUM	LOW	N	N	N	N	Group	AMERICAN	AFEV	0	1	1
PEP	MEDIUM	LOW	Y	Y	N	N	Group	HEALTHY	AF	0	1	0
POK	HIGH	LOW	Y	N	N	N	Group	ASIAN	AF	0	1	0
MIN	MEDIUM	LOW	N	Y	Y	N	Group	AMERICAN	EV	0	0	1
DON	MEDIUM	LOW	N	N	Y	N	Standalone	ARABIC	EV	0	0	1
WO	LOW	LOW	N	N	N	N	Group	ASIAN	AF	0	1	0
HAA	MEDIUM	LOW	N	Y	N	N	Standalone	DESERT	EV	0	0	1
PAS	LOW	LOW	N	Y	N	N	Group	ITALIAN	EV	0	0	1
BIR	MEDIUM	LOW	N	Y	N	N	Group	INDIAN	AF	0	1	0
NAT	LOW	LOW	N	Y	N	N	Standalone	AMERICAN	AFEV	0	1	1
LOT	HIGH	LOW	N	N	N	N	Group	DESERT	AFEV	0	1	1
OG	LOW	LOW	N	Y	N	N	Group	INTERNATIONAL	EV	0	0	1
HEA	MEDIUM	LOW	Y	N	N	N	Group	DESERT	AFEV	0	1	1
PIZ	HIGH	LOW	N	Y	N	N	Standalone	ITALIAN	EV	0	0	1
NKD	MEDIUM	LOW	N	Y	N	N	Standalone	ITALIAN	EV	0	0	1
PAP	MEDIUM	LOW	N	Y	N	N	Standalone	ITALIAN	EV	0	0	1
POP	LOW	LOW	N	Y	N	N	Standalone	DESERT	EV	0	0	1
GEN	MEDIUM	LOW	Y	Y	Y	N	Group	DESERT	AFEV	0	1	1
BAR	MEDIUM	LOW	Y	N	N	N	Group	HEALTHY	AF	0	1	0
POK	MEDIUM	LOW	Y	N	N	N	Group	ASIAN	AF	0	1	0
PAS	HIGH	LOW	N	N	N	N	Standalone	DESERT	AFEV	0	1	1
ULT	MEDIUM	LOW	N	Y	N	N	Group	INTERNATIONAL	AFEV	0	1	1
PRO	MEDIUM	LOW	N	Y	N	N	Group	ITALIAN	EV	0	0	1
NOB	LOW	LOW	N	Y	N	N	Group	AMERICAN	EV	0	0	1
OLD	MEDIUM	LOW	N	N	N	N	Standalone	MEXICAN	AFEV	0	1	1
GYM	LOW	MEDIUM	Y	Y	Y	N	Group	HEALTHY	AF	0	1	0
BIR	MEDIUM	MEDIUM	N	Y	N	N	Standalone	INDIAN	AF	0	1	0
LAL	MEDIUM	MEDIUM	N	N	N	N	Group	AMERICAN	AFEV	0	1	1

THE MODEL:

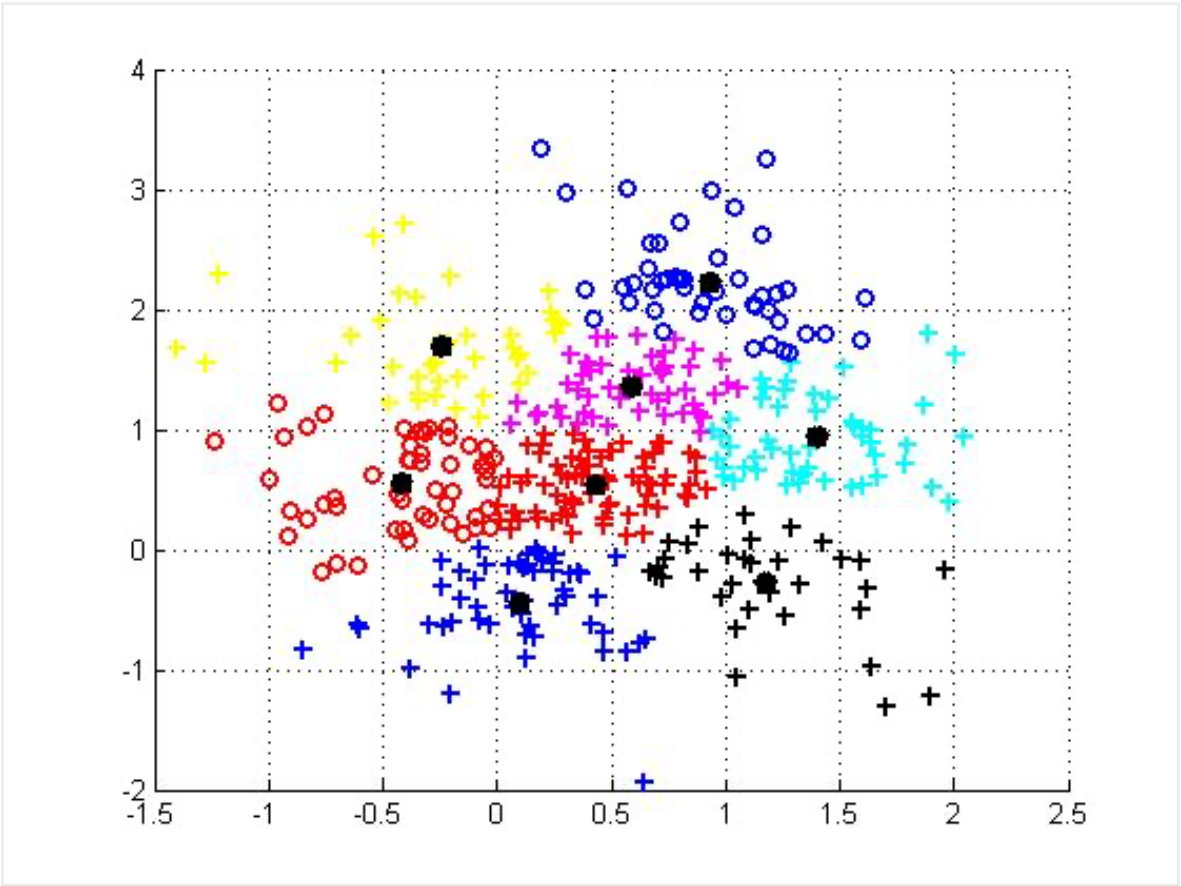
```
kmeans = KMeans(n_clusters = 5, init = 'k-means++', random_state = 42,
kmeans.fit(df1)
y_kmeans = kmeans.fit_predict(df1)
y_kmeans
```

We have applied **K-means clustering** algorithm to find groups which have not been explicitly labeled in the data. This can be **used** to confirm business assumptions about what types of groups exist or to identify unknown groups in complex data sets.

Using the elbow method, we can deduce that five is the optimum number of clusters to be considered in our project



Each cluster has its own properties and it differ from the fact that some clusters might have many common aspect.



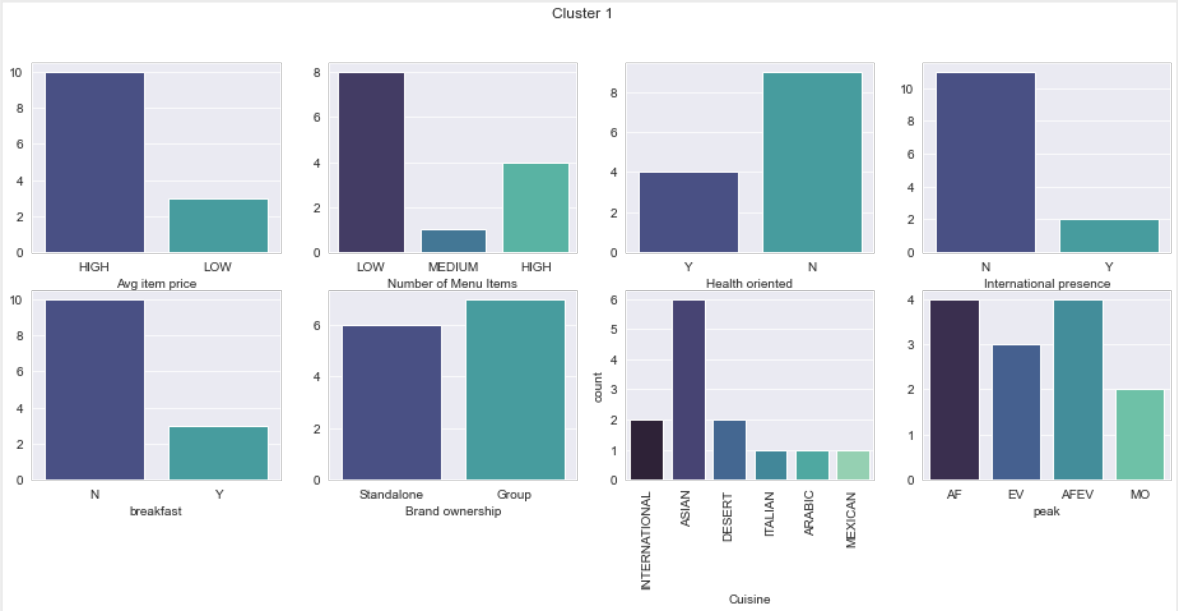
FRESH RESTUARANT	PASTEL res	RS CHINESE	SHAWARMA AND
SUSHI SPECIAL	LITTLE YOYO	CAKE ONE	FISH AND FISH
POKE RES	MANMAN	PIZZA MONALISA	HK POKE
			SUSHI SUSHI

The clusters can be intersecting at some points , however the group of brands creating one cluster are the ones that are closer to the centroids with the least sum square of error.

THE CLUSTERS:

Cluster 1

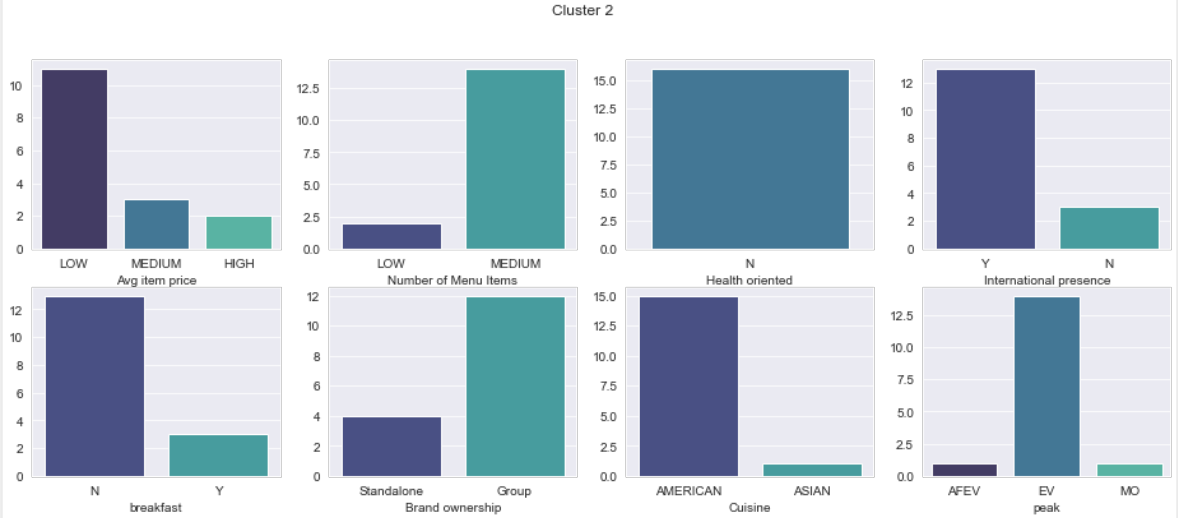
As we can see from the chart of the average item price, we can say that this cluster has many brands with the price at the higher end, more than 40 AED, they mostly have low number of menu items, where by the heath orientation is not a primary option, Asian cuisine is dominance in this group.



Cluster 2 — Fast Food brands

Economic options with a price range between 20 and 40 AED, it has variety of menu item that can reach over 40 items per menu. clearly this cluster is not health oriented and it focus on mass selling as they have regional/international presence. most of the bands are part of a group and the pro-dominate cuisine is American and the most of the sales occurs at the launch time.

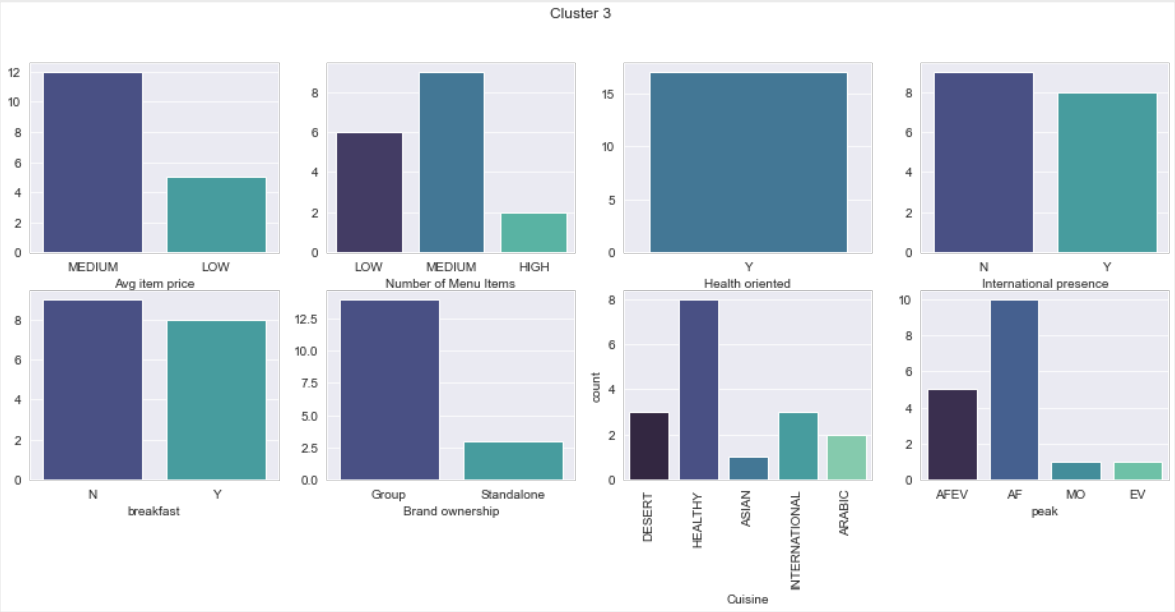
SUCHI ROLLS & CO	XYSLIDERS	GGBURGER	BURGERS & FRIES
RIBS FAMOUS	SPICY CHICKEN	FRUIT GREEN	BURGER NEW
POLUN CHICKEN	SPECIAL SLIDERS	SUPER BURGERS	MEAT BALLS
DELICIOUS COOKIES	BURGER 99	GREE CURRY	POPPOP



Cluster 3 - Healthy brands

The price is in the acceptable range of the average buyer, between 20 and 40 dirhams. the most common feature of the brand is the orientation towards healthy meals and serving three meals per day that include breakfast. mostly the brands are owned by groups.

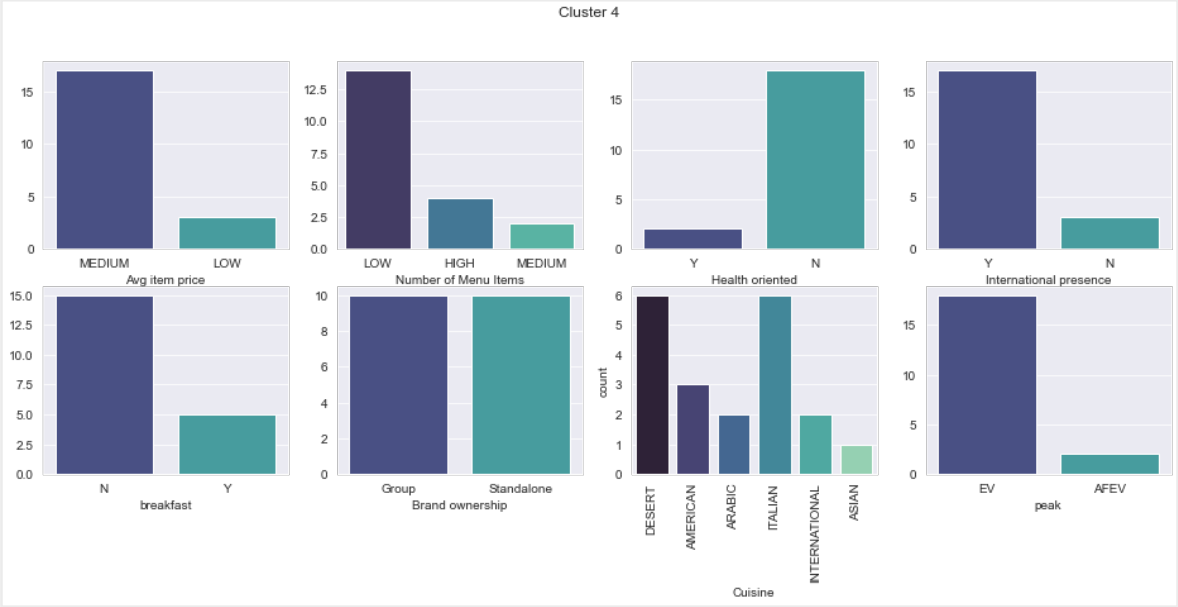
SUCHI ROLLS & CO	XYSLIDERS	GGBURGER	BURGERS & FRIES
RIBS FAMOUS	SPICY CHICKEN	FRUIT GREEN	BURGER NEW
POLUN CHICKEN	SPECIAL SLIDERS	SUPER BURGERS	MEAT BALLS
DELICIOUS COOKIES	BURGER 99	GREE CURRY	POPPOP
SALAD SPECIAL			



Cluster 4

With the dominance of Italian and desert cuisines the group of the brands are on the side of more calories than healthy, most the sales occures in the evening and the price of the items is considered relatively low.

FRESH RESTUARANT	PASTEL res	RS CHINESE	SHAWARMA AND
SUSHI SPECIAL	LITTLE YOYO	CAKE ONE	FISH AND FISH
POKE RES	MANMAN	PIZZA MONALISA	HK POKE
CAKE PLUS	VVV PIZZA	BURGER TOP	SUSHI SUSHI

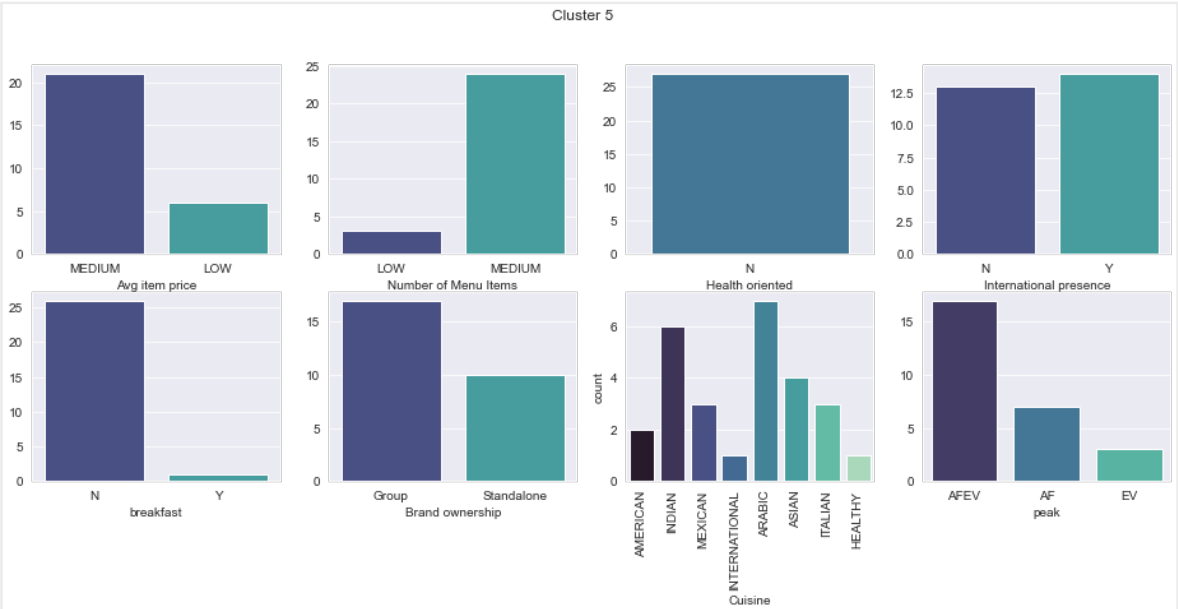


Cluster5

Most arabic and indian brands falls in this category with a medium priced items range between 20 and 40 dirhams, the indications are not leaning toward healthy options however the brands has a good international presence as well. Most of the brands don't cater for breakfast.

FRESH RESTUARANT	PASTEL res	RS CHINESE	SHAWARMA AND
SUSHI SPECIAL	LITTLE YOYO	CAKE ONE	FISH AND FISH
SUCHI ROLLS & CO	XYSLIDERS	GGBURGER	BURGERS & FRIES
RIBS FAMOUS	SPICY CHICKEN	FRUIT GREEN	BURGER NEW
POLUN CHICKEN	SPECIAL SLIDERS	SUPER BURGERS	MEAT BALLS
DELICIOUS COOKIES	BURGER 99	GREE CURRY	POPPOP
	PIZZA MORE	WOK NOW	CHINESE BEST

∑



CONCLUSION:

In conclusion, The clustering model has been successful in finding the similarities and differences in each of the groups, as we can see there are some categories with significant effects that are shaping the groups persona.
This MVP is the first mile stone for many projects that can be utilized by different teams to compare, assess and analyze.
The goal of this project is not to find the clusters or group the brands, the purpose will only be achieved when this data become of use in different projects, i.e comparing the different clusters with the features that are related to like (retention , sales, discounts, revenue, profit,..)