

情報可視化ソフトウェア Hidden によるデータ分析

ピザ百名店について

1.はじめに

私は食べることが大好きで日頃から食べログなどのグルメサイトをよく利用するので、今回は食べログ百名店を分析対象にすることにした。中でも私はピザが好きなので今回はピザ百名店を取り上げた。

< 食べログについて >

参考文献①によると、食べログは月間約 8,998 万人以上が利用する日本最大級のグルメサイトであり、独自のランキングやユーザーの口コミ・写真をもとにお店探しから予約までを簡単に行うことができるものである。

< 百名店について >

参考文献②によると、百名店とは食べログユーザーから高い評価を集めた 100 店を、さまざまなジャンルごとに発表するグルメアワードであり、現時点で 29 ジャンルが掲載されている。

2.作成したデータについて

以下に述べる A から O までの 15 項目を 100 店舗分調査し、100 行 15 列のデータを作成した。

< 項目一覧 >

- A ピザの種類数
- B マルゲリータの価格(税込)
- C 総合評価
- D 料理・味
- E サービス
- F 雰囲気
- G 酒・ドリンク
- H コストパフォーマンス
- I 口コミ数
- J 保存数
- K 投稿写真数
- L 百名店評価

M オープン日からの経過年数(満)

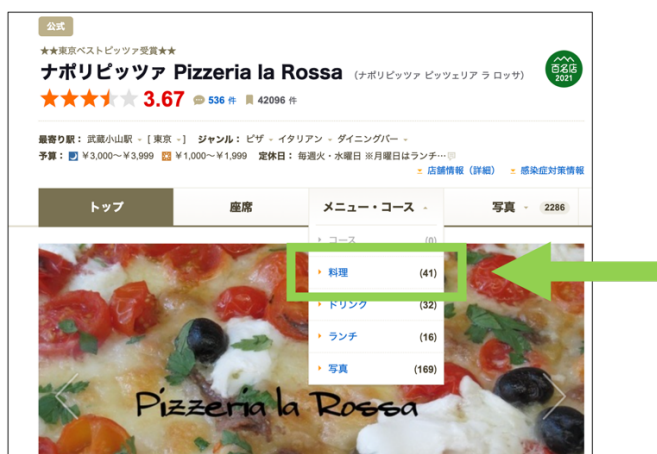
N 所属地域

O 受賞歴

< 各項目の詳細 >

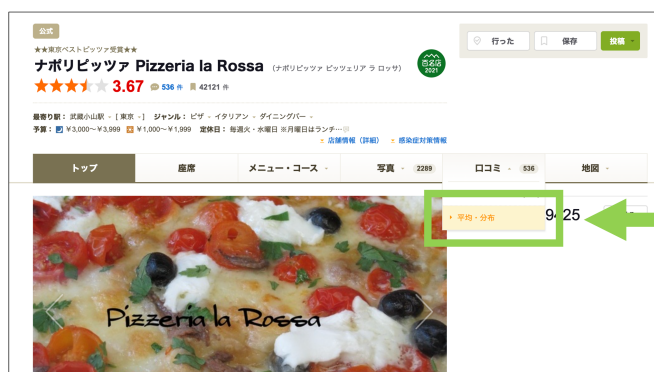
・ A ピザの種類数 と B マルゲリータの価格(税込)

それぞれのお店のページの「メニュー・コース」→「料理」を選択し、ピザの種類数とマルゲリータの税込価格を調べた。ほとんどのお店でマルゲリータの記載があったが、数店舗ほどメニューにないお店があったため、その場合は他のシンプルなピザの値段を代用した。(以下の図は参考文献③より)



・ C 総合評価 ～ H コストパフォーマンス

総合評価からコストパフォーマンスまではそれぞれのお店のページの「口コミ」→「平均・分布」を選択しところにある、右下の図の評価平均の点数を上から順にピックアップしたものである。この点数の算出方法については参考文献④に詳しい説明があり、ユーザー影響度や口コミ数を考慮し独自の算出方法が取られている。点数の見方としては 4.00 以上が全体の TOP500 前後でかなりの高評価、3.50 以上が全体の約 3%で高い評価となっている。(以下の図は共に参考文献③より)



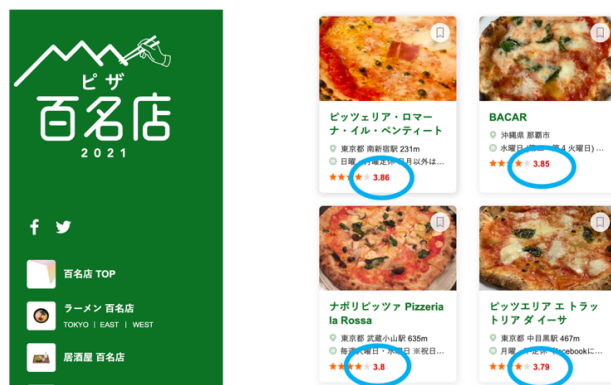
・I 口コミ数 と J 保存数 と K 投稿写真数

口コミ数と保存数と投稿写真数は以下の図(参考文献③より)の青丸の部分からピックアップした。



・L 百名店評価

参考文献⑤中の選出店一覧で見ることができる点数(青丸で囲んだ部分)を百名店評価とした。(以下の図は参考文献⑤より)



・M オープン日からの経過年数(満)

オープン日からの経過年数(満)はそれぞれのお店のページの店舗情報のところに「オープン日」の記載があるため、そこから現在までの満年数を計算して求めた。

・N 所属地域

所属地域を以下の通りに数値化した。数字は各地域に属する店舗数が少ない順に割り振った。東京の店舗数が圧倒的に多かったため東京だけは関東地方とは分けて扱うことにした。

- 中国・四国地方…1
- 北海道・東北地方…2
- 九州・沖縄地方…3
- 中部地方…4
- 近畿地方…5

関東地方(東京を除く)…6

東京…7

・O 受賞歴

受賞歴はそれぞれのお店のページの店舗情報のところに「受賞・選出歴」の記載があるため、百名店を受賞した回数を利用することにした。ちなみに以下の例では4という数字が入ることになる。(図は参考文献③より)

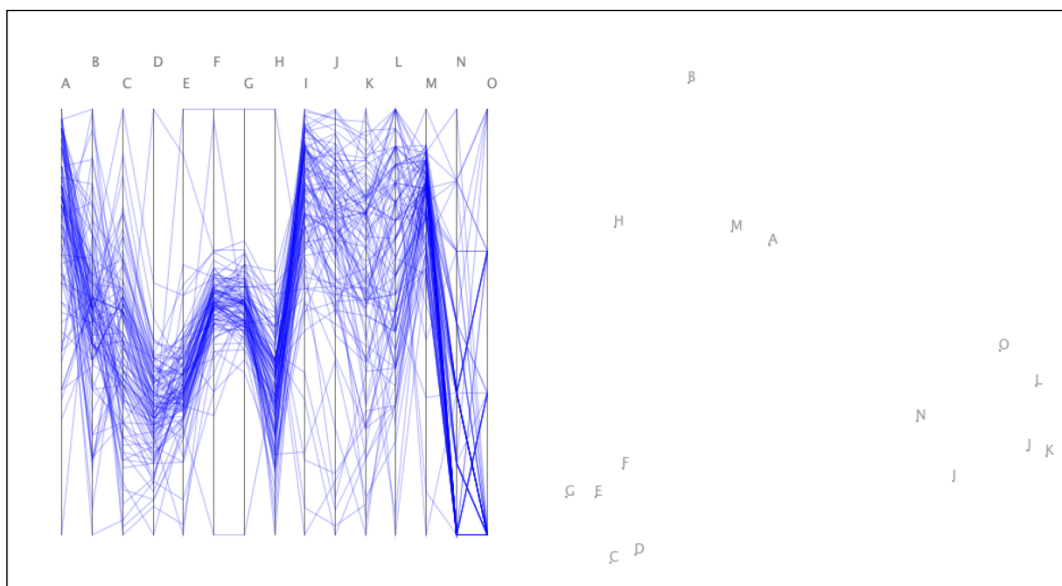
店舗情報 (詳細)	
店舗基本情報	
店名	ナポリピッツァ Pizzeria la Rossa (ナポリピッツァ ピッツェリア ラ ロッサ)
受賞・選出歴	   

< 作成した CSV ファイル(の一部) >

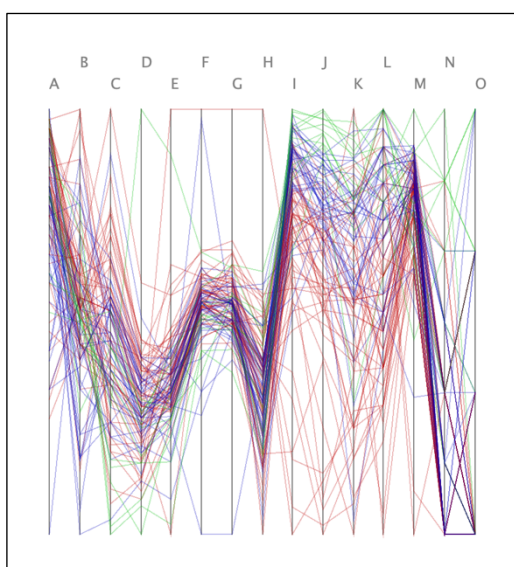
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
2	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
3		13	1804	3.81	4.05	3.08	3.43	3.34	3.31	229	55724	1307	3.86	24	7
4		2	1620	3.93	4.09	3.66	3.82	3.6	3.57	514	36658	2458	3.85	14	3
5		4	1430	3.77	3.91	4.1	3.4	3.45	3.77	1159	66836	4110	3.81	21	6
6		17	1408	3.81	3.88	3.55	3.57	3.47	3.71	403	41496	2258	3.8	11	7
7		14	1650	3.51	3.68	3.04	3.27	3.01	3.18	1328	106884	4184	3.79	12	7
8		2	1650	3.63	3.83	2.95	3.42	3.08	3.22	1412	95949	3264	3.79	15	7
9		5	1580	3.84	3.99	3.49	3.61	3.36	3.44	724	52418	3434	3.79	24	7
10		7	2355	3.62	3.73	3.4	3.52	3.29	3.38	999	108942	3727	3.79	10	7
11		3	1260	3.7	3.87	3.41	3.52	3.38	3.44	786	54360	2748	3.78	21	7
12		31	1600	3.76	3.88	3.5	3.59	3.39	3.45	621	47981	3738	3.78	11	7
13		25	1650	3.61	3.75	3.22	3.57	3.23	3.19	1184	94007	3386	3.77	23	7
14		3	1500	3.66	3.76	3.22	3.32	3.23	3.77	701	57589	2614	3.77	15	7
15		2	1575	3.8	3.98	3.28	3.46	3.23	3.62	561	53565	1637	3.77	16	7
16		13	1500	3.73	3.83	3.51	3.6	3.47	3.69	494	45666	2874	3.77	8	7
17		11	1650	3.73	3.93	3.25	3.48	3.27	3.25	510	47097	1850	3.77	21	7
18		21	1800	3.78	3.86	3.48	3.5	3.36	3.55	329	37269	1643	3.77	12	7
19		5	1728	3.79	3.86	3.49	3.62	3.36	3.82	734	36591	4334	3.76	13	4
20		23	1600	3.69	3.77	3.43	3.52	3.32	3.62	419	47402	2069	3.76	12	7

3.可視化の結果と分析

まずはクラスタリングを1としてそのまま結果を表示すると以下のようなった。



このままでは見づらいためクラスタリングを2,3,4…と変更してみると3のときが比較的に見やすいことがわかった。(左下はクラスタリング=3のときのPCPの折れ線)

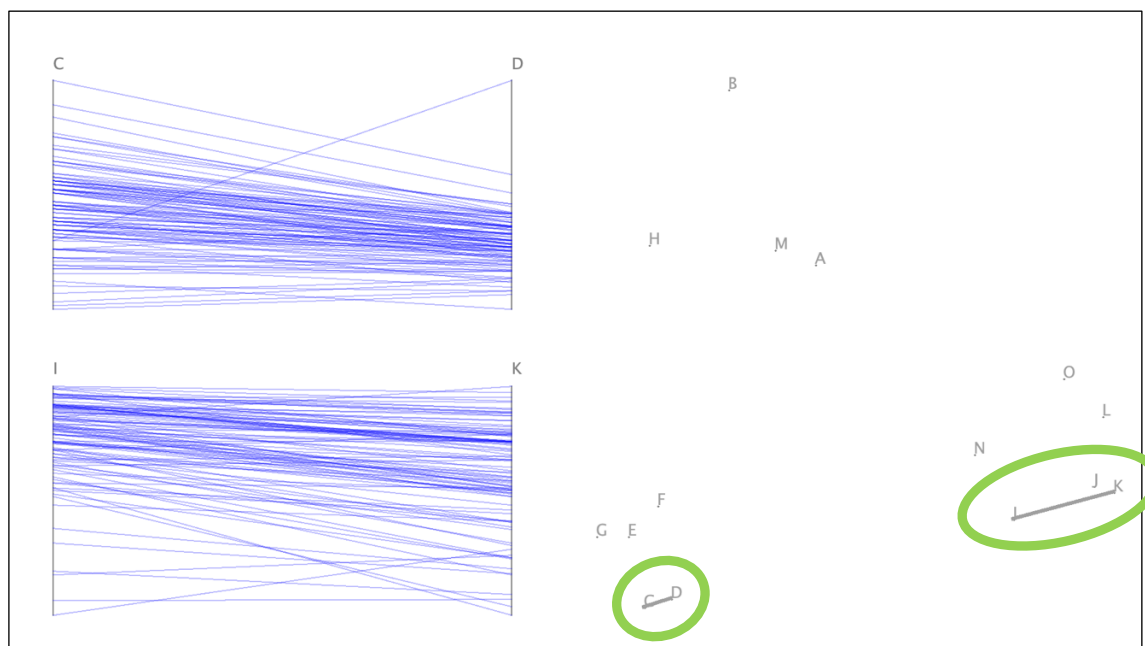


I, J, K, Lの部分に着目すると色のグラデーションが綺麗に見えていることに気が付く。これは折れ線の交差が少なく正の相関が見られるためであると考えられる。詳しくは後ほど分析する。

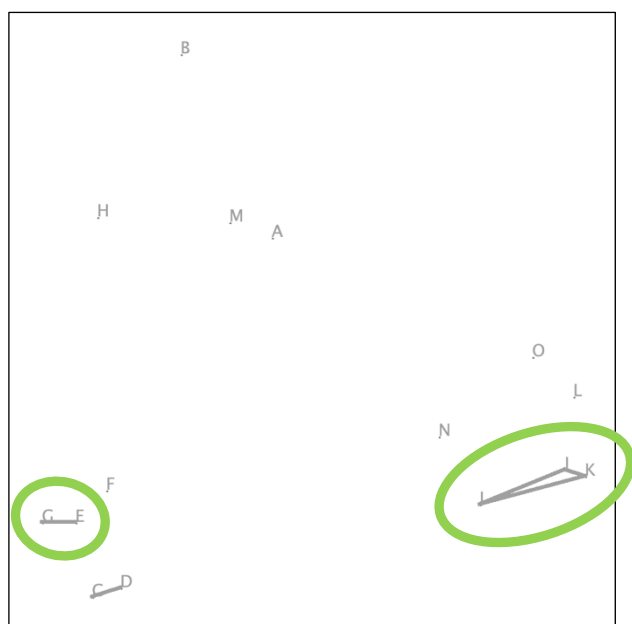
← 以降この図を“折れ線(1)”と呼ぶ

以下では右上のスライダーを操作して相関の強いデータを順に取り上げて分析していく。右側の散布図を見ていくとまず始めにCとD、IとKがそれぞれ線分で結合された。つまりこれらには最も強い相関があるということがわかる。続いて左側の折れ線を確認するとCとD、IとKの間にはともに正の相関があることがわかる。Cの総合評価とDの料理・味はどちらも食べログユーザーによる評価であるため、料理の味に満足したユーザーが総合評価を高くすることは容易に想像がつく。またIの口コミ数とKの投稿写真数について

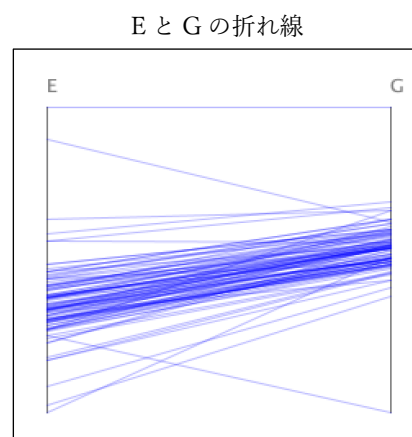
も、同一人物が口コミと同時に写真を投稿することが多いと推測されるためこれらの間に強い正の相関があることはすぐに納得できる。



さらにスライダーを上昇させていくと次はEとG、そしてIとJ、JとKがそれぞれ線分で結合され右下にIとJとKの三角形が現れた。

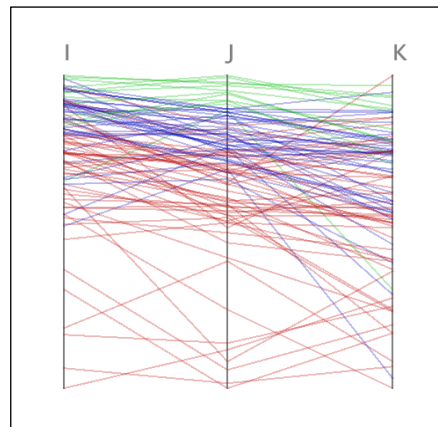


EとGの相関をPCPの折れ線で確認してみると正の相関が見られた。EのサービスとGの酒・ドリンクに正の相関があることは不思議ではないが、これら2つの項目がまず始めに結合されたことは面白い事実だと感じた。



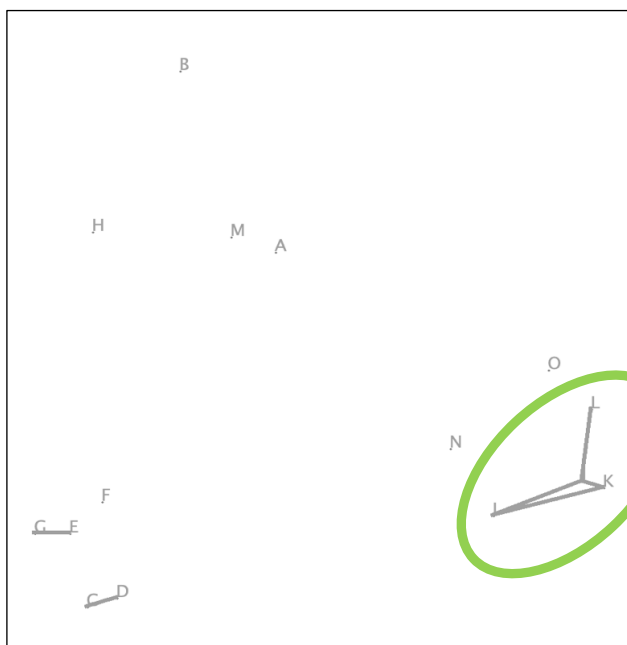
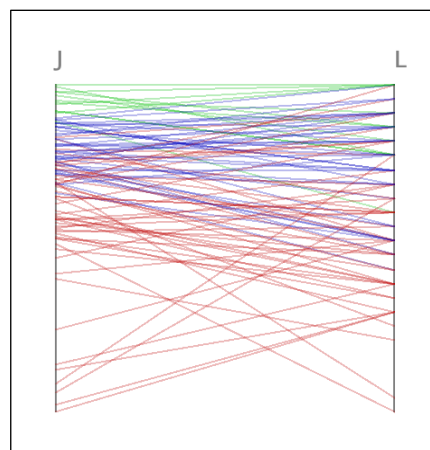
続いてIとJとKの相関を折れ線で確認してみるとIとJ、JとKの間にも正の相関が見られた。これらの折れ線はクラスタリングを3に設定した場合が最も見やすかったためその結果を右に表示する。先ほどIとKの間に正の相関が見られたことから、Iの口コミ数とJの保存数とKの投稿写真数の3項目は正の相関の強さで結びついていることがわかった。しかしよく見てみると交差している線も複数見られることからJ保存数は多いがK投稿写真数は少ないといった例外もあるということがわかった。

IとJとKの折れ線 (クラスタリング=3)

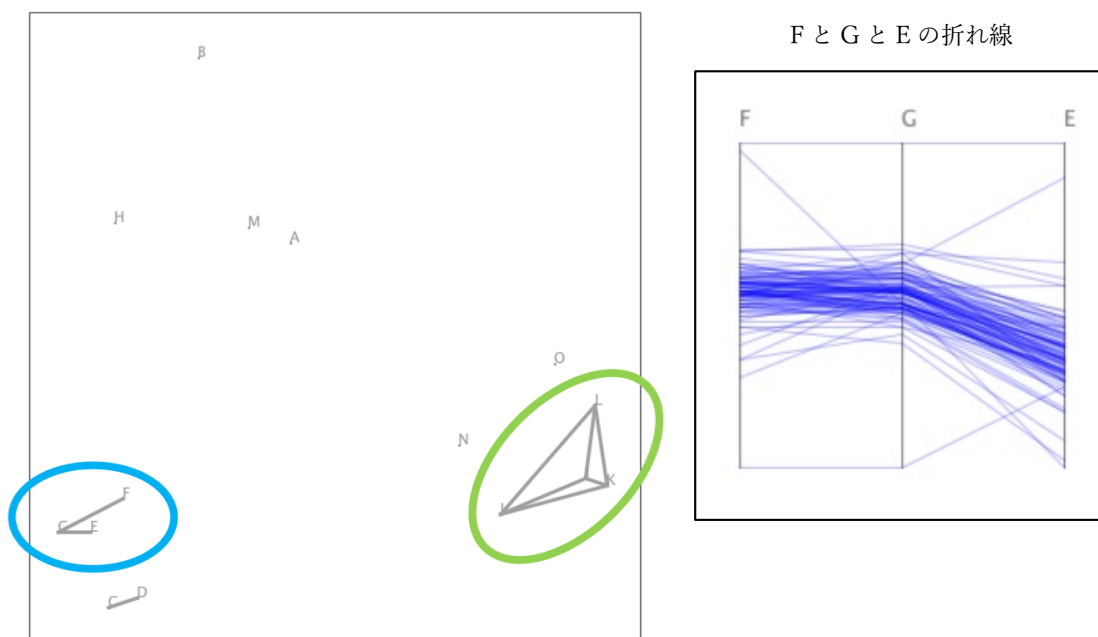


さらに右上のスライダーを上昇させていくと今度はIとLが線分で結ばれ、IとJとKとLが一つのグループになった。PCPの折れ線を確認すると、JとLには正の相関があることがわかった。つまりこのグループは正の相関の強さで結びついていることがわかる。そのためp.5の折れ線(1)でもI, J, K, Lの部分のグラデーションが綺麗に見えたのだと思う。Lは百名店評価であり、参考文献②によるとこれは食べログユーザーによる評価、つまりC総合評価～Hコストパフォーマンスに基づいていることがわかった。C～Hはユーザ影響度や口コミ数を考慮して算出されているため、Iの口コミ数とLの百名店評価に正の相関が見られるのだと考えた。

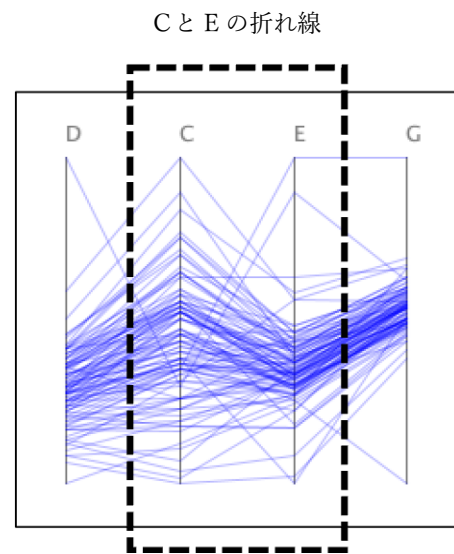
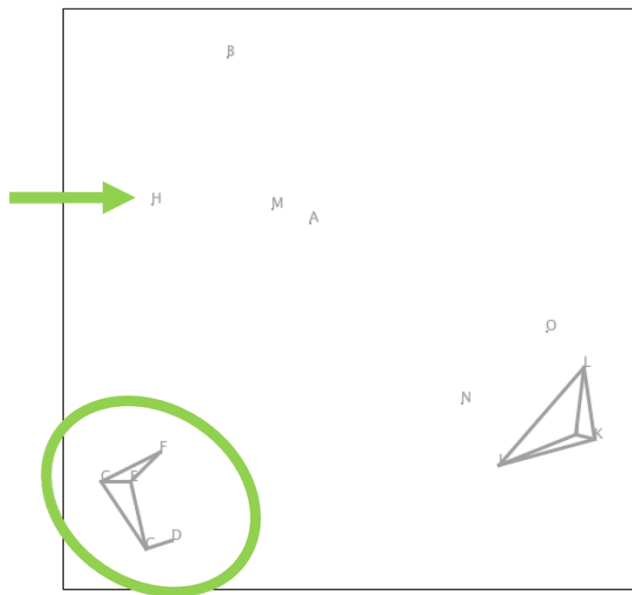
JとLの折れ線 (クラスタリング=3)



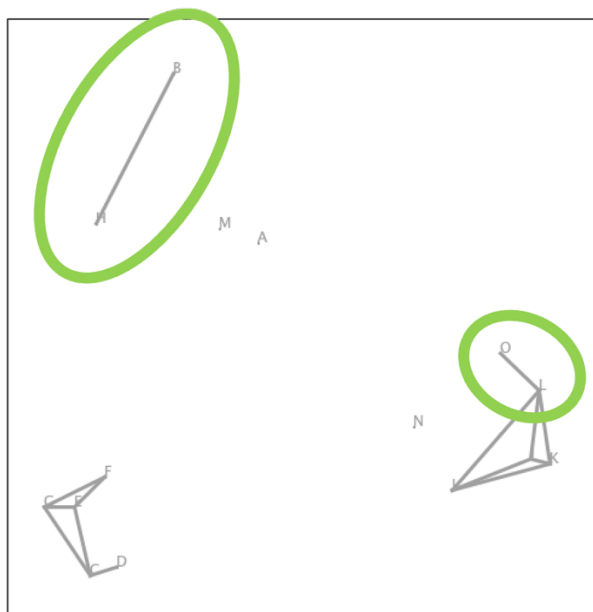
さらにスライダーを動かしていくと次はFとG、LとJ、LとKがそれぞれ線分で結合された。これにより緑色で囲まれたI, J, K, Lの部分は完全グラフのようになり、全ての2項目との間に相関があるということがわかった。青色で囲まれた部分に着目すると、以下の折れ線よりFとGの間にも正の相関が見られるということがわかった。F 雰囲気と G 酒・ドリンクと E サービスはいずれも食べログユーザーによるお店の評価であるため、これらに正の相関があることは納得がいく。



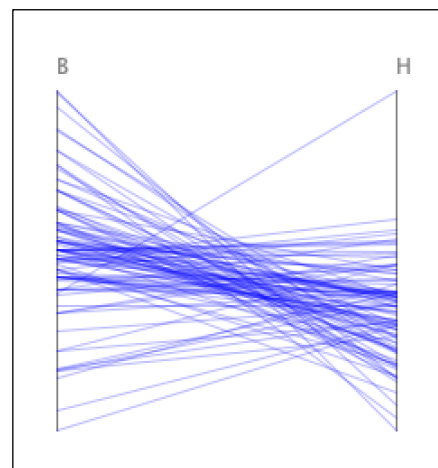
さらに右上のスライダーを上昇させていくと次はCとE、CとGがそれぞれ線分で結合された。折れ線よりC 総合評価とE サービスの間にも正の相関があるということがわかった。これにより、食べログユーザーによる6つの評価項目(C~H)のうち、Hのコストパフォーマンスを除く5つの項目には正の相関が見られるということがわかった。散布図を見ても、Hのコストパフォーマンスがかなり離れたところにあり、これだけは少し他の評価項目と性質が異なるということも明らかになった。



さらに進めていくと今度はBとH、LとOが線分で結合された。



BとHの折れ線

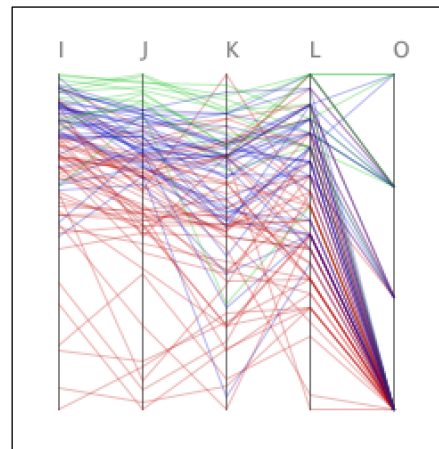


BとHの相関を折れ線グラフで確認してみると綺麗な負の相関になっていることがわかった。Bはマルゲリータの価格であり、Hは食べログユーザーがお店のコストパフォーマンスを評価した数値である。この結果から、マルゲリータの価格が安いほどコストパフォーマンスの数値が高くなる傾向があるということがわかった。このことから、お店側としては高い価格設定で客にコストパフォーマンスが良いと評価させることは難しいということが明らかになった。

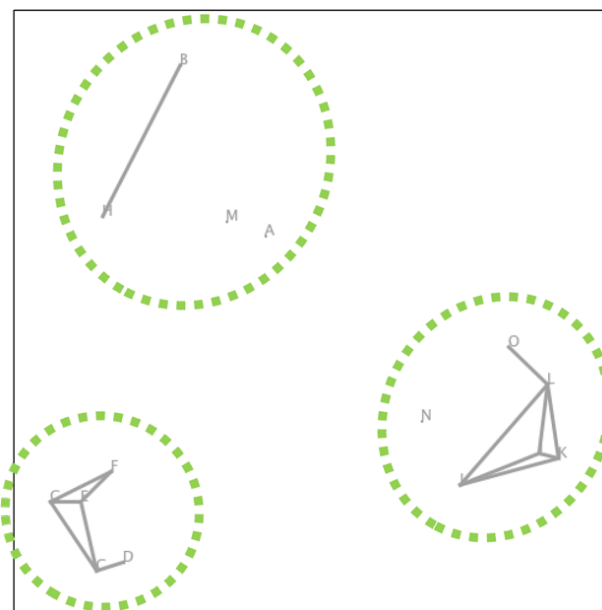
L と O の相関を折れ線で確認してみると正の相関が見られることがわかった。O は受賞歴であり最多は 4 回であった。初選出のお店はもちろん 1 回であるので、O には 1 から 4 までのいずれかの数字が入っている。そのため O の軸に着目すると綺麗に四箇所に分かれている。

一番始めの折れ線(1)から I, J, K, L の部分に正の相関が見られるということはわかっていたが、右図より新たに O についてもこれらと正の相関があるということが明らかになった。しかしここまで来てようやく結合されたため相関の強さはそれほどでもないということもわかった。

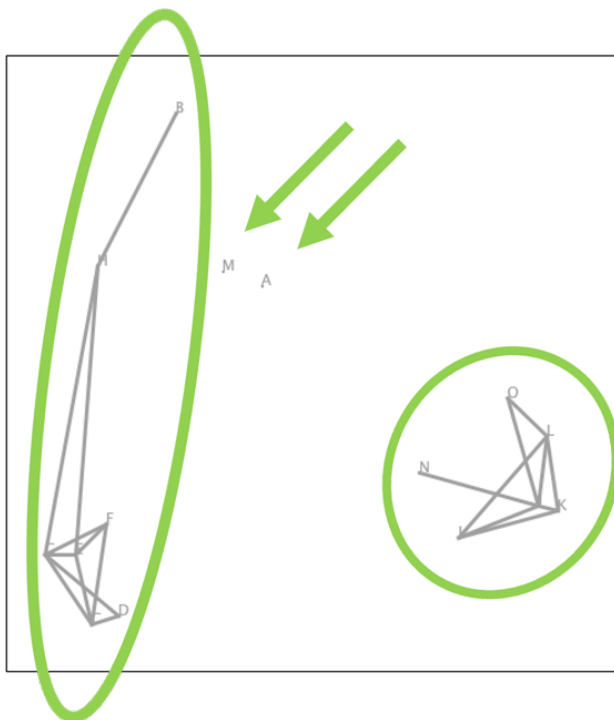
L と O の折れ線 (クラスタリング=3)



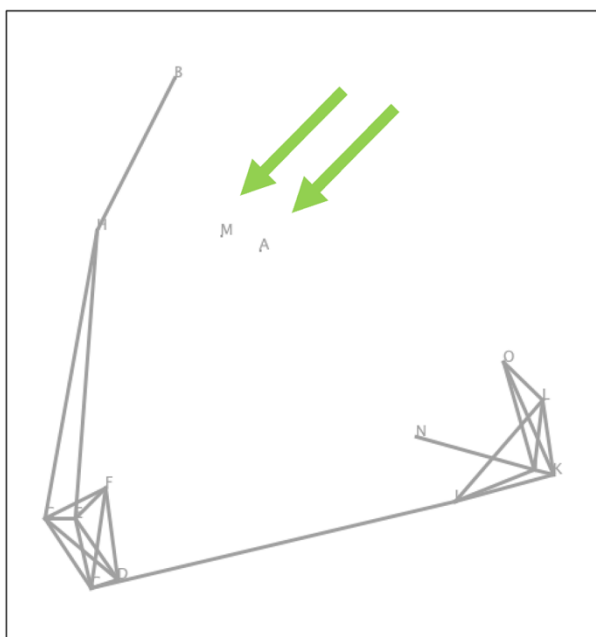
ここまでくると、散布図の見た目から以下のような 3 つのグループに分かれるのではないかと予想したが、実際は違っていた。



バーを上げていくと左上の B-N と左下のグループが線分で結ばれ全体が 2 つのグループに分かれた。緑の矢印で示した A のピザの種類数と M のオープン日からの経過年数はまだ残っている。



さらにバーを上げていくと上の 2 つのグループがつながり、緑の矢印で示した A ピザの種類数と M オープン日からの経過年数だけが残る形となった。このことから、ピザの種類数とオープン日からの経過年数は他のどの項目とも相関がほとんど見られないということがわかった。



ピザの種類数が多いほどお客さんの満足度が上がり、Cの総合評価も高くなるのではないかと思っていたが、実際にはこれらの間に相関はないということがわかり驚いた。しかし改めて考えてみると、お店の評価に関わる重大な要素は自分が食べた料理に対する満足度だと思うので、ピザの種類数は関係がないことにも納得がいった。

また、オープン日からの経過年数については、長いほどIの口コミ数やOの受賞歴が増えていったりするのではないかと思っていたが相関は見られなかった。そこでさらに調べてみると、食べログは、2005年3月に現在のサービスを開始し(参考文献⑥より)、2017年6月26日に食べログ百名店をスタートした(参考文献⑦より)ことがわかった。つまりそれ以前にオープンしていたお店に関しては、経過年数が何年であろうと食べログ上での差はないということがわかった。これによりMのオープン日からの経過年数は他の項目との相関が弱かったのだらうと考えた。

4.結論

ピザ百名店に関する15項目のデータを可視化した結果、C総合評価とD料理・味、I口コミ数とK投稿写真数の間に最も強い正の相関が見られることがわかった。

また、I口コミ数、J保存数、K投稿写真数、L百名店評価の全ての2項目との間には正の相関があることもわかった。特にLの百名店評価は、Iの口コミ数と一番に結合されたことから、百名店に選ばれるためには口コミ数を伸ばすことが重要ではないかと考えられる。

さらに、食べログユーザーによる6つの評価項目(C~H)のうち、Hのコストパフォーマンスを除く5つの項目には正の相関が見られ、HコストパフォーマンスはBマルゲリータの価格と負の相関があることもわかった。

最後にAピザの種類数とMオープン日からの経過年数が残ったことから、予想に反してこれらは他のどの項目とも相関が弱いということも明らかになった。

5.感想

まずテーマを決めるところから悩んだが、「ピザ百名店」にしたことは大正解だった。100店舗分の15項目全てのデータを集めることはとても大変だったが、百名店に選ばれる人気のあるお店のホームページを順番に1つ1つ見ていくことができて楽しかった。また、100行15列の表にまとめられたデータを見ただけではわからなかったことが、可視化によって徐々に明らかになっていくところがとても面白かった。データを有効に活用するために可視化はとても重要な役割を担うということがわかり興味が湧いてきた。15項目をどのように次元削減し2変数にしているのかや、左側の折れ線をどうやって瞬時に自動的に作成しているのかなどの裏側が気になった。

参考文献

- ① 「食べログの特長と便利な使い方」
<https://tabelog.com/help/beginner/>
- ② 「食べログ百名店」
<https://award.tabelog.com/hyakumeiten>
- ③ 「ナポリピッツァ Pizzeria la Rossa」
<https://tabelog.com/tokyo/A1317/A131702/13130666/>
- ④ 「点数・ランキングについて」
<https://tabelog.com/help/score/>
- ⑤ 「ピザ百名店 2021」
<https://award.tabelog.com/hyakumeiten/pizza/2021?sort=random>
- ⑥ 株式会社カカコム 「『食べログ』飲食店向けサービスを本格化」
<https://corporate.kakaku.com/press/release/20090311>
- ⑦ 株式会社カカコム 「高い評価を集めた 100 店を選出する『食べログ百名店』を開始」
<https://corporate.kakaku.com/press/release/20170626>