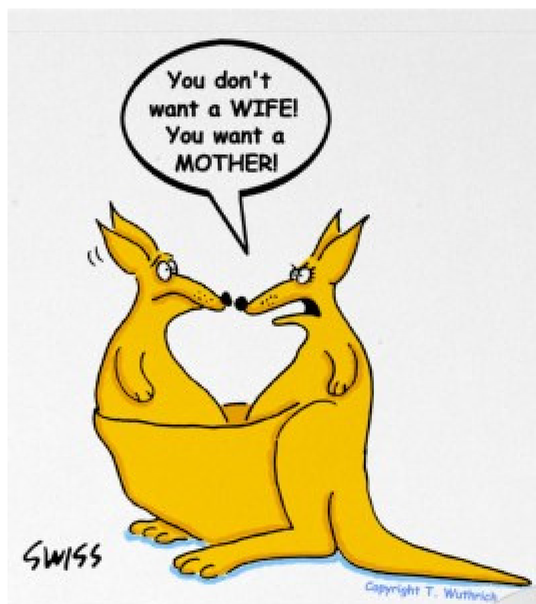
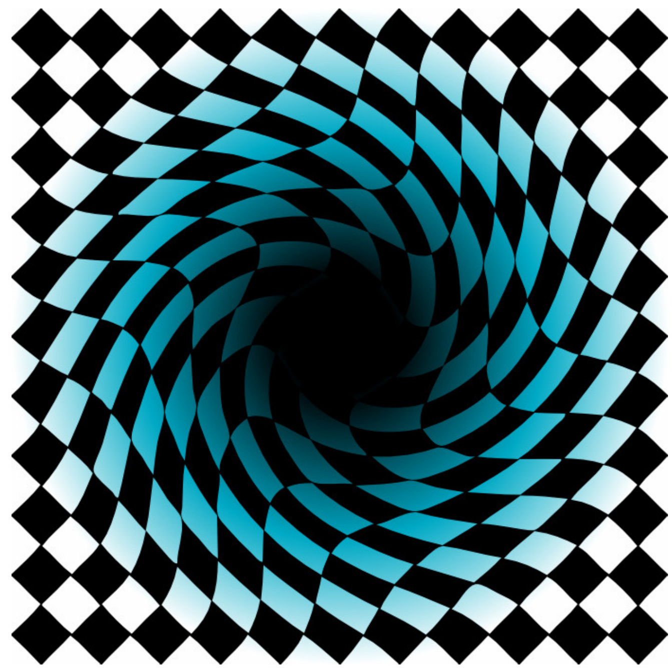
Dimension Reduction

Principal Components Analysis





Mechanics

Input: X_1**,** X_2 ,… , X_p
Output: PC_1, PC_2 ,… , PC_p (Ultimately we'll use a subset)

The i^{th} principal component is a weighted average:

$$PC_i = a_{i1} X_1 + a_{i2} X_2 + \ldots + a_{ip} X_p$$

Weights chosen such that:

1. PCs are ordered by their variance (PC_1 has largest variance)
2. Pairs of PCs have correlation = 0
3. For each PC, sum of squared weights =1

Example: Business School Programs

Univ	SAT	Top10	Accept	SFRatio	Expenses	GradRate
Brown	1310	89	22	13	22,704	94
CalTech	1415	100	25	6	63,575	81
CMU	1260	62	59	9	25,026	72
Columbia	1310	76	24	12	31,510	88
Cornell	1280	83	33	13	21,864	90
Dartmouth	1340	89	23	10	32,162	95
Duke	1315	90	30	12	31,585	95
Georgetown	1255	74	24	12	20,126	92
Harvard	1400	91	14	11	39,525	97
JohnsHopkins	1305	75	44	7	58,691	87
MIT	1380	94	30	10	34,870	91
Northwestern	1260	85	39	11	28,052	89
NotreDame	1255	81	42	13	15,122	94
PennState	1081	38	54	18	10,185	80
Princeton	1375	91	14	8	30,220	95
Purdue	1005	28	90	19	9,066	69
Stanford	1360	90	20	12	36,450	93
TexasA&M	1075	49	67	25	8,704	67
UCBerkeley	1240	95	40	17	15,140	78
UChicago	1290	75	50	13	38,380	87
UMichigan	1180	65	68	16	15,470	85
UPenn	1285	80	36	11	27,553	90
UVA	1225	77	44	14	13,349	92
UWisconsin	1085	40	69	15	11,857	71
Yale	1375	95	19	11	43,514	96

Use PCA to:

1) Reduce # columns
2) Identify relations between columns
3) Visualize universities in 2D

Source: US News & World Report, Sept 18 1995

PCA in XLMiner

Data Reduction & Exploration

Output specifies whether covariance or correlation matrix used (here – correlation matrix).

Principal Components

	Components					
Variable	**1**	**2**	**3**	**4**	**5**	**6**
SAT	0.45774868	0.03968045	0.18703876	0.13124055	0.02064597	-0.8580547
Top10	0.42714444	-0.19993152	0.49780852	0.37489522	0.48201644	0.39607504
Accept	-0.42430812	0.32089293	-0.15627895	0.06128667	0.80109364	-0.21693356
SFRatio	-0.39064837	-0.43256435	0.60608089	-0.50739086	0.07682328	-0.17204805
Expenses	0.3625232	0.63448638	0.20474122	-0.62340063	0.07254726	0.17376293
GradRate	0.37940401	-0.51555371	-0.53247261	-0.43863374	0.33810937	0.00353743

Variance	4.61208487	0.78681612	0.28656188	0.16378011	0.12430621	0.02645062
Variance%	76.86808014	13.11360168	4.77603149	2.72966838	2.07177019	0.44084364
Cum%	76.86808014	89.98168182	94.75771332	97.48738098	99.5591507	100
P-value	0	0.00000004	0.00073126	0.00263538	0.00140999	1

Reducing data dimension

Principal Components

	Components					
Variable	**1**	**2**	**3**	**4**	**5**	**6**
SAT	0.45774868	0.03968045	0.18703876	0.13124055	0.02064597	-0.8580547
Top10	0.42714444	-0.19993152	0.49780852	0.37489522	0.48201644	0.39607504
Accept	-0.42430812	0.32089293	-0.15627895	0.06128667	0.80109364	-0.21693356
SFRatio	-0.39064837	-0.43256435	0.60608089	-0.50739086	0.07682328	-0.17204805
Expenses	0.3625232	0.63448638	0.20474122	-0.62340063	0.07254726	0.17376293
GradRate	0.37940401	-0.51555371	-0.53247261	-0.43863374	0.33810937	0.00353743

Variance	4.61208487	0.78681612	0.28656188	0.16378011	0.12430621	0.02645062
Variance%	76.86808014	13.11360168	4.77603149	2.72966838	2.07177019	0.44084364
Cum%	76.86808014	89.98168182	94.75771332	97.48738098	99.5591507	100
P-value	0	0.00000004	0.00073126	0.00263538	0.00140999	1

PC_1 captures _____ % of the information

The first two PCs capture _____%

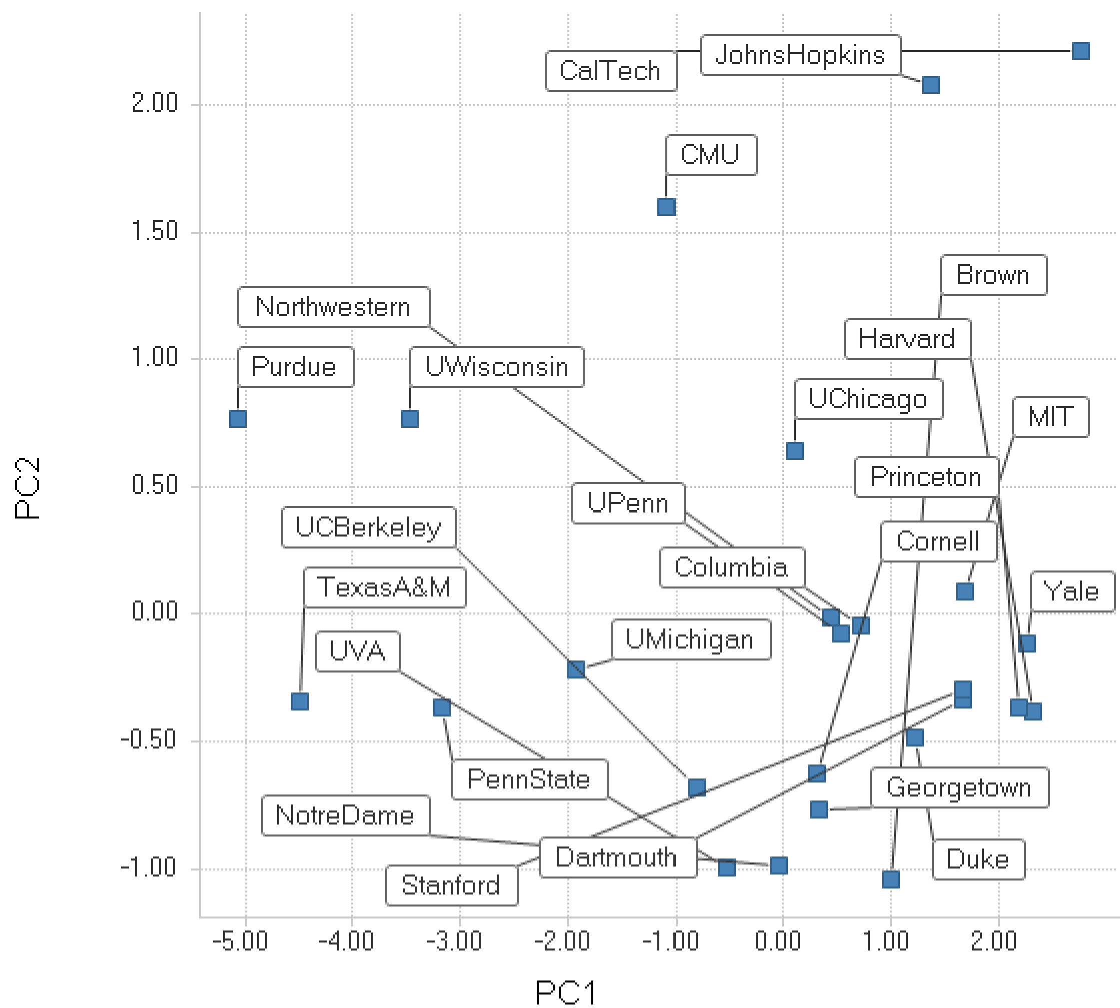
XLMiner: Computing scores

Scores given for each PC

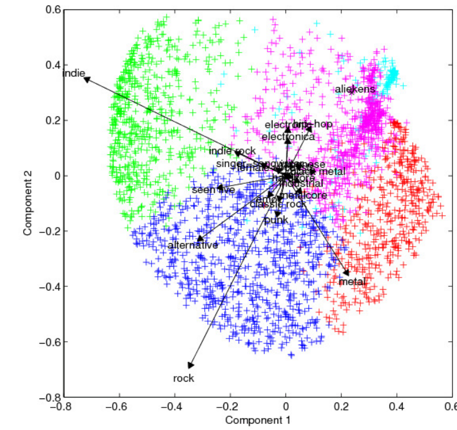
Recall: PC1 & PC2 are uncorrelated (*r = 0*)

Row Id.	1	2
1	0.98947096	-1.04280615
2	2.76521754	2.21340251
3	-1.08998942	1.5982517
4	0.72675508	-0.04133511
5	0.30561018	-0.62240905
6	1.66241097	-0.33740574
7	1.2216301	-0.48106378
8	0.33190566	-0.76930493
9	2.32618284	-0.37872922
10	1.37492549	2.07669187
11	1.69122922	0.08645435
12	0.44174835	-0.01090807
13	-0.03942522	-0.98881435
14	-3.168396	-0.36701241
15	2.19108367	-0.36428159
16	-5.06847715	0.76415795
17	1.66530418	-0.29942313
18	-4.48564911	-0.3405683
19	-0.80598319	-0.68478525
20	0.09578693	0.63730472
21	-1.92351854	-0.22022633
22	0.53133249	-0.07798085
23	-0.52146798	-0.99661624
24	-3.47699881	0.76273364
25	2.25931191	-0.11532623

Score plot (score2 vs score1) using Spotfire



SVD (similar to PCA)



Data Matrix: **M = U Σ V'**

Using an example from the Wikipedia page:

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 2 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 4 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & \sqrt{5} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \sqrt{0.2} & 0 & 0 & 0 & \sqrt{0.8} \\ 0 & 0 & 0 & 1 & 0 \\ -\sqrt{0.8} & 0 & 0 & 0 & \sqrt{0.2} \end{bmatrix}$$

Books rated by users

	User 1	User 2	User 3	User 4
Book 1	1	4		1
Book 2	2			5
Book 3		4	1	4
Book 4	1			
Book 5	3	5	5	1
Book 6		3	2	

Original matrix:

$$\begin{pmatrix} 1 & 4 & 0 & 1 \\ 2 & 0 & 0 & 5 \\ 0 & 4 & 1 & 4 \\ 1 & 0 & 0 & 0 \\ 3 & 5 & 5 & 1 \\ 0 & 3 & 2 & 0 \end{pmatrix}$$

Rebuilt with k=4, the ma

Approximation in rounde

$$\begin{pmatrix} 0 & 4 & 1 & 1 \\ 2 & 0 & 0 & 5 \\ 0 & 4 & 1 & 4 \\ 0 & 0 & 0 & 0 \\ 3 & 5 & 5 & 1 \\ 0 & 3 & 2 & 0 \end{pmatrix}$$

Need only
k columns of U
k values of Σ
k rows of V'

http://journal.batard.info/post/2009/04/08/svd-fun-profit

This week's online discussion

Discussion: Reducing Dimension of Mobile Survey Data

Prof. Galit Shmueli

Mar 21 at 3:29pm

16 20

Consider a sample from a dataset on mobile us
mobile users in India. Take a look at the data sa
Choose one of the points below and post a t
thread with the adequate point.

1. What approaches would you take to reduce
   which method you would apply to which col
2. Suppose the goal is *describing* the relations
   potential factors (customer demographics, t
   apply PCA? Which columns would you app
   describe the relationship?
3. For *predicting* service switching (churn), if w
   space, what information would we need to p
   PCA reduce the number of questions that w

Column Name	Description
serialnum	ID of respondent
StartDate	Survey start date/time
EndDate	Survey end date/time
SurveyDuration (Hrs.Min)	Survey Duration
Completed	Whether the survey was completed (1=yes)
Num Mobiles	Q1 Do you currently own one or more Mobiles? □ No. I use landline only. (1) □ One - with single SIM (2) □ One handset with two SIMs (3) □ Two handsets (4) □ More than two handsets or more than two SIMs (5)
Mobile Type (Primary)	What type of mobile phone handset do you own? If you own more than one, ple □ Basic phone without internet capability (1) □ Smartphone (non-touchscreen) (2) □ Smartphone with touch-screen (3) □ Tablet with phone features (4)
Service Provider (Primary)	Who is your current service provider for your primary mobile phone? □ Airtel (1) □ Reliance (2) □ Idea (3) □ Vodafone (4) □ Tata DOCOMO (5) □ Tata Indicom (6) □ Aircel (7) □ BSNL / MTNL (8) □ Uninor (9) □ Virgin Mobile (10) □ Other (11)
Network duration	Q41 How long have you been on this network? □ Less than 6 months (1) □ 6 months to 1 year (2) □ 1 to 2 years (3) □ More than 2 years (4)
Mobile Service	Q42 Which mobile service type do you use for your primary mobile phone? □ GSM (1) □ CDMA (2) □ Not sure/Don't know. (3)
Provider-Network Coverage	Rate your service provider on this area (1=poor, 2=below average, 3=average,
Provider-Call quality	Rate your service provider on this area (1=poor, 2=below average, 3=average,
Provider-Call charges	Rate your service provider on this area (1=poor, 2=below average, 3=average,
Provider-Roaming charges	Rate your service provider on this area (1=poor, 2=below average, 3=average,
Provider-Customer support	Rate your service provider on this area (1=poor, 2=below average, 3=average,
Provider-Offers and promotions	Rate your service provider on this area (1=poor, 2=below average, 3=average,
Provider-Easy bill payment, varied recharge options, etc	Rate your service provider on this area (1=poor, 2=below average, 3=average

Use compressed data in modeling

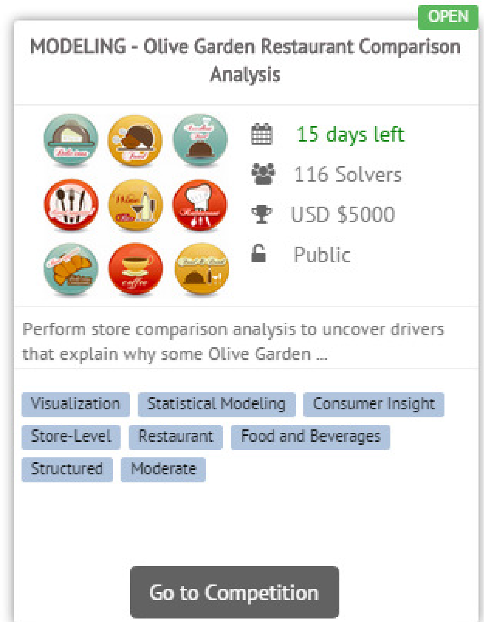
For predicting?

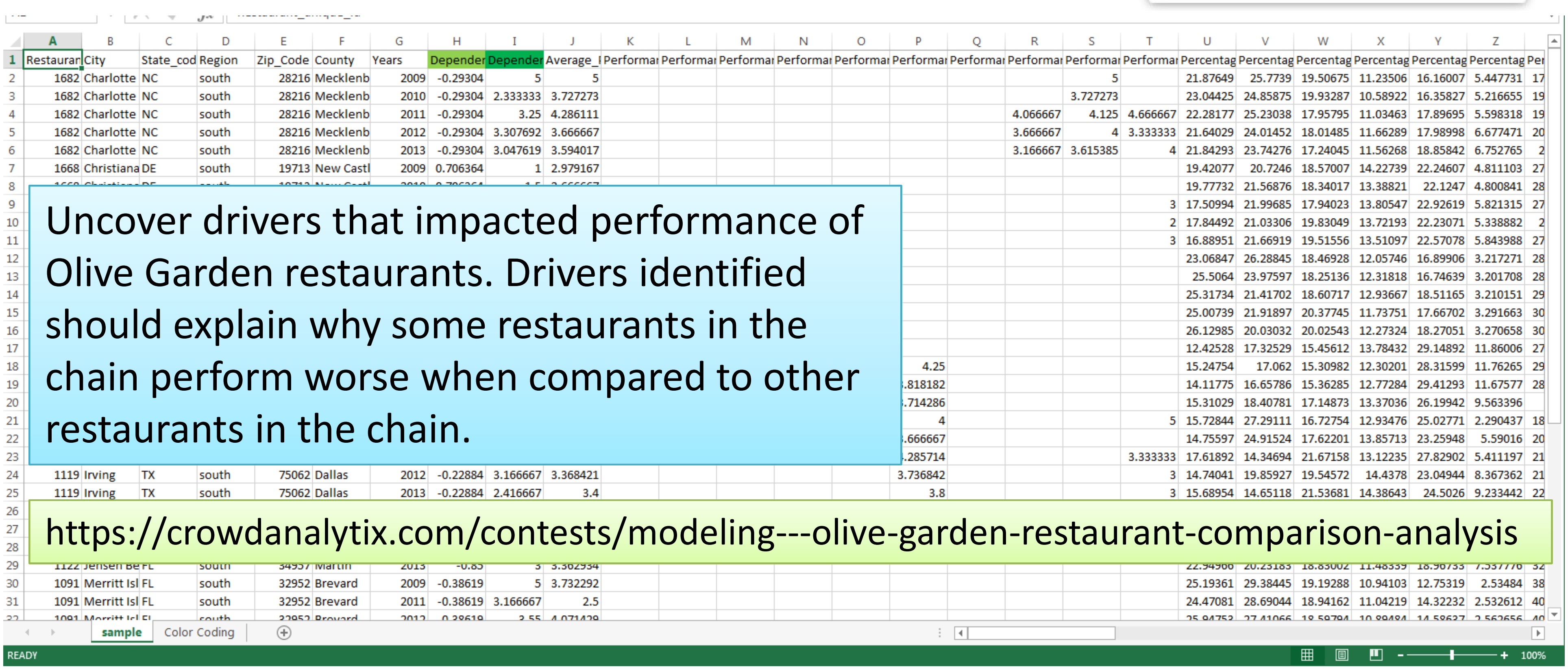
For explaining?

Data Mining Contest (Crowdanalytics)

Each restaurant, each year

Lots of variables!





	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	
1	Restauran	City	State_cod	Region	Zip_Code	County	Years	Depender	Depender	Average_I	Performar	Performar	Performar	Performar	Performar	Performar	Performar	Performar	Performar	Performar	Percentag	Percentag	Percentag	Percentag	Percentag	Percentag	Per
2	1682	Charlotte	NC	south	28216	Mecklenb	2009	-0.29304	5	5									5		21.87649	25.7739	19.50675	11.23506	16.16007	5.447731	17
3	1682	Charlotte	NC	south	28216	Mecklenb	2010	-0.29304	2.333333	3.727273									3.727273		23.04425	24.85875	19.93287	10.58922	16.35827	5.216655	19
4	1682	Charlotte	NC	south	28216	Mecklenb	2011	-0.29304	3.25	4.286111								4.066667	4.125	4.666667	22.28177	25.23038	17.95795	11.03463	17.89695	5.598318	19
5	1682	Charlotte	NC	south	28216	Mecklenb	2012	-0.29304	3.307692	3.666667								3.666667	4	3.333333	21.64029	24.01452	18.01485	11.66289	17.98998	6.677471	20
6	1682	Charlotte	NC	south	28216	Mecklenb	2013	-0.29304	3.047619	3.594017								3.166667	3.615385	4	21.84293	23.74276	17.24045	11.56268	18.85842	6.752765	2
7	1668	Christiana	DE	south	19713	New Castl	2009	0.706364	1	2.979167											19.42077	20.7246	18.57007	14.22739	22.24607	4.811103	27
8																					19.77732	21.56876	18.34017	13.38821	22.1247	4.800841	28
9																				3	17.50994	21.99685	17.94023	13.80547	22.92619	5.821315	27
10																				2	17.84492	21.03306	19.83049	13.72193	22.23071	5.338882	2
11																				3	16.88951	21.66919	19.51556	13.51097	22.57078	5.843988	27
12																					23.06847	26.28845	18.46928	12.05746	16.89906	3.217271	28
13																					25.5064	23.97597	18.25136	12.31818	16.74639	3.201708	28
14																					25.31734	21.41702	18.60717	12.93667	18.51165	3.210151	29
15																					25.00739	21.91897	20.37745	11.73751	17.66702	3.291663	30
16																					26.12985	20.03032	20.02543	12.27324	18.27051	3.270658	30
17																					12.42528	17.32529	15.45612	13.78432	29.14892	11.86006	27
18																4.25					15.24754	17.062	15.30982	12.30201	28.31599	11.76265	29
19																.818182					14.11775	16.65786	15.36285	12.77284	29.41293	11.67577	28
20																.714286					15.31029	18.40781	17.14873	13.37036	26.19942	9.563396	
21																4				5	15.72844	27.29111	16.72754	12.93476	25.02771	2.290437	18
22																.666667					14.75597	24.91524	17.62201	13.85713	23.25948	5.59016	20
23																.285714				3.333333	17.61892	14.34694	21.67158	13.12235	27.82902	5.411197	21
24	1119	Irving	TX	south	75062	Dallas	2012	-0.22884	3.166667	3.368421						3.736842				3	14.74041	19.85927	19.54572	14.4378	23.04944	8.367362	21
25	1119	Irving	TX	south	75062	Dallas	2013	-0.22884	2.416667	3.4						3.8				3	15.68954	14.65118	21.53681	14.38643	24.5026	9.233442	22
30	1091	Merritt Isl	FL	south	32952	Brevard	2009	-0.38619	5	3.732292											25.19361	29.38445	19.19288	10.94103	12.75319	2.53484	38
31	1091	Merritt Isl	FL	south	32952	Brevard	2011	-0.38619	3.166667	2.5											24.47081	28.69044	18.94162	11.04219	14.32232	2.532612	40

Uncover drivers that impacted performance of Olive Garden restaurants. Drivers identified should explain why some restaurants in the chain perform worse when compared to other restaurants in the chain.

https://crowdanalytix.com/contests/modeling---olive-garden-restaurant-comparison-analysis