

# Projet BIG DATA :

## Utilisateurs YELP



Vicky Wang  
Maxence Genet  
M2 IES

# Introduction



# Introduction

## YELP :

- plateforme en ligne créée en 2004
- permettant aux utilisateurs de noter, commenter, recommander des commerces locaux
- commerces locaux : restaurants, cafés, coiffeurs ou services divers



# SOMMAIRE

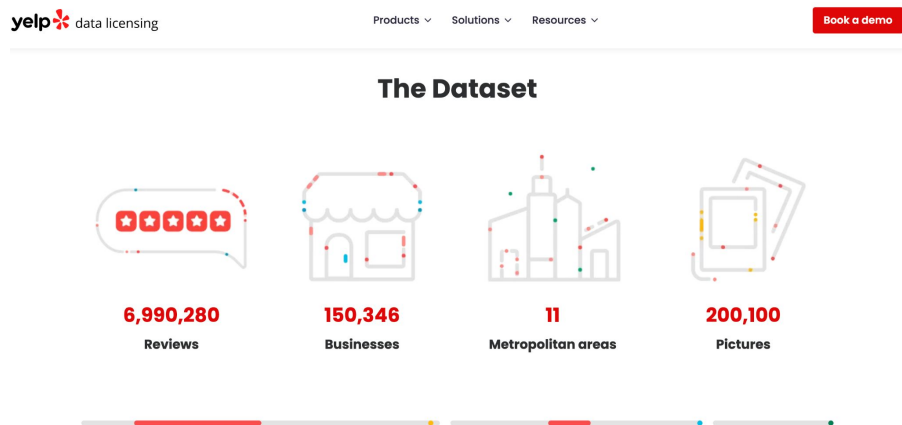
1. Présentation de la base de données
2. Construction de l'entrepôt de données
3. Construction des flux ETL
4. Requêtage
5. Rapport d'analyse
6. Conclusion

# 1. Présentation de la base de données

# 1. Présentation de la base de données

- Yelp Open Dataset Challenge  $\Rightarrow$  2 bases de données en open data
- 1 base contenant les informations sur les établissements, les utilisateurs et leurs interactions via les avis
- 1 autre base contenant les informations relatives aux photos

$\Rightarrow$  données recensées entre 2004 et 2022



# 1. Présentation de la base de données

- business.json (établissements)
- review.json (avis)
- et user.json (utilisateurs)

⇒ Ces tables sont reliés entre eux par des identifiants (business\_id, user\_id)

⇒ Focus sur l'état de la Californie aux US



# Table : Business

business_id	name	address	city	state	postal_code	latitude	longitude	stars	review_count	is_open	categories	hours		
Pns2l4eNsfO8kk	Abby Rappoport	1616 Chapala St	Santa Barbara	CA	93101	34.4266787	-119.7111968	5.0	7		0 Doctors, Traditional Chinese Medicine, Naturopathic/Holistic,			
noByYNI DLQAr	H&M	827-833 State S	Santa Barbara	CA	93101	34.4202089255	-119.700459738	3.0	24		1 Women's Clothir	{"Monday": "11:0-18:0", "Tuesday": "11:0-18:0",		
IDtLPgUrqorppq	Helena Avenue	1131 Anacapa St,	Santa Barbara	CA	93101	34.4144445	-119.6906718	4.0	389		1 Food, Restaurar	{"Monday": "0:0-0:0", "Tuesday": "8:0-14:0",		
nUqrF-h9S7myC	Iron Horse Auto	825 Cacique St	Santa Barbara	CA	93103	34.4196198	-119.6770319	4.5	32		1 Towing, Automot	{"Monday": "0:0-0:0", "Tuesday": "8:0-17:0",		
bYjnX_J1bHZob	Tinkle Belle Diaper Service		Santa Barbara	CA	93101	34.420334	-119.7107494	5.0	17		1 Parenting Class	{"Monday": "0:0-0:0", "Tuesday": "10:0-18:0",		
SZU9c8V2GuRE	Santa Barbara S 230 Stearns Wh		Santa Barbara	CA	93101	34.4087147	-119.6850187	4.0	2404		1 Live/Raw Food,	{"Monday": "0:0-0:0", "Tuesday": "11:0-21:0",		

Variables composant la table **Business** :

- business\_id : identifiant du commerce
- name : nom du commerce
- adress, city, psotal\_code latitude, longitude et state
- stars : nombre d'étoile sur 5
- review\_count : nombre de notes
- is\_open : activité des commerces
- categories
- hours : horaires des commerces

⇒ au total 5203 commerces





# Table : Review

review_id	user_id	business_id	stars	useful	funny	cool	date
ZKvDG2sBvHVd wSTuiTk-sKNdcf B5XSoSG3SfvQ			3.0		1	1	0 2016-03-30 22:4
pUycOfUwM8vq; 59MxRhNVhU9M gebiRewfieSdt1			3.0		0	0	0 2016-07-25 07:3
eCiWBf1CJ0Zdv OhECKhQEexFy vC2qm1y3Au5cz			4.0		0	0	0 2013-09-04 03:4

Variables composant la table **Review** :

- review\_id : identifiant des commentaires
  - user\_id : identifiant des utilisateurs
  - business\_id : identifiant des commerces
  - stars : nombre d'étoile sur 5
  - useful, funny, cool : type de commentaires
  - date : date de publication du commentaire
- ⇒ au total 348 857 commentaires



# Table : User

user_id	name	review_count	yelping_since	useful	funny	cool	fans	average_stars	compliment_hot	compliment_mor	compliment_prol	compliment_cuti	compliment_list	compliment_noti	compliment_plai	compliment_coo	compliment_funi	compliment_writ	compliment_photo
qVc8ODYUSSZj	Walker	585	2007-01-25 16:4	7217		1259	5994	267 3.91	250	65	55	56	18	232	844	467	467	239	180
SZDeASXq7o05	Gwen	224	2005-11-29 04:3	512		330	299	28 4.27	24	4	1	6	2	12	16	26	26	10	9
q_QQ5kBBwCc	Jane	1221	2005-03-14 20:2	14953		9940	11211	1357 3.85	1713	163	191	361	147	1212	5696	2543	2543	815	323
iYzhPPqnrjJkg1	Chris	4	2010-11-03 18:5	1		0	1	4.0	0	0	0	0	0	0	0	0	0	0	0
QF1Kus8iwLW	Catherine	607	2009-04-27 20:2	4573		3714	4149		1020	79	68	29	23	144	407	808	808	587	723

Variables composant la table **User** :

- user\_id : identifiant des utilisateurs
  - name : nom des utilisateurs
  - review\_count : nombre de commentaires publiés
  - yelping\_since : date de création du compte utilisateur
  - useful, funny, cool : comptage du type de commentaires publiées par l'utilisateur
  - average\_stars : note moyenne qu'à donner l'utilisateur
  - compliment\_"" : détails du commentaire de l'utilisateur
- ⇒ 155 949 utilisateurs Yelp

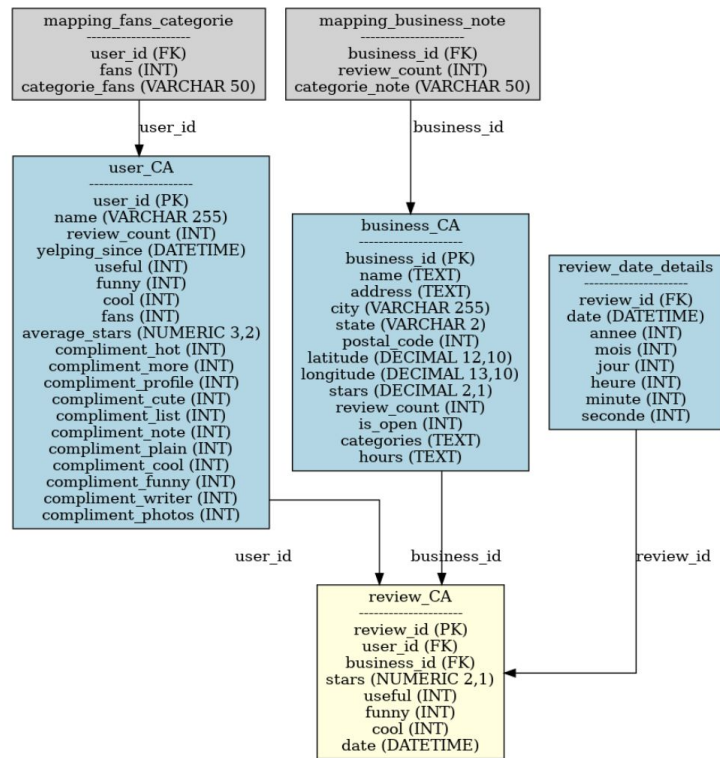


## 2. Construction de l'entrepôt de données

# Création d'une nouvelle base : MySQL

Base de données YELP :

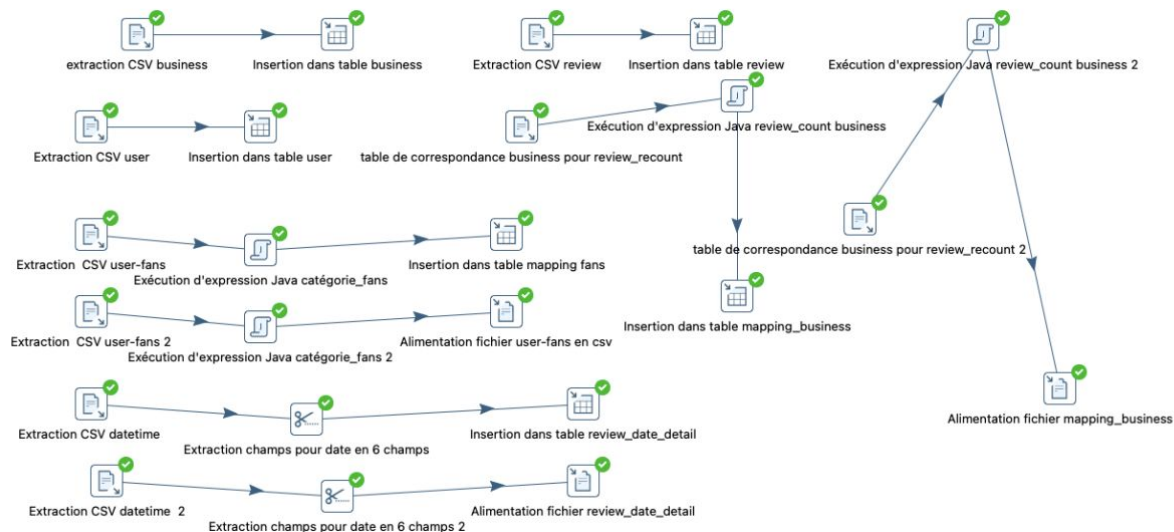
- table vide avec dimension des variables dans MySQL
- intégration des différentes tables utiles à notre analyse
- review\_CA : table reliant les différentes interactions entre commerces, utilisateurs et commentaires
- les principales clés primaires : user\_id, business\_id, review\_id



### 3. Construction des flux ETL

# Intégration avec Pentaho

- Extraction -> transformation  
-> alimentation/insertion
- Création de table de correspondance et de dimension à partir de Pentaho.
- Doublon de requête pour exploiter nos données sur différentes



## 4. Requêtage



# Analyse des utilisateurs Yelp les plus influents

	critere	user_id	name	review_count	fans
▶	Top Review	Hi10sGSZNxQH3NLYWSZ1oA	Fox	17473	3493
▶	Top Fans	hizGc5W1tBHPghM5YKCAtg	Katie	1825	3642

Entre activité (**reviews**) et notoriété (**fans**) :

- Fox est celui qui a rédigé le plus grand nombre d'avis  $\Rightarrow$  17 473 reviews
- Katie possède le plus grand nombre de fans avec 3 642 personnes qui la suivent.  
 $\Rightarrow$  Bien qu'elle ait rédigé moins d'avis (1 825), elle bénéficie d'une forte popularité.



# Répartition des utilisateurs par catégorie de fans et par année (Roll up)

Evolution du profil des utilisateurs au fil du temps :

- Montée en popularité de la plateforme en 2007
- En 2007 :
  - 1520 utilisateurs avaient "Très peu de fans" avec 5032 fans
  - 68 utilisateurs appartenait à la catégorie "Très nombreux fans" avec un total de 18 716 fans.

annee	categorie_fans	nb_users	total_fans
2004	Aucun fan	6	0
2004	Beaucoup de fans	1	54
2004	Nombre modéré de fans	2	28
2004	Très nombreux fans	3	3698
2004	Très peu de fans	8	25
2005	Aucun fan	31	0
2005	Beaucoup de fans	24	1693
2005	Nombre modéré de fans	56	1301
2005	Très nombreux fans	20	7027
2005	Très peu de fans	77	321
2006	Aucun fan	203	0
2006	Beaucoup de fans	70	4789
2006	Nombre modéré de fans	285	6502
2006	Très nombreux fans	68	18716
2006	Très peu de fans	500	1764
2007	Aucun fan	721	0
2007	Beaucoup de fans	166	11226
2007	Nombre modéré de fans	649	15047
2007	Très nombreux fans	138	36677
2007	Très peu de fans	1520	5032

# Analyse de l'évolution des avis à travers une moyenne mobile

Tendance sur la satisfaction des clients dans le temps :

- pour Alameda Park : la moyenne reste stable entre 2016 et 2020
- pour Stewart's De Rooting and Plumbing : la moyenne mobile varie plus

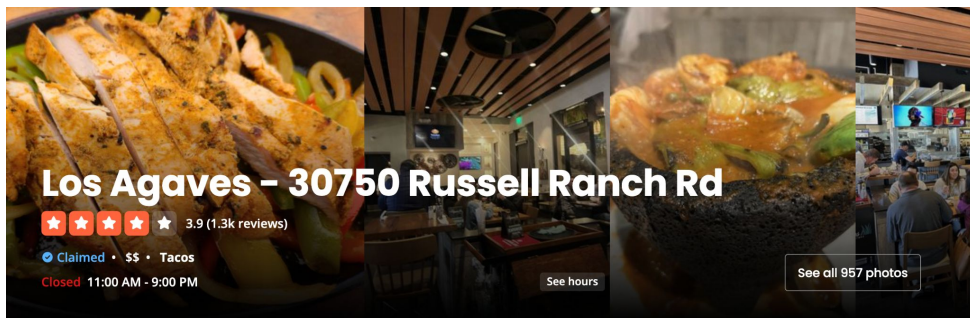
⇒ L'utilisation de la moyenne mobile permet de repérer des évolutions progressives de chaque commerce

business_id	business_name	review_year	avg_stars	moving_avg
--O3ip9NpXTKD4oBS1pY2A	Alameda Park	2016	4.16667	4.727778000
--O3ip9NpXTKD4oBS1pY2A	Alameda Park	2017	4.60000	4.716161818
--O3ip9NpXTKD4oBS1pY2A	Alameda Park	2018	4.55556	4.702778333
--O3ip9NpXTKD4oBS1pY2A	Alameda Park	2019	4.85714	4.714652307
--O3ip9NpXTKD4oBS1pY2A	Alameda Park	2020	5.00000	4.735034285
-06ngMH_Ejkm_6HQBYxB7g	Stewart's De Rooting & Plumbing	2016	5.00000	3.041667500
-06ngMH_Ejkm_6HQBYxB7g	Stewart's De Rooting & Plumbing	2017	3.66667	3.166668000
-06ngMH_Ejkm_6HQBYxB7g	Stewart's De Rooting & Plumbing	2018	3.66667	3.250001666
-06ngMH_Ejkm_6HQBYxB7g	Stewart's De Rooting & Plumbing	2019	3.50000	3.285715714
-06ngMH_Ejkm_6HQBYxB7g	Stewart's De Rooting & Plumbing	2020	4.00000	3.375001250
-0hxpklpBh2T0tvdM1mSlw	Brookstone	2016	3.00000	1.666666666
-0hxpklpBh2T0tvdM1mSlw	Brookstone	2017	2.25000	1.812500000
-0hxpklpBh2T0tvdM1mSlw	Brookstone	2018	3.00000	2.050000000
-1g8Qb6t_mSX_ak1thMmrQ	FreeWalkingTourSB	2017	5.00000	5.000000000
-1g8Qb6t_mSX_ak1thMmrQ	FreeWalkingTourSB	2018	5.00000	5.000000000
-1g8Qb6t_mSX_ak1thMmrQ	FreeWalkingTourSB	2019	4.94444	4.981480000
-1g8Qb6t_mSX_ak1thMmrQ	FreeWalkingTourSB	2020	5.00000	4.986110000
-1ze-oWDnrGAzvAg56QXUA	GraphicInk	2016	5.00000	5.000000000
-1ze-oWDnrGAzvAg56QXUA	GraphicInk	2017	5.00000	5.000000000
-1ze-oWDnrGAzvAg56QXUA	GraphicInk	2018	5.00000	5.000000000
-1ze-oWDnrGAzvAg56QXUA	GraphicInk	2019	5.00000	5.000000000
-3Aooxlkg38UyUdlz5oXdw	Chase Restaurant	2016	3.43243	3.303183333
-3Aooxlkg38UyUdlz5oXdw	Chase Restaurant	2017	4.14286	3.387151000
-3Aooxlkg38UyUdlz5oXdw	Chase Restaurant	2018	3.64103	3.410230909
-3Aooxlkg38UyUdlz5oXdw	Chase Restaurant	2019	3.36585	3.406532500
-3Aooxlkg38UyUdlz5oXdw	Chase Restaurant	2020	3.12329	3.384744615
-3kTgLg2lUr2PbaokvcnmQ	Santa Barbara Balayage by Kar...	2016	5.00000	5.000000000
-3kTgLg2lUr2PbaokvcnmQ	Santa Barbara Balayage by Kar...	2019	5.00000	5.000000000
-6jvfSJGprbfBD2QrS9zQw	Mesa Produce	2017	5.00000	5.000000000
-6jvfSJGprbfBD2QrS9zQw	Mesa Produce	2018	5.00000	5.000000000
-6jvfSJGprbfBD2QrS9zQw	Mesa Produce	2019	5.00000	5.000000000
-6L_z3fD1ieoJb0FJohw	Channel Islands Outfitters	2016	4.96000	4.855238000

# Analyse des avis reçus par l'établissement Los Agaves

⇒ Los Agaves est le commerce qui a récolté le plus grand nombre de notes

⇒ moyenne de 4% étoiles



★ Write a review   Add photo   Share   Save

## Menu

### Popular Dishes



View full menu >

## Order Food

Takeout   Delivery

No Fees   Pick up in 30-40 mins

Start Order

review_id	review_date	review_stars	business_stars
pOeeHDXyycVdjLjzH6kM_Q	2008-09-13	5.0	4.0
rKhaxkXmkLb_l2a0vrqhyw	2008-12-09	5.0	4.0
vRLEaNFsSzUbGyzVzpjboQA	2009-02-17	4.0	4.0
q_aaC1CRT4r3x_g3kkmwA	2009-03-03	4.0	4.0
iUp8rg1uDwS0_NH4y6yl6Q	2009-03-03	3.0	4.0
OcAwYgy7qM9LjH-hdGig9A	2009-03-03	5.0	4.0
BO8t754Mz4lhry_DYpuWPQ	2009-03-22	4.0	4.0
oLcnEk9-1xUvadcWu1O_oA	2009-03-26	4.0	4.0
boOqEwGIL7VelgUilYrz8g	2009-04-01	5.0	4.0
XowB6i3TTCe0frOd06TAOW	2009-04-06	5.0	4.0
WnQNFZCxrMm_bRbSupJ...	2009-04-09	4.0	4.0
J9NR006NBh5kFzbQz-gASQ	2009-05-12	5.0	4.0
n2lR_SwlPjdK5DPoAlKr1Q	2009-06-01	5.0	4.0
S1CoNb71TTLRwvMomsc14Q	2009-06-10	4.0	4.0
I_jVdWRjNOQ9igAE1CcUdw	2009-06-15	5.0	4.0
bCXraFz2qhN7-2LXVbTefw	2009-06-24	1.0	4.0
GSgJ69Q-icTGc_bkRPodgw	2009-07-11	3.0	4.0
iIT9quS6_oyEO5vrNCt-6g	2009-07-29	5.0	4.0
cQZ1YImSNfCkJE7fpXxw5A	2009-09-20	3.0	4.0
wEguML63fqHwOTHg09wqA	2009-10-13	4.0	4.0
alidOR98f9CuQE8nQLLP1g	2009-10-17	3.0	4.0
NsnZ1_Qny8i5xEIRKKI8Ww	2009-11-24	5.0	4.0
9NXpUy7xCx4TkHEWw0N...	2010-01-08	5.0	4.0
lr5bq9lj5jIP7RAIgYWbTA	2010-01-18	5.0	4.0
Q2JAWpZZ8EoGZYFVHdq...	2010-01-18	4.0	4.0
_dNln-3iWI_9LuYFep5MsA	2010-01-22	5.0	4.0
I-rkn8cvKPuhda9MRkMRLQ	2010-02-12	5.0	4.0
Fyj_QOamM54ddRoQ5_G...	2010-03-07	4.0	4.0
4Pg5CzUbQE-baFe772F-9w	2010-03-15	5.0	4.0
3nsB3cJAM9S7e64cJPLoUQ	2010-03-18	4.0	4.0
lt5oRh3fyBNKoSOELeyvAw	2010-03-20	5.0	4.0
slVuu2E1-FbooEvrLqYMuQ	2010-03-25	5.0	4.0

# Analyse de la première et de la dernière note reçue par les établissements

Evolution des évaluations données aux

établissements en comparant leur première et leur dernière note reçue :

- pour **H&M** : la première note datant de 2011 était  $\frac{4}{5}$  et la dernière note datant de 2020 est  $\frac{5}{5}$   
⇒ perte de la satisfaction client
- pour **Santa Barbara Shellfish Company** : la première note datant de 2005 était de  $\frac{5}{5}$  et la dernière note datant de 2022 reste  $\frac{5}{5}$   
⇒ satisfaction constante des clients

business_id	name	first_review_st...	first_review_date	last_review_sta...	last_review_date
Pns2l4eNst0Bkk83dixA6A	Abby Bappoport, I.A.C. CMO	5.0	2012-05-02 18:07:38	5.0	2015-03-16 03:43:08
noByYNI DLQAr9ccqxdlDw	H&M	4.0	2011-06-24 03:17:08	2.0	2020-07-26 18:51:31
lDILPgUrqorppqSLdMhZQ	Helena Avenue Bakery	4.0	2016-07-11 14:42:07	4.0	2022-01-16 22:28:56
nUqF-h9S7myCovNDecOvw	Iron Horse Auto Body	5.0	2014-08-04 01:20:45	5.0	2021-12-03 21:02:27
bYinX_J1bHZob10DsSfkgQ	Tinkle Belle Diaper Service	5.0	2017-02-27 18:11:18	5.0	2021-08-07 20:15:58
SZU9c8V2GuREDN5KgyHFJw	Santa Barbara Shellfish Company	5.0	2005-07-07 21:56:53	5.0	2022-01-19 05:21:58
QZU7TczBb30XaPvVCkXg	805 Ink	5.0	2010-08-15 01:20:39	5.0	2022-01-01 23:47:37
25Uww0C0wvF9CZ_3B6vWtA	Enjoy The Mountain	5.0	2015-05-05 05:07:23	3.0	2021-10-23 16:16:26
xF9r1XbMvEOsJeHlmFhlw	Weddings in Santa Barbara	5.0	2009-08-31 06:09:18	5.0	2019-04-18 19:34:00
4xhGQGDGqU60BlznBjquA	California Tacos and Taproom	5.0	2018-12-06 04:12:23	5.0	2021-11-28 02:37:25
Rad7b6MFOEmh41CdREMO	Isla Vista Community Bike Center	5.0	2018-01-10 08:30:20	5.0	2018-10-22 23:33:07
ifjluUv4VASwmFqEp8cWlQ	Marty's Pizza	4.0	2007-08-19 08:08:32	5.0	2022-01-07 05:35:59
VeFtrEZ4lWacrQg6Eq4cg	Cal Taco	5.0	2007-01-13 22:49:12	1.0	2021-11-21 03:05:15
4pS7qnJ7_DCoJB-EnI7KA	Prop and Decor Outlet: The Tent...	5.0	2017-12-29 23:46:28	5.0	2020-12-05 17:52:50
X7FQ5k29A_RRyep7TOfjtw	Dawna Ara, DACM, LAC	5.0	2009-09-30 00:46:57	5.0	2021-06-14 03:22:29
bdfZdB2MTXIT6-RBjSlpOg	Pho Bistro	2.0	2007-07-19 10:13:03	4.0	2021-11-03 00:59:16
VEbHYioBtoPiIOWntE_DBA	Tienda Ho	4.0	2008-06-08 02:01:54	5.0	2021-09-13 10:13:32
xwSWUJcQkzTF6Hnm_I_Mcgg	Rusty's Pizza Parlor	3.0	2006-12-29 05:03:23	1.0	2021-11-08 02:31:21
82suwumWp1MEq04QhS9DQ	Surreal Virtual Reality Studio	5.0	2019-10-10 15:08:15	5.0	2020-09-14 02:42:45
uHprVoxTV7CwKJdaEEf3aA	Challenge Asphalt Paving	1.0	2012-12-20 16:30:02	5.0	2021-07-22 03:31:53
-kY_HDP7lMvGI-kBIZVU4A	Dune Coffee Roasters - Anacapa	5.0	2012-10-10 19:27:50	3.0	2021-12-08 05:03:29
18eWJFJbXyR9J_5xlcRLYA	Siam Elephant	4.0	2009-03-13 01:00:16	5.0	2022-01-17 05:30:32
6jomGWEI4rylmrWpGuUQQQ	Santa Barbara Athletic Club	1.0	2010-03-24 17:31:45	1.0	2021-12-06 17:34:40
Ed7RUOj9MTU9U-Y73bYlw	Run Montecito-Summerland	4.0	2011-07-17 05:01:41	5.0	2021-12-19 21:30:13
VelrgRMOK_0ZTolzUd2kUA	Franceschi Park	5.0	2009-04-13 19:00:02	4.0	2021-11-02 06:28:18
BkmVHg6HjHWMcOfXcd6VQ	Hair By Audrey Johnson	5.0	2013-05-11 16:13:22	5.0	2016-10-08 00:43:37
Jb1MIURq6lIKZ244yoirA	Jessie Sessions - Berkshire Hat...	5.0	2019-07-11 02:51:49	5.0	2021-11-02 18:29:14
B5XSoSG3SvQGikEQG1tISQ	Los Padres National Forest	5.0	2008-04-27 14:11:55	5.0	2020-10-20 18:30:12
vLT1KtrA9bWvjFOg-0xVlg	Pleology Pizzeria	4.0	2016-08-27 00:04:50	1.0	2020-02-28 04:05:01
-ujBP1Dw0j1-Flaz97-LXQ	Lama Dog Tap Room	5.0	2016-05-14 17:24:52	4.0	2021-12-01 19:04:15
wYROX7w9fkQrWUS2YyY9g	Will Nelson Fitness	5.0	2017-08-17 22:51:01	5.0	2018-06-16 19:00:09
BMTIMWDR3zcR5T30GFDoXQ	CA Pro Home Inspection	5.0	2015-10-03 19:53:56	5.0	2016-08-16 16:49:18
kH0Ycn_I7SniwcmHsGluA	The Ranch Grill at Palomar	4.0	2007-06-05 05:31:32	3.0	2011-06-19 00:05:29

## 5. Rapport d'analyse



# Dashboard “Categories”

## Categories

Please select a category

categories: Vegetarian, Gluten-Free, Food, Restaurants, Coffee & Tea, American (New), Vegan, ... (146) ▾



**Shops count**

185



**Open shops**

115



**Review count**

24 550

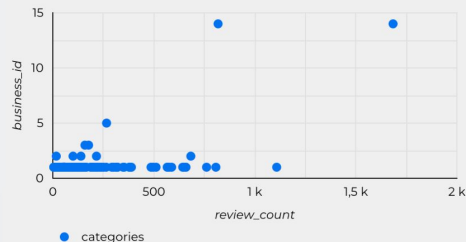


**AVG Stars**

3,77

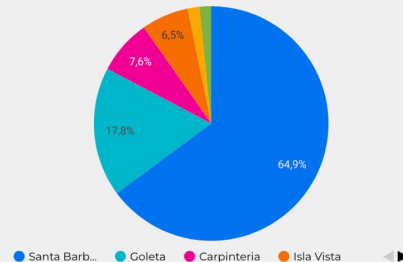
Please select a point on the graph

**AVG review by category**



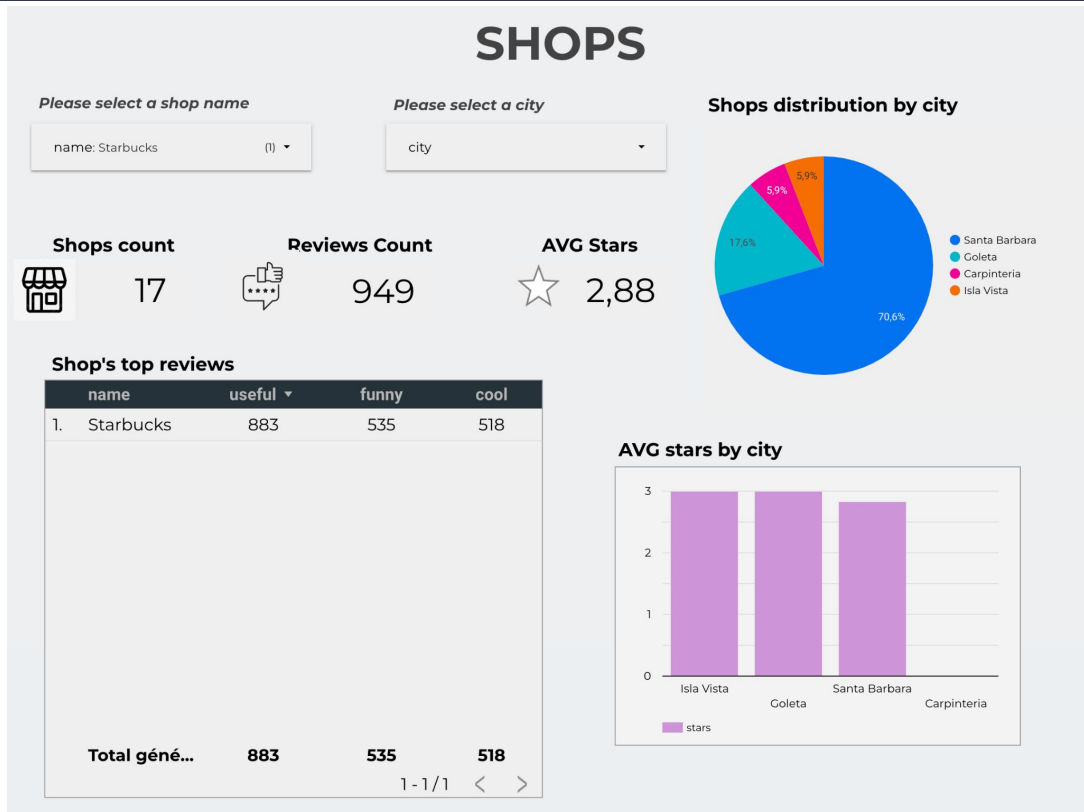
Scatter plot showing the total number of reviews by business category

**Business distribution by city**



Pie chart showing the distribution of businesses by city in California listed on Yelp

# Dashboard “Shops”



# Dashboard “Popular User”

## POPULAR USER

Please select a shop name

name: Starbucks

(1) ▾

Please select a review star

stars ▾

Please select a city

city ▾

Total popular user

29

Avg review by popular user



3.0

Column chart of comments sent to popular...

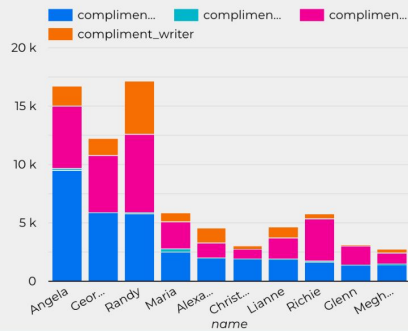


table of ratings given to a business by popular users

	Business name	User name	city	Business star	Review star
1.	Starbucks	Lianne	Santa Barb...	2,5	3
2.	Starbucks	Abbey	Santa Barb...	3	2
3.	Starbucks	Stephanie	Santa Barb...	3	2
4.	Starbucks	Ashley	Santa Barb...	3	2
5.	Starbucks	Jen	Santa Barb...	null	1
6.	Starbucks	Aurora	Santa Barb...	null	1
7.	Starbucks	Ashley	Goleta	3	1
8.	Starbucks	Michelle	Goleta	3	1
9.	Starbucks	Faye	Goleta	3	1
10.	Starbucks	Archie	Goleta	3	1



## 6. Conclusion



# Conclusion

Evolution des évaluations données aux établissements en comparant leur première et leur dernière note reçue :

- pour **H&M** : la première note datant de 2011 était  $\frac{4}{5}$  et la dernière note datant de 2020 est  $\frac{2}{5}$   
⇒ perte de la satisfaction client
- pour **Santa Barbara Shellfish Company** : la première note datant de 2005 était de 5/5 et la dernière note datant de 2022 reste 5/5  
⇒ satisfaction constante des clients