Maxence Genet
Vicky WANG
M2 IES

Projet BIG DATA :

Analyse des utilisateurs YELP

Université Paris Nanterre

Année universitaire 2024-2025

SOMMAIRE

1. Introduction

Ce projet a été réalisé en binôme par Maxence GENET et Vicky WANG. Nous avons choisi ensemble la base de données Yelp. Le choix s'est fait de manière naturelle, car nous cherchions une base riche, variée et réaliste pour permettre des analyses intéressantes autour de la satisfaction client, de la réputation en ligne et du comportement des utilisateurs. Cette base de données nous intéresse car nous voulions ouvrir une boutique plus tard. Cela nous servirait de template pour faire des études de marché poussées.

Une fois la base sélectionnée, nous avons réfléchi ensemble à la manière de nous répartir les tâches tout en gardant une dynamique de travail collaboratif. Nous avons d'abord travaillé à deux sur des scripts en amont pour filtrer et réduire la base initiale, qui était très (voire trop) volumineuse. Cela nous a permis d'extraire les données les plus pertinentes et de faciliter les étapes suivantes.

Maxence s'est ensuite occupé de la partie transformation des données avec Pentaho. Il a mis en place les différents flux d'intégration, créé les tables de l'entrepôt de données et réalisé les requêtes SQL nécessaires à l'analyse. De son côté, Vicky a pris en charge une grande partie du travail sur Power BI. Elle a conçu les visualisations interactives et structuré les tableaux de bord pour mettre en valeur les résultats obtenus.

Tout au long du projet, nous avons travaillé en collaboration, en nous aidant mutuellement sur les différentes étapes.

2. Présentation de la base de données

Le jeu de données utilisé dans ce projet provient du site officiel de Yelp, dans le cadre du Yelp Open Dataset Challenge, accessible à l'adresse https://www.yelp.com/dataset. Yelp propose deux bases en open data : nous avons choisi la première, plus légère, qui ne contient pas les données relatives aux photos. Ce choix s'est fait à la fois pour des raisons de faisabilité technique et pour nous permettre de nous concentrer sur l'essentiel : les établissements, les utilisateurs et leurs interactions via les avis.
Le dataset Yelp est composé de plusieurs fichiers au format JSON, chacun représentant un type d'objet : business.json (établissements), review.json (avis) et user.json (utilisateurs). Ces fichiers sont bien structurés et reliés entre eux par des identifiants (business_id, user_id), ce qui facilite leur exploitation relationnelle.
Nous avons été motivés par ce choix de dataset car il permet non seulement de travailler sur un volume conséquent et réaliste, mais aussi d'explorer une problématique proche de nos intérêts personnels. En effet, à plus long terme, certains membres du groupe envisagent d'ouvrir leur propre établissement. Yelp est une plateforme incontournable en matière de visibilité locale, et la capacité à analyser ses données constitue un atout stratégique. Ce projet nous sert ainsi de modèle réutilisable pour faire des études de marché, analyser les zones géographiques, comprendre les attentes des clients ou identifier les tendances à partir des retours des utilisateurs.
Étant donné le volume important des fichiers, nous avons dû concevoir des scripts Python pour filtrer les données et en réduire la taille, tout en gardant une cohérence entre les

entités. Nous avons commencé par le fichier business.json, en filtrant uniquement les établissements situés dans l'État de Californie (state = CA). Cette extraction nous a permis de constituer la table business_CA.
À partir de cette table, nous avons extrait tous les identifiants business_id pour sélectionner les avis correspondants dans review.json, formant ainsi la table review_CA. Enfin, nous avons collecté les user_id apparaissant dans les avis afin de filtrer les utilisateurs pertinents dans user.json, ce qui a permis de générer la table user_CA.
Ces trois tables principales (business_CA, review_CA, user_CA) constituent la base de notre entrepôt de données. Leur structuration progressive nous a permis de réduire les volumes tout en maintenant une logique relationnelle robuste, essentielle pour la modélisation décisionnelle et l'analyse multidimensionnelle qui suit.

3. Construction de l'entrepôt de données et flux ETL

3.1 Préparation des données sources

Le dataset Yelp est initialement fourni au format JSON, réparti en plusieurs fichiers volumineux (business, review, user, etc.). Dans le but de faciliter son exploitation et de se concentrer sur un périmètre géographique cohérent, nous avons filtré les établissements situés exclusivement dans l'État de la Californie (state = CA) à partir du fichier business.json.
À partir des business_id extraits, nous avons récupéré dans review.json les avis associés, constituant ainsi la table review_CA. Enfin, les user_id figurant dans ces avis nous ont permis d'isoler les utilisateurs pertinents dans user.json, donnant naissance à la table user_CA.
Une fois ces trois jeux de données extraits, nous avons converti les fichiers JSON en CSV structurés en supprimant les colonnes non nécessaires à notre analyse et en standardisant les formats (types numériques, format de date). Ces fichiers CSV ont ensuite servi de base pour l'intégration dans l'entrepôt.

3.2 Création des tables MySQL (schéma relationnel)

Avant toute insertion de données, nous avons conçu un script SQL pour créer les tables vides de notre entrepôt dans une base MySQL nommée yelp. Ce script définit la structure relationnelle de notre entrepôt avec les clés primaires et étrangères nécessaires.
La table principale review_CA joue le rôle de table de faits. Elle regroupe les avis laissés par les utilisateurs sur les établissements, avec les variables suivantes :

- review_id (clé primaire)
- user_id (clé étrangère vers user_CA)
- business_id (clé étrangère vers business_CA)
- stars, useful, funny, cool
- date (date de publication de l'avis)

Les dimensions associées sont les suivantes :

- user_CA, qui contient les informations complètes des utilisateurs : leur nom, ancienneté (yelping_since), nombre de fans, moyenne des étoiles attribuées, et l'ensemble des compliments reçus sur la plateforme. La clé primaire est user_id.

- business_CA, qui décrit les établissements Yelp : nom, adresse, ville, coordonnées géographiques, catégories, horaires, nombre d'avis et moyenne des notes. La clé primaire est business_id

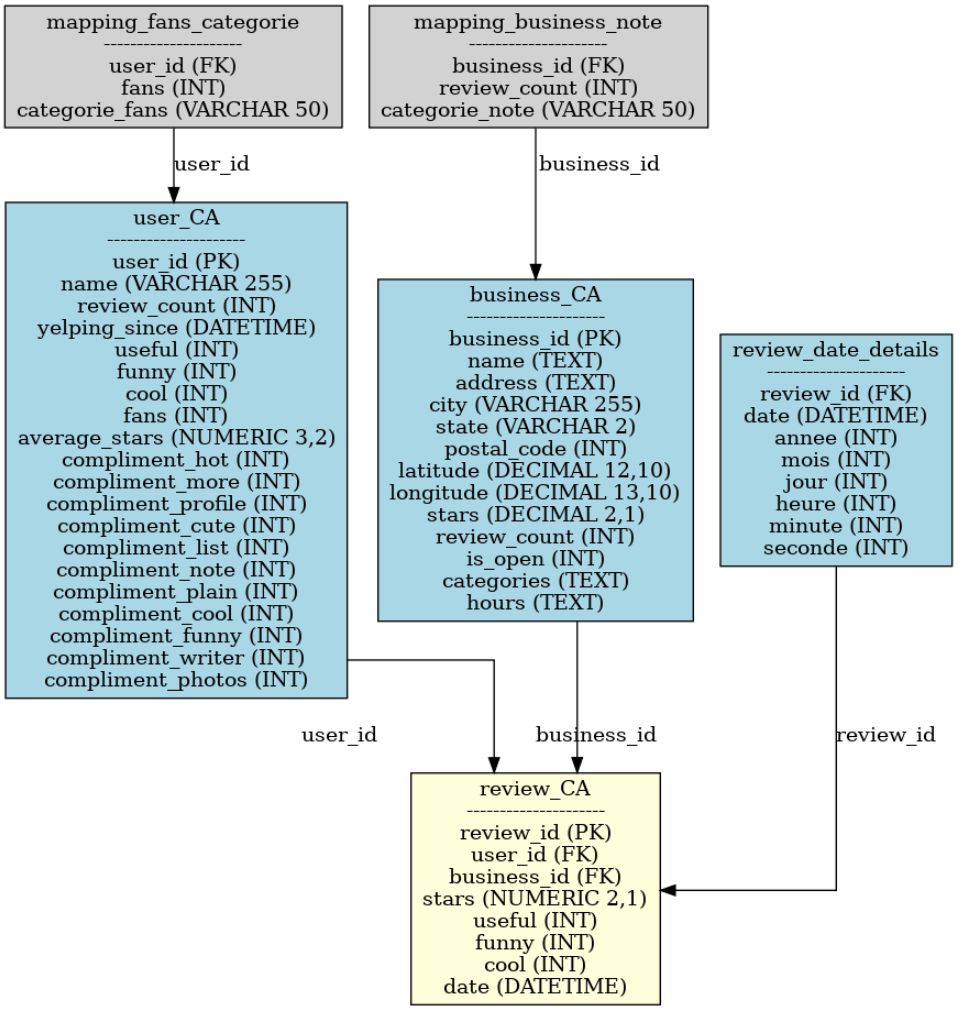
- review_date_details, table temporelle dérivée de review_CA, qui reprend le champ date et le décompose en année, mois, jour, heure, minute, seconde. Elle est liée à review_CA par la clé review_id

Deux tables de correspondance viennent enrichir l'entrepôt avec des éléments catégoriels supplémentaires :

- mapping_business_note : permet de catégoriser les établissements selon leur volume d'avis (review_count) et de leur attribuer une étiquette analytique (categorie_note). Elle est reliée à business_CA via business_id.

- mapping_fans_categorie : classe les utilisateurs selon leur popularité en fonction du nombre de fans (categorie_fans). Elle est reliée à user_CA via user_id.

Schéma relationnel de l'entrepôt de données :



3.3 Intégration des données avec Pentaho

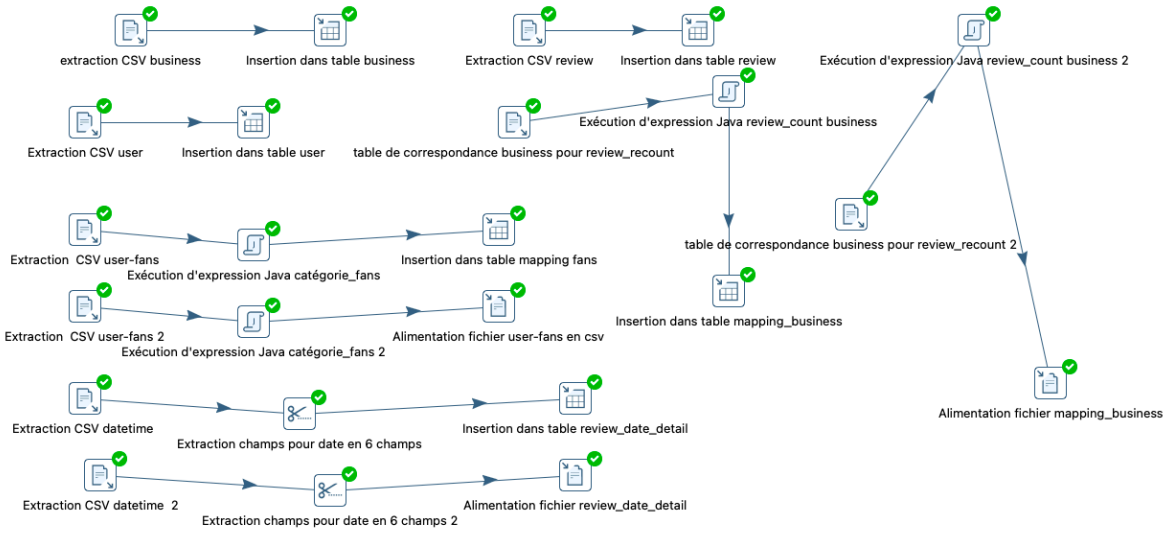
Une fois les tables créées dans MySQL, nous avons utilisé Pentaho Data Integration pour construire les flux ETL qui assurent l'insertion automatisée des données dans l'entrepôt.
Les fichiers CSV sont d'abord lus avec le composant CSV file input. Des transformations sont ensuite appliquées, incluant :

- Le filtrage de valeurs nulles ou invalides
- La conversion des types (notamment pour les dates)
- Le renommage ou l'harmonisation des colonnes
- L'extraction de champs spécifiques pour alimenter les tables auxiliaires

Les données sont ensuite chargées dans MySQL avec les composants Table Output (ou Insert/Update pour les cas nécessitant des mises à jour sans duplication).

3.4 Organisation des flux ETL

Nous avons structuré nos transformations .ktr par table source, avec un fichier principal par entité (user_CA.ktr, business_CA.ktr, review_CA.ktr...) et des transformations complémentaires pour les extractions spécifiques (review_date_details.ktr, mapping_fans_categorie.ktr...).
Cette organisation modulaire facilite la maintenance et permet de rejouer rapidement un flux en cas de mise à jour des données sans devoir retraiter l'ensemble.



Dans cette transformation, nous avons alimenté à la fois des tables dans notre entrepôt de données et des fichiers CSV. Ces fichiers CSV (qui sont les tables de correspondances des dimensions) ont été nécessaires pour pouvoir exploiter les données avec Looker car nous ne pouvions pas utiliser Power BI sur nos machines ni connecter directement Looker à notre entrepôt. Nous avons donc lié Looker à notre Google Drive pour lire les CSV, avant d'alimenter notre base de données avec ces mêmes fichiers. C'est pourquoi nous avons des transformations en doublons. De plus, la table de correspondance review_date_detail n'a pas pu être lue dans Looker car le logiciel n'arrivait pas à lire le fichier.

- Table business_CA : Pour la table business_CA, nous avons extrait les données depuis un fichier CSV contenant les informations des établissements, puis nous les avons insérées dans l'entrepôt de données.

- Table review_CA : Pour la table review_CA, nous avons également extrait un fichier CSV contenant les avis des utilisateurs, que nous avons insérés dans l'entrepôt. À partir de ces données, nous avons calculé le nombre d'avis par établissement grâce à une expression Java, et créé des correspondances entre les avis et les établissements. Ces correspondances ont été utilisées pour enrichir la table MAPPING_BUSINESS_NOTE.

- Table user_CA : La table user_CA a été alimentée à partir d'un fichier CSV contenant les profils des utilisateurs. Nous avons inséré ces données dans l'entrepôt afin de conserver des informations détaillées comme le nombre de reviews, la date d'inscription, les compliments reçus ou encore le score moyen attribué aux établissements.

- Table MAPPING_FANS_CATEGORIE : nous avons extrait les données utilisateurs avec leur nombre de fans, puis utilisé une expression Java pour les classer dans des catégories (fan == 0 : « aucun fan » ; fan < 10 : « très peu de fans » ; fans entre 10 et 50 : « nombre modéré de fans » ; fans entre 50 et 100 : « beaucoup de fans » ; fans > 100 : « très nombreux fans »). Ces données ont été insérées dans la table de correspondance et également exportées au format CSV.

- Table MAPPING_BUSINESS_NOTE : À partir de la table review_CA, nous avons calculé le nombre d'avis par établissement grâce à une expression Java, et créé des correspondances entre les avis et les établissements. Ces correspondances ont été utilisées pour enrichir la table MAPPING_BUSINESS_NOTE (inférieur à 50 = très peu de notes, entre 50 et 100 = peu de notes, et plus de 100 = noté convenablement).

- Table REVIEW_DATE_DETAIL : nous avons extrait les timestamps des avis et les avons décomposés en plusieurs champs (année, mois, jour, heure, minute, seconde). Ces informations temporelles ont été insérées dans une table dédiée et également exportées en fichier CSV.

4. Requêtage

Après cette phase de transformation et d'alimentation de notre entrepôt de données ainsi que des fichiers CSV, nous allons désormais passer à l'étape des requêtes. Ces requêtes, parfois complexes, vont nous permettre d'explorer et de comprendre les grandes tendances présentes dans notre base de données.

4.1 Analyse des utilisateurs les plus influents

critere	user_id	name	review_count	fans
Top Review	Hi10sGSZNxQH3NLyWSZ1oA	Fox	17473	3493
Top Fans	hizGc5W1tBHPghM5YKCAtg	Katie	1825	3642

Cette requête permet d'identifier deux utilisateurs clés dans la base de données selon deux critères différents. L'utilisateur nommé Fox est celui qui a rédigé le plus grand nombre d'avis, avec un total de 17 473 reviews. Cela en fait l'utilisateur le plus actif en termes de contribution sur la plateforme. De son côté, l'utilisatrice Katie possède le plus grand nombre de fans, avec 3 642 personnes qui la suivent. Bien qu'elle ait rédigé moins d'avis (1 825), elle bénéficie d'une forte popularité. Cette distinction met en évidence deux formes d'influence différentes sur Yelp : l'activité et la notoriété.

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

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

Cette requête présente une analyse de la répartition des utilisateurs selon leur catégorie de fans, pour chaque année d'inscription sur Yelp. On observe différentes catégories comme "Aucun fan", "Très peu de fans", "Nombre modéré de fans", "Beaucoup de fans" et "Très nombreux fans", construites à partir du nombre de fans de chaque utilisateur.

Pour chaque combinaison année–catégorie, la requête fournit deux indicateurs : le nombre d'utilisateurs (nb_users) et le nombre total de fans (total_fans) associés à ces utilisateurs. Par exemple, en 2007, 1520 utilisateurs avaient "Très peu de fans", totalisant 5032 fans, tandis que 68 utilisateurs appartenaient à la catégorie "Très nombreux fans" avec un total de 18 716 fans.

Cette analyse permet de repérer l'évolution du profil des utilisateurs au fil du temps. On remarque que certaines années, comme 2007 et 2008, enregistrent un nombre plus élevé

d’utilisateurs dans toutes les catégories, ce qui reflète une montée en popularité de la plateforme. Les utilisateurs les plus influents (ayant beaucoup ou très nombreux fans) restent peu nombreux, mais concentrent une part importante du total de fans, ce qui confirme leur rôle central dans la dynamique communautaire de Yelp.

4.3 Analyse de l’évolution des avis à travers une moyenne mobile

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
-0hxpkIpBh2T0tvdM1mSIw	Brookstone	2016	3.00000	1.666666666
-0hxpkIpBh2T0tvdM1mSIw	Brookstone	2017	2.25000	1.812500000
-0hxpkIpBh2T0tvdM1mSIw	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
-3AooxIkg38UyUdlz5oXdw	Chase Restaurant	2016	3.43243	3.303183333
-3AooxIkg38UyUdlz5oXdw	Chase Restaurant	2017	4.14286	3.387151000
-3AooxIkg38UyUdlz5oXdw	Chase Restaurant	2018	3.64103	3.410230909
-3AooxIkg38UyUdlz5oXdw	Chase Restaurant	2019	3.36585	3.406532500
-3AooxIkg38UyUdlz5oXdw	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_z3ftD1iepJb0FfJqhw	Channel Islands Outfitters	2016	4.96000	4.855238000

Cette requête permet d’étudier l’évolution des notes moyennes attribuées à différents établissements au fil des années en calculant une moyenne mobile. Pour chaque ligne, on retrouve l’identifiant de l’établissement, son nom, l’année de l’avis, la note moyenne obtenue cette année-là et la moyenne mobile, qui lisse les variations annuelles.

Cette méthode est utile pour observer des tendances sur la satisfaction des clients dans le temps. Par exemple, pour Alameda Park, la moyenne mobile reste stable entre 2016 et 2020, ce qui indique une constance dans la perception des avis. En revanche, pour Stewart's De Rooting and Plumbing, la moyenne mobile varie davantage, ce qui peut traduire des changements dans la qualité perçue par les clients.

L’utilisation de la moyenne mobile permet ainsi de repérer des évolutions progressives, que ce soit des améliorations continues ou des dégradations dans la note moyenne attribuée à un établissement. Elle apporte une vision plus régulière et lisible de la satisfaction des utilisateurs dans le temps.

4.3 Analyse des avis reçus par l’établissement Los Agaves

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
vRLEaNFSzUbGyzVzpjboQA	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
BO8t754Mz4Ihry_DYpuWPQ	2009-03-22	4.0	4.0
oLcnEk9-1xUvadcWu1O_oA	2009-03-26	4.0	4.0
boOqEwGIL7VeIgUilYrz8g	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
S1CoNb71TTLRwvMomscI4Q	2009-06-10	4.0	4.0
l_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
wEguML63fqHWoTHg09wkqA	2009-10-13	4.0	4.0
alidOR98f9CuQE8nQLLP1g	2009-10-17	3.0	4.0
NsnZ1_Qny8i5xElRKKI8Ww	2009-11-24	5.0	4.0
9NXpUy7xCx4TkHEWW0N…	2010-01-08	5.0	4.0
lr5bq9lj5jIP7RAlgYWbTA	2010-01-18	5.0	4.0
Q2JAWpZZ8EoGZYFVHdq…	2010-01-18	4.0	4.0
_dNln-3iWl_9LuYFep5MsA	2010-01-22	5.0	4.0
l-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-FbooEyrLqYMuQ	2010-03-25	5.0	4.0

Cette requête affiche l’ensemble des avis attribués à l’établissement Los Agaves, qui fait partie des établissements ayant reçu le plus grand nombre de notes dans la base de données. Pour chaque avis, on retrouve la date, la note donnée par l’utilisateur, ainsi que la note moyenne de l’établissement au moment de la publication de cet avis.

On remarque que de nombreux utilisateurs ont attribué la note maximale de 5 étoiles, alors que la note moyenne de l’établissement est souvent légèrement inférieure, autour de 4 étoiles. Cela montre que l’établissement bénéficie d’une perception globalement positive, avec une forte proportion d’avis très favorables.

L’analyse des écarts entre les notes individuelles et la moyenne générale permet d’évaluer la stabilité de la qualité perçue, ainsi que l’évolution de la satisfaction client au fil du temps. Le fait que cet établissement continue de recevoir des notes élevées renforce son image de valeur sûre auprès des utilisateurs.

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

business_id	name	first_review_st...	first_review_date	last_review_sta...	last_review_date
Pns2l4eNsfO8kk83dixA6A	Abby Rappoport, LAC, CMQ	5.0	2012-05-02 18:07:38	5.0	2015-03-16 03:43:08
noByYNtDLQAra9ccqxdfDw	H&M	4.0	2011-06-24 03:17:08	2.0	2020-07-26 18:51:31
IDtLPgUrqorrpqSLdfMhZQ	Helena Avenue Bakery	4.0	2016-07-11 14:42:07	4.0	2022-01-16 22:28:56
nUqrF-h9S7myCcvNDecOvw	Iron Horse Auto Body	5.0	2014-08-04 01:20:45	5.0	2021-12-03 21:02:27
bYjnX_J1bHZob10DoSFkqQ	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
QZU7TcrztBb3tXaPbVCkXg	805 Ink	5.0	2010-06-15 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
xF9r1XbMvEOsJeHlmFhIvw	Weddings in Santa Barbara	5.0	2009-08-31 06:09:18	5.0	2019-04-18 19:34:00
4xhGQGdGqU60BIznBjqnuA	California Tacos and Taproom	5.0	2018-12-06 04:12:23	5.0	2021-11-28 02:37:25
Rad7bl6MFOEemh41CdREMQ	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	5.0	2007-08-19 08:08:32	5.0	2022-01-07 05:35:59
VeFfrEZ4iWaecrQg6Eq4cg	Cal Taco	4.0	2007-01-13 22:49:12	1.0	2021-11-21 03:05:15
4pS7qnjJ7_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-RBjSIpQg	Pho Bistro	2.0	2007-07-19 10:13:03	4.0	2021-11-03 00:59:16
VEbHYioBfoPilOWntE_DBA	Tienda Ho	4.0	2008-06-08 02:01:54	5.0	2021-09-13 10:13:32
xwSWtJcQkzTF6HNm_IMgcg	Rusty's Pizza Parlor	3.0	2006-12-29 05:03:23	1.0	2021-11-08 02:31:21
82suwumWp1MEq04QhiS9DQ	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_HDP7IMvGl-kBIZVU4A	Dune Coffee Roasters - Anacapa	5.0	2012-10-10 19:27:50	3.0	2021-12-08 05:03:29
18eWJFJbXyR9j_5xfcRLYA	Siam Elephant	4.0	2009-03-13 01:00:16	5.0	2022-01-17 05:30:32
6jomGWEl4rylmrWPguUQQQ	Santa Barbara Athletic Club	1.0	2010-03-24 17:31:45	1.0	2021-12-06 17:34:40
Ed7RUOj9MTUf9U-Y73bYlw	Run Montecito-Summerland	4.0	2011-07-17 05:01:41	5.0	2021-12-19 21:30:13
VeIrgRMOK_0ZToziUd2kUA	Franceschi Park	5.0	2009-04-13 19:00:02	4.0	2021-11-02 06:28:18
BkmVHg6HjHWMc0OFxcd6VQ	Hair By Audrey Johnson	5.0	2013-05-11 16:13:22	5.0	2016-10-08 00:43:37
Jb1MlURq6ltKZ244tyoirA	Jessie Sessions - Berkshire Hat...	5.0	2019-07-11 02:51:49	5.0	2021-11-02 18:29:14
B5XSoSG3SfvQGtKEGQ1tSQ	Los Padres National Forest	5.0	2008-04-27 14:11:55	5.0	2020-10-20 18:30:12
vLT1KtrA9bWvjFOg-0xVlg	Pieology Pizzeria	4.0	2016-08-27 00:04:50	1.0	2020-02-28 04:05:01
-ujBP1Dw0j1-Ffaz97-LXQ	Lama Dog Tap Room	5.0	2016-05-14 17:24:52	4.0	2021-12-01 19:04:15
wYROX7rw9fkQrWUS2YyY9g	Will Nelson Fitness	5.0	2017-08-17 22:51:01	5.0	2018-06-16 19:00:09
BMTtMWDR3zcR5T30GFDoXQ	CA Pro Home Inspection	5.0	2015-10-03 19:53:56	5.0	2016-08-16 16:49:18
kHQXo-IZSpixnctmHsGlvA	The Beach Grill at Padaro	4.0	2007-06-05 05:31:32	3.0	2011-06-19 00:05:29

Cette requête permet d’observer l’évolution des évaluations données aux établissements en comparant leur première et leur dernière note reçue, ainsi que les dates correspondantes. Pour chaque établissement, on retrouve son identifiant, son nom, la note de la première review, sa date, puis la note de la dernière review, et sa date.

Cette analyse met en lumière les trajectoires de réputation. Certains établissements, comme “H&M” ou “Iron Horse Auto Body”, commencent avec une note de 4.0 et conservent cette même note au fil des années, ce qui suggère une stabilité dans la qualité perçue. D’autres, comme “Helena Avenue Bakery”, débutent avec une note élevée, mais voient leur dernière note baisser, ce qui peut traduire une dégradation dans l’expérience client. À l’inverse, des établissements comme “Marty’s Pizza” ou “Rusty’s Pizza Parlor” reçoivent encore récemment des avis très positifs, ce qui peut indiquer une amélioration continue ou un maintien d’un haut niveau de satisfaction.

L’écart entre la première et la dernière note permet ainsi d’identifier les établissements dont la qualité perçue s’est améliorée, s’est détériorée, ou est restée stable au fil du temps. Ce type de requête est particulièrement utile pour détecter les évolutions de réputation sur le long terme.

4.5 Top 5 des établissements les plus populaires parmi les 10 catégories les plus fréquentes

business_id	name	city	stars	review_count	review_year	category
3MieDW5uihkPvXqnxAlGmg	SB Buggie	Santa Barbara	5.0	231	2016	Active Life
xuhCaQ1gnWyHY0tFSsrh2g	Santa Barbara Art Glass	Santa Barbara	5.0	204	2016	Active Life
YbnJYHNp_fHbI-hcFg48vQ	Santa Barbara Adventure Company	Santa Barbara	5.0	195	2016	Active Life
DboqYyH-S8pV6WxaF9Plow	Cal Coast Adventures	Santa Barbara	5.0	184	2016	Active Life
-mHLlBqekJe_VC61oul9yA	Segway of Santa Barbara	Santa Barbara	5.0	165	2016	Active Life
3MieDW5uihkPvXqnxAlGmg	SB Buggie	Santa Barbara	5.0	231	2017	Active Life
xuhCaQ1gnWyHY0tFSsrh2g	Santa Barbara Art Glass	Santa Barbara	5.0	204	2017	Active Life
YbnJYHNp_fHbI-hcFg48vQ	Santa Barbara Adventure Company	Santa Barbara	5.0	195	2017	Active Life
DboqYyH-S8pV6WxaF9Plow	Cal Coast Adventures	Santa Barbara	5.0	184	2017	Active Life
-mHLlBqekJe_VC61oul9yA	Segway of Santa Barbara	Santa Barbara	5.0	165	2017	Active Life
3MieDW5uihkPvXqnxAlGmg	SB Buggie	Santa Barbara	5.0	231	2018	Active Life
xuhCaQ1gnWyHY0tFSsrh2g	Santa Barbara Art Glass	Santa Barbara	5.0	204	2018	Active Life
YbnJYHNp_fHbI-hcFg48vQ	Santa Barbara Adventure Company	Santa Barbara	5.0	195	2018	Active Life
DboqYyH-S8pV6WxaF9Plow	Cal Coast Adventures	Santa Barbara	5.0	184	2018	Active Life
-mHLlBqekJe_VC61oul9yA	Segway of Santa Barbara	Santa Barbara	5.0	165	2018	Active Life
3MieDW5uihkPvXqnxAlGmg	SB Buggie	Santa Barbara	5.0	231	2019	Active Life
xuhCaQ1gnWyHY0tFSsrh2g	Santa Barbara Art Glass	Santa Barbara	5.0	204	2019	Active Life
YbnJYHNp_fHbI-hcFg48vQ	Santa Barbara Adventure Company	Santa Barbara	5.0	195	2019	Active Life
DboqYyH-S8pV6WxaF9Plow	Cal Coast Adventures	Santa Barbara	5.0	184	2019	Active Life
-mHLlBqekJe_VC61oul9yA	Segway of Santa Barbara	Santa Barbara	5.0	165	2019	Active Life
3MieDW5uihkPvXqnxAlGmg	SB Buggie	Santa Barbara	5.0	231	2020	Active Life
xuhCaQ1gnWyHY0tFSsrh2g	Santa Barbara Art Glass	Santa Barbara	5.0	204	2020	Active Life
YbnJYHNp_fHbI-hcFg48vQ	Santa Barbara Adventure Company	Santa Barbara	5.0	195	2020	Active Life
DboqYyH-S8pV6WxaF9Plow	Cal Coast Adventures	Santa Barbara	5.0	184	2020	Active Life
-mHLlBqekJe_VC61oul9yA	Segway of Santa Barbara	Santa Barbara	5.0	165	2020	Active Life
jMzcn59A2OYZ3zv2BoDdOw	Fernando's Smog Check	Goleta	5.0	193	2016	Automotive
l2CuGoQb4ZFFQNnivegqqA	Super Value Smog	Santa Barbara	5.0	123	2016	Automotive

Cette requête identifie, pour chaque année, les cinq établissements ayant reçu le plus grand nombre de reviews, parmi les dix catégories les plus représentées dans la base de données. Elle permet de croiser popularité, mesurée par le nombre d’avis, et appartenance à une catégorie très présente dans la base.

On constate que la catégorie Active Life domine largement cette sélection, avec des établissements situés principalement à Santa Barbara. Des noms comme SB Buggie, Santa Barbara Adventure Company, Cal Coast Adventures, Segway of Santa Barbara ou Santa Barbara Art Glass apparaissent de manière répétée chaque année. Ces établissements ont tous obtenu une note de 5.0, ce qui montre un niveau élevé de satisfaction client combiné à une grande visibilité.

Cette constance dans le classement reflète une forte réputation et une activité soutenue. En 2016, deux établissements de la catégorie Automotive figurent également parmi les plus populaires, ce qui montre que d’autres catégories peuvent ponctuellement se démarquer.

Les 10 catégories les plus populaires sont :

Cette analyse permet de repérer les établissements les plus influents dans les catégories majeures et de suivre leur évolution d'année en année.

5. Rapport d'analyse

5.1 Dashboard "Categories"

Ce premier rapport d'analyse vise à étudier la **répartition des commerces par catégorie** au sein de la base de données Yelp, filtrée sur l'État de la Californie. Il permet d'avoir une vue synthétique de l'offre commerciale en fonction des catégories d'activités renseignées par Yelp.
L'objectif est de comprendre :

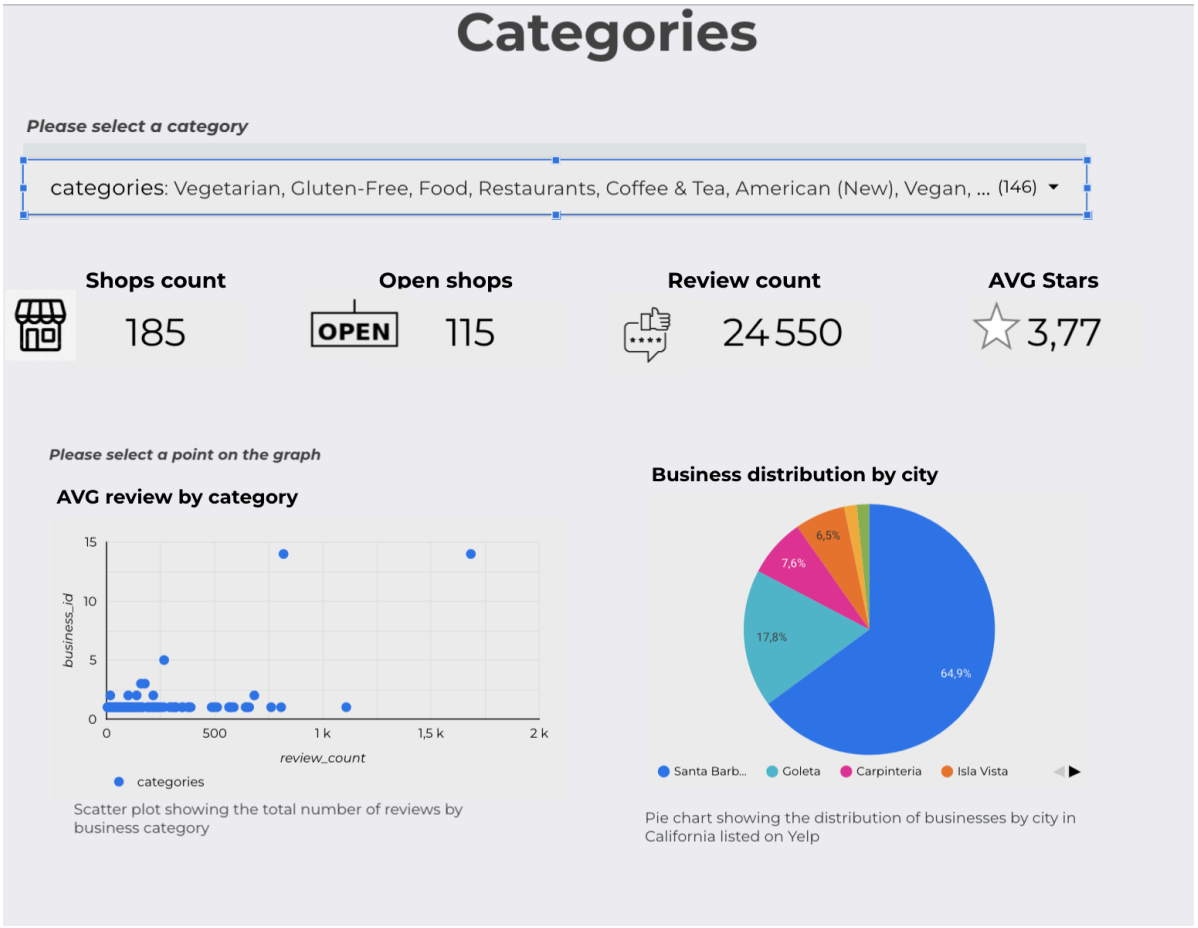
- La quantité d'établissements présents dans chaque catégorie
- La performance globale de ces catégories en termes de volume d'avis et de notes
- Leur répartition géographique dans les principales villes

Le dashboard est interactif et permet de filtrer par catégorie, avec les indicateurs suivants mis en avant :

Indicateur	**Description**
Shops count	Nombre total d'établissements dans la catégorie sélectionnée
Open shops	Nombre d'établissements actuellement ouverts (is_open = 1)
Review count	Nombre total d'avis associés à ces établissements
AVG Stars	Note moyenne globale sur 5 attribuée aux établissements

Deux visualisations principales complètent cette vue :

- Un nuage de points (scatter plot) qui permet de visualiser la densité des reviews par business dans chaque catégorie
- Un camembert représentant la répartition des établissements par ville (ex : Santa Barbara, Goleta…)



L'exemple affiché ici concerne la sélection de toutes les catégories contenant le mot "Coffee", comme *Coffee & Tea*, *Vegan*, *Restaurants*, ou encore *American (New)*.
On observe un total de 185 établissements, dont 115 sont ouverts, soit environ 62 % d'activité en cours.
Ces établissements ont généré 24 550 reviews, avec une note moyenne de 3,77/5.
La majorité des établissements de cette catégorie sont situés à Santa Barbara (64,9 %), suivis par Goleta (17,8 %) et Carpinteria (7,6 %).
Le scatter plot révèle que quelques établissements très populaires peuvent recevoir jusqu'à 2 000 reviews, tandis que la majorité restent en dessous de 500.

5.2. Dashboard "Shops"

Ce second rapport d'analyse se concentre sur les commerces pris individuellement, en permettant de filtrer dynamiquement sur un nom de commerce précis (ex. : Starbucks) ainsi que sur la ville. Il permet de zoomer sur un établissement particulier, d'étudier son impact sur Yelp et de comparer ses performances selon les villes.

L'objectif est de comprendre :

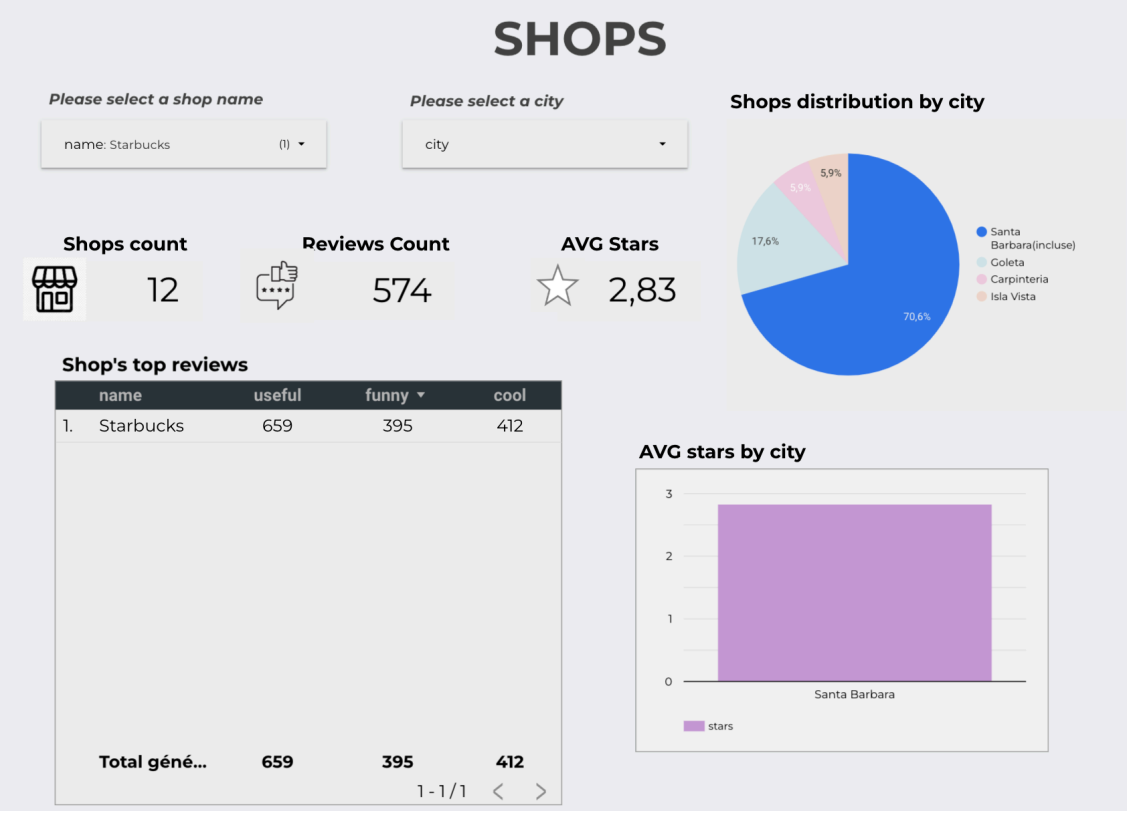
- Le nombre de points de vente correspondant à un commerce donné

- Le niveau d’activité communautaire autour de ce commerce (volume d’avis, utilité perçue, etc.)
- La répartition géographique des établissements de la marque
  La satisfaction client moyenne par ville

Le tableau de bord est interactif : il permet de filtrer les résultats par nom d’établissement (name) et par ville (city). Il met en avant trois indicateurs clés :

Indicateur	Description
Shops count	Nombre total d’établissements correspondant au nom sélectionné
Reviews count	Nombre total d’avis laissés pour ce commerce
AVG Stars	Note moyenne globale sur 5 attribuée aux établissements concernés

Trois visualisations principales viennent enrichir cette analyse. Un camembert affiche la répartition géographique des établissements sélectionnés, ce qui permet d’identifier les villes les plus représentées. Un graphique à barres compare la note moyenne obtenue dans chaque ville, mettant en lumière d’éventuelles disparités de satisfaction client selon les zones. Enfin, un tableau de synthèse liste les principales interactions autour du commerce : nombre de votes "useful", "funny" et "cool", reflétant l’engagement des utilisateurs autour de ses établissements.



L’exemple affiché ici concerne l’établissement Starbucks, filtré à partir du nom du commerce via le sélecteur interactif.
On observe un total de 12 établissements Starbucks enregistrés dans la région étudiée, ayant généré 574 avis au total.
La note moyenne attribuée aux Starbucks est faible, avec un score de 2,83 / 5, ce qui peut montrer une insatisfaction modérée des clients.
Le camembert de répartition géographique montre que la majorité des établissements Starbucks (70,6 %) sont situés à Santa Barbara, suivis de Goleta (17,6 %) et d’autres villes comme Carpinteria ou Isla Vista.
Le tableau récapitulatif des top reviews indique que Starbucks a cumulé 659 avis jugés utiles, 395 jugés drôles, et 412 cool, ce qui montre une forte interaction communautaire, même si la note moyenne reste en dessous de la moyenne observée sur d’autres commerces.
Enfin, le graphe de répartition des notes moyennes par ville montre peu de variation dans la satisfaction client, avec une moyenne générale légèrement en dessous de 3 à Santa Barbara.

5.3 Dashboard “Popular User”

Ce troisième rapport d’analyse se concentre sur les utilisateurs populaires de Yelp, en filtrant dynamiquement les données selon un commerce spécifique (par exemple Starbucks), la note attribuée, ainsi que la ville. Il permet ainsi de croiser les performances d’un établissement donné avec le comportement des utilisateurs les plus suivis ou les plus influents de la plateforme.

Un filtrage a été établi en amont pour ne retenir que les utilisateurs ayant plus de 100 fans et ayant rédigé plus de 100 avis. Ce critère permet de se concentrer uniquement sur les profils les plus actifs et les plus reconnus par la communauté Yelp.

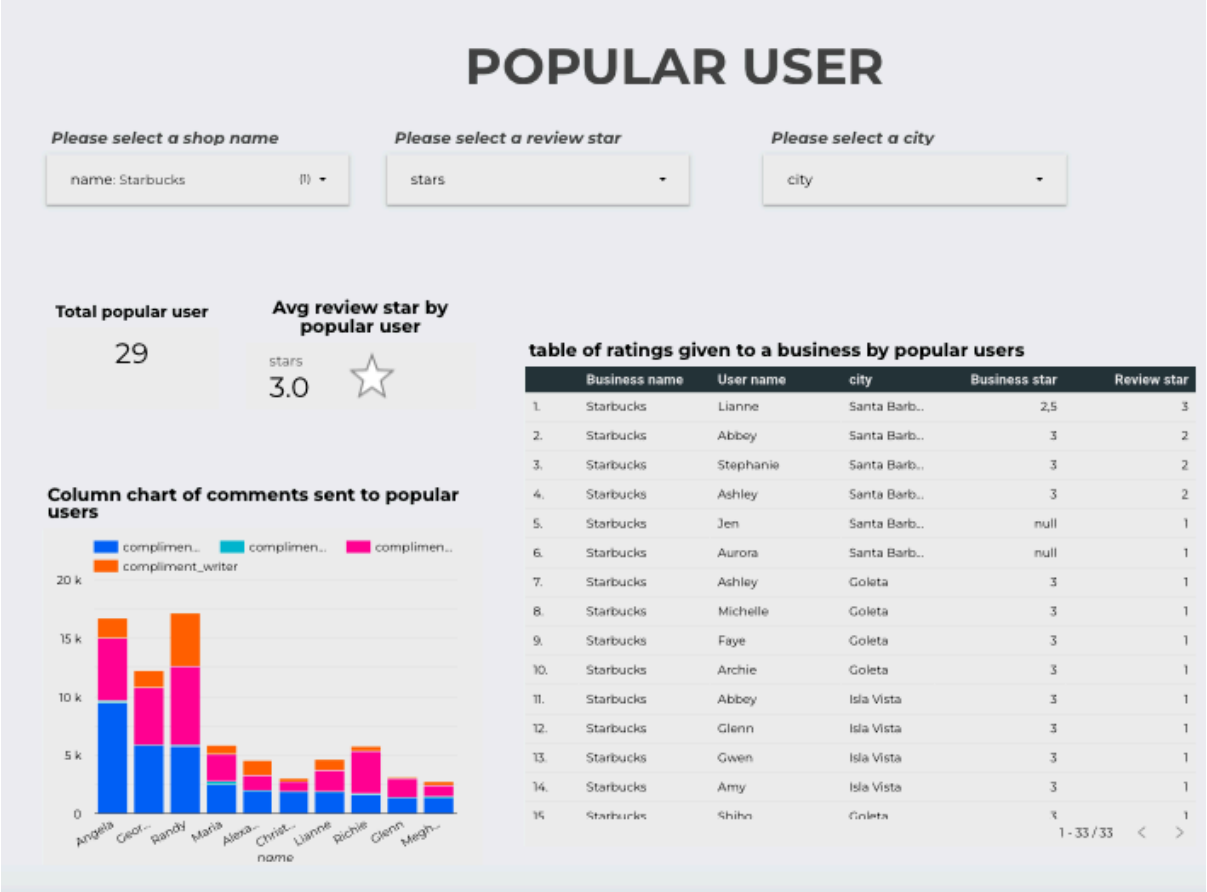
L’objectif principal est de comprendre :

- le volume d’interactions entre un commerce donné et les utilisateurs populaires
- le niveau moyen de satisfaction exprimé par ces utilisateurs
- les écarts entre les notes des utilisateurs influents et la note moyenne de l’établissement
- le type de commentaires adressés à ces utilisateurs

Le tableau de bord est interactif. Il propose trois filtres permettant de sélectionner le nom d’un établissement, une note spécifique et une ville. Trois indicateurs clés sont mis en avant :

Indicateurs	Descritption
Total popular user	nombre d'utilisateurs populaires ayant laissé un avis sur le commerce sélectionné
Avg review star by popular user	note moyenne donnée par ces utilisateurs
Table of ratings	tableau détaillant les avis des utilisateurs populaires avec comparaison entre la note qu'ils ont laissé et la note moyenne du commerce

Deux visualisations principales complètent l'analyse. Un graphique en barres montre la répartition des types de compliments reçus par les utilisateurs populaires ayant interagi avec le commerce sélectionné. Cela donne un aperçu qualitatif de leur profil, comme leur humour, leur pertinence ou leur style. Un tableau de détail liste chaque avis, avec le nom du commerce, le nom de l'utilisateur, la ville, la note moyenne de l'établissement et la note attribuée par l'utilisateur populaire.



Dans l'exemple affiché, l'établissement sélectionné est Starbucks. On observe que 29 utilisateurs populaires ont laissé un avis, avec une note moyenne de 3,0 sur 5. Cette note est relativement modérée, ce qui peut indiquer une certaine exigence de la part de ces

profils ou une expérience client perfectible. Le tableau indique que plusieurs avis sont plus généreux que la note globale, mais d'autres sont en décalage. Les données de compliments révèlent que certains utilisateurs reçoivent une grande quantité d'interactions positives, ce qui peut renforcer leur crédibilité au sein de la plateforme.

6.Difficulté

Tout au long du projet, nous avons été confrontés à plusieurs difficultés techniques liées principalement aux limites de nos machines et aux outils disponibles. Nos ordinateurs n'étaient pas très puissants, ce qui a rendu le traitement des données initiales ainsi que l'exécution des scripts relativement longs et parfois instables.

L'un des premiers obstacles a été lié à Power BI. Étant tous les deux sur Mac, nous n'avons pas pu installer la version classique du logiciel. La version en ligne, bien plus limitée, ne permettait d'importer que de petits fichiers CSV. Nous avons même tenté d'installer une machine virtuelle Windows pour contourner ce problème, mais nos ordinateurs étaient trop lents pour faire fonctionner Power BI dans de bonnes conditions. L'environnement devenait inutilisable, ce qui nous a poussés à chercher une alternative.

Nous nous sommes donc tournés vers Looker Studio (anciennement Google Data Studio), qui propose un accès en ligne et gratuit. Cependant, cet outil présente également plusieurs contraintes. Il ne permet pas de se connecter directement à un serveur SQL local, ce qui nous a obligés à extraire manuellement nos données sous forme de fichiers CSV, puis à les déposer dans Google Drive. Ce processus est non seulement plus long, mais limite aussi fortement les volumes de données exploitables.

Looker est par ailleurs un outil assez lent, surtout lorsqu'il est utilisé de manière prolongée ou avec des jeux de données trop importants. Nous avons dû faire des extractions partielles, ce qui a empêché une exploitation complète de notre base, notamment la table datetime que nous n'avons pas pu intégrer. Contrairement à Power BI, Looker ne permet pas l'utilisation de requêtes complexes (comme en SQL ou en DAX) de manière native, ce qui réduit fortement les possibilités d'analyse avancée.

En parallèle, l'utilisation de Pentaho a également montré ses limites. Certaines opérations, comme le découpage ou l'explosion de champs (par exemple pour la colonne catégorie), ne peuvent pas être réalisées facilement. Les scripts disponibles dans Pentaho sont eux-mêmes limités, ce qui oblige parfois à faire les manipulations à la main. Compte tenu du volume important de données, cela n'était tout simplement pas envisageable.

7. Conclusion

Ce projet nous a permis de mettre en œuvre l'ensemble des étapes d'un processus décisionnel autour d'un jeu de données réel, en partant de l'extraction brute jusqu'à la visualisation interactive. À travers la base Yelp, nous avons exploré des problématiques

concrètes liées à la satisfaction client, à l'influence des utilisateurs, et à la notoriété des commerces locaux.

Malgré les contraintes matérielles et techniques rencontrées, nous avons su adapter nos outils et notre méthodologie pour mener à bien le projet. La limitation de Power BI sur Mac, les restrictions de Looker Studio, les lourdeurs de traitement dans Pentaho, ainsi que les contraintes de volume nous ont obligés à faire preuve de flexibilité et à ajuster nos choix techniques. Ces difficultés ont été formatrices, car elles nous ont poussés à mieux comprendre les limites de chaque outil et à trouver des solutions alternatives viables, en particulier via l'extraction ciblée de données et le travail collaboratif.

Nous avons ainsi pu construire un entrepôt de données structuré, concevoir des requêtes pertinentes pour explorer les dynamiques présentes dans la base, et produire des tableaux de bord interactifs exploitables.