Projet Transversal MBDIA | SQL

Travail réalisé par :

Hamza NAMOUCHI

Encadré par :

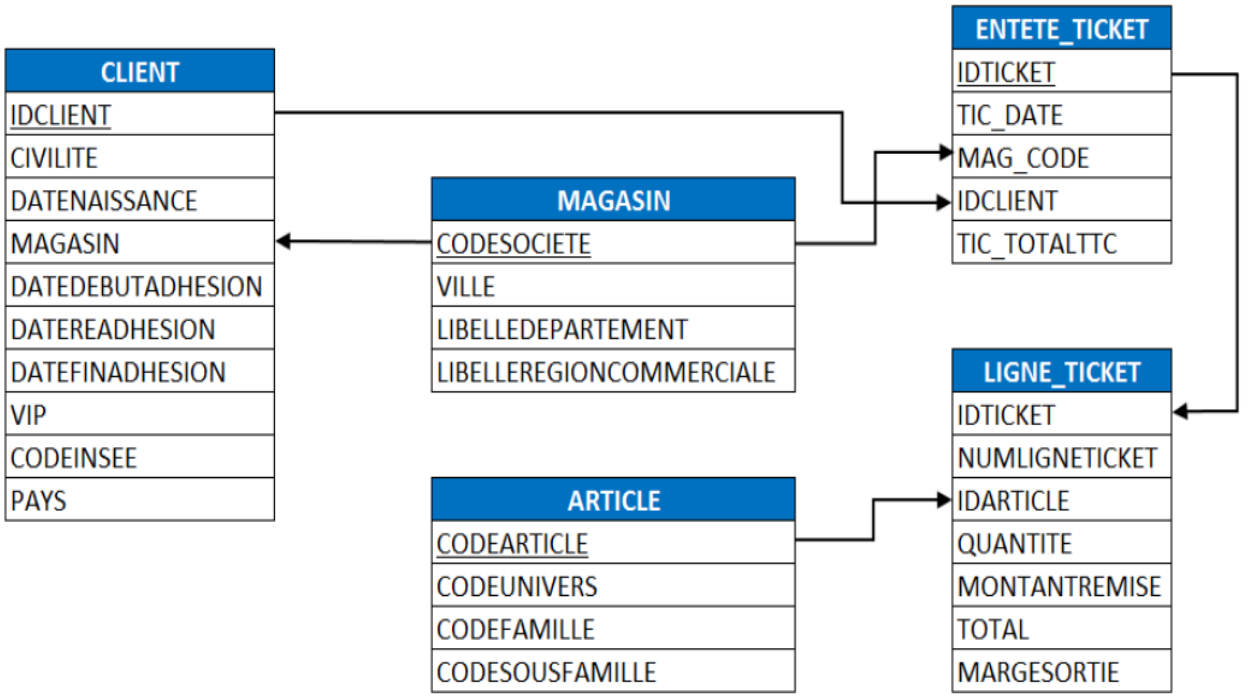
Mme Juliette LEMAINS

MBAESG

BIG DATA & INTELLIGENCE ARTIFICIELLE

Contexte

Une société X a envoyé ses données client ainsi que les achats sur l'année N-2 (2016) et N-1 (2017).
Nos données se représentent sous la forme de 5 tables principales (Client, ref_article, ref_magasin, entête_ticket et lignes_ticket) dont le schéma relationnel est le suivant :



On fixe le **01/01/2018** comme date d'étude pour ne pas biaiser les données (âge, période d'adhésion et type de clients).

On a commencé par l'importation de nos tables à travers le script donné dans le fichier SCRIPT_IMPORT_DATA_TRANSVERSE.sql.

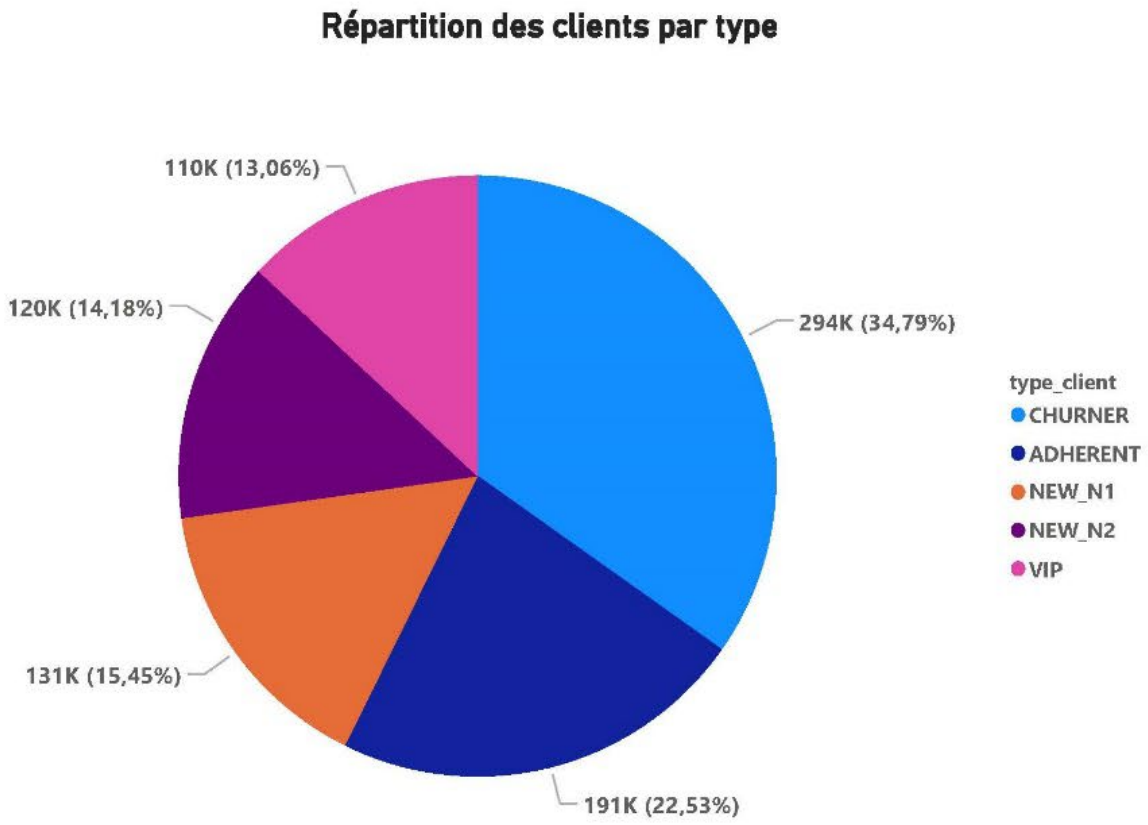
Tout au long du projet, on était amené à créer des subsets avec moins de lignes pour le cas des tables entete_ticket et lignes_ticket vu qu'elles sont trop volumineuses et on a parfois créé des copies de table pour pouvoir avancer avec la copie et laisser la table originale sans modification.

Notre outil de Data Viz était **Power BI**.

1) Etude globale

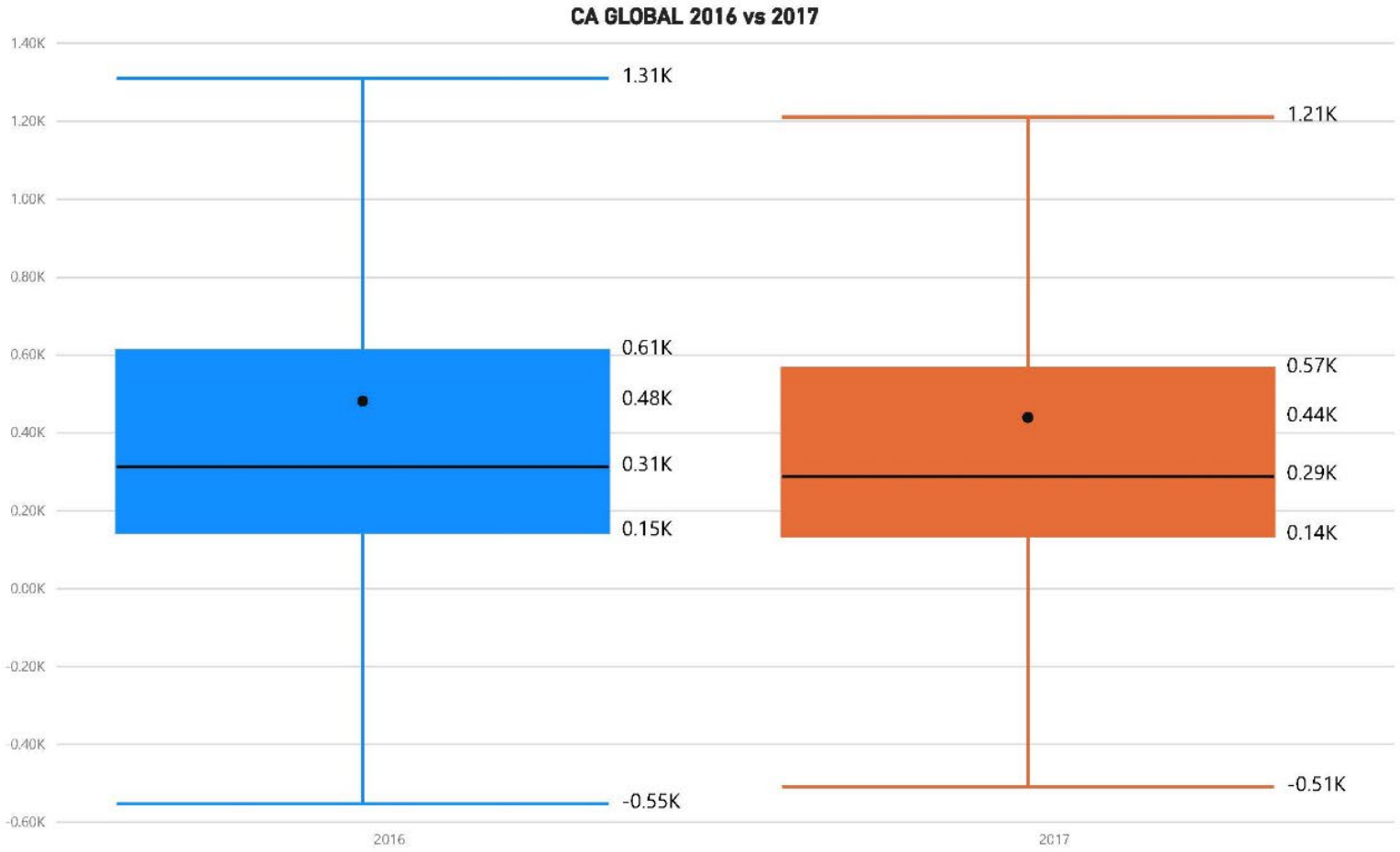
a. Répartition des clients par type

On a testé si chaque client est-il VIP ou pas. Puis, on a construit une nouvelle colonne 'Type_client' qui répartit les clients en VIP, NEW_N2 (adhésion en 2016), NEW_N1 (adhésion en 2017), adhérent (toujours présent dans la base à la date d'étude) ou bien Churner (client parti avant la date d'étude).



b. Comportement du chiffre d'affaires par client 2016 vs 2017

On a requêté notre base pour avoir le CA (somme de 'tic_totalttc' de la table entete_ticket) par année par client.
La boite à moustache ci-dessous nous montre le comportement du CA durant 2016 contre 2017.
NB : Whisker Type = 1.5 IQR (IQR = Interquartile Range = $Q_3 - Q_1$)



c. Répartition des clients par âge x sexe

Après la construction de la colonne ‘Âge’ et la correction de la colonne ‘Civilité’, il s’est avéré qu’il y a plus de 300K clients sans âge (valeur non renseignée).

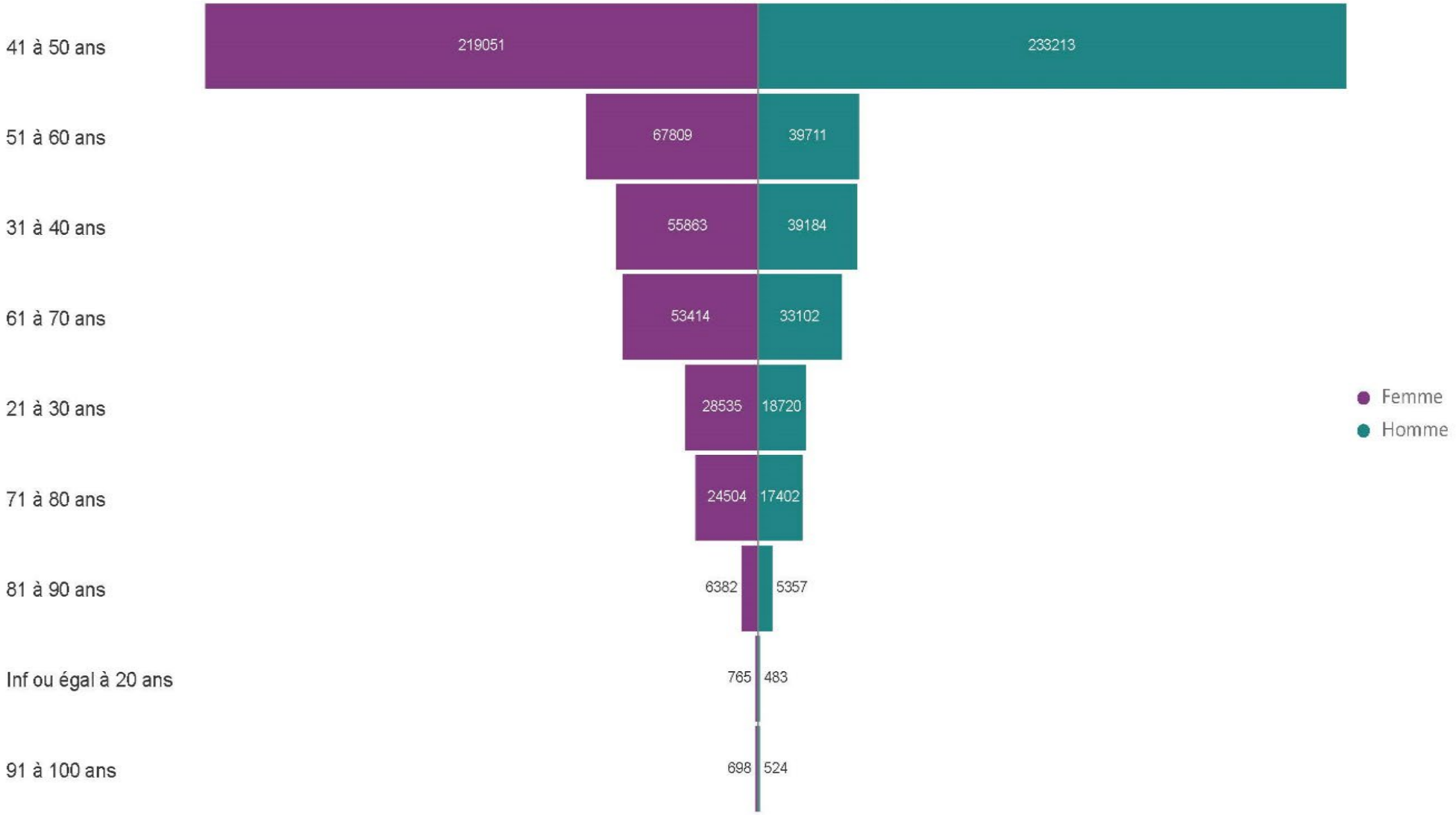
C’est une partie de clients qu’on ne peut pas l’ignorer donc on a choisi de remplacer leurs âges par la moyenne des âges dans toute la table client qui est **50 ans**.

Cela étant, on a procédé notre étude sur l’intervalle d’âge [18, 100ans] qui nous parait tout à fait logique.

On suppose que l’adhésion vise les clients majeurs et que 100 ans est une borne largement supérieure à l’espérance de vie en France en 2018 (85,3 ans pour les femmes contre 79,4 pour les hommes selon La Direction de la recherche, des études, de l’évaluation et des Statistiques).

À la fin de cette partie, on a construit une colonne ‘tranche_age’ qui répartit les clients sur 9 tranches d’âges (inf ou égal à 20ans, 21 à 30 ans, etc…) vu que ce n’était pas évident de représenter 83 lignes d’âges ainsi que le sexe dans un seul graphique.

Répartition des clients par tranche d'âge x sexe



2) Etude par magasin

a. Résultat par magasin

À ce niveau d'étude, on s'intéressait à construire deux colonnes supplémentaires ('actif2016' et 'actif2017') qui vont nous renseigner si le client est présent ou non sur l'année en question.

Pour cette partie, on a fait 3 requêtes ;

- Première requête qui donne le nombre des clients, les clients actifs sur 2016 vs 2017 ainsi que l'évolution des clients actifs sur ces deux années répartis par magasin.
- Deuxième requête qui nous renseigne sur le CA 2016 par magasin.
- Troisième requête qui nous renseigne sur le CA 2017 par magasin.

Notre logiciel de Data Viz s'occupera du reste telles que la fusion des trois tables, la construction de la colonne 'Différence CA 2016vs2017' ainsi que la colonne 'indice_évolution'.

Le résultat se représente sous la forme du tableau suivant :

Répartition des clients par magasin

magasin	Somme de count	Somme de sum2016	Somme de sum2017	Somme de évolution_clients	Tot_ttc_2016	Somme de Tot_ttc_2017	Somme de Diff tot_ttc_2016vs2017	Indice_évolution
BAR	1320	1091	1143	4,77 %	455 609,35	511 120,31	55 510,96	Positif
CAG	9027	5744	8024	39,69 %	1 697 422,14	1 792 635,33	95 213,19	Positif
DUM	6757	5792	5858	1,14 %	2 208 888,88	2 458 296,77	249 407,89	Positif
EPN	8790	6861	7516	9,55 %	1 748 055,54	1 821 401,02	73 345,48	Positif
MOB	18912	16293	16364	0,44 %	5 977 004,79	6 384 418,04	407 413,25	Positif
RMA	6481	2964	6479	118,59 %	1 108 075,43	2 656 633,94	1 548 558,51	Positif
SAL	3120	2469	2547	3,16 %	605 824,60	664 168,62	58 344,02	Positif
SNO	10819	8155	9493	16,41 %	2 071 762,57	2 292 897,87	221 135,30	Positif
SSM	23552	19119	19579	2,41 %	6 413 411,93	6 842 402,95	428 991,02	Positif
VIT	14653	12032	12033	0,01 %	3 007 212,43	3 089 915,26	82 702,83	Positif
VIV	3629	2858	3008	5,25 %	923 415,88	973 409,27	49 993,39	Positif
ALB	12814	11019	10516	-4,56 %	2 960 491,65	3 049 735,09	89 243,44	Moyen
ALM	16471	13958	13580	-2,71 %	4 036 788,61	4 113 500,34	76 711,73	Moyen
BEA	20704	17270	17667	2,30 %	5 739 924,27	5 566 069,91	-173 854,36	Moyen
BEC	7963	6620	6296	-4,89 %	1 470 501,12	1 485 815,73	15 314,61	Moyen
BLA	11721	9573	9109	-4,85 %	2 143 924,90	2 255 599,64	111 674,74	Moyen
BRE	10089	8119	8079	-0,49 %	2 089 464,49	2 121 890,22	32 425,73	Moyen
CLA	23659	19460	19753	1,51 %	5 847 741,89	5 639 065,55	-208 676,34	Moyen
ECU	15398	12906	12895	-0,09 %	4 112 227,44	4 274 237,60	162 010,16	Moyen
EST	1273	579	1125	94,30 %	0,00	0,00	0,00	Moyen
HAG	4776	4052	3849	-5,01 %	801 071,08	848 128,87	47 057,79	Moyen
IAB	16872	13831	13730	-0,73 %	3 222 142,17	3 239 040,89	16 898,72	Moyen
MAC	14628	11985	11654	-2,76 %	2 798 360,47	2 821 781,51	23 421,04	Moyen
MOU	24356	19646	19875	1,17 %	6 349 088,35	6 045 145,65	-303 942,70	Moyen

magasin	Somme de count	Somme de sum2016	Somme de sum2017	Somme de évolution_clients	Tot_ttc_2016	Somme de Tot_ttc_2017	Somme de Diff tot_ttc_2016vs2017	Indice_évolution
OBE	9696	8311	8049	-3,15 %	1 793 099,24	1 796 715,77	3 616,53	Moyen
PRI	26909	22374	21275	-4,91 %	5 528 750,33	5 557 473,28	28 722,95	Moyen
QUE	11686	9503	9201	-3,18 %	1 783 723,09	1 898 024,29	114 301,20	Moyen
SEY	25931	21931	21420	-2,33 %	4 336 406,70	4 400 118,80	63 712,10	Moyen
SJV	11060	8986	8952	-0,38 %	2 254 727,34	2 414 626,02	159 898,68	Moyen
SLM	7889	6471	6421	-0,77 %	1 366 815,64	1 431 044,59	64 228,95	Moyen
SMR	12527	10936	10331	-5,53 %	3 044 483,67	3 063 414,70	18 931,03	Moyen
SUR	23741	19622	18719	-4,60 %	4 807 772,47	4 966 887,74	159 115,27	Moyen
THO	10518	9025	8778	-2,74 %	2 897 265,27	2 979 303,98	82 038,71	Moyen
VIC	11546	9494	9276	-2,30 %	2 114 766,50	2 133 378,76	18 612,26	Moyen
AVI	15424	13034	12720	-2,41 %	4 065 939,61	3 951 027,58	-114 912,03	Négatif
BSN	10298	8741	6946	-20,54 %	2 020 922,33	1 665 792,86	-355 129,47	Négatif
CLI	4452	4137	3240	-21,68 %	918 075,64	794 102,31	-123 973,33	Négatif
DIJ	12582	10723	10087	-5,93 %	3 043 839,41	2 900 305,72	-143 533,69	Négatif
FEG	10545	9048	8432	-6,81 %	2 347 398,26	2 161 775,78	-185 622,48	Négatif
FRV	15352	13089	12356	-5,60 %	3 922 133,37	3 764 133,92	-157 999,45	Négatif
GAI	16092	14098	13168	-6,60 %	6 143 169,42	5 666 500,73	-476 668,69	Négatif
GAP	8131	7055	6754	-4,27 %	2 080 513,48	1 926 180,97	-154 332,51	Négatif
GEX	19261	16991	16330	-3,89 %	7 388 908,23	6 957 685,71	-431 222,52	Négatif
HEI	23480	19901	19232	-3,36 %	5 834 749,58	5 796 841,23	-37 908,35	Négatif
LAB	11041	9121	8608	-5,62 %	2 053 101,79	2 021 049,86	-32 051,93	Négatif
MAN	11335	9993	8981	-10,13 %	2 719 688,26	2 461 591,49	-258 096,77	Négatif
MET	17962	15783	14736	-6,63 %	4 826 410,13	4 633 844,77	-192 565,36	Négatif
MUL	14236	12507	11205	-10,41 %	3 332 191,16	3 172 041,51	-160 149,65	Négatif
NEV	7004	5758	5702	-0,97 %	1 987 411,68	1 769 606,94	-217 804,74	Négatif
ORL	9873	8267	8118	-1,80 %	2 695 002,71	2 467 711,98	-227 290,73	Négatif
PEG	10996	9378	8689	-7,35 %	2 118 864,43	2 044 229,97	-74 634,46	Négatif
PEP	7595	6250	5730	-8,32 %	1 315 595,21	1 225 988,65	-89 606,56	Négatif
POC	10169	8675	7959	-8,25 %	2 406 910,18	2 235 013,54	-171 896,64	Négatif
PON	11772	10007	9187	-8,19 %	2 577 362,88	2 363 867,85	-213 495,03	Négatif
RAV	11374	9867	8939	-9,41 %	2 625 700,87	2 302 863,76	-322 837,11	Négatif
SCH	5521	5082	4316	-15,07 %	979 934,81	624 613,92	-355 320,89	Négatif
SEM	10347	8868	7863	-11,33 %	1 959 365,37	1 781 800,21	-177 565,16	Négatif
SGL	8837	7619	6538	-14,19 %	1 575 548,19	1 450 191,34	-125 356,85	Négatif
SMA	5406	4808	4024	-16,31 %	1 741 264,00	1 490 879,74	-250 384,26	Négatif
STE	11986	10373	9417	-9,22 %	2 526 952,24	2 279 578,34	-247 373,90	Négatif
STR	2943	2468	2271	-7,98 %	670 530,02	642 991,56	-27 538,46	Négatif
VAL	16674	14424	13176	-8,65 %	3 459 987,91	2 986 765,69	-473 222,22	Négatif
VAR	12934	11898	10581	-11,07 %	3 150 351,97	2 701 641,48	-448 710,49	Négatif
VEN	10528	9108	8675	-4,75 %	3 354 293,20	3 182 815,70	-171 477,50	Négatif
VIB	24607	20269	19281	-4,87 %	4 402 349,51	4 323 459,04	-78 890,47	Négatif
VIF	14169	12178	11330	-6,96 %	2 709 178,85	2 586 192,37	-122 986,48	Négatif
VLG	22504	19541	18577	-4,93 %	6 212 516,50	5 653 657,78	-558 858,72	Négatif
Total	**844717**	**708108**	**685762**	**-28,83 %**		**193 575 042,61**	**-3 307 338,82**	

b. Distance Client / Magasin

On commence par télécharger les données GPS des villes et code-insee pour pouvoir calculer la distance client-magasin. On garde que 4 colonnes ‘code_insee’, ‘commune’, ‘lat’ et ‘long’.

Chaque point GPS se compose de coordonnées Latitude et Longitude donc il fallait construire deux colonnes ‘lat_client’ et ‘long_client’ pour chaque client ainsi que deux colonnes ‘lat_mag’ et ‘long_mag’ pour chaque magasin.

Coordonnées magasin :

On a dû corriger manuellement quelques valeurs de la colonne ‘ville’ de la table ref_magasin pour que ce soit adéquat avec les communes de notre base GPS (voir script).

Coordonnées clients :

À cette étape, on a trouvé 121.518 lignes clients (14,38% de la base) sans coordonnées gps parce leur code insee soit il n’existe pas soit il est mal saisi.

Pour 94.158 clients (11,15% de la base), le code insee est non valide, notre premier réflexe était de supprimer le ‘0’ qui existe au début de plusieurs valeurs de codeinsee (87.309 cas).

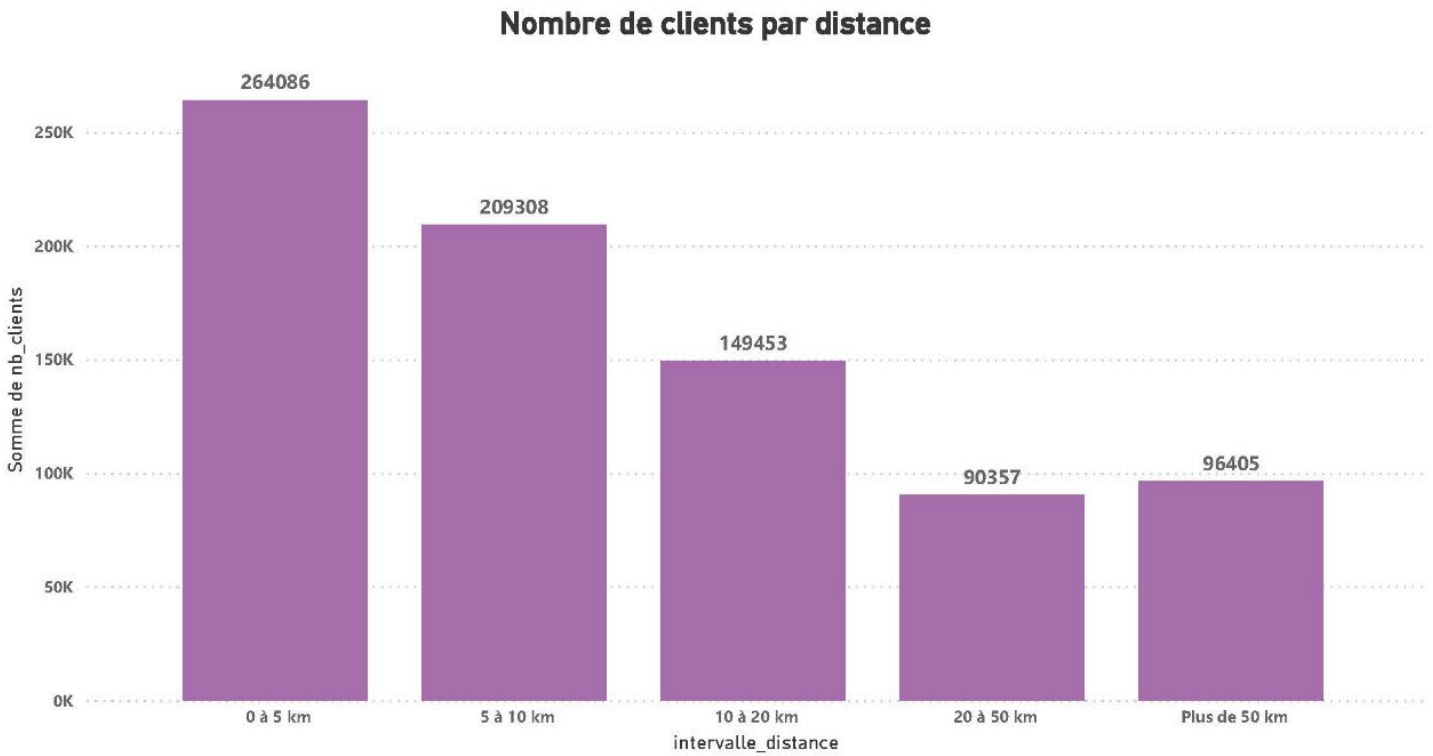
Pour finir, il nous reste que 35108 clients avec des codes insee invalides ou manquants (4,15% de la base qu’on va éventuellement les éliminer de notre étude.

Après ce traitement, on importe les coordonnées des magasins sur la table client et on procède à la construction de la colonne ‘distance’ qui prend en entrée 4 variables (‘lat_client’, ‘long_client’, ‘lat_mag’ et ‘long_mag’) et renvoie comme résultat la distance en Km selon la formule suivante :

{Distance = ACOS(SIN (RADIANS (lat_client))*SIN(RADIANS(lat_mag))+COS(RADIANS(lat_client))*COS(RADIANS(lat_mag))*COS(RADIANS(long_client-long_mag)))*6371 }.

On finit avec l’ajout d’une colonne ‘Intervalle_distance’ dans le but de répartir nos distances sous forme de 5 intervalles (0 à 5km, 5 à 10km, 10 à 20km, 20 à 50km et plus de 50km).

Nos clients sont répartis par distance ci-après :



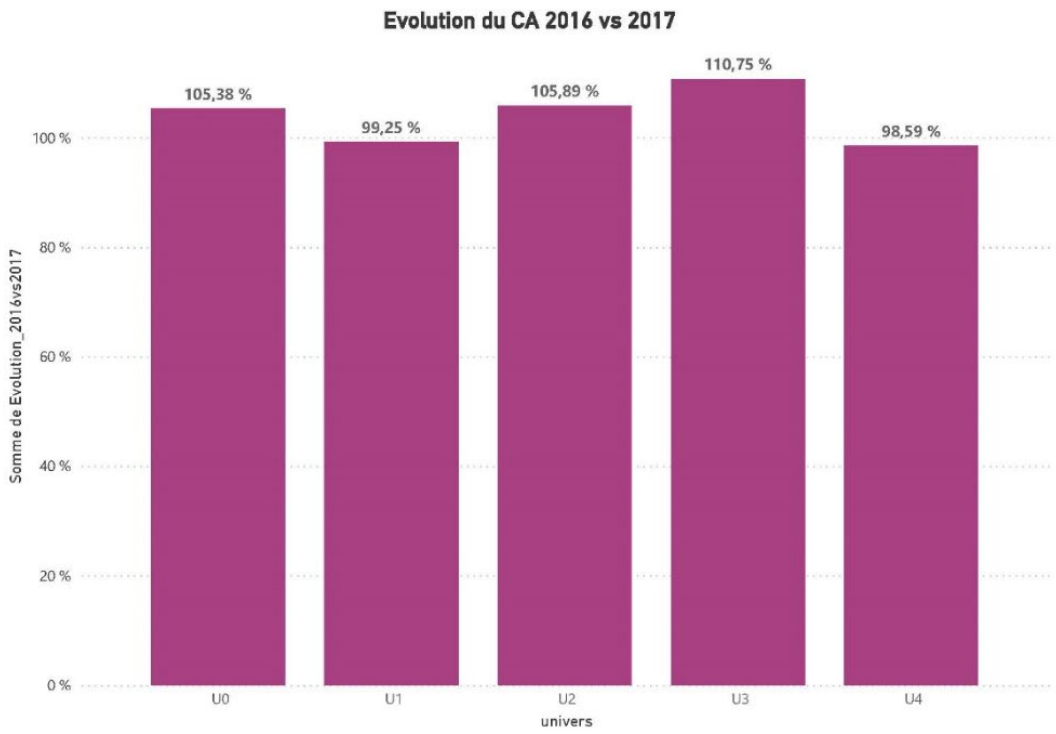
3) Etude par univers

a. Etude par univers

À partir des 3 tables ref_article, entete_ticket et lignes_ticket et en utilisant deux Innerjoin, on a eu le CA réparti par univers sur chacune des années 2016 et 2017 ;

	année double precision	univers character varying (15)	ca double precision
1	2016	COUPON	20856916.5299963
2	2016	U0	55234976.9400012
3	2016	U1	617844461.369420
4	2016	U2	191011310.910031
5	2016	U3	221353867.580028
6	2016	U4	427084699.780045
7	2017	U0	52414068.7000016
8	2017	U1	622514691.519361
9	2017	U2	180391397.900021
10	2017	U3	199862140.400002
11	2017	U4	433208279.360068

Dans le but d'optimiser le temps du travail, on a pivoté ce tableau par rapport à la colonne 'Année' sur notre outil de data viz. Puis, on a calculé l'évolution 2016/2017 du CA ce qui nous a généré l'histogramme suivant ;



b. Top par univers

Pour cette dernière partie, on a construit une table avec toutes les familles par univers avec leurs chiffres d'affaires, il s'agit d'une table de 25 lignes contenant toutes les 25 familles possibles et leurs CA listées par univers sous la forme suivante ;

	univers character varying (15)	famille character varying (15)	ca double precision
1	COUPON	COUPON	20856916.529996574
2	U0	160	107164665.71002889
3	U0	230	31611.57000000011
4	U0	400	10902.5
5	U0	900	441865.86000000004
6	U1	010	148837506.1800139
7	U1	020	418428107.91022515
8	U1	030	98429079.87999614
9	U1	040	55341703.12998754

Total rows: 25 of 25 Query complete 00:00:00.172

On s'intéresse après à sélectionner les top 5 familles par univers.

	univers character varying (15)	famille character varying (15)	ca double precision
1	U0	160	107164665.71002889
2	U1	020	418428107.91022515
3	U2	090	151333903.5900073
4	U3	120	366508301.0905412
5	U4	220	826328750.4096882

Ce tableau nous permet de représenter nos 5 top familles comme suit ;