مدرسة علوم المعلومات
ⵜⴰⵙⴷⴰⵡⵉⵜ ⵏ ⵜⵎⵓⵙⵙⵏⵉⵏ ⴷ ⵜⵏⵢⵎⵉⵙⵏ
ECOLE DES SCIENCES
DE L'INFORMATION

Rapport Du Projet Data Warehouse

BI Solution for Telecom Customer Churn

Réalisé par : HILMI ELMEHDI

Professeur : Mme Rhanoui

1- Domaine :

Télécommunication.

2- Contexte fonctionnel :

Toute entreprise de télécommunication souhaite maximiser le nombre de clients. Pour atteindre cet objectif, il est important non seulement d'essayer d'en attirer de nouveaux, mais aussi de conserver ceux qui existent déjà. Fidéliser un client coûtera moins cher à l'entreprise que d'en attirer un nouveau. De plus, un nouveau client peut être faiblement intéressé par les services aux entreprises et il sera difficile de travailler avec lui, alors que les anciens clients disposent déjà des données nécessaires sur l'interaction avec le service.

Besoin d'analyse :

Le besoin de cette étude est de :

- Comprendre les raisons de résiliations des clients pour améliorer la qualité des offres et le service clientèle.
- Analyser, de décrire et de prédire les données liées au désabonnement des clients afin de faire des tableaux de bord analytiques et prévisions à l'avenir pour faciliter la tâche de prise de décision pour la compagnie de télécommunication.

3- Jeu de données utilisés :

« IBM Teclo_Churn », Un ensemble de six fichiers Excel qui contiennent les informations sur une société de télécommunication fictive Telco qui a fourni des services de téléphonie résidentielle et Internet à 7043 clients en Californie au troisième trimestre de l'année 2020. Des informations sur les clients désabonnés, restés et ceux qui viennent de s'abonner. D'autres informations démographiques et géographiques importantes sont incluses pour chaque client, le type de contrat, les services choisis, le statut de désabonnement…etc

Et aussi d'autres fichiers, un qui s'agit d'une calendrier de 2020 et l'autre qui contient des données sur tous les Zip Code de Californie.

- Telco_customer_churn_demographics.xlsx
- Telco_customer_churn_location.xlsx
- Telco_customer_churn_location.xlsx
- Telco_customer_churn_population.xlsx
- Telco_customer_churn_services.xlsx
- Telco_customer_churn_status.xlsx
- DimDates.xlsx
- Zipcode.xlsx

Technologies :

1. SQL server management studio :

SQL Server Management Studio est une application logicielle lancée pour la première fois avec Microsoft SQL Server 2005 qui est utilisé pour configurer, gérer et administrer tous les composants au sein de Microsoft serveur SQL.

2. SQL Server Integration Services :

SSIS est un composant du logiciel de base de données Microsoft SQL Server qui peut être utilisé pour effectuer un large éventail de tâches d'extraction, de transformation et de chargement de données.

3. Microsoft Visual Studio :

Nous utiliserons Visual studio pour SQL Server Data Tools (SSDT) est un outil de développement moderne pour création de bases de données relationnelles SQL Server, bases de données dans Azure SQL, données Analysis Services (AS) les modèles, les packages Integration Services (IS) et les rapports Reporting Services (RS).

4. Power BI :

Power BI est une solution de Business Intelligence développée par Microsoft pour permettre aux entreprises d'agréger, d'analyser et de visualiser les données en provenance de sources multiples.

Indicateurs clé de performance – KPIs :

- **Description des axes de l'analyse :**

Axe de l'analyse	Description	Values
Customer	Information sur le client	HA4485-99, 22, Male, Married
Localisation	Zip code, ville, state lié au client.	225694, Los Angelos, California
Service	Les services utilisés par chaque client	Phone Service, Multiple Lines, Internet Service
Date	Représente la date et le trimestre de l'inscription ou de désabonnement du client	02/04/2020, Q2

Indicateurs – KPIs :

- **Le taux de désabonnement global** : un indicateur qui permet de calculer le pourcentage de personnes ayant décidé de se désabonner d'une compagnie de télécommunication.
- **Le taux de désabonnement basé sur** :
  - Démographie (genre – retraité – âge…)
  - Services (Internet – phone - Streaming - mode de paiements – type de
  - contrat…)
  - Géographie (State – ville – code postale…).
  - Sur la raison de désabonnement (Churn category)
  - La valeur à vie du client (pour surveiller ceux qui ont un CLTV élevé).
- **Score de satisfaction client :** un indicateur de la qualité de service / produit et évidemment de la satisfaction client.
- **Taux de fidélisation de client :** permet justement de mesurer la fidélisation client

Taux de fidélisation = (1 – Taux d'attrition) x 100

- **Taux de rétention de client :** est un chiffre qui indique la capacité d'une entreprise à fidéliser efficacement ses clients. Pour le calculer il faut préciser une période.

CRR = [(CF-CN) ÷ CD] x 100

  - **Nombre de clients retenus au début de la période (CD)**
  - **Nombre de clients retenus à la fin de la période (CF)**
  - **Nombre de nouveaux clients acquis durant la période (CN)**

- **La valeur à vie du client (Customer Lifetime Value - CLTV) :** s'agit d'une métrique qui donne la somme des gains nets générés par les clients tout au long de leur activité avec l'entreprise.

4- Matrice dimensionnelle d'analyse des indicateurs

	DImCustomer	DimService	DimLocalisation	DimDates
Taux de désabonnement	**X**	**X**	**X**	**X**
Taux de fidélisation	**X**	**X**	**X**	**X**
Taux de satisfaction de client	**X**			
Taux de rétention de client	**X**			**X**
Valeur à vie de client	**X**			

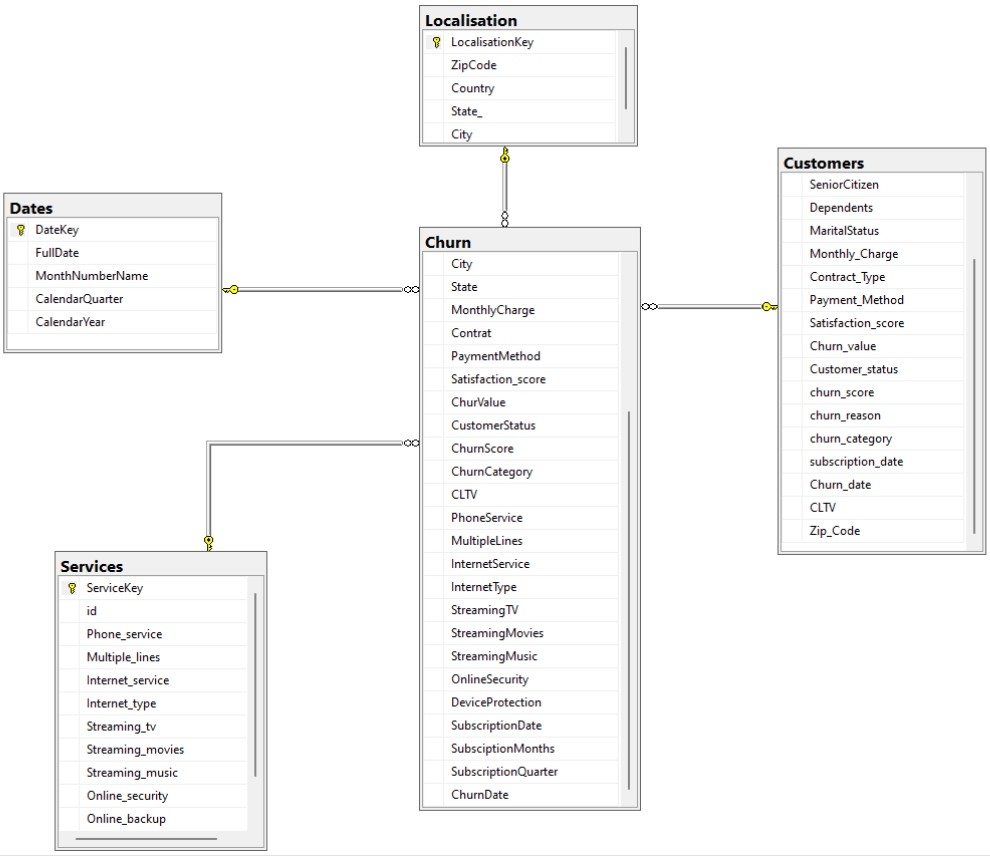
5- Modélisation du datawarehouse :

Dans cette partie du projet, nous mettrons en œuvre l'entrepôt de données, en utilisant le schéma en étoile, qui est l'approche la plus utilisée pour développer des entrepôts de données, il se compose d'une table de faits faisant référence à un certain nombre de table de dimension.

Notre entrepôt de données va contenir un table de fait *''Churn''*, et quatre dimensions :

DimDates, DimCustomers, DimService et finalement *DimLocalisation*.

Les attributs de chaque table sont représentés sur le schéma ci-dessous :



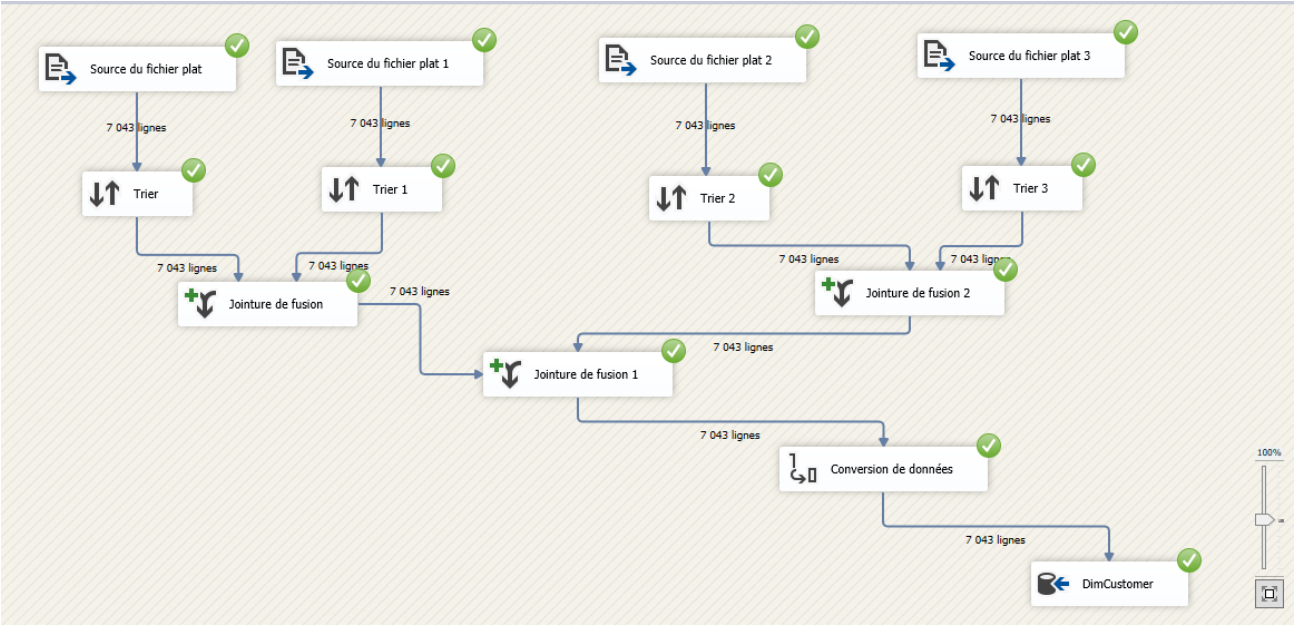
6- ETL :

Extract, **T**ransform, **L**oad est la procédure générale de copier des données d'un ou de plusieurs sources dans un système de destination qui représente les données différemment des source ou dans un contexte différent de la source. Dans ce projet, nous extrairons les données des fichiers plats ci-dessous, et chargez-les dans *''ChurnDW''* Datawarehouse créé

sur SQL Server Management Studio (SSMS), nous devons remplir les quatre dimensions ainsi que la table des faits, pour ce faire nous utiliserons SQL Server Data Tools (SSDT).

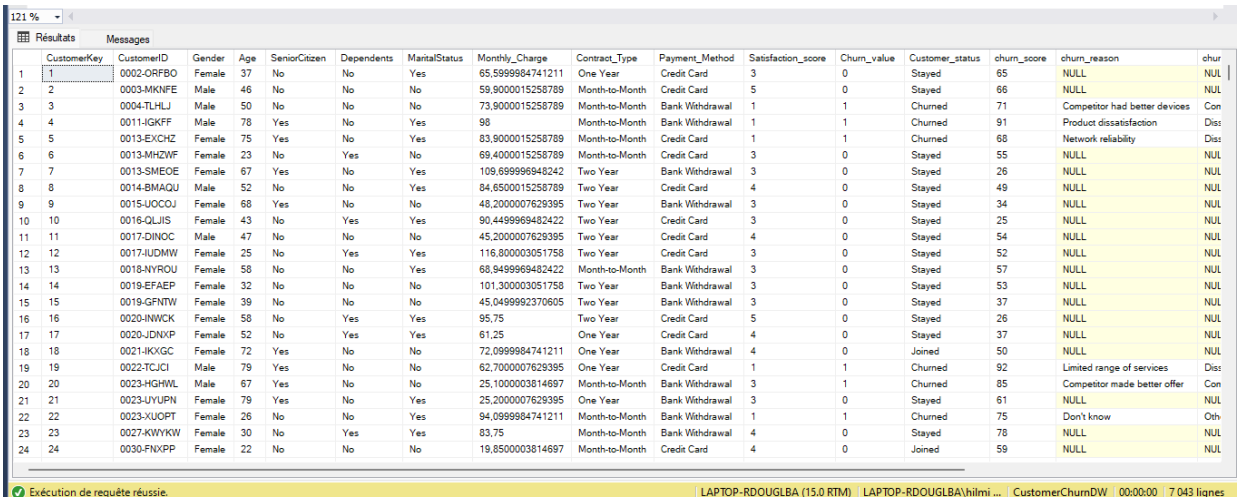
- **La dimension Client :**

Pour remplir cette dimension, on va fusionner les fichiers plats par des jointures et par la suite effectuer une conversion de données pour les rendre compatible avec ceux définies sur la base de données sur Management Studio.



Vérification de la persistance des données :

```
select * from dbo.Customers;
```



121 %

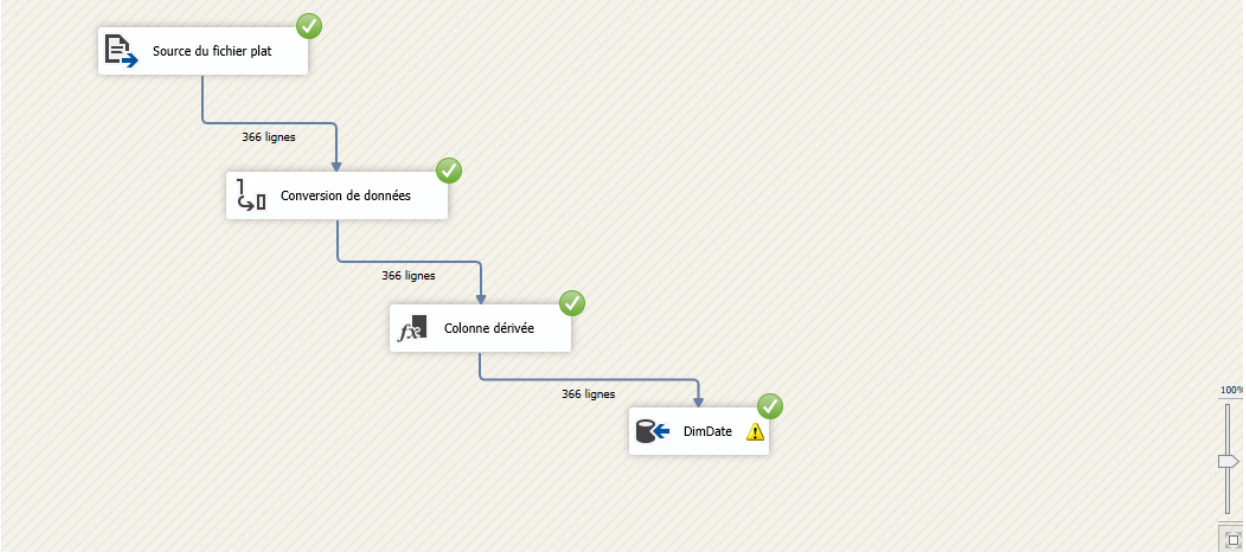
Résultats | Messages

	CustomerKey	CustomerID	Gender	Age	SeniorCitizen	Dependents	MaritalStatus	Monthly_Charge	Contract_Type	Payment_Method	Satisfaction_score	Churn_value	Customer_status	churn_score	churn_reason	chur
1	1	0002-ORFBO	Female	37	No	No	Yes	65,5999984741211	One Year	Credit Card	3	0	Stayed	65	NULL	NUL
2	2	0003-MKNFE	Male	46	No	No	No	59,9000015258789	Month-to-Month	Credit Card	5	0	Stayed	66	NULL	NUL
3	3	0004-TLHLJ	Male	50	No	No	No	73,9000015258789	Month-to-Month	Bank Withdrawal	1	1	Churned	71	Competitor had better devices	Con
4	4	0011-IGKFF	Male	78	Yes	No	Yes	98	Month-to-Month	Bank Withdrawal	1	1	Churned	91	Product dissatisfaction	Diss
5	5	0013-EXCHZ	Female	75	Yes	No	Yes	83,9000015258789	Month-to-Month	Credit Card	1	1	Churned	68	Network reliability	Diss
6	6	0013-MHZWF	Female	23	No	Yes	No	69,4000015258789	Month-to-Month	Credit Card	3	0	Stayed	55	NULL	NUL
7	7	0013-SMEOE	Female	67	Yes	No	Yes	109,699996948242	Two Year	Bank Withdrawal	3	0	Stayed	26	NULL	NUL
8	8	0014-BMAQU	Male	52	No	No	Yes	84,6500015258789	Two Year	Credit Card	4	0	Stayed	49	NULL	NUL
9	9	0015-UOCOJ	Female	68	Yes	No	No	48,2000007629395	Two Year	Bank Withdrawal	3	0	Stayed	34	NULL	NUL
10	10	0016-QLJIS	Female	43	No	Yes	Yes	90,4499969482422	Two Year	Credit Card	3	0	Stayed	25	NULL	NUL
11	11	0017-DINOC	Male	47	No	No	No	45,2000007629395	Two Year	Credit Card	4	0	Stayed	54	NULL	NUL
12	12	0017-IUDMW	Female	25	No	Yes	Yes	116,800003051758	Two Year	Credit Card	3	0	Stayed	52	NULL	NUL
13	13	0018-NYROU	Female	58	No	No	Yes	68,9499969482422	Month-to-Month	Bank Withdrawal	3	0	Stayed	57	NULL	NUL
14	14	0019-EFAEP	Female	32	No	No	No	101,300003051758	Two Year	Bank Withdrawal	3	0	Stayed	53	NULL	NUL
15	15	0019-GFNTW	Female	39	No	No	No	45,0499992370605	Two Year	Bank Withdrawal	3	0	Stayed	37	NULL	NUL
16	16	0020-INWCK	Female	58	No	Yes	Yes	95,75	Two Year	Credit Card	5	0	Stayed	26	NULL	NUL
17	17	0020-JDNXP	Female	52	No	Yes	Yes	61,25	One Year	Credit Card	4	0	Stayed	37	NULL	NUL
18	18	0021-IKXGC	Female	72	Yes	No	No	72,0999984741211	One Year	Bank Withdrawal	4	0	Joined	50	NULL	NUL
19	19	0022-TCJCI	Male	79	Yes	No	No	62,7000007629395	One Year	Credit Card	1	1	Churned	92	Limited range of services	Diss
20	20	0023-HGHWL	Male	67	Yes	No	No	25,1000003814697	Month-to-Month	Bank Withdrawal	3	1	Churned	85	Competitor made better offer	Con
21	21	0023-UYUPN	Female	79	Yes	No	Yes	25,2000007629395	One Year	Bank Withdrawal	3	0	Stayed	61	NULL	NUL
22	22	0023-XUOPT	Female	26	No	No	Yes	94,0999984741211	Month-to-Month	Bank Withdrawal	1	1	Churned	75	Don't know	Oth
23	23	0027-KWYKW	Female	30	No	Yes	Yes	83,75	Month-to-Month	Bank Withdrawal	4	0	Stayed	78	NULL	NUL
24	24	0030-FNXPP	Female	22	No	No	No	19,8500003814697	Month-to-Month	Credit Card	4	0	Joined	59	NULL	NUL

Exécution de requête réussie. | LAPTOP-RDOUGLBA (15.0 RTM) | LAPTOP-RDOUGLBA\hilmi ... | CustomerChurnDW | 00:00:00 | 7 043 lignes

Le processus est similaire pour les autres dimensions.

- **La dimension Date:**



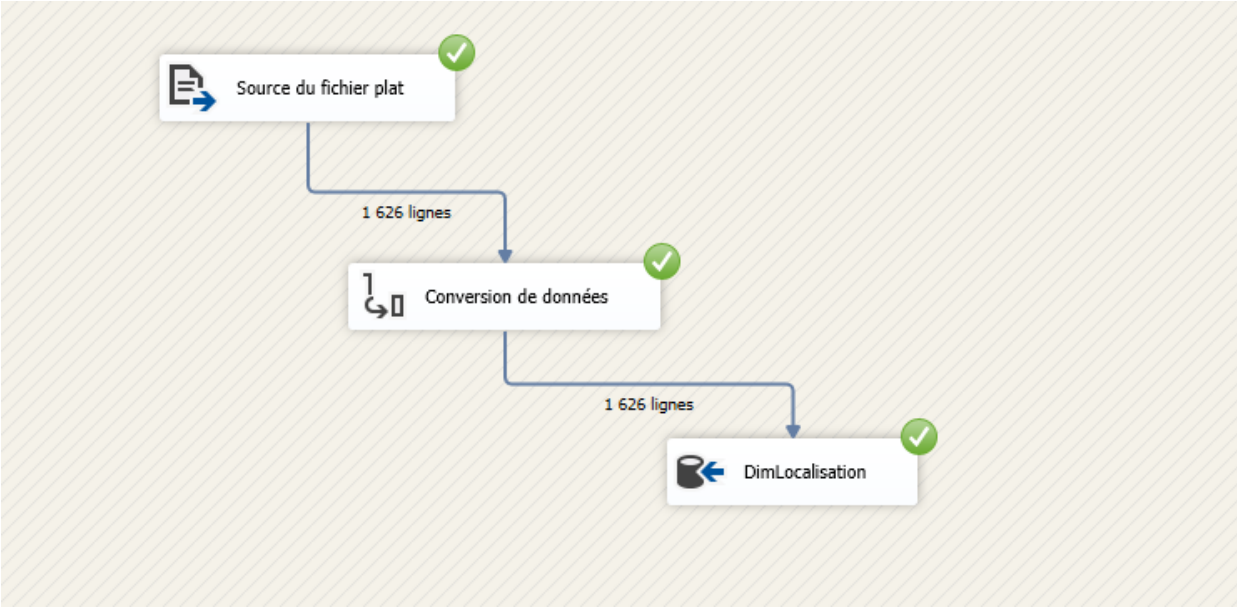

```
select * from dbo.Dates;
```

Résultats | Messages

	DateKey	FullDate	MonthNumberName	CalendarQuarter	CalendarYear
1	1099	2020-01-01	1	Q1	2020
2	1100	2020-01-02	1	Q1	2020
3	1101	2020-01-03	1	Q1	2020
4	1102	2020-01-04	1	Q1	2020
5	1103	2020-01-05	1	Q1	2020
6	1104	2020-01-06	1	Q1	2020
7	1105	2020-01-07	1	Q1	2020
8	1106	2020-01-08	1	Q1	2020
9	1107	2020-01-09	1	Q1	2020
10	1108	2020-01-10	1	Q1	2020
11	1109	2020-01-11	1	Q1	2020
12	1110	2020-01-12	1	Q1	2020
13	1111	2020-01-13	1	Q1	2020
14	1112	2020-01-14	1	Q1	2020
15	1113	2020-01-15	1	Q1	2020
16	1114	2020-01-16	1	Q1	2020
17	1115	2020-01-17	1	Q1	2020
18	1116	2020-01-18	1	Q1	2020
19	1117	2020-01-19	1	Q1	2020
20	1118	2020-01-20	1	Q1	2020
21	1119	2020-01-21	1	Q1	2020
22	1120	2020-01-22	1	Q1	2020
23	1121	2020-01-23	1	Q1	2020
24	1122	2020-01-24	1	Q1	2020
25	1123	2020-01-25	1	Q1	2020

Exécution de requête réussie.

- **La dimension Localisation:**



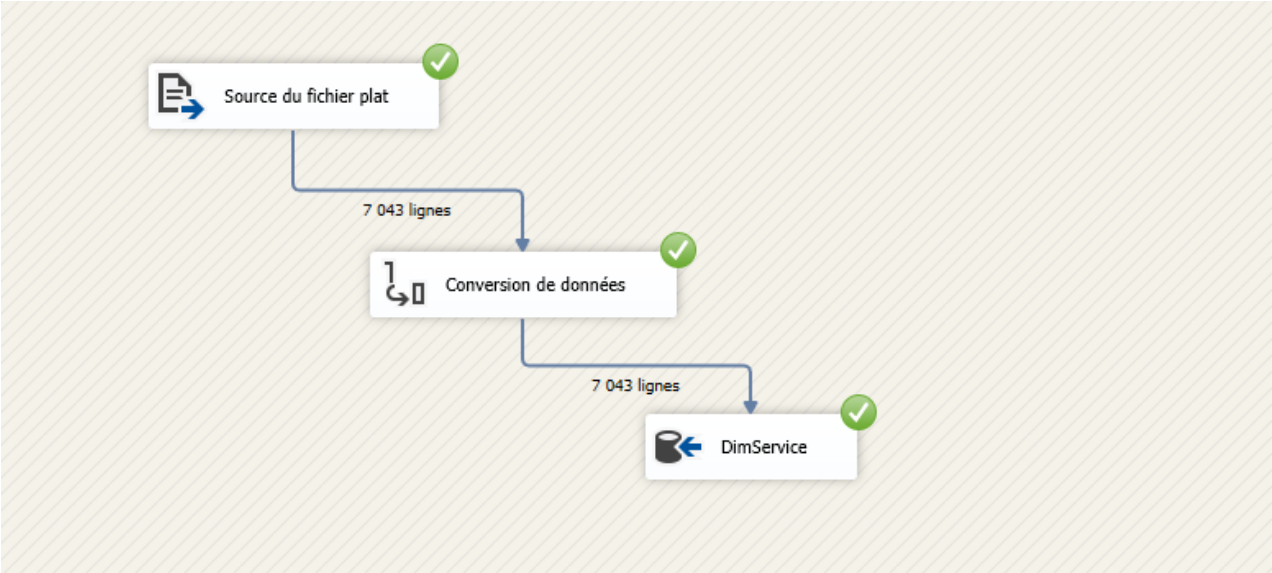

```
select * from dbo.Localisation;
```

Résultats | Messages

	LocalisationKey	ZipCode	Country	State_	City
1	1	90022	United States	California	Los Angeles
2	2	90063	United States	California	Los Angeles
3	3	90065	United States	California	Los Angeles
4	4	90303	United States	California	Inglewood
5	5	90602	United States	California	Whittier
6	6	90660	United States	California	Pico Rivera
7	7	90720	United States	California	Los Alamitos
8	8	91024	United States	California	Sierra Madre
9	9	91106	United States	California	Pasadena
10	10	91107	United States	California	Pasadena
11	11	91605	United States	California	North Hollywood
12	12	91722	United States	California	Covina
13	13	91732	United States	California	El Monte
14	14	91746	United States	California	La Puente
15	15	91748	United States	California	Rowland Heights
16	16	91764	United States	California	Ontario
17	17	91901	United States	California	Alpine
18	18	92122	United States	California	San Diego
19	19	92274	United States	California	Thermal
20	20	92509	United States	California	Riverside
21	21	92544	United States	California	Hemet
22	22	92557	United States	California	Moreno Valley
23	23	92571	United States	California	Perris
24	24	92807	United States	California	Anaheim
25	25	92823	United States	California	Brea

Exécution de requête réussie.

- **La dimension Service:**




```
select * from dbo.Services;
```

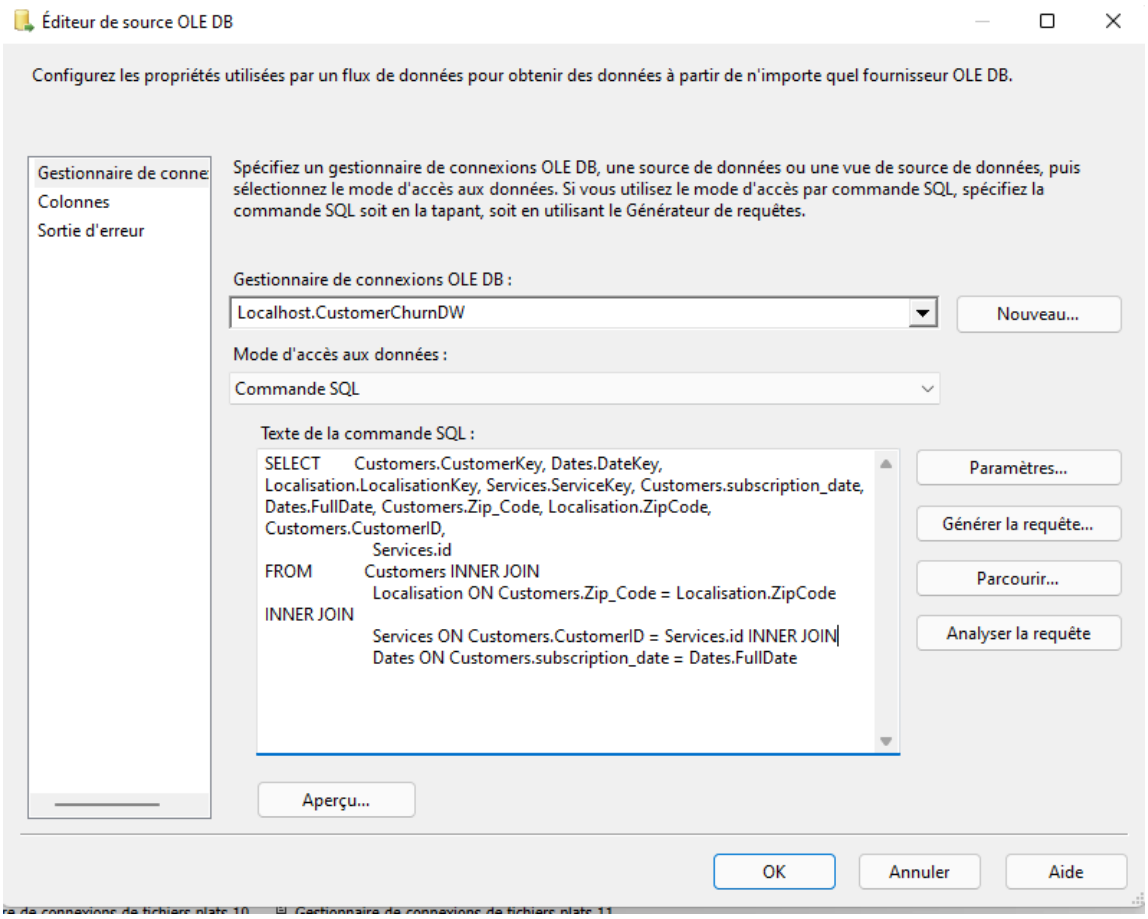
Résultats | Messages

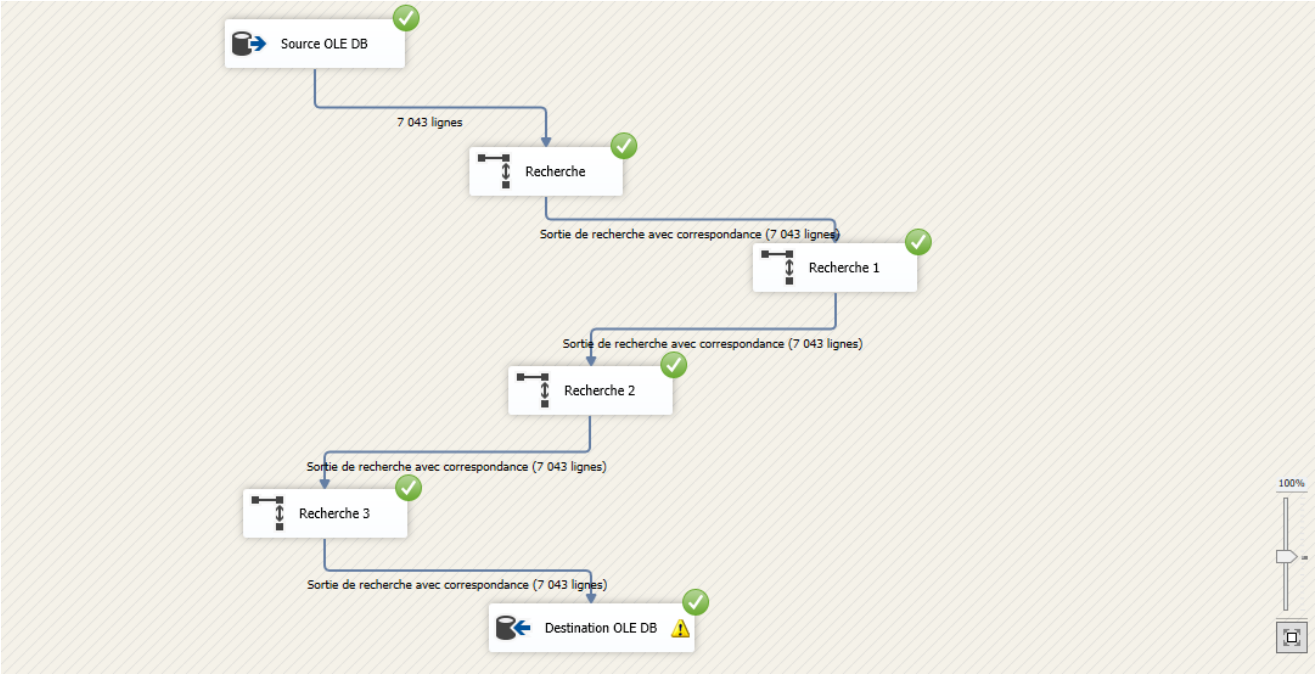
	ServiceKey	id	Phone_service	Multiple_lines	Internet_service	Internet_type	Streaming_tv	Streaming_movies	Streaming_music	Online_security	Online_backup	Device_protection
1	1	8779-QRDMV	0	0	1	DSL	0	1	0	0	0	1
2	2	7495-OOKFY	1	1	1	Fiber Optic	0	0	0	0	1	0
3	3	1658-BYGOY	1	1	1	Fiber Optic	1	1	1	0	0	0
4	4	4598-XLKNJ	1	0	1	Fiber Optic	1	1	0	0	1	1
5	5	4846-WHAFZ	1	1	1	Fiber Optic	0	0	0	0	0	0
6	6	4412-YLTKF	1	1	1	Fiber Optic	0	0	0	0	0	1
7	7	0390-DCFDQ	1	0	1	Fiber Optic	0	0	0	0	0	0
8	8	3445-HXXGF	0	0	1	DSL	0	1	0	0	1	1
9	9	2656-FMOKZ	1	1	1	Fiber Optic	0	0	0	0	0	0
10	10	2070-FNEXE	1	0	1	Fiber Optic	0	0	0	1	0	0
11	11	0094-OIFMO	1	0	1	Fiber Optic	1	1	0	0	1	0
12	12	9947-OTFQU	1	0	1	Fiber Optic	0	0	0	0	1	0
13	13	9514-JDSKI	0	0	1	DSL	1	0	0	0	1	0
14	14	7273-TEFQD	0	0	1	DSL	0	1	1	0	0	1
15	15	3606-TWKGI	1	1	1	Cable	1	1	0	0	1	1
16	16	4385-GZQXV	1	0	1	Fiber Optic	1	1	1	0	0	0
17	17	3488-PGMQJ	1	0	1	Fiber Optic	0	0	0	0	1	0
18	18	7534-BFESC	1	1	1	Fiber Optic	0	0	0	0	0	0
19	19	8098-LLAZX	1	1	1	Fiber Optic	1	1	0	0	0	0
20	20	0265-EDXBD	1	1	1	Fiber Optic	0	0	0	0	0	0
21	21	2840-XANRC	1	0	1	Cable	1	1	1	0	0	1
22	22	5020-ZSTTY	1	1	1	DSL	1	1	1	1	1	0
23	23	5804-LEPIM	1	0	1	Fiber Optic	0	0	0	0	0	0
24	24	0623-IIHUG	0	0	1	Cable	1	0	0	1	0	0
25	25	9057-MSWCO	0	0	1	Cable	0	0	0	1	0	0

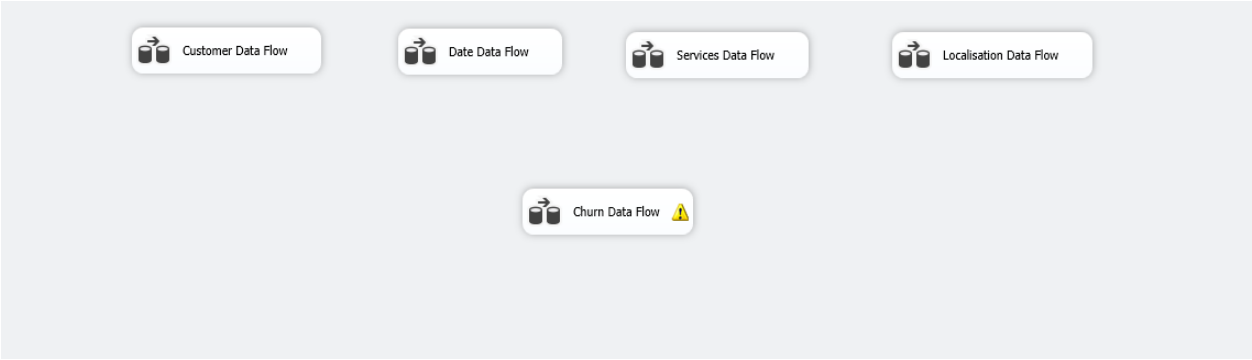
Exécution de requête réussie. | LAPTOP-RDOUGLBA (15.0 RTM) | LAPTOP-RDOUGLBA\hilmi ... | CustomerChurnD

- **Table de fait [Churn] :**

Requête dans Source OLE DB :


Éditeur de source OLE DB
Configurez les propriétés utilisées par un flux de données pour obtenir des données à partir de n'importe quel fournisseur OLE DB.
Gestionnaire de conne
Colonnes
Sortie d'erreur
Spécifiez un gestionnaire de connexions OLE DB, une source de données ou une vue de source de données, puis sélectionnez le mode d'accès aux données. Si vous utilisez le mode d'accès par commande SQL, spécifiez la commande SQL soit en la tapant, soit en utilisant le Générateur de requêtes.
Gestionnaire de connexions OLE DB :
Localhost.CustomerChurnDW
Nouveau...
Mode d'accès aux données :
Commande SQL
Texte de la commande SQL :
SELECT Customers.CustomerKey, Dates.DateKey,
Localisation.LocalisationKey, Services.ServiceKey, Customers.subscription_date,
Dates.FullDate, Customers.Zip_Code, Localisation.ZipCode,
Customers.CustomerID,
Services.id
FROM Customers INNER JOIN
Localisation ON Customers.Zip_Code = Localisation.ZipCode
INNER JOIN
Services ON Customers.CustomerID = Services.id INNER JOIN
Dates ON Customers.subscription_date = Dates.FullDate
Paramètres...
Générer la requête...
Parcourir...
Analyser la requête
Aperçu...
OK
Annuler
Aide


Source OLE DB
7 043 lignes
Recherche
Sortie de recherche avec correspondance (7 043 lignes)
Recherche 1
Sortie de recherche avec correspondance (7 043 lignes)
Recherche 2
Sortie de recherche avec correspondance (7 043 lignes)
Recherche 3
Sortie de recherche avec correspondance (7 043 lignes)
Destination OLE DB
100%




```sql
select * from dbo.Churn
```

121 %

Résultats | Messages

	urnKey	CustomerKey	DateKey	LocalisationKey	ServiceKey	CustomerID	Gender	Age	MaritalStatus	City	State	MonthlyCharge	Contrat	PaymentMethod	Satisfaction_score	ChurValue	CustomerStatus
1		1	1194	770	4007	0002-ORFBO	Female	37	Yes	Frazier Park	California	65,5999984741211	One Year	Credit Card	3	0	Stayed
2		2	1192	871	4789	0003-MKNFE	Male	46	No	Glendale	California	59,9000015258789	Month-to-Month	Credit Card	5	0	Stayed
3		3	1196	1161	1902	0004-TLHLJ	Male	50	No	Costa Mesa	California	73,9000015258789	Month-to-Month	Bank Withdrawal	1	1	Churned
4		4	1231	331	396	0011-IGKFF	Male	78	Yes	Martinez	California	98	Month-to-Month	Bank Withdrawal	1	1	Churned
5		5	1189	221	369	0013-EXCHZ	Female	75	Yes	Camarillo	California	83,9000015258789	Month-to-Month	Credit Card	1	1	Churned
6		6	1222	594	3310	0013-MHZWF	Female	23	No	Midpines	California	69,4000015258789	Month-to-Month	Credit Card	3	0	Stayed
7		7	1184	566	708	0013-SMEOE	Female	67	Yes	Lompoc	California	109,699996948242	Two Year	Bank Withdrawal	3	0	Stayed
8		8	1231	49	4226	0014-BMAQU	Male	52	Yes	Napa	California	84,6500015258789	Two Year	Credit Card	4	0	Stayed
9		9	1264	557	696	0015-UOCOJ	Female	68	No	Simi Valley	California	48,2000007629395	Two Year	Bank Withdrawal	3	0	Stayed
10		10	1218	77	5576	0016-QLJIS	Female	43	Yes	Sheridan	California	90,4499969482422	Two Year	Credit Card	3	0	Stayed
11		11	1237	1380	3819	0017-DINOC	Male	47	No	Rancho Santa Fe	California	45,2000007629395	Two Year	Credit Card	4	0	Stayed
12		12	1270	920	5247	0017-IUDMW	Female	25	Yes	Sunnyvale	California	116,800003051758	Two Year	Credit Card	3	0	Stayed
13		13	1178	983	5615	0018-NYROU	Female	58	Yes	Antelope	California	68,9499969482422	Month-to-Month	Bank Withdrawal	3	0	Stayed
14		14	1138	1144	3791	0019-EFAEP	Female	32	No	La Mesa	California	101,300003051758	Two Year	Bank Withdrawal	3	0	Stayed
15		15	1193	450	5132	0019-GFNTW	Female	39	No	Los Olivos	California	45,0499992370605	Two Year	Bank Withdrawal	3	0	Stayed
16		16	1203	446	4035	0020-INWCK	Female	58	Yes	Woodlake	California	95,75	Two Year	Credit Card	5	0	Stayed
17		17	1270	248	6393	0020-JDNXP	Female	52	Yes	Point Reyes Station	California	61,25	One Year	Credit Card	4	0	Stayed
18		18	1332	384	509	0021-IKXGC	Female	72	No	San Marcos	California	72,0999984741211	One Year	Bank Withdrawal	4	0	Joined
19		19	1229	132	161	0022-TCJCI	Male	79	No	Daly City	California	62,7000007629395	One Year	Credit Card	1	1	Churned
20		20	1149	250	289	0023-HGHWL	Male	67	No	Morgan Hill	California	25,1000003814697	Month-to-Month	Bank Withdrawal	3	1	Churned
21		21	1158	571	716	0023-UYUPN	Female	79	Yes	Ahwahnee	California	25,2000007629395	One Year	Bank Withdrawal	3	0	Stayed
22		22	1275	1232	2448	0023-XUOPT	Female	26	Yes	Carnelian Bay	California	94,0999984741211	Month-to-Month	Bank Withdrawal	1	1	Churned
23		23	1242	29	6129	0027-KWYKW	Female	30	Yes	Alpaugh	California	83,75	Month-to-Month	Bank Withdrawal	4	0	Stayed
24		24	1366	776	3024	0030-FNXPP	Female	22	No	Keeler	California	19,8500003814697	Month-to-Month	Credit Card	4	0	Joined

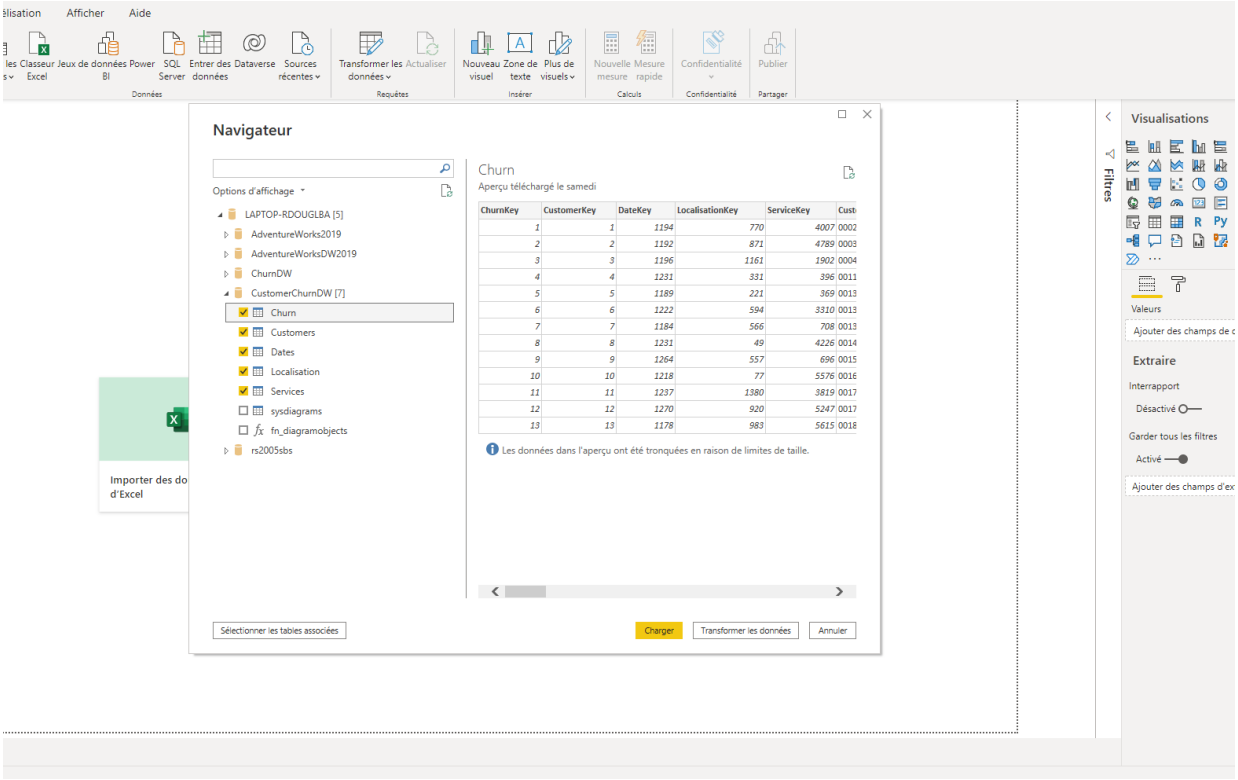
Exécution de requête réussie. | LAPTOP-RDOUGLBA (15.0 RTM) | LAPTOP-RDOUGLBA\hilmi ... | CustomerChurnDW | 00:00:00 | 7 043 lignes

7- Analyse et Reporting

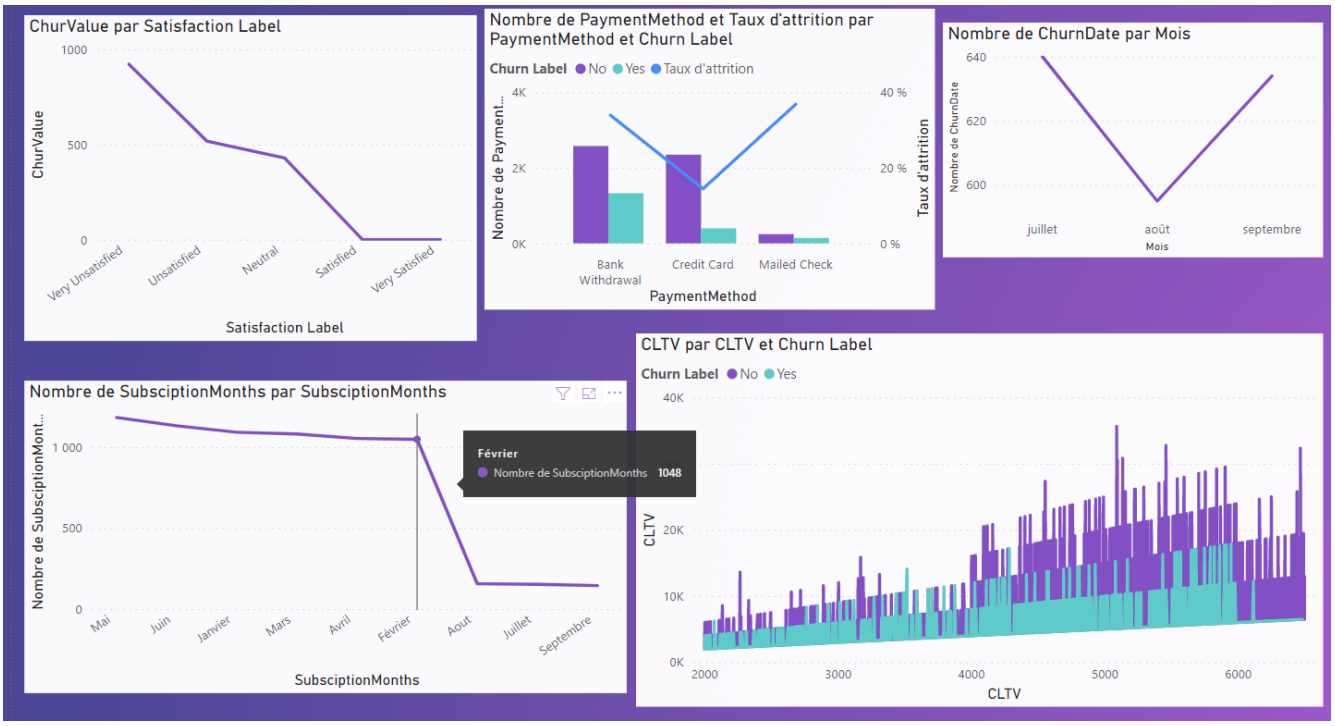
Pour cette partie, nous allons construire notre analyse et notre tableau de bord en utilisant la visualisation sur python et Power BI:

Power BI :

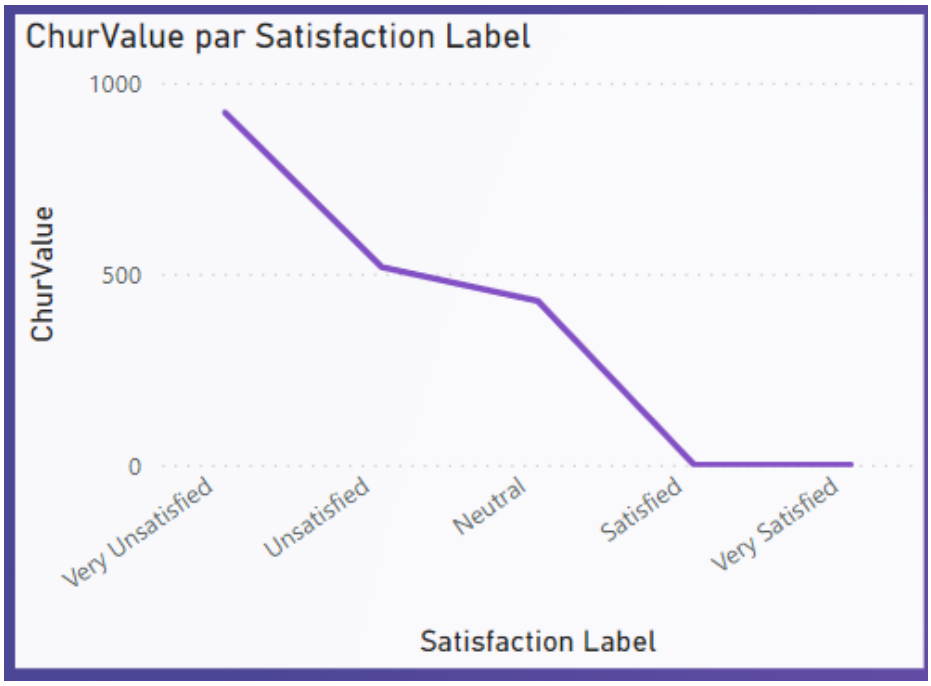
En premier lieu, on importe notre entrepôt de données sur Power BI afin de l'exploiter en calculant nos indicateurs et créant des tableaux de bords.



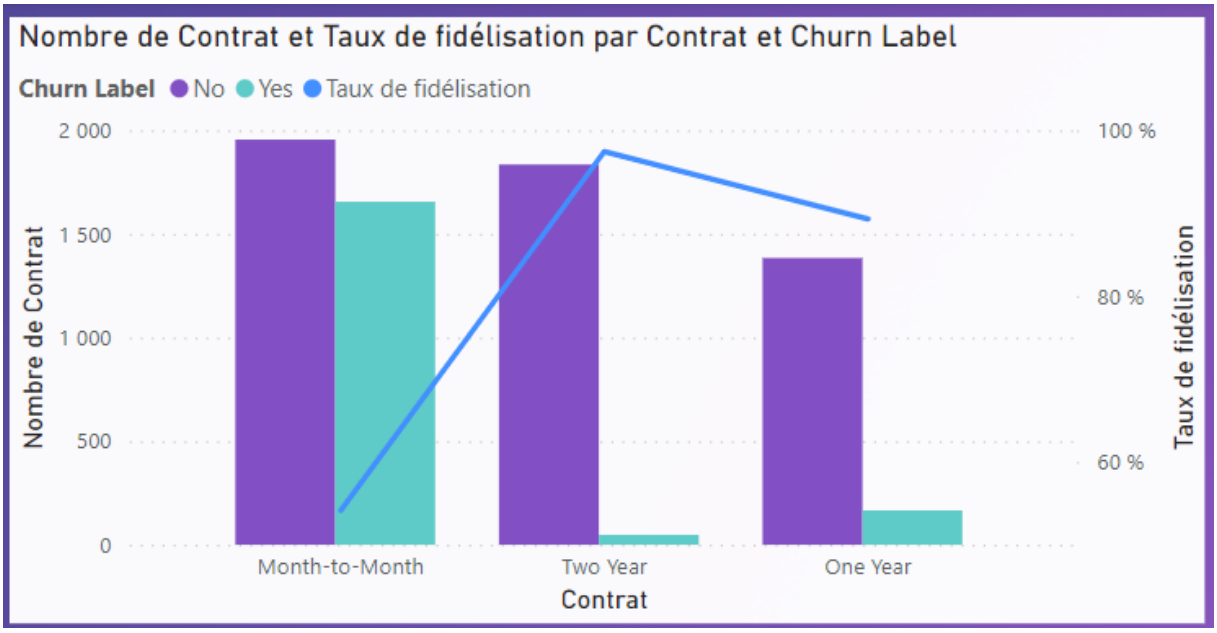
Dashboard 1 : Présente les différentes KPIs : Taux d'attrition , Taux de rétention de client, Taux de fidélisation, Taux de satisfaction.



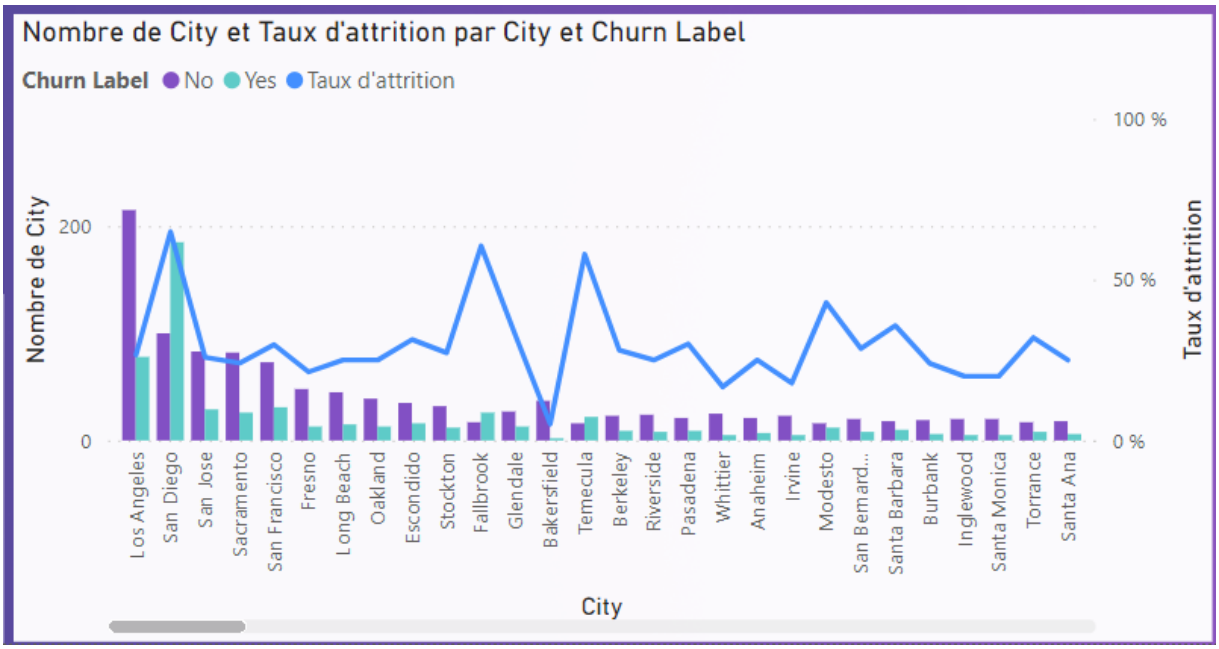
Dashboard 2



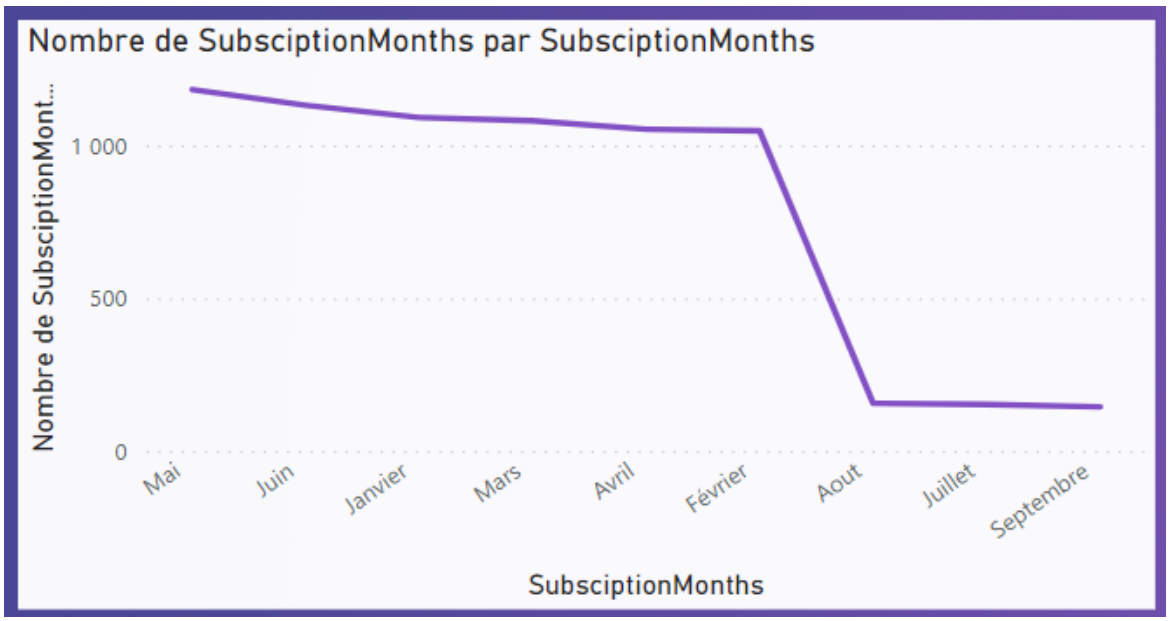
On remarque que les clients qui se désabonnent sont non satisfait



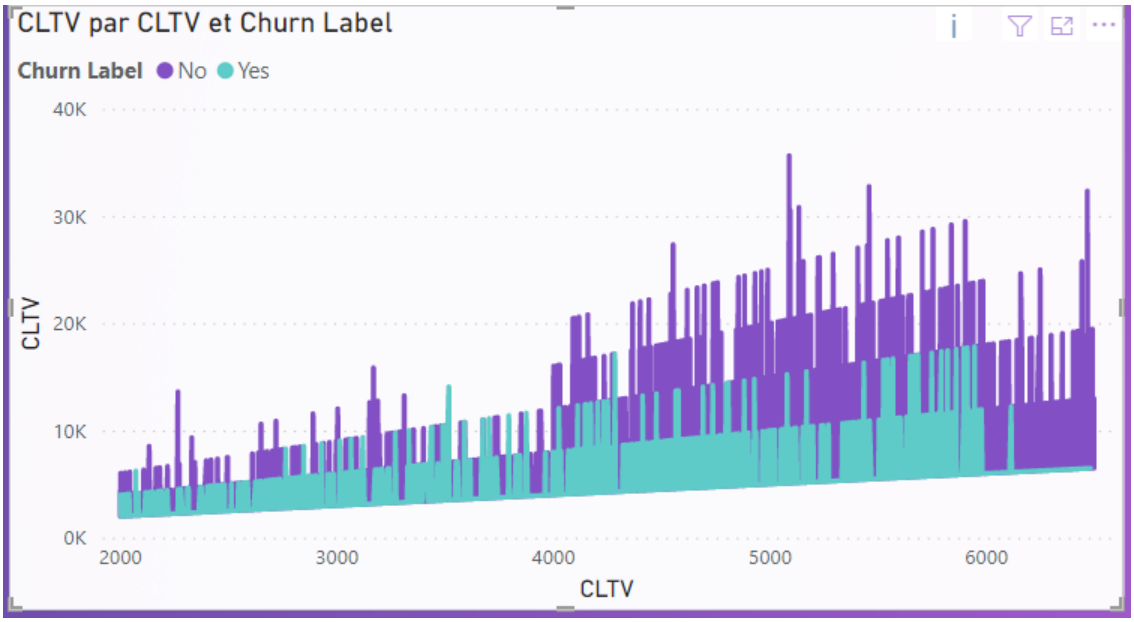
Cette figure représente le taux de fidélisation par type de contrat, et on remarque que les clients avec une contrat d'un an ou deux ans sont les plus susceptible de rester abonnés.



Cette figure représente le taux d'attrition dans chaque ville.



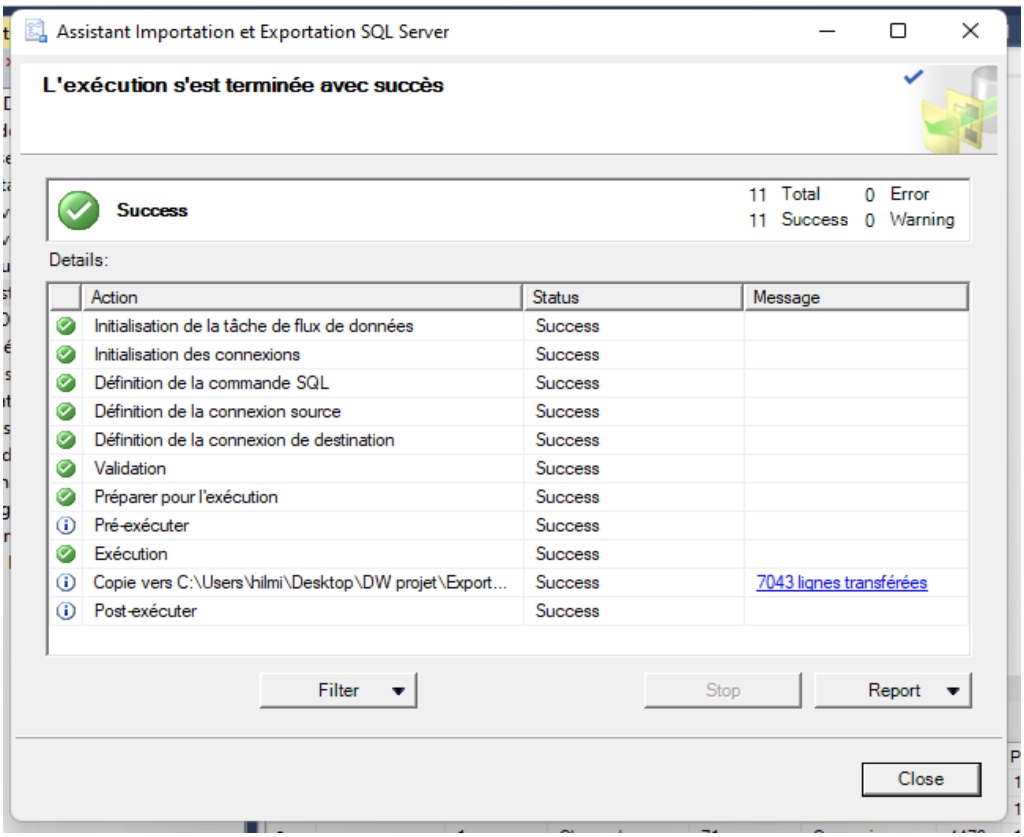
Cette figure présente le nombre de subscription dans chaque mois



Cette figure présente la densité de CLTV par un filtre de désabonnement.

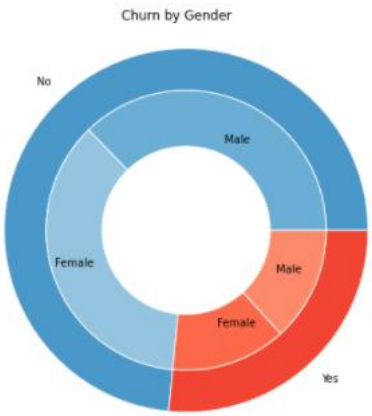
Python : Visualisation et Analyse.

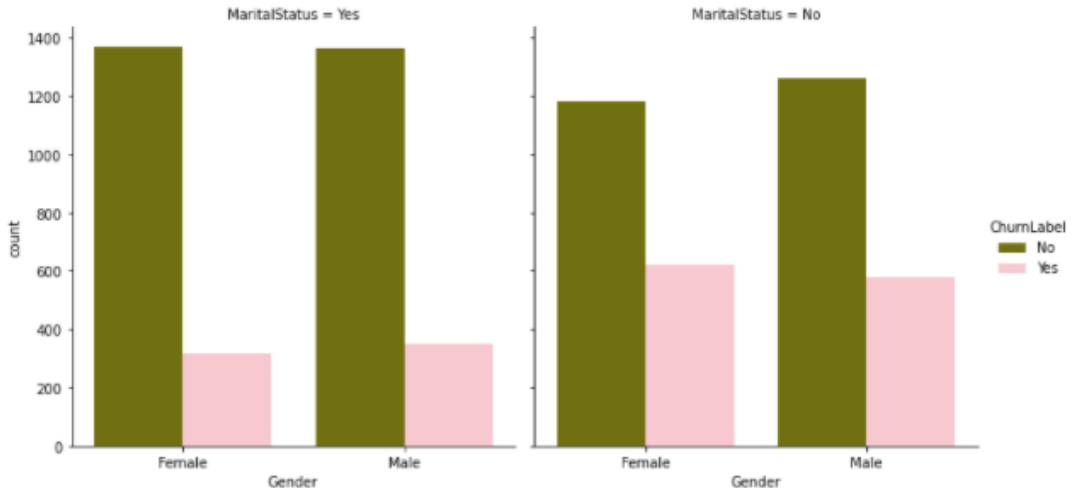
On exporte la table de fait réalisé sous la format CSV afin d'effectuer une analyse descriptive et prédictive qui vise à explorer et comprendre l'ensemble de données, les raisons de désabonnement et réaliser un modèle prédictif d'attrition des clients.



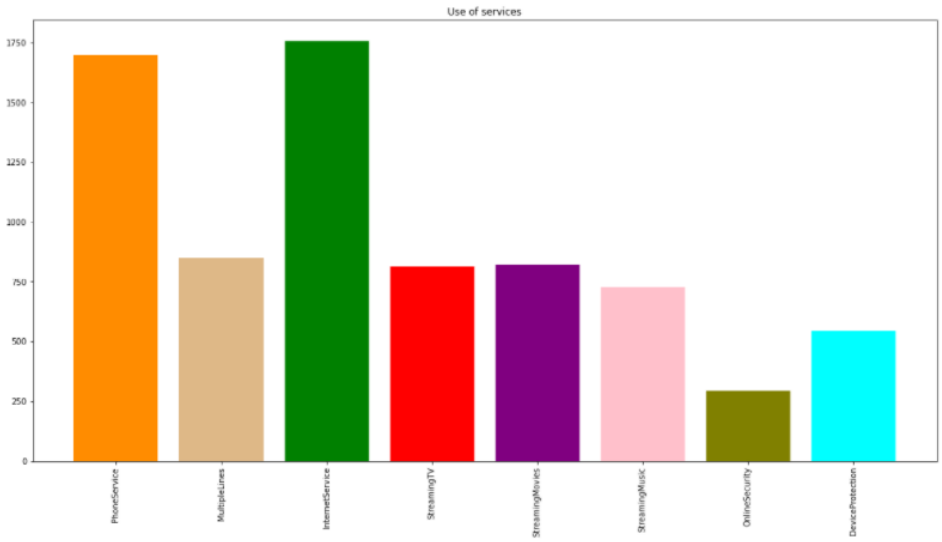
- **Data Visualization :**

Une visualisation générale des données.

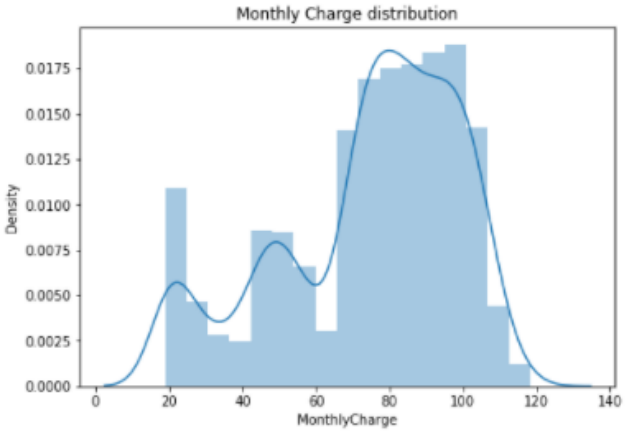


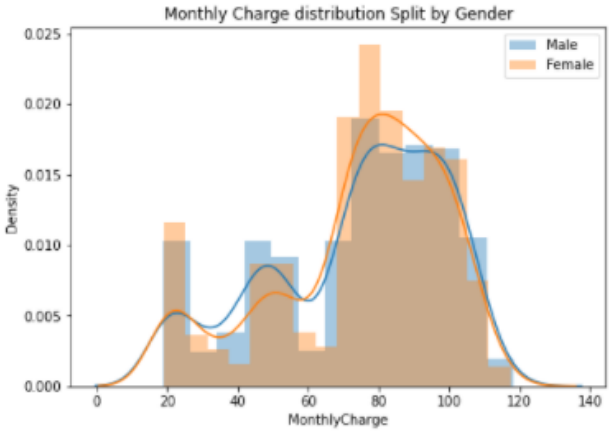


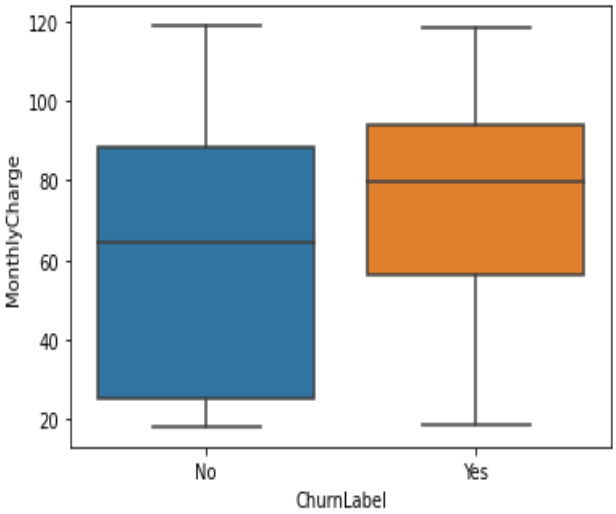
D'après la visualisation ci-dessus, les femmes et les hommes qui ne sont pas marriés sont plus susceptibles de se désabonner.

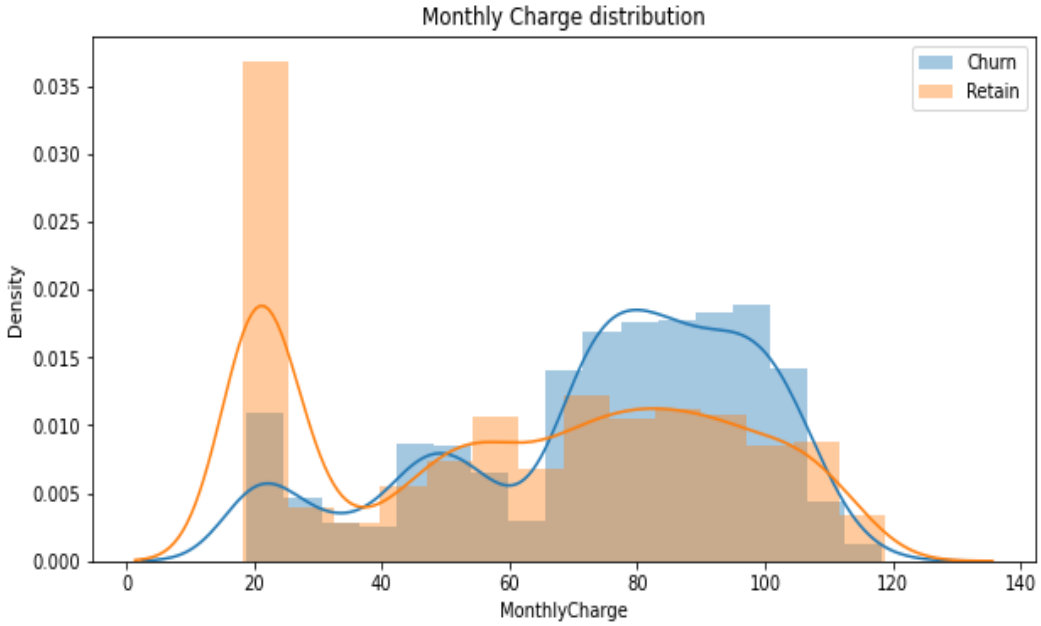


Le taux de désabonnement est plus élevé dans le service internet puis le service phone. Il se peut que la qualité de ces deux services est assez médiocre.

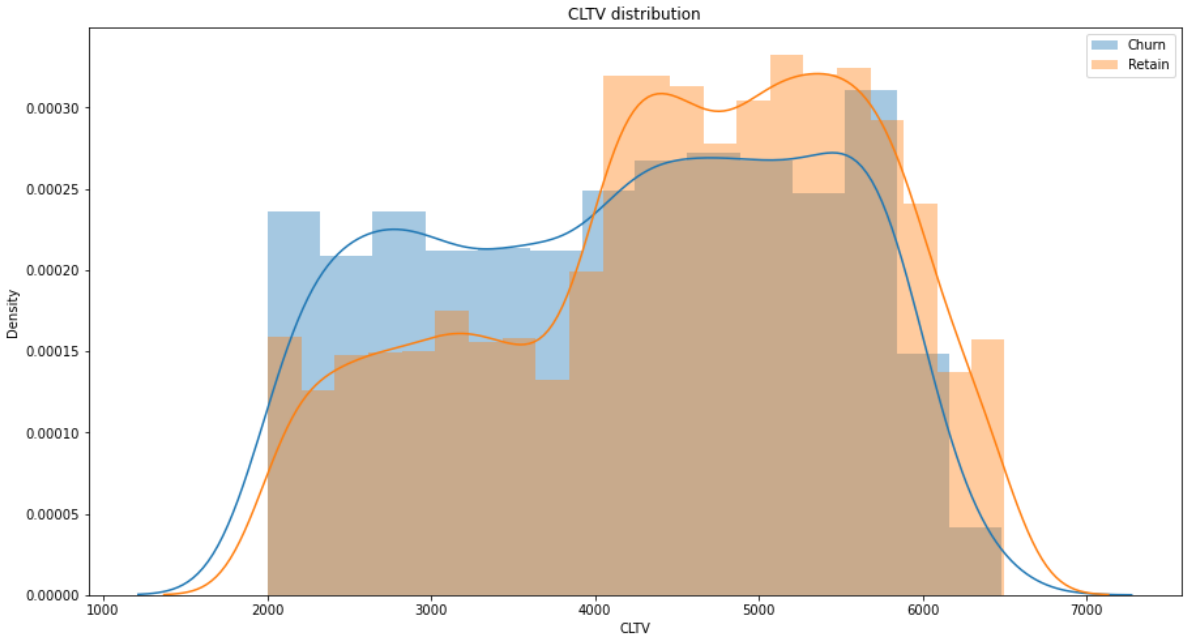








D'après la visualisation ci-dessus, les clients qui effectuent des transactions par mois en moyenne de plus de 60 $ seront plus susceptibles de se désabonner que la transaction moyenne inférieure.



Les clients avec un CLTV plus de 4000 sont plus susceptible de conserver l'abonnement.

➔ Plus la valeur du CLTV est élevée, plus le client est précieux. Les clients de grande valeur doivent être surveillés pour le taux de désabonnement.

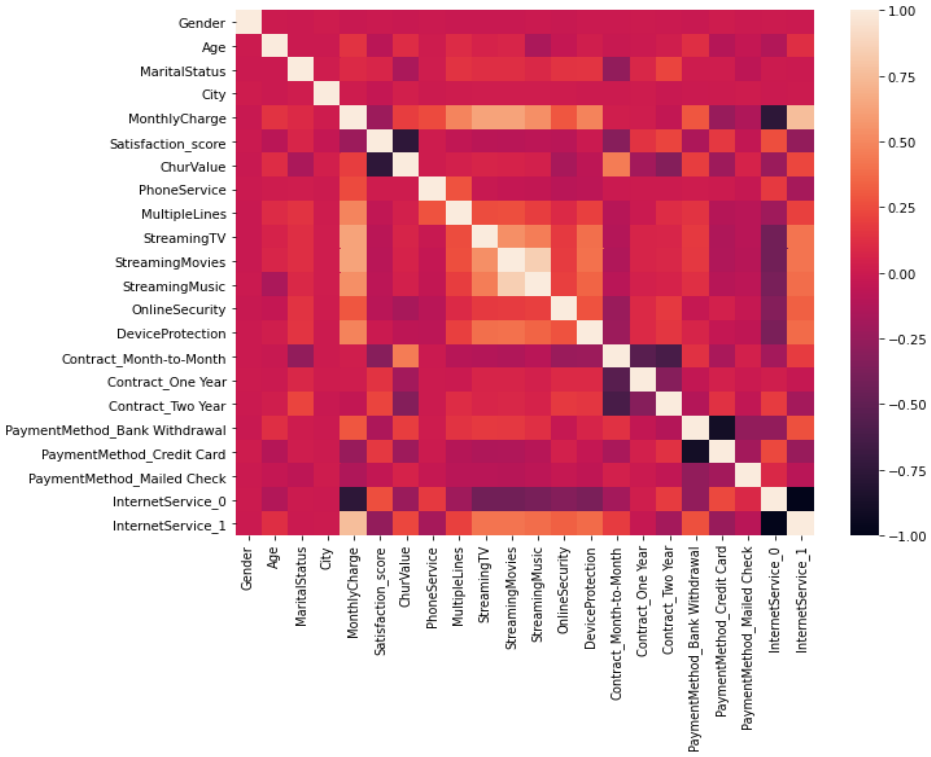
- **Analyse prédictive**

Dans la dernière partie du projet, nous mettrons en œuvre des solutions prédictives pour la compagnies Telco afin d'explorer les modèles en analysant les faits actuels et historiques et faire des prédictions sur la situation des futurs clients.

Après le pré-traitement et la normalisation des données, on détermine la corrélation entre variables :

Entrée [48]:
```
plt.figure(figsize=(10,8))
sns.heatmap(data2.corr())
```

Out[48]: <AxesSubplot:>

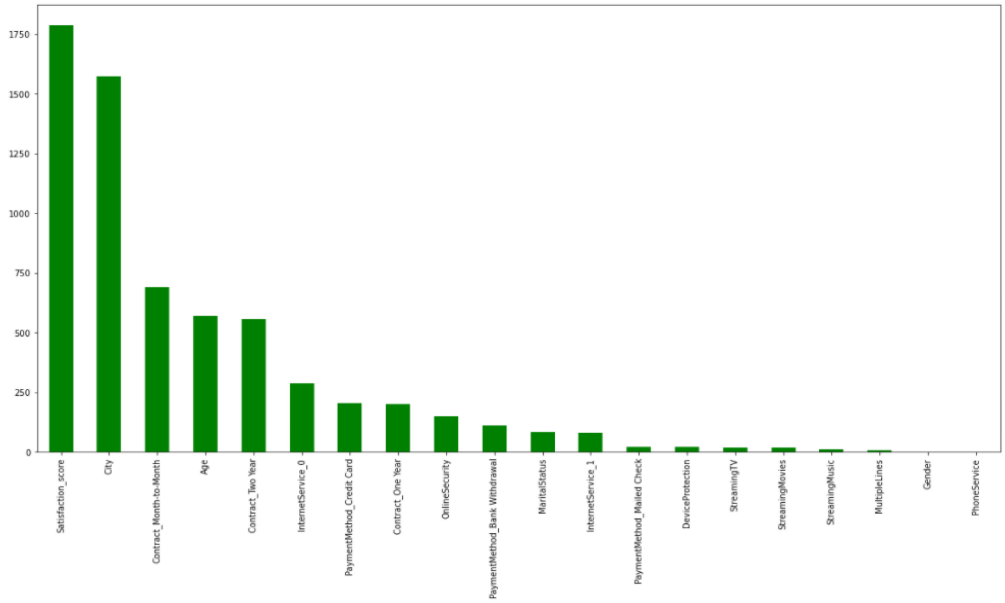


On choisit les variables les plus corrélées avec la variable cible en utilisant le SelectKBest avec un test Chi-square.

Entrée [64]:
```
# SelectKBest avec le test Chi-square (categorial vs categorial)

from sklearn.feature_selection import chi2
from sklearn.feature_selection import SelectKBest

X_cat = data2.drop(['ChurValue','MonthlyCharge'],axis = 1)
Y_cat = data2.ChurValue
fs = SelectKBest(score_func = chi2, k="all")
X_f = fs.fit_transform(X_cat, Y_cat)
chi2_scores_df = pd.DataFrame(fs.scores_, index = X_cat.columns, columns= ['Score'])
chi2_scores_df['Score'].sort_values(ascending=False).plot.bar(figsize=(20,10), color = 'green')
chi2_scores_df
```



On fixe **K=5**

Entrée [57]:
```
from sklearn.feature_selection import chi2
from sklearn.feature_selection import SelectKBest

X_cat = data2.drop(['ChurValue','MonthlyCharge'],axis = 1)
Y_cat = data2.ChurValue
fs = SelectKBest(score_func = chi2, k=5)
X_f = fs.fit_transform(X_cat, Y_cat)

X_best = fs.get_feature_names_out(X_cat.columns)
X_best = list(X_best)
X_best
```

Out[57]:
```
['Age',
 'City',
 'Satisfaction_score',
 'Contract_Month-to-Month',
 'Contract_Two Year']
```

Entrée [58]:
```
X = data2[X_best+['MonthlyCharge']]
y=data2.ChurValue

X_train_val, X_test, y_train_val, y_test = train_test_split(X,y,test_size=.2, stratify=y)
```

Entrée [59]:
```
model = LogisticRegression(random_state=42, max_iter =10000)
model.fit(X_train_val,y_train_val)
```

Out[59]:
```
LogisticRegression(max_iter=10000, random_state=42)
```

Entrée [60]:
```
from sklearn.metrics import accuracy_score
accuracy_score(y_test,model.predict(X_test))
```

Out[60]:
```
0.936124911284599
```

On obtient une accuracy de 0.94.

```
Entrée [61]: from sklearn.metrics import classification_report, confusion_matrix, plot_confusion_matrix

plt.figure(figsize=(10,10))
plot_confusion_matrix(model,X_test,y_test)
print(classification_report(y_test,model.predict(X_test)))
```

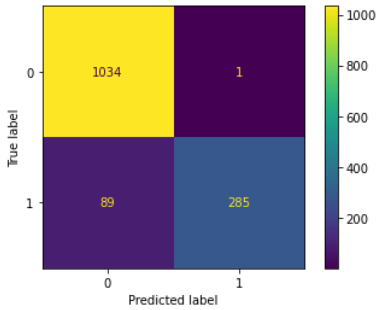
```
              precision    recall  f1-score   support

           0       0.92      1.00      0.96      1035
           1       1.00      0.76      0.86       374

    accuracy                           0.94      1409
   macro avg       0.96      0.88      0.91      1409
weighted avg       0.94      0.94      0.93      1409
```

```
C:\ProgramData\Anaconda3\lib\site-packages\sklearn\utils\deprecation.py:87: FutureWarning: Function plot_confusion_matrix is de
precated; Function `plot_confusion_matrix` is deprecated in 1.0 and will be removed in 1.2. Use one of the class methods: Confu
sionMatrixDisplay.from_predictions or ConfusionMatrixDisplay.from_estimator.
  warnings.warn(msg, category=FutureWarning)
```

```
<Figure size 720x720 with 0 Axes>
```



En affichant le rapport de classification, on remarque que le modèle implémenté effectue de bonne prédiction et avait de bonne score au niveau de tous les métriques d'évaluation.

Conclusion

La prise de décision est un rôle clé dans le succès de toute entreprise, ces décisions affectent directement l'entreprise, et donc une mauvaise décision par une source mal informée conduit à des résultats désastreux. Les décisions sont prises au niveau individuel jusqu'au niveau organisationnel. Cette solution BI que nous avons mise en place améliorera sûrement la décisions prises par les dirigeants de la compagnie Telco. Les systèmes de BI permettront en effet tous les gestionnaires de la compagnie Telco avec suffisamment d'informations et les rendre capables de la prise de décision.