

**Sommaire**

PAGE

Remerciement 4

Introduction 5-6

CHAPITRE 1 – Techniques et programmes requis 7-8

CHAPITRE 2 – La gestion de base de données 9-14

CHAPITRE 3 – Etude statistique 15-33

Informations sur Datafraame 15-17

Filtrage et nettoyage des données 18

Comprendre le chiffre d’affaire de l’entreprise 19-24

Clients actifs pour chaque mois 25-29

Taux de rétention mensuel 30-32

Le taux de désabonnement (Churn Rate) 32-33

Conclusion 33

CHAPITRE 4 – La segmentation par la méthode RFM 34-47

Représentation de RFM 34

Récence 35-38

Fréquence 39-41

Montant 41-43

Segmentation Globale 43-47

Conclusion 47

CHAPITRE 5 – La valeur client (CLV) 48-63

Définition et calcul de CLV 48-52

Modèle bi-varié 52-59

Modèle multivarié 59-63

Conclusion générale 64-65

***Remerciement***

En tout premier lieu, Nous tenons à remercier toute l’équipe

Pédagogique de l'université Mohammed VI Polytechnique et les

intervenants professionnels responsables de la formation Sciences des

données.

Avant de débuter ce rapport, nous profitons de l’occasion pour remercier

tout d’abord nos professeurs **Abdellah MASSAQ**, **Idir OUASSOU**

et **M’hammed ELKAHOUI** pour leurs engagement à nos côtés et qui n’ont

pas cessé de nous accompagner pendant la durée du projet et d’avoir

incités à travailler en mettant à notre disposition leurs expériences et

leurs compétences. A travers leurs conseils concernant les missions

évoquées dans ce rapport, qu’ils nous ont apporté lors des

différents suivis.

**Indroduction :**

**Introduction**

Dans le marché commercial, il y’a plusieurs entreprises qui génèrent une quantité importante des données. Cette vaste base de données de transactions clients doit être analysée pour concevoir des stratégies rentables.

Vue que la majorité des clients ont des besoins différents (ce qu'ils font, de quelle façon ils achètent, à quel moment ils achètent, de combien d'argent ils disposent…). Donc, avec cette augmentation de commandes, c’est difficile de comprendre les exigences de chaque client. Pour cela on a décidé de choisir ce sujet comme un projet de fin de semestre. Pour créer des segments qui jouent un meilleur rôle dans le regroupement de ces clients, qui servent à :

* Reconnaître les meilleurs clients.
* Déterminer nos fidèles clients.
* Savoir les clients qui sont au bord du désabonnement.
* Quand et comment développer des nouveaux produits.
* Le moment convenable pour diffuser des annonces et des promotions

On peut réaliser ces taches grâce aux plusieurs techniques, on cite : **RFM**



Pour mieux déterminer quels clients dépenser de l’argent et dans quelle mesure, RFM est une méthodologie de marketing axée sur l’évaluation de la valeur de vos clients, puis en utilisant ces valeurs pour aider à guider vos décisions. Ce faisant, vous serez en mesure d’avoir une approche plus précise et pertinente pour chaque client, tout en améliorant votre rentabilité et votre expérience clientèle.

**Chapitre1 : Techniques et programmes requis.**

Pour réaliser ce projet nous avons implémenté nos modèles par plusieurs langages et logiciels, on cite :

**Rstudio :**

Pour prendre les bonnes décisions, il est nécessaire de disposer de statistiques d’excellente qualité, et c’est pour ce but nous avons utilisé R-Studio pour les études statistiques surtout la régression linéaire.

**Python :**

**Python** permet de créer toutes sortes de programmes, comme des jeux, des logiciels, etc. Il est possible d'associer des bibliothèques à Python afin d'étendre ses possibilités, et c’est le cas pour notre projet ; nous avons installé plusieurs bibliothèques comme : Numpy, Matplotlib, Pandas, Sklearn ...Etc.

**Anaconda :**

**Anaconda** est une distribution libre et open source des langages de programmation, Pour faire simple, c’est tout simplement une console.

**Jupyter :**

**Jupyter** est une interface pour python. On peut accéder à ce logiciel en passant par le navigateur d’Anaconda.

**Xampp :**

**XAMPP** est un ensemble de logiciels permettant de mettre en place facilement un serveur Web et un serveur FTP. Il s’agit d’une distribution de logiciels libres (X Apache MySQL Perl PHP) offrant une bonne souplesse d’utilisation, réputée pour son installation simple et rapide.

**PhpMyAdmin :**

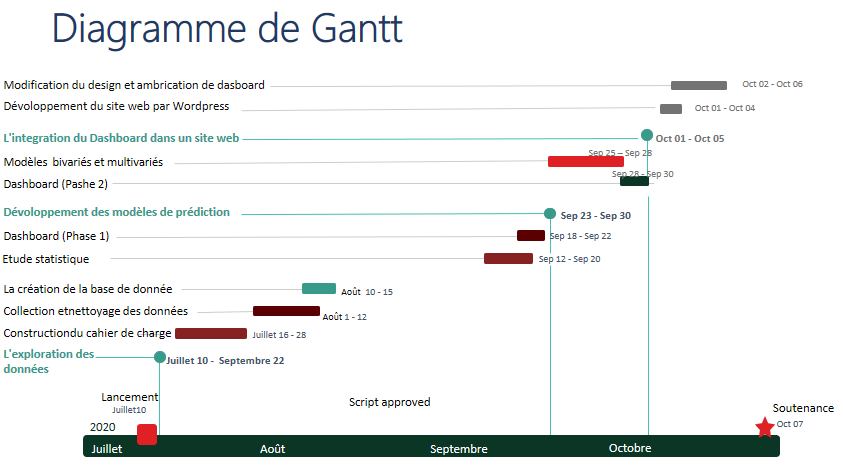
**PhpMyAdmin** est une application web qui permet de gérer un serveur de bases de données MySQL. Dans un environnement multi-utilisateurs, cette interface écrite en PHP permet également de donner à un utilisateur un accès à ses propres bases de données.

**Powerbi :**

C’est un outil très fort et puissant pour visualiser différents types de graphes comme des dashboards et fournir de meilleures informations sur les données.

**WordPress :**

Nous avons intégré notre dashboard dans un site qu’on a le créer avec wordpress, pour qu’il soit public et pour qu’il reste interactif.

**Diagramme de Gantt :**

**Chapitre 2 : La gestion de la base de données**

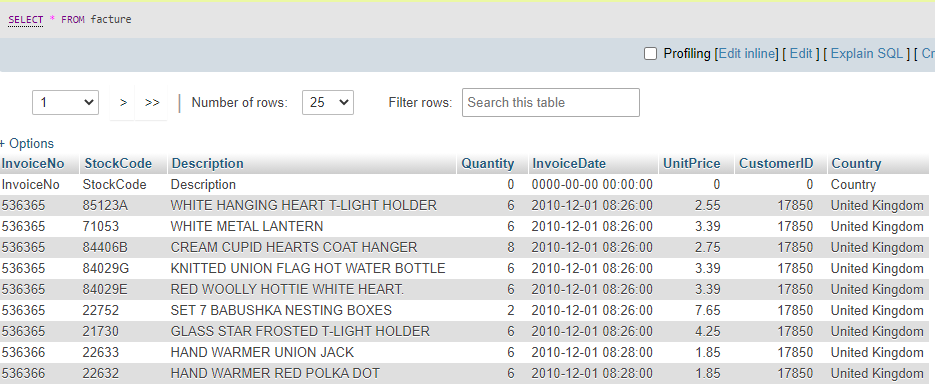
Afin de simplifier et faciliter la gestion de notre dataframe, nous allons conceptualiser une base de donnée pour effectuer des requêtes pour obtenir des informations sur ce que contiennent les tables qui compose la base de donnée. Nous commencerons par des requêtes très simples pour ensuite les complexifier. Les requêtes seront effectuées dans la base de données en utilisant Workbench.



**Importation du fichier CSV a notre table facture :**

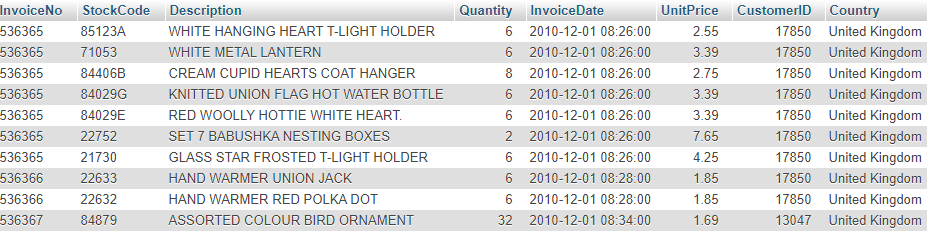
LOAD DATA INFILE ‘data.csv’ INTO TABLE facture FIELDS TERMINATED BY ‘,’ ENCLOSED BY ‘ ” ’ LINES TERMINATED BY ‘\n’;

### **Visualisons la table facture :**



On constate que lorsqu’on a importé notre fichier CSV en MySQL, la ligne des colonnes a été importée aussi.

Donc on doit les supprimer en utilisant la commande ‘DELETE’



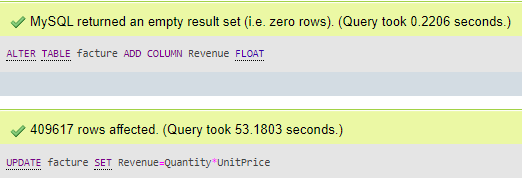
On va augmenter la taille de la colonne **‘description’** de la table facture qui contient le nom des produits, car il y a des produits qui ont des noms très long.

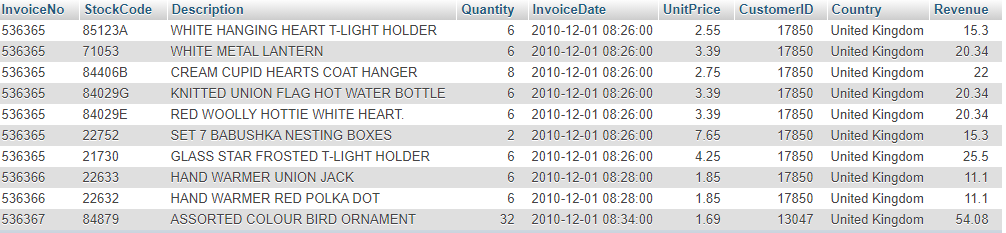
La commande **ALTER TABLE** permet de réaliser ce travail :



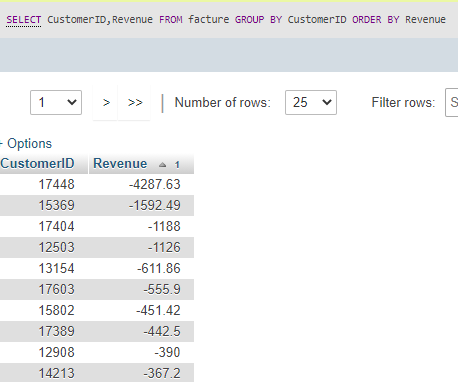
Pour créer le revenu de chaque transaction on doit multiplier la quantité par le prix unitaire de chaque produit :

Premièrement on va créer un nouveau colonne ‘revenue’ de type ‘FLOAT’ en utilisons l’outil ‘ALTER’ et après on va multiplier la quantité par le prix unitaire pour obtenir le revenu total pour chaque transaction.





La fonction **‘GROUP BY’** permet de regrouper la colonne ‘Revenue’ par l’identifiant des clients.

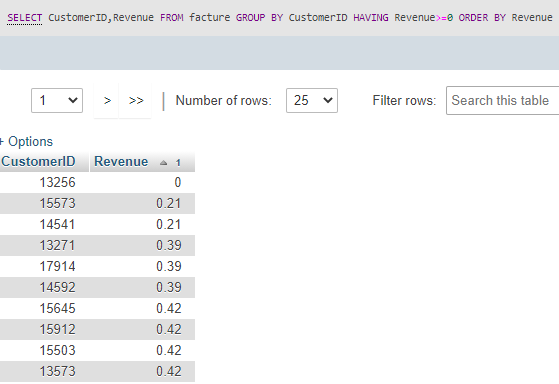


On remarque qu’il y’a des quantités négatives car le tableau contient des revenus négatifs.

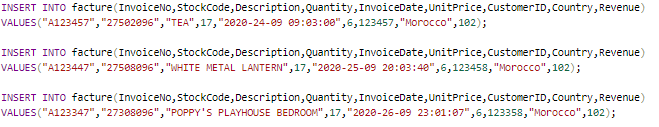
Mais pourquoi ?

Prenant le cas des achats qui nous cause la perte d’argent, par exemple y’a un client qui passe une commande et après quand il reçoit son produit il demande d’annuler la commande, à cause des raisons de l’état de produit ou autres. Nous, quand nous faisons un remboursement pour lui, nous perdons de l’argent parce que nous avons payé pour l’expédition, l’emballage et la publicité etc.

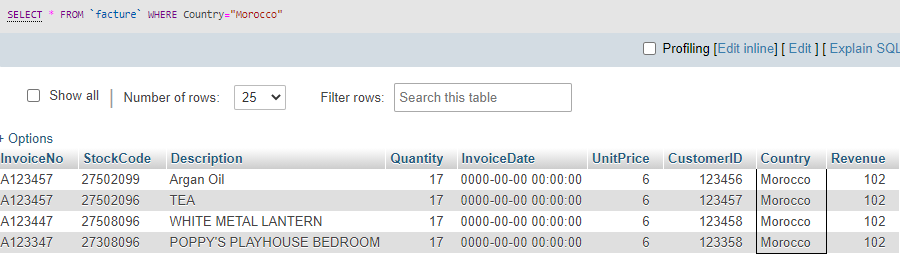
Par la suite, si on veut extraire les revenus, il est préférable d’utiliser la commande ‘HAVING’ pour visualiser juste les revenus et les quantités positives.



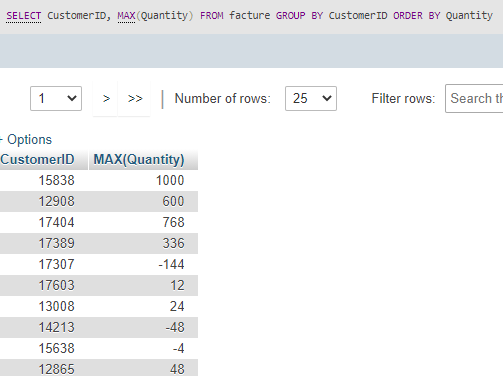
Nous allons insérer quelques enregistrements dans la table facture :



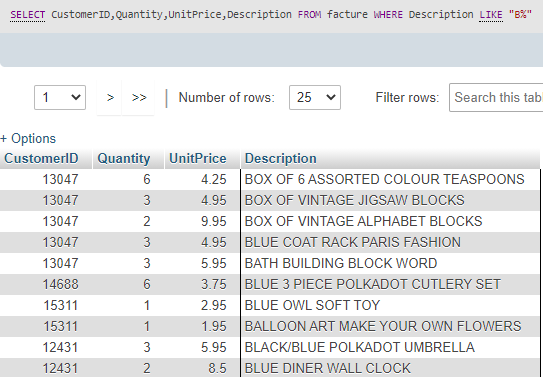
Et voici les commandes des clients marocains :



La commande suivante permet d’obtenir le maximum des quantités achetées pour chaque client.



Pour afficher les produits dont le nom commence par la lettre ‘B’ on va utiliser l’outil **‘LIKE’.**



**Conclusion du premier chapitre :**

Pour ce chapitre, nous avons utilisé **PhpMyAdmin** comme interface de MySQL pour exécuter des requêtes en gérant notre base et au même temps appliquant ce que nous avons exploré dans le module Bases de données.

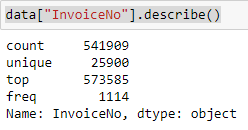
**Chapitre3 :**[**Étude**](https://fr.wikipedia.org/wiki/Statistique#%C3%89tude_d'une_seule_variable) **Statistique**



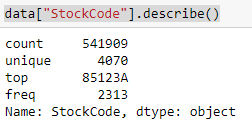
1-Définition de dataframe :

Notre dataframe représente des données d’une entreprise d’e-commerce qui répertorie les achats effectués par environ 4000 clients sur une période d'un an (du 01/12/2010 au 09/12/2011). Il contient huit colonnes :

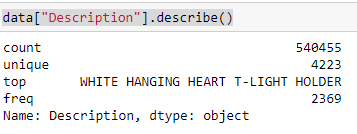
InvoiceNo : Est une colonne qui indique le numéro de la facture ; un nombre entier à six chiffres attribués de manière unique à chaque transaction. Si ce code commence par la lettre 'c', il signifie une annulation de la commande.



StockCode : Et un code unique pour chaque produit, un nombre entier à cinq chiffres.



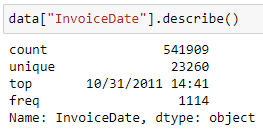
Description : C’est le nom du produit.



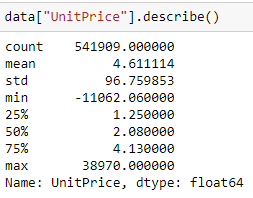
Quantity : C’est la colonne qui contient les quantités de chaque produit par transaction.



InvoiceDate : La colonne InvoiceDate retourne le jour et l'heure auxquels chaque transaction a été générée.

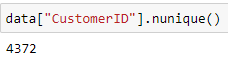


UnitPrice : Est la colonne qui montre le prix du produit par unité.

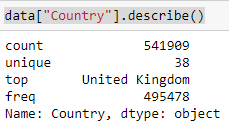


CustomerID : C’est un nombre entier à cinq chiffres attribués de manière unique à chaque client.

La commande ci-dessous nous donne le nombre de clients :



Country : Il s’agit du nom du pays où chaque client réside.

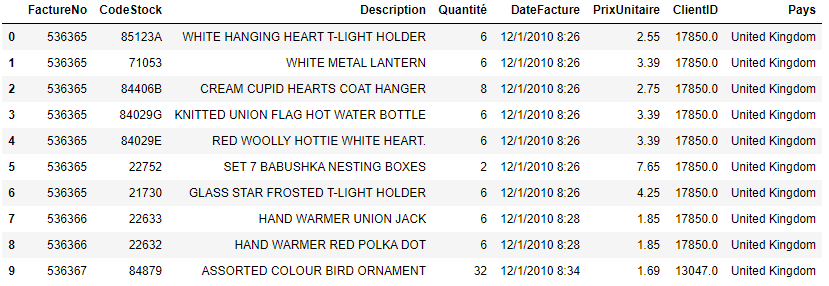


On remarque que le meilleur pays au niveau de nombre des clients est **Royaume-Uni.**

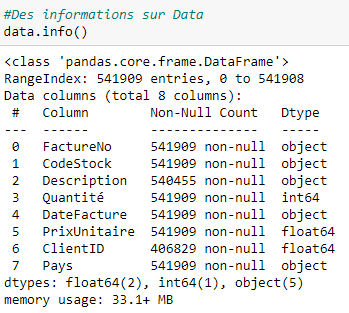
Avant d’afficher les dix premières lignes, on va tout d’abord renommer les colonnes en français en utilisons la commande ‘Rename’.



La figure suivante représente le résultat final de notre dataframe :



2-Quelques informations sur notre dataframe :



3-Filtration et Nettoyage des données :

* Pour accéder à tous les composants de la date (Jours,Mois,années),   
  on doit changer le type de la colonne 'InvoiceDate'. En utilisant la fonction ‘to\_datetime’ et la fonction ‘dt.strftime’ pour extraire l’année et le mois de chaque achat.



* Remplacer les Nan par des 0 par la commande fillna(0). 
* La suppression des doublons en utilisant la commande drop\_duplicates() . 

**4-Comprendre le chiffre d’affaire de l’entreprise :**

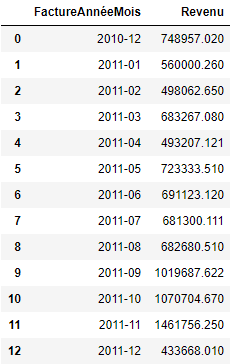
Pour créer le revenu total de chaque client on doit multiplier la quantité par le prix unitaire de chaque produit :



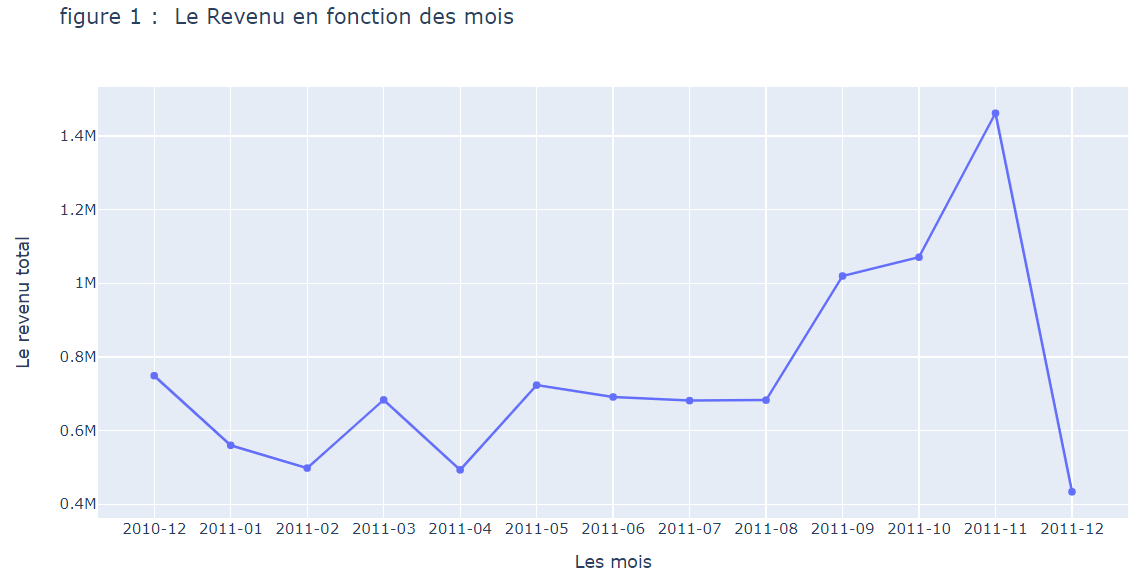
Les données suivantes représentent le regroupement du revenu par mois



L’ajout d’une autre dataframe ‘data\_revenue’ pour visualiser le graphe du revenu.

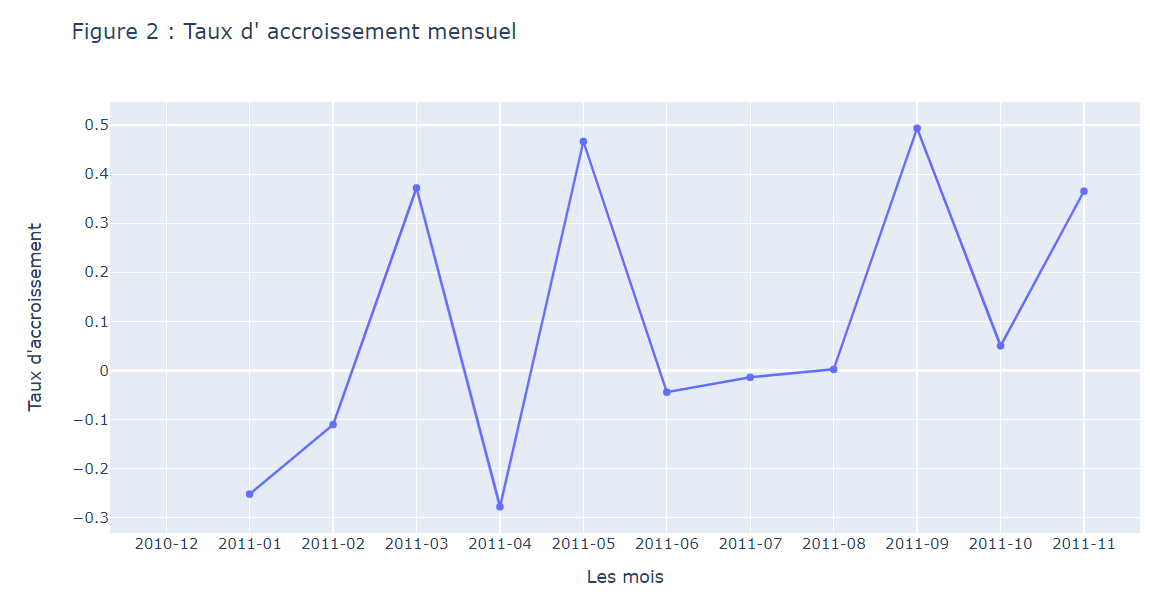


Le graphe suivant montre l’évolution du revenu total en fonction du temps :



Selon ces statistiques on remarque que l’évolution du revenu est en cascade, le maximum et le minimum du revenu ont été enregistrés respectivement en Novembre (1.5M$) et Décembre(0.4M$).

La commande **pct\_change** sert à transférer les données d'une colonne en proportion et c'est le cas pour le revenu Pour visualiser la variation de cette proportion. 

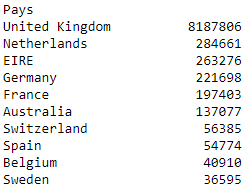


On observe tout d’abord qu’il y’a une forte augmentation du taux d’accroissement dans les mois (2011/01 – 2011-03), (2011/04 - 2011-05), (2011/08 – 2011/09), (2011/10 – 2011/11). Cela montre que cette entreprise ait connu une augmentation dans son chiffre d’affaire dans ces mois, par contre dans les autres mois le taux d’accroissance est négatif, ce qui explique que l’entreprise a rencontrée dans ces périodes des problèmes.

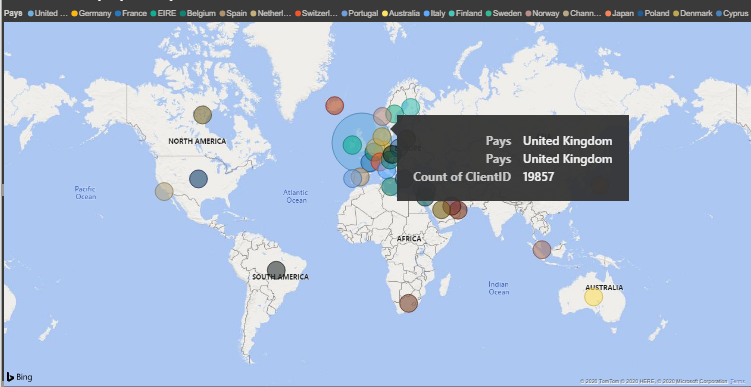
Remarque importante :

D’après les figures 1 et 2, on remarque que lorsque le taux d’accroissement augmente cela ne signifie pas forcément que les revenus de la société augmentent.

La fonction **‘Groupby’** nous permet de visualiser le revenu des clients trié par pays :



La carte géographique suivante indique la distribution des commandes par pays

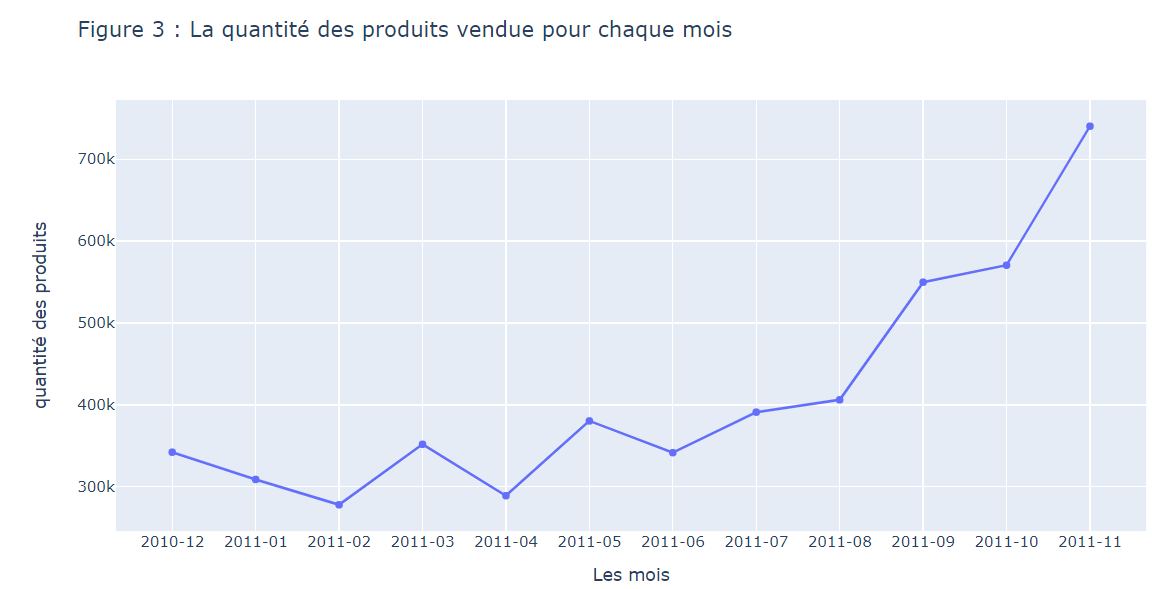


La Carte nous affirme que la majorité des commandes viennent de Royaume-Uni.

**Remarque :**

Dans la suite nous nous concentrons sur ce pays pour maximiser la précision de notre modèle : 

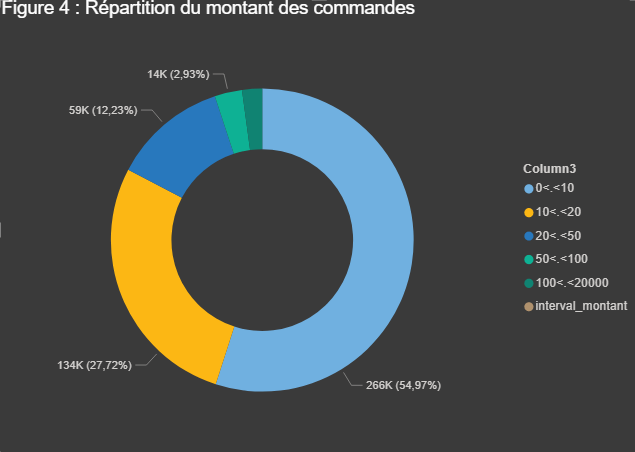
Pour accéder à la quantité vendue, c’est plus praticable de la dessiner en fonction des mois.



**Les chiffres nous montrent** que la quantité des produits augmente (le maximum est 750k pour Novembre 2011) sauf pour février et avril on relève une petite chute.

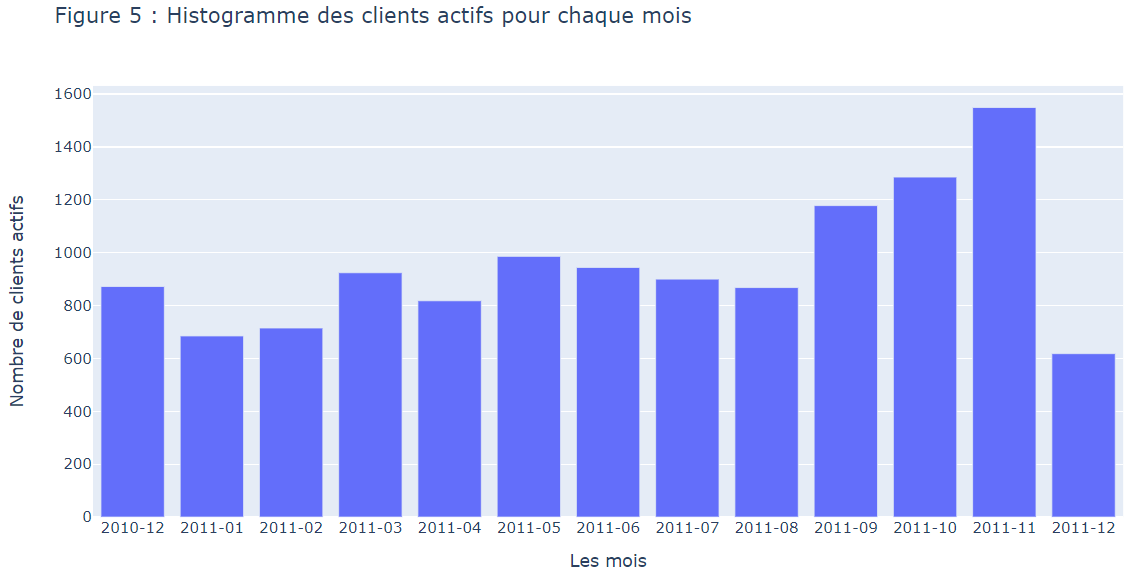
Pour conclure, on peut dire que les clients de cette entreprise adorent les services et les produits présentés par cette entreprise, c’est ce que construira une relation continue de fidélité.

* **Répartition des montants des commandes :**



Le camembert réalisé représente la répartition du montant des commandes. En effet, le graphique apparaitre une faible proportion des commandes des produits à prix élevé (plus de 100$). Par contre dans les produits à bas prix, une large proportion est accordée à le pouvoir d'achat des clients.

5-Clients actifs pour chaque mois



L’histogramme représente le nombre de clients actifs en fonction des mois, c’est-à-dire les clients qu'ont achetés plus d'une fois dans une période bien précise.

Dans les mois 2010/12 – 2011-08 on remarque que l’allure est un peu régulière et croissante, il varie entre 700 et 900. Au contraire, dans les mois 2011/09 – 2011-11 il y’a une importante augmentation de nombre des clients actifs, ce qui explique l’accroissement des revenus de l’entreprise dans cette même période (figure1) et finalement après le mois 2011/12 le nombre des clients actif diminue une autre fois 600 clients exactement comme la figure1, mais cette corrélation n'est pas toujours correcte.

La quantité totale vendue par mois :

L’histogramme ci-dessous représente la quantité totale vendue en fonction des mois.



On remarque que dans les mois 2010/12 – 2011-08 les bâtons sont régulière et constante avec des petites variations et ils commencent à monter et descendre respectivement dans les mois (2011/09 – 2011-11) (2011-11 – 2011-12). Et une autre fois, on revient avec la même conclusion, c’est qu’il y’a une relation entre la quantité vendue et les revenus. Mais pour l'instant, nous ne savons pas si c'est toujours vrai ou non.

6-Clients nouveaux & existants :

Comme vous le savez, au fil du temps, toutes les entreprises cherchent gagner de nouveaux clients, mais aussi à maintenir la relation avec les clients existants. Cependant, chaque entreprise devrait se concentrer sur un seul segment soit les nouveaux ou les existants.

Le choix du segment qui vous convient varie d'un projet à l'autre. Il n'y a pas de règle absolue.

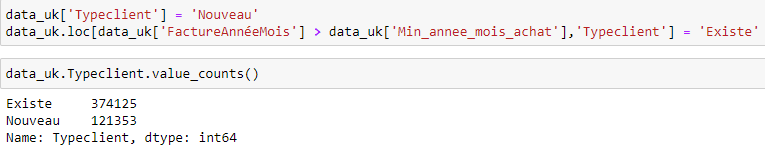
Dans la suite on va définir le segment le plus rentable pour notre entreprise.

Le premier achat de nos clients **:** 



Pour séparer les cas on va ajouter une autre colonne de type Catégorique contenant l’état de chaque client soit **Existant** ou **Nouveau,** mais tout d’abord on va combiner data\_uk et min\_achat en une seule dataframe **(data\_uk = pd.merge (data\_uk, min\_achat, on='CustomerID'))**, afin de dessiner la courbe de chaque catégorie en fonction du revenu.

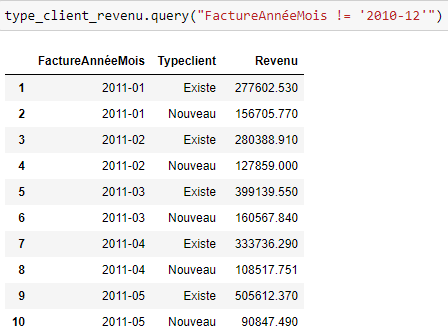
- Si le premier achat (min) d'un client a été effectué avant la date actuelle donc on va le considérer comme un client existant.

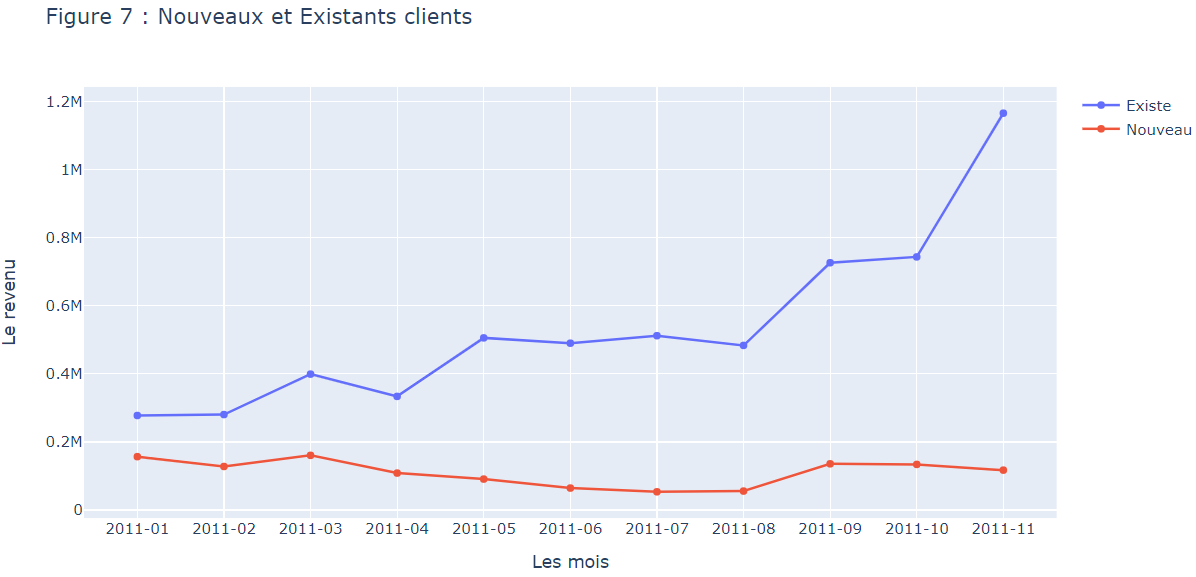


On observe que qu’on a 374125 existants clients et 121353 nouveaux clients. Pour conclure, nous dirons que les nouveaux jouent un rôle très important pour augmenter le revenu de l’entreprise.

Pour récupérer Le revenu total de chaque catégorie en fonction des mois on va bénéficier de la fonction groupby : 

On va supprimer les données de 12-2010 dans type\_client\_revenu parce que c’est le premier mois de lancement du projet, donc il y a aucun client en tant qu'existant :



Pour clarifier la présentation de nos catégories, nous proposons le graphe suivant :

La différence est très claire entre l’évolution du revenu pour les nouveaux et existants clients.

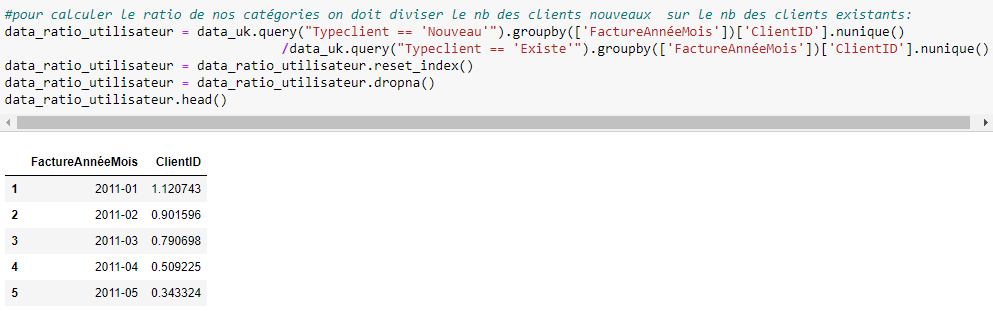
En analysant dans le détail le graphique on se rend compte que le profit de l’entreprise est évolué grâce aux transactions et commandes des existants clients et très bas pour les nouveaux.

Comme action et décision d’analyseurs de l’entreprise, ils vont offrir des promotions pour encourager les clients de joindre, profiter de ses services et ses produits et au même temps proposer des suggestions d’abonnement pour les clients existants.

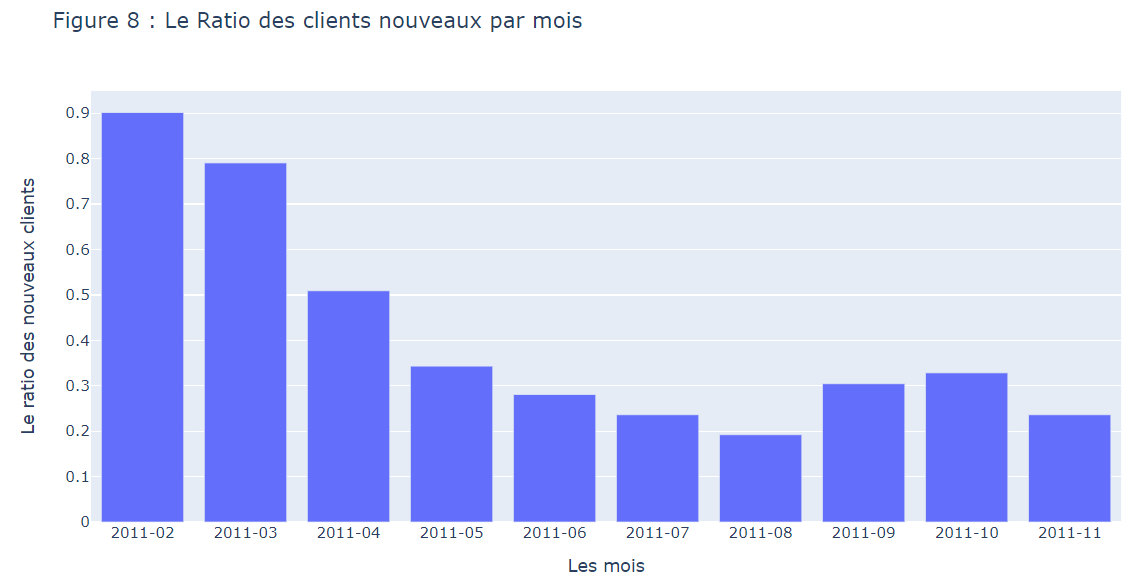
* **Les ratios des nouveaux et existants clients :**

Pour affirmer les résultats obtenus dans cette section, on va voir les ratios de ces deux types des clients dans notre dataframe.

Pour calculer le ratio de nos catégories on doit diviser le nombre des clients nouveaux sur le nb des clients existants :



Traçons ces données pour avoir une meilleure vue sur les deux catégories :



Pour février, le ratio de nouveaux clients est de 90%, car tous les clients acquis par l'entreprise sont des nouveaux. Mais comme prévu, ce ratio a commencé à diminuer après ce mois et se situait autour de 20%.

**7-Taux de rétention mensuel :**

Maintenant, nous aborderons le taux de rétention mensuel. Mais que représente ce taux ? et quels sont ses avantages par rapport à notre entreprise ?

Le taux de rétention indique à quel point votre service s'adapte au marché. Il s’agit du pourcentage de clients qui manifestent leur satisfaction envers votre compagnie par diverses actions : nouveaux achats, achats supplémentaires, upgrades ou réabonnements.

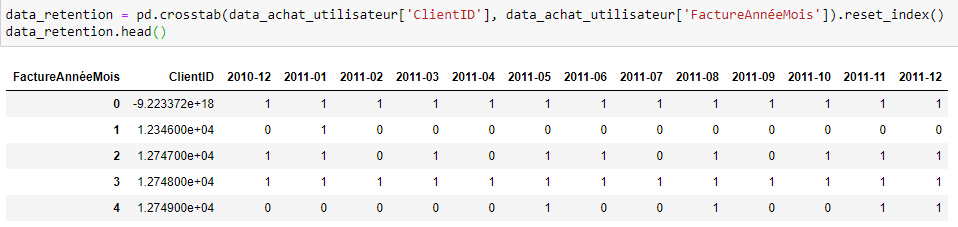
C’est une démarche qui nous permet de vérifier que nous ne perdons pas de clients aussi rapidement.

Pour visualiser le taux de rétention mensuel, nous devons calculer le nombre de clients conservé le mois précédent.  
**Taux de rétention mensuel = clients retenus à partir du mois précédent / Total des clients actifs.**

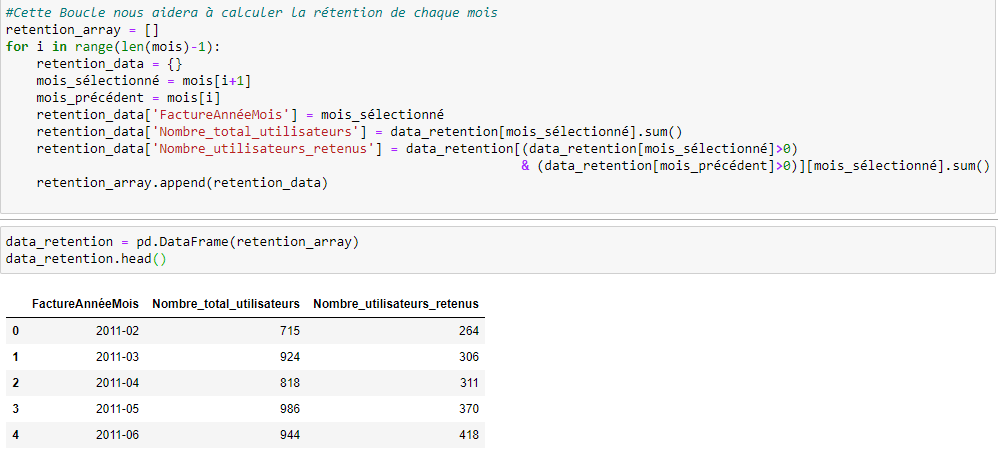
**Commençons par le calcul du nombre de clients retenus à partir du mois précédent et le total des clients actifs.**



**Dans le tableau au-dessus nous avons chaque client attaché au budget qu'il consacre à nos produits chaque mois.**

**Nous allons créer le tableau de contingence, afin que nous puissions savoir si un client effectue un achat dans le mois spécifié (1 si un client effectue un achat, 0 sinon)** ****

Enfin, nous calculerons les totaux des colonnes pour connaître le nombre de clients retenus par rapport au nombre total de clients.



**Une lecture du tableau en dessus :**

Prenons la première ligne : Parmi 715 clients, 354 ont déjà achète le mois dernier.

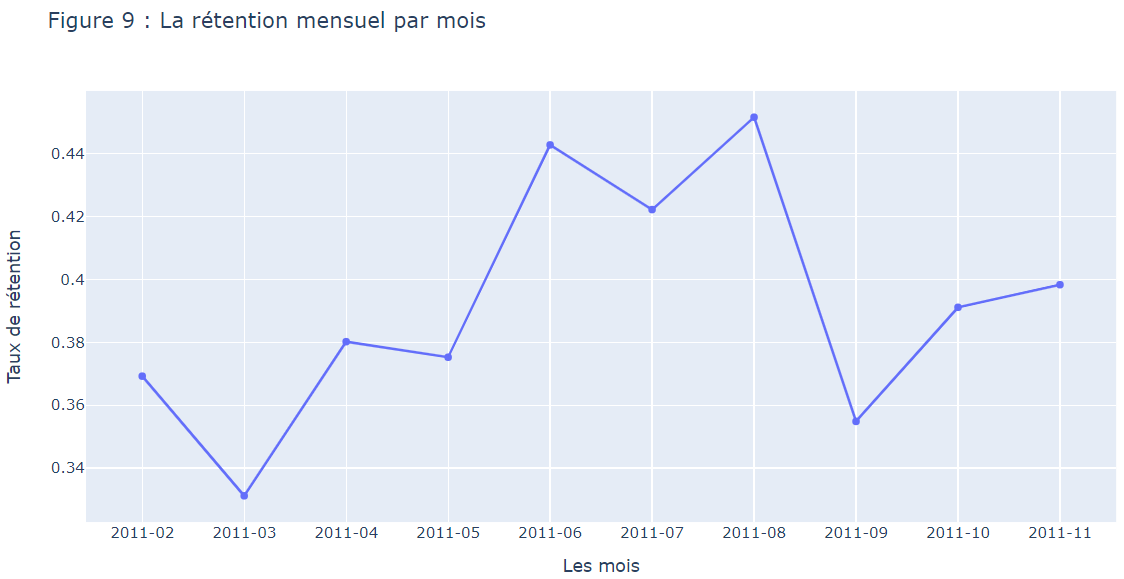
**Remarque :**

Dès les cinq premières lignes, on remarque que le nombre de clients retenus a commencé à augmenter de mois en mois, mais est-ce que c’est le cas pour tous les mois ?

Pour répondre à cette question, nous devons tracer le ratio des clients retenus par rapport au nombre total des clients.

Commençons par calculer ce ratio :





D’après ce graphe, on remarque que le progrès de la rétention est en cascade, à chaque mois l’entreprise reçoit énormément de commandes, est c’est ce qui justifie l’augmentation de la quantité des produits vendue chaque mois.

La décroissance constatée pour septembre 2011 est due à l’accumulation du revenu de nouveaux clients.

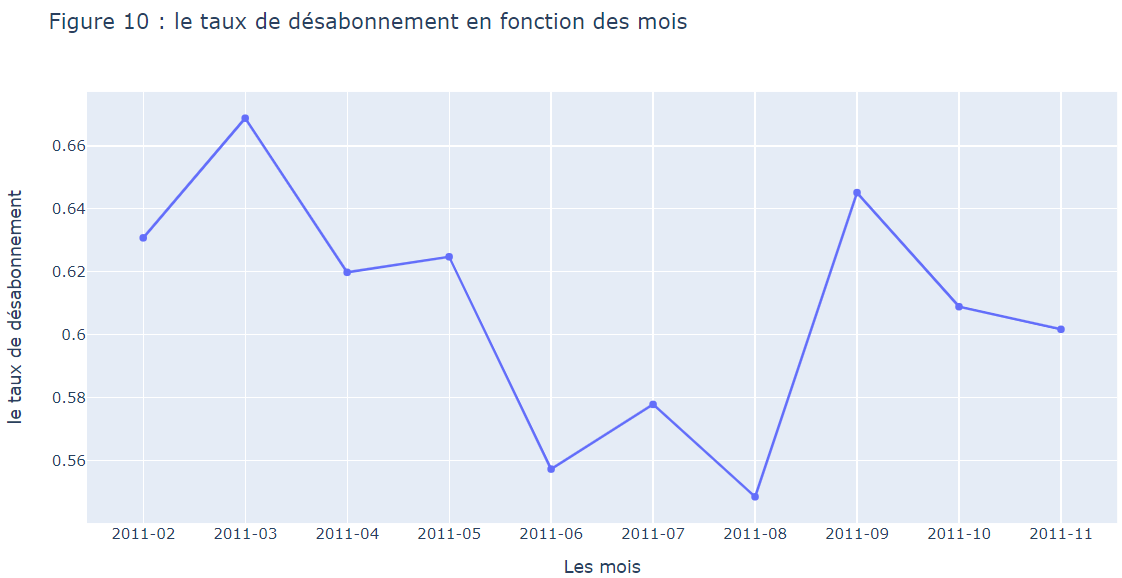
**8-Le taux de désabonnement :**

Nous avons traité la rétention et pour l’adverse on va analyser ce qu’on appelle le taux de désabonnement.

**Néanmoins,** **Qu'est-ce que signifie le taux de désabonnement ?**

Le taux de désabonnement, également appelé taux d'attrition ou de désabonnement de la clientèle, est le taux auquel les clients arrêtent de faire affaire avec une entité et de passer une commande.



La courbe suivante représente le taux de désabonnement :

Basant sur cette représentation, on peut dire que le taux de désabonnement a commencé à diminuer à partir de Mars, et ça commence à augmenter à dater d'août.

Faisons une liaison entre le taux de rétention et le taux de désabonnement, on note que la raison pour laquelle la rétention diminue entre août et septembre est l’accroissement de taux de désabonnement.

**Conclusion du 2ème Chapitre :**

Dans ce chapitre, nous avons découvert de nombreuses fonctionnalités et paramètres qui nous ont aidés à découvrir l’entreprise et de savoir quelle est sa puissance et les talents nécessaires qu’il doit améliorer pour augmenter par exemple le taux de rétention, le pourcentage des nouveaux clients et diminuer le taux de désabonnement.

On cite, les interprétations suivantes :

* Le revenu a été maximisé en Novembre (1461756$) donc les clients sont actifs dans ce mois et
* La majorité des commandes viennent de Royaume-Uni.
* 55% des commandes leurs montant est compris entre 0 et 10$.
* Le profit de l’entreprise est évolué grâce aux transactions et commandes des existants clients et très bas pour les nouveaux.
* Le ratio des nouveaux clients est très élevé pour février.

**Chapitre 4 : La segmentation par la méthode RFM**

**3-1. Introduction :**



La segmentation RFM permet aux spécialistes du marketing de cibler des clusters spécifiques de clients avec des communications beaucoup plus pertinentes pour leur comportement particulier et de générer ainsi des taux de réponse beaucoup plus élevés, ainsi qu'une fidélité accrue et une valeur à vie des clients.  
Comme d'autres méthodes de segmentation, la segmentation RFM est un moyen puissant d'identifier des groupes de clients pour un traitement spécial. RFM est synonyme de récence, fréquence et monétaire.

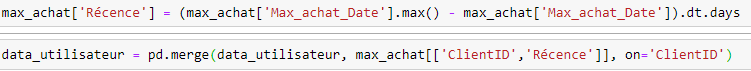
# 3-1-1. Récence :

# Récence : la date du dernier achat. Notez bien que l’on part du principe qu’une personne qui a acheté récemment chez vous a plus de chances de revenir commander chez vous.

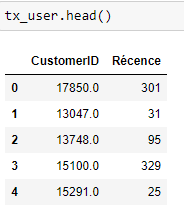
# Tout d’abord, on va créer une nouvelle dataframe dans laquelle on va intégrer chaque paramètre d’RFM avec leur cluster :

# Nous devons utiliser la fonction max pour extraire la date du dernier achat :

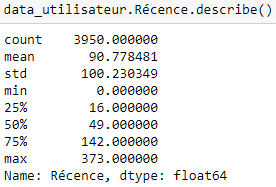
# 

Voilà ce que la récence signifie ; c'est le présent moins le dernier achat, et puisque nos données reviennent des années 2010/2011, donc le présent est le maximum des achats : On a exploité la fonction **dt.days** pour afficher les résultats sous forme des chiffres indiquant les jours.

Ce sont les cinq premières lignes. Prenons le premier cas, le client qui a 17850 comme identifiant, cela fait 301 jours qu'il n'a pas acheté chez nous, et donc un.



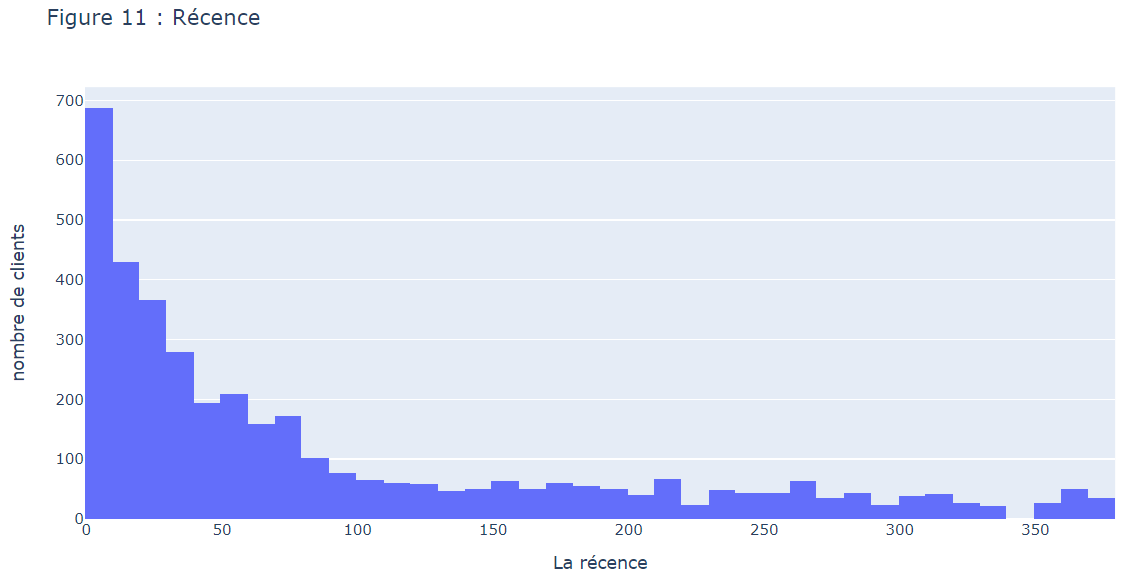
Utilisons l'outil **‘describe’** pour connaitre comment la récence est distribuée autour de nos données



D’après ces chiffres on constate que la moyenne des récences pour nos clients est 90jours avec un écart type d'environ 100 jours. En moyenne, la durée qui s'écoule jusqu'à le nouvel achat d'un client présente un écart d'environ 100 jours par rapport à la moyenne (90 jours).

**Remarque :** Même si la moyenne est de 90 jours de récence, la médiane est de 49.

Ci-dessus l'histogramme pour nous montrer comment la récence a été distribuée autour de nos clients.



L’allure indique une distribution décalée à gauche de la médiane, et donc une queue de distribution étalée vers la droite.

Maintenant, il est vrai que nous savons la récence de chaque client. Cela signifie que nous pouvons savoir la dernière fois qu'il a acheté chez nous.

Passons à la réalité, comme vous savez que nous avons plus de 4000 clients, supposons que nous souhaitons offrir une promotion aux clients qui n'ont pas acheté chez nous depuis 90 jours (récence égale à 90), comment pouvons-nous les distinguer des autres catégories ? Très difficile, n'est-ce pas ? C'est pourquoi dans ce qui suit nous présenterons un outil très puissant, qui permet de grouper ces clients dans des catégories spécifiques.

* **K-means :**

K-means (ou K-moyennes) : C’est l’un des algorithmes de clustering les plus répandus. Il permet d’analyser un jeu de données caractérisées par un ensemble de descripteurs, afin de regrouper les données “similaires” en groupes (ou clusters).

La similarité entre deux données peut être inférée grâce à la “distance” séparant leurs descripteurs ; ainsi deux données très similaires sont deux données dont les descripteurs sont très proches. Cette définition permet de formuler le problème de partitionnement des données comme la recherche de K “données prototypes”, autour desquelles peuvent être regroupées les autres données.

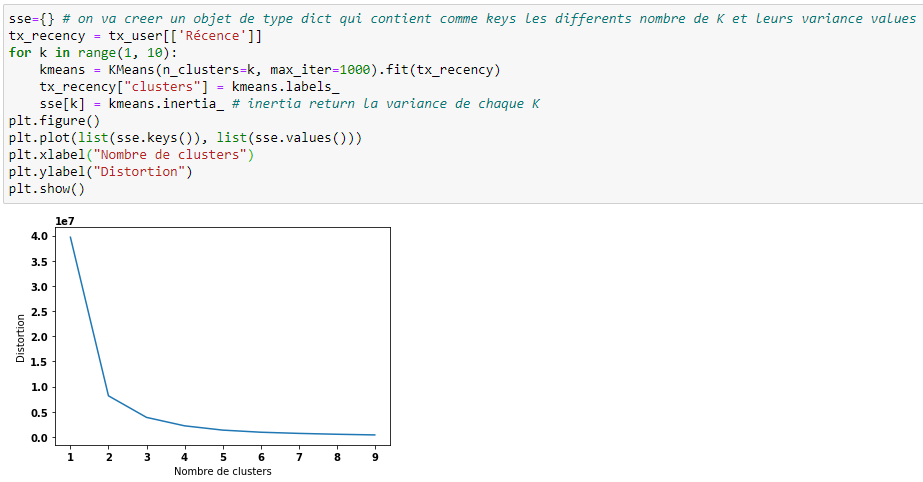
Comment définir le nombre des groupes ?

Pour un même jeu de données, il n’existe pas un unique clustering possible. La difficulté résidera donc à choisir un nombre de clusters K qui permettra de mettre en lumière des patterns intéressants entre les données. Malheureusement, il n’existe pas de procédé automatisé pour trouver le bon nombre de clusters.

La méthode la plus usuelle pour choisir le nombre de clusters est de lancer K-Means avec différentes valeurs de K et de calculer la variance des différents clusters. La variance est la somme des distances entre chaque centre d’un cluster et les différentes observations inclues dans le même cluster. Ainsi, on cherche à trouver un nombre de clusters K de telle sorte que les clusters retenus minimisent la distance entre leurs centres (centroids) et les observations dans le même cluster.

Et c'est ce que nous viendrons de faire dans le graphe suivant :

Ce graphique est souvent appelé elbow graphe

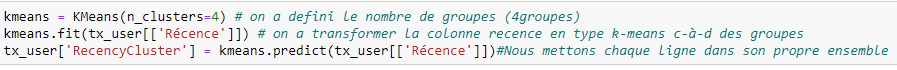


Le k optimal pour vos données est celui du nombre de clusters à partir duquel la variance ne se réduit plus significativement. En effet, la “chute” de la courbe de variance (distortion) entre 1 et 3 clusters est significativement plus grande que celle entre 5 clusters et 9 clusters.

Finalement, le choix (au vu de ce graphique), entre 3 ou 4 clusters, reste un peu flou et à votre discrétion. Le choix se fera en fonction de votre jeu de données et ce que vous cherchez à accomplir. Finalement, Il n’y a pas de solution unique à un problème de clustering.

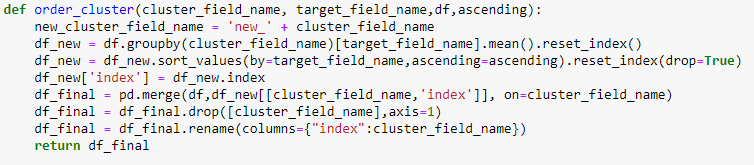
Pour notre cas, on a choisi n\_clusters=4.

Maintenant, quand nous connaissons le nombre de K approprié pour nos données, nous pouvons commencer à appliquer la méthode k-means.



Voilà, nous avons réussi à construire des groupes en utilisant la méthode des k-means. Il ne nous reste plus qu'à les classer selon un ordre.

Cette fonction nous aidera à le faire. Elle prend deux paramètres, cluster\_field\_name (l’objet qu’on veut classifier), target\_field\_name (l’ordre de classification), df (dataframe) et asending (croissant ou décroissant)



Maintenant on va répéter le même processus pour la fréquence et le montant.

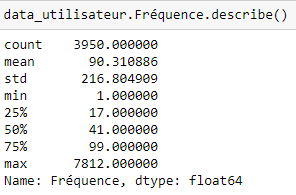
**3-1-2. Fréquence**

**Fréquence :** le nombre d’achats réalisé sur une période donnée. Plus un client achète régulièrement chez vous, plus il y a de chances pour qu’il achète à nouveau. Vous l’avez compris, on analyse ici son niveau de fidélité.

# Pour créer des groupes de fréquences, nous devons trouver le nombre total de commandes pour chaque client. Calculez d'abord cela et voyez à quoi ressemble la fréquence dans notre base de données clients :

Ce sont les cinq premières lignes. Prenons le premier cas, le client qui a 12748 comme identifiant a acheté 4642 fois chez nous.

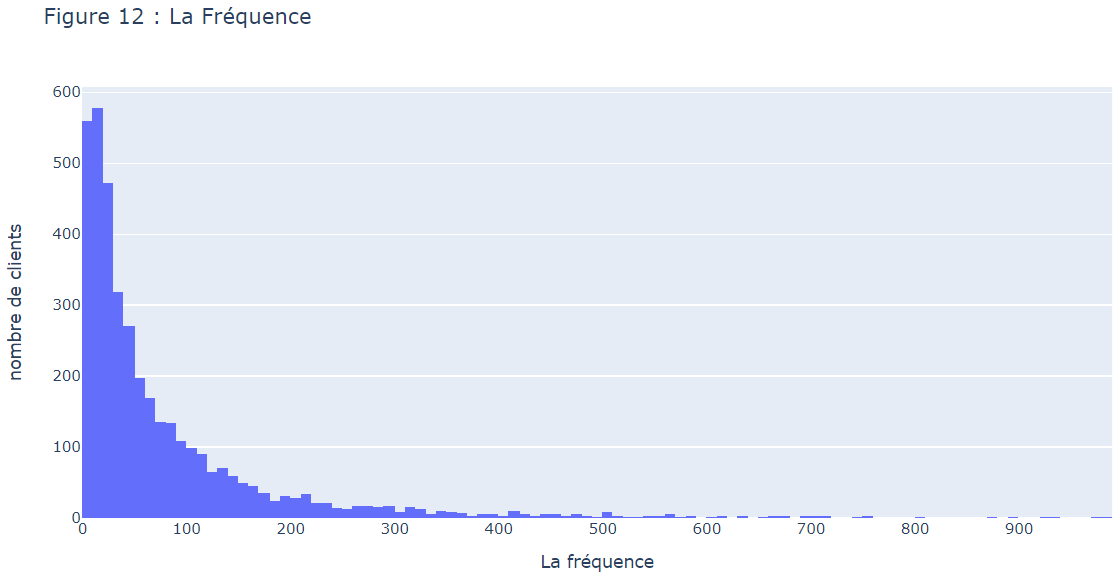
Nous utilisons l'outil ‘describe’ pour voir comment la fréquence est distribuée autour de nos données :



On voit que la moyenne des fréquences pour nos clients est 90 achats avec un écart type d'environ 216 achats. En moyenne, le nombre des achats passe un client chez nous présente un écart d'environ 216 commandes par rapport à la moyenne (90 commandes).

**Remarque :** même si la moyenne est de 90 jours de récence, la médiane est de 41

Ci-dessus l'histogramme pour nous montrer comment est la distribution de la fréquence parmi nos clients.

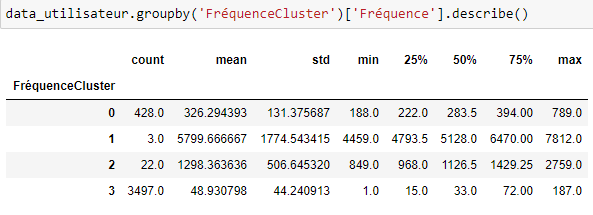


L’allure indique une distribution décalée à gauche de la médiane, et donc une queue de distribution étalée vers la droite.

Nous regroupons ces données de fréquence en groupes.



Appelons la fonction **order\_cluser** pour les classifier : 

On a regroupé chaque cluster par **[Fréquence]** en ajoutant la fonction **describe** pour décrire chaque groupe : 

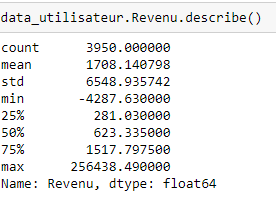
En comptant sur ce tableau, il est très clair l’utilité des clusters à travers les vastes différences entre les moyennes de chaque cluster et l’espacement entre toutes les échelles.

# 3-1-3. Montant :

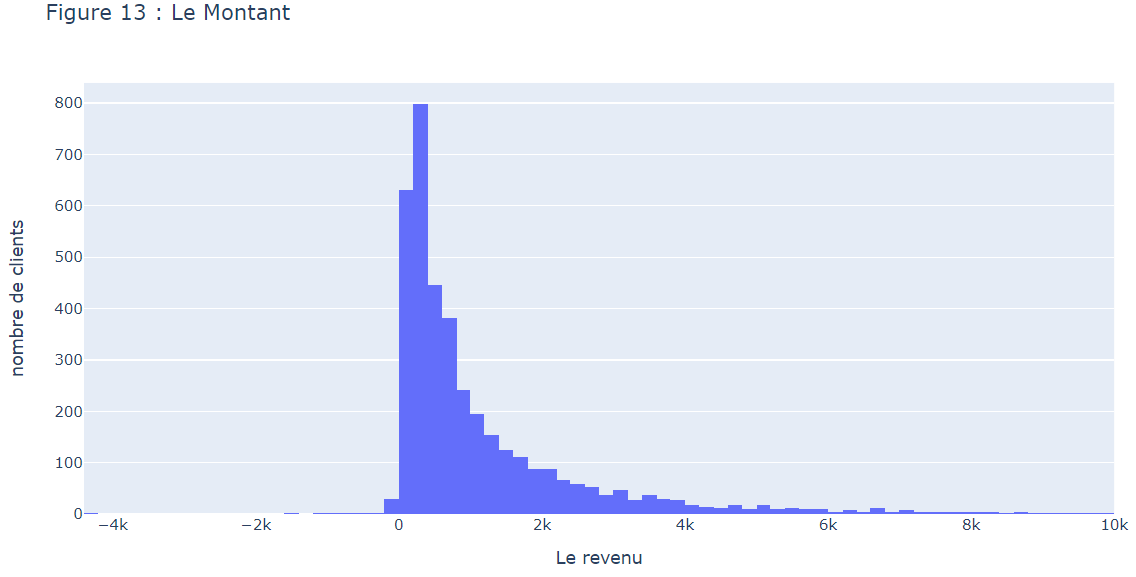
**Montant :** la somme des achats cumulés sur une période donnée. Les gros acheteurs répondent mieux que les petits.

On mesure ici la valeur client.

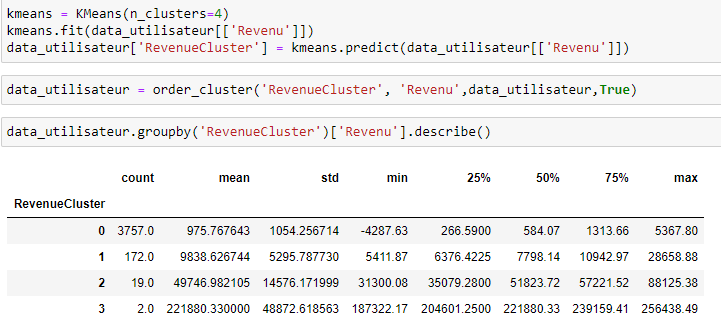
**Rappel :** Le revenu du client est le prix unitaire de chaque produit multiplié par la quantité vendue : 



La moyenne du revenu est 1708.14 $ mais la médiane est 623.33 $, la différence de la distance entre la moyenne et les points est l’écart type=6548.6.

L'histogramme suivant nous montre la variation du montant ; comme échantillon **le revenu < 10000** : 

Nous remarquons que la majorité des clients ont un revenu entre 0 et 2000 $.

-Regroupons maintenant ces données de revenu en groupes et appelons la fonction **order\_cluser** pour l’arrangement des points : 

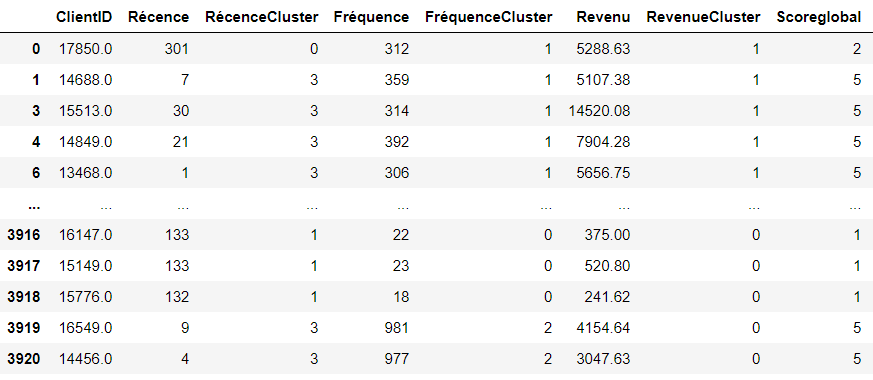
Jusqu’à ici on a appliqué les clusters de la récence, la fréquence et le montant, mais l’inconvénient c’est que on a séparé les composant d’RFM et on a traité chaque élément tout seul.

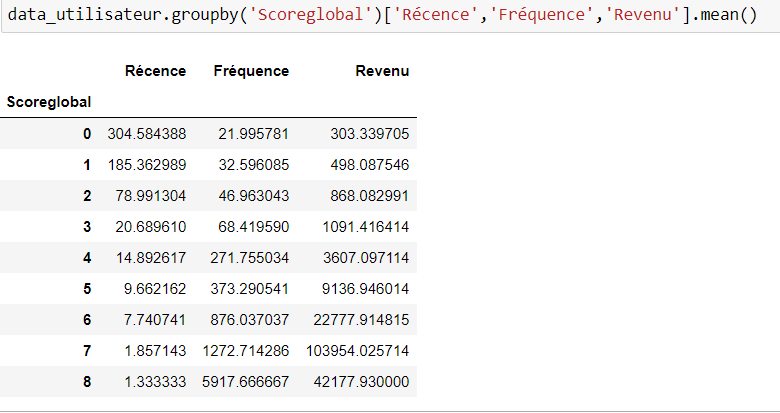
Ce qui demande beuacoup d’analyses et d’interprétation.

# Pour éviter cet obstacle on va créer ce qu’on appelle : Segmentation Globale.

# 3-2. Segmentation Globale :

Pour globaliser et généraliser la visualisation on va utiliser ‘Scoreglobal’ comme sommation de RécenceCluster, FréquenceCluster et RevenuCluster : 

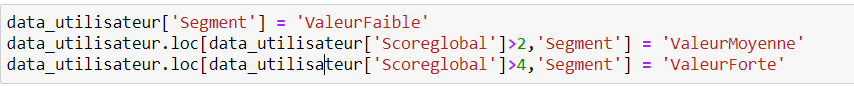
Et donc Data\_utilisateur contient les colonnes suivantes :

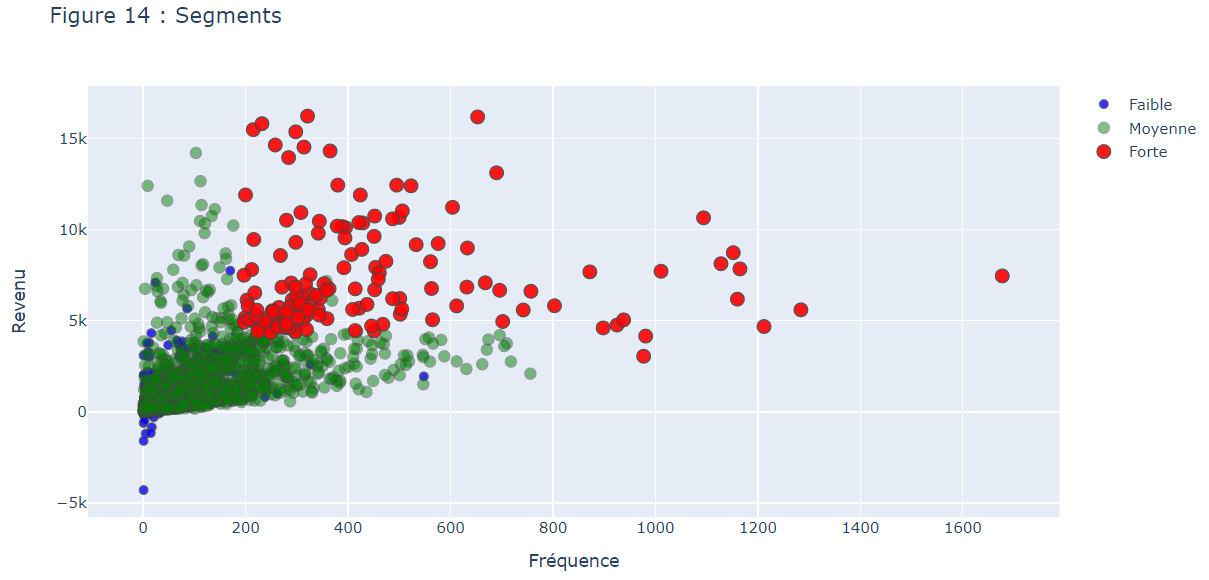
Et pour mieux comprendre la moyenne de récence, fréquence et revenu de chaque nombre de ce score on va utiliser la fonction **groupby :**

# On peut aussi utiliser la fonction count() pour savoir nombre de clients inclus dans chaque échelle et dans l’exemple suivant c’est pour la récence :

Afin de simplifier la colonne scoreglobal, il est préférable de créer une autre colonne **“Segment”** de type catégorique en affectant à chaque client un des mots suivants :

* ValeurFaible
* ValeurMoyenne
* ValeurForte



**Tout de suite nous allons designer des courbes pour nos segments :** 

* **Revenu VS Fréquence :**

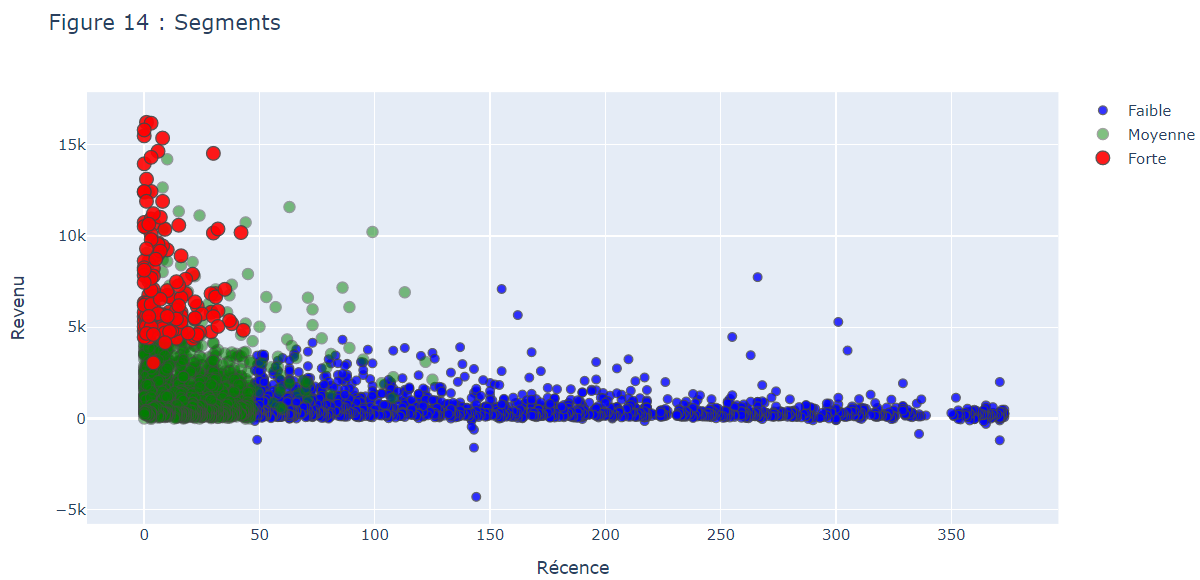
La figure 13 représente La distribution de chaque ensemble.

**Les clients avec valeur faible :** ont une fréquence entre 0 et 200 et un revenu entre –100$ et 800$.

**Les clients avec valeur moyenne :** ont une fréquence entre 0 et 700 et un revenu entre 0$ et 7000$.

**Les clients avec valeur forte :** ont une fréquence entre 200 et 1200 et un revenu entre 5000$ et 15000$.

* **Revenu VS Récence :**



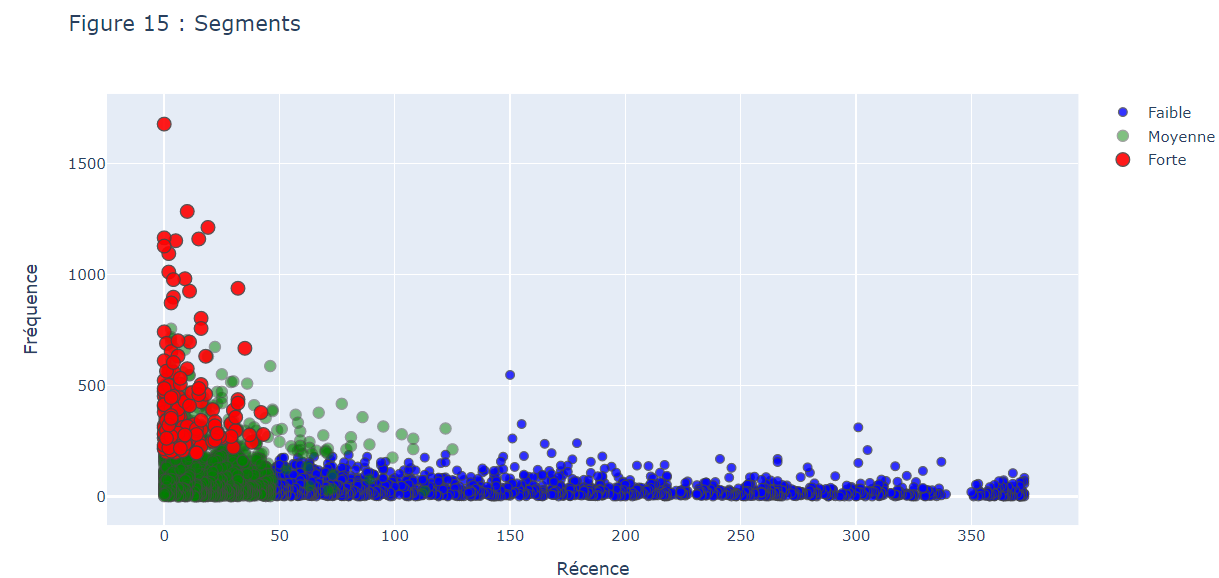
La figure 14 représente La distribution de chaque ensemble.

**Les clients avec valeur faible :** ont une Récence entre 50 et 400 et un revenu entre –100$ et 800$.

**Les clients avec valeur moyenne :** ont une Récence entre 0 et 150 et un revenu entre 0$ et 7000$.

**Les clients avec valeur forte :** ont une Récence entre 0 et 50 et un revenu entre 5000$ et 16000$.

* **Récence VS Fréquence :**



La figure 15 représente La distribution de chaque ensemble.

**Les clients avec valeur faible :** ont une Récence entre 50 et 400 et une Fréquence entre 0 et 100.

**Les clients avec valeur moyenne :** ont une Récence entre 0 et 130 et une Fréquence entre 0 et 700.

**Les clients avec valeur forte :** ont une Récence entre 0 et 50 et une Fréquence entre 200 et 1200.

**Conclusion de la 3ème Partie :**

Nous avons créé notre module pour prédire à quelle catégorie chaque client appartient :

* La moyenne des récences pour nos clients est 90jours.
* 720 clients ont la récence entre 0 et 9 jours.
* La moyenne des fréquences pour nos clients est 91 achats.
* La majorité des clients ont une fréquence entre 0 et 200.
* La majorité des clients ont un revenu entre 0et 2000$.

**Chapitre 5 : LA VALEUR NETTE À VIE DES CLIENTS (CLV) :**

**Définition :**

La valeur client est le terme français qu’on peut traduire par « Customer lifetime value ». C’est une notion qui mesure à un instant 't' la somme des revenus générés sur la durée de vie d’un client, passée et à venir.

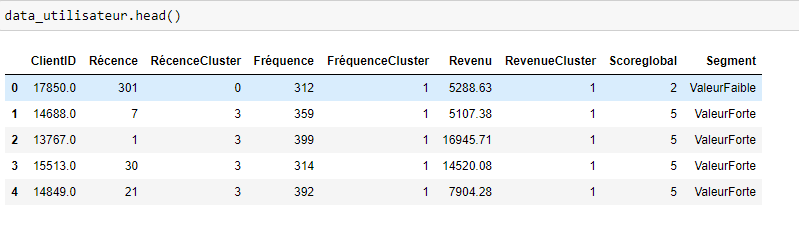
CLV est utilisée en marketing et en gestion afin d’appréhender la rentabilité d’un client sur toute sa période de vie : les revenus qu’il génère (ventes ou marge) peuvent ainsi être comparés aux dépenses et aux moyens mobilisés pour l’acquérir et le fidéliser.

' La valeur client ' est une valeur monétaire qui représente le montant des revenus ou des bénéfices qu’un client donnera à l’entreprise au cours de la période de la relation ». CLV démontre les implications de l’acquisition de clients à long terme par rapport aux clients à court terme. Cette technique peut vous aider à répondre aux questions les plus importantes sur les ventes à chaque entreprise :

Comment identifier les clients les plus rentables ? Comment une entreprise peut-elle offrir le meilleur produit et gagner le plus d’argent ? Comment segmenter les clients rentables ? Combien de budget faut-il dépenser pour acquérir des clients ?

Il y’a beaucoup d’approches disponibles pour le calcul CLV nous avons besoin de données historiques des clients, mais vous ne serez pas en mesure de calculer pour les nouveaux clients. Pour résoudre ce problème Business Analyste développe des modèles d’apprentissage automatique pour prédire le CLV des nouveaux clients. Explorons l’approche suivante pour le calcul CLV :

Créez un modèle de régression pour les clients existants. Prenez les données récentes sur six mois comme variables indépendantes et les revenus totaux sur trois ans comme variable dépendante et établissez un modèle de régression à partir de ces données.

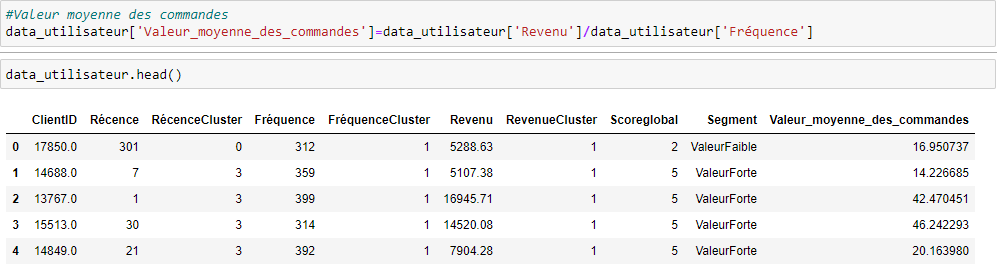
On va exploiter notre dataframe 'data\_utilisateur' :

**On va appliquer la relation suivante :**

* + **CLV = ((Valeur moyenne des commandes x Fréquence d’achat) /Taux de désabonnement) x Marge bénéficiaire.**

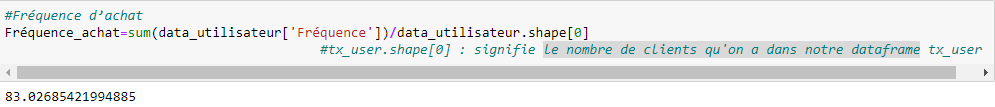
**Commençons par le calcul de 'Valeur moyenne des commandes'.**

**Pour calculer cette valeur on doit diviser le revenu total pour chaque client sur le nombre d’achat qu’il a passé (Fréquence).**



**Calculons la Fréquence d’achat :**

**Comment ? Il faut juste diviser la fréquence sur le nombre de clients qu'on a dans notre dataframe.**



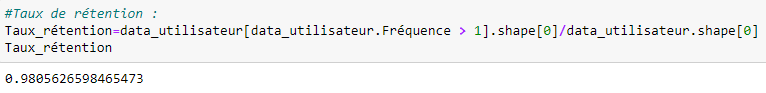
**Taux de désabonnement :**

**Taux de désabonnement, lorsqu'il est appliqué à une base clientèle, fait référence à la proportion de clients contractuels ou des abonnés qui quittent un fournisseur au cours d'une période donnée. Il est considéré comme un indicateur possible d'insatisfaction des clients.**

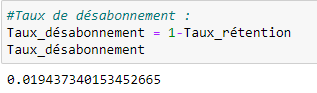
**Tout simplement, c'est juste l'opposé du taux de rétention que nous avons couvert dans la partie précédente.**

**Rappel : Le taux de clients retenus à partir du mois précédent.**

**Alors pour calculer le taux de désabonnement, il faut tout d’abord calculer le taux de rétention.**

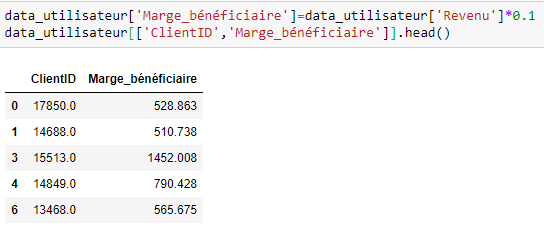


**Le taux de désabonnement est juste ‘1- taux de rétention’.**



La marge bénéficiaire :

La marge bénéficiaire est le ratio de rentabilité couramment utilisé. Il représente le pourcentage des ventes totales gagnées comme gain (Profit). Et puisque nous n'avons pas ces informations dans nos données, nous supposerons que notre entreprise a environ 10% de profit sur la vente totale.



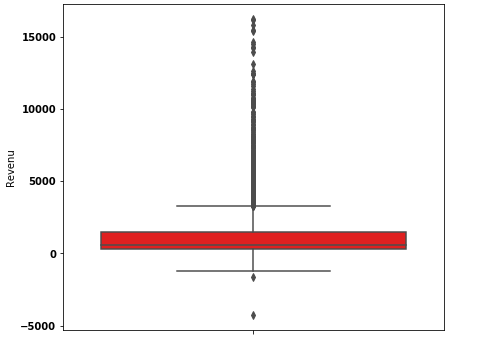
**Voilà, maintenant nous avons implémenté tous les paramètres pour calculer 'CLV'.**

**Mais avant de faire cela, nous supprimerons les valeurs aberrantes car dans ce qui suit, nous utiliserons le CLV pour créer un modèle de prédiction, c'est pourquoi nous devons supprimer cette valeur pour augmenter la précision de notre modèle.**

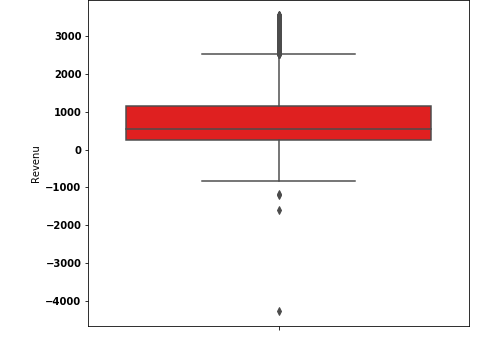
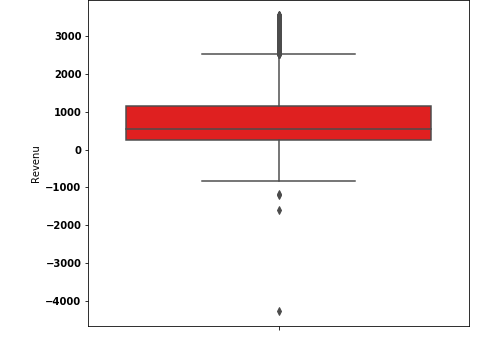
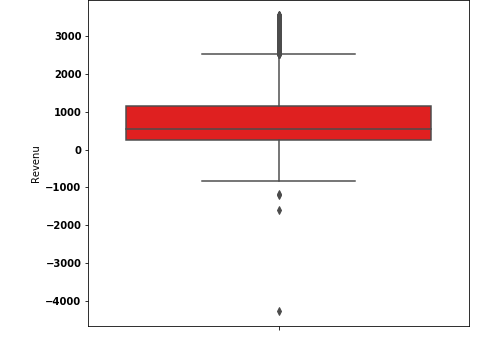
**Et pour y parvenir, nous supprimerons les quartiles de notre data. On utilise la commande :**



**Le Boxplot avant de supprimer les valeurs aberrantes :**



**Le Boxplot après la suppression des valeurs aberrantes :**



**Nous avons réussi à nous débarrasser de la plupart des valeurs aberrantes .**

**Calculons maintenant le CLV.**



**Lecture de tableau :**

**Le client qui a 13047 comme un identifiant a dépensé 67104$ profit pendant la période de vie où il était en relation avec notre entreprise.**

**1-Modèle bi-varié pour prédire la quantité totale vendu en mois prochain :**

Nature des variables :

La variable explicative X est : la quantité totale vendu en mois actuel.

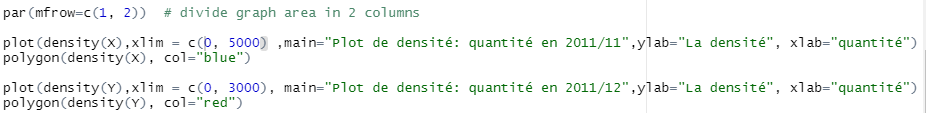
La variable dépendante Y est : la quantité totale vendu en mois prochain.

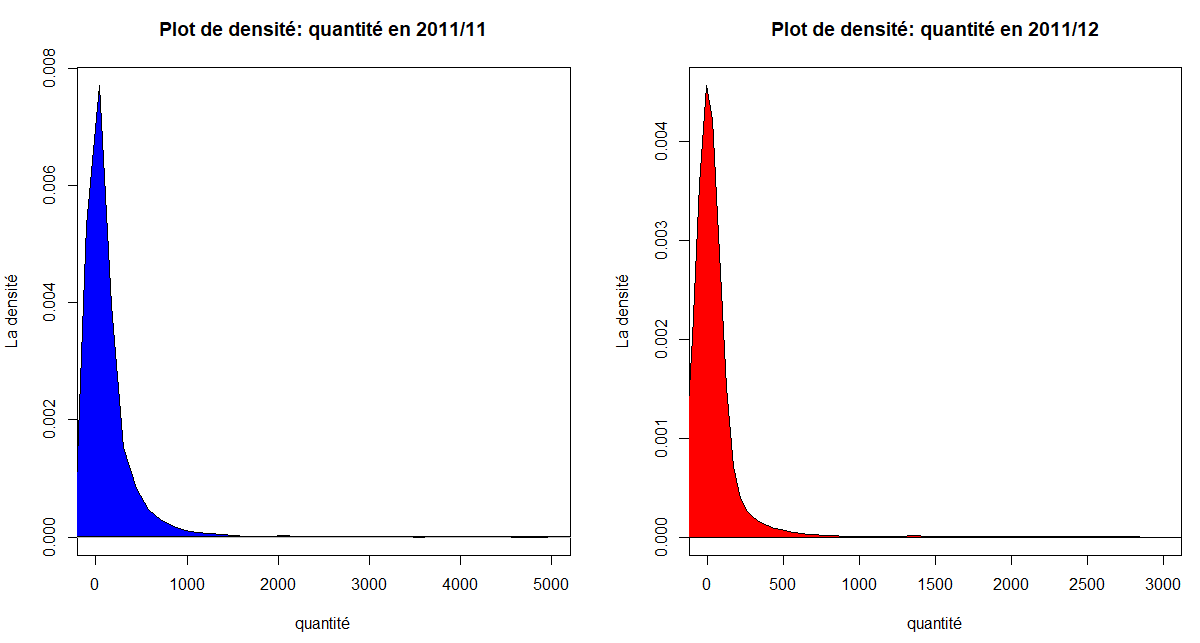
Le coefficient de corrélation linéaire :



0.8694977 est la valeur obtenue. Cela signifie qu’il existe une relation affine importante entre la quantité totale vendu en mois actuel augmente et la quantité totale vendu en mois prochain.

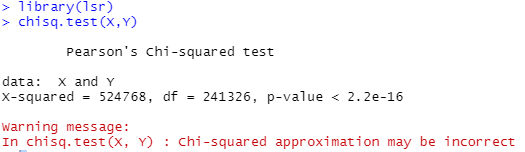
Calculons les densités :





-Le chi-deux χ2 :

> install.packages("lsr")



Le chi-deux χ2 est différent de zéro (x-squared = 2.2e-16) donc les deux variables « X » et « Y » sont dépendantes

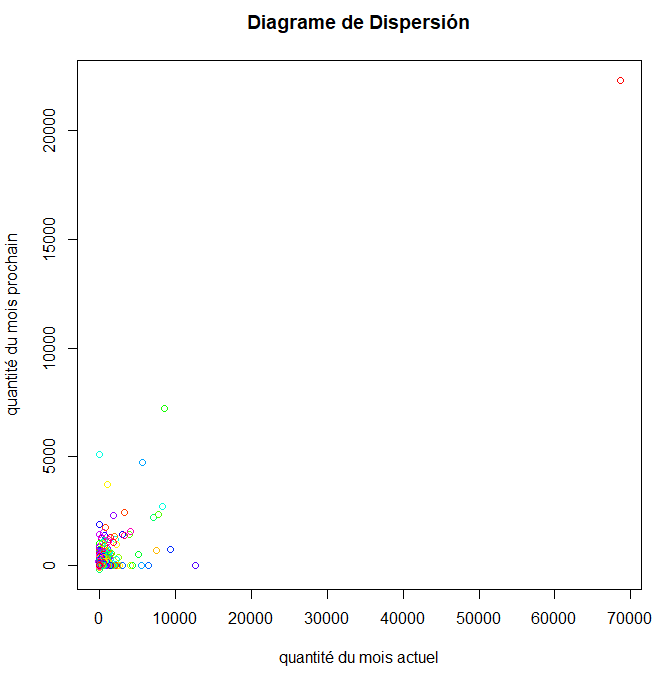
-Le coefficient V de Cramer :



Le coefficient V de Cramer égale à 0.5999528 donc il y a une forte corrélation entre les deux variables.

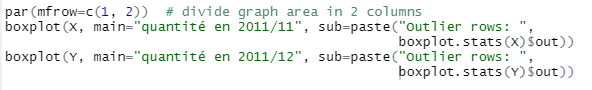
Le diagramme de dispersion :

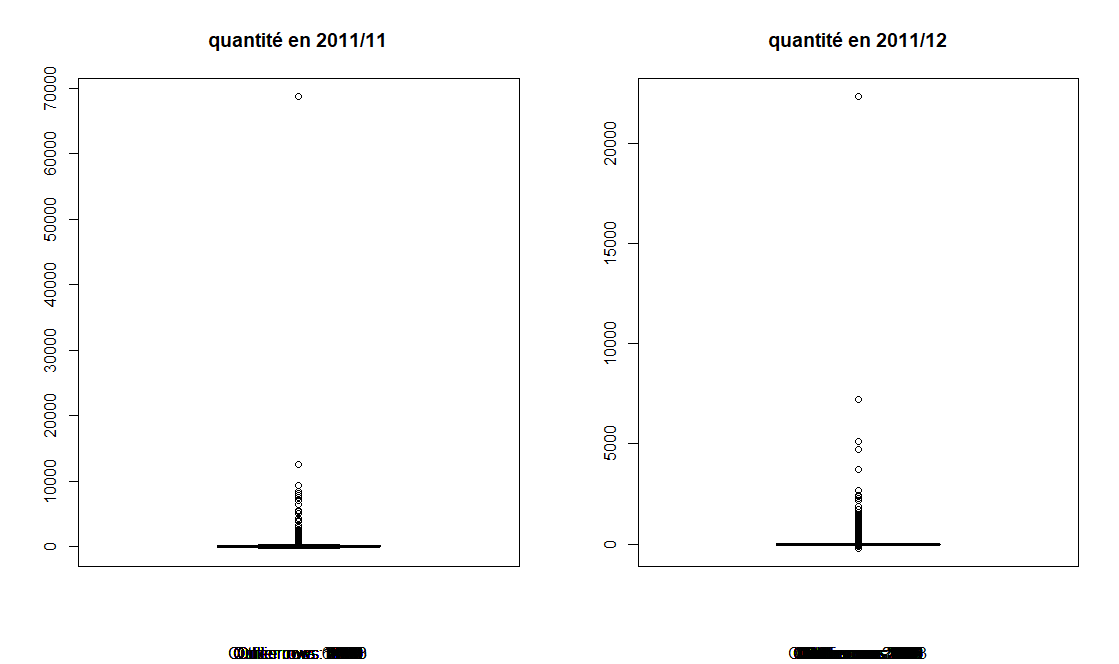
> plot(X,Y)



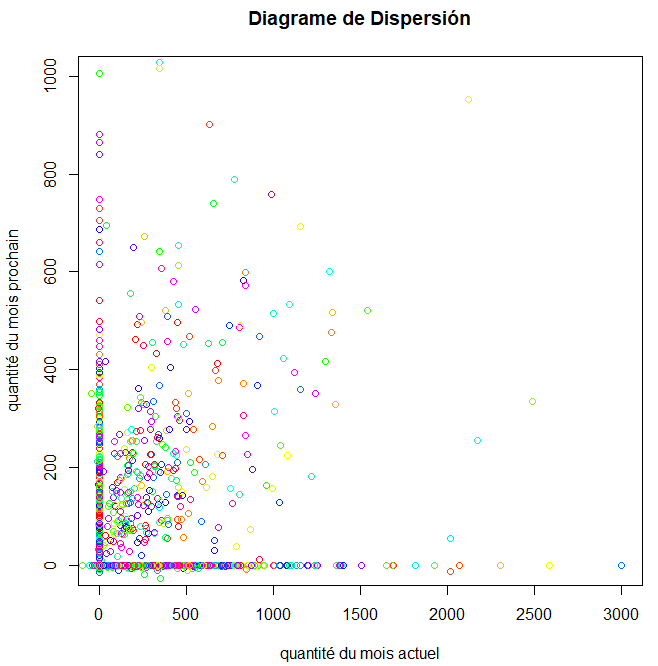
Notre graphe est mal construisé car Il y a des valeurs aberrantes dans nos deux variables.

Cette commande permet de construire un boxplot pour définir les valeurs aberrantes





Après l’élimination des valeurs aberrantes on obtient un graphe plus lisible et compréhensible :



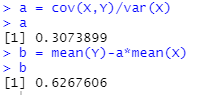
**Est-ce que le nuage de points suggère une forme de liaison particulière ?**

La forme du diagramme de dispersion indique une relation linéaire positive entre les distances obtenues avec X et Y. En effet, quand la quantité totale vendue en mois actuel augmente, la quantité totale vendue en mois prochain augmente également.

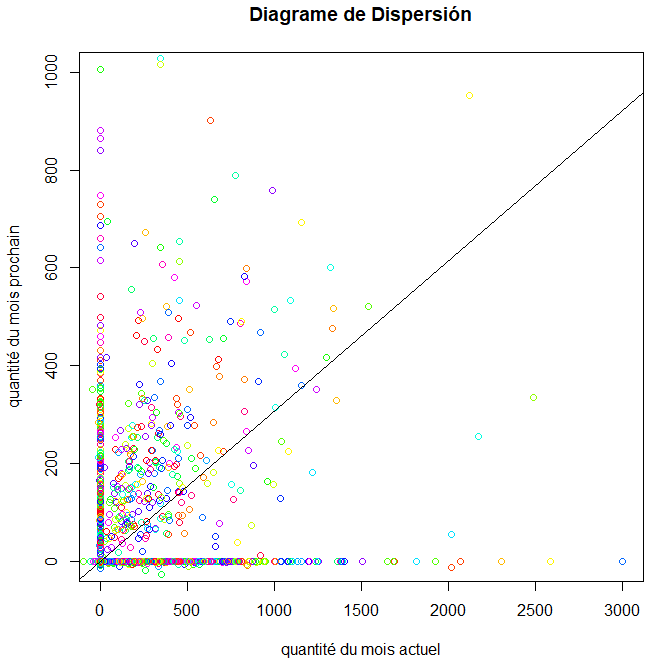
Le calcul des quantités suivantes :



Les coefficients a et b de la droite des moindres carrés :



Voilà la fonction de la droite des moindres carrés :



Y = 0,3073899 \* X + 0,6267606

Exemple de prévision :

On va donner à X qui représente la quantité totale vendu en mois actuel une valeur et on va obtenir une valeur de Y qui représente la quantité totale vendu en mois prochain

Si X=2096 articles, Y= 644 articles.

* Le coefficient de variation totale (CVT)

**> SVT<-3951\*var(Y)#n\*var(Y)**

**> SVT**

**[1] 712435195**

* Le coefficient de variation expliquée (CVE)

**> SVE <- 3951\*a\*a\*var(X)**

**> SVE**

**[1] 538619652**

* Le coefficient de variation résiduelle (CVR)

**> SVR<-SVT-SVE**

**> SVR**

**[1] 173815543**

* Le coefficient de détermination (R)

> R<-1-SCR/SCT

> R

[1] 0.7560262

Conclusion : R² est très proche de 1, cela signifie que l’équation de la droite de régression est capable de déterminer presque 75 % de la distribution des points. Cela signifie alors que le modèle mathématique utilisé, ainsi que les paramètres a et b calculés sont ceux qui déterminent la distribution des points.

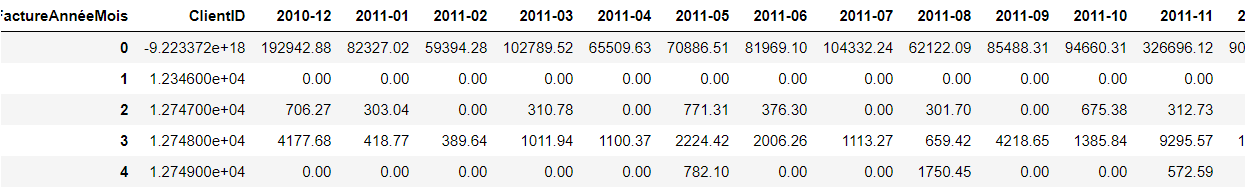
**2-Modèle multivarié de prédiction pour CLTV :**

Comme nous avons mentionné au-dessus, que jusqu’à maintenant on ne peut pas calculer le CLV pour des nouveaux clients. Pour cela, on va construire un modèle de régression linéaire pour préaviser cette valeur pour des nouveaux clients.

Extraire l'année et le mois de chaque achat :

Le tableau pivot prend les colonnes comme entrée et regroupe les entrées dans une table bi-dimensionnelle de manière à fournir une synthèse multidimensionnelle des données :

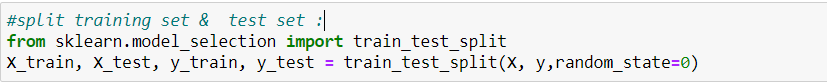
Ici l’indexation est faite par la colonne **ClientID** ,et comme colonnes on a utilisé **FactureAnnéeMois** ,la fonction d’agrégation est la somme des revenus.



L’ajout d’une colonne CLV considéré comme la somme des revenus de chaque client durant tous les mois :

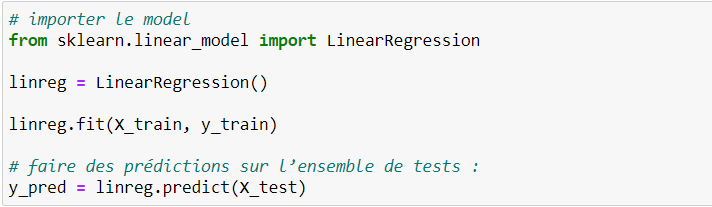
Nous avons utilisé la fonction **iloc** pour sommer à partir de la deuxième colonne.

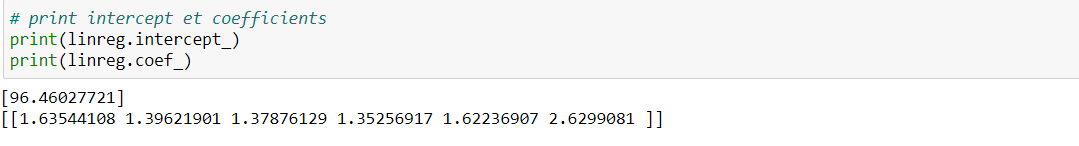


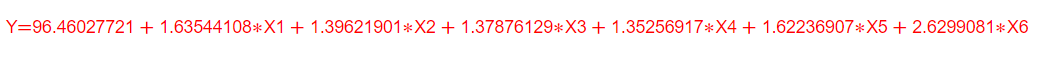
Fractionnons le jeu de données à l’aide de la fonction train\_test\_split() : 

**Développement de modèle :**

Tout d’abord, nous allons importer le module régression linéaire et créer un objet régression linéaire. Ensuite, nous adapterons notre modèle sur l’ensemble de train en utilisant la fonction fit () et effectuerons la prédiction sur l’ensemble d’essai en utilisant la fonction de prévision ().

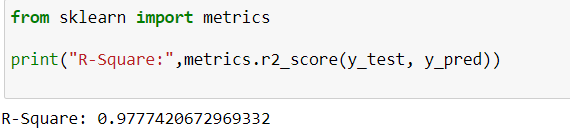




Enfin, nous avons réussi à construire notre équation de prédiction : 

**Dans quelle mesure le modèle convient-il aux données ?**

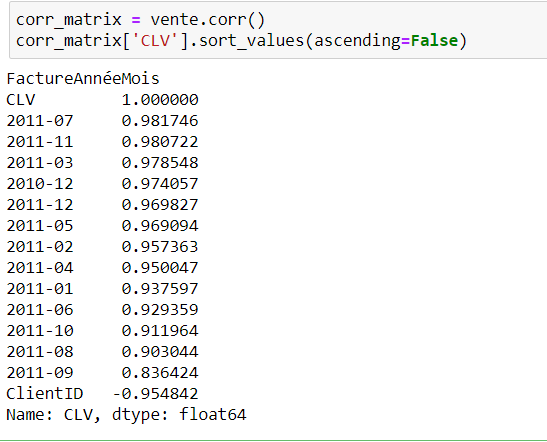
Afin d’évaluer l’ajustement global du modèle linéaire, nous utilisons la valeur R-squared :



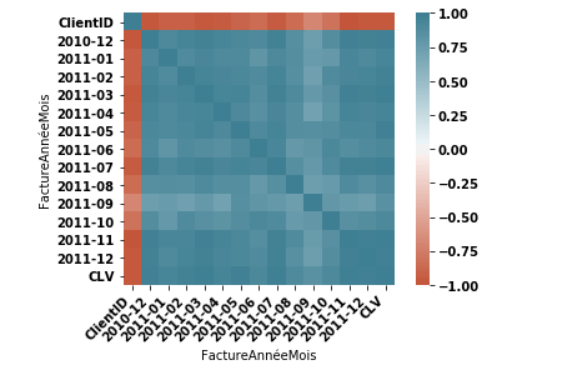
Super ! Notre modèle est fort en précisioon et permet de mieux adapter les données. Car R-Square est proche de 1.

* **La matrice de corrélation :**

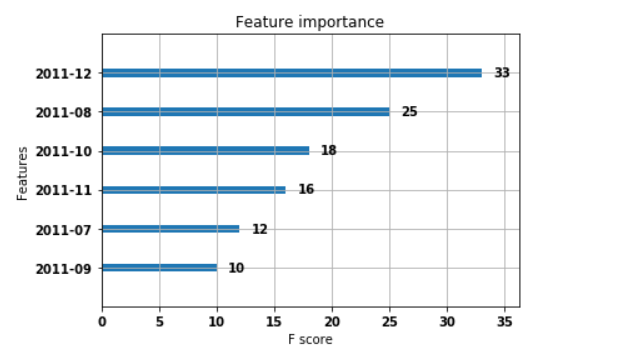
**La matrice de corrélation** est utilisée pour évaluer la dépendance entre plusieurs variables en même temps. Le résultat est une table contenant les coefficients de corrélation entre chaque variable et les autres.



La représentation graphique de la matrice de corrélation est la suivante :



Nous sommes passionnés de savoir quels sont les Xi qui influencent sur Y.

Nous dessinons la figure suivante : 

* On observe que Décembre est le mois le plus sensible qui a une grande importance dans notre modèle.

Exemple de prévision :

96.4602+706.27\*1.6354+ 303.04\*1.3962+0\*1.3787+ 310.78\*1.3525+ 0\*1.6223+ 771.31\*2.6299=4123.39$

**Conclusion**

Ce projet vise principalement à créer des algorithmes puissants qui transforment une base de données de clients bruts en données très compréhensibles qui contiennent plusieurs groupes et chaque groupe contient ses propres clients de manière à ce que les clients du groupe entier aient un intérêt et des comportements particuliers. Cela va contribuer à aider les entreprises surtout au Maroc et partout dans le monde à prendre des décisions basées sur des calculs scientifiques concrets, par deux étudiants qui sont très motivés et talentueux dans le domaine de sciences des données et aide à la décision.

Dans la première partie de ce projet, l'évaluation et le prétraitement des données ont été effectués. Cette partie était l'une des étapes les plus difficiles à faire pour continuer car il y avait 8 colonnes avec plus de 500000 lignes à analyser et toutes n'avaient pas de description. De nombreuses valeurs manquantes et des informations manquantes sur les attributs ont été identifiées. La colonne revenue que nous avons créée a joué un rôle majeur, elle a ensuite été utilisée dans les parties des modèles de clustering et de prédiction et aussi on l'exploiter pour faire des études statistiques très importantes de nombreuses fonctionnalités et paramètres :

-définir les revenues totales des clients nouveau et existants dans lequel on a trouvé que la majorité des transactions ont été effectués par les clients existants

-Nous avons noté que la majorité des commandes viennent de Royaume-Uni.

Dans la deuxième partie, nous avons utilisé l'outil de clustering le plus puissant 'KMeans' pour regrouper nos clients en 3 clusters majeurs (valeur faible, valeur moyenne et valeur forte) identifiés 1 cluster est la clientèle cible de l'entreprise (valeur forte : haute fréquence et montant, faible récence.). Ce sont des personnes qui nous achètent fréquemment de grandes quantités tout le temps.

Enfin, une régression linéaire a été sélectionnée et paramétrée pour construire 2 modèles supervisés afin de prédire le chiffre d'affaires du mois suivant et la valeur de la durée de vie du client sur le nouvel ensemble de données clients. La performance résultante des 2 algorithmes d'apprentissage supervisé est supérieure à 85,9%.

Ce projet nous a été bénéfique de plusieurs manières. Il nous a ouvert au monde professionnel qui implique la responsabilité et la discipline ainsi qu'un travail d'équipe collaboratif, c'était l'occasion d'améliorer nos connaissances en études de travail statistique et de maîtriser les langages de programmation comme Python, R, SQL etc. pour lesquels être utile sur le plan professionnel.

Enfin, nous avions d'autres algorithmes dans notre cahier de charge, parmi eux la prévision du taux de désabonnement, la prévision du prochain jour d'achat, la prévision des ventes, etc. Malheureusement, faute de temps, nous n'avons pas été en mesure de les construire tous et nous nous sommes concentrés sur les plus importants. Donc pour cela le travail ne se terminera pas ici. Et ce n'est que le début et c'est sûr que nous les développerons tous avec une seule idée : Aider les entreprises au Maroc.

Et voilà, il nous reste que vous remerciez pour l’aide et l'intérêt que vous avez apporté à ce projet, et pour le temps que vous nous avez consacré tout au long de cette période, sachant répondre à toutes nos interrogations. Nous avons particulièrement apprécié la diligence et la conscience professionnelle avec laquelle vous l'avez mené car vous avez également fait preuve, votre disponibilité a été une grande richesse très formatrice.

Vous êtes des professionnelles investis, compétents.

Votre énergie professionnelle est contagieuse et nous a grandement motivé pour mener à bien la fin de ce projet.

Les ressources du projet :

Le drive pour télécharger : Jupyter notebook (code python), dataframe, code R pour les statistiques bivariée et plein d'autres complémentaires du rapport :

<https://drive.google.com/drive/folders/1AETEihcM7q020ngfkF_eD5uwWe1s3ZtA?usp=sharing>

Notre site web pour visualiser le Dashboard en ligne :

<https://customersegmentation.000webhostapp.com/lsd/>

**Les cours et tutoriels que nous avons consommés pour réussir ce projet :**

K-Means clustering qu’on a l’utilisé dans le RFM segmentation :

<https://www.coursera.org/learn/data-science-k-means-clustering-python>

<https://www.youtube.com/playlist?list=PLSe8yMfKE4UmVizxXR8fiBtlDoiAFkr5i>

Cours de Power Bi :

<https://opdhsblobprod03.blob.core.windows.net/contents/6121d12283d64e6380411be32ab7c8b4/2bd9d158f29e3c04ff7e4c3b672c02f1?sv=2018-03-28&sr=b&si=ReadPolicy&sig=uJVbgM2JyJHcDkMFd9btQ%2BtJt88Nq44wU1Gn9cYjyXw%3D&st=2020-10-04T10%3A03%3A11Z&se=2020-10-05T10%3A13%3A11Z>

Cours Python (études statistiques, étude Bi variée régression/multivariée) :

<https://www.edx.org/course/analyzing-data-with-python>

Wordpress pour la création du site web :

<https://www.youtube.com/playlist?list=PL5BcU-_5Oa_ocJ_cyNKqhqJICMEJM9uqy>