Analyse du fichier de paiement

CameliaBalAG.csv – Camélia Tech – Avril 2016

# Introduction

Le fichier intitulé ‘CameliaBalAG.csv’ contient environ 6 Go de données de paiement, soit 62.523.908 lignes. Ce fichier csv décrit une facture par ligne et est constitué de plusieurs colonnes d’attributs pour ces factures. Ces attributs décrivent de façon presque complète la facture, comportant des informations telles que l’identifiant du payeur, la date d’émission ou encore le montant.

Le premier point important de cette analyse est le choix du domaine d’étude. En effet les factures présentes dans le fichier sont échelonnées entre 1989 et 2015. Les factures sont néanmoins majoritairement concentrées dans les années 2010-2015, à hauteur de plus de 98 %. Les factures extérieures à cet intervalle présentent en outre des comportements irrationnels, tels que des retards de paiement de près de dix ans, ou des montants excessivement élevés. Dès lors, dans cette analyse nous nous concentrerons uniquement sur les factures dont la date d’émission est postérieure à la date du 1er Janvier 2010.

Dans ce document, nous allons donc effectuer une première analyse de ce fichier, et commencer à mettre en relation son contenu avec le contenu d’autres fichiers de la même base de données, afin de vérifier la cohérence de cette dernière. L’analyse du document sera de ce fait découpée en cinq parties distinctes.

Dans un premier temps nous ferons la description des colonnes une par une, explicitant des distributions et autres informations intéressantes. La seconde partie explicitera les relations entre les colonnes, établissant une analyse plus poussée du fichier. Enfin, dans un dernier temps nous poserons les questions restant actuellement sans réponse et terminerons notre analyse.

# Analyse des colonnes

Le fichier qui nous intéresse est découpé en colonnes, dont voici la liste exhaustive, contenant le nom du champ, une brève description qu’il contient et enfin le type de données :

* Entrep\_id identifiant de l’entreprise (int)
* DatePiece date d’émission de la facture (str)
* DateEcheance date d’échéance de la facture (str)
* DateDernierPaiement date du paiement définitif de la facture (str)
* MontantPieceEur montant de la facture en euro (int)
* MontantLitige montant du litige (int)
* Devise nature de la devise (str)
* DateInsert date de l’entrée dans la base de données (str)

Nous allons désormais passer à l’analyse particulière de chacune de ces colonnes.

## L’identifiant de l’entreprise

[‘entrep\_id’]

Cette colonne contient des données de type int. Deux analyses sont possibles sur cette colonne : la distribution pure des identifiants et celle du nombre de ligne par identifiant, c’est-à-dire du nombre de factures par entreprise. Nous allons mener ces deux analyses.

### La distribution des identifiants

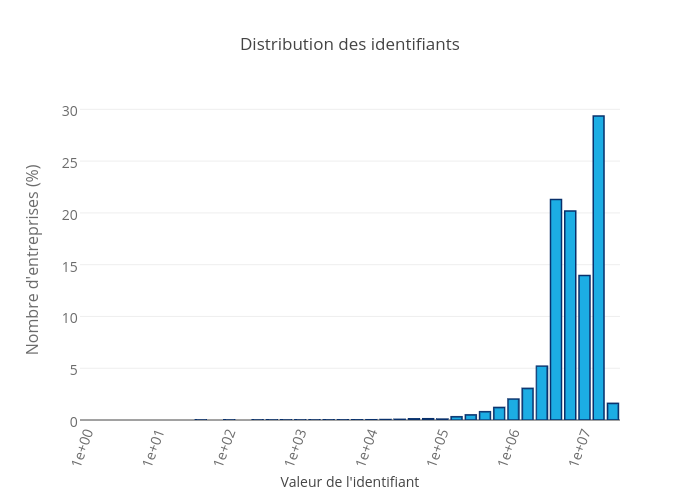
Les données présentes dans la colonne sont exclusivement des nombres compris entre 0 et 26.648.215. Un simple comparatif avec le fichier ‘CameliaEtab.csv’ nous permet de nous rendre compte que tous les identifiants présents dans le fichier sont également présents dans le reste de la base de données, à l’exception de l’identifiant 0. Nous avons donc pris soin de supprimer les lignes contenant cet identifiant, cela ne concernaient cependant qu’un nombre limité de factures, moins de 0.001 %.

Voici la distribution des identifiants présents dans la base de données ainsi que les informations relatives à cette distribution :

Nombre d’entreprise : 716.348

Identifiant minimal : 6

Identifiant maximal : 26.648.215



Nous remarquons ainsi que les identifiants des entreprises sont pour la plupart des grands nombres à quelques exceptions. Il est à noter que ce graphe possède des axes logarithmiques permettant une plus grande visibilité de la distribution.

### La distribution du nombre de factures par entreprise

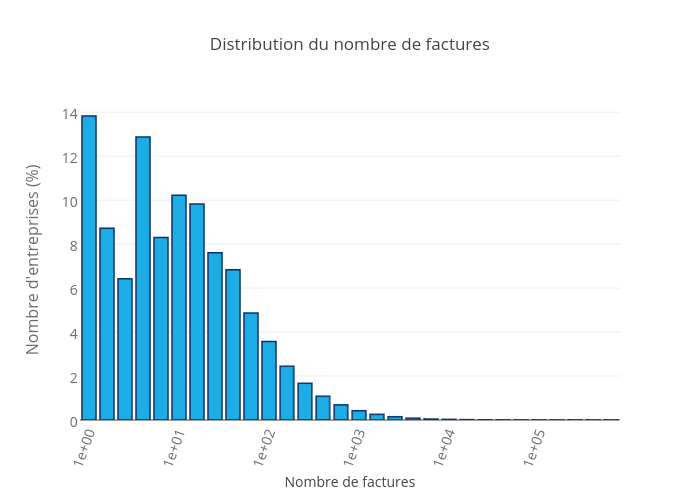
La colonne ‘entrep\_id’ nous permet également de faire une analyse du nombre de factures par entreprise. Ce paramètre est important car il permet de savoir si l’on possède assez de connaissance pour évaluer un modèle précis sur une entreprise. En effet il est impensable de produire un modèle se basant uniquement sur une dizaine de factures.

La distribution du nombre de factures ainsi que les informations relatives sont exposées ci-dessous :

Nombre moyen de factures : 87

Nombre minimal de factures : 1

Nombre maximal de factures : 814.375



De même que pour le précédent graphe, les axes sont ici logarithmiques pour une meilleure visualisation des extrémités. On remarque notamment que près de deux tiers des entreprises possèdent moins de dix factures, néanmoins plus de 80% des factures appartiennent aux grosses entreprises. Un clivage est éventuellement réalisable ici, séparant les entreprises ayant peu de factures aux entreprises en ayant plus. Dans notre modèle une élimination des entreprises ne possédant pas assez de factures est donc envisageable.

## Les différentes dates des factures

[‘datePiece’, ‘dateEcheance’, ‘dateDernierPaiement’]

Ces trois colonnes sont à priori des *strings*, des chaînes de caractères convertibles en date selon le format ‘AAAA-MM-JJ’. Nous allons procéder à une analyse pure du contenu des colonnes puis nous ferons des observations sur les relations entre les colonnes, notamment la proportion de factures payées à temps ou encore le retard moyen des factures.

### La distribution des factures à travers les ans

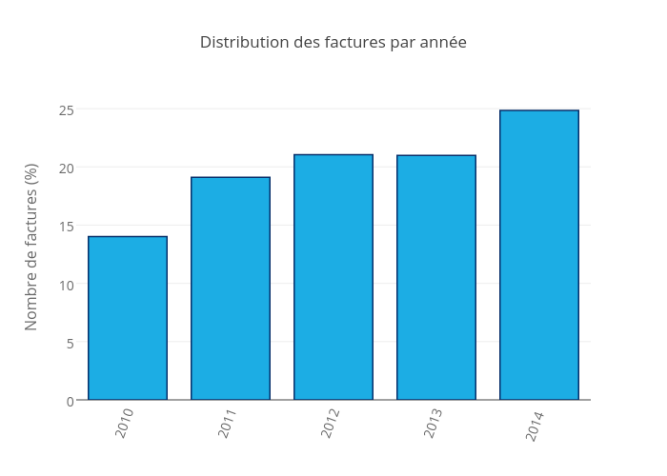
Il y a relativement peu de mauvaises entrées pour les deux premières colonnes, c’est-à-dire que seules 1.500 entrées possèdent des dates incorrectes, soit moins de 0.002%. En revanche 95% des factures sont considérées comme non payées, car elles possèdent la valeur ‘0000-00-00’ dans le champ ‘dateDernierPaiement’, ce qui est une valeur très importante.

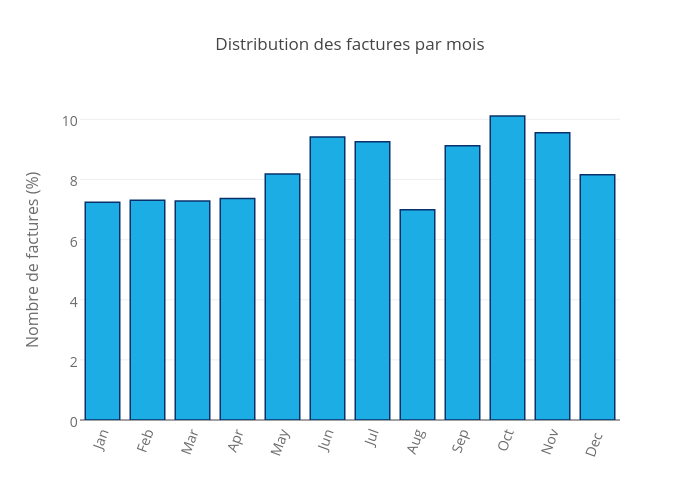
Nous remarquons aussi un léger taux de factures comportant des dates incohérentes. En effet, 635.116 factures, soit environ 1% du total, présentent des dates incohérentes, c’est-à-dire une date d’échéance ou de dernier paiement antérieure à la date d’émission. Pour des raisons pratiques nous nous contentons de supprimer ces entrées pour l’instant.

Rappelons enfin que nous avons pris la liberté d’épurer le fichier des 2% des factures dont la date d’émission était antérieure à 2010. Ainsi, si l’on analyse le champ ‘datePiece’ nous pouvons obtenir les données suivantes :

Date la plus ancienne : 2010-01-01

Date la plus récente : 2014-12-31

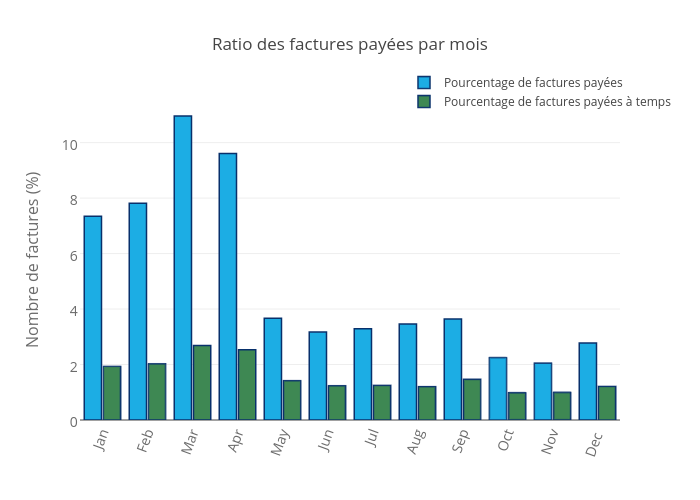
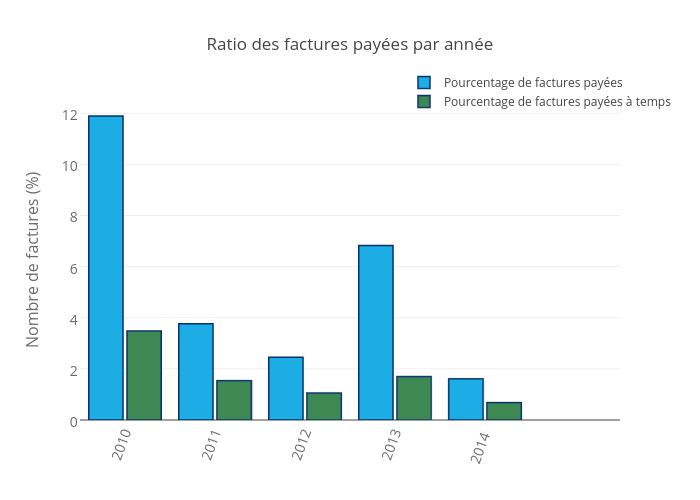
****



### Le ratio de factures payées

Comme considéré précédemment, seules 5% des factures sont considérées comme payées, puisqu’elles possèdent une date de dernier paiement qui est non nulle. Même s’il concerne plus de trois millions de factures, ce ratio est extrêmement faible et présente un fait qui mérite d’être souligné dans cette analyse.

Dès lors, il est tout de même possible de tracer des distributions de ratio de factures payées selon l’année ou le mois.

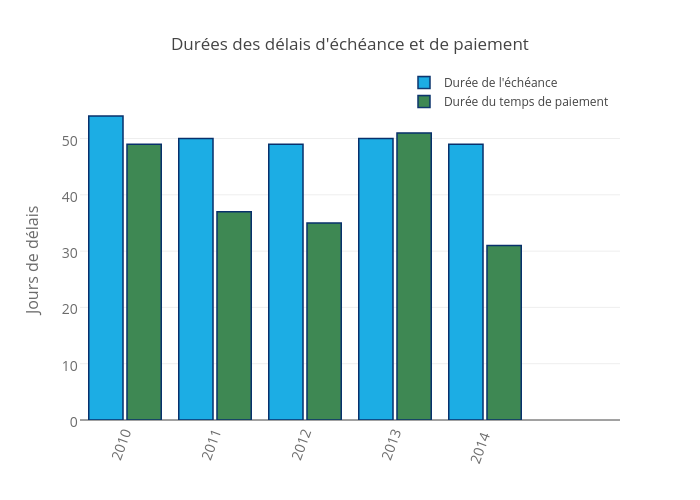


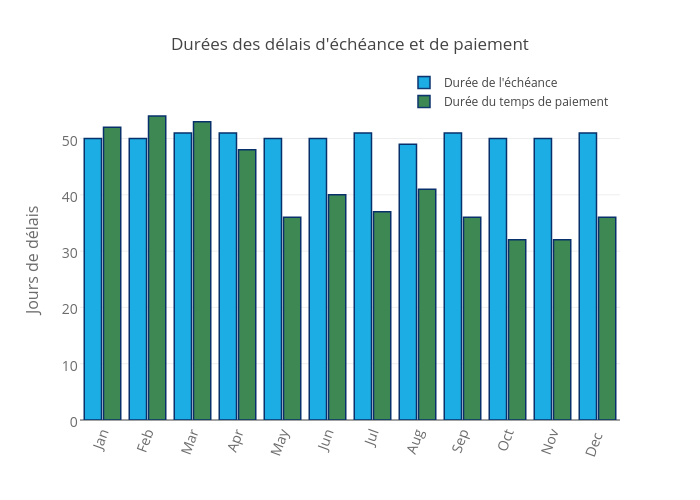
### Les délais : échéances et retards moyens

Enfin l’analyse conjointe des trois colonnes permet également d’étudier les délais. Par exemple le délai d’échéance moyen, ou le retard moyen avec lequel sont payées les factures. Ces statistiques sont intéressantes puisqu’elle constitue le cœur du futur programme de prédiction. Tout d’abord analysons les échéances.

La Loi LME publiée au journal officiel depuis 2009 impose une échéance de 45 jours fin de mois. Il y a deux façons de calculer cette échéance, mais globalement l’échéance est de maximum 60 jours. Sur notre domaine d’étude, c’est-à-dire les factures datant de 2010 au moins, l’échéance moyenne est de 50 jours ce qui est en adéquation avec la loi LME.

De même on peut analyser les délais entre l’émission de la facture et le paiement définitif s’il a lieu. On remarque dans un premier temps que le ratio de factures payées à temps avoisine les 50% sur l’ensemble des données payées, avec une valeur réelle de 42.5%.

****



## Le montant des factures

[‘montantPieceEur’]

La cinquième colonne du fichier contient un entier qui donne le montant de la facture en Euros dans la quasi-totalité des cas. Ce montant est intéressant car il permet de discriminer les types de factures et d’essayer d’établir des relations avec le retard de paiement.

L’analyse de cette colonne indique que les données sont plutôt propres, au sens où il s’agit de nombres entiers positifs dans plus de 99.999 % des cas. Effectivement seules 36 entrées sur les 60 millions ne correspondent pas à ces critères. Nous nous sommes donc permis de les retirer de la base de données.

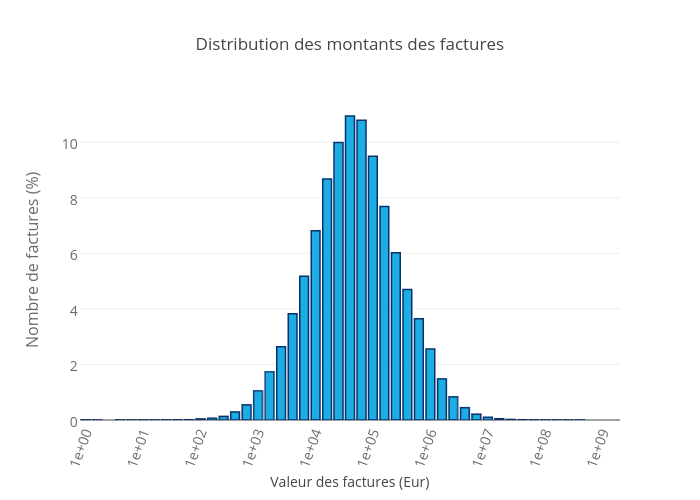
L’étude de la colonne donne les résultats suivants :

Montant moyen des factures : 293.515 €

Montant minimal : 1 €

Montant maximal : 2.140.992.800 €

Et la distribution des montants est donnée sur le graphe ci-dessous :



De même que les premiers graphes de cette analyse les axes sont logarithmiques pour une meilleure visualisation des données. Ce qui s’apparente à une gaussienne en première approximation est donc une gaussienne logarithmique.

Une question fondamentale est soulevée par ces données. En effet, le montant moyen est de 300k€ et le montant maximal de plus de 2 milliards d’euros. Ce qui nous amène à nous demander quelle est la nature de ces factures et quel est le domaine d’application du produit de prédiction que nous serons amenés à concevoir. Dans les fait, il également possible de cliver les données selon le montant des factures, néanmoins la plus grosse proportion des factures possède un montant de l’ordre de 100k€ à 10M€, ce qui est réellement impressionnant pour une facture.

## L’analyse des autres colonnes

[‘montantLitige’, ‘devise’,’dateInsert’]

Les trois dernières colonnes du fichier contiennent des informations supplémentaires qui ne sont malheureusement pas pertinentes car trop souvent incomplètes.

### La colonne ‘montantLitige’

La première de ces colonnes supplémentaires représente le montant du litige engendré par la facture. Même si le sens de cette valeur peut rester ambigu, il pose cependant un problème quant à son utilisation, puisque seulement 285.801 entrées en possède un non-nul. Cela ne représente que 0.4% du nombre total d’entrées dans le fichier, ce qui est vraiment peu.

### La colonne ‘devise’

La colonne devise représente le nom de la devise dans laquelle la facture a été émise. Il est à noter que cette colonne semble dépourvue d’intérêt puisqu’elle contient la chaîne de caractère ‘EUR’ dans 99.9% des cas. Les cas restants contiennent des dates qui font sûrement référence à la colonne suivante et donc à une erreur de formatage de la ligne. Comme la colonne suivante ne présente également à première vue que peu d’intérêt, nous ne supprimons pas ces lignes mal formatées sur la fin.

### La colonne ‘dateInsert’

A priori cette colonne présente la date d’insertion de l’entrée dans la base de données. Ces dates sont souvent identiques, ce qui montre que la base de données s’est construite de manière itérative. Le seul intérêt de cette colonne serait à priori de détecter les groupes d’entrées au comportement curieux et de se demander si elles ont été insérées en même temps. Néanmoins la complexité de calcul n’est pas justifiée par rapport au résultat obtenu, et nous avons décidé de laisser le contenu de cette colonne de côté dans une première approximation.

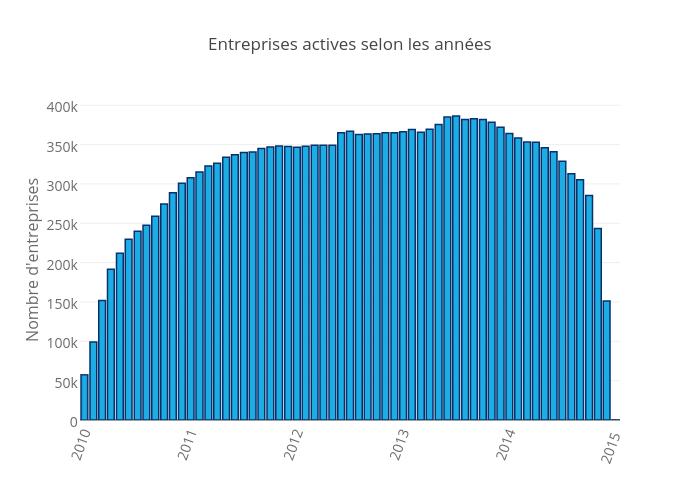
# Analyse globale du fichier

Après avoir passé en revue les colonnes, explicité leur contenu et tracé les distributions correspondantes, nous sommes désormais plus à même de décrire le fichier de façon plus globale. Dans cette ultime partie, nous établirons donc une analyse poussée du fichier dans sa globalité, mettant en relation les différentes colonnes.

## Les entreprises au fil des ans

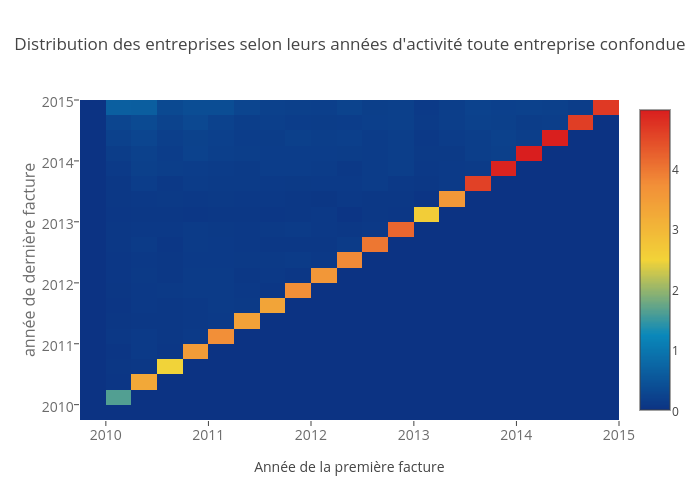
La première jointure que nous pouvons faire sur les différentes colonnes est la visualisation de l’évolution des entreprises au fil des ans. Effectivement, nous pouvons collecter l’ensemble des factures pour une entreprise donnée et en déduire ses années de fonctionnement. Nous ne considèrerons ici que les factures présentes dans le fichier, donc une entreprise sera pour nous fonctionnelle depuis sa première facture jusqu’à se dernière.

Le graphe ci-dessous affiche ces données.



Le nombre d’entreprises semble donc équilibré au fil des années, avec une légère infériorité pour l’année 2010 qui peut probablement être expliquée par le fait que c’est le début de la base de données. Il est important de considérer le fait que la plupart des entreprises ne possèdent qu’une seule facture dans le fichier et donc qu’à ce titre, elles n’apparaissent que dans une seule colonne de ce graphe.

Une autre visualisation est possible pour mieux comprendre ce phénomène. Sur le graphique ci-dessous, l’axe des abscisses présente la date de début d’activité de l’entreprise au sens de sa première facture, tandis que l’axe des ordonnées sa date de clôture, donc de dernière facture.

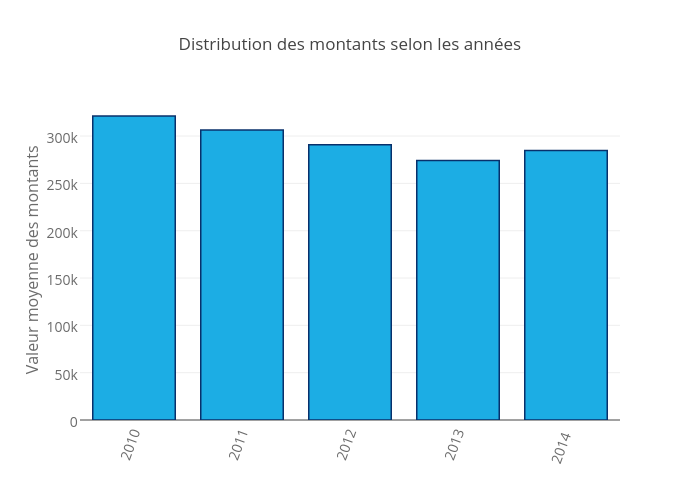


Le premier fait que l’on voit est l’importance de la diagonale qui prouve le phénomène expliqué plus haut des entreprises éphémères. Il est néanmoins également possible de voir une zone d’influence en haut à gauche du graphique, soulignant le fait que le reste des entreprises est actif sur l’ensemble du temps d’étude.

## Les montants au fil des ans

De même que pour les entreprises il est possible de faire une analyse des montants au fil des ans. Cela permet de vérifier que le montant est bien constant au fil des ans et au moins les factures semblent être réparties de façon similaire au cours des années, ce qui est bon à savoir puisque dès lors il nous est possible de ne pas cliver les comportements selon les années.

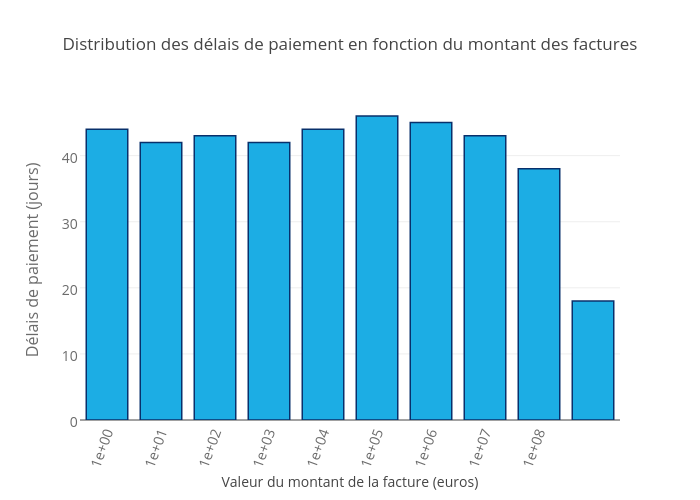
Le graphe ci-dessous nous donne la distribution des factures au cours des ans.

****

## Le retard selon le montant

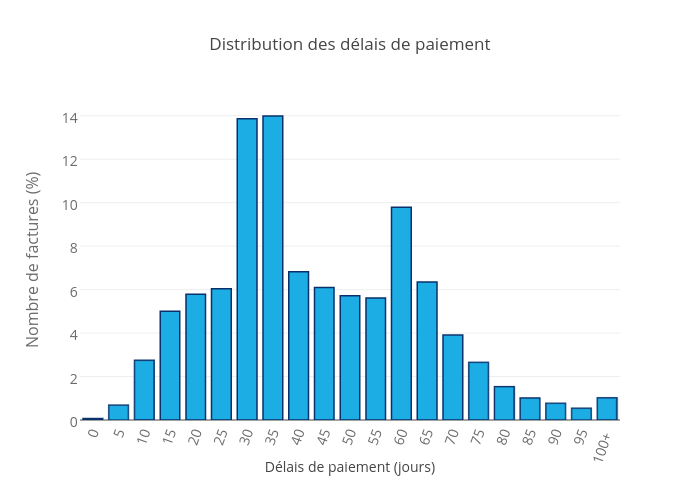
Cette analyse s’avère aussi à priori importante dans le sens où l’on pourrait éventuellement prévoir une corrélation entre le fait que le montant d’une facture est élevé et le fait qu’il faut plus de temps pour la payer. En revanche, à la visualisation du graphe ci-dessous, on se rend compte que les délais moyens selon les montants sont similaires et donc que le critère n’est pas si déterminant que cela.

Cette analyse nous permet de faire remonter une question précédente : quelle est le type de facture auquel nous avons à faire ? Dans tous les cas, l’idée de cette corrélation n’est pas mise de côté tout à fait pour l’instant, peut-être ressurgira-t-elle après certains clivages mentionnés plus haut.

****

## La distribution des retards

Enfin dans notre analyse il pourrait être intéressant d’analyser la distribution des retards.



Il est très certainement possible d’extraire des comportements particuliers de ce graphique, cela pourrait être une base adéquate pour commencer les études de prédiction.

# Conclusion

Nous avons mené une première analyse d’ensemble du fichier, répertoriant les colonnes et traçant les distributions. Nous avons donc à présent un bon panorama sur le contenu du fichier et des données qu’il contient. Il est certes encore possible d’extraire d’autres statistiques plus précises du fichier, néanmoins notre étude se limitait à la vision globale et n’a donc pas été poussée au maximum.

Nous pouvons de plus nous poser d’ores et déjà un certain nombre de questions :

* Pourquoi seules 5% des factures sont payées, ou du moins pourquoi 95% ne possèdent pas de date de dernier paiement ?
* Comment se fait-il que le montant moyen des factures s’élève à 300k€, et que le montant maximal dépasse aisément les 2 milliards d’euros ?
* Finalement quel type de factures est présent dans ce fichier, et quel comportement cherche-t-on précisément à identifier.

Ce document contenait une analyse d’introduction aux données, et selon les réactions et commentaires engendré par son contenu, un second document d’analyse et de données sera éventuellement publié par la suite, détaillant et mettant en lumière les points importants pour notre modèle de prédiction. Aussi chaque commentaire ou appréciation est la bienvenue pour nous aider à mieux comprendre et assimiler le large contenu de ce fichier.