2 - 5 mars 2020

PROJET DATA VISUALISATION

“Transparence - Santé”

Equipe TEC

Thibaut Fiolka

Emilie Viard

Clément Lamarque

Ecole IA Microsoft powered by Simplon

Sommaire

I) Introduction.

Un groupe de journalistes territoriaux décide de passer au crible les données publiées par le ministère de la santé dans le cadre de la transparence entre les entreprises et les acteurs du domaine de la santé.

La base de données publique « Transparence - Santé » précise, pour chaque type de lien d'intérêts, des informations pour:

- les conventions: l'identité des parties concernées, le montant et l'organisateur, le nom, la date et le lieu de la manifestation;
- des avantages directs et indirects: l'identité des parties concernées, le montant, la nature et la date de chaque avantage dès lors que le montant de chaque avantage est supérieur ou égal à 10 euros TTC;
- les rémunérations: l'identité des parties, la date du versement, le montant dès lors qu'il est supérieur ou égal à 1 da€ TTC. Le cas échéant, le bénéficiaire final de la rémunération ou de l'avantage est renseigné par l'entreprise.

Dans le cadre des recherches sur la possibilité d'effectuer des études plus approfondies, notre entreprise est sollicitée pour son expertise en statistiques et survol des données.

Notre étude se porte dans un premier temps sur des données concernant les actions des entreprises situées pour $\frac{3}{4}$ d'entre elles en france.

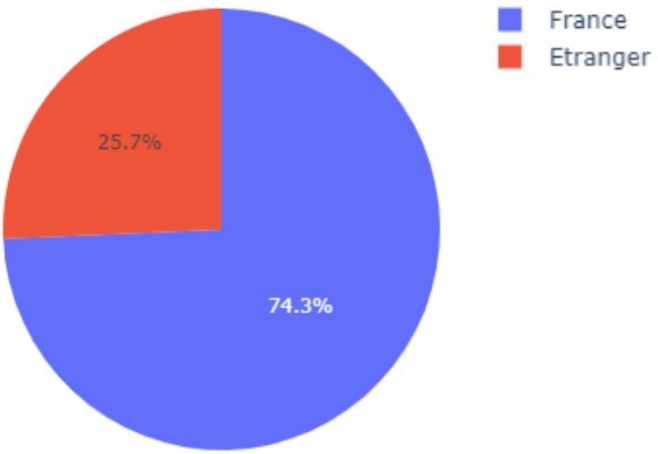

Figure 1 - Répartition des entreprises françaises et étrangères.

On y retrouve la dispersion des 2328 entreprises françaises dont 377 sont situées à Paris. Leur disposition en france ne montre pas une grande discordance avec les attendus, excepté pour le département du Rhône (69) contenant 159 entreprises.

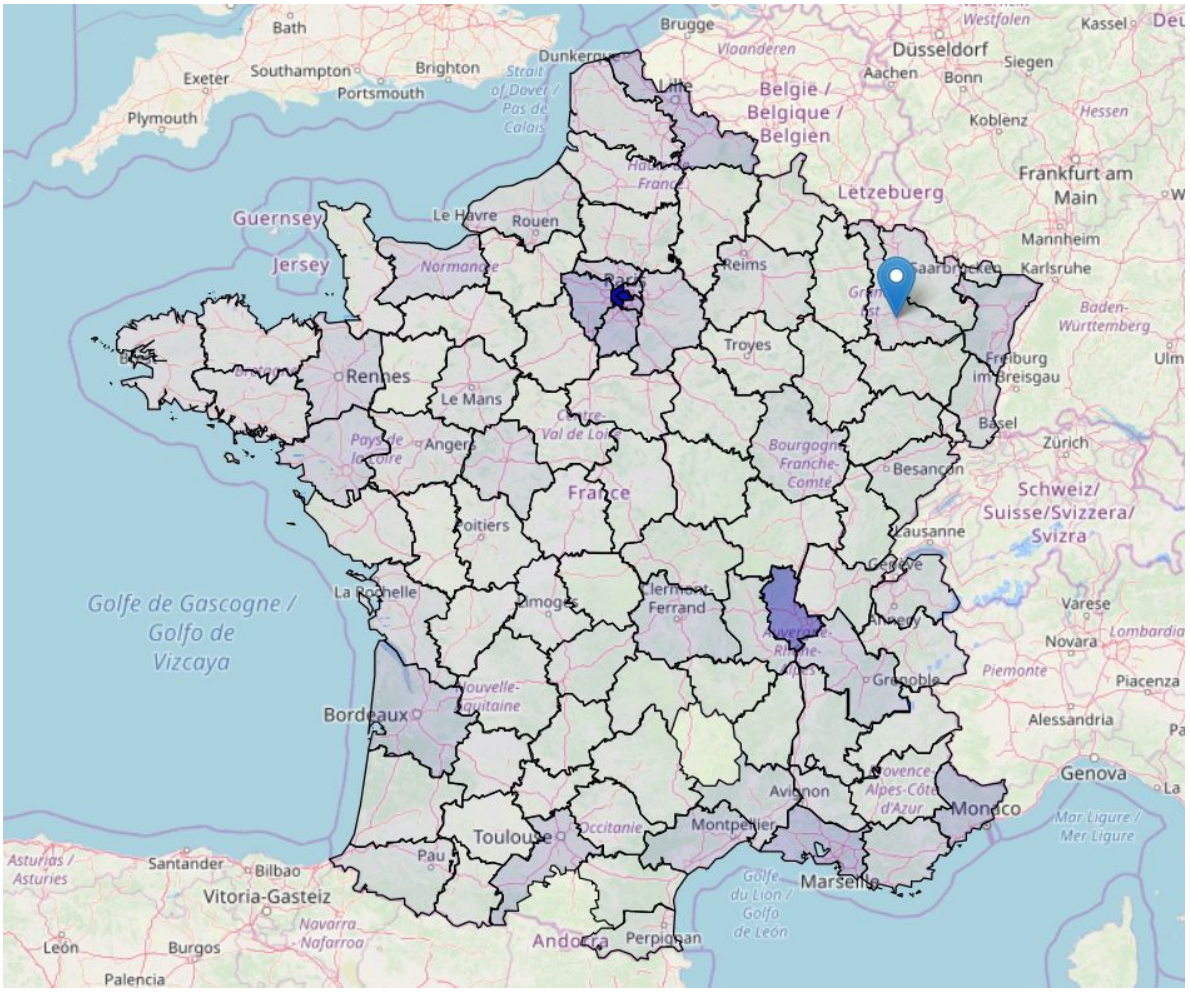

Figure 2 - Répartition des 2 328 entreprises françaises.

Les données mises à notre disposition sont basées majoritairement sur des entreprises françaises. Dans le cadre d'une première analyse pour les journalistes territoriaux, nous pourrons potentiellement omettre des données caractérisées par les entreprises étrangères suivant les études qui seront portées.

II) Justification des choix sur les langages.

Des intérêts se présentent dans les langages Pythons et R pour charger ou analyser le jeu de données:

- Le jeu de données a été chargé sous R en raison de la taille des BD. R est en effet plus à même de traiter les données volumineuses que Python, qui est réservé aux tâches répétitives et aux bases plus légères.

  Le langage de R met à disposition des commandes simplifiées dans une étude statistique

- Python n'est pas le plus profitable pour le chargement du jeu de données. Cependant, dispose d'un grand nombre de possibilitées pour les représenter.

Notre équipe a décidé de travailler sur les deux logiciels pour étudier plus précisément leurs intérêts et la correspondance des résultats obtenus.

L'étude sur le choix des intérêts d'un langage se passe dans un premier temps sur les méthodes possibles pour effectuer des nettoyages du jeu des données. En prenant tout particulièrement en compte:

- la recherche et suppression des lignes pluri-présentes.
- une étude des colonnes présentants un intérêt (rôles, valeurs manquantes)
- les modifications du format des valeurs identiques saisies différemment. (exemples: "101 avenue Anatole France" ayant différentes écritures, " 54000" transformé en "54000", ...)

Le langage R démontre un très grand intérêt pour des actions habituelles et couramment utilisées. Cependant, lorsque les actions demandent une précision et un tri extrêmement affiné, le langage de Python est aujourd'hui privilégié.

III) Analyse des données.

1) Statistique.

Une première étude statistique se base sur la répartition des différents montants TTC en nous basant sur les quartiles. Pour les trois documents associés aux déclarations, voici ce qui est obtenu avec R:

```
> summary(avantage$avant_montant_ttc)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
    10.0     23.0     40.0    143.9     60.0 3000000.0
> summary(convention$conv_montant_ttc)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
  -16707       25       57     1880      299 33655638  4181071
> summary(remuneration$remu_montant_ttc)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     10     120     600    5338    1800 4843939
```

Figure 1 - Répartition des quartiles sur les montants TTC.

Tout d'abord, pour les avantages moyens d'une déclaration, on retrouve les quartiles relativement proches les uns des autres et, bien en dessous de la moyenne. Cela laisse suggérer que globalement, les montants présents dans les avantages sont peu élevés et qu'il existe des cas où celui-ci est extrêmement élevé.
Ensuite, pour les conventions, on retrouve des cas similaires à ce qui était vu précédemment voir plus qu'accentués (Q3 = 299 et Moy = 1 880). L'information donnée concernant le minimum est surprenante en vue du fait qu'elle est négative. Une réaction a été faite en excluant les éléments ayant une valeur TTC négative bien qu'aucune déduction n'ai vu le jour. Cependant, la validité de cet élément peut être étudié plus en profondeur avec les autres facteurs pris en compte dans les données pouvant potentiellement combler et valider le fait qu'il soit négatif.
Enfin, pour les rémunérations, on tombe sur un cas similaire à précédemment sur la répartition des montants TTC perçus.

Notre deuxième étude a un grand intérêt dans le cadre des variables quantitatives. Cependant, pour notre cas, l'instruction faite nous apporte des informations très intéressantes sur les valeurs uniques de nos variables qualitatives. Cette action nous permettant d'identifier très rapidement une cohésion partielle des valeurs prises par nos variable.

```
> df_status(remuneration)
                 variable q_zeros p_zeros q_na p_na q_inf p_inf    type unique
1  entreprise_identifiant       0    0.00    0 0.00     0     0  factor    828
...
8            benef_prenom       4    0.00    0 0.00     0     0  factor   9358
9      benef_qualite_code       0    0.00    0 0.00     0     0  factor     24
10               qualite       0    0.00    0 0.00     0     0  factor     24
11         benef_adresse1       9    0.00    6 0.00     0     0  factor  71105
12         benef_adresse2     262    0.06    4 0.00     0     0  factor  23489
13         benef_adresse3     398    0.09    0 0.00     0     0  factor   9521
14         benef_adresse4     292    0.06    0 0.00     0     0  factor   1446
...
34   remu_convention_liee       0    0.00    0 0.00     0     0  factor 350244
```

Figure 2 - Quantité des valeurs uniques prises par les variables considérées.

En raison du fait que la quasi-intégralité des données sont des variables qualitatives et non quantitatives mais aussi en vue des expériences passées. Une analyse approfondie des données n'a pas été effectuée. Cependant, une légère étude aurait pu voir le jour en étudiant une corrélation entre la quantité des versements et le montant total perçu par les entreprises.

2) Analyse associée aux entreprises.

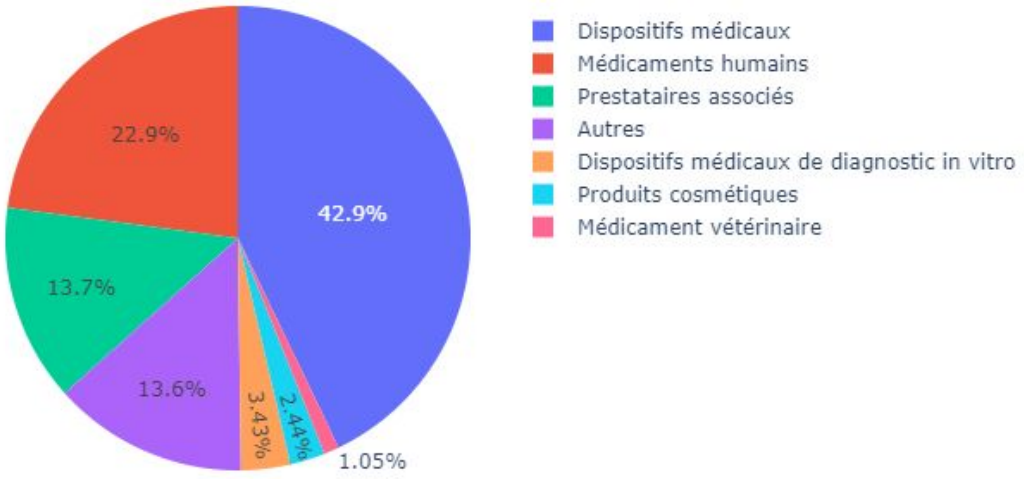

Figure 1 - Nombre d'entreprises par secteur

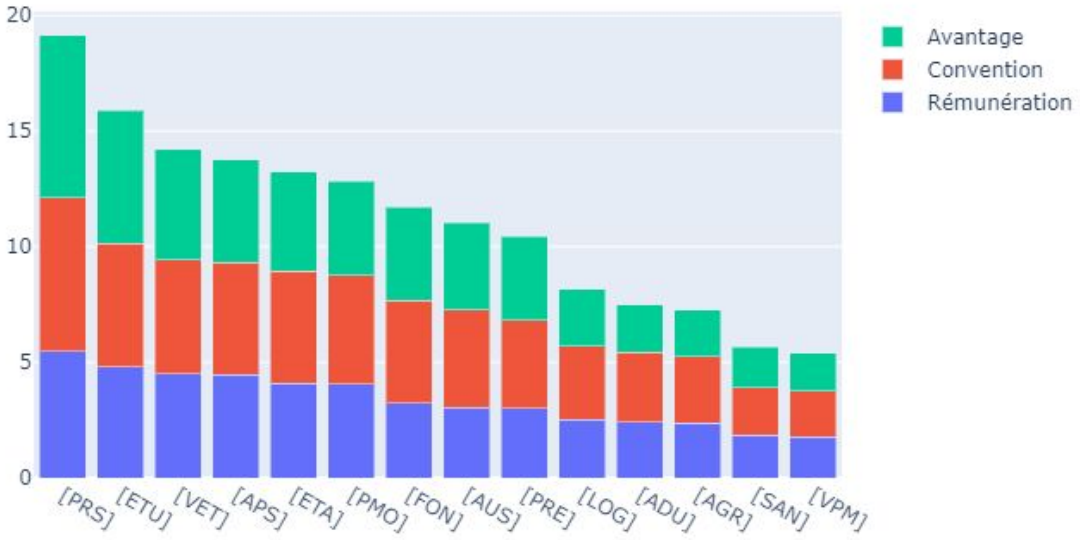

Figure 2 - Nombre de rémunérations par catégorie de bénéficiaires

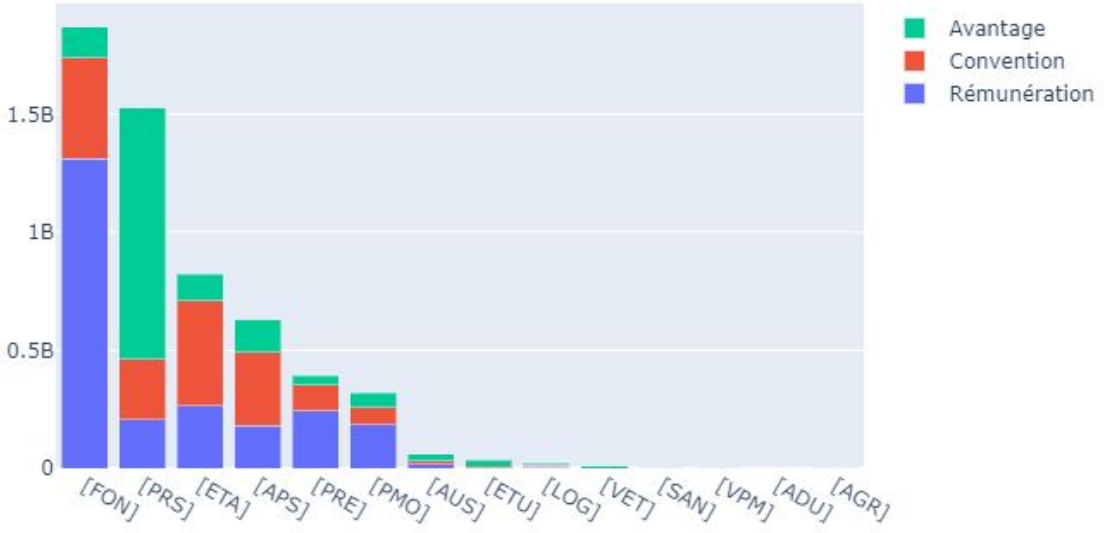

Figure 3 - Montant perçu par catégorie de bénéficiaire

On peut voir sur la figure 1 que la majorité des entreprises qui investissent sont du secteur des dispositifs médicaux ou dans le domaine du médicament. On peut aussi remarquer, grâce à la figure 2, que la grande majorité des bénéficiaires de ces investissements sont des professionnels de santé. Cependant, la figure 3 nous montre que ce sont les fondations qui profitent de la plus grande rémunération.

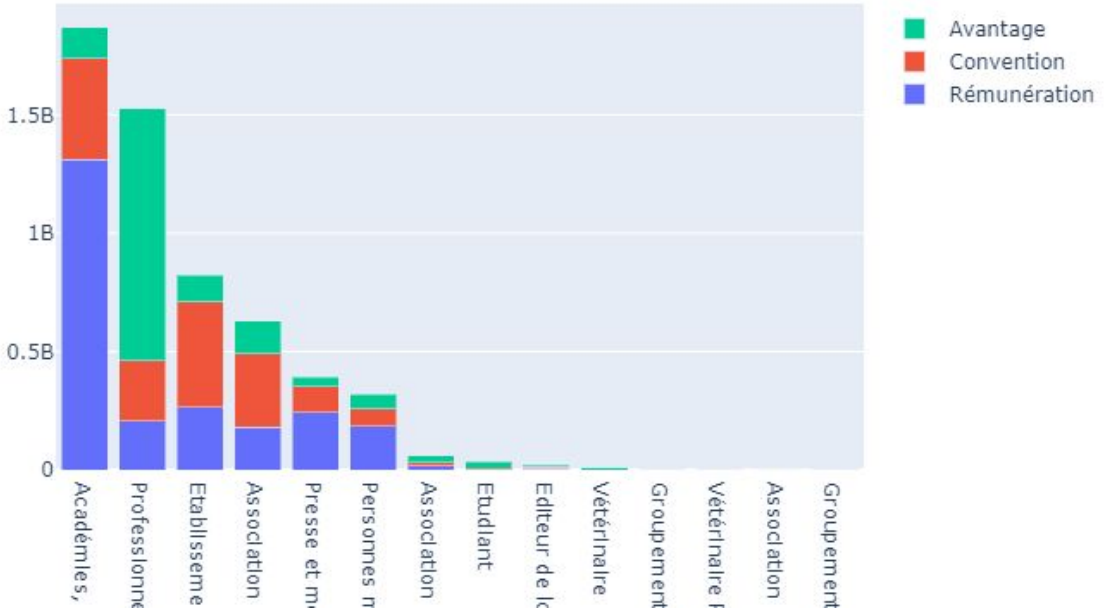

Figure 4 - Montant investi par catégorie d'entreprise

Sur la figure 4, on peut observer que les académies, fondations, sociétés savantes et organismes de conseil profitent principalement de rémunérations, tandis que les professionnels de santé profitent davantage d'avantages.

IV) Etude temporelle des rémunérations

1) Tendances annuelles

Les montants TTC apparaissant dans les différentes tables étaient les principales données quantitatives pouvant être exploitées. Nous avons fait le choix de réaliser une analyse statistique des montants versés par les entreprises par années afin de dégager certaines tendances entre 2010 et 2019.

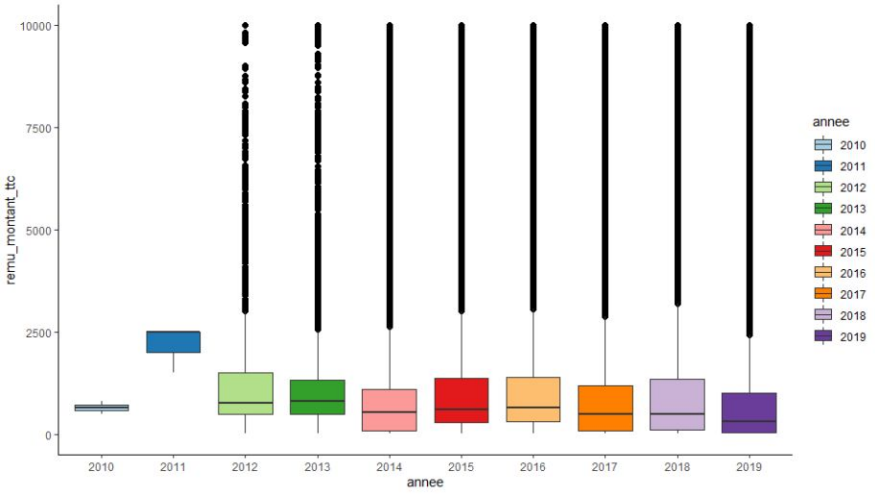

Figure 1 - montants des rémunérations versées par année.

Les boîtes à moustaches (Figure 1 ci-dessus) mettent en évidence des valeurs aberrantes à compter de 2012, tandis que la médiane a baissé progressivement au fil des années. Les courbes de densité montrent des tendances différentes pour chaque année, avec toutefois des modes qui coïncident (Figure 2 ci-après).

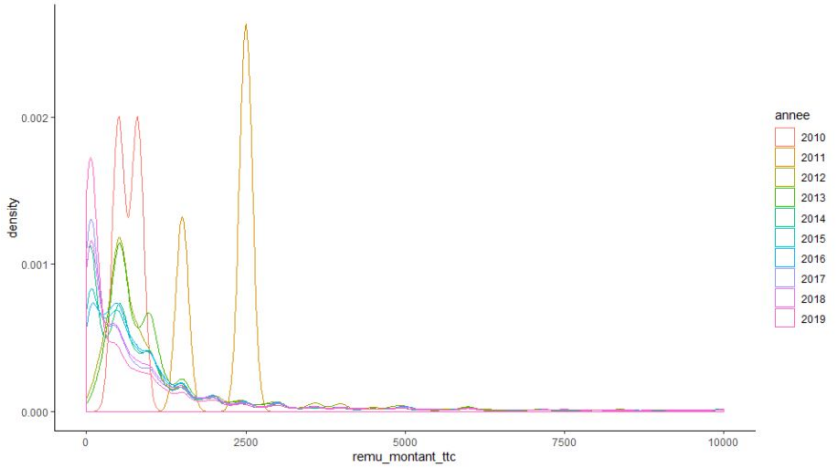

Figure 2 - densité des montants des rémunérations par année.

2) Tendances mensuelles

Nous avons creusé les variations mensuelles des rémunérations versées par les entreprises, en analysant les montants totaux versés par mois et par an entre 2010 et 2019 selon les catégories de bénéficiaires (cf Appendix 1).

Les “académies, fondations, sociétés savantes et organismes de conseils” reçoivent la plus grande proportion des rémunérations globales, représentant près de 50% du total à compter de 2014. En revanche, les parts des professionnels de santé et des établissements de santé ont diminué progressivement au fil des années.

Nous avons ensuite tenté de dégager une tendance générale quant aux rémunérations versée sur une année “type”. D’une part, nous avons calculé les moyennes pour chaque mois de toutes les rémunérations individuelles, et d’autre part nous avons établi la moyenne des totaux versés chaque mois dans la décennie. Le dernier calcul est ainsi une moyenne des diagrammes présentés en Appendix 1.

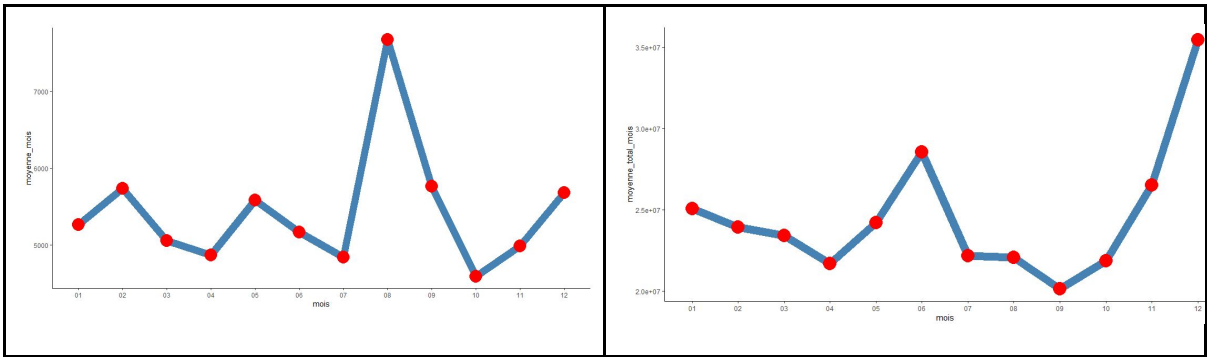

Figure 3 - tendance mensuelle des montants individuels (à gauche) et moyenne des rémunérations totales versés par mois (à droite) entre 2010 et 2019.

Les deux graphiques, bien que représentant le calcul de moyennes pour les mêmes données, montrent des tendances différentes.

Une opposition dans les tendances croisées pour les mois de juin et d’août a notamment attiré notre attention. D’une part Le montant total moyen des rémunérations est élevé en juin tandis que la rémunération moyenne est faible. L’inverse est constaté pour le mois d’août.

Cette distorsion peut par exemple s’expliquer par la variabilité des rémunérations à petits montants. En effet, la raréfaction des petites rémunérations face à une proportion inchangée de grands montants provoquerait l’envolée de la moyenne d’août, tandis qu’une vague de petites rémunérations entraînerait l’effet inverse en juin. Ce point pourrait faire l’objet d’une étude plus poussée pour mettre les résultats en parallèle avec la densité des rémunérations par mois et confirmer cette hypothèse le cas échéant.

3) Etude de cas : GlaxoSmithKline (“GSK”)

Nous avons analysé les rémunérations versées par les entreprises réparties par catégorie et ventilées temporellement de 2010 à 2019 (cf. Appendix 1). Le mois de juin 2014 a attiré notre attention par la proportion anormale de rémunération globale à destination de la Presse et Média (figure 1 ci-après):

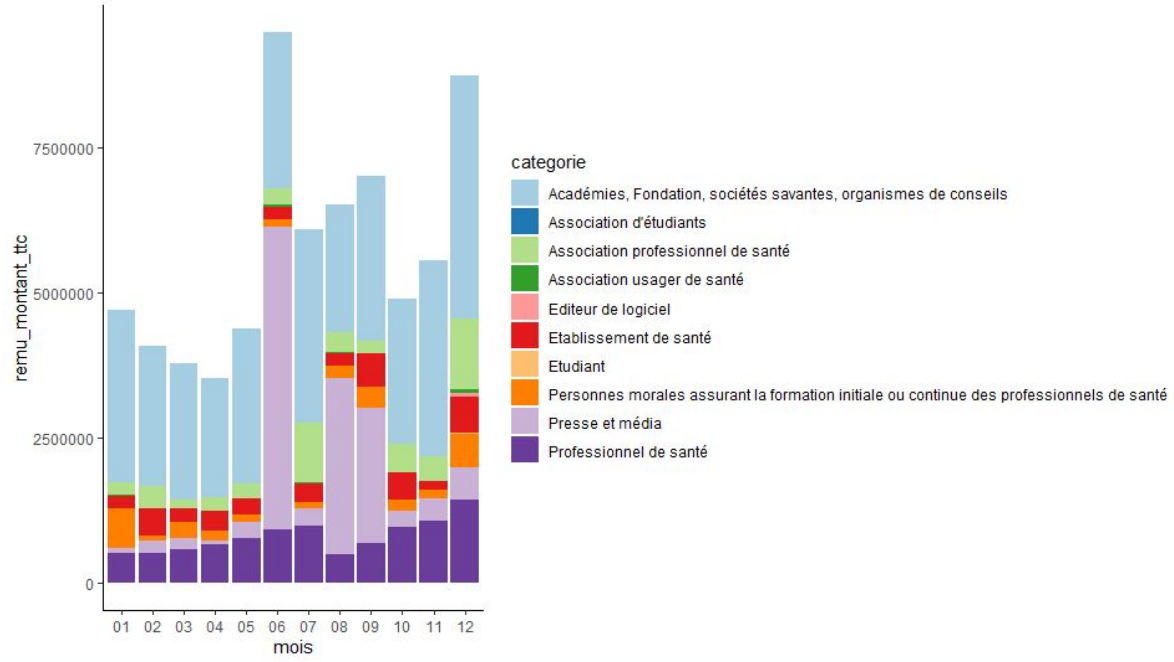

Figure 1 - Rémunération totale par catégorie et par mois (2014).

Une recherche plus approfondie a permis d’établir que les six entreprises qui ont versé des rémunérations à la catégorie Presse et les médias en juin 2014 figurent dans le top 10 des sommes les plus importantes versées toutes catégories confondues (figure 2 ci-après):

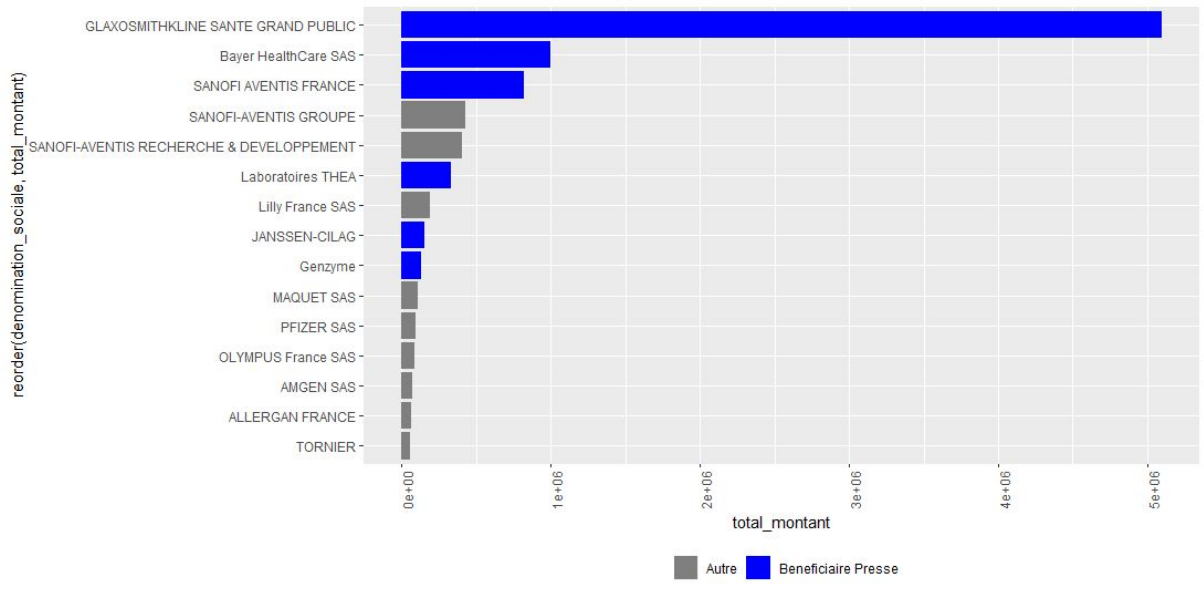

Figure 2- Rémunération 2014 par mois et par catégorie.

On constate un décalage significatif entre le premier du classement et les autres entreprises: le total des sommes versées en juin 2014 par GlaxoSmithKline Santé Grand Public est disproportionné vis-à-vis des autres budgets.

L'hypothèse selon laquelle les montants alloués par GSK seraient responsables de la distorsion graphique du mois de juin 2014 a été formulée. Afin de confirmer cette hypothèse, les rémunérations versées par GSK ont été retirées des données avant de réaliser un nouveau diagramme (ci-dessous à droite):

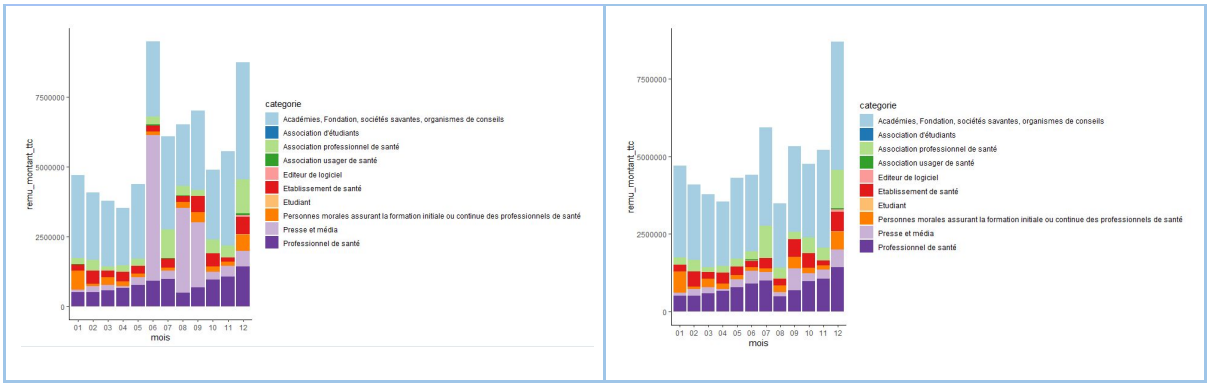

Figure 3 - Comparaison des rémunérations globales versées en 2014 par catégorie, montants versés par GSK inclus (à gauche) et exclus (à droite).

Dans le nouveau diagramme, les proportions des montants alloués à la catégorie "Presse et médias" sont en adéquation avec les tendances générales. Cette comparaison démontre donc à suffisance que les montants versés par GSK en juin 2014 étaient des données aberrantes (au même titre que les montants versés en août et en septembre 2014).

Au vu de ce qui précède, il semblerait que GSK ait été en liens étroits avec la presse et les médias pendant cette période, en déployant un budget conséquent. Ceci est à mettre en perspective avec des événements marquants concernant l'entreprise.

En poursuivant les recherches journalistiques, nous avons en effet appris que GSK avait reçu une mise en garde de la FDA (Food and Drug Administration des États-Unis) via une lettre datée du 12 juin 2014. Cette lettre faisait suite à une inspection que la FDA avait effectuée au début avril 2014 dans une des leurs usines de fabrication de vaccins antigrippaux sise à Sainte-Foy (Québec)[1].

En conclusion, il est à noter qu'une analyse statistique temporelle permet de révéler bien plus qu'une tendance générale.

[1] https://ca.gsk.com/fr-ca/salle-de-presse/communiqués-de-presse/2014/énoncé-de-gsk-lettre-de-mise-en-garde-de-la-fda-usine-de-fabrication-de-sainte-foy/

V) Gestion de projet.

1) Diagramme de Gantt

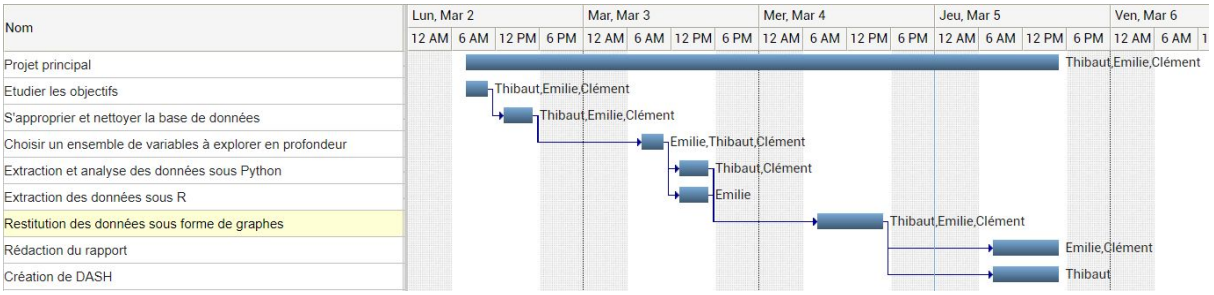

Figure 1 - Diagramme de Gantt

Nous nous sommes répartis les tâches comme ci-dessus. Nous avons mis du temps à déterminer sur quelles valeurs nous allions nous baser afin d'effectuer notre étude.

2)Difficultés rencontrées

Au début du projet, nous avons manqué d'organisation. Nous avons pris beaucoup de temps à comprendre à quoi correspondaient les données, et à établir un plan de recherche pertinent sans être trop ambitieux compte-tenu de la contrainte de temps à respecter.

Finalement, nous avons su dégager deux analyses. La première porte sur les entreprises et les rémunérations des différents intervenants dans les conventions. La deuxième est temporelle et s'intéresse à la répartition des fonds médicaux sur une année.

VI) Conclusion et perspectives.

L'équipe TEC a fait le choix d'utiliser R et python pour mener l'étude des données : les membres ont ainsi exploré les capacités offertes par les deux systèmes et sont par conséquent montés en compétences. En outre, l'équipe a déployé des méthodes pour traiter la base de données, incluant notamment la suppression des doublons, la recherche des données manquantes et le tri des données pertinentes.

La contribution technique du projet a par ailleurs consisté en l'analyse de variables, la réalisation de graphiques (diagrammes en bâtons, pie-chart, boîtes à moustache, courbes de densité, graphiques linéaires), l'interprétation des supports de visualisation précités ainsi que la revue des données aberrantes. Enfin, la connaissance des plateformes et des environnements de travail n'était pas en reste, puisque l'équipe a restitué ses travaux sur github et sur dashboard.

La mobilisation des compétences n'aurait toutefois pu être fructueuse sans un travail d'équipe efficient. Dans un premier temps, il nous a fallu développer une bonne coordination pour structurer nos actions. Tel était l'un des défis initiaux de la première journée en essayant en parallèle de cerner les données à exploiter. Par la suite, les échanges concernant les différentes étapes à effectuer ont été étudiées quotidiennement.

L'équipe a relevé des défis tout au long de ce projet et a identifié des potentiels de perfectionnement. Dans le cadre de nos capacités concernant des analyses faites sur les données, une amélioration peut être recherchée dans la méthode de préparation des données brutes (*data wrangling*), notamment concernant leur structuration (*tidy, reshaping*), leur enrichissement et leur validation.

L'étude menée se base en outre sur une unique partie des données. Une deuxième possibilité d'amélioration concerne le fait d'effectuer les recherches en prenant les différents tableaux pour essayer de tisser des liens entre leurs parties. Pour une analyse plus aboutie des données, nous devons continuer l'émergence des outils statistiques et de visualisation.

Enfin, une dernière piste de progression serait l'acquisition de clés d'interprétation et des grilles de lecture supplémentaires. Cela nous permettrait d'une part de diriger des études statistiques avec pertinence, et d'autre part d'anticiper l'exploitation de celles-ci par le corps journalistique.

APPENDIX 1:
Diagrammes des rémunérations par catégorie, mois et année