

Marchés publics et base FOPPA

Groupe 6

BURGEAT Jérôme
LUO Yingqi
ZEPHIR Marie

24 septembre 2024

Master 2 informatique
ILSEN/IA

UE Business intelligence & Systèmes décisionnels
ECUE Application Business Intelligence

Responsable
Vincent Labatut

Sommaire

Titre	1
Sommaire	2
1 Présentation	3
1.1 Contexte	3
1.2 Organisation	3
1.3 Implémentation	4
2 Données	5
2.1 Caractéristiques	5
2.2 Visualisation des données	8
2.2.1 Lots	8
2.3 Nettoyage	24
2.3.1 Les applications de nettoyage des données	24
2.4 Analyse descriptive	28
2.4.1 Lots	28
2.4.2 Criteria	41
2.4.3 Agents	42
3 Questionnements	45
3.1 Les flux de communication mesurés en termes de nombre d'échanges et en termes d'argent	45
3.1.1 Quelles entreprises communiquent le plus avec quelles autres entreprises ?	46
3.1.2 Quelles entreprises ont le plus de flux monétaire sortants/entrants ? Y a-t-il des PME parmi elles ?	47
3.1.3 D'où provient l'argent qui arrive/sort en plus grande quantité ? (en termes de départements français)	49
3.1.4 Quelles sont les activités les plus actives ou les plus lucratives en termes de code CPV ?	50
4 Extension	53
4.1 Méthode d'enrichissement	53
4.2 Analyse	55
5 Conclusion	56
6 Lexique	57

1 Présentation

1.1 Contexte

Le projet vise à nettoyer et analyser les données de marchés publics fournies afin de révéler les informations et les modèles sous-jacents, et d'essayer de répondre aux questions pertinentes dans le domaine en question à travers les données. Les données volumineuses utilisées dans ce projet proviennent des données publiques de marchés publics en France (base de données FOPPA¹), comprenant des avis d'appel d'offres et d'attribution entre 2010 et 2020. L'objectif principal est de réaliser une analyse descriptive des données de marchés publics pour répondre à diverses questions sur les marchés publics. Cette analyse permet une flexibilité dans le choix des orientations d'analyse, mais l'objectif global est de comprendre les caractéristiques, les tendances et les anomalies potentielles dans les marchés publics français.

Lien Github : <https://github.com/Luo-Ying/Application-BI-Feuille-de-donn-es-/tree/main>

1.2 Organisation

Le groupe est composé de trois personnes, notamment :

- Jérôme BURGEAT
- Yingqi LUO
- Marie ZEPHIR

Tâches	Membres
Écriture du code	Tous les membres
Génération des graphiques	Tous les membres
Nettoyage de la base de données	ZEPHIR
APIs	BURGEAT & LUO
Rédaction du rapport	Tous les membres
Rédaction du requirement.txt	ZEPHIR & LUO
Rédaction du README.md	LUO

Table 1. Répartition des tâches

- La bibliothèque `sqlite3` permet aux développeurs Python de créer, d'accéder et de gérer des bases de données SQLite directement depuis Python. Elle a été utilisée pour lier la base de données FOPPA [[potin:hal-03796734](#)]. Cette bibliothèque n'a pas été utilisée lors de nos séances de TP.
- La bibliothèque `beautifulsoup4` nous permet de parcourir, rechercher et modifier l'arbre de syntaxe abstraite d'un document HTML ou XML. Elle est conçue pour des tâches de web scraping, c'est-à-dire pour extraire des données à partir de pages web. Cette bibliothèque n'a pas non plus été utilisée en séances de TP.
- La bibliothèque `lxml` est utile pour le parsing et la manipulation de documents XML et HTML. Elle a été utilisée pour les démarches d'API pour l'extension également. Cette bibliothèque n'a pas été vue en cours.

1. <https://ted.europa.eu/en/>

- La bibliothèque `matplotlib` est largement utilisée pour la création de visualisations statiques, animées et interactives en Python. Elle offre une interface de haut niveau pour dessiner des graphiques attractifs et informatifs, avec un grand souci du détail pour la qualité de finition. Cette bibliothèque a été utilisée en TP de fouille de données avec Stéphane Huet.
- La bibliothèque `numpy` est connue pour ses capacités de manipulation de tableaux multidimensionnels, appelés `ndarray`. Elle fournit un ensemble complet d'outils pour travailler avec ces tableaux de manière efficace et est fondamentale pour le calcul scientifique en Python. Nous l'avons également utilisée en TP de fouille de données.
- La bibliothèque `pandas` facilite la manipulation et l'analyse de données. Elle offre des structures de données et des fonctions d'analyse de données robustes, flexibles et faciles à utiliser, rendant le traitement de données tabulaires, de séries temporelles ou de matrices multidimensionnelles simple et intuitif. Elle a également été vue en TP de fouille de données.
- La bibliothèque `requests` est utilisée pour envoyer tous types de requêtes HTTP facilement. Elle nous a été utile pour les requêtes API. Elle n'a pas été vue en TP.
- La bibliothèque `seaborn` permet la visualisation de données Python basée sur Matplotlib. Elle fournit une interface de haut niveau pour dessiner des graphiques statistiques attrayants et informatifs. Elle est utile pour avoir des graphiques plus intéressants pour l'utilisateur. Nous l'avons vue en TP de fouille de données.
- La bibliothèque `tabulate` permet de formater des tableaux de données en texte lisible et bien formaté, prêts à être imprimés dans la console ou à être utilisés dans des fichiers texte. Elle est particulièrement utile pour afficher les résultats de requêtes de bases de données, des données structurées ou tout autre ensemble de données tabulaires en Python de manière claire et esthétique. Elle est utilisée pour l'extraction et la visualisation des données de la base de données. Elle n'a pas été vue en TP.
- La bibliothèque `geopandas` est conçue pour la manipulation et l'analyse de données spatiales ou géographiques. Elle a été utile lors du questionnement pour la visualisation des résultats. Elle n'a pas été vue en TP.
- La bibliothèque `xlrd` permet de lire des données et de formater des informations à partir de fichiers Excel (avec des extensions `.xls` ou `.xlsx`, c'est-à-dire les versions d'Excel 97-2003 et les versions plus récentes jusqu'à Excel 2010). Elle s'est avérée utile lors de l'extension avec d'autres bases de données grâce à l'API. Elle n'a pas été vue en TP.

1.3 Implémentation

Nous avons trois parties de script globalement :

- Les scripts pour dessiner les graphiques en exploitant les données.
- Les scripts pour nettoyer les données.
- Les scripts pour répondre aux questionnements.

Les scripts pour dessiner les graphiques en exploitant les données. Cette partie de programmes sera lancée avant le nettoyage et après le nettoyage, afin de générer les graphiques à partir des données brutes et des données nettoyées. Elle contient les scripts suivantes :

- *script_single.py* - ce script a pour but de générer les graphiques avec les données individuelles récupérées depuis la base de données.
- *script_pair.py* - ce script a pour objectif de générer des graphiques à partir des données récupérées en paires depuis la base de données.

Les scripts pour nettoyer les données. Cette partie du programme sera lancée une fois les graphiques générés à partir des données brutes, à la fois en attribut individuel et en paire. Elle contient les scripts suivants :

- *script_clean_variables_api_tedeuropa.py* - Ce script a pour but de nettoyer les données en appliquant nos propres algorithmes.
- *script_clean_variables_manually.py* - Ce script utilise l'API du site EU tenders pour vérifier si les données sont bien mises à jour. Dans le cas contraire, nous les corigeons.

Rem. Les explications des scripts et des algorithmes pour la partie nettoyage seront bien expliquées dans la section dédiée au nettoyage [2.3](#)

Les scripts pour répondre aux questionnements. Cette partie du programme sera lancée à la fin de tous les scripts pour générer les graphiques et les fichiers de sortie afin de répondre aux questions posées. Ils produiront des fichiers contenant les réponses en utilisant les algorithmes nécessaires à partir des données nettoyées. Il contient les scripts suivants :

- *script_varsAssociation* - Ce script fusionne les tableaux nécessaires pour calculer les listes d'ordre des agents ou des départements et utilise l'API SIREN pour montrer les informations de manière plus détaillée des agents. Son but est de démontrer et d'analyser les résultats obtenus des questions posées.
- *script_cpv_par_domaine* - Ce script fusionne également les tableaux nécessaires pour calculer les listes d'ordre des activités (identifiées par le code CPV) afin de répondre aux questions posées.

2 Données

2.1 Caractéristiques

Nous avons 5 fichiers différents qui contiennent 5 tableaux :

- **Lots** - représentant les lots faisant l'objet des marchés publics.
- **Criteria** - représentant les critères d'attribution associés aux lots de la table Lots.
- **Agents** - représentant les agents économiques jouant le rôle d'acheteur ou de fournisseur dans un marché public.
- **Names** - contenant les noms secondaires attribués aux agents économiques de la table Agents, en plus de leur nom principal.
- **LotBuyers** et **LotSuppliers** - contenant le *lotId* et *agentId*, qui permettent de relier les tables **Agents** et **Lots** et ainsi d'implémenter les relations correspondantes.

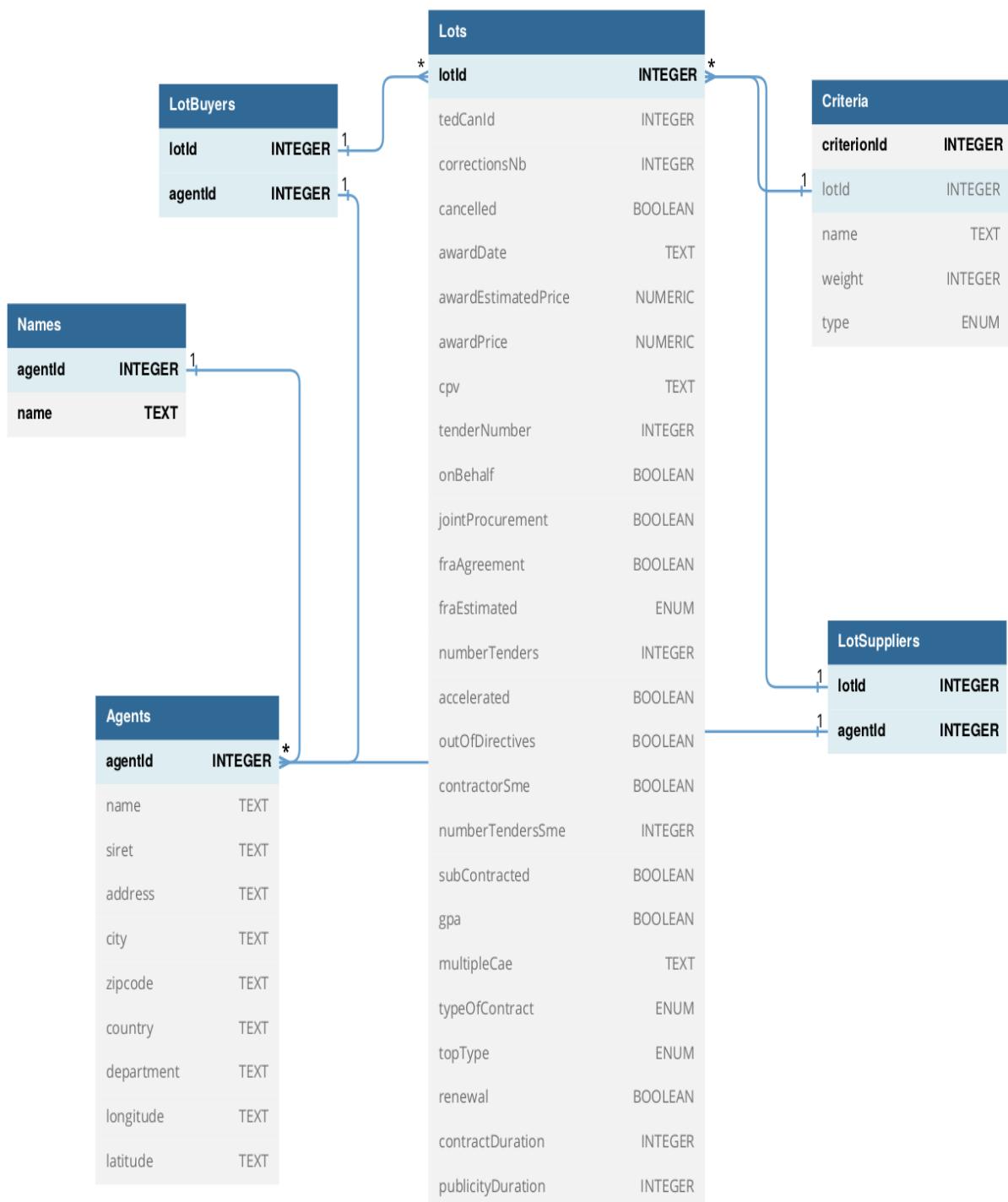


Figure 1. Relation des tableaux

Attribut	Signification
<i>lotId</i>	Clé primaire
<i>tedCanId</i>	ID TED de l'avis d'attribution du lot
<i>correctionsNb</i>	Nombre de correctifs publiés pour ce lot
<i>cancelled</i>	Booléen indiquant si l'appel d'offres du lot a été annulé
<i>awardDate</i>	Date à laquelle la décision d'attribution a été prise pour ce lot
<i>awardEstimatedPrice</i>	Estimation du prix du lot effectuée dans l'appel d'offres
<i>awardPrice</i>	Prix effectif du lot indiqué dans l'avis d'attribution
<i>cpv</i>	Code Common Procurement Vocabulary du lot
<i>numberTenders</i>	Nombre d'offres effectuées pour le lot
<i>onBehalf</i>	Booléen indiquant que l'acheteur est un groupement
<i>jointProcurement</i>	Booléen indiquant s'il s'agit d'un marché conjoint
<i>fraAgreement</i>	Booléen indiquant si le lot fait partie d'un accord-cadre
<i>fraEstimated</i>	Champ suggérant l'existence d'un accord-cadre (le cas échéant)
<i>lotsNumber</i>	Numéro du lot dans la notice d'attribution
<i>accelerated</i>	Booléen indiquant l'utilisation de la procédure rapide
<i>outOfDirectives</i>	Booléen indiquant un avis d'attribution sans appel d'offres associé
<i>contractorSme</i>	Booléen indiquant si le gagnant est une PME
<i>numberTendersSme</i>	Nombre d'offres issues de PME pour ce lot
<i>subContracted</i>	Booléen indiquant si le lot est sous-traité
<i>gpa</i>	Booléen indiquant un lien avec l'Accord sur les Marchés Publics
<i>multipleCae</i>	Booléen indiquant si l'avis d'attribution liste plusieurs acheteurs
<i>typeOfContract</i>	Contrat de fournitures (S), travaux (W), ou services (U)
<i>topType</i>	Type de procédure d'attribution
<i>renewal</i>	Possibilité de renouveler le contrat
<i>contractDuration</i>	Durée du contrat
<i>publicityDuration</i>	Durée de la période de publicité de l'appel d'offres
<i>totalLots</i>	Nombre total de lots dans la notice d'attribution après nettoyage

Table 2. Lots

Attribut	Signification
<i>agentId</i>	Clé primaire
<i>name</i>	Nom principal de l'agent (cf. Table 4)
<i>siret</i>	Numéro d'identification unique dans la base SIRENE
<i>address</i>	Voie et numéro dans l'adresse postale
<i>city</i>	Ville dans l'adresse postale
<i>zipcode</i>	Code postal dans l'adresse
<i>country</i>	État membre
<i>department</i>	Département français
<i>longitude</i>	Position spatiale (X)
<i>latitude</i>	Position spatiale (Y)

Table 3. Agents

Attribut	Signification
<i>agentId</i>	Clé étrangère désignant l'agent concerné
<i>name</i>	Nom (secondaire) associé à cet agent

Table 4. Names

Attribut	Signification
<i>criterionId</i>	Clé primaire
<i>lotId</i>	Clé étrangère désignant le lot concerné
<i>name</i>	Représentation textuelle du critère
<i>weight</i>	Importance du critère dans le processus d'attribution
<i>type</i>	Catégorie du critère

Table 5. Criteria

Attribut	Signification
<i>lotId</i>	Clé étrangère indiquant le lot concerné
<i>agentId</i>	Clé étrangère désignant l'agent concerné

Table 6. LotsBuyers et LotsSuppliers

Remarque : Les informations supplémentaires retrouverons dans la section 6.

2.2 Visualisation des données

2.2.1 Lots

Section variable unitaire

- **awardDate**

Les données de awardDate sont situés entre 2010 et 2020. En se basant sur les valeurs fournies, plus de 80% des attributions de lots ont été effectués entre 2008 et 2020. De plus, le calcule des pourcentages des dates d'attribution de contrats nuls ou non sont les suivants : sur 1 380 965 lignes, 86.47% sont instanciés (1 194 294) et 13.52% dates vides (186 671).

- **awardEstimatedPrice**

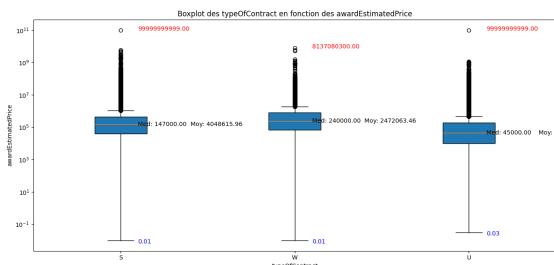


Figure 2. Boxplot des awardEstimatedPrice en fonction des types de contrats

- Le contrat de type S (Fourniture) a une médiane de 147 000 et une moyenne de 4 048 615,96. Il existe un grand nombre de valeurs aberrantes, la plus élevée approchant 10^{11} .
- Le contrat de type W (Travail) a une médiane de 2 400 000 et une moyenne de 24 727 063,46. Il existe également un grand nombre de valeurs aberrantes, le plus élevé dépassant 10^9 .
- Le contrat de type U (Service) a une médiane de 45 000 et une moyenne de 2 358 436,05. Il existe également un grand nombre de valeurs aberrantes, la plus élevée étant également autour de 10^{11} .
- Ces valeurs aberrantes peuvent indiquer la présence de prix de récompense extrêmes dans les données, ou peuvent être des erreurs de saisie. Lors de l'analyse des données, il est nécessaire d'enquêter davantage sur les raisons de ces valeurs aberrantes et de décider si elles doivent être incluses dans l'analyse. De plus, étant donné que la distribution des données à une grande étendue numérique, l'axe des

ordonnées du diagramme en boîte utilise une échelle logarithmique pour mieux visualiser les données à large éventail numérique.

- **awardPrice**

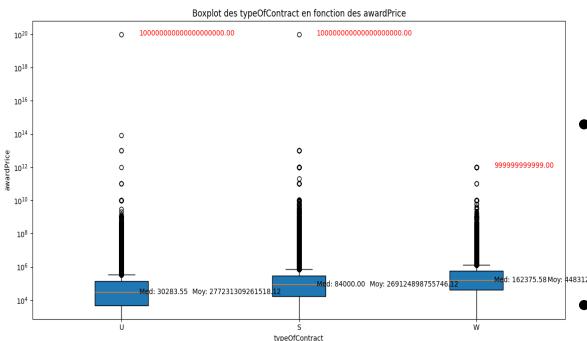
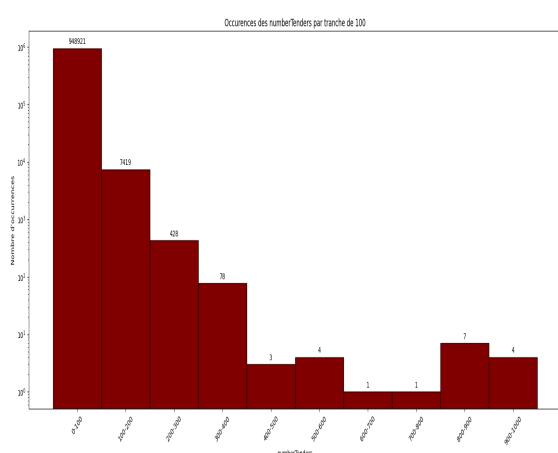


Figure 3. Boxplot des awardPrice en fonction des types de contrats

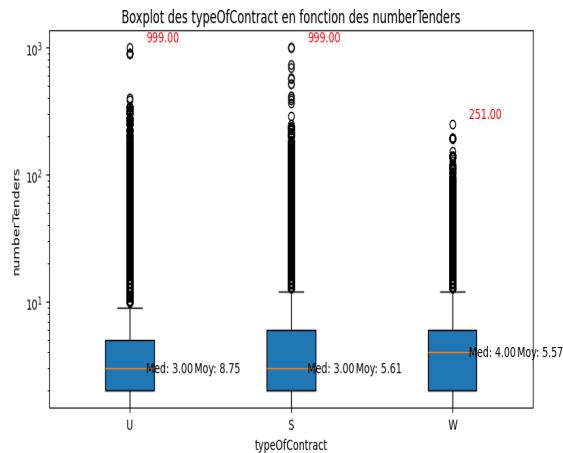
- Le contrat de type U (Service) a une médiane de 30283.55 et une moyenne de 277231309261518.12. Il existe un grand nombre de valeurs aberrantes, la plus élevée approchant 10^{20} .
- Le contrat de type S (Fourniture) a une médiane de 84 000 et une moyenne de 269124898755746.12. Il existe également un grand nombre de valeurs aberrantes, le plus élevé dépassant 10^{20} .
- Le contrat de type de W (Travaux) a une médiane de 162375.58 et une moyenne de 44831290.28. Il existe également un grand nombre de valeurs aberrantes, la plus élevée étant également autour de 10^{12} (Beaucoup moins que les deux autres types de contrats).

- Il y a beaucoup d'anomalies que l'on peut remarquer dans ces données. Premièrement, le prix du type de contrat est énormément inférieur à celui des deux autres types. Deuxièmement, les médianes sont souvent nettement plus basses que les moyennes, ce qui signifie qu'il existe vraiment des données qui sont énormément grandes et n'ont aucun sens. Troisièmement, les chiffres maximum de awardPrice sont presque 10^9 fois plus élevés que ceux de awardEstimatedPrice, ce qui n'est pas du tout normal.

- **numberTenders**



(a) Distribution des numberTenders par nombre d'occurrences



(b) Distribution des numberTenders par type de contrat

Figure 4. Distribution des numberTenders

Le graphique à gauche montre le nombre d'occurrences du nombre d'offres (numberTenders) dans différentes plages. L'axe des ordonnées du graphique à barres utilise

une échelle logarithmique pour mieux visualiser les différences d'ordre de grandeur des données. On peut voir sur le graphique que la plupart des nombres d'offres se concentrent dans la plage de 0 à 100, et que le nombre d'occurrences diminue rapidement avec l'augmentation du nombre d'offres.

Le graphique à droite est un diagramme en boîte qui montre la distribution du nombre d'offres (numberTenders) pour trois types de contrats différents (U (Service), S (Fourniture), W (Travaux)). L'axe des ordonnées du diagramme en boîte utilise également une échelle logarithmique. On peut observer que :

- La médiane du type de contrat Services (U) est d'environ 3, avec une moyenne d'environ 8,75. Il y a quelques valeurs aberrantes, la plus élevée étant de 999.
- La médiane du type de contrat Fournitures (S) est d'environ 3, avec une moyenne d'environ 5,61. Il y a moins de valeurs aberrantes, mais une d'entre elles est de 999.
- La médiane du type de contrat Travaux (W) est d'environ 4, avec une moyenne d'environ 5,57. Il y a quelques valeurs aberrantes, bien que la différence par rapport aux deux autres contrats soit plus faible, la valeur aberrante la plus élevée étant de 251.

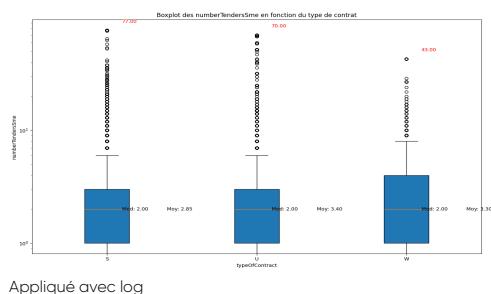
Les médianes et les moyennes des trois contrats sont assez proches, ce qui suggère que la plupart des fluctuations de données puissent ne pas être très importantes. Cependant, étant donné que les valeurs maximales sont assez éloignées, il est possible qu'il y ait des exceptions ou des erreurs apparentes, nécessitant une analyse plus approfondie pour déterminer si elles représentent réellement la situation ou si elles devraient être exclues de l'ensemble de données pour éviter les biais dans les résultats d'analyse.

• **lotsNumber**

Initialement, "lotsNumber" est considéré dans le sujet comme « Nombre total de lots dans la notice d'attribution ». En l'occurrence, "lotsNumber" ne représente pas le nombre total de lots d'un marché, mais est plutôt un identifiant ou un numéro assigné à chaque lot individuel. Ainsi, chaque lot d'un marché donné se voit attribuer un lotsNumber unique qui le distingue des autres lots au sein du même marché.

Sur 1380965 entrées, seulement 77.10% (1 068 827) ont des valeurs qui semblent correctes (un nombre lui est associé), 21.39% (295 427) sont vides et 1.21% (16 711) possèdent des valeurs erronées (un ou plusieurs caractères ou symboles).

- **numberTendersSme**



- Les contrats de type S (fournitures) : la médiane (Med) est de 2,00, la moyenne (Moy) est de 2,85, les valeurs aberrantes sont réparties sur une plage plus large, atteignant jusqu'à 77,00.
- Les contrats de type U (services) : la médiane (Med) est de 2,00, la moyenne (Moy) est de 3,40, il y a plus de valeurs aberrantes, atteignant jusqu'à 70,00.
- Les contrats de type W (travaux) : la médiane (Med) est de 2,00, la moyenne (Moy) est de 3,30, il y a moins de valeurs aberrantes, atteignant jusqu'à 43,00.

À partir de ce diagramme en boîte, nous pouvons observer les points suivants :

- La médiane de tous les trois types de contrats est de 2, ce qui indique qu'au moins la moitié des contrats ont un nombre d'offres de 2 ou moins.
- La moyenne est légèrement supérieure à la médiane, ce qui pourrait être dû à la présence de valeurs aberrantes élevées qui augmentent la moyenne.
- La présence de valeurs aberrantes indique que certains contrats ont un nombre d'offres beaucoup plus élevé que la majorité des contrats.
- La valeur aberrante la plus élevée pour les contrats de type S est de 77, pour les contrats de type U est de 70, et pour les contrats de type W est de 43, ce qui suggère que les contrats de type S et U pourraient avoir plus de cas de nombres d'offres extrêmement élevés.

La présence de valeurs aberrantes peut nécessiter des investigations supplémentaires pour déterminer si elles représentent des comportements de marché spécifiques ou des anomalies dans les données.

- **typeOfContract**

"typeOfContract" se réfère aux types de contrats soient catégorisés comme suit :

Identifiant	Description	Fréquence
S	Fournitures	40.16%
U	Services	39.83%
W	Travaux	20.00%

Table 7. Fréquence des typeOfContract

- S (Fournitures) : Ces contrats concernent l'achat de biens physiques ou matériels. Cela peut inclure des fournitures de bureau, des équipements informatiques, des véhicules, du mobilier, des machines, etc.
- U (Services) : Ces contrats couvrent des prestations de services, qui peuvent être des services de conseil, de maintenance, d'assurance, d'éducation, de santé, des services juridiques, informatiques, de nettoyage, etc.
- W (Travaux) : Ce sont des contrats pour des travaux de construction ou de génie civil. Cela englobe la construction de bâtiments, de routes, de ponts, la rénovation

ou la réparation d'infrastructures existantes, etc.

• **contractDuration**

"contractDuration" est la durée du contrat dont les valeurs numériques sont exprimées en mois. De plus, seulement 1/3 des valeurs présentes dans la colonne sont vides.

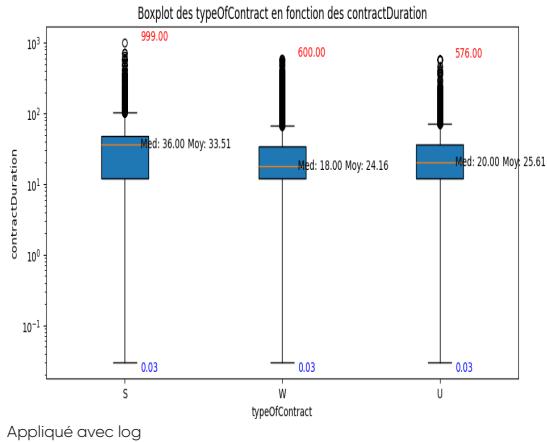


Figure 5. Boxplot des contractDuration en fonction des types de contrats

- Fournitures (S) : La moitié des contrats de fournitures ont une durée de moins de 36 mois.
Valeur aberrante : 999 mois
- Travaux (W) : La moitié des contrats de travaux ont une durée inférieure à 18 mois.
Valeur aberrante : 600 mois
- Services (U) : 20 mois, légèrement supérieur à celui des travaux, mais inférieur aux fournitures.
Valeur aberrante : 576 mois

• **publicityDuration**

"publicityDuration" est une mesure de la durée pendant laquelle un appel d'offres est annoncé publiquement avant la clôture du processus de soumission des offres. La durée de publicité est exprimée en jours.

Cette période de publicité est cruciale dans les marchés publics, car elle garantit la transparence du processus d'attribution et donne aux potentiels soumissionnaires suffisamment de temps pour découvrir l'appel d'offres, préparer et déposer leurs propositions.

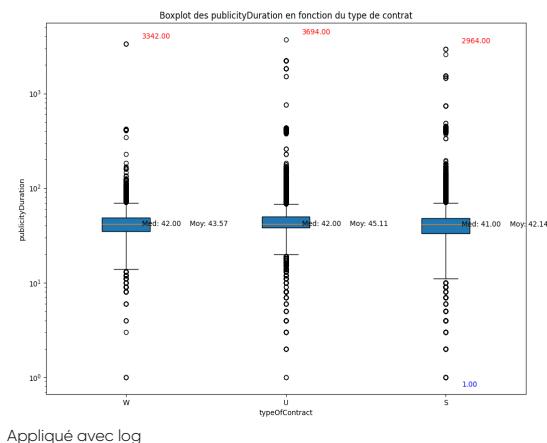


Figure 6. Boxplot des publicityDuration en fonction des types de contrats

La moitié des appels d'offres pour chaque type de contrat a une période de publicité d'environ 40 jours. Les moyennes sont légèrement plus élevées, indiquant que quelques durées de publicité très longues tirent la moyenne vers le haut. Les valeurs extrêmes indiquées par les points noirs suggèrent des anomalies, avec certains appels d'offres ayant des périodes de publicité allant jusqu'à plusieurs années, ce qui peut être atypique et mériterait une investigation supplémentaire pour comprendre leur justesse ou identifier des erreurs potentielles dans les données.

Section paire de variables

- awardEstimatedPrice & awardPrice

Nom	Nombre	Pourcentage
awardEstimatedPrice non vide et awardPrice vide	0	0 %
awardEstimatedPrice vide et awardPrice non vide	765 029	55.40 %
Les deux champs non vides	189 613	13.73 %
Les deux champs vides	426 323	30.87 %

Table 8. Répartition des valeurs pour awardEstimatedPrice et numberTenders

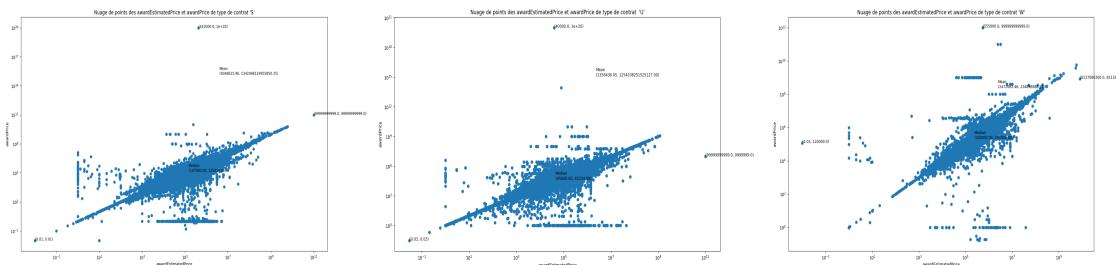


Figure 7. Distribution des awardPrice vs awardEstimatedPrice par type de contrat

D'après les images indiquées ci-dessus, nous pouvons remarquer que, de manière générale, les awardEstimatedPrice et awardPrice sont bien corrélés. De plus, il est évident que sur la plupart des marchés, les transactions sont presque toujours conclues au prix initial définit (Il est évident qu'il peut y avoir des cas spécifiques résultant de variations des prix réglementés), mais la présence du prix initial n'est pas nécessaire [PrixDemandePublic]. Pour les trois types de contrats, dans la plupart des cas, le prix réel correspond au prix estimé, mais il existe également des écarts significatifs. Ces écarts peuvent être dus à une estimation inexacte, des changements dans la portée du projet, des variations des conditions du marché ou d'autres facteurs. Pour les points de données qui s'éloignent de la diagonale, une analyse supplémentaire peut être nécessaire pour comprendre les raisons de ces différences.

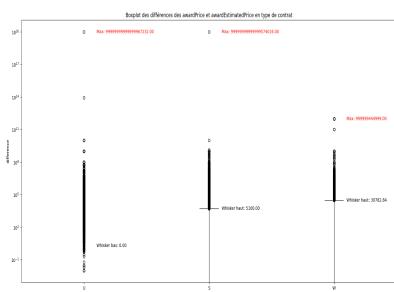


Figure 8. Boxplot des awardPrice vs awardEstimatedPrice par type de contrat

À partir du diagramme de nuage de point précédent, nous avons remarqué qu'il y a des données aberrantes dans la base de données, ce qui a entraîné une distribution de points aberrante. Maintenant, en combinant le diagramme en boîte à gauche montrant la différence entre awardEstimatedPrice et awardPrice, nous pouvons clairement voir qu'il y a de nombreuses données aberrantes dans la base de données, en particulier certaines données aberrantes extrêmement élevées. En raison de la présence de ces données aberrantes, après l'application de la fonction logarithmique, le diagramme en boîte a été étiré vers le haut au point où il est impossible d'effectuer une analyse. Par conséquent, il est évident que certaines données sont complètement anormales et ne doivent pas être prises en compte dans le contexte de l'analyse appliquée.

- **awardEstimatedPrice & numberTenders**

Les trois graphiques de dispersion que compare le nombre d'offres reçues pour un lot (numberTenders) à l'estimation du prix de ce lot (awardEstimatedPrice) pour les contrats de fournitures (S), de services (U) et de travaux (W).

Nom	Nombre	Pourcentage
awardEstimatedPrice non vide et numberTenders vide	798 959	57.86%
awardEstimatedPrice vide et numberTenders non vide	31 706	2.30%
Les deux champs non vides	157907	11.43%
Les deux champs vides	392393	28.41%

Table 9. Répartition des valeurs pour awardEstimatedPrice et numberTenders

Pour chaque type de contrat (S, U, W), il existe un regroupement de points le long de la gamme basse à moyenne du nombre d'offres, ce qui pourrait suggérer que la plupart des lots reçoivent un nombre relativement faible à modéré d'offres.

La corrélation ne semble pas exister entre le nombre d'offres et l'estimation du prix des lots. En effet, dans un ensemble de données idéal, il est attendu de voir un certain modèle ou tendance. Par exemple, des lots avec des prix estimés plus élevés pourraient attirer plus d'offres en raison de leur valeur plus grande, ce qui n'est pas clairement visible dans les graphiques.

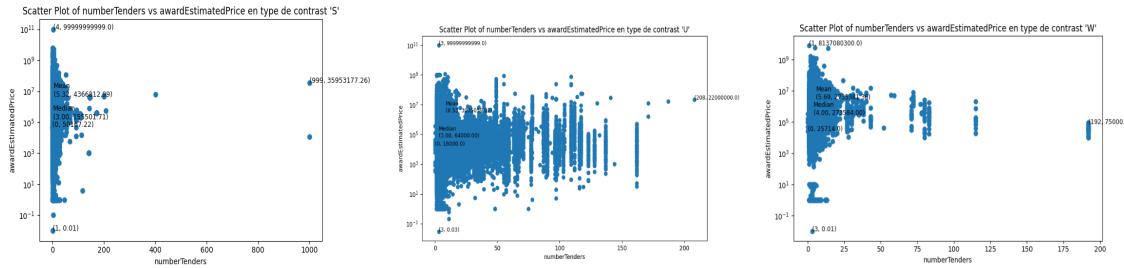


Figure 9. Distribution des awardEstimatedPrice vs numberTenders par type de contrat

Les valeurs aberrantes suggèrent que certains lots sont exceptionnels par rapport à la majorité des lots. Ces valeurs aberrantes nécessitent une analyse plus approfondie pour déterminer si elles sont correctes ou si elles résultent d'erreurs de saisie.

- **awardEstimatedPrice & accelerated**

La variable awardEstimatedPrice représente l'estimation financière pour un lot spécifique dans un appel d'offres, tandis que accelerated est un indicateur booléen qui signifie si une procédure accélérée a été utilisée pour ce lot.

Nom	Nombre	Pourcentage
awardEstimatedPrice non vide et accelerated vide	2 099	0.15%
awardEstimatedPrice vide et accelerated non vide	189 157	13.70%
Les deux champs non vides	456	0.03%
Les deux champs vides	1189253	86.12%

Table 10. Répartition des valeurs pour awardEstimatedPrice et accelerated

Seulement 456 lots ont des informations à la fois sur l'estimation du prix et l'utilisation de la procédure accélérée, et la grande majorité des lots n'ont pas d'informations sur l'une ou l'autre ou les deux de ces variables.

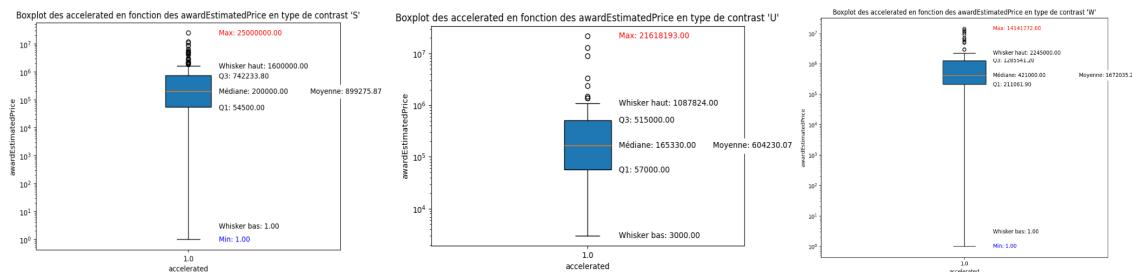


Figure 10. Boxplot des awardEstimatedPrice vs accelerated par type de contrat

Dans le premier diagramme pour les contrats de fournitures (S), on observe une médiane d'environ 200 000 avec des prix allant jusqu'à un maximum extrême de 2 500 000. Les valeurs du premier et troisième quartile montrent une distribution assez large des prix estimés.

Pour les services (U), la médiane est autour de 165 330 avec des prix maximum encore plus élevés, suggérant que les services peuvent avoir des estimations de prix potentiellement plus élevées que les fournitures.

Dans le troisième diagramme, dédié aux travaux (W), la médiane est plus élevée que pour les fournitures et les services, s'élevant à environ 421 000, ce qui peut refléter la nature

généralement coûteuse des projets de construction et de travaux.

Les lots soumis à une procédure accélérée présentent une large gamme d'estimations de prix pour chaque type de contrat, avec quelques valeurs extrêmes qui pourraient indiquer des cas particuliers ou des erreurs de saisie.

- **awardEstimatedPrice & numberTendersSme**

Nom	Nombre	Pourcentage
awardEstimatedPrice non vide et numberTendersSme vide	28 622	2.03%
awardEstimatedPrice vide et numberTendersSme non vide	174 209	12.37%
Les deux champs non vides	15404	1.09%
Les deux champs vides	1162730	82.51%

Table 11. Répartition des valeurs pour awardEstimatedPrice et numberTendersSme

Ces pourcentages montrent qu'une grande partie des lots manquent d'une ou des deux informations. Seulement un petit pourcentage des lots ont des informations complètes pour ces deux variables, ce qui limite les analyses possibles sur la relation entre ces deux facteurs.

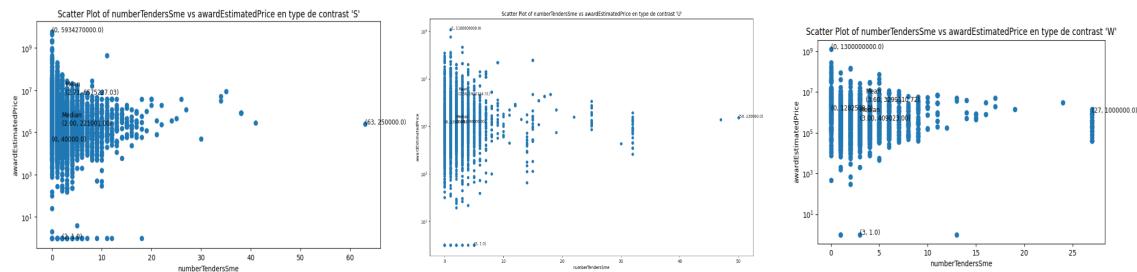


Figure 11. Distribution des awardEstimatedPrice vs numberTendersSme par type de contrat

Les points sont largement dispersés sans suivre une direction spécifique, ce qui suggère qu'il n'y ait pas de relation forte et cohérente entre le nombre d'offres de PME et l'estimation du prix du lot. De plus, il n'y a pas de forme ou de tendance qui indiquerait une corrélation positive ou négative.

Les valeurs extrêmes, qui sont très éloignées de la majorité des autres points, peuvent influencer la perception de la corrélation. Cependant, même en les excluant, aucun motif clair ne se dégage.

- **awardPrice & cpv**

Contrairement aux autres statistiques précédentes, awardPrice et cpv ont peu de valeurs nulles en commun.

Deux cas de figures pourraient se distinguer entre les deux variables :

- Attract du lot : Les lots avec des prix estimés plus élevés pourraient être plus attractifs pour toutes les entreprises, y compris les PME, en raison du potentiel de profit plus important. Toutefois, si l'estimation du prix est trop élevée, cela pourrait dissuader les PME qui pourraient ne pas avoir les ressources pour gérer de gros contrats.

- Compétitivité : Les PME pourraient être plus compétitives dans certaines gammes de prix où elles peuvent offrir de meilleurs prix ou une spécialisation plus pointue par rapport à des entreprises plus grandes.

Nom	Nombre	Pourcentage
awardPrice non vide et cpv vide	426 304	31.14%
awardPrice vide et cpv non vide	52	<0.01%
Les deux champs non vides	954590	69.78%
Les deux champs vides	19	<0.01%

Table 12. Répartition des valeurs pour awardPrice et cpv

Chaque boîte représente la répartition des prix attribués pour un groupe spécifique de CPV.

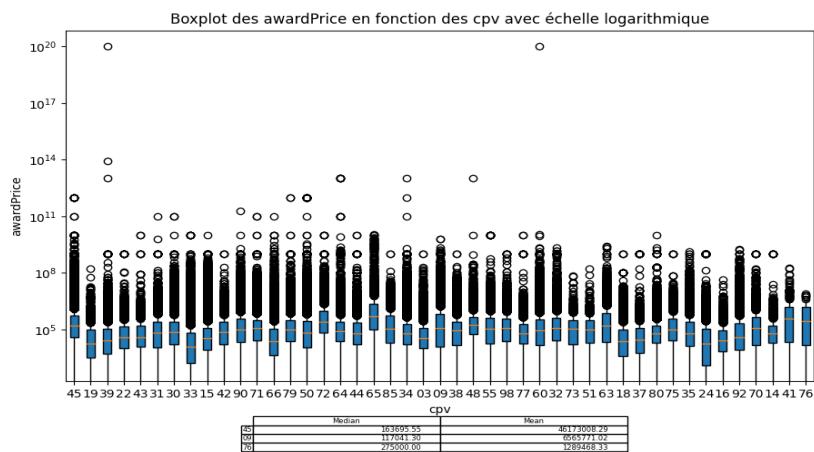


Figure 12. Boîte à moustache des occurrences des cpv par division par type de contrat

Un élément commun à tous les graphiques est la présence de valeurs aberrantes, indiquées par des points isolés au-dessus des moustaches supérieures. Ces points signalent des lots dont les prix attribués s'écartent significativement de la distribution typique, suggérant soit des cas uniques de contrats avec des estimations de coûts exceptionnellement élevés ou bas, soit des erreurs de saisie des données.

Deux valeurs aberrantes particulièrement notables : 39 (Meubles, aménagements, appareils électroménagers et produits de nettoyage et 60 (Services de transport (à l'exclusion du transport des déchets)) dominent les autres par leur ampleur, indiquant des prix attribués qui surpassent de loin les médianes et moyennes de leur groupe de CPV respectif.

- awardPrice & les variables booléennes

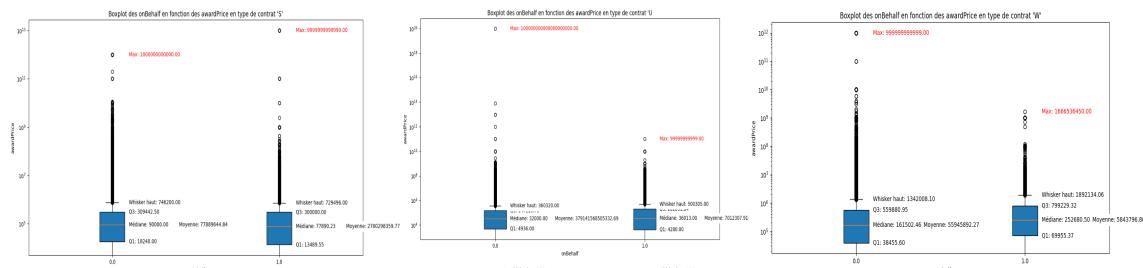


Figure 13. Distribution des awardPrice vs onBehalf par type de contrat

Le graphique ci-dessus montre les boîte à moustache de awardPrice et onBehalf pour différents types de contrats. En observant le graphique, nous pouvons remarquer que :

- Les valeurs maximales de tous les trois types de contrats sont très élevées, la valeur maximale du type de contrat 'U' (services) atteignant 10^{20} , ce qui pourrait indiquer la présence de valeurs aberrantes ou d'erreurs extrêmes dans les données.
- Les valeurs des médianes et des quartiles dans les données sont relativement basses, indiquant que la plupart des prix attribués (awardPrice) sont bas ou dans la norme.
- La moyenne est élevée en raison des valeurs maximales extrêmes, ce qui suggère une possible asymétrie des données.
- À part les remarques sur les anomalies, nous constatons que la médiane du type de contrat 'W' (travaux) est relativement élevée par rapport aux autres types de contrat, ce qui semble cohérent.

Ici, nous montrons que pour awardPrice et onBehalf en termes de la variable booléenne. Puisque les autres variables booléennes (jointProcurement, fraAgreement, accelerated, outOfDirectives, numberTendersSme, subContracted, gpa, multipleCae et renewal) présentent les mêmes problèmes, nous avons donc décidé de les présenter après le nettoyage des données.

- awardPrice & topType

Dans cette partie, nous étudions la relation entre le prix des marchés publics (awardPrice) et le type de procédure d'attribution (topType). Nous analysons la répartition des types de procédure d'attribution en fonction des prix effectifs des attributions de lots.

Nom	Nombre	Pourcentage
awardPrice non vide et topType vide	56	0,004%
awardPrice vide et topType non vide	426 188	30,86%
Les deux champs non vides	954586	69,12%
Les deux champs vides	135	0,010%

Table 13. Répartition des valeurs pour awardPrice et topType

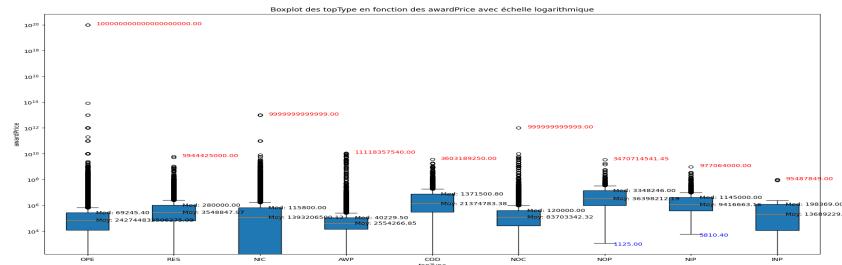


Figure 14. Boîte à moustache des awardPrice vs topType

En examinant le tableau ci-dessus, nous pouvons conclure que les résultats obtenus lors de l'analyse de ces deux variables sont basés plus de 60% de la base de données. Les graphiques présentent la relation définie précédemment. Nous avons choisi de représenter les données sur une échelle logarithmique afin de mieux mettre en évidence la grande disparité des valeurs dans les prix effectifs. Les points clés que nous observons sont les suivants :

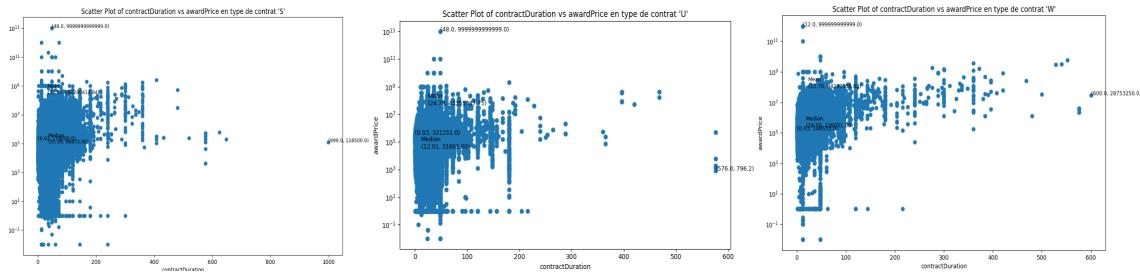
- Procédures ouvertes (OPE) : Les médianes de ces catégories sont relativement basses comparées aux moyennes, suggérant que la majorité des lots ont des prix plus modestes, mais que les moyennes sont tirées vers le haut par les valeurs aberrantes.
- Négocie avec un appel à la concurrence (NIC) et Procédure restreinte (RES) : Bien que la médiane soit légèrement plus élevée que dans les OPE, ces catégories présentent également des valeurs aberrantes exceptionnelles qui dominent l'échelle de prix, ce qui indique des cas extrêmes ou des erreurs de saisie des données telles que 999999999999 pour le prix effectif.
- Partenariat innovant (INP) : Bien que ces catégories montrent moins de données, les valeurs aberrantes sont toujours présentes et sont très élevées par rapport aux médianes, ce qui peut indiquer des projets uniques de grande envergure, ce qui est peu probable.
- Nous observons que pour les contrats de fourniture de biens et de services (NOP) et les contrats de travaux (NIP), les valeurs des prix effectifs commencent à des niveaux relativement élevés, à partir de 1 125€ et 5 810€ respectivement. Ceci suggère que pour ces types de contrats, il est courant que les appels d'offres débutent avec ces montants. Il est à noter qu'aucun appel d'offres avec des valeurs inférieures n'est répertorié dans notre base de données.

En ce qui concerne les valeurs aberrantes, on observe plusieurs cas extrêmes avec des montants de prix étonnamment élevés, allant jusqu'à des nombres astronomiques. Ces valeurs peuvent indiquer des projets d'une envergure exceptionnelle ou, plus probablement, des erreurs dans les données. En effet, des valeurs comme "10000000000000000000000.00" ou "9999999999.00" semblent trop spécifiques et trop élevées pour représenter des montants réels de marchés publics et suggèrent plutôt des erreurs de saisie ou de format.

• awardPrice & contractDuration

Dans cette partie, nous étudions la relation entre le prix des marchés publics (awardPrice) et la durée de contrat (contractDuration). Nous analysons la répartition des durées des contrats en fonction des prix effectifs des attributions de lots.

Nom	Nombre	Pourcentage
awardPrice non vide et contractDuration vide	344 177	24,92%
awardPrice vide et contractDuration non vide	232 303	16,82%
Les deux champs non vides	610465	44,20%
Les deux champs vides	194020	14,05%

Table 14. Répartition des valeurs pour awardPrice et contractDuration**Figure 15.** Distribution des awardPrice vs contractDuration par type de contrat

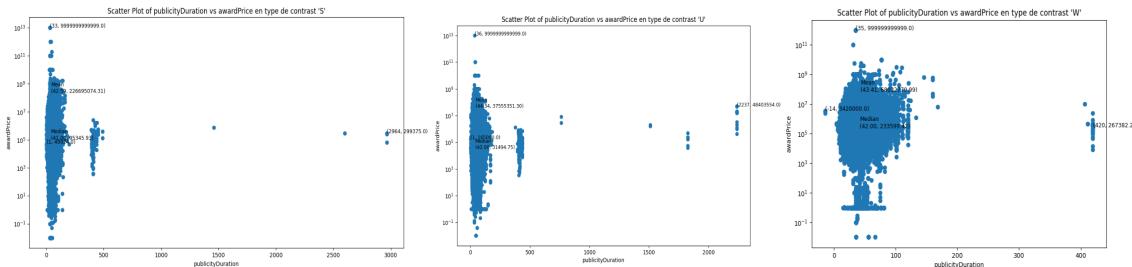
En examinant le tableau ci-dessus, nous pouvons conclure que les résultats obtenus lors de l'analyse de ces deux variables sont basés sur 44,20% de la base de données. Les graphiques présentent la relation définie précédemment par type de contrat. Nous avons choisi de représenter les données sur une échelle logarithmique afin de mieux mettre en évidence la grande disparité des valeurs dans les prix effectifs. Les points clés que nous observons sont les suivants :

- Sur les trois graphiques du contrat fournitures 'S', travaux 'W', services 'U', la concentration des points vers des durées de contrat plus courtes pourrait indiquer une préférence pour des attributions de contrat plus rapides. Cependant, les valeurs aberrantes élevées suggèrent l'existence de contrats avec des prix d'attribution exceptionnellement élevés tels que 9999999999999, qui pourrait être des erreurs lors de la saisie de l'utilisateur, car de tels montants sont inimaginables. Les valeurs aberrantes au niveau de l'axe des abscisses sont aussi très bizarres, avoir des contrats d'une durée de plus de 600 mois qui représente 50 ans, est inimaginable. Une raison pour cela pourrait être le fait que l'utilisateur s'est trompé lors de la saisie, vu que publicityDuration est en jours, il a peut-être cru que pour cette variable aussi, c'était en jour.
- La concentration moyenne des durées de publicité se situe approximativement au même niveau pour les trois, mais les points sont indiscernables en raison de la grande disparité et des valeurs aberrantes dans les données. Un nettoyage est nécessaire pour permettre une meilleure visualisation et une analyse précise de cette paire de variables. Actuellement, nous n'observons pas de corrélation directe entre les variables avant le nettoyage.

• awardPrice & publicityDuration

Dans cette partie, nous étudions la relation entre le prix des marchés publics (awardPrice) et la durée de publicité (publicityDuration). Nous analysons la répartition des durées de publicité en fonction des prix effectifs des attributions de lots.

Nom	Nombre	Pourcentage
awardPrice non vide et publicityDuration vide	235 927	17,08%
awardPrice vide et publicityDuration non vide	316 659	22,93%
Les deux champs non vides	718715	52,04%
Les deux champs vides	109664	7,941%

Table 15. Répartition des valeurs pour awardPrice et publicityDuration**Figure 16.** Distribution des awardPrice vs publicityDuration par type de contrat

En examinant le tableau ci-dessus, nous pouvons conclure que les résultats obtenus lors de l'analyse de ces deux variables sont basés sur plus de 50% des données de la base de données. Les graphiques présentent la relation définie précédemment par type de contrat. Nous avons choisi de représenter les données sur une échelle logarithmique afin de mieux mettre en évidence la grande disparité des valeurs dans les prix effectifs. Les points clés que nous observons sont les suivants :

- Dans le premier graphique pour les trois types de contrat fournitures 'S', services 'U', travaux 'W', on observe une concentration de points autour des faibles durées de publicité, indiquant que la majorité des lots avec ce type de contrat ont tendance à être attribués à des prix plus bas après des périodes de publicité courtes. La moyenne et la médiane de awardPrice sont éloignées, ce qui pourrait indiquer une distribution asymétrique avec une longue queue de valeurs plus élevées. Les valeurs aberrantes extrêmes suggèrent que quelques lots ont été attribués à des prix anormalement élevés ou après des périodes de publicité inhabituellement longues, tels que 3 000 jours, revenant à 100 mois. Il y a dû avoir une erreur lors du calcul de la durée de publicité. Nous avons des valeurs aberrantes extrêmement éloignées de la concentration, ce qui rend l'analyse moins concluante, car on n'arrive plus à différencier les points qui se trouvent dans la concentration.
- Le type de contrat travaux 'W' a une durée moyenne de publicité deux fois plus courtes que les deux autres avec un prix effectif 100 fois moins que les autres.

Pour chaque type de contrat, le nettoyage des données et deuxième analyse est nécessaire pour avoir plus de détails sur la concentration des données. Visuellement, il semble y avoir une certaine concentration de données autour des durées de publicité plus courtes et des prix d'attribution plus bas, mais cela n'indique pas nécessairement une corrélation directe.

- **numberTenders & numberTendersSme**

Dans cette partie, nous allons étudier la relation entre le nombre total d'offres effectuées (numberTenders) et ceux émis par des petites et moyennes entreprises (numberTendersSme). Nous examinerons la répartition des offres soumises par les PME parmi l'ensemble des offres.

Nom	Nombre	Pourcentage
numberTenders non vide et numberTendersSme vide	912840	66,10%
numberTenders vide et numberTendersSme non vide	0	0,000%
Les deux champs non vides	44026	3,188%
Les deux champs vides	424099	30,71%

Table 16. Répartition des valeurs pour numberTenders et numberTendersSme

Dans le tableau, nous constatons que pour plus de 60% des offres émises, nous n'avons pas d'information sur la présence ou l'absence de PME. En revanche, pour toutes les offres émises par des PME, nous disposons du nombre total d'offres émises. Cela signifie que moins de 4% des données peuvent être utilisées pour visualiser la répartition des offres soumises par les PME parmi l'ensemble des offres.

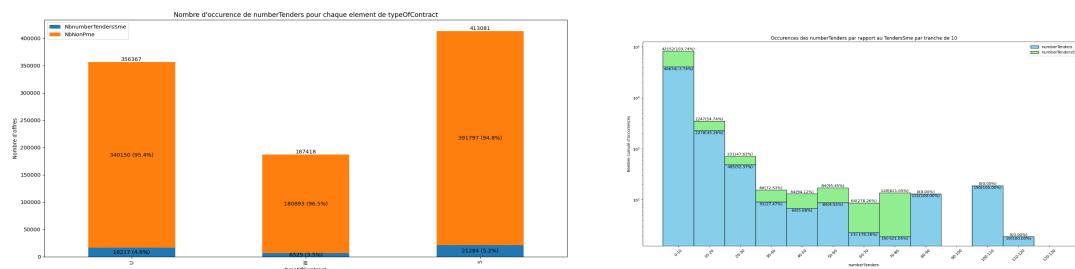


Figure 17. Distribution des numberTenders vs numberTendersSme

Les deux graphiques montrent comment les offres soumises par les petites et moyennes entreprises (PME) se répartissent parmi l'ensemble des offres disponibles. Le graphique de gauche illustre cette répartition selon les types de contrats, tandis que celui de droite la présente par tranches de 10. Voici les points clés que nous observons :

- Le nombre d'offres soumises par les PME est significativement inférieur par rapport au nombre total d'offres, comme le montre le graphique de gauche. Cette différence pourrait refléter les défis que les PME rencontrent pour concurrencer dans des marchés dominés par de grandes entreprises ou pour répondre à des appels d'offres complexes.
- Le type de contrat concernant les fournitures(S) est celui qui reçoit le plus d'offres ainsi que le plus d'offres venant des PME.
- Sur le graphique de droite, on observe que la majorité des offres se situent dans les tranches inférieures (0-40), ce qui indique que la plupart des appels d'offres reçoivent un nombre relativement faible d'offres. Cela pourrait suggérer une compétition limitée ou un marché fortement segmenté.
- Toutefois, nous observons des valeurs incompréhensibles comme par exemple pour la tranche de 0 à 10, nous avons plus d'offres émises par des PME que le total d'offres émises, ce qui est aberrant. Cela pourrait être des fraudes.

- **numberTenders & topType**

Dans cette section, nous explorons la relation entre deux variables : le nombre d'offres soumises pour un lot (*numberTenders*) et le type de procédure d'attribution (*topType*). Pour rappel, *numberTenders* représente le nombre d'offres effectuées, tandis que *topType* désigne la méthode d'attribution employée. Notre objectif est d'analyser comment le nombre d'offres varie selon le type de procédure utilisé.

Nom	Nombre	Pourcentage
numberTenders non vide et topType vide	40	0,002%
numberTenders vide et topType non vide	423948	30,69%
Les deux champs non vides	956826	69,29%
Les deux champs vides	158923	0,011%

Table 17. Répartition des valeurs pour numberTenders et topType

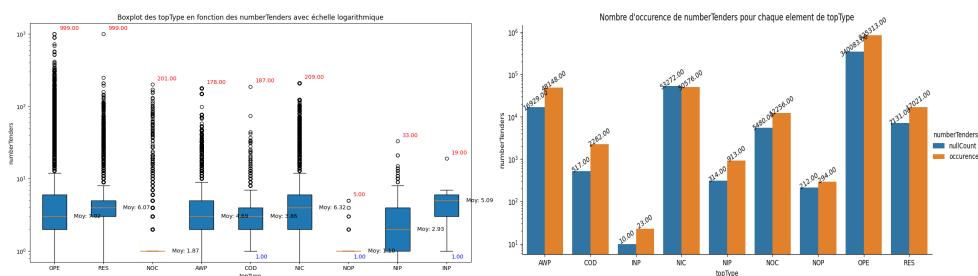


Figure 18. Distribution des numberTenders vs topType

La figure gauche qui représente la distribution du nombre d'offres par rapport aux types de procédures sous format boîte à moustaches. Alors que la figure de droite représente la même distribution sous format histogramme. Les conclusions suivantes ont pu être faites :

- Certains types de procédures, comme les procédures ouvertes (OPE), les procédures restreintes (RES), les procédures négociées avec un appel à la concurrence (NIC), ont une moyenne de soumissions et un nombre de soumissions plus élevée que d'autres, ce qui pourrait indiquer une plus grande ouverture ou une plus grande accessibilité de ces types de marchés aux fournisseurs. D'autre part, les types de procédures négociées sans appel à la concurrence préalable (NOC, NOP) présentent des moyennes et un nombre de soumissions nettement plus basses, ce qui peut signifier une spécialisation ou des exigences plus strictes, réduisant ainsi le nombre de fournisseurs potentiels.
- Les valeurs aberrantes, particulièrement élevées pour les types OPE et RES, soulignent que certains appels d'offres attirent un nombre exceptionnellement élevé de propositions telles que 999. Ces cas pourraient être dus à des marchés particulièrement attractifs ou à des appels d'offres ouverts sans restriction significative à l'entrée, mais cela semble peu probable. Sinon, cela pourrait être des erreurs lors du remplissage du champ par l'utilisateur ou une fraude quelconque.
- La présence de valeurs aberrantes basses pour des types tels que NOP suggère l'existence d'appels d'offres où peu d'offres sont soumises, possiblement à cause d'exigences très spécifiques ou de la nature moins attrayante de ces marchés.

2.3 Nettoyage

2.3.1 Les applications de nettoyage des données

Les étapes de nettoyage appliquées :

- **API Ted**

Dans cette partie, je vais uniquement me concentrer sur la correction des données.

L'appel API TED permet de récupérer des informations à partir de fichiers XML correspondants aux "tedCanId", puis mettre à jour la base de données avec les nouvelles données récupérées.

Après avoir récupéré tous les tedCanId uniques de la table Lots dans la base de données, le script compare les informations existantes dans la base de données avec celles extraites du fichier XML pour chaque champ (la liste des champs est consultable [ici](#)).

En cas de différence, si la valeur du XML n'est pas vide, il y a un enregistrement de la modification dans un fichier CSV "historique_modifications.csv" et une mise à jour la base de données avec la valeur du XML.

- **Nettoyage manuel**

D'après la discussion précédente sur les données brutes, nous avons remarqué les erreurs ou des données manquantes. Cette partie mentionne comment le nettoyage manuel des différents problèmes identifiés dans la base de données a été faite. Nous avons utilisé plusieurs méthodes, telles que la substitution ou la suppression. Chacune de ces méthodes est importante pour obtenir un résultat cohérent et homogène. Les différentes façons de procéder au nettoyage pour chacune des variables sont expliquées dans cette section.

- **awardDate**

Pour la variable en question, nous avons constaté des anomalies concernant les années de décision d'attribution. Les données couvrent normalement la période de 2010 à 2020, cependant, nous avons identifié 26 721 dates de décision antérieures à 2010 (représentant 1,93% des données), et seulement 3 dates postérieures à 2020 (moins de 0,0002% des données). Ces valeurs sont non seulement inhabituelles, mais aussi incohérentes. Nous avons donc présumé qu'elles étaient probablement erronées. Par ailleurs, en analysant d'autres variables, nous avons remarqué que la variable tedCanId comprenait l'année de publication de l'appel d'offres suivie de son identifiant TED. Nous avons jugé plus adéquat de remplacer ces valeurs aberrantes par l'année de publication sur le TED, plutôt que de les mettre à zéro, car cela aurait faussement impliqué que ces lots n'ont pas fait l'objet d'une décision d'attribution, ce qui n'est pas le cas. Ainsi, afin de maintenir l'intégrité des données, nous avons procédé à une substitution logique. Les années supérieures à 2020 et inférieures à 2010 ont été remplacées par les quatre premières valeurs de leur tedCanId respectif (correspondant à l'année de publication).

- **accelerated** La variable accelerated indique si la procédure rapide a été utilisée lors de l'appel d'offres. Lors de la vérification des données, nous avons remarqué que seules des valeurs de 1 ou nulles étaient présentes. Les champs étaient donc

renseignés uniquement lorsqu'une procédure rapide avait été effectivement utilisée. Nous nous sommes alors demandé si le fait que les champs soient remplis uniquement lorsqu'il est vrai signifiait que les personnes ne les remplissaient tout simplement pas lorsque cela était faux. En suivant une logique similaire à celle d'une case à cocher, les utilisateurs cochent la case uniquement lorsque c'est vrai, donc elle reste à nulle lorsque cela est faux. Au départ, seulement 0,18% de la base de données était renseigné avec des valeurs uniquement à vrai, ce qui rendait les analyses effectuées avec cette variable peu informatives. Afin d'obtenir un plus grand nombre de données, suivant la logique que nous avons mentionnée, nous avons remplacé les données non renseignées, donc vides, par des zéros (faux). Avec cette modification, nous utilisons l'intégralité des données pour l'analyse.

- **contractDuration** La variable contractDuration nous informe sur la durée des contrats, exprimée en mois. Les données brutes fournies par FOPPA contiennent des valeurs incohérentes et aberrantes. Selon certains extraits de loi sur la durée des contrats que nous avons consulté, la plupart des contrats ne peuvent excéder une durée de 8 ans, soit 96 mois [**PrixDureeContratHuitAns**], tandis que dans certains cas exceptionnels, la durée peut aller jusqu'à 12 ans, soit 144 mois [**PrixDureeContratDouzeAns**]. En tenant compte de ces informations, nous avons établi une limite de 145 mois. Lors de l'analyse des données brutes, nous avons remarqué que 1465 entrées dépassaient cette limite, soit 0,10% des données. Cependant, parmi ces 0,10% des données, de nombreuses étaient très incohérentes, comme des valeurs de 999, qui semblent être des erreurs de saisie ou des fraudes, ou encore des valeurs de 600 correspondants à plus de 50 ans de contrat. Étant donné que la durée de contrat n'est pas la seule variable en jeu, nous avons également examiné une deuxième variable, la durée de publicité, exprimée en jours. Nous nous sommes demandés si les utilisateurs avaient peut-être commis une erreur d'échelle lors de la saisie des données. Si la durée du contrat a été saisie en jours, cela pourrait expliquer les valeurs aberrantes. Par exemple, 720 jours correspondent à environ 24 mois, ce qui semble plus logique.

Suivant cette logique, nous avons remplacé toutes les durées de contrat supérieures à 145 mois par leur division par 30 (considérant qu'il s'agit d'erreurs de saisie en jours).

Nous avons également remarqué un grand nombre de valeurs aberrantes concernant les durées minimales des contrats. Sur les 1572 valeurs, représentant 0,11% du total, qui étaient inférieures à 1 mois, une analyse approfondie a montré que la plupart étaient justifiées, comme dans le cas des services de transport (hors transport des déchets), identifiés par le code CPV 60. Cependant, il y avait 370 cas où la durée des contrats dans le secteur des travaux de construction et du BTP (code CPV 45) était inférieure à 1 mois, ce qui semblait incohérent. Par conséquent, nous avons décidé de filtrer ces valeurs en les définissant à null pour les lots associés au code CPV 45 et ayant une durée inférieure à 1 mois. Cette démarche visait à obtenir des données plus homogènes et cohérentes.

- **lotsNumber** En examinant le sujet, nous avons noté une contradiction entre la définition de la variable donnée et les données réelles de la base de données. Le sujet précise que cette variable devrait représenter le nombre total de lots dans la notice d'attribution, mais dans la base de données, elle correspondait plutôt

au numéro de chaque lot dans cette notice. Pour rectifier cette incohérence, nous avons créé une nouvelle colonne appelée **totalLot**, dans laquelle nous avons enregistré le nombre total de lots pour chaque avis d'attribution en comptant le nombre d'occurrences de chaque identifiant de marché (**tedCanId**) dans les données. Ensuite, nous avons rempli cette colonne en effectuant une jointure par l'identifiant de marché (**tedCanId**).

- **publicityDuration** La variable **publicityDuration** indique la durée de publicité pour l'appel d'offres, mesurée en jours. Nous avons identifié des valeurs aberrantes au niveau des valeurs minimales. Les valeurs inférieures à zéro étaient incohérentes et pourraient résulter d'une erreur lors du calcul. Comme elles étaient peu nombreuses (seulement 6 valeurs, soit moins de 0,004% de nos données), supprimer ces valeurs n'aurait pas d'impact significatif sur nos analyses. Pour les valeurs se situant entre 0 et 5, nous les avons remplacées par 5. En effet, selon le texte de loi sur les marchés publics [**DureePublicteUn**][**DureePublicteDeux**], le délai minimal, même en tenant compte des cas extrêmes d'urgence, est de 5 jours. Comme nous disposions de plus de données pour ces cas (141 occurrences), nous avons jugé plus cohérent de les substituer par le délai minimal. En ce qui concerne les valeurs aberrantes supérieures à 144, approximativement équivalent à un semestre (6 mois), nous avons observé qu'elles étaient situées au-dessus de la moustache dans nos analyses, indiquant qu'elles étaient effectivement aberrantes. Nous avons ainsi identifié 606 de ces valeurs, représentant 0,043% des données concernées. Avec un si petit nombre de données, nous avons choisi de remplacer ces valeurs aberrantes par 0.
- **awardPrice & awardEstimatedPrice** Les variables **awardPrice** et **awardEstimatedPrice** présentent la plus grande disparité dans leurs valeurs, ainsi que le plus grand nombre de valeurs aberrantes. Pour éviter de fausser les données lors du traitement des valeurs aberrantes, nous avons suivi une approche logique pour chacune de ces variables, comme décrit ci-dessous :

Prenons l'exemple de la variable **awardPrice**, bien que la même logique ait été appliquée à **awardEstimatedPrice**. Nous avons regroupé tous les **awardPrice** non vides selon plusieurs critères que nous avons jugés importants pour déterminer le prix effectif d'un lot. Nous avons formé des groupes de similarités en considérant les lots qui appartiennent à la même division du CPV (les 2 premiers chiffres du code CPV), qui ont la même valeur de **subContracted**, une durée de contrat similaire par tranche de 12 mois (soit un an), et enfin qui ont le même type de contrat. Cette sélection nous a permis d'obtenir des groupes de **awardPrice** avec leurs **lotId** correspondants et leurs groupes respectifs. Nous avons observé plusieurs centaines de groupes de similarités.

Pour chaque groupe, nous avons calculé les limites supérieures et inférieures de la plage de valeurs (*moustache haute* et *moustache basse*). Ensuite, nous avons remplacé les valeurs d'**awardPrice** de chaque lot par le montant de la *moustache basse* de son groupe respectif s'il était inférieur, ou par le montant de la *moustache haute* s'il était supérieur. Les lots dont le prix effectif était inférieur à 100 ou supérieur à 10 000 000, considérés comme des valeurs aberrantes incompréhensibles, ont été remplacés par la médiane de leur groupe respectif.

Après cette étape, les valeurs étaient nettement moins dispersées au sein des

groupes de similarités. Cependant, certains lots pouvaient encore être considérés comme aberrants malgré cela. Pour résoudre ce problème, une deuxième manipulation a été effectuée pour supprimer les valeurs qui restaient en dehors des seuils fixés, soit en dessous de 100 ou au-dessus de 10 000 000.

Une fois que *awardPrice* et *awardEstimatedPrice* ont été nettoyées, nous avons examiné la différence entre elles. Nous avons remarqué que parfois, il y avait des écarts extrêmes entre les prix estimés et les prix effectifs dans les annonces d'attribution. Ces écarts sont incohérents dans certains cas où le prix effectif est de 99 999 999 et le prix estimé est de 40 000, ou vice-versa. Pour résoudre cette incohérence, nous avons utilisé le même système de regroupement par similarité, mais cette fois-ci en prenant la différence entre *awardPrice* et *awardEstimatedPrice*, uniquement pour les différences positives. Nous avons appliqué les mêmes critères de regroupement, en tenant compte du code CPV (les 2 premiers chiffres), de la valeur de *subContracted*, de la durée du contrat par tranche de 12 mois (soit un an), et enfin du type de contrat. Ensuite, nous avons calculé les moustaches haute et basse pour chaque groupe respectif, et ainsi substitué les valeurs aberrantes dans leur groupe par la moustache haute si elles étaient au-dessus de celle-ci, ou par la moustache basse si elles étaient en dessous. Dans ce processus, nous n'avons pas remplacé les valeurs par la médiane, car nous n'avions plus de valeurs aberrantes en dessous de 100 et au-dessus de 10 000 000 grâce au nettoyage précédent.

Nous avons ainsi obtenu des données avec beaucoup moins de disparités, les valeurs ayant été substituées par des valeurs similaires. Cette technique nous a également permis d'éviter la perte de données en les mettant toutes à zéro. Nous avons constaté que 13% des données de *awardPrice* non nulles ont été mises à zéro, ainsi que 12% des données de *awardEstimatedPrice* non nulles.

- **numberTenders** Pour cette variable, nous avons utilisé la même méthode de substitution de valeurs aberrantes en regroupant les données similaires. En examinant le nombre d'offres émises par les entreprises pour les lots, nous avons repéré plusieurs valeurs incohérentes et aberrantes, notamment quatre occurrences à 999. Ce chiffre semble peu réaliste pour le nombre d'offres soumises lors d'un appel d'offres. Pour remédier à cela, nous avons regroupé les données en considérant les lots partageant les mêmes caractéristiques, telles que la catégorie du code CPV, le niveau de sous-traitance, la durée du contrat sur une base annuelle et le type de contrat. Ensuite, nous avons corrigé les valeurs aberrantes en remplaçant celles qui dépassaient les limites définies par les "moustaches" hautes et basses dans chaque groupe.

Une fois les données traitées, nous avons également examiné le champ *numberTendersSme*, car dans plusieurs cas, le nombre d'offres émises par des PME était supérieur au nombre total d'offres émises. Nous avons constaté que ces données étaient incohérentes. Par conséquent, nous avons procédé à un traitement des champs *numberTenders* et *numberTendersSme*. Lorsque le nombre d'offres émises par des PME était supérieur au nombre total d'offres émises (*numberTendersSme* > *numberTenders*), nous avons remplacé la valeur de *numberTendersSme* par celle de *numberTenders*. Cette décision a été prise après avoir constaté que seulement 3,188% des données contenaient à la fois ces deux champs non nuls. Mettre ces valeurs à null aurait entraîné une perte de données, ce qui aurait été probléma-

tique étant donné la faible quantité de données déjà disponibles. Nous avons donc jugé que cette solution était la plus cohérente pour ce cas.

2.4 Analyse descriptive

Variable 1	Variable 2	Corrélation
jointProcurement	multipleCae	0.99
onBehalf	jointProcurement	0.90
onBehalf	multipleCae	0.89
numberTenders	numberTendersSme	0.79
cpv	typeOfContract_U	0.71
topType_NIC	topType_OPE	0.66

Variable 1	Variable 2	Corrélation
jointProcurement	multipleCae	0.99
awardEstimatedPrice	awardPrice	0.97
onBehalf	jointProcurement	0.90
onBehalf	multipleCae	0.89
numberTenders	numberTendersSme	0.87
cpv	typeOfContract_U	0.71
topType_NIC	topType_OPE	0.66

Table 18. Tableaux du calcul de corrélation avant (gauche) et après (droite) le nettoyage

2.4.1 Lots

Section variables unitaires

- **cancelled**

"cancelled" permet de savoir si un si l'appel d'offres du lot a été annulé. Le faible nombre d'annulations (280 soit à peine 0.02%) par rapport au nombre total de contrats est la raison pour laquelle ces cas ne sont pas exclus des autres analyses. Comme ils représentent une petite proportion de l'ensemble des données, ils ne représentent pas un impact significatif sur les tendances générales.

- **awardDate**

Contrairement à avant, l'année 2010 a la hauteur la plus élevée. Cela indique que cette année-là, il y a eu le plus grand nombre de décisions d'attribution comparé aux autres années affichées. La baisse graduelle des décisions d'attribution de lots jusqu'en 2016, pourrait être mise en relation avec [les développements législatifs et réglementaires dans le domaine des marchés publics en France](#). En 2016, plusieurs réformes importantes ont été introduites dans la législation française sur les marchés publics, en réponse aux directives européennes visant à moderniser et à rendre plus efficaces les procédures de passation des marchés publics.

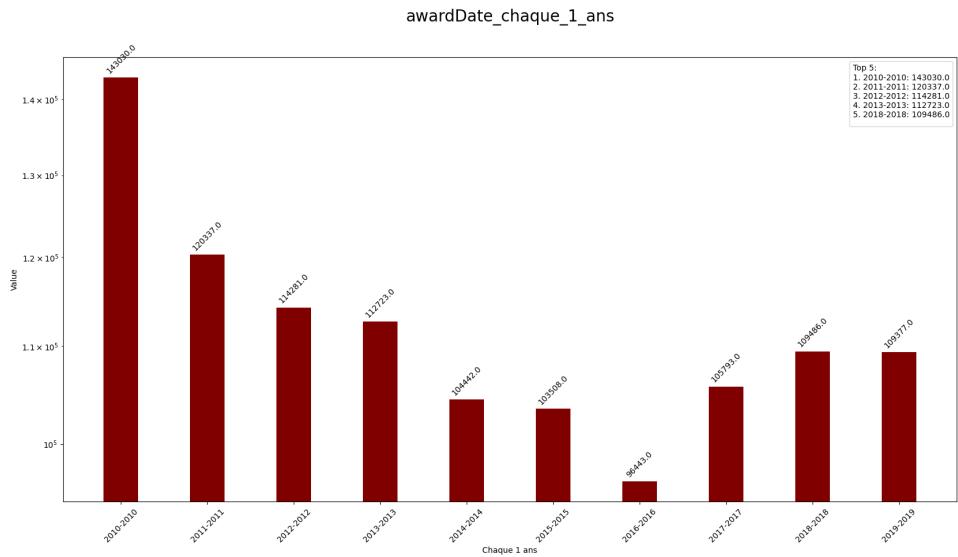


Figure 19. Distribution des awardDate chaque année

La mise en œuvre de ces nouvelles réglementations a pu influencer la manière dont les marchés étaient attribués, avec des objectifs tels que la simplification des procédures, l'amélioration de l'accès pour les PME et l'encouragement de l'innovation. Ces réformes ont pu entraîner une réduction temporaire du nombre de marchés attribués pendant que les entités adjudicatrices s'adaptaient aux nouveaux processus et exigences. De plus, cela a pu également refléter une période d'ajustement où les entreprises apprenaient à naviguer dans le nouveau cadre réglementaire avant de soumettre des offres ou de participer à des appels d'offres.

• awardEstimatedPrice

Les contrats de travaux ont tendance à avoir des estimations de prix plus élevées et plus variées, ce qui peut refléter la complexité et la diversité des projets de construction et de génie civil. Les contrats de services et de fournitures ont des médianes plus basses, mais les valeurs extrêmes hautes dans ces catégories pourraient indiquer des cas particuliers où des biens ou services spécifiques sont extrêmement coûteux.

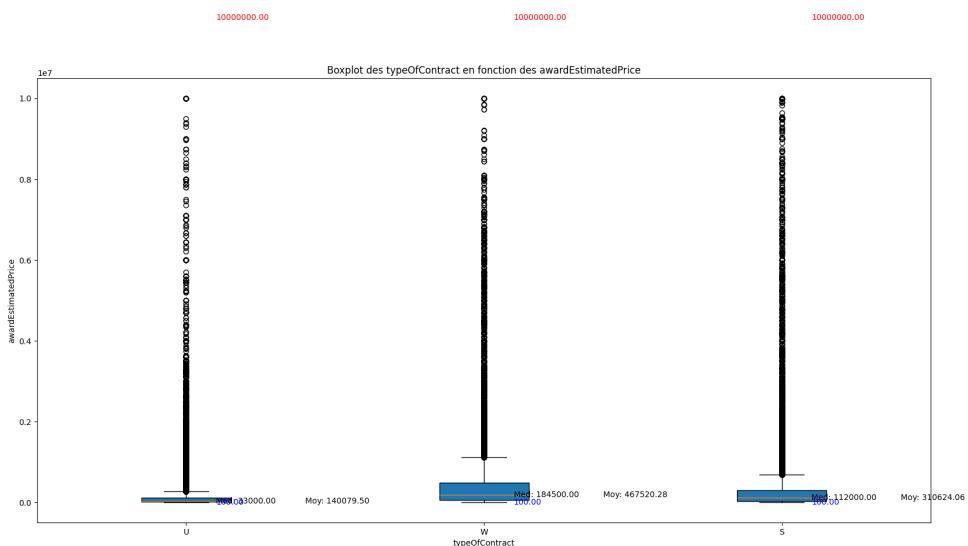


Figure 20. Boxplot des awardEstimatedPrice en fonction des types de contrats

- Services (U) : La médiane est très basse comparée aux autres, et il y a peu de points hors des moustaches, ce qui suggère que la plupart des contrats de services aient des estimations de prix basses à modérées, avec quelques exceptions significatives.
- Travaux (W) : La médiane est la plus élevée que celui des services, indiquant que les prix estimés des travaux sont généralement plus élevés. La présence de nombreux points hors des moustaches indique aussi une grande variabilité avec un certain nombre de lots ayant des estimations très élevées, probablement dû à des projets d'envergure ou très spécialisés.
- Fournitures (S) : La médiane relativement basse, mais avec des valeurs extrêmes suggérant que certains contrats de fournitures ont des estimations de coûts exceptionnellement élevées.

• awardPrice

Dans l'ensemble, ces tendances indiquent que les projets de travaux (W) sont généralement les plus coûteux, suivis par certains contrats de services (U) et de fournitures (S) qui présentent également des coûts élevés, mais avec moins de fréquence.

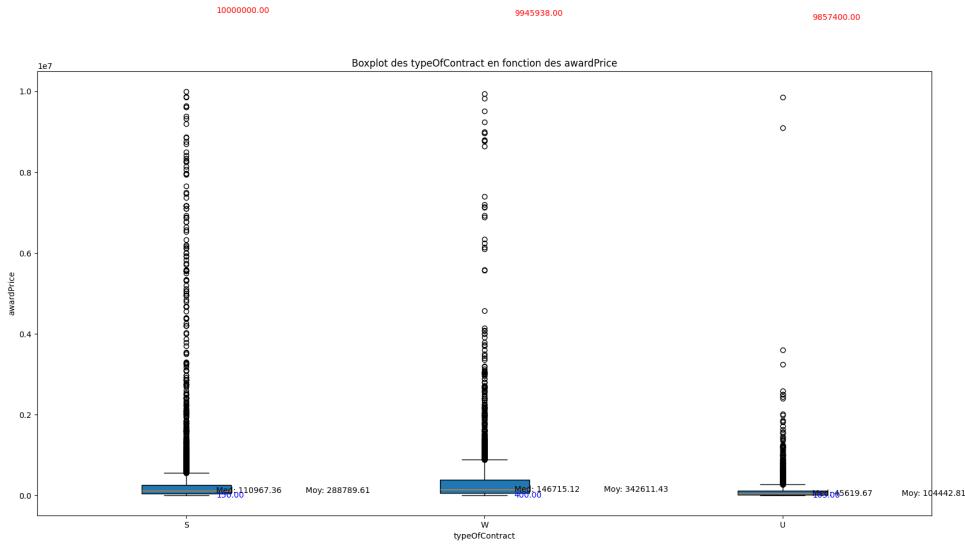


Figure 21. Boxplot des awardPrice en fonction des types de contrats

- Services (U) : La médiane plus basse et moins de points extrêmes par rapport aux travaux, suggérant que les services ont tendance à être moins coûteux que les grands projets de travaux, bien qu'il y ait encore des cas où le prix effectif peut être considérablement élevé.
- Travaux (W) : La médiane est plus élevée que pour les fournitures, ce qui est attendu étant donné que les contrats de travaux englobent souvent des projets de grande envergure. Le nombre élevé de valeurs extrêmes souligne la grande variabilité des coûts associés à ces contrats.
- Fournitures (S) : La médiane est relativement basse, indiquant que pour la majorité des contrats de fournitures, le prix effectif est modeste. Il y a toutefois quelques

valeurs extrêmes indiquant des contrats de fourniture de coût élevé.

- **cpv**

Le **CPV** (Common Procurement Vocabulary) ou Vocabulaire commun pour les marchés est utilisé par l'Union européenne pour les marchés publics. Il s'agit d'un système de classification qui aide à harmoniser les références utilisées par les fournisseurs et les acheteurs pour décrire le sujet des marchés publics.

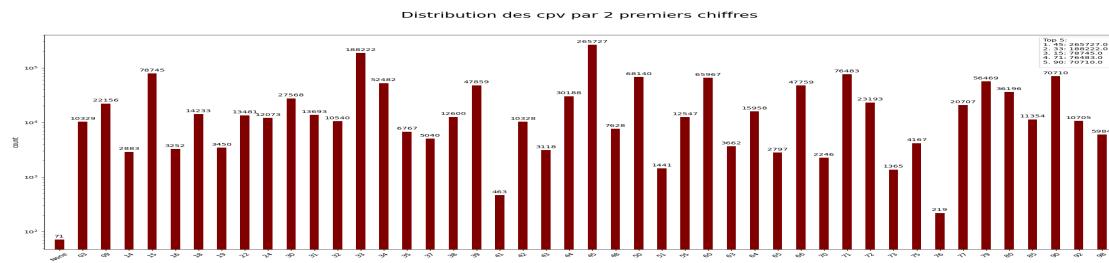


Figure 22. Distribution des CPV par divisions

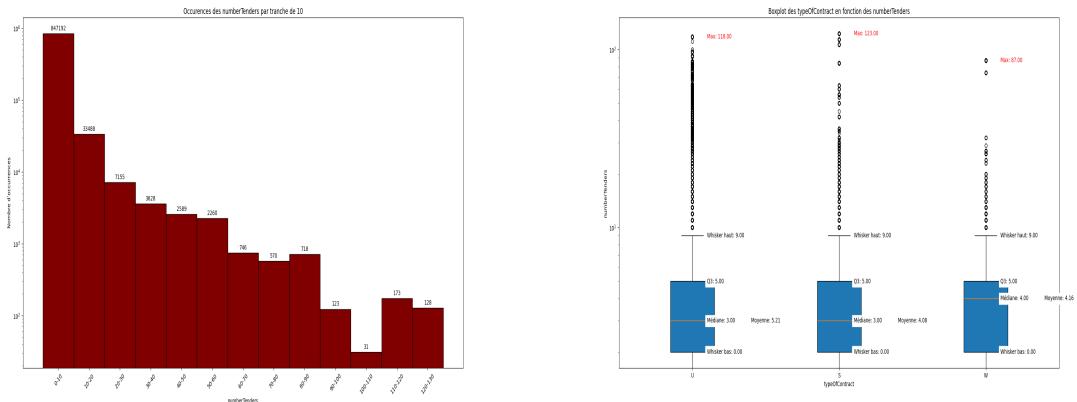
L'histogramme ci-dessus montre la fréquence des codes CPV regroupés par les deux premiers chiffres avec une échelle logarithmique. Certaines catégories CPV sont nettement plus fréquentes que d'autres, comme on peut le voir avec les barres les plus hautes. Ainsi, il est possible d'analyser les catégories de CPV les plus courantes.

CPV	Position	Fréquence
45 2	1	265 727 (22.25%)
33 3	2	188 222 (15.76%)
15 4	3	78 745 (6.59%)
71 5	4	76 483 (6.40%)
90 6	5	70 710 (5.92%)

Table 19. Classement des fréquences de CPV

- **numberTenders**

2. Travaux de construction
3. Matériels médicaux, pharmaceutiques et produits de soins personnels
4. Produits alimentaires, boissons, tabac et produits connexes
5. Services d'architecture, services de construction, services d'ingénierie et services d'inspection
6. Services d'évacuation des eaux usées et d'élimination des déchets, services d'hygiénisation et services relatifs à l'environnement



(a) Distribution des numberTenders par nombre d'occurrences

(b) Distribution des numberTenders par type de contrat

Figure 23. Distribution des numberTenders

L'histogramme montre une fréquence élevée pour la tranche 0-10 et que les fréquences diminuent graduellement pour les tranches supérieures, cela suggère que la plupart des lots reçoivent un nombre relativement faible d'offres, et que moins de lots reçoivent un grand nombre d'offres.

Le cas surprenant est que la fréquence s'élève de nouveau après la tranche 100-110. Cela s'explique avec le diagramme moustache, dont les contrats de services (U), ceux de fournitures (S) et ceux de travaux (W), ont une médiane se situe autour de 3 à 4 offres par lot, indiquant qu'une majorité de lots reçoivent peu d'offres. Cette tendance commune suggère que quel que soit le type de contrat, les marchés sont en général pas être très compétitifs ou que les appels d'offres ne parviennent pas à attirer un grand nombre d'entreprises.

Malgré une tendance générale vers un faible nombre d'offres, certains lots reçoivent un nombre beaucoup plus important d'offres, comme en témoignent les valeurs élevées pour chacun des types de contrat. Cela reflète des lots particulièrement attractifs, soit par leur valeur, soit par leur nature.

- **fraEstimated**

fraEstimated est un champ suggérant l'existence d'un accord-cadre (le cas échéant). Nous pouvons remarquer qu'il n'y a pas beaucoup d'instances pour cette cible, surtout celles qui ont plusieurs points indiquants. Dans quelques cas, des lots sont quand même indiqués en catégorie tout seul pour 'K', 'A' et 'C', mais rarement indiqués à plusieurs endroits et plus clairement. La plupart du temps, dans les cas généraux, ce champ est vide (82,72 % sont nuls).

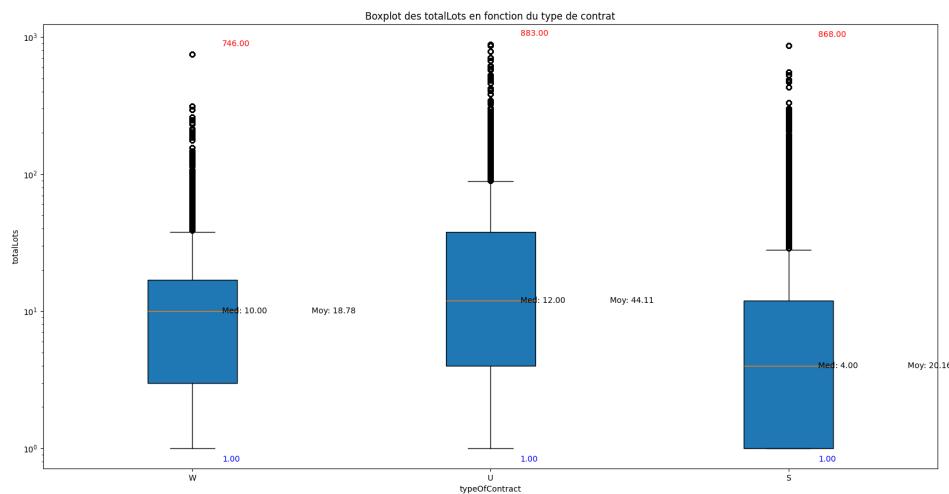
- **totalLots**

Le graphique en boîte à moustaches montre la distribution du nombre total de lots (totalLots) en fonction des types de contrat : travaux (W), services (U) et fournitures (S). Les valeurs minimales sont toutes à 1, ce qui est logique, car chaque avis doit comprendre au moins un lot. La présence de whiskers inférieurs courts pour les trois types de contrats montre que la majorité des avis comprennent un nombre de lots relativement

Group	NbFraEstimated
A	65 544 (4.75%)
AC	8 893 (0.64%)
C	65 021 (4.17%)
K	73 061 (5.29%)
KA	20 307 (1.47%)
KAC	2 521 (0.18%)
KC	3 363 (0.24%)

Table 20. Pourcentage de chaque catégorie de fraEstimated

restreint.

**Figure 24.** Boxplot des totalLots en fonction des types de contrats

La médiane pour les contrats de travaux et de services est respectivement de 10 et 12, ce qui indique que la moitié des avis d'attribution pour ces types de contrats comprennent 10 lots ou moins pour les travaux et 12 lots ou moins pour les services. Pour les contrats de fournitures, la médiane est plus basse, à 4, suggérant que les avis de marchés de fournitures ont tendance à avoir moins de lots en général.

Les moyennes sont plus élevées que les médianes pour les trois types de contrats, ce qui suggère que des valeurs extrêmement élevées de totalLots tirent la moyenne vers le haut. Cela est particulièrement visible avec les avis d'attribution comportant un nombre de lots extrêmement élevé, dépassant 800 lots pour les contrats de travaux et de services. Cela pourrait indiquer quelques marchés particulièrement grands ou segmentés en de nombreux petits lots.

- **numberTendersSme**

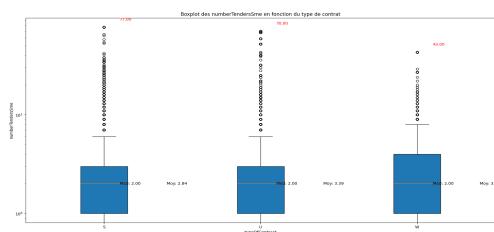


Figure 25. Boxplot des totalLots en fonction des types de contrats

- Après le nettoyage des données, nous constatons que les données n'ont pas beaucoup changé malgré les efforts de nettoyage. La seule différence notable concerne la médiane pour le type de contrat fournitures 'S', qui est passée de 2,85 à 2,84. Par conséquent, nous pouvons effectuer la même analyse descriptive que celle effectuée avant le nettoyage sur ces données.

- **topType**

topType correspond au type de procédure d'attribution utilisée dans les marchés publics, indiquant la méthode selon laquelle un marché est attribué à un ou plusieurs fournisseurs.

Identifiant	Fréquence
OPE	84.39% (1 165 396)
NIC	7.52% (103 848)
AWP	4.71% (65 077)
RES	1.75% (24 152)
NOC	1.28% (17 736)
COD	0.20% (2 799)
NIP	0.08% (1 227)
NOP	0.03% (506)
INP	<0.01% (33)

Table 21. Fréquence des typeOfContract

- **contractDuration**

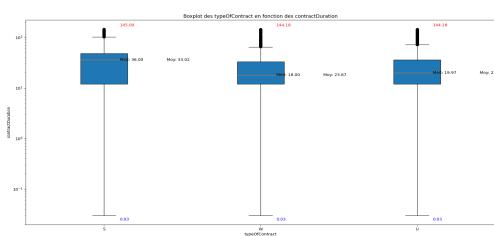


Figure 26. Boxplot des contractDuration en fonction des types de contrats

- Après avoir nettoyé les données, nous constatons une réduction de l'écart entre les différentes valeurs. L'axe vertical n'est plus en échelle logarithmique, ce qui rend les valeurs aberrantes plus cohérentes. Nous observons également une légère diminution de la médiane pour le type de contrat services 'U'. Avec ces ajustements, les mêmes analyses que celles effectuées avant le nettoyage des données peuvent être réalisées sur ce graphique.

- **publicityDuration**

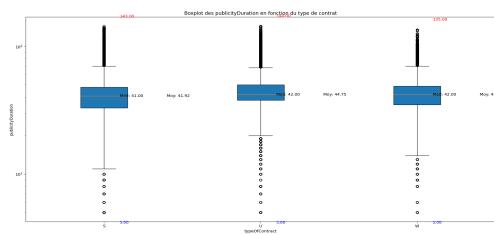


Figure 27. Boxplot des publicityDuration en fonction des types de contrats

- Tout comme le contractDuration après le nettoyage, nous constatons une réduction de l'écart entre les différentes valeurs. Avec ces ajustements, les mêmes analyses que celles effectuées avant le nettoyage des données peuvent être réalisées sur ce graphique.

Section paire de variables

- **awardEstimatedPrice & awardPrice**

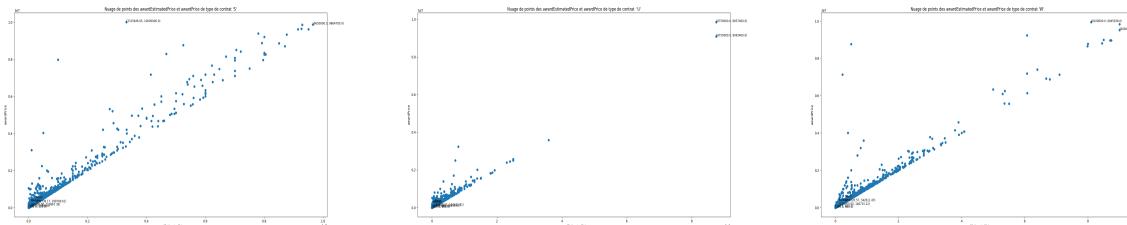


Figure 28. Distribution des awardEstimatedPrice vs awardPrice

Après avoir nettoyé les données, nous remarquons une réduction significative de l'écart entre elles. En effet, les valeurs ont été ramenées d'une échelle de 10^{12} à une échelle de 10^7 sur l'axe des ordonnées, ce qui représente une réduction considérable de l'amplitude des données. Cette diminution de l'écart entre les valeurs permet de mettre en évidence plus clairement la corrélation entre les deux variables étudiées. Cependant, malgré ce nettoyage, nous constatons encore la présence de quelques valeurs aberrantes, notamment dans le cas des contrats de type 'U'. En comparaison avec les autres types de contrats, il apparaît que ceux de type 'U' présentent généralement des appels d'offres avec des prix estimés et des prix effectifs moins importants.

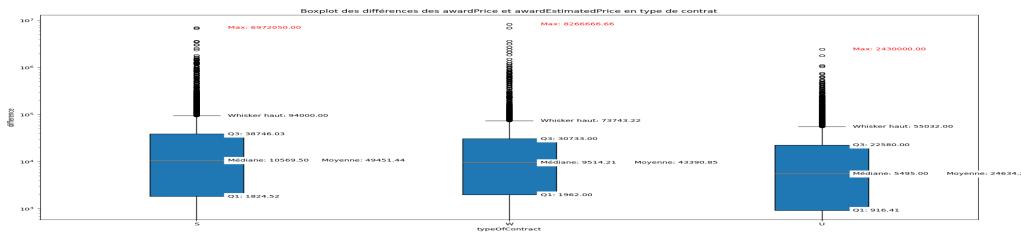


Figure 29. Boxplot des différences entre les awardPrice et les awardEstimatedPrice par type de contrat

Sur cette figure, nous observons également une disparité moins significative concernant la différence entre les deux variables. Les boîtes à moustaches sont mieux visibles, permettant ainsi de remarquer que, pour le type de services 'U', la médiane et la moyenne sont moins élevées que pour les deux autres.

- **awardEstimatedPrice & numberTenders**

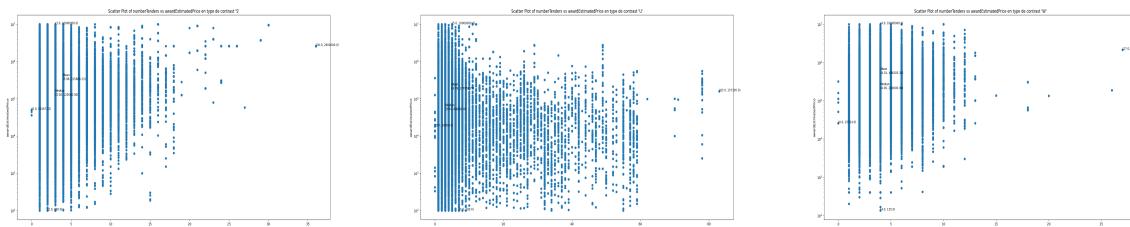


Figure 30. Distribution des awardEstimatedPrice vs numberTenders

Après avoir effectué le nettoyage des données, nous constatons une dispersion moindre des valeurs. Cette clarification nous permet de mieux comprendre comment les offres se répartissent par rapport aux prix estimés selon les différents types de contrats, confirmant ainsi l'absence de corrélation directe entre ces deux variables. En ce qui concerne l'analyse descriptive, nous parvenons aux mêmes conclusions que celles obtenues avant le nettoyage des données.

- **awardEstimatedPrice & accelerated**

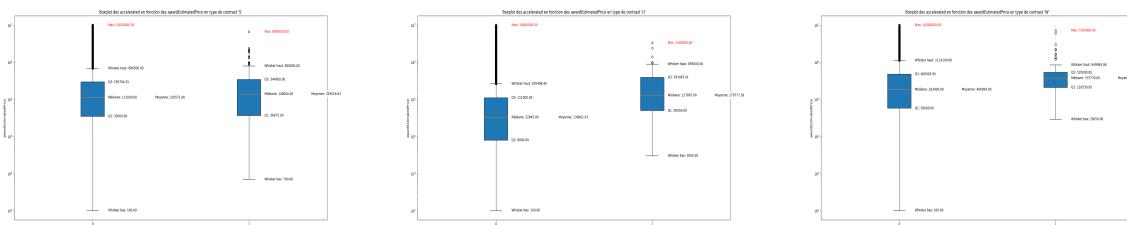


Figure 31. Boîte à moustache des awardEstimatedPrice vs accelerated

Après avoir nettoyé les données en incluant les zéros pour la série "accelerated" et les données "awardEstimatedPrice", nous avons maintenant deux ensembles distincts de données représentés par des boîtes à moustaches pour chaque type de contrat. Dans

les données d'origine, nous avons observé la présence de valeurs extrêmes très élevées pour chaque type de contrat, avec des prix estimés dépassant largement la majorité des données. Désormais, les médianes et les moyennes reflètent mieux l'ensemble des données, car elles ne sont plus affectées par ces valeurs extrêmes. Ce processus rapide n'a pas entraîné d'augmentations significatives des prix estimés.

- **awardEstimatedPrice & numberTendersSme**

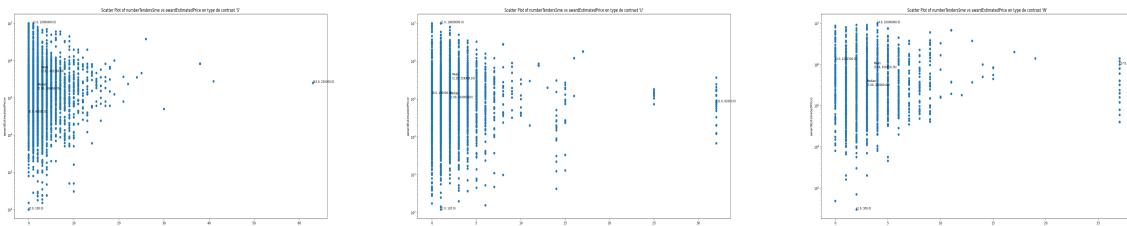


Figure 32. Distribution des awardEstimatedPrice vs numberTendersSme

Après avoir effectué le nettoyage des données, nous constatons une dispersion moindre des valeurs. Cette clarification nous permet de mieux comprendre comment les offres venant des PME se répartissent par rapport aux prix estimés selon les différents types de contrats, confirmant ainsi l'absence de corrélation directe entre ces deux variables. En ce qui concerne l'analyse descriptive, nous parvenons aux mêmes conclusions que celles obtenues avant le nettoyage des données.

- **awardPrice & cpv**

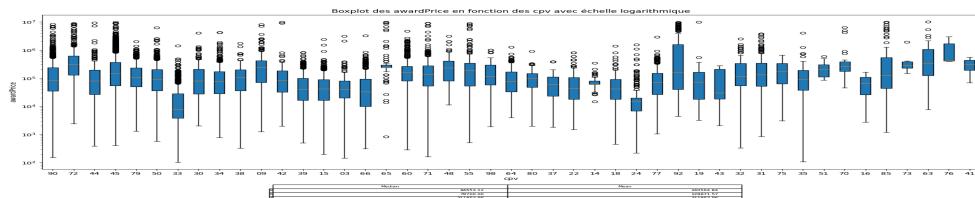


Figure 33. Boîte à moustache des awardPrice vs cpv

Les CPVs avec des whiskers bas élevés et une médiane haute comprennent des services liés à l'industrie du pétrole et du gaz (CPV 76), des services de recherche et de conseil (CPV 73), et l'approvisionnement en eau (CPV 41). Ces domaines ont tendance à impliquer des coûts supérieurs en raison de leur spécialisation, de leur réglementation stricte, ou de leur importance critique. Ils méritent une évaluation approfondie pour assurer une gestion financière adéquate et l'optimisation des dépenses publiques.

D'autre part, des CPVs tels que ceux liés à l'équipement de sécurité (CPV 35), aux services d'architecture et d'ingénierie (CPV 71), et aux services environnementaux (CPV 90) affichent des whiskers bas très bas et une médiane moyenne, suggérant une compétition accrue et des coûts modérés, reflétant peut-être des domaines où des économies d'échelle sont réalisables ou où la standardisation permet des prix compétitifs.

Enfin, les CPVs avec des whiskers bas très bas et une médiane basse, comme les produits chimiques (CPV 24) et les matériels médicaux (CPV 33), indiquent des marchés où les prix sont généralement plus bas et standardisés, ce qui peut signifier des économies pour les achats publics. Ces catégories pourraient bénéficier d'une veille concurrentielle pour maintenir les coûts bas tout en assurant la qualité des produits et services achetés.

Les domaines tels que l'infrastructure, la technologie, la santé et les services environnementaux ont connu différentes intensités de sollicitation, ce qui a modelé le paysage des marchés publics en France entre 2010 et 2020.

- **awardPrice & les variables booléennes**

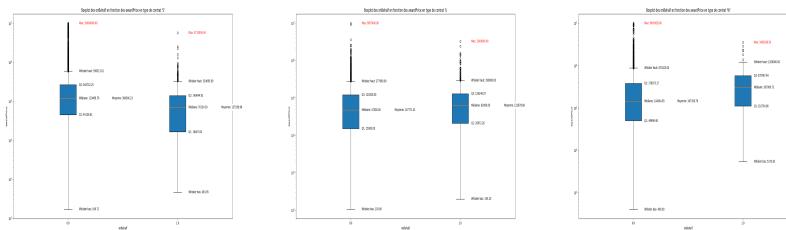


Figure 34. Boîte à moustache des awardPrice vs onBehalf

Dans cette représentation graphique, l'échelle des prix des récompenses (awardPrice) est réduite, ce qui signifie qu'il y a moins de valeurs aberrantes ou incohérentes. Cette tendance est également observée pour toutes les autres variables booléennes liées aux prix effectifs. Malgré ces modifications, les conclusions que nous pouvons tirer à partir de ces données restent les mêmes que celles de la série initiale.

- **awardPrice & topType**

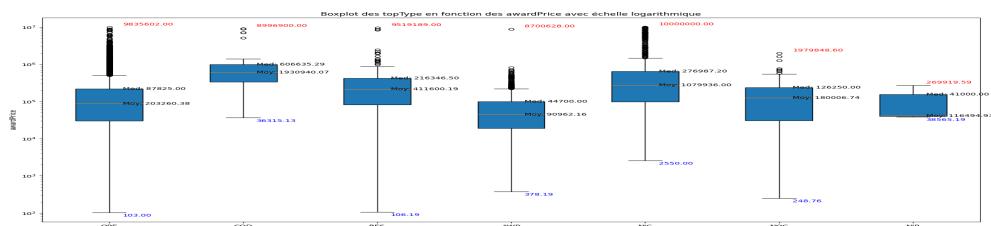


Figure 35. Boîte à moustache des awardPrice vs topType

Les valeurs extrêmes ont été nettoyées, ce qui donne des boîtes à moustaches qui reflètent une distribution des prix plus cohérente et probablement plus précise. Les médianes des différents types de contrats semblent plus uniformes après le nettoyage, ce qui pourrait indiquer que les montants typiques des contrats sont similaires à travers les différents types de procédures. Les moyennes semblent moins influencées par les valeurs aberrantes, suggérant une distribution des prix plus stable et prévisible. Avec des données nettoyées, il est plus facile d'interpréter les tendances sans être faussé par des valeurs anormales, ce qui est essentiel pour l'analyse de fraude ou d'irrégularités.

Nous pouvons également remarquer certaines modifications dans les statistiques. En examinant la boîte à moustaches pour la procédure Dialogue compétitif (COD), nous observons que la moyenne, la médiane et la dispersion de la longueur des moustaches ont considérablement diminué. Cela suggère une réduction de la disparité des données. Il semble que la plupart des prix effectifs pour ce type de procédure soient similaires. Une tendance similaire est également observée pour la procédure Négocié avec un appel à la concurrence (NIP).

- **awardPrice & contractDuration**

Bien que l'on puisse supposer que des contrats plus longs pourraient coûter plus cher en raison d'un engagement prolongé, les données ne montrent pas nécessairement cette tendance. Les décisions d'attribution des prix des lots semblent être basées sur une variété de facteurs autres que la simple durée du contrat.

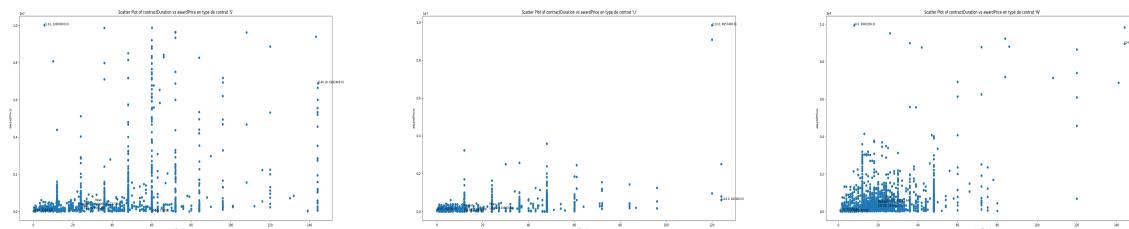


Figure 36. Distribution des awardPrice vs contractDuration

Des exceptions notables sont présentes, surtout dans les contrats de travaux (W), où quelques contrats de longue durée se distinguent par des montants de lots significativement élevés. Cela pourrait indiquer des projets d'envergure nécessitant des investissements conséquents en termes de temps et de ressources financières.

- **awardPrice & publicityDuration**

Il y a une grande dispersion des prix pour toutes les durées de publicité dans les trois types de contrats. Cela suggère qu'il n'y ait pas de corrélation directe et forte entre la durée de la publicité et le prix du lot attribué.

Pour les contrats de services et de fournitures, la plupart des lots ont une durée de publicité relativement courte avec des prix variés. Pour les travaux, les données semblent un peu plus dispersées sur la durée de publicité, mais avec une tendance similaire de prix variés.

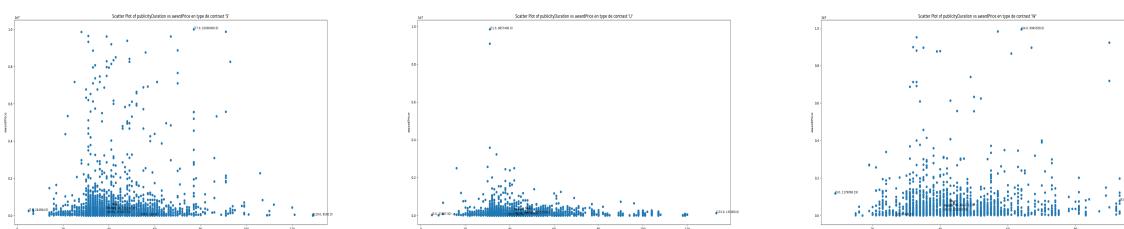


Figure 37. Distribution des awardPrice vs publicityDuration

Aucun modèle clair n'émerge indiquant une tendance linéaire ou exponentielle entre la durée de publicité et le prix du lot. Cela suggère que d'autres facteurs, en dehors de la durée de publicité, ont une influence significative sur le prix final des lots.

En d'autres termes, les lots avec des périodes de publicité plus longues ne sont pas systématiquement associés à des prix d'attribution plus élevés ou plus bas.

- **numberTenders & numberTendersSme**

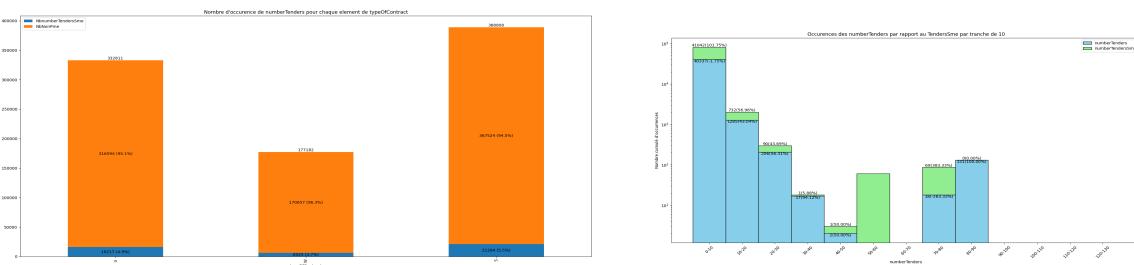


Figure 38. Distribution des numberTenders vs numberTendersSme

Sur le premier graphique, nous voyons une comparaison directe du nombre d'offres issues de PME par rapport au nombre total d'offres pour différents types de contrats. La majorité écrasante des offres ne proviennent pas de PME, ce qui peut indiquer que les grandes entreprises ou les entités autres que les PME sont plus actives dans la soumission d'offres, ou que les PME rencontrent des barrières à l'entrée sur le marché des marchés publics.

Le second graphique illustre la répartition des offres reçues en fonction de leur nombre par lot. Il montre que les tranches avec le plus petit nombre d'offres sont les plus communes. La distinction entre les offres issues de PME et les autres offres n'est pas immédiatement apparente ici, mais on peut observer que les lots recevant peu d'offres semblent avoir une plus grande proportion d'offres de PME. Ceci pourrait suggérer que les PME ont tendance à participer à des appels d'offres moins compétitifs ou à des marchés de taille plus petite.

- **numberTenders & topType**

La comparaison entre différents types de procédures d'attribution des marchés publics et le nombre d'offres reçues pour chaque procédure permettent d'évaluer la compétitivité et l'accessibilité de chaque type de procédure.

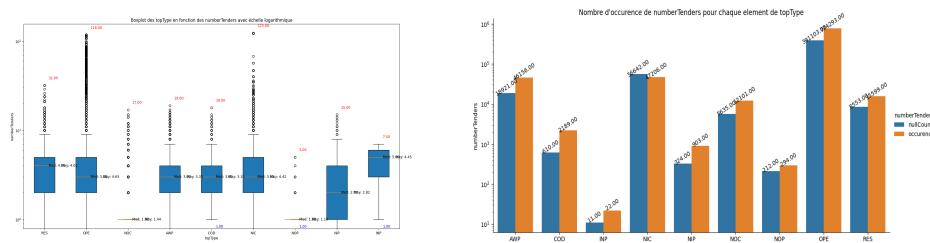


Figure 39. Distribution des numberTenders vs topType

En ce qui concerne le nombre d'offres, les procédures ouvertes (OPE) et restreintes (RES)

reçoivent le plus grand nombre d'offres, ce qui suggère une forte compétitivité et un large intérêt de la part des fournisseurs.

D'autre part, les procédures négociées sans publication préalable (AWP, NOC/NOP) et les partenariats innovants (INP) reçoivent moins d'offres, ce qui indique une compétition plus faible. Ces procédures sont susceptibles d'être utilisées dans des circonstances spéciales où la rapidité ou l'expertise spécialisée est requise, limitant potentiellement le nombre de participants éligibles ou intéressés.

Les dialogues compétitifs (COD) et les procédures négociées avec publication (NIC/NIP) se situent dans une gamme intermédiaire en termes de nombre d'offres, ce qui peut refléter un équilibre entre accessibilité et spécificité des exigences du marché.

2.4.2 Criteria

- **type** Chaque lot est associé à des critères d'attribution répartis en six catégories : le prix, la technique, le délai, d'autres critères, l'impact environnemental et l'impact social. Dans notre base de données, chaque critère est systématiquement classé dans l'une de ces catégories, ce qui signifie qu'il n'y a aucune valeur manquante pour cette variable.

Catégorie	Valeur
TECHNICAL	1098843
PRICE	1086700
OTHER	384431
DELAY	158923
ENVIRONMENTAL	152911
SOCIAL	28600

Table 22. Distribution des type par tranche de 10 en ordre décroissant

Lorsque nous analysons le tableau de distribution des catégories, nous notons que :

- Catégories Techniques et Prix : Ces deux catégories affichent les plus grands nombres. Cette prédominance indique que les critères techniques et de coût sont les plus fréquemment pris en compte dans les appels d'offres, soulignant leur importance cruciale dans la décision d'attribution des marchés publics.
- Catégories Environnementales et Sociales : Ces catégories montrent le nombre le plus faible, signalant quelles sont bien moins prioritaires comparées aux critères techniques et de prix. Ce constat pourrait être une opportunité pour les entreprises de se différencier en mettant en avant leurs performances dans ces domaines.

• weight

Pour chaque lot, il y a un critère d'attribution avec une importance relative par rapport aux autres lots dans le processus d'attribution. Cette importance est représentée par la variable *weight* (poids en français). Nous avons la totalité des poids pour chaque critère, soit aucune valeur vide pour cette variable.

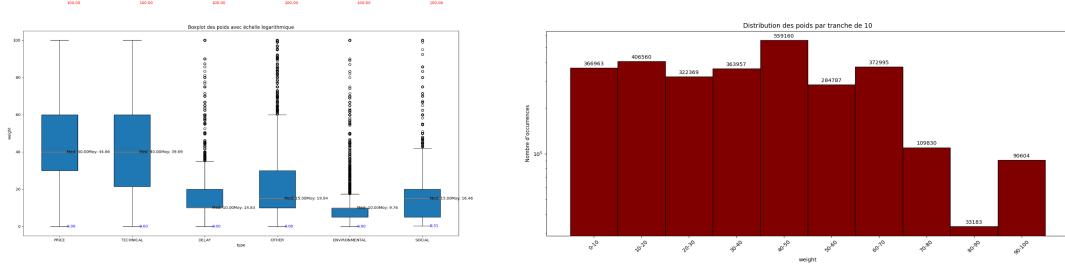


Figure 40. Distribution des weight

La figure de gauche illustre la distribution des poids attribués aux critères de sélection des appels d'offres dans six catégories : prix, technique, délai, autres, environnemental, social avec une subdivision en environnemental et social.

Les points notables dans ce graphique sont comme suit :

- Les médianes varient, montrant que les critères techniques et de prix sont généralement considérés comme les plus importants (médianes respectivement à 44.66 et 39.69), suivis par les critères de délai, environnementaux et sociaux (médianes respectivement à 14.83, 19.94 et 16.46). Cela suggère une tendance à valoriser la technicité et le coût par rapport à l'impact social ou environnemental.
- Les distributions des poids dans les catégories techniques et de prix sont également plus étendues, indiquant une variation significative dans l'importance accordée à ces critères entre différents appels d'offres. En revanche, la catégorie 'délai' montre une concentration des poids autour de valeurs plus basses, ce qui peut signifier que les délais sont un facteur de considération standard et moins sujet à variation dans l'importance relative avec beaucoup d'exceptions pour des situations extrêmes.
- Les catégories 'environnemental' et 'social' présentent une distribution plus resserrée avec de petites médianes, mais avec des valeurs aberrantes significatives. Ceci peut indiquer que, bien que ces critères ne soient généralement pas les plus valorisés, quelques rares appels d'offres les considèrent comme essentiels, voir même extrêmement important.

L'image sur la droite démontre la distribution des poids par tranche de 10. Les points clés de ce graphique-ci sont :

- Les poids inférieurs à 50 ont un nombre d'occurrences significatif, ce qui indique que dans la plupart des appels d'offres, aucun critère unique ne domine le processus d'évaluation.
- Les poids supérieurs à 60 ont un nombre d'occurrences plus faible, ce qui pourrait indiquer que les critères avec de tels poids sont très rares et peuvent signaler une procédure d'appel d'offres très spécifique.

2.4.3 Agents

• Siret

Le numéro **SIRET** est composé de 14 chiffres et est structuré de la manière suivante :

- Les 9 premiers chiffres constituent le numéro SIRENE (Système d'Identification du Répertoire des Entreprises), qui identifie l'entreprise elle-même.
- Les 5 chiffres suivants, appelés NIC (Numéro Interne de Classement), identifient

l'établissement spécifique de l'entreprise.

En base de données, nous observons 24% de valeurs vides pour cette variable.

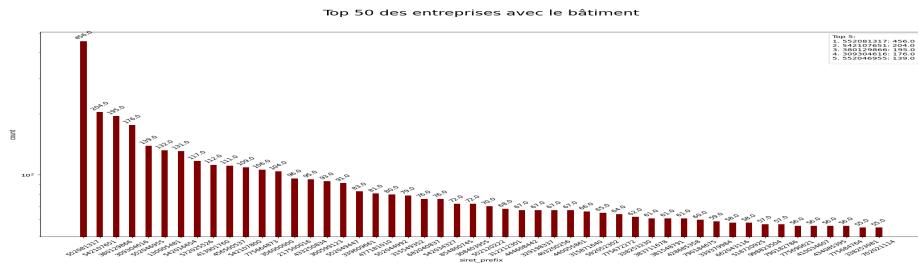


Figure 41. Distribution des siren

L'image illustre comment les 50 entreprises les plus fréquentes sont réparties en termes de nombre d'agents. L'axe vertical utilise une échelle logarithmique pour représenter ces nombres d'agents, ce qui permet de gérer la grande disparité entre les comptages. Quant à l'axe horizontal, il indique les numéros de SIRENE des entreprises. Cette approche est choisie pour identifier directement les entreprises plutôt que des établissements spécifiques au sein de celles-ci.

Ici nous observons plusieurs point clés :

- L'entité avec le plus grand nombre d'agent est le 552081317, soit, ELECTRICITE DE FRANCE comptant 456 agents économiques. Ceci est une valeur aberrante certes, ce qui peut indiquer un monopole ou un avantage concurrentiel significatif. Toutefois la raison est que EDF une entreprise d'énergie intégrée et l'un des plus grands producteurs d'électricité au monde. En tant qu'une des plus grandes entreprises françaises, elle emploie un grand nombre de personnes dans divers domaines, notamment la production d'électricité, la distribution, la commercialisation, la recherche et le développement. Elle touche un grand nombre de domaine et avoir un tel nombre d'agent est normal pour une si grande entreprise.
- Les entreprises de ce top 50 compte, les grandes entreprises d'énergie, les entreprises de communication tel que Orange qui est troisième dans le classement avec 195 agents économiques, les assurances, la construction, le transport, etc.

Lors de l'analyse de cette variable, il n'y a pas de constatation particulièrement aberrante concernant le nombre d'agents économiques par numéro de SIRENE.

• department

Les départements[[frwiki:212672035](#)] mentionnés dans ce document correspondent exclusivement à des départements français. Comme nous avons de nombreux agents économiques qui sont basés à l'étranger, il est normal d'avoir un grand pourcentage, soit 30.28% de valeurs vides pour cette variable. Sur la base de cette information, nous avons élaboré plusieurs diagrammes dans le but d'analyser leur répartition au sein de la table **Agents**.

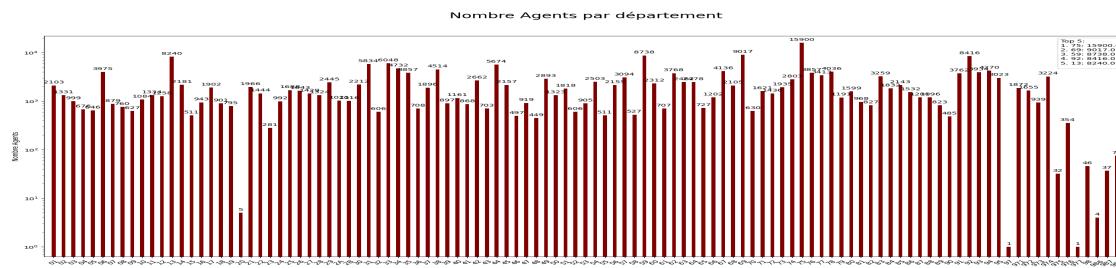


Figure 42. Distribution des department

Le graphique ci-dessus illustre la répartition des agents économiques par département. L'axe des ordonnées est représenté sur une échelle logarithmique, il représente le nombre d'agents. L'axe des abscisses quant à lui représente le numéro du département.

Plusieurs points clés émergent de cette visualisation :

- Le département ayant le plus d'agents économiques est le 75, soit, Paris. Il contient 1.7 fois plus d'agents économiques que le reste du top 5. Toutefois, il est logique de s'attendre à ce que Paris ait un nombre élevé d'acteurs économiques participant aux marchés publics, en raison de sa concentration de population, de ses entreprises et de son importance économique en tant que capitale et centre d'affaires majeur en France.
- Nous observons que le département numéro 20 ne compte que 5 agents économiques, ce qui semble aberrant étant donné qu'il s'agit de la Corse, une région et un département. La Corse se compose de deux départements : la Haute-Corse (2B), qui compte 1016 agents économiques, et la Corse-du-Sud (2A), qui en compte 1021. Les deux départements affichent des chiffres assez élevés en termes d'agents économiques, ce qui semble être dans la norme. Ainsi, pour évaluer le nombre total d'agents économiques dans le département 20, il convient de prendre en compte à la fois le 2A et le 2B, ce qui nous donne au final un nombre d'agents économiques compréhensible.
- Les départements avec le moins d'agents économiques sont les départements d'outre-mer (DOM) de France, représentés par le code 97, avec seulement 1 agent. Étant donné que nous considérons également les départements commençant par 97, il est normal d'en tenir compte. De plus, le code 977 est attribué à un territoire français d'outre-mer appelé Saint-Barthélemy, où il n'y a également qu'un seul agent économique. Saint-Barthélemy est une île située dans les Caraïbes, faisant partie des Antilles françaises. Elle bénéficie d'un statut particulier en tant que collectivité d'outre-mer (COM). Enfin, le code 986 correspond à Wallis-et-Futuna, un territoire d'outre-mer de la France situé dans le Pacifique Sud, où l'on dénombre 4 agents économiques. En raison de la petite taille et de la population réduite des codes 977 et 986, il est probable que le nombre d'agents économiques soit moins élevé.

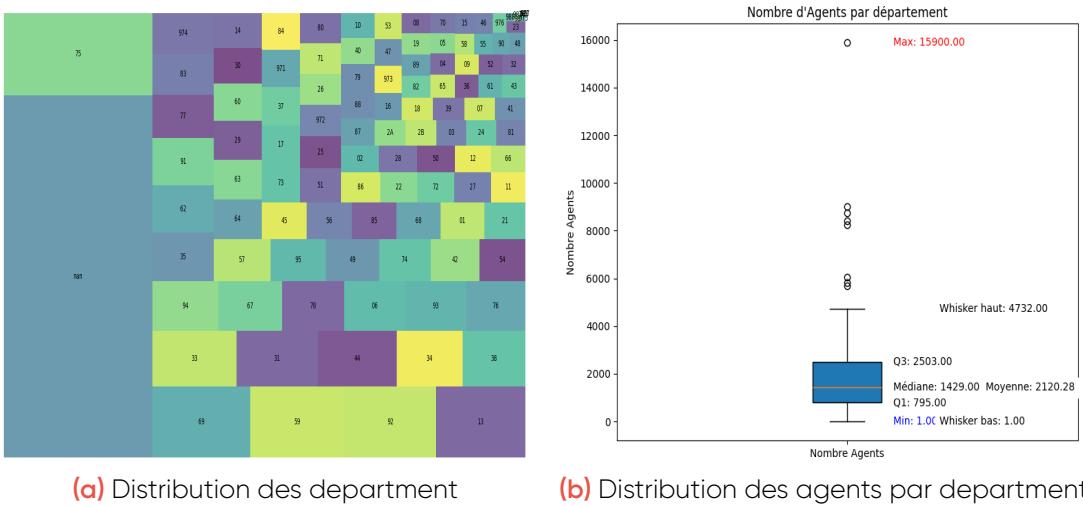


Figure 43. Distribution des department

Concernant la distribution des agents par département avec un diagramme de boîte à moustache. Les observations que nous pouvons ajouter grâce à ce diagramme est :

- L'analyse des quartiles dénote une médiane de 1429 agents, avec un premier quartile à 795 et un troisième quartile à 2503, ce qui indique que 50% des départements ont entre 795 et 2503 agents. La valeur moyenne, située à 2120.28, est supérieure à la médiane, ce qui suggère une distribution asymétrique avec une queue longue à droite.
- Nous observons que Paris est considéré comme un cas extrême, ce qui pourrait être expliqué par la raison donnée précédemment.

Lors de l'analyse de cette variable, il n'y a pas de constatation particulièrement aberrante concernant le nombre d'agents économiques par département. Les quelques valeurs hors normes que nous observons ont une explication légitime.

3 Questionnements

3.1 Les flux de communication mesurés en termes de nombre d'échanges et en termes d'argent

- **Méthodes**

Pour répondre aux questions suivantes, nous avons fusionné les tableaux de bases de données Lots, LotBuyers, LotSuppliers, Agents. Ensuite, nous avons regroupé les mêmes éléments dans leurs groupes respectifs en fonction des types de questions demandées (groupe des acheteurs, groupe des fournisseurs, groupe des départements, groupe des communications en paire avec le même acheteur et fournisseur, groupe des codes CPV), en comptant le nombre de fois qu'ils apparaissent dans les tableaux dans le but de compter en termes d'échanges. Pour ce qui est de l'argent, nous avons calculé la somme des awardPrice pour tous les flux d'échange pour chaque objet. Enfin, nous les trions par ordre décroissant, et pour les résultats de chaque question, nous conservons les 50 meilleurs résultats pour chaque théme.

Les listes obtenues des 50 meilleures entreprises dans les résultats ont été récupérées avec des informations plus détaillées grâce à l'API SIRENE⁷. Elles ont été restructurées et

7. https://api.gouv.fr/les-api/sirene_v3

sauvegardées sous forme de fichiers CSV qui se trouvent dans le répertoire /Output/top50 du projet.

3.1.1 Quelles entreprises communiquent le plus avec quelles autres entreprises ?

• Résultats

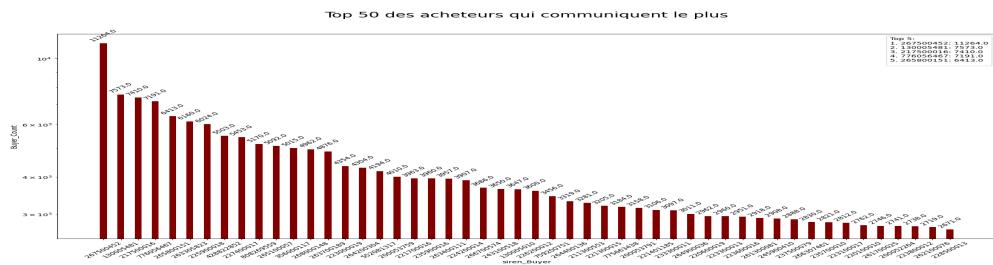


Figure 44. Les top 50 des acheteurs qui communiquent le plus

Fichier csv qui contient les informations des entreprises plus détaillées : [résultat.csv](#)

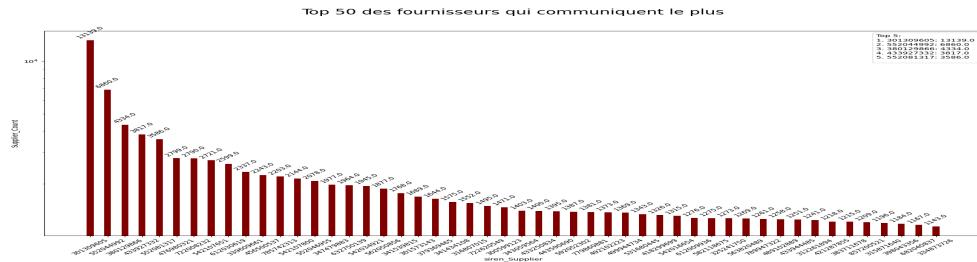


Figure 45. Les top 50 des fournisseurs qui communiquent le plus

Fichier csv qui contient les informations des entreprises plus détaillées : [rезультат.csv](#)

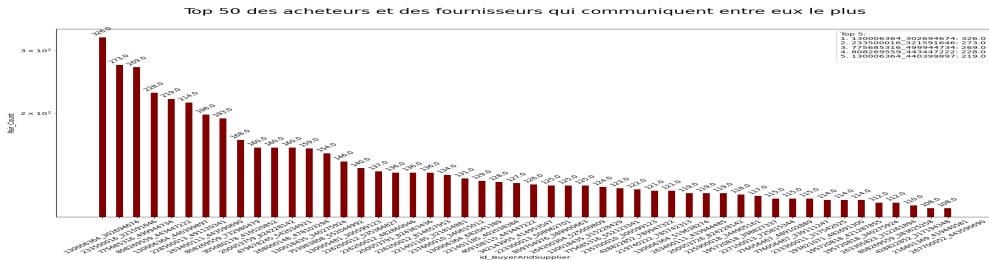


Figure 46. Les top 50 des acheteurs et des fournisseurs qui communiquent entre eux le plus

Fichier csv qui contient les informations des entreprises plus détaillées : [résultat.csv](#)

Le résultat des top 50 acheteurs (3.1.1) et des fournisseurs (4.4) contient les informations des entreprises pendant les années de 2010 à 2020. Les acheteurs les plus actifs sont souvent des organismes publics (il y a 27 entreprises actives dans "Administration publique générale", suivi de "Activités hospitalières" pour 15 entreprises), ce qui nous permet de comprendre que ces entreprises publiques fournissant des services publics sont souvent de grands acheteurs actifs. Quant aux fournisseurs les plus actifs, ils sont souvent plus diversifiés. Cela montre que sur le marché public, nous demandons souvent une diversité de services. Parmi toutes les activités les plus actives, des services très demandés sur le marché public incluent "assurance", "Commerce de gros (commerce interentreprises) alimentaire non spécialisé",

"Analyses, essais et inspections techniques", des services d'installation, etc.

Dans les communications en paires des acheteurs et des fournisseurs (45. Dans le fichier CSV résultant, chaque paire de lignes correspond à un couple formé d'un acheteur et d'un fournisseur. L'acheteur correspond à la ligne au-dessus du fournisseur.), on constate que de nombreux fournisseurs proposent des services dans le domaine de la "Formation continue d'adultes", et leurs acheteurs sont souvent des régions (gouvernements) ou des organismes publics tels que "Pôle Emploi". Nous pouvons observer que le gouvernement réalise de nombreuses transactions et activités pour acheter et redistribuer des services visant à réinsérer les individus sur le marché du travail. Outre les formations, le fichier de résultats indique également que les gouvernements investissent souvent dans les voyages et le transport. Nous pouvons également remarquer que des entreprises fournissant des activités dans le domaine alimentaire ("Commerce de gros (commerce interentreprises) alimentaire spécialisé divers") apparaissent fréquemment, et que les activités de leurs acheteurs sont variées ("Administration publique (tutelle) des activités économiques", "Enseignement de disciplines sportives et d'activités de loisirs", "Administration publique générale", "Enseignement secondaire technique ou professionnel"). Cela indique que le secteur alimentaire conserve souvent une grande part de marché et qu'il compte souvent des clients variés.

Les entreprises sont souvent créées il y a environ 30 ans en moyenne. Toutes ces entreprises sont cotées en bourse.

Il y a seulement une entreprise (siren : 552081317, "ELECTRICITE DE FRANCE") qui apparaît à la fois dans la liste des 50 acheteurs les plus actifs et dans celle des 50 fournisseurs les plus actifs. Cela semble cohérent, étant donné qu'EDF est une très grande entreprise d'énergie en France avec des opérations diversifiées. Elle peut être à la fois un acheteur majeur de divers biens et services nécessaires à ses activités, ainsi qu'un fournisseur important de produits ou services liés à l'énergie.

3.1.2 Quelles entreprises ont le plus de flux monétaire sortants/entrants ? Y a-t-il des PME parmi elles ?

- Résultats

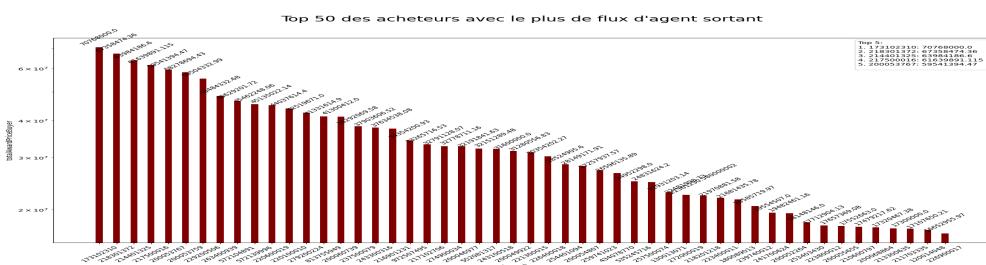


Figure 47. Les top 50 des acheteurs avec le plus de flux d'argent sortant

Fichier csv qui contient les informations des entreprises plus détaillées : [résultat.csv](#)



Figure 48. Les top 50 des fournisseurs avec le plus de flux d'argent entrant

Fichier csv qui contient les informations des entreprises plus détaillées : [résultat.csv](#)

Les deux fichiers présentent les 50 principaux acheteurs ayant les flux monétaires sortants les plus élevés, ainsi que les 50 principaux fournisseurs avec les flux entrants les plus importants. Ces fichiers fournissent également des détails supplémentaires sur ces agents, obtenus via l'API SIRENE, notamment leur catégorie d'entreprise, leur activité principale, leur effectif d'unité légale, etc.

En examinant à la fois le graphique et le fichier CSV, plusieurs points clés ont été observés pour cette question :

- Acheteurs (Top 50 des acheteurs avec le plus de flux d'argent sortant) :
 - Les entités répertoriées sont principalement des administrations publiques, comme indiqué par leurs activités principales 'Administration publique générale', nous avons 34 observations pour cette activité, soit 68% du top 50. La présence dominante d'entités publiques dans la liste des acheteurs reflète la nature des marchés publics où des fonds substantiels sont alloués à des projets gouvernementaux ou publics.
 - Les tailles des entreprises listées s'étendent également, avec une présence notable d'entreprises intermédiaires (ETI) et de grandes entreprises (GE), avec plus de 74% du top 50%. Nous avons 24% de PME.
 - Nous avons observé plusieurs cas intéressants dans notre classement. Par exemple, le Centre Communal d'Action Sociale (263400939) se classe en 9ème position malgré sa petite taille. Il s'agit d'un établissement public, dirigé par un conseil d'administration présidé par le maire de la commune. Cette organisation apporte un soutien essentiel aux personnes âgées, aux personnes handicapées, aux enfants et aux personnes en difficulté. Elle gère divers services sociaux et médico-sociaux, fournit une assistance alimentaire et un hébergement d'urgence, ainsi que des programmes d'intégration et de soutien à l'emploi. Son implication dans de grands projets sociaux et de rénovation d'infrastructures peut expliquer sa position dans le top 50 des communes.
 - Un autre exemple intrigant est celui des Résidences du Quercy Blanc (200060739), classées 16ème avec 37 millions d'euros d'achats. Il s'agit d'établissements médicalisés pour personnes âgées. Ce chiffre d'achat semble inhabituellement élevé pour seulement deux établissements de notre base de données. Il est possible qu'ils achètent une quantité importante de médicaments et de services hospitaliers, mais cela reste tout de même surprenant.
 - Enfin, nous avons la Société Publique Locale Delta 3 (434078770), une petite

entreprise spécialisée dans l'ingénierie et les études techniques, avec seulement 6 à 9 employés en 2021. Malgré sa taille modeste, elle a réalisé des achats de plus de 24 millions d'euros. Cette disparité pourrait être un indicateur de fraude.

- Fournisseurs (Top 50 des fournisseurs avec le plus de flux d'agent entrant) :
 - Les fournisseurs comprennent des entreprises de divers secteurs, tels que la gestion d'installations sportives, la production d'électricité, et le commerce inter-entreprises.
 - Les tailles des entreprises listées s'étendent également, avec une présence notable d'entreprises intermédiaires (ETI) et de grandes entreprises (GE), avec plus de 76% du top 50%. Nous avons 18% de PME.
 - La diversité des entreprises et la présence de grandes entreprises sont cohérentes avec la tendance des marchés publics à attribuer des contrats de grande valeur à des entités bien établies ayant la capacité de gérer de grands projets.
 - Une entreprise qui se démarque est la société 532010576 (SELIA) avec seulement 2 employés recensés. Bien qu'elle soit active dans la production d'électricité, un secteur très demandé, le faible effectif suscite des doutes quant à sa position dans le top 50. Cette situation pourrait indiquer une possible fraude.
 - Une autre entreprise, également avec 2 employés, est présente dans le top 50 : la société 732014964 (FONCIER CONSEIL - SOCIETE EN NOM COLLECTIF, FONCIER CONSEIL SNC, FONCIER CONSEIL SOC EN NOM COLLECTIF). Spécialisée dans la construction d'autres ouvrages de génie civil, elle a été fondée le 1er janvier 1973, ce qui témoigne d'une expertise significative dans ce domaine. Toutefois, le nombre d'employés répertoriés pourrait être inexact.

3.1.3 D'où provient l'argent qui arrive/sort en plus grande quantité ? (en termes de départements français)

- **Résultats**

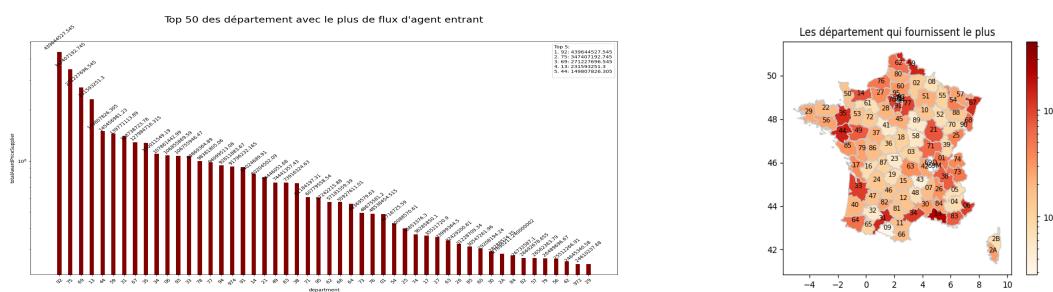


Figure 49. Les départements avec le plus de flux d'argent entrants

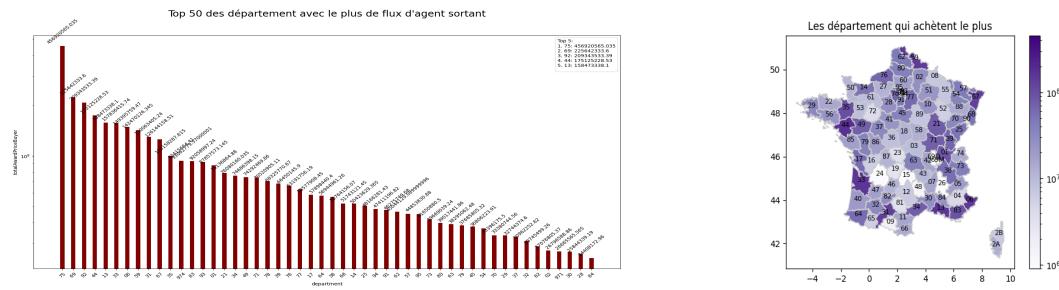


Figure 50. Les départements avec le plus de flux d'argent sortants

Les résultats ci-dessus nous montrent les flux d'argent entrant et sortant par département, on peut remarquer une certaine corrélation entre les deux.

En effet, leurs ordres varient peu et les deux graphiques sont très similaires.

Ainsi, le top 5 contient les mêmes départements seulement leurs ordres diffèrent (argent entrant : 92 75 69 13 44, argent sortant : 75 69 92 44 13).

Les départements en question sont :

- **92 Haut de seine dont la ville principale est Nanterre et contient le quartier de la Défense**
- **75 Paris capitale de la France**
- **69 Rhône avec Lyon en ville principale**
- **13 Bouche du rhône contenant Marseille**
- **44 Loire-Atlantique avec Nantes**

On peut donc remarquer que les flux d'argent se concentrent principalement autour des grandes métropoles françaises.

Aussi en comparant avec les départements de localisation des plus grands acheteurs et fournisseurs avec les plus gros flux d'argent, on peut remarquer une corrélation faible.

En effet certaines de ces entreprises se trouvent bien des les départements remarqués précédemment comme **Suez** dans le 92 et **EDF** dans le 75 mais aussi que d'autres entreprises sont présentent dans des départements bas dans les Top 50 avec **ACTION DEVELOPPEMENT LOISIR** dans le 14 et **Sopra Steria** dans le 74.

3.1.4 Quelles sont les activités les plus actives ou les plus lucratives en termes de code CPV ?

Les activités les plus actives en termes d'échanges entre acheteurs et fournisseurs, illustrées par le flux des CPVs, montrent les secteurs où il y a une forte interaction et un volume important de transactions. Les codes CPV avec le plus de flux peuvent indiquer des marchés avec une forte demande ou une concurrence élevée, où de nombreux fournisseurs.

Quant aux activités les plus lucratives, elles sont représentées par les codes CPV qui génèrent les revenus les plus élevés pour les fournisseurs. Ces secteurs sont généralement ceux où les contrats sont de grande valeur, parfois en raison de la complexité ou de la spécialisation des biens et services demandés.

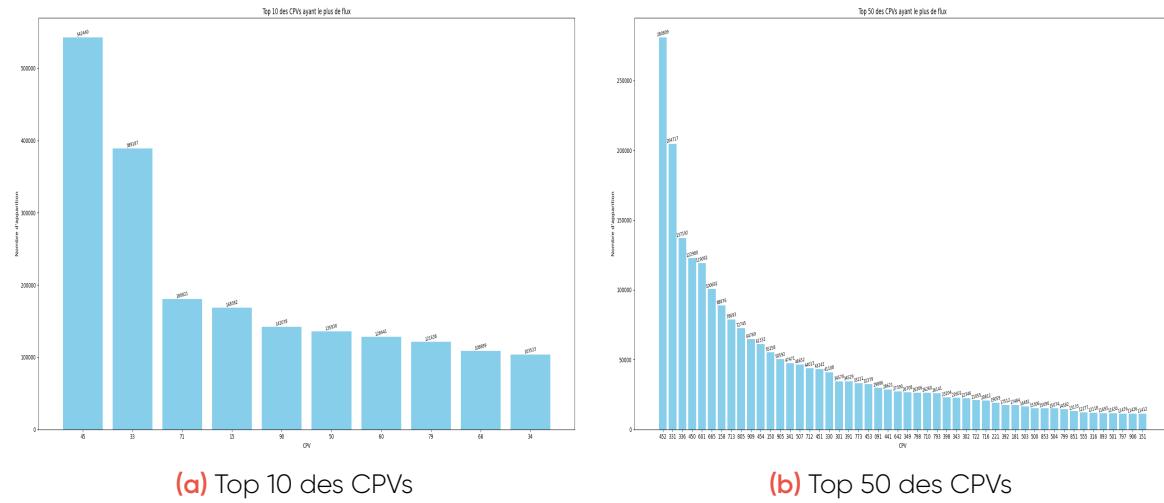


Figure 51. Top des flux d'échanges entre acheteurs et fournisseurs

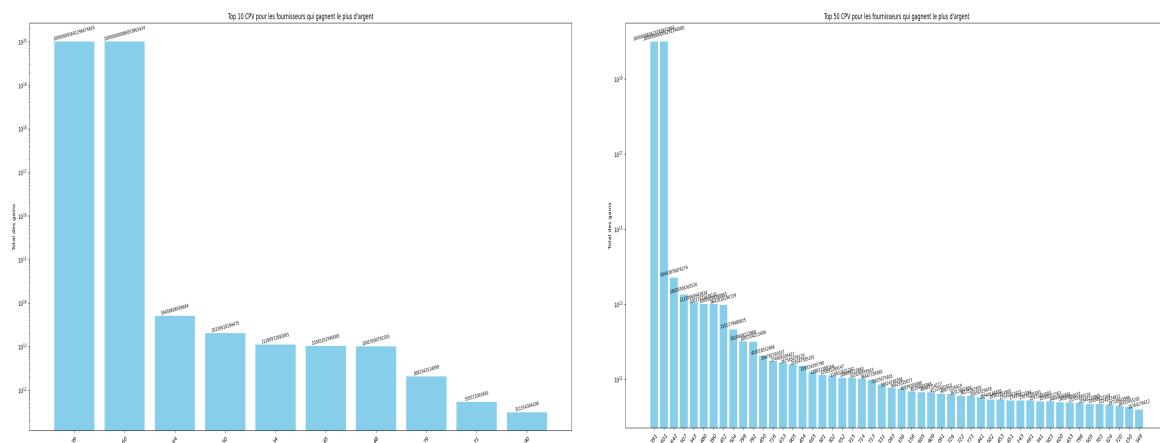
En observant les flux de CPVs, les activités les plus actives semblent être celles où il y a un nombre significatif d'échanges entre les acheteurs et les fournisseurs.

Les codes CPV qui dominent les échanges et les revenus dans les marchés publics révèlent des tendances et des priorités dans les dépenses gouvernementales :

- Construction (CPV 45, 452, 450) : Ces catégories dominent clairement les autres et montre l'importance des infrastructures et des travaux de construction comme moteur économique principal. Cela inclut la construction de bâtiments publics, d'infrastructures routières, et d'autres travaux de génie civil.
- Santé (CPV 33, 331, 336) : Le secteur médical et pharmaceutique est également un des CPVs dominant. Cela traduit la priorité donnée à la santé publique. Cela comprend l'achat d'équipements médicaux et de produits pharmaceutiques, probablement stimulé par l'évolution démographique et les impératifs de santé, comme la gestion des crises sanitaires.
- Services professionnels (CPV 71) : Les services d'architecture, d'ingénierie et d'inspection sont en demande constante, indiquant une nécessité d'expertise technique et de supervision dans la mise en œuvre des projets publics.

Ces tendances indiquent que les gouvernements se concentrent sur la construction et la rénovation d'infrastructures, la santé publique, l'environnement et la prestation de services essentiels.

Maintenant, concentrerons-nous sur les tendances des codes CPV pour les fournisseurs qui génèrent le plus d'argent. Ces derniers mettent en évidence la valeur élevée de certains secteurs d'activité dans les marchés publics.



(a) Distribution des numberTenders par nombre d'occurrences

(b) Distribution des numberTenders par type de contrat

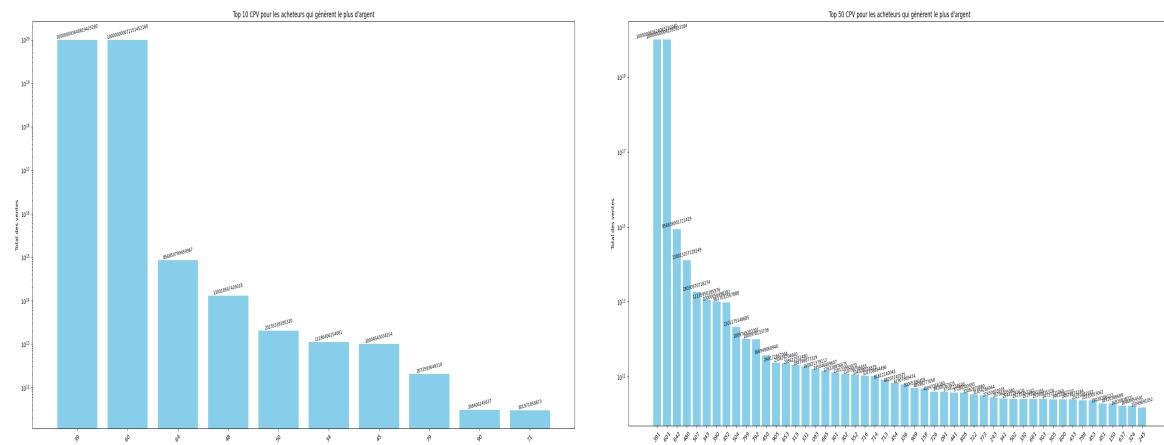
Figure 52. Top des fournisseurs qui gagnent le plus d'argent

Les deux premiers CPVs (Travaux de construction et Matériels médicaux et pharmaceutiques) sont presque les mêmes que précédemment. Viennent s'ajouter les services liés aux transports (CPV 60, 601), la télécommunication et les services postaux (CPV 64, 642) et les services de réparation et d'entretien (CPV 50, 507).

Cela s'explique puisque la dépendance croissante aux technologies de l'information et de la communication pour les opérations gouvernementales et le besoin d'assurer la connectivité à travers le pays et que la maintenance est un domaine constant, nécessaire pour préserver la valeur des investissements dans les infrastructures et les équipements.

Les données indiquent que les acheteurs publics dépensent des sommes importantes dans divers secteurs, reflétant leurs besoins et priorités. Certains CPVs apparaissent à la fois dans les activités les plus actives et les plus lucratives, indiquant des domaines où les acheteurs publics sont particulièrement investis.

De nouveaux, les CPVs qui reviennent le plus sont Travaux de construction (CPV 45 et 452) et Transport et services connexes (CPV 60 et 601). Les nouveaux CPVs remarquables sont les meubles et équipements de bureau (CPV 391). Cette nouvelle apparition suggère des investissements significatifs dans l'aménagement des espaces et la fourniture d'équipements pour les services publics.



(a) Distribution des numberTenders par nombre d'occurrences

(b) Distribution des numberTenders par type de contrat

Figure 53. Top des acheteurs qui dépensent le plus d'argent

L'analyse des flux des CPVs révèle les priorités des dépenses publiques et les domaines d'activité économique les plus dynamiques. Les secteurs de la construction et de la santé, en particulier, bénéficient d'une attention continue, traduisant des besoins fondamentaux en infrastructures et en soins médicaux. L'entretien des bâtiments et l'investissement dans le mobilier et l'équipement de bureau soulignent l'importance accordée aux environnements de travail et aux services administratifs.

Les dépenses dans les services de transport et de télécommunications reflètent une volonté d'améliorer la connectivité et la mobilité, essentielles pour le développement socio-économique. Ces tendances mettent en lumière la volonté d'optimiser et de moderniser les services publics pour répondre efficacement aux besoins des citoyens et des entreprises.

4 Extension

- Cette section porte sur l'enrichissement de la base avec des données externes, et l'analyse de la base étendue ainsi obtenue.

4.1 Méthode d'enrichissement

Le processus de mise à jour de la base de données avec des informations provenant de fichiers XML du site TED est une tâche complexe qui implique plusieurs étapes détaillées pour garantir l'intégrité et la précision des données.

- Extraction des Identifiants Uniques : Le script commence par récupérer tous les identifiants uniques (`tedCanId`) des lots listés dans la base de données existante. Ces identifiants servent à récupérer les fichiers XML correspondants depuis le site TED. Contrairement à la base de donnée (une concaténation entre année et id), la structure XML est : « `id-année` ».
- Récupération et Analyse des Fichiers XML : Chaque identifiant est transformé en un nom de fichier spécifique qui est utilisé pour télécharger le fichier XML correspondant du site TED. Ces fichiers sont ensuite lus et analysés en utilisant la bibliothèque `BeautifulSoup`, qui est conçue pour naviguer et chercher dans les données structurées de XML.
- Extraction et Vérification des Informations : Des fonctions spécifiques sont déployées pour extraire des données telles que la date d'attribution, l'estimation et le prix réel du

lot, les numéros CPV (classification des produits), et d'autres informations contractuelles directement des fichiers XML. Ces données sont extraites en tenant compte des diverses balises et attributs qui peuvent différer d'un fichier à l'autre.

4. Mise à Jour de la Base de Données : Le script compare les informations extraites avec celles déjà présentes dans la base de données. Si une différence est détectée, la base de données est mise à jour avec les nouvelles informations.
5. Historisation : Chaque mise à jour est enregistrée dans un fichier CSV. Cela inclut l'ancienne et la nouvelle valeur, fournissant ainsi un historique complet des modifications pour des vérifications ultérieures et une transparence accrue.

La difficulté principale de ce processus réside dans la variabilité de la structure des fichiers XML. Le nom des balises, leurs attributs et leur emplacement dans l'arborescence du document peuvent changer, ce qui nécessite une programmation flexible pour gérer ces variations. Le script doit donc comporter une logique robuste pour identifier correctement ces balises malgré les variations, ce qui peut être réalisé à travers une combinaison de recherche de balises spécifiques, d'analyse d'attributs et de gestion d'exceptions pour les cas non conformes.

Attribut	Étiquette
awardDate	CONTRACT_AWARD_DATE DATE_CONCLUSION_CONTRACT
awardEstimatedPrice	INITIAL_ESTIMATED_TOTAL_VALUE_CONTRACT VALUE_COST ESTIMATED_TOTAL
awardPrice	VALUE VAL_OBJECT
cpv	ORIGINAL_CPV OBJECT_DESCR
numberTenders	NB_TENDERS_RECEIVED OFFERS RECEIVED NUMBER
fraAgreement	CONCLUSION_FRAMEWORK AGREEMENT
lotsNumber	AWARD_CONTRACT BIB_DOC_S CONTRACT_NUMBER LOT_NO LOT_NUMBER
accelerated	ACCELERATED_PROC
numberTendersSme	NB_TENDERS_RECEIVED_SME
gpa	RP_REGULATION CONTRACT_COVERED_GPA
typeOfContract	NC_CONTRACT_NATURE
renewal	RENEWAL
contractDuration	DURATION
publicityDuration	DATE_RECEIPT_TENDERS - DATE_DISPATCH_NOTICE

Table 23. Répartition des valeurs pour awardEstimatedPrice et numberTenders

4.2 Analyse

Par rapport aux changements TED, uniquement 3 702 fichiers (environ 1%) ont été générés par rapport aux nombres de tedCanId uniques 410 283. Parmi ces fichiers, il y a eu 2 081 changements.

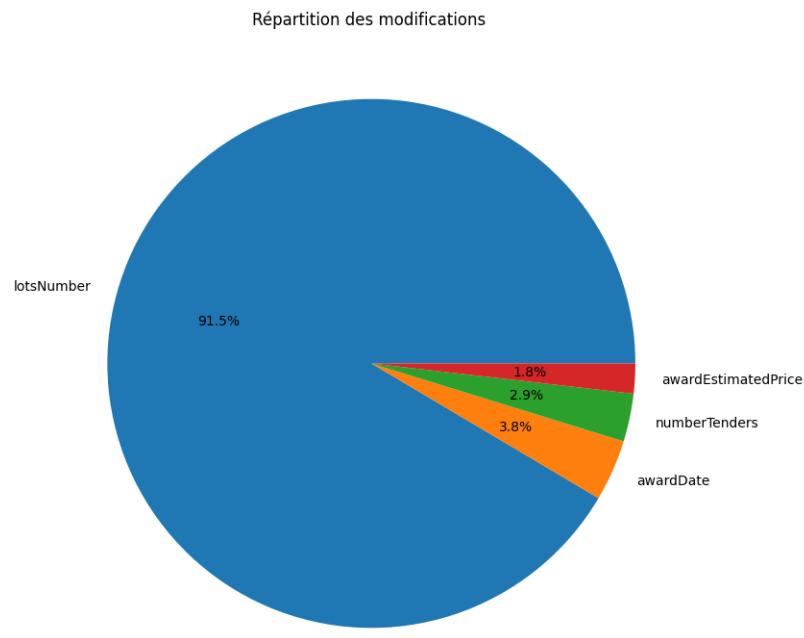


Figure 54. Diagramme circulaire des modifications de la base de données

La grande majorité des changements (91.5%) concerne lotsNumber. Cela concerne principalement des corrections où la valeur précédente était vide, voire incorrecte dans de rares cas. Les autres changements, bien qu'ils représentent une part plus petite, indiquent des modifications sur des informations.

Ces changements, bien qu'en pourcentage plus faible, peuvent avoir un impact significatif sur l'analyse des données, reflétant des ajustements importants concernant l'estimation des prix, le nombre d'offres reçues, et les dates d'attribution des marchés publics.

tedCanId	Attribut	Ancienne Valeur	Nouvelle Valeur
Différence importante			
20106782	awardEstimatedPrice	64 224 124	64 224.125
20106782	awardEstimatedPrice	9 417 558	9 417.558
Différence faible			
201024344	awardEstimatedPrice	3 054 148.2	3 054 148.15
201024344	awardEstimatedPrice	7 990 417	7 990 417.2

Table 24. Extrait des modifications engendrées par l'appel API

En termes d'impact, les modifications importantes nécessitent une attention particulière, car elles affectent fortement l'analyse des tendances des coûts et la planification budgétaire.

Les petites modifications, bien qu'elles puissent sembler négligeables, sont également importantes, car elles reflètent la précision et l'attention aux détails dans la gestion des données.

5 Conclusion

Nous avons pu apercevoir à travers le rapport sur les marchés publics et la base FOPPA, un cas de fouille de données grâce à une approche méthodique et collaborative. L'effort collectif du groupe a permis de nettoyer et d'analyser en profondeur les données issues des marchés publics en France, en se concentrant particulièrement sur la période allant de 2010 à 2020. L'utilisation d'API et les manipulations manuelles a permis de révéler des tendances, des caractéristiques et des anomalies importantes qui éclairent la compréhension des dynamiques des marchés publics.

L'analyse des données, enrichie par des visualisations graphiques perspicaces, a mis en évidence des aspects clés tels que les types de contrats les plus courants, les secteurs d'activité les plus impliqués et les variations dans les pratiques d'attribution. De plus, les questionnements initiaux ont été abordés avec rigueur, permettant de dégager des insights sur les flux de communication et les échanges monétaires entre les entreprises, ainsi que sur l'implication des PME dans les marchés publics.

Le partage des connaissances et des compétences au sein du groupe a été un facteur clé de succès, chaque membre apportant une expertise unique au projet.

6 Lexique

Marché public [WikipediaMarchépublic] : Un marché public est un contrat conclu à titre onéreux entre un acheteur public (appelé pouvoir adjudicateur dans le droit de l'Union européenne) et des personnes publiques ou privées, et qui répond aux besoins de cet acheteur public en matière de fournitures, services et travaux. Un marché public peut être passé par différents types d'acheteurs publics : les collectivités publiques (État central, entité fédérée, collectivité territoriale, agence publique spécialisée) ou des personnes morales assimilées à des acheteurs publics. Le site des publications des annonces de la France : Marches-publics.gouv.fr.

Lot : Un lot est une unité autonome qui est attribuée séparément. L'allotissement consiste à diviser un marché public en lots qui sont des unités autonomes. Chaque lot, issu de ce fractionnement, correspond à un marché distinct faisant l'objet d'un marché séparé et s'oppose en cela au marché unique. Par exemple, lorsqu'une opération comporte de divers travaux, des lots peuvent être établis correspondant aux divers ouvrages, spécialités et usages professionnels.

TED(Tenders Electronic Daily) : Le Tenders Electronic Daily (TED) est un site dédié à la publication des appels d'offres et des avis d'attribution liés aux marchés publics. Par conséquent, ce site héberge des documents relatifs à tous les marchés publics dont le coût estimé est supérieur au Seuil. En outre, il peut également héberger des contrats inférieurs à ce seuil, mais une telle publication n'est pas obligatoire.

Correction notice(AVIS RECTIFICATIF) : L'attribut *correctionsNb* du tableau *Lots* indique le nombre de fois que le lot a été corrigé. Par conséquent, les lots peuvent présenter des erreurs. Lorsque cela se produit, le pouvoir adjudicateur est tenu de notifier ces erreurs aux candidats (les potentiels fournisseurs) en leur adressant un avis rectificatif, lequel est publié sur les mêmes canaux d'informations que le lot. Outre les erreurs, une telle publication peut également être déclenchée par des modifications notables des conditions. Une correction peut entraîner une extension de la période d'acceptation.

CPV : Le Code CPV est obligatoirement renseigné dans les avis de marchés. Il est composé d'un vocabulaire principal servant à définir l'objet d'un marché ainsi que d'un vocabulaire supplémentaire permettant d'introduire des données qualitatives complémentaires. Le vocabulaire principal repose sur une structure arborescente de codes comptant jusqu'à 9 chiffres (un code à 8 chiffres plus un chiffre de contrôle) auxquels correspond un intitulé qui décrit le type de fournitures, de travaux ou de services, objet du marché. Dans cette base de donnée, le code cpv sont à 8 chiffre.

fraEstimated : La relation au champ Accord-cadre FRA_ESTIMATED indique la relation (possible) détectée automatiquement entre l'avis et un accord-cadre :

- K : mot-clé "cadre" trouvé dans le titre ou la description de l'avis
- A : plusieurs attributions ont été données pour un lot
- C : la plupart des avis qui suivent cet avis sont marqués comme accord-cadre

onBehalf : Indique que l'acheteur est un regroupement, souvent utilisé lorsque plusieurs petites entreprises se regroupent pour répondre à un appel d'offres, partageant ainsi les coûts associés à la proposition.

accelerated : Un booléen indiquant l'utilisation de la procédure rapide dans la base de

donnée. Pour indiquer que si la procédure a été accélérée [[WikipediaMarchépublic](#)].

Marché conjoint : Le marché conjoint est un type de groupement. Un marché est conjoint lorsque chacun des opérateurs économiques membres du groupement s'engage à exécuter la ou les prestations qui sont susceptibles de lui être attribuées dans le marché.

Accord-cadre : L'accord-cadre fait partie des techniques d'achat mobilisables par l'acheteur pour faire la présélection d'opérateurs économiques susceptibles de répondre à son besoin, ou permettre la présentation des offres ou leur sélection. Il permet de présélectionner un ou plusieurs opérateurs économiques en vue de conclure un contrat qui fixe tout ou partie des règles relatives aux commandes à passer au cours d'une période donnée.

Out of directives : C'est un avis d'attribution qui a été émis sans qu'un appel d'offres ait été lancé. En d'autres termes, il signale qu'un contrat a été attribué directement à un fournisseur sans passer par une procédure d'appel d'offres, ce qui peut parfois être justifié par des circonstances particulières telles que l'urgence, l'absence de concurrence, ou des critères spécifiques définis par la législation régissant les marchés publics.

Government Procurement Agreement (GPA) : En général, dans le contexte des marchés publics, le terme "gpa" est souvent utilisé pour désigner l'Accord sur les marchés publics (AMP) ou Government Procurement Agreement (GPA) en anglais.

L'attribut "gpa" en tant que booléen pourrait être utilisé pour indiquer si un marché ou un contrat est lié à l'Accord sur les marchés publics. Cet accord, qui est un accord plurilatéral négocié dans le cadre de l'Organisation mondiale du commerce (OMC), concerne l'accès aux marchés publics entre ses membres. Si "gpa" est défini comme vrai (true), cela signifie que le marché ou le contrat en question est soumis aux règles et obligations énoncées dans l'Accord sur les marchés publics. Sinon, cela signifierait qu'il n'est pas couvert par cet accord.

ContractorSme : Les données pour le champ *ContractorSme* indiquent si le(s) gagnant(s) sont des PME. Elles contiennent des valeurs telles que "N-Y-N-Y", qui représentent plusieurs gagnants et indiquent s'ils sont une PME, où "N" signifie "Non" et "Y" signifie "Oui". Chaque valeur "N" ou "Y" correspond à un gagnant. Cependant, l'ordre des entreprises gagnantes pour le lot est manquant. Par conséquent, nous ne pouvons pas savoir quelle valeur correspond à quelle entreprise lorsque plusieurs valeurs sont présentes.

topType : Type de procédure d'attribution, il contient :

- AWP : Attribution sans publication préalable d'un avis de marché;
- COD : Dialogue compétitif;
- NOC/NOP : Négocié sans appel à la concurrence préalable;
- NIC/NIP : Négocié avec un appel à la concurrence;
- OPE : Procédure ouverte;
- RES : Procédure restreinte;
- INP : Partenariat innovant.

Renouvable d'un contrat : La possibilité de renouveler le contrat après l'expiration de la période initiale.

Siret:Le numéro de SIRET est Le numéro SIRET (Système d'Identification du Répertoire des Entreprises et de leurs Établissements) est un identifiant unique attribué à chaque établissement en France[1]. Il est utilisé pour identifier de manière précise une entité économique, qu'il s'agisse d'une entreprise, d'une association, d'une administration publique ou d'un autre

organisme.