Nicolas DUFOUR

Le travail le plus sexy du XXIe siècle à travers le monde.

Le métier de Data Scientist a été nommé Job le plus sexy du XXIème siècle par le magazine Harvard Business en 2012. C'est un métier qui mélange mathématiques, informatique et capacité d'analyse pour donner du sens aux données qui affluent de partout. Au courant de la décennie, les entreprises ont cherché à recruter des professionnels de la donnée dans la plupart des corps de métiers. L'Union Européenne estime qu'en Europe, les travailleurs de la donnée vont doubler entre 2016 et 2020 pour atteindre 10,4 millions de personnes travaillant dans ce milieu. En même temps, l'économie de la data européenne va elle aussi doubler de valeur pour atteindre 739 milliards d'euros. Avec le manque de talents pouvant effectuer ces tâches, le métier de Data Scientist est parmi les professions les mieux payés. Cependant, cette profession est inégalement rémunérée à travers le monde. Nous allons nous poser la question suivante :

Quelles différences peut-on observer dans le salaire des Data Scientist à travers le monde ?

Data Scientist was named the Sexiest Job of the 21st Century by Harvard Business magazine in 2012. It's a job that blends mathematics, computer science, and analytical skills to make sense of the data that's pouring in from everywhere. Over the decade, companies have been looking to recruit data professionals in most sectors. The European Union estimates that the number of data workers in Europe will double between 2016 and 2020 to reach 10.43 million. At the same time, the data economy will also double in value to €739 billion. With the lack of talent to perform these tasks, this has made the job of Data Scientist one of the highest paid professions. However, this profession is uneven across the world. We will ask ourselves the following question:

What differences can we observe in the compensation of data scientists around the world?

1. Données et Sources

1.1. Origine des données

Les données proviennent du site Kaggle. Kaggle est une plateforme communautaire de Data Scientist qui est la référence dans le milieu. Elle centralise des bases de données ainsi que des défis autour de ces bases de données. Ces données ont été récoltées en 2017 auprès de 16 000 travailleurs de la donnée lors d'une enquête visant à définir un panorama du métier de Data Scientist à travers le monde avec plus d'une centaine de questions allant des études suivies, aux outils utilisés et surtout, composante qui nous intéresse le plus, le salaire. Nous avons aussi comme source le FMI (Fond Monétaire International) pour le PPA (Parité de pouvoir d'achat)
De plus, cette étude est supportée par un rapport de l'union européenne sur le marché européen de la données de 2017.

1.2. Nettoyage des données

Nos données provenant d'une étude participative, certaines données peuvent être manquantes ou aberrantes. Ainsi, nous allons d'abord supprimer toutes les entrées qui ne renseignent pas correctement le salaire. Puis, nous allons aussi supprimer les salaires trop petits et trop élevés. Ainsi nous ne considérerons que les salaires annuels entre 10 000 USD et 2 000 000 USD. Si l'on a pris une borne inférieure si petite, c'est pour ne pas pénaliser les pays en voie de développement aux salaires plus faibles.
Nous allons ensuite considérer les pays ayant eu plus de 80 réponses pour avoir une certaine significativité statistique dans notre étude.

2. Transformations et outils statistiques

2.1. Normalisation des revenus par la parité de pouvoir d'achat.

Nous disposons des revenus en provenance de diverses origines à travers le monde dans différentes monnaies. Nous disposons des taux de change vers le dollar américain, mais cependant cette conversion n'est pas interprétable en soi. En effet, 1 dollar n'a pas le même pouvoir d'achat à travers le monde. Pour cela, nous allons nous baser sur un indicateur appelé la parité du pouvoir d'achat. Cet indice détermine combien de biens peut-on utiliser pour acheter avec la monnaie voulue par rapport à une autre. La monnaie de référence utilisée ici sera le dollar américain. Celui que l'on va utiliser provient du FMI.
Ainsi, nous allons normaliser nos revenus avec la formule suivante :

$$Revenus_{normalise} = \frac{Revenu_{local} * Change_{local \to dollar}}{PPA_{local \to dollar}}$$

2.2. Test Statistique de Mood

Nous allons être amenés dans cette étude à étudier une différence entre les médianes de notre jeu de données. Pour cela, on dispose d'un test statistique appelé test de Mood. C'est un cas particulier du test du khi-deux. En effet nous allons tester l'hypothèse suivante :

Soit 2 jeux de données X_1 et X_2 de médianes respective m_1 et m_2 .
On veut tester l'hypothèse nulle suivante :

$$H_0 : m_1 = m_2$$

contre l'hypothèse alternative :

$$H_1 : m_1 \neq m_2$$

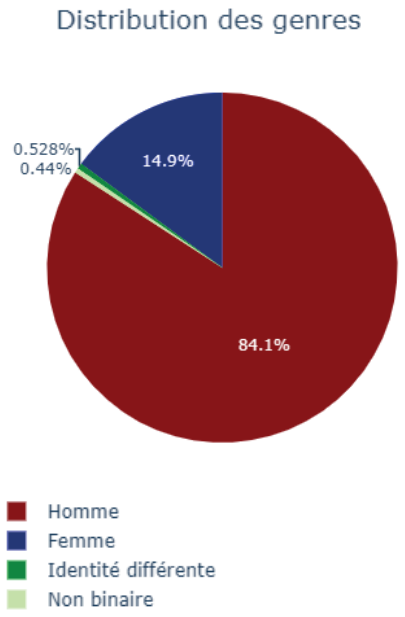
Pour cela, le test de Mood calcule une médiane globale des 2 jeux de données et nous vérifions si chaque élément appartient à un jeu de données de telle médiane grâce à un test du khi-deux.

3. Profil type du Data Scientist

Avant de continuer notre étude pays par pays, nous allons d'abord essayer de dresser un profil type des répondants à l'enquête. Ceci pourra nous donner une idée du profil type du Data Scientist.

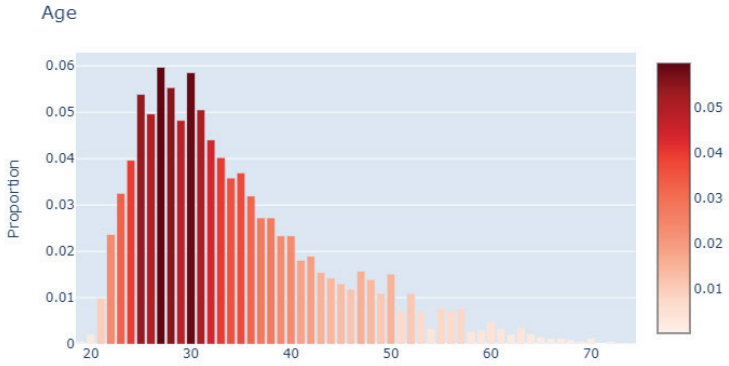
3.1. Genre

On voit que le métier de Data Scientist est majoritairement masculin.



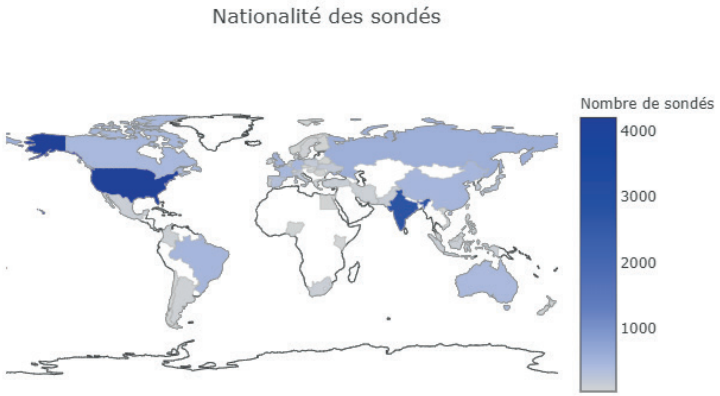
3.2. Age

Le métier de Data Scientist étant relativement récent, la plupart des professionnels se répartissent autour de la trentaine. On voit cependant une partie non négligeable de la profession qui à plus de 40 ans. Ce sont les personnes qui se sont convertis d'autres métiers liés à l'informatique et aux mathématiques qui ont sauté le pas pour devenir Data Scientist.



3.3. Nationalités

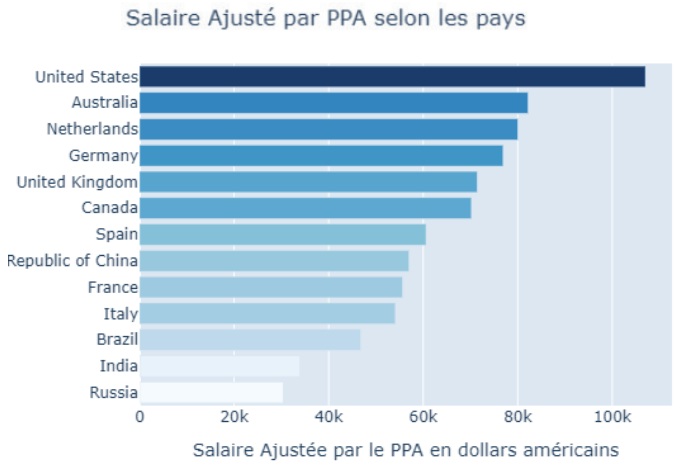
On s'intéresse ici à la nationalité des sondés. On peut voir que les pays qui ont répondu sont principalement les pays riches ou ayant une économie émergente. Pour le reste de notre étude, nous ne conserverons que les pays ayant plus de 80 sondés. On se retrouve avec la liste de pays suivante : États-Unis, Inde, Royaume-Uni, Allemagne, France, Brésil, Canada, Espagne, Australie, Russie, Italie, Chine et Pays-Bas.



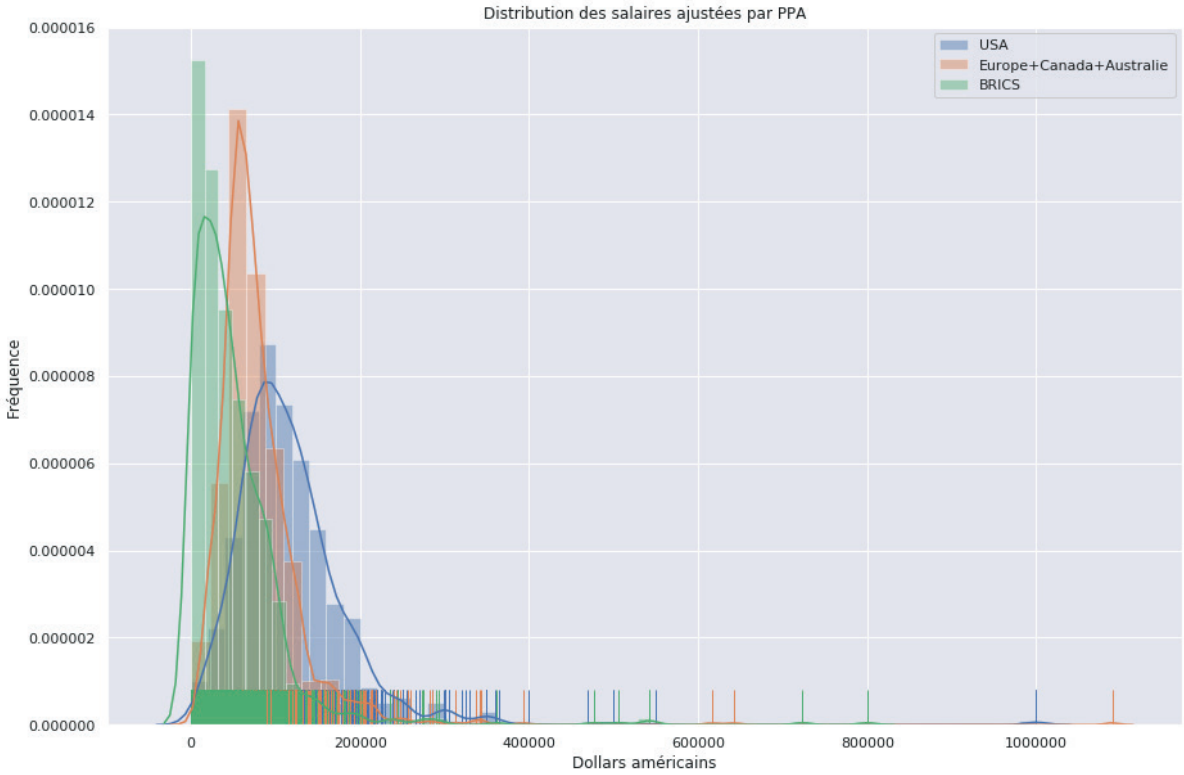
4. Étude des différences de salaire par pays

4.1. Revenu médian par pays normalisé avec PPA

Le graphe représente, le revenu en dollars ramené au pouvoir d'achat de chaque pays par le PPA. On voit que les USA ont le plus grand revenu. La Russie et l'Inde eux, ont les plus faibles revenus. Pour la suite de notre étude, nous allons faire 3 types de groupe. Le premier les USA, le second, l'Europe avec le Canada et l'Australie et finalement le dernier, les BRICS avec la Chine, le Brésil, l'Inde et la Russie. Le terme BRICS définit en effet les économies émergentes qui sont constituées des 4 pays plus l'Afrique du Sud.



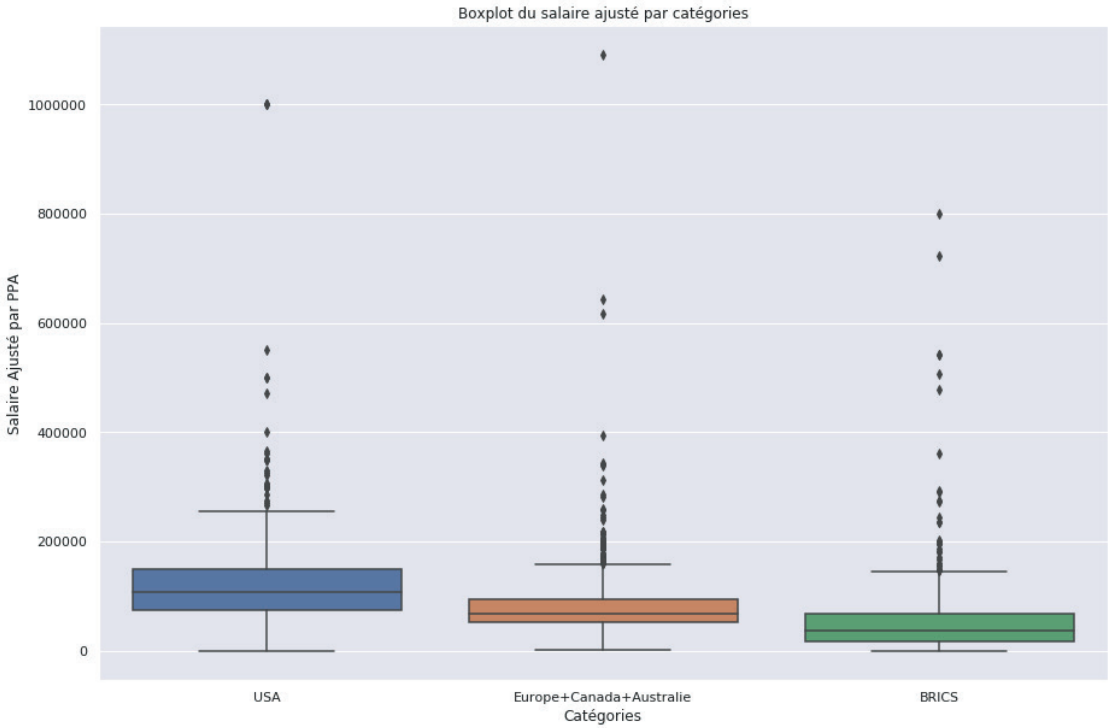
On peut observer les distributions suivantes pour les salaires normalisés par PPA:



On se rend compte que les USA ont une plus grande variance autour de leur médiane. Les BRICS eux ont une asymétrie gauche forte. On a donc beaucoup de petits salaires. L'Europe, Canada et Australie, eux ont une distribution centrée de faible variance. Il y a donc moins de variabilité dans les salaires européens par rapport aux salaires américains. Ceci est en grande partie dû aux très grands salaires américains qui existent moins en Europe/Canada/Australie.

4. Étude des différences de salaire par pays

Pour y voir plus clair, regardons les diagrammes en boîtes de ces distributions :



Les revenus normalisés par PPA médians suivants :

BRICS	Europe/Canada/ Australie	USA
37 021$	68 181$	107 000$

Nous allons effectuer un test de Mood pour vérifier que l'on a bien une différence statistique.

Test de Mood	BRICS = EU/ CA/AU	EU/CA/AU = USA	USA= BRICS
p-value	1.35e-45	1.22e-72	9.55e-118

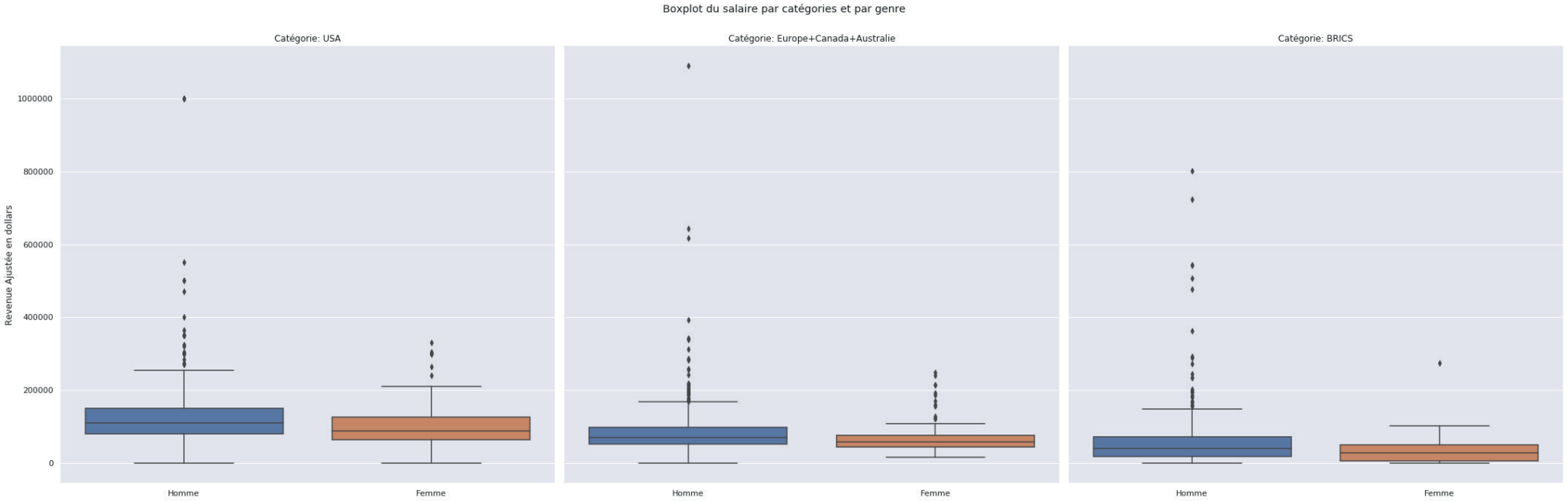
Ce test statistique nous permet de rejeter l'hypothèse nulle et donc admettre qu'il existe des différences de salaire entre les 3 régions. En effet nos p-value sont suffisamment petites pour rejeter l'hypothèse nulle avec une probabilité presque certaine.

Ceci confirme nos hypothèses précédentes et on voit de plus que les BRICS ont un revenu médian plus faible mais le 3e quartier statistique est similaire à celui de l'Europe. Ainsi, le quart supérieur de professionnels bien payés se situe au même salaire que l'Europe. C'est sur les professionnels peu payés où les salaires sont plus bas.

4. Étude des différences de salaire par pays

4.2. Différence en fonction du genre

On a vu que le métier de Data Scientist était surtout exercé par les hommes. On peut se poser la question de la différence de salaire entre homme et femmes. Nous n'avons pas pu réaliser pas les études statistiques liées au genre et les différences d'homme/femme car le nombre de réponses de ces catégories n'est pas suffisant pour avoir de l'importance statistique. Comme la distribution des revenus varie fortement entre régions, il peut être intéressant d'étudier séparément chaque région. On obtient les diagrammes en boîtes suivants :



4. Étude des différences de salaire par pays

Les revenus normalisés par PPA médians sont :

Revenus normalisés par PPA médians	BRICS	Europe/ Canada/ Australie	USA
Homme	39 548$	70 247$	110 000$
Femme	28 248$	58 974$	89 250$

Nous allons pour chaque région faire le test statistique pour voir si la différence des médianes entre hommes et femmes est réelle dans la profession.

Revenus normalisés par PPA médians	BRICS	Europe/ Canada/ Australie	USA
p-value de la différence de médiane des salaires homme femme	0.016	1e-4	1e-5

Pour une probabilité de rejet de 0.05, on peut rejeter les 3 hypothèses nulles et donc, on a des différences de salaire médian dans les 3 régions. On remarquera quand même que l'on n'a pas autant de certitude pour les BRICS mais cela est dû à un nombre de données moins important que les 2 autres régions.

On observe 2 choses :

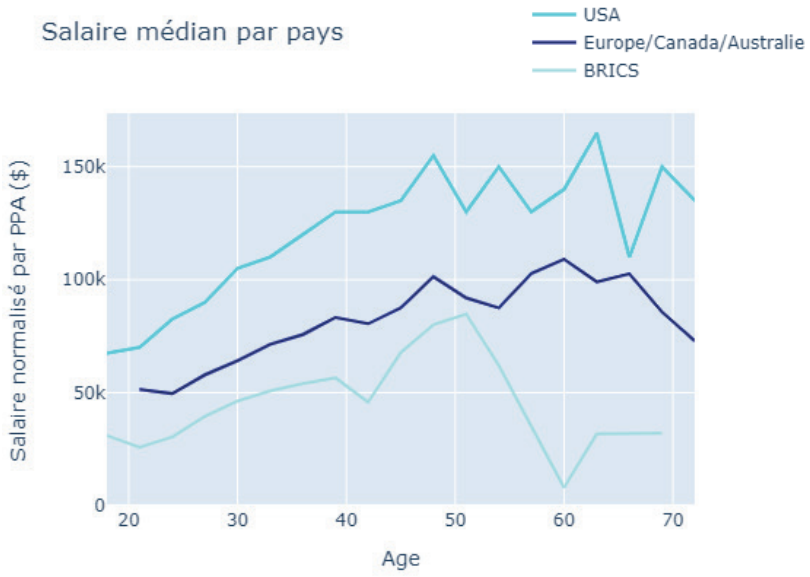
Premièrement, le salaire médian est plus bas pour les femmes dans les 3 régions.

Deuxièmement, les femmes ont beaucoup moins de valeurs extrêmes supérieures par rapport aux hommes. Les très hauts salaires sont réservés aux hommes. Ceci est encore plus vrai dans les BRICS où l'on a très peu de valeurs extrêmes supérieures pour les femmes.

4. Étude des différences de salaire par pays

4.3. Progression du salaire avec l'âge

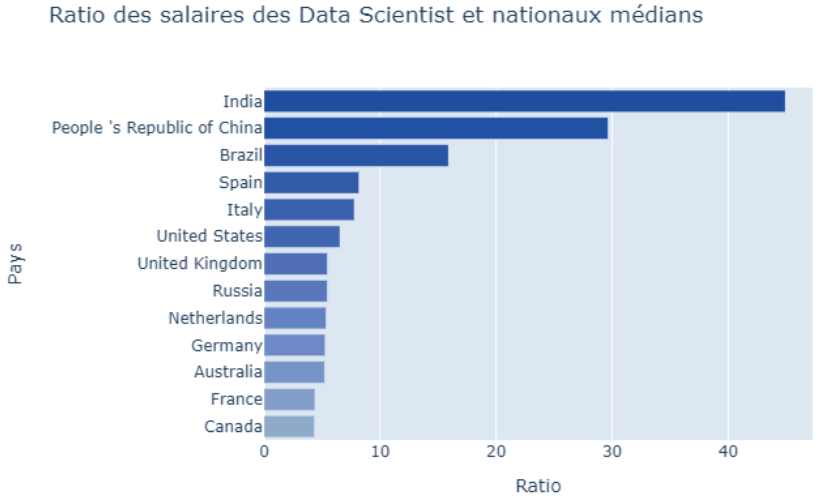
Nous allons maintenant étudier comment évolue le salaire en fonction de l'âge du Data Scientist. Nous allons encore une fois compartimenter notre étude dans les 3 régions. On va aussi regrouper les âges par tranche de 3 pour diminuer la variance de notre courbe. On obtient le graphique suivant :



On se rend compte que le salaire augmente avec l'âge. Plus le Data Scientist prend de l'expérience et mieux il est payé. À noter que l'on a toujours une grande variabilité à partir de 40 ans car nos sondés sont proches de 30 ans. Ainsi à partir de 40 ans, on a peu de données et donc une plus forte variance.

4.4. Comparaison au salaire médian national

Nous allons comparer le salaire des Data Scientist par rapport aux salaires médians nationaux pour voir comment se situe le métier par rapport au reste de la population. Cependant, nous ne disposons des revenus médians par pays que pour l'année 2013. Donc, nous avons utilisé l'inflation pour simuler les revenus médians de 2017.
Les graphiques suivant présentent les résultats par pays (en pourcentage par rapport aux salaires moyens):



On peut voir que les Data Scientist des BRICS gagnent beaucoup plus que les citoyens moyens de leurs pays. Ceci est principalement dû à des salaires minimum très bas et beaucoup de travail peu payé. On voit aussi que la France et le Canada sont les pays où le métier de Data Scientist est le moins bien rémunéré par rapport au revenu moyen du pays.

5. Conclusion

On a pu voir que le métier de Data Scientist est particulièrement présent dans les grandes puissances économiques comme l'Europe, les BRICS ou l'Amérique du Nord. C'est un métier exercé par une population assez jeune et majoritairement par des hommes. Dans cette étude, nous avons dû convertir les salaires en utilisant une unité commune: le dollar. Pour cela, nous nous sommes basés sur la correction du PPA (Parité du pouvoir d'achat). Les rémunérations sont inégales entre pays mais aussi à l'intérieur d'un pays : Les femmes et les jeunes ont des positions qui sont moins bien rémunérées que la médiane. Le traitement des données a aussi montré que les Data Scientist des pays des BRICS ont une position plus favorable par rapport au citoyen médian de leurs pays que dans les pays Occidentaux.

6. Critique et voie d'approfondissement

Cette étude peut être améliorée sur plusieurs points. Premièrement elle est basée sur le PPA (Parité du pouvoir d'achat). Cette correction n'est pas parfaite car elle se base sur le prix d'un panier d'articles prédéfini. À travers le monde, les besoins de base ne sont pas nécessairement les mêmes.

L'autre critique, c'est que cette étude se base sur les statistiques des pays. Or au sein des pays les variations peuvent être très fortes que ce soit sur le PPA ou les salaires (par exemple dans la Silicon Valley). Un approfondissement de ce sujet serait l'etude des différences sur un panel de grande villes où la data-science est populaire. (Paris, Amsterdam, Singapour, San Francisco etc…)

Sources

Étude du marché Européen : https://ec.europa.eu/digital-single-market/en/news/final-results-european-data-market-study-measuring-size-and-trends-eu-data-economy

Données Data Scientist : https://www.kaggle.com/kaggle/kaggle-survey-2017

PPA par pays : https://www.imf.org/en/Publications/WEO/Issues/2017/09/19/world-economic-outlook-october-2017#Chapter%201

OCDE inflation : https://data.oecd.org/price/inflation-cpi.htm

OCDE salaire median : https://stats.oecd.org/Index.aspx?DataSetCode=IDD

Brown, G. W.; Mood, A. M. On Median Tests for Linear Hypotheses. Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 159--166, University of California Press, Berkeley, Calif., 1951. https://projecteuclid.org/euclid.bsmsp/1200500226