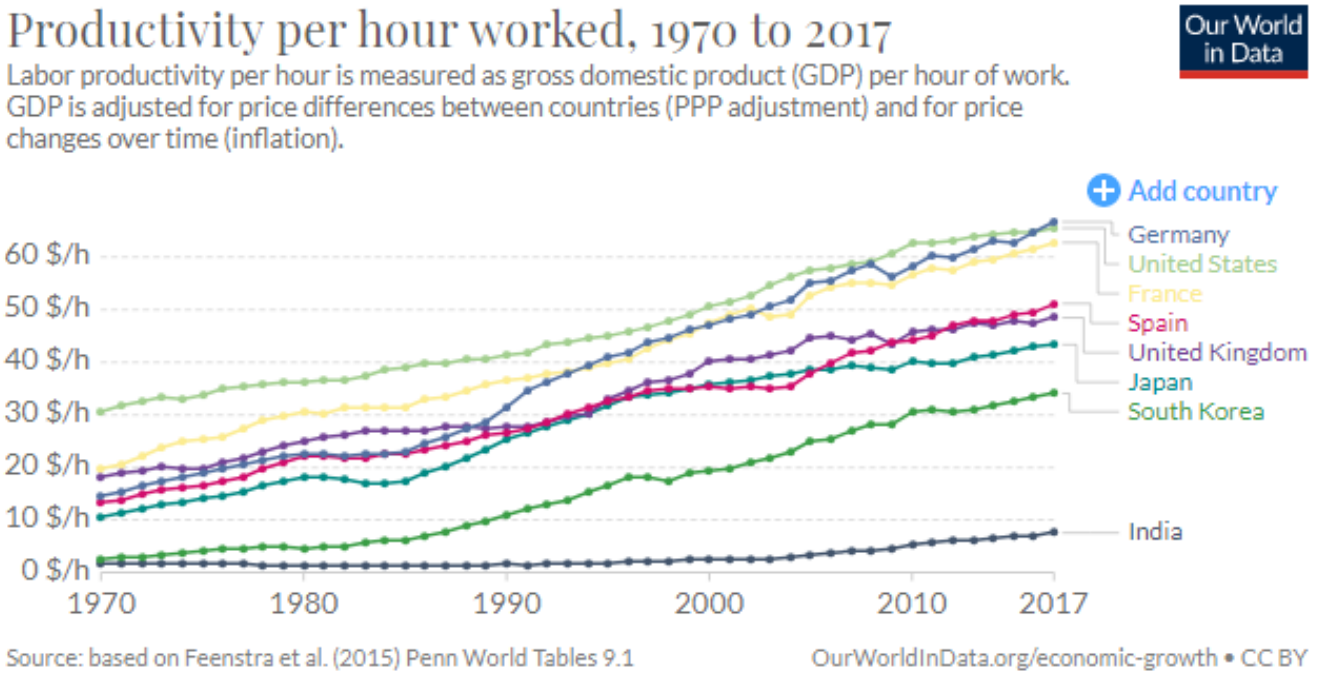
Projet BIUM : Quelles sont les causes de la productivité ?

MEETOOA Kevin
SAHLI Oussama

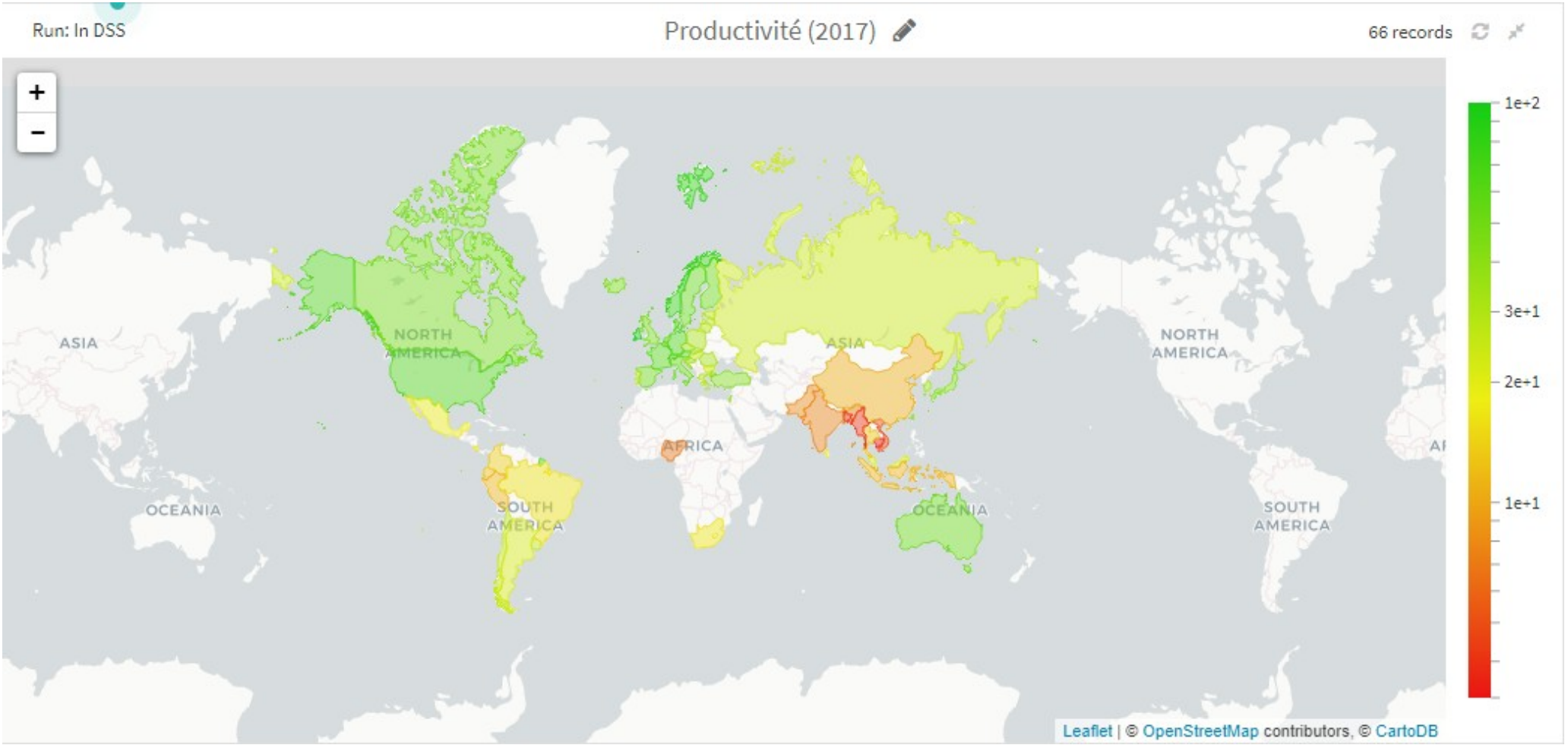
Définition de la productivité

- On utilise la définition de l'OCDE
- Productivité d'un pays: PIB horaire moyen par personne active
- L'inflation est prise en compte



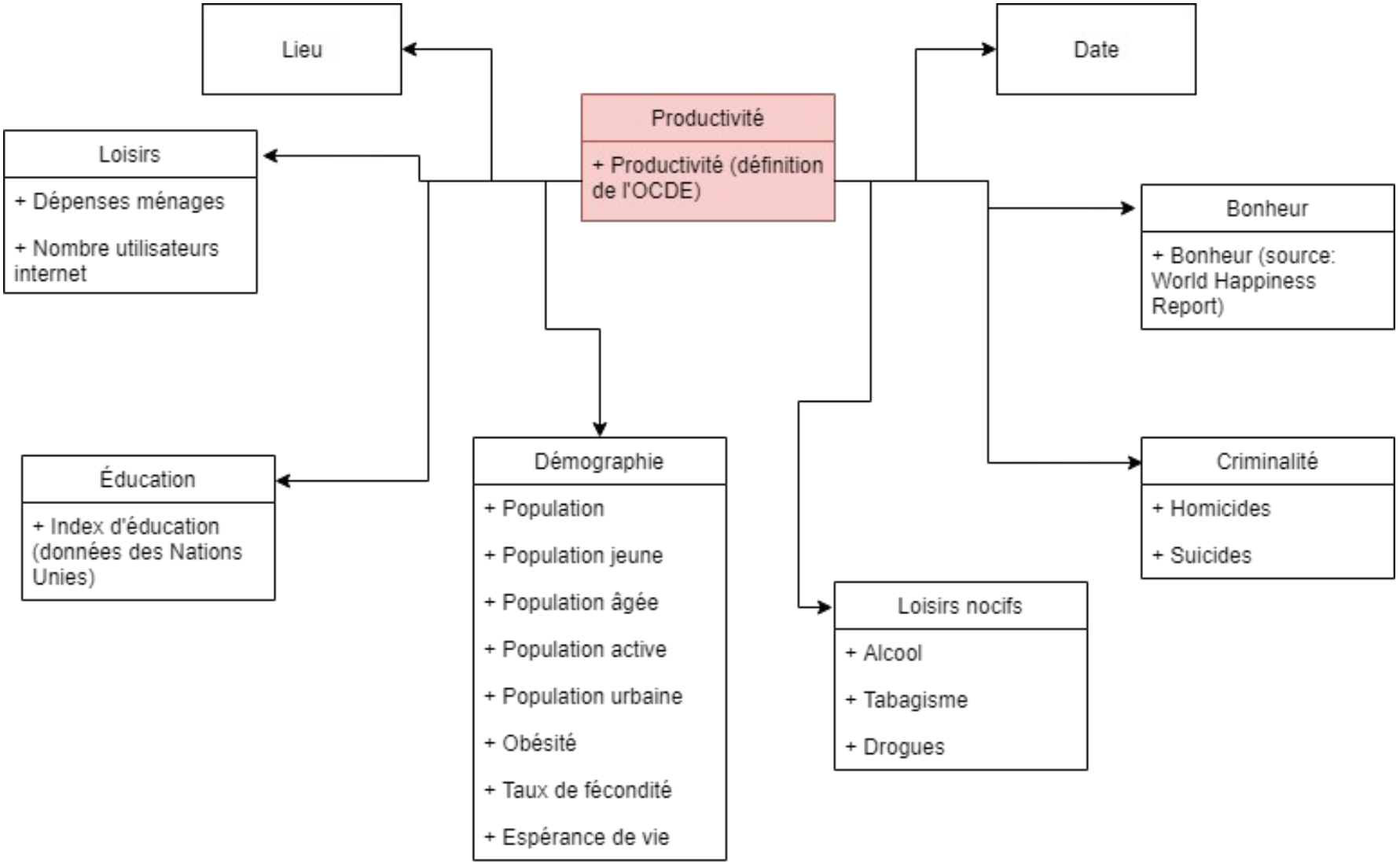
MEETOOA Kevin
SAHLI Oussama

Carte mondiale de la productivité (2017)



MEETOOA Kevin
SAHLI Oussama

Schéma en étoile et axes d'analyse



MEETOOA Kevin
SAHLI Oussama

Datasets

- Principale source de données : https://ourworldindata.org/

- Autres sources de données : Kaggle, data.world

- Forme la plus fréquente des données OurWorldInData: Une ligne par pays et année

	Standard	Standard	Standard	Standard
1	Entity	Code	Year	Per capita CO_2 emissions (tonnes per capita)
2	Afghanistan	AFG	1800	0
3	Afghanistan	AFG	1801	0
4	Afghanistan	AFG	1802	0
5	Afghanistan	AFG	1803	0
6	Afghanistan	AFG	1804	0
7	Afghanistan	AFG	1805	0
8	Afghanistan	AFG	1806	0
9	Afghanistan	AFG	1807	0
10	Afghanistan	AFG	1808	0

Nettoyage et préparation des données

- Autre format rencontré dans les datasets : Une ligne par pays, une colonne par année

	Standard	Standard	Standard	Standard	Standard	Standard	Standard	Standard
1	Entity	Code	1960	1961	1962	1963	1964	1965
2	Aruba	ABW						
3	Afghanistan	AFG	414.371	491.378	689.396	707.731	839.743	1008.425
4	Angola	AGO	550.05	454.708	1180.774	1151.438	1224.778	1188.108
5	Albania	ALB	2024.184	2280.874	2464.224	2082.856	2016.85	2174.531
6	Andorra	AND						
7	United Arab Emirates	ARE	11.001	11.001	18.335	22.002	18.335	22.002

- On transforme ces données selon le schéma défini dans le slide précédent (on « transforme les colonnes en lignes »)

Transformation des données

- On normalise toutes les données afin d'obtenir des valeurs comprises entre 0 et 1

- Formule utilisée pour un ensemble de données $x_1, x_2, \ldots, x_n$

(pour une année donnée) : $x_{normalise} = \frac{x - x_{min}}{x_{max} - x_{min}}$

- On reviendra sur l'importance de la normalisation dans les slides suivants

Illustration : Avant/après normalisation :

	Standard	Standard	Standard	Standard
1	Entity	Code	Year	Productivity
2	Brazil	BRA	2005	10.218953
3	Singapore	SGP	2005	42.10561
4	Sri Lanka	LKA	2005	8.7077026
5	Japan	JPN	2005	38.687721
6	Indonesia	IDN	2005	5.0369921
7	Brazil	BRA	2017	16.340326
8	Singapore	SGP	2017	48.246395
9	Sri Lanka	LKA	2017	18.845987
10	Japan	JPN	2017	43.353165
11	Indonesia	IDN	2017	11.26875

	Standard	Standard	Standard	Standard
1	Country	Code	year	Productivity
2	Brazil	BRA	2005	0.13979374450861304
3	Singapore	SGP	2005	1.0
4	Sri Lanka	LKA	2005	0.09902474675215771
5	Japan	JPN	2005	0.9077956181366019
6	Indonesia	IDN	2005	0.0
7	Brazil	BRA	2017	0.13715248767194343
8	Singapore	SGP	2017	1.0
9	Sri Lanka	LKA	2017	0.2049139960102922
10	Japan	JPN	2017	0.8676705885407254
11	Indonesia	IDN	2017	0.0

Chargement des données dans la BD SQL distante

- On crée une table SQL pour chaque dimension ainsi qu'une table pour le fait

- Productivite (ID_PRODUCTIVITE, productivite, ID_CRIMINALITE*, ID_EDUCATION*, ID_LOISIRNOCIF*, ID_LOISIR*, ID_DEMOGRAPHIE*, ID_DATE*, ID_LIEU*) ⋕
- Criminalite (ID_CRIMINALITE, code, country, year, homicides, suicides)
- Education (ID_EDUCATION, code, country, year, education_index)
- LoisirNocif (ID_LN, code, country, year, nbDecesDrogues, nbDecesTabac, taux_alcool)
- Demographie (ID_DEMO, code, country, year, population, population_jeune, population_vieux, population_active, population_urbaine, esperance_vie, obesite, taux_fecondite)
- Date (ID_DATE, year)
- Lieu (ID_LIEU, country, code)

Chargement des données dans la BD SQL distante

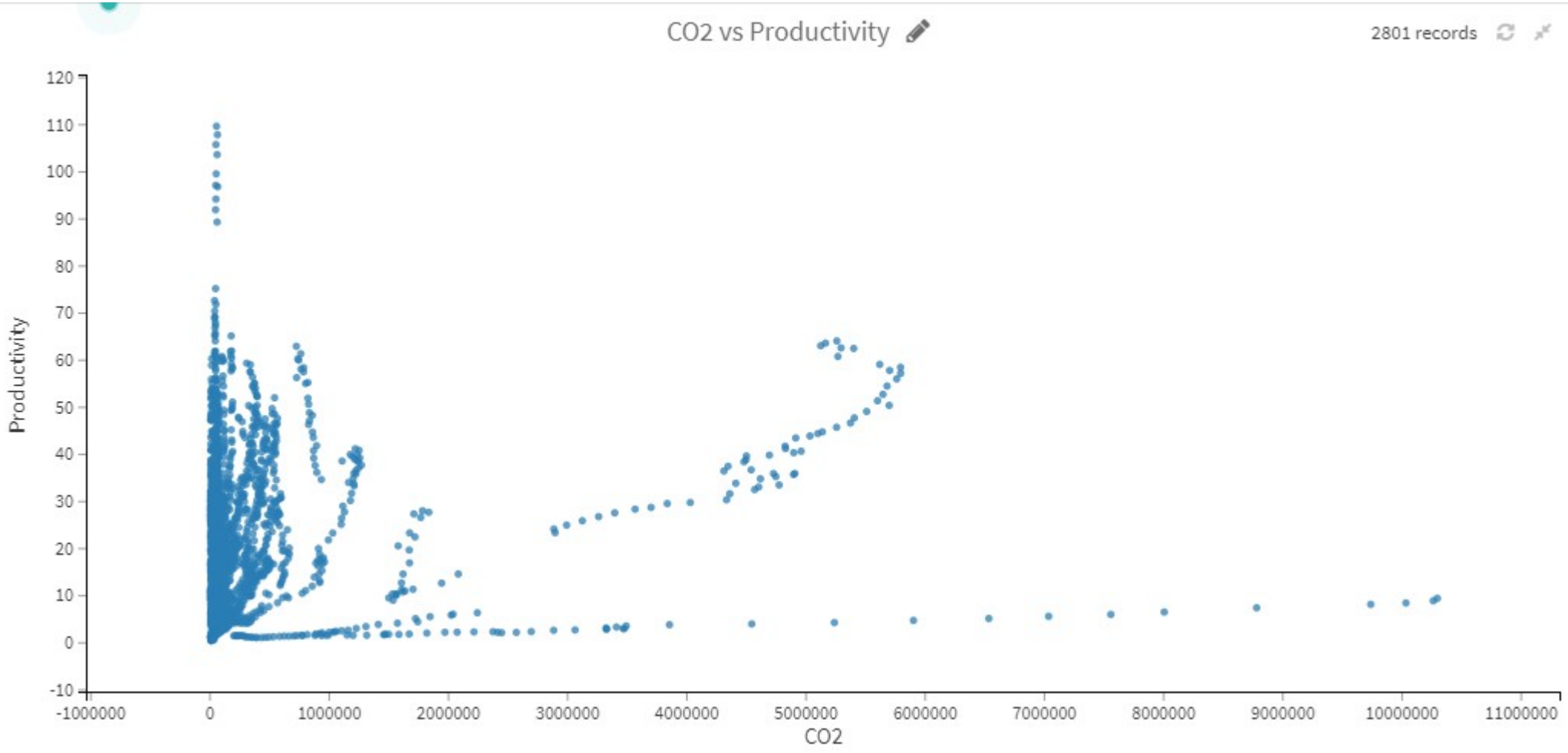
- Illustration :

```
mysql> desc productivite;
+-----------------+---------------+------+-----+---------+-------+
| Field           | Type          | Null | Key | Default | Extra |
+-----------------+---------------+------+-----+---------+-------+
| id_crim         | int(11)       | NO   | MUL | NULL    |       |
| id_education    | int(11)       | NO   | MUL | NULL    |       |
| id_ln           | int(11)       | NO   | MUL | NULL    |       |
| id_loisir       | int(11)       | NO   | MUL | NULL    |       |
| id_demography   | int(11)       | NO   | MUL | NULL    |       |
| id_date         | int(11)       | NO   | MUL | NULL    |       |
| id_lieu         | int(11)       | NO   | MUL | NULL    |       |
| Productivite    | decimal(20,18)| NO   |     | NULL    |       |
| id_productivity | int(11)       | NO   | PRI | NULL    |       |
+-----------------+---------------+------+-----+---------+-------+
9 rows in set (0.01 sec)

mysql> select count(*) from productivite;
+----------+
| count(*) |
+----------+
|     1122 |
+----------+
1 row in set (0.01 sec)
```

```
mysql> desc criminalite;
+-----------+----------------+------+-----+---------+-------+
| Field     | Type           | Null | Key | Default | Extra |
+-----------+----------------+------+-----+---------+-------+
| Code      | varchar(8)     | NO   |     | NULL    |       |
| Country   | varchar(32)    | NO   |     | NULL    |       |
| Homicides | decimal(21,19) | NO   |     | NULL    |       |
| id_crim   | int(11)        | NO   | PRI | NULL    |       |
| suicides  | decimal(21,19) | NO   |     | NULL    |       |
| year      | int(11)        | NO   |     | NULL    |       |
+-----------+----------------+------+-----+---------+-------+
6 rows in set (0.01 sec)

mysql> select count(*) from criminalite;
+----------+
| count(*) |
+----------+
|     5488 |
+----------+
1 row in set (0.01 sec)
```

Premières analyses : Corrélations

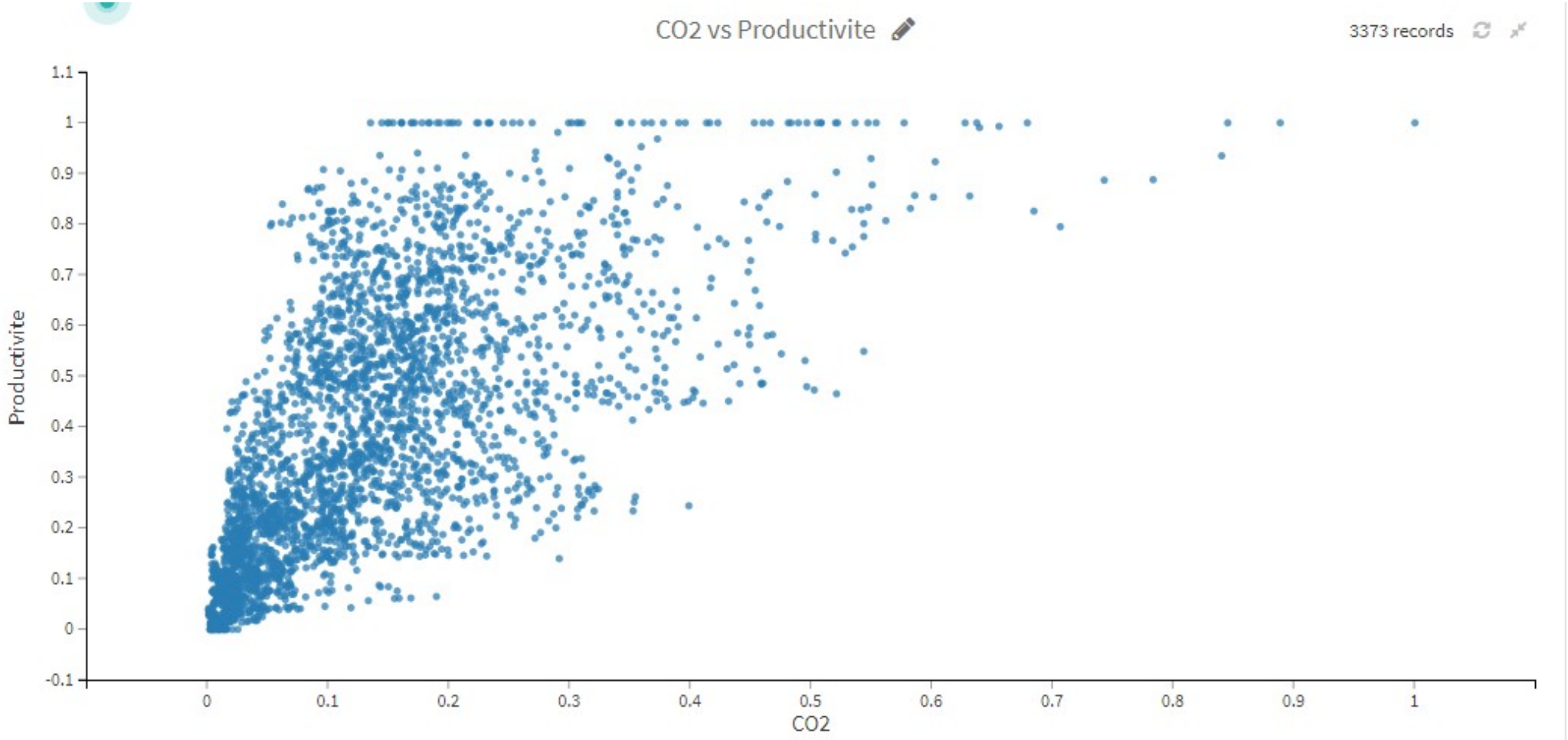
Productivité en fonction des émissions CO2 sans normalisation



MEETOOA Kevin
SAHLI Oussama

Corrélations

Productivité en fonction des émissions CO2 avec normalisation



MEETOOA Kevin
SAHLI Oussama

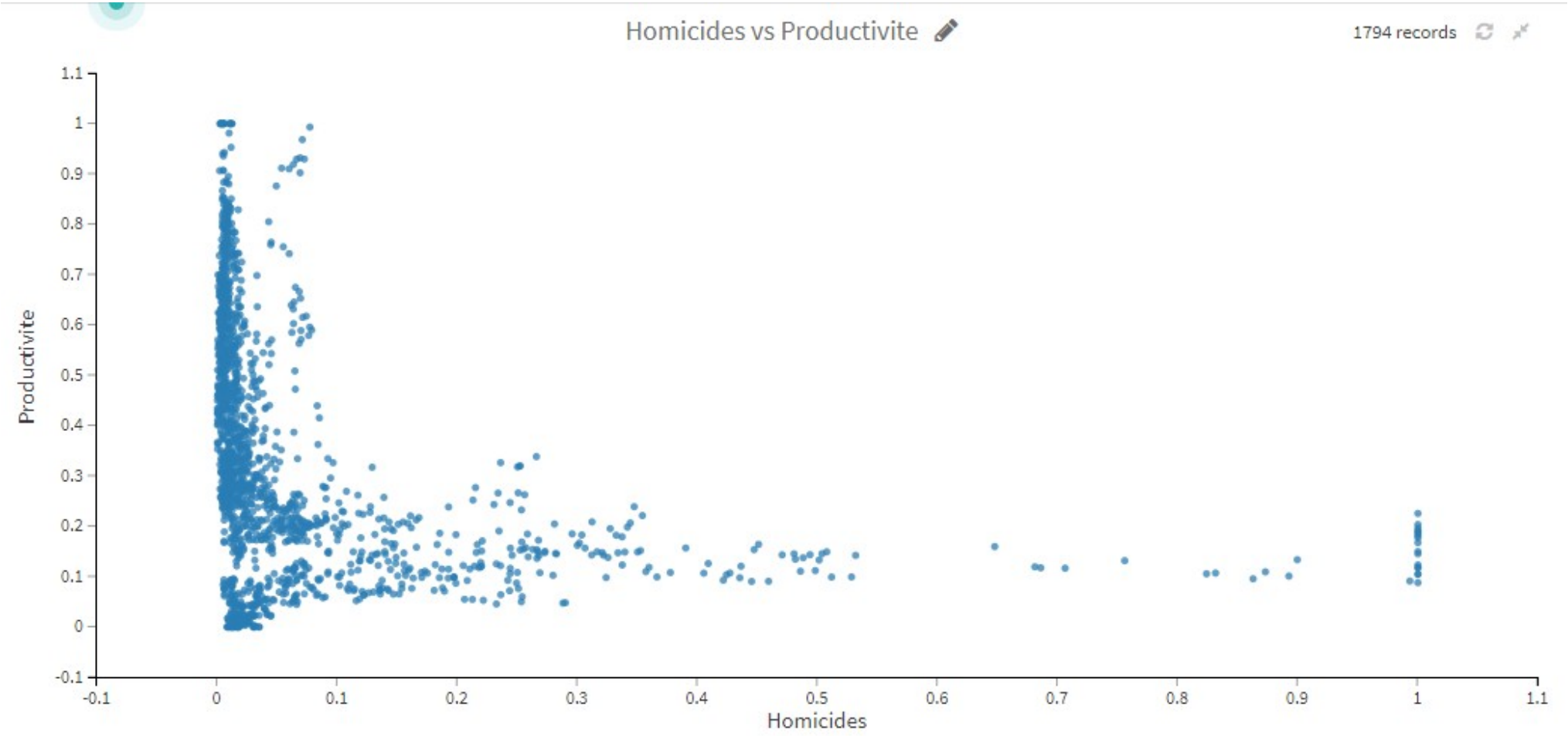
Impact de la normalisation

- La normalisation a du sens : On ne peut pas comparer des données ayant des échelles de grandeur différentes

- La normalisation est effectuée par année et pour chaque pays, afin de limiter l'impact des phénomènes de type inflation

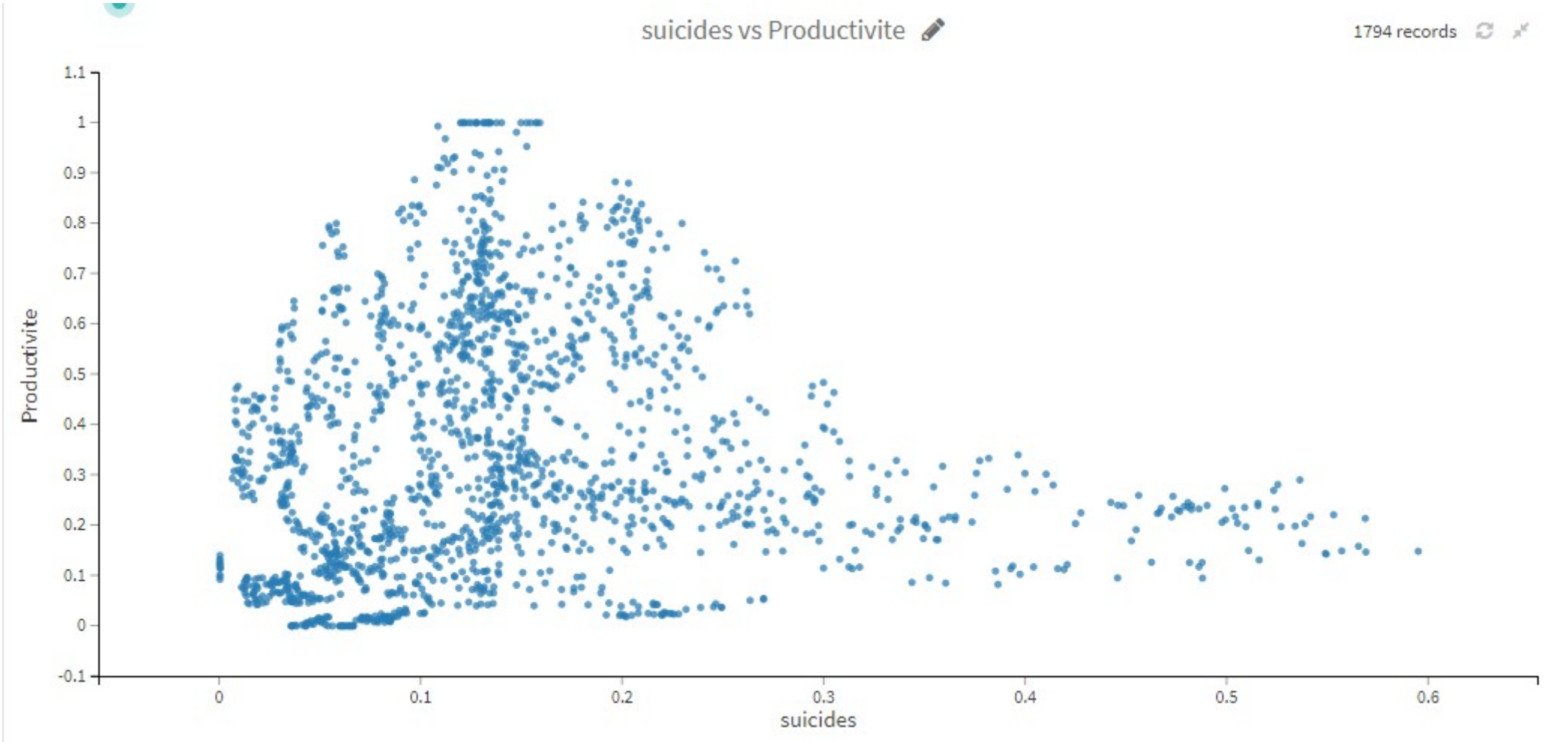
- À partir de maintenant, on travaillera uniquement avec des données normalisées

Corrélations : Productivité en fonction du nombre d'homicides



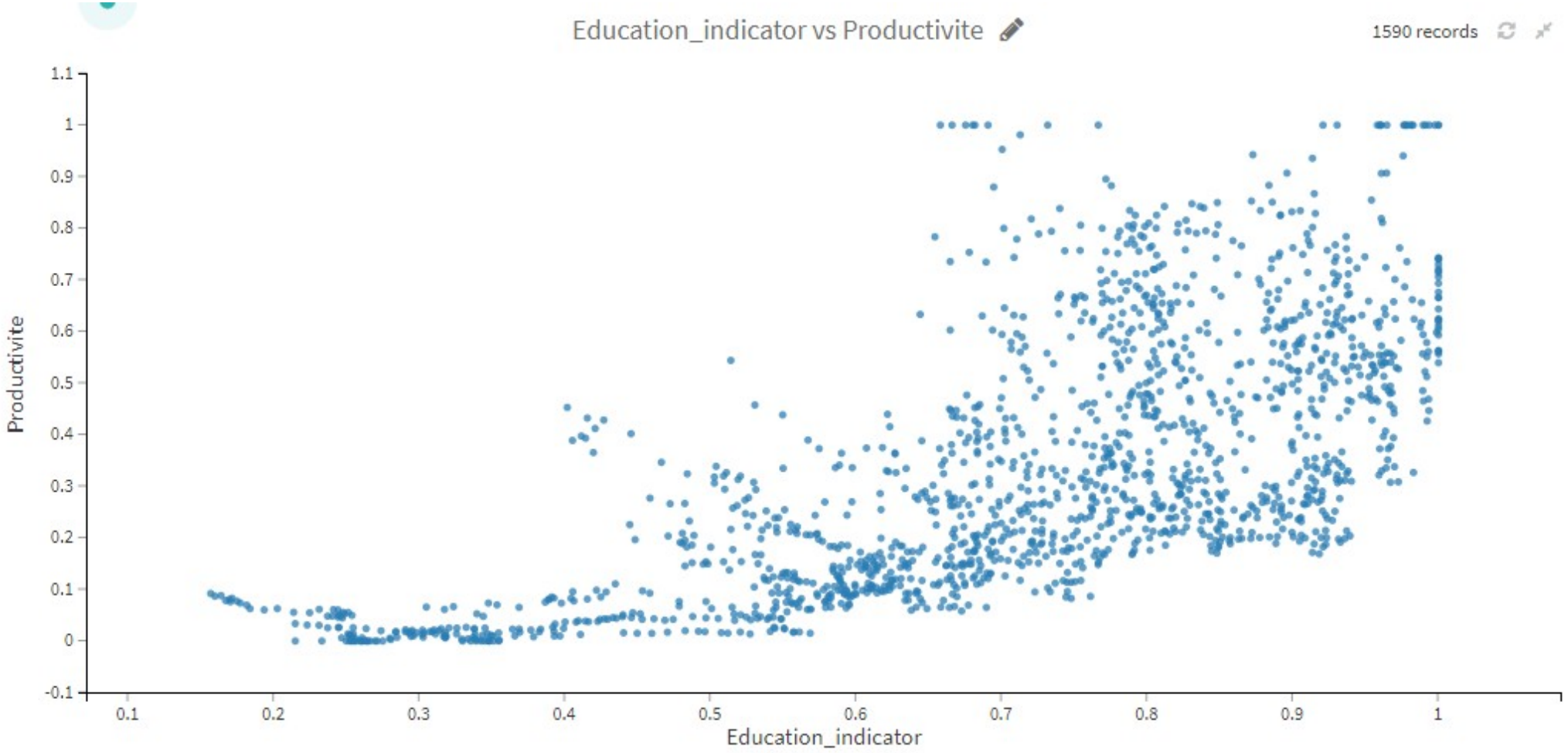
MEETOOA Kevin
SAHLI Oussama

Corrélations : Productivité en fonction du nombre de suicides



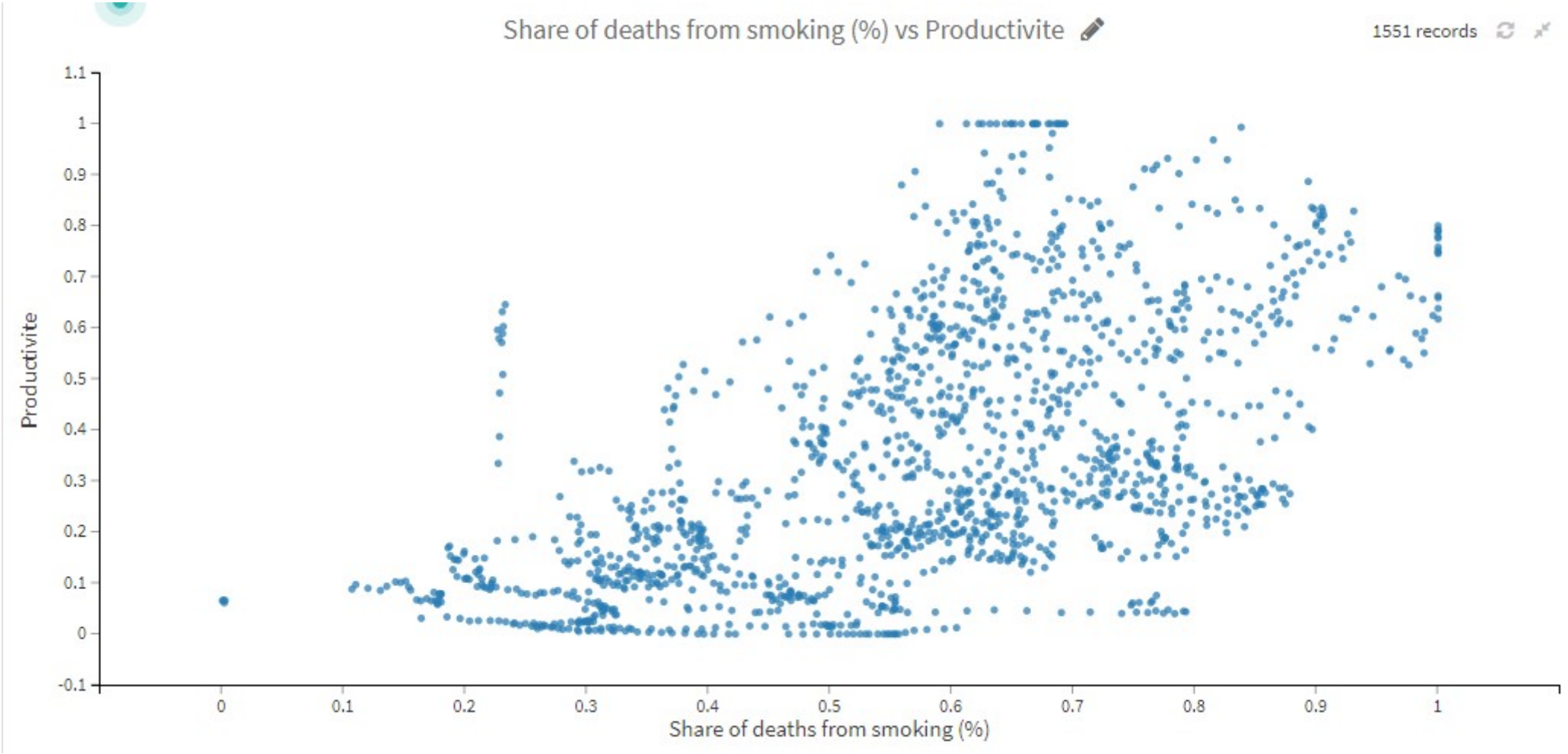
MEETOOA Kevin
SAHLI Oussama

Corrélations : Productivité en fonction de l'index d'éducation



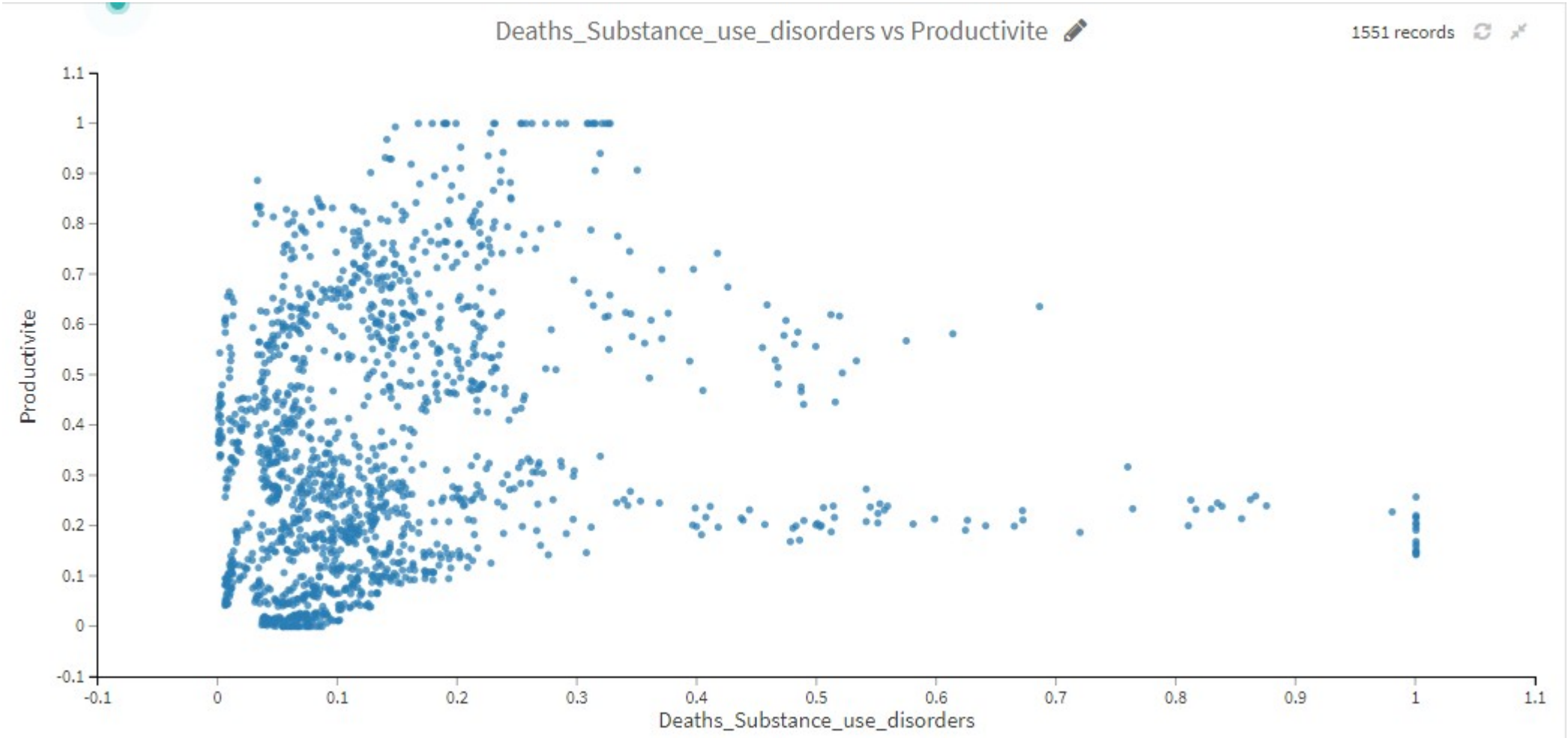
MEETOOA Kevin
SAHLI Oussama

Corrélations : Productivité en fonction du nombre de décès liés au tabagisme



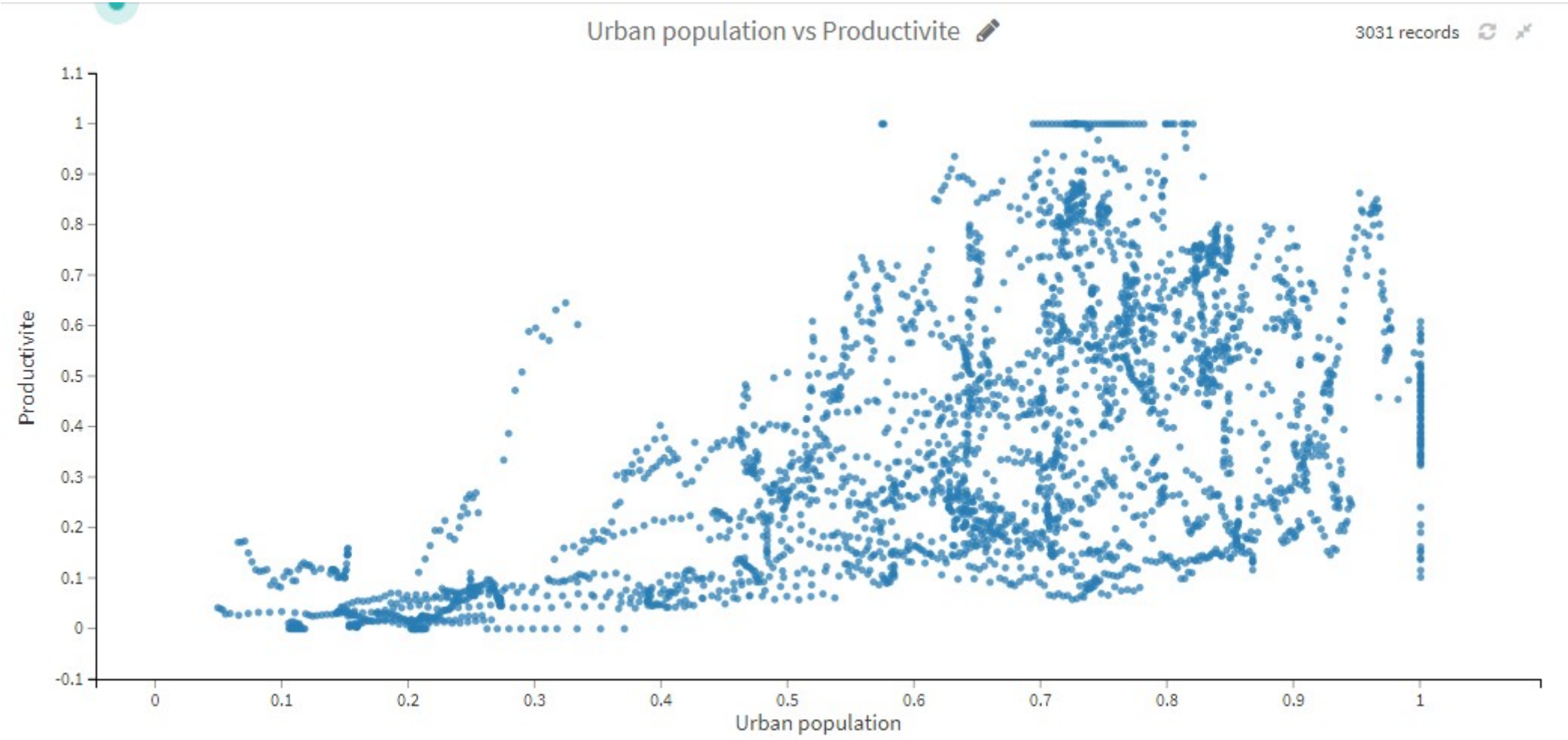
MEETOOA Kevin
SAHLI Oussama

Corrélations : Productivité en fonction du nombre de décès liés à la drogue



MEETOOA Kevin
SAHLI Oussama

Corrélations : Productivité en fonction de la population urbaine



MEETOOA Kevin
SAHLI Oussama

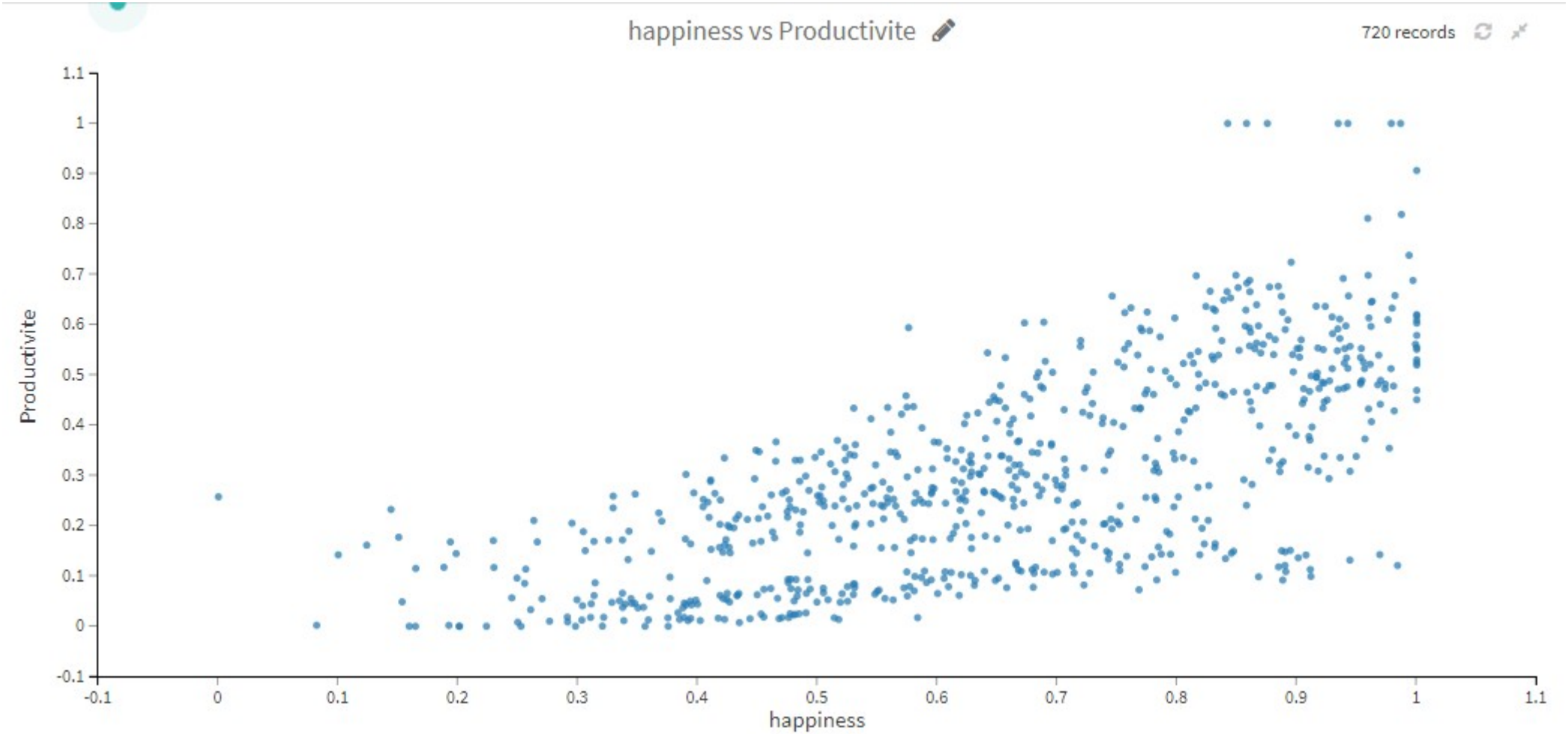
Corrélations : Productivité en fonction de l'espérance de vie

MEETOOA Kevin
SAHLI Oussama

Corrélations : Productivité en fonction du taux d'obésité

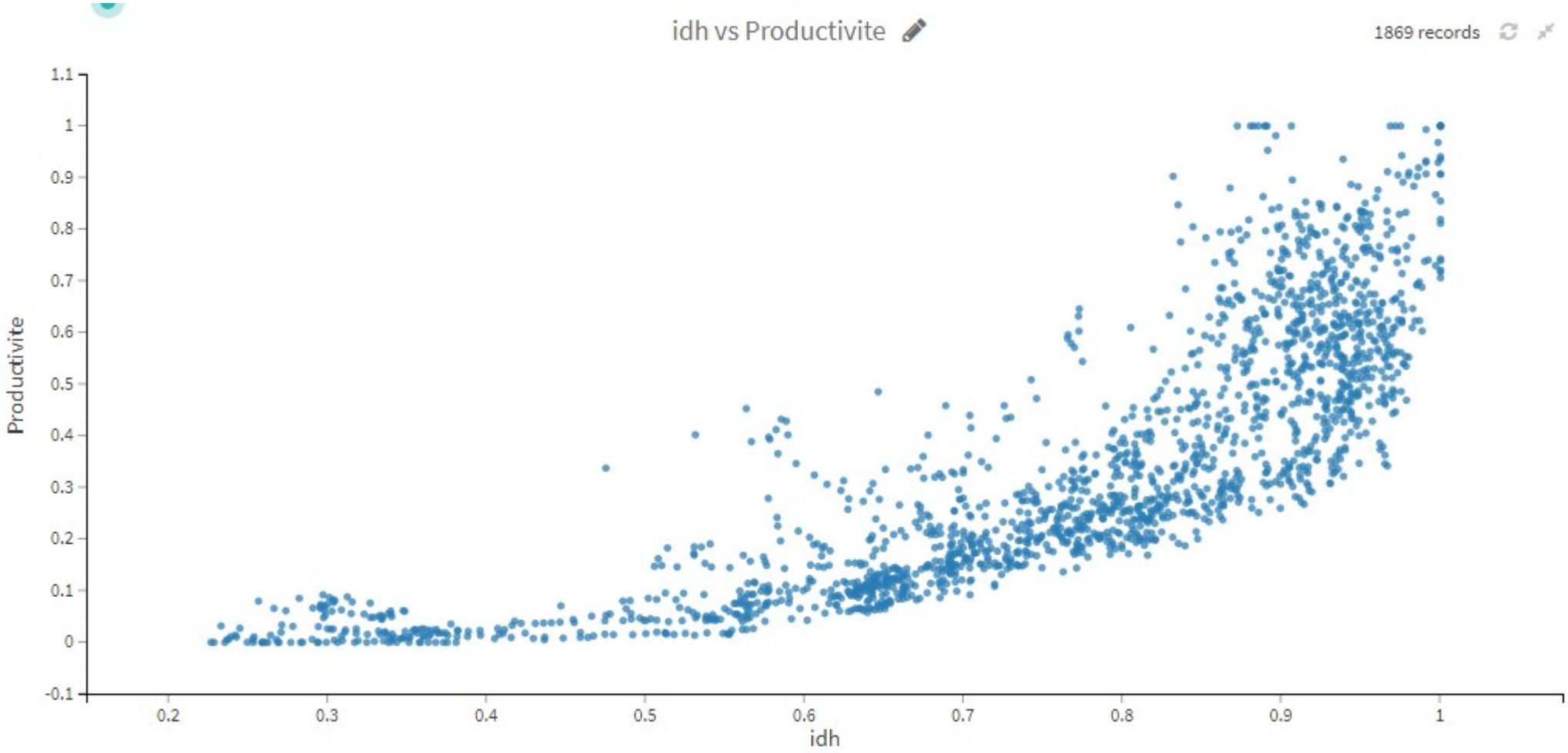
MEETOOA Kevin
SAHLI Oussama

Corrélations : Productivité en fonction du bonheur



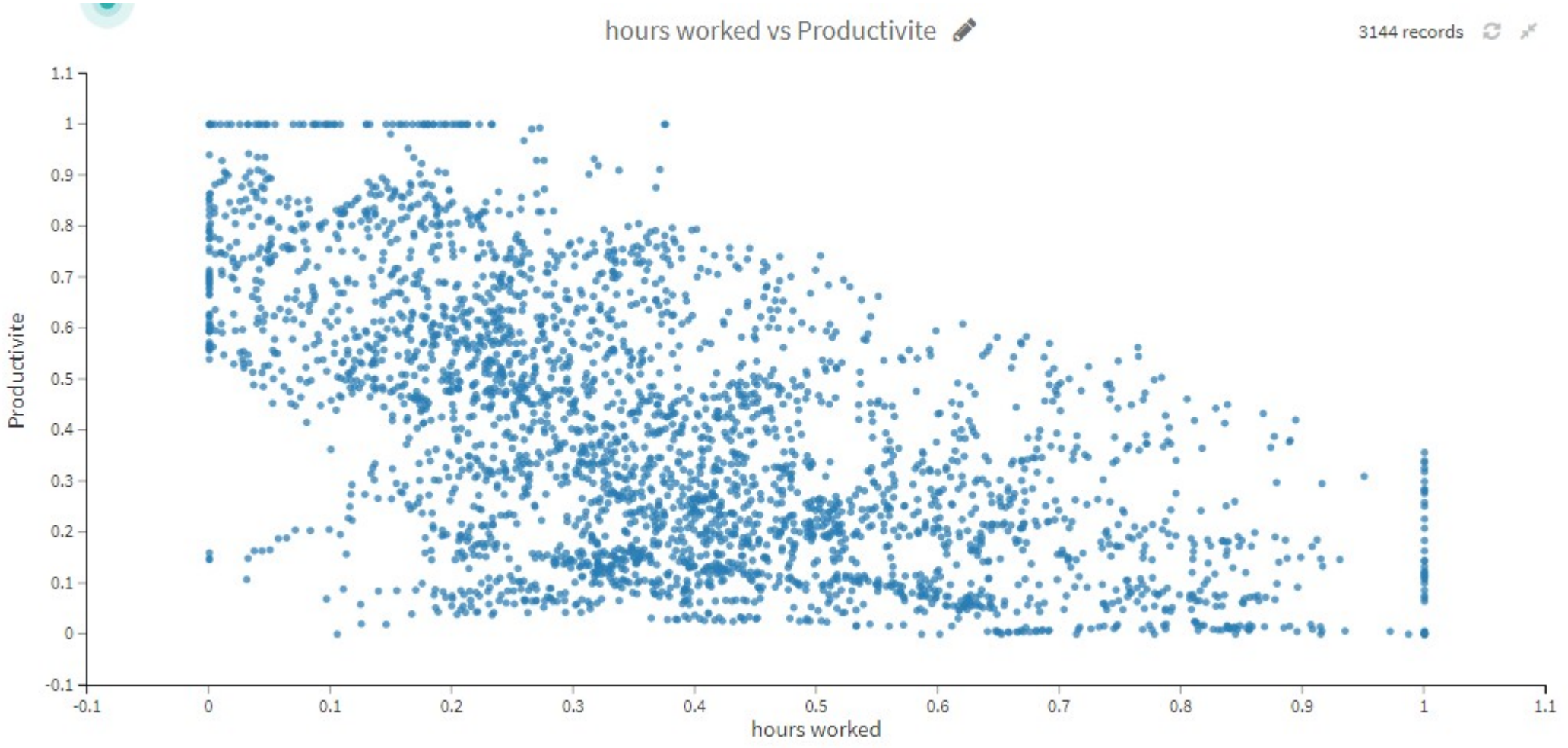
MEETOOA Kevin
SAHLI Oussama

Corrélations : Productivité en fonction de l'IDH



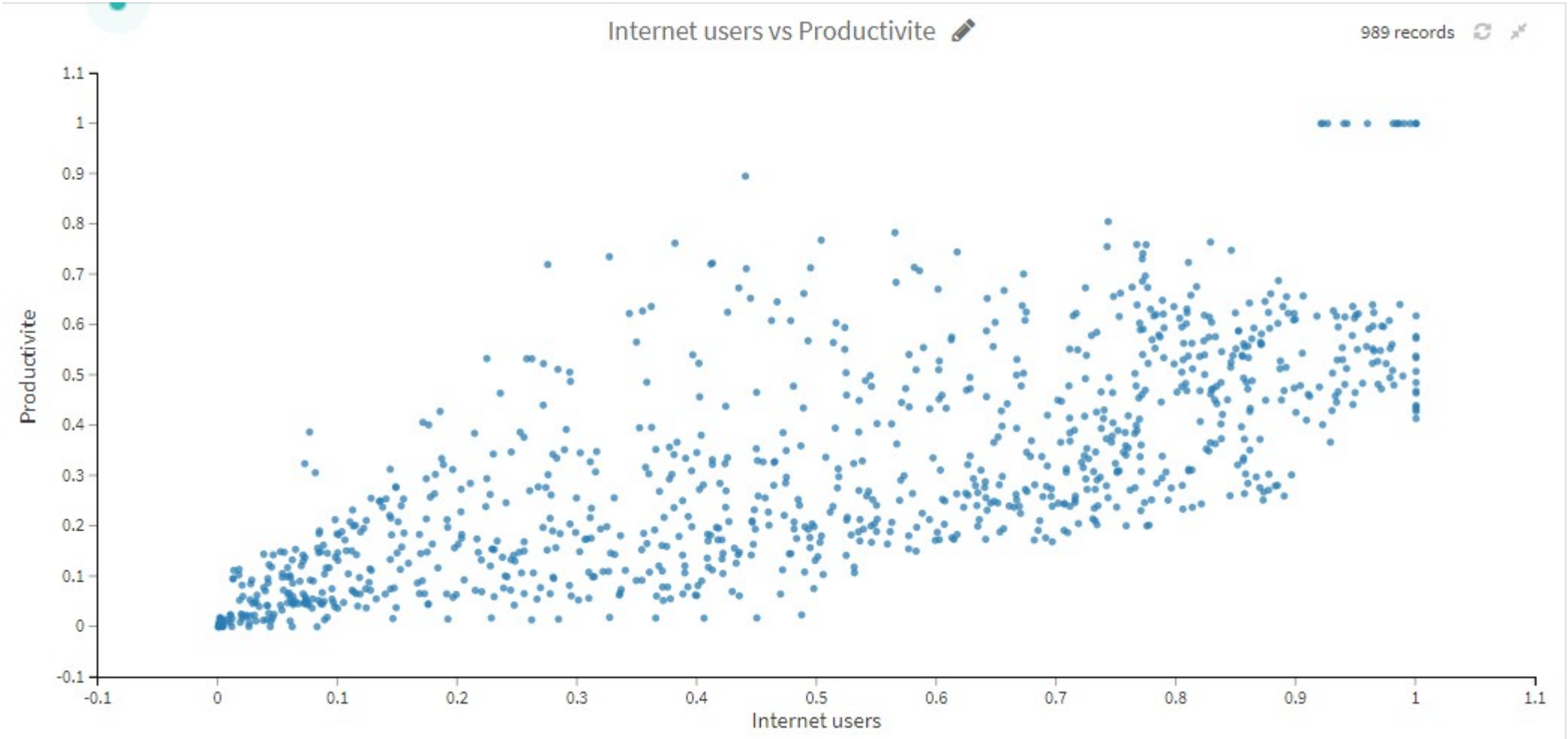
MEETOOA Kevin
SAHLI Oussama

Corrélations : Productivité en fonction du nombre d'heures travaillées



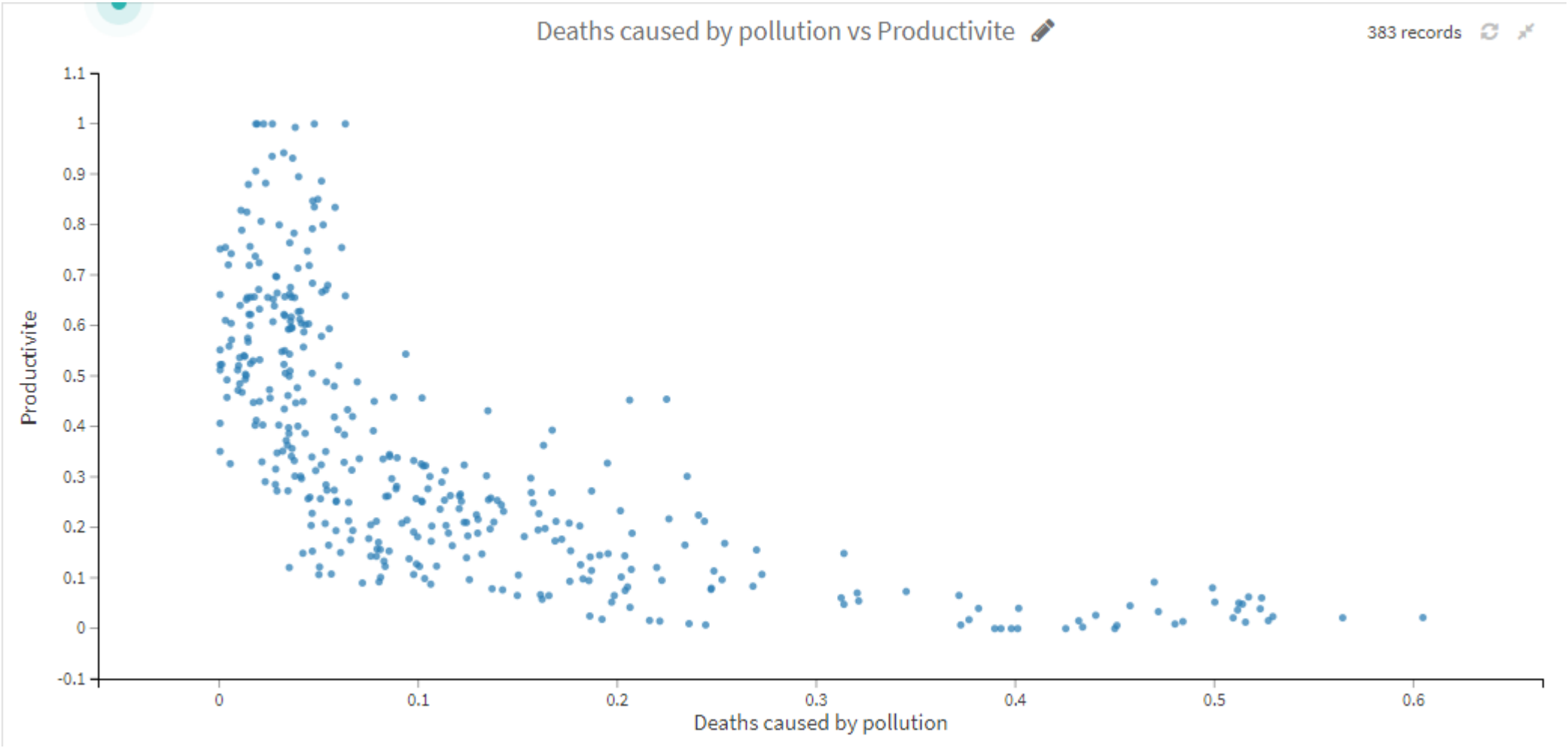
MEETOOA Kevin
SAHLI Oussama

Corrélations : Productivité en fonction de la proportion d'utilisateurs d'internet



MEETOOA Kevin
SAHLI Oussama

Corrélations : Productivité en fonction du nombre de décès liés à la pollution



MEETOOA Kevin
SAHLI Oussama

Récapitulatif : Coefficients de corrélation

Variable	Corrélation avec la productivité
IDH	0.799
Proportion d'utilisateurs d'internet	0.729
Espérance de vie	0.689
Bonheur	0.689
Émissions de CO2	0.688
Éducation	0.672
Décès liés à la pollution	-0.662
Nombre d'heures travaillées	-0.594
Urbanisation	0.583
Alcool	0.539
Homicides	0.3
Obésité	0.28
Nombre de suicides	0.04

Autre analyse : Groupement des pays (ce qu'on aurait voulu faire)

- On peut essayer de faire des analyses plus fines en regroupant les pays selon certains critères (économiques par exemple)

- On peut essayer de dégager les causes principales de la productivité pour chacun des groupes

- Intérêt : Par exemple, les causes de la productivité peuvent être différentes dans les pays qui sont développés (au sens de l'IDH) et ceux qui ne le sont pas

Suite de l’analyse

- Une fois les critères trouvés, on peut essayer d’analyser chacun des critères (évolution temporelle, dépendance avec d’autres critères…)

- On peut ensuite espérer avoir une réponse au moins partielle à notre problématique, même si la réponse est différente selon le groupement de pays que l’on regarde.

Problème

- Parmi les critères trouvés, il faut savoir si pour un critère donné, le critère est une cause de productivité ou une conséquence de la productivité…

- Exemple : On peut penser qu'une importante proportion d'utilisateurs d'internet est une conséquence d'une grande productivité et non une cause

Conclusion

- Critères les plus importants : IDH, bonheur, éducation, nombre d'heures travaillées...

- Les critères retenus sont assez « classiques », il aurait pu être intéressant de trouver des critères « inattendus » mais nous n'avons pas réussi à trouver des données intéressantes

- Exemple : Nous n'avons pas réussi à trouver des données par pays pour le temps passé devant la télé/les séries télé/les jeux vidéos