Recomendando Papers - Análisis y Visualización

Integrantes: Juliana Benitez, Leonardo Latini

Dataset

El dataset está compuesto por metadata de papers subidos a la pagina https://arxiv.org de las categorias Matemática, Física y Computación, entre los años 2017 y 2018. Este dataset cuenta con la metadata de 229448 publicaciones y fue creado utilizando OAI protocol for metadata harvesting (OAI-PMH), que permite acceder a la metadata de los articulos publicados en Arxiv. El dataset está provisto en formato .csv, separado por ';' (punto y coma).

Está formado por las siguientes columnas:¶

- Id [object]: Identificador unico del artículo
- Title [object]: Título del artículo
- Abstract [object]: Resumen del artículo
- Fields [object]: Áreas a las que pertenece el artículo (math, physics or cs).
- Categories [object]: Categorias o tags de la publicación (para mas referencias ver: https://arxiv.org/, https://arxiv.org/archive/math, https://arxiv.org/corr/subjectclasses
- Authors [object]: Apellido y nombre de los autores del artículo, separados por coma.
- Doi [object] (Digital Object Identifier) es una forma de identificar un objeto digital.
- Journal [object]: Journal donde fue publicado el artículo (En caso que no sea un pre-print).
- Created [datetime64[ns]]: Fecha en que fue subida la primera versión del artículo.
- Year [int64]: Año en que fue subida la primera versión del artículo.
- Month [int64]: Mes en que fue subida la primera versión del artículo.
- Day [int64]: Dia en que fue subida la primera versión del artículo.
- Abstract_Length [int64]: Cantidad de caracteres en el abstract.
- Title_Length [int64]: Cantidad de caracteres en el título.
- Number_Authors [int64]: Cantidad de autores.
- Number_Fields [int64]: Número de Áreas.
- Number_Categories [int64]: Número de Categorias.

1.Cantidad de papers publicados por día y por mes. Calcular estadísticos como media, moda y mediana

Media de papers publicados por día:

```
7401.56
```

Mediana de papers publicados por día:

```
7460.0
```

Moda de papers publicados por día:

```
0     4578
1     6751
2     7019
3     7072
4     7106
5     7167
```

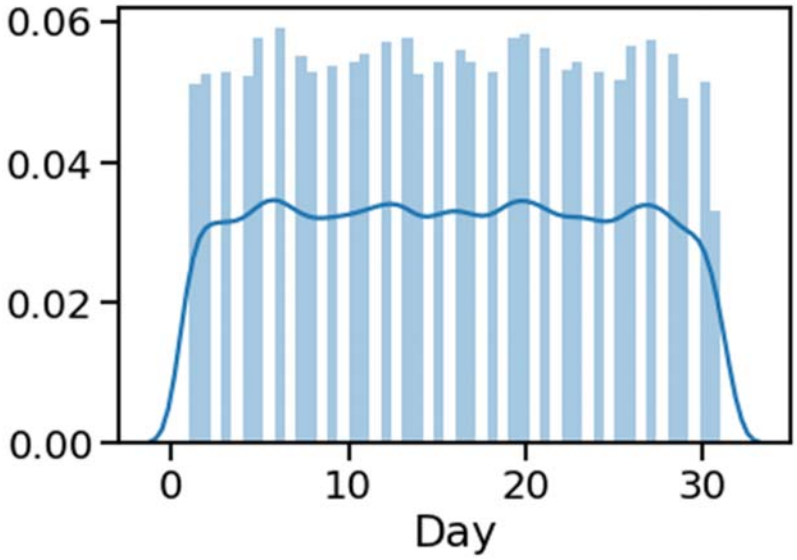
```
6     7221
7     7227
8     7250
9     7272
10    7276
11    7282
12    7302
13    7398
14    7448
15    7460
16    7463
17    7473
18    7563
19    7608
20    7629
21    7706
22    7726
23    7775
24    7844
25    7900
26    7924
27    7934
28    7937
29    8018
30    8119
```
Hay múltiples modas para esta variable

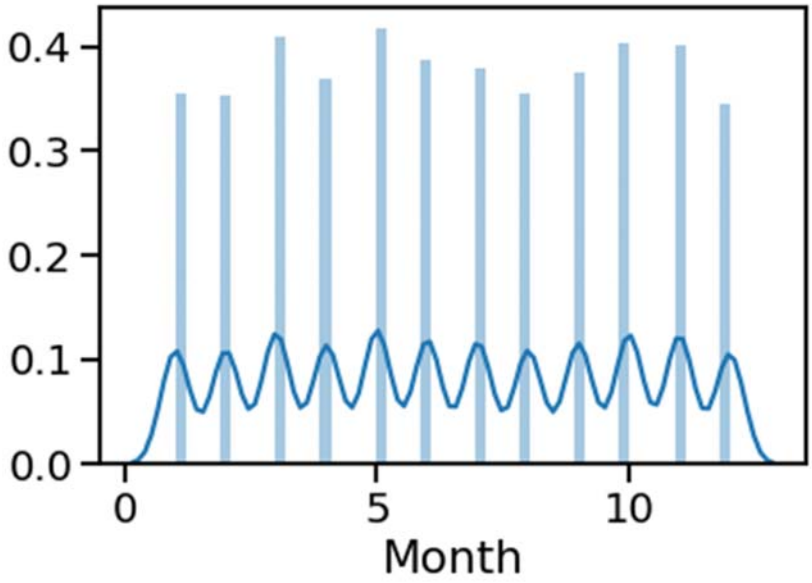
Media de papers publicados por mes:

```
19120.666666666668
```
Mediana de papers publicados por mes:

```
19059.5
```
Moda de papers publicados por mes:

```
0     17358
1     17841
2     17860
3     17875
4     18646
5     18967
6     19152
7     19524
8     20201
9     20379
10    20607
11    21038
```
Hay múltiples modas para esta variable





En líneas generales se aprecia una distribución prácticamente uniforme a lo largo de los meses para los 2 años incluidos en el dataset. Las cantidades de papers que se crean mensualmente rondan entre un mínimo de 17358 y un máximo de 21038. Analizando a nivel de días, vemos que la creación de papers desde los días 1 al 30 se asemeja a una distribución uniforme, rondando entre los 6751 y los 8119. Y en los dias 31 disminuye a valores de 4578, lo cuales son bastante más bajos que los observados en los restantes días.

2.

Calcular estadísticos como la moda, media, mediana y desviación estándar de las variables 'Title_Length' y 'Number_Authors'. ¿Responden a alguna distribución conocida?¿ Existen outliers?

Media de variable 'Title_Length':

`75.14`

Mediana de variable 'Title_Length':

`72.0`

Moda de variable 'Title_Length':

`71`

Desvío estándar de variable 'Title_Length':

`27.59`

Moda, mediana y media no coinciden en la variable Longitud del título, con lo cual para esta variable hay asimetría positiva.

Media de variable ‘Number_Authors':

`4.87321310275095`

Mediana de variable ‘Number_Authors':

`3.0`

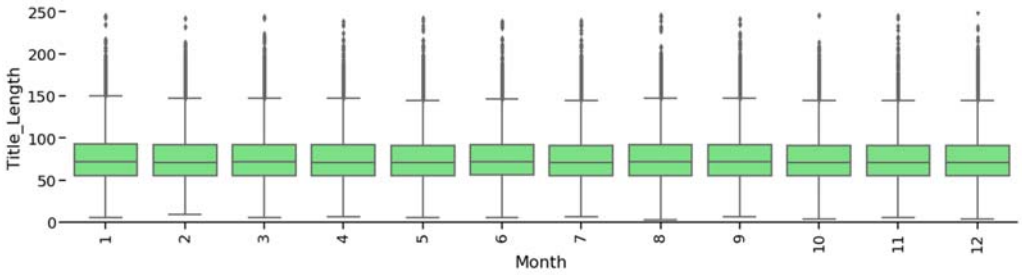
Moda de variable ‘Number_Authors':

`2`

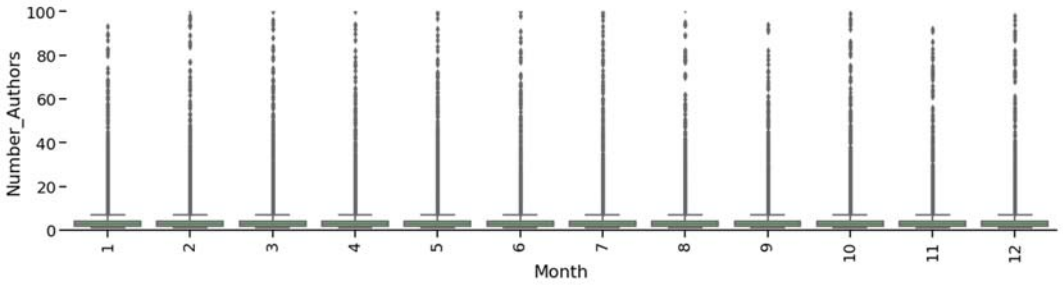
Desvío estándar de variable ’Number_Authors':

`25.44`

Lo mismo ocurre para esta variable, Número de autores, donde la distribución muestra una asimetría positiva.

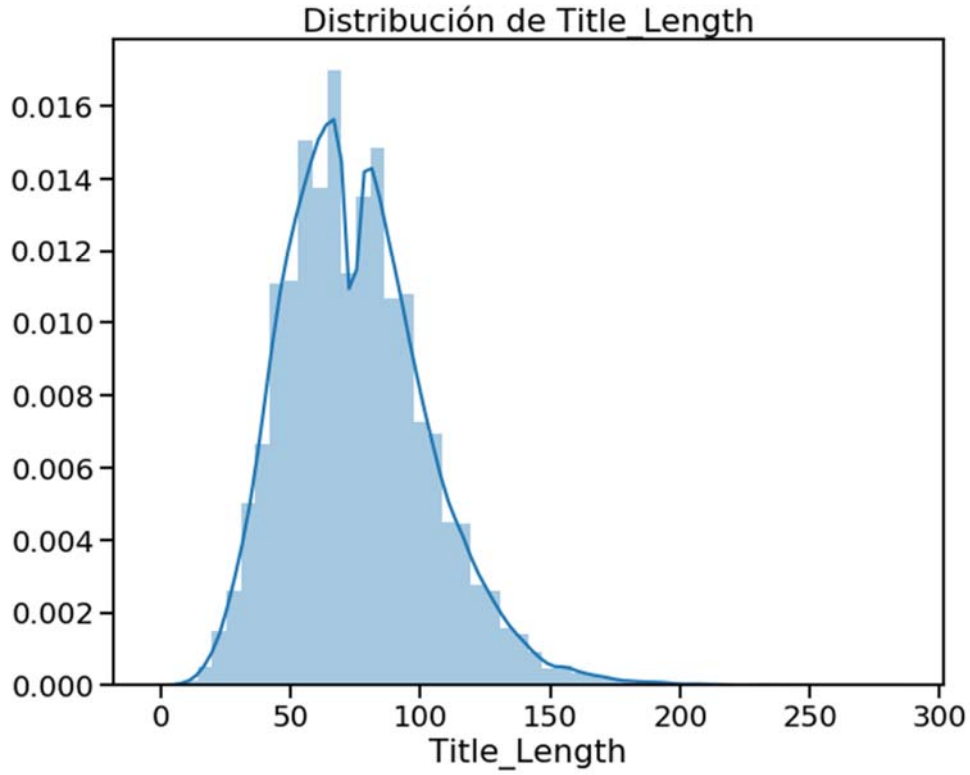


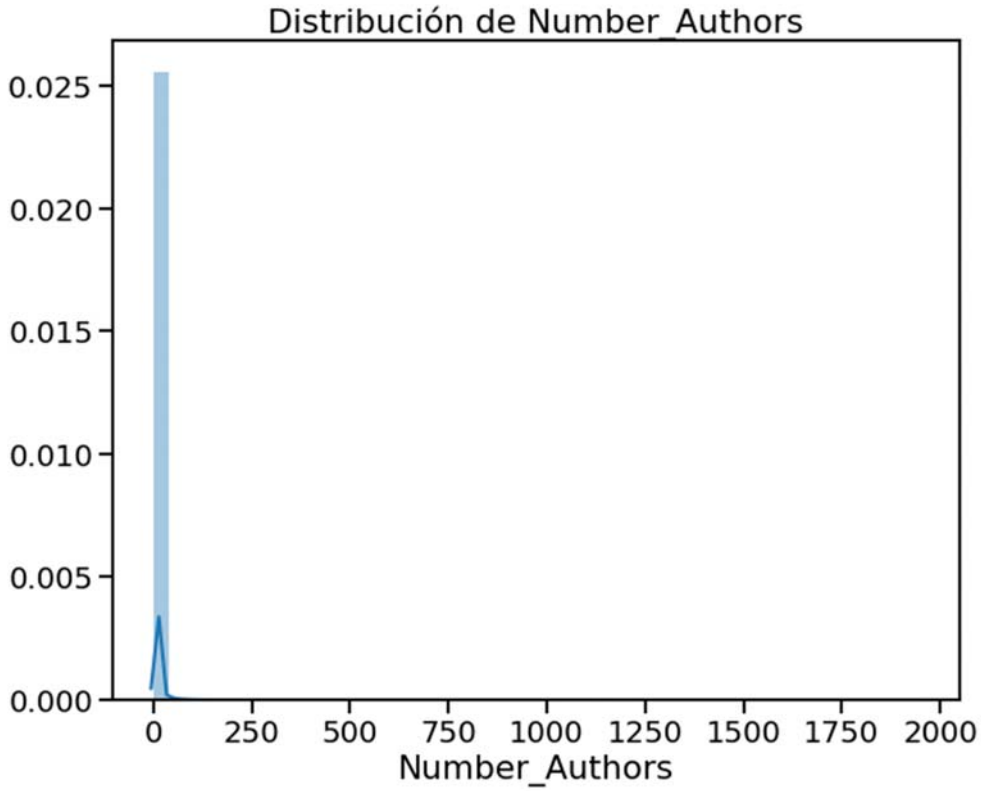
Con el grafico de boxplot se ve más claramente la mediana de longitud de título y que en todos los meses, los outliers rondan más o menos los mismos valores.



La mediana de autores es similar para los meses del año, pero hay muchos puntos de outliers

La distribución de Title Length se asemeja a una bimodal, en un test como shapiro wilk, se confirma que la distribución no es normal (aunque para volúmenes de datos tan grandes como estos el test no es apto para confirmar normalidad).



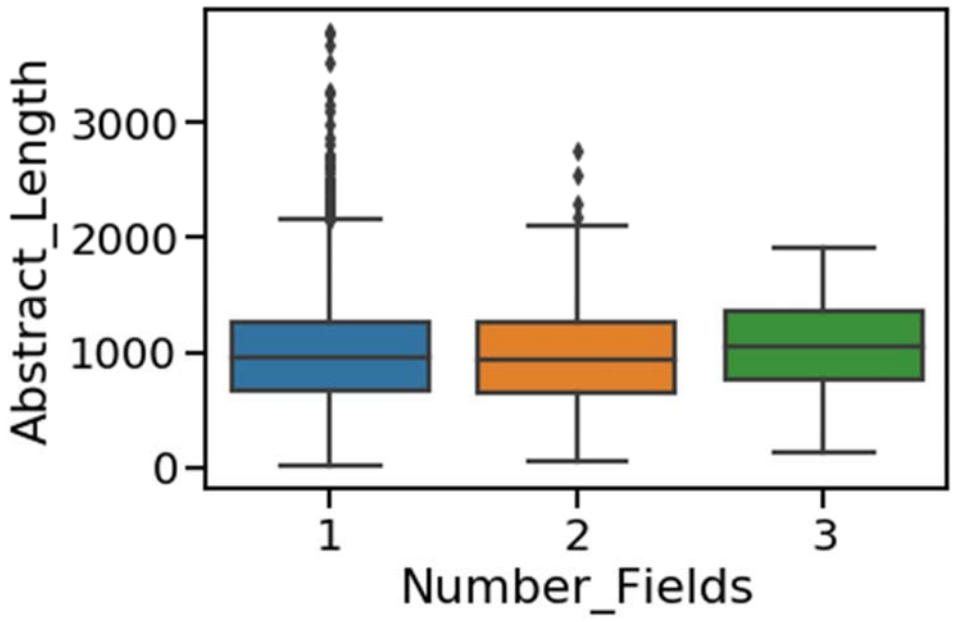


En este caso, la distribución del número de autores muestra una distribución muy asimétrica, con alta densidad en un bajo número de autores, pero hay casos de número total de autores que está en el rango de 200-1950, lo cual puede ser común en estas disciplinas estudiadas, pero no en otras ramas de la ciencia. Estos casos de altos números de autores, que son estadísticamente outliers, pero observando el set de datos se ve que corresponden a colaboraciones masivas en proyectos por ejemplo del CERN.

La longitud de los títulos de los papers tiene un rango de 277, que va desde longitud 3 hasta longitud 280. Presenta una longitud promedio de 75.137443 y la longitud que más se repite es la longitud de 71.

Según se muestra gráficamente se observan outliers en esta variable.

De todos modos, los outliers de longitud en cada uno de los meses son similares. El número de autores de los papers nos presenta un promedio de 4.8732 siendo 2 la cantidad de autores que más se repite.



En este grafico se aprecia que, en las tres disciplinas, la mediana de longitud del abstract (cuantificada como número de caracteres) es similar, y esto nos sugiere que se relaciona con las condiciones o guías que las revistas indican para el formato para la publicación del artículo.

3.

Dividan en 4 partes el dataset y vayan calculando bayes con respecto a 2 variables aleatorias, usando los resultados de cada iteración / partición para calcular. El objetivo es simular que los datos que van llegando en cada iteración recalculan la probabilidad.

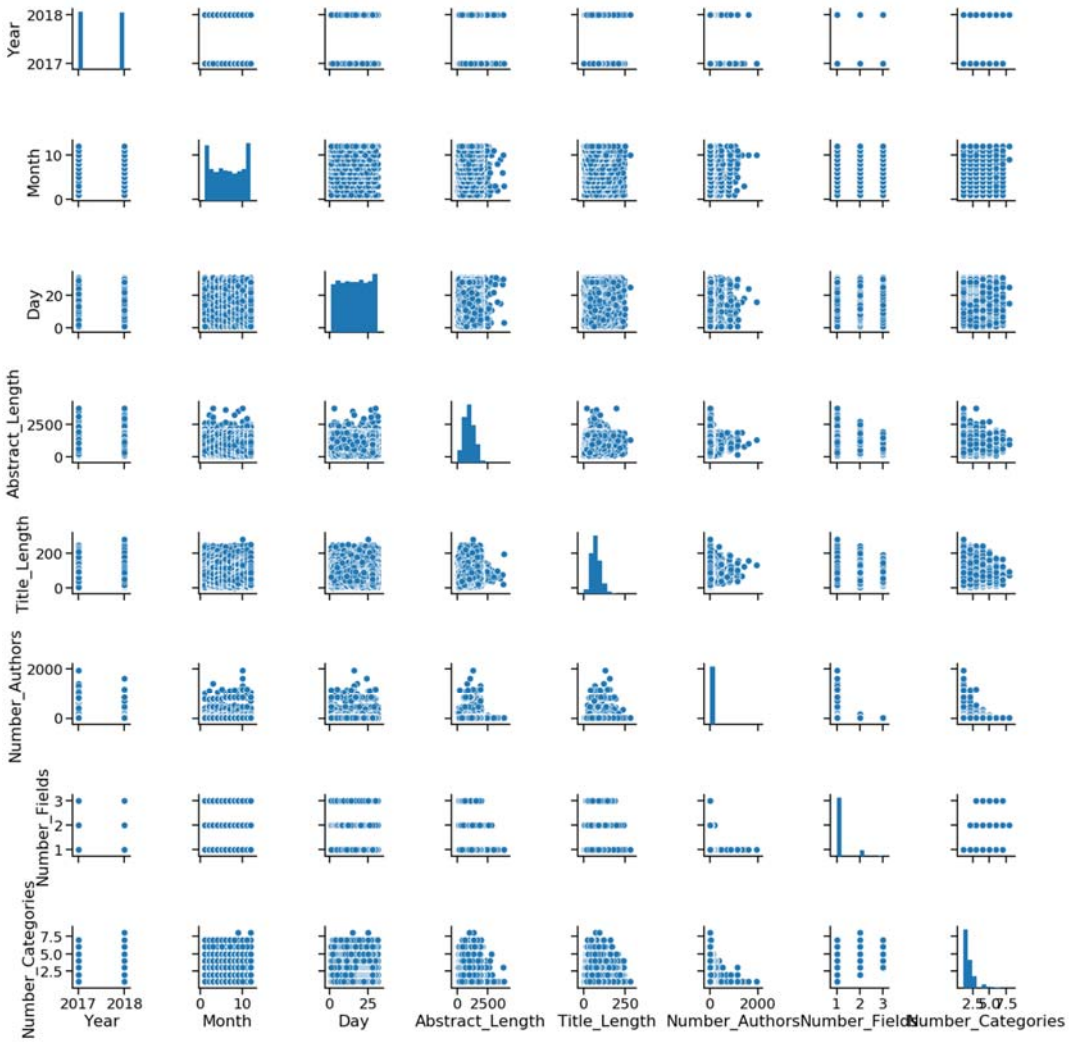
```
from scipy import stats

##estas variables aleatorias son la media y el desvio estandar del numero de autores

#dataset1 = numpy.random.normal(4.873213, 25.445353, size=int(dataset.shape[0]/4))

print(int(dataset1.shape[0]/4))

loc, scale= stats.norm.fit(dataset1)

print('media', loc, ' - ds ', scale)


n= stats.norm(loc = loc, scale = scale)

n.rvs(4)
```

```
14340
media 4.727762253279168  - ds  25.485537810527195
array([ 17.90590857, -35.24792936,   6.44041158, -14.76086791])
```

4.

Análisis de cuáles son las variables más correlacionadas. (Considerar solo las variables numéricas)

Creamos subset de las variables estrictamente numéricas (aunque incluyan por separado a los componentes del elemento ‘Created’ (fechas en formato datetime64))



Utilizando la función pairplot para explorar los datos que son variables numéricas puede observarse:

* Para los dos años registrados en la base de datos el número de papers subidos al sitio es muy similares (en ambos el número de artículos es superior a 110000)

* Entre el grafico que genera pairplot y el histograma hay una discrepancia; en el pairplot muestra los mayores valores en mes 1 y 12 (enero y diciembre respectivamente), cuando en el histograma la cantidad de archivos subido no varía tanto de mes a mes (aunque si se nota que marzo y mayo tiene mayores valores que otros meses)

* Hay casos de números de autores altos, de más de 200 autores, incluso hay uno de 1945 autores.

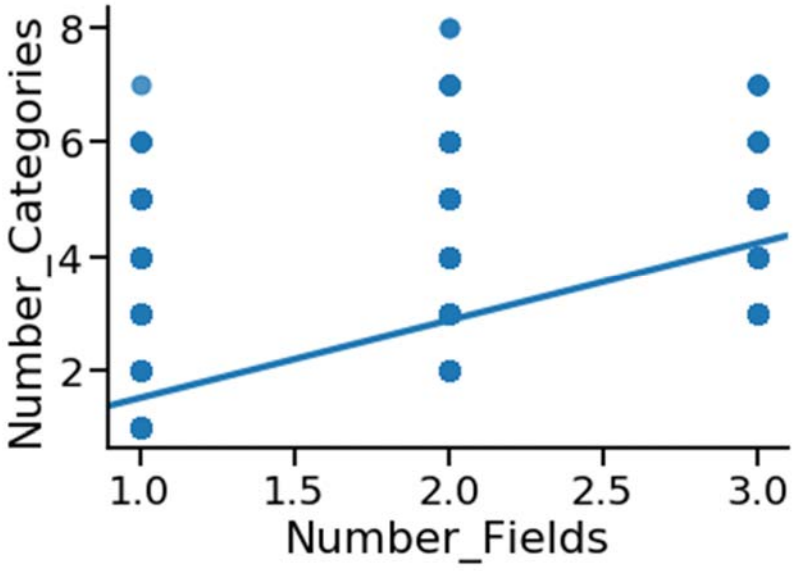
* Dado que solo hay tres posibles campos (Matemática, Física y Ciencia de la Computación), las relaciones con otras variables no parecen ser en ningún caso informativas

* Abstract length y title length podrían llegar a tener una distribución normal

	Year	Month	Day	Abstract_Length	Title_Length	Number_Authors	Number_Fields	Number_Categories
Year	1	-0.0410763	-0.00734567	0.0192596	0.00573531	-0.00607989	-0.0153083	0.013934
Month	-0.0410763	1	-0.0178859	-0.000688231	-0.00507148	0.00201909	-0.0152933	0.00793939
Day	-0.00734567	-0.0178859	1	-0.00287909	-0.00476983	-0.00217966	0.000102479	0.00407856
Abstract_Length	0.0192596	-0.000688231	-0.00287909	1	0.21648	0.0430795	-0.00116475	0.0435823
Title_Length	0.00573531	-0.00507148	-0.00476983	0.21648	1	0.0556969	-0.0239659	-0.034807
Number_Authors	-0.00607989	0.00201909	-0.00217966	0.0430795	0.0556969	1	-0.0256061	-0.0191926
Number_Fields	-0.0153083	-0.0152933	0.000102479	-0.00116475	-0.0239659	-0.0256061	1	0.454287
Number_Categories	0.013934	0.00793939	0.00407856	0.0435823	-0.034807	-0.0191926	0.454287	1

Con esta matriz se ve mejor las correlaciones: principalmente entre las variables **Abstract_Length** - **Title_Length** y **Number_Categories**-**Number_Fields**. Probablemente estas correlaciones se hacen con Pearson; creemos que sería más correcto usar correlaciones no paramétricas como Spearman.

Correlación a nivel visualización de las variables **Number_Fields** y **Number_Categories.**



Dadas las características de las variables número de categorías y número de campos de estudio, que son valores numéricos enteros, el grafico y la correlación puede indicarse como significativa, aunque no es suficientemente informativo o relevante.

5.

Calcular la probabilidad marginal y conjunta, y la correlación entre otras dos variables, por ejemplo 'Number_fields' y ' Number_Categories'. Representar visualmente la probabilidad conjunta entre los valores posibles de las variables elegidas.

Number_Categories	1	2	3	4	5	6	7	8	All
Number_Fields									
1	130897	53152	18489	4506	915	27	1	0	207987
2	0	9425	6377	3374	1351	301	30	2	20860
3	0	0	262	166	126	36	11	0	601
All	130897	62577	25128	8046	2392	364	42	2	229448

Probabilidad marginal para la variable Number_Categories

```
probabilidad de pertenecer a una categoría=  0.5705
probabilidad de pertenecer a dos categorías=  0.2727
```

probabilidad de pertenecer a tres categorias= 0.1095
probabilidad de pertenecer a cuatro categorias= 0.0351
probabilidad de pertenecer a cinco categorias= 0.0104
probabilidad de pertenecer a seis categorias= 0.0016
probabilidad de pertenecer a siete categorias= 0.0002
probabilidad de pertenecer a ocho categorias= 0.0

probabilidad marginal para la variable Number_Fields

probabilidad de pertenecer a un campo de estudio= 0.9065
probabilidad de pertenecer a dos campos de estudio= 0.0909
probabilidad de pertenecer a tres campos de estudio= 0.0026

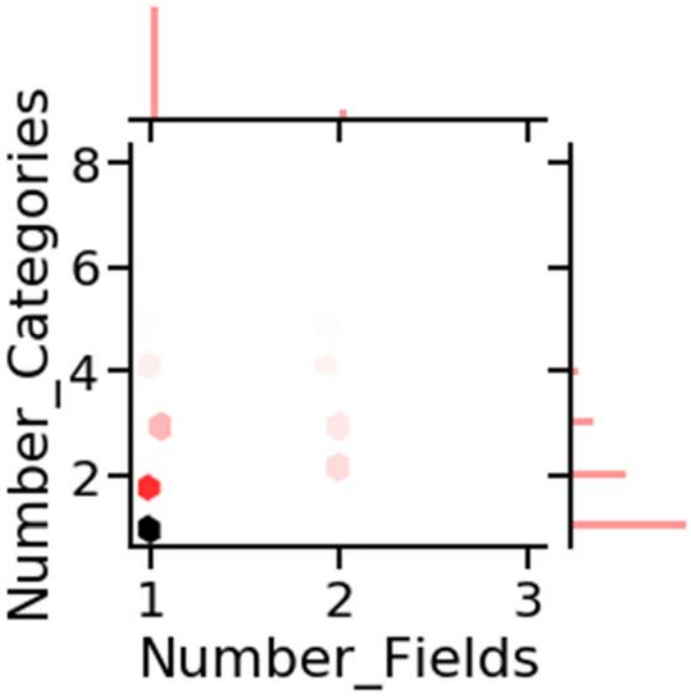
Probabilidades conjuntas

probabilidad de pertenecer a un campo de estudio y una categoría= 0.5705
probabilidad de pertenecer a un campo de estudio y dos categorías= 0.2317
probabilidad de pertenecer a un campo de estudio y tres categorías= 0.0806
probabilidad de pertenecer a un campo de estudio y cuatro categorías= 0.0196
probabilidad de pertenecer a un campo de estudio y cinco categorías= 0.004
probabilidad de pertenecer a un campo de estudio y seis categorías= 0.0001
probabilidad de pertenecer a un campo de estudio y siete categorías= 0.0
probabilidad de pertenecer a un campo de estudio y ocho categorías= 0.0

probabilidad de pertenecer a dos campos de estudio y una categoría= 0.0
probabilidad de pertenecer a dos campos de estudio y dos categorías= 0.0411
probabilidad de pertenecer a dos campos de estudio y tres categorías= 0.0278
probabilidad de pertenecer a dos campos de estudio y cuatro categorías= 0.0147
probabilidad de pertenecer a dos campos de estudio y cinco categorías= 0.0059
probabilidad de pertenecer a dos campos de estudio y seis categorías= 0.0013
probabilidad de pertenecer a dos campos de estudio y siete categorías= 0.0001
probabilidad de pertenecer a dos campos de estudio y ocho categorías= 0.0001

probabilidad de pertenecer a tres campos de estudio y una categoría= 0.0

```
probabilidad de pertenecer a tres campos de estudio y dos
categorías= 0.0
probabilidad de pertenecer a tres campos de estudio y tres
categorías= 0.0011
probabilidad de pertenecer a tres campos de estudio y cuatro
categorías= 0.0007
probabilidad de pertenecer a tres campos de estudio y cinco
categorías= 0.0005
probabilidad de pertenecer a tres campos de estudio y seis
categorías= 0.0002
probabilidad de pertenecer a tres campos de estudio y siete
categorías= 0.0
probabilidad de pertenecer a tres campos de estudio y ocho
categorías= 0.0
```



Concluimos entonces que es un set de datos bastante consistente, por lo menos en lo que respecta a la mayoría de las variables estudiadas. La excepción es la variable número de autores, donde dada la naturaleza de los campos estudiados, existen colaboraciones masivas y el número de autores es alto respecto a otras ramas de la ciencia. Por otra parte, se observa que los artículos en su mayoría, son asignados a pocas categorías dentro de los 3 grandes campos de estudio del dataset.