Philippe Prince-Tritto

EXPLORACIÓN DE DATOS JURÍDICOS Y CLASIFICACIÓN DE DOCUMENTOS CON REGRESIÓN LOGÍSTICA MULTINOMIAL

[0] INTRODUCCIÓN

Con el desarrollo de las tecnologías de la información, ahora es posible utilizar los datos para tomar decisiones informadas en diversos campos, desde el sector del automóvil hasta la medicina y el transporte. Sin embargo, la ciencia de datos ha demostrado su utilidad en áreas que antes estaban reservadas únicamente a la intervención humana. Al estudiar el lenguaje, la comunidad científica se ha interesado desde los inicios de la informática en cómo explotar los datos que contiene.

El lenguaje es un objeto ambiguo para el científico: por un lado, contiene información valiosa, por otro, esta información es difícil de explotar debido a su plasticidad y su estructura abierta. Por lo tanto, para disponer de herramientas de apoyo a la toma de decisiones que exploten los datos contenidos en el lenguaje, es necesario que estos datos se estructuren según un método formal para que puedan ser explotados por las técnicas estadísticas. En otras palabras, es necesario un tratamiento previo para dar a los distintos conceptos presentes en el lenguaje un valor numérico que pueda servir de base para los cálculos si queremos automatizar determinadas tareas a partir de los tesoros enterrados en las formulaciones del lenguaje natural.

Lo que acabamos de describir es un problema de clasificación bien conocido en un sector en particular: las ciencias jurídicas. Varias iniciativas han pretendido clasificar el lenguaje jurídico para automatizar determinadas tareas. En la presente investigación nos situamos en este contexto para verificar una sencilla hipótesis:

Si tenemos determinados conceptos identificados en los documentos jurídicos, ¿sería posible hacer que estos datos nos ayuden en las tareas de clasificación de dichos documentos?

El tema tiene un cierto interés social. En efecto, observar las correlaciones y otros vínculos estadísticos entre los datos textuales para poder identificar un tipo de documento a partir de los conceptos contenidos en un documento tiene una doble ventaja.

Por un lado, agiliza el trabajo del poder judicial, ya sea el juez, el abogado u otros actores del sector, al proporcionarles herramientas para organizar sus datos de forma automática. De este modo, el profesional del derecho puede disponer de un tiempo precioso que no dedicará a la organización de los documentos, sino al cumplimiento de los criterios de eficacia impuestos por los textos internacionales, que es un elemento clave del Estado de Derecho.

Por otra parte, el riesgo de error humano se reduce cuando se utilizan herramientas para perfeccionar el trabajo repetitivo, y los vínculos encontrados en un conjunto de datos siempre pueden explotarse más para comprobar el trabajo a priori o a posteriori, de modo que no se omita ningún elemento jurídico en un documento que se presentará en el contexto de un litigio.

El presente trabajo se sitúa en un contexto empírico en el sentido de que los datos que se analizarán proceden de la práctica jurídica, lo más cerca posible del terreno. De hecho, nos basaremos en los datos recogidos en el día a día de un bufete de abogados que se ocupa de los litigios de divorcio, en primera instancia, en el juzgado de la ciudad de México.

Es importante entender que los abogados son personas ocupadas que no pueden perder su valioso tiempo en tareas tecnológicas. En este sentido, la tecnología debe adaptarse a sus prácticas y no al revés.

El conjunto de datos que tenemos procede, por tanto, de datos recogidos con prácticas imperfectas: los documentos habrán sido fotografiados por los alumnos, con una calidad variable (borrosos, sin alinear), y luego se habrá hecho una extracción mediante técnicas de Reconocimiento Óptico de Caracteres (OCR en inglés) para que los textos sean al menos utilizables. Sin embargo, muchas palabras son ilegibles para la máquina, por lo que tuvimos que seguir trabajando en los datos para identificar los conceptos a mano, utilizando la experiencia humana.

Lo que es posible hacer en un despacho de abogados en este contexto y en un contexto de falta de tiempo, y que puede serles útil, es anonimizar los documentos para que puedan ser reutilizados como plantillas dentro del despacho. La base de datos de la que disponemos se compone, por tanto, de documentos de litigios de divorcio, cuyos datos personales han sido identificados manualmente con el fin de anonimizarlos.

[0.1] OBJETIVOS

Nos centraremos en explorar los datos que tenemos para establecer vínculos entre el número de ocurrencia de los conceptos de anonimización y la clasificación de un documento en materia de litigio de divorcio.

El reto de nuestro trabajo en el presente proyecto es determinar si, a pesar de estas etiquetas de anonimización un tanto comunes, podríamos hacer que los datos hablaran para ayudar a los profesionales del derecho a clasificar los documentos basándose únicamente en su anonimato.

Se trata, por tanto, de responder a una pregunta experimental y el objetivo del presente trabajo se cumplirá sea cual sean los resultados de desempeño porque, más allá de una lógica de rendimiento, se tratará:

- De sacar las conclusiones necesarias en cuanto a la utilidad de los datos por parte de los bufetes de abogados.
- Se podrá identificar si, para automatizar las tareas de clasificación, los abogados necesitan cambiar su práctica en cuanto a la identificación de conceptos para identificar conceptos adicionales a solo los que permiten anonimizar sus documentos (práctica común), o si la mera anonimización de sus documentos es suficiente para llevar a cabo esta automatización.

[0.2] ANTECEDENTES

Como lo mencionamos anteriormente, contamos con un conjunto de datos que consiste en una lista de documentos de casos de divorcio presentados en primera instancia en la ciudad de México. Estos documentos han sido anotados a mano para identificar conceptos que permiten la anonimización de los mismos. En los documentos se anotaron un total de 70 conceptos de anonimización. Cada documento tiene una categoría general y un tipo, pero nos centraremos únicamente en su tipo (Type).

A partir de este conjunto de datos, generamos para cada documento la frecuencia con la que aparece cada concepto. Debido al gran número de conceptos, sólo nos quedamos con los 10 más frecuentes para limitar la complejidad del análisis estadístico:

Tipo	Descripción
Type	variable categórica

wcount	variable numérica
Apellido_actor	variable numérica
Nombre_actor	variable numérica
Nombre_demandado	variable numérica
Apellido_demandado	variable numérica
Nombre_autorizado	variable numérica
Apellido_autorizado	variable numérica
Num_expediente	variable numérica
Nombre_otro	variable numérica
Apellido_otro	variable numérica
Num_juzgado	variable numérica

De igual forma, hemos reducido el número de categorías de la variable de Type, para quedarnos con 4 categorías que permitirán la clasificación de los documentos según las 10 variables numéricas antes mencionadas.

De los 334 documentos iniciales, esta reducción resulta en 62 ensayos, lo cuál será el tamaño de la muestra que usaremos en este proyecto.

En este primer acercamiento, reducimos el número de variables de forma arbitraría para intentar plantear una primera hipótesis. Sin embargo, es importante señalar que la muestra que hemos elegido es susceptible de evolucionar en función de las necesidades del presente estudio, principalmente en el caso de que las distintas pruebas realizadas revelen la imposibilidad de producir trabajos relevantes para la(s) hipótesis que vamos a probar.

Intentaremos demostrar nuestra hipótesis separando varios pasos que nos permitan avanzar en la demostración :

1. En primer paso, nos centraremos en la elaboración de una **descripción de los datos**, con el objetivo de identificar los análisis que se pueden realizar.
2. En un segundo paso, realizaremos un **análisis de varianza** para comprobar el impacto de las diferentes variables numéricas en la clasificación del tipo de documento
3. En un tercer paso:
    a. Comprobaremos la correlación entre las variables numéricas buscando similitudes en su linealidad para determinar si podemos descartar en los análisis posteriores ciertas variables por su coeficiente de correlación
    b. Veremos si el recuento de estas variables no oculta simplemente una variable correlacionada: el número de palabras en cada uno de los documentos.
4. En un cuarto paso, realizaremos un **análisis de datos categóricos** mediante una prueba ANOVA, para ver si la variable tipo de documento no podría ser clasificada según las variables independientes identificadas como teniendo una distribución normal.
5. En una quinta y última etapa, estableceremos un **análisis multivariante** para determinar un perfil mediante un modelo de regresión lineal generalizado con el fin de estimar la probabilidad de que un conjunto de variables de recuento X conduzca a la clasificación de un documento en un tipo Y.

[1] DESCRIPCIÓN DE DATOS

El conjunto de datos se encuentra en el Anexo 1. Estableceremos en la presente sección las tablas de frecuencia de los conceptos para cada tipo de documento, así como la media, la mediana, la desviación estándar y el coeficiente de asimetría, tanto para cada tipo de documento como para las variables aleatorias de recuento de conceptos en su conjunto.

Asimismo, trataremos de identificar las variables numéricas con una distribución de probabilidad normal, con el fin de identificar los análisis que podremos llevar a cabo.

[1.1] DESCRIPCIÓN DE LAS VARIABLES

[1.1.1] Variable categórica "Type"

Type	Tamaño de la muestra
Contestación de demanda	21
Manifestación de parte	14
Solicitud de copias certificadas	14
Consignación billete de deposito	13
Total de muestras	*62*

[1.1.2] Variables Aleatorias numéricas

Disponemos del recuento total de palabras en cada documento (*wcount*).

Las variables *Apellido_actor, Nombre_actor, Nombre_demandado, Apellido_demandado, Nombre_autorizado, Apellido_autorizado, Num_expediente, Nombre_otro, Apellido_otro, Num_juzgado*, representan el conteo del concepto jurídico al cual se refieren en un documento dado. Por tanto, toman valores enteros. Sus posibles valores constituyen un conjunto infinito numerable, pero que no puede rebasar el número total de palabras de cada documento que constituye el ensayo. Por lo anterior, cumplen con la definición de una variable discreta de recuento al no poder tomar cualquier valor de $\mathbb{R}$, lo cual puede ser problemático para las análisis posteriores que llevaremos a cabo.

Decidimos entonces transformar estos datos para que estas variables sean continúas.

Para esta tarea, establecimos las tablas de frecuencia de estas variables, transformándolas en variables continúas haciendo el cociente entre el conteo de concepto y el número total de palabras en cada documento. Podemos entonces señalar la siguiente descripción de datos:

```
datos.describe()
```

	wcount	B-apellido_actor	B-nombre_actor	B-nombre_demandado	B-apellido_demandado	B-nombre_autorizado	B-apellido_autorizado	B-num_expediente	B-nombre_otro	B-apellido_otro	B-num_juzgado
count	62.000000	62.000000	62.000000	62.000000	62.000000	62.000000	62.000000	62.000000	62.000000	62.000000	62.000000
mean	1056.241935	0.005946	0.005799	0.003618	0.003586	0.006184	0.005898	0.003562	0.002646	0.002652	0.005980
std	1883.739903	0.005932	0.005546	0.003966	0.004176	0.020646	0.020759	0.004577	0.003566	0.003700	0.010027
min	49.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	190.750000	0.001328	0.001652	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	309.500000	0.004397	0.004301	0.002502	0.001887	0.000000	0.000000	0.001417	0.000537	0.000597	0.002355
75%	947.000000	0.009604	0.008886	0.006509	0.006474	0.000386	0.000386	0.005054	0.004760	0.004760	0.005355
max	10538.000000	0.027523	0.027523	0.020202	0.020202	0.146739	0.149457	0.014634	0.010309	0.014085	0.034146

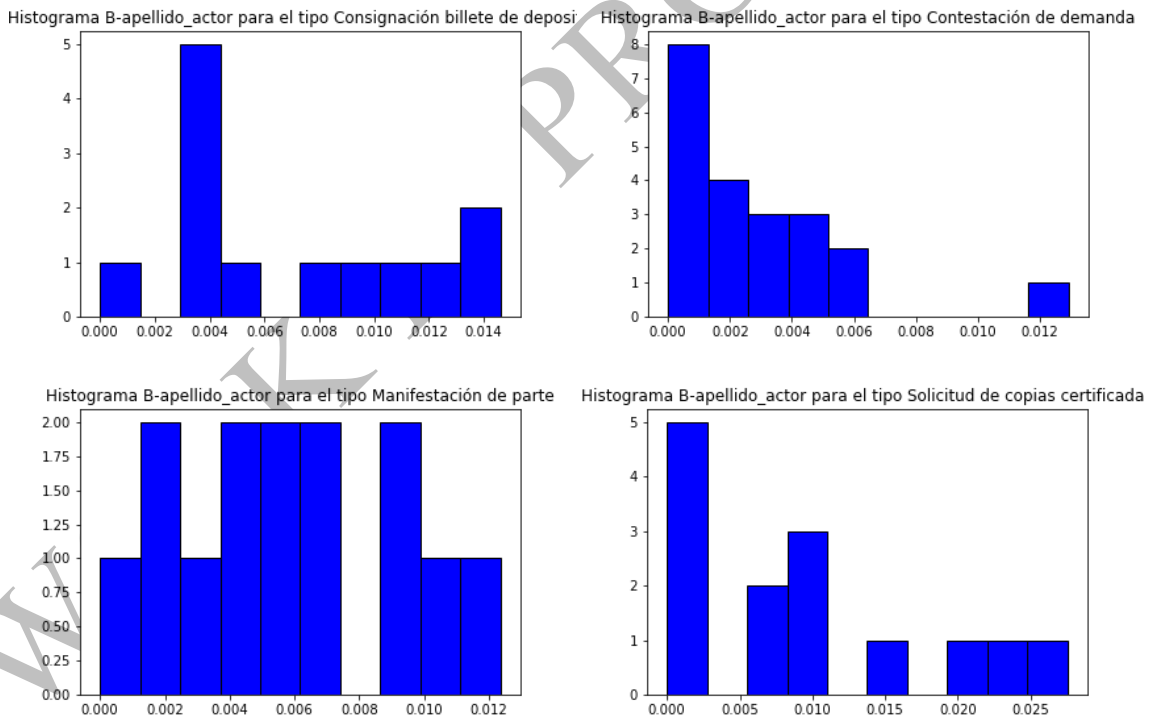
Vemos con la descripción de los datos la media (*mean*) para cada uno de las columnas numéricas, asi como su desviación estándar (*std*), y sus valores minimos (*min*) y máximos (*max*).

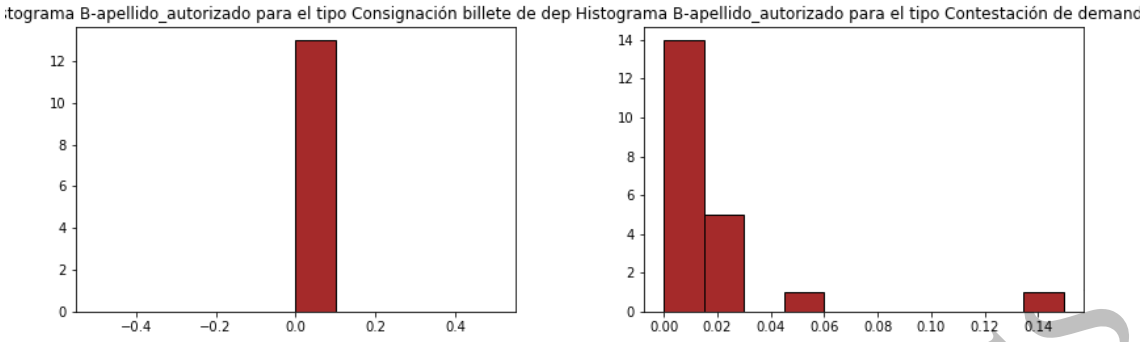
Esta transformación nos permite llevar a cabo las siguientes etapas:

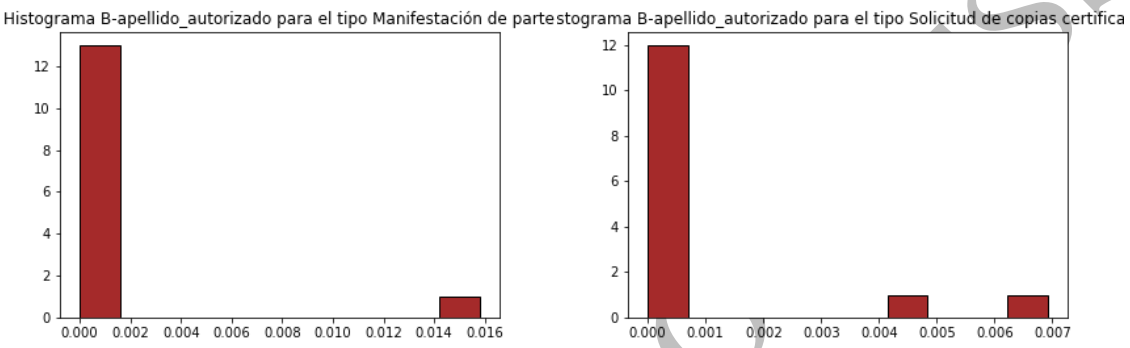
- Para cada tipo de documento buscaremos examinar su histograma para ver si se ajusta a una distribución normal primero haciendo una prueba de Shapiro-Wilks y, si no se ajusta, llevar a cabo una transformación Box-Cox para normalizar la distribución y que se ajuste a una normal. En caso de que los pasos anteriores no permiten ajustar la variable a una distribución normal, podremos en el análisis de varianza elegir examinar con una prueba ANOVA la relación entre la aparición del concepto y el tipo de documento.
- En caso de que no logremos que se ajuste a una distribución normal, tendremos dos opciones para el análisis de varianza:
    - Descartar la variable para el análisis
    - Llevar a cabo una prueba no paramétrica de Kruskal-Wallis para determinar la relación entre la aparición del concepto y el tipo de documento.

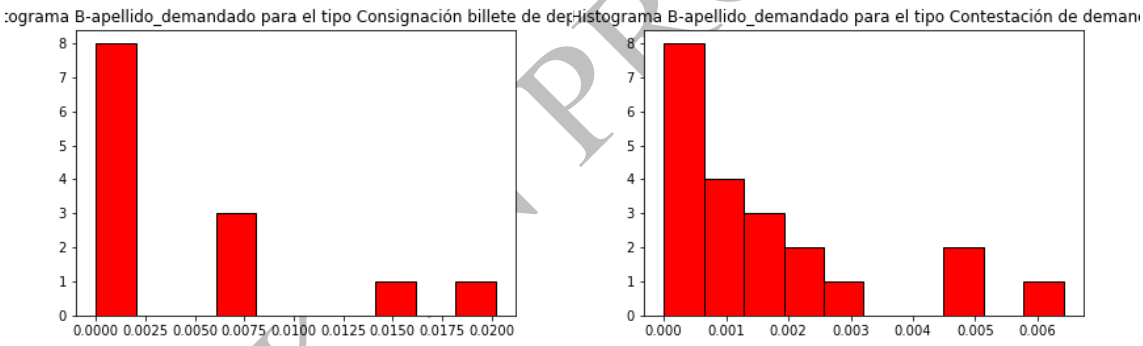
1.1.2.1 Histógramas

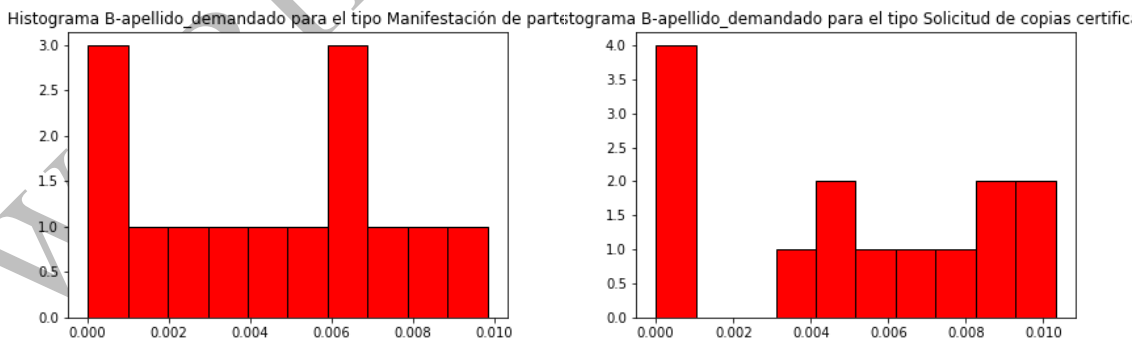
Vamos a graficar, para cada uno de los 4 tipos de documentos, el histógrama de cada una de las variables para ver si alguna tiene una distribución normal. Puesto que son 4 tipos y que las variables numéricas de conceptos de anonimización son 10, graficaremos 40 histógramas.

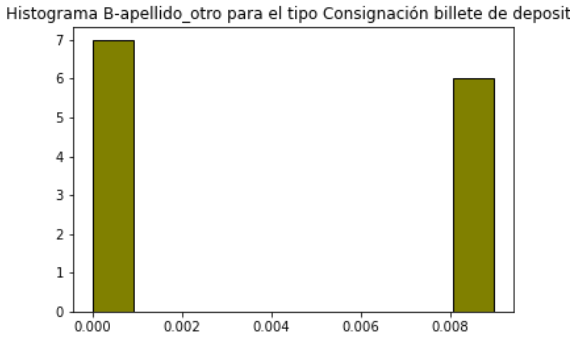


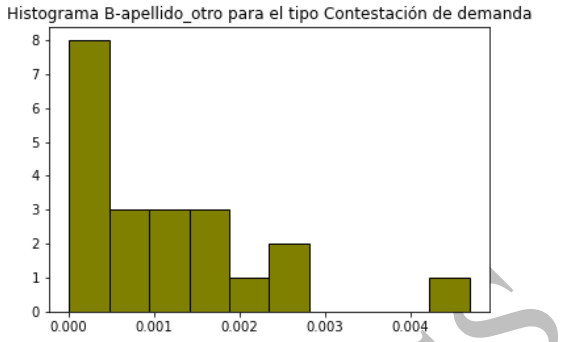


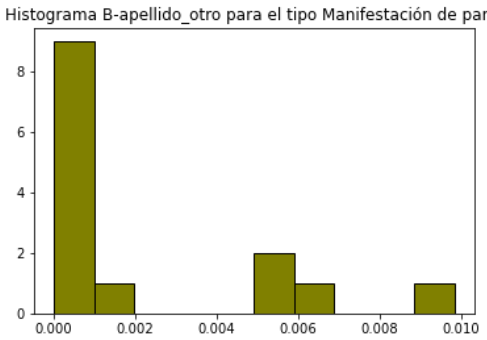


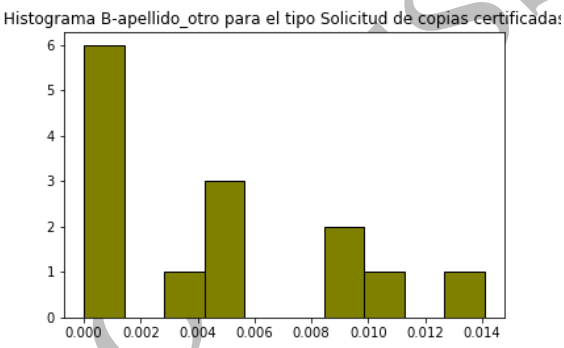


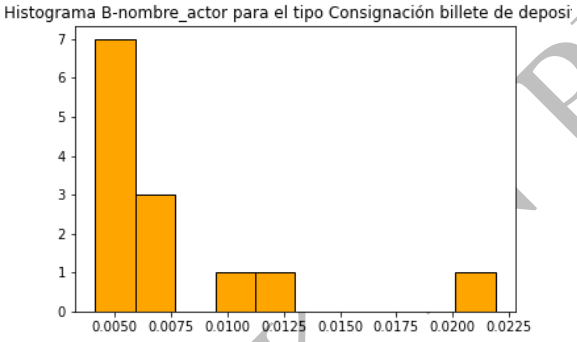


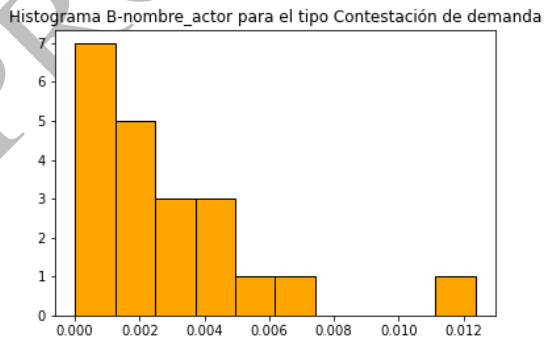


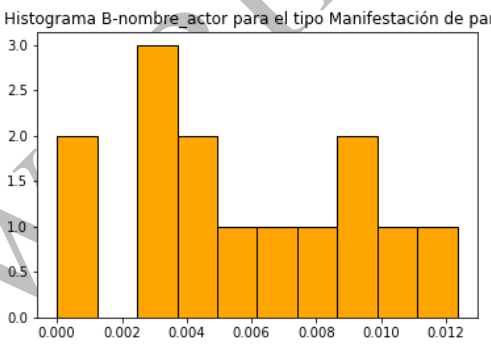


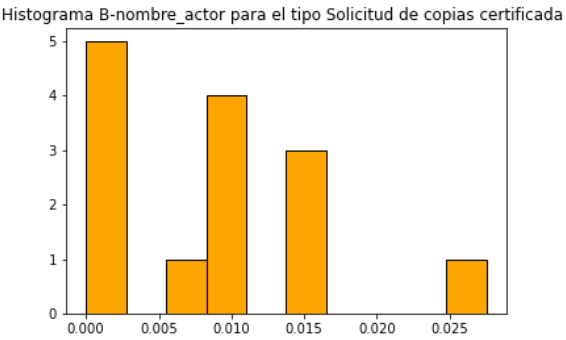


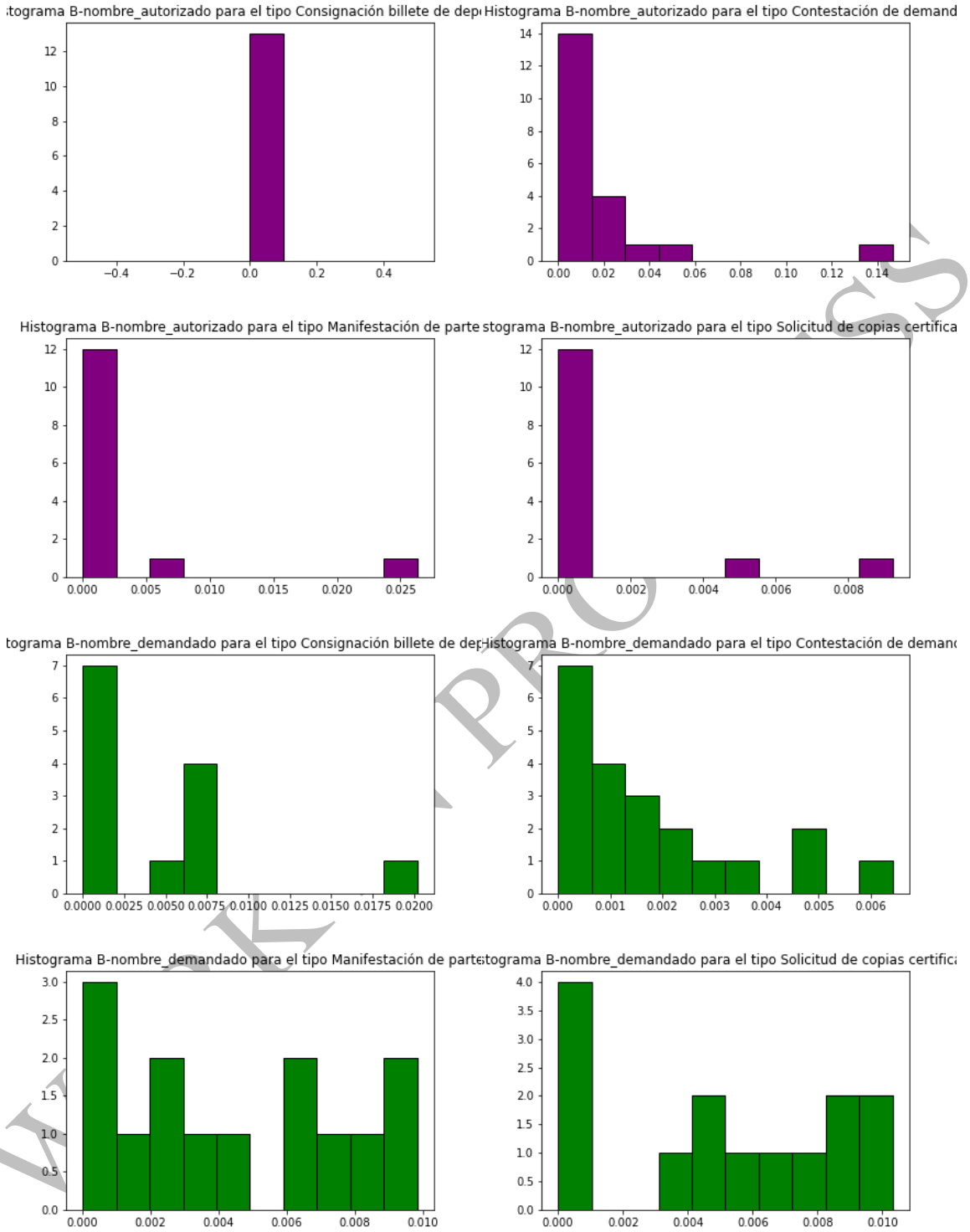


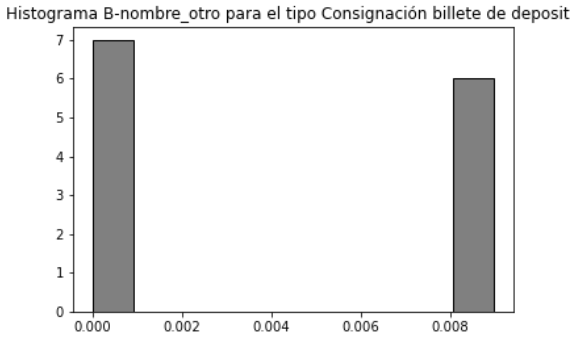


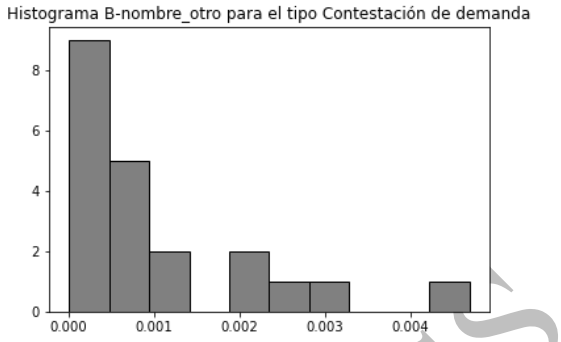


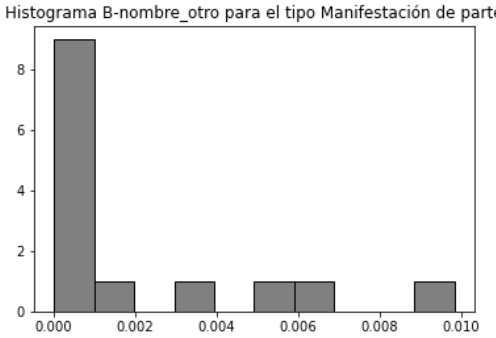


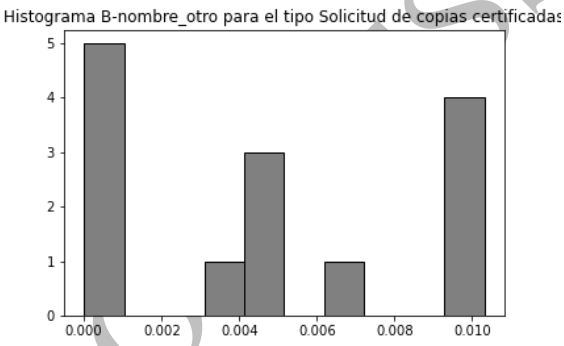


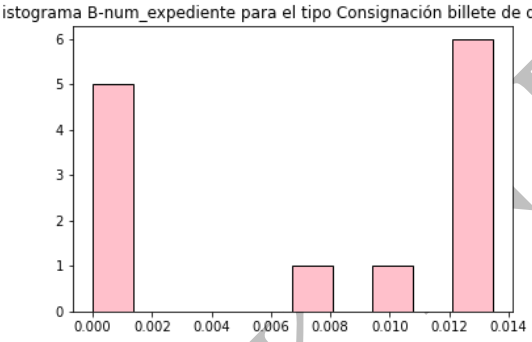

stograma B-nombre_autorizado para el tipo Consignación billete de dep
Histograma B-nombre_autorizado para el tipo Contestación de demand
Histograma B-nombre_autorizado para el tipo Manifestación de parte
stograma B-nombre_autorizado para el tipo Solicitud de copias certifica
stograma B-nombre_demandado para el tipo Consignación billete de dep
Histograma B-nombre_demandado para el tipo Contestación de demand
Histograma B-nombre_demandado para el tipo Manifestación de parte
stograma B-nombre_demandado para el tipo Solicitud de copias certifica

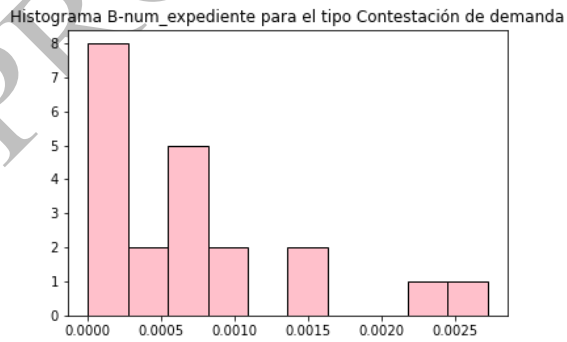


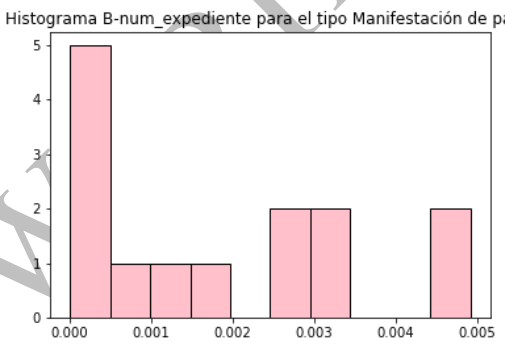


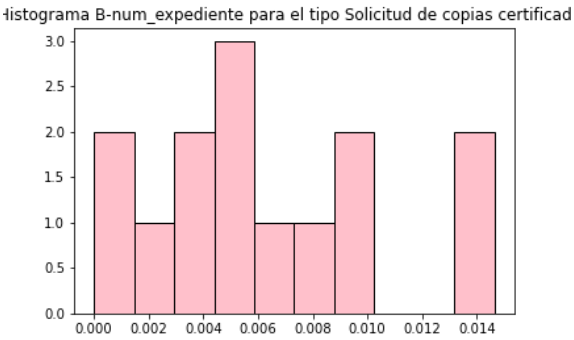


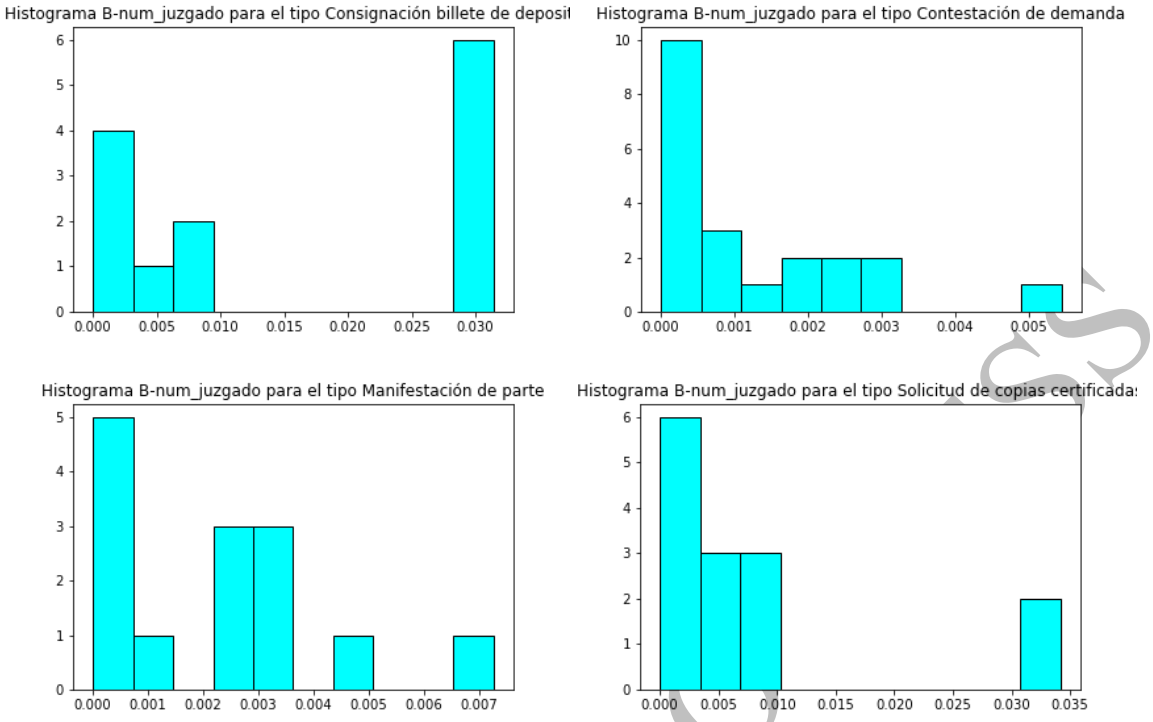








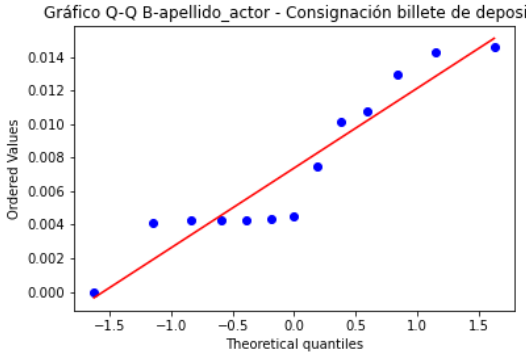


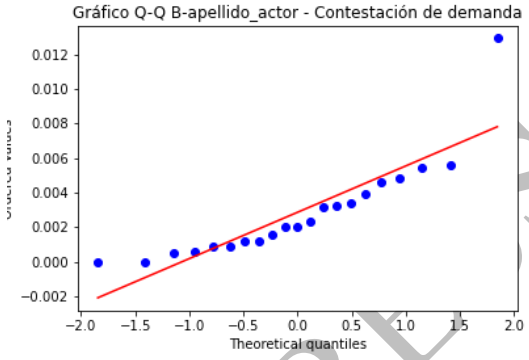


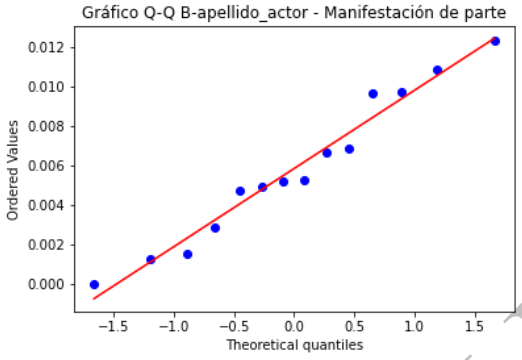
Visualmente, ciertas de las pares tipo-variable tienen una distribución que se parece a la normal. Sin embargo, lo necesitamos verificar de una forma más precisa.

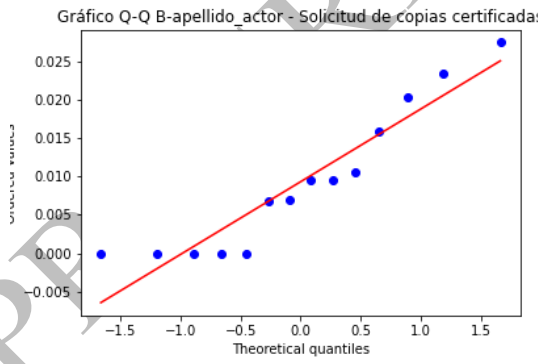
1.1.2.2 Gráfico Quantile-Quantile

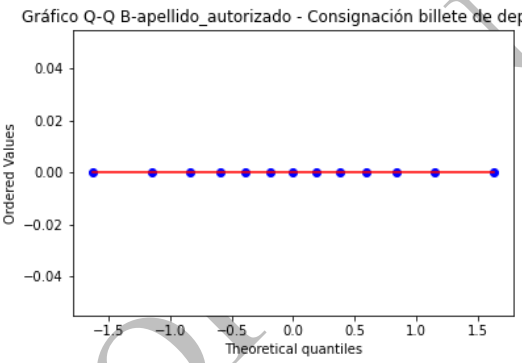
El segundo método que vamos a utilizar para ver si nuestros datos tienen una distribución normal es el gráfico quantile-quantile. Mientras más cercanos estén los puntos a la línea recta, más parecida será su distribución a la normal.

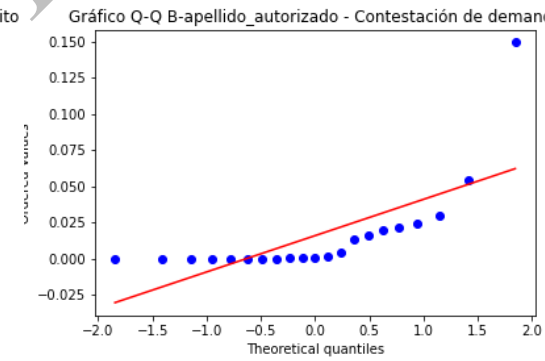


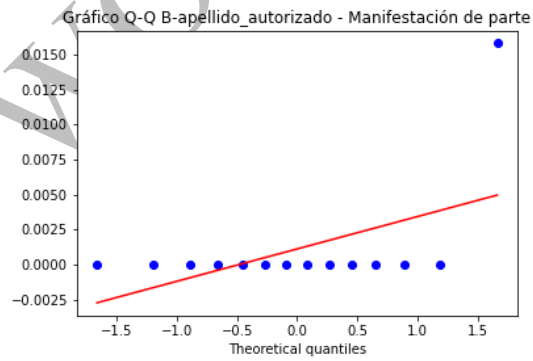


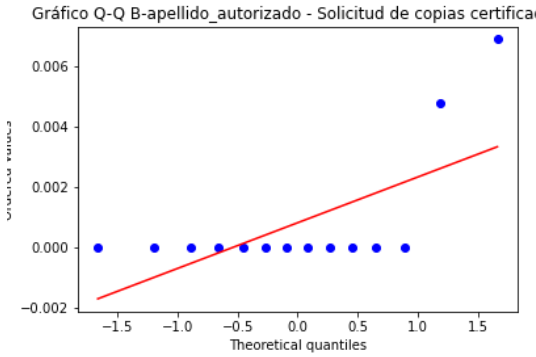


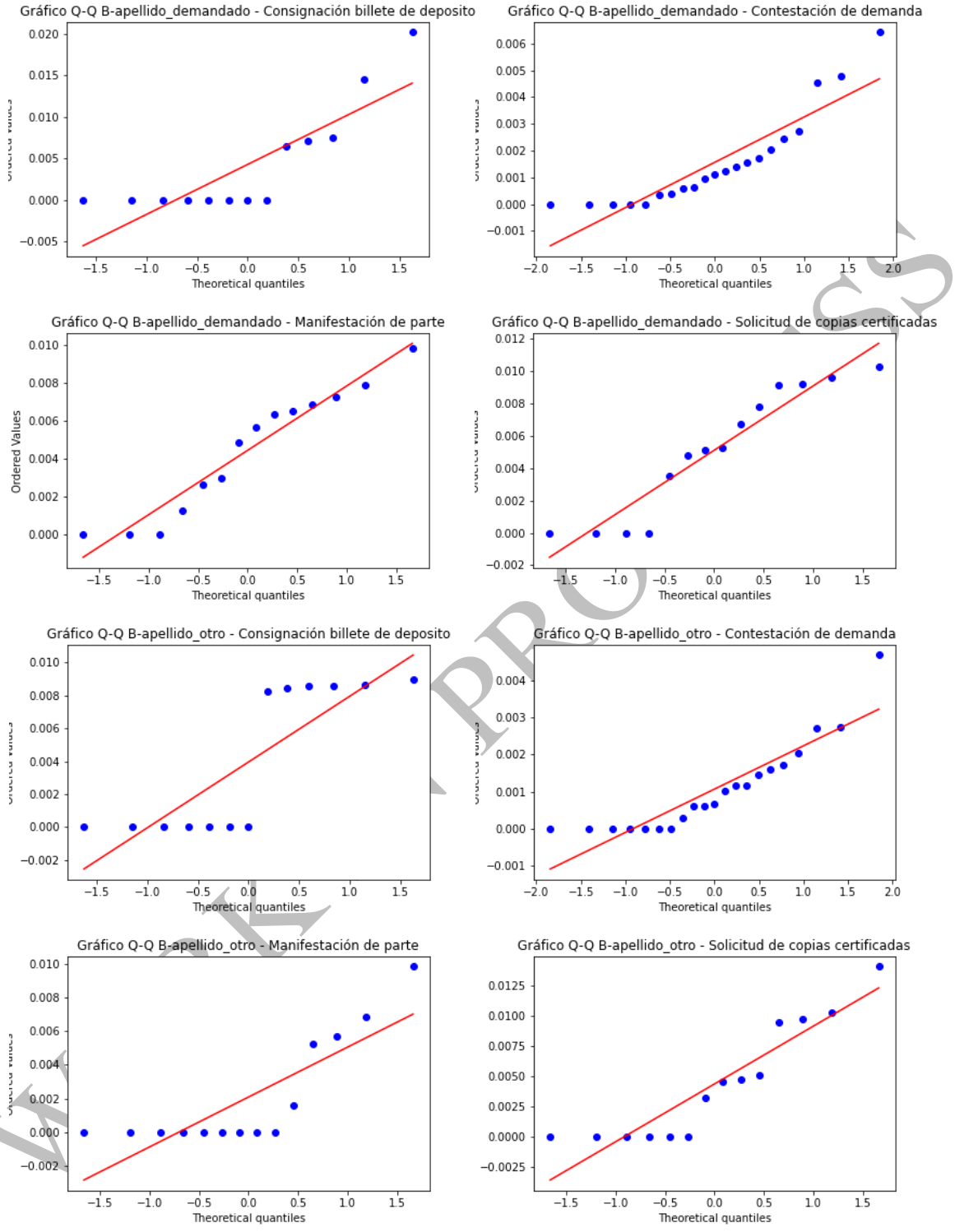


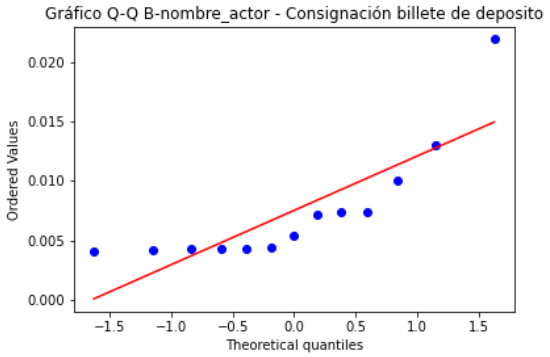


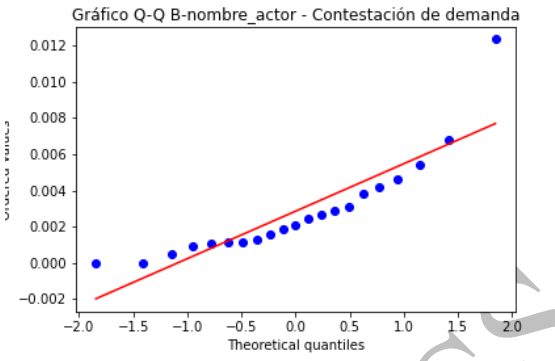


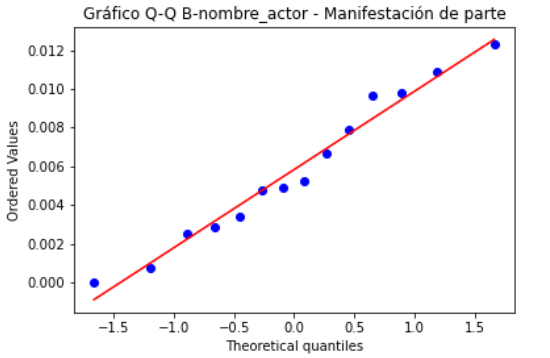


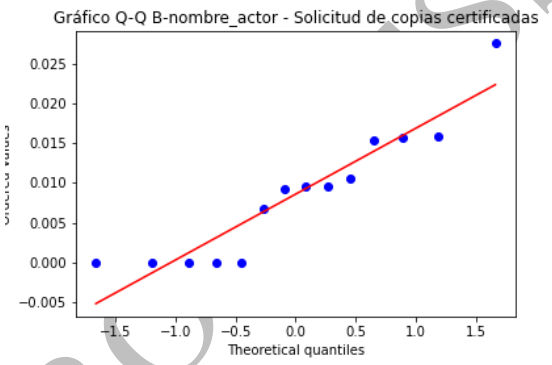


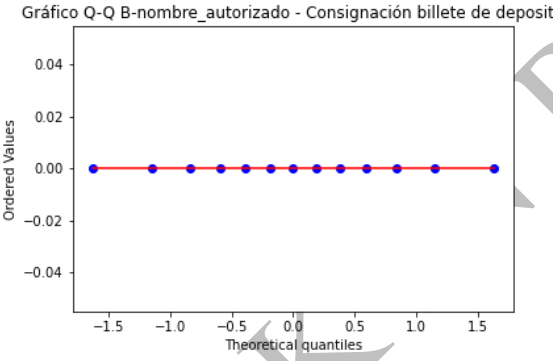


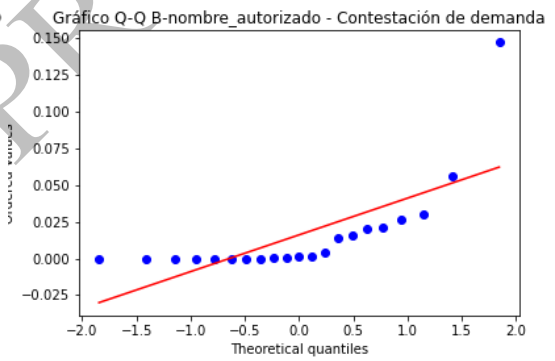


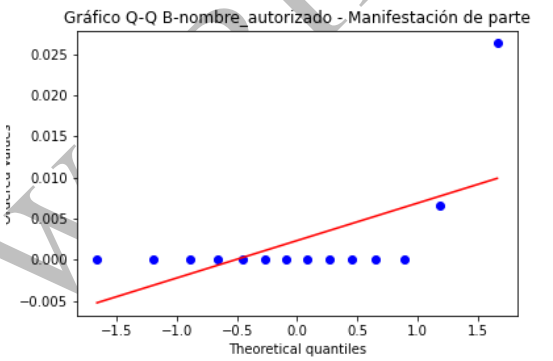


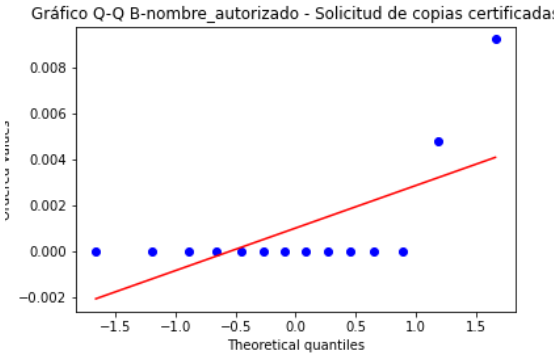












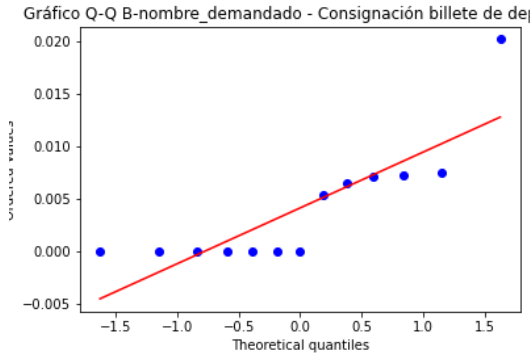


Gráfico Q-Q B-nombre_demandado - Consignación billete de deposito
Theoretical quantiles

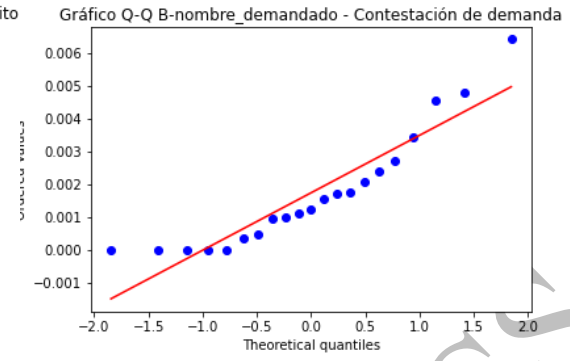


Gráfico Q-Q B-nombre_demandado - Contestación de demanda
Theoretical quantiles

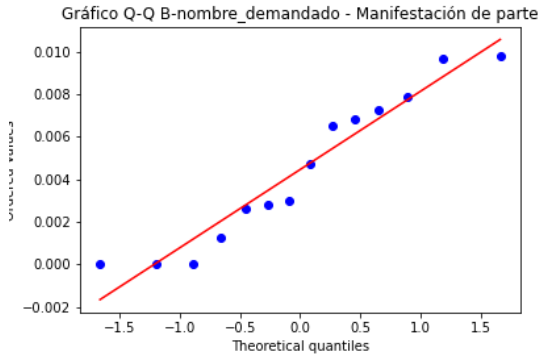


Gráfico Q-Q B-nombre_demandado - Manifestación de parte
Theoretical quantiles

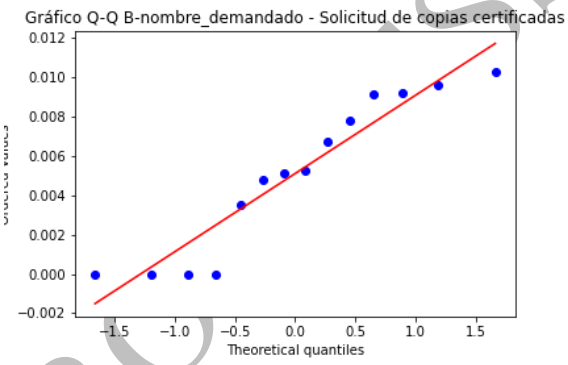


Gráfico Q-Q B-nombre_demandado - Solicitud de copias certificadas
Theoretical quantiles

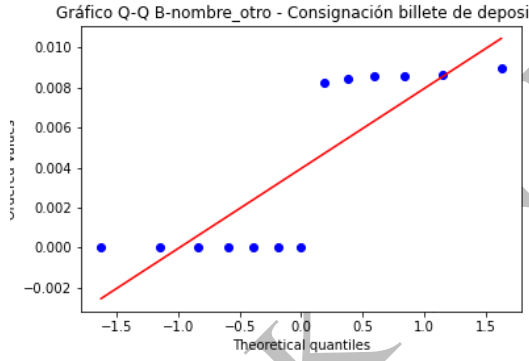


Gráfico Q-Q B-nombre_otro - Consignación billete de deposito
Theoretical quantiles

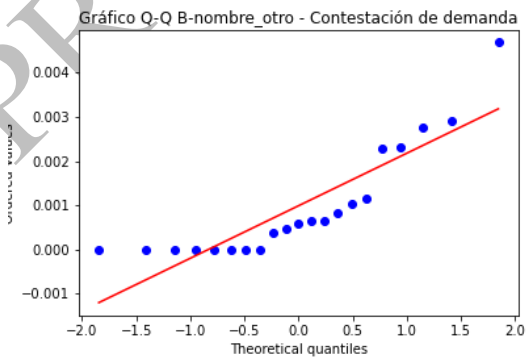


Gráfico Q-Q B-nombre_otro - Contestación de demanda
Theoretical quantiles

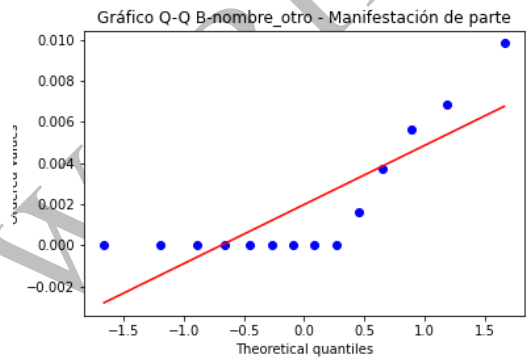


Gráfico Q-Q B-nombre_otro - Manifestación de parte
Theoretical quantiles

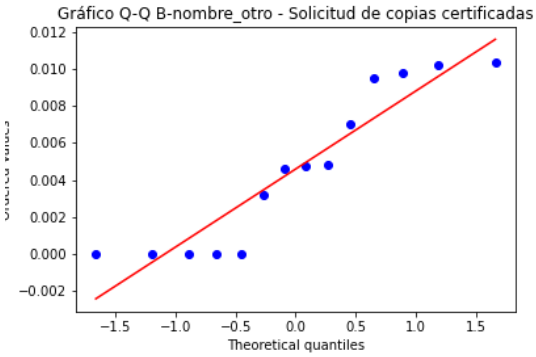
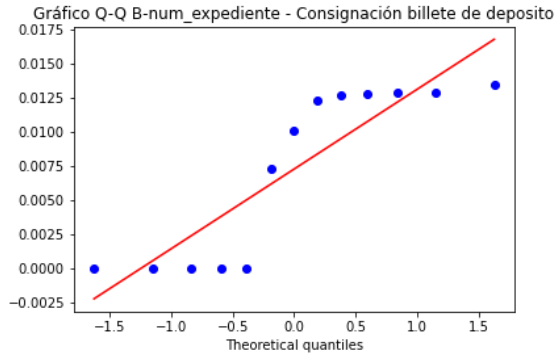
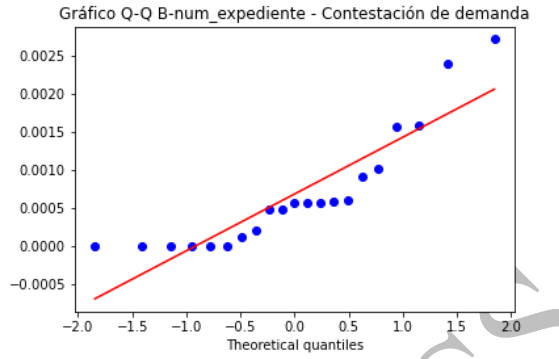
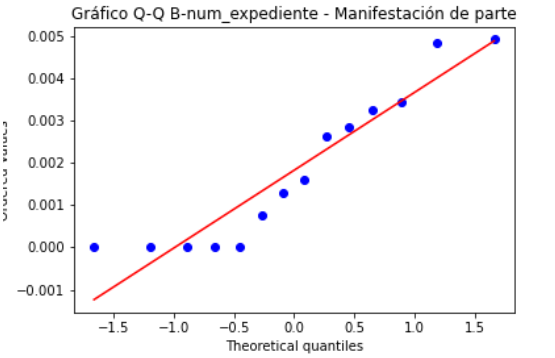
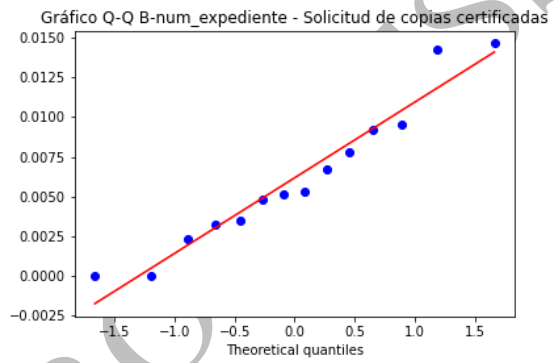


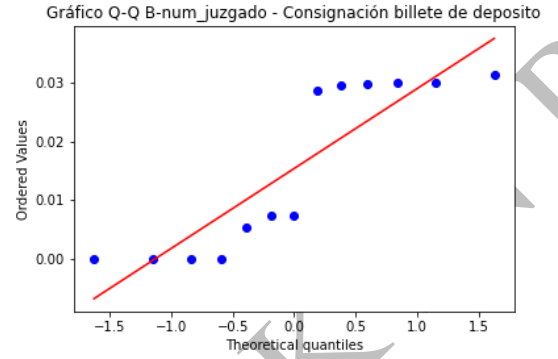
Gráfico Q-Q B-nombre_otro - Solicitud de copias certificadas
Theoretical quantiles

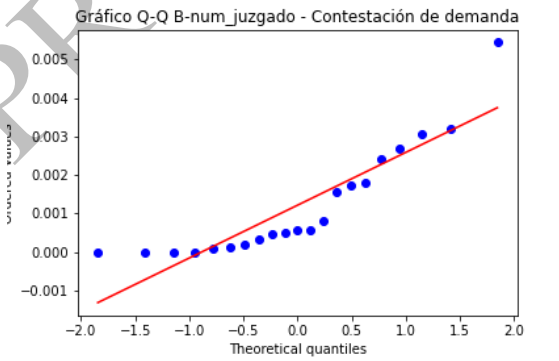


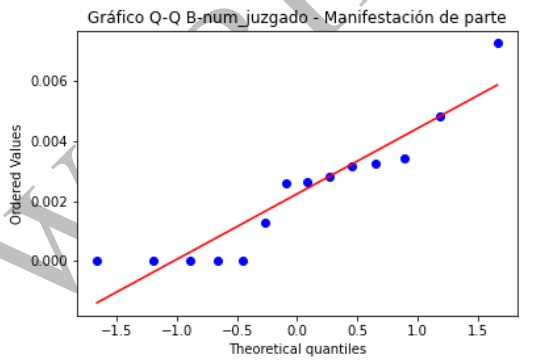


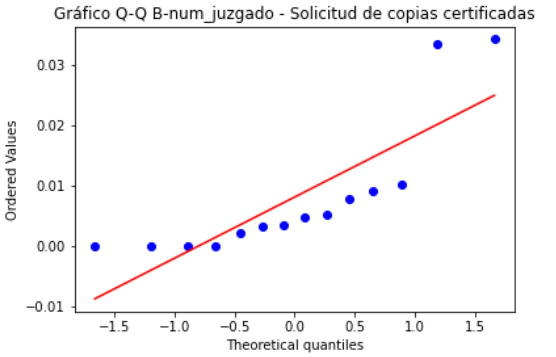












Para tener un resultado matemático más preciso que una evaluación visual, haremos ahora una prueba de Shapiro-Wilks.

1.1.2.3 Prueba Shapiro-Wilks

El estadístico de Shapiro-Wilks es un valor calculado con nuestros datos y el p-value se obtiene utilizando dicho estadístico. Si el p-value es mayor a 0.05, se asume que se trata de una distribución normal.

Lista de variables que se conforman con una distribución normal:

1. Solicitud de copias certificadas para B-apellido_actor que tiene un p-value de 0.053
2. Solicitud de copias certificadas para B-nombre_actor que tiene un p-value de 0.052
3. Solicitud de copias certificadas para B-nombre_demandado que tiene un p-value de 0.061
4. Solicitud de copias certificadas para B-apellido_demandado que tiene un p-value de 0.061
5. Solicitud de copias certificadas para B-num_expediente que tiene un p-value de 0.390
6. Manifestación de parte para B-apellido_actor que tiene un p-value de 0.706
7. Manifestación de parte para B-nombre_actor que tiene un p-value de 0.761
8. Manifestación de parte para B-nombre_demandado que tiene un p-value de 0.175
9. Manifestación de parte para B-apellido_demandado que tiene un p-value de 0.237
10. Consignación billete de depósito para B-apellido_actor que tiene un p-value de 0.084
11. Consignación billete de depósito para B-nombre_autorizado que tiene un p-value de 1.000
12. Consignación billete de depósito para B-apellido_autorizado que tiene un p-value de 1.000

Lista de variables que NO se conforman con una distribución normal:

1. Solicitud de copias certificadas para B-nombre_autorizado que tiene un p-value de 0.000
2. Solicitud de copias certificadas para B-apellido_autorizado que tiene un p-value de 0.000
3. Solicitud de copias certificadas para B-nombre_otro que tiene un p-value de 0.020
4. Solicitud de copias certificadas para B-apellido_otro que tiene un p-value de 0.017
5. Solicitud de copias certificadas para B-num_juzgado que tiene un p-value de 0.000
6. Contestación de demanda para B-apellido_actor que tiene un p-value de 0.000
7. Contestación de demanda para B-nombre_actor que tiene un p-value de 0.001
8. Contestación de demanda para B-nombre_demandado que tiene un p-value de 0.008
9. Contestación de demanda para B-apellido_demandado que tiene un p-value de 0.001
10. Contestación de demanda para B-nombre_autorizado que tiene un p-value de 0.000
11. Contestación de demanda para B-apellido_autorizado que tiene un p-value de 0.000
12. Contestación de demanda para B-num_expediente que tiene un p-value de 0.001
13. Contestación de demanda para B-nombre_otro que tiene un p-value de 0.000
14. Contestación de demanda para B-apellido_otro que tiene un p-value de 0.002
15. Contestación de demanda para B-num_juzgado que tiene un p-value de 0.001
16. Manifestación de parte para B-nombre_autorizado que tiene un p-value de 0.000
17. Manifestación de parte para B-apellido_autorizado que tiene un p-value de 0.000
18. Manifestación de parte para B-num_expediente que tiene un p-value de 0.038
19. Manifestación de parte para B-nombre_otro que tiene un p-value de 0.000
20. Manifestación de parte para B-apellido_otro que tiene un p-value de 0.000
21. Manifestación de parte para B-num_juzgado que tiene un p-value de 0.047
22. Consignación billete de depósito para B-nombre_actor que tiene un p-value de 0.001
23. Consignación billete de depósito para B-nombre_demandado que tiene un p-value de 0.001
24. Consignación billete de depósito para B-apellido_demandado que tiene un p-value de 0.001
25. Consignación billete de depósito para B-num_expediente que tiene un p-value de 0.001

26. Consignación billete de depósito para B-nombre_otro que tiene un p-value de 0.000
27. Consignación billete de depósito para B-apellido_otro que tiene un p-value de 0.000
28. Consignación billete de depósito para B-num_juzgado que tiene un p-value de 0.002

1.1.2.4 Intento de normalización de los datos

Para lograr que nuestros datos tengan una distribución normal, existen varios métodos y usaremos la transformación de Yeo-Johnson, puesto que tenemos ciertos valores a 0 que no son posibles de tratar con una transformación de Box-Cox.

El aplicar el método de Yeo-Johnson no es garantía de que los datos adopten una distribución normal, por lo que es necesario volver a llevar a cabo las pruebas de normalidad para asegurar nos de haberlo logrado.

Esta transformación aplica la función que contiene la variable lamba, la cual tiene un valor que va de -5 a +5. Cuando lambda tine un valor igual a cero, aplica la función logaritmo a nuestros datos. Una vez que se tienen todos los valores nuevos, el método hace pruebas para ver cual valor de lambda que resultó ser el óptimo.

$$\psi(\lambda, y) = \begin{cases} ((y+1)^{\lambda} - 1)/\lambda & \text{if } \lambda \neq 0, y \geq 0 \\ \log(y+1) & \text{if } \lambda = 0, y \geq 0 \\ -[(-y+1)^{2-\lambda} - 1)]/(2-\lambda) & \text{if } \lambda \neq 2, y < 0 \\ -\log(-y+1) & \text{if } \lambda = 2, y < 0 \end{cases}$$

Llevamos este análisis en "1_Descr_normal.ipynb" y volvimos a graficar los histogramas, el gráfico Quantile-Quantile, y a hacer una prueba de Shapiro-Wilks después de la transformación. Con base en esto, pudimos entonces normalizar ciertos datos y establecer inicialmente las siguientes posibilidades en términos de análisis de varianza:

1. ['Solicitud de copias certificadas', 'B-apellido_actor'] se puede comparar con ['Contestación de demanda', 'B-apellido_actor'] mediante ANOVA
2. ['Solicitud de copias certificadas', 'B-nombre_actor'] se puede comparar con ['Contestación de demanda', 'B-nombre_actor'] mediante ANOVA",
3. ['Solicitud de copias certificadas', 'B-nombre_demandado'] se puede comparar con ['Contestación de demanda', 'B-nombre_demandado'] mediante ANOVA",
4. ['Solicitud de copias certificadas', 'B-apellido_demandado'] se puede comparar con ['Contestación de demanda', 'B-apellido_demandado'] mediante ANOVA",
5. ['Manifestación de parte', 'B-apellido_actor'] se puede comparar con ['Contestación de demanda', 'B-apellido_actor'] mediante ANOVA",
6. ['Manifestación de parte', 'B-nombre_actor'] se puede comparar con ['Contestación de demanda', 'B-nombre_actor'] mediante ANOVA",
7. ['Manifestación de parte', 'B-nombre_demandado'] se puede comparar con ['Contestación de demanda', 'B-nombre_demandado'] mediante ANOVA",
8. ['Manifestación de parte', 'B-apellido_demandado'] se puede comparar con ['Contestación de demanda', 'B-apellido_demandado'] mediante ANOVA",
9. ['Consignación billete de deposito', 'B-apellido_actor'] se puede comparar con ['Contestación de demanda', 'B-apellido_actor'] mediante ANOVA",

10. ['Consignación billete de deposito', 'B-nombre_autorizado'] se puede comparar con ['Contestación de demanda', 'B-nombre_autorizado'] mediante ANOVA",
11. ['Solicitud de copias certificadas', 'B-nombre_actor'] se puede comparar con ['Consignación billete de deposito', 'B-nombre_actor'] mediante KRUSKAL-WALLIS",
12. ['Solicitud de copias certificadas', 'B-nombre_demandado'] se puede comparar con ['Consignación billete de deposito', 'B-nombre_demandado'] mediante KRUSKAL-WALLIS",
13. ['Solicitud de copias certificadas', 'B-apellido_demandado'] se puede comparar con ['Consignación billete de deposito', 'B-apellido_demandado'] mediante KRUSKAL-WALLIS",
14. ['Solicitud de copias certificadas', 'B-num_expediente'] se puede comparar con ['Contestación de demanda', 'B-num_expediente'] mediante KRUSKAL-WALLIS",
15. ['Solicitud de copias certificadas', 'B-num_expediente'] se puede comparar con ['Manifestación de parte', 'B-num_expediente'] mediante KRUSKAL-WALLIS",
16. ['Solicitud de copias certificadas', 'B-num_expediente'] se puede comparar con ['Consignación billete de deposito', 'B-num_expediente'] mediante KRUSKAL-WALLIS",
17. ['Manifestación de parte', 'B-nombre_actor'] se puede comparar con ['Consignación billete de deposito', 'B-nombre_actor'] mediante KRUSKAL-WALLIS",
18. ['Manifestación de parte', 'B-nombre_demandado'] se puede comparar con ['Consignación billete de deposito', 'B-nombre_demandado'] mediante KRUSKAL-WALLIS",
19. ['Manifestación de parte', 'B-apellido_demandado'] se puede comparar con ['Consignación billete de deposito', 'B-apellido_demandado'] mediante KRUSKAL-WALLIS",
20. ['Consignación billete de deposito', 'B-nombre_autorizado'] se puede comparar con ['Solicitud de copias certificadas', 'B-nombre_autorizado'] mediante KRUSKAL-WALLIS",
21. ['Consignación billete de deposito', 'B-nombre_autorizado'] se puede comparar con ['Manifestación de parte', 'B-nombre_autorizado'] mediante KRUSKAL-WALLIS",
22. ['Consignación billete de deposito', 'B-apellido_autorizado'] se puede comparar con ['Solicitud de copias certificadas', 'B-apellido_autorizado'] mediante KRUSKAL-WALLIS",
23. ['Consignación billete de deposito', 'B-apellido_autorizado'] se puede comparar con ['Contestación de demanda', 'B-apellido_autorizado'] mediante KRUSKAL-WALLIS",
24. ['Consignación billete de deposito', 'B-apellido_autorizado'] se puede comparar con ['Manifestación de parte', 'B-apellido_autorizado'] mediante KRUSKAL-WALLIS",
25. ['Contestación de demanda', 'B-nombre_actor'] se puede comparar con ['Consignación billete de deposito', 'B-nombre_actor'] mediante KRUSKAL-WALLIS",
26. ['Contestación de demanda', 'B-nombre_demandado'] se puede comparar con ['Consignación billete de deposito', 'B-nombre_demandado'] mediante KRUSKAL-WALLIS",
27. ['Contestación de demanda', 'B-apellido_demandado'] se puede comparar con ['Consignación billete de deposito', 'B-apellido_demandado'] mediante KRUSKAL-WALLIS",
28. ['Contestación de demanda', 'B-nombre_autorizado'] se puede comparar con ['Solicitud de copias certificadas', 'B-nombre_autorizado'] mediante KRUSKAL-WALLIS",
29. ['Contestación de demanda', 'B-nombre_autorizado'] se puede comparar con ['Manifestación de parte', 'B-nombre_autorizado'] mediante KRUSKAL-WALLIS"]

Sin embargo, una verificación manual de los datos reveló que, si bien ciertas distribuciones de las variables independientes aparecían como normal, eso era debido a una modificación demasiada abrupta de los datos que los dejaba en cero, por lo que los volvía imposible explotar con cualquier técnica (ver 'df_yeojohnson.xlsx')

No fue posible tampoco normalizar los datos con una transformación de Box-Cox que necesitaba valores estrictamente positivos, por lo que llegamos a la conclusión que tendremos que analizar la varianza de los datos iniciales ('datos_analisis.xlsx') con técnicas no paramétricas (Kruskal-Wallis), y sólamente llevar a cabo un análisis categórico de los datos con ANOVA para las variables que sí seguían una distribución normal sin transformación:

Manifestación de parte	Consignación billete de deposito	Apellido actor	ANOVA
Solicitud de copias certificadas	Manifestación de parte	Apellido actor	ANOVA
Solicitud de copias certificadas	Consignación billete de deposito	Apellido actor	ANOVA
Solicitud de copias certificadas	Manifestación de parte	Nombre actor	ANOVA
Solicitud de copias certificadas	Manifestación de parte	Nombre demandado	ANOVA

Pudimos también notar que las variables 'nombre_autorizado' y 'apellido_autorizado' están a cero para todas las muestras de Consignación billete de deposito, por lo cual no será relevante llevar a cabo un análisis paramétrico o no paramétrico para compararlas con otros documentos.

[2] Análisis de Varianza

El análisis de varianza es un paso indispensable para determinar la relevancia entre las variables independientes y dependientes. Existen pruebas paramétricas como la ANOVA, pero esta necesita una distribución normal de los datos.

Para los datos que siguen una distribución no paramétrica, llevaremos entonces una prueba de Kruskal-Wallis con Python (GeeksforGeeks, 2022), es una prueba no paramétrica y una alternativa al ANOVA. Por no paramétrica, se entiende que no se supone que los datos procedan de una distribución determinada. El objetivo principal de esta prueba es determinar si existe una diferencia estadística entre las medianas de dos grupos independientes. Sin embargo, empezaremos por una prueba comparando los 4 grupos para cada una de las variables independientes.

Hipótesis:
La prueba de Kruskal-Wallis tiene las hipótesis nula y alternativa que se exponen a continuación:

- La hipótesis nula (H0): La mediana es la misma para todos los grupos de datos.
- La hipótesis alternativa: (H1): La mediana no es igual para todos los grupos de datos.

Consideremos un ejemplo en el que queremos determinar si la frecuencia de apariencia de diferentes conceptos de anonimización provoca una diferencia en la clasificación del tipo de documento en un litigio de divorcio. Con base en la sección anterior, decidimos optar por grupos de 3 o 4 documentos de litigios en materia de divorcio de primera instancia de la Ciudad de México. Ahora, cada grupo se analizará a la luz de un concepto de anonimización. Se han calculado las frecuencia de apariencia de estos conceptos en el paso anterior.

Necesitamos un p-value inferior a 0.05 para rechazar la hipótesis nula.

Prueba de las variables independientes en su conjunto:

Los resultados de las pruebas de Kruskal-Wallis son los siguientes para cada una de las variables independientes en su totalidad (independientemente del tipo de documento):

- Para la variable X_apellido_actor, obtenemos un estadístico de Kruskal-Wallis de 9.008665713144007 con un p-value de 0.029175892749961824 entre los 4 tipos de documentos.
    - Puesto que el p-value < 0.05, podemos rechazar la hipótesis nula para la varibale X_apellido_actor. La variable X_num_juzgado conduce a diferencias estadísticamente significativas entre los 4 tipos de documento de divorcio.
- Para la variable X_apellido_autorizado, obtenemos un estadístico de Kruskal-Wallis de 24.341693988752766 con un p-value de 2.119516782772439e-05 entre los 4 tipos de documentos.
    - Puesto que el p-value < 0.05, podemos rechazar la hipótesis nula para la varibale X_apellido_autorizado. La variable X_num_juzgado conduce a diferencias estadísticamente significativas entre los 4 tipos de documento de divorcio.
- Para la variable X_apellido_demandado, obtenemos un estadístico de Kruskal-Wallis de 6.768958620847607 con un p-value de 0.07963803895839695 entre los 4 tipos de documentos.
    - Puesto que el p-value >= 0.05, NO podemos rechazar la hipótesis nula para la varibale X_apellido_demandado. La variable X_num_juzgado NO conduce a diferencias estadísticamente significativas entre los 4 tipos de documento de divorcio.
- Para la variable X_apellido_otro, obtenemos un estadístico de Kruskal-Wallis de 2.4872763768166557 con un p-value de 0.477594916155089 entre los 4 tipos de documentos.
    - Puesto que el p-value >= 0.05, NO podemos rechazar la hipótesis nula para la varibale X_apellido_otro. La variable X_num_juzgado NO conduce a diferencias estadísticamente significativas entre los 4 tipos de documento de divorcio.
- Para la variable X_nombre_actor, obtenemos un estadístico de Kruskal-Wallis de 10.99373419961653 con un p-value de 0.011759805425559374 entre los 4 tipos de documentos.
    - Puesto que el p-value < 0.05, podemos rechazar la hipótesis nula para la varibale X_nombre_actor. La variable X_num_juzgado conduce a diferencias estadísticamente significativas entre los 4 tipos de documento de divorcio.
- Para la variable X_nombre_autorizado, obtenemos un estadístico de Kruskal-Wallis de 19.217182923065288 con un p-value de 0.0002465353864332242 entre los 4 tipos de documentos.
    - Puesto que el p-value < 0.05, podemos rechazar la hipótesis nula para la varibale X_nombre_autorizado. La variable X_num_juzgado conduce a diferencias estadísticamente significativas entre los 4 tipos de documento de divorcio.
- Para la variable X_nombre_demandado, obtenemos un estadístico de Kruskal-Wallis de 6.532758312036238 con un p-value de 0.08837951109782628 entre los 4 tipos de documentos.
    - Puesto que el p-value >= 0.05, NO podemos rechazar la hipótesis nula para la varibale X_nombre_demandado. La variable X_num_juzgado NO conduce a diferencias estadísticamente significativas entre los 4 tipos de documento de divorcio.
- Para la variable X_nombre_otro, obtenemos un estadístico de Kruskal-Wallis de 4.716180661876431 con un p-value de 0.1937991775124675 entre los 4 tipos de documentos.

  - Puesto que el p-value >= 0.05, NO podemos rechazar la hipótesis nula para la varibale X_nombre_otro. La variable X_num_juzgado NO conduce a diferencias estadísticamente significativas entre los 4 tipos de documento de divorcio.
- Para la variable X_num_expediente, obtenemos un estadístico de Kruskal-Wallis de 14.314769108423286 con un p-value de 0.0025065509474043054 entre los 4 tipos de documentos.
  - Puesto que el p-value < 0.05, podemos rechazar la hipótesis nula para la varibale X_num_expediente. La variable X_num_juzgado conduce a diferencias estadísticamente significativas entre los 4 tipos de documento de divorcio.
- Para la variable X_num_juzgado, obtenemos un estadístico de Kruskal-Wallis de 8.836592083410453 con un p-value de 0.03154425859722744 entre los 4 tipos de documentos.
  - Puesto que el p-value < 0.05, podemos rechazar la hipótesis nula para la varibale X_num_juzgado. La variable X_num_juzgado conduce a diferencias estadísticamente significativas entre los 4 tipos de documento de divorcio.

Prueba de las variables independientes para cada par de documentos:

Realizamos ahora la prueba entre tipos documentos, comparando cada vez un documento con el otro (60 comparaciones en total), para examinar si hay diferencias estadísticamente significativas cuando comparamos sólamente dos documentos entre sí. El objetivo es descartar las variables que no permitiran diferenciar entre los documentos. Los resultados numéricos están disponibles en '2_Analisis_Varianza.ipynb' y '2_varianza.xlsx'. Los resumimos en la siguiente tabla, donde:

- Una celda **verde** significa una diferencia estadísticamente significativa entre los tipos de documentos con base en la variable independiente

- Una celda **roja** significa la ausencia de diferencia estadísticamente significativa entre los tipos de documentos con base en la variable independiente

comparación	X_apellido_actor	X_apellido_autorizado	X_apellido_demandado	X_apellido_otro	X_nombre_actor	X_nombre_autorizado	X_nombre_demandado	X_nombre_otro	X_num_expediente	X_num_juzgado
4 tipos	1	1	0	0	1	1	0	0	1	1
['Solicitud de copias certificadas'] y ['Contestación de demanda']	0	1	1	0	0	1	1	1	1	1
['Solicitud de copias certificadas'] y ['Manifestación de parte']	0	0	0	0	0	0	0	0	1	0
['Solicitud de copias certificadas'] y ['Consignación billete de deposito']	0	0	0	0	0	0	0	0	0	0
['Contestación de demanda'] y ['Manifestación de parte']	1	1	1	0	1	1	1	0	0	0
['Contestación de demanda'] y ['Consignación billete de deposito']	1	1	0	0	1	1	0	0	0	1
['Manifestación de parte'] y ['Consignación billete de deposito']	0	0	0	0	0	0	0	0	0	1

[3] CORRELACIÓN DE DATOS

[3.1] CORRELACIONES ENTRE NOMBRE Y APELLIDOS

En la descripción de datos, hemos podido constatar que ciertas distribuciones eran fuertemente correlacionadas, particularmente los nombre y apellidos de las partes en los documentos. A continuación, trataremos de estimar el grado de correlación entre estas variables para definir si todas están necesarias en nuestro análisis.

Usaremos para esto el coeficiente de correlación de Pearson y haremos las pruebas con Python (Sparrow, 2022).

$$r = \frac{COV(x, y)}{\sigma_x \sigma_y}$$

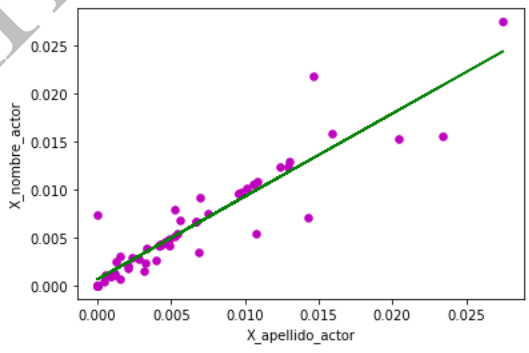
El código esta disponible en '3_Correlacion.ipynb'. Llegamos a los siguientes resultados:

- Coeficiente de correlación de Pearson para X_apellido_actor y X_nombre_actor: 0.925
- Coeficiente de correlación de Pearson para X_apellido_demandado y X_nombre_demandado: 0.942
- Coeficiente de correlación de Pearson para X_apellido_otro y X_nombre_otro: 0.930

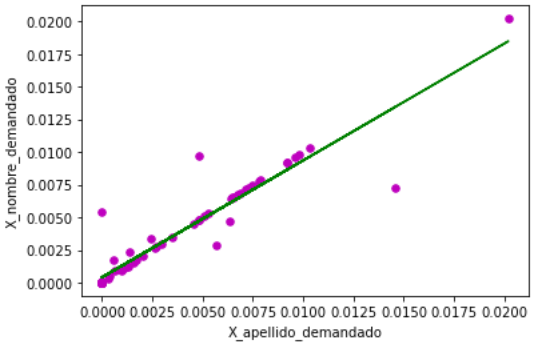
Si aplicamos el coeficiente de correlación de Pearson para cada uno de estos conjuntos de datos, encontramos que es casi idéntico.

Sin embargo, no podemos concluir inmediatamente que si el coeficiente de correlación de Pearson es alto, entonces existe una relación lineal entre ellos. Sólo podremos agrupar las variables si se trata de una relación lineal, que ahora comprobaremos gráficamente (GeeksforGeeks, 2022).

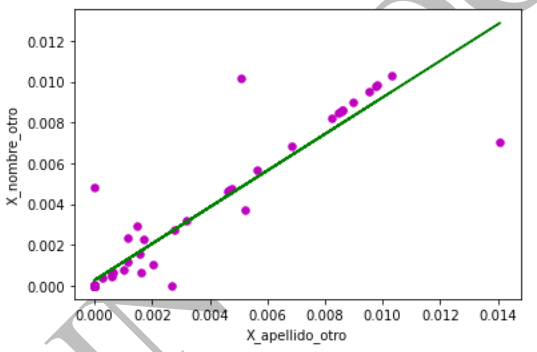
Coeficientes estimados:
X_apellido_actor = 0.0006577754458830321
X_nombre_actor = 0.8646804496547732



Coeficientes estimados:
X_apellido_demandado = 0.00040991511108850066
X_nombre_demandado = 0.8948501260519858



Coeficientes estimados:
X_apellido_otro = 0.0002699248157775765
X_nombre_otro = 0.8960754077807938



Observamos gráficamente que si bien las variables estan fuertemente correlacionadas, es deseable seguir tratandolas independientemente por la dispersión que tienen puesto que en ciertos casos tienen valores diferentes que pueden influir para discriminar entre un tipo de documento u otro.

[3.2] CÁLCULO DE LA SIGNIFICACIÓN ENTRE EL NÚMERO DE LAS PALABRAS Y LAS DIFERENTES VARIABLES INDEPENDIENTES

Necesitamos verificar también que los vínculos entre los datos que estamos intentando probar no resultan de una mera correlación escondida entre el largo del documento (su número de palabras) y la frecuencia de apariencia de los conceptos de anonimización.

Para determinar el valor de p asociado al coeficiente de correlación, se calcula el valor de t mediante la siguiente fórmula.

$$t = \frac{r}{\sqrt{(1 - r^2)/(N - 2)}}$$

A este valor t se le asocia el p-value en función del grado de libertad, al igual que en la prueba t.

Para la correlación, no tenemos que realizar ningún cálculo particular para averiguar el tamaño del efecto. Sólo nos fijamos en el valor del coeficiente y lo interpretamos según las directrices de Cohen (1988):

- Alrededor de r = 0,30 e inferior >> Correlación Baja
- Alrededor de r = 0,50 >> Correlación media
- Alrededor de r = 0,70 y más >> Fuerte correlación

Obtuvimos los siguientes resultados:

1. Coeficiente de correlación de Pearson para X_apellido_actor y wcount: -0.259
2. Coeficiente de correlación de Pearson para X_nombre_actor y wcount: -0.259
3. Coeficiente de correlación de Pearson para X_nombre_demandado y wcount: -0.228
4. Coeficiente de correlación de Pearson para X_apellido_demandado y wcount: -0.231
5. Coeficiente de correlación de Pearson para X_nombre_autorizado y wcount: -0.028
6. Coeficiente de correlación de Pearson para X_apellido_autorizado y wcount: -0.022
7. Coeficiente de correlación de Pearson para X_num_expediente y wcount: -0.312
8. Coeficiente de correlación de Pearson para X_nombre_otro y wcount: -0.223
9. Coeficiente de correlación de Pearson para X_apellido_otro y wcount: -0.208
10. Coeficiente de correlación de Pearson para X_num_juzgado y wcount: -0.242

Observamos que, excepto para el Número de expediente, el Coeficiente de correlación de Pearson indica una correlación negativa baja. Tendremos que interpretar los resultados con cautela puesto que la correlación tampoco es nula entre el número de palabras en un documento y la frecuencia de apariencia de los conceptos de anonimización.

[4] ANÁLISIS DE DATOS CATEGÓRICOS

En está sección, haremos un análisis categórico con pruebas ANOVA para las variables identificadas en la primera sección que tienen una distribución normal. El código es consultable y ejecutable en el archivo '4_Análisis_categóricos.ipynb'.

Los tipos de documentos que se podrán probar con el análisis ANOVA frente a las variables independientes que siguen una distribución normal son los siguientes:

	Tipo 1	Tipo 2	Variable
0	Manifestación de parte	Consignación billete de deposito	X_apellido_actor
1	Solicitud de copias certificadas	Manifestación de parte	X_apellido_actor
2	Solicitud de copias certificadas	Consignación billete de deposito	X_apellido_actor
3	Solicitud de copias certificadas	Manifestación de parte	X_nombre_actor
4	Solicitud de copias certificadas	Manifestación de parte	X_nombre_demandado

Una ANOVA unidireccional (Análisis de varianza (ANOVA) con Python, s. f.) utiliza las siguientes hipótesis nulas y alternativas:

- H0 (hipótesis nula): μ 1 = μ 2 = μ 3 =… = μ k (todas las medias poblacionales son iguales)
- H1 (hipótesis alternativa): al menos una media poblacional es diferente del resto

El ANOVA es bastante robusto a la falta de homodedasticidad si el diseño es equilibrado (mismo número de observaciones por grupo), entonces omitiremos el test de homodedasticidad en nuestro caso.

1. Variable Apellido Actor

Compararemos la varianza de la variable Apellido Actor para la clasificación entre:
- Manifestación de parte y Consignación billete de deposito
- Solicitud de copias certificadas, Manifestación de parte
- Solicitud de copias certificadas y Consignación billete de deposito.

2. Compararemos la varianza de la variable Nombre Actor para la clasificación entre Solicitud de copias certificadas, Manifestación de parte
3. Compararemos la varianza de la variable Nombre Demandado para la clasificación entre Solicitud de copias certificadas, Manifestación de parte

[4.1] Variable Apellido Actor

Obtuvimos los siguientes resultados:

- Manifestación de parte y Consignación billete de deposito:
  - F_onewayResult(statistic=0.8765537417918144, pvalue=0.3581040560313825)
- Solicitud de copias certificadas y Manifestación de parte:
  - F_onewayResult(statistic=1.654712643763896, pvalue=0.20965589620984962)
- Solicitud de copias certificadas y Consignación billete de deposito:
  - F_onewayResult(statistic=0.45659736789649935, pvalue=0.5054215776916644)

En cada uno de los casos, dado que el p-value no es menor que 0.05, no podemos rechazar la hipótesis nula. Esto significa que no tenemos evidencia suficiente para decir que el apellido del actor permite diferenciar entre los documentos. Eso confirma la prueba no paramétrica realizada en la sección 2.

[4.2] Nombre Actor para la clasificación entre Solicitud de copias certificadas y Manifestación de parte

Obtuvimos los siguientes resultados:

- Solicitud de copias certificadas y Manifestación de parte
  - F_onewayResult(statistic=1.2692613937320105, pvalue=0.270196498717093)

Dado que el p-value no es menor que 0.05, no podemos rechazar la hipótesis nula. Esto significa que no tenemos evidencia suficiente para decir que el nombre del actor permite diferenciar entre los documentos Solicitud de copias certificadas y Manifestación de parte. Eso confirma la prueba no paramétrica realizada en la sección 2.

[4.3] NOMBRE DEMANDADO PARA LA CLASIFICACIÓN ENTRE SOLICITUD DE COPIAS CERTIFICADAS Y MANIFESTACIÓN DE PARTE

Obtuvimos los siguientes resultados:

- Solicitud de copias certificadas y Manifestación de parte
    - F_onewayResult(statistic=0.21240803951588183, pvalue=0.6487185571428585)

Dado que el p-value no es menor que 0.05, no podemos rechazar la hipótesis nula. Esto significa que no tenemos evidencia suficiente para decir que el nombre del demandado permite diferenciar entre los documentos Solicitud de copias certificadas y Manifestación de parte. Eso confirma la prueba no paramétrica realizada en la sección 2.

[5] ANÁLISIS MULTIVARIANTE

La presente sección tiene como objetivo llevar a cabo un último análisis antes de brindar una conclusión general a la hipótesis formulada en la conclusión.

En esta fase de nuestra investigación, hemos podido identificar los siguientes puntos:
- La prueba no paramétrica de Kruskal-Wallis nos permitió rechazar las variables apellido_demandado, apellido_otro, nombre_actor, nombre_autorizado, nombre_demandado, nombre_otro para esta etapa del análisis
- Las pruebas no paramétricas han determinado que las variables apellido_actor, apellido_autorizado, nombre_actor, nombre_autorizado, num_expediente y num_juzgado podrían ser significativa para diferenciar entre los 4 tipos de documentos.
- Sin embargo, la comparación de las medias de todos los grupos (cuando era posible usar la prueba ANOVA) o de las medianas de todos los grupos (con las pruebas Kruskal-Wallis) reveló ciertas imposibilidades:
    - La variable apellido_actor sólo permitirá diferenciar entre los tipos:
        - Contestación de demanda y Manifestación de parte
        - Contestación de demanda y Consignación billete de deposito
    - La variable apellido_autorizado sólo permitiran diferenciar entre los tipos:
        - Contestación de demanda y Manifestación de parte
        - Contestación de demanda y Consignación billete de deposito
        - Solicitud de copias certificadas y Contestación de demanda
    - La variable num_expediente sólo permitirá diferenciar entre los tipos:
        - Solicitud de copias certificadas y Manifestación de parte
        - Solicitud de copias certificadas y Consignación billete de deposito
    - La variable num_juzgado sólo permitirá diferenciar entre los tipos:
        - Solicitud de copias certificadas y Contestación de demanda
        - Contestación de demanda y Consignación billete de deposito
        - Manifestación de parte y Consignación billete de deposito

En otras palabras, estas 4 variables independientes pueden servir para una regresión logística multinomial que llevaremos a cabo en Python (Real Python, 2022) con el código disponible en '5_analisis_multivariante.ipynb':

Es importante precisar que estas 4 variables independientes:

- No permiten independientemente diferenciar entre los 4 tipos de documentos
- En su conjunto podrían servir a la diferenciación entre los 4 tipos de documentos, excepto para una diferenciación entre Solicitud de copias certificadas y Consignación billete de deposito

Por esta razón, el modelo de regresión logística que vamos a utilizar sólamente servirá para diferenciar entre los tipos de documentos, pero en caso de que los clasifique como Solicitud de copias certificadas o Consignación billete de deposito, entonces no podremos establecer que el documento pertenezca a una u otra categoría porque ninguna variable independiente nos permite hacer esta diferencia. Por tanto, tenemos que restringir el alcance del presente estudio y configurar el modelo de regresión multinomial para que pueda entregar un resultado de tal forma que clasifique entre:

- Contestación de demanda
- Manifestación de parte

El problema de clasificación, por tanto, es binario.

La clasificación binaria es un algoritmo de aprendizaje supervisado que clasifica las nuevas observaciones en una de dos clases que representaremos con 1 y 0:

- Contestación de demanda = 1
- Manifestación de parte = 0

Evaluación del clasificador binario

- Si el modelo predice con éxito que el documento es Contestación de demanda, este caso se denominará True Positive (TP).
- Si el modelo predice con éxito que el documento es Manifestación de parte, este caso se denominará True Negative (TN).

El clasificador binario también puede identificar erróneamente los documentos:

- Si un documento es clasificado como Manifestación de parte erroneamente, este error se denomina False Negative (FN).
- Del mismo modo, si un documento es clasificado como Contestación de demanda cuando es en realidad una Manifestación de parte, este error se denomina False Positive (FP)

Podemos evaluar entonces el clasificador binario basándonos en los siguientes parámetros:

- True Positive (TP): El documento es Contestación de demanda y el modelo predice 1.
- False Positive (FP): El documento es Manifestación de parte y el modelo predice 1.
- True Negative (TN): El documento es Manifestación de parte y el modelo predice 0.
- False Negative (FN): El documento es Contestación de demanda y el modelo predice 0.

Una vez obtenidos estos valores, podemos calcular la puntuación de precisión del clasificador binario de la siguiente manera:

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

A continuación se presenta una matriz de confusión que representa los parámetros anteriores:

		PREDICTED	
		Positive	Negative
ACTUAL	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Muchos métodos utilizan la clasificación binaria. Usaremos la regresión logística multivariante para el presente estudio.

1. Datos: Tenemos 21 documentos Contestación de demanda, denotados como 1, y 14 documentos Manifestación de parte, denotados como 0.
2. Definir las variables independientes y la variable dependiente: Almacenaremos las filas de los conceptos de anonimización en una variable X y la clase correspondiente de esas observaciones (0 o 1) en una variable Y.
3. Dividir el conjunto de datos en conjuntos de entrenamiento y de prueba: Utilizamos el 75% de los datos para el entrenamiento y el 25% para las pruebas.
4. Normalizamos después de dividir los datos.
5. Ajustamos un modelo de regresión logística a los datos de entrenamiento
6. Hicimos predicciones con el conjunto de prueba
7. Calculamos la precisión comparando los valores reales y los valores previstos

Pudimos entonces calcular el rendimiento del modelo comparando las predicciones del modelo con los verdaderos valores objetivo, que reservamos en la variable y_test.

Entonces calculamos la matriz de confusión para obtener los parámetros necesarios:

- True Positive (TP) = 5
- False Positive (FP) = 1
- True Negative (TN) = 3
- False Negative (FN) = 0

Con estos valores, pudimos calcular una precisión de 88.9% para el clasificador binario con este modelo de regresión logística multivariante.

[6] CONCLUSIONES

Todo este análisis nos ha demostrado que el simple uso de conceptos de anonimización no es una garantía para la clasificación de los documentos de litigios de divorcio. En efecto, las pruebas mostraron que los conceptos podían correlacionarse con la longitud del documento, y que no todos los conceptos permitían, en este contexto, inferir un vínculo estadísticamente significativo con el tipo de documento.

Sin embargo, tras discriminar las variables superfluas, pudimos obtener resultados alentadores en cuanto a la identificación de los documentos Contestación de demanda y Manifestación de parte sobre la base de sólo 4 variables independientes que, a priori, no tienen nada que ver con los conceptos habitualmente utilizados por los abogados para clasificar sus documentos.

Se trata de una serendipia significativa, sobre todo porque el tamaño de la muestra era relativamente pequeño.

Por lo tanto, el presente estudio invita a realizar más análisis, no sólo con los 70 conceptos de la base de datos completa, sino también con toda la gama de tipos de documentos. Además, es probable que el aumento del número de observaciones sea un requisito indispensable. De hecho, aunque el presente trabajo muestra pistas prometedoras para la clasificación de documentos jurídicos de divorcio mediante el trabajo habitual de anonimización de documentos, los resultados sólo pueden ser concluyentes con datos más amplios.

Anexo 1

#	Type	wcount	Apellido actor	Nombre actor	Nombre demandado	Apellido demandado	Nombre autorizado	Apellido autorizado	Num expediente	Nombre otro	Apellido otro	Num juzgado
1	Solicitud de copias certificadas	97	0	0	0	0	0	0	0	0	0	0
2	Solicitud de copias certificadas	97	0	0	1	1	0	0	0	1	1	0
3	Contestación de demanda	363	0	0	0	0	0	0	0	1	1	0
4	Contestación de demanda	860	1	1	0	0	0	0	0	2	1	0
5	Contestación de demanda	370	0	0	0	0	0	0	0	0	1	1
6	Contestación de demanda	1247	4	3	8	8	37	37	3	0	0	3
7	Contestación de demanda	10538	9	11	5	4	5	5	0	4	3	1
8	Contestación de demanda	3053	12	8	1	1	0	0	0	2	2	0
9	Contestación de demanda	976	2	2	2	2	0	1	1	1	2	3
10	Contestación de demanda	368	2	2	1	1	54	55	1	0	0	2
11	Contestación de demanda	1728	1	2	0	0	36	38	1	0	0	3
12	Contestación de demanda	640	1	2	0	0	1	1	1	3	3	1
13	Contestación de demanda	8394	10	11	13	13	1	1	1	4	5	1
14	Contestación de demanda	2053	1	1	7	5	8	9	1	6	3	1
15	Manifestación de parte	611	3	3	6	6	0	0	3	6	6	0
16	Manifestación de parte	49	0	0	0	0	0	0	0	0	0	0
17	Manifestación de parte	551	6	6	4	4	0	0	0	0	0	4
18	Manifestación de parte	707	2	2	2	4	0	0	2	4	4	2
19	Manifestación de parte	1338	2	1	4	4	0	0	1	5	7	0

20	Manifestación de parte	786	1	2	1	1	0	0	1	0	0	1
21	Manifestación de parte	292	2	1	2	2	0	0	1	2	2	1
22	Consignación billete de deposito	234	1	1	0	0	0	0	3	2	2	7
23	Consignación billete de deposito	223	1	1	0	0	0	0	3	2	2	7
24	Consignación billete de deposito	233	1	1	0	0	0	0	3	2	2	7
25	Consignación billete de deposito	232	1	1	0	0	0	0	3	2	2	7
26	Consignación billete de deposito	243	1	1	0	0	0	0	3	2	2	7
27	Consignación billete de deposito	236	1	1	0	0	0	0	3	2	2	7
28	Contestación de demanda	1740	8	8	3	3	35	35	1	4	3	1
29	Solicitud de copias certificadas	205	0	0	0	0	0	0	3	2	2	7
30	Solicitud de copias certificadas	210	0	0	0	0	0	0	3	2	2	7
31	Contestación de demanda	627	2	1	3	3	35	34	1	0	0	2
32	Manifestación de parte	81	1	1	0	0	0	0	0	0	0	0
33	Consignación billete de deposito	140	2	1	1	1	0	0	0	0	0	0
34	Consignación billete de deposito	137	2	3	1	2	0	0	1	0	0	1
35	Manifestación de parte	380	2	3	3	3	10	6	1	0	0	1
36	Consignación billete de deposito	186	2	1	1	0	0	0	0	0	0	1
37	Consignación billete de deposito	154	2	2	1	1	0	0	0	0	0	0
38	Consignación billete de deposito	135	0	1	0	0	0	0	0	0	0	0
39	Consignación billete de deposito	134	1	1	1	1	0	0	0	0	0	1
40	Consignación billete de deposito	99	1	1	2	2	0	0	1	0	0	0
41	Contestación de demanda	1733	4	5	3	1	24	23	1	2	2	0
42	Solicitud de copias certificadas	284	3	3	1	1	0	0	1	2	4	1
43	Contestación de demanda	2076	7	8	2	2	0	0	1	0	0	1
44	Contestación de demanda	4931	10	9	6	6	5	4	3	4	5	4
45	Contestación de demanda	1780	23	22	2	2	47	44	1	0	0	1
46	Contestación de demanda	2205	2	2	10	10	35	36	2	0	0	4
47	Contestación de demanda	3104	15	13	3	2	0	0	0	2	5	1
48	Solicitud de copias certificadas	312	3	3	3	3	0	0	1	1	1	1
49	Solicitud de copias certificadas	209	2	2	1	1	0	0	2	1	1	1
50	Solicitud de copias certificadas	196	4	3	1	1	0	0	1	2	1	2
51	Solicitud de copias certificadas	189	3	3	1	1	0	0	1	0	0	1
52	Solicitud de copias certificadas	128	3	2	1	1	0	0	1	0	0	1
53	Manifestación de parte	307	3	3	2	2	0	0	1	0	0	1
54	Solicitud de copias certificadas	434	3	4	4	4	4	3	1	2	2	1
55	Manifestación de parte	207	2	2	2	1	0	0	1	0	0	1
56	Solicitud de copias certificadas	149	1	1	1	1	0	0	1	0	0	0
57	Solicitud de copias certificadas	109	3	3	1	1	0	0	1	0	0	1
58	Manifestación de parte	150	1	1	0	0	1	0	0	0	0	0
59	Solicitud de copias certificadas	208	0	0	0	0	1	1	1	1	0	0
60	Contestación de demanda	5014	28	34	12	7	0	0	1	3	3	1

61	Manifestación de parte	383	2	2	1	1	0	0	0	0	0	1
62	Manifestación de parte	632	3	3	3	4	0	0	1	1	1	2

Los archivos completos del proyecto están comprimidos. Se puede verificar la integridad del archivo comparando el MD5 que debe de revelar el siguiente valor:

e650101144e314d03cc3cd835f77344a

BIBLIOGRAFÍA

GeeksforGeeks. (2022, 28 marzo). *How to Perform a Kruskal-Wallis Test in Python*. Recuperado 29 de septiembre de 2022, de https://www.geeksforgeeks.org/how-to-perform-a-kruskal-wallis-test-in-python/

Sparrow, J. (2022, 6 julio). *Python – Test de corrélation de Pearson entre deux variables – Acervo Lima*. Recuperado 29 de septiembre de 2022, de https://fr.acervolima.com/test-de-correlation-python-pearson-entre-deux-variables/

GeeksforGeeks. (2022b, agosto 22). *Linear Regression (Python Implementation)*. Recuperado 29 de septiembre de 2022, de https://www.geeksforgeeks.org/linear-regression-python-implementation/

Análisis de varianza (ANOVA) con Python. (s. f.). Recuperado 29 de septiembre de 2022, de https://www.cienciadedatos.net/documentos/pystats09-analisis-de-varianza-anova-python.html