Maestría en Inteligencia Artificial

Aprendizaje Automático Avanzado
Transformación de variables numéricas

Dr. Gaddiel Desirena López

Transformación: Es una función que mapea los elementos del conjunto X al conjunto Y.

- En términos estadísticos, se usan para estabilizar la varianza.
- Para variables correlacionadas, se observa dispersión variable en función de la magnitud de la observación.
- Una vez trabajados los datos (usando una regresión, red neuronal, etc.), se realiza la transformación inversa para poder interpretar correctamente los datos.

1. Transformación logaritmo y recíproco

Función logaritmo: Una forma de visualizar la función logaritmo es como el inverso de la exponencial, es decir

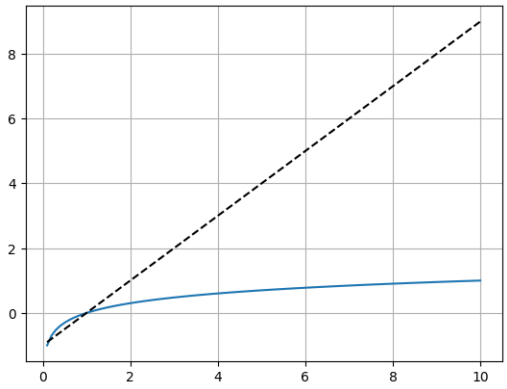
$$a^{x'} = x \quad \leftrightarrow \quad \log_a(x) = x'.$$

- Comprime el rango numérico alto y expande el rango bajo.
- Los valores pequeños, entre (0,1) los mapea al intervalo $(-\infty,0)$.
- Específicamente, la función $\log_{10}(x)$, mapea los intervalos mostrados en la Tabla 1 y el gráfico se muestra en la Figura 1.

Tabla 1: Mapeos de la función $\log_{10}$.

x	$\log_{10}(x)$
$[1, 10]$	$[0, 1]$
$[10, 100]$	$[1, 2]$
$\ldots$	

Figura 1: Gráfica de la función $\log_{10}(x)$.

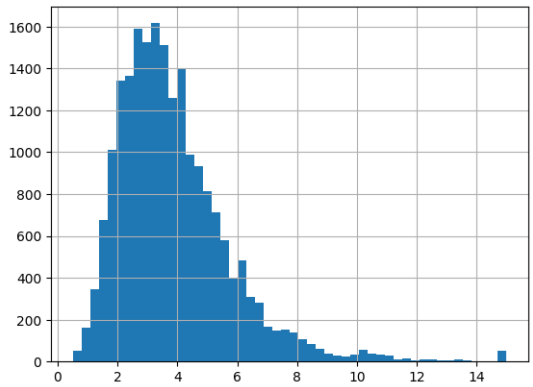


- Para que el mapeo de los valores más pequeños corresponda a cero, una buena práctica es desplazar las observaciones de tal manera que $x \in [1, \infty)$.
- Es útil para lidiar con números positivos con una distribución de cola pesada. La Tabla 2 muestra la sección inicial y final de datos numéricos ordenados con sesgo positivo (sesgo=1.646657), mientras que la Figura 2 muestra el histograma de estos datos.

Tabla 2: Fragmento de datos con cola pesada.

0	8.3252
1	8.3014
2	7.2574
3	5.6431
4	3.8462
20635	1.5603
20636	2.5568
20637	1.7000
20638	1.8672
20639	2.3886

Figura 2: Histograma de los datos originales.

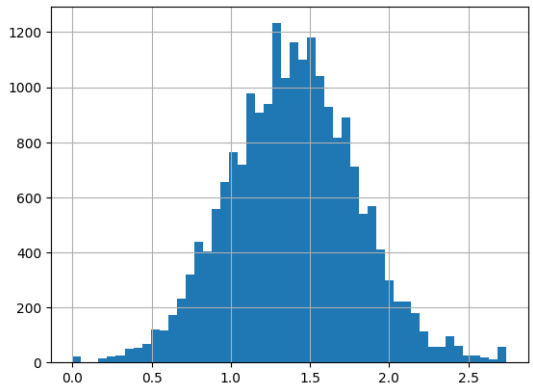


- Los datos originales se desplazan a 1 y se hacen pasar por la función logaritmo, resultando los mostrados en la Tabla 3, con un sesgo de 0.078282, y su histograma en la Figura 3.

Tabla 3: Fragmento de los datos transformados.

0	2.177623
1	2.174922
2	2.048660
3	1.815346
4	1.469325
20635	0.722900
20636	1.117401
20637	0.788503
20638	0.861750
20639	1.060807

Figura 3: Histograma de los datos transformados.

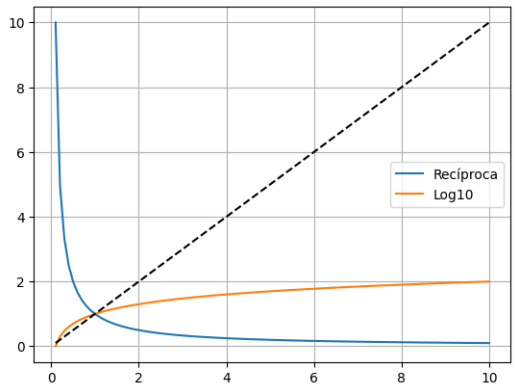


Recíproco: Otra transformación útil para variables con datos con un sesgo muy grande es la función recíproca

$$x' = 1/x$$

- Es más potente que la función logarítmica. Los valores grandes se atenúan más y los valores menores que 1 crecen más rápido.
- A modo de comparación. Con valores menores que uno, por ejemplo $x = 0.1$, la función logaritmo regresa $\log(0.1) = -1$, que en magnitud es más pequeño que el obtenido con la función recíproco $10 = 1/0.1$. Además, al evaluar ambas funciones con $x = 10$, se obtiene $1 = \log(10)$ y $0.1 = 1/10$, donde se ve que los valores grandes son más atenuados con la función recíproco. La Figura 4 muestra los detalles entre el intervalo $[0.1, 10]$.

Figura 4: Comparación de la función recíproca con la logarítmica.



- Para una variable con sesgo positivo muy pronunciado, de 97.639561, aplicando la transformación logarítmica se compensa el sesgo a 2.662098, mientras que con la transformación reciproca, el sesgo resultante es de 0.531376. Se muestran los histogramas de la transformación logarítmica y recíproca respectivamente en la Figura 5 y en la Figura 6.

Figura 5: Histograma de los datos transformados con $log(x)$.

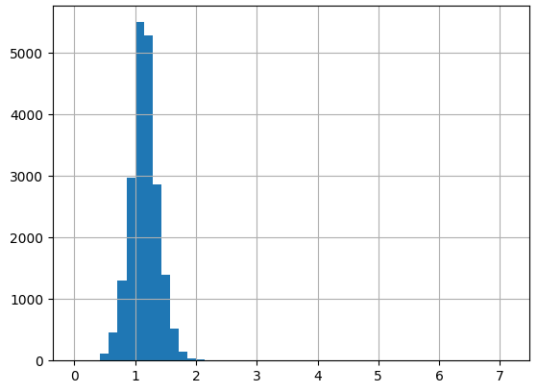
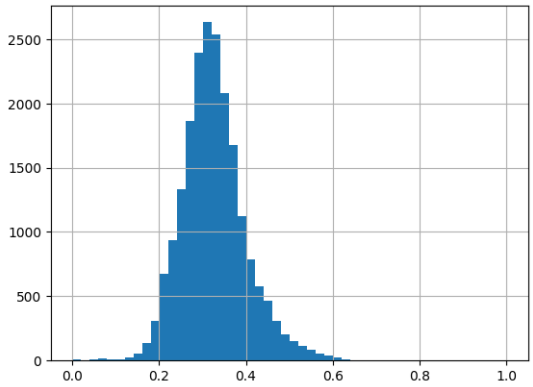


Figura 6: Histograma de los datos transformados con $1/x$.



Existen muchas otras transformaciones para compensar el sesgo positivo como las raíces $x' = \sqrt{x}$, $x' = \sqrt[3]{x}$; la función *logit*: $x' = ln(\frac{x}{1-x})$, y la transformación $x' = arcsin(\sqrt{x})$, estas últimas validas para $x \in (0, 1)$ [2].

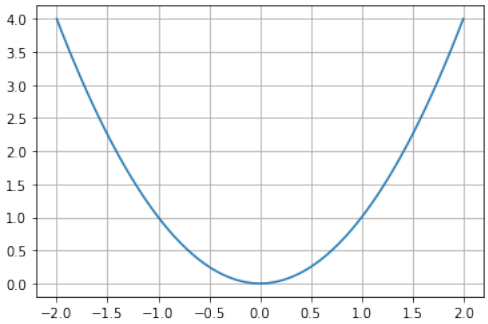
2. Transformación cuadrática y cúbica

Función cuadrática: Si se tienen datos cuya distribución presenta una cola cargada al lado izquierdo, en contraposición a la raíz cuadrada, considere la función cuadrática

$$x' = x^2.$$

- Los valores pequeños, entre (-1,1) son atenuados, los valores mayores que uno, en magnitud, se amplifican. La Figura 7 muestra los detalles.

Figura 7: Gráfica de la función x^2.

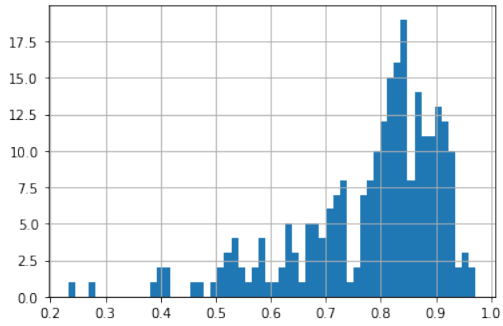


- Es útil para lidiar con valores que presentan un sesgo negativo (cola a la izquierda). La Tabla 4 muestra el inicio y el final de una distribución de datos cuyo histograma presenta un sesgo negativo, éste se muestra en la Figura 8.
- Igual que las transformaciones logarítmica y recíproca, una buena práctica es desplazar los datos numéricos por encima de 1.

Tabla 4: Fragmento de datos con sesgo negativo.

0	0.232
1	0.279
2	0.390
3	0.393
4	0.400
265	0.950
266	0.951
267	0.952
268	0.962
269	0.972

Figura 8: Histograma de los datos originales.

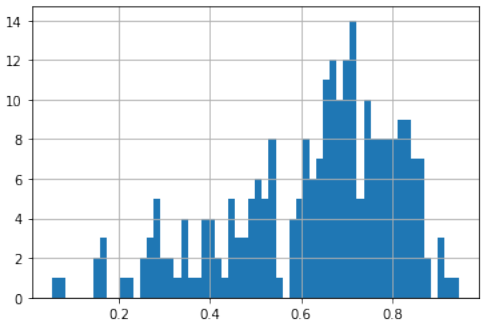


- Los datos originales se hacen pasar por la función cuadrática, resultando los mostrados en la Tabla 5, y su histograma en la Figura 9.

Tabla 5: Fragmento de los datos transformados por la función cuadrada.

0	0.054
1	0.078
2	0.152
3	0.154
4	0.160
265	0.903
266	0.904
267	0.907
268	0.926
269	0.945

Figura 9: Histograma de los datos transformados por x^2.



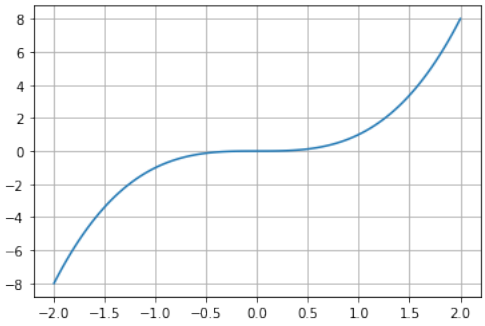
Función cúbica: Una transformación más de potencia. Tiene el mismo efecto que la función cuadrática: para variables con densidad de distribución que presentan una cola más alargada del lado izquierdo,

las distribuye de forma que los valores pequeños sean más frecuentes. La transformación cúbica se expresa como:

$$x' = x^3.$$

- Es más potente que la transformación cuadrática. Los números pequeños, al multiplicarse por ellos mismos, se hacen cada vez más pequeños, el resultante se atenúa aun más, es decir atenúa directamente el resultado de la función cuadrática. En cambio, para números mayores que uno, la función cuadrática se ve amplificada proporcionalmente por el valor a ser transformado.
- Mantiene el signo de los valores negativos. Se debe considerar que **los valores negativos, menores que menos uno, también crecerán en magnitud** por lo que se alejarán aún más hacia el lado izquierdo, por lo que se consigue el efecto contrario. La Figura 10 muestra el resultado de ser evaluada en el intervalo $[-2, 2]$.

Figura 10: Gráfica de la función x^3.

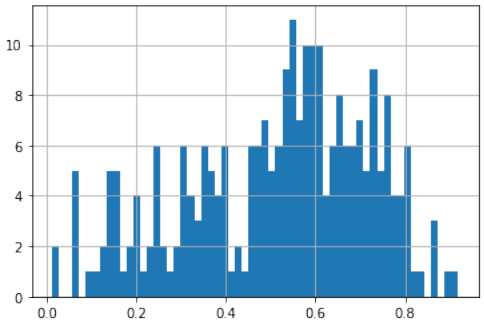


- Los primeros cinco y los últimos cinco elementos transformados se muestran en la Tabla 6, mientras que la representación gráfica de su frecuencia se muestra en la Figura 11.

Tabla 6: Fragmento de los datos transformados por la función cúbica.

0	0.013
1	0.022
2	0.059
3	0.061
4	0.064
265	0.859
266	0.859
267	0.863
268	0.890
269	0.918

Figura 11: Histograma de los datos transformados por x^3.



3. Jerarquía en las transformaciones de potencia

Se pueden englobar la mayoría de las transformaciones en la siguiente expresión:

$$x' = x^\lambda,$$

únicamente variando λ, si $\lambda = -1$, la transformación corresponde a la transformación recíproca; si $\lambda = 2$, a la transformación cuadrad, etcétera. Por lo que esta transformación compensa tanto sesgos positivos como negativos y, dependiendo de que tan sesgada es la variable a transformar, el exponente λ se irá alejando cada vez más de 1, como lo muestra la tabla 7. Esto evitando, naturalmente, $\lambda = 0$ y $\lambda = 1$; en la siguiente sección se incluirá el exponente en cero con una liguera modificación en la familia de transformaciones.

Tabla 7: Jerarquía de elección y efecto de la potencia λ.

Sesgo (s)	**Potencia** (λ)
$s\acute{}0$	$\lambda\check{}1$
$s<0$	$\lambda>1$
$s>0$	$\lambda<0$
$s\acute{}0$	$\lambda\acute{}0$

4. Transformación Box–Cox

Para escoger el mejor λ en la familia de transformaciones de potencia, se realiza una modificación de modo que, al hacer $\lambda \to 0$, la transformación tiene a $\ln(x)$ [4]. Esta modificación es la propuesta por los profesores David Cox y George Box para definir un criterio de optimización continuo con respecto a λ:

$$x' = \frac{x^\lambda - 1}{\lambda}.$$

Explorando el límite

$$\lim_{\lambda\to 0} \frac{x^\lambda - 1}{\lambda} = \lim_{\lambda\to 0} \ln(x) x^\lambda;$$

por lo que se tiene la transformación Box–Cox:

$$x' = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \ln(x) & \lambda = 0 \end{cases}. \tag{1}$$

Para calcular la potencia óptima se suponen los siguientes puntos:

- Los elementos en el conjunto a transformar (X) son independientes entre sí,
- todos los elementos a transformar tienen la misma distribución desconocida $f(x)$,
- se conocen la media (μ_x) y desviación estándar (σ_x) de X.
- La variable transformada tiene distribución Normal,
- la media aritmética (μ) y la desviación estándar (σ) de la variable transformada son desconocidas.

La potencia, λ, se calcula como aquella que maximiza la probabilidad conjunta de que la variable transformada tenga distribución Normal, es decir

$$\max_\theta \prod_{i=1}^{n} P\{x < x_i|\theta\} = \frac{1}{(\sqrt{2\pi}\sigma)^n} \int_{-\infty}^{x} e^{-\frac{(\mu - x')^2}{2\sigma^2}} dx,$$

donde $\theta = [\mu, \sigma, \lambda]^T$ y x' calculada como en (1). Box y Cox desmuestran en su artículo que los parámetros θ se pueden calcular por máxima log–verosimilitud, sin embargo, μ y σ se estiman a partir de μ_x y σ_x para la distribución Normal de forma iterativa calculando λ.

Las características de la transformación son las siguientes:

- Se define únicamente para valores de x positivos. Preferentemente $x \in [1, \infty)$.
- Es una transformación flexible, si se escoge $\lambda = 0$ se tiene la transformación logarítmica.
- Con $0 < \lambda < 1$ se obtienen las transformaciones por raíces ($x' = \sqrt{x}$, $x' = \sqrt[3]{x}$, etc.).
- Al escoger $\lambda < 0$ se obtienen las transformaciones potencia de la función recíproca ($x' = 1/x^2$, $x' = 1/\sqrt{x}$, etc.), específicamente con $\lambda = -1$ se consigue la transformación recíproco.
- Finalmente, para $\lambda > 1$ la transformación se vuelve del tipo cuadrática o cúbica. Las gráficas con diferentes λ se muestran en la Figura 12.
- Como ejemplo, la Tabla 8 muestra los datos ordenados de una distribución cuya frecuencia se ve en la Figura 13.

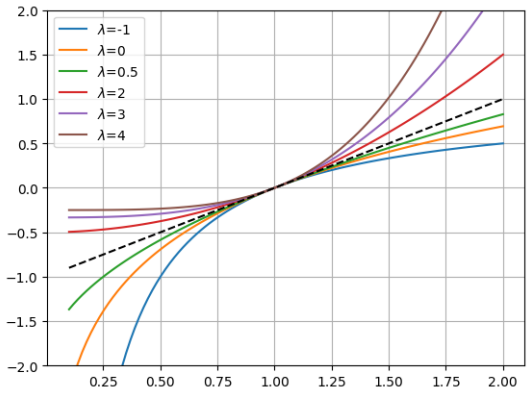
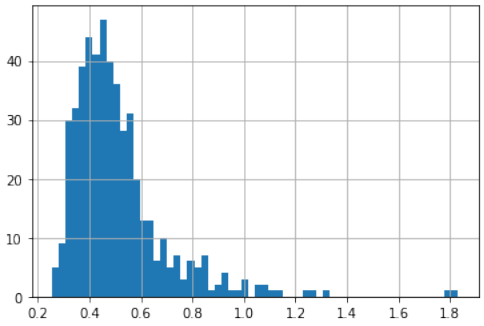


Figura 12: Gráfica de la transformación Box–Cox

Tabla 8: Fragmento de los datos sin transformar.

0	0.255
1	0.262
2	0.268
3	0.276
4	0.281
495	1.235
496	1.259
497	1.308
498	1.801
499	1.832

Figura 13: Histograma de los datos sin transformar.



- Intuitivamente se escoge $\lambda = 0$ para realizar una transformación logarítmica, ésta se muestra en la Figura 14.
- Los datos se transforman nuevamente con $\lambda = -0.9328$ obteniendo una distribución con una apariencia más cercana a la Gaussiana (Figura 15).

Figura 14: Histograma de los datos transformados por Box–Cox con $\lambda = 0$.

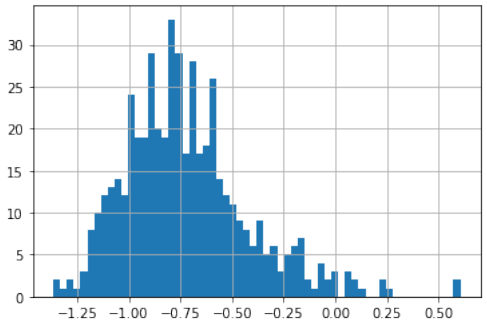
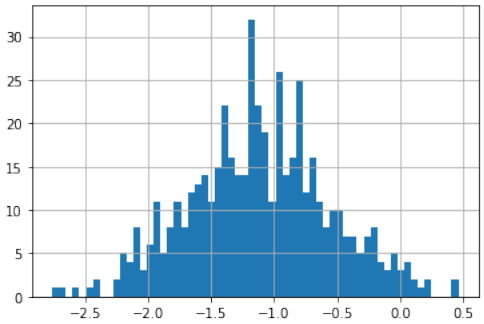


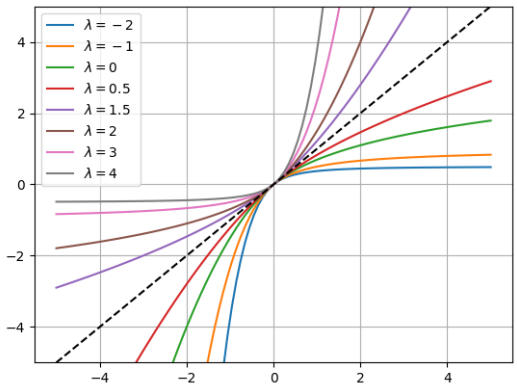
Figura 15: Histograma de los datos transformados por Box–Cox con $\lambda = -0.9328$.



5. Transformación Yeo–Johnson

Una extensión de la transformación Box–Cox para variables con valores negativos es la transformación Yeo–Johnson [5]

Figura 16: Representación gráfica de la transformación Yeo–Johnson.



- Siempre se puede desplazar la variable de forma que presente solo valores positivos para usar Box–Cox.
- La transformación se define como

$$x' = \begin{cases} \dfrac{(x+1)^{\lambda}-1}{\lambda} & x \geq 0,\ \lambda \neq 0 \\ \ln(x+1) & x \geq 0,\ \lambda = 0 \\ -\dfrac{(-x+1)^{2-\lambda}-1}{2-\lambda} & x < 0,\ \lambda \neq 2 \\ -\ln(-x+1) & x < 0,\ \lambda = 2 \end{cases}$$

- Note que cada curva en la Figura 16 se mantiene por debajo o por encima de la identidad $x = x$ (línea discontinua). Esto asegura que la compensación del sesgo es la misma para valores negativos y positivos.
- El primer criterio para identificar la función a ser ejecutada es el signo del elemento x después se usa λ bajo el mismo criterio que Box–Cox para estabilizar la varianza.
- Figura 16 muestra la evaluación de esta transformación en el intervalo $[-4, 4]$.
  - Note la continuidad entre valores negativos y positivos.
  - Tiene ambos efectos: Atenuar valores grandes ($\lambda = \{0.0, 0.5\}$) y amplificar valores grandes ($\lambda = \{1.5, 2.0\}$).
  - La función que se ejecuta cambia dependiendo del signo de x. Si se escoge $\lambda = 0$, la función con $x \geq 0$ será logaritmo, mientras que para $x < 0$ será cuadrática.

- Considere la densidad de distribución de la Figura 17, contiene valores negativos y un sesgo negativo, la Figura 18 muestra el efecto de la transformación con $\lambda = 3.074$

Figura 17: Densidad de distribución de una variable sin transformar.

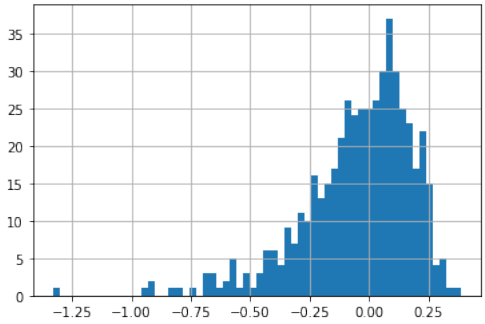
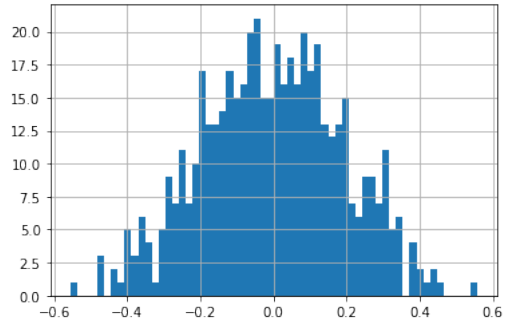


Figura 18: Histograma de la variable transformada con Yeo–Johnson usando $\lambda = 3.074$.



Referencias

[1] `https://www.cienciasinseso.com/transformacion-de-datos/`

[2] Kuhn, M. & Johnson K. (2019). Feature Engineering and Selection: A Practical Approach for Predictive Models. CRC Press. 121–126.

[3] Zheng A. & Casari A. (2018). Feature Engineering for Machine Learning. Principles and Techniques for Data Scientist. O'Reilly. 15–29.

[4] Box, G. E. P. & Cox, D. R. (1964). An Analysis of transformations.Journal of the Royal Statistical Society, Serie B, 26. 211–252.

[5] Yeo, I. K., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. Biometrika, 87(4), 954–959.

[6] Galli, S. (2022). Python feature engineering cookbook: over 70 recipes for creating, engineering, and transforming features to build machine learning models. Packt Publishing Ltd. 77–107.