Tarea 1

Autor

Juan Román Uribe

Sebastián Mena Aliaga

Profesor

Christian Araya Muñoz

Modelación Estadística Aplicaciones Multidisciplinaria

Magister en Estadística, PUCV

Agosto 2021

Introducción

El presente trabajo tiene como objetivo realizar un Análisis Exploratorio de Datos (AED) y responder, a través de evidencia estadística (pruebas de hipótesis), a las preguntas de la tarea 1, que a su vez fueron planteadas por la asesoría de la universidad hacía un puerto de camiones en Valparaíso, con el fin de comprender con mayor objetividad el comportamiento de los tiempo de los camiones que cumplen su estadía en dicho puerto.

1 Pregunta 1

1.1 Diccionario de Datos

A continuación se muestran los descriptivos de las principales variables utilizadas en el *script* de R y, en consecuencia, el nombre utilizado para las variables dentro de esta tarea:

- tiempo_1 : Tiempo en minutos desde que el camión entra al recinto, hasta que el camión se posiciona en el pórtico de control de acceso.
- tiempo_2 : Minutos desde que el camión entra y sale del pórtico de control de acceso.
- tiempo_3 : Minutos de espera y atención en la oficina 1.
- tiempo_4 : Minutos de espera y atención en la oficina 2.
- tiempo_5 : Minutos desde que el camión sale de oficina 2 y luego sale del recinto.
- tipo_carga : Tipo de cargamento que transporta el camión.
- turno : Turno en cual ingresa el camión.
- responsable : Responsable en registrar tiempos de tiempo_2.
- responsable_O1: Responsable en registrar tiempos de delta_3 (Oficina 1).
- Responsable_O2: Responsable en registrar tiempos de delta_4 (Oficina 2).

1.2 Análisis Exploratorio de Datos (AED)

En esta sección se desarrolla una AED para cada *tiempo_n* y el tipo de carga de los camiones, a partir de gráficos y resúmenes estadístico, y, a su vez, se verificará la presencia de datos atípicos y valores perdidos (NA) para cada variable, y se tratarán según corresponda.

1.2.1 AED sobre tiempo_1

Al analizar la variable de *tiempo*_1, se presencia una alta cantidad de tiempos no registrados (NA) en la hora de acceso, con un total de 471 NA observados, además se observan 89 *outliers* cuando se omiten los 471 datos perdidos. A modo de corrección de los casos con datos perdidos, en primer lugar, se trata la sustitución de cada NA por la media ($Media = 4.996$), sin embargo, tal como se observa a la izquierda de la figura 1 (antes de la corrección de NA) y a la derecha (luego de corregir los NA por la media), la corrección provoca un cambio brusco en la disposición de los datos en el histograma. Esto último se debe al uso de un estadístico, como la media, que es muy sensible ante muestras con alta presencia de *outliers*, por lo tanto se rechaza su uso como sustituto para los datos perdidos. Cabe mencionar que, para la figura 1, dada su alta presencia de datos atípicos, se ha corregido deliberadamente el tamaño del eje X para facilitar la visualización del gráfico.

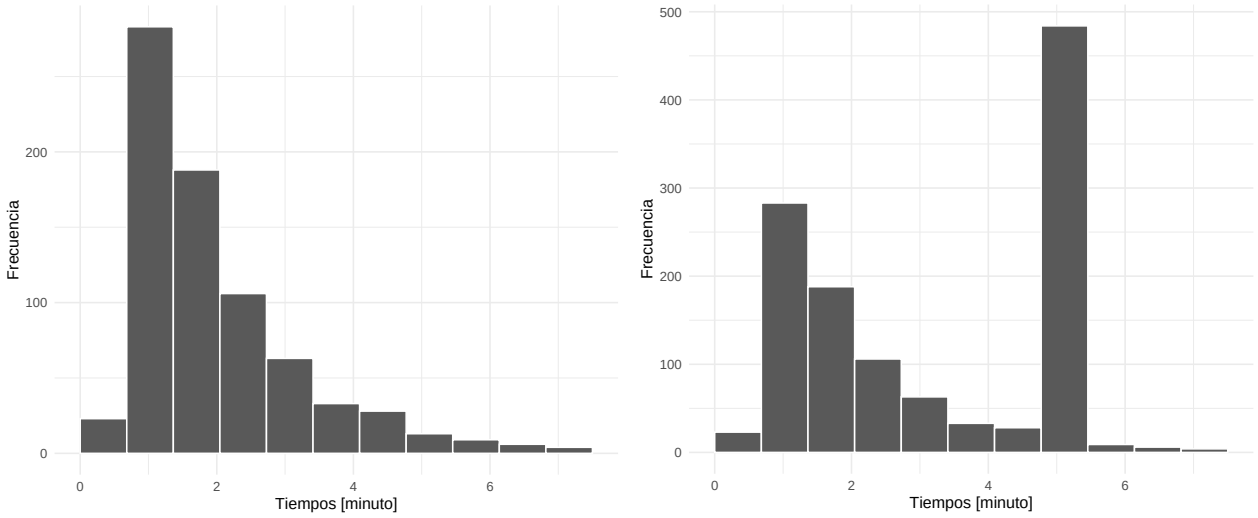

Figure 1: Izquierda, sin corrección de NA. Derecha, NA corregido por media

Se decide corregir los valores perdidos el valor de la mediana ($Mediana = 1.65$), debido a su resistencia ante la presencia de *outliers*, logrando un comportamiento de la curva más similar a la data sin NA, ver figura 2.

Min	0.333
1er Qu.	1.400
Mediana	1.650
Media	3.778
3er Qu.	2.062
Max	414.250
Desv.	15.490
CV	3.083

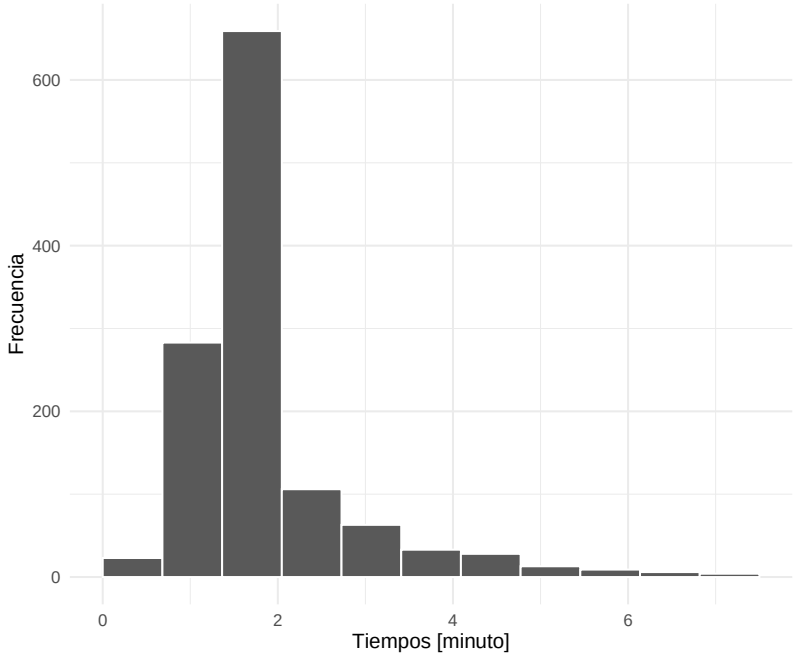

Figure 2: Resumen estadístico e histograma de tiempo_1, *outliers* corregido q por mediana

Si comparamos el gráfico teórico de la curva de densidad de la exponencial (a cualquier parámetro de esta) con el histograma de la muestra (figura 2), se puede observar que no pareciera tener un comportamiento exponencial, sin embargo, un análisis visual queda limitado a la hora de fundamentar un hecho objetivamente, es por ello que se pone a prueba la hipótesis de que los datos siguen una distribución exponencial, a partir de un test de bondad de ajuste. Al realizar el test de Kolmogorov-Smirnov, utilizando el parámetro estimado de la exponencial $\lambda = 0.2647$, y con una confianza del 95%; los resultados entregan un estadístico $D = 0.333$ y un $p - valor < 2.2x10^{-16}$, lo que implicaría un contundente rechazo a la hipótesis. No obstante, se sospecha que los *outliers* podrían tener una alta implicancia en los resultados del test, por lo que se emplea otras técnica para evaluar la distribución de la muestra, tal como los gráficos de Q-Q Plot, P-P Plot, y tal como se puede observar en el gráfico de la curva de densidad empírica y teórica, y el Q-Q Plot de la figura 3, la dificultad de aplicar el test con la presencia de los datos atípicos.

Para profundizar el análisis, se remueven los 187 *outliers* de los 1294 datos de la muestra, valor considerado aceptable dada la cantidad de datos que quedan luego de la sustracción. Luego se gráfica la curva empírica y teórica de la exponencial, sin los *outliers* (ver figura 4), se puede observar en el gráfico de densidad una alta concentración de datos entre el tiempo 1.5 y 2 minutos, esto debido a la decisión anterior de corregir los 471 datos faltantes por un único valor de la mediana, evidenciando un problema en dicha acción y planteándose la oportunidad de aplicar métodos más sofisticados en el tratamiento de los faltantes, análisis que queda fuera de los alcances esta tarea.

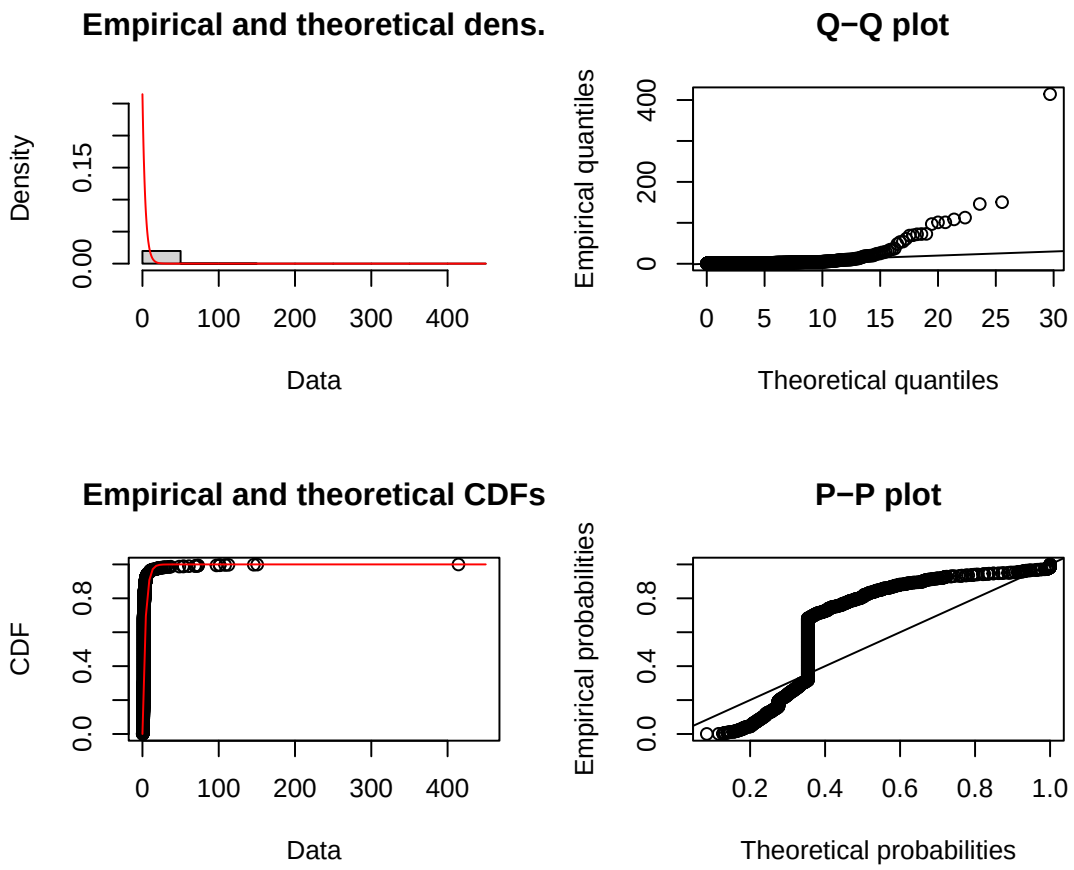

Figure 3: Izquierda, gráficos teóricos de la curva exponencial en contraste a la muestra. Derecha, Q-Q PLot y P-P Plot. Muestra con *outliers*.

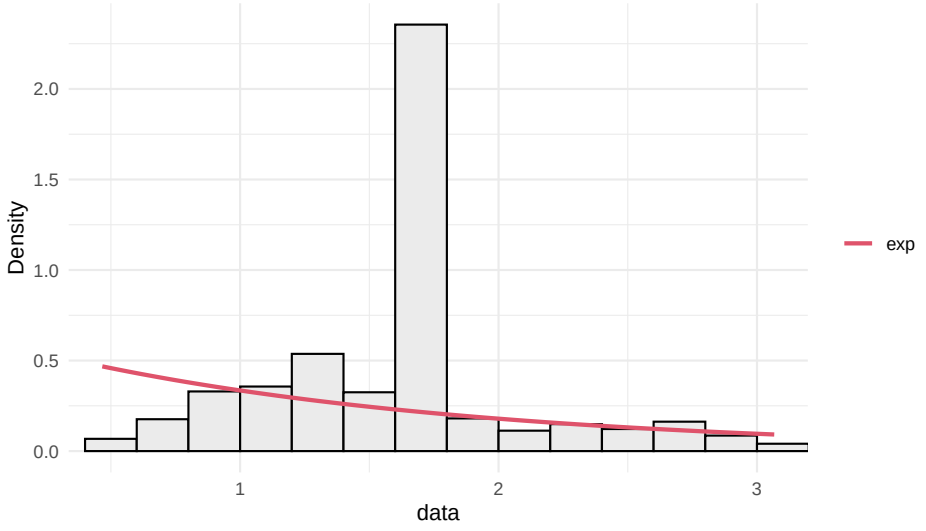

Figure 4: Gráficos teóricos de la curva exponencial en contraste a la muestra, sin *outliers*.

1.2.2 AED sobre tiempo_2

Para el caso del *tiempo*_2, se observan 69 *outliers*, incluyendo un caso totalmente atípico de un valor de tiempo negativo de -0.41667, siendo este valor totalmente imposible dada la naturaleza del ejercicio, por lo que se sustituye dicho valor por la mediana al igual que la variable anterior ($Mediana = 1.65$). Luego de corregir el valor negativo por su mediana se presenta, en la figura 5, el resumen estadístico e histograma del para el *tiempo*_2, corrigiendo además el tamaño del eje X para facilitar la visualización del gráfico, se observa poco probable que los datos sigan una distribución exponencial.

Min	0.333
1er Qu.	0.933
Mediana	1.333
Media	1.840
3er Qu.	1.812
Max	238.8
Desv.	7.093
CV	3.854

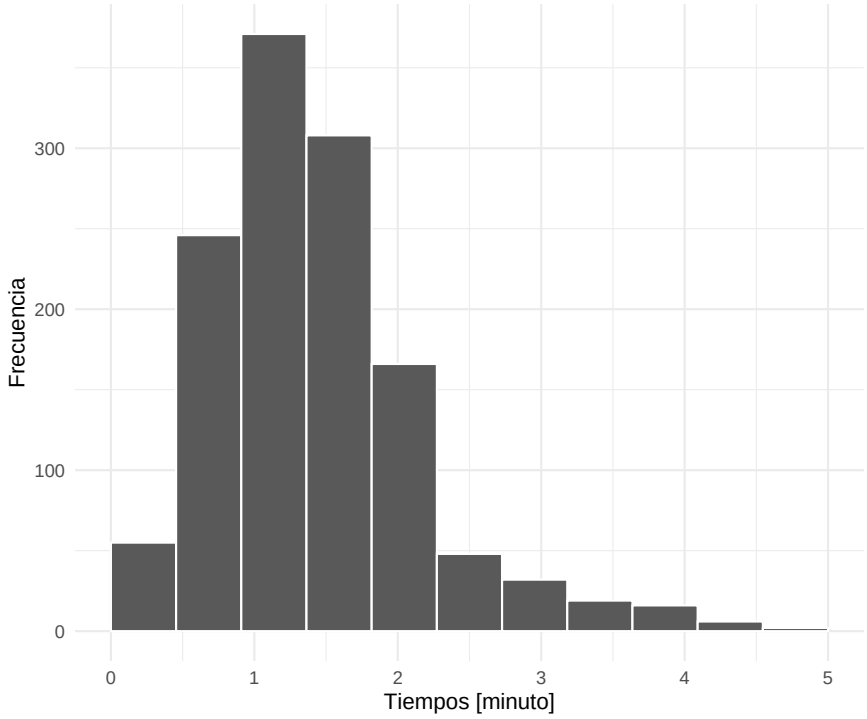

Figure 5: Resumen estadístico e histograma de tiempo_2, *outliers* corregido q por mediana

Al igual que en el caso del *tiempo*_1, la presencia de *outliers* en la muestra genera ruido a la hora de probar la hipótesis de que los datos sigan una distribución exponencial, por lo que, para este caso son removidos de la muestra, además que su remoción no considera un cambio brusco en el tamaño de la muestra. Al realizar el test de Kolmogorov-Smirnov, utilizando el parámetro estimado de la exponencial $\lambda = 0.747$, y con una confianza del 95%; los resultados entregan un estadístico $D = 0.269$ y un $p - valor < 2.2x10^{-16}$, y complementando los resultados del test, tal como se observa en los gráficos de Q-Q Plot,P-P Plot, y de densidad teórica versus empírica (ver figura 6), se puede argumentar categóricamente que los datos no se distribuyen de forma exponencial. Al realizar un análisis un poco más profundo probando diversas distribuciones, se ha concluido que la distribución logística con parámetros de localización 1.3098 y escala de 0.3373, es la más adecuada para describir esta muestra de datos, sustentado con K-S de $p - valor = 0.0862$ y evidencia gráfica en la figura

7.

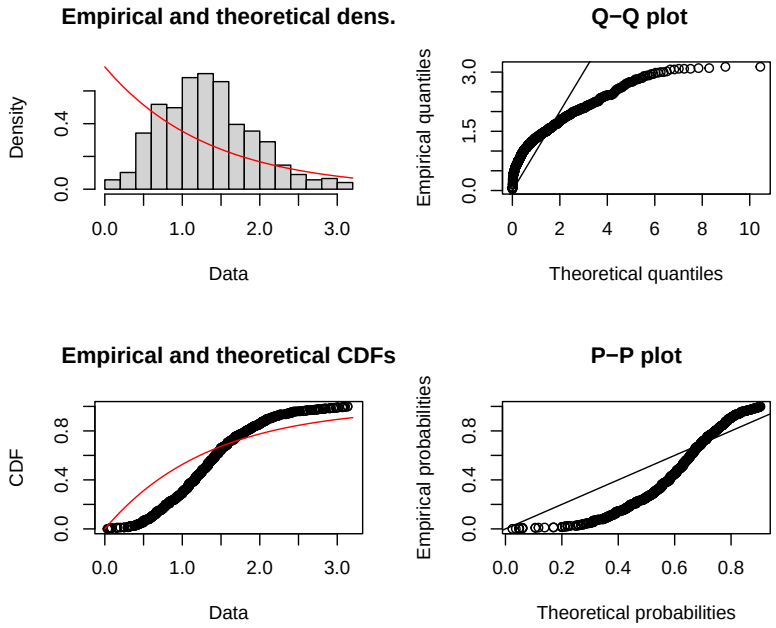

Figure 6: Izquierda, gráficos teóricos de la curva exponencial en contraste a la muestra. Derecha, Q-Q PLot y P-P Plot. Muestra sin *outliers*.

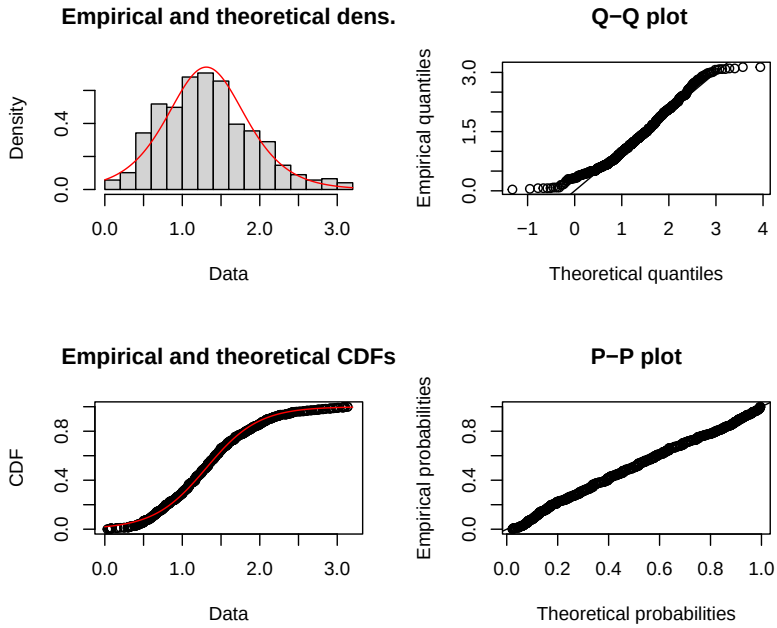

Figure 7: Izquierda, gráficos teóricos de la curva logística en contraste a la muestra. Derecha, Q-Q PLot y P-P Plot. Muestra sin *outliers*.

1.2.3 AED sobre tiempo_3

Al realizar la exploración para *tiempo_3* (ver figura 8), en primer lugar, no se observan datos atípicos. Al estudiar la posibilidad de un comportamiento exponencial, a través del test K-S, se obtienen con un parámetro $\lambda = 0.119$, un estadístico $D = 0.130$ y un $p - valor < 2.2x10^{-16}$, además al comprar gráficamente la curva teórica y experimental, se tiene argumentos suficientes para rechazar el comportamiento exponencial.

Min	0.000
1er Qu.	2.007
Mediana	8.315
Media	8.396
3er Qu.	13.610
Max	23.670
Desv.	6.445
CV	0.767

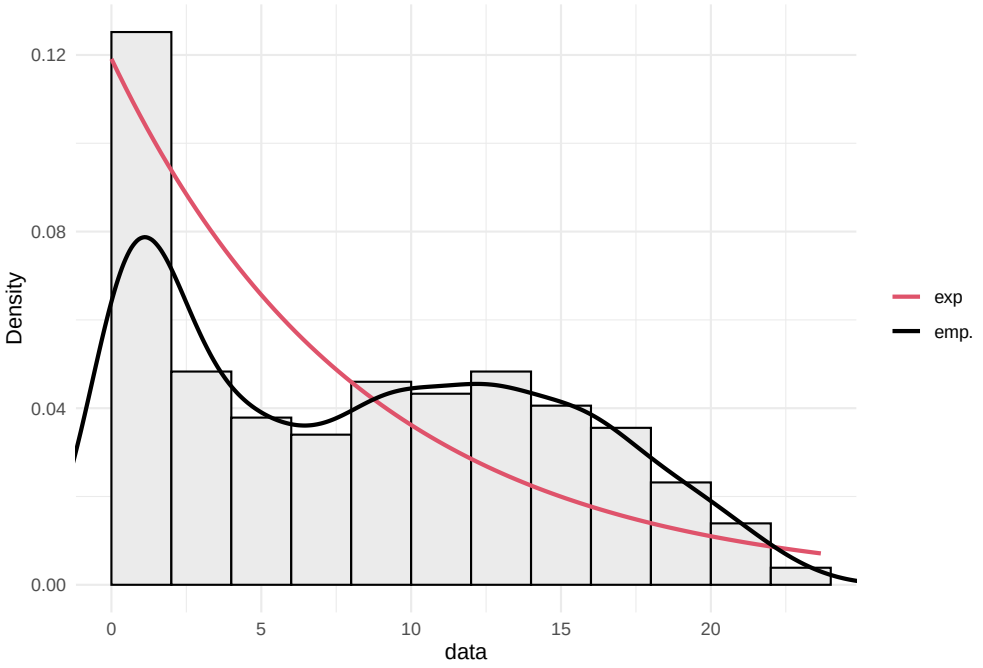

Figure 8: Resumen estadístico e histograma de tiempo_3.

1.2.4 AED sobre tiempo_4

De la muestra, en el histograma de la figura 9, se observa un posible comportamiento normal, sin embargo, esta suposición se cae al observar la evidencia estadística de K-S con un parámetro $\lambda = 0.044$, un estadístico $D = 0.112$ y un $p - valor < 1.7x10^{-14}$. La muestra no posee outliers.

1.2.5 AED sobre tiempo_5

Del *tiempo_5* se puede observar, de la figura 10, que los datos no siguen una distribución exponencial, supuesto que se corrobora con un 95% de confianza al realizar el test K-S con parámetro $\lambda = 1.959$, los resultados de $D = 0.176$ y $p - valor = 2.2x10^{-16}$, rechazando la hipótesis de un comportamiento exponencial.

Min	0.000
1er Qu.	3.445
Mediana	17.045
Media	22.639
3er Qu.	36.540
Max	83.570
Desv.	21.308
CV	0.491

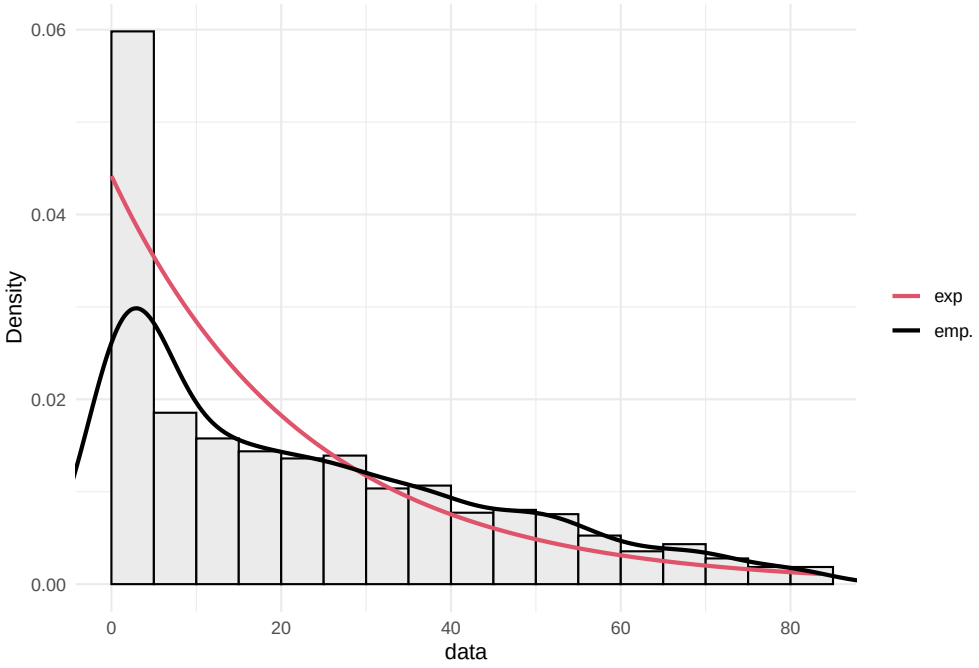

Figure 9: Resumen estadístico e histograma de tiempo_4

Min	0.000
1er Qu.	0.2500
Mediana	0.5200
Media	0.5104
3er Qu.	0.7600
Max	1.0000
Desv.	0.293
CV	0.575

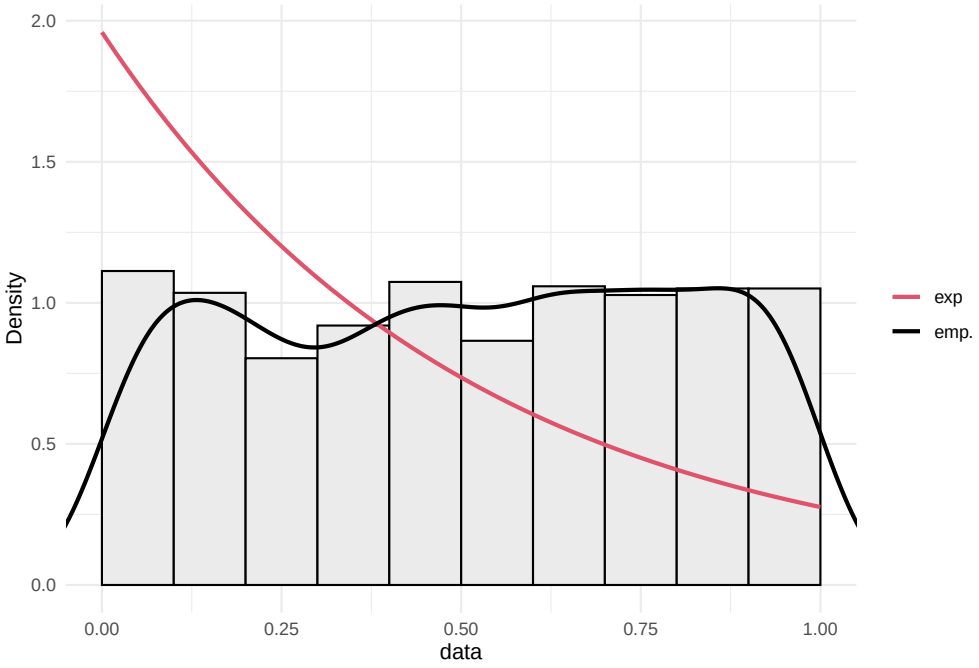

Figure 10: Resumen estadístico e histograma de tiempo_5

1.2.6 AED el tiempo de ciclo

Se define el tiempo de ciclo como la suma total de las 5 variables de tiempos del sistema. La variable cuenta con 6 *outliers*, datos que se retiran de la muestra. Gráficamente, como se puede observar en el histograma de la figura 11, esta variable pareciera tener un comportamiento exponencial pero, tal como ocurre en la variable *tiempo_4*, la hipotesis se rechaza al realizar el test K-S con parámetro $\lambda = 0.027$, los resultados de $D = 0.080$ y $p - valor = 1.346x10^{-7}$.

Min	1.827
1er Qu.	11.492
Mediana	31.280
Media	36.292
3er Qu.	56.224
Max	119.603
Desv.	27.548
CV	0.759

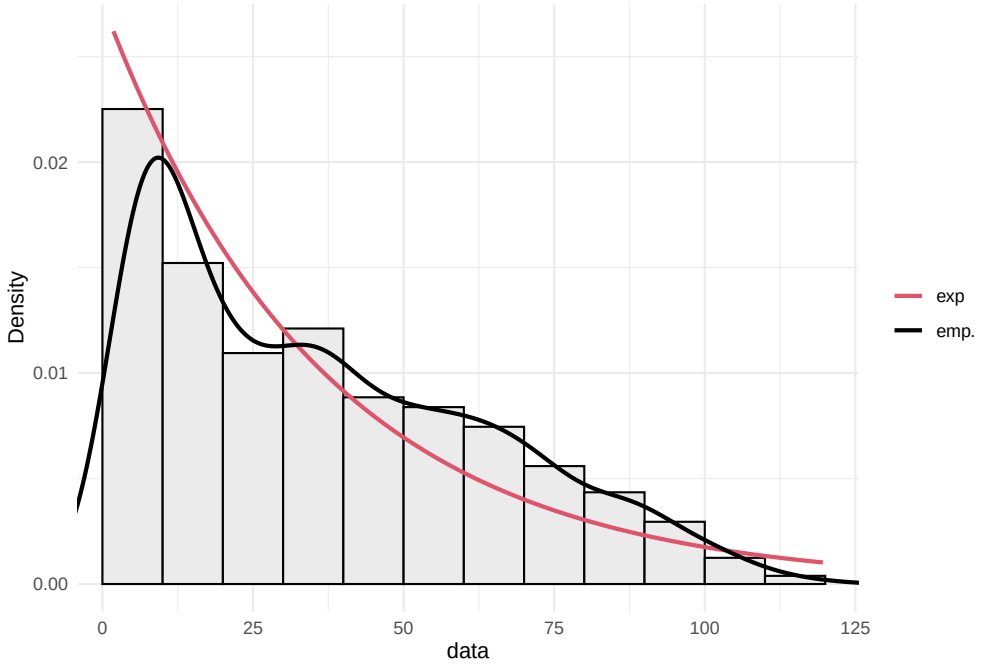

Figure 11: Resumen estadístico e histograma de tiempo de ciclo

1.2.7 AED sobre tipo de carga

Al analizar la variable de tipo de carga se observa principalmente que la carga de medicamento es lo más transportado, y los bienes de lujo como el cargamento menos transportado, registrado en la muestra

Tipo de Carga	Cantidad
No perecibles	109
Bienes de lujo	20
Medicamentos	587
Textil	261
Textos	317

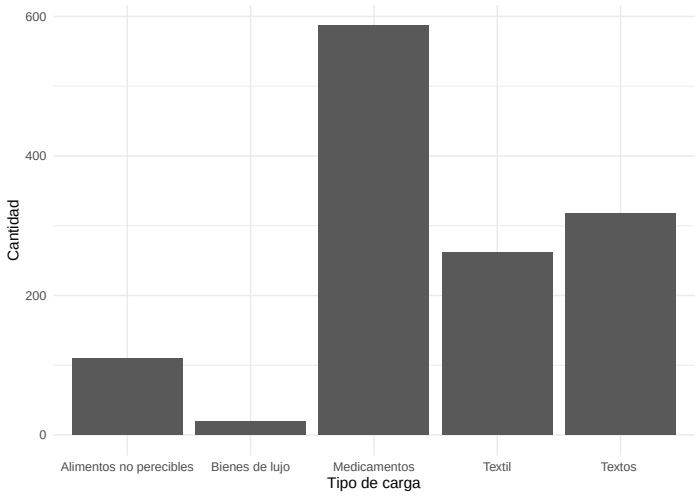

Figure 12: Resumen estadístico e histograma de delta_5

1.2.8 Diferencias significativas en los pórticos de entrada y salida

En esta subsección se desea probar si existen diferencias significativas en los tiempos de atención en los pórticos a la entrada y salida, osea, entre las variable *tiempo_1* y *tiempo_5*. Si comenzamos con un análisis visual, observando los histogramas de las figuras 4 y 10, nos podremos percatar a *priori* que no existe un comportamiento similar entre las variables, sin embargo, ante un estudio estadísticos rigurosos se hace necesario el uso de test de hipótesis.

A partir de un análisis K-S para probar la normalidad de ambas variables (ver tabla 1), se concluye que los datos no se distribuyen ni normal, ni exponencial como se evidenció anteriormente. Entonces, para evaluar diferencia significativas entre las muestras se opta por el uso del test no perimétricos de rangos con signos de Wilcoxon. La hipótesis que se evalúa es:

$$H_0 : Me_{tiempo\_1} - Me_{tiempo\_2} = 0$$
$$H_1 : Me_{tiempo\_1} - Me_{tiempo\_2} \neq 0$$

El test de Wilcoxon calcula un estadístico $W = 1401373$ y un $p - valor < 2.2x10^{-16}$, por lo tanto, con una confianza del 95% se descarta que sus medidas de posición sean idénticas, osea que posee diferencias significativas en los tiempos de atención.

Variable	Estadístico	p-valor
tiempo_1	0.126	$p - valor = 2.2x10^{-16}$
tiempo_5	0.074	$p - valor = 1.165x10^{-6}$

Table 1: Test K-S para evaluar normalidad de tiempos 1 y 2

2 Pregunta 2

2.1 ¿Es posible aseverar que existen diferencias en los tiempos de atención en los pórticos de ingreso, dependiendo del turno?

Los subconjuntos de la variable de tiempo_2 separados por los turnos T1, T2 y T3 poseen las cantidades muéstrales de 564, 495 y 235. Ninguno de los subconjuntos, luego de evaluar la bondad de ajuste de K-S, evidencia un comportamiento normal o exponencial, por otro lado, al realizar un test de homogeneidad de varianza de Levene centrada en la mediana para los 3 turnos, se obtiene los resultados de estadístico $F = 0.771$ y un $p - valor = 0.463$, por lo tanto, se concluye que no existen diferencias significativas entre las varianzas de los grupos. Por último, se observa en los boxplot de la figura 13 que existe una cierta asimetría hacía la derecha en los tres turnos, por lo tanto, se cumplen las condiciones para evaluar diferencias significativas entre los turnos utilizando el test de rango de Kruskal-Wallis. Al evaluar el test se obtiene un $p - valor = 0.0001028$, implicando que existen diferencias significativas en al menos dos de los tres grupos.

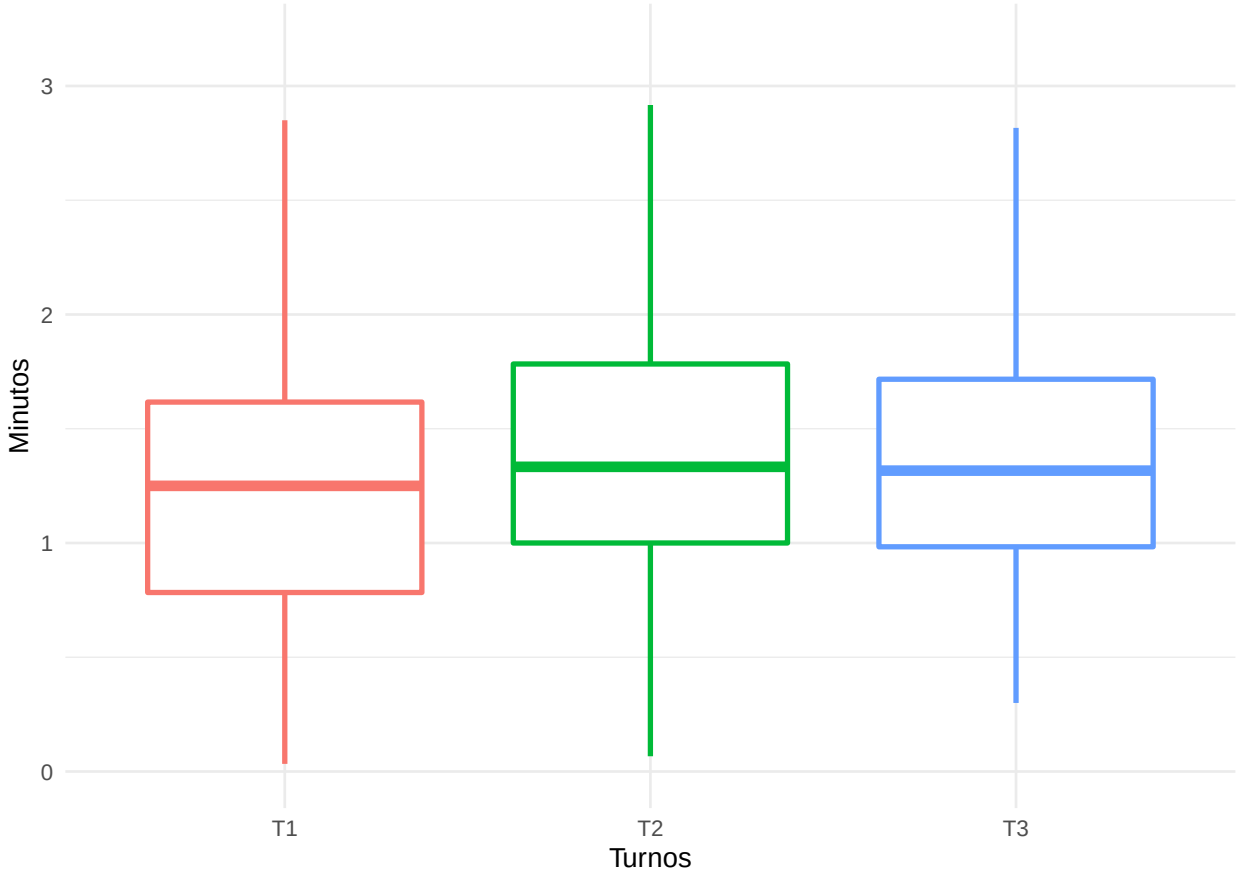

Figure 13: Boxplots de tiempo_2 por cada turno

3 Pregunta 3

3.1 ¿Es posible asegurar que no existe un sesgo asociado al investigador en cada estación en donde se midieron tiempos?

El sesgo en una investigación se asocia a errores sistemáticos dada alguna debilidad en la metodología de muestreo de los datos o en el procesamiento de estos, su efecto se mide cualitativamente. Por otro lado, se tiene el error aleatorio relacionado con variaciones producidas por el azar, que puede medirse cuantitativamente.

Respecto al error sistemático, error por el cual se pregunta en este apartado, se puede subdividir en tres tipos: dado el fenómeno medido, por instrumento de medición o dado por el investigador que realiza la medición. Desde este punto de vista, y dada la evidencia obtenida a partir del AED de la pregunta 1, tales como los NAs de la variable tiempo_1 y el valor negativo del tiempo_2, se sostiene que debe existir un sesgo asociado al investigador al momento de registrar los datos.

4 Pregunta 4

4.1 ¿Existe alguna relación entre el turno de ingreso al recinto y el tipo de carga que transporta el camión?

Dado que se desea estudiar una relación entre dos variables categóricas, se comienza el análisis a partir de una tabla de contingencia que contiene las cantidad entre las variables, tal como se observa en la tabla 2, los bienes de lujos cuentan con 20 registros, descartando el uso del test χ^2 para variables categóricas, y optando por el test exacto de Fisher dada su precisión para frecuencias bajas, sin embargo, es importante señalar que tradicionalmente el test de Fisher se utiliza para tablas 2x2 pero, es posible extender el tamaño de la tabla para el test utilizando simulaciones del p valor.

La hipótesis nula que se desea contrastar es si las variable de turno posee alguna relación con la de tipo de carga, y según los resultados del test de Fisher basado en la simulación del p-valor, se obtiene un $p - valor = 0.0004$, rechazando la hipótesis que las variables estén relacionadas.

	Alim. no perecibles	**Bienes de lujo**	**Medicamentos**	**Textil**	**Textos**	
T1	19	12	251	123	159	564
T2	64	3	191	104	133	495
T3	26	5	145	34	25	235
	109	20	587	261	317	1294

Table 2: Tabla de contingencia para turnos y tipo de carga

4.2 ¿Existe alguna diferencia en los tiempos de ciclo de un camión en el recinto, de acuerdo al tipo de carga que transporta?

Al estudiar el ciclo de un camión en el recinto por tipo de carga que transporta, se puede aseverar a partir de test de Kruskal-Wallis que los datos no siguen una distribución de normalidad o exponencial, además, realizando un test de homogeneidad de varianza a partir del test Levene centrada en la mediana, obteniéndose un estadístico $F = 0.770$ y un $p-valor = 0.545$, se puede concluir que no existen diferencias significativas en las varianzas entre tipos de carga.

Al visualizar los boxplot de la figura 14, se observa un posible comportamiento similar entre las categorías, sin embargo, al realizar el test de K-S para evaluar diferencias entre los tiempos de ciclo de las categorías, se puede aseverar con un $p - valor = 3.215x10^{-10}$ que existen diferencias significativas entre las categorías. Es de interés además, para profundizar análisis futuro, verificar diferencias entre pares de categorías, con la finalidad de detallar aún más el entendimiento de los datos.

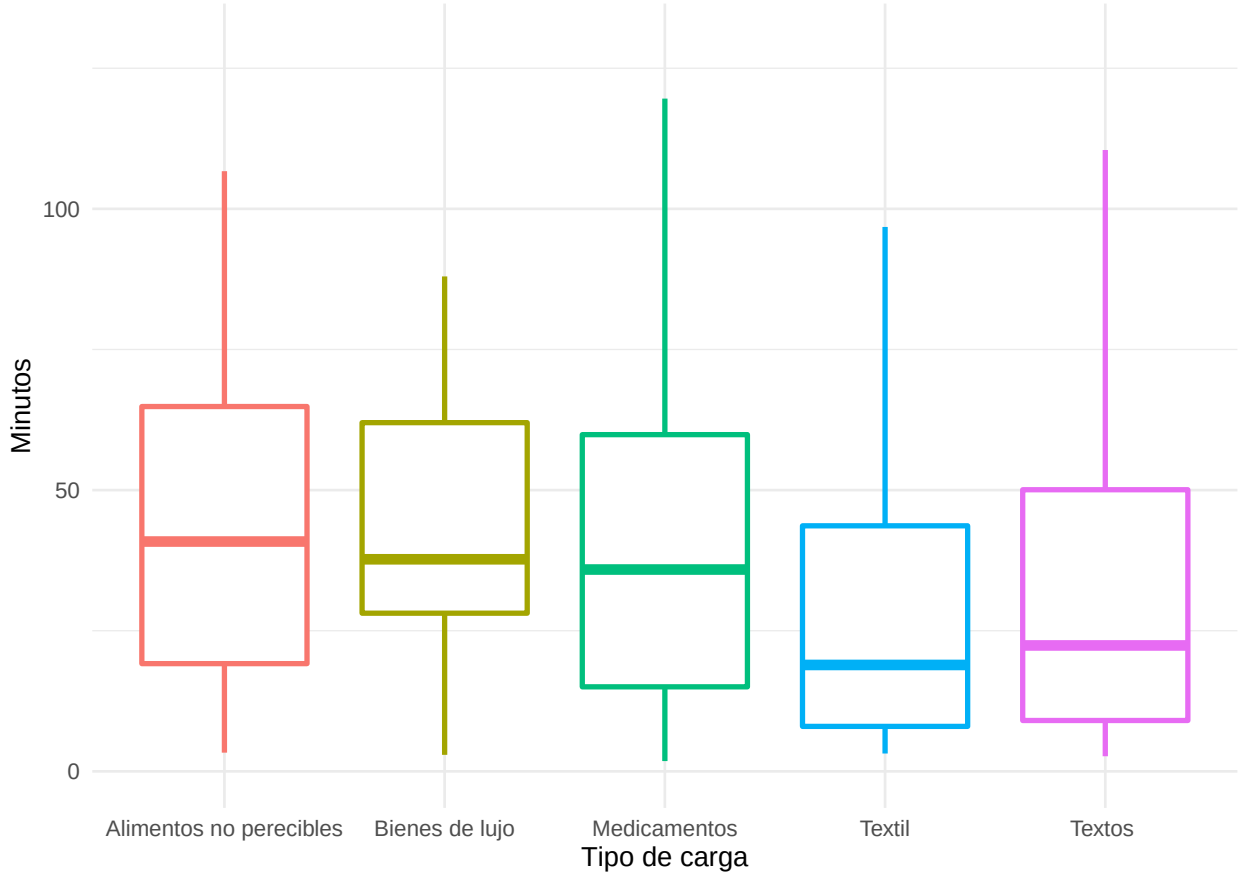

Figure 14: Boxplot de tiempos por tipo de carga.

5 Pregunta 5

5.1 ¿Existe evidencia suficiente que respalde algún tipo de asociación entre los tiempos de atención en los pórticos de ingreso, los minutos de atención en la oficina 1 (sin considerar tiempo de espera) y en la oficina 2?

Dado que anteriormente se ha rechazado la hipótesis tanto de normalidad y exponencial de los datos, se descarta aplicar el coeficiente de correlación de Pearson, y se opta por la alternativa no perimétrica, el coeficiente de correlación por rangos de Spearman para este análisis. Los resultados obtenidos del test (ver tabla 3) indican que, con una confianza del 95%, existe una correlación entre las variables de tiempo. Se puede observar gráficamente en la figura 16 la alta correlación que existe entre las variables tiempo_3 y tiempo_4.

Prueba	**p-valor**	ρ
tiempo_1 $\sim$ tiempo_3	0.0085	-0.0730
tiempo_1 $\sim$ tiempo_4	0.0057	-0.0767
tiempo_3 $\sim$ tiempo_4	$< 2.2x10^{-16}$	0,9708

Table 3: Test de correlación por método Spearman

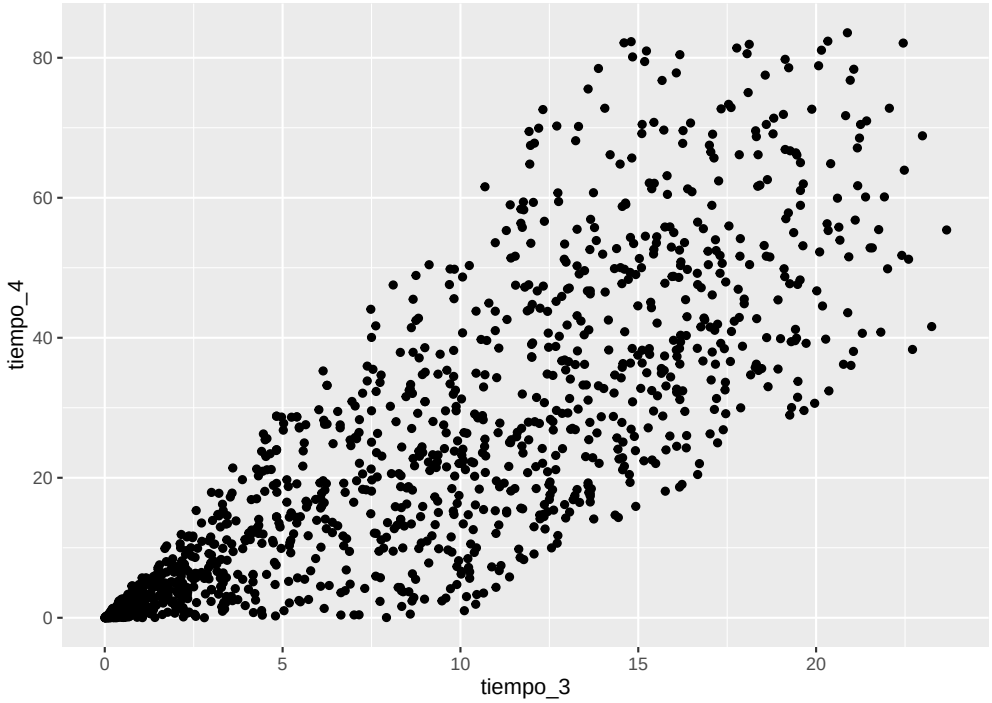

Figure 15: Gráfico de puntos para tiempo_3 y tiempo_4

6 Pregunta 6

6.1 ¿Qué estación podría estar generando los principales cuellos de botella del circuito, de acuerdo con la evidencia?

Se comienza con un análisis descriptivo para entender los cuellos de botellas de los tiempos del sistema y, dado que los tiempos 1 y 2 cuentan con una alta presencia de *outliers*, se trabajará con la mediana para obtener conclusiones sistema. Al calcular la mediana de los tiempos del 1 al 5, se obtienen los resultados de 3.778, 1.839, 8.396, 22.639 y 0.510 de la mediana respectivamente, además, al visualizar los boxplot de la figura 16, se puede observar que la variable tiempo_4 (tiempo en oficia 2) es la que aparentemente genera el mayor cuello de botella en el sistema. Sin embarga, tal como se ha planteado anteriormente, el análisis visual queda limitado por los sentidos, por lo que un contraste objetivo se ve fortalecido a través de un análisis estadístico de pruebas de hipótesis, en este caso se opta por utilizar Kruskal-Wallis para validar la hipótesis:

$$\begin{aligned} H_0: &\quad Me_{tiempo\_4} - Me_{tiempo\_3} \leq 0 \\ H_1: &\quad Me_{tiempo\_4} - Me_{tiempo\_3} > 0 \end{aligned}$$

Los resultados obtenidos del test de K-S con la alternativa hacía la derecha, son $D = 0.056$ y $p - valor = 0.016$, por lo tanto, a un $\alpha = 0.05$ se rechaza la hipótesis nula a favor de que el tiempo_4 es mayor al tiempo_3. En conclusión, es el tiempo en la oficina 2 la que genera un mayor cuello de botella en el sistema.

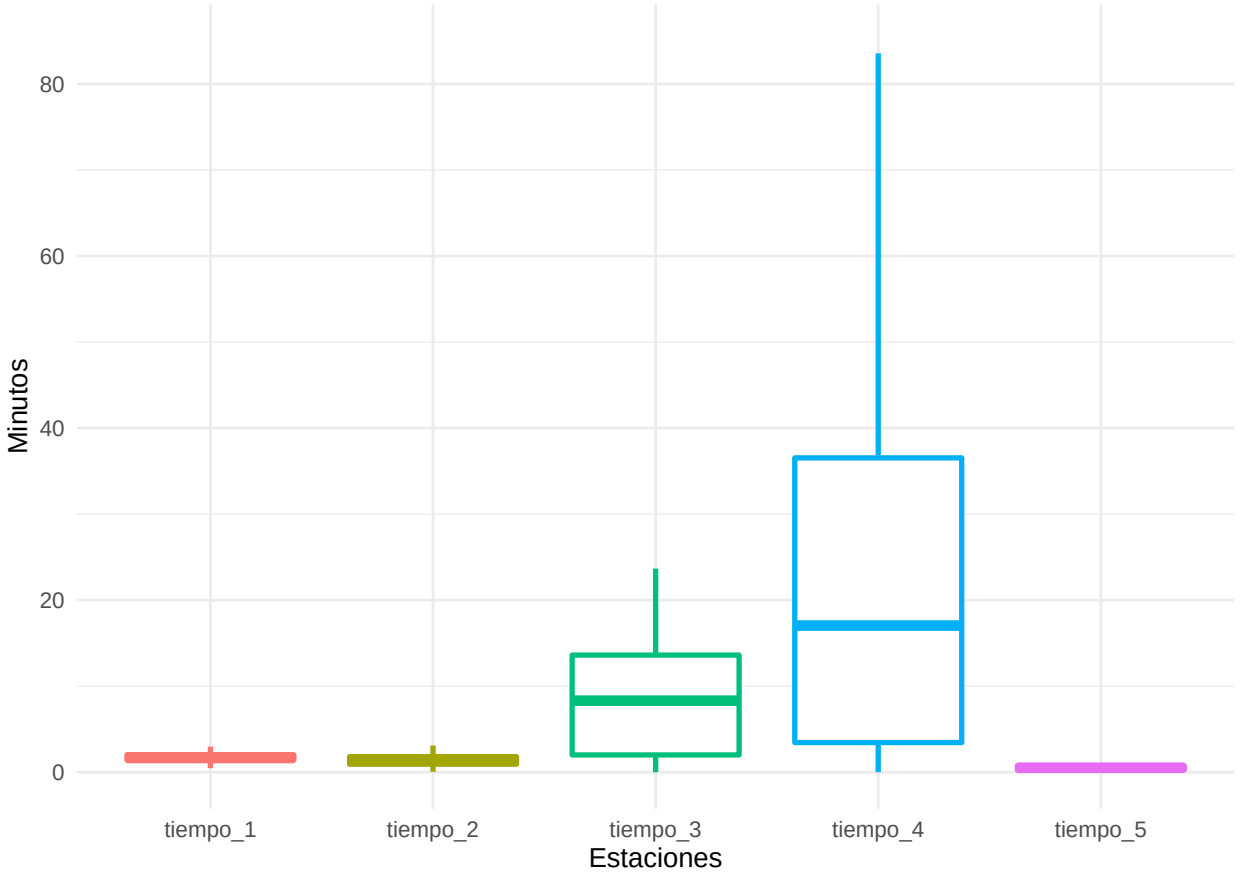

Figure 16: Boxplot de los tiempos del sistema.