Instituto Politécnico Nacional

Escuela Superior de Cómputo

Proyecto 2

Materia: Programación para Ciencia de Datos

Integrantes:

- **De Luna Ocampo Yanina**
- **Perea Samaniego Jesús Giovanni**
- **// Ruíz Aguilar Alex Gerardo**

Docente:

- **Galindo Durán Cristal Karina**

Grupo:

- **3AM1**

Fecha de entrega: 27/10/2021

ÍNDICE

Descripción de los datos

La base de datos a analizar corresponde al número de personas que ingresan a cada estación del metro diariamente del 1 de enero del 2021 al 30 de septiembre de 2021. Este dataset se llama “afluencia_metro” y está integrado por 53,236 registros.

Cada registro a su vez se compone de los atributos: índice, fecha (en formato compacto), día, mes, año, línea, estación y afluencia. Cada uno de los atributos enlistados sigue los parámetros de comportamiento descritos a continuación:

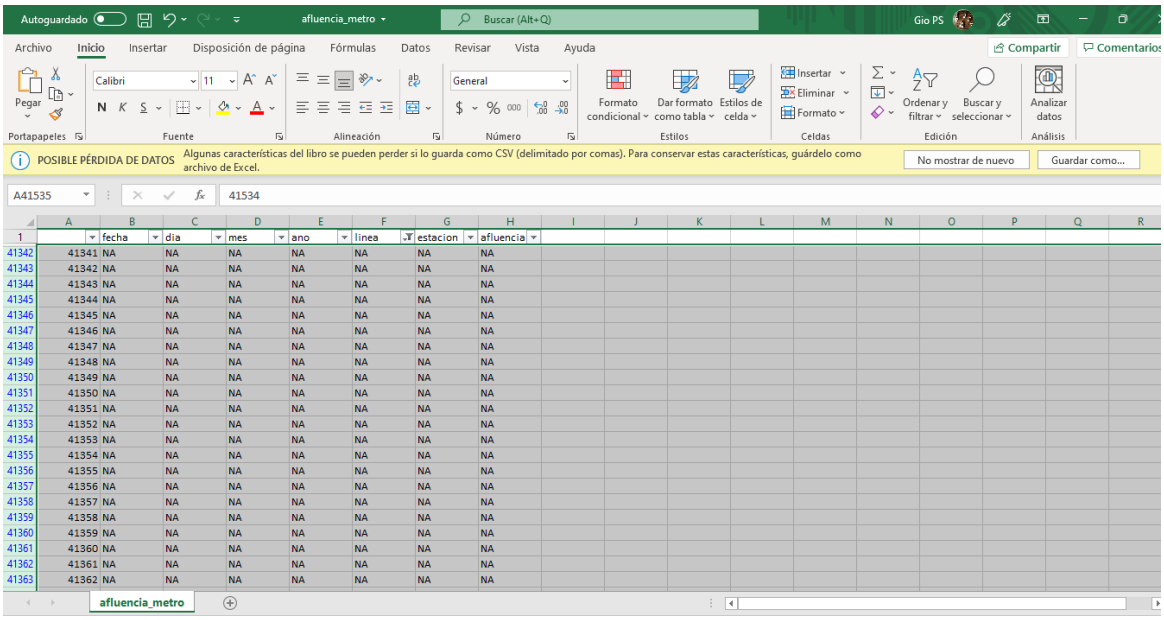
- Índice: Esta columna funge como un índice de conteo de cada uno de los registros. Es un número único y creciente. Solo puede tomar valores numéricos enteros a partir de uno.
- Fecha: Este atributo está integrado por día, mes y año descrito por la siguiente notación: dd/mm/aaaa. Cada uno de los datos atómicos que la componen pueden asumir los siguientes valores: día, del 1 al 30 o 31; mes, del 1 al 12 y en el año, solamente 2021.
- Día: Esta columna se refiere a los dígitos del día del mes. Este atributo puede tomar valores del 1 al 30 o 31 dependiendo el mes al que pertenece el día de registro. Este atributo nos permite conocer la afluencia diaria de pasajeros.
- Mes: Se refiere al mes del año. Es una columna de tipo caracter y que toma el nombre complete de cada mes iniciando con mayúscula. Hace referencia a la afluencia de pasajeros que abordan las estaciones por bloque cada mes.

- Año: Esta columna se refiere al año que, para esta base de datos, asume un único valor: 2021.
- Línea: Esta columna se refiere a la cantidad de líneas que tiene este medio de transporte. Los nombres reales de las líneas van del 1-9, A, B y 12. Sin embargo, para los fines de este análisis, se reemplazará el nombre de las líneas A y B, por 10 y 11. El resto de los nombres permanecen con su valor numérico.
- Estación: Cada una de estas tiene un nombre estipulado para poder diferenciarlas y así podernos mover a través de la ciudad. Originalmente, esta variable asume el nombre asignado al momento de su creación.
- Afluencia: Esta es la columna que indica la cantidad de personas que ingresaron diariamente en cada estación del metro. Los valores que puede tomar esta columna son de tipo entero no negativos.

Pre-procesamiento de los datos

Antes de comenzar con el análisis de los datos fue necesario hacer un proceso de limpieza, filtrado y reestructuración de los datos disponibles.

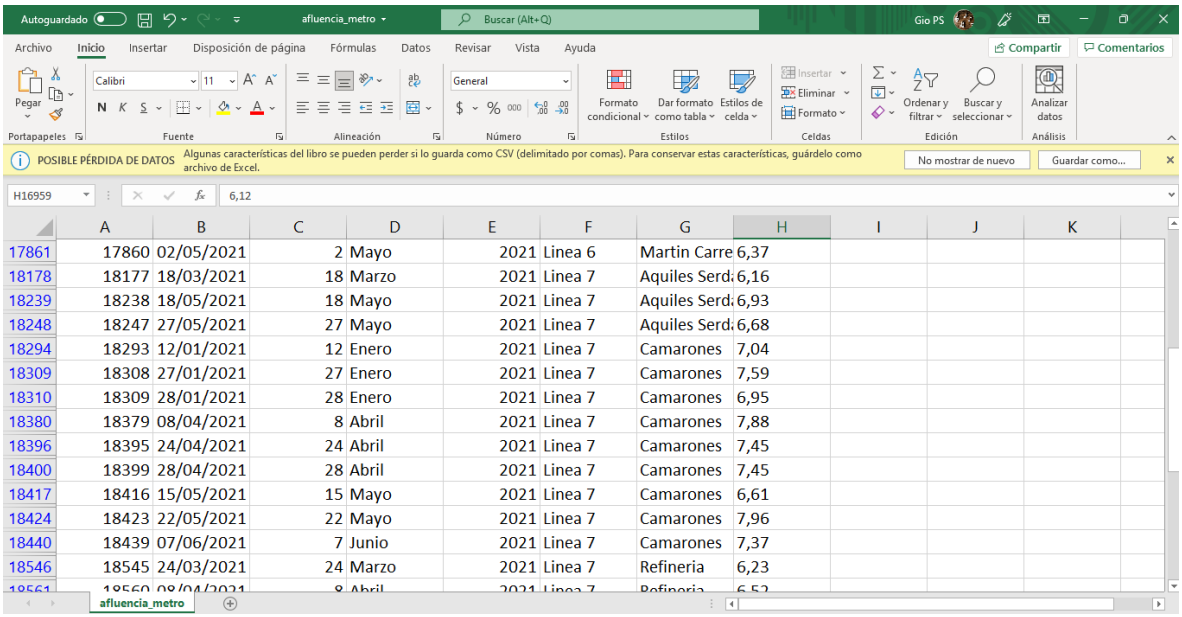
En la columna de “Afluencia” había varios registros en blanco, los cuales si se dejaban de esa manera podrían haber ocasionado problemas para determinar ciertos parámetros como la media o la varianza. Por un lado, las funciones de R los hubieran tomado como valores NA lo cual hubiera generado que todos nuestros cálculos se fueran a nulo. En el caso de la media, los datos faltantes podrían haber decrementado en gran medida su valor. Para ello, a todos los valores de la “Afluencia” que no estaban registrados se les asignó el valor de cero.



		fecha	dia	mes	ano	linea	estacion	afluencia
41342	41341	NA	NA	NA	NA	NA	NA	NA
41343	41342	NA	NA	NA	NA	NA	NA	NA
41344	41343	NA	NA	NA	NA	NA	NA	NA
41345	41344	NA	NA	NA	NA	NA	NA	NA
41346	41345	NA	NA	NA	NA	NA	NA	NA
41347	41346	NA	NA	NA	NA	NA	NA	NA
41348	41347	NA	NA	NA	NA	NA	NA	NA
41349	41348	NA	NA	NA	NA	NA	NA	NA
41350	41349	NA	NA	NA	NA	NA	NA	NA
41351	41350	NA	NA	NA	NA	NA	NA	NA
41352	41351	NA	NA	NA	NA	NA	NA	NA
41353	41352	NA	NA	NA	NA	NA	NA	NA
41354	41353	NA	NA	NA	NA	NA	NA	NA
41355	41354	NA	NA	NA	NA	NA	NA	NA
41356	41355	NA	NA	NA	NA	NA	NA	NA
41357	41356	NA	NA	NA	NA	NA	NA	NA
41358	41357	NA	NA	NA	NA	NA	NA	NA
41359	41358	NA	NA	NA	NA	NA	NA	NA
41360	41359	NA	NA	NA	NA	NA	NA	NA
41361	41360	NA	NA	NA	NA	NA	NA	NA
41362	41361	NA	NA	NA	NA	NA	NA	NA
41363	41362	NA	NA	NA	NA	NA	NA	NA

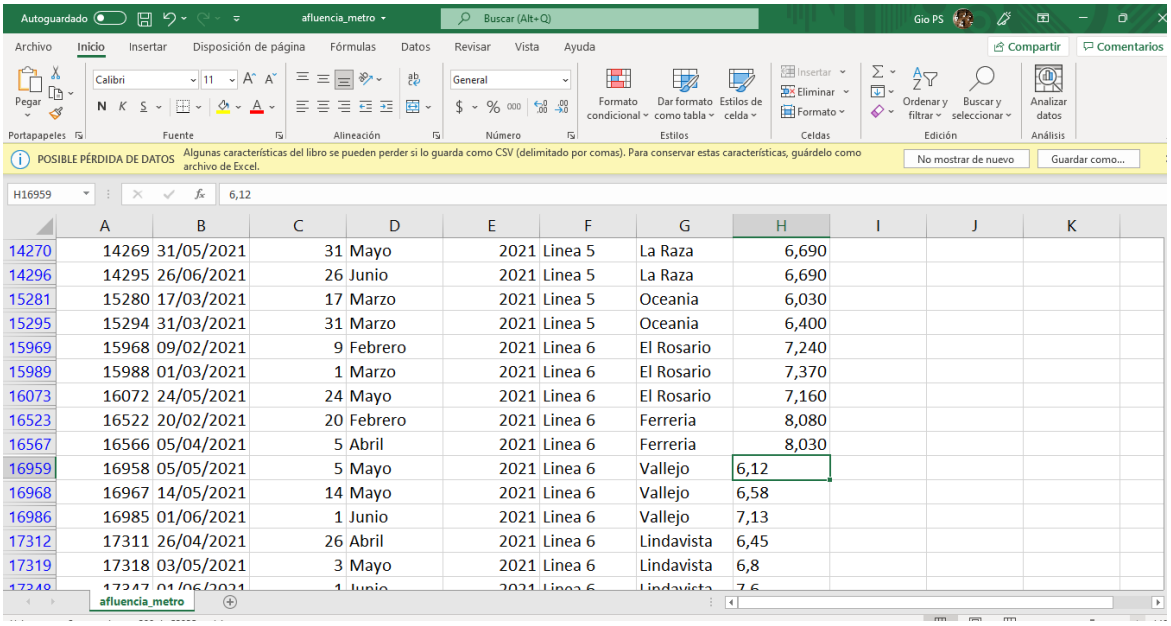
Imagen 1. Valores no registrados (NA) de la base de datos afluencia_metro original.

En la misma columna de “Afluencia”, había varios registros incompletos que debieron corregirse. El registro de muchos números en tal columna estaba expresado como: “2,14”, “10,49” o “7,11”. Contextualizando, el uso de la coma en el documento se daba como carácter de separación cada tres dígitos, los números como los mostrados arriba carecen de ceros a la derecha. En estos casos, se completó con uno o dos ceros para cumplir tres dígitos después de la coma. Esto se hizo a fin de no afectar los cálculos de las medidas de tendencia central, de posición y de dispersión.



	A	B	C	D	E	F	G	H	I	J	K
17861	17860	02/05/2021	2	Mayo	2021	Linea 6	Martin Carre	6,37			
18178	18177	18/03/2021	18	Marzo	2021	Linea 7	Aquiles Serd	6,16			
18239	18238	18/05/2021	18	Mayo	2021	Linea 7	Aquiles Serd	6,93			
18248	18247	27/05/2021	27	Mayo	2021	Linea 7	Aquiles Serd	6,68			
18294	18293	12/01/2021	12	Enero	2021	Linea 7	Camarones	7,04			
18309	18308	27/01/2021	27	Enero	2021	Linea 7	Camarones	7,59			
18310	18309	28/01/2021	28	Enero	2021	Linea 7	Camarones	6,95			
18380	18379	08/04/2021	8	Abril	2021	Linea 7	Camarones	7,88			
18396	18395	24/04/2021	24	Abril	2021	Linea 7	Camarones	7,45			
18400	18399	28/04/2021	28	Abril	2021	Linea 7	Camarones	7,45			
18417	18416	15/05/2021	15	Mayo	2021	Linea 7	Camarones	6,61			
18424	18423	22/05/2021	22	Mayo	2021	Linea 7	Camarones	7,96			
18440	18439	07/06/2021	7	Junio	2021	Linea 7	Camarones	7,37			
18546	18545	24/03/2021	24	Marzo	2021	Linea 7	Refineria	6,23			

Imagen 2. Corrección del registro de los números como el 8,900 en algunos casos solo están registrados como 8,9.



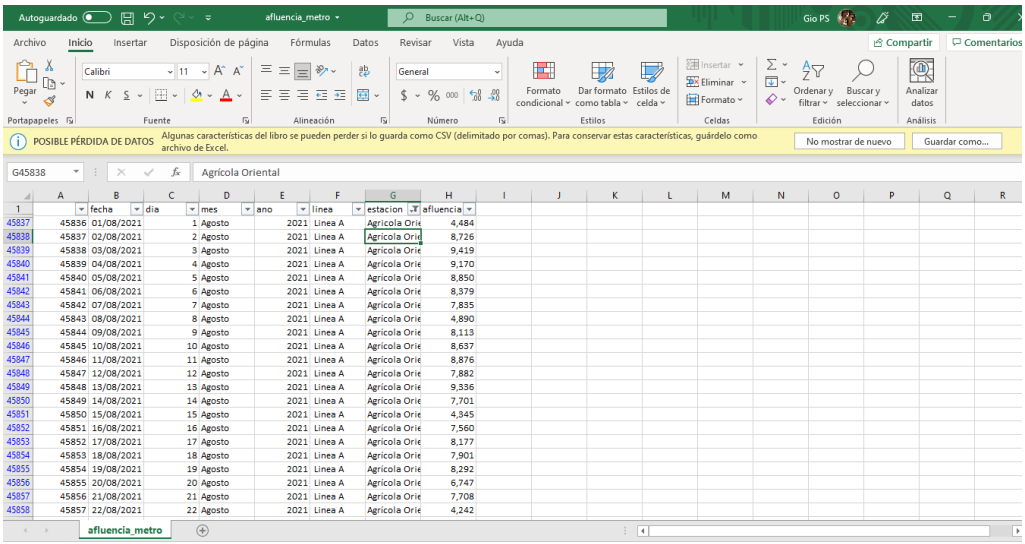
	A	B	C	D	E	F	G	H	I	J	K
14270	14269	31/05/2021	31	Mayo	2021	Linea 5	La Raza	6,690			
14296	14295	26/06/2021	26	Junio	2021	Linea 5	La Raza	6,690			
15281	15280	17/03/2021	17	Marzo	2021	Linea 5	Oceania	6,030			
15295	15294	31/03/2021	31	Marzo	2021	Linea 5	Oceania	6,400			
15969	15968	09/02/2021	9	Febrero	2021	Linea 6	El Rosario	7,240			
15989	15988	01/03/2021	1	Marzo	2021	Linea 6	El Rosario	7,370			
16073	16072	24/05/2021	24	Mayo	2021	Linea 6	El Rosario	7,160			
16523	16522	20/02/2021	20	Febrero	2021	Linea 6	Ferreria	8,080			
16567	16566	05/04/2021	5	Abril	2021	Linea 6	Ferreria	8,030			
16959	16958	05/05/2021	5	Mayo	2021	Linea 6	Vallejo	6,12			
16968	16967	14/05/2021	14	Mayo	2021	Linea 6	Vallejo	6,58			
16986	16985	01/06/2021	1	Junio	2021	Linea 6	Vallejo	7,13			
17312	17311	26/04/2021	26	Abril	2021	Linea 6	Lindavista	6,45			
17319	17318	03/05/2021	3	Mayo	2021	Linea 6	Lindavista	6,8			

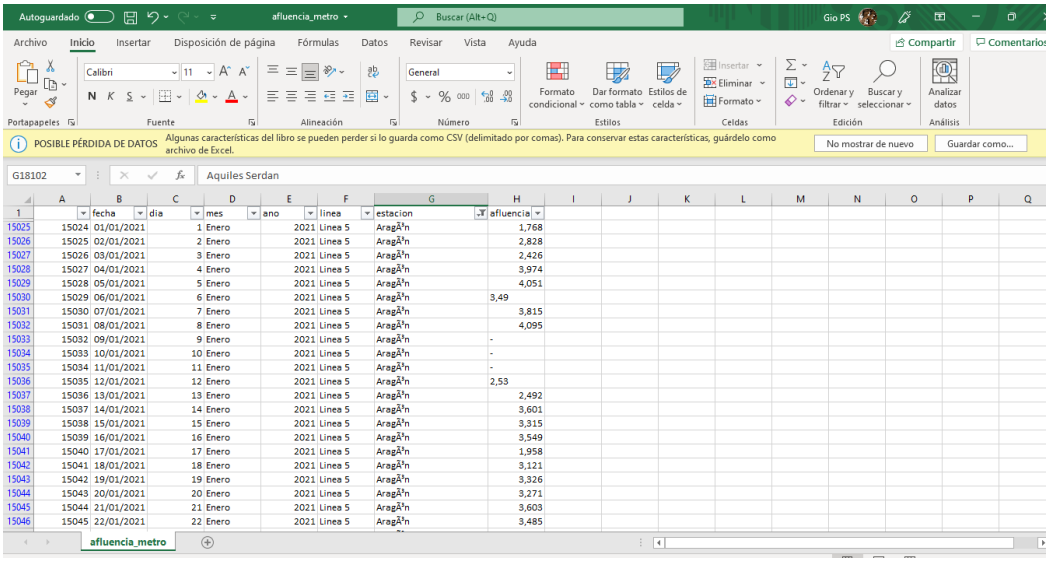
Imagen 3. La “,” está siendo usada como un símbolo para separar las cifras en grupos de tres, no como un punto decimal. Esto se puede saber ya que en otras celdas la coma es usada de forma correcta en la separación de números grandes cada tres dígitos y ya que no puede haber un número de personas fraccionario como 8.9 personas, la afluencia es una columna de números enteros.

En línea con el párrafo anterior, había datos cuestionables en la columna “Afluencia”. Un dato cuestionable es que los ingresos en ciertas estaciones estaban registrados con el valor de 0, 1, 2,4, 5 y en general, valores de menos de 100, cuando, de forma empírica, sabemos que el número de personas que ingresan a una estación del metro en CDMX es muy poco probable que sea menor a 1000. Por ende, se asumió que los datos menores a 100 eran registros incorrectos. Al no contar con la información del archivo original o poder inferirla de forma contextual (como en el caso de registros separados por coma), se optó por eliminar estos datos, en tanto, no proporcionan información clara y fidedigna, y alterarían los parámetros estadísticos.

Para finalizar con la limpieza de los datos de la columna “Afluencia” se eliminaron los registros iguales a cero. Como vimos arriba, los registros iguales a cero eran registros nulos, no había información de los ingresos en una estación en tales días. Esto puede deberse debido a la pérdida del archivo, una omisión en el registro de ingresos, etcétera. Para que estos datos faltantes, no alteraran el promedio de la afluencia general, la desviación estándar o el cálculo de cuartiles, se optó por eliminarlos. No proporcionaban más información a la explicada en estas líneas.

En la columna “Estación”, los nombres de estas se repetían con ciertas diferencias entre los caracteres. Por ello, se tuvo que homogeneizar el registro de todas las estaciones. A fin de evitar problemas con el procesamiento de datos por parte de RStudio, a todas las estaciones cuyo nombre se acentúa se optó por cambiar tal letra por la misma letra sin acentuar.





Imagenes 4. Corrección ortográfica y gramatical del nombre de las estaciones del metro, ya que la base de datos despliega caracteres especiales cuando encuentra letras acentuadas y ñ's.

Para usar las estaciones como variable independiente "x" esta debía ser de tipo numérico, lo cual de origen no es. Por tanto, a lado derecho de la columna "Estación" se agregó una nueva columna llamada "ID_estación". En esta columna se muestra un número único que se le asignó a cada estación como forma de identificación. Los valores de los id van del 1 al 163, que obedecen al número de estaciones en el STCM. Para la asignación de este id, se enlistaron las estaciones y se ordenaron de forma alfabética. Así, la asignación de identificadores quedó de la siguiente manera:

id_estacion	nom_estacion
1	20 de Noviembre
2	Acatitla
3	Aculco
4	Agricola Oriental
5	Allende
6	Apatlaco
7	Aquiles Serdan
8	Aragon
9	Atlalilco
10	Auditorio
11	Autobuses del Norte
12	Azcapotzalco
13	Balbuena
14	Balderas
15	Barranca del Muerto
16	Bellas Artes
17	Blvd Pto Aereo
18	Bondojito
19	Bosque de Aragon
20	Buenavista
21	Calle 11
22	Camarones
23	Canal de San Juan
24	Canal del Norte
25	Candelaria
26	Centro Medico
27	Cerro de la Estrella
28	Chabacano
29	Chapultepec
30	Chilpancingo
31	Ciudad Azteca
32	Ciudad Deportiva
33	Colegio Militar

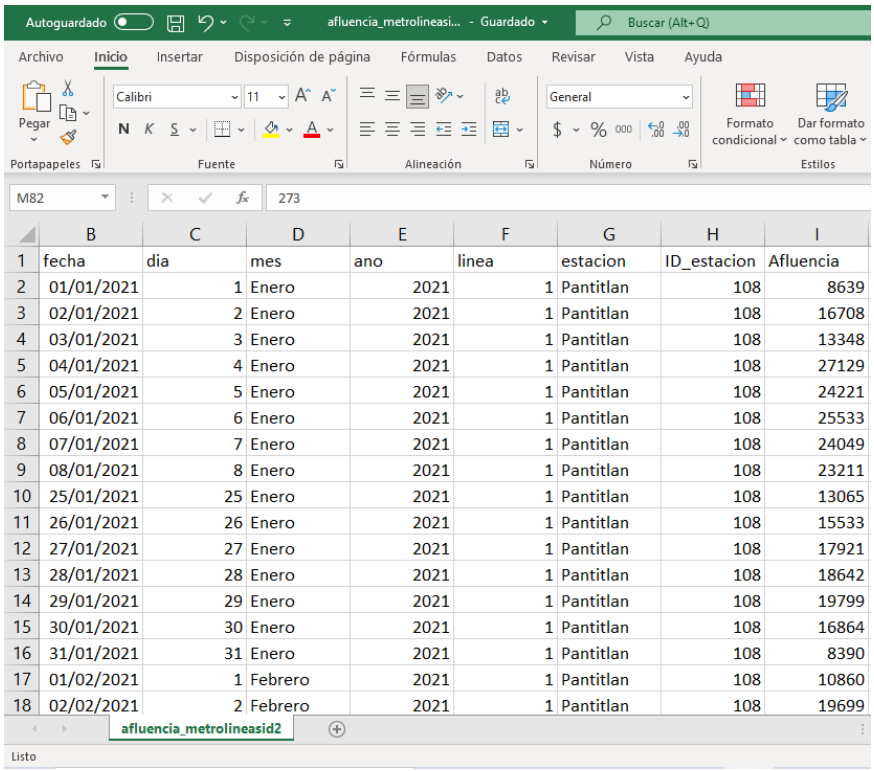
id_estacion	nom_estacion
85	Los Reyes
86	Martin Carrera
87	Merced
88	Mexicaltzingo
89	Miguel A. de Q.
90	Misterios
91	Mixcoac
92	Mixiuhca
93	Moctezuma
94	Morelos
95	Muzquiz
96	Nativitas
97	Nezahualcóyotl
98	Niños Heroes
99	Nopalera
100	Normal
101	Norte 45
102	Obrera
103	Observatorio
104	Oceania
105	Olimpica
106	Olivos
107	Panteones
108	Pantitlan
109	Parque de los Venados
110	Patriotismo
111	Peñon Viejo
112	Periferico Oriente
113	Pino Suarez
114	Plaza Aragon
115	Polanco
116	Politecnico
117	Popotla

34	Constitucion de 1917
35	Constituyentes
36	Consulado
37	Copilco
38	Coyoacan
39	Coyuya
40	Cuatro Caminos
41	Cuauhtemoc
42	Cuitlahuac
43	Culhuacan
44	Division del Norte
45	Doctores
46	Dvo 18 de marzo
47	Dvo Oceania
48	Ecatepec
49	Eduardo Molina
50	Eje Central
51	El Rosario
52	Ermita
53	Escuadron 201
54	Etiopia
55	Eugenia
56	Ferreria
57	Fray Servando
58	Garibaldi
59	General Anaya
60	Gomez Farias
61	Guelatao
62	Guerrero
63	Hangares
64	Hidalgo
65	Hospital General
66	Impulsora
67	Indios Verdes

118	Portales
119	Potrero
120	Puebla
121	Refineria
122	Revolucion
123	Ricardo Flores Magon
124	Rio de los Remedios
125	Romero Rubio
126	Salto del Agua
127	San Andres Tomatlan
128	San Antonio
129	San Antonio Abad
130	San Cosme
131	San Joaquin
132	San Juan de Letran
133	San Lazaro
134	San Pedro de los Pinos
135	Santa Anita
136	Santa Marta
137	Sevilla
138	Tacuba
139	Tacubaya
140	Talisman
141	Tasqueña
142	Tepalcates
143	Tepito
144	Terminal Aerea
145	Tezonco
146	Tezozomoc
147	Tlahuac
148	Tlaltenco
149	Tlatelolco
150	UAMI
151	Universidad

68	Inst. del Petroleo
69	Insurgentes
70	Insurgentes Sur
71	Isabel la Catolica
72	Iztacalco
73	Iztapalapa
74	Jamaica
75	Juanacatlan
76	Juarez
77	La Paz
78	La Raza
79	La Viga
80	La Villa-Basilica
81	Lagunilla
82	Lazaro Cardenas
83	Lindavista
84	Lomas Estrella

152	Valle Gomez
153	Vallejo
154	Velodromo
155	Viaducto
156	Villa de Aragon
157	Villa de Cortes
158	Viveros
159	Xola
160	Zapata
161	Zapotitlan
162	Zaragoza
163	Zocalo



	B	C	D	E	F	G	H	I
1	fecha	dia	mes	ano	linea	estacion	ID_estacion	Afluencia
2	01/01/2021	1	Enero	2021	1	Pantitlan	108	8639
3	02/01/2021	2	Enero	2021	1	Pantitlan	108	16708
4	03/01/2021	3	Enero	2021	1	Pantitlan	108	13348
5	04/01/2021	4	Enero	2021	1	Pantitlan	108	27129
6	05/01/2021	5	Enero	2021	1	Pantitlan	108	24221
7	06/01/2021	6	Enero	2021	1	Pantitlan	108	25533
8	07/01/2021	7	Enero	2021	1	Pantitlan	108	24049
9	08/01/2021	8	Enero	2021	1	Pantitlan	108	23211
10	25/01/2021	25	Enero	2021	1	Pantitlan	108	13065
11	26/01/2021	26	Enero	2021	1	Pantitlan	108	15533
12	27/01/2021	27	Enero	2021	1	Pantitlan	108	17921
13	28/01/2021	28	Enero	2021	1	Pantitlan	108	18642
14	29/01/2021	29	Enero	2021	1	Pantitlan	108	19799
15	30/01/2021	30	Enero	2021	1	Pantitlan	108	16864
16	31/01/2021	31	Enero	2021	1	Pantitlan	108	8390
17	01/02/2021	1	Febrero	2021	1	Pantitlan	108	10860
18	02/02/2021	2	Febrero	2021	1	Pantitlan	108	19699

Imagen 5. Base de datos con la nueva columna “ID_estacion” y con las adecuaciones previas.

Diagrama de flujo

Desarrollo del análisis estadístico

Medidas de tendencia central, de dispersión y de posición

Cálculo de las medidas de tendencia central, de dispersión y de posición.

Las medidas de tendencia central y de dispersión de la afluencia del metro son:	
Media:	11421.33
Mediana:	8732
Moda:	6683, 6748, 6930, 8755

Interpretación

En promedio, 11,421 personas ingresan diariamente al metro de la ciudad de México por cada una de las 163 estaciones de este sistema de transporte en lo que llevamos del 2021. Cabe resaltar que es muy probable que en años anteriores y en los que vienen, el número promedio de personas sea mayor, ya que 2020 y 2021 son años atípicos, por el contexto de la pandemia de COVID-19. En este entendido, muchas personas trabajan desde su casa o asisten a clases de manera virtual. Por ende, la afluencia en el transporte no es la misma que la cotidiana. Aunque se han

ido retomando ciertas actividades, al día de hoy, no podemos decir que el 100% de las personas están acudiendo a la oficina o la escuela y por ende, el promedio, muy probablemente, aumentará cuando se retomen las actividades al 100% de su afluencia.

El dato que divide exactamente a la mitad la distribución de las afluencias de personas es de 8732. Finalmente, el dato del número de personas que ingresan a las estaciones del metro que más se repite es: 6683, 6748, 6930 y 8755. Lo cual implica que tenemos una distribución multimodal.

Medidas de dispersión	
Varianza:	99231657
Desviación estándar:	9961.509
Coeficiente de variación:	0.8721843

Interpretación

Al revisar la varianza y la desviación estándar podría parecer que los datos están muy alejados de la media, pero si vemos la desviación estándar que está en las mismas unidades que la media, los datos no se alejan mucho más allá del valor promedio que es de 11,421.33. Los datos grandes no deben sorprendernos, ya que los registros de las afluencias oscilan entre las 1000 personas hasta un máximo de 88,734 de ingresos únicos en una estación.

Cuartiles:				
0%	25%	50%	75%	100%
11	4975	8732	14208	88734

Deciles:									
0%	10%	20%	30%	40%	50%	60%	70%	80%	90%
11.0	2900.0	4268.0	5745.0	7156.0	8732.0	10598.0	12829.0	15966.0	23175.5

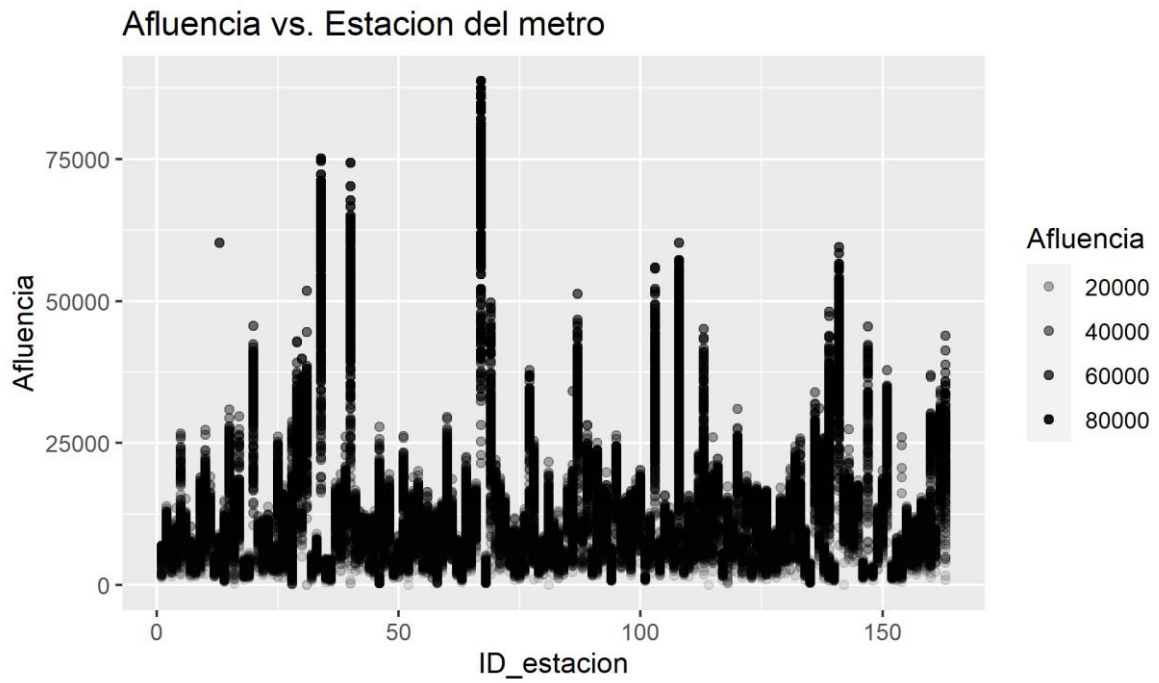
100%
88734.0

Interpretación

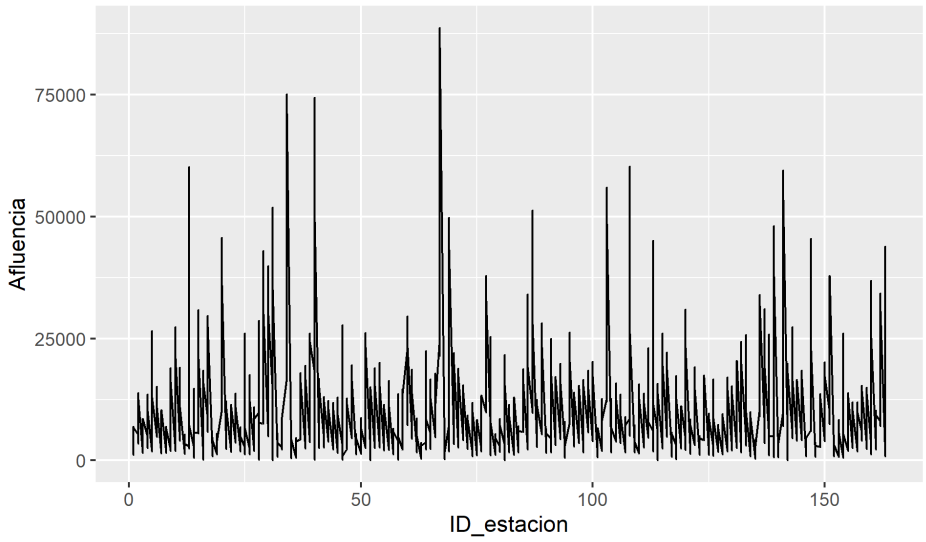
En línea con lo expresado anteriormente, los datos se encuentran contenidos en un intervalo muy grande de datos. El primer dato es el ingreso de 11 personas, lo cual obedece a los datos cuestionables de la base de datos, donde encontramos errores en el registro de los números e incluso valores nulos, pero como no tenemos los archivos fuente de primera mano o un indicio que nos ayude a corregir tal registro, se dejaron así. En línea con el resultado de la mediana, el cuartil del 50% es el mismo valor que el de la mediana. Finalmente, el dato del 100% nos permite ver que el rango de los datos es muy grande, aunque tampoco es la norma, ya que vemos el brinco del rango 75% al del 100%. Por ende, el número de registros con valores superiores a 14 mil no son la mayoría

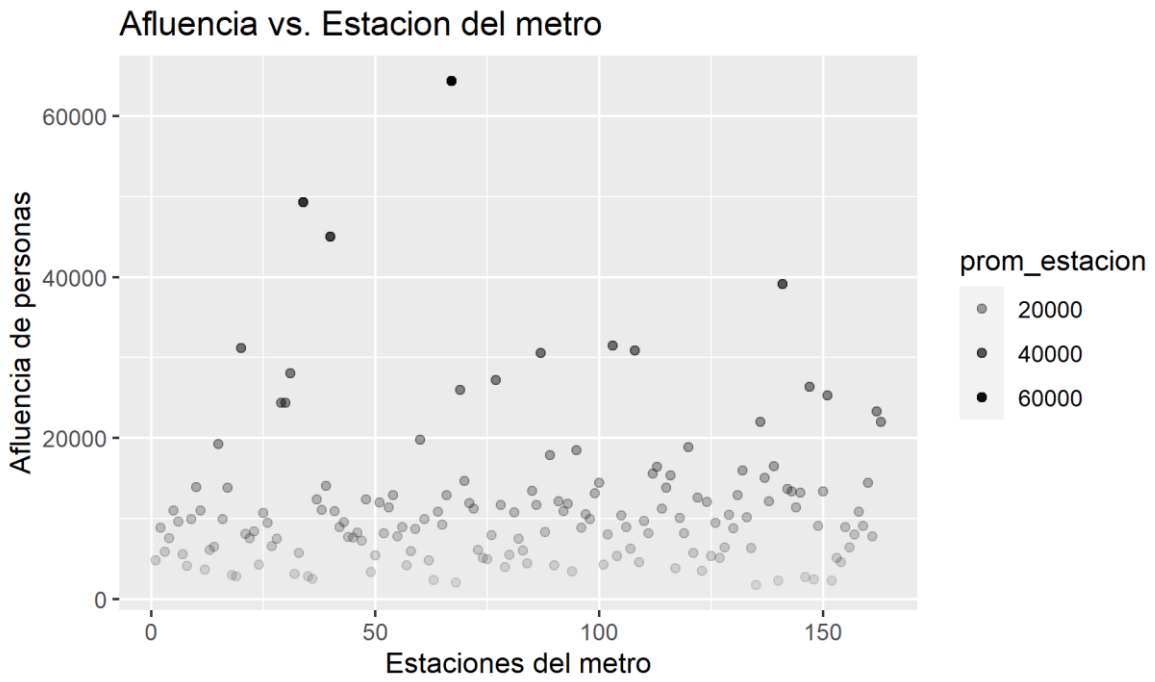
Este último punto, puede visualizarse con mayor claridad al ver los datos en deciles. Los valores superiores a 23,180 ingresos se encuentran a penas en el último 10% de los valores de la variable aleatoria. De forma análoga con lo anterior, los ingresos de menos de 2900 personas son apenas el 10% del dataset. Con lo cual aseguramos que las medidas de tendencia central y de dispersión sean más o menos regulares entre 4,000 y 16,000 personas.

Gráfico de dispersión

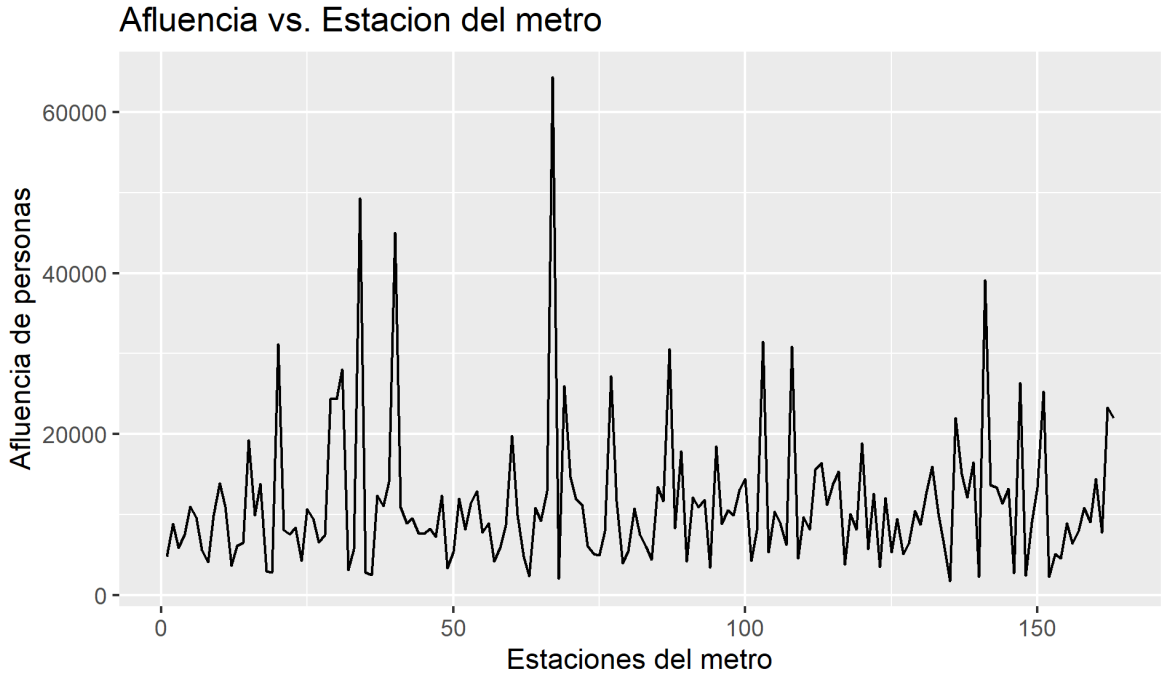


Gráficos 1. El gráfico de dispersión “Estación de metro vs. Afluencia” nos permite ver la manera en que se distribuye la cantidad de personas que ingresan por cada una de las 163 estaciones del metro. En el eje x se ha usado el identificador numérico de cada estación a fin de que fuera más fácil su interpretación. Del lado derecho, se incluye una tabla de afluencia donde los puntos más difuminados son ingresos más pequeños, en tanto, los más opacos representan una cantidad mayor de personas que ingresan. Esta intensidad de la opacidad permite obtener información más puntual de qué número representa cada punto del gráfico o sobre qué valor se encuentra.





Gráficos 2. Los gráficos de dispersión “Afluencia vs. Estación del metro” mostrados en este apartado son una versión más simplificada de los mostrados arriba, esta vez tomando a la Afluencia de personas como un promedio. Para esta versión, se obtuvo el promedio de afluencia de las personas en cada estación. Con ello, se busca facilitar la visualización y entendimiento de los datos mostrados arriba, ya que por cada estación solo se muestra el punto promedio de ingreso en lugar de un punto por cada día del año.



Regresión y correlación lineal

Coeficientes:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.721e+01	3.187e-01	242.25	<2e-16 ***
datosarchivo$Afluencia	3.003e-04	2.103e-05	14.28	<2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error:	46.1 on 48424 degrees of freedom
Multiple R-squared:	0.004193, Adjusted R-squared: 0.004173
F-statistic:	203.9 on 1 and 48424 DF, p-value: < 2.2e-16

Interpretación

La ecuación que describe la relación entre las variables:

x=Estación del metro

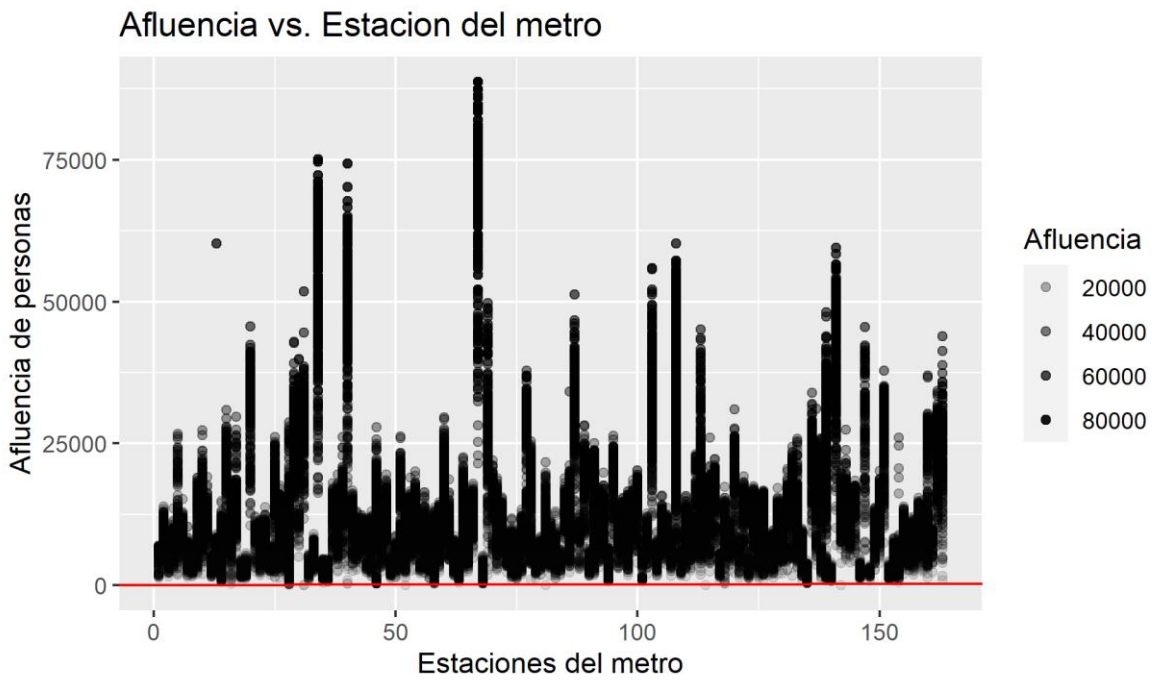
y= Número de pasajeros que ingresa en esa estación (afluencia)

es:

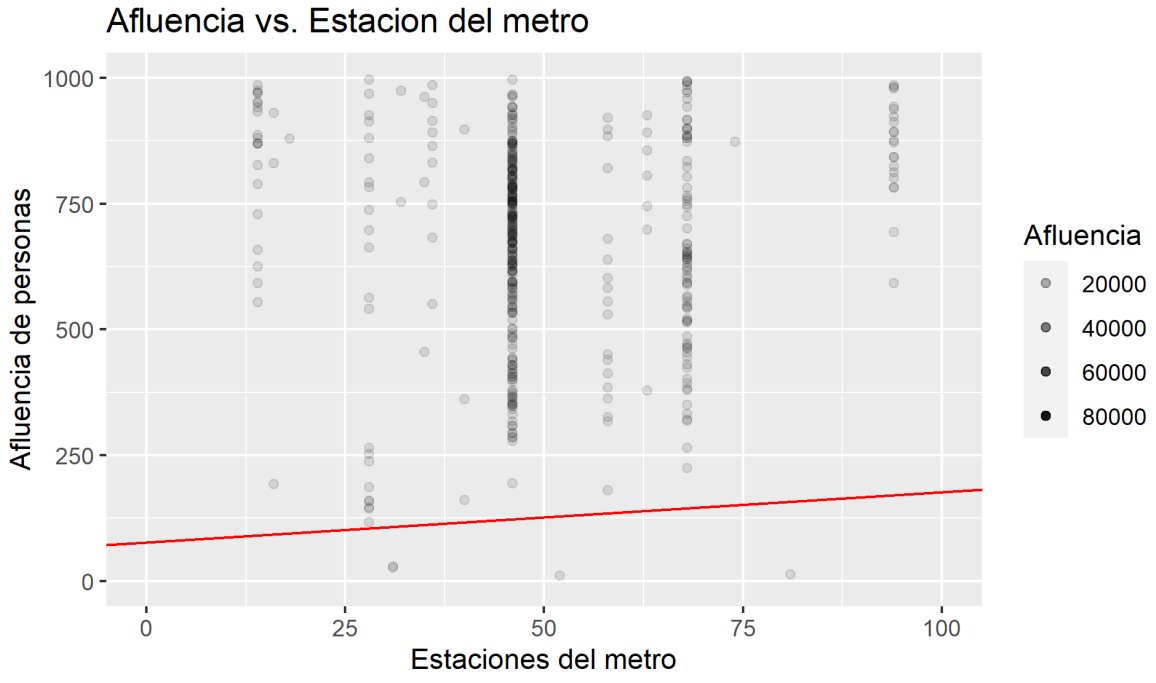
$$y=0.0003003041x + 77.21026$$

El modelo encontrado muestra una relación positiva entre ambas variables, es decir, que la estación realmente influye en el número de pasajeros que ingresan al metro. Más allá de ello, a partir del análisis del coeficiente p, generado entre los “Coeficientes”, podemos ver que esta toma un valor 0.00000000000000022. Dicho valor es menor a la hipótesis nula del test de normalidad según la cual el valor de $p<0.5$ para que exista relación entre las variables. Por ende, los datos siguen una distribución normal. Para verificar esto último, posteriormente se hará la prueba de normalidad de los datos. En conclusión, de lo anterior, podemos decir que en función de la estación del metro que se esté analizando esta define en gran medida la cantidad de pasajeros que ingresa al metro.

Gráfica de línea de regresión lineal



Gráficos 3. La línea de regresión lineal podemos ver que se localiza en la parte inferior de los registros de las afluencias. En la versión más acotada de la gráfica (grafica de abajo), se ve que esta línea de regresión cruza con el eje y en el punto 77.21026.



Prueba de normalidad

Para analizar la normalidad de los datos de la afluencia, se optó por utilizar las pruebas de Kolmogorov, ya que tenemos más de 5000 registros. Para ello, se utilizó la función shapiro.test() la cual nos arrojó los siguientes resultados:

Lilliefors (Kolmogorov-Smirnov) normality test

data: datosarchivo$Afluencia

D = 0.14475, p-value < 2.2e-16

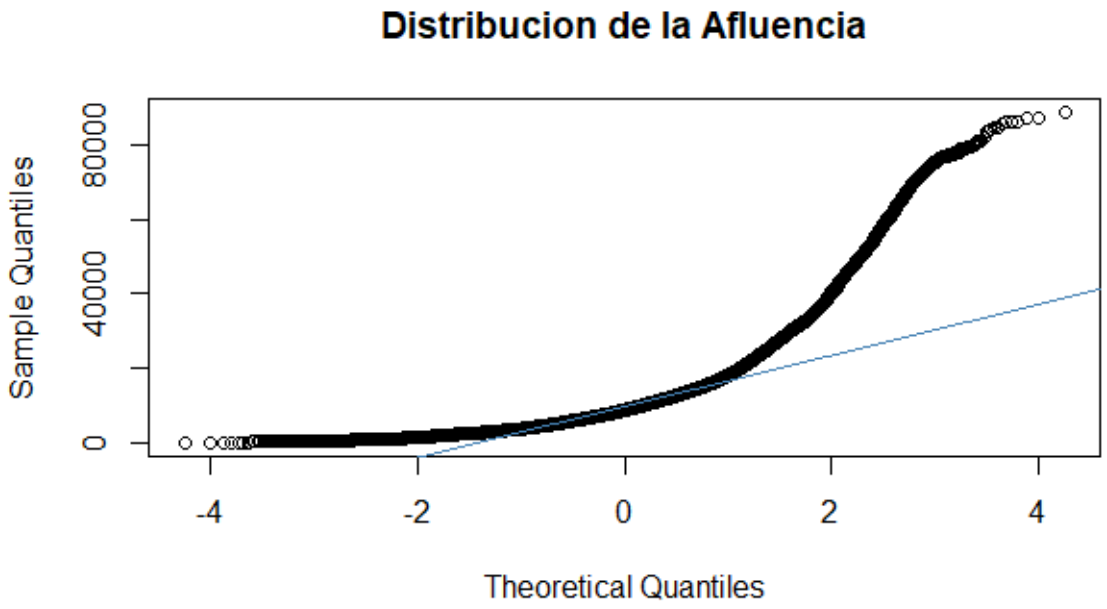


Gráfico 4: Distribución de la afluencia de personas por cada estación de metro.

Interpretación

Tal como se había comprobado al obtener la ecuación de relación entre variables, el p-value es mucho menor a 0.5, lo cual demuestra que los datos de la afluencia de personas en el metro siguen una distribución normal.

Covarianza y Coeficiente de correlación

El valor de la covarianza entre los datos es de: 29799.67

cov(estaciones,afluencia)

29799.67

Matriz de covarianza		
	estaciones	Afluencia
estaciones	2134.168	29799.67
afluencia	29799.67	99231656.69

El coeficiente de correlación de las variables Estaciones del metro y Afluencia es: 0.06475481.

Pearson's product-moment correlation

data: datosarchivo$ID_estacion and datosarchivo$Afluencia

t = 14.28, df = 48424, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.05588047 0.07361893

sample estimates:

cor 0.06475481

Matriz de correlaciones		
	estaciones	Afluencia
estaciones	1.00000000	0.06475481
afluencia	0.06475481	1.00000000

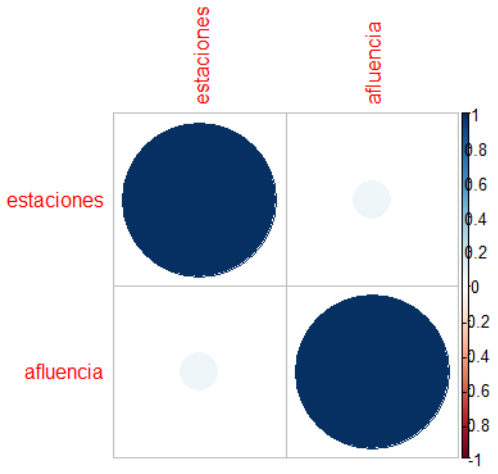


Gráfico 5. Matriz de correlación de forma gráfica.

Interpretación

El coeficiente de correlación de las variables Estaciones del metro y Afluencia, de acuerdo con el método de Pearson es de 0.06475481. Esto se interpreta como una correlación prácticamente nula. Por ende, un investigador o científico de datos debería considerar relacionar a la afluencia de los datos con otra variable tal como el tiempo (mes, año), la línea del metro o cualquier otra disponible en la base de datos, a fin de encontrar una relación más significativa entre variables.

Distribución de probabilidad

La variable aleatoria de este experimento aleatorio es de tipo discreto, en tanto que el número de personas que pueden confluir a una estación es un número finito numerable y entero. Por otro lado, tenemos ya definido un número promedio de personas que pueden llegar a una estación del metro (un espacio) diariamente. Por ende, el tipo de distribución que mejor se adapta o puede describir a nuestros datos es una distribución de Poisson.

De esta forma, la variable X es el número de la estación por la que ingresan los usuarios. La media de este valor se toma como lambda. La media de las estaciones es de 80.64013. Finalmente, la función de densidad de X determinara la

probabilidad de que las personas ingresen al metro por una estación específica o por una serie de estaciones en un día.

Por ejemplo, podemos calcular la probabilidad de que las personas ingresen por la x=45.

A partir de la aplicación de la función para obtener la distribución de probabilidad de Poisson para x=45, obtenemos: 0.000004959402.

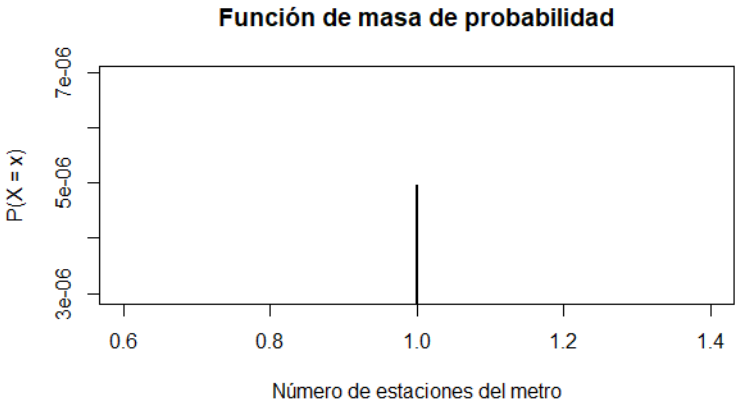


Gráfico 6. La gráfica muestra el valor de la probabilidad de x=45, en el eje x podemos ver que solo estamos calculando la probabilidad de un registro. Por ello, el valor en el eje x es 1 y el de y es la probabilidad.

Para un nuevo valor como x=97, la probabilidad de ocurrencia es = 0.008526337.

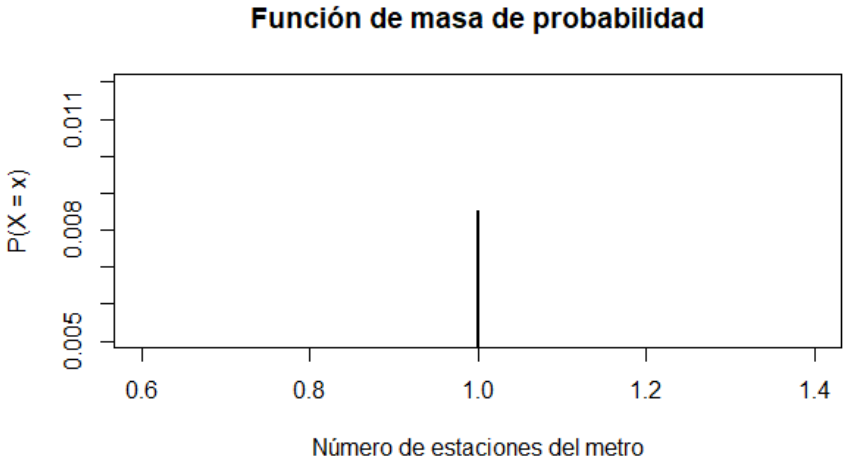


Gráfico 7. La gráfica muestra el valor de la probabilidad de x=97. Al igual que en el caso anterior, solo estamos calculando la probabilidad de un registro. Por ello, el valor en el eje x es 1 y el de y es la probabilidad.

Como se puede apreciar al ingresar solo un posible valor de x, las probabilidades asumen un valor muy bajo o prácticamente nulo, esto se debe a que el número de registros es muy grande, 48427 en total (tras la depuración hecha en el preprocesamiento).

Para encontrar probabilidades más grandes, habría que determinarse la probabilidad para un intervalo de datos o buscar una función de distribución acumulada.

Por ejemplo, para calcular la probabilidad de un rango de valores como de 51:80 obtendremos una serie de probabilidades que asumen los siguientes valores:

[1] "0.0001051728" "0.0001630990" "0.0002481571" "0.0003705819" "0.0005433413"

[6] "0.0007824127" "0.0011069099" "0.0015389889" "0.0021034621" "0.0028270577"

[11] "0.0037372837" "0.0048608879" "0.0062219466" "0.0078396655" "0.0097260255"

[16] "0.0118834542" "0.0143027359" "0.0169613897" "0.0198227347" "0.0228358276"

[21] "0.0259363962" "0.0290488111" "0.0320890403" "0.0349684382" "0.0375981260"

[26] "0.0398936555" "0.0417796054" "0.0431937547" "0.0440905070" "0.0444433034"

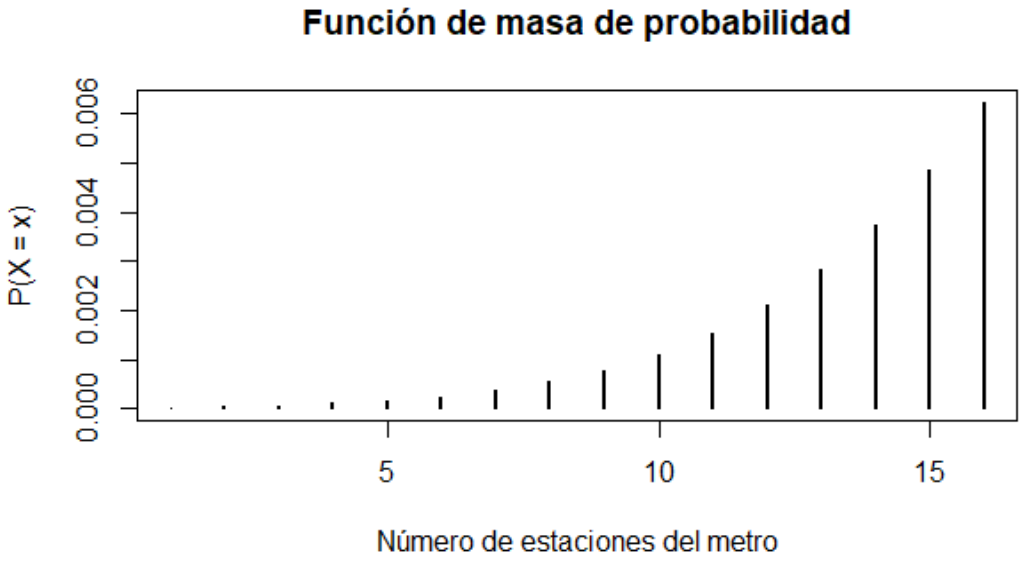


Gráfico 8. La gráfica muestra el valor de la probabilidad de x=51 hasta x=80. Al igual que en el caso anterior, solo estamos calculando la probabilidad de un registro. Por ello, el valor en el eje x va desde 1 hasta 16 que es el número de probabilidades que estamos calculando.

Conclusión