



3CM1-Programación para Ciencia de Datos
Docente: Galindo Durán Cristal Karina

PROYECTO 2 ANÁLISIS DE AFLUENCIA EN EL METRO

De Luna Ocampo Yanina
Perea Samaniego Jesus Giovanni

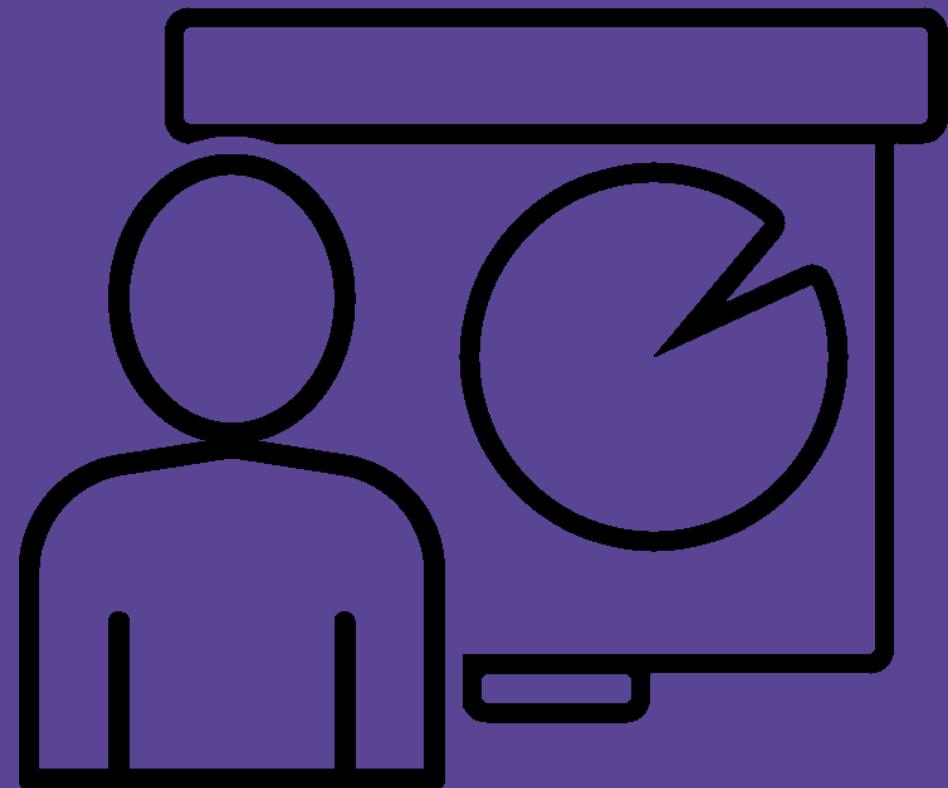


CONTENIDO

- DESCRIPCIÓN DE LOS DATOS
- PRE-PROCESAMIENTO DE LOS DATOS
- DEFINICIÓN DE LAS VARIABLES
- DIAGRAMA DE FLUJO
- DESARROLLO DEL ANÀLISIS ESTADÍSTICO
- GRÀFICOS DE DISPERSIÓN (COMPLETO)
- CONCLUSIONES
- REFERENCIAS



DESCRIPCIÓN DE LOS DATOS.



01

Nuestros datos corresponden al número de personas que ingresan a cada estación del metro diariamente del 1 de enero del 2021 al 30 de septiembre de 2021.

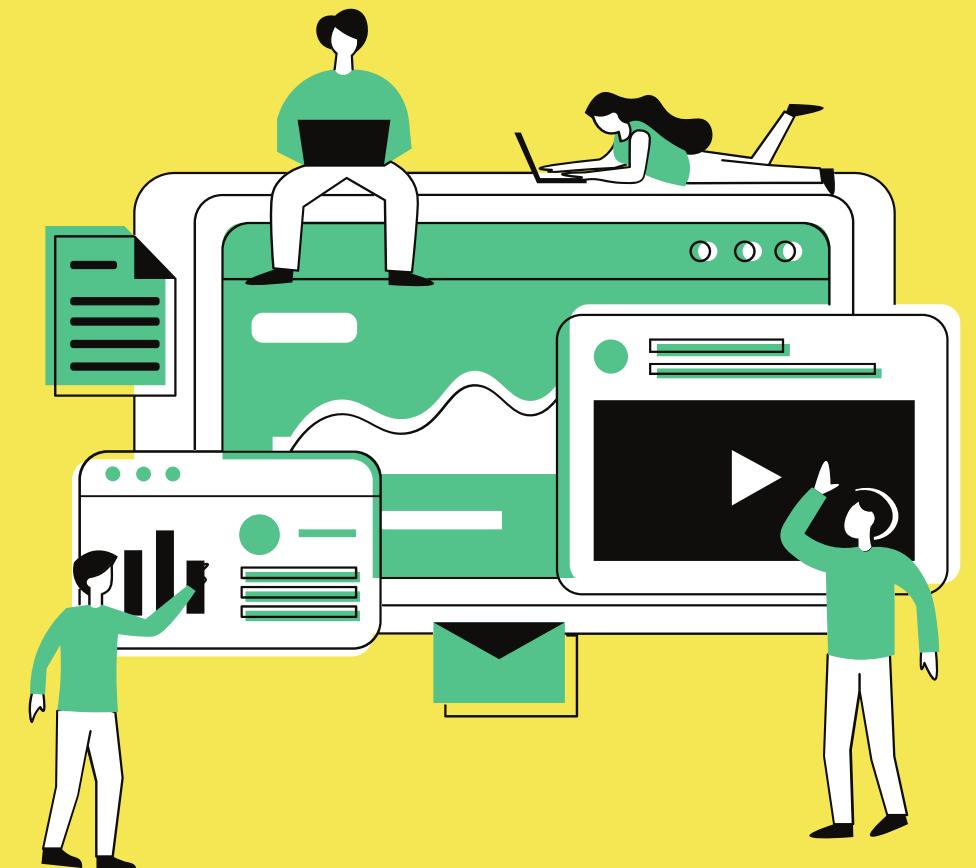
02

Están conformados por índice (valores = a partir de 1), día (valores = del 1 al 30 o 31), mes (valores = del 1 al 12), año (valores = 2021), línea (valores = del 1 al 9, A, B y 12), estación (valores = nombre), afluencia (valores = enteros no negativos) y fecha (valores = del 1 al 30 o 31).

PRE-PROCESAMIENTO DE LOS DATOS.

PARA ESTA SECCIÓN SEGUIMOS LOS SIGUIENTES PASOS PARA PODER OBTENER UNA MEJOR INTERPRETACIÓN FINAL.

1. QUITAR LOS DATOS EN BLANCO.
2. CORREGIMOS REGISTROS INCOMPLETOS, COMPLETAMOS LOS INCOMPLETOS Y QUITAMOS LOS REGISTROS ERRONEOS (REGISTROS CON VALOR DE 0 O MENORES A 1000).
3. PARA EL CASO DE LAS ESTACIONES, QUITAMOS LOS TRANSBORDOS, SERÍAN NOMBRES REPETIDOS A FIN DE EVITAR PROBLEMAS CON EL PROCESAMIENTO DE DATOS.
4. AGREGAMOS UNA COLUMNA LLAMADA ID_ESTACIÓN, QUE HACE REFERENCIA A UN DICCIONARIO DE DATOS, CON EL FIN DE ASIGNARLE NÚMEROS DEL 1 AL 163 A LAS ESTACIONES PARA FACILITAR NUESTRO TRABAJO.



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.

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

DEFINICIÓN DE LAS VARIABLES.

VARIABLE INDEPENDIENTE



X, es la estación del metro representada por su ID.

VARIABLE DEPENDIENTE



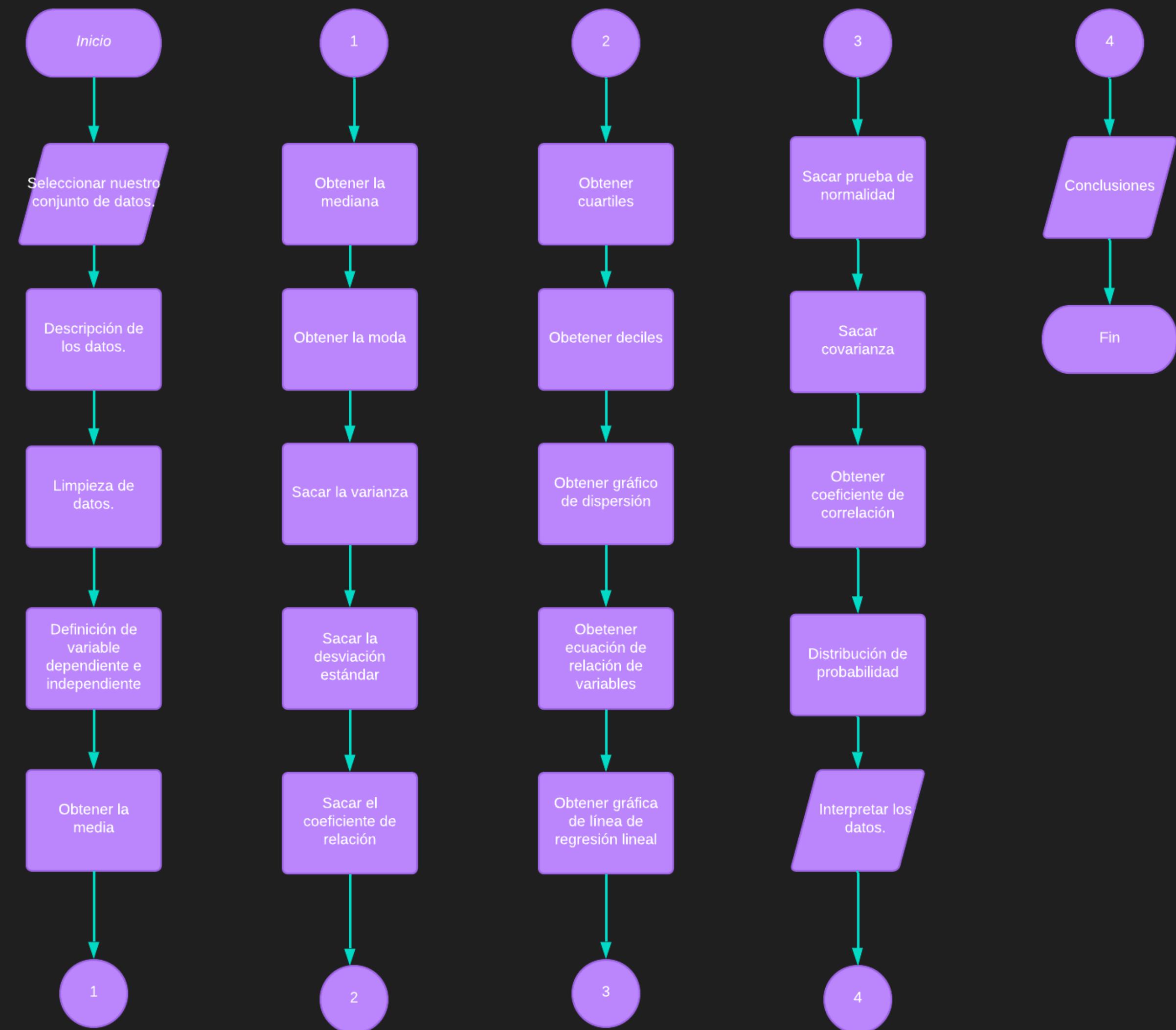
Y, es la afluencia de las personas.

RELACIÓN ENTRE VARIABLES



Relacionamos la existencia entre la estación del metro y la cantidad de personas que ingresan al SCTM.

DIAGRAMA DE FLUJO



DESARROLLO DEL ANÁLISIS ESTADÍSTICO

MEDIDAS DE TENDENCIA CENTRAL



MEDIA

11421.33

MEDIANA

8732

MODA

6683, 6748, 6930, 8755

MEDIDAS DE DISPERSIÓN



VARIANZA

99231657

DESV. EST.

9961.509

MODA

0.8721843

```
1estimadoresestadisticos<-function(vectordedatos){  
2  
3  cat("Media: ", mean(vectordedatos))  
4  cat("\nMediana: ", median(vectordedatos))  
5  cat("\nModa: ", mfv(vectordedatos))  
6  cat("\n\nMedidas de dispersion")  
7  cat("\nVarianza: ", var(vectordedatos))  
8  cat("\nDesviacion estandar: ", sd(vectordedatos))  
9  cat("\nCoeficiente de correlacion: ", desvestR/promR)  
10  
11}
```

CUARTILES Y DECILES

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

Deciles:									
0%	10%	20%	30%	40%	50%	60%	70%	80%	90%
11.	2900.	4268.	5745.	7156.	8732.	10598.	12829.	15966.	23175.
0	0	0	0	0	0	0	0	0	5

100%
88734.
0



```
1 cat("\nLas medidas de tendencia central y de dispersion de la  
afluencia del metro son:")  
2 estimadoresestadisticos(datosarchivo$Afluencia)  
3 cat("\nCuartiles: ")  
4 cuartilR<-quantile(datosarchivo$Afluencia)  
5 cuartilR  
6 cat("\nDeciles: ")  
7 decilR<-quantile(datosarchivo$Afluencia, prob=seq(0,1,length=11))  
8 decilR
```

GRÁFICOS DE DISPERSIÓN

GRÁFICO DE DISPERSIÓN

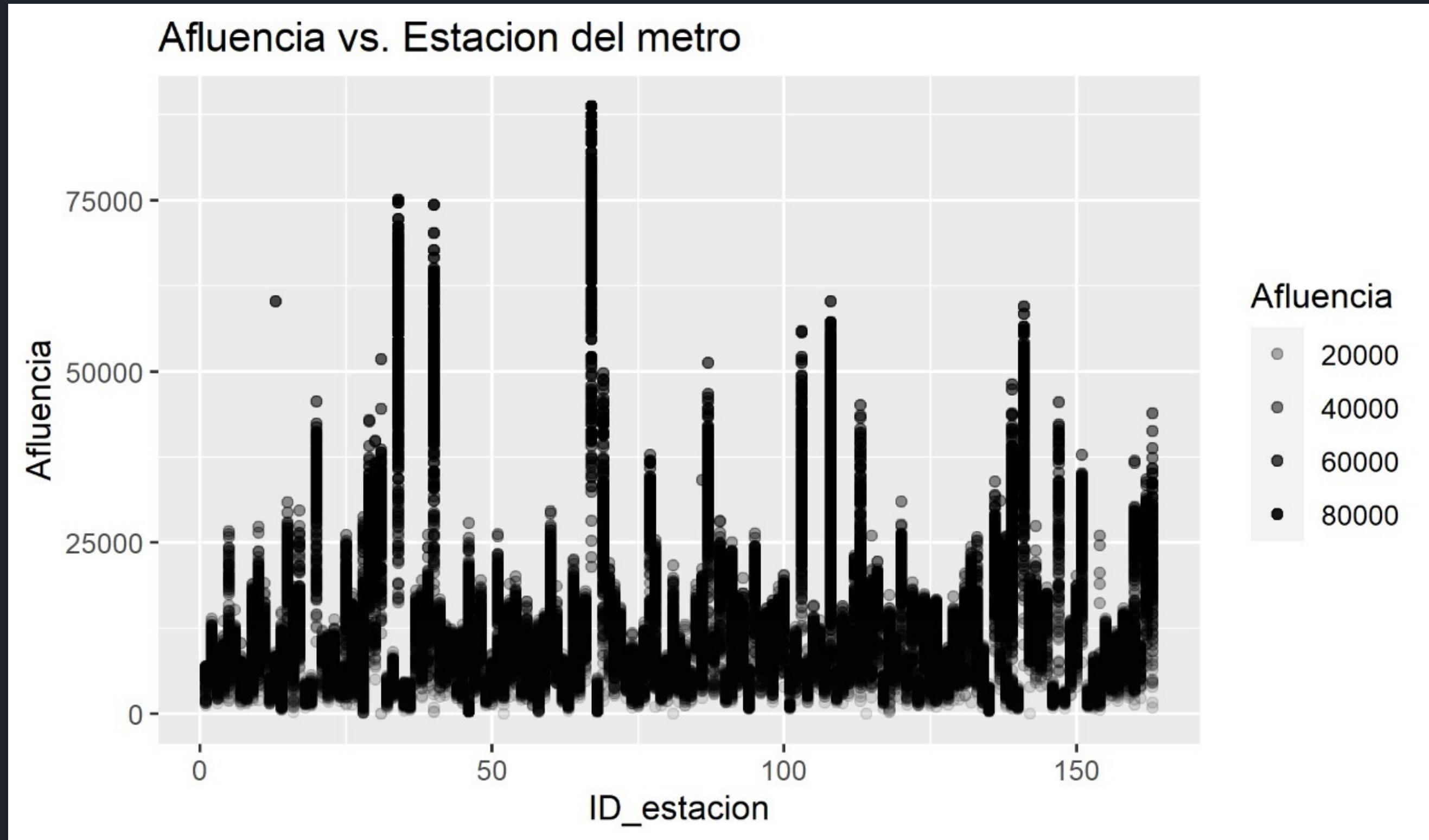
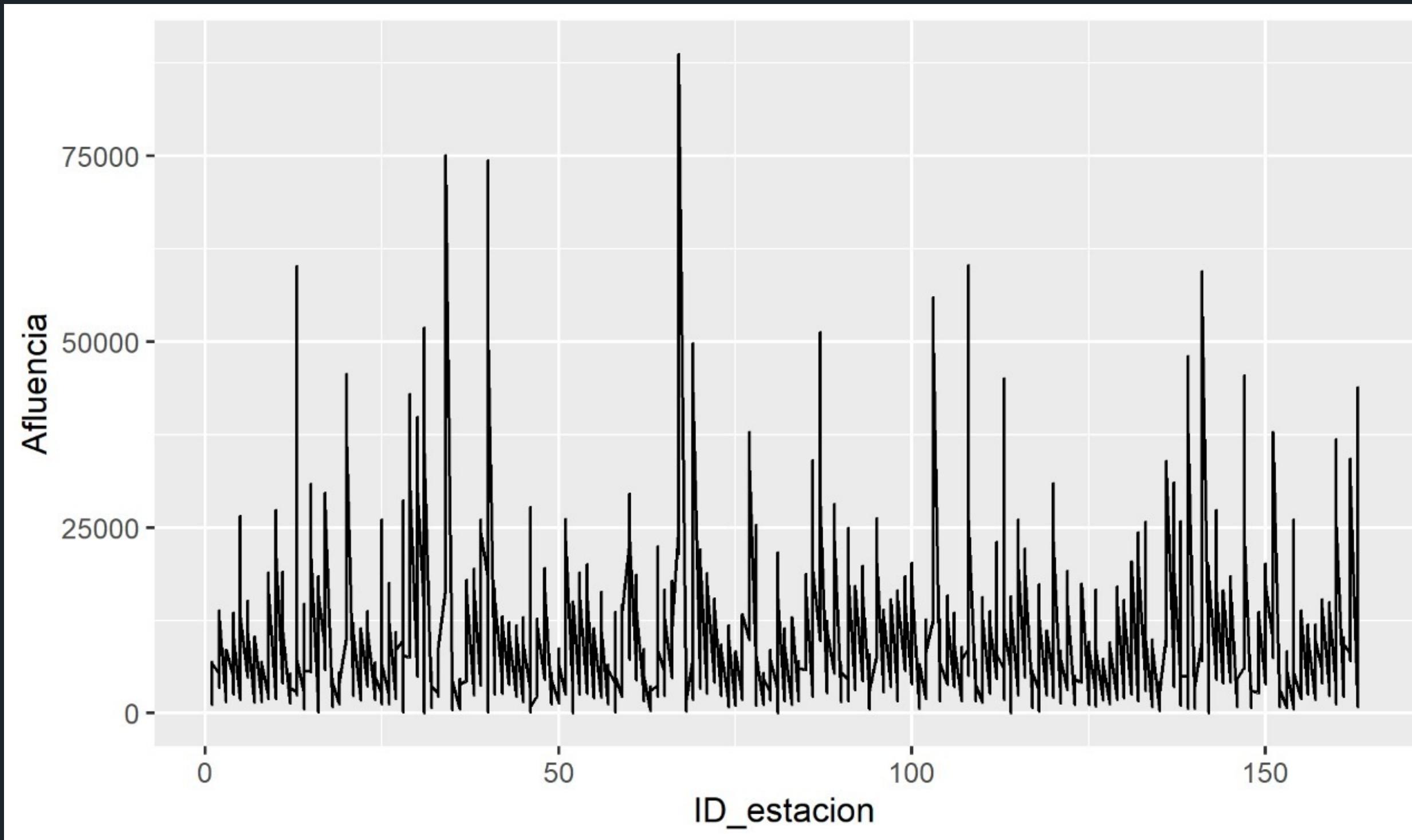


GRÁFICO DE DISPERSIÓN





```
1 ##Grafico de Dispersion
2 graficadispersion<-
  ggplot(datosarchivo,aes(x=ID_estacion,y=Afluencia))
3 graficadispersion+geom_point(aes(alpha = Afluencia))
4 ggsave("graficodispersioncompleto.png")
5 graficadispersion<-
  ggplot(datosarchivo,aes(x=ID_estacion,y=Afluencia))
6 graficadispersion + geom_line()
7 ggsave("2graficodispersioncompletolinea.png")
```

GRÁFICO DE DISPERSIÓN

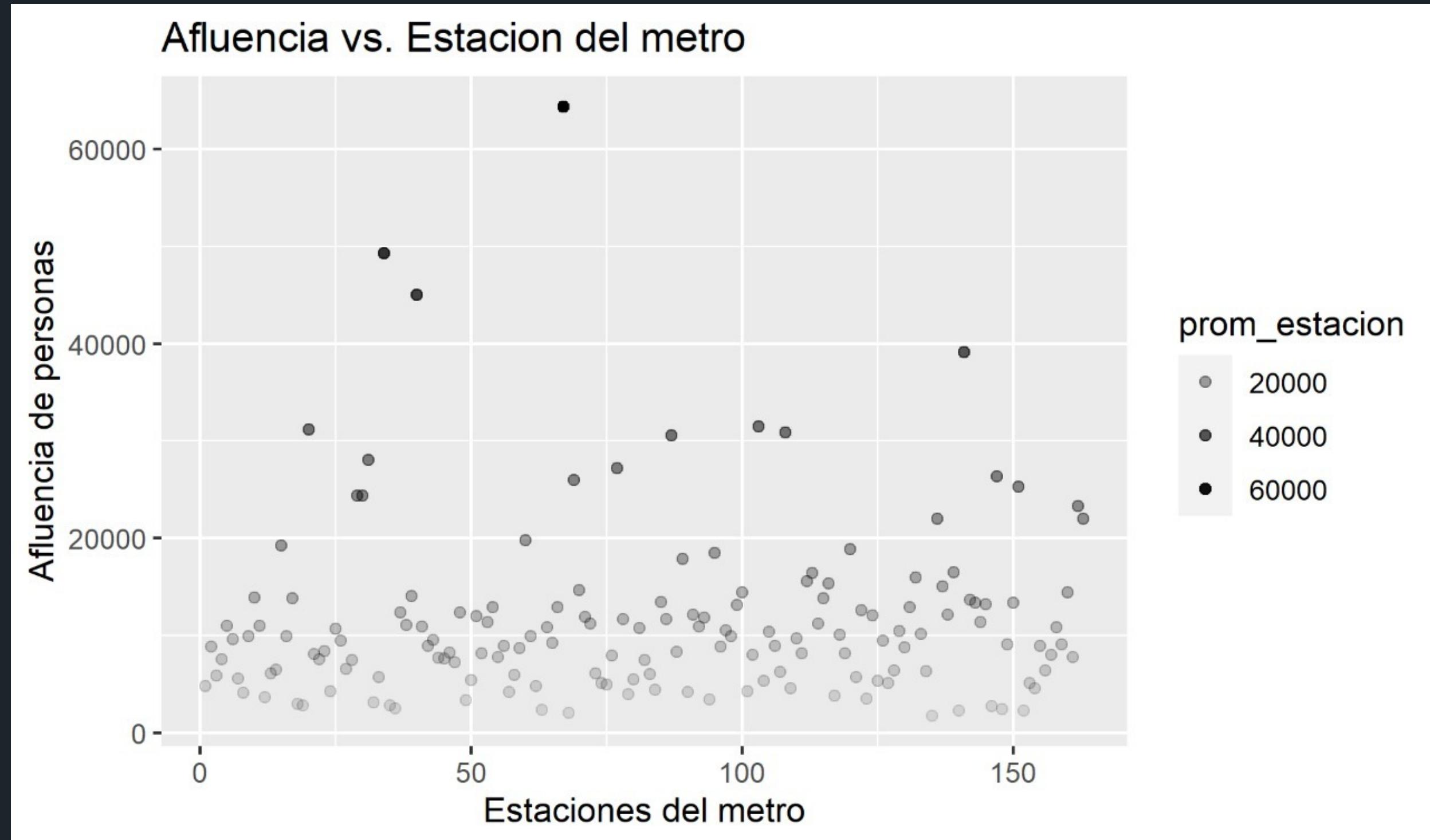
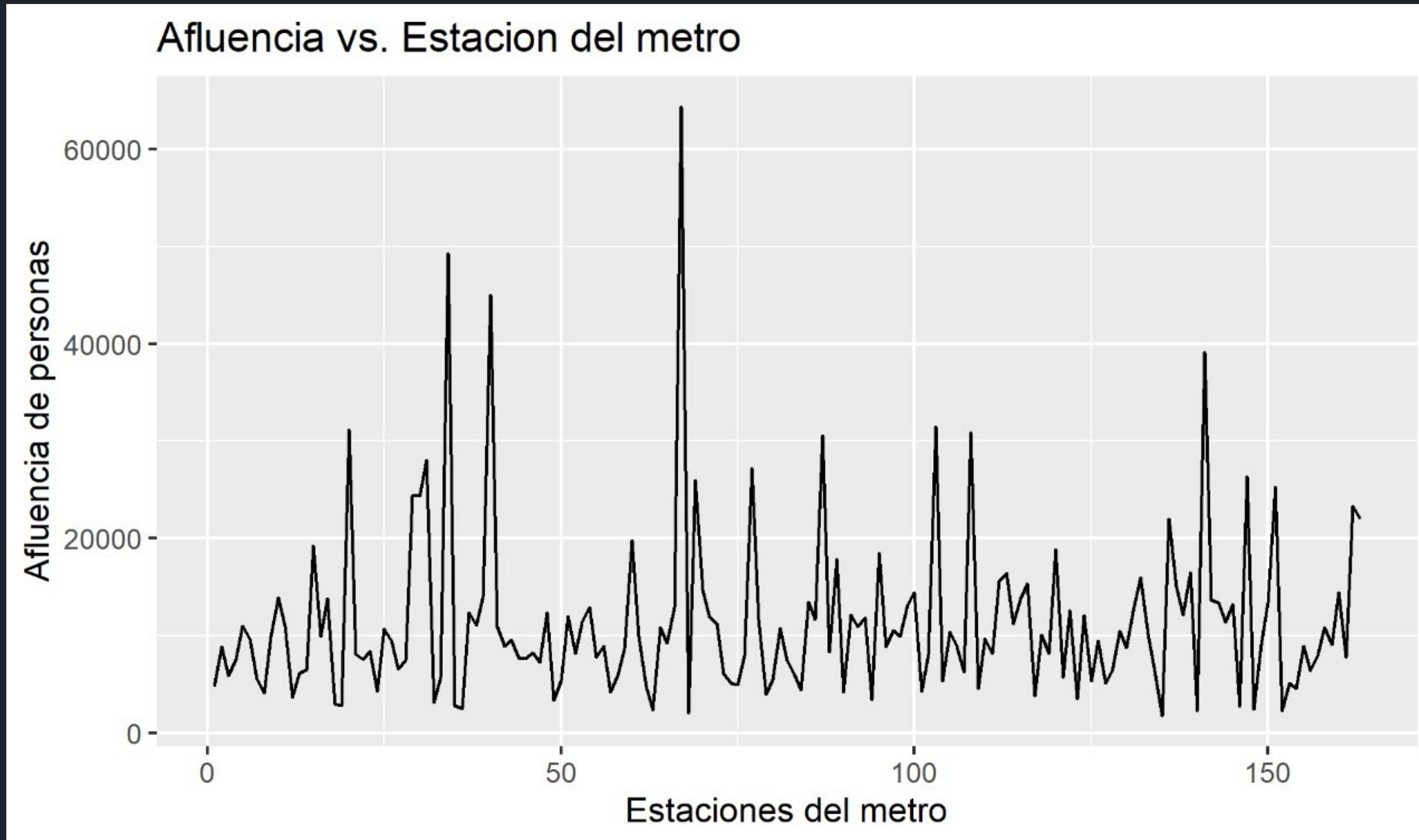


GRÁFICO DE DISPERSIÓN





```
1 graficadispersion<-
  ggplot(datosarchivo,aes(x=id_estacion,y=prom_estacion))+  

  2   ggttitle("Afluencia vs. Estacion del metro")  

    +xlab("Estaciones del metro")+ylab("Afluencia de personas")  

  3 ##graficadispersion+geom_point(aes(alpha = promestacion))  

  4 graficadispersion+geom_point(aes(alpha = prom_estacion))  

  5 ggsave("3graficodispersionpromedio.png")  

  6 graficadispersion+xlab("Estaciones del metro")  

    +ylab("Afluencia de personas") + geom_line()  

  7 ggsave("4graficodispersionlinea.png")
```

REGRESIÓN Y CORREALCIÓN 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



```
1 ##4Generacion del modelo lineal
2 modelolineal<-
  lm(datosarchivo$ID_estacion~datosarchivo$Afluencia,data=datosarchivo)
3 summary(modelolineal)
```

GRÁFICO DE DISPERSIÓN

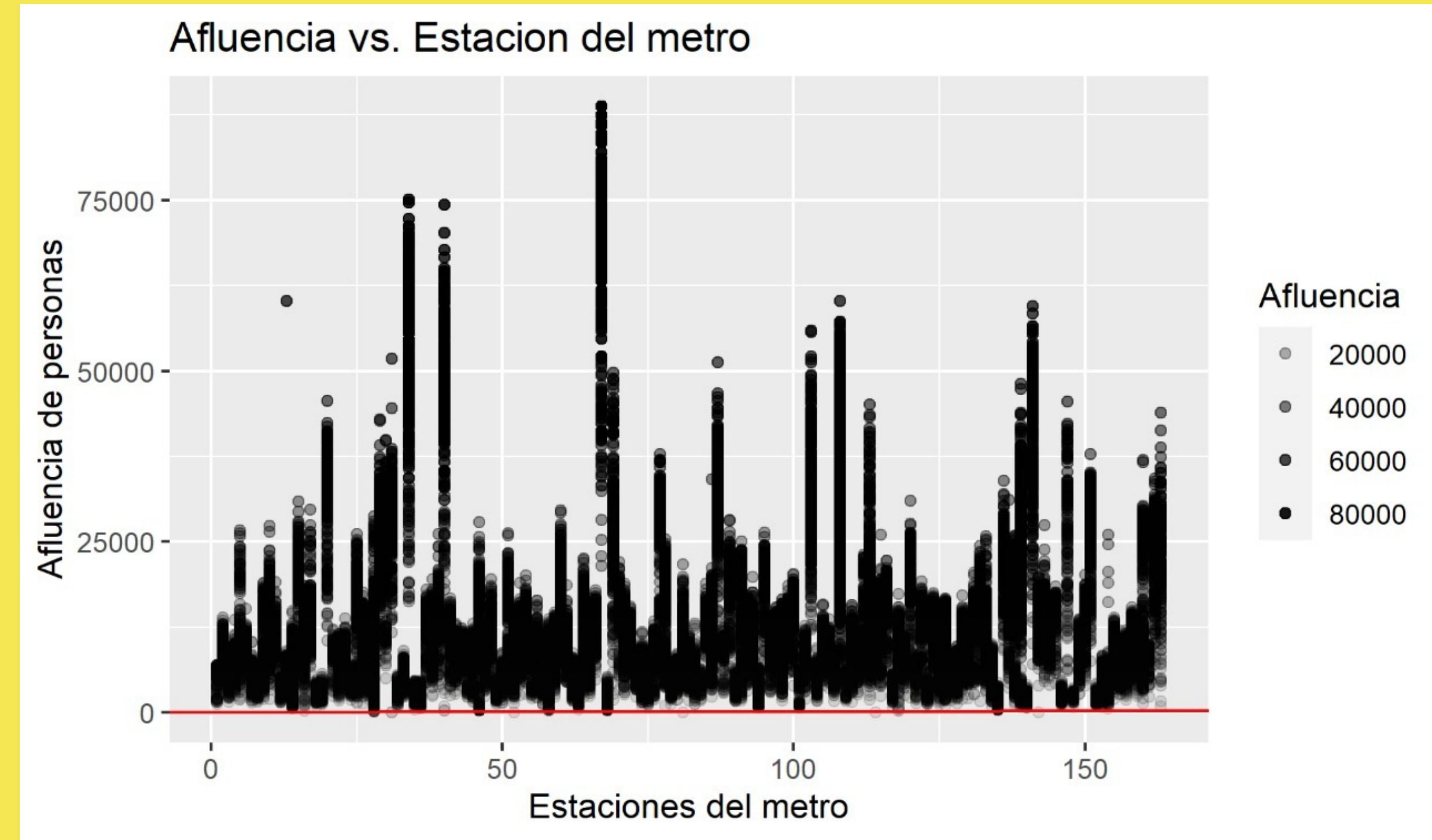
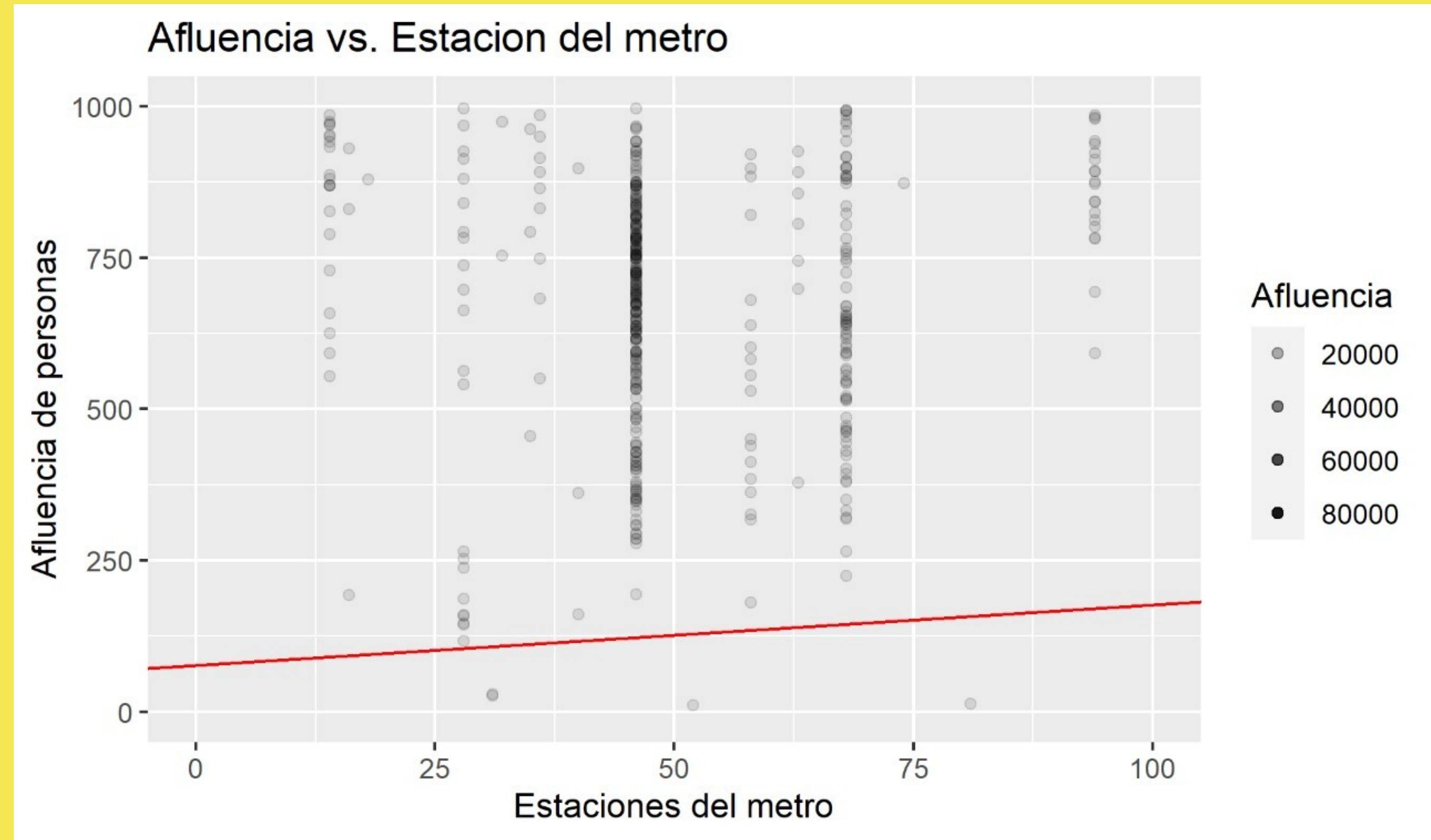
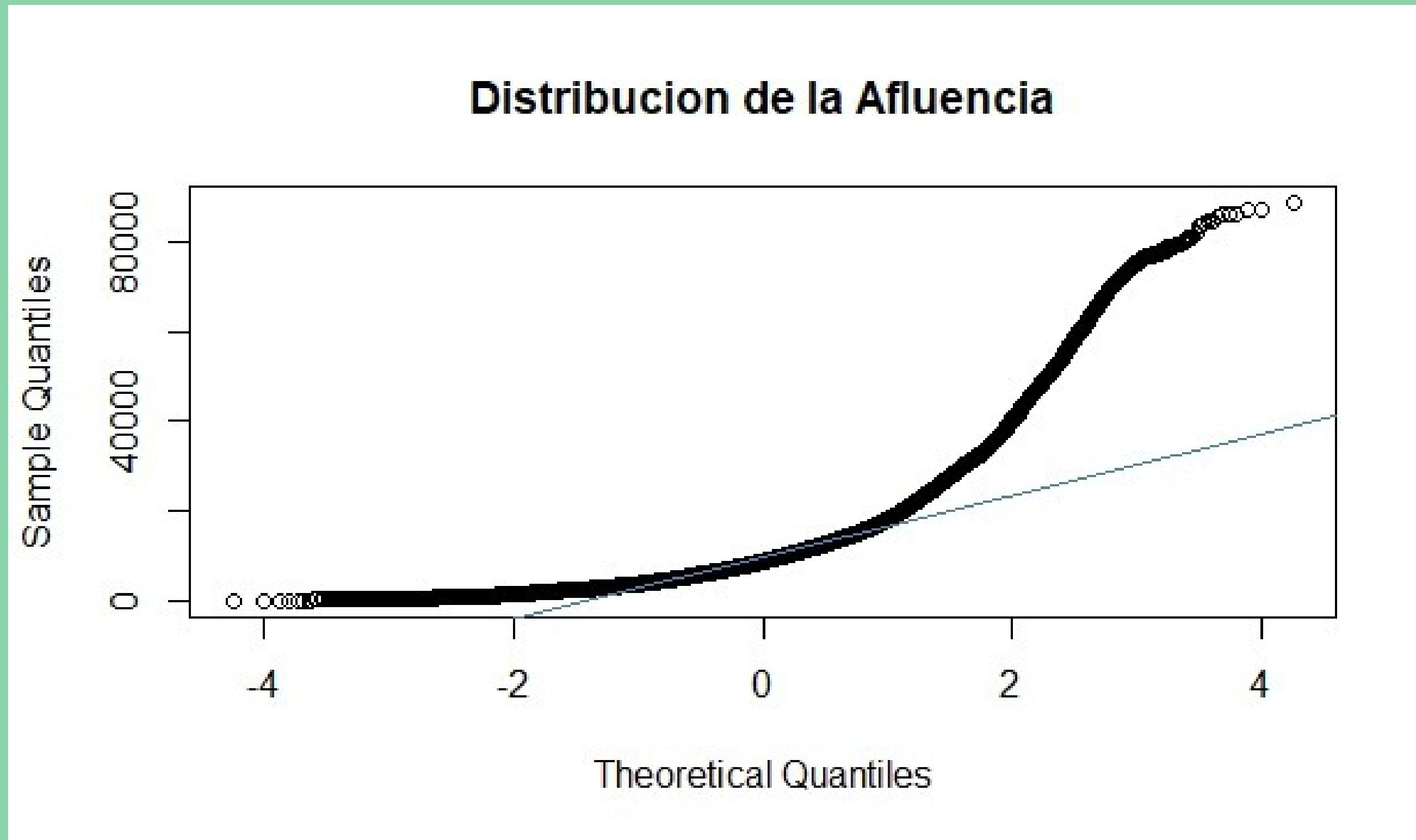


GRÁFICO DE DISPERSIÓN



```
1##5 Grafico de Dispersion con linea de regresion lineal
2graficadispersion<-ggplot(datosarchivo,aes(x=ID_estacion,y=Afluencia))
+geom_point(aes(alpha = Afluencia))
3
4##Linea de la ecuacion lineal
5graficadispersion+geom_abline(intercept = 77.21026, color="Red")+
6  xlab("Estaciones del metro")+ylab("Afluencia de personas")
+ggtitle("Afluencia vs. Estacion del metro")
7
8ggsave("5graficodispersionconlm.png")
9
10graficadispersion+xlab("Estaciones del metro")+ylab("Afluencia de
personas")
11
12graficadispersion+ggtitle("Afluencia vs. Estacion del metro")
13graficadispersion+ xlim(0,100)+ylim(0,1000)+geom_abline(intercept =
77.21026, color="Red")+
14  xlab("Estaciones del metro")+ylab("Afluencia de personas")
+ggtitle("Afluencia vs. Estacion del metro")
15ggsave("6graficodispersionconlm2.png")
```

PRUEBAS DE NORMALIDAD





```
1 ##6 Determinacion de normalidad
2 lillie.test(datosarchivo$Afluencia)
3 qqnorm(datosarchivo$Afluencia, main="Distribucion de la Afluencia")
4 qqline(datosarchivo$Afluencia, col="steelblue")
5 ##ggsave("7graficonormalidad.png")
```



COVARIANZA Y COEFICIENTE DE CORRELACIÓN

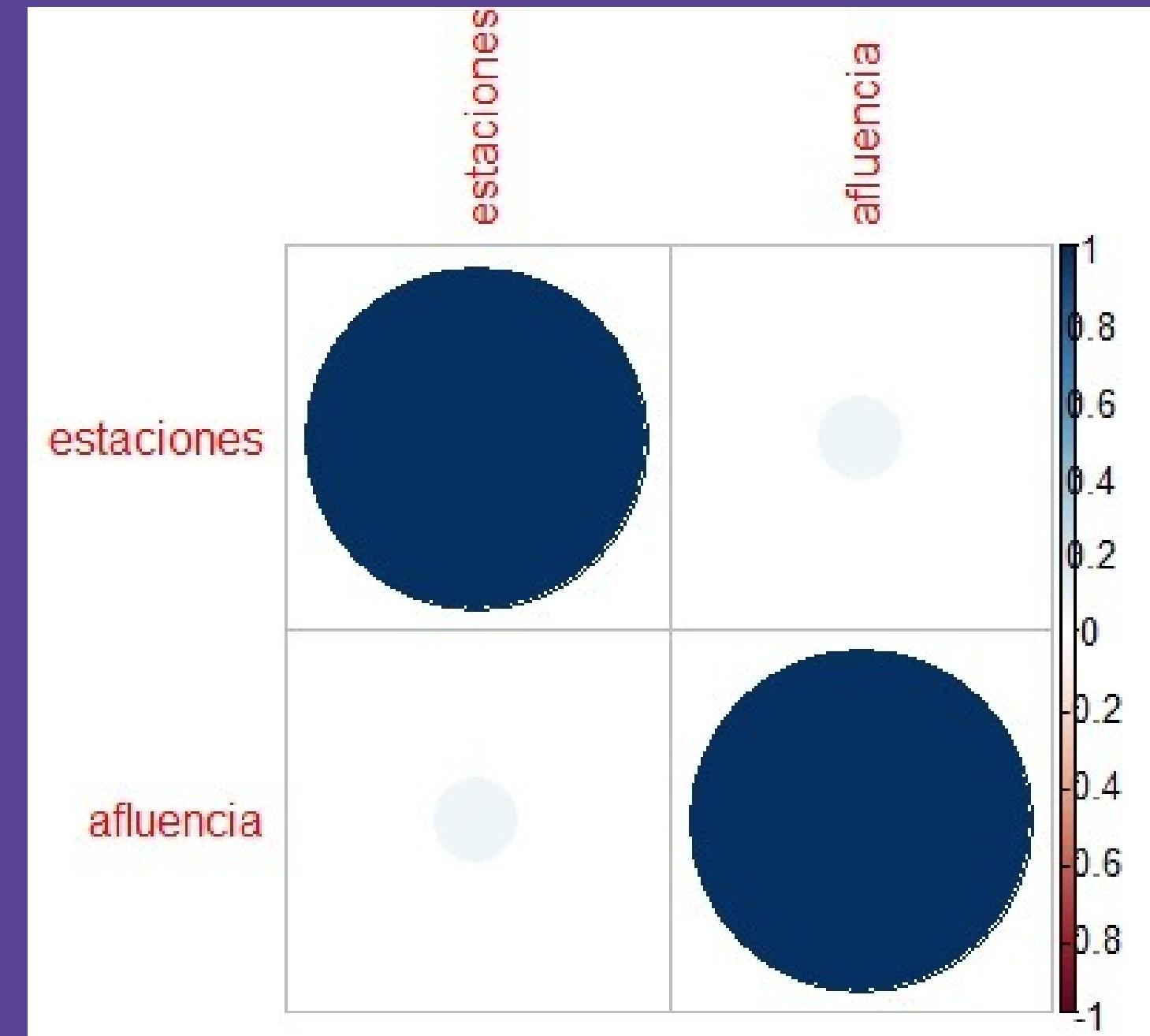
Matriz de covarianza		
	estaciones	afluencia
estaciones	2134.168	29799.67
afluencia	29799.67	99231656.69

Matriz de correlaciones		
	estaciones	Afluencia
estaciones	1.00000000	0.06475481
afluencia	0.06475481	1.00000000



```
1 ##7 Determinacion del coeficiente de correlacion
2 coefCor<-cor.test(x=datosarchivo$ID_estacion, y=datosarchivo$Afluencia,
method='pearson')
3 ##coefCor<-round(coefCor, digits = 2)
4 cat("El coeficiente de correlacion de las variables Estaciones del metro
y Afluencia es: ")
5 coefCor
```

COVARIANZA Y COEFICIENTE DE CORRELACIÓN



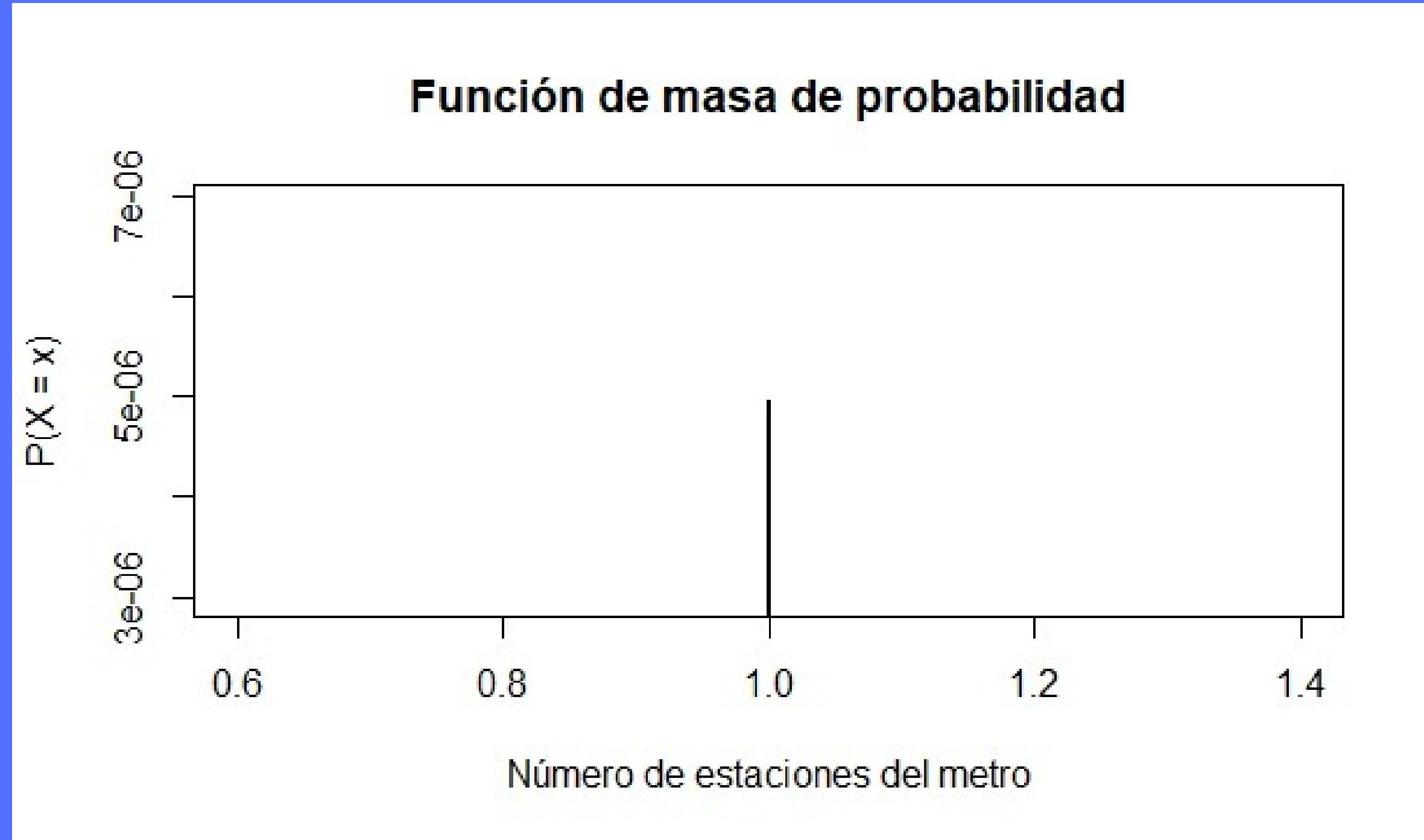


```
1 ##8 Matriz de correlaciones
2 estaciones<-c(datosarchivo$ID_estacion)
3 afluencia<-c(datosarchivo$Afluencia)
4 datosjuntos<-data.frame(estaciones,afluencia)
5 ##matrizCor<-cor(datosjuntos, method = "pearson")
6 matrizCor<-round(cor(datosjuntos, method = "pearson"),2)
7 matrizCor
8 corrplot(matrizCor,method = "circle")
```



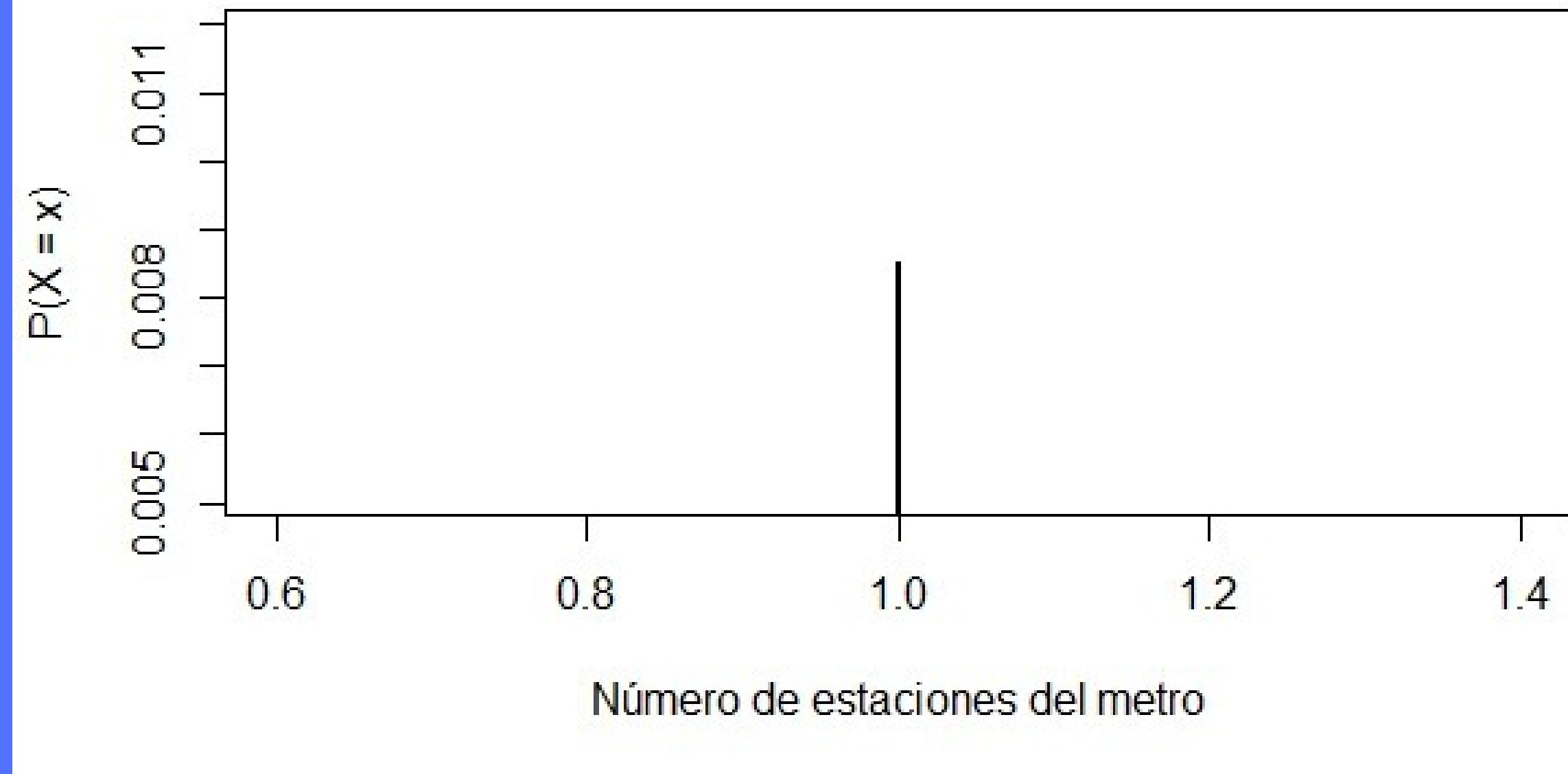
```
1 ##8.1MatrizdeCovarianzas
2 cov(estaciones,afluencia)
3 cov(datosjuntos)
4 covarianza<-cov(datosjuntos,method = "pearson")
5 covarianza
```

DISTRIBUCIÓN DE PROBABILIDAD

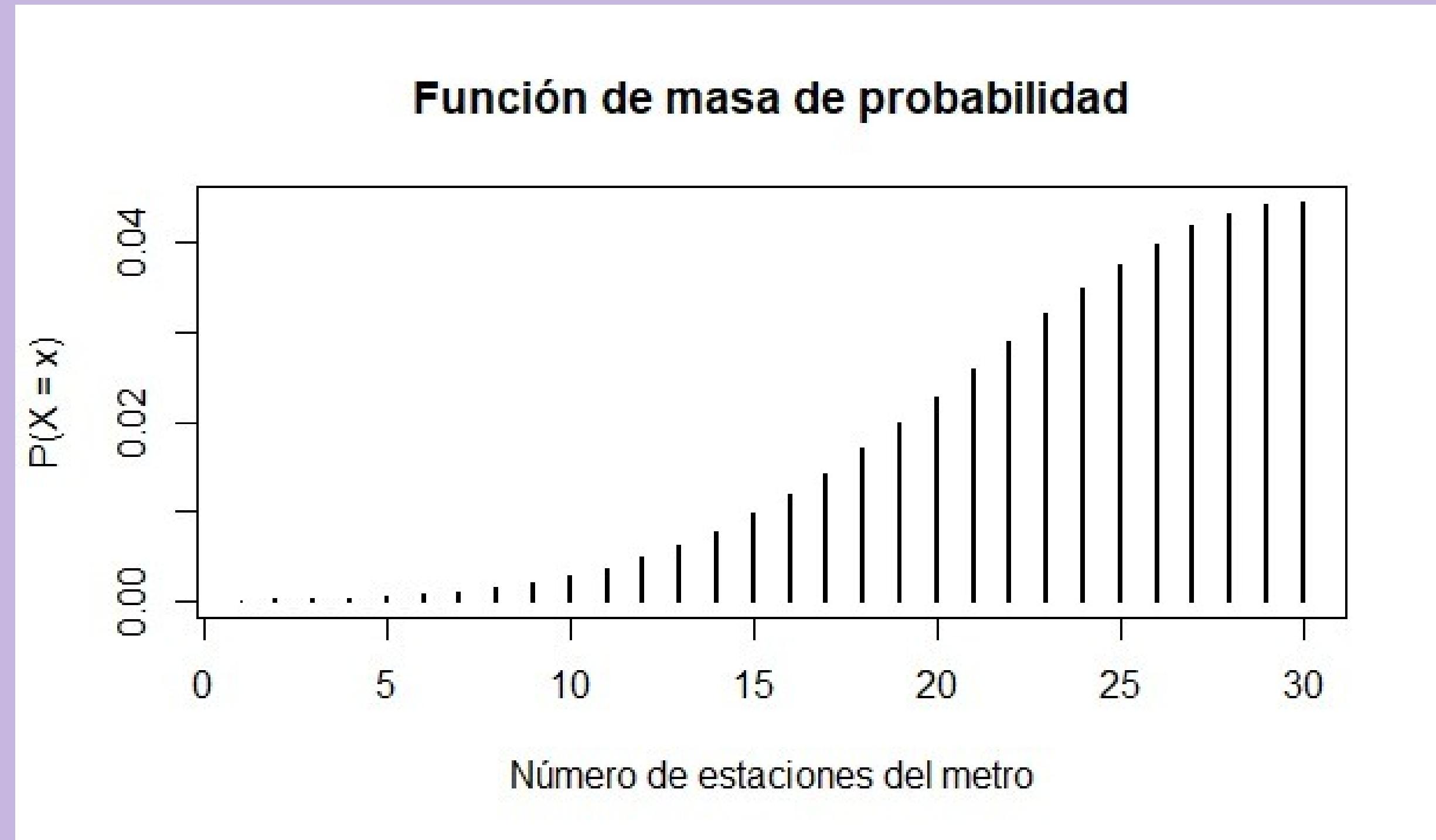


DISTRIBUCIÓN DE PROBABILIDAD

Función de masa de probabilidad

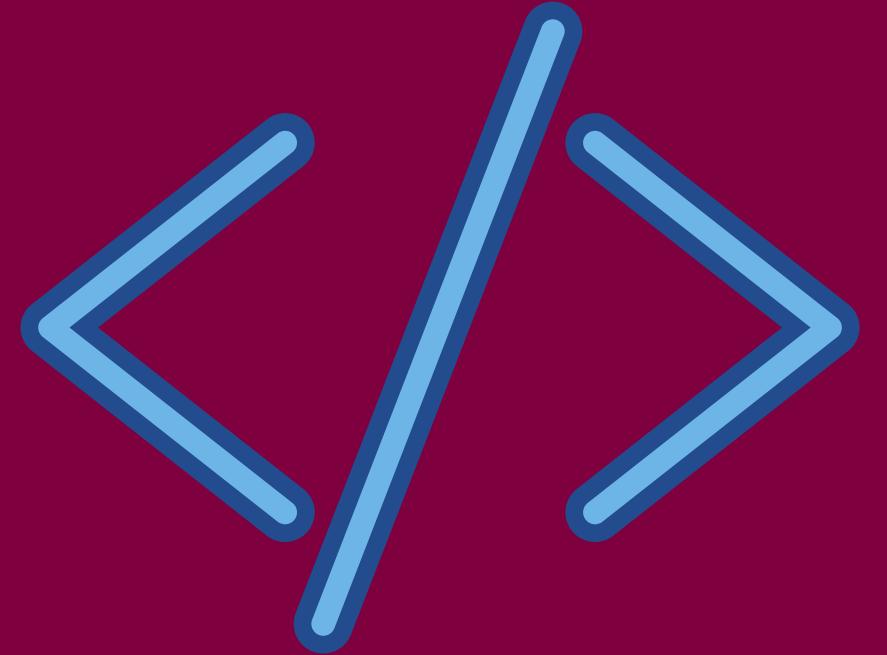


FUNCIÓN MASA DE PROBABILIDAD

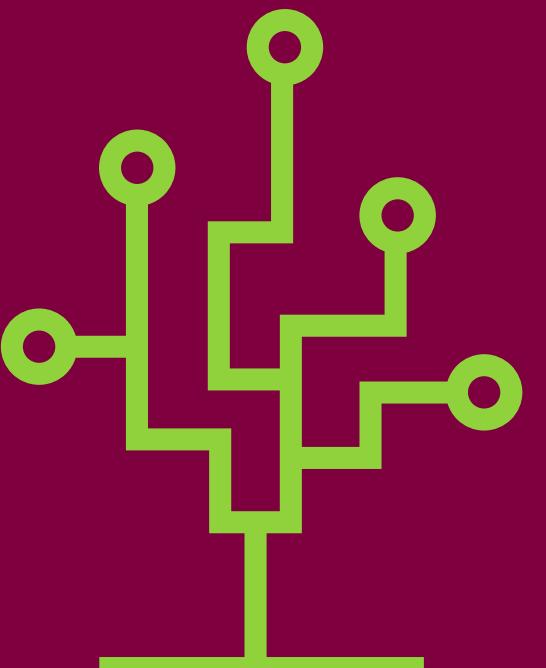




```
1 ##9DistribucionPoisson
2 lamda<-mean(datosarchivo$ID_estacion)
3 funcionPoisson<-dpois(51:80, lamda)
4 format(funcionPoisson, scientific = FALSE)
5
6 graficaPoisson<-plot(funcionPoisson, type = "h", lwd = 2, main =
  "Función de masa de probabilidad", ylab = "P(X = x)", xlab =
  "Número de estaciones del metro")
7 graficaPoisson
```



CONCLUSIONES



MUCHAS GRACIAS!