Analisis de bases y manipulación:

library(readr)

library(tidyverse)

library(lubridate)

#Es muy facil conectar con sistemas de bases de datos con tidyverse

# Leer la base de datos que acabamos de importar

datos <- read\_csv("/Volumes/External/carpetas-de-investigacion-pgj-cdmx.csv")

# <- significa igual en R, lo puedo usar para asignar valores etc etc...

#Vamos a aplicar funciones: cuantos datos tengo con tally(.) ?

tally(datos)

#de otra forma usando pipe %>% aplica la funcion del lado derecho al objeto del

#lado izquierdo: base %>% tally()

datos %>% tally()

#Ahora vamos a calcular la media: glimpse() nos resume los datos

datos %>% glimpse()

datos[1,2] # accedemos a la entrada de una matriz

datos$delito # vemos la columna delito

#vamos a hacer que lea la fecha estructurada

datos <- datos %>% mutate(fecha\_nueva=ymd\_hms(fecha\_hechos))

#Vamos a hacer una nueva columna de la base de datos y me voy quedar con

#Solo la fecha sin la hora.

datos <- datos %>% mutate(fecha=date(fecha\_nueva))

#Ahora queremos agrupar por fecha y contar

conteo\_delitos <- datos %>% group\_by(fecha) %>% tally()

#Calculamos la media

mean(conteo\_delitos$n)

media <- conteo\_delitos %>% summarise(mean(n))

desv.est <- conteo\_delitos %>% summarise(sd(n))

x.barra <- mean(conteo\_delitos$n)

n.longi<- length(conteo\_delitos$n)

media

desv.est

x.barra

n.longi

#Calculo del Skewness=Asimetria

coefAsim <- (1/desv.est^3)\*mean((conteo\_delitos$n - x.barra)^3)

coefAsim

#mediana

median(conteo\_delitos$n)

conteo\_delitos%>% summarise(mean(n), median(n), sum(n), var(n), mad(n), sd(n))

#Dia promedio que ocurren los delitos

datos <- datos %>% mutate(`Día de la semana` = weekdays(fecha))

conteo.dia <- datos %>% group\_by(`Día de la semana`) %>% tally()

conteo.dia <- conteo.dia %>% mutate(NumDia= NA)

#conteo.dia <- #conteo.dia %>% mutate(NumDia=if\_else(`Día de la semana`=="Friday", 5, NumDia))

conteo.dia$NumDia<- c(5,1,6,7,4,2,3)

#Los estadísticos

mean(conteo\_delitos$n)

mad(conteo\_delitos$n) #promedio(|x\_i - xbarra|)

median(conteo\_delitos$n)

quantile(conteo\_delitos$n, 0.25)

?var

n <- length(conteo\_delitos$n)

(n-1)/n\*var(conteo\_delitos$n)

cuantil <- quantile(conteo\_delitos$n, 0.25)

cuantil[1] + 20

as.numeric(cuantil[1])

library(moments)

conteo\_delitos %>% summarise(IQR(n), kurtosis(n))

IQR(conteo\_delitos$n)

ggplot(conteo\_delitos) +

geom\_boxplot(aes(x = n))

ggplot(conteo\_delitos) +

geom\_boxplot(aes(y = n), color = "#3ba17d", fill = "deepskyblue4") +

labs(

title = "Distribución de las carpetas de investigación de delitos en PGJ",

subtitle = "Datos de CDMX, diciembre 2018",

x = "",

y = "Número de carpetas abiertas por día",

caption = "Datos de la AIP"

)

conteo\_tipo <- datos %>% filter(str\_detect(categoria\_delito, "ROBO"))

conteo\_tipo <- conteo\_tipo %>% group\_by(delito) %>% tally()

ggplot(conteo\_tipo) +

geom\_col(aes(x = delito, y = n)) +

theme(axis.text.x = element\_text(angle = 90, size = 3)) +

labs(

title = "Careptas de investigación iniciadas en PGJ por robo",

subtitle = "Datos de CDMX, diciembre 2018",

x = "Tipo de robo",

y = "Conteo de casos",

caption = "Datos de la AIP"

)

setwd("/cloud/project")

ggsave("Grafica\_barras.pdf", width = 6, height = 4)

ggplot(datos) +

geom\_point(aes(x = longitud, y = latitud), size = 1,

color = "purple", alpha = 0.1) + #Alpha = transparencia de 0 a 1

theme\_minimal()

ggplot(conteo\_delitos) +

geom\_line(aes(x = fecha, y = n)) +

geom\_point(aes(x = fecha, y = n), color = "red") +

theme\_classic() +

geom\_label(aes(x = dmy("25/12/2018"), y = 425),

label = "Navidad") +

labs(

x = "Fecha de aperturas de la carpeta",

y = "Número de carpetas abiertas",

title = "Carpetas de investigación abiertas en PGJ",

subtitle = "Datos abiertos CDMX, Diciembre 2018",

caption = "Fuente: AIP"

)

#EJEMPLO

año.primaria <- sample(c(1,2,3), 1000, replace = T)

#No tiene sentido un boxplot con estos datos:

ggplot() +

geom\_boxplot(aes(x = año.primaria)) +

geom\_point(aes(x = año.primaria, y = 0), position = "jitter")

ggplot() +

geom\_bar(aes(x = año.primaria))

#Tabla de contingencias x= Alcaldia, y=Año de inicio

table(datos$alcaldia\_hechos, datos$ao\_inicio)

addmargins(table(datos$alcaldia\_hechos, datos$ao\_inicio))

#Tabla de frecuencias

table(datos$alcaldia\_hechos, datos$ao\_inicio) %>% prop.table() %>% addmargins()

data("mtcars")

datos.coches<- mtcars

ggplot(datos.coches) +

geom\_point(aes(x=wt, y=mpg), color= "red") + theme\_bw()

x<- datos.coches$wt

y<- datos.coches$mpg

#Calculamos el coseno con producto punto/ norma

# %\*% es la multiplicacion de HADAMARD

producto.punto <- t(x) %\*% y

coseno <- producto.punto/(sqrt(sum(x^2))\*sqrt(sum(y^2)))

coseno

#Un coseno de 1 significaria paralelos i.e que la relación entre x, y es LINEAL.

#Un coseno de 0 significa que son ortogonales, no podemos describir de manera

#lineal la relación x,y

#Esto es el coseno pero de x- media, y-media (centralizado)

cor(x, y, method="pearson")

cor(x,y, method = "spearman")

calidad\_alimentos <- factor(c("malo", "bueno", "bueno", "regular", "bueno",

"bueno"), order= TRUE, levels= c("malo", "regular",

"bueno"))

calidad\_servicio <- factor(c("1 estrella", "4 estrellas", "5 estrellas", "2 estrellas", "5 estrellas",

"4 estrellas"), order= TRUE, levels= c("1 estrella", "2 estrellas",

"3 estrellas", "4 estrellas", "5 estrellas"))

# Falta comentar esto

ggplot()+geom\_point(aes(x=calidad\_alimentos, y=calidad\_servicio,

size= calidad\_alimentos, color=calidad\_servicio))

cor(as.numeric(calidad\_alimentos), as.numeric(calidad\_servicio), method= "spearman")

cor(as.numeric(calidad\_alimentos), as.numeric(calidad\_servicio), method= "kendall")

#Corremos la regresión lineal

ggplot(datos.coches)+ geom\_point(aes(x= wt, y= mpg))+

geom\_smooth(aes(x= wt, y= mpg), method= "lm",

formula= y ~ x, se=FALSE)+ theme\_minimal()

hora.dia<- factor(c("MedioDía", "Anochecer", "Medianoche", "Amanecer"), order=TRUE, levels = c("Amanecer", "MedioDía", "Anochecer", "Medianoche"))

cantidad.aves<-factor(c(200,100,0,100), order= TRUE)

cor(as.numeric(hora.dia), as.numeric(cantidad.aves), method= "spearman")

ggplot()+geom\_point(aes(x=hora.dia, y=cantidad.aves,

size= hora.dia, color=cantidad.aves))

a<-c(-1/2,1/2,3/2,-3/2)

b<-c(7/4,-1/4,-5/4,-1/4)

producto.punto1 <- t(a) %\*% b

producto.punto1

(sqrt(sum(a^2))\*sqrt(sum(b^2)))

Crear Bases de Datos:

library(readr)

library(tidyverse)

library(lubridate)

nsim<-1000 #Cantidad de simulaciones

N<- 10000 #Tamaño de población

n<- 100 #tamaño de la muestra por cada simulación

base.datos<- data.frame(x=rpois(N,lambda= 4))#Esta es la U en nuestra notacion

media.U <-mean(base.datos$x)

media.muestra<- rep(NA,nsim)

#Ciclo a través de las 1000 simulaciones

for(i in 1:nsim){

muestra<- sample(base.datos$x, n, replace=FALSE)#Muestreo simple sin remplazo

media.muestra[i] <- mean(muestra)}#cada uno de estos es un xbarra}

mean(media.muestra)

media.U

ggplot()+

geom\_histogram(aes(x=media.muestra, y=..density..), fill="red")+

geom\_vline(aes(xintercept=media.U), linetype= "dashed")

#Se puede ver que estamos aproximando demasiado bien la media de U

#Ahora vamos a calcular la varianza de la xbarra\_s

var.xbarra<- ((1-n/N)/n)\*var(base.datos$x)

var.muestral.xbarra<- rep(NA,nsim)

for(i in 1:nsim){

muestra <- sample(base.datos$x, n, replace=FALSE)

var.muestral.xbarra[i]<- ((1-n/N)/n)\*var(muestra)

}

ggplot()+

geom\_density(aes(x=var.muestral.xbarra), color="red")+

geom\_vline(aes(xintercept=var.xbarra))

x.data<-c(1,2,3)

datos.nuevos<- c()

for (i in 1:length(x.data)){

datos.nuevos<- c(datos.nuevos, x[i]^i)

}

mean(datos.nuevos)

DENLE UNA PASADA A:

* r-coder.com
* cookbook-r.com
* kaggle.com
* rpubs.com
* Cursos de R que creo que están buenos:
* Curso UNAM, antes era gratuito, ahora ya no sé: [https://www.coursera.org/learn/intro-data-science-programacion-estadistica-r (Enlaces a un sitio externo.)](https://www.coursera.org/learn/intro-data-science-programacion-estadistica-r)
* En particular, a mi me gustan los cursos de KIrill Emerenko en Udemy:
  + Básico: [https://www.udemy.com/course/r-programming/ (Enlaces a un sitio externo.)](https://www.udemy.com/course/r-programming/)
  + Intermedio: <https://www.udemy.com/course/r-analytics/>