Instituto Politécnico Nacional

Escuela Superior de Computo

Programación para la ciencia de datos.

Cristal Karina Galindo Durán

Practica 6:

ESTADISTICA DESCRIPTIVA

Vianey Maravilla Pérez

3AM1

Unidad temática a la que corresponde la práctica.

II Análisis exploratorio de Datos

***Objetivo**.*

Realizar scripts en Lenguaje R que permita implementar análisis de regresión lineal simple, correlación; así como, obtener la matriz de correlación.

***Introducción**.*

La regresión lineal simple y la correlación son métodos estadísticos que estudian la relación existente entre dos variables. De forma específica la correlación lineal cuantifica el como están relacionadas dos variables, mientras que la regresión lineal simple genera un modelo teniendo como base la relación entre ambas variables (dependiente e independiente).

En esta actividad se incluyen un conjunto de ejercicios que le permites al discente poner en práctica conceptos sobre la regresión lineal y correlación; así como, la obtención de la matriz de correlación en Lenguaje R.

Material o equipo necesario para la práctica.

- *Computadora*
- *Internet*
- *Lenguaje R y R Studio.*

Ejercicios:

1. Descargar del siguiente enlace: https://archive.ics.uci.edu/ml/datasets/Parking+Birmingham
2. Importarlo a R
3. Verificar la normalidad para las variables utilizadas (gráfica y función)
4. Realizar el diagrama de dispersión entre las variables
5. Encontrar el modelo que mejor se ajuste a los datos
6. Obtener el coeficiente de correlación que mejor se adecue a la distribución de los datos
7. Obtener la matriz de correlación entre las variables analizadas
8. Interpretar resultados

Consideraciones:

Para lograr tener un buen desarrollo en esta practica se tiene que tomar en cuenta que debemos utilizar lillie.test(x) e instalar una paquetería llamada "nortest".

Para poder interpretar los resultados de manera correcta debemos tener un buen desarrollo para eso lo visto en clase hará demasiada participación, así como también unas investigaciones a parte para poder limpiar datos, es decir, poder identificar los datos negativos y repetidos para así limpiarlos, luego entonces, tenemos que investigar como poder agrupar los datos.

Con esos pasos previos ya tendremos lista la información para comenzar la ejecución de la práctica número 6.

Procedimiento:

```
"Practica número 6
1.  Descargar del siguiente enlace: https://archive.ics.uci.edu/ml/datasets/Parking+Birmingham
2.  Importarlo a R
3.  Verificar la normalidad para las variables utilizadas (gráfica y función)
4.  Realizar el diagrama de dispersión entre las variables
5.  Encontrar el modelo que mejor se ajuste a los datos
6.  Obtener el coeficiente de correlación que mejor se adecue a la distribución de los datos
7.  Obtener la matriz de correlación entre las variables analizadas
8.  Interpretar resultados
Hecho por: Maravilla Pérez Vianey 3AM1"

#Instalamos el paquete y las libcerias a usar

library(nortest)
library(ggplot2)
library(psych)
library(ggcorrplot)
library(dplyr)


#Importamos nuestro archivo .csv desde su ubicación para así tener los datos en el código

archivo<- read.csv("C:/Users/viane/Desktop/ESCOM/3.-TERCER SEMESTRE/PROGRAMACION PARA LAS CIENCIAS DE DATOS/Códigos/dataset.csv")

#Visualizamos los datos.

View(archivo)

#Agrupamos los datos para poder obtener la concurrencia, es el mejor método

archivo <- archivo %>%
select(SystemCodeNumber, Capacity, Occupancy)  %>%
group_by(SystemCodeNumber) %>%
summarize(mean(Capacity), mean(Occupancy))

#Visualizamos los datos nuevamente

View(archivo)

#Cambiamos los nombres de nuestras columnas

colnames(archivo) <- c("Id" , "Capacity" , "Occupancy")

# Visualizamos los datos nuevamente

View(archivo)

#Hacemos las pruebas de normalidad para poder hacer la graficacion
#Capacidad y Ocupados
hist(archivo$Capacity, col = "purple", xlab = "Capacidad", border = "green", main = "Histograma -> Capacidad")
qqnorm(archivo$Capacity, pch=19, col = "purple")
qqline(archivo$Capacity, col = "green", lwd=2)

hist(archivo$Occupancy,col="purple", xlab = "Ocupados", border = "green", main = "Histograma -> Ocupados")
qqnorm(archivo$Occupancy, pch=19, col="purple")
qqline(archivo$Occupancy, col="green", lwd=2)



#Implementamos lillie.test que nos indica la profesora, junto con shapiro.test.

lillie.test(archivo$Capacity) #Capacidad
lillie.test(archivo$Occupancy) #Ocupabilidad
shapiro.test(archivo$Capacity) #Capacidad
shapiro.test(archivos$Occupancy) #Ocupabilidad

#Implementamos o en este caso hacemos la grafica de dispersion para poder interpretarla

plot(archivo$Capacity, archivo$Occupancy)
disp = ggplot (archivo, aes(x=Capacity, y=Occupancy))
```

```r
disp + geom_point()

fmd <- lm(archivo$Occupancy ~ archivo$Capacity, archivo)
summary(fmd)
disp + geom_point() + geom_smooth(method = "lm", colour = "red")


#Coeficiente de dispersion con Spearman sin distribucion normal

cor(archivo[2:3], method = "Spearman")
ggcorrplot(round(cor(archivo[2:3], method = "Spearman"), 2))

pairs.panels(archivos[2:3], method = "Spearman")
archivo %>%
ggplot(aes(x = Id, y = Occupancy)) + geom_bar(stat = "identity", fill= "#AEF8CD", alpha=.8, width= .5)
xlab("") + theme_bw()

```

Resultado:

```
> library(nortest)
> library(ggplot2)
> library(psych)
> library(ggcorrplot)
> library(dplyr)
> archivo<- read.csv("C:/Users/viane/Desktop/ESCOM/3.-TERCER SEMESTRE/PROGRAMACION PARA LAS CIENCIAS DE DATOS/Códigos/dataset.csv")
Warning messages:
1: In doTryCatch(return(expr), name, parentenv, handler) :
  display list redraw incomplete
2: In doTryCatch(return(expr), name, parentenv, handler) :
  invalid graphics state
3: In doTryCatch(return(expr), name, parentenv, handler) :
  invalid graphics state
> View(archivo)
```

	SystemCodeNumber	Capacity	Occupancy	LastUpdated
1	BHMBCCMKT01	577	61	2016-10-04 07:59:42
2	BHMBCCMKT01	577	64	2016-10-04 08:25:42
3	BHMBCCMKT01	577	80	2016-10-04 08:59:42
4	BHMBCCMKT01	577	107	2016-10-04 09:32:46
5	BHMBCCMKT01	577	150	2016-10-04 09:59:48
6	BHMBCCMKT01	577	177	2016-10-04 10:26:49
7	BHMBCCMKT01	577	219	2016-10-04 10:59:48
8	BHMBCCMKT01	577	247	2016-10-04 11:25:47
9	BHMBCCMKT01	577	259	2016-10-04 11:59:44
10	BHMBCCMKT01	577	266	2016-10-04 12:29:45
11	BHMBCCMKT01	577	269	2016-10-04 13:02:48
12	BHMBCCMKT01	577	263	2016-10-04 13:29:45
13	BHMBCCMKT01	577	238	2016-10-04 14:02:47
14	BHMBCCMKT01	577	215	2016-10-04 14:29:49
15	BHMBCCMKT01	577	192	2016-10-04 14:57:13
16	BHMBCCMKT01	577	165	2016-10-04 15:30:14
17	BHMBCCMKT01	577	162	2016-10-04 16:04:12
18	BHMBCCMKT01	577	143	2016-10-04 16:31:14
19	BHMBCCMKT01	577	54	2016-10-05 07:57:17
20	BHMBCCMKT01	577	59	2016-10-05 08:30:15
21	BHMBCCMKT01	577	71	2016-10-05 09:04:19
22	BHMBCCMKT01	577	83	2016-10-05 09:30:15

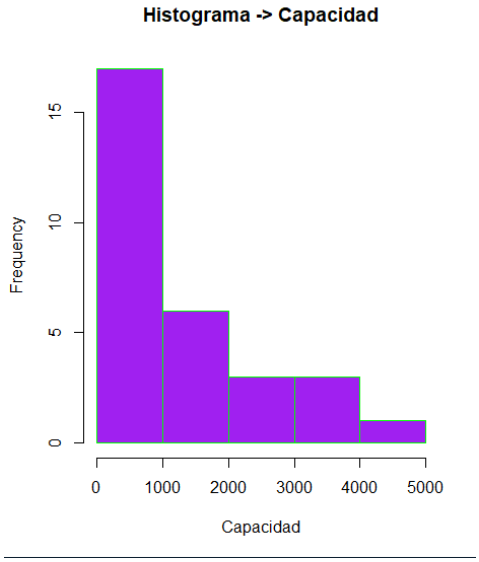
```
> archivo <- archivo %>%
+ select(SystemCodeNumber, Capacity, Occupancy)  %>%
+ group_by(SystemCodeNumber) %>%
+ summarize(mean(Capacity), mean(Occupancy))
> View(archivo)
```

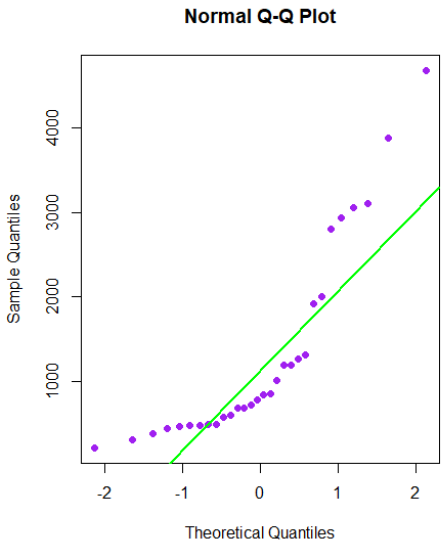
	SystemCodeNumber	mean(Capacity)	mean(Occupancy)
1	BHMBCCMKT01	577	162.02973
2	BHMBCCPST01	317	136.03683
3	BHMBCCSNH01	863	572.12133
4	BHMBCCTHL01	387	288.35747
5	BHMBRCBRG01	1010	647.36425
6	BHMBRCBRG02	1194	564.11636
7	BHMBRCBRG03	849	243.39123
8	BHMBRTARC01	496	385.21591
9	BHMEURBRD01	470	302.49314
10	BHMEURBRD02	220	136.55408
11	BHMMBMMBX01	687	477.30183
12	BHMNCPHST01	1200	557.68674
13	BHMNCPLDH01	720	505.19659
14	BHMNCPNHS01	500	355.48748
15	BHMNCPNST01	485	285.93826
16	BHMNCPPLS01	450	86.66460
17	BHMNCPRAN01	600	387.27150
18	Broad Street	690	436.15930
19	Bull Ring	3053	1454.86762
20	NIA Car Parks	1268	207.26163
21	NIA North	480	35.42593
22	NIA South	788	196.35714

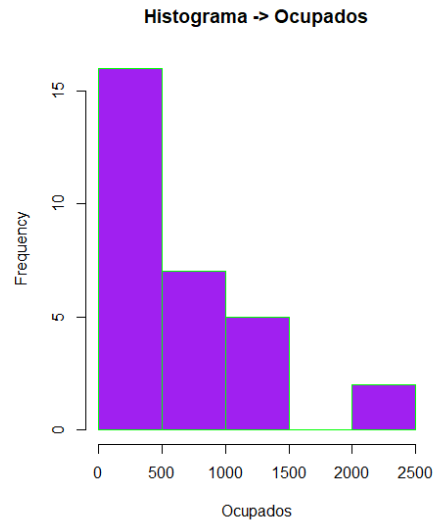
```
> colnames(archivo) <- c("Id" , "Capacity" , "Occupancy")
> View(archivo)
```

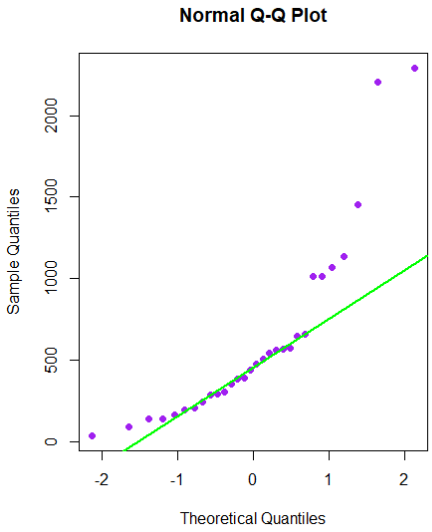
	Id	Capacity	Occupancy
1	BHMBCCMKT01	577	162.02973
2	BHMBCCPST01	317	136.03683
3	BHMBCCSNH01	863	572.12133
4	BHMBCCTHL01	387	288.35747
5	BHMBRCBRG01	1010	647.36425
6	BHMBRCBRG02	1194	564.11636
7	BHMBRCBRG03	849	243.39123
8	BHMBRTARC01	496	385.21591
9	BHMEURBRD01	470	302.49314
10	BHMEURBRD02	220	136.55408
11	BHMMBMMBX01	687	477.30183
12	BHMNCPHST01	1200	557.68674
13	BHMNCPLDH01	720	505.19659
14	BHMNCPNHS01	500	355.48748
15	BHMNCPNST01	485	285.93826
16	BHMNCPPLS01	450	86.66460
17	BHMNCPRAN01	600	387.27150
18	Broad Street	690	436.15930
19	Bull Ring	3053	1454.86762
20	NIA Car Parks	1268	207.26163
21	NIA North	480	35.42593
22	NIA South	788	196.35714

```
> colnames(archivo) <- c("Id" , "Capacity" , "Occupancy")
> View(archivo)
> #Hacemos las pruebas de normalidad para poder hacer la graficacion
> #Capacidad y Ocupados
> hist(archivo$Capacity, col = "purple", xlab = "Capacidad", border = "green", main = "Histograma -> Capacidad")
> qqnorm(archivo$Capacity, pch=19, col = "purple")
> qqline(archivo$Capacity, col = "green", lwd=2)
> hist(archivo$Occupancy,col="purple", xlab = "Ocupados", border = "green", main = "Histograma -> Ocupados")
> qqnorm(archivo$Occupancy, pch=19, col="purple")
> qqline(archivo$Occupancy, col="green", lwd=2)
```



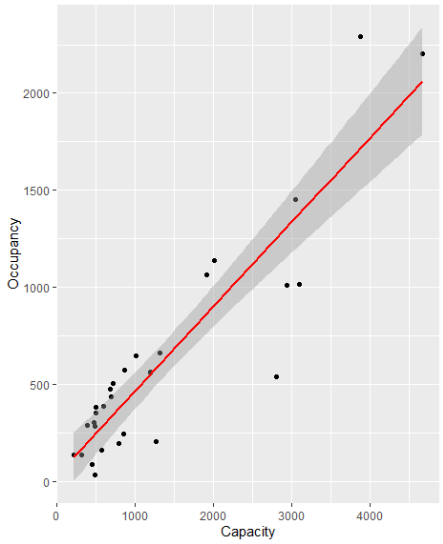




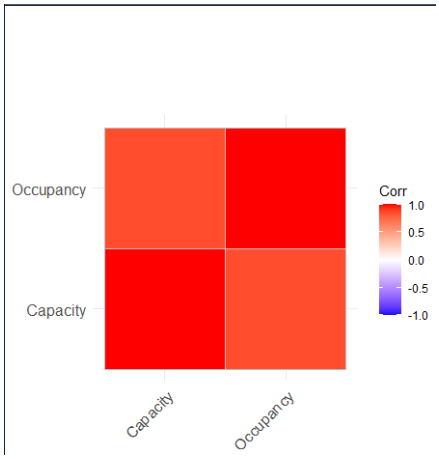


```
> lillie.test(archivo$Capacity) #Capacidad

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  archivo$Capacity
D = 0.23668, p-value = 0.0001656

> lillie.test(archivo$Occupancy) #Ocupabilidad

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  archivo$Occupancy
D = 0.23162, p-value = 0.0002619

> shapiro.test(archivo$Capacity) #Capacidad

        Shapiro-Wilk normality test

data:  archivo$Capacity
W = 0.7952, p-value = 5.366e-05

> shapiro.test(archivo$Occupancy) #Ocupabilidad

        Shapiro-Wilk normality test

data:  archivo$Occupancy
W = 0.79326, p-value = 4.954e-05
```

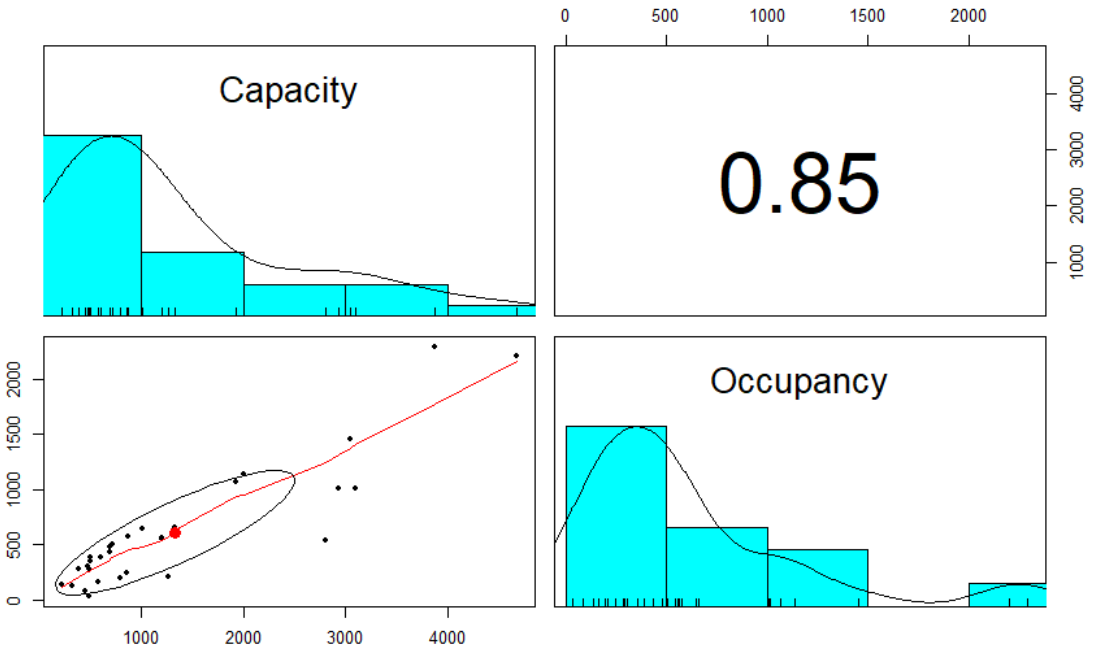
```
> plot(archivo$Capacity, archivo$Occupancy)
> disp = ggplot (archivo, aes(x=Capacity, y=Occupancy))
> disp + geom_point()
> fmd <- lm(archivo$Occupancy ~ archivo$Capacity, archivo)
> summary(fmd)

Call:
lm(formula = archivo$Occupancy ~ archivo$Capacity, data = archivo)

Residuals:
    Min      1Q  Median      3Q     Max
-709.86 -137.47   59.17  140.70  573.92

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      33.98358   67.15010   0.506    0.617
archivo$Capacity  0.43381    0.03809  11.390 5.05e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 241 on 28 degrees of freedom
Multiple R-squared:  0.8225,    Adjusted R-squared:  0.8162
F-statistic: 129.7 on 1 and 28 DF,  p-value: 5.045e-12

> disp + geom_point() + geom_smooth(method = "lm", colour = "red")
`geom_smooth()` using formula 'y ~ x'
```

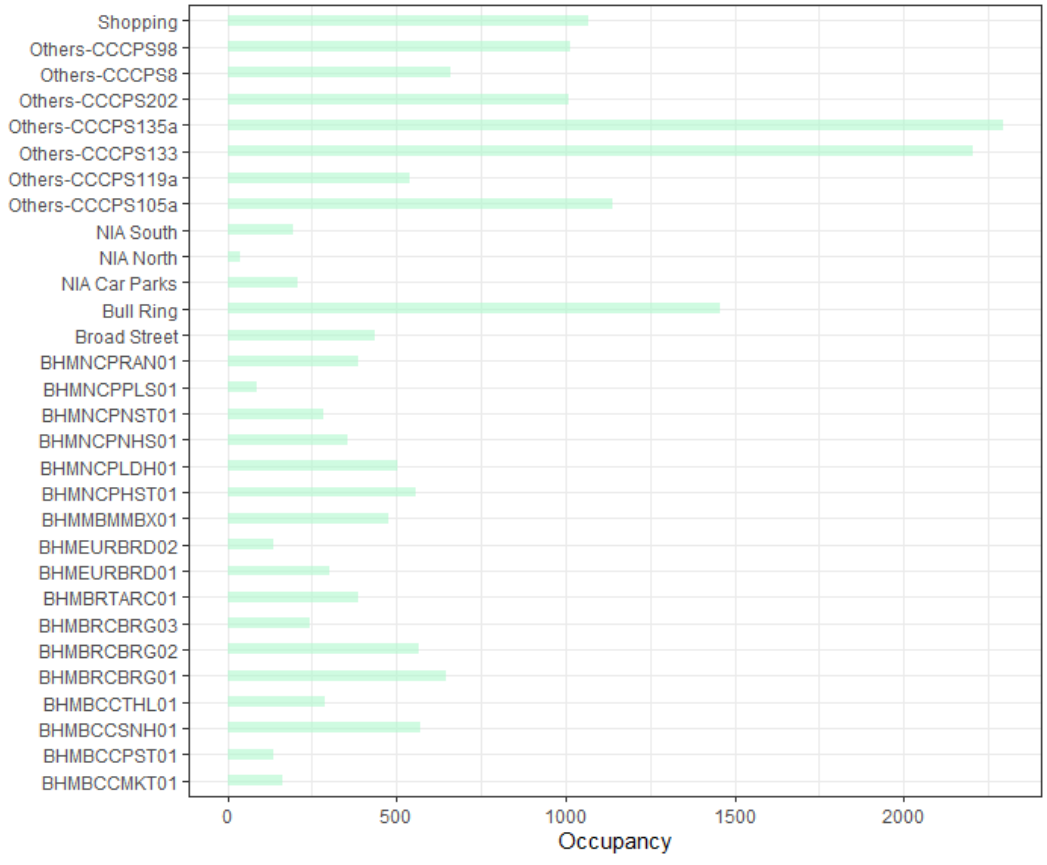



```
> cor(archivo[2:3], method = "spearman")
           Capacity Occupancy
Capacity  1.0000000 0.8478309
Occupancy 0.8478309 1.0000000
> ggcorrplot(round(cor(archivo[2:3], method = "spearman"), 2))
```




```
> pairs.panels(archivo[2:3], method = "spearman")
> archivo %>%
+ ggplot(aes(x = Id, y = Occupancy)) + geom_bar(stat = "identity", fill= "#AFF8CD", alpha=.6, width= .4) + coord_flip() + xlab("") + theme_bw()
>
> archivo %>%
+ ggplot(aes(x = Id, y = Occupancy)) + geom_bar(stat = "identity", fill= "#AFF8CD", alpha=.6, width= .4) + coord_flip() + xlab("") + theme_bw()
```


Capacity
0.85
Occupancy
0
500
1000
1500
2000
1000
2000
3000
4000


Shopping
Others-CCCPS98
Others-CCCPS8
Others-CCCPS202
Others-CCCPS135a
Others-CCCPS133
Others-CCCPS119a
Others-CCCPS105a
NIA South
NIA North
NIA Car Parks
Bull Ring
Broad Street
BHMNCPRAN01
BHMNCPPLS01
BHMNCPNST01
BHMNCPNHS01
BHMNCPLDH01
BHMNCPHST01
BHMMBMMBX01
BHMEURBRD02
BHMEURBRD01
BHMBRTARC01
BHMBRCBRG03
BHMBRCBRG02
BHMBRCBRG01
BHMBCCTHL01
BHMBCCSNH01
BHMBCCPST01
BHMBCCMKT01
0
500
1000
1500
2000
Occupancy

Conclusiones:

En esta práctica como vimos, importamos datos desde una descarga directa de la web, verificamos la normalidad para las variables de los datos en este caso la gráfica y función, obtuvimos el coeficiente de correlación, la matriz para poder verificar la ocupabilidad del estacionamiento, así como también la capacidad del mismo, como vimos en las gráficas tienen ocupado distintas áreas del mismo, fue una práctica complicada pues la mayoría de mis compañeros (incluyéndome) no supimos como abordar el script, sin embargo, se concluyo de manera exitosa.