Universidad del Desarrollo
Universidad de Excelencia

Taller 1

Taller de Inteligencia de Negocios

Docente:

Mauricio Herrera

Integrantes:

Jan Frese

Jorge Ramírez

Problema 1

Para la solución de las preguntas correspondientes al Problema 1, se utilizará la base de datos “PSU_Postulantes.csv”.

Antes de responder todas las preguntas, se instalarán las librerías necesarias, al igual que la localización de la base de datos empleada junto con su ruta de acceso a través del siguiente código:

```
setwd("C:/Users/jorge/OneDrive/Documentos/TIN/")

getwd()

rm(list=ls())

graphics.off()

PSU <- read.csv("PSU_Postulantes.csv",header = T)

View(PSU)

nrow(PSU)

#install.packages("readxl")

#install.packages("dplyr")

#install.packages("httr")

#install.packages("lubridate")

#install.packages("ggplot2")

#install.packages("RColorBrewer")

#install.packages("plotly")

library(readxl)

library(dplyr)

library(httr)

library(lubridate)

library(ggplot2)

library(RColorBrewer)

library(plotly)
```

Pregunta 1:

Se realizará la limpieza de datos en la categoría de “Sexo”. Agrupando así solo dos grupos: “Hombre” y “Mujer”. Además, se realizará dicha limpieza a las variables “psu_mat” y “psu_leng” correspondiente a los puntajes PSU de matemáticas y lenguaje.

El código se realizó de la siguiente manera:

```
#Ejercicio 1
### limpieza datos sexo ###
unique(PSU$Sexo)
PSU$Sexo[PSU$Sexo == 'm'] = 'Mujer'
PSU$Sexo[PSU$Sexo == 'M'] = 'Mujer'
PSU$Sexo[PSU$Sexo == 'mujer'] = 'Mujer'
PSU$Sexo[PSU$Sexo == 'h'] = 'Hombre'
PSU$Sexo[PSU$Sexo == 'H'] = 'Hombre'
PSU$Sexo[PSU$Sexo == 'hombre'] = 'Hombre'
unique(PSU$Sexo)
PSU$Sexo <- factor(PSU$Sexo)
levels(PSU$Sexo)
which(is.na(PSU$Sexo))
PSU <- PSU[-137,]
View(PSU)
### limpieza datos psu_leng ###
which(is.na(PSU$psu_leng))
# Como no existen NA en los datos de esta variable, no hay necesidad de realizar una limpieza
```

En primer lugar, se procede a la limpieza por género, se agrupan todos los diminutivos de “Sexo” en dos grupos, los cuales fueron mencionados anteriormente. Como solamente se registró un valor NA, correspondiente al dato 137, se procede a eliminar.

Tras esto, como segunda medida, se procede a hacer la limpieza de los datos de la PSU de matemáticas, de la siguiente manera:

```
### limpieza datos psu_mat ###
which(is.na(PSU$psu_mat))
mean(PSU$psu_mat, na.rm = TRUE)
mate_hom <- mean(PSU$psu_mat[which(PSU$Sexo =='Hombre')],na.rm=TRUE)
round(mate_hom)
mate_muj <- mean(PSU$psu_mat[which(PSU$Sexo == "Mujer")],na.rm = TRUE)
round(mate_muj)
PSU$psu_mat[which(PSU$Sexo == "Hombre" & is.na(PSU$psu_mat))] <- round(mate_hom)
PSU$psu_mat[which(PSU$Sexo == "Mujer" & is.na(PSU$psu_mat))] <- round(mate_muj)
which(is.na(PSU$psu_mat))
View(PSU)
```

Durante este paso, se generaban 2 situaciones:

-Primero había una gran cantidad de valores no disponibles (NA), por lo que se asignaron los promedios de “psu_mat”, tanto de hombres como mujeres a estas casillas según corresponda.

-Segundo, los puntajes PSU no contienen decimales, por lo que se redondea el valor obtenido al entero más cercano. Los promedios resultantes para cada género fueron 617 para los hombres y 615 para las mujeres.

Finalmente, para los puntajes de la PSU de lenguaje, no fue necesario hacer limpieza, ya que no existían valores no disponibles (NA) en esta categoría.

```
### limpieza datos psu_leng ###
which(is.na(PSU$psu_leng))
# Como no existen NA en los datos de esta variable, no hay necesidad de realizar una limpieza
```

Pregunta 2:

Para este problema, se agrupan “Hombre” y “Mujer” en dos “tablas” distintas, para así obtener la cantidad de postulantes en cada categoría (118 postulantes de cada género), luego se realiza un gráfico de barras con ambos valores para cada género (fig1).

```
# Ejercicio 2
unique(PSU$Sexo)
PSU <- arrange(PSU, desc(Sexo))
mujeres <- PSU%>%
        filter(Sexo == "Mujer")
nrow(mujeres)
hombres <- PSU%>%
         filter(Sexo == "Hombre")
nrow(hombres)
View(PSU)
ggplot(PSU, aes(Sexo, fill = Sexo)) + geom_bar(position = 'dodge')
```

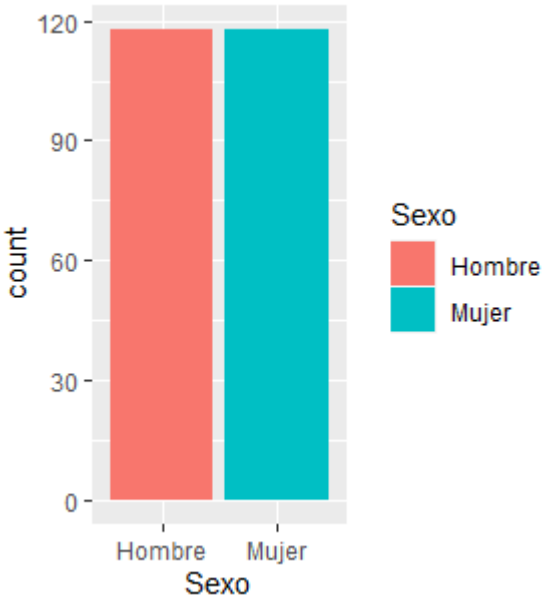

fig1.

Pregunta 3:

Para calcular los promedios y desviaciones estándar de cada género para ambas PSU (matemática y lenguaje), se ejecutan los siguientes códigos, redondeando los valores de cada PSU debido a que no contemplan decimales los puntajes de la prueba.

```
# Ejercicio 3

### promedio y desviacion estandar de psu_mat ###
round(mean(hombres$psu_mat))
round(sd(hombres$psu_mat))
round(mean(mujeres$psu_mat))
round(sd(mujeres$psu_mat))

### promedio y desviacion estandar de psu_leng ###
round(mean(hombres$psu_leng))
round(sd(hombres$psu_leng))
round(mean(mujeres$psu_leng))
round(sd(mujeres$psu_leng))
```

Obteniendo de esto, los siguientes resultados:

- Promedios y Desviación Estándar **Hombres**:
  -psu_mat: 617 ; desv_est: 59
  -psu_leng: 616 ; desv_est: 53

- Promedios y Desviación Estándar **Mujeres**:
  -psu_mat: 615 ; desv_est: 51
  -psu_leng: 611 ; desv_est: 54

Pregunta 4:

Se formula una prueba t-student, con los siguientes valores:

- ‘a’: Valores “psu_mat” Hombres
- ‘b’: Valores “psu_mat” Mujeres
- ‘c’: Valores “psu_leng” Hombres
- ‘d’: Valores “psu_leng” Mujeres

```
# Ejercicio 4

a <- hombres$psu_mat

b <- mujeres$psu_mat

t.test(a,b) # no hay diferencias significativas (pvalor>0.05)

c <- hombres$psu_leng

d <- mujeres$psu_leng

t.test(c,d) # no hay diferencias significativas (pvalor>0.05)
```

Para la PSU de matemáticas, el p-valor de la prueba arrojó un valor de 0.7905, en base a esto, como el p-valor es mayor que 0.05 no hay diferencias significativas entre los promedios de Hombres y Mujeres

Para la PSU de lenguaje, el p-valor de la prueba arrojó un valor de 0.4516. Dado que su p-valor es mayor a 0.05, no existen diferencias significativas en los puntajes de cada género.

Como conclusión, no hay diferencias significativas en los puntajes de la PSU entre los hombres y mujeres que realizaron la prueba.

Pregunta 5:

Se realiza el diagrama de cajas y bigotes (boxplot) para la PSU de matemáticas diferenciado en género (fig2).

ggplot(data = PSU) + geom_boxplot(aes(x = Sexo, y = psu_mat, fill = Sexo))

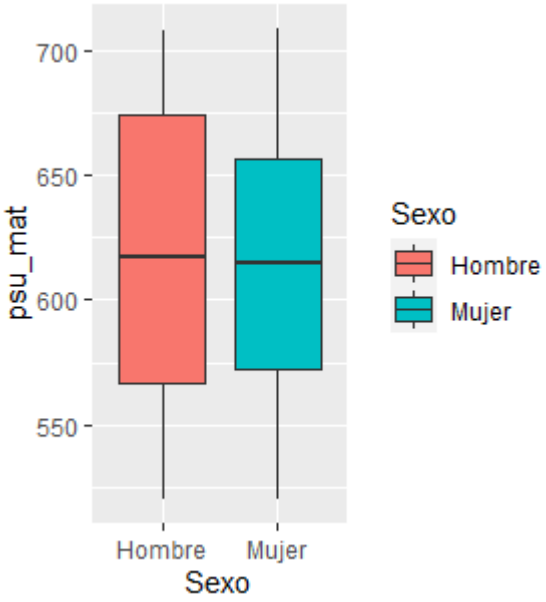

fig2.

Como el diagrama está separado por cuantiles (4 para ser más específicos), notamos que la mediana, tanto para Hombres como Mujeres, no varía mucho.

Por otra parte, el valor mínimo está por debajo de los 550 puntos para ambos sexos, mientras que el valor máximo en ambos supera los 700 puntos.

Las Mujeres tuvieron una mayor densidad de postulantes con puntajes cercanos a 550, como también a los 700, mientras que la mayor de cantidad de postulantes Hombres, se encontraba en los cuantiles por debajo y encima de la mediana (alrededor de 600 y 650 puntos).

Que básicamente reafirma lo obtenido de la prueba t-student, donde no se pueden apreciar diferencias significativas entre los resultados de ambos géneros.

Problema 2

Para la solución de las preguntas correspondientes al Problema 2, se utilizará la base de datos “pizza_delivery.csv”.

Antes de responder todas las preguntas, se instalan las librerías necesarias, junto con la localización de la base de datos empleada:

```
setwd("C:/Users/jorge/OneDrive/Documentos/TIN/")
getwd()
rm(list=ls())
graphics.off()
  Pizza <- read.csv("pizza_delivery.csv",header = T)
View(Pizza)
#install.packages("readxl")
#install.packages("dplyr")
#install.packages("httr")
#install.packages("lubridate")
#install.packages("ggplot2")
#install.packages("RColorBrewer")
#install.packages("plotly")
library(readxl)
library(dplyr)
library(httr)
library(lubridate)
library(ggplot2)
library(RColorBrewer)
library(plotly)
```

Pregunta 1:

Para cambiar el tipo de moneda de dólares (predeterminado de la base de datos) a pesos chilenos (tomando el valor actual de cambio de 791,8 pesos el dólar), se ejecuta el siguiente código:

```
#Ejercicio 1

Pizza$bill <- round(791.8*Pizza$bill)
View(Pizza)
```

Como el valor de los pesos chilenos no contempla decimales, se optó por redondear al entero más cercano esta transformación.

Pregunta 2:

Para esta pregunta, se procede a revisar si existen valores nulos dentro de la base en las categorías “Bill”, “Branch” y “Pizzas”. Como no existen valores nulos, se procede al siguiente paso.

Para términos prácticos, se crearon tres tablas distintas correspondientes a los distintos tipos de “Branch” registrados en la base de datos (“East”, “West” y “Centre”). Luego se calculó el precio a pagar, junto con la cantidad de pizzas pedidas según cada “Branch”.

```
#Ejercicio 2


which(is.na(Pizza$bill))
which(is.na(Pizza$branch))
which(is.na(Pizza$pizzas))


unique(Pizza$branch)
East <- Pizza %>%
        filter(branch == "East")
West <- Pizza %>%
        filter(branch == "West")
Centre <- Pizza %>%
        filter(branch == "Centre")


round(mean(East$bill))
round(mean(West$bill))
round(mean(Centre$bill))


round(mean(East$pizzas))
round(mean(West$pizzas))
round(mean(Centre$pizzas))
```

Como no existen valores decimales en el valor de los pesos chilenos y tampoco en la cantidad de pizzas pedidas (uno no puede pedir 2,5 pizzas por mencionar un ejemplo), se optó por redondear estos valores.

Del código se obtiene:

Promedios de Pago por sucursal:

-East: $ 29.092

-West: $ 35.009

-Centre: $ 37.298

Promedios de pizzas ordenadas por sucursal:

-East: 2 pizzas.

-West: 3 pizzas.

-Centre: 3 pizzas.

Pregunta 3:

Se construye un diagrama de cajas y bigotes para los tiempos de despacho, separados por zonas (“Branch”) (fig3):

 # Ejercicio 3

ggplot(data = Pizza) + geom_boxplot(aes(x = branch, y = time, fill = branch))

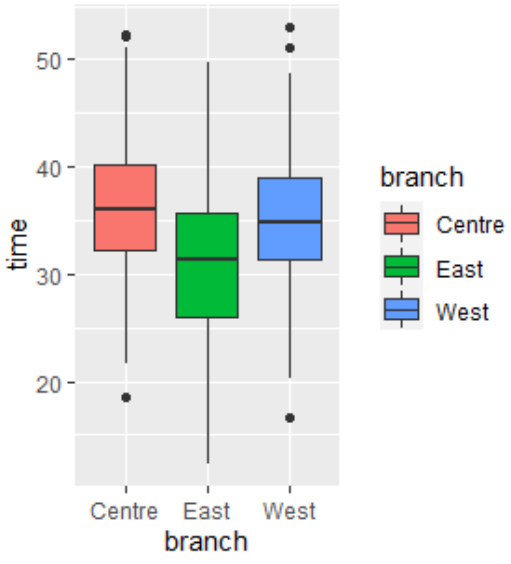

fig3.

Del diagrama, podemos observar que la mediana de la sucursal de “East” es la menor de las tres, es decir, tiene menor tiempo de despacho.

Pregunta 4:

Para comparar el desempeño de los tiempos de entrega de pizzas según operador, primero, se realiza un diagrama de cajas y bigotes (fig4). En segundo lugar, se filtra la base de datos para ambas operadoras (Melissa y Laura), con el fin de realizar una prueba t-student para analizar los resultados, con la variable ‘j’ correspondiente a Melissa y la variable ‘f’ correspondiente a Laura. Finalmente, se calculan los valores de los cuantiles para los datos registrados en ambas operadoras:

```
# Ejercicio 4

# primera comparacion
unique(Pizza$operator)
ggplot(data = Pizza) + geom_boxplot(aes(x = operator, y = time, fill = operator))
```

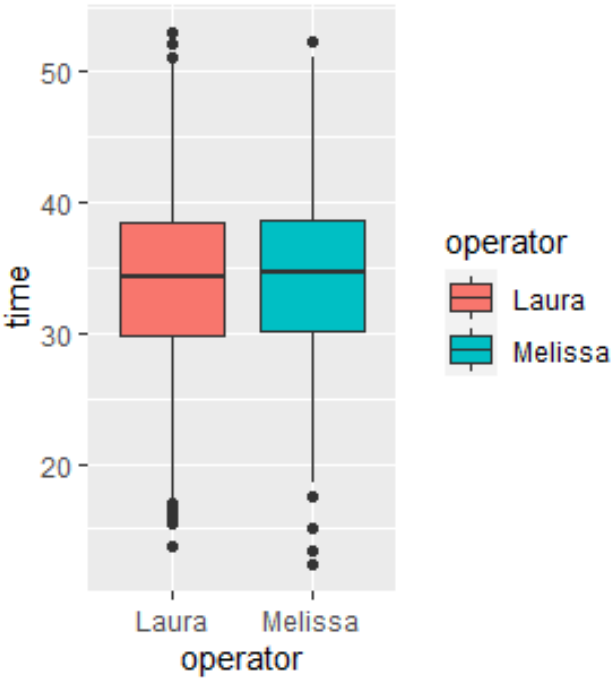

fig4.

```
# segunda comparacion
Melissa <- Pizza %>%
  filter(operator == "Melissa")
Laura <- Pizza %>%
  filter(operator == "Laura")

j <- Melissa$time
f <- Laura$time
t.test(j,f)

# tercera comparacion
quantile(Melissa$time, 0.25)
quantile(Melissa$time, 0.5)
quantile(Melissa$time, 0.75)
quantile(Melissa$time, 1)
quantile(Laura$time, 0.25)
quantile(Laura$time, 0.5)
quantile(Laura$time, 0.75)
quantile(Laura$time, 1)
```

Respecto a la primera comparación, se puede afirmar que no hay grandes diferencias entre la mediana de los tiempos de entrega entre operadoras.También se puede observar que Laura presenta una mayor cantidad de tiempos atípicos/excepcionales de entrega. Finalmente, se puede concluir que, tanto Laura como Melissa han participado en pedidos de Pizza de manera similar.

En la segunda comparación, se puede observar que el p-valor del test es mayor a 0,05 y, por lo tanto, no existen diferencias significativas entre los desempeños de tiempo de entrega entre las operadoras.

En la última comparación, los valores de los cuantiles de cada operadora son los siguientes:

Melissa

30,20963 → primer cuantil (25%)

34,6415 → segundo cuantil (50%)

38,63497 → tercer cuantil (75%)

52,31672 → cuarto cuantil (100%)

Laura

29,87681 → primer cuantil (25%)

34,30399 → segundo cuantil (50%)

38,389 → tercer cuantil (75%)

53,09626 → cuarto cuantil (100%)

Como se puede apreciar, los valores de cada cuantil son bastante similares entre cada operadora, lo cual reafirma el diagrama de cajas y bigotes y el test t-student, los cuales concluyen que no hay diferencias significativas entre cada operadora según el desempeño em los tiempos de entrega de la pizza.