Visualització de dades (Enginyeria de Dades – EE - UAB)
Examen Segon Parcial – 3 Juny 2022
RESPOSTES MODEL A

Nom i Cognom:_____

NIU:_____ Grup de Matrícula:_____

Només es permet l'ús d'internet per l'accés al campus virtual en el moment de descarregar el full d'enunciats y d'entregar l'examen.

Sólo se permite el uso de internet para el acceso al campus virtual en el momento de descargar la hoja de enunciados y de entregar el examen.

PARTE 1 (2,5 pt)

Dataset: GDP-LifeExp-Population.csv

**** Lee los tres enunciados de esta parte antes de empezar ****

1.1. (0,5 pt) Filtra las observaciones que tengan valores NAN o null.

Identifica outliers (valores extremos que distorsionen la muestra) en la variable *Population* y filtra las observaciones que no sean comparables al resto.

RESPOSTA:

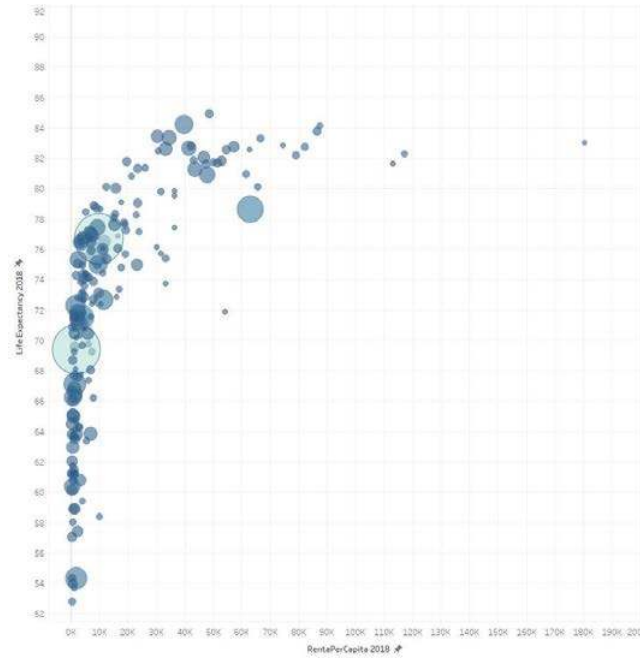
Nulls: Andorra, Eritrea, Gibraltar, N. Korea, Gibraltar, F. Polynesia, S. Sudan, Venezuela
Outliers: World, OCDE members

1.2. (1 pt) Haz un **bubble chart** con las variables *Life expectancy 2018*, *PIB per capita USD 2018* y *Population 2018*, donde *Population* sea el tamaño de los círculos.

Los círculos deben estar coloreados por la variable de tu elección que creas que más aporta a la visualización.

Añade las etiquetas, leyendas, y transparencias necesarias para que la gráfica sea legible y comprensible.

RESPOSTA:



1.3. (0,5 pt) ¿Por qué un *bubblechar*t es una visualización apropiada en este caso? Razona brevemente según el tipo y el número de atributos.

RESPOSTA:

Tenemos tres atributos cuantitativos y un número alto de observaciones (200). Un bubble chart sirve para graficar la relación entre tres atributos cuantitativos, dos en la posición X e Y, y un tercero en el tamaño de los círculos. Además es un tipo de gráfica que escala bien a cientos de observaciones, a diferencia de otras, como las gráficas de barras.

1.4. (0,5 pt) ¿Qué variable usas para colorear los datos y qué escala de color utilizas? Razona brevemente porqué la variable y porqué la escala.

RESPOSTA:

Cualquiera de las 3 cuantitativas serviría, aunque la más apropiada será *Population* para reforzar el canal más débil, que es el tamaño de los círculos. La opción incorrecta sería usar el país.

PART 2 (3,5 pt)

Dataset: `starwars` **de R per l'exercici 2.1**

Dataset: `statsNBA2008.csv`. **Aquest dataset té 21 atributs de 50 jugadors de la NBA. Recull les estadístiques de la NBA del 2008. I estudiarem les components principals d'alguns dels atributs en l'exercici 2.2.**

2.1 (2,5 pt) Volem representar els pesos dels personatges de starwars (dades incloses en R) en un mapa d'arbre (*treemap*) que us permeti contestar les preguntes de l'apartat d. El mapa d'arbre (*treemap*) podeu fer-lo en R o en qualsevol altre llenguatge, utilitzant les llibreries que us semblin convenients.

a) Feu el següent data massage: Descarteu les files del dataframe on la variable massa ('mass') o la variable especie ('species') continguin un valor Nan i reemplaceu els Nans de la columna gènere per 'none'. (0,5 pt)

b) Feu un mapa d'arbre. Ara bé, heu d'explicar-ho bé: tot narrant com és un *treemap* en general - redacteu els passos que heu de fer per construir la visualització triada. Poseu llegenda i títol i mostreu el gràfic (1,25 pt)

c) Canvieu la paleta de color. Si treballeu en R podeu posar manualment la paleta de color `c('#DCB0F2','#87C55F','#9EB9F3')` o posar una altra paleta adient que us agradi. En tots els casos, si treballeu amb R, en un altre llenguatge o amb Tableau, presenteu en l'apartat anterior el gràfic que us ha sortit amb el color per defecte i aquí amb una paleta diferent (0,25 pt)

d) Un cop tingueu la visualització, contesteu: Quin és el personatge amb més massa? De quina espècie i gènere és? Digueu el nom dels dos robots ('droids') amb menys pes, quin gènere són? (0,25 pt)

e) Expliqueu en quin cas són òptims el mosaic, el treemap i un paral·lel set (sense necessitat de fer cap visualització) i quines diferències principals hi ha entre ells. (0,25 pt)

RESPOSTA:

a)

```
> library ("tidyverse")
> data = starwars %>% drop_na(mass) %>% drop_na(species) %>% replace_na(list(gender = "none"))
```

b) Com vam veure a classe, un mapa d'arbre és un dibuix rectangular dividit en caselles, i cada casella representa una sola observació. Vam veure que era una bona manera de mostrar dades jeràrquiques mitjançant rectangles imbricats. I l'àrea relativa de cada casella expressava una variable contínua. També vam veure que era òptim quan hi ha com a màxim dues variables d'agrupació, per tant no en definirem més.

Per tant, una possibilitat amb aquest dataset seria:

- *Per definir el color i actuar doncs com un 'grup pare' utilitzaríem la variable "gender" (Agafem aquesta com grup de color, ja que sol té 3 nivells ("masculine", "feminine" i "none" i serà fàcil entendre la llegenda, que si posem "species" que en té més).*
- *Com a 'subgrup' utilitzaríem la especie (species, que té més nivells que gènere).*
- *Com a variable que descrigui l'àrea de les caselles triarem "mass". Aquesta tria ve donada perquè ens estan preguntant sobre la massa dels personatges. I de totes les variables que ens pregunten, és l'única numèrica.*
- *I com a 'label'/nivell, escollim el nom ("name").*

Si ho fem amb R, el primer que hem de fer és carregar les llibreries necessàries. La llibreria especifica aquí és treemapify (abans necessitarem tenir instal·lat el paquet com vam veure al seminari corresponent). També haurem de fer ús de geom_treemap.

Per fer un treemap basic doncs:

```
> library ("treemapify")
```

d) El personatge amb més pes és en Jabba Desilijic Tiure, masculí i de l'espècie hut (Hutt)

Els robots amb menys massa són el R2-D2 i el R5-D4, són masculins

e) *Quan s'especifiquen les proporcions d'acord amb múltiples variables d'agrupació, les representacions de mosaics, els mapes d'arbres o els conjunts paral·lels (paral·lel sets) són aproximacions de visualització útils. Ara bé, els mosaics assumeixen que tots els nivells d'una variable d'agrupament es poden combinar amb tots els nivells d'una altra variable d'agrupació, mentre que els mapes d'arbres no fan aquesta suposició.*

Els mapes d'arbres com el que hem fet, funcionen bé fins i tot si les subdivisions d'un grup són completament diferents de les subdivisions d'un altre.

Finalment, els conjunts paral·lels funcionen millor que les trames de mosaic o els mapes d'arbres quan hi ha més de dues variables d'agrupació.

2.2. (1 pt) Voleu saber quines són les components principals de les variables numerades a sota del dataset *statsNBA2008.csv.* Podeu fer-ho en qualsevol llenguatge, però expliqueu perquè heu escollit la visualització que heu escollit, els passos que heu seguit per fer-la, mostreu-la, i extreieu alguna/es conclusió/ns del vostre gràfic.

- **PTS (percentatge de punts de l'equip, punts per joc)**
- **FGM (cistelles de camp realitzats, percentatge de cistelles de camp)**
- **FGA (tirs de camp intentats, número d'intencions de cistelles de camp)**
- **DRB (rebots defensius)**
- **ORB (rebots ofensius)**
- **TRB (rebots totals)**

RESPOSTA:

Llegim dades i carreguem llibreries:

```
> NBA=read_csv('statsNBA2008.csv')
```

Primer ens preparem el dataframe que ens interessa, vam veure al seminari 5 les tibles de R i com utilitzar-les. Una manera que es podria fer servir, seria

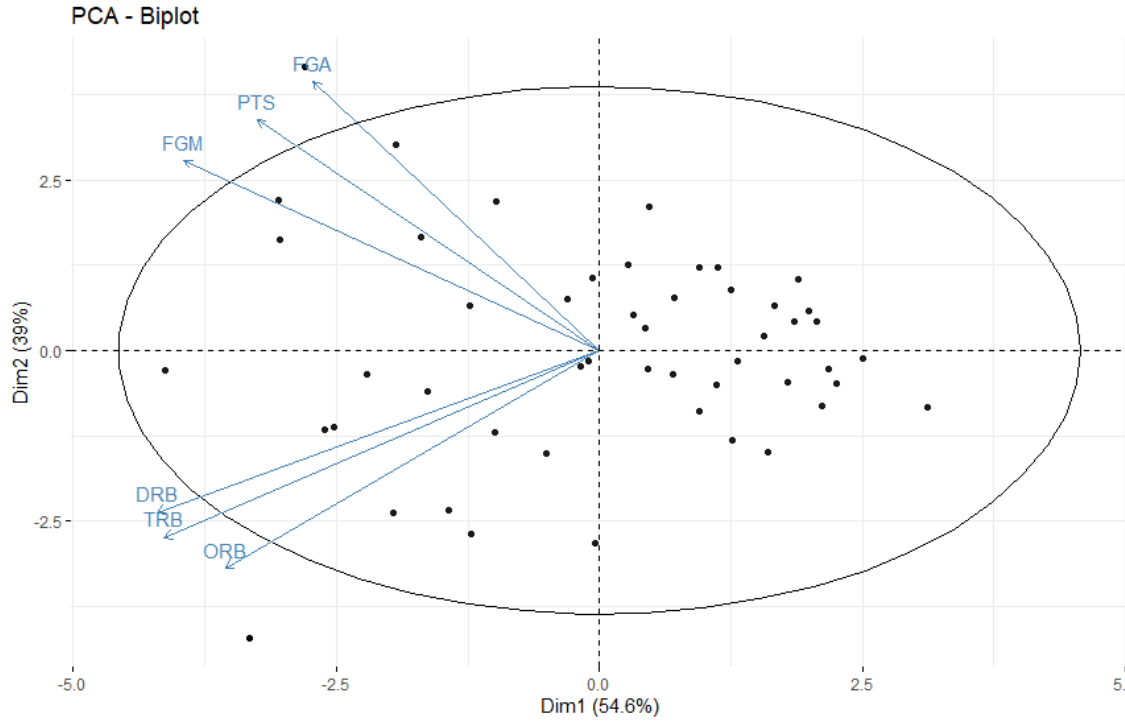
```
> NBA_tb <- as_tibble(NBA)
> NBA_interes<-NBA_tb[,c('PTS','FGM','FGA','DRB','ORB','TRB')]
```

També ho podríem fer utilitzant select de dplyr com havíem vist en classes anteriors.
Assignem a una nova variable anomenada NBA_pca la matriu resultant d'aplicar la funció *prcomp() vista en la primera part de la classe d'avui. Centrem les variables a zero utilitzant l'argument 'center=TRUE', i les escalem com al seminari per a que tinguin variança igual a 1, fent us de l'argument 'scale. =TRUE'.*

```
> NBA_pca<-prcomp(NBA_interes, center=TRUE, scale.=TRUE)
```

Podem per exemple mostrar un biplot

```
> fviz_pca_biplot (NBA_pca, geom.ind="point", geom.var = c("arrow", "text"), addEllipses = TRUE, legend.tittle="Groups")
```



- *PC1 correspon al 54.6% de la variança total.*
- *PC2 correspon al 39% de la variança total.*

Per tant, coneixent la posició d'una mostra en relació amb (només) les components PC1 i PC2, podem tenir una visió molt precisa de la seva ubicació en relació amb altres mostres. Això és degut a que només PC1 i PC2 poden explicar el 93.6% de la variança.

PART 3 (4 pt)

Dataset: `Suicide rates 1985-2016.csv`

Agafarem el dataset de noms del nombre de suïcidis en diferents països entreels anys 1985 i 2016 segmentat per diferents variables com sexe, edat, generació, població, etc. Utilitzeu les llibreries (plotly, gganimate, shiny, etc.) que creieu convenients i dibuixeu les gràfiques que us facin falta.

RESPOSTA:

```
> library(tidyverse)
```

```
> library(dplyr)
> library(plotly)
> library(shiny)

> getwd()
> setwd("C:/Users/enric/Documents/R")
> SWorld <- read.csv('./Suicide rates 1985-2016.csv')
> str(SWorld)
# Renombrar alguns camps amb caracters erronis (també es pot utilitzar la
comanda rename):
> names(SWorld)[1] <- "country"
> names(SWorld)[10] <- "gdp_for_year"
> names(SWorld)[11] <- "gdp_per_capita"
> str(SWorld)
```

3.1. (1 pt) Mostra el codi i una gràfica de línies de l'evolució del nombre de suïcidis per anys i generació. Digues quines generacions són capdavanteres i en quin període d'anys ho són.

RESPOSTA:

PAS 1: DATA MASSAGING: Seleccionar variables i agrupar per generació i any

```
> SWorldG <- SWorld %>% select(generation, year, suicides_no, population,
gdp_per_capita) %>% group_by(generation, year) %>% summarise(numeroS =
sum(suicides_no), poblacio=sum(population), PIB=mean(gdp_per_capita))
```

O bé...

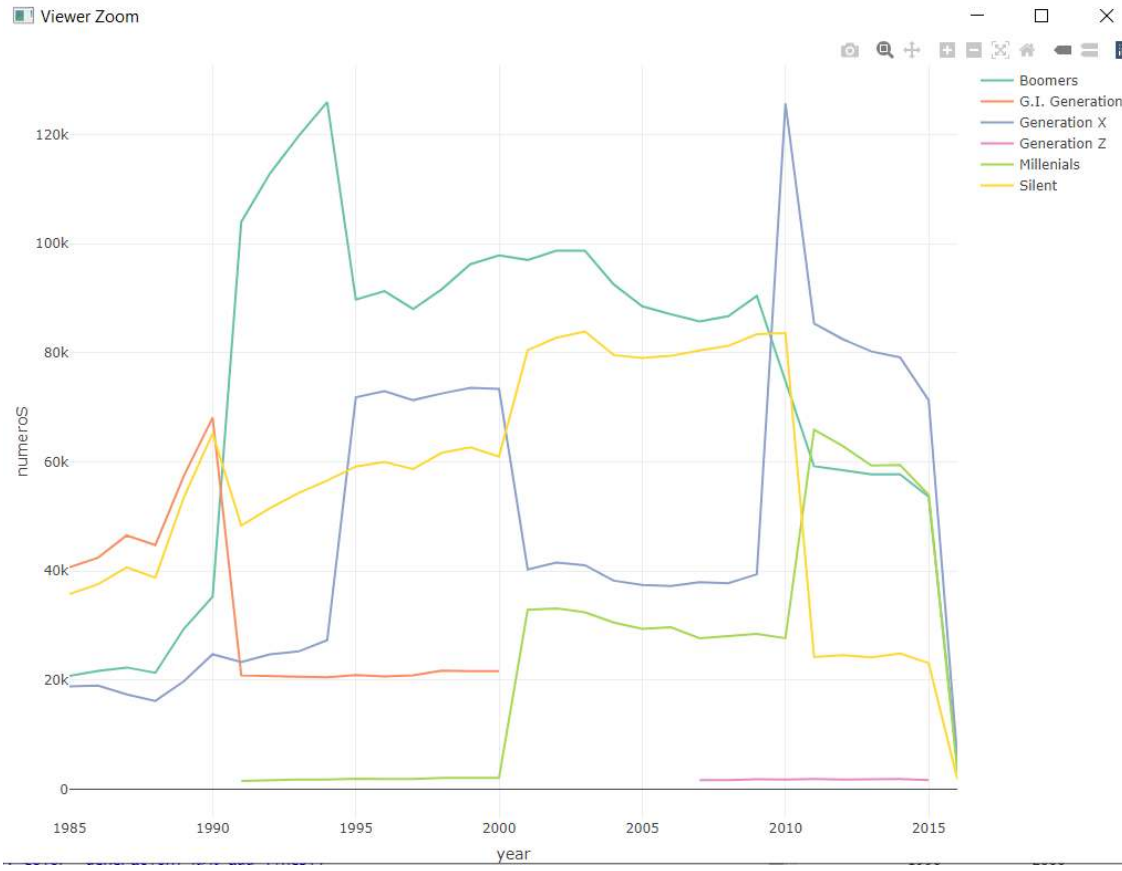
```
> SWorldG <- SWorld %>% group_by(generation, year) %>% summarise(numeroS
= sum(suicides_no), poblacio=sum(population), PIB=mean(gdp_per_capita))
```

```
# PAS 2a: GRAFICA DE LINIES AMB ggplotly
> ggplotWG <- ggplot(SWorldG, aes(x=year, y=numeroS, color=generation)) +
geom_line()
> ggplotly(ggplotWG)

# PAS 2b: GRAFICA DE LINIES AMB plotly
> plot_ly(SWorldG, x=~year, y=~numeroS, color=~generation) %>% add_lines()
>

# PAS 2c: GRÀFICA DE LINIES AMB shiny
ui <- fluidPage(   selectizeInput(  inputId = "generat",
                   label = "Selecciona una Generacio:",
                   choices = unique(SWorldG$generation),
                   selected = c("Silent"),
                   multiple = TRUE
                       ),
      plotlyOutput(outputId = "p")
        )
server <- function(input, output, ...)
   {   output$p <- renderPlotly (
       {plot_ly(SWorldG, x = ~year, y = ~numeroS, color=~generation) %>%
```

```
                    filter(generation %in% input$generat) %>%
                    group_by(generation) %>%
                    add_lines()
        })
    }
shinyApp(ui, server)
```

GENERACIONS CAPDAVANTERES:

- GI GENERATION: 1985-1990
- BOOMERS: 1991-2009
- GENERATION X: 2010-2016



3.2. (0,5 pts) D'aquesta, obteniu la següent informació:

a) Classificació per generacions l'any 2011.

RESPOSTA:

POSICIÓ	GENERACIÓ	Nombre de Suïcidis
1	GENERATION X	85.345
2	MILLENIALS	65.873
3	BOOMERS	59.178
4	SILENT	24.209

5	GENERATION Z	1.879
6		

b) Nombre màxim de suïcidis per generació.

RESPOSTA:

GENERACIÓ	ANY	Nombre de Suïcidis
BOOMERS	1994	125.932
GENERATION X	2012	125.681
GENERATION Z	2014	1.882
G.I. GENERATION	1990	68.118
MILLENIALS	2011	65.873
SILENT	2003	83.902

3.3. (1 pt) Mostra el codi i una gràfica d'un scatter plot 3D sobre el nombre de suïcidis en DONES per a cada 100k, població i PIB per càpita (`gdp_per_capita`).

RESPOSTA:

PAS 1: DATA MASSAGING:Seleccionar variables i agrupar per generació i any

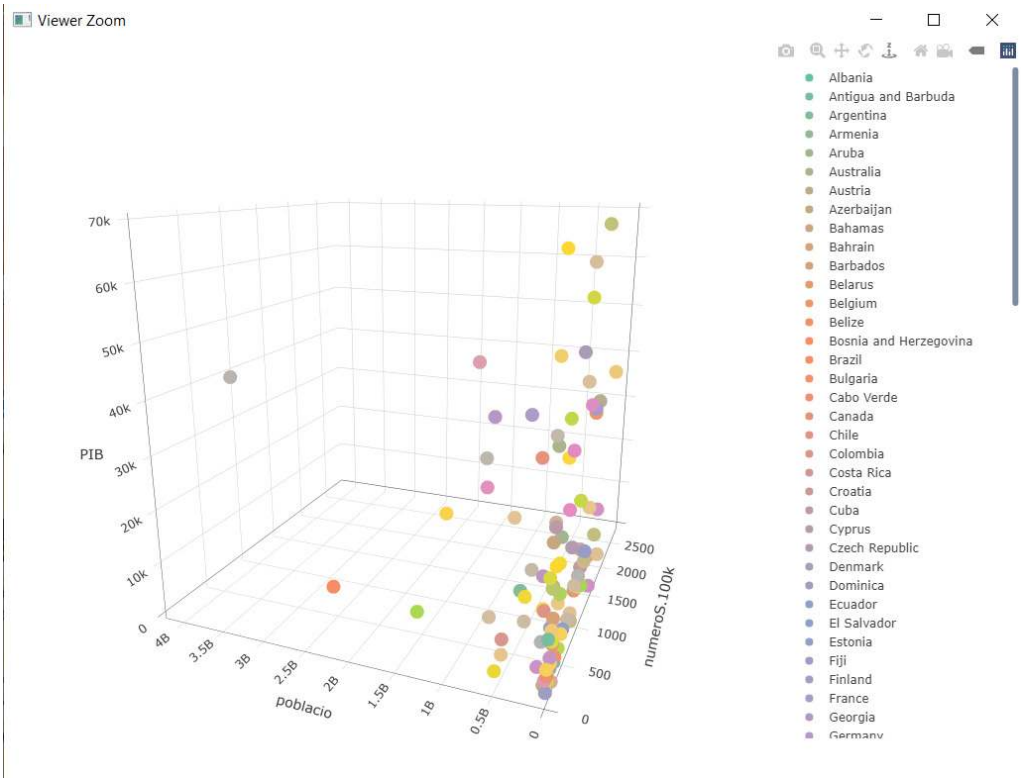
```
> SWorldD <- SWorld %>% filter(sex == "female") %>% select(country, year, suicides.100k.pop, population, gdp_per_capita) %>% group_by(country) %>% summarise(numeroS.100k = sum(suicides.100k.pop), poblacio=sum(population), PIB=mean(gdp_per_capita))
```

O bé...

```
SWorldD <- SWorld %>% filter(sex == "female") %>% group_by(country) %>% summarise(numeroS.100k = sum(suicides.100k.pop), poblacio=sum(population), PIB=mean(gdp_per_capita))
```

PAS 2: GRAFICA DE SCATTER 3D AMB `plotly`

```
> plot_ly(SWorldD, x=~numeroS.100k, y=~poblacio, z=~PIB, color=~country, type='scatter3d', mode='markers')
```



3.4. (0,5 pts): D'aquesta gràfica, obteniu la següent informació:

a) **Dades (país, suicidis.100k.pop, població i PIB per càpita) dels 4 països amb més població. Trobes a faltar algun país?. Màxim 2.**

RESPOSTA:

PAÍS	SUICIDIS.100k.pop	Població	PIB per càpita
USA	869,81	41.113.698	39.269,6
Brazil	394,84	24.654.8B	6.091,484
Russian Federation	1879,37	19.80711B	6.518.815
Japan	2546,84	18.847.87B	36.397,55

b) **Dades (país, suicidis.100k.pop, població i PIB per càpita) del país amb més PIB per càpita.**

RESPOSTA:

PAÍS	SUICIDIS.100k.pop	Població	PIB per càpita
Luxemburg	1455,66	6.577.184	68.798,39

4 (1 pt) Classifica les següents preguntes o mesures que s'utilitzen en tests d'Usabilitat o d'Experiència d'Usuari (UX) en les següents dues categories:

a) **Instrumentals / Usabilitat / Pragmàtiques**

b) **No instrumentals / Hedònics / Emocionals**

I digues en quin tipus de test s'inclou cada pregunta o mesura.

PREGUNTES:

1. ***Complicated 1 2 3 4 5 6 7 Simple***
2. ***The design looks attractive***
3. ***Non-inclusive 1 2 3 4 5 6 7 Inclusive***
4. ***It was easy to find the information I needed***
5. ***I thought the System was easy to use***
6. ***The product is creatively designed***
7. ***Conservative 1 2 3 4 5 6 7 Innovative***
8. ***The product is stylish: Strongly Disagree 1 2 3 4 5 6 Strongly Agree***
9. ***This System has all the functions and capabilities I expect it to have***
10. ***I found the various functions in this System very well integrated***

RESPOSTA:

PREGUNTA	**CATEGORIA**	**TIPUS DE TEST**
1	**a)**	**Attrakdiff, Pragmatic**
2	**b)**	**meCUE, A2, Aesthetic**
3	**b)**	**Attrakdiff, Hedonic**
4	**a)**	**PSSUQ**
5	**a)**	**SUS**
6	**b)**	**meCUE, A1, Aesthetic**
7	**b)**	**Attrakdiff, Hedonic**
8	**b)**	**meCUE, A3, Aesthetic**
9	**a)**	**PSSUQ**
10	**a)**	**SUS**