Carrera de Computación
Fundamentos de Análisis de Datos
Examen Parcial B2 – Práctica

Respuestas

PREGUNTA 1: EDA

Script R

En el cuadro verde pegue el texto de los comandos R usados para lo solicitado en la Pregunta 1

```
library(dplyr)
library(ggplot2)

getwd()
df = read.csv("datasetParcial_2b.csv")

########################
# EXPLORACIÓN
########################

dim(df) # 32581    12
colnames(df)
head(df)
str(df)
summary(df)
class(df)

unique(df$Id)
unique(df$Age)
unique(df$Income)
unique(df$Home)
unique(df$Emp_length)
unique(df$Intent)
unique(df$Amount)
unique(df$Rate)
unique(df$Status)
unique(df$Percent_income)
unique(df$Default)
unique(df$Cred_length)

# Clasificacion de variables
# Numericas
# -> Id, Age, Icome, Emp_length, Amount, Rate, Percent_income, Cred_length
# Categoricas
# -> Home, Intent, Status, Default
# Bimodales

########################
# ANALISIS EDA
########################

# Valores nulos
sapply(df, function(x) sum(is.na(x) | !nzchar(x))) # Cantidad de valores faltantes y
nulos

# Inspeccionar observaciones con valores faltantes
df[is.na(df$Emp_length) | !nzchar(df$Emp_length),]
df[is.na(df$Rate) | !nzchar(df$Rate),]
```

```
df_clean = df %>% select(-1)

# Elimine datos Faltantes -> Emp_length
df_clean = df_clean %>% filter(!is.na(Emp_length) | !nzchar(df$Emp_length) )
df_clean = df_clean %>% filter(!is.na(Rate) | !nzchar(Rate) )

sapply(df_clean_2, function(x) sum(is.na(x) | !nzchar(x)))

# Elimine datos Atipicos para una de las variables
# Mostrar cuartiles
quantile(df_clean$Age, na.rm = TRUE)

# Calcular IQR
Q1 <- quantile(df_clean$Age,0.25, na.rm = TRUE)
Q1
Q3 <- quantile(df_clean$Age,0.75, na.rm = TRUE)
Q3
IQR <- Q3 - Q1
IQR

# Calcular límites
limite_inf <- Q1 - 1.5 * IQR
limite_sup <- Q3 + 1.5 * IQR
limite_inf
limite_sup

# Detectar outliers usando los límites inferior y superior
outliers <- df_clean$Age[df_clean$Age < limite_inf | df_clean$Age > limite_sup]
print(outliers)

# Cantidad absoluta y relativa de outliers
sum(df_clean$Age < limite_inf | df_clean$Age > limite_sup)
mean(df_clean$Age < limite_inf | df_clean$Age > limite_sup)

# Observaciones con valores atípicos
df_out <- df_clean %>% filter(Age < limite_inf | Age > limite_sup)
df_out

# Visualización con boxplot (diagrama de caja)
boxplot(df_clean$Age, main = "Boxplot de Age",
        ylab = "Precio del pasaje")
boxplot.stats(df_clean$Age)$out

# Eliminar Outliers
df_clean <- df_clean %>%
  filter(Age >= limite_inf & Age <= limite_sup)

sd(df_clean$Age)
quantile(df_clean$Age)
boxplot(df_clean$Age)
```

Interprete los valores atípicos

¿Qué análisis puede realizar respecto a los valores atípicos detectados

Con el análisis de los valores atípicos se pudo evidenciar que existían registros con más de 100 años de edad, con lo cual se procedió a eliminar ya que así se nos recalcó. Ya que si no los tratábamos nos iban a ocasionar problemas en los análisis que realicemos a continuación. Porque también se los podía imputar de diferentes formas más por ejemplo aplicando mediana, raíz cuadrada y logaritmo

PREGUNTA 2: ANÁLISIS DE COMPONENTES PRINCIPALES (PCA)

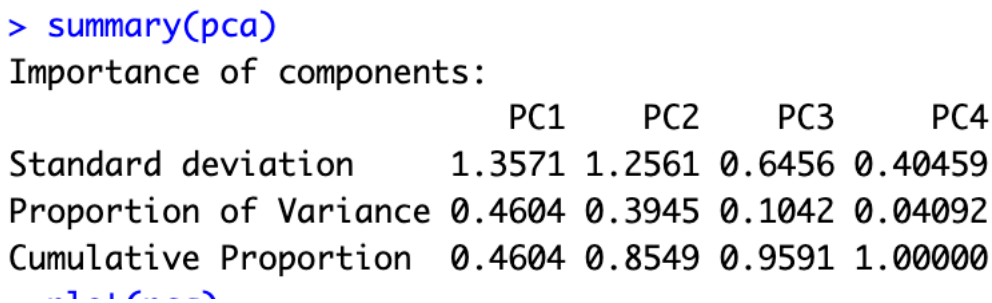
Script R

En el cuadro verde pegue el texto de los comandos R usados para lo solicitado en la Pregunta 2

```
#########################
# ANALISIS PCA
#########################
str(df_clean)

# Calcular PCA sobre varibles continuas
df_pca = df_clean %>% select(Amount, Age, Cred_length, Percent_income)
str(df_pca)

pca <- prcomp(df_pca, center = TRUE, scale. = TRUE)

# Ver la varianza explicada (importancia de componentes)
summary(pca)

plot(pca)

# Ver la matriz de rotación
print(pca$rotation)

# Convertir los datos de PCA a un data frame
pca_data <- as.data.frame(pca$x)

# Añadir las etiquetas
pca_data$TipoVivienda <- df_clean$Home

# Crear un gráfico de dispersión para analizar El proposito del Prestamo
# en el contexto de las dos primeras componentes PC1 y PC2
ggplot(pca_data, aes(x = PC1, y = PC2, colour = TipoVivienda)) +
  geom_point() +
  labs(title = "PCA del Conjunto de Datos de Prestamos",
       x = "Primer Componente Principal",
       y = "Segundo Componente Principal") +
  theme_minimal()
```

Resumen del PCA

Pegue la imagen del cuadro resumen del PCA realizado



```
> summary(pca)
Importance of components:
                          PC1    PC2    PC3     PC4
Standard deviation     1.3571 1.2561 0.6456 0.40459
Proportion of Variance 0.4604 0.3945 0.1042 0.04092
Cumulative Proportion  0.4604 0.8549 0.9591 1.00000
```

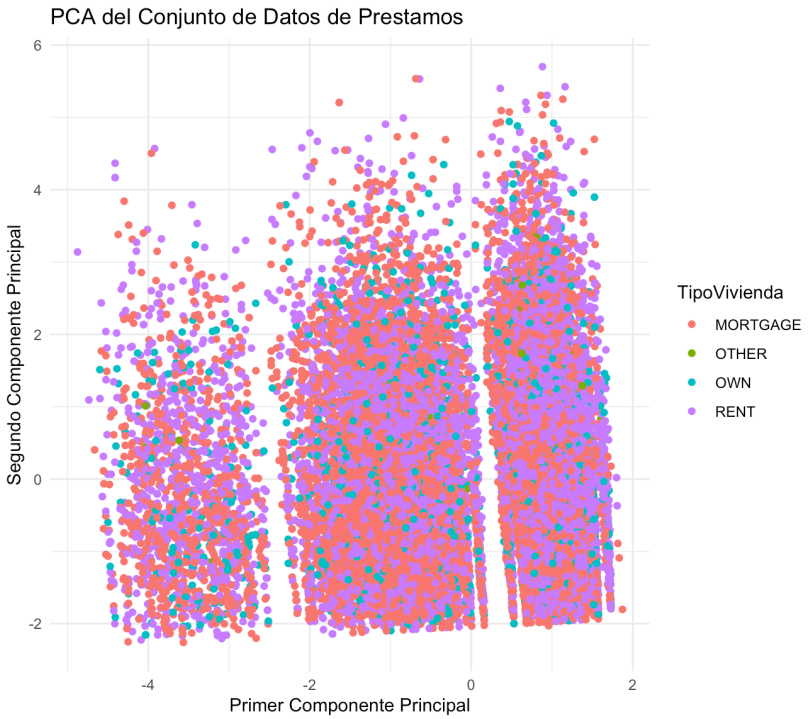
Matriz de rotación

Pegue la imagen de la matriz de rotación del PCA realizado

```
> # Ver la matriz de rotación
> print(pca$rotation)
                      PC1         PC2          PC3         PC4
Amount         -0.08770185  0.70235444 -0.7050322738 -0.04400133
Age            -0.70479082 -0.04444900 -0.0007962126  0.70802087
Cred_length    -0.70397139 -0.04461223  0.0870304723 -0.70346267
Percent_income -0.00160104  0.70903616  0.7038142908  0.04371048
```

Gráfico de las primeras componentes principales

Pegue la imagen del gráfico de las componentes principales



Interprete los resultados

¿Qué análisis puede realizar respecto a lo resultados del PCA realizado?

Con los resultados obtenido se puede observar, gracias al resumen del PCA que los primeros componentes representan el 85.48% de los datos los cuales son las variables Amount y Age. Lo cual nos indica que se puede reducir su dimensionalidad de 4 a 2 variables sin perder información relevante.

Mientras que con la matriz de rotación se puede observar que en Componente 1 la que tiene más peso es la variable Age y en el Componente 2 es la variable Percent_Income

Con la gráfica podemos interpretar:

- Cada punto representa una observación (una solicitud de préstamo) transformada en sus coordenadas

principales (PC1 y PC2).

- Los colores indican el Tipo de vivienda de quien solicita el préstamo
- PC1 y PC2 representan combinaciones lineales de las variables originales y juntas explican gran parte de la variabilidad del dataset.

Se puede interpretar que las personas que solicitan mas prestamos son las personas que cuenta con una vivienda de tipo RENTA y despues le sigue las personas que tienen una vivienda tipo HIPOTECA Y despues le sigue las personas que tienen una vivienda tipo PROPIA y otras muy pocas de tipo OTHER.

.-