Problem Statement

1. Use the link given below and locate the bank marketing dataset. https://archive.ics.uci.edu/ml/machine-learning-databases/00222/

Perform the below operations:

a) **Create a visual for representing missing values in the dataset.**

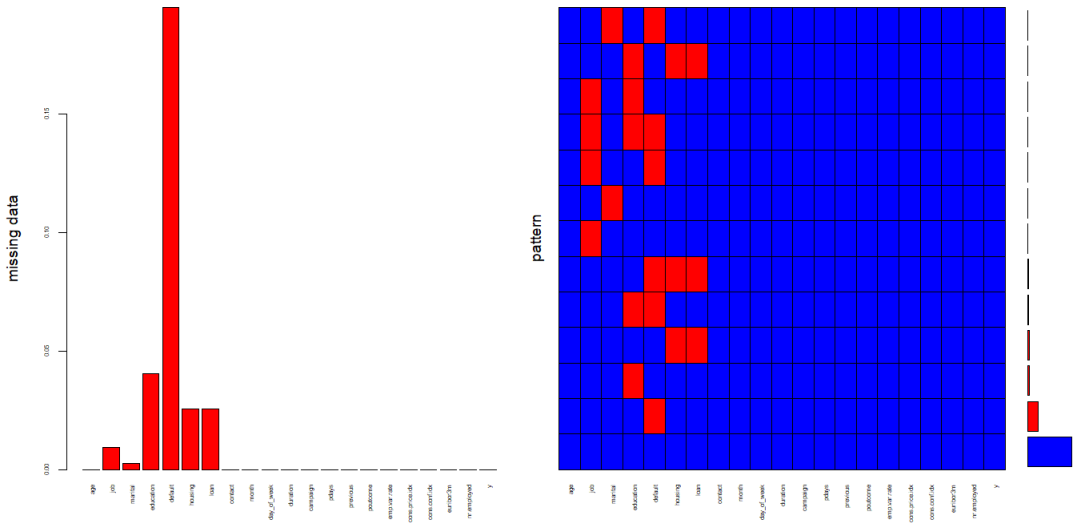
Answer :

```
>bank <- read.csv bank <- read.csv("C:/Users/Vikram/Desktop/Acad/bank-additional.csv",sep=";")

> View(bank)
> dim(bank)
[1] 4119 21
> str(bank))


> library(psych)
> psych::describe(bank)

 > library(VIM)
 >
 > missing <- bank
 > missing[missing == "unknown"] <- NA
 >
 > aggr(missing, col=c('blue', 'red'),
 + numbers=TRUE, sortvars= TRUE,
 + labels=names(missing), cex.axis=0.5,
 + gap=3, ylab=c("missing data","pattern"))
 Warning message:
 In plot.aggr(res, ...) : not enough horizontal space to display freq
 >
 > sapply(missing, function(x) sum(is.na(x)))
```

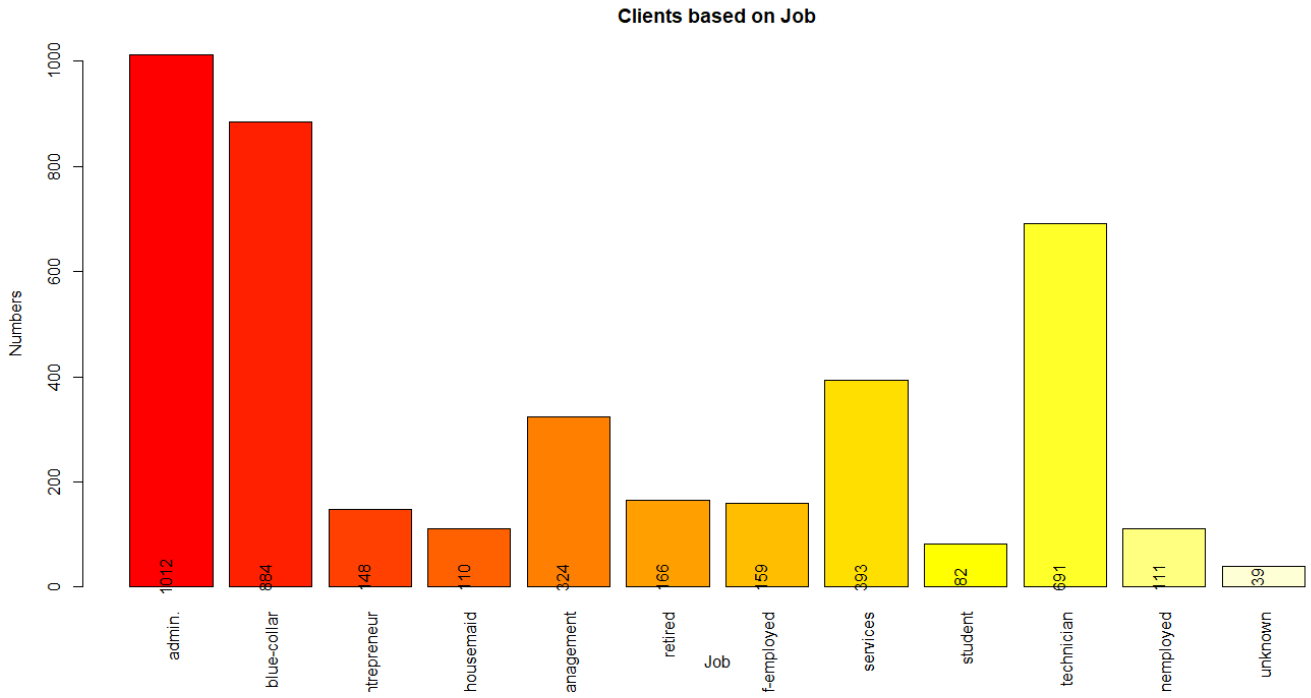


b) **Show a distribution of clients based on a job.**

Answer :

```
> t <- table(bank$job)
> # distribution in tabular form
> t
```

```
> title <- barplot(t, xlab = "Job", ylab = "Numbers", main = "Clients based on Job",
+ col = heat.colors(12), las=3)
> text(title, 0, t, pos = 3, srt = 90)
```



c) Check whether is there any relation between Job and Marital Status?

Answer :

```
> chisq.test(missing$job, missing$marital)

        Pearson's Chi-squared test

data: missing$job and missing$marital
X-squared = 395.28, df = 20, p-value < 2.2e-16
```

Ho : There is NO association between Job and Marital Status
Since P Value is less than 0.05, there is association between Job and Marital status at 95% confidence level.
Since NA values are very less, they are omitted.

d) **Check whether is there any association between Job and Education**?

Answer :

```
> chisq.test(missing$job, missing$education)

        Pearson's Chi-squared test

data: missing$job and missing$education
X-squared = 3429, df = 60, p-value < 2.2e-16
```

Ho : There is NO association between Job and Education.
Since the P value is less than 0.05, there is association between Job and Education at 95% confidence level.Since NA values are very less, they are omitted