R Markdown Data science

2023-10-24

Student Name: Manav Saini

Student ID: 20608893

Mutual Funds India detailed Dataset (Source Kaggle)

The dataset is downloaded from Kaggle.This is a comprehensive data set about Indian mutual funds which contains a mix of categorical as well as numerical variables providing an opportunity to go in depth and explore as well as predict the relationships and risk factors that these investment vehicles present.

```
#Read the file and store it in the variable named CD.
CD = read.csv("/Users/manavsaini/Downloads/comprehensive_mutual_funds_data.csv")
```

Checking the first few rows.

As observed, the data 20 variables of both numeric and categorical variables. 15 of these variables are numerical variables and 10 are continues data values #more suitable towards the regression based tasks.

```
head(CD)
```

```
##                                       scheme_name min_sip min_lumpsum
## 1 Aditya Birla SL Active Debt Multi-Mgr FoF-Dir Growth     100         100
## 2                        Aditya Birla SL Arbitrage Fund    1000        1000
## 3       Aditya Birla SL Asset Allocator FoF-Dir Growth    1000        1000
## 4     Aditya Birla SL Bal Bhavishya Yojna - Dir Growth     500        1000
## 5               Aditya Birla SL Balanced Advantage Fund     100         100
## 6 Aditya Birla SL Banking&Financial Services-DirGrowth    1000        1000
##   expense_ratio fund_size_cr fund_age_yr       fund_manager sortino alpha    sd
## 1          0.27           10          10     Kaustubh Gupta    0.32  2.24  9.39
## 2          0.36         4288          10   Lovelish Solanki    1.33  1.53  0.72
## 3          0.53          157          10 Vinod Narayan Bhat    3.44  2.67 10.58
## 4          0.76          637           4        Atul Penkar    2.18 -6.37 14.99
## 5          0.61         6386          10       Mohit Sharma    3.69  1.99 10.38
## 6          1.17         2384           9        Dhaval Gala    2.07  1.24 25.53
##   beta sharpe risk_level                          amc_name rating
## 1 0.01   0.24          3 Aditya Birla Sun Life Mutual Fund      3
## 2 0.56    1.1          1 Aditya Birla Sun Life Mutual Fund      3
## 3 0.67   1.42          5 Aditya Birla Sun Life Mutual Fund      3
## 4 0.85    0.9          6 Aditya Birla Sun Life Mutual Fund      2
## 5 0.68   1.39          6 Aditya Birla Sun Life Mutual Fund      4
## 6 0.96   0.97          6 Aditya Birla Sun Life Mutual Fund      2
##            category                                  sub_category returns_1yr
```

```
## 1             Other                                    FoFs Domestic         4.0
## 2            Hybrid                          Arbitrage Mutual Funds         5.6
## 3             Other                                    FoFs Domestic         2.0
## 4 Solution Oriented                                  Childrens Funds        -0.7
## 5            Hybrid Dynamic Asset Allocation or Balanced Advantage         4.5
## 6            Equity               Sectoral / Thematic Mutual Funds         5.3
##   returns_3yr returns_5yr
## 1         6.5         6.9
## 2         4.8         5.5
## 3        18.9         9.7
## 4        17.1          NA
## 5        18.6         9.7
## 6        24.6         9.2
```

Checking the attributes using the "str" function in R.

#The file has 20 variables of 814 observations. It is rich enough to predict and #use various regression as well as classification techniques.

```
str(CD)
```

```
## 'data.frame':    814 obs. of  20 variables:
##  $ scheme_name  : chr  "Aditya Birla SL Active Debt Multi-Mgr FoF-Dir Growth" "Aditya Birla SL Arbit
##  $ min_sip      : int  100 1000 1000 500 100 1000 1000 1000 100 100 ...
##  $ min_lumpsum  : int  100 1000 1000 1000 100 1000 1000 1000 100 100 ...
##  $ expense_ratio: num  0.27 0.36 0.53 0.76 0.61 1.17 0.37 1.29 0.31 0.69 ...
##  $ fund_size_cr : num  10 4288 157 637 6386 ...
##  $ fund_age_yr  : int  10 10 10 4 10 9 10 10 10 8 ...
##  $ fund_manager : chr  "Kaustubh Gupta" "Lovelish Solanki" "Vinod Narayan Bhat" "Atul Penkar" ...
##  $ sortino      : chr  "0.32" "1.33" "3.44" "2.18" ...
##  $ alpha        : chr  "2.24" "1.53" "2.67" "-6.37" ...
##  $ sd           : chr  "9.39" "0.72" "10.58" "14.99" ...
##  $ beta         : chr  "0.01" "0.56" "0.67" "0.85" ...
##  $ sharpe       : chr  "0.24" "1.1" "1.42" "0.9" ...
##  $ risk_level   : int  3 1 5 6 6 6 3 6 2 4 ...
##  $ amc_name     : chr  "Aditya Birla Sun Life Mutual Fund" "Aditya Birla Sun Life Mutual Fund" "Adity
##  $ rating       : int  3 3 3 2 4 2 4 0 4 4 ...
##  $ category     : chr  "Other" "Hybrid" "Other" "Solution Oriented" ...
##  $ sub_category : chr  "FoFs Domestic" "Arbitrage Mutual Funds" "FoFs Domestic" "Childrens Funds" ..
##  $ returns_1yr  : num  4 5.6 2 -0.7 4.5 5.3 4.5 -10.3 4.9 8.7 ...
##  $ returns_3yr  : num  6.5 4.8 18.9 17.1 18.6 24.6 6.8 29.4 7.1 9 ...
##  $ returns_5yr  : num  6.9 5.5 9.7 NA 9.7 9.2 7.3 9.3 7.4 7.2 ...
```

Converting Categorical to Numerical.

```
# Create a categorical variable
#We selectively converted the categorical variables into factors.
CD$amc_name = as.factor(CD$amc_name)
CD$category = as.factor(CD$category)
CD$sub_category = as.factor(CD$sub_category)
```

```r
#checking if the variables have been converted to factors successfully.
str(CD)
```

```
## 'data.frame':    814 obs. of  20 variables:
##  $ scheme_name  : chr  "Aditya Birla SL Active Debt Multi-Mgr FoF-Dir Growth" "Aditya Birla SL Arbit
##  $ min_sip      : int  100 1000 1000 500 100 1000 1000 1000 100 100 ...
##  $ min_lumpsum  : int  100 1000 1000 1000 100 1000 1000 1000 100 100 ...
##  $ expense_ratio: num  0.27 0.36 0.53 0.76 0.61 1.17 0.37 1.29 0.31 0.69 ...
##  $ fund_size_cr : num  10 4288 157 637 6386 ...
##  $ fund_age_yr  : int  10 10 10 4 10 9 10 10 10 8 ...
##  $ fund_manager : chr  "Kaustubh Gupta" "Lovelish Solanki" "Vinod Narayan Bhat" "Atul Penkar" ...
##  $ sortino      : chr  "0.32" "1.33" "3.44" "2.18" ...
##  $ alpha        : chr  "2.24" "1.53" "2.67" "-6.37" ...
##  $ sd           : chr  "9.39" "0.72" "10.58" "14.99" ...
##  $ beta         : chr  "0.01" "0.56" "0.67" "0.85" ...
##  $ sharpe       : chr  "0.24" "1.1" "1.42" "0.9" ...
##  $ risk_level   : int  3 1 5 6 6 6 3 6 2 4 ...
##  $ amc_name     : Factor w/ 39 levels "Aditya Birla Sun Life Mutual Fund",..: 1 1 1 1 1 1 1 1 1 ..
##  $ rating       : int  3 3 3 2 4 2 4 0 4 4 ...
##  $ category     : Factor w/ 5 levels "Debt","Equity",..: 4 3 4 5 3 2 1 2 1 1 ...
##  $ sub_category : Factor w/ 38 levels "Aggressive Hybrid Mutual Funds",..: 18 2 18 4 10 34 3 34 7 8
##  $ returns_1yr  : num  4 5.6 2 -0.7 4.5 5.3 4.5 -10.3 4.9 8.7 ...
##  $ returns_3yr  : num  6.5 4.8 18.9 17.1 18.6 24.6 6.8 29.4 7.1 9 ...
##  $ returns_5yr  : num  6.9 5.5 9.7 NA 9.7 9.2 7.3 9.3 7.4 7.2 ...
```

Potential research questions:

Regression Analysis:

Predicting Riskiness of Mutual Funds: Can we predict if a mutual fund policy is risky based on attributes like "expense_ratio," "fund_size_cr," "fund_age_yr," "sortino," "alpha," "sd," "beta," "sharpe," "amc_name," "rating," and "category"? This analysis could help investors make informed decisions when choosing mutual funds.

Classification Analysis:

Categorizing Mutual Funds: Can we classify mutual funds into categories like "Low Risk," "Medium Risk," and "High Risk" based on the given attributes. This classification can help investors choose funds that align with their risk tolerance.