Structure of R programming

---

title: "CSV Data Analysis in R - project00- dataset HR-attrition"

format: html

editor: visual

{r}

install.packages("tidyverse")

library(tidyverse)

data <- read\_csv("original.csv")

# View first few rows

head(data)

glimpse(data)

##str(data)

summary(data)

##names(data) ## column names

data$Education <- factor(data$Education,

levels = c(1, 2, 3, 4),

labels = c("under diploma", "diploma", "bachlor", "Master"),

ordered = TRUE)

data$EmployeeNumber <- as.factor(data$EmployeeNumber)

data$EnvironmentSatisfaction<- as.factor(data$EnvironmentSatisfaction, ordered = TRUE)

data$JobInvolvement <- as.factor(data$JobInvolvement, ordered = TRUE )

data$JobLevel <- as.factor(data$JobLevel, ordered = TRUE)

data$BusinessTravel <- as.factor(data$BusinessTravel,ordered = TRUE)

data$EmployeeCount <- as.factor(data$EmployeeCount)

data$JobSatisfaction <- as.factor(data$JobSatisfaction,ordered = TRUE)

data$PerformanceRating <- as.factor(data$PerformanceRating,ordered = TRUE)

data$RelationshipSatisfaction <- as.factor(data$RelationshipSatisfaction,ordered = TRUE)

# Example: mean of a column

mean(data$MonthlyIncome, na.rm = TRUE)

# Summary of all numeric columns

summary\_table <- data %>%

summarise(across(where(is.numeric),list(mean = mean,median = median, sd = sd), na.rm = TRUE))

# write a summary table in format\_csv()

summary\_table %>%

summarise(across(where(is.numeric), list(mean = mean,median = median sd = sd), na.rm = TRUE))

write\_csv(summary\_table, "summary\_table.csv")

{r}

view(data)

print(head(data, 10))

| **Original Columns** | **Normalized Table** | **Explanation** |
| --- | --- | --- |
| EmployeeNumber, Age, Gender, MaritalStatus, etc. | Employees | Employee personal attributes |
| JobRole, Department, JobLevel | JobRoles | Job roles are reused and have shared departments |
| Education, EducationField | EducationLevels | Education type and field normalization |
| Attrition, MonthlyIncome, OverTime | WorkData | Business data that may change over time |

| **Original Columns** | **Normalized Table** | **Explanation** |
| --- | --- | --- |
| EmployeeNumber, Age, Gender, MaritalStatus, etc. | Employees | Employee personal attributes |
| JobRole, Department, JobLevel | JobRoles | Job roles are reused and have shared departments |
| Education, EducationField | EducationLevels | Education type and field normalization |
| Attrition, MonthlyIncome, OverTime | WorkData | Business data that may change over time |

New analysis step by step:Ç

Project : data analysis project on "HR Analytics - Employee Attrition & Performance"

key tasks:

**Project Objectives:**

* Exploratory Data Analysis (EDA)
* Database Normalization (to 3NF)
* SQL Queries
* NoSQL Use Case (MongoDB)
* Data Preparation (Power Query/M)
* Data Modeling (Power BI, OLAP Star Schema)
* Identification of Business Questions
* Analytical Indicator Reporting
* Visualizations

**Deliverables (PowerPoint Report):**

* Project Presentation, Conclusion, and Recommendations
* Data Dictionary
* EDA Summary
* Normalization Process (3NF presentation)
* ER Diagram
* T-SQL Script Excerpts
* NoSQL Use Case Explanation
* Visual Comparison of Data (before/after transformations)
* OLAP Star Schema Image
* List of Metrics with Explanations
* Key Visualizations

**Tools Mentioned:**

* R and VSCode (for initial analysis if desired)
* SQL Server, SQL Server Management Studio (SSMS), T-SQL
* MongoDB and MongoDB Compass
* Power Query (M Language)
* Power BI Desktop (DAX Language)

**Data:**

* IBM HR Analytics Employee Attrition & Performance dataset (WA\_Fn-UseC\_-HR-Employee-Attrition.csv)

**Let's start with the first crucial step: Exploratory Data Analysis (EDA) and understanding your dataset.**

Before we jump into any tools, let's load and inspect the WA\_Fn-UseC\_-HR-Employee-Attrition.csv file. This will help us understand its structure, identify potential issues, and get a feel for the data we're working with.

**Step 1: Initial Data Inspection and Understanding**

Please confirm that you have access to a Python environment (e.g., Jupyter Notebook, VS Code with Python extension) where we can run some basic commands. If not, we can use a free online environment like Google Colab.

Once you're ready, we'll perform the following:

1. **Load the dataset:** We'll use the pandas library for this.
2. **Display the first few rows:** To get a quick glance at the data.
3. **Check data types:** To understand how each column is interpreted.
4. **Get a summary of descriptive statistics:** To see the distribution of numerical data.
5. **Check for missing values:** This is crucial for data cleaning.
6. **Identify unique values in categorical columns:** To understand the range of categories.

Are you ready to begin with Step 1, or do you have any questions before we proceed?