

Dataset: HRDataset_v14.csv

Background: HR Analytics

In this notebook, I explore an HR dataset to analyze employee structure and organizational trends.

The focus is on data reorganization, filtering, and multi-index structures using Python and Pandas.

I applied techniques for data cleaning, transforming and managing DataFrames and Series, and building insights about workforce composition and turnover.

Skills demonstrated: Data cleaning, multi-index DataFrames, filtering, organizational analysis, data persistence

Question 1 - Initial loading and exploration

- 1. The pandas library was imported; the "read_csv" function was called to read the dataset in question; The pandas function "set_option" was used here with the "display.max_columns" and "None" arguments in order to show all the columns in the original dataset.
- 2. In this context, the "info()" method was used to bring the names, types, and number of columns in the dataset.

After analysis we can identify the "ManagerName", "PerformanceScore" and "Gender ID" columns that should be converted to the "category" type.

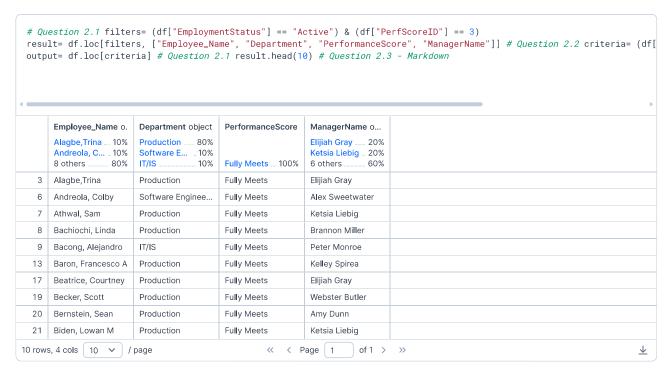
3. As a result of the operation performed, an object of type Pandas Series was selected, now representing a "PerformanceScore" column with its associated index.

```
# Questão 1.1
 import pandas as pd
 df= pd.read_csv("HRDataset_v14.csv")
 pd.set_option("display.max_columns", None)
 # Questão 1.2
 df.info()
 # Questão 1.3
 performance_series= df["PerformanceScore"]
 performance_series.head(10)
 df.head()
 <class 'pandas.core.frame.DataFrame'>
 RangeIndex: 311 entries, 0 to 310
 Data columns (total 36 columns):
 # Column
                             Non-Null Count Dtype
                             311 non-null object
 0
     Employee_Name
     EmpID
                             311 non-null
 1
                                            int64
 2
     MarriedID
                             311 non-null int64
                             311 non-null
 3
     MaritalStatusID
                                             int64
                               311 non-null
                                              int64
                             311 non-null
 5
     EmpStatusID
                                              int64
 6
    DeptID
                             311 non-null int64
     PerfScoreID
                             311 non-null
                                            int64
     FromDiversityJobFairID
 8
                              311 non-null
                                              int64
                               311 non-null
 10
    Termd
                              311 non-null
                                              int64
 11 PositionID
                             311 non-null
                                             int64
 12 Position
                             311 non-null
                                              object
                              311 non-null
 13 State
                                              object
 14 Zip
                              311 non-null
                                              int64
                             311 non-null
 15 DOB
                                              object
 16 Sex
                             311 non-null
                                              object
 17 MaritalDesc
                             311 non-null
                              311 non-null
 18 CitizenDesc
                                              object
 19 HispanicLatino
                               311 non-null
                                              object
 20
     RaceDesc
                              311 non-null
                                              object
 21 DateofHire
                              311 non-null
                                              object
 22 DateofTermination
                              104 non-null
                                              object
 23 TermReason
                               311 non-null
                                              object
 24
     EmploymentStatus
                               311 non-null
                                              object
                                                                                                                                    Р
      Employee_Name o. EmplO int64
                                          MarriedID int64
                                                             MaritalStatusID i...
                                                                              GenderID int64
                                                                                                 EmpStatusID int64
                                                                                                                   DeptID int64
   0 Adinolfi, Wilson K
                                   10026
                                                         0
                                                                           Λ
                                                                                             1
                                                                                                                                 5
                                                                                                                                  3
   1
      Ait Sidi, Karthikey...
                                   10084
                                                         1
                                                                                             1
                                                                                                               5
                                   10196
   2
      Akinkuolie, Sarah
                                                         1
                                                                           1
                                                                                             0
                                                                                                               5
                                                                                                                                 5
      Alagbe,Trina
                                   10088
                                                         1
                                                                           1
                                                                                             0
                                                                                                               1
                                                                                                                                  5
   4
      Anderson, Carol
                                   10069
                                                         0
                                                                           2
                                                                                             0
                                                                                                               5
                                                                                                                                 5

                                                                      of 1 > >>
5 rows, 36 cols 10 v / page
                                                                                                                                  \overline{+}
```

Question 2 - Selection of data for performance analysis

- 1. In this code we applied the filters requested in the question, applied the ".loc[]" function so that it brought only the records that were equivalent to the filters and visualized the result through the ".head()" method.
- 2. In this code block, only employees who have an absence number greater than 10 and who have a salary greater than 80000 have been filtered
- 3. To perform filtering in pure Python (without the Pandas library), we would start by opening the CSV file using the standard library csv module; we would go line by line using a "for"; we would apply the conditions of the question, more than 10 absences and salary greater than 80,000; We would store the result in a list.



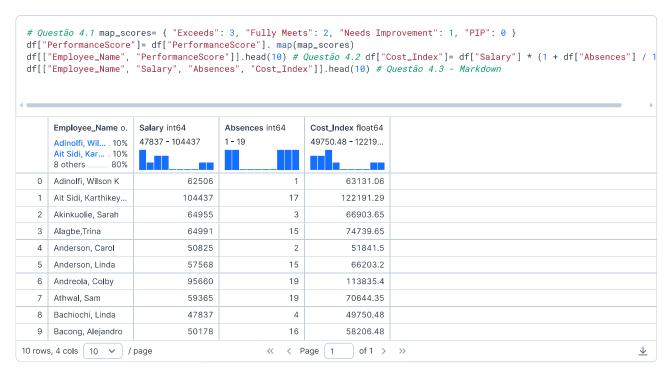
Question 3 - Structure with multiple indexes and segmentation by manager

- 1. The mentioned columns have been selected and stored in a new dataframe called "novo_df".
- 2. The Pandas "set_index" function was called to configure the "Department" and "ManagerName" columns as the indexes of the new dataframe.
- 3. The use of multi-indexes in this context helps to organize the data according to hierarchical levels, such as "Department" "Manager" "Employee"; it facilitates the visualization of the dataset and ends up facilitating the interpretation and understanding of the data; It facilitates data aggregation and in addition, there is data optimization, without the need to create additional columns or multiple filters.



Question 4 - Transformations and adjustments

- 1. In this question, we store in the variable map_scores a dictionary corresponding the value to its numerical correspondent and then apply the ".map" function in the "PerformanceScore" column using the variable defined above as a parameter.
- 2. In the code below, the expression "Salary * (1 + Absences / 100)" is used with the "Salary" and "Absences" columns as a parameter, returning a value that indicates the total cost of the employee (salary + absences) to the company.
- 3. Adjusted and calculated metrics such as the "Cost_Index" column help managers make better decisions; make data more reliable with regard to the reality of the company and its employees; It allows you to observe the company's actual spending and adds a weight related to behavior (absences).



Question 5 - Persistence of data

- 1. The dataframe created in question 3 and modified in question 4, with the column "Cost_Index" was brought into this code and transformed into a standard CSV (with a comma) and another with a semicolon (;).
- 2. The pandas "read_csv" method was used here to read the two CSVs and the "display" method to show the first 5 rows of each df returned.
- 3. If the tab used in the save is not the same as the one informed in the reading, Pandas (or other software) can read all columns as a single string or even generate incorrect extra columns; errors also happen when the encoding in which the file was saved is not the same as the one in which it is being read, as an example: file saved in "UTF-8", but read in "Latin-1" can lead to errors such as "André" (with the acute accent) appearing as another character.

