Primary EDA point :

* The dataset is relatively clean with no missing values.
* Three columns (EmployeeCount, Over18, StandardHours) are constant and will likely be removed or ignored in further analysis and modeling as they provide no predictive power.
* We have a mix of numerical and categorical variables, with some categorical variables encoded as integers. These might need conversion to factors for proper statistical treatment and visualization.
* The Attrition variable is binary, making this a classification problem.

Next level analysis: relationship between other variables with Attrition

The 03\_ Justificacao\_NormalizacaoBD.docx document outlines the normalization process applied to your HR employee dataset, moving from an initial single table to a Third Normal Form (3NF) structure. The primary goals of this process were to organize information, reduce redundancy, and ensure data integrity.

Here's a breakdown of how the normalization was performed according to your document:

### Justification of the Normalization Process

#### Second Normal Form (2NF)

The second step focused on eliminating partial dependencies related to the primary key, which was identified as 'EmployeeNumber'. A partial dependency occurs when a non-key attribute is dependent on only

*part* of a composite primary key. However, since 'EmployeeNumber' is a single-column primary key, the document's explanation of "eliminating partial dependencies" refers to identifying attributes that describe something *other* than the employee directly and moving them to new, specialized tables.

The 03\_ Justificacao\_NormalizacaoBD.docx document outlines the normalization process applied to your HR employee dataset, moving from an initial single table to a Third Normal Form (3NF) structure. The primary goals of this process were to organize information, reduce redundancy, and ensure data integrity.

Here's a breakdown of how the normalization was performed according to your document:

### Justification of the Normalization Process

The original dataset started as a single table in SQL Server Management Studio (SSMS) with 35 columns, containing various employee-related data such as age, job role, satisfaction levels, and salary.

#### First Normal Form (1NF)

To achieve 1NF, the document states that it was verified that all columns in the original table contained atomic values. This means no single cell held multiple values, nor were there any columns containing lists or composite fields.

* **Example provided:** The 'MaritalStatus' column contained only one marital status per employee (e.g., 'Single', 'Married', or 'Divorced').
* **Conclusion for 1NF:** The original structure of the dataset already met the criteria for 1NF.

#### Second Normal Form (2NF)

The second step focused on eliminating partial dependencies related to the primary key, which was identified as 'EmployeeNumber'. A partial dependency occurs when a non-key attribute is dependent on only

*part* of a composite primary key. However, since 'EmployeeNumber' is a single-column primary key, the document's explanation of "eliminating partial dependencies" refers to identifying attributes that describe something *other* than the employee directly and moving them to new, specialized tables.

