# CitData source:

|  |  |
| --- | --- |
| **Employee Attrition, Kaggle** | |
| Source | The source of this dataset is a Kaggle dataset called [Employee Attrition](https://www.kaggle.com/datasets/HRAnalyticRepository/employee-attrition-data). This is a fictionalised dataset for the purposes of trying to predict employee attrition in a Canadian company.  So, whilst not ideal that the dataset is fictional, the reality is most companies would be unwilling to share their employee data publicly which would make sourcing an employee-focused project very difficult. |
| Collection | The dataset was created six years ago by a user called Lyndon Sundmark. The people related data was generated by an online fake data generator. The user used a combination of fake data generators, R, and the application of randomised functions in order to create the fictionalised data. |
| Contents | The dataset contains information relating to employees at a Canadian organisation. There are 49,654 rows and 18 columns, although an initial glance indicates several duplicate rows. The variables cover various variables such as birthdate, age, length of service, location, gender, department, and termination status. With permission, the dataset was changed from Canada to Europe. |
| Limitations | There are a few limitations with this dataset which we need to be cognizant of:   * The most significant limitation for this dataset is the fact that it’s fictional, it would have been better to have access to a company’s real employee data. * The data spans employees who have joined between 2006 and 2015. If this were real data we would question the applicability of it as a dataset given the time lag, however, since it’s fictional it is satisfactory for our project. * There are some variables it would have been interesting to include, such as salary, and other Personal Identifiable Information such as ethnicity, disability status etc. to see if these have any impact on retention rates. It would also be interesting to use global data. |
| Ethics | Obviously if this were real data then there would be many ethical considerations to bear in mind. Information about a person that can be used to uniquely establish that persons’ identity is called personally identifiable information, or PII, and is protected by many laws across the globe. Some ethical considerations to bear in mind would be:   * The company – or employer – who collected the data is responsible for ensuring that the data is kept protected. * The employer must ensure that they only collect data which is reasonable and necessary. * The employer must also ensure that employees are comfortable with their PII being used in this way. If, for example, employees shared their data for a specific purpose and you as the employer want to use it for something else, you would need to firstly reconfirm with employees that they are happy for their data to be used in such a way. This could be in the form of an attestation. * The employer must also ensure that only certain individuals have access to the PII, although aggregated data which protects the identification of the employees may be shared more broadly. A general rule is that aggregated data which has fewer than five individuals in it is too small a group to report on given that it could be easy to identify individuals. |
| Relevancy | This dataset is very relevant to the project as it enables me to look at variables that impact whether an employee will stay at a company. |

# Data Profile

* [Consistency and Wrangling Checks](file:///C:\Users\rutha\OneDrive\CareerFoundry\7.%20Final%20Project\FinalProject_ConsistencyWrangling.zip)

The dataset has 49,653 rows and 18 columns. During the data quality checks, no duplicates or missing values were found, so the number of rows remains 49,653. I removed three unnecessary columns and added in a column which provides the population for each city location, based on census data from [Statistic Canada](https://www.statcan.gc.ca/en/start). Some column names were amended to adhere to consistent naming conventions, and some datatypes were reduced to mitigate against future memory constraints.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Variable* | *Description* | *Quant/Qual* | *Structured/*  *Unstructured* | *Time-variant/invariant* | *Nominal/ordinal/*  *Discrete/continuous* |
| EmployeeID | Unique employee number as an identifier | Qualitative | Unstructured | Invariant | Nominal |
| Recorddate\_key | When the information for the employee was taken | Qualitative | Structured | Variant | Ordinal |
| Birthdate\_key | Date the employee was born | Qualitative | Structured | Variant | Ordinal |
| Orighiredate\_key | Data the employee joined the organisation | Qualitative | Structured | Variant | Ordinal |
| Terminationdate\_key | Date the employee left the organisation, not applicable for those still employed. | Qualitative | Structured | Variant | Ordinal |
| Age | How old the employee was for that year | Quantitative | Structured | Variant | Discrete |
| Length\_of\_service | How long the employee was employed (status date – hire year) | Quantitative | Structured | Variant | Discrete |
| City\_name | Where the fictious employee was located | Qualitative | Structured | Invariant | Nominal |
| Department\_name | Which department the employee worked in | Qualitative | Structured | Invariant | Nominal |
| Job\_title | Their job title | Qualitative | Structured | Invariant | Nominal |
| Store\_name | The numeric name of the store they worked in (1 through 46) | Qualitative | Structured | Invariant | Nominal |
| Gender\_short | Their gender as a single letter (F or M) | Qualitative | Structured | Invariant | Nominal |
| Gender\_full | Their gender as a full work (Female or Male) | Qualitative | Structured | Invariant | Nominal |
| Termreason\_desc | The reason for their departure: retirement, resignation, or not applicable | Qualitative | Structured | Invariant | Nominal |
| Termtype\_desc | Whether their termination was voluntary or involuntary, or not applicable | Qualitative | Structured | Invariant | Nominal |
| STATUS\_YEAR | The year for that record of data | Qualitative | Structured | Variant | Nominal |
| STATUS | Whether the employee was active or terminated | Qualitative | Structured | Invariant | Nominal |
| BUSINESS\_UNIT | Whether the employee worked in head office or not | Qualitative | Structured | Invariant | Nominal |

# Questions

1. How many employees have been terminated?
2. How many employees are still active?
3. What variables impact attrition at our fictional Canadian company?
   1. Age
   2. Seniority
   3. Store
   4. City population
   5. Gender
4. Is the voluntary or involuntary nature of the termination impacted by any of the above variables?