**Task 6.1**

Nicole Chiu

July 17, 2023

Data Summary

* Data Source: The data set is available and downloadable from Kaggle (<https://www.kaggle.com/datasets/HRAnalyticRepository/employee-attrition-data>).
* Data Collection: The dataset is fictional and generated using fake data generators, random number functions, and the randomize function within R. The dataset was last updated six years ago.
* Data Content: The dataset contains fictional employee data of a Canadian company, with employees who’s start dates range between 2006 and 2015.
* Why this data set: Because I work in the Diversity, Equity, and Inclusion space, I was most curious to complete the final achievement project around people-centered data, particularly diversity-related data, such as ethnicity, gender, etc. However, this data is sensitive and very unlikely to be available in an open data source. This dataset still lacks all the diversity-related data that I am interested in; however, companies would still be unwilling to disclose their employee data and so this fictional dataset is close to what I would be interested in – looking at whether an employee is likely to stay or leave a company.
* Data Profile:
  + 18 columns and 49,653 rows
  + No missing values
  + No duplicates
  + Removed two columns: “recorddate\_key” and “gender\_short”

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variables** | **Description** | **Time-variant/ invariant** | **Structured/ Unstructured** | **Qualitative/ Quantitative** | **Nominal/Ordinal**  **Discrete/ Continuous** |
| EmployeeID | Employee’s ID number, unique identifier | Invariant | Unstructured | Qualitative | Nominal |
| Recorddate\_key | The date that employee data was recorded | Variant | Structured | Qualitative | Ordinal |
| Birthdate\_key | The employee’s birthdate | Variant | Structured | Qualitative | Ordinal |
| Orighiredate\_key | Employee’s start date | Variant | Structured | Qualitative | Ordinal |
| Terminationdate\_key | Employee’s end date, if applicable | Variant | Structured | Qualitative | Ordinal |
| Age | The employee’s age at record date | Variant | Structured | Qualitative | Discrete |
| Length\_of\_service | Tenure of employee (status year – hire year) | Variant | Structured | Quantitative | Discrete |
| City\_name | Employee’s city | Invariant | Structured | Qualitative | Nominal |
| Department\_name | Employee’s department | Invariant | Structured | Qualitative | Nominal |
| Job\_title | Employee’s job title, or role | Invariant | Structured | Qualitative | Nominal |
| Store\_name | Employee’s store of employment – numeric | Invariant | Structured | Qualitative | Nominal |
| Gender\_short | Employee’s gender, abbreviated as M or F | Invariant | Structured | Qualitative | Nominal |
| Gender\_full | Employee’s gender | Invariant | Structured | Qualitative | Nominal |
| Termreason\_desc | Reason for employee’s departure (retirement, resignation, or not applicable) | Invariant | Structured | Qualitative | Nominal |
| Termtype\_desc | The type of termination: voluntary, involuntary, or not applicable | Invariant | Structured | Qualitative | Nominal |
| STATUS\_YEAR | Year that the employee data was recorded | Variant | Structured | Qualitative | Nominal |
| STATUS | Employee’s status of active or terminated | Invariant | Structured | Qualitative | Nominal |
| BUSINESS\_UNIT | Employee’s business unit: head office or otherwise | Invariant | Structured | Qualitative | Nominal |

Limitations & Ethics

* Limitations:
  + The most significant limitation is that this dataset is randomized and fictional. The analysis performed on this made-up dataset will need to be applied to real-life people data of an organization when possible.
  + Though gender is included, other diversity variables would have been helpful to get a full DEI picture around employee attrition. For example, race/ethnicity would be most helpful, along with sexuality identity, if possible, and disability status.
  + This data is from one Canadian (fictional) organization. Data from organizations in the same business sector or similar type of organization would help benchmark this organization to see how they are doing compared to similar organizations.
  + The data includes employees who started between 2006 and 2015. As it is now 2023, more up-to-date, recent data would be helpful, particularly if we consider any significant events that occurred in Canada, or even global events (e.g., pandemic in 2020) and to make relevant decisions or recommendations based on the data.

Questions to Explore

* How many women have been terminated compared to men?
  + How many of these terminations were voluntary vs. involuntary?
* How many men vs. women are managers?
* How many millennials and gen-z (vs. older generation) are managers?
* How does age, gender, and/or role impact attrition?