NATIONAL CENTRE FOR SCIENTIFIC RESEARCH "DEMOKRITOS"

MSc in Data Science
Big Data Mining
Academic Year: 2020-2021

UNIVERSITY OF THE PELOPONNESE

Theodoros Efthymiadis

Maria Moutti

CREATE YOUR OWN DATASET & ADD NOISE

TABLE OF CONTENT

Table of Contents

INTRODUCTION

Human Resource analytics (HR Analytics) is defined as the area in the field of analytics that deals with people analysis and applying analytical process to the human capital within the organization to improve employee performance and improving employee retention.

PROBLEM FORMULATION

GENERAL DESCRIPTION

We don't start by finding data, but by asking questions:

1) What personality traits / psychometric indicators predict good employee performance in the workplace?
2) Do different departments / career paths favour different psychometric profiles?

We are working in the HR analytics department of a large organization. We will acquire the necessary data from our past and present employees and use them as a training set for predictive modelling. The trained models will be used to predict the performance of future recruitments, depending on their professional profile, psychometric profile and the position they are applying for. Moreover, our model can be used to relocate existing employees that are performing poorly to different roles or departments that might suit them better. In both cases, such a model can be a very useful business decision support tool for any HR manager.

DATA SOURCES

In order to solve this problem, there are four main data categories that are required, which will be described in detail:

1) Personal – Professional profile [required for past, present and future employees]
2) Psychometric Indicators [required for past, present and future employees (recruitment process will involve the appropriate psychometric test)]
3) Company Department Structure
4) Performance/evaluation data (KPIs) for the last 5 years, both global and department – specific [required for past and present employees]

Personal – Professional profile attributes

First Name: string (free text)
Last Name: string (free text)
Gender: will be constrained to (M, F, Other)
Date of Birth: date
Job title: string (free text)
Department: string (constrained: department must exist in the database)
Marital Status: will be constrained to (married, single)
Number of children: integer
Date of Hire: date
Currently working here: Boolean
Recruiter: string (constrained: The recruiter should be included in company employees)
Salary: float

Academic Level: string constrained to High school, technical, BSc, MSc, PHD
Experience in years: integer
Number of previous employers: integer

Psychometric Indicators attributes

Big Five personality assessment -> 5 psychometric traits rated in a scale of (1-100). Each trait can be further dispatched into 2 sub – traits called facets, also in a scale of (1-100). It's important to note that the scale of each trait is NOT calculated as the average score of its two facets. That's why we keep both the main traits and the sub – facets stored for analysis.

For a quick overview: https://en.wikipedia.org/wiki/Hierarchical_structure_of_the_Big_Five

- **Conscientiousness**: the ability to be organized, make plans and work diligently towards the accomplishment of goals
- ➢ **Orderliness**: the tendency to follow routines and plans, both in organizing space and time. Associated with the inability to tolerate chaos and surprises
- ➢ **Industriousness**: the tendency to apply standards, follow timetables and work hard, focused and efficiently to meet deadlines
- **Neuroticism**: the sensitivity to experience of negative emotions, such as pain, sadness, anger, fear, low self – esteem, stress and anxiety
- ➢ **Withdrawal**: the ability to handle uncertain, complex situations without being put down by fear, stress, worry and disappointment and become avoidant
- ➢ **Volatility**: the tendency to change mood quickly and lose composure when unfortunate events happen
- **Extraversion**: the tendency to experience positive emotion and be socially assertive
- ➢ **Enthusiasm**: the tendency to experience excitement, hope and optimism, as well as to easily engage in social situations and approach people fast
- ➢ **Assertiveness**: the tendency to trust your own opinion, put forward your ideas, take charge and lead the way, often dominating social situations
- **Openness to Experience**: the tendency to pursue new experiences, be creative and engage in artistic, abstract and intellectual endeavours
- ➢ **Intellect (NOT IQ)**: need for abstract ideas, concepts and novelty. The tendency to constantly explore and accurately formulate solutions to challenging problems
- ➢ **Openness**: interest in creativity, art and beauty. Tendency to be imaginative, visionary and sensitive to colours, architecture and music
- **Agreeableness**: the tendency to be compliant, care about others and avoid conflict
- ➢ **Compassion**: tendency to feel other people pain and to try to help them through self – sacrifice or creation of 'everybody wins' situations
- ➢ **Politeness**: tendency to be obedient, compliant and submissive and avoid confrontation and conflict

Department data

<u>Different Departments</u>:

Sales, Product, Finance, HR, Legal, Strategy, Technology

Each department will be a different table in our database schema, with information about its past and present employees. Employees' records of employees who have left the company more than 20 years ago will be deleted

Performance data (KPIs) for the last 5 years

- Global KPIs (common for all departments):
    - time of the employee with the company
    - number of promotions
    - Bonus received (euro)
    - Overtime (hours): integer
    - Chargeability (% work hours spent on income generating activities)
    - performance = {low, medium, high} (general annual evaluation)
    - percentile within department: top (85th-100th)/average (15th- 85th)/min (0th - 15th)

- Department – specific KPIs:
    - Sales:
        - Total Sales (Euro): float
        - Number of clients asking for the specific employee: integer
    - Product:
        - number of defects/time period in the products
        - number of complaining customers: integer
    - Finance:
        - Sum of non – servicing obligations (Euros): float
    - HR:
        - Total time of hired employees in the company: float
        - Average recruitment time: float
        - Number of recruited employees that were fired: integer
    - Legal:
        - Number of successful lawsuits: integer
        - Number of disputes amicably resolved (without court): integer
    - Strategy:
        - Total Sales (Euro): float
        - Number of different Teams participated in: integer
        - Number of projects participated in: integer
    - Technology:
        - Number of problematic code commitments: integer

Noise

Identified Sources of noise:

Typographic Errors:
This is the most common source of noise. All the information is provided to the system by manual or semi – automated data entry, which is prone to mistakes.
Wrong Updates / No Updates:
A significant amount of the aforementioned attributes, such as annual evaluation metrics of the last 5 years, 'marital status', or 'job title' are time – dependant. Therefore, regular and accurate updates of the database are required. These updates can be another major source of noise.
Job title not representative:
Often, a job title might not be representative of the actual activity of an employee. This might lead to wrong predictions.
Psychometric Indicators not fully representative:
The psychometric indicators often aren't perfect, therefore, being unable to describe the personality of an employee with complete accuracy.
Personal bias during the psychometric test (intentional or not):
Subjects' answers to psychometric tests are often biased. This can be attributed to three main reasons. Either they don't really know themselves that well, or they answer the questions based on an idealised version of themselves. Finally, they might be trying to manipulate the test results to get a better evaluation.
Chargeability: difficulty in quantifying in several departments
In some departments, like HR, it is difficult to identify what activities are revenue generating and what activities aren't.
Biased evaluation process:
The evaluation is administered by employees within the company, which may be biased to favour specific employees for various reasons.
Missing values (server crush/phishing):
A server crush or a phishing attempt might destroy a significant portion of the company's database, which may not be fully recoverable, resulting in a number of missing values.
Recently Hired Employees:
We mentioned that full evaluation metrics for a time period of 5 years will be stored. However, some employees might be working for the company for less than 5 years, resulting in missing values for the rest of the 5 year period.
Irrelevant Experience:
The years of experience of an employee stored in the database might correspond to non – relevant experience for their current position

DATA SCHEMA

Our data schema will contain:

A) 1 table for the Employees, including their professional and personal profile data,
B) 1 table for the psychometric indicators of each employee,

C) 1 table for each department (7 departments in total), containing the global and the department specific evaluation metrics for the employees of each department.

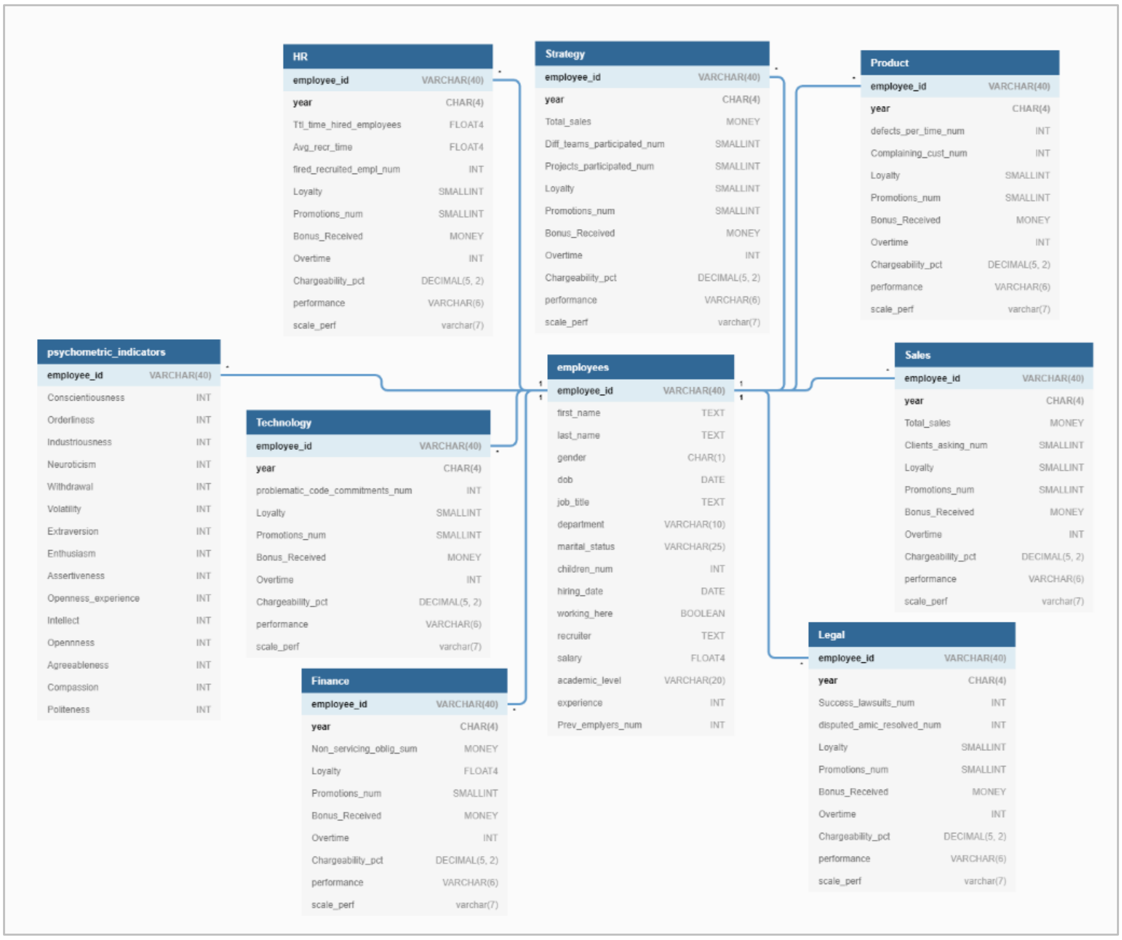
Therefore, there will be 9 tables in total:

EMPLOYEES							
Employee id(PK)	First Name	Last Name	Gender	Date of Birth	Job Title	Department	…

Psychometric Indicators							
Employee id (FK)	Conscientiousness	Orderliness	Industriousness	Neuroticism	Withdrawal	Volatility	…

Sales							
Employee id (FK)	Total Sales (Euro)	Number of clients asking for the specific employee	Time in the company (loyalty)	number of promotions	Bonus received (euro)	Chargeability	…

Similar tables will be created for all 7 departments. Each table will include all global evaluation KPIs and the department specific KPI's. The resulting relational schema is illustrated in the figure below:



HR	
employee_id	VARCHAR(40)
year	CHAR(4)
Ttl_time_hired_employees	FLOAT4
Avg_recr_time	FLOAT4
fired_recruited_empl_num	INT
Loyalty	SMALLINT
Promotions_num	SMALLINT
Bonus_Received	MONEY
Overtime	INT
Chargeability_pct	DECIMAL(5, 2)
performance	VARCHAR(6)
scale_perf	varchar(7)

Strategy	
employee_id	VARCHAR(40)
year	CHAR(4)
Total_sales	MONEY
Diff_teams_participated_num	SMALLINT
Projects_participated_num	SMALLINT
Loyalty	SMALLINT
Promotions_num	SMALLINT
Bonus_Received	MONEY
Overtime	INT
Chargeability_pct	DECIMAL(5, 2)
performance	VARCHAR(6)
scale_perf	varchar(7)

Product	
employee_id	VARCHAR(40)
year	CHAR(4)
defects_per_time_num	INT
Complaining_cust_num	INT
Loyalty	SMALLINT
Promotions_num	SMALLINT
Bonus_Received	MONEY
Overtime	INT
Chargeability_pct	DECIMAL(5, 2)
performance	VARCHAR(6)
scale_perf	varchar(7)

psychometric_indicators	
employee_id	VARCHAR(40)
Conscientiousness	INT
Orderliness	INT
Industriousness	INT
Neuroticism	INT
Withdrawal	INT
Volatility	INT
Extraversion	INT
Enthusiasm	INT
Assertiveness	INT
Openness_experience	INT
Intellect	INT
Opennness	INT
Agreeableness	INT
Compassion	INT
Politeness	INT

Technology	
employee_id	VARCHAR(40)
year	CHAR(4)
problematic_code_commitments_num	INT
Loyalty	SMALLINT
Promotions_num	SMALLINT
Bonus_Received	MONEY
Overtime	INT
Chargeability_pct	DECIMAL(5, 2)
performance	VARCHAR(6)
scale_perf	varchar(7)

employees	
employee_id	VARCHAR(40)
first_name	TEXT
last_name	TEXT
gender	CHAR(1)
dob	DATE
job_title	TEXT
department	VARCHAR(10)
marital_status	VARCHAR(25)
children_num	INT
hiring_date	DATE
working_here	BOOLEAN
recruiter	TEXT
salary	FLOAT4
academic_level	VARCHAR(20)
experience	INT
Prev_emplyers_num	INT

Sales	
employee_id	VARCHAR(40)
year	CHAR(4)
Total_sales	MONEY
Clients_asking_num	SMALLINT
Loyalty	SMALLINT
Promotions_num	SMALLINT
Bonus_Received	MONEY
Overtime	INT
Chargeability_pct	DECIMAL(5, 2)
performance	VARCHAR(6)
scale_perf	varchar(7)

Finance	
employee_id	VARCHAR(40)
year	CHAR(4)
Non_servicing_oblig_sum	MONEY
Loyalty	FLOAT4
Promotions_num	SMALLINT
Bonus_Received	MONEY
Overtime	INT
Chargeability_pct	DECIMAL(5, 2)
performance	VARCHAR(6)
scale_perf	varchar(7)

Legal	
employee_id	VARCHAR(40)
year	CHAR(4)
Success_lawsuits_num	INT
disputed_amic_resolved_num	INT
Loyalty	SMALLINT
Promotions_num	SMALLINT
Bonus_Received	MONEY
Overtime	INT
Chargeability_pct	DECIMAL(5, 2)
performance	VARCHAR(6)
scale_perf	varchar(7)

DATA GENERATION PROCESS

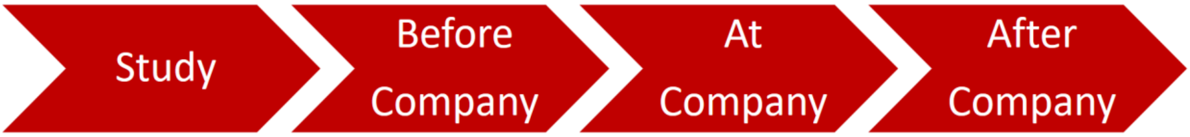
PERSONAL AND PROFESSIONAL PROFILE

We suppose that our data set will represent the current state of the company on 1st January 2021, after all evaluations have been concluded and reported for 2020. Perhaps some lazy or overloaded department managers didn't manage to complete and record the evaluations. This will be explored as a source of noise.

Firstly, we will generate the data set trying to apply certain constraints that make sense and create a 'perfect' data set without logical inconsistencies. Probable inconsistencies will be introduced later through the noise insertion processes.

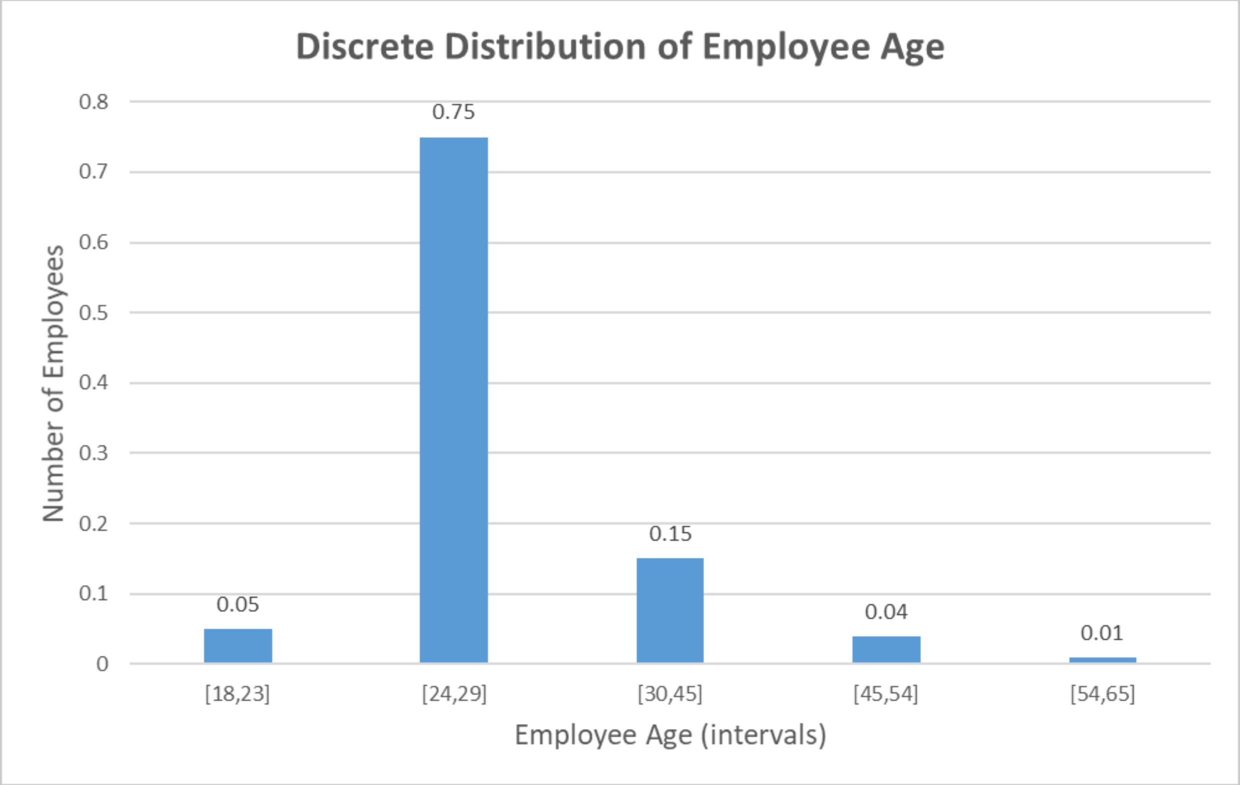
Our data set generation will begin by creating realistic employee profiles that take into consideration possible constraints regarding age, date of hire, academic level, work experience, previous employers etc. For instance, it's unlikely that an 18 years old person holds a PhD and has three previous employers. Moreover, a 73 years old person wouldn't be working in our company. In order to tackle these issues, while also keeping the complexity of the data generation process minimal, we will base our data generation on a number of assumptions. These assumptions will be described while explaining gradually the data generation process.

A sequence of stages is used to model the career of the employee and the timeline of our records:



- The 'Study' period refers to the years invested in academia and starts at the age of 18
- The 'Before Company' period refers to the years between the completion of employee's studies and the hiring date.
- The 'At Company' period refers to the time spent working at the company. This period can last up to the employee's 65th birthday.
- The 'After Company' period refers to the period from employee's resignation, retirement or firing until his/her records are deleted from the company's database. This period can last up to 20 years.

The first step to randomly generate a new employee is to create a random unique employee ID (uuid) and fake first and last name. Then, we should appoint a gender to the employee. It would be impractical to use a string parser and an NLP model to predict the gender based on the random generated name within this project. Therefore, the gender will be randomly generated with equal probability between 'M' for male and 'F' for female. As a result, 50% of the gender values will probably be wrong. In a similar fashion, the marital status will be randomly appointed with equal probability between 'Single' and 'Married'. Moreover, the age of the employee should be defined. The age of the employees working in our company ranges from 18 to 65. Moreover, the age is modelled through the following discrete distribution:



This distribution is based on the presumption that our company is involved in tech sector, where the majority of junior employees hold at least a Bachelor degree and, therefore, occupy the class between 24 and 29 years old. This distribution will be referred hereinafter as the '**Employee Age Distribution**'. It's important to make a distinction between past and current employees. The age of past employees refers to their age when they left the company. As a result, the Employee Age Distribution will be used to randomly estimate their age. However, their real age is calculated by adding the duration of their 'After Company' period to the last age record stored in our DB (calculated by the Employee Age Distribution). For example, an employee that left the company 12 years ago at age of 43 is now 55 years old. Based on the real employee age, a random birthday date will be generated.

Finally, the number of children of the employee will be calculated using a normal distribution that is dependent on the real age of the employee:

$$(Number\ of\ Children) = round\left(max\left[0, normal\left(1{,}5 * \min\left(1, \frac{age}{35}\right), 0{,}5\right)\right]\right)$$

This normal distribution has a mean value of 1,5 and a standard deviation of 0,5. This means that 99,7% of the samples will fall within the range $1{,}5\ \pm 3 * 0{,}5 = [0{,}3]$. The factor $\min\left(1, \frac{age}{35}\right)$ will reduce the average number of children for the younger employees.

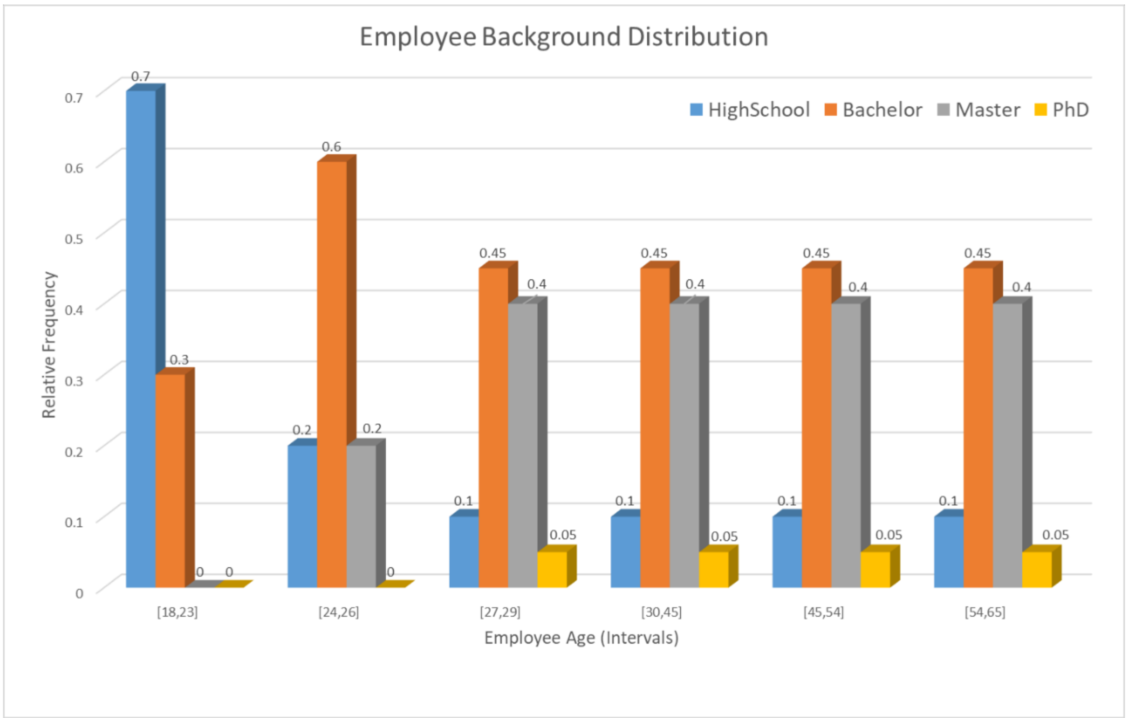
Everything mentioned so far can be summarized in the following key points:

- Use the 'Employee Age Distribution' to calculate the age interval of the employee.
- Use a uniform distribution in that time interval to estimate the exact age of the employee.

- Use a uniform distribution to decide if the employee is current or past. Given that the company is old and keeps a lot of records, we assume that 20% of the records refer to current employees and 80% refer to past employees (have left the company).
- If reference is made to a past employee, we use a uniform distribution in the range of [0,20] to estimate the duration of the 'After Company' period for the specific employee and add it to the employee age to calculate his/her real age.
- Based on the real age, use the faker library to generate a random birth date.
- Based on the real age, use the normal distribution that was described above to estimate the number of children.

Based on the age of the employees, their academic level will be randomly calculated based on a number of assumptions:

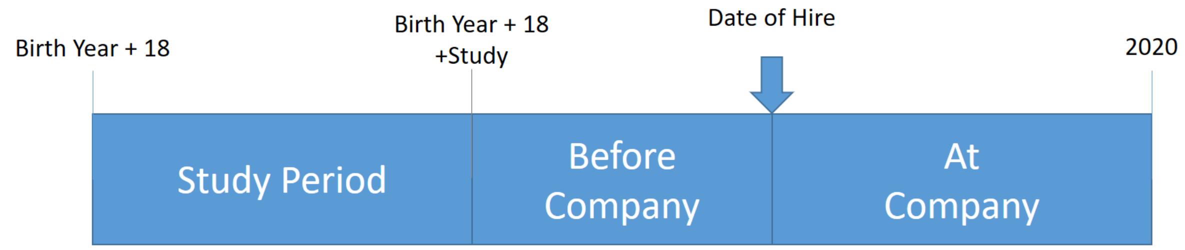
- Any academic title was acquired at the beginning of their career, starting at the age of 18.
- Higher academic titles were acquired immediately following lower ones, without any breaks or working experience in-between.
- Bachelor degrees always last 3 years
- Masters Degrees always last 2 years
- PhDs always last 4 years.
- As a result, an employee with academic level of 'MSc' has spent 5 year studying, while a 'PhD' has spent 9 years studying and is, therefore, at least 27 years old.
- The academic level of each employee is modelled through the following discrete distribution:



The different age intervals are mostly similar to the ones of the Employee Age Distribution, but do not overlap 100%. This distribution will be referred hereinafter as the '**Employee Background Distribution**'.

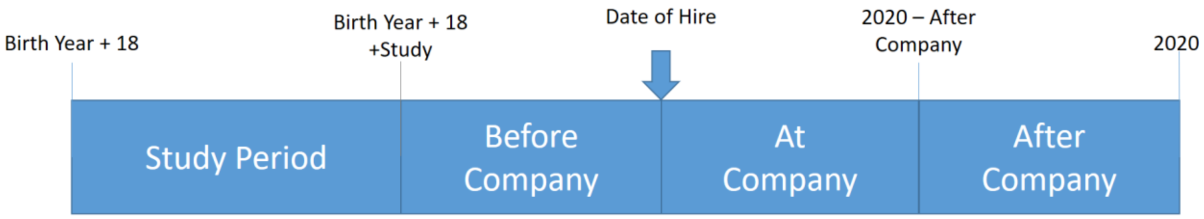
By knowing the academic background of the employee, we know how much time he/she spent studying and, in combination with the year of birth, this number allows for the calculation of the beginning of the employee's professional career. We can now randomly estimate the date that the employee was hired, as follows:

If the employee still works in the company:



$$(birth\ year) + 18 + (study\ period) < (year\ of\ hire) < 2020$$

If the employee has left the company:



$$(birth\ year) + 18 + (study\ period) < (year\ of\ hire) < 2020 - (years\ the\ employee\ left)$$

The year of hire will be randomly generated within the time interval mentioned above. Then, a random date of hire will be generated within the year of hire that was defined in the previous step.

After calculating the date of hire, we will estimate the working experience of the employee based on the length of the periods 'Before Company' and 'At Company'. We assume that 80% of the 'Before Company' period was spent working for other employers.

$$(Work\ Experience) = (Exp\ Before\ Company) + (Exp\ At\ Company)$$

The working experience Before Company is calculated using the same formula for past and current employees, while the experience At Company uses a slightly different formula:

$$(Exp\ Before\ Company) = 0.8\ [(Hire\ Year) - (birth\ year) + 18 + (study\ period)]$$

If the employee still works in the company:

$$(Exp\ At\ Company) = 2020 - (Hire\ Year)$$

If the employee has left the company:

$$(Exp\ At\ Company) = 2020 - (years\ the\ employee\ left) - (Hire\ Year)$$

Moreover, we will assume that the employee worked on average for 3 years for each previous employer. To account for shorter occupation periods, we will use the following formula:

$$(Number\ of\ previous\ employers) = int\left(\frac{(Exp\ Before\ Company)}{3}\right) + 1$$

By applying this formula, an employee with 4 years of experience Before Company will be considered to have 2 previous employers. The working experience divided by 3 will result to 1,333 previous employers, which will be rounded and increased by one to give the final output equal to 2.

Based on previous working experience, academic background and time in the company, the annual salary of the employee will be calculated using the following formulas:

$$(Salary\ Hired) = random\big(basic\ income, \big[basic\ income + y_{exp} * 2000 + y_{study} * 500\big]\big)$$

Where:

Salary Hired: The salary that the employee negotiated during the hiring process.

Basic Income: This depends on the country. For our example, it will be set to 10000 Euros/year.

y_{exp}: The working experience of the employee (in years) before he/she was hired.

y_{study}: The academic experience of the employee (in years) before he/she was hired.

The salary is randomly generated between the two extreme values to capture the variance of the ability of different candidates to negotiate their salary.

$$Salary = random\big([Salary\ Hired], \big[(Salary\ Hired) * (1+e)^{years\ in\ company}\big]\big)$$

Where:

Salary: The current employee salary.

e: The maximum percentage of annual salary raise based on the company's policy. In our example, it will be set to 10%.

years in company: The number of years the employee has worked for the company.

Once again, the random data generation within the interval described by the equation captures the ability of different employees to negotiate their annual salary range.

As a next step, we will randomly generate the department of the employee. This will be achieved by sampling from the list of existing departments using a uniform distribution. The potential departments, as described earlier, are Sales, Product, Finance, HR, Legal, Strategy and Technology.

Furthermore, we will randomly generate the job title of the employee from a list of pre – defined job titles that we created for the departments in our company. This list is illustrated in the next table:

Department Name	Job Titles
Sales	'Director of Sales', 'Sales Manager', 'Area Sales Manager', 'Sales Executive', 'Sales Representative', 'Brand Ambassador', 'Sales Associate'
Product	'Production Manager', 'Production Technician', 'Product Integration Assistant', 'Product Communications Planner', 'Product Brand Associate', 'Product Implementation Manager', 'Product Creative Analyst'
Finance	'Pricing Analyst', 'Financial Analyst', 'Credit Risk Analyst', 'Portfolio Analyst', 'Investment Manager', 'Credit Risk Manager', 'Finance Manager'
HR	'Recruiting Manager', 'Recruiting Assistant', 'Talent Consultant', 'Benefits Counselor', 'Retention Specialist', 'Workforce Analyst', 'HR Coordinator'
Legal	'Resolution Specialist', 'Legal Analyst', 'Legal Research Analyst', 'Manager Legal', 'Defense Attorney', 'Patent Attorney', 'Attorney General'
Strategy	'CIO', 'CEO', 'Strategy Director', 'Strategic Planner', 'Business Strategy Manager', 'Strategy Analyst', 'Business Planner'
Technology	'Information Security Manager', 'IT Support', 'IT Director', 'Software Engineer', 'Database Administrator', 'Network Engineer', 'Software Engineering Manager'

The final attribute of the 'personal and professional profile' data is the name and ID of the recruiter that hired the employee. The recruiter must meet two conditions:

- He/ She must be working in 'HR'.
- He/ She should have joined the company before the recruited employee.

For each employee, based on his/ her recruitment date, we start by identifying all possible recruiters. Then, we select randomly one of them using a uniform distribution. If there are no possible recruiters, the selected field is handled as 'NULL'.

PSYCHOMETRIC INDICATORS

The psychometric indicators are randomly generated in the range [0,100] using a uniform distribution. The only applicable constraint is that the value of each aspect of the Big Five scale should be located between the values of its two corresponding facets. For example, 'Agreeableness' is comprised of 'Compassion' and 'Politeness'. As a consequence, the value of 'Agreeableness' is randomly generated between the values of 'Compassion' and 'Politeness'.

EVALUATION METRICS STRUCTURE

All evaluation metrics for each employee are kept in the database table under the department of the employee. These data include both universal and department-specific evaluation metrics. The data are kept in the database for the last 5 years of employment, if available. In order to model this in our data schema in an efficient manner, each record in the department tables (e.g. HR) will use the combination of 'employee ID' and 'Year' as a primary key. In that regard, up to maximum 5 records will be referring to each employee.

GLOBAL EVALUATION METRICS (Common for all departments)

For each year, we will calculate the number of years that the employee has already worked for the company and store it in the 'Loyalty' attribute.

By assuming that each employee gets promoted on average once every 4 years, the number of promotions of an employee will be calculated as follows:

$$(Number\ of\ Promotions) = int\left(\frac{Loyalty}{4}\right)$$

The annual bonus will be randomly generated in the range of (0% - 30%) of current annual salary.

The annual overtime in hours will be randomly generated in the range of (0% - 20%) of annual working hours. The annual working hours will be estimated as follows:

$$(Annual\ Working\ Hours) = \left[months * \left(\frac{days}{month}\right) - \frac{(off\ days)}{year}\right] * \left(\frac{hours}{day}\right)$$

$$(Annual\ Working\ Hours) = [12 * 21 - 25] * 8 = 1816\ \frac{hours}{year}$$

The chargeability percentage will be randomly generated between 0 and 100.

The percentile of the employee within the department for the current year (namely his/her performance in comparison to his/her co-workers) will be calculated with the use of a uniform generator in the range (0,100).

- If the generator value is smaller than 15, the employee will be categorised to the 'Bottom 15%' and will also be assigned a performance value 'Low'.
- If the generator value is greater than 85, then the employee will be categorised to the 'Top 15%' and will also be assigned a performance value 'High'.
- In any other case, the employee will be categorised to the 'Mid 70%' percentile and will also be randomly assigned a performance value of 'Low', 'Medium' or 'High' with an equal probability.

DEPARTMENT SPECIFIC EVALUATION METRICS

There is a number of evaluation metrics specific to each department. These have to be calculated on a separate basis.

HR:

Each HR employee is evaluated based on the total time that the employees recruited by him/her worked for the company. This can be directly calculated by joining the 'Recruiter ID' attribute of each company's employee with the 'ID' attribute of each employee working in HR. After identifying all employees that were hired by each HR recruiter, we sum the time intervals that each of them spent in the company.

The average recruitment time will be randomly generated between 1 and 12 months.

Finally, in order to estimate how many of the recruits were fired, we will assume that 20% of the recruits that have left the company were fired.

Sales:

The amount of total annual sales will be randomly generated between 1000 and 100.000 Euro.

The number of clients that explicitly ask for the specific sales employee will be randomly generated between 0 and 5 and stored in the attribute of 'Clients Asking'.

Product:

The number of defects in the products per year will be randomly generated between 10 and 50.

The number of complaining clients will be randomly generated between 0 and 20.

Finance:

The total number of non – servicing obligations will be randomly generated between 0 and 10.000.

Legal:

The number of successful lawsuits will be randomly generated between 0 and 3.

The number of disputes amicably resolved will be randomly generated between 0 and 6.

Strategy:

The amount of total annual sales will be randomly generated between 1000 and 100.000 Euro. This attribute accounts for the fact that strategic managers are often involved in sales as well.

The number of different teams that the employee participated in will be randomly generated between 1 and 10.

The number of different projects that the employee participated in will be randomly generated between 1 and 20.

Technology:

The number of problematic code commitments will be randomly generated between 0 and 20.

NOISE INSERTION

There are several potential sources of noise that can be present in our data set. Some of them are directly evident, such as typographic errors and missing values, while others, such as personal bias when evaluating employees, are implicit and should be handled during the stage of data analysis. Our noise insertion process is only dealing with the former. There are 3 sources of noise that will be modelled:

- Typographic errors due to human error. This category may refer to individual wrong characters or entire strings.
- Random missing values due to human error or hardware malfunctioning.
- Entire records missing due to human error.

Typographic errors:

This is the most common source of error at data entry. In order to model single character or single digit errors we use a function that iterates through elements of a data frame and randomly replaces characters or digits with other, randomly generated characters or digits. However, not all elements of a data frame are parsed, but only those that are more likely to be recorded through a manual data entry process. We assume that the chance of a typographic error is 5% (this is customizable in the python script). Moreover, there are cases that an entire string might be incorrectly inserted into the database, for example when the employee who is conducting the data entry process confuses the names of two different employees. Finally, there are cases at data entry when an employee might accidentally choose the wrong value among a set of existing values. For instance, the record of an employee that works for Sales department might be assigned to the Technology department instead. We assume that such confusions happen with a probability of 2% (also customizable).

Random missing values:

Random missing values are usual in most database systems for multiple reasons. In our analysis, we will assume that any value in the database could be 'NaN' with a probability of 2% (customizable). This is implemented in the python code through a function that iterates over all elements of a data frame and randomly assigns 'NaN' values to them.

Entire records missing:

This type of error refers to the case where a manager forgot to provide the evaluation for a specific employee, or the report was not registered in the database. Moreover, this might account for the case that someone is hired through a referral, skips the interview process and, therefore, doesn't take a psychometric test. This type of value is modelled through a function that parses all records of a data frame and randomly drops some of them. We assume that the probability for a missing record is 2%.

PROGRAMMING IMPLEMENTATION

The implementation of everything mentioned above was executed through the development of 3 Python scripts. The scripts were developed in a virtual environment using the Pycharm software. In order to reproduce the results, it is suggested that the user creates a similar virtual environment and installs all necessary python libraries that are listed in the 'README.txt' file. Pay extra attention to the version of the 'numpy' library, because the latest version is prone to a number of bugs and should not be preferred.

After setting up the virtual environment and installing the libraries, the scripts should be executed in a specific order:

- Run the 'Personal_Profile.py' script by using the command line. It will produce an excel file in the current folder containing the personal information and psychometric profile of all employees. The file is named 'folder_path/employees.xlsx'.
- Run the 'Departments.py' script by using the command line. This will read the excel file that was created in the previous step, assign the employees to random departments, estimate their evaluation metrics and append the information back to the 'employees.xlsx' file. This version of the file is also provided in the deliverable.
- Run the 'Noise_Insertion.py' script by using the command line. This will read the 'employees.xlsx' file, insert noise to the data and provide an output file named 'noisy_employees.xlsx'. This file is also provided in the deliverable.

A large number of seed functions were used to control the random data generation from various python libraries. The results should be completely reproducible.