Data Generation

1. Dataset Scenario

   The dataset is generated to simulate the human resource data in a company. In reality, the personal information such as name, address, phone number, etc. would be also included in the dataset. However, these variables might not yield meaningful insights for the data analytics. In addition, it makes sense to censor or eliminate personal data if the data analytics task is outsourced to others. Even one single employee in the HR department cannot have the access to all the personal information of the whole company. Thus, in our scenario, we create 13 variables that are available for data analytics to make meaningful analysis. The overview of variable and their characteristics is introduced in the following section.

2. Overview of Variables

   The table below gives an overview of the variables including the possible value we assigned to each variable and the basic idea we have in our mind when designing it. The detailed design mechanism and the characteristics of each variable are presented in the next section.

ID	Variable Name	Data Type	Data Values
1	Age	Numeric	[20, 67] (integer)
2	Gender	Character	0 – female 1 – male
3	Department	Character	1 – Executive office 2 – R&D 3 – IT 4 – Marketing 5 – HR 6 – Sales 7 – Accounting
4	Position	Character	1 – Employee 2 – Manager 3 – Executive Manager 4 – Chef Executive Manager
5	Education	Character	1 – High school 2 – Bachelor 3 – Master 4 - Doctor

6	MaritalStatus	Character	1 – Single 2 – Married
7	Overtime	Numeric	[0, 35] (integer)
8	WorkingHours	Numeric	[140, 195] (integer)
9	JobTenure	Numeric	[0, 47] (integer)
10	Salary	Numeric	Integer
11	SatisfactionLevel	Numeric	[0, 100] (integer)
12	EQ	Numeric	[70, 150]()
13	IQ	Numeric	[70, 135]

Table 1. Variables Overview

3. Characteristics of Variables

3.1. Age

Age represents the age structure of all employees of the company in the scenario. Data is normal distributed ranging from 20 to 67.

3.2. Gender

The imaginary scenario is created with a clear gender imbalance in technical departments such as R&D and IT. In these departments 78% of the employees are male. On the other hand, 64% of employees in Marketing, HR, and Accounting are female. For other departments the normal ratio of 52:48 for male and female is used.

3.3. Department

There are 7 different departments introduced in this data set. The position of Executive Office is staffed with 3 people in the company. R&D and IT department respectively has 20% of employees. The other 4 departments share the rest 60% people equally.

3.4. Position

There is exactly one Chief Executive Manager in the whole company who belongs to Executive Office. Each department has one Executive Manager and 20% staffs as managers. The rest are all employees.

3.5. Education

Education refers to common levels of education that can be attained. Since the education process takes time, older people have larger

probability to have higher education. However, due to the historical access to education and modern trend in the higher education，employees older than 50 have a less educational background and the level varies from a younger generation.

3.6. Marital Status

MaritalStatus is assumed that the older a person is, the higher probability that he/she is married. In the scenario, 70% of employees older than 40 years are married. For employees between the age of 30 and 40, 60% of them are married. There are 60% singles in the age between 20 and 30. The ones 20 years old or younger are mostly single.

3.7. Overtime

Overtime refers to the time each employees work exceeds standard working hours (160 hours) per month. Moreover, it is assumed that the overworking hours of different departments and positions can differ. At first 7 departments are divided into two groups: higher overtime group includes R&D, IT, Marketing and Sales； lower overtime group includes Executive office, HR, and Accounting departments. Secondly, in each group that the higher the position of an employee the longer he/she works. The unique Chief Executive Manager has the highest overtime in the company. Based on the assumption, there will be 6 clusters regarding *Overtime* and *Position*.

Department	**Position**	**Overtime Range**
R&D, IT, Marketing, and Sales departments	Employee	[5, 15]
	Manager	[15, 25]
	Executive Manager	[25, 30]
Executive office, HR, and Accounting departments	Employee	0
	Manager	[1, 3]
	Executive Manager	[3, 5]
Chief Executive Manager		[31, 35]

Table 2. Range of Overtime

3.8. WorkingHours

WorkingHours is the overall working time of one employee per month. It equals to normal working hours plus overtime. The standard working time is 160 hours per month. It is assumed that employees are allowed to work less than standard time but not less than 140 hours. The value of *WorkingHours* with zero *Overtime* distributes between 140 and 160. Otherwise, the overall working hours are 160 plus *Overtime*.

3.9. JobTenure

Job tenure represents how many years the person have worked in the company. The people who worked less than one year are counted as 0. This variable is affected by *Age* and *Education*. The career path starts at *Age* 20 after graduating high school. For each next level in education specific duration of the educational process is considered. To get a bachelor degree usually takes 3 years. A master and doctor degree needs 5 and 8 years respectively after high school. Thus the formula of the maximum Job Tenure a person can have in this company is as follow.

$$JobTenure = Age - 20 - Education$$

3.10. Salary

In this scenario, Salary is dependent on *Position*, *JobTenure*, *WorkingHours* and *Overtime*. Employees are first grouped by *Position* and *JobTenure*. Within the group, *Salary* is linear correlated to *WorkingHours* and *Overtime*. Salary is composed of basic pay, pay for normal working hours and pay for working overtime. The detailed formula is written in Table 3.The salary of chief executive manager is always 20,000 per month as an outlier in this variable.

Position	**Job Tenure**	**Salary**
Employee	**≤ 7**	1000 + 10×(*WorkingHours* - *Overtime*) + 15×*Overtime*
	> 7	1200 + 10×(*WorkingHours* - *Overtime*) + 15×*Overtime*
Manager	**≤ 7**	2000 + 20×(*WorkingHours* - *Overtime*) + 30×*Overtime*
	> 7	2200 + 20×(*WorkingHours* - *Overtime*) + 30×*Overtime*
Executive	**≤ 7**	3000 + 30×(*WorkingHours* - *Overtime*) + 45×*Overtime*

Manager	> 7	3200 + 30×(*WorkingHours* - *Overtime*) + 45×*Overtime*
Chief Executive Manager		20000

Table 3. Formula of calculating salary

3.11. SatisfactionLevel

MaritalStatus, *Overtime*, *Age* and *Salary* are the variables influence employees' satisfaction level. The decision tress as Figure 1 is implemented in this variable.

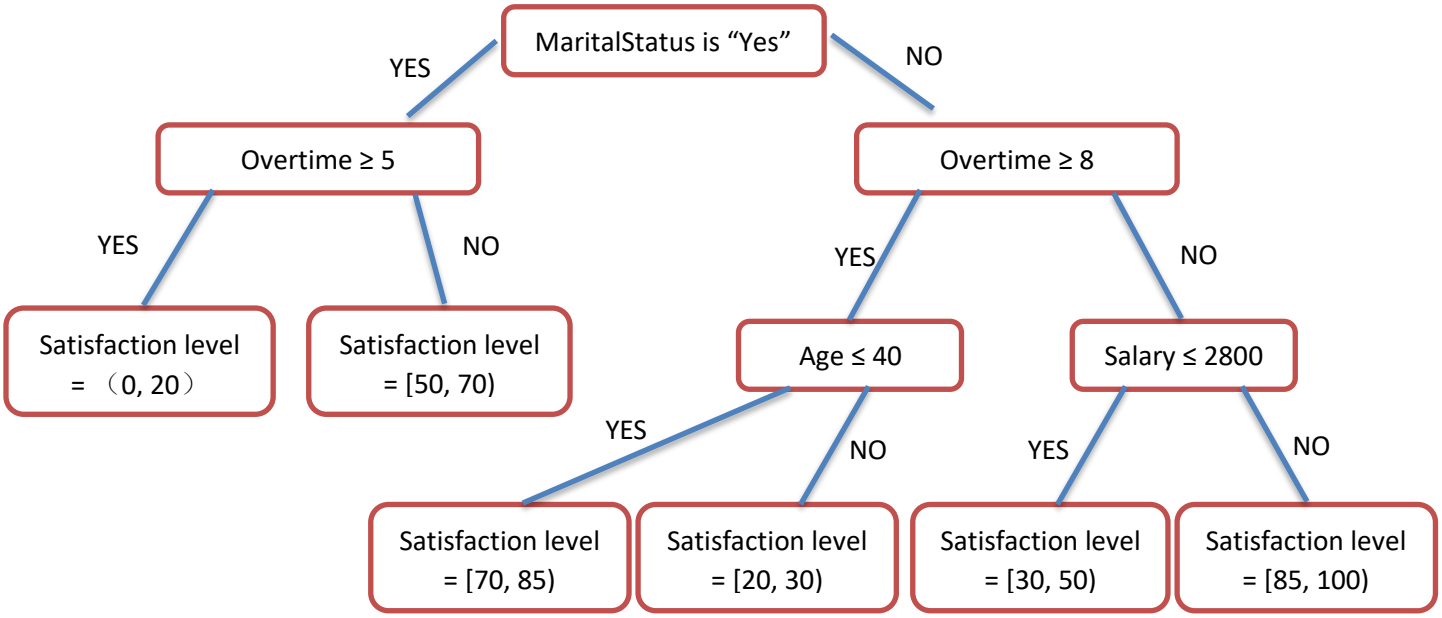

Figure 1: Decision tree of SatisfactionLevel

3.12. EQ & IQ

For *EQ* and *IQ*, the idea is to generate variables that are correlated to *SatisfactionLevel*. It is assumed that *SatisfactionLevel* has a positive relation with *EQ* but negative relation with *IQ*. So, first, the *SatisfactionLevel* is combined with *IQ* and *EQ* in a data frame in order to use Cholesky decomposition to get independence correlation. Then, the correlations are enforced by the pre-defined correlation matrix. The correlations between *(SatisfactionLevel, EQ)* is 0.7 and -0.6 for *(SatisfactionLevel, IQ)*.