**Experiment No.1**

**Title:** Understanding of the Data

**Batch: A4 Roll No.: 1914078 Experiment No.:1**

**Aim**: Understanding of the Data

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Resources needed:** Any RDBMS, EXCEL, Data storage tool

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Theory:**

In order to make data ready for data mining process, data exploration is essential step to develop a high-level understanding of the data. Data exploration includes in detail analysis of attributes and their data values and visualization. It aimed at identifying possible relationship between two or more variables/objects.

Broadly classifying, there are two types of attributes, numeric and categorical.

**Categorical Attribute:**

In categorical, each value represents some kind of category, code, or state. Categorical variables are either nominal or ordinal, depending on the extent of information the numerical coding provides.

The values of a nominal attribute are symbols or names of things. Nominal means “relating to names.”

E.g. hair color and occupation are two attributes describing person objects.

Possible values for hair color are black, brown, blond, red, auburn, gray, and white. For occupation, possible values are teacher, dentist, programmer, farmer etc.

An ordinal attribute is an attribute with possible values that have a meaningful order or ranking among them, but the magnitude between successive values is not known. For example, grade attribute with values A+, A,A-, B, C; Student\_progress attribute with values Good, average , poor. The central tendency of an ordinal attribute can be represented by its mode and its median (the middle value in an ordered sequence), but the mean cannot be defined.

Nominal, binary, and ordinal attributes are qualitative. That is, they describe a feature of an object without giving an actual size or quantity. The values of such qualitative attributes are typically words representing categories.

**Numeric Attributes:**

A numeric attribute is quantitative; that is, it is a measurable quantity, represented in integer or real values. Numeric attributes can be interval-scaled or ratio-scaled.

*Interval-Scaled Attributes:*

Interval-scaled attributes are measured on a scale of equal-size units. The values of interval-scaled attributes have order and can be positive, 0, or negative. Thus, in addition to providing a ranking of values, such attributes allow us to compare and quantify the difference between values.

For example, temperature, humidity attributes

*Ratio-Scaled Attributes:*

A ratio-scaled attribute is a numeric attribute with an inherent zero-point. That is, if a measurement is ratio-scaled, we can speak of a value as being a multiple (or ratio) of another value. In addition, the values are ordered, and we can also compute the difference between values, as well as the mean, median, and mode.

For example, *years of experience*

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

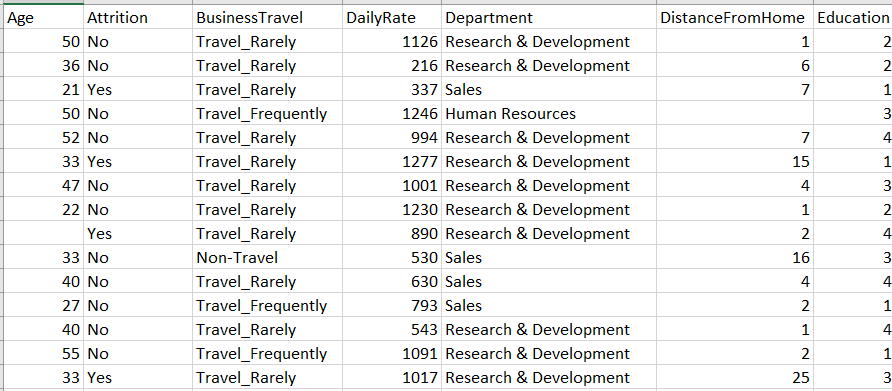
**Procedure / Approach /Algorithm / Activity Diagram:**

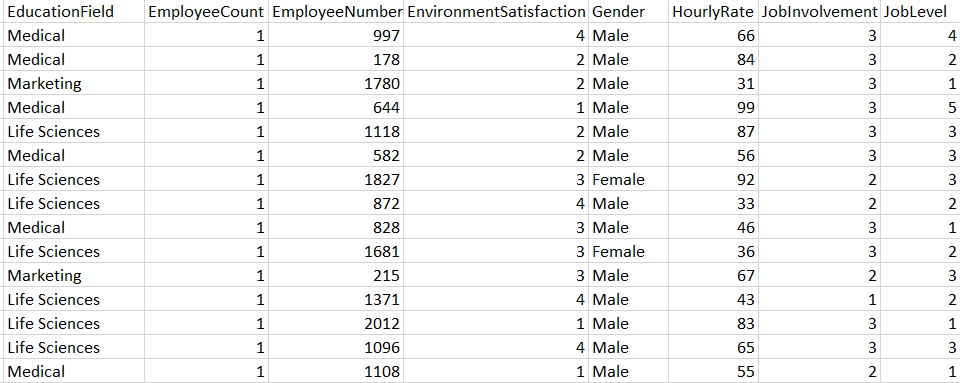
1. Download the large dataset for the purpose of exploration and ensure that dataset has variety of attributes; number of attributes must be at least 25.
2. Identify the category of each attribute from the dataset which you have created.
3. Identify the attributes which can provide any kind of useful information either collectively or as an individual. Also, discuss the about the information provided by the attribute and how it will be computed?

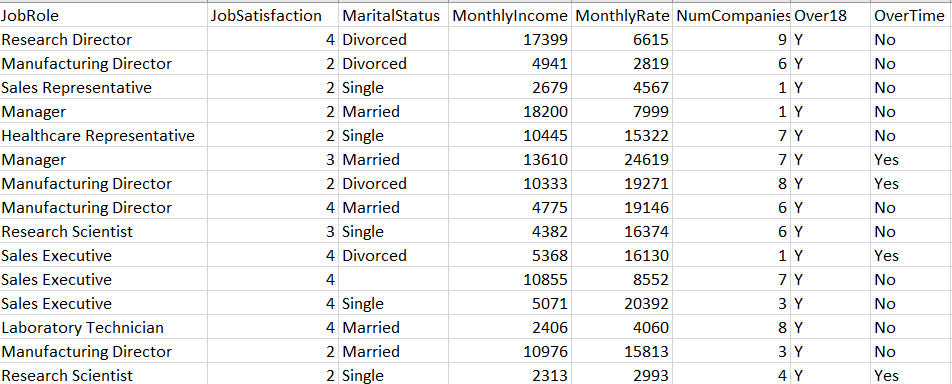
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

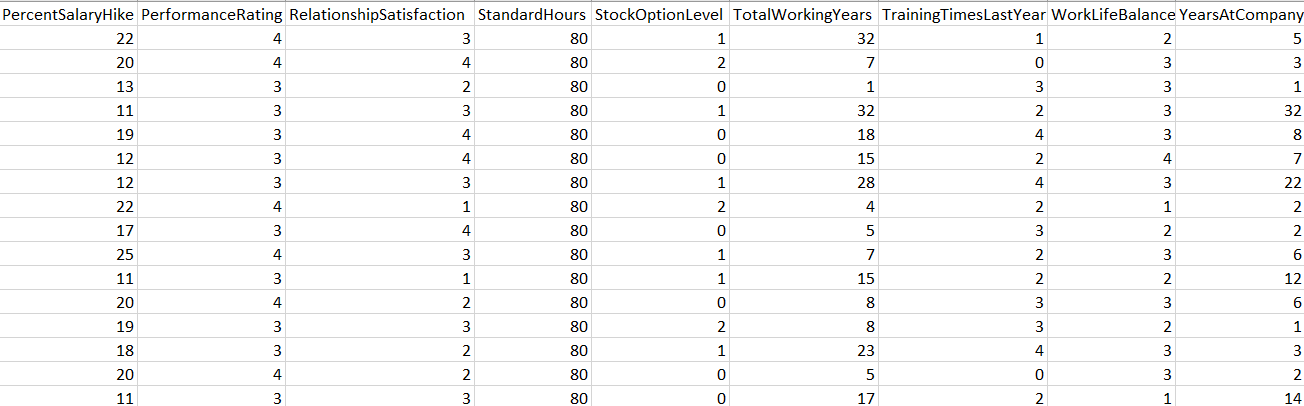
**Results: (Program printout with output / Document printout as per the format)**

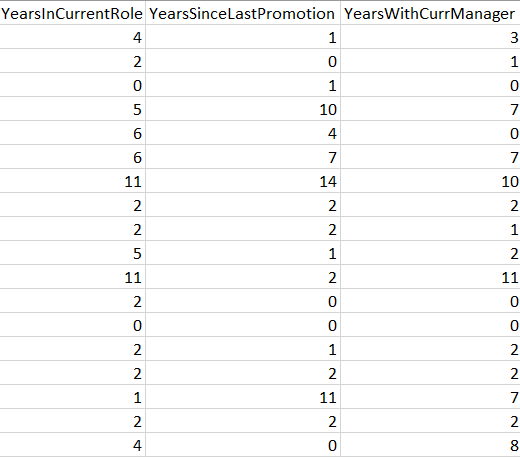
**Task 1**

****

****

****

****

****

**Task 2**

|  |  |  |
| --- | --- | --- |
| **Column** | **Type** | **Attribute** |
| Age | Quantitative | Numeric |
| Attrition | Qualitative | Binary |
| BusinessTravel | Qualitative | Nominal |
| DailyRate | Quantitative | Numeric |
| Department | Qualitative | Nominal |
| DistanceFromHome | Quantitative | Numeric |
| Education | Quantitative | Numeric |
| EducationField | Qualitative | Nominal |
| EmployeeCount | Quantitative | Continuous |
| EmployeeNumber | Qualitative | Nominal |
| EnvironmentSatisfaction | Quantitative | Ratio |
| Gender | Qualitative | Binary |
| HourlyRate | Quantitative | Numeric |
| JobInvolvement | Quantitative | Ratio |
| JobLevel | Quantitative | Ratio |
| JobRole | Qualitative | Nominal |
| JobSatisfaction | Quantitative | Ratio |
| MaritalStatus | Qualitative | Binary |
| MonthlyIncome | Quantitative | Numeric |
| MonthlyRate | Quantitative | Numeric |
| NumCompaniesWorked | Quantitative | Numeric |
| Over18 | Qualitative | Binary |
| OverTime | Qualitative | Binary |
| PercentSalaryHike | Quantitative | Numeric |
| PerformanceRating | Quantitative | Ratio |
| RelationshipSatisfaction | Quantitative | Ratio |
| StandardHours | Quantitative | Continuous |
| StockOptionLevel | Quantitative | Ratio |
| TotalWorkingYears | Quantitative | Numeric |
| TrainingTimesLastYear | Quantitative | Numeric |
| WorkLifeBalance | Quantitative | Ratio |
| YearsAtCompany | Quantitative | Numeric |
| YearsInCurrentRole | Quantitative | Numeric |
| YearsSinceLastPromotion | Quantitative | Numeric |
| YearsWithCurrManager | Quantitative | Numeric |

**Task 3**

* MonthlyIncome, MonthlyRate, NumCompaniesWorked are important fields for knowing more information about the employee resume.
* EnvironmentSatisfaction, WorkLifeBalance, RelationshipSatisfaction, JobInvolvement, JobSatisfaction and StandardHours are key fields to know about the work ethic about the employee.
* YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, TotalWorkingYears, TrainingTimesLastYear and PerformanceRating gives important information regarding the efficiency of the employee.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Questions:**

* + 1. Compare Discrete and Continuous Attributes. Give at least 5 examples of each.

**ANS:**

|  |  |
| --- | --- |
| **DISCRETE DATA** | **CONTINOUS DATA** |
| Discrete data is the type of data that has clear spaces between values. | Continuous data is data that falls in a continuous sequence. |
| Discrete data is countable | Continuous data is measurable. |
| Discrete data is information that can only take certain values. These values don’t have to be whole but they are fixed values. | Continuous data is data that can take any value. |
| This type of data is often represented using tally charts, bar charts or pie charts. | This data is best shown on a line graph as this type of graph can show how the data changes over a given period of time. |
| **EG:** Days with temperature > specific degree, no. of students arrived late, whether a guy will take policy or not, zip codes, profession. | **EG:** temperature, height, weight, avg daily temperature, litres of water in a tank, time per delivery, speed. |

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Outcomes:** CO1: Summarize the data

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Conclusion:** We were able to classify the attributes/features of dataset into types as well as on basis of usefulness.

**Grade: AA / AB / BB / BC / CC / CD /DD**

Signature of faculty in-charge with date

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**References:**

Books/ Journals/ Websites:

1. Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3nd Edition