**Excercise-7**

**Part-A:- Data Transformation**

1. Create at dataset (excel/vector)/download of your choice having at least 200 observations (numerical).

2. Implement normalization as

a. Scale the data in between 0-1

(i) Calculate the mean and SD for both dataset (original and post normalized

dataset)

b. standardize the data (Z-normalization)

(i) Calculate the mean and SD for both dataset (original and post normalized

dataset)

c. scale the in between given range, i.e 10-30

(i) Calculate the mean and SD for both dataset (original and post normalized

dataset)

**Part-B:**

1. Create a dataset in excel (name, Country\_region\_id, salary, age, smoking), where:

Name: string

Country\_region\_id: categorical (fill numeric value only: against four categories), ex

(north:1, South:2, East:3, West:4)

Salary: numeric (min-400, max 15000)

Age: numeric (min-10, max 100)

Smoking: Categorical (yes, no)

2. Fill some values (at least 20) in excel.

3. Load the data (either using R/Python)

4. Calculate and display number of person “Country\_region\_id” wise. Such as (ex):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Country\_region\_id | 1 | 2 | 3 | 4 |
| No of persons | 5 | 11 | 1 | 3 |

5. Calculate and display number of person “who smoke”.

6. Displays the data type of each field using command. Observe that is it giving correct result?

|  |  |  |
| --- | --- | --- |
| name | string | ? |
| Country\_region\_id | categorical | ? |
| Salary | numeric | ? |
| Age | numeric | ? |
| smoking | Categorical | ? |

7. if not then convert it, such as (Country\_region\_id into categorical). 1,2,3,4 values do not have numeric meaning. These are category.

8. Similarly, Smoking must be categorical (result from step-6 might be displaying it string). Convert it categorical.

9. Now again, Calculate and display number of person “Country\_region\_id” wise.

10. Now again, Calculate and display number of person “who smoke”.

11. Display all the entered data.

12. Rescale numeric filed (age and salary).

13. Again display complete data.

14. Display data for fields: age and salary.

15. Calculate the mean of age.

16. Display complete data (all fields) where age is greater than age\_mean.