DATA PREPROCESCING

Cleaning the Reason for Absence

Reason for Absence	Date	Transportation Expense	Distance to Work	Age	Daily Work Ave
26	07/07/2015	289	36	33	23
0	14/07/2015	118	13	50	23
23	15/07/2015	179	51	38	23
7	16/07/2015	279	5	39	23
23	23/07/2015	289	36	33	23
23	10/07/2015	179	51	38	23
22	17/07/2015	361	52	28	23
23	24/07/2015	260	50	36	23
19	06/07/2015	155	12	34	23
22	13/07/2015	235	11	37	23

	Reason_1	Reason_2	Reason_3	Reason_4	Date	Tra
0	0	0	0	1	07/07/2015	
1	0	0	0	0	14/07/2015	
2	0	0	0	1	15/07/2015	
3	1	0	0	0	16/07/2015	
4	0	0	0	1	23/07/2015	

VARIOUS DISEASES	1	Certain infectious and parasitic diseases
	2	Neoplasms
	3	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
	4	Endocrine, nutritional and metabolic diseases
	5	Mental and behavioral disorders
	6	Diseases of the nervous system
	7	Diseases of the eye and adnexa
	8	Diseases of the ear and mastoid process
	9	Diseases of the circulatory system
	10	Diseases of the respiratory system
	11	Diseases of the digestive system
	12	Diseases of the skin and subcutaneous tissue
	13	Diseases of the musculoskeletal system and connective tissue
	14	Diseases of the genitourinary system
PREGNANCY	15	Pregnancy, childbirth and the puerperium
	16	Certain conditions originating in the perinatal period
	17	Congenital malformations, deformations and chromosomal abnormalities
POISONING	18	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
	19	Injury, poisoning and certain other consequences of external causes
	20	External causes of morbidity and mortality
	21	Factors influencing health status and contact with health services
LIGHT REASON	22	Patient follow-up
	23	Medical consultation
	24	Blood donation
	25	Laboratory examination
	26	Unjustified absence
	27	Physiotherapy
	28	Dental consultation

289	36	33	239.554	30	1	2	1	2		

DATA PREPROCESCING

Cleaning the Reason for Absence

Reason for Absence	Date	Transportation Expense	Distance to Work	Age	Daily W
26	07/07/2015	289	36	33	
0	14/07/2015	118	13	50	
23	15/07/2015	179	51	38	
7	16/07/2015	279	5	39	
23	23/07/2015	289	36	33	
23	10/07/2015	179	51	38	
22	17/07/2015	361	52	28	
23	24/07/2015	260	50	36	
19	06/07/2015	155	12	34	
22	13/07/2015	235	11	37	

	Reason_1	Reason_2	Reason_3	Reason_4	Date
0	0	0	0	1	07/07/2015
1	0	0	0	0	14/07/2015
2	0	0	0	1	15/07/2015
3	1	0	0	0	16/07/2015
4	0	0	0	1	23/07/2015

1. Create a separate dataframe, containing dummy values for ALL available reasons

```
reason_columns = pd.get_dummies(df['Reason for Absence'], drop_first = True)
```

2. Split reason_columns into 4 types

```
reason_type_1 = reason_columns.loc[:,1:14].max(axis=1)
reason_type_2 = reason_columns.loc[:,15:17].max(axis=1)
reason_type_3 = reason_columns.loc[:,18:21].max(axis=1)
reason_type_4 = reason_columns.loc[:,22:].max(axis=1)
```

3. To avoid multicollinearity, drop the 'Reason for Absence' column from df

```
df = df.drop(['Reason for Absence'], axis = 1)
```

4. Concatenate df and the 4 types of reason for absence

```
df = pd.concat([df, reason_type_1, reason_type_2, reason_type_3, reason_type_4], axis = 1)
```

5. Assign names to the 4 reason type columns

```
column_names = ['Date', 'Transportation Expense', 'Distance to Work', 'Age',
                'Daily Work Load Average', 'Body Mass Index', 'Education', 'Children',
                'Pet', 'Absenteeism Time in Hours', 'Reason_1', 'Reason_2', 'Reason_3', 'Reason_4']
df.columns = column_names
```

6. Re-order the columns in df

```
column_names_reordered = ['Reason_1', 'Reason_2', 'Reason_3', 'Reason_4', 'Date', 'Transportation
Expense',
                          'Distance to Work', 'Age', 'Daily Work Load Average', 'Body Mass Index', 'Education',
                          'Children', 'Pet', 'Absenteeism Time in Hours']
df = df[column_names_reordered]
```