Employee Absenteeism

Manjunath Nagendra

10 January 2019

Contents

Chapter 1

Introduction

1.1 Problem Statement

XYZ is a courier company. As we appreciate that human capital plays an important role in collection, transportation and delivery. The company is passing through genuine issue of Absenteeism. Absenteeism here is based on many attributes that the employee exhibits. The objective of this project is to address the following questions:

a. What changes company should bring to reduce the number of absenteeism?
b. How much losses every month can we project in 2011 if same trend of absenteeism continues?

1.2 Data

Our task is to build a regression model which predicts Absenteeism of employees in hours depending on certain attributes which they exhibit. Given below is a sample of data set about employee absenteeism.

Table 1.1 Employee Absenteeism Data (Columns: 1-10)

ID	Reason for absence	Month of absence	Day of the week	Seasons	Transportation expense	Distance from Residence to Work	Service time	Age	Work load Average/day
11	26	7	3	1	289	36	13	33	239554
36	0	7	3	1	118	13	18	50	239554
3	23	7	4	1	179	51	18	38	239554
7	7	7	5	1	279	5	14	39	239554
11	23	7	5	1	289	36	13	33	239554
3	23	7	6	1	179	51	18	38	239554
10	22	7	6	1		52	3	28	239554
20	23	7	6	1	260	50	11	36	239554
14	19	7	2	1	155	12	14	34	239554
1	22	7	2	1	235	11	14	37	239554
20	1	7	2	1	260	50	11	36	239554
20	1	7	3	1	260	50	11	36	239554
20	11	7	4	1	260	50	11	36	239554
3	11	7	4	1	179	51	18	38	239554

Table 1.2: Employee Absenteeism Data (Columns: 11-21)

Hit target	Disciplinary failure	Education	Son	Social drinker	Social smoker	Pet	Weight	Height	Body mass index	Absenteeism time in hours
97	0	1	2	1	0	1	90	172	30	4
97	1	1	1	1	0	0	98	178	31	0
97	0	1	0	1	0	0	89	170	31	2
97	0	1	2	1	1	0	68	168	24	4
97	0	1	2	1	0	1	90	172	30	2
97	0	1	0	1	0	0	89	170	31	
97	0	1	1	1	0	4	80	172	27	8
97	0	1	4	1	0	0	65	168	23	4
97	0	1	2	1	0	0	95	196	25	40
97	0	3	1	0	0	1	88	172	29	8
97	0	1	4	1	0	0	65	168	23	8
97	0	1	4	1	0	0	65	168	23	8
97	0	1	4	1	0	0	65	168	23	8
97	0	1	0	1	0	0	89	170	31	1

Attribute Information:

1. Individual identification (ID)

2. Reason for absence (ICD).

Absences attested by the International Code of Diseases (ICD) stratified into 21 categories (I to XXI) as follows: I certain infectious and parasitic diseases

II Neoplasms

III Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism

IV Endocrine, nutritional and metabolic diseases

V Mental and behavioral disorders

VI Diseases of the nervous system

VII Diseases of the eye and adnexa

VIII Diseases of the ear and mastoid process

IX Diseases of the circulatory system

X Diseases of the respiratory system

XI Diseases of the digestive system

XII Diseases of the skin and subcutaneous tissue

XIII Diseases of the musculoskeletal system and connective tissue

XIV Diseases of the genitourinary system XV Pregnancy, childbirth and the puerperium

XVI Certain conditions originating in the perinatal period

XVII Congenital malformations, deformations and chromosomal abnormalities XVIII Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified

XIX Injury, poisoning and certain other consequences of external causes

XX External causes of morbidity and mortality

XXI Factors influencing health status and contact with health services. And 7 categories without (CID) patient follow-up (22), medical consultation (23), blood donation (24), laboratory examination (25), unjustified absence (26), physiotherapy (27), dental consultation (28).

3. Month of absence

4. Day of the week (Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6))

5. Seasons (summer (1), autumn (2), winter (3), spring (4))

6. Transportation expense

7. Distance from Residence to Work (kilometers)

8. Service time

9. Age

10. Work load Average/day

11. Hit target

12. Disciplinary failure (yes=1; no=0)

13. Education (high school (1), graduate (2), postgraduate (3), master and doctor (4))

14. Son (number of children)

15. Social drinker (yes=1; no=0)

16. Social smoker (yes=1; no=0)

17. Pet (number of pet)

18. Weight

19. Height

20. Body mass index

21. Absenteeism time in hours (target)

Chapter 2

Methodology

2.1 Pre Processing

Before we even process the data or build a model above it, we need to have a good understanding about our dataset. Initial exploration of data set is the key step to building a good model. Initial data exploration involves cleaning of data like removing or replacing missing values, discovering few patterns to maximize insight, understanding the dimensions and visualization through plots and graphs. This whole process is termed as Exploratory Data Analysis.

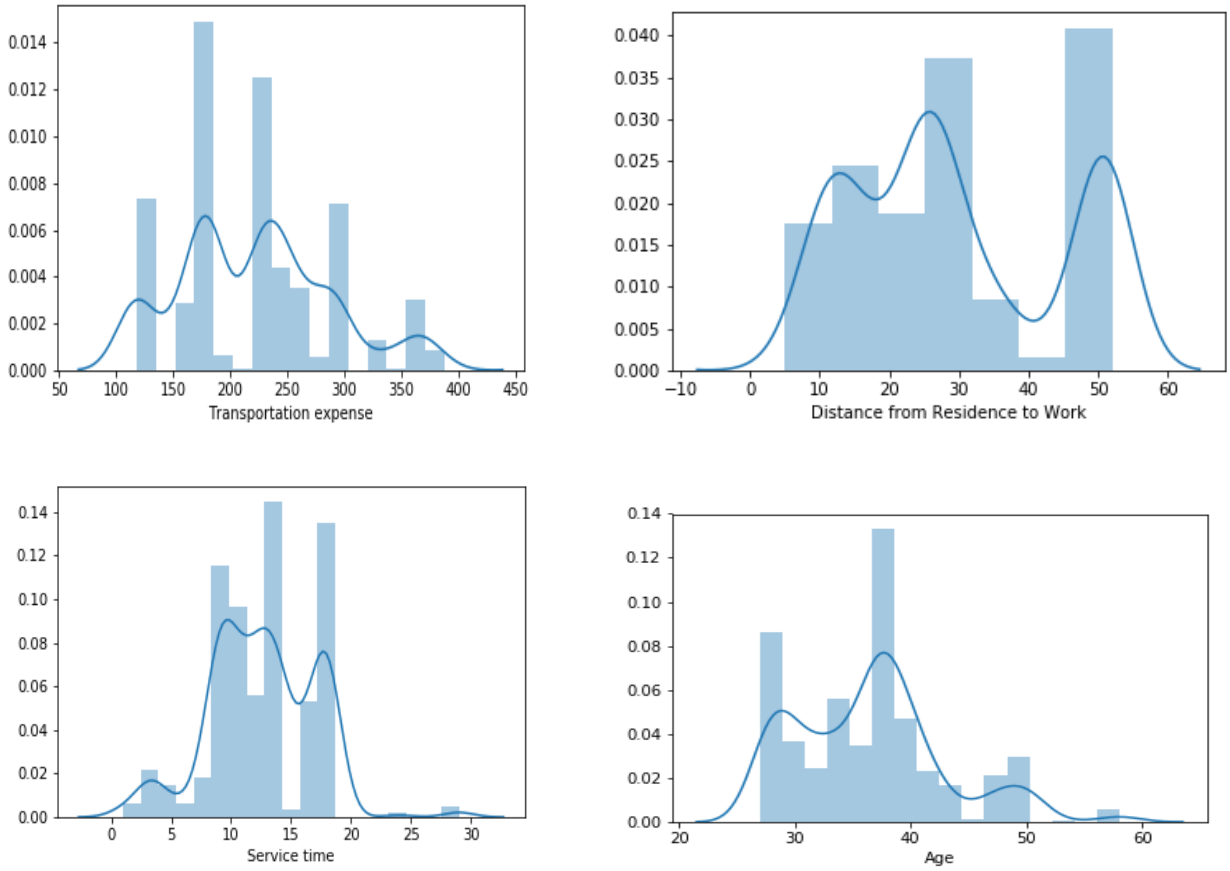
To begin with, let's start analyzing the data type present in the data set from Table 2.1. On an over view, we observe that all data present here are numeric and we have missing values to deal with.

Table 2.1: Employee Absenteeism dataset description

```
RangeIndex: 740 entries, 0 to 739
Data columns (total 21 columns):
ID                                740 non-null int64
Reason for absence                737 non-null float64
Month of absence                  739 non-null float64
Day of the week                   740 non-null int64
Seasons                           740 non-null int64
Transportation expense            733 non-null float64
Distance from Residence to Work   737 non-null float64
Service time                      737 non-null float64
Age                               737 non-null float64
Work load Average/day             730 non-null float64
Hit target                        734 non-null float64
Disciplinary failure              734 non-null float64
Education                         730 non-null float64
Son                               734 non-null float64
Social drinker                    737 non-null float64
Social smoker                     736 non-null float64
Pet                               738 non-null float64
Weight                            739 non-null float64
Height                            726 non-null float64
Body mass index                   709 non-null float64
Absenteeism time in hours         718 non-null float64
dtypes: float64(18), int64(3)
```

Depending upon the nature of the data, dataset can be categorized as continuous and categorical dataset. Variables which could be categorized as continuous variables are 'Transportation expense', 'Distance from Residence to Work', 'Service time', 'Age', 'Work load Average/day ','Hit target', 'Weight', 'Height', 'Body mass index', 'Absenteeism time in hours' and rest of the variables which could be categorized are ‘ID’, ‘Reason for absence’, ‘Month of absence’, ‘Day of the week’,’ Seasons’, ‘Disciplinary failure’, ‘Education’, ‘Son’, ‘Social drinker’, ‘Social smoker’, ‘Pet’ fall under categorical variables.

From Figure 2.1 we can make out most of the continuous variables are normally distributed.



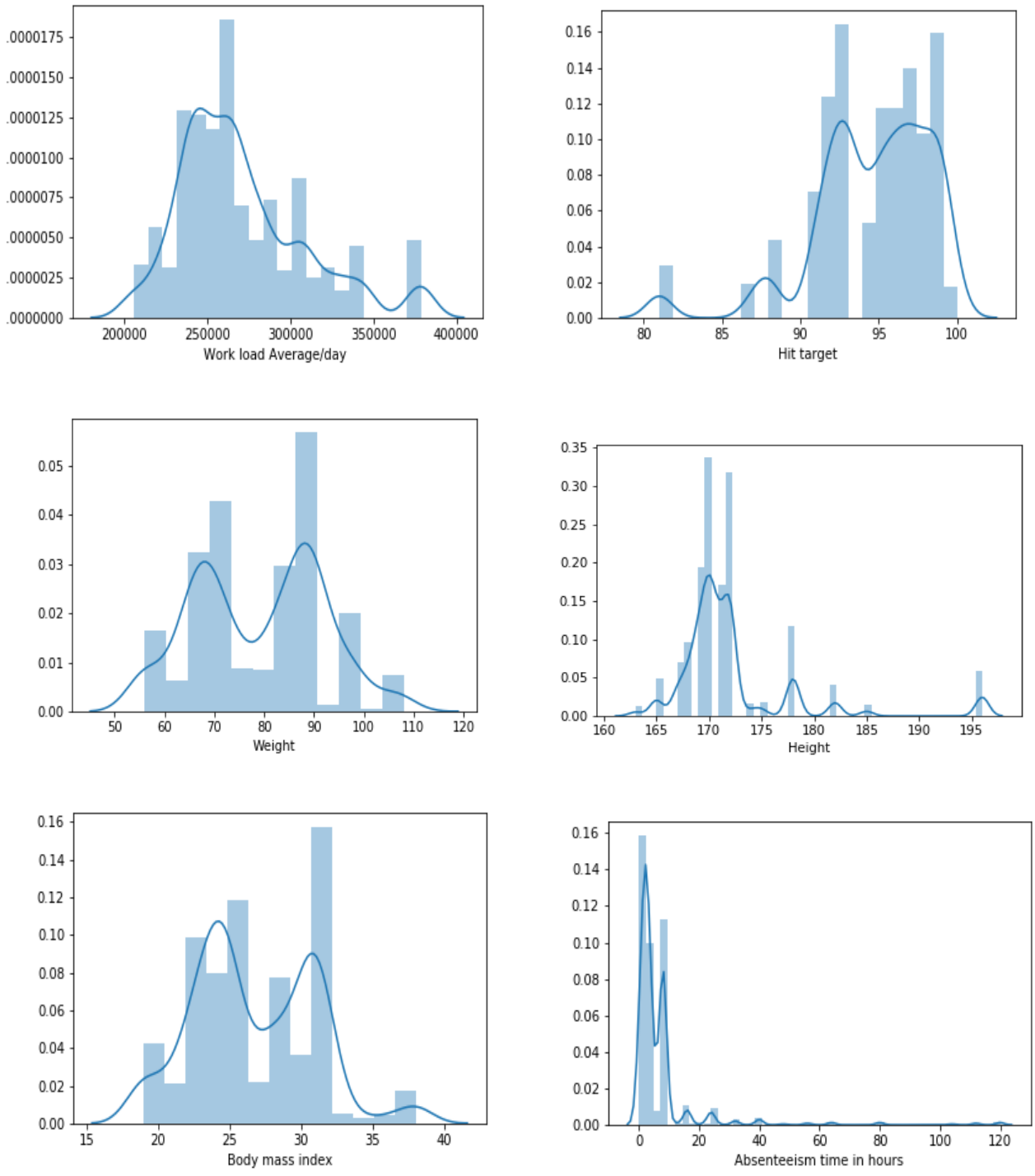


Figure 2.1: Distribution plot with kernel density for Employee Absenteeism Dataset

2.1.1 Missing Values

Addressing missing values is an important task of cleaning data before building a model on them. Missing values might happen either because of Human error, or clients refuse to answer, optional box in questionnaire or system would have failed to capture the data. Cleaning of data involves checking for missing values under independent variables and taking care of them. First we need to check the percentage of missing values for all predictors. If predictors have above 30 % of missing values those predictors should be considered to be deleted but end results might be assumption. Otherwise, the missing values should be imputed by central statistical way i.e median, mean (continuous variables) or distance based method Knn-imputation. Table 2.2 represents missing value percentage for all predictors. After testing Knn-imputation is chosen as the best method to impute missing values for this project.

Table 2.2: Missing value percentage of independent variables

	Variables	Missing_Percentage
0	Body mass index	4.189189
1	Absenteeism time in hours	2.972973
2	Height	1.891892
3	Work load Average/day	1.351351
4	Education	1.351351
5	Transportation expense	0.945946
6	Son	0.810811
7	Disciplinary failure	0.810811
8	Hit target	0.810811
9	Social smoker	0.540541
10	Age	0.405405
11	Reason for absence	0.405405
12	Service time	0.405405
13	Distance from Residence to Work	0.405405
14	Social drinker	0.405405
15	Pet	0.270270
16	Weight	0.135135
17	Month of absence	0.135135
18	Seasons	0.000000
19	Day of the week	0.000000
20	ID	0.000000

2.1.2 Outlier Analysis

Next step is to check for outliers as they impact on outcome of results. We will use boxplot to visualize outliers for all continuous variables as illustrated in Figure 2.2. From Figure 2.2 we can make out few of the continuous variables have outliers. Once we check for outliers, we either need to drop those values or impute as they contain values falling away from the actual bunch of values. Outliers are replaced with null values which creates missing values for respective predictors. Here in this project once outliers are replaced with null values, those null values are filled using Knn-imputation.

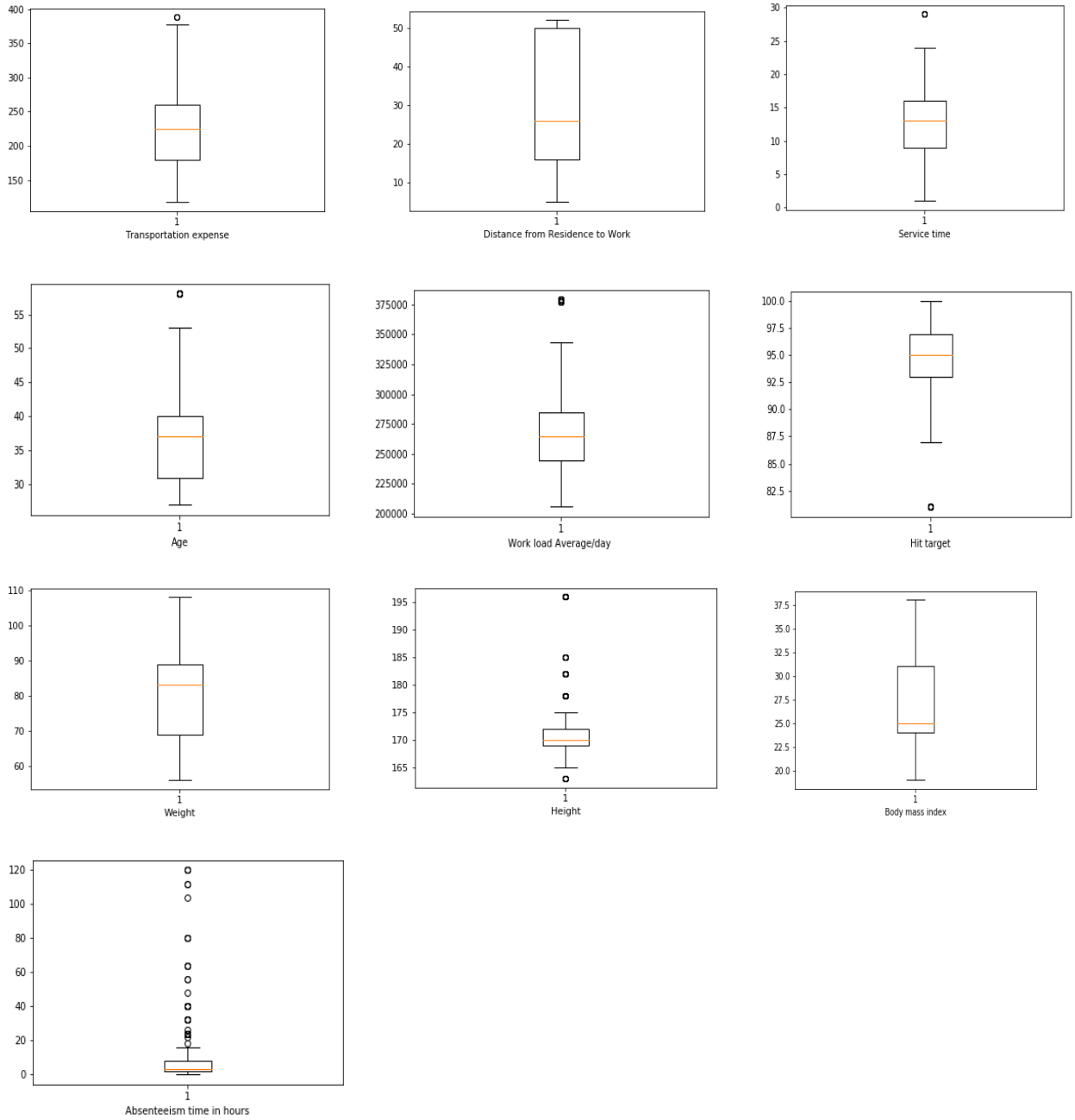


Figure 2.2: Boxplot for each predictor

2.1.3 Feature Selection

Before we get into model building, we need to assess the importance of each predictor or independent variables. Then select only relevant variables needed for model construction by eliminating other variables which does not add any value for our classification problem.

The methods used here to do Feature Selection are classified as:

a. **Correlation Analysis**. For all the continuous variable we will build a head map to inspect which variables are highly correlated to each other. We will use corr() method to calculate the correlation between each continuous variables and pass those values to head map to visualize them as in Figure 2.3.
b. **Chi-Square test of independence**: For all the categorical variables we will use chi-square test to test relationship between 2 categorical variables. From chi-square test, we make 2 hypothesis one is *Null Hypothesis*: 2 variables (one of them is target variable) are independent variables and other is *Alternate hypothesis:* 2 variables (one of them is target variable) are not independent variables. Variables which has p-value above 0.05 will be dropped for model construction further.

Figure 2.3: Heat Map to show Correlation Matrix of Continuous Variables

From Figure 2.3, we can see that 2 variable are highly correlated like *Weight* and *Body mass index* where *Weight* is most likely calculated from *Body mass index* or vice versa. So we drop any one of the correlated pair for further modelling. Here I am considering variable *Body mass index* to be dropped as they won't add any value during model development.

2.1.4 Feature Scaling

Another aspect under Pre-Processing is to normalize the data in the dataset before moving to model building phase so that the data are in proportion with one another. We can do Feature Scaling on continuous variables by 2 methods:

a. Normalization - convert all categorical variable values between 0 and 1
b. Standardization (Z-score) – Standardization can be applied on continuous variables only when they are normally distributed.

We can observe From Figure 2.1 that most variables are normally distributed. Here I am choosing Normalization method to get values of all the predictors under same scale within a range from 0 and 1.

After Data Pre-Processing, Dataset will be

a. Free of missing values.
b. Outliers will be replaced with values within variable/feature range.
c. Only those features/variables are selected which add value for predicting target variable.
d. All variables values are normalized/standardized.

After this stage, dataset is ready to build Machine Learning/Statistical model on them to predict our target variable.

2.2 Modeling

2.2.1 Model Selection

Since our target variable is continuous in nature, we are dealing with regression problem here. So we can use below algorithms to build a regression model on our pre-processed dataset and predict our target variable how many hours does employee go absent.

a. Decision Tree Regressor.
b. Random Forest Regressor.
c. Linear Regression.

Let's go one by one and check which algorithm performs better on this dataset to predict our target variable. Performance of above mentioned algorithms will be calculated with help of *Error Metrics* which we will discuss further.

2.2.1.1 Decision Tree Regressor

As Decision Tree is a predictive model based on series of Boolean test, can be used for regression which uses concept of *Information Gain* i.e variables with highest *Information gain* is used to spilt at root node first.

2.2.1.2 Random Forest Regressor

Random Forest is an ensemble technique which uses n-Decision Trees to predict target variable which works on the concept of bagging. It uses *gini-index* to select parent node and split from randomly selected small set of variables (sqrt (total independent variable)). From the n-trees generated, new test case will traverse tree according to the condition it carries and displays relevant output.

2.2.1.3 Linear Regression

Linear Regression is used only for regression purpose. Here regression coefficient for independent variable is calculated for each category. Statistical technique Least Square Estimators are used here.

Chapter 3

Conclusion

3.1 Model Evaluation

Since we have few models now to predict our target variable, we need to decide which one to choose. There are several criteria exist for evaluating and comparing model performance. As we have regression model here, we will choose MAE (Mean Absolute Error) and RMSE (Root Mean Square Error) to evaluate our model.

3.1.1 Mean Absolute Error

MAE is one of the known metric to evaluate a regression model once we train the model and predicted the test values. Error is the difference between actual value and predicted value of the target variable. After that we take the absolute value of the error and take mean of all the values of the error.

$$MAE = \frac{1}{n}\sum|y_j - \overline{y_j}|$$

3.1.2 Root Mean Square Error

Root Mean Square Error is another important metric to evaluate regression model. It is similar to MAE but it is similar to standard deviation of error.

$$RMSE = \sqrt{1/n\sum\left(y_j - \overline{y_j}\right)^2}$$

MAE and RMSE is one of the best metric because both displays the predicted error but does not consider direction into account.

So let's see how our models performed with these two metrics considered.

Note: I have completed building model for this project both in R and Python. But I am publishing scores here in this report of only the best though difference between both are very minimal. Also note all the values are normalized.

Decision Tree Regression

Decision Tree Regressor	
MAE	0.13
RMSE	0.18

Random Forest Regression

Random Forest Regressor	
MAE	0.12
RMSE	0.17

Linear Regression

Linear Regression	
MAE	0.14
RMSE	0.18

3.2 Model Selection

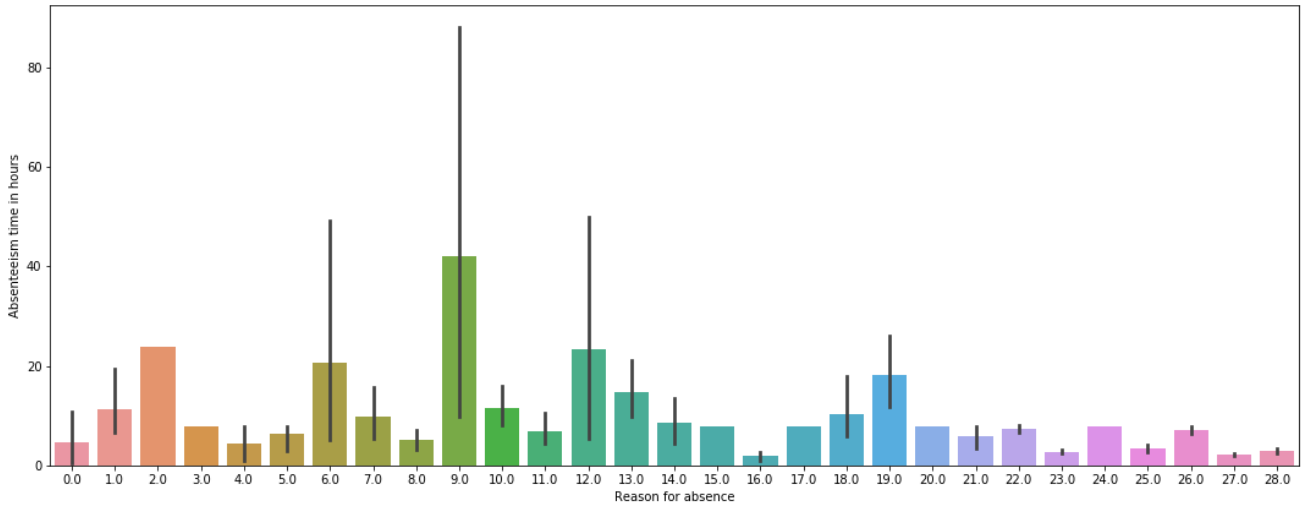
Considering both MAE and RMSE, we can choose any of the above algorithms as the error difference are minimal but I am considering Random Forest Regressor for this model as it has the minimum value of RMSE.

Using the above model we can predict the Employee Absenteeism in hours give the necessary predictors.

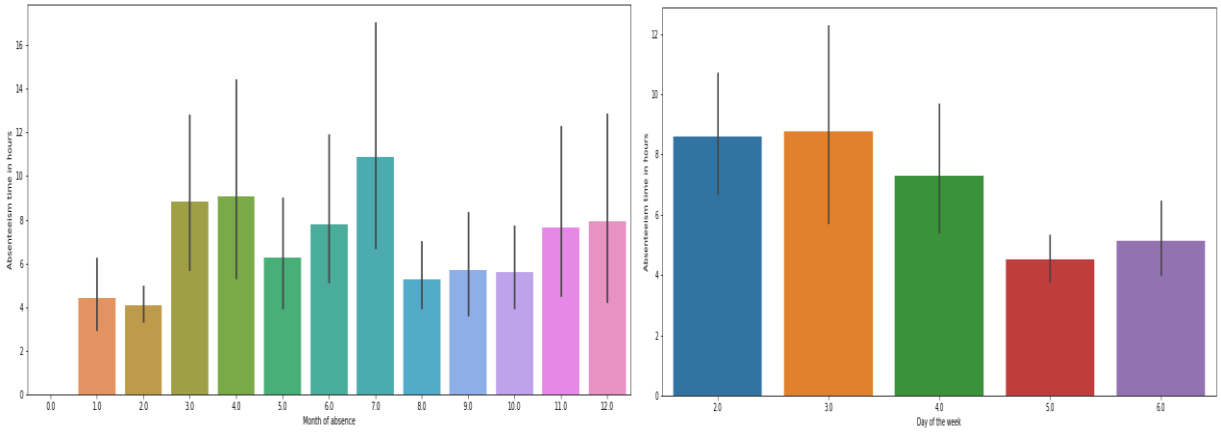
3.3 Answer to asked question

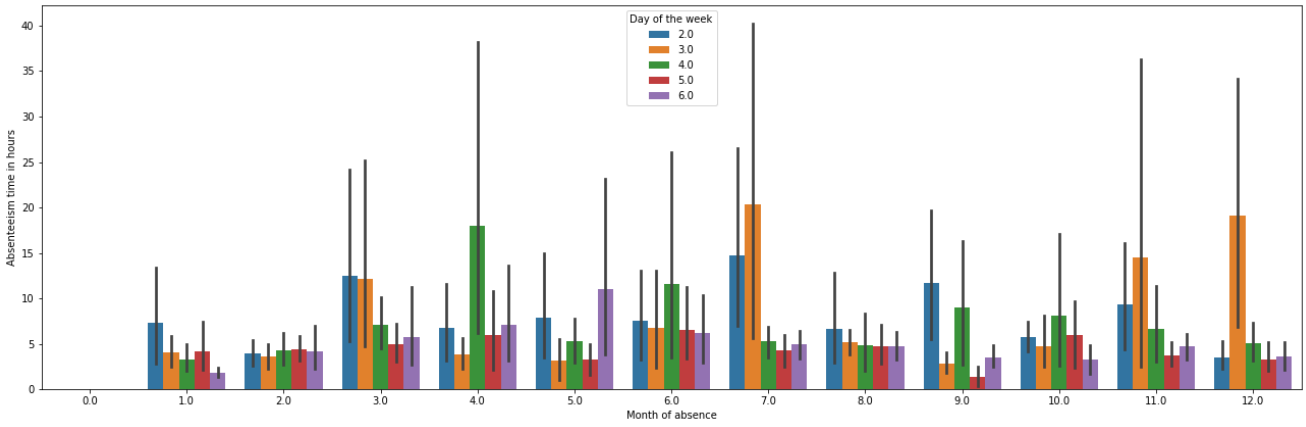
What changes company should bring to reduce the number of absenteeism?

1. Employees are absent as they are more prone to Diseases of the circulatory system (code 9) as being the reason for absenteeism. Company must investigate on this health issue. And also employees are not much absent for the reasons which are generic/follow up (last 7) in nature.

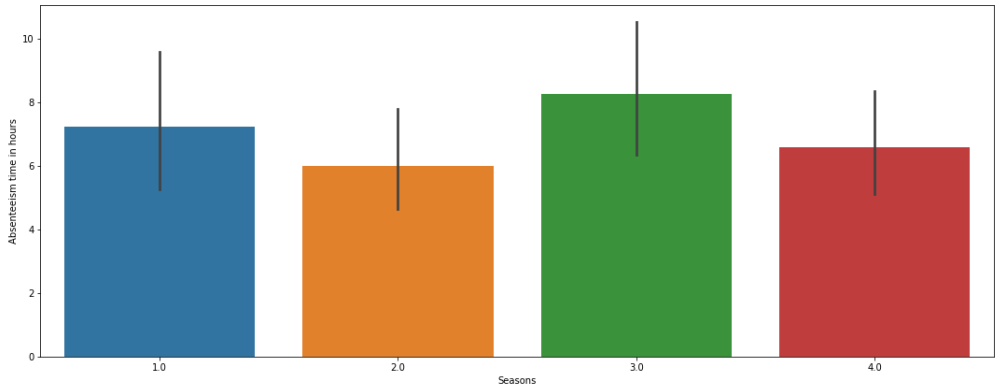


2. Employees seem to be absent most likely in the month of July. And they happen to be more absent more on first two days of the week (Monday and Tuesday).

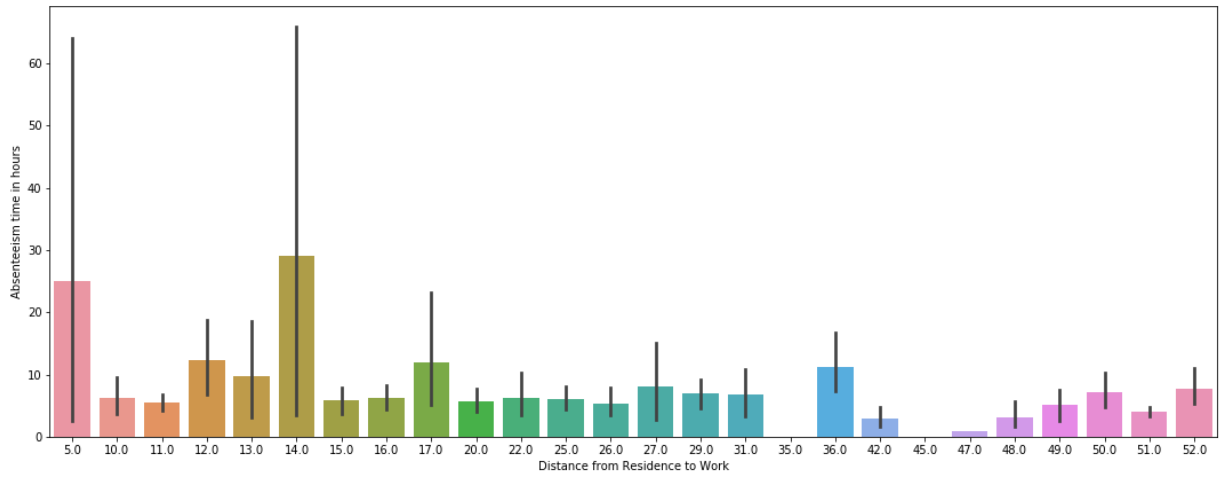




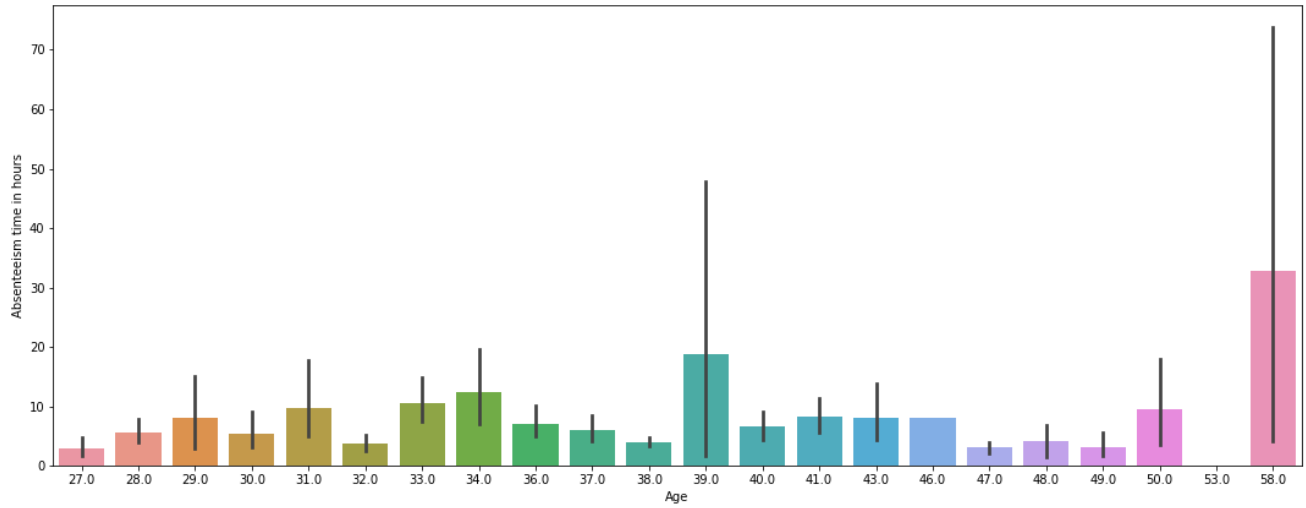
3. Employees are most likely more absent during winter season.



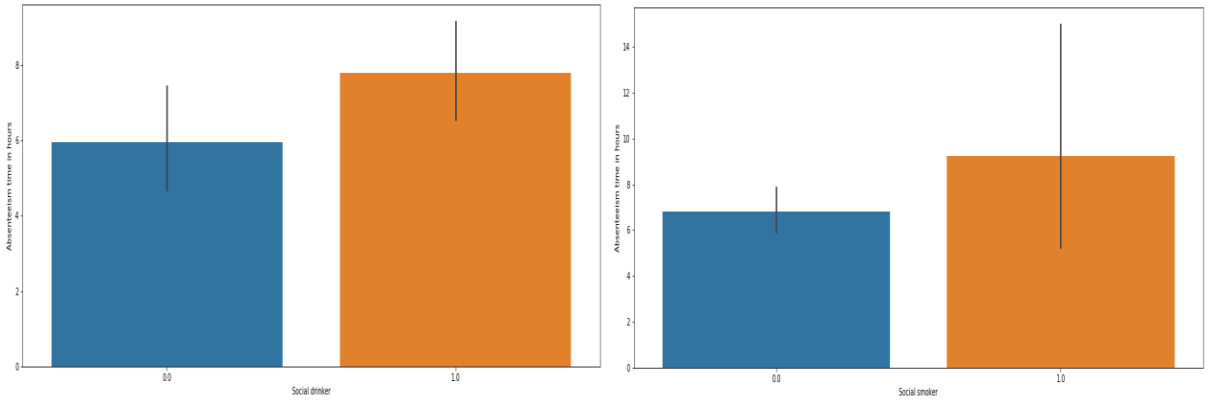
4. Employees are most likely to be absent when their residence is close to work which leads to frequent breaks.



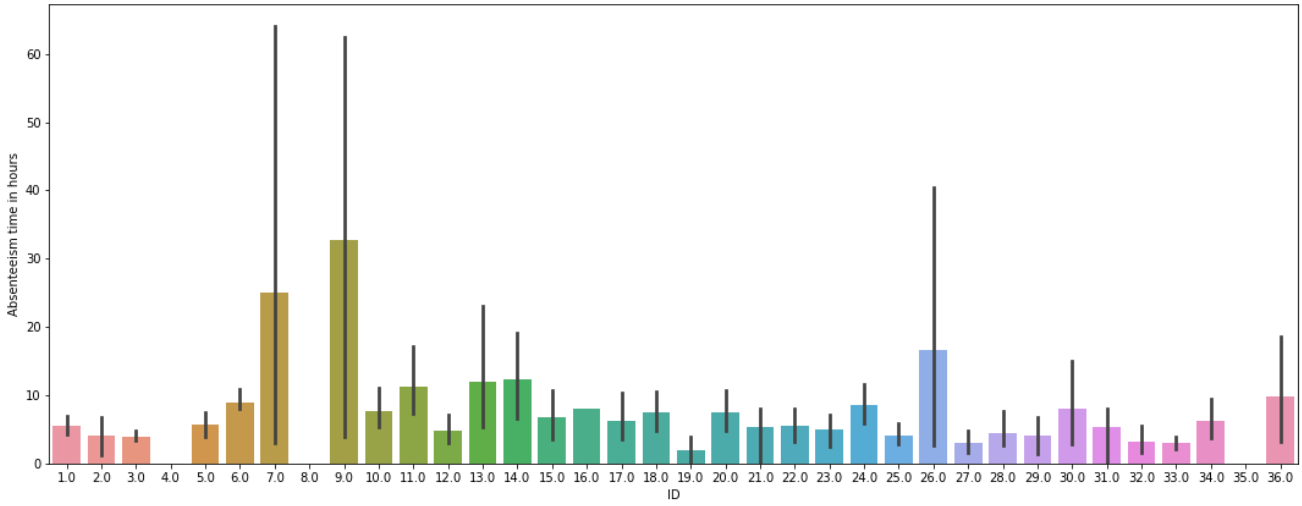
5. Employees who are old above 55 should be taken care of as they are most likely to be absent.



6. Employees who are most likely a social drinker and smoker are tend to be absent more as they will be taking much breaks.



7. Company should reward employees who are never absent like ID 4,8, 35 which in turn motivate other employees not to skip work.



8. Employees who have more than 2 children are tend to be more absent.