Absenteeism time Analysis

Sanket Agrawal

13th April, 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. The company has shared it dataset and requested to have an answer on the following areas:

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

1.2 Data

Dataset Details: Dataset is a time-series multivariate with a total number of 20 independent variables and one dependent variables. Out of twenty independent variables there are ten categorical variables and 10 continuous variables. The dataset also has missing values.

Attribute Information:
1Individual 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 behavioural 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 musculo-skeletal 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)
XXII. patient follow-up
XXIII. medical consultation
XXIV. blood donation
XXV. laboratory examination
XXVI. unjustified absence
XXVII. physiotherapy
XXVIII. dental consultation
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)

Sample data: The table 1.1 shows the a sample instance of the data.

Table 1.1: Sample instance from the dataset

ID	Reason for absence	Month of absence	Day of the week	Seasons	Transportation expense	Distance from Residence to Work
11	26	7	3	1	289	36
Service time	**Age**	**Work load Average/day**	**Hit target**	**Disciplinary failure**	**Education**	**Son**
13	33	239554	97	0	1	2
Social drinker	**Social smoker**	**Pet**	**Weight**	**Height**	**Body mass index**	**Absenteeism time in hours**
1	0	1	90	172	30	4

Data Characteristic: The table 1.2 shows the data characteristic such as count mean first, second, third quartile values etc.

Table 1.2: Columns and their statistical details.

	ID	Reason for absence	Month of absence	Day of the week	Seasons	Transportation expense	Distance from Residence to Work
count	740	737	739	740	740	733	737
mean	18.017568	19.188602	6.31935	3.914865	2.544595	221.035471	29.667571
std	11.021247	8.437493	3.435948	1.421675	1.111831	66.954179	14.848124
min	1	0	0	2	1	118	5
25%	9	13	3	3	2	179	16
50%	18	23	6	4	3	225	26
75%	28	26	9	5	4	260	50
max	36	28	12	6	4	388	52
	Service time	**Age**	**Work load Average/day**	**Hit target**	**Disciplinary failure**	**Education**	**Son**
count	737	737	730	734	734	730	734
mean	12.565807	36.449118	271188.8603	94.587193	0.053134	1.29589	1.017711
std	4.389813	6.480148	38981.88087	3.792705	0.224453	0.676965	1.094928
min	1	27	205917	81	0	1	0
25%	9	31	244387	93	0	1	0
50%	13	37	264249	95	0	1	1
75%	16	40	284853	97	0	1	2
max	29	58	378884	100	1	4	4
	Social drinker	**Social smoker**	**Pet**	**Weight**	**Height**	**Body mass index**	**Absenteeism time in hours**
count	737	736	738	739	726	709	718

mean	0.567164	0.07337	0.746612	79.063599	172.152893	26.684062	6.977716
std	0.495805	0.260919	1.319726	12.86863	6.081065	4.292819	13.476962
min	0	0	0	56	163	19	0
25%	0	0	0	69	169	24	2
50%	1	0	0	83	170	25	3
75%	1	0	1	89	172	31	8
max	1	1	8	108	196	38	120

Unique Value Counts: The table 1.3 shows the Columns on the left side and the unique value counts on the right.

Table 1.3: Columns with their value counts

Column Name	Unique Value Counts
ID	36
Reason for absence	28
Month of absence	13
Day of the week	5
Seasons	4
Transportation expense	25
Distance from Residence to Work	26
Service time	19
Age	23
Work load Average/day	39
Hit target	14
Disciplinary failure	3
Education	5
Son	6
Social drinker	3
Social smoker	3
Pet	7
Weight	27
Height	15
Body mass index	18
Absenteeism time in hours	20

Chapter 2

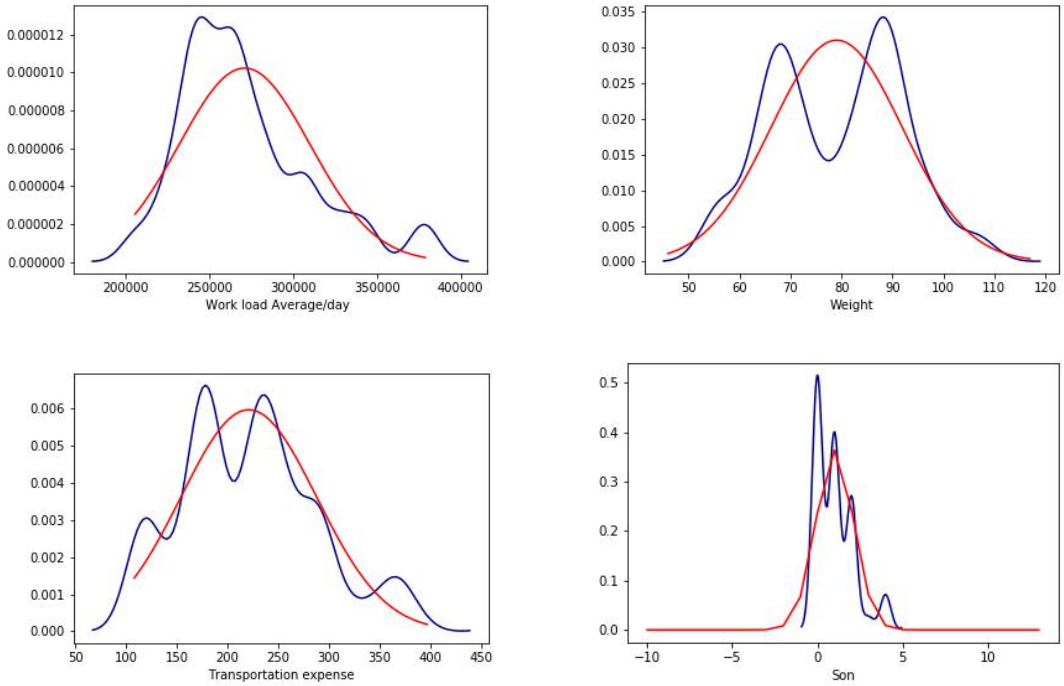
Methodology

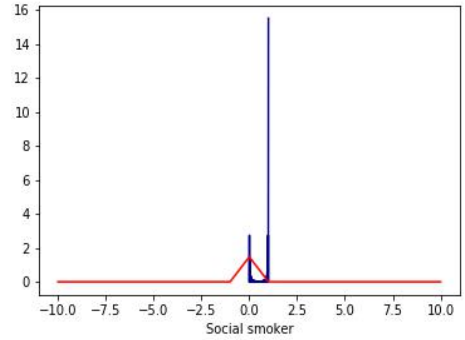
2.1 Pre-Processing

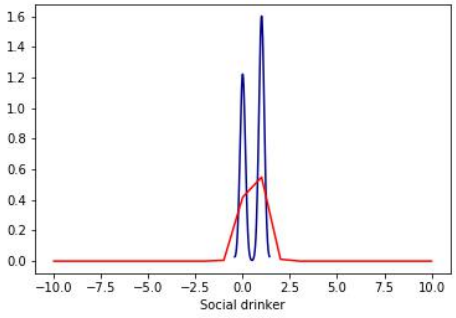
The categorical data is already in numerical form so we don't need to perform any sort of conversion to get the categorical codes out of text. Also, it is observed that column "Reason of Absence" and "Month of absence" has zero in it which is not a category so we can replace it with nan.

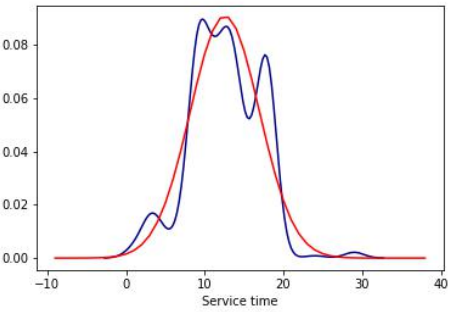
2.1.1 Uni variate Analysis

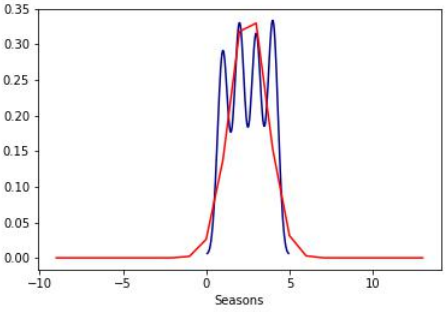
The figure 2.1 shows the data distribution of the variables(blue line) and the normal distribution curve(red line) with the variable mean and standard deviation.
We can see that the continuous variables follows the normal distribution but the categorical variables doesn't.

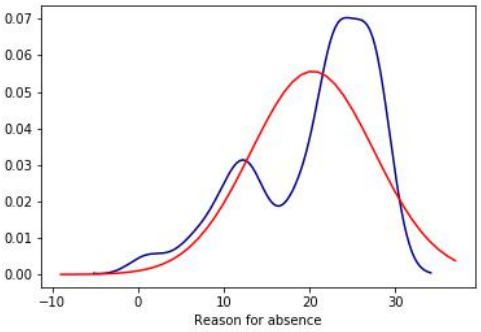


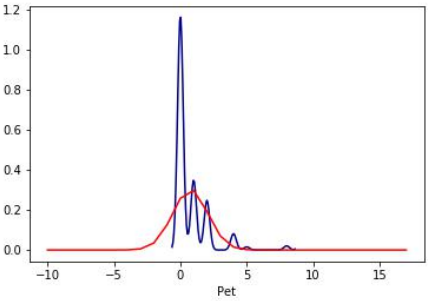

Social smoker

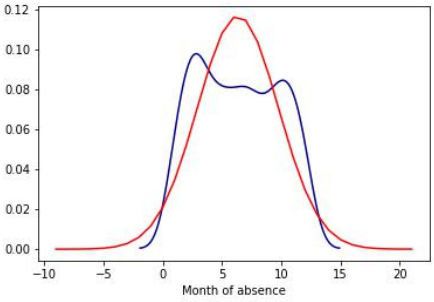

Social drinker

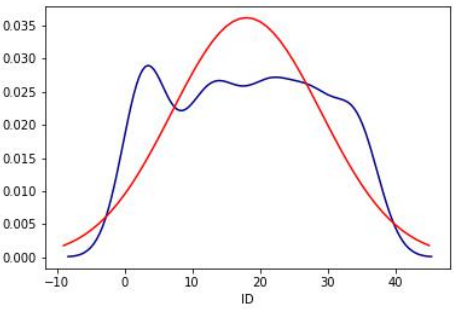

Service time

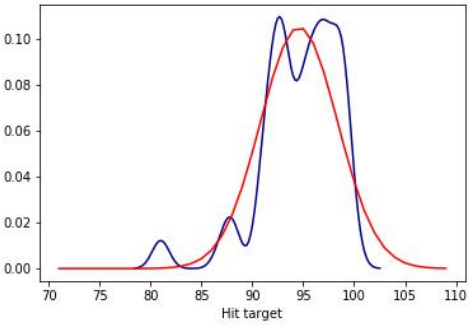

Seasons

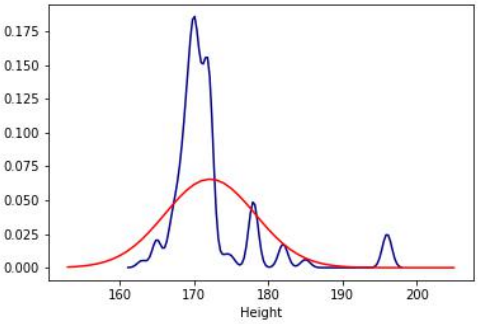

Reason for absence


Pet


Month of absence


ID


Hit target


Height

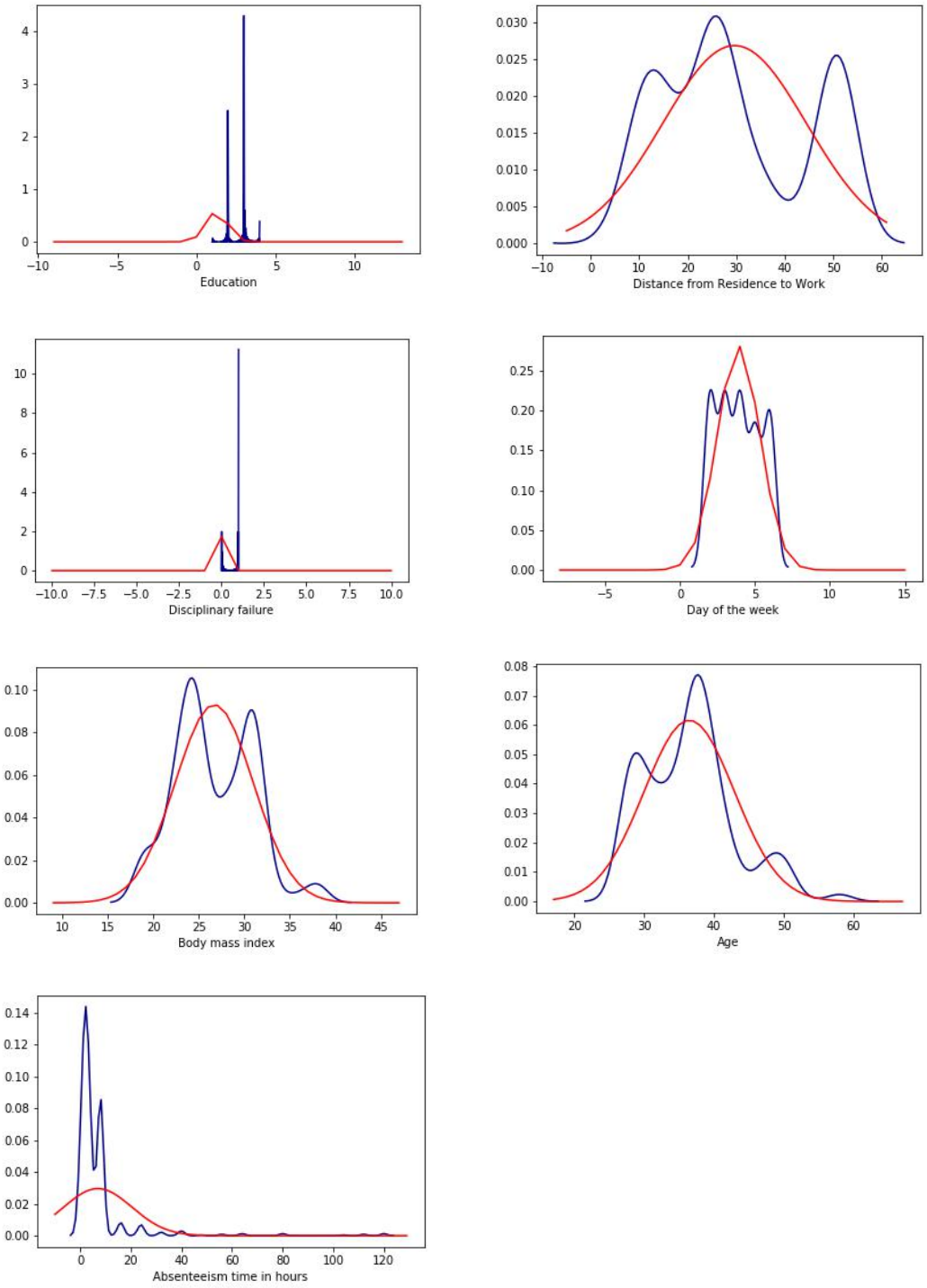


Figure 2.1: Probability Density Function of different Variables

2.2 Missing Value Analysis

The table 2.1 shows the variables and the missing value percentage. The highest number of missing values is in Reason of absence and columns seasons, day of the week, and ID doesn't has any missing values.

Table 2.1: Variables with their missing value percentage

Variables	Missing_percentage
Reason for absence	6.216216
Body mass index	4.189189
Absenteeism time in hours	2.972973
Height	1.891892
Work load Average/day	1.351351
Education	1.351351
Transportation expense	0.945946
Son	0.810811
Disciplinary failure	0.810811
Hit target	0.810811
Social smoker	0.540541
Month of absence	0.540541
Age	0.405405
Service time	0.405405
Distance from Residence to Work	0.405405
Social drinker	0.405405
Pet	0.27027
Weight	0.135135
Seasons	0
Day of the week	0
ID	0

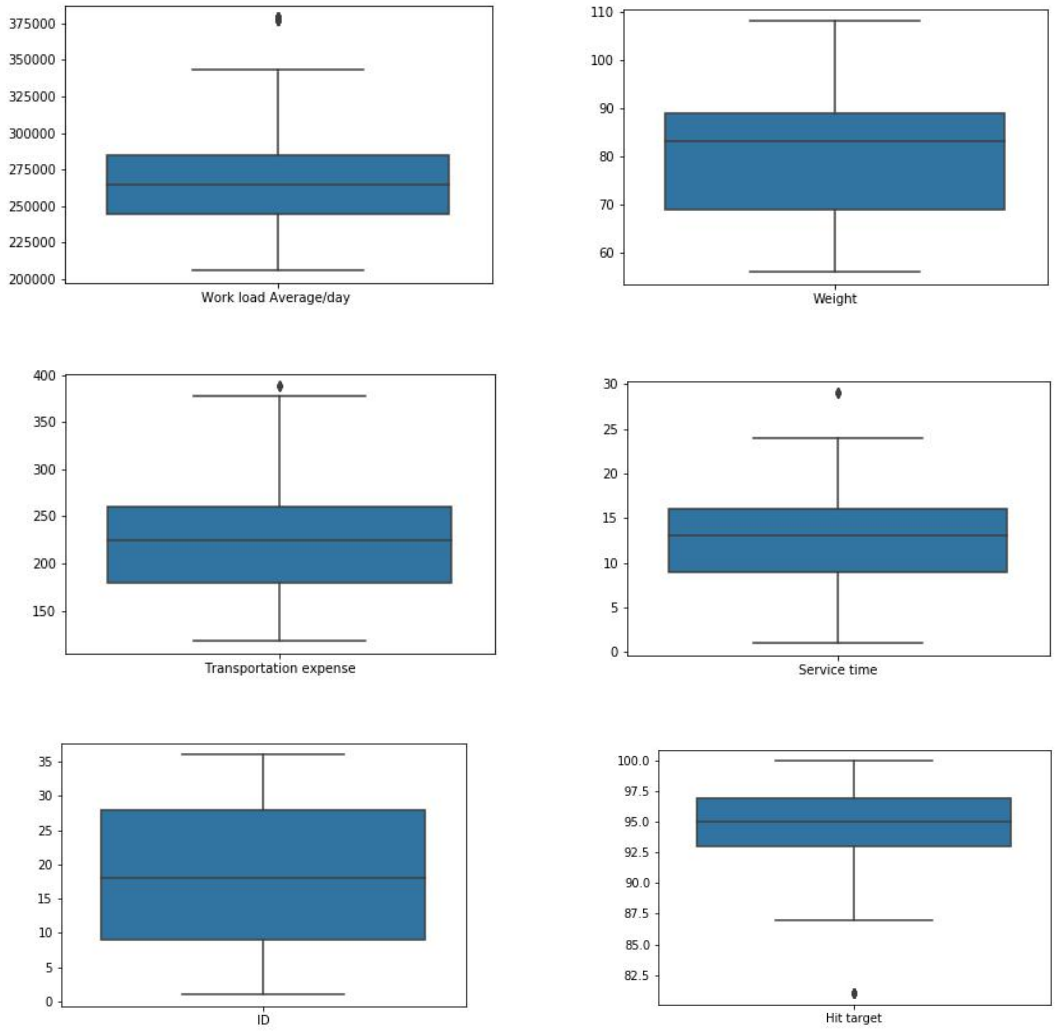
2.3 Outlier Analysis

Outliers are those instances whose values doesn't t follows the regular pattern. Either these are the erroneous cases or they are the special cases which we need to consider separately. Here, in this case we considered the outliers as the erroneous cases and has replaced with the nan. We choose to replace instances which has values less than the first quartile value - 1.5 times the IQR or greater than third quartile value + 1.5 times the IQR are considered as the outliers.
Below are the number of outliers replaced with nan for each columns:

- For column Transportation expense number of outliers replaced with NaN are 3

- For column Distance from Residence to Work number of outliers replaced with NaN is 0
- For column Service time number of outliers replaced with NaN are 5
- For column Age number of outliers replaced with NaN are 8
- For column Work load Average/day number of outliers replaced with NaN are 31
- For column Hit target number of outliers replaced with NaN are 19
- For column Weight number of outliers replaced with NaN is 0
- For column Height number of outliers replaced with NaN are 119

Figure 2.2 has the box plots for the Continuous variables showing the outliers.



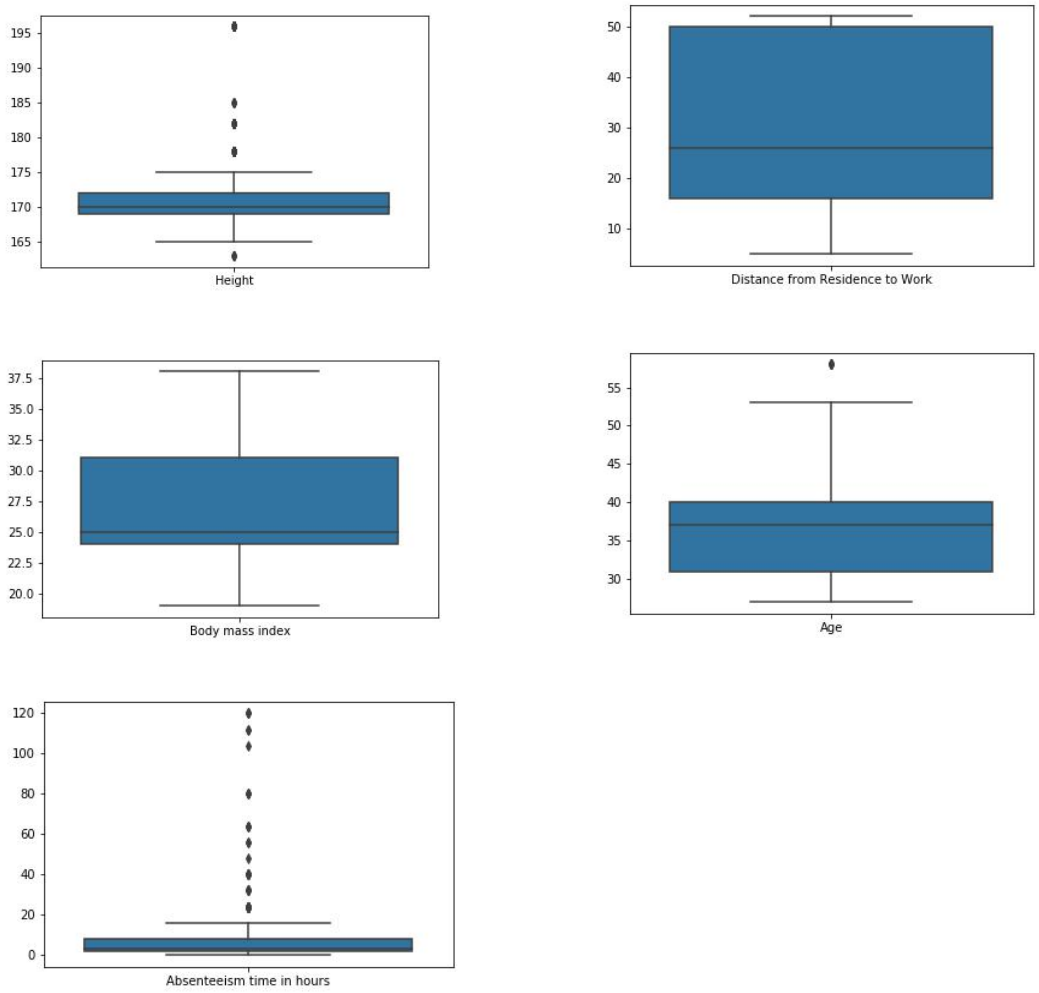


Figure 2.2: Box plots for different Variables

2.4 Impute Missing Value

Now, after the outlier detection we have more number of NaNs in our data. So, we need to remove these nan values. There are many algorithms to impute the values such as, mean, median, mode, or KNN. We reject the thought of using KNN because the KNN finds the nearest instances using MSE. Since using MSE without normalization or standardization ends in finding the wrong neighbors because of the values variables such as service time which has high values. This reduces the effect of other independent variables while finding the neighbors.
Also, we saw a pattern in the data. For each ID the columns such as distance of home and work, transportation expenses, education, son, social drinker, social smoker etc does not change in one year. So, we replaced the nans with the mean of the column value for that particular ID. Table 2.2 shows the value counts for two IDs.

Table 2.2: Shows the Columns for the given IDs with their respective value counts.

Counts\ID	3	34
ID	1	1
Reason for absence	13	14
Month of absence	12	12
Day of the week	5	5
Seasons	4	4
Transportation expense	1	1
Distance from Residence to Work	1	1
Service time	1	1
Age	1	1
Work load Average/day	28	24
Hit target	13	12
Disciplinary failure	2	1
Education	1	1
Son	1	1
Social drinker	1	1
Social smoker	1	1
Pet	1	1
Weight	1	1
Height	1	2
Absenteeism time in hours	8	10

The columns which are bold usually does not changes in a short period of time such as a year.
We dropped all those instances where the target variable is nan.

2.5 Feature Selection

Feature selection is an important and next step after the removal of outliers and imputation of NaNs. In feature selection we choose the independent variables which are really independent to each other and contributes maximum for predicting the dependent variable. For this purpose we have used three tests:

- ANOVA Test
- Chi-square test
- Correlation matrix

The ANOVA test is used here to see the contribution of each categorical variable in predicting the target/dependent variable. Below are the values of the test performed.

Table 2.3: Variables and their ANOVA score values calculated with respect to target variable

Categorical Variable	ANOVA Score Value
Reason for absence	3.78E-34
Month of absence	0.00132978
Day of the week	0.01483054
Seasons	0.00402139
Disciplinary failure	2.16E-290
Education	0.48776212
Social drinker	0.00200266
Social smoker	0.42398522
Son	3.34299369e-07
Pet	0.0576525

We can drop all those independent variables whose value is greater than 0.05. As these variables and the dependent variable are almost independent to each other. Thus, gives no or small contribution in predicting the dependent variable but increases the complexity of the dataset. Thus, removing these variables reduces the overall model complexity.

Hence, we remove columns "Social Smoker", "Pet", and "Education".

Second test we used is the correlation matrix between all the conti nuous variables. If the correlation value is very high then we drop of the two variables. As the information incorporated by these varia bles is redundant. Thus, we choose to drop body mass index as it has correlation of 0.91 with weight. Also, considering the fact that the body mass index is derived from weight and height we can re move it(Figure 2.3).

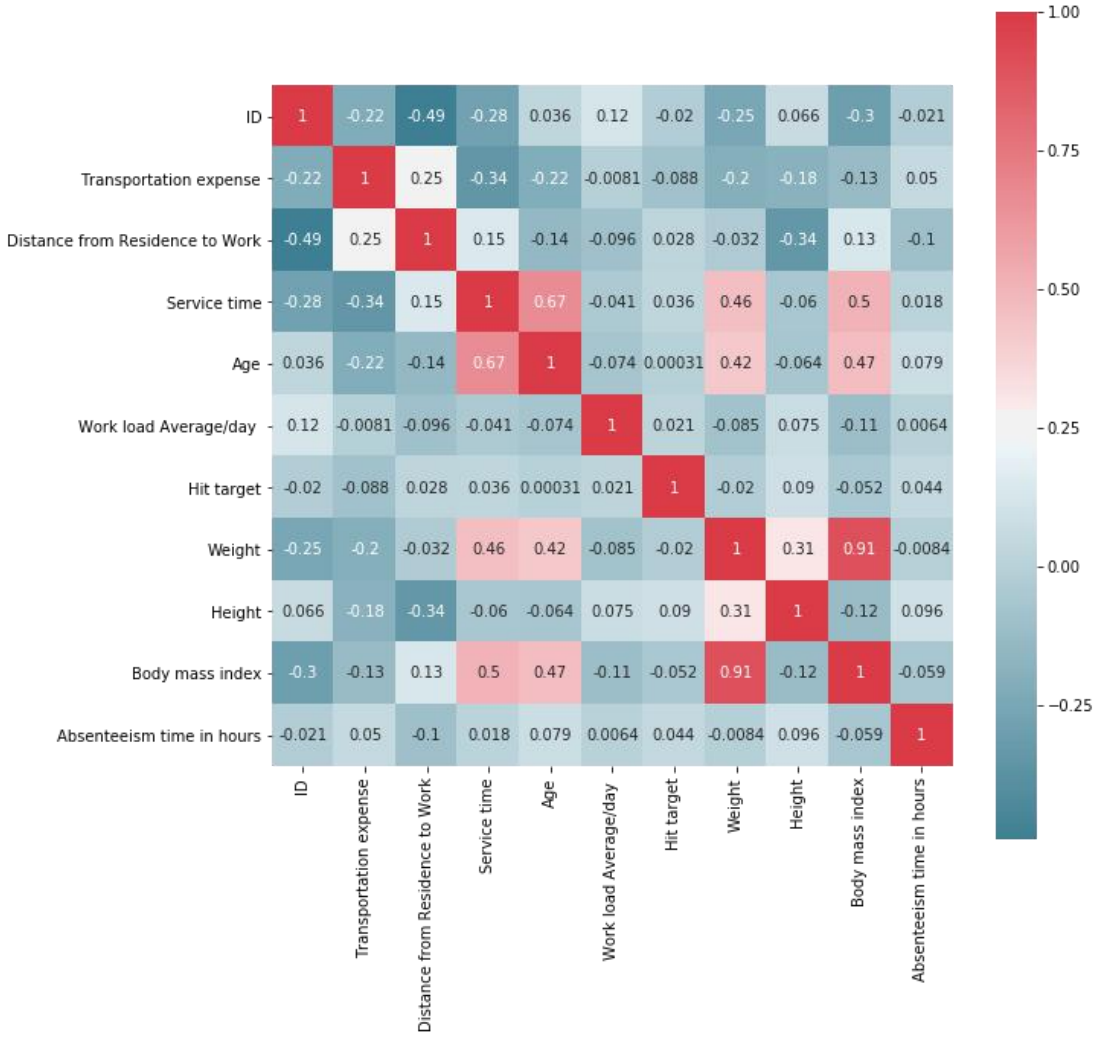


Figure 2.3: Correlation Matrix of Continuous Variables

Third test is the chi-square values between all the categorical variables. If the value is greater than 0.05 than the variables are independent. Here in figure 2.4, we can see that "Day of the week" has values greater than 0.05 for most of the independent variables thus it contains most of the information.

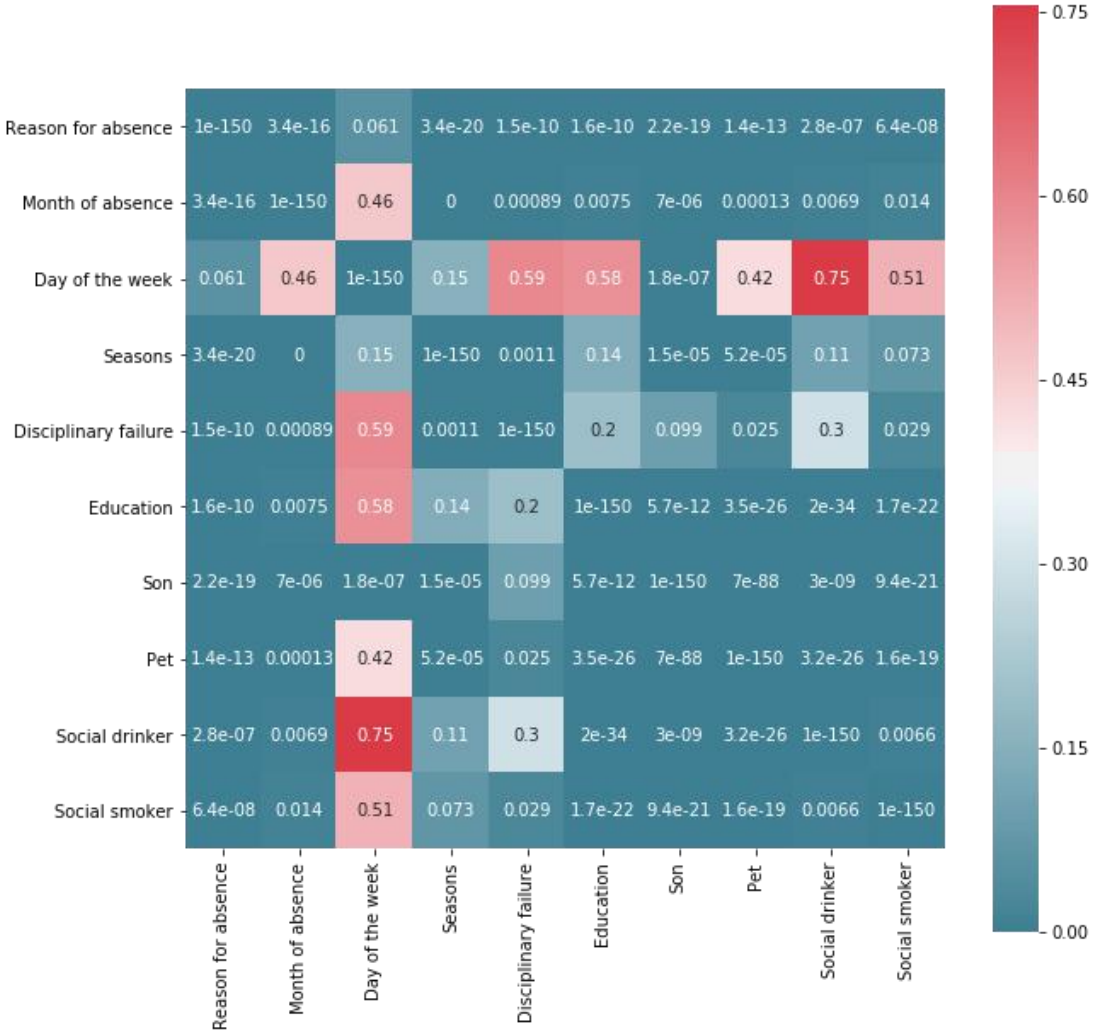


Figure 2.4: Chi-square matrix for categorical Variables

2.6 Model

We have used three regressive models for the prediction of the dependent variable Absenteeism time. For prediction we have used the normalization technique, converted the categorical features into the dummy variables. We tested all the three models, namely Linear regression, Decision tree regressor, and Random forest regressor, firstly without dummy variables and without normalization, secondly with normalization and thirdly with both normalization and dummy variables.

First model i.e., Linear regressor tries to fit a straight line which has the minimum loss function value. Here the loss function used is the MSE(mean squared error).

Second model i.e., Decision tree computes a tree based on a specified criterion, here MSE. We have used other configurations as follows: maximum features to be 8 and random state as 1. We set the random state so that we get the same answer each time. Rest of the hyper-parameters are kept with the default values.

Lastly, we used the an ensembler, Random forest, for predicting the values. The configurations used for it are as follows: Number of trees used to estimate the values as 25, max features as 11, max depth as 10, random state as 1.

Table 2.4 shows the importance values given to each independent variable by the decision tree regressor:

Table 2.4: Variables with their importance values by decision tree regressor

Variables	Importance Value
Day of the week	0.23051986
Reason for absence	0.18915196
Month of absence	0.15729824
Work load Average/day	0.08366446
ID	0.06846845
Distance from Residence to Work	0.055564314
Seasons	0.045025006
Age	0.04463034
Height	0.038617946
Social drinker	0.021893933
Transportation expense	0.019458419
Son	0.019167522
Hit target	0.014129688
Weight	0.008909071
Disciplinary failure	0.002397224
Service time	0.001103569

Chapter 3

Conclusion

3.1 Model Evaluation

3.1.1 MSE(Mean Squared Error)

MSE or mean squared error is an error metric for regression problems. It is calculated as the squared difference of the predicted and target values. The lower the value the better is the performance.

Linear Regression:

Table 3.1: MSE values by Linear Regressor

	Without any transformations	With normalization	With normalization & dummy variables
Training Dataset	158.5892426	158.5892426	3.89E-25
Validation Dataset	160.269305	1768727355	3.92E-25

Clearly, the loss in the third column is minimum. Also, linear regression works by minimizing the MSE so the normalized dataset will help in this case.

Decision Tree:

Table 3.2: MSE values by Decision Tree Regressor

	Without any transformations	With normalization	With normalization & dummy variables
Training Dataset	8.909694555	8.909694555	0.00E+00
Validation Dataset	294.869213	297.7638889	1.65E+02

Similarly, as in the case of linear regression here also the third column has the minimum loss.

Random Forest:

Table 3.3: MSE values by Random forest Regressor

	Without any transformations	With normalization	With normalization & dummy variables
Training Dataset	37.46308451	37.46066405	6.40E+00
Validation Dataset	170.1000651	296.5525826	4.25E+01

Clearly, the values in third column has the minimum values and thus the minimum loss.

3.1.2 R2 Score

R2 score is calculated by subtraction of 1 and ratio of residual sum of squares ((y_true - y_prediction) ** 2).sum() and the total sum of squares ((y_true - y_true.mean()) ** 2).sum(). The best value is 1. Signifying that the residual sum of squares is zero.

Linear Regression:

Table 3.4: R2 score values by Linear Regressor

	Without any transformations	With normalization	With normalization & dummy variables
Training Dataset	0.171329944	0.171329944	1
Validation Dataset	-0.015672297	11208928.72	1

Clearly, the values in third column has the best values.

Decision Tree:

Table 3.5: R2 score values by Decision tree Regressor

	Without any transformations	With normalization	With normalization & dummy variables
Training Dataset	0.953444528	0.953444528	1
Validation Dataset	0.868670304	-0.887014691	-0.042864829

Here, the R2 score without any transformation is giving the best scores.

Random Forest:

Table 3.6: R2 score values by Random Forest Regressor

	Without any transformations	With normalization	With normalization & dummy variables
Training Dataset	0.804245636	0.804258283	0.966563853
Validation Dataset	-0.077972628	-0.879338298	0.730702299

Here, the values in third column has the best values.

3.2 Model Selection

From the above tables we can infer that the Linear Regression with normalization and dummy variables has the minimum training and validation loss and highest R2 values.

3.3 Suggestion for the Firm

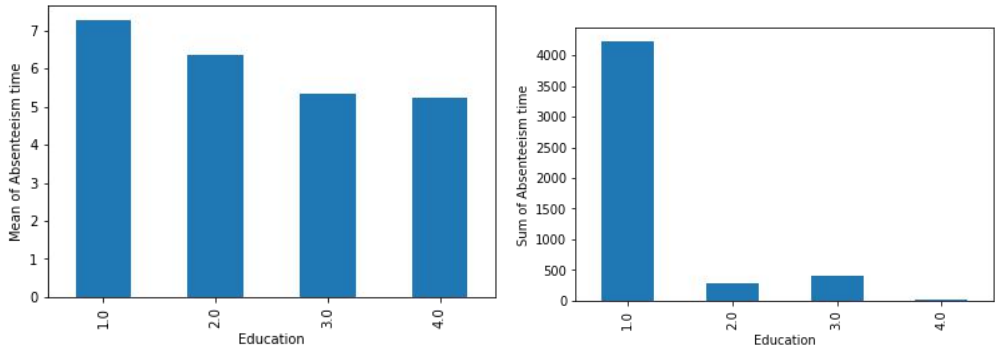


Figure 3.1: Education vs mean and total Absenteeism time(in hours)

Inferences: Figure 3.1 shows the relationship between Absenteeism time and education. The left graph shows the mean of Absenteeism time and education and the right graph shows the sum of Absenteeism time and education. Both the graphs indicates that the employees with the higher education qualifications are more productive.
Suggestion: Hire more of graduates as this will keep the employees with average pay and less absenteeism time.

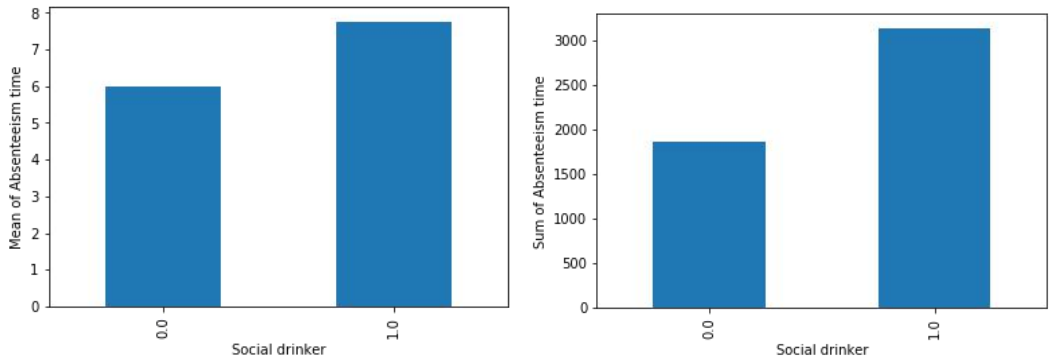


Figure 3.2: Social drinker vs mean and total Absenteeism time(in hours)

Inference: Figure 3.2 shows the relationship between Absenteeism time and drinking habits. The left graph shows the mean of Absenteeism time and drinking habit and the right graph shows the sum of Absenteeism time and drinking habit. Both the graphs indicates that the employees who drinks are absent **twice** the time who doesn't drinks.
Suggestion: Some non-smoking initiatives can be started within the firm.

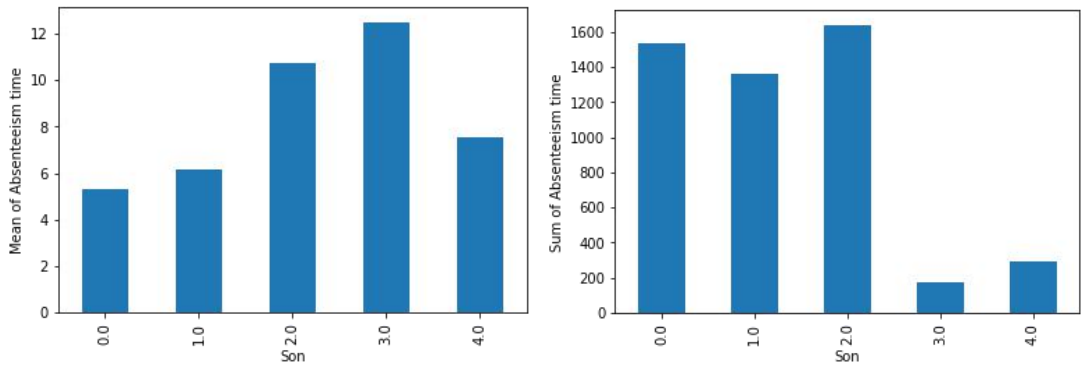


Figure 3.3: No. of Son vs mean and total Absenteeism time(in hours)

Inference: Figure 3.3 shows the relationship between Absenteeism time and number of children. The left graph shows the mean of Absenteeism time and number of children and the right graph shows the sum of Absenteeism time and number of children. The graph on the left shows that on an average the employees having more number of children are absent for . But on the other hand there are more number of employees, being absent, having zero to two children.

Suggestion: Some kind of family planning awareness can be initiated within the firm.

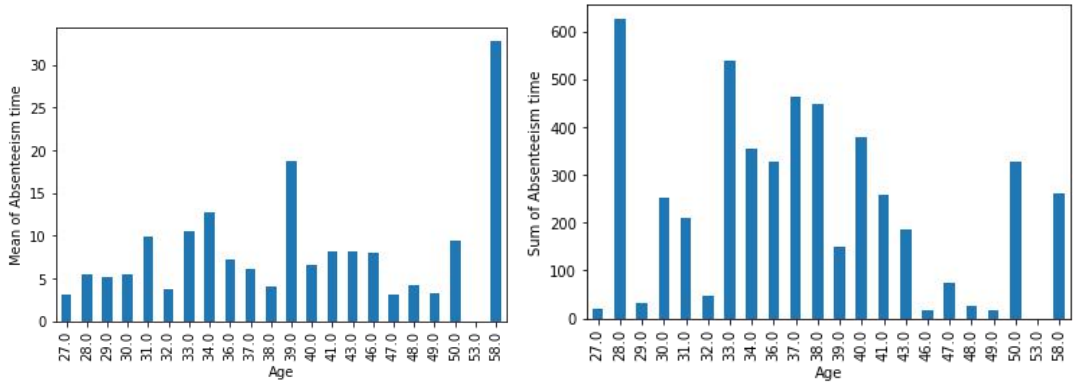


Figure 3.4 Age vs mean and total Absenteeism time(in hours)

Inference: Figure 3.4 shows the relationship between Absenteeism time and age. The left graph shows the mean of Absenteeism time and age and the right graph shows the sum of Absenteeism time and age. On an average people of age around 39 are absent. And employees of age around 28 has the highest frequency.

Suggestion: Employees of age group 28, and 39 should be warned.

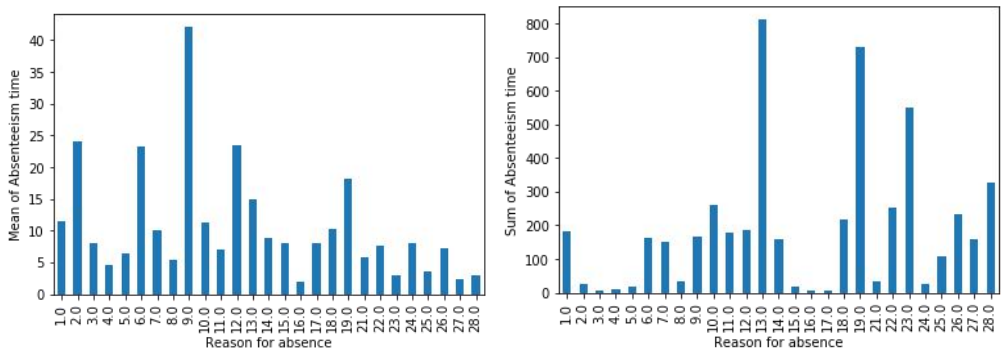


Figure 3.5: Reason of absence vs mean and total Absenteeism time(in hours)

Inference: Figure 3.5 shows the relationship between Absenteeism time and reason for absence. The left graph shows the mean of Absenteeism time and reason for absence and the right graph shows the sum of Absenteeism time and reason for absence. Most of the people were absent due to Diseases of the musculo-skeletal, injury, poisoning, and blood donation.

Suggestion: Firm can start some kind of checkups on weekends instead of working days. Also, Some health awareness programs, like yoga, can also be initiated in the firm.

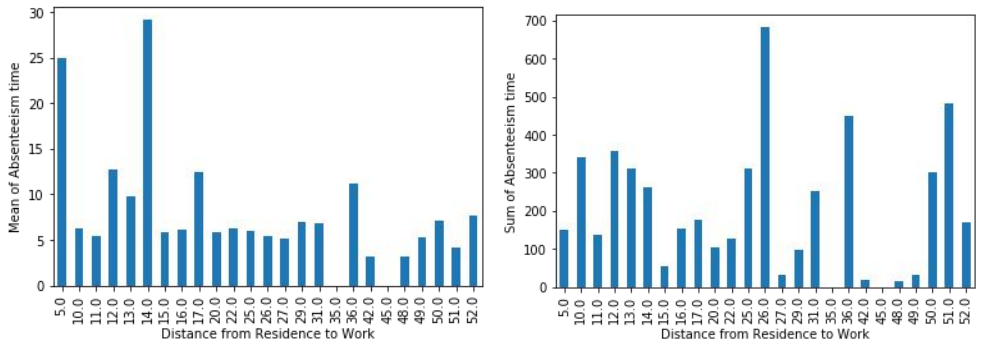


Figure 3.6: Distance from home to work vs mean and total Absenteeism time(in hours)

Inference: Figure 3.6 shows the relationship between Absenteeism time and distance from residence to work. The left graph shows the mean of Absenteeism time and distance from residence to work and the right graph shows the sum of Absenteeism time and distance from residence to work. Employees living at a distant places were absent for longer time duration.

Suggestion: Firm can promote bike/car pooling which not only will reduce the commute time in comparison to public transport but also, it will reduce their average commute cost.

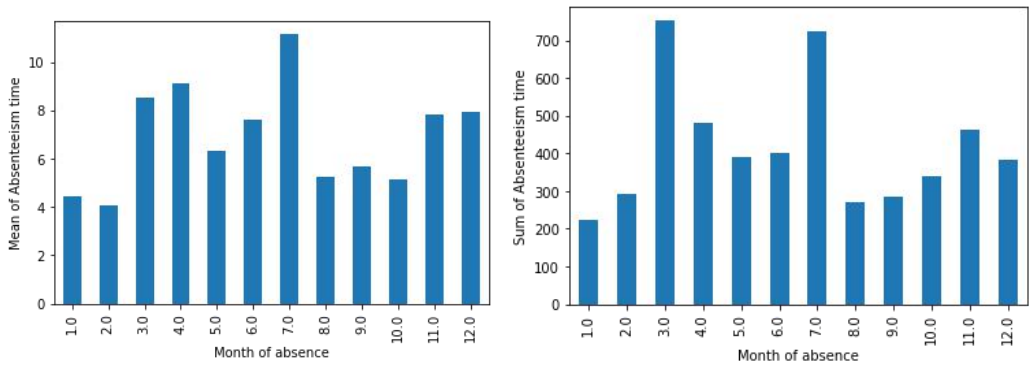


Figure 3.7: Month of absence vs mean and total Absenteeism time(in hours)

Inference: Figure 3.7 shows the relationship between Absenteeism time and months. The left graph shows the mean of Absenteeism time and months and the right graph shows the sum of Absenteeism time and months. Clearly most of the employees were absent in the month of March and July.

Suggestion: Company can start hourly payment to some employees for these months.

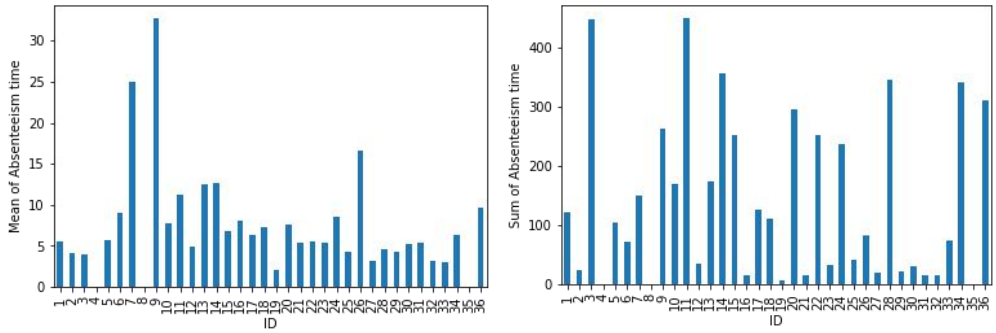


Figure 3.8: Employee ID vs mean and total Absenteeism time(in hours)

Inference: Figure 3.8 shows the relationship between Absenteeism time and employee ID. The left graph shows the mean of Absenteeism time and employee ID and the right graph shows the sum of Absenteeism time and employee ID.

Suggestion: Employees with IDs 11, 28, 36 must be warned and employees with ID 4, 8, 35 must be rewarded with some incentives at the start of next year to reduce the absenteeism time.

3.4 Trend

When there is a pattern in the data we using predictive algorithms just like we used in the above sections. But some times the pattern tends to repeat itself after a certain period of time. By analyzing the values over a period of time we can deduct the trend it follows.

In this scenario we have the data for just one year. So, we trained a linear regression model on the first 12 months data to find the trend and at the same time to predict the Absenteeism time for the upcoming year. Table 3.7 shows the Expected Absenteeism time in hours if the same trend follows. Here, the months 1 to 12 are actual values and the values from 13 to 24 shows the Absenteeism time in hours if the same trend follows predicted by the model.

Table 3.7: Month of absence and total Absenteeism time(in hours)

Month	Sum of Absenteeism time
1	222
2	294
3	752
4	482
5	392
6	403
7	724
8	272
9	284
10	340
11	463
12	382
13	403.1363636
14	400.9265734
15	398.7167832
16	396.506993
17	394.2972028
18	392.0874126
19	389.8776224
20	387.6678322
21	385.458042
22	383.2482517
23	381.0384615
24	378.8286713

Below we illustrated the trend in Absenteeism time with time:

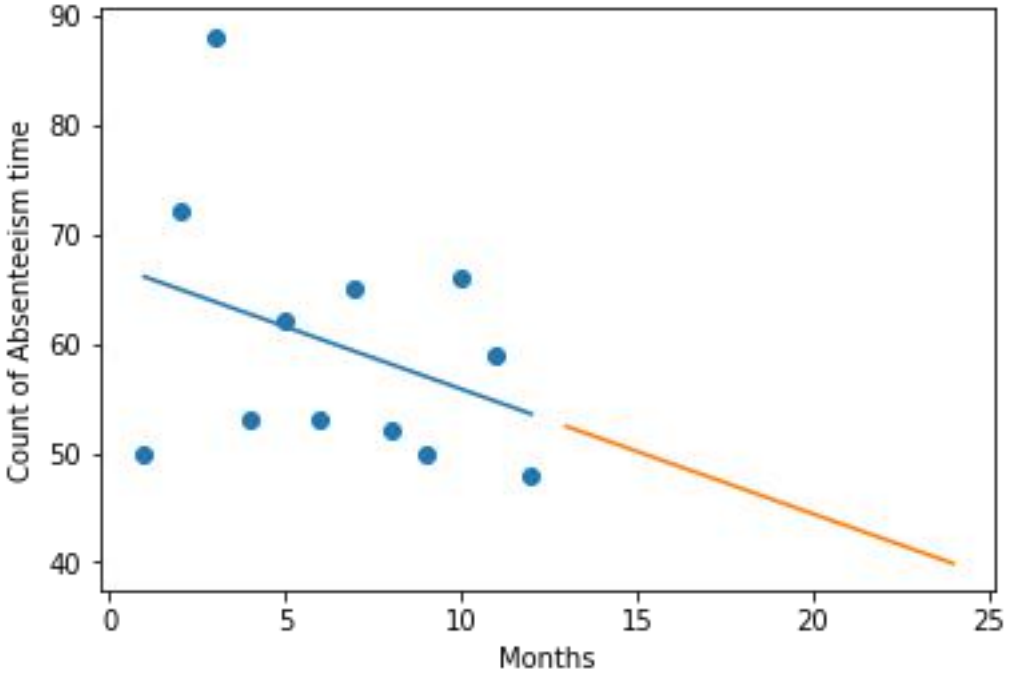


Figure 3.9: Month of absence vs count of employees absent in that month

Figure 3.9 shows the trend of absenteeism time with respect to months. It is evident that the count of employees taking time off will reduce in the next year*.

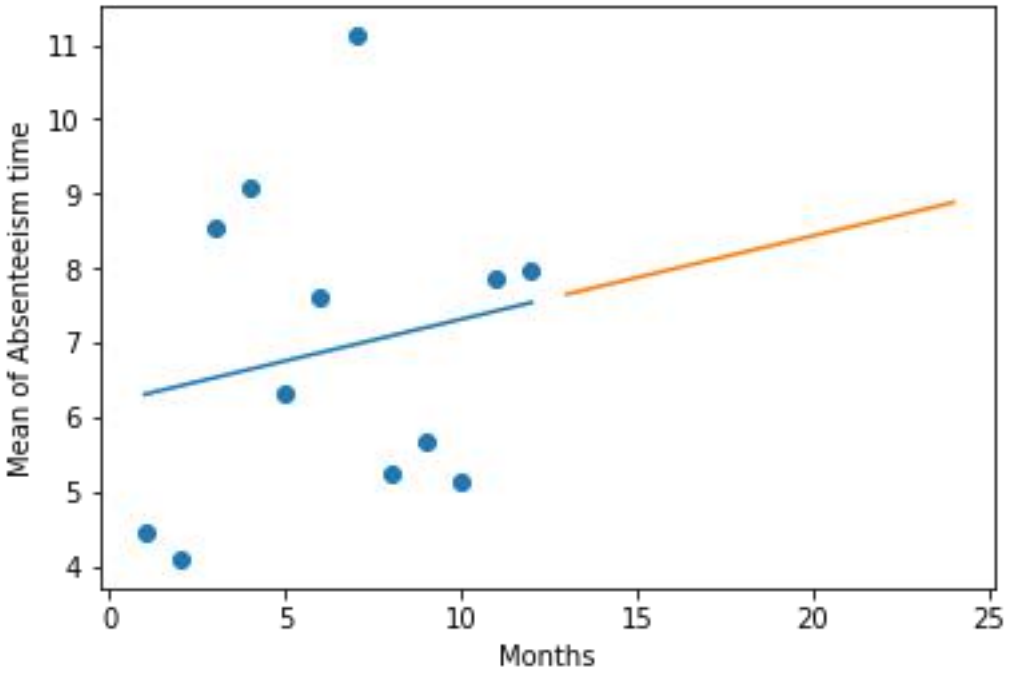


Figure 3.10: Month of absence vs mean of Absenteeism time(in hours)

Figure 3.10 shows the trend of mean absenteeism time with respect to months. It is evident that the mean time will rise in the next year. But

from the trend in number of employees we can say that the mean is rising but the number is decreasing. Thus, it indicates that the average time per employee will rise in the upcoming months*.

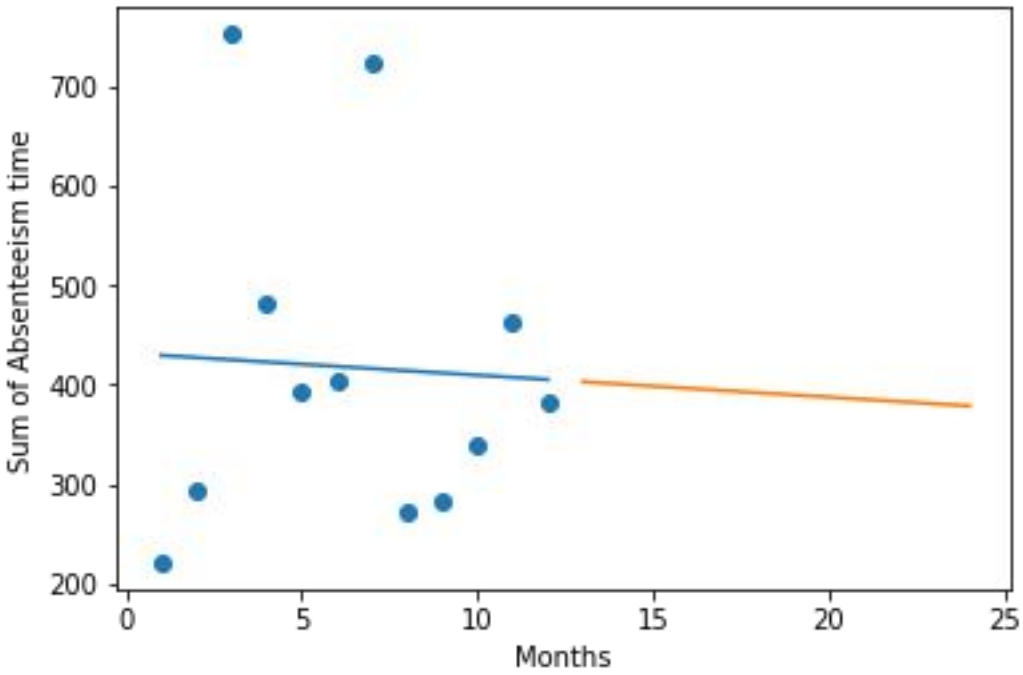


Figure 3.11: Month of absence vs mean of Absenteeism time(in hours)

Figure 3.11 shows the trend of total absenteeism time with respect to months. It is evident that the total time will reduce in the next months*. Thus, increasing the productivity of the firm.

* These trends may or may not follow as the data was available just for one year.

Appendix A - Other Deductions

1. Firm should not give warning to employee ID 3 by just seeing that he was absent for the most of the time as:

- lives far from the office - 51km(maximum - 52)
- No disciplinary failure
- February suffered with Diseases of the musculoskeletal system and connective tissue and following months went for its treatment and consultation

2. Employees of age 38 has very low mean absenteeism time and high service time. Also, as a matter of fact employees don't have any children and they are graduates.

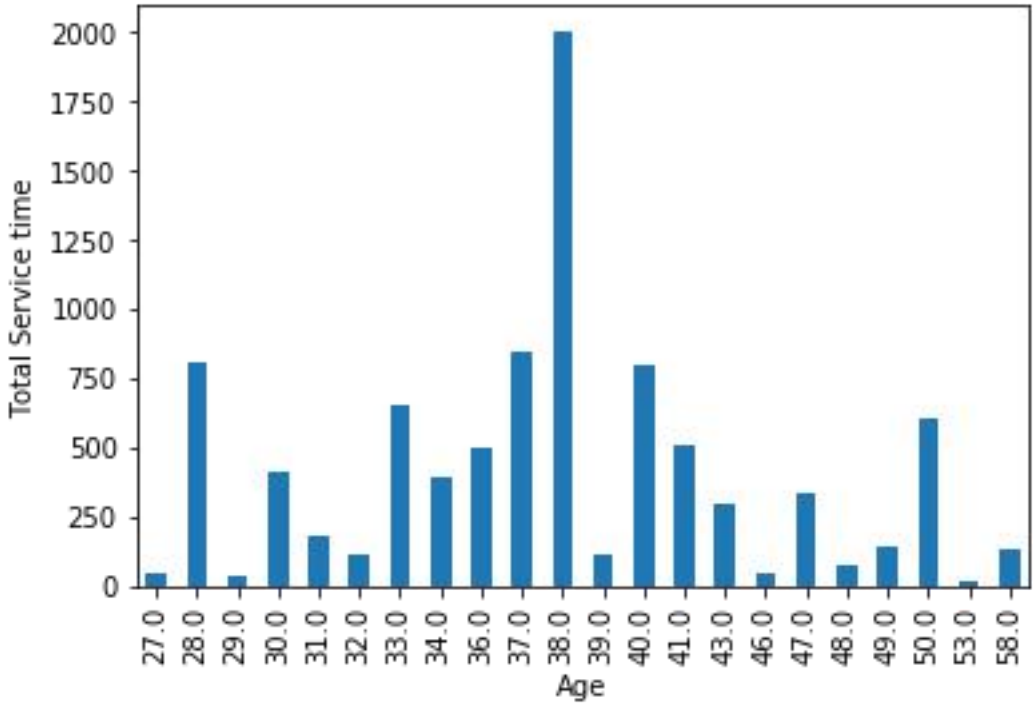


Figure 1: Age vs Total Service time

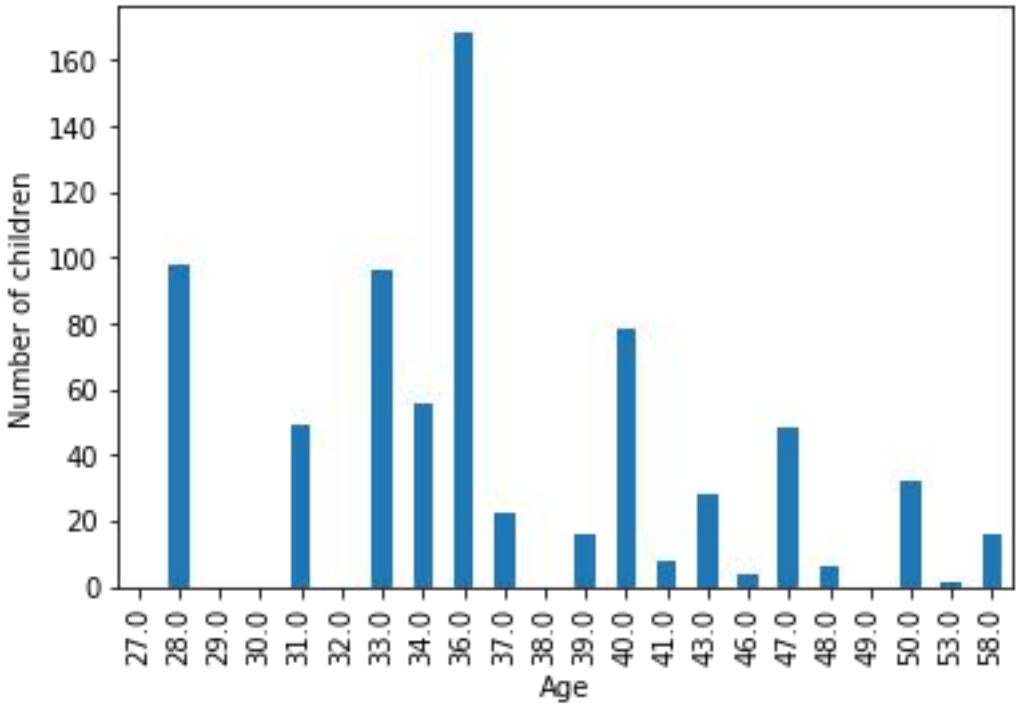


Figure 2: Age vs Number of children

3. Employee with ID 5 has the highest percentage of absence with reason: "unjustified absence"

References

Garrett Grolemund and Hadley Wickham. 2016. R for Data Science. ISBN: 9781491910399

Jake VanderPlas. 2016. Python Data Science Handbook: Essential Tools for Working with Data. ISBN: 9781491912058