EMPLOYEE ABSENTEEISM

Ataf Ahmed

25.01.2019

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

The objective of this project is to study employee details and behaviour and find the causes and patterns that cause absenteeism. The dataset has all the records of absenteeism including the requisite information of the employee for that instant.

An overview of the dataset:-

(i) Our dataset contains 740 observations and 21 features.

(ii) Variable Description:-

S.no	Variable Name	Description
1.	ID	Employee id
2.	Reason for absence	Absences attested by the International Code of Diseases- Total 28 categories
3.	Seasons	Summer-1 Autumn-2 Winter-3 Spring-4
4.	Day of the week	Day of the week
5.	Month of absence	Month of the year
6.	Transportation Expense	Fare
7.	Distance from residence to work	Distance
8.	Service time	Time of service
9.	Age	Age
10.	Workload Average/day	Average workload of that employee
11.	Hit target	Hit target
12.	Disciplinary failure	0-No 1-Yes
13.	Education	1-High school 2-Graduate

		3-Post graduate 4-Master and doctor
14.	Son	Number of children
15.	Social drinker	0-No, 1-Yes
16.	Social smoker	0-No, 1-Yes
17.	Pet	Number of pets
18.	Weight	Weight
19.	Height	Height
20.	Body mass index	Body mass index
21.	Absenteeism time in hours	Absenteeism time in hours

Absenteeism time in hours is our target variable here, while rest all others are our predictors or independent variables.

Chapter 2

Methodology

2.1 Preprocessing

The first step towards building any predictive model is Exploratory Data Analysis. We need to look at our data , its size, shape, the type of variables in it, the distributions of various variables and their respective relationship with one another.

Thus after loading the data we look at its shape and the data types of the variables.

- data = pd.read_excel("work".xls)
- data.shape
- data.dtypes.

We found that our data has a shape of 740 x 21 and all the variables in our dataset are of the numeric type .

2.1.1 Missing Values Analysis

Next we look if our dataset has any missing values present in it as they can heavily influence our training model. Missing values if present need to be handled in an appropriate way for building a good model.

Our dataset returned missing values present in 18 of the 21 variables.We need to impute the missing values with the right method.

Our dataset has 36 unique values of ID, which can be used to impute the missing values in a lot of variables. We found the ID of the missing value and imputed with the record of the same attribute given at other instances. This approach worked well for features which have personal information of the employees.

While 'Reason of Absenteeism' was used to impute 'Absenteeism time in hours', 'Workload' and 'hit target' was imputed by their means.

2.1.2 Outlier Analysis

Drawing the boxplot of all the numerical variables to look for outliers.

Fig 2.1 shows the boxplot of outliers

7 of the numerical features returned outliers in them. It would be better if we impute these outliers as it can affect our model negatively.

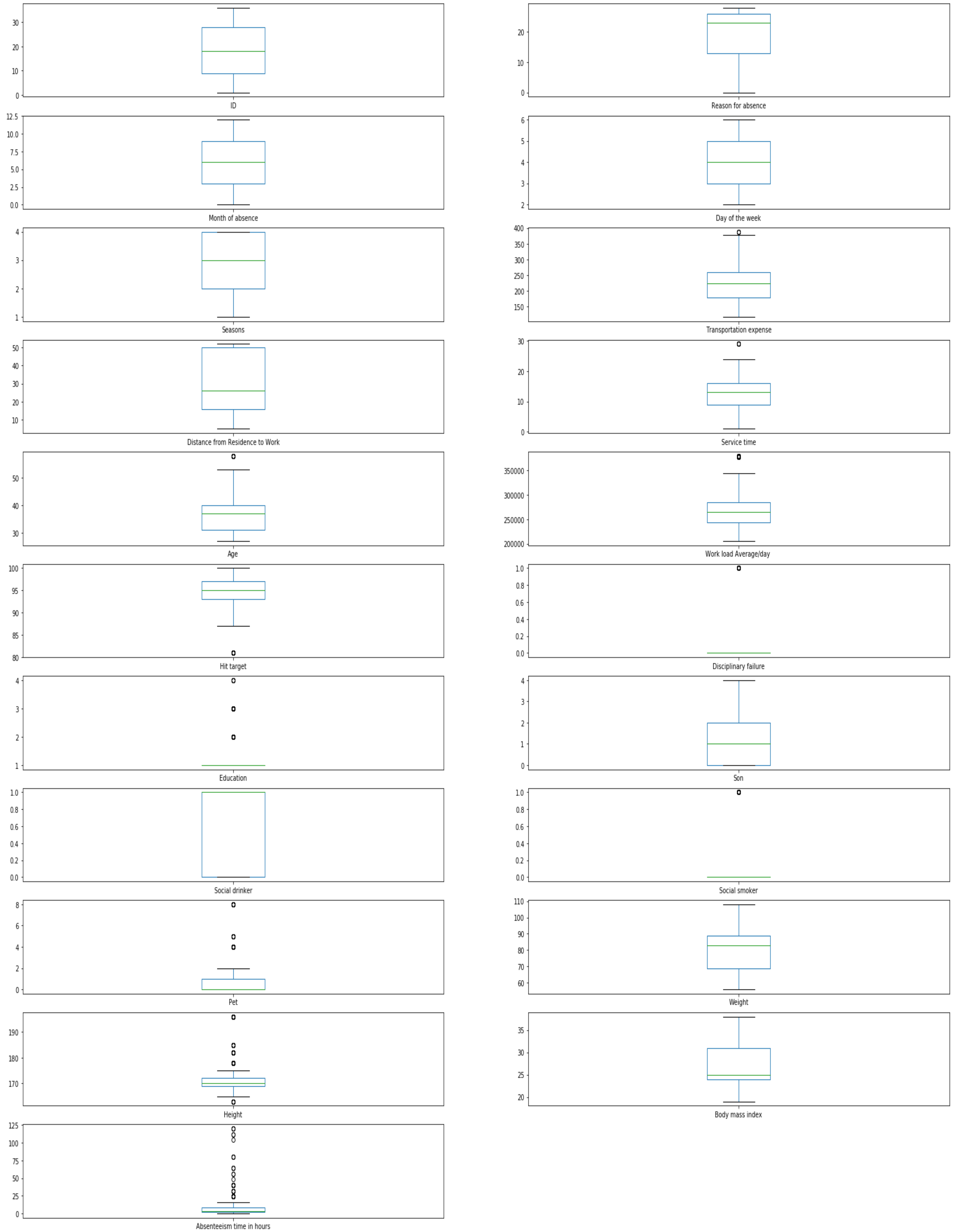


Fig 2.1 Boxplot for outliers.

Our first step while dealing with outliers would be to find them and replace them with NA.

Once all outliers are replaced with NA we can apply suitable technique to impute them – mean, median, k-nn imputation being some of them. We have used median to impute the NA values.

Checking the dataset again after imputation:-

- data.isna().sum()

Our data returned zero missing values so we can proceed further with our analysis.

2.1.3 Converting to Categorical

All the 21 features in our dataset were of float type. However after a sneak peak in the data we found that some of those features were actually categorical type and needed to be converted to factors for better analysis.

- #CONVERTING DATATYPES
- data['ID'] = data['ID'].astype('category')
- data['Month of absence'] = data['Month of absence'].astype('category')
- data['Reason for absence'] = data['Reason for absence'].astype('category')
- data['Day of the week'] = data['Day of the week'].astype('category')
- data['Seasons'] = data['Seasons'].astype('category')
- data['Disciplinary failure'] = data['Disciplinary failure'].astype('category')
- data['Education'] = data['Education'].astype('category')
- data['Son'] = data['Son'].astype('category')
- data['Social drinker'] = data['Social drinker'].astype('category')
- data['Social smoker'] = data['Social smoker'].astype('category')
- data['Pet'] = data['Pet'].astype('category')

2.2 Understanding Absenteeism

The primary objective of our project is to understand absenteeism and the major factors causing it, while suggesting ways to reduce it.

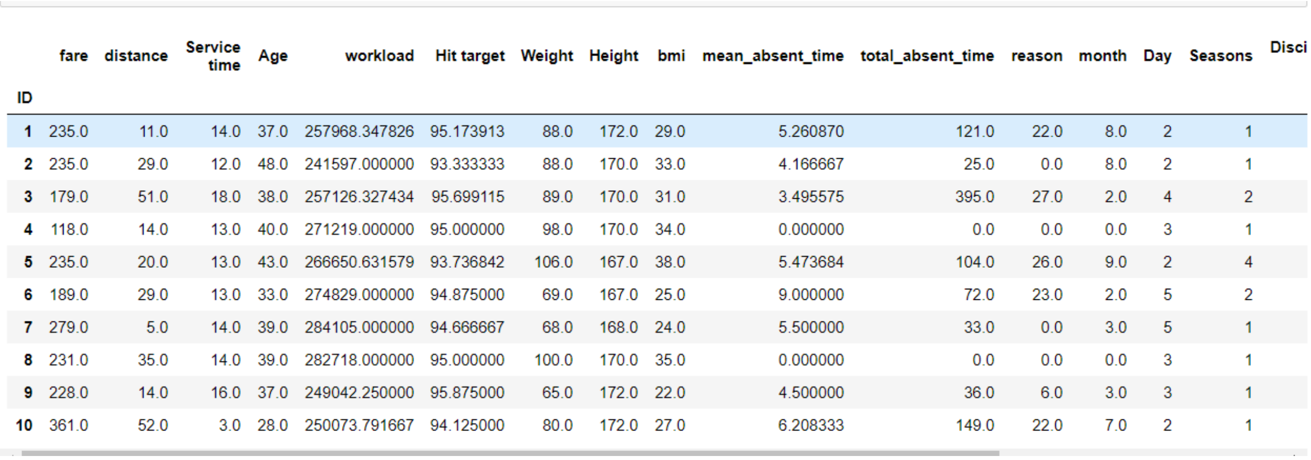
2.2.1 Absenteeism by Employee ID

Since our dataset has only 36 unique values of ID i.e. records of absenteeism of 36 employees in the company. This can greatly help our exploratory data analysis as we can study absenteeism trends of each employee differently.

I have created a new dataframe to do so having 36 rows, each row being record of each employee.

While the columns have all the other features of our original dataset, I have also included two new columns- **'mean_absent_time'** & **'total_absent_time'.**

The new dataset is named as **id_data**.



ID	fare	distance	Service time	Age	workload	Hit target	Weight	Height	bmi	mean_absent_time	total_absent_time	reason	month	Day	Seasons	Disci
1	235.0	11.0	14.0	37.0	257968.347826	95.173913	88.0	172.0	29.0	5.260870	121.0	22.0	8.0	2	1	
2	235.0	29.0	12.0	48.0	241597.000000	93.333333	88.0	170.0	33.0	4.166667	25.0	0.0	8.0	2	1	
3	179.0	51.0	18.0	38.0	257126.327434	95.699115	89.0	170.0	31.0	3.495575	395.0	27.0	2.0	4	2	
4	118.0	14.0	13.0	40.0	271219.000000	95.000000	98.0	170.0	34.0	0.000000	0.0	0.0	0.0	3	1	
5	235.0	20.0	13.0	43.0	266650.631579	93.736842	106.0	167.0	38.0	5.473684	104.0	26.0	9.0	2	4	
6	189.0	29.0	13.0	33.0	274829.000000	94.875000	69.0	167.0	25.0	9.000000	72.0	23.0	2.0	5	2	
7	279.0	5.0	14.0	39.0	284105.000000	94.666667	68.0	168.0	24.0	5.500000	33.0	0.0	3.0	5	1	
8	231.0	35.0	14.0	39.0	282718.000000	95.000000	100.0	170.0	35.0	0.000000	0.0	0.0	0.0	3	1	
9	228.0	14.0	16.0	37.0	249042.250000	95.875000	65.0	172.0	22.0	4.500000	36.0	6.0	3.0	3	1	
10	361.0	52.0	3.0	28.0	250073.791667	94.125000	80.0	172.0	27.0	6.208333	149.0	22.0	7.0	2	1	

Fig 2.2 Employee wise data

Fig 2.2 shows a part of the new dataset, comprising the new features **mean_absent_time** and **total_absent_time.** Features like month, season and reason for absence were created by their respective mode for that particular employee.

- MOST FREQUENT ABSENTEE

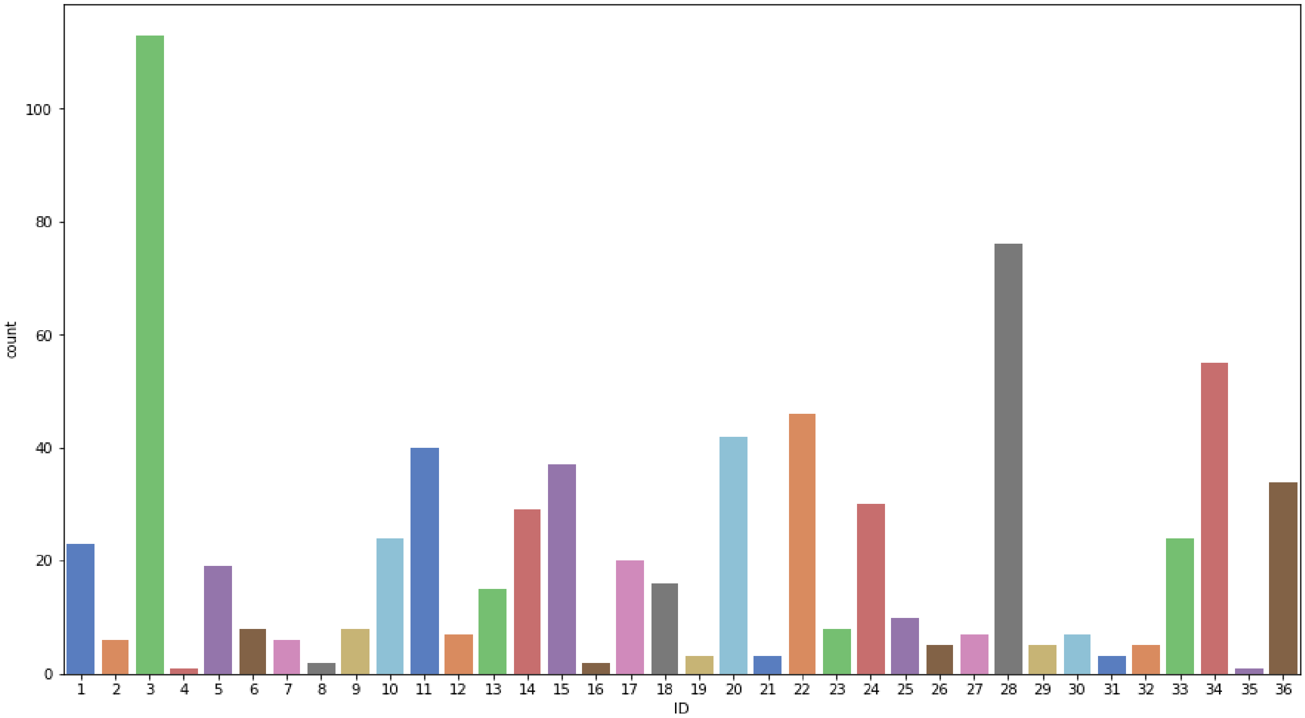


Fig 2.3 Countplot of ID wise absenteeism

Employee ID 3 has a huge instances of absenteeism followed by ID 28 and 34.

Employee 4 and 35 have the least instances of absenteeism.

- MOST HOURS ABSENT

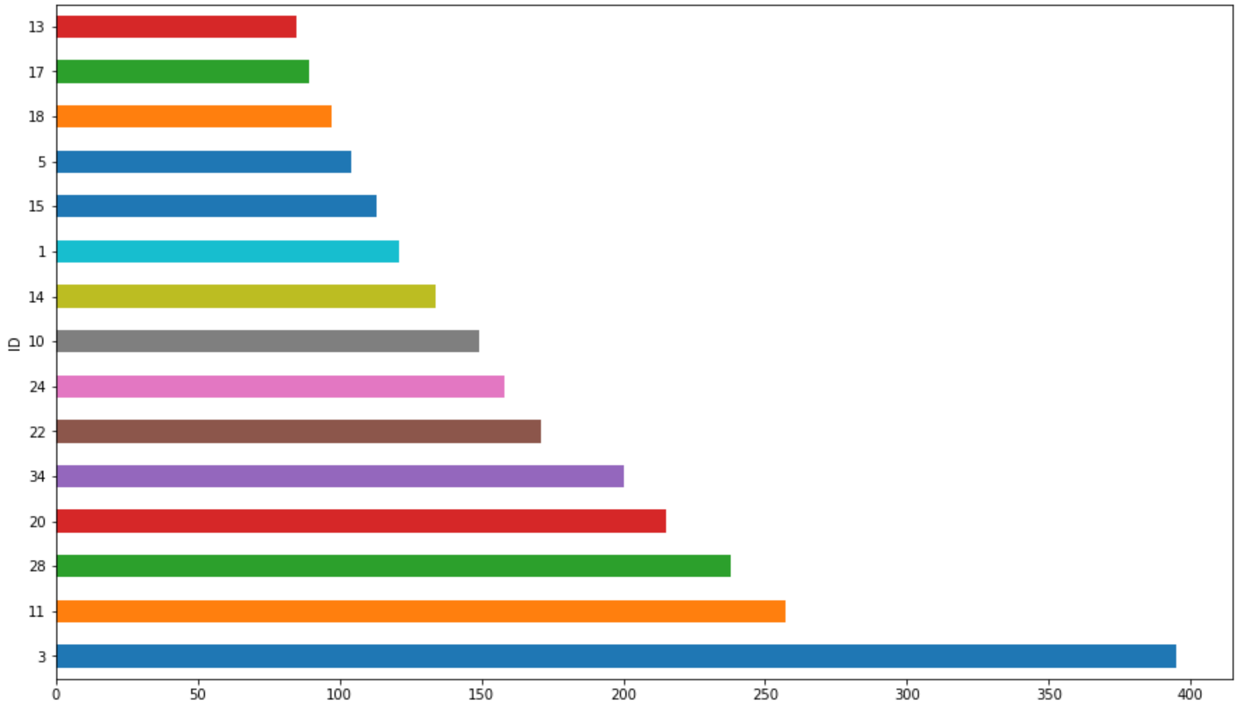


Fig 2.4 total absenteeism hours v/s ID

Obviously employee no. 3 has the most hours absent with around 400 hours, which is a huge number.

He is followed by 11, 28 , 20 & 34.

We can have a closer look at these 5 employees who have caused the most absenteeism accounting for nearly **45% of total absenteeism.**

- EXAMINING THE EMPLOYEES WITH MOST ABSENTEEISM

  On further examining the employees with most absenteeism we found a few conclusions that could be the cause for that employee:-

  1. ID 3 - Distance from residence to work is 51 km, which is very high and mode reason is 27.
  2. ID 11 – Mode reason is 19 and has 2 sons.
  3. ID 28 – Mode reason is 23 , has 2 pets.
  4. ID 20 – Mode reason is 28, distance from work is 50km.

2.2.2 Absenteeism and Reason for Absence

Our next step takes us to analysing the absenteeism trends by the reason for absence as mentioned by the ICD.

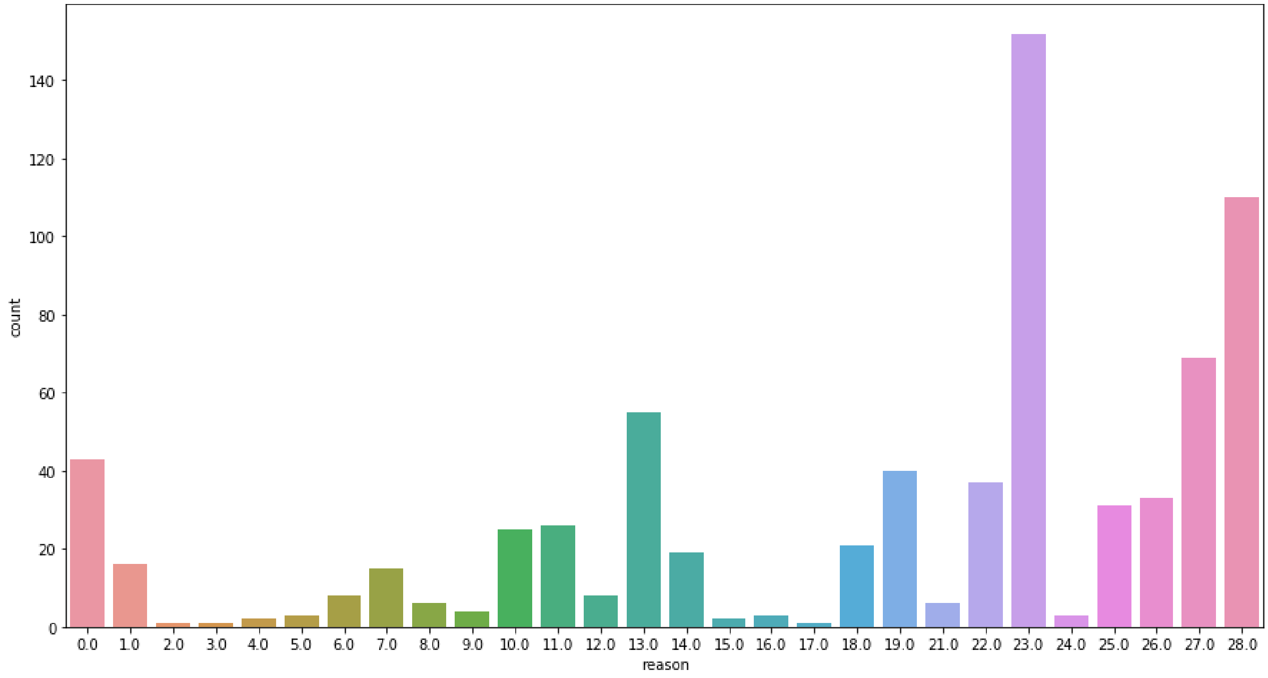


Fig 2.5 Frequency distribution of Reason for absence

Reason **23, 28 , 27, 13 & 0** have the most instances.

- Total absenteeism hours and reason

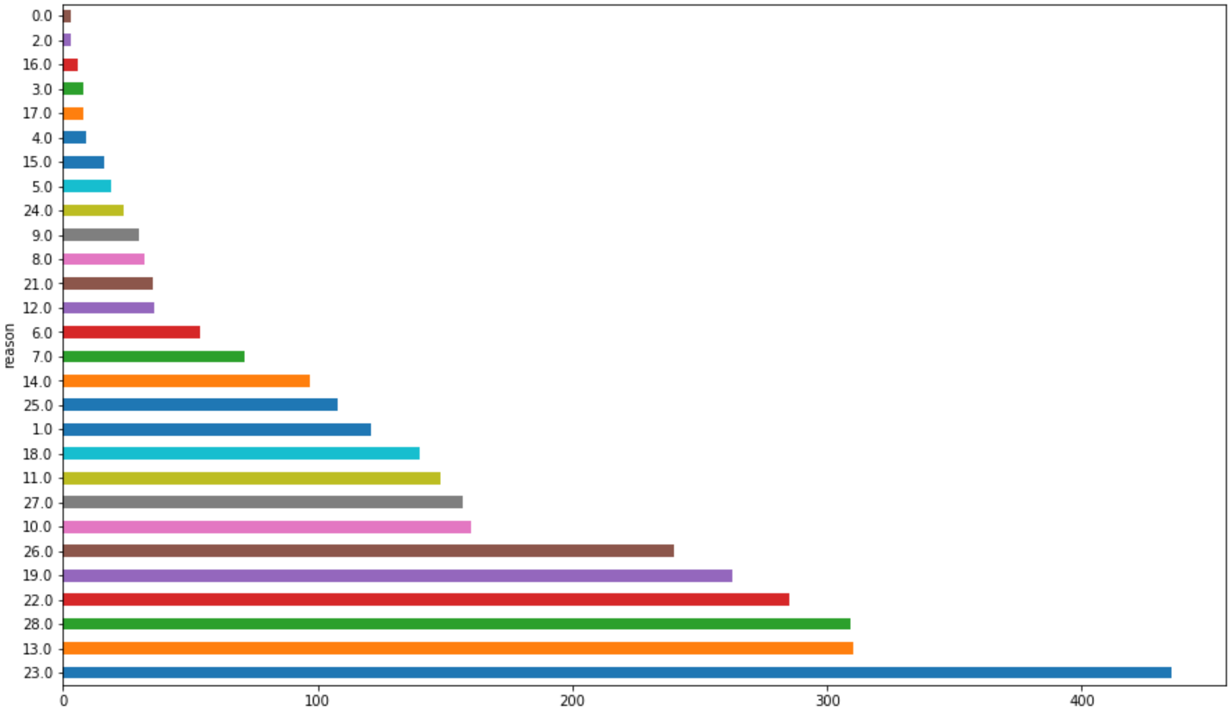


Fig 2.6 Total absenteeism hours and reason

Reason **23,13,28,22,19 & 26** contribute to almost half of the total absenteeism.

2.2.3 Absenteeism by All other Features

We will try now to analyse absenteeism by the other features present in our dataset with the help of grouping and visualisations.

1. **CATEGORICAL VARIABLES**
   - Absenteeism per month

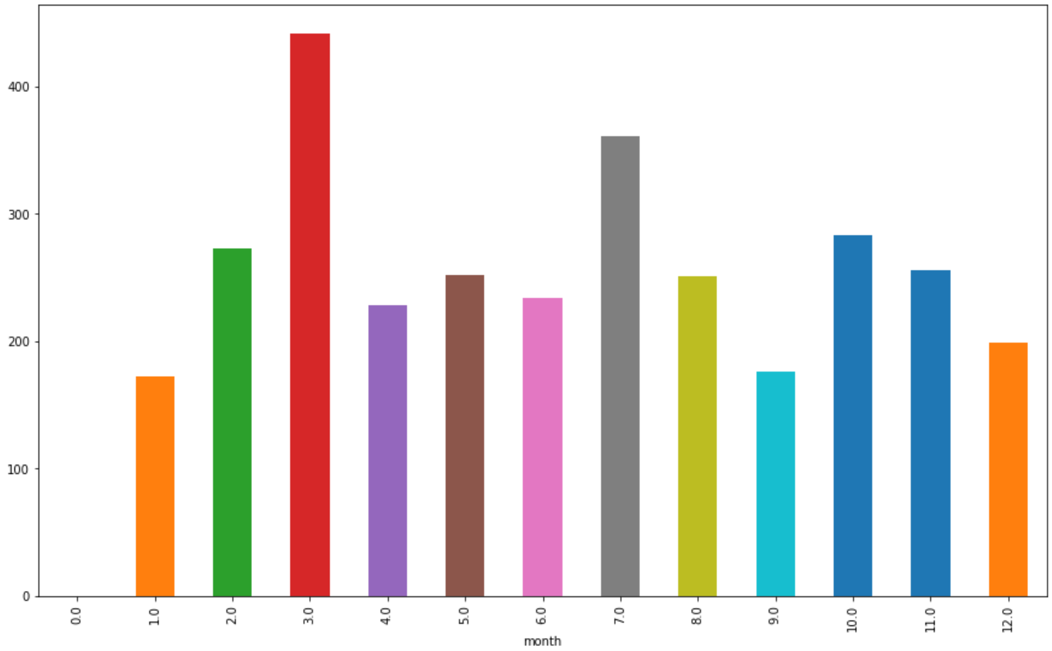


Fig 2.7 Total absenteeism hours per month

March has the highest absenteeism time followed by July.

- Absenteeism by weekday

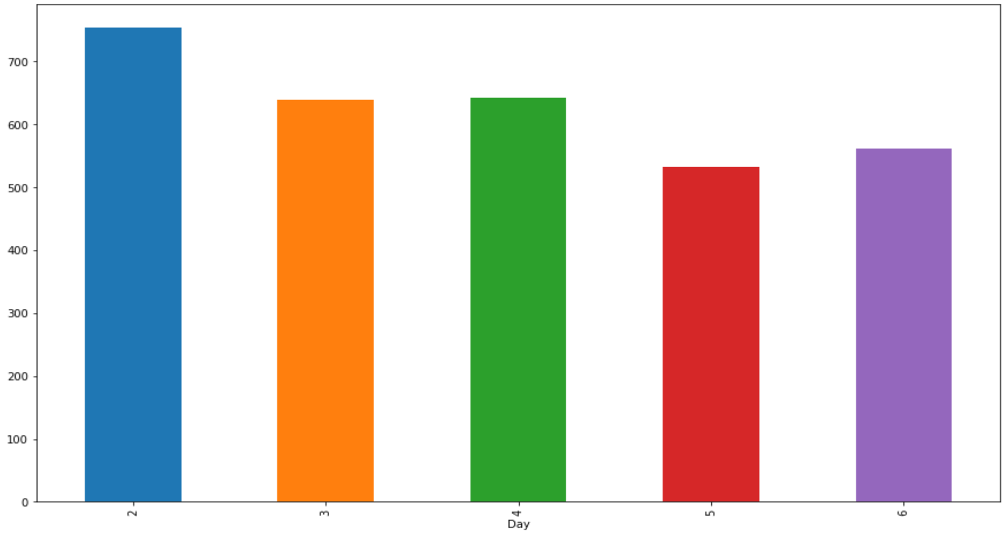


Fig 2.8 Absenteeism time by day of the week

Understandably Monday has the most hours of absenteeism.

- Absenteeism by Seasons

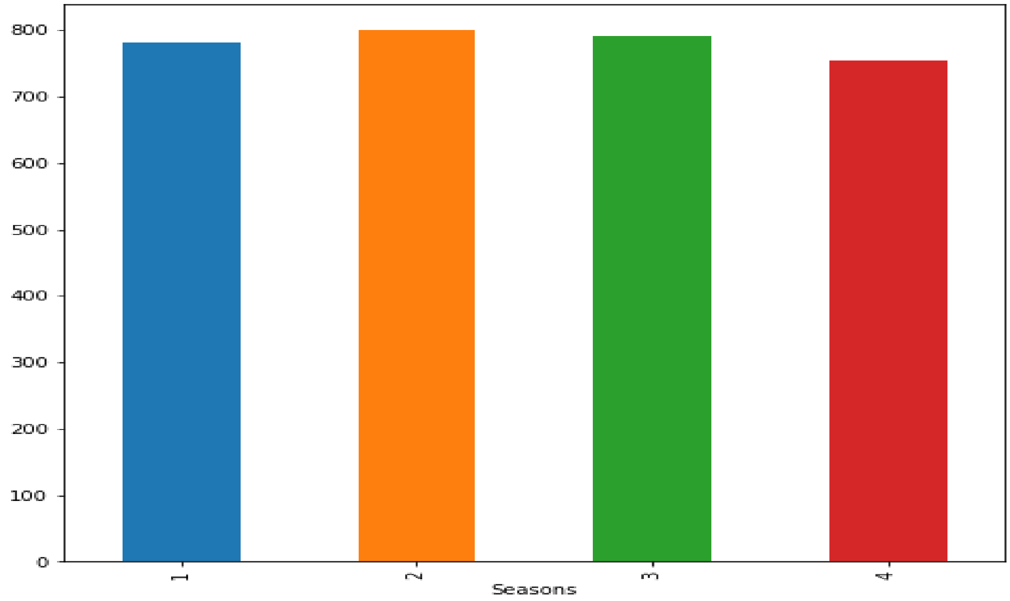


Fig 2.9 Absenteeism time by Seasons

- Absenteeism by Education

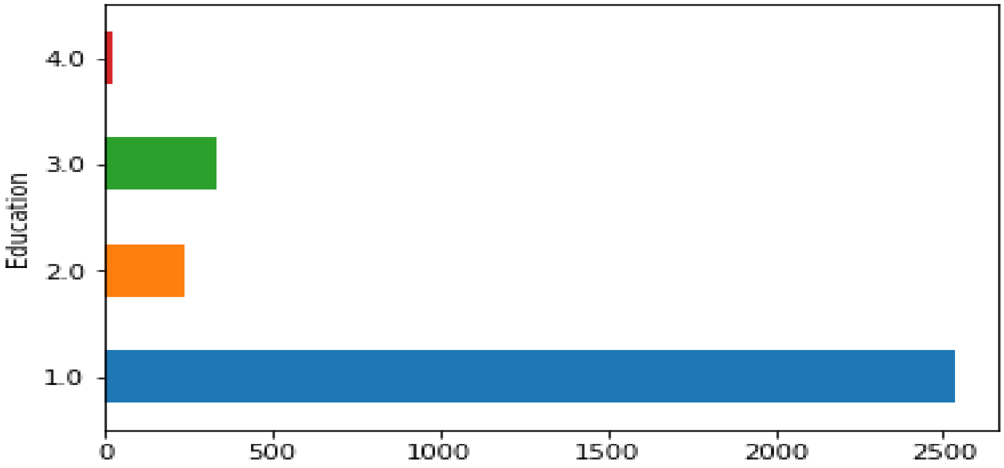


Fig 2.10 Total absenteeism time by Education

People with Education 1 constitute to about 90% of the total absenteeism.

However it is worth mentioning the Education distribution of the employees, with 28 of the 36 employees having education as 1, which makes it the obvious factor for absenteeism.

- Absenteeism by Sons

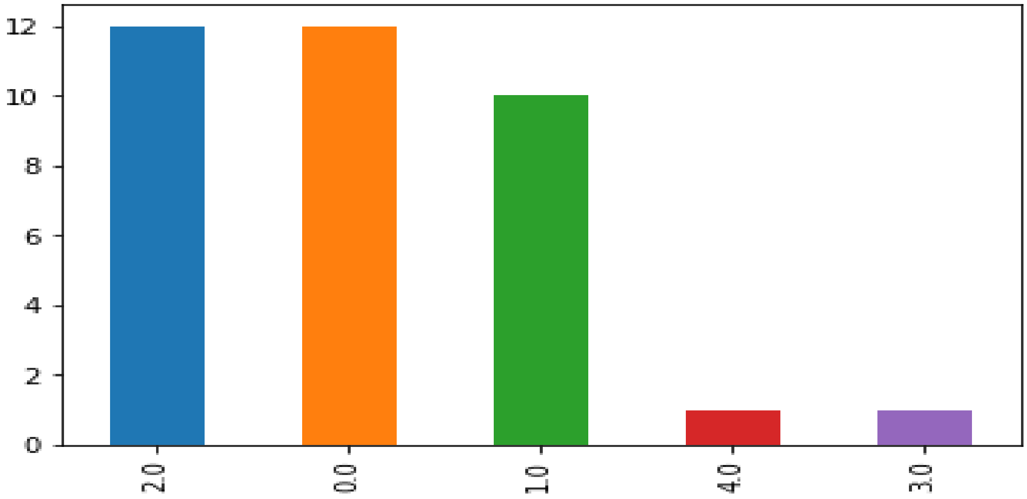


Fig 2.11 Number of Sons of Employees

12 employees each have 2 & 0 sons. 10 employees have 1 son, while 1 employee has 3 and 1 has 4 sons. Thus we can expect employees with 2,1 & 0 dominating absenteeism.

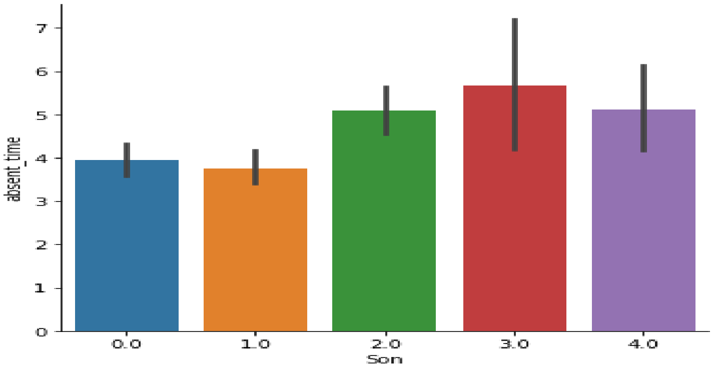


Fig 2.12 Median values of absenteeism By number of Sons

The mean value of absenteeism is most for the single employee with 3 sons, while employees with 0 & 1 sons have the least mean time of absenteeism.

- Absenteeism by Disciplinary failure

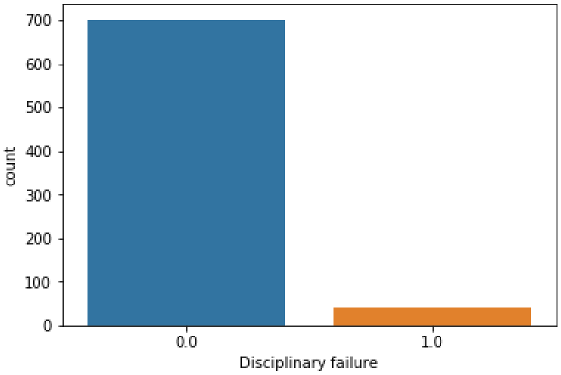


Fig 2.13 Count of disciplinary failure

As the instances of disciplinary failure are negligible, we cannot consider it as a useful predictor.

- Absenteeism by Social smoker

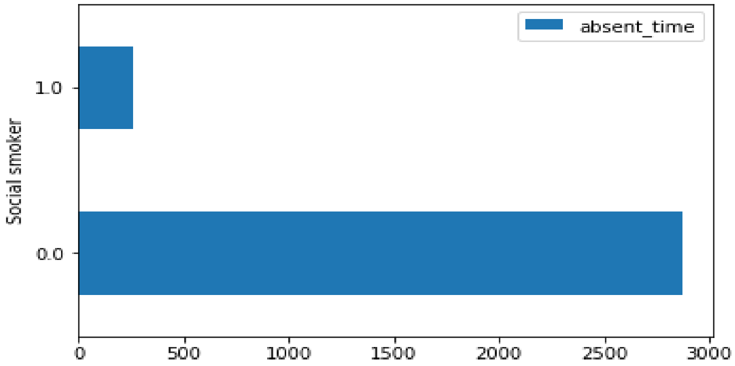


Fig 2.14 Absenteeism time by smoker

Non-smokers contribute to about 90% of absenteeism , but 32 of the 36 employees are non-smokers.

- Absenteeism by Social drinker

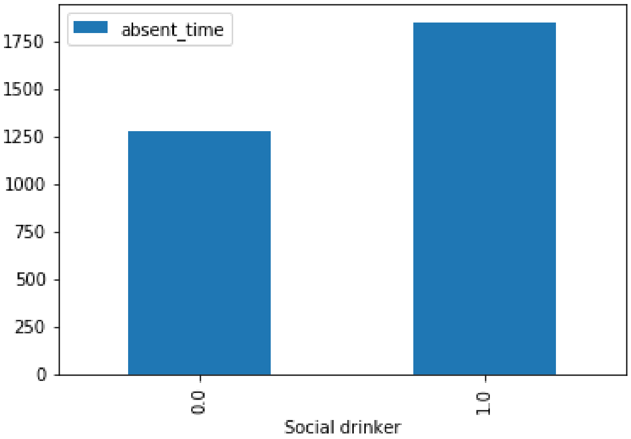
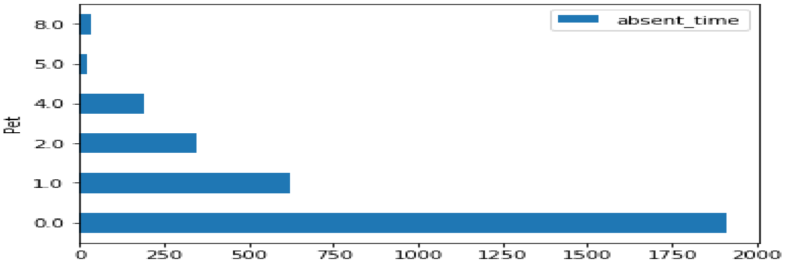


Fig 2.15 Absenteeism time by drinker

Social drinkers contribute majorly to the absenteeism but 29 of 36 employees are social drinkers.

- Absenteeism by Pets

Fig2.16 Absenteeism time by pets



People with 0 pets have about 75% of the total absenteeism time, however 19 people have 0 pets.

Thus it can be said that more pets contribute to lesser absenteeism.

2. NUMERICAL FEATURES

- Absenteeism by Age

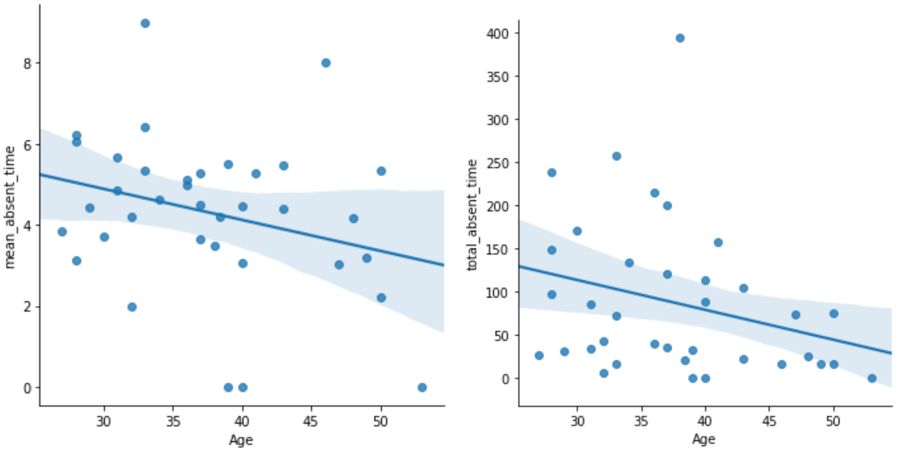


Fig 2.17 Age and absenteeism time

We can clearly see that the mean_absent_time and total_absent_time are both decreasing with increasing Age. Young people are more absent.

- Absenteeism by Service time

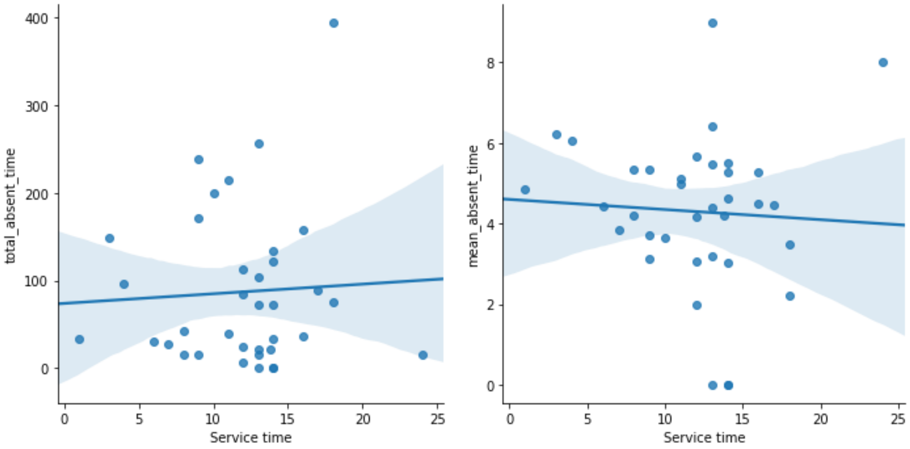


Fig 2.18 Service time and absenteeism

No clear trend here.

- Absenteeism by Transportation Expense

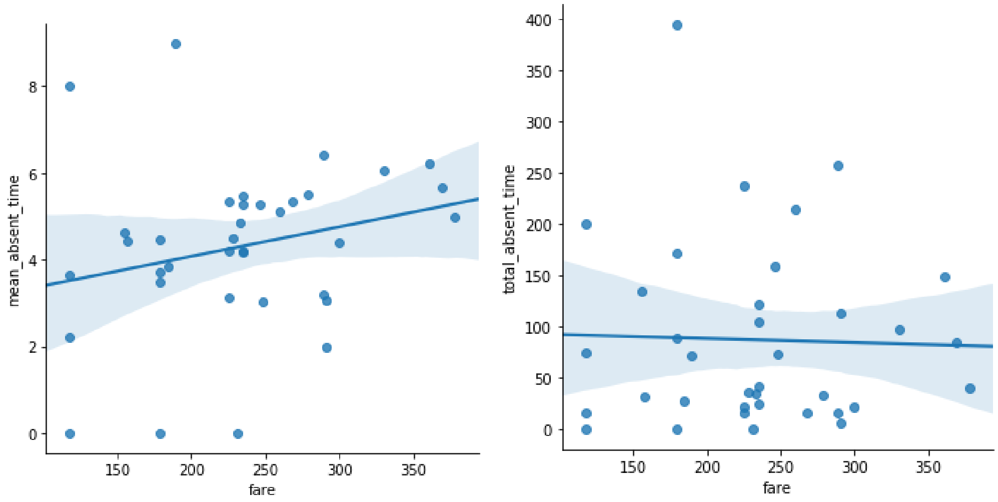


Fig 2.19 Absenteeism by transportation expense

While total absent time is unaffected by fare , the mean absent time is showing an increase with increase in fare.

- Absenteeism by Distance from residence to work

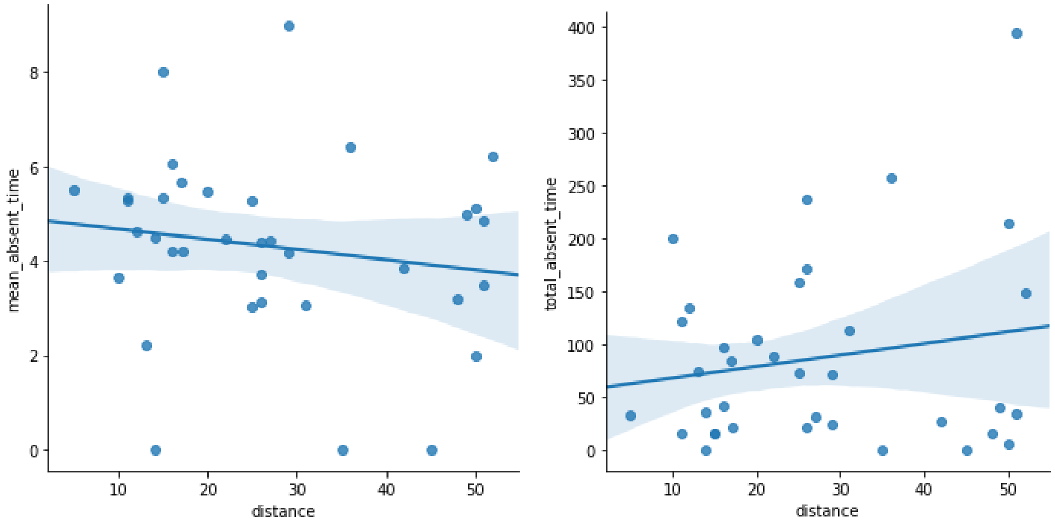


Fig 2.20 Absenteeism by distance

Total absent time is showing an increase with increasing distance while mean absent time is showing a gradual decrease with increase in the distance from residence to work.

- Absenteeism by Workload average/day

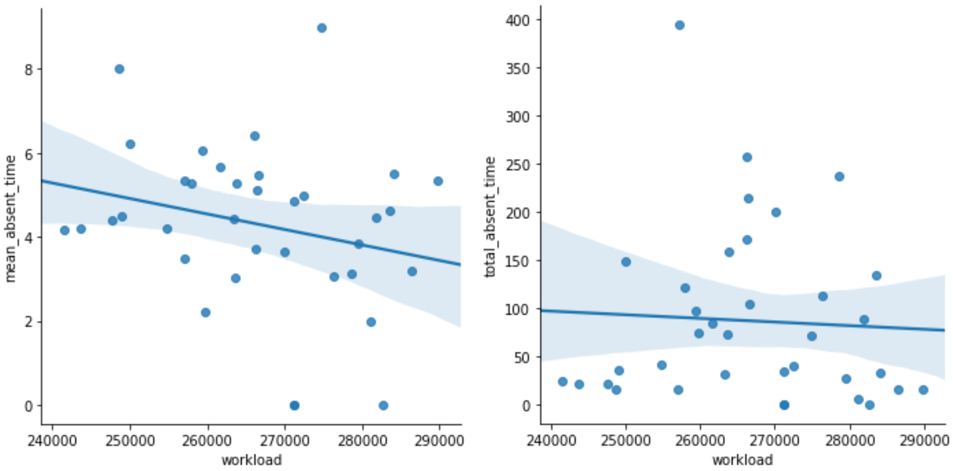


Fig 2.21 Absenteeism by workload

Mean absent time is showing a gradual decrease with increasing workload.

- Absenteeism by Hit target

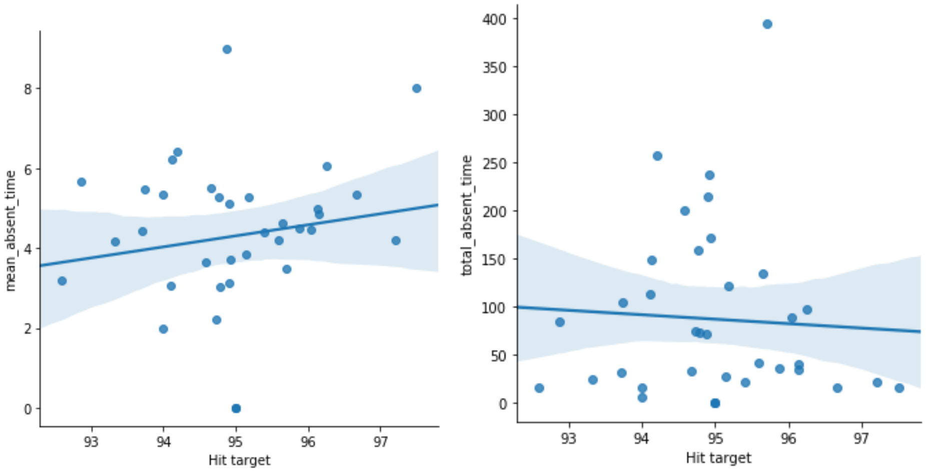


Fig 2.22 Absenteeism by Hit target

Mean absent time is increasing with Hit target.

- Absenteeism by body mass index

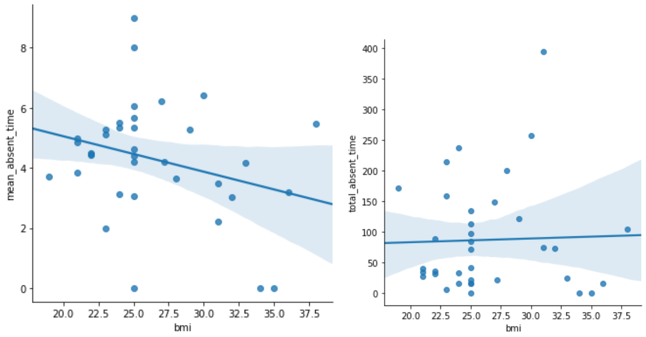


Fig 2.23 Bmi

Mean absent time is less for people with high body mass index.

2.3 Causes of Absenteeism & Possible Measures to Reduce it

After examining absenteeism in the above section we can surely list the patterns that are causing high amount of absenteeism and can also suggest ways to possibly reduce them.

1.Employee ID – Employee number 3,11,28 & 20 have been the contributor of about 50% of total absenteeism time. While they may have reasons to justify their absence the company can definitely have discussions with them personally regarding this.

2.Reason for absence- Reason 23,13,28,22 & 19 have the maximum absent time.

23 – Medical consultation :- More details regarding the consultation should be taken by the company in this case.

13- Disease of the Musculoskeletal system and connective tissues- The company can conduct seminars educating it's employees on maintaining muscle and skeleton health.

28- Dental Consultation:- Monthly free dental checkup at the company can be conducted.

26-Unjustified absence:- This should be made non acceptable.

3. Month & Day of the Week – March and Monday show very large absenteeism, so a small incentive working in March and on Mondays can reduce absenteeism.

4.Age & Service time- we found that younger people tend to be more absent in this company.

5. Distance & Transportation expense – More distance and more fare contribute to more absenteeism . To tackle this company can encourage it's employees to have residences not too far.

6. Hit target – More hit target had more mean absent time. Distributing the target more evenly might solve this problem.

2.4 Feature Selection

Selecting which features to feed in our model is an important step in model building. We want to train our model with the features that influence our target variable and help in the prediction. Features carrying no or little importance would only add to the complexity of our model. Moreover features which are highly dependent on each other also add to the complexity of the model and make our models weak. Hence we want to select the independent variables such that they have an influence on the target variable and have little influence on one another.

2.4.1 Correlation Analysis

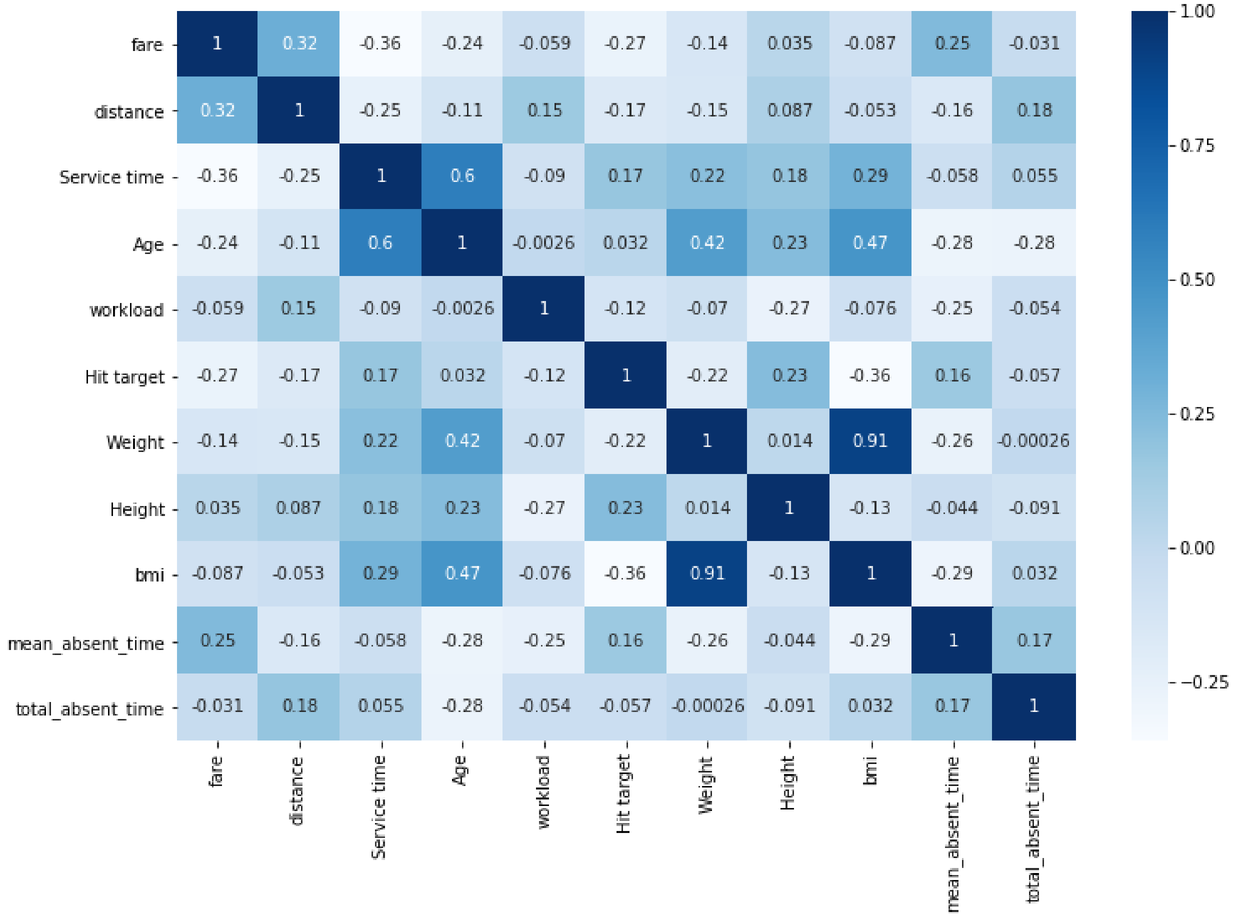


Fig 2.24 Heatmap of the continuous features

Fig 2.24 shows a correlation heatmap of all numerical variables in our dataset.

The heatmap was drawn using python's seaborn library.

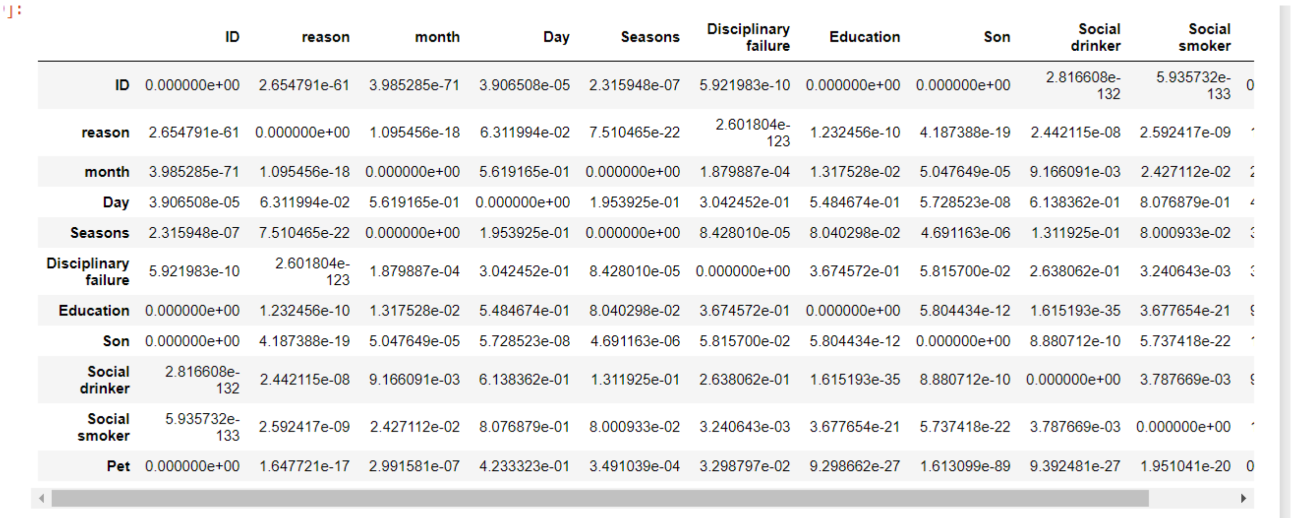
From the heatmap a few inferences can be drawn:-

1. Weight has the high positive correlation with bmi.
2. No independent variable is showing high correlation with target variable.
3. Service time, workload, height, fare are showing very low correlation with the target variable and thus they can be dropped.

2.4.2 Chi-Sqaure Test

Pearson's chi-squared test (χ^2) is a statistical test applied to sets of categorical data to evaluate how likely it is that any observed difference between the sets arose by chance.

We performed the chi-square test on our categorical independent variables to examine which are correlated to each other. The test was judged by p-value. If the p-value is less than 0.05 we concluded that the two features are correlated to each other and one of them needs to be dropped.



	ID	reason	month	Day	Seasons	Disciplinary failure	Education	Son	Social drinker	Social smoker
ID	0.000000e+00	2.654791e-61	3.985285e-71	3.906508e-05	2.315948e-07	5.921983e-10	0.000000e+00	0.000000e+00	2.816608e-132	5.935732e-133
reason	2.654791e-61	0.000000e+00	1.095456e-18	6.311994e-02	7.510465e-22	2.601804e-123	1.232456e-10	4.187388e-19	2.442115e-08	2.592417e-09
month	3.985285e-71	1.095456e-18	0.000000e+00	5.619165e-01	0.000000e+00	1.879887e-04	1.317528e-02	5.047649e-05	9.166091e-03	2.427112e-02
Day	3.906508e-05	6.311994e-02	5.619165e-01	0.000000e+00	1.953925e-01	3.042452e-01	5.484674e-01	5.728523e-08	6.138362e-01	8.076879e-01
Seasons	2.315948e-07	7.510465e-22	0.000000e+00	1.953925e-01	0.000000e+00	8.428010e-05	8.040298e-02	4.691163e-06	1.311925e-01	8.000933e-02
Disciplinary failure	5.921983e-10	2.601804e-123	1.879887e-04	3.042452e-01	8.428010e-05	0.000000e+00	3.674572e-01	5.815700e-02	2.638062e-01	3.240643e-03
Education	0.000000e+00	1.232456e-10	1.317528e-02	5.484674e-01	8.040298e-02	3.674572e-01	0.000000e+00	5.804434e-12	1.615193e-35	3.677654e-21
Son	0.000000e+00	4.187388e-19	5.047649e-05	5.728523e-08	4.691163e-06	5.815700e-02	5.804434e-12	0.000000e+00	8.880712e-10	5.737418e-22
Social drinker	2.816608e-132	2.442115e-08	9.166091e-03	6.138362e-01	1.311925e-01	2.638062e-01	1.615193e-35	8.880712e-10	0.000000e+00	3.787669e-03
Social smoker	5.935732e-133	2.592417e-09	2.427112e-02	8.076879e-01	8.000933e-02	3.240643e-03	3.677654e-21	5.737418e-22	3.787669e-03	0.000000e+00
Pet	0.000000e+00	1.647721e-17	2.991581e-07	4.233323e-01	3.491039e-04	3.298797e-02	9.298662e-27	1.613099e-89	9.392481e-27	1.951041e-20

Fig 2.25 Chi-Square Test

p-value < 0.05:-

1. Seasons and month
2. ID with Education, Son,Pet.

Thus we can drop seasons and ID for our model.

2.5 Feature Scaling

Feature scaling is a method used to standardize the range of independent variables or features of data.

The continuous features in our dataset are not normally distributed and hence we decided to use normalization to scale our features.

2.6 Feature Engineering

Feature Engineering is defined as the process of using domain knowledge of the data to create features that make machine learning algorithms work better.

While there is not much scope in our project to perform feature engineering but we will use 'One-Hot Encoding' on our predictor categorical variables.

A **one hot encoding** is a representation of categorical variables as binary vectors. This first requires that the categorical values be mapped to integer values. Then, each integer value is represented as a binary vector that is all zero values except the index of the integer, which is marked with a 1.

We take this approach rather than integer encoding because:-

1. One hot encoding removes the chances of unnecessary weighing observations of a feature even when the categories of a feature do not have linear weighted relationship.

2. Our model has less features and it is not a computational expense to improve our model by increasing a few features.

While Random Forest can handle both numerical and categorical data, linear regression cannot and we are using the one-hot encoding to feed our regression model.

For the Random Forest model we will feed the model with categorical data.

Pandas 'get_dummies' function does our required task easily.

2.7 Modelling

2.7.1 Model Selection

The first thing to consider when deciding the choice of the model is the type of target variable.

In our problem our target variable is of numeric type and it falls under regression type of problem.

There are a lot of algorithms to perform regression but we will use a linear regression and random forests regression to build our model and will use corresponding evaluation metrics to measure the performance of our model.

Before building a model we will split our Predictor(X) set and target(Y) set into training and test data for validation of our model. We will train our model on the training data and predict for the test data and would evaluate the model accordingly.

The size of the training set chosen here is 75% of all the data, while we will test our model on the remaining 25%.

For building a model Python's scikit-learn (aka sklearn) library has been used, which provides numerous machine learning algorithms and evaluation metrics as well.

Splitting the data using train_test_split in sklearn:

- x_train, x_test, y_train, y_test = train_test_split(x,y,test_size=0.25, random_state=124)

Now we have 2 pairs of object for building our model, each containing predictors and target features.

2.7.2 Linear Regression

We begin our modelling with the simplest of models, i.e. multiple linear regression.

Linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables.

- m= sm.OLS(y_train,x_train.astype(float)).fit()
- m.summary()
- p = m.predict(x_test)
- np.sqrt(mean_squared_error(y_test,p))

2.7.3 Random Forest Regressor

It is an ensemble technique that builds multiple decision trees and merges them together to get a more accurate and stable prediction. Random forests are a form of bagging, and the averaging over trees can substantially reduce instability that might otherwise result.

Applying Random Forest:

- rf=RandomForestRegressor(n_estimators=300,bootstrap=True,max_depth=30,oob_score=True, random_state = 42)
- rf.fit(x_train, y_train)
- pre = rf.predict(x_test)
- rf.score(x_train,y_train)

Parameters used:

1. n_estimators = we choose 300 trees in our model after experimenting with a few values
2. bootstrap = enabling bootstrapping
3. max_depth = the maximum depth of the tree was chosen to be 30.

Chapter 3

Conclusion

3.1 Model Evaluation

There are multiple performance evaluation metrics present, but for our case we have used the following

1. R-squared – It tells us how much variability of the target feature is explained by our predictors.
2. RMSE - Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit.
3. Feature Importance – The random forest regressor has a special function called feature importance which tells us the importance of predictor variables for the outcome.

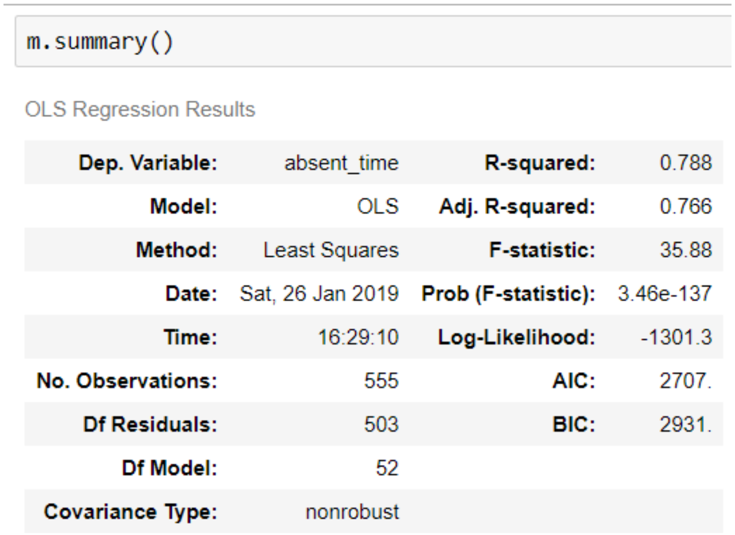
3.2 Model Performance

1. Linear Regression

R-squared – 0.788

Adjusted R-squared- 0.766

RMSE - 2.73



```
m.summary()
```

OLS Regression Results

Dep. Variable:	absent_time	**R-squared:**	0.788
Model:	OLS	**Adj. R-squared:**	0.766
Method:	Least Squares	**F-statistic:**	35.88
Date:	Sat, 26 Jan 2019	**Prob (F-statistic):**	3.46e-137
Time:	16:29:10	**Log-Likelihood:**	-1301.3
No. Observations:	555	**AIC:**	2707.
Df Residuals:	503	**BIC:**	2931.
Df Model:	52		
Covariance Type:	nonrobust		

Fig 3.1 Summary of OLS model

Around 76% of variance in our dependent variable is explained by the independent variables of our regression model. The RMSE value is 2.73, which is also good.

2. Random Forest

R-squared – 0.86

RMSE – 2.87

Feature importance:

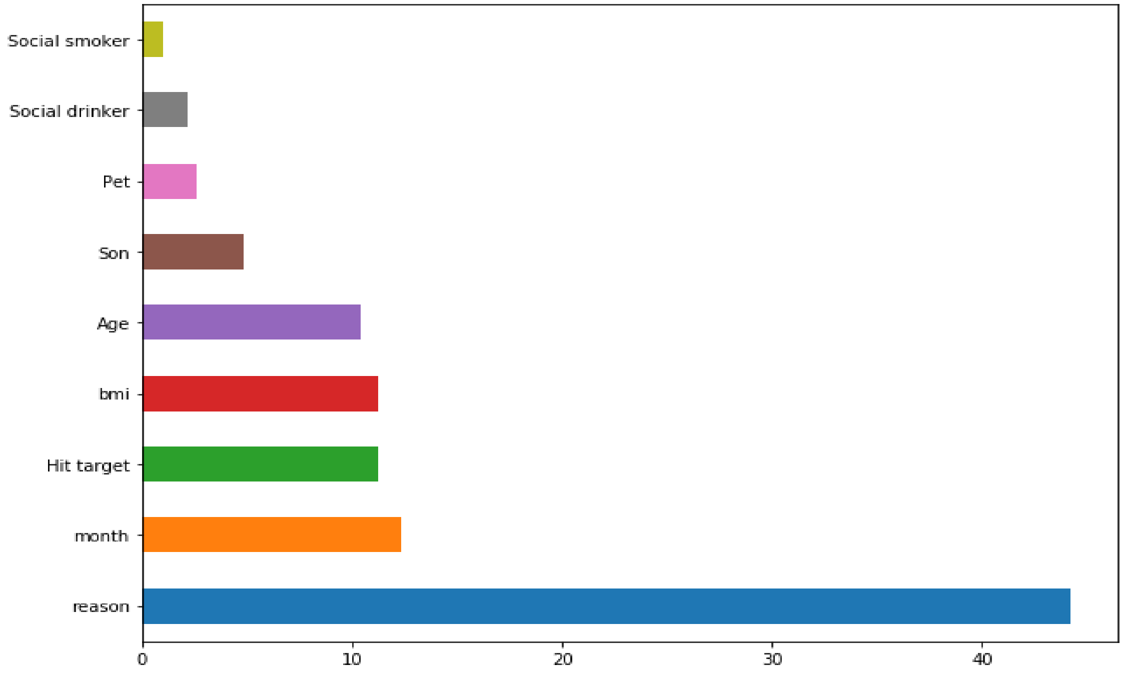


Fig 3.2 Feature importance by percentage in our model

2.3 Model Selection

Both the models are performing pretty well here and have comparable results.

Either of them can be chosen, but I'm chosing the Random Forest model here as it also explains me the feature importances.

2.4 Loss Projection

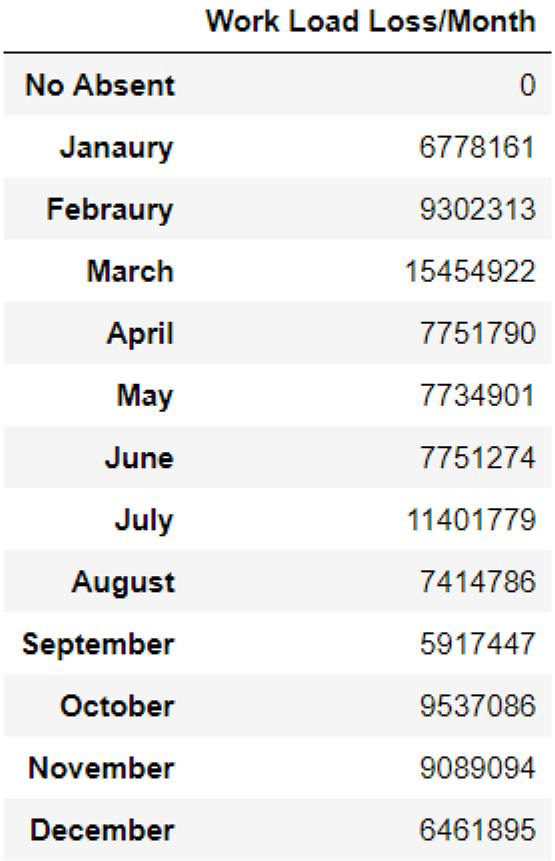
The company had asked how much losses it per month would incur if the same trend continues in the year 2011.

For doing the same we created a new dataframe having column workload loss per month.

Since we have the daily workload average and the absent time , and assuming employees work 8 hours a day we can project workload loss per month.

Daily workload loss = [(workload/day) / working hours] x Absent time

Thus the monthly workload loss came out to be:



	Work Load Loss/Month
No Absent	0
Janaury	6778161
Febraury	9302313
March	15454922
April	7751790
May	7734901
June	7751274
July	11401779
August	7414786
September	5917447
October	9537086
November	9089094
December	6461895

Fig 3.4 Workload per month loss projection