2. Exploratory Data Analysis :

The dataset has 737 x 10 dimensions, where out of 10 there is one dependent variable called *absenttime* in the dataset. After reviewing dataset we noticed that there is no year feature, although it is mentioned in the description that the data is from 2007 to 2010, so we added a new feature *year*. We split the data set into train and test with dimensions 600x10 and 137x10 respectively. Below is the table 1 generated using *describe* function of library *psych*. From the summary table we can observe that there are no missing values in the dataset,

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
month	1	737	6.35	3.42	6	6.31	4.45	1	12	11	0.07	-1.27	0.13
day	2	737	3.91	1.42	4	3.89	1.48	2	6	4	0.10	-1.29	0.05
season	3	737	2.55	1.11	3	2.56	1.48	1	4	3	-0.04	-1.35	0.04
transexp	4	737	221.51	66.96	225	218.09	68.20	118	388	270	0.39	-0.33	2.47
distance	5	737	29.62	14.84	26	29.40	16.31	5	52	47	0.31	-1.26	0.55
servtime	6	737	12.55	4.39	13	12.75	5.93	1	29	28	0.00	0.65	0.16
age	7	737	36.42	6.46	37	35.80	5.93	27	58	31	0.69	0.42	0.24
childen	8	737	1.02	1.10	1	0.85	1.48	0	4	4	1.08	0.72	0.04
bmi	9	737	26.66	4.27	25	26.63	4.45	19	38	19	0.30	-0.31	0.16
absenttime	10	737	6.95	13.35	3	4.35	2.97	0	120	120	5.69	38.27	0.49

Table 1: Summary table of AbsenteeismData dataset

whereas the standard deviation of *transexp* is very high and for *distance* feature and *absenttime* dependent variable is high compared to other variables. The skewness in absenttime can be observed due to potential outliers present in data, as the maximum value it has is 120 whereas the median is 3 and the mean is 6.68.

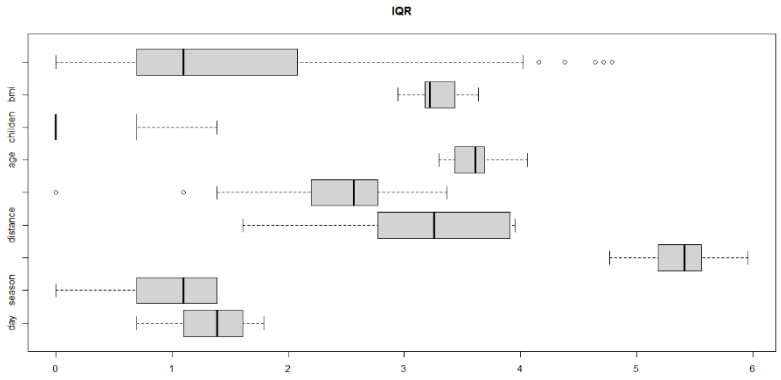


Fig 1: Box plots

From the figure on left we can clearly see that there are outliers to absenttime and servtime.

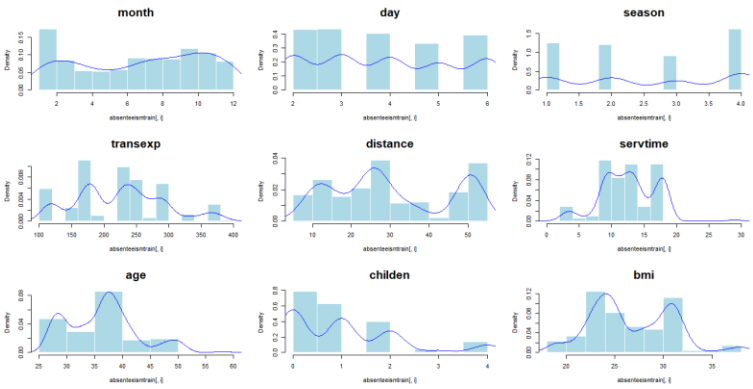


Fig 2: Distribution and density of features.

From the figure 2 plots we can say that none of the features are normally distributed, where month has 12 distinct values, day has 4, season has etc.

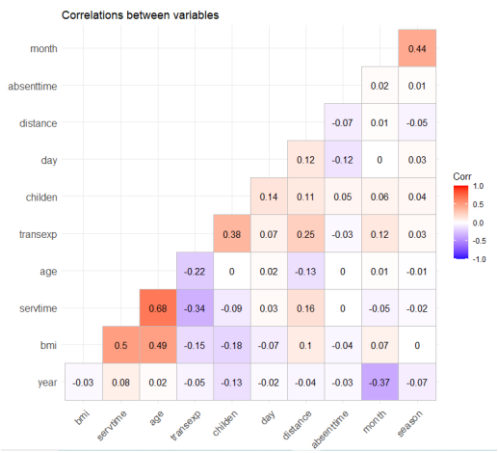
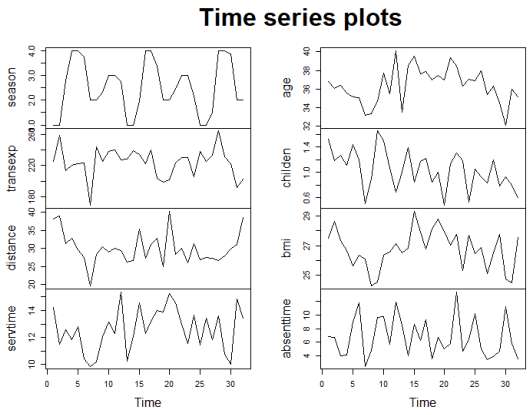


Fig 3: Correlation of features

In figure 3 from the correlation matrix we can observe that the *servtime* & *age* have the highest correlation, 0.68. *bmi* have the second and third highest correlation with *servtime* and *age*, 0.5 & 0.49 respectively. Whereas servtime and *transexp* have the highest correlation in negative, ie -0.34. All the other combinations seems to be uncorrelated with each other.

Fig 4 : Time series plots of features



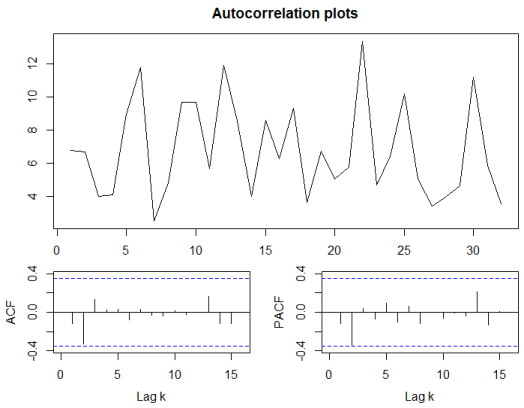
After exploratory analysis we converted the training data set into time series.

Table 2 : Stationary test results

Test	p-Value
ADF	0.04377
KPSS	0.1
Box-Ljung	0.4828

From the stationarity test we came to the conculsion that the data is stationary.

Fig 5 : Autocorrelaation plot



From Autocorrelation analysis it is clear that in our time series there are no components like trend, seasonality or cyclic.

3. Fit a statistical model & Forecasting:

1. ARIMA Model

Due to the high correlation of absenttime with transexp and children, we applied ARIMA model, where transexp and children were used as regressors individually and in combination. Following are the results of ARIMA model:

Table 3: ARIMA results

	Regressors			
	arima_1	*arima_2*	*arima_3*	*arima_4*
ARIMA	No	children	transexp	children+transexp
AIC	162.32	160.67	160.55	160.5
BIC	165.25	165.07	163.48	163.43

For all the combinations we received ARIMA(0,0,0) model as result, in qqplot the residuals are more scattered and there are no lags in ACF test, so we rejected this model.

2. Linear regression model

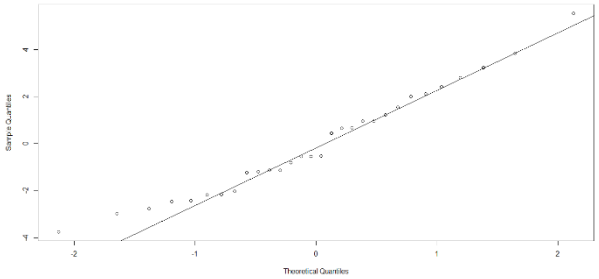
Table 4: TSLM results

Model	p-Value	Regressors	
tslm_1	3.141	*children*	>0.05
tslm_2	0.03579	*transexp*	<0.05

We also applied a linear regression model where we used transexp and children as regressor and we got respective p-values as mentioned in table 4, so tslm_2 is more significant, but upon doing the residual analysis for it in qqplot the residuals are more scattered and there are no lags in ACF test, so we rejected this model.

3. Neural Network Model

Fig 6: QQNORM plot



We tried NNAR model where we get NNAR(2,2) and sigma^2 = 4.952, so to check the significance of model we did residual analysis. In qqnorm plot it is clearly visible that the residual points are very close to the qqline and acf and pcf significancant lags which proves its significance.

Forecasting

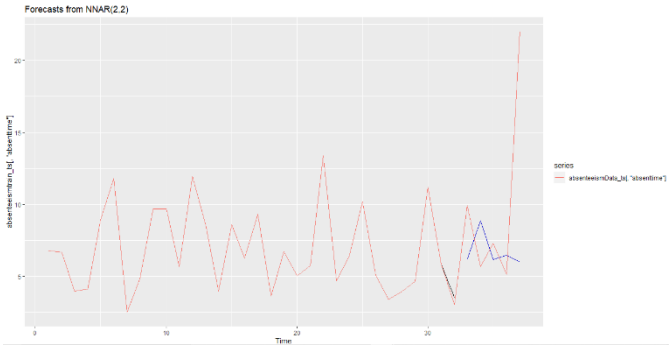


Fig 7: Forecast graph of prdicted values

Form forecasted graph on left we can claim that it shows better promising prediction values than the other models.

Weakness

Inorder to do in modelling and forecasting we required to have more data values which we certainly don't have in this case. The prediction seems not be accurate because of same issue. We didn't had AIC and BIC values for neural network model which made us to resort with residual analysis only to check its significance.