Machine Learning

ITE2011

Absenteeism at Work

FINAL REPORT

Submitted By

Name	Registration No.
Rahul Ghanghas	16BIT0065
Aishmita Kakkar	16BIT0319

Faculty – Prof. PRABUKUMAR M

Slot – F2+TF2

VIT®
UNIVERSITY
(Estd. u/s 3 of UGC Act 1956)
VELLORE ■ CHENNAI
www.vit.ac.in

January 2019

Introduction

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 its dataset and has requested to have an answer on the following areas:

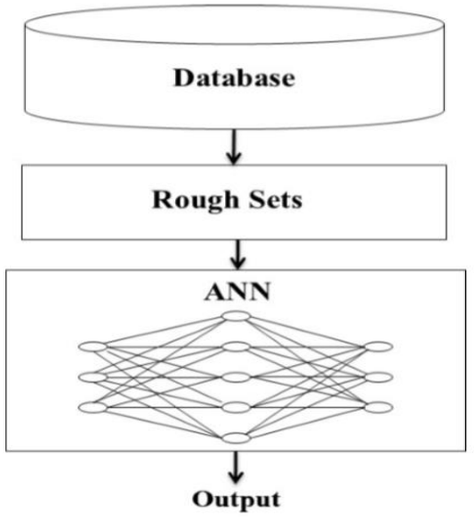
1. What changes company should bring to reduce the number of absenteeism?
2. How much loss every month can we project in 2011 if same trend of absenteeism continues?

Abstract

The high competitiveness in the market, professional development combined with the development of organizations and the pressure to reach increasingly audacious goals, create increasingly overburdened employees and end up acquiring some disturbance in the state of health related to the type of work activity, including depression considered the evil of the 21st century, taking employees to absenteeism. Absenteeism is defined as absence to work as expected, represents for the company the loss of productivity and quality of work.

The purpose of this project is to apply an artificial neural network to prediction of absenteeism at work. The database used in the experiment has 21 attributes and 741 records from documents that prove that they are absent from work and were collected from July 2007 to July 2010. The methodological synthesis of the paper consists of the modeling of an Artificial Neural Network (ANN), the 21 attributes were reduced to 11 attributes through the Rough Sets, and these attributes were used in the experiments to prediction of absenteeism.

ANN they are models consisting of simple processing units, called artificial neurons, these models are inspired by the structure of the brain and aim to simulate human behavior, such as learning, association, generalization and abstraction when submitted to training. The experiments with the ANN presented the expected results in prediction of absenteeism at work. Therefore, it is concluded that the ANN can be applied in the prediction of absenteeism at work.



Attribute Information:

1. Individual identification (ID)
2. Reason for absence (ICD).

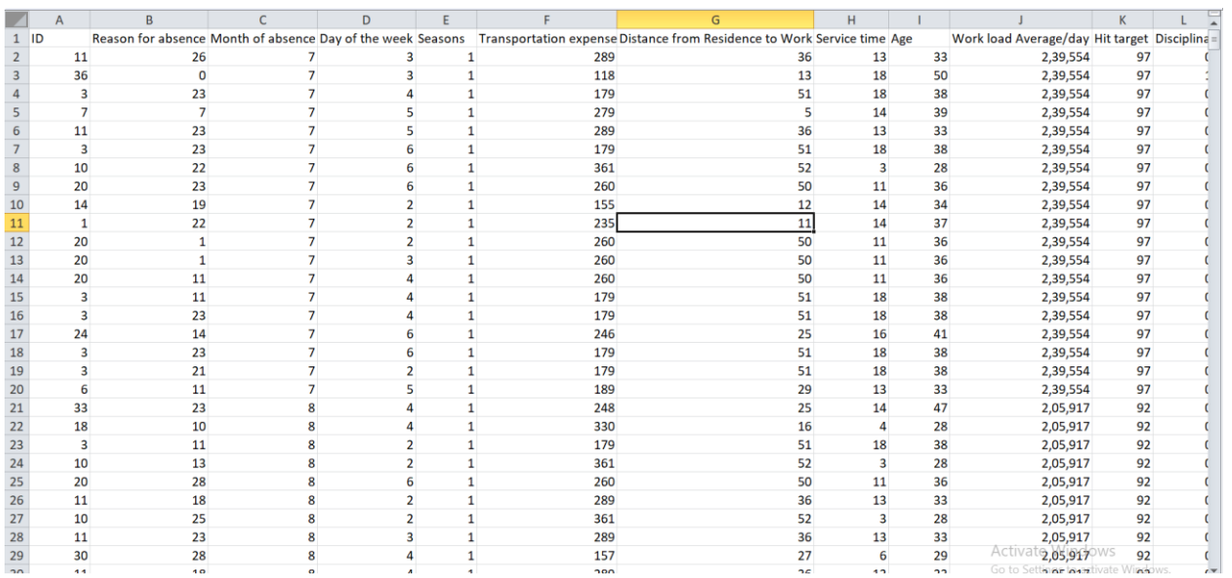
Absences attested by the International Code of Diseases (ICD) stratified into 21 categories (I to XXI) as follows:

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 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 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)

Database (Time-Series Data)



	A	B	C	D	E	F	G	H	I	J	K	L
1	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	Hit target	Disciplina
2	11	26	7	3	1	289	36	13	33	2,39,554	97	0
3	36	0	7	3	1	118	13	18	50	2,39,554	97	1
4	3	23	7	4	1	179	51	18	38	2,39,554	97	0
5	7	7	7	5	1	279	5	14	39	2,39,554	97	0
6	11	23	7	5	1	289	36	13	33	2,39,554	97	0
7	3	23	7	6	1	179	51	18	38	2,39,554	97	0
8	10	22	7	6	1	361	52	3	28	2,39,554	97	0
9	20	23	7	6	1	260	50	11	36	2,39,554	97	0
10	14	19	7	2	1	155	12	14	34	2,39,554	97	0
11	1	22	7	2	1	235	11	14	37	2,39,554	97	0
12	20	1	7	2	1	260	50	11	36	2,39,554	97	0
13	20	1	7	3	1	260	50	11	36	2,39,554	97	0
14	20	11	7	4	1	260	50	11	36	2,39,554	97	0
15	3	11	7	4	1	179	51	18	38	2,39,554	97	0
16	3	23	7	4	1	179	51	18	38	2,39,554	97	0
17	24	14	7	6	1	246	25	16	41	2,39,554	97	0
18	3	23	7	6	1	179	51	18	38	2,39,554	97	0
19	3	21	7	2	1	179	51	18	38	2,39,554	97	0
20	6	11	7	5	1	189	29	13	33	2,39,554	97	0
21	33	23	8	4	1	248	25	14	47	2,05,917	92	0
22	18	10	8	4	1	330	16	4	28	2,05,917	92	0
23	3	11	8	2	1	179	51	18	38	2,05,917	92	0
24	10	13	8	2	1	361	52	3	28	2,05,917	92	0
25	20	28	8	6	1	260	50	11	36	2,05,917	92	0
26	11	18	8	2	1	289	36	13	33	2,05,917	92	0
27	10	25	8	2	1	361	52	3	28	2,05,917	92	0
28	11	23	8	3	1	289	36	13	33	2,05,917	92	0
29	30	28	8	4	1	157	27	6	29	2,05,917	92	0

	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Service time	Age	Work load Average/day	Hit target	Disciplinar	Education	Son	Social drin	Social smc	Pet	Weight	Height	Body mass index	Absenteeism time in hours
2	13	33	2,39,554	97	0	1	2	1	0	1	90	172	30	4
3	18	50	2,39,554	97	1	1	1	1	0	0	98	178	31	0
4	18	38	2,39,554	97	0	1	0	1	0	0	89	170	31	2
5	14	39	2,39,554	97	0	1	2	1	1	0	68	168	24	4
6	13	33	2,39,554	97	0	1	2	1	0	1	90	172	30	2
7	18	38	2,39,554	97	0	1	0	1	0	0	89	170	31	2
8	3	28	2,39,554	97	0	1	1	1	0	4	80	172	27	8
9	11	36	2,39,554	97	0	1	4	1	0	0	65	168	23	4
10	14	34	2,39,554	97	0	1	2	1	0	0	95	196	25	40
11	14	37	2,39,554	97	0	3	1	0	0	1	88	172	29	8
12	11	36	2,39,554	97	0	1	4	1	0	0	65	168	23	8
13	11	36	2,39,554	97	0	1	4	1	0	0	65	168	23	8
14	11	36	2,39,554	97	0	1	4	1	0	0	65	168	23	8
15	18	38	2,39,554	97	0	1	0	1	0	0	89	170	31	1
16	18	38	2,39,554	97	0	1	0	1	0	0	89	170	31	4
17	16	41	2,39,554	97	0	1	0	1	0	0	67	170	23	8
18	18	38	2,39,554	97	0	1	0	1	0	0	89	170	31	2
19	18	38	2,39,554	97	0	1	0	1	0	0	89	170	31	8
20	13	33	2,39,554	97	0	1	2	0	0	2	69	167	25	8
21	14	47	2,05,917	92	0	1	2	0	0	1	86	165	32	2
22	4	28	2,05,917	92	0	2	0	0	0	0	84	182	25	8
23	18	38	2,05,917	92	0	1	0	1	0	0	89	170	31	1
24	3	28	2,05,917	92	0	1	1	1	0	4	80	172	27	40
25	11	36	2,05,917	92	0	1	4	1	0	0	65	168	23	4
26	13	33	2,05,917	92	0	1	2	1	0	1	90	172	30	8
27	3	28	2,05,917	92	0	1	1	1	0	4	80	172	27	7
28	13	33	2,05,917	92	0	1	2	1	0	1	90	172	30	1
29	6	29	2,05,917	92	0	1	0	1	1	0	75	185	22	4

Literature Review

1. According to **Dakely C.A. (1948)** "Absenteeism is the ratio of the number of production man-days or shifts lost to the total number of production scheduled to work". The labour bureau (1962) defines absenteeism as the total shifts lost because of absence as percentage of the total number of man shifts scheduled to work.

2. According to **Likewise Hackett J.D. (1929)** defines it as "the temporary cessation of work for not less than one whole working day initiative of the worker when his presence is expected by the employer". Similarly encyclopaedia of social science observes "Absenteeism as the time lost in industrial establishment by avoidable or unavoidable absence of employees. The time lost by the strikes or by lateness amounting to an hour or two is not usually included".

3. **Ernest B. Akyeampong** has written a research paper Trends and seasonality in Absenteeism. In this paper the author focus on that at which time period the employees are more absent. In this paper he said that illness-related absences are highly seasonal, reaching a peak during the winter months (December to February) and a trough during the summer (June to August). The high incidence in winter is likely related to the prevalence of communicable diseases at that time, especially colds and influenza. The low incidence during the summer may be partly because many employees take their vacation during these months. Because of survey design, those who fall ill during vacation will likely report vacation rather than sickness or disability as the main reason for being away

from work.Compared with the annual average, part-week absences are roughly 30% more prevalent in the winter months and almost 20% less soduring the summer months. Seasonality is much less evident in fullweek absences.

4. Mariajosé Romero and Young-Sun Lee has written a research paper A National Portrait of Chronic Absenteeism in the Early Grades. In this paper he focused on the following points:

(i) How widespread is the Problem of Early Absenteeism?

(ii) Does Family Incomes Impact Early Absenteeism?

(iii) What is the Impact of Early Absenteeism on Academic Achievement?

5. **Wolter H.J. Hassink& Pierre Koning (2009**) find statistically significant differences in absence patterns across groups of workers with different eligibility statuses depending on their attendance records and whether they had previously won. One finding is that absenteeism rose among workers who, having won already, were ineligible for further participation. Nevertheless, and although the reduction in firm-wide absence associated with the lottery drifted from2.4 percentage points to 1.1 percentage points after seven months, the authors conclude that the lottery was of net benefit to the firm.

6. **Morten Nordberg and Knut Røed** have written a research paper Absenteeism, Health Insurance, and Business Cycles. In this he wants to evaluate how the economic environment affects worker absenteeism and he also isolate the causal effects of business cycle developments on work-resumption prospects for ongoing absence spells, by conditioning on the state of the business cycle at the moment of entry into sickness absence.

- Business cycle improvements yield lower work-resumption rates forepersons that are absent, and higher relapse rates for persons who have already resumed work.
- Absence sometimes represents a health investment, in the sense that longer absence now reduces the subsequent relapse propensity.
- The work-resumption rate increases when sickness benefits are exhausted, but that work-resumptions at this point tend to be short-lived.

Drawing on the compatibility principle in attitude theory, we found that overall job attitude (job satisfaction and organizational commitment) provides increasingly powerful prediction of more integrative behavioral criteria (focal performance, contextual performance, lateness, absence, and turnover combined). The principle was sustained by a combination of metaanalysis and structural equations showing better fit of unified versus diversified models of meta-analytic correlations between those criteria. Overall job attitude strongly predicted a higher-order behavioral construct, defined as desirable contributions made to one's work role.

The attitudinal and behavioral effects of being promoted and being rejected for promotion were examined in a quasi-experiment conducted at an international bank in Hong Kong. Promoted tellers who had more internal louses of control (LOC) maintained improved attitudes across 3- and 18-month posttest intervals. Attitudes returned to baseline levels by the second posttest among external-LOC individuals who had been promoted. There was no change in attitudes among people passed over for promotion. Absenteeism and job performance both decreased among promotes. The implications for the administration of promotions are considered.

7. **C.Swarnalatha And G.Sureshkrishna:(2013)** Absenteeism – A Menace To Organization In Building Job Satisfaction Among Employees In Automotive Industries In India, Absenteeism Results In Financial Losses Both Because Of The Resultant Reduction In Productivity And The Cost Of Sick Leave Benefits Or Others Are Paid As Wages For No Work. Absenteeism Reduces The Satisfaction Level Of The Employee And Makes Him Unsecured About His Job In The Organization.

8. **Prakash K. Kannan:(2012)** A Study On Absenteeism Of Employees Among Food Retailing In Coimbatore, Their Study Concludes That Absenteeism Can Be Reduced To A Great Extent If The Management Takes Initiative In Making The Workers Feel Responsible Towards Their Job By Introducing Various Motivational Schemes.

OVERVIEW OF THE DATASET:

	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	Hit target
1	11	26.0	7.0	3	1	289.0	36.0	13.0	33.0	239554.0	97.0
2	36	0.0	7.0	3	1	118.0	13.0	18.0	50.0	239554.0	97.0
3	3	23.0	7.0	4	1	179.0	51.0	18.0	38.0	239554.0	97.0
4	7	7.0	7.0	5	1	279.0	5.0	14.0	39.0	239554.0	97.0
5	11	23.0	7.0	5	1	289.0	36.0	13.0	33.0	239554.0	97.0

Fig: 1.1 Columns 1to 11 of the dataset

Disciplinary failure	Education	Son	Social drinker	Social smoker	Pet	Weight	Height	Body mass index	Absenteeism time in hours
0.0	1.0	2.0	1.0	0.0	1.0	90.0	172.0	30.0	4.0
1.0	1.0	1.0	1.0	0.0	0.0	98.0	178.0	31.0	0.0
0.0	1.0	0.0	1.0	0.0	0.0	89.0	170.0	31.0	2.0
0.0	1.0	2.0	1.0	1.0	0.0	68.0	168.0	24.0	4.0
0.0	1.0	2.0	1.0	0.0	1.0	90.0	172.0	30.0	2.0

Fig 1.2 Columns 11 to 21 of the dataset

As we can see, there are 8 categorical variables and 12 numeric variables. The dependent variable or the target variable is Absenteeism time in hours, i.e. the number of hours employees remaining absent.

Exploratory Data Analysis (EDA)

Missing Value Analysis:

ID	0
Reason for absence	3
Month of absence	1
Day of the week	0
Seasons	0
Transportation expense	7
Distance from Residence to Work	3
Service time	3
Age	3
Work load Average/day	10
Hit target	6
Disciplinary failure	6
Education	10
Son	6
Social drinker	3
Social smoker	4
Pet	2
Weight	1
Height	14
Body mass index	31
Absenteeism time in hours	22

Fig: 2.1 Missing values in variables of the raw dataset

Fig 2.1 shows the missing values in variables of the raw dataset, mostly the missing values are less than 10 in all variables except Height, Body mass index, and Absenteeism time.

Few insights on the missing values:

1] Taking all the rows containing one or more than one NA came out to be 101. Thus, although there were 145 missing values in total, number of rows came out to be 101 because some rows had more than 1 NA.

2] Dropping more than 100 rows in an 800+ rows dataset was not feasible, so performed correlation analysis beforehand to check multi collinearity. The result showed Body mass index was redundant variable which had the most missing values, so dropped it before any imputation task. Now, it was fine to drop all NAs since BMI had gone, and there were no missing values in absenteeism time in hours as this was our target variable and all NAs were deleted from it, because it is not good to impute values in target variable as advised, this ultimately made us left with only few missing values in 20s. Although this was an option, but all the proceedings are done by imputing other variables and dropping BMI and missing rows of target variable.

Visualization of the data:

We cannot undermine the fact that, though simple, barplots present the best and swiftest results in getting us the feel of dataset. Thus, will be seeing important variables' plots, mostly with our target variable. The result of graphs can change to an extent that it can lead to changes in answer if we use imputed/processed dataset for plotting graphs.

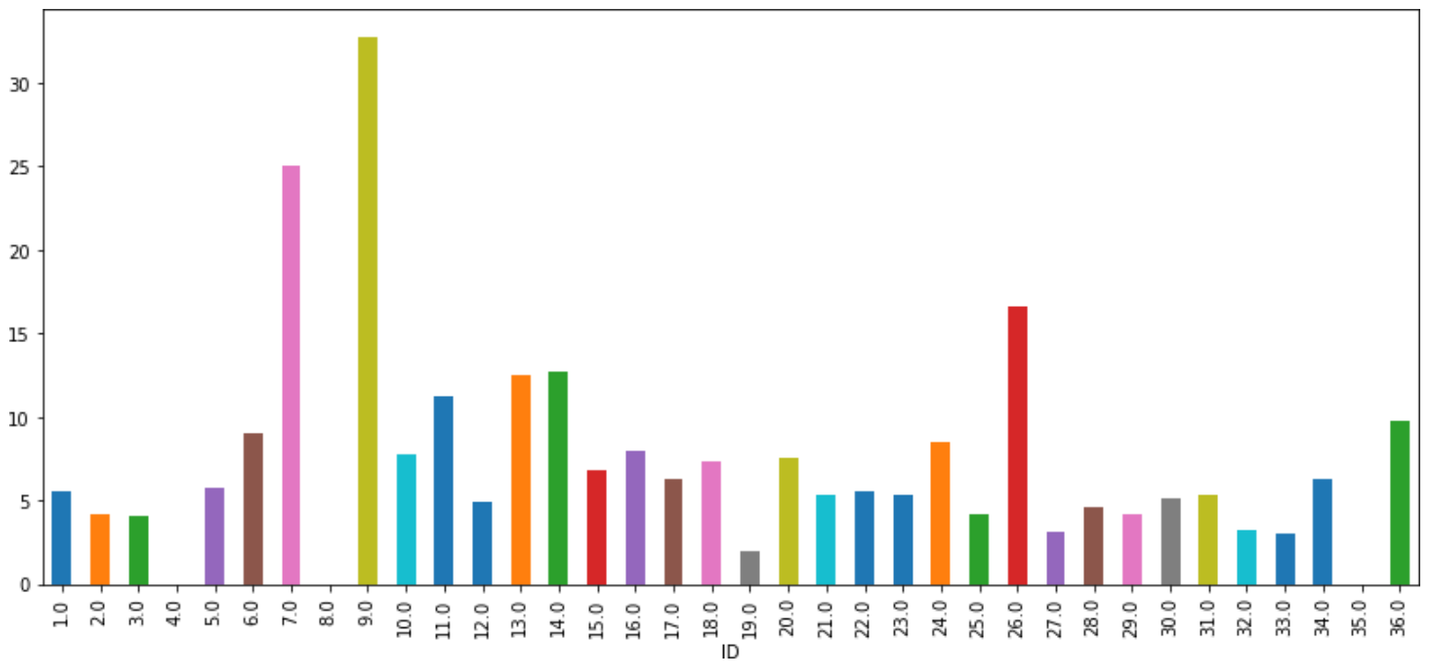


Fig: 2.2a Mean absenteeism hours for each ID

The figure 2.2a shows the mean of hours that persons with ID following from 1 to 36

were absent for the whole period. It shows that person with ID: 9 and 7 had the most number of hours remaining as absent, while ID:4, 8, 35 had the least or Zero number of hours of absenteeism.

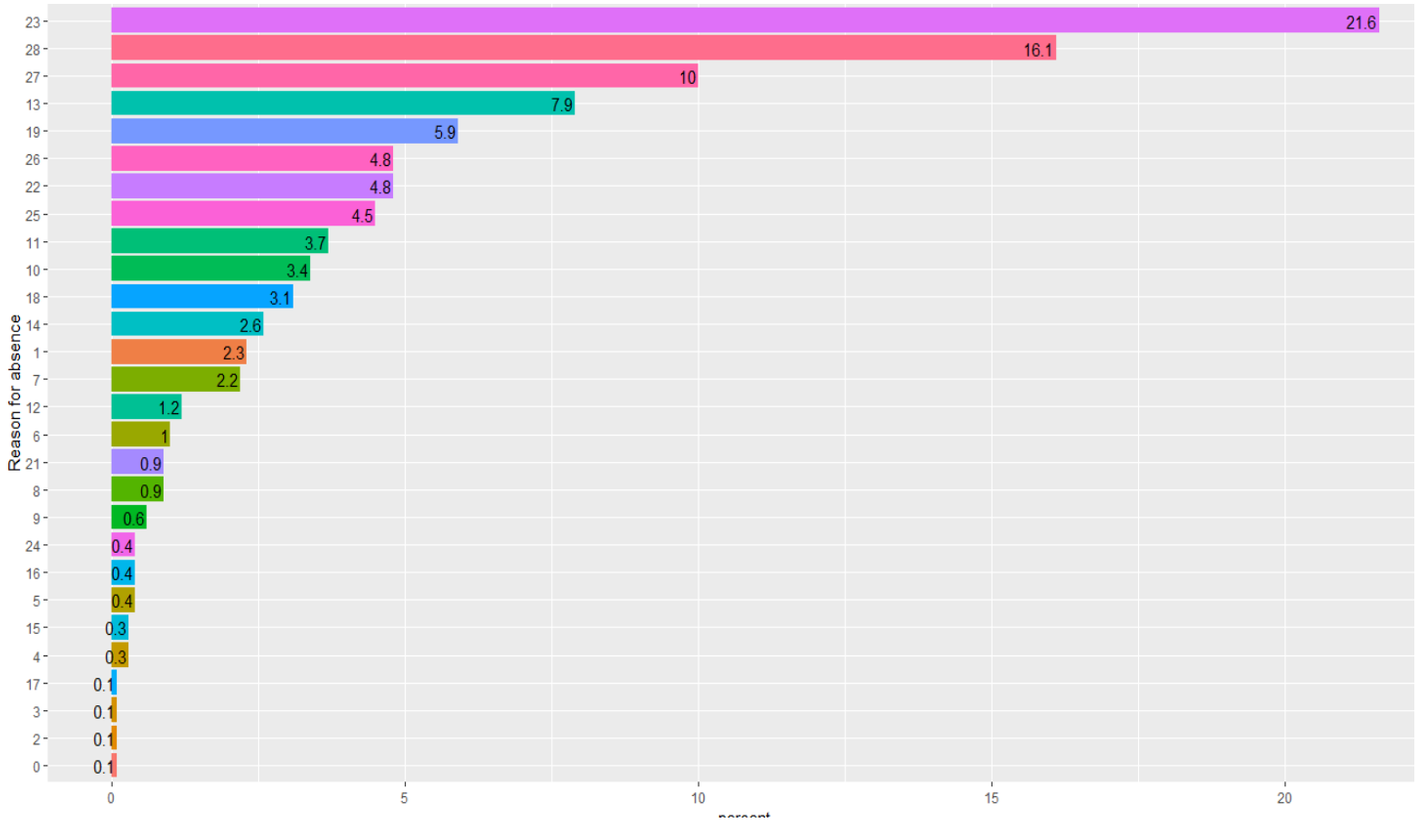


Fig: 2.2b(i) Percentage of the reason of absence used by employees

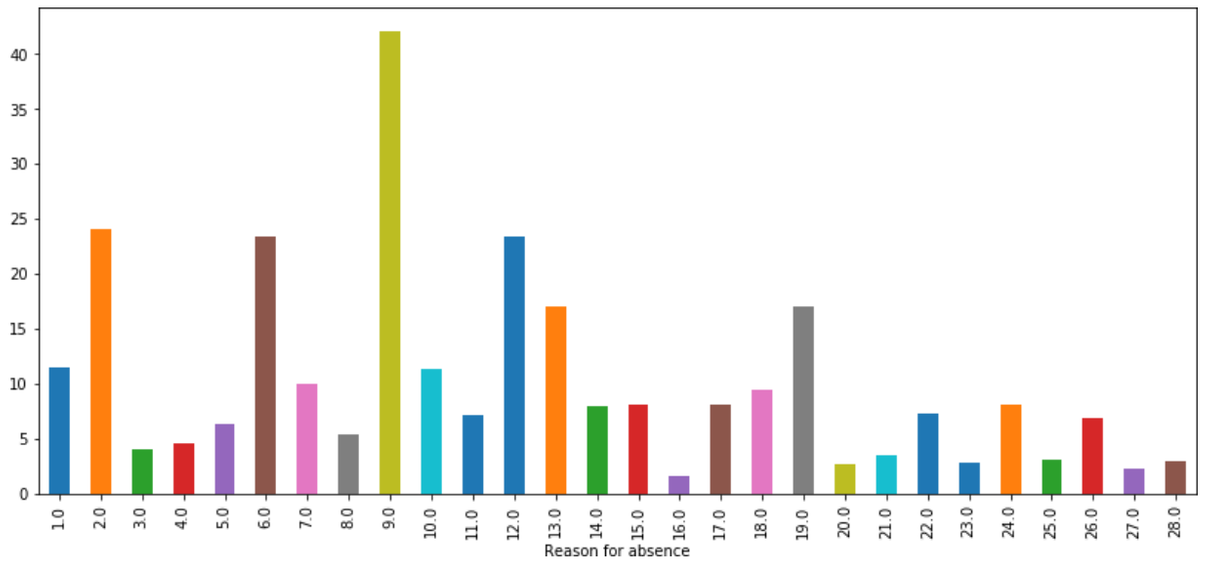


Fig: 2.2b(ii) Mean absenteeism hours for each reason of absence

Fig 2.2a(i) shows the distribution of reason of absence in percentages used by employees over the whole period. It shows that reason 23(medical consultation) is the most prominent reason amongst all, as it is the most evident reason and the most generic one, apart from this 4.8 percent of employees haven't stated their reason of absenteeism which is reason 26(unjustified absence). The least reason stated by employees for being absent are 3 (Diseases of the blood and blood- forming organs and certain disorders involving the immune mechanism), 2(Neoplasms) etc. these reasons are also anyhow very uncommon to see in our daily life. Note that, there is 0.1 percent of reason of absence number 0, which should not have been there was removed later, thus by careful scrutinization of each and every variable, all the rows of all variables were imputed which had absurd values.

Fig 2.2b(ii) shows the mean of number of hours that employees have been absent for particular reason. It shows that reason 9(Diseases of the circulatory system) has the most longing effect on the number of hours of absenteeism on employees, and the least ones were 16 (Certain conditions originating in the perinatal period), 21(External causes of morbidity and mortality) etc.

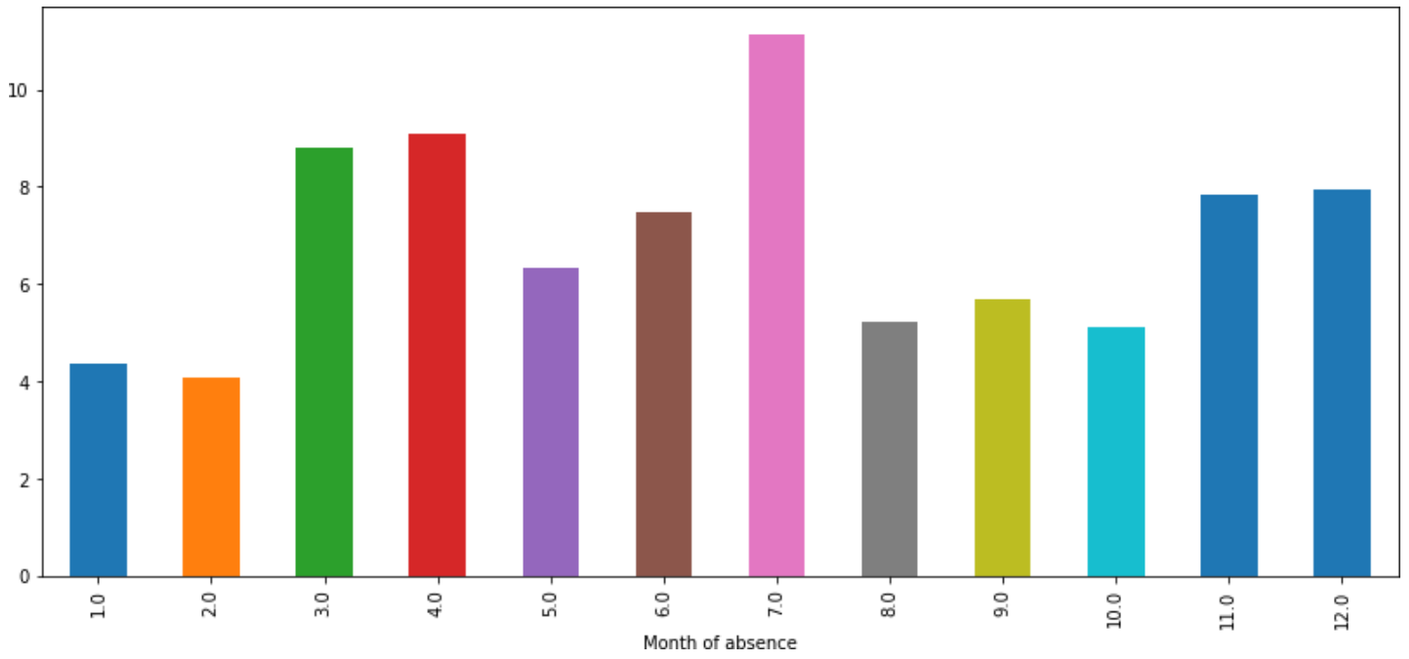


Fig: 2.2c Mean absenteeism hours for each Month of absence

Fig 2.2c shows the mean of absenteeism hours spread over months. It suggests that most number of skipping of hours is seen in the month of July followed by march and so on, and the least being in January.

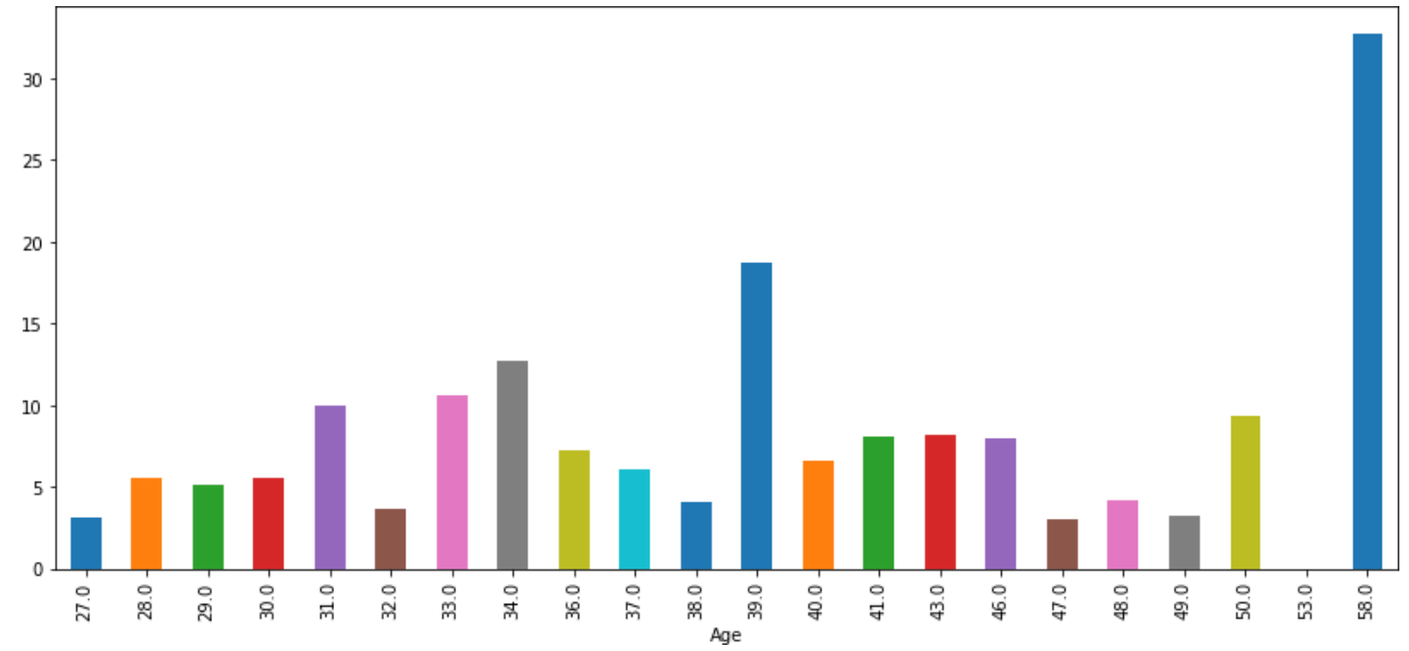
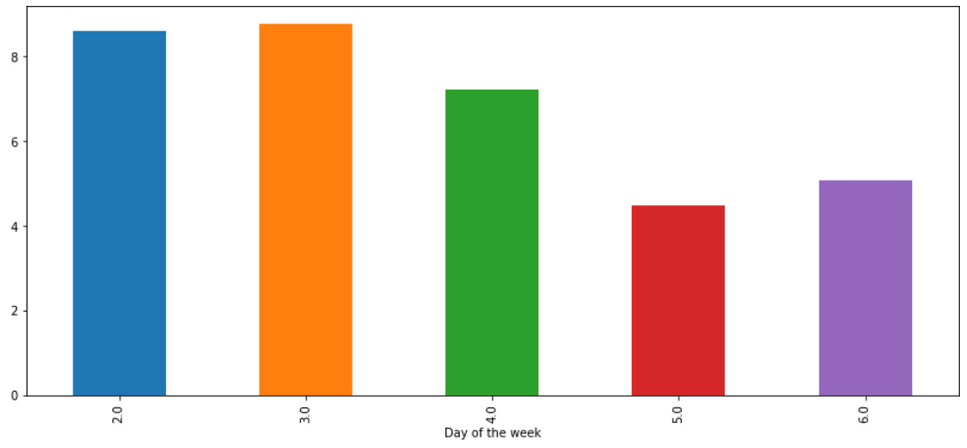


Fig: 2.2d Mean absenteeism hours for each Age Value

Fig 2.2d shows the mean of absenteeism hours spread over different Aged people. It suggests that most number of skipping of hours is seen in elderly people, while the youngest lot seems good in remaining present.



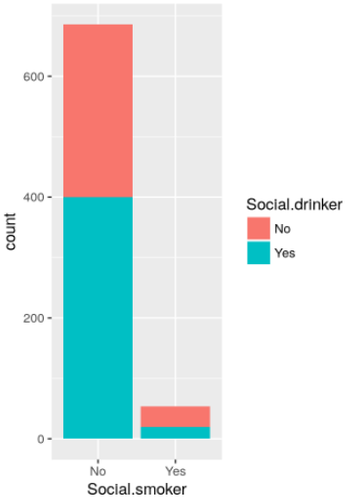
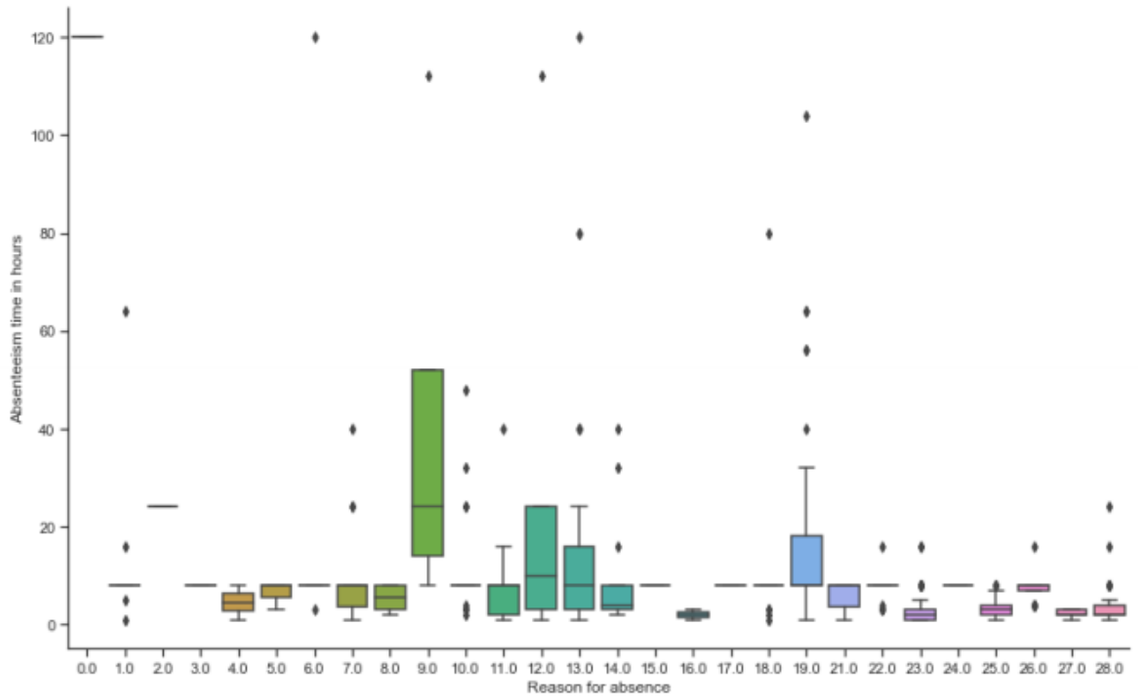


Fig: 2.2e Mean absenteeism hours for each reason of absence Fig: 2.2f shows the count of instances s of smoking and drinking

Fig 2.2e shows that Monday and Tuesday had the most amount of absenteeism hours, while Thursday, middle day of the week, had the least hours of absenteeism.

Fig 2.2f shows that around 50% of people are social drinkers, while very less percentage of people ~10% are social smoke

1. Reason for absence
   Pattern between 'Reason for absence' and 'Absenteeism time in hours' may be used to impute missing values in 'Reason for absence'.
   Box plot for 'Reason for absence' and 'Absenteeism time in hours':



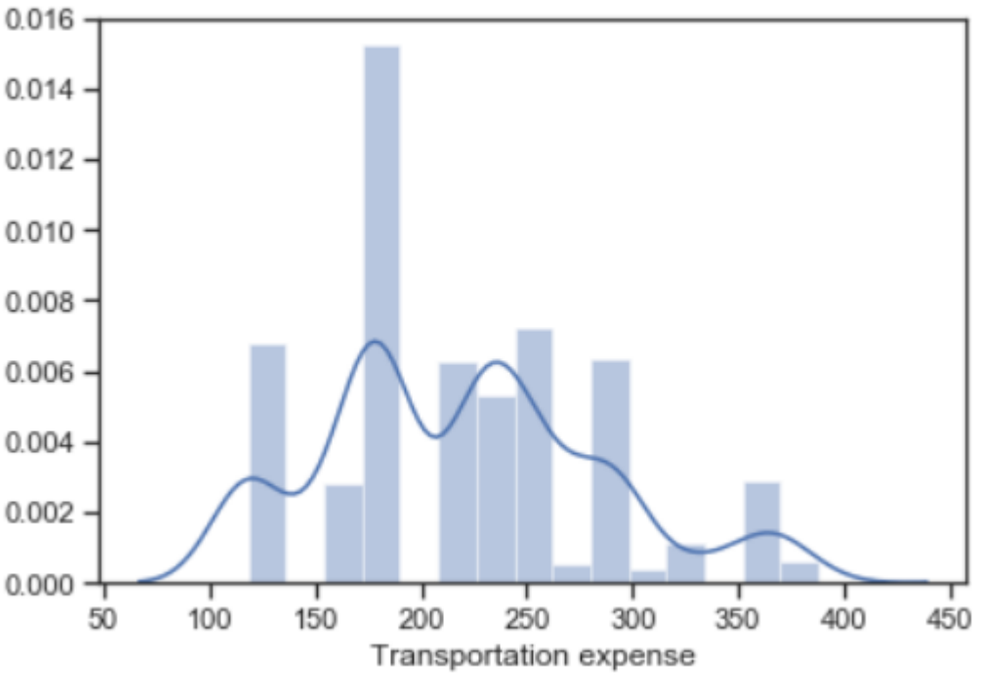
Category 27 of 'Reason for absence' is taking < 10 hrs of 'Absenteeism time in hours'. So, null values of 'Reason for absence' are put equal to 27 since 'Absenteeism time in hours' for the observations having null values is < 10 hrs. Zero category of 'Reason for absence' column has been put equal to category 26(i.e. unjustified absence).

2. Month of absence - Putting 'Month of absence' null value equal to 10.
3. Transportation expense 'Transportation expense' depends on 'Distance from Residence to Work' so we will use 'Distance from Residence to Work' values to impute missing values in 'Transportation expense'.
   If 'Distance from Residence to Work' value is 51, then 'Transportation expense' is equal to 179. If 'Distance from Residence to Work' value is 50, then 'Transportation expense' is equal to 260. If 'Distance from Residence to Work' value is 52, then 'Transportation expense' is equal to 361. If 'Distance from Residence to Work' value is 11, then 'Transportation expense' is equal to 235. If 'Distance from Residence to Work' value is 31, then 'Transportation expense' is equal to 291.
4. Distance from Residence to Work 'ID' column has been used to impute missing value for 'Distance from Residence to Work'.
5. Service time 'ID' column has been used to impute missing value for 'Service time'.
6. Age 'ID' column has been used to impute missing value for 'Age'.
7. Work Load Average/day 'Work load Average/day' values are dependent upon 'Month of absence' and 'Hit target' values.
8. Hit target 'Hit target' values are dependent upon 'Month of absence' and 'Work load Average/day' values.
9. Disciplinary failure 'Disciplinary failure' missing values have been put to 0.
10. Education 'ID' column has been used to impute missing value for 'Education'.

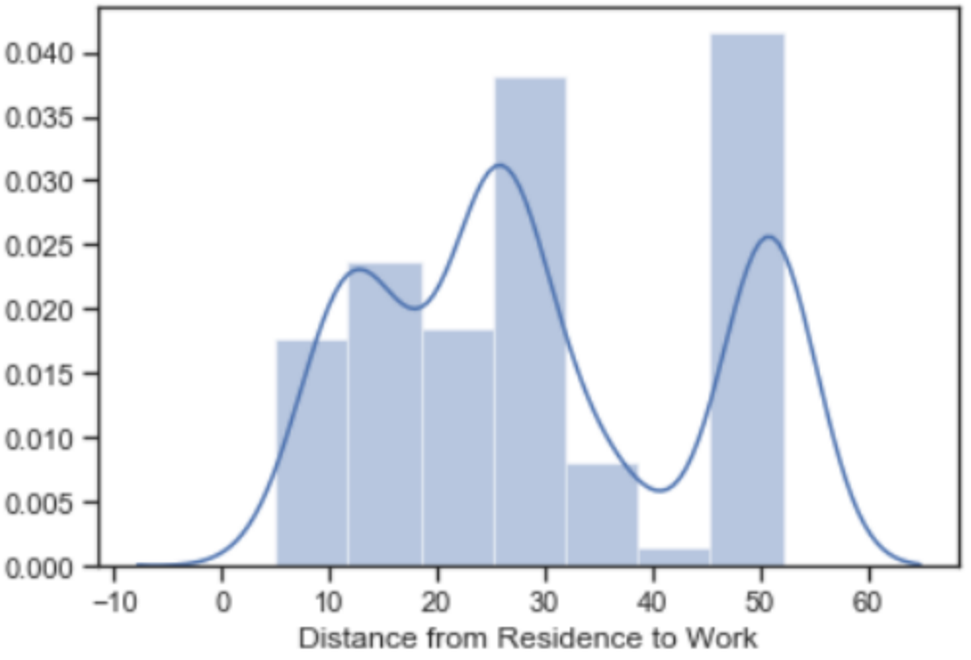
11. Son 'ID' column has been used to impute missing value for ‘Son’.
12. Social drinker 'ID' column has been used to impute missing value for ‘Social drinker’.
13. Social smoker 'ID' column has been used to impute missing value for ‘Social smoker’.
14. Pet 'ID' column has been used to impute missing value for ‘Pet’.
15. Weight 'ID' column has been used to impute missing value for ‘Weight’.
16. Height 'ID' column has been used to impute missing value for ‘Height’.
17. Body mass index 'ID' column has been used to impute missing value for ‘Body mass index’.
18. Absenteeism time in hours 'Reason for absence' column has been used to impute missing value for ‘Absenteeism time in hours’. All variables missing values have been imputed.

Distributions

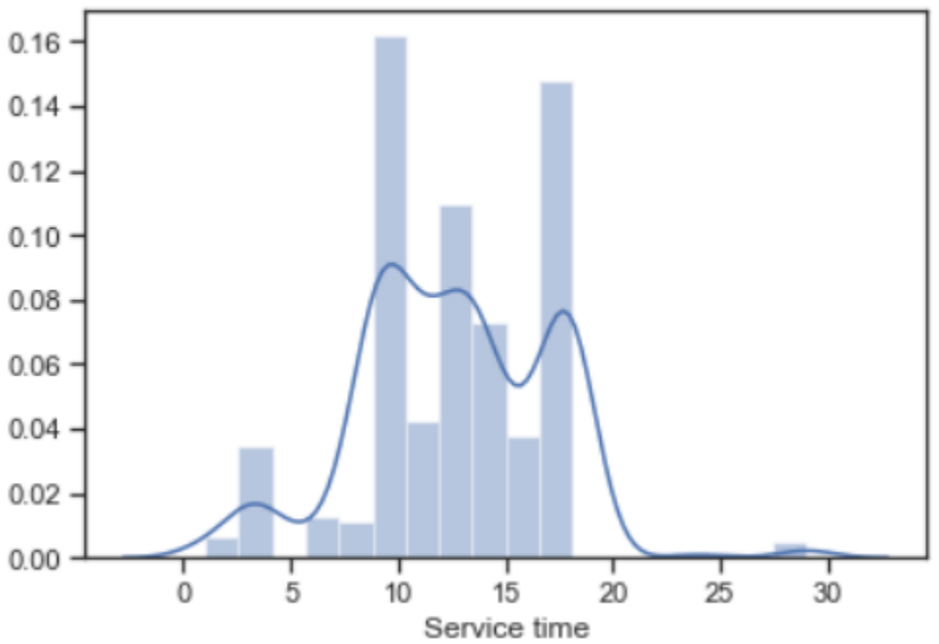
1. Transportation expense



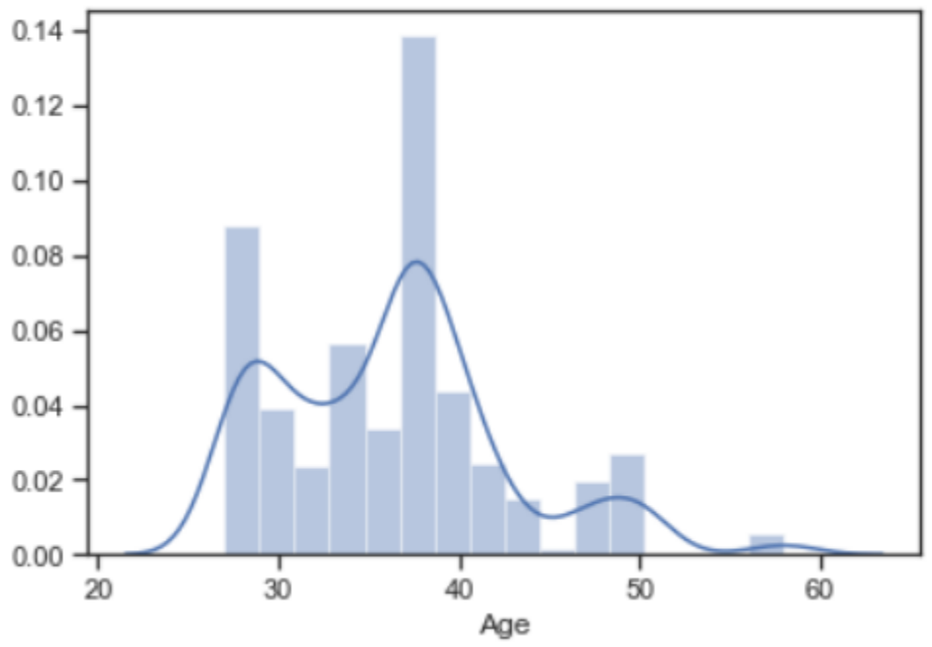
2. Distance from Residence to Work



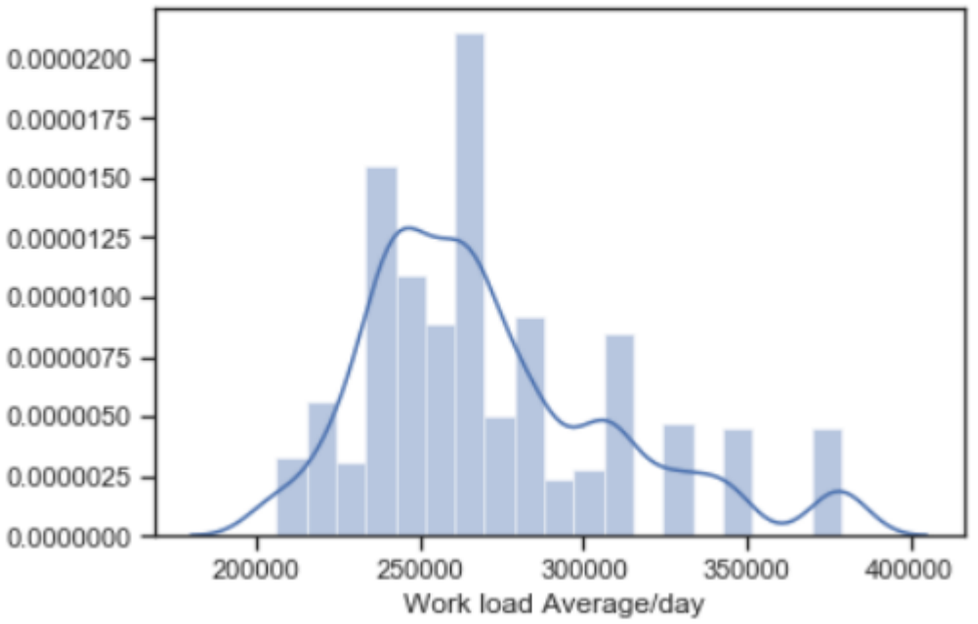
3. Service time



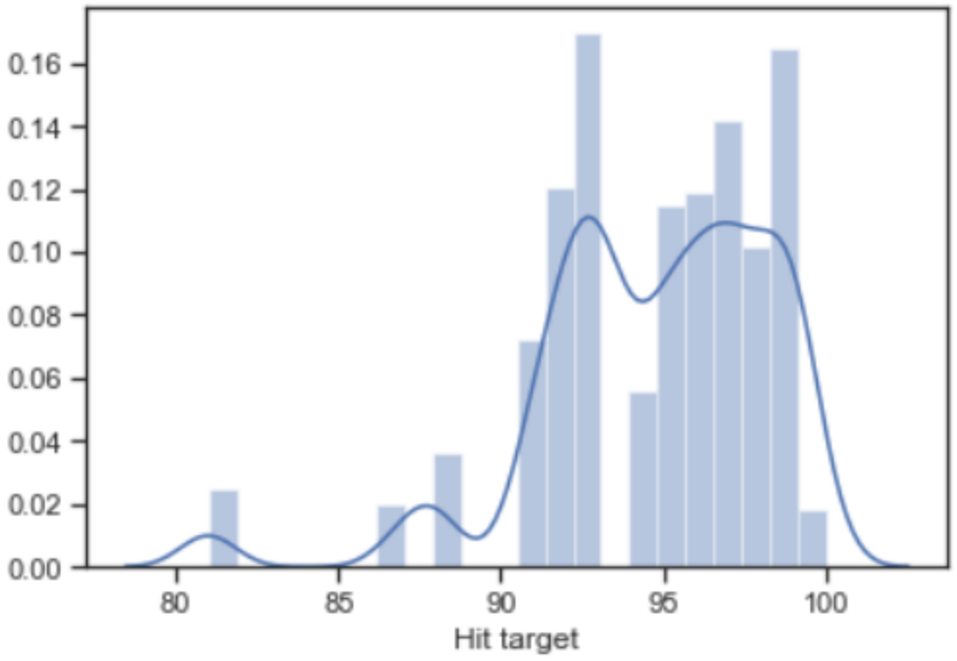
4. Age



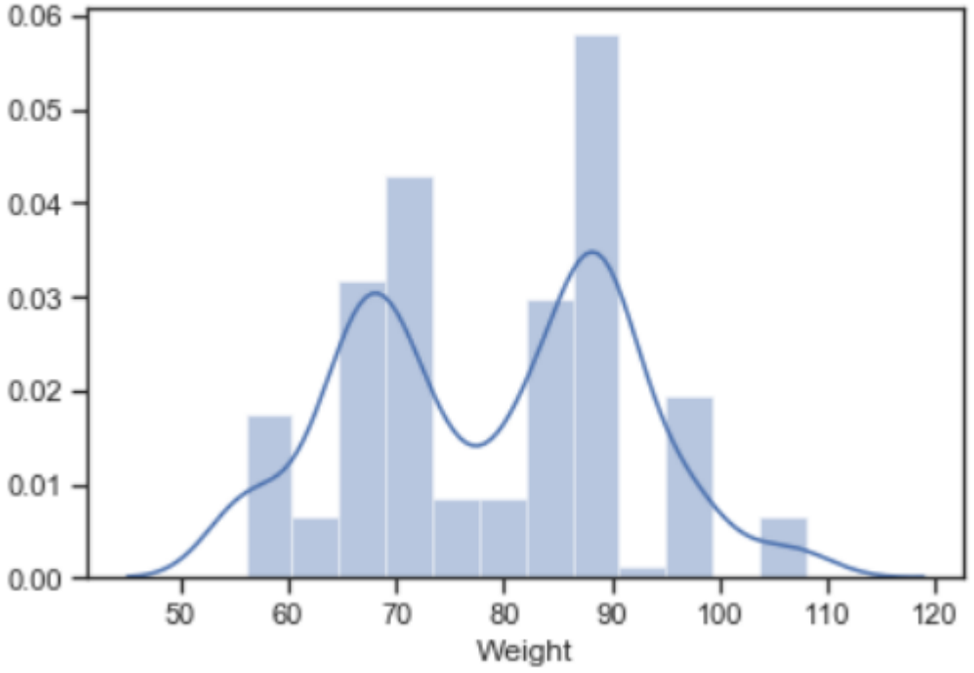
5. Work load Average/day



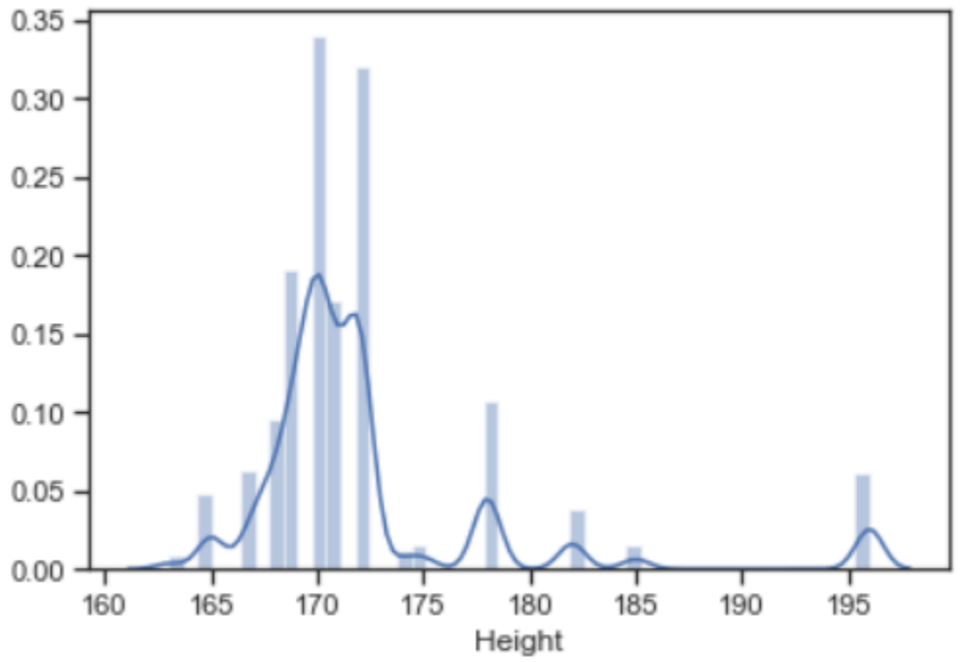
6. Hit target



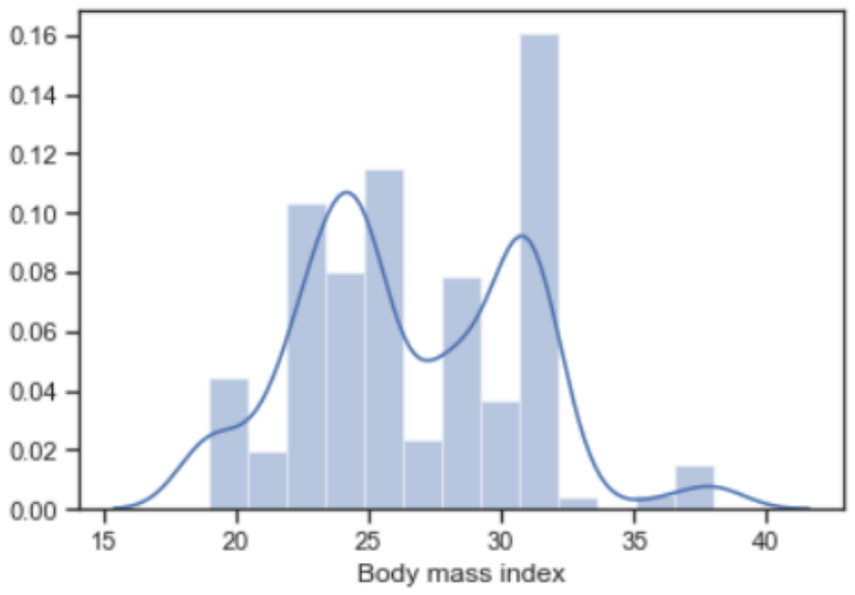
7. Weight



8. Height



9. Body mass index



Chi-square test was done for correlation between categorical variables:

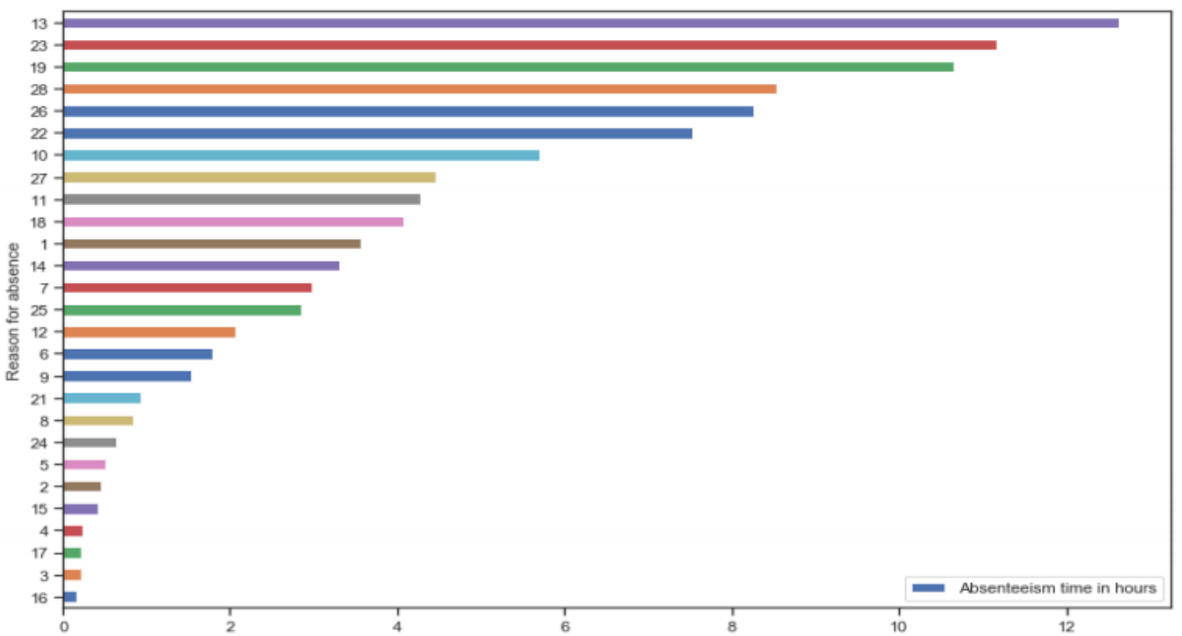
	Reason for absence	Month of absence	Day of the week	Seasons	Disciplinary failure	Education	Son	Social drinker	Social smoker	Pet
Reason for absence	0.000000e+00	2.328978e-15	6.131313e-02	5.901286e-21	4.768228e-13	1.749968e-10	1.741446e-18	1.711278e-08	1.914852e-08	9.402040e-19
Month of absence	2.328978e-15	0.000000e+00	6.087582e-01	0.000000e+00	4.180979e-01	8.639891e-03	3.123398e-05	2.363939e-02	3.194987e-02	5.667973e-05
Day of the week	6.131313e-02	6.087582e-01	0.000000e+00	3.948001e-01	2.460672e-01	5.439849e-01	2.223057e-09	3.575265e-01	8.227146e-01	4.046197e-01
Seasons	5.901286e-21	0.000000e+00	3.948001e-01	0.000000e+00	2.425953e-02	6.286186e-02	1.059010e-05	1.981972e-01	1.569992e-01	1.772464e-04
Disciplinary failure	4.768228e-13	4.180979e-01	2.460672e-01	2.425953e-02	0.000000e+00	9.094792e-01	4.545640e-01	6.721064e-01	2.660185e-03	7.200002e-01
Education	1.749968e-10	8.639891e-03	5.439849e-01	6.286186e-02	9.094792e-01	0.000000e+00	9.146983e-12	3.391862e-34	1.984123e-24	1.176649e-29
Son	1.741446e-18	3.123398e-05	2.223057e-09	1.059010e-05	4.545640e-01	9.146983e-12	0.000000e+00	3.706723e-09	4.132435e-21	2.225197e-88
Social drinker	1.711278e-08	2.363939e-02	3.575265e-01	1.981972e-01	6.721064e-01	3.391862e-34	3.706723e-09	0.000000e+00	1.515194e-02	1.196121e-26
Social smoker	1.914852e-08	3.194987e-02	8.227146e-01	1.569992e-01	2.660185e-03	1.984123e-24	4.132435e-21	1.515194e-02	0.000000e+00	5.706486e-14
Pet	9.402040e-19	5.667973e-05	4.046197e-01	1.772464e-04	7.200002e-01	1.176649e-29	2.225197e-88	1.196121e-26	5.706486e-14	0.000000e+00

Dropping Seasons since p-value of 'Seasons' versus 'Month of absence' is 0.00 (<0.05) rejecting null hypothesis that the two variables are independent.

No two variables have correlation coeff. > 0.95 so we will not drop any continuous independent variables.

Relationships of categorical independent variables with dependent variable

1. 'Reason for absence' Vs. 'Absenteeism time in hours'



Top 3 categories in order of Absenteeism time are:

A. Category 13 : Diseases of the musculoskeletal system and connective tissue - 12.62 % of total time

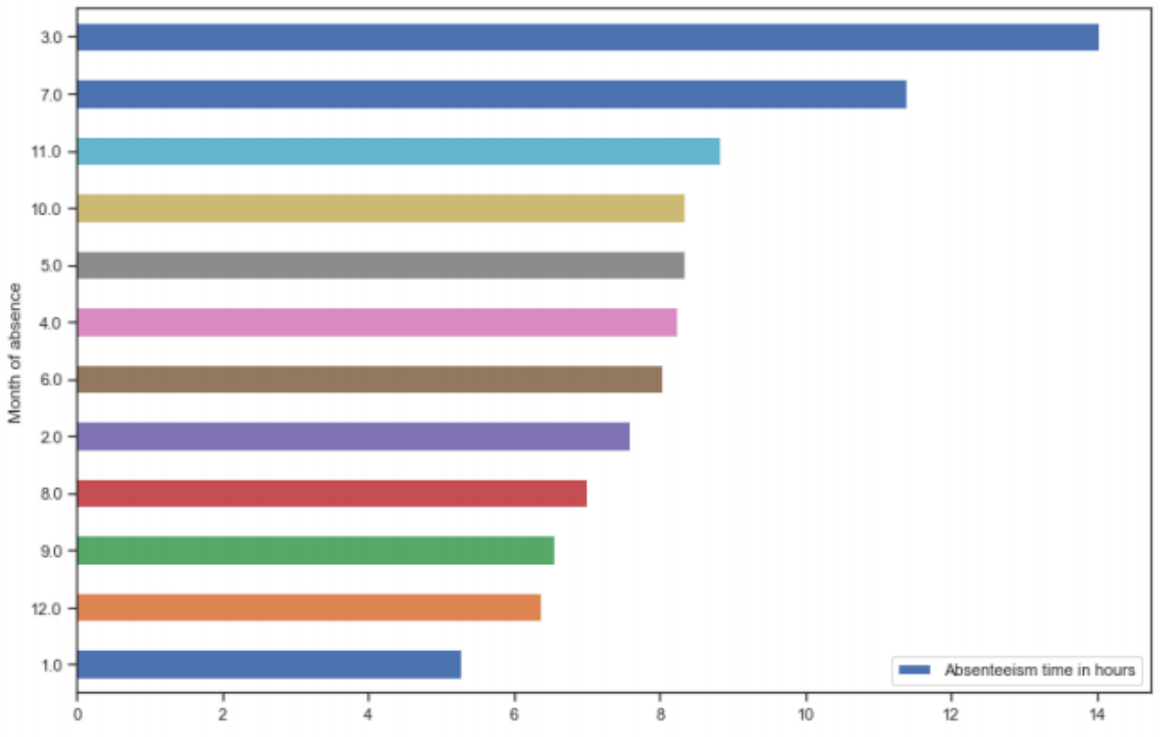
B. Category 23 : medical consultation - 11.17 % of total time

C. Category 19 : Injury, poisoning and certain other consequences of external causes - 10.64 % of total time

D. Category 28 : dental consultation - 8.53 % 0f total time

E. Category 26 : unjustified absence - 8.27 % of total time

2. 'Month of absence' Vs. 'Absenteeism time in hours'



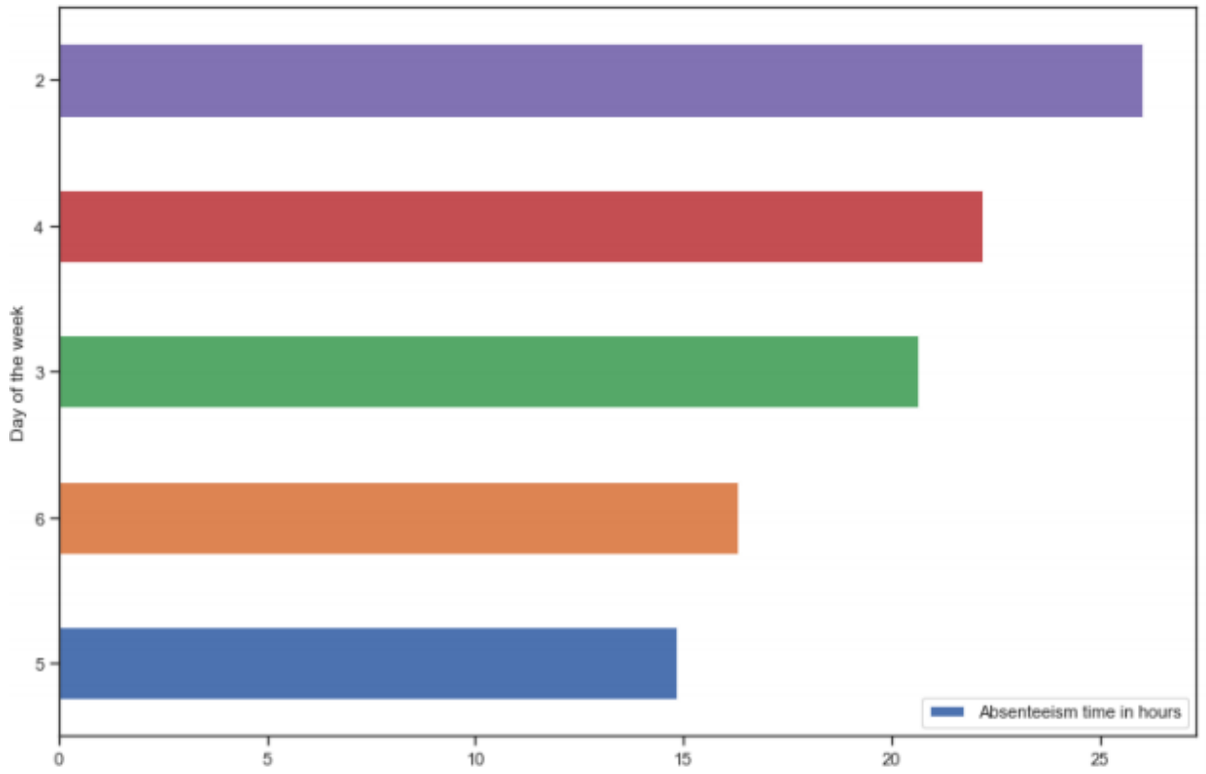
Top 3 months in order of Absenteeism time are:

A. Month 3 : March - 14.02 % of total time

B. Month 7 : July - 11.38 % of total time

C. Month 11 : November - 8.82 % of total time

3. 'Day of the week' Vs. 'Absenteeism time in hours'



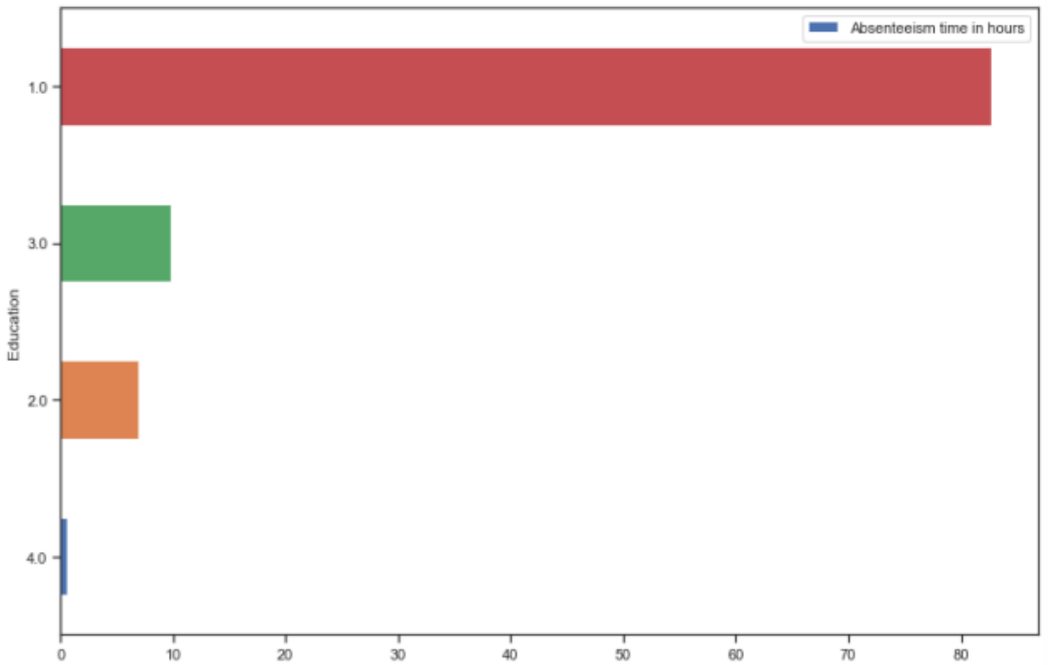
Top 3 days in order of Absenteeism time are:

A. Day 2 : Monday - 26.02 % of total time

B. Day 4 : Wednesday - 22.19 % of total time

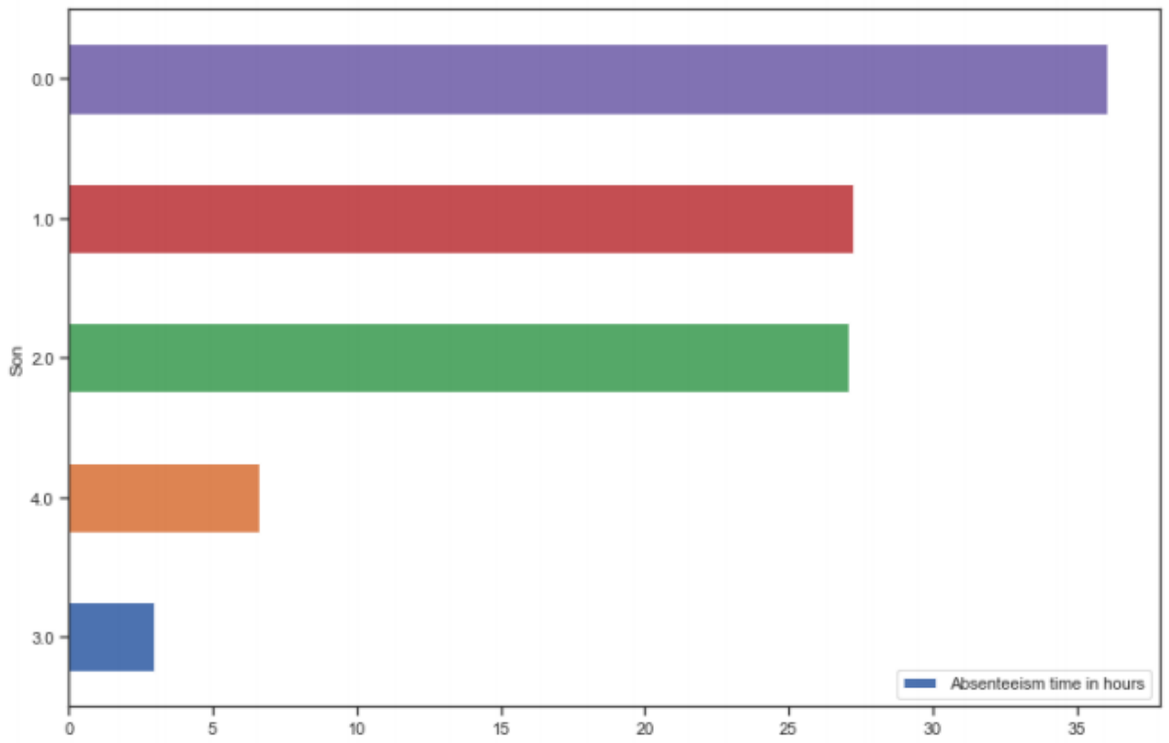
C. Day 3 : Tuesday - 20.63 % of total time

4. 'Education' Vs. 'Absenteeism time in hours'



82.69 % of absenteeism time is contributed by people having high school education. This may be due to majority of people having high school education. No conclusion may be drawn from this graph.

5. 'Son' Vs. 'Absenteeism time in hours'



Top 3 categories in order of Absenteeism time are:

A. Category 0 : No son - 36.06 % of total time

B. Category 1 : One son - 27.23 % of total time

C. Category 2 : Two sons - 27.08 % of total time People with no son are taking most of absenteeism time.

Feature Selection

For selecting the feature and understanding the inter-relationships between the variables, we performed three tests.

Correlation analysis:

Fig: 2.3a shows the correlation analysis table where we can deduce that Body Mass Index and Weight are highly correlated, also service time and age are somewhat positively correlated. The variable that gets dropped is Body Mass index, since it had the most number of missing values, thus helping us in reducing the pain of filling artificial data.

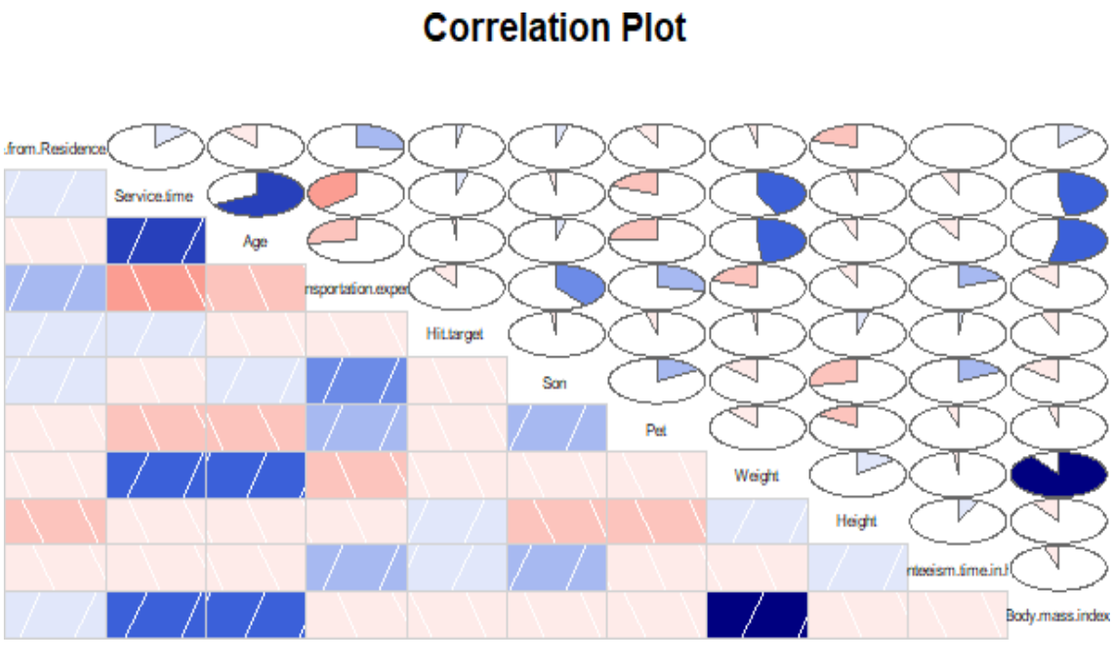


Fig: 2.3a Correlation analysis table heatmap

VIF:

VIF helps in quantifying the extent of correlation between one predictor and the other predictors in a model. It helps not only in determining collinearity with one variable, but helps in making out correlation of one variable with more than one variable.

Distance from Residence to Work	1.681130
Service time	3.371883
Age	2.424660
Work load Average/day	1.049937
Transportation expense	1.597151
Hit target	1.043096
Son	1.255561
Pet	1.580640
Weight	157.814366
Height	28.786233
Body mass index	147.832865
Absenteeism time in hours	1.048464

Fig 2.3b(i) VIF before variable selection

Distance from Residence to Work	1.600629
Service time	3.240114
Age	2.306970
Work load Average/day	1.048436
Transportation expense	1.591382
Hit target	1.042957
Son	1.251408
Pet	1.509672
Weight	1.645911
Height	1.484519
Absenteeism time in hours	1.046549

Fig 2.3b(ii) VIF after variable selection

Fig 2.3b(i) shows the VIF values of different variables before removing Body Mass Index, notice that Weight, Height and BMI are above 10, suggesting that these three have some connection between them which cannot be seen by correlation plot. Also, BMI and weight have the most prominent connection, which can also be concluded by r value of correlation analysis between BMI and Weight. In Fig 2.3b(ii), it can be seen that when BMI is removed all the inflated variation settles down.

Using Extra-tree Regressor:

The Extra-Tree method (extremely randomized trees method), its main objective is to further randomizing tree building in the context of input features. This method is quite similar to random forest, but unlike random forests, in Extra tree regressor splits are selected in random instead of using some criteria like in Random Forest.

Fig 2.3c shows that reason for absence, day of the week and work load are amongst the main features in deciding absenteeism, while variables like education and social smoker doesn't really helps in determining our target variable. However, we would include all the variables except BMI since the variables are less and are carrying at least some importance in determining our target variable

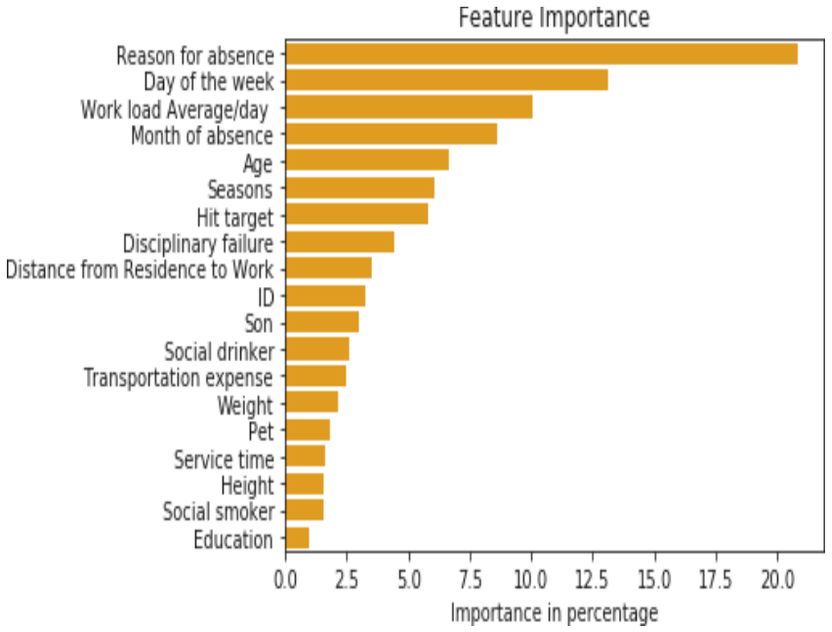


Fig 2.3c Importance of variable in percentage using Extra trees regresssor

Outlier Analysis

Outlier detection and treatment is always a tricky part especially when our dataset is small. The box plot method detects outlier if any value is present greater than (Q3 + (1.5 * IQR)) or less than (Q1 – (1.5 * IQR)).

Fig 2.4a shows the boxplot of all numerical variables, showcasing us the distribution of the data across.

Fig 2.4b shows the number of outliers for each variable. Now, by seeing all the variables individually and checking the nature of outliers, it was observed that most of the outliers like height, number of pets, age etc. were feasible data, so only imputed absenteeism in hours variable by KNN and by MICE method in RStudio

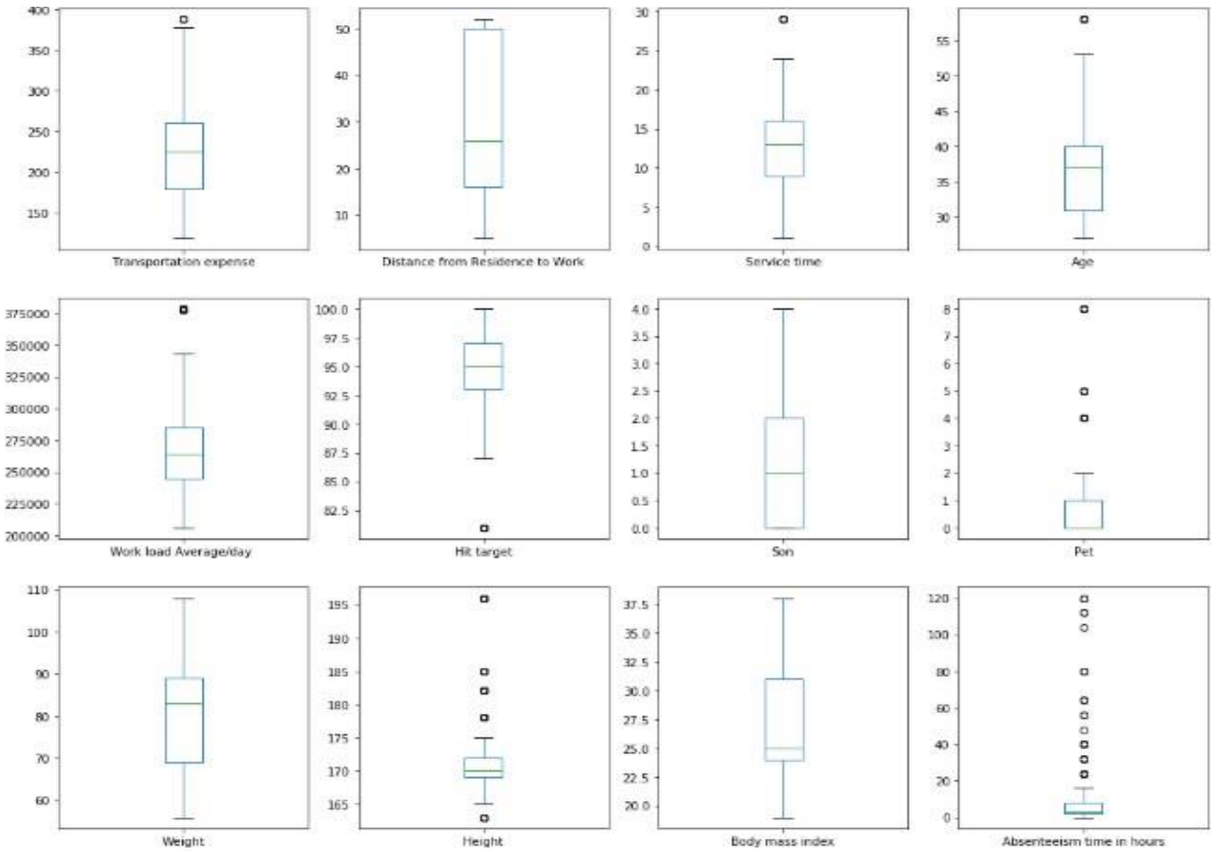


Fig: 2.4a Box Plot of all numerical variables

:

	0
ID	0
Reason for absence	0
Month of absence	0
Day of the week	0
Seasons	0
Transportation expense	3
Distance from Residence to Work	0
Service time	5
Age	8
Work load Average/day	29
Hit target	19
Disciplinary failure	0
Education	0
Son	0
Social drinker	0
Social smoker	0
Pet	42
Weight	0
Height	114
Absenteeism time in hours	43

Fig: 2.4b Outlier data information for each variable

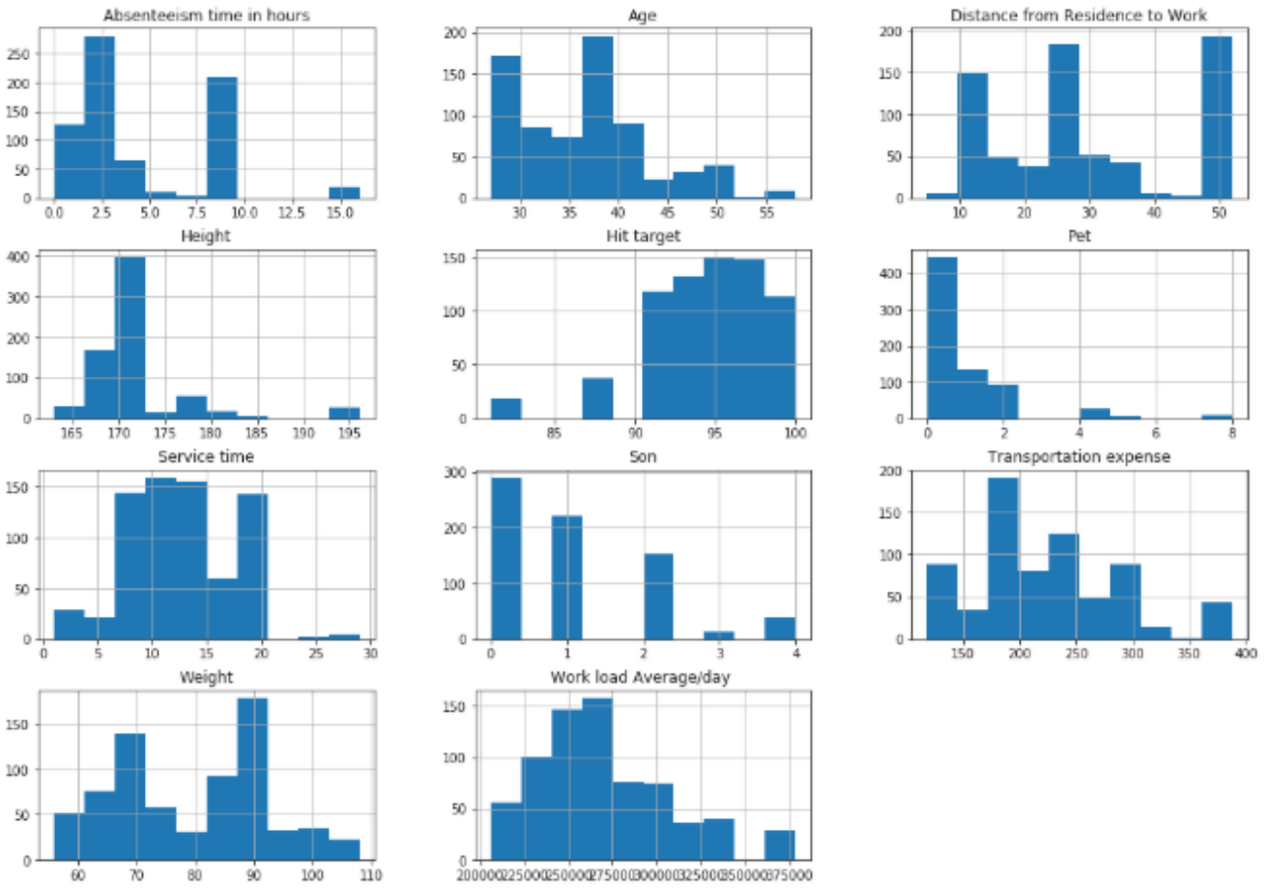


Fig: 2.4d Histograms of all numerical variables after outlier analysis

Fig. 2.4d shows the histograms plot of all numerical features suggesting that none of the variable is following normal distribution, and also by transformation the curve shape didn't change, thus going by original values only and no transformation.

Feature Scaling:

It was observed that many numerical variables had different ranges of value, some huge and some small, so just to stay assured that large values don't dominate our model, fature scaling was performed and all variables' values were brought in between 0 to 1 by using the below formula.

$$Value_{new} = \frac{Value - minValue}{maxValue - minValue}$$

Model Development

All the models were developed and tested to see their performance on data. The performance metric used was RMSE, as it is the most widely used performance metric, and helps in getting or visualizing deviations easily. Although, MAPE is also a good metric and it provides good a visual enticement to the client as it is very easy to read and work upon.

In all the models, Result-train RMSE means RMSE value obtained when the model is applied to the data from which it was created. Result-test RMSE means RMSE value obtained when the model is applied to different dataset(20%) of the bigdataset.

Linear Regression:

In Fig 3.1a, the summary of linear regression model is show. The adjusted R-square value(~0.79) is pretty good, and most of our data is explaining the target variable. All other parameters also seem to be pretty good from the OLS Summary table.

==========

==========

Linear Regression Result-
test

RMSE: 3.480437989373406

Result-train

RMSE: 2.3221722321368317

==========

==========

The train and test RMSE are 2.32 and 3.48 respectively, which are fine, and the model is overall acceptable with good parameters' value.

OLS Regression Results

Dep. Variable:	Absenteeism time in hours	**R-squared:**	0.807
Model:	OLS	**Adj. R-squared:**	0.784
Method:	Least Squares	**F-statistic:**	35.06
Date:	Fri, 28 Sep 2018	**Prob (F-statistic):**	2.46e-146
Time:	20:02:15	**Log-Likelihood:**	-1302.1
No. Observations:	574	**AIC:**	2726.
Df Residuals:	513	**BIC:**	2992.
Df Model:	61		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Transportation expense	1.3097	0.621	2.108	0.036	0.089	2.531
Distance from Residence to Work	0.0701	0.562	0.125	0.901	-1.034	1.174
Service time	0.1185	1.296	0.091	0.927	-2.427	2.664
Age	-1.3471	0.797	-1.690	0.092	-2.913	0.219
Work load Average/day	1.6122	0.573	2.815	0.005	0.487	2.737
Hit target	1.8350	0.839	2.188	0.029	0.188	3.482
Son	0.8702	0.491	1.773	0.077	-0.094	1.834
Pet	-1.3783	0.899	-1.533	0.126	-3.145	0.388

Weight	1.1267	0.665	1.693	0.091	-0.181	2.434
Height	-1.8893	0.859	-2.199	0.028	-3.577	-0.202
Reason for absence_2.0	-0.7065	2.666	-0.265	0.791	-5.945	4.532
Reason for absence_3.0	2.1236	2.626	0.809	0.419	-3.035	7.283
…..						
Education_4.0	1.1559	1.351	0.855	0.393	-1.499	3.811
Social drinker_1.0	0.4152	0.364	1.140	0.255	-0.300	1.131
Social smoker_1.0	0.2872	0.554	0.518	0.604	-0.802	1.376

Omnibus:	149.241	**Durbin-Watson:**	1.916
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	594.890
Skew:	1.135	**Prob(JB):**	6.62e-130
Kurtosis:	7.441	**Cond. No.**	62.0

Warnings:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Fig 3.1a Summary of linear regression model

Ridge and Lasso Regression:

Ridge or lasso are forms of regularized linear regressions. Lasso calculates its value by reducing the power of unnecessary variables or least important variables, while ridge on the other hand tends to bring all variables' contribution in building models. Both of these are generally used when we want to use a more complex or when we have to overcome overfitting issues from linear model. Anyways, since we did not have much issue with all the above stated problems in linear model, it is ok to pick linear model as a better choice compared to these two. Nonetheless, ridge has definitely performed well here and is almost equivalent to linear model in terms of performance.

==========

==========

Ridge Result-test

RMSE: 3.4796849319660326

Result-train

RMSE: 2.334055816414068

==========

==========

Lasso Result-test

RMSE: 4.058357539349357

Result-train

RMSE: 3.232807836883631

==========

==========

SVR:

Support Vector Regressor(Gausian Kernel) has a test value of ~4, hence the model is not apt for building in our dataset.

==========

==========

SVR

Result-test

RMSE: 4.040058113370883

Result-train

RMSE: 2.993197635737354

==========

==========

KNN:

The test value RMSE for this model is ~3.56, which is fine as compared to other models, so to c heck the variation of result by varying the number of neighbors, a graph is plotted between RMS E value and number of neighbors in Fig.3.4b. According to figure, k = 5(which luckily is also th e default value) seems the best, with RMSE of ~3.56

==========

==========

KNeighborsRegressor Result-test

RMSE: 3.566121453655529

Result-train

RMSE: 2.4403261334564816

==========

==========

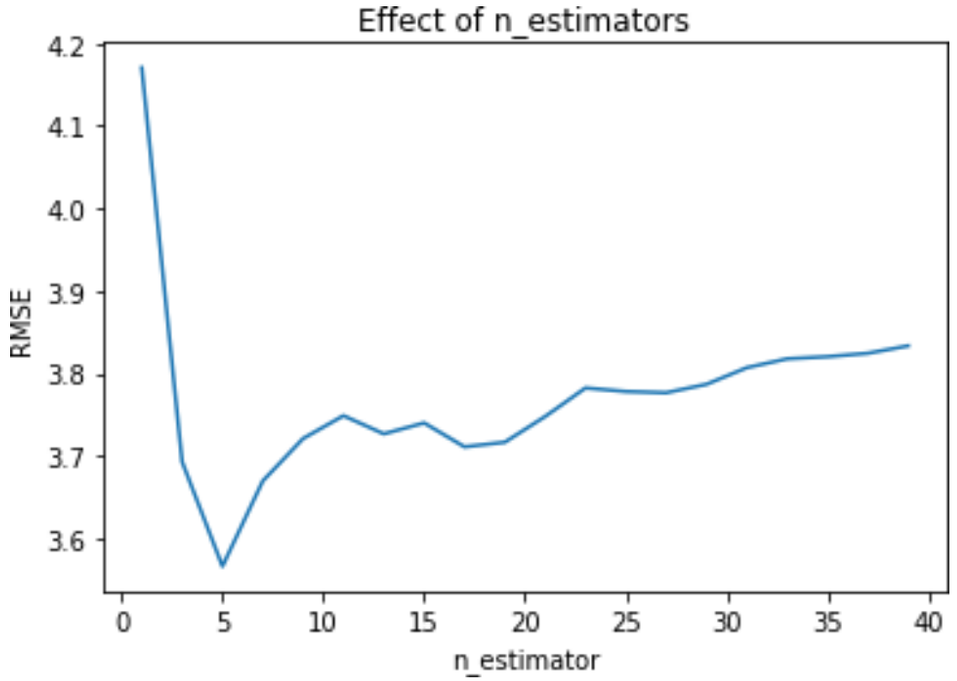


Fig 3.4b Number of nearest neighbors in KNN v/s RMSE values

Decision Tree Regressor:

The values of train RMSE is close to 0 which suggests that there is over-fitting problem in this model, thus a regularized model of Decision tree is needed to treat this.

==========

==========

DecisionTreeRegressor Result-
test

RMSE: 4.346134936801766

Result-train

RMSE: 0.26807482443947966

==========

==========

Random Forest Regressor:

The solution to the previous model of decision tree gets exactly served by the regularized method, Random Forest. Notice that regularized method helped when there was over-fitting in traditional method, like in the earlier where linear model did not have over-fitting problem,, regularized ridge and lasso didn't come much of handy in improving score. But, here in random forest the score has increased considerably from 4.36 to 3.45 with train value also getting better than decision tree's model.

==========

==========

RandomForestRegressor
Result-test

RMSE: 3.454816155486407

Result-train

RMSE: 1.1326756427399138

==========

==========

The Fig.3.4c depicts the RMSE values with number of trees, which is the least in 20-25 and 160- 180 range with RMSE of around 3.36, this makes the model of choice to be higher.

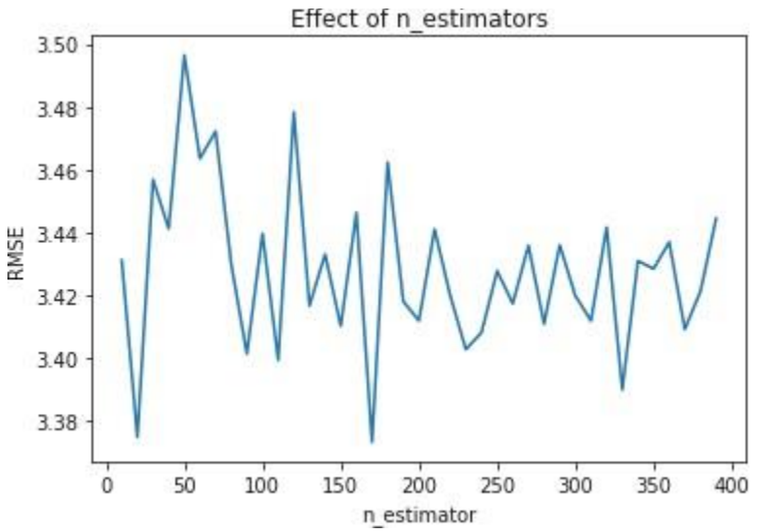


Fig 3.4c Number of trees in Random Forest v/s RMSE values

ADA Boost Regressor and Gradient Boost Regressor:

Both ADA boost and Gradient boost are boosting algorithms which means that they convert a set of weak learners into a single strong learner. They both initialize a strong learner (usually a decision tree) and iteratively create a weak learner that is added to the strong learner. They just differ on how they create the weak learners during the iterative process.

==========

==========

AdaBoostRegressor
Result-test

RMSE: 3.7049464047044403

Result-train

RMSE: 2.7963536152480746

==========

==========

GradientBoostingRegressor Result-
test

RMSE: 3.4747885667678045

Result-train

RMSE: 1.9479853430273673

==========

==========

XGBoost Regressor:

XGBoost uses a few computational tricks to speed up gradient descent and line search components, as well as a penalty function.

==========

==========

XGBRegressor
Result-test

RMSE: 3.4722148845178946

Result-train

RMSE: 2.0042449748723015

==========

==========

Hyperparameter tuning:

Here we will perform parameter tuning of XGBoost Regressor using GridSearchCV and will then see how our model performs on tuned parameters.

The value comes out to be:

{'gamma': 0, 'learning_rate': 0.01, 'max_depth': 3, 'n_estimators': 4

00, 'random_state': 1, 'subsample': 0.7}

After building the model on tuned parameters, it was seen that not significant improvement was there in the score, moreover it decreased suggesting that default parameters were better in model building. This might have occurred because the range of values input for GridSearchCV was not vast as it would have taken a long computational time, thus it missed on the best parameters.

Nevertheless, in actual condition, by increasing the range, this problem hardly occurs.

Final Inference and answers

Inference:

From all the above models, it was seen that Random Forest, Linear regression and all boosting m ethods performed well, so selection of any one of these would just be satisfactory, but the main f ocus lies in the preprocessing steps involved, the reason being, there were about two big imputat ions/changes in the dataset, one while dealing with missing values and another while dealing wit h outliers. Also, many ways of feature selection and binning of categories were available. Juggli ng with all these parameters and applying various permutations and combinations on the same, p lus the domain knowledge would have helped near the best model. Only after this, the tuning an d model building should have been done and overseen. Thus, there are several other ways and m ixed approaches we could employ till we can eliminate previous models and reach to an optial m odel.

The Fig 4.1a below shows and sums up the test MSE in a graphical plot of all the models used.

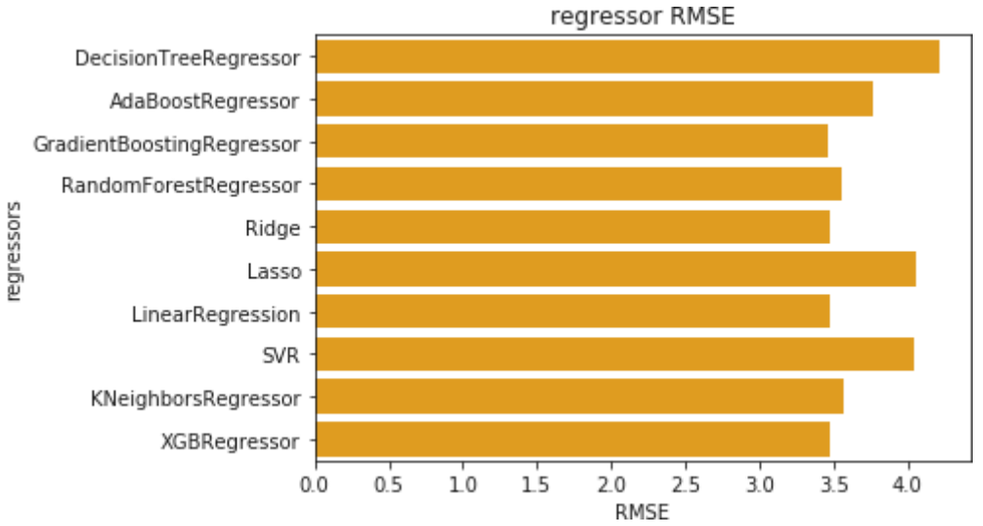


Fig. 4.1a Model v/s the test RMSE score

Answers:

Q1. What changes company should bring to reduce the number of absenteeism?

-The changes company can bring to reduce absenteeism hours is by cutting or improving in the areas where the absenteeism hours gets affected. For e.g. by observing graphs and plots in visualizations section, it was seen that oldest aged group of ~58 had tremendous shoot in absenteeism, thus employing younger batch can be helpful. The most number of instances(~21%) for the reason for absence was for medical consultation, thus having a half-yearly or quarterly checkup, or a doctor solely for company employees can also be appointed. It was also observed that work load was one of the important factors during the model building phase, so by bringing the work load to an optimized point which makes the employees avert absenteeism and the profits of the company also doesn't degrade much can be put in work. There was about 5% of instances where employees didn't justify the reason of absence, so stringent rules can be brought up against such employees. Like these, many other small changes can be brought up by the company, but broadly the above stated reasons would suffice. If absenteeism is a serious concern faced by the company, then it can definitely invest in the areas stated while compromising little on profits.

Q2. How much loss every month can we project in 2011 if same trend of absenteeism continues?

-If the same trend of absenteeism continues according to the dataset, then the losses incurred can be projected and be seen in Fig. 4.1b which shows that most of the losses were projected in the month of july, followed by march and april. The least was seen in the month of February.

The projection of losses was measured by grouping by means of service time per month, work load average/day, and number of absenteeism hours. The formula used was (absenteeism time in hours) / (Service time), which measured the part out of whole which were rendered absent by the employees, and this factor was multiplied by average work load(or units of work to be done in days), this gets us the final average losses per month.

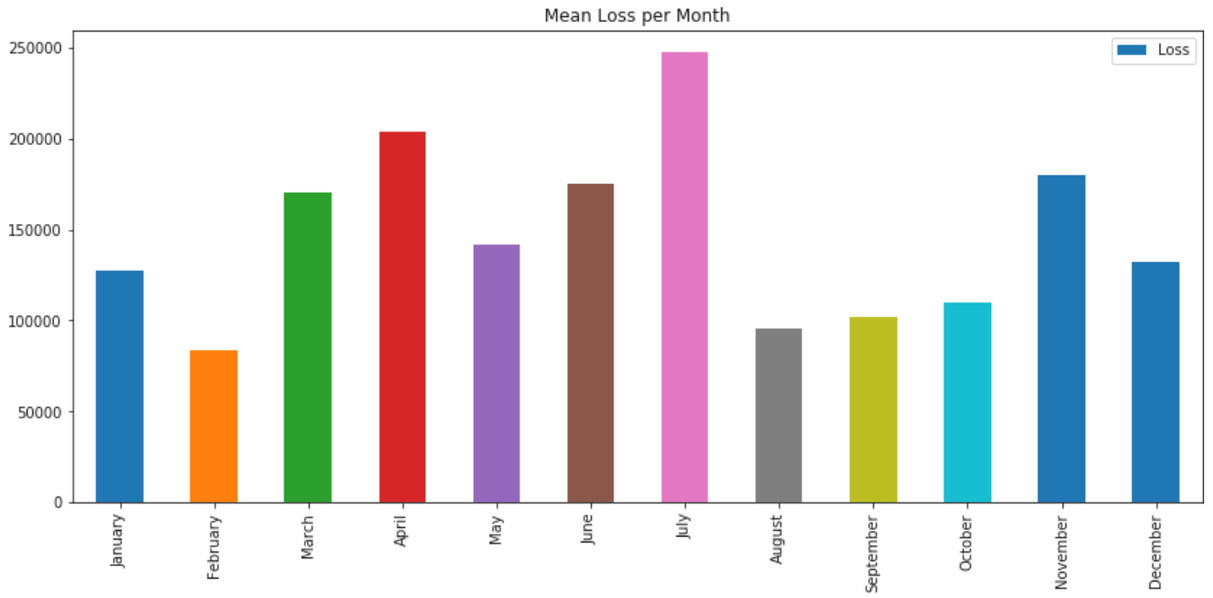


Fig. 4.2b Mean Loss per month due to absenteeism

Time Series - Arima Model

```
from statsmodels.tsa.stattools import adfuller
dftest = adfuller(ts,autolag='AIC')
dfoutput = pd.Series(dftest[0:4], index=['Test Statistic','p-value','#Lags Used','Number of Observations Used'])
for key,value in dftest[4].items():
   dfoutput['Critical Value (%s)'%key] = value
dfoutput
ts_log = np.log(ts)
plt.plot(ts_log)
ts_diff = ts_log - ts_log.shift()
ts_diff
plt.plot(ts_diff)
ts_diff.fillna(0,inplace=True)
dftest = adfuller(ts_diff)
dfoutput = pd.Series(dftest[0:4], index=['Test Statistic','p-value','#Lags Used','Number of Observations Used'])
for key,value in dftest[4].items():
```

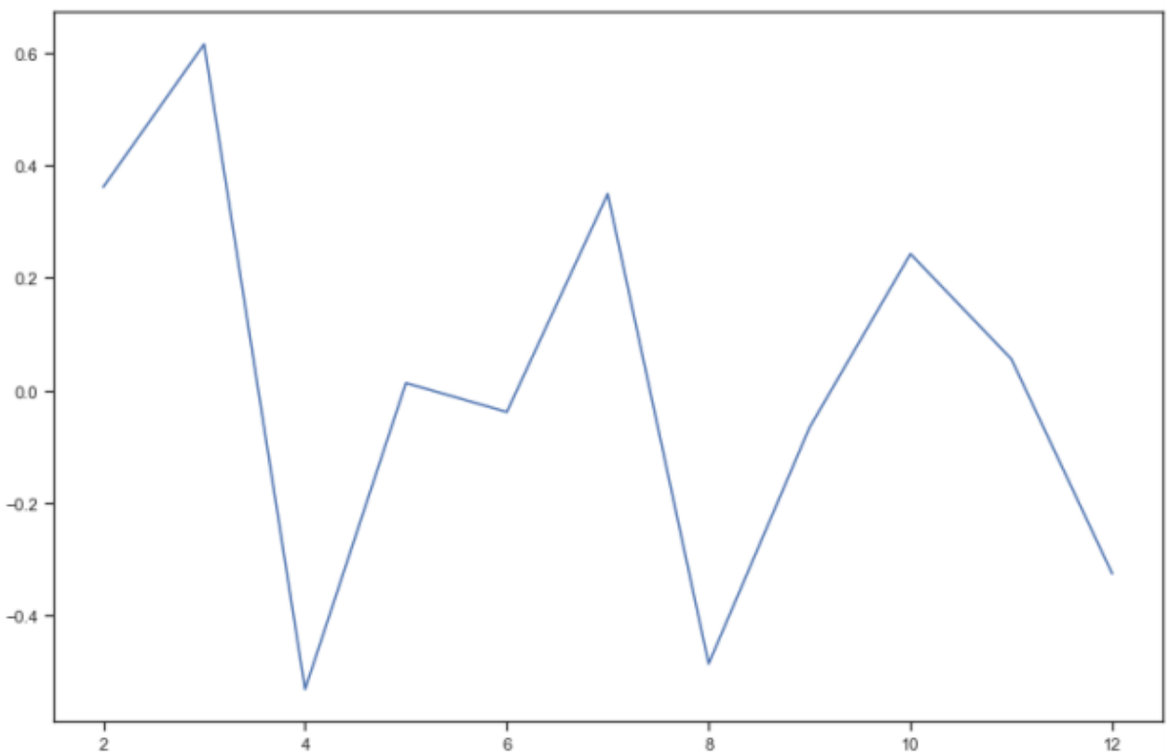
```
    dfoutput['Critical Value (%s)'%key] = value
dfoutput
```

First, Dickey Fuller test has been done to check if time series is stationary or not. Results were:

```
Test Statistic                  -0.000000
p-value                          0.958532
#Lags Used                       7.000000
Number of Observations Used      4.000000
Critical Value (1%)             -7.355441
Critical Value (5%)             -4.474365
Critical Value (10%)            -3.126933
dtype: float64
```

Since Test Statistic > Critical Values for 1%, 5% & 10%, time series is not stationary.

We will be first taking log of time series and then subtracting a shifted (single lag) log time series from log series. Now, time series plot is as under:



Dickey Fuller test was done again after taking log & differencing on log.

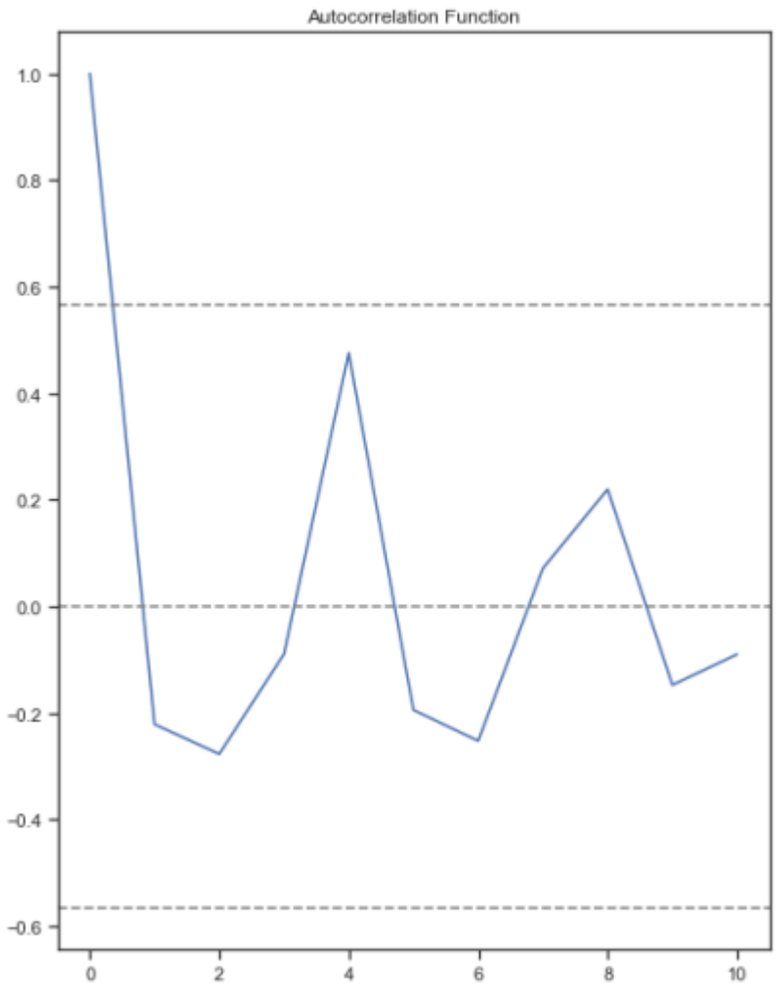
```
Test Statistic                 -3.069239
p-value                         0.028915
#Lags Used                      1.000000
Number of Observations Used    10.000000
Critical Value (1%)            -4.331573
Critical Value (5%)            -3.232950
Critical Value (10%)           -2.748700
dtype: float64
```

Since Test Statistic(-3.069239) < Critical Value for 10% (-2.748700), timeseries ts_diff is stationary.

Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) were plotted to find the values of p (AR order) and q (MA order).

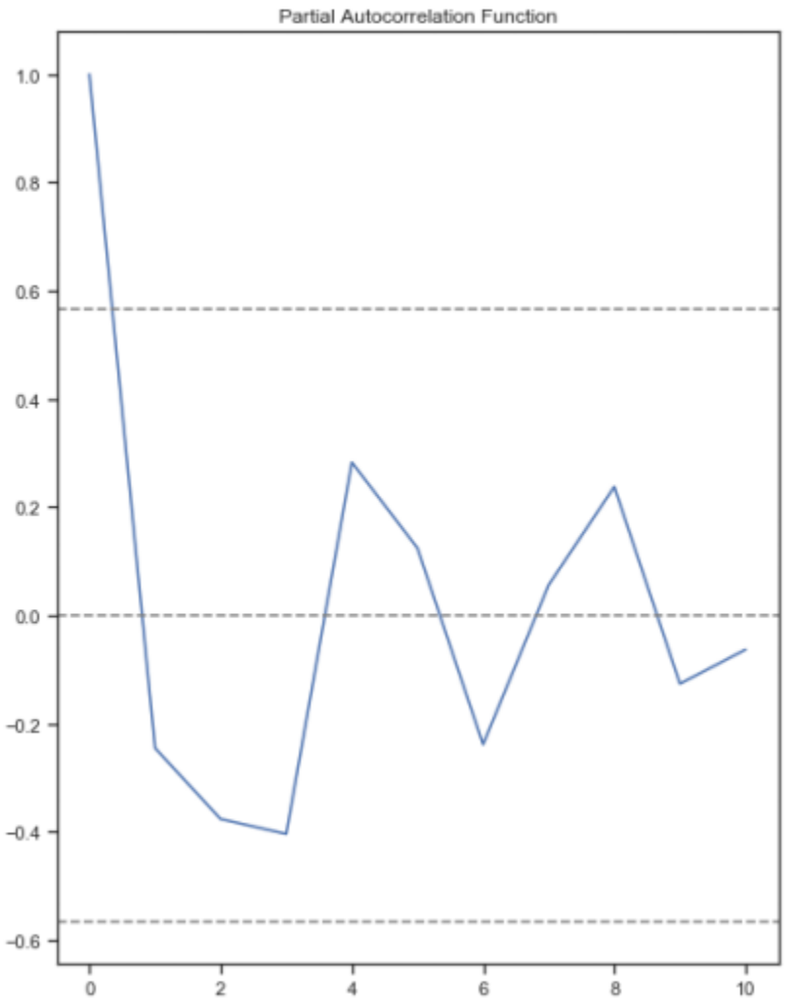
ACF plot

```
from statsmodels.tsa.stattools import acf, pacf
lag_acf = acf(ts_diff, nlags=10)
lag_pacf = pacf(ts_diff, nlags=10, method='ols')
#Plot ACF:
plt.subplot(121)
plt.plot(lag_acf)
plt.axhline(y=0,linestyle='--',color='gray')
plt.axhline(y=-1.96/np.sqrt(len(ts_diff)),linestyle='--',color='gray')
plt.axhline(y=1.96/np.sqrt(len(ts_diff)),linestyle='--',color='gray')
plt.title('Autocorrelation Function')
plt.tight_layout()
```



PACF plot

```
#Plot PACF:
plt.subplot(122)
plt.plot(lag_pacf)
plt.axhline(y=0,linestyle='--',color='gray')
plt.axhline(y=-1.96/np.sqrt(len(ts_diff)),linestyle='--',color='gray')
plt.axhline(y=1.96/np.sqrt(len(ts_diff)),linestyle='--',color='gray')
plt.title('Partial Autocorrelation Function')
plt.tight_layout()
```



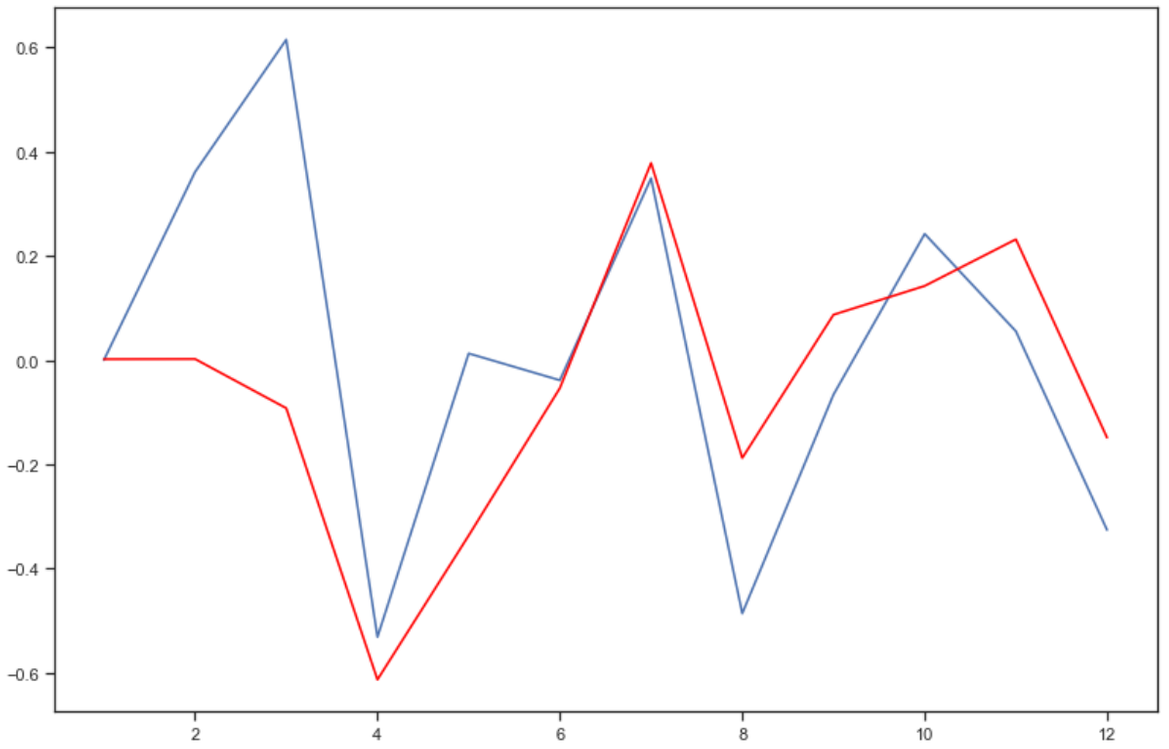
It was found from the above charts that p=0 & q=0. ARIMA stands for Auto-Regressive Integrated Moving Averages. The predictors depend on the parameters (p, d, q) of the ARIMA model:

1. Number of AR (Auto-Regressive) terms (p): AR terms are just lags of dependent variable.
2. Number of MA (Moving Average) terms (q): MA terms are lagged forecast errors in prediction equation.
3. Number of Differences (d): These are the number of differences. We will have to check for several combinations of p & q to decide which model is best for forecasting. ARIMA model was applied for several combinations of p, d, q and Residual Sum of Squares(RSS) was calculated to check which combination gives lowest RSS.

```
from statsmodels.tsa.arima_model import ARIMA
```

```
model = ARIMA(ts_diff, order=(3, 0, 0))
results_AR = model.fit(disp=-1)
results_AR.fittedvalues
```

```
plt.plot(ts_diff) plt.plot(results_AR.fittedvalues, color='red')
```



Finding sum of Residual Sum of Squares(RSS) for evaluation of ARIMA model

```
RSS = (results_AR.fittedvalues-ts_diff)**2 RSS.fillna(0,inplace=True) sum(RSS)
```

```
preds = results_AR.predict(start=12,end=24)
```

```
preds
preds = preds[1:]
predictions_ARIMA_diff = preds
predictions_ARIMA_diff
predictions_ARIMA_diff_cumsum = predictions_ARIMA_diff.cumsum()
ts_diff
predictions_ARIMA_diff_cumsum
```

```
predictions_ARIMA_log = pd.Series(4.8, index=range(13,25))
predictions_ARIMA_log
predictions_ARIMA_log = predictions_ARIMA_log.add(predictions_ARIMA_diff_cumsum,fill_value=0)

sum(ts_log)/12
predictions_ARIMA_log
predictions_ARIMA = np.exp(predictions_ARIMA_log)
predictions_ARIMA

plt.plot(ts)
plt.plot(predictions_ARIMA)
```

Plot of original time series and forecast values: