CASCADE CUP'20

DATA ANALYSIS REPORT

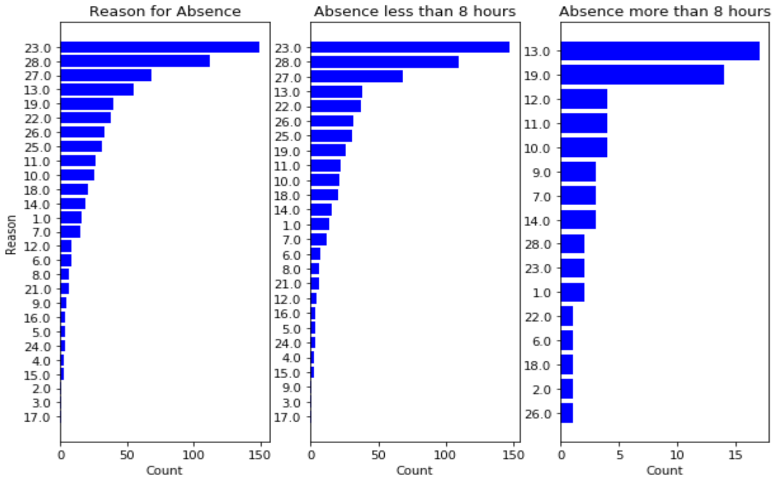
Team - The Insightful_2

Roshan Nayak
roshannayak610@gmail.com

Ullas Kannantha
ullaskannantha2@gmail.com

Absenteeism is any failure to report for or remain at work as scheduled, regardless of the reason. Absenteeism can have a severe impact on the workplace. We are trying find various factors that might play a role to absenteeism in the workplace.

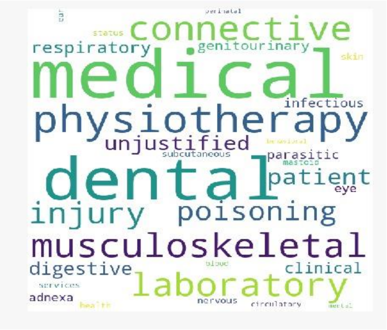
Which is the most common reason for Absence?



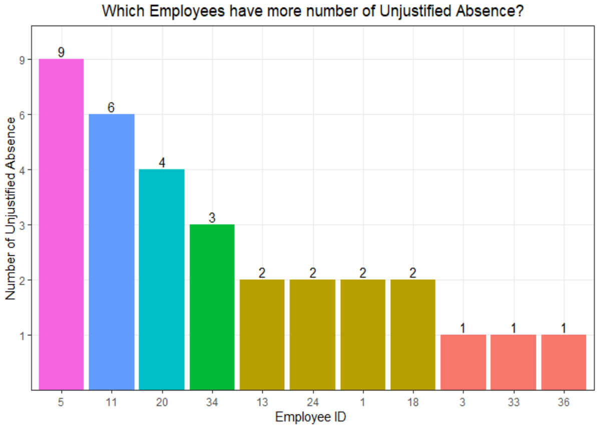
Leftmost plot shows the count of Reson for Absence. From the other two plots we can infer that Absents with more than 8 hours was due to some serious health issues and the ones with less than or equal to 8 hours were majorly due to less serious issues like medical or dental consultation.

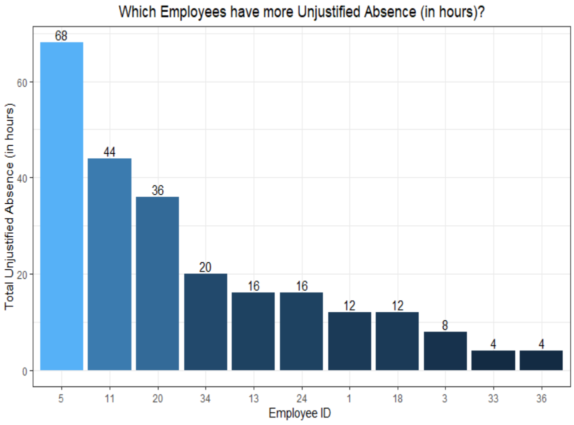
This word cloud gives us an idea about which terms were used more often in the reasons for absence. We see that words “medical” (*Reason 23*) and “dental” (*Reason 28*) were used many times.

(We modified the variable *Reason for Absence*. We shotened the name to fit it neatly in the wordcloud)

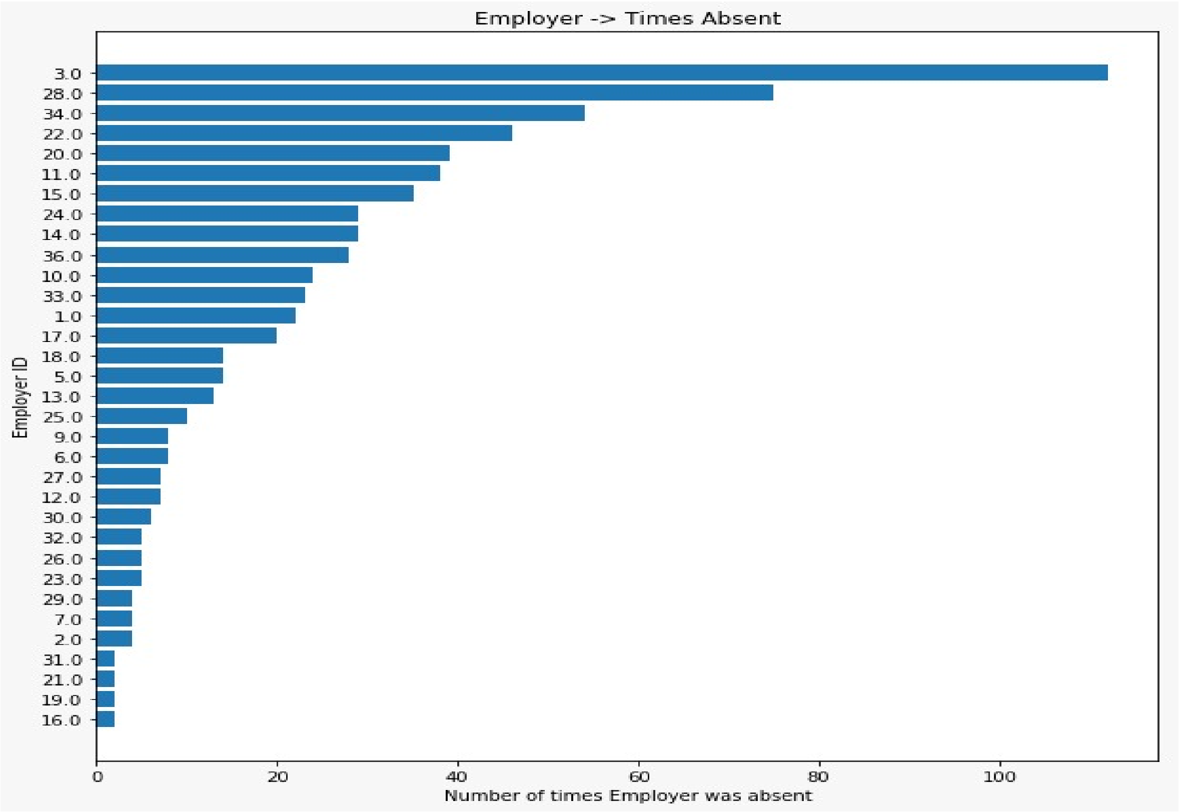


Unjustified Absence



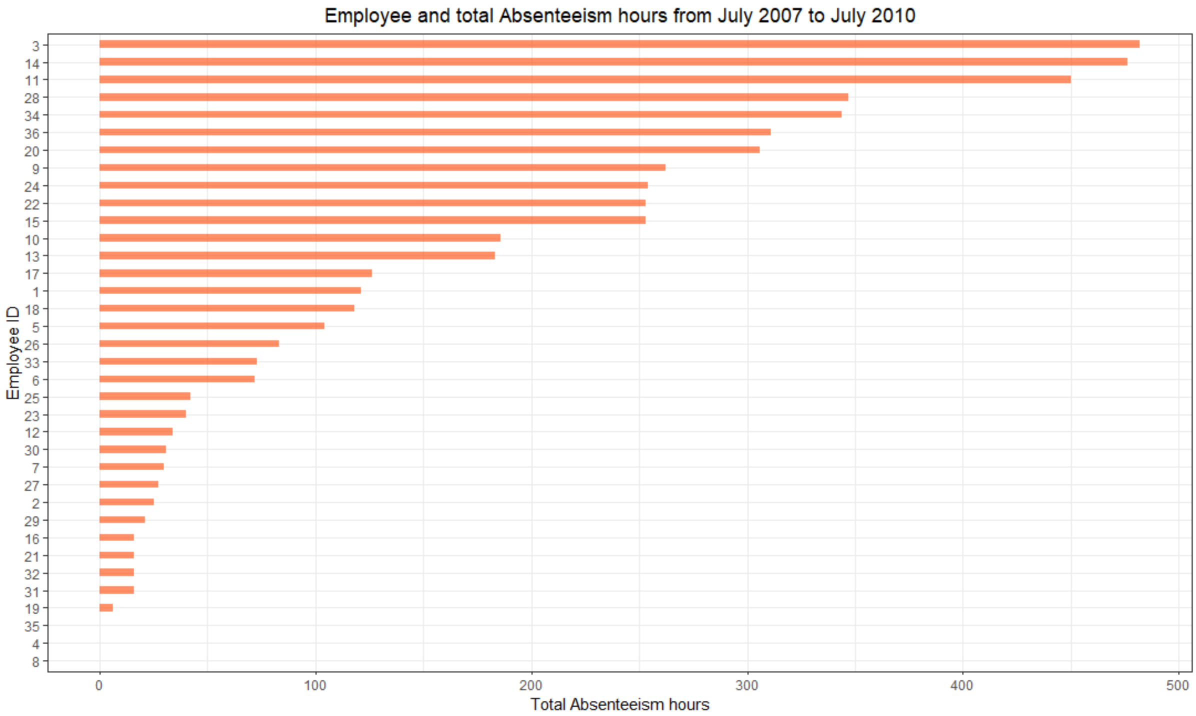


Employee 5 has more number of *Unjustified Absence* as well as long unjustified absence.



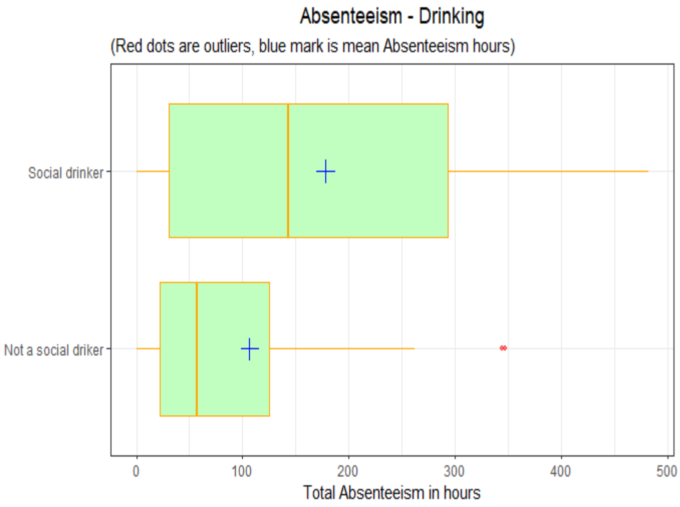
Employee *3* has more *number of absence.* Employee *16,19, 21* and *31* having low number of absence.

(We created a new variable by grouping employee ID and find his/her number of absence – It will be used further)

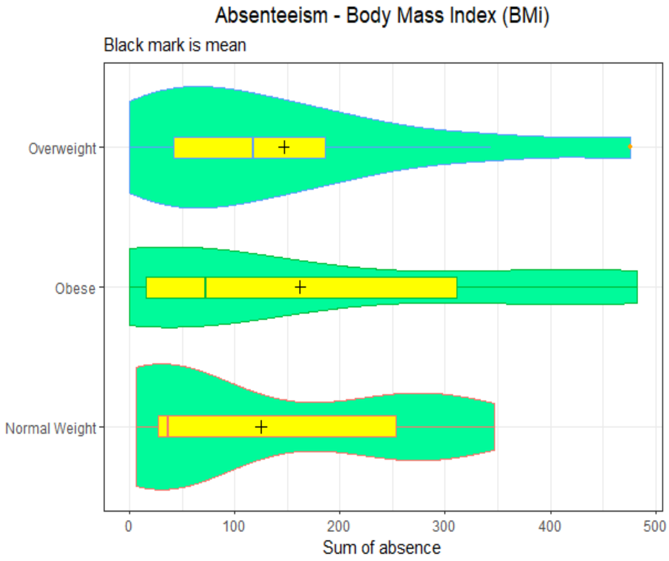


Employee *3*, *14* and *11* have more *number of absence.* Employee *19, 35, 4* and *8* having low total absence hours.

(We created a new variable by grouping employee ID and finding their total absence hour - It will be used further)

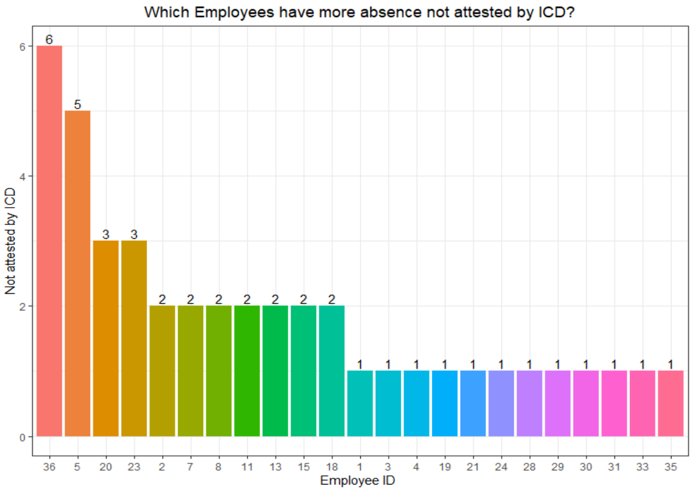


This is a boxplot showing how *drinking* is related to Absenteeism. We see that a Social drinker has more mean Total Absenteeism hours.



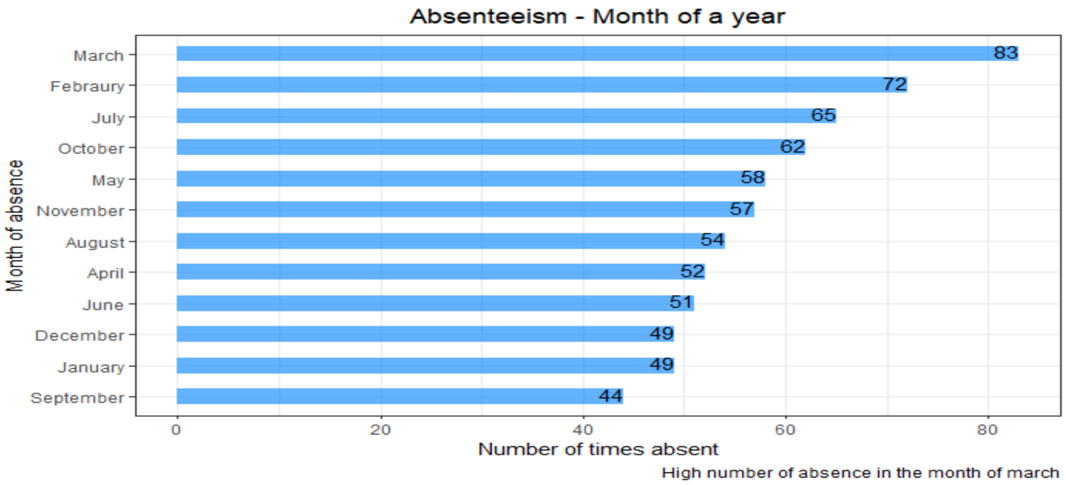
This violen plot is showing how *body mass* of a person correlates to Abesnteeism. We see tha Obese people have high mean Total Absenteeism hours.

(We categorized BMI 18.5–24.9 as Normal weight, BMI 25–29.9 as Overweight, BMI of 30 or greater as Obese. Reference:https://www.nhlbi.nih.gov/health/educational/lose_wt/BMI/bmicalc.htm)

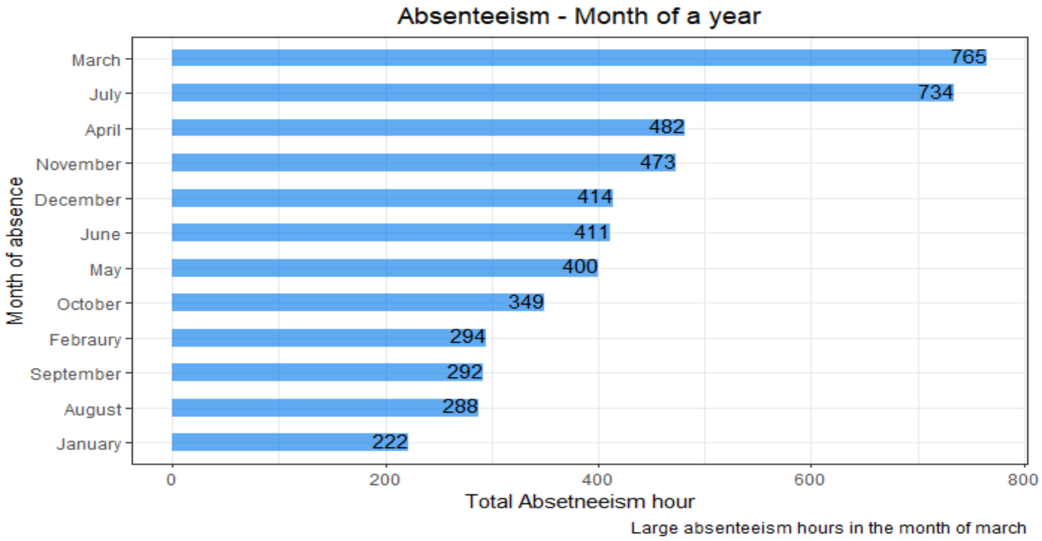


This bar plot shows which employee has high *Reason of absence* which are not attested by International Code of Diseases (ICD). We see that employee 36 has high number of absence which are not attested by ICD.

(Not attested by ICD were named as ‘0’ in Reason for absence in the dataset)

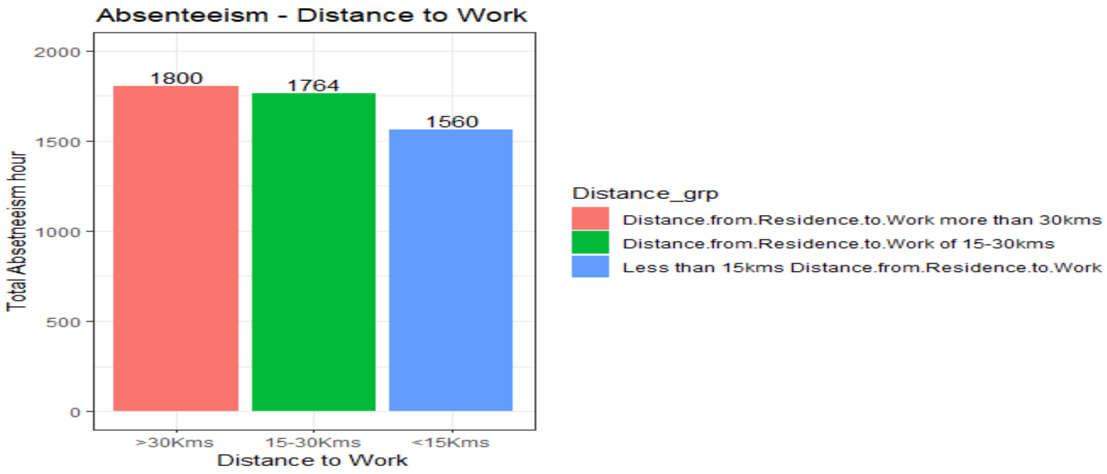


Above plot shows how *Month* of a year relates to Absenteeism (*Number of absence*)



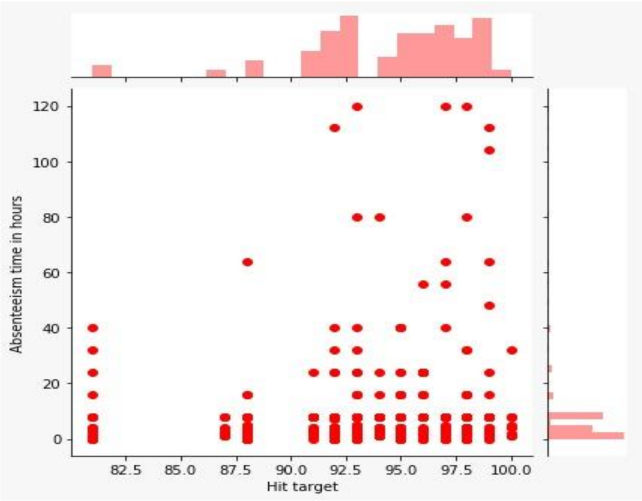
Above plot shows how *Month* of a year relates to Absenteeism (*Total time of absence*)

(For both the plots above each month number (i.e., 1, 2...) is converted into respective month name)

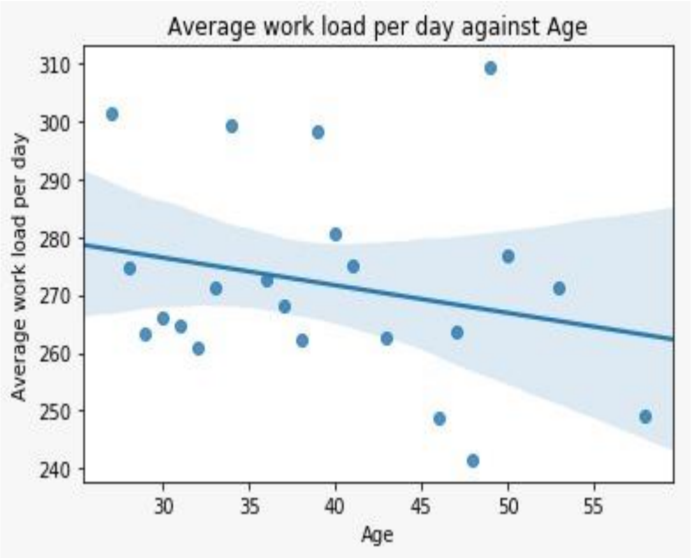


Above visualization shows how *distance to work* correlates with Absenteeism (Total Absence hours)

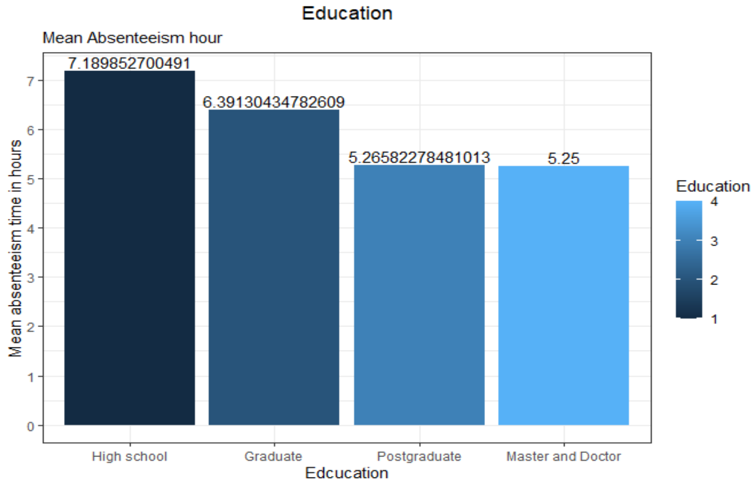
(Distance to work is categorized as indicated in the plot)



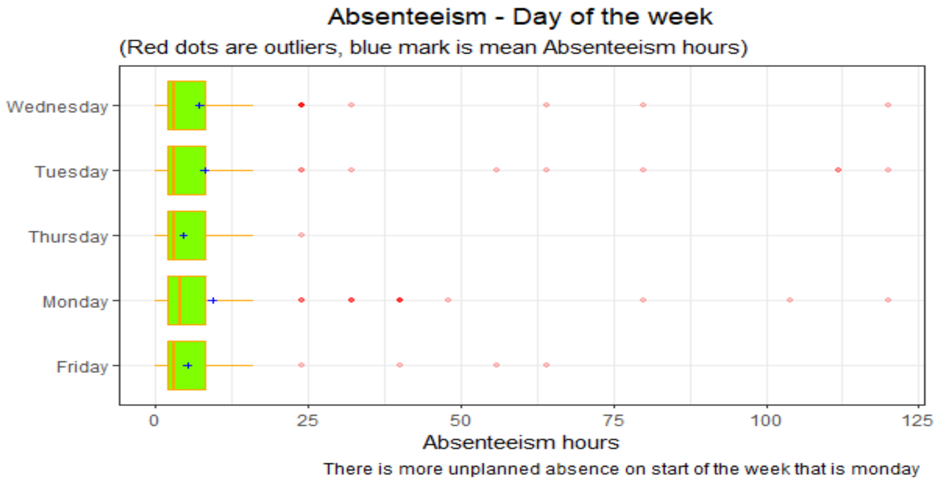
This visualization shows how *Hit target* and Absenteeism time in hours are correlated.



Here we are finding how *work load* and *Age* are correlated. There is negative correlation between age and work load. This would have been more clearer if it was not for one outlier here.



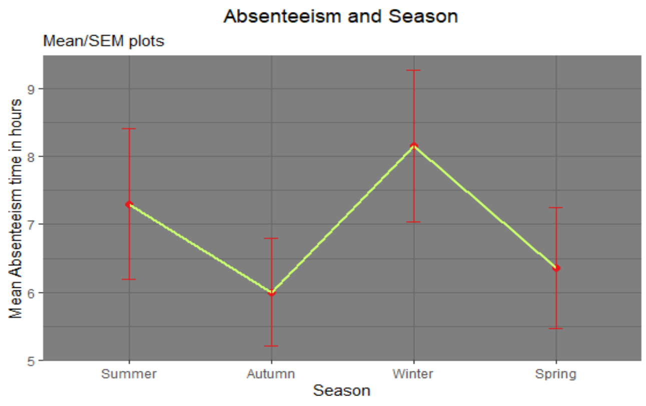
This barplot shows how *education* level influences Absenteeism.
Here we see that employee with highest education (Master and Doctor) has low mean Absenteeism time in hours. We also see that employee with lowest education level in the dataset (High school) has high mean Absenteeism time in hours.



There is more unplanned absence on start of the week that is monday

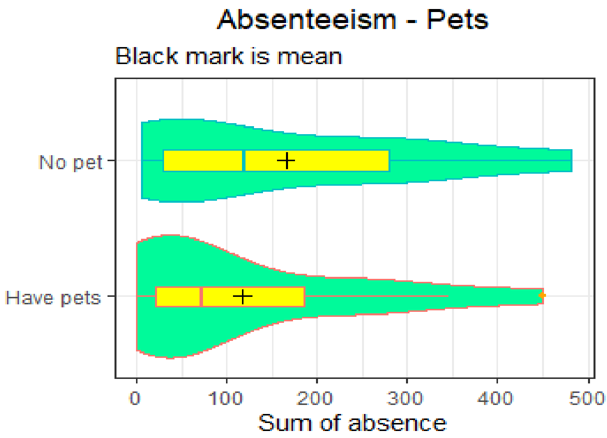
Above box plot shows how *Day of the week* and Absenteeism are correlated. We see that Monday has high mean high mean Absenteeism hour indicated by '+' and it gradually decreases over the week with lowest mean Absenteeism on Friday.

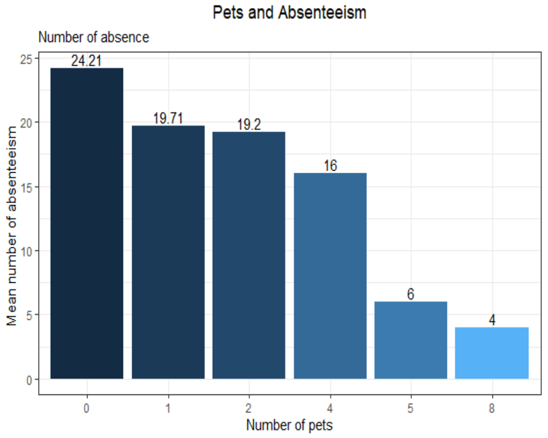
(We have renamed weeks which was indicated with number (1, 2, 3....) with their respective names)



This mean plot with error bars shows Absenteeism in each *season*. Here the error bars represent Confidence Interval (C I). We see that winter has high mean Absenteeism hour whereas Autumn has lowest.

(We have converted numerical naming of seasons (1, 2, 3...) to their respective names)

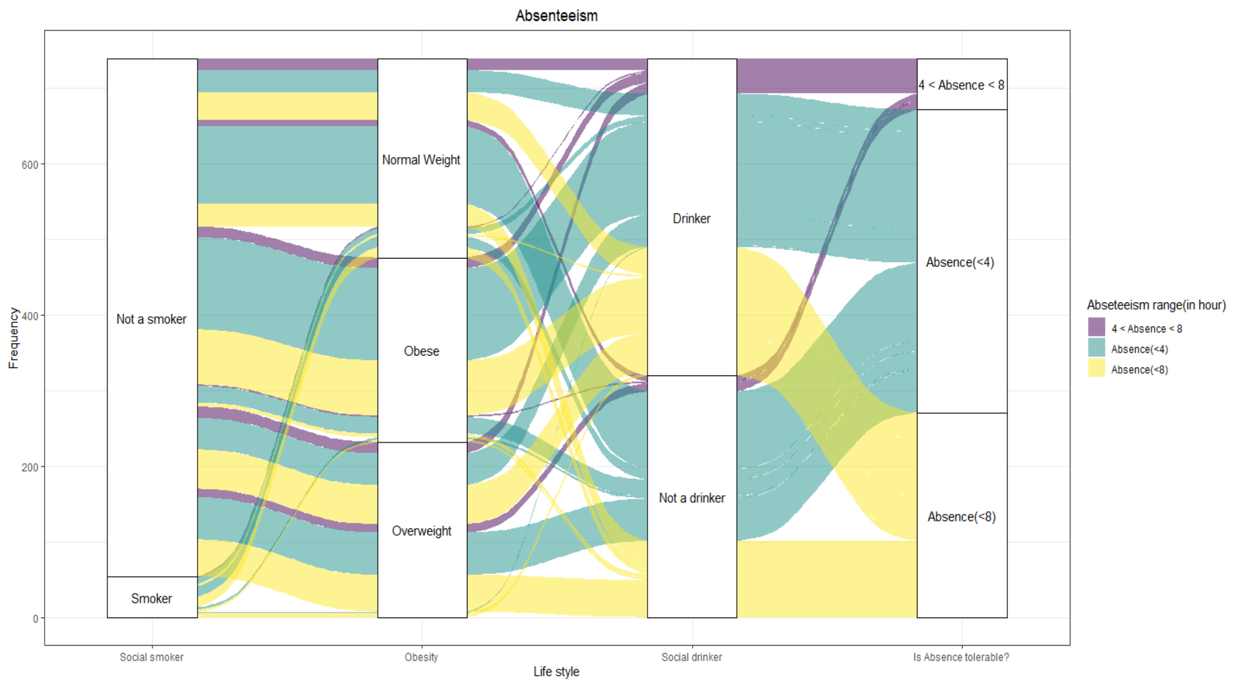




In

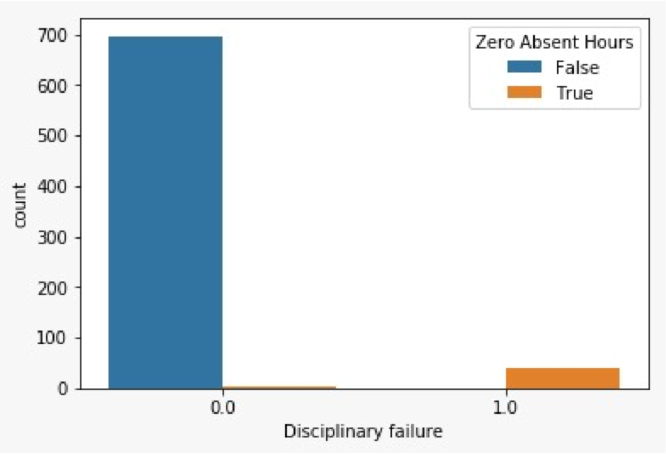
The above Violin plot we see the relation between *pets* at home and Absenteeism. We see that people with pets have low mean (indicated by '+') total absence hour while people without pets have high mean absence hour. Bar plot shows mean number of absences for each number of pets.

(Here we categorized employee with '0' pet as No pet and employee with more than 0 pet as Have pets)

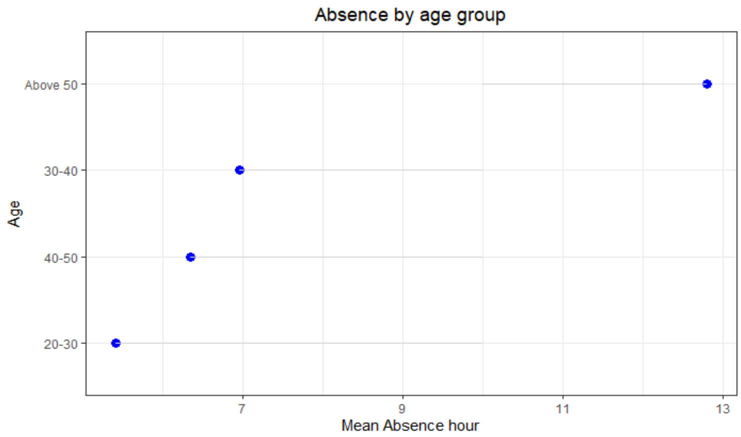


Above Alluvial diagram shows the relation among *Smoking habit*, *drinking habit*, *body mass* and *Absenteeism*. This visualization is about the personality of an employee and Absenteeism.

(Alluvial diagram is created with `ggalluvial` package (R package), generating `ggplot2` graphs)

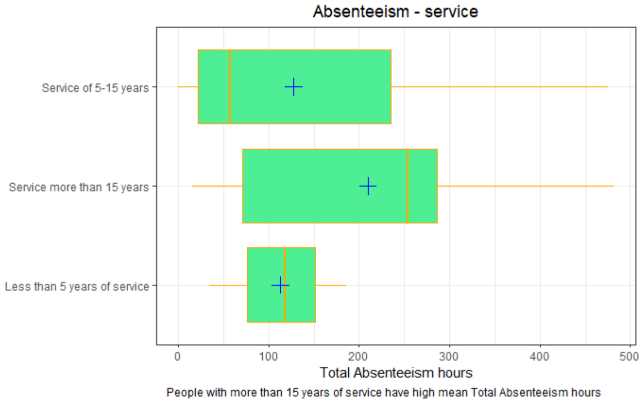


This visualization shows how *Disciplinary failure* is related to Absenteeism. We see that employees with Disciplinary failure as '0' has high number of absence hour which are above 0. We also see that employees with Disciplinary failure as '1' have high number of zero absence hour.



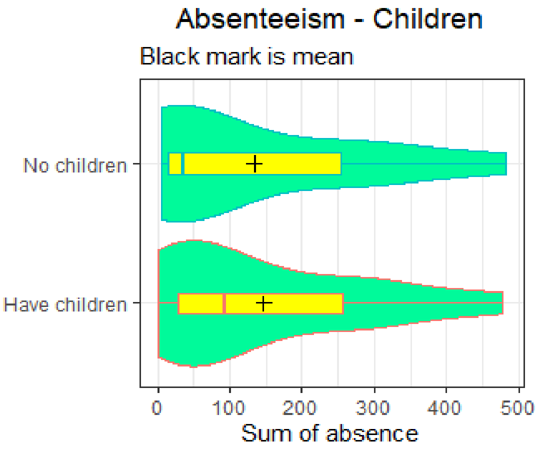
This visualization shows how *age* and *Absenteeism* are correlated. We see that employees above 50 years have high mean absence hour and it decreases as the age decreases with age group 20-30 having low mean absence hour.

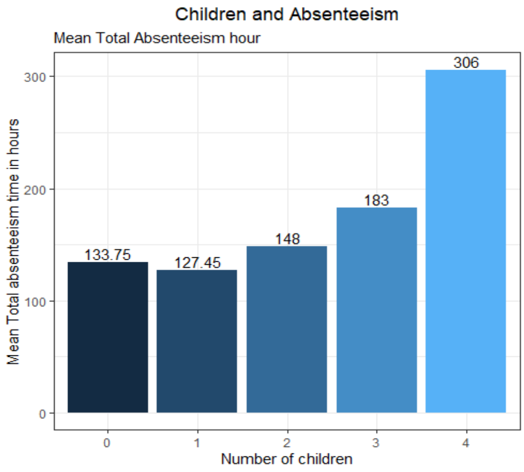
(Here we categorized the age as indicated in the visualization)



This box plot shows how *service time* is correlated with Absenteeism. We see that people with more than 15 years of service has high mean (indicated by '+') total absenteeism hour.

(We created a new variable which indicates the category of service as shown in this visualization)





This bar plot shows how *number of children* of an employee relates to Absenteeism. We see that employees with a greater number of children have high mean (indicated by '+') Absenteeism hour and it decreased as the number of children decreased.

Summary

- **Number of visualizations: 25.**
- **Number of attributes considered: 20.**
- **New attributes created:** *Total sum of absence hour, Total number of absences, Pet or No Pet category, Children or No Children category, Body mass category, Service time category, Age category, Distance to work category* and *Absence time category*.
- We found *reason of absence* is most important feature**.**
- We considered many criterions which might have correlation with Absenteeism. Mainly we considered how *family of an employee* affects Absenteeism and how *Employee life style* affects Absenteeism.
- **Attributes not considered: (2)** *Transportation expense* and *Height.* We found these attributes had low importance.

*All
you need is lots and lots of data and lots of information about what the right answer is, and you'll be able to train a big neural net to do what you want.*

-Geoffrey Hinton

"We can only see a short distance ahead, but we can see plenty there that needs to be done."

-Alan Turing

We thank Trell and Consulting & Analytics Club, IIT Guwahati for hosting "Cascade Cup - The Ultimate Data Science Challenge". We have learnt a lot during all 3 stages and we are thankful for giving us this opportunity to get hands on experience with real world data.

Team - The Insightful_2

Roshan Nayak
roshannayak610@gmail.com

Ullas Kannantha
ullaskannantha2@gmail.com