Cascade Cup Final Round

Data Analysis Report

TEAM:DataFreaks

Shreya Sajal ,IIT Guwahati

Shubham Mondal ,IIT Guwahati

Exploratory Data Analysis

Introduction

In this EDA report we have analyzed the **Absenteeism (unplanned absences or habitual pattern of absence from a duty or obligation without good reason)** in a company whose database was created with records of absenteeism at work from July 2007 to July 2010 at a courier company in Brazil.

Through the analysis, we have tried to explore all the important columns that would give an insight into the factors associated with the problem of absenteeism. The motive is to help the company devise strategies to deal with the problem.

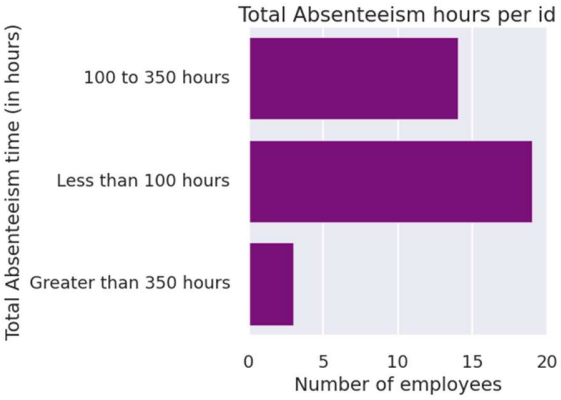
Cleaning and Preprocessing

Removed the **34 duplicates** and corrected mistyped values in the 'ID', 'Reason for absence' and 'Month of absence' columns

- The Reason for absence column in the dataset contains 28 categories, marked from 1 to 28. The 20th category was originally marked as 0 which was corrected later.
- ID had a column corresponding to 28^{th} ID marked as 29
- Three rows had 'Month of absence' value 0 that was replaced with the help of Work Load Avg per day column to the closest corresponding months

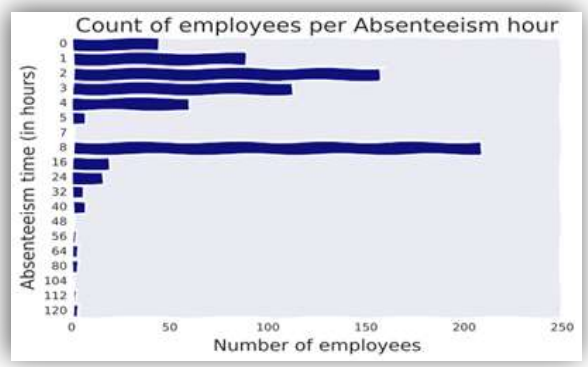
Absenteeism hours distribution

THE PLOT SHOWS THE DISTRIBUTION OF TOTAL ABSENTEEISM HOURS (summing all instances) BY 36 DIFFERENT EMPLOYEES OVER THE ENTIRE DATASET.



- The highest frequency of employees fall in the range of total absenteeism hours less than 100 and the lowest seen absenteeism range is more than 350 hours (expected, as unplanned absence of around half a month by any employee would only be in some extraordinary circumstances).

THE PLOT BELOW SHOWS THE DISTRIBUTION OF THE 19 DIFFERENT ABSENTEEISM TIME VALUES OVER THE ENTIRE DATASET

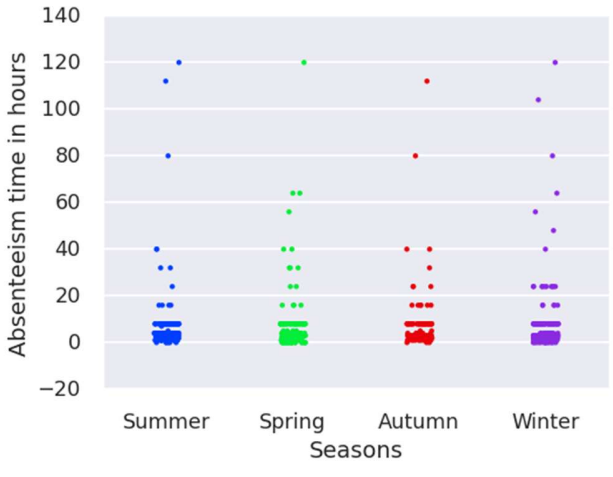


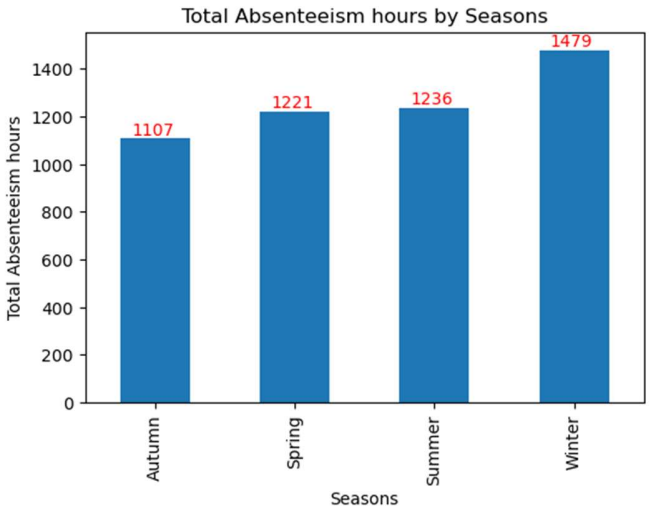
- It can be seen that 8 hours of unplanned absence by the employees was witnessed the maximum times in the company. As the dataset has 36 unique IDs of employees, multiple employees, at different instances, contributed their share of absenteeism hours, adding up to the numbers on y-axis, the maximum share being for 8 hours.

- The 20 plus hours of absenteeism has very low frequencies among the employees.

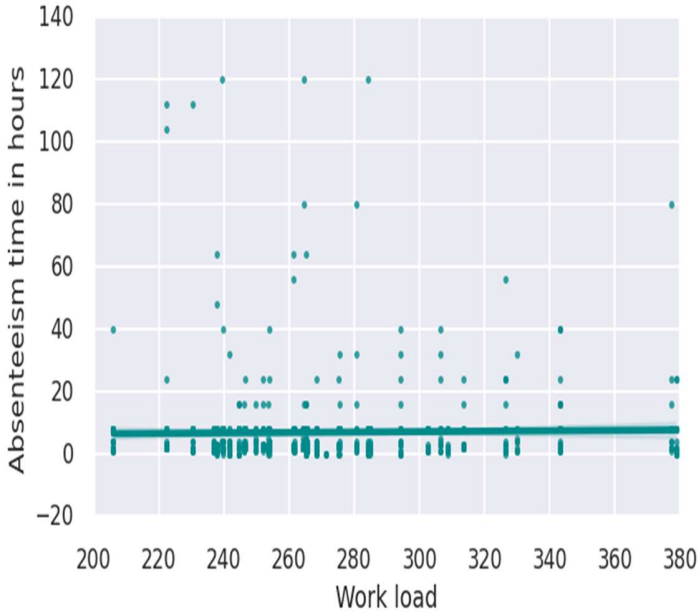
SEASON

All the four seasons: Summer, Spring, Autumn and Winter show almost similar contribution, yet spring having the highest count in the dataset.
Looking at the total absenteeism hours by seasons, we can see that the highest total unplanned absence is seen in winter, after grouping the data by seasons.



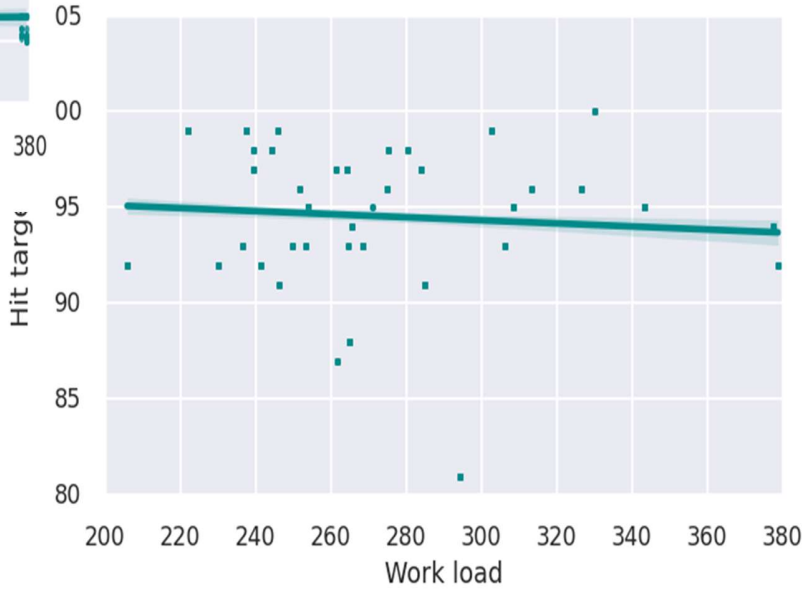


WORK LOAD AVERAGE/DAY:



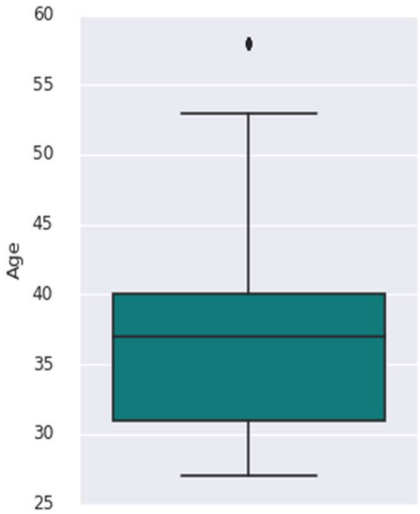
- The average workload per day shows little positive correlation with Absenteeism time in hours.
- The hit target value shows negative correlation with work load average per day.
- This result is somewhat practical as the target hitting will decrease with increased workload.

Reducing the workload average per day may help the company increase the hit target value and reduce the absenteeism hours as well.



Age of The Employee:

The boxplot shows the age distribution of the 36 employees. Half of the employees are below 37, 75% below 40, minimum age being 27 and highest 58 among all.

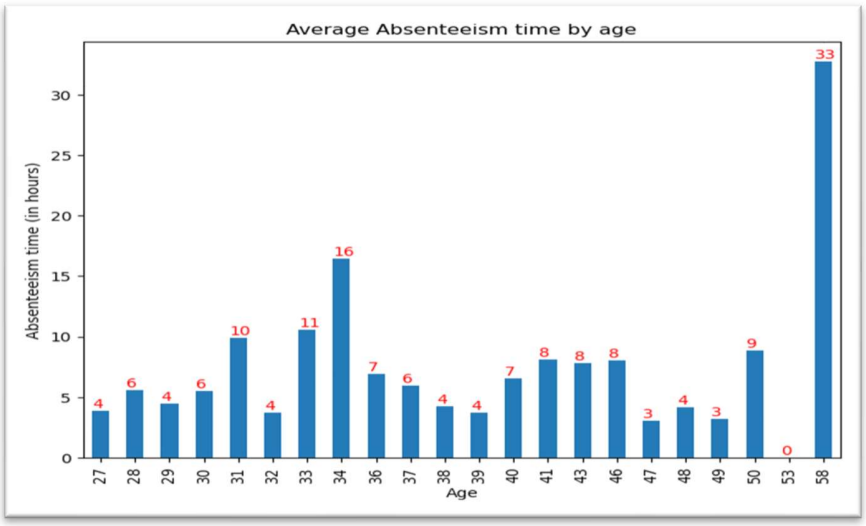


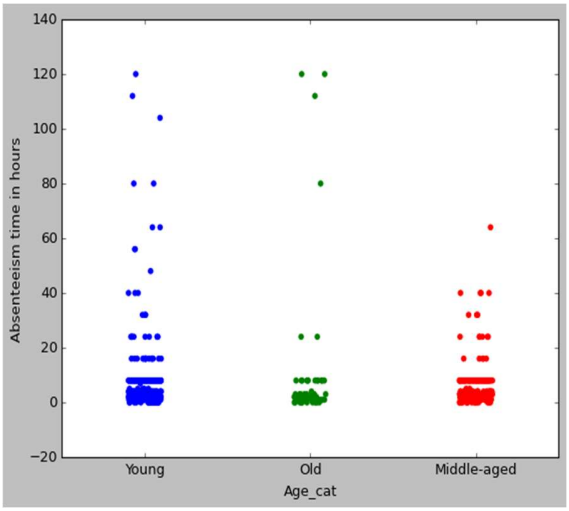
The plot below shows the variation of the average absenteeism time with the age of the employees. We grouped the dataset on the basis of age and then calculating the mean of the absenteeism hours corresponding to a particular age.

In the third graph, the categories were divided as: less than 37 years-Young,37-47 years-middle aged,47 above -old.

From the graphs, no strict trend is obtained in absenteeism hours with variation in age.

- The middle-aged group is seen to wrap its per-instance absenteeism contribution in 50 hours. The absenteeism hours by young employees cover wider range.





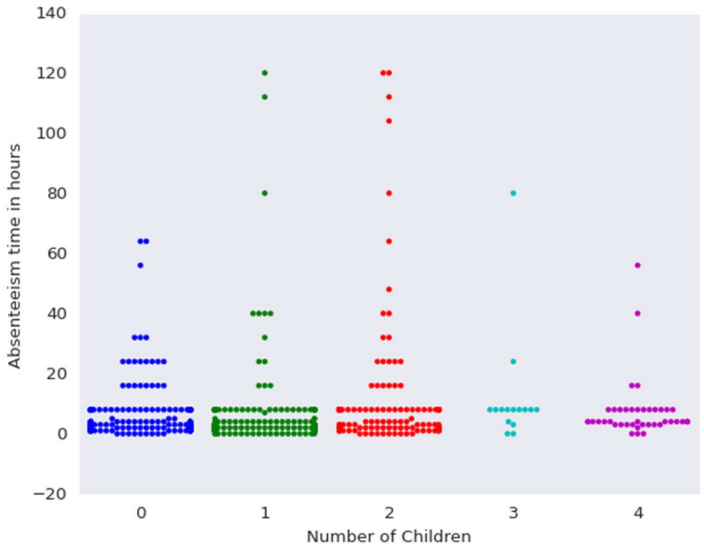
The 47-above age group does not fully cover higher (20+) absenteeism hours but still has some points in 80+ range indicating special reasons.

Number of Children:

The employees with number of children 1 or 2 populate the above 80 absenteeism hours.

In the absenteeism range, the upper limit is seen to increase with 0 to 1 child and then decrease from 2 to 4 children.

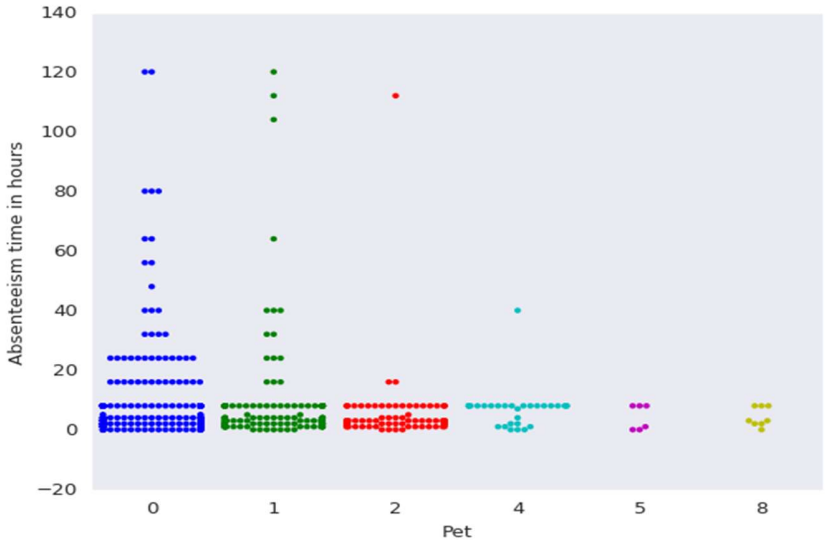
Looks like having more than two children implies lesser reasons for absence than a single child or two children.



Number of Pets:

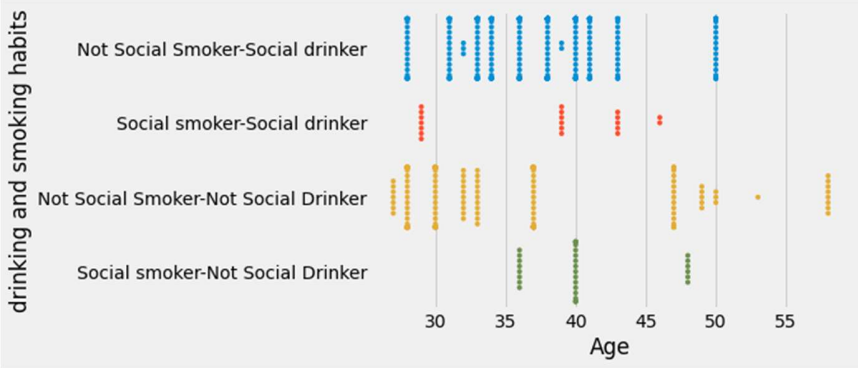
In the absenteeism range, the upper limit is clearly seen to decrease with increase in the number of pets the employee has.

Even though the data points are less for higher values of number of pets, the available points are seen to majorly lie lower ranges.

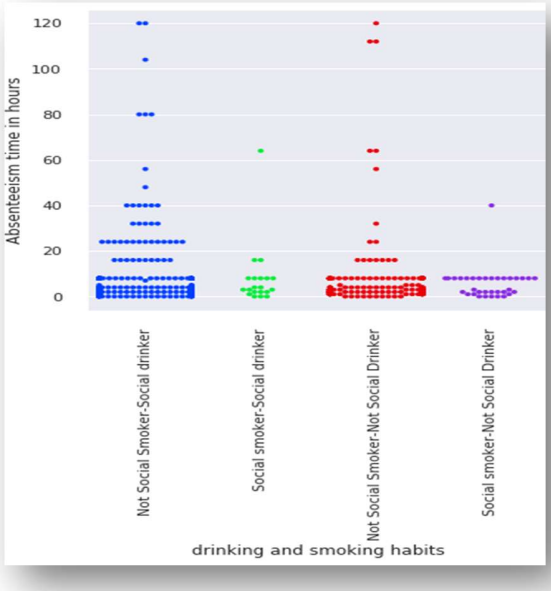


Drinking and smoking habits:

Distribution with Age:The swarmplot shows the distribution of the people with 4 different drinking and smoking habits across the age groups.

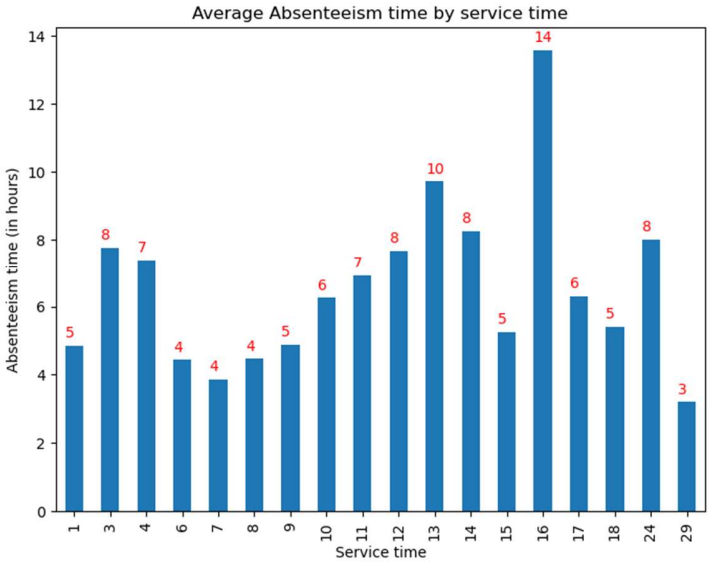


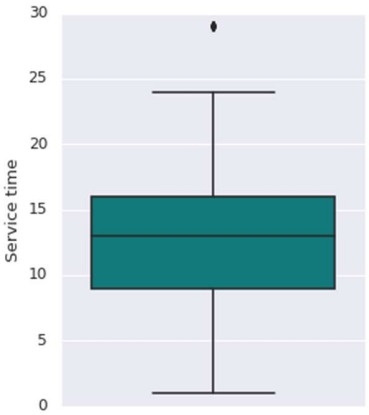
- The elder ones with age more than 47 are neither smokers not drinkers.(they have had enough)
- The younger people with age less than 37 (the median age) either have no such habits or are exclusively social drinkers and both drinker and smokers in a few instances.(getting bad habits)
- Major part of the people having both drinking and smoking habits fall in the mid age group.And,almost no instances of people free of both these habits is seen in this group.(bad habits at their peak)



Distribution with Absenteeism time in hours:

- The higher absenteeism hours in the data is majorly contributed by exclusive drinkers or neither drinker nor smokers.Not much can be concluded as the classes are highly imbalanced.
- Max absenteeismDrinkers&smokers>Max absenteeism hour-only smokers

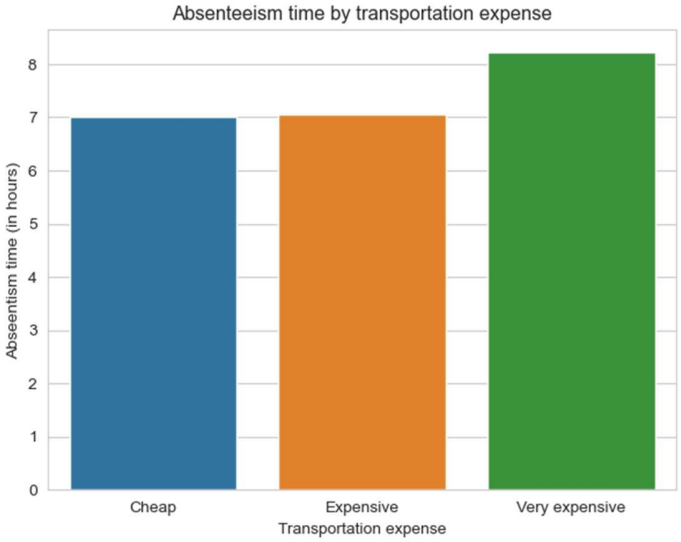




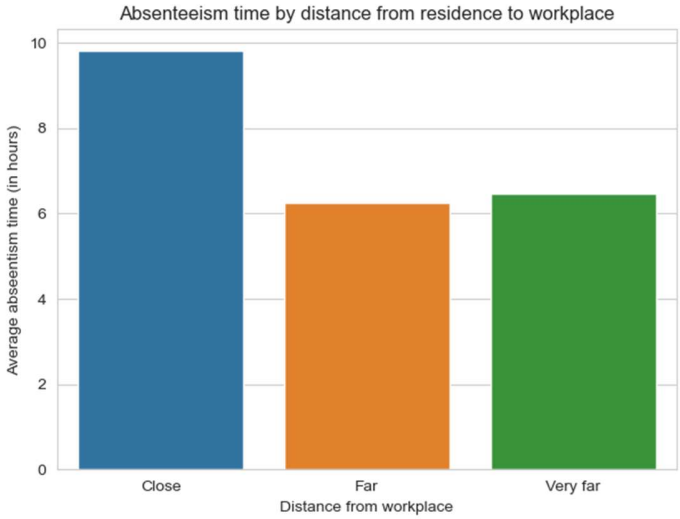
Service Time:

- Half of the employees are with less than 12 years of service experience as would be expected as half of them are below 37 years of age.
- 1 employee is seen with 29 years of service experience and contributing 3 hours of average absenteeism.
- Seems to shows little positive correlation with absence hours. In the mid -range ,the absenteeism increases with service years till peak at 16 years

Distance from work to home and Transportation expenses:

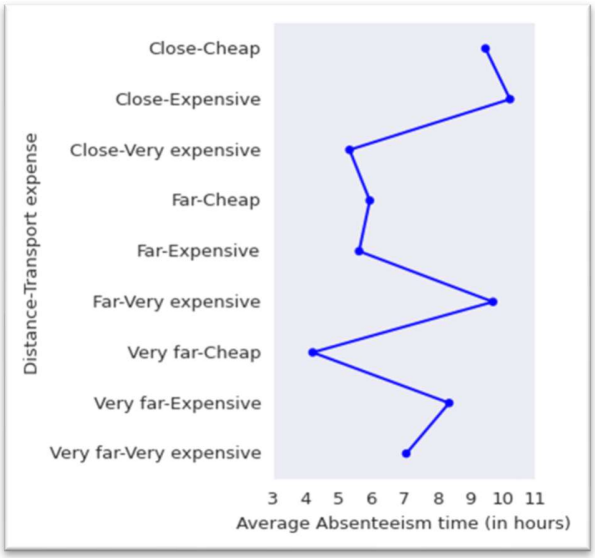


The first two graphs show the distribution of average absenteeism hours with transportation expense and distance of workplace to home of the employees.



How do we know which of the two features is more significantly impacting the average absenteeism hours?

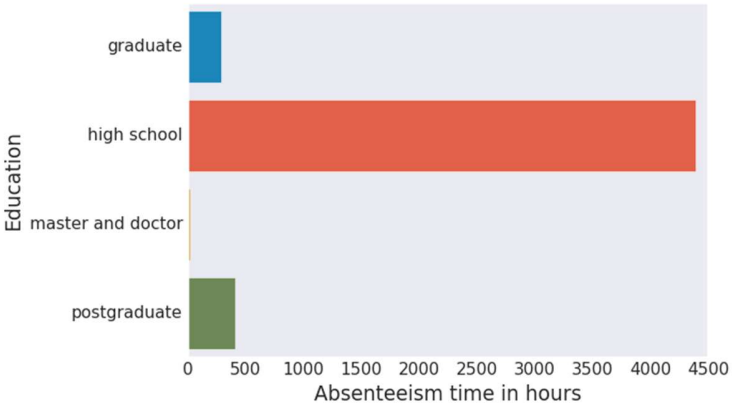
Here,Cheap:Less than 200,expensive:200-300,very expensive:300 plus



Here,Close:less than 15km,Far:15-35km,Very far:35 plus km

To get further insights, we combined the two features and grouped the dataset by the newly created feature and obtained the plot below. So, what can be inferred from this?The plot does not show quite expected trends suggesting that there are features more important than this in determining absenteeism.

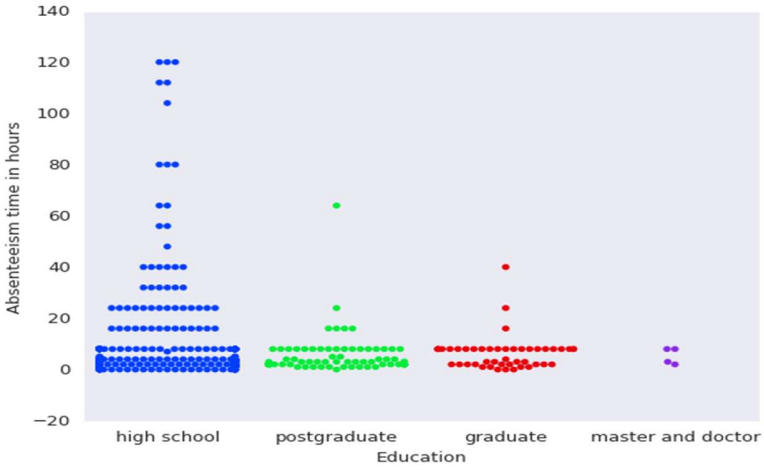
Of the 9 categories,the ones topping the avg absenteeism are Close-Expensive,Far-Very expensive,Close-Cheap,Very far-Expensive and the one with lowest is Very far-Cheap.So we can see that the transportation cost plays a more important role than distance



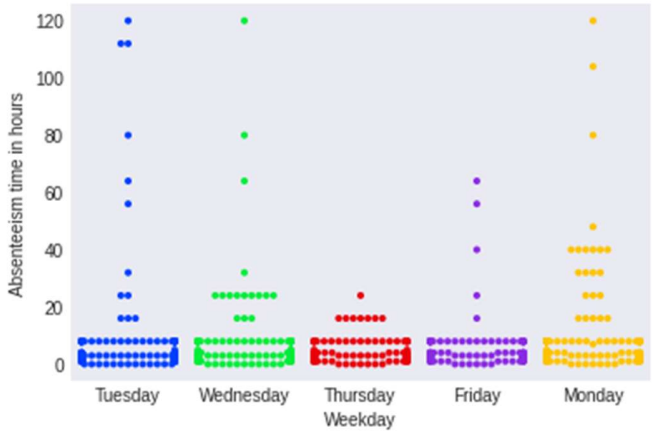
Employee Education:

- The employee education classes are highly imbalanced and don't follow any orderly trend

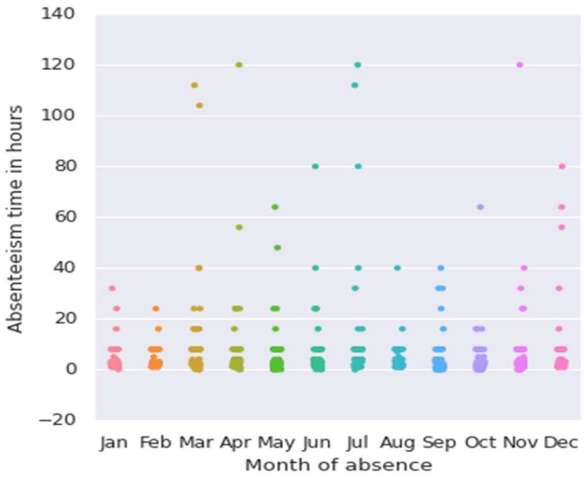
- The swarmplot shows the distribution and range trends: quite high absenteeism hour ranges for high school level educated employees, almost similar intermediate range for graduate and postgraduate and the least for master and doctor level educated employees.



DAY OF THE WEEK AND MONTH:

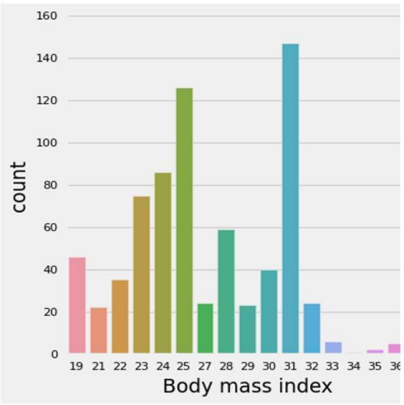


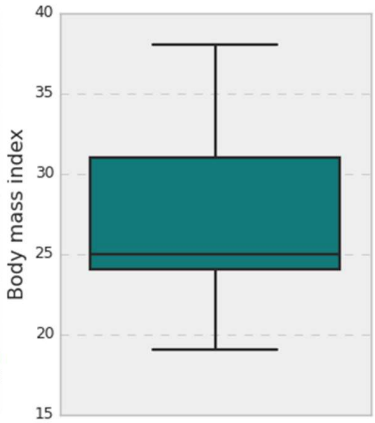
- The month and day of the week also seem to impact the unplanned absence hours in employees.
- In months,the beiginning of the year sees lower range of unplanned absence hours(maybe some new year resolutions) but with the proceeding months the range is seen to widen with similar distribution seen in March to July .
- Towards the end of the year the absenteeism hours range can be seen to comparatively increase(time for new resolutions)
- Grouping our data by weekdays,Monday,Tuesday and Wednesday witness higher hours of employee unplanned absence.
- Thursday has the lowest range of absenteeism hours wrapping under 25 hours with a sudden widening hour range on Friday(maybe greed of 1 unplanned+2 sanctioned offs continuously till Monday trapped some)

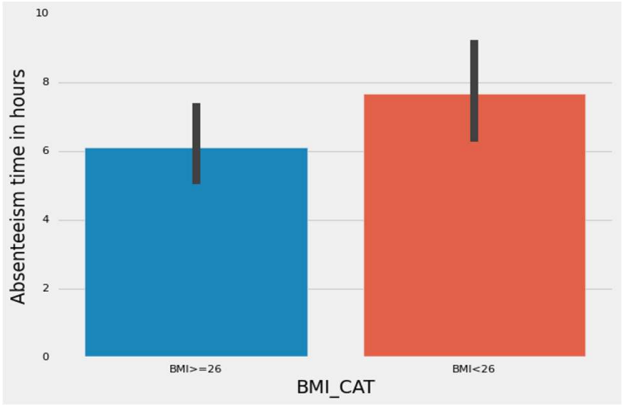


BODY MASS INDEX:

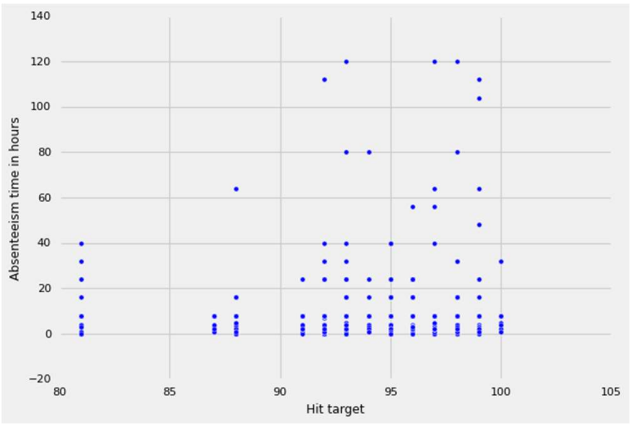
- The first two plots show the distribution of Body mass index (BMI)of employees over the entire dataset.
- Half of the employees have BMI less than 25 and as we know half of the employees are young (less than 37yrs old), this makes sense (with some exceptional points).
- Dividing the BMI field with threshold body mass index value 26, we get the last barplot that shows some negative correlation between BMI and absenteeism hrs.





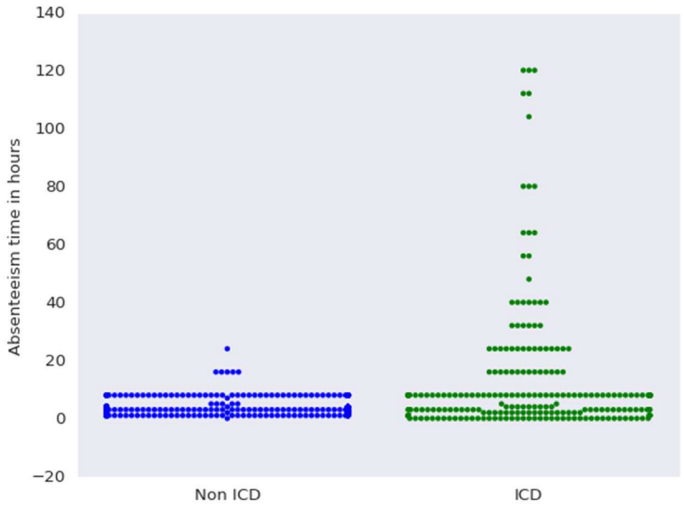


HIT TARGET:

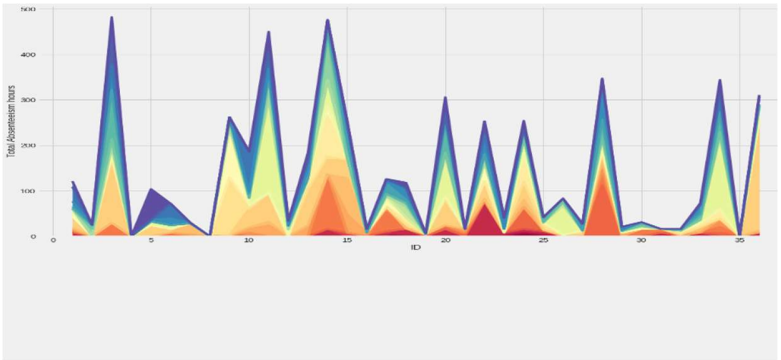


- For higher values of target hit, higher absenteeism hours is seen.
- The below 87 target hit instances have absence hours under 40 whereas 90 above have ranges upto 120 hours per instance

REASON FOR ABSENCE:

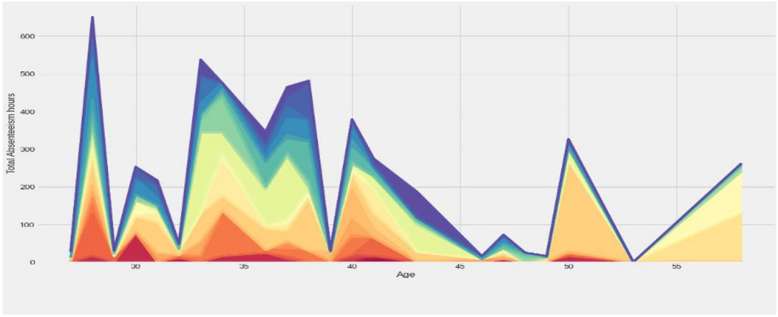


- The swarmplot on right shows the absenteeism hour distribution with the reason of absence.
- Here ICD is a list of 21 categories of absences attested by the International Code of Diseases (ICD).
- The Non-ICD are 7 categories without (ICD)
- The ICD category absence reason in green has higher ranges of absenteeism hours than the Non-ICD ones. This is similar to what we would expect as the reasons in the 21 categories are serious enough to require more attention and time.
- **The ICD categories are covered in the yellow to red shade of the spectrum and non-ICD in blue in the right-side graph and the graph in the following page.**



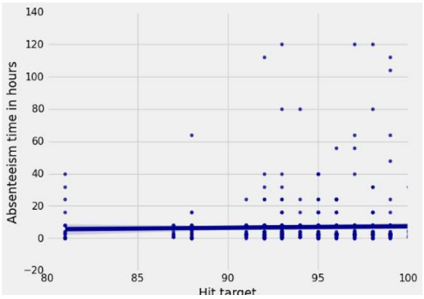
- The graph shows distribution of absence reasons with total absenteeism hours (y-axis) over all instances per employee ID(x-axis).

- The major portion of reasons for larger absence hours is covered with ICD reasons
- The graph below shows a similar distribution except for age being on the x-axis.
- Here we can notice an increase in the ICD reasons for similar absenteeism hours with increase in age.

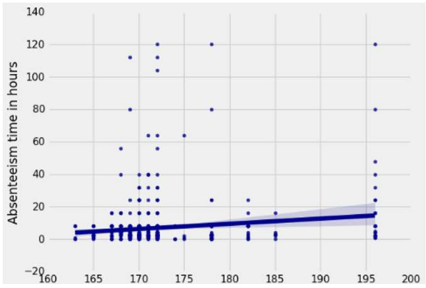


- The conclusions from two graphs can be established as: Higher hours absence reasons are majorly ICD .With higher age the percentage of ICD absence increases when compared to similar absence hours for lower ages.

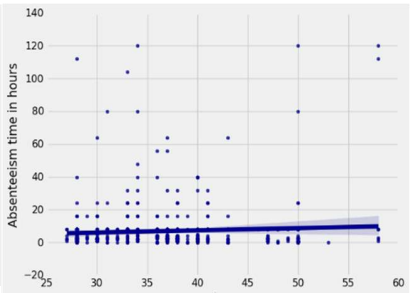
Fields positively correlated with absenteeism:



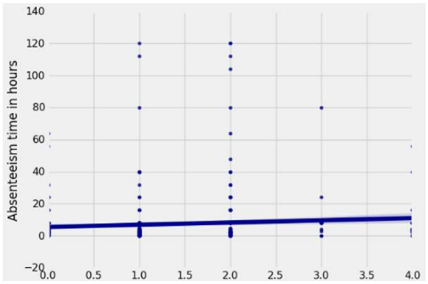
Hit Target



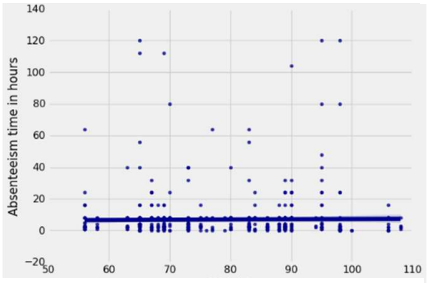
Height



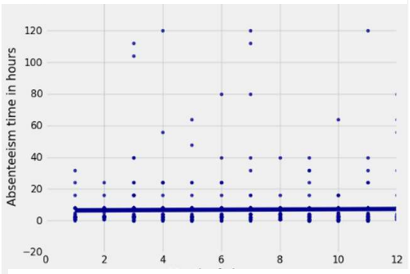
Age



Children



Weight

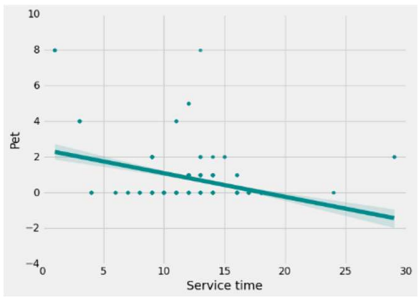


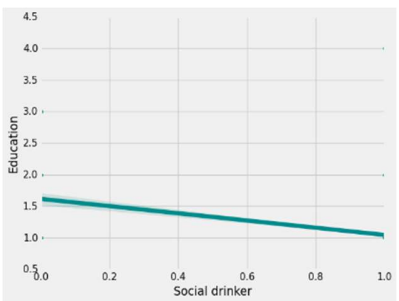
Month of Absence

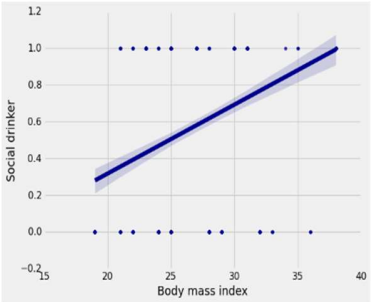
- From the correlation matrix and a heatmap, these fields were found to have, although low, but positive correlation with the absenteeism hours.
- These observations are close to our previous conclusions above

Social Drinker

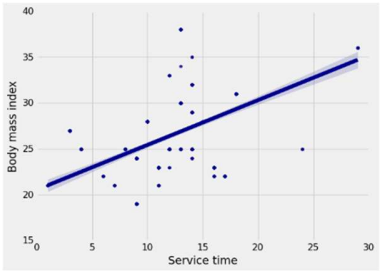
Other interesting and significant correlations:

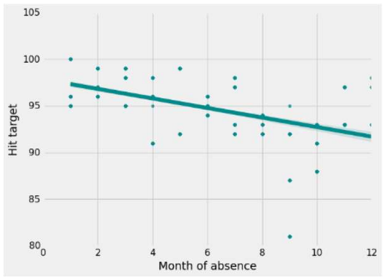






Other than having correlation with absenteeism hours, some other fields in the dataset were found to display striking correlations with each other.

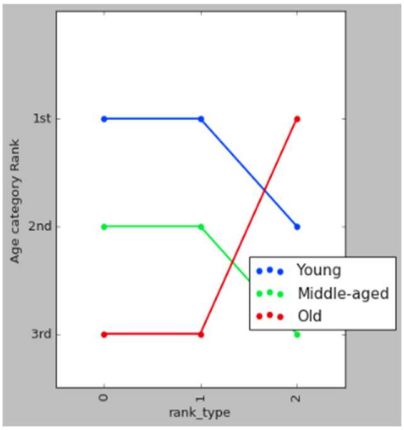


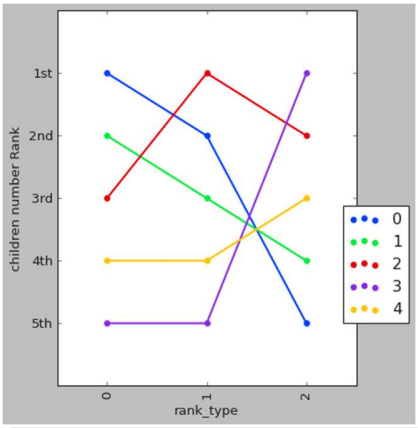


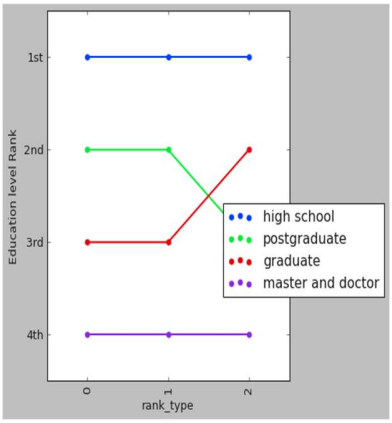
- The ones in cyan color display negative correlation and the blue ones display positive.
- With proceeding months in a year, the target hit value is seen to decrease.
- The Employees with higher service period have higher body mass indices
- The number of pets is seen to decrease with increase in the service period of an employee.
- Social drinking habits display prominence in people with higher BMI and lower education
- Note that these are just correlations and not causations that we are implying.

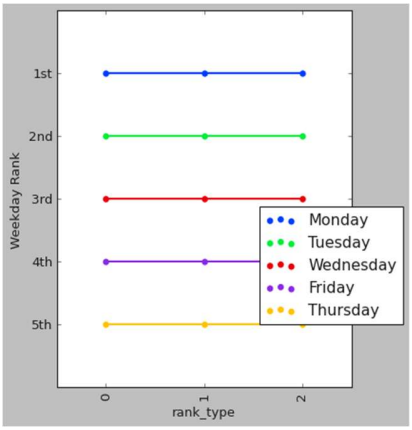
RANKING FEATURES CATEGORICALLY ON 3 METRICS:

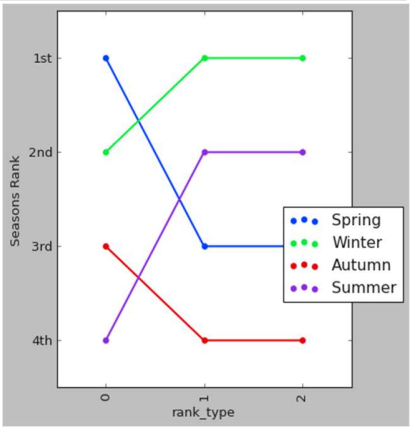
- The three ranking metrics are **VALUE COUNTS (0), TOTAL ABSENTEEISM HOURS (1) AND AVERAGE ABSENTEEISM HOURS (2)** respectively.
- The rankings obtained are more or less similar to our conclusions whie discussing these features.











THANKYOU